



# Some basic notation and background

Regression

Brian Caffo, PhD  
Johns Hopkins Bloomberg School of Public Health

# Some basic definitions

- In this module, we'll cover some basic definitions and notation used throughout the class.
- We will try to minimize the amount of mathematics required for this class.
- No calculus is required.

# Notation for data

- We write  $X_1, X_2, \dots, X_n$  to describe  $n$  data points.
- As an example, consider the data set  $\{1, 2, 5\}$  then
  - $X_1 = 1, X_2 = 2, X_3 = 5$  and  $n = 3$ .
- We often use a different letter than  $X$ , such as  $Y_1, \dots, Y_n$ .
- We will typically use Greek letters for things we don't know. Such as,  $\mu$  is a mean that we'd like to estimate.
- We will use capital letters for conceptual values of the variables and lowercase letters for realized values.
  - So this way we can write  $P(X_i > x)$ .
  - $X_i$  is a conceptual random variable.
  - $x$  is a number that we plug into.

# The empirical mean

- Define the empirical mean as

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

- Notice if we subtract the mean from data points, we get data that has mean 0. That is, if we define

$$\tilde{X}_i = X_i - \bar{X}.$$

The the mean of the  $\tilde{X}_i$  is 0.

- This process is called "centering" the random variables.
- The mean is a measure of central tendency of the data.
- Recall from the previous lecture that the mean is the least squares solution for minimizing

$$\sum_{i=1}^n (X_i - \mu)^2$$

# The empirical standard deviation and variance

- Define the empirical variance as

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \left( \sum_{i=1}^n X_i^2 - n\bar{X}^2 \right)$$

- The empirical standard deviation is defined as  $S = \sqrt{S^2}$ . Notice that the standard deviation has the same units as the data.
- The data defined by  $X_i$ 's have empirical standard deviation 1. This is called "scaling" the data.
- The empirical standard deviation is a measure of spread.
- Sometimes people divide by  $n$  rather than  $n - 1$  (the latter produces an unbiased estimate.)

# Normalization

- The the data defined by

$$Z_i = \frac{X_i - \bar{X}}{s}$$

have empirical mean zero and empirical standard deviation 1.

- The process of centering then scaling the data is called "normalizing" the data.
- Normalized data are centered at 0 and have units equal to standard deviations of the original data.
- Example, a value of 2 form normalized data means that data point was two standard deviations larger than the mean.

# The empirical covariance

- Consider now when we have pairs of data,  $(X_i, Y_i)$ .
- Their empirical covariance is

$$\text{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n-1} \left( \sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y} \right)$$

- Some people prefer to divide by  $n$  rather than  $n - 1$  (the latter produces an unbiased estimate.)
- The correlation is defined is

$$\text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{S_x S_y}$$

where  $S_x$  and  $S_y$  are the estimates of standard deviations for the  $X$  observations and  $Y$  observations, respectively.

# Some facts about correlation

- $\text{Cor}(X, Y) = \text{Cor}(Y, X)$
- $-1 \leq \text{Cor}(X, Y) \leq 1$
- $\text{Cor}(X, Y) = 1$  and  $\text{Cor}(X, Y) = -1$  only when the X or Y observations fall perfectly on a positive or negative sloped line, respectively.
- $\text{Cor}(X, Y)$  measures the strength of the linear relationship between the X and Y data, with stronger relationships as  $\text{Cor}(X, Y)$  heads towards -1 or 1.
- $\text{Cor}(X, Y) = 0$  implies no linear relationship.



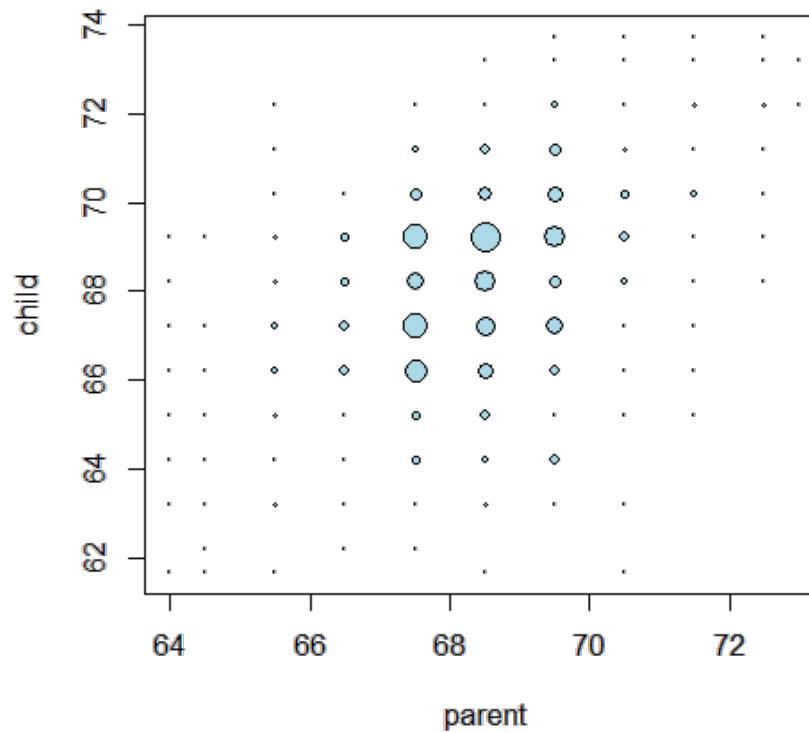
# Least squares estimation of regression lines

Regression via least squares

Brian Caffo, Jeff Leek and Roger Peng  
Johns Hopkins Bloomberg School of Public Health

# General least squares for linear equations

Consider again the parent and child height data from Galton



# Fitting the best line

- Let  $Y_i$  be the  $i^{\text{th}}$  child's height and  $X_i$  be the  $i^{\text{th}}$  (average over the pair of) parents' heights.
- Consider finding the best line
  - Child's Height =  $\beta_0 + \text{Parent's Height } \beta_1$
- Use least squares

$$\sum_{i=1}^n \{Y_i - (\beta_0 + \beta_1 X_i)\}^2$$

- How do we do it?

# Let's solve this problem generally

- Let  $\mu_i = \beta_0 + \beta_1 X_i$  and our estimates be  $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ .
- We want to minimize

$$\dagger \sum_{i=1}^n (Y_i - \mu_i)^2 = \sum_{i=1}^n (Y_i - \hat{\mu}_i)^2 + 2 \sum_{i=1}^n (Y_i - \hat{\mu}_i)(\hat{\mu}_i - \mu_i) + \sum_{i=1}^n (\hat{\mu}_i - \mu_i)^2$$

- Suppose that

$$\sum_{i=1}^n (Y_i - \hat{\mu}_i)(\hat{\mu}_i - \mu_i) = 0$$

then

$$\dagger = \sum_{i=1}^n (Y_i - \hat{\mu}_i)^2 + \sum_{i=1}^n (\hat{\mu}_i - \mu_i)^2 \geq \sum_{i=1}^n (Y_i - \hat{\mu}_i)^2$$

# Mean only regression

- So we know that if:

$$\sum_{i=1}^n (Y_i - \hat{\mu}_i)(\hat{\mu}_i - \mu_i) = 0$$

where  $\mu_i = \beta_0 + \beta_1 X_i$  and  $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$  then the line

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X$$

is the least squares line.

- Consider forcing  $\beta_1 = 0$  and thus  $\hat{\beta}_1 = 0$ ; that is, only considering horizontal lines
- The solution works out to be

$$\hat{\beta}_0 = \bar{Y}.$$

# Let's show it

$$\begin{aligned}\sum_{i=1}^n (Y_i - \hat{\mu}_i)(\hat{\mu}_i - \mu_i) &= \sum_{i=1}^n (Y_i - \hat{\beta}_0)(\hat{\beta}_0 - \beta_0) \\ &= (\hat{\beta}_0 - \beta_0) \sum_{i=1}^n (Y_i - \hat{\beta}_0)\end{aligned}$$

Thus, this will equal 0 if  $\sum_{i=1}^n (Y_i - \hat{\beta}_0) = n\bar{Y} - n\hat{\beta}_0 = 0$

Thus  $\hat{\beta}_0 = \bar{Y}$ .

# Regression through the origin

- Recall that if:

$$\sum_{i=1}^n (Y_i - \hat{\mu}_i)(\hat{\mu}_i - \mu_i) = 0$$

where  $\mu_i = \beta_0 + \beta_1 X_i$  and  $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$  then the line

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X$$

is the least squares line.

- Consider forcing  $\beta_0 = 0$  and thus  $\hat{\beta}_0 = 0$ ; that is, only considering lines through the origin
- The solution works out to be

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n Y_i X_i}{\sum_{i=1}^n X_i^2}.$$

# Let's show it

$$\begin{aligned}\sum_{i=1}^n (Y_i - \hat{\mu}_i)(\hat{\mu}_i - \mu_i) &= \sum_{i=1}^n (Y_i - \hat{\beta}_1 X_i)(\hat{\beta}_1 X_i - \beta_1 X_i) \\ &= (\hat{\beta}_1 - \beta_1) \sum_{i=1}^n (Y_i X_i - \hat{\beta}_1 X_i^2)\end{aligned}$$

Thus, this will equal 0 if  $\sum_{i=1}^n (Y_i X_i - \hat{\beta}_1 X_i^2) = \sum_{i=1}^n Y_i X_i - \hat{\beta}_1 \sum_{i=1}^n X_i^2 = 0$

Thus

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n Y_i X_i}{\sum_{i=1}^n X_i^2}.$$

# Recapping what we know

- If we define  $\mu_i = \beta_0$  then  $\hat{\beta}_0 = \bar{Y}$ .
  - If we only look at horizontal lines, the least squares estimate of the intercept of that line is the average of the outcomes.
- If we define  $\mu_i = X_i\beta_1$  then  $\hat{\beta}_1 = \frac{\sum_{i=1}^n Y_i X_i}{\sum_{i=1}^n X_i^2}$ 
  - If we only look at lines through the origin, we get the estimated slope is the cross product of the X and Ys divided by the cross product of the Xs with themselves.
- What about when  $\mu_i = \beta_0 + \beta_1 X_i$ ? That is, we don't want to restrict ourselves to horizontal lines or lines through the origin.

# Let's figure it out

$$\begin{aligned}\sum_{i=1}^n (Y_i - \hat{\mu}_i)(\hat{\mu}_i - \mu_i) &= \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)(\hat{\beta}_0 + \hat{\beta}_1 X_i - \beta_0 - \beta_1 X_i) \\ &= (\hat{\beta}_0 - \beta_0) \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) + (\beta_1 - \hat{\beta}_1) \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) X_i\end{aligned}$$

Note that

$$0 = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = n\bar{Y} - n\hat{\beta}_0 - n\hat{\beta}_1 \bar{X} \text{ implies that } \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Then

$$\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) X_i = \sum_{i=1}^n (Y_i - \bar{Y} + \hat{\beta}_1 \bar{X} - \hat{\beta}_1 X_i) X_i$$

# Continued

$$= \sum_{i=1}^n \{(Y_i - \bar{Y}) - \hat{\beta}_1(X_i - \bar{X})\} X_i$$

And thus

$$\sum_{i=1}^n (Y_i - \bar{Y}) X_i - \hat{\beta}_1 \sum_{i=1}^n (X_i - \bar{X}) X_i = 0.$$

So we arrive at

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y}) X_i}{\sum_{i=1}^n (X_i - \bar{X}) X_i} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})} = \text{Cor}(Y, X) \frac{\text{Sd}(Y)}{\text{Sd}(X)}.$$

And recall

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

# Consequences

- The least squares model fit to the line  $Y = \beta_0 + \beta_1 X$  through the data pairs  $(X_i, Y_i)$  with  $Y_i$  as the outcome obtains the line  $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$  where

$$\hat{\beta}_1 = \text{Cor}(Y, X) \frac{\text{Sd}(Y)}{\text{Sd}(X)} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

- $\hat{\beta}_1$  has the units of  $Y/X$ ,  $\hat{\beta}_0$  has the units of  $Y$ .
- The line passes through the point  $(\bar{X}, \bar{Y})$
- The slope of the regression line with  $X$  as the outcome and  $Y$  as the predictor is  $\text{Cor}(Y, X)\text{Sd}(X)/\text{Sd}(Y)$ .
- The slope is the same one you would get if you centered the data,  $(X_i - \bar{X}, Y_i - \bar{Y})$ , and did regression through the origin.
- If you normalized the data,  $\left\{ \frac{X_i - \bar{X}}{\text{Sd}(X)}, \frac{Y_i - \bar{Y}}{\text{Sd}(Y)} \right\}$ , the slope is  $\text{Cor}(Y, X)$ .

# Revisiting Galton's data

Double check our calculations using R

```
y <- galton$child  
x <- galton$parent  
beta1 <- cor(y, x) * sd(y) / sd(x)  
beta0 <- mean(y) - beta1 * mean(x)  
rbind(c(beta0, beta1), coef(lm(y ~ x)))
```

```
(Intercept)      x  
[1,]    23.94 0.6463  
[2,]    23.94 0.6463
```

# Revisiting Galton's data

Reversing the outcome/predictor relationship

```
beta1 <- cor(y, x) * sd(x) / sd(y)
beta0 <- mean(x) - beta1 * mean(y)
rbind(c(beta0, beta1), coef(lm(x ~ y)))
```

	(Intercept)	y
[1,]	46.14	0.3256
[2,]	46.14	0.3256

# Revisiting Galton's data

Regression through the origin yields an equivalent slope if you center the data first

```
yc <- y - mean(y)
xc <- x - mean(x)
beta1 <- sum(yc * xc) / sum(xc ^ 2)
c(beta1, coef(lm(y ~ x))[2])
```

```
x
0.6463 0.6463
```

# Revisiting Galton's data

Normalizing variables results in the slope being the correlation

```
yn <- (y - mean(y))/sd(y)
xn <- (x - mean(x))/sd(x)
c(cor(y, x), cor(yn, xn), coef(lm(yn ~ xn))[2])
```

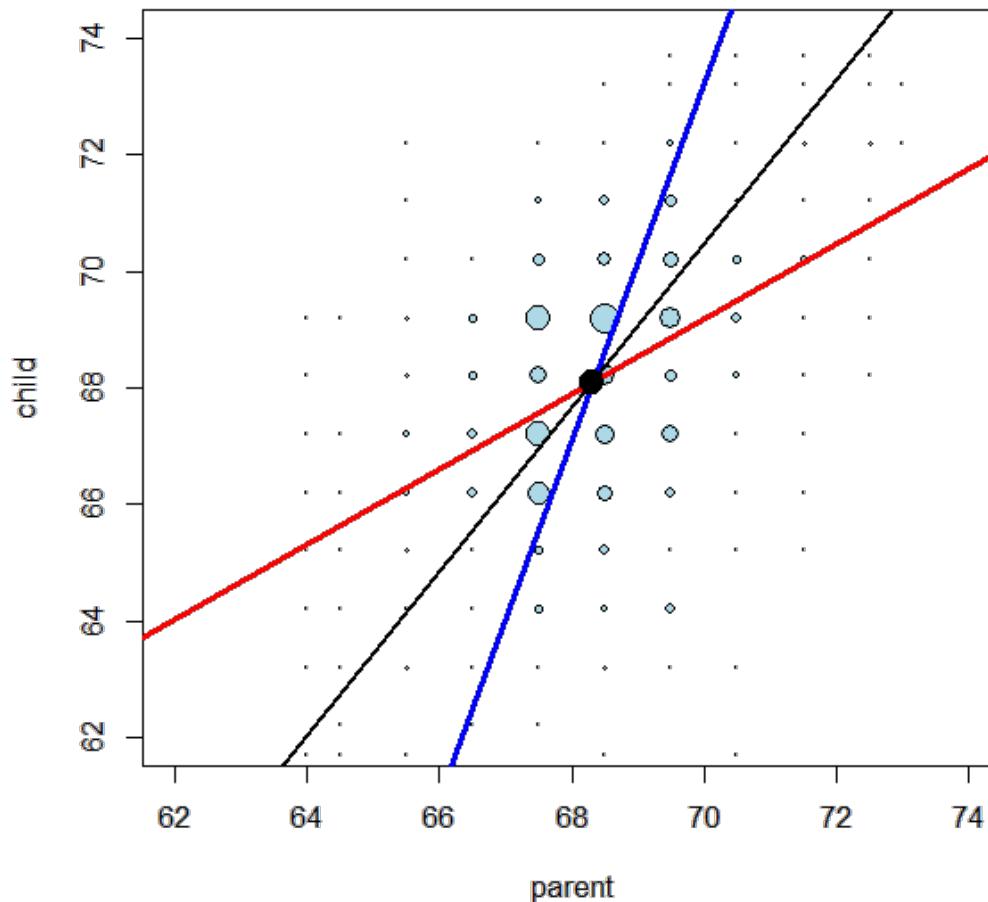
```
xn
0.4588 0.4588 0.4588
```

# Plotting the fit

- Size of points are frequencies at that X, Y combination.
- For the red lie the child is outcome.
- For the blue, the parent is the outcome (accounting for the fact that the response is plotted on the horizontal axis).
- Black line assumes  $\text{Cor}(Y, X) = 1$  (slope is  $Sd(Y)/Sd(x)$ ).
- Big black dot is  $(\bar{X}, \bar{Y})$ .

The code to add the lines

```
abline(mean(y) - mean(x) * cor(y, x) * sd(y) / sd(x),  
      sd(y) / sd(x) * cor(y, x),  
      lwd = 3, col = "red")  
abline(mean(y) - mean(x) * sd(y) / sd(x) / cor(y, x),  
      sd(y) cor(y, x) / sd(x),  
      lwd = 3, col = "blue")  
abline(mean(y) - mean(x) * sd(y) / sd(x),  
      sd(y) / sd(x),  
      lwd = 2)  
points(mean(x), mean(y), cex = 2, pch = 19)
```





# **Historical side note, Regression to Mediocrity**

Regression to the mean

Brian Caffo, Jeff Leek, Roger Peng PhD  
Johns Hopkins Bloomberg School of Public Health

# A historically famous idea, Regression to the Mean

- Why is it that the children of tall parents tend to be tall, but not as tall as their parents?
- Why do children of short parents tend to be short, but not as short as their parents?
- Why do parents of very short children, tend to be short, but not as short as their child? And the same with parents of very tall children?
- Why do the best performing athletes this year tend to do a little worse the following?

# Regression to the mean

- These phenomena are all examples of so-called regression to the mean
- Invented by Francis Galton in the paper "Regression towards mediocrity in hereditary stature" The Journal of the Anthropological Institute of Great Britain and Ireland , Vol. 15, (1886).
- Think of it this way, imagine if you simulated pairs of random normals
  - The largest first ones would be the largest by chance, and the probability that there are smaller for the second simulation is high.
  - In other words  $P(Y < x|X = x)$  gets bigger as  $x$  heads into the very large values.
  - Similarly  $P(Y > x|X = x)$  gets bigger as  $x$  heads to very small values.
- Think of the regression line as the intrinsic part.
  - Unless  $\text{Cor}(Y, X) = 1$  the intrinsic part isn't perfect

# Regression to the mean

- Suppose that we normalize  $X$  (child's height) and  $Y$  (parent's height) so that they both have mean 0 and variance 1.
- Then, recall, our regression line passes through  $(0, 0)$  (the mean of the  $X$  and  $Y$ ).
- If the slope of the regression line is  $\text{Cor}(Y, X)$ , regardless of which variable is the outcome (recall, both standard deviations are 1).
- Notice if  $X$  is the outcome and you create a plot where  $X$  is the horizontal axis, the slope of the least squares line that you plot is  $1/\text{Cor}(Y, X)$ .

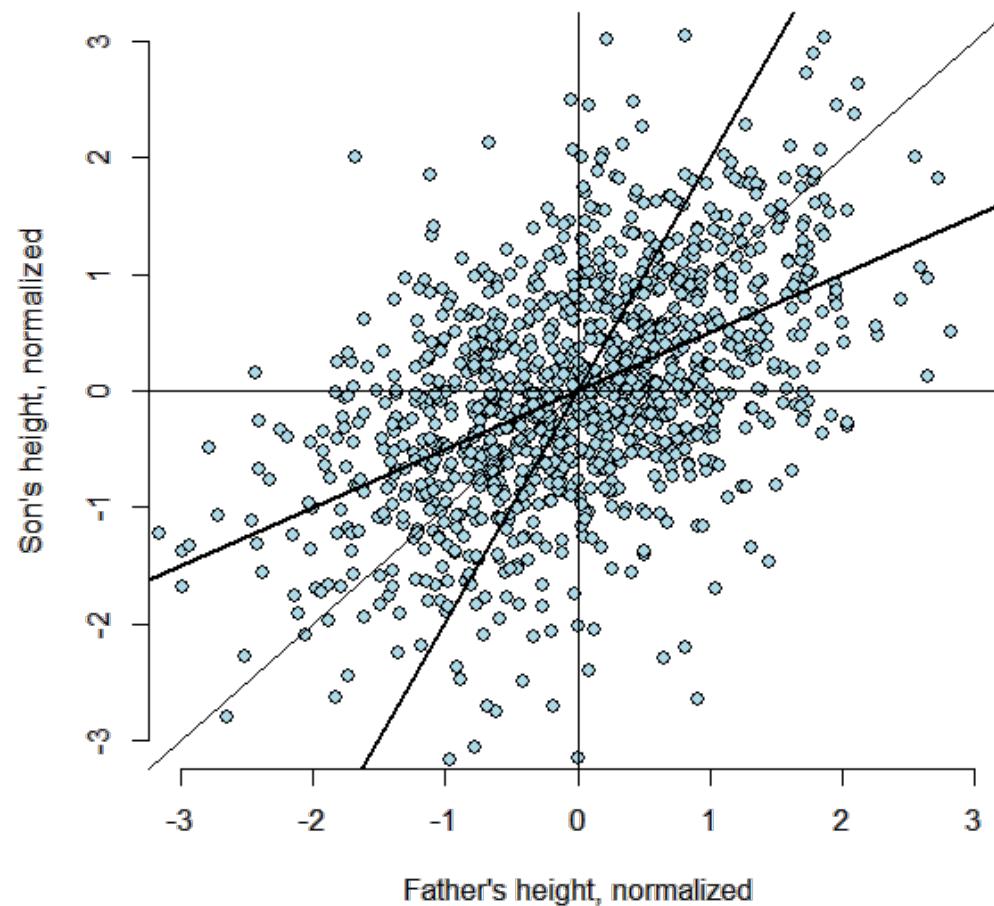
# Normalizing the data and setting plotting parameters

```
library(UsingR)
data(father.son)
y <- (father.son$sheight - mean(father.son$sheight)) / sd(father.son$sheight)
x <- (father.son$fheight - mean(father.son$fheight)) / sd(father.son$fheight)
rho <- cor(x, y)
myPlot <- function(x, y) {
  plot(x, y,
       xlab = "Father's height, normalized",
       ylab = "Son's height, normalized",
       xlim = c(-3, 3), ylim = c(-3, 3),
       bg = "lightblue", col = "black", cex = 1.1, pch = 21,
       frame = FALSE)
}
```

# Plot the data, code

```
myPlot(x, y)
abline(0, 1) # if there were perfect correlation
abline(0, rho, lwd = 2) # father predicts son
abline(0, 1 / rho, lwd = 2) # son predicts father, son on vertical axis
abline(h = 0); abline(v = 0) # reference lines for no relationship
```

# Plot the data, results



# Discussion

- If you had to predict a son's normalized height, it would be  $\text{Cor}(Y, X) * X_i$
- If you had to predict a father's normalized height, it would be  $\text{Cor}(Y, X) * Y_i$
- Multiplication by this correlation shrinks toward 0 (regression toward the mean)
- If the correlation is 1 there is no regression to the mean (if father's height perfectly determines child's height and vice versa)
- Note, regression to the mean has been thought about quite a bit and generalized



# Statistical linear regression models

Brian Caffo, Jeff Leek, Roger Peng  
Johns Hopkins Bloomberg School of Public Health

# Basic regression model with additive Gaussian errors.

- Least squares is an estimation tool, how do we do inference?
- Consider developing a probabilistic model for linear regression

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Here the  $\epsilon_i$  are assumed iid  $N(0, \sigma^2)$ .
- Note,  $E[Y_i | X_i = x_i] = \mu_i = \beta_0 + \beta_1 x_i$
- Note,  $\text{Var}(Y_i | X_i = x_i) = \sigma^2$ .
- Likelihood equivalent model specification is that the  $Y_i$  are independent  $N(\mu_i, \sigma^2)$ .

# Likelihood

$$>(\beta, \sigma) = \prod_{i=1}^n \left\{ (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2\sigma^2} (y_i - \mu_i)^2\right) \right\}$$

so that the twice the negative log (base e) likelihood is

$$-2 \log\{>(\beta, \sigma)\} = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \mu_i)^2 + n \log(\sigma^2)$$

## Discussion

- Maximizing the likelihood is the same as minimizing -2 log likelihood
- The least squares estimate for  $\mu_i = \beta_0 + \beta_1 x_i$  is exactly the maximum likelihood estimate (regardless of  $\sigma$ )

# Recap

- Model  $Y_i = \mu_i + \epsilon_i = \beta_0 + \beta_1 X_i + \epsilon_i$  where  $\epsilon_i$  are iid  $N(0, \sigma^2)$
- ML estimates of  $\beta_0$  and  $\beta_1$  are the least squares estimates

$$\hat{\beta}_1 = \text{Cor}(Y, X) \frac{\text{Sd}(Y)}{\text{Sd}(X)} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

- $E[Y | X = x] = \beta_0 + \beta_1 x$
- $\text{Var}(Y | X = x) = \sigma^2$

# Interpreting regression coefficients, the itc

- $\beta_0$  is the expected value of the response when the predictor is 0

$$E[Y|X = 0] = \beta_0 + \beta_1 \times 0 = \beta_0$$

- Note, this isn't always of interest, for example when  $X = 0$  is impossible or far outside of the range of data. ( $X$  is blood pressure, or height etc.)
- Consider that

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i = \beta_0 + a\beta_1 + \beta_1(X_i - a) + \epsilon_i = \tilde{\beta}_0 + \beta_1(X_i - a) + \epsilon_i$$

So, shifting your  $X$  values by value  $a$  changes the intercept, but not the slope.

- Often  $a$  is set to  $\bar{X}$  so that the intercept is interpreted as the expected response at the average  $X$  value.

# Interpreting regression coefficients, the slope

- $\beta_1$  is the expected change in response for a 1 unit change in the predictor

$$E[Y | X = x + 1] - E[Y | X = x] = \beta_0 + \beta_1(x + 1) - (\beta_0 + \beta_1x) = \beta_1$$

- Consider the impact of changing the units of X.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i = \beta_0 + \frac{\beta_1}{a} (X_i a) + \epsilon_i = \beta_0 + \tilde{\beta}_1 (X_i a) + \epsilon_i$$

- Therefore, multiplication of X by a factor a results in dividing the coefficient by a factor of a.
- Example: X is height in m and Y is weight in kg. Then  $\beta_1$  is kg/m. Converting X to cm implies multiplying X by 100cm/m . To get  $\beta_1$  in the right units, we have to divide by 100cm/m to get it to have the right units.

$$Xm \times \frac{100\text{cm}}{\text{m}} = (100X)\text{cm} \text{ and } \beta_1 \frac{\text{kg}}{\text{m}} \times \frac{1\text{m}}{100\text{cm}} = \left( \frac{\beta_1}{100} \right) \frac{\text{kg}}{\text{cm}}$$

# Using regression coefficients for prediction

- If we would like to guess the outcome at a particular value of the predictor, say  $X$ , the regression model guesses

$$\hat{\beta}_0 + \hat{\beta}_1 X$$

- Note that at the observed value of  $X$ s, we obtain the predictions

$$\hat{\mu}_i = \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

- Remember that least squares minimizes

$$\sum_{i=1}^n (Y_i - \mu_i)$$

for  $\mu_i$  expressed as points on a line

# Example

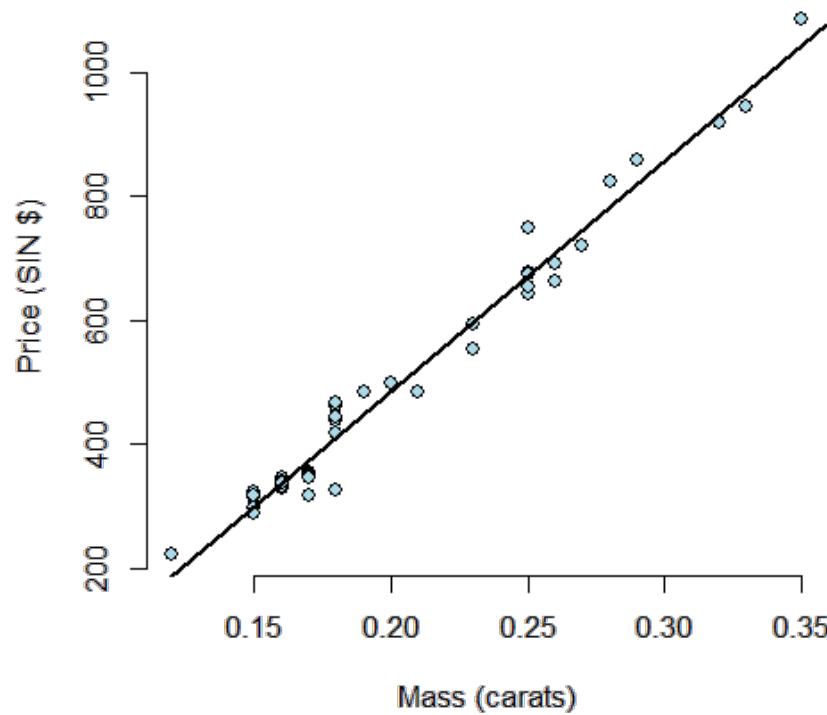
## diamond data set from UsingR

Data is diamond prices (Singapore dollars) and diamond weight in carats (standard measure of diamond mass, 0.2 g). To get the data use `library(UsingR); data(diamond)`

Plotting the fitted regression line and data

```
data(diamond)
plot(diamond$carat, diamond$price,
      xlab = "Mass (carats)",
      ylab = "Price (SIN $)",
      bg = "lightblue",
      col = "black", cex = 1.1, pch = 21, frame = FALSE)
abline(lm(price ~ carat, data = diamond), lwd = 2)
```

# The plot



# Fitting the linear regression model

```
fit <- lm(price ~ carat, data = diamond)  
coef(fit)
```

(Intercept)	carat
-259.6	3721.0

- We estimate an expected 3721.02 (SIN) dollar increase in price for every carat increase in mass of diamond.
- The intercept -259.63 is the expected price of a 0 carat diamond.

# Getting a more interpretable intercept

```
fit2 <- lm(price ~ I(carat - mean(carat)), data = diamond)  
coef(fit2)
```

(Intercept)	I(carat - mean(carat))
500.1	3721.0

Thus \$500.1 is the expected price for the average sized diamond of the data (0.2042 carats).

# Changing scale

- A one carat increase in a diamond is pretty big, what about changing units to 1/10th of a carat?
- We can just do this by just dividing the coefficient by 10.
  - We expect a 372.102 (SIN) dollar change in price for every 1/10th of a carat increase in mass of diamond.
- Showing that it's the same if we rescale the Xs and refit

```
fit3 <- lm(price ~ I(carat * 10), data = diamond)
coef(fit3)
```

```
(Intercept) I(carat * 10)
-259.6      372.1
```

# Predicting the price of a diamond

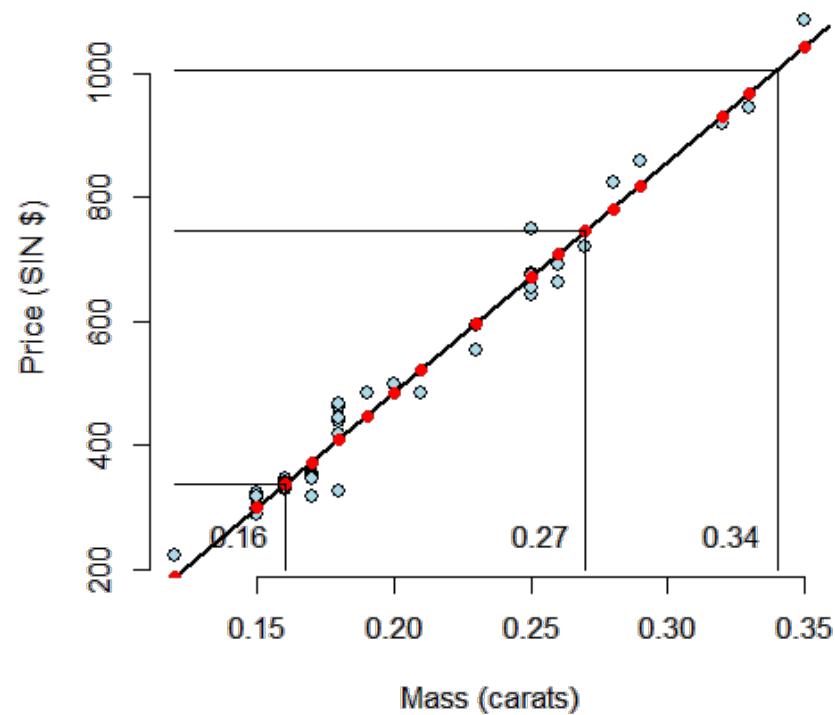
```
newx <- c(0.16, 0.27, 0.34)  
coef(fit)[1] + coef(fit)[2] * newx
```

```
[1] 335.7 745.1 1005.5
```

```
predict(fit, newdata = data.frame(carat = newx))
```

```
1      2      3  
335.7 745.1 1005.5
```

Predicted values at the observed Xs (red) and at the new Xs (lines)





# Residuals and residual variation

Brian Caffo, Jeff Leek and Roger Peng  
Johns Hopkins Bloomberg School of Public Health

# Residuals

- Model  $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$  where  $\epsilon_i \sim N(0, \sigma^2)$ .
- Observed outcome  $i$  is  $Y_i$  at predictor value  $X_i$
- Predicted outcome  $i$  is  $\hat{Y}_i$  at predictor value  $X_i$  is

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

- Residual, the difference between the observed and predicted outcome

$$e_i = Y_i - \hat{Y}_i$$

- The vertical distance between the observed data point and the regression line
- Least squares minimizes  $\sum_{i=1}^n e_i^2$
- The  $e_i$  can be thought of as estimates of the  $\epsilon_i$ .

# Properties of the residuals

- $E[e_i] = 0$ .
- If an intercept is included,  $\sum_{i=1}^n e_i = 0$
- If a regressor variable,  $X_i$ , is included in the model  $\sum_{i=1}^n e_i X_i = 0$ .
- Residuals are useful for investigating poor model fit.
- Positive residuals are above the line, negative residuals are below.
- Residuals can be thought of as the outcome ( $Y$ ) with the linear association of the predictor ( $X$ ) removed.
- One differentiates residual variation (variation after removing the predictor) from systematic variation (variation explained by the regression model).
- Residual plots highlight poor model fit.

# Code

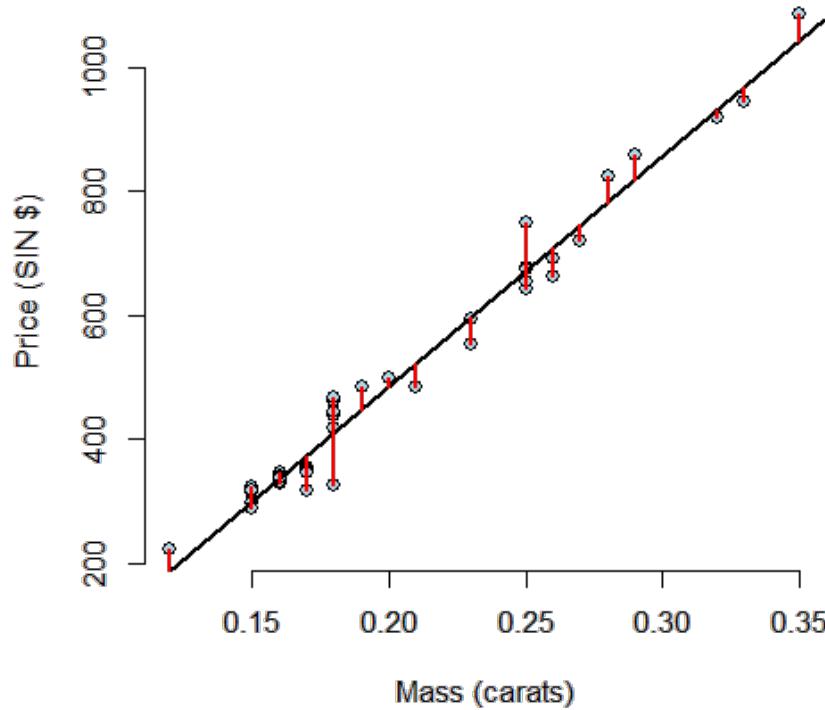
```
data(diamond)
y <- diamond$price; x <- diamond$carat; n <- length(y)
fit <- lm(y ~ x)
e <- resid(fit)
yhat <- predict(fit)
max(abs(e - (y - yhat)))
```

```
[1] 9.486e-13
```

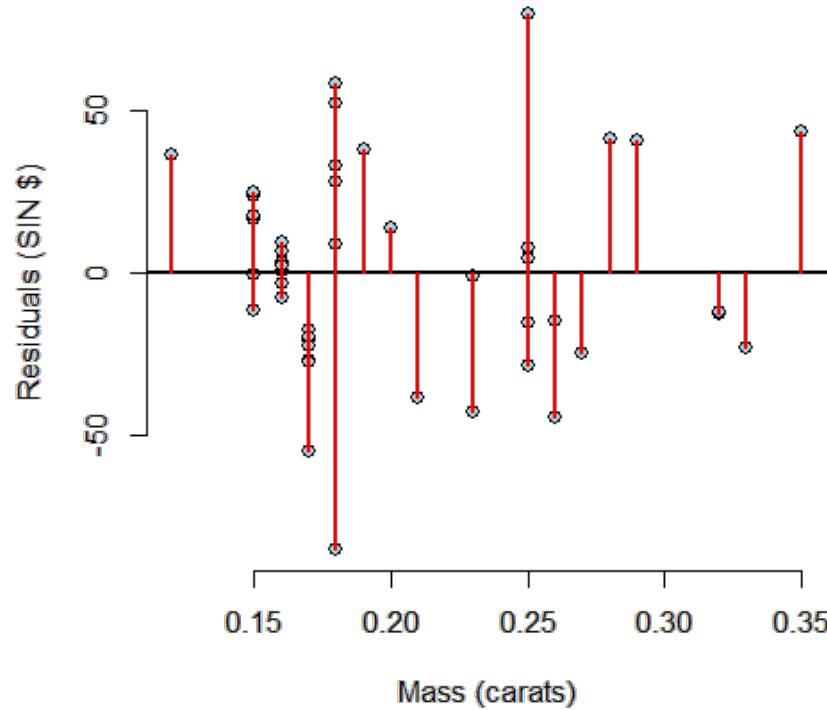
```
max(abs(e - (y - coef(fit)[1] - coef(fit)[2] * x)))
```

```
[1] 9.486e-13
```

# Residuals are the signed length of the red lines

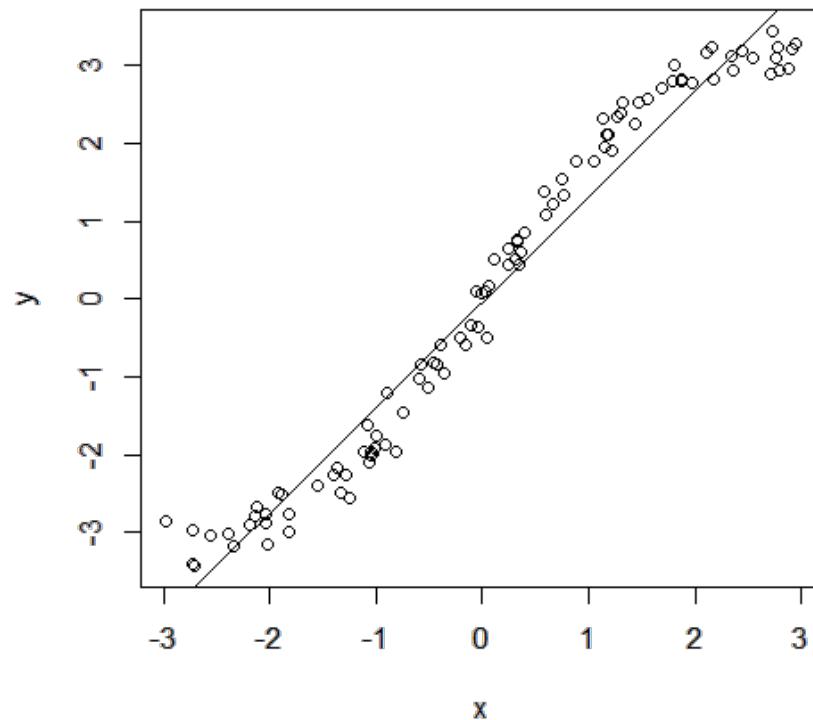


# Residuals versus X

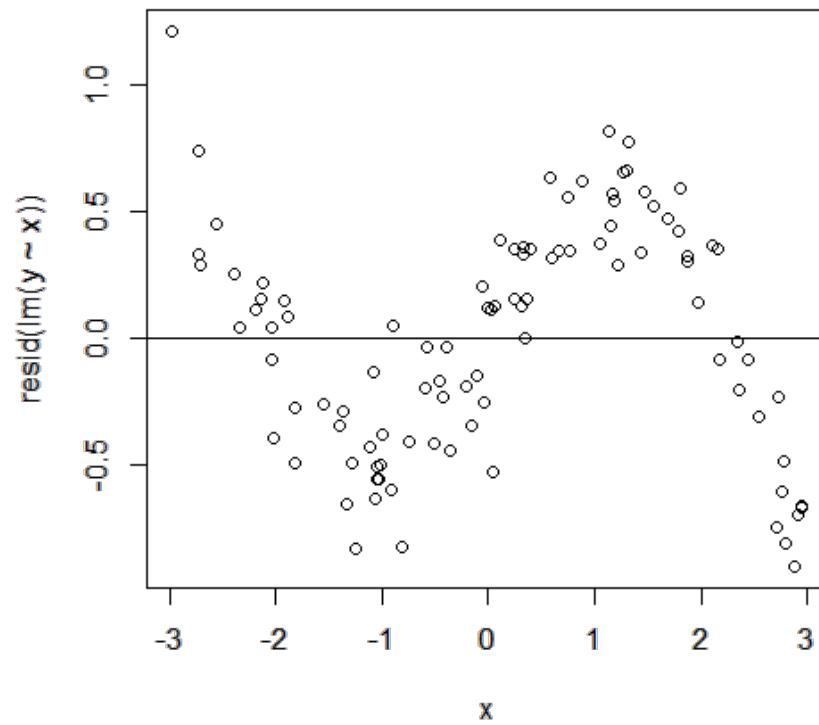


# Non-linear data

```
x <- runif(100, -3, 3); y <- x + sin(x) + rnorm(100, sd = .2);  
plot(x, y); abline(lm(y ~ x))
```

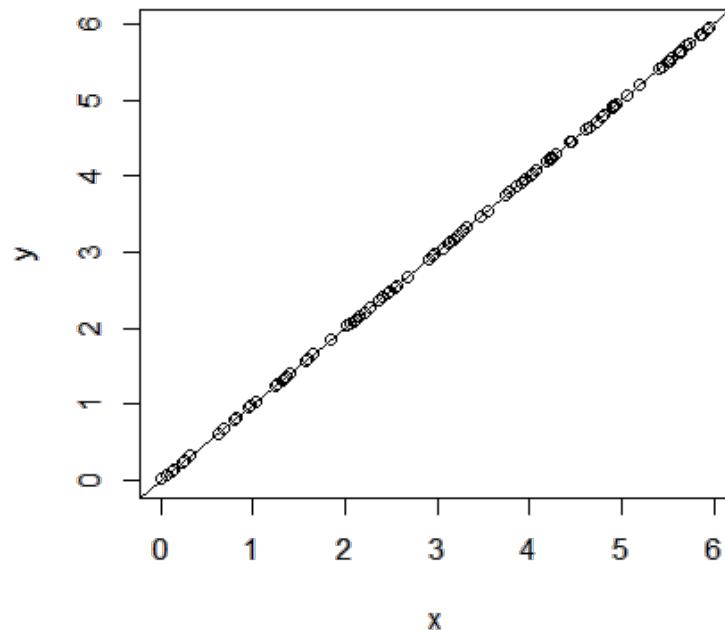


```
plot(x, resid(lm(y ~ x)));
abline(h = 0)
```



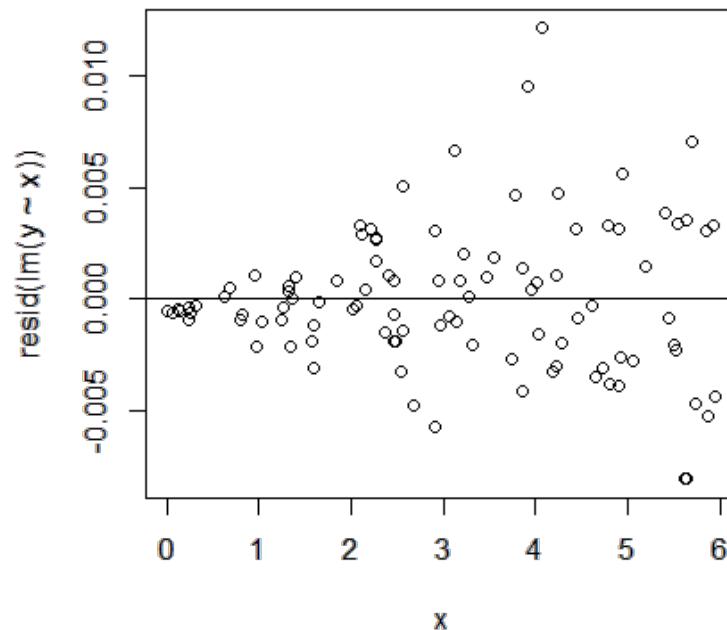
# Heteroskedasticity

```
x <- runif(100, 0, 6); y <- x + rnorm(100, mean = 0, sd = .001 * x);  
plot(x, y); abline(lm(y ~ x))
```



# Getting rid of the blank space can be helpful

```
plot(x, resid(lm(y ~ x)));
abline(h = 0)
```



# Estimating residual variation

- Model  $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$  where  $\epsilon_i \sim N(0, \sigma^2)$ .
- The ML estimate of  $\sigma^2$  is  $\frac{1}{n} \sum_{i=1}^n e_i^2$ , the average squared residual.
- Most people use

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2.$$

- The  $n - 2$  instead of  $n$  is so that  $E[\hat{\sigma}^2] = \sigma^2$

# Diamond example

```
y <- diamond$price; x <- diamond$carat; n <- length(y)  
fit <- lm(y ~ x)  
summary(fit)$sigma
```

```
[1] 31.84
```

```
sqrt(sum(resid(fit)^2) / (n - 2))
```

```
[1] 31.84
```

# Summarizing variation

$$\begin{aligned}\sum_{i=1}^n (Y_i - \bar{Y})^2 &= \sum_{i=1}^n (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \\ &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + 2 \sum_{i=1}^n (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2\end{aligned}$$

---

## Scratch work

$$(Y_i - \hat{Y}_i) = \{Y_i - (\bar{Y} - \hat{\beta}_1 \bar{X}) - \hat{\beta}_1 X_i\} = (Y_i - \bar{Y}) - \hat{\beta}_1 (X_i - \bar{X})$$

$$(\hat{Y}_i - \bar{Y}) = (\bar{Y} - \hat{\beta}_1 \bar{X} - \hat{\beta}_1 X_i - \bar{Y}) = \hat{\beta}_1 (X_i - \bar{X})$$

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = \sum_{i=1}^n \{(Y_i - \bar{Y}) - \hat{\beta}_1 (X_i - \bar{X})\} \{\hat{\beta}_1 (X_i - \bar{X})\}$$

$$= \hat{\beta}_1 \sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X}) - \hat{\beta}_1^2 \sum_{i=1}^n (X_i - \bar{X})^2$$

$$= \hat{\beta}_1^2 \sum_{i=1}^n (X_i - \bar{X})^2 - \hat{\beta}_1^2 \sum_{i=1}^n (X_i - \bar{X})^2 = 0$$

# Summarizing variation

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

Or

Total Variation = Residual Variation + Regression Variation

Define the percent of total variation described by the model as

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

# Relation between $R^2$ and $r$ (the correlation)

Recall that  $(\hat{Y}_i - \bar{Y}) = \hat{\beta}_1(X_i - \bar{X})$  so that

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \hat{\beta}_1^2 \frac{\sum_{i=1}^n (X_i - \bar{X})}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \text{Cor}(Y, X)^2$$

Since, recall,

$$\hat{\beta}_1 = \text{Cor}(Y, X) \frac{Sd(Y)}{Sd(X)}$$

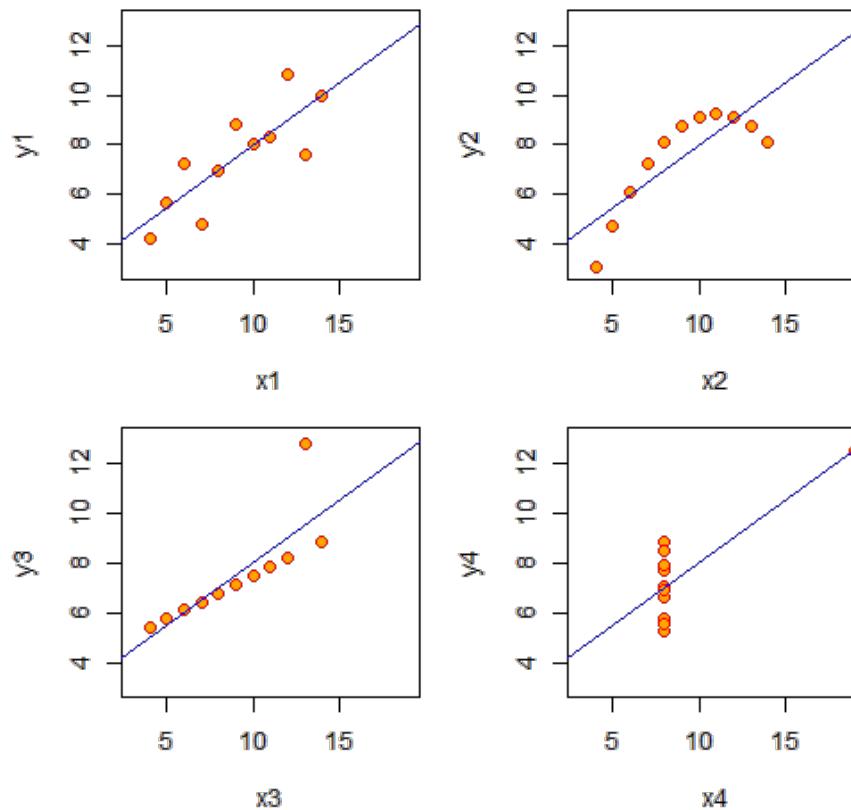
So,  $R^2$  is literally  $r$  squared.

# Some facts about $R^2$

- $R^2$  is the percentage of variation explained by the regression model.
- $0 \leq R^2 \leq 1$
- $R^2$  is the sample correlation squared.
- $R^2$  can be a misleading summary of model fit.
  - Deleting data can inflate  $R^2$ .
  - (For later.) Adding terms to a regression model always increases  $R^2$ .
- Do `example(anscombe)` to see the following data.
  - Basically same mean and variance of X and Y.
  - Identical correlations (hence same  $R^2$  ).
  - Same linear regression relationship.

# **data(anscombe) ; example(anscombe)**

Anscombe's 4 Regression data sets





# Multivariable regression

Brian Caffo, Roger Peng and Jeff Leek  
Johns Hopkins Bloomberg School of Public Health

# Multivariable regression analyses

- If I were to present evidence of a relationship between breath mint usage (mints per day, X) and pulmonary function (measured in FEV), you would be skeptical.
  - Likely, you would say, 'smokers tend to use more breath mints than non smokers, smoking is related to a loss in pulmonary function. That's probably the culprit.'
  - If asked what would convince you, you would likely say, 'If non-smoking breath mint users had lower lung function than non-smoking non-breath mint users and, similarly, if smoking breath mint users had lower lung function than smoking non-breath mint users, I'd be more inclined to believe you'.
- In other words, to even consider my results, I would have to demonstrate that they hold while holding smoking status fixed.

# Multivariable regression analyses

- An insurance company is interested in how last year's claims can predict a person's time in the hospital this year.
  - They want to use an enormous amount of data contained in claims to predict a single number. Simple linear regression is not equipped to handle more than one predictor.
- How can one generalize SLR to incorporate lots of regressors for the purpose of prediction?
- What are the consequences of adding lots of regressors?
  - Surely there must be consequences to throwing variables in that aren't related to Y?
  - Surely there must be consequences to omitting variables that are?

# The linear model

- The general linear model extends simple linear regression (SLR) by adding terms linearly into the model.

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_p X_{pi} + \epsilon_i = \sum_{k=1}^p X_{ik} \beta_j + \epsilon_i$$

- Here  $X_{1i} = 1$  typically, so that an intercept is included.
- Least squares (and hence ML estimates under iid Gaussianity of the errors) minimizes

$$\sum_{i=1}^n \left( Y_i - \sum_{k=1}^p X_{ki} \beta_j \right)^2$$

- Note, the important linearity is linearity in the coefficients. Thus

$$Y_i = \beta_1 X_{1i}^2 + \beta_2 X_{2i}^2 + \dots + \beta_p X_{pi}^2 + \epsilon_i$$

is still a linear model. (We've just squared the elements of the predictor variables.)

# How to get estimates

- The real way requires linear algebra. We'll go over an intuitive development instead.
- Recall that the LS estimate for regression through the origin,  $E[Y_i] = X_{1i}\beta_1$ , was  $\sum X_i Y_i / \sum X_i^2$ .
- Let's consider two regressors,  $E[Y_i] = X_{1i}\beta_1 + X_{2i}\beta_2 = \mu_i$ .
- Also, recall, that if  $\hat{\mu}_i$  satisfies

$$\sum_{i=1} (Y_i - \hat{\mu}_i)(\hat{\mu}_i - \mu_i) = 0$$

for all possible values of  $\mu_i$ , then we've found the LS estimates.

$$\sum_{i=1}^n (Y_i - \hat{\mu}_i)(\hat{\mu}_i - \mu_i) = \sum_{i=1}^n (Y_i - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}) \left\{ X_{1i}(\hat{\beta}_1 - \beta_1) + X_{2i}(\hat{\beta}_2 - \beta_2) \right\}$$

- Thus we need

$$1. \sum_{i=1}^n (Y_i - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}) X_{1i} = 0$$

$$2. \sum_{i=1}^n (Y_i - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}) X_{2i} = 0$$

- Hold  $\hat{\beta}_1$  fixed in 2. and solve and we get that

$$\hat{\beta}_2 = \frac{\sum_{i=1}^n (Y_i - X_{1i}\hat{\beta}_1) X_{2i}}{\sum_{i=1}^n X_{2i}^2}$$

- Plugging this into 1. we get that

$$0 = \sum_{i=1}^n \left\{ Y_i - \frac{\sum_j X_{2j} Y_j}{\sum_j X_{2j}^2} X_{2i} + \beta_1 \left( X_{1i} - \frac{\sum_j X_{2j} X_{1j}}{\sum_j X_{2j}^2} X_{2i} \right) \right\} X_{1i}$$

# Continued

- Re writing this we get

$$0 = \sum_{i=1}^n \left\{ e_{i,Y|X_2} - \hat{\beta}_1 e_{i,X_1|X_2} \right\} X_{1i}$$

where  $e_{i,a|b} = a_i - \frac{\sum_{j=1}^n a_j b_j}{\sum_{j=1}^n b_j^2}$   $b_i$  is the residual when regressing  $b$  from  $a$  without an intercept.

- We get the solution

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n e_{i,Y|X_2} e_{i,X_1|X_2}}{\sum_{i=1}^n e_{i,X_1|X_2} X_1}$$

- But note that

$$\begin{aligned}\sum_{i=1}^n e_{i,X_1|X_2}^2 &= \sum_{i=1}^n e_{i,X_1|X_2} \left( X_{1i} - \frac{\sum_j X_{2j} X_{1j}}{\sum_j X_{2j}^2} X_{2i} \right) \\ &= \sum_{i=1}^n e_{i,X_1|X_2} X_{1i} - \frac{\sum_j X_{2j} X_{1j}}{\sum_j X_{2j}^2} \sum_{i=1}^n e_{i,X_1|X_2} X_{2i}\end{aligned}$$

But  $\sum_{i=1}^n e_{i,X_1|X_2} X_{2i} = 0$ . So we get that

$$\sum_{i=1}^n e_{i,X_1|X_2}^2 = \sum_{i=1}^n e_{i,X_1|X_2} X_{1i}$$

Thus we get that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n e_{i,Y|X_2} e_{i,X_1|X_2}}{\sum_{i=1}^n e_{i,X_1|X_2}^2}$$

# Summing up fitting with two regressors

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n e_{i,Y|X_2} e_{i,X_1|X_2}}{\sum_{i=1}^n e_{i,X_1|X_2}^2}$$

- That is, the regression estimate for  $\beta_1$  is the regression through the origin estimate having regressed  $X_2$  out of both the response and the predictor.
- (Similarly, the regression estimate for  $\beta_2$  is the regression through the origin estimate having regressed  $X_1$  out of both the response and the predictor.)
- More generally, multivariate regression estimates are exactly those having removed the linear relationship of the other variables from both the regressor and response.

# Example with two variables, simple linear regression

- $Y_i = \beta_1 X_{1i} + \beta_2 X_{2i}$  where  $X_{2i} = 1$  is an intercept term.
- Then  $\frac{\sum_j X_{2j} X_{1j}}{\sum_j X_{2j}^2} X_{2i} = \frac{\sum_j X_{1j}}{n} = \bar{X}_1$ .
- $e_{i,X_1|X_2} = X_{1i} - \bar{X}_1$ .
- Similarly  $e_{i,Y|X_2} = Y_i - \bar{Y}$ .
- Thus

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n e_{i,Y|X_2} e_{i,X_1|X_2}}{\sum_{i=1}^n e_{i,X_1|X_2}^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = Cor(X, Y) \frac{Sd(Y)}{Sd(X)}$$

# The general case

- The equations

$$\sum_{i=1}^n (Y_i - X_{1i}\hat{\beta}_1 - \dots - X_{ip}\hat{\beta}_p)X_k = 0$$

for  $k = 1, \dots, p$  yields  $p$  equations with  $p$  unknowns.

- Solving them yields the least squares estimates. (With obtaining a good, fast, general solution requiring some knowledge of linear algebra.)
- The least squares estimate for the coefficient of a multivariate regression model is exactly regression through the origin with the linear relationships with the other regressors removed from both the regressor and outcome by taking residuals.
- In this sense, multivariate regression "adjusts" a coefficient for the linear impact of the other variables.

# Fitting LS equations

Just so I don't leave you hanging, let's show a way to get estimates. Recall the equations:

$$\sum_{i=1}^n (Y_i - X_{1i}\hat{\beta}_1 - \dots - X_{ip}\hat{\beta}_p)X_k = 0$$

If I hold  $\hat{\beta}_1, \dots, \hat{\beta}_{p-1}$  fixed then we get that

$$\hat{\beta}_p = \frac{\sum_{i=1}^n (Y_i - X_{1i}\hat{\beta}_1 - \dots - X_{i,p-1}\hat{\beta}_{p-1})X_{ip}}{\sum_{i=1}^n X_{ip}^2}$$

Plugging this back into the equations, we wind up with

$$\sum_{i=1}^n (e_{i,Y|X_p} - e_{i,X_1|X_p}\hat{\beta}_1 - \dots - e_{i,X_{p-1}|X_p}\hat{\beta}_{p-1})X_k = 0$$

# We can tidy it up a bit more, though

Note that

$$X_k = e_{i,X_k|X_p} + \frac{\sum_{i=1}^n X_{ik}X_{ip}}{\sum_{i=1}^n X_{ip^2}} X_p$$

and  $\sum_{i=1}^n e_{i,X_j|X_p} X_{ip} = 0$ . Thus

$$\sum_{i=1}^n (e_{i,Y|X_p} - e_{i,X_1|X_p} \hat{\beta}_1 - \dots - e_{i,X_{p-1}|X_p} \hat{\beta}_{p-1}) X_k = 0$$

is equal to

$$\sum_{i=1}^n (e_{i,Y|X_p} - e_{i,X_1|X_p} \hat{\beta}_1 - \dots - e_{i,X_{p-1}|X_p} \hat{\beta}_{p-1}) e_{i,X_k|X_p} = 0$$

# To sum up

- We've reduced  $p$  LS equations and  $p$  unknowns to  $p - 1$  LS equations and  $p - 1$  unknowns.
  - Every variable has been replaced by its residual with  $X_p$ .
  - This process can then be iterated until only Y and one variable remains.
- Think of it as follows. If we want an adjusted relationship between  $y$  and  $x$ , keep taking residuals over confounders and do regression through the origin.
  - The order that you do the confounders doesn't matter.
  - (It can't because our choice of doing  $p$  first was arbitrary.)
- This isn't a terribly efficient way to get estimates. But, it's nice conceptually, as it shows how regression estimates are adjusted for the linear relationship with other variables.

# Demonstration that it works using an example

Linear model with two variables and an intercept

```
n <- 100; x <- rnorm(n); x2 <- rnorm(n); x3 <- rnorm(n)
y <- x + x2 + x3 + rnorm(n, sd = .1)
e <- function(a, b) a - sum( a * b ) / sum( b ^ 2) * b
ey <- e(e(y, x2), e(x3, x2))
ex <- e(e(x, x2), e(x3, x2))
sum(ey * ex) / sum(ex ^ 2)
```

```
[1] 1.004
```

```
coef(lm(y ~ x + x2 + x3 - 1)) #the -1 removes the intercept term
```

x	x2	x3
1.0040	0.9899	1.0078

# Showing that order doesn't matter

```
ey <- e(e(y, x3), e(x2, x3))
ex <- e(e(x, x3), e(x2, x3))
sum(ey * ex) / sum(ex ^ 2)
```

```
[1] 1.004
```

```
coef(lm(y ~ x + x2 + x3 - 1)) #the -1 removes the intercept term
```

x	x2	x3
1.0040	0.9899	1.0078

# Residuals again

```
ey <- resid(lm(y ~ x2 + x3 - 1))
ex <- resid(lm(x ~ x2 + x3 - 1))
sum(ey * ex) / sum(ex ^ 2)
```

```
[1] 1.004
```

```
coef(lm(y ~ x + x2 + x3 - 1)) #the -1 removes the intercept term
```

x	x2	x3
1.0040	0.9899	1.0078

# Interpretation of the coefficient

$$E[Y|X_1 = x_1, \dots, X_p = x_p] = \sum_{k=1}^p x_k \beta_k$$

So that

$$\begin{aligned} E[Y|X_1 = x_1 + 1, \dots, X_p = x_p] - E[Y|X_1 = x_1, \dots, X_p = x_p] \\ = (x_1 + 1)\beta_1 + \sum_{k=2}^p x_k + \sum_{k=1}^p x_k \beta_k = \beta_1 \end{aligned}$$

So that the interpretation of a multivariate regression coefficient is the expected change in the response per unit change in the regressor, holding all of the other regressors fixed.

In the next lecture, we'll do examples and go over context-specific interpretations.

# Fitted values, residuals and residual variation

All of our SLR quantities can be extended to linear models

- Model  $Y_i = \sum_{k=1}^p X_{ik}\beta_k + \epsilon_i$  where  $\epsilon_i \sim N(0, \sigma^2)$
- Fitted responses  $\hat{Y}_i = \sum_{k=1}^p X_{ik}\hat{\beta}_k$
- Residuals  $e_i = Y_i - \hat{Y}_i$
- Variance estimate  $\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^n e_i^2$
- To get predicted responses at new values,  $x_1, \dots, x_p$ , simply plug them into the linear model  $\sum_{k=1}^p x_k \hat{\beta}_k$
- Coefficients have standard errors,  $\hat{\sigma}_{\hat{\beta}_k}$ , and  $\frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}_{\hat{\beta}_k}}$  follows a  $T$  distribution with  $n - p$  degrees of freedom.
- Predicted responses have standard errors and we can calculate predicted and expected response intervals.

# Linear models

- Linear models are the single most important applied statistical and machine learning technique, *by far.*
- Some amazing things that you can accomplish with linear models
  - Decompose a signal into its harmonics.
  - Flexibly fit complicated functions.
  - Fit factor variables as predictors.
  - Uncover complex multivariate relationships with the response.
  - Build accurate prediction models.



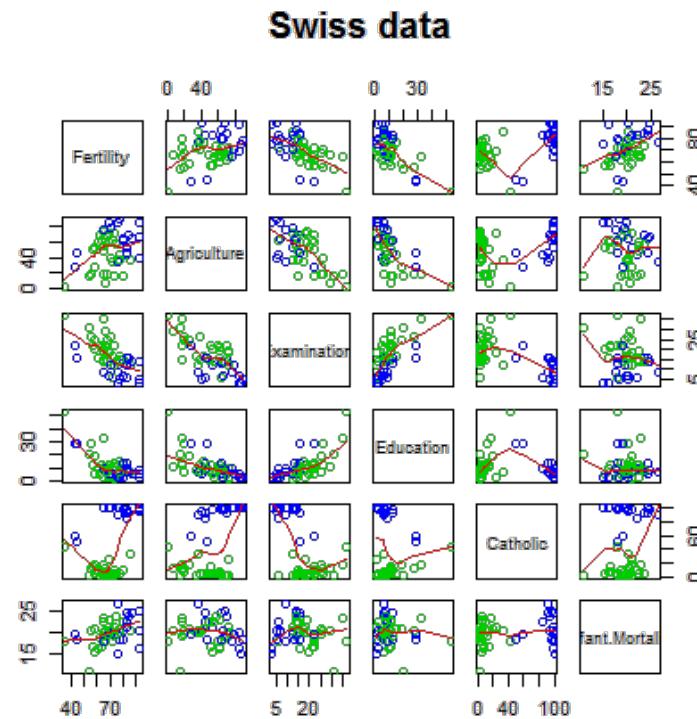
# Multivariable regression examples

Regression Models

Brian Caffo, Jeff Leek and Roger Peng  
Johns Hopkins Bloomberg School of Public Health

# Swiss fertility data

```
library(datasets); data(swiss); require(stats); require(graphics)
pairs(swiss, panel = panel.smooth, main = "Swiss data", col = 3 + (swiss$Catholic > 50))
```



# ?Swiss

## Description

Standardized fertility measure and socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888.

A data frame with 47 observations on 6 variables, each of which is in percent, i.e., in [0, 100].

- [,1] Fertility Ig, ‘common standardized fertility measure’
- [,2] Agriculture % of males involved in agriculture as occupation
- [,3] Examination % draftees receiving highest mark on army examination
- [,4] Education % education beyond primary school for draftees.
- [,5] Catholic % ‘catholic’ (as opposed to ‘protestant’).
- [,6] Infant.Mortality live births who live less than 1 year.

All variables but ‘Fertility’ give proportions of the population.

# Calling lm

```
summary(lm(Fertility ~ . , data = swiss))
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	66.9152	10.70604	6.250	1.906e-07
Agriculture	-0.1721	0.07030	-2.448	1.873e-02
Examination	-0.2580	0.25388	-1.016	3.155e-01
Education	-0.8709	0.18303	-4.758	2.431e-05
Catholic	0.1041	0.03526	2.953	5.190e-03
Infant.Mortality	1.0770	0.38172	2.822	7.336e-03

# Example interpretation

- Agriculture is expressed in percentages (0 - 100)
- Estimate is -0.1721.
- We estimate an expected 0.17 decrease in standardized fertility for every 1% increase in percentage of males involved in agriculture in holding the remaining variables constant.
- The t-test for  $H_0 : \beta_{\text{Agri}} = 0$  versus  $H_a : \beta_{\text{Agri}} \neq 0$  is significant.
- Interestingly, the unadjusted estimate is

```
summary(lm(Fertility ~ Agriculture, data = swiss))$coefficients
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	60.3044	4.25126	14.185	3.216e-18
Agriculture	0.1942	0.07671	2.532	1.492e-02

How can adjustment reverse the sign of an effect? Let's try a simulation.

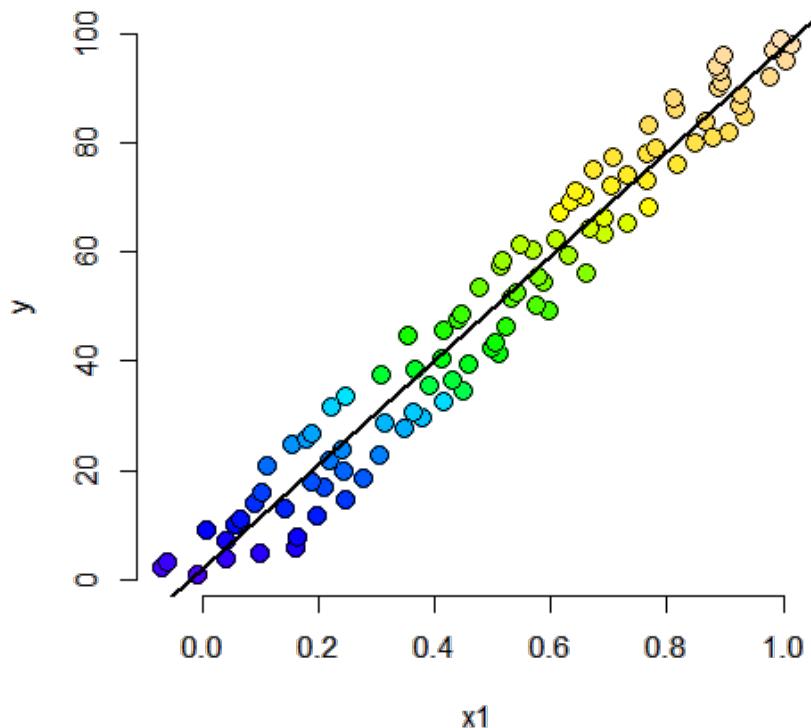
```
n <- 100; x2 <- 1 : n; x1 <- .01 * x2 + runif(n, -.1, .1); y = -x1 + x2 + rnorm(n, sd = .01)
summary(lm(y ~ x1))$coef
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.618	1.200	1.349	1.806e-01
x1	95.854	2.058	46.579	1.153e-68

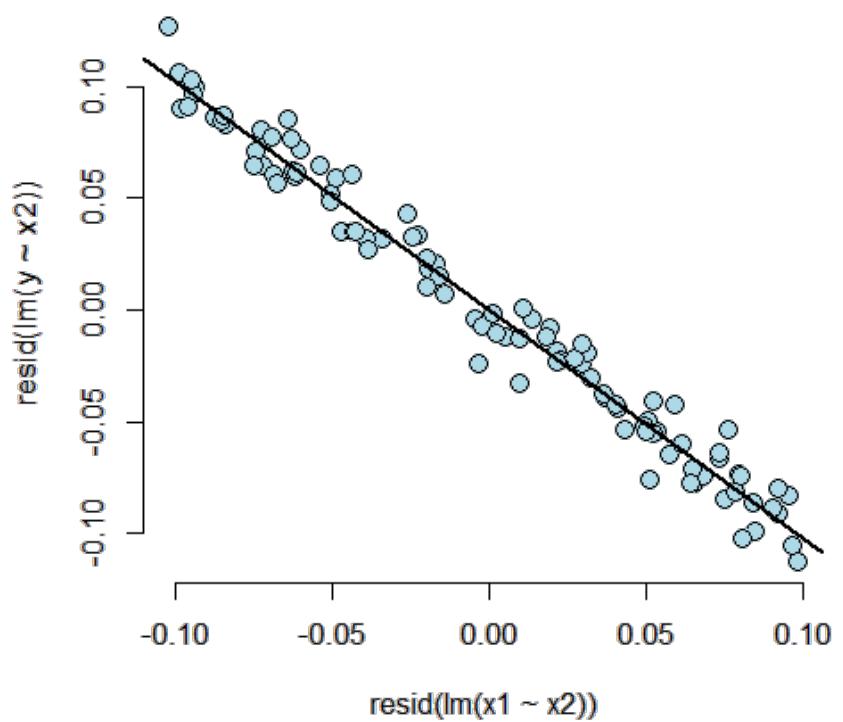
```
summary(lm(y ~ x1 + x2))$coef
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.0003683	0.0020141	0.1829	8.553e-01
x1	-1.0215256	0.0166372	-61.4001	1.922e-79
x2	1.0001909	0.0001681	5950.1818	1.369e-271

**Unadjusted, color is X2**



**Adjusted**



# Back to this data set

- The sign reverses itself with the inclusion of Examination and Education, but of which are negatively correlated with Agriculture.
- The percent of males in the province working in agriculture is negatively related to educational attainment (correlation of -0.6395) and Education and Examination (correlation of 0.6984) are obviously measuring similar things.
  - Is the positive marginal an artifact for not having accounted for, say, Education level? (Education does have a stronger effect, by the way.)
- At the minimum, anyone claiming that provinces that are more agricultural have higher fertility rates would immediately be open to criticism.

# What if we include an unnecessary variable?

$z$  adds no new linear information, since it's a linear combination of variables already included. R just drops terms that are linear combinations of other terms.

```
z <- swiss$Agriculture + swiss$Education  
lm(Fertility ~ . + z, data = swiss)
```

Call:

```
lm(formula = Fertility ~ . + z, data = swiss)
```

Coefficients:

(Intercept)	Agriculture	Examination	Education	Catholic
66.915	-0.172	-0.258	-0.871	0.104
Infant.Mortality	$z$			
1.077	NA			

# Dummy variables are smart

- Consider the linear model

$$Y_i = \beta_0 + X_{i1}\beta_1 + \epsilon_i$$

where each  $X_{i1}$  is binary so that it is a 1 if measurement  $i$  is in a group and 0 otherwise. (Treated versus not in a clinical trial, for example.)

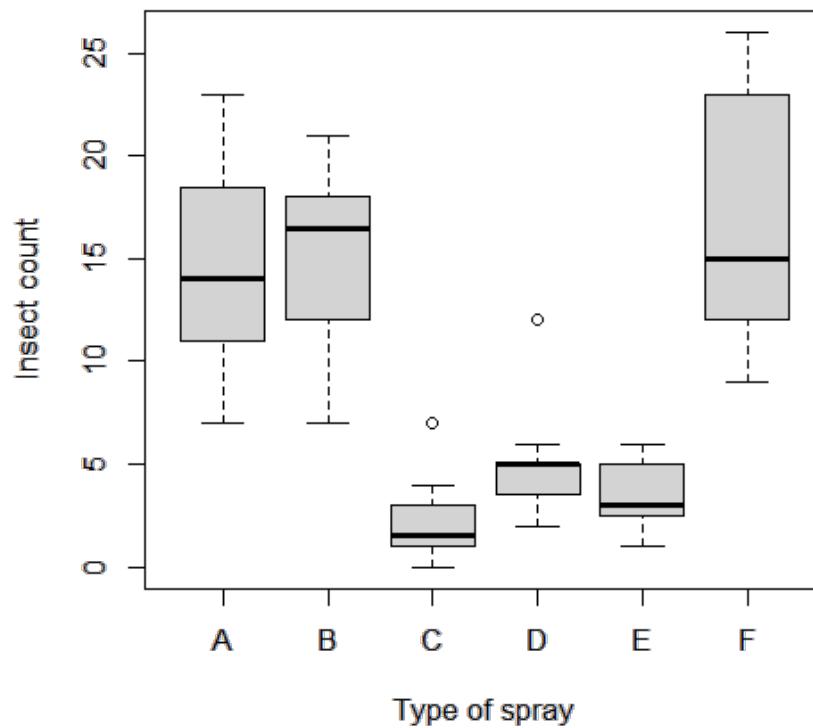
- Then for people in the group  $E[Y_i] = \beta_0 + \beta_1$
- And for people not in the group  $E[Y_i] = \beta_0$
- The LS fits work out to be  $\hat{\beta}_0 + \hat{\beta}_1$  is the mean for those in the group and  $\hat{\beta}_0$  is the mean for those not in the group.
- $\beta_1$  is interpreted as the increase or decrease in the mean comparing those in the group to those not.
- Note including a binary variable that is 1 for those not in the group would be redundant. It would create three parameters to describe two means.

# More than 2 levels

- Consider a multilevel factor level. For didactic reasons, let's say a three level factor (example, US political party affiliation: Republican, Democrat, Independent)
- $Y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \epsilon_i$ .
- $X_{i1}$  is 1 for Republicans and 0 otherwise.
- $X_{i2}$  is 1 for Democrats and 0 otherwise.
- If  $i$  is Republican  $E[Y_i] = \beta_0 + \beta_1$
- If  $i$  is Democrat  $E[Y_i] = \beta_0 + \beta_2$ .
- If  $i$  is Independent  $E[Y_i] = \beta_0$ .
- $\beta_1$  compares Republicans to Independents.
- $\beta_2$  compares Democrats to Independents.
- $\beta_1 - \beta_2$  compares Republicans to Democrats.
- (Choice of reference category changes the interpretation.)

# Insect Sprays

InsectSprays data



# Linear model fit, group A is the reference

```
summary(lm(count ~ spray, data = InsectSprays))$coef
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	14.5000	1.132	12.8074	1.471e-19
sprayB	0.8333	1.601	0.5205	6.045e-01
sprayC	-12.4167	1.601	-7.7550	7.267e-11
sprayD	-9.5833	1.601	-5.9854	9.817e-08
sprayE	-11.0000	1.601	-6.8702	2.754e-09
sprayF	2.1667	1.601	1.3532	1.806e-01

# Hard coding the dummy variables

```
summary(lm(count ~  
          I(1 * (spray == 'B')) + I(1 * (spray == 'C')) +  
          I(1 * (spray == 'D')) + I(1 * (spray == 'E')) +  
          I(1 * (spray == 'F'))  
          , data = InsectSprays))$coef
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	14.5000	1.132	12.8074	1.471e-19
I(1 * (spray == "B"))	0.8333	1.601	0.5205	6.045e-01
I(1 * (spray == "C"))	-12.4167	1.601	-7.7550	7.267e-11
I(1 * (spray == "D"))	-9.5833	1.601	-5.9854	9.817e-08
I(1 * (spray == "E"))	-11.0000	1.601	-6.8702	2.754e-09
I(1 * (spray == "F"))	2.1667	1.601	1.3532	1.806e-01

# What if we include all 6?

```
lm(count ~  
  I(1 * (spray == 'B')) + I(1 * (spray == 'C')) +  
  I(1 * (spray == 'D')) + I(1 * (spray == 'E')) +  
  I(1 * (spray == 'F')) + I(1 * (spray == 'A')), data = InsectSprays)
```

Call:

```
lm(formula = count ~ I(1 * (spray == "B")) + I(1 * (spray ==  
  "C")) + I(1 * (spray == "D")) + I(1 * (spray == "E")) + I(1 *  
  (spray == "F")) + I(1 * (spray == "A")), data = InsectSprays)
```

Coefficients:

(Intercept)	I(1 * (spray == "B"))	I(1 * (spray == "C"))	I(1 * (spray == "D"))	
14.500		0.833	-12.417	-9.583
I(1 * (spray == "E"))	I(1 * (spray == "F"))	I(1 * (spray == "A"))		
-11.000		2.167		NA

# What if we omit the intercept?

```
summary(lm(count ~ spray - 1, data = InsectSprays))$coef
```

	Estimate	Std. Error	t value	Pr(> t )
sprayA	14.500	1.132	12.807	1.471e-19
sprayB	15.333	1.132	13.543	1.002e-20
sprayC	2.083	1.132	1.840	7.024e-02
sprayD	4.917	1.132	4.343	4.953e-05
sprayE	3.500	1.132	3.091	2.917e-03
sprayF	16.667	1.132	14.721	1.573e-22

```
unique(ave(InsectSprays$count, InsectSprays$spray))
```

```
[1] 14.500 15.333 2.083 4.917 3.500 16.667
```

# Summary

- If we treat Spray as a factor, R includes an intercept and omits the alphabetically first level of the factor.
  - All t-tests are for comparisons of Sprays versus Spray A.
  - Empirical mean for A is the intercept.
  - Other group means are the itc plus their coefficient.
- If we omit an intercept, then it includes terms for all levels of the factor.
  - Group means are the coefficients.
  - Tests are tests of whether the groups are different than zero. (Are the expected counts zero for that spray.)
- If we want comparisons between, Spray B and C, say we could refit the model with C (or B) as the reference level.

# Reordering the levels

```
spray2 <- relevel(InsectSprays$spray, "C")
summary(lm(count ~ spray2, data = InsectSprays))$coef
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.083	1.132	1.8401	7.024e-02
spray2A	12.417	1.601	7.7550	7.267e-11
spray2B	13.250	1.601	8.2755	8.510e-12
spray2D	2.833	1.601	1.7696	8.141e-02
spray2E	1.417	1.601	0.8848	3.795e-01
spray2F	14.583	1.601	9.1083	2.794e-13

# Doing it manually

Equivalently

$$\text{Var}(\hat{\beta}_B - \hat{\beta}_C) = \text{Var}(\hat{\beta}_B) + \text{Var}(\hat{\beta}_C) - 2\text{Cov}(\hat{\beta}_B, \hat{\beta}_C)$$

```
fit <- lm(count ~ spray, data = InsectSprays) #A is ref
bbmbc <- coef(fit)[2] - coef(fit)[3] #B - C
temp <- summary(fit)
se <- temp$sigma * sqrt(temp$cov.unscaled[2, 2] + temp$cov.unscaled[3,3] - 2 *temp$cov.unscaled[2,3])
t <- (bbmbc) / se
p <- pt(-abs(t), df = fit$df)
out <- c(bbmbc, se, t, p)
names(out) <- c("B - C", "SE", "T", "P")
round(out, 3)
```

B - C	SE	T	P
13.250	1.601	8.276	0.000

# Other thoughts on this data

- Counts are bounded from below by 0, violates the assumption of normality of the errors.
  - Also there are counts near zero, so both the actual assumption and the intent of the assumption are violated.
- Variance does not appear to be constant.
- Perhaps taking logs of the counts would help.
  - There are 0 counts, so maybe  $\log(\text{Count} + 1)$
- Also, we'll cover Poisson GLMs for fitting count data.

# Example - Millennium Development Goal 1

[http://www.un.org/millenniumgoals/pdf/MDG\\_FS\\_1\\_EN.pdf](http://www.un.org/millenniumgoals/pdf/MDG_FS_1_EN.pdf)

[http://apps.who.int/gho/athena/data/GHO/WHOSIS\\_000008.csv?profile=text&filter=COUNTRY::SEX:](http://apps.who.int/gho/athena/data/GHO/WHOSIS_000008.csv?profile=text&filter=COUNTRY::SEX:)

# WHO childhood hunger data

```
#download.file("http://apps.who.int/gho/athena/data/GHO/WHOSIS_000008.csv?profile=text&filter=COUNTRY:*  
hunger <- read.csv("hunger.csv")  
hunger <- hunger[hunger$Sex!="Both sexes",]  
head(hunger)
```

	Indicator	Data.Source	PUBLISH.STATES	Year	WHO.region		
	Country	Sex	Display.Value	Numeric	Low	High	Comments
1	Children aged <5 years underweight (%)	NLIS_310044	Published	1986	Africa		
2	Children aged <5 years underweight (%)	NLIS_310233	Published	1990	Americas		
3	Children aged <5 years underweight (%)	NLIS_312902	Published	2005	Americas		
5	Children aged <5 years underweight (%)	NLIS_312522	Published	2002	Eastern Mediterranean		
6	Children aged <5 years underweight (%)	NLIS_312955	Published	2008	Africa		
8	Children aged <5 years underweight (%)	NLIS_312963	Published	2008	Africa		
1	Senegal	Male	19.3	19.3	NA	NA	NA
2	Paraguay	Male	2.2	2.2	NA	NA	NA
3	Nicaragua	Male	5.3	5.3	NA	NA	NA
5	Jordan	Female	3.2	3.2	NA	NA	NA
6	Guinea-Bissau	Female	17.0	17.0	NA	NA	NA
8	Ghana	Male	15.7	15.7	NA	NA	NA

# Plot percent hungry versus time

```
lm1 <- lm(hunger$Numeric ~ hunger$Year)  
plot(hunger$Year,hunger$Numeric,pch=19,col="blue")
```



# Remember the linear model

$$Hu_i = b_0 + b_1 Y_i + e_i$$

$b_0$  = percent hungry at Year 0

$b_1$  = decrease in percent hungry per year

$e_i$  = everything we didn't measure

# Add the linear model

```
lm1 <- lm(hunger$Numeric ~ hunger$Year)
plot(hunger$Year,hunger$Numeric,pch=19,col="blue")
lines(hunger$Year, lm1$fitted, lwd=3,col="darkgrey")
```



# Color by male/female

```
plot(hunger$Year,hunger$Numeric,pch=19)
points(hunger$Year,hunger$Numeric,pch=19,col=( (hunger$Sex=="Male")*1+1 ))
```



# Now two lines

$$HuF_i = bf_0 + bf_1 YF_i + ef_i$$

$bf_0$  = percent of girls hungry at Year 0

$bf_1$  = decrease in percent of girls hungry per year

$ef_i$  = everything we didn't measure

$$HuM_i = bm_0 + bm_1 YM_i + em_i$$

$bm_0$  = percent of boys hungry at Year 0

$bm_1$  = decrease in percent of boys hungry per year

$em_i$  = everything we didn't measure

# Color by male/female

```
lmM <- lm(hunger$Numeric[hunger$Sex=="Male"] ~ hunger$Year[hunger$Sex=="Male"])
lmF <- lm(hunger$Numeric[hunger$Sex=="Female"] ~ hunger$Year[hunger$Sex=="Female"])
plot(hunger$Year,hunger$Numeric,pch=19)
points(hunger$Year,hunger$Numeric,pch=19,col=((hunger$Sex=="Male")*1+1))
lines(hunger$Year[hunger$Sex=="Male"],lmM$fitted,col="black",lwd=3)
lines(hunger$Year[hunger$Sex=="Female"],lmF$fitted,col="red",lwd=3)
```



# Two lines, same slope

$$Hu_i = b_0 + b_1 \mathbb{I}(\text{Sex}_i = " \text{Male} ") + b_2 Y_i + e_i^*$$

$b_0$  - percent hungry at year zero for females

$b_0 + b_1$  - percent hungry at year zero for males

$b_2$  - change in percent hungry (for either males or females) in one year

$e_i^*$  - everything we didn't measure

# Two lines, same slope in R

```
lmBoth <- lm(hunger$Numeric ~ hunger$Year + hunger$Sex)
plot(hunger$Year,hunger$Numeric,pch=19)
points(hunger$Year,hunger$Numeric,pch=19,col=((hunger$Sex=="Male")*1+1))
abline(c(lmBoth$coeff[1],lmBoth$coeff[2]),col="red",lwd=3)
abline(c(lmBoth$coeff[1] + lmBoth$coeff[3],lmBoth$coeff[2] ),col="black",lwd=3)
```



# Two lines, different slopes (interactions)

$$Hu_i = b_0 + b_1 \mathbb{I}(\text{Sex}_i = "Male") + b_2 Y_i + b_3 \mathbb{I}(\text{Sex}_i = "Male") \times Y_i + e_i^+$$

$b_0$  - percent hungry at year zero for females

$b_0 + b_1$  - percent hungry at year zero for males

$b_2$  - change in percent hungry (females) in one year

$b_2 + b_3$  - change in percent hungry (males) in one year

$e_i^+$  - everything we didn't measure

# Two lines, different slopes in R

```
lmBoth <- lm(hunger$Numeric ~ hunger$Year + hunger$Sex + hunger$Sex*hunger$Year)
plot(hunger$Year,hunger$Numeric,pch=19)
points(hunger$Year,hunger$Numeric,pch=19,col=((hunger$Sex=="Male")*1+1))
abline(c(lmBoth$coeff[1],lmBoth$coeff[2]),col="red",lwd=3)
abline(c(lmBoth$coeff[1] + lmBoth$coeff[3],lmBoth$coeff[2] + lmBoth$coeff[4]),col="black",lwd=3)
```



# Two lines, different slopes in R

```
summary(lmBoth)
```

Call:

```
lm(formula = hunger$Numeric ~ hunger$Year + hunger$Sex + hunger$Sex *  
    hunger$Year)
```

Residuals:

Min	1Q	Median	3Q	Max
-25.91	-11.25	-1.85	7.09	46.15

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	603.5058	171.0552	3.53	0.00044 ***
hunger\$Year	-0.2934	0.0855	-3.43	0.00062 ***
hunger\$SexMale	61.9477	241.9086	0.26	0.79795
hunger\$Year:hunger\$SexMale	-0.0300	0.1209	-0.25	0.80402

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.2 on 944 degrees of freedom

Multiple R-squared: 0.0318, Adjusted R-squared: 0.0287

# Interpreting a continuous interaction

$$E[Y_i | X_{1i} = x_1, X_{2i} = x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Holding  $X_2$  constant we have

$$E[Y_i | X_{1i} = x_1 + 1, X_{2i} = x_2] - E[Y_i | X_{1i} = x_1, X_{2i} = x_2] = \beta_1 + \beta_3 x_2$$

And thus the expected change in  $Y$  per unit change in  $X_1$  holding all else constant is not constant.  $\beta_1$  is the slope when  $x_2 = 0$ . Note further that:

$$\begin{aligned} & E[Y_i | X_{1i} = x_1 + 1, X_{2i} = x_2 + 1] - E[Y_i | X_{1i} = x_1, X_{2i} = x_2 + 1] \\ & - E[Y_i | X_{1i} = x_1 + 1, X_{2i} = x_2] - E[Y_i | X_{1i} = x_1, X_{2i} = x_2] \\ & = \beta_3 \end{aligned}$$

Thus,  $\beta_3$  is the change in the expected change in  $Y$  per unit change in  $X_1$ , per unit change in  $X_2$ .

Or, the change in the slope relating  $X_1$  and  $Y$  per unit change in  $X_2$ .

# Example

$$Hu_i = b_0 + b_1 In_i + b_2 Y_i + b_3 In_i \times Y_i + e_i^+$$

$b_0$  - percent hungry at year zero for children with whose parents have no income

$b_1$  - change in percent hungry for each dollar of income in year zero

$b_2$  - change in percent hungry in one year for children whose parents have no income

$b_3$  - increased change in percent hungry by year for each dollar of income - e.g. if income is \$10,000, then change in percent hungry in one year will be

$$b_2 + 1e4 \times b_3$$

$e_i^+$  - everything we didn't measure

**Lot's of care/caution needed!**



# Multivariable regression

Regression

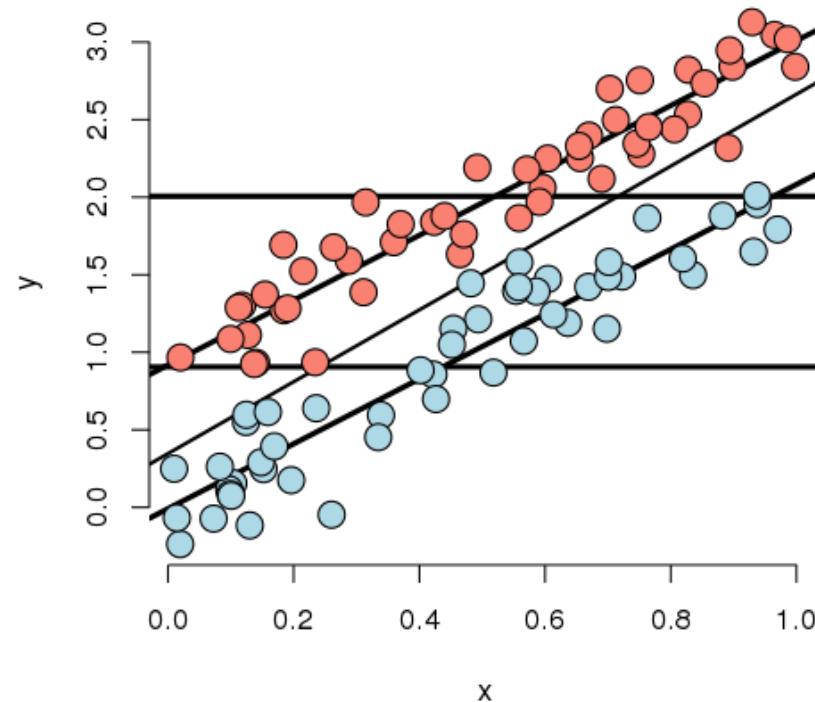
Brian Caffo, Jeff Leek, Roger Peng  
Johns Hopkins Bloomberg School of Public Health

# Consider the following simulated data

Code for the first plot, rest omitted (See the git repo for the rest of the code.)

```
n <- 100; t <- rep(c(0, 1), c(n/2, n/2)); x <- c(runif(n/2), runif(n/2));
beta0 <- 0; beta1 <- 2; tau <- 1; sigma <- .2
y <- beta0 + x * beta1 + t * tau + rnorm(n, sd = sigma)
plot(x, y, type = "n", frame = FALSE)
abline(lm(y ~ x), lwd = 2)
abline(h = mean(y[1 : (n/2)]), lwd = 3)
abline(h = mean(y[(n/2 + 1) : n]), lwd = 3)
fit <- lm(y ~ x + t)
abline(coef(fit)[1], coef(fit)[2], lwd = 3)
abline(coef(fit)[1] + coef(fit)[3], coef(fit)[2], lwd = 3)
points(x[1 : (n/2)], y[1 : (n/2)], pch = 21, col = "black", bg = "lightblue", cex = 2)
points(x[(n/2 + 1) : n], y[(n/2 + 1) : n], pch = 21, col = "black", bg = "salmon", cex = 2)
```

# Simulation 1

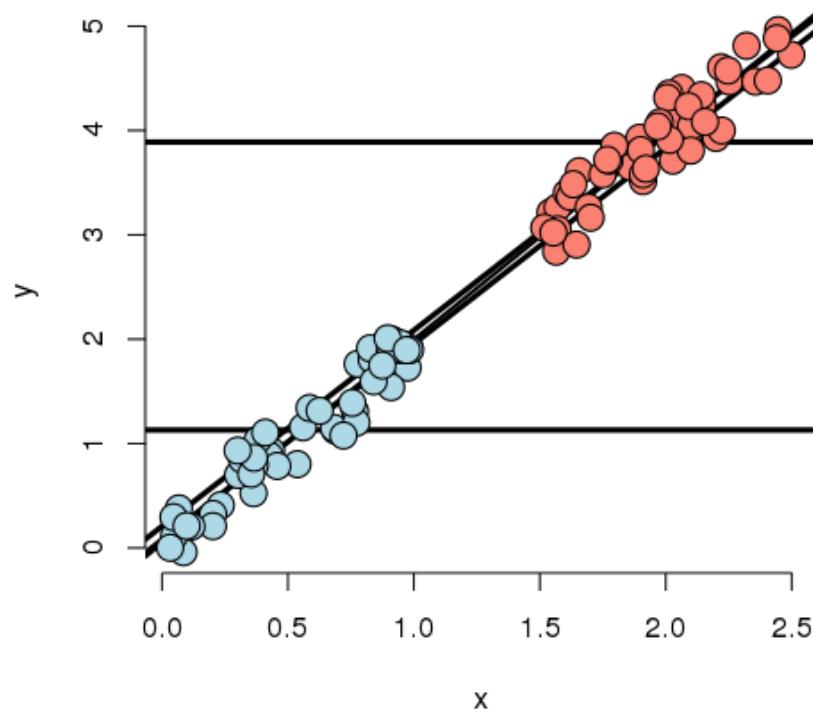


# Discussion

Some things to note in this simulation

- The X variable is unrelated to group status
- The X variable is related to Y, but the intercept depends on group status.
- The group variable is related to Y.
  - The relationship between group status and Y is constant depending on X.
  - The relationship between group and Y disregarding X is about the same as holding X constant

# Simulation 2

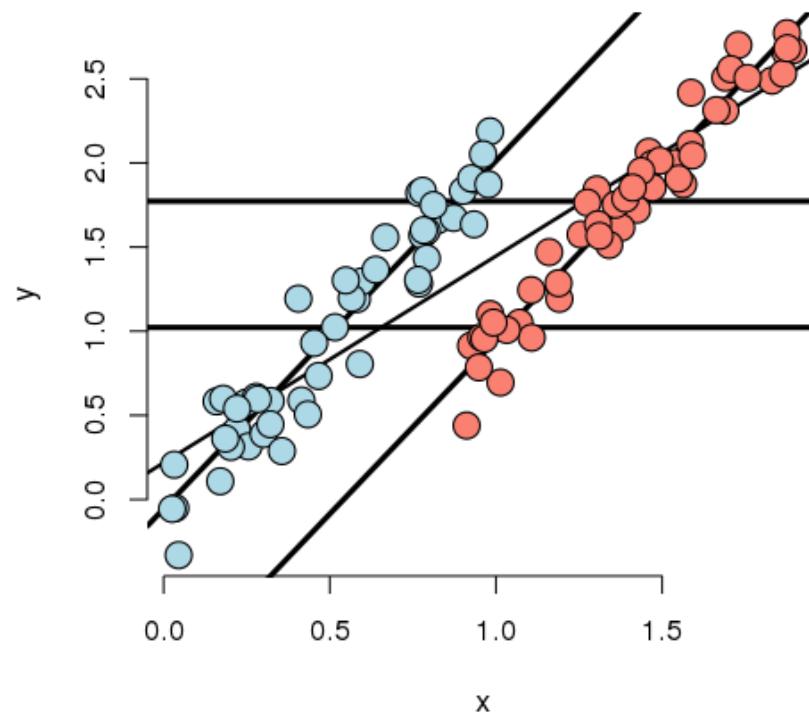


# Discussion

Some things to note in this simulation

- The X variable is highly related to group status
- The X variable is related to Y, the intercept doesn't depend on the group variable.
  - The X variable remains related to Y holding group status constant
- The group variable is marginally related to Y disregarding X.
- The model would estimate no adjusted effect due to group.
  - There isn't any data to inform the relationship between group and Y.
  - This conclusion is entirely based on the model.

# Simulation 3

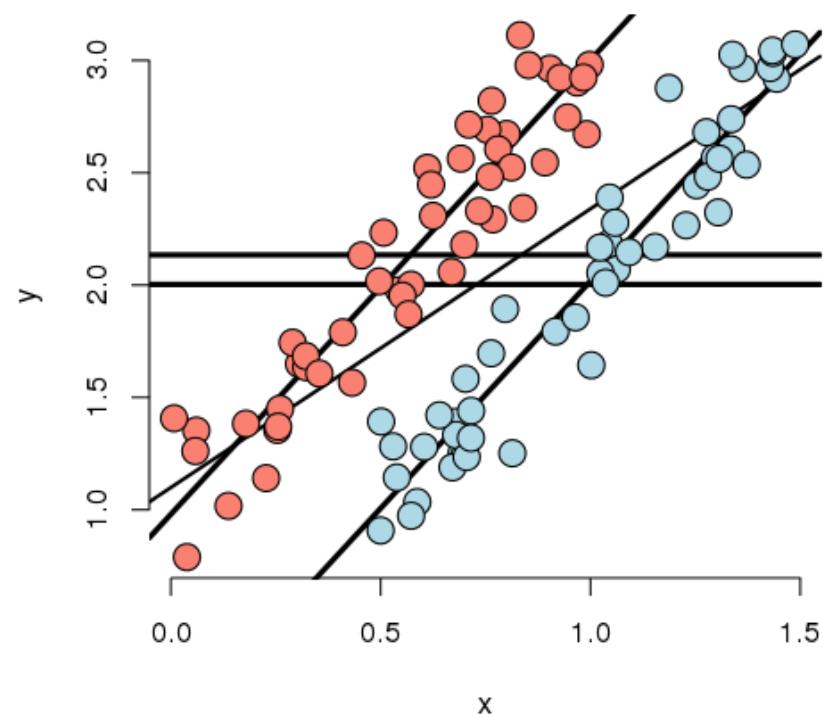


# Discussion

Some things to note in this simulation

- Marginal association has red group higher than blue.
- Adjusted relationship has blue group higher than red.
- Group status related to X.
- There is some direct evidence for comparing red and blue holding X fixed.

# Simulation 4

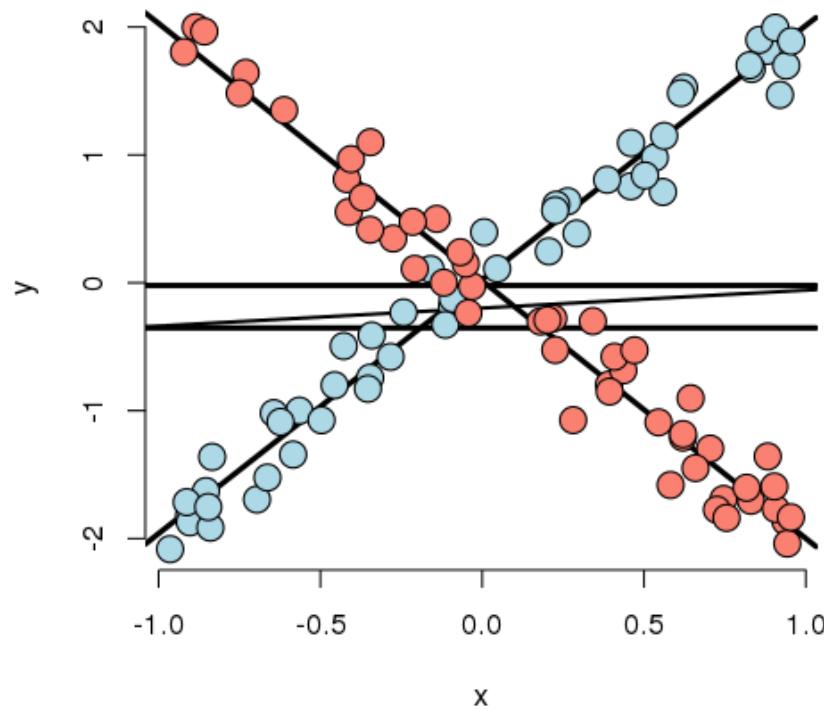


# Discussion

Some things to note in this simulation

- No marginal association between group status and Y.
- Strong adjusted relationship.
- Group status not related to X.
- There is lots of direct evidence for comparing red and blue holding X fixed.

# Simulation 5

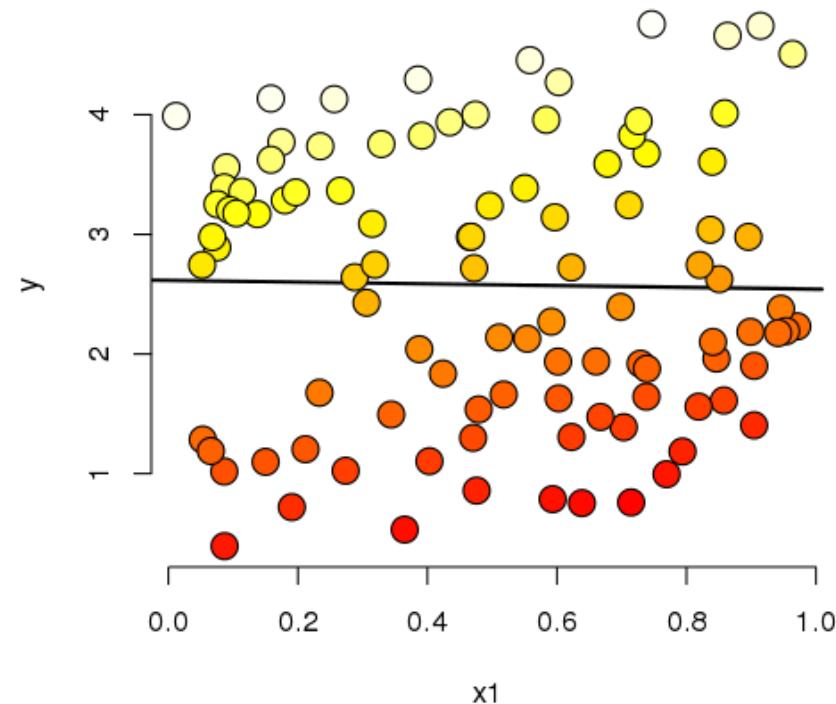


# Discussion

Some things to note from this simulation

- There is no such thing as a group effect here.
  - The impact of group reverses itself depending on X.
  - Both intercept and slope depends on group.
- Group status and X unrelated.
  - There's lots of information about group effects holding X fixed.

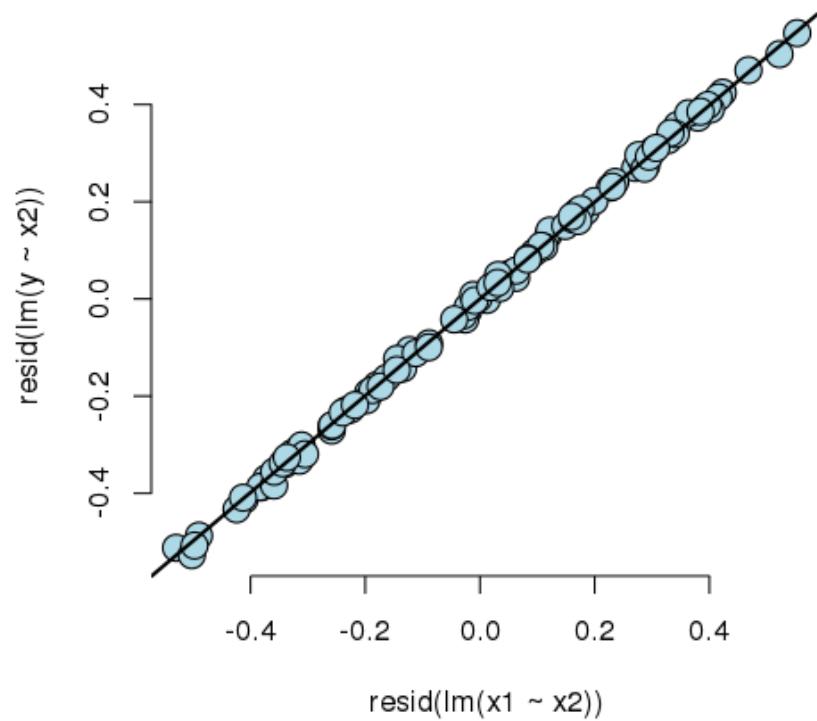
## Simulation 6



Do this to investigate the bivariate relationship

```
library(rgl)  
plot3d(x1, x2, y)
```

## Residual relationship



# Discussion

Some things to note from this simulation

- X1 unrelated to X2
- X2 strongly related to Y
- Adjusted relationship between X1 and Y largely unchanged by considering X2.
  - Almost no residual variability after accounting for X2.

# Some final thoughts

- Modeling multivariate relationships is difficult.
- Play around with simulations to see how the inclusion or exclusion of another variable can change analyses.
- The results of these analyses deal with the impact of variables on associations.
  - Ascertaining mechanisms or cause are difficult subjects to be added on top of difficulty in understanding multivariate associations.



# Residuals, diagnostics, variation

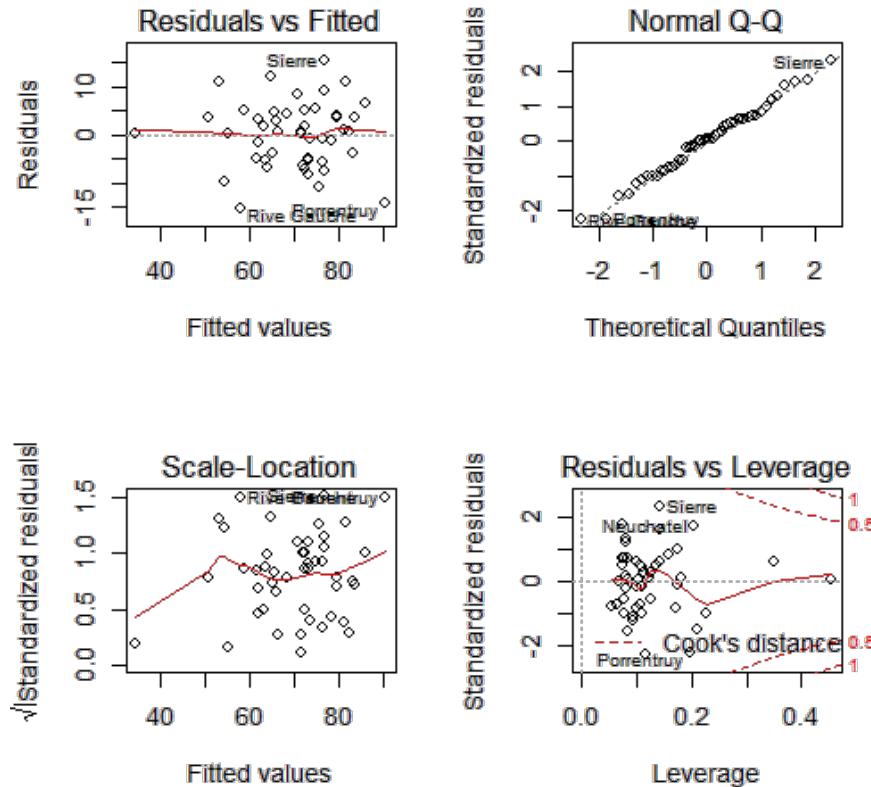
Regression

Brian Caffo, Jeff Leek, Roger Peng  
Johns Hopkins Bloomberg School of Public Health

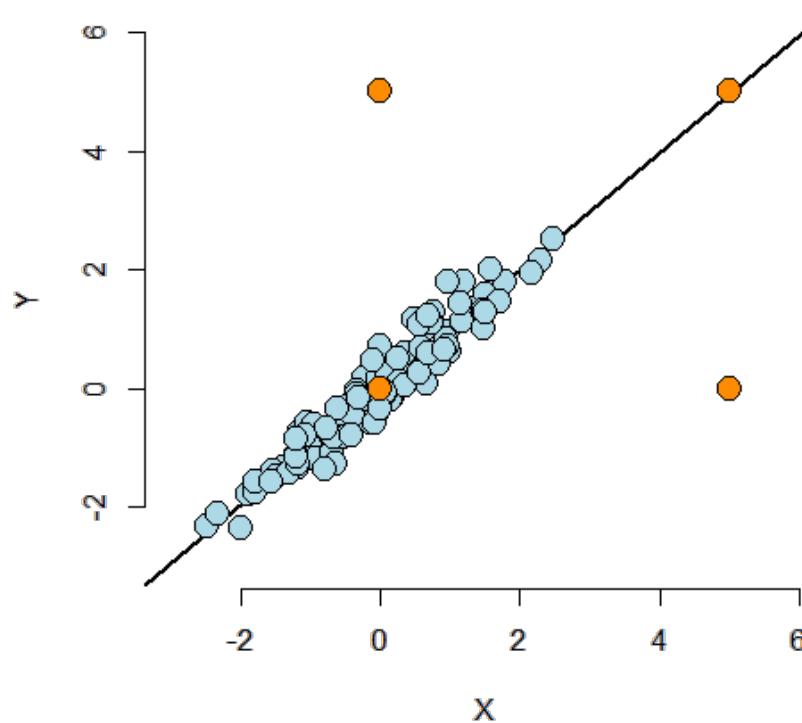
# The linear model

- Specified as  $Y_i = \sum_{k=1}^p X_{ik} \beta_j + \epsilon_i$
- We'll also assume here that  $\epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$
- Define the residuals as  $e_i = Y_i - \hat{Y}_i = Y_i - \sum_{k=1}^p X_{ik} \hat{\beta}_j$
- Our estimate of residual variation is  $\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n-p}$ , the  $n - p$  so that  $E[\hat{\sigma}^2] = \sigma^2$

```
data(swiss); par(mfrow = c(2, 2))
fit <- lm(Fertility ~ . , data = swiss); plot(fit)
```



# Influential, high leverage and outlying points



# Summary of the plot

Calling a point an outlier is vague.

- Outliers can be the result of spurious or real processes.
- Outliers can have varying degrees of influence.
- Outliers can conform to the regression relationship (i.e being marginally outlying in X or Y, but not outlying given the regression relationship).
  - Upper left hand point has low leverage, low influence, outliers in a way not conforming to the regression relationship.
  - Lower left hand point has low leverage, low influence and is not to be an outlier in any sense.
  - Upper right hand point has high leverage, but chooses not to exert it and thus would have low actual influence by conforming to the regression relationship of the other points.
  - Lower right hand point has high leverage and would exert it if it were included in the fit.

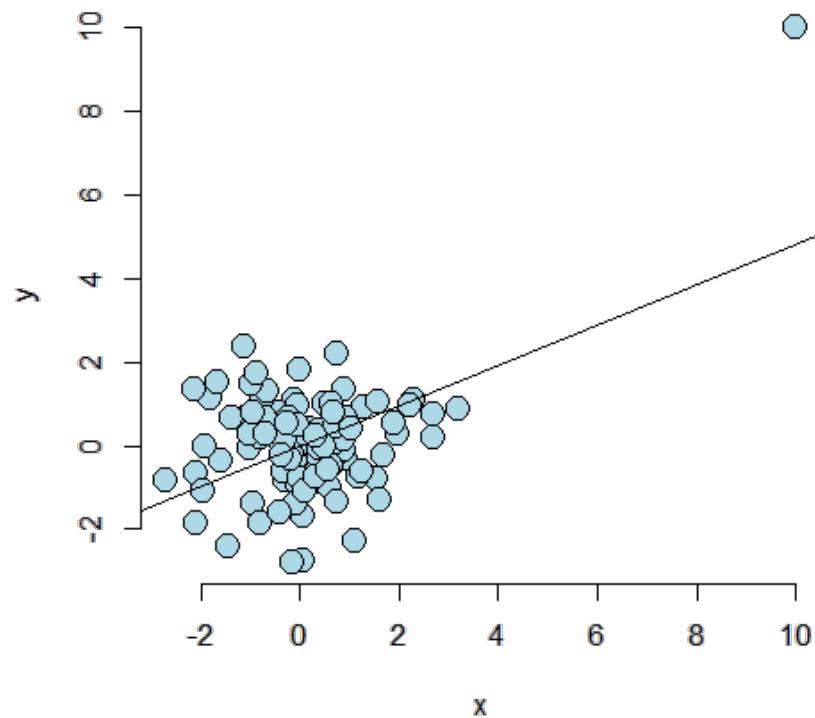
# Influence measures

- Do `?influence.measures` to see the full suite of influence measures in stats. The measures include
  - `rstandard` - standardized residuals, residuals divided by their standard deviations)
  - `rstudent` - standardized residuals, residuals divided by their standard deviations, where the  $i^{\text{th}}$  data point was deleted in the calculation of the standard deviation for the residual to follow a t distribution
  - `hatvalues` - measures of leverage
  - `dffits` - change in the predicted response when the  $i^{\text{th}}$  point is deleted in fitting the model.
  - `dfbetas` - change in individual coefficients when the  $i^{\text{th}}$  point is deleted in fitting the model.
  - `cooks.distance` - overall change in teh coefficients when the  $i^{\text{th}}$  point is deleted.
  - `resid` - returns the ordinary residuals
  - `resid(fit) / (1 - hatvalues(fit))` where `fit` is the linear model fit returns the PRESS residuals, i.e. the leave one out cross validation residuals - the difference in the response and the predicted response at data point  $i$ , where it was not included in the model fitting.

# How do I use all of these things?

- Be wary of simplistic rules for diagnostic plots and measures. The use of these tools is context specific. It's better to understand what they are trying to accomplish and use them judiciously.
- Not all of the measures have meaningful absolute scales. You can look at them relative to the values across the data.
- They probe your data in different ways to diagnose different problems.
- Patterns in your residual plots generally indicate some poor aspect of model fit. These can include:
  - Heteroskedasticity (non constant variance).
  - Missing model terms.
  - Temporal patterns (plot residuals versus collection order).
- Residual QQ plots investigate normality of the errors.
- Leverage measures (hat values) can be useful for diagnosing data entry errors.
- Influence measures get to the bottom line, 'how does deleting or including this point impact a particular aspect of the model'.

# Case 1



# The code

```
n <- 100; x <- c(10, rnorm(n)); y <- c(10, c(rnorm(n)))
plot(x, y, frame = FALSE, cex = 2, pch = 21, bg = "lightblue", col = "black")
abline(lm(y ~ x))
```

- The point `c(10, 10)` has created a strong regression relationship where there shouldn't be one.

# Showing a couple of the diagnostic values

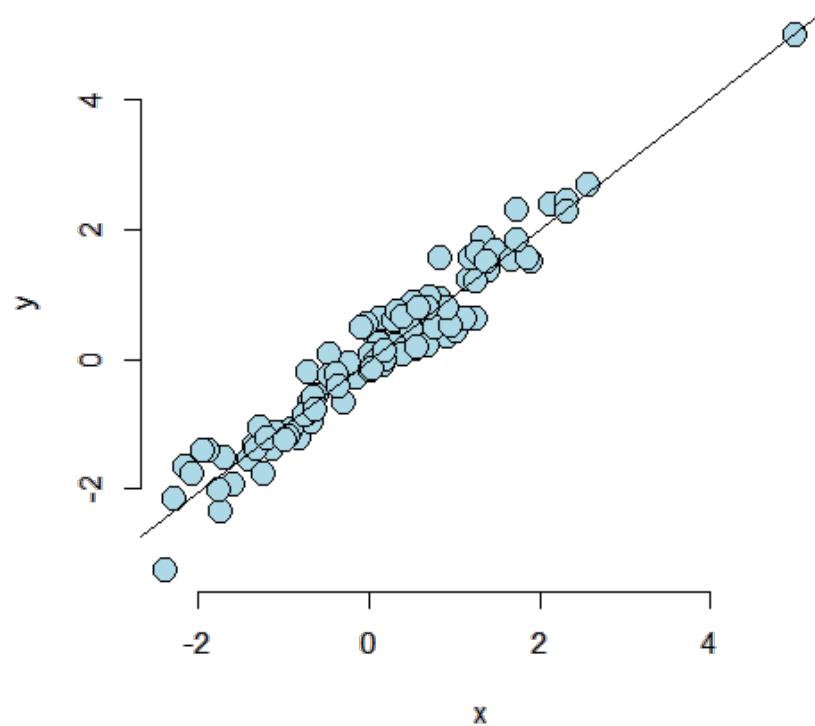
```
fit <- lm(y ~ x)
round(dfbetas(fit)[1 : 10, 2], 3)
```

1	2	3	4	5	6	7	8	9	10
6.007	-0.019	-0.007	0.014	-0.002	-0.083	-0.034	-0.045	-0.112	-0.008

```
round(hatvalues(fit)[1 : 10], 3)
```

1	2	3	4	5	6	7	8	9	10
0.445	0.010	0.011	0.011	0.030	0.017	0.012	0.033	0.021	0.010

## Case 2



# Looking at some of the diagnostics

```
round(dfbetas(fit2)[1 : 10, 2], 3)
```

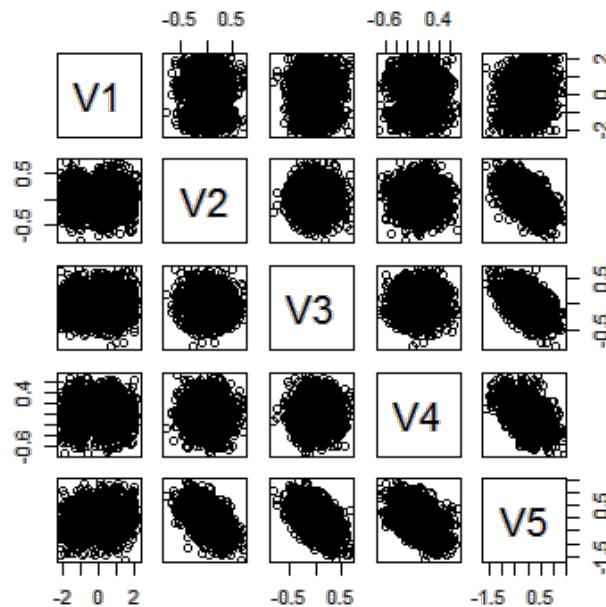
1	2	3	4	5	6	7	8	9	10
-0.072	-0.041	-0.007	0.012	0.008	-0.187	0.017	0.100	-0.059	0.035

```
round(hatvalues(fit2)[1 : 10], 3)
```

1	2	3	4	5	6	7	8	9	10
0.164	0.011	0.014	0.012	0.010	0.030	0.017	0.017	0.013	0.021

# Example described by Stefanski TAS 2007 Vol 61.

```
## Don't everyone hit this server at once. Read the paper first.  
dat <- read.table('http://www4.stat.ncsu.edu/~stefanski/NSF_Supported/Hidden_Images/orly_owl_files/orly  
pairs(dat)
```



# Got our P-values, should we bother to do a residual plot?

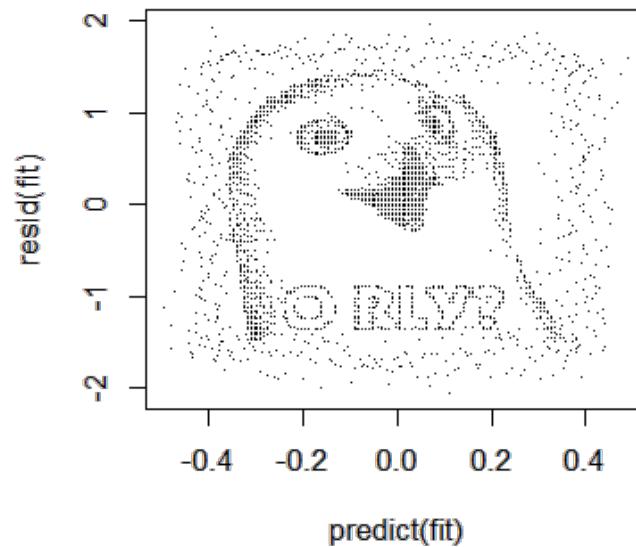
```
summary(lm(V1 ~ . - 1, data = dat))$coef
```

	Estimate	Std. Error	t value	Pr(> t )
V2	0.9856	0.12798	7.701	1.989e-14
V3	0.9715	0.12664	7.671	2.500e-14
V4	0.8606	0.11958	7.197	8.301e-13
V5	0.9267	0.08328	11.127	4.778e-28

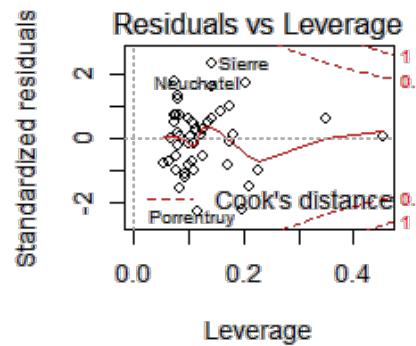
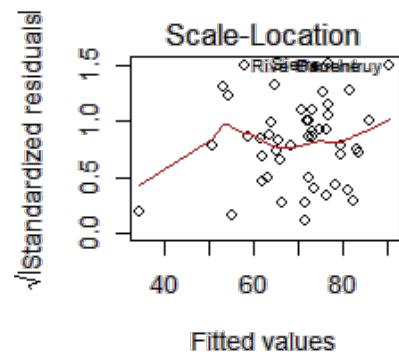
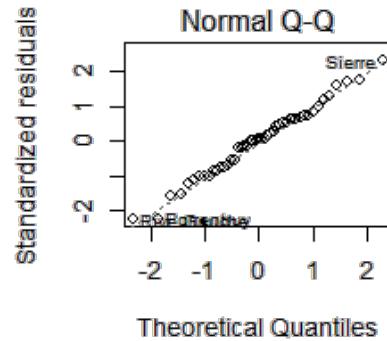
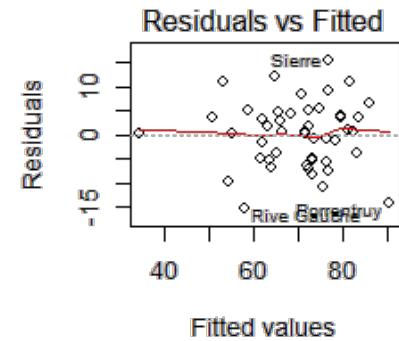
# Residual plot

P-values significant, O RLY?

```
fit <- lm(V1 ~ . - 1, data = dat); plot(predict(fit), resid(fit), pch = '.')
```



# Back to the Swiss data





# Multiple variables

Regression

Brian Caffo, Jeff Leek, Roger Peng  
Johns Hopkins Bloomberg School of Public Health

# Multivariable regression

- We have an entire class on prediction and machine learning, so we'll focus on modeling.
  - Prediction has a different set of criteria, needs for interpretability and standards for generalizability.
  - In modeling, our interest lies in parsimonious, interpretable representations of the data that enhance our understanding of the phenomena under study.
  - A model is a lens through which to look at your data. (I attribute this quote to Scott Zeger)
  - Under this philosophy, what's the right model? Whatever model connects the data to a true, parsimonious statement about what you're studying.
- There are nearly uncontable ways that a model can be wrong, in this lecture, we'll focus on variable inclusion and exclusion.
- Like nearly all aspects of statistics, good modeling decisions are context dependent.
  - A good model for prediction versus one for studying mechanisms versus one for trying to establish causal effects may not be the same.

# The Rumsfeldian triplet

*There are known knowns. These are things we know that we know. There are known unknowns. That is to say, there are things that we know we don't know. But there are also unknown unknowns. There are things we don't know we don't know.* Donald Rumsfeld

In our context

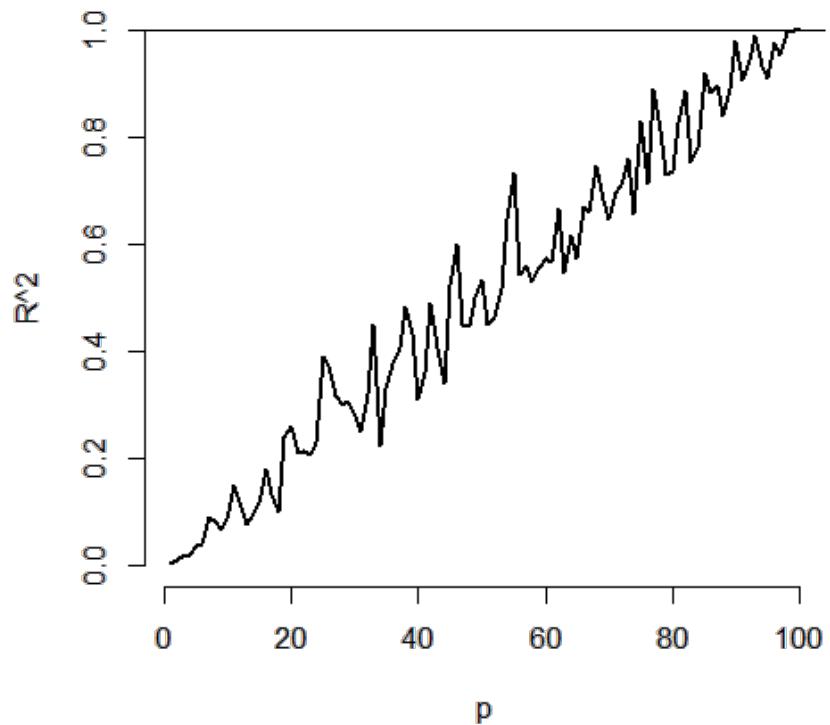
- (Known knowns) Regressors that we know we should check to include in the model and have.
- (Known Unknowns) Regressors that we would like to include in the model, but don't have.
- (Unknown Unknowns) Regressors that we don't even know about that we should have included in the model.

# General rules

- Omitting variables results in bias in the coefficients of interest - unless their regressors are uncorrelated with the omitted ones.
  - This is why we randomize treatments, it attempts to uncorrelate our treatment indicator with variables that we don't have to put in the model.
  - (If there's too many unobserved confounding variables, even randomization won't help you.)
- Including variables that we shouldn't have increases standard errors of the regression variables.
  - Actually, including any new variables increasese (actual, not estimated) standard errors of other regressors. So we don't want to idly throw variables into the model.
- The model must tend toward perfect fit as the number of non-redundant regressors approaches n.
- $R^2$  increases monotonically as more regressors are included.
- The SSE decreases monotonically as more regressors are included.

# Plot of $R^2$ versus n

For simulations as the number of variables included equals increases to  $n = 100$ . No actual regression relationship exist in any simulation



# Variance inflation

```
n <- 100; nosim <- 1000
x1 <- rnorm(n); x2 <- rnorm(n); x3 <- rnorm(n);
betas <- sapply(1 : nosim, function(i){
  y <- x1 + rnorm(n, sd = .3)
  c(coef(lm(y ~ x1))[2],
    coef(lm(y ~ x1 + x2))[2],
    coef(lm(y ~ x1 + x2 + x3))[2])
})
round(apply(betas, 1, sd), 5)
```

x1	x1	x1
0.02839	0.02872	0.02884

# Variance inflation

```
n <- 100; nosim <- 1000
x1 <- rnorm(n); x2 <- x1/sqrt(2) + rnorm(n) /sqrt(2)
x3 <- x1 * 0.95 + rnorm(n) * sqrt(1 - 0.95^2);
betas <- sapply(1 : nosim, function(i){
  y <- x1 + rnorm(n, sd = .3)
  c(coef(lm(y ~ x1))[2],
    coef(lm(y ~ x1 + x2))[2],
    coef(lm(y ~ x1 + x2 + x3))[2])
})
round(apply(betas, 1, sd), 5)
```

x1	x1	x1
0.03131	0.04270	0.09653

# Variance inflation factors

- Notice variance inflation was much worse when we included a variable that was highly related to  $x_1$ .
- We don't know  $\sigma$ , so we can only estimate the increase in the actual standard error of the coefficients for including a regressor.
- However,  $\sigma$  drops out of the relative standard errors. If one sequentially adds variables, one can check the variance (or sd) inflation for including each one.
- When the other regressors are actually orthogonal to the regressor of interest, then there is no variance inflation.
- The variance inflation factor (VIF) is the increase in the variance for the  $i$ th regressor compared to the ideal setting where it is orthogonal to the other regressors.
  - (The square root of the VIF is the increase in the sd ...)
- Remember, variance inflation is only part of the picture. We want to include certain variables, even if they dramatically inflate our variance.

# Revisiting our previous simulation

```
##doesn't depend on which y you use,  
y <- x1 + rnorm(n, sd = .3)  
a <- summary(lm(y ~ x1))$cov.unscaled[2,2]  
c(summary(lm(y ~ x1 + x2))$cov.unscaled[2,2],  
  summary(lm(y~ x1 + x2 + x3))$cov.unscaled[2,2]) / a
```

```
[1] 1.895 9.948
```

```
temp <- apply(betas, 1, var); temp[2 : 3] / temp[1]
```

```
x1    x1  
1.860 9.506
```

# Swiss data

```
data(swiss);
fit1 <- lm(Fertility ~ Agriculture, data = swiss)
a <- summary(fit1)$cov.unscaled[2,2]
fit2 <- update(fit, Fertility ~ Agriculture + Examination)
fit3 <- update(fit, Fertility ~ Agriculture + Examination + Education)
c(summary(fit2)$cov.unscaled[2,2],
  summary(fit3)$cov.unscaled[2,2]) / a
```

```
[1] 1.892 2.089
```

# Swiss data VIFs,

```
library(car)
fit <- lm(Fertility ~ . , data = swiss)
vif(fit)
```

Agriculture	Examination	Education	Catholic Infant.	Mortality
2.284	3.675	2.775	1.937	1.108

```
sqrt(vif(fit)) #I prefer sd
```

Agriculture	Examination	Education	Catholic Infant.	Mortality
1.511	1.917	1.666	1.392	1.052

# What about residual variance estimation?

- Assuming that the model is linear with additive iid errors (with finite variance), we can mathematically describe the impact of omitting necessary variables or including unnecessary ones.
  - If we underfit the model, the variance estimate is biased.
  - If we correctly or overfit the model, including all necessary covariates and/or unnecessary covariates, the variance estimate is unbiased.
  - However, the variance of the variance is larger if we include unnecessary variables.

# Covariate model selection

- Automated covariate selection is a difficult topic. It depends heavily on how rich of a covariate space one wants to explore.
  - The space of models explodes quickly as you add interactions and polynomial terms.
- In the prediction class, we'll cover many modern methods for traversing large model spaces for the purposes of prediction.
- Principal components or factor analytic models on covariates are often useful for reducing complex covariate spaces.
- Good design can often eliminate the need for complex model searches at analyses; though often control over the design is limited.
- If the models of interest are nested and without lots of parameters differentiating them, it's fairly uncontroversial to use nested likelihood ratio tests. (Example to follow.)
- My favorite approach is as follows. Given a coefficient that I'm interested in, I like to use covariate adjustment and multiple models to probe that effect to evaluate it for robustness and to see what other covariates knock it out. This isn't a terribly systematic approach, but it tends to teach you a lot about the the data as you get your hands dirty.

# How to do nested model testing in R

```
fit1 <- lm(Fertility ~ Agriculture, data = swiss)
fit3 <- update(fit, Fertility ~ Agriculture + Examination + Education)
fit5 <- update(fit, Fertility ~ Agriculture + Examination + Education + Catholic + Infant.Mortality)
anova(fit1, fit3, fit5)
```

## Analysis of Variance Table

Model 1: Fertility ~ Agriculture

Model 2: Fertility ~ Agriculture + Examination + Education

Model 3: Fertility ~ Agriculture + Examination + Education + Catholic +  
Infant.Mortality

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	45	6283				
2	43	3181	2	3102	30.2	8.6e-09 ***
3	41	2105	2	1076	10.5	0.00021 ***
---						
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						



# Generalized linear models

Regression Models

Brian Caffo, Jeff Leek, Roger Peng  
Johns Hopkins Bloomberg School of Public Health

# Linear models

- Linear models are the most useful applied statistical technique. However, they are not without their limitations.
  - Additive response models don't make much sense if the response is discrete, or strictly positive.
  - Additive error models often don't make sense, for example if the outcome has to be positive.
  - Transformations are often hard to interpret.
  - There's value in modeling the data on the scale that it was collected.
  - Particularly interpretable transformations, natural logarithms in specific, aren't applicable for negative or zero values.

# Generalized linear models

- Introduced in a 1972 RSSB paper by Nelder and Wedderburn.
- Involves three components
  - An *exponential family* model for the response.
  - A systematic component via a linear predictor.
  - A link function that connects the means of the response to the linear predictor.

# Example, linear models

- Assume that  $Y_i \sim N(\mu_i, \sigma^2)$  (the Gaussian distribution is an exponential family distribution.)
- Define the linear predictor to be  $\eta_i = \sum_{k=1}^p X_{ik} \beta_k$ .
- The link function as  $g$  so that  $g(\mu) = \eta$ .
  - For linear models  $g(\mu) = \mu$  so that  $\mu_i = \eta_i$
- This yields the same likelihood model as our additive error Gaussian linear model

$$Y_i = \sum_{k=1}^p X_{ik} \beta_k + \epsilon_i$$

where  $\epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$

# Example, logistic regression

- Assume that  $Y_i \sim \text{Bernoulli}(\mu_i)$  so that  $E[Y_i] = \mu_i$  where  $0 \leq \mu_i \leq 1$ .
- Linear predictor  $\eta_i = \sum_{k=1}^p X_{ik} \beta_k$
- Link function  $g(\mu) = \eta = \log\left(\frac{\mu}{1-\mu}\right)$   $g$  is the (natural) log odds, referred to as the **logit**.
- Note then we can invert the logit function as

$$\mu_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} \quad \text{and} \quad 1 - \mu_i = \frac{1}{1 + \exp(\eta_i)}$$

Thus the likelihood is

$$\prod_{i=1}^n \mu_i^{y_i} (1 - \mu_i)^{1-y_i} = \exp\left(\sum_{i=1}^n y_i \eta_i\right) \prod_{i=1}^n (1 + \eta_i)^{-1}$$

# Example, Poisson regression

- Assume that  $Y_i \sim \text{Poisson}(\mu_i)$  so that  $E[Y_i] = \mu_i$  where  $0 \leq \mu_i$
- Linear predictor  $\eta_i = \sum_{k=1}^p X_{ik} \beta_k$
- Link function  $g(\mu) = \eta = \log(\mu)$
- Recall that  $e^x$  is the inverse of  $\log(x)$  so that

$$\mu_i = e^{\eta_i}$$

Thus, the likelihood is

$$\prod_{i=1}^n (y_i!)^{-1} \mu_i^{y_i} e^{-\mu_i} \propto \exp\left( \sum_{i=1}^n y_i \eta_i - \sum_{i=1}^n \mu_i \right)$$

# Some things to note

- In each case, the only way in which the likelihood depends on the data is through

$$\sum_{i=1}^n y_i \eta_i = \sum_{i=1}^n y_i \sum_{k=1}^p X_{ik} \beta_k = \sum_{k=1}^p \beta_k \sum_{i=1}^n X_{ik} y_i$$

Thus if we don't need the full data, only  $\sum_{i=1}^n X_{ik} y_i$ . This simplification is a consequence of choosing so-called 'canonical' link functions.

- (This has to be derived). All models achieve their maximum at the root of the so called normal equations

$$0 = \sum_{i=1}^n \frac{(Y_i - \mu_i)}{\text{Var}(Y_i)} W_i$$

where  $W_i$  are the derivative of the inverse of the link function.

# About variances

$$0 = \sum_{i=1}^n \frac{(Y_i - \mu_i)}{\text{Var}(Y_i)} W_i$$

- For the linear model  $\text{Var}(Y_i) = \sigma^2$  is constant.
- For Bernoulli case  $\text{Var}(Y_i) = \mu_i(1 - \mu_i)$
- For the Poisson case  $\text{Var}(Y_i) = \mu_i$ .
- In the latter cases, it is often relevant to have a more flexible variance model, even if it doesn't correspond to an actual likelihood

$$0 = \sum_{i=1}^n \frac{(Y_i - \mu_i)}{\phi\mu_i(1 - \mu_i)} W_i \quad \text{and} \quad 0 = \sum_{i=1}^n \frac{(Y_i - \mu_i)}{\phi\mu_i} W_i$$

- These are called 'quasi-likelihood' normal equations

# Odds and ends

- The normal equations have to be solved iteratively. Resulting in  $\hat{\beta}_k$  and, if included,  $\hat{\phi}$ .
- Predicted linear predictor responses can be obtained as  $\hat{\eta} = \sum_{k=1}^p X_k \hat{\beta}_k$
- Predicted mean responses as  $\hat{\mu} = g^{-1}(\hat{\eta})$
- Coefficients are interpreted as

$$g(E[Y|X_k = x_k + 1, X_{\sim k} = x_{\sim k}]) - g(E[Y|X_k = x_k, X_{\sim k} = x_{\sim k}]) = \beta_k$$

or the change in the link function of the expected response per unit change in  $X_k$  holding other regressors constant.

- Variations on Newton/Raphson's algorithm are used to do it.
- Asymptotics are used for inference usually.
- Many of the ideas from linear models can be brought over to GLMs.



# Generalized linear models, binary data

Regression models

Brian Caffo, Jeff Leek and Roger Peng  
Johns Hopkins Bloomberg School of Public Health

# Key ideas

- Frequently we care about outcomes that have two values
  - Alive/dead
  - Win/loss
  - Success/Failure
  - etc
- Called binary, Bernoulli or 0/1 outcomes
- Collection of exchangeable binary outcomes for the same covariate data are called binomial outcomes.

# Example Baltimore Ravens win/loss

## Ravens Data

```
download.file("https://dl.dropboxusercontent.com/u/7710864/data/ravensData.rda"
              , destfile=".~/data/ravensData.rda",method="curl")
load("./data/ravensData.rda")
head(ravensData)
```

	ravenWinNum	ravenWin	ravenScore	opponentScore
1	1	W	24	9
2	1	W	38	35
3	1	W	28	13
4	1	W	34	31
5	1	W	44	13
6	0	L	23	24

# Linear regression

$$RW_i = b_0 + b_1 RS_i + e_i$$

$RW_i$  - 1 if a Ravens win, 0 if not

$RS_i$  - Number of points Ravens scored

$b_0$  - probability of a Ravens win if they score 0 points

$b_1$  - increase in probability of a Ravens win for each additional point

$e_i$  - residual variation due

# Linear regression in R

```
lmRavens <- lm(ravensData$ravenWinNum ~ ravensData$ravenScore)  
summary(lmRavens)$coef
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.2850	0.256643	1.111	0.28135
ravensData\$ravenScore	0.0159	0.009059	1.755	0.09625

# Odds

Binary Outcome 0/1

$$RW_i$$

Probability (0,1)

$$\Pr(RW_i | RS_i, b_0, b_1)$$

Odds  $(0, \infty)$

$$\frac{\Pr(RW_i | RS_i, b_0, b_1)}{1 - \Pr(RW_i | RS_i, b_0, b_1)}$$

Log odds  $(-\infty, \infty)$

$$\log\left(\frac{\Pr(RW_i | RS_i, b_0, b_1)}{1 - \Pr(RW_i | RS_i, b_0, b_1)}\right)$$

# Linear vs. logistic regression

Linear

$$RW_i = b_0 + b_1 RS_i + e_i$$

or

$$E[RW_i | RS_i, b_0, b_1] = b_0 + b_1 RS_i$$

Logistic

$$\Pr(RW_i | RS_i, b_0, b_1) = \frac{\exp(b_0 + b_1 RS_i)}{1 + \exp(b_0 + b_1 RS_i)}$$

or

$$\log\left(\frac{\Pr(RW_i | RS_i, b_0, b_1)}{1 - \Pr(RW_i | RS_i, b_0, b_1)}\right) = b_0 + b_1 RS_i$$

# Interpreting Logistic Regression

$$\log\left(\frac{\Pr(\text{RW}_i|\text{RS}_i, b_0, b_1)}{1 - \Pr(\text{RW}_i|\text{RS}_i, b_0, b_1)}\right) = b_0 + b_1 \text{RS}_i$$

$b_0$  - Log odds of a Ravens win if they score zero points

$b_1$  - Log odds ratio of win probability for each point scored (compared to zero points)

$\exp(b_1)$  - Odds ratio of win probability for each point scored (compared to zero points)

# Odds

- Imagine that you are playing a game where you flip a coin with success probability  $p$ .
- If it comes up heads, you win  $X$ . If it comes up tails, you lose  $Y$ .
- What should we set  $X$  and  $Y$  for the game to be fair?

$$E[\text{earnings}] = Xp - Y(1 - p) = 0$$

- Implies

$$\frac{Y}{X} = \frac{p}{1 - p}$$

- The odds can be said as "How much should you be willing to pay for a  $p$  probability of winning a dollar?"
  - (If  $p > 0.5$  you have to pay more if you lose than you get if you win.)
  - (If  $p < 0.5$  you have to pay less if you lose than you get if you win.)

# Visualizing fitting logistic regression curves

```
x <- seq(-10, 10, length = 1000)
manipulate(
  plot(x, exp(beta0 + beta1 * x) / (1 + exp(beta0 + beta1 * x)),
       type = "l", lwd = 3, frame = FALSE),
  beta1 = slider(-2, 2, step = .1, initial = 2),
  beta0 = slider(-2, 2, step = .1, initial = 0)
)
```

# Ravens logistic regression

```
logRegRavens <- glm(ravensData$ravenWinNum ~ ravensData$ravenScore, family="binomial")
summary(logRegRavens)
```

Call:

```
glm(formula = ravensData$ravenWinNum ~ ravensData$ravenScore,
family = "binomial")
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.758	-1.100	0.530	0.806	1.495

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.6800	1.5541	-1.08	0.28
ravensData\$ravenScore	0.1066	0.0667	1.60	0.11

(Dispersion parameter for binomial family taken to be 1)

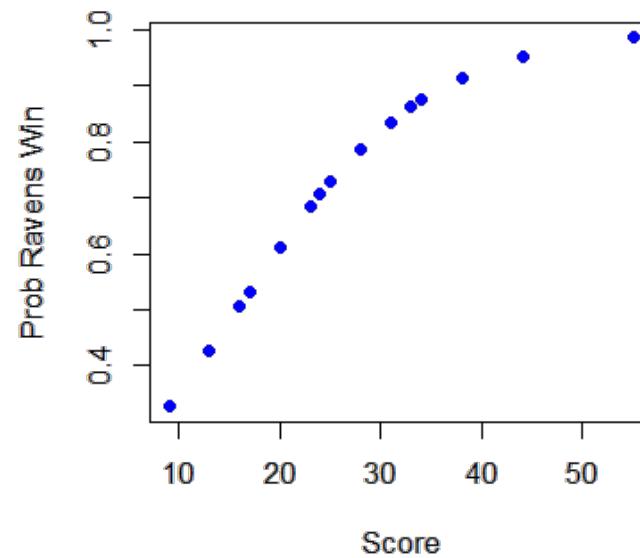
Null deviance: 24.435 on 19 degrees of freedom

Residual deviance: 20.895 on 18 degrees of freedom

AIC: 24.89

# Ravens fitted values

```
plot(ravensData$ravenScore, logRegRavens$fitted, pch=19, col="blue", xlab="Score", ylab="Prob Ravens Win")
```



# Odds ratios and confidence intervals

```
exp(logRegRavens$coeff)
```

	(Intercept)	ravensData\$ravenScore
	0.1864	1.1125

```
exp(confint(logRegRavens))
```

	2.5 %	97.5 %
(Intercept)	0.005675	3.106
ravensData\$ravenScore	0.996230	1.303

# ANOVA for logistic regression

```
anova(logRegRavens, test="Chisq")
```

Analysis of Deviance Table

Model: binomial, link: logit

Response: ravensData\$ravenWinNum

Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			19	24.4	
ravensData\$ravenScore	1	3.54	18	20.9	0.06 .
---					
Signif. codes:	0	'***'	0.001	'**'	0.01
		'*'	0.05	'. '	0.1
					' '
					1

# Interpreting Odds Ratios

- Not probabilities
- Odds ratio of 1 = no difference in odds
- Log odds ratio of 0 = no difference in odds
- Odds ratio < 0.5 or > 2 commonly a "moderate effect"
- Relative risk  $\frac{\Pr(\text{RW}_i | \text{RS}_i=1)}{\Pr(\text{RW}_i | \text{RS}_i=0)}$  often easier to interpret, harder to estimate
- For small probabilities RR  $\approx$  OR but **they are not the same!**

[Wikipedia on Odds Ratio](#)

# Further resources

- [Wikipedia on Logistic Regression](#)
- [Logistic regression and glms in R](#)
- Brian Caffo's lecture notes on: [Simpson's paradox](#), [Case-control studies](#)
- [Open Intro Chapter on Logistic Regression](#)



# Count outcomes, Poisson GLMs

Regression Models

Brian Caffo, Jeffrey Leek, Roger Peng  
Johns Hopkins Bloomberg School of Public Health

# Key ideas

- Many data take the form of counts
  - Calls to a call center
  - Number of flu cases in an area
  - Number of cars that cross a bridge
- Data may also be in the form of rates
  - Percent of children passing a test
  - Percent of hits to a website from a country
- Linear regression with transformation is an option

# Poisson distribution

- The Poisson distribution is a useful model for counts and rates
- Here a rate is count per some monitoring time
- Some examples uses of the Poisson distribution
  - Modeling web traffic hits
  - Incidence rates
  - Approximating binomial probabilities with small  $p$  and large  $n$
  - Analyzing contingency table data

# The Poisson mass function

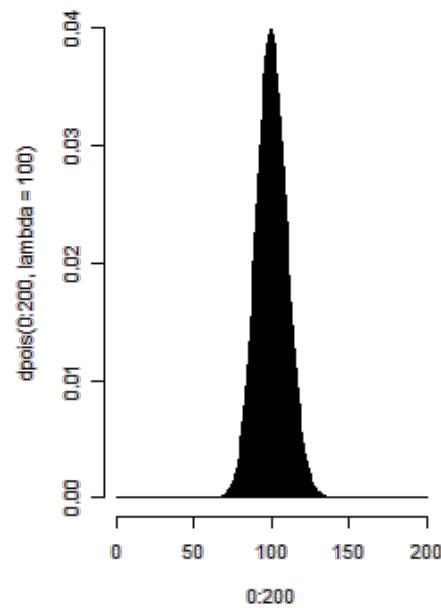
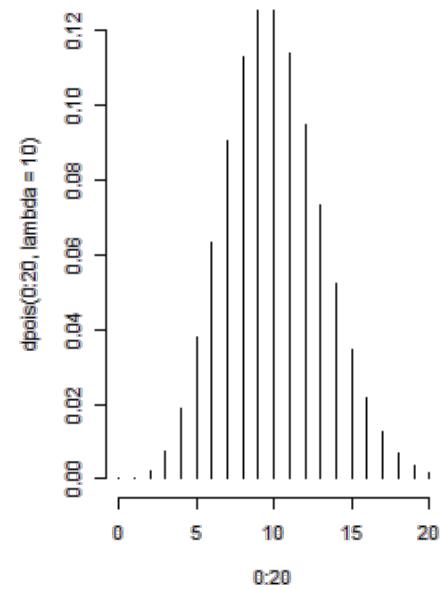
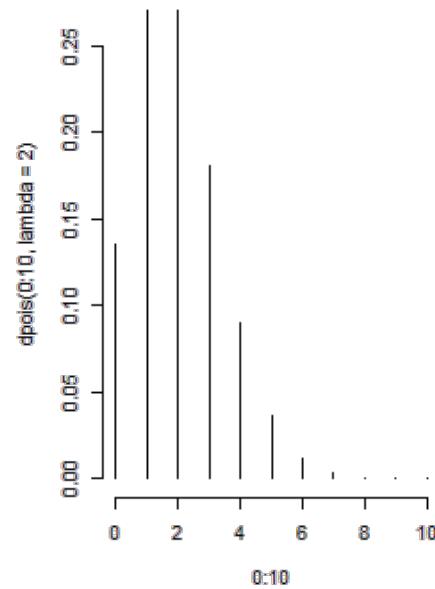
- $X \sim \text{Poisson}(t\lambda)$  if

$$P(X = x) = \frac{(t\lambda)^x e^{-t\lambda}}{x!}$$

For  $x = 0, 1, \dots$

- The mean of the Poisson is  $E[X] = t\lambda$ , thus  $E[X/t] = \lambda$
- The variance of the Poisson is  $\text{Var}(X) = t\lambda$ .
- The Poisson tends to a normal as  $t\lambda$  gets large.

```
par(mfrow = c(1, 3))
plot(0 : 10, dpois(0 : 10, lambda = 2), type = "h", frame = FALSE)
plot(0 : 20, dpois(0 : 20, lambda = 10), type = "h", frame = FALSE)
plot(0 : 200, dpois(0 : 200, lambda = 100), type = "h", frame = FALSE)
```



# Poisson distribution

Sort of, showing that the mean and variance are equal

```
x <- 0 : 10000; lambda = 3  
mu <- sum(x * dpois(x, lambda = lambda))  
sigmasq <- sum((x - mu)^2 * dpois(x, lambda = lambda))  
c(mu, sigmasq)
```

```
[1] 3 3
```

# Example: Leek Group Website Traffic

- Consider the daily counts to Jeff Leek's web site

<http://biostat.jhsph.edu/~jleek/>

- Since the unit of time is always one day, set  $t = 1$  and then the Poisson mean is interpreted as web hits per day. (If we set  $t = 24$ , it would be web hits per hour).

# Website data

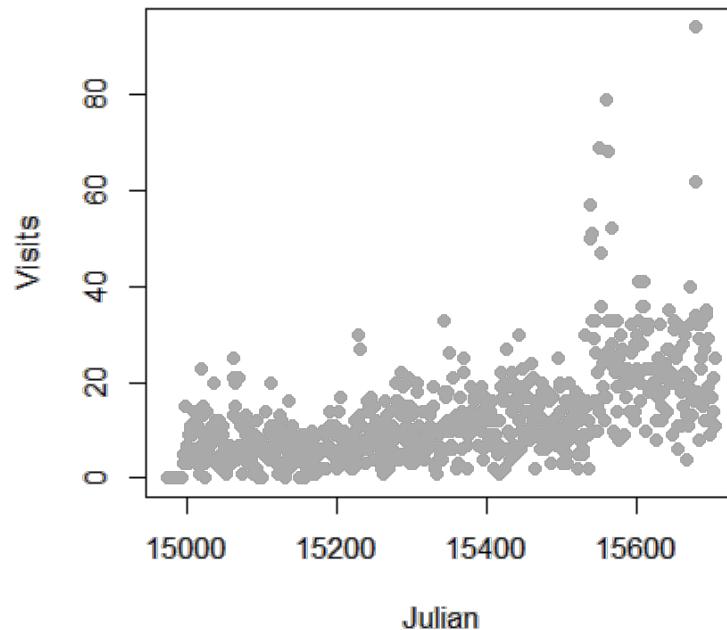
```
download.file("https://dl.dropboxusercontent.com/u/7710864/data/gaData.rda", destfile = "./data/gaData.rda")
load("./data/gaData.rda")
gaData$ julian <- julian(gaData$date)
head(gaData)
```

	date	visits	simplystats	julian
1	2011-01-01	0	0	14975
2	2011-01-02	0	0	14976
3	2011-01-03	0	0	14977
4	2011-01-04	0	0	14978
5	2011-01-05	0	0	14979
6	2011-01-06	0	0	14980

<http://skardhamar.github.com/rga/>

# Plot data

```
plot(gaData$julian,gaData$visits,pch=19,col="darkgrey",xlab="Julian",ylab="Visits")
```



# Linear regression

$$NH_i = b_0 + b_1 JD_i + e_i$$

$NH_i$  - number of hits to the website

$JD_i$  - day of the year (Julian day)

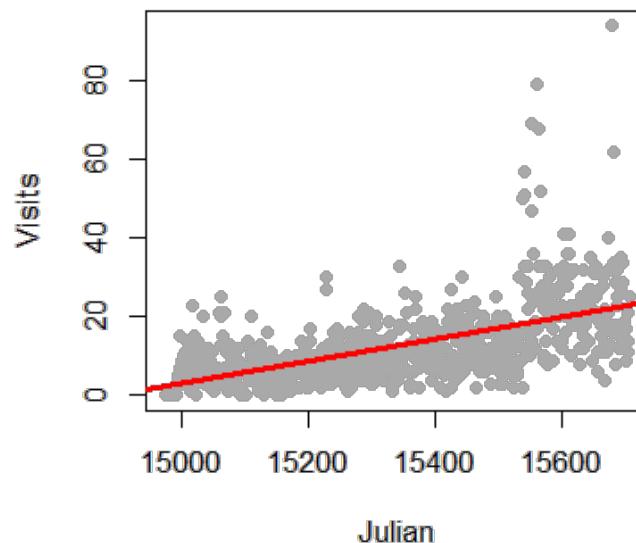
$b_0$  - number of hits on Julian day 0 (1970-01-01)

$b_1$  - increase in number of hits per unit day

$e_i$  - variation due to everything we didn't measure

# Linear regression line

```
plot(gaData$julian,gaData$visits,pch=19,col="darkgrey",xlab="Julian",ylab="Visits")
lm1 <- lm(gaData$visits ~ gaData$julian)
abline(lm1,col="red",lwd=3)
```



# Aside, taking the log of the outcome

- Taking the natural log of the outcome has a specific interpretation.
- Consider the model

$$\log(NH_i) = b_0 + b_1 JD_i + e_i$$

$NH_i$  - number of hits to the website

$JD_i$  - day of the year (Julian day)

$b_0$  - log number of hits on Julian day 0 (1970-01-01)

$b_1$  - increase in log number of hits per unit day

$e_i$  - variation due to everything we didn't measure

# Exponentiating coefficients

- $e^{E[\log(Y)]}$  geometric mean of Y.
  - With no covariates, this is estimated by  $e^{\frac{1}{n} \sum_{i=1}^n \log(y_i)} = (\prod_{i=1}^n y_i)^{1/n}$
- When you take the natural log of outcomes and fit a regression model, your exponentiated coefficients estimate things about geometric means.
- $e^{\beta_0}$  estimated geometric mean hits on day 0
- $e^{\beta_1}$  estimated relative increase or decrease in geometric mean hits per day
- There's a problem with logs when you have zero counts, adding a constant works

```
round(exp(coef(lm(I(log(gaData$visits + 1)) ~ gaData$julian))), 5)
```

(Intercept)	gaData\$julian
0.000	1.002

# Linear vs. Poisson regression

Linear

$$NH_i = b_0 + b_1 JD_i + e_i$$

or

$$E[NH_i | JD_i, b_0, b_1] = b_0 + b_1 JD_i$$

Poisson/log-linear

$$\log(E[NH_i | JD_i, b_0, b_1]) = b_0 + b_1 JD_i$$

or

$$E[NH_i | JD_i, b_0, b_1] = \exp(b_0 + b_1 JD_i)$$

# Multiplicative differences

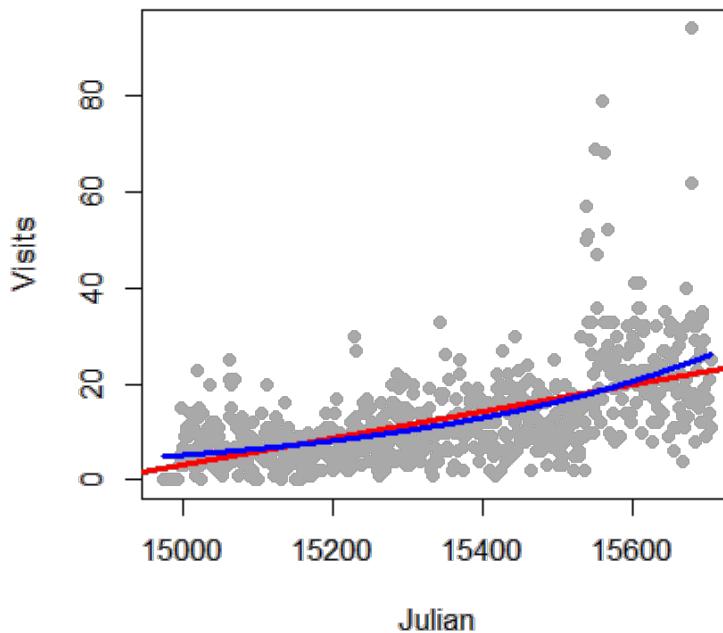
$$E[NH_i|JD_i, b_0, b_1] = \exp(b_0 + b_1 JD_i)$$

$$E[NH_i|JD_i, b_0, b_1] = \exp(b_0) \exp(b_1 JD_i)$$

If  $JD_i$  is increased by one unit,  $E[NH_i|JD_i, b_0, b_1]$  is multiplied by  $\exp(b_1)$

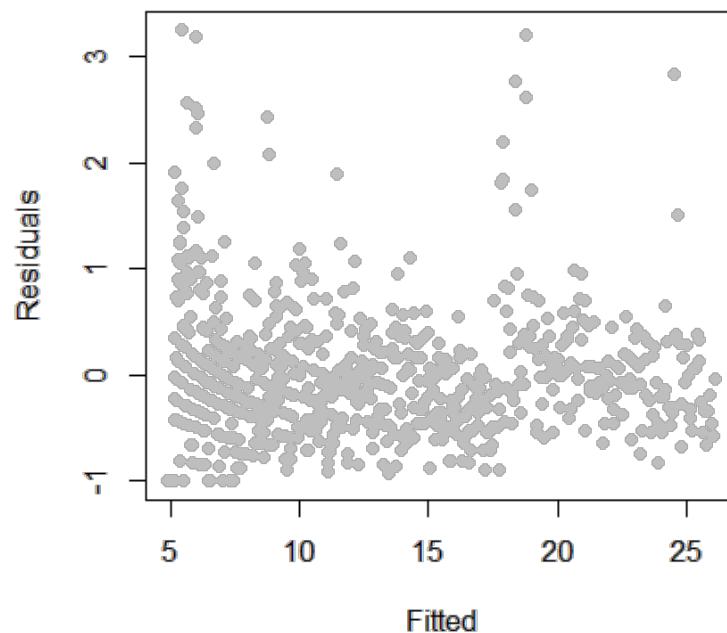
# Poisson regression in R

```
plot(gaData$julian,gaData$visits,pch=19,col="darkgrey",xlab="Julian",ylab="Visits")
glm1 <- glm(gaData$visits ~ gaData$julian,family="poisson")
abline(lm1,col="red",lwd=3); lines(gaData$julian,glm1$fitted,col="blue",lwd=3)
```



# Mean-variance relationship?

```
plot(glm1$fitted,glm1$residuals,pch=19,col="grey",ylab="Residuals",xlab="Fitted")
```



# Model agnostic standard errors

```
library(sandwich)
confint.agnostic <- function (object, parm, level = 0.95, ...)
{
  cf <- coef(object); pnames <- names(cf)
  if (missing(parm))
    parm <- pnames
  else if (is.numeric(parm))
    parm <- pnames[parm]
  a <- (1 - level)/2; a <- c(a, 1 - a)
  pct <- stats:::format.perc(a, 3)
  fac <- qnorm(a)
  ci <- array(NA, dim = c(length(parm), 2L), dimnames = list(parm,
                                                             pct))
  ses <- sqrt(diag(sandwich::vcovHC(object)))[parm]
  ci[] <- cf[parm] + ses %o% fac
  ci
}
```

<http://stackoverflow.com/questions/3817182/vcovhc-and-confidence-interval>

# Estimating confidence intervals

```
confint(glm1)
```

	2.5 %	97.5 %
(Intercept)	-34.34658	-31.159716
gaData\$julian	0.00219	0.002396

```
confint.agnostic(glm1)
```

	2.5 %	97.5 %
(Intercept)	-36.362675	-29.136997
gaData\$julian	0.002058	0.002528

# Rates

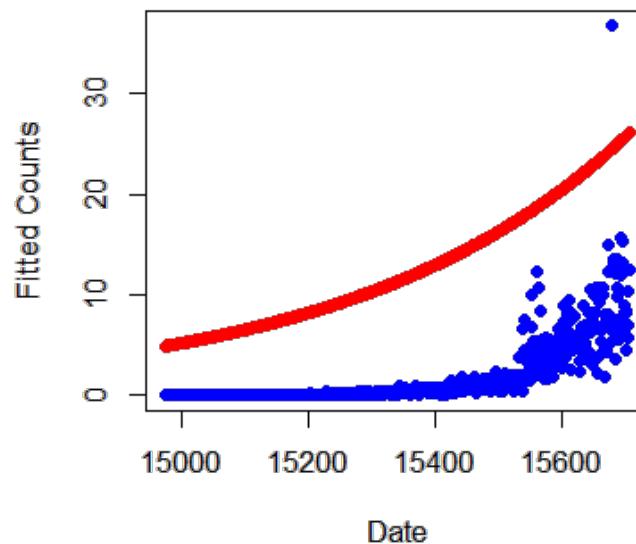
$$E[NHSS_i|JD_i, b_0, b_1]/NH_i = \exp(b_0 + b_1 JD_i)$$

$$\log(E[NHSS_i|JD_i, b_0, b_1]) - \log(NH_i) = b_0 + b_1 JD_i$$

$$\log(E[NHSS_i|JD_i, b_0, b_1]) = \log(NH_i) + b_0 + b_1 JD_i$$

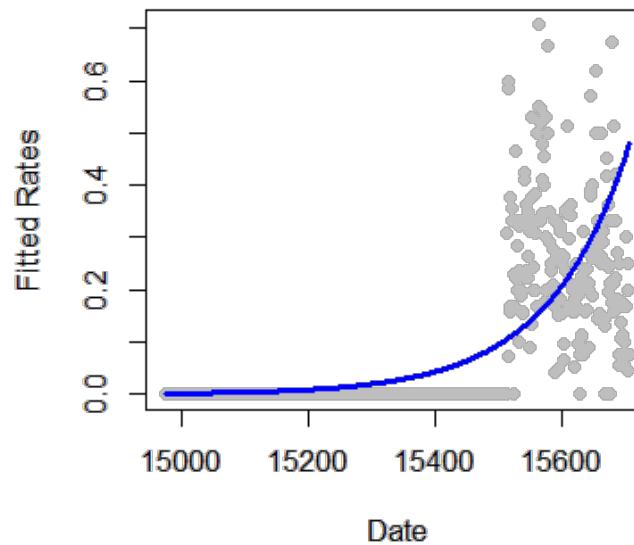
# Fitting rates in R

```
glm2 <- glm(gaData$simplystats ~ julian(gaData$date), offset=log(visits+1),  
            family="poisson", data=gaData)  
plot(julian(gaData$date), glm2$fitted, col="blue", pch=19, xlab="Date", ylab="Fitted Counts")  
points(julian(gaData$date), glm1$fitted, col="red", pch=19)
```



# Fitting rates in R

```
glm2 <- glm(gaData$simplystats ~ julian(gaData$date), offset=log(visits+1),  
            family="poisson", data=gaData)  
plot(julian(gaData$date), gaData$simplystats/(gaData$visits+1), col="grey", xlab="Date",  
     ylab="Fitted Rates", pch=19)  
lines(julian(gaData$date), glm2$fitted/(gaData$visits+1), col="blue", lwd=3)
```



# More information

- [Log-linear models and multiway tables](#)
- [Wikipedia on Poisson regression](#), [Wikipedia on overdispersion](#)
- [Regression models for count data in R](#)
- [pscl package](#) - the function *zeroinfl* fits zero inflated models.



# Hodgepodge

Regression models

Brian Caffo, Jeff Leek, Roger Peng  
Johns Hopkins Bloomberg School of Public Health

# How to fit functions using linear models

- Consider a model  $Y_i = f(X_i) + \epsilon$ .
- How can we fit such a model using linear models (called scatterplot smoothing)
- Consider the model

$$Y_i = \beta_0 + \beta_1 X_i + \sum_{k=1}^d (x_i - \xi_k)_+ \gamma_k + \epsilon_i$$

where  $(a)_+ = a$  if  $a > 0$  and 0 otherwise and  $\xi_1 \le \dots \le \xi_d$  are known knot points.

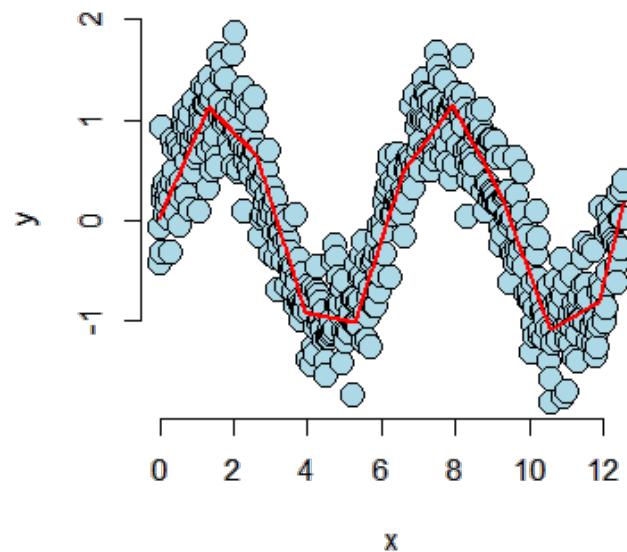
- Prove to yourself that the mean function

$$\beta_0 + \beta_1 X_i + \sum_{k=1}^d (x_i - \xi_k)_+ \gamma_k$$

is continuous at the knot points.

# Simulated example

```
n <- 500; x <- seq(0, 4 * pi, length = n); y <- sin(x) + rnorm(n, sd = .3)
knots <- seq(0, 8 * pi, length = 20);
splineTerms <- sapply(knots, function(knot) (x > knot) * (x - knot))
xMat <- cbind(1, x, splineTerms)
yhat <- predict(lm(y ~ xMat - 1))
plot(x, y, frame = FALSE, pch = 21, bg = "lightblue", cex = 2)
lines(x, yhat, col = "red", lwd = 2)
```

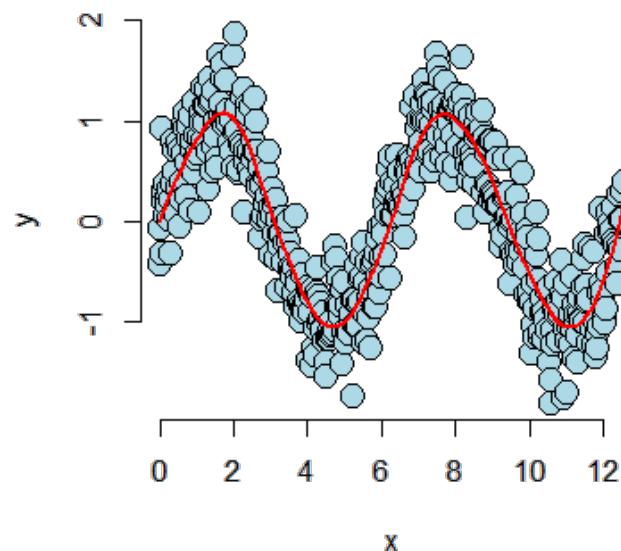


# Adding squared terms

- Adding squared terms makes it continuously differentiable at the knot points.
- Adding cubic terms makes it twice continuously differentiable at the knot points; etcetera.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \sum_{k=1}^d (x_i - \xi_k)_+^2 \gamma_k + \epsilon_i$$

```
splineTerms <- sapply(knots, function(knot) (x > knot) * (x - knot)^2)
xMat <- cbind(1, x, x^2, splineTerms)
yhat <- predict(lm(y ~ xMat - 1))
plot(x, y, frame = FALSE, pch = 21, bg = "lightblue", cex = 2)
lines(x, yhat, col = "red", lwd = 2)
```

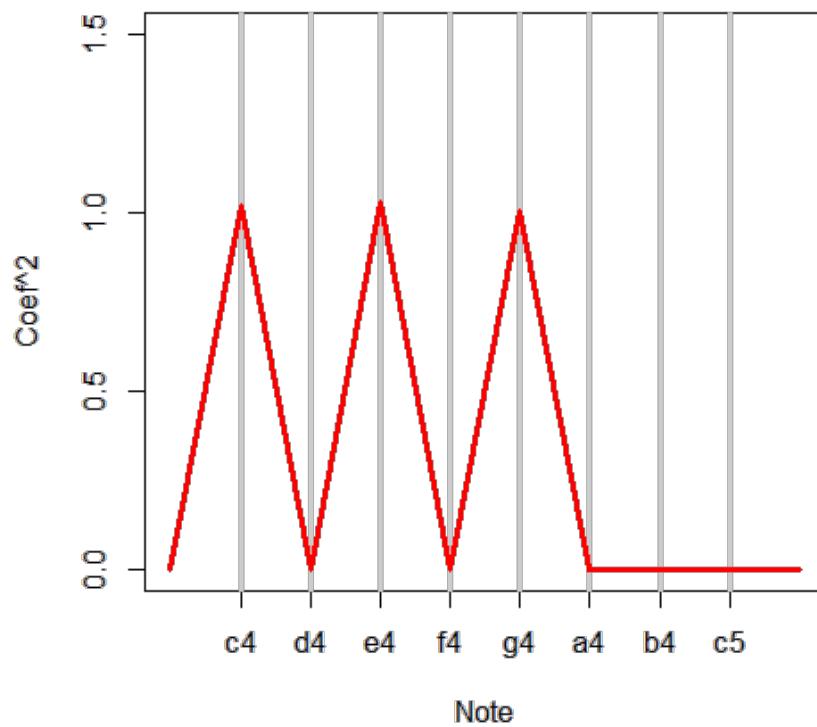


# Notes

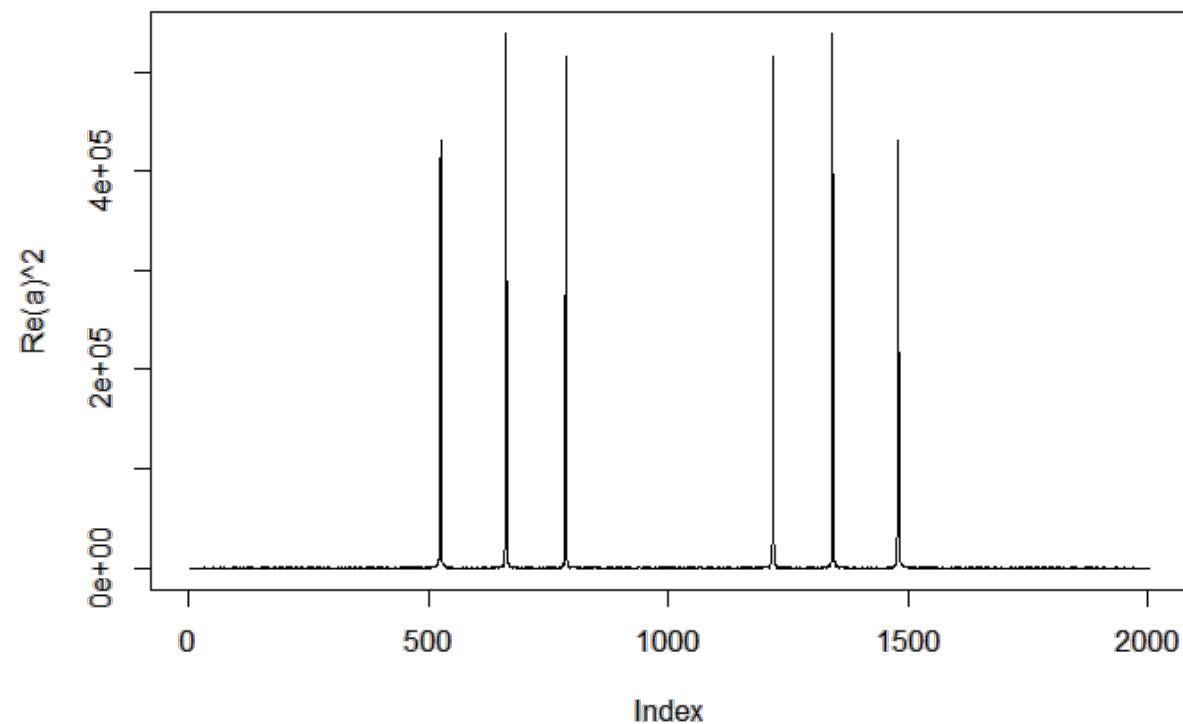
- The collection of regressors is called a basis.
  - People have spent **a lot** of time thinking about bases for this kind of problem. So, consider this as just a teaser.
- Single knot point terms can fit hockey stick like processes.
- These bases can be used in GLMs as well.
- An issue with these approaches is the large number of parameters introduced.
  - Requires some method of so called regularization.

# Harmonics using linear models

```
##Chord finder, playing the white keys on a piano from octave c4 - c5
notes4 <- c(261.63, 293.66, 329.63, 349.23, 392.00, 440.00, 493.88, 523.25)
t <- seq(0, 2, by = .001); n <- length(t)
c4 <- sin(2 * pi * notes4[1] * t); e4 <- sin(2 * pi * notes4[3] * t);
g4 <- sin(2 * pi * notes4[5] * t)
chord <- c4 + e4 + g4 + rnorm(n, 0, 0.3)
x <- sapply(notes4, function(freq) sin(2 * pi * freq * t))
fit <- lm(chord ~ x - 1)
```



```
##(How you would really do it)
a <- fft(chord); plot(Re(a)^2, type = "l")
```



# **Regression III: Advanced Methods**

William G. Jacoby  
*Michigan State University*

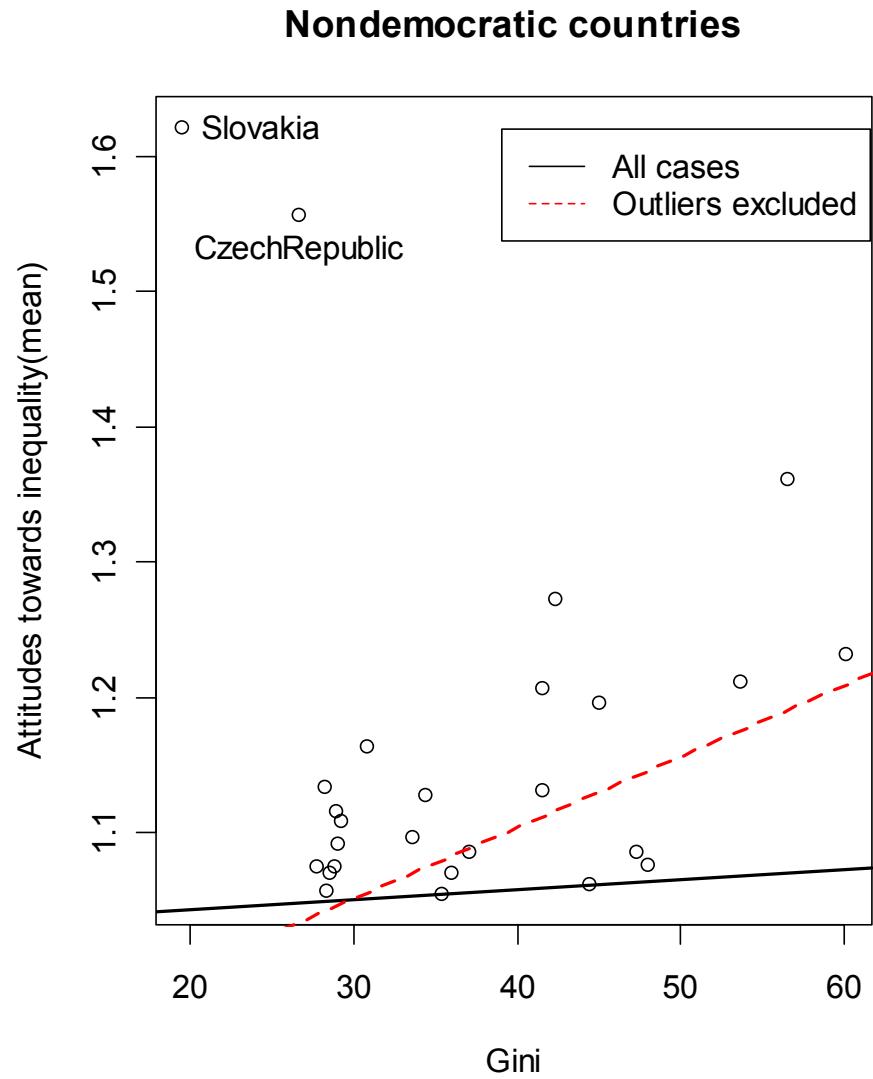
<http://polisci.msu.edu/jacoby/icpsr/regress3>

# Outlying Observations: Why pay attention?

- Can cause us to misinterpret patterns in plots
  - Outliers can affect visual resolution of remaining data in plots (forces observations into “clusters”)
  - Temporary removal of outliers, and/or transformations can “spread out” clustered observations and bring in the outliers (if not removed)
- More importantly, separated points can have a strong *influence* on statistical models—deleting outliers from a regression model can sometimes give completely different results
  - Unusual cases can substantially influence the fit of the OLS model—**Cases that are both outliers and high leverage exert influence on both the slopes and intercept of the model**
  - Outliers may also indicate that our model fails to capture important characteristics of the data

# Ex 1. Influence and Small Samples: Inequality Data (1)

- Small samples are especially vulnerable to outliers—there are fewer cases to counter the outlier
- With Czech Republic and Slovakia included, there is no relationship between Attitudes towards inequality and the Gini coefficient
- If these cases are removed, we see a positive relationship



# Ex 1. Influence and Small Samples: Inequality Data (2)

## Model including all cases

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.0283	0.1278	8.05	0.0000
gini	0.0007	0.0028	0.27	0.7908
gdp	0.0000	0.0000	2.19	0.0387

Residual standard error: 0.138

Multiple R-Squared: 0.175

## Model excluding Czech Rep. & Slovakia

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.8931	0.0578	15.45	0.0000
gini	0.0053	0.0013	4.07	0.0005
gdp	0.0000	0.0000	1.69	0.1050

Residual standard error: 0.0602

Multiple R-Squared: 0.462

# R script for Ex. 1

```
Weakliem2<-read.table('C:/data/Weakliem2.txt', header=T)
attach(Weakliem2)

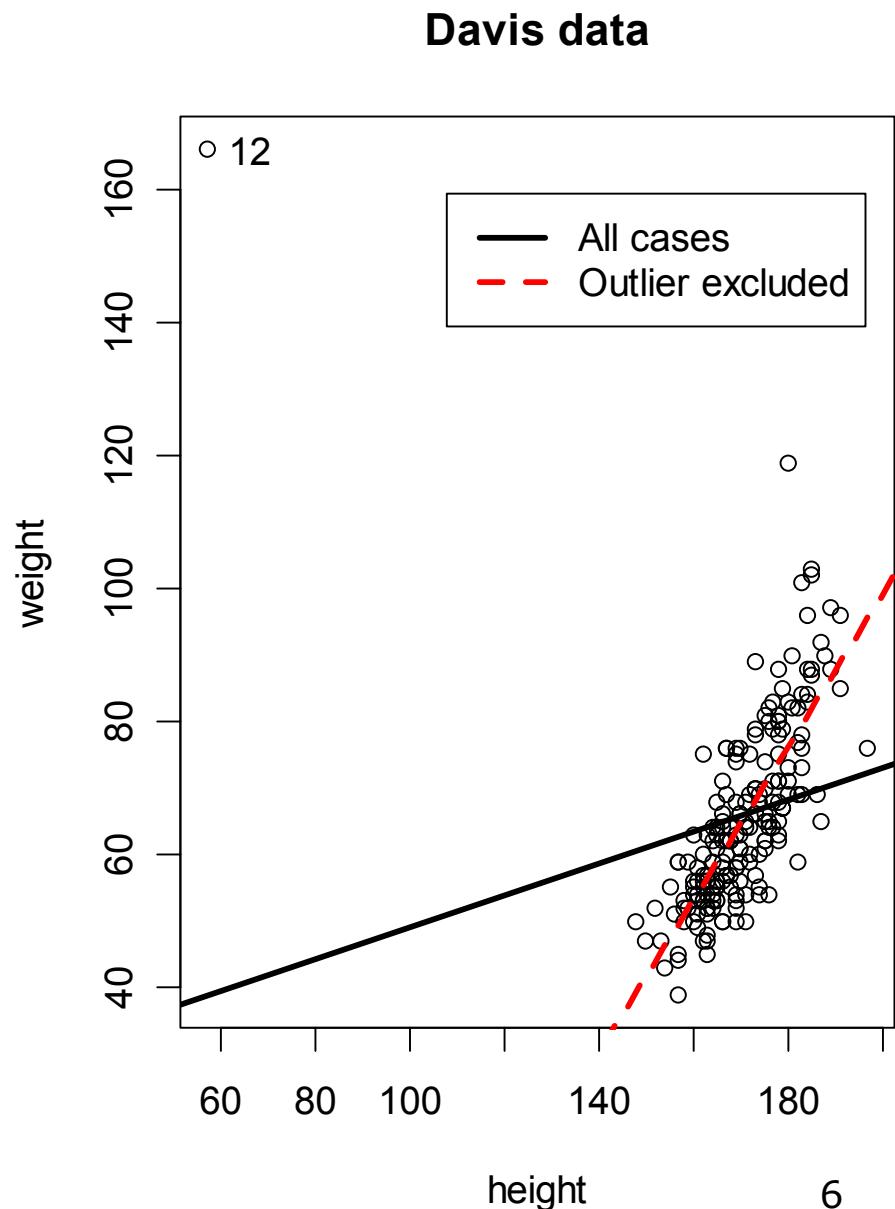
plot(gini, secpay, main='Nondemocratic countries', xlab='Gini',
ylab='Attitudes towards inequality(mean)')

Weakliem.model1<-lm(secpay~gini+gdp)
abline(Weakliem.model1, lwd=2, lty=1, col=1)
identify(gini,secpay, row.names(Weakliem2))
#"identify" returns cases 7, 26 as outliers
Weakliem.model2<-update(Weakliem.model1, subset=-c(7,26))
abline(Weakliem.model2, lwd=2, lty=2, col=2)
legend(locator(1), lty=1:2, col=1:2,
       legend=c('All cases', 'Outliers excluded'))

library(xtable)#Prints LaTeX code for the output table
print(xtable(Weakliem.model1))
print(xtable(Weakliem.model2))
```

## Ex 2. Influence and Small Samples: Davis Data (1)

- These data are the Davis data in the `car` package
- It is clear that observation 12 is ***influential***
- The model including observation 12 does a poor job of representing the trend in the data; The model excluding observation 12 does much better
- The output on the next slide confirms this



## Ex 2. Influence and Small Samples: Davis Data (2)

### Model including all cases

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	25.2662	14.9504	1.69	0.0926
height	0.2384	0.0877	2.72	0.0072

Residual standard error: 14.86

Multiple R-Squared: 0.0359

### Model excluding observation #12

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-130.7470	11.5627	-11.31	0.0000
height	1.1492	0.0677	16.98	0.0000

Residual standard error: 8.523

Multiple R-Squared: 0.594

## R script for Ex. 2

```
>library(car)
>data(Davis)
>attach(Davis)
>davis.model.1<-lm(repwt~weight)

>plot(height, weight, main="Davis data")
>Model1<-lm(weight~height)
>identify(height, weight, row.names(Davis))
    #observation 12 returned as outlier
>abline(Model1, lty=1, col=1, lwd=3)
>Model2<-update(Model1, subset=-12)
>abline(Model2, lty=2, col=2, lwd=3)
>legend(locator(1), lty=1:2, col=1:2, lwd=3,
  legend=c('All cases', 'Outlier excluded'))
```

# Types of Unusual Observations (1)

## 1. Regression Outliers

- An observation that is unconditionally unusual in either its Y or X value is called a ***univariate outlier***, but it is not necessarily a regression outlier
- ***A regression outlier is an observation that has an unusual value of the dependent variable Y, conditional on its value of the independent variable X***
  - In other words, for a regression outlier, neither the X nor the Y value is necessarily unusual on its own
- A regression outlier will have a large residual but not necessarily affect the regression slope coefficient

# Types of Unusual Observations (2)

## 2. Cases with Leverage

- An observation that has an unusual X value—i.e., it is far from the mean of X—has *leverage* on (i.e., the potential to influence) the regression line
- The further away from the mean of X (either in a positive or negative direction), the more leverage an observation has on the regression fit
- High leverage does not necessarily mean that it influences the regression coefficients
  - It is possible to have a high leverage and yet follow straight in line with the pattern of the rest of the data

# Types of Unusual Observations (3)

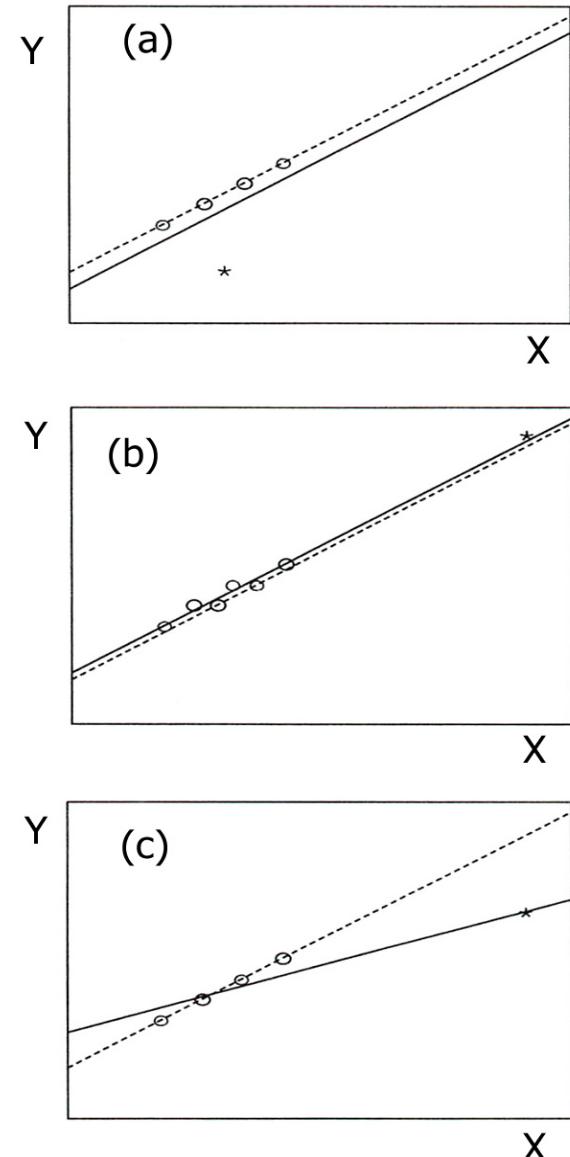
## 3. *Influential Observations*

- Only when an observation has ***high leverage*** and is an ***outlier in terms of Y-value*** will it strongly influence the regression line
  - In other words, it must have an unusual  $X$ -value with an unusual  $Y$ -value *given its  $X$ -value*
- In such cases both the intercept and slope are affected, as ***the line chases the observation***

***Influence=Leverage X Discrepancy***

# Types of Unusual Observations (4)

- **Figure (a): Outlier without influence.** Although its Y value is unusual given its X value, it has little influence on the regression line because it is in the middle of the X-range
- **Figure (b) High leverage** because it has a high value of X. However, because its value of Y puts it in line with the general pattern of the data it has **no influence**
- **Figure (c): Combination of discrepancy (unusual Y value) and leverage (unusual X value)** results in strong influence. When this case is deleted both the slope and intercept change dramatically.



Adapted from Figure 11.1 (Fox, 1997)

# Assessing Leverage: Hat Values (1)

- Most common measure of leverage is the **hat-value**,  $h_i$ ,
- The name hat-values results from their calculation based on the fitted values (Y-hat):

$$\begin{aligned}\hat{Y}_i &= h_{1j}Y_1 + h_{2j}Y_2 + \cdots + h_{nj}Y_n \\ &= \sum_{i=1}^n h_{ij}Y_i\end{aligned}$$

- Recall that the **Hat Matrix**,  $\mathbf{H}$ , projects the Y's onto their predicted values:

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{X}\mathbf{b} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \mathbf{H}\mathbf{y}\end{aligned}$$

$$\mathbf{H}_{(n \times n)} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

# Assessing Leverage: Hat Values (2)

- If  $h_{ij}$  is large, the  $i$ th observation has a substantial impact on the  $j$ th fitted value
- Since  $\mathbf{H}$  is symmetric and idempotent, the diagonal entries represent both the  $i_{th}$  row and the  $i_{th}$  column:

$$\begin{aligned} h_i &= \mathbf{h}_i' \mathbf{h}_i \\ &= \sum_{j=1}^n h_{ij}^2 \end{aligned}$$

- This implies, then, that  $h_i = h_{ii}$
- As a result, the hat value  $h_i$  measures the ***potential leverage of  $Y_i$  on all the fitted values***

# Properties of Hat-Values

- The average hat-value is:  $\bar{h} = (k + 1)/n$
- Hat values are bounded between  $1/n$  and 1
- In simple regression hat values measure distance from the mean of X:

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^n (X_j - \bar{X})^2}$$

- In multiple regression,  $h_i$  measures the distance from the centroid point of X's (point of means)
- **Rule of Thumb:**
  - ***Hat values exceeding about twice the average hat-value should be considered noteworthy***
  - With large sample sizes, however, this cut-off is unlikely to identify any observations regardless of whether they deserve attention

# Hat Values in Multiple Regression

- The diagram to the right shows elliptical contours of hat values for two independent variables
- As the contours suggest, hat values in multiple regression take into consideration the *correlational* and *variational* structure of the X's
- As a result, outliers in multi-dimensional X-space are high leverage observations—*i.e.*, the **dependent variable values are irrelevant in calculating  $h_i$**

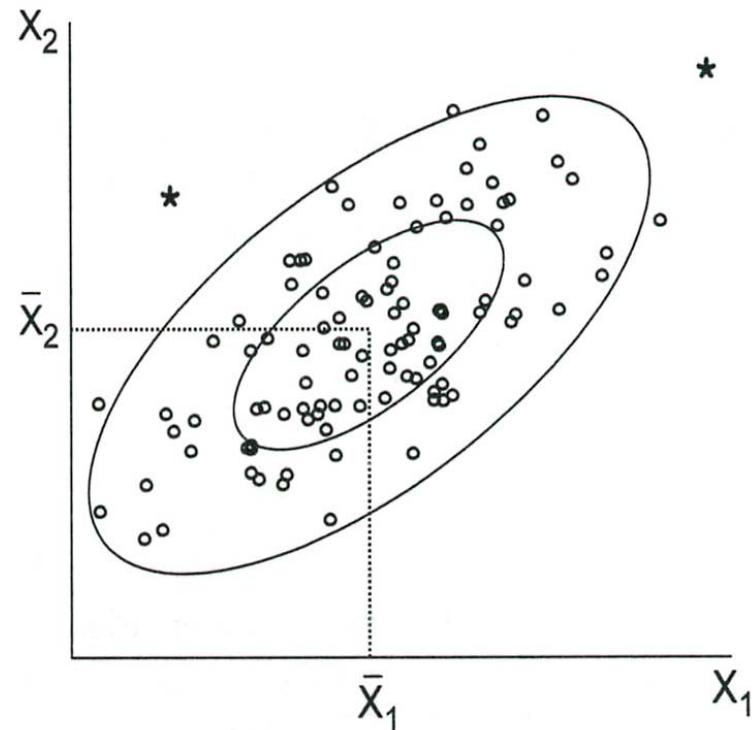


Figure 11.3 from Fox (1997)

# Leverage and Hat Values: Inequality data revisited (1)

- We start by fitting the model to the complete dataset
- Recall that, looking at the scatterplot of Gini and attitudes, we identified two possible outliers (Czech Republic and Slovakia)
- With these included in the model there was no apparent effect of Gini on attitudes:

Model including all cases

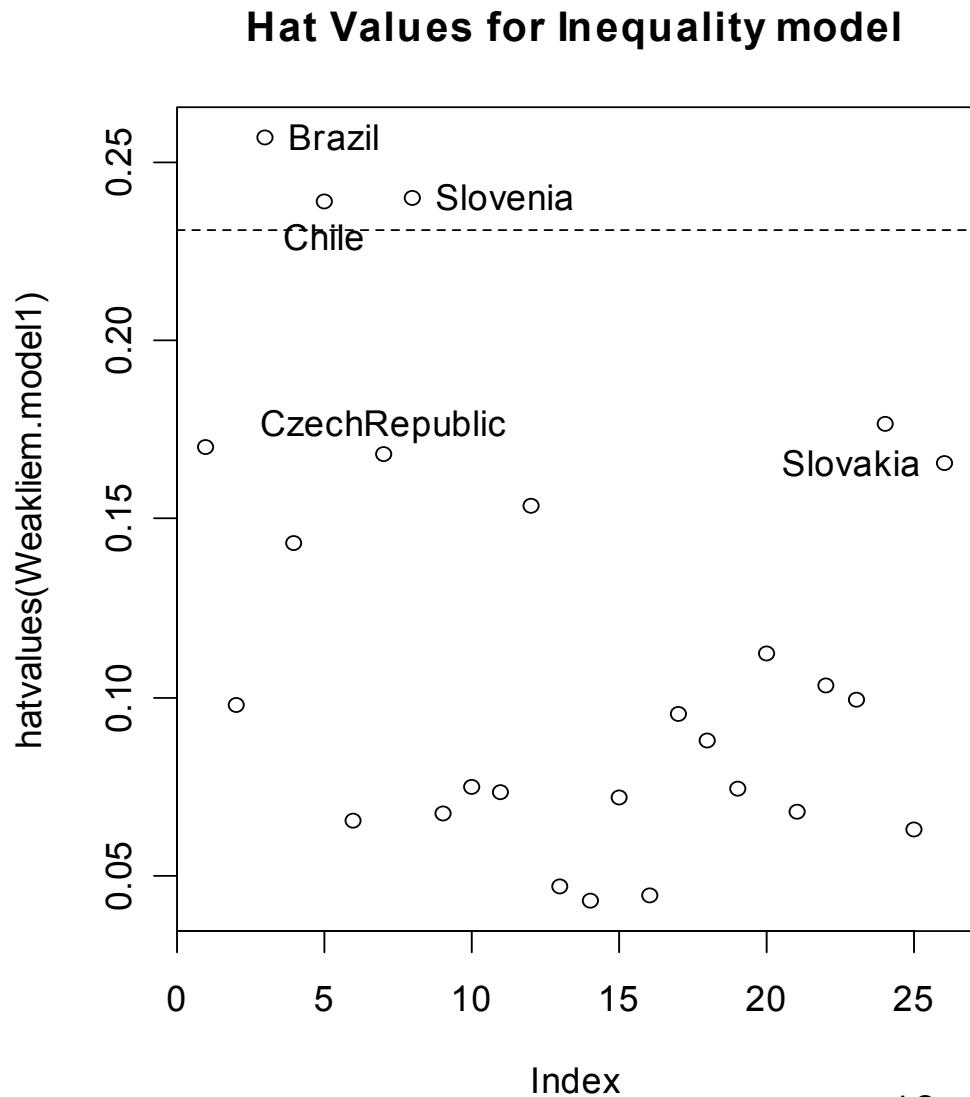
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.0283	0.1278	8.05	0.0000
gini	0.0007	0.0028	0.27	0.7908
gdp	0.0000	0.0000	2.19	0.0387

# R script for plot of Hat Values

```
>library(car)
>plot(hatvalues(Weakliem.model1) ,
      main="Hat Values for Inequality model")
>abline(h=c(2,3)*3/length(secpay), lty=2)
#"h" signifies horizontal line
#the average hat value=(k+1)/n.
#A rule of thumb is that 2*average hat value
#for large samples, and 3*average hat value
#for small samples should be examined
>identify(1:length(secpay) ,
           hatvalues(Weakliem.model1) ,
           row.names(Weakliem2))
```

# Leverage and Hat Values: Inequality data revisited (2)

- Several countries have large hat values, suggesting that they have unusual X values
- Notice that there are several that have much higher hat values than the Czech Republic and Slovakia
- These cases have ***high leverage, but not necessarily high influence***



# Formal Tests for Outliers: Standardized Residuals

- Unusual observations typically have large residuals but not necessarily so—***high leverage observations can have small residuals because they pull the line towards them:***

$$V(E_i) = \sigma_{\varepsilon}^2(1 - h_i)$$

- Standardized residuals provide one possible, though unsatisfactory, way of detecting outliers:

$$E'_i = \frac{E_i}{S_E \sqrt{1 - h_i}}$$

- The numerator and denominator are not independent and thus  $E'_i$  does not follow a t-distribution: If  $|E_i|$  is large, the standard error is also large:

$$S_E = \sqrt{\sum E_i^2 / (n - k - 1)}$$

# Studentized Residuals (1)

- If we refit the model deleting the  $i$ th observation we obtain an estimate of the standard deviation of the residuals  $S_{E(-1)}$  (standard error of the regression) that is based on the  $n-1$  observations
- We then calculate the *studentized residuals*  $E_i^*$ 's, which have an independent numerator and denominator:

$$E_i^* = \frac{E_i}{S_{E(-i)} \sqrt{1 - h_i}}$$

- Studentized residuals follow a t-distribution with  $n-k-2$  degrees of freedom
- We might employ this method when we have several cases that might be outliers
- Observations that have a studentized residual outside the  $\pm 2$  range are considered statistically significant at the 95%  $\alpha$  level

# Studentized Residuals (2)

- An alternative, but equivalent, method of calculating studentized residuals is the so-called ‘mean-shift’ outlier model:

$$Y = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k + \gamma D + \varepsilon$$

Here D is a dummy regressor coded 1 for observation  $i$  and 0 for all other observations.

- We test the null hypothesis that the outlier  $i$  does not differ from the rest of the observations,  $H_0: \gamma=0$ , by calculating the  $t$ -test:

$$t_0 = \frac{\tilde{\gamma}}{\widehat{SE}(\tilde{\gamma})}$$

- The test statistic is the studentized residual  $E_i^*$  and is distributed as  $t_{n-k-2}$
- This method is most suitable when, after looking at the data, we have determined that a particular case might be an outlier

# Studentized Residuals (3)

## The Bonferroni adjustment

- Since we are selecting the furthest outlier, it is not legitimate to use a simple t-test
  - We would expect that 5% of the studentized residuals would be beyond  $t_{.025} \pm 2$  by chance alone
- To remedy this we can make a **Bonferroni adjustment** to the  $p$ -value.
  - The Bonferroni  $p$ -value for the largest outlier is:  
 $p=2np$  where  $p$  is the unadjusted  $p$ -value from a  $t$ -test with  $n-k-2$  degrees of freedom
- A special  $t$ -table is needed if you do this calculation by hand, but the `outlier.test` function in the `car` package for **R** will give it to you automatically

# Studentized Residuals (4)

## An Example of the Outlier Test

- The Bonferroni-adjusted outlier test in `car` tests the ***largest absolute studentized residual***.
- Recalling our *inequality model*:

```
> outlier.test(Weakliem.model1)
max|rstudent| df unadjusted p Bonferroni p
4.317504 22 0.0002778084 0.007223019
```

Observation: 26

```
> row.names(Weakliem2)[26]
[1] "Slovakia"
```

- It is now quite clear that Slovakia (observation 26) is an outlier, but as of yet we have not assessed whether it influences the regression line

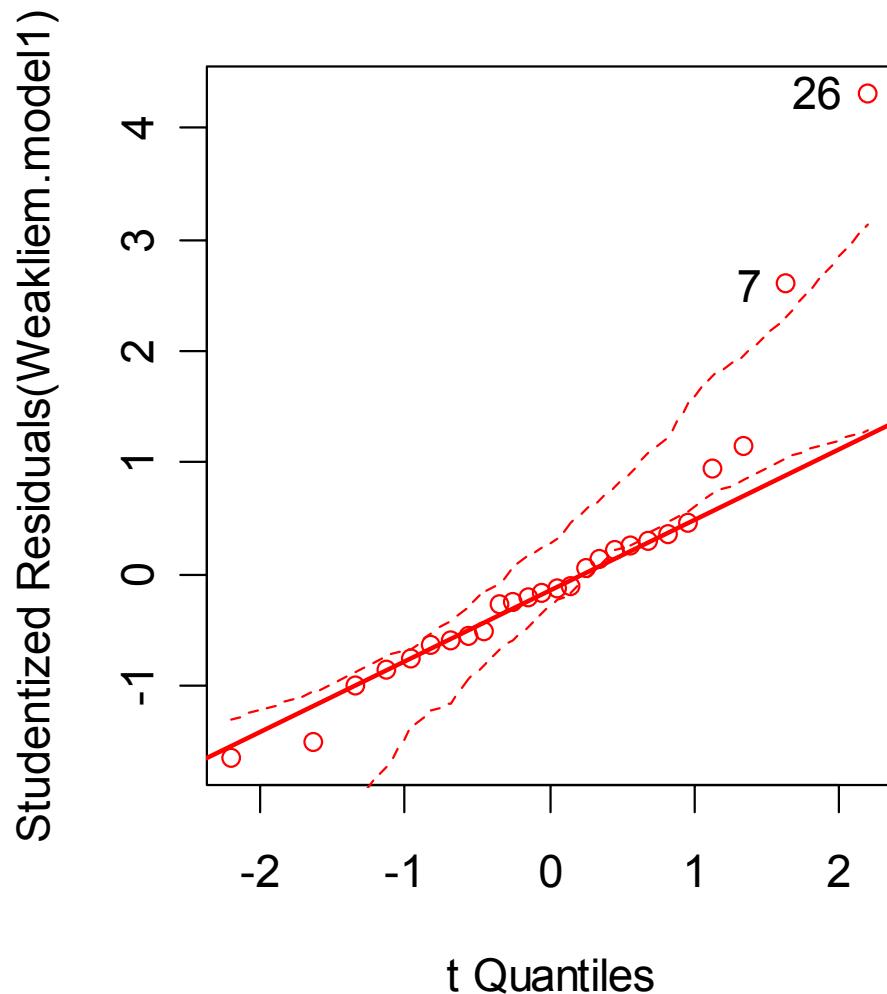
# Quantile Comparison Plots (1)

- Recall that we used quantile comparison plots to compare the distribution of a single variable to the t-distribution, assessing whether the distribution of the variable showed a departure from normality
- Using the same technique, we can compare the distribution of the studentized residuals from our regression model to the t-distribution
- Observations that stray outside of the 95% confidence envelope are statistically significant outliers

```
>library(car)  
>qq.plot(Weakliem.model1, simulate=T,  
         labels=row.names(Weakliem2))  
          #simulate=T specifies a bootstrap  
          # 95% confidence envelope
```

# Quantile-Comparison Plots (2): Example: Inequality data

- Here we can again see that two cases appear to be outliers: 7 and 26, which represent the Czech Republic and Slovakia



# Influential Observations: DFBeta and DFBetas (1)

- Recall that ***an influential observation is one that combines discrepancy with leverage***
- Therefore, examine how regression coefficients change if outliers are omitted from the model
- We can use  $D_{ij}$  (often termed ***DFBeta<sub>ij</sub>***) to do so:

$$D_{ij} = B_j - B_{j(-i)}$$

for  $i = 1, \dots, n$  and  $j = 0, 1, \dots, k$

where the  $B_j$  are for all the data and the  $B_{j(-i)}$  are with the  $i$ th observation removed

- $D^*_{ij}$  (Dfbetas<sub>ij</sub>) standardizes the measure, by dividing by  $S_{Bj(-i)}$
- A standard cut-off for an influential observation is:

$$D^*_{ij} = 2 n^{-0.5}$$

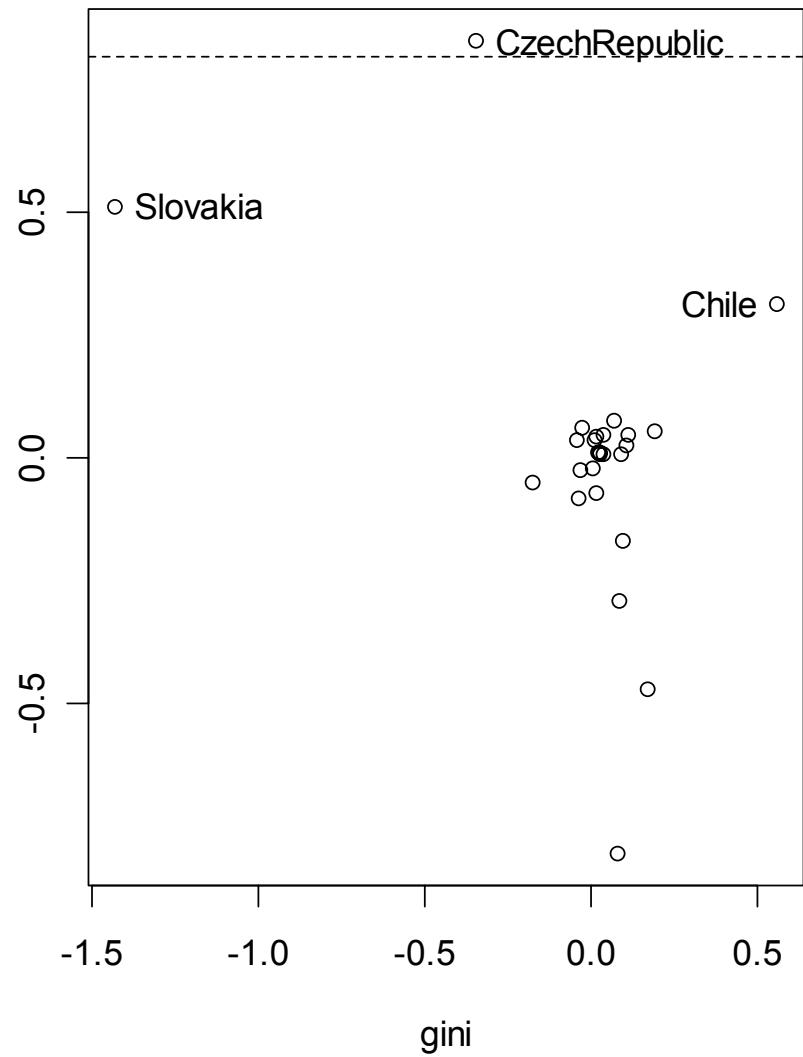
# R script for DFBetas plot

```
>library(car)
>Weakliem.dfbetas<-dfbetas(Weakliem.model1)
>plot(Weakliem.dfbetas[,c(2,3)],
      main="DFBetas for the Gini and GDP coefficients")
      #c(2,3) specifies the coefficients of interest
>abline(h=2/sqrt(length(Weakliem2)), lty=2)
      #adds the rule of thumb cut-off line
>identify(Weakliem.dfbetas[,2], Weakliem.dfbetas[,3],
           row.names(Weakliem2))
```

# Influential Observations: DFBetas (2)

- We see here Slovakia makes the ***gdp coefficient larger*** and the ***coefficient for gini smaller***
- The Czech Republic also makes the ***coefficient for gdp larger***
- A problem with ***DFBetas*** is that each observation has several measures of influence—one for each coefficient  $n(k+1)$  different measures
- ***Cook's D*** overcomes the problem by presenting a single summary measure for each observation

DFBetas for the Gini and GDP coefficients



# Cook's Distance (Cook's D)

- Cook's D measures the 'distance' between  $B_j$  and  $B_{j(-i)}$  by calculating an F-test for the hypothesis that  $\beta_j = B_{j(-i)}$ , for  $j=0,1,\dots,k$ . An F statistic is calculated for each observation as follows:

$$D_i = \frac{E_i'^2}{k+1} \times \frac{h_i}{1-h_i}$$

where  $h_i$  is the hat-value for each observation and  $E_i'$  is the standardized residual

- The first fraction **measures discrepancy**; the second fraction **measures leverage**
- There is **no significance test** for  $D_i$  (i.e., the F value here measures only distance) but a cut-off rule of thumb is:

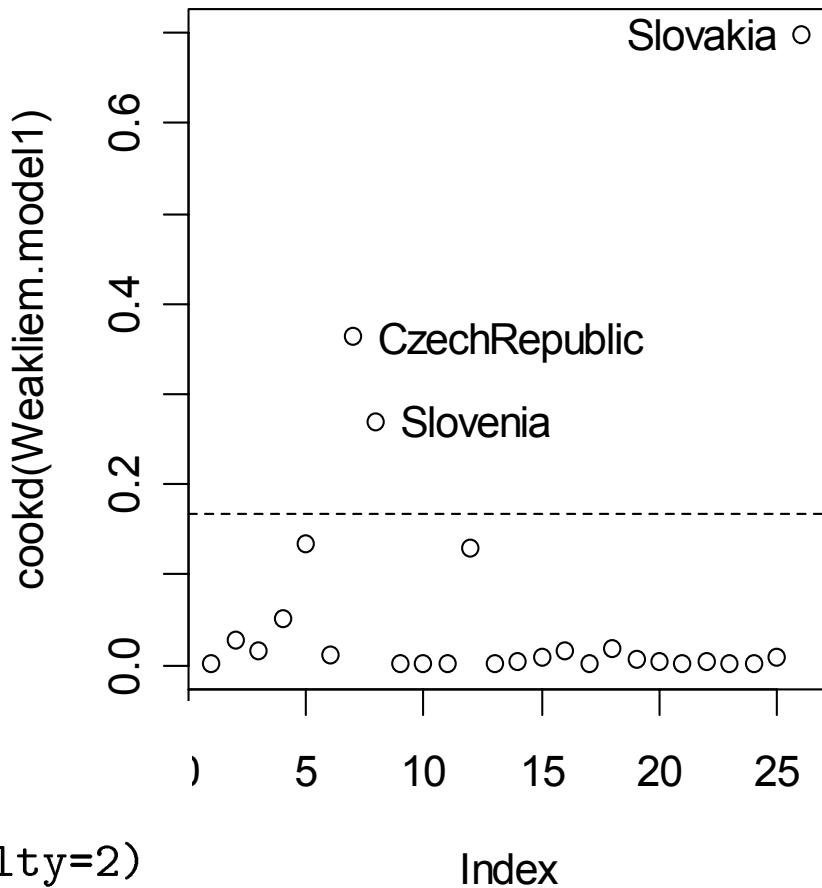
$$D_i > \frac{4}{n - k - 1}$$

- The cut-off is useful, but there is no substitute for examining **relative discrepancies** in plots of Cook's D versus cases, or of  $E_i^*$  against  $h_i$

# Cook's D: An Example

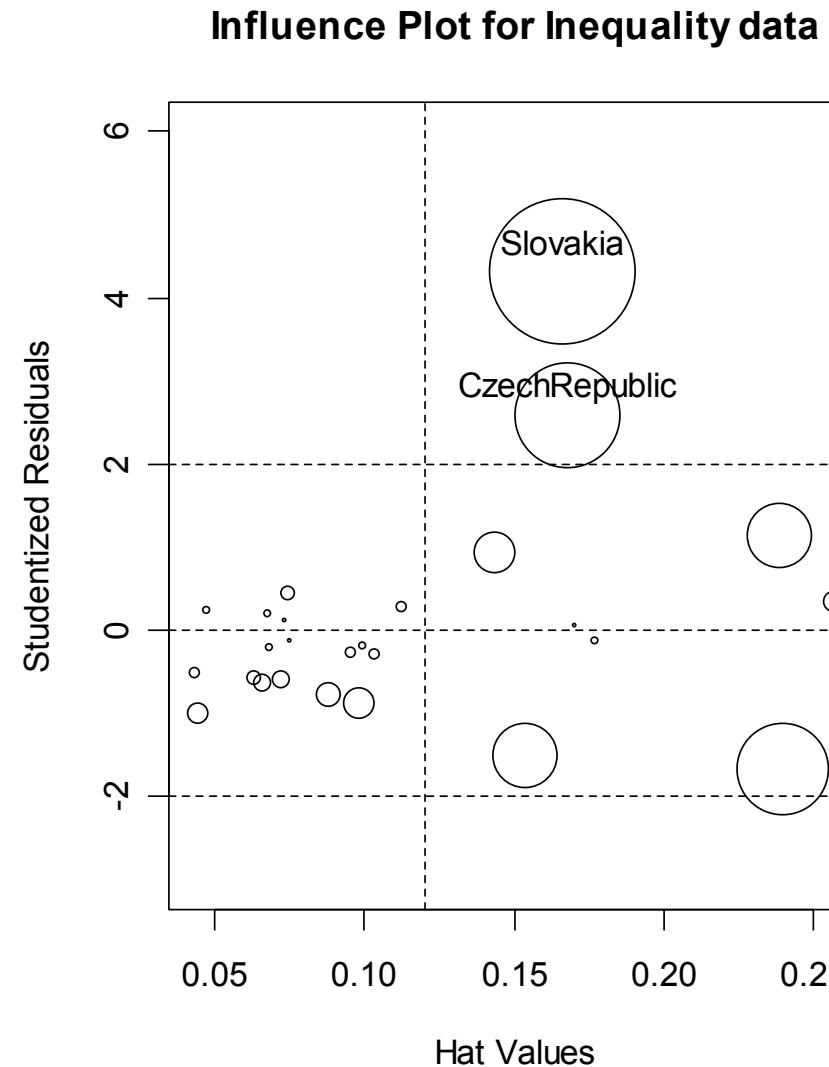
- We can see from this plot of Cook's D against the case numbers, that Slovakia has an unusually high level of influence on the regression surface
- The Czech Republic and Slovenia also stand out

```
> library(car)
> plot(cookd(Weakliem.model1))
> abline(h=4/length(Weakliem2), lty=2)
> identify(1:26, cookd(Weakliem.model1),
           row.names(Weakliem2))
[1] 7 8 26
```



# Influence Plot (or “bubble plot”)

- Displays ***studentized residuals, hat-values*** and ***Cook's D*** on a single plot
- The horizontal axis represents the *hat-values*; the vertical axis represents the *studentized residuals*; circles for each observation represent the relative size of the Cook's D
  - The *radius* is proportional to the square root of Cook's D, and thus ***the areas are proportional to the Cook's D***



# R-script for the Influence Plot

```
plot(hatvalues(Weakliem.model1),  
      rstudent(Weakliem.model1), ylim=c(-3,6), type='n',  
      main="Influence Plot for Inequality data",  
      xlab="Hat Values",  
      ylab="Studentized Residuals")  
cook<-sqrt(cookd(Weakliem.model1))  
points(hatvalues(Weakliem.model1),  
       rstudent(Weakliem.model1), cex=10*cook/max(cook))  
abline(v=3/25, lty=2)#line for hatvalues  
abline(h=c(-2,0,2), lty=2)  
#lines for studentized residuals  
identify(hatvalues(Weakliem.model1),  
         rstudent(Weakliem.model1), row.names(Weakliem2))
```

# Joint Influence (1)

- Subsets of cases can jointly influence a regression line, or can offset each other's influence
- Cook's D can help us determine joint influence if there are relatively few influential cases.
  - That is, we can delete cases sequentially, updating the model each time and exploring the Cook's Ds again
  - This approach is impractical if there are potentially a large number of subsets to explore, however
- ***Added-variable plots*** (also called ***partial-regression plots***) provide a more useful method of assessing joint influence

## Joint influence (2)

- The heavy solid represent the regression with all cases included; The broken line is the regression with the asterisk deleted;The light solid line is for the regression with both the plus and asterisk deleted
- Depending on where the jointly influential cases lie, they can have different effects on the regression line.
- (a) and (b) are jointly influential because they change the regression line when included together.
- The observations in (c) offset each other and thus have little effect on the regression line

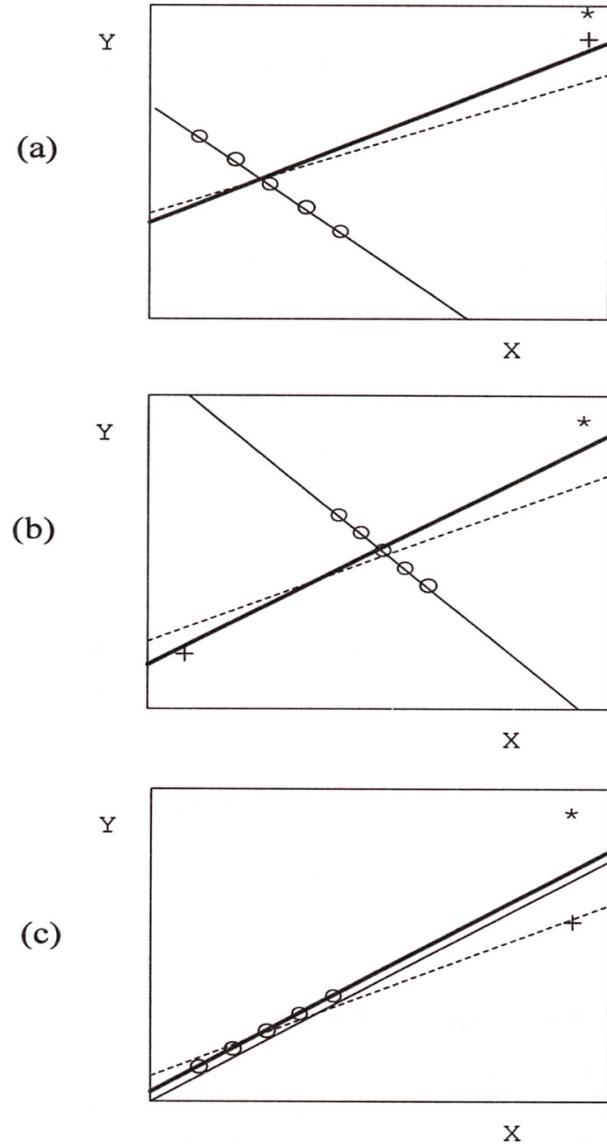


Figure 11.4 from Fox (1997)

# Added-Variable Plots (1) (or partial regression plots)

- Let  $Y_i^{(1)}$  represent the residuals from the least-squares regression of  $Y$  on all of the  $X$ 's except for  $X_1$ :

$$Y_i = A^{(1)} + B_2^{(1)}X_{i2} + \cdots + B_k^{(1)}X_{ik} + Y_i^{(1)}$$

- Similarly,  $X_i^{(1)}$  are the residuals from the regression of  $X_1$  on all other  $X$ 's:

$$X_{i1} = C^{(1)} + D_2^{(1)}X_{i2} + \cdots + D_k^{(1)}X_{ik} + X_i^{(1)}$$

- These two equations determine the residuals  $Y^{(1)}$  and  $X^{(1)}$  as parts of  $Y$  and  $X_1$  that remain when the effects of  $X_2, \dots, X_k$  are removed

# Added-Variable Plots (2) (or partial regression plots)

- Residuals  $Y^{(1)}$  and  $X^{(1)}$  have the following properties:
  1. Slope of the regression of  $Y^{(1)}$  on  $X^{(1)}$  is the least-squares slope  $B_1$  from the full multiple regression
  2. Residuals from the regression of  $Y^{(1)}$  on  $X^{(1)}$  are the same as the residuals from the full regression:

$$Y_i^{(1)} = B_1 X_1^{(1)} + E_i$$

- 3. Variation of  $X^{(1)}$  is the conditional variance of  $X_1$  holding the other X's constant. Consequently, except for the  $df$  the standard error from the partial simple regression is the same as the multiple regression SE of  $B_1$ .

$$\widehat{SE(B_1)} = \frac{S_E}{\sqrt{\sum X_i^{(1)2}}}$$

# Added-Variable Plots (3)

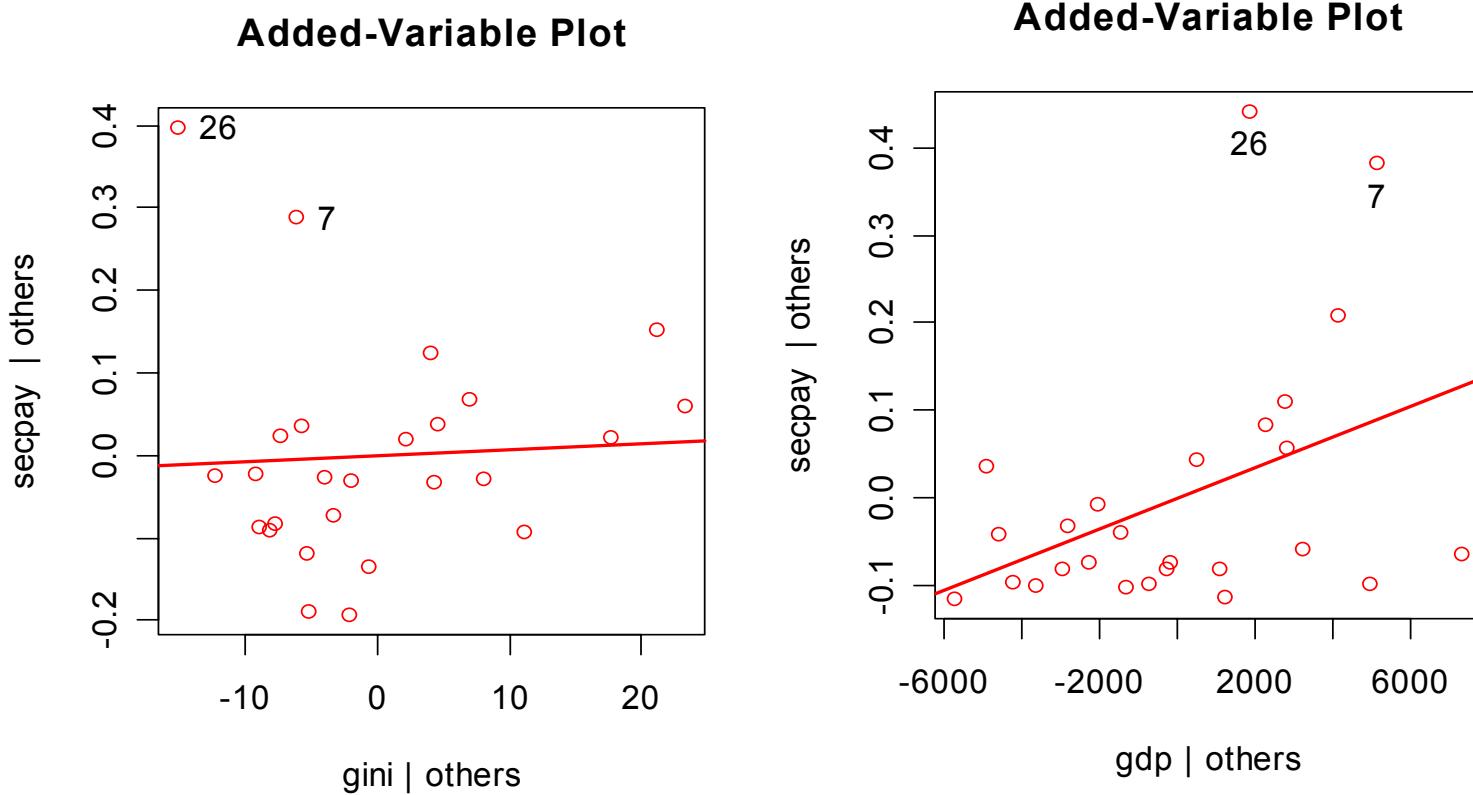
## An Example

- Once again recalling the outlier model from the Inequality data (Weakliem.model1)
- A plot of  $Y^{(1)}$  against  $X^{(1)}$  allows us to examine the leverage and influence of cases on  $B_1$ 
  - we make one plot for each  $X$
- These plots also gives us an idea of the precision of our slopes ( $B_1 \dots B_k$ )

```
#Added variable plots (partial regression plot)
>library(car)
>av.plots(Weakliem.model1)
#This allows you to choose the
#variables interactively
>leverage.plot(Weakliem.model1, "gini")
#This method you choose the
#variable of interest
```

# Added-Variable Plots (4)

## Example cont'd



- We see here that cases 7 (Czech Republic) and 26 (Slovakia) have unusually high Y values given their X's
- Because they are on the extreme of the X-range as well, they are most likely influencing both slopes

# Unusual Observations and their impact on Standard Errors

- Depending on their location, unusual observations can either increase or decrease standard errors
- Recall that the standard error for a slope is as follows:

$$\widehat{SE}(B) = \frac{s_E}{\sqrt{\sum(X_i - \bar{X})^2}}$$

- An observation with **high leverage** (i.e., an X-value far from the mean of X) increases the size of the denominator, and thus **decreases the standard error**
- A regression outlier (i.e., a point with a large residual) that does not have leverage (i.e., it does not have an unusual X-value) does not change the slope coefficients but will **increase the standard error**

# Unusual cases: Solutions?

- Unusual observations may reflect miscoding, in which case the observations can be rectified or deleted entirely
- Outliers are sometimes of substantive interest:
  - If only a few cases, we may decide to deal separately with them
  - Several outliers may reflect model misspecification—*i.e.*, an important explanatory variable that accounts for the subset of the data that are outliers has been neglected
- Unless there are strong reasons to remove outliers we may decide to keep them in the analysis and use alternative models to OLS, for example **robust regression**, which down weight outlying data.
  - Often these models give similar results to an OLS model that omits the influential cases, because they assign very low weight to highly influential cases

# **Summary (1)**

- Small samples are especially vulnerable to outliers—there are fewer cases to counter the outlier
- Large samples can also be affected, however, as shown by the “marital coital frequency” example
- Even if you have many cases, and your variables have limited ranges, miscodes that could influence the regression model are still possible
- Unusual cases are only influential when they are both unusual in terms of their Y value given their X (outlier), and when they have an unusual X-value (leverage):

**Influence = Leverage X Discrepancy**

## Summary (2)

- We can test for outliers using ***studentized residuals*** and ***quantile-comparison plots***
- Leverage is assessed by exploring the **hat-values**
- Influence is assessed using ***DFBetas*** and, preferably ***Cook's Ds***
- ***Influence Plots*** (or bubble plots) are useful because they display the studentized residuals, hat-values and Cook's distances all on the same plot
- Joint influence is best assessed using ***Added-Variable Plots*** (or partial-regression plots)

# 31. SIMPLE LINEAR REGRESSION

## VI: LEVERAGE AND INFLUENCE

These topics are not covered in the text, but they are important.

### Leverage

If the data set contains outliers, these can affect the least-squares fit.

To study the impact on the fitted line of moving a single data point, see the website at:

<http://www.stat.sc.edu/~west/javahtml/Regression.html>

If a given data point (say, the  $i^{\text{th}}$  one) is moved up or down, the corresponding fitted value  $\hat{y}_i$  will move proportionally to the change in  $y_i$ . The proportionality constant is called leverage, and denoted in Minitab by  $h_i$ . We get a value of the leverage  $h_i$  for each data point.

The leverage of a given of the data point measures the impact that  $y_i$  has on  $\hat{y}_i$ .

The further  $x_i$  is from  $\bar{x}$ , the larger  $h_i$ , and therefore the more sensitive  $\hat{y}_i$  is to changes in  $y_i$ .

So points with very large and very small  $x$  values have more leverage than points with intermediate  $x$  values.

If for some reason a point with high leverage also happens to be far from the least squares line which would be fitted to the remaining data points (i.e., if the point is an outlier), then we may need to take some action, e.g., delete the point, reconsider whether the model is reasonable, see if there was a recording error, etc.

It can be shown that the  $h_i$  are all between 0 and 1.

In practice  $h_i$  is considered large if it exceeds  $4/n$ .

## Influence Diagnostics

An observation is **influential** if the estimates change substantially when the point is omitted.

- Leverage depends only on the  $x$ 's, not on the  $y$ 's.
- A point with high leverage may or may not be influential.
- A point with low leverage may or may not be influential.
- Looking at residuals may not reveal influential points, since an outlier, particularly if it occurs at a point of high leverage, will tend to drag the fitted line along with it and therefore it may have a small residual. This phenomenon is called **masking**.

A more direct measure of the influence of the  $i^{\text{th}}$  data point is given by **Cook's D statistic**, which measures the sum of squared deviations between the observed  $\hat{y}$  values and the hypothetical  $\hat{y}$  values we would get if we deleted the  $i^{\text{th}}$  data point.

Observations with  $D_i > 1$  should be examined carefully.

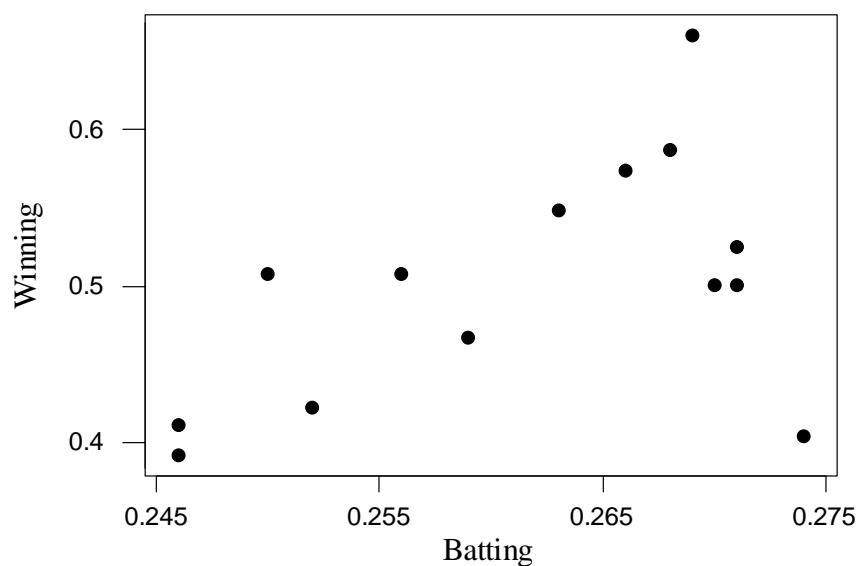
**Eg:** Consider the Team Batting Average ( $x$ ) and Team Winning Percentage ( $y$ ) for the 14 teams in the American League in 1986. The data file is **Baseball86.MTP**

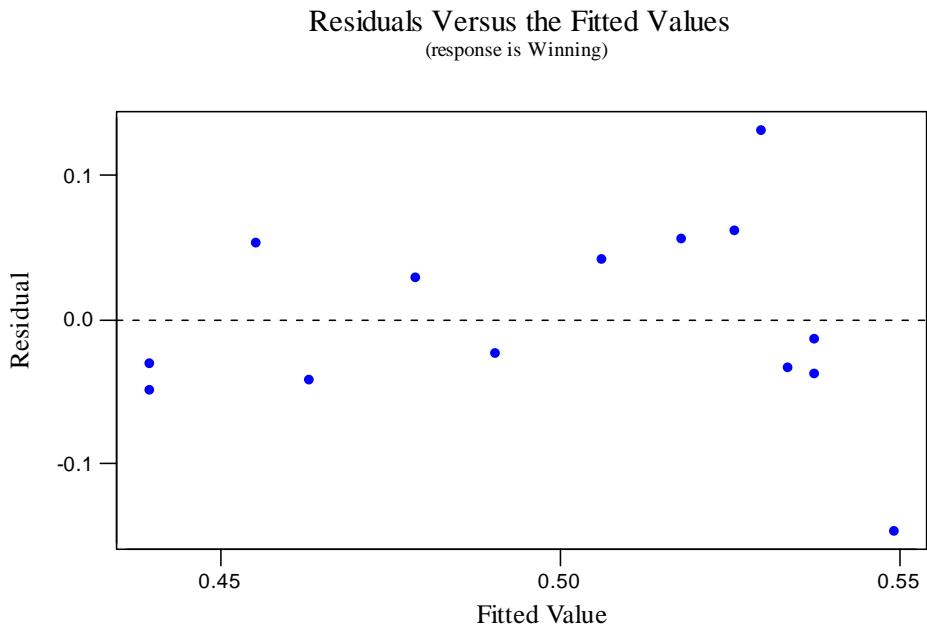
The scatterplot shows some indication of a positive linear association, although some of the teams with high batting averages have surprisingly low winning percentages. These teams are Cleveland, Milwaukee, Toronto, and Minnesota (the most extreme case).

The residual plot confirms that the linear model is far from perfect.

<b>Team</b>	<b>Team Batting Average (x)</b>	<b>Team Winning Percentage (y)</b>
Baltimore	.266	.574
Boston	.269	.661
California	.256	.508
Chicago	.246	.410
Cleveland	.271	.500
Detroit	.259	.467
Kansas City	.250	.508
Milwaukee	.271	.525
Minnesota	.274	.403
New York	.268	.587
Oakland	.252	.422
Seattle	.246	.391
Texas	.263	.548
Toronto	.270	.500

Scatterplot of Winning vs. Batting





The points which were "surprisingly low" in the scatterplot now show up as strongly negative residuals, indicating that for these teams, their winning percentages fall short of what would be predicted by a linear regression model. Another problem is that the residuals indicate an overall upward trend. This is a sign that the outliers have "dragged down" the fitted line.

The fitted model is  $\hat{y} = -0.5245 + 3.919x$ .

The  $p$ -value for  $\beta_1$  is 0.070, and  $R^2$  is 0.248, indicating a weak to moderate linear association.

## Regression Analysis

The regression equation is  
Winning = - 0.524 + 3.92 Batting

Predictor	Coef	SE Coef	T	P
Constant	-0.5245	0.5154	-1.02	0.329
Batting	3.919	1.969	1.99	0.070
S	0.07017	R-Sq = 24.8%	R-Sq(adj) = 18.5%	

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	0.019496	0.019496	3.96	0.070
Residual Error	12	0.059089	0.004924		
Total	13	0.078585			

Predicted Values

Fit	StDev Fit	95.0% CI	95.0% PI
0.4944	0.0190	( 0.4530, 0.5358)	( 0.3360, 0.6528)

Incidentally, if we delete the outlier teams, the  $p$ -value goes down to 0.000 and  $R^2$  goes up to 0.821.

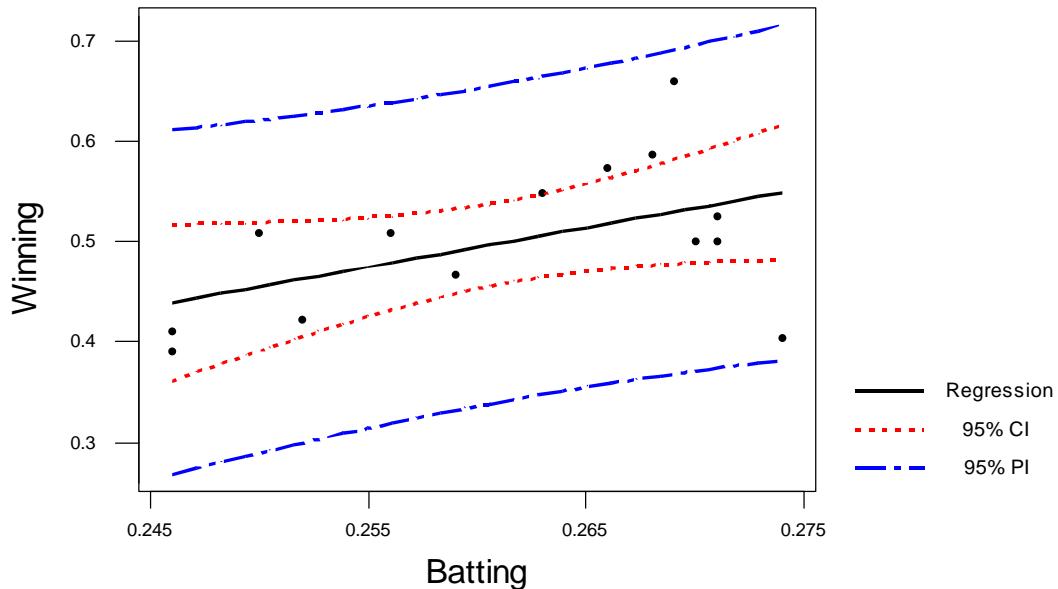
So the linear relationship *is* strong for the remaining 10 teams.

We next examine the Minitab "Fitted Line Plot".

This gives a scatterplot, together with the fitted line, and (an option for) 95% confidence and prediction intervals. Note that the confidence intervals are wider at the ends.

## Regression Plot

$Y = -5.2E-01 + 3.91887X$   
R-Sq = 24.8 %



Next, we compute the leverage and Cook's D statistics.

In Minitab, use Stat → Regression → Regression → Storage.  
Click boxes for Hi (leverage) and Cook's Distance.

The point for Minnesota (Case 9) has a leverage of 0.1945, which does not exceed  $4/n = 0.29$ , and therefore would not be considered extremely high.

It has a Cook's D of 0.65, which does not exceed 1, and so would not be considered an outlier by this criterion.

But the unusualness of Minnesota is partially masked by Cleveland, Milwaukee and Toronto. If we leave out all four teams, the results change drastically. In general, Cook's D can be "fooled" by multiple outliers.

## American League Baseball, 1986

<b>Team</b>	<b>Batting</b>	<b>Winning</b>	<b>H1</b>	<b>COOK1</b>
Baltimore	0.266	0.574	0.087380	0.033503
Boston	0.269	0.661	0.115737	0.259203
California	0.256	0.508	0.095257	0.010122
Chicago	0.246	0.410	0.260676	0.042267
Cleveland	0.271	0.500	0.142520	0.027700
Detroit	0.259	0.467	0.076352	0.005014
Kansas City	0.250	0.508	0.175603	0.073092
Milwaukee	0.271	0.525	0.142520	0.003083
Minnesota	0.274	0.403	0.194509	0.651305
New York	0.268	0.587	0.104709	0.049751
Oakland	0.252	0.422	0.142520	0.033177
Seattle	0.246	0.391	0.260676	0.114114
Texas	0.263	0.548	0.073201	0.015146
Toronto	0.270	0.500	0.128341	0.019360

### Regression Analysis

The regression equation is  
 $Winning = -0.524 + 3.92 \text{ Batting}$

Predictor	Coef	SE Coef	T	P
Constant	-0.5245	0.5154	-1.02	0.329
Batting	3.919	1.969	1.99	0.070

S = 0.07017      R-Sq = 24.8%      R-Sq(adj) = 18.5%

### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	0.019496	0.019496	3.96	0.070
Residual Error	12	0.059089	0.004924		
Total	13	0.078585			

### Predicted Values

Fit	StDev Fit	Fit	95.0% CI	95.0% PI
0.4944	0.0190	( 0.4530, 0.5358)	( 0.3360, 0.6528)	

## Regression Analysis

BASEBALL DATA, WITHOUT MINNESOTA, CLEVELAND, MILWAUKEE, TORONTO

The regression equation is

$$\text{Winning} = -1.79 + 8.93 \text{ Batting}$$

Predictor	Coef	SE Coef	T	P
Constant	-1.7913	0.3792	-4.72	0.001
Batting	8.928	1.472	6.07	0.000

$$S = 0.03895 \quad R-\text{Sq} = 82.1\% \quad R-\text{Sq}(\text{adj}) = 79.9\%$$

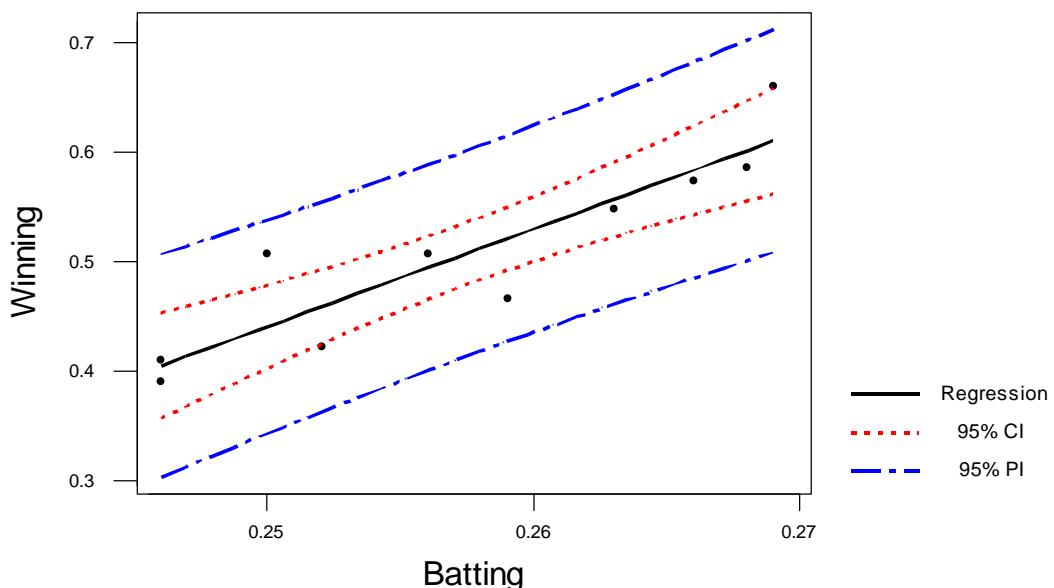
Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	0.055835	0.055835	36.80	0.000
Residual Error	8	0.012139	0.001517		
Total	9	0.067974			

## Baseball: Four Cases Omitted

$$Y = -1.79134 + 8.92791X$$

$$R-\text{Sq} = 82.1\%$$



---

# 7

# Dummy-Variable Regression

---

One of the serious limitations of multiple-regression analysis, as presented in Chapters 5 and 6, is that it accommodates only quantitative response and explanatory variables. In this chapter and the next, I will explain how qualitative explanatory variables, called *factors*, can be incorporated into a linear model.<sup>1</sup>

The current chapter begins with an explanation of how a *dummy-variable regressor* can be coded to represent a *dichotomous* (i.e., two-category) factor. I proceed to show how a set of dummy regressors can be employed to represent a *polytomous* (many-category) factor. I next describe how interactions between quantitative and qualitative explanatory variables can be represented in dummy-regression models and how to summarize models that incorporate interactions. Finally, I explain why it does not make sense to standardize dummy-variable and interaction regressors.

## 7.1 A Dichotomous Factor

---

Let us consider the simplest case: one dichotomous factor and one quantitative explanatory variable. As in the two previous chapters, assume that relationships are *additive*—that is, that the partial effect of each explanatory variable is the same regardless of the specific value at which the other explanatory variable is held constant. As well, suppose that the other assumptions of the regression model hold: The errors are independent and normally distributed, with zero means and constant variance.

The general motivation for including a factor in a regression is essentially the same as for including an additional quantitative explanatory variable: (1) to account more fully for the response variable, by making the errors smaller, and (2) even more important, to avoid a biased assessment of the impact of an explanatory variable, as a consequence of omitting another explanatory variable that is related to it.

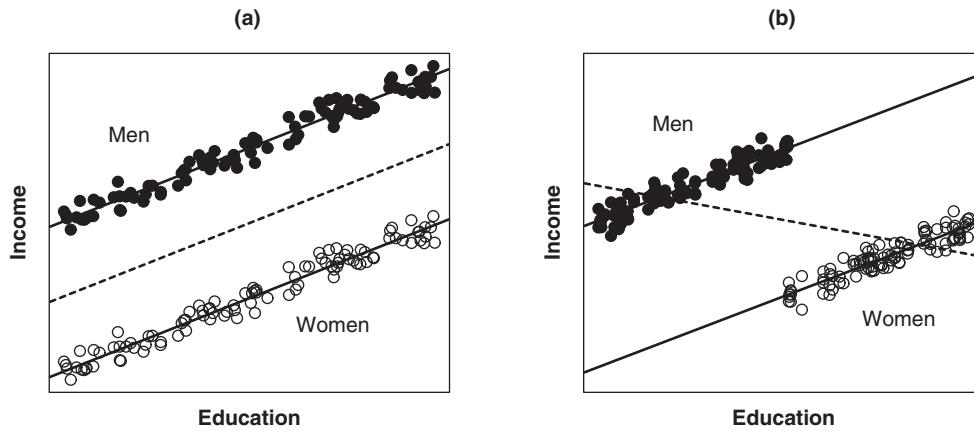
For concreteness, suppose that we are interested in investigating the relationship between education and income among women and men. Figure 7.1(a) and (b) represents two small (idealized) populations. In both cases, the within-gender regressions of income on education are parallel. Parallel regressions imply additive effects of education and gender on income: Holding education constant, the “effect” of gender is the vertical distance between the two regression lines, which—for parallel lines—is everywhere the same. Likewise, holding gender constant, the “effect” of education is captured by the within-gender education slope, which—for parallel lines—is the same for men and women.<sup>2</sup>

In Figure 7.1(a), the explanatory variables gender and education are unrelated to each other: Women and men have identical distributions of education scores (as can be seen by projecting the points onto the horizontal axis). In this circumstance, if we ignore gender and regress income on education alone, we obtain the same slope as is produced by the separate within-gender

---

<sup>1</sup>Chapter 14 deals with qualitative *response* variables.

<sup>2</sup>I will consider nonparallel within-group regressions in Section 7.3.



**Figure 7.1** Idealized data representing the relationship between income and education for populations of men (filled circles) and women (open circles). In (a), there is no relationship between education and gender; in (b), women have a higher average level of education than men. In both (a) and (b), the within-gender (i.e., partial) regressions (solid lines) are parallel. In each graph, the overall (i.e., marginal) regression of income on education (ignoring gender) is given by the broken line.

regressions. Because women have lower incomes than men of equal education, however, by ignoring gender we inflate the size of the errors.

The situation depicted in Figure 7.1(b) is importantly different. Here, gender and education are related, and therefore if we regress income on education alone, we arrive at a biased assessment of the effect of education on income: Because women have a higher average level of education than men, and because—for a given level of education—women's incomes are lower, on average, than men's, the overall regression of income on education has a *negative* slope even though the within-gender regressions have a *positive* slope.<sup>3</sup>

In light of these considerations, we might proceed to partition our sample by gender and perform separate regressions for women and men. This approach is reasonable, but it has its limitations: Fitting separate regressions makes it difficult to estimate and test for gender differences in income. Furthermore, if we can reasonably assume parallel regressions for women and men, we can more efficiently estimate the common education slope by pooling sample data drawn from both groups. In particular, if the usual assumptions of the regression model hold, then it is desirable to fit the common-slope model by least squares.

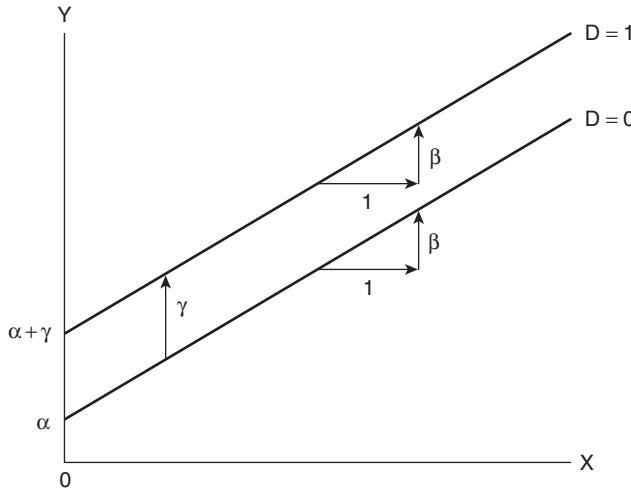
One way of formulating the common-slope model is

$$Y_i = \alpha + \beta X_i + \gamma D_i + \varepsilon_i \quad (7.1)$$

where  $D$ , called a *dummy-variable regressor* or an *indicator variable*, is coded 1 for men and 0 for women:

$$D_i = \begin{cases} 1 & \text{for men} \\ 0 & \text{for women} \end{cases}$$

<sup>3</sup>That marginal and partial relationships can differ in sign is called *Simpson's paradox* (Simpson, 1951). Here, the marginal relationship between income and education is negative, while the partial relationship, controlling for gender, is positive.



**Figure 7.2** The additive dummy-variable regression model. The line labeled  $D = 1$  is for men; the line labeled  $D = 0$  is for women.

Thus, for women the model becomes

$$Y_i = \alpha + \beta X_i + \gamma(0) + \varepsilon_i = \alpha + \beta X_i + \varepsilon_i$$

and for men

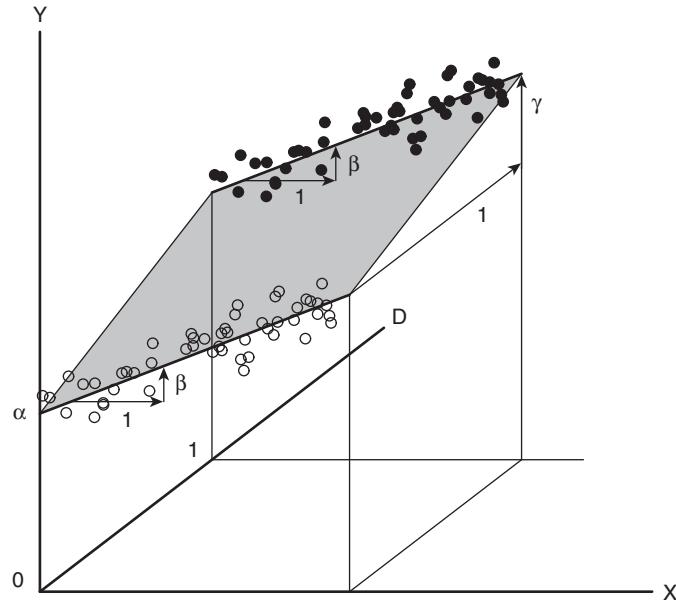
$$Y_i = \alpha + \beta X_i + \gamma(1) + \varepsilon_i = (\alpha + \gamma) + \beta X_i + \varepsilon_i$$

These regression equations are graphed in Figure 7.2.

This is our initial encounter with an idea that is fundamental to many linear models: the distinction between *explanatory variables* and *regressors*. Here, *gender* is a qualitative explanatory variable (i.e., a factor), with categories *male* and *female*. The dummy variable  $D$  is a regressor, representing the factor gender. In contrast, the quantitative explanatory variable *education* and the regressor  $X$  are one and the same. Were we to transform education, however, prior to entering it into the regression equation—say, by taking logs—then there would be a distinction between the explanatory variable (*education*) and the regressor (*log education*). In subsequent sections of this chapter, it will transpire that an explanatory variable can give rise to several regressors and that some regressors are functions of more than one explanatory variable.

Returning to Equation 7.1 and Figure 7.2, the coefficient  $\gamma$  for the dummy regressor gives the difference in intercepts for the two regression lines. Moreover, because the within-gender regression lines are parallel,  $\gamma$  also represents the constant vertical separation between the lines, and it may, therefore, be interpreted as the expected income advantage accruing to men when education is held constant. If men were *disadvantaged* relative to women with the same level of education, then  $\gamma$  would be *negative*. The coefficient  $\alpha$  gives the intercept for women, for whom  $D = 0$ ; and  $\beta$  is the common within-gender education slope.

Figure 7.3 reveals the fundamental geometric “trick” underlying the coding of a dummy regressor: We are, in fact, fitting a regression plane to the data, but the dummy regressor  $D$  is defined only at the values 0 and 1. The regression plane intersects the planes  $\{X, Y|D = 0\}$  and  $\{X, Y|D = 1\}$  in two lines, each with slope  $\beta$ . Because the difference between  $D = 0$  and  $D = 1$  is one unit, the difference in the  $Y$ -intercepts of these two lines is the slope of the plane in the  $D$  direction,



**Figure 7.3** The geometric “trick” underlying dummy regression: The linear regression plane is defined only at  $D = 0$  and  $D = 1$ , producing two regression lines with slope  $\beta$  and vertical separation  $\gamma$ . The hollow circles represent women, for whom  $D = 0$ , and the solid circles men, for whom  $D = 1$ .

that is  $\gamma$ . Indeed, Figure 7.2 is simply the projection of the two regression lines onto the  $\{X, Y\}$  plane.

Essentially similar results are obtained if we instead code  $D$  equal to 0 for men and 1 for women, making men the *baseline* (or *reference*) category (see Figure 7.4): The *sign* of  $\gamma$  is reversed, because it now represents the difference in intercepts between women and men (rather than vice versa), but its *magnitude* remains the same. The coefficient  $\alpha$  now gives the income intercept for men. It is therefore immaterial which group is coded 1 and which is coded 0, as long as we are careful to interpret the coefficients of the model—for example, the sign of  $\gamma$ —in a manner consistent with the coding scheme that is employed.

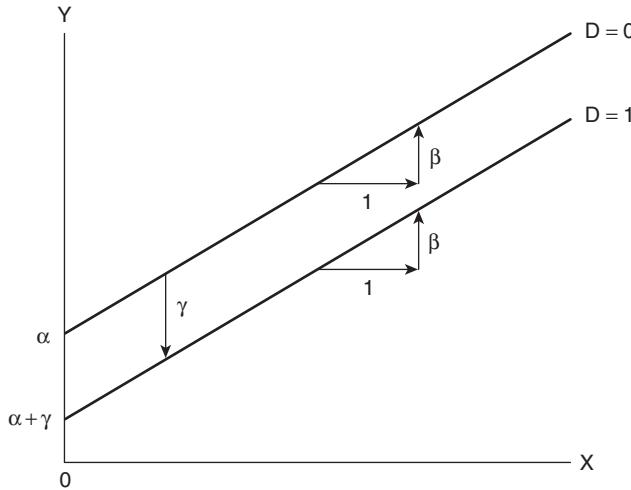
To determine whether gender affects income, controlling for education, we can test  $H_0: \gamma = 0$ , either by a  $t$ -test, dividing the estimate of  $\gamma$  by its standard error, or, equivalently, by dropping  $D$  from the regression model and formulating an incremental  $F$ -test. In either event, the statistical-inference procedures of the previous chapter apply.

Although I have developed dummy-variable regression for a single quantitative regressor, the method can be applied to any number of quantitative explanatory variables, as long as we are willing to assume that the slopes are the same in the two categories of the factor—that is, that the regression surfaces are parallel in the two groups. In general, if we fit the model

$$Y_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \gamma D_i + \varepsilon_i$$

then, for  $D = 0$ , we have

$$Y_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i$$



**Figure 7.4** The additive dummy-regression model coding  $D = 0$  for men and  $D = 1$  for women (cf., Figure 7.2).

and, for  $D = 1$ ,

$$Y_i = (\alpha + \gamma) + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

A dichotomous factor can be entered into a regression equation by formulating a dummy regressor, coded 1 for one category of the factor and 0 for the other category. A model incorporating a dummy regressor represents parallel regression surfaces, with the constant vertical separation between the surfaces given by the coefficient of the dummy regressor.

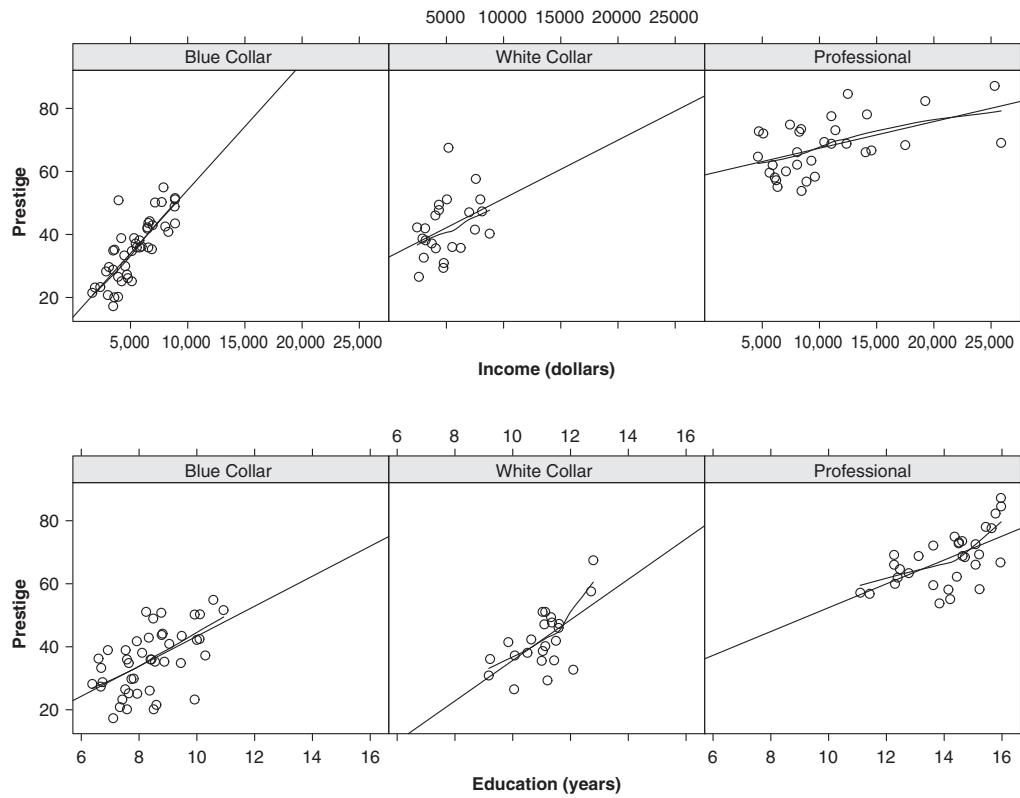
## 7.2 Polytomous Factors

The coding method of the previous section generalizes straightforwardly to polytomous factors. By way of illustration, recall (from the previous chapter) the Canadian occupational prestige data. I have classified the occupations into three rough categories: (1) professional and managerial occupations, (2) “white-collar” occupations, and (3) “blue-collar” occupations.<sup>4</sup>

Figure 7.5 shows conditioning plots for the relationship between prestige and each of income and education within occupational types.<sup>5</sup> The partial relationships between prestige and the explanatory variables appear reasonably linear, although there seems to be evidence that the income slope varies across the categories of type of occupation (a possibility that I will pursue in the next section of the chapter). Indeed, this change in slope is an explanation of the nonlinearity in the relationship between prestige and income that we noticed in Chapter 4. These conditioning

<sup>4</sup>Although there are 102 occupations in the full data set, several are difficult to classify and consequently were dropped from the analysis. The omitted occupations are athletes, babysitters, farmers, and “newsboys,” leaving us with 98 observations.

<sup>5</sup>In the preceding chapter, I also included the gender composition of the occupations as an explanatory variable, but I omit that variable here. Conditioning plots are described in Section 3.3.4.



**Figure 7.5** Conditioning plots for the relationships between prestige and each of income (top panel) and education (bottom panel) by type of occupation, for the Canadian occupational prestige data. Each panel shows the linear least-squares fit and a lowess smooth with a span of 0.9. The graphs labeled “Professional” are for professional and managerial occupations.

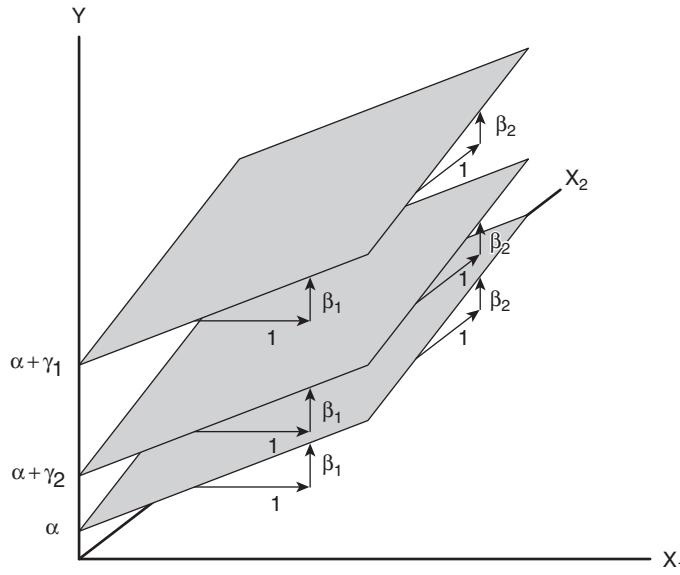
plots do not tell the whole story, however, because the income and education levels of the occupations are correlated, but they give us a reasonable initial look at the data. Conditioning the plot for income by level of education (and vice versa) is out of the question here because of the small size of the data set.

The *three*-category occupational-type factor can be represented in the regression equation by introducing *two* dummy regressors, employing the following coding scheme:

Category	$D_1$	$D_2$	
Professional and managerial	1	0	(7.2)
White collar	0	1	
Blue collar	0	0	

A model for the regression of prestige on income, education, and type of occupation is then

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \varepsilon_i \quad (7.3)$$



**Figure 7.6** The additive dummy-regression model with two quantitative explanatory variables  $X_1$  and  $X_2$  represents parallel planes with potentially different intercepts in the  $\{X_1, X_2, Y\}$  space.

where  $X_1$  is income and  $X_2$  is education. This model describes three parallel regression planes, which can differ in their intercepts:

$$\begin{aligned} \text{Professional: } & Y_i = (\alpha + \gamma_1) + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \\ \text{White collar: } & Y_i = (\alpha + \gamma_2) + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \\ \text{Blue collar: } & Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \end{aligned}$$

The coefficient  $\alpha$ , therefore, gives the intercept for blue-collar occupations;  $\gamma_1$  represents the constant vertical difference between the parallel regression planes for professional and blue-collar occupations (fixing the values of education and income); and  $\gamma_2$  represents the constant vertical distance between the regression planes for white-collar and blue-collar occupations (again, fixing education and income). Assuming, for simplicity, that all coefficients are positive, and that  $\gamma_1 > \gamma_2$ , the geometry of the model in Equation 7.3 is illustrated in Figure 7.6.

Because blue-collar occupations are coded 0 for both dummy regressors, “blue collar” implicitly serves as the baseline category to which the other occupational-type categories are compared. The choice of a baseline category is essentially arbitrary, for we would fit precisely the same three regression planes regardless of which of the three occupational-type categories is selected for this role. The values (and meaning) of the individual dummy-variable coefficients  $\gamma_1$  and  $\gamma_2$  depend, however, on which category is chosen as the baseline.

It is sometimes natural to select a particular category as a basis for comparison—an experiment that includes a “control group” comes immediately to mind. In this instance, the individual dummy-variable coefficients are of interest, because they reflect differences between the “experimental” groups and the control group, holding other explanatory variables constant.

In most applications, however, the choice of a baseline category is entirely arbitrary, as it is for the occupational prestige regression. We are, therefore, most interested in testing the null hypothesis of no effect of occupational type, controlling for education and income,

$$H_0: \gamma_1 = \gamma_2 = 0 \tag{7.4}$$

but the individual hypotheses  $H_0: \gamma_1 = 0$  and  $H_0: \gamma_2 = 0$ —which test, respectively, for differences between professional and blue-collar occupations and between white-collar and blue-collar occupations—are of less intrinsic interest.<sup>6</sup> The null hypothesis in Equation 7.4 can be tested by the incremental-sum-of-squares approach, dropping the two dummy variables for type of occupation from the model.

I have demonstrated how to model the effects of a three-category factor by coding two dummy regressors. It may seem more natural to treat the three occupational categories symmetrically, coding *three* dummy regressors, rather than arbitrarily selecting one category as the baseline:

Category	$D_1$	$D_2$	$D_3$	
Professional and managerial	1	0	0	(7.5)
White collar	0	1	0	
Blue collar	0	0	1	

Then, for the  $j$ th occupational type, we would have

$$Y_i = (\alpha + \gamma_j) + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

The problem with this procedure is that there are too many parameters: We have used four parameters ( $\alpha, \gamma_1, \gamma_2, \gamma_3$ ) to represent only three group intercepts. As a consequence, we could not find unique values for these four parameters even if we knew the three population regression lines. Likewise, we cannot calculate unique least-squares estimates for the model because the set of three dummy variables is perfectly collinear; for example, as is apparent from the table in Equation 7.5,  $D_3 = 1 - D_1 - D_2$ .

In general, then, for a polytomous factor with  $m$  categories, we need to code  $m - 1$  dummy regressors. One simple scheme is to select the last category as the baseline and to code  $D_{ij} = 1$  when observation  $i$  falls in category  $j$ , and 0 otherwise:

Category	$D_1$	$D_2$	...	$D_{m-1}$	
1	1	0	...	0	
2	0	1	...	0	(7.6)
:	:	:		:	
$m-1$	0	0	...	1	
$m$	0	0	...	0	

A polytomous factor can be entered into a regression by coding a set of 0/1 dummy regressors, one fewer than the number of categories of the factor. The “omitted” category, coded 0 for all dummy regressors in the set, serves as a baseline to which the other categories are compared. The model represents parallel regression surfaces, one for each category of the factor.

<sup>6</sup>The essential point here is not that the separate hypotheses are of *no* interest but that they are an arbitrary subset of the pairwise differences among the categories. In the present case, where there are three categories, the individual hypotheses represent two of the three pairwise group comparisons. The third comparison, between professional and white-collar occupations, is not *directly* represented in the model, although it is given indirectly by the difference  $\gamma_1 - \gamma_2$ . See Section 7.2.1 for an elaboration of this point.

When there is more than one factor, and if we assume that the factors have additive effects, we can simply code a set of dummy regressors for each. To test the hypothesis that the effect of a factor is nil, we delete its dummy regressors from the model and compute an incremental  $F$ -test of the hypothesis that all the associated coefficients are 0.

Regressing occupational prestige ( $Y$ ) on income ( $X_1$ ) and education ( $X_2$ ) produces the fitted regression equation

$$\hat{Y} = -7.621 + 0.001241X_1 + 4.292X_2 \quad R^2 = .81400$$

$$(3.116) \quad (0.000219) \quad (0.336)$$

As is common practice, I have shown the estimated standard error of each regression coefficient in parentheses beneath the coefficient. The three occupational categories differ considerably in their average levels of prestige:

Category	Number of Cases	Mean Prestige
Professional and managerial	31	67.85
White collar	23	42.24
Blue collar	44	35.53
All occupations	98	47.33

Inserting dummy variables for type of occupation into the regression equation, employing the coding scheme shown in Equation 7.2, produces the following results:

$$\hat{Y} = -0.6229 + 0.001013X_1 + 3.673X_2 + 6.039D_1 - 2.737D_2$$

$$(5.2275) \quad (0.000221) \quad (0.641) \quad (3.867) \quad (2.514)$$

$$R^2 = .83486 \quad (7.7)$$

The three fitted regression equations are, therefore,

$$\begin{aligned} \text{Professional: } \hat{Y} &= 5.416 + 0.001013X_1 + 3.673X_2 \\ \text{White collar: } \hat{Y} &= -3.360 + 0.001013X_1 + 3.673X_2 \\ \text{Blue collar: } \hat{Y} &= -0.623 + 0.001013X_1 + 3.673X_2 \end{aligned}$$

Note that the coefficients for both income and education become slightly smaller when type of occupation is controlled. As well, the dummy-variable coefficients (or, equivalently, the category intercepts) reveal that when education and income levels are held constant statistically, the difference in average prestige between professional and blue-collar occupations declines greatly, from  $67.85 - 35.53 = 32.32$  points to 6.04 points. The difference between white-collar and blue-collar occupations is reversed when income and education are held constant, changing from  $42.24 - 35.53 = +6.71$  points to  $-2.74$  points. That is, the greater prestige of professional occupations compared with blue-collar occupations appears to be due mostly to differences in education and income between these two classes of occupations. While white-collar occupations have greater prestige, on average, than blue-collar occupations, they have lower prestige than blue-collar occupations of the same educational and income levels.<sup>7</sup>

To test the null hypothesis of no partial effect of type of occupation,

$$H_0: \gamma_1 = \gamma_2 = 0$$

<sup>7</sup>These conclusions presuppose that the additive model that we have fit to the data is adequate, which, as we will see in Section 7.3.5, is not the case.

we can calculate the incremental  $F$ -statistic

$$\begin{aligned} F_0 &= \frac{n - k - 1}{q} \times \frac{R_1^2 - R_0^2}{1 - R_1^2} \\ &= \frac{98 - 4 - 1}{2} \times \frac{.83486 - .81400}{1 - .83486} = 5.874 \end{aligned} \quad (7.8)$$

with 2 and 93 degrees of freedom, for which  $p = .0040$ . The occupational-type effect is therefore statistically significant but (examining the coefficient standard errors) not very precisely estimated. The education and income coefficients are several times their respective standard errors, and hence are highly statistically significant.

### 7.2.1 Coefficient Quasi-Variances\*

Consider a dummy-regression model with  $p$  quantitative explanatory variables and an  $m$ -category factor:

$$Y_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \cdots + \gamma_{m-1} D_{i,m-1} + \varepsilon_i$$

The dummy-variable coefficients  $\gamma_1, \gamma_2, \dots, \gamma_{m-1}$  represent differences (or *contrasts*) between each of the other categories of the factor and the reference category  $m$ , holding constant  $X_1, \dots, X_p$ . If we are interested in a comparison between any other two categories, we can simply take the difference in their dummy-regressor coefficients. Thus, in the preceding example (letting  $C_1 \equiv \hat{\gamma}_1$  and  $C_2 \equiv \hat{\gamma}_2$ ),

$$C_1 - C_2 = 5.416 - (-3.360) = 8.776$$

is the estimated average difference in prestige between professional and white-collar occupations of equal income and education.

Suppose, however, that we want to know the standard error of  $C_1 - C_2$ . The standard errors of  $C_1$  and  $C_2$  are available directly in the regression “output” (Equation 7.7), but to compute the standard error of  $C_1 - C_2$ , we need in addition the estimated sampling covariance of these two coefficients. That is,<sup>8</sup>

$$\text{SE}(C_1 - C_2) = \sqrt{\hat{V}(C_1) + \hat{V}(C_2) - 2 \times \hat{C}(C_1, C_2)}$$

where  $\hat{V}(C_j) = [\text{SE}(C_j)]^2$  is the estimated sampling variance of coefficient  $C_j$ , and  $\hat{C}(C_1, C_2)$  is the estimated sampling covariance of  $C_1$  and  $C_2$ . For the occupational prestige regression,  $\hat{C}(C_1, C_2) = 6.797$ , and so

$$\text{SE}(C_1 - C_2) = \sqrt{3.867^2 + 2.514^2 - 2 \times 6.797} = 2.771$$

We can use this standard error in the normal manner for a  $t$ -test of the difference between  $C_1$  and  $C_2$ .<sup>9</sup> For example, noting that the difference exceeds twice its standard error suggests that it is statistically significant.

<sup>8</sup>See Appendix D on probability and estimation. The computation of regression-coefficient covariances is taken up in Chapter 9.

<sup>9</sup>Testing all differences between pairs of factor categories raises an issue of simultaneous inference, however. See the discussion of Scheffé confidence intervals in Section 9.4.4.

Although computer programs for regression analysis typically report the covariance matrix of the regression coefficients if asked to do so, it is not common to include coefficient covariances in published research along with estimated coefficients and standard errors, because with  $k + 1$  coefficients in the model, there are  $k(k + 1)/2$  variances and covariances among them—a potentially large number. Readers of a research report are therefore put at a disadvantage by the arbitrary choice of a reference category in dummy regression, because they are unable to calculate the standard errors of the differences between all pairs of categories of a factor.

*Quasi-variances* of dummy-regression coefficients (Firth, 2003; Firth & De Menezes, 2004) speak to this problem. Let  $\tilde{V}(C_j)$  denote the quasi-variance of dummy coefficient  $C_j$ . Then,

$$\text{SE}(C_j - C_{j'}) \approx \sqrt{\tilde{V}(C_j) + \tilde{V}(C_{j'})}$$

The squared relative error of this approximation for the contrast  $C_j - C_{j'}$  is

$$\text{RE}_{jj'} \equiv \frac{\tilde{V}(C_j - C_{j'})}{\widehat{V}(C_j - C_{j'})} = \frac{\tilde{V}(C_j) + \tilde{V}(C_{j'})}{\widehat{V}(C_j) + \widehat{V}(C_{j'}) - 2 \times \widehat{C}(C_j, C_{j'})}$$

The approximation is accurate for this contrast when  $\text{RE}_{jj'}$  is close to 1, or, equivalently, when

$$\log(\text{RE}_{jj'}) = \log[\tilde{V}(C_j) + \tilde{V}(C_{j'})] - \log[\widehat{V}(C_j) + \widehat{V}(C_{j'}) - 2 \times \widehat{C}(C_j, C_{j'})]$$

is close to 0. The quasi-variances  $\tilde{V}(C_j)$  are therefore selected to minimize the sum of squared log relative errors of approximation over all pairwise contrasts,  $\sum_{j < j'} [\log(\text{RE}_{jj'})]^2$ . The resulting errors of approximation are typically very small (Firth, 2003; Firth & De Menezes, 2004).

The following table gives dummy-variable coefficients, standard errors, and quasi-variances for type of occupation in the Canadian occupational prestige regression:

Category	$C_j$	$\text{SE}(C_j)$	$\tilde{V}(C_j)$
Professional	6.039	3.867	8.155
White collar	-2.737	2.514	-0.4772
Blue collar	0	0	6.797

I have set to 0 the coefficient (and its standard error) for the baseline category, blue collar. The negative quasi-variance for the white-collar coefficient is at first blush disconcerting (after all, ordinary variances cannot be negative), but it is not wrong: The quasi-variances are computed to provide accurate variance approximations for coefficient *differences*; they do not apply directly to the coefficients themselves. For the contrast between professional and white-collar occupations, we have

$$\text{SE}(C_1 - C_2) \approx \sqrt{8.155 - 0.4772} = 2.771$$

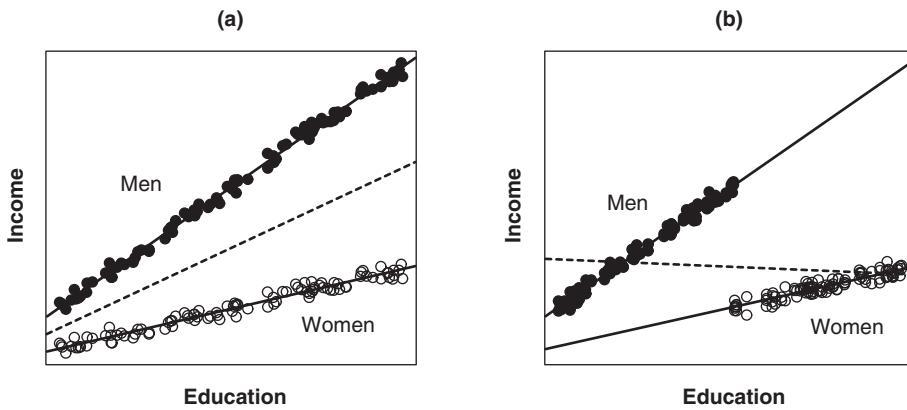
Likewise, for the contrast between professional and blue-collar occupations,

$$C_1 - C_3 = 6.039 - 0 = 6.039$$

$$\text{SE}(C_1 - C_3) \approx \sqrt{8.155 + 6.797} = 3.867$$

Note that in this application, the quasi-variance “approximation” to the standard error proves to be exact, and indeed this is necessarily the case when there are just three factor categories, because there are then just three pairwise differences among the categories to capture.<sup>10</sup>

<sup>10</sup>For the details of the computation of quasi-variances, see Chapter 15, Exercise 15.11.



**Figure 7.7** Idealized data representing the relationship between income and education for populations of men (filled circles) and women (open circles). In (a), there is no relationship between education and gender; in (b), women have a higher average level of education than men. In both cases, the within-gender regressions (solid lines) are not parallel—the slope for men is greater than the slope for women—and, consequently, education and gender interact in affecting income. In each graph, the overall regression of income on education (ignoring gender) is given by the broken line.

### 7.3 Modeling Interactions

Two explanatory variables are said to *interact* in determining a response variable when the partial effect of one depends on the value of the other. The additive models that we have considered thus far therefore specify the *absence* of interactions. In this section, I will explain how the dummy-variable regression model can be modified to accommodate interactions between factors and quantitative explanatory variables.<sup>11</sup>

The treatment of dummy-variable regression in the preceding two sections has assumed parallel regressions across the several categories of a factor. If these regressions are *not* parallel, then the factor interacts with one or more of the quantitative explanatory variables. The dummy-regression model can easily be modified to reflect these interactions.

For simplicity, I return to the contrived example of Section 7.1, examining the regression of income on gender and education. Consider the hypothetical data shown in Figure 7.7 (and contrast these examples with those shown in Figure 7.1 on page 121, where the effects of gender and education are additive). In Figure 7.7(a) [as in Figure 7.1(a)], gender and education are independent, because women and men have identical education distributions; in Figure 7.7(b) [as in Figure 7.1(b)], gender and education are related, because women, on average, have higher levels of education than men.

It is apparent in both Figure 7.7(a) and Figure 7.7(b), however, that the within-gender regressions of income on education are not parallel: In both cases, the slope for men is larger than the slope for women. Because the effect of education varies by gender, education and gender interact in affecting income.

It is also the case, incidentally, that the effect of gender varies by education. Because the regressions are not parallel, the relative income advantage of men changes (indeed, grows) with

<sup>11</sup>Interactions between factors are taken up in the next chapter on analysis of variance; interactions between quantitative explanatory variables are discussed in Section 17.1 on polynomial regression.

education. Interaction, then, is a symmetric concept—that the effect of education varies by gender implies that the effect of gender varies by education (and, of course, vice versa).

The simple examples in Figures 7.1 and 7.7 illustrate an important and frequently misunderstood point: *Interaction* and *correlation* of explanatory variables are empirically and logically distinct phenomena. Two explanatory variables can interact *whether or not* they are related to one another statistically. Interaction refers to the manner in which explanatory variables *combine* to affect a response variable, not to the relationship *between* the explanatory variables themselves.

Interaction and correlation of explanatory variables are empirically and logically distinct phenomena. Two explanatory variables can interact whether or not they are related to one another statistically. Interaction refers to the manner in which explanatory variables combine to affect a response variable, not to the relationship between the explanatory variables themselves.

### 7.3.1 Constructing Interaction Regressors

We could model the data in Figure 7.7 by fitting separate regressions of income on education for women and men. As before, however, it is more convenient to fit a combined model, primarily because a combined model facilitates a test of the gender-by-education interaction. Moreover, a properly formulated unified model that permits different intercepts and slopes in the two groups produces the same fit to the data as separate regressions: The full sample is composed of the two groups, and, consequently, the residual sum of squares for the full sample is minimized when the residual sum of squares is minimized in each group.<sup>12</sup>

The following model accommodates different intercepts and slopes for women and men:

$$Y_i = \alpha + \beta X_i + \gamma D_i + \delta(X_i D_i) + \varepsilon_i \quad (7.9)$$

Along with the quantitative regressor  $X$  for education and the dummy regressor  $D$  for gender, I have introduced the *interaction regressor*  $XD$  into the regression equation. The interaction regressor is the *product* of the other two regressors; although  $XD$  is therefore a function of  $X$  and  $D$ , it is not a *linear* function, and perfect collinearity is avoided.<sup>13</sup>

For women, model (7.9) becomes

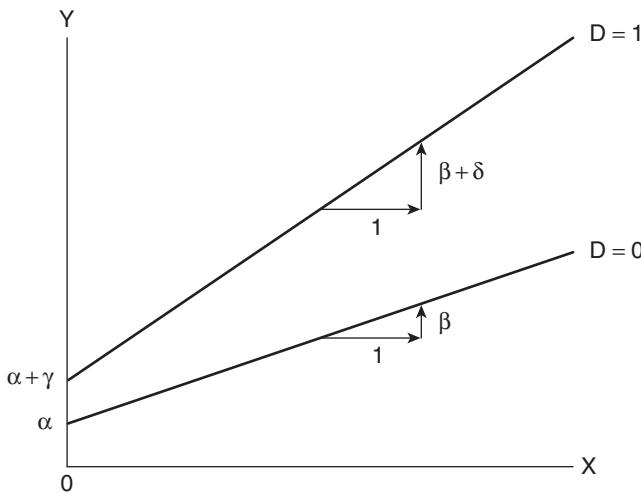
$$\begin{aligned} Y_i &= \alpha + \beta X_i + \gamma(0) + \delta(X_i \cdot 0) + \varepsilon_i \\ &= \alpha + \beta X_i + \varepsilon_i \end{aligned}$$

and for men

$$\begin{aligned} Y_i &= \alpha + \beta X_i + \gamma(1) + \delta(X_i \cdot 1) + \varepsilon_i \\ &= (\alpha + \gamma) + (\beta + \delta)X_i + \varepsilon_i \end{aligned}$$

<sup>12</sup>See Exercise 7.4.

<sup>13</sup>If this procedure seems illegitimate, then think of the interaction regressor as a new variable, say  $Z \equiv XD$ . The model is linear in  $X$ ,  $D$ , and  $Z$ . The “trick” of introducing an interaction regressor is similar to the trick of formulating dummy regressors to capture the effect of a factor: In both cases, there is a distinction between explanatory variables and regressors. Unlike a dummy regressor, however, the interaction regressor is a function of *both* explanatory variables.



**Figure 7.8** The dummy-variable regression model with an interaction regressor. The line labeled  $D = 1$  is for men; the line labeled  $D = 0$  is for women.

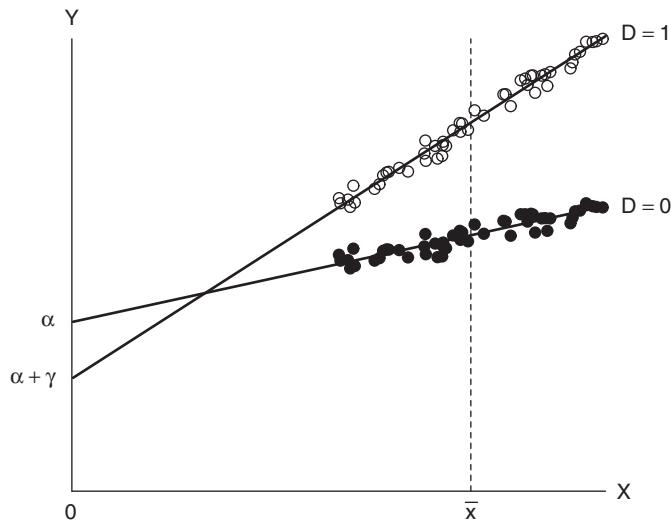
These regression equations are graphed in Figure 7.8: The parameters  $\alpha$  and  $\beta$  are, respectively, the intercept and slope for the regression of income on education among women (the baseline category for gender);  $\gamma$  gives the difference in intercepts between the male and female groups; and  $\delta$  gives the difference in slopes between the two groups. To test for interaction, therefore, we may simply test the hypothesis  $H_0: \delta = 0$ .

Interactions can be incorporated by coding interaction regressors, taking products of dummy regressors with quantitative explanatory variables. The resulting model permits different slopes in different groups—that is, regression surfaces that are not parallel.

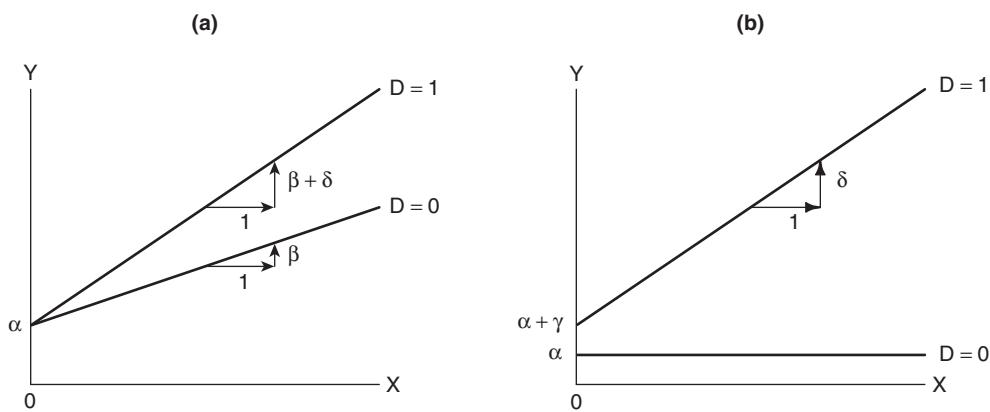
In the additive, no-interaction model of Equation 7.1 and Figure 7.2, the dummy-regressor coefficient  $\gamma$  represents the *unique* partial effect of gender (i.e., the expected income difference between men and women of equal education, regardless of the value at which education is fixed), while the slope  $\beta$  represents the *unique* partial effect of education (i.e., the within-gender expected increment in income for a one-unit increase in education, for both women and men). In the interaction model of Equation 7.9 and Figure 7.8, in contrast,  $\gamma$  is no longer interpretable as the unqualified income difference between men and women of equal education.

Because the within-gender regressions are not parallel, the separation between the regression lines changes; here,  $\gamma$  is simply the separation at  $X = 0$ —that is, above the origin. It is generally no more important to assess the expected income difference between men and women of 0 education than at other educational levels, and therefore the difference-in-intercepts parameter  $\gamma$  is not of special interest in the interaction model. Indeed, in many instances (although not here), the value  $X = 0$  may not occur in the data or may be impossible (as, for example, if  $X$  is weight). In such cases,  $\gamma$  has no literal interpretation in the interaction model (see Figure 7.9).

Likewise, in the interaction model,  $\beta$  is not the unqualified partial effect of education, but rather the effect of education among women. Although this coefficient *is* of interest, it is not necessarily



**Figure 7.9** Why the difference in intercepts does not represent a meaningful partial effect for a factor when there is interaction: The difference-in-intercepts parameter  $\gamma$  is *negative* even though, within the range of the data, the regression line for the group coded  $D = 1$  is *above* the line for the group coded  $D = 0$ .



**Figure 7.10** Two models that violate the principle of marginality: In (a), the dummy regressor  $D$  is omitted from the model  $E(Y) = \alpha + \beta X + \delta(XD)$ ; in (b), the quantitative explanatory variable  $X$  is omitted from the model  $E(Y) = \alpha + \gamma D + \delta(XD)$ . These models violate the principle of marginality because they include the term  $XD$ , which is a higher-order relative of both  $X$  and  $D$  (one of which is omitted from each model).

more important than the effect of education among men ( $\beta + \delta$ ), which does not appear *directly* in the model.

### 7.3.2 The Principle of Marginality

Following Nelder (1977), we say that the separate partial effects, or *main effects*, of education and gender are *marginal* to the education-by-gender interaction. In general, we neither test nor interpret the main effects of explanatory variables that interact. If, however, we can rule out interaction either on theoretical or on empirical grounds, then we can proceed to test, estimate, and interpret the main effects.

As a corollary to this principle, it does not generally make sense to specify and fit models that include interaction regressors but that omit main effects that are marginal to them. This is not to say that such models—which violate the *principle of marginality*—are uninterpretable: They are, rather, not broadly applicable.

The principle of marginality specifies that a model including a *high-order term* (such as an interaction) should normally also include the “lower-order relatives” of that term (the main effects that “compose” the interaction).

Suppose, for example, that we fit the model

$$Y_i = \alpha + \beta X_i + \delta(X_i D_i) + \varepsilon_i$$

which omits the dummy regressor  $D$ , but includes its “higher-order relative”  $XD$ . As shown in Figure 7.10(a), this model describes regression lines for women and men that have the same intercept but (potentially) different slopes, a specification that is peculiar and of no substantive interest. Similarly, the model

$$Y_i = \alpha + \gamma D_i + \delta(X_i D_i) + \varepsilon_i$$

graphed in Figure 7.10(b), constrains the slope for women to 0, which is needlessly restrictive.

### 7.3.3 Interactions With Polytomous Factors

The method for modeling interactions by forming product regressors is easily extended to polytomous factors, to several factors, and to several quantitative explanatory variables. I will use the Canadian occupational prestige regression to illustrate the application of the method, entertaining the possibility that occupational type interacts both with income ( $X_1$ ) and with education ( $X_2$ ):

$$\begin{aligned} Y_i = & \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 D_{i1} + \gamma_2 D_{i2} \\ & + \delta_{11} X_{i1} D_{i1} + \delta_{12} X_{i1} D_{i2} + \delta_{21} X_{i2} D_{i1} + \delta_{22} X_{i2} D_{i2} + \varepsilon_i \end{aligned} \quad (7.10)$$

Note that we require one interaction regressor for each product of a dummy regressor with a quantitative explanatory variable. The regressors  $X_1 D_1$  and  $X_1 D_2$  capture the interaction between income and occupational type;  $X_2 D_1$  and  $X_2 D_2$  capture the interaction between education and

occupational type. The model therefore permits different intercepts and slopes for the three types of occupations:

$$\begin{aligned} \text{Professional: } Y_i &= (\alpha + \gamma_1) + (\beta_1 + \delta_{11})X_{i1} + (\beta_2 + \delta_{21})X_{i2} + \varepsilon_i \\ \text{White collar: } Y_i &= (\alpha + \gamma_2) + (\beta_1 + \delta_{12})X_{i1} + (\beta_2 + \delta_{22})X_{i2} + \varepsilon_i \\ \text{Blue collar: } Y_i &= \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \end{aligned} \quad (7.11)$$

Blue-collar occupations, which are coded 0 for both dummy regressors, serve as the baseline for the intercepts and slopes of the other occupational types. As in the no-interaction model, the choice of baseline category is generally arbitrary, as it is here, and is inconsequential. Fitting the model in Equation 7.10 to the prestige data produces the following results:

$$\begin{aligned} \widehat{Y}_i &= 2.276 + 0.003522X_1 + 1.713X_2 + 15.35D_1 - 33.54D_2 \\ &\quad (7.057) \quad (0.000556) \quad (0.927) \quad (13.72) \quad (17.54) \\ &\quad - 0.002903X_1D_1 - 0.002072X_1D_2 \\ &\quad (0.000599) \quad (0.000894) \\ &\quad + 1.388X_2D_1 + 4.291X_2D_2 \\ &\quad (1.289) \quad (1.757) \\ R^2 &= .8747 \end{aligned} \quad (7.12)$$

This example is discussed further in the following section.

### 7.3.4 Interpreting Dummy-Regression Models With Interactions

It is difficult in dummy-regression models with interactions (and in other complex statistical models) to understand what the model is saying about the data simply by examining the regression coefficients. One approach to interpretation, which works reasonably well in a relatively straightforward model such as Equation 7.12, is to write out the implied regression equation for each group (using Equation 7.11):

$$\begin{aligned} \text{Professional: } \widehat{\text{Prestige}} &= 17.63 + 0.000619 \times \text{Income} + 3.101 \times \text{Education} \\ \text{White collar: } \widehat{\text{Prestige}} &= -31.26 + 0.001450 \times \text{Income} + 6.004 \times \text{Education} \\ \text{Blue collar: } \widehat{\text{Prestige}} &= 2.276 + 0.003522 \times \text{Income} + 1.713 \times \text{Education} \end{aligned} \quad (7.13)$$

From these equations, we can see, for example, that income appears to make much more difference to prestige in blue-collar occupations than in white-collar occupations, and has even less impact on prestige in professional and managerial occupations. Education, in contrast, has the largest impact on prestige among white-collar occupations, and has the smallest effect in blue-collar occupations.

An alternative approach (from Fox, 1987, 2003; Fox & Andersen, 2006) that generalizes readily to more complex models is to examine the high-order terms of the model. In the illustration, the high-order terms are the interactions between income and type and between education and type.

- Focusing in turn on each high-order term, we allow the variables in the term to range over their combinations of values in the data, fixing other variables to typical values. For example, for the interaction between type and income, we let type of occupation take on successively the categories blue collar, white collar, and professional (for which the dummy regressors

$D_1$  and  $D_2$  are set to the corresponding values given in Equation 7.6), in combination with income values between \$1500 and \$26,000 (the approximate range of income in the Canadian occupational prestige data set); education is fixed to its average value in the data,  $\bar{X}_2 = 10.79$ .

- We next compute the fitted value of prestige at each combination of values of income and type of occupation. These fitted values are graphed in the “effect display” shown in the upper panel of Figure 7.11; the lower panel of this figure shows a similar effect display for the interaction between education and type of occupation, holding income at its average value. The broken lines in Figure 7.11 give  $\pm 2$  standard errors around the fitted values—that is, approximate 95% pointwise confidence intervals for the effects.<sup>14</sup> The nature of the interactions between income and type and between education and type is readily discerned from these graphs.

### 7.3.5 Hypothesis Tests for Main Effects and Interactions

To test the null hypothesis of no interaction between income and type,  $H_0: \delta_{11} = \delta_{12} = 0$ , we need to delete the interaction regressors  $X_1 D_1$  and  $X_1 D_2$  from the full model (Equation 7.10) and calculate an incremental  $F$ -test; likewise, to test the null hypothesis of no interaction between education and type,  $H_0: \delta_{21} = \delta_{22} = 0$ , we delete the interaction regressors  $X_2 D_1$  and  $X_2 D_2$  from the full model. These tests, and tests for the main effects of income, education, and occupational type, are detailed in Tables 7.1 and 7.2: Table 7.1 gives the regression sums of squares for several models, which, along with the residual sum of squares for the full model,  $RSS_1 = 3553$ , are the building blocks of the incremental  $F$ -tests shown in Table 7.2. Table 7.3 shows the hypothesis tested by each of the incremental  $F$ -statistics in Table 7.2.

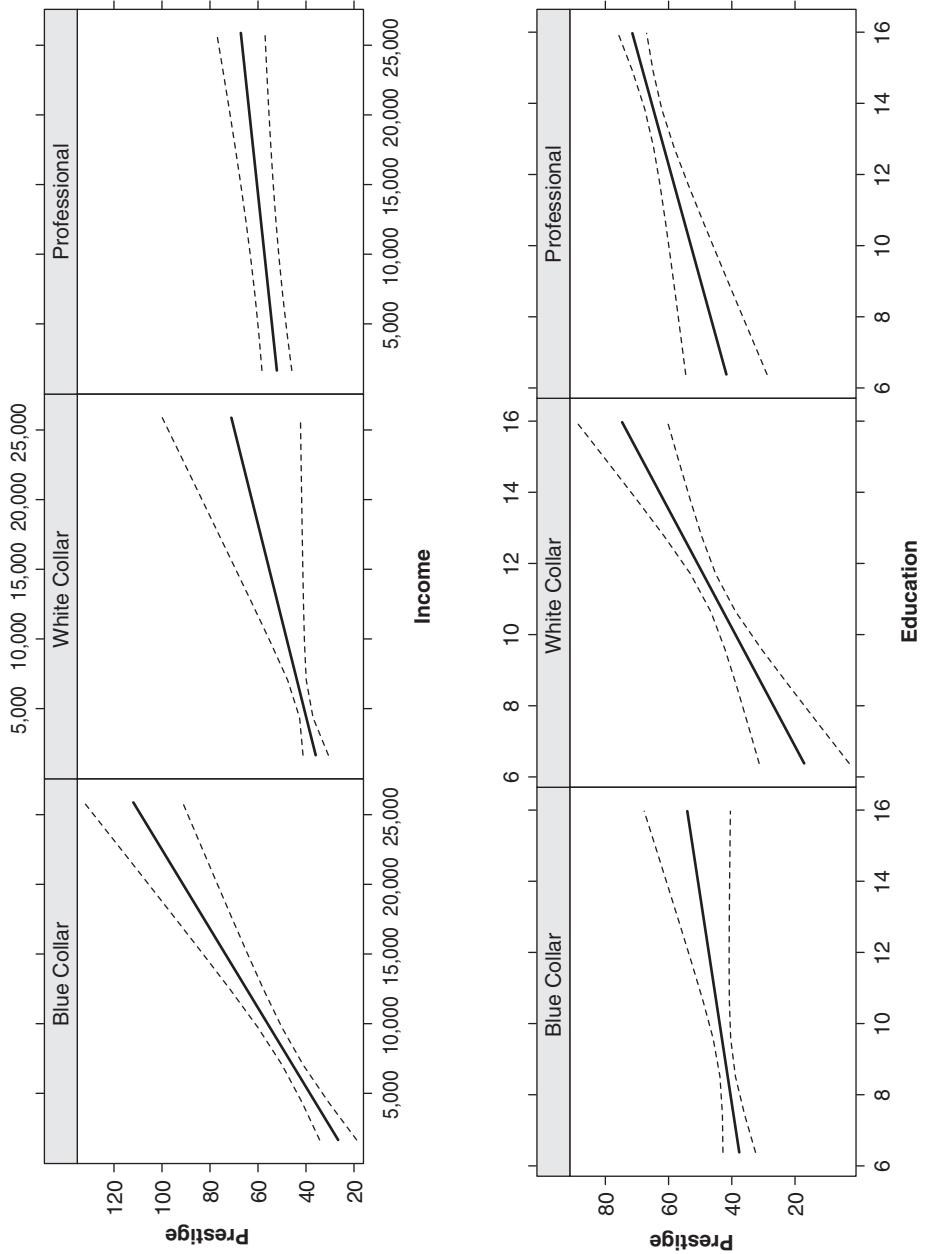
Although the analysis-of-variance table (Table 7.2) conventionally shows the tests for the main effects of education, income, and type before the education-by-type and income-by-type interactions, the structure of the model makes it sensible to examine the interactions first: Conforming to the principle of marginality, the test for each main effect is computed assuming that the interactions that are higher-order relatives of the main effect are 0 (as shown in Table 7.3). Thus, for example, the test for the income main effect assumes that the income-by-type interaction is absent (i.e., that  $\delta_{11} = \delta_{12} = 0$ ), but not that the education-by-type interaction is absent ( $\delta_{21} = \delta_{22} = 0$ ).<sup>15</sup>

The principle of marginality serves as a guide to constructing incremental  $F$ -tests for the terms in a model that includes interactions.

In this case, then, there is weak evidence of an interaction between education and type of occupation, and much stronger evidence of an income-by-type interaction. Considering the small number of cases, we are squeezing the data quite hard, and it is apparent from the coefficient standard errors (in Equation 7.12) and from the effect displays in Figure 7.11 that the interactions are not precisely estimated. The tests for the main effects of income, education, and type, computed assuming that the higher-order relatives of each such term are absent, are all highly statistically

<sup>14</sup>For standard errors of fitted values, see Exercise 9.14.

<sup>15</sup>Tests constructed to conform to the principle of marginality are sometimes called “type-II” tests, terminology introduced by the SAS statistical software package. This terminology, and alternative tests, are described in the next chapter.



**Figure 7.11** Income-by-type (upper panel) and education-by-type (lower panel) “effect displays” for the regression of prestige on income, education, and type of occupation. The solid lines give fitted values under the model, while the broken lines give 95% pointwise confidence intervals around the fit. To compute fitted values in the upper panel, education is set to its average value in the data; in the lower panel, income is set to its average value.

**Table 7.1** Regression Sums of Squares for Several Models Fit to the Canadian Occupational Prestige Data

Model	Terms	Parameters	Regression Sum of Squares	df
1	$I, E, T, I \times T, E \times T$	$\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2, \delta_{11}, \delta_{12}, \delta_{21}, \delta_{22}$	24,794.	8
2	$I, E, T, I \times T$	$\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2, \delta_{11}, \delta_{12}$	24,556.	6
3	$I, E, T, E \times T$	$\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2, \delta_{21}, \delta_{22}$	23,842.	6
4	$I, E, T$	$\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2$	23,666.	4
5	$I, E$	$\alpha, \beta_1, \beta_2$	23,074.	2
6	$I, T, I \times T$	$\alpha, \beta_1, \gamma_1, \gamma_2, \delta_{11}, \delta_{12}$	23,488.	5
7	$E, T, E \times T$	$\alpha, \beta_2, \gamma_1, \gamma_2, \delta_{21}, \delta_{22}$	22,710.	5

NOTE: These sums of squares are the building blocks of incremental  $F$ -tests for the main and interaction effects of the explanatory variables. The following code is used for “terms” in the model:  $I$ , income;  $E$ , education;  $T$ , occupational type.

**Table 7.2** Analysis-of-Variance Table, Showing Incremental  $F$ -Tests for the Terms in the Canadian Occupational Prestige Regression

Source	Models Contrasted	Sum of Squares	df	F	p
Income	3–7	1132.	1	28.35	<.0001
Education	2–6	1068.	1	26.75	<.0001
Type	4–5	592.	2	7.41	<.0011
Income $\times$ Type	1–3	952.	2	11.92	<.0001
Education $\times$ Type	1–2	238.	2	2.98	.056
Residuals		3553.	89		
Total		28,347.	97		

**Table 7.3** Hypotheses Tested by the Incremental  $F$ -Tests in Table 7.2

Source	Models Contrasted	Null Hypothesis
Income	3–7	$\beta_1 = 0   \delta_{11} = \delta_{12} = 0$
Education	2–6	$\beta_2 = 0   \delta_{21} = \delta_{22} = 0$
Type	4–5	$\gamma_1 = \gamma_2 = 0   \delta_{11} = \delta_{12} = \delta_{21} = \delta_{22} = 0$
Income $\times$ Type	1–3	$\delta_{11} = \delta_{12} = 0$
Education $\times$ Type	1–2	$\delta_{21} = \delta_{22} = 0$

significant. In light of the strong evidence for an interaction between income and type, however, the income and type main effects are not really of interest.<sup>16</sup>

The degrees of freedom for the several sources of variation add to the total degrees of freedom, but—because the regressors in different sets are correlated—the sums of squares do not add to the total sum of squares.<sup>17</sup> What is important here (and more generally) is that sensible hypotheses are tested, not that the sums of squares add to the total sum of squares.

## 7.4 A Caution Concerning Standardized Coefficients

In Chapter 5, I explained the use—and limitations—of standardized regression coefficients. It is appropriate to sound another cautionary note here: Inexperienced researchers sometimes report standardized coefficients for dummy regressors. As I have explained, an *unstandardized* coefficient for a dummy regressor is interpretable as the expected response-variable difference between a particular category and the baseline category for the dummy-regressor set (controlling, of course, for the other explanatory variables in the model).

If a dummy-regressor coefficient is standardized, then this straightforward interpretation is lost. Furthermore, because a 0/1 dummy regressor cannot be increased by one standard deviation, the usual interpretation of a standardized regression coefficient also does not apply. Standardization is a linear transformation, so many characteristics of the regression model—the value of  $R^2$ , for example—do not change, but the standardized coefficient itself is not directly interpretable. These difficulties can be avoided by standardizing only the response variable and *quantitative* explanatory variables in a regression, leaving dummy regressors in 0/1 form.

A similar point applies to interaction regressors. We may legitimately standardize a quantitative explanatory variable *prior* to taking its product with a dummy regressor, but to standardize the interaction regressor itself is not sensible: The interaction regressor cannot change independently of the main-effect regressors that compose it and are marginal to it.

It is not sensible to standardize dummy regressors or interaction regressors.

## Exercises

**Exercise 7.1.** Suppose that the values  $-1$  and  $1$  are used for the dummy regressor  $D$  in Equation 7.1 instead of  $0$  and  $1$ . Write out the regression equations for men and women, and explain how the parameters of the model are to be interpreted. Does this alternative coding of the

<sup>16</sup>We tested the occupational type main effect in Section 7.2 (Equation 7.8 on page 129), but using an estimate of error variance based on Model 4, which does not contain the interactions. In Table 7.2, the estimated error variance is based on the full model, Model 1. Sound general practice is to use the largest model fit to the data to estimate the error variance even when, as is frequently the case, this model includes effects that are not statistically significant. The largest model necessarily has the smallest residual sum of squares, but it also has the fewest residual degrees of freedom. These two factors tend to offset one another, and it usually makes little difference whether the estimated error variance is based on the full model or on a model that deletes nonsignificant terms. Nevertheless, using the full model ensures an unbiased estimate of the error variance.

<sup>17</sup>See Section 10.2 for a detailed explanation of this phenomenon.

dummy regressor adequately capture the effect of gender? Is it fair to conclude that the dummy-regression model will “work” properly as long as two distinct values of the dummy regressor are employed, one each for women and men? Is there a reason to prefer one coding to another?

**Exercise 7.2.** Adjusted means (based on Section 7.2): Let  $\bar{Y}_1$  represent the (“unadjusted”) mean prestige score of professional occupations in the Canadian occupational prestige data,  $\bar{Y}_2$  that of white-collar occupations, and  $\bar{Y}_3$  that of blue-collar occupations. Differences among the  $\bar{Y}_j$  may partly reflect differences among occupational types in their income and education levels. In the dummy-variable regression in Equation 7.7, type-of-occupation differences are “controlled” for income and education, producing the fitted regression equation

$$\hat{Y} = A + B_1X_1 + B_2X_2 + C_1D_1 + C_2D_2$$

Consequently, if we fix income and education at particular values—say,  $X_1 = x_1$  and  $X_2 = x_2$ —then the fitted prestige scores for the several occupation types are given by (treating “blue collar” as the baseline type):

$$\begin{aligned}\hat{Y}_1 &= (A + C_1) + B_1x_1 + B_2x_2 \\ \hat{Y}_2 &= (A + C_2) + B_1x_1 + B_2x_2 \\ \hat{Y}_3 &= \quad A \quad + B_1x_1 + B_2x_2\end{aligned}$$

- (a) Note that the *differences* among the  $\hat{Y}_j$  depend only on the dummy-variable coefficients  $C_1$  and  $C_2$  and not on the values of  $x_1$  and  $x_2$ . Why is this so?
- (b) When  $x_1 = \bar{X}_1$  and  $x_2 = \bar{X}_2$ , the  $\hat{Y}_j$  are called *adjusted means* and are denoted  $\tilde{Y}_j$ . How can the adjusted means  $\tilde{Y}_j$  be interpreted? In what sense is  $\tilde{Y}_j$  an “adjusted” mean?
- (c) Locate the “unadjusted” and adjusted means for women and men in each of Figures 7.1(a) and (b) (on page 121). Construct a similar figure in which the difference between adjusted means is *smaller* than the difference in unadjusted means.
- (d) Using the results in the text, along with the mean income and education values for the three occupational types, compute adjusted mean prestige scores for each of the three types, controlling for income and education. Compare the adjusted with the unadjusted means for the three types of occupations and comment on the differences, if any, between them.

**Exercise 7.3.** Can the concept of an adjusted mean, introduced in Exercise 7.2, be extended to a model that includes interactions? If so, show how adjusted means can be found for the data in Figure 7.7(a) and (b) (on page 131).

**Exercise 7.4.** Verify that the regression equations for each occupational type given in Equation 7.13 (page 136) are identical to the results obtained by regressing prestige on income and education *separately* for each of the three types of occupations. Explain why this is the case.

## Summary

---

- A dichotomous factor can be entered into a regression equation by formulating a dummy regressor, coded 1 for one category of the variable and 0 for the other category. A model incorporating a dummy regressor represents parallel regression surfaces, with the constant separation between the surfaces given by the coefficient of the dummy regressor.
- A polytomous factor can be entered into a regression by coding a set of 0/1 dummy regressors, one fewer than the number of categories of the factor. The “omitted” category, coded

0 for all dummy regressors in the set, serves as a baseline to which the other categories are compared. The model represents parallel regression surfaces, one for each category of the factor.

- Interactions can be incorporated by coding interaction regressors, taking products of dummy regressors with quantitative explanatory variables. The model permits different slopes in different groups—that is, regression surfaces that are not parallel.
- *Interaction* and *correlation* of explanatory variables are empirically and logically distinct phenomena. Two explanatory variables can interact *whether or not* they are related to one another statistically. Interaction refers to the manner in which explanatory variables *combine* to affect a response variable, not to the relationship *between* the explanatory variables themselves
- The principle of marginality specifies that a model including a high-order term (such as an interaction) should normally also include the lower-order relatives of that term (the main effects that “compose” the interaction). The principle of marginality also serves as a guide to constructing incremental *F*-tests for the terms in a model that includes interactions, and for examining the effects of explanatory variables.
- It is not sensible to standardize dummy regressors or interaction regressors.

[Home](#)[Academic](#)[Downloads](#)[About Us](#)

Compare Statistical Tools  
From \$4.99  
[Learn More](#)

## Applied Statistics Handbook

[Table of Contents](#) | [Purchase Info](#)



Applied Statistics Handbook  
\$4.99  
[Learn More](#)



### Adjusted R<sup>2</sup>

Adjusted R<sup>2</sup> is used to compensate for the addition of variables to the model. As more independent variables are added to the regression model, unadjusted R<sup>2</sup> will generally increase but there will never be a decrease. This will occur even when the additional variables do little to help explain the dependent variable. To compensate for this, adjusted R<sup>2</sup> is corrected for the number of independent variables in the model. The result is an adjusted R<sup>2</sup> than can go up or down depending on whether the addition of another variable adds or does not add to the explanatory power of the model. Adjusted R<sup>2</sup> will always be lower than unadjusted.

It has become standard practice to report the adjusted R<sup>2</sup>, especially when there are multiple models presented with varying numbers of independent variables.

$$\bar{R}^2 = \left( R^2 - \frac{k}{n-1} \right) \frac{n-1}{n-k-1}$$

$$\bar{R}^2 = \left( .948 - \frac{2}{5-1} \right) \frac{5-1}{5-2-1}, \quad \bar{R}^2 = (.948 - .50)(2), \quad \bar{R}^2 = .806$$


---



Google™

Web

[www.acastat.com](http://www.acastat.com)

[Product Brochure](#) | [Workbook](#) | [Stat Handbook](#) | [Links](#) | [Site License](#) | [Documentation](#) | [Updates](#) | [Privacy](#) | [Contact Us](#) | [About Us](#)

Copyright © 2012, AcaStat Software. All Rights Reserved.

Mac and the Mac logo are trademarks of Apple Inc. Windows and the Windows logo are trademarks of Microsoft Corp. Both are registered in the U.S. and other countries.

Cross Validated is a question and answer site for people interested in statistics, machine learning, data analysis, data mining, and data visualization. It's 100% free, no registration required.

[Take the 2-minute tour](#) 

## Adjusted vs. unadjusted effects in regression

Is "unadjusted" basically just simple linear regression whereas "adjusted" is multiple regression? For example, looking at the effect of x on y adjusting for other variables like a, b and c versus not adjusting for them.

[regression](#) | [multiple-regression](#)

edited Oct 3 '11 at 15:48

 whuber ♦ 80.7k 8 135 283

asked Oct 3 '11 at 15:26

 question 112 1 6

Yes, that is my understanding – [Peter Flom](#) ♦ Oct 3 '11 at 16:25

I agree. And apparently "yes" isn't long enough to be a valid answer. – [Karl](#) Oct 3 '11 at 21:22

### 1 Answer

Since based on the comments "Yes" isn't long enough to be an answer:

Yes.

When a regression reports an unadjusted estimate, it's just a regression of X on Y with no other covariates. An adjusted estimate is the same regression of X on Y in the presence of at least one covariate.

answered Oct 3 '11 at 23:18

 Fomite 11.3k 2 35 77

# Influence Measures for CART Classification Trees

Avner Bar-Hen\*, Servane Gey†, Jean-Michel Poggi‡

## Abstract

This paper deals with measuring the influence of observations on the results obtained with CART classification trees. To define the influence of individuals on the analysis, we use influence measures to propose criterions to measure the sensitivity of the CART classification tree analysis. The proposals, based on jackknife trees, are organized around two lines: influence on predictions and influence on partitions. In addition, the analysis is extended to the pruned sequences of CART trees to produce a CART specific notion of influence.

A numerical example, the well known spam dataset, is presented to illustrate the notions developed throughout the paper. A real dataset relating the administrative classification of cities surrounding Paris, France, to the characteristics of their tax revenues distribution, is finally analyzed using the new influence-based tools.

## 1 Introduction

Classification And Regression Trees (CART; Breiman *et al.* (1984) [4]) have proven to be very useful in various applied contexts mainly because models can include numerical as well as nominal explanatory variables and because models can be easily represented (see for example Zhang and Singer (2010) [21], or Bel *et al.* (2009) [2]). Because

---

\*Laboratoire MAP5, Université Paris Descartes & EHESP, Rennes, France

†Laboratoire MAP5, Université Paris Descartes, France

‡Laboratoire de Mathématiques, Université Paris Sud, Orsay, France and Université Paris Descartes, France

CART is a nonparametric method as well as it provides data partitioning into distinct groups, such tree models have several additional advantages over other techniques: for example input data do not need to be normally distributed, predictor variables are not supposed to be independent, and non linear relationships between predictor variables and observed data can be handled.

It is well known that CART appears to be sensitive to perturbations of the learning set. This drawback is even a key property to make resampling and ensemble-based methods (as bagging and boosting) effective (see Gey and Poggi (2006) [12]). To preserve interpretability of the obtained model, it is important in many practical situations to try to restrict to a single tree. The stability of decision trees is then clearly an important issue and then it is important to be able to evaluate the sensitivity of the data on the results. Recently Briand *et al.* (2009) [5] propose a similarity measure between trees to quantify it and use it from an optimization perspective to build a less sensitive variant of CART. This view of instability related to bootstrap ideas can be also examined from a local perspective. Following this line, Bousquet and Elisseeff (2002) [3] studied the stability of a given method by replacing one observation in the learning sample with another one coming from the same model.

The aim of this paper is to focus on individual observations diagnosis issues rather than model properties or variable selection problems. The use of an influence measure is a classical diagnostic method to measure the perturbation induced by a single element, in other terms we examine stability issue through jackknife. We use decision trees to perform diagnosis on observations.

Many authors derived asymptotic normality of the influence functions under weak assumptions. Variance of the asymptotic normal distribution is generally estimated through resampling techniques. Therefore these results could be used to obtain a threshold to decide whether an observation is an outlier or not. Apart the question of the rate of convergence to normal distribution, we prefer not to use this approach and to propose descriptive statistics that allows to come back to the data in order to decide if an influent observation is an outlier or not. The outline is the following. Section 2 recalls first some general background on the so-called CART method. Then it introduces some influence measures for CART based on predictions, or on partitions, and finally an influence measure more deeply related to CART method in-

volving sequences of pruned trees. Section 3 contains an illustrative numerical application on the spam dataset (see Hastie, Tibshirani and Friedman (2009) [13]). Section 4 explores an original dataset relating the administrative classification of cities surrounding Paris, France, to the characteristics of their tax revenues distribution, by using the new influence-based tools. Finally Section 5 opens some perspectives.

## 2 Methods and Notations

### 2.1 CART classification trees

Let us briefly recall, following Bel *et al.* [2], some general background on Classification And Regression Trees (CART). For more detailed presentation see Breiman *et al.* [4] or, for a simple introduction, see Venables and Ripley (2002) [19]. The data are considered as an independent sample of the random variables  $(X^1, \dots, X^p, Y)$ , where the  $X^k$ 's are the explanatory variables and  $Y$  is the categorical variable to be explained. CART is a rule-based method that generates a binary tree through recursive partitioning that splits a subset (called a node) of the data set into two subsets (called sub-nodes) according to the minimization of a heterogeneity criterion computed on the resulting sub-nodes. Each split is based on a single variable. Remark that even if from a theoretical point of view CART methodology also allows for more general splits, most of the packages that implement CART only consider univariate splits, we adopt here this restriction. Some variables may be used several times while others may not be used at all. Let us consider a decision tree  $T$ . When  $Y$  is a categorical variable a class label is assigned to each terminal node (or leaf) of  $T$ . Hence  $T$  can be viewed as a mapping to assign a value  $\widehat{Y}_i = T(X_i^1, \dots, X_i^p)$  to each observation. The growing step leading to a deep maximal tree is obtained by recursive partitioning of the training sample by selecting the best split at each node according to some heterogeneity index, such that it is equal to 0 when there is only one class represented in the node to be split, and is maximum when all classes are equally frequent. The two most popular heterogeneity criteria are the Shannon entropy and the Gini index. Among all binary partitions of each set of values of the explanatory variables at a node  $t$ , the principle of CART is to split  $t$  into two sub-nodes  $t_-$  and  $t_+$  according to a threshold on one of the variables (or a subset of the labels for categorical variables), such that the reduction of heterogeneity between a node and

the two sub-nodes is maximized. The growing procedure is stopped when there is no more admissible splitting. Each leaf is assigned to the most frequent class of its observations. Of course, such a maximal tree (denoted by  $T_{max}$ ) generally overfits the training data and the associated prediction error  $R(T_{max})$ , with

$$R(T) = \mathbb{P}(T(X^1, \dots, X^p) \neq Y), \quad (1)$$

is typically large. Since the goal is to build from the available data a tree  $T$  whose prediction error is as small as possible, in a second stage the tree  $T_{max}$  is pruned to produce a subtree  $T'$  whose expected performance is close to the minimum of  $R(T')$  over all binary subtrees  $T'$  of  $T_{max}$ . Since the joint distribution  $\mathbb{P}$  of  $(X^1, \dots, X^p, Y)$  is unknown, the pruning is based on the penalized empirical risk  $\hat{R}_{pen}(T)$  to balance optimistic estimates of empirical risk by adding a complexity term that penalizes larger subtrees. More precisely the empirical risk is penalized by a complexity term, which is linear in the number of leaves of the tree:

$$\hat{R}_{pen}(T) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{T(X_i^1, \dots, X_i^p) \neq Y_i} + \alpha |T| \quad (2)$$

where  $\mathbb{I}$  is the indicator function,  $n$  the total number of observations,  $\alpha$  a positive penalty constant,  $|T|$  denotes the number of leaves of the tree  $T$  and  $Y_i$  is the  $i$ th random realization of  $Y$ .

The *R* package *rpart* provides both the sequence of subtrees pruned from a deep maximal tree and a final tree selected from this sequence by using the 1-SE rule (see [4]). The maximal tree is constructed by using the Gini index (default) and stops when the minimum number of observations in a leaf is reached, or if the misclassification rate of the branch provided by splitting a node is too small. The penalized criterion used in the pruning of *rpart* is  $\hat{R}_{pen}$  defined by (2). The cost complexity parameter denoted by  $cp$  corresponds to the temperature  $\alpha$  used in the original penalized criterion (2) divided by the misclassification rate of the root of the tree. The pruning step leads to a sequence  $\{T_1; \dots; T_K\}$  of nested subtrees (where  $T_K$  is reduced to the root of the tree) associated with a nondecreasing sequence of temperatures  $\{cp_1; \dots; cp_K\}$ . Then, the selection step is based on the 1-SE rule: using cross-validation, *rpart* computes 10 estimates of each prediction error  $R(T_k)$ , leading to average misclassification errors  $\{\hat{R}_{cv}(T_1); \dots; \hat{R}_{cv}(T_K)\}$  and their respective standard deviations

$\{SE(T_1); \dots; SE(T_K)\}$ . Finally, the selected subtree corresponds to the maximal index  $k_1$  such that

$$\hat{R}_{cv}(T_{k_1}) \leq \hat{R}_{cv}(T_{k_0}) + SE(T_{k_0}),$$

where  $\hat{R}_{cv}(T_{k_0}) = \min_{1 \leq k \leq K} \hat{R}_{cv}(T_k)$ .

## 2.2 Influence measures for CART

Let  $X = (X^1, \dots, X^p) \in \mathcal{X}$  be the vector of the explanatory variables, and consider that the data are independent realizations  $\mathcal{L} = \{(x_1, y_1); \dots; (x_n, y_n)\}$  of  $(X, Y) \in \mathcal{X} \times \{1; \dots; J\}$ . The dependent variable  $Y$  is assumed to be a categorical variable with  $J$  unordered categories. For a given tree  $T = T(\mathcal{L})$ , we denote any node of  $T$  by  $t$ . Hence, we use the following notations:

- $\tilde{T}$  the set of leaves of  $T$ , and  $|T|$  the number of leaves of  $T$ ,
- the empirical distribution  $p_{x,T}$  of  $Y$  conditionally to  $X = x$  and  $T$ , is defined by: for  $j = 1, \dots, J$ ,  $p_{x,T}(j) = p(j|t)$  the proportion of the class  $j$  in the leaf of  $T$  in which  $x$  falls.

We denote by  $T$  the tree obtained from the complete sample  $\mathcal{L}$ , while  $(T^{(-i)})_{1 \leq i \leq n}$  denote jackknife trees obtained from  $(\mathcal{L} \setminus \{(X_i, Y_i)\})_{1 \leq i \leq n}$ .

For a given tree  $T$ , two different main aspects are of interest: the predictions delivered by the tree or the partition associated with the tree, the second highlights the tree structure while the first focuses on its predictive performance. This distinction is classical and already examined for example by Miglio and Soffritti (2004) [14] recalling some proximity measures between classification trees and promoting the use of a new one mixing the two aspects. We then derive two kinds of influence measures for CART based on jackknife trees: influence on predictions and influence on partitions. In addition, the analysis is extended to the pruned sequences of CART trees to produce a CART specific notion of influence.

Our main purpose is to provide descriptive statistics for CART classification trees. The six proposed indices are complementary and give some insight on particular data. Last part of this section is devoted to a discussion about the pros and cons of each index.

Influence analysis is a classical tool in data analysis. For example, discriminant analysis has been studied by Campbell (1978) [6], Critchley and Vitiello (1991) [8] for the linear case and Croux and Joossens (2005) [9] for the quadratic one. For linear discrimination influence functions on the error rate, or the prediction error of binary classifiers, were considered in [10, 11]. Extension to nonparametric supervised classification is not easy since it is difficult to obtain an exact form for influence function.

### 2.2.1 Influence on predictions

We propose three influence measures based on predictions.

The first, closely related to the resubstitution estimate of the prediction error (replacing the response by its prediction provided by the reference tree), evaluates the impact of a single change on all the predictions, is defined by: for  $i = 1, \dots, n$

$$I_1(x_i) = \sum_{k=1}^n \mathbb{1}_{T(x_k) \neq T^{(-i)}(x_k)}, \quad (3)$$

*i.e.*  $I_1(x_i)$  is the number of observations for which the predicted label changes using the jackknife tree  $T^{(-i)}$  instead of the reference tree  $T$ . The second, closely related to the leave-one-out estimate of the prediction error, is: for  $i = 1, \dots, n$

$$I_2(x_i) = \mathbb{1}_{T(x_i) \neq T^{(-i)}(x_i)}, \quad (4)$$

*i.e.*  $I_2(x_i)$  indicates if  $x_i$  is classified in a different way by  $T$  and  $T^{(-i)}$ . This index is obtained like the leave-one-out estimate but the response is replaced by its prediction provided by the reference tree.

The third one measures the distance between the distribution of the label in the nodes where  $x_i$  falls: for  $i = 1, \dots, n$

$$I_3(x_i) = d(p_{x_i, T}, p_{x_i, T^{(-i)}}), \quad (5)$$

where  $p_{x_i, T}$  is  $p_{x, T}$  for  $x = x_i$  and  $d$  is a distance (or a divergence) between probability distributions.

$I_1$  and  $I_2$  are based on the predictions only while  $I_3$  is based on the distribution of the labels in each leaf.

To compute  $I_3$ , several distances or divergences can be used. In this

paper, we use the total variation distance: for  $p$  and  $q$  two distributions on  $\{1; \dots; J\}$ ,

$$d(p, q) = \max_{A \in \{1; \dots; J\}} |p(A) - q(A)| = 2^{-1} \sum_{j=1}^J |p(j) - q(j)|$$

**Remark 1.** The Kullback-Leibler divergence (relied to the Shannon entropy index) and the Hellinger distance (relied to the Gini index) can also be used instead of the total variation distance.

### 2.2.2 Influence on partitions

We propose two influence measures based on partitions:  $I_4$  measuring the variations on the number of leaves in each partition, and  $I_5$  based on the quantification of the difference between the two partitions. These indices are computed in the following way: for  $i = 1, \dots, n$

$$I_4(x_i) = |T^{(-i)}| - |T|, \quad (6)$$

$$I_5(x_i) = 1 - J(\tilde{T}, \tilde{T}^{(-i)}), \quad (7)$$

where  $J(\tilde{T}, \tilde{T}^{(-i)})$  is the Jaccard similarity between the partitions of  $\mathcal{L}$  respectively defined by  $\tilde{T}^{(-i)}$  and  $\tilde{T}$ .

Recall that, for two partitions  $C_1$  and  $C_2$  of  $\mathcal{L}$ , the Jaccard coefficient  $J(C_1, C_2)$  is computed as

$$J(C_1, C_2) = \frac{a}{a + b + c},$$

where  $a$  counts the number of pairwise points of  $\mathcal{L}$  belonging to the same leaf in both partitionings,  $b$  the number of pairwise points belonging to the same leaf in  $C_1$ , but not in  $C_2$ , and  $c$  the number of pairwise points belonging to the same leaf in  $C_2$ , but not in  $C_1$ . The more similar  $C_1$  and  $C_2$ , the closer  $J(C_1, C_2)$  to 1.

**Remark 2.** Such influence index can be generated using different similarities between partitions. For a detailed analysis and comparisons, see [18].

### 2.2.3 Influence based on subtrees sequences

Another way to inspect the dataset is to consider the complexity cost constant, penalizing bigger trees in the pruning step of the CART tree design, as a tuning parameter. It allows to scan the data and sort them with respect to their influence on the CART tree.

Let us consider on the one hand the sequence of subtrees based on the complete dataset, denoted by  $T$ , and on the other hand the  $n$  jackknife sequences of subtrees based on the jackknife subsamples  $\mathcal{L} \setminus \{(X_i, Y_i)\}$ , denoted by  $T_{cp_j}^{(-i)}$ . Suppose that the sequence  $T$  contains  $K_T$  elements, and that each sequence  $T_{cp_j}^{(-i)}$  contains  $K_{T^{(-i)}}$  elements ( $i = 1, \dots, n$ ). This leads to a total of  $N_{cp} \leq K_T + \sum_{1 \leq i \leq n} K_{T^{(-i)}}$  distinct values  $\{cp_1; \dots; cp_{N_{cp}}\}$  of the cost complexity parameter in increasing order from  $cp_1$  to  $cp_{N_{cp}} = \max_{1 \leq j \leq N_{cp}} cp_j$ .

Then, for each value  $cp_j$  of the complexity and each observation  $x_i$ , we compute the binary variable  $\mathbb{1}_{T_{cp_j}(x_i) \neq T_{cp_j}^{(-i)}(x_i)}$  that tells us if the reference and jackknife subtrees corresponding to the same complexity  $cp_j$  provide different predicted labels for the removed observation  $x_i$ . Thus we define influence measure  $I_6$  as the number of complexities for which these predicted labels differ: for  $i = 1, \dots, n$

$$I_6(x_i) = \sum_{j=1}^{N_{cp}} \mathbb{1}_{T_{cp_j}(x_i) \neq T_{cp_j}^{(-i)}(x_i)}. \quad (8)$$

**Remark 3.** Since the jackknife sequences of subtrees do not change for many observations, usually we obtain that  $N_{cp} \ll K_T + \sum_{1 \leq i \leq n} K_{T^{(-i)}}$ .

### 2.2.4 Comparison of the influence measures

$I_2$  is a local index while  $I_1$  is a global index taking into account the whole sample.  $I_6$  is a weighted version of  $I_2$  since  $I_6$  is a sum of the  $I_2$  over the various values of  $c_p$ . Index  $I_3$  can be viewed as a generalization of  $I_1$  and  $I_2$ . The choice of total variation is subjective as stated in Remark 1.

Change in the CART topology is quantified by  $I_4$  and  $I_5$ .  $I_4$  is the difference between the number of leaves of the reference tree and the number of leaves of the jackknife tree. As noted in Remark 2 the Jaccard index can be replaced by any similarity measure.

The range of possible values of  $I_2$  and  $I_4$  is low while  $I_1$ ,  $I_5$  and  $I_6$  are more variable.

These various indices are complementary and easy to compute. They give different kind of information and detection of outlier should be based on the study of the various indices even if, from a practical point of view, information on prediction is generally more useful than information on partition of CART classification tree.

## 3 Illustration on Spam Dataset

### 3.1 Spam dataset

The *spam* data, publicly available at `ftp.ics.uci.edu`, consists of information from 4601 email messages, in a study to screen email for "spam" (i.e. junk email). The data are presented in details in [13, p. 301]. The response variable is binary, with values nonspam or spam, and there are 57 predictors: 54 given by the percentage of words in the email that match a given word or a given character, and three related to uninterrupted sequences of capital letters: the average length, the length of the longest one and the sum of the lengths of uninterrupted sequences. The objective was to design an automatic spam detector that could filter out spam before clogging the users' mailboxes. This is a supervised learning problem.

### 3.2 Reference tree and jackknife trees

The reference tree is obtained using the *R* package *rpart* (see [15], [19]) and accepting the default values for all the parameters (mainly the Gini heterogeneity function to grow the maximal tree and pruning thanks to 10-fold cross-validation).

The reference tree based on all observations, namely  $T$ , is given in Figure 1. This tree has 7 leaves, obtained from splits involving the variables *charDollar*, *remove*, *hp*, *charExclamation*, *capitalTotal*, and *free* (in order of appearance in the tree). Each leaf is labeled by the prediction of  $Y$  (spam or nonspam) and the distribution of  $Y$  inside the node (for example, the third leaf is almost pure: it contains 1 nonspam and 20 spams). These make the tree easy to interpret and to describe: indeed, from the direct interpretation of the path from the root to the third leaf: an email containing many occurrences of

”!” and ”free” is almost always a spam.

To compute the influence of one observation we compute  $T^{(-i)}$  the tree based on all observations except the observation  $i$  ( $i = 1, \dots, 4601$ ). We then obtain a collection of 4601 trees, which can be described with respect to the reference tree  $T$  in many different ways. This description is carried out according to various filters, namely: the retained variables, the number of leaves of the final tree, the observations involved in the differences. The variables *charDollar*, *charExclamation*, *hp* and *remove* are always present. The variable *capitalLong* is present in 88 trees, the variable *capitalTotal* is present in 4513 trees and the variable *free* is present in 4441 trees. This indicates instable clades within  $T$ . In addition variables *free* and *capitalLong* as well as *capitalLong* and *capitalTotal* never appear simultaneously, while the 4441 trees containing *free* also contain *capitalTotal*. This highlights the variability among the 4601 trees. All the differences are explained by the presence (or not) of the variable *free*. Indeed, removing one observation from node generated from variable *free* is enough to remove the split and to merge the two nodes leading to misclassification. There are 77 observations  $x_i$  classified differently by  $T$  and the corresponding jackknife tree  $T^{(-i)}$ , and 160 jackknife trees with one less leaf than  $T$ . All the other jackknife trees have the same number of leaves as  $T$ . A careful examination of the highlighted observations leads to a spam and a nonspam mails that define the second split of the reference tree: the threshold on the variable *remove* is the middle of their corresponding values.

### 3.3 Influence measures

The influence measures  $I_1$ ,  $I_2$ ,  $I_3$ ,  $I_4$ ,  $I_5$  and  $I_6$  respectively defined by (3), (4), (5), (6), (7) and (8) are computed on the *spam* dataset by using the jackknife trees computed in paragraph 3.2. The results are summarized in the following paragraphs.

#### 3.3.1 Influence on predictions

Indices  $I_1$  and  $I_3$  computed on the 163 observations of the *spam* dataset for which  $I_1$  is nonzero are given in Figure 2.

There are 77 observations for which  $I_2 = 1$ , that is for which  $x_i$  is classified differently by  $T$  and the corresponding jackknife tree  $T^{(-i)}$ ,

Figure 1: CART reference tree for the *spam* dataset.



Figure 2: Influence indices based on predictions for the *spam* dataset.

while 163 jackknife trees lead to a nonzero number of observations for which the predicted label changes. These observations correspond to observations leading to jackknife trees sufficiently perturbed from  $T$  to have a different shape. The 77 aforesaid observations lead to a total variation distance between  $p_{x_i, T}$  and  $p_{x_i, T(-i)}$  larger than 0.6. The others lead to a total variation distance smaller than 0.1. The 86

remaining jackknife trees for which the number of observations differently labeled is nonzero provide total variation distances from 0.016 to 0.1.

$I_1$  and  $I_2$  are based on the predictions only while  $I_3$  takes into account the distribution of the labels in each leaf. For example, the contribution of  $I_3$  with respect to  $I_2$  is that some observations are classified similarly by  $T$  and  $T^{(-i)}$ , but actually lead to conditional probability distributions  $p_{x_i, T^{(-i)}}$  largely different from  $p_{x_i, T}$ . These observations are not sufficiently important in the construction of  $T$  to perturb the classification, but play some role in the tree instability.

### 3.3.2 Influence on partitions

There are 160 jackknife trees having one less leaf than  $T$ , and 163 leading to a partition  $\tilde{T}^{(-i)}$  different from  $\tilde{T}$ . Among the 160 aforesaid trees, 139 lead to a partition  $\tilde{T}^{(-i)}$  of dissimilarity larger than 0.05. Hence there are 21 trees with one less leaf than  $T$ , but leading to a partition  $\tilde{T}^{(-i)}$  not far from  $\tilde{T}$ . The others perturb sufficiently  $T$  to change the partition consequently. Let us note that all jackknife trees partitions are of dissimilarity smaller than 0.06 from  $\tilde{T}$ . This is due to the very local perturbations around  $x_i$ .

Let us emphasize that the 163 trees leading to a partition  $\tilde{T}^{(-i)}$  different from  $\tilde{T}$  correspond exactly to the 163 trees leading to a nonzero number of mails classified differently. This shows an unexpected behavior of CART on this dataset: different partitions lead to predictors assigning different labels on the training sample.

### 3.3.3 Influence based on subtrees sequences

The pruned subtrees sequences contain around six elements and they represent  $N_{cp} = 27$  distinct values  $\{cp_1; \dots; cp_{27}\}$  of the cost complexity parameter (from 0.01 to 0.48).

The distribution of influence index  $I_6$  is given in Table 1.

$I_6$	0	1	2	3	4	7	12	13	14	17	18	21
Nb. Obs.	2768	208	1359	79	62	1	1	66	30	2	23	2

Table 1: Frequency distribution of the influence index  $I_6$  over the 4601 emails.

The number of actual values of  $I_6$  is small. These values organize the data as nested level sets in decreasing order of influence. For 60% of the observations, predictions are the same all along the pruned subtrees sequences, making these observations not influential for  $I_6$ . There are 123 observations leading to different predictions for at least half of the pruned subtrees, and 2 observations change prediction labels of trees for 21 complexities (among the 27 cps). Let us remark that these two most influential mails for  $I_6$  are the same mails influential for  $I_2$  and  $I_4$ .

Index  $I_6$  can be examined in a structural way by locating the observation on the topology of the reference tree. Of course, graphical software tools based on such idea can be useful to screen a given data set.

**Remark 4.** Let us recall that the influence is measured with respect to a reference model and of course more instable reference tree automatically increases the number of individuals that can be categorized as influential. In fact, even if it is implicit, influence measures are relative to a given model. Typically, increasing the number of leaves of the reference tree automatically promotes new observations as influential.

## 4 Exploring the Paris Tax Revenues dataset

### 4.1 Dataset and reference tree

#### 4.1.1 Dataset

We apply the tools presented in the previous section on tax revenues of households in 2007 from the 143 cities surrounding Paris. Cities are grouped into four counties (corresponding to the french administrative “département”). The PATARE data (PAris TAx REvenues) are freely available on <http://www.data-publica.com/data>. For confidentiality reason we do not have access to the tax revenues of the individual households but we have characteristics of the distribution of the tax revenues per city. Paris has 20 ”arrondissements” (corresponding to districts), Seine-Saint-Denis is located at the north of Paris and has 40 cities, Hauts-de-Seine is located at the west of Paris and has 36 cities, and Val-de-Marne is located at the south of Paris and has 48 cities. For each city, we have the first and the 9th deciles (named respectively

$D_1$  and  $D_9$ ), the quartiles (named respectively  $Q_1$ ,  $Q_2$  and  $Q_3$ ), the mean, and the percentage of the tax revenues coming from the salaries and treatments (named  $PtSal$ ). Figure 3 gives the summary statistics for each variable per county.

Basically we tried to predict the county of the city with the characteristics of the tax revenues distribution. This is a supervised classification problem where the explained variable is quaternary.

We emphasize that this information cannot be easily retrieved from the explanatory variables considered without the county information. Indeed, the map (see Figure 4) of the cities drawn according to a  $k$ -means ( $k=4$ ) clustering (each symbol is associated with a cluster) superimposed with the borders of the counties, exhibits a poor recovery of counties through clusters.

#### 4.1.2 Reference tree

Figure 5 shows the reference tree. Each leaf is labelled by two informations: the predicted county and the distribution of the cities over the 4 counties. For example the second leaf gives 0/17/1/3 meaning that it contains 0 districts from Paris, 17 cities from Hauts-de-Seine, 1 from Seine-Saint-Denis and 3 from Val-de-Marne.

All the 5 terminal nodes located on the left subtree below the root are homogeneous since the distributions are almost pure, while half the nodes of the right subtree are highly heterogeneous. The first split involves the last decile of the income distribution and then discriminates the counties with rich cities from counties mainly constituted with poorer cities. More precisely, the labels mainly distinguish Paris and Hauts-de-Seine on the left from Seine-Saint-Denis on the right, while Val-de-Marne appears in both sides.

The extreme quantiles are sufficient to separate richest from poorest counties while more global predictors are useful to further discriminate between intermediate cities. Indeed the splits on the left part are mainly based on the deciles  $D_1$ ,  $D_9$  and  $PtSal$  is only used to separate Hauts-de-Seine from Val-de-Marne. The splits on the right part are based on all the dependent variables but involve  $PtSal$  and mean variables to separate Seine-Saint-Denis from Val-de-Marne. Let us notice that predictors  $Q_1$  and  $Q_2$  do not appear.

Surprisingly, the predictions given by the reference tree are generally correct (the resubstitution misclassification rate calculated from the

confusion matrix given in Table 2, is equal to 24.3%). Since the cities within each county are very heterogeneous, we look for the cities which perturb the reference tree.

Predicted		Paris	Haut de Seine	Seine Saint Denis	Val de Marne
Actual	Paris	20	0	0	0
	Haut de Seine	0	30	1	5
Seine Saint Denis		1	4	28	7
Val de Marne		3	9	5	30

Table 2: Confusion matrix: actual versus predicted county, using the CART reference tree.

After this quick inspection of the reference tree avoiding careful inspection of the cities inside the leaves, let us focus on the influential cities highlighted by the previously defined indices. In the sequel the cities (which are the individuals) are written in italics to be clearly distinguished from counties which are written between quotation marks.

## 4.2 Influential observations

### 4.2.1 Presentation

There are 44 cities classified differently by  $T$  and the corresponding jackknife tree  $T^{(-i)}$ , and 44 jackknife trees having a different number of leaves than  $T$ . The frequency distribution of the difference in the number of leaves, summarized by the influence index  $I_4$ , is given in Table 3. The aforesaid cities are given in Table 5 of the appendix, classified by their respective labels in the dataset. **DC** denotes the set of cities classified differently by  $T$  and  $T^{(-i)}$ , and **DNF** denotes the set of cities for which  $T^{(-i)}$  has not the same number of leaves as  $T$ . Let us emphasize that the cardinality of  $\mathbf{DC} \cup \mathbf{DNF}$  is equal to 63: 19 cities are classified differently by trees having the same number of leaves, while 18 cities lead to jackknife trees having different number of leaves, but classifying the corresponding cities in the same way.

Indices  $I_1$  and  $I_3$  computed on the 75 observations of the PATARE dataset for which  $I_1$  is nonzero are given in Figure 6.

$I_4$	-3	-2	-1	0	1
Nb. Obs.	1	8	25	99	10

Table 3: Frequency distribution of the influence index  $I_4$  over the 143 cities.

There are 44 observations classified differently by  $T$  and  $T^{(-i)}$ , while 75 jackknife trees lead to a nonzero number of observations for which predicted labels change. These 75 jackknife trees contain the 63 trees for which the number of leaves changes or classifying the corresponding city differently. There are 13 cities for which the total variation distance between the distributions defined by  $T$  and  $T^{(-i)}$  respectively is larger than 0.5. Among these 13 cities, 2 lead to jackknife trees at distance 1 from  $T$ : *Asnieres sur Seine* (from “Hauts-de-Seine”) and *Paris 13eme* (from “Paris”). The value of  $I_1$  and  $I_2$  at these 2 points is equal to 1, meaning that each city provides a jackknife tree sufficiently close to  $T$  to unchange the classification, except for the removed city. In fact, if we compare the 2 jackknife trees with  $T$ , we can notice that the thresholds in the second split for *Asnieres sur Seine*, and in the first split for *Paris 13eme*, are slightly moved. It suffices to classify on the one hand *Asnieres sur Seine* in the pure leaf “Paris”, and on the other hand to remove *Paris 13eme* from this leaf to classify it in “Seine-Saint-Denis”. This explains the astonishing value of 1 for the corresponding total variation distances.

Influence index  $I_5$  on the 44 observations of the PATARE dataset for which  $I_4$  is nonzero is given in Figure 7.

There are 44 observations leading to jackknife trees having number of leaves different from  $T$ . Two cities lead to jackknife trees providing partitions at distance larger than 0.5 from  $T$ : *Neuilly Plaisance* and *Villemonble* (both from “Seine-Saint-Denis”). When removed, these 2 cities change drastically the value of the threshold in the first split, what implies also drastic changes in the rest of the tree. Moreover, the corresponding jackknife trees have 2 less leaves than  $T$ , what obviously increases the Jaccard similarity between  $T$  and each jackknife tree.

The frequency distribution of influence index  $I_6$  over the 143 cities of the PATARE dataset is given in Table 4.

There are 29 different values of complexities in the reference and jac-

$I_6$	0	1	2	3	4	6	7	9	10	12	13	14	16	17	21	24	26
Nb. Obs.	7	44	10	17	9	2	14	5	1	3	3	10	7	6	2	1	2

Table 4: Frequency distribution of influence index  $I_6$  over the 143 cities.

cknife trees sequences. Two cities change prediction labels of trees for 26 complexities: *Asnieres-sur-Seine* and *Villemomble*. In the decreasing order of influence, one city changes labels of trees for 24 complexities, and 2 cities for 21 complexities: *Paris 13eme*, and *Bry-sur-Marne* (from “Val-de-Marne”), *Rueil-Malmaison* (from “Hauts-de-Seine”). These 5 observations change labels for more than 72.4% of the complexities. 61 observations change labels of trees for less than 6.9% of the complexities.

Let us notice that *Asnieres-sur-Seine* and *Paris 13eme* have already been selected as influential for  $I_3$ , and *Villemomble* for  $I_5$ . Nevertheless, the behaviours of  $I_1$ ,  $I_2$  and  $I_3$  at points *Bry-sur-Marne* and *Rueil-Malmaison* are comparable to behaviours at points *Asnieres-sur-Seine* and *Paris 13eme*:  $I_1$  and  $I_2$  are equal to 1, and  $I_3$  is equal to 0.66, meaning that these 2 cities belong to the 13 cities for which  $I_3$  is larger than 0.5. Let us also remark that only *Montreuil* (from “Seine-Saint-Denis”) is among the 13 aforesaid cities, but not in the 26 cities listed above and selected as influential for  $I_4$  and  $I_6$ . The value of  $I_3$  at this point is equal to 0.52.

#### 4.2.2 Interpretation

One can find in Figure 8 the influential cities, with respect to the two indices  $I_4$  and  $I_6$ , located in the reference tree. Let us notice that most of the selected cities have also been selected by influence indices  $I_1$ ,  $I_2$ ,  $I_3$  and  $I_5$ , so we refer only to  $I_4$  and  $I_6$  in what follows.

Let us emphasize that only three cities among the 26 influential cities quoted in Figure 8 are misclassified when using the reference tree.

Index  $I_4$  highlights cities (*Noisy-le-Grand*, *Bagnoeux*, *Le Blanc Mesnil*, *Le Bourget*, *Neuilly-Plaisance*, *Noisy-le-Sec*, *Sevran*, *Vaujours* and *Villemomble*) far from Paris and of middle or low social level. All the cities having an index of -3 or -2 are located in nodes of the right part of the reference tree whereas the rich cities are concentrated in the

left part.

Index  $I_6$  highlights cities for which two parts of the city can be distinguished: a popular one with a low social level and a rich one of high social level. They are located in the right part of the reference tree (for the higher values of  $I_6 = 26, 24$  and  $21$ : *Asnieres sur Seine, Villemomble, Paris 13eme, Bry-sur-Marne* and *Rueil-Malmaison*) as well as in the left part (for moderate values of  $I_6 = 16$  and  $17$ : *Chatenay-Malabry, Clamart, Fontenay aux Roses, Gagny, Livry-Gargan, Vanves, Chevilly-Larue, Gentilly, Le Perreux sur Marne, Le Pre-Saint-Gervais, Maisons-Alfort, Villeneuve-le-Roi, Vincennes* and the particularly interesting city *Villemomble* for  $I_6 = 26$ ). Indeed, we can notice that only *Villemomble* is highlighted both by  $I_4$  and  $I_6$ .

To explore the converse, we inspect now the list of the 51 cities associated with lowest values of  $I_6$  ( $0$  and  $1$ ) which can be considered as the less influential, the more stable. It can be easily seen that it corresponds to the 16 rich district of Paris downtown (*Paris 1er* to *12eme* and *Paris 14eme* to *Paris 16eme*) and mainly cities near Paris or directly connected by the RER line transportation.

It should be noted that the influence indices cannot be easily explained neither by central descriptors like the mean or the median Q2 nor by dispersion descriptors as Q3-Q1 and D9-D1. Bimodality seems the key property to explain high values of the influence indices.

In addition, coming back to the non supervised analysis, one may notice that influential observations for PCA (Principal Component analysis) are not related to influential cities detected using  $I_6$  index. Indeed, Figure 9 contains the two first principal components capturing more than 95% of the total variance. PCA has been performed on the correlation matrix of all the predictors. Each city is represented in this plane by a symbol of size proportional to the  $I_6$  index. Hence one can see that the points influential for PCA (those far from the origin) are generally of small influence for influence index  $I_6$ .

#### 4.2.3 Spatial interpretation

To end this study, a map is useful to capture the spatial interpretation and complement the previous comments which need some prior knowledge about the sociology of the Paris area. In Figure 10, the 143 cities are represented by a circle proportional to the influence index

$I_6$  and a spatial interpolation is performed using 4 grey levels. This map shows that Paris is stable, and that each surrounding county contains a stable area: the richest or the poorest cities. What is remarkable is that the white as well as the gray areas are clustered.

## 5 Perspectives

Two directions for future work can be sketched.

First, the tools developed in this paper for the classification case can be generalized for the regression case. The instability is smoother in the regression case since the data are adjusted thanks to a surface rather than a frontier. Then the number of false classifications is replaced with the sum of square residuals typically which is more sensitive to perturbations but the differences between the full tree and the jackknife ones are more stringent in the classification case. Some classical problems, like outlier detection has been intensively studied in the regression case and a lot of solutions have been developed around the ideas of robust regression (see Rousseeuw (1984) [16]). A complete scheme for comparison can be retrieved from Chèze and Poggi [7], where a tree-based algorithm has been developed and compared intensively to well known competitive methods including robust regression.

Another direction is to focus on model stability and robustness rather than centering the analysis around individuals. A first idea could be, following Bar-Hen *et al.* (2008) [1], to build the most robust tree by iteratively remove the most influential observation until stabilisation between reference and jackknife trees. A second one is to consider, starting from the  $I_6$  index but summing on the observations instead of the complexities, the percentage of observations differently classified by the reference and jackknife subtrees at fixed complexity. This is out of the scope of this article.

Finally we proposed influence measures for CART classification tree but this work could be extended to resampled version of CART, for example bagged tree.

## 6 Appendix

In Table 5 one can find, for each county, three sets of cities for the PATARE dataset for which the jackknife tree differs from the reference tree.

	<b>DNF</b> $\cap$ <b>DC</b>	<b>DNF</b> $\setminus$ <b>DC</b>	<b>DC</b> $\setminus$ <b>DNF</b>
<b>Hauts-de-Seine</b>	Boulogne Billancourt, Bourg la Reine, Clichy, Garches, Meudon, Neuilly sur Seine, Saint Cloud, Sceaux, Ville d'Avray.	Bagnoeux.	Asnieres sur Seine, Chatenay Malabry, Clamart, Fontenay aux Roses, Nanterre, Rueil Malmaison, Vanves.
<b>Paris</b>	Paris 18eme, Paris 19eme, Paris 20eme.	$\emptyset$	Paris 13eme.
<b>Seine-Saint-Denis</b>	Le Blanc Mesnil, Le Bourget, Neuilly sur Marne, Noisy le Grand, Noisy le Sec, Sevran, Villemomble, Villepinte.	Aulnay sous Bois, Neuilly Plaisance, Rosny sous Bois, Tremblay en France, Vaujours.	Gagny, Le Pre Saint Gervais, Livry Gargan, Montreuil.
<b>Val-de-Marne</b>	Alfortville, Chevilly Larue, Gentilly, Le Perreux sur Marne, Vincennes.	Arcueil, Bonneuil sur Marne, Cachan, Champigny sur Marne, Choisy le Roi, Creteil, Fontenay sous Bois, Mandres les Roses, Orly, Saint Maurice, Villejuif, Vitry sur Seine.	Boissy Saint Leger, Bry sur Marne, Limeil Brevennes, Maisons Alfort, Valenton, Villeneuve le Roi, Villiers sur Marne.

Table 5: **DNF**: for the 4 counties, cities from the PATARE dataset for which the corresponding jackknife tree  $T^{(-i)}$  has not the same number of leaves as CART reference tree  $T$ . **DC**: cities classified differently by  $T$  and  $T^{(-i)}$ .

## References

- [1] Bar-Hen, A., Mariadassou, M., Poursat, M.-A. and Vandenkorrenhuyse, Ph. (2008). *Influence Function for Robust Phylogenetic Re-*

*constructions.* Molecular Biology and Evolution, 25(5), 869-873.

- [2] Bel, L., Allard, D., Laurent, J.M., Cheddadi, R. and Bar-Hen, A. (2009). *CART algorithm for spatial data: application to environmental and ecological data.* Computat. Stat. and Data Anal., 53(8), 3082-3093.
- [3] Bousquet, O., Elisseeff, A. (2002). *Stability and generalization.* J. Machine Learning Res. 2, 499–526.
- [4] Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. *Classification and regression trees.* Chapman & Hall (1984).
- [5] Briand, B., Ducharme, G. R., Parache, V. and Mercat-Rommens, C. (2009). *A similarity measure to assess the stability of classification trees,* Comput. Stat. Data Anal., 53(4), 1208–1217.
- [6] Campbell, N.A. (1978). *The influence function as an aid in outlier detection in discriminant analysis.*, Appl. Statist., 27, 251–258.
- [7] Chèze, N. and Poggi, J.M. (2006). *Outlier detection by boosting regression trees.* Journal of Statistical Research of Iran (JSRI), 3, 1–21.
- [8] Critchley, F. and Vitiello, C. (1991). *The influence of observations on misclassification probability estimates in linear discriminant analysis.*, Biometrika, 78, 677–690.
- [9] Croux, C. and Joossens, K. (2005). *Influence of observations on the misclassification probability in quadratic discriminant analysis.*, Journal of Multivariate Analysis, 96(2), 384–403.
- [10] Croux, C., Filzmoser, P., and Joossens, K. (2008). *Classification Efficiencies for Robust Linear Discriminant Analysis*, Statistica Sinica, 18(2), 581–599
- [11] Croux, C., Haesbroeck, G., and Joossens, K. (2008). *Logistic Discrimination using Robust Estimators: an influence function approach*, The Canadian Journal of Statistics, 36(1), 157–174.
- [12] Gey, S. and Poggi, J.M. (2006). *Boosting and instability for regression trees.* Comput. Stat. Data Anal., 50(2), 533-550.
- [13] Hastie, T.J., Tibshirani, R.J. and Friedman, J.H. (2009). *The elements of statistical learning: data mining, inference and prediction.* Third edition, Springer, New-York.

- [14] Miglio, R. and Soffritti, G. (2004). *The comparison between classification trees through proximity measures*. Comput. Stat. Data Anal., 45(3), 577–593.
- [15] R Development Core Team *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0 (2009). URL <http://www.R-project.org/>.
- [16] Rousseeuw, P. (1984). *Least median of squares regression*, J. Amer. Statist. Assoc., 79, 871-880.
- [17] Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. Wiley, Interscience, New York.
- [18] Youness, G. and Saporta, G. (2009). *Comparing partitions of two sets of units based on the same variables*. Advances in Data Analysis and Classification, 4(1), 53-64.
- [19] Venables, W. N., and Ripley, B.D. (2002). *Modern Applied Statistics with S*. Fourth Edition, Springer.
- [20] Verboven, S. and Hubert, M. (2005). *LIBRA: a MATLAB library for robust analysis*, Chemometrics and Intelligent Laboratory Systems, 75, 127-136.
- [21] Zhang, H. and Singer, B. H. (2010). *Recursive Partitioning and Applications*, 2<sup>nd</sup> edition, Springer.

Figure 3: PATARE dataset: boxplots of the variables per county (“département” in France) and zip codes for counties.

Figure 4: Spatial representation of  $k$ -means ( $k=4$ ) clustering of the PATARE dataset cities (each symbol is associated with a cluster).

Figure 5: CART reference tree for the PATARE dataset.

Figure 6: Influence indices based on predictions for PATARE dataset cities.

Figure 7: Influence index based on partitions for PATARE dataset cities.

Figure 8: Influential cities located on the CART reference tree.

Figure 9: Plane of the two first principal components: Cities are represented by symbols proportional to influence index  $I_6$ .

Figure 10: The 143 cities are represented by a circle proportional to the influence index  $I_6$  and a spatial interpolation is performed using 4 grey levels.

# Chapter 7: Dummy variable regression

Why include a qualitative independent variable? . . . . .	2
<b>Simplest model</b>	<b>3</b>
Simplest case . . . . .	4
Example (continued) . . . . .	5
Possible solution: separate regressions . . . . .	6
Independent variable vs. regressor . . . . .	7
Common slope model . . . . .	8
Testing . . . . .	9
<b>More general models</b>	<b>10</b>
More than one quantitative independent variable . . . . .	11
Polytomous independent variables . . . . .	12
Example (continued) . . . . .	13
Testing with polytomous independent variable . . . . .	14
R commands . . . . .	15
More than one qualitative independent variable . . . . .	16
<b>Interaction</b>	<b>17</b>
Definition . . . . .	18
Interaction vs. correlation . . . . .	19
Constructing regressors . . . . .	20
Testing . . . . .	21
Principle of marginality . . . . .	22
Polytomous independent variables . . . . .	23
Hypothesis tests . . . . .	24
Standardized estimates . . . . .	25
Interaction between categorical variables . . . . .	26

## Why include a qualitative independent variable?

- We are interested in the effect of a qualitative independent variable (for example: do men earn more than women?)
- We want to better predict/describe the dependent variable. We can make the errors smaller by including variables like gender, race, etc.
- Qualitative variables may be confounding factors. Omitting them may cause biased estimates of other coefficients.

2 / 26

## Simplest model

3 / 26

### Simplest case

- Example:
  - ◆ Dependent variable: income
  - ◆ One quantitative independent variable: education
  - ◆ One dichotomous (can take two values) independent variable: gender
- Assume effect of either independent variable is the same, regardless of the value of the other variable (additivity, parallel regression lines) - See pictures from book.
- Usual assumptions on statistical errors: independent, zero means, constant variance, normally distributed, fixed  $X$ 's or  $X$  independent of statistical errors.

4 / 26

### Example (continued)

- Suppose that we are interested in the effect of education on income, and that gender has an effect on income.
- See pictures from book.
- Scenario 1: Gender and education are uncorrelated
  - ◆ Gender is not a confounding factor
  - ◆ Omitting gender gives correct slope estimate, but larger errors
- Scenario 2: Gender and education are correlated
  - ◆ Gender is a confounding factor
  - ◆ Omitting gender gives biased slope estimate, and larger errors

5 / 26

### Possible solution: separate regressions

- Fit separate regression for men and women
- Disadvantages:
  - ◆ How to test for the effect of gender?
  - ◆ If it is reasonable to assume that regressions for men and women are parallel, then it is more efficient to use all data to estimate the common slope.

6 / 26

### Independent variable vs. regressor

- $Y = \text{income}$ ,  $X = \text{education}$ ,  $D = \text{regressor for gender}$ :

$$D_i = \begin{cases} 1 & \text{for men} \\ 0 & \text{for women} \end{cases}$$

- Independent variable = real variables of interest
- Regressor = variable put in the regression model
- In general, regressors are functions of the independent variables. Sometimes regressors are equal to the independent variables.

7 / 26

### Common slope model

- $Y_i = \alpha + \beta X_i + \gamma D_i + \epsilon_i$

- For women ( $D_i = 0$ ):

$$Y_i = \alpha + \beta X_i + \gamma \cdot 0 + \epsilon_i = \alpha + \beta X_i + \epsilon_i$$

- For men ( $D_i = 1$ ):

$$Y_i = \alpha + \beta X_i + \gamma \cdot 1 + \epsilon_i = (\alpha + \gamma) + \beta X_i + \epsilon_i$$

- See picture from book.

- What are the interpretations of  $\alpha$ ,  $\beta$  and  $\gamma$ ?

- What happens if we code  $D = 1$  for women and  $D = 0$  for men?

8 / 26

## Testing

- Test the partial effect of gender:
  - ◆  $H_0 : \gamma = 0, H_a : \gamma \neq 0$
  - ◆ Same as before:  
Compute  $t$ -statistic or incremental F-test
- Test the partial effect of education:
  - ◆  $H_0 : \beta = 0, H_a : \beta \neq 0$
  - ◆ Same as before:  
Compute  $t$ -statistic or incremental F-test
- Cystic fibrosis example.

9 / 26

## More general models

10 / 26

### More than one quantitative independent variable

- All methods go through, as long as we assume parallel regression surfaces.
- Model:  $Y_i = \alpha + \beta_1 X_{i1} + \dots + \beta_k X_{ik} + \gamma D_i + \epsilon_i$ .
- Women ( $D_i = 0$ ):

$$\begin{aligned} Y_i &= \alpha + \beta_1 X_{i1} + \dots + \beta_k X_{ik} + \gamma \cdot 0 + \epsilon_i \\ &= \alpha + \beta_1 X_{i1} + \dots + \beta_k X_{ik} + \epsilon_i \end{aligned}$$

- Men ( $D_i = 1$ ):

$$\begin{aligned} Y_i &= \alpha + \beta_1 X_{i1} + \dots + \beta_k X_{ik} + \gamma \cdot 1 + \epsilon_i \\ &= (\alpha + \gamma) + \beta_1 X_{i1} + \dots + \beta_k X_{ik} + \epsilon_i \end{aligned}$$

- Interpretation of  $\alpha, \beta_1, \dots, \beta_k, \gamma$ .

11 / 26

## Polytomous independent variables

- Qualitative variable with more than two categories
- Example: Duncan data:
  - ◆ Dependent variable:  $Y = \text{prestige}$
  - ◆ Quantitative independent variables:  $X_1 = \text{income}$  and  $X_2 = \text{education}$
  - ◆ Qualitative independent variable: type (bc, prof, wc)

- $D_1$  and  $D_2$  are regressors for type:

Type	$D_1$	$D_2$
Blue collar (bc)	0	0
Professional (prof)	1	0
White collar (wc)	0	1

- If there are  $p$  categories, use  $p - 1$  dummy regressors.  
What happens if we use  $p$  regressors?

12 / 26

## Example (continued)

- $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \epsilon$
- Blue collar ( $D_{i1} = 0$  and  $D_{i2} = 0$ ):

$$\begin{aligned} Y_i &= \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 \cdot 0 + \gamma_2 \cdot 0 + \epsilon_i \\ &= \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \end{aligned}$$

- Professional ( $D_{i1} = 1$  and  $D_{i2} = 0$ ):

$$\begin{aligned} Y_i &= \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 \cdot 1 + \gamma_2 \cdot 0 + \epsilon_i \\ &= (\alpha + \gamma_1) + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \end{aligned}$$

- White collar ( $D_{i1} = 0$  and  $D_{i2} = 1$ ):

$$\begin{aligned} Y_i &= \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 \cdot 0 + \gamma_2 \cdot 1 + \epsilon_i \\ &= (\alpha + \gamma_2) + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \end{aligned}$$

13 / 26

## Testing with polytomous independent variable

- Test partial effect of type, i.e., the effect of type controlling for income and education.
- $H_0 : \gamma_1 = \gamma_2 = 0$
- $H_a$ : at least one  $\gamma_j \neq 0$ ,  $j = 1, 2$ .
- Incremental F-test:
  - ◆ Null model:  
$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$
  - ◆ Full model:  
$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma_1 D_1 + \gamma_2 D_2 + \epsilon$$
- What do the individual p-values in `summary(lm())` mean?
- First look at F-test, then at individual p-values

14 / 26

## R commands

- Creating dummy variables by hand:

```
D1 <- (type=="prof")*1
D2 <- (type=="wc")*1
m1 <- lm(prestige~education+income+D1+D2)
```
- Letting R do things automatically:

```
m1 <- lm(prestige~education+income+type)
m1 <- lm(prestige~education+income+factor(type))
```
- The use of `factor()`:
  - ◆ `factor()` is not needed in this example, because the coding of the categories is in words: "bc", "prof", "wc".
  - ◆ It is essential to use `factor()` if the coding of the categories is numerical!
  - ◆ To be safe, you can always use `factor`.
- Example R-code

15 / 26

## More than one qualitative independent variable

- Example:  $Y = \text{prestige}$ ,  $X_1 = \text{income}$ ,  $X_2 = \text{education}$ ,

Type	$D_1$	$D_2$
Blue collar	0	0
Professional	1	0
White collar	0	1

and

Gender	$D_3$
Women	0
Men	1

- $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma_1 D_1 + \gamma_2 D_2 + \gamma_3 D_3 + \epsilon$

- What is the equation for men with professional jobs? And for women with white collar jobs?

16 / 26

## Interaction

17 / 26

### Definition

- Two variables are said to *interact* in determining a dependent variable if the partial effect of one depends on the value of the other.
- So far we only studied models without interaction.
- Interaction between a quantitative and a qualitative variable means that the regression surfaces are not parallel. See picture.
- Interaction between two qualitative variables means that the effect of one of the variables depends on the value of the other variable. Example: the effect of type of job on prestige is bigger for men than for women.
- Interaction between two quantitative variables is a bit harder to interpret, and we may consider that later.

18 / 26

### Interaction vs. correlation

- First, note that in general, the *independent* variables are *not independent* of each other.
- Correlation:  
Independent variables are statistically related to each other.
- Interaction:  
Effect of one independent variable on the dependent variable depends on the value of the other independent variable.
- Two independent variables can interact whether or not they are correlated.

19 / 26

## Constructing regressors

- $Y = \text{income}$ ,  $X = \text{education}$ ,  $D = \text{dummy for gender}$
- $Y_i = \alpha + \beta X_i + \gamma D_i + \delta(X_i D_i) + \epsilon_i$
- Note  $X \cdot D$  is a new regressor. It is a function of  $X$  and  $D$ , but not a linear function. Therefore we do not get perfect collinearity.
- Women ( $D_i = 0$ ):

$$Y_i = \alpha + \beta X_i + \gamma \cdot 0 + \delta(X_i \cdot 0) + \epsilon_i = \alpha + \beta X_i + \epsilon_i$$

- Men ( $D_i = 1$ )

$$\begin{aligned} Y_i &= \alpha + \beta X_i + \gamma \cdot 1 + \delta(X_i \cdot 1) + \epsilon_i \\ &= (\alpha + \gamma) + (\beta + \delta)X_i + \epsilon_i \end{aligned}$$

- Interpretation of  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$ .

20 / 26

## Testing

- Testing for interaction is testing for a difference in slope between men and women.  $H_0 : \delta = 0$  and  $H_a : \delta \neq 0$ .
- What is the difference between:
  - ◆ The model with interaction
  - ◆ Fitting two separate regression lines for men and women

21 / 26

## Principle of marginality

- If interaction is significant, do not test or interpret main effects:
  - ◆ First test for interaction effect.
  - ◆ If no interaction, test and interpret main effects.
- If interaction is included in the model, main effects should also be included.
- See pictures of models that violate the principle of marginality.

22 / 26

## Polytomous independent variables

- Create interaction regressors by taking the products of all dummy variable regressors and the quantitative variable.
- Example:
  - ◆  $Y = \text{prestige}$ ,  $X_1 = \text{education}$ ,  $X_2 = \text{income}$
  - ◆  $D_1, D_2 = \text{dummies for type of job}$

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma_1 D_1 + \gamma_2 D_2 \\ + \delta_{11} X_1 D_1 + \delta_{12} X_1 D_2 + \delta_{21} X_2 D_1 + \delta_{22} X_2 D_2 + \epsilon$$

- Interpretation of parameters

23 / 26

## Hypothesis tests

- When testing for main effects and interactions, follow principle of marginality
- Use incremental F-test
- Examples in R-code

24 / 26

## Standardized estimates

- Do not standardize dummy-regressor coefficients.
- Dummy regressor coefficient has clear interpretation.
- By standardizing it, this interpretation gets lost. Therefore we don't standardize dummy regressor coefficients.
- Also, don't standardize interaction regressors. You can standardize the quantitative independent variable before taking its product with the dummy regressor.

25 / 26

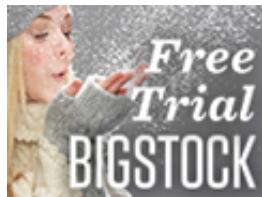
## Interaction between categorical variables

- Example: Does reproduction reduce lifespan of male fruitflies?
- Experiment:
  - ◆ male flies with 1 pregnant (not receptive) female per day
  - ◆ male flies with 8 pregnant females per day
  - ◆ male flies with 1 virgin (receptive) female per day
  - ◆ male flies with 8 virgin females per day
  - ◆ male flies without females
- Each group contains 25 fruitflies
- Available information:
  - ◆ Thorax length in mm
  - ◆ Percentage of time sleeping
  - ◆ Longevity in days
- See plots

26 / 26



By John Gruber



Free 7-Day Trial from Bigstock. Your first 35 stock photo downloads are free. No strings attached.

ADS VIA THE DECK [Ads via The Deck](#)

## Markdown: Basics

- [Main](#)
- [Basics](#)
- [Syntax](#)
- [License](#)
- [Dingus](#)

### GETTING THE GIST OF MARKDOWN'S FORMATTING SYNTAX

This page offers a brief overview of what it's like to use Markdown. The [syntax page](#) provides complete, detailed documentation for every feature, but Markdown should be very easy to pick up simply by looking at a few examples of it in action. The examples on this page are written in a before/after style, showing example syntax and the HTML output produced by Markdown.

It's also helpful to simply try Markdown out; the [Dingus](#) is a web application that allows you type your own Markdown-formatted text and translate it to XHTML.

**Note:** This document is itself written using Markdown; you can [see the source for it by adding '.text' to the URL](#).

### PARAGRAPHS, HEADERS, BLOCKQUOTES

A paragraph is simply one or more consecutive lines of text, separated by one or more blank lines. (A blank line is any line that looks like a blank line — a line containing nothing but spaces or tabs is considered blank.) Normal paragraphs should not be indented with spaces or tabs.

Markdown offers two styles of headers: *Setext* and *atx*. Setext-style headers for `<h1>` and `<h2>` are created by “underlining” with equal signs (=) and hyphens (-), respectively. To create an atx-style header, you put 1-6 hash marks (#) at the beginning of the line — the number of hashes equals the resulting HTML header level.

Blockquotes are indicated using email-style '>' angle brackets.

Markdown:

```
A First Level Header  
=====
```

```
A Second Level Header  
-----
```

```
Now is the time for all good men to come to  
the aid of their country. This is just a  
regular paragraph.
```

```
The quick brown fox jumped over the lazy  
dog's back.
```

```
### Header 3
```

```
> This is a blockquote.  
>  
> This is the second paragraph in the blockquote.  
>  
> ## This is an H2 in a blockquote
```

Output:

```
<h1>A First Level Header</h1>  
  
<h2>A Second Level Header</h2>  
  
<p>Now is the time for all good men to come to  
the aid of their country. This is just a  
regular paragraph.</p>  
  
<p>The quick brown fox jumped over the lazy  
dog's back.</p>  
  
<h3>Header 3</h3>  
  
<blockquote>  
  <p>This is a blockquote.</p>  
  
  <p>This is the second paragraph in the blockquote.</p>  
  
<h2>This is an H2 in a blockquote</h2>  
</blockquote>
```

## Phrase Emphasis

Markdown uses asterisks and underscores to indicate spans of emphasis.

Markdown:

```
Some of these words *are emphasized*.  
Some of these words _are emphasized also_.
```

Use two asterisks for **strong emphasis**.  
 Or, if you prefer, use two underscores instead.

**Output:**

```
<p>Some of these words <em>are emphasized</em>.  

Some of these words <em>are emphasized also</em>.</p>

<p>Use two asterisks for <strong>strong emphasis</strong>.  

Or, if you prefer, <strong>use two underscores instead</strong>.</p>
```

## LISTS

Unordered (bulleted) lists use asterisks, pluses, and hyphens (\*, +, and –) as list markers. These three markers are interchangeable; this:

- \* Candy.
- \* Gum.
- \* Booze.

this:

- + Candy.
- + Gum.
- + Booze.

and this:

- Candy.
- Gum.
- Booze.

all produce the same output:

```
<ul>
<li>Candy.</li>
<li>Gum.</li>
<li>Booze.</li>
</ul>
```

Ordered (numbered) lists use regular numbers, followed by periods, as list markers:

1. Red
2. Green
3. Blue

**Output:**

```
<ol>
<li>Red</li>
<li>Green</li>
<li>Blue</li>
</ol>
```

If you put blank lines between items, you'll get `<p>` tags for the list item text. You can create multi-paragraph list items by indenting the paragraphs by 4 spaces or 1 tab:

- \* A list item.  
With multiple paragraphs.
- \* Another item in the list.

Output:

```
<ul>
<li><p>A list item.</p>
<p>With multiple paragraphs.</p></li>
<li><p>Another item in the list.</p></li>
</ul>
```

## Links

Markdown supports two styles for creating links: *inline* and *reference*. With both styles, you use square brackets to delimit the text you want to turn into a link.

Inline-style links use parentheses immediately after the link text. For example:

```
This is an [example link] (http://example.com) .
```

Output:

```
<p>This is an <a href="http://example.com/">
example link</a>.</p>
```

Optionally, you may include a `title` attribute in the parentheses:

```
This is an [example link] (http://example.com/ "With a Title") .
```

Output:

```
<p>This is an <a href="http://example.com/" title="With a Title">
example link</a>.</p>
```

Reference-style links allow you to refer to your links by names, which you define elsewhere in your document:

```
I get 10 times more traffic from [Google][1] than from
[Yahoo][2] or [MSN][3].
```

```
[1]: http://google.com/      "Google"
[2]: http://search.yahoo.com/ "Yahoo Search"
[3]: http://search.msn.com/   "MSN Search"
```

Output:

```
<p>I get 10 times more traffic from <a href="http://google.com/">
Google</a> than from <a href="http://search.yahoo.com/">
Yahoo</a> or <a href="http://search.msn.com/">
MSN</a>.</p>
```

The title attribute is optional. Link names may contain letters, numbers and spaces, but are *not* case sensitive:

```
I start my morning with a cup of coffee and
[The New York Times] [NY Times].
```

```
[ny times]: http://www.nytimes.com/
```

Output:

```
<p>I start my morning with a cup of coffee and
<a href="http://www.nytimes.com/">The New York Times</a>.</p>
```

## Images

Image syntax is very much like link syntax.

Inline (titles are optional):

```
![alt text] (/path/to/img.jpg "Title")
```

Reference-style:

```
![alt text][id]
```

```
[id]: /path/to/img.jpg "Title"
```

Both of the above examples produce the same output:

```

```

## Code

In a regular paragraph, you can create code span by wrapping text in backtick quotes. Any ampersands (&) and angle brackets (< or >) will automatically be translated into HTML entities. This makes it easy to use Markdown to write about HTML example code:

```
I strongly recommend against using any `<blink>` tags.
```

```
I wish SmartyPants used named entities like `&mdash;` instead of decimal-encoded entites like `—`.
```

Output:

```
<p>I strongly recommend against using any
<code>&lt;blink&gt;</code> tags.</p>
```

```
<p>I wish SmartyPants used named entities like
<code>&mdash;</code> instead of decimal-encoded
entites like <code>&#8212;</code>. </p>
```

To specify an entire block of pre-formatted code, indent every line of the block by 4 spaces or 1 tab. Just like with code spans, &, <, and > characters will be escaped automatically.

Markdown:

If you want your page to validate under XHTML 1.0 Strict, you've got to put paragraph tags in your blockquotes:

```
<blockquote>
    <p>For example.</p>
</blockquote>
```

Output:

```
<p>If you want your page to validate under XHTML 1.0 Strict,
you've got to put paragraph tags in your blockquotes:</p>
```

```
<pre><code>&lt;blockquote&gt;
    &lt;p&gt;For example.&lt;/p&gt;
&lt;/blockquote&gt;
</code></pre>
```

 **Search**

# Unusual and influential data

Chapter 11 . . . . .	2
What to do with unusual data? . . . . .	3
Unusual data points . . . . .	4
<b>Leverage points</b>	<b>5</b>
Leverage . . . . .	6
Leverage . . . . .	7
<b>Regression outliers</b>	<b>8</b>
Residuals . . . . .	9
Standardized/studentized residuals. . . . .	10
Testing for outliers . . . . .	11
<b>Influential points</b>	<b>12</b>
Influence . . . . .	13
Joint influence . . . . .	14
Some more useful R-commands . . . . .	15

## Chapter 11

- Unusual data points:
  - ◆ What to do with them?
  - ◆ Leverage: hat values
  - ◆ Outliers: standardized/studentized residuals
  - ◆ Influence: Cook's distance
  - ◆ Added variable plots can help find clusters of points that are jointly influential

2 / 15

### What to do with unusual data?

- Neither ignore them, nor throw them out without thinking
- Check for data entry errors
- Think of reasons why observation may be different
- Change the model
- Fit model with and without the observations to see the effect
- Robust regression

3 / 15

### Unusual data points

- Univariate outlier:
  - ◆ Unusual value for one of the  $X$ 's or for  $Y$
- Leverage point: point with unusual combination of independent variables
- Regression outlier:
  - ◆ Large residual (in absolute value)
  - ◆ The value of  $Y$  *conditional* on  $X$  is unusual
- Influential point: points with large influence on the regression coefficients
- Influence = Leverage  $\times$  'Outlyingness'
- See examples

4 / 15

### Leverage

- Leverage is measured by the so-called “hat values”
- Hat values:  $\hat{Y}_j = h_{1j}Y_1 + \dots + h_{nj}Y_n = \sum_{i=1}^n h_{ij}Y_i$
- In matrix notation,  $h_{ij}$  are the elements of the hat matrix  $H = X(X^T X)^{-1}X^T$ .  $H$  is called the hat matrix since  $\hat{Y} = HY$ .
- The weight  $h_{ij}$  captures the contribution of  $Y_i$  to the fitted value  $\hat{Y}_j$
- The number  $h_i \equiv h_{ii} = \sum_{j=1}^n h_{ij}^2$  summarizes the leverage of  $Y_i$  on *all* fitted values
- Note the dependent variable  $Y$  is not involved in the computation of the hat values

6 / 15

### Leverage

- Range of the hat values:  $1/n \leq h_i \leq 1$
- Average of the hat values:  $\bar{h} = (k+1)/n$
- Rule of thumb: leverage is large if  $h_i > 2(k+1)/n$ . Draw a horizontal line at this value
- R-function: `hatvalues()`
- See example

7 / 15

### Residuals

- Residuals:  $E_i = Y_i - \hat{Y}_i$ . R-function `resid()`.
- Even if statistical errors have constant variance, the residuals do not have constant variance:  $V(E_i) = \sigma_\epsilon^2(1 - h_i)$ .
- Hence, high leverage points tend to have small residuals, which makes sense because these points can ‘pull’ the regression line towards them.

9 / 15

## Standardized/studentized residuals

- We can compute versions of the residuals with constant variance:
  - ◆ Standardized residuals  $E'_i$  and studentized residuals  $E_i^*$ :

$$E'_i = \frac{E_i}{S_E\sqrt{1-h_i}} \quad \text{and} \quad E_i^* = \frac{E_i}{S_{E(-i)}\sqrt{1-h_i}}.$$

- ◆ Here  $S_{E(-i)}$  is an estimate of  $\sigma_\epsilon$  when leaving out the  $i$ th observation.
- ◆ R-functions `rstandard()` and `rstudent()`.

10 / 15

## Testing for outliers

- Look at studentized residuals by eye.
- If the model is correct, then  $E_i^*$  has t-distribution with  $n - k - 2$  degrees of freedom.
- If the model is true, about 5% of observations will have studentized residuals outside of the ranges  $[-2, 2]$ . It is therefore reasonable to draw horizontal lines at  $\pm 2$ .
- We can use Bonferroni test to determine if largest studentized residual is an outlier: divide your cut-off for significant p-values (usually 0.05) by  $n$ .

11 / 15

## Influential points

12 / 15

### Influence

- Influence = Leverage  $\times$  'Outlyingness'
- Cook's distance:

$$D_i = \frac{h_i}{1-h_i} \times \frac{E_i'^2}{k+1}$$

- Cook's distance measures the difference in the regression estimates when the  $i$ th observation is left out
- Rule of thumb: Cook's distance is large if  $D_i > 4/(n - k - 1)$
- R-command: `cooks.distance()`

13 / 15

### Joint influence

- See example
- Use added variable plots to detect this

14 / 15

## Some more useful R-commands

- `identify()`: to identify points in the plot
- `plot(m)`: gives 4 plots:
  - ◆ Residuals against fitted values
  - ◆ QQ-plot of standardized residuals
  - ◆ Scale-location plot
  - ◆ Cook's distance plot
- `influence.measures(m)`: contains various measures of influence.

15 / 15



SATURDAY, MARCH 3, 2007

## Comparative Model Testing and Nested Models

As we've discussed, part of the latest assignment requires you to engage in comparative model testing. Specifically, you will run your model both with and without directed paths from three university properties (public/private status, years of existence, and endowment [square-root transformed]) to their Undergraduate Quality (UQ).

The more parsimonious model is, of course, the one without the additional paths. To override the preference for parsimony, therefore, you will have to show that the additional paths, as a set, significantly reduce the overall model chi-square, thus improving model fit. As you move along in your careers, you may wish to adopt additional criteria, such as whether the reduction in chi-square appears substantively large in addition to being statistically significant, but for now, we'll use statistically significant change as our criterion.

You can display your results in a table, as follows:

Model.....X<sub>2</sub>.....df....

Model w/ fewer parameters.....----

Model w/ added parameters.....----

Delta (change).....----

The chi-square change score (the top chi-square minus the bottom one) can be treated like any other chi-square value and be referred to a chi-square table, with degrees of freedom equal to

DR. REIFMAN'S...

[Faculty Webpage](#)[Intro Stats Page](#)

SEM OVERVIEW PAGES

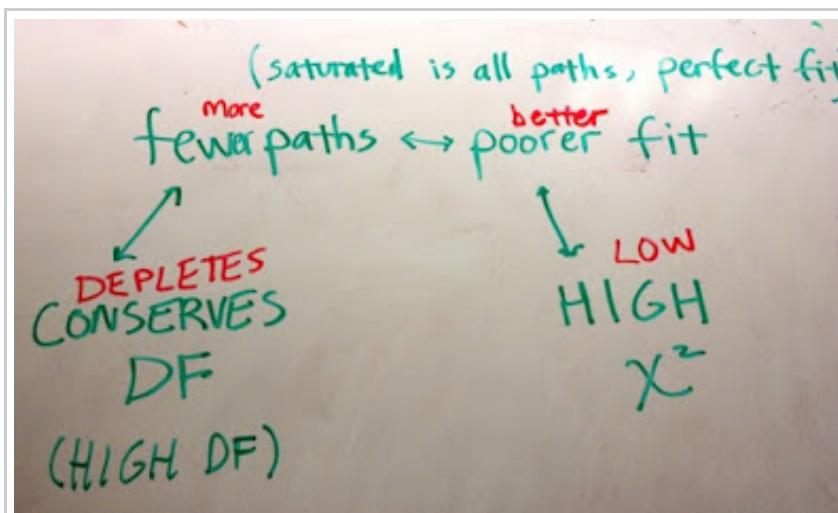
[U. Kentucky / PIRE](#)[Kenny \(UConn\)](#)[Newsom \(Portland State\)](#)[Rigdon \(Georgia State\)](#)[StructuralEquations.org](#)[SEM Pros and Cons](#)

SPECIFIC SEM ISSUES

[CFA -- Use of Different Programs](#)[Causality and SEM \(Bollen & Pearl Working Paper\)](#)[Bharmal Medication article -- Good for reviewing how to count up degrees of freedom](#)[Model Fit \(Kenny\)](#)[Correlating Residual Variances](#)[Non-Positive Definite \(Error Message\)](#)[Sample Size \(Muthén & Muthén\)](#)[Sample Size \(Westland\)](#)[Sample Size \(Wolf et al.\)](#)[Missing Data in SEM \(Enders\)](#)

delta df (top df minus bottom df).

**UPDATE, March 11, 2012:** Xiaohui photographed the explanation I diagrammed on the board, linking number of paths in a model, goodness of fit, chi-square, and degrees of freedom. A key point was to demonstrate that if one model has a higher chi-square than another model, it will also have a higher number of degrees of freedom. **All of the green phrases go together: a model with fewer paths (which preserves a higher df) will have a poorer fit and thus a higher chi-square.** The red terms represent the opposite of the green terms, and thus **the red terms go together, as well: a model with more paths (which depletes the df) will lead to a better fit and thus a lower chi-square.**



**UPDATE, March 5, 2008:** Kristina photographed the decision-tree I drew on the board, to augment our discussion of comparative model testing. Here it is (you can click on the image to enlarge it).

Missing Data in SEM (Newsom)

Meta-Analytic SEM

YouTube Videos on Many SEM Topics (Gaskin)

OPTIONAL BOOKS FOR STUDENTS SEEKING ADDITIONAL PERSPECTIVE

Farbrigar & Wegener (EFA)

Barbara Byrne (SEM/AMOS)

Rex Kline (SEM)

Schumacker & Lomax (SEM)

Handbook of SEM (multiple contributors; R. Hoyle, editor)

SEM-BASED DISSERTATIONS/THESSES FROM FORMER STUDENTS OF THE CLASS (SOME HAVE TTU RESTRICTED VIEWING)

Joy Cheng

Yoona Chin

Sothy Eng

Kyle Gillett

Stephanie Haygood

Branden Henline

Kristina Keyton

NaYeon Lee

Andrea McCourt

Adam Munk

Megan Oka

Damon Rappleyea

Hye-Sun Ro

Brittney Schrick

Xiaohui Tang

Shera Thomas-Jackson

Mitsue Uchida

AMOS INFORMATION

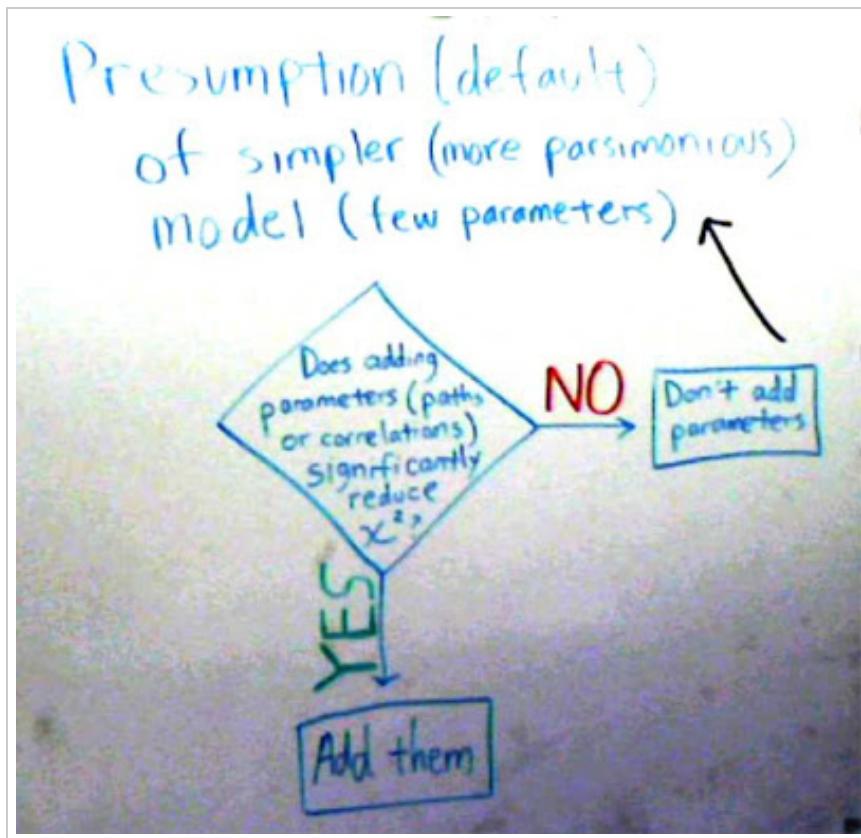
AMOS Development Corp.

AMOS Tutorial (U. Texas)

Citing AMOS in Your Papers

On the Hub (Student Software Discounts)

Videos on Using AMOS for Advanced Applications



And now, back to our regular programming...

An important condition for being able to conduct comparative model tests is that the two models being compared to each other must possess the property of **nestedness**. Two models are nested if they can be converted from one to the other either by *only adding parameters* to one to obtain the other, or *only removing parameters* from one to obtain the other. By *parameters*, we mean anything that is freely estimated in SEM (e.g., structural paths, non-directional correlations). If you start with one model and convert it to a new, second model by both *adding and subtracting* parameters from the initial model, the two models will *not* fulfill the criteria for nestedness and thus cannot be compared via the delta chi-square test.

The following two diagrams provide examples of nested and non-nested models.

#### PATH ANALYSIS

Ants in Argentina  
U. of Exeter (UK)

#### BIDIRECTIONAL ARROWS, 2SLS, INSTRUMENTAL VARIABLES

Chang & Chen  
John Fox

#### FACTOR ANALYSIS (EXPLORATORY)

Statsoft Electronic Textbook  
Garson (NC State)  
MacCallum (UNC)  
Parallel Analysis, for Determining No. of Factors (O'Connor)  
Factor Rotation (Mathworks)

#### LONGITUDINAL/CAUSALITY

Longitudinal Notes (Reifman Methods Class)  
Causality Notes (Reifman Methods Class)  
Correlation & Causality Blog

#### JOURNALS, ARTICLES

SEM (the journal)  
Special SEM Issue of Personality & Individual Differences (May 2007)

#### MISCELLANEOUS

SEMNET Discussion Forum  
Dataset Archive (ICPSR)  
Garson Overall Stats Page  
Least-Squares Visualizer  
Correlation-Covariance Conversion Formula  
Award Statement for "SEM The Musical"

TTU RESOURCES

[Online Syllabi for All Courses](#)  
[Academic Calendars](#)  
[Final Exam Schedules](#)

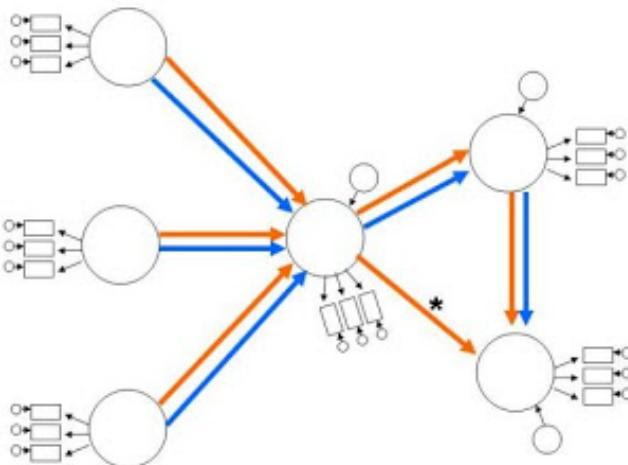
## BLOG ARCHIVE

- 2014 (2)
- 2013 (1)
- 2012 (3)
- 2011 (1)
- 2010 (5)
- 2009 (5)
- 2008 (8)
- ▼ 2007 (16)
  - April (4)
  - ▼ March (2)

[Comparative Model Testing  
and Nested Models](#)

- Negative Variances (Heywood Cases)
- February (5)
- January (5)

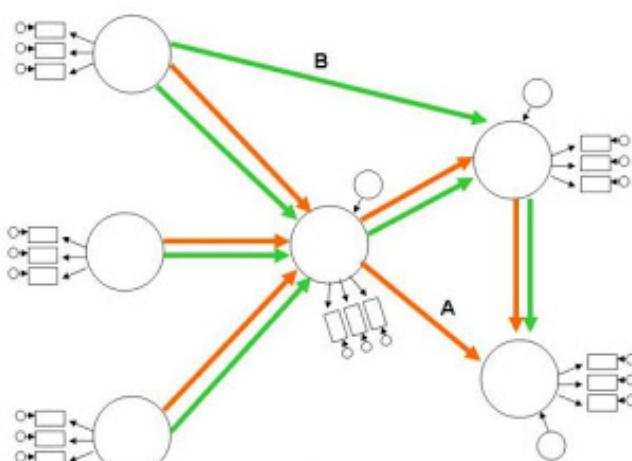
WHAT SONGS FROM PREVIOUS YEARS DO YOU WANT TO SING AT SEM THE MUSICAL 6? (YOU MAY VOTE FOR UP TO FIVE.)



Blue model nested within red model. Situation could come about by...

- Starting with blue model and adding starred path to get red model OR
- Starting with red model and removing starred path to get blue model.

Nestedness = Making changes by only adding parameters or only subtracting them.

**NOT NESTED**

Red model converted into green model by removing A path and adding B.

An analogous situation exists in multiple regression. You can do a delta R-square test to see, for example, if a model with predictor set A, B, C, D, and E accounts for significantly more variance in the dependent variable than does predictor set A, B, and C. ABC is contained -- that is nested -- within ABCDE, thus permitting the statistical comparison. You could not, however, test whether predictor set ABCDF accounts for more variance than set ABCDE, because the change in models would have required both dropping a predictor and adding one. If ABCDE was the starting point, we would have dropped E and added F.

We'll use the following article to delve more deeply into comparative model testing:

Bryant, A. L., Schulenberg, J., Bachman, J. G., O'Malley, P. M., & Johnston, L. D. (2000). Understanding the links among school misbehavior, academic achievement, and cigarette use: A national

panel study of adolescents. *Prevention Science, 1, 71-87.*

POSTED BY ALAN AT 7:29 PM

Newer Post

Home

Older Post

At Least Three

Gotta Fix it to One (It Ain't Free)

Constrain, 'strain, 'strain

You've Got to Check Your R-M-S-E-~~A~~

You've Had a Bad Fit

Your Model's Only One

It Do Run Run

Count 'em Up

AMOS is Ideal

Equal (You've Got Me Constrained to You)

Once You Work in AMOS

Pyramid of Success

Hey Hey Heywood Cases

Nestedness

Maximum Likelihood

Votes so far: 0

Roll ~~closed~~ open

## Equations in R Markdown



Josh Paulson

October 13, 2013 03:22

Recently viewed articles

## Equations in R Markdown

With [R Markdown](#), you can embed LaTeX and MathML equations directly into your document. Equations are displayed using the [MathJax](#) JavaScript library. Note that this library is loaded from the MathJax website so readers of your document must be online to see the rendered equations.

### LaTeX Inline Equations

To include an inline LaTeX equation you enclose the equation in \$ delimiters, for

example:

```

1
2 The Arithmetic mean is equal to  $\frac{1}{n} \sum_{i=1}^n x_i$ , or
3 the summation of n numbers divided by n.

```

In order to avoid conflicts with currency specifications, the following syntactic rules apply to the use of \$ delimiters:

- The equation text must be directly attached to the \$ characters with no whitespace in between.
- The closing \$ must not be followed by a number, letter, or back-tick.
- The equation text can contain at most two line breaks.

To prevent a \$ from being treated as an equation delimiter you can escape it using a backslash (e.g. \\$).

Inline equations can span multiple lines. Note however that within the RStudio editor you won't see syntax highlighting for inline equations that span across lines.

### LaTeX Display Equations

To include a LaTeX display equation you enclose the equation in \$\$ delimiters, for

example:

```

1
2 $$
3 \begin{aligned}
4 \dot{x} &= \sigma(y-x) \\
5 \dot{y} &= \rho x - y - xz \\
6 \dot{z} &= -\beta z + xy
7 \end{aligned}
8 $$
9

```

Display equations can span an arbitrary number of lines and unlike inline equations do not have a requirement that the equation text be directly attached to the delimiters.

Display equations are rendered as a centered block element within the generated web page.

[R Markdown](#)[Using R Markdown](#)

Related articles

[Using R Markdown](#)[Customizing Markdown Rendering](#)[Creating Notebooks from R Scripts](#)[R Markdown Specification](#)[Using Sweave and knitr](#)

## Alternative Syntax for LaTeX Equations

The syntax described above for inline and display equations is based on conventions used for embedding equations in LaTeX documents. It's also compatible with the syntax used by org-mode and the pandoc markdown engine. However, depending on how you are publishing web content you may wish to use one the alternate syntaxes described below.

### WP LaTeX

The [WP LaTeX](#) WordPress plugin supports a variation of the traditional \$ and \$\$ delimiters for embedding equations. For example:

```

1 $latex P(E) = {n \choose k} p^k (1-p)^{n-k}$
2
3
4 $$\begin{aligned}
5 \dot{x} &= \sigma(y-x) \\
6 \dot{y} &= \rho x - y - xz \\
7 \dot{z} &= -\beta z + xy
8 \end{aligned}$$
9
10 $$
11

```

You might choose to use this syntax if you intend to eventually publish your markdown into a WordPress blog.

### MathJax Native

You can also use the native MathJax delimiters for inline and display equations. For example:

```

1 \[ P(E) = {n \choose k} p^k (1-p)^{n-k} \]
2
3
4 \[
5 \begin{aligned}
6 \dot{x} &= \sigma(y-x) \\
7 \dot{y} &= \rho x - y - xz \\
8 \dot{z} &= -\beta z + xy
9 \end{aligned}\]
10 \]
11

```

Note that when R Markdown processes \$ or \$latex style delimiters within a document they are written into the target HTML file using the native MathJax delimiters shown above.

## MathML Equations

To insert MathML equations, wrap your equation inside a standard `<math>` tag. For example, to insert the quadratic formula you would use:

```

<math xmlns="http://www.w3.org/1998/Math/MathML" display="block">
<mrow>
<mi>x</mi>
<mo>=</mo>
<mfrac>
<mrow>
<mo>\pm</mo>
<mi>b</mi>
<mo>\sqrt{b^2 - 4ac}</mo>
</mrow>
<mn>2</mn>
</mfrac>

```

```
<mo>+</mo>
<mn>4</mn>
<mi>a</mi>
<mi>c</mi>
</mrow>
</msqrt>
</mrow>
<mrow>
<mn>2</mn>
<mi>a</mi>
</mrow>
</mfrac>
</mrow>
</math>
```

## Related Topics

- [Using R Markdown](#)
- [Customizing Markdown Rendering](#)
- [R Markdown Specification](#)
- [Creating Notebooks from R Scripts](#)

Was this article helpful?  
0 out of 0 found this helpful



f t in g+

Have more questions? [Submit a request](#)

## Comments

Article is closed for comments.

# Hat matrix

From Wikipedia, the free encyclopedia

In statistics, the **hat matrix**,  $H$ , sometimes also called **projection matrix**, maps the vector of observed values to the vector of fitted values. It describes the influence each observed value has on each fitted value.<sup>[1]</sup> The diagonal elements of the hat matrix are the leverages, which describe the influence each observed value has on the fitted value for that same observation.

If the vector of observed values is denoted by  $\mathbf{y}$  and the vector of fitted values by  $\hat{\mathbf{y}}$ ,

$$\hat{\mathbf{y}} = H\mathbf{y}.$$

As  $\hat{\mathbf{y}}$  is usually pronounced "y-hat", the hat matrix is so named as it "puts a hat on  $\mathbf{y}$ ".

Suppose that we wish to solve a linear model using linear least squares. The model can be written as

$$\mathbf{y} = X\beta + \epsilon,$$

where  $X$  is a matrix of explanatory variables (the design matrix),  $\beta$  is a vector of unknown parameters to be estimated, and  $\epsilon$  is the error vector.

## Contents

- 1 Uncorrelated errors
- 2 Correlated errors
- 3 Blockwise formula
- 4 See also
- 5 References

## Uncorrelated errors

For uncorrelated errors, the estimated parameters are

$$\hat{\beta} = (X^\top X)^{-1} X^\top \mathbf{y},$$

so the fitted values are

$$\hat{\mathbf{y}} = X\hat{\beta} = X(X^\top X)^{-1} X^\top \mathbf{y}.$$

Therefore the hat matrix is given by

$$H = X(X^\top X)^{-1} X^\top.$$

In the language of linear algebra, the hat matrix is the orthogonal projection onto the column space of the design matrix  $X$ . (Note that  $(X^\top X)^{-1} X^\top$  is the pseudoinverse of  $X$ .)

The hat matrix corresponding to a linear model is symmetric and idempotent, that is,  $H^2 = H$ . However, this is not always the case; in locally weighted scatterplot smoothing (LOESS), for example, the hat matrix is in general neither symmetric nor idempotent.

The formula for the vector of residuals  $\mathbf{r}$  can be expressed compactly using the hat matrix:

$$\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - H\mathbf{y} = (I - H)\mathbf{y}.$$

The covariance matrix of the residuals is therefore, by error propagation, equal to  $(I - H)^\top \Sigma (I - H)$ , where  $\Sigma$  is the covariance matrix of the errors (and by extension, the observations as well). For the case of linear models with independent and identically distributed errors in which  $\Sigma = \sigma^2 I$ , this reduces to  $(I - H)\sigma^2$ .<sup>[1]</sup>

For linear models, the trace of the hat matrix is equal to the rank of  $X$ , which is the number of independent parameters of the linear model. For other models such as LOESS that are still linear in the observations  $\mathbf{y}$ , the hat matrix can be used to define the effective degrees of freedom of the model.

The hat matrix has a number of useful algebraic properties.<sup>[2][3]</sup> Practical applications of the hat matrix in regression analysis include leverage and Cook's distance, which are concerned with identifying observations which have a large effect on the results of a regression.

## Correlated errors

The above may be generalized to the case of correlated errors. Suppose that the covariance matrix of the errors is  $\Sigma$ . Then since

$$\hat{\boldsymbol{\beta}} = (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} \mathbf{y},$$

the hat matrix is thus

$$H = X (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1},$$

and again it may be seen that  $H^2 = H$

## Blockwise formula

Suppose the design matrix  $C$  can be decomposed by columns as  $C = [A, B]$ . Define the Hat operator as  $H(X) = X (X^\top X)^{-1} X^\top$ . Similarly, define the residual operator as  $M(X) = I - H(X)$ . Then the Hat matrix of  $C$  can be decomposed as follows:

$$H(C) = H(A) + H(M(A)B)<sup>[4]</sup>$$

There are a number of applications of such a partitioning. The classical application has  $A$  a column of all ones, which allows one to analyze the effects of adding an intercept term to a regression. Another use is in the fixed effects model, where  $A$  is a large sparse matrix of the dummy variables for the fixed effect terms. One can use

this partition to compute the hat matrix of  $\mathbf{C}$  without explicitly forming the matrix  $\mathbf{C}$ , which might be too large to fit into computer memory.

## See also

- Moore–Penrose pseudoinverse
- Studentized residuals
- Effective degrees of freedom
- Idempotent matrix

## References

1. <sup>a b</sup> Hoaglin, David C.; Welsch, Roy E. (February 1978), "The Hat Matrix in Regression and ANOVA", *The American Statistician* **32** (1): 17–22, doi:10.2307/2683469 (<http://dx.doi.org/10.2307%2F2683469>), JSTOR 2683469 (<https://www.jstor.org/stable/2683469>)
2. <sup>a</sup> Gans, P. (1992) *Data Fitting in the Chemical Sciences*, Wiley. ISBN 978-0-471-93412-7
3. <sup>a</sup> Draper, N.R., Smith, H. (1998) *Applied Regression Analysis*, Wiley. ISBN 0-471-17082-6
4. <sup>a</sup> Rao, C. Radhakrishna; Toutenburg, Shalabh, Heumann (2008). *Linear Models and Generalizations* (3rd ed.). Berlin: Springer. p. 323. ISBN 978-3-540-74226-5.

Retrieved from "[http://en.wikipedia.org/w/index.php?title=Hat\\_matrix&oldid=598734186](http://en.wikipedia.org/w/index.php?title=Hat_matrix&oldid=598734186)"

Categories: Statistical terminology | Regression analysis | Matrices

- This page was last modified on 8 March 2014 at 19:33.
- Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.



---

The Hat Matrix in Regression and ANOVA

Author(s): David C. Hoaglin and Roy E. Welsch

Reviewed work(s):

Source: *The American Statistician*, Vol. 32, No. 1 (Feb., 1978), pp. 17-22

Published by: [American Statistical Association](#)

Stable URL: <http://www.jstor.org/stable/2683469>

Accessed: 18/01/2012 13:48

---

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at

<http://www.jstor.org/page/info/about/policies/terms.jsp>

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.



American Statistical Association is collaborating with JSTOR to digitize, preserve and extend access to *The American Statistician*.

<http://www.jstor.org>

# The Hat Matrix in Regression and ANOVA

DAVID C. HOAGLIN AND ROY E. WELSCH\*

In least-squares fitting it is important to understand the influence which a data  $y$  value will have on each fitted  $y$  value. A projection matrix known as the hat matrix contains this information and, together with the Studentized residuals, provides a means of identifying exceptional data points. This approach also simplifies the calculations involved in removing a data point, and it requires only simple modifications in the preferred numerical least-squares algorithms.

KEY WORDS: Analysis of variance; Regression analysis; Projection matrix; Outliers; Studentized residuals; Least-squares computations.

## 1. Introduction

In fitting linear models by least squares it is very often useful to determine how much influence or leverage each data  $y$  value ( $y_i$ ) can have on each fitted  $y$  value ( $\hat{y}_i$ ). For the fitted value  $\hat{y}_i$  corresponding to the data value  $y_i$ , the relationship is particularly straightforward to interpret, and it can reveal multivariate outliers among the carriers (or  $x$  variables) which might otherwise be difficult to detect. The desired information is available in the hat matrix, which gives each fitted value  $\hat{y}_i$  as a linear combination of the observed values  $y_j$ . (The term "hat matrix" is due to John W. Tukey, who introduced us to the technique about ten years ago.) The present article derives and discusses the hat matrix and gives an example to illustrate its usefulness.

Section 2 defines the hat matrix and derives its basic properties. Section 3 formally examines two familiar examples, while Section 4 gives a numerical example. In practice one must, of course, consider the actual effect of the data  $y$  values in addition to their leverage; we discuss this in terms of the residuals in Section 5. Section 6 then sketches how the hat matrix can be obtained from two accurate numerical algorithms used for solving least-squares problems.

## 2. Basic Properties

We are concerned with the linear model

$$\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad (2.1)$$

which summarizes the dependence of the *response*  $y$

\* David C. Hoaglin is Senior Analyst, Abt Associates, 55 Wheeler Street, Cambridge, MA 02138, and Research Associate, Department of Statistics, Harvard University, Cambridge, MA 02138. Roy E. Welsch is Associate Professor of Operations Research and Management, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139. This work was supported in part by NSF Grant SOC75-15702 to Harvard University and by NSF Grant 76-14311 DSS to the National Bureau of Economic Research.

on the *carriers*  $X_1, \dots, X_p$  in terms of the data values  $y_i$  and  $x_{i1}, \dots, x_{ip}$  for  $i = 1, \dots, n$ . (We refrain from thinking of  $X_1, \dots, X_p$  as independent variables because they are often not independent in any reasonable sense.) In fitting the model (2.1) by least squares (assuming that  $X$  has rank  $p$  and that  $E(\boldsymbol{\epsilon}) = \mathbf{0}$  and  $\text{var}(\boldsymbol{\epsilon}) = \sigma^2 I_n$ ), we usually obtain the fitted or predicted values from  $\hat{y} = \mathbf{X}\mathbf{b}$ , where  $\mathbf{b} = (X^T X)^{-1} X^T \mathbf{y}$ . From this it is simple to see that

$$\hat{y} = X(X^T X)^{-1} X^T \mathbf{y}. \quad (2.2)$$

To emphasize the fact that (when  $X$  is fixed) each  $\hat{y}_i$  is a linear function of the  $y_j$ , we write (2.2) as

$$\hat{y} = H\mathbf{y}, \quad (2.3)$$

where  $H = X(X^T X)^{-1} X^T$ . The  $n \times n$  matrix  $H$  is known as the hat matrix simply because it maps  $\mathbf{y}$  into  $\hat{y}$ . Geometrically, if we represent the data vector  $\mathbf{y}$  and the columns of  $X$  as points in Euclidean  $n$  space, then the points  $X\boldsymbol{\beta}$  (which we can obtain as linear combinations of the column vectors) constitute a  $p$  dimensional subspace. The fitted vector  $\hat{y}$  is the point of that subspace nearest to  $\mathbf{y}$ , and it is also the perpendicular projection of  $\mathbf{y}$  into the subspace. Thus  $H$  is a projection matrix. Also familiar is the role which  $H$  plays in the covariance matrices of  $\hat{y}$  and of  $\mathbf{r} = \mathbf{y} - \hat{y}$ :

$$\text{var}(\hat{y}) = \sigma^2 H, \quad (2.4)$$

$$\text{var}(\mathbf{r}) = \sigma^2(I - H). \quad (2.5)$$

For the data analyst, the element  $h_{ij}$  of  $H$  has a direct interpretation as the amount of leverage or influence exerted on  $\hat{y}_i$  by  $y_j$  (regardless of the actual value of  $y_j$ , since  $H$  depends only on  $X$ ). Thus a look at the hat matrix can reveal sensitive points in the design, points at which the value of  $y$  has a large impact on the fit (Huber 1975). In using the word "design" here, we have in mind both the standard regression or ANOVA situation, in which the values of  $X_1, \dots, X_p$  are fixed in advance, and the situation in which  $y$  and  $X_1, \dots, X_p$  are sampled together. The simple designs, such as two-way analysis of variance, give good control over leverage (as we shall see in Section 3); and with fixed  $X$  one can examine, and perhaps modify, the experimental conditions in advance. When the carriers are sampled, one can at least determine whether the observed  $X$  contains sensitive points and consider omitting them if the corresponding  $y$  value seems discrepant. Thus we use the hat matrix to identify "high-leverage points." If this notion is to be really useful, we must make it more precise.

The influence of the response value  $y_i$  on the fit is most directly reflected in its leverage on the corre-

sponding fitted value  $\hat{y}_i$ , and this is precisely the information contained in  $h_{ii}$ , the corresponding diagonal element of the hat matrix. We can easily imagine fitting a simple regression line to data  $(x_i, y_i)$ , making large changes in the  $y$  value corresponding to the largest  $x$  value, and watching the fitted line follow that data point. In this one-carrier problem or in a two-carrier problem a scatter plot will quickly reveal any  $x$  outliers, and we can verify that they have relatively large diagonal elements  $h_{ii}$ . When  $p > 2$ , scatter plots may not reveal multivariate outliers, which are separated in  $p$  space from the bulk of the  $x$  points but do not appear as outliers in a plot of any single carrier or pair of carriers, and the diagonal of the hat matrix is a source of valuable diagnostic information. In addition to being somewhat easier to understand, the diagonal elements of  $H$  can be less trouble to compute, store, and examine, especially if  $n$  is moderately large. Thus attention focuses primarily (often exclusively) on the  $h_{ii}$ , which we shall sometimes abbreviate  $h_i$ . We next examine some of their properties.

As a projection matrix,  $H$  is symmetric and idempotent ( $H^2 = H$ ), as we can easily verify from the definition following (2.3). Thus we can write

$$h_{ii} = \sum_{j=1}^n h_{ij}^2 = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2, \quad (2.6)$$

and it is immediately clear that  $0 \leq h_{ii} \leq 1$ . These limits are helpful in understanding and interpreting  $h_{ii}$ , but they do not yet tell us when  $h_{ii}$  is large. We know, however, that the eigenvalues of a projection matrix are either zero or one and that the number of nonzero eigenvalues is equal to the rank of the matrix. In this case,  $\text{rank}(H) = \text{rank}(X) = p$ , and hence  $\text{trace}(H) = p$ , i.e.,

$$\sum_{i=1}^n h_i = p. \quad (2.7)$$

The average size of a diagonal element of the hat matrix, then, is  $p/n$ . Experience suggests that a reasonable rule of thumb for large  $h_i$  is  $h_i > 2p/n$ . Thus we determine high-leverage points by looking at the diagonal elements of  $H$  and paying particular attention to any  $x$  point for which  $h_i > 2p/n$ . Usually we treat the  $n$  values  $h_i$  as a batch of numbers and bring them together in a stem-and-leaf display (as we shall illustrate in Section 4). For a more refined screening when the model includes the constant carrier and the rows of  $X$  are sampled from a  $(p - 1)$  variate Gaussian distribution, we could use the fact that (for any single  $h_i$ )  $[(n - p)(h_i - 1/n)]/[(p - 1)(1 - h_i)]$  has an  $F$  distribution on  $p - 1$  and  $n - p$  degrees of freedom.

From (2.6), we can also see that whenever  $h_{ii} = 0$  or  $h_{ii} = 1$ , we have  $h_{ij} = 0$  for all  $j \neq i$ . These two extreme cases can be interpreted as follows. First, if  $h_{ii} = 0$ , then  $\hat{y}_i$  must be fixed at zero by design—it is not affected by  $y_i$  or by any other  $y_j$ . A point with  $x = 0$  when the model is a straight line through the origin

provides a simple example. Second, when  $h_{ii} = 1$ , we have  $\hat{y}_i = y_i$ —the model always fits this data value exactly. In effect, the model dedicates a parameter to this particular observation (as is sometimes done explicitly by adding a dummy variable to remove an outlier).

Now that we have developed the hat matrix and a number of its properties, we turn to three examples, two designed and one sampled. We then discuss (in Section 5) how to handle  $y_i$  when  $h_{ii}$  indicates a high-leverage point.

### 3. Formal Examples

To illustrate the hat matrix and develop our intuition, we begin with two familiar examples in which the calculations can be done by simple algebra.

The usual regression line,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

has

$$X = \begin{pmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{pmatrix}^T,$$

and a few steps of algebra give

$$h_{ij} = \frac{1}{n} + [(x_i - \bar{x})(x_j - \bar{x})] / \left[ \sum_{k=1}^n (x_k - \bar{x})^2 \right]. \quad (3.1)$$

Next we examine the relationship between structure and leverage in a simple balanced design: a two-way table with  $R$  rows and  $C$  columns and one observation per cell. (Behnken and Draper (1972) discuss variances of residuals in several more complicated designs. It is straightforward to find  $H$  through (2.5).) The usual model for the  $R \times C$  table is

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij},$$

with the constraints  $\alpha_1 + \dots + \alpha_R = 0$  and  $\beta_1 + \dots + \beta_C = 0$ ; here  $n = RC$  and  $p = R + C - 1$ . We could, of course, write this model in the form of (2.1), but it is simpler to preserve the subscripts  $i$  and  $j$  and to denote an element of the hat matrix as  $h_{ij,kl}$ . When we recall that

$$\hat{y}_{ij} = y_{i\cdot} + y_{\cdot j} - y_{..}, \quad (3.2)$$

(a dot in place of a subscript indicates the average with respect to that subscript), it is straightforward to obtain

$$h_{ij,ij} = 1/C + (1/R) - (1/RC) = (R + C - 1)/RC; \quad (3.3)$$

$$h_{ij,il} = (R - 1)/RC, \quad l \neq j; \quad (3.4)$$

$$h_{ij,kj} = (C - 1)/RC, \quad k \neq i; \quad (3.5)$$

$$h_{ij,kl} = -(1/RC), \quad k \neq i, l \neq j. \quad (3.6)$$

From (3.3) we see that all the diagonal elements of  $H$  are equal, as we would expect in a balanced design. Further, (3.3) through (3.6) show that  $\hat{y}_{ij}$  will be affected by any change in  $y_{kl}$  for any values of  $k$  and  $l$ .

#### 4. A Numerical Example

In this section we examine the hat matrix in a regression example, emphasizing (either here or in Section 5) the connections between it and other sources of diagnostic information. We use a ten-point example, for which we can easily present  $H$  in full. In a larger data set, we would generally work with only the diagonal elements,  $h_i$ . Welsch and Kuh (1977) discuss a larger example.

The data for this example come from Draper and Stoneman (1966); we reproduce it in Table 1. The response is strength, and the carriers are the constant, specific gravity, and moisture content. To probe the relationship between the nonconstant carriers, we plot moisture content against specific gravity (Figure A). In this plot, point 4, with coordinates (0.441, 8.9), is to some extent a bivariate outlier (its value is not extreme for either carrier), and we should expect it to have substantial leverage on the fit. Indeed, if this point were absent, it would be considerably more difficult to distinguish the two carriers.

1. Data on Wood Beams

beam number	specific gravity	moisture content	strength
1	0.499	11.1	11.14
2	0.558	8.9	12.74
3	0.604	8.8	13.13
4	0.441	8.9	11.51
5	0.550	8.8	12.38
6	0.528	9.9	12.60
7	0.418	10.7	11.13
8	0.480	10.5	11.70
9	0.406	10.5	11.02
10	0.467	10.7	11.41

The hat matrix for this  $X$  appears in Table 2, and a stem-and-leaf display (Tukey 1972b, 1977) of the diagonal elements (rounded to multiples of .01) is as follows:

0	
1	559
2	456
3	2
4	22
5	
6	0

2. The Hat Matrix for the Wood Beam Data (lower triangle omitted by symmetry)

i	1	2	3	4	5	6	7	8	9	10
j										
1	.418	-.002	.079	-.274	-.046	.181	.128	.222	.050	.242
2		.242	.292	.136	.243	.128	-.041	.033	-.035	.004
3			.417	-.019	.273	.187	-.126	.044	-.153	.004
4				.604	.197	-.038	.168	-.022	.275	-.028
5					.252	.111	-.030	.019	-.010	-.010
6						.148	.042	.117	.012	.111
7							.262	.145	.277	.174
8								.154	.120	.168
9									.315	.148
10										.187

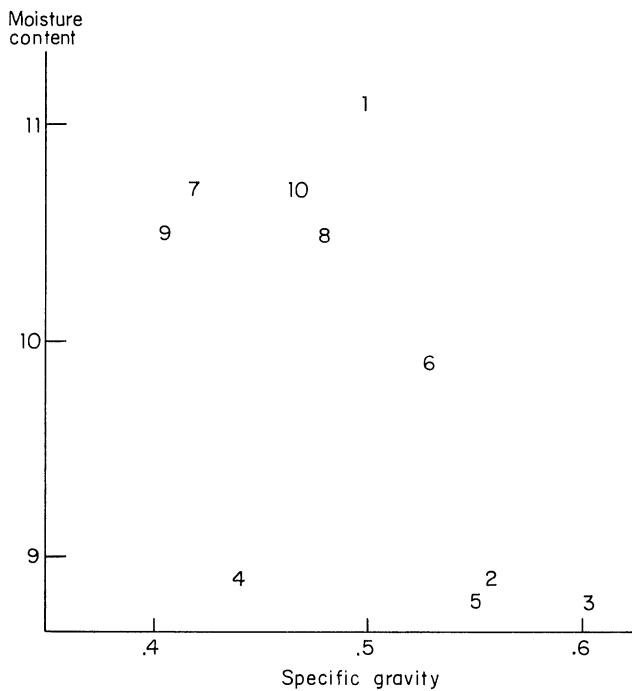


Figure A. The Two Carriers for the Wood Beam Data (Plotting symbol is beam number.).

We note that  $h_4$  is the largest diagonal element and that it just exceeds the level ( $2p/n = 6/10$ ) set by our rough rule of thumb. Examining  $H$  element by element, we find that it responds to the other qualitative features of Figure A. For example, the relatively high leverage of points 1 and 3 reflects their position as extremes in the scatter of points. The moderate negative value of  $h_{1,4}$  is explained by the positions of points 1 and 4 on opposite sides of the rough sloping band where the rest of the points lie. The moderate positive values of  $h_{1,8}$  and  $h_{1,10}$  show the mutually reinforcing positions of these three points. The central position of point 6 accounts for its low leverage. Other noticeable values of  $h_{ij}$  have similar explanations.

Having identified point 4 as a high-leverage point in this data set, it remains to investigate the effect of its position and response value on the fit. Does the model fit well at point 4, or should this point be set aside? We turn to these questions next.

## 5. Bringing in the Residuals

So far we have examined the design matrix  $X$  for evidence of points where the data value  $y$  has high leverage on the fitted value  $\hat{y}$ . If such influential points are present, we must still determine whether they have had any adverse effects on the fit. A discrepant value of  $y$ , especially at an influential design point, may lead us to set that entire observation aside (planning to investigate it in detail separately) and refit without it, but we emphasize that such decisions cannot be made automatically. As we can see for the regression line, with  $h_{ij}$  given by (3.1), the more extreme design points generally provide the greatest information on certain coefficients (in this case, the slope), and omitting such an observation may substantially reduce the precision with which we can estimate those coefficients. If we delete row  $i$ , that is,  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$ , from the design matrix  $X$  and denote the result by  $X_{(i)}$ , then (Rao 1965, p. 29), except for the constant factor  $\sigma^2$ , the covariance matrix of  $\mathbf{b}$  is

$$(X_{(i)}^T X_{(i)})^{-1} = (X^T X)^{-1} + (X^T X)^{-1} \mathbf{x}_i \mathbf{x}_i^T (X^T X)^{-1} / (1 - h_i). \quad (5.1)$$

The presence of  $(1 - h_i)$  in the denominator shows how removing a high-leverage point may increase the variance of coefficient estimates. Alternatively, the accuracy of the apparently discrepant point may be beyond question, so that dismissing it as an outlier would be unacceptable. In both these situations, then, the apparently discrepant point may force us to question the adequacy of the model.

In detecting discrepant  $y$  values, we always examine the residuals,  $r_i = y_i - \hat{y}_i$ , using such techniques as a scatterplot against each carrier, a scatterplot against  $\hat{y}$ , and a normal probability plot. (Anscombe (1973) has discussed and illustrated some of these.) When there is substantial variation among the  $h_i$  values, (2.5) indicates that we should allow for differences in the variances of the  $r_i$  (Anscombe and Tukey 1963) and look at  $r_i/(1 - h_i)^{1/2}$ . This adjustment puts the residuals on an equal footing, but it is often more convenient to use the *standardized residual*,  $r_i/(s(1 - h_i)^{1/2})$ , where  $s^2$  is the residual mean square.

For diagnostic purposes, we would naturally ask about the size of the residual corresponding to  $y_i$  when data point  $i$  has been omitted from the fit. That is, we base the fit on the remaining  $n - 1$  data points and then predict the value for  $y_i$ . This residual is  $y_i - \mathbf{x}_i \hat{\beta}_{(i)}$ , where  $\hat{\beta}_{(i)}$  is the least-squares estimate of  $\beta$  based on all the data except data point  $i$ . (These residuals are also the basis of Allen's (1974) PRESS criterion for selecting variables in regression.) Similarly  $s_{(i)}^2$  is the residual mean square for the "not- $i$ " fit, and the standard deviation of  $y_i - \mathbf{x}_i \hat{\beta}_{(i)}$  is estimated by  $s_{(i)} [1 + \mathbf{x}_i (X_{(i)}^T X_{(i)})^{-1} \mathbf{x}_i^T]^{1/2}$ . We now define the *Studentized residual*:

$$r_i^* = \frac{y_i - \mathbf{x}_i \hat{\beta}_{(i)}}{s_{(i)} [1 + \mathbf{x}_i (X_{(i)}^T X_{(i)})^{-1} \mathbf{x}_i^T]^{1/2}}. \quad (5.2)$$

Since the numerator and denominator in (5.2) are independent,  $r_i^*$  has a  $t$  distribution on  $n - p - 1$  degrees of freedom, and we can readily assess the significance of any single Studentized residual. (Of course,  $r_i^*$  and  $r_j^*$  will not be independent.) In actually calculating the Studentized residuals we can save a great deal of effort by observing that the quantities we need are readily available. Straightforward algebra using (5.1) turns (5.2) into

$$r_i^* = r_i / (s_{(i)} (1 - h_i)^{1/2}), \quad (5.3)$$

and we can obtain  $s_{(i)}$  from

$$(n - p - 1)s_{(i)}^2 = (n - p)s^2 - r_i^2 / (1 - h_i). \quad (5.4)$$

Once we have the diagonal elements of  $H$ , the rest is simple.

Our diagnostic strategy, then, is to examine the  $h_i$  for high-leverage design points and the  $r_i^*$  for discrepant  $y$  values. These two aspects of the search for troublesome data points are complementary; neither is sufficient by itself. When  $h_i$  is small,  $r_i^*$  may be large because  $r_i$  is large, but the impact of  $y_i$  on the fit or on the coefficients may be minor. And when  $h_i$  is large,  $r_i^*$  may still be moderate or small because  $y_i$  is consistent with the model and the rest of the data.

Just how to combine the information from  $h_i$  and  $r_i^*$  is a matter of judgment. We prefer the more detailed grasp of the data which comes from looking at the  $h_i$  and the  $r_i^*$  separately. For diagnostic purposes, a practice which we recommend is to tag as exceptional any data point for which  $h_i$  or  $r_i^*$  is significant at the 10 percent level. To decide whether an exceptional point is actually damaging, one would then use a criterion which is appropriate in the context of the data. Two likely criteria are the change in coefficients,  $\hat{\beta} - \hat{\beta}_{(i)}$ , easily calculated from

$$\hat{\beta} - \hat{\beta}_{(i)} = (X^T X)^{-1} \mathbf{x}_i^T r_i / (1 - h_i); \quad (5.5)$$

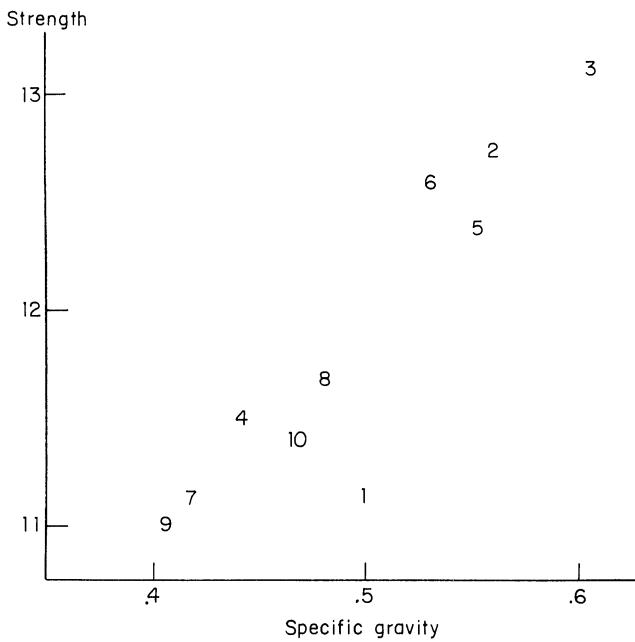
and the change in fit at point  $i$ ,  $\mathbf{x}_i(\hat{\beta} - \hat{\beta}_{(i)})$ , which simply reduces to  $h_i r_i / (1 - h_i)$ . (The size of such changes would customarily be compared to some suitable measure of scale.) For both of these criteria it is easy to determine the effect of setting aside an exceptional point without recalculation.

To continue our diagnosis of the wood beam example, we plot strength against specific gravity in Figure B and strength against moisture content in Figure C. With the exception of beam 1, the first of these looks quite linear and well-behaved. In the second plot we see somewhat more scatter, and beam 4 stands apart from the rest. Table 3 gives  $r_i$ ,  $(1 - h_i)^{1/2}$ ,  $s_{(i)}$ , and the Studentized residuals  $r_i^*$ . Among the  $r_i^*$ , beam 1 appears as a clear stray ( $p < .02$ ), and beam 6 also deserves attention ( $p < .1$ ). Since beam 4 is known to have high leverage ( $h_i = .604$ ), we continue to investigate it.

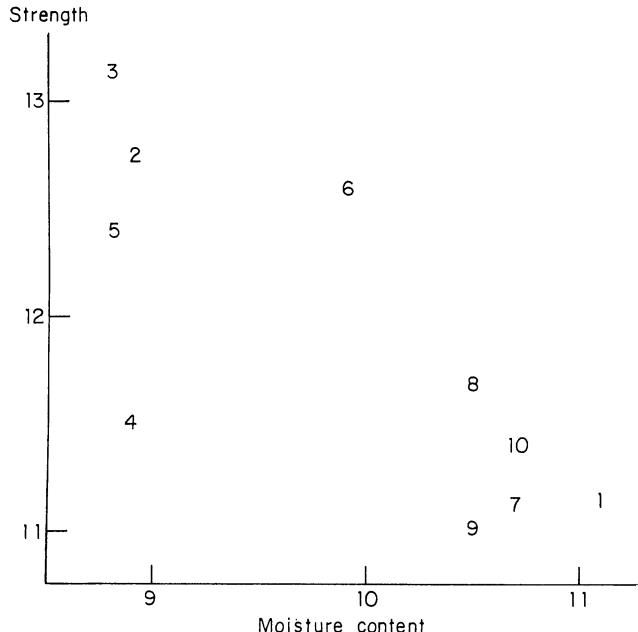
The fit for the full data is

$$\hat{y} = 10.302 + 8.495(SG) - 0.2663(MC), \quad (5.6)$$

with  $s = 0.2753$ ; and when we set aside beams 1, 4,



**Figure B.** Strength versus Specific Gravity for the Wood Beam Data (Plotting symbol is beam number.).



**Figure C.** Strength versus Moisture Content for the Wood Beam Data (Plotting symbol is beam number.).

and 6 in turn, we find  $\hat{\beta} - \hat{\beta}_{(i)}$  to be  $(2.710, -1.772, -0.1932)^T$ ,  $(-2.109, 1.695, 0.1242)^T$ , and  $(-0.642, 0.748, 0.0329)^T$ , respectively. The estimated standard errors for  $\hat{\beta}_0$ ,  $\hat{\beta}_1$ , and  $\hat{\beta}_2$  are 1.896, 1.784, and 0.1237, so that setting aside either beam 1 or beam 4 causes each coefficient to change by roughly 1.0 to 1.5 in standard-error units. Thus we should be reluctant to include these data points. By comparison, removing beam 6 leads to changes only about 25 percent as large.

Similarly, the change in fit at point  $i$ ,  $x_i(\hat{\beta} - \hat{\beta}_{(i)})$ , is  $-0.319$  for beam 1,  $-0.256$  for beam 4, and  $0.078$  for

### 3. Studentized Residuals and Related Quantities for the Wood Beam Data

i	$r_i$	$h_i$	$(1-h_i)^{1/2}$	$s_{(i)}$	$r_i^*$
1	-.444	.418	.763	.179	-3.254
2	.069	.242	.871	.296	.267
3	.041	.417	.764	.297	.182
4	-.167	.604	.629	.277	-.961
5	-.250	.252	.865	.273	-1.058
6	.450	.148	.923	.221	2.203
7	.127	.262	.859	.291	.509
8	.117	.154	.920	.293	.436
9	.066	.315	.828	.296	.270
10	-.009	.187	.902	.298	-.033

beam 6. Dividing each of these by the estimated standard error of  $\hat{y}_i$  ( $s\sqrt{h_i}$  from (2.4)) yields  $-1.790$ ,  $-1.196$ , and  $0.737$ , respectively. On the whole these are not as substantial as the coefficient changes, but beam 1 and (to a lesser extent) beam 4 are still fairly damaging.

We have used two sources of diagnostic information, the diagonal elements of the hat matrix and the Studentized residuals, to identify data points which may have an unusual impact on the results of fitting the linear model (2.1) by least squares. We must interpret this information as clues to be followed up to determine whether a particular data point is discrepant, but not as automatic guidance for discarding observations. Often the circumstances surrounding the data will provide explanations for unusual behavior, and we will be able to reach a much more insightful analysis. Judgment and external sources of information can be important at many stages. For example, if we were trying to decide whether to include moisture content in the model for the wood beam data (the context in which Draper and Stone-man (1966) introduced this example), we would have to give close attention to the effect of beam 4 on the correlation between the carriers as well as the correlation between the coefficients. Such considerations do not readily lend themselves to automation and are an important ingredient in the difference between data analysis and data processing (Tukey 1972a).

## 6. Computation

Since we find the hat matrix (at least the diagonal elements  $h_i$ ) a very worthwhile diagnostic addition to the information usually available in multiple regression, we now briefly describe how to obtain  $H$  from the more accurate numerical techniques for solving least-squares problems. Just as these techniques provide greater accuracy by not forming  $X^T X$  or solving the normal equations directly, we do not calculate  $H$  according to the definition.

For most purposes the method of choice is to represent  $X$  as

$$X = Q R, \quad (6.1)$$

(with  $Q$  an orthogonal transformation and  $R = [\tilde{R}^T$ ,

$0^T]^T$ , where  $\tilde{R}$  is  $p \times p$  upper triangular) and obtain  $Q$  as a product of Householder transformations. Substituting (6.1) and the special structure of  $R$  into the definition of  $H$ , we see that

$$H = Q \begin{bmatrix} I_p & 0 \\ 0 & 0 \end{bmatrix} Q^T. \quad (6.2)$$

With a modest increase in computation cost, a simple modification of the basic algorithm yields  $H$  as a by-product. If  $n$  is large, we can arrange to calculate and store only the  $h_i$ .

Finally we mention the singular-value decomposition,

$$\underset{n \times p}{X} = \underset{n \times p}{U} \underset{p \times p}{\Sigma} \underset{p \times p}{V^T}, \quad (6.3)$$

where  $U^T U = I_p$ ,  $\Sigma$  is diagonal, and  $V$  is orthogonal. If this more elaborate approach is used (e.g., when  $X$  might not be of full rank), we can calculate the hat matrix from

$$H = U U^T. \quad (6.4)$$

These and other decompositions are discussed by Golub (1969). For a recent account of numerical techniques in solving linear least-squares problems, we recommend the book by Lawson and Hanson (1974).

[Received October 18, 1976. Revised June 9, 1977.]

## References

- Allen, D. M. (1974), "The Relationship Between Variable Selection and Data Augmentation and a Method for Prediction," *Technometrics*, 16, 125-127.
- Anscombe, F. J. (1973), "Graphs in Statistical Analysis," *The American Statistician*, 27, 17-21.
- , and Tukey, J. W. (1963), "The Examination and Analysis of Residuals," *Technometrics*, 5, 141-160.
- Behnken, D. W., and Draper, N. R. (1972), "Residuals and Their Variance Patterns," *Technometrics*, 14, 101-111.
- Draper, N. R., and Stoneman, D. M. (1966), "Testing for the Inclusion of Variables in Linear Regression by a Randomisation Technique," *Technometrics*, 8, 695-699.
- Golub, G. H. (1969), "Matrix Decompositions and Statistical Calculations," in *Statistical Computation*, eds. R. C. Milton and J. A. Nelder, New York: Academic Press.
- Huber, P. J. (1975), "Robustness and Designs," in *A Survey of Statistical Design and Linear Models*, ed. J. N. Srivastava, Amsterdam: North-Holland Publishing Co.
- Lawson, C. L., and Hanson, R. J. (1974), *Solving Least Squares Problems*, Englewood Cliffs, N.J.: Prentice-Hall.
- Rao, C. R. (1965), *Linear Statistical Inference and Its Applications*, New York: John Wiley & Sons.
- Tukey, J. W. (1972a), "Data Analysis, Computation and Mathematics," *Quarterly of Applied Mathematics*, 30, 51-65.
- (1972b), "Some Graphic and Semigraphic Displays," in *Statistical Papers in Honor of George W. Snedecor*, ed. T. A. Bancroft, Ames, Iowa: Iowa State University Press.
- (1977), *Exploratory Data Analysis*, Reading, Mass.: Addison-Wesley Publishing Co.
- Welsch, R. E., and Kuh, E. (1977), "Linear Regression Diagnostics," Working Paper 173, Cambridge, Mass.: National Bureau of Economic Research.

# Interpreting the Coefficients on Dummy Variables

## *Definition of a Dummy Variable*

A dummy variable is an artificial variable constructed such that it takes the value unity (one) whenever the category it represents occurs, and zero otherwise.

## *Interpretation of the Coefficient on a Dummy Variable, Reference Category Method*

Suppose D is a 0-1 variable in the regression model

$$\hat{Y} = b_0 + b_1 D + b_2 X$$

Relative to the regression line for the reference category, where D=0, the line for the category where D=1 is parallel and  $b_1$  units higher.

## *Interpretation of the Coefficients on Dummy Variables, All Categories Included*

Suppose  $D_1, D_2, \dots, D_k$  are a set of mutually exclusive and exhaustive dummy variables for a factor that has  $k$  categories in the regression model

$$\hat{Y} = b_1 D_1 + b_2 D_2 + \dots + b_k D_k + b_{k+1} X .$$

Then each of the coefficients,  $b_1, b_2, \dots, b_k$ , is the intercept estimate for the category that the respective dummy variable represents.

## *Interpretation of the Interaction Coefficient, Reference Category Method*

Suppose D is a 0-1 dummy variable, and DX is the product of the dummy variable D and the continuous variable X in the regression model

$$\hat{Y} = b_0 + b_1 D + b_2 X + b_3 DX .$$

Then  $b_3$  is the estimate of the difference in the slope between the two categories, that is, the estimated slope for the reference category where  $D = 0$  is  $b_2$ , and the estimated slope for the category where  $D = 1$  is  $b_2 + b_3$ .

# Chapter 18

## Regression - hockey sticks, broken sticks, piecewise, change points

### Contents

---

<b>18.1</b>	<b>Hockey-stick, piecewise, or broken-stick regression</b>	<b>1146</b>
18.1.1	Example: Nenana River Ice Breakup Dates	1147
<b>18.2</b>	<b>Searching for the change point</b>	<b>1151</b>
18.2.1	Change point model for the Nenana River Ice Breakup	1152
<b>18.3</b>	<b>What is the first time that a treatment mean differ from a control mean</b>	<b>1157</b>
18.3.1	How long does a bait last for attracting ants?	1158

---

A simple regression analysis assumes that the change in response is the same across the range of  $X$  values. In some cases, a model where the slope changes in different parts of the  $X$  space may be biologically more realistic.

This chapter examines two cases of fitting regression lines with breaks in the slope. In the first case, the location of the change in slope is known in advance; the second cases also estimates the location of the change, also known as the change point problem.

The examples in this chapter look at cases with a single change point – the extension to multiple change points (both known and unknown) is straightforward. Similarly, the change from linear to quadratic lines is also straightforward.

A related method, a spline fit to the data, where a flexible curve is fit between (evenly) spaced knot points that is a like a non-parametric curve fit is explored in a different chapter.

### 18.1 Hockey-stick, piecewise, or broken-stick regression

In this section, the location of the change point is known. The statistical model is:

$$Y = \beta_0 + \beta_1(X) + \beta_2(X - C)^+ + \epsilon$$

where  $\beta_0$  is the intercept,  $\beta_1$  is the slope before the change point  $C$ , and  $\beta_2$  is the DIFFERENCE in slope after the change point. The slope after the change point is  $\beta_1 + \beta_2$ . The variable  $(X - C)^+$  is a derived variable which takes the value of 0 for values of  $X$  less than  $C$  and the values  $X - C$  for values of  $X$  greater than  $C$ . This is usually created using a Formula Editor based on the actual data.

The hypothesis of interest is  $H : \beta_2 = 0$  which indicates no change in slope between  $X < C$  and  $X > C$ .

Because the value of  $C$  is specified in advance, ordinary least-squares can be used to fit the model. Most computer packages can easily fit this model.

### 18.1.1 Example: Nenana River Ice Breakup Dates

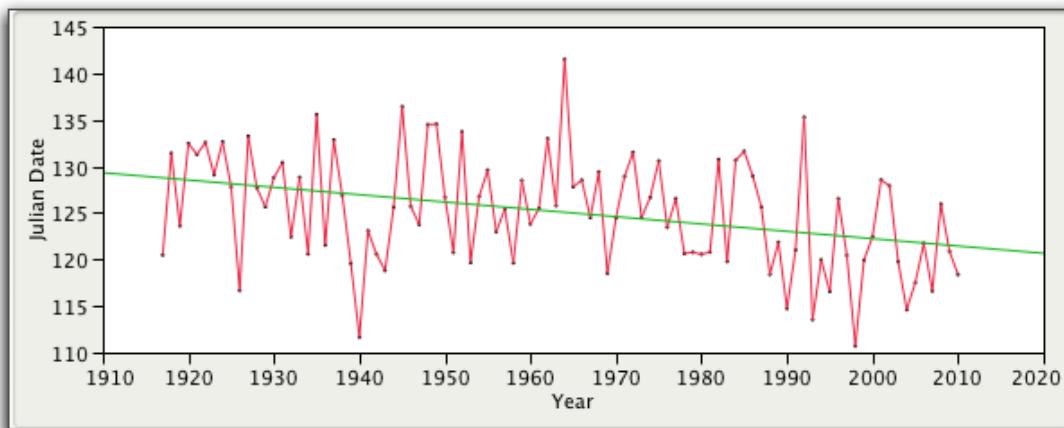
The Nenana river in the Interior of Alaska usually freezes over during October and November. The ice continues to grow throughout the winter accumulating an average maximum thickness of about 110 cm, depending upon winter weather conditions. The Nenana River Ice Classic competition began in 1917 when railroad engineers bet a total of 800 dollars, winner takes all, guessing the exact time (month, day, hour, minute) ice on the Nenana River would break up. Each year since then, Alaska residents have guessed at the timing of the river breakup. A tripod, connected to an on-shore clock with a string, is planted in two feet of river ice during river freeze-up in October or November. The following spring, the clock automatically stops when the tripod moves as the ice breaks up. The time on the clock is used as the river ice breakup time. Many factors influence the river ice breakup, such as air temperature, ice thickness, snow cover, wind, water temperature, and depth of water below the ice. Generally, the Nenana river ice breaks up in late April or early May (historically, April 20 to May 20). The time series of the Nenana river ice breakup dates can be used to investigate the effects of climate change in the region.

In 2010, the jackpot was almost \$300,000 and the ice went out at 9:06 on 2010-04-29. In 2012, the jackpot was over \$350,000 and the ice went out at 19:39 on 2012-04-23 - as reported at <http://www.cbc.ca/news/offbeat/story/2012/05/02/alaska-ice-contest.html>. The latest winner, Tommy Lee Waters, has also won twice before, but never has been a solo winner. Waters spent time drilling holes in the area to measure the thickness of the ice. Altogether he spent \$5,000 on tickets for submitting guesses (he purchased every minute of the afternoon of 23 April) and spent an estimated 1,200 hours working out the math by hand. And, it was also his birthday! (What are the odds?) You too can use statistical methods to gain fame and fortune!

More details about the Ice Classic are available at <http://www.nenanaakiceclassic.com>.

The data are available in the *nenana.csv* data file available in the the Sample Program Library at <http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms>. It is also available in the *nenana.jmp* data file.

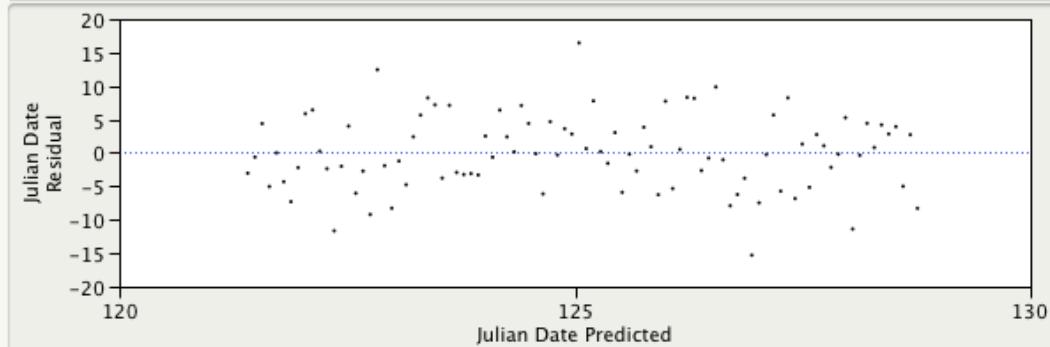
A simple regression line fit to the time of break up with year as the predictor show evidence of a decline over time (i.e. the time of breakup is tending to occur earlier) and there is no evidence of auto-correlation.



### Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t	Lower 95%	Upper 95%
Intercept	280.27503	42.0968	6.66	<.0001*	196.66716	363.8829
Year	-0.079037	0.021438	-3.69	0.0004*	-0.121614	-0.03646

### Residual by Predicted Plot



### Durbin-Watson

Durbin- Watson	Number of Obs.	AutoCorrelation	Prob<DW
1.9345786	94	0.0194	0.3356

A closer inspection of the top graph gives the impression that until about 1970, the regression line was “flat” and only after 1970 did the time of breakup seem to decrease.

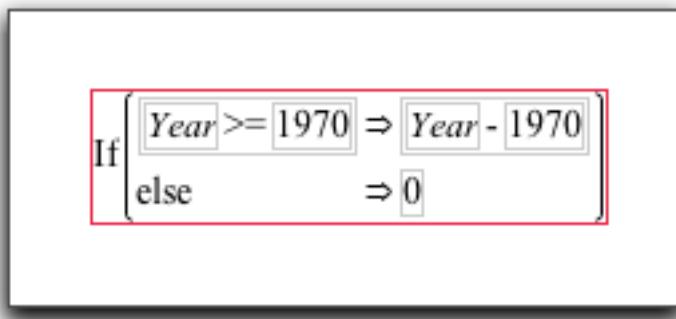
A broken stick model (separate slopes in the pre-1970 and the post-1970 eras) can be easily fit. The statistical model is:

$$JulianDate = \beta_0 + \beta_1(Year) + \beta_2(Year - 1970)^+ + \epsilon$$

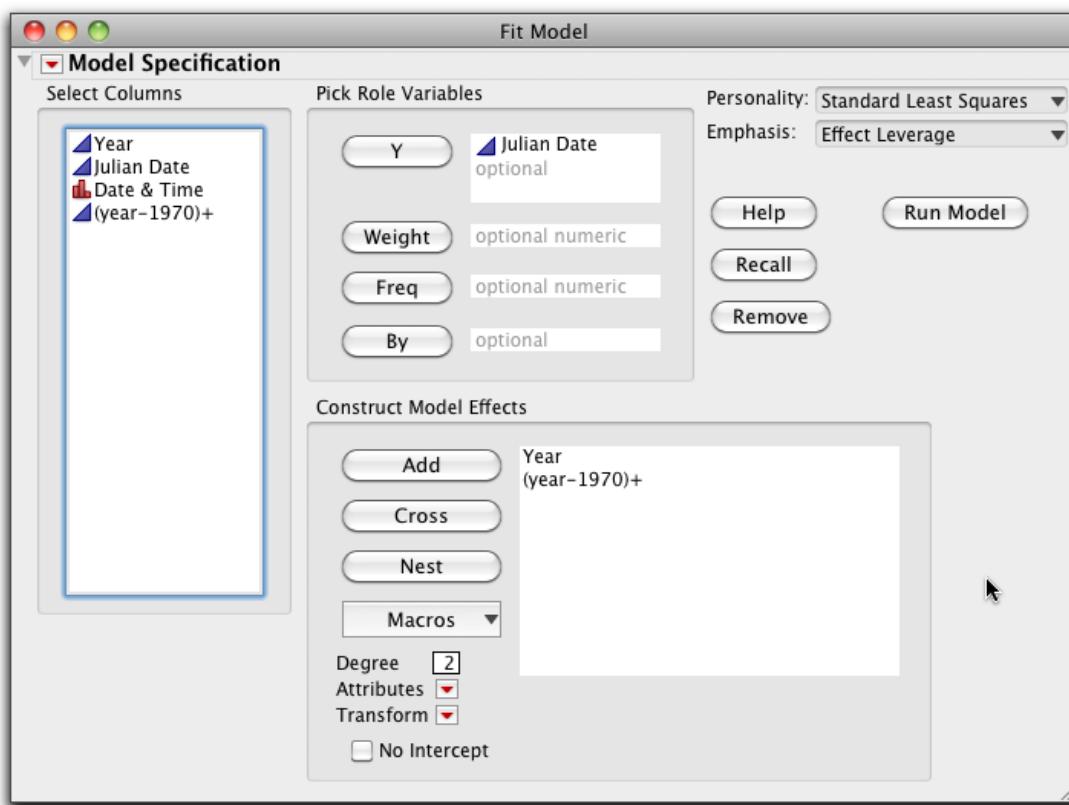
where  $JulianDate$  is the date of breakup,  $Year$  is the calendar year. The parameters to be estimated are  $\beta_0$  the intercept,  $\beta_1$  the change in the breakup prior to the change point,  $\beta_2$  the change in slope after the change point which is assumed to happen in 1970. The term  $(Year - 1970)^+$  takes the value 0 if the

argument is negative (i.e. before 1970); and the value of the argument if it is positive.

To fit this model, we need to create a new variable that is zero for the pre-1970 period and equal to  $(year - 1970)$  in the post 1970 period. This is easily created in *JMP* using the Formula Editor:



The change point model with a known change point is then fit using standard multiple regression. In *JMP*, this is done using the *Analyze->Fit Model* platform:



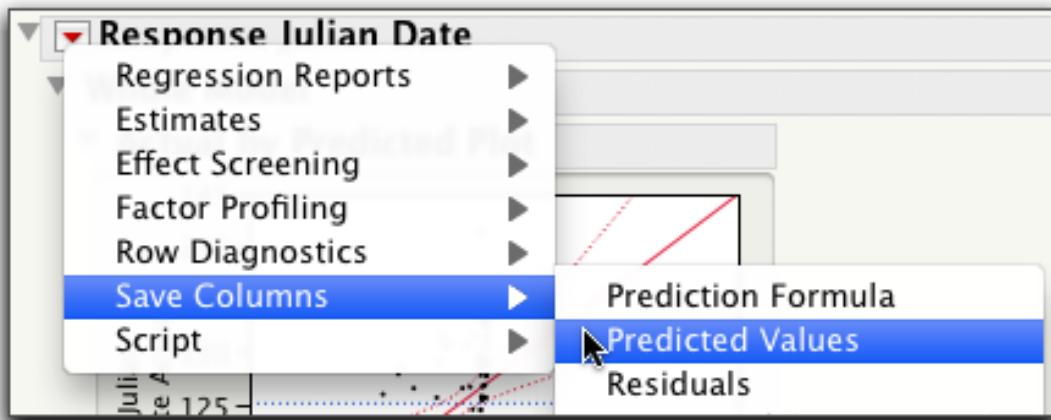
which gives the estimates:

Parameter Estimates						
Term	Estimate	Std Error	t Ratio	Prob> t	Lower 95%	Upper 95%
Intercept	146.49586	78.71607	1.86	0.0660	-9.863939	302.85565
Year	-0.010133	0.040417	-0.25	0.8026	-0.090417	0.0701507
(year-1970) <sup>+</sup>	-0.173596	0.086854	-2.00	0.0486*	-0.346121	-0.001072

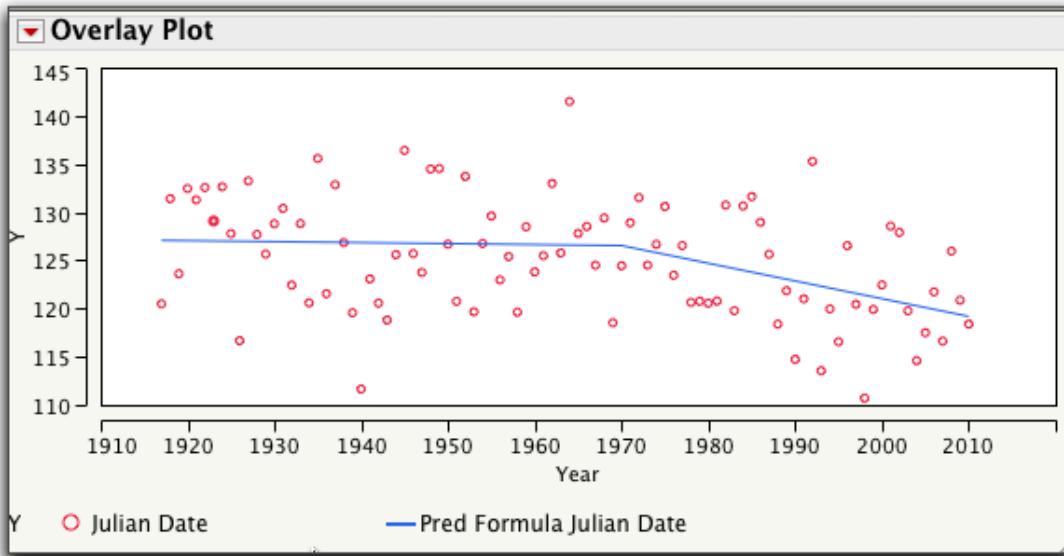
A test for differential slopes in the two eras is then equivalent to a test if  $\beta_2 = 0$ .

In this case the *p*-value for the  $\beta_2$  coefficient (associated with the  $(year - 1970)^+$  variable) is just under 0.05 providing some evidence of a different slope in the two eras.

A plot of the fitted model is obtained by saving the predicted values to the data table:



and then plotting the actual data and the fitted points on the same graph using the *Graph->Overlay* platform:



Confidence intervals for the MEAN response in a particular year (not likely of interest in this example) and for the individual responses in a particular year are generated in the usual way.

Note that the estimated slope for the pre-1970 era is not statistically different from 0. If you wanted to fit a model where the line was flat (i.e. the slope was 0) in the pre-1970 era, this is done by using only the  $(year - 1970)^+$  variable. Many of the automatically generated plots look odd (e.g. all of the points appear to be replotted at 1970), the intercept has a different interpretation in the two models because  $year = 0$  has a different definition in the two models, but if the fitted model is plotted against the original year variable everything works out properly. In this particular case, the two latter models give predicted lines that are almost identical. It is quite RARE that you would fit a line where the slope is known to be zero in practice, but see the next example for a case where it is sensible.

## 18.2 Searching for the change point

. In the previous section on segmented regression (also known as hockey-stick regression or broken-stick regression), the locations of the break are assumed to be known. In many cases, the location of the break is not known, and it is of interest to estimate the break point as well.

The problems of identifying changes at unknown times and of estimating the location of changes is known as “the change-point problem”. Numerous methodological approaches have been implemented in examining change-point models. Maximum-likelihood estimation, Bayesian estimation, isotonic regression, piecewise regression, quasi-likelihood and non-parametric regression are among the methods which have been applied to resolving challenges in change-point problems. Grid-searching approaches have also been used to examine the change-point problem. A review of the literature especially as it applies to regression problem (as of 2008) is available at: <http://biostats.bepress.com/cgi/viewcontent.cgi?article=1075&context=cobra>.

The standard change-point problem in regression models consists of

- testing the null hypothesis that no change in regimes has taken place against the alternative that observations were generated by two (or possibly more) distinct regression equations, and
- estimating the two regimes that gave rise to the data.

There are two common models. First are models where the regression line is continuous at the break point, and models where the regression line can be discontinuous. In these notes, we only consider the continuous case.

This problem has a long history. A nice summary and treatment of the problem is available in

Toms, J. D. and Lesperance, M L. (2003).  
 Piecewise regression: A tool for identifying ecological thresholds.  
*Ecology*, 84, 2034-2041  
<http://dx.doi.org/10.1890/02-0472>.

The change point model starts with the broken-stick model seen earlier, i.e.

$$Y = \beta_0 + \beta_1(X) + \beta_2(X - C)^+ + \epsilon$$

where  $Y$  is the response variable,  $X$  is the covariate, and  $C$  is the change point, i.e. where the break occurs. This model is appropriate where there is an abrupt transition at the break point, but a smooth

transition may be more realistic for some data. One drawback of this model is that convergence problems can occur in locating  $C$  when the data are sparse around the neighborhood of  $C$ .

Toms and Lesperance (2003) review the use of model with gentler transitions, e.g. the hyperbolic tangent model or the bent-cable model. The bent-cable regression model was recently developed by Chiu, Lockhart and Routledge (2006, Bent-cable regression theory and application, Journal of the American Statistical Association, 101, 542-553). The bent-cable regression model fits a smooth transition between the two linear parts of the model. The latter is also applicable to regression models where the  $X$  variable is time and auto-correlation may be present<sup>1</sup>.

The simple piece-wise linear model can be fit using the *Analyze->Modelling ->NonLinear* platform of *JMP*.

### 18.2.1 Change point model for the Nenana River Ice Breakup

Refer to the previous section about details on the Nenana River Ice Breakup contest. Rather than specifying a break point at 1970, we will fit the change point model to estimate the change point.

The data are available in the *nenana.csv* data table in the Sample Program Library at <http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms>.

The statistical model is:

$$\text{JulianDate} = \beta_0 + \beta_1(\text{Year}) + \beta_2(\text{Year} - C)^+ + \epsilon$$

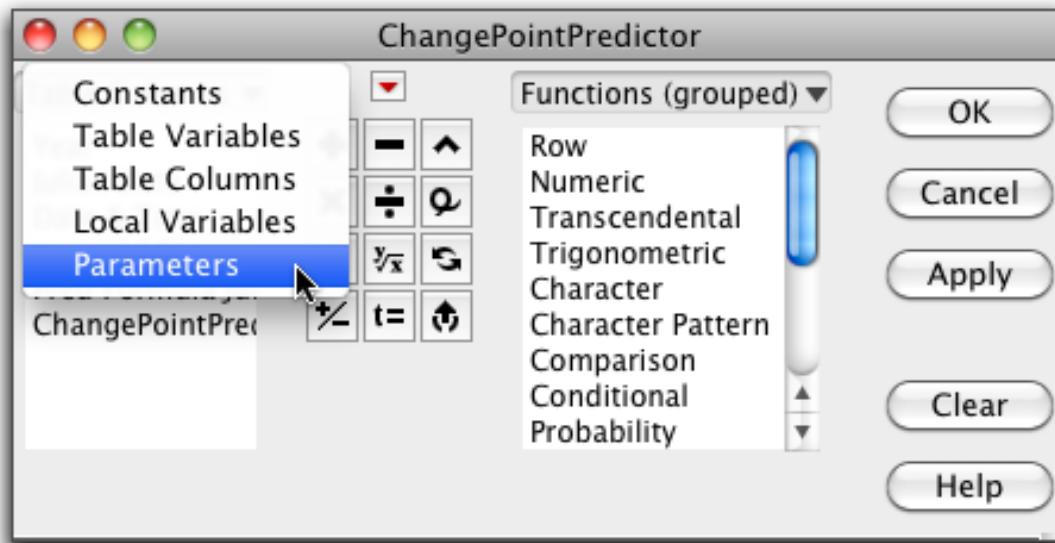
where  $\text{JulianDate}$  is the date of breakup,  $\text{Year}$  is the calendar year. The parameters to be estimated are  $\beta_0$  the intercept,  $\beta_1$  the change in the breakup prior to the change point,  $\beta_2$  the change in slope after the breakup, and  $C$  the change point.

We first need to define the parameters of the model ( $\beta_0, \beta_1, \beta_2, C$ ) and the predicted value in terms of the parameters of the model. We start by creating a new column in the data table *ChangePointPredictor* and start the Formula Editor.

New parameters are defined (along with initial starting guesses), by using the drop-down menu in the top left of the formula editor:

---

<sup>1</sup> Chiu, G. S. and Lockhart, R. L. (2010). Bent-cable regression with auto-regressive noise. Canadian Journal of Statistics, 38, 386-407. <http://dx.doi.org/10.1002/cjs.10070>

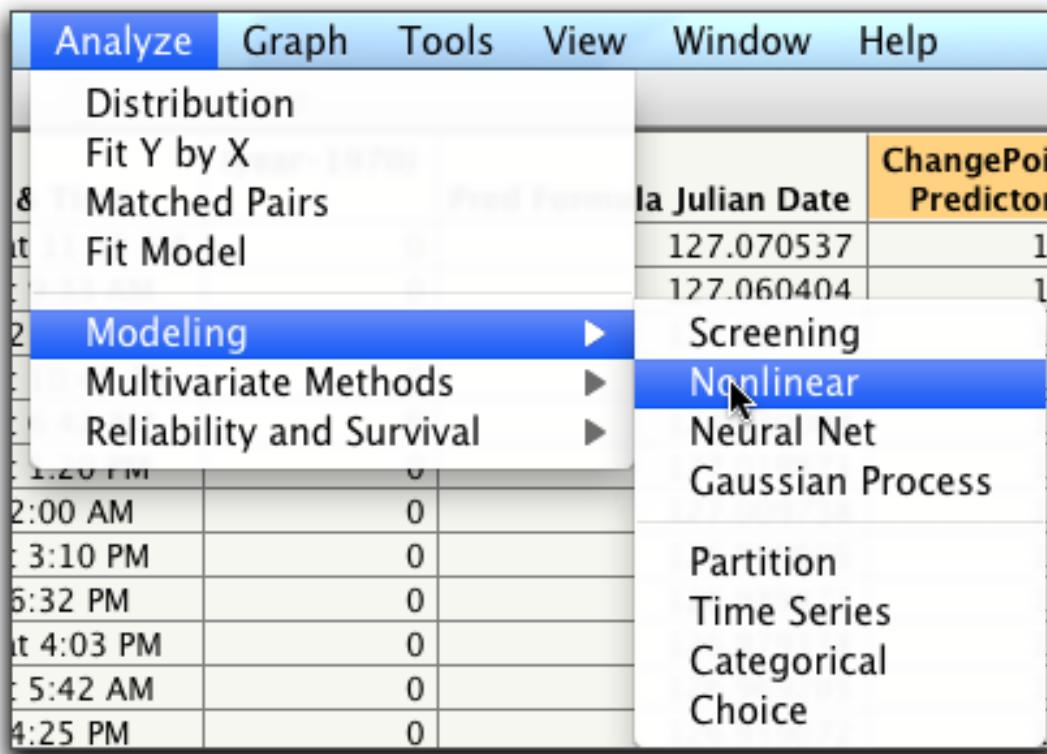


Click on the *New Parameters* item and create the four parameters and their initial values (based on the results from the previous example). The choice of initial values is not that crucial. Then create the predicted value in terms of the parameters and the columns in the data table:

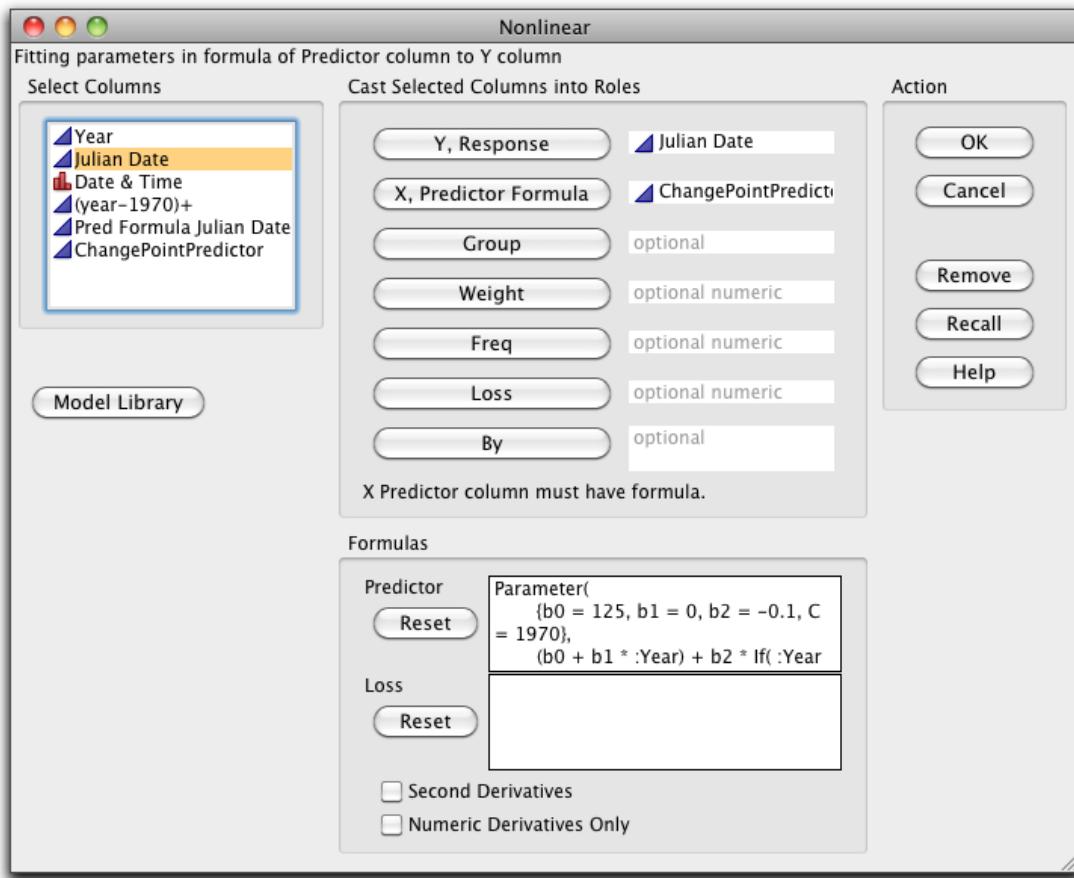
$$(b_0 + b_1 * \text{Year}) + b_2 * \text{If}[\text{Year} \geq C, \text{Year} - C, 0]$$

Notice the use of the *If* function to adjust for the break point. You can switch back and forth between the parameters, data table columns, etc. using the drop down menu in the top right of the formula editor. When you are finished, close the Formula Editor, and the data table will be updated with initial predictions based on the initial values specified.

Select the *Analyze->Modelling ->NonLinear* platform:



Specify the predicted value and  $Y$  variables appropriately:



Notice that the formula for the predictions is displayed.

This brings up the *Analyze->Modelling ->NonLinear* platform control panel. The initial fit is displayed. Press the *Go* button to find the non-linear least squares fit.

**Control Panel**

Converged in Gradient

	Criterion	Current	Stop Limit
Go	Iteration	6	60
Stop	Obj Change	1.0485853e-7	1e-15
Step	Relative Gradient	1.636061e-13	0.000001
	Gradient	1.90711e-10	0.000001

Reset

Parameter	Current Value	Lock
b0	134.0604714	<input type="checkbox"/>
b1	-0.00370333	<input type="checkbox"/>
b2	-0.170525936	<input type="checkbox"/>
C	1967.1558103	<input type="checkbox"/>

Approximate standard errors are also presented at the bottom of the output:

Parameter	Estimate	ApproxStdErr
b0	134.0604714	103.096019
b1	-0.00370333	0.05308602
b2	-0.170525936	0.08672183
C	1967.1558103	13.5792228

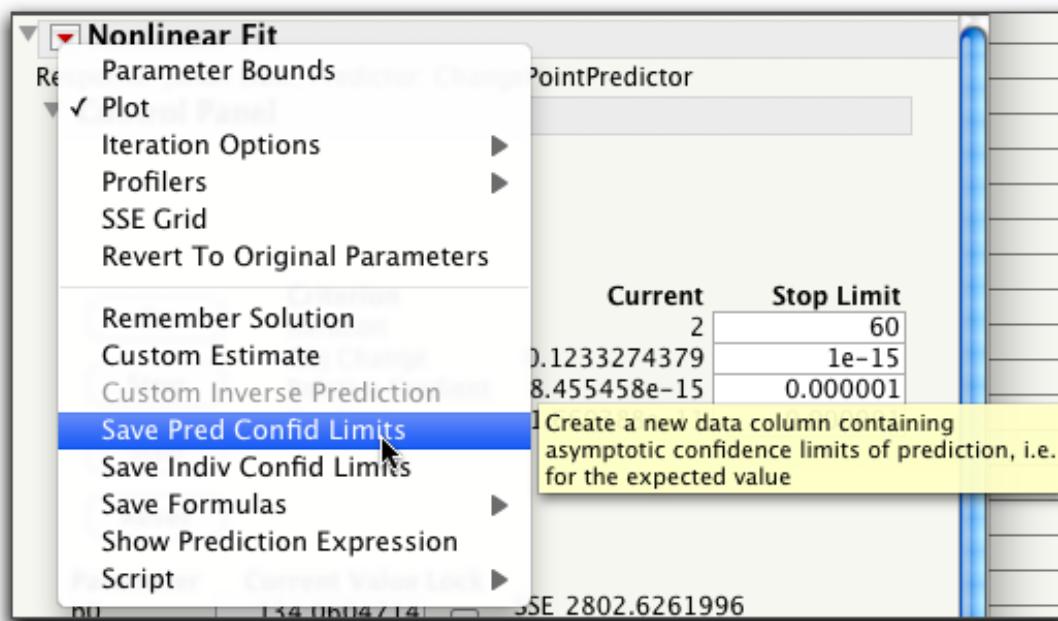
The non-linear least squares algorithm appears to have converged at the estimates listed in the table. The estimated change-point of 1967 is close to the value of 1970 “guess-timated” earlier.

The standard errors are based on large-sample theory. In order to compute a 95% confidence interval for the break point, you could use the standard  $estimate \pm 2(se)$ , but in small samples, the resulting confidence intervals may not perform well. Toms and Lesperance (2003) recommend that a likelihood ratio confidence interval be computed. *JMP* attempts to compute profile-likelihood confidence intervals when you press the *Confidence Interval* button which gives:

Parameter	Estimate	ApproxStdErr	Lower CL	Upper CL
b0	134.0604714	103.096019	.	.
b1	-0.00370333	0.05308602	.	.
b2	-0.170525936	0.08672183	-0.3634129	.
C	1967.1558103	13.5792228	1917.06467	.

In this case, the profile intervals fail to give upper and lower bounds because the slope after the change point is just on the boundary of statistical significance at ( $\alpha = 0.05$ ). If you change the confidence coefficient from 95% to 90%, the procedure is able to find confidence bounds on the  $C$  parameter. Consequently, there may or may not be a change point. Notice that the lower boundary of the confidence interval for  $C$  is quite far below the point estimate!

Confidence intervals for the mean response and prediction intervals for a future response are obtained in the usual way and are interpreted in the same way as in ordinary regression. In *JMP*, these are obtained by clicking on the red triangle:



The *Analyze->Modelling ->NonLinear* platform also allows you to “play” with the estimates to investigate the sensitivity of the fit to the parameters. The *Profiler* option under the red triangle is also useful in these cases.

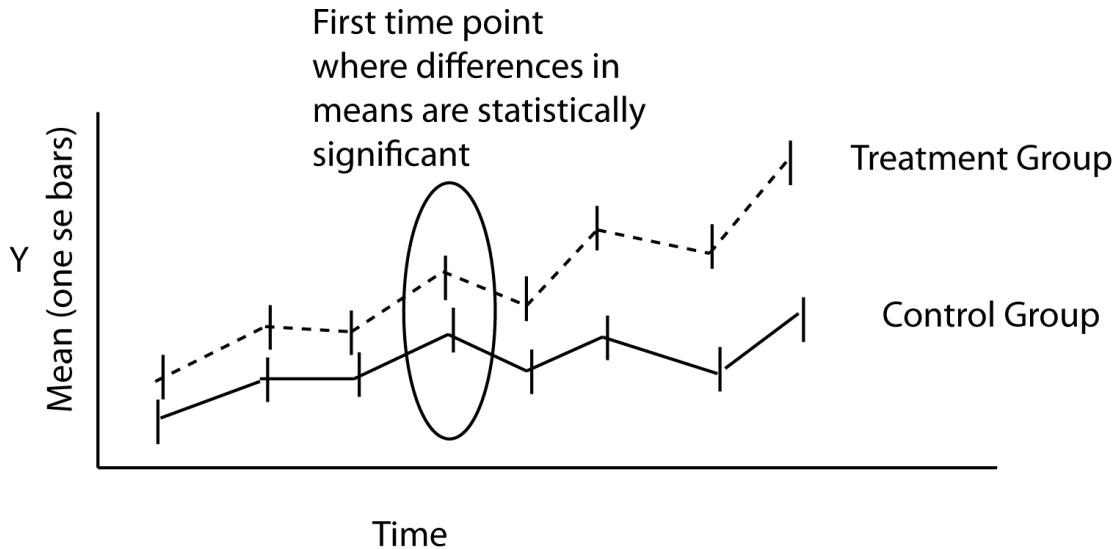
### 18.3 What is the first time that a treatment mean differ from a control mean

A fairly common request in our Statistical Consulting Service is for help in finding the time at which some treatment gives a difference in response from a control. For example, a group of animals may be fed a control diet and are measured over time, while another group of animals are fed an experimental

diet and are measured over time. At which point do the responses between the two groups start to differ?

Let us assume, for simplicity, that separate animals are measured at each time point so that the problem of longitudinal data are ignored. For example, suppose that animals must be sacrificed at each time point to measure the response. A naive analysis starts by plotting the means of the two groups over time and searching for the first time point at which the two means are statistically different:

**An illustration of a naive and WRONG method to search  
for a change point in an experiment.**



This is NOT A VALID ANALYSIS! The problem is that the estimate of the change point for this analysis will depend on the sample size and the alpha level (the cutoff to declare statistical significance). If the sample size is small in each group, then the standard error bars are larger, and the estimated change point tends to be larger than if the sample size is large and the standard errors are smaller. If the alpha level is chosen to be  $\alpha = 0.10$  rather than  $\alpha = 0.05$ , then it is easier to detect an effect and so the estimated change point would once again shift.

The actual change point does NOT depend on sample size! All that should happen is that the estimated precision of the change point problem should be worse for smaller sample sizes than for larger sample sizes.

The proper way to search for a change point is to find the DIFFERENCE or  $\log(\text{RATIO})$  of the means at each time point and then apply the change point analysis to the difference or  $\log(\text{ratio})$ . A model where the difference in means is forced to be zero prior to the unknown change point may be a suitable alternate model.

### 18.3.1 How long does a bait last for attracting ants?

This example is based on a project by Nate Derstine of Biological Sciences at Simon Fraser University. The data are simulated, but illustrative of the process.

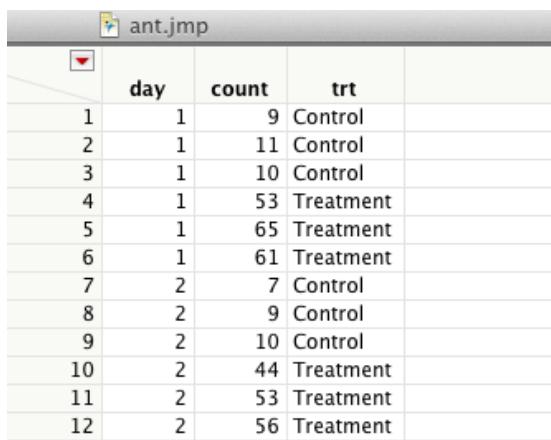
Considered one of the worst invasive pest ants, the electric ant, or little fire ant (*Wasmannia auropunctata*) has negatively impacted both biodiversity and agriculture. Its distribution is nearly pantropical,

and greenhouse infestations have been reported as far north as Canada and the United Kingdom. Current *W. auropunctata* detection methods commonly employ a food item like peanut butter. An alternative detection method may use pheromone attractants. For *W. auropunctata*, a one-way trap containing an alarm pheromone has been successfully used to detect little fire ant populations in macadamia nut orchards. What is the longevity of this type of pheromone lure as used in a unique one-way ant trap?

At the beginning of the experiment, 180 control traps and 180 treatment traps were prepared. On each day, three traps of each type were randomized to locations in the orchard where the ant species were known to be present. 24 hours later, the traps were retrieved and the number of ants captured counted, and the trap is discarded.

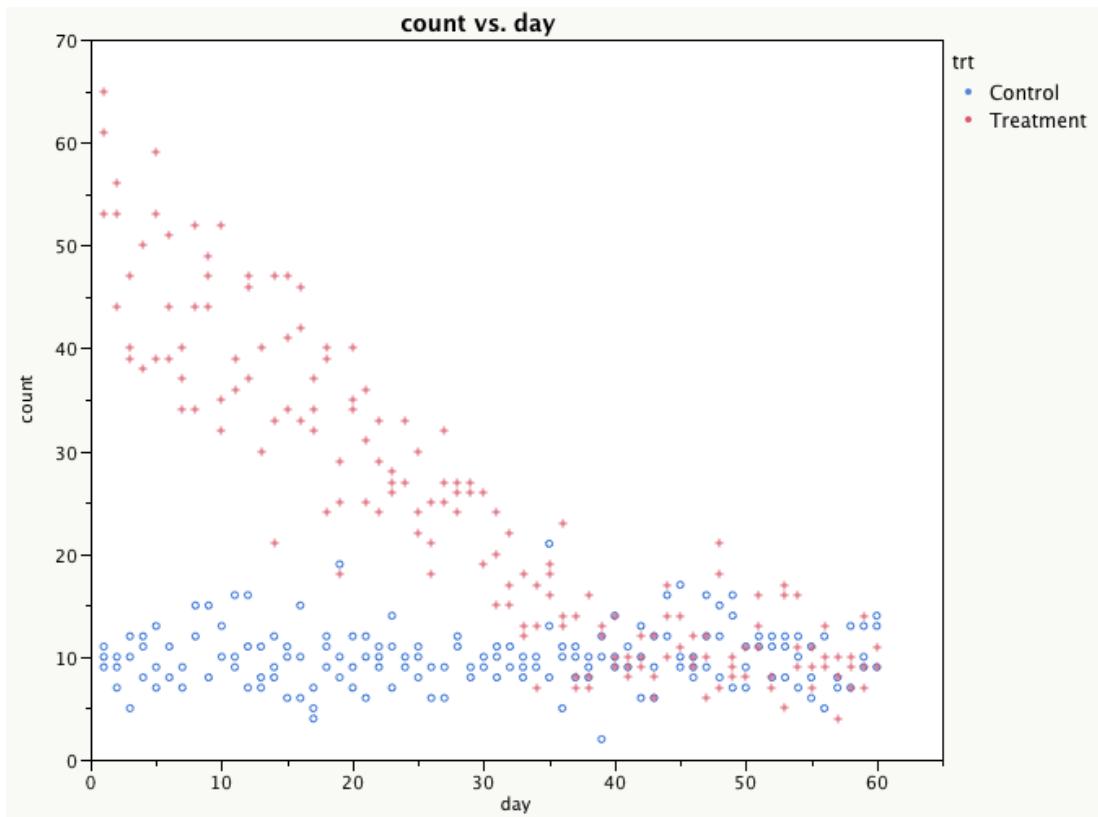
Because separate traps were prepared for each day and randomized each day, each observation can be treated as independent of other observations. This avoids the complications of repeated measures if the same lure is used for multiple days or the same locations used for the experiment – the analysis ideas are similar, but some care is needed to deal with potential correlation in the responses taken from the same trap at the same location over time.

The data is available in the *ants.csv* file in the Sample Program Library at <http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms>. The data are also available in a *JMP* datafile. Part of the raw data are shown below:



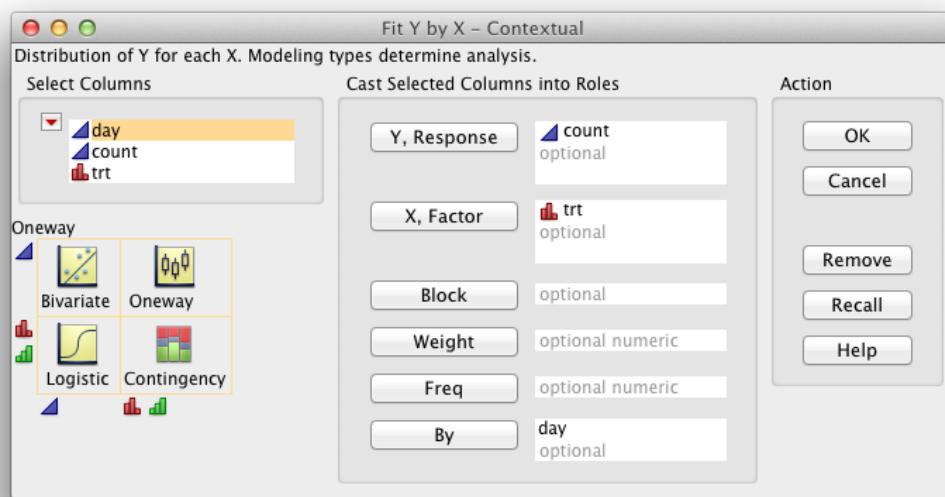
	day	count	trt
1	1	9	Control
2	1	11	Control
3	1	10	Control
4	1	53	Treatment
5	1	65	Treatment
6	1	61	Treatment
7	2	7	Control
8	2	9	Control
9	2	10	Control
10	2	44	Treatment
11	2	53	Treatment
12	2	56	Treatment

Start by plotting the data in the usual way using the *Analyze->Fit Y-by-X* platform after assigning markers to the row based on the *trt* variable:



It would appear from the plot, that the pheromone loses its effectiveness somewhere between day 30 and day 50, but the actual time point is difficult to see because of the noise in the data.

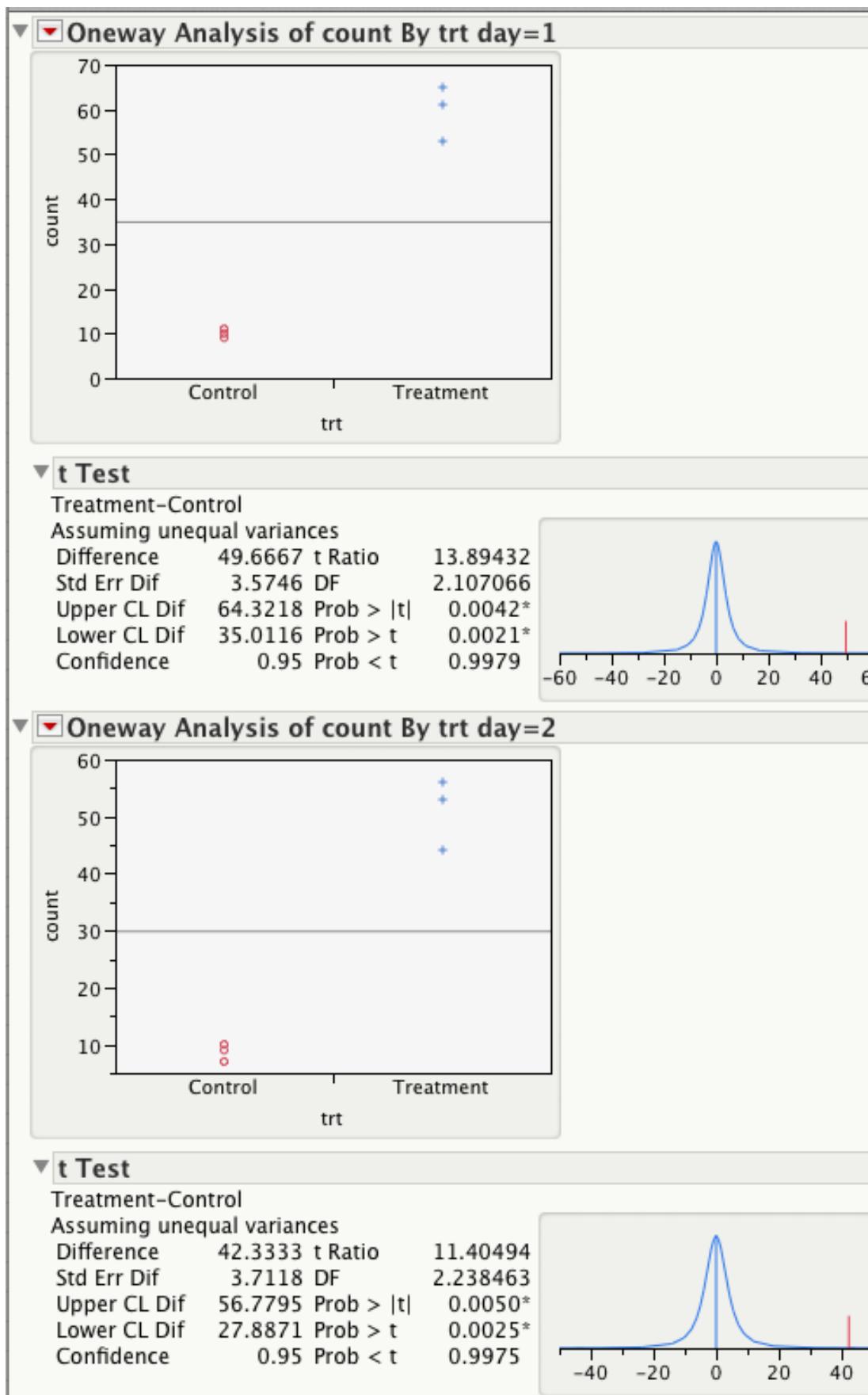
A naive analysis would do a t-test on each day and see when a statistically significant difference was detected. This takes a bit of work in *JMP*, but is relatively straightforward. First, use the *Analyze->Fit Y-by-X* platform to do a t-test for each day, comparing the mean count for the two treatments. Notice the use of day as a *By* variable.



This gives a separate plot by *Day*. To fit the t-test for each day, use the **Command** key, click on the red-triangle, and select t-test:

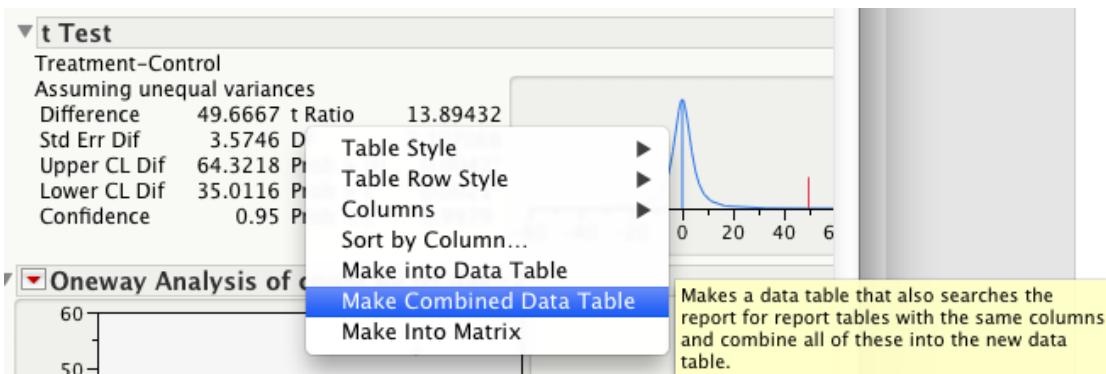


You now have a separate report for EACH day:



## CHAPTER 18. REGRESSION - HOCKEY STICKS, BROKEN STICKS, PIECEWISE, CHANGE POINTS

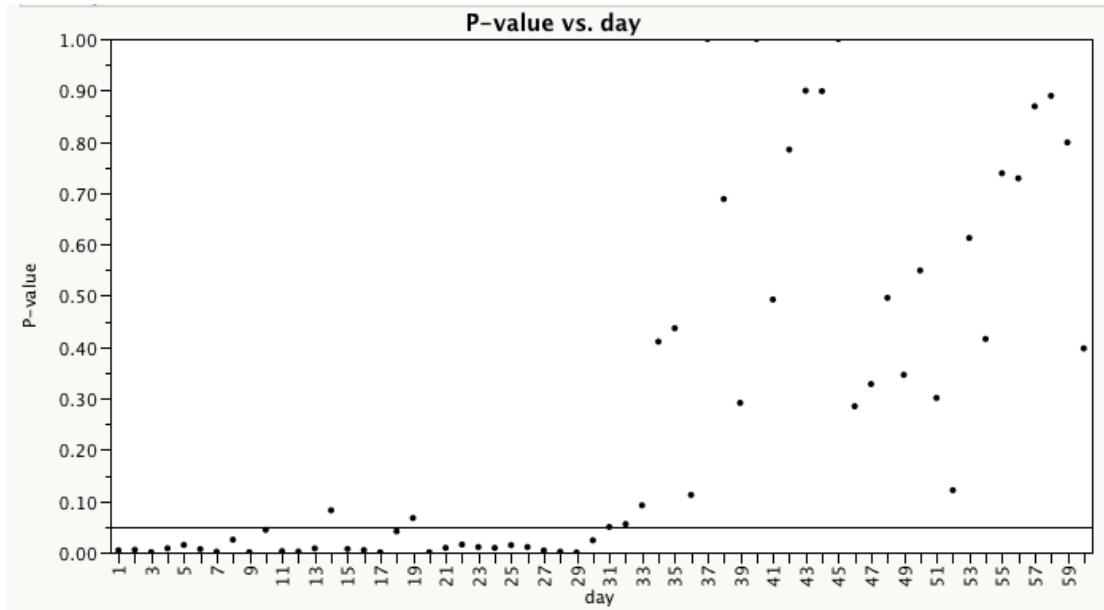
We want to extract the p-value for each day. Right click in ANY of the t-test reports and choose *Make Combined Data Table*:



This creates a MESSY data table:

	day	X	Y	Test Assumption	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
1	1	trt	count	Unequal Variances	Difference	49.6667	• t Ratio	13.894321643	•	•
2	1	trt	count	Unequal Variances	Std Err Diff	3.5746	• DF	2.10706604	•	•
3	1	trt	count	Unequal Variances	Upper CL Dif	64.3218	• Prob >  t	0.0042	•	•
4	1	trt	count	Unequal Variances	Lower CL Dif	35.0116	• Prob > t	0.0021	•	•
5	1	trt	count	Unequal Variances	Confidence	•	0.95 Prob < t	0.9979	•	•
6	2	trt	count	Unequal Variances	Difference	42.3333	• t Ratio	11.404936679	•	•
7	2	trt	count	Unequal Variances	Std Err Diff	3.7118	• DF	2.2384626583	•	•
8	2	trt	count	Unequal Variances	Upper CL Dif	56.7795	• Prob >  t	0.0050	•	•
9	2	trt	count	Unequal Variances	Lower CL Dif	27.8871	• Prob > t	0.0025	•	•
10	2	trt	count	Unequal Variances	Confidence	•	0.95 Prob < t	0.9975	•	•

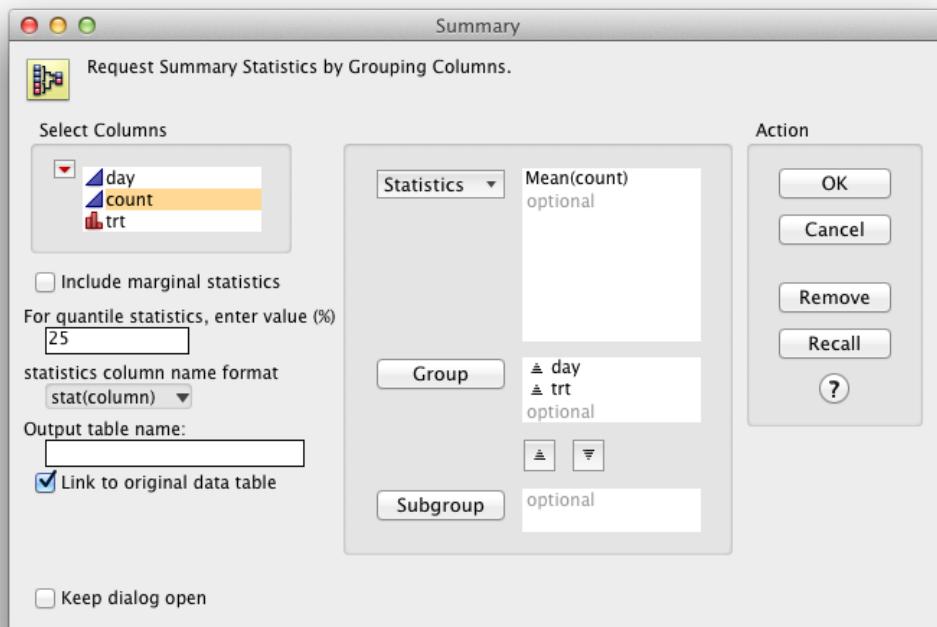
Finally, we select only the rows where *Column 1* contains the string “Upper CL Dif” and then plot *Column 6* vs. day, adjust the axes to get the plot:



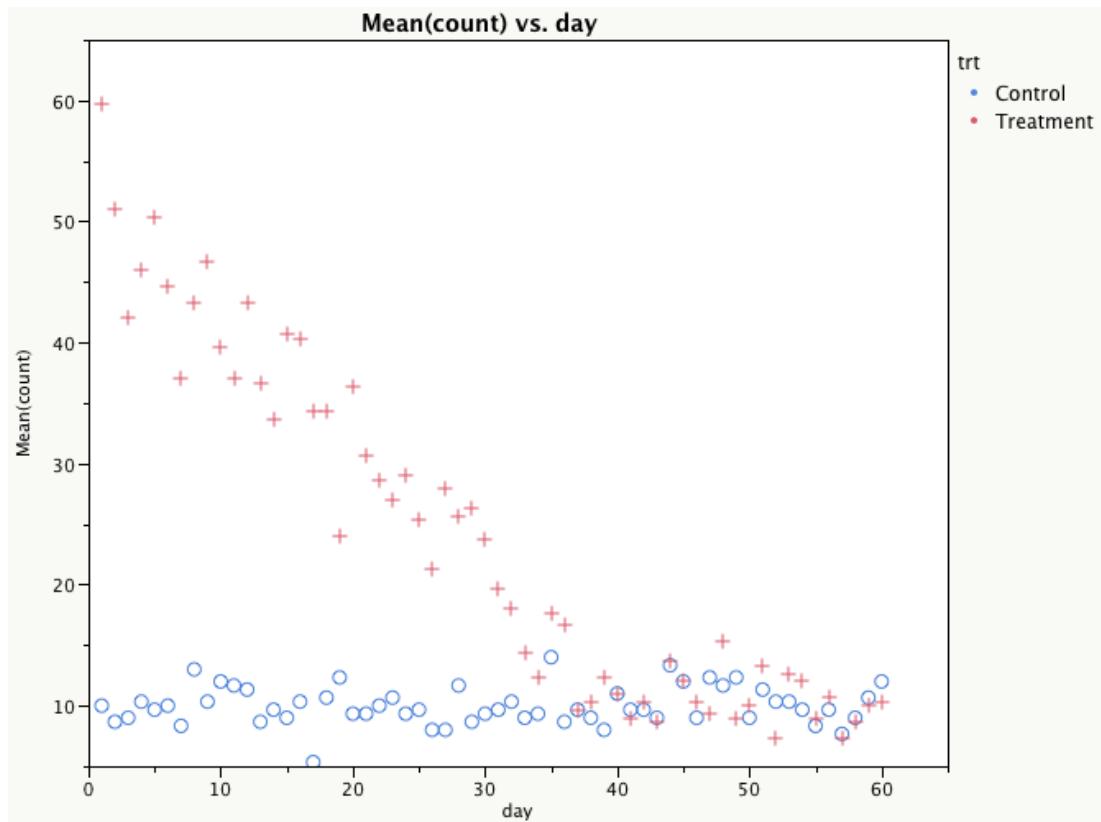
Based on this naive analysis, it would appear that the pheromone is effective only out to about day 30, but the previous graph shows that the change point should likely occur around day 40ish. Also note that this naive analysis failed to detect a difference in the mean count captured just after day 10 and just

before day 20, but the previous graph shows this is likely a false negative. This again illustrates the perils of the naive analysis.

The change-point method of analysis starts by finding the average for each day for treatment and control.

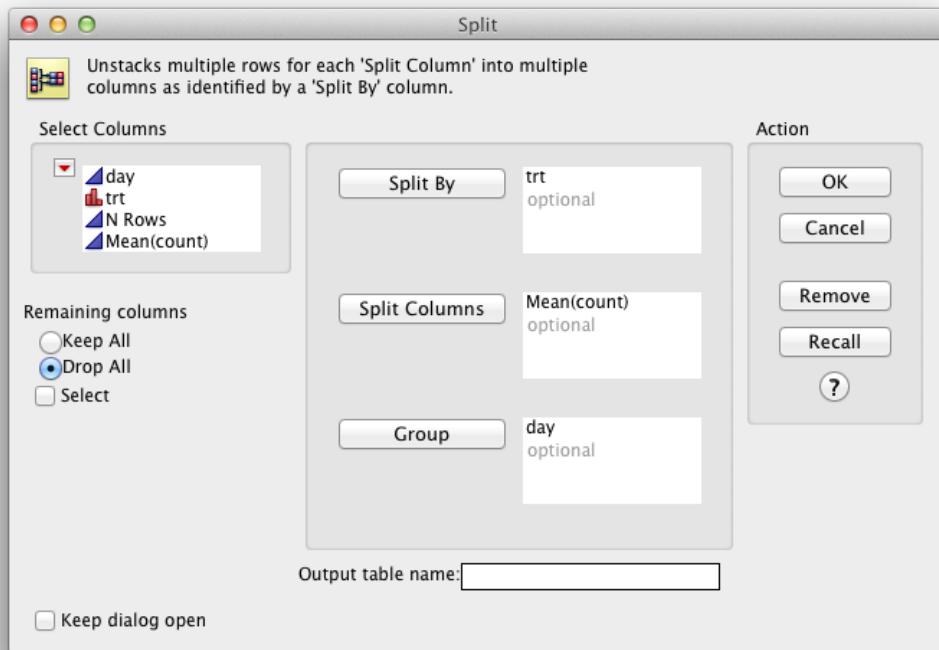


and then plotting the results to give:



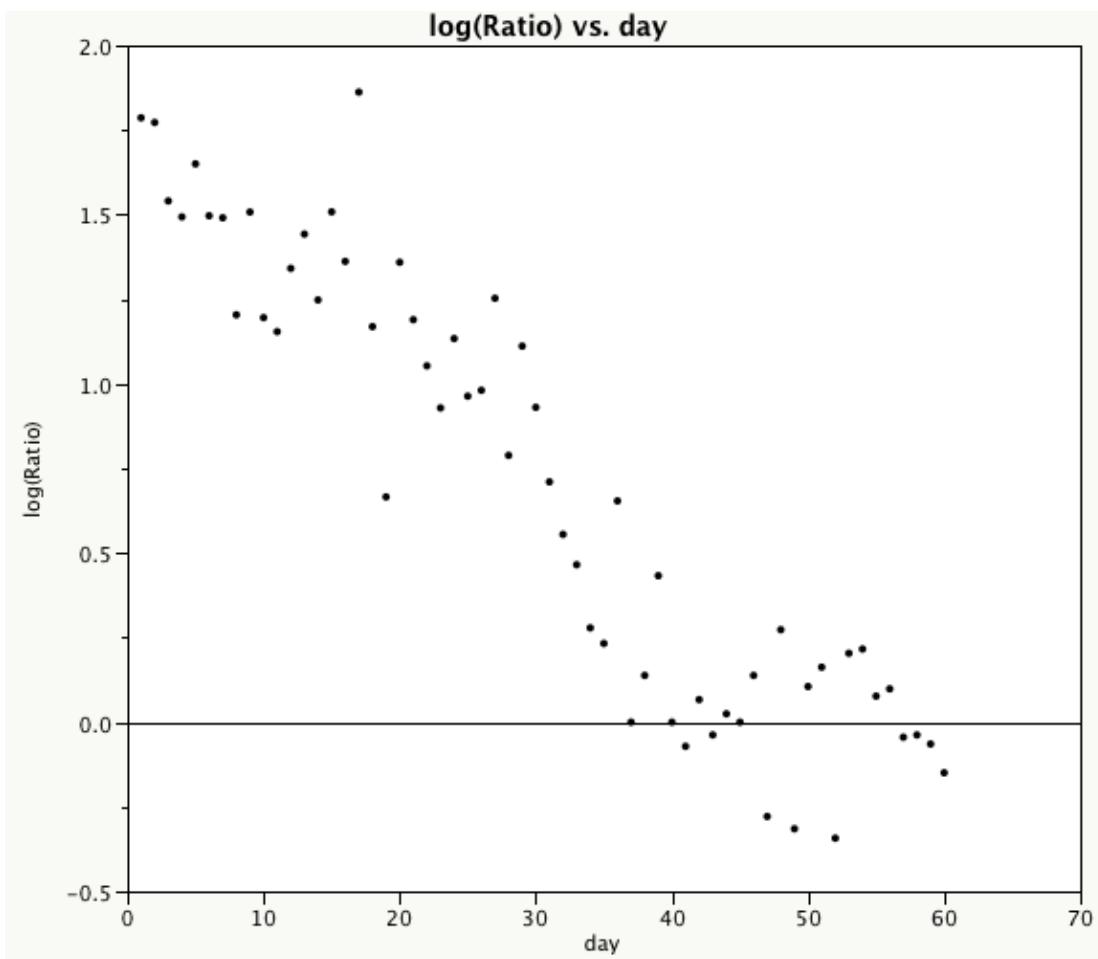
This plot again shows that the pheromone lasts to about day 40ish.

Finally, we find the log(ratio) of the treatment mean to the control mean and plot this ratio by day. We start in *JMP* by splitting the data and then creating a new formula variable:



	day	Control	Treatment	log(Ratio)
1	1	10	59.6666666667	1.78618842
2	2	8.6666666667	51	1.77234138
3	3	9	42	1.54044504
4	4	10.3333333333	46	1.49326648
5	5	9.6666666667	50.3333333333	1.64998401

and then plotting the  $\log(ratio)$ :



Note that a  $\log(ratio) = 0$  implies that the mean counts are the same. The change point appears to be around 40 days again.

The change point model is

$$\log(ratio)_i = \beta_0 + \beta_1(Day_i) + \epsilon_i$$

if  $Day_i$  is less than the change point (CP), and

$$\log(ratio)_i = \beta_0 + \beta_1(Day_i) + \beta_2(Day_i - CP)^+ + \epsilon_i$$

where  $(x)^+$  takes the value 0 if  $x < 0$  and  $x$  if  $x > 0$  as outlined earlier in Section 18.2.

We can fit a change-point model to the  $\log(ratio)$  using the *NonLinear* platform like we did in Section 18.2.1. Refer to the previous example on how to set up the *NonLinear* model.

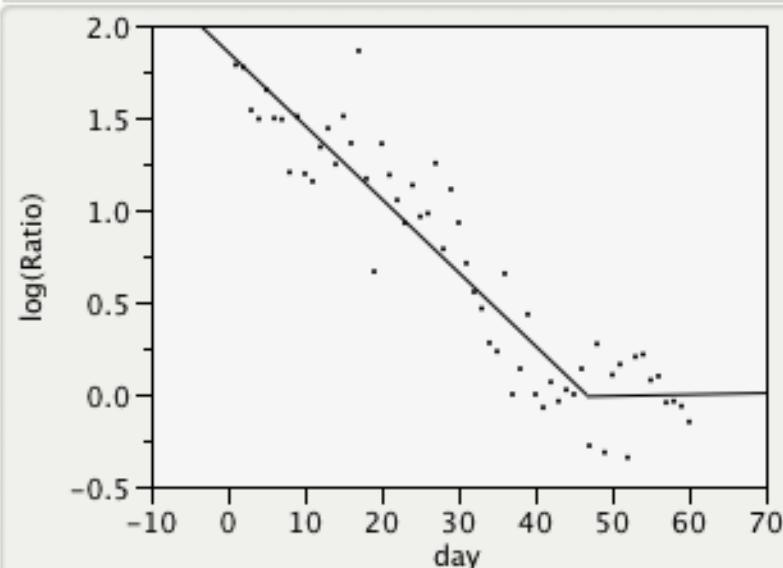
The final estimates are:

### Solution

	SSE	DFE	MSE	RMSE
	2.9267755943	56	0.0522638	0.2286129
Parameter	Estimate	ApproxStdErr	Lower CL	Upper CL
b0	1.8551947817	0.06852862	1.72534934	1.99917177
b1	-0.039936124	0.00253897	.	.
b2	0.0407329133	0.01536807	0.01460111	0.07360056
CP	46.761285714	3.36520996	38.9540849	52.1064838

Solved By: Analytic Gauss–Newton

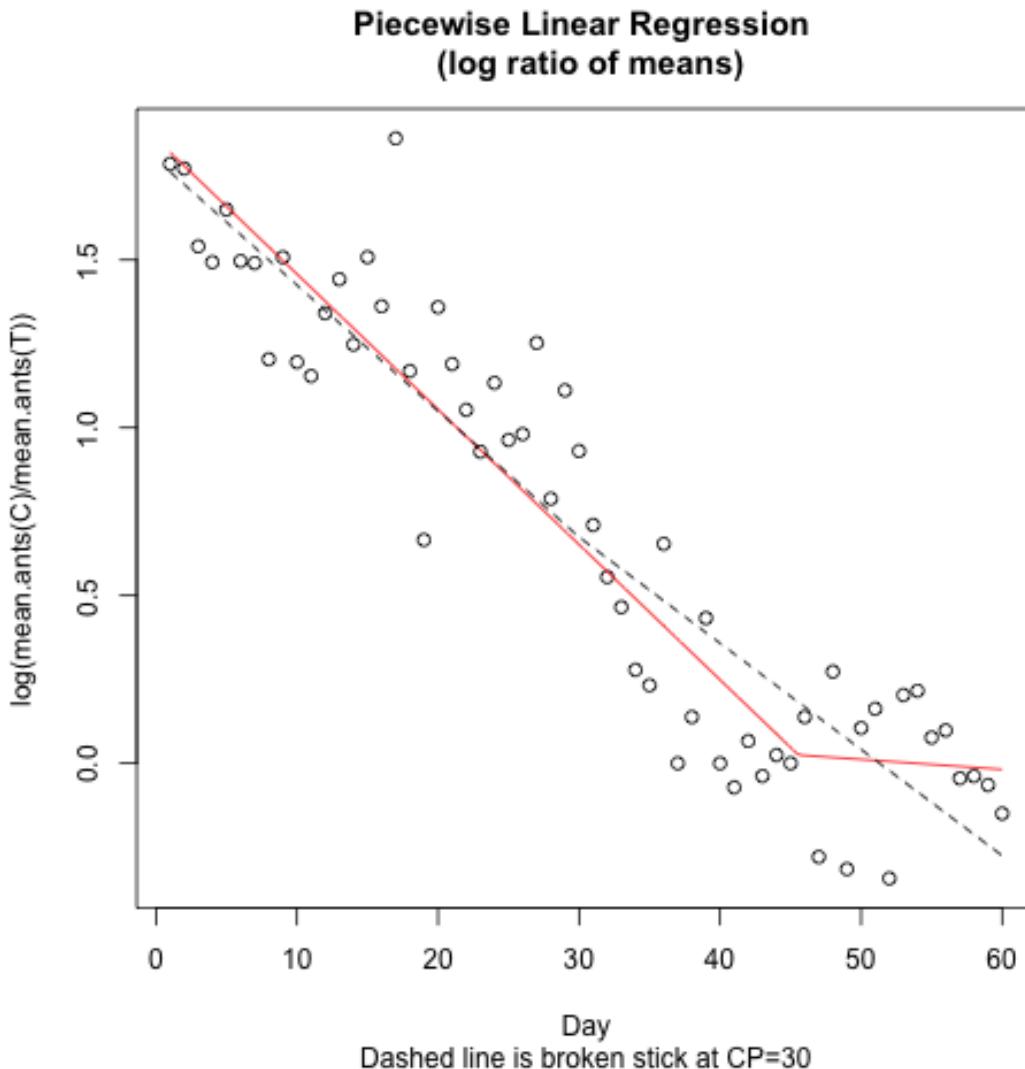
### Plot



The results are not entirely satisfactory<sup>2</sup> - the estimated change point is around 46 days which seems sensible, but the 95% confidence interval for the change point is very wide! Why was the change point estimated so poorly?

The problem is that there are no constraints on the slope after the change point. So a model where the change point is, say around 30, gives almost the same fit as the model where the change point is around 45:

<sup>2</sup>Note that SAS, R and JMP give slightly different answers because R uses maximum likelihood estimation, while SAS and JMP uses non-linear least squares. The programs also compute the confidence intervals differently – R uses a bootstrap approach while JMP and SAS use a delta-method approximation. The results are all asymptotically equivalent.



In these cases, what is needed is a broken-stick model where the slope AFTER the change point is forced to be zero. This model is surprisingly easy to fit:

$$Y_i = \beta_0 + \beta_1 \min(CP, Day_i)$$

where  $CP$  is the change point. Of course, the model must be fit with different values for the CP to find the best fit.

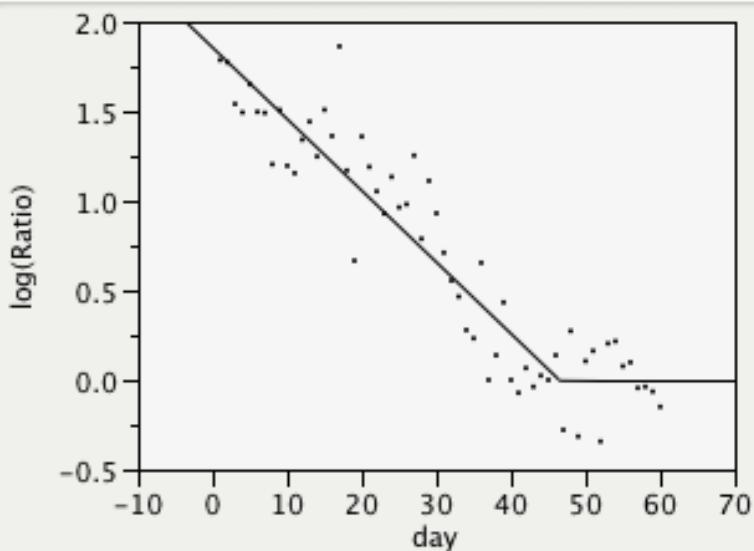
The non-linear least-squares procedure can be easily modified to force the slope to be 0 after the change point. Consult the *JMP* script for details. The final results is:

▼ Solution

	SSE	DFE	MSE	RMSE
	2.9269200279	57	0.0513495	0.2266042
Parameter	Estimate	ApproxStdErr	Lower CL	Upper CL
b0	1.8551947817	0.06792651	1.72661703	1.99727587
b1	-0.039936124	0.00251666	-0.0455432	.
CP	46.626837663	2.26354135	42.1114843	51.0141011

Solved By: Analytic Gauss–Newton

▼ Plot



The revised fit has an estimated change point around 47 days with a 95% confidence interval for the change point between 43 and 51 days – much tighter than the previous broken-stick change-point model.

The same basic methods are employed in the cases where the same trap and/or location is repeatedly used over time. As a general rule, you would want multiple traps of each type and try and randomize the locations over time as much as possible. It is not necessary to measure the traps every day. For example, at the start of the experiment, you could take a daily measurement every 5 days, and then switch to more intensive monitoring (i.e. daily) after about 30 days.

The broken-stick model is a simplification of reality (there likely isn't such a sharp change at the change point), but will serve as a close approximation to the underlying process.

## Lecture 13. Use and Interpretation of Dummy Variables

Stop worrying for 1 lecture and learn to appreciate the uses that “dummy variables” can be put to

Using dummy variables to measure average differences

Using dummy variables when more than 2 discrete categories

Using dummy variables for policy analysis

Using dummy variables to net out seasonality

## Use and Interpretation of Dummy Variables

Dummy variables – **where the variable takes only one of two values** – are useful tools in econometrics, since often interested in variables that are *qualitative* rather than *quantitative*

## Use and Interpretation of Dummy Variables

Dummy variables – **where the variable takes only one of two values** – are useful tools in econometrics, since often interested in variables that are *qualitative* rather than *quantitative*

In practice this means interested in variables that split the sample into two distinct groups in the following way

## Use and Interpretation of Dummy Variables

Dummy variables – **where the variable takes only one of two values** – are useful tools in econometrics, since often interested in variables that are *qualitative* rather than *quantitative*

In practice this means interested in variables that split the sample into two distinct groups in the following way

$D = 1$  if the criterion is satisfied

## Use and Interpretation of Dummy Variables

Dummy variables – **where the variable takes only one of two values** – are useful tools in econometrics, since often interested in variables that are *qualitative* rather than *quantitative*

In practice this means interested in variables that split the sample into two distinct groups in the following way

- $D = 1$  if the criterion is satisfied
- $D = 0$  if not

## Use and Interpretation of Dummy Variables

Dummy variables – **where the variable takes only one of two values** – are useful tools in econometrics, since often interested in variables that are *qualitative* rather than *quantitative*

In practice this means interested in variables that split the sample into two distinct groups in the following way

$D = 1$       if the criterion is satisfied  
 $D = 0$       if not

Eg. Male/Female

## Use and Interpretation of Dummy Variables

Dummy variables – **where the variable takes only one of two values** – are useful tools in econometrics, since often interested in variables that are *qualitative* rather than *quantitative*

In practice this means interested in variables that split the sample into two distinct groups in the following way

$D = 1$  if the criterion is satisfied  
 $D = 0$  if not

Eg. Male/Female  
so that the dummy variable “Male” would be coded  
1 if male

## Use and Interpretation of Dummy Variables

Dummy variables – **where the variable takes only one of two values** – are useful tools in econometrics, since often interested in variables that are *qualitative* rather than *quantitative*

In practice this means interested in variables that split the sample into two distinct groups in the following way

$D = 1$  if the criterion is satisfied  
 $D = 0$  if not

Eg. Male/Female  
so that the dummy variable “Male” would be coded  
1 if male  
and 0 if female

## Use and Interpretation of Dummy Variables

Dummy variables – **where the variable takes only one of two values** – are useful tools in econometrics, since often interested in variables that are *qualitative* rather than *quantitative*

In practice this means interested in variables that split the sample into two distinct groups in the following way

$D = 1$  if the criterion is satisfied  
 $D = 0$  if not

Eg. Male/Female  
so that the dummy variable "Male" would be coded  
1 if male  
and 0 if female

(though could equally create another variable "Female" coded 1 if female and 0 if male)

Example: Suppose we are interested in the gender pay gap

Example: Suppose we are interested in the gender pay gap

Model is  $\ln W = b_0 + b_1 \text{Age} + b_2 \text{Male}$

where Male = 1 or 0

Example: Suppose we are interested in the gender pay gap

Model is  $\ln W = b_0 + b_1 \text{Age} + b_2 \text{Male}$

where Male = 1 or 0

For men therefore the predicted wage

$$\hat{\ln W}_{men} = \hat{b}_0 + \hat{b}_1 \hat{\text{Age}} + \hat{b}_2 * (1)$$

Example: Suppose we are interested in the gender pay gap

Model is  $\ln W = b_0 + b_1 \text{Age} + b_2 \text{Male}$

where Male = 1 or 0

For men therefore the predicted wage

$$\begin{array}{ccccccc} & ^{\wedge} & & ^{\wedge} & & ^{\wedge} & & ^{\wedge} \\ \ln W_{men} & = & b_0 & + & b_1 & Age & + & b_2 * (1) \end{array}$$

$$\begin{array}{ccccccc} & ^{\wedge} & & ^{\wedge} & & ^{\wedge} & \\ & b_0 & + & b_1 & Age & + & b_2 \end{array}$$

Example: Suppose we are interested in the gender pay gap

Model is  $\ln W = b_0 + b_1 \text{Age} + b_2 \text{Male}$

where Male = 1 or 0

For men therefore the predicted wage

$$\begin{aligned} \hat{\ln W}_{men} &= b_0 + b_1 \text{Age} + b_2 * (1) \\ &= b_0 + b_1 \text{Age} + b_2 \end{aligned}$$

For women

$$\hat{\ln W}_{women} = b_0 + b_1 \text{Age} + b_2 * (0)$$

**Example:** Suppose we are interested in the gender pay gap

Model is  $\ln W = b_0 + b_1 \text{Age} + b_2 \text{Male}$

where Male = 1 or 0

For men therefore the predicted wage

$$\begin{aligned} \ln W_{men} &= b_0 + b_1 \text{Age} + b_2 * (1) \\ &= b_0 + b_1 \text{Age} + b_2 \end{aligned}$$

For women

$$\begin{aligned} \ln W_{women} &= b_0 + b_1 \text{Age} + b_2 * (0) \\ &= b_0 + b_1 \text{Age} \end{aligned}$$

Remember that OLS predicts the mean or average value of the dependent variable

$$\bar{\hat{Y}} = \bar{Y}$$

(see lecture 2)

So in the case of a regression model with log wages as the dependent variable,  $\ln W = b_0 + b_1 \text{Age} + b_2 \text{Male}$

the average of the fitted values equals the average of log wages

$$\bar{\hat{\ln}(W)} = \bar{\ln W}$$

Remember that OLS predicts the mean or average value of the dependent variable

$$\bar{\hat{Y}} = \bar{Y}$$

(see lecture 2)

Remember that OLS predicts the mean or average value of the dependent variable

$$\bar{\hat{Y}} = \bar{Y}$$

(see lecture 2)

So in the case of a regression model with log wages as the dependent variable,  $\ln W = b_0 + b_1 \text{Age} + b_2 \text{Male}$

Remember that OLS predicts the mean or average value of the dependent variable

$$\bar{\hat{Y}} = \bar{Y}$$

(see lecture 2)

So in the case of a regression model with log wages as the dependent variable,  $\ln W = b_0 + b_1 \text{Age} + b_2 \text{Male}$

the average of the fitted values equals the average of log wages

$$\bar{\hat{\ln}(W)} = \bar{\ln W}$$

So the (average) difference in pay between men and women is then

$$\ln W^{\text{men}} - \ln W^{\text{women}}$$

So the (average) difference in pay between men and women is then

$$\bar{\ln W^{\text{men}}} - \bar{\ln W^{\text{women}}} = \bar{\ln W_{\text{men}}} - \bar{\ln W_{\text{women}}}$$

So the (average) difference in pay between men and women is then

$$\begin{aligned} \bar{\wedge} & & \bar{\wedge} \\ \text{Ln}W^{\text{men}} - \text{Ln}W^{\text{women}} &= \text{Ln}W_{\text{men}} - \text{Ln}W_{\text{women}} \\ & \wedge \quad \wedge \quad \wedge \quad \wedge \quad \wedge \\ &= b_0 + b_1 \text{Age} + b_2 - b_0 + b_1 \text{Age} \end{aligned}$$

The (average) difference in pay between men and women is then

$$\begin{aligned} \bar{\ln W}_{men} - \bar{\ln W}_{women} &= \bar{\ln W}_{men} - \bar{\ln W}_{women} \\ &= \hat{b}_0 + \hat{b}_1 \hat{Age} + \hat{b}_2 - \hat{b}_0 + \hat{b}_1 \hat{Age} \\ &= \hat{b}_2 \end{aligned}$$

which is just the coefficient on the male dummy variable

The (average) difference in pay between men and women is then

$$\begin{aligned} \bar{\ln W^{\text{men}}} - \bar{\ln W^{\text{women}}} &= \bar{\ln W_{\text{men}}} - \bar{\ln W_{\text{women}}} \\ &= b_0 + b_1 \bar{\text{Age}} + b_2 - b_0 + b_1 \bar{\text{Age}} \\ &= b_2 \end{aligned}$$

which is just the coefficient on the male dummy variable

It also follows that the constant,  $b_0$ , measures the intercept of default group (women) with age set to zero and  $b_0 + b_2$  is the intercept for men

The (average) difference in pay between men and women is then

$$\begin{aligned} \bar{\ln W_{men}} - \bar{\ln W_{women}} &= \bar{\ln W_{men}} - \bar{\ln W_{women}} \\ &= b_0 + b_1 \bar{Age} + b_2 - b_0 + b_1 \bar{Age} + b_2 \\ &= b_2 \end{aligned}$$

which is just the coefficient on the male dummy variable

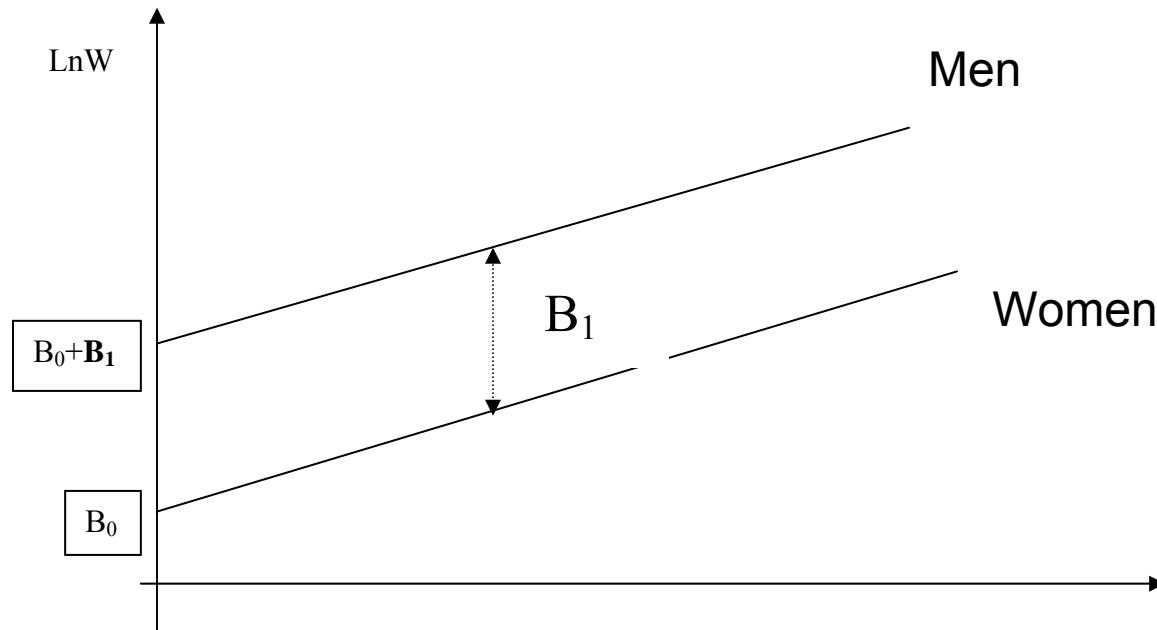
So the coefficients on dummy variables measure the average difference between the group coded with the value "1" and the group coded with the value "0" (the "default" or "base group" )

It also follows that the constant,  $b_0$ , now measures the notional value of the dependent variable (in this case log wages) of the default group (in this case women) with age set to zero

and  $b_0 + b_2$  is the intercept and notional value of log wages at age zero for men

So to measure ***average*** difference between two groups

$$\ln W = \beta_0 + \beta_1 \text{Group Dummy}$$



A simple regression of the log of hourly wages on age using the data set ps4data.dta gives

. reg lh wage age	Source	SS	df	MS	Number of obs =	12098
	Model	75.4334757	1	75.4334757	F( 1, 12096) =	235.55
	Residual	3873.61564	12096	.320239388	Prob > F =	0.0000
	Total	3949.04911	12097	.326448633	R-squared =	0.0191
					Adj R-squared =	0.0190
					Root MSE =	.5659
lh wage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.0070548	.0004597	15.348	0.000	.0061538	.0079558
_cons	1.693719	.0186945	90.600	0.000	1.657075	1.730364

Now introduce a male dummy variable (1= male, 0 otherwise) as an **intercept dummy**. This specification says the slope effect (of age) is the same for men and women, but that the intercept (or the **average difference** in pay between men and women) is different

. reg lhw age male	Source	SS	df	MS	Number of obs =	12098
	Model	264.053053	2	132.026526	F( 2, 12095) =	433.34
	Residual	3684.99606	12095	.304671026	Prob > F =	0.0000
	Total	3949.04911	12097	.326448633	R-squared =	0.0669
					Adj R-squared =	0.0667
					Root MSE =	.55197
lhw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.0066816	.0004486	14.89	0.000	.0058022	.0075609
male	.2498691	.0100423	24.88	0.000	.2301846	.2695537
_cons	1.583852	.0187615	84.42	0.000	1.547077	1.620628

Hence

$$\begin{aligned} \text{average wage difference between men and women} \\ =(b_0 - (b_0 + b_2)) = b_2 = 25\% \text{ more on average} \end{aligned}$$

Note that if we define a dummy variables as female (1= female, 0 otherwise) then

. reg lh wage age female				Number of obs = 12098		
Source	SS	df	MS	F( 2, 12095) = 433.34		
Model	264.053053	2	132.026526	Prob > F	=	0.0000
Residual	3684.99606	12095	.304671026	R-squared	=	0.0669
Total	3949.04911	12097	.326448633	Adj R-squared	=	0.0667
				Root MSE	=	.55197
lh wage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.0066816	.0004486	14.894	0.000	.0058022	.0075609
female	-.2498691	.0100423	-24.882	0.000	-.2695537	-.2301846
_cons	1.833721	.0190829	96.093	0.000	1.796316	1.871127

The coefficient estimate on the dummy variable is the same but the sign of the effect is reversed (now negative). This is because the reference (default) category in this regression is now men

Model is now  $\text{LnW} = b_0 + b_1\text{Age} + b_2\text{female}$

so constant,  $b_0$ , measures average earnings of default group (men)  
and  $b_0 + b_2$  is average earnings of women

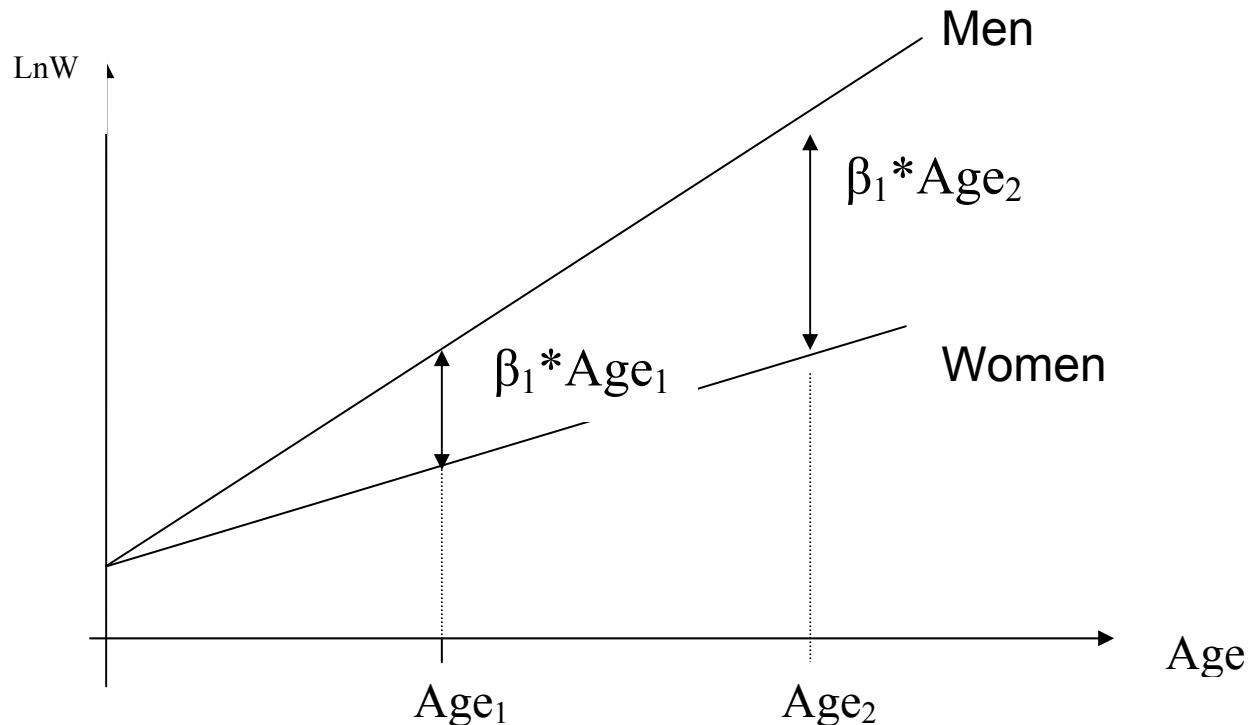
So now

$$\begin{aligned} \text{average wage difference between men and women} \\ = (b_0 - (b_0 + b_2)) = b_2 = -25\% \text{ less on average} \end{aligned}$$

Hence it does not matter which way the dummy variable is defined as long as you are clear as to the appropriate reference category.

2) To measure ***Difference in Slope Effects*** between two groups

$$\ln W = \beta_0 + \beta_1 \text{Group Dummy} * \text{Slope Variable}$$



(Dummy Variable Interaction Term)

We need to consider an **interaction term** – multiply slope variable (age) by dummy variable.

Now consider an **interaction term** – multiply slope variable (age) by dummy variable.

Model is now  $\text{LnW} = b_0 + b_1\text{Age} + b_2\text{Female} * \text{Age}$

Now consider an **interaction term** – multiply slope variable (age) by dummy variable.

Model is now  $\text{LnW} = b_0 + b_1\text{Age} + b_2\text{Female} * \text{Age}$

This means that (slope) age effect is different for the 2 groups

Now consider an **interaction term** – multiply slope variable (age) by dummy variable.

Model is now  $\text{LnW} = b_0 + b_1\text{Age} + b_2\text{Female} * \text{Age}$

This means that (slope) age effect is different for the 2 groups

$$\frac{d\text{LnW}}{d\text{Age}} = b_1 \quad \text{if female} = 0$$

Now consider an **interaction term** – multiply slope variable (age) by dummy variable.

Model is now  $\text{LnW} = b_0 + b_1\text{Age} + b_2\text{Female} * \text{Age}$

This means that (slope) age effect is different for the 2 groups

$$\begin{aligned}\frac{d\text{LnW}}{d\text{Age}} &= b_1 && \text{if female} = 0 \\ &= b_1 + b_2 && \text{if female} = 1\end{aligned}$$

```

. g femage=female*age          /* command to create interaction term */

. reg lh wage age femage
Source |       SS           df           MS
-----+-----
Model | 283.289249      2  141.644625
Residual | 3665.75986 12095   .3030806
-----+-----
Total | 3949.04911 12097   .326448633

Number of obs = 12098
F( 2, 12095) = 467.35
Prob > F = 0.0000
R-squared = 0.0717
Adj R-squared = 0.0716
Root MSE = .55053

-----+
lh wage |     Coef.    Std. Err.      t    P>|t|    [95% Conf. Interval]
-----+
age |   .0096943   .0004584    21.148  0.000    .0087958   .0105929
femage |  -.006454   .0002465   -26.188  0.000   -.0069371  -.005971
_cons |   1.715961   .0182066    94.249  0.000    1.680273  1.751649

```

So effect of 1 extra year of age on earnings

$$\begin{aligned}
&= .0097 \text{ if male} \\
&= (.0097 - .0065) \text{ if female}
\end{aligned}$$

Can include both an intercept and a slope dummy variable in the same regression to decide whether differences were caused by differences in intercepts or the slope variables

```

. reg lh wage age female femage
Source |       SS           df           MS
-----+-----
Model | 283.506857      3  94.5022855
Residual | 3665.54226 12094   .303087668
-----+-----
Total | 3949.04911 12097   .326448633

Number of obs = 12098
F( 3, 12094) = 311.80
Prob > F = 0.0000
R-squared = 0.0718
Adj R-squared = 0.0716
Root MSE = .55053

-----+
lh wage |     Coef.    Std. Err.      t    P>|t|    [95% Conf. Interval]
-----+
age |   .0100393   .0006131    16.376  0.000    .0088376   .011241
female |   .0308822   .0364465     0.847  0.397   -.0405588   .1023233
femage |  -.0071846   .0008968   -8.012  0.000   -.0089425  -.0054268
_cons |   1.701176   .0252186    67.457  0.000    1.651743  1.750608

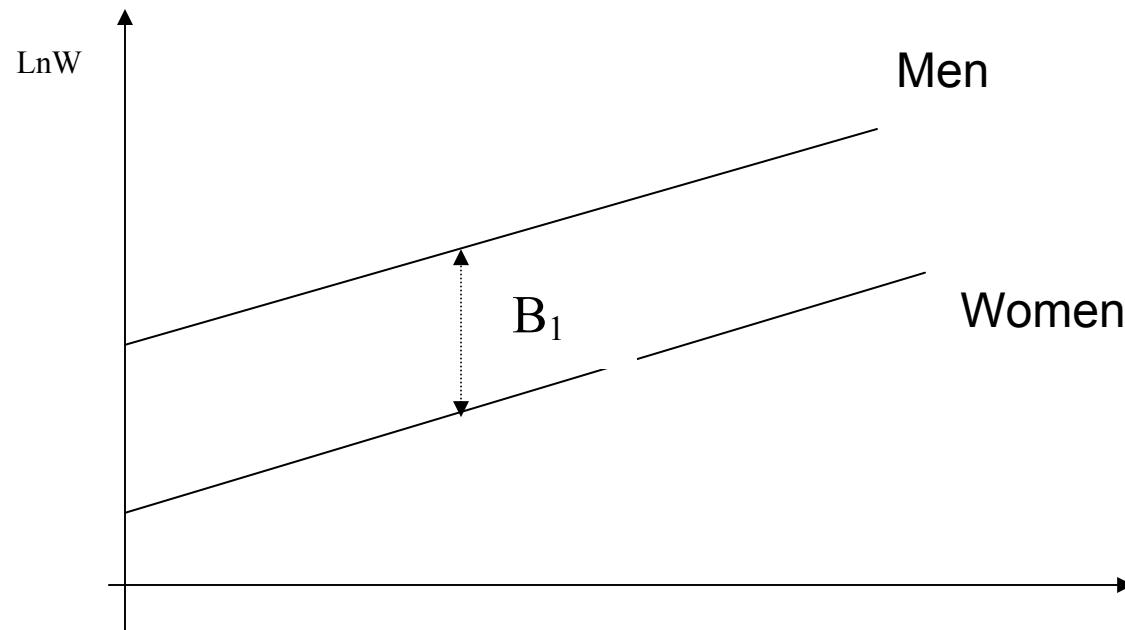
```

In this example the average differences in pay between men and women appear to be driven by factors which cause the slopes to differ (ie the rewards to extra years of experience are much lower for women than men)- Note that this model is equivalent to running separate regressions for men and women – since allowing both intercept and slope to vary

## Using & Understanding Dummy Variables

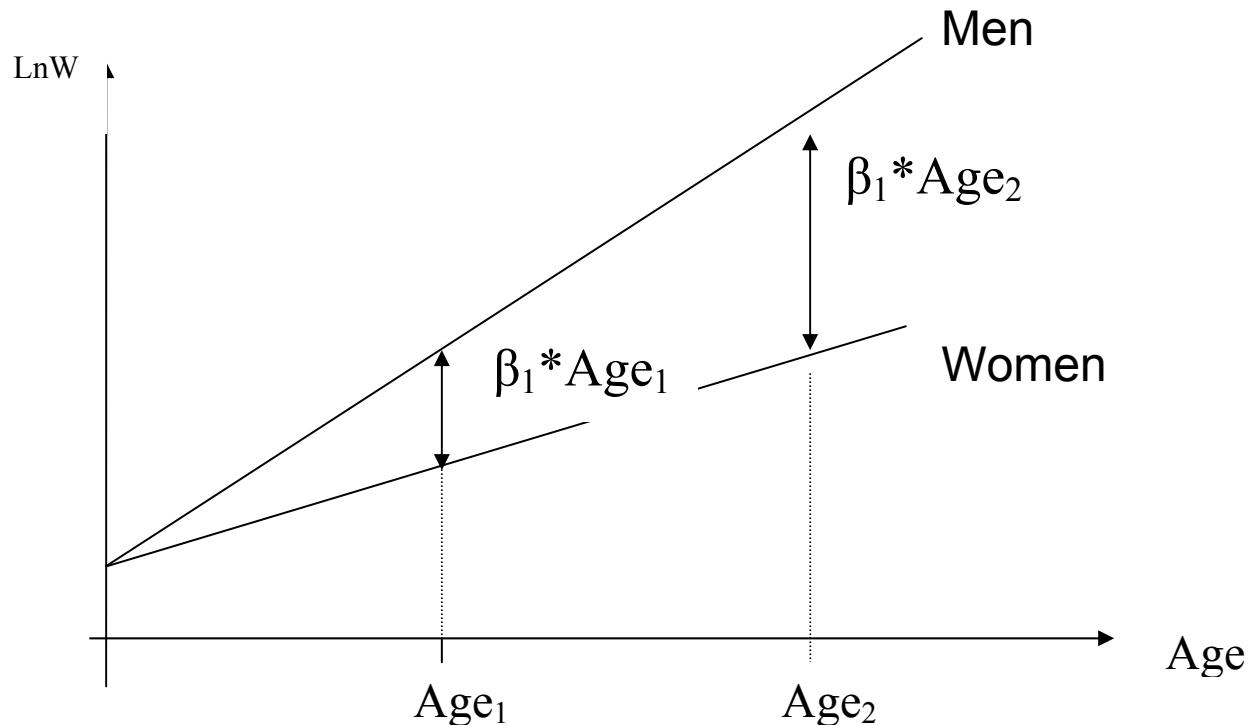
To measure **average** difference between two groups

$$\ln W = \beta_0 + \beta_1 \text{Group Dummy}$$



2) To measure ***Difference in Slope Effects*** between two groups

$$\ln W = \beta_0 + \beta_1 \text{Group Dummy} * \text{Slope Variable}$$



(Dummy Variable Interaction Term)

## **The Dummy Variable Trap**

The general principle of dummy variables can be extended to cases where there are several (but not infinite) discrete groups/categories

## **The Dummy Variable Trap**

The general principle of dummy variables can be extended to cases where there are several (but not infinite) discrete groups/categories

In general just define a dummy variable for each category

## **The Dummy Variable Trap**

The general principle of dummy variables can be extended to cases where there are several (but not infinite) discrete groups/categories

In general just define a dummy variable for each category

Eg if no. groups = 3 (North, Midlands, South)

## The Dummy Variable Trap

The general principle of dummy variables can be extended to cases where there are several (but not infinite) discrete groups/categories

In general just define a dummy variable for each category

Eg if no. groups = 3 (North, Midlands, South)

Then define

$D_{\text{North}}$  = 1 if live in the North, 0 otherwise

## The Dummy Variable Trap

The general principle of dummy variables can be extended to cases where there are several (but not infinite) discrete groups/categories

In general just define a dummy variable for each category

Eg if no. groups = 3 (North, Midlands, South)

Then define

$D_{\text{North}}$       = 1 if live in the North, 0 otherwise

$D_{\text{Midlands}}$       = 1 if live in the Midlands, 0 otherwise

## The Dummy Variable Trap

The general principle of dummy variables can be extended to cases where there are several (but not infinite) discrete groups/categories

In general just define a dummy variable for each category

Eg if no. groups = 3 (North, Midlands, South)

Then define

- $D_{\text{North}}$       = 1 if live in the North, 0 otherwise
- $D_{\text{Midlands}}$     = 1 if live in the Midlands, 0 otherwise
- $D_{\text{South}}$        = 1 if live in the South, 0 otherwise

## The Dummy Variable Trap

The general principle of dummy variables can be extended to cases where there are several (but not infinite) discrete groups/categories

In general just define a dummy variable for each category

Eg if no. groups = 3 (North, Midlands, South)

Then define

- $D_{\text{North}}$       = 1 if live in the North, 0 otherwise
- $D_{\text{Midlands}}$     = 1 if live in the Midlands, 0 otherwise
- $D_{\text{South}}$        = 1 if live in the South, 0 otherwise

However

## The Dummy Variable Trap

The general principle of dummy variables can be extended to cases where there are several (but not infinite) discrete groups/categories

In general just define a dummy variable for each category

Eg if no. groups = 3 (North, Midlands, South)

Then define

- |                       |  |
|-----------------------|--|
| $D_{\text{North}}$    | = 1 if live in the North, 0 otherwise    |
| $D_{\text{Midlands}}$ | = 1 if live in the Midlands, 0 otherwise |
| $D_{\text{South}}$    | = 1 if live in the South, 0 otherwise    |

However

As a rule should always include **one less** dummy variable in the model than there are categories, otherwise will introduce multicollinearity into the model

## Example of Dummy Variable Trap

Suppose interested in estimating the effect of (5) different qualifications on pay

A regression of the log of hourly earnings on dummy variables for each of 5 education categories gives the following output

. reg lh wage age postgrad grad highint low none						
Source	SS	df	MS		Number of obs	= 12098
Model	932.600688	5	186.520138		F( 5, 12092)	= 747.70
Residual	3016.44842	12092	.249458189		Prob > F	= 0.0000
Total	3949.04911	12097	.326448633		R-squared	= 0.2362
					Adj R-squared	= 0.2358
					Root MSE	= .49946
-----						
lh wage	Coef.	Std. Err.	t	P> t	[ 95% Conf. Interval]	
age	.010341	.0004148	24.931	0.000	.009528	.0111541
postgrad	(dropped)					
grad	-.0924185	.0237212	-3.896	0.000	-.1389159	-.045921
highint	-.4011569	.0225955	-17.754	0.000	-.4454478	-.356866
low	-.6723372	.0209313	-32.121	0.000	-.7133659	-.6313086
none	-.9497773	.0242098	-39.231	0.000	-.9972324	-.9023222
_cons	2.110261	.0259174	81.422	0.000	2.059459	2.161064

Since in this example there are 5 possible education categories  
(postgrad, graduate, higher intermediate, low and no qualifications)

Since in this example there are 5 possible education categories (postgrad, graduate, higher intermediate, low and no qualifications)

5 dummy variables exhaust the set of possible categories, so the sum of these 5 dummy variables is always one for each observation in the data set.

Since in this example there are 5 possible education categories (postgrad, graduate, higher intermediate, low and no qualifications)

5 dummy variables exhaust the set of possible categories, so the sum of these 5 dummy variables is always one for each observation in the data set.

Obs.	Constant	postgrad	grad	higher	low	noquals	Sum
------	----------	----------	------	--------	-----	---------	-----

Since in this example there are 5 possible education categories (postgrad, graduate, higher intermediate, low and no qualifications)

5 dummy variables exhaust the set of possible categories, so the sum of these 5 dummy variables is always one for each observation in the regression data set.

Obs.	Constant	postgrad	grad	higher	low	noquals	Sum
1	1	1	0	0	0	0	2

Since in this example there are 5 possible education categories (postgrad, graduate, higher intermediate, low and no qualifications)

5 dummy variables exhaust the set of possible categories, so the sum of these 5 dummy variables is always one for each observation in the regression data set.

Obs.	Constant	postgrad	grad	higher	low	noquals	Sum
1	1	1	0	0	0	0	1

Since in this example there are 5 possible education categories (postgrad, graduate, higher intermediate, low and no qualifications)

5 dummy variables exhaust the set of possible categories, so the sum of these 5 dummy variables is always one for each observation in the regression data set.

Obs.	Constant	postgrad	grad	higher	low	noquals	Sum
1	1	1	0	0	0	0	1
2	1	0	1	0	0	0	

Since in this example there are 5 possible education categories (postgrad, graduate, higher intermediate, low and no qualifications)

5 dummy variables exhaust the set of possible categories, so the sum of these 5 dummy variables is always one for each observation in the regression data set.

Obs.	Constant	postgrad	grad	higher	low	noquals	Sum
1	1	1	0	0	0	0	1
2	1	0	1	0	0	0	1

Since in this example there are 5 possible education categories (postgrad, graduate, higher intermediate, low and no qualifications)

5 dummy variables exhaust the set of possible categories, so the sum of these 5 dummy variables is always one for each observation in the regression data set.

Obs.	Constant	postgrad	grad	higher	low	noquals	Sum
1	1	1	0	0	0	0	1
2	1	0	1	0	0	0	1
3	1	0	0	0	0	1	1

Since in this example there are 5 possible education categories (postgrad, graduate, higher intermediate, low and no qualifications)

5 dummy variables exhaust the set of possible categories, so the sum of these 5 dummy variables is always one for each observation in the regression data set.

Obs.	Constant	postgrad	grad	higher	low	noquals	Sum
1	1	1	0	0	0	0	1
2	1	0	1	0	0	0	1
3	1	0	0	0	0	1	1

Given the presence of a constant using 5 dummy variables leads to pure multicollinearity, because the sum=1 which is the same as the value of the constant)

Since in this example there are 5 possible education categories (postgrad, graduate, higher intermediate, low and no qualifications)

5 dummy variables exhaust the set of possible categories, so the sum of these 5 dummy variables is always one for each observation in the regression data set.

Obs.	Constant	postgrad	grad	higher	low	noquals	Sum
1	1	1	0	0	0	0	1
2	1	0	1	0	0	0	1
3	1	0	0	0	0	1	1

Given the presence of a constant using 5 dummy variables leads to pure multicollinearity, (the sum=1 = value of the constant)

Since in this example there are 5 possible education categories (postgrad, graduate, higher intermediate, low and no qualifications)

5 dummy variables exhaust the set of possible categories, so the sum of these 5 dummy variables is always one for each observation in the data set.

Obs.	Constant	postgrad	grad	higher	low	noquals	Sum
1	1	1	0	0	0	0	1
2	1	0	1	0	0	0	1
3	1	0	0	0	0	1	1

Given the presence of a constant using 5 dummy variables leads to pure multicollinearity, (the sum=1 = value of the constant)

Since in this example there are 5 possible education categories (postgrad, graduate, higher intermediate, low and no qualifications)

5 dummy variables exhaust the set of possible categories, so the sum of these 5 dummy variables is always one for each observation in the data set.

Obs.	constant	postgrad	grad	higher	low	noquals	Sum
1	1	1	0	0	0	0	1
2	1	0	1	0	0	0	1
3	1	0	0	0	0	1	1

Given the presence of a constant using 5 dummy variables leads to pure multicollinearity, (the sum=1 = value of the constant)

So can't include all 5 dummies and the constant in the same model

Solution: drop one of the dummy variables. Then sum will no longer equal one for **every** observation in the data set.

Solution: drop one of the dummy variables. Then sum will no longer equal one for **every** observation in the data set.

Suppose drop the no quals dummy in the example above, the model is then

Obs.	Constant	postgrad	grad	higher	low
1	1	1	0	0	0
2	1	0	1	0	0
3	1	0	0	0	0

Solution: drop one of the dummy variables. Then sum will no longer equal one for **every** observation in the data set.

Suppose drop the no quals dummy in the example above, the model is then

Obs.	Constant	postgrad	grad	higher	low	Sum of dummies
1	1	1	0	0	0	
2	1	0	1	0	0	
3	1	0	0	0	0	

Solution: drop one of the dummy variables. Then sum will no longer equal one for **every** observation in the data set.

Suppose drop the no quals dummy in the example above, the model is then

Obs.	Constant	postgrad	grad	higher	low	Sum of dummies
1	1	1	0	0	0	1
2	1	0	1	0	0	
3	1	0	0	0	0	

Solution: drop one of the dummy variables. Then sum will no longer equal one for **every** observation in the data set.

Suppose drop the no quals dummy in the example above, the model is then

Obs.	Constant	postgrad	grad	higher	low	Sum of dummies
1	1	1	0	0	0	1
2	1	0	1	0	0	1
3	1	0	0	0	0	

Solution: drop one of the dummy variables. Then sum will no longer equal one for **every** observation in the data set.

Suppose drop the no quals dummy in the example above, the model is then

Obs.	Constant	postgrad	grad	higher	low	Sum of dummies
1	1	1	0	0	0	1
2	1	0	1	0	0	1
3	1	0	0	0	0	0

Solution: drop one of the dummy variables. Then sum will no longer equal one for **every** observation in the data set.

Suppose drop the no quals dummy in the example above, the model is then

Obs.	Constant	postgrad	grad	higher	low	Sum of dummies
1	1	1	0	0	0	1
2	1	0	1	0	0	1
3	1	0	0	0	0	0

and so the sum is no longer collinear with the constant

Solution: drop one of the dummy variables. Then sum will no longer equal one for **every** observation in the data set.

Suppose drop the no quals dummy in the example above, the model is then

Obs.	Constant	postgrad	grad	higher	low	Sum of dummies
1	1	1	0	0	0	1
2	1	0	1	0	0	1
3	1	0	0	0	0	0

and so the sum is no longer collinear with the constant

Doesn't matter which one you drop, though convention says drop the dummy variable corresponding to the most common category.

Solution: drop one of the dummy variables. Then sum will no longer equal one for **every** observation in the data set.

Suppose drop the no quals dummy in the example above, the model is then

Obs.	Constant	postgrad	grad	higher	low	Sum of dummies
1	1	1	0	0	0	1
2	1	0	1	0	0	1
3	1	0	0	0	0	0

and so the sum is no longer collinear with the constant

Doesn't matter which one you drop, though convention says drop the dummy variable corresponding to the most common category.

However changing the “default” category does change the coefficients, since all dummy variables are measured relative to this default reference category

Example: Dropping the postgraduate dummy (which Stata did automatically before when faced with the dummy variable trap) just replicates the above results. All the education dummy variables pay effects are measured relative to the missing postgraduate dummy variable (which effectively is now picked up by the constant term)

. reg lhw age grad highint low none						
Source	SS	df	MS	Number of obs = 12098		
Model	932.600688	5	186.520138	F( 5, 12092) = 747.70		
Residual	3016.44842	12092	.249458189	Prob > F = 0.0000		
Total	3949.04911	12097	.326448633	R-squared = 0.2362		
				Adj R-squared = 0.2358		
				Root MSE = .49946		
lhw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.010341	.0004148	24.93	0.000	.009528	.0111541
grad	-.0924185	.0237212	-3.90	0.000	-.1389159	-.045921
highint	-.4011569	.0225955	-17.75	0.000	-.4454478	-.356866
low	-.6723372	.0209313	-32.12	0.000	-.7133659	-.6313086
none	-.9497773	.0242098	-39.23	0.000	-.9972324	-.9023222
_cons	2.110261	.0259174	81.42	0.000	2.059459	2.161064

coefficients on education dummies are all negative since all categories earn less than the default group of postgraduates

Changing the default category to the no qualifications group gives

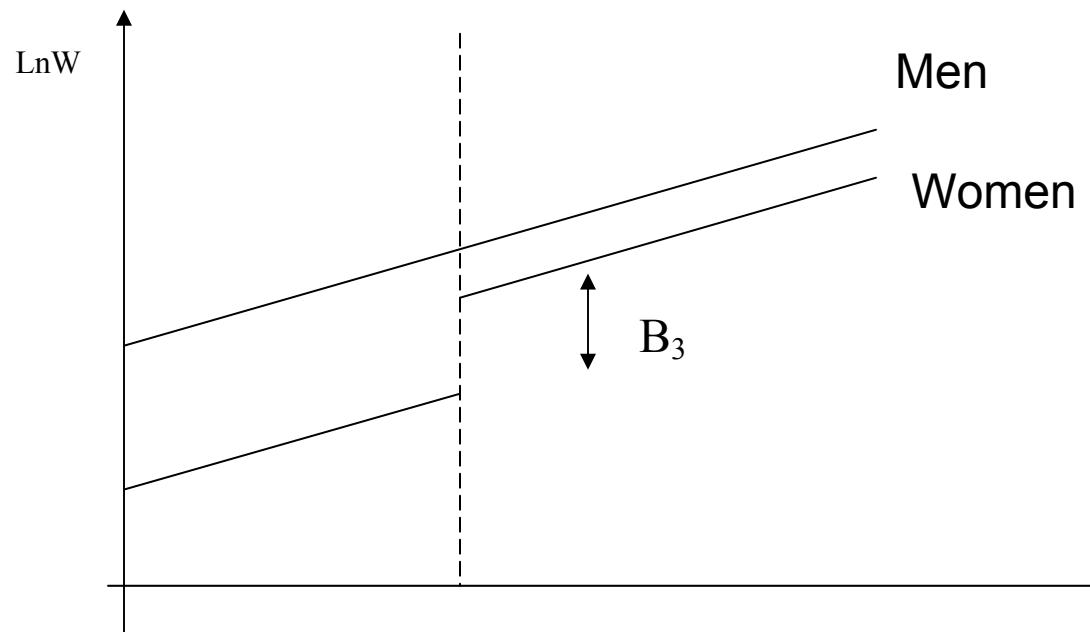
. reg lhw age postgrad grad highint low						
Source	SS	df	MS	Number of obs = 12098		
Model	932.600688	5	186.520138	F( 5, 12092) = 747.70		
Residual	3016.44842	12092	.249458189	Prob > F = 0.0000		
Total	3949.04911	12097	.326448633	R-squared = 0.2362		
				Adj R-squared = 0.2358		
				Root MSE = .49946		
lhw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.010341	.0004148	24.93	0.000	.009528	.0111541
postgrad	.9497773	.0242098	39.23	0.000	.9023222	.9972324
grad	.8573589	.0189204	45.31	0.000	.8202718	.894446
highint	.5486204	.0174109	31.51	0.000	.5144922	.5827486
low	.2774401	.0151439	18.32	0.000	.2477555	.3071246
_cons	1.160484	.0231247	50.18	0.000	1.115156	1.205812

and now the coefficients are all positive (relative to those with no qual.)

## Dummy Variables and Policy Analysis

One important practical use of a regression is to try and evaluate the “treatment effect” of a policy intervention.

- 3) To Measure Effects of ***Change in the Average Behaviour*** of two groups, one subject to a policy the other not (the Difference-in-Difference Estimator)



Treatment/Policy affects only **a sub-section** of the population  
Eg A drug, EMA, Change in Tuition Fees, Minimum Wage

and may lead to a change in behaviour for the treated group - as captured by a change in the intercept (or slope) **after** the intervention (treatment) takes place

## Dummy Variables and Policy Analysis

One important practical use of a regression is to try and evaluate the “treatment effect” of a policy intervention.

Usually this means comparing outcomes for those affected by a policy that is of concern to economists

Eg a law on taxing cars in central London – creates a “treatment” group, (eg those who drive in London) and those not, (the “control” group).

Other examples targeted tax cuts, minimum wages,  
area variation in schooling practices, policing

In principle one could set up a dummy variable to denote membership of the treatment group (or not) and run the following regression

$$\ln W = a + b * \text{Treatment Dummy} + u \quad (1)$$

In principle one could set up a dummy variable to denote membership of the treatment group (or not) and run the following regression

$$\ln W = a + b * \text{Treatment Dummy} + u \quad (1)$$

where Treatment = 1 if exposed to a treatment = 0 if not

```
reg price newham if time>3 & (newham==1 | croydon==1)
reg price newham if time<=3 & (newham==1 | croydon==1)
reg price newham after afternew if time>3 & (newham==1 | croydon==1)
```

Problem: a single period regression of the dependent variable on the “treatment” variable as in (1) will **not** give the desired treatment effect.

Problem: a single period regression of the dependent variable on the “treatment” variable as in (1) will **not** give the desired treatment effect.

This is because there may always have been a different value for the treatment group even before the policy intervention took place

Problem: a single period regression of the dependent variable on the “treatment” variable as in (1) will **not** give the desired treatment effect.

This is because there may always have been a different value for the treatment group even before the policy intervention took place

ie. Could estimate

$$\ln W = a + b * \text{Treatment Dummy} + u$$

in the period before any treatment took place

Problem: a single period regression of the dependent variable on the “treatment” variable as in (1) will **not** give the desired treatment effect.

This is because there may always have been a different value for the treatment group even before the policy intervention took place

ie. Could estimate

$$\ln W = a + b * \text{Treatment Dummy} + u$$

in the period before any treatment took place

but whatever the effect observed it cannot be caused by the treatment since these events are observed before the treatment took place.

Problem: a single period regression of the dependent variable on the “treatment” variable as in (1) will **not** give the desired treatment effect.

This is because there may always have been a different value for the treatment group even before the policy intervention took place

ie. Could estimate

$$\ln W = a + b^* \text{Treatment Dummy} + u$$

in the period before any treatment took place

but whatever the effect observed it cannot be caused by the treatment since these events are observed before the treatment took place.

The idea instead is then to compare the **change** in Y for the treatment group who experienced the shock with the change in Y of the control group who did not

Problem: a single period regression of the dependent variable on the “treatment” variable as in (1) will **not** give the desired treatment effect.

This is because there may always have been a different value for the treatment group even before the policy intervention took place

ie. Could estimate

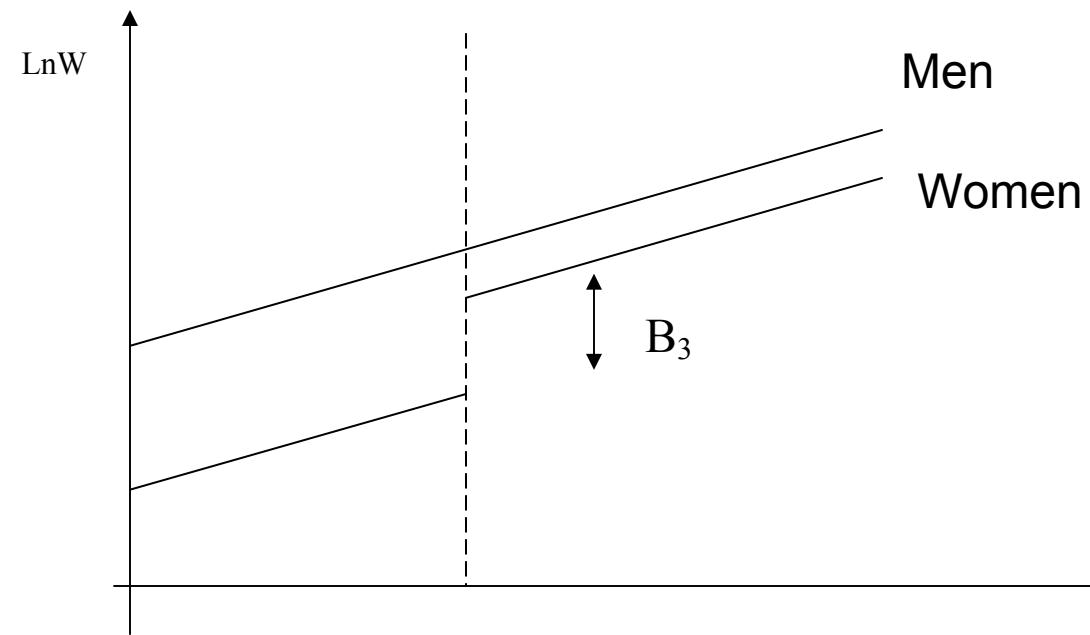
$$\ln W = a + b^* \text{Treatment Dummy} + u$$

in the period before any treatment took place

but whatever the effect observed it cannot be caused by the treatment since these events are observed before the treatment took place.

The idea instead is then to compare the **change** in Y for the treatment group who experienced the shock with the change in Y of the control group who did not

- the “**difference in difference estimator**”



Men

Women

$B_3$

If the change for Treatment group reflects both

$$[Y_t^2 - Y_t^1] = \text{Effect of Policy} + \text{other influences}$$

If the change for Treatment group reflects both

$$[Y_{t^2} - Y_{t^1}] = \text{Effect of Policy} + \text{other influences}$$

but the change for control group is only caused by

If the change for Treatment group reflects both

$$[Y_t^2 - Y_t^1] = \text{Effect of Policy} + \text{other influences}$$

but the change for control group is only caused by

$$[Y_c^2 - Y_c^1] = \text{Effect of other influences}$$

If the change for Treatment group reflects both

$$[Y_{t^2} - Y_{t^1}] = \text{Effect of Policy} + \text{other influences}$$

but the change for control group is only caused by

$$[Y_{c^2} - Y_{c^1}] = \text{Effect of other influences}$$

then the difference in differences

If the change for Treatment group reflects both

$$[Y_t^2 - Y_t^1] = \text{Effect of Policy} + \text{other influences}$$

but the change for control group is only caused by

$$[Y_c^2 - Y_c^1] = \text{Effect of other influences}$$

then the difference in differences

$$[Y_t^2 - Y_t^1] - [Y_c^2 - Y_c^1] =$$

Effect of Policy + other influences

If the change for Treatment group reflects both

$$[Y_t^2 - Y_t^1] = \text{Effect of Policy} + \text{other influences}$$

but the change for control group is only caused by

$$[Y_c^2 - Y_c^1] = \text{Effect of other influences}$$

then the difference in differences

$$[Y_t^2 - Y_t^1] - [Y_c^2 - Y_c^1] =$$

$$\text{Effect of Policy} + \text{other influences} - \text{Effect of other influences}$$

If the change for Treatment group reflects both

$$[Y_t^2 - Y_t^1] = \text{Effect of Policy} + \text{other influences}$$

but the change for control group is only caused by

$$[Y_c^2 - Y_c^1] = \text{Effect of other influences}$$

then the difference in differences

$$[Y_t^2 - Y_t^1] - [Y_c^2 - Y_c^1] =$$

$$\text{Effect of Policy} + \text{other influences} - \text{Effect of other influences}$$

$$= \text{Effect of Policy}$$

If the change for Treatment group reflects both

$$[Y_t^2 - Y_t^1] = \text{Effect of Policy} + \text{other influences}$$

but the change for control group is only caused by

$$[Y_c^2 - Y_c^1] = \text{Effect of other influences}$$

then the difference in differences

$$[Y_t^2 - Y_t^1] - [Y_c^2 - Y_c^1] =$$

$$\text{Effect of Policy} + \text{other influences} - \text{Effect of other influences}$$

$$= \text{Effect of Policy}$$

(assuming the effect of other influences is the same for both treatment and control groups )

If the change for Treatment group reflects both

$$[Y_t^2 - Y_t^1] = \text{Effect of Policy} + \text{other influences}$$

but the change for control group is only caused by

$$[Y_c^2 - Y_c^1] = \text{Effect of other influences}$$

then the difference in differences

$$[Y_t^2 - Y_t^1] - [Y_c^2 - Y_c^1] =$$

$$\text{Effect of Policy} + \text{other influences} - \text{Effect of other influences}$$

$$= \text{Effect of Policy}$$

(assuming the effect of other influences is the same for both treatment and control groups )

Hence the need to try and choose a control group that is similar to the treatment group (apart from the experience of the treatment)

In practice this estimator can be obtained by combining (pooling) the data over the periods before and after and running the following regression

$$\ln W = a + a_2 \text{After} + b_1 \text{Treatment Dummy} + b_2 \text{After} * \text{Treatment Dummy}$$

In practice this estimator can be obtained by combining (pooling) the data over the periods before and after and running the following regression

$$\ln W = a + a_2 \text{After} + b_1 \text{Treatment Dummy} + b_2 \text{After} * \text{Treatment Dummy}$$

where now

**After** is a dummy variable

- = 1 if data observed after the treatment
- = 0 if data observed before the treatment

In practice this estimator can be obtained by combining (pooling) the data over the periods before and after and running the following regression

$$\ln W = a + a_2 \text{After} + b_1 \text{Treatment Dummy} + b_2 \text{After} * \text{Treatment Dummy}$$

where now

**After** is a dummy variable

- = 1 if data observed after the treatment
- = 0 if data observed before the treatment

a is the average wage of the control group in the base year,  
a<sub>2</sub>, is the average wage of the control group in the second year,  
b<sub>1</sub> gives the difference on wages between treatment and control group  
in the base year

b<sub>2</sub> is the “difference in difference” estimator – the change in wages for  
the treatment group relative to the control in the second period.

Why ?

$$\ln W = a + a_2 \text{After} + b_1 \text{Treatment Dummy} + b_2 \text{After} * \text{Treatment Dummy}$$

$$\ln W = a + a_2 \text{After} + b_1 \text{Treatment Dummy} + b_2 \text{After} * \text{Treatment Dummy}$$

If After=0 and Treatment Dummy = 0

$$\ln W = a + a_2 \text{After} + b_1 \text{Treatment Dummy} + b_2 \text{After} * \text{Treatment Dummy}$$

If After=0 and Treatment Dummy = 0

$$\ln W = a + a_2 * 0 + b_1 * 0 + b_2 * 0$$

$$\ln W = a + a_2 \text{After} + b_1 \text{Treatment Dummy} + b_2 \text{After} * \text{Treatment Dummy}$$

If After=0 and Treatment Dummy = 0

$$\begin{aligned}\ln W &= a + a_2 * 0 + b_1 * 0 + b_2 * 0 \\ &= a\end{aligned}$$

$$\ln W = a + a_2 \text{After} + b_1 \text{Treatment Dummy} + b_2 \text{After} * \text{Treatment Dummy}$$

If After=0 and Treatment Dummy = 0

$$\begin{aligned}\ln W &= a + a_2 * 0 + b_1 * 0 + b_2 * 0 \\ &= a\end{aligned}$$

Similarly

If After =0 and Treatment Dummy = 1,  $\ln W = a + b_1$

$$\ln W = a + a_2 \text{After} + b_1 \text{Treatment Dummy} + b_2 \text{After} * \text{Treatment Dummy}$$

If After=0 and Treatment Dummy = 0

$$\begin{aligned}\ln W &= a + a_2 * 0 + b_1 * 0 + b_2 * 0 \\ &= a\end{aligned}$$

Similarly

If After =0 and Treatment Dummy = 1,  $\ln W = a + b_1$

If After =1 and Treatment Dummy = 0,  $\ln W = a + a_2$

$$\ln W = a + a_2 \text{After} + b_1 \text{Treatment Dummy} + b_2 \text{After} * \text{Treatment Dummy}$$

If After=0 and Treatment Dummy = 0

$$\begin{aligned}\ln W &= a + a_2 * 0 + b_1 * 0 + b_2 * 0 \\ &= a\end{aligned}$$

Similarly

If After =0 and Treatment Dummy = 1,  $\ln W = a + b_1$

If After =1 and Treatment Dummy = 0,  $\ln W = a + a_2$

If After =0 and Treatment Dummy = 1,  $\ln W = a + a_2 + b_1 + b_2$

$$\ln W = a + a_2 \text{After} + b_1 \text{Treatment Dummy} + b_2 \text{After} * \text{Treatment Dummy}$$

If After=0 and Treatment Dummy = 0

$$\begin{aligned}\ln W &= a + a_2 * 0 + b_1 * 0 + b_2 * 0 \\ &= a\end{aligned}$$

Similarly

If After =0 and Treatment Dummy = 1,  $\ln W = a + b_1$

If After =1 and Treatment Dummy = 0,  $\ln W = a + a_2$

If After =0 and Treatment Dummy = 1,  $\ln W = a + a_2 + b_1 + b_2$

So the change in wages for the treatment group is

$$\ln W = a + a_2 \text{After} + b_1 \text{Treatment Dummy} + b_2 \text{After} * \text{Treatment Dummy}$$

If After=0 and Treatment Dummy = 0

$$\begin{aligned}\ln W &= a + a_2 * 0 + b_1 * 0 + b_2 * 0 \\ &= a\end{aligned}$$

Similarly

If After =0 and Treatment Dummy = 1,  $\ln W = a + b_1$

If After =1 and Treatment Dummy = 0,  $\ln W = a + a_2$

If After =0 and Treatment Dummy = 1,  $\ln W = a + a_2 + b_1 + b_2$

So the change in wages for the treatment group is

$$(a + a_2 + b_1 + b_2)$$

$$\ln W = a + a_2 \text{After} + b_1 \text{Treatment Dummy} + b_2 \text{After} * \text{Treatment Dummy}$$

If After=0 and Treatment Dummy = 0

$$\begin{aligned}\ln W &= a + a_2 * 0 + b_1 * 0 + b_2 * 0 \\ &= a\end{aligned}$$

Similarly

If After =0 and Treatment Dummy = 1,  $\ln W = a + b_1$

If After =1 and Treatment Dummy = 0,  $\ln W = a + a_2$

If After =0 and Treatment Dummy = 1,  $\ln W = a + a_2 + b_1 + b_2$

So the change in wages for the treatment group is

$$(a + a_2 + b_1 + b_2) - (a + b_1)$$

$$\ln W = a + a_2 \text{After} + b_1 \text{Treatment Dummy} + b_2 \text{After} * \text{Treatment Dummy}$$

If After=0 and Treatment Dummy = 0

$$\begin{aligned}\ln W &= a + a_2 * 0 + b_1 * 0 + b_2 * 0 \\ &= a\end{aligned}$$

Similarly

If After =0 and Treatment Dummy = 1,  $\ln W = a + b_1$

If After =1 and Treatment Dummy = 0,  $\ln W = a + a_2$

If After =0 and Treatment Dummy = 1,  $\ln W = a + a_2 + b_1 + b_2$

So the change in wages for the treatment group is

$$(a + a_2 + b_1 + b_2) - (a + b_1) = a_2 + b_2$$

$$\ln W = a + a_2 \text{After} + b_1 \text{Treatment Dummy} + b_2 \text{After} * \text{Treatment Dummy}$$

If After=0 and Treatment Dummy = 0

$$\begin{aligned}\ln W &= a + a_2 * 0 + b_1 * 0 + b_2 * 0 \\ &= a\end{aligned}$$

Similarly

If After =0 and Treatment Dummy = 1,  $\ln W = a + b_1$

If After =1 and Treatment Dummy = 0,  $\ln W = a + a_2$

If After =0 and Treatment Dummy = 1,  $\ln W = a + a_2 + b_1 + b_2$

So the change in wages for the treatment group is

$$(a + a_2 + b_1 + b_2) - (a + b_1) = a_2 + b_2$$

and the change in wages for the control group is

$$(a + a_2) - (a)$$

$$\ln W = a + a_2 \text{After} + b_1 \text{Treatment Dummy} + b_2 \text{After} * \text{Treatment Dummy}$$

If After=0 and Treatment Dummy = 0

$$\begin{aligned}\ln W &= a + a_2 * 0 + b_1 * 0 + b_2 * 0 \\ &= a\end{aligned}$$

Similarly

If After =0 and Treatment Dummy = 1,  $\ln W = a + b_1$

If After =1 and Treatment Dummy = 0,  $\ln W = a + a_2$

If After =0 and Treatment Dummy = 1,  $\ln W = a + a_2 + b_1 + b_2$

So the change in wages for the treatment group is

$$(a + a_2 + b_1 + b_2) - (a + b_1) = a_2 + b_2$$

and the change in wages for the control group is

$$(a + a_2) - (a) = a_2$$

$$\ln W = a + a_2 \text{After} + b_1 \text{Treatment Dummy} + b_2 \text{After} * \text{Treatment Dummy}$$

If After=0 and Treatment Dummy = 0

$$\ln W = a + a_2 * 0 + b_1 * 0 + b_2 * 0$$

$= a$

Similarly

If After =0 and Treatment Dummy = 1,  $\ln W = a + b_1$

If After =1 and Treatment Dummy = 0,  $\ln W = a + a_2$

If After =0 and Treatment Dummy = 1,  $\ln W = a + a_2 + b_1 + b_2$

So the change in wages for the treatment group is

$$(a + a_2 + b_1 + b_2) - (a + b_1) = a_2 + b_2$$

and the change in wages for the control group is

$$(a + a_2) - (a) = a_2$$

so the "difference in difference" estimator

= Change in wages for treatment – change in wages for control

$$= (a_2 + b_2) - (a_2) = b_2$$

$$\ln W = a + a_2 \text{After} + b_1 \text{Treatment Dummy} + b_2 \text{After} * \text{Treatment Dummy}$$

If After=0 and Treatment Dummy = 0

$$\begin{aligned}\ln W &= a + a_2 * 0 + b_1 * 0 + b_2 * 0 \\ &= a\end{aligned}$$

Similarly

If After =0 and Treatment Dummy = 1,  $\ln W = a + b_1$

If After =1 and Treatment Dummy = 0,  $\ln W = a + a_2$

If After =0 and Treatment Dummy = 1,  $\ln W = a + a_2 + b_1 + b_2$

So the change in wages for the treatment group is

$$(a + a_2 + b_1 + b_2) - (a + b_1) = a_2 + b_2$$

and the change in wages for the control group is

$$(a + a_2) - (a) = a_2$$

so the “difference in difference” estimator

= Change in wages for treatment – change in wages for control

$$\ln W = a + a_2 \text{After} + b_1 \text{Treatment Dummy} + b_2 \text{After} * \text{Treatment Dummy}$$

If After=0 and Treatment Dummy = 0

$$\begin{aligned}\ln W &= a + a_2 * 0 + b_1 * 0 + b_2 * 0 \\ &= a\end{aligned}$$

Similarly

If After =0 and Treatment Dummy = 1,  $\ln W = a + b_1$

If After =1 and Treatment Dummy = 0,  $\ln W = a + a_2$

If After =0 and Treatment Dummy = 1,  $\ln W = a + a_2 + b_1 + b_2$

So the change in wages for the treatment group is

$$(a + a_2 + b_1 + b_2) - (a + b_1) = a_2 + b_2$$

and the change in wages for the control group is

$$(a + a_2) - (a) = a_2$$

so the “difference in difference” estimator

$$\begin{aligned}&= \text{Change in wages for treatment} - \text{change in wages for control} \\ &= (a_2 + b_2) - (a_2)\end{aligned}$$

$$\ln W = a + a_2 \text{After} + b_1 \text{Treatment Dummy} + b_2 \text{After} * \text{Treatment Dummy}$$

If After=0 and Treatment Dummy = 0

$$\begin{aligned}\ln W &= a + a_2 * 0 + b_1 * 0 + b_2 * 0 \\ &= a\end{aligned}$$

Similarly

If After =0 and Treatment Dummy = 1,  $\ln W = a + b_1$

If After =1 and Treatment Dummy = 0,  $\ln W = a + a_2$

If After =0 and Treatment Dummy = 1,  $\ln W = a + a_2 + b_1 + b_2$

So the change in wages for the treatment group is

$$(a + a_2 + b_1 + b_2) - (a + b_1) = a_2 + b_2$$

and the change in wages for the control group is

$$(a + a_2) - (a) = a_2$$

so the “difference in difference” estimator

$$\begin{aligned}&= \text{Change in wages for treatment} - \text{change in wages for control} \\ &= (a_2 + b_2) - (a_2) = b_2\end{aligned}$$

House prices in East End rise after Olympic win - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Getting Started Latest Headlines

home | treasury | about hbos | economics | community

HBOS plc logo

Back to home page

HBOS plc home > Media Centre > House prices in East End rise after Olympic win

Take me to... 635.00 -24.00p  
20/02/2008 12:40  
search this site Go Our share price...

# Halifax press release

## House prices in East End rise after Olympic win

### Friday 2nd February 2007

With 2,000 days to go to the start of the 2012 Olympics, new research from Halifax Estate Agents shows that house prices in three London postal districts close to the site of 2012 Olympics games have risen by more than 15%, or at least £35,000, since London's winning bid was announced.

Across London, house prices have risen by 15% since Q2 2005.

The best performance has been in Leytonstone (E11), which saw a 23% (£50,714) increase in its average house price since mid 2005 followed by Hackney (E8) with a 21% (£48,578) increase and Clapton (E5) 18% (£38,895) rise.

Seven areas close to the Olympic site recorded house price increases of more than 10% since Q2 2006, while all areas close to the Games site have seen at least a £15,000 rise in their average house price. (Table 1)

Stratford (E15), the focal point for Olympic construction activity, saw an 8% (£16,801) increase in its average price since June 2005 to £225,652.

#### Previous Olympic host cities have seen strengthening house prices

Each of the previous four host cities (Barcelona, Atlanta, Sydney, Athens) have seen house prices rise by more than the national average over the five year period in the run-up to the Olympic games, the main period of Olympic related development activity. These host cities averaged house price increases of 66% over five years against an average rise in host nation house prices of 47% - a differential of 19 percentage points. (Table 2)

#### Key Findings

**London (See Table 1)**

- The postal district near the Olympic site with the highest average house price is Leytonstone (E11) - £275,827. The next most expensive Olympic areas are Hackney (E8) £274,948 and Clapton (E5) £258,394

Done

start Auto Sync TextPad - [W... Microsoft... Adobe Acrobat... Stata/SE 10.0... Windows... Microsoft... House prices i... Desktop

EN 12:57 Wednesday 20/02/2008

**Example** In July 2005, London won the rights to host the 2012 Olympic games. Shortly afterward there were media reports that house prices were rising “fast” in areas close to the Olympic site. Can evaluate whether this was true by using Newham as the Treatment area (the borough in which the Olympic site is located) and a similar London borough further away from the site as a “control”.

The data set *olympics.dta* has monthly data on house prices over time in Newham & Hounslow. The dummy variable “newham” takes the value 1 if the house price observation is from Newham and 0 if not. The dummy variable “after” takes the value 1 if the month was after the Olympic announcement and 0 otherwise. The interaction term “afternew”

g afternew=after\*newham

takes the value 1 only if the month is after the event and the observation is in Newham. The coefficient on this term will be the difference-in-difference estimator (the differential effect of the Olympic bid on house prices in newham relative to the control area of croydon).

. reg price after if newham==1					
Source	SS	df	MS	Number of obs = 81	
Model	3.8272e+10	1	3.8272e+10	F( 1, 79) = 38.09	
Residual	7.9385e+10	79	1.0049e+09	Prob > F = 0.0000	
Total	1.1766e+11	80	1.4707e+09	R-squared = 0.3253	
				Adj R-squared = 0.3167	
				Root MSE = 31700	
price	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
after	53378.93	8649.343	6.17	0.000	36162.84 70595.02
_cons	165035	3962.462	41.65	0.000	157147.9 172922

House prices were indeed higher in Newham after the Olympic announcement, but...

. reg price after if hounslow==1					
Source	SS	df	MS	Number of obs = 80	
Model	2.9394e+10	1	2.9394e+10	F( 1, 78) = 36.75	
Residual	6.2388e+10	78	799846080	Prob > F = 0.0000	
Total	9.1782e+10	79	1.1618e+09	R-squared = 0.3203	
				Adj R-squared = 0.3115	
				Root MSE = 28282	
price	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]

after	46857.27	7729.537	6.06	0.000	31468.94	62245.6
_cons	205399.7	3563.14	57.65	0.000	198306.1	212493.4

they were also higher in Hounslow

Moreover the annual rate of growth of house prices (approximated by the log of the 12 month change) was not significantly different in the after period

. reg dlogp after if newham==1	Source   SS df MS	Number of obs = 72
		F( 1, 70) = 0.74
	Model   .025162236 1 .025162236	Prob > F = 0.3931
	Residual   2.38491217 70 .034070174	R-squared = 0.0104
	Total   2.4100744 71 .03394471	Adj R-squared = -0.0037
		Root MSE = .18458
	dlogp   Coef. Std. Err. t P> t  [95% Conf. Interval]	
	after   -.0440185 .0512209 -0.86 0.393 -.1461754 .0581385	
	_cons   .0868633 .0248889 3.49 0.001 .037224 .1365027	

and the difference-in-difference analysis confirms that there was no differential house price growth between the two areas. It seems claims of a house price effect were exaggerated.

reg logp after newham afternew if newham==1   hounslow==1	Source   SS df MS	Number of obs = 161
		F( 3, 157) = 40.90
	Model   3.70463956 3 1.23487985	Prob > F = 0.0000
	Residual   4.74020159 157 .030192367	R-squared = 0.4387
	Total   8.44484115 160 .052780257	Adj R-squared = 0.4280
		Root MSE = .17376
	logp   Coef. Std. Err. t P> t  [95% Conf. Interval]	
	after   .2172745 .0474896 4.58 0.000 .1234735 .3110755	
	newham   -.2308987 .0308383 -7.49 0.000 -.2918101 -.1699872	
	afternew   .0868757 .0671047 1.29 0.197 -.0456688 .2194202	
	_cons   12.22069 .0218916 558.24 0.000 12.17745 12.26393	

## **Using Dummy Variables to capture Seasonality in Data**

Can also use dummy variables to pick out and control for seasonal variation in data

ELMR\_Feb08.pdf (application/pdf Object) - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Getting Started Latest Headlines

Save a Copy Search Select ABC Sign Y!

http://www.statistics.gov.uk/elmr/02\_08/downloads/ELMR\_Feb08.pdf

Economic & Labour Market Review | Vol 2 | No 2 | February 2008

## Key time series

### National accounts aggregates

Last updated: 23/01/08

Seasonally adjusted

	£ million		Indices (2003 = 100)								
	At current prices		Value indices at current prices				Chained volume indices			Implied deflators <sup>3</sup>	
	Gross domestic product (GDP) at market prices	Gross value added (GVA) at basic prices	GDP at market prices <sup>1</sup>	GVA at basic prices	Gross national disposable income at market prices <sup>2</sup>	GDP at market prices	GVA at basic prices	GDP at market prices	GVA at basic prices		
	YBHA	ABML	YBEU	YBEX	YBFP	YBEZ	CGCE	YBGB	CGBV		
2002	1,055,793	937,323	94.4	94.3	97.1	97.3	97.3	97.0	97.0		
2003	1,118,245	993,507	100.0	100.0	100.0	100.0	100.0	100.0	100.0		
2004	1,184,296	1,051,934	105.9	105.9	103.4	103.3	103.3	102.6	102.5		
2005	1,233,976	1,096,629	110.3	110.4	104.2	105.2	105.2	104.9	104.9		
2006	1,303,573	1,158,871	116.6	116.6	105.8	108.2	108.3	107.7	107.7		
2007					111.6	111.7					
2002 Q1	259,054	229,737	92.7	92.5	95.9	96.4	96.5	96.1	95.9		
2002 Q2	262,774	233,372	94.0	94.0	96.2	97.0	96.9	96.9	97.0		
2002 Q3	265,836	236,103	95.1	95.1	98.3	97.7	97.6	97.4	97.4		
2002 Q4	268,129	238,111	95.9	95.9	98.2	98.2	98.1	97.7	97.7		
2003 Q1	272,953	242,612	97.6	97.7	99.4	98.8	98.8	98.9	98.9		
2003 Q2	277,119	246,427	99.1	99.2	98.9	99.3	99.3	99.8	99.9		
2003 Q3	281,996	250,492	100.9	100.9	100.0	100.4	100.4	100.4	100.5		
2003 Q4	286,177	253,976	102.4	102.3	101.7	101.5	101.6	100.9	100.7		
2004 Q1	288,912	256,106	103.3	103.1	101.9	102.2	102.2	101.1	100.9		
2004 Q2	295,066	262,094	105.5	105.5	103.2	103.1	103.2	102.3	102.3		
2004 Q3	297,941	264,732	106.6	106.6	103.0	103.5	103.5	102.9	103.0		
2004 Q4	302,377	269,002	108.2	108.3	105.4	104.1	104.2	103.9	104.0		

Waiting for www.statistics.gov.uk...

62 of 72

## **Using Dummy Variables to capture Seasonality in Data**

Can also use dummy variables to pick out and control for seasonal variation in data

The idea is to include a set of dummy variables for each quarter (or month or day) which will then net out the average change in a variable resulting from any seasonal fluctuations

## Using Dummy Variables to capture Seasonality in Data

Can also use dummy variables to pick out and control for seasonal variation in data

The idea is to include a set of dummy variables for each quarter (or month or day) which will then net out the average change in a variable resulting from any seasonal fluctuations

$$Y_t = b_0 + b_1 Q1 + b_2 Q2 + b_3 Q3 + b_4 X + u_t$$

## Using Dummy Variables to capture Seasonality in Data

Can also use dummy variables to pick out and control for seasonal variation in data

The idea is to include a set of dummy variables for each quarter (or month or day) which will then net out the average change in a variable resulting from any seasonal fluctuations

$$Y_t = b_0 + b_1 Q1 + b_2 Q2 + b_3 Q3 + b_4 X + u_t$$

Hence the coefficient on the quarterly dummy Q1  
=1 if data belong to the 1<sup>st</sup> quarter of the year (Jan-Mar)  
= 0 otherwise

## Using Dummy Variables to capture Seasonality in Data

Can also use dummy variables to pick out and control for seasonal variation in data

The idea is to include a set of dummy variables for each quarter (or month or day) which will then net out the average change in a variable resulting from any seasonal fluctuations

$$Y_t = b_0 + b_1 Q1 + b_2 Q2 + b_3 Q3 + b_4 X + u_t$$

Hence the coefficient on the quarterly dummy Q1  
=1 if data belong to the 1<sup>st</sup> quarter of the year (Jan-Mar)  
= 0 otherwise

gives the level of Y in the 1<sup>st</sup> quarter of the year relative to the constant (Q4 level of Y) averaged over all Q1 observations in the data set

## Using Dummy Variables to capture Seasonality in Data

Can also use dummy variables to pick out and control for seasonal variation in data

The idea is to include a set of dummy variables for each quarter (or month or day) which will then net out the average change in a variable resulting from any seasonal fluctuations

$$Y_t = b_0 + b_1 Q1 + b_2 Q2 + b_3 Q3 + b_4 X + u_t$$

Hence the coefficient on the quarterly dummy Q1  
=1 if data belong to the 1<sup>st</sup> quarter of the year (Jan-Mar)  
= 0 otherwise

gives the level of Y in the 1<sup>st</sup> quarter of the year relative to the constant (Q4 level of Y) averaged over all Q1 observations in the data set

Series net of seasonal effects are said to be “seasonally adjusted”

It may also be useful to model an economic series as a combination of seasonal and a trend component

It may also be useful to model an economic series as a combination of seasonal and a trend component

$$Y_t = b_0 + b_1 Q1 + b_2 Q2 + b_3 Q3 + b_4 \text{Trend} + u_t$$

It may also be useful to model an economic series as a combination of seasonal and a trend component

$$Y_t = b_0 + b_1 Q1 + b_2 Q2 + b_3 Q3 + b_4 \text{Trend} + u_t$$

where Trend = 1 in year 1

It may also be useful to model an economic series as a combination of seasonal and a trend component

$$Y_t = b_0 + b_1Q1 + b_2Q2 + b_3Q3 + b_4\text{Trend} + u_t$$

where Trend      = 1 in year 1  
                      = 2 in year 2

It may also be useful to model an economic series as a combination of seasonal and a trend component

$$Y_t = b_0 + b_1 Q1 + b_2 Q2 + b_3 Q3 + b_4 \text{Trend} + u_t$$

where Trend      = 1 in year 1  
                  = 2 in year 2  
                  = T in year T

It may also be useful to model an economic series as a combination of seasonal and a trend component

$$Y_t = b_0 + b_1 Q1 + b_2 Q2 + b_3 Q3 + b_4 \text{Trend} + u_t$$

where Trend      = 1 in year 1  
                      = 2 in year 2  
                      = T in year T

since  $dY_t/d\text{Trend} = b_4$

given that the coefficient measures the unit change in y for a unit change in the trend variable

and the units of measurement in this case are years

It may also be useful to model an economic series as a combination of seasonal and a trend component

$$Y_t = b_0 + b_1 Q1 + b_2 Q2 + b_3 Q3 + b_4 \text{Trend} + u_t$$

where Trend      = 1 in year 1  
                      = 2 in year 2

                      = T in year T

since  $dY_t/d\text{Trend} = b_4$

given that the coefficient measures the unit change in y for a unit change in the trend variable

and the units of measurement in this case are years

then in the model above the trend term measures the annual change in the Y variable net of any seasonal influences

Department for Transport - Tomorrow's roads: safer for everyone - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Getting Started Latest Headlines

<http://www.dft.gov.uk/pgr/roadsafety/strategytargetsperformance/tomorrowsroadssafeforeveryone?page=2>

Home | Accessibility | Cymraeg | Contact us | Help | What's new | A to Z index | Site map

Search Advanced search Search other DfT sites

About DfT Policy, guidance and research Press office Consultations Freedom of Information Transport for you In your area Popular pages

DfT home > Policy, guidance and research > Road safety > Strategy, targets and performance

**Tomorrow's roads: safer for everyone**

> Back to contents Print this page Print all pages Download PDF

**Chapter 1 - Introduction**

**Road accidents**

**1.1** Road accidents cause immense human suffering. Every year, around 3,500 people are killed on Britain's roads and 40,000 are seriously injured. In total, there are over 300,000 road casualties, in nearly 240,000 accidents, and about fifteen times that number of non-injury incidents. This represents a serious economic burden; the direct cost of road accidents involving deaths or injuries is thought to be in the region of £3 billion a year.

**1.2** Nevertheless, Britain has had - relatively speaking - remarkable success in reducing road casualties. And this is despite the vast growth in traffic since the beginning of the last century. In 1930 there were only 2.3 million motor vehicles in Great Britain, but over 7,000 people were killed in road accidents. Today, there are over 27 million vehicles on our roads but far fewer road deaths.

**Indices of traffic and casualties: 1949-1998**

**1.3** In 1987 a target was set to reduce road casualties by one-third by 2000 compared with the average for 1981-85. We have more than achieved this target for reducing deaths and serious injuries. Road deaths have fallen by 39% and serious injuries by 45% and we are now one of the safest countries in Europe and indeed the world. However, there has not been any such steep decline in the number of accidents, nor in the number of slight injuries, although improvements in vehicle design have helped to reduce the severity of injuries to car occupants.

**The new targets**

**1.4** There is no reason for us to be complacent. We know we can reduce road casualties still further. That is why we are setting a new 10-year target and launching this new road safety strategy. We need new targets to help everyone to focus on achieving a further substantial improvement in road safety over the next 10 years. By 2010 we want to achieve, compared with the average for 1994-98:

- a 40% reduction in the number of people killed or seriously injured in road accidents;

See also

> Child road safety: achieving the 2010 target - Full report (PDF 432 kb)

Done

start Auto Sync Lecture Hand... Lecture 12\_U... Stata/SE 10.... Stata Graph .... KINGSTON (E:) Department f... Adobe Acrobat... Desktop 11:35 Thursday 28/02/2008

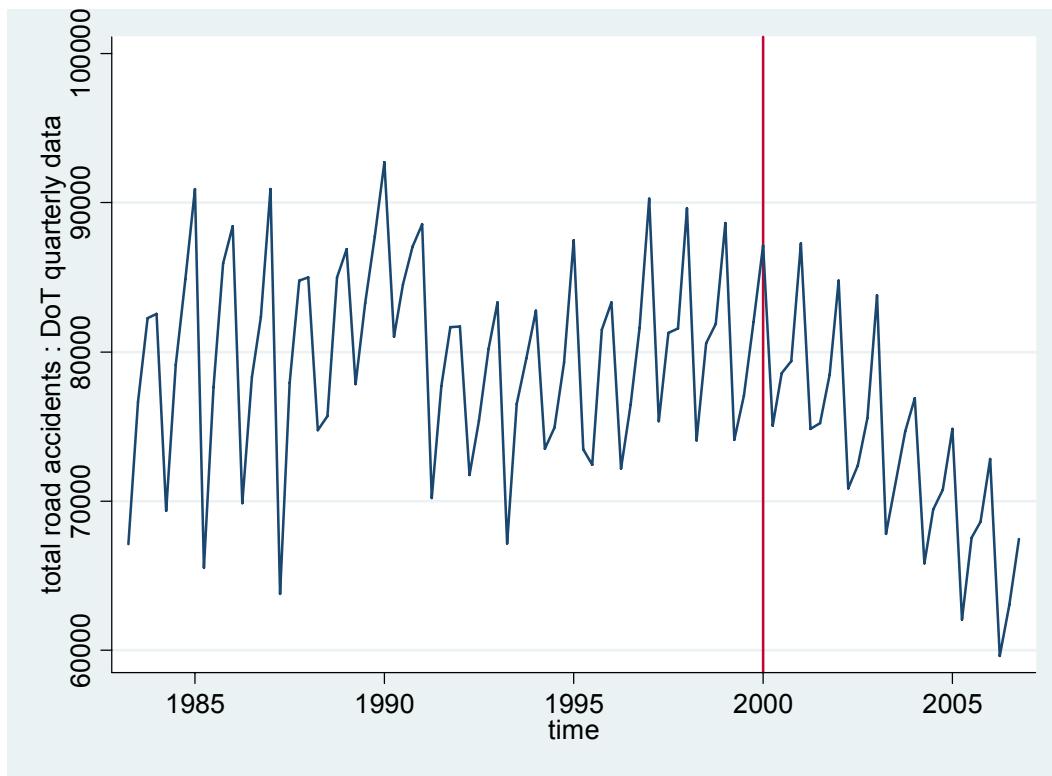
## The new targets

**1.4** There is no reason for us to be complacent. We know we can reduce road casualties still further. That is why we are setting a new 10-year target and launching this new road safety strategy. We need new targets to help everyone to focus on achieving a further substantial improvement in road safety over the next 10 years. By 2010 we want to achieve, compared with the average for 1994-98:

- a 40% reduction in the number of people killed or seriously injured in road accidents;
- a 50% reduction in the number of children killed or seriously injured; and
- a 10% reduction in the slight casualty rate, expressed as the number of people slightly injured per 100 million vehicle kilometres.

The data set accidents.dta (on the course web site) contains quarterly information on the number of road accidents in the UK from 1983 to 2006

```
twoway (line acc time, xline(2000) )
```



The graph shows that road accidents vary more **within** than **between** years

Can see seasonal influence from a regression of number of accidents on 3 dummy variables (1 for each quarter minus the default category – which is the 4<sup>th</sup> quarter)

```
. list acc year quart time Q1 Q2 Q3 Q4, clean
```

	acc	year	quart	time	Q1	Q2	Q3	Q4
1.	67135	1983	Q1	1983.25	1	0	0	0
2.	76622	1983	Q2	1983.5	0	1	0	0
3.	82277	1983	Q3	1983.75	0	0	1	0
4.	82550	1983	Q4	1984	0	0	0	1
5.	69362	1984	Q1	1984.25	1	0	0	0
6.	79124	1984	Q2	1984.5	0	1	0	0

A regression of road accident numbers on quarterly dummies (q4=winter is default given by constant term at 85249 accidents, on average in the 4<sup>th</sup> quarter) shows accidents are significantly less likely to happen outside the fourth quarter (October-December). On average there are 14,539 fewer accidents in the first quarter of the year than in the last

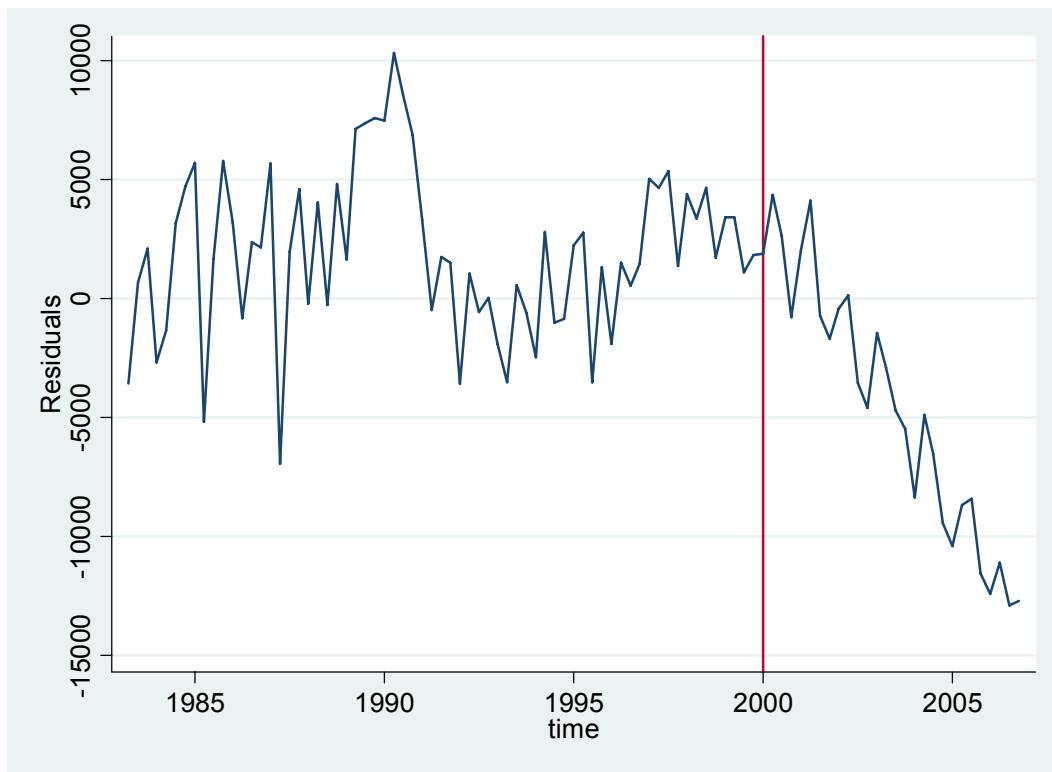
```
. reg acc Q1 Q2 Q3
```

Source	SS	df	MS	Number of obs	=	95
Model	2.6976e+09	3	899214117	F( 3, 91)	=	34.16
Residual	2.3957e+09	91	26326242.3	Prob > F	=	0.0000
Total	5.0933e+09	94	54184365.9	R-squared	=	0.5296
				Adj R-squared	=	0.5141
				Root MSE	=	5130.9
acc	Coef.	Std. Err.	t	P> t	[95% Conf.	Interval]
Q1	-14539.44	1497.179	-9.71	0.000	-17513.4	-11565.48
Q2	-9292.567	1497.179	-6.21	0.000	-12266.53	-6318.604
Q3	-5074.609	1497.179	-3.39	0.001	-8048.572	-2100.646
_cons	85249.61	1069.869	79.68	0.000	83124.45	87374.77

Saving residual values after netting out the influence of the seasons is the basis for the production of “**seasonally adjusted**” data (better guide to underlying trend), used in many official government statistics.

Can get a sense of how this works with the following command after a regression

```
. predict rhat, resid  
/* saves the residuals in a new variable with the name "rhat" */
```



Graph of the residuals is much smoother than the original series – it should be since much of the seasonality has been taken out by the dummy variables. The graph also shows that once seasonality accounted for, there is little evidence in a change in the number of road accidents over time until the year 2000

To model both seasonal and trend components of an economic series, simply include both seasonal dummies and a time trend in the regression model

$$Y_t = b_0 + b_1 Q_1 + b_2 Q_2 + b_3 Q_3 + b_4 \text{TREND} + u_t$$

. reg acc Q1 Q2 Q3 year

Source	SS	df	MS	Number of obs	=	95
--------	----	----	----	---------------	---	----

					F( 4, 90) =	45.39
Model	3.4052e+09	4	851308410		Prob > F	= 0.0000
Residual	1.6881e+09	90	18756630.6		R-squared	= 0.6686
Total	5.0933e+09	94	54184365.9		Adj R-squared	= 0.6538
					Root MSE	= 4330.9
acc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
Q1	-14340.33	1264.153	-11.34	0.000	-16851.79	-11828.87
Q2	-9093.455	1264.153	-7.19	0.000	-11604.92	-6581.995
Q3	-4875.497	1264.153	-3.86	0.000	-7386.958	-2364.037
year	-398.2231	64.83547	-6.14	0.000	-527.0301	-269.4161
_cons	879306.5	129285.1	6.80	0.000	622459.1	1136154

Can see that there is a downward trend in road accidents (of around 400 a year over the whole sample period) net of any seasonality. Could also use dummy variable interactions to test whether this trend is stronger after 2000. How?

Can also use seasonal dummy variables to check whether an apparent association between variables is in fact caused by seasonality in the data

Source	SS	df	MS	Number of obs	=	71
Model	236050086	1	236050086	F( 1, 69) =		6.19
Residual	2.6325e+09	69	38151620.6	Prob > F	=	0.0153
Total	2.8685e+09	70	40978741.5	R-squared	=	0.0823
				Adj R-squared	=	0.0690
				Root MSE	=	6176.7
acc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
du	-4104.777	1650.228	-2.49	0.015	-7396.892	-812.662
_cons	79558.78	768.3058	103.55	0.000	78026.06	81091.51

The regression suggests a negative association between the change in the unemployment rate and the level of accidents (a 1 percentage point rise in the unemployment rate leads to a fall in the number of accidents by 4104 if this regression is to be believed)

Might this be in part because seasonal movements in both data series are influencing the results (the unemployment rate also varies seasonally, typically higher in q1 of each year)

```
. reg acc du q2-q4
```

Source	SS	df	MS	Number of obs = 71 F( 4, 66) = 47.37 Prob > F = 0.0000 R-squared = 0.7417 Adj R-squared = 0.7260 Root MSE = 3350.8			
Model	2.1275e+09	4	531865433				
Residual	741050172	66	11228032.9				
Total	2.8685e+09	70	40978741.5				
acc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]		
du	-1030.818	1009.324	-1.02	0.311	-3045.999	984.3627	
q2	5132.594	1266.59	4.05	0.000	2603.766	7661.422	
q3	10093.64	1174.291	8.60	0.000	7749.089	12438.18	
q4	14353.92	1212.479	11.84	0.000	11933.13	16774.72	
_cons	72488.21	834.607	86.85	0.000	70821.87	74154.56	

Can see if add quarterly seasonal dummy variables then apparent effect of unemployment disappears.

# Likelihood-ratio test

From Wikipedia, the free encyclopedia

In statistics, a **likelihood ratio test** is a statistical test used to compare the fit of two models, one of which (the *null model*) is a special case of the other (the *alternative model*). The test is based on the likelihood ratio, which expresses how many times more likely the data are under one model than the other. This likelihood ratio, or equivalently its logarithm, can then be used to compute a p-value, or compared to a critical value to decide whether to reject the null model in favour of the alternative model. When the logarithm of the likelihood ratio is used, the statistic is known as a **log-likelihood ratio statistic**, and the probability distribution of this test statistic, assuming that the null model is true, can be approximated using **Wilks' theorem**.

In the case of distinguishing between two models, each of which has no unknown parameters, use of the likelihood ratio test can be justified by the Neyman–Pearson lemma, which demonstrates that such a test has the highest power among all competitors.<sup>[1]</sup>

## Contents

- 1 Use
- 2 Background
- 3 Simple-vs-simple hypotheses
- 4 Definition (likelihood ratio test for composite hypotheses)
  - 4.1 Interpretation
  - 4.2 Distribution: Wilks' theorem
- 5 Examples
  - 5.1 Coin tossing
- 6 References
- 7 External links

## Use

Each of the two competing models, the null model and the alternative model, is separately fitted to the data and the log-likelihood recorded. The test statistic (often denoted by  $D$ ) is twice the difference in these log-likelihoods:

$$\begin{aligned} D &= -2 \ln \left( \frac{\text{likelihood for null model}}{\text{likelihood for alternative model}} \right) \\ &= -2 \ln(\text{likelihood for null model}) + 2 \ln(\text{likelihood for alternative model}) \end{aligned}$$

The model with more parameters will always fit at least as well (have an equal or greater log-likelihood). Whether it fits significantly better and should thus be preferred is determined by deriving the probability or p-value of the difference  $D$ . Where the null hypothesis represents a special case of the alternative hypothesis, the probability distribution of the test statistic is approximately a chi-squared distribution with degrees of freedom equal to

$df2 - df1$ .<sup>[2]</sup> Symbols  $df1$  and  $df2$  represent the number of free parameters of models 1 and 2, the null model and the alternative model, respectively. The test requires nested models, that is: models in which the more complex one can be transformed into the simpler model by imposing a set of constraints on the parameters.<sup>[3]</sup>

For example: if the null model has 1 parameter and a log-likelihood of  $-8024$  and the alternative model has 3 parameters and a log-likelihood of  $-8012$ , then the probability of this difference is that of chi-squared value of  $+2 \cdot (8024 - 8012) = 24$  with  $3 - 1 = 2$  degrees of freedom. Certain assumptions<sup>[4]</sup> must be met for the statistic to follow a chi-squared distribution and often empirical p-values are computed.

## Background

The **likelihood ratio**, often denoted by  $\Lambda$  (the capital Greek letter lambda), is the ratio of the likelihood function varying the parameters over two different sets in the numerator and denominator. A **likelihood ratio test** is a statistical test for making a decision between two hypotheses based on the value of this ratio.

It is central to the Neyman–Pearson approach to statistical hypothesis testing, and, like statistical hypothesis testing in general, is both widely used and criticized.

## Simple-vs-simple hypotheses

A statistical model is often a parametrized family of probability density functions or probability mass functions  $f(x|\theta)$ . A simple-vs-simple hypotheses test has completely specified models under both the null and alternative hypotheses, which for convenience are written in terms of fixed values of a notional parameter  $\theta$ :

$$H_0 : \theta = \theta_0,$$

$$H_1 : \theta = \theta_1.$$

Note that under either hypothesis, the distribution of the data is fully specified; there are no unknown parameters to estimate. The likelihood ratio test statistic can be written as:<sup>[5][6]</sup>

$$\Lambda(x) = \frac{L(\theta_0|x)}{L(\theta_1|x)} = \frac{f(x|\theta_0)}{f(x|\theta_1)}$$

or

$$\Lambda(x) = \frac{L(\theta_0 | x)}{\sup\{ L(\theta | x) : \theta \in \{\theta_0, \theta_1\} \}},$$

where  $L(\theta|x)$  is the likelihood function, and  $\sup$  is the Supremum function. Note that some references may use the reciprocal as the definition.<sup>[7]</sup> In the form stated here, the likelihood ratio is small if the alternative model is better than the null model and the likelihood ratio test provides the decision rule as:

If  $\Lambda > c$ , do not reject  $H_0$ ;

If  $\Lambda < c$ , reject  $H_0$ ;

Reject with probability  $q$  if  $\Lambda = c$ .

The values  $c$ ,  $q$  are usually chosen to obtain a specified significance level  $\alpha$ , through the relation:  $q \cdot P(\Lambda = c | H_0) + P(\Lambda < c | H_0) = \alpha$ . The Neyman-Pearson lemma states that this likelihood ratio test is the most powerful among all level  $\alpha$  tests for this problem.<sup>[1]</sup>

## Definition (likelihood ratio test for composite hypotheses)

A null hypothesis is often stated by saying the parameter  $\theta$  is in a specified subset  $\Theta_0$  of the parameter space  $\Theta$ .

$$H_0 : \theta \in \Theta_0$$

$$H_1 : \theta \in \Theta_0^C$$

The likelihood function is  $L(\theta|x) = f(x|\theta)$  (with  $f(x|\theta)$  being the pdf or pmf), which is a function of the parameter  $\theta$  with  $x$  held fixed at the value that was actually observed, *i.e.*, the data. The **likelihood ratio test statistic** is<sup>[8]</sup>

$$\Lambda(x) = \frac{\sup\{L(\theta|x) : \theta \in \Theta_0\}}{\sup\{L(\theta|x) : \theta \in \Theta\}}.$$

Here, the **supp** notation refers to the supremum function.

A **likelihood ratio test** is any test with critical region (or rejection region) of the form  $\{x | \Lambda \leq c\}$  where  $c$  is any number satisfying  $0 \leq c \leq 1$ . Many common test statistics such as the  $Z$ -test, the  $F$ -test, Pearson's chi-squared test and the  $G$ -test are tests for nested models and can be phrased as log-likelihood ratios or approximations thereof.

## Interpretation

Being a function of the data  $x$ , the likelihood ratio is therefore a statistic. The **likelihood ratio test** rejects the null hypothesis if the value of this statistic is too small. How small is too small depends on the significance level of the test, *i.e.*, on what probability of Type I error is considered tolerable ("Type I" errors consist of the rejection of a null hypothesis that is true).

The numerator corresponds to the maximum likelihood of an observed outcome under the null hypothesis. The denominator corresponds to the maximum likelihood of an observed outcome varying parameters over the whole parameter space. The numerator of this ratio is less than the denominator. The likelihood ratio hence is between 0 and 1. Low values of the likelihood ratio mean that the observed result was less likely to occur under the null hypothesis as compared to the alternative. High values of the statistic mean that the observed outcome was nearly as likely to occur under the null hypothesis as compared to the alternative, and the null hypothesis cannot be rejected.

## Distribution: Wilks' theorem

If the distribution of the likelihood ratio corresponding to a particular null and alternative hypothesis can be explicitly determined then it can directly be used to form decision regions (to accept/reject the null hypothesis). In most cases, however, the exact distribution of the likelihood ratio corresponding to specific hypotheses is very difficult to determine. A convenient result, attributed to Samuel S. Wilks, says that as the sample size  $n$  approaches  $\infty$ , the test statistic  $-2 \log(\Lambda)$  for a nested model will be asymptotically  $\chi^2$ -distributed with

degrees of freedom equal to the difference in dimensionality of  $\Theta$  and  $\Theta_0$ .<sup>[4]</sup> This means that for a great variety of hypotheses, a practitioner can compute the likelihood ratio  $\Lambda$  for the data and compare  $-2 \log(\Lambda)$  to the  $\chi^2$  value corresponding to a desired statistical significance as an approximate statistical test.

## Examples

### Coin tossing

An example, in the case of Pearson's test, we might try to compare two coins to determine whether they have the same probability of coming up heads. Our observation can be put into a contingency table with rows corresponding to the coin and columns corresponding to heads or tails. The elements of the contingency table will be the number of times the coin for that row came up heads or tails. The contents of this table are our observation  $X$ .

	Heads	Tails
Coin 1	$k_{1H}$	$k_{1T}$
Coin 2	$k_{2H}$	$k_{2T}$

Here  $\Theta$  consists of the parameters  $p_{1H}, p_{1T}, p_{2H}$ , and  $p_{2T}$ , which are the probability that coins 1 and 2 come up heads or tails. The hypothesis space  $H$  is defined by the usual constraints on a distribution,  $0 \leq p_{ij} \leq 1$ , and  $p_{iH} + p_{iT} = 1$ . The null hypothesis  $H_0$  is the subspace where  $p_{1j} = p_{2j}$ . In all of these constraints,  $i = 1, 2$  and  $j = H, T$ .

Writing  $n_{ij}$  for the best values for  $p_{ij}$  under the hypothesis  $H$ , maximum likelihood is achieved with

$$n_{ij} = \frac{k_{ij}}{k_{iH} + k_{iT}}.$$

Writing  $m_{ij}$  for the best values for  $p_{ij}$  under the null hypothesis  $H_0$ , maximum likelihood is achieved with

$$m_{ij} = \frac{k_{1j} + k_{2j}}{k_{1H} + k_{2H} + k_{1T} + k_{2T}},$$

which does not depend on the coin  $i$ .

The hypothesis and null hypothesis can be rewritten slightly so that they satisfy the constraints for the logarithm of the likelihood ratio to have the desired nice distribution. Since the constraint causes the two-dimensional  $H$  to be reduced to the one-dimensional  $H_0$ , the asymptotic distribution for the test will be  $\chi^2(1)$ , the  $\chi^2$  distribution with one degree of freedom.

For the general contingency table, we can write the log-likelihood ratio statistic as

$$-2 \log \Lambda = 2 \sum_{i,j} k_{ij} \log \frac{n_{ij}}{m_{ij}}.$$

## References

1. ^ **a b** Neyman, Jerzy; Pearson, Egon S. (1933). "On the Problem of the Most Efficient Tests of Statistical Hypotheses". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **231** (694–706): 289–337. doi:10.1098/rsta.1933.0009 (<http://dx.doi.org/10.1098%2Frsta.1933.0009>). JSTOR 91247 (<https://www.jstor.org/stable/91247>).
2. ^ Huelsenbeck, J. P.; Crandall, K. A. (1997). "Phylogeny Estimation and Hypothesis Testing Using Maximum Likelihood". *Annual Review of Ecology and Systematics* **28**: 437–466. doi:10.1146/annurev.ecolsys.28.1.437 (<http://dx.doi.org/10.1146%2Fannurev.ecolsys.28.1.437>).
3. ^ An example using phylogenetic analyses is described at Huelsenbeck, J. P.; Hillis, D. M.; Nielsen, R. (1996). "A Likelihood-Ratio Test of Monophyly". *Systematic Biology* **45** (4): 546. doi:10.1093/sysbio/45.4.546 (<http://dx.doi.org/10.1093%2Fsysbio%2F45.4.546>).
4. ^ **a b** Wilks, S. S. (1938). "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses". *The Annals of Mathematical Statistics* **9**: 60–62. doi:10.1214/aoms/1177732360 (<http://dx.doi.org/10.1214%2Faoms%2F1177732360>).
5. ^ Mood, A.M.; Graybill, F.A. (1963) *Introduction to the Theory of Statistics*, 2nd edition. McGraw-Hill ISBN 978-0070428638 (page 286)
6. ^ Kendall, M.G., Stuart, A. (1973) *The Advanced Theory of Statistics, Volume 2*, Griffin. ISBN 0852642156 (page 234)
7. ^ Cox, D. R. and Hinkley, D. V *Theoretical Statistics*, Chapman and Hall, 1974. (page 92)
8. ^ Casella, George; Berger, Roger L. (2001) *Statistical Inference*, Second edition. ISBN 978-0534243128 (page 375)

## External links

- Practical application of likelihood ratio test described (<http://www.itl.nist.gov/div898/handbook/apr/section2/apr233.htm>)
- Richard Lowry's Predictive Values and Likelihood Ratios (<http://faculty.vassar.edu/lowry/clin2.html>) Online Clinical Calculator

Retrieved from "[http://en.wikipedia.org/w/index.php?title=Likelihood-ratio\\_test&oldid=612229509](http://en.wikipedia.org/w/index.php?title=Likelihood-ratio_test&oldid=612229509)"

Categories: Statistical ratios | Statistical tests | Statistical theory

- 
- This page was last modified on 9 June 2014 at 15:52.
  - Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.

# Statistics 512: Applied Linear Models

## Topic 5

### Topic Overview

This topic will cover

- Diagnostics and Remedial Measures
- Influential Observations and Outliers

### Chapter 9: Regression Diagnostics

We now have more complicated models. The ideas (especially with regard to the residuals) of Chapter 3 still apply, but we will also concern ourselves with the detection of outliers and influential data points. The following are often used for the identification of such points and can be easily obtained from SAS:

- Studentized deleted residuals
- Hat matrix diagonals
- Dffits, Cook's D, DFBETAS
- Variance inflation factor
- Tolerance

### Life Insurance Example

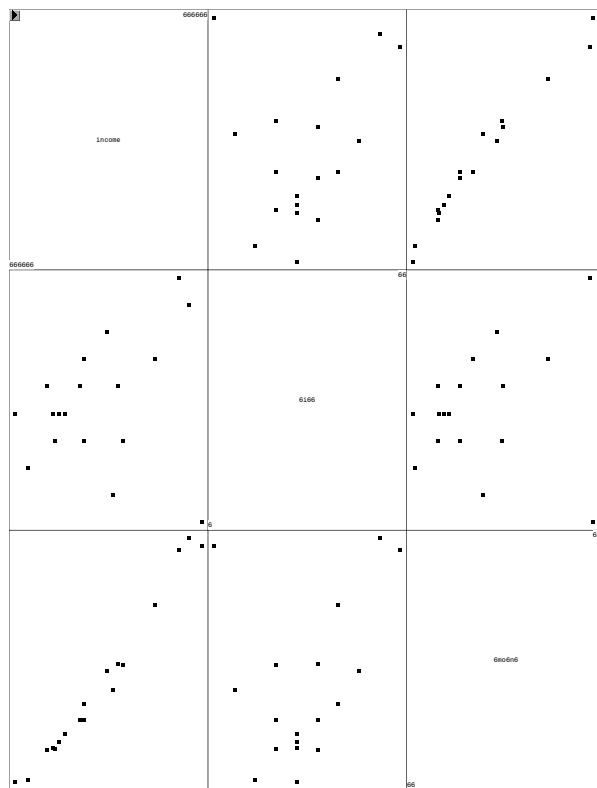
- We will use this as a running example in this topic.
- References: page 364 in NKNW and `nknw364.sas`.
- $Y$  = amount of insurance (in \$1000)
- $X_1$  = Average Annual Income (in \$1000)
- $X_2$  = Risk Aversion Score (0-10)
- $n = 18$  managers were surveyed.

```
data insurance;
infile 'H:\System\Desktop\Ch09ta01.dat';
input income risk amount;
proc reg data=insurance;
model amount=income risk/r influence;
```

Just to get oriented...

Analysis of Variance						Pr > F
Source	DF	Sum of Squares		Mean Square	F Value	
		Model	Error		<.0001	
Corrected Total	17	176324	2405.14763	160.34318		
Root MSE		12.66267	R-Square	0.9864		
Dependent Mean		134.44444	Adj R-Sq	0.9845		
Coeff Var		9.41851				
Parameter Estimates						
Variable	DF	Parameter Estimate		Standard Error		Pr >  t
		-205.71866		11.39268	-18.06	
income	1	6.28803		0.20415	30.80	<.0001
risk	1	4.73760		1.37808	3.44	0.0037

Model is significant and  $R^2 = 0.9864$  – quite high – both variables are significant.

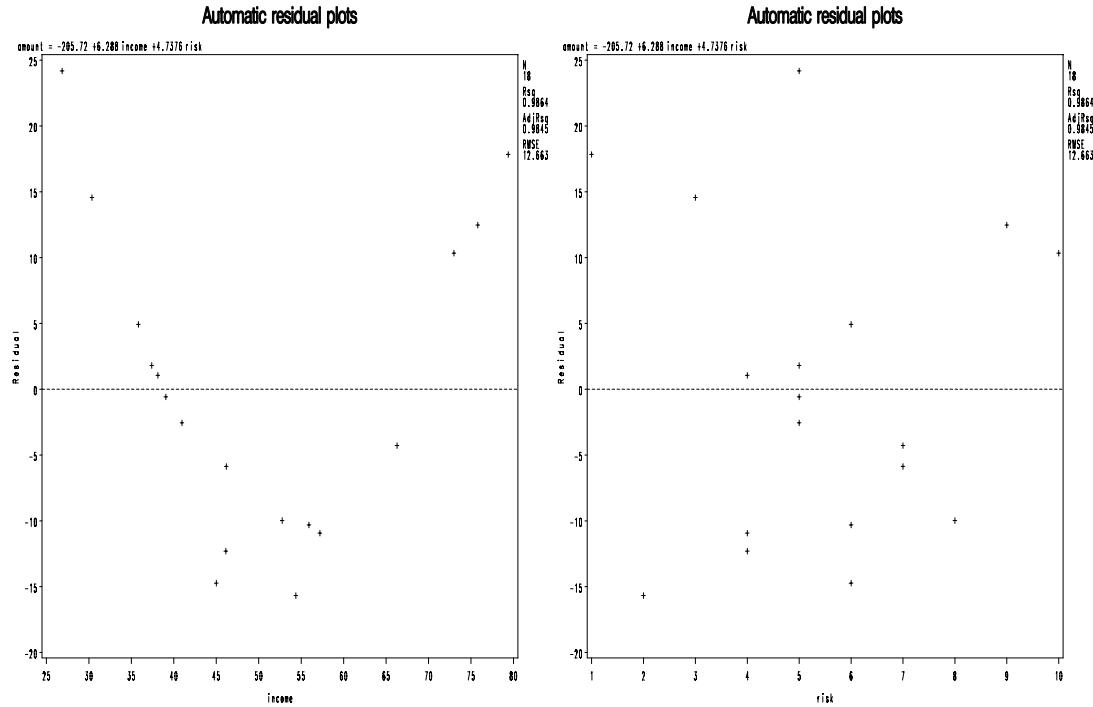


## The Usual Residual Plots

The `plot` statement generates the following two residual plots (in the past we have used `gplot` to create these). These residuals are for the full model. Note the weird syntax

`r.*(income risk)`. It prints the estimated equation and the  $R^2$  on it automatically, which is kind of nice. This is an alternative to saving the residuals and using `gplot`, although you have less control over the output.

```
title1 'Insurance';
proc reg data=insurance;
  model amount=income risk/r partial;
  plot r.*(income risk);
```



It looks like there is something quadratic going on with *income* in the full model. The residuals for *risk* look okay.  
(We should also do a qqplot.)

## Types of Residuals

### Regular Residuals

- $e_i = Y_i - \hat{Y}_i$  (the usual).
- These are given in the SAS output under the heading “Residual” when you use the `r` option in the `model` statement, and to store them use `r = (name)` in an `output` statement.

### Studentized Residuals

- $e_i^* = \frac{e_i}{\sqrt{MSE \times (1-h_{i,i})}}$

- *Studentized* means divided by its standard error. (When you ignore the  $h_{i,i}$  and just divide by Root MSE they are called *semistudentized residuals*.)
- Recall that  $s^2\{\mathbf{e}\} = \text{MSE}(\mathbf{I} - \mathbf{H})$ , so that  $s^2\{e_i\} = \text{MSE}(1 - h_{i,i})$ . These follow a  $t_{(n-p)}$  distribution if all assumptions are met.
- Studentized residuals are shown in the SAS output under the heading “Student Residual.” In the output, “Residual” / “Std Error Residual” = “Student Residual”. SAS also prints a little bar graph of the studentized residuals so you can identify large ones quickly.
- In general, values larger than about 3 should be investigated. (The actual cutoff depends on a  $t$  distribution and the sample size; see below.) These are computed using the ‘r’ option and can be stored using `student=(name)`.

## Studentized Deleted Residuals

- The idea: delete case  $i$  and refit the model. Compute the predicted value and residual for case  $i$  using this model. Compute the “studentized residual” for case  $i$ . (Don’t do this literally.)
- We use the notation  $(i)$  to indicate that case  $i$  has been deleted from the computations.
- $d_i = Y_i - \hat{Y}_{i(i)}$  is the deleted residual. (Also used for PRESS criterion)
- Interestingly, it can be calculated from the following formula without re-doing the regression with case  $i$  removed. It turns out that  $d_i = \frac{e_i}{(1-h_{i,i})}$ , where  $h_{i,i}$  is the  $i$ th diagonal element of the Hat matrix  $\mathbf{H}$ . Its estimated variance is  $s^2\{d_i\} = \frac{\text{MSE}_{(i)}}{(1-h_{i,i})}$ .
- The studentized deleted residual is  $t_i = \frac{d_i}{\sqrt{s^2\{d_i\}}} = \frac{e_i}{(1-h_{i,i})} \sqrt{\frac{(1-h_{i,i})}{\text{MSE}_{(i)}}} = \frac{e_i}{\sqrt{\text{MSE}_{(i)}(1-h_{i,i})}}$ .
- $\text{MSE}_{(i)}$  can be computed by solving this equation:  $(n-p)\text{MSE} = (n-p-1)\text{MSE}_{(i)} + \frac{e_i^2}{1-h_{i,i}}$ .
- The  $t_i$  are shown in the SAS output under the heading “Rstudent”, and the  $h_{i,i}$  under the heading “Hat Diag H”. To calculate these, use the `influence` option and to store them use `rstudent=(name)`.
- We can use these to test (using a Bonferroni correction for  $n$  tests) whether the case with the largest studentized residual is an outlier (see page 374).

```
proc reg data=insurance;
  model amount=income risk/r influence;
```

Obs	Dep Var	Output Statistics						Cook's D
		Std Error	Student	-2-1	0	1	2	
1	91.0000	-14.7311	12.216	-1.206	**			0.036
2	162.0000	-10.9321	12.009	-0.910	*			0.031
3	11.0000	24.1845	11.403	2.121		****		0.349
4	240.0000	-4.2780	11.800	-0.363				0.007
5	73.0000	-2.5522	12.175	-0.210				0.001
6	311.0000	10.3417	10.210	1.013		**		0.184
7	316.0000	17.8373	7.780	2.293		****		2.889
8	154.0000	-9.9763	11.798	-0.846	*			0.036
9	164.0000	-10.3084	12.239	-0.842	*			0.017
10	54.0000	1.0560	12.009	0.0879				0.000
11	53.0000	4.9301	11.878	0.415				0.008
12	326.0000	12.4728	10.599	1.177		**		0.197
13	55.0000	1.8081	12.050	0.150				0.001
14	130.0000	-15.6744	11.258	-1.392	**			0.171
15	112.0000	-5.8634	12.042	-0.487				0.008
16	91.0000	-12.2985	12.162	-1.011	**			0.029
17	14.0000	14.5636	11.454	1.271		**		0.120
18	63.0000	-0.5798	12.114	-0.0479				0.000

### Test for Outliers Using Studentized Deleted Residuals

- should use the Bonferroni correction since you are looking at all  $n$  residuals
- studentized deleted residuals follow a  $t_{(n-p-1)}$  distribution since they are based on  $n-1$  observations
- If a studentized deleted residual is bigger in magnitude than  $t_{n-p-1}(1 - \frac{\alpha}{2n})$  then we identify the case as a possible outlier based on this test.
- In our example, take  $\alpha = 0.05$ . Since  $n = 18$  and  $p = 3$ , we use  $t_{14}(0.9986) \approx 3.6214$ .
- None of the observations may be called an outlier based on this test.
- Note that if we neglected to use the Bonferroni correction our cutoff would be 2.1448 which would detect obs. 3 and 7, but this would not be correct.
- Note that “identifying an outlier” does not mean that you then automatically remove the observation. It just means you should take a closer look at that observation and check for reasons why it should possibly be removed. It could also mean that you have problems with normality and/or constant variance in your dataset and should consider a transformation.

### What to Look For

When we examine the residuals we are looking for

- Outliers

- Non-normal error distributions
- Influential observations

## Other Measures of Influential Observations

The `influence` option calculates a number of other quantities. We won't spend a whole lot of time on these, but you might be wondering what they are.

Obs	Cook's Hat Diag		DFFITS	Output Statistics		
	D	H		Intercept	income	risk
1	0.036	0.0693	-0.3345	-0.1179	0.1245	-0.1107
2	0.031	0.1006	-0.3027	-0.0395	-0.1470	0.1723
3	0.349	0.1890	1.1821	0.9594	-0.9871	0.1436
4	0.007	0.1316	-0.1369	0.0770	-0.0821	-0.0410
5	0.001	0.0756	-0.0580	-0.0394	0.0286	0.0011
6	0.184	0.3499	0.7437	-0.5298	0.3048	0.5125
7	2.889	0.6225	3.5292	-0.3649	2.6598	-2.6751
8	0.036	0.1319	-0.3263	0.0816	0.0254	-0.2452
9	0.017	0.0658	-0.2212	0.0308	-0.0672	-0.0366
10	0.000	0.1005	0.0284	0.0238	-0.0138	-0.0092
11	0.008	0.1201	0.1490	0.0863	-0.1057	0.0536
12	0.197	0.2994	0.7801	-0.5820	0.4495	0.4096
13	0.001	0.0944	0.0468	0.0348	-0.0294	0.0014
14	0.171	0.2096	-0.7423	-0.2706	-0.2656	0.6269
15	0.008	0.0957	-0.1543	-0.0164	0.0532	-0.0953
16	0.029	0.0775	-0.2934	-0.1810	0.0258	0.1424
17	0.120	0.1818	0.6129	0.5803	-0.3608	-0.2577
18	0.000	0.0849	-0.0141	-0.0101	0.0080	-0.0001
*	0.826	0.3333	0.8165	1 (or 0.4714)		

## Cook's Distance

- This measures the influence of case  $i$  on all of the  $\hat{Y}_i$ 's. It is a standardized version of the sum of squares of the differences between the predicted values computed with and without case  $i$ .
- Large values suggest an observation has a lot of influence. Cook's D values are obtained via the '`r`' option in the `model` statement and can be stored with `cookd=(name)`.
- here "large" means larger than the 50th percentile of the  $F_{p,n-p}$  distribution; for our example  $F_{3,15}(0.5) = 0.826$  .

## Hat Matrix Diagonals

- $h_{i,i}$  is a measure of how much  $Y_i$  is contributing to the prediction of  $\hat{Y}_i$ . This depends on the distance between the  $X$  values for the  $i$ th case and the means of the  $X$  values. Observations with extreme values for the predictors will have more influence.

- $h_{i,i}$  is sometimes called the *leverage* of the  $i$ th observation. It always holds that  $0 \leq h_{i,i} \leq 1$  and  $\sum h_{i,i} = p$ .
- A large value of  $h_{i,i}$  suggests that the  $i$ th case is distant from the center of all  $X$ 's. The average value is  $p/n$ . Values far from this average (say, twice as large) point to cases that should be examined carefully because they may have a substantial influence on the regression parameters.
- For our example,  $\frac{2p}{n} = \frac{6}{18} = 0.333$  so values larger than 0.333 would be considered large. Observations #6, #7, and maybe #12 seem to have a lot of influence. These can be further examined with the next set of influence statistics.
- The hat matrix diagonals are displayed with the `influence` option and can be stored with `h=(name)` .

#### DEFITS

- Another measure of the influence of case  $i$  on its own fitted value  $\hat{Y}_i$ . It is a standardized version of the difference between  $\hat{Y}_i$  computed with and without case  $i$ . It is closely related to  $h_{i,i}$  (consult the text for formula if you are interested). Values larger than 1 (for small to medium size datasets) or  $2\sqrt{\frac{p}{n}}$  (for large datasets) are considered influential. (In our example,  $2\sqrt{\frac{p}{n}} = 0.816$  but this is a small dataset so we would use 1).
- these are calculated with the `influence` option and can be stored with `dffits=(name)`.

#### DFBETAS

- A measure of the influence of case  $i$  on each of the regression coefficients.
- It is a standardized version of the difference between the regression coefficient computed with and without case  $i$ .
- Values larger than 1 (for small-to-medium datasets) or  $\frac{2}{\sqrt{n}}$  (for large datasets) are considered influential. In this example  $\frac{2}{\sqrt{n}} = 0.4714$ , but we would use 1 as a cutoff.
- According to all these measures, observation #7 appears to be influential. This is not surprising because it has the smallest risk (1) and the highest income (79.380) of all the observations.

## Measures of Multicollinearity

We already know about several identifying factors in dealing with multicollinearity:

- regression coefficients change greatly when predictors are included/excluded from the model
- significant  $F$ -test but *no* significant  $t$ -tests for  $\beta$ 's (ignoring intercept)

- regression coefficients that don't "make sense", i.e. don't match scatterplot and/or intuition
- Type I and II  $SS$  very different
- predictors that have pairwise correlations

There are two other numerical measures that can be used: `vif` and `tolerance`

### Variance Inflation Factor

- The VIF is related to the variance of the estimated regression coefficients.
- $VIF_k = \frac{1}{1-R_k^2}$  where  $R_k^2$  is the squared multiple correlation obtained in a regression where all other explanatory variables are used to predict  $X_k$ . We calculate it for each explanatory variable.
- If this  $R_k^2$  is large that means  $X_k$  is well predicted by the other  $X$ 's. One suggested rule is that a value of 10 or more for VIF indicates excessive multicollinearity. This corresponds to an  $R_k^2$  of  $\geq 0.9$ . Use the `vif` option to the `model` statement.

### Tolerance

- $TOL = 1 - R_k^2 = \frac{1}{VIF}$ . A tolerance of  $< 0.1$  is the same as a  $VIF > 10$ , indicating excessive multicollinearity. Use the `TOL` option to the `model` statement. Described in comment on p 388.

Typically you would look at either `vif` or `tol`, not both.

```
proc reg data=insurance;
  model amount=income risk/tol vif;

                                Parameter Estimates
Variable      Tolerance      Inflation
Intercept          .              0
income            0.93524     1.06925
risk              0.93524     1.06925
```

These values are quite acceptable.

### Partial Regression Plots

- Also called partial residual plots, added variable plots or adjusted variable plots.
- Related to partial correlations, they help you figure out the net effect of  $X_i$  on  $Y$ , given that other variables are in the model.

- One plot for each  $X_i$ . To get the plot, run two regressions. In the first, use the other  $X$ 's to predict  $Y$ . In the second use the other  $X$ 's to predict  $X_i$ . Then plot the residuals from the first regression against the residuals from the second regression. The correlation of these residuals was called the *partial correlation coefficient*.
- A linear pattern in this type of plot indicates that the variable would be useful in the model, and the slope is its regression coefficient. The plots shows the strength of a marginal relationship between  $Y$  and  $X_i$  in the full model. If the partial residual plot for  $X_i$  appears “flat”,  $X_i$  may not need to be included in the model. If they appear like a straight line (with non-zero slope), then that suggests  $X_i$  should be included as a linear term, etc.
- Nonlinear relationships, heterogeneous variances, and outliers may also be detected in these plots.
- In SAS, the ‘partial’ option in the `model` statement can be used to get a partial residual plot. This is not a very good plot (useful for first glance, but not something you would want to publish), so it is useful to know how to create a better one.

Coding for the poor resolution plot (they're kind of ugly):

```
proc reg data=insurance;
  model amount=income risk/r partial;
```

(The number labels on the plot are the first digit of income because we said “`id income`”.) The axes are labelled `amount` and `income`, but we are actually plotting the residuals for `amount` (predicted by `risk`) vs. the residuals for `income` (when predicted by `risk`)

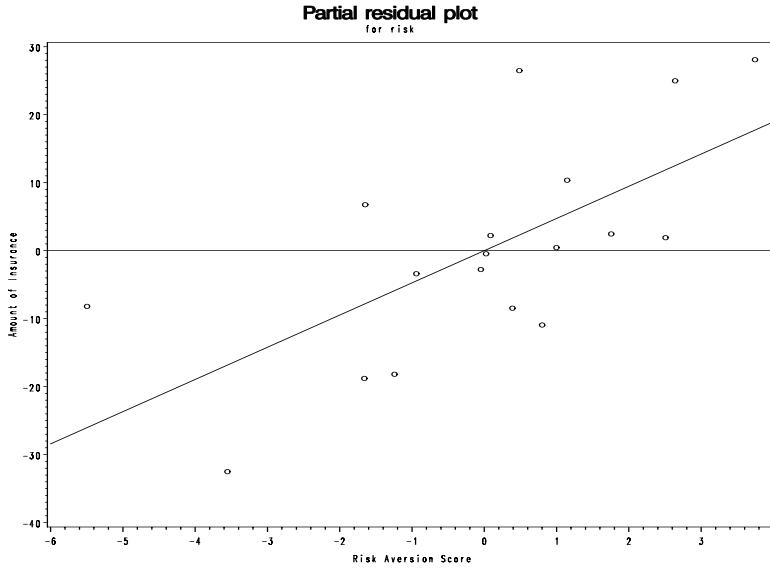
(The number labels on the plot are the first digit of `income` because we said “`id income`”.)

## Obtaining Partial Regression Plots

```
title1 'Partial residual plot';
title2 'for risk';
symbol1 v=circle i=rl;
axis1 label=('Risk Aversion Score');
axis2 label=(angle=90 'Amount of Insurance');
proc reg data=insurance;
  model amount risk = income;
  output out=partialrisk r=resamt resrisk;
proc gplot data=partialrisk;
  plot resamt*resrisk / haxis=axis1 vaxis=axis2 vref = 0;
run;
```

The  $y$ -axis has the residuals for the model `insur = income`. The  $x$ -axis has the residuals for the model `risk = income` (i.e. treat `risk` as a  $Y$ -variable).

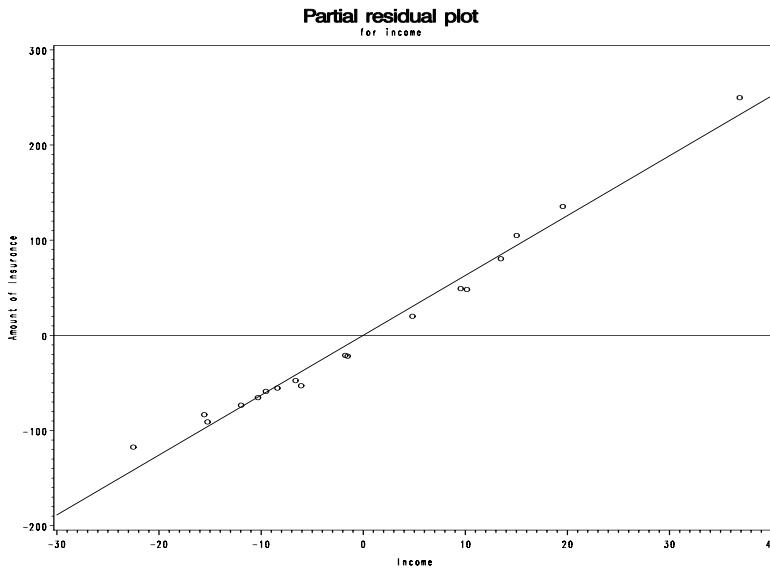
The residuals compared to the horizontal line are the residuals for the model that omits `risk` as a variable. The residuals compared to the “regression” line are the residuals for the



model that includes *risk* as a variable. Are the points closer to the regression line than to the x-axis? This helps decide if there is much to be gained (i.e. smaller residuals) by including *risk* in the model. In this case *risk* clearly should be included.

Similar code for *income*:

```
axis3 label=('Income');
title2 'for income';
proc reg data=insurance;
model amount income = risk;
      output out=partialincome r=resamt resinc;
proc gplot data=partialincome;
plot resamt*resinc / haxis=axis3 vaxis=axis2 vref = 0;
```



The resulting plot has on the y-axis the residuals for the model `insur = risk`, and the x-axis has the residuals for the model `income = risk`. This is the same as the text plot.

This plot shows, first of all, that *income* is clearly needed in the model. Secondly, we can see that the effect of *income* (when *risk* is included) is *mostly* linear. Third, a close look shows that the residuals curve a bit around the straight line, so that there is a quadratic effect. However, the quadratic effect is small compared to the linear one. A quadratic term will improve the fit of the model, but it may not improve it *much*. We would have to weigh the improved fit vs. the interpretability and possible multicollinearity problems when deciding on the final model.

Here's what happens when we include the square of (centered) income:

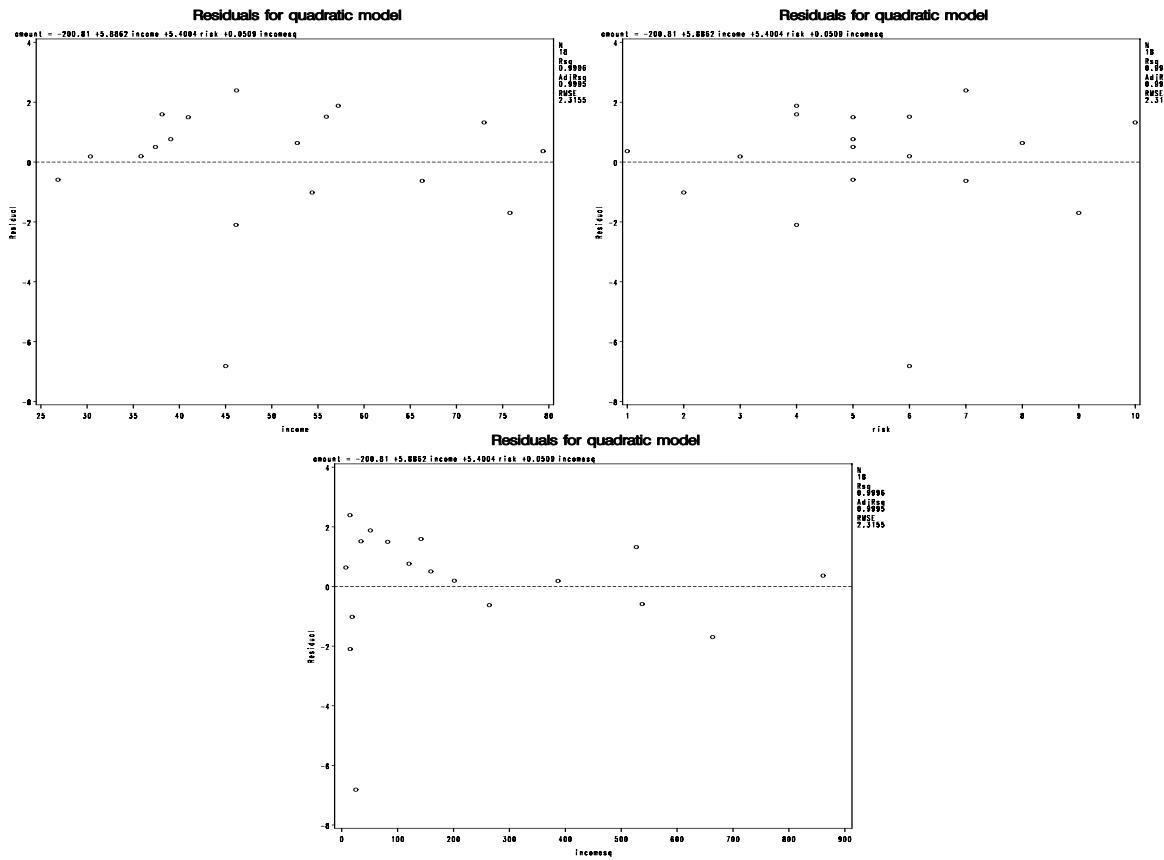
```
data quad;
  set insurance;
  sinc = income;
proc standard data=quad out=quad mean=0;
  var sinc;
data quad;
  set quad;
  incomesq = sinc*sinc;
title1 'Residuals for quadratic model';
proc reg data=quad;
  model amount = income risk incomesq / r vif;
  plot r.*(income risk incomesq);
```

Analysis of Variance					
	DF	Sum of Squares	Mean Square	F Value	Pr > F
Source					
Model	3	176249	58750	10958.0	<.0001
Error	14	75.05895	5.36135		
Corrected Total	17	176324			

Root MSE	2.31546	R-Square	0.9996
Dependent Mean	134.44444	Adj R-Sq	0.9995
Coeff Var	1.72224		

Parameter Estimates						
	Parameter	Standard Estimate	Error	t Value	Pr >  t	Variance Inflation
Variable	DF					
Intercept	1	-200.81134	2.09649	-95.78	<.0001	0
income	1	5.88625	0.04201	140.11	<.0001	1.35424
risk	1	5.40039	0.25399	21.26	<.0001	1.08627
incomesq	1	0.05087	0.00244	20.85	<.0001	1.26657

For the two-variable model,  $R^2$  was 0.9864, so while this is an improvement, it does not make a big difference. Our assumptions are now more closely met, which is good, but it also appears an outlier now exists where it did not before.



## Regression Diagnostics Summary

Check normality of the residuals with a normal quantile plot.

Plot the residuals versus predicted values, versus each of the  $X$ 's and (when appropriate) versus time

Examine the partial regression plots for each  $X$  variable.

Examine

- the studentized deleted residuals (RSTUDENT in the output)
- The hat matrix diagonals
- Dffits, Cook's D, and the DFBETAS
- Check observations that are extreme on these measures relative to the other observations
- Examine the tolerance or VIF for each  $X$

If there are variables with low tolerance / high VIF, or if any of the other indications of multicollinearity problems are present, you may need to do some model building:

- Recode variables
- Variable selection

# Remedial Measures (Chapter 10)

- Weighted Regression
- Robust Regression
- Nonparametric Regression
- Bootstrapping

## Weighted Regression

### Maximum Likelihood

$$\begin{aligned}
 Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i, \quad \text{Var}(\epsilon_i) = \sigma_i^2 \\
 Y_i &\sim N(\beta_0 + \beta_1 X_i, \sigma_i^2) \\
 f_i &= \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma_i}\right)^2} \\
 L &= f_1 \times f_2 \times \cdots \times f_n - \text{likelihood function}
 \end{aligned}$$

- Variance is no longer constant
- Maximization of  $L$  with respect to  $\beta$ 's.
- Equivalent to minimization of  $\sum \frac{1}{\sigma_i^2} (Y_i - \beta_0 - \beta_1 X_{i,1} - \dots - \beta_{p-1} X_{i,p-1})^2$

## Weighted Least Squares

- Used to deal with unequal variances:

$$\sigma^2\{\epsilon\} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

- Least squares minimizes the sum of the squared residuals. For WLS, we minimize instead the sum of the squared residuals each multiplied by an appropriate weight. If the error variances are known, the weights are  $w_i = 1/\sigma_i^2$ .
- Otherwise the variances need to be estimated (see discussion pages 403-405).
- The regression coefficients with weights are:  $\mathbf{bw} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{W}\mathbf{Y})$  where  $\mathbf{W}$  is a diagonal matrix of weights.
- In SAS, use a 'weight' statement in PROC REG.

## Drawbacks to Weighted Least Squares

No clear interpretation for  $MSE$ .  $MSE$  will be close to 1 if error variance is modeled well.

## Advantages to Weighted Least Squares

Improved parameter estimates, and CI's. Valid inference in presence of heteroscedasticity.

## Determining the Weights

We try to find a relationship between the absolute residual and another variable and use this as a model for the standard deviation; or similarly for the squared residual and the variance. Sometimes it is necessary to use grouped data or approximately grouped data to estimate the variance. With a model for the standard deviation or the variance, we can approximate the optimal weights. Optimal weights are proportional to the inverse of the variance as shown above. If the data have many observations for each value of  $X$  we can get a variance estimate at each value (this happens frequently in ANOVA).

## NKNW Example

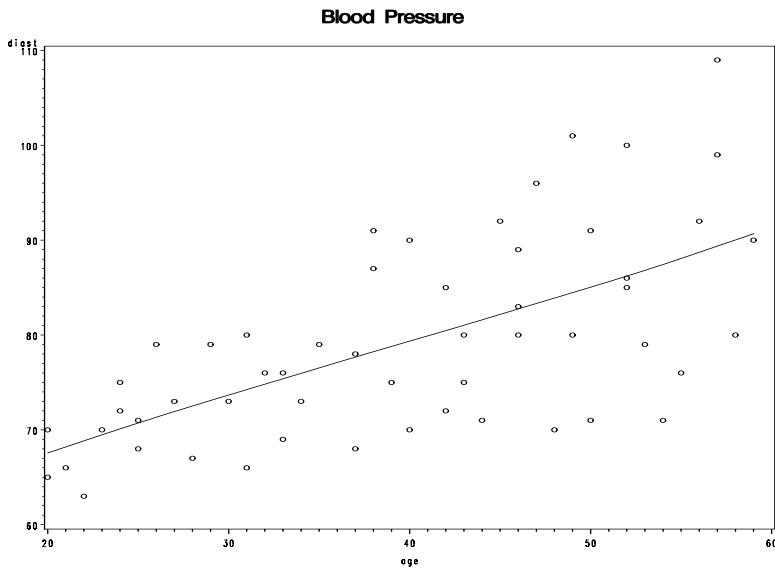
- NKNW p 406 (`nknw406.sas`)
- $Y$  is diastolic blood pressure
- $X$  is age
- $n = 54$  healthy adult women aged 20 to 60 years old

```
data pressure;
  infile 'H:\System\Desktop\Ch10ta01.dat';
  input age diast;
proc print data=pressure;
title1 'Blood Pressure';
symbol1 v=circle i=sm70;
proc sort data=pressure;
  by age;
proc gplot data=pressure;
  plot diast*age;
```

This clearly has non-constant variance. Run the (unweighted) regression to get residuals.

```
proc reg data=pressure;
  model diast=age / clb;
  output out=diag r=resid;
```

Source	Analysis of Variance				
	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2374.96833	2374.96833	35.79	<.0001



Error	52	3450.36501	66.35317
Corrected Total	53	5825.33333	

Root MSE	8.14575	R-Square	0.4077
Dependent Mean	79.11111	Adj R-Sq	0.3963
Coeff Var	10.29659		

Variable	DF	Parameter	Parameter Estimates				
			Standard	Error	t Value	Pr >  t	95% Confidence Limits
Intercept	1	56.15693	3.99367	14.06	<.0001	48.14304	64.17082
age	1	0.58003	0.09695	5.98	<.0001	0.38548	0.77458

Use the output data set to get the absolute and squared residuals. Plot each of them (vs.  $X$ ) with a smoother.

```

data diag;
set diag;
absr=abs(resid);
sqrr=resid*resid;

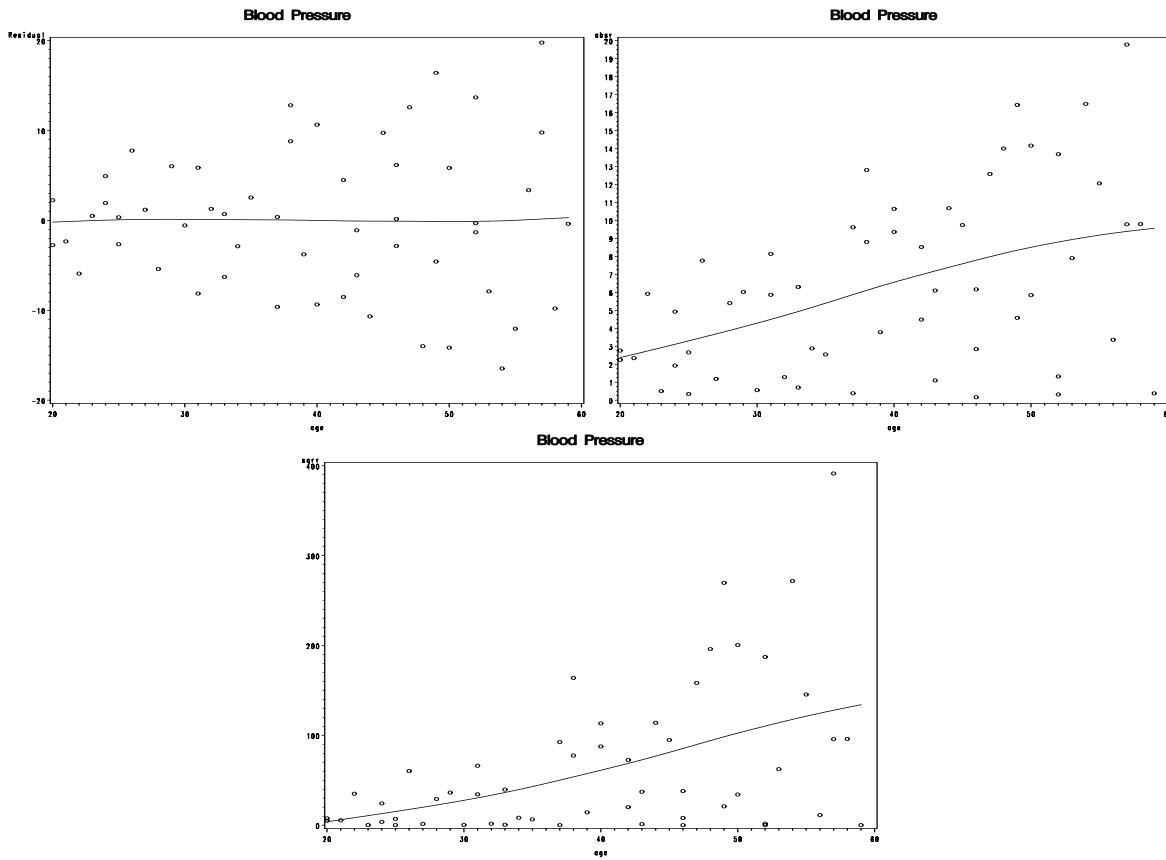
proc gplot data=diag;
plot (resid absr sqrr)*age;

```

The absolute value of the residuals appears to have a fairly linear relationship with  $age$  (it appears more linear than does the graph of squared residuals vs.  $age$ ). Thus, we will model standard deviation as a linear function of  $age$ . (If the second graph was more linear we would model variance instead.) We will model the absolute residuals as a function of  $age$ , and use the predicted values of that regression as weights.

Predict the standard deviation (absolute value of the residual):

```
proc reg data=diag;
```



```

model absr=age;
output out=findweights p=shat;
data findweights;
  set findweights;
  wt=1/(shat*shat);

```

We always compute the weights as the reciprocal of the estimated variance. Regression with weights:

```

proc reg data=findweights;
  model diast=age / clb p;
  weight wt;
  output out = weighted p = predict;

```

Analysis of Variance						
Source	DF	Sum of		Mean Square	F Value	Pr > F
		Squares	Square			
Model	1	83.34082	83.34082	83.34082	56.64	<.0001
Error	52	76.51351	1.47141			
Corrected Total	53	159.85432				
Root MSE		1.21302	R-Square	0.5214		
Dependent Mean		73.55134	Adj R-Sq	0.5122		
Coeff Var		1.64921				

Variable	DF	Parameter	Parameter Estimates					
			Standard Estimate	Error	t Value	Pr >  t	95% Confidence	Limits
Intercept	1	55.56577	2.52092	22.04	<.0001	50.50718	60.62436	
age	1	0.59634	0.07924	7.53	<.0001	0.43734	0.75534	

## Other Methods

### Robust Regression

- Basic idea is to have a procedure that is not sensitive to outliers.
- Alternatives to least squares, minimize either the sum of absolute values of residuals or the median of the squares of residuals.
- Do weighted regression with weights based on residuals, and iterate.
- See Section 10.3 for details.

### Nonparametric Regression

- Several versions
- We have used e.g. `i=sm70`
- Interesting theory
- All versions have some smoothing parameter similar to the 70 in `i=sm70`.
- Confidence intervals and significance tests not fully developed.

### Bootstrap

- Very important theoretical development that has had a major impact on applied statistics
- Based on simulation
- Sample *with* replacement from the data or residuals and get the distribution of the quantity of interest
- CI usually based on quantiles of the sampling distribution

### Model Validation

Three approaches to checking the validity of the model.

- Collect new data: does it fit the model?
- Compare with theory, other data, simulation.
- Use some of the data for the basic analysis (“training set”) and some for validity check.

# Qualitative Explanatory Variables (Chapter 11)

Example include

- Gender as an explanatory variable
- Placebo versus treatment
- Insurance Co. example from previous notes (Type of company)

## Two Categories

Recall from Topic 4 (General Linear Tests):

- Model:  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$
- When  $X_1 = 0$ ,  $\beta_1$  and  $\beta_3$  terms disappear:  $Y = \beta_0 + \beta_2 X_2 + \epsilon$ . For this group,  $\beta_0$  is the intercept, and  $\beta_2$  is the slope.
- When  $X_1 = 1$ ,  $\beta_1$  and  $\beta_3$  terms are incorporated into the intercept and  $X_2$  coefficient:

$$Y = (\beta_0 + \beta_1) + (\beta_2 + \beta_3)X_2 + \epsilon$$

- For this group,  $\beta_0 + \beta_1$  is the intercept, and  $\beta_2 + \beta_3$  is the slope.
- $H_0 : \beta_1 = \beta_3 = 0$  is the hypothesis that the regression lines are the same.
- $H_0 : \beta_1 = 0$  hypothesizes the two intercepts are equal.
- $H_0 : \beta_3 = 0$  hypothesizes the two slopes are equal.

## More Complicated Models

- If a categorical (qualitative) variable has  $k$  possible values we need  $k - 1$  indicator variables in order to describe it.
- These can be defined in many different ways; we will do this in Chapter 16 (ANOVA).
- We also can have several categorical explanatory variables, plus interactions, etc.
- Example: Suppose we have a variable *speed* for which 3 levels (high, medium, low) are possible. Then we would need two indicator variables (e.g.  $X_1 = \text{medium}$  and  $X_2 = \text{high}$ ) to describe the situation.

speed	$X_1$	$X_2$
low	0	0
medium	1	0
high	0	1

## Piecewise Linear Model

At some (known) point or points, the slope of the relationship changes. We can describe such a model with indicator variables.

Examples:

- tax brackets
- discount prices for bulk quantities
- overtime wages

### Piecewise Linear Model Example

- NKNW page 476 (`nknw476.sas`)
- $Y$  = unit cost,  $X_1$  = lot size,  $n = 8$
- We have reason to believe that a linear model is appropriate, but a slope change should be allowed at  $X_1 = 500$ . (Note the ‘bending’ in the plot.)
- We can do this by including an indicator variable  $X_2$  that is 1 if  $X_1$  is bigger than 500 and 0 otherwise and allowing it to interact with  $X_1$ .

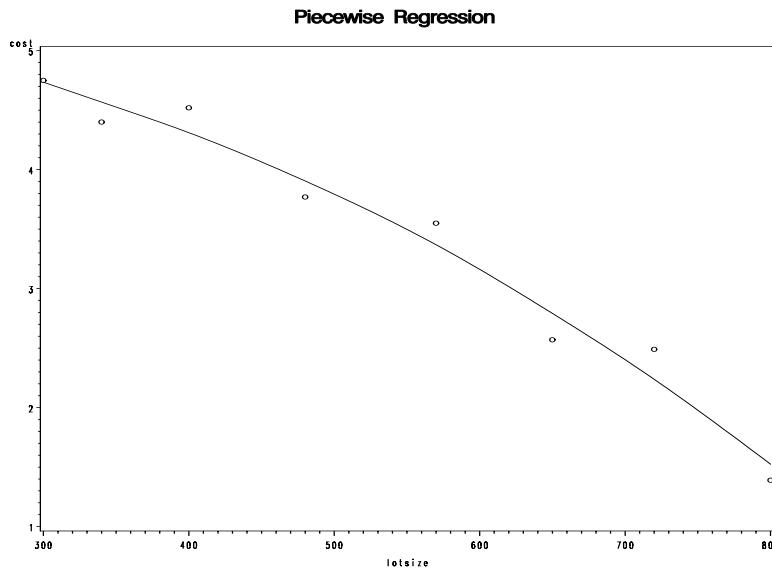
```
data piecewise;
  infile 'H:\System\Desktop\Ch11ta06.dat';
  input cost lotsize;
  symbol1 v=circle i=sm70 c=black;
  proc sort data=piecewise; by lotsize;
  proc gplot data=piecewise;
    plot cost*lotsize;
```

## Piecewise Model

Define a new variable  $X_2$  which is 0 when  $X_1 \leq 500$  and 1 when  $X_1 > 500$ . Then create an adjusted interaction term  $X_3 = X_2(X_1 - 500)$ . This uses  $-500X_2$  to indicate the change in intercept and the product  $X_1X_2$  to find the change in slope. Note that there is only one parameter since the two lines must join at  $X_1 = 500$ . We will not use  $X_2$  explicitly in the model, just the interaction term  $X_3$ . Thus the model is

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \epsilon \\ &= \beta_0 + \beta_1 X_1 + \beta_2 X_2(X_1 - 500) + \epsilon \\ &= \beta_0 - 500\beta_2 X_2 + \beta_1 X_1 + \beta_2 X_1 X_2 + \epsilon \\ &= \begin{cases} \beta_0 + \beta_1 X_1 & X_2 = 0 \quad (X_1 \leq 500) \\ (\beta_0 - 500\beta_2) + (\beta_1 + \beta_2)X_1 & X_2 = 1 \quad (X_1 > 500) \end{cases} \end{aligned}$$

Our model has



- An intercept ( $\beta_0$ )
- A coefficient for lot size (the slope  $\beta_1$ )
- An additional explanatory variable that will add a constant to the slope whenever lot size is greater than 500.

```
data piecewise; set piecewise;
  if lotsize le 500
    then cslope=0;
  if lotsize gt 500
    then cslope=lotsize-500;
proc print data=piecewise;
```

Obs	cost	lotsize	cslope
1	4.75	300	0
2	4.40	340	0
3	4.52	400	0
4	3.77	480	0
5	3.55	570	70
6	2.57	650	150
7	2.49	720	220
8	1.39	800	300

The variable `cslope` is our  $X_3$ . Run the regression:

```
proc reg data=piecewise;
  model cost=lotsize cslope;
  output out=pieceout p=costhat;
```

Source	DF	Analysis of Variance		F Value	Pr > F
		Sum of Squares	Mean Square		

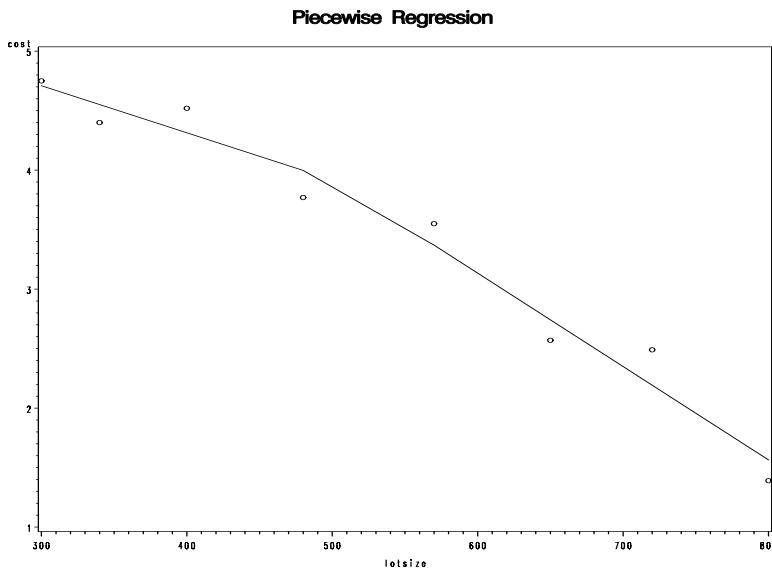
Model	2	9.48623	4.74311	79.06	0.0002
Error	5	0.29997	0.05999		
Corrected Total	7	9.78620			
Root MSE		0.24494	R-Square	0.9693	
Dependent Mean		3.43000	Adj R-Sq	0.9571	
Coeff Var		7.14106			
Parameter Estimates					
Parameter Standard					
Variable	DF	Estimate	Error	t Value	Pr >  t
Intercept	1	5.89545	0.60421	9.76	0.0002
lotsize	1	-0.00395	0.00149	-2.65	0.0454
cslope	1	-0.00389	0.00231	-1.69	0.1528

Plot data with fitted values:

```

symbol1 v=circle i=none c=black;
symbol2 v=none i=join c=black;
proc sort data=pieceout; by lotsize;
proc gplot data=pieceout;
  plot (cost costhat)*lotsize/overlay;

```



# Likelihood-ratio test

---

## related topics

{math, number, function}  
 {rate, high, increase}  
 {game, team, player}  
 {theory, work, human}  
 {disease, patient, cell}  
 {album, band, music}

In statistics, a **likelihood ratio test** is used to compare the fit of two models, one of which is nested within the other. This often occurs when testing whether a simplifying assumption for a model is valid, as when two or more model parameters are assumed to be related.

Both models are fitted to the data and their log-likelihood recorded. The test statistic (usually denoted  $D$ ) is twice the difference in these log-likelihoods:

The model with more parameters will always fit at least as well (have a greater log-likelihood). Whether it fits significantly better and should thus be preferred can be determined by deriving the probability or p-value of the obtained difference  $D$ . In many cases, the probability distribution of the test statistic can be approximated by a chi-square distribution with  $(df_1 - df_2)$  degrees of freedom, where  $df_1$  and  $df_2$  are the degrees of freedom of models 1 and 2

## related documents

Average  
 Estimator  
 Conditional probability  
 Reinforcement learning  
 Poisson distribution  
 Interpolation search  
 Mersenne twister  
 Golden ratio base  
 Linear  
 Monotonic function  
 Automata theory  
 Borel algebra  
 Graftal  
 Special linear group  
 Monotone convergence theorem  
 Generating trigonometric tables

respectively.

The test requires nested models, that is, models in which the more complex one can be transformed into the simpler model by imposing a set of linear constraints on the parameters.

In a concrete case, if model 1 has 1 free parameter and a log-likelihood of 8012 and the alternative model has 3 degrees of freedom and a LL of 8024, then the probability of this difference is that of chi-square of  $24 = 2 \cdot (8024 - 8012)$  under  $2 = 3 - 1$  degrees of freedom. Certain assumptions must be met for the statistic to follow a chi-squared distribution and often empirical p-values are computed.

### **Contents**

- [1 Background](#)
- [2 Simple-versus-simple hypotheses](#)
- [3 Definition \(likelihood ratio test for composite hypotheses\)](#)
  - [3.1 Interpretation](#)
  - [3.2 Approximation](#)
- [4 Examples](#)
  - [4.1 Coin tossing](#)
- [5 Criticism](#)
- [6 See also](#)
  - [6.1 Context](#)
- [7 References](#)
- [8 External links](#)

## **Background**

De Morgan's laws  
Column space  
Discriminant  
Controllability  
Isomorphism theorem  
NaN

Generalized Riemann hypothesis  
Stirling's approximation  
Conjunctive normal form  
Sylow theorems  
Jacobi symbol  
Free variables and bound variables  
Laurent series

Kolmogorov space

The **likelihood ratio**, often denoted by  $\Lambda$  (the capital [Greek letter lambda](#)), is the ratio of the [likelihood function](#) varying the parameters over two different sets in the numerator and denominator. A **likelihood-ratio test** is a statistical test for making a decision between two hypotheses based on the value of this ratio.

It is central to the [Neyman–Pearson](#) approach to statistical hypothesis testing, and, like statistical hypothesis testing generally, is both widely used and much criticized; see [Criticism](#), below.

## Simple-versus-simple hypotheses

[Full article ▶](#)

---

The article content of this page came from Wikipedia and is governed by CC-BY-SA.

## Likelihood Ratio Tests

Likelihood ratio tests (LRTs) have been used to compare two **nested** models. The form of the test is suggested by its name,

$$\text{LRT} = -2 \log_e \left( \frac{\mathcal{L}_s(\hat{\theta})}{\mathcal{L}_g(\hat{\theta})} \right),$$

the ratio of two likelihood functions; the simpler model ( $s$ ) has fewer parameters than the general ( $g$ ) model. Asymptotically, the test statistic is distributed as a chi-squared random variable, with degrees of freedom equal to the difference in the number of parameters between the two models.

Likelihood ratio tests compare two models provided the simpler model is a special case of the more complex model (i.e., “nested”). LRTs can be presented as a difference in the log-likelihoods (recall that  $\log(A/B) = \log A - \log B$ ) and this is often handy as they can be expressed in terms of deviance. Then,

$$\begin{aligned}\text{LRT} &= -2 \left( \log_e(\mathcal{L}_s) - \log_e(\mathcal{L}_g) \right) \\ &= -2 \log_e(\mathcal{L}_s) + 2 \log_e(\mathcal{L}_g) \\ &= \text{deviance}_s - \text{deviance}_g.\end{aligned}$$

Thus, the LRT can be computed as a difference in the deviance for the two models (ignoring the term for the saturated model). This is convenient as the deviance is a value of interest in other respects.

Say we flipped two coins, we could have

$$H_0: p_1 = p_2 \equiv p \quad (\text{the simpler model, 1 parameter}),$$

vs. the alternative of different probabilities of heads, hence

$$H_a: p_1 \neq p_2. \quad (\text{the more general model, 2 parameters}).$$

Compute MLEs under both models and compute the deviances ( $D$ ),

$$D_s = -2\log_e(\mathcal{L}(\hat{p})) , \quad K_s = 1, \text{ the number of parameters}$$

$$D_g = -2\log_e(\mathcal{L}(\hat{p}_1, \hat{p}_2)) , \quad K_g = 2, \text{ the number of parameters.}$$

The the LRT =  $D_s - D_g$ ,  $df = K_g - K_s = 1$ .

Thus, this test statistic is approximately  $\chi^2$  with 1 df under the null hypothesis. The approximation improves as sample size increases. Note, too that the log-likelihood for the saturated model is a constant and the same for both of the above models; thus it was deleted in this example.

Testing of null hypotheses has seen decreasing use in many areas of applied science over the past 2 decades. We will make some reference to LRTs so that students can better understand existing literature that makes use of these methods. Problems with statistical hypothesis testing will be outlined at a latter point.

## canoe.moore

Christopher Moore's Blog

# Linear mixed-effects regression p-values in R: A likelihood ratio test function

By Christopher Moore on September 7, 2010 9:18 PM | [7 Comments](#)

Following [Douglas Bates' advice](#) for those required to produce  $p$ -values for fixed-effects terms in a mixed-effects model, I wrote a function to perform a likelihood ratio test for each term in an `lmer()` object. Bates has championed the notion that calculating the  $p$ -values for fixed effects is not trivial. That's because with unbalanced, multilevel data, the [denominator degrees of freedom used to penalize certainty are unknown](#) (i.e., we're uncertain about how uncertain we should be). As the author of `lme4`, the foremost mixed-effects modeling package in R, he has practiced what he preaches by declining to [approximate denominator degrees of freedom as SAS does](#). Bates contends that alternative inferential approaches make  $p$ -values unnecessary. I agree with that position, focusing instead on information criteria and effect sizes. However, as a program evaluator, I recognize that some stakeholders find  $p$ -values useful for understanding findings.

A likelihood ratio test can be used to test  $H_0 : \beta = 0$  if the sample size is large. According to [Fitzmaurice, Laird, and Ware](#), twice the difference between the maximized log-likelihoods of two nested models,  $G^2 = 2(\hat{l}_{\text{full}} - \hat{l}_{\text{reduced}})$ , represents the degree to which the reduced model is inadequate. When the sample size is large,  $G^2 \sim \chi^2$  with degrees of freedom equal to the difference in parameters between the full and reduced models,  $m_{\text{full}} - m_{\text{reduced}}$ . The  $p$ -value is  $\Pr(\chi^2 > G^2 | H_0)$ .

What if the sample size is not large? A  $p$ -value based on  $\chi^2$  will be too liberal (i.e., the type I error rate will exceed the nominal  $p$ -value). More conservatively, we might say that  $G^2 \sim F$  with  $m_{\text{full}} - m_{\text{reduced}}$  numerator and  $N_{\text{effective}} - m_{\text{full}} - 1$  denominator degrees of freedom. According to [Snijders and Bosker](#), the effective sample size lies somewhere between  $M$  total micro-observations (i.e., at level one) and  $N$  clusters randomly sampled in earlier stages (i.e., at higher levels). Formally, the effective sample size is  $N_{\text{effective}} = \frac{Nn}{1+(n-1)\rho}$ , where  $n$  observations are nested within each cluster and intraclass correlation is  $\rho = \frac{\tau^2}{\tau^2 + \sigma^2}$ . Even if  $M = Nn$  is large,  $N_{\text{effective}}$  (and statistical power) could be quite small if  $N$  is small and  $\rho$  is large:  $\frac{5*100}{1+(100-1)*0.8} = 6.2$ . Unbalanced designs, modeling three or more levels, and cross-level interactions add to our uncertainty about the denominator degrees of freedom.

The function I wrote chews up the `lmer()` model call and concatenates the frame and model matrix slots, after which it iteratively fits (via maximum likelihood instead of restricted ML) models reduced by each fixed effect and compares them to the full model, yielding a vector of  $p$ -values based on  $\chi^2(1)$ . As the example shows, the function can handle shortcut formulas whereby lower order terms are implied by an interaction term. The function doesn't currently handle weights, `glmer()` objects, or on-the-fly transformations of the dependent variable [e.g., `log(dep.var) ~ ...`]. The accuracy of resulting  $p$ -values depends on large sample properties, as discussed above, so I don't recommend using the function with small samples. I'm working on another function that will calculate  $p$ -values based on the effective sample size estimated from intraclass correlation. I will post that function in a future entry. I'm sure the following function could be improved, but I wanted to go ahead share it with other applied researchers whose audience likes  $p$ -values. Please [let me know](#) if you see ways to make it better.

```
p.values.lmer <- function(x) {
  summary.model <- summary(x)
  data.lmer <- data.frame(model.matrix(x))
  names(data.lmer) <- names(fixef(x))
  names(data.lmer) <- gsub(pattern=":", x=names(data.lmer), replacement=". ", fixed=T)
  names(data.lmer) <- ifelse(names(data.lmer)=="(Intercept)", "Intercept", names(data.lmer))
  string.call <- strsplit(x=as.character(x@call), split=" + (", fixed=T)
  var.dep <- unlist(strsplit(x=unlist(string.call)[2], " ~ ", fixed=T))[1]
  vars.fixef <- names(data.lmer)
  formula.ranef <- paste(" + (", string.call[[2]][-1], sep="")
  formula.ranef <- paste(formula.ranef, collapse=" ")
  formula.full <- as.formula(paste(var.dep, " ~ -1 +", paste(vars.fixef, collapse=" + "), formula.ranef))
  data.ranef <- data.frame(x@frame[, which(names(x@frame) %in% names(ranef(x))))]
  names(data.ranef) <- names(ranef(x))
  data.lmer <- data.frame(x@frame[, 1], data.lmer, data.ranef)
  names(data.lmer)[1] <- var.dep
  out.full <- lmer(formula.full, data=data.lmer, REML=F)
  p.value.LRT <- vector(length=length(vars.fixef))
  for(i in 1:length(vars.fixef)) {
    formula.reduced <- as.formula(paste(var.dep, " ~ -1 +", paste(vars.fixef[-i], collapse=" + "), formula.ranef))
    out.reduced <- lmer(formula.reduced, data=data.lmer, REML=F)
    print(paste("Reduced by:", vars.fixef[i]))
    print(out.LRT <- data.frame(anova(out.full, out.reduced)))
  }
}
```

## Categories

[About me \(1\)](#)  
[Personal \(23\)](#)  
[Praxes \(45\)](#)

## Monthly Archives

[March 2014 \(1\)](#)  
[July 2013 \(1\)](#)  
[April 2013 \(1\)](#)  
[March 2013 \(1\)](#)  
[November 2012 \(1\)](#)  
[October 2012 \(1\)](#)  
[September 2012 \(1\)](#)  
[May 2012 \(1\)](#)  
[October 2011 \(1\)](#)  
[September 2011 \(1\)](#)  
[August 2011 \(1\)](#)  
[July 2011 \(2\)](#)  
[June 2011 \(1\)](#)  
[January 2011 \(1\)](#)  
[October 2010 \(2\)](#)  
[September 2010 \(3\)](#)  
[August 2010 \(3\)](#)  
[July 2010 \(1\)](#)  
[June 2010 \(1\)](#)  
[April 2010 \(1\)](#)  
[March 2010 \(3\)](#)  
[February 2010 \(3\)](#)  
[January 2010 \(1\)](#)  
[November 2009 \(1\)](#)  
[October 2009 \(1\)](#)  
[September 2009 \(4\)](#)  
[August 2009 \(1\)](#)  
[July 2009 \(2\)](#)  
[June 2009 \(1\)](#)  
[May 2009 \(3\)](#)  
[March 2009 \(1\)](#)  
[February 2009 \(3\)](#)  
[January 2009 \(2\)](#)  
[November 2008 \(1\)](#)  
[October 2008 \(3\)](#)  
[September 2008 \(3\)](#)  
[August 2008 \(2\)](#)  
[July 2008 \(3\)](#)  
[April 2008 \(5\)](#)

## Recent Entries

[Multilevel modeling: From lme4 to Mplus](#)  
[Measures of Effective Teaching weighting schemes: An improved plot](#)  
[Applying generalizability theory with R: A package](#)  
[Which schools are closing achievement gaps in Minnesota?](#)  
[Evaluation 2012 presentation](#)

```

p.value.LRT[i] <- round(out.LRT[2, 7], 3)
}
summary.model@coefs <- cbind(summary.model@coefs, p.value.LRT)
summary.model@methTitle <- c("\n", summary.model@methTitle,
                           "\n(p-values from comparing nested models fit by maximum likelihood)")
print(summary.model)
}

library(lme4)
library(SASmixed)
lmer.out <- lmer(strength ~ Program * Time + (Time|Subj), data=Weights)
p.values.lmer(lmer.out)

```

Yields:

```

Linear mixed model fit by REML
(p-values from comparing nested models fit by maximum likelihood)
Formula: strength ~ Program * Time + (Time | Subj)
Data: Weights
AIC BIC logLik deviance REMLdev
1343 1383 -661.7    1313    1323

Random effects:
Groups   Name        Variance Std.Dev. Corr
Subj     (Intercept) 9.038486 3.00641
          Time       0.031086 0.17631 -0.118
Residual            0.632957 0.79559
Number of obs: 399, groups: Subj, 57

Fixed effects:
            Estimate Std. Error t value p.value.LRT
(Intercept) 79.99018  0.68578 116.64000 0.000
ProgramRI    0.07009  1.02867  0.07000 0.944
ProgramWI    1.11526  0.95822  1.16000 0.235
Time        -0.02411  0.04286 -0.56000 0.564
ProgramRI:Time 0.12902  0.06429  2.01000 0.043
ProgramWI:Time 0.18397  0.05989  3.07000 0.002

Correlation of Fixed Effects:
             (Intr) PrgrRI PrgrWI Time   PrRI:T
ProgramRI -0.667
ProgramWI -0.716  0.477
Time      -0.174  0.116  0.125
PrgrRI:Tm 0.116 -0.174 -0.083 -0.667
PrgrWI:Tm 0.125 -0.083 -0.174 -0.716  0.477

```

Categories: [Praxes](#)

## 7 Comments

[Christopher Desjardins](#) | September 23, 2010 9:42 PM | [Reply](#)

This is very useful. However, I would recommend that readers also consider using a fit index, such as the AIC or BIC, or in the case of comparing nested models using the anova() function. I know that lots of folks rely on p-values for decision making, I am quite skeptical of them as you know especially when compared against some arbitrary 0.05 cutpoint. In fact, I would argue using alpha = 0.05 to base decisions is bad practice indeed. Instead, I recommend examining and exploring alternative hypotheses via model fitting as I feel they more accurately capture the inherent uncertainty underlying statistics. Or at a minimum just presenting the p-values and letting the readers decide. However, I do understand as an applied researcher and someone who works for the government, that interested parties are less interested in the underlying academics and more about whether such and such effect is 'significant' or not.

Cheers,

Chris

[Steve Brady](#) | October 20, 2010 2:49 PM | [Reply](#)

This works well for singular random effects.

Can the function be modified to handle nested random effects? At the moment, when I try it with nested random effects, I receive the following error:

Error in names(data.ranef) 'names' attribute [6] must be the same length as the vector [2]

[Ling Cui](#) | May 1, 2012 5:16 PM | [Reply](#)

Thank you very much for sharing this; very handy and useful!

[elmar tobi](#) | August 28, 2012 9:10 AM | [Reply](#)

Thank you very very much. I have been using a anova and  $2*(1-pnorm(abs(t-value)))$  where possible. P-values may indeed not be perfect but in genetics its the norm...

A very happy lmer user!

[Jaap Denissen](#) | August 30, 2013 6:10 AM | [Reply](#)

[Applying generalizability theory with R](#)  
[Estimating congeneric reliability with R](#)  
[Converting Rasch units \(RITs\) to thetas](#)  
[Urban canoeing: Dining at The Sample Room](#)  
[Urban canoeing: North/Northeast Minneapolis](#)

## Links

[50 Rivers Project](#)  
[Adventure Learning at University of Idaho \(AL@UI\)](#)  
[Christopher David Desjardin's Blog](#)  
[Citizen-Statistician](#)  
[CodeCogs Online LaTeX Equation Editor](#)  
[Dad vs Wild](#)  
[Fiscal Issues & Geeky Stuff](#)  
[Jeremy Y. Wang: A Grad Student's Thoughts on Mind, Brain, Education and Science](#)  
[KBEM Jazz 88, operated by Minneapolis Public Schools](#)  
[KFAI Radio Without Boundaries](#)  
[My Kitchen and Food Adventures](#)  
[Southern Poverty Law Center](#)  
[stacyboemiller: A Little Bit of Chaos. A Lot of Creativity.](#)  
[Surly Bikes Blog](#)  
[Statistical Modeling, Causal Inference, and Social Science](#)  
[Talkin' 'Bout the Next Generation](#)  
[Wilder Research](#)

## Search

This is great, thanks!

I spent a great deal of time looking for a way to obtain p-values for multilevel models. I used to work with pvals.fnc from the languageR package, but this no longer works with newer versions of R.

The function written by Christopher works like a charm and is consistent with state-of-the-art methodological advice. I needed to tweak the function a bit, however, pertaining to three lines of code:

```
p.value.LRT[i] summary.model$coefficients summary.model$methTitle
```

[Christopher Moore](#) | [November 2, 2013 9:27 AM](#) | [Reply](#)

Thanks to readers for informing me that the p.values.lmer() function no longer works with the new version of lme4. Please see earlier comments and Laurie Samuels' suggestions below for adapting the function work with the new version. I have no plans to update the function myself because I am too busy writing my dissertation.

Hi Christopher-

Thank you so much for posting your code for getting p-values from lme4 ([http://blog.lib.umn.edu/moor0554/canoemoore/2010/09/lmer\\_p-values\\_lrt.html](http://blog.lib.umn.edu/moor0554/canoemoore/2010/09/lmer_p-values_lrt.html)) --- I have found it really useful.

I recently updated my lme4 to the latest version, and there are apparently some changes that make the new version incompatible with code used for the old version... ([http://lme4.r-forge.r-project.org/misc/lme4\\_conversion.html](http://lme4.r-forge.r-project.org/misc/lme4_conversion.html)). So I wanted to let you know the two things I changed in your code to make it work with the new version:

1. The p.value.LRT[i] line now needs to be "p.value.LRT[i]
2. On the next two lines, the @ notation doesn't work with the latest version of lme4, so @coefs and @methTitle now need to be \$coefficients and \$methTitle, respectively.

I think this is what Jaap Denissen's comment was saying, but I didn't understand the comment until after I had worked through everything. I would have just posted all this as a comment myself so that other people could see it, but I kept failing the captcha...

Thank you again!

Laurie Samuels

Vanderbilt University

Laurie Samuels replied to [comment from Christopher Moore](#) | [November 14, 2013 3:53 PM](#) | [Reply](#)

It looks like the autoformatting on the blog is having trouble with the R code in my original email. In the p.value.LRT line, the only change is that the 7 needs to be changed to an 8.

## Leave a comment

Name

Email Address

URL

Remember personal info?

Comments (You may use HTML tags for style)

Captcha:



Type the characters you see in the picture above.

[Preview](#) [Submit](#)

13/6/2014

## Linear mixed-effects regression p-values in R: A likelihood ratio test function - canoe.moore

Powered by Movable Type Enterprise

This blog is licensed under a Creative Commons License.



# Likelihood Ratio Tests

Previously, we have employed a program (`Jtest`) to test an hypothesis involving many parameters (multiple constraints). This test, which involves comparing parameter estimates with hypothesized values, is known as a Wald test. We also used the `Jtest` program to compare two different estimators in a Hausman test. There is an alternative method for conducting joint tests, which can be implemented easily both from within and outside of Gauss. In this handout, we will first discuss this method in general and then see how it works in the familiar context of a linear regression model.

## 1 Motivation

We have already discussed that in large samples it is desirable to estimate models by selecting estimates so as to maximize the likelihood function. Intuitively, since the true parameter values generated the data, in large samples, the data are most likely to have been generated by the true parameter values. According, if we view the likelihood function as the probability of the data, as previously discussed it should seem reasonable that we should select estimates so as to maximize the likelihood.

With estimates of the parameters in a model selected to maximize the likelihood, it is natural to employ the maximized likelihood as a measure of how well the model fits the data. For example, suppose that we postulate a model, and select parameters to maximize the likelihood (the "probability of the data"). However, suppose that the maximized likelihood is quite small. In other words, while we selected parameter estimates to maximize it, the probability of the data coming from the postulated model is small. In this case, one would naturally say that the model provides a "poor" fit to the data. On the other hand, if the maximized likelihood is very large, then the data are very likely to have come from the postulated model. In this case, the model provides a "good" fit to the data.

With the maximized likelihood as our measure of data-fit, it should seem intuitive that we might compare models on the basis of how well they fit the data. In this handout, we will develop a test based on this principle. While this principle will apply to a very wide range of models, here we will examine it in terms of the probit model . With  $V \equiv X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + \beta_4$ , write

this model as:

$$Q = \begin{cases} 1 : & V + u > 0 \\ 0 : & V + u \leq 0 \end{cases}$$

Here, for example,  $X_1$  might be the price of the product,  $X_2$  the price of a close substitute, and  $X_3$  an individual's income, etc.

Employing the probit model above, suppose that we want to test whether or not  $X_2$  and  $X_3$  jointly belong in the model. Here, we are testing:

$$H_0 : \beta_2 = 0, \beta_3 = 0 \text{ vs. } H_1 : \text{Not } H_0.$$

To formulate a test, suppose that we first maximize the likelihood under the null hypothesis,  $H_0$ , by imposing the constraints under the null hypothesis. To do this, we would maximize the likelihood with  $X_2$  and  $X_3$  left out of the model. For this constrained case, write the likelihood (omitting  $X_2$  and  $X_3$  from the model) as:

$$L_0(\beta_1, \beta_4) \equiv \text{Likelihood under } H_0.$$

With  $(\hat{\beta}_1, \hat{\beta}_4)$  as the maximum likelihood estimates under  $H_0$ , the maximized likelihood under  $H_0$  is given as:

$$\hat{L}_0 \equiv L_0(\hat{\beta}_1, \hat{\beta}_4)$$

Similarly, for the unconstrained case (where all variables are included), define:

$$L_1(\beta_1, \beta_2, \beta_3, \beta_4) \equiv \text{Likelihood under } H_1$$

With  $(\bar{\beta}_1, \bar{\beta}_2, \bar{\beta}_3, \bar{\beta}_4)$  as the maximum likelihood estimates under  $H_1$ , the maximized likelihood under  $H_1$  is given as:

$$\hat{L}_1 \equiv L_1(\bar{\beta}_1, \bar{\beta}_2, \bar{\beta}_3, \bar{\beta}_4).$$

In maximizing a likelihood, it is equivalent to maximize the log likelihood instead. With the likelihood being a product of terms, its log will be more tractable as a sum of terms. In testing, it also turns out to be easier to work with log likelihoods. Define the likelihood-ratio test-statistic as:

$$LRT = 2 * \left[ \ln(\hat{L}_1) - \ln(\hat{L}_0) \right] = 2 * \ln\left(\hat{L}_1/\hat{L}_0\right).$$

Here, LRT has two properties that we require for test statistics. First, it naturally measures the difference in how the two models fit the data, where the likelihood is used to measure data-fit. We would then reject  $H_0$  if LRT is very large, which would mean that the model fits the data much better under  $H_1$  than under  $H_0$ . As a second required property, we must know the distribution of the test statistic under the null-hypothesis. Here, an amazing result is true in large samples. Namely, under the null hypothesis and when the sample size is large, LRT has a chi-squared distribution with  $r$  degrees of freedom. The degrees of freedom,  $r$ , is equal to the number of restrictions under the null hypothesis (2 in the example above).

With the rejection region now given by:

$$LRT > c,$$

we could now find the P-value, which is that significance level such that  $c=LRT$  in the sample. Alternatively, at any given significance level, we could look of  $c$  in a  $\chi^2(r)$  table. In the remainder of this handout, we will return to the more familiar linear model and show that likelihood ratio tests in such models are equivalent to the same types of tests you have done previously.

## 2 Likelihood Ratio Tests: Linear Models

The tests outlined above apply whenever we estimate models by maximum likelihood estimation. In this manner, these tests apply to the probit, ordered, and censored models that we have examined. To further motivate and explain these tests, in this section we will apply them to the linear model. Consider the following demand model without censoring:

$$Q \equiv X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + \beta_4 + u,$$

where the error,  $u$ , is distributed as  $N(0, \sigma^2)$ . In what follows, for simplicity we will first assume that the error variance,  $\sigma^2$ , is known. Subsequently, we will argue that in large samples it does not matter whether or not this variance is known.

To simplify the forms of the likelihoods in this case, it is useful to employ sums of squared residuals under the null and alternative hypotheses.

Accordingly, let:

$$\begin{aligned} R_0 &\equiv \sum [Q - X_1\hat{\beta}_1 + \hat{\beta}_4]^2 \\ R_1 &\equiv \sum [Q - X_1\bar{\beta}_1 + X_2\bar{\beta}_2 + X_3\bar{\beta}_3 + \bar{\beta}_4]^2 \end{aligned}$$

In the linear model with normal errors, OLS estimates (which by definition minimize a sum of squared residuals) are also maximum likelihood estimates. From discussions we have had in class, we may then write the maximized likelihoods as:

$$\begin{aligned} \hat{L}_0 &= C \exp \left( -\frac{1}{2} [R_0/\sigma^2] \right) \\ \hat{L}_1 &= C \exp \left( -\frac{1}{2} [R_1/\sigma^2] \right), \end{aligned}$$

where C is a constant. **You should make sure that you understand why this is the case.** The statistic LRT defined above is given in this case as:

$$\begin{aligned} LRT &= 2 * [Ln(L_1^*) - Ln(L_0^*)] \\ &= 2 * \left[ \left( -\frac{1}{2} [R_1/\sigma^2] \right) - \left( -\frac{1}{2} [R_0/\sigma^2] \right) \right] \\ &= [R_0 - R_1] / \sigma^2, \quad R_0 \succeq R_1 \end{aligned}$$

We reject the null hypothesis if LRT is large, which means that the sum of squared residuals under  $H_0$  is much larger than that under  $H_1$ . In this case, the model would fit the data much better under  $H_1$ , which would lead us to reject  $H_0$ . As to the distribution of LRT, it can be shown that:

$$R_0/\sigma^2 \text{ is distributed as } \chi^2(N - K_0),$$

where N is the sample size and  $K_0$  is the number of parameters estimated under  $H_0$ . Notice that we subtract one degree of freedom for each parameter that we estimate. Similarly,

$$R_1/\sigma^2 \text{ is distributed as } \chi^2(N - K_1),$$

where  $K_1$  is the number of parameters estimated in the unconstrained case. Again, we subtract one degree of freedom for every parameter we estimation (accounting for the degree of "estimation uncertainty"). Employing a

property of Chi-square variables:

$$\begin{aligned}[R_0 - R_1] / \sigma^2 &\text{ is distributed as } \chi^2(r), \\ r &= (N - K_0) - (N - K_1) = K_1 - K_0.\end{aligned}$$

The above discussion assumed that we knew the disturbance variance,  $\sigma^2$ . Denote  $\hat{\sigma}^2$  as a consistent estimator for  $\sigma^2$ . In Econ 322, when  $\sigma^2$  was not known, we replaced it with an estimate. With a minor adjustment (dividing the numerator of LRT by its degrees of freedom), the resulting test statistic had an F distribution. Here, we will assume the sample size is large, in which case  $\hat{\sigma}^2$  is probably close to  $\sigma^2$ . As a result, in large samples it can be shown that the above test is unaffected if we  $\hat{\sigma}^2$  replaces  $\sigma^2$  above. We can still use the  $\chi^2$  tables; no further adjustments are required.

In large samples, you would find that the F-test would give the same results as the  $\chi^2(r)$  test described above. Namely, when we reject under one test, we reject under the other. When we fail to reject under one test, we fail to reject under the other. The P-values under both tests are equivalent. It is in this sense that we say that the two tests are equivalent. The F-tests you did in the past may be viewed as versions of the Likelihood-Ratio test.

# Comparing Nested Models

Two models are *nested* if one model contains all the terms of the other, and at least one additional term.

The larger model is the *complete* (or *full*) model, and the smaller is the *reduced* (or *restricted*) model.

Example: with two independent variables  $x_1$  and  $x_2$ , possible *terms* are  $x_1$ ,  $x_1x_2$ ,  $x_1^2$ , and so on.

Consider three models:

- First order:

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2;$$

- Interaction:

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2;$$

- Full second order:

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2.$$

The first order model is nested within both the Interaction model and the Full second order model.

The Interaction model is nested within the Full second order model.

We usually want to use the simplest (most *parsimonious*) model that adequately fits the observed data.

One way to decide is by testing

- $H_0$ : reduced model is adequate;
- $H_a$ : full model is better.

When the full model has exactly *one* more term than the reduced model, we can use a *t*-test.

E.g., testing the Interaction model

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2;$$

against the First order model

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

- $H_0$ : “reduced model is adequate” is the same as  $H_0 : \beta_3 = 0$ .
- So the usual *t*-statistic is the relevant test statistic.

When the full model has *more than one* additional term, we use an *F*-test, which generalizes the *t*-test.

Basic idea: fit both models, and test whether the full model fits significantly better than the reduced model:

$$F = \frac{\left( \frac{\text{Drop in SSE}}{\text{Number of extra terms}} \right)}{s^2 \text{ for full model}}$$

where SSE is the sum of squared residuals.

When  $H_0$  is true,  $F$  follows the *F*-distribution, which we use to find the *P*-value.

E.g., testing the Full second order model

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2;$$

against the Interaction model

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2.$$

Here  $H_0$  is  $\beta_4 = \beta_5 = 0$ , and  $H_a$  is the opposite.

In R, the `lm()` method is not convenient for carrying out this test; `aov()` is better.

```
summary(aov(Cost ~ Weight + Distance + I(Weight * Distance) +
             I(Weight^2) + I(Distance^2), express))
    Df Sum Sq Mean Sq F value Pr(>F)
Weight          1 270.55 270.55 1380.001 2.17e-15 ***
Distance        1 143.63 143.63  732.616 1.72e-13 ***
I(Weight * Distance) 1 31.27 31.27  159.487 4.84e-09 ***
I(Weight^2)      1   3.80   3.80   19.383 0.000602 ***
I(Distance^2)   1   0.09   0.09    0.451 0.512657
Residuals       14  2.74   0.20
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1     1
```

---

```
summary(aov(Cost ~ Weight + Distance + I(Weight * Distance),
            express))
    Df Sum Sq Mean Sq F value Pr(>F)
Weight          1 270.55 270.55  652.59 2.14e-14 ***
Distance        1 143.63 143.63  346.45 2.89e-12 ***
I(Weight * Distance) 1 31.27 31.27   75.42 1.88e-07 ***
Residuals       16  6.63   0.41
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1     1
```

These are *sequential* sums of squares, adding each term to the model in order.

See Residuals line in each set of results:

$$\text{SSE}(\text{Full second order}) = 2.74,$$

$$\text{SSE}(\text{Interaction}) = 6.63,$$

so

$$F = \frac{(6.63 - 2.74)/2}{0.20} = 9.75, P < .01$$

You can, in fact, calculate  $F$  from the output for the full model:

- Note that, because the terms are added sequentially, the sums of squares for the common terms (Weight, Distance, and Weight \* Distance) are the same in both models.
- In the reduced model, the extra terms (Weight<sup>2</sup> and Distance<sup>2</sup>) have gone away.
- Their *combined* sum of squares,  $3.80 + 0.09 = 3.89$ , is exactly the increase in SSE,  $6.63 - 2.74 = 3.89$ .

So we can also calculate

$$F = \frac{(\text{Sum Sq for Weight}^2 + \text{Sum Sq for Distance}^2)/2}{\text{Mean Square for Residuals}}$$

using only the output for the full model.

Note that  $F$  was calculated imprecisely, because of rounding.

We can get more digits using `print(summary(...), digits = 8)` for example, or calculate  $F$  to full precision:

```
s = summary(aov(Cost ~ Weight + Distance + I(Weight * Distance) +
                  I(Weight^2) + I(Distance^2), express))[[1]]
sum(s[c("I(Weight^2)", "I(Distance^2)", "Sum Sq")]) / 2 /
  s["Residuals", "Mean Sq"]
```

We could use the same  $F$ -test when there is only one additional term in the full model, based on just one line in the ANOVA table, provided it is the last term in the formula.

It appears very different from the  $t$ -test described earlier.

Some matrix algebra shows that it is, in fact, exactly the same test:

- The  $F$ -statistic is exactly the square of the  $t$ -statistic.
- The  $F$  critical values are exactly the squares of the (two-sided)  $t$  critical values.
- So the  $P$ -value is exactly the same.

# Complete Example: Road Construction Cost

Data from the Florida Attorney General's office

$y$  = successful bid;

$x_1$  = DOT engineer's estimate of cost

$x_2$  = indicator of fixed bidding:

$$x_2 = \begin{cases} 1 & \text{if fixed} \\ 0 & \text{if competitive} \end{cases}$$

Get the data and plot them:

```
flag = read.table("Text/Exercises&Examples/FLAG.txt",
                  header = TRUE)
pairs(flag[,-1])
```

Section 4.14 suggests beginning with the full second order model, and simplifying it as far as possible (but no further!).

We'll take the opposite approach: begin with the first order model, and complicate it as far as necessary.

Because  $x_2$  is a dummy variable, the first order model is a pair of parallel straight lines:

```
summary(lm(COST ~ DOTEST + STATUS, flag))
```

## First order model

Call:

```
lm(formula = COST ~ DOTESEN + STATUS, data = flag)
```

Residuals:

Min	1Q	Median	3Q	Max
-2199.94	-73.83	7.76	53.68	1722.42

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )		
(Intercept)	-20.537724	26.817718	-0.766	0.444558		
DOTESEN	0.930781	0.009744	95.519	< 2e-16 ***		
STATUS	166.357224	49.287822	3.375	0.000864 ***		
---						
Signif. codes:	0 ***	0.001 **	0.01 *	0.05 .	0.1	1

Residual standard error: 306.3 on 232 degrees of freedom

Multiple R-squared: 0.9755, Adjusted R-squared: 0.9752

F-statistic: 4610 on 2 and 232 DF, p-value: < 2.2e-16

Both variables look important.

The DOTESt coefficient is close to 1, so the winning bids roughly track the estimated cost.

The positive STATUS coefficient means the line for STATUS = 1 is higher than the line for STATUS = 0.

Are the slopes different? Try the interaction model:

```
summary(lm(COST ~ DOTESt * STATUS, flag))
```

## Interaction model

Call:

```
lm(formula = COST ~ DOTESEN * STATUS, data = flag)
```

Residuals:

Min	1Q	Median	3Q	Max
-2143.12	-43.21	1.39	40.17	1765.99

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )		
(Intercept)	-6.428025	26.208287	-0.245	0.806		
DOTESEN	0.921338	0.009723	94.755	< 2e-16 ***		
STATUS	28.673189	58.661711	0.489	0.625		
DOTESEN:STATUS	0.163282	0.040431	4.039	7.32e-05 ***		
---						
Signif. codes:	0 ***	0.001 **	0.01 *	0.05 .	0.1	1

Residual standard error: 296.7 on 231 degrees of freedom

Multiple R-squared: 0.9771, Adjusted R-squared: 0.9768

F-statistic: 3281 on 3 and 231 DF, p-value: < 2.2e-16

The interaction term is highly significant: reject the first order model in favor of the interaction model.

The slopes are 0.921338 for  $\text{STATUS} = 0$ , and  $0.921338 + 0.163282 = 1.08462$  for  $\text{STATUS} = 1$ .

So the competitive auctions are won with bids that fall slightly below the estimated cost, while the fixed winning bids fall slightly above the estimated cost.

To validate the interaction model, we compare it with (finally!) the full second order model.

### Note

When some variables are qualitative, the “full second order model” consists of the full second order (i.e., quadratic) model in the *quantitative* variables, plus the *interactions* of those terms with the *qualitative* variables:

```
summary(lm(COST ~ DOTEST + STATUS + I(DOTEST * STATUS) +  
          I(DOTEST^2) + I(DOTEST^2 * STATUS), flag))
```

## Full second order model

Call:

```
lm(formula = COST ~ DOTESEN + STATUS + I(DOTESEN * STATUS) +  
    I(DOTESEN^2) + I(DOTESEN^2 * STATUS), data = flag)
```

Residuals:

Min	1Q	Median	3Q	Max
-2143.50	-35.38	1.27	46.58	1771.19

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )		
(Intercept)	-2.972e+00	3.089e+01	-0.096	0.92344		
DOTESEN	9.155e-01	2.917e-02	31.385	< 2e-16 ***		
STATUS	-3.673e+01	7.477e+01	-0.491	0.62375		
I(DOTESEN * STATUS)	3.242e-01	1.192e-01	2.721	0.00702 **		
I(DOTESEN^2)	7.191e-07	3.404e-06	0.211	0.83288		
I(DOTESEN^2 * STATUS)	-3.576e-05	2.478e-05	-1.443	0.15041		
---						
Signif. codes:	0 ***	0.001 **	0.01 *	0.05 .	0.1	1

Residual standard error: 296.6 on 229 degrees of freedom

Multiple R-squared: 0.9773, Adjusted R-squared: 0.9768

F-statistic: 1970 on 5 and 229 DF, p-value: < 2.2e-16

## Full second order model, ANOVA

```
summary(aov(COST ~ DOTEST + STATUS + I(DOTEST * STATUS) +
             I(DOTEST^2) + I(DOTEST^2 * STATUS), flag))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
DOTEST	1	864038187	864038187	9818.947	< 2e-16 ***
STATUS	1	1069006	1069006	12.148	0.000589 ***
I(DOTEST * STATUS)	1	1435733	1435733	16.316	7.32e-05 ***
I(DOTEST^2)	1	15	15	0.000	0.989487
I(DOTEST^2 * STATUS)	1	183210	183210	2.082	0.150411
Residuals	229	20151321	87997		

---

Signif. codes: 0 \*\*\* 0.001 \*\* 0.01 \* 0.05 . 0.1 1

The significance of the two terms that were added is tested using an  $F$  statistic with 2 degrees of freedom in the numerator; the value is only slightly greater than 1, and is completely consistent with the null hypothesis that the interaction model is adequate.

# Model Training and Tuning

## Contents

- [Model Training and Parameter Tuning](#)
  - [An Example](#)
- [Basic Parameter Tuning](#)
- [Customizing the Tuning Process](#)
  - [Pre-Processing Options](#)
  - [Alternate Tuning Grids](#)
  - [Plotting the Resampling Profile](#)
  - [The trainControl Function](#)
  - [Alternate Performance Metrics](#)
  - [Choosing the Final Model](#)
- [Extracting Predictions and Class Probabilities](#)
- [Exploring and Comparing Resampling Distributions](#)
  - [Within-Model](#)
  - [Between-Models](#)
- [Fitting Models Without Parameter Tuning](#)

## Model Training and Parameter Tuning

The [caret](#) package has several functions that attempt to streamline the model building and evaluation process.

The `train` function can be used to

- evaluate, using resampling, the effect of model tuning parameters on performance
- choose the "optimal" model across these parameters
- estimate model performance from a training set

First, a specific model must be chosen. Currently, 150 are available using [caret](#); see [train Model List](#) or [train Models By Tag](#) for details. On these pages, there are lists of tuning parameters that can potentially be optimized. [User-defined models](#) can also be created.

The first step in tuning the model (line 1 in the algorithm above) is to choose a set of parameters to evaluate. For example, if fitting a Partial Least Squares (PLS) model, the number of PLS components to evaluate must be specified.

```

1 Define sets of model parameter values to evaluate
2 for each parameter set do
3   for each resampling iteration do
4     Hold-out specific samples
5     [Optional] Pre-process the data
6     Fit the model on the remainder
7     Predict the hold-out samples
8   end
9   Calculate the average performance across hold-out predictions
10 end
11 Determine the optimal parameter set
12 Fit the final model to all the training data using the optimal parameter set

```

Once the model and tuning parameter values have been defined, the type of resampling should be also be specified. Currently,  $k$ -fold cross-validation (once or repeated), leave-one-out cross-validation and bootstrap (simple estimation or the 632 rule) resampling methods can be used by `train`. After resampling, the process produces a profile of performance measures is available to guide the user as to which tuning parameter values should be chosen. By default, the function automatically chooses the tuning parameters associated with the best value, although different algorithms can be used (see details below below).

### An Example

The Sonar data are available in the [mlbench](#) package. Here, we load the data:

```

library(mlbench)
data(Sonar)
str(Sonar[, 1:10])

```

```

'data.frame': 208 obs. of 10 variables:
 $ V1 : num 0.02 0.0453 0.0262 0.01 0.0762 0.0286 0.0317 0.0519 0.0223 0.0164 ...
 $ V2 : num 0.0371 0.0523 0.0582 0.0171 0.0666 0.0453 0.0956 0.0548 0.0375 0.0173 ...

```

```
$ V3 : num  0.0428 0.0843 0.1099 0.0623 0.0481 ...
$ V4 : num  0.0207 0.0689 0.1083 0.0205 0.0394 ...
$ V5 : num  0.0954 0.1183 0.0974 0.0205 0.059 ...
$ V6 : num  0.0986 0.2583 0.228 0.0368 0.0649 ...
$ V7 : num  0.154 0.216 0.243 0.11 0.121 ...
$ V8 : num  0.16 0.348 0.377 0.128 0.247 ...
$ V9 : num  0.3109 0.3337 0.5598 0.0598 0.3564 ...
$ V10: num  0.211 0.287 0.619 0.126 0.446 ...
```

The function `createDataPartition` can be used to create a stratified random sample of the data into training and test sets:

```
library(caret)
set.seed(998)
inTraining <- createDataPartition(Sonar$Class, p = 0.75, list = FALSE)
training <- Sonar[inTraining, ]
testing <- Sonar[-inTraining, ]
```

We will use these data illustrate functionality on this (and other) pages.

## Basic Parameter Tuning

By default, simple bootstrap resampling is used for line 3 in the algorithm above. Others are available, such as repeated  $K$ -fold cross-validation, leave-one-out etc. The function `trainControl` can be used to specify the type of resampling:

```
fitControl <- trainControl(## 10-fold CV
                           method = "repeatedcv",
                           number = 10,
                           ## repeated ten times
                           repeats = 10)
```

More information about `trainControl` is given in [a section below](#).

The first two arguments to `train` are the predictor and outcome data objects, respectively. The third argument, `method`, specifies the type of model (see [train Model List](#) or [train Models By Tag](#)). To illustrate, we will fit a boosted tree model via the `gbm` package. The basic syntax for fitting this model using repeated cross-validation is shown below:

```
set.seed(825)
gbmFit1 <- train(Class ~ ., data = training,
                  method = "gbm",
                  trControl = fitControl,
                  ## This last option is actually one
                  ## for gbm() that passes through
                  verbose = FALSE)
gbmFit1
```

```
Stochastic Gradient Boosting

157 samples
 60 predictors
 2 classes: 'M', 'R'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 10 times)

Summary of sample sizes: 142, 142, 140, 142, 142, 141, ...
```

Resampling results across tuning parameters:

	interaction.depth	n.trees	Accuracy	Kappa	Accuracy SD	Kappa SD
1		50	0.8	0.5	0.1	0.2
1		100	0.8	0.6	0.1	0.2
1		200	0.8	0.6	0.09	0.2
2		50	0.8	0.6	0.1	0.2
2		100	0.8	0.6	0.09	0.2
2		200	0.8	0.6	0.1	0.2
3		50	0.8	0.6	0.09	0.2
3		100	0.8	0.6	0.09	0.2
3		200	0.8	0.6	0.08	0.2

```
Tuning parameter 'shrinkage' was held constant at a value of 0.1
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were n.trees = 150, interaction.depth = 3 and shrinkage = 0.1.
```

For a gradient boosting machine (GBM) model, there are three main tuning parameters:

- number of iterations, i.e. trees, (called `n.trees` in the `gbm` function)
- complexity of the tree, called `interaction.depth`
- learning rate: how quickly the algorithm adapts, called `shrinkage`

The default values tested for this model are shown in the first two columns (shrinkage is not shown because the grid set of candidate models all use a value of 0.1 for this tuning parameter). The column labeled "Accuracy" is the overall agreement rate averaged over cross-validation iterations. The agreement standard deviation is also calculated from the cross-validation results. The column "Kappa" is Cohen's (unweighted) Kappa statistic averaged across the resampling results. train works with specific models (see [train Model List](#) or [train Models By Tag](#)). For these models, train can automatically create a grid of tuning parameters. By default, if  $p$  is the number of tuning parameters, the grid size is  $3^p$ . As another example, regularized discriminant analysis (RDA) models have two parameters (gamma and lambda), both of which lie on  $[0, 1]$ . The default training grid would produce nine combinations in this two-dimensional space.

There are several [notes](#) regarding specific model behaviors for train. There is additional functionality in train that is described in the next section.

## Customizing the Tuning Process

There are a few ways to customize the process of selecting tuning/complexity parameters and building the final model.

### Pre-Processing Options

As previously mentioned, train can pre-process the data in various ways prior to model fitting. The function `preProcess` is automatically used. This function can be used for centering and scaling, imputation (see details below), applying the spatial sign transformation and feature extraction via principal component analysis or independent component analysis.

To specify what pre-processing should occur, the `train` function has an argument called `preProcess`. This argument takes a character string of methods that would normally be passed to the `method` argument of the [preProcess function](#). Additional options to the `preProcess` function can be passed via the `trainControl` function.

These processing steps would be applied during any predictions generated using `predict.train`, `extractPrediction` or `extractProbs` (see details later in this document). The pre-processing would **not** be applied to predictions that directly use the `object$finalModel` object.

For imputation, there are three methods currently implemented:

- $k$ -nearest neighbors takes a sample with missing values and finds the  $k$  closest samples in the training set. The average of the  $k$  training set values for that predictor are used as a substitute for the original data. When calculating the distances to the training set samples, the predictors used in the calculation are the ones with no missing values for that sample and no missing values in the training set.
- another approach is to fit a bagged tree model for each predictor using the training set samples. This is usually a fairly accurate model and can handle missing values. When a predictor for a sample requires imputation, the values for the other predictors are fed through the bagged tree and the prediction is used as the new value. This model can have significant computational cost.
- the median of the predictor's training set values can be used to estimate the missing data.

If there are missing values in the training set, PCA and ICA models only use complete samples.

### Alternate Tuning Grids

The tuning parameter grid can be specified by the user. The argument `tuneGrid` can take a data frame with columns for each tuning parameter. The column names should be the same as the fitting function's arguments. For the previously mentioned RDA example, the names would be `gamma` and `lambda`. `train` will tune the model over each combination of values in the rows.

For the boosted tree model, we can fix the learning rate and evaluate more than three values of `n.trees`:

```
gbmGrid <- expand.grid(interaction.depth = c(1, 5, 9),
                        n.trees = (1:30)*50,
                        shrinkage = 0.1)
nrow(gbmGrid)
set.seed(825)
gbmFit2 <- train(Class ~ ., data = training,
                  method = "gbm",
                  trControl = fitControl,
                  verbose = FALSE,
                  ## Now specify the exact models
                  ## to evaluate:
                  tuneGrid = gbmGrid)
gbmFit2

Stochastic Gradient Boosting

157 samples
 60 predictors
 2 classes: 'M', 'R'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 10 times)

Summary of sample sizes: 142, 142, 140, 142, 142, 141, ...

Resampling results across tuning parameters:

interaction.depth  n.trees  Accuracy  Kappa  Accuracy SD  Kappa SD
1                 50        0.77     0.53    0.1          0.2
1                 100       0.78     0.56    0.095      0.19
1                 150       0.79     0.58    0.094      0.19
1                 200       0.79     0.58    0.094      0.19
```

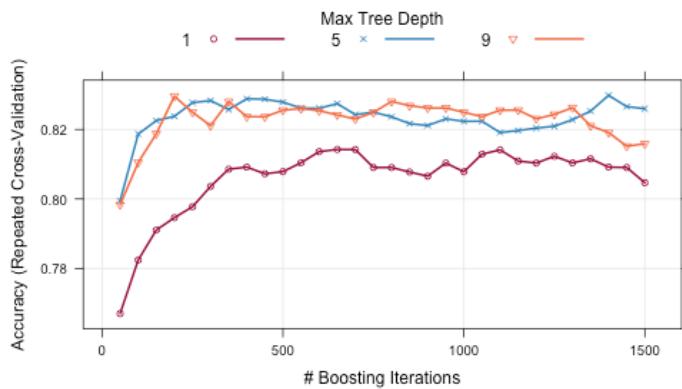
```
:          :          :          :          :
9       1200     0.82     0.64     0.092    0.19
9       1200     0.82     0.64     0.09      0.18
9       1300     0.83     0.65     0.088      0.18
9       1400     0.82     0.64     0.092      0.19
9       1400     0.82     0.63     0.095      0.19
9       1400     0.82     0.63     0.092      0.19
9       1500     0.82     0.63     0.093      0.19
```

Tuning parameter 'shrinkage' was held constant at a value of 0.1  
 Accuracy was used to select the optimal model using the largest value.  
 The final values used for the model were n.trees = 1400, interaction.depth = 5 and shrinkage = 0.1.

## Plotting the Resampling Profile

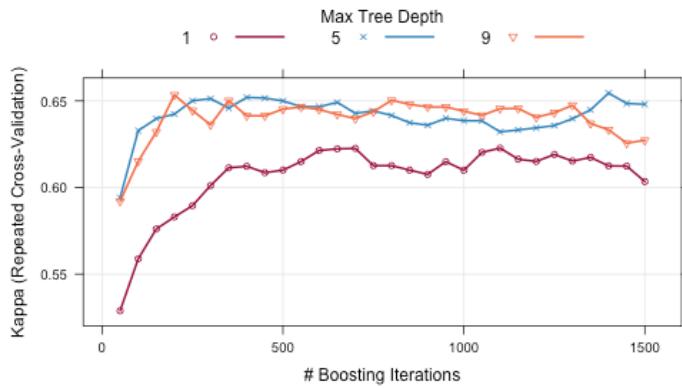
The plot function can be used to examine the relationship between the estimates of performance and the tuning parameters. For example, a simple invocation of the function shows the results for the first performance measure:

```
trellis.par.set(caretTheme())
plot(gbmFit2)
```



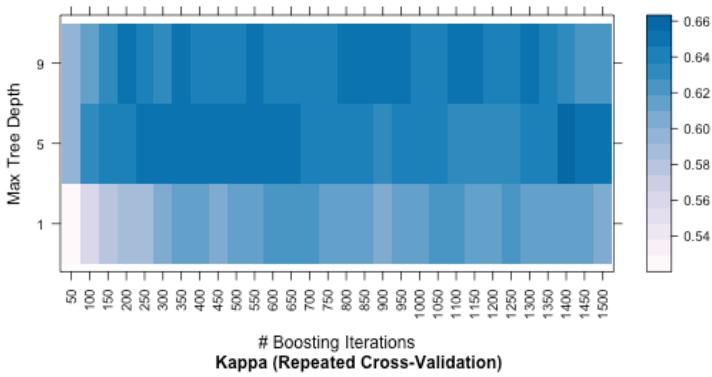
Other performance metrics can be shown using the metric option:

```
trellis.par.set(caretTheme())
plot(gbmFit2, metric = "Kappa")
```



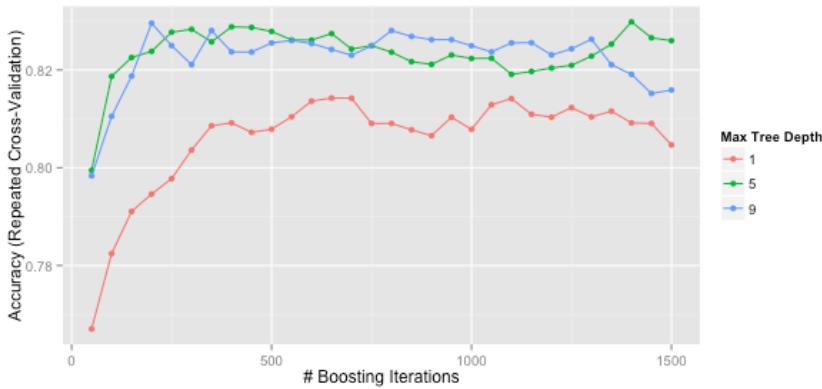
Other types of plot are also available. See ?plot.train for more details. The code below shows a heatmap of the results:

```
trellis.par.set(caretTheme())
plot(gbmFit2, metric = "Kappa", plotType = "level",
scales = list(x = list(rot = 90)))
```



A ggplot method can also be used:

```
ggplot(gbmFit2)
```



There are also plot functions that show more detailed representations of the resampled estimates. See `?xyplot.train` for more details.

From these plots, a different set of tuning parameters may be desired. To change the final values without starting the whole process again, the `update.train` can be used to refit the final model. See `?update.train`

## The trainControl Function

The function `trainControl` generates parameters that further control how models are created, with possible values:

- **method:** The resampling method: `"boot"`, `"cv"`, `"LOOCV"`, `"LGOCV"`, `"repeatedcv"`, `"timeslice"`, `"none"` and `"oob"`. The last value, out-of-bag estimates, can only be used by random forest, bagged trees, bagged earth, bagged flexible discriminant analysis, or conditional tree forest models. GBM models are not included (the `gbm` package maintainer has indicated that it would not be a good idea to choose tuning parameter values based on the model OOB error estimates with boosted trees). Also, for leave-one-out cross-validation, no uncertainty estimates are given for the resampled performance measures.
- **number and repeats:** number controls with the number of folds in  $K$ -fold cross-validation or number of resampling iterations for bootstrapping and leave-group-out cross-validation. repeats applied only to repeated  $K$ -fold cross-validation. Suppose that `method = "repeatedcv"`, `number = 10` and `repeats = 3`, then three separate 10-fold cross-validations are used as the resampling scheme.
- **verboseIter:** A logical for printing a training log.
- **returnData:** A logical for saving the data into a slot called `trainingData`.
- **p:** For leave-group out cross-validation: the training percentage
- For `method = "timeslice"`, `trainControl` has options `initialWindow`, `horizon` and `fixedWindow` that govern how [cross-validation can be used for time series data](#).
- **classProbs:** a logical value determining whether class probabilities should be computed for held-out samples during resample.
- **index and indexOut:** optional lists with elements for each resampling iteration. Each list element is the sample rows used for training at that iteration or should be held-out. When these values are not specified, `train` will generate them.
- **summaryFunction:** a function to compute alternate performance summaries.
- **selectionFunction:** a function to choose the optimal tuning parameters, and examples.
- **PCAthresh, ICAcomp and k:** these are all options to pass to the `preProcess` function (when used).
- **returnResamp:** a character string containing one of the following values: `"all"`, `"final"` or `"none"`. This specifies how much of the resampled performance measures to save.
- **allowParallel:** a logical that governs whether `train` should [use parallel processing \(if available\)](#).

There are several other options not discussed here.

## Alternate Performance Metrics

The user can change the metric used to determine the best settings. By default, RMSE and  $R^2$  are computed for regression while accuracy and Kappa are computed for classification. Also by default, the parameter values are chosen using RMSE and accuracy, respectively for regression and classification. The `metric` argument of

the train function allows the user to control which the optimality criterion is used. For example, in problems where there are a low percentage of samples in one class, using metric = "Kappa" can improve quality of the final model.

If none of these parameters are satisfactory, the user can also compute custom performance metrics. The trainControl function has a argument called summaryFunction that specifies a function for computing performance. The function should have these arguments:

- data is a reference for a data frame or matrix with columns called `obs` and `pred` for the observed and predicted outcome values (either numeric data for regression or character values for classification). Currently, class probabilities are not passed to the function. The values in data are the held-out predictions (and their associated reference values) for a single combination of tuning parameters. If the classProbs argument of the trainControl object is set to `TRUE`, additional columns in `data` will be present that contains the class probabilities. The names of these columns are the same as the class levels. Also, if weights were specified in the call to train, a column called `weights` will also be in the data set.
- lev is a character string that has the outcome factor levels taken from the training data. For regression, a value of `NULL` is passed into the function.
- model is a character string for the model being used (i.e. the value passed to the method argument of train).

The output to the function should be a vector of numeric summary metrics with non-null names. By default, train evaluate classification models in terms of the predicted classes. Optionally, class probabilities can also be used to measure performance. To obtain predicted class probabilities within the resampling process, the argument classProbs in trainControl must be set to `TRUE`. This merges columns of probabilities into the predictions generated from each resample (there is a column per class and the column names are the class names).

As shown in the last section, custom functions can be used to calculate performance scores that are averaged over the resamples. Another built-in function, `twoClassSummary`, will compute the sensitivity, specificity and area under the ROC curve:

```
head(twoClassSummary)

1 function (data, lev = NULL, model = NULL)
2 {
3   require(pROC)
4   if (!all(levels(data[, "pred"]) == levels(data[, "obs"])))
5     stop("levels of observed and predicted data do not match")
6   rocObject <- try(pROC::roc(data$obs, data[, lev[1]]), silent = TRUE)
```

To rebuild the boosted tree model using this criterion, we can see the relationship between the tuning parameters and the area under the ROC curve using the following code:

```
fitControl <- trainControl(method = "repeatedcv",
                            number = 10,
                            repeats = 10,
                            ## Estimate class probabilities
                            classProbs = TRUE,
                            ## Evaluate performance using
                            ## the following function
                            summaryFunction = twoClassSummary)
set.seed(825)
gbmFit3 <- train(Class ~ ., data = training,
                  method = "gbm",
                  trControl = fitControl,
                  verbose = FALSE,
                  tuneGrid = gbmGrid,
                  ## Specify which metric to optimize
                  metric = "ROC")
gbmFit3
```

```
Stochastic Gradient Boosting

157 samples
 60 predictors
 2 classes: 'M', 'R'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 10 times)

Summary of sample sizes: 142, 142, 140, 142, 142, 141, ...

Resampling results across tuning parameters:
```

interaction.depth	n.trees	ROC	Sens	Spec	ROC SD	Sens SD	Spec SD
1	50	0.86	0.83	0.69	0.1	0.15	0.17
1	100	0.87	0.84	0.73	0.093	0.14	0.16
1	150	0.88	0.85	0.74	0.088	0.13	0.17
1	200	0.88	0.86	0.75	0.084	0.14	0.17
:	:	:	:	:	:	:	:
9	1200	0.91	0.89	0.76	0.075	0.12	0.16
9	1200	0.91	0.89	0.75	0.074	0.12	0.15
9	1300	0.91	0.89	0.75	0.075	0.11	0.15
9	1400	0.91	0.89	0.76	0.073	0.11	0.15
9	1400	0.91	0.88	0.76	0.073	0.12	0.15
9	1400	0.91	0.87	0.77	0.072	0.13	0.15
9	1500	0.92	0.87	0.78	0.072	0.13	0.15

Tuning parameter 'shrinkage' was held constant at a value of 0.1

```
ROC was used to select the optimal model using the largest value.
The final values used for the model were n.trees = 1500, interaction.depth = 9 and shrinkage = 0.1.
```

In this case, the average area under the ROC curve associated with the optimal tuning parameters was 0.916 across the 100 resamples.

## Choosing the Final Model

Another method for customizing the tuning process is to modify the algorithm that is used to select the "best" parameter values, given the performance numbers. By default, the train function chooses the model with the largest performance value (or smallest, for mean squared error in regression models). Other schemes for selecting model can be used. [Breiman et al\(1984\)](#) suggested the "one standard error rule" for simple tree-based models. In this case, the model with the best performance value is identified and, using resampling, we can estimate the standard error of performance. The final model used was the simplest model within one standard error of the (empirically) best model. With simple trees this makes sense, since these models will start to over-fit as they become more and more specific to the training data.

train allows the user to specify alternate rules for selecting the final model. The argument selectionFunction can be used to supply a function to algorithmically determine the final model. There are three existing functions in the package: best is chooses the largest/smallest value, oneSE attempts to capture the spirit of [Breiman et al\(1984\)](#) and tolerance selects the least complex model within some percent tolerance of the best value. See `?best` for more details.

User-defined functions can be used, as long as they have the following arguments:

- `x` is a data frame containing the tune parameters and their associated performance metrics. Each row corresponds to a different tuning parameter combination.
- `metric` a character string indicating which performance metric should be optimized (this is passed in directly from the `metric` argument of `train`).
- `maximize` is a single logical value indicating whether larger values of the performance metric are better (this is also directly passed from the call to `train`).

The function should output a single integer indicating which row in `x` is chosen.

As an example, if we chose the previous boosted tree model on the basis of overall accuracy, we would choose: `n.trees = 1500, interaction.depth = 9, shrinkage = 0.1`. However, the scale in this plots is fairly tight, with accuracy values ranging from 0.863 to 0.916. A less complex model (e.g. fewer, more shallow trees) might also yield acceptable accuracy.

The tolerance function could be used to find a less complex model based on  $(x - x_{\text{best}})/x_{\text{best}} \times 100$ , which is the percent difference. For example, to select parameter values based on a 2% loss of performance:

```
whichTwoPct <- tolerance(gbmFit3$results, metric = "ROC", tol = 2, maximize = TRUE)
cat("best model within 2 pct of best:\n")

best model within 2 pct of best:

gbmFit3$results[whichTwoPct, 1:6]

shrinkage interaction.depth n.trees      ROC    Sens   Spec
32          0.1                  5     100 0.8988 0.864 0.765
```

This indicates that we can get a less complex model with an area under the ROC curve of 0.899 (compared to the "pick the best" value of 0.916).

The main issue with these functions is related to ordering the models from simplest to complex. In some cases, this is easy (e.g. simple trees, partial least squares), but in cases such as this model, the ordering of models is subjective. For example, is a boosted tree model using 100 iterations and a tree depth of 2 more complex than one with 50 iterations and a depth of 8? The package makes some choices regarding the orderings. In the case of boosted trees, the package assumes that increasing the number of iterations adds complexity at a faster rate than increasing the tree depth, so models are ordered on the number of iterations then ordered with depth. See `?best` for more examples for specific models.

## Extracting Predictions and Class Probabilities

As previously mentioned, objects produced by the `train` function contain the "optimized" model in the `finalModel` sub-object. Predictions can be made from these objects as usual. In some cases, such as `pls` or `gbm` objects, additional parameters from the optimized fit may need to be specified. In these cases, the `train` objects uses the results of the parameter optimization to predict new samples. For example, if predictions were create using `predict.gbm`, the user would have to specify the number of trees directly (there is no default). Also, for binary classification, the predictions from this function take the form of the probability of one of the classes, so extra steps are required to convert this to a factor vector. `predict.train` automatically handles these details for this (and for other models).

Also, there are very few standard syntaxes for model predictions in R. For example, to get class probabilities, many `predict` methods have an argument called `type` that is used to specify whether the classes or probabilities should be generated. Different packages use different values of `type`, such as `"prob"`, `"posterior"`, `"response"`, `"probability"` or `"raw"`. In other cases, completely different syntax is used.

For `predict.train`, the `type` options are standardized to be `"class"` and `"prob"` (the underlying code matches these to the appropriate choices for each model. For example:

```
predict(gbmFit3, newdata = head(testing))

[1] R R R M M
Levels: M R
```

```
predict(gbmFit3, newdata = head(testing), type = "prob")

      M          R
1 1.957e-06 0.9999980
2 4.824e-14 1.0000000
3 1.118e-32 1.0000000
4 1.528e-10 1.0000000
5 9.997e-01 0.0003027
6 9.995e-01 0.0004851
```

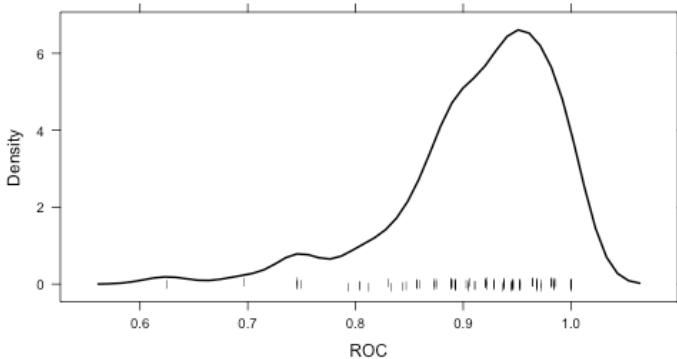
## Exploring and Comparing Resampling Distributions

### Within-Model

There are several [lattice](#) functions than can be used to explore relationships between tuning parameters and the resampling results for a specific model:

- xyplot and stripplot can be used to plot resampling statistics against (numeric) tuning parameters.
- histogram and densityplot can also be used to look at distributions of the tuning parameters across tuning parameters.

For example, the following statements create a density plot:



Note that if you are interested in plotting the resampling results across multiple tuning parameters, the option `resamples = "all"` should be used in the control object.

### Between-Models

The [caret](#) package also includes functions to characterize the differences between models (generated using `train`, `sbf` or `rfe`) via their resampling distributions. These functions are based on the work of [Hothorn et al. \(2005\)](#) and [Eugster et al \(2008\)](#).

First, a support vector machine model is fit to the Sonar data. The data are centered and scaled using the `preProc` argument. Note that the same random number seed is set prior to the model that is identical to the seed used for the boosted tree model. This ensures that the same resampling sets are used, which will come in handy when we compare the resampling profiles between models.

```
set.seed(825)
svmFit <- train(Class ~ ., data = training,
                 method = "svmRadial",
                 trControl = fitControl,
                 preProc = c("center", "scale"),
                 tuneLength = 8,
                 metric = "ROC")
svmFit

Support Vector Machines with Radial Basis Function Kernel

157 samples
 60 predictors
 2 classes: 'M', 'R'

Pre-processing: centered, scaled
Resampling: Cross-Validated (10 fold, repeated 10 times)

Summary of sample sizes: 142, 142, 140, 142, 142, 141, ...

Resampling results across tuning parameters:

      C    ROC   Sens   Spec   ROC SD   Sens SD   Spec SD
      0.2   0.9   0.7   0.7   0.08   0.1   0.2
      0.5   0.9   0.8   0.8   0.06   0.1   0.2
      1     0.9   0.9   0.8   0.06   0.1   0.1
      2     0.9   0.9   0.8   0.05   0.1   0.2
      4     0.9   0.9   0.8   0.05   0.1   0.2
```

```
8   0.9  0.9   0.8   0.05   0.1    0.2
20  0.9  0.9   0.8   0.06   0.09   0.2
30  0.9  0.9   0.8   0.06   0.1    0.2
```

Tuning parameter 'sigma' was held constant at a value of 0.01234  
 ROC was used to select the optimal model using the largest value.  
 The final values used for the model were sigma = 0.01 and C = 8.

Also, a regularized discriminant analysis model was fit.

```
set.seed(825)
rdaFit <- train(Class ~ ., data = training,
                  method = "rda",
                  trControl = fitControl,
                  tuneLength = 4,
                  metric = "ROC")
rdaFit
```

Regularized Discriminant Analysis

```
157 samples
60 predictors
2 classes: 'M', 'R'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 10 times)

Summary of sample sizes: 142, 142, 140, 142, 142, 141, ...

Resampling results across tuning parameters:
```

gamma	lambda	ROC	Sens	Spec	ROC SD	Sens SD	Spec SD
0	0	0.8	0.8	0.8	0.1	0.1	0.2
0	0.3	0.8	0.8	0.8	0.1	0.1	0.2
0	0.7	0.8	0.8	0.8	0.1	0.1	0.2
0	1	0.9	0.8	0.8	0.1	0.1	0.2
0.3	0	0.9	0.9	0.8	0.07	0.1	0.2
0.3	0.3	0.9	0.9	0.8	0.08	0.1	0.1
0.3	0.7	0.9	0.9	0.8	0.07	0.1	0.2
0.3	1	0.9	0.9	0.8	0.08	0.1	0.2
0.7	0	0.9	0.9	0.7	0.08	0.1	0.2
0.7	0.3	0.9	0.9	0.7	0.09	0.1	0.2
0.7	0.7	0.9	0.9	0.7	0.09	0.1	0.2
0.7	1	0.9	0.9	0.7	0.09	0.1	0.2
1	0	0.7	0.7	0.6	0.1	0.2	0.2
1	0.3	0.7	0.7	0.6	0.1	0.2	0.2
1	0.7	0.7	0.7	0.6	0.1	0.2	0.2
1	1	0.7	0.7	0.7	0.1	0.2	0.2

ROC was used to select the optimal model using the largest value.  
 The final values used for the model were gamma = 0.3 and lambda = 0.

Given these models, can we make statistical statements about their performance differences? To do this, we first collect the resampling results using resamples.

```
resamps <- resamples(list(GBM = gbmFit3,
                           SVM = svmFit,
                           RDA = rdaFit))
resamps

Call:
resamples.default(x = list(GBM = gbmFit3, SVM = svmFit, RDA = rdaFit))

Models: GBM, SVM, RDA
Number of resamples: 100
Performance metrics: ROC, Sens, Spec
Time estimates for: everything, final model fit
```

```
summary(resamps)

Call:
summary.resamples(object = resamps)

Models: GBM, SVM, RDA
Number of resamples: 100

ROC
  Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
GBM 0.625  0.889  0.933 0.916  0.968   1    0
SVM 0.696  0.918  0.961 0.946  0.984   1    0
RDA 0.635  0.857  0.929 0.907  0.954   1    0

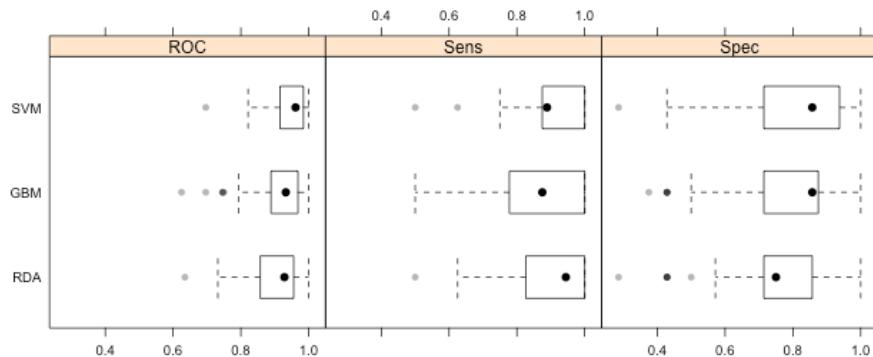
Sens
  Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
GBM 0.5  0.778  0.875 0.867      1    1    0
```

SVM	0.5	0.875	0.889	0.910	1	1	0
RDA	0.5	0.851	0.944	0.903	1	1	0
<b>Spec</b>							
GBM	0.375	0.714	0.857	0.782	0.875	1	0
SVM	0.286	0.714	0.857	0.815	0.906	1	0
RDA	0.286	0.714	0.750	0.761	0.857	1	0

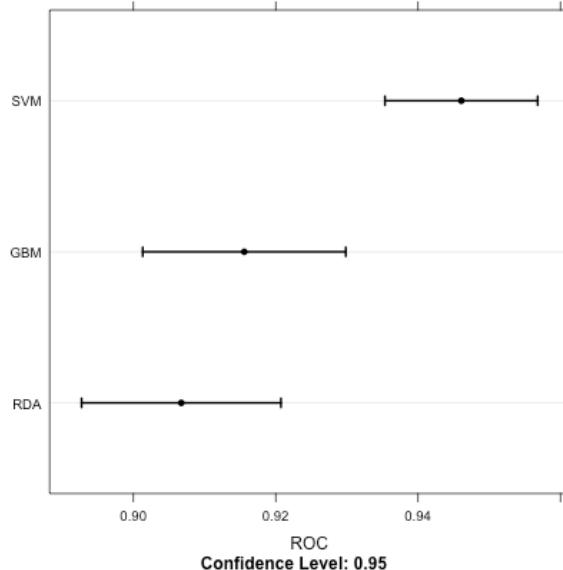
Note that, in this case, the option `resamples = "final"` should be used in the control objects.

There are several lattice plot methods that can be used to visualize the resampling distributions: density plots, box-whisker plots, scatterplot matrices and scatterplots of summary statistics. For example:

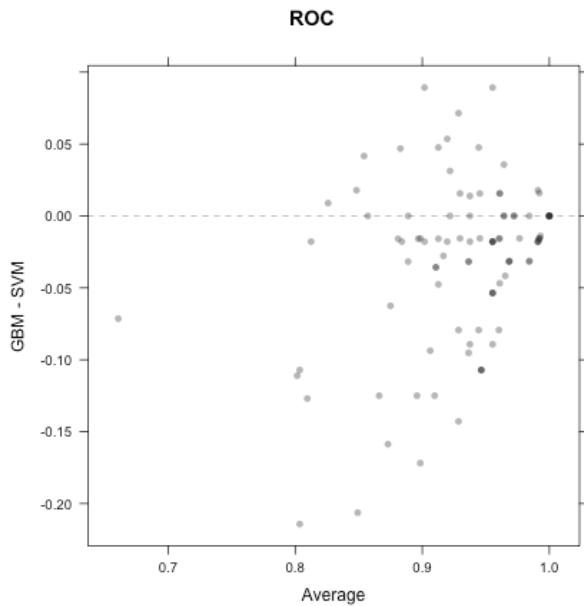
```
trellis.par.set(theme1)
bwplot(resamps, layout = c(3, 1))
```



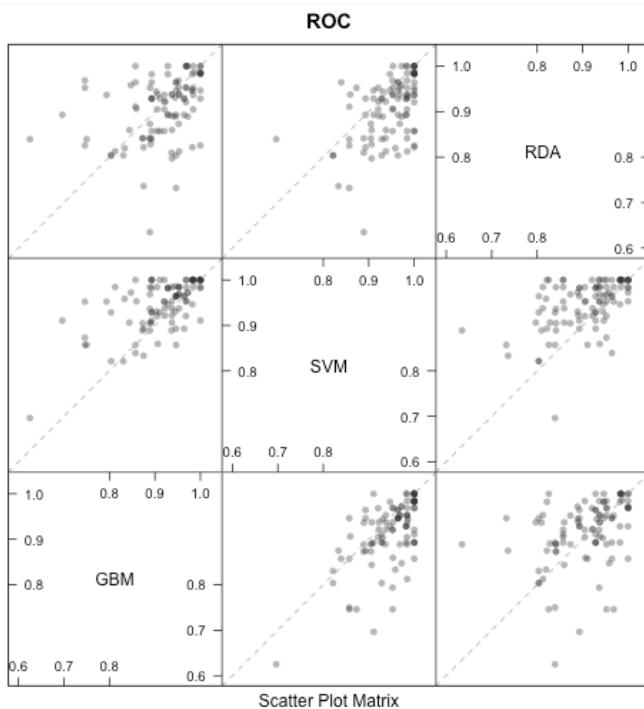
```
trellis.par.set(caretTheme())
dotplot(resamps, metric = "ROC")
```



```
trellis.par.set(theme1)
xyplot(resamps, what = "BlandAltman")
```



```
splom(resamps)
```



Other visualizations are available in `densityplot.resamples` and `parallel.resamples`

Since models are fit on the same versions of the training data, it makes sense to make inferences on the differences between models. In this way we reduce the within-resample correlation that may exist. We can compute the differences, then use a simple *t*-test to evaluate the null hypothesis that there is no difference between models.

```
difValues <- diff(resamps)
difValues

Call:
diff.resamples(x = resamps)

Models: GBM, SVM, RDA
Metrics: ROC, Sens, Spec
Number of differences: 3
p-value adjustment: bonferroni

summary(difValues)

Call:
summary.diff.resamples(object = difValues)

p-value adjustment: bonferroni
Upper diagonal: estimates of the difference
```

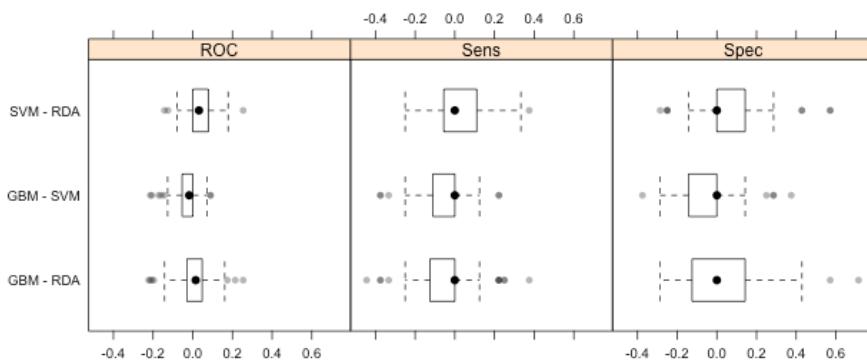
```
Lower diagonal: p-value for H0: difference = 0
```

ROC		
GBM	SVM	RDA
GBM	-0.03050	0.00885
SVM	1.81e-06	0.03935
RDA	0.824	5.02e-08

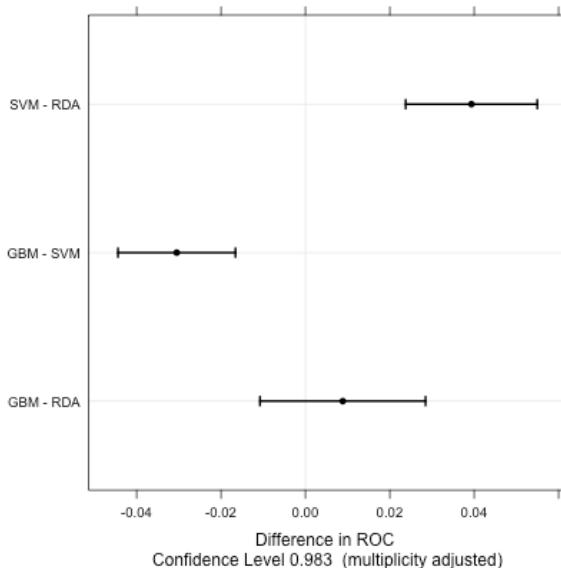
Sens		
GBM	SVM	RDA
GBM	-0.04333	-0.03611
SVM	0.000983	0.00722
RDA	0.042556	1.000000

Spec		
GBM	SVM	RDA
GBM	-0.0334	0.0211
SVM	0.05664	0.0545
RDA	0.71443	0.00414

```
trellis.par.set(theme1)
bwplot(difValues, layout = c(3, 1))
```



```
trellis.par.set(caretTheme())
dotplot(difValues)
```



## Fitting Models Without Parameter Tuning

In cases where the model tuning values are known, train can be used to fit the model to the entire training set without any resampling or parameter tuning. Using the `method = "none"` option in `trainControl` can be used. For example:

```
fitControl <- trainControl(method = "none", classProbs = TRUE)
set.seed(825)
gbmFit4 <- train(Class ~ ., data = training,
  method = "gbm",
  trControl = fitControl,
  verbose = FALSE,
  ## Only a single model can be passed to the
  ## function when no resampling is used:
  tuneGrid = data.frame(interaction.depth = 4,
```

```

n.trees = 100,
shrinkage = .1),
metric = "ROC")
gbmFit4

Stochastic Gradient Boosting

157 samples
60 predictors
2 classes: 'M', 'R'

No pre-processing
Resampling: None

```

Note that `plot.train`, `resamples`, `confusionMatrix.train` and several other functions will not work with this object but `predict.train` and others will:

```

predict(gbmFit4, newdata = head(testing))

[1] R R R R M M
Levels: M R

predict(gbmFit4, newdata = head(testing), type = "prob")

      M          R
1 0.4060186 0.5940
2 0.0301124 0.9699
3 0.0009584 0.9990
4 0.0399826 0.9600
5 0.9614988 0.0385
6 0.7496510 0.2503

```

- **Links**

[train Model List](#)

- **Topics**

- [Main Page](#)
- [Data Sets](#)
- [Visualizations](#)
- [Pre-Processing](#)
- [Data Splitting](#)
- [Miscellaneous Model Functions](#)
- [Model Training and Tuning](#)
- [train Model List](#)
- [train Models By Tag](#)
- [train Models By Similarity](#)
- [Using Custom Models](#)
- [Variable Importance](#)
- [Feature Selection](#)
- [Other Functions](#)
- [Parallel Processing](#)
- [Adaptive Resampling](#)

Created on Sat May 31 2014 using caret version 6.0-29 and R version 3.0.3 (2014-03-06).

## Testing Nested Models

- Two models are *nested* if both contain the same terms and one has at least one additional term.
- Example:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon \quad (1)$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \epsilon \quad (2)$$

- Model (1) is *nested within* model (2).
- Model (1) is the **reduced** model and model (2) is the **full** model.

## Testing Nested Models (cont'd)

- How do we decide whether the more complex (full) model contributes additional information about the association between  $y$  and the predictors?
- In example above, this is equivalent to testing  $H_0 : \beta_4 = \beta_5 = 0$  versus  $H_a : \text{at least one } \beta \neq 0$ .
- Test consists in comparing the SSE for the reduced model ( $SSE_R$ ) and the SSE for the complete model ( $SSE_C$ ).
- $SSE_R > SSE_C$  always so question is whether the drop in SSE from fitting the complete model is ‘large enough’.

## Testing Nested Models (cont'd)

- We use an  $F$ -test to compare nested models, one with  $k$  parameters (reduced) and another one with  $k + p$  parameters (complete or full).
- Hypotheses:  $H_0 : \beta_{k+1} = \beta_{k+2} = \dots = \beta_{k+p} = 0$  versus  $H_a$  : At least one  $\beta \neq 0$ .
- Test statistic: 
$$F = \frac{(SSE_R - SSE_C) / \text{\# of additional } \beta's}{SSE_C / [n - (k + p + 1)]}$$
- At level  $\alpha$ , we compare the  $F$ -statistic to an  $F_{\nu_1, \nu_2}$  from table, where  $\nu_1 = p$  and  $\nu_2 = n - (k + p + 1)$ .
- If  $F \geq F_{\alpha, \nu_1, \nu_2}$ , reject  $H_0$ .

## Testing Nested Models (cont'd)

- See Example 4.10 on page 233.
- Steps are:
  1. Fit complete model with  $k + p$   $\beta$ 's and get  $SSE_C$ .
  2. Fit reduced model with  $k$   $\beta$ 's and get  $SSE_R$ .
  3. Set up hypotheses and choose  $\alpha$  value.
  4. Compute  $F$ —statistic and compare to table  $F_{\alpha, \nu_1, \nu_2}$ .
- If test leads to rejecting  $H_0$ , then at least one of the additional terms in the model contributes information about the response.

## Testing Nested Models (cont'd)

- Parsimonious models are preferable to big models as long as both have similar predictive power.
- A parsimonious model is one with a small number of predictors.
- If models are not nested, cannot use the  $F$ –test above to choose between one and another. Must rely on other sample statistics such as  $R_a^2$  and  $RMSE$ .
- In the end, choice of model is subjective.

## Lecture 11: AV plots, hypothesis testing and nested models

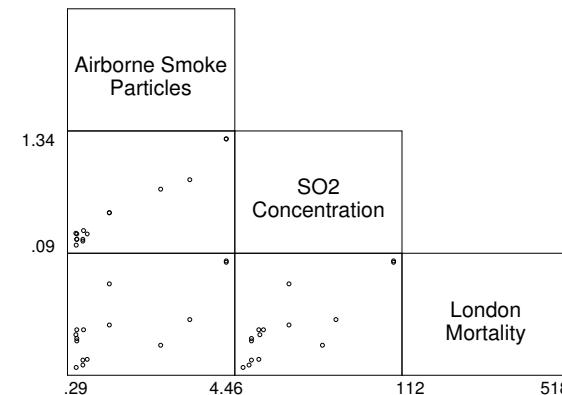
Sandy Eckel  
seckel@jhsph.edu

9 May 2008

1

## Another Example: Mortality

### Smoke, pollution & London mortality data



2

## Mortality Example: Model

Let:

- $Y$  = the daily mortality for London (*deaths*)
- $X_1$  = airborne smoke particles ( $\text{mg/m}^3$ ) (*smoke*)
- $X_2$  =  $\text{SO}_2$  (ppm) (*so2*)

### Model:

- Systematic:  $Y_i = \beta_0 + \beta_1(X_{1i}-2) + \beta_2(X_{2i}-5) + \varepsilon_i$
- Random:  $\varepsilon_i \sim N(0, \sigma^2)$
- Mortality is a linear function of the concentration of airborne smoke particles *AND* the  $\text{SO}_2$  level

3

## Mortality Example: Results

### Model:

$$E(Y | X) = \beta_0 + \beta_1(X_1 - 2) + \beta_2(X_2 - 5)$$

Source	SS	df	MS	Number of obs	=	15
Model	205097.531	2	102548.765	F( 2, 12 )	=	36.57
Residual	33654.2025	12	2804.51687	Prob > F	=	0.0000
Total	238751.733	14	17053.6952	R-squared	=	0.8590
				Adj R-squared	=	0.8355
				Root MSE	=	52.958

deaths	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
smokecenter	-220.3244	58.14314	-3.79	0.003	-347.0074 -93.64135
so2center	1051.816	212.5959	4.95	0.000	588.6096 1515.023
(Intercept)	174.7703	29.16174	5.99	0.000	111.2323 238.3083

4

## Mortality Example

### Inference: Overall F test

- Overall F-Test:
  - Are ANY of the covariates significant?
- $H_0: \beta_1 = \beta_2 = 0;$
- $F_{obs}: (2,12) = 36.57;$
- p-value = 0.0000
- Decision: At least one of the  $\beta$ 's are nonzero

5

## Mortality Example

### Coefficient inference: individual 95% C.I. & t-tests

- |   |   |
|---|---|
| $\beta_0$ <ul style="list-style-type: none"> <li>▪ <math>b_0 = 174.8</math><br/>95% CI: (111.2, 238.3)</li> <li>▪ <math>H_0: \beta_0 = 0</math></li> <li>▪ <math>t_{obs}: (12) = 5.99</math></li> <li>▪ p-value = 0.000</li> </ul>    | $\beta_2$ <ul style="list-style-type: none"> <li>▪ <math>b_2 = 1051.8</math><br/>95% CI: (588.6, 1515.0)</li> <li>▪ <math>H_0: \beta_2 = 0</math></li> <li>▪ <math>t_{obs}: (12) = 4.95</math></li> <li>▪ p-value = 0.000<br/>means p-value &lt; 0.001</li> </ul> |
| $\beta_1$ <ul style="list-style-type: none"> <li>▪ <math>b_1 = -220.3</math><br/>95% CI: (-347.0, -93.6)</li> <li>▪ <math>H_0: \beta_1 = 0</math></li> <li>▪ <math>t_{obs}: (12) = -3.79</math></li> <li>▪ p-value = 0.003</li> </ul> |   |

6

## Mortality Example

### Parameter Estimates Interpretation

- $b_0$ : when smoke particles and  $SO_2$  are at their average levels, (2 mg/m<sup>3</sup>, and 0.5 ppm respectively), the estimated mean number of deaths is 174.8 / day
- $b_1$ : the estimated mean mortality is 22 deaths/day lower on days when particles are 0.1 mg/m<sup>3</sup> higher *if SO<sub>2</sub> is unchanged*
- $b_2$  : (*You do!*)

7

## Mortality Example

### Association between x and y

- The estimate for airborne smoke particles is  $b_1 = -220$ , implying that smoke particles and mortality have a *negative* relationship
  - i.e. an *increase* in smoke particles is associated with a *decrease* in mortality, after adjusting for  $SO_2$  levels.

8

## Mortality Example Negative Association??

- BUT WAIT!
- Look at the plot of *deaths vs smoke* presented previously. Shouldn't the relationship be *positive* instead?!
- Let's run Simple Linear Regressions (SLRs) of mortality on smoke & SO<sub>2</sub> and see what we get

9

## SLR Models

- $Y$  = the daily mortality for London (*deaths*)
- $X_1$  = airborne smoke particles (mg/m<sup>3</sup>) (*smoke*)
- $X_2$  = SO<sub>2</sub> (ppm) (*so2*)
- Smoke:
  - 1)  $Y_i = \beta_0 + \beta_1(X_1 - 2) + \varepsilon_i$
  - 2)  $\varepsilon_i \sim N(0, \sigma^2)$
- SO<sub>2</sub>:
  - 1)  $Y_i = \beta_0^* + \beta_1^*(X_2 - .5) + \varepsilon_i^*$
  - 2)  $\varepsilon_i^* \sim N(0, \sigma^{2*})$

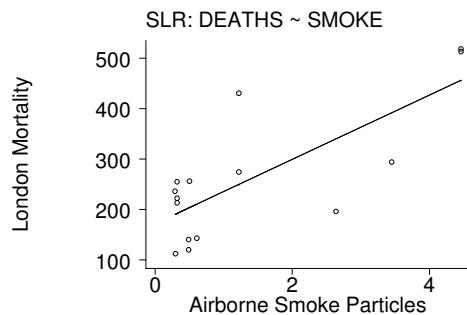
10

## SLR: Deaths ~ Smoke

Parameter Estimates:  $b_0 = 299.3$   
 $b_1 = 63.8$  (*is positive?!!*)

Amount of variation described:  $R^2 = SSM / SST = 57\%$

Residual Variability left over, (undescribed by this SLR):  
 $SSE = 1023002.216$



11

## SLR: Death ~ SO<sub>2</sub>

Parameter Estimates:  $b_0 = 256.2$   
 $b_1 = 272.2$

Amount of variation described:  $R^2 = SSM / SST = 69\%$

Residual Variability left over, (undescribed by this SLR):  
 $SSE = 73924.6211$



12

## Confounding in this Example

Recall our parameter interpretations:

- $\beta_1$  = Expected change in mortality on days when particles are 0.1 mg/m<sup>3</sup> higher *if SO<sub>2</sub> is unchanged*
- Suppose we examine the relationship between smoke particle concentrations and SO<sub>2</sub> levels, (SLR):

13

## SLR: Smoke ~ SO<sub>2</sub>



14

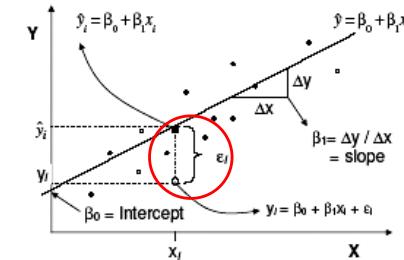
## Confounding

- Smoke particle concentrations and SO<sub>2</sub> levels are highly related! How can we talk about *changing* smoke particle concentrations while *leaving SO<sub>2</sub> levels unchanged??*
- This is 'confounding'!
  - both covariates are related to the outcome and to each other
- Confounding is the reason we found differences between the SLR models and the MLR model
- We'll visualize this relationship using 'Added Variable Plots'

15

## Recall Residuals: part "left over"

- Residuals are deviations (what's 'left over') in the response (Y) after removing what was expected given the predictor (X)
- The residuals are the part of Y that can't be predicted by X!



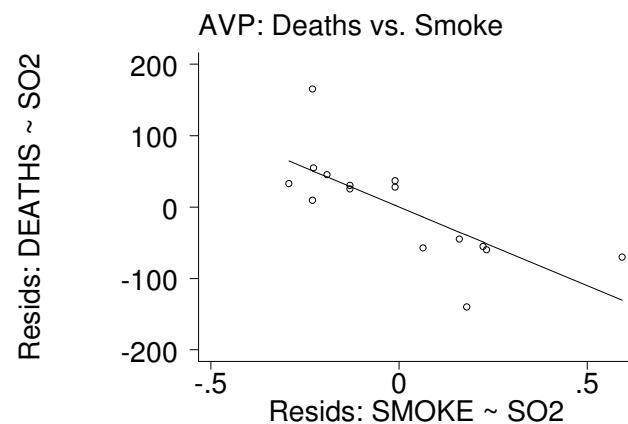
16

## Adjusted Variable Plots Idea

- Explain all the signal we can in London daily mortality using SO<sub>2</sub> levels
- Explain all the signal we can in smoke particle concentrations using SO<sub>2</sub> levels
- Explain everything that's 'left over' in mortality with everything that's 'left over' in smoke particle concentrations. The slope of this line will be the MLR coefficient!

17

## Mortality Example Adjusted Variable Plot



18

## Recipe for AV plot

Recipe for obtaining the MLR slope for X<sub>1</sub> from an AV plot (adjusted for X<sub>2</sub>):

1. Regress Y on X<sub>2</sub>, save residuals as: R<sub>Y|X<sub>2</sub></sub>
2. Regress X<sub>1</sub> on X<sub>2</sub>, save residuals as: R<sub>X<sub>1</sub>|X<sub>2</sub></sub>
3. Plot R<sub>Y|X<sub>2</sub></sub> vs R<sub>X<sub>1</sub>|X<sub>2</sub></sub> (Adjusted Variable Plot )

Regress R<sub>Y|X<sub>2</sub></sub> on R<sub>X<sub>1</sub>|X<sub>2</sub></sub>:

$$R_{Y|X_2} = \beta_0^* + \beta_1^* R_{X_1|X_2} + \epsilon$$

19

## Notes on AV Plots

- β<sub>1</sub>\* is identical to the coefficient of X<sub>1</sub> from an MLR of Y on X<sub>1</sub> and X<sub>2</sub>
- β<sub>0</sub>\* (intercept) is always 0
- The AV Plot display may be misleading if Y and/or X<sub>1</sub> are not linearly related to the other predictors

20

## AV Plot Recipe for Mortality Example

- Regress deaths on (centered)  $\text{SO}_2$ , save residuals
  - Removes the effects of  $\text{SO}_2$  on mortality
$$\text{Deaths} = 272 + 256 \text{SO}_{2c} + R_{Y|X2}$$
- Regress Smoke on  $\text{SO}_2$  (both centered), save residuals
  - Removes the effects of  $\text{SO}_2$  on smoke particles
$$\text{Smoke}_c = -.44 + 3.6 \text{SO}_{2c} + R_{X1|X2}$$
- Regress  $R_{Y|X2}$  on  $R_{X1|X2}$ 
  - regress deaths *adjusted for  $\text{SO}_2$*  on smoke particles *adjusted for  $\text{SO}_2$*
- $R_{Y|X2} = 0.0 - 220 R_{X1|X2}$

21

## AV plot interpretation

- Parameter from this last regression:  $\beta_1^* = -220$  is the same as the related parameter from the MLR of deaths on smoke particles *and*  $\text{SO}_2$
- $$E(\text{Deaths}) = \beta_0 + \beta_1(\text{smoke}-2) + \beta_2(\text{SO}_2-.5)$$
$$= 174.8 - 220 (\text{smoke} - 2) + 1052 (\text{SO}_2 - 0.5)$$
- This helps in our interpretations of  $\beta_1$ : the effect of airborne smoke particles on daily mortality after having removed (adjusted for) the effects of  $\text{SO}_2$ 
  - This is what is usually meant by the term 'adjustment'

22

## MLR and Scientific Inference

- The **single most important idea** today may be the realization that MLR can shift interpretations markedly!
- From SLR of the air pollution data:  
$$E(\text{Deaths}) = 299 + 64(\text{smoke}-2)$$
  - Expected deaths **increase** by an estimated 64 per  $\text{mg}/\text{m}^3$  increase in British smoke

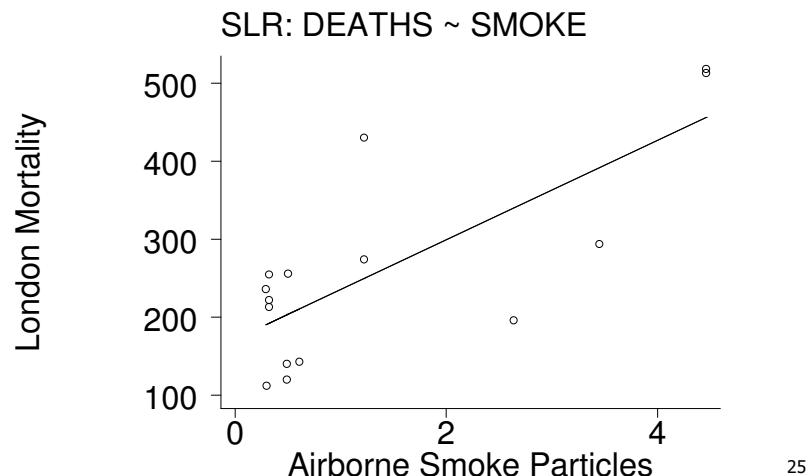
23

## MLR and Scientific Inference

- From MLR of the air pollution data:  
$$E(\text{Deaths}) = 174.8 - 220(\text{smoke}-2) + 1052(\text{SO}_2-.5)$$
  - *Controlling for  $\text{SO}_2$* , expected deaths **decrease** 220 per  $\text{mg}/\text{m}^3$  of British smoke
- Interpretation and value of a regression coefficient depends critically on what other variables are in the model !!

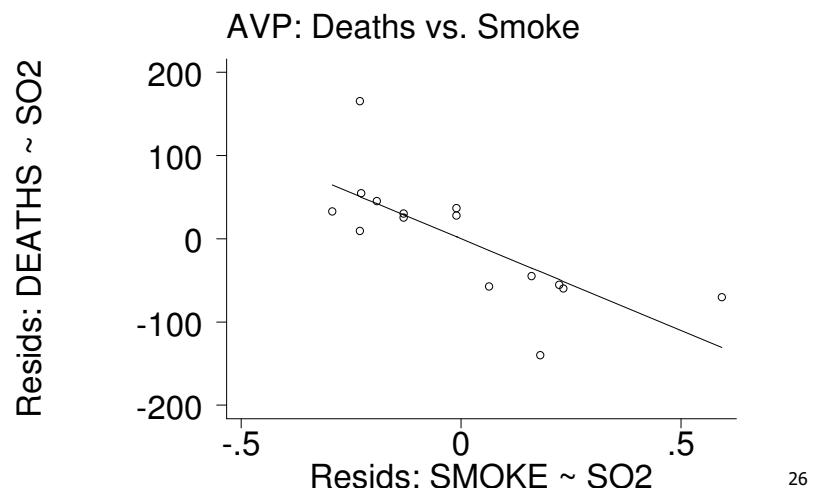
24

## Simple Linear Regression



25

## Multiple Linear Regression



26

## Types of predictors in regression

- primary predictor
  - always in model
- other predictor(s)
  - can we improve prediction after adjusting for primary predictor?
  - interaction may be a component here
- potential confounder(s) (i.e., demographics)
  - only important if they change the effect of the primary predictor
  - commonly: age, gender, SES, race, etc...

27

## Nested models

Definition: One model is nested within another if the **parent model** contains the 'original' set of variables and is *nested* within the **extended model** that contains the original set of variables plus additional variables

28

## Nested models

### Deciding whether to include variables

If the 'new variable(s)' are:

- another predictor(s)
  - assess with t-test in extended model if single variable
  - assess with F-test if two or more variables
- potential confounder(s)
  - compare CI of primary predictor in parent model to see whether new estimate of primary predictor coefficient is significantly different

29

## Dataset

- Class health dataset
  - Outcome: number of credits
  - Primary predictor
    - housing (on or off campus)
  - Other predictors
    - health status (good/excellent or fair/poor)
    - year in school

30

## Models

### ▪ Parent Model (Model 1)

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 (\text{Housing}_i)$$

1 if on-campus  
0 if off-campus

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
housing	.1666667	.6761572	0.25	0.807	-1.228853 1.562187
(Intercept)	16.2	.5135783	31.54	0.000	15.14003 17.25997

### ▪ Extended Model (Model 2)

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 (\text{Housing}_i) + \hat{\beta}_2 (\text{Healthgood}_i)$$

1 if excellent/good  
0 if fair/poor

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
housing	.1541237	.6860262	0.22	0.824	-1.26503 1.573277
healthgood	.4139175	.7124214	0.58	0.567	-1.059838 1.887673
(Intercept)	15.9366	.6904955	23.08	0.000	14.5082 17.365

31

## Comparing models 1 and 2

- Model 1 is nested in model 2
- Model 2 contains only one extra variable (healthgood), so use a t-test to decide whether to include healthgood
  - p=0.567 >  $\alpha=0.05$  tests  $H_0: \beta_2=0$
  - Fail to reject  $H_0$
  - Conclude model 2 is no better than model 1

32

## What if we add more than one variable?

- The t-test on each row only tests that variable *in the presence of everything else in the model*
- When more than one variable is added at a time, the t-test is not sufficient
  - The t-test only tests one variable at a time
  - Use the F-test instead to compare nested models that differ by more than one variable

33

## When would more than one variable need to be added??

- Many modeling scenarios require adding more than one variable at once to go from the parent model to the extended model
- Commonly occurs when categorical variable needs to be added

34

## Why do we need to specially code a categorical predictor?

- A categorical predictor (such as year in program) cannot be added as a single variable
  - If we add year (1, 2, 3, or 4) to the model in its original form, then software thinks it is a continuous predictor
  - As a continuous predictor, the difference in mean number of credits taken would be assumed to change by a constant amount for each additional year

35

## Correct coding of a categorical predictor

- A categorical predictor should always be recoded as a set of dummy variables
  - Choose one category as the reference group
  - For each **other** category, create a dummy variable for membership in that category
  - You can have **R** do this automatically for you with the command **factor (mycatvar)** within your linear regression command

36

## Example

- Year1 = reference group (no dummy variable for this group)
- **Year2** = 1 for those in year 2, 0 else
- **Year34** = 1 for those in yr 3/4, 0 else
  - very few observations, so categories were combined
- In in year 3: Year2=0, Year34=1
- For a first year: Year2=0, Year34=0

37

## Model 3

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1(\text{Housing}_i) + \hat{\beta}_2(\text{Year2}_i) + \hat{\beta}_3(\text{Year34}_i)$$

	credits	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
+	housing	-1.402299	.8537457	-1.64	0.115	-3.172859 .3682613
	year2	.7068966	.7215468	0.98	0.338	-.7894999 2.203293
	year34	-2.10197	1.087462	-1.93	0.066	-4.357228 .1532874
(Intercept)		17.34483	.9268436	18.71	0.000	15.42267 19.26698

- We cannot evaluate Year using the t-test for each row, because two variables are needed to define Year and the t tests are separate
- We must use an F-test to evaluate Year by comparing the residual sums of squares (RSS) in the parent model and in the nested model.

38

## Comparing model RSS and Residual df

### PARENT: MODEL 1

Source	SS	df	MS	RSS <sub>parent</sub>	Residual df <sub>parent</sub>
Model	.176282088	1	176282088		
Residual	69.6333335	24	2.9013889		
Total	69.8096156	25	2.79238462		

Number of obs = 26  
 F( 1, 24) = 0.06  
 Prob > F = 0.8074  
 R-squared = 0.0025  
 Adj R-squared = -0.0390  
 Root MSE = 1.7033

### EXTENDED: MODEL 3

Source	SS	df	MS	RSS <sub>extended</sub>	Residual df <sub>extended</sub>
Model	19.9853465	3	6.66178216		
Residual	49.8242691	22	2.26473951		

Number of obs = 26  
 F( 3, 22) = 2.94  
 Prob > F = 0.0555  
 R-squared = 0.2863  
 Adj R-squared = 0.1890  
 Root MSE = 1.5049

39

## The F-test for nested models

H<sub>0</sub>: all new  $\beta$ 's=0 in population

H<sub>A</sub>: at least one new  $\beta$  is not 0 in population

Numerator of F-statistic:

$$(RSS_{\text{parent}} - RSS_{\text{extended}}) / (\text{number variables added})$$

Denominator of F-statistic:

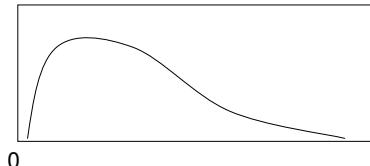
$$RSS_{\text{extended}} / (\text{residual df}_{\text{extended}})$$

$$F_{\text{obs}} = \frac{(69.6 - 49.8)}{\frac{49.8}{22}} = 4.4$$

40

## The F table

- Recall: the F distribution is very similar to the  $\chi^2$  distribution



- F distribution is automatically 2-sided (like  $\chi^2$ )
- df change the shape of the F distribution (like  $\chi^2$ ), but now there are two sets of df: the numerator df and the denominator df

41

## The F table

- numerator df: # of variables added = 2
- denominator df: residual df<sub>extended</sub> = 22
- Using  $\alpha=0.05$ , find  $F_{cr}$ 
  - Find quantile in R, using appropriate degrees of freedom  
`> qf(.05, 2, 22, lower.tail=F)`  
[1] 3.443357
- $F_{cr}=3.44 < F_{obs}=4.4$
- So, our p-value <  $\alpha$

42

## Conclusion using the F-test

- Reject  $H_0$ : conclude that adding year improves prediction after adjusting for housing
  - Notice: both individual t tests were not statistically significant, but F test was still significant
  - Must always use F test to evaluate multiple X's at once

43

## The F test: notes

- The F test *can* be used to compare any two nested models
- If only one variable is added, it's easier to compare the models using the t test for that variable
  - $t^2=F$  if one variable is added

44

## The F test: how to in R

- Fit parent model  
`fit.parent <- lm(y ~ x1 + x2)`
- Fit the extended model (parent model is nested within the extended model)  
`fit.extend <- lm(y ~ x1 + x2 + x3 + x4)`
- Perform the F-test  
`print(anova(fit.parent, fit.extend))`

### Example output:

Analysis of Variance Table

```
Model 1: y ~ x1 + x2
Model 2: y ~ x1 + x2 + x3 + x4
  Res.Df   RSS Df Sum of Sq    F Pr(>F)
1     650 110.65
2     648 109.51  2      1.14 3.3718 0.03493 *
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

45

## Nested Models

- Comparing nested models
  - 1 new variable: use t test for that variable
  - 2+ new variables: use F test
- Categorical predictor
  - set one group as reference
  - create dummy variable for other groups
  - include/exclude all dummy variables
  - evaluate categorical predictor with F test

46

## Lecture 11 Summary

- Hypothesis tests in linear regression
  - Overall F-test
  - Individual coefficient 95% CI and t-tests
- F-tests for nested models
- AV plots
  - visualizing the relationship between the outcome and a continuous predictor after adjusting for the effects of a third variable

47

# Matrix decompositions for regression analysis

Douglas Bates

2010-09-07 Tue

## Contents

<b>1</b>	<b>Matrix decompositions</b>	<b>1</b>
1.1	Orthogonal matrices . . . . .	1
1.1.1	Preserving lengths . . . . .	1
1.1.2	The determinant of an orthogonal matrix . . . . .	1
1.2	The QR decomposition . . . . .	1
<b>2</b>	<b>Least squares estimation</b>	<b>2</b>
2.1	Spherical normal distributions and least squares estimators . . . . .	2
2.2	Comparison to the usual text-book formulas . . . . .	3
2.3	R functions related to the QR decomposition . . . . .	3
<b>3</b>	<b>Related matrix decompositions</b>	<b>4</b>
3.1	The Cholesky decomposition . . . . .	4
3.2	Evaluation of the Cholesky decomposition . . . . .	4
3.3	The singular value decomposition . . . . .	5

# 1 Matrix decompositions

## 1.1 Orthogonal matrices

An *orthogonal*  $n \times n$  matrix,  $\mathbf{Q}$  has the property that its transpose is its inverse. That is,

$$\mathbf{Q}'\mathbf{Q} = \mathbf{Q}\mathbf{Q}' = \mathbf{I}_n$$

That is, the columns of  $\mathbf{Q}$  must be orthogonal to each other and all have unit length, and the same for the rows.

Two consequences of this property are that the transformation from  $\mathbb{R}^n$  to  $\mathbb{R}^n$  determined by  $\mathbf{Q}$  or  $\mathbf{Q}'$  preserves lengths and that the determinant of  $\mathbf{Q}$  is  $\pm 1$ .

### 1.1.1 Preserving lengths

For any  $\mathbf{x} \in \mathbb{R}^n$

$$\|\mathbf{Q}\mathbf{x}\|^2 = (\mathbf{Q}\mathbf{x})'\mathbf{Q}\mathbf{x} = \mathbf{x}'\mathbf{Q}'\mathbf{Q}\mathbf{x} = \mathbf{x}'\mathbf{x} = \|\mathbf{x}\|^2$$

Thus the transformation determined by  $\mathbf{Q}$  or by  $\mathbf{Q}'$  must be a *rigid* transformation, composed of reflections or rotations.

### 1.1.2 The determinant of an orthogonal matrix

The determinants of  $n \times n$  matrices  $\mathbf{A}$  and  $\mathbf{B}$ , written  $|\mathbf{A}|$  and  $|\mathbf{B}|$ , are scalars with the property that

$$|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|$$

Furthermore, the determinant of a *diagonal* matrix or a *triangular* matrix is simply the product of its diagonal elements. Also,  $|\mathbf{A}| = |\mathbf{A}'|$ .

From these properties we can derive

$$1 = |\mathbf{I}_n| = |\mathbf{QQ}'| = |\mathbf{Q}||\mathbf{Q}'| = |\mathbf{Q}|^2 \Rightarrow |\mathbf{Q}| = \pm 1.$$

As described on the Wikipedia page, the determinant of  $\mathbf{Q}$  is the volume of the parallelepiped spanned by the columns (or by the rows) of  $\mathbf{Q}$ . We know that the columns of  $\mathbf{Q}$  are *orthonormal* hence they span a unit volume. The sign indicates whether the transformation preserves orientation. In two dimensions a rotation preserves orientation and a reflection reverses the orientation.

## 1.2 The QR decomposition

An  $n \times p$  matrix  $\mathbf{X}$  has a QR decomposition consisting of an orthogonal  $n \times n$  matrix  $\mathbf{Q}$  and an  $n \times p$  matrix  $\mathbf{R}$  that is zero below the main diagonal. In the cases we consider in regression analysis the matrix  $\mathbf{X}$  is the model matrix where  $n$  is the number of observations,  $p$  is the number of coefficients (or *parameters*) in the model and  $n \geq p$ . Our basic model is that the observed responses  $\mathbf{y}$  are the realization of a vector-valued random variable  $\mathcal{Y}$  with distribution

$$\mathcal{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}_0, \sigma^2 \mathbf{I}_n)$$

for some unknown value  $\boldsymbol{\beta}_0$  of the coefficient vector  $\boldsymbol{\beta}$ .

The QR decomposition of the model matrix  $\mathbf{X}$  is written

$$\mathbf{X} = \mathbf{Q} \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix} = [\mathbf{Q}_1 \quad \mathbf{Q}_2] \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix} = \mathbf{Q}_1 \mathbf{R} \quad (1)$$

where  $\mathbf{R}$  is a  $p \times p$  upper triangular matrix,  $\mathbf{Q}_1$  is the first  $p$  columns of  $\mathbf{Q}$  and  $\mathbf{Q}_2$  is the last  $n - p$  columns of  $\mathbf{Q}$ .

That fact that matrices  $\mathbf{Q}$  and  $\mathbf{R}$  must exist is proved by construction. The matrix  $\mathbf{Q}$  is the product of  $p$  *Householder reflections* (see the Wikipedia page for the QR decomposition). The process is similar to the Gram-Schmidt orthogonalization process, but more flexible and numerically stable. If the diagonal elements of  $\mathbf{R}$  are all non-zero (in practice this means that none of them is very small in absolute value) then  $\mathbf{X}$  has *full column rank* and the columns of  $\mathbf{Q}_1$  form an *orthonormal basis* of the *column span* of  $\mathbf{X}$ .

The implementation of the QR decomposition in R guarantees that any near-zero elements on the diagonal of  $\mathbf{R}$  are rearranged by column permutation to occur in the trailing columns. That is, if the rank of  $\mathbf{X}$  is  $k < p$  then the first  $k$  columns of  $\mathbf{Q}$  form an orthogonal basis for the column span of  $\mathbf{X}\mathbf{P}$  where  $\mathbf{P}$  is a  $p \times p$  permutation matrix.

Our text often mentions rank-deficient cases where  $\text{rank}(\mathbf{X}) = k < p$ . In practice these occur rarely because the process of building the model matrix in R involves a considerable amount of analysis of the model formula to remove the most common cases of rank deficiency. Nevertheless, rank deficiency can occur and is detected and handled in the `lm` function in R.

## 2 Least squares estimation

The reason that we are interested in a QR decomposition of  $\mathbf{X}$  is because the probability model is linked to geometric properties of the response  $\mathbf{y}$  and the column span of  $\mathbf{X}$ .

### 2.1 Spherical normal distributions and least squares estimators

A distribution of the form

$$\mathcal{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}_0, \sigma^2 \mathbf{I}_n) \quad (2)$$

is called a *spherical normal* distribution because the contours of constant probability density are spheres centered at  $\mathbf{X}\boldsymbol{\beta}_0$ . In other words, the probability density function of  $\mathcal{Y}$ , evaluated at  $\mathbf{y}$  is determined by the distance,  $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0\|^2$  with larger distances corresponding to lower densities.

The *maximum likelihood estimate*,  $\hat{\boldsymbol{\beta}}$  of the coefficient vector,  $\boldsymbol{\beta}$ , is

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \quad (3)$$

These estimates are also called the *least squares estimates* of  $\boldsymbol{\beta}$ .

Because multiplication by an orthogonal matrix like  $\mathbf{Q}'$  preserves lengths we can write

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = \arg \min_{\boldsymbol{\beta}} \|\mathbf{Q}'\mathbf{y} - \mathbf{Q}'\mathbf{X}\boldsymbol{\beta}\|^2 = \arg \min_{\boldsymbol{\beta}} \|\mathbf{c}_1 - \mathbf{R}\boldsymbol{\beta}\|^2 + \|\mathbf{c}_2\|^2 \quad (4)$$

where  $\mathbf{c}_1 = \mathbf{Q}'\mathbf{y}$  is the first  $p$  elements of  $\mathbf{Q}\mathbf{y}$  and  $\mathbf{c}_2 = \mathbf{Q}'\mathbf{y}$  is the last  $n - p$  elements. If  $\text{rank}(\mathbf{X}) = p$  then  $\text{rank}(\mathbf{R}) = p$  and  $\mathbf{R}^{-1}$  exists and we can write  $\hat{\boldsymbol{\beta}} = \mathbf{R}^{-1}\mathbf{c}_1$  (although you don't actually calculate  $\mathbf{R}^{-1}$  to solve the triangular linear system  $\mathbf{R}\hat{\boldsymbol{\beta}} = \mathbf{c}_1$  for  $\hat{\boldsymbol{\beta}}$ ).

In a model fit by the `lm` or `aov` functions in R there is a component `e$ffects` which is  $\mathbf{Q}'\mathbf{y}$ . The component `$qr` is a condensed form of the QR decomposition of the model matrix  $\mathbf{X}$ . The matrix  $\mathbf{R}$  is embedded in there but the matrix  $\mathbf{Q}$  is a virtual matrix represented as a product of Householder reflections and not usually evaluated explicitly.

## 2.2 Comparison to the usual text-book formulas

Most text books state that the least squares estimates are

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad (5)$$

giving the impression that  $\hat{\boldsymbol{\beta}}$  is calculated this way. It isn't.

If you substitute  $\mathbf{X} = \mathbf{Q}_1\mathbf{R}$  in 5 you get

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'\mathbf{Q}'_1\mathbf{y} = \mathbf{R}^{-1}(\mathbf{R}')^{-1}\mathbf{R}'\mathbf{Q}'_1\mathbf{y} = \mathbf{R}^{-1}\mathbf{Q}'_1\mathbf{y},$$

our previous result.

Whenever you see  $\mathbf{X}'\mathbf{X}$  in a formula you should mentally replace it by  $\mathbf{R}'\mathbf{R}$  and similarly replace  $(\mathbf{X}'\mathbf{X})^{-1}$  by  $\mathbf{R}^{-1}(\mathbf{R}')^{-1}$  then see if you can simplify the result.

For example, the variance of the least squares estimator  $\hat{\boldsymbol{\beta}}$  is

$$(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X}) = \sigma^2\mathbf{R}^{-1}(\mathbf{R}^{-1})'$$

The R function `chol2inv` calculates  $\mathbf{R}^{-1}(\mathbf{R}^{-1})'$  directly from  $\mathbf{R}$  (not a big deal in most cases but when  $p$  is very large it should be faster and more accurate than evaluating  $\mathbf{R}^{-1}$  explicitly). The determinant of  $\mathbf{X}'\mathbf{X}$  is

$$|\mathbf{X}'\mathbf{X}| = |\mathbf{R}'\mathbf{R}| = |\mathbf{R}|^2 = \left( \prod_{i=1}^p r_{i,i} \right)^2$$

The fitted values  $\hat{\mathbf{y}}$  are  $\mathbf{Q}_1\mathbf{Q}'_1\mathbf{y}$  and thus the *hat matrix* (which puts a "hat" on  $\mathbf{y}$  by transforming it to  $\hat{\mathbf{y}}$ ) is the  $n \times n$  matrix  $\mathbf{Q}_1\mathbf{Q}'_1$ . Often we are interested in the diagonal elements of the hat matrix, which are the sums of the squares of rows of  $\mathbf{Q}_1$ . (In practice you don't want to calculate the entire  $n \times n$  hat matrix just to get the diagonal elements when  $n$  could be very large.)

The residuals,  $\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}}$ , are calculated as  $\hat{\mathbf{e}} = \mathbf{Q}_2\mathbf{Q}'_2\mathbf{y}$ .

The matrices  $\mathbf{Q}_1\mathbf{Q}'_1$  and  $\mathbf{Q}_2\mathbf{Q}'_2$  are *projection matrices*, which means that they are symmetric and *idempotent*. (A square matrix  $\mathbf{A}$  is idempotent if  $\mathbf{AA} = \mathbf{A}$ .) When  $\text{rank}(\mathbf{X}) = p$ , the hat matrix  $\mathbf{Q}_1\mathbf{Q}'_1$  projects any vector in  $\mathbb{R}^n$  onto the column span of  $\mathbf{X}$ . The other projection,  $\mathbf{Q}_2\mathbf{Q}'_2$ , is onto the subspace orthogonal to the column span of  $\mathbf{X}$  (see the figure on the front cover of the text).

## 2.3 R functions related to the QR decomposition

As mentioned above, every time you fit a linear model with `lm` or `aov` or `lm.fit`, the returned object contains a `$qr` component. This is a condensed form of the decomposition, only slightly larger than  $\mathbf{X}$  itself. Its class is "qr".

There are several extractor functions for a "qr" object: `qr.R()`, `qr.Q()` and `qr.X()`, which regenerates the original matrix. By default `qr.Q()` returns the matrix called  $\mathbf{Q}_1$  above with  $p$

columns but you can specify the number of columns desired. Typical alternative choices are  $n$  or  $\text{rank}(\mathbf{X})$ .

The `$rank` component of a "qr" object is the computed rank of  $\mathbf{X}$  (and, hence, of  $\mathbf{R}$ ). The `$pivot` component is the permutation applied to the columns. It will be  $1:p$  when  $\text{rank}(\mathbf{X}) = p$  but when  $\text{rank}(\mathbf{X}) < p$  it may be other than the identity permutation.

Several functions are applied to a "qr" object and a vector or matrix. These include `qr.coef()`, `qr.qy()`, `qr.qty()`, `qr.resid()` and `qr.fitted()`. The `qr.qy()` and `qr.qty()` functions multiply an  $n$ -vector or an  $n \times m$  matrix by  $\mathbf{Q}$  or  $\mathbf{Q}'$  without ever forming  $\mathbf{Q}$ . Similarly, `qr.fitted()` creates  $\mathbf{Q}_1\mathbf{Q}'_1\mathbf{x}$  and `qr.resid()` creates  $\mathbf{Q}_2\mathbf{Q}'_2\mathbf{x}$  without forming  $\mathbf{Q}$ .

The `is.qr()` function tests an object to determine if it is of class "qr".

### 3 Related matrix decompositions

#### 3.1 The Cholesky decomposition

The Cholesky decomposition of a positive definite symmetric matrix, which means a  $p \times p$  symmetric matrix  $\mathbf{A}$  such that  $\mathbf{x}'\mathbf{A}\mathbf{x} > 0$  for all non-zero  $\mathbf{x} \in \mathbb{R}^p$  is of the form

$$\mathbf{A} = \mathbf{R}'\mathbf{R} = \mathbf{L}\mathbf{L}'$$

where  $\mathbf{R}$  is an upper triangular  $p \times p$  matrix and  $\mathbf{L} = \mathbf{R}'$  is lower triangular. The two forms are the same decomposition: it is just a matter of whether you want  $\mathbf{R}$ , the factor on the right, or  $\mathbf{L}$ , the factor on the left. Generally statisticians write the decomposition as  $\mathbf{R}'\mathbf{R}$ .

The decomposition is only determined up to changes in sign of the rows of  $\mathbf{R}$  (or, equivalently, the columns of  $\mathbf{L}$ ). For definiteness we require positive diagonal elements in  $\mathbf{R}$ .

When  $\text{rank}(\mathbf{X}) = p$  the Cholesky decomposition  $\mathbf{R}$  of  $\mathbf{X}'\mathbf{X}$  is the equal to the matrix  $\mathbf{R}$  from the QR decomposition up to changes in sign of rows. The matrix  $\mathbf{X}'\mathbf{X}$  matrix is obviously symmetric and it is positive definite because

$$\mathbf{x}'(\mathbf{X}'\mathbf{X})\mathbf{x} = \mathbf{x}'(\mathbf{R}'\mathbf{R})\mathbf{x} = \|\mathbf{Rx}\|^2 \geq 0$$

with equality only when  $\mathbf{Rx} = \mathbf{0}$ , which, when  $\text{rank}(\mathbf{R}) = p$ , implies that  $\mathbf{x} = \mathbf{0}$ .

#### 3.2 Evaluation of the Cholesky decomposition

The R function `chol()` evaluates the Cholesky decomposition. As mentioned above `chol2inv()` creates  $(\mathbf{X}'\mathbf{X})^{-1}$  directly from the Cholesky decomposition of  $\mathbf{X}'\mathbf{X}$ .

Generally the QR decomposition is preferred to the Cholesky decomposition for least squares problems because there is a certain loss of precision when forming  $\mathbf{X}'\mathbf{X}$ . However, when  $n$  is very large you may want to build up  $\mathbf{X}'\mathbf{X}$  using blocks of rows. Also, if  $\mathbf{X}$  is *sparse* it is an advantage to use sparse matrix techniques to evaluate and store the Cholesky decomposition.

The `Matrix` package for R provides even more capabilities related to the Cholesky decomposition, especially for sparse matrices.

For everything we will do in Statistics 849 the QR decomposition should be the method of choice.

### 3.3 The singular value decomposition

Another decomposition related to orthogonal matrices is the singular value decomposition (or SVD) in which the matrix  $\mathbf{X}$  is reduced to a diagonal form

$$\mathbf{X} = \mathbf{U}_1 \mathbf{D} \mathbf{V}' = \mathbf{U} \begin{bmatrix} \mathbf{D} \\ \mathbf{0} \end{bmatrix} \mathbf{V}'$$

where  $\mathbf{U}$  is an  $n \times n$  orthogonal matrix,  $\mathbf{D}$  is a  $p \times p$  diagonal matrix with non-negative diagonal elements (which are called the *singular values* of  $\mathbf{X}$ ) and  $\mathbf{V}$  is a  $p \times p$  orthogonal matrix. As for  $\mathbf{Q}$  and  $\mathbf{Q}_1$ ,  $\mathbf{U}_1$  consists of the first  $p$  columns of  $\mathbf{U}$ . For definiteness we order the diagonal elements of  $\mathbf{D}$ , which must be non-negative, in decreasing order.

The singular value decomposition of  $\mathbf{X}$  is related to the eigendecomposition or spectral decomposition of  $\mathbf{X}'\mathbf{X}$  because

$$\mathbf{X}'\mathbf{X} = \mathbf{V} \mathbf{D} \mathbf{U}_1' \mathbf{U}_1 \mathbf{D} \mathbf{V}' = \mathbf{V} \mathbf{D}^2 \mathbf{V}'$$

implying that the eigenvalues of  $\mathbf{X}'\mathbf{X}$  are the squares of the singular values of  $\mathbf{X}$  and the right singular vectors, which are the columns of  $\mathbf{V}$  are also the eigenvectors of  $\mathbf{X}'\mathbf{X}$

Calculation of the SVD is an iterative (as opposed to a direct) computation and potentially more computing intensive than the QR decomposition, although modern methods for evaluating the SVD are very good indeed.

Symbolically we can write the least squares solution in the full-rank case as

$$\hat{\boldsymbol{\beta}} = \mathbf{V} \mathbf{D}^{-1} \mathbf{U}_1' \mathbf{y}$$

where  $\mathbf{D}^{-1}$  is a diagonal matrix whose diagonal elements are the inverses of the diagonal elements of  $\mathbf{D}$ .

The pseudo-inverse or *generalized inverse* of  $\mathbf{X}$ , written  $\mathbf{X}^-$  is calculated from the pseudo-inverse of the diagonal matrix,  $\mathbf{D}$ . In theory the diagonal elements of  $\mathbf{D}^-$  are  $1/d_{i,i}$ ,  $i$  when  $d_{i,i} \neq 0$  and 0 when  $d_{i,i} = 0$ . However, we can't count on  $d_{i,i}$  being 0 even when, in theory, it should be. You need to decide when the singular values are close to zero, which is actually a very difficult problem. At best we can some heuristics, based on the ratio of  $d_{i,i}/d_{1,1}$  to decide when a diagonal element is "effectively zero".

The use of the pseudo-inverse seems to be a convenient way to handle rank-deficient  $\mathbf{X}$  matrices but, as mentioned above, the best way to handle rank-deficient  $\mathbf{X}$  matrices is not to produce them in the first place.

# Math 141

## Lecture 24: Model Comparisons and The F-test

Albyn Jones<sup>1</sup>

<sup>1</sup>Library 304

jones@reed.edu

[www.people.reed.edu/~jones/courses/141](http://www.people.reed.edu/~jones/courses/141)

# Nested Models

Two linear models are **Nested** if one (the *restricted model*) is obtained from the other (the *full model*) by setting some parameters to zero (i.e. removing terms from the model), or some other constraint on the parameters.

We can compare nested models fit to the same dataset with the F test.

## Example

```
# Full Model  
Mfull <- lm(Y ~ X + W + Z + T,  
              data = MyDataSet)  
  
# Restricted Model  
Mres <- lm(Y ~ X + W, data = MyDataSet )
```

Fitting the restricted model is equivalent to forcing  $\beta_Z = \beta_T = 0$  in the full model.

# Comparing Nested Models

The crucial question is whether the residual sum of squares for the restricted model ( $RSS_R$ ) is substantially larger than the residual sum of squares for the full model ( $RSS_F$ ).

R. A. Fisher worked out the distribution of a ratio of the two under the null hypothesis that the restricted model is correct, which typically corresponds to the statement that some parameters are zero.

As usual, this story depends on the residuals having at least an approximately normal distribution.

# The F-Test

Assuming model validity, the F-ratio (F is for Fisher, by the way)

$$F_{df_N, df_F} = \frac{(RSS_R - RSS_F)/(df_R - df_F)}{RSS_F / df_F}$$

has an  $F$  distribution with degrees of freedom  $(df_N, df_F)$  if the restricted model is correct.

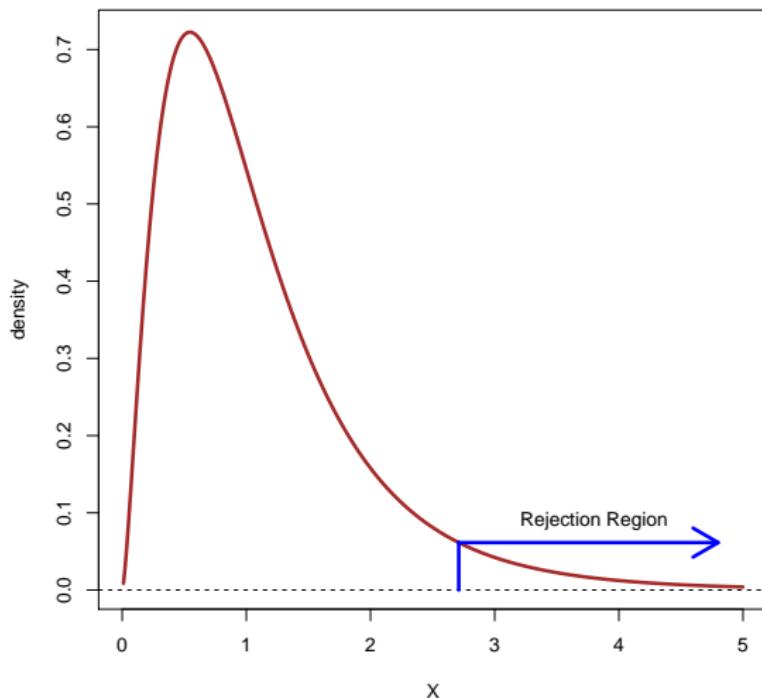
Note:  $df_N = df_R - df_F$ , and  $df_F$  and  $df_R$  are residual df from the two models.

Reject: if  $F > qf(.95, df_N, df_F)$

Note:  $df_R - df_F$  is always the number of constraints on the parameters that converts the full model to the restricted model.

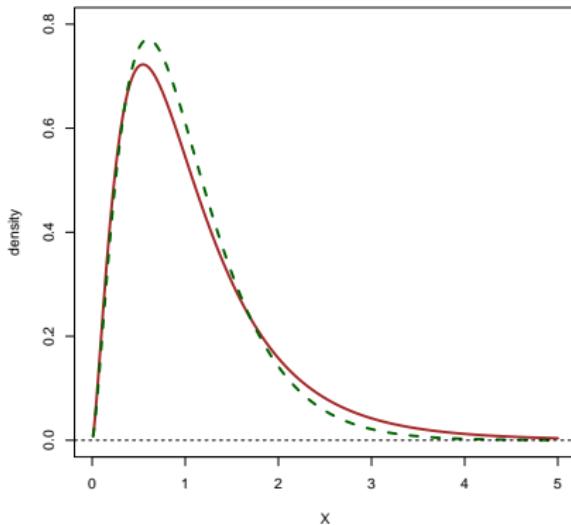
# The F density

F(5,20) density



# Analogy

$F(5,20)$  density vs  $\text{Chisq}(5)/5$



$F_{k,n}$  is to  $\chi_k^2$  as  $t_n$  is to  $N(0, 1)$ . The denominator estimates  $\sigma^2$ .  
If we knew  $\sigma^2$ , the ratio would have a  $\chi^2$  distribution.

## Connection to the t Distribution: $F_{1,k}$ is $t_k^2$

```
lm(formula = ht18 ~ ht2, data = Berkeley)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	32.1203	26.6572	1.205	0.233
ht2	1.5998	0.3031	5.278	2.2e-06
---				

Residual standard error: 7.572

on 56 degrees of freedom

F-statistic: 27.86 on 1 and 56 DF, p-value: 2.2e-06

```
> 5.278^2
```

```
[1] 27.85728
```

# Example, CPS wage data summary

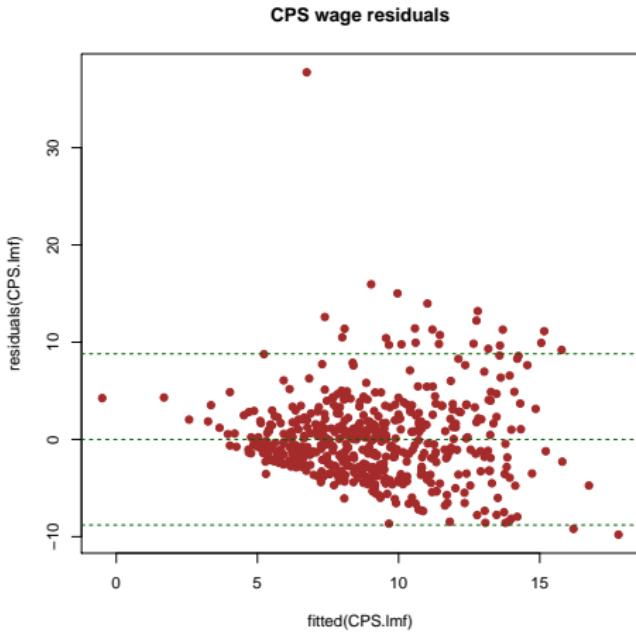
Call:

```
lm(formula = wage ~ race*sex + educ + age + union,  
    data = CPS)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-7.09772	1.46620	-4.841	< .001
raceW	0.06191	0.89024	0.070	0.94458
sexM	0.59693	1.10624	0.540	0.58970
educ	0.82717	0.07405	11.170	< .001
age	0.10481	0.01672	6.268	< .001
union	1.59479	0.51016	3.126	0.00187
raceW:sexM	1.77023	1.17363	1.508	0.13207

# Plot Residuals!



What Next?

# Example, CPS log(wage) data summary

Call:

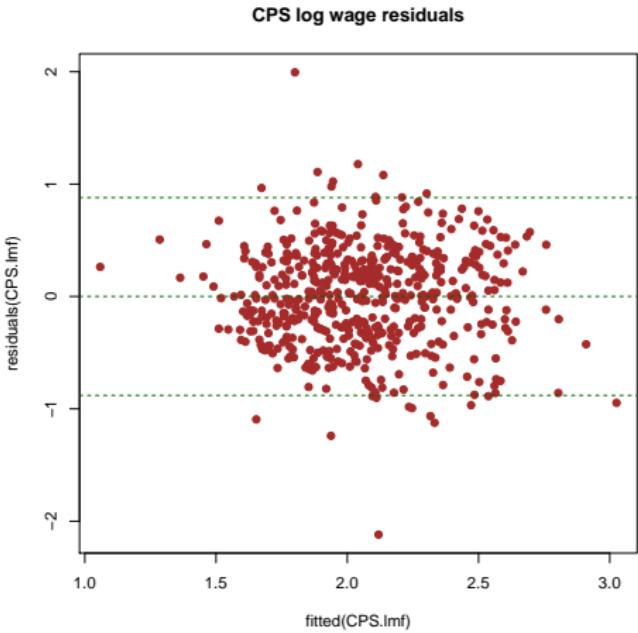
```
lm(formula = log(wage) ~ race*sex + educ +  
    age + union, data = CPS)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.330807	0.147882	2.237	0.0257
raceW	0.033692	0.089791	0.375	0.7076
sexM	0.112103	0.111577	1.005	0.3155
educ	0.085200	0.007469	11.407	< .01
age	0.011585	0.001686	6.869	< .01
union	0.221588	0.051455	4.306	< .01
raceW:sexM	0.133519	0.118374	1.128	0.2599

Residual standard error: 0.4446 on 527 df

# Plot Residuals Again!



Better?

## Looking for a parsimonious model?

None of the coefficients for race, sex, and the race\*sex interaction were statistically significantly different from zero.  
Let's fit a restricted model, dropping those non-significant explanatory variables.

## Example, CPS log(wage) Restricted Model

Call:

```
lm(formula = log(wage) ~ educ + age + union,  
   data = CPS)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.511462	0.127785	4.003	< .01
educ	0.084898	0.007694	11.034	< .01
age	0.010742	0.001728	6.217	< .01
union	0.260260	0.052135	4.992	< .01

Residual standard error: 0.4593 on 530 df

# Model Comparison

```
> anova(CPS.loglmr,CPS.loglmf)
```

Analysis of Variance Table

Model 1: log(wage) ~ educ + age + union

Model 2: log(wage) ~ race\*sex + educ + age + union

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	530	111.81				
2	527	104.16	3	7.6478	12.898	3.836e-08

**Say WHAT?** None of the omitted coefficients were statistically significantly different from 0! How can this happen?

# The Null and Alternative Hypotheses

What is  $H_0$ ?

The restricted model is correct. Informally: the restricted model fits as well as the full model.

Formally:

$H_0$ : coefficients for the omitted terms are all 0.

Formally:

$H_1$ : at least one omitted coefficient is not zero.

# Important!

Individual t-tests are testing a null hypothesis for a single coefficient

$$H_0 : \beta = 0$$

given we have controlled for the other variables in the model!

# What was missing?

```
formula = log(wage) ~ sex + educ + age + union,  
                    data = CPS)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.352998	0.126995	2.780	0.00564
sexM	0.228555	0.039400	5.801	< .001
educ	0.085473	0.007468	11.445	< .001
age	0.011727	0.001686	6.957	< .001
unionUnion	0.210191	0.051331	4.095	< .001

# What happened?

The race\*sex interaction was a distraction!

```
> cor(sex=="M", sex=="M" & race=="W")  
[1] 0.8638632
```

Strongly correlated explanatory variables can be distractors, each does part of the work of predicting the response, neither seems important when the other is included.

# Interpretation

The coefficient for the dummy variable for Males was about .23.  
What does that mean?

All other factors held equal, the difference between  $\log(wage)$  for males and  $\log(wage)$  for females is .23:

$$\log(W) = \text{OtherStuff} + .23 \cdot \text{sexM}$$

Therefore

$$W = e^{\text{OtherStuff} + .23 \cdot \text{sexM}} = e^{\text{OtherStuff}} e^{.23 \cdot \text{sexM}}$$

The dummy variable sexM is 1 for males and 0 for females, so the difference is the multiplicative factor

$$e^{.23} \approx 1.26$$

Conclusion: Males with the same education level, age and Union status get paid about 26% more than corresponding females with the same covariate values.

# R will try to prevent silliness

```
> anova(CPS.loglmr, CPS.lmf)
```

Analysis of Variance Table

Response: log(wage)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
educ	1	21.481	21.4807	101.821	< .001
age	1	9.898	9.8976	46.916	< .001
union	1	5.257	5.2573	24.920	< .001
Residuals	530	111.811	0.2110		
---					

Warning message:

In anova.lmlist(object, ...) :

models with response "wage" removed because  
response differs from model 1

# Michelson's Data, full model

```
> summary(MF)
```

Call:

```
lm(formula = Speed ~ Run, data = Michelson)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	299909.00	16.60	18067.739	0.000
Run2	-53.00	23.47	-2.258	0.026
Run3	-64.00	23.47	-2.726	0.007
Run4	-88.50	23.47	-3.770	0.000
Run5	-77.50	23.47	-3.301	0.001

# Michelson's Data, restricted model

```
> Run1 <- Michelson$Run == 1
```

```
> summary(MR)
```

Call:

```
lm(formula = Speed ~ Run1, data = Michelson)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.998e+05	8.283e+00	36197.63	< 2e-16
Run1TRUE	7.075e+01	1.852e+01	3.82	0.000234

# Model Comparison!

```
> anova(MR, MF)
```

Analysis of Variance Table

Model 1: Speed ~ Run1

Model 2: Speed ~ Run

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	98	537935				
2	95	523510	3	14425	0.8726	0.4582

What was  $H_0$ , and what do we conclude?

**The F test compares nested models fit to the same dataset.**

It allows us to test hypotheses involving multiple parameters simultaneously.

If you wish to conclude that a collection of coefficients are all zero, or none of a subset of your explanatory variables predict the response, an F-test is the appropriate tool.



# Quick-R

accessing the power of R

## Statistics

[Descriptive Statistics](#)

[Frequencies & Crosstabs](#)

[Correlations](#)

[t-tests](#)

[Nonparametric Statistics](#)

[Multiple Regression](#)

[Regression Diagnostics](#)

[ANOVA/MANOVA](#)

[\(M\)ANOVA Assumptions](#)

[Resampling Stats](#)

[Power Analysis](#)

[Using With and By](#)

## R in Action



[R in Action](#) significantly expands upon this material. Use promo code **ria38** for a 38% discount.

## Top Menu

[Home](#)

[The R Interface](#)

[Data Input](#)

[Data Management](#)

[Basic Statistics](#)

[Advanced Statistics](#)

[Basic Graphs](#)

[Advanced Graphs](#)

[Blog](#)

## Regression Diagnostics

An excellent review of regression diagnostics is provided in John Fox's aptly named [Overview of Regression Diagnostics](#). Dr. Fox's `car` package provides advanced utilities for regression modeling.

```
# Assume that we are fitting a multiple linear regression
# on the MTCARS data
library(car)
fit <- lm(mpg~disp+hp+wt+drat, data=mtcars)
```

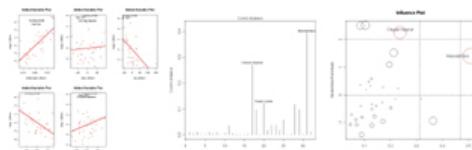
This example is for **exposition only**. We will ignore the fact that this may not be a great way of modeling the this particular set of data!

## Outliers

```
# Assessing Outliers
outlierTest(fit) # Bonferonni p-value for most extreme obs
qqPlot(fit, main="QQ Plot") #qq plot for studentized resid
leveragePlots(fit) # leverage plots
```

## Influential Observations

```
# Influential Observations
# added variable plots
av.plots(fit)
# Cook's D plot
# identify D values > 4/(n-k-1)
cutoff <- 4/((nrow(mtcars)-length(fit$coefficients)-2))
plot(fit, which=4, cook.levels=cutoff)
# Influence Plot
influencePlot(fit, id.method="identify", main="Influence Plot",
sub="Circle size is proportional to Cook's Distance" )
```

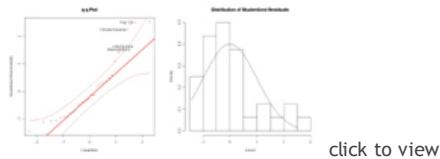


click to view

## Non-normality

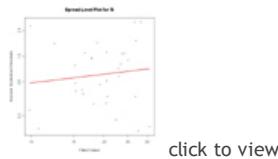
```
# Normality of Residuals
# qq plot for studentized resid
```

```
qqplot(fit, main="QQ Plot")
# distribution of studentized residuals
library(MASS)
sresid <- studres(fit)
hist(sresid, freq=FALSE,
     main="Distribution of Studentized Residuals")
xfit<-seq(min(sresid),max(sresid),length=40)
yfit<-dnorm(xfit)
lines(xfit, yfit)
```



## Non-constant Error Variance

```
# Evaluate homoscedasticity
# non-constant error variance test
ncvTest(fit)
# plot studentized residuals vs. fitted values
spreadLevelPlot(fit)
```

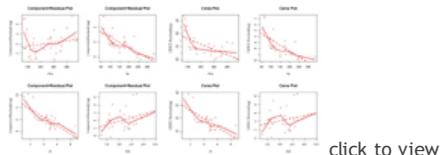


## Multi-collinearity

```
# Evaluate Collinearity
vif(fit) # variance inflation factors
sqrt(vif(fit)) > 2 # problem?
```

## Nonlinearity

```
# Evaluate Nonlinearity
# component + residual plot
crPlots(fit)
# Ceres plots
ceresPlots(fit)
```



## Non-independence of Errors

```
# Test for Autocorrelated Errors
durbinWatsonTest(fit)
```

## Additional Diagnostic Help

The `gvlma()` function in the `gvlma` package, performs a global validation of linear model assumptions as

well separate evaluations of skewness, kurtosis, and heteroscedasticity.

```
# Global test of model assumptions
library(gvlma)
gvmmodel <- gvlma(fit)
summary(gvmmodel)
```

## Going Further

If you would like to delve deeper into regression diagnostics, two books written by John Fox can help:  
[Applied regression analysis and generalized linear models \(2nd ed\)](#) and [An R and S-Plus companion to applied regression](#).

Copyright © 2014 [Robert I. Kabacoff, Ph.D.](#) | [Sitemap](#)

Designed by [WebTemplateOcean.com](#)

**mtcars** {datasets}

## R Documentation

**Motor Trend Car Road Tests****Description**

The data was extracted from the 1974 *Motor Trend* US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

**Usage**

```
mtcars
```

**Format**

A data frame with 32 observations on 11 variables.

[, 1]	mpg	Miles/(US) gallon
[, 2]	cyl	Number of cylinders
[, 3]	disp	Displacement (cu.in.)
[, 4]	hp	Gross horsepower
[, 5]	drat	Rear axle ratio
[, 6]	wt	Weight (lb/1000)
[, 7]	qsec	1/4 mile time
[, 8]	vs	V/S
[, 9]	am	Transmission (0 = automatic, 1 = manual)
[,10]	gear	Number of forward gears
[,11]	carb	Number of carburetors

**Source**

Henderson and Velleman (1981), Building multiple regression models interactively. *Biometrics*, **37**, 391–411.

**Examples**

```
require(graphics)
pairs(mtcars, main = "mtcars data")
coplot(mpg ~ disp | as.factor(cyl), data = mtcars,
       panel = panel.smooth, rows = 1)
```

---

[Package *datasets* version 2.16.0 [Index](#)]

**The R Datasets Package**

**Documentation for package 'datasets' version 2.15.3**

- [DESCRIPTION file.](#)

### [Help Pages](#)

[A](#) [B](#) [C](#) [D](#) [E](#) [F](#) [H](#) [I](#) [J](#) [L](#) [M](#) [N](#) [O](#) [P](#) [Q](#) [R](#) [S](#) [T](#) [U](#) [V](#) [W](#)

#### [datasets-package](#)

The R Datasets Package

-- A --

<a href="#">ability.cov</a>	Ability and Intelligence Tests
<a href="#">airmiles</a>	Passenger Miles on Commercial US Airlines, 1937-1960
<a href="#">AirPassengers</a>	Monthly Airline Passenger Numbers 1949-1960
<a href="#">airquality</a>	New York Air Quality Measurements
<a href="#">anscombe</a>	Anscombe's Quartet of 'Identical' Simple Linear Regressions
<a href="#">attenu</a>	The Joyner-Boore Attenuation Data
<a href="#">attitude</a>	The Chatterjee-Price Attitude Data
<a href="#">austres</a>	Quarterly Time Series of the Number of Australian Residents

-- B --

<a href="#">beaver1</a>	Body Temperature Series of Two Beavers
<a href="#">beaver2</a>	Body Temperature Series of Two Beavers
<a href="#">beavers</a>	Body Temperature Series of Two Beavers
<a href="#">BJsales</a>	Sales Data with Leading Indicator
<a href="#">BJsales.lead</a>	Sales Data with Leading Indicator
<a href="#">BOD</a>	Biochemical Oxygen Demand

-- C --

<a href="#">cars</a>	Speed and Stopping Distances of Cars
<a href="#">ChickWeight</a>	Weight versus age of chicks on different diets
<a href="#">chickwts</a>	Chicken Weights by Feed Type
<a href="#">CO2</a>	Carbon Dioxide Uptake in Grass Plants
<a href="#">co2</a>	Mauna Loa Atmospheric CO2 Concentration
<a href="#">crimtab</a>	Student's 3000 Criminals Data

-- D --

[datasets](#)

The R Datasets Package

[discoveries](#)

Yearly Numbers of Important Discoveries

[DNase](#)

Elisa assay of DNase

--- E ---

[esoph](#)

Smoking, Alcohol and (O)esophageal Cancer

[euro](#)

Conversion Rates of Euro Currencies

[euro.cross](#)

Conversion Rates of Euro Currencies

[eurodist](#)

Distances Between European Cities

[EuStockMarkets](#)

Daily Closing Prices of Major European Stock Indices, 1991-1998

--- F ---

[faithful](#)

Old Faithful Geyser Data

[fdeaths](#)

Monthly Deaths from Lung Diseases in the UK

[Formaldehyde](#)

Determination of Formaldehyde

[freeny](#)

Freeny's Revenue Data

[freeny.x](#)

Freeny's Revenue Data

[freeny.y](#)

Freeny's Revenue Data

--- H ---

[HairEyeColor](#)

Hair and Eye Color of Statistics Students

[Harman23.cor](#)

Harman Example 2.3

[Harman74.cor](#)

Harman Example 7.4

--- I ---

[Indometh](#)

Pharmacokinetics of Indomethacin

[infert](#)

Infertility after Spontaneous and Induced Abortion

[InsectSprays](#)

Effectiveness of Insect Sprays

[iris](#)

Edgar Anderson's Iris Data

[iris3](#)

Edgar Anderson's Iris Data

[islands](#)

Areas of the World's Major Landmasses

--- J ---

[JohnsonJohnson](#)

Quarterly Earnings per Johnson &amp; Johnson Share

--- L ---

[LakeHuron](#)

Level of Lake Huron 1875-1972

[ldeaths](#)

Monthly Deaths from Lung Diseases in the UK

[lh](#)

Luteinizing Hormone in Blood Samples

[LifeCycleSavings](#)

Intercountry Life-Cycle Savings Data

[Loblolly](#)

Growth of Loblolly pine trees

[longley](#)

Longley's Economic Regression Data

[lynx](#)

Annual Canadian Lynx trappings 1821-1934

-- M --

[mdeaths](#)

Monthly Deaths from Lung Diseases in the UK

[morley](#)

Michelson Speed of Light Data

[mtcars](#)

Motor Trend Car Road Tests

-- N --

[nhtemp](#)

Average Yearly Temperatures in New Haven

[Nile](#)

Flow of the River Nile

[nottem](#)

Average Monthly Temperatures at Nottingham, 1920-1939

[npk](#)

Classical N, P, K Factorial Experiment

-- O --

[occupationalStatus](#)

Occupational Status of Fathers and their Sons

[Orange](#)

Growth of Orange Trees

[OrchardSprays](#)

Potency of Orchard Sprays

-- P --

[PlantGrowth](#)

Results from an Experiment on Plant Growth

[precip](#)

Annual Precipitation in US Cities

[presidents](#)

Quarterly Approval Ratings of US Presidents

[pressure](#)

Vapor Pressure of Mercury as a Function of Temperature

[Puromycin](#)

Reaction Velocity of an Enzymatic Reaction

-- Q --

[quakes](#)

Locations of Earthquakes off Fiji

-- R --

[randu](#)

Random Numbers from Congruential Generator RANDU

[rivers](#)

Lengths of Major North American Rivers

[rock](#)

Measurements on Petroleum Rock Samples

-- S --

[Seatbelts](#)

Road Casualties in Great Britain 1969-84

[sleep](#)

Student's Sleep Data

[stack.loss](#)

Brownlee's Stack Loss Plant Data

[stack.x](#)

Brownlee's Stack Loss Plant Data

[stackloss](#)

Brownlee's Stack Loss Plant Data

[state](#)

US State Facts and Figures

<a href="#">state.abb</a>	US State Facts and Figures
<a href="#">state.area</a>	US State Facts and Figures
<a href="#">state.center</a>	US State Facts and Figures
<a href="#">state.division</a>	US State Facts and Figures
<a href="#">state.name</a>	US State Facts and Figures
<a href="#">state.region</a>	US State Facts and Figures
<a href="#">state.x77</a>	US State Facts and Figures
<a href="#">sunspot.month</a>	Monthly Sunspot Data, from 1749 to "Present"
<a href="#">sunspot.year</a>	Yearly Sunspot Data, 1700-1988
<a href="#">sunspots</a>	Monthly Sunspot Numbers, 1749-1983
<a href="#">swiss</a>	Swiss Fertility and Socioeconomic Indicators (1888) Data

-- T --

<a href="#">Theoph</a>	Pharmacokinetics of Theophylline
<a href="#">Titanic</a>	Survival of passengers on the Titanic
<a href="#">ToothGrowth</a>	The Effect of Vitamin C on Tooth Growth in Guinea Pigs
<a href="#">treering</a>	Yearly Treering Data, -6000-1979
<a href="#">trees</a>	Girth, Height and Volume for Black Cherry Trees

-- U --

<a href="#">UCBAmissions</a>	Student Admissions at UC Berkeley
<a href="#">UKDriverDeaths</a>	Road Casualties in Great Britain 1969-84
<a href="#">UKgas</a>	UK Quarterly Gas Consumption
<a href="#">UKLungDeaths</a>	Monthly Deaths from Lung Diseases in the UK
<a href="#">USAccDeaths</a>	Accidental Deaths in the US 1973-1978
<a href="#">USAArrests</a>	Violent Crime Rates by US State
<a href="#">USJudgeRatings</a>	Lawyers' Ratings of State Judges in the US Superior Court
<a href="#">USPersonalExpenditure</a>	Personal Expenditure Data
<a href="#">uspop</a>	Populations Recorded by the US Census

-- V --

<a href="#">VADeaths</a>	Death Rates in Virginia (1940)
<a href="#">volcano</a>	Topographic Information on Auckland's Maunga Whau Volcano

-- W --

<a href="#">warpbreaks</a>	The Number of Breaks in Yarn during Weaving
<a href="#">women</a>	Average Heights and Weights for American Women
<a href="#">WorldPhones</a>	The World's Telephones
<a href="#">WWWusage</a>	Internet Usage per Minute

# Regression Diagnostics with R

Anne Boomsma

Department of Statistics & Measurement Theory  
University of Groningen

April 30, 2014

`Regrdiag.R.tex`

# Regression Diagnostics with R

Anne Boomsma

*Department of Statistics & Measurement Theory, University of Groningen*

## 1. Introduction

In our opinion, the best start for regression applications in R is either Faraway's (2005) book *Linear models with R*, or Fox's (2002) *R and S-Plus companion to applied regression*. In this document, we present an overview of regression diagnostics using material from chapter four of Faraday's book mainly. When running through the examples, the power of the R environment will become unmistakably clear, especially in the versatility of its graphical options.

First, install the packages `faraway` (Faraway), `car` (Fox), and `lmtest` (R) from a Comprehensive R Archive Network (CRAN) mirror by choosing [Packages → Install packages](#) at the upper tool bar of RGui (R's Graphical user interface). Next, load the `faraway` package, and from that package data frame `savings`.

```
> library(faraway)                                # loading package 'faraway'  
> data(savings)                                  # documentation on data set 'sexab'
```

The command `attach(savings)` is not strictly necessary in the sequel, nor recommended here: for some commands, country labels would vanish in the output.

```
> ? savings                                     # documentation of "Savings rates"
```

This data frame contains the savings rates in  $n = 50$  countries (source: Belsley, Kuh & Welsch, 1980). The data are averaged over the period 1960–1970. The data frame ( $50 \times 5$ ) contains the following objects or variables:

<code>sr</code>	savings rate – personal saving divided by disposable income
<code>pop15</code>	percent population under age of 15
<code>pop75</code>	percent population over age of 75
<code>dpi</code>	per-capita disposable income in dollars
<code>ddpi</code>	percent growth rate of <code>dpi</code>

```
> savings # list complete data frame 'savings'
```

The linear regression model **M1** for response variable savings rate **sr** is specified and estimated as follows:

```
> M1 <- lm(sr ~ pop15 + pop75 + dpi + ddpi, data=savings)
> (M1_sum <- summary(M1)) # summary of estimated model
```

Call:

```
lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = savings)
```

Residuals:

Min	1Q	Median	3Q	Max
-8.242	-2.686	-0.249	2.428	9.751

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	28.566087	7.354516	3.88	0.00033 ***
pop15	-0.461193	0.144642	-3.19	0.00260 **
pop75	-1.691498	1.083599	-1.56	0.12553
dpi	-0.000337	0.000931	-0.36	0.71917
ddpi	0.409695	0.196197	2.09	0.04247 *
---				
Signif. codes:	0 *** 0.001 ** 0.01 * 0.05 . 0.1			1

Residual standard error: 3.8 on 45 degrees of freedom

Multiple R-squared: 0.338, Adjusted R-squared: 0.28

F-statistic: 5.76 on 4 and 45 DF, p-value: 0.00079

▷ Check whether the details of this summary are well understood.

```
> options(show.signif.stars=F) # suppress stars of significance
> options(digits=4) # set numbers of significant digits
```

The fitted values  $\hat{Y}_i$  and the residuals  $e_i$  can be obtained as follows:

```
> fitted(M1) # predicted Y:  $\hat{Y}_i$ 
> residuals(M1) # residuals  $e_i$ 
> which.max(abs(residuals(M1))) # largest absolute residual  $|e_i|$ ?
```

Zambia

46

▷ Diagnostic purpose of residuals: locating large errors in prediction.

## 2. Checking model assumptions

We need to inspect the validity of the main assumptions of the linear regression model. This refers, first of all, to the (conditional) distribution of the model's errors terms  $\epsilon_i$ : homogeneous variance, normality, and independence. Analysis of observed residuals  $e_i$  may help to evaluate the plausibility of these assumptions. Checking for unusual and influential observations is another part of regression diagnostics. In addition, the validity of the structural model itself, i.e., its linearity  $E(Y) = \mathbf{X}\beta$  and the selection of explanatory variables, should be examined.

### 2.1 Constant variance

- Residual plot:  $\hat{Y}_i$  against  $e_i$

```
> par(las=1)                                # horizontal style of axis labels
> plot(fitted(M1), residuals(M1), xlab="Fitted", ylab="Residuals")
> abline(h=0, col="red")                      # draws a horizontal red line at y = 0
```

There are a number of specific plot diagnostics for an `lm()` object, which allow for standard plotting jobs—all available in the built-in `stats` package.

```
> ? plot.lm
> plot(M1, which=1)                          #  $\hat{Y}_i$  against  $e_i$ 
> plot(M1, ask=TRUE)                         # all six standard lm() plots available
```

- Absolute residual plot:  $\hat{Y}_i$  against  $|e_i|$

```
> plot(fitted(M1), abs(residuals(M1)), xlab="Fitted", lab="|Residuals|")
```

This plot is designed to check for constant variance only.

- Absolute residual plot:  $\hat{Y}_i$  against  $\text{sqrt}(\text{standardized } |e_i|)$

```
> plot(M1, which=3)                          # R's standardized residuals scale-location plot
```

- Quick and dirty test

Faraway (2005) mentions the following  $F$ -test as a quick way to check non-constant variance by a regression of  $|e_i|$  on  $\hat{Y}_i$ , where  $|e_i|$  is the response and  $\hat{Y}_i$  the explanatory variable.

```
> summary(lm(abs(residuals(M1)) ~ fitted(M1)))
```

```

Call:
lm(formula = abs(residuals(M1)) ~ fitted(M1))

Residuals:
    Min      1Q  Median      3Q     Max 
-2.8395 -1.6078 -0.3493  0.6625  6.7036 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.8398    1.1865   4.079 0.000170 ***
fitted(M1)   -0.2035    0.1185  -1.717 0.092501  
                                                        
Residual standard error: 2.163 on 48 degrees of freedom
Multiple R-Squared:  0.05784,    Adjusted R-squared:  0.03821 
F-statistic: 2.947 on 1 and 48 DF,  p-value: 0.0925

```

It turns out that the absolute residuals  $|e_i|$  are not predicted very well by  $\hat{Y}_i$ , which is roughly satisfying. Hence, we conclude that there does not seem to be a problem with the constant variance assumption.

- Interpretation of residual plots

For a proper evaluation of residual plots it may be helpful to generate some artificial plots for situations where true relationships are known.

```

> par(mfrow=c(3,3))          # setting a plot device, here a (3 x 3) matrix
                                # with three plots in each of the three rows;
                                # first, an empty plot window pops up

> for(i in 1:9) plot(1:50,rnorm(50))          # constant variance
> for(i in 1:9) plot(1:50,(1:50)*rnorm(50))    # strong heterogeneity
> for(i in 1:9) plot(1:50,sqrt((1:50))*rnorm(50)) # mild heterogeneity
> for(i in 1:9) plot(1:50,cos((1:50)*pi/25)+rnorm(50)) # non-linearity

```

## 2.2 Normality

- Q-Q plots

Observed ordered residuals  $e_i$  (the sample quantiles at the  $y$ -axis) are plotted against expected normal quantiles  $\Phi^{(-1)}[i/(n+1)]$  at the  $x$ -axis, where  $\Phi(x)$  is the standard normal distribution function, i.e.,  $\Phi(x) = \Pr(X < x)$ . Recall that  $e_i \sim \mathcal{N}(0, \sigma^2)$ .

For the `savings` data we try the following:

```
> par(mfrow=c(1,1))                                # reset plotting device
> qqnorm(residuals(M1), ylab="Residuals")          # Q-Q plot
> qqline(residuals(M1))                            # line through Q1 and Q3
```

- Interpretation of Q-Q plots

To get an idea of the variation to be expected in a Q-Q plot, inspect the plots generated for a number of probability distributions. In the examples below, we use the standard normal, the lognormal, Student's  $t$  with one degree of freedom, and the uniform  $\mathcal{U}(0, 1)$  distribution, respectively. Nine independent pseudo-random samples of size 50 are generated from each distribution. For each sample, a Q-Q plot with a quartile-line is produced.

```
> par(mfrow=c(3,3))
> for(i in 1:9) x = rnorm(50); qqnorm(x); qqline(x)
  # i.e., standard normal distribution (symmetric)

> for(i in 1:9) x = rlnorm(50); qqnorm(x); qqline(x)
  # lognormal distribution (long right tail, skew to right)

> for(i in 1:9) x = rt(50,1); qqnorm(x); qqline(x)
  # Student t-distribution with one df (heavy tails, platykurtic)

> for(i in 1:9) x = runif(50); qqnorm(x); qqline(x)
  # uniform (0,1) distribution (short tails, leptokurtic)
```

If the errors  $\epsilon_i$  are not normal, the least squares estimates may not be optimal. They will still be best linear unbiased estimates, but other robust estimators may be more effective. Tests and confidence intervals may not be exact. Long-tailed distributions in particular, cause large inaccuracies. Mild non-normality may be safely ignored, according to Faraway (2005, p. 59), but we may need more specificity here.

- Histograms and box plots

Histograms and box plots graphs are also suitable for checking normality, along with descriptive statistics like skewness and kurtosis, for example.

```
> par(mfrow=c(1,1))
> hist(residuals(M1))
> boxplot(residuals(M1))
```

- Shapiro-Wilks normality test

```
> shapiro.test(residuals(M1))

Shapiro-Wilk normality test

data: residuals(M1)
W = 0.987, p-value = 0.8524
```

The null hypothesis is that the residuals have a normal distribution. The *p*-value of the test statistic is large in this example. It thus follows that the null hypothesis is not rejected. Faraway (2005) only recommends this test in conjunction with a Q-Q plot. For large samples the test may be too sensitive, and for small samples its power may be too small – the usual dilemma.

### 2.3 Independent errors

The data set `airquality` from the `datasets` package serves as a more appropriate illustration here than the `savings` data. The data are daily air quality measurements in New York, from May to September 1973 (source: Chambers, Cleveland, Kleiner & Tukey, 1983). We have a data frame with `n = 153` observations on 6 numerical variables.

<code>Ozone</code>	Ozone (ppb – in parts per billion particles)
<code>Solar.R</code>	Solar R (Solar radiation in Langleys)
<code>Wind</code>	Wind (mph)
<code>Temp</code>	Temperature (degrees F)
<code>Month</code>	Month (1–12)
<code>Day</code>	Day of month (1–31)

```
> airquality                                # notice missing values (NAs)
> attach(airquality)
> names(airquality)
```

- Scatter plots

Take a look at scatter plots first. The function `pairs()` produces a matrix of scatter plots for all pairs of variables in a data frame.

```
> pairs(airquality, panel=panel.smooth)        # matrix of scatter plots
```

Inspection of correlations for linear relationships (listwise deletion of missing cases), given these scatter plots, can be illustrative too.

```
> round(cor(airquality, use="complete.obs"), digits=2)
```

Next a linear regression model `M2` for `Ozone` is fitted to the data, where `Month` and `Day` are not used as linear predictors.

```
> M2 <- lm(Ozone ~ Solar.R + Wind + Temp, data=airquality,
+   na.action=na.exclude)
> summary(M2)                                # summary of the estimated linear model

Call:
lm(formula = Ozone ~ Solar.R + Wind + Temp, data = airquality,
    na.action = na.exclude)

Residuals:
    Min      1Q  Median      3Q     Max 
-40.485 -14.219 -3.551  10.097  95.619 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -64.34208   23.05472  -2.791  0.00623  
Solar.R       0.05982    0.02319   2.580  0.01124  
Wind         -3.33359    0.65441  -5.094 1.52e-06  
Temp          1.65209    0.25353   6.516 2.42e-09  

Residual standard error: 21.18 on 107 degrees of freedom
Multiple R-Squared:  0.6059,    Adjusted R-squared:  0.5948 
F-statistic: 54.83 on 3 and 107 DF,  p-value: < 2.2e-16

> table(complete.cases(airquality))  # number of complete cases
```

We notice that the data frame has missing values. There are 111 complete cases only. The default with respect to missing values for regression analysis in R is to omit any case that contains a missing value. The option `na.action=na.exclude` does not use cases with missing values in the computation but keeps track of which cases are missing in the residual, fitted values and other quantities.

Residual diagnostics show some non-constant variance and non-linearity—see the previous `pairs()` plots. Therefore, a logarithmic transformation of the response variable `Ozone` is made, resulting in model `M2_log`.

- Transformation of the response variable

```
> M2_log <- lm(log(Ozone) ~ Solar.R + Wind + Temp, airquality,
+   na.action=na.exclude)
> summary(M2_log)                                # summary of the estimated linear model

Call:
lm(formula = log(Ozone) ~ Solar.R + Wind + Temp, data = airquality,
    na.action = na.exclude)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.061929 -0.299696 -0.002312  0.307559  1.235783 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.2621323  0.5535669 -0.474 0.636798  
Solar.R       0.0025152  0.0005567  4.518 1.62e-05  
Wind        -0.0615625  0.0157130 -3.918 0.000158  
Temp         0.0491711  0.0060875  8.077 1.07e-12  

Residual standard error: 0.5086 on 107 degrees of freedom
Multiple R-Squared:  0.6644,    Adjusted R-squared:  0.655 
F-statistic: 70.62 on 3 and 107 DF,  p-value: < 2.2e-16
```

- ▷ Notice the improvement of fit of model `M2_log` over that of model `M2`, where `Ozone` was untransformed.

We now check for correlated error terms. Recall that there is a time component in the `airquality` data.

- Index plot of residuals  $e_i$ , i.e., a plot of  $e_i$  against time

```
> par(las=1, mfrow=c(1,1))
> plot(residuals(M2_log), ylab="Residuals")
> abline(h=0)
```

If there was serial correlation, we would see either long runs of residuals above or below the line for positive correlation, or greater than normal fluctuations for negative correlation. Unless the effects are strong, they may be difficult to detect. Therefore, it is often better to plot successive residuals.

- Plot of successive residuals  $e_i$  [1,152] against  $e_{i+1}$  [2,154]

```
> plot(residuals(M2_log)[-153], residuals(M2_log)[-1],
+   xlab=expression(hat(epsilon)[i]), ylab=expression(hat(epsilon)[i+1]))
```

No obvious problem with correlated errors is shown. There is an outlier though, which we may try to identify. Is there really only one outlier?

```
> identify(residuals(M2_log)[-153], residuals(M2_log)[-1], n=4)
```

- Regression of  $e_{i+1}$  [response] on  $e_i$  [explanatory variable]

```
> summary(lm(residuals(M2_log)[-1] ~ -1 + residuals(M2_log)[-153]))
```

Call:

```
lm(formula = residuals(M2_log)[-1] ~ -1 + residuals(M2_log)[-153])
```

Residuals:

Min	1Q	Median	3Q	Max
-2.07274	-0.28953	0.02583	0.32256	1.32594

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
residuals(M2_log)[-153]	0.1104	0.1053	1.048	0.297

Residual standard error: 0.5078 on 91 degrees of freedom

Multiple R-Squared: 0.01193, Adjusted R-squared: 0.001073

F-statistic: 1.099 on 1 and 91 DF, p-value: 0.2973

This regression model of successive residuals omits the intercept term, `-1`, because the mean of the residuals is zero, by definition.

Clearly, there is no substantive correlation (take the square root of R-Squared, which gives 0.10922), also to be shown as follows:

```
> cor(residuals(M2_log)[-1], residuals(M2_log)[-153], use="complete.obs")
[1] 0.1092547
```

- Durbin-Watson test

The function for this test statistic is implemented in the `lmtest` package.

```
> library(lmtest)
Loading required package: zoo                      # a message that can be ignored
# zoo means Z's Ordered Observations

> dwtest(Ozone ~ Solar.R + Wind + Temp, data=na.omit(airquality))

Durbin-Watson test

data: Ozone ~ Solar.R + Wind + Temp
DW = 1.9355, p-value = 0.3347
alternative hypothesis: true autocorrelation is greater than 0
```

The  $p$  value indicates that there is no evidence of correlated errors, but the results should be viewed with skepticism because of the omission of the missing values, according to Faraway (2005). Interestingly, Faraway does not show the test results for `log(Ozone)`, which are slightly worse (`DW = 1.8068, p-value = 0.1334`).

In general, if the errors appear to be correlated, we can use generalized least squares estimation, implemented by the function `gls()`.

- Runs test

A runs test is an alternative to the Durbin-Watson test. The function `runs.test()` in the package `tseries` computes the runs test statistic for randomness of a dichotomous (binary) data series `x`. Its application is not appropriate here, because of missing values `NAs`.

### 3. Detecting unusual observations

The search for unusual, weird data points and influential observations is as important as checking model assumptions, if not a more crucial task indeed. For illustrations, we return to the `savings` data set in the `faraway` package, and to model `M1` as defined on page 2.

#### 3.1 Leverage points

First, notice that the function `influence()` returns values from four vectors or matrices.

- `hat`: a vector containing the diagonal of the `hat` matrix (see Boomsma, 2010)—the diagonal elements are the so-called leverage points  $h_i$ .

- **coefficients**: unless `do.coef` is `FALSE`, a matrix whose  $i$ th row contains the resulting change in the estimated coefficients when the  $i$ th case is dropped from the regression.
- **sigma**: a vector whose  $i$ th element contains the estimate of the residual standard deviation obtained when the  $i$ th case is dropped from the regression.
- **wt.res**: a vector of `weighted` (or for class `glm` rather `deviance`) residuals.

For more details, use the following commands:

```
> help(influence)                                # details of the four vectors/matrices
> M1_inf <- influence(M1)                      # the listed influence information
> M1_inf$hat                                     # leverages  $h_i$  of savings data
> summary(M1_inf$hat)

      Min. 1st Qu. Median   Mean 3rd Qu.   Max.
0.03730 0.06427 0.07502 0.10000 0.09702 0.53150
```

- The purpose of leverages  $h_i$  is to detect outliers in explanatory variables  $X_j$ .

Outliers, according to Stevens (1992), are values of  $h_i > 2p/n$ ; here  $2p/n = 10/50 = 0.20$ .

```
> which(M1_inf$hat > 0.20)

Ireland      Japan      United States      Libya
      21          23            44            49

> sum(M1_inf$hat)                                # sum equals number of predictors
[1] 5
```

As an efficient alternative, the function `hatvalues()` could be used, as recommended in the R documentation of `influence()`.

```
> hatvalues(M1); sum(hatvalues(M1))
```

- Half-normal plots for leverages

Plot the data against the positive normal quantiles. We are looking for outliers. The function `halfnorm()` is implemented in the `faraway` package.

```
> par(mfrow=c(1,1))
> countries <- rownames(savings)                 # stores names of countries
```

```
> halfnorm(lm.influence(M1)$hat, labs=countries, ylab="Leverages")
```

In this half-normal plot, the labels of countries having the two largest leverages are shown by default, see `help(halfnorm)`.

R has a function for `lm()` objects, plotting leverage points against standardized residuals (as defined by R), and ranges of Cooks's distances.

```
> plot(M1, which=5) # leverage against R's standardized residuals
```

### 3.2 Outliers

- Standardized residuals

```
> M1_sum <- summary(M1) # linear model 'M1' for savings rate 'sr'  
> M1_sum # summary of estimated model, as shown before
```

Call:

```
lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = savings)
```

Residuals:

Min	1Q	Median	3Q	Max
-8.2422	-2.6857	-0.2488	2.4280	9.7509

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	28.5660865	7.3545161	3.884	0.000334
pop15	-0.4611931	0.1446422	-3.189	0.002603
pop75	-1.6914977	1.0835989	-1.561	0.125530
dpi	-0.0003369	0.0009311	-0.362	0.719173
ddpi	0.4096949	0.1961971	2.088	0.042471

Residual standard error: 3.803 on 45 degrees of freedom

Multiple R-Squared: 0.3385, Adjusted R-squared: 0.2797

F-statistic: 5.756 on 4 and 45 DF, p-value: 0.0007904

```
> M1_sum$sig # sqrt(MSE)  
[1] 3.802669
```

The statistic **sig** gives an estimate of residual standard error, as it is called; in fact, **sqrt(MSE)** is an unbiased estimator of the standard deviation of  $\epsilon_i$ .

```
> zresid <- residuals(M1)/(M1$sum$sig)           # standardized residuals
> zresid                                         # standardized residuals and country names
> qqnorm(zresid, ylab="Standardized Residuals") # Q-Q plot
> abline(0,1)                                     # line 'y = x'
```

Under normality, we expect the points to follow the diagonal line **y = x**, approximately. Compare this Q-Q plot with that for the unstandardized residuals, shown earlier.

- ▷ Diagnostic purpose of standardized residuals: locating large errors in prediction.

For an overview of the names of the arguments that can be selected from the summary table of the function **lm**, use the following commands:

```
> ? summary.lm                                # summarizing linear model fits
> M1$sum$r.squared                           # R-squared
> M1$sum$adj.r.squared                        # adjusted R-squared
```

- **Studentized residuals**

```
> stud <- residuals(M1)/(M1$sum$sig*sqrt(1 - M1$inf$hat))
> stud                                         # Studentized residuals and country names
> qqnorm(stud, ylab="Studentized Residuals") # Q-Q plot
> abline(0,1)                                 # line 'y = x'
```

Here too, as the Studentized residuals are standardized, we expect the points to follow the diagonal line **y = x**, approximately, if normality holds.

- ▷ Diagnostic purpose of Studentized residuals: detection of outliers in response variable  $Y$ .

Notice that the Studentized residuals **stud**, as defined above, equal the standardized residuals as computed by function **rstandard()** in R.

```
> rstandard(M1)                                # standardized residuals in R
```

R has a different convention than usual (as in SPSS, for example) in defining standardized [function **rstandard()**] and Studentized residuals [function **rstUDENT()**], respectively. When computing these residuals for the  $i$ -th data point, the R function **rstandard()** uses an unbiased estimator of the standard deviation of the observed residuals (not the standard deviation of the error terms  $\epsilon_i$ , but that of  $e_i$ ). This now accounts for the fact that **rstandard** is equivalent with the usual formula for Studentized residuals (see Boomsma, 2010). The R function

`rstudent()`, on the other hand, calculates residuals from a regression where all points are used except observation  $i$ . The general idea of the functions `rstandard()` and `rstudent()` is to “renormalize the residuals to have unit variance, using an overall and leave-one-out measure of error variance respectively”; see `help(influence.measures)`.

With this knowledge, for plotting objectives we might as well use a fancier plotting function:

```
> plot(M1, which=2) # R's standardized residuals Q-Q plot
```

It should also be noticed that the package `stats` incorporates the general function `influence.measures(model)`, which covers a set of subfunctions for regression (leave-one-out deletion) diagnostics. Some of these functions will be addressed now. Again, for an overview see the R documentation:

```
> ? influence.measures # regression deletion diagnostics
```

- **Jackknifed Studentized residuals**

► Diagnostic purpose of Studentized deleted residuals: detection of influential observations, as well as for validation purposes.

```
> jack <- rstudent(M1) # leave-one-out Studentized residuals
> jack[which.max(abs(jack))]
```

```
Zambia
2.853558
```

This value of **2.85**, the largest Studentized deleted residual, is pretty large for a standard normal scale. But is it an outlier, we should ask. We could test whether this observation is an outlier, using a Student’s  $t$ -statistic with  $n - p - 1$  degrees of freedom, where  $p$  is the number of predictors in an intercept model. If we would use a Bonferroni correction to have a minimal overall  $\alpha$  level of 0.05, and a significance level  $\alpha/n$  for each individual test, the critical Bonferroni value is computed as follows. Notice that for the `savings` data  $n = 50$  and  $p = 5$ , hence  $df = 44$ .

```
> qt(.05/(50*2), 44) # quantile for two-sided alpha = 0.05
# in a Student's t-distribution with df = 44
[1] -3.525801
```

Since **2.85** is less than **3.52**, we conclude that Zambia is not an outlier.

The `car` package contains a Bonferroni outlier test which just calculates the very thing:

```
> library(car)
> outlier.test(M1)                                # equivalent result from the "car" package

max|rstudent| = 2.8536, degrees of freedom = 44,
unadjusted p = 0.0065667, Bonferroni p = 0.32833

Observation: Zambia
```

### 3.3 Influential Observations

- Cook's distance

Cook's distance measure is a combination of a residual effect and leverage, as shown by Equation 19 in Boomsma (2010). This combination leads to influence.

- ▷ Diagnostic purpose of Cook's distance measure: the detection of influential observations; detection of the joint influence of outliers, both in the response variable  $Y$  and the explanatory variables  $X_j$ .

A half-normal plot can be used to identify influential observations.

```
> (cook <- cooks.distance(M1))
> countries <- rownames(savings)
> halfnorm(cook, 3, labs=countries, ylab="Cook's distance")
> which.max(cook)

Libya
49
```

There are efficient alternative options in the `car` package:

```
> plot(cookd(M1))
> identify(1:50, cookd(M1), countries)
```

But there is also a diagnostic `lm` plotting function from R itself, providing direct identifying information:

```
> plot(M1, which=4)                                # Cook's distance measure
```

We can also plot leverage points against Cook's distance.

```
> plot(M1, which=6) # leverage against Cook's distance
```

If we exclude **Lybia**, we can examine how the fit of the linear regression model changes.

```
> M1_L <- lm(sr ~ pop15 + pop75 + dpi + ddpi, data=savings,
+   subset=(cook < max(cook)))
> summary(M1_L) # linear model estimates without Libya
```

Call:

```
lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = savings,
subset = (cook < max(cook)))
```

Residuals:

Min	1Q	Median	3Q	Max
-8.0699	-2.5408	-0.1584	2.0934	9.3732

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	24.5240460	8.2240263	2.982	0.00465
pop15	-0.3914401	0.1579095	-2.479	0.01708
pop75	-1.2808669	1.1451821	-1.118	0.26943
dpi	-0.0003189	0.0009293	-0.343	0.73312
ddpi	0.6102790	0.2687784	2.271	0.02812

Residual standard error: 3.795 on 44 degrees of freedom

Multiple R-Squared: 0.3554, Adjusted R-squared: 0.2968

F-statistic: 6.065 on 4 and 44 DF, p-value: 0.0005617

```
> M1_inf <- influence(M1); M1_inf$coef
```

Recall that in the coefficients matrix **M1\_inf\$coef**, the *i*th row contains the change in the estimated coefficients which results when the *i*th case is dropped from the regression.

```
> M1_inf$coef [,2]
```

The second column of **M1\_inf\$coef** is related to the regression coefficient of **pop15**, the first explanatory variable (after the intercept term).

```
> plot(M1_inf$coef [,2], ylab="Change in pop15 coefficient")
> abline(h=0)
> identify(1:50, M1_inf$coef [,2], countries) # identify plotted points
# use 'Esc' to leave plot
```

Here, we have plotted the change in the second parameter estimate when a single case is left out. The `identify()` function was used to identify plotted points. The country with the largest change could also be identified with the following command:

```
> which.max(abs(M1_inf$coef[,2]))
```

```
Japan
23
```

The previous plot should be repeated for the other coefficients. In the last plot, **Japan** is an influential observation. We might therefore examine the effect of removing this country from the sample data.

```
> M1_J <- lm(sr ~ pop15 + pop75 + dpi + ddpi, data=savings,
+   subset=(countries != "Japan"))
> summary(M1_J)                                # linear model estimates without Japan

Call:
lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = savings,
subset = (countries != "Japan"))

Residuals:
    Min      1Q  Median      3Q     Max 
-7.9969 -2.5918 -0.1150  2.0318 10.1571 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 23.9401714  7.7839968   3.076  0.00361  
pop15       -0.3679015  0.1536296  -2.395  0.02096  
pop75       -0.9736743  1.1554502  -0.843  0.40397  
dpi        -0.0004706  0.0009191  -0.512  0.61116  
ddpi        0.3347486  0.1984457   1.687  0.09871  

Residual standard error: 3.738 on 44 degrees of freedom
Multiple R-Squared: 0.277,      Adjusted R-squared: 0.2113 
F-statistic: 4.214 on 4 and 44 DF,  p-value: 0.005649
```

- ▷ Compare the results of this model with those of the full model.

#### 4. Checking the structure of the model

In this section we check whether the systematic part of the model,  $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ , is correct. Questions under investigation here are, for example: Does the linearity assumption hold? What is the effect of  $X_j$  on  $Y$ ?

- Added variable plot or partial regression plot

We could regress the  $X$ s on  $Y$  without explanatory variable  $X_j$ , and get residuals  $\hat{\delta}$  which represent  $Y$  with the other  $X$ -effect ( $j' \neq j$ ) taken out. Similarly, if we regress  $X_j$  on all  $X$  except  $X_j$ , we get residuals  $\hat{\gamma}$ , which represent  $X_j$  with the other  $X$ -effects taken out. The added variable plot shows  $\hat{\delta}$  against  $\hat{\gamma}$ . Look for non-linearity, outliers, and influential observations in the plot.

The estimated slope of a line fitted to this plot is  $b_j$ . The partial regression plot shows the marginal relationship between the response and an explanatory variable, after the effect of the other explanatory variables has been removed (partialled out). We focus here on the relationship between one predictor, `pop15`, and the response `sr`.

```
> delta <- residuals(lm(sr ~ pop75 + dpi + ddpi, data=savings))
> gamma <- residuals(lm(pop15 ~ pop75 + dpi + ddpi, data=savings))
> plot(gamma, delta, xlab="pop15 residuals", ylab="savings residuals")

> M1d <- lm(delta ~ gamma)           # linearity between residuals?
> coef(M1d)
  (Intercept)      gamma
  5.425926e-17 -4.611931e-01

> coef(M1)                      # coefficients of the full linear model
  (Intercept)      pop15      pop75       dpi       ddpi
  28.5660865407 -0.4611931471 -1.6914976767 -0.0003369019  0.4096949279

> abline(coef(M1d) ["(Intercept)", "gamma"], col="red")
> abline(0,coef(M1) ["pop15"], col="blue")
```

The added variable plot function `av.plots()` in the `car` package does a similar job. The reader might have inferred by now that `car` is the acronym of *Companion to Applied Regression*, the (2002) book of John Fox.

```
> av.plots(lm(sr ~ pop15 + pop75 + dpi + ddpi, data=savings))

> av.plots(M1)                      # for short
? av.plots
```

The help documentation of `av.plots()` shows options for variable selection and point identification.

- ▷ Notice that the slope in the residual plot and the slope for `pop15` in the full regression model are the same.

- **Partial residual plot**

This competitor of the added variable plot, plots  $e_i + b_j X_{ij}$  against  $X_j$ . Again, the estimated slope will be  $b_j$ . Partial residual plots are better for the detection of linearity, added variable plots are better for the detection of outliers and influential data points.

```
> plot(savings$pop15, residuals(M1)+coef(M1) ["pop15"]*savings$pop15,
+   xlab="pop15", ylab="Savings Adjusted")
> abline(0,coef(M1) ["pop15"])
```

More directly, the partial residual plot function `prplot()` from the `faraway` package can be used, which provides the same result.

Notice that the source file `wilcox14.R` contains a (different) function with the label `prplot()`, which might cause interaction problems (error messages) at some point — check with command `fix(prplot)`.

```
> source("wilcox14.R")          # load source file 'wilcox14.R'
> prplot(M1, 1)                # partial residual plot, where the second
                                # argument indexes the explanatory variable
```

The function `cr.plots` [`component + residual` (`partial residual`) `plots`] in the `car` package could also be used.

```
> cr.plots(lm(sr ~ pop15 + pop75 + dpi + ddpi, data=savings),
+   variable="pop15")
> cr.plots(M1, variable="pop15") # for short
```

It appears from these plots that there are different relationships in two groups: a group with a low percentage of the population under 15 years (`pop15`), and a group with a high percentage of `pop15`. A division could be made at `pop15 = 35`. We could, therefore, perform two separate analyses, one for each group. First we identify the groups, as follows:

```
> subset(savings, pop15 < 35)      # rich countries, it seems
> subset(savings, pop15 > 35)      # poor countries

> M1_low <- lm(sr ~ pop15+pop75+dpi+ddpi, data=savings, subset=(pop15 < 35))
> M1_high <- lm(sr ~ pop15+pop75+dpi+ddpi, data=savings, subset=(pop15 > 35))
```

```
> summary(M1_low)
```

Call:

```
lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = savings,  
subset = (pop15 < 35))
```

Residuals:

Min	1Q	Median	3Q	Max
-5.5890	-1.5015	0.1165	1.8857	5.1466

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	23.9617950	8.0837502	2.964	0.00716
pop15	-0.3858976	0.1953686	-1.975	0.06092
pop75	-1.3277421	0.9260627	-1.434	0.16570
dpi	-0.0004588	0.0007237	-0.634	0.53264
ddpi	0.8843944	0.2953405	2.994	0.00668

Residual standard error: 2.772 on 22 degrees of freedom

Multiple R-Squared: 0.5073, Adjusted R-squared: 0.4177

F-statistic: 5.663 on 4 and 22 DF, p-value: 0.002734

```
> summary(M1_high)
```

Call:

```
lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = savings,  
subset = (pop15 > 35))
```

Residuals:

Min	1Q	Median	3Q	Max
-5.55105	-3.51012	0.04428	2.67638	8.49830

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-2.4339689	21.1550278	-0.115	0.910
pop15	0.2738537	0.4391910	0.624	0.541
pop75	-3.5484769	3.0332806	-1.170	0.257
dpi	0.0004208	0.0050001	0.084	0.934
ddpi	0.3954742	0.2901012	1.363	0.190

Residual standard error: 4.454 on 18 degrees of freedom

Multiple R-Squared: 0.1558, Adjusted R-squared: -0.03185

F-statistic: 0.8302 on 4 and 18 DF, p-value: 0.5233

- Try to interpret the results of these analyses, and draw appropriate conclusions. Notice, for example, the different estimates of the residual standard errors in the two groups, and the different R-squared values.

## 5. More diagnostics

The general suite of functions `influence.measures(model)` also contains the functions `dfits(model)`, `dfbeta(model)`, `dfbetas(model)` and `dfbetas(model)`, as described by Boomsma (2010).

In Section 4 we have not discussed regression diagnostics with respect to the problem of multicollinearity. In practice, this potential problem should not be left unattended, of course. Many of the regression diagnostics described above can also be used for generalized linear model fitting. The `stats` package contains the workhorse function `gls()`, by which we can work with non-normal error distributions — like the families of binomial, Poisson and gamma distributions — and link functions as well.

## References

- Belsley, D.A., Kuh, E., & Welsch, R.E. (1980). *Regression diagnostics*. New York: Wiley.
- Boomsma, A. (2010). *An overview of regression diagnostics*. Unpublished manuscript, Department of Statistics & Measurement Theory, University of Groningen.
- Chambers, J.M., Cleveland, W.S., Kleiner, B., & Tukey, P.A. (1983). *Graphical methods for data analysis*. Belmont, CA: Wadsworth.
- Faraway, J.J. (2002). *Practical regression and anova using R*. Unpublished manuscript. Retrieved May 20, 2009, from <http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf>. [Data sets and scripts are also directly available from <http://www.maths.bath.ac.uk/~jjf23/book/>.]
- Faraway, J.J. (2005). *Linear models with R*. Boca Raton, FL: Chapman & Hall/CRC.
- Fox, J. (2002). *An R and S-Plus companion to applied regression*. Thousand Oaks, CA: Sage.
- Stevens, J. (1992). *Applied multivariate statistics for the social sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

*Regression Diagnostics with R*

Copyright © 2014 by Anne Boomsma, Vakgroep Statistiek & Meettheorie, Rijksuniversiteit Groningen

Alle rechten voorbehouden. Niets in deze uitgave mag worden verveelvoudigd, opgeslagen in een geautomatiseerd gegevensbestand, en/of openbaar gemaakt, in enige vorm of op enige wijze, hetzij electronisch, mechanisch, door fotocopie, microfilm of op enige andere manier, zonder voorafgaande schriftelijke toestemming van de auteur.

*All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the author.*

## Regression and Correlation

Topics Covered:

- Dependent and independent variables.
- Scatter diagram.
- Correlation coefficient.
- Linear Regression line.

by Dr.I.Namestnikova

## Introduction

Regression analysis is used to model and analyse numerical data consisting of values of an **independent variable  $X$**  (the variable that we fix or choose deliberately) and **dependent variable  $Y$** .

The main purpose of finding a relationship is that the knowledge of the relationship may enable events to be predicted and perhaps controlled.

## Correlation coefficient

To measure the strength of the linear relationship between  $X$  and  $Y$  the **sample correlation coefficient  $r$**  is used.

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}},$$

$$S_{xy} = n \sum xy - \sum x \sum y,$$

$$S_{xx} = n \sum x^2 - (\sum x)^2, \quad S_{yy} = n \sum y^2 - (\sum y)^2$$

Where  $x$  and  $y$  observed values of variables  $X$  and  $Y$  respectively.

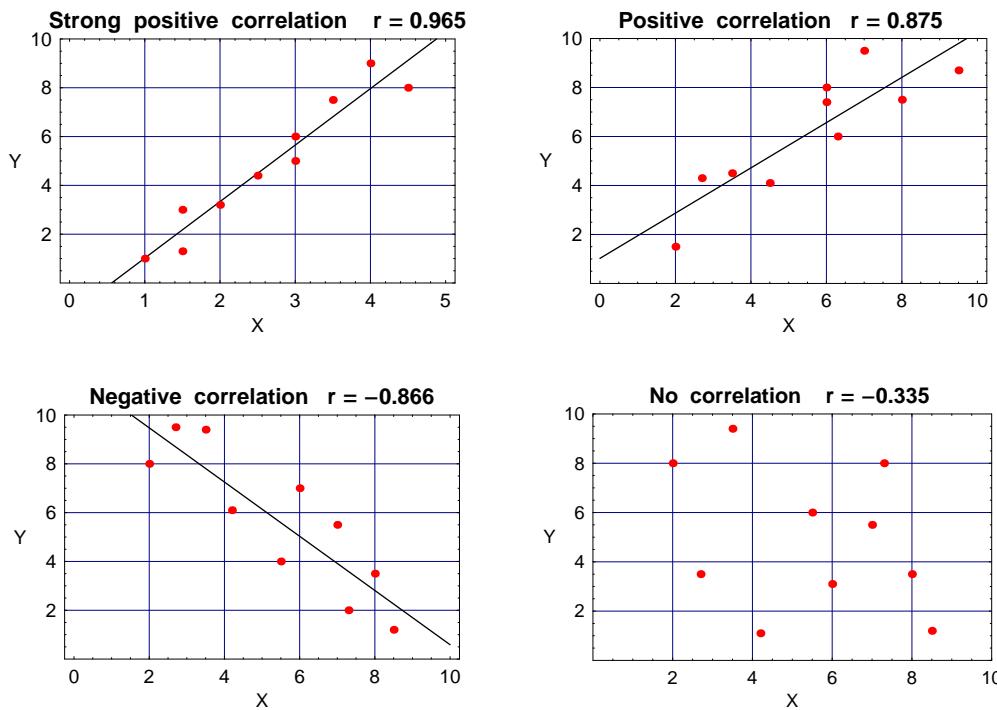
### Important notes

1. If the calculated  $r$  value is **positive** then the slope will **rise** from left to right on the graph. If the calculated value of  $r$  is **negative** the slope will **fall** from left to right.
2. The  $r$  value will **always** lie between **-1** and **+1**. If you have an  $r$  value outside of this range you have made an error in the calculations.
3. Remember that a correlation does not necessarily demonstrate a causal relationship. A significant correlation only shows that two factors vary in a related way (positively or negatively).
4. The formula above can be rewritten as

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}, \quad \sigma_x = \sqrt{\frac{1}{n} \sum x^2 - \bar{x}^2}, \quad \sigma_y = \sqrt{\frac{1}{n} \sum y^2 - \bar{y}^2}$$
$$\sigma_{xy} = \frac{1}{n} \sum xy - \bar{x}\bar{y}, \quad \bar{x} = \frac{1}{n} \sum x, \quad \bar{y} = \frac{1}{n} \sum y$$

## Scatter Diagrams

**Scatter diagrams** are used to graphically represent and compare two sets of data. The **independent variable** is usually plotted on the **X** axis. The **dependent variable** is plotted on the **Y** axis. By looking at a scatter diagram, we can see whether there is any connection (**correlation**) between the two sets of data. A scatter plot is a useful summary of a set of bivariate data (two variables), usually drawn before working out a linear correlation coefficient or fitting a regression line. It gives a good visual picture of the relationship between the two variables, and aids the interpretation of the correlation coefficient or regression model.



From plots one can see that if the more the points tend to cluster around a straight line and the higher the correlation (the stronger the **linear relationship** between the two variables). If there exists a random scatter of points, there is no relationship between the two variables (very low or zero correlation).

Very low or zero correlation could result from a non-linear relationship between the variables. If the relationship is in fact non-linear (points clustering around a curve, not a straight line), the correlation coefficient will not be a good measure of the strength. A scatter plot will also show up a **non-linear relationship** between the two variables and whether or not there exist any outliers in the data.

### Example 1

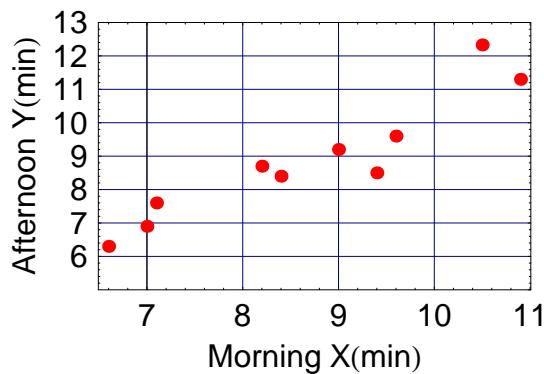
Determine on the basis of the following data whether there is a relationship between the time, in minutes, it takes a person to complete a task in the morning  $X$  and in the late afternoon  $Y$ .

Morning ( $x$ ) (min)	8.2	9.6	7.0	9.4	10.9	7.1	9.0	6.6	8.4	10.5
Afternoon ( $y$ ) (min)	8.7	9.6	6.9	8.5	11.3	7.6	9.2	6.3	8.4	12.33

### Solution

The data set consists of  $n = 10$  observations.

#### Step 1.



To construct the scatter diagram for the given data set to see any correlation between two sets of data.

From the scatter diagram we can conclude that it is likely that there is a linear relationship between two variables.

#### Step 2.

Set out a table as follows and calculate all required values  $\sum x$ ,  $\sum y$ ,  $\sum x^2$ ,  $\sum y^2$ ,  $\sum xy$ .

Morning ( $x$ ) (min)	Afternoon ( $y$ ) (min)	$x^2$	$y^2$	$xy$
8.2	8.7	67.24	75.69	71.34
9.6	9.6	92.16	92.16	92.16
7.0	6.9	49.00	47.61	48.30
9.4	8.5	88.36	72.25	79.90
10.9	11.3	118.81	127.69	123.17
7.1	7.6	50.41	57.76	53.96
9	9.2	81.00	84.64	82.80
6.6	6.3	43.56	39.69	41.58
8.4	8.4	70.56	70.56	70.56
10.5	12.33	110.25	151.29	129.465
$\sum x = 86.7$	$\sum y = 88.8$	$\sum x^2 = 771.35$	$\sum y^2 = 819.34$	$\sum xy = 792.92$

### Step 3.

Calculate

$$S_{xy} = n \sum xy - \sum x \sum y = 10 \times 792.92 - 86.7 \times 88.8 = 230.24$$

$$S_{xx} = n \sum x^2 - (\sum x)^2 = 10 \times 771.35 - (86.7)^2 = 196.61$$

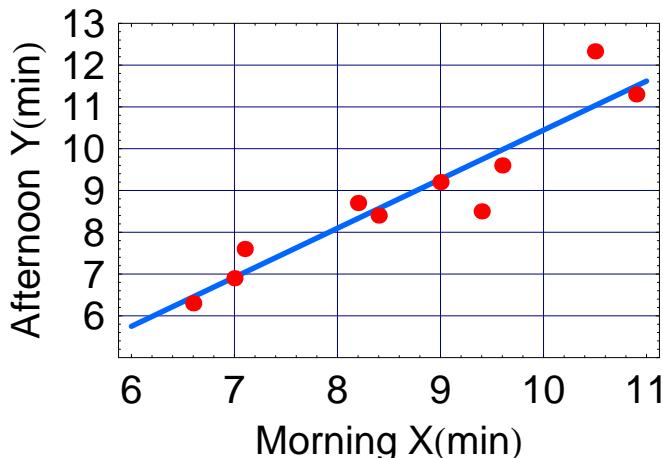
$$S_{yy} = n \sum y^2 - (\sum y)^2 = 10 \times 819.34 - 88.8^2 = 307.96$$

### Step 4

Finally we obtain **correlation coefficient r**

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{230.24}{\sqrt{196.61 \times 307.96}} = 0.9357$$

The correlation coefficient is closed to 1 therefore the linear relationship exists between the two variables.



It would be tempting to try to fit a line to the data we have just analysed - producing an **equation** that shows the relationship, The method for this is called **linear regression**. By using linear regression method the line of best fit is

$$\text{Regression equation: } y = 1.171x - 1.273$$

This line is shown in blue on the above graph. How to find this equation one can see in the next section.

## Linear regression analysis: fitting a regression line to the data

When a scatter plot indicates that there is a **strong linear relationship** between two variables (confirmed by **high correlation coefficient**), we can fit a straight line to this data which may be used to predict a value of the dependent variable, given the value of the independent variable.

Recall that the equation of a **regression line** (straight line) is

$$y = a + bx$$

$$b = \frac{S_{xy}}{S_{xx}} \quad a = \bar{y} - b\bar{x} = \frac{\sum_i y_i - b \sum_i x_i}{n}$$

To illustrate the technique, let us consider the following data.

### Example 2

Suppose that we had the following results from an experiment in which we measured the growth of a cell culture (as optical density) at different pH levels.

pH	3	4	4.5	5	5.5	6	6.5	7	7.5
Optical density	0.1	0.2	0.25	0.32	0.33	0.35	0.47	0.49	0.53

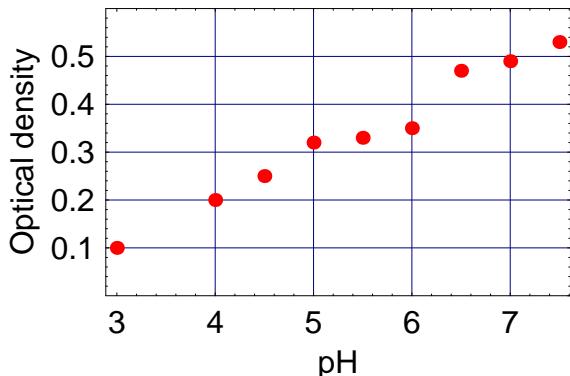
Find the equation to fit these data.

### Solution

We can follow the same procedures for correlation, as before.

The data set consists of  $n = 9$  observations.

**Step 1.** To construct the scatter diagram for the given data set to see any correlation between two sets of data.



These results suggest a linear relationship.

**Step 2.** Set out a table as follows and calculate all required values  $\sum x$ ,  $\sum y$ ,  $\sum x^2$ ,  $\sum y^2$ ,  $\sum xy$ .

pH ( $x$ )	Optical density( $y$ )	$x^2$	$y^2$	$xy$
3	0.1	9	0.01	0.3
4	0.2	16	0.04	0.8
4.5	0.25	20.25	0.0625	1.125
5	0.32	25	0.1024	1.6
5.5	0.33	30.25	0.1089	1.815
6	0.35	36	0.1225	2.1
6.5	0.47	42.25	0.2209	3.055
7	0.49	49	0.2401	3.43
7.5	0.53	56.25	0.2809	3.975
$x = 49$	$y = 3.04$	$x^2 = 284$	$y^2 = 1.1882$	$xy = 18.2$
$\bar{x} = 5.444$	$\bar{y} = 0.3378$			

### Step 3.

Calculate

$$S_{xy} = n \sum xy - \sum x \sum y = 9 \times 18.2 - 49 \times 3.04 \\ = 163.8 - 148.96 = 14.84.$$

$$S_{xx} = n \sum x^2 - (\sum x)^2 = 2556 - 2401 = 155.$$

$$S_{yy} = n \sum y^2 - (\sum y)^2 = 10.696 - 9.242 = 1.454$$

### Step 4.

Finally we obtain **correlation coefficient  $r$**

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{14.84}{\sqrt{155 \times 1.454}} = 0.989$$

The correlation coefficient is closed to 1 therefore it is likely that the linear relationship exists between the two variables. To verify the correlation  $r$  we can run a hypothesis test.

## Step 5. A hypothesis test

- **Hypothesis** about the **population** correlation coefficient  $\rho$

1. The **null hypothesis**  $H_0 : \rho = 0$ .
2. The **alternative hypothesis**  $H_A : \rho \neq 0$ .

- **Distribution of test statistic.** When  $H_0$  is true ( $\rho = 0$ ) and the assumption are met, the appropriate test statistic is distributed as **Student's  $t$  distribution**

( the **test statistics** is  $t = r \sqrt{\frac{n - 2}{1 - r^2}}$  with  $n - 2$  degrees of freedom).

The number of degrees of freedom is two less than the number of points on the graph ( $9 - 2 \equiv 7$  degrees of freedom in our example because we have 9 points).

- **Decision rule.** If we let  $\alpha = 0.025$ ,  $2\alpha = 0.05$ , the critical values of  $t$  in the present example are  $\pm 2.365$  (e.g. see John Murdoch, "Statistical tables for students of science, engineering, psychology, business, management, finance", 1998, Macmillan, 79 p., Table 7).

If, from our data, we compute a value of  $t$  that is either greater or equal to **2.365** or less than or equal to **-2.365**, we will reject the null hypothesis.

- **Calculation of test statistic.**

$$t = 0.989 \sqrt{\frac{7}{1 - 0.989^2}} = 17.69$$

- **Statistical decision.** Since the computed value of the test statistic exceed the critical value of  $t$ , we **reject** the null hypothesis.

- **Conclusion.** We conclude that there is a **very highly significant positive correlation** between pH and growth as measured by optical density of the cell culture.

### Step 6.

Now we want to use **regression analysis** to find the line of best fit to the data. We have done nearly all the work for this in the calculations above.

The **regression equation** for  $y$  on  $x$  is:  $y = bx + a$  where  $b$  is the **slope** and  $a$  is the **intercept** (the point where the line crosses the  $y$ -axis)

We calculate  $b$  and  $a$  as:

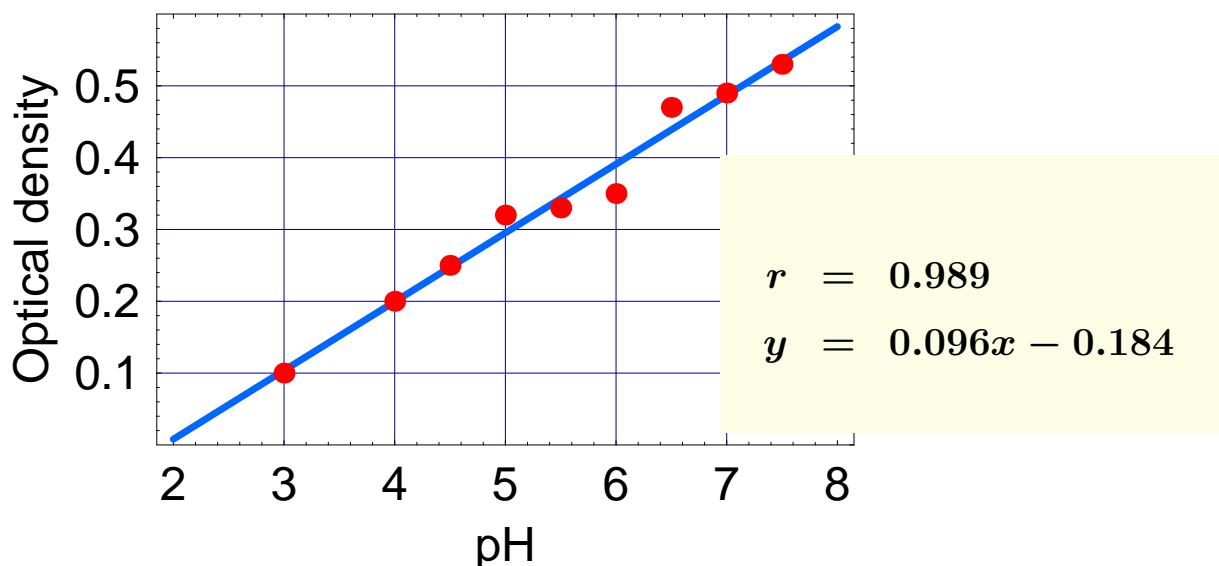
$$b = \frac{S_{xy}}{S_{xx}} = \frac{14.84}{155} = 0.096$$

$$\begin{aligned} a &= \bar{y} - b\bar{x} \\ &= 0.3378 - 0.096 \cdot 5.444 \\ &= 0.3378 - 0.5226 = -0.184 \end{aligned}$$

So the equation for the line of best fit is:

$$y = 0.096x - 0.184$$

(to 3 decimal places).



### Example 3

The tensile strength of a cable for upper-limb prosthesis was investigated. Stainless steel cable is commonly available in three sizes (diameters): **1.19 mm**, **1.59 mm** and **2.38 mm**. Four tests were performed for each diameter size and the results are given in the table below

Cable diameter ( $mm$ )	Cable cross area ( $mm^2$ )	Tensile strength (KN)
1.19	1.1122	1.27
	1.1122	1.45
	1.1122	1.43
	1.1122	1.36
1.59	1.9856	2.20
	1.9856	2.56
	1.9856	2.38
	1.9856	2.45
2.38	4.4488	4.58
	4.4488	5.03
	4.4488	5.67
	4.4488	4.39

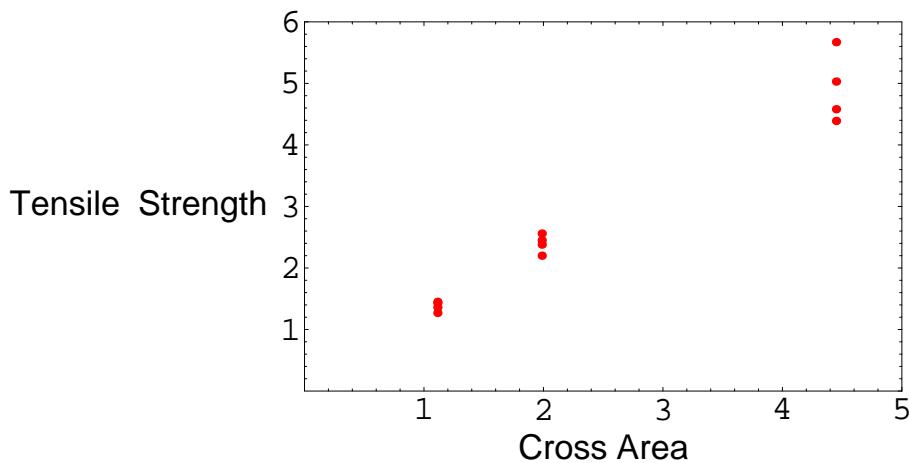
Let be  $X$  = Cable cross area ( $mm^2$ ),  $Y$  = Tensile strength (KN)

1. Construct a scatter diagram to illustrate these results.
2. Calculate the correlation coefficient for the data and comment on the result.
3. Obtain the least squares estimates for the sample regression equation of "Tensile strength" on "Cable cross area".
4. Estimate the tensile strength for cable with cross area 3.
5. Comment on the suitability of using the sample regression equation to estimate the tensile strength for cable with cross area 5.

## Solution

We can follow the same procedure, as before. The data set consists of  $n = 12$  observations.

1. To construct the scatter diagram for the given data set to see any correlation between two sets of data.



These results may suggest a linear relationship.

2. Set out a table as follows and calculate all required values  $\sum x$ ,  $\sum y$ ,  $\sum x^2$ ,  $\sum y^2$ ,  $\sum xy$ .

Cable cross area ( $x$ )	Tensile strength ( $y$ )	$x^2$	$y^2$	$xy$
1.1122	1.27	1.237	1.613	1.413
1.1122	1.45	1.237	2.103	1.613
1.1122	1.43	1.237	2.045	1.591
1.1122	1.36	1.237	1.850	1.513
1.9856	2.20	3.943	4.840	4.368
1.9856	2.56	3.943	6.554	5.083
1.9856	2.38	3.943	5.664	4.726
1.9856	2.45	3.943	6.003	4.867
4.4488	4.58	19.792	20.976	20.376
4.4488	5.03	19.792	25.301	22.378
4.4488	5.67	19.792	32.149	25.225
4.4488	4.39	19.792	19.272	19.530
$x = 30.19$	$y = 34.77$	$x^2 = 99.89$	$y^2 = 128.37$	$xy = 112.68$
$\bar{x} = 2.5158$	$\bar{y} = 2.8975$			

Calculate

$$S_{xy} = n \sum xy - \sum x \sum y = 12 \times 112.68 - 30.19 \times 34.77 \\ = 1352.16 - 1049.71 = 302.454$$

$$S_{xx} = n \sum x^2 - (\sum x)^2 = 1198.68 - 911.436 = 287.244$$

$$S_{yy} = n \sum y^2 - (\sum y)^2 = 1540.44 - 1208.95 = 331.487$$

Finally we obtain **correlation coefficient  $r$**

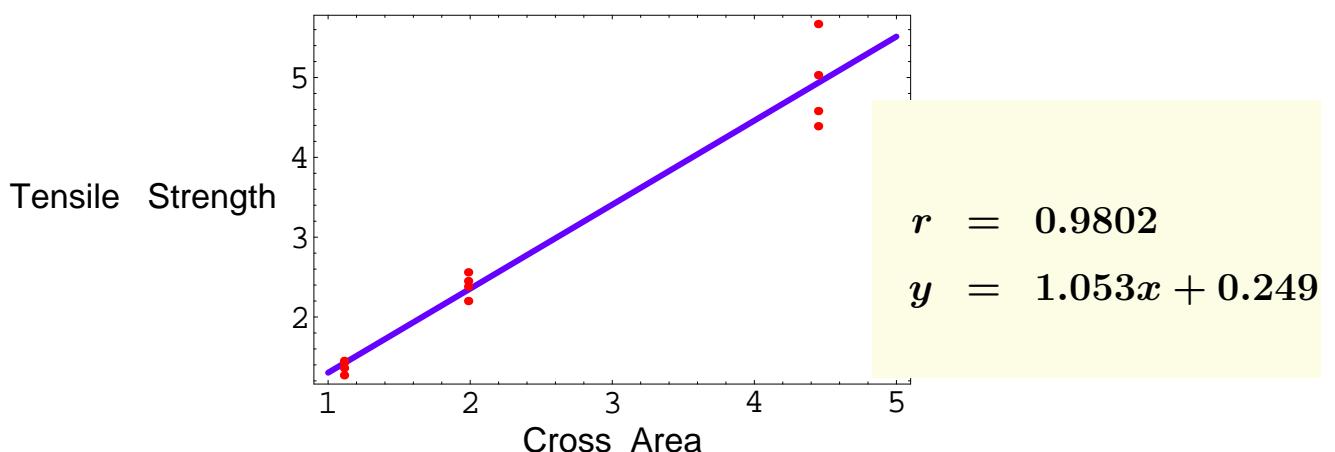
$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{302.454}{\sqrt{287.244 \times 331.487}} = 0.9802$$

The correlation coefficient is closed to 1 therefore it is likely that the linear relationship exists between the two variables. To verify the correlation we can run a hypothesis test.

**3.** We calculate  $b$  and  $a$  as:

$$b = \frac{S_{xy}}{S_{xx}} = \frac{302.454}{287.244} = 1.053 \\ a = \bar{y} - b\bar{x} = 2.8975 - 1.053 \cdot 2.5158 = 0.249$$

So the equation for the line of best fit is:  $y = 1.053x + 0.249$  (to 3 decimal places).



4. To estimate the tensile strength for cable with cross area **3** we need to substitute  $x = 3$  into the regression equation  $y(3) = 1.053 \times 3 + 0.249 = 3.408$
5. The sample regression equation is not suitable to estimate the tensile strength for cable with cross area **5** because this value is outside the test range ( $1.1 \leq x \leq 4.45$ ).

#### Example 4

To make a prosthesis we must know the force that acts in it as the person moves. This force depends on the adjacent musculature. Records of the variation with time of the force in hip joints during level walking show two maximum values in the stance phase of each cycle. A total of **16** subjects took part in the study. An indication of the variation in average of the maxima hip joint force with body weight  **$W$**  and the ratio of stride length  **$L$**  to height  **$H$**  are given in the table below.

$\frac{WL}{H}$ (kg)	33.6	41.4	43.3	44.1	45.6	46.0	49.8	53.2	53.8	54.7	55.2	58.3	59.7	62.2	66.3	72.1
Mean hip joint force $F$ (kN)	1.400	1.300	1.050	1.320	1.200	1.107	1.560	2.070	2.200	1.730	1.870	2.520	2.370	2.640	2.380	2.850

Let be  $X = \frac{WL}{H}$  and mean hip joint force  $Y$

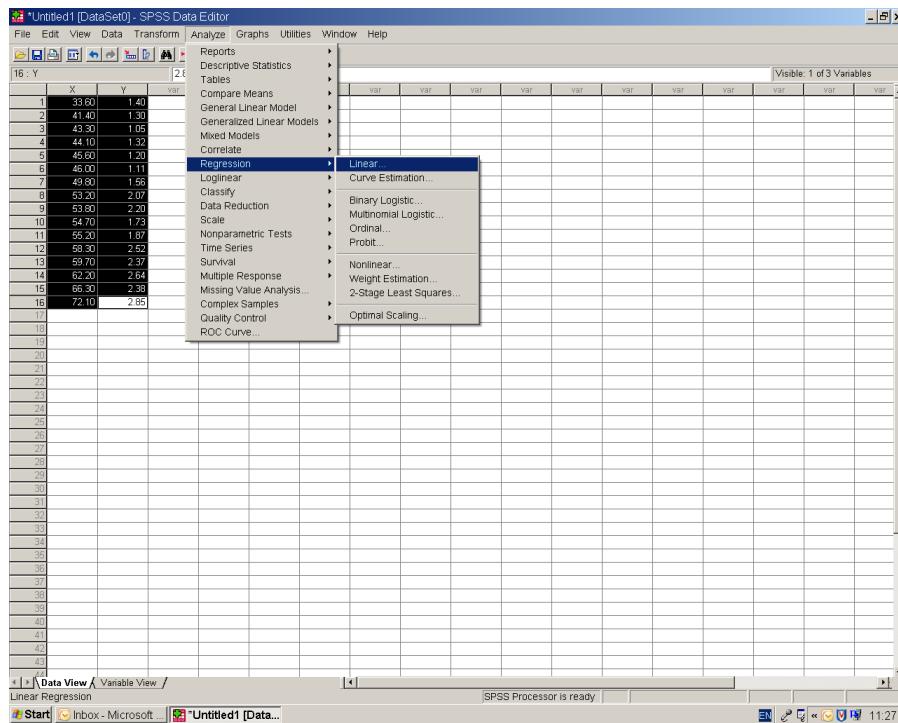
- Construct a scatter diagram to illustrate these results.
- Calculate the correlation coefficient for the data and comment on the result.
- Obtain the least squares estimates for the sample regression equation of " $\frac{WL}{H}$ " on "Mean hip joint force".

#### Solution

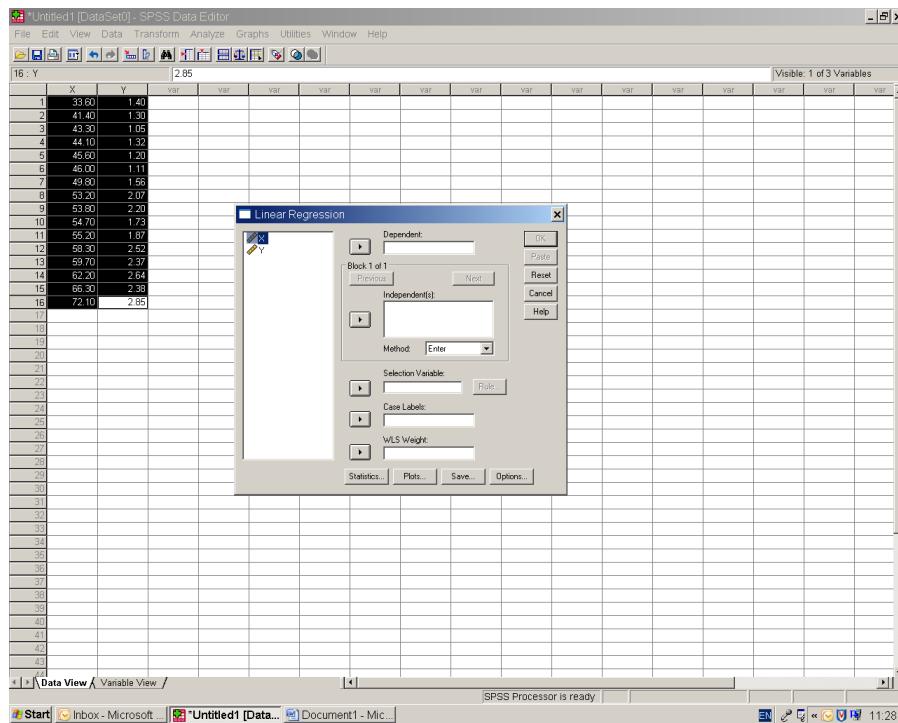
We can follow the same procedure, as before. Another option is to use SPSS or Excel.

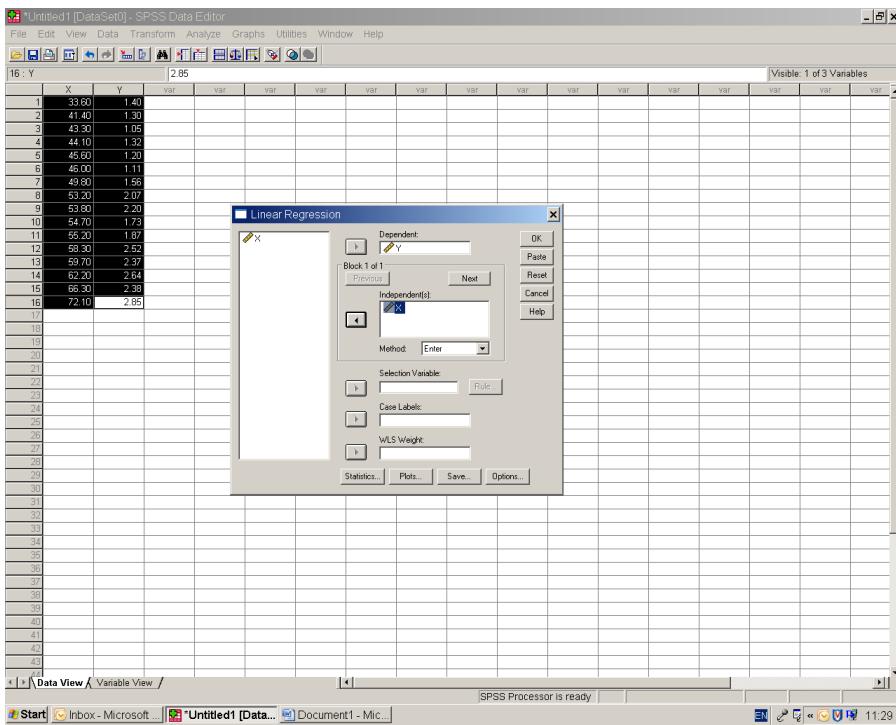
## Example of using SPSS for Regression Analysis

Open Data.sav. Click Analyze, click Regression and click Linear.



In the **Linear Regression** dialog box click **X** and click an arrow to move **X** into **Independent Variable** list. In the **Linear Regression** dialog box click **Y** and click an arrow to move **Y** into **Dependent Variable** list.





Then click OK

**REGRESSION**  
 /MISSING LISTWISE  
 /STATISTICS COEFF OUTS R ANOVA  
 /CRITERIA=PIN(.05) POUT(.10)  
 /NOORIGIN  
 /DEPENDENT Y  
 /METHOD=ENTER X .

**♦ Regression**

**Variables Entered/Removed<sup>a</sup>**

Model	Variables Entered	Variables Removed	Method
1	X		Enter

a. All requested variables entered.  
 b. Dependent Variable: Y

**Model Summary**

Model	R	R Square	Adjusted R Square	Std Error of the Estimate
1	.8883*	.788	.773	28338

a. Predictors: (Constant), X

You can use the **Output Viewer** to browse results.

The screenshot shows the SPSS Output Viewer window titled "SPSSexampleOutput1.spo [Document2] - SPSS Viewer". The menu bar includes File, Edit, View, Data, Transform, Insert, Format, Analyze, Graphs, Utilities, Window, and Help. The toolbar contains various icons for file operations like Open, Save, Print, and zoom. The left pane displays a tree view of the output structure:

- Output
- Log
- Regression
  - Title
  - Notes
  - Active Dataset
  - Variables Entered/Removed
  - Model Summary
  - ANOVA
  - Coefficients
- Interactive Graph
  - Title
  - Notes
  - Active Dataset
  - Scatterplot

The right pane displays the following tables:

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.888 <sup>a</sup>	.788	.773	.29336

a. Predictors: (Constant), X

**ANOVA<sup>b</sup>**

Model	Sum of Squares	df	Mean Square	F	Sig.
1	Regression 4.170	1	4.170	51.938	.000 <sup>a</sup>
	Residual 1.124	14	.080		
	Total 5.294	15			

a. Predictors: (Constant), X  
b. Dependent Variable: Y

**Coefficients<sup>a</sup>**

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1	(Constant) -.917	.390		-2.350	.034
	X .053	.007	.888	7.207	.000

a. Dependent Variable: Y

SPSS Processor is unavailable

You can add a scatter plot. Click **Graphs**, click **Iteractive**, than click **Scatterplot**. The scatter plot can be edited, just double click on it.

\*Untitled1 [DataSet0] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

16 : Y 2.85 Visible: 2 of 2 Variables

	X	Y	var																
1	33.60	1.40																	
2	41.40	1.30																	
3	43.30	1.05																	
4	44.10	1.32																	
5	45.60	1.20																	
6	46.00	1.11																	
7	49.80	1.56																	
8	53.20	2.07																	
9	53.80	2.20																	
10	54.70	1.73																	
11	55.20	1.87																	
12	58.30	2.52																	
13	59.70	2.37																	
14	62.20	2.64																	
15	66.30	2.38																	
16	72.10	2.85																	
17																			
18																			
19																			
20																			
21																			
22																			
23																			
24																			
25																			
26																			
27																			
28																			
29																			
30																			
31																			
32																			
33																			
34																			
35																			
36																			
37																			
38																			
39																			
40																			
41																			
42																			
43																			
44																			

Data View Variable View / SPSS Processor is ready

Output1 [Document1] - SPSS Viewer

File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help

Interactive Legacy Dialogs Chart Builder... Bar... Dot... Line... Ribbon... Drop-Line... Area... Pie Boxplot... Error Bar... Histogram... Scatterplot...

REGRESSION  
/MISSING LISTWISE  
/STATISTICS COEFF OUTS R ANOVA  
/CRITERIA=PIN(.05) POUT(.10)  
/NOORIGIN  
/DEPENDENT Y  
/METHOD=ENTER X .

Regression

[DataSet0]

Variables Entered/Removed<sup>b</sup>

Model	Variables Entered	Variables Removed	Method
1	X <sup>a</sup>	.	Enter

a. All requested variables entered.  
b. Dependent Variable: Y

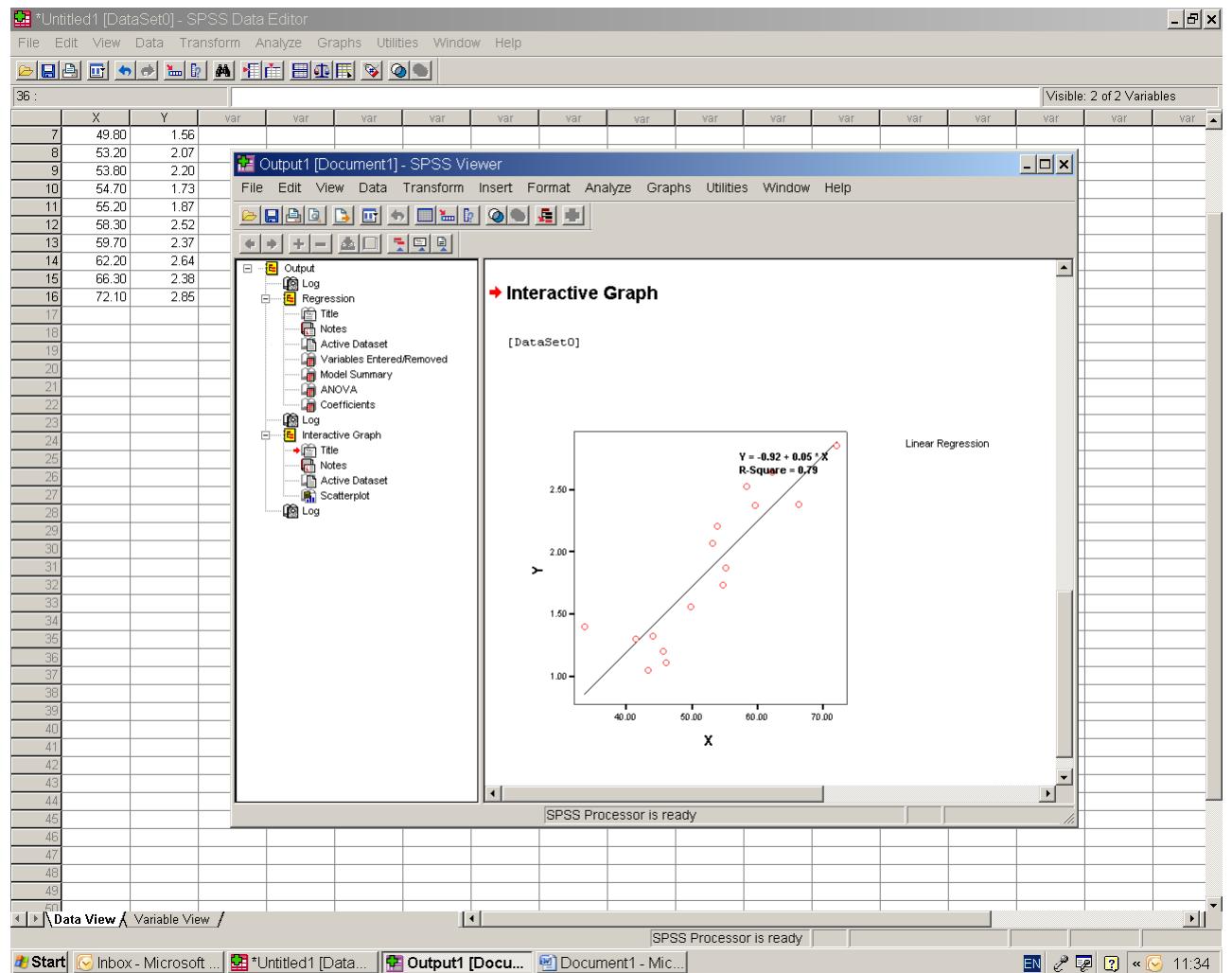
Model Summary

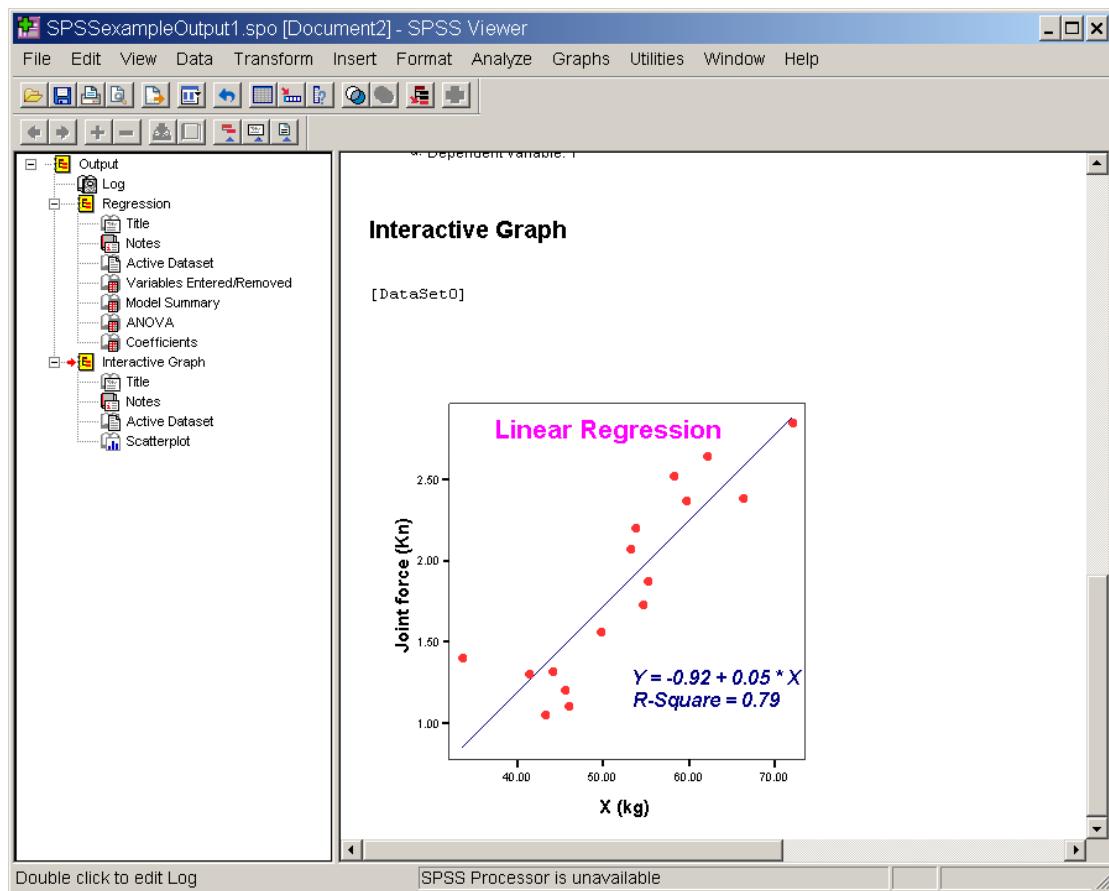
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.888a	.788	.773	.28336

a. Predictors: (Constant), X

Interactive Scatterplot SPSS Processor is ready

Start Inbox - Microsoft ... \*Untitled1 [Data... Output1 [Docu... Document1 - Mic... SPSS Processor is ready EN 11:31





# Regression through the Origin

## KEYWORDS:

*Teaching;*  
*Regression;*  
*Analysis of variance;*  
*Statistical software packages.*

*Joseph G. Eisenhauer*

Canisius College, Buffalo, USA.  
e-mail: eisenhauer@canisius.edu

## Summary

This article describes situations in which regression through the origin is appropriate, derives the normal equation for such a regression and explains the controversy regarding its evaluative statistics. Differences between three popular software packages that allow regression through the origin are illustrated using examples from previous issues of *Teaching Statistics*.

## ◆ INTRODUCTION ◆

Although ordinary least-squares (OLS) regression is one of the most familiar statistical tools, far less has been written – especially in the pedagogical literature – on regression through the origin (RTO). Indeed, the subject is surprisingly controversial. The present note highlights situations in which RTO is appropriate, discusses the implementation and evaluation of such models and compares RTO functions among three popular statistical packages. Some examples gleaned from past *Teaching Statistics* articles are used as illustrations. For expository convenience, OLS and RTO refer here to linear regressions obtained by least-squares methods with and without a constant term, respectively.

## ◆ MODEL SELECTION: ◆ WHEN IS RTO APPROPRIATE?

Textbooks rarely discuss RTO other than to caution against dropping the constant term from a regression, on the grounds that imposing any such restriction can only diminish the model's fit to the data. There are, however, circumstances in which RTO is appropriate or even necessary.

First, RTO may be unavoidable if transformations of the OLS model are needed to correct violations of the Gauss–Markov assumptions. Consider, for example, the simple linear regression of  $Y$  on  $x$

$$Y_i = \beta_0 + \beta_1 x_i + e_i \quad (1)$$

where  $\beta_0$  is the intercept,  $\beta_1$  is the slope and  $e_i$  denotes the  $i$ th residual. Lagging observations and taking first differences (i.e. subtracting each observation from its successor) to correct for serial correlation in the errors requires transforming equation (1) into an RTO equation of the form

$$Y_i - Y_{i-1} = \beta_1(x_i - x_{i-1}) + (e_i - e_{i-1})$$

Alternatively, applying weighted least squares to correct for heteroscedasticity will result in a model with no intercept if the weighting factor ( $z$ ) is not an independent variable. In that case,  $\beta_0$  becomes a coefficient and equation (1) is replaced by a multiple linear regression without a constant:

$$Y_i/z_i = \beta_0(1/z_i) + \beta_1(x_i/z_i) + (e_i/z_i)$$

Even without such transformations, however, there are often strong a priori reasons for believing that  $Y = 0$  when  $x = 0$ , and therefore omitting the constant. Indeed, Theil (1971, p. 176) contends 'From an economic point of view, a constant term usually has little or no explanatory virtues'. While that may be a slight exaggeration – it is easy to find examples in which an intercept does matter – there are certainly cases in which economic theory posits the absence of a constant. The widely used Cobb–Douglas production function, for example, relates output ( $Y$ ) to capital ( $K$ ) and labour ( $L$ ) according to  $Y = K^{\beta_1} L^{\beta_2}$ , and taking logarithms yields  $\ln Y = \beta_1 \ln K + \beta_2 \ln L$ ; imposing a constant

on this model would imply an unrealistic ability to manufacture goods without resources. An agricultural example is provided by Chambers and Dunstan (1986), who regress sugar cane harvests on farmland acreage; clearly, if no land is cultivated, there will be no crop. Casella (1983, p. 150) suggests an engineering example in which gasoline usage is a simple linear function of vehicular weight; he reasons that, in principle, a weightless vehicle would consume no fuel, so ‘considering the physical constraints ... it seems most appropriate to fit a line through the origin’. And Adelman and Watkins (1994) apply RTO to the valuation of mineral deposits. Of course, similar instances can be found in almost any discipline; some ornithological and nutritional examples are discussed below.

Even when theory proscribes a constant, however, careful consideration of the observed range of data is needed. As Hocking (1996, p. 177) points out, ‘if the data are far from the origin, we have no evidence that the linearity applies over this expanded range. For example, the response may increase exponentially near the origin and then stabilize into a near linear response in the region of typical inputs.’ Alternatively, observations at the origin may represent a discontinuity from an otherwise linear function with a positive or negative intercept. Under those circumstances, knowing that  $Y = 0$  when  $x = 0$  is insufficient justification for RTO.

If there is uncertainty regarding the appropriateness of including an intercept, several diagnostic devices can provide guidance. Most obviously, one can run the OLS regression and test the null hypothesis  $H_0 : \beta_0 = 0$  using the Student’s  $t$  statistic to determine whether the intercept is significant. Alternatively, Hahn (1977) suggests running the regression with and without an intercept, and comparing the standard errors to decide whether OLS or RTO provides a superior fit. And Casella (1983) suggests artificially creating an extra observation – a leverage point – that pulls the OLS regression line naturally through the origin. Unless the data set is small and the observations cluster near the origin, any such leverage point is likely to be an outlier but, if it appears to be a plausible extrapolation of the actual data, one may conclude that RTO is an acceptable model. Unfortunately, there are infinitely many such leverage points that could be chosen for that exercise, and the reasonableness of RTO will depend on which point is used.

---

## ◆ IMPLEMENTATION AND ◆ EVALUATION OF RTO

---

In one respect, RTO is merely a special case of OLS, and the absence of the constant is actually a simplification. Indeed, minimizing the sum of squared errors for the simple linear RTO model

$$Y_i = \beta x_i + e_i$$

involves far less calculation than it does for the OLS model of equation (1). The problem

$$\min_{\beta} \sum (Y_i - \beta x_i)^2 = \sum Y_i^2 - 2\beta \sum x_i Y_i + \beta^2 \sum x_i^2$$

has only one normal equation or first-order condition

$$-2 \sum x_i Y_i + 2\hat{\beta} \sum x_i^2 = 0$$

and the easily derived second-order condition,  $2\sum x_i^2 > 0$ , clearly guarantees a minimum. From the normal equation, the estimated slope of the regression line is

$$\hat{\beta} = \sum x_i Y_i / \sum x_i^2$$

as noted by, for example, Pettit and Peers (1991). (For weighted versions, see Turner, 1960.)

Unfortunately, the RTO residuals will usually have a nonzero mean, because forcing the regression line through the origin is generally inconsistent with the best fit. The proper method for evaluating RTO has long been disputed (see, for example, Marquardt and Snee 1974; Maddala 1977; Gordon 1981). To appreciate the controversy, note the familiar identity

$$(Y_i - \bar{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}) \quad (2)$$

where  $\bar{Y}$  denotes the mean of the dependent variable and  $\hat{Y}_i$  is the  $i$ th fitted value. Squaring both sides and summing across all observations gives

$$\begin{aligned} \sum (Y_i - \bar{Y})^2 &= \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \bar{Y})^2 \\ &\quad + 2\sum (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) \end{aligned}$$

but, as is well known, the cross-product term is equal to zero in the case of OLS. The remaining terms therefore constitute the usual analysis of variance decomposition

$$\Sigma(Y_i - \bar{Y})^2 = \Sigma(Y_i - \hat{Y}_i)^2 + \Sigma(\hat{Y}_i - \bar{Y})^2 \quad (3)$$

where the left-hand side is the sum of squares total (SST), the first term on the right is the sum of squares due to error (SSE) and the final term is the sum of squares due to regression (SSR). The coefficient of determination for OLS is then defined by the ratio of SSR to SST

$$R^2 = \frac{\Sigma(\hat{Y}_i - \bar{Y})^2}{\Sigma(Y_i - \bar{Y})^2}$$

or equivalently

$$R^2 = 1 - \frac{\Sigma(Y_i - \hat{Y}_i)^2}{\Sigma(Y_i - \bar{Y})^2} \quad (4)$$

Some authors maintain that because this diagnostic measure is based on an identity, it should not depend on the inclusion or exclusion of a constant term in the regression. From that perspective, equation (4) is equally valid for RTO and OLS.

However, when there is no constant in the regression,  $\Sigma(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})$  will generally take a nonzero value, so equation (3) is not a valid basis for analysis of variance in RTO. And if the RTO model provides a sufficiently poor fit, the data may exhibit more variation around the regression line than around  $\bar{Y}$ , in which case  $\Sigma(Y_i - \hat{Y}_i)^2 > \Sigma(Y_i - \bar{Y})^2$ . Heedlessly applying equation (4) would then result in an implausibly negative (and thus uninterpretable) coefficient of determination as well as a negative  $F$  ratio. Moreover, it is often argued that defining SST as the sum of squared deviations from the mean is inappropriate when the regression line is forced through the origin but does not necessarily pass through  $(\bar{x}, \bar{Y})$ ; when so viewed, equation (2) is replaced by the identity

$$(Y_i - 0) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - 0) \quad (2')$$

Squaring and summing yields

$$\Sigma Y_i^2 = \Sigma(Y_i - \hat{Y}_i)^2 + \Sigma \hat{Y}_i^2 + 2\Sigma \hat{Y}_i(Y_i - \hat{Y}_i)$$

but the final (cross-product) term in this equation equals zero under RTO, because

$$\begin{aligned} \Sigma \hat{Y}_i(Y_i - \hat{Y}_i) &= \sum \hat{\beta} x_i (Y_i - \hat{\beta} \bar{x}_i) = \hat{\beta} [\sum x_i Y_i - \hat{\beta} \sum x_i^2] \\ &= \hat{\beta} [\sum x_i Y_i - (\sum x_i Y_i / \sum x_i^2) \sum x_i^2] = 0 \end{aligned}$$

Thus, equation (3) is replaced by

$$\Sigma Y_i^2 = \Sigma(Y_i - \hat{Y}_i)^2 + \Sigma \hat{Y}_i^2 \quad (3')$$

Applying equation (3') rather than equation (3) to RTO, one finds that SSE is unchanged, but SST =  $\Sigma Y_i^2$  and SSR =  $\Sigma \hat{Y}_i^2$ . Redefining SST and SSR in this manner results in

$$R^2 = \frac{\Sigma \hat{Y}_i^2}{\Sigma Y_i^2} \quad (4')$$

a strictly non-negative coefficient of determination that equals or exceeds the measure in equation (4). Of course, these definitions also affect the adjusted  $R^2$  and  $F$  statistics, but do not alter the standard error of the regression ( $S_e$ ). Note that, without a constant, the degrees of freedom for SST, SSR and SSE are  $n$ ,  $k$  and  $n - k$ , respectively, where  $n$  is the sample size and  $k$  is the number of independent variables; thus,  $S_e = \sqrt{SSE/(n - k)}$  regardless of how SST is defined.

The controversy over SST is not merely academic: practitioners (and students) running RTO will obtain various outputs depending on which computer packages they use. Indeed, as Prvan et al. (2002, p. 74) observed in a recent comparison of Minitab, SPSS and Excel, ‘Obtaining a simple linear regression is easy in all three packages, and all three give the standard output options (such as regression through the origin)’. But in fact the three packages all give different outputs for RTO! Two illustrative examples are provided below.

## ◆ EXAMPLES ◆

In Kimber’s essay on the shape of birds’ eggs, egg height is regressed on width, both with and without an intercept (Kimber 1995). Her study of 281 species can be approximately replicated using the 96 observations provided in the Data Bank section of the Summer 1990 issue of *Teaching Statistics* (Data Bank 1990). Regardless of which computer package is used, OLS yields the following output:

$$\begin{aligned} \text{Height} &= -1.774 + 1.444 \text{ Width} \\ &\quad [0.001] [0.000] \\ S_e &= 2.123 \quad F = 5720.076 \quad R^2 = 0.984 \quad \bar{R}^2 = 0.984 \end{aligned}$$

where two-tailed  $p$ -values are shown in brackets below the estimates. Notice that the intercept is statistically significant; although it is, of course, impossible for an egg to have a zero width, the intercept may nevertheless be important, as it represents the extrapolation of the regression line back to the vertical axis. The effect of removing the intercept can be seen by running RTO. If Excel is used, as it was by Kimber, RTO yields

$$\text{Height} = 1.382 \text{ Width}$$

[0.000]

$$S_e = 2.251 \quad F = 5073.616 \quad R^2 = 0.9816 \quad \bar{R}^2 = 0.9711$$

which indicates a poorer fit by all diagnostic measures: the standard error,  $F$  and  $R^2$  (adjusted and unadjusted). However, the SPSS linear regression procedure without an intercept yields

$$\text{Height} = 1.382 \text{ Width}$$

[0.000]

$$S_e = 2.251 \quad F = 24,283.995 \quad R^2 = 0.996 \quad \bar{R}^2 = 0.996$$

Notice that the regression equation and standard error are the same in the two programs, but the  $F$  and  $R^2$  statistics are different. Indeed, in SPSS these statistics seem to indicate a better fit without the intercept than with it. The discrepancy between software packages arises because Excel is based on equations (3) and (4), while the RTO function in SPSS uses equations (3') and (4'). The SPSS output, however, is accompanied by the disclaimer ‘For regression through the origin (the no-intercept model), R Square measures the proportion of the variability in the dependent variable about the origin explained by regression. This CANNOT be compared to R Square for models which include an intercept’ [emphasis in original].

To make matters more confusing still, SPSS offers a nonlinear regression option, which requires a model statement and initial parameter values. If one uses the nonlinear option but specifies the linear model and a reasonable initial value for the slope, this option yields results identical to those for Excel – that is, it applies equation (4) to compute  $R^2$ ! Meanwhile, the Minitab option for RTO gives the same regression equation and standard error as Excel and SPSS, but reports neither the  $F$  nor the  $R^2$  statistic. However, Minitab’s ANOVA table, from which  $F$  and  $R^2$  would be derived, is based on equation (3').

Because Excel and the nonlinear option in SPSS apply equation (4) regardless of whether an intercept is present, it is easy (and perhaps instructive for students) to construct examples that generate negative  $R^2$  and  $F$  statistics for regressions through the origin using these packages. (One need only construct a line with a large intercept and then estimate it without the intercept.) Extreme cases of that sort can provide a springboard for discussion, and make a compelling argument for using equation (4') rather than equation (4) to evaluate RTO.

The same issues arise, of course, in multiple linear regressions. Consider the nutritional study conducted by Johnson (1995): the caloric contents of various foods are regressed on their fat, protein and carbohydrate contents. For the 13 foods in his sample, OLS yields

$$\text{Calories} = 4.446 + 8.715 \text{ Fat} + 4.044 \text{ Protein}$$

[0.395] [0.000] [0.000]

+ 3.841 Carbohydrates

[0.000]

$$S_e = 6.97 \quad F = 232 \quad R^2 = 0.987 \quad \bar{R}^2 = 0.983$$

regardless of which statistical software is used. But here the constant is insignificant and, as Johnson observes, nutritional theory indicates that a constant is inappropriate for this regression. In SPSS, removing the constant gives

$$\text{Calories} = 8.888 \text{ Fat} + 4.266 \text{ Protein}$$

[0.000] [0.000]

+ 3.978 Carbohydrates

[0.000]

$$S_e = 6.90 \quad F = 1459.66 \quad R^2 = 0.998 \quad \bar{R}^2 = 0.997$$

with all diagnostics indicating an improved fit. Minitab and Excel produce the same equation but different diagnostics. Minitab again reports only  $S_e$ , while Excel generates

$$\text{Calories} = 8.888 \text{ Fat} + 4.266 \text{ Protein}$$

[0.000] [0.000]

+ 3.978 Carbohydrates

[0.000]

$$S_e = 6.895 \quad F = 236.5 \quad R^2 = 0.986 \quad \bar{R}^2 = 0.883$$

In contrast to the previous example, the Excel output now seems more confusing than the SPSS output. Notice that Excel’s  $R^2$  and adjusted  $R^2$  statistics for RTO indicate a worse fit, while its  $S_e$  and  $F$  statistics indicate a better fit, compared to the OLS model.

Given these inconsistencies, Hocking (1996, p. 178) notes: 'It is natural to ask if there is a measure analogous to  $R^2$  for the no-intercept model. We suggest the square of the sample correlation between observed and predicted values'. It can easily be shown that this measure is equal to the unadjusted coefficient of determination for the OLS model. It therefore gives an interpretable measure of the quality of an RTO model, but does not help in comparing RTO with OLS. For that purpose, the best measures appear to be the  $p$ -value of the OLS constant and the standard errors of the OLS and RTO regressions. Using these measures, the constant should be retained in the eggs example given above, but not in the nutrition example.

## ◆ CONCLUSION ◆

Regression through the origin is an important and useful tool in applied statistics, but it remains a subject of pedagogical neglect, controversy and confusion. Hopefully, this synthesis provides some clarity. However, in the light of the unresolved debate, perhaps the strongest conclusion to be drawn from this review is that the practice of statistics remains as much an art as it is a science, and the development of statistical judgment is therefore as important as computational skill.

### Acknowledgements

The author would like to thank Donald Dale, Scott Trees and Luigi Ventura, whose discussions of results in another paper prompted him to write this one. He also thanks the editor and anonymous referees for providing helpful comments. Any errors are his own.

### References

- Adelman, M.A. and Watkins, G.C. (1994). Reserve asset values and the Hotelling valuation principle: further evidence. *Southern Economic Journal*, **61**(1), 664–73.
- Casella, G. (1983). Leverage and regression through the origin. *American Statistician*, **37**(2), 147–52.
- Chambers, R.L. and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, **73**(3), 597–604.
- Data Bank (1990). Birds, eggs, and databases. *Teaching Statistics*, **12**(2), 62–3.
- Gordon, H.A. (1981). Errors in computer packages: least squares regression through the origin. *The Statistician*, **30**(1), 23–9.
- Hahn, G.J. (1977). Fitting regression models with no intercept term. *Journal of Quality Technology*, **9**(2), 56–61.
- Hocking, R.R. (1996). *Methods and Applications of Linear Models: Regression and Analysis of Variance*. New York: John Wiley.
- Johnson, R. (1995). A multiple regression project. *Teaching Statistics*, **17**(2), 64–6.
- Kimber, H. (1995). The 'golden egg'. *Teaching Statistics*, **17**(2), 34–7.
- Maddala, G.S. (1977). *Econometrics*. New York: McGraw-Hill.
- Marquardt, D.W. and Snee, R.D. (1974). Test statistics for mixture models. *Technometrics*, **16**(4), 533–7.
- Pettit, L.I. and Peers, H.W. (1991). An example not to be followed? *Teaching Statistics*, **13**(1), 8.
- Prvan, T., Reid, A. and Petocz, P. (2002). Statistical laboratories using Minitab, SPSS, and Excel: a practical comparison. *Teaching Statistics*, **24**(2), 68–75.
- Theil, H. (1971). *Principles of Econometrics*. New York: John Wiley.
- Turner, M.E. (1960). Straight line regression through the origin. *Biometrics*, **16**(3), 483–5.

# Simple Linear Regression with R

## Getting and Opening Data Files

We will use an example data set from *Regression Analysis by Example* (4th ed.) by Chatterjee and Hadi (Wiley, New York, 2006). Go to the web site for this book at <http://www.ilr.cornell.edu/~hadi/rabe4/>. We will use the computer repair data. In this study a random sample of service call records for a computer repair operation were examined and the length of each call (in minutes) and the number of components repaired or replaced were recorded. The data are in file P027.txt. Follow the directions on the book's home page to download this and save it in the R folder on your computer. Then read the file into R as shown below. (header=TRUE means the file contains names for the variables on the first line.)

```
> repairs = read.table("P027.txt", header=TRUE)
> attach(repairs)
> repairs
   Minutes Units
1       23     1
2       29     2
3       49     3
4       64     4
5       74     4
6       87     5
7       96     6
8       97     6
9      109    7
10      119    8
11      149    9
12      145    9
13      154   10
14      166   10
```

## Simple Plots for Each Variable

Of course, the first step is to look at your data.

```
> stem(Minutes)

The decimal point is 2 digit(s) to the right of the |

 0 | 23
 0 | 5679
 1 | 0012
 1 | 5557

> stem(Units)

The decimal point is at the |

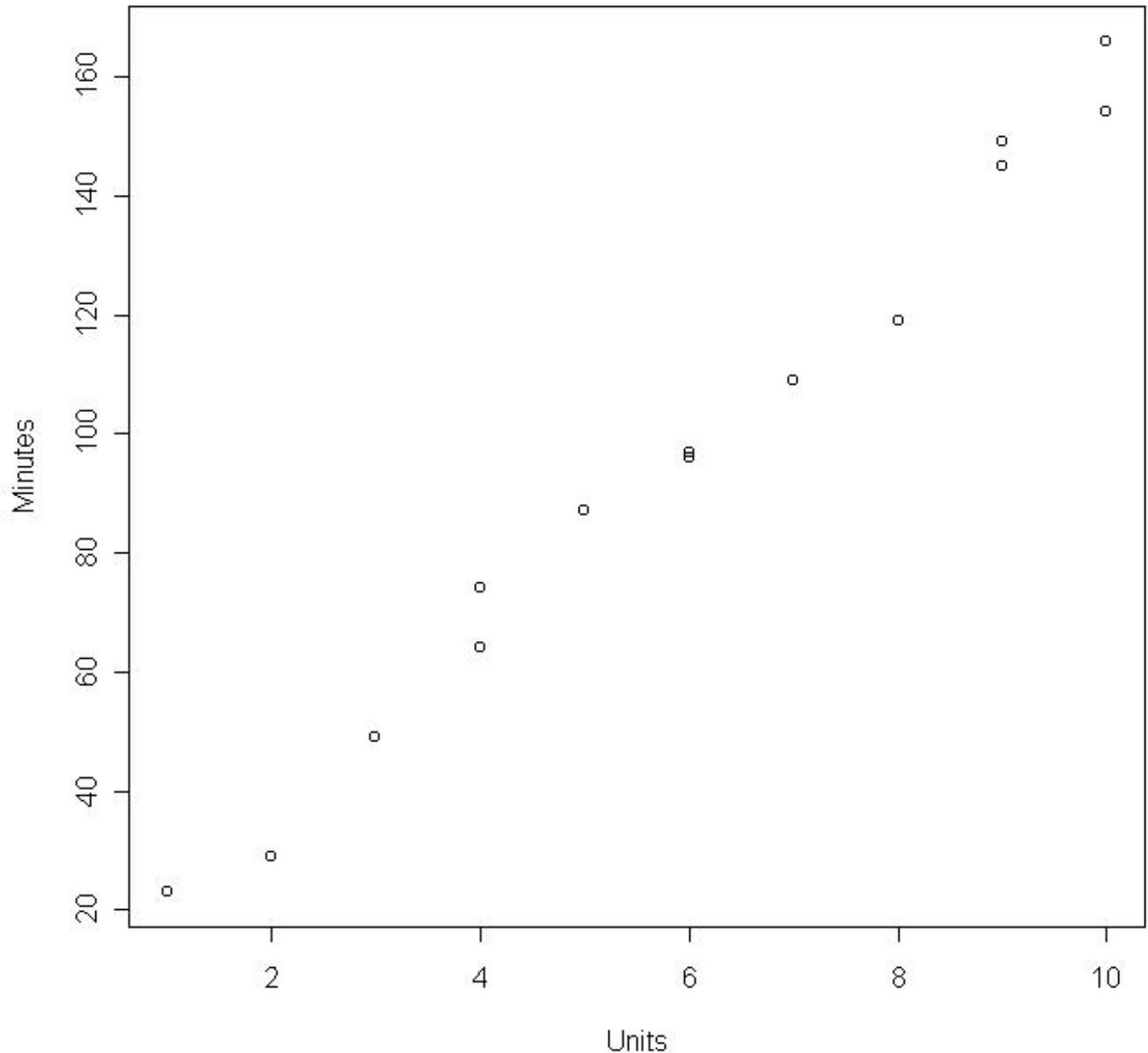
 0 | 0
 2 | 00
 4 | 000
 6 | 000
 8 | 000
10 | 00
```

We could have made histograms or boxplots. We simply want to see if there are any peculiarities in the data for

each variable by itself before we look into relationships between variables. We see none here.

## Scatterplots

```
> plot(Units,Minutes)
```



Note that the first variable in the `plot` command is plotted on the horizontal axis. We are not surprised to see that the length of a service call increases with the number of components repaired or replaced.

## Correlation and Covariance

```
> cor(Units,Minutes)
[1] 0.9936987
> cov(Units,Minutes)
[1] 136
```

## Running the Regression

The regression command is `lm` for linear model. We will store that model in a variable called `model`. The order of the variables is dependent followed by a tilde "~" followed by a list of independent variables.

```
> model = lm(Minutes ~ Units)
> model

Call:
lm(formula = Minutes ~ Units)

Coefficients:
(Intercept)      Units
        4.162     15.509

> summary(model)

Call:
lm(formula = Minutes ~ Units)

Residuals:
    Min      1Q  Median      3Q      Max 
-9.2318 -3.3415 -0.7143  4.7769  7.8033 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.162     3.355     1.24    0.239    
Units       15.509     0.505    30.71 8.92e-13 ***  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.392 on 12 degrees of freedom
Multiple R-Squared:  0.9874,    Adjusted R-squared:  0.9864 
F-statistic: 943.2 on 1 and 12 DF,  p-value: 8.916e-13
```

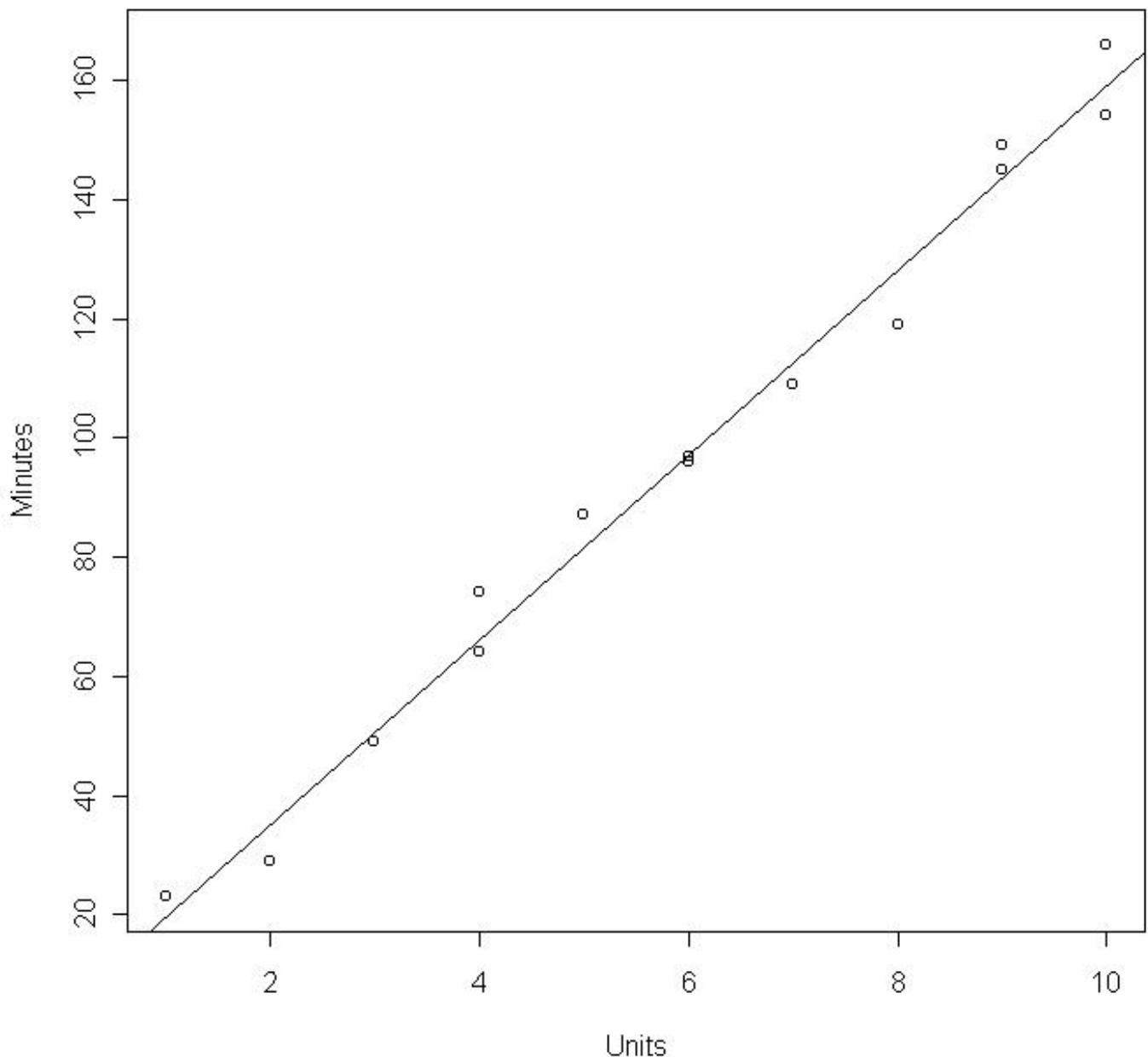
The regression equation is  $\text{minutes} = 4.162 + 15.509 * \text{units}$ . The "t values" test the hypotheses that the corresponding population parameters are 0. Usually we test whether the slope is zero because if it is then the model is not much use. Here the

p

-value for that test is "8.92e-13" which is to say  $8.92 \times 10^{-13}$  or 0.000000000000892, so we would reject the hypothesis that the slope is zero. If you wish to test a nonzero value, subtract it from the coefficient in the regression output (15.509) and divide the result by the coefficient's s.e. (0.505). (Use a calculator for this.) Similarly, if you want confidence intervals, use the coefficient plus or minus the product of its s.e. with a t-value for the desired confidence level and 12 degrees of freedom. (Use a calculator for this.) This also works for the intercept (4.162) using its s.e. (3.355).

To plot the regression line on the scatterplot, type

```
> abline(model)
```



You can cut and paste R output into your own reports but note that the text windows on the statistics.com Assignments page will only accept text input. So, of the output examples above, the scatterplots could *not* be pasted there. All the text that appears showing our interaction with R *can* be pasted into Assignments. To copy the contents of a graphics window (say for a report you are writing with your word processor), first click on File in the graph window, then select any of the first three options.

## Regression through the Origin

To fit a regression line through the origin (i.e.,  $\text{intercept}=0$ ) redo the regression but this time include that 0 in the model specification.

```
> model2 = lm(Minutes ~ 0 + Units)
> summary(model2)
```

```
Call:
lm(formula = Minutes ~ 0 + Units)
```

Residuals:

Min	1Q	Median	3Q	Max
-9.5955	-2.4733	0.4417	5.0243	9.7023

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
Units	16.0744	0.2213	72.63	<2e-16 ***
---				
Signif. codes:	0	'***'	0.001	'**'
			0.01	'*'
			0.05	'..'
			0.1	'.'
			1	''

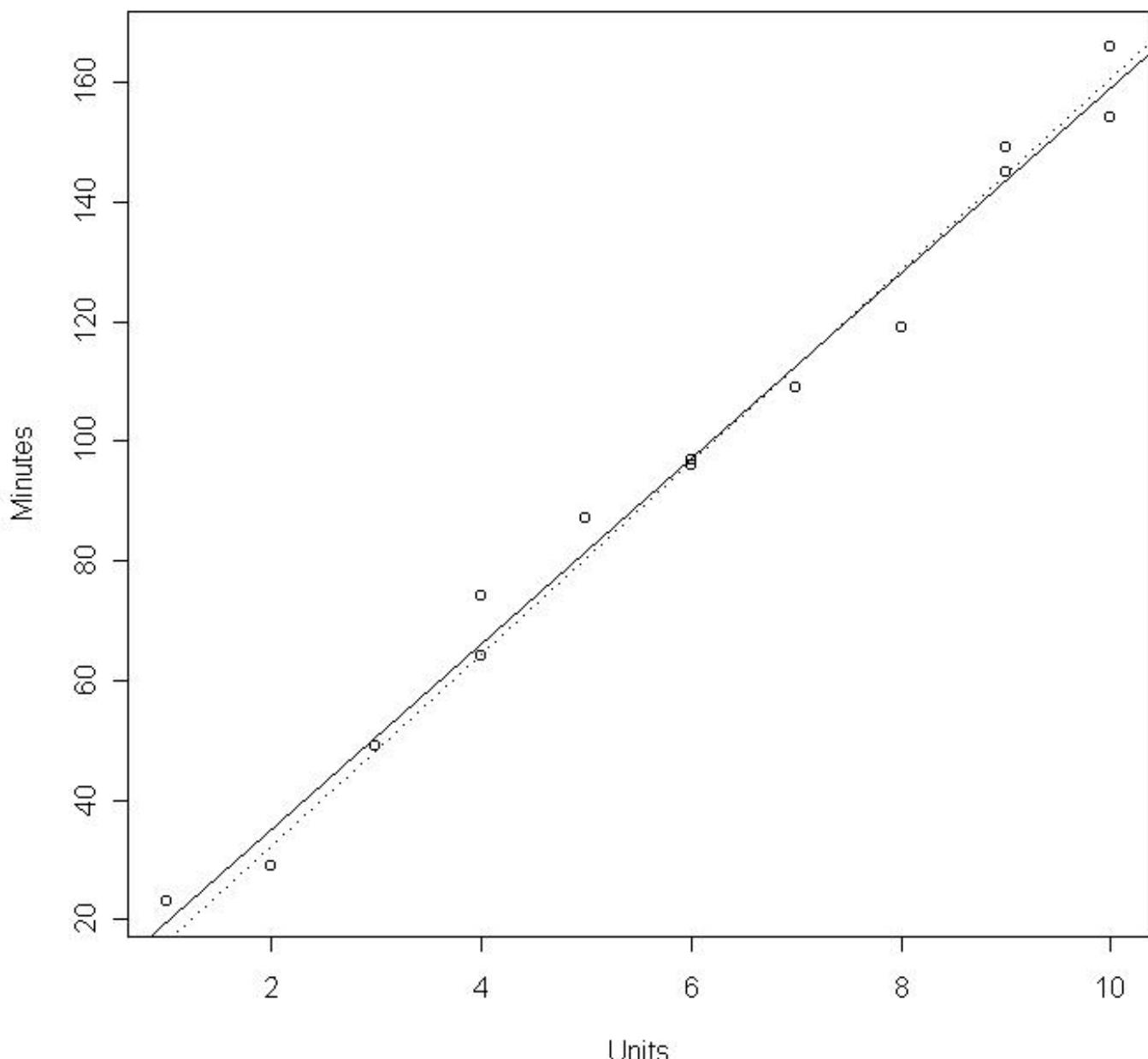
Residual standard error: 5.502 on 13 degrees of freedom

Multiple R-Squared: 0.9975, Adjusted R-squared: 0.9974

F-statistic: 5274 on 1 and 13 DF, p-value: < 2.2e-16

or  $\text{minutes} = 16.0744 * \text{units}$ . We can add this line to the graph to see how different it is.

```
> abline(model2, lty = "dotted")
```



Not much.

## Predictions

We predicted the length of a service call with four components repaired or replaced, then got confidence intervals for the prediction of a single observation and for the mean of all observations with `Units=4` (and based on our first model).

```
> predict(model, newdata=data.frame(Units=4))  
[1] 66.19674  
> predict(model, newdata=data.frame(Units=4), interval = "pred")  
    fit      lwr      upr  
[1,] 66.19674 53.83936 78.55413  
> predict(model, newdata=data.frame(Units=4), interval = "confidence")  
    fit      lwr      upr  
[1,] 66.19674 62.36271 70.03077
```

The syntax is tortured and will not be explained here.

---

© 2007, 2008 statistics.com

Mathematics Stack Exchange is a question and answer site for people studying math at any level and professionals in related fields. It's 100% free, no registration required.

[Take the 2-minute tour](#) ×

## Why the sum of residuals equals 0 when we do a sample regression by OLS?

That's my question, I have looking round online and people post a formula by they don't explain the formula. Could anyone please give me a hand with that ? cheers

(statistics) (statistical-inference)

asked Sep 15 '13 at 7:37

 **Maximilian1988**  
314 3 8

### 3 Answers

If the OLS regression contains a constant term, i.e. if in the regressor matrix there is a regressor of a series of ones, then the sum of residuals is *exactly* equal to zero, as a matter of algebra.

I will show this for the simple regression, it is trivial to see that it holds for the multivariate regression.

Specify the regression model

$$y_i = a + bx_i + u_i, \quad i = 1, \dots, n$$

Then the OLS estimator  $(\hat{a}, \hat{b})$  minimizes the sum of squared residuals, i.e.

$$(\hat{a}, \hat{b}) : \sum_{i=1}^n (y_i - \hat{a} - \hat{b}x_i)^2 = \min$$

For the OLS estimator to be the *argmin* of the objective function, it must be the case as a necessary condition, that the first partial derivatives with respect to  $a$  and  $b$ , evaluated at  $(\hat{a}, \hat{b})$  equal zero. For our result, we need only consider the partial w.r.t.  $a$ :

$$\frac{\partial}{\partial a} \sum_{i=1}^n (y_i - a - bx_i)^2 \Big|_{(\hat{a}, \hat{b})} = 0 \Rightarrow -2 \sum_{i=1}^n (y_i - \hat{a} - \hat{b}x_i) = 0$$

But  $y_i - \hat{a} - \hat{b}x_i = \hat{u}_i$ , i.e. is equal to the residual, so we have that

$$\sum_{i=1}^n (y_i - \hat{a} - \hat{b}x_i) = \sum_{i=1}^n \hat{u}_i = 0$$

The above also implies that if the regression specification does *not* include a constant term, then the sum of residuals will not, in general, be zero.

answered Sep 17 '13 at 21:40

 **Alecos Papadopoulos**  
5,296 1 4 17

Thansk a lot, I didnt see your answer till now, really appreciate it :) – [Maximilian1988](#) Sep 19 '13 at 4:13

Sum of residuals doesn't exactly equal 0. However it is a very reasonable assumption that the expectation of the residuals will be 0. This is similar to the case of unbiased estimation, where we want the bias to be 0. Here the residual  $y_i - \beta_0 - \beta_1 x_i$  are sometimes negative sometimes positive, but we hope that their overall sum will be 0, so that the estimation is good enough.

answered Sep 15 '13 at 13:01

 **Abishanka Saha**  
1,792 1 9

take the estimated values from the line of best fit and use these y values to subtract from the original y values then add them up. if it is a good line of best fit then it should approach zero, but bad lines of best fit will be much less or more than zero

answered Oct 16 '13 at 1:35



kasey

1

# What Are Degrees of Freedom?

Shanta Pandey and Charlotte Lyn Bright

**A**s we were teaching a multivariate statistics course for doctoral students, one of the students in the class asked, "What are degrees of freedom? I know it is not good to lose degrees of freedom, but what are they?" Other students in the class waited for a clear-cut response. As we tried to give a textbook answer, we were not satisfied and we did not get the sense that our students understood. We looked through our statistics books to determine whether we could find a more clear way to explain this term to social work students. The wide variety of language used to define *degrees of freedom* is enough to confuse any social worker! Definitions range from the broad, "Degrees of freedom are the number of values in a distribution that are free to vary for any particular statistic" (Healey, 1990, p. 214), to the technical:

Statisticians start with the number of terms in the sum [of squares], then subtract the number of mean values that were calculated along the way. The result is called the degrees of freedom, for reasons that reside, believe it or not, in the theory of thermodynamics. (Norman & Streiner, 2003, p. 43)

Authors who have tried to be more specific have defined degrees of freedom in relation to sample size (Trochim, 2005; Weinbach & Grinnell, 2004), cell size (Salkind, 2004), the number of relationships in the data (Walker, 1940), and the difference in dimensionalities of the parameter spaces (Good, 1973). The most common definition includes the number or pieces of information that are free to vary (Healey, 1990; Jaccard & Becker, 1990; Pagano, 2004; Warner, 2008; Wonnacott & Wonnacott, 1990). These specifications do not seem to augment students' understanding of this term. Hence, degrees of freedom are conceptually difficult but are important to report to understand statistical analysis. For example, without degrees of freedom, we are unable to calculate or to understand any

underlying population variability. Also, in a bivariate and multivariate analysis, degrees of freedom are a function of sample size, number of variables, and number of parameters to be estimated; therefore, degrees of freedom are also associated with statistical power. This research note is intended to comprehensively define degrees of freedom, to explain how they are calculated, and to give examples of the different types of degrees of freedom in some commonly used analyses.

## DEGREES OF FREEDOM DEFINED

In any statistical analysis the goal is to understand how the variables (or parameters to be estimated) and observations are linked. Hence, degrees of freedom are a function of both sample size ( $N$ ) (Trochim, 2005) and the number of independent variables ( $k$ ) in one's model (Toothaker & Miller, 1996; Walker, 1940; Yu, 1997). The degrees of freedom are equal to the number of independent observations ( $N$ ), or the number of subjects in the data, minus the number of parameters ( $k$ ) estimated (Toothaker & Miller, 1996; Walker, 1940). A parameter (for example, slope) to be estimated is related to the value of an independent variable and included in a statistical equation (an additional parameter is estimated for an intercept in a general linear model). A researcher may estimate parameters using different amounts or pieces of information, and the number of independent pieces of information he or she uses to estimate a statistic or a parameter are called the degrees of freedom ( $df$ ) (HyperStat Online, n.d.). For example, a researcher records income of  $N$  number of individuals from a community. Here he or she has  $N$  independent pieces of information (that is,  $N$  points of incomes) and one variable called income ( $k$ ); in subsequent analysis of this data set, degrees of freedom are associated with both  $N$  and  $k$ . For instance, if this researcher wants to calculate sample variance to understand the extent to which incomes vary in this community, the degrees of freedom equal  $N - k$ . The relationship between sample size and degrees of freedom is

positive; as sample size increases so do the degrees of freedom. On the other hand, the relationship between the degrees of freedom and number of parameters to be estimated is negative. In other words, the degrees of freedom decrease as the number of parameters to be estimated increases. That is why some statisticians define degrees of freedom as the number of independent values that are left after the researcher has applied all the restrictions (Rosenthal, 2001; Runyon & Haber, 1991); therefore, degrees of freedom vary from one statistical test to another (Salkind, 2004). For the purpose of clarification, let us look at some examples.

### A Single Observation with One Parameter to Be Estimated

If a researcher has measured income ( $k = 1$ ) for one observation ( $N = 1$ ) from a community, the mean sample income is the same as the value of this observation. With this value, the researcher has some idea of the mean income of this community but does not know anything about the population spread or variability (Wonnacott & Wonnacott, 1990). Also, the researcher has only one independent observation (income) with a parameter that he or she needs to estimate. The degrees of freedom here are equal to  $N - k$ . Thus, there is no degree of freedom in this example ( $1 - 1 = 0$ ). In other words, the data point has no freedom to vary, and the analysis is limited to the presentation of the value of this data point (Wonnacott & Wonnacott, 1990; Yu, 1997). For us to understand data variability,  $N$  must be larger than 1.

### Multiple Observations ( $N$ ) with One Parameter to Be Estimated

Suppose there are  $N$  observations for income. To examine the variability in income, we need to estimate only one parameter (that is, sample variance) for income ( $k$ ), leaving the degrees of freedom of  $N - k$ . Because we know that we have only one parameter to estimate, we may say that we have a total of  $N - 1$  degrees of freedom. Therefore, all univariate sample characteristics that are computed with the sum of squares including the standard deviation and variance have  $N - 1$  degrees of freedom (Warner, 2008).

Degrees of freedom vary from one statistical test to another as we move from univariate to bivariate and multivariate statistical analysis, depending on the nature of restrictions applied even when

sample size remains unchanged. In the examples that follow, we explain how degrees of freedom are calculated in some of the commonly used bivariate and multivariate analyses.

### Two Samples with One Parameter (or $t$ Test)

Suppose that the researcher has two samples, men and women, or  $n_1 + n_2$  observations. Here, one can use an independent samples  $t$  test to analyze whether the mean incomes of these two groups are different. In the comparison of income variability between these two independent means (or  $k$  number of means), the researcher will have  $n_1 + n_2 - 2$  degrees of freedom. The total degrees of freedom are the sum of the number of cases in group 1 and group 2 minus the number of groups. As a case in point, see the SAS and SPSS outputs of a  $t$  test comparing the literacy rate (LITERACY, dependent variable) of poor and rich countries (GNPSPLIT, independent variable) in Table 1. All in all, SAS output has four different values of degrees of freedom (two of which are also given by SPSS). We review each of them in the following paragraphs.

The first value for degrees of freedom under  $t$  tests is 100 (reported by both SAS and SPSS). The two groups of countries (rich and poor) are assumed to have equal variances in their literacy rate, the dependent variable. This first value of degrees of freedom is calculated as  $n_1 + n_2 - 2$  (the sum of the sample size of each group compared in the  $t$  test minus the number of groups being compared), that is,  $64 + 38 - 2 = 100$ .

For the test of equality of variance, both SAS and SPSS use the  $F$  test. SAS uses two different values of degrees of freedom and reports folded  $F$  statistics. The numerator degrees of freedom are calculated as  $n_1 - 1$ , that is  $64 - 1 = 63$ . The denominator degrees of freedom are calculated as  $n_2 - 1$  or  $38 - 1 = 37$ . These degrees of freedom are used in testing the assumption that the variances in the two groups (rich and poor countries, in our example) are not significantly different. These two values are included in the calculations computed within the statistical program and are reported on SAS output as shown in Table 1. SPSS, however, computes Levene's weighted  $F$  statistic (see Table 1) and uses  $k - 1$  and  $N - k$  degrees of freedom, where  $k$  stands for the number of groups being compared and  $N$  stands for the total number of observations in the sample; therefore, the degrees of freedom associated with the Levene's  $F$  statistic

**Table 1: SAS and SPSS Outputs of a t Test Comparing the Literacy Rate (Dependent Variable) of Poor and Rich Countries (Independent Variable) with Reported Degrees of Freedom (N = 102)**

The SAS System										SPSS SYSTEM										
The TTEST Procedure										Group Statistics										
Variable	GNPSPLIT	Statistics				GNPSPLIT				N		M		SD		SEM				
		Lower CL	Upper CL	Mean	Mean	Lower CL	Upper CL	Mean	Std Dev	0 (poor)	64	46.563	25.647	88.974	18.0712	3.2059	2.9315			
LITERACY	poor(0)	64	40.156	46.563	52.969	21.846	25.647			1 (rich)	38									
LITERACY	rich(1)	38	83.034	88.974	94.914	14.733	18.071													
LITERACY	Diff (1-2)	-51.81	-42.41	-33.01	20.325	23.135														
Statistics										Levene's Test for Equality of Variances										
LITERACY										F									p	
LITERACY		Equal variances assumed				Equal variances not assumed					14.266									.000
LITERACY		Maximum				Maximum				t test for Equality of Means										
		Upper CL	Std Dev	Std Err	5	95				t		M	SE							
Variable	GNPSPLIT											Difference								
LITERACY	poor(0)	31.062	3.2059																	
LITERACY	Rich(1)	23.38	2.9315																	
LITERACY	Diff (1-2)	26.854	4.7379																	
T-Tests																				
Variable	Method	Variances	DF	t Value	Pr >  t															
LITERACY	Pooled	Equal	100	-8.95	<.0001															
LITERACY	Satterthwaite	Unequal	97	-9.76	<.0001															
LITERACY		Equality of Variances				Equality of Variances														
Variable	Method	Num DF	Den DF	F Value	Pr > F															
LITERACY	Folded F	63	37	2.01	0.0236															

Note: CL (SAS output) or CI (SPSS output) = confidence interval. GNPSPLIT = Countries are divided into rich (1) and poor (0) based on their gross national product (GNP) per capita in 1980. SEM = standard error of the mean.

are the same (that is,  $k - 1 = 2 - 1 = 1$ ,  $N - k = 102 - 2 = 100$ ) as the degrees of freedom associated with “equal” variance test discussed earlier, and therefore SPSS does not report it separately.

If the assumption of equal variance is violated and the two groups have different variances as is the case in this example, where the folded  $F$  test or Levene’s  $F$  weighted statistic is significant, indicating that the two groups have significantly different variances, the value for degrees of freedom (100) is no longer accurate. Therefore, we need to estimate the correct degrees of freedom (SAS Institute, 1985; also see Satterthwaite, 1946, for the computations involved in this estimation).

We can estimate the degrees of freedom according to Satterthwaite’s (1946) method by using the following formula:

$$df \text{ Satterthwaite} =$$

$$\frac{(n_1 - 1)(n_2 - 1)}{(n_1 - 1) \left[ 1 - \frac{S_1^2 n_2}{(S_1^2 n_2 + S_2^2 n_1)} \right]^2 + (n_2 - 1) \left[ \frac{S_1^2 n_2}{(S_1^2 n_2 + S_2^2 n_1)} \right]^2}$$

where  $n_1$  = sample size of group 1,  $n_2$  = sample size of group 2, and  $S_1$  and  $S_2$  are the standard deviations of groups 1 and 2, respectively. By inserting subgroup data from Table 1, we arrive at the more accurate degrees of freedom as follows:

$$\begin{aligned} & \frac{(64 - 1)(38 - 1)}{(64 - 1) \left[ 1 - \frac{25.65^2 \times 38}{(25.65^2 \times 38) + 18.07^2 \times 64} \right]^2 + (38 - 1) \left[ \frac{25.65^2 \times 38}{(25.65^2 \times 38) + 18.07^2 \times 64} \right]^2} \\ &= \frac{2331}{63 \left[ 1 - \frac{25000.96}{(25000.96) + 20897.59} \right]^2 + 37 \left[ \frac{25000.96}{(25000.96) + 20897.59} \right]^2} \\ &= \frac{2331}{63 \left[ 1 - \frac{25000.96}{45898.55} \right]^2 + 37 \left[ \frac{25000.96}{45898.55} \right]^2} \\ &= \frac{2331}{63[1 - .5447]^2 + 37[.5447]^2} \\ &= \frac{2331}{63 \times .207 + 37 \times .2967} = \frac{2331}{24.0379} = 96.97 \end{aligned}$$

Because the assumption of equality of variances is violated, in the previous analysis the Satterthwaite’s

value for degrees of freedom, 96.97 (SAS rounds it to 97), is accurate, and our earlier value, 100, is not. Fortunately, it is no longer necessary to hand calculate this as major statistical packages such as SAS and SPSS provide the correct value for degrees of freedom when the assumption of equal variance is violated and equal variances are not assumed. This is the fourth value for degrees of freedom in our example, which appears in Table 1 as 97 in SAS and 96.967 in SPSS. Again, this value is the correct number to report, as the assumption of equal variances is violated in our example.

### Comparing the Means of $g$ Groups with One Parameter (Analysis of Variance)

What if we have more than two groups to compare? Let us assume that we have  $n_1 + \dots + n_g$  groups of observations or countries grouped by political freedom (FREEDOMX) and that we are interested in differences in their literacy rates (LITERACY, the dependent variable). We can test the variability of  $g$  means by using the analysis of variance (ANOVA). The ANOVA procedure produces three different types of degrees of freedom, calculated as follows:

- The first type of degrees of freedom is called the *between-groups degrees of freedom* or *model degrees of freedom* and can be determined by using the number of group means we want to compare. The ANOVA procedure tests the assumption that the  $g$  groups have equal means and that the population mean is not statistically different from the individual group means. This assumption reflects the null hypothesis, which is that there is no statistically significant difference between literacy rates in  $g$  groups of countries ( $\mu_1 = \mu_2 = \mu_3$ ). The alternative hypothesis is that the  $g$  sample means are significantly different from one another. There are  $g - 1$  model degrees of freedom for testing the null hypothesis and for assessing variability among the  $g$  means. This value of model degrees of freedom is used in the numerator for calculating the  $F$  ratio in ANOVA.

- The second type of degrees of freedom, called the *within-groups degrees of freedom* or *error degrees of freedom*, is derived from subtracting the model degrees of freedom from the corrected total degrees of freedom. The within-groups

degrees of freedom equal the total number of observations minus the number of groups to be compared,  $n_1 + \dots + n_g - g$ . This value also accounts for the denominator degrees of freedom for calculating the  $F$  statistic in an ANOVA.

- Calculating the third type of degrees of freedom is straightforward. We know that the sum of deviation from the mean or  $\Sigma(Y_i - \bar{Y}) = 0$ . We also know that the total sum of squares or  $\Sigma(Y_i - \bar{Y})^2$  is nothing but the sum of  $N^2$  deviations from the mean. Therefore, to estimate the total sum of squares  $\Sigma(Y_i - \bar{Y})^2$ , we need only the sum of  $N - 1$  deviations from the mean. Therefore, with the total sample size we can obtain the total degrees of freedom, or corrected total degrees of freedom, by using the formula  $N - 1$ .

In Table 2, we show the SAS and SPSS output with these three different values of degrees of freedom using the ANOVA procedure. The dependent variable, literacy rate, is continuous, and the independent variable, political freedom or FREEDOMX, is nominal. Countries are classified into three groups on the basis of the amount of political freedom each country enjoys: Countries that enjoy high political freedom are coded as 1 ( $n = 32$ ), countries that enjoy moderate political freedom are coded as 2 ( $n = 34$ ), and countries that enjoy no political freedom are coded as 3 ( $n = 36$ ). The mean literacy rates (dependent variable) of these groups of countries are examined. The null hypothesis tests the assumption that there is no significant difference in the literacy rates of these countries according to their level of political freedom.

The first of the three degrees of freedom, the between-groups degrees of freedom, equals  $g - 1$ . Because there are three groups of countries in this analysis, we have  $3 - 1 = 2$  degrees of freedom. This accounts for the numerator degrees of freedom in estimating the  $F$  statistic. Second, the within-groups degrees of freedom, which accounts for the denominator degrees of freedom for calculating the  $F$  statistic in ANOVA, equals  $n_1 + \dots + n_g - g$ . These degrees of freedom are calculated as  $32 + 34 + 36 - 3 = 99$ . Finally, the third degrees of freedom, the total degrees of freedom, are calculated as  $N - 1$  ( $102 - 1 = 101$ ). When reporting  $F$  values and their respective degrees of freedom, researchers should report them as follows: The independent and the

dependent variables are significantly related [ $F(2, 99) = 16.64, p < .0001$ ].

### Degrees of Freedom in Multiple Regression Analysis

We skip to multiple regression because degrees of freedom are the same in ANOVA and in simple regression. In multiple regression analysis, there is more than one independent variable and one dependent variable. Here, a parameter stands for the relationship between a dependent variable ( $Y$ ) and each independent variable ( $X$ ). One must understand four different types of degrees of freedom in multiple regression.

- The first type is the *model (regression) degrees of freedom*. Model degrees of freedom are associated with the number of independent variables in the model and can be understood as follows:

A null model or a model without independent variables will have zero parameters to be estimated. Therefore, predicted  $Y$  is equal to the mean of  $Y$  and the degrees of freedom equal 0.

A model with one independent variable has one predictor or one piece of useful information ( $k = 1$ ) for estimation of variability in  $Y$ . This model must also estimate the point where the regression line originates or an intercept. Hence, in a model with one predictor, there are  $(k + 1)$  parameters— $k$  regression coefficients plus an intercept—to be estimated, with  $k$  signifying the number of predictors. Therefore, there are  $[(k + 1) - 1]$ , or  $k$  degrees of freedom for testing this regression model.

Accordingly, a multiple regression model with more than one independent variable has some more useful information in estimating the variability in the dependent variable, and the model degrees of freedom increase as the number of independent variables increase. The null hypothesis is that all of the predictors have the same regression coefficient of zero, thus there is only one common coefficient to be estimated (Dallal, 2003). The alternative hypothesis is that the regression coefficients are not zero and that each variable explains a different amount of variance in the dependent variable. Thus, the researcher must estimate  $k$  coefficients plus the intercept. Therefore,

**Table 2: Analysis of Variance (ANOVA) of Literacy Rates According to Countries' Levels of Freedom with Reported Degrees of Freedom (N = 102)**

SAS SYSTEM							SPSS SYSTEM						
The ANOVA Procedure				LITERACY									
Dependent Variable: LITERACY		Percent of adult population literate b		Sum of Squares		F Value	Pr > F	Sum of Squares		F			
Source	DF	Mean Square						df	M <sup>2</sup>				
Model	2	24253.48284	12126.74142	16.64	<.0001			Between Groups	24253.483	2			
Error	99	72156.09559	728.84945					Within Groups	72156.096	99			
Corrected Total	101	96409.57843						Total	96409.578	101			
R-Square													
	0.251567	43.29061	26.99721					Post Hoc Tests					
Source													
	DF	Anova SS	Mean Square	LITERACY	Mean			Homogeneous Subsets					
FREEDOMX													
	2	24253.48284	12126.74142	16.64	<.0001			LITERACY					
The ANOVA Procedure													
Duncan's Multiple Range Test <sup>1</sup> for LITERACY													
Alpha													
								3 (=not free)	36	47.417			
Error Degrees of Freedom								2 (=partly free)	34	57.529			
								1 (=free)	32	84.313			
Error Mean Square													
								p		.126			
Harmonic Mean of Cell Sizes													
								Subset ( $\alpha = .05$ )					
Number of Means <sup>2</sup>													
								N	1	2			
Critical Range													
								Subset ( $\alpha = .05$ )					
Duncan Grouping													
		Mean <sup>3</sup>	N					Subset ( $\alpha = .05$ )					
A		84.313	32					Subset ( $\alpha = .05$ )					
								Subset ( $\alpha = .05$ )					
B		57.529	34					Subset ( $\alpha = .05$ )					
								Subset ( $\alpha = .05$ )					
B		47.417	36					Subset ( $\alpha = .05$ )					
								Subset ( $\alpha = .05$ )					

Means for groups in homogeneous subsets are displayed.

<sup>a</sup>Uses Harmonic Mean Sample Size = 33.921.

<sup>b</sup>The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Note: <sup>1</sup>Duncan's multiple range test for literacy controls the Type I comparisonwise error rate, not the experimentwise error rate. <sup>2</sup>Cell sizes are not equal. <sup>3</sup>Means with the same letter are not significantly different. <sup>4</sup>Coef Var = coefficient of variation. <sup>5</sup>S = Sum of Squares.

there are  $(k + 1) - 1$  or  $k$  degrees of freedom for testing the null hypothesis (Dallal, 2003). In other words, the model degrees of freedom equal the number of useful pieces of information available for estimation of variability in the dependent variable.

- The second type is the *residual*, or *error, degrees of freedom*. Residual degrees of freedom in multiple regression involve information of both sample size and predictor variables. In addition, we also need to account for the intercept. For example, if our sample size equals  $N$ , we need to estimate  $k + 1$  parameters, or one regression coefficient for each of the predictor variables ( $k$ ) plus one for the intercept. The residual degrees of freedom are calculated  $N - (k + 1)$ . This is the same as the formula for the error, or within-groups, degrees of freedom in the ANOVA. It is important to note that increasing the number of predictor variables has implications for the residual degrees of freedom. Each additional parameter to be estimated costs one residual degree of freedom (Dallal, 2003). The remaining residual degrees of freedom are used to estimate variability in the dependent variable.
- The third type of degrees of freedom is the *total*, or *corrected total, degrees of freedom*. As in ANOVA, this is calculated  $N - 1$ .
- Finally, the fourth type of degrees of freedom that SAS (and not SPSS) reports under the parameter estimate in multiple regression is worth mentioning. Here, the null hypothesis is that there is no relationship between each independent variable and the dependent variable. The degree of freedom is always 1 for each relationship and therefore, some statistical software, such as SPSS, do not bother to report it.

In the example of multiple regression analysis (see Table 3), there are four different values of degrees of freedom. The first is the regression degrees of freedom. This is estimated as  $(k + 1) - 1$  or  $(6 + 1) - 1 = 6$ , where  $k$  is the number of independent variables in the model. Second, the residual degrees of freedom are estimated as  $N - (k + 1)$ . Its value here is  $99 - (6 + 1) = 92$ . Third, the total degrees of freedom are calculated  $N - 1$  (or  $99 - 1 = 98$ ). Finally, the degrees of freedom shown under parameter estimates for each parameter always equal

1, as explained above.  $F$  values and the respective degrees of freedom from the current regression output should be reported as follows: The regression model is statistically significant with  $F(6, 92) = 44.86, p < .0001$ .

## Degrees of Freedom in a Nonparametric Test

Pearson's chi square, or simply the chi-square statistic, is an example of a nonparametric test that is widely used to examine the association between two nominal level variables. According to Weiss (1968) "the number of degrees of freedom to be associated with a chi-square statistic is equal to the number of independent components that entered into its calculation" (p. 262). He further explained that each cell in a chi-square statistic represents a single component and that an independent component is one where neither observed nor expected values are determined by the frequencies in other cells. In other words, in a contingency table, one row and one column are fixed and the remaining cells are independent and are free to vary. Therefore, the chi-square distribution has  $(r - 1) \times (c - 1)$  degrees of freedom, where  $r$  is the number of rows and  $c$  is the number of columns in the analysis (Cohen, 1988; Walker, 1940; Weiss, 1968). We subtract one from both the number of rows and columns simply because by knowing the values in other cells we can tell the values in the last cells for both rows and columns; therefore, these last cells are not independent.

As an example, we ran a chi-square test to examine whether gross national product (GNP) per capita of a country (GNPSPLIT) is related to its level of political freedom (FREEDOMX). Countries (GNPSPLIT) are divided into two categories—rich countries or countries with high GNP per capita (coded as 1) and poor countries or countries with low GNP per capita (coded as 0), and political freedom (FREEDOMX) has three levels—free (coded as 1), partly free (coded as 2), not free (coded as 3) (see Table 4). In this analysis, the degrees of freedom are  $(2 - 1) \times (3 - 1) = 2$ . In other words, by knowing the number of rich countries, we would automatically know the number of poor countries. But by knowing the number of countries that are free, we would not know the number of countries that are partly free and not free. Here, we need to know two of the three components—for instance, the number of countries that are free and partly free—so that we will know the number of countries

**Table 3: Multiple Regression of Literacy Rates on Educational and Social Variables with Reported Degrees of Freedom (N = 99)**

SAS SYSTEM		SPSS SYSTEM			
The REG Procedure		Model Summary			
<b>Model: MODEL1</b>					
Dependent Variable: LITERACY perc of adult population literate b					
Number of Observations Read	192				
Number of Observations Used	99				
Number of Observations with Missing Values	93				
Analysis of Variance					
Source	Sum of Squares	Mean Square	F Value		
Model	70570	11762	44.86		
Error	24120	262.17651	<.0001		
Corrected Total	98	94690			
Root MSE	16.19187	R-Square	0.7453		
Dependent Mean	62.02020	Adj R-Sq	0.7287		
Coeff Var	26.10741				
Parameter Estimates					
Variable	Label	Parameter Estimate	Standard Error		
Intercept	Intercept	29.72514	9.27642		
free	Free Countries	1	3.20		
partfree	Partly Free Countries	1	0.00005		
EDFUND84	education expenditures 1984 b	1	1.08		
PUPILS80	PUPILS PER TEACHER 1980 ISD	1	-1.62820		
SCHOOL80	PRIMARY SCHOOL STUDENTS 1980 ISD	1	0.00009		
GNP80	GNP PER CAPITA 1980 ISD	1	-0.69716		
			t		
		(Constant)	29.725		
		free	5.396		
		partfree	-1.628		
		EDFUND84	4.83E-005		
		PUPILS80	-.697		
		SCHOOL80	.583		
		GNP80	.001		
			β		
		B	SE		
			t		
			p		

\*Predictors: (Constant), GNP80, partfree, EDFUND84, SCHOOL80, PUPILS80, free

Analysis of Variance: Dependent Variable = Literacy

\*Predictors: (Constant), GNP80, partfree, EDFUND84, SCHOOL80, PUPILS80, free

Coefficients: Dependent Variable = Literacy

Model		Unstandardized Coefficients		Standardized Coefficients		t	p
		B	SE	β	Standardized Coefficients		
1	(Constant)	29.725	9.276	3.204	.002		
	free	5.396	5.000	.082	1.079		
	partfree	-1.628	4.128	-.025	-.394		
	EDFUND84	4.83E-005	.000	.034	.584		
	PUPILS80	-.697	.154	-.310	-4.522		
	SCHOOL80	.583	.068	.524	8.615		
	GNP80	.001	.001	.158	1.917		

Note: Perc = percentage. Coeff Var = coefficient of variation. Adj = adjusted. GNP = gross national product.

**Table 4: Chi Square of GNPSPLIT and Level of Freedom in Countries with Reported Degrees of Freedom (N = 124)**

SAS SYSTEM		SPSS SYSTEM	
Case Processing Summary		Cases	
		Valid	Missing
GNPSPLIT	N	%	N
GNPSPLIT * FREEDOMX	124	64.6	68
Total	124	100.0	192
GNPSPLIT			
Frequency			
Percent			
Row Pct			
Col Pct	1 ,	2 ,	3 ,
Total			
FREEDOMX(FREEDOM INDEX ISD)			
Table of GNPSPLIT by FREEDOMX			
GNPSPLIT			
Frequency			
Percent			
Row Pct			
Col Pct	1 ,	2 ,	3 ,
Total			
ffffffffff'ffffffffffff'ffffffffffff'			
0 ,	10 ,	31 ,	35 ,
, 8.06 ,	25.00 ,	28.23 ,	61.29
, 13.16 ,	40.79 ,	46.05 ,	
, 28.57 ,	75.61 ,	72.92 ,	
ffffffffff'ffffffffffff'ffffffff'			
1 ,	25 ,	10 ,	13 ,
, 20.16 ,	8.06 ,	10.48 ,	38.71
, 52.08 ,	20.83 ,	27.08 ,	
, 71.43 ,	24.39 ,	27.08 ,	
ffffffffff'ffffffffffff'ffffffffffff'			
Total	35	41	48
, 28.23	33.06	38.71	100.00
Frequency Missing = 68			
Statistics for Table of GNPSPLIT by FREEDOMX			
Statistic	DF	Value	Prob
Chi-Square	2	22.0708	<.0001
Likelihood Ratio Chi-Square	2	22.0179	<.0001
Mantel-Haenszel Chi-Square	1	14.8569	0.0001
Phi Coefficient		0.4219	
Contingency Coefficient		0.3887	
Cramer's V		0.4219	
Effective Sample Size = 124			
Frequency Missing = 68			

\*Zero cells (0.0%) have expected count less than 5. The minimum expected count is 13.55.

	Value	df	Asymptotic, <i>p</i> (2-tailed)
Pearson $\chi^2$	22.071*	2	.000
Likelihood Ratio	22.018	2	.000
Linear-by-Linear Association	14.857	1	.000
Valid Cases ( <i>N</i> )	124		

that are not free. Therefore, in this analysis there are two independent components that are free to vary, and thus the degrees of freedom are 2.

Readers may note that there are three values under degrees of freedom in Table 4. The first two values are calculated the same way as discussed earlier and have the same values and are reported most widely. These are the values associated with the Pearson chi-square and likelihood ratio chi-square tests. The final test is rarely used. We explain this briefly. The degree of freedom for the Mantel-Haenszel chi-square statistic is calculated to test the hypothesis that the relationship between two variables (row and column variables) is linear; it is calculated as  $(N - 1) \times r^2$ , where  $r^2$  is the Pearson product-moment correlation between the row variable and the column variable (SAS Institute, 1990). This degree of freedom is always 1 and is useful only when both row and column variables are ordinal.

## CONCLUSION

Yu (1997) noted that "degree of freedom is an intimate stranger to statistics students" (p. 1). This research note has attempted to decrease the strangeness of this relationship with an introduction to the logic of the use of degrees of freedom to correctly interpret statistical results. More advanced researchers, however, will note that the information provided in this article is limited and fairly elementary. As degrees of freedom vary by statistical test (Salkind, 2004), space prohibits a more comprehensive demonstration. Anyone with a desire to learn more about degrees of freedom in statistical calculations is encouraged to consult more detailed resources, such as Good (1973), Walker (1940), and Yu (1997).

Finally, for illustrative purposes we used World Data that reports information at country level. In our analysis, we have treated each country as an independent unit of analysis. Also, in the analysis, each country is given the same weight irrespective of its population size or area. We have ignored limitations that are inherent in the use of such data. We warn readers to ignore the statistical findings of our analysis and take away only the discussion that pertains to degrees of freedom. **SWR**

## REFERENCES

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Dallal, G. E. (2003). *Degrees of freedom*. Retrieved May 30, 2006, from <http://www.tufts.edu/~gdallal/dof.htm>
- Good, I. J. (1973). What are degrees of freedom? *American Statistician*, 27, 227-228.
- Healey, J. F. (1990). *Statistics: A tool for social research* (2nd ed.). Belmont, CA: Wadsworth.
- HyperStat Online. (n.d.). *Degrees of freedom*. Retrieved May 30, 2006, from <http://davidmlane.com/hyperstat/A42408.html>
- Jaccard, J., & Becker, M. A. (1990). *Statistics for the behavioral sciences* (2nd ed.). Belmont, CA: Wadsworth.
- Norman, G. R., & Streiner, D. L. (2003). *PDQ statistics* (3rd ed.). Hamilton, Ontario, Canada: BC Decker.
- Pagano, R. R. (2004). *Understanding statistics in the behavioral sciences* (7th ed.). Belmont, CA: Wadsworth.
- Rosenthal, J. A. (2001). *Statistics and data interpretation for the helping professions*. Belmont, CA: Wadsworth.
- Runyon, R. P., & Haber, A. (1991). *Fundamentals of behavioral statistics* (7th ed.). New York: McGraw-Hill.
- Salkind, N. J. (2004). *Statistics for people who (think they) hate statistics* (2nd ed.). Thousand Oaks, CA: Sage Publications.
- SAS Institute. (1985). *SAS user's guide: Statistics, version 5.1*. Cary, NC: SAS Institute.
- SAS Institute. (1990). *SAS procedure guide, version 6* (3rd ed.). Cary, NC: SAS Institute.
- Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2, 110-114.
- Toothaker, L. E., & Miller, L. (1996). *Introductory statistics for the behavioral sciences* (2nd ed.). Pacific Grove, CA: Brooks/Cole.
- Trochim, W.M.K. (2005). *Research methods: The concise knowledge base*. Cincinnati: Atomic Dog.
- Walker, H. W. (1940). Degrees of freedom. *Journal of Educational Psychology*, 31, 253-269.
- Warner, R. M. (2008). *Applied statistics*. Thousand Oaks, CA: Sage Publications.
- Weinbach, R. W., & Grinnell, R. M., Jr. (2004). *Statistics for social workers* (6th ed.). Boston: Pearson.
- Weiss, R. S. (1968). *Statistics in social research: An introduction*. New York: John Wiley & Sons.
- Wonnacott, T. H., & Wonnacott, R. J. (1990). *Introductory statistics* (5th ed.). New York: John Wiley & Sons.
- Yu, C. H. (1997). *Illustrating degrees of freedom in terms of sample size and dimensionality*. Retrieved November 1, 2007, from <http://www.creative-wisdom.com/computer/sas/df.html>

**Shanta Pandey, PhD**, is associate professor, and **Charlotte Lyn Bright, MSW**, is a doctoral student, George Warren Brown School of Social Work, Washington University, St. Louis. The authors are thankful to a student whose curious mind inspired them to work on this research note. Correspondence concerning this article should be sent to Shanta Pandey, George Warren Brown School of Social Work, Washington University, St. Louis, MO 63130; e-mail: [pandey@wustl.edu](mailto:pandey@wustl.edu).

Original manuscript received August 29, 2006

Final revision received November 15, 2007

Accepted January 16, 2008

*Copyright of Social Work Research is the property of National Association of Social Workers and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.*



# Variable Importance Assessment in Regression: Linear Regression versus Random Forest

Ulrike GRÖMPING

Relative importance of regressor variables is an old topic that still awaits a satisfactory solution. When interest is in attributing importance in linear regression, averaging over orderings methods for decomposing  $R^2$  are among the state-of-the-art methods, although the mechanism behind their behavior is not (yet) completely understood. Random forests—a machine-learning tool for classification and regression proposed a few years ago—have an inherent procedure of producing variable importances. This article compares the two approaches (linear model on the one hand and two versions of random forests on the other hand) and finds both striking similarities and differences, some of which can be explained whereas others remain a challenge. The investigation improves understanding of the nature of variable importance in random forests. This article has supplementary material online.

KEY WORDS: Linear model; Random forest; Variable importance.

## 1. INTRODUCTION

Variable importance in regression is an important topic in applied statistics that keeps coming up in spite of critics who basically claim that the question should not have been asked in the first place (cf., e.g., Ehrenberg 1990; Christensen 1992; Stufken 1992). Grömping (2007) recently provided an overview and a detailed discussion of the properties of two methods that derive variable importance in linear regression based on variance decomposition. Chevan and Sutherland (1991) proposed “Hierarchical Partitioning” for more general univariate regression situations; Theil and Chung (1988) discussed information-based approaches including multivariate linear regression.

Recently, random forests have received a lot of attention in biostatistics and other fields. They are popular, because they can handle large numbers of variables with relatively small numbers of observations and in addition provide an assessment of variable importance (cf., e.g., Breiman 2001; Ishwaran 2007; Strobl et al. 2007). There are several recent articles on variable importance in random forests: van der Laan (2006) introduced a general concept based on causal effects. Ishwaran (2007) attempted to get a theoretical handle on MSE reduction in Random Forests

based on Breiman et al.’s (1984) Classification And Regression Trees (called RF-CART in the sequel) by investigating the behavior of closed-form expressions of a modified version; of course, a modification for tractability comes with the risk of sacrificing generalizability. For uncorrelated regressors, Strobl et al. (2007) demonstrated that variable importance metrics in RF-CART are biased under relevant circumstances and introduced a different type of forest which does not exhibit this bias in the uncorrelated regressor situations they simulated. For correlated regressors, Strobl et al. (2008) found that the proposed solution of Strobl et al. (2007) does not solve all issues. Thus, they proposed “conditional variable importance” as a modification to the algorithm for determining variable importance in random forests.

Random forests can be used for classification and regression, as was already suggested by Breiman 2001; random survival forests have also been proposed (cf., e.g., Hothorn et al. 2004; Ishwaran et al. 2008). This article investigates the regression situation and compares the newly proposed variable importance measures from two specific types of random forests to the more classical tools for linear regression models. The focus is on inter-regressor correlation as an important determinant of the behavior of variable importance metrics. Here, the random forest variable importance approach can benefit from the somewhat more advanced understanding of what happens in linear models.

Linear regression is a classical parametric method which requires explicit modeling of nonlinearities and interactions, if necessary. It is known to be reasonably robust, if the number of observations  $n$  is distinctly larger than the number of variables  $p$  ( $n \gg p$ ). With more variables than observations ( $p > n$  or even  $p \gg n$ ), linear regression breaks down, unless shrinkage methods are used like ridge regression (Hoerl and Kennard 1970), the lasso (Tibshirani 1996), or the elastic net as a combination of both (Zou and Hastie 2005). Random forests, on the other hand, are nonparametric and allow nonlinearities and interactions to be learned from the data without any need to explicitly model them. Also, they have been reported to work well not only for the  $n \gg p$  setting but also for data mining in the  $p \gg n$  setting. Reasons for usage of variable importances also differ in the two scenarios (cf. Section 7 for a detailed discussion). This article concentrates on the  $n \gg p$  situation. Nevertheless, the findings will also have implications for  $p \gg n$  variable selection applications and will shed some light into the black box of random forests.

The next section briefly introduces the example dataset which will be used in Sections 4 and 5 for illustration and method comparison. Section 3 presents the linear model along

Ulrike Grömping is Professor, Department II—Mathematics, Physics, Chemistry, BHT Berlin—University of Applied Sciences, Luxemburger Str. 10, D-13353 Berlin, Germany (E-mail: [groemping@bht-berlin.de](mailto:groemping@bht-berlin.de)).

with its relative importance metrics, whereas Section 4 introduces two variants of random regression forests and their associated variable importance metric. Section 5 uses the example data to compare the different methods. Section 6 presents a simulation study under systematically varied correlation structures and coefficient vectors like those in the article by Grömping (2007) that compares (i) variance decomposition based on averaging over orderings (Lindeman, Merenda, and Gold 1980; Kruskal 1987a, 1987b; Feldman 2005) and (ii) random forest variable importance metrics for two types of forests. Interpretation of results is followed by pointing out areas of interest for further research. The final Section 7 discusses in detail the purpose-specific conceptual needs for variable importance metrics in both linear model and random forest.

## 2. SWISS FERTILITY EXAMPLE

The R software (R Development Core Team 2008) contains a small socio-demographic dataset (“swiss”) on Fertility in 47 Swiss provinces in 1888 that is suitable for demonstrating the varying behaviors of different approaches. A larger dataset including the same variables for 182 provinces is available online (Switzerland Socio-economic variables 1870 to 1930, <http://opr.princeton.edu/archive/pefp/switz.asp>) and has been used here for better stability of results. The set of variables has been kept the same: Fertility rate in the married population (Fertility), percentage of male population in agriculture jobs (Agriculture), percentage of draftees with highest grade in an army exam (Examination), percentage of draftees with more than primary school education (Education), percentage of catholics (Catholic), and percentage of children who did not survive their first year (Infant.Mortality). It is not entirely clear how the smaller and the larger dataset are related (the maximum of some variables is slightly larger in the smaller dataset).

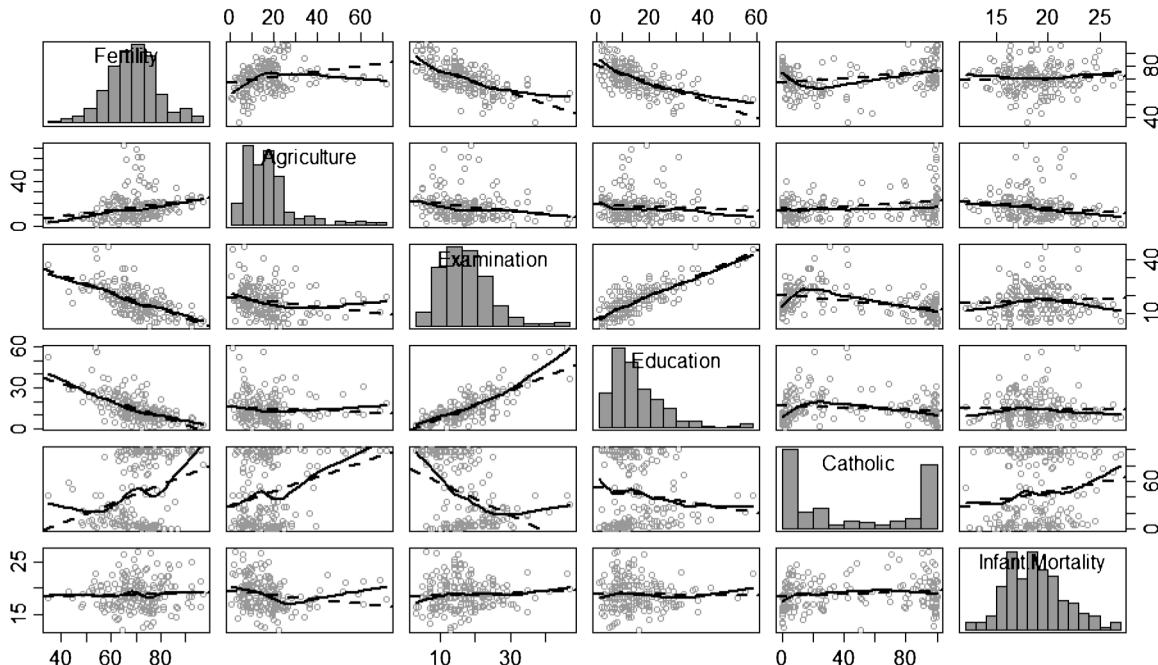


Figure 1. Scatterplot matrix of Swiss Fertility data with linear and loess lines.

Because the analyses are only intended as examples, this has not been further pursued. Figure 1 provides an overview of the bivariate relations for the data.

## 3. LINEAR REGRESSION AND IMPORTANCE METRICS

The linear regression model is considered in its usual form

$$Y = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_p + \varepsilon, \\ \beta_0, \beta_1, \dots, \beta_p \text{ fixed and unknown}, \quad (1)$$

where the random variables  $X_j$ ,  $j = 1, \dots, p$ , denote  $p$  regressor variables and the random variable  $\varepsilon$  denotes an error term, which is uncorrelated to the regressors and has expectation 0 and variance  $\sigma^2 > 0$ . The regressor variances are denoted as  $v_j$ ,  $j = 1, \dots, p$ , the inter-regressor correlations as  $\rho_{jk}$ , and the  $p \times p$  covariance matrix between regressors is assumed to be positive definite so that any sample regressor matrix with  $n > p$  rows is of full column rank with probability 1. Model (1) implies the conditional moments  $E(Y|X_1, \dots, X_p) = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_p$  and  $\text{var}(Y|X_1, \dots, X_p) = \text{var}(\varepsilon|X_1, \dots, X_p) = \sigma^2$  and the marginal variance model

$$\text{var}(Y) = \sum_{j=1}^p \beta_j^2 v_j + 2 \sum_{j=1}^{p-1} \sum_{k=j+1}^p \beta_j \beta_k \sqrt{v_j v_k} \rho_{jk} + \sigma^2. \quad (2)$$

Formula (2) depends on  $\beta_j$  and  $v_j$  through  $\beta_j \sqrt{v_j}$  only, the estimated version of which is equivalent to using the standardized coefficient, because division by the standard deviation of the response is not relevant when looking at one response only.

The first two summands of (2) constitute the part of the variance that is explained by the regressors, and  $R^2$  from a linear

model with  $n$  independent observations is consistent for the proportion of the first two summands of (2) in the total  $\text{var}(Y)$ . Variable importance methods that decompose  $R^2$  thus have to decompose the first two summands of (2). It is well known that this variance—or equivalently the model sum of squares or  $R^2$ —can be uniquely decomposed in case of uncorrelated regressors, but that different methods lead to different results for correlated regressors. In particular, the increase in  $R^2$  allocated to a certain regressor  $X_k$  depends on which regressors are already present in the model before adding  $X_k$ . For any particular order of regressors, a unique allocation can be determined by allocating each regressor the increase in  $R^2$  (or explained variance) when it is added to the model. Lindeman, Merenda, and Gold (1980, henceforth, LMG) proposed to average such order-dependent allocations over all  $p!$  orderings for a fair unique assessment. This approach (independently proposed also by Kruskal (1987a, 1987b)) will be termed LMG throughout this article. Feldman (2005) proposed a modified version with data-dependent weights that favor strong predictors (called PMVD for Proportional Marginal Variance Decomposition following Feldman). For mathematical detail on LMG and PMVD, see, for example, the article by Grömping 2007. Note that LMG has been proposed under various different names, for example, dominance analysis (Budescu 1993) or Shapley value regression (Lipovetsky and Conklin 2001). In this article, LMG and PMVD are compared to random forest variable importance assessments.

## 4. RANDOM FORESTS FOR REGRESSION

A forest is an ensemble of trees—like in real life. Breiman (2001) introduced the general concept of random forests and proposed one specific instance of this concept, which we will consider as RF-CART in the following. A further instance proposed by Strobl et al. (2007), RF-CI, will also be introduced below. The types of trees used in these two concepts will be presented in the following subsection, before introducing their integration into a forest approach in Section 4.2.

### 4.1 Regression Trees

A regression tree (cf. Figure 2 for an example) is built by recursively partitioning the sample (= the “root node”) into more

and more homogeneous groups, so-called nodes, down to the “terminal nodes.” Each split is based on the values of one variable and is selected according to a splitting criterion. Once a tree has been built, the response for any observation can be predicted by following the path from the root node down to the appropriate terminal node of the tree, based on the observed values for the splitting variables, and the predicted response value simply is the average response in that terminal node. For example, the left tree shown in Figure 2, modeling Fertility based on five candidate regressors, would predict the Fertility for a Swiss province with Education 20% and Agriculture 5% as 55.69% (go left on both splits). The regression function thus estimated from a tree is a multidimensional step function. This article considers binary trees only, that is, trees that split a parent node into two children at any step.

#### 4.1.1 CART Trees

The CART algorithm proposed by Breiman et al. (1984) chooses the split for each node such that maximum reduction in overall node impurity is achieved, where impurity is measured as the total sum of squared deviations from node centers. CART first grows a tree very large and subsequently “prunes” it, that is, cuts off branches that do not add to predictive performance according to a pruning criterion that can differ from the splitting criterion. The reason for pruning a large tree instead of growing a small tree only in the first place is an improvement in the predictive performance of the tree (stopping too early might miss out on later improvements). If only one tree is built, care is needed to make sure the tree does not overfit the data. For this purpose, the degree of pruning is typically decided based on cross-validation. If CART trees are used in random forests, they are typically grown quite large, and no pruning is done (cf. right tree in Figure 2); for details, see Section 4.2.

#### 4.1.2 Conditional Inference Trees

The splitting approach in CART trees has been known for a long time to be unfair in the presence of regressor variables of different types, categorical variables with different numbers of categories, or differing numbers of missing values (cf., e.g., Breiman 1984; Shih and Tsai 2004). To avoid this variable selection bias, Hothorn, Hornik, and Zeileis (2006b) proposed to use multiplicity-adjusted conditional tests (cf. Hothorn et al.

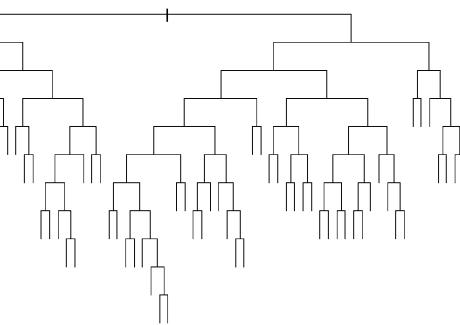
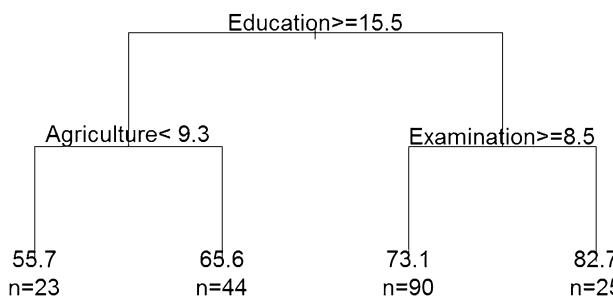


Figure 2. Individual tree and tree used in a forest (computed with R-package rpart; Therneau and Atkinson 1997). *Left:* A CART tree for Fertility for the Swiss data. Splits are labeled with the splitting criterion. Split condition true goes to the left, false to the right. *Right:* An unpruned CART tree for the same data that might be used in a forest (nodes of size 5 or less are terminal).

2006a) rather than maximum impurity reduction as the splitting criterion. The key idea of their approach is to untangle variable selection and split selection, similarly to the well-known CHAID approach (Kass 1980) for classification. Thus, for any node to be split, the procedure conducts a global permutation test of the null hypothesis of no association between any of the regressors and the response within the node. If this global null hypothesis is not rejected, the node is not split and becomes a terminal node. Otherwise, individual null hypotheses of no association to the response are tested for each variable, and the variable with the smallest  $p$ -value is selected for splitting. Subsequently, the splitting rule is determined based on the selected variable. With this method, pruning is not needed, because trees are not grown any further, once there is no further statistically significant split. For the Swiss Fertility data, the conditional tree is similar to the left CART tree in Figure 2. Analogously to CART, where pruning is usually omitted when growing forests, conditional inference trees for forests are also per default grown larger than those used in single tree analyses (cf. next section).

## 4.2 Random Forests

A random forest consists of a large number,  $n_{tree}$ , of trees, for example 1000. Theoretical results on random forests (Breiman 2001; Ishwaran 2007) are asymptotic in the number of trees, and a large number of trees has been reported to be particularly important when interest is in diagnostic quantities like variable importance (cf., e.g., Breiman 2002). A random forest is random in two ways: (i) each tree is based on a random subset of the observations, and (ii) each split within each tree is created based on a random subset of  $mtry$  candidate variables. Trees are quite unstable, so that this randomness creates differences in individual trees' predictions. The overall prediction of the forest is the average of predictions from the individual trees—because individual trees produce multidimensional step functions, their average is again a multidimensional step function that can nevertheless predict smooth functions because it aggregates a large number of different trees. For visualization of forest results, main effects and interaction plots similar to what has been proposed by Friedman (1991) for MARS can be used; cf. also Section 5 below.

### 4.2.1 RF-CART and RF-CI

RF-CART, that is, random forests based on CART trees, are most well known, because they have already been proposed in the fundamental article by Breiman (2001). Recently, Strobl et al. (2007) proposed to base random forests on the conditional inference trees discussed in Section 4.1.2. To highlight that the key difference lies in the different type of underlying trees (CART versus CI = conditional inference), this approach is called RF-CI in the following. Here, the two forest types have been applied as implemented in the R-packages randomForest (Liaw and Wiener 2002 based on Breiman 2001, 2002) and party (function cforest; Strobl et al. 2007). The default number of trees ( $n_{tree} = 500$ ) is identical for both forest variants. The number  $mtry$  of variables to be considered for each split is a tuning parameter, which has the default floor( $p/3$ ) for

RF-CART. With  $mtry = 1$ , the splitting variable would be determined completely at random, whereas  $mtry = p$  would eliminate one aspect of randomness for the forest, and it has been recommended to try half and twice the default as well (Liaw and Wiener 2002). For RF-CI,  $mtry$  has no meaningful default. Choice of  $mtry$  is further discussed in connection with simulation results (cf. Section 6.3).

RF-CART and RF-CI use different default sampling approaches: RF-CART uses a with-replacement sample of size  $n$ , RF-CI a without-replacement sample of size  $0.632 * n$ . According to Strobl et al. (2007), this difference is inconsequential in the setting investigated here with continuous regressors only. The defaults of RF-CART and RF-CI also result in very different tree sizes: RF-CART uses large unpruned trees that are grown with a lower limit for the size of nodes to be considered for splitting (nodes of size 5 or less are not split), for example, like the one on the right in Figure 2. RF-CI trees for forests are per default built without significance testing (tests without multiplicity adjustment are conducted at a default significance level of 100%), limiting tree size by restricting minimum split size for a node to 20 and minimum node size to 7. This stricter criterion, together with the sampling approach, implies that trees in RF-CI are per default substantially smaller than those in RF-CART (e.g., 8 to 11 terminal nodes for a typical tree within a RF-CI forest of the example data, in comparison to about 50 to 70 terminal nodes for RF-CART trees). Settings for both RF-CART and RF-CI could be adjusted such that the trees become larger or smaller by adjusting node splitting criteria, or—in case of RF-CI—by introducing a significance level smaller than 1. Increasing the minimum node size for RF-CART has been proposed by Segal, Barbour, and Grant (2004) for improving prediction accuracy, and selected simulation results with this modification will be discussed in Section 6.3.

### 4.2.2 Mean Squared Error

According to random sampling of observations, regardless whether with or without replacement, (an average of) 36.8% of the observations are not used for any individual tree—that is, are “out of the bag” = OOB for that tree. The accuracy of a random forest's prediction can be estimated from these OOB data as

$$\text{OOB-MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y}_{i \text{ OOB}})^2,$$

where  $\bar{y}_{i \text{ OOB}}$  denotes the average prediction for the  $i$ th observation from all trees for which this observation has been OOB. Analogously to linear regression, with the overall sum of squares  $SST = \sum_{i=1}^n (y_i - \bar{y})^2$  defined in the usual way,  $\text{OOB-}R^2$  can be obtained as  $1 - \text{OOB-MSE}/SST$ .

### 4.2.3 Variable Importance

A very well-known variable importance metric in CART trees and random forests is the so-called Gini importance for classification and its analogue, average impurity reduction, for regression forests. However, because of impurity's bias for selecting split variables, the resulting variable importance metrics are of course also biased (cf., e.g., Shih and Tsai 2004;

Strobl et al. 2007). Breiman (2002) suggested reduction in MSE when permuting a variable (Method 1), called “MSE reduction” in the following, as the method of choice. Although he discarded this metric for excessive variability in situations with many regressors in a later version of the Random Forests manual (Breiman 2003), permutation-based MSE reduction has been adopted as the state-of-the-art approach by various authors (Diaz-Uriarte and Alvarez de Andrés 2006; Ishwaran 2007; Genuer, Poggi, and Tuleau 2008; Strobl et al. 2008). Therefore, this permutation-based “MSE reduction” is also used as the random forest importance criterion in this article. It is determined as follows: For tree  $t$ , the OOB mean squared error is calculated as the average of the squared deviations of OOB-responses from their respective predictions:

$$\text{OOBMSE}_t = \frac{1}{n_{\text{OOB},t}} \sum_{\substack{i=1: \\ i \in \text{OOB}_t}}^n (y_i - \hat{y}_{i,t})^2,$$

where the  $\hat{\cdot}$  indicates predictions,  $\text{OOB}_t = \{i : \text{observation } i \text{ is OOB for tree } t\}$ , that is, summation is over OOB observations only, and  $n_{\text{OOB},t}$  is the number of OOB observations in tree  $t$ . If regressor  $X_j$  does not have predictive value for the response, it should not make a difference if the values for  $X_j$  are randomly permuted in the OOB data before the predictions are generated. Thus,

$$\begin{aligned} \text{OOBMSE}_t(X_j \text{ permuted}) \\ = \frac{1}{n_{\text{OOB},t}} \sum_{\substack{i=1: \\ i \in \text{OOB}_t}}^n (y_i - \hat{y}_{i,t}(X_j \text{ permuted}))^2 \end{aligned} \quad (3)$$

should not be substantially larger (or might by chance even be smaller) than  $\text{OOBMSE}_t$ . For each variable  $X_j$  in each tree  $t$ , the difference  $\text{OOBMSE}_t(X_j \text{ permuted}) - \text{OOBMSE}_t$  is calculated based on one permutation of the variable’s out-of-bag data for the tree (this difference is of course 0 for a variable that happens to be not involved in any split of tree  $t$ ). The MSE reduction according to regressor  $X_j$  for the complete forest is obtained as the average over all  $ntree$  trees of these differences. It is worthwhile to realize that the thus-calculated MSE reduction is NOT the same as the reduction in the forest’s MSE by having variable  $X_j$  available (versus not having  $X_j$  in the set of

explanatory variables). This was also pointed out by Ishwaran et al. (2008).

Whereas LMG and PMVD naturally decompose  $R^2$ , such natural decomposition does not occur for forests’ MSE reduction. Thus, for comparison purposes, all variable importance metrics in this article have been normalized to sum to 100%.

## 5. ALL METRICS APPLIED TO THE SWISS FERTILITY DATA

For the Swiss Fertility data, a linear model has been fitted using quadratic effects for Agriculture and Catholic and linear effects for the other three regressors, based on the impressions from Figure 1. The random forests have been fitted with default settings, apart from an increased  $ntree$  to ensure stability of the variable importance assessment (cf. Breiman 2002 or Genuer, Poggi, and Tuleau 2008 for recommended  $ntree$  values). In addition to the default  $mtry = 1$ ,  $mtry = 2$  has been run for comparison.

Table 1 shows normalized variable importance metrics from all approaches. Figure 3 shows effects plots for the linear model and for RF-CART. Within the linear model, LMG and PMVD allocations are almost identical, apart from the split within the only pair of strongly correlated regressors, Examination and Education (correlation: 0.79): Here, LMG gives Examination the benefit of the doubt, whereas PMVD allocates almost no contribution to Examination. PMVD’s bootstrap confidence interval for the share of Examination (not shown) includes the full LMG interval for the same share in this example (cf. also Grömping 2007 for a discussion of the variance properties of PMVD and LMG); that is, PMVD is extremely variable here. Importances from the two types of forest behave somewhat differently both from each other and from the linear model assessments. None of the forest metrics shows any similarity to PMVD regarding Examination’s share, whereas the RF-CI allocation for Education is almost as extreme as PMVD’s with  $mtry = 2$ . Table 1 also shows that the RF-CI assessment strongly depends on  $mtry$ , whereas the RF-CART assessment is more stable over  $mtry$ . This behavior will also be confirmed by the simulation study of Section 6. Results on  $mtry$  from the literature will be discussed in that context (Section 6.3).

Table 1. Relative importance of the five effects for “Fertility” normalized to sum 100%\* in linear regression and random forest models.

Response:	Random Forest MSE reduction, $ntree = 2000$					
	Linear model** ( $R^2 = 61.3\%$ )		$mtry = 1^{***}$		$mtry = 2^{***}$	
Fertility	PMVD	LMG	RF-CART	RF-CI	RF-CART	RF-CI
Agriculture	21.3	22.0	26.1	20.7	31.2	14.4
Examination	1.0	25.6	22.9	28.8	20.2	31.8
Education	56.3	31.5	28.5	35.9	30.1	44.9
Catholic	18.2	18.3	17.6	12.9	13.3	8.6
Infant.Mortality	3.3	2.7	4.9	1.7	5.2	0.3
Total	100.0	100.0	100.0	100.0	100.0	100.0

\*Normalization to sum 100% is not recommended for data analysis purposes, but is helpful for making metrics’ relative assessments comparable.

\*\* Agriculture and Catholic quadratic, the other variables linear; calculation with R-package relaimpo (cf. Grömping 2006).

\*\*\*  $mtry$  is the number of candidate variables randomly selected for each split in each tree.

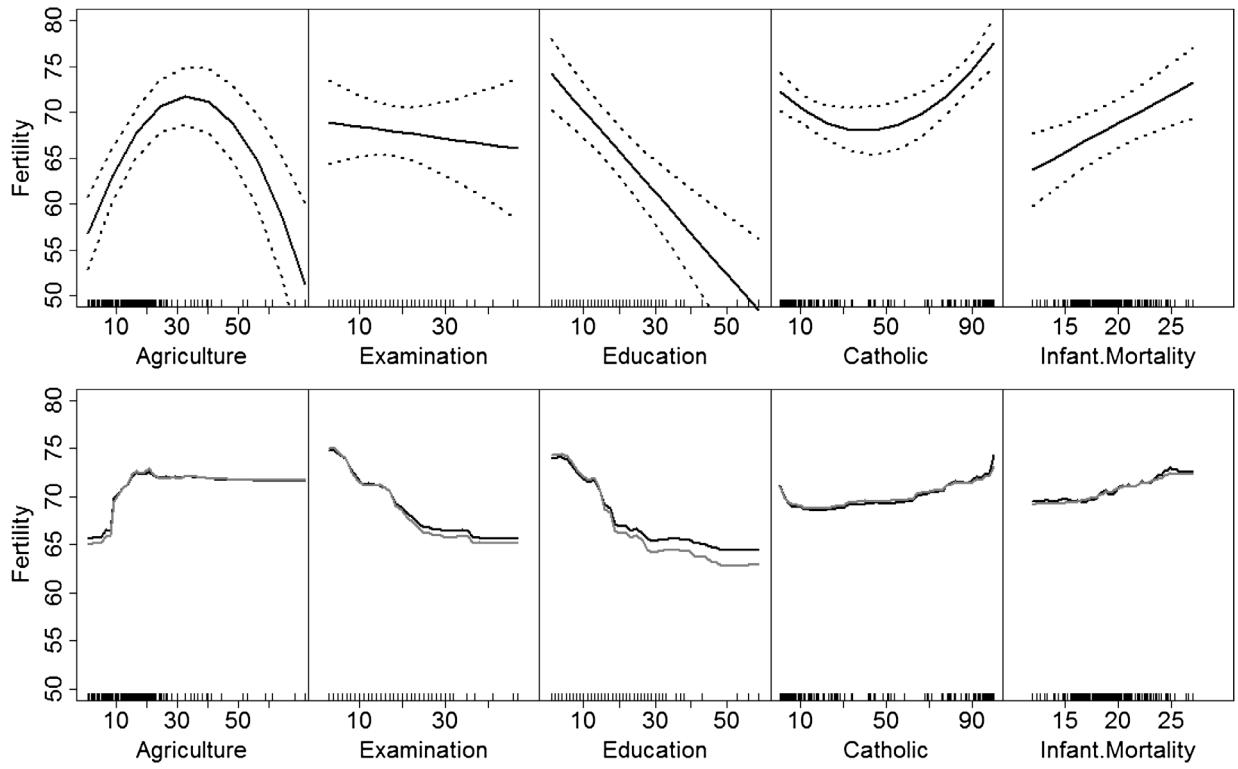


Figure 3. Main effects plot for the linear model (top, with 95% bands) and RF-CART ( $mtry = 1$  (black) and  $mtry = 2$  (gray)). Rugs at the bottom represent individual data values for the 182 Swiss provinces.

Of course, with this real data example, there is no guarantee that the data are adequately described by the chosen linear model with three linear and two quadratic effects and without interactions. Differences between allocations from the linear model and the two forest approaches in the example analysis may be due both to method differences and to deviations between the postulated linear model structure and the true model that the nonparametric forest approaches have attempted to estimate.

## 6. SIMULATION STUDY

To focus on the method-related differences only, this section reports a simulation study for truly linear models. Note that forests can only approximate linear models, because they always fit step functions. However, it has been proven that they closely approximate any smooth expectation model for large samples and large numbers of trees (Ishwaran 2007) over a finite closed rectangular regressor space. Of course, the linear model is on its home turf, whereas the forests are far less parsimonious for fitting a truly linear model. This is reflected by much larger dispersion of forest-based variable importance metrics, which is, however, not in the focus of this study. It can be conjectured that, under the default settings used here, RF-CART is more efficient in approximating a continuous model than RF-CI because of the larger number of nodes in each tree.

### 6.1 Simulation Scenarios

The simulation scenarios have been chosen as a subset from those in Grömping (2007); for these scenarios, LMG and PMVD have already been investigated, and their true values (in terms of limits when increasing the sample size to infinity) can be calculated. Here, 100 samples of size  $n = 1000$  each have been simulated in the following way: Four regressors  $X_1, \dots, X_4$  have been generated as  $n$  independent observation vectors from four-variate normal distributions with expectations 0, variances 1, and correlations  $\text{corr}(X_j, X_k) = \rho^{|j-k|}$  with  $\rho = -0.9$  to 0.9 in steps of 0.1,  $j, k = 1, \dots, 4$ . Note that negative values of  $\rho$  provide mixed-sign correlation structures among regressors, that is, positive and negative values of  $\rho$  yield structurally different correlation matrices. Responses have been created from the regressors using seven different vectors of true model coefficients:  $\beta_1 = (4, 1, 1, 0.3)^T$ ,  $\beta_2 = (1, 1, 1, 0.3)^T$ ,  $\beta_3 = (4, 1, 0, 0)^T$ ,  $\beta_4 = (1, 1, 1, 1)^T$ ,  $\beta_5 = (1.2, 1, 1, 0.3)^T$ ,  $\beta_6 = (1, 1, 1, 0)^T$ ,  $\beta_7 = (4, 3.5, 3, 2.5)^T$ , adding normal random error with  $\sigma^2$  such that  $R^2 = 50\%$ . The two forest methods have been implemented with 500 trees for each forest using their previously given default settings (cf. Section 4.2). The aforementioned instability of MSE reductions (Breiman 2002; Genuer, Poggi, and Tuleau 2008) is not an issue here, because variation is not the topic of investigation and averages over 100 simulation runs are reasonably stable. Simulation code for implementing similar studies—adjusted to run with the current version of R-package party for RF-CI—can be found online as supplemental material.

## 6.2 Descriptive Method Comparisons

This section describes simulation results in terms of average normalized variable importances from random forests and true normalized values for LMG and PMVD—against which their estimates are consistent. LMG and PMVD values are shown as reference curves to indicate that they are the theoretical limits as opposed to averages over simulation runs. This is not meant to indicate that they represent an overall gold standard. A key difference between LMG and PMVD can be observed in Figure 4: For LMG, the importance allocated to the regressor with the largest coefficient decreases substantially in favor of the importances allocated to the other regressors with increasing degree of correlation. This is called “equalizing behavior” in the following.

Average variable importance from the forests with  $mtry = 1$  (Figure 4) is found to be quite similar to LMG. However, the similarities are far from perfect, and average variable importance from RF-CI shows a slight tendency toward PMVD. With increasing  $mtry$  (cf. Figures 5 and 6), variable importance for RF-CI becomes more and more similar to PMVD, whereas variable importance for RF-CART is *relatively* stable over  $mtry$  and remains similar to LMG.

Whereas RF-CI shows a much stronger dependence on  $mtry$  than RF-CART, dependence of allocations on the correlation parameter depends on the situation: LMG and RF-CART show a much stronger dependence, for example, in Figure 5, whereas the dependence is stronger for PMVD and particularly for RF-CIs in Figure 6. The most striking feature for Figure 6

is the strong correlation-dependence of allocations to the first three  $X$ 's for RF-CI. This is much stronger than for PMVD and increases with increasing  $mtry$ , so that the regressor with the largest coefficient loses its first rank to the second regressor for a high positive correlation parameter and high  $mtry$ , whereas the regressor with the smallest coefficient overtakes the two medium regressors for a strongly negative correlation parameter (and thus a mixed-sign correlation pattern) with increasing  $mtry$ . This behavior is not yet understood.

Apart from the correlation pattern and the tuning parameter  $mtry$ , the sample size is also relevant for variable importance allocations: For simulations with smaller sample size ( $n = 100$ , not shown)—for which good approximation of a linear model by a random forest is not guaranteed—RF-CI allocations were far less similar to PMVD even in situations where agreement is almost perfect for  $n = 1000$  (e.g.,  $\beta_1$  with  $mtry = 4$ ). In case of RF-CART, a closer similarity to LMG has been observed for smaller samples, whereas larger samples tend to be more equalizing than LMG between weaker and stronger regressors (but not for regressors with no influence) for low degrees of correlation (cf., e.g., Figure 5).

## 6.3 Discussion of the Dependence on $mtry$

The tuning parameter  $mtry$  deserves special attention. Breiman (2001) recommended  $mtry = \sqrt{p}$  for classification forests and observed that larger numbers would be needed for situations for which many irrelevant inputs confuse the picture. He indicated that  $mtry \ll p$  should improve predictive performance

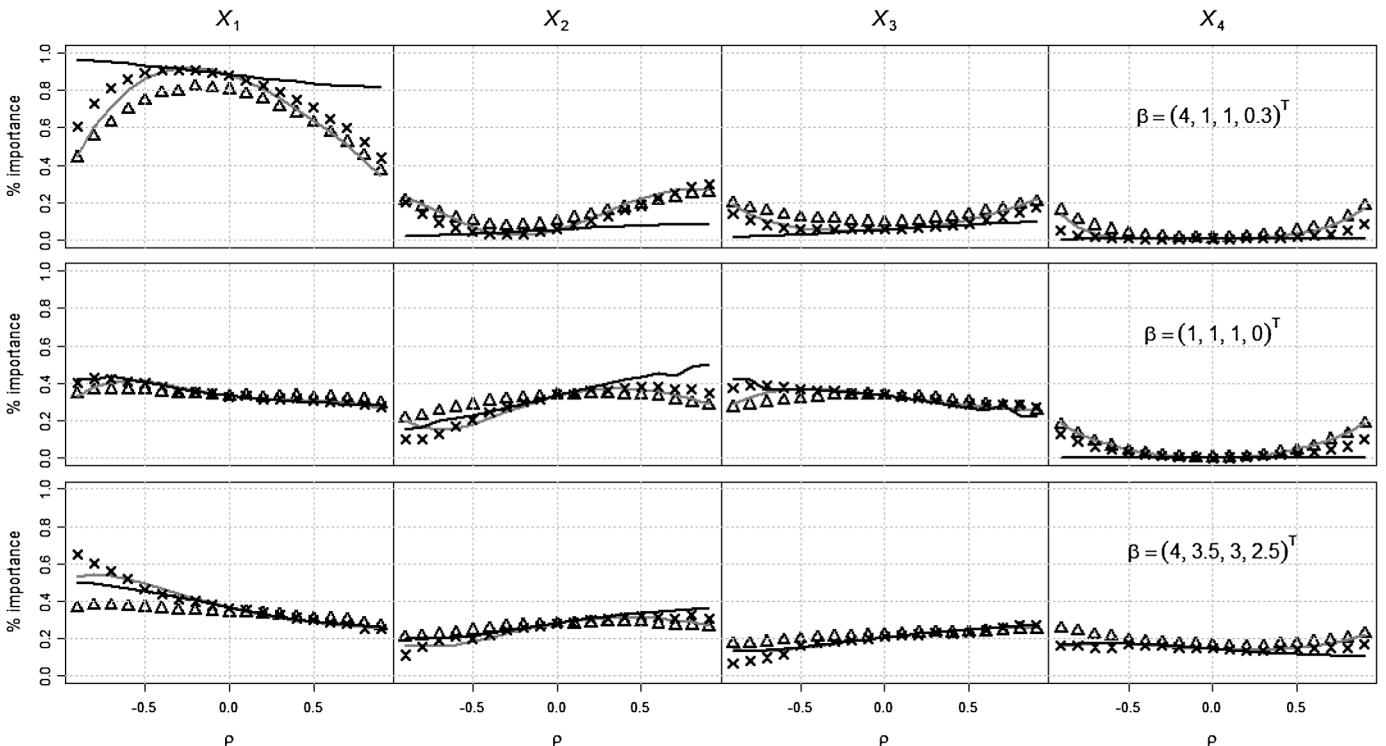


Figure 4. Average normalized importances for the four  $X$ 's from 100 simulated datasets based on  $mtry = 1$  variables per split,  $\beta_1 = (4, 1, 1, 0.3)^T$  (top),  $\beta_6 = (1, 1, 1, 0)^T$  (middle), and  $\beta_7 = (4, 3.5, 3, 2.5)^T$  (bottom),  $\text{corr}(X_j, X_k) = \rho^{|j-k|}$  with  $\rho = -0.9$  to  $0.9$  in steps of  $0.1$ . Gray line: true normalized LMG allocation; black line: true normalized PMVD allocation.  $\Delta$ : variable importance (% MSE reduction) from RF-CART,  $\times$ : variable importance (% MSE reduction) from RF-Cl.

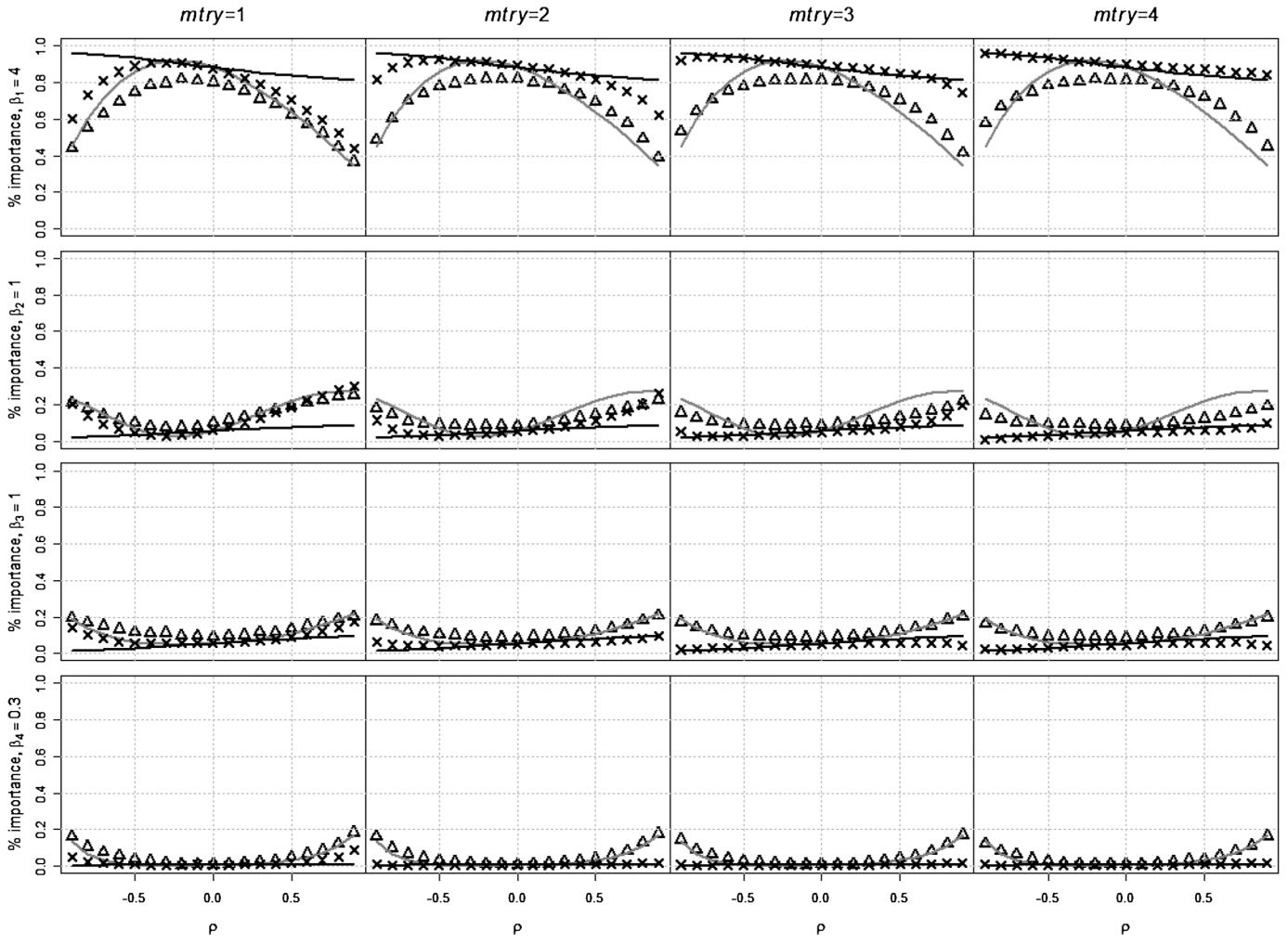


Figure 5. Average normalized importances for the four  $X$ 's (top to bottom:  $X_1$  to  $X_4$ ) from 100 simulated datasets for  $mtry = 1, 2, 3, 4$  (left to right) with  $\beta_1 = (4, 1, 1, 0.3)^T$ ,  $\text{corr}(X_j, X_k) = \rho^{|j-k|}$  with  $\rho = -0.9$  to  $0.9$  in steps of  $0.1$ . Gray line: true normalized LMG allocation; black line: true normalized PMVD allocation.  $\triangle$ : variable importance (% MSE reduction) from RF-CART,  $\times$ : variable importance (% MSE reduction) from RF-Cl.

versus using all variables, because lower correlation between individual trees improves prediction accuracy. He also observed this effect to be much weaker in regression forests, and consequently the proposed default of  $p/3$  for regression forests (Liaw and Wiener 2002) is much larger than  $\sqrt{p}$  for large numbers of variables. Recently Genauer, Poggi, and Tuleau (2008) presented simulation results that indicate better prediction accuracy with  $mtry = p$  for the artificial data from the article by Friedman (1991). Within our simulation study, the  $mtry$  value that minimized OOB-MSE depended on both the model and the correlation between regressors (but was generally  $mtry = 1$  or  $mtry = 2$ , and in most cases performance differences were small).

Diaz-Uriarte and Alvarez de Andrés (2006) investigated  $mtry$ 's impact on variable importance for classification forests and found the default values to work satisfactorily and independently of the settings of other parameters like  $ntree$ . The empirical investigations in this article have shown that the choice of  $mtry$  can substantially affect the allocated importance in random forests for regression. The reasons for this behavior are

most obvious when considering  $mtry = 1$ : By random choice, a regressor with no relation to the response, which would have never been selected for a split given any competition, will by chance sometimes become the basis of splitting. If the regressor is unrelated to both the response and the other regressors, it will only get a weak share allocated both in RF-CART and in RF-Cl's (e.g., Figure 4,  $X_4$  in the middle scenario for  $\rho = 0$ ). This is because even if such a regressor has by chance been made the basis of a split, the % MSE reduction will be on average zero. Now, consider a variable with no conditional influence of its own (coefficient 0 in the model) but a strong genuine correlation to one or more of the regressors that do have a nonzero coefficient. With  $mtry = 1$ , such a variable will sometimes be the only candidate for a split and will as such—because of its covariance with the response—create splits that do reduce impurity and will also decrease % MSE for OOB cases; that is, the permuted variable will perform worse than the original variable for the OOB cases because the split based on this variable picks up a real influence of the correlated regressor(s) that is lost otherwise for the particular tree. With  $mtry = 2$  or larger, one single

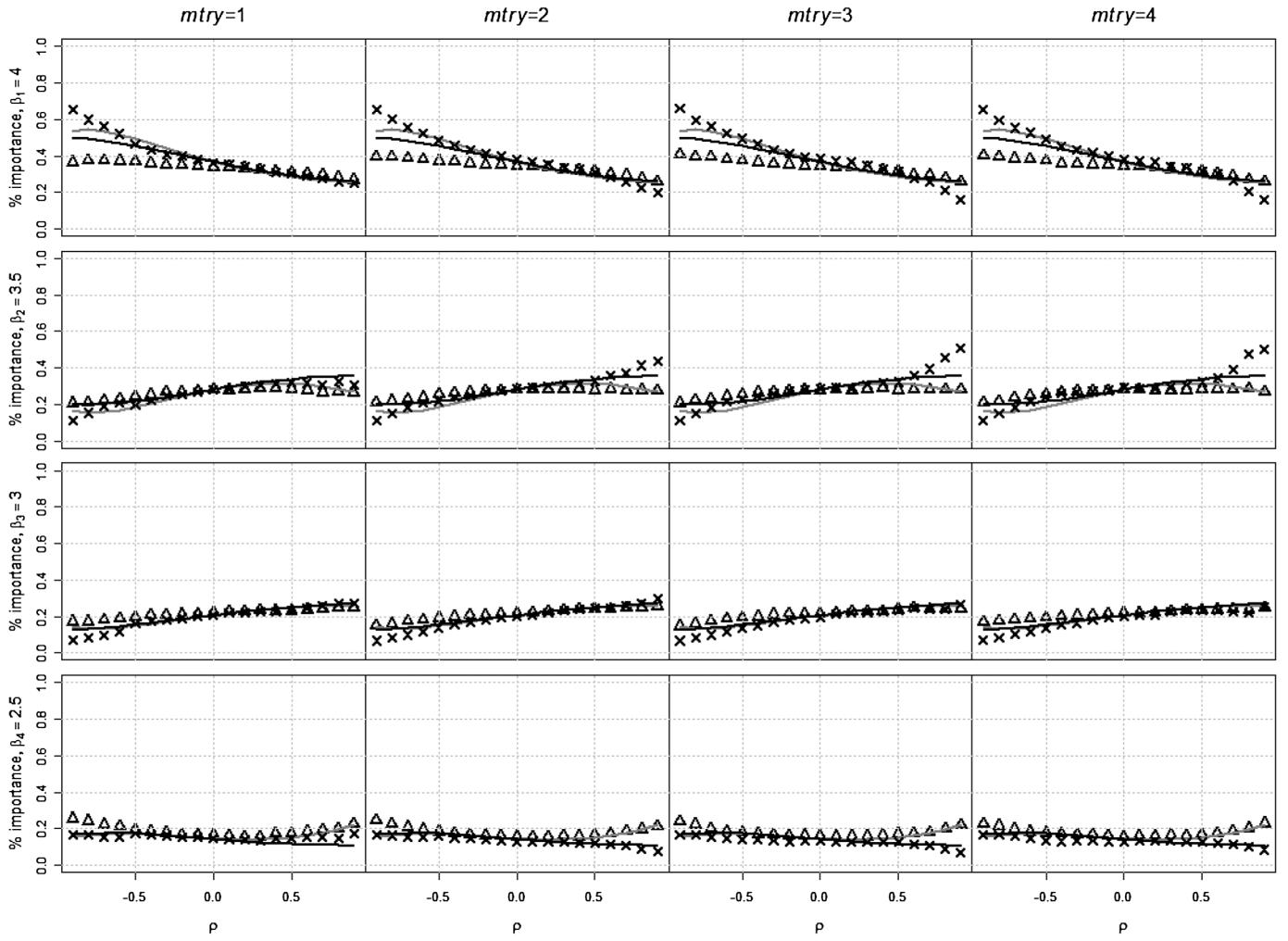


Figure 6. Average normalized importances for the four  $X$ 's (top to bottom:  $X_1$  to  $X_4$ ) from 100 simulated datasets for  $mtry = 1, 2, 3, 4$  (left to right) with  $\beta_7 = (4, 3.5, 3, 2.5)^T$ ,  $\text{corr}(X_j, X_k) = \rho^{|j-k|}$  with  $\rho = -0.9$  to 0.9 in steps of 0.1. Gray line: true normalized LMG allocation; black line: true normalized PMVD allocation.  $\Delta$ : variable importance (% MSE reduction) from RF-CART,  $\times$ : variable importance (% MSE reduction) from RF-CI.

weak regressor is usually in competition with another stronger regressor and will not win often, unless it is related to a stronger regressor: A weak regressor will stand in for a related stronger regressor, if the stronger regressor is neither sufficiently accommodated for in previous splits of the tree nor present in the current candidate set. With  $mtry$  increasing, it becomes less and less likely that a weak regressor will become the splitter. RF-CI will likely have stopped splitting before the weak regressors come into play with larger  $mtry$ . Therefore, it is plausible that the weaker effects will get lower allocations for larger values of  $mtry$  in RF-CI. This phenomenon has also been observed by Strobl et al. (2008, sec. 2.2) in the classification context. These considerations do not, however, capture the full nature of RF-CI allocations; especially, the behavior in Figure 6 that has been described above still awaits explanation.

Allocations from RF-CART were far less dependent on  $mtry$ . Conjecturing this behavior to be due to the much larger size of individual trees, a few simulations with minimum node sizes for splits of 2, 21, and 51 instead of the default 6 were run. These show that RF-CART with limited tree size (= larger min-

imum node size for splits) exhibits a much stronger dependence of allocations on  $mtry$ . The effect of reducing tree size was a strong increase in the equalizing behavior for small values of  $mtry$ , even for uncorrelated regressors. Note that—as previously mentioned for the standard tree setup—this equalizing behavior concerns strong and weak regressors but not regressors with no real effect. Also note that limiting tree size does not cause RF-CART to behave like RF-CI for  $mtry = p$ ; RF-CART remains more equalizing than RF-CI.

#### 6.4 Aspects for Further Research

As pointed out in the previous subsection, whereas it is reasonably well understood why strong variables substantially gain importance relative to weaker variables with increasing  $mtry$  in RF-CIs, other aspects of how variable importances depend on  $mtry$  should be subject to further investigation (e.g., behavior of RF-CI allocations in Figure 6). Also, it would be interesting to separate the effects of tree size from the effects of using  $p$ -values rather than maximum impurity reductions in comparing RF-CART to RF-CI. Furthermore, whereas it is convincing

that  $p$ -values are good at fairly treating unimportant variables of all types (i.e., null variables), it is not so clear that they are unbiased at representing the relative merits of several important variables of different strengths. The difficulty for investigating this lies in the fact that there is no easily identifiable gold-standard importance assessment of the strength for non-null variables (cf. also the following section).

The shape of the regressor space might be another interesting topic for further research. Contrary to linear models, forests (of course) do not only react to the mean- and covariance-structure of response and regressors but also to further aspects of the data. For example, variable importance in a forest has been found to depend on the shape of the regressor space: in a simulation (not shown) for the scenario of Section 6.1 with  $\beta_1$  and uncorrelated regressors distributed uniformly over a four-dimensional cube, variable importances differ depending on whether the cube sides are parallel to the coordinate axes or rotated (cf. graph and program in the supplemental material). It would be interesting to investigate this further.

## 7. THE CONCEPT OF VARIABLE IMPORTANCE

Variable importance is not very well defined as a concept. Even in the well-developed linear model and the standard  $n \gg p$  situation, there is no theoretically defined variable importance metric in the sense of a parametric quantity that a variable importance estimator *should* try to estimate. In the absence of a clearly agreed true value, ad hoc proposals for empirical assessment of variable importance have been made, and desirability criteria for these have been formulated, for example, “decomposition” of  $R^2$  into “nonnegative contributions attributable to each regressor” has been postulated (cf., e.g., Grömping 2007 for more detail). Popular approaches for empirically assessing a variable’s importance include squared correlations (a completely marginal approach) and squared standardized coefficients (an approach conditional on all other variables in the model), as critically discussed, for example, by Darlington (1968). In line with Johnson and Lebreton’s (2004) (vague) definition of relative importance, LMG and PMVD account for both marginal and conditional aspects of importance by averaging over  $R^2$  contributions from models with different variables preceding the respective variable of interest (and thus conditioned upon). LMG has been heuristically introduced and found a justification as a Shapley value later on (e.g., Stufken 1992) whereas PMVD was introduced by Feldman (2005) to satisfy the “exclusion” property, which requests that a variable with coefficient 0 should be allocated zero importance. Thus, PMVD is closer to a conditional perspective than LMG, in that it honors conditional unimportance of a variable given all others.

With their request that both conditional and marginal aspects need to be reflected in measuring relative importance, Johnson and Lebreton (2004) aimed for explanatory importance. Depending on the research question at hand, the focus of the variable importance assessment in regression can be explanatory or predictive importance or a mixture of both; cf. the article by Grömping (2007) for a more detailed discussion. To clearly differentiate between these two foci, consider the following very

simple examples of a causal chain:

$$X_2 \rightarrow X_1 \rightarrow Y, \quad (4)$$

$$X_2 \leftarrow X_1 \rightarrow Y. \quad (5)$$

In causal chain (4),  $X_2$  indirectly influences the response; in causal chain (5),  $X_2$  is correlated to the response but does not influence it. If all relations are linear, a linear model for  $Y$  with regressors  $X_1$  and  $X_2$  will have a zero coefficient for  $X_2$  in both (4) and (5), that is, conditional on  $X_1$ ,  $X_2$  does not contribute anything to the prediction. Thus, in the presence of sufficient data, both coefficient-based approaches and PMVD would allocate importance zero to  $X_2$ . This makes perfect sense for prediction purposes, and in causal chain (5) also for explanatory importance; however, most people would disagree that  $X_2$  is unimportant for  $Y$  in an explanatory or causal sense in causal chain (4).

In random forest applications, variable importance is typically used for variable selection; random forests are especially popular for  $p \gg n$  scenarios. In parallel to the distinction between explanatory/causal importance and predictive importance in conventional regression models, variable selection can serve two different objectives (cf., e.g., Diaz-Uriarte and Alvarez de Andrés 2006; Genuer, Poggi, and Tuleau 2008), namely (a) to identify important variables highly related to the response variable for explanatory and interpretation purposes (parallel to explanatory/causal) or (b) to identify a small number of variables sufficient for a good prediction of the response variable (parallel to prediction). In prediction or variable selection with purpose (b), one would strive to avoid redundancy and obtain a parsimonious prediction model. It is not so important that the model contains all relevant variables, as long as prediction works well. For example, ideally one would only select  $X_1$  in causal chains (4) or (5), but selection of  $X_2$  alone would also be acceptable, if the relation between  $X_1$  and  $X_2$  were sufficiently strong. (Because the linear model or the forest itself does not contain information about structural relationships between variables, it can be very difficult or even impossible to distinguish between the causal chains shown in (4) and (5) and the respective causal chains with the two  $X$ ’s in swapped roles.) On the other hand, for explanatory/causal importance or identification of potentially important variables in variable selection, it would be considered detrimental if an important variable is missed because of a low variable importance allocation, even though another highly correlated variable might well stand in for this variable in terms of prediction. Thus, one would certainly want the variable selection method to find both  $X_1$  and  $X_2$  in causal chain (4), whereas one would prefer not to consider  $X_2$  as important in causal chain (5). However, because the two causal chains are indistinguishable for a linear model or a random forest, one would have to accept identifying  $X_2$  as important in (5) as the price for being able to find it in (4).

The desire for an adequate variable selection method for purpose (a) was the starting point for Zou and Hastie’s (2005) introduction of the elastic net: They modified the lasso, which is known to have a tendency to select one representative of a group of strongly correlated variables only (i.e., to work well for purpose (b)), into showing a “grouping property,” that is,

a tendency to select correlated variables together. Put simply, grouping is achieved by biasing coefficients such that for highly correlated regressors coefficients with higher absolute values are shrunk toward the origin and coefficients with lower absolute values are “shrunk” away from the origin, so that the whole group ends up to be chosen together. “Grouping” of correlated regressors is conceptually close to what has been previously called LMG’s “equalizing” behavior. Thus, purpose (a) is served better because of a reduced risk of missing an important variable in the presence of further variables highly correlated with it.

Within the range of squared (standardized) coefficients at the conditional extreme, over PMVD, LMG to squared marginal correlations at the marginal extreme, simulations showed that RF-CART and RF-CI with small  $mtry$  behave similarly to LMG, that is, balance between conditional and marginal approach leaning toward marginal, whereas RF-CI with large  $mtry$  behaves more similarly to PMVD, that is, also balances between marginal and conditional approach leaning toward the conditional end. Strobl et al. (2008) positioned themselves at the conditional extreme by considering it “bias” that variables with the same coefficients receive different importances due to inter-regressor correlations; thus, following their logic, one would also have to reject the idea of predicting  $Y$  based on  $X_2$  instead of  $X_1$  in causal chains (4) or (5) above, except for the limiting case for which there is a perfectly deterministic relation between the two regressors in the data. Of course, with  $p \gg n$  variable selection, unbiased estimation of all coefficients is impossible—additional constraints or penalties (like in the elastic net) reintroduce estimability but also bias into the estimates. Strobl et al. (2008) suggested getting closer to conditional importance by using RF-CI (instead of RF-CART) together with a conditional instead of an unconditional permutation algorithm for the OOB data in (3). However, their approach does not attack a key driver of marginality in the random forest approach: as long as  $mtry < p$ , correlated variables will more or less frequently stand in for stronger predictors and thus act as splitters for marginal reasons. Results from the simulation study suggest that using RF-CI with  $mtry = p$  might already go a long way toward making RF-CI more conditional.

Both random forest variable importances and LMG and PMVD are based on shares of the explained model variance or reductions in error variation by using the true instead of the permuted values for a variable in prediction (again, beware of misinterpreting forest’s MSE reduction; cf. also Section 4.2.3). Achen (1982) referred to measures based on the response’s dispersion as “dispersion importance” (cf. also Johnson and Lebreton 2004; Grömping 2007; Genuer, Poggi, and Tuleau 2008). It is obvious from (2) that the variance induced by regressors depends on both the coefficients and the correlations between regressors. For example, a group of regressors with positive regression coefficients and large positive inter-regressor correlations (e.g., 0.9) contributes more than twice as much to the variance (2) than another group of uncorrelated regressors with the same absolute magnitude of regression coefficients, according to the impact of the second summand. Thus, an importance metric based on dispersion must be expected to reflect this dependence on inter-regressor correlations, allocating higher importance to correlated regressors. If this is unwanted, one may

go to great lengths to eliminate the correlation influences—for example, by introducing intricate data-dependent weights as in PMVD or a conditional permutation scheme as in the work of Strobl et al. 2008—or one might try to find an altogether different basis for variable importance allocation, for example, standardized coefficients in the linear model; unfortunately, there is no unbiased equivalent for these in the  $p \gg n$  situation.

## SUPPLEMENTAL MATERIAL

**Simulation programs:** A preparation program (`utilityPrograms.R`, R program code) loads all required packages and provides calculation routines that are used in the simulation. The simulation function is provided in a separate file (`simulationProgram.R`, R program code). After running the previous two programs, (`callSimulationProgram.R`, R program code) can be adapted to do the simulations that the user is interested in. Comments within the programs give further instructions. (`programs.zip`)

**Dependence of variable importance in random forests on the shape of the regressor space:** As pointed out in Section 6.4, the simulations have indicated that variable importance in random forests depends on the shape of the regressor space. This is further explained in this supplement. (`Supplement_shape.pdf`, Acrobat file)

[Received September 2008. Revised August 2009.]

## REFERENCES

- Achen, C. H. (1982), *Interpreting and Using Regression*, Beverly Hills, CA: Sage.
- Breiman, L. (2001), “Random Forests,” *Machine Learning*, 45, 5–32.
- (2002), “Manual on Setting Up, Using, and Understanding Random Forests V3.1,” unpublished manuscript, available at [http://oz.berkeley.edu/users/breiman/Using\\_random\\_forests\\_V3.1.pdf](http://oz.berkeley.edu/users/breiman/Using_random_forests_V3.1.pdf).
- (2003), “Manual on Setting Up, Using, and Understanding Random Forests V4.0,” unpublished manuscript, available at [ftp://ftp.stat.berkeley.edu/pub/users/breiman/Using\\_random\\_forests\\_v4.0.pdf](ftp://ftp.stat.berkeley.edu/pub/users/breiman/Using_random_forests_v4.0.pdf).
- Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984), *Classification and Regression Trees*, Belmont, CA: Wadsworth.
- Budescu, D. V. (1993), “Dominance Analysis: A New Approach to the Problem of Relative Importance of Predictors in Multiple Regression,” *Psychological Bulletin*, 114, 542–551.
- Chevan, A., and Sutherland, M. (1991), “Hierarchical Partitioning,” *The American Statistician*, 45, 90–96.
- Christensen, R. (1992), Comment on “Hierarchical Partitioning,” by A. Chevan and M. Sutherland, *The American Statistician*, 46, 74.
- Darlington, R. B. (1968), “Multiple Regression in Psychological Research and Practice,” *Psychological Bulletin*, 69, 161–182.
- Díaz-Uriarte, R., and Alvarez de Andrés, S. (2006), “Gene Selection and Classification of Microarray Data Using Random Forest,” *BMC Bioinformatics*, 7, 3.
- Ehrenberg, A. S. C. (1990), “The Unimportance of Relative Importance,” *The American Statistician*, 44, 260.
- Feldman, B. (2005), “Relative Importance and Value,” unpublished manuscript, available at [http://www.prismalytics.com/docs/RelativeImportance050319.pdf](http://www.prismanalytics.com/docs/RelativeImportance050319.pdf).
- Friedman, J. (1991), “Multivariate Additive Regression Splines,” *The Annals of Statistics*, 19, 82–91.

- Genuer, R., Poggi, J.-M., and Tuleau, C. (2008), "Random Forests: Some Methodological Insights," Research Report 6729, Institut National de Recherche en Informatique et en Automatique. ISSN 0249–6399.
- Grömping, U. (2006), "Relative Importance for Linear Regression in R: The Package relaimpo," *Journal of Statistical Software*, 17, 1. Available at <http://www.jstatsoft.org/v17/i01/>.
- (2007), "Estimators of Relative Importance in Linear Regression Based on Variance Decomposition," *The American Statistician*, 61, 139–147.
- Hoerl, A. E., and Kennard, R. W. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55–67.
- Hothorn, T., Hornik, K., van de Wiel, M. A., and Zeileis, A. (2006a), "A Lego System for Conditional Inference," *The American Statistician*, 60, 257–263.
- Hothorn, T., Hornik, K., and Zeileis, A. (2006b), "Unbiased Recursive Partitioning: A Conditional Inference Framework," *Journal of Computational and Graphical Statistics*, 15, 651–674.
- Hothorn, T., Lausen, B., Benner, A., and Radespiel-Tröger, M. (2004), "Bagging Survival Trees," *Statistics in Medicine*, 23 (1), 77–91.
- Ishwaran, H. (2007), "Variable Importance in Binary Regression Trees and Forests," *Electronic Journal of Statistics*, 1, 519–537.
- Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008), "Random Survival Forests," *The Annals of Applied Statistics*, 2, 841–860.
- Johnson, J. W., and Lebreton, J. M. (2004), "History and Use of Relative Importance Indices in Organizational Research," *Organizational Research Methods*, 7, 238–257.
- Kass, G. (1980), "An Exploratory Technique for Investigating Large Quantities of Categorical Data," *Applied Statistics*, 29, 119–127.
- Kruskal, W. (1987a), "Relative Importance by Averaging Over Orderings," *The American Statistician*, 41, 6–10.
- (1987b), Correction to "Relative Importance by Averaging Over Orderings," *The American Statistician*, 41, 341.
- Liaw, A., and Wiener, M. (2002), "Classification and Regression by randomForest," *R News*, 2, 18–22.
- Lindeman, R. H., Merenda, P. F., and Gold, R. Z. (1980), *Introduction to Bivariate and Multivariate Analysis*, Glenview, IL: Scott, Foresman.
- Lipovetsky, S., and Conklin, M. (2001), "Analysis of Regression in Game Theory Approach," *Applied Stochastic Models in Business and Industry*, 17, 319–330.
- R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. Available at <http://www.R-project.org>, ISBN 3-900051-07-0.
- Segal, M. R., Barbour, J. D., and Grant, R. M. (2004), "Relating HIV-1 Sequence Variation to Replication Capacity via Trees and Forests," *Statistical Applications in Genetics and Molecular Biology*, 3, article 2.
- Shih, Y.-S., and Tsai, H.-W. (2004), "Variable Selection Bias in Regression Trees With Constant Fits," *Computational Statistics and Data Analysis*, 45, 595–607.
- Strobl, C., Boulesteix, A., Kneib, T., Augustin, T., and Zeileis, A. (2008), "Conditional Variable Importance for Random Forests," *BMC Bioinformatics*, 9, 307.
- Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007), "Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution," *BMC Bioinformatics*, 8, 25.
- Stufken, J. (1992), "On Hierarchical Partitioning," *The American Statistician*, 46, 70–71.
- Theil, H., and Chung, C.-F. (1988), "Information-Theoretic Measures of Fit for Univariate and Multivariate Linear Regressions," *The American Statistician*, 42, 249–252.
- Therneau, T., and Atkinson, E. (1997), "An Introduction to Recursive Partitioning Using the rpart Routines," Technical Report 61, Mayo Clinic, Section of Statistics.
- Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Ser. B*, 58, 267–288.
- van der Laan, M. (2006), "Statistical Inference for Variable Importance," *The International Journal of Biostatistics*, 2, 1008–1008.
- Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society, Ser. B*, 67, 301–320.

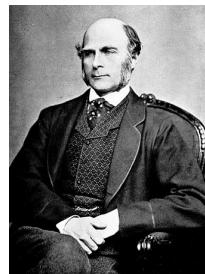


# Introduction to regression

## Regression

Brian Caffo, Jeff Leek and Roger Peng  
Johns Hopkins Bloomberg School of Public Health

# A famous motivating example



(Perhaps surprisingly, this example is still relevant)

A screenshot of a journal article from the European Journal of Human Genetics. The title is "Predicting human height by Victorian and genomic methods". The article is authored by Karl E. Kendrew, David J. Paterson, and Michael J. Bamshad. It discusses the use of Victorian-era data from the Royal Anthropological Institute's Stature Project to predict human height, comparing it with modern genomic approaches. The page includes a sidebar with journal navigation links and a footer with the EJHG logo.

<http://www.nature.com/ejhg/journal/v17/n8/full/ejhg20095a.html>

Predicting height: the Victorian approach beats modern genomics

# Questions for this class

- Consider trying to answer the following kinds of questions:
  - To use the parents' heights to predict childrens' heights.
  - To try to find a parsimonious, easily described mean relationship between parent and children's heights.
  - To investigate the variation in childrens' heights that appears unrelated to parents' heights (residual variation).
  - To quantify what impact genotype information has beyond parental height in explaining child height.
  - To figure out how/whether and what assumptions are needed to generalize findings beyond the data in question.
  - Why do children of very tall parents tend to be tall, but a little shorter than their parents and why children of very short parents tend to be short, but a little taller than their parents? (This is a famous question called 'Regression to the mean'.)

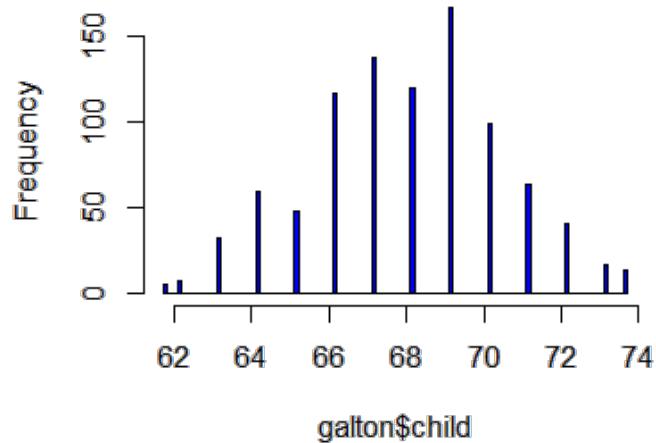
# Galton's Data

- Let's look at the data first, used by Francis Galton in 1885.
- Galton was a statistician who invented the term and concepts of regression and correlation, founded the journal Biometrika, and was the cousin of Charles Darwin.
- You may need to run `install.packages("UsingR")` if the `UsingR` library is not installed.
- Let's look at the marginal (parents disregarding children and children disregarding parents) distributions first.
  - Parent distribution is all heterosexual couples.
  - Correction for gender via multiplying female heights by 1.08.
  - Overplotting is an issue from discretization.

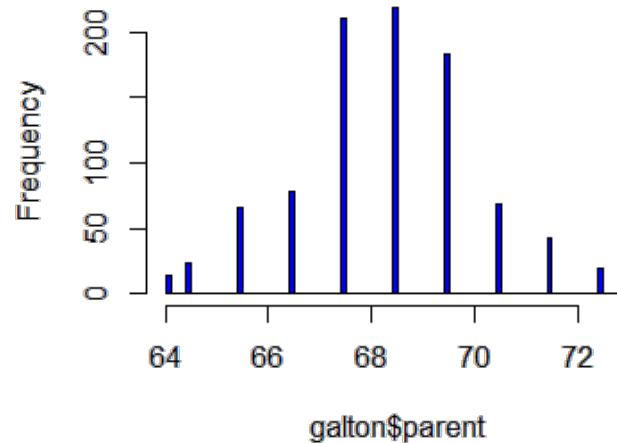
# Code

```
library(UsingR); data(galton)
par(mfrow=c(1,2))
hist(galton$child,col="blue",breaks=100)
hist(galton$parent,col="blue",breaks=100)
```

Histogram of galton\$child



Histogram of galton\$parent



# Finding the middle via least squares

- Consider only the children's heights.
  - How could one describe the "middle"?
  - One definition, let  $Y_i$  be the height of child  $i$  for  $i = 1, \dots, n = 928$ , then define the middle as the value of  $\mu$  that minimizes

$$\sum_{i=1}^n (Y_i - \mu)^2$$

- This is physical center of mass of the histogram.
- You might have guessed that the answer  $\mu = \bar{X}$ .

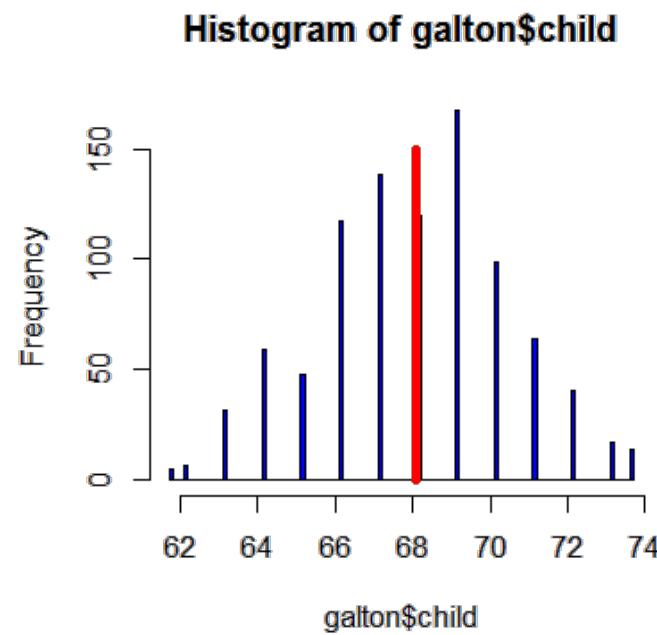
# Experiment

Use R studio's manipulate to see what value of  $\mu$  minimizes the sum of the squared deviations.

```
library(manipulate)
myHist <- function(mu){
  hist(galton$child,col="blue",breaks=100)
  lines(c(mu, mu), c(0, 150),col="red",lwd=5)
  mse <- mean((galton$child - mu)^2)
  text(63, 150, paste("mu = ", mu))
  text(63, 140, paste("MSE = ", round(mse, 2)))
}
manipulate(myHist(mu), mu = slider(62, 74, step = 0.5))
```

# The least squares estimate is the empirical mean

```
hist(galton$child,col="blue",breaks=100)
meanChild <- mean(galton$child)
lines(rep(meanChild,100),seq(0,150,length=100),col="red",lwd=5)
```

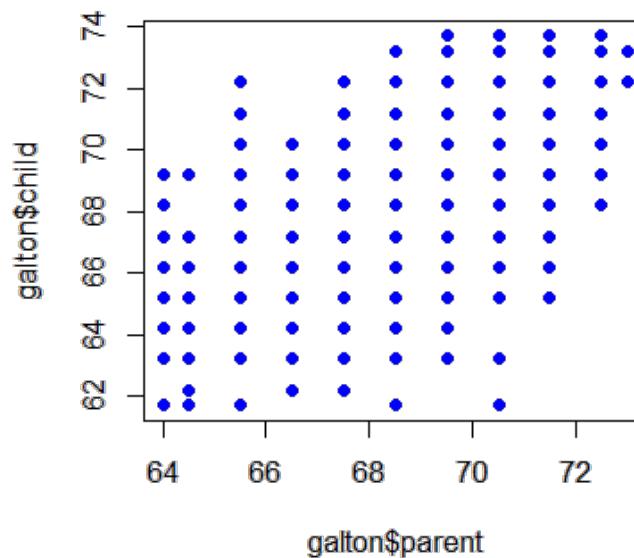


The math follows as:

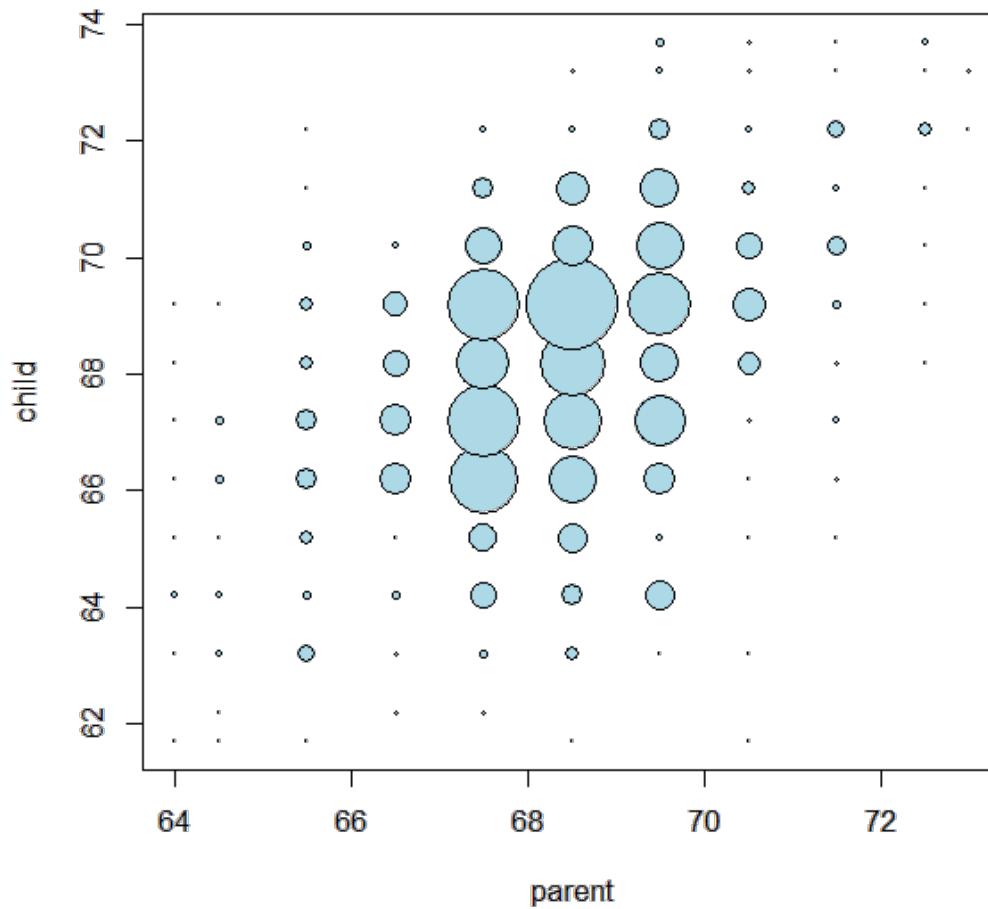
$$\begin{aligned}\sum_{i=1}^n (Y_i - \mu)^2 &= \sum_{i=1}^n (Y_i - \bar{Y} + \bar{Y} - \mu)^2 \\&= \sum_{i=1}^n (Y_i - \bar{Y})^2 + 2 \sum_{i=1}^n (Y_i - \bar{Y})(\bar{Y} - \mu) + \sum_{i=1}^n (\bar{Y} - \mu)^2 \\&= \sum_{i=1}^n (Y_i - \bar{Y})^2 + 2(\bar{Y} - \mu) \sum_{i=1}^n (Y_i - \bar{Y}) + \sum_{i=1}^n (\bar{Y} - \mu)^2 \\&= \sum_{i=1}^n (Y_i - \bar{Y})^2 + 2(\bar{Y} - \mu) \left( \sum_{i=1}^n Y_i - n\bar{Y} \right) + \sum_{i=1}^n (\bar{Y} - \mu)^2 \\&= \sum_{i=1}^n (Y_i - \bar{Y})^2 + \sum_{i=1}^n (\bar{Y} - \mu)^2 \\&\geq \sum_{i=1}^n (Y_i - \bar{Y})^2\end{aligned}$$

# Comparing childrens' heights and their parents' heights

```
plot(galton$parent,galton$child,pch=19,col="blue")
```



Size of point represents number of points at that (X, Y) combination (See the Rmd file for the code).



# Regression through the origin

- Suppose that  $X_i$  are the parents' heights.
- Consider picking the slope  $\beta$  that minimizes

$$\sum_{i=1}^n (Y_i - X_i \beta)^2$$

- This is exactly using the origin as a pivot point picking the line that minimizes the sum of the squared vertical distances of the points to the line
- Use R studio's manipulate function to experiment
- Subtract the means so that the origin is the mean of the parent and children's heights

```

myPlot <- function(beta){
  y <- galton$child - mean(galton$child)
  x <- galton$parent - mean(galton$parent)
  freqData <- as.data.frame(table(x, y))
  names(freqData) <- c("child", "parent", "freq")
  plot(
    as.numeric(as.vector(freqData$parent)),
    as.numeric(as.vector(freqData$child)),
    pch = 21, col = "black", bg = "lightblue",
    cex = .15 * freqData$freq,
    xlab = "parent",
    ylab = "child"
  )
  abline(0, beta, lwd = 3)
  points(0, 0, cex = 2, pch = 19)
  mse <- mean( (y - beta * x)^2 )
  title(paste("beta = ", beta, "mse = ", round(mse, 3)))
}
manipulate(myPlot(beta), beta = slider(0.6, 1.2, step = 0.02))

```

# The solution

In the next few lectures we'll talk about why this is the solution

```
lm(I(child - mean(child)) ~ I(parent - mean(parent)) - 1, data = galton)
```

Call:

```
lm(formula = I(child - mean(child)) ~ I(parent - mean(parent)) -  
  1, data = galton)
```

Coefficients:

```
I(parent - mean(parent))  
  0.646
```

# Visualizing the best fit line

Size of points are frequencies at that X, Y combination

