Exploration librairie swirl : Multivar Examples

```
> 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: 6
                                                                          0%

 MultiVar_Examples. (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.)
                              ====                                        4%
 In this lesson, we'll look at some examples of regression models with more than one variable.
```

We'll begin with the Swiss data which we've taken the liberty to load for you. This data is part of R's datasets package. It was gathered in 1888, a time of demographic change in Switzerland, and measured six quantities in 47 French-speaking provinces of Switzerland. We used the code from the slides (the R function pairs) to display here a 6 by 6 array of scatterplots showing pairwise relationships between the variables. All of the variables, except for fertility, are proportions of population. For example, "Examination" shows the percentage of draftees receiving the highest mark on an army exam, and "Education" the percentage of draftees with education beyond primary school.

```
                    =========                                       9%
```

From the plot, which is NOT one of the factors measured?

1: Infant Mortality
2: Fertility
3: Catholic
4: Obesity

Selection: 4

You are really on a roll!

```
                 =============                                      13%
```

First, use the R function lm to generate the linear model "all" in which Fertility is the variable dependent on all the others. Use the R shorthand "." to represent the five independent variables in the formula passed to lm. Remember the data is "swiss".

```
> all <- lm(Fertility~.,swiss)
```

Nice work!

```
               =================                                    17%
```

Now look at the summary of the linear model all.

```
> summary(all)

Call:
lm(formula = Fertility ~ ., data = swiss)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.2743  -5.2617   0.5032   4.1198  15.3213 

Coefficients:
                 Estimate Std. Error t value Pr(>t)
(Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
Agriculture      -0.17211    0.07030  -2.448  0.01873 *
Examination      -0.25801    0.25388  -1.016  0.31546
Education        -0.87094    0.18303  -4.758 2.43e-05 ***
Catholic          0.10412    0.03526   2.953  0.00519 **
Infant.Mortality  1.07705    0.38172   2.822  0.00734 **
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.165 on 41 degrees of freedom
Multiple R-squared:  0.7067,            Adjusted R-squared:  0.671
F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10

 You got it right!
                    ======================                                  22%
 Recall that the Estimates are the coeffients of the independent variables of the linear model
 (all of which are percentages) and they reflect an estimated change in the dependent
variable  (fertility) when the corresponding independent variable changes. So, for every 1%
increase in  percent of males involved in agriculture as an occupation we expect a .17
decrease in  fertility, holding all the other variables constant; for every 1% increase in
Catholicism, we  expect a .10 increase in fertility, holding all other variables constant.
                  ==========================                                26%
 The "*" at the far end of the row indicates that the influence of Agriculture on Fertility is
 significant. At what alpha level is the t-test of Agriculture significant?

1: 0.01
2: 0.1
3: R doesn't say
4: 0.05

Selection: 1

 Not quite! Try again.
 Look at the "Signif. codes" line in the summary output.

1: R doesn't say
2: 0.01
3: 0.1
4: 0.05

Selection: 4

 Keep up the great work!
                ==============================                              30%
 Now generate the summary of another linear model (don't store it in a new variable) in
which  Fertility depends only on agriculture.

> summary(lm(Fertility~Agriculture,swiss))

Call:
lm(formula = Fertility ~ Agriculture, data = swiss)

Residuals:
    Min      1Q  Median      3Q     Max 
-25.5374  -7.8685  -0.6362   9.0464  24.4858 

Coefficients:
           Estimate Std. Error t value Pr(>t)    
(Intercept) 60.30438    4.25126  14.185   <2e-16 ***
Agriculture  0.19420    0.07671   2.532   0.0149 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.82 on 45 degrees of freedom
Multiple R-squared:  0.1247,              Adjusted R-squared:  0.1052 
F-statistic: 6.409 on 1 and 45 DF,  p-value: 0.01492

 You got it right!
               ===================================                         35%
 What is the coefficient of agriculture in this new model?

1: *
2: 60.30438
3: 0.19420
4: 0.07671

Selection: 3

 All that hard work is paying off!
          ========================================                         39%
 The interesting point is that the sign of the Agriculture coefficient changed from negative
(when all the variables were included in the model) to positive (when the model only
considered Agriculture). Obviously the presence of the other factors affects the influence
Agriculture has on Fertility.
      ============================================                         43%
 Let's consider the relationship between some of the factors. How would you expect level
 Education and performance on an Examination to be related?

1: They would be correlated
2: I would not be able to guess without more information
3: They would be uncorrelated

Selection: 1

 You nailed it! Good job!
  ================================================                         48%
 Now check your intuition with the R command "cor". This computes the correlation between
 Examination and Education.

> cor(Examination,Education)
```

```
Error in is.data.frame(y) : object 'Education' not found
> cor(swiss$Examination,swiss$Education)
[1] 0.6984153
```

You got it right!

```
  =======================================================                    52%
```

The correlation of .6984 shows the two are correlated. Now find the correlation between Agriculture and Education.

```
> cor(swiss$Agriculture,swiss$Education)
[1] -0.6395225
```

Excellent job!

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

The negative correlation (-.6395) between Agriculture and Education might be affecting Agriculture's influence on Fertility. I've loaded and sourced the file swissLMs.R in your working directory. In it is a function makelms() which generates a sequence of five linear models. Each model has one more independent variable than the preceding model, so the first has just one independent variable, Agriculture, and the last has all 5. I've tried loading the source code in your editor. If I haven't done this, open the file manually so you can look at the code.

```
 ===============================================================             61%
```

Now run the function makelms() to see how the addition of variables affects the coefficient of Agriculture in the models.

```
> makelms()
Agriculture Agriculture Agriculture Agriculture Agriculture
 0.1942017   0.1095281  -0.2030377  -0.2206455  -0.1721140
```

Perseverance, that's the answer.

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

The addition of which variable changes the sign of Agriculture's coefficient from positive to negative?

```
1: Infant.Mortality
2: Education
3: Catholic
4: Examination

Selection: 2
```

That's a job well done!

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

Now we'll show what happens when we add a variable that provides no new linear information to a model. Create a variable ec that is the sum of swiss$Examination and swiss$Catholic.

```
> ec <- swiss$Examination + swiss$Catholic
```

```
 You are really on a roll!
  ===================================================================  74%
 Now generate a new model efit with Fertility as the dependent variable and the remaining 5
of  the original variables AND ec as the independent variables. Use the R shorthand ". + ec"
for  the righthand side of the formula.

> efit <- lm(Fertility~. + ec,swiss)

 Your dedication is inspiring!
====================================================================   78%
 We'll see that R ignores this new term since it doesn't add any information to the model.
====================================================================   83%
 Subtract the efit coefficients from the coefficients of the first model you created, all.

> all$coefficients-efit$coefficients
    (Intercept)     Agriculture     Examination       Education       Catholic
             0               0               0               0              0
Infant.Mortality              ec
             0              NA
Warning message:
In all$coefficients - efit$coefficients :
  longer object length is not a multiple of shorter object length

 Excellent work!
====================================================================   87%
 Which is the coefficient of ec?

1: I haven't a clue.
2: 0
3: NA

Selection: 3

 You are quite good my friend!
====================================================================   91%
 This tells us that

1: R is really cool
2: Adding ec zeroes out the coefficients
3: Adding ec doesn't change the model

Selection: 3

 You got it!
====================================================================   96%
 Congrats! You've concluded this first lesson on multivariable linear models.
```

```
==================================================================================  100%
 Would you like to receive credit for completing this course on Coursera.org?

1: Yes
2: No

Selection: 1
What is your email address? jeanluc.bellier@sfr.fr
What is your assignment token? fWLPLK2rfXTmqioF
Grade submission succeeded!
 That's the answer I was looking for.
 You've reached the end of this lesson! Returning to the main menu...
```