Multiple Regression

Robin Donatello

Last Updated 2017-11-05 12:33:22

Contents

This topic is discussed in depth in PMA5, Chapter 7. The UCLA IDRE has constructed [textbook examples] for this chapter.

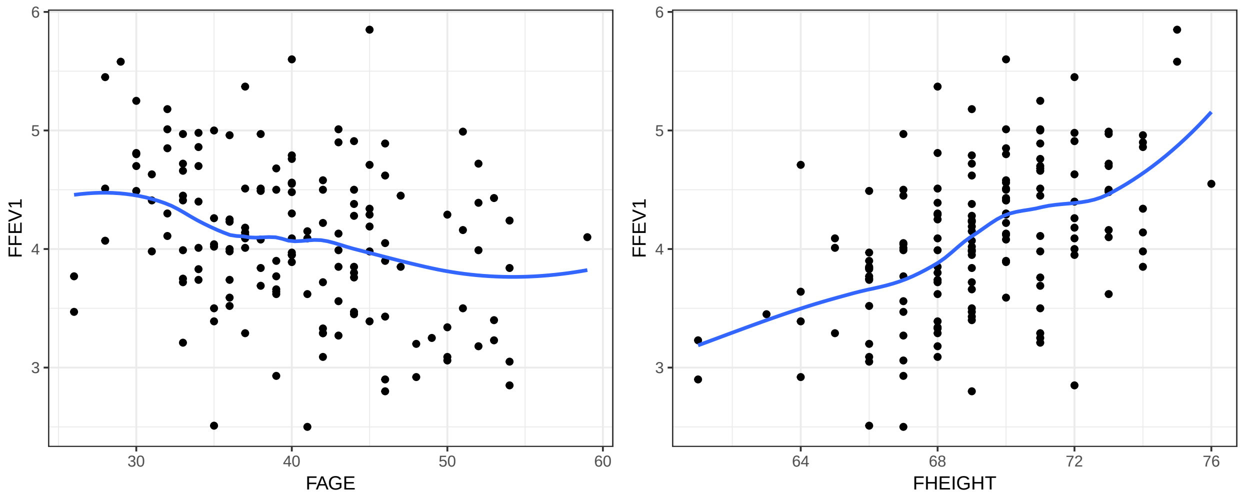
Purpose

The general purpose of multiple regression is to learn more about the relationship between several independent or predictor variables and a quantitative dependent variable. Multiple regression procedures are very widely used in research. In general, this inferential tool allows us to ask (and hopefully answer) the general question "*what is the best predictor of…*", and does "*additional variable A*" or "*additional variable B" confound the relationship between my explanatory and response variable?*"

- Educational researchers might want to learn about the best predictors of success in high-school.
- Sociologists may want to find out which of the multiple social indicators best predict whether or not a new immigrant group will adapt to their new country of residence.
- Biologists may want to find out which factors (i.e. temperature, barometric pressure, humidity, etc.) best predict caterpillar reproduction.

Consider three variables that measure lung function: Age, Height, and FEV1 (The amount of air exhaled during the first second of a forced breath).

```r
a <- ggplot(fev, aes(y=FFEV1, x=FAGE)) +
        geom_point() + geom_smooth(se=FALSE) + theme_bw()
b <- ggplot(fev, aes(y=FFEV1, x=FHEIGHT)) +
        geom_point() + geom_smooth(se=FALSE) + theme_bw()
grid.arrange(a, b, ncol=2)
```



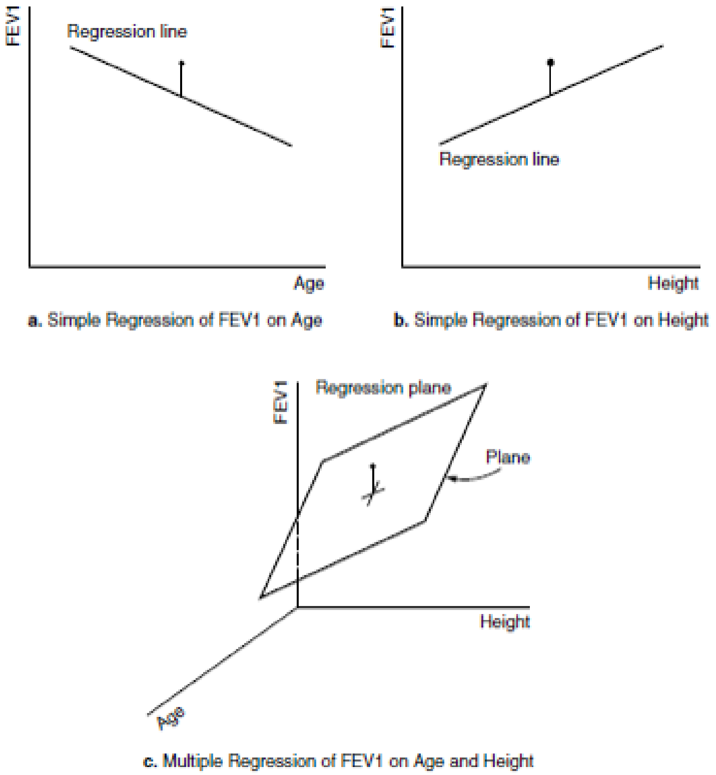
Multiple Linear Regression,

- Extends simple linear regression.
- Describes a linear relationship between a single continuous Y variable, and several X variables.
- Predicts 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., $log(x), x^2$.



a. Simple Regression of FEV1 on Age

b. Simple Regression of FEV1 on Height

c. Multiple Regression of FEV1 on Age and Height

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

Figure 1:

Mathematical Model

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

The assumptions on the residuals ϵ_i still hold:

- 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) 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:

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

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 the regression coefficient estimates that minimize the sum of squared errors can be calculated as:

$$\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x} \quad \text{and} \quad \hat{\beta_1} = r\frac{s_y}{s_x}$$

For multiple linear regression, the fitted values $\hat{y_i}$ are calculated as the linear combination of x's and β's, $\sum_{i=1}^{p} X_{ij}\beta_j$. The sum of the squared residual errors (the distance between the observed point y_i and the fitted value) now has the following form:

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

Or in matrix notation

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

The details of methods to calculate the Least Squares estimate of β's is left to a course in mathematical statistics.

Continued Example: Lung Function

In PMA5 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`.

```r
fev.ht.model <- lm(FFEV1 ~ FHEIGHT , data=fev)
summary(fev.ht.model)
```

```
## 
## 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
```

```r
round(confint(fev.ht.model),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.

- 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`.

```r
fev1.age.model <- lm(FFEV1 ~ FAGE, data=fev)
summary(fev1.age.model)
```

```
## 
## 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
```

```
confint(fev1.age.model)
```

```
##                   2.5 %      97.5 %
## (Intercept)  4.67237168  5.86037675
## FAGE        -0.04381897 -0.01464154
```

For every one year older the father gets, his FEV1 significantly decreases by 0.03 (95% CI 0.02, 0.04) liters (p = .00001).

```
fev1.ht.model <- lm(FFEV1 ~ FHEIGHT, data=fev)
summary(fev1.ht.model)
```

```
## 
## 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
```

```
confint(fev1.ht.model)
```

```
##                   2.5 %     97.5 %
## (Intercept) -6.36315502 -1.8102499
## FHEIGHT      0.08526328  0.1509472
```

For every inch taller a father is, his FEV1 significantly increases by 0.11 (95%CI 0.09, 0.15) 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
```

```
confint(mv_model)
```

```
##                   2.5 %      97.5 %
## (Intercept) -5.00919751 -0.51229620
## FAGE        -0.03922545 -0.01405323
## FHEIGHT      0.08319434  0.14559974
```

Holding height constant, a father who is one year older is expected to have a FEV value 0.03 (0.01, 0.04) liters less than another man (p<.0001).

Holding height constant, a father who is 1cm taller than another man is expected to have a FEV value of 0.11 (.08, 0.15) liter greater than the other man (p<.0001).

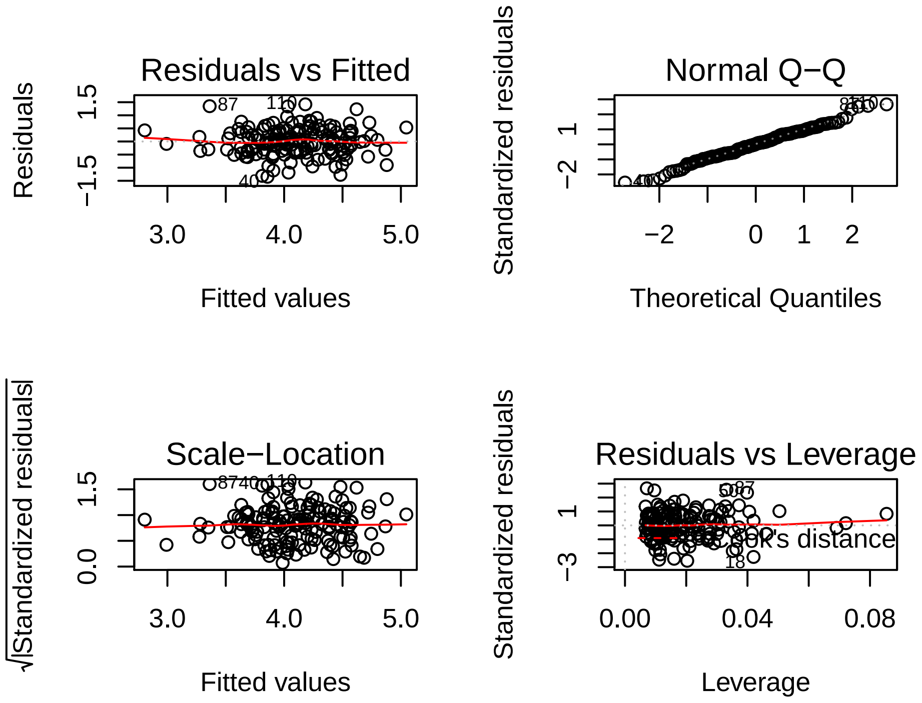
For the model that includes age, the coefficient for height is now 0.11, which is interpreted as the rate of change of FEV1 as a function of height **after adjusting for age**. This is also called the **partial regression coefficient** of FEV1 on height after adjusting for age.

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

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)
```



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 (PMA5 Ch. 14)

Interlude: The necessity of tidy data for analysis.

The data on Lung function originally was recorded in *wide* format, with separate variables for mother's and father's FEV1 score (`MFEV1` and `FFEV`). In this format, the data is one record per family.

```
head(fev)
```

```
##   ID AREA FSEX FAGE FHEIGHT FWEIGHT FFVC FFEV1 MSEX MAGE MHEIGHT MWEIGHT
## 1  1    1    1   53      61     161  391  3.23    2   43      62     136
## 2  2    1    1   40      72     198  441  3.95    2   38      66     160
## 3  3    1    1   26      69     210  445  3.47    2   27      59     114
## 4  4    1    1   34      68     187  433  3.74    2   36      58     123
## 5  5    1    1   46      61     121  354  2.90    2   39      62     128
## 6  6    1    1   44      72     153  610  4.91    2   36      66     125
##   MFVC MFEV1 OCSEX OCAGE OCHEIGHT OCWEIGHT OCFVC OCFEV1 MCSEX MCAGE
## 1  370  3.31     2    12       59      115   296   2.79    NA    NA
## 2  411  3.47     1    10       56       66   323   2.39    NA    NA
## 3  309  2.65     1     8       50       59   114   1.11    NA    NA
## 4  265  2.06     2    11       57      106   256   1.85     1     9
## 5  245  2.33     1    16       61       88   260   2.47     2    12
## 6  349  3.06     1    15       67      100   389   3.55     1    13
##   MCHEIGHT MCWEIGHT MCFVC MCFEV1 YCSEX YCAGE YCHEIGHT YCWEIGHT YCFVC
## 1       NA       NA    NA     NA    NA    NA       NA       NA    NA
## 2       NA       NA    NA     NA    NA    NA       NA       NA    NA
## 3       NA       NA    NA     NA    NA    NA       NA       NA    NA
## 4       49       56   159   1.30    NA    NA       NA       NA    NA
## 5       60       85   268   2.34     2    10       50       53   154
## 6       57       87   276   2.37     2    10       55       72   195
##   YCFEV1
## 1     NA
## 2     NA
## 3     NA
## 4     NA
## 5   1.43
## 6   1.69
```

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
```

The UCLA IDRE [textbook examples] shows you how to reshape this data long to wide using `varstocases`.

Stratification & Interactions

Recall when testing for a moderator, we fit models *stratified* on the potential moderating variable. In doing so, we were examining the regression equations for each subgroup of the population and seeing if the relationship between the response and explanatory variables *changed* for at least one subgroup.

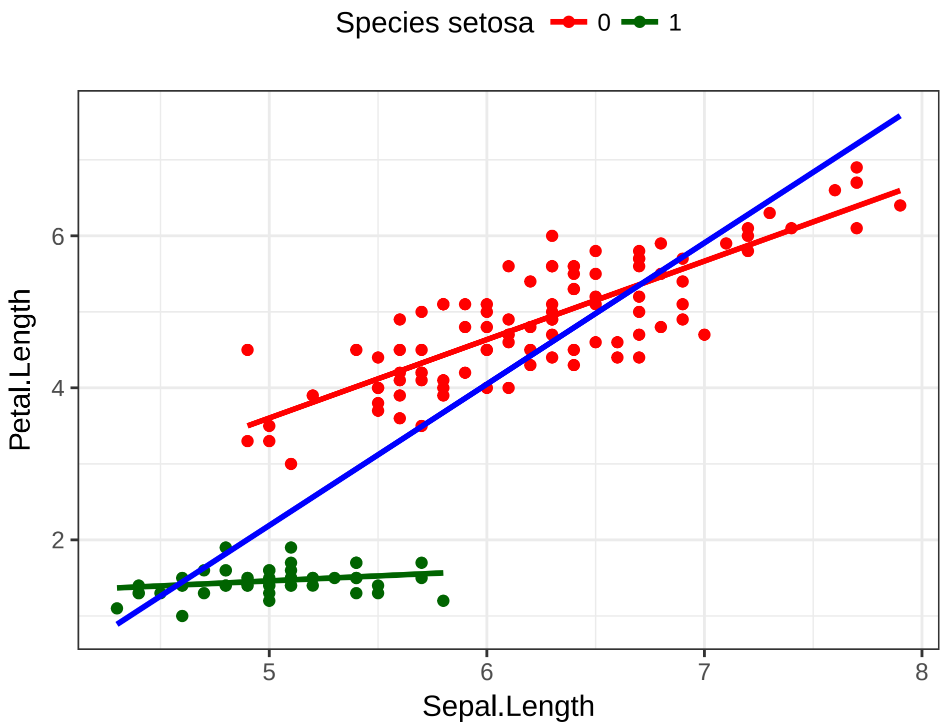
Consider the relationship between the length of an iris petal, and the length of it's sepal. Earlier we found that the iris species modified this relationship. Lets consider a binary indicator variable for species that groups *veriscolor* and *virginica* together.

```r
iris$setosa <- ifelse(iris$Species=="setosa", 1, 0)
table(iris$setosa, iris$Species)
```

```
##
##     setosa versicolor virginica
##   0      0         50        50
##   1     50          0         0
```

Within the *setosa* species, there is little to no relationship between sepal and petal length. For the other two species, the relationship looks still significantly positive, but in the combined sample there appears to be a strong positive relationship (blue).

```r
ggplot(iris, aes(x=Sepal.Length, y=Petal.Length, col=as.factor(setosa))) +
            geom_point() + theme_bw() + theme(legend.position="top") +
            scale_color_manual(name="Species setosa", values=c("red", "darkgreen")) +
            geom_smooth(se=FALSE, method="lm") +
            geom_smooth(aes(x=Sepal.Length, y=Petal.Length), col="blue", se=FALSE, method='lm')
```



The mathematical model describing the relationship between Petal length (Y), and Sepal length (X), for species *setosa* (s) versus not-setosa (n), is written as follows:

$$Y_{is} \sim \beta_{0s} + \beta_{1s} * x_i + \epsilon_{is} \qquad \epsilon_{is} \sim \mathcal{N}(0, \sigma_s^2)$$
$$Y_{in} \sim \beta_{0n} + \beta_{1n} * x_i + \epsilon_{in} \qquad \epsilon_{in} \sim \mathcal{N}(0, \sigma_n^2)$$

In each model, the intercept, slope, and variance of the residuals can all be different. This is the unique and powerful feature of stratified models. The downside is that each model is only fit on the amount of data in that particular subset. Furthermore, each model has 3 parameters that need to be estimated: β_0, β_1, and σ^2, for a total of 6 for the two models. The more parameters that need to be estimated, the more data we need.

Simplififcation of interaction models

If we only care about how species changes the relationship between petal and sepal length, we can fit what is called an **interaction** model. Interactions are mathematically represented as a multiplication between the two variables that are interacting: here it is sepal length (x_1) and species (x_2). Note that both *main effects* of sepal length, and species are also included in the model.

$$Y_i \sim \beta_0 + \beta_1 x_i + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i}$$

When $x_2 = 0$, the record is on an iris not from the *setosa* species.

$$Y_i \sim \beta_0 + \beta_1 x_i + \beta_2(0) + \beta_3 x_{1i}(0)$$

which simplifies to

$$Y_i \sim \beta_0 + \beta_1 x_i$$

When $x_2 = 1$, the record is on an iris of the *setosa* species.

$$Y_i \sim \beta_0 + \beta_1 x_i + \beta_2(1) + \beta_3 x_{1i}(1)$$

which simplifies to

$$Y_i \sim (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_i$$

These then **are** the stratified models! Each sub-model has a different intercept and slope, but we only had to estimate 4 parameters instead of 6.

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

```r
summary(lm(Petal.Length ~ Sepal.Length + setosa + Sepal.Length*setosa, data=iris))
```

```
##
## Call:
## lm(formula = Petal.Length ~ Sepal.Length + setosa + Sepal.Length *
##     setosa, data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.96754 -0.19948 -0.01386  0.22597  1.05479
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1.55571    0.37509  -4.148 5.68e-05 ***
## Sepal.Length         1.03189    0.05957  17.322  < 2e-16 ***
## setosa               2.35877    0.88266   2.672  0.00839 **
## Sepal.Length:setosa -0.90026    0.17000  -5.296 4.28e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3929 on 146 degrees of freedom
## Multiple R-squared:  0.9515, Adjusted R-squared:  0.9505
## F-statistic: 954.1 on 3 and 146 DF,  p-value: < 2.2e-16
```

The coefficient b_3 for the interaction term is significant, confirming that species changes the relationship between sepal length and petal length. How we interpret this, and the other coefficients will be discussed later.

Adding more covariates to the model

What if we now wanted to include other predictors in the model? How does sepal length relate to petal length after controlling for petal width? We add the variable for petal width into the model

```
summary(lm(Petal.Length ~ Sepal.Length + setosa + Sepal.Length*setosa + Petal.Width, data=iris))
```

```
##
## Call:
## lm(formula = Petal.Length ~ Sepal.Length + setosa + Sepal.Length *
##     setosa + Petal.Width, data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.83519 -0.18278 -0.01812  0.17004  1.06968
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -0.86850    0.27028  -3.213  0.00162 **
## Sepal.Length         0.66181    0.05179  12.779  < 2e-16 ***
## setosa               1.83713    0.62355   2.946  0.00375 **
## Petal.Width          0.97269    0.07970  12.204  < 2e-16 ***
## Sepal.Length:setosa -0.61106    0.12213  -5.003 1.61e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2769 on 145 degrees of freedom
## Multiple R-squared:  0.9761, Adjusted R-squared:  0.9754
## F-statistic:  1478 on 4 and 145 DF,  p-value: < 2.2e-16
```

So far, petal width, and the combination of species and sepal length are both significantly associated with petal length.

Note of caution: Stratification implies that the stratifying variable interacts with all other variables. So if we were to go back to the stratified model where we fit the model of petal length on sepal length AND petal width, stratified by species, we would be implying that species interacts with both sepal length and petal width.

Categorical Predictors

This topic is also discussed in more detail in PMA5 Chapter 9.3.

Example: Iris species

Let's continue to model the length of the iris petal based on the length of the sepal, controlling for species. But here we'll keep species as a categorical variable. What happens if we just put the variable in the model?

```
summary(lm(Petal.Length ~ Sepal.Length + Species, data=iris))
```

```
##
## Call:
## lm(formula = Petal.Length ~ Sepal.Length + Species, data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.76390 -0.17875  0.00716  0.17461  0.79954
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1.70234    0.23013  -7.397 1.01e-11 ***
## Sepal.Length       0.63211    0.04527  13.962  < 2e-16 ***
## Speciesversicolor  2.21014    0.07047  31.362  < 2e-16 ***
## Speciesvirginica   3.09000    0.09123  33.870  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2826 on 146 degrees of freedom
## Multiple R-squared:  0.9749, Adjusted R-squared:  0.9744
## F-statistic:  1890 on 3 and 146 DF,  p-value: < 2.2e-16
```

Examine the coefficient names, `Speciesversicolor` and `Speciesvirginica`. R (and most software packages) automatically take a categorical variable and turn it into a series of binary indicator variables. Let's look at what the software program does in the background. Below is a sample of the iris data. The first column shows the row number, specifically I am only showing 2 sample rows from each species. The second column is the value of the sepal length, the third is the binary indicator for if the iris is from species *versicolor*, next the binary indicator for if the iris is from species *virginica*, and lastly the species as a 3 level categorical variable (which is what we're used to seeing at this point.)

	Sepal.Length	Speciesversicolor	Speciesvirginica	Species
1	5.1	0	0	setosa
2	4.9	0	0	setosa
51	7	1	0	versicolor
52	6.4	1	0	versicolor
101	6.3	0	1	virginica
102	5.8	0	1	virginica

Factor variable coding

- Most commonly known as "Dummy coding". Not an informative term to use.
- 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.

For the iris example, 2 indicator variables are created for *versicolor* and *virginica*. Interpreting the regression coefficients are going to be **compared to the reference group**. In this case, it is species *setosa*.

The mathematical model is now written as follows, where x_1 is Sepal Length, x_2 is the indicator for *versicolor*, and x_3 the indicator for *virginica*

$$Y_i \sim \beta_0 + \beta_1 x_i + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i$$

Let's look at the regression coefficients and their 95% confidence intervals from the main effects model again.

```r
main.eff.model <- lm(Petal.Length ~ Sepal.Length + Species, data=iris)
pander(main.eff.model)
```

Table 2: Fitting linear model: Petal.Length ~ Sepal.Length + Species

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.702	0.2301	-7.397	1.005e-11
Sepal.Length	0.6321	0.04527	13.96	1.121e-28
Speciesversicolor	2.21	0.07047	31.36	9.646e-67
Speciesvirginica	3.09	0.09123	33.87	4.918e-71

```r
pander(confint(main.eff.model))
```

	2.5 %	97.5 %
(Intercept)	-2.157	-1.248
Sepal.Length	0.5426	0.7216
Speciesversicolor	2.071	2.349
Speciesvirginica	2.91	3.27

In this *main effects* model, Species only changes the intercept. The effect of species is not multiplied by Sepal length. The interpretations are the following:

- b_1: After controlling for species, Petal length significantly increases with the length of the sepal (0.63, 95% CI 0.54-0.72, p<.0001).
- b_2: *Versicolor* has on average 2.2cm longer petal lengths compared to *setosa* (95% CI 2.1-2.3, p<.0001).
- b_3: *Virginica* has on average 3.1cm longer petal lengths compared to *setosa* (95% CI 2.9-3.3, p<.0001).

Interactions between Q*B and Q*C

Lastly let's look at how to fit and interpret a model with an interaction between a categorical and a continuous variable. Recall an interaction **changes the relationship** between an explanatory variable and the response variable.

Q*B

Let's revisit the interaction model where x_1 is Sepal Length and x_2 is the indicator for *setosa*.

$$Y_i \sim \beta_0 + \beta_1 x_i + \beta_2 x_{2i} + \beta_3 x_{1i} * x_{2i}$$

```r
summary(lm(Petal.Length ~ Sepal.Length + setosa + Sepal.Length*setosa, data=iris))
```

```
##
## Call:
## lm(formula = Petal.Length ~ Sepal.Length + setosa + Sepal.Length *
##     setosa, data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.96754 -0.19948 -0.01386  0.22597  1.05479
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1.55571    0.37509  -4.148 5.68e-05 ***
## Sepal.Length         1.03189    0.05957  17.322  < 2e-16 ***
## setosa               2.35877    0.88266   2.672  0.00839 **
## Sepal.Length:setosa -0.90026    0.17000  -5.296 4.28e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3929 on 146 degrees of freedom
## Multiple R-squared:  0.9515, Adjusted R-squared:  0.9505
## F-statistic: 954.1 on 3 and 146 DF,  p-value: < 2.2e-16
```

The main effects (b_1, b_2) cannot be interpreted by themselves when there is an interaction in the model.

- If $x_2 = 0$, then the effect of x_1 on Y simplifies to: β_1
  - b_1 The effect of sepal length on petal length **for non-setosa species of iris** (`setosa=0`)
  - For non-setosa species, the petal length increases 1.03cm for every additional cm of sepal length.
- If $x_2 = 1$, then the effect of x_1 on Y model simplifies to: $\beta_1 + \beta_3$
  - For setosa species, the petal length increases by `1.03-0.9=0.13` cm for every additional cm of sepal length.

Don't remember how I got these model simplifications? See this section

Q*C

Let's up the game now and look at the full interaction model with a categorical version of species. Recall x_1 is Sepal Length, x_2 is the indicator for *versicolor*, and x_3 the indicator for *virginica* .

$$Y_i \sim \beta_0 + \beta_1 x_i + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{1i} x_{2i} + \beta_5 x_{1i} x_{3i} + \epsilon_i$$

```r
summary(lm(Petal.Length ~ Sepal.Length + Species + Sepal.Length*Species, data=iris))
```

```
##
## Call:
## lm(formula = Petal.Length ~ Sepal.Length + Species + Sepal.Length *
##     Species, data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68611 -0.13442 -0.00856  0.15966  0.79607
```

```
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       0.8031     0.5310   1.512    0.133
## Sepal.Length                      0.1316     0.1058   1.244    0.216
## Speciesversicolor                -0.6179     0.6837  -0.904    0.368
## Speciesvirginica                 -0.1926     0.6578  -0.293    0.770
## Sepal.Length:Speciesversicolor    0.5548     0.1281   4.330 2.78e-05 ***
## Sepal.Length:Speciesvirginica     0.6184     0.1210   5.111 1.00e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2611 on 144 degrees of freedom
## Multiple R-squared:  0.9789, Adjusted R-squared:  0.9781
## F-statistic:  1333 on 5 and 144 DF,  p-value: < 2.2e-16
```

The slope of the relationship between sepal length and petal length is calculated as follows, for each species:

- *setosa* $(x_2 = 0, x_3 = 0) : b_1 = 0.13$
- *versicolor* $(x_2 = 1, x_3 = 0) : b_1 + b_2 + b_4 = 0.13 + 0.55 = 0.68$
- *virginica* $(x_2 = 0, x_3 = 1) : b_1 + b_3 + b_5 = 0.13 + 0.62 = 0.75$

Compare this to the estimates gained from the stratified model:

```
coef(lm(Petal.Length ~ Sepal.Length, data=subset(iris, Species=="setosa")))
```

```
##  (Intercept) Sepal.Length
##    0.8030518    0.1316317
```

```
coef(lm(Petal.Length ~ Sepal.Length, data=subset(iris, Species=="versicolor")))
```

```
##  (Intercept) Sepal.Length
##    0.1851155    0.6864698
```

```
coef(lm(Petal.Length ~ Sepal.Length, data=subset(iris, Species=="virginica")))
```

```
##  (Intercept) Sepal.Length
##    0.6104680    0.7500808
```

They're the same! Proof that an interaction is equivelant to stratification.

So why do an interaction? Why not stratify?

Stratification implies that the stratifying variable interacts with all other variables. Even variables that the variable is not directly interacting with.

E.g. the stratified model below

- $Y = A + B + C + D + C * D$, when D=1
- $Y = A + B + C + D + C * D$, when D=0

is the same as the following interaction model:

- $Y = A + B + C + D + A * D + B * D + C * D$