Exploration librairie swirl : Multivar Examples 3

```
> swirl()

 Welcome to swirl! Please sign in. If you've been here before, use the same name as you did
then. If you are new, call yourself something unique.

What shall I call you? jlbellier

 Would you like to continue with one of these lessons?

1: Regression Models Least Squares Estimation
2: No. Let me start something new.

Selection: 2

 Please choose a course, or type 0 to exit swirl.

1: Regression Models
2: Statistical Inference
3: Take me to the swirl course repository!

Selection: 1

 Please choose a lesson, or type 0 to return to course menu.

 1: Introduction                            2: Residuals
 3: Least Squares Estimation                4: Residual Variation
 5: Introduction to Multivariable Regression 6: MultiVar Examples
 7: MultiVar Examples2                      8: MultiVar Examples3
 9: Residuals Diagnostics and Variation     10: Variance Inflation Factors
11: Overfitting and Underfitting            12: Binary Outcomes
13: Count Outcomes

Selection: 8

                                                                              0%

 MultiVar_Examples3. (Slides for this and other Data Science courses may be found at github
 https://github.com/DataScienceSpecialization/courses. If you care to use them, they must be
downloaded as a zip file and viewed locally. This lesson corresponds to
Regression_Models/02_02_multivariateExamples.)
                        ===                                                   3%
```

This is the third and final lesson in which we'll look at regression models with more than one independent variable or predictor. We'll begin with WHO hunger data which we've taken the liberty to load for you. WHO is the World Health Organization and this data concerns young children from around the world and rates of hunger among them which the organization compiled over a number of years. The original csv file was very large and we've subsetted just the rows which identify the gender of the child as either male or female. We've read the data into the data frame "hunger" for you, so you can easily access it.

```
                    =====                                                   5%
```

As we did in the last lesson let's first try to get a better understanding of the dataset. Use the R function dim to find the dimensions of hunger.

```
> dim(hunger)
[1] 948  13
```

You are really on a roll!

```
                  ========                                                  8%
```

How many samples does hunger have?

```
> 948
[1] 948
```

Excellent job!

```
                ===========                                                11%
```

Now use the R function names to find out what the 13 columns of hunger represent.

```
> names(hunger)
 [1] "X"              "Indicator"      "Data.Source"    "PUBLISH.STATES" "Year"           "WHO.region"
 [7] "Country"        "Sex"            "Display.Value"  "Numeric"        "Low"            "High"
[13] "Comments"
```

Your dedication is inspiring!

```
               =============                                               14%
```

The Numeric column for a particular row tells us the percentage of children under age 5 who were underweight when that sample was taken. This is one of the columns we'll be focussing on in this lesson. It will be the outcome (dependent variable) for the models we generate.

```
             ================                                              16%
```

Let's first look at the rate of hunger and see how it's changed over time. Use the R function lm to generate the linear model in which the rate of hunger, Numeric, depends on the predictor, Year. Put the result in the variable fit.

```
> fit <- lm(Numeric~Year, hunger)
```

You are doing so well!

```
            ===================                                            19%
```

Now look at the coef portion of the summary of fit.

```
> summary(fit)$coef
             Estimate  Std. Error   t value     Pr(>t)
(Intercept) 634.479660 121.1445995  5.237375 2.007699e-07
Year         -0.308397   0.0605292 -5.095012 4.209412e-07

 Nice work!
                 ====================                                        22%
 What is the coefficient of hunger$Year?

1: 634.47966
2: 0.06053
3: 121.14460
4: -0.30840

Selection: 4

 That's the answer I was looking for.
                ======================                                       24%
 What does the negative Estimate of hunger$Year show?

1: As time goes on, the rate of hunger decreases
2: As time goes on, the rate of hunger increases
3: I haven't a clue

Selection: 1

 Great job!
               ==========================                                    27%
 What does the intercept of the model represent?

1: the number of children questioned in the survey
2: the percentage of hungry children at year 0
3: the number of hungry children at year 0

Selection: 2

 You got it right!
              ============================                                   30%
 Now let's use R's subsetting capability to look at the rates of hunger for the different genders
to see  how, or even if, they differ.  Once again use the R function lm to generate the linear
model in which the  rate of hunger (Numeric) for female children depends on Year. Put the
result in the variable lmF. You'll  have to use the R construct x[hunger$Sex=="Female"] to pick
out both the correct Numerics and the correct  Years.

> lmF <- lm(Numeric[hunger$Sex == "Female"]~Year[hunger$Sex == "Female"], hunger)

 You are amazing!
```

```
          ================================                              32%
```

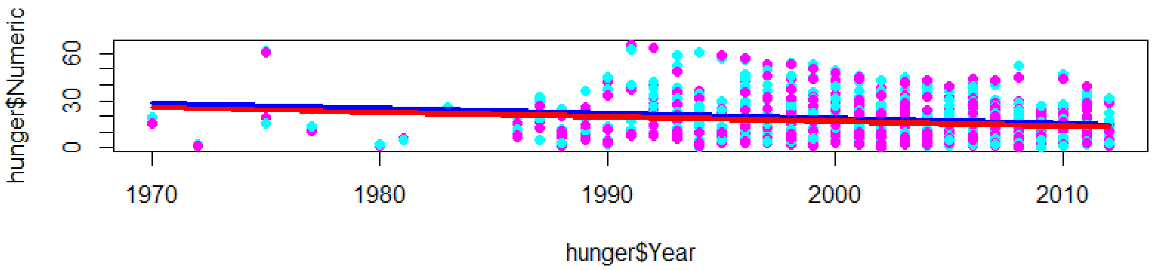
Do the same for male children and put the result in lmM.

```
> lmM <- lm(Numeric[hunger$Sex == "Male"]~Year[hunger$Sex == "Male"], hunger)
```

Keep up the great work!

```
        ====================================                            35%
```

Now we'll plot the data points and fitted lines using different colors to distinguish between males (blue) and females (pink).



```
       ======================================                           38%
```

We can see from the plot that the lines are not exactly parallel. On the right side of the graph (around the year 2010) they are closer together than on the left side (around 1970). Since they aren't parallel, their slopes must be different, though both are negative. Of the following R expressions which would confirm that the slope for males is negative?

1: lmM$coef[1]
2: lmM$coef[2]
3: lmF$coef[2]

Selection: ***2***

Keep working like that and you'll get there!

```
     =========================================                          41%
```

Now instead of separating the data by subsetting the samples by gender we'll use gender as another predictor to create the linear model lmBoth. Recall that to do this in R we place a plus sign "+" between the independent variables, so the formula looks like dependent ~ independent1 + independent2.

```
   ============================================                         43%
```

Create lmBoth now. Numeric is the dependent, Year and Sex are the independent variables. The data is "hunger". For lmBoth, make sure Year is first and Sex is second.

```
> lmBoth <- lm(Numeric~Year+Sex, hunger)
```

You are amazing!

```
  ==============================================                        46%
```

Now look at the summary of lmBoth with the R command summary.

```
> summary(lmBoth)

Call:
lm(formula = Numeric ~ Year + Sex, data = hunger)

Residuals:
    Min      1Q  Median      3Q     Max 
-25.472 -11.297  -1.848   7.058  45.990 

Coefficients:
            Estimate Std. Error t value Pr(>t)
(Intercept) 633.5283   120.8950   5.240 1.98e-07 ***
Year         -0.3084     0.0604  -5.106 3.99e-07 ***
SexMale       1.9027     0.8576   2.219   0.0267 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.2 on 945 degrees of freedom
Multiple R-squared:  0.03175,	Adjusted R-squared:  0.0297 
F-statistic: 15.49 on 2 and 945 DF,  p-value: 2.392e-07

 You nailed it! Good job!
  ===================================================                 49%
 Notice that three estimates are given, the intercept, one for Year and one for Male. What
happened to the estimate for Female? Note that Male and Female are categorical variables
hence they are factors in this model. Recall from the last lesson (and slides) that R treats  the
first (alphabetical) factor as the reference and its estimate is the intercept which  represents
the percentage of hungry females at year 0. The estimate given for the factor Male  is a
distance from the intercept (the estimate of the reference group Female). To calculate  the
percentage of hungry males at year 0 you have to add together the intercept and the male
estimate given by the model.
  ======================================================              51%
 What percentage of young Males were hungry at year 0?

1: 633.2199
2: I can't tell since the data starts at 1970.
3: 635.431
4: 1.9027

Selection: 4

 Not quite! Try again.
 The intercept is the percentage of females hungry at year 0 and the intercept plus
hunger$SexMale is the percentage of males hungry at year 0.

1: 633.2199
2: 1.9027
```

3: I can't tell since the data starts at 1970.
4: 635.431

*Selection: **4***

Your dedication is inspiring!

```
  ==========================================================                    54%
```

What does the estimate for hunger$Year represent?

1: the annual decrease in percentage of hungry males
2: the annual decrease in percentage of hungry females
3: the annual decrease in percentage of hungry children of both genders

*Selection: **2***

Keep trying!
The model looked at all the data and didn't specify which gender to consider.

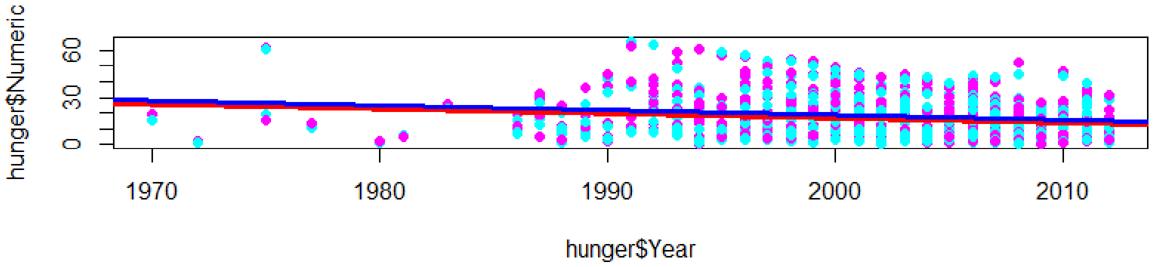
1: the annual decrease in percentage of hungry males
2: the annual decrease in percentage of hungry females
3: the annual decrease in percentage of hungry children of both genders

*Selection: **3***

You got it right!

```
  ============================================================                  57%
```

Now we'll replot the data points along with two new lines using different colors. The red line will have the female intercept and the blue line will have the male intercept.



```
 ===============================================================               59%
```

The lines appear parallel. This is because

1: they have slopes that are very close
2: they have the same slope
3: I have no idea

*Selection: **2***

That's correct!

```
  ==============================================================               62%
```

Now we'll consider the interaction between year and gender to see how that affects changes in rates of hunger. To do this we'll add a third term to the predictor portion of our model formula, the product of year and gender.

```
  =================================================================            65%
```

Create the model lmInter. Numeric is the outcome and the three predictors are Year, Sex, and Sex*Year. The data is "hunger".

```
> lmInter <- lm(Numeric~Year*Sex, hunger)
```

You nailed it! Good job!

```
  ======================================================================       68%
```

Now look at the summary of lmInter with the R command summary.

```
> summary(lmInter)
```

```
Call:
lm(formula = Numeric ~ Year * Sex, data = hunger)

Residuals:
    Min      1Q  Median      3Q     Max 
-25.913 -11.248  -1.853   7.087  46.146 

Coefficients:
              Estimate Std. Error t value Pr(>t)
(Intercept)  603.50580  171.05519   3.528 0.000439 ***
Year          -0.29340    0.08547  -3.433 0.000623 ***
SexMale       61.94772  241.90858   0.256 0.797946
Year:SexMale  -0.03000    0.12087  -0.248 0.804022
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.21 on 944 degrees of freedom
Multiple R-squared:  0.03181,	Adjusted R-squared:  0.02874 
F-statistic: 10.34 on 3 and 944 DF,  p-value: 1.064e-06
```

Your dedication is inspiring!

```
 =======================================================================       70%
```

What is the percentage of hungry females at year 0?

```
1: 61.94772
2: 603.5058
3: The model doesn't say.

Selection: 2
```

Excellent work!

=== 73%

What is the percentage of hungry males at year 0?

1: 603.5058
2: The model doesn't say.
3: 665.4535
4: 61.94772

Selection: ***3***

You are quite good my friend!

=== 76%

What is the annual change in percentage of hungry females?

1: -0.03000
2: The model doesn't say.
3: 0.08547
4: -0.29340

Selection: ***4***

You're the best!

=== 78%

What is the annual change in percentage of hungry males?

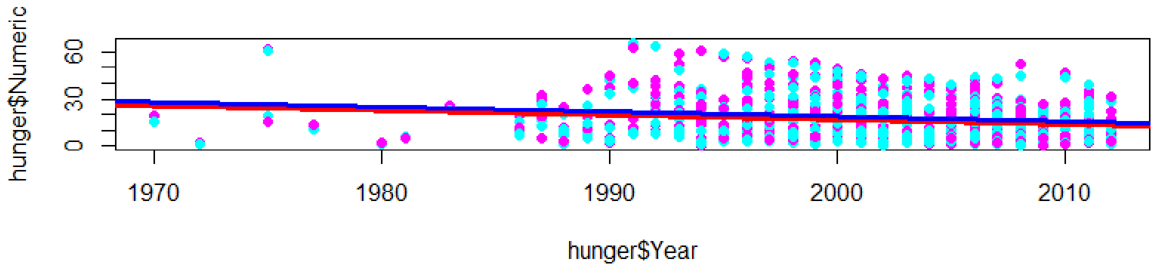
1: -0.32340
2: 0.12087
3: The model doesn't say.
4: -0.03000

Selection: ***1***

Great job!

=== 81%

Now we'll replot the data points along with two new lines using different colors to distinguish between the genders.



=== 84%

Which line has the steeper slope?

1: Female
2: They look about the same
3: Male

Selection: **3**

Great job!

== **86%**

Finally, we note that things are a little trickier when we're dealing with an interaction between predictors which are continuous (and not factors). The slides show the underlying algebra, but we can summarize.

== **89%**

Suppose we have two interacting predictors and one of them is held constant. The expected change in the outcome for a unit change in the other predictor is the coefficient of that changing predictor + the coefficient of the interaction * the value of the predictor held constant.

== **92%**

Suppose the linear model is Hi = b0 + (b1*Ii) + (b2*Yi)+ (b3*Ii*Yi) + ei. Here the H's represent the outcomes, the I's and Y's the predictors, neither of which is a category, and the b's represent the estimated coefficients of the predictors. We can ignore the e's which represent the residuals of the model. This equation models a continuous interaction since neither I nor Y is a category or factor. Suppose we fix I at some value and let Y vary.

== **95%**

Which expression represents the change in H per unit change in Y given that I is fixed at 5?

1: b0+b2
2: b1+5*b3
3: b2+b3*5
4: b2+b3*Y

Selection: **3**

Nice work!

== **97%**

Congratulations! You've finished this final lesson in multivariable regression models.

=== **100%**