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 ϵ_i are assumed iid $N(0, \sigma^2)$.
- Note, $E[Y_i \mid X_i = x_i] = \mu_i = \beta_0 + \beta_1 x_i$
- Note, $Var(Y_i \mid X_i = x_i) = \sigma^2$.
- 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 maximimum likelihood estimate (regardless of σ)

Recap

- Model $Y_i = \mu_i + \epsilon_i = \beta_0 + \beta_1 X_i + \epsilon_i$ where ϵ_i are iid $N(0, \sigma^2)$
- ML estimates of β_0 and β_1 are the least squares estimates

$$\hat\beta_1 = Cor(Y, X)\frac{Sd(Y)}{Sd(X)} \quad \hat\beta_0 = \bar Y - \hat\beta_1 \bar X$$

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

Interpretting regression coefficients, the itc

- β_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 you X values by value a changes the intercept, but not the slope.

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

Interpretting regression coefficients, the slope

- β_1 is the expected change in response for a 1 unit change in the predictor

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

- 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 β_1 is kg/m. Converting X to cm implies multiplying X by $100cm/m$. To get β_1 in the right units, we have to divide by $100cm/m$ to get it to have the right units.

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

Using regression coeficients 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 Xs, 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 μ_i expressed as points on a line

Example

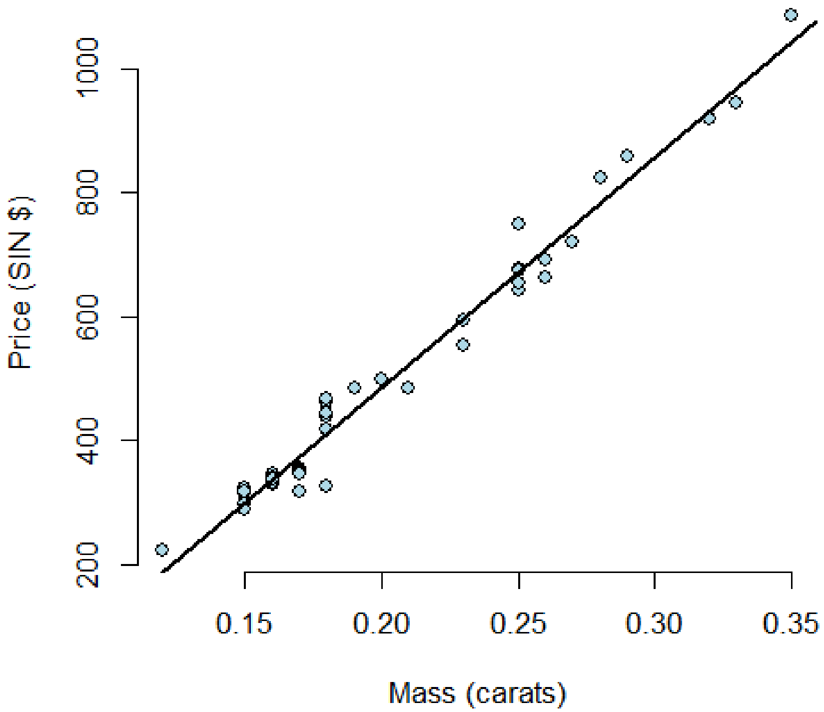
`diamond` data set from `UsingR`

Data is diamond prices (Signapore 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 coeficient 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

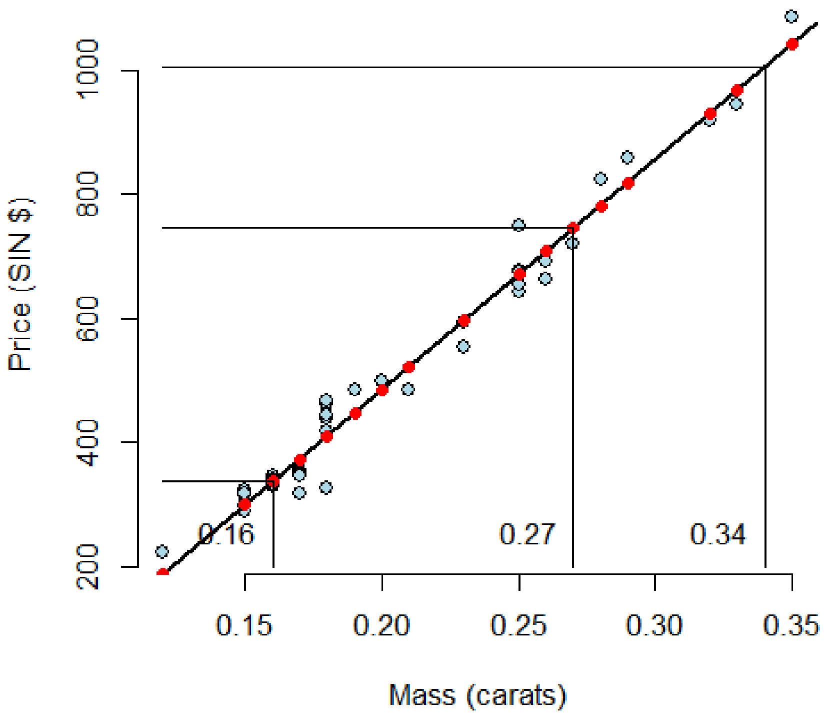
```
newx <- c(0.16, 0.27, 0.34)
coef(fit)[1] + coef(fit)[2] * newx
```

```
[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 valuve X_i is

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

- Residual, the 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 ϵ_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

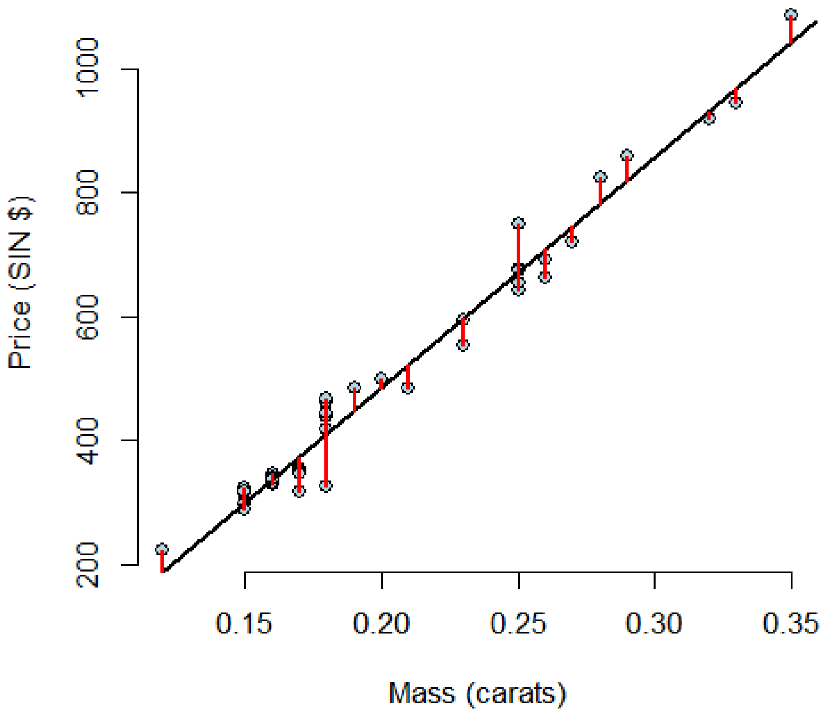
```
data(diamond)
y <- diamond$price; x <- diamond$carat; n <- length(y)
fit <- lm(y ~ x)
e <- resid(fit)
yhat <- predict(fit)
max(abs(e -(y - yhat)))
```

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

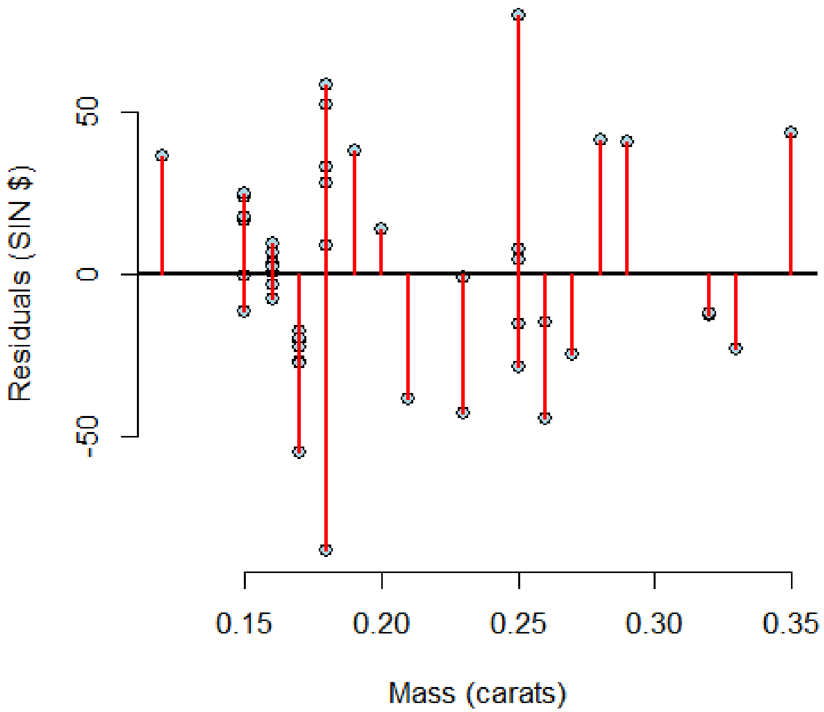
```
max(abs(e - (y - coef(fit)[1] - coef(fit)[2] * x)))
```

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

Residuals are the signed length of the red lines

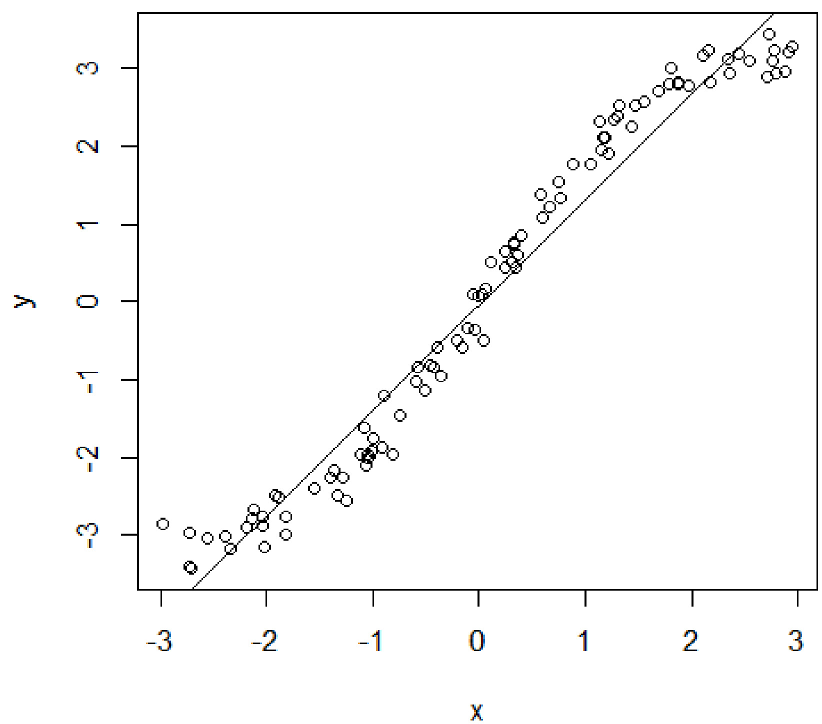


Residuals versus X

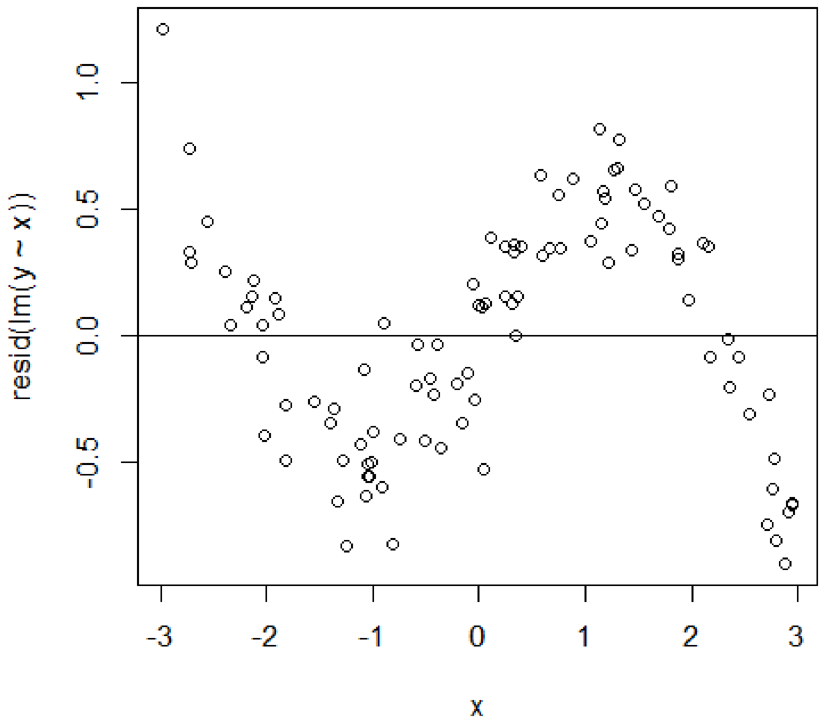


Non-linear data

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

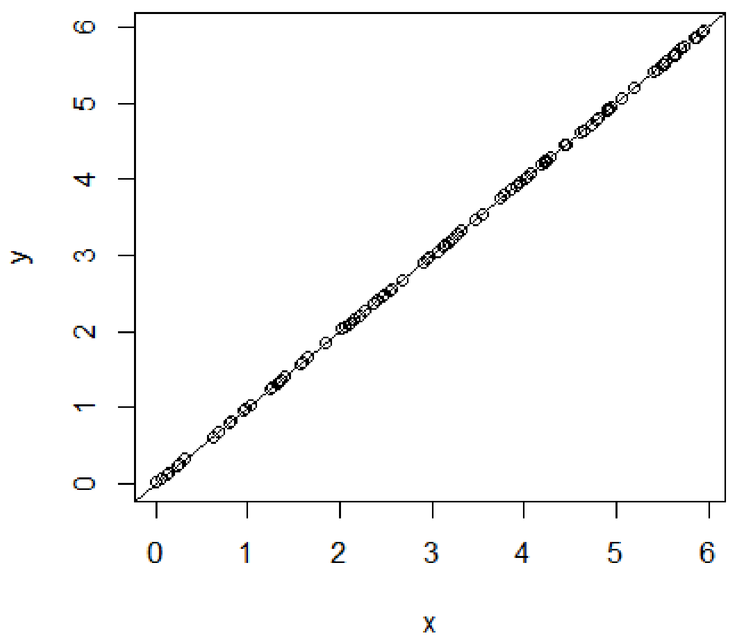


```
plot(x, resid(lm(y ~ x)));
abline(h = 0)
```



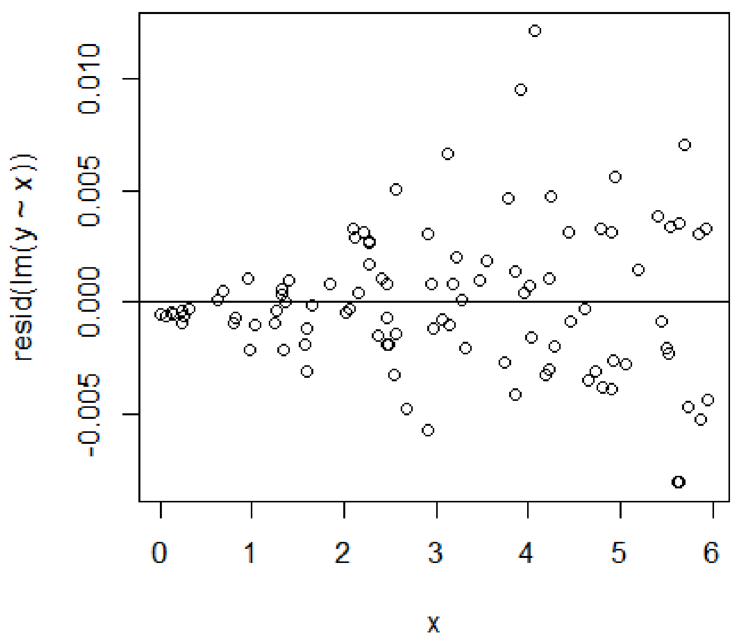
Heteroskedasticity

```
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 σ^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

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

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

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 varation 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 corrrelation)

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

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

Since, recall,

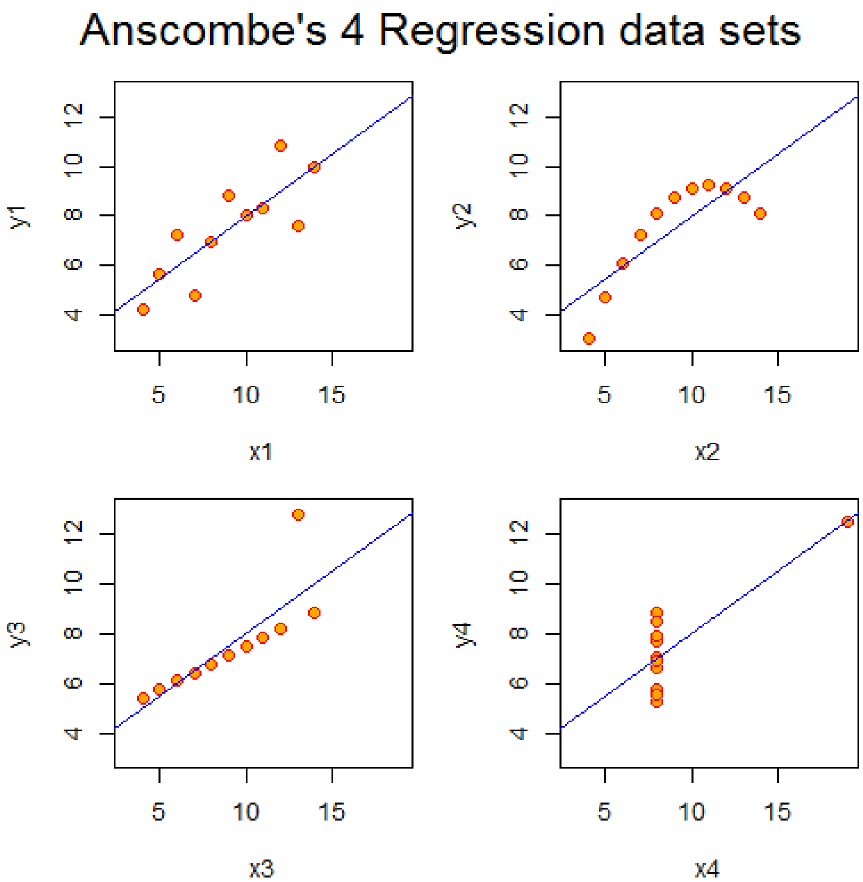
$$\hat{\beta}_1 = Cor(Y, X) \frac{Sd(Y)}{Sd(X)}$$

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)



Inference in regression

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

Recall our model and fitted values

- Consider the model

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

- $\epsilon \sim N(0, \sigma^2)$.
- We assume that the true model is known.
- We assume that you've seen confidence intervals and hypothesis tests before.
- $\hat\beta_0 = \bar Y - \hat\beta_1 \bar X$
- $\hat\beta_1 = Cor(Y, X) \frac{Sd(Y)}{Sd(X)}$.

Review

- Statistics like $\frac{\hat \theta - \theta}{\hat \sigma_{\hat \theta}}$ often have the following properties.
  1. Is normally distributed and has a finite sample Student's T distribution if the estimated variance is replaced with a sample estimate (under normality assumptions).
  2. Can be used to test $H_0 : \theta = \theta_0$ versus $H_a : \theta >, <, \neq \theta_0$.
  3. Can be used to create a confidence interval for θ via $\hat \theta \pm Q_{1-\alpha/2} \hat \sigma_{\hat \theta}$ where $Q_{1-\alpha/2}$ is the relevant quantile from either a normal or T distribution.
- In the case of regression with iid sampling assumptions and normal errors, our inferences will follow very similarily to what you saw in your inference class.
- We won't cover asymptotics for regression analysis, but suffice it to say that under assumptions on the ways in which the X values are collected, the iid sampling model, and mean model, the normal results hold to create intervals and confidence intervals

Standard errors (conditioned on X)

$$
\begin{aligned}
\mathrm{Var}(\hat{\beta}_1) &= \mathrm{Var}\left(\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right) \\
&= \frac{\mathrm{Var}\left(\sum_{i=1}^{n} Y_i(X_i - \bar{X})\right)}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2} \\
&= \frac{\sum_{i=1}^{n} \sigma^2 (X_i - \bar{X})^2}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2} \\
&= \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}
\end{aligned}
$$

Results

- $\sigma^2_{\hat\beta_1} = Var(\hat\beta_1) = \sigma^2 / \sum_{i=1}^n (X_i - \bar X)^2$
- $\sigma^2_{\hat\beta_0} = Var(\hat\beta_0) = \left(\frac{1}{n} + \frac{\bar X^2}{\sum_{i=1}^n (X_i - \bar X)^2}\right)\sigma^2$
- In practice, σ is replaced by its estimate.
- It's probably not surprising that under iid Gaussian errors

$$\frac{\hat\beta_j - \beta_j}{\hat\sigma_{\hat\beta_j}}$$

follows a t distribution with $n-2$ degrees of freedom and a normal distribution for large n.

- This can be used to create confidence intervals and perform hypothesis tests.

Example diamond data set

```r
library(UsingR); data(diamond)
y <- diamond$price; x <- diamond$carat; n <- length(y)
beta1 <- cor(y, x) * sd(y) / sd(x)
beta0 <- mean(y) - beta1 * mean(x)
e <- y - beta0 - beta1 * x
sigma <- sqrt(sum(e^2) / (n-2))
ssx <- sum((x - mean(x))^2)
seBeta0 <- (1 / n + mean(x) ^ 2 / ssx) ^ .5 * sigma
seBeta1 <- sigma / sqrt(ssx)
tBeta0 <- beta0 / seBeta0; tBeta1 <- beta1 / seBeta1
pBeta0 <- 2 * pt(abs(tBeta0), df = n - 2, lower.tail = FALSE)
pBeta1 <- 2 * pt(abs(tBeta1), df = n - 2, lower.tail = FALSE)
coefTable <- rbind(c(beta0, seBeta0, tBeta0, pBeta0), c(beta1, seBeta1, tBeta1, pBeta1))
colnames(coefTable) <- c("Estimate", "Std. Error", "t value", "P(>|t|)")
rownames(coefTable) <- c("(Intercept)", "x")
```

Example continued

```
coefTable
```

```
            Estimate Std. Error t value   P(>|t|)
(Intercept)   -259.6      17.32  -14.99 2.523e-19
x             3721.0      81.79   45.50 6.751e-40
```

```
fit <- lm(y ~ x);
summary(fit)$coefficients
```

```
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)   -259.6      17.32  -14.99 2.523e-19
x             3721.0      81.79   45.50 6.751e-40
```

Getting a confidence interval

```
sumCoef <- summary(fit)$coefficients
sumCoef[1,1] + c(-1, 1) * qt(.975, df = fit$df) * sumCoef[1, 2]
```

```
[1] -294.5 -224.8
```

```
sumCoef[2,1] + c(-1, 1) * qt(.975, df = fit$df) * sumCoef[2, 2]
```

```
[1] 3556 3886
```

With 95% confidence, we estimate that a 0.1 carat increase in diamond size results in a 355.6 to 388.6 increase in price in (Singapore) dollars.

Prediction of outcomes

- Consider predicting Y at a value of X
  - Predicting the price of a diamond given the carat
  - Predicting the height of a child given the height of the parents
- The obvious estimate for prediction at point x_0 is

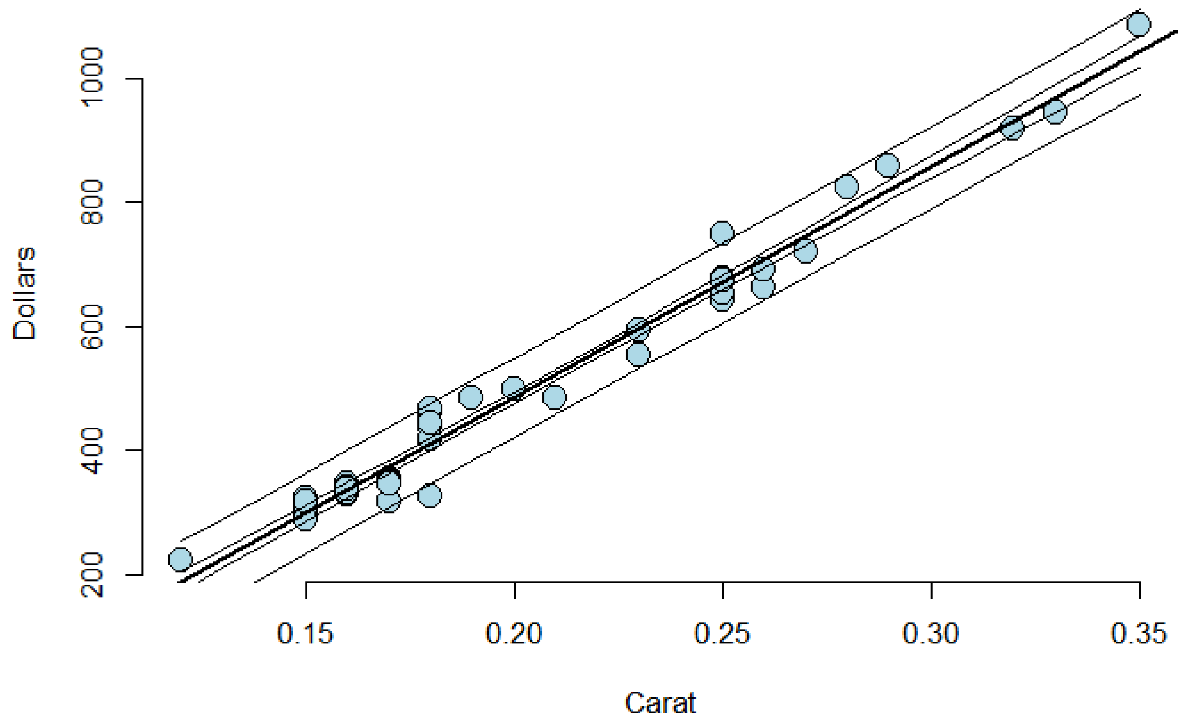
$$\hat\beta_0 + \hat\beta_1 x_0$$

- A standard error is needed to create a prediction interval.
- There's a distinction between intervals for the regression line at point x_0 and the prediction of what a y would be at point x_0.
- Line at x_0 se, $\hat\sigma\sqrt{\frac{1}{n} + \frac{(x_0 - \bar X)^2}{\sum_{i=1}^n (X_i - \bar X)^2}}$
- Prediction interval se at x_0, $\hat\sigma\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar X)^2}{\sum_{i=1}^n (X_i - \bar X)^2}}$

Plotting the prediction intervals

```
plot(x, y, frame=FALSE,xlab="Carat",ylab="Dollars",pch=21,col="black", bg="lightblue", cex=2)
abline(fit, lwd = 2)
xVals <- seq(min(x), max(x), by = .01)
yVals <- beta0 + beta1 * xVals
se1 <- sigma * sqrt(1 / n + (xVals - mean(x))^2/ssx)
se2 <- sigma * sqrt(1 + 1 / n + (xVals - mean(x))^2/ssx)
lines(xVals, yVals + 2 * se1)
lines(xVals, yVals - 2 * se1)
lines(xVals, yVals + 2 * se2)
lines(xVals, yVals - 2 * se2)
```

Plotting the prediction intervals



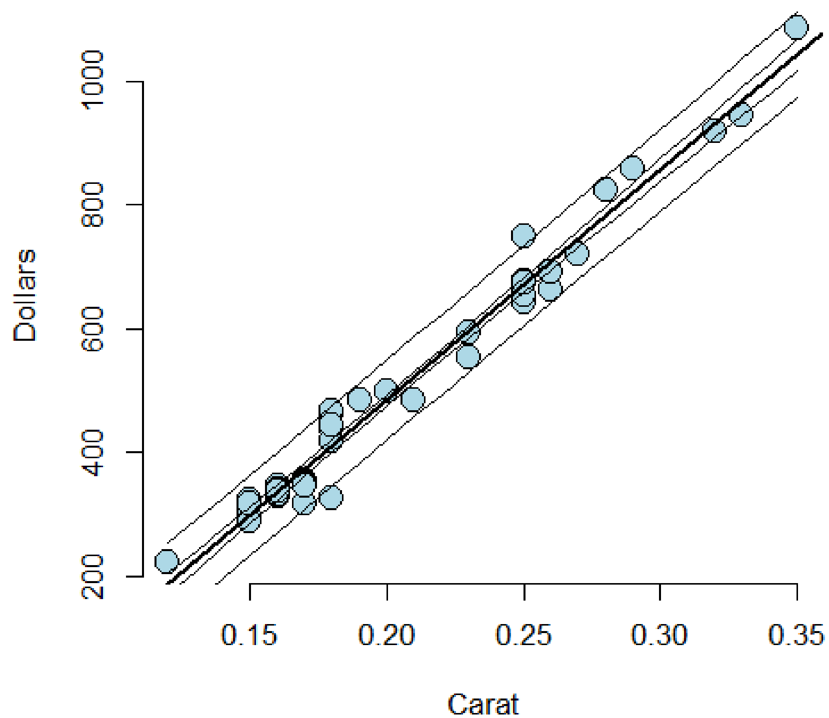
Discussion

- Both intervals have varying widths.
  - Least width at the mean of the Xs.
- We are quite confident in the regression line, so that interval is very narrow.
  - If we knew β_0 and β_1 this interval would have zero width.
- The prediction interval must incorporate the variabilibity in the data around the line.
  - Even if we knew β_0 and β_1 this interval would still have width.

In R

```
newdata <- data.frame(x = xVals)
p1 <- predict(fit, newdata, interval = ("confidence"))
p2 <- predict(fit, newdata, interval = ("prediction"))
plot(x, y, frame=FALSE,xlab="Carat",ylab="Dollars",pch=21,col="black", bg="lightblue", cex=2)
abline(fit, lwd = 2)
lines(xVals, p1[,2]); lines(xVals, p1[,3])
lines(xVals, p2[,2]); lines(xVals, p2[,3])
```

In R



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 useage (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 incoporate 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.

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

- 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

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

  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 μ_i, then we've found the LS estimates.

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

- 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

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

- 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 $e_{i,a|b} = a_i - \frac{\sum_{j=1}^{n} a_j b_j}{\sum_{i=1}^{n} b_j^2} b_i$ 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

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

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

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 β_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 β_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 $\frac{\sum_j X_{2j} X_{1j}}{\sum_j X_{2j}^2} X_{2i} = \frac{\sum_j X_{1j}}{n} = \bar{X}_1$.
- $e_{i,X_1|X_2} = X_{1i} - \bar{X}_1$.
- Simiarly $e_{i,Y|X_2} = Y_i - \bar{Y}$.
- 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

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

  for $k = 1, \ldots, p$ 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:

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

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

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

Plugging this back into the equations, we wind up with

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

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

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

is equal to

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

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 conceputally, 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 coeficient

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

So that

$$E[Y|X_1 = x_1 + 1, \ldots, X_p = x_p] - E[Y|X_1 = x_1, \ldots, X_p = x_p]$$

$$= (x_1 + 1)\beta_1 + \sum_{k=2}^{p} x_k + \sum_{k=1}^{p} x_k \beta_k = \beta_1$$

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, $x_1, \ldots, x_p$, 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 techniqe, *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.

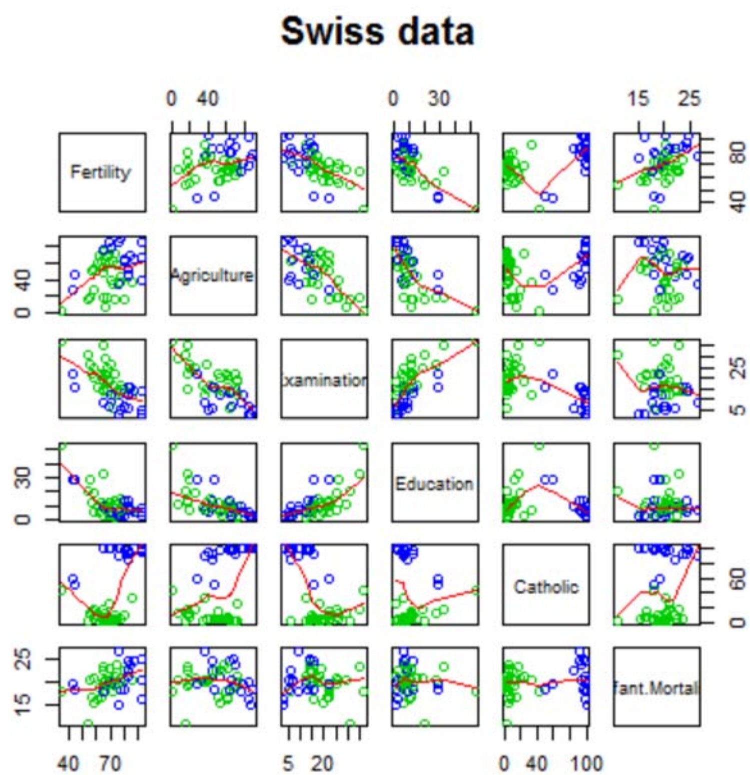
Multivariable regression examples

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 $H_0 : \beta_{Agri} = 0$ versus $H_a : \beta_{Agri} \neq 0$ 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
```

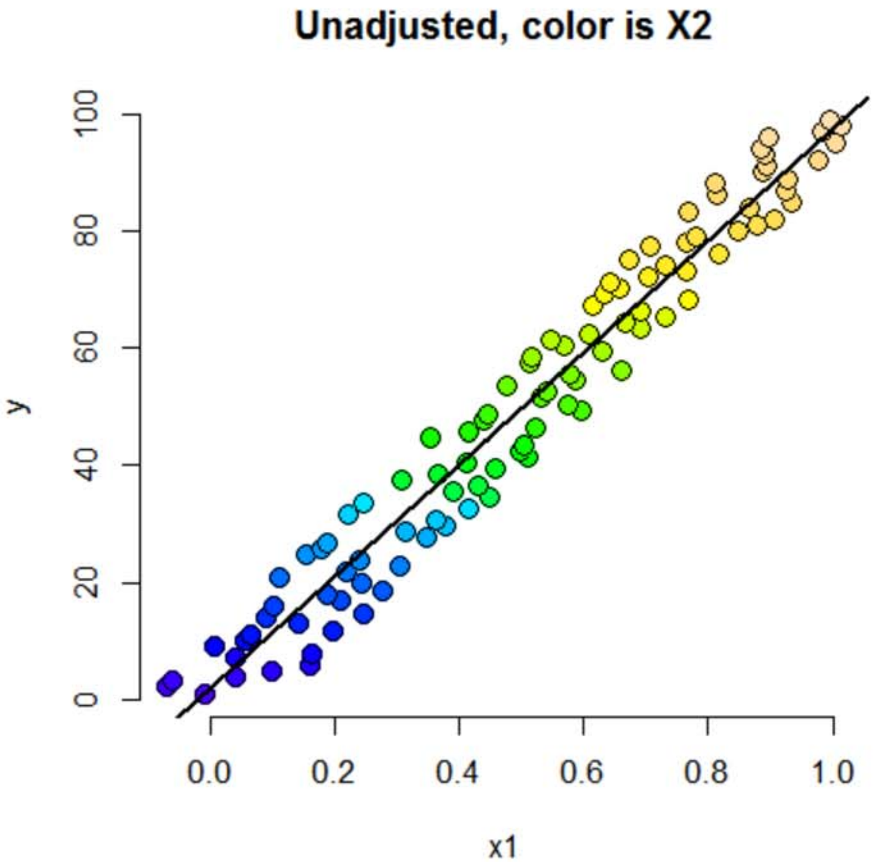
How can adjustment reverse the sign of an effect? Let's try a simulation.

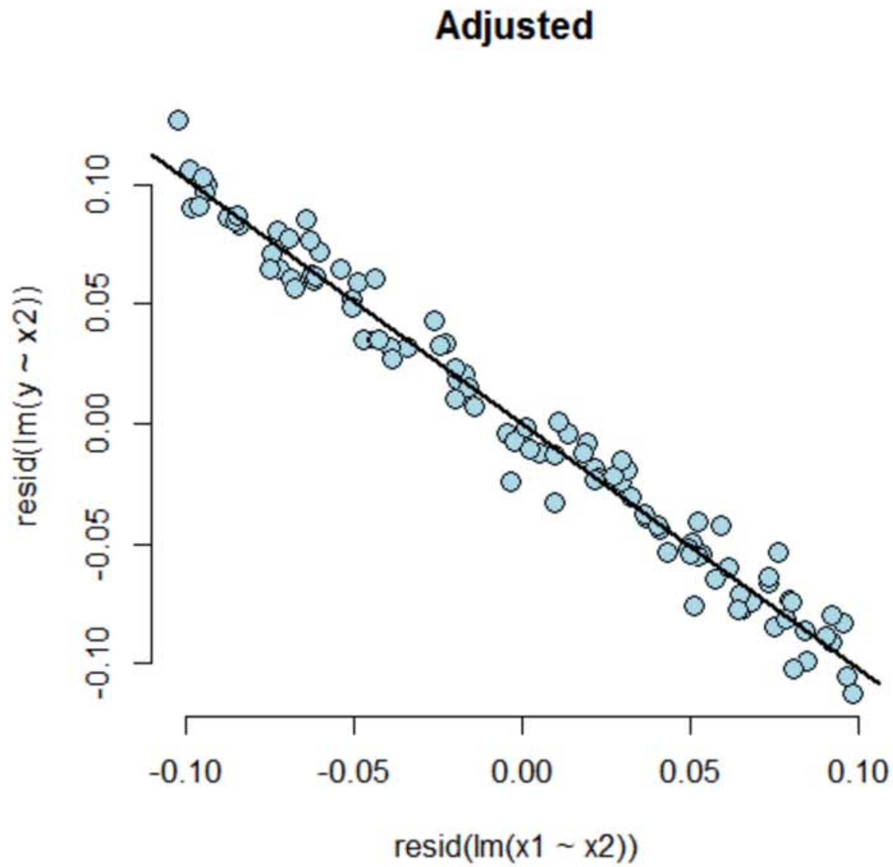
```
n <- 100; x2 <- 1 : n; x1 <- .01 * x2 + runif(n, -.1, .1); y = -x1 + x2 + rnorm(n, sd = .01)
summary(lm(y ~ x1))$coef
```

```
             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
y
x1
0
20
40
60
80
100
0.0
0.2
0.4
0.6
0.8
1.0


Adjusted
resid(lm(y ~ x2))
resid(lm(x1 ~ x2))
-0.10
-0.05
0.00
0.05
0.10

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

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

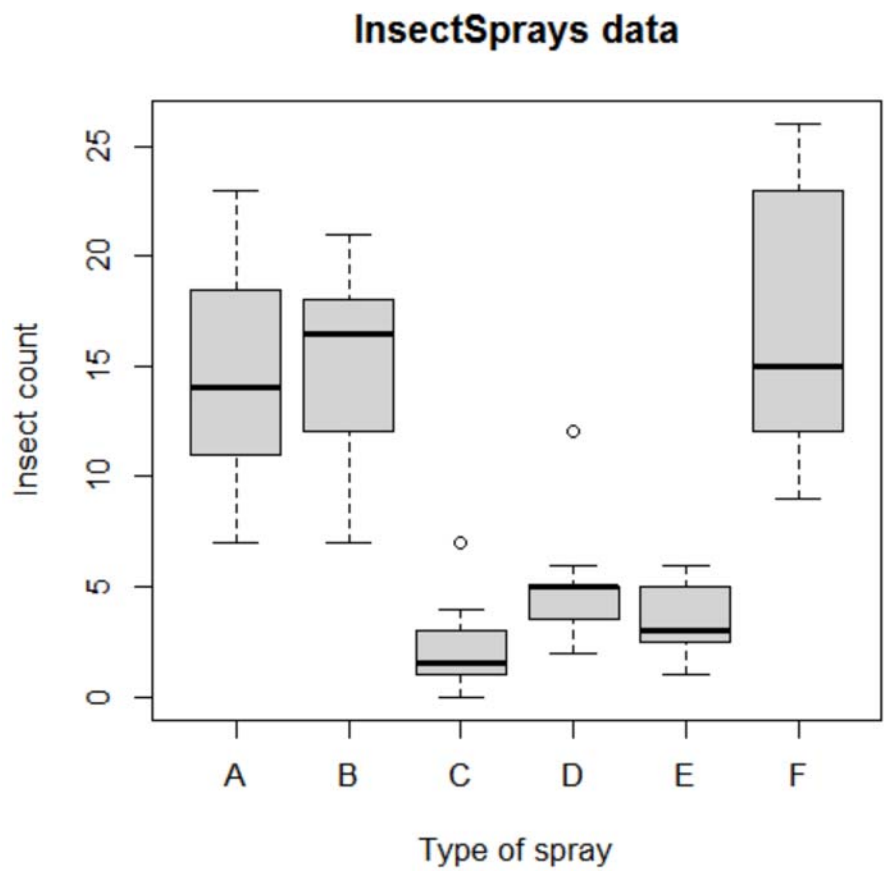
  where each X_{i1} is binary so that it is a 1 if measurement i is in a group and 0 otherwise. (Treated versus not in a clinical trial, for example.)

- 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.
- β_1 is interpretted 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 $E[Y_i] = \beta_0$.
- β_1 compares Republicans to Independents.
- β_2 compares Democrats to Independents.
- $\beta_1 - \beta_2$ compares Republicans to Democrats.
- (Choice of reference category changes the interpretation.)

Insect Sprays



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.
  - Emprirical 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

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

```
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(Count + 1)
- Also, we'll cover Poisson GLMs for fitting count data.

Example - Millenium Development Goal 1

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:*

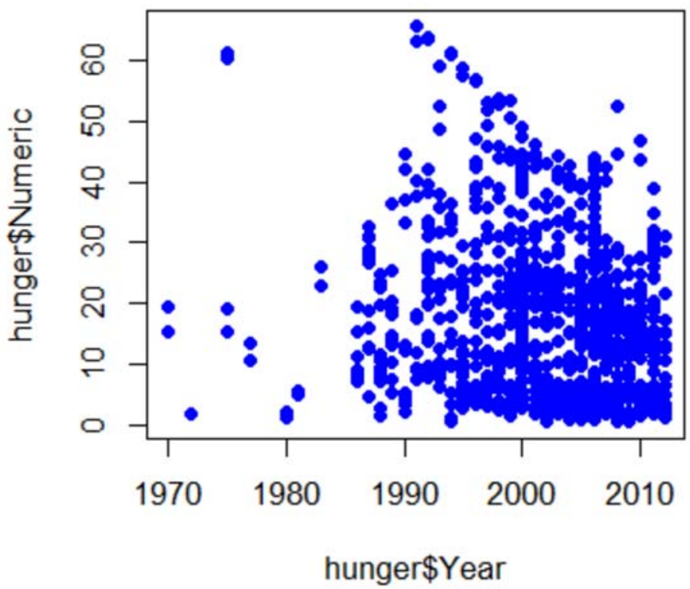
WHO childhood hunger data

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

```
                               Indicator Data.Source PUBLISH.STATES Year              WHO.region
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
        Country    Sex Display.Value Numeric Low High Comments
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

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

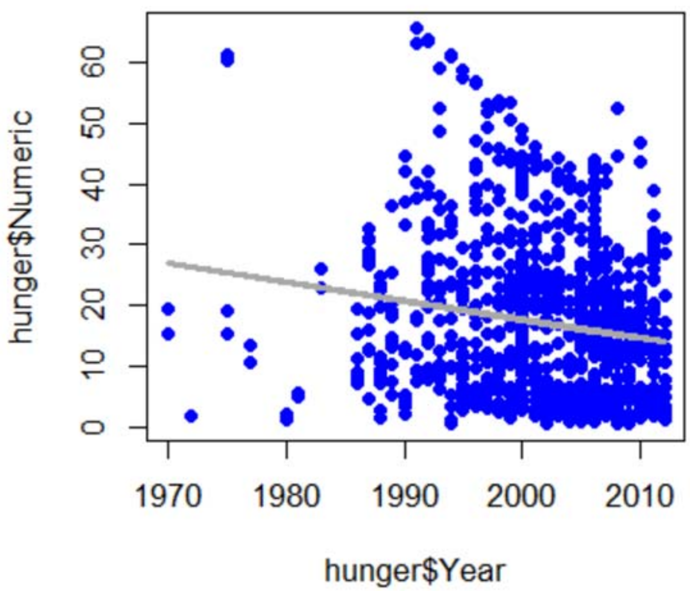
b_0 = percent hungry at Year 0

b_1 = decrease in percent hungry per year

e_i = everything we didn't measure

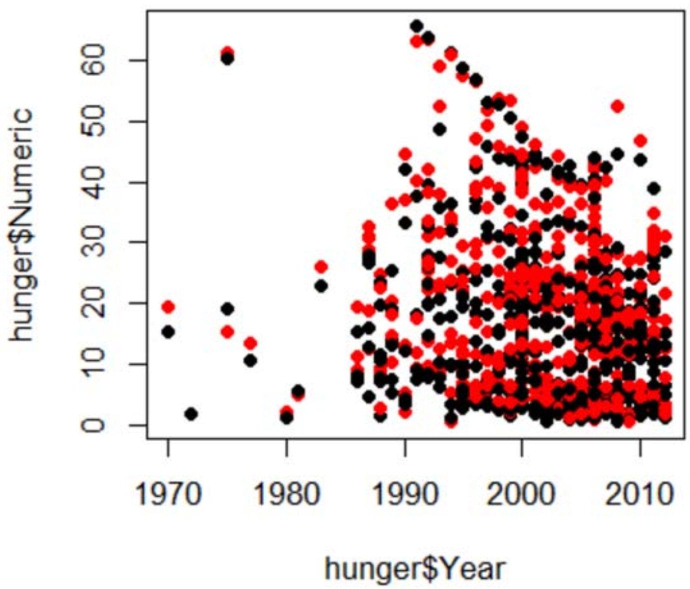
Add the linear model

```
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

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

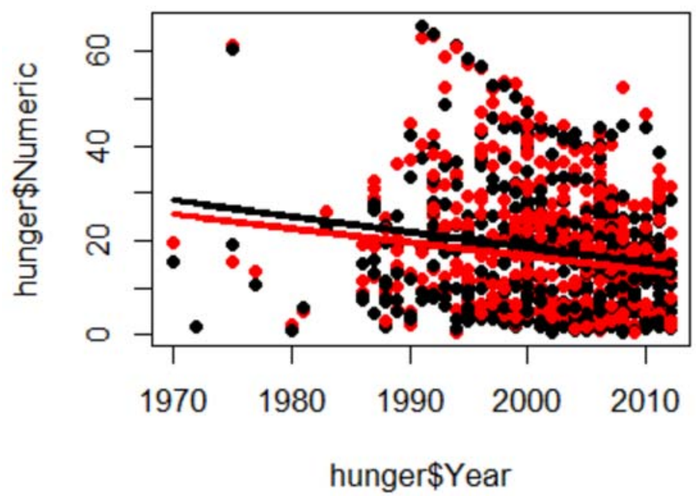
bm_0 = percent of boys hungry at Year 0

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

$$Hu_i = b_0 + b_1 1(Sex_i = " Male ") + b_2 Y_i + e_i^*$$

b_0 - percent hungry at year zero for females

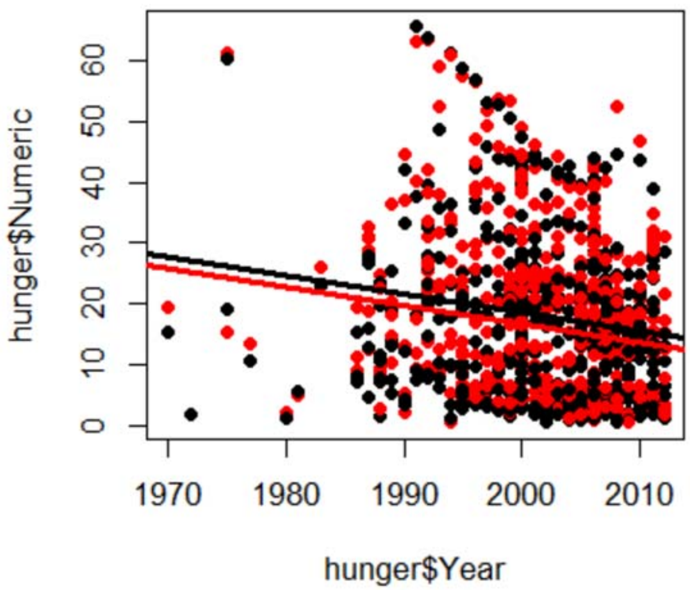
$b_0 + b_1$ - percent hungry at year zero for males

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)

$$Hu_i = b_0 + b_1 1(Sex_i =" Male ") + b_2 Y_i + b_3 1(Sex_i =" Male ") \times Y_i + e_i^+$$

b_0 - percent hungry at year zero for females

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

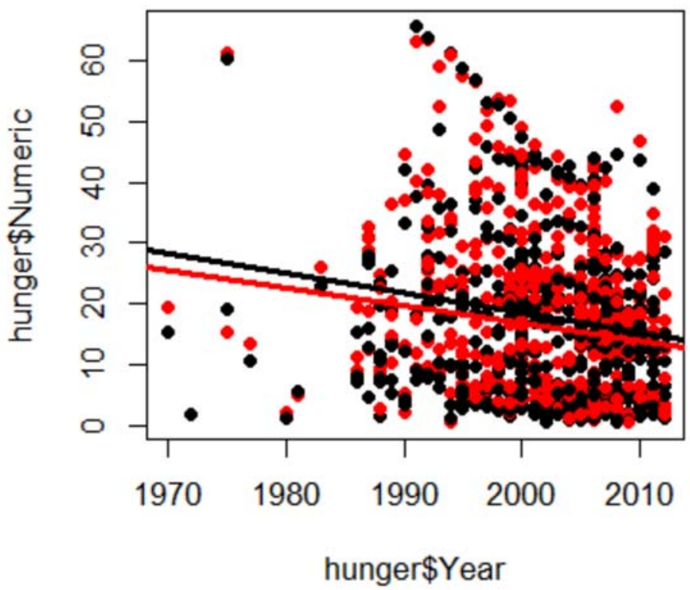
b_2 - change in percent hungry (females) in one year

$b_2 + b_3$ - 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
F-statistic: 10.3 on 3 and 944 DF,  p-value: 1.06e-06
```

Interpretting 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. β_1 is the slope when $x_2 = 0$. Note further that:

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

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

$$= \beta_3$$

Thus, β_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!