Lec 03: Multiple Linear Regression Analysis

MATH 456 - Spring 2016

Assigned Reading

Afifi: Chapter 7

Multiple Regression and Correlation (*Afifi Ch 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

[top]

- Mean of y values at any given x_i is: $E(y|x_i) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$
- Variance of Y values at any set of values of X's is σ^2 (same for all X's)
- Y values are normally distributed at any given X (need for inference)

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.

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

Estimation of parameters

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.

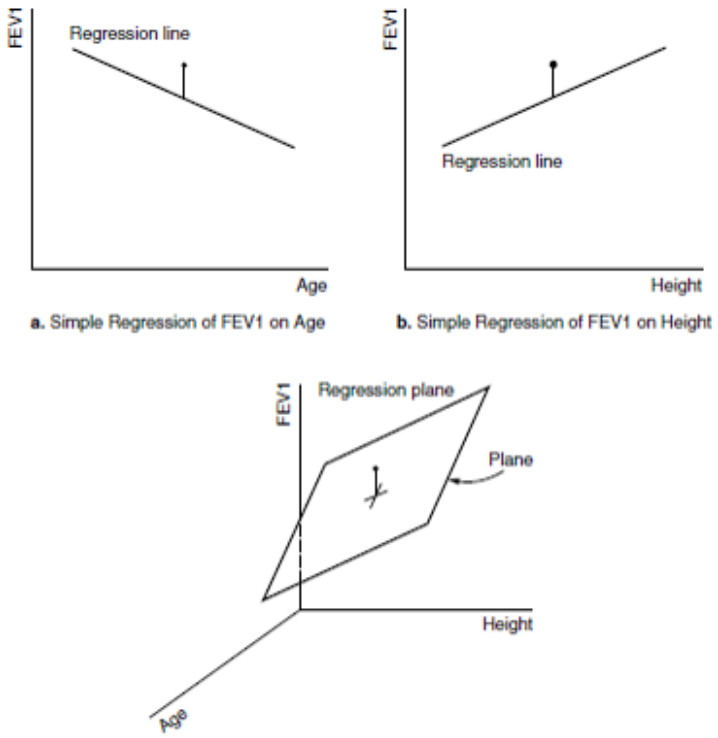


Figure 7.1: *Hypothetical Representation of Simple and Multiple Regression Equations of FEV1 on Age and Height*

Figure 1: Figure 7.1

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

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

[top]

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
```

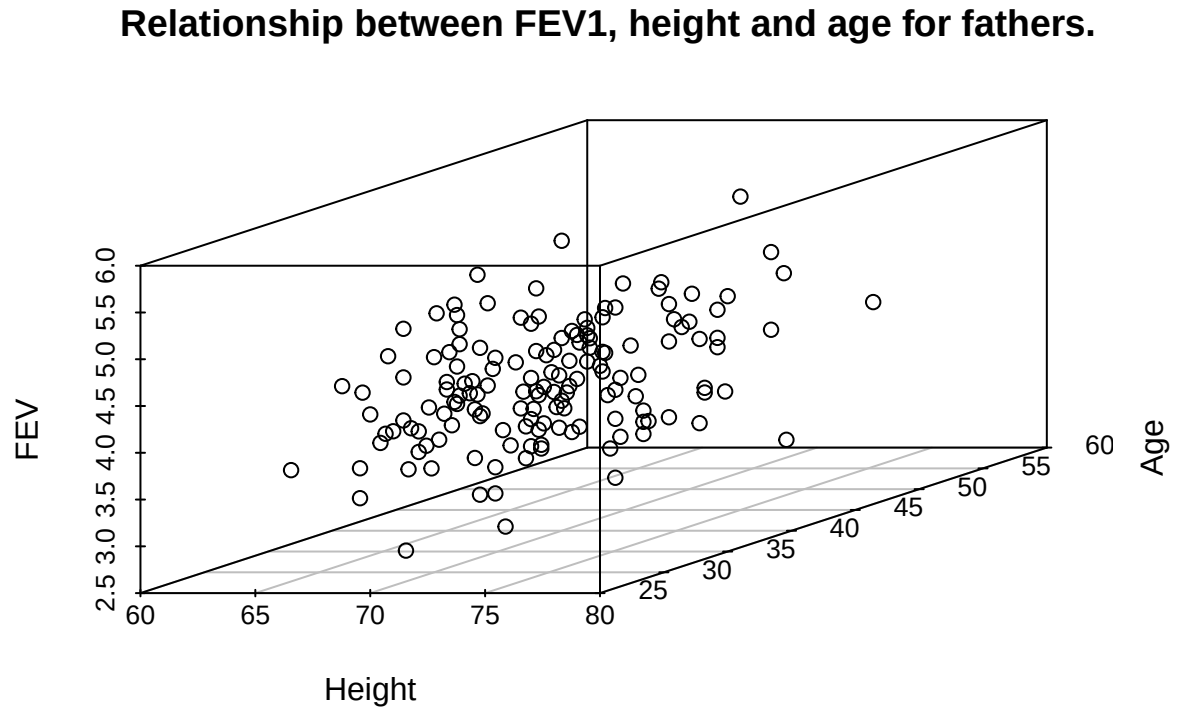
This model concludes that FEV1 in fathers significantly increases by 0.12 (95% CI: 0.09, 0.15) liters per additional inch in height (p<.0001). Looking at the multiple R^2 (correlation of determination), this simple model explains 25% of the variance seen in the outcome y.

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

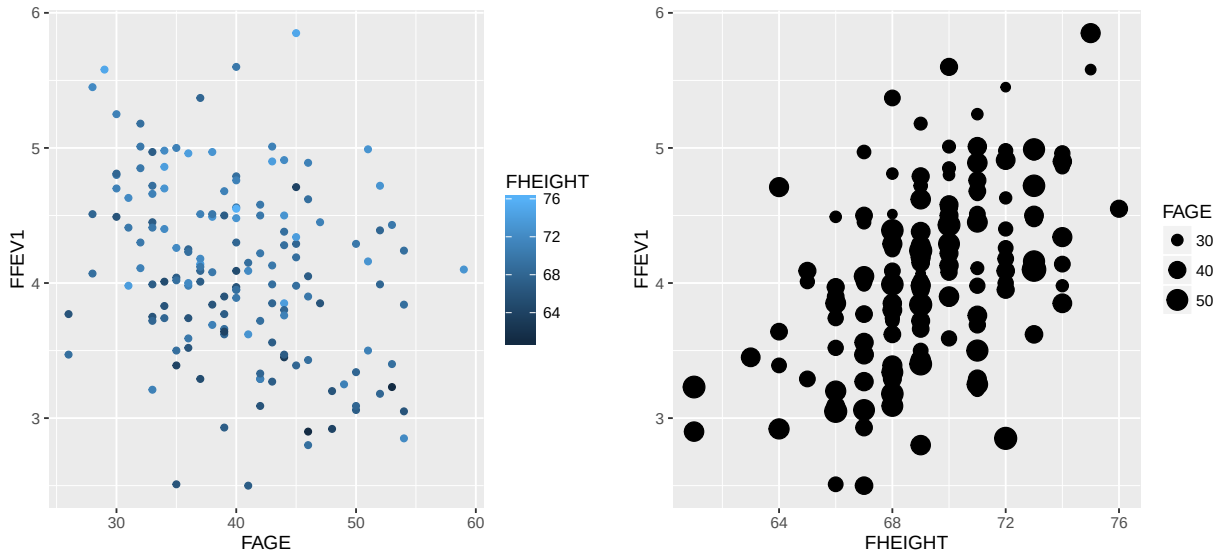
```
library(scatterplot3d)
scatterplot3d(x=fev$FHEIGHT, y=fev$FAGE, z=fev$FFEV1,
              xlab="Height", ylab="Age", zlab="FEV",
              main="Relationship between FEV1, height and age for fathers.")
```



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

[top]

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
```

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

[top]

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
```

Interpretation of regression coefficients

[top]

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. 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.

- 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)
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")
```

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.

Other parameters estimated

[top]

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
```

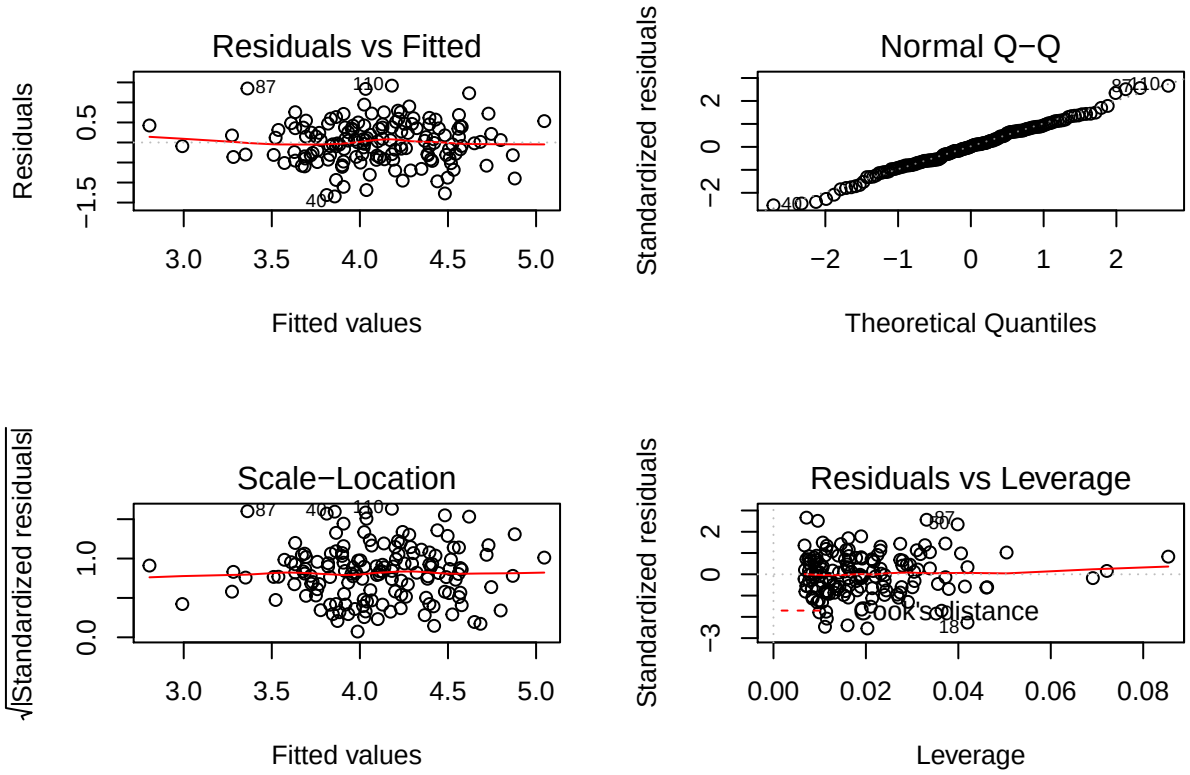
- The correlation of a variable with itself is always 1, hence the 1's down the diagonal.
- Each entry on the off diagonal is the simple correlation value: ρ_{12}.
- This matrix is also symmetric.

Regression diagnostics and transformations

[top]

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)

Interactions between variables

[top]

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 &

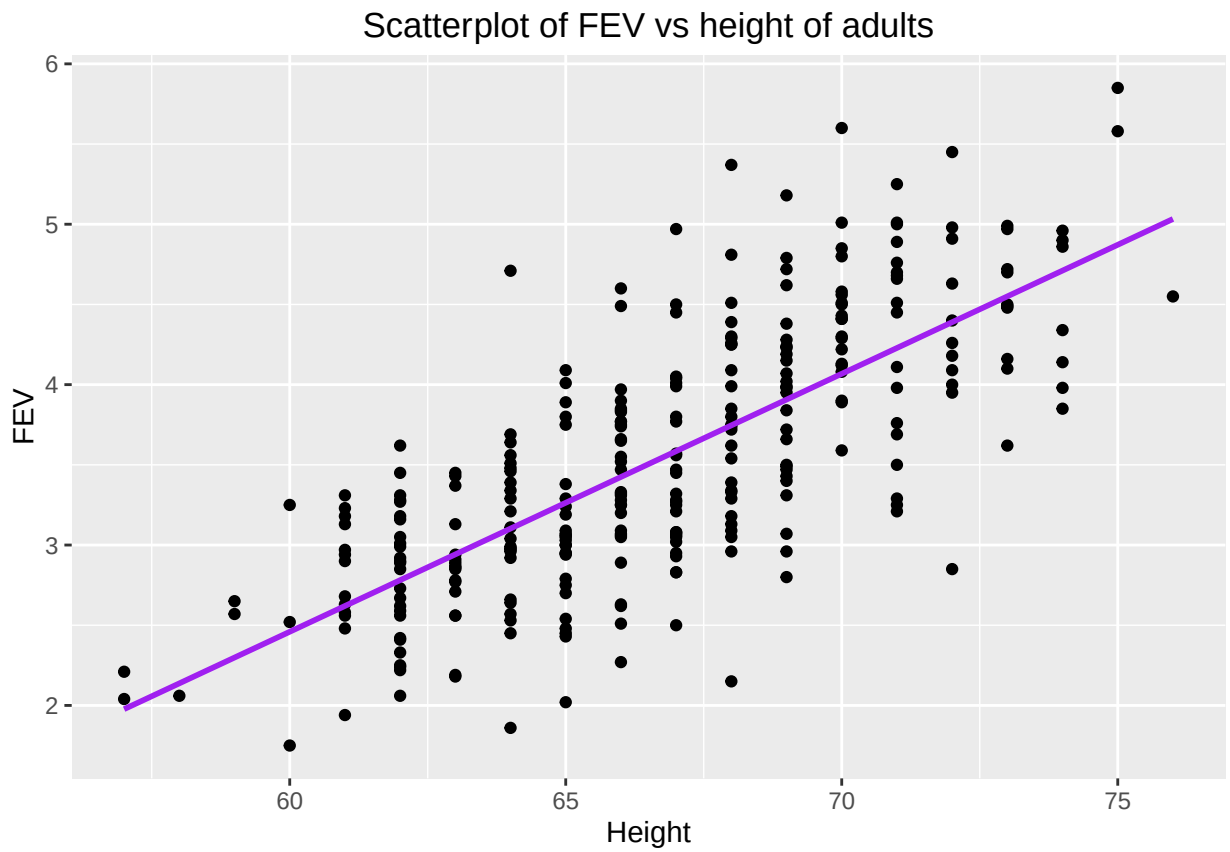
For demonstration purposes, consider the situation where the data came to us as *long* format with a separate variable for gender instead of in *wide* format with separate variables for mother and father characteristics. Here I combine height, FEV and age for both males and females into a single long data set.

```
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
```

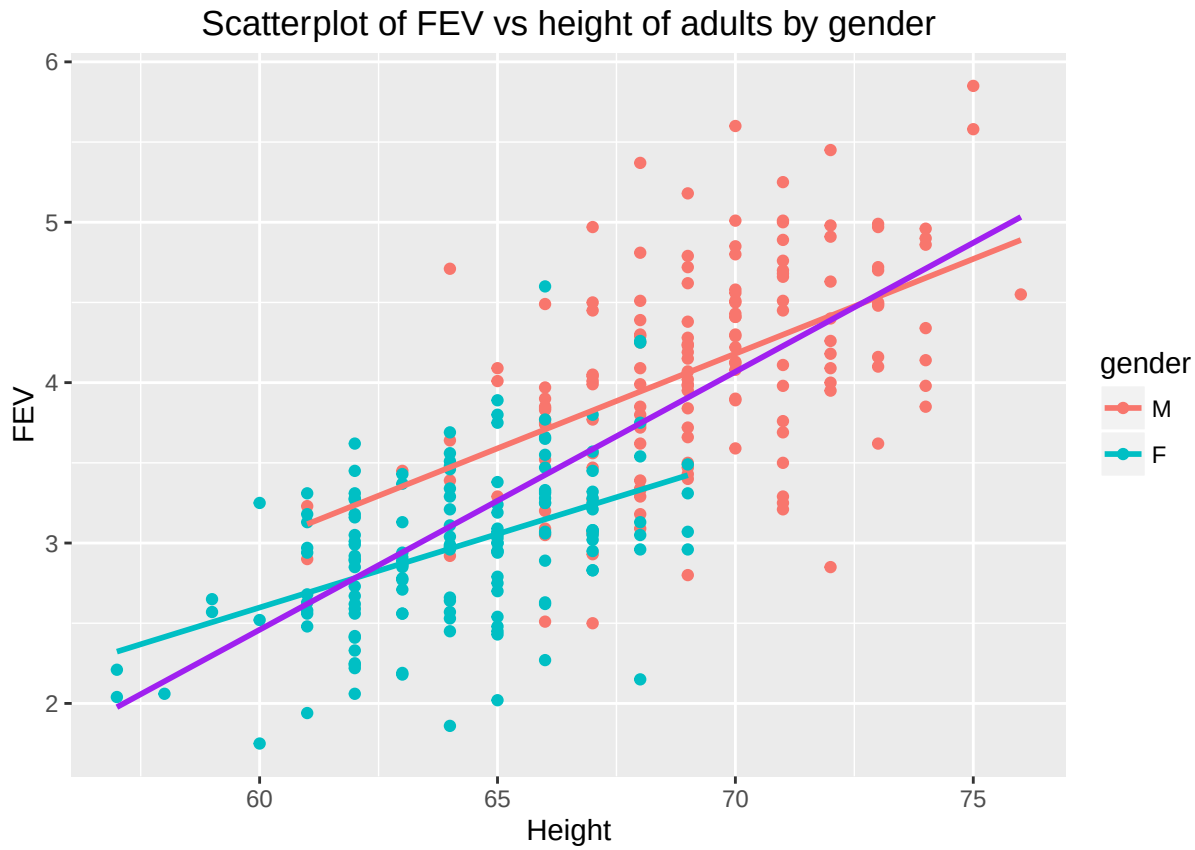
What does the relationship between height and FEV look like for the overall sample?

```
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")
```



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 depends on gender. These variables are said to **interact** with each other.

Recap of a main effects model

[top]

First let's examine the *main effects* model, that is the model without any interaction terms.

$$FEV1 \sim \beta_0 + \beta_1 * height + \beta_2 * gender$$

```
summary(lm(FEV ~ ht + gender, data=fev2))
```

```
##
## Call:
## lm(formula = FEV ~ ht + gender, data = fev2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.53494 -0.31541  0.00351  0.31137  1.42796
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.27944    0.76883  -4.266 2.68e-05 ***
## ht           0.10645    0.01108   9.603  < 2e-16 ***
## genderF     -0.57014    0.08157  -6.989 1.82e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5031 on 297 degrees of freedom
## Multiple R-squared:  0.6098, Adjusted R-squared:  0.6071
## F-statistic:   232 on 2 and 297 DF,  p-value: < 2.2e-16
```

The first thing to notice is the coefficient label: `genderF`. The gender variable is a factor variable, and is automatically *dummy 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. In this case there is only one other option, Male, so when `genderF=0` the record states male.

Interpretations

- For every inch taller someone is, their FEV increases by 0.1 liters.
- Females have a FEV measurement of 0.57 lower compared to males of the same height.

Adding interaction terms to a model

[top]

The full interaction model can be written as:

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

To add an interaction term to a linear model in R you use the `*` operator between the interaction variables.

```
intx_model <- lm(FEV ~ ht + gender + ht*gender, data=fev2)
summary(intx_model)
```

```
##
## Call:
## lm(formula = FEV ~ ht + gender + ht * gender, data = fev2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.56688 -0.31451  0.00643  0.31791  1.45205
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.08670    1.02715  -3.979 8.72e-05 ***
## ht           0.11811    0.01482   7.970 3.46e-14 ***
## genderF      1.18326    1.48297   0.798    0.426
## ht:genderF  -0.02642    0.02231  -1.184    0.237
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5027 on 296 degrees of freedom
## Multiple R-squared:  0.6116, Adjusted R-squared:  0.6077
## F-statistic: 155.4 on 3 and 296 DF,  p-value: < 2.2e-16
```

```
confint(intx_model)
```

```
##                    2.5 %      97.5 %
## (Intercept) -6.10815291 -2.06525202
## ht           0.08894215  0.14726834
## genderF     -1.73524201  4.10175441
## ht:genderF  -0.07032214  0.01748734
```

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.

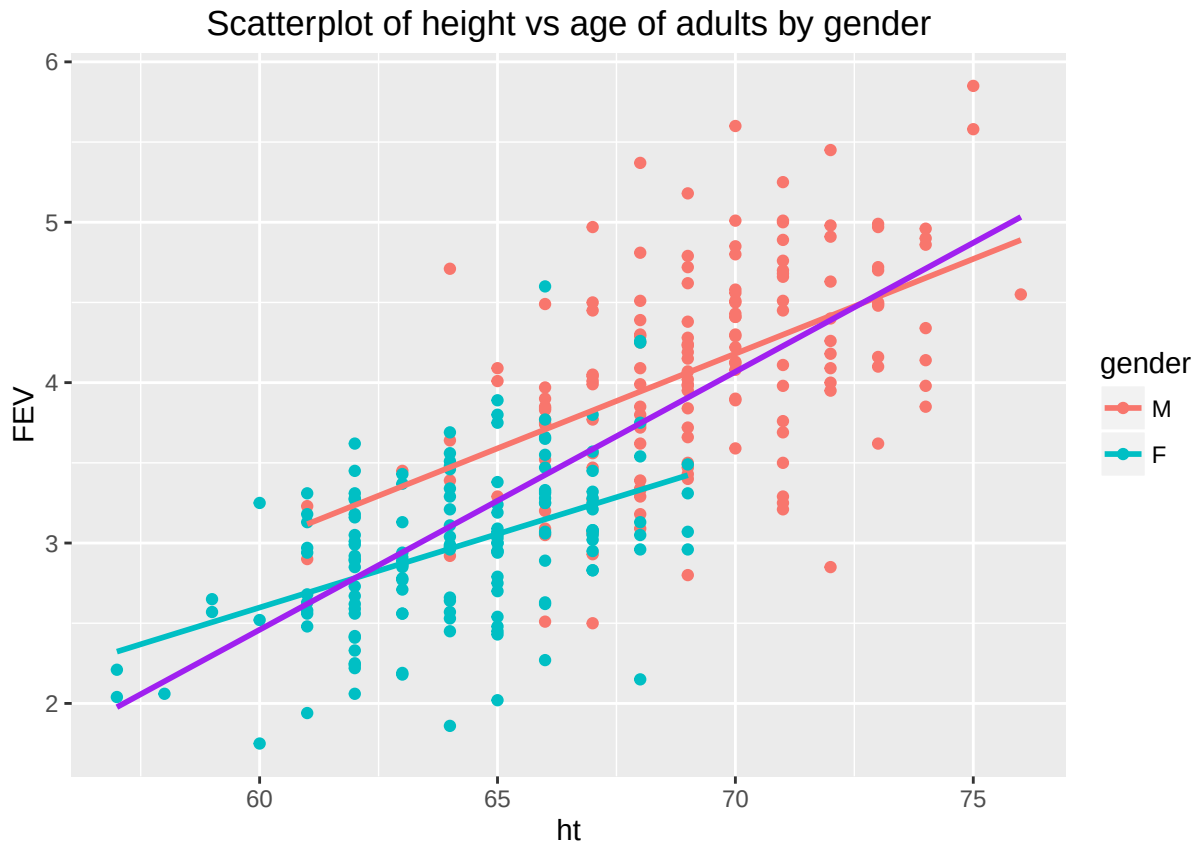
Stratification

[top]

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).

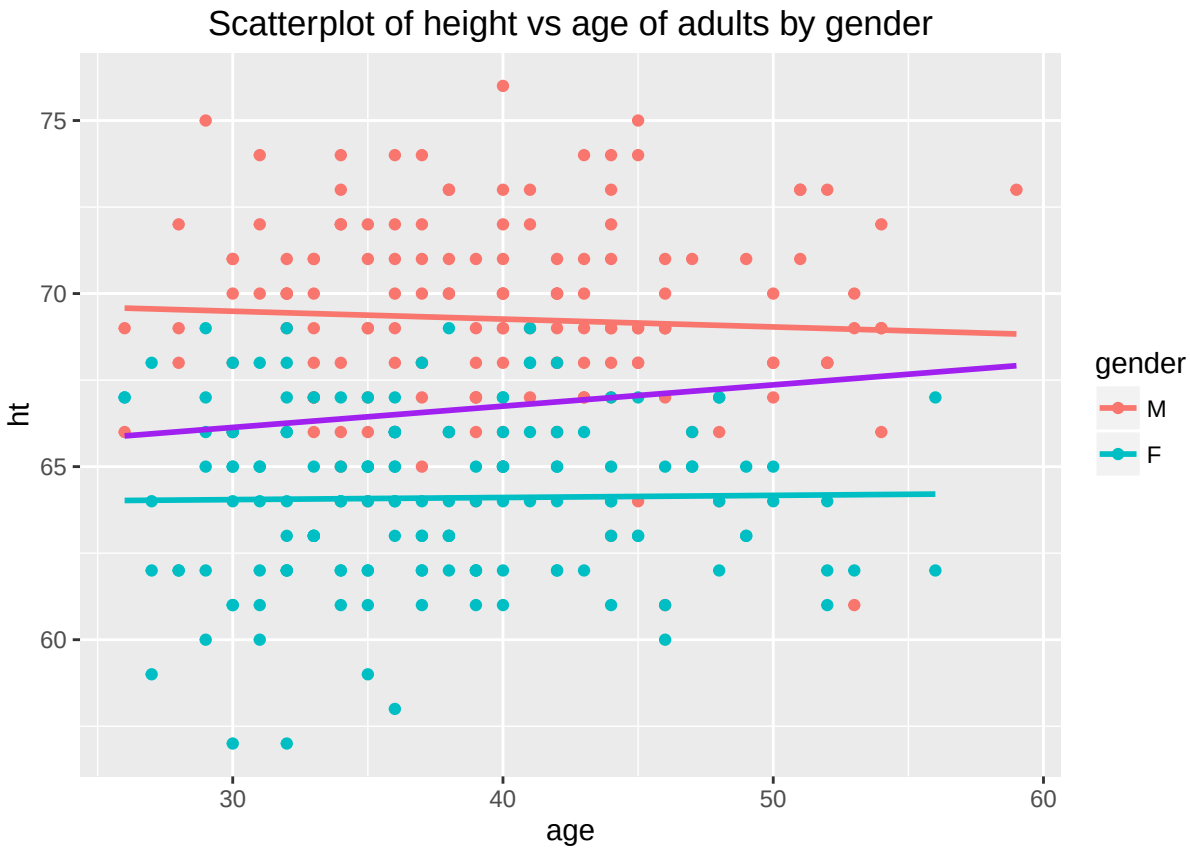
$$FEV1_M \sim \beta_{0M} + \beta_{1M} * height + \beta_{2M} * age$$
$$FEV1_F \sim \beta_{0M} + \beta_{1F} * height + \beta_{2F} * age$$

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



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 and run

—>

```
# 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")
```

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")
```

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")
```

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.

```r
# 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")
```

1

2

o.pred

3.586

2.555

m.pred

3.990

3.000

f.pred

3.301

2.531

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

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")
```

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

[top]

- 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?