Multiple Linear Regression Analysis

Robin Donatello

MATH 456 - Spring 2016

Navbar: [Home] [Schedule] [Data] [Week 4 Overview] [HW Info] [Google Group] [Google Drive]

Multiple Regression

Associated Reading

Afifi: Chapter 7

Aims

- Extend simple linear regression.
- Describe linear relationship between a single continuous Y variable, and several X variables.
- Draw inferences regarding this relationship.
- Predict Y from $X_1, X_2, \ldots, X_P$.

Now it's no longer a 2D regression *line*, but a p dimensional regression plane.

Types of X variables

- Fixed: The levels of X are selected in advance with the intent to measure the affect on an outcome Y.
- Variable: Random sample of individuals from the population is taken and X and Y are measured on each individual.
- X's can be continuous or discrete (categorical)
- X's can be transformations of other X's, e.g., polynomial regression

Mathematical Model

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \epsilon_i$$

The Gauss-Markov assumptions still hold here. Recall these concern the set of error random variables, ϵ_i:

- They have mean zero
- They are homoscedastic, that is all have the same finite variance: $Var(\epsilon_i) = \sigma^2 < \infty$
- Distinct error terms are uncorrelated: (Independent) $\text{Cov}(\epsilon_i, \epsilon_j) = 0, \forall i \neq j$.

The regression model relates y to a function of $\mathbf{X}$ and β, where $\mathbf{X}$ is a nxp matrix of p covariates on n observations and β is a length p vector of regression coefficients. In matrix notation this looks like:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

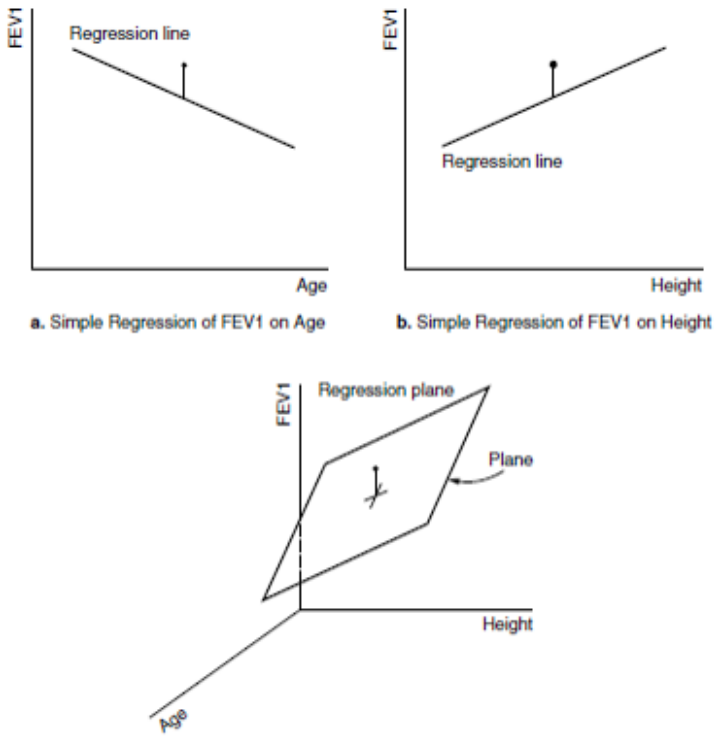


Figure 7.1: *Hypothetical Representation of Simple and Multiple Regression Equations of FEV1 on Age and Height*

Figure 1: Figure 7.1

Parameter Estimation

The goal of regression analysis is to minimize the residual error. That is, to minimize the difference between the value of the dependent variable predicted by the model and the true value of the dependent variable.

$$\epsilon_i = \hat{y}_i - y_i$$

The method of least squares accomplishes this by finding parameter estimates β_0 and β_1 that minimized the sum of the squared residuals:

$$\sum_{i=1}^{n} \epsilon_i$$

For simple linear regression these are found to be

$$\hat{\beta}_0 = \bar{y} * \hat{\beta}_1 \bar{x} \quad \text{and} \quad \hat{\beta}_1 = r\frac{s_y}{s_x}$$

For multiple linear regression the function to minimize is

$$\sum_{i=1}^{n} |y_i - \sum_{i=1}^{p} X_{ij}\beta_j|^2$$

Or in matrix notation

$$||\mathbf{y} - \mathbf{X}\beta||^2$$

The details of methods to solve these minimization functions are left to a course in mathematical statistics, however we will return to this notation.

Continued Example: Lung Function

In Chapter 6 the data for fathers from the lung function data set were analyzed. These data fit the variable-X case. Height was used as the X variable in order to predict `FEV`.

```
fev <- read.delim("C:/GitHub/MATH456/data/Lung_020716.txt", sep="\t", header=TRUE)
summary(lm(FFEV1 ~ FHEIGHT , data=fev))
```

```
##
## Call:
## lm(formula = FFEV1 ~ FHEIGHT, data = fev)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.56688 -0.35290  0.04365  0.34149  1.42555
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.08670    1.15198  -3.548 0.000521 ***
## FHEIGHT      0.11811    0.01662   7.106 4.68e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.5638 on 148 degrees of freedom
## Multiple R-squared:  0.2544, Adjusted R-squared:  0.2494
## F-statistic:  50.5 on 1 and 148 DF,  p-value: 4.677e-11
```

```
round(confint(lm(FFEV1 ~ FHEIGHT , data=fev)),2)
```

```
##             2.5 % 97.5 %
## (Intercept) -6.36  -1.81
## FHEIGHT      0.09   0.15
```

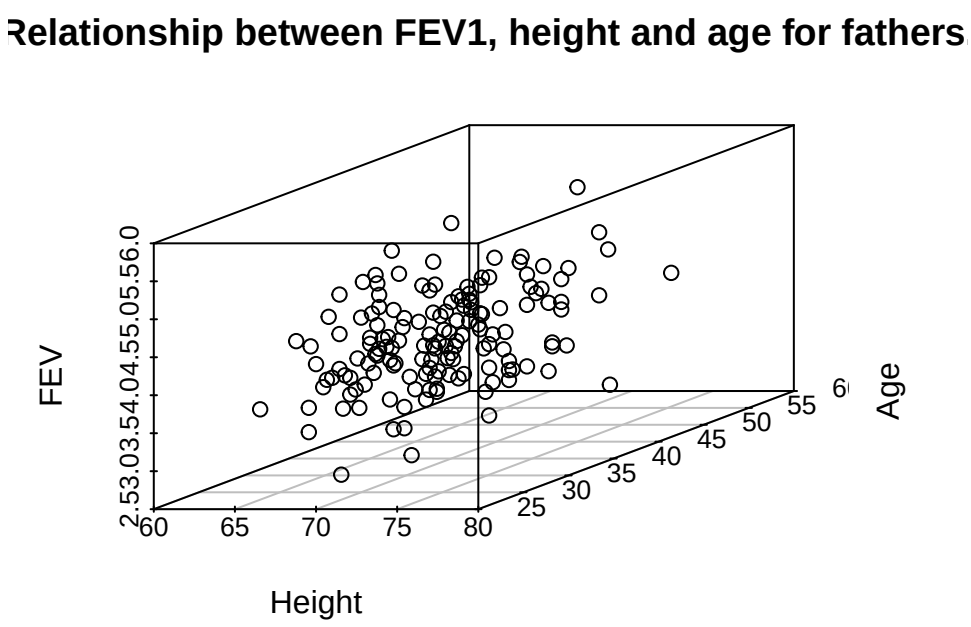
This model concludes that FEV1 in fathers significantly increases by 0.12 (95% CI: 0.09, 0.15) liters per additional inch in height (p<.0001). Looking at the multiple R^2 (correlation of determination), this simple model explains 25% of the variance seen in the outcome y.

However, FEV tends to decrease with age for adults, so we should be able to predict it better if we use both height and age as independent variables in a multiple regression equation.

First let's see different ways to graphically explore the relationship between three characteristics simultaneously.

3D scatterplots

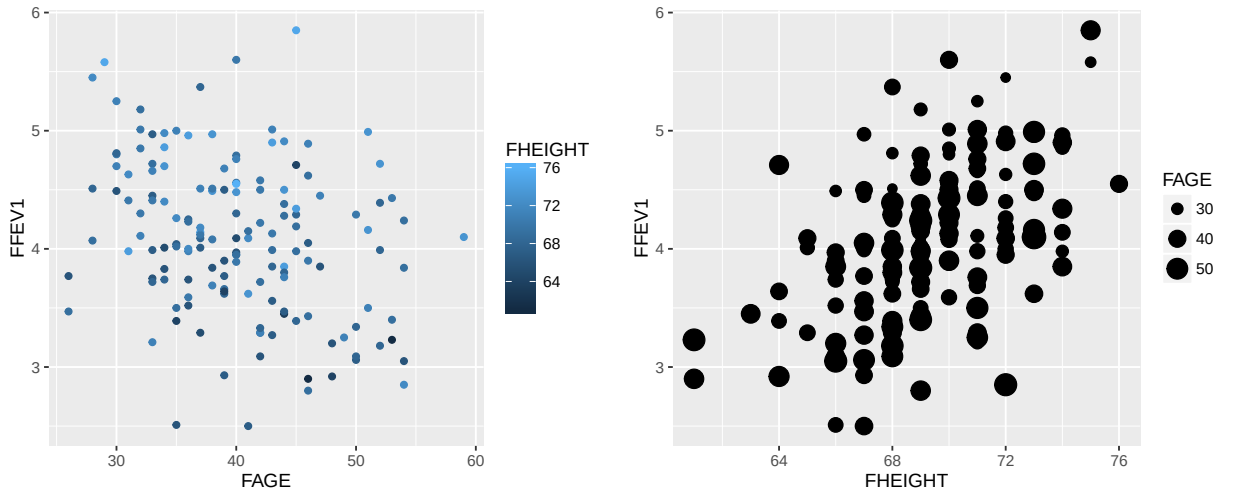
```
library(scatterplot3d)
scatterplot3d(x=fev$FHEIGHT, y=fev$FAGE, z=fev$FFEV1,
              xlab="Height", ylab="Age", zlab="FEV",
              main="Relationship between FEV1, height and age for fathers.")
```



See http://www.statmethods.net/graphs/scatterplot.html for two simple ways to create interactive, spinning 3D scatterplots.

Controlling the color, or size of points using the third characteristic

```
library(gridExtra)
a <- qplot(y=FFEV1, x=FAGE, color=FHEIGHT, data=fev)
b <- qplot(y=FFEV1, x=FHEIGHT, size=FAGE, data=fev)
grid.arrange(a, b, ncol=2)
```



The scatterplot of FEV against age demonstrates the decreasing trend of FEV as age increases, and the increasing trend of FEV as height increases. The third color however is pretty scattered across the plot. There is no obvious trend observed.

- What direction do you expect the slope coefficient for age to be? For height?

Model fitting

Simple Linear Regression

Let's examine the bivarate relationship of FEV1 (forced expiratory volume in 1 minute) for fathers `FFEV1` on their age `FAGE`.

```
summary(lm(FFEV1 ~ FAGE, data=fev))
```

```
##
## Call:
## lm(formula = FFEV1 ~ FAGE, data = fev)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.73332 -0.46620 -0.01332  0.42572  1.89899
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.266374   0.300590  17.520  < 2e-16 ***
## FAGE        -0.029230   0.007382  -3.959 0.000116 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6209 on 148 degrees of freedom
## Multiple R-squared:  0.09578,    Adjusted R-squared:  0.08967
## F-statistic: 15.68 on 1 and 148 DF,  p-value: 0.0001163
```

For every one year older the father gets, his FEV1 significantly decreases by 0.03 liters (p = .00001).

```
summary(lm(FFEV1 ~ FHEIGHT, data=fev))
```

```
##
## Call:
## lm(formula = FFEV1 ~ FHEIGHT, data = fev)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.56688 -0.35290  0.04365  0.34149  1.42555
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.08670    1.15198  -3.548 0.000521 ***
## FHEIGHT      0.11811    0.01662   7.106 4.68e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5638 on 148 degrees of freedom
## Multiple R-squared:  0.2544, Adjusted R-squared:  0.2494
## F-statistic:  50.5 on 1 and 148 DF,  p-value: 4.677e-11
```

For every inch taller a father is, his FEV1 significantly increases by 0.11 liters (p < .0001).

Multiple Linear Regression

Fitting a regression model in R with more than 1 predictor is trivial. Just add each variable to the right hand side of the model notation connected with a +.

```
mv_model <- lm(FFEV1 ~ FAGE + FHEIGHT, data=fev)
summary(mv_model)
```

```
##
## Call:
## lm(formula = FFEV1 ~ FAGE + FHEIGHT, data = fev)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.34708 -0.34142  0.00917  0.37174  1.41853
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.760747   1.137746  -2.427   0.0165 *
## FAGE        -0.026639   0.006369  -4.183 4.93e-05 ***
## FHEIGHT      0.114397   0.015789   7.245 2.25e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5348 on 147 degrees of freedom
## Multiple R-squared:  0.3337, Adjusted R-squared:  0.3247
## F-statistic: 36.81 on 2 and 147 DF,  p-value: 1.094e-13
```

A father who is one year older is expected to have a FEV value 0.03 liters less than another father of the same height (p<.0001).

A father who is the same age as another father is expected to have a FEV value of 0.11 liter greater than another father of the same age who is one inch shorter (p<.0001).

For the model that includes age, the coefficient for height is now 0.11, which is interpreted as the rate of change of FEV1 as a function of height **after adjusting for age**. This is also called the **partial regression coefficient** of FEV1 on height after adjusting for age.

Both height and age are significantly associated with FEV in fathers (p<.0001 each).

ANOVA for regression

For a global test to see whether or not the regression model is helpful in predicting the values of y, we can use an ANOVA. This is the same as testing that all $\beta_j, j = 1, \ldots, p$ are all equal to 0. Let's look at what we get if we wrap the ANOVA function, `aov()` around the linear model results.

```
summary(aov(mv_model))
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## FAGE          1   6.04   6.044   21.13 9.17e-06 ***
## FHEIGHT       1  15.01  15.013   52.49 2.25e-11 ***
## Residuals   147  42.04   0.286
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We get the sums of squares (SS) for each predictor individually, not combined into a SS for regression and a SS residuals. So where do we find this global test that this model is better than using no predictors at all? At the very last line in the summary of the linear model results.

```
summary(mv_model)
```

```
##
## Call:
## lm(formula = FFEV1 ~ FAGE + FHEIGHT, data = fev)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.34708 -0.34142  0.00917  0.37174  1.41853
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.760747   1.137746  -2.427   0.0165 *
## FAGE        -0.026639   0.006369  -4.183 4.93e-05 ***
## FHEIGHT      0.114397   0.015789   7.245 2.25e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5348 on 147 degrees of freedom
## Multiple R-squared:  0.3337, Adjusted R-squared:  0.3247
## F-statistic: 36.81 on 2 and 147 DF,  p-value: 1.094e-13
```

Predictions

What is the predicted FEV1 for a 30-year-old male whose height was 70 inches? There are multiple ways to evaluate this calculation.

1. Manual calculation by plugging in the value of each variable individually

```
betas <- mv_model$coefficients
betas
```

```
## (Intercept)        FAGE     FHEIGHT
## -2.76074686 -0.02663934  0.11439704
```

```
betas[1] + betas[2]*30 + betas[3]*70
```

```
## (Intercept)
##    4.447866
```

Interpretation: We would expect a a 30-year-old male whose height was 70 inches to have an FEV1 value of 4.45 liters.

2. In matrix notation a new vector `x.new` is created and multiplied by the vector of coefficients.

```
x.new <- c(1, 30, 70)
x.new %*% betas
```

```
##          [,1]
## [1,] 4.447866
```

3. Using the `predict()` function. This requires the `newdat` input to be a a `data.frame` object.

```
x.pred <- data.frame(cbind(FAGE = 30, FHEIGHT = 70))
predict(mv_model, newdata=x.pred)
```

```
##        1
## 4.447866
```

Compare this value to the single predictor equation with just height:

```
slr.mod <- lm(FFEV1 ~ FHEIGHT, data=fev)
predict(slr.mod, newdata = data.frame(FHEIGHT=70))
```

```
##        1
## 4.180665
```

This value of 4.18 is the rate of change of FEV1 for fathers as a function of height when no other variables are taken into account.

Comparing coefficients

- Problem: Values of $\hat{\beta}_j$ are NOT directly comparable for any $j = 1, \ldots, p$.
- Solution: Standardized coefficients:

$$\hat{\beta}_j = \hat{\beta}_j \frac{s_{x_j}}{s_y}$$

- These coefficients are the same as if you standardized the data X and Y prior to conducting the regression.
- The larger the magnitude of the standardized coefficient, the more X_i directly contributes to the prediction of y.
- The standardized slope coefficient is the amount of change in the mean of the standardized y values when the value of X is increased by one standard deviation, keeping the other X variables constant.

Interlude: Reporting regression coefficients using tables in Markdown

There are not many options for making readable tables in Markdown. For better control you would be best suited by switching to LaTeX and Sweave (both can be done in R Studio). Here is a reference webpage that shows three basic options. http://kbroman.org/knitr_knutshell/pages/figs_tables.html

Here is an example using `xtable`. I first round the results of the table to 2 digits, then format the p-value to read <.0001 instead of a super small digit. Then I wrap `xtable()` around that output.

```
library(xtable)
options(xtable.comment=FALSE)
coeff.out <- round(summary(mv_model)$coef, 2)
coeff.out[,4] <- ifelse(coeff.out[,4]<.0001, "<.0001", coeff.out[,4])
tab <- xtable(coeff.out)
names(tab)[3:4] <- c("t", "p-value")
#print(tab, type="html")
print(tab)
```

	Estimate	Std. Error	t	p-value
(Intercept)	-2.76	1.14	-2.43	0.02
FAGE	-0.03	0.01	-4.18	<.0001
FHEIGHT	0.11	0.02	7.25	<.0001

R specifics: Rounding has to come first because the `ifelse()` changed the data type in the entire matrix from numeric to character, and once it's a character you can't round. There are ways around this but I am not going to discuss it here.

Estimated Covariance

For the variable-X case, several additional parameters are needed to characterize the full joint distribution $f(x, y)$. A matrix of the estimated covariances between the parameter estimates βj's can be obtained in R by using the `vcov()` function on the model output.

```
round(vcov(mv_model), digits=3)
```

```
##             (Intercept)   FAGE FHEIGHT
## (Intercept)       1.294 -0.002  -0.017
## FAGE             -0.002  0.000   0.000
## FHEIGHT          -0.017  0.000   0.000
```

Note: This is NOT the same as the variance-covariance matrix of the actual data. This can be found using the variance `var()` function. For the sake of example I put a third covariate in the list: `FWEIGHT`.

```
library(dplyr)
x.vars <- fev %>% select(FAGE, FHEIGHT, FWEIGHT, FFEV1) # requires dplyr
round(var(x.vars), digits=3)
```

```
##           FAGE FHEIGHT FWEIGHT  FFEV1
## FAGE    47.472  -1.075  -3.649 -1.388
## FHEIGHT -1.075   7.724  34.695  0.912
## FWEIGHT -3.649  34.695 573.798  2.067
## FFEV1   -1.388   0.912   2.067  0.423
```

- The variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2$ are found down the diagonal.
- The covariances σ_{ij} are found on the off diagonal entries.

- This matrix is symmetric

Similarly the correlation matrix can be found using the `cor()` function.

```
round(cor(x.vars), digits=3)
```

```
##           FAGE FHEIGHT FWEIGHT  FFEV1
```

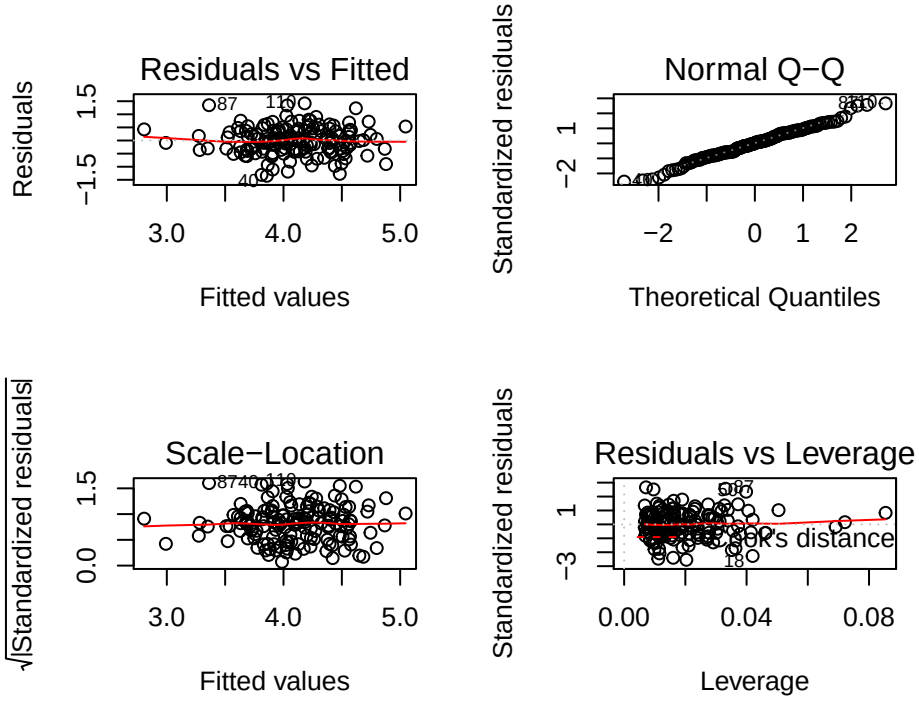
```
## FAGE     1.000  -0.056  -0.022 -0.309
## FHEIGHT -0.056   1.000   0.521  0.504
## FWEIGHT -0.022   0.521   1.000  0.133
## FFEV1   -0.309   0.504   0.133  1.000
```

- The correlation of a variable with itself is always 1, hence the 1's down the diagonal.
- Each entry on the off diagonal is the simple correlation value: ρ_{12}.
- This matrix is also symmetric.

Model Diagnostics

The same set of regression diagnostics can be examined to identify any potential influential points, outliers or other problems with the linear model.

```
par(mfrow=c(2,2))
plot(mv_model)
```



The textbook provides more details about tools to detect outliers and influential points by examining the following measures:

- studentized residuals
- leverage
- DFFITS
- Cooks distance

At least one reference website on how to visualize these measures can be found here: http://www.statmethods.net/stats/rdiagnostics.html

Multicollinearity

- Occurs when some of the X variables are highly intercorrelated.
- Affects estimates and their SE's (p. 143)
- Look at tolerance, and its inverse, the Variance Inflation Factor (VIF)
- Need tolerance < 0.01, or VIF > 100.

```
library(car)
vif(mv_model)
```

```
##     FAGE  FHEIGHT
## 1.003163 1.003163
```

```
tolerance = 1/vif(mv_model)
tolerance
```

```
##      FAGE   FHEIGHT
## 0.9968473 0.9968473
```

- Solution: use variable selection to delete some X variables.
- Alternatively, use Principal Components (Ch. 14)

Interaction Models

Consider a model with only two predictors: X_i and x_j.

$$E(y|x_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

A model of this form is said to be *additive*. If the additive terms for these variables do not completely specify their effects on the dependent variable Y, then interaction of X_1 and X_2 is said to be present.

Interlude: Reshaping the Lung data from wide to long.

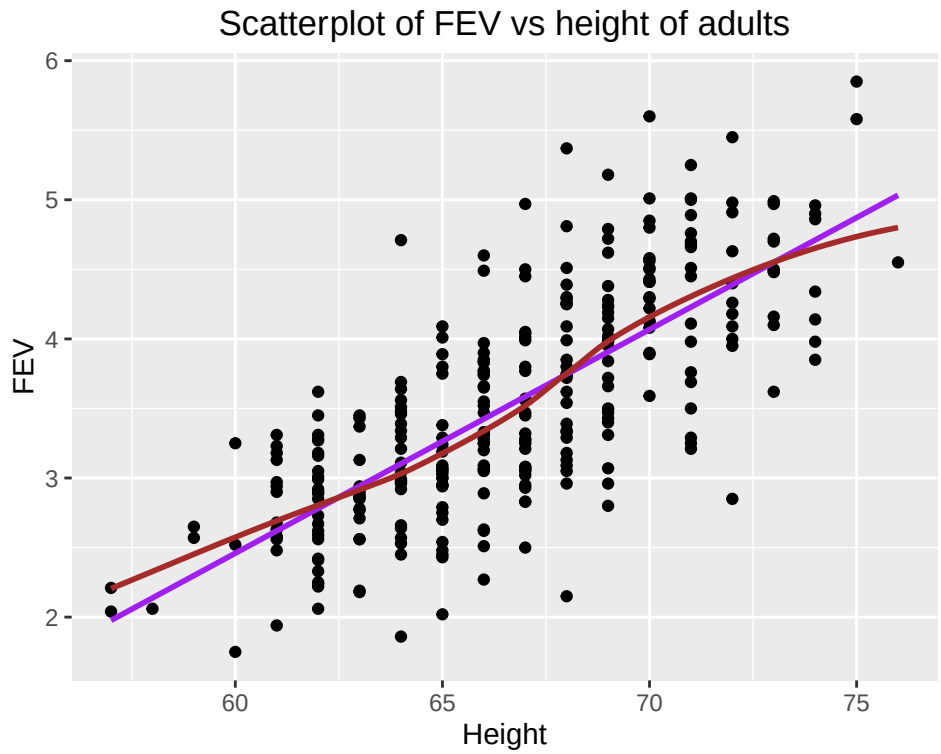
The data on Lung function came to us in *wide* format, with separate variables for mother's and father's FEV1 score (`MFEV1` and `FFEV`). To analyze the effect of gender on FEV, the data need to be in *long* format, with a single variable for `FEV` and a separate variable for gender. The following code chunk demonstrates one method of combining data on height, gender, age and FEV1 for both males and females.

```
fev2 <- data.frame(gender = c(fev$FSEX, fev$MSEX),
                   FEV = c(fev$FFEV1, fev$MFEV1),
                   ht = c(fev$FHEIGHT, fev$MHEIGHT),
                   age = c(fev$FAGE, fev$MAGE))
fev2$gender <- factor(fev2$gender, labels=c("M", "F"))
head(fev2)
```

```
##   gender  FEV ht age
## 1      M 3.23 61  53
## 2      M 3.95 72  40
## 3      M 3.47 69  26
## 4      M 3.74 68  34
## 5      M 2.90 61  46
## 6      M 4.91 72  44
```

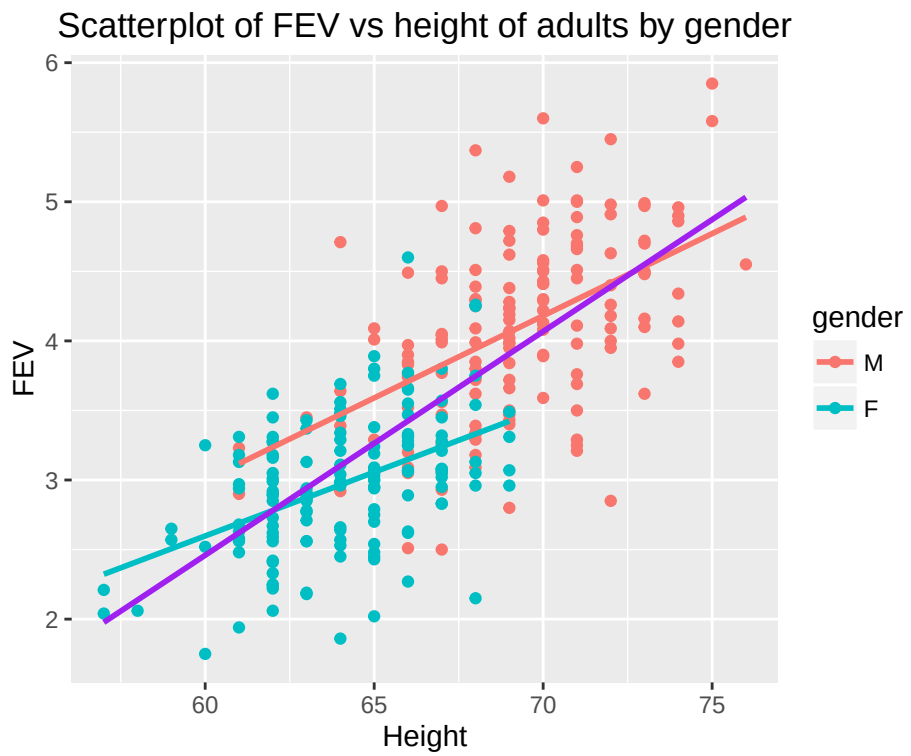
What does the relationship between height and FEV look like for the overall sample?

```
ggplot(data=fev2, aes(x=ht, y=FEV)) + geom_point() + xlab("Height") +
  ggtitle("Scatterplot of FEV vs height of adults") +
  geom_smooth(se=FALSE, method='lm', col="purple") +
  geom_smooth(se=FALSE, col="brown")
```



However if we examine the relationship separately by gender we see a slightly different story.

```
ggplot(data=fev2, aes(x=ht, y=FEV, col=gender)) + geom_point() + xlab("Height") +
  ggtitle("Scatterplot of FEV vs height of adults by gender") +
  geom_smooth(se=FALSE, method='lm') +
  geom_smooth(aes(x=ht, y=FEV), col="purple", se=FALSE, method='lm')
```



If we put numeric summaries to these bivariate relationships we see that the correlation between FEV and height overall is `cor(fev2$FEV, fev2$ht)` $= 0.74$, but for males it is `cor(fev$FFEV, fev$FHEIGHT)` $= 0.5$, and for females the correlation is `cor(fev$MFEV, fev$MHEIGHT)` $= 0.46$.

Conclusion: The relationship between FEV1 and height may depend on gender. These variables are said to **interact** with each other. In other words, gender changes the relationship between height and FEV1.

Consider the linear model of FEV on gender(x_1), height(x_2) and age(x_3) where gender interacts with height.

$$FEV1 \sim \beta_0 + \beta_1 * gender + \beta_2 * height + \beta_3 * age + \beta_4 * gender * height$$

According to the codebook, when `gender = 0` the record is on a male, and when `gender = 1` the record is on a female. The model for males then simplifies to:

$$FEV1 \sim \beta_0 + \beta_2 * height + \beta_3 * age$$

and the model for females would be:

$$FEV1 \sim (\beta_0 + \beta_1) + (\beta_2 + \beta_4) * height + \beta_3 * age$$

Interactions are fit in `R` by simply multiplying `*` the two variables together in the model statement.

```
intx.model <- lm(FEV ~ gender + ht + age + gender*ht, data=fev2)
summary(intx.model)
```

```
##
## Call:
## lm(formula = FEV ~ gender + ht + age + gender * ht, data = fev2)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -1.32892 -0.29985 -0.00487  0.32476  1.41863
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.92254    0.99654  -2.933  0.00362 **
## genderF      0.82968    1.41007   0.588  0.55672
## ht           0.11485    0.01409   8.152 1.03e-14 ***
## age         -0.02339    0.00407  -5.746 2.27e-08 ***
## genderF:ht  -0.02210    0.02121  -1.042  0.29813
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4775 on 295 degrees of freedom
## Multiple R-squared:  0.6507, Adjusted R-squared:  0.646
## F-statistic: 137.4 on 4 and 295 DF,  p-value: < 2.2e-16
```

```
confint(intx.model)
```

```
##                   2.5 %      97.5 %
## (Intercept) -4.88377440 -0.96131560
## genderF     -1.94539206  3.60474683
## ht           0.08712363  0.14257543
## age         -0.03139883 -0.01537858
## genderF:ht  -0.06383574  0.01963119
```

- Using the output from R after fitting the interaction model specified above, write out the regression equation for males and females separately.

The first thing to notice is the coefficient label: `genderF`. Since the variable for gender is a factor variable, is automatically *reference coded* into a new variable that is not present on the data set but only as part of the linear model function. This variable `genderF` is a 1 if the gender on record is female, and 0 otherwise.

The p-value for the interaction term `ht:genderF` is large, and the confidence interval for this parameter covers zero, so there is no indication that an interaction exists. That is, there is not enough reason to believe that gender significantly affects the relationship between height and FEV.

Reminder Just as in a Two-Way ANOVA with an interaction term, the main effects cannot be interpreted directly when there is an interaction in the model. This means you **cannot** interpret the direct effect of height or gender on FEV1 using just β_1 or β_3.

- What is the effect of height on FEV1 for females?

Later on we will talk about how categorical variables with multiple factor levels are reference coded for entry into models.

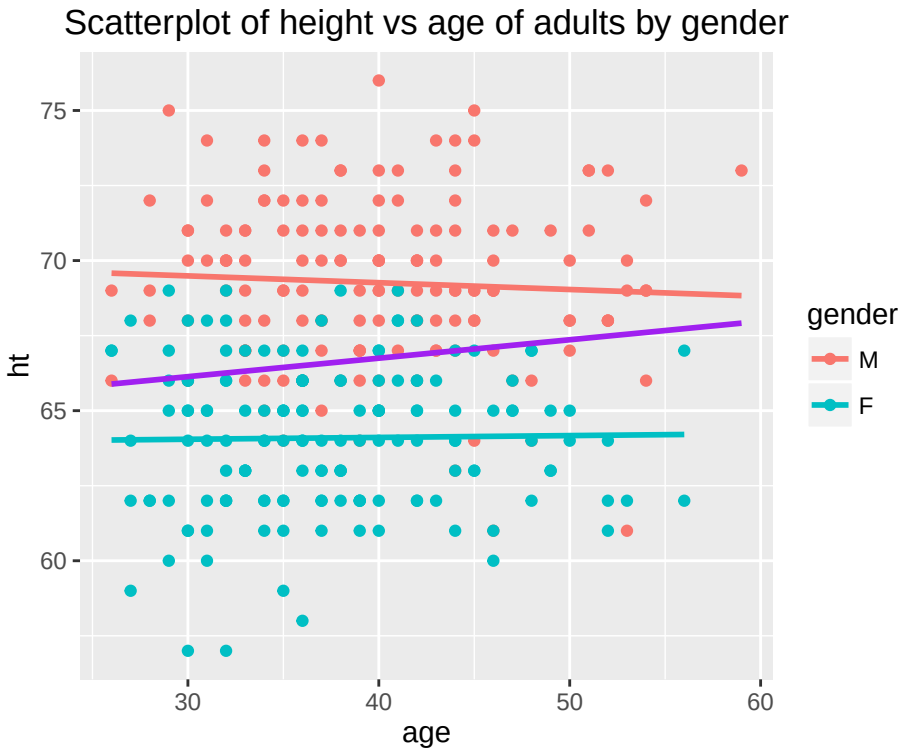
Stratification

Sometimes it is desirable to examine equations for subgroup of the population. Consider the relationship between age, height and FEV by gender. We write the model with the same set of covariates on each strata (gender).

$$FEV1_M \sim \beta_{0M} + \beta_{1M} * height + \beta_{2M} * age$$
$$FEV1_F \sim \beta_{0M} + \beta_{1F} * height + \beta_{2F} * age$$

Example: Within each gender there exists a negative correlation between age and height. However in the combined sample this appears to be a positive correlation.

```
ggplot(data=fev2, aes(x=age, y=ht, col=gender)) + geom_point() +
  ggtitle("Scatterplot of height vs age of adults by gender") +
  geom_smooth(se=FALSE, method='lm') +
  geom_smooth(aes(x=age, y=ht), col="purple", se=FALSE, method='lm')
```



This is similar to Simpsons Paradox, where there are a number of additional examples of this situation including the UC Berkeley gender bias lawsuit.

Since gender affects the relationship between FEV and both height and age, the appropriate model would then be:

$$FEV1 \sim \beta_0 + \beta_1 * gender + \beta_2 * height + \beta_3 * age + \beta_4 * gender * height + \beta_5 * gender * age$$

If we let gender $= 0$ if the record is on a male, and gender $= 1$ if the record is on a female, then the model for males would be:

$$FEV1 \sim \beta_0 + \beta_2 * height + \beta_3 * age$$

and the model for females would be:

$$FEV1 \sim (\beta_0 + \beta_1) + (\beta_2 + \beta_4) * height + (\beta_3 + \beta_5) * age$$

Instead of running the model on the full set of data and then calculating the correct coefficient for each gender we *stratify* the model by fitting the model on each subgroup separately.

```
# subset the data
M <- subset(fev2, gender=="M")
F <- subset(fev2, gender=="F")
# Overall model
overall_model <- lm(FEV ~ age + ht, data=fev2)
# run stratified models
male_model   <- lm(FEV ~ age + ht, data=M)
female_model <- lm(FEV ~ age + ht, data=F)
```

The overall model indicates that FEV decreases with age and increases with height (p<.0001 each)

```
tab <- xtable(summary(overall_model), digits=3)
#print(tab, type="html")
print(tab)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-6.737	0.563	-11.960	0.000
age	-0.019	0.004	-4.186	0.000
ht	0.165	0.008	19.785	0.000

Model on females only:

```
tabf <- xtable(summary(female_model), digits=3)
#print(tabf, type="html")
print(tabf)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.211	0.896	-2.468	0.015
age	-0.020	0.005	-3.963	0.000
ht	0.093	0.014	6.757	0.000

Model on males only:

```
tabm <- xtable(summary(male_model), digits=3)
#print(tabm, type="html")
print(tabm)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.761	1.138	-2.427	0.016
age	-0.027	0.006	-4.183	0.000
ht	0.114	0.016	7.245	0.000

Create a prediction of FEV for two cases, a 30 year old who is 66cm tall, and a 50 year old who is 62cm tall.

```
# create new data frame using the overall mean height
new.data <- data.frame(ht = c(66, 62), age=c(30, 50))
# predict on each model
o.pred <- predict(overall_model, new.data)
m.pred <- predict(male_model, new.data)
f.pred <- predict(female_model, new.data)
pred <- xtable(rbind(o.pred, m.pred, f.pred), digits=3)
#print(pred, type="html")
print(pred)
```

Column number 1 is a prediction of FEV for a 30 year old who is 66cm tall, column number 2 is a prediction of FEV for a 50 year old who is 66cm tall. Notice the overall prediction

	1	2
o.pred	3.586	2.555
m.pred	3.990	3.000
f.pred	3.301	2.531

Even though the equations for males and females look quite similar, the predicted FEV1 for females of the same height and age as a male is expected to be less. The overall prediction is right in between the two estimates.

Note of caution: Stratification implies that the stratifying variable interacts with all other variables.

Testing equality of individual coefficients between groups

To compare the regression coefficients for men and women we could simply compare the sign and magnitude of the standardized regression coefficients.

If an interaction exists, then the two coefficients from the stratified models would be equal.

To test the null hypothesis that the effect of height on FEV is the same across genders, we compute the following test statistic:

$$Z = \frac{\beta_{2M} - \beta_{2F}}{\sqrt{Var(\beta_{2M}) + Var(\beta_{2F})}}$$

This Z statistic follows the standard normal distribution, $\mathcal{N}(0,1)$ and so you can use the `pnorm()` function to calculate the p-value for the test.

$$Z = \frac{0.114397 - 0.092593}{\sqrt{0.015789^2 + 0.013704^2}}$$

```
z = (0.114397 - 0.092593)/(sqrt(0.015789^2 + 0.013704^2))
z
```

```
## [1] 1.042917
```

```
2*(1-pnorm(z))
```

```
## [1] 0.2969868
```

*`pnorm(z)` takes the value of the test statistic as the argument **`z`**, and returns the probability of a random variable being **below** the test statistic. This can be thought of as the area to the **left** under the normal probability distribution curve. The p-value for a statistical test is the probability of observing a test statistic equal to or greater than the one observed. Since our test statistic is positive, we are interested in P(Z>1.04). Thus we want to calculate the area is the **right** under the normal probability distribution. Since the result of `pnorm` gives us the left and we want the right, and the full area under the curve adds up to 1, we simply calculate `1-pnorm()` to find the area to the right. Lastly, since this is a two tailed test we double the tail area to calculate the p-value of the test in question*

With a large p-value of 0.30 there is insufficient evidence to believe that the relationship between FEV and height differs by gender.

Testing using a CI

You can also attempt to test for a difference in slope coefficients by comparing the confidence intervals for the parameters. *Note: I am showing how I wrangled the linear model output into a table that is easily read for your info only. This is not mandatory but it looks nice.*

```
mci <- paste("(", round(confint(male_model)[,1],2),", ",
             round(confint(male_model)[,2],2), ")", sep="")
fci <- paste("(", round(confint(female_model)[,1],2),", ",
             round(confint(female_model)[,2],2), ")", sep="")
out <- cbind(Male = mci, Female = fci)[-1,]
rownames(out) <- c("Age", "Height")
out <- xtable(out)
#print(out, type="html")
print(out)
```

	Male	Female
Age	(-0.04, -0.01)	(-0.03, -0.01)
Height	(0.08, 0.15)	(0.07, 0.12)

- If CIs *do not* overlap then slopes are significantly different from each other.
- Since CIs do overlap, then two slopes *may or may not* be significantly different from each other.

For both age and height the CI's for the slopes overlap a large amount, so I would suspect that there is no significant difference in the coefficients between models. This finding corroborates the formal statistical test done earlier.

What to watch out for

- See cautions for simple regression including violations of assumptions, outliers, influential points
- Need representative sample
- Multicollinearity: coefficient of any one variable can vary widely, depending on what others are included in the model
- Missing values: Even more important here
  - Default method is complete case analysis
  - If any variable in the model has missing data, the entire record is excluded.
- Number of observations in sample should be large enough relative to the number of parameters that are being estimated.
- This includes the variances and covariances of the parameter estimates.

On Your Own

1. Fit the regression model for the fathers using `FFVC` as the dependent variable and age and height as the independent variables. Write the results for this regression model so they would be suitable for inclusion in a report. Include a table of results.
2. Confirm that this F-test in the model results is the correct one to use by manually calculating the F statistic using an ANOVA table. Confirm the degrees of freedom in both the numerator and denominator are correct, as well as the calculation of the p-value. If you are not familiar with the `qf()` function in R to find the probability under the F distribution, here are some helpful resources in addition to your classmates.
   - http://www.r-tutor.com/elementary-statistics/probability-distributions/f-distribution
   - https://www.youtube.com/watch?v=PZiVe5DMJWA
3. Fit a regression model for females using `MFVC` as the dependent variable and age and height as the independent variables. Summarize the results in a tabular form.
4. Test whether the effect of age and height on FVC for males are significantly different than for females.

5. Using the data on births from North Carolina (`NCbirths`), create a model of the weight of the baby at birth in pounds using the mothers age, smoking habit, and the number of hospital visits during pregnancy as dependent variables. Interpret the regression coefficients in context of the problem and include 95% confidence intervals and p-values in your discussion.
6. Find a 95% prediction interval for a 30-year-old smoking mother with 16 visits to the doctor during her pregnancy.
7. Test for an interaction between smoking habit and the mothers age. Include a plot similar to the one shown in the lecture notes to support your findings.
8. Using the Parental HIV data, generate a variable that represents the sum of the variables describing the neighborhood where the adolescent lives (`NGHB1-NGHB11`). Is the age at which adolescents start smoking different for girls compared to boys, after adjusting for the score describing the neighborhood?

Categorical Predictors

Associated Reading

Afifi: Chapter 9.3, Harrel Ch 2

Factor variable coding

- Most commonly known as "Dummy coding"
- Better used term: Indicator variable
- Math notation: **I(gender == "Female")**.
- A.k.a reference coding
- For a nominal X with K categories, define K indicator variables.
  - Choose a reference (referent) category:
  - Leave it out
  - Use remaining K-1 in the regression.
  - Often, the largest category is chosen as the reference category.

Example: Religion against income and depression

Consider a log-linear model for the effect of marital status (X_2) on log income while controlling for age(X_1). This is called a log-linear model because the outcome has been log transformed.

$$log(Y_i) = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2$$

Forget why we log-transformed income? Re-visit the notes on Data Screening and Transformations http://norcalbiostat.github.io/MATH456/notes/lec01_data_prep.html

```
dep <- read.table("C:/GitHub/MATH456/data/Depress_041616.txt", sep="\t", header=TRUE)
names(dep) <- tolower(names(dep)) # I hate all captal variable names
levels(dep$marital)
```

```
## [1] "Divorced"      "Married"       "Never Married" "Separated"
## [5] "Widowed"
```

Marital status has 5 levels, so we would need 4 indicator variables. R always uses the first level of a factor variable as the reference level.

- Let $x_2 = 1$ when `marital='Married'`, and 0 otherwise,
- let $x_3 = 1$ when `marital='Never Married'`, and 0 otherwise,
- let $x_4 = 1$ when `marital='Separated'`, and 0 otherwise,
- let $x_5 = 1$ when `marital='Widowed'`, and 0 otherwise.

The mathematical model would look like:

$$log(Y)|X \sim \beta_0 + \beta_1 * x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$$

Two levels of interpretation here.

1. The outcome is log transformed, so the interpretation has to be back-transformed.
2. The coefficients for the other levels of the categorical variable are in *comparison* to the reference level.

Interpretation of log-linear models Calculate the change in Y that corresponds to a one unit change in x_1. Since marital status is remaining constant, I will exclude it from the calculations below to save space and not to detract from the main point.

Write each equation down

$$log(Y)|x_1 = \beta_0 + \beta_1 x_1$$
$$log(Y)|(x_1 + 1) = \beta_0 + \beta_1 (x_1 + 1)$$

Find the difference

$$(log(Y)|x_1) - (log(Y)|(x_1 + 1)) = (\beta_0 + \beta_1 x_1) - (\beta_0 + \beta_1 (x_1 + 1))$$

and simplify.

$$log(\frac{Y|x_1}{Y|x_1 + 1}) = \beta_1$$
$$\frac{Y|x_1}{Y|x_1 + 1} = e^{\beta_1}$$

Each 1-unit increase in x_j multiplies the expected value of Y by $e^{\hat{\beta}_j}$.

Interpretation: $100\hat{\beta}_j$ is the expected **percentage** change in Y for a unit increase in x_j.

The nice thing about factor variables in R, is that the appropriate indicator variables are automatically created for you by the linear model (`lm()`) function.

```
summary(lm(log(income) ~ age + marital,data=dep))
```

```
##
## Call:
## lm(formula = log(income) ~ age + marital, data = dep)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.62643 -0.46829  0.01535  0.45280  1.48175
##
```

```
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             3.065696   0.166760  18.384  < 2e-16 ***
## age                    -0.009043   0.002919  -3.098 0.002143 **
## maritalMarried          0.416653   0.124698   3.341 0.000944 ***
## maritalNever Married -0.183156   0.141354  -1.296 0.196109
## maritalSeparated     -0.394544   0.223431  -1.766 0.078482 .
## maritalWidowed       -0.278352   0.173159  -1.607 0.109042
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7059 on 288 degrees of freedom
## Multiple R-squared:  0.1971, Adjusted R-squared:  0.1832
## F-statistic: 14.14 on 5 and 288 DF,  p-value: 2.236e-12
```

- For every year older, a persons income decreases by 1%. (`exp(-0.009)` = 0.99)
- Married individuals have a 52% higher income compared to those who are divorced. (`exp(-0.417)` = 1.52)
- Those who have never been married have 16% lower income compared to those who are divorced. (`exp(-0.183)` = 0.83)
- Separated individuals have 32% lower income compared to those who are divorced. (`exp(-0.394)` = 0.67)
- Widowed individuals have 24% lower income compared to those who are divorced. (`exp(-0.278)` = 0.76)

Other references on how to interpret regression parameters when they have been log transformed:

- http://www.ats.ucla.edu/stat/mult_pkg/faq/general/log_transformed_regression.htm
- http://www.kenbenoit.net/courses/ME104/logmodels2.pdf

On Your Own

On Your Own

Create a model to analyze the relationship of education status to depression level as measured by CESD after controlling for age. Combine all education levels below a HS graduate into one reference category called "Up to HS" prior to analysis.

This is a seemingly simple request, but there are a lot of steps you must do to correctly analyze this question.

1. Ensure that you are using the analyzable version of the depression data set. It may be helpful to confirm that your recodes are correct by comparing your data management code file to mine dm_depress located on our course website.
2. Reference your Ch3 homework (or the solutions) if you need help collapsing educational categories.
3. Ensure that R is treating "Up to HS" as the reference category for education level. If it is not, use the `levels` argument of the `factor()` function to reorder your factor levels. This is also presented in the Ch3 solutions.
4. Consider a transformation of `CESD`. Explain and justify using graphical measures why you chose to, or chose not to, transform CESD prior to modeling.
5. Check the model fit by examining the residuals to see if the assumption that $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is upheld.
6. Identify any potential outliers. Explain why you think they are outliers. Examine their standardized residuals and leverage values.If any seem to stand out or have high values for either measure, exclude them from the analysis and re-run the model.
7. Once you have finalized your model, interpret ALL coefficients in context of the problem. State if any are significantly predictive of the outcome, provide p-values in your conclusion.

8. Does this model do well at all in predicting CESD? Answer this question using both the coefficient of determination and the ANOVA test of overall global fit (testing that all β's are 0)

Splines & other non-linear terms

Associated Reading

- Afifi Section 9.4
- Harrell 2.4.3, pg 39 http://biostat.mc.vanderbilt.edu/tmp/course.pdf
- Harrell ch2 from second edition pdf in shared GDrive.
- https://www.youtube.com/watch?v=o_d4hmKhmsQ
- http://www.r-bloggers.com/thats-smooth/

Example 1: Simulated data.

Example data pulled from _http://faculty.washington.edu/heagerty/Courses/b571/homework/spline-tutorial.q_

Suppose we have a predictor that takes the values 1:24

```
x <- c(1:24)
```

and there is an outcome variable that is predicted by the variable X, but in some non-linear fashion:

```
mu <- 10 + 5 * sin( x * pi / 24 )  - 2 * cos( (x-6)*4/24 )
```

But there is always some amount of error associated with real data.

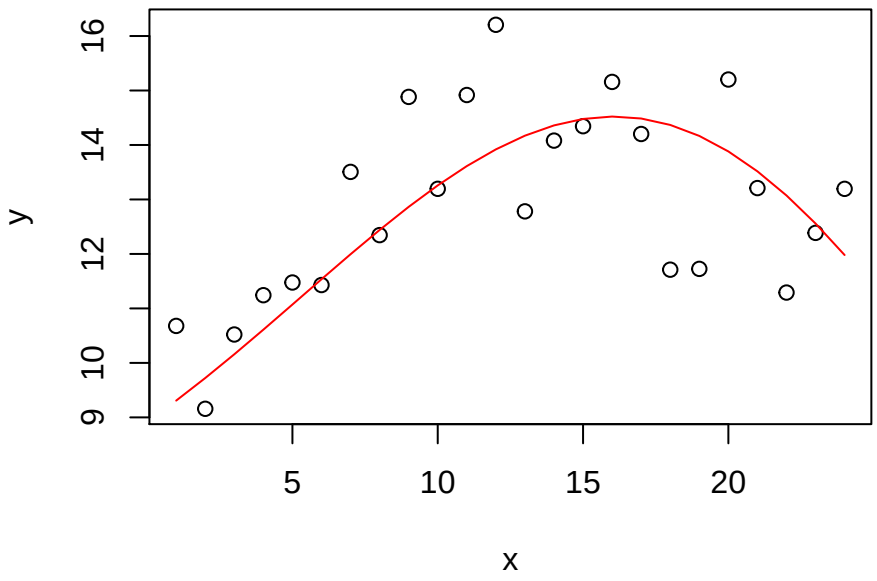
```
set.seed(42)
eee <- rnorm(length(mu))
```

So our simulated data then is the true trend + the noise.

```
y <- mu + eee
```

Let's look at the data, and the real mean trend without the random noise.

```
plot(y~x)
lines(x, mu, col="red" )
```



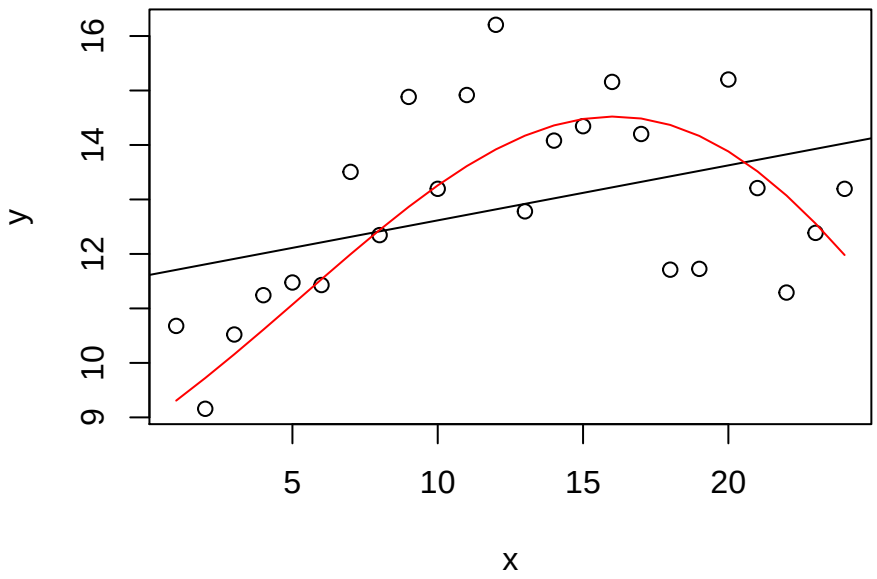
Let's look at ways to fit a model to this data.

Linear

Ignore the trend and fit a linear model.

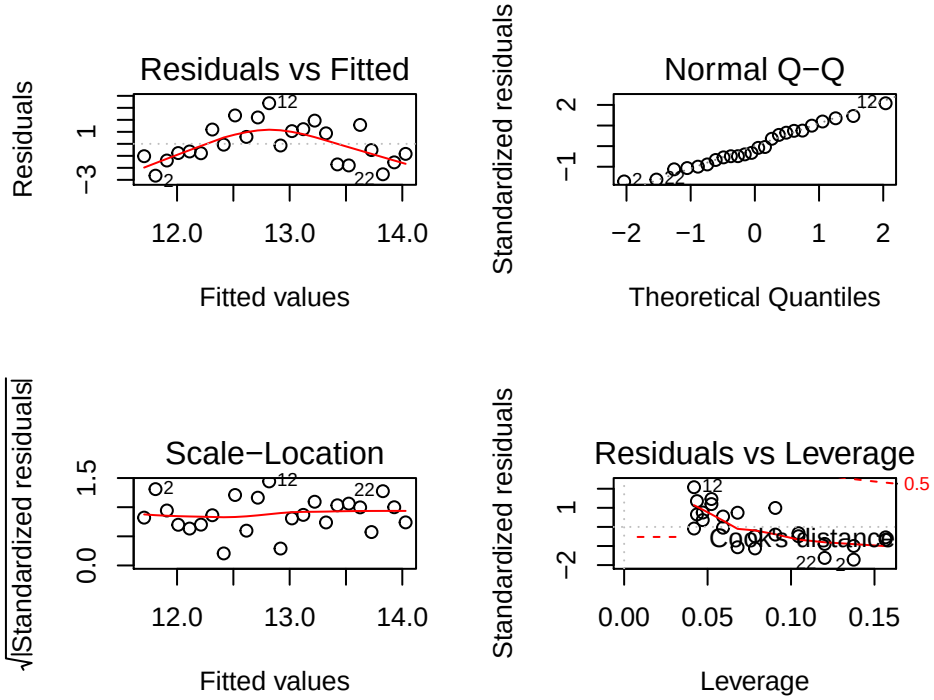
$$E(Y|X) = \beta_0 + \beta_1 X$$

```
fit.slr <- lm(y~x)
plot(y~x)
abline(fit.slr)
lines(x, mu, col="red" )
```



Undoubtedly not a good fit. Examining the residuals shows the non-constant variance clearly.

```
par(mfrow=c(2,2))
plot(fit.slr)
```



Piecewise linear splines

We allow the x axis to be divided into intervals, with a linear model fit within each interval. The breakpoints between intervals are called *knots*. This is where you are allowing the slope of the line to change. For example to break the x-axis into three sections we would use 2 knots. The model would look like.

$$E(Y|X) = \beta_0 + \beta_1 X + \beta_2 (X-a)_+ + \beta_3 (X-b)_+$$

where $(u)_+$ contains the value of u when u is positive, and 0 otherwise.

Let's put knots at 6, 12, and 18.

```
x6 <- (x-6)
x6[ x6<0 ] <- 0

x12 <- (x-12)
x12[ x12<0 ] <- 0

x18 <- (x-18)
x18[ x18<0 ] <- 0
```

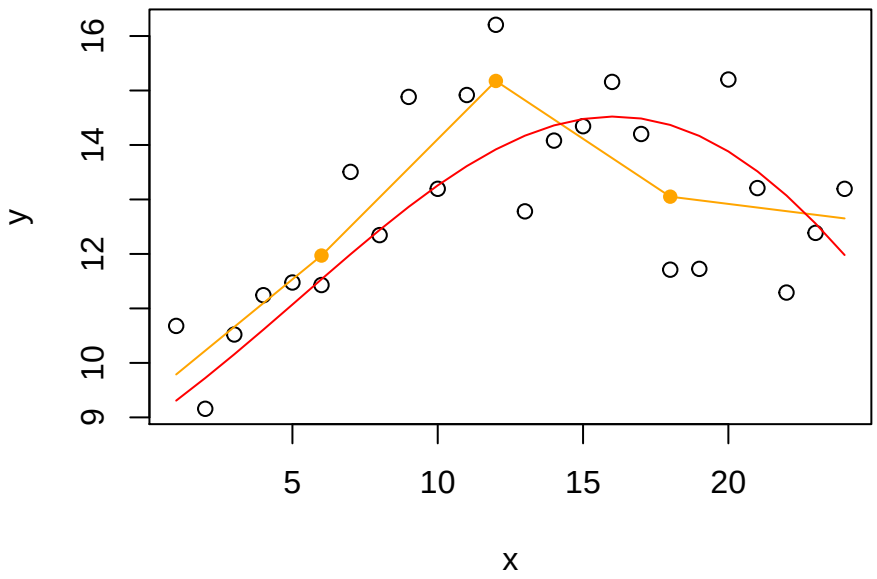
What does this data look like now?

```
t(cbind(x, x6, x12, x18)[8:20,])
```

```
##     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13]
## x      8    9   10   11   12   13   14   15   16    17    18    19    20
## x6     2    3    4    5    6    7    8    9   10    11    12    13    14
## x12    0    0    0    0    0    1    2    3    4     5     6     7     8
## x18    0    0    0    0    0    0    0    0    0     0     0     1     2
```

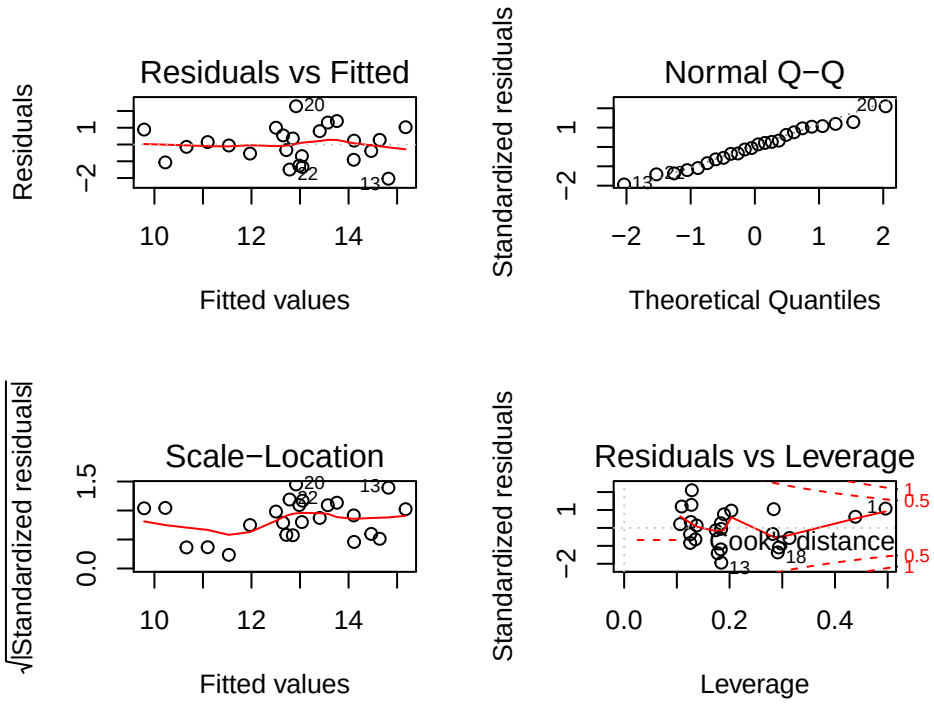
Now let's fit this model.

```
fit.lin.spline <- lm(y ~ x + x6 + x12 + x18)
plot(y~x)
lines(x, predict(fit.lin.spline), col="orange")
points(c(6, 12, 18), predict(fit.lin.spline)[c(6, 12, 18)], pch=16, col="orange")
lines(x, mu, col="red" )
```



Much closer than the linear model, but it still lacks the curvature that is present in the data. The residual plots look much better already.

```
par(mfrow=c(2,2))
plot(fit.lin.spline)
```



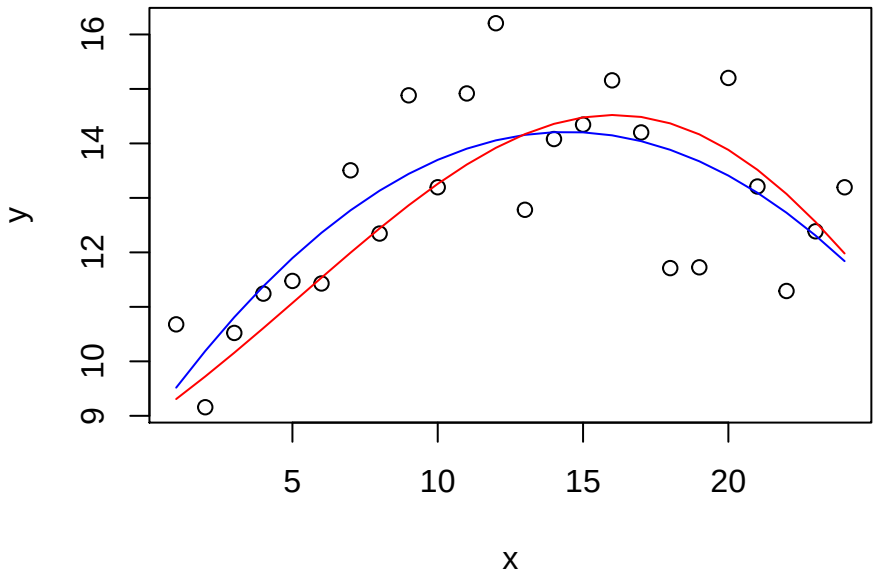
Powers

A non-linear effect can be as simple as adding a covariate at some power.

$$E(Y|X) = \beta_0 + \beta_1 X + \beta_2 X^2$$

Testing $H_0 : \beta_2 = 0$ tests the null hypothesis that the effect of X_1 on Y is linear vs the effect is quadratic.

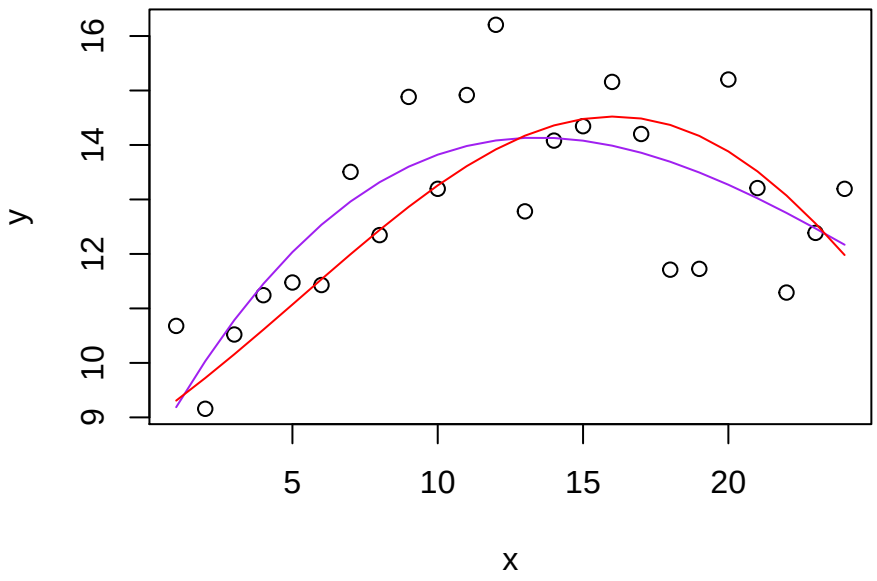
```
x.squared <- x^2
fit.sq <- lm(y~x + x.squared)
plot(y~x)
lines(x, predict(fit.sq), col="blue")
lines(x, mu, col="red" )
```



A cubic term could also be added.

$$E(Y|X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 x^3$$

```
x.cubed <- x^3
fit.cubic <- lm(y~x + x.squared + x.cubed)
plot(y~x)
lines(x, predict(fit.cubic), col="purple")
lines(x, mu, col="red" )
```



Adding this cubic term allows for another "wiggle" in the fitted line.

Cubic splines

Combining the two concepts allows for a very flexible polynomial model.

$$E(Y|X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 (X-a)_+^3 + \beta_5 (X-b)_+^3$$

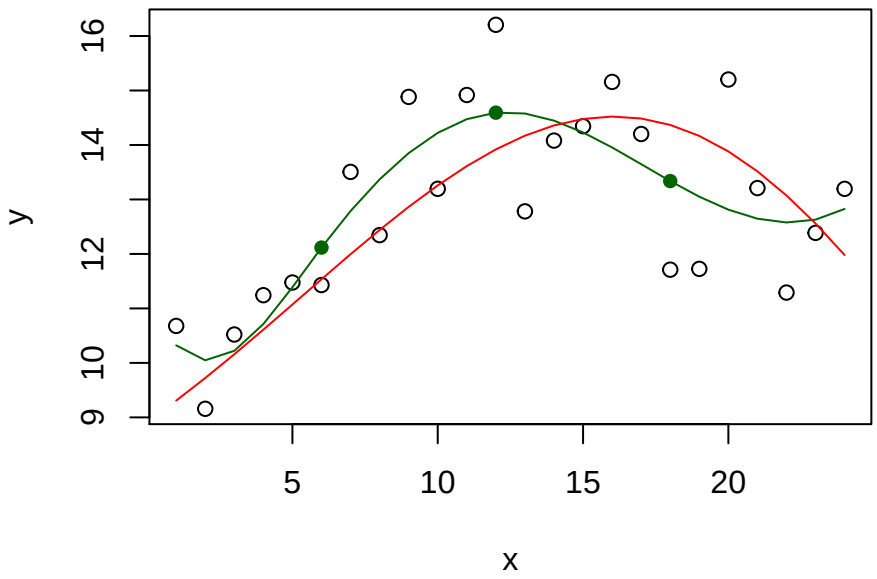
Using the knots at 6, 12, and 18 let's fit a cubic spline.

```
x6.cubed <- x6^3
x12.cubed <- x12^3
x18.cubed <- x18^3

fit.cub.spline <- lm(y ~ x + x.squared + x.cubed + x6.cubed + x12.cubed + x18.cubed)
```

Replot and look at the fitted model.

```
plot(y~x)
lines(x, predict(fit.cub.spline), col="darkgreen")
points(c(6, 12, 18), predict(fit.cub.spline)[c(6, 12, 18)], pch=16, col="darkgreen")
lines(x, mu, col="red" )
```



It seems like our model is fitting the data better, but sometimes there is a balance between a flexible model, and overfitting the data (when your model fits each point better than the true underlying average.)

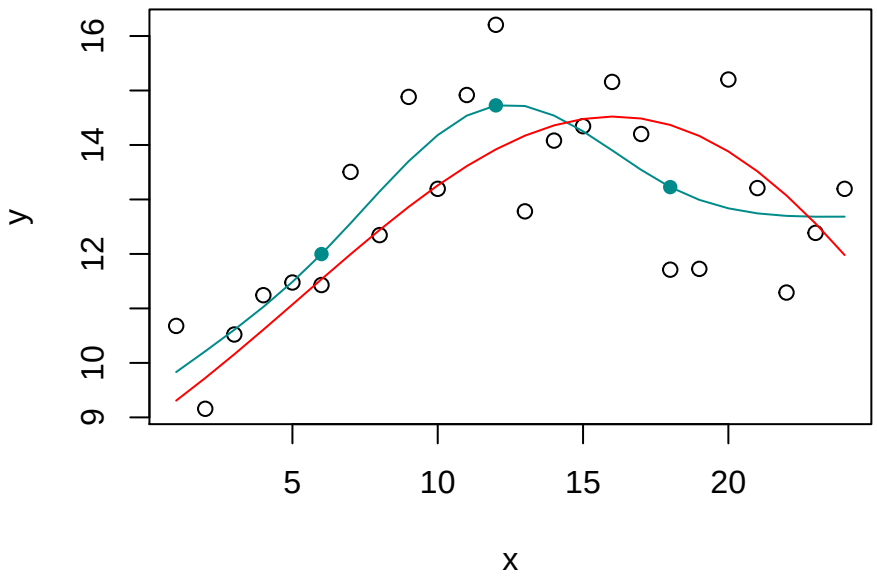
Natural splines

Also called *natural splines*, these models constrain the model to be linear in the tails. The model is difficult to write, and fit by hand so we will use the `splines` package.

```
library(splines)

fit.ns = lm( y ~ ns(x, knots=c(6,12,18) ) )

plot(y~x)
lines(x, predict(fit.ns), col="darkcyan")
points(c(6, 12, 18), predict(fit.ns)[c(6, 12, 18)], pch=16, col="darkcyan")
lines(x, mu, col="red" )
```



There are other methods of model fitting under the umbrella of *Nonparametric Regression*, these include kernel smoothing, smoothing splines, and the familiar LOWESS (locally weighted scatterplot smoothing) and LOESS (Local regression) models.

On Your Own

On Your Own

1. Using the `cars` data set built into R, build a model to predict the distance a car takes to stop based on how fast it was going.

2. Using the family lung function data, build a model to predict FEV1 to height for the oldest child.