# Problem Set 1:Getting Started Key

Claire Duquennois

**Group Member 1: Your name here** Tory Burford
**Group Member 2: Your name here** Tesfa Asefa **Group Member 3: Your name here** Shreiya Venkatesan

```
#This is a comment, alone in a chunk.
```

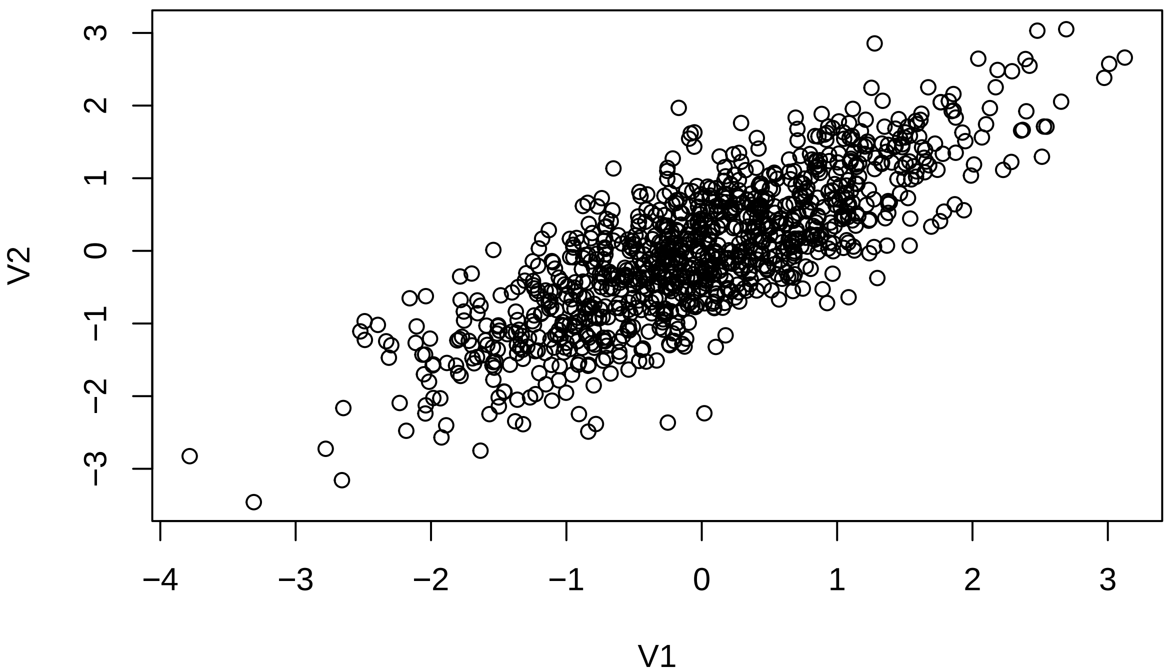
In the following chunk I generate simulated data consisting of two correlated variables. Notice the chunk setting is set to suppress the warning messages that are generated when R loads a package.

```
library(MASS)
library(ggplot2)

out <- as.data.frame(mvrnorm(1000, mu = c(0,0),
                     Sigma = matrix(c(1,0.8,0.8,1), ncol = 2),
                     empirical = TRUE))
cor(out)
```

```
##     V1  V2
## V1 1.0 0.8
## V2 0.8 1.0
```

```
plot(out)
```




Next I generate a randomly distributed error term and I calculate the outcome variable which depends on both V1 and V2 and some noise:

$$Y = \beta_1 V_1 + \beta_2 V_2 + \epsilon$$

```
out$error<-rnorm(1000, mean=0, sd=1)

#The data generating process
B1<-3
B2<-6

out$Y<-out$V1*B1+out$V2*B2+out$error
```

**TO DO: For the questions below write the needed code and a written response to the question.**

## 0.1 Question:

**Write a chunk in which you regress $Y$ on $V_1$ and $V_2$. Are your estimates of $\beta_1$ and $\beta_2$ biased?**

```
model<-lm(Y~V1+V2, out)
summary(model)
```

```
##
## Call:
## lm(formula = Y ~ V1 + V2, data = out)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2102 -0.6588 -0.0278  0.6620  3.7297
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.02991    0.03146   0.951    0.342
## V1           2.98524    0.05246  56.904   <2e-16 ***
## V2           6.02996    0.05246 114.942   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9949 on 997 degrees of freedom
## Multiple R-squared:  0.9868, Adjusted R-squared:  0.9868
## F-statistic: 3.738e+04 on 2 and 997 DF,  p-value: < 2.2e-16
```

**Answer: Our results are not biased because we have accounted for all variables that affect Y.**

## 0.2 Question:

**Write a chunk in which you regress $Y$ on $V_1$ only. Is your estimate of $\beta_1$ biased?**

```
model2<-lm(Y~V1, out)
summary(model2)
```

```
##
## Call:
## lm(formula = Y ~ V1, data = out)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.3464  -2.7012   0.0199   2.6002  12.5871
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.02991    0.11871   0.252    0.801
## V1           7.80921    0.11877  65.752   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.754 on 998 degrees of freedom
## Multiple R-squared:  0.8125, Adjusted R-squared:  0.8123
## F-statistic:  4323 on 1 and 998 DF,  p-value: < 2.2e-16
```

**Answer: Our estimate of $\beta_1$ is biased because we failed to account for the second variable. This means that $\beta_1$ is capturing some of the effect from $V_2$**

## 0.3 Question:

**Generate a new variable $Y_{adj}$ such that $Y_{adj} = Y - \beta_2 * V_2$. Then regress $Y_{adj}$ on $V_1$. Is your estimate of $\beta_1$ biased? Can you explain why/why not?**

```
out$Y_adj<-out$Y-6*out$V2
model3<-lm(Y_adj~V1, out)
summary(model3)
```

```
##
## Call:
## lm(formula = Y_adj ~ V1, data = out)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1888 -0.6528 -0.0304  0.6544  3.7445
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.02991    0.03145   0.951    0.342
## V1           3.00921    0.03147  95.634   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9945 on 998 degrees of freedom
## Multiple R-squared:  0.9016, Adjusted R-squared:  0.9015
## F-statistic:  9146 on 1 and 998 DF,  p-value: < 2.2e-16
```

**Answer: Our estimate of $\beta_1$ is not biased because we accounted for V2 before we ran the regression, which means $\beta_1$ for Y_adjusted is only capturing the effect of V1.**

## 0.4 Question:

**The code below generates a scatter plot and regression line for the relationship between $V_1$ and $Y$ as well as $V_1$ and $Y_{adj}$. Submit an improved visualization of this data. Hint: you will need to delete the # to get the code to run**

```
## create sample for the graph
install.packages("dplyr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:MASS':
##
##     select

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

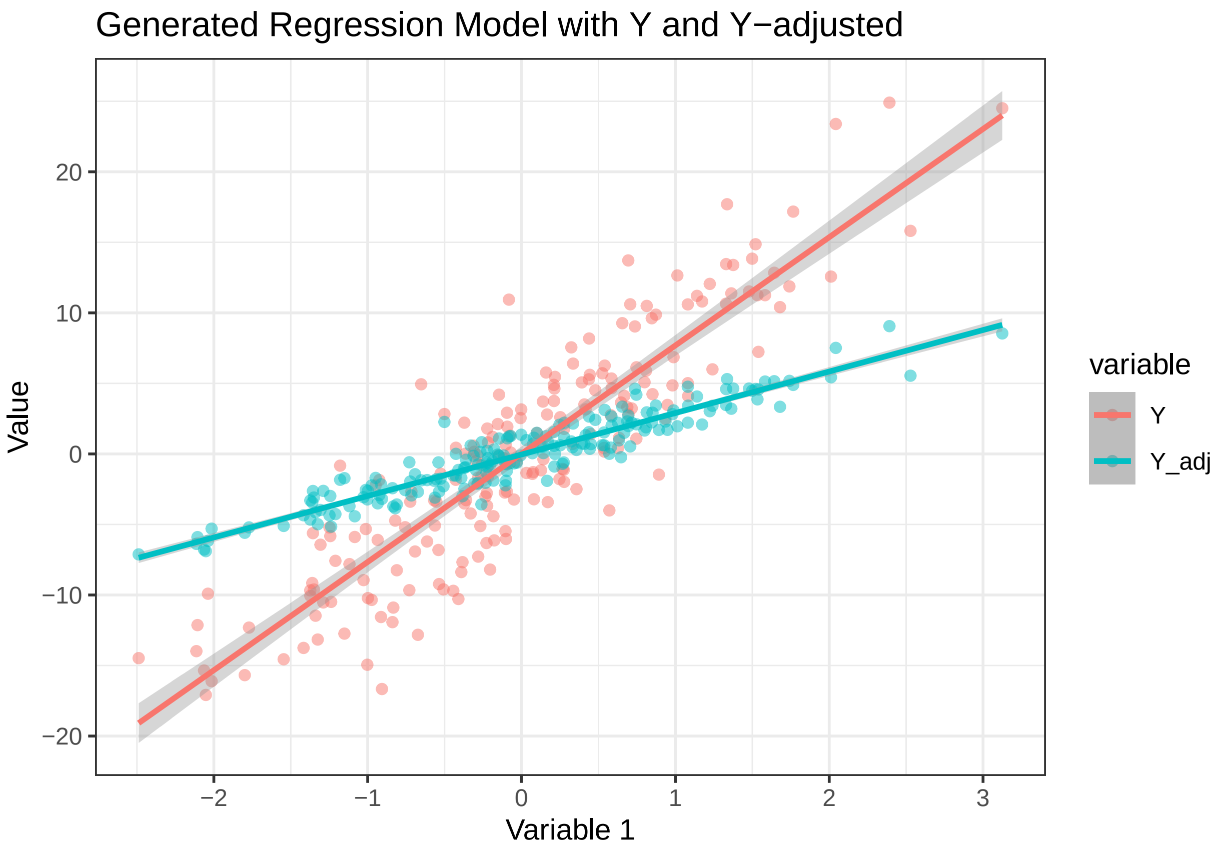
```r
samp <- sample_n(out, 200)

plotted<-ggplot(samp, aes(V1, y = value, color = variable)) +
    geom_point(aes(y = Y, col = "Y"),  alpha=.5) +
    geom_point(aes(y = Y_adj, col = "Y_adj"), alpha=.5)+
    geom_smooth(method='lm', aes(y = Y, col = "Y"))+
    geom_smooth(method='lm', aes(y = Y_adj, col = "Y_adj")) + theme_bw() + labs(x="Variable 1", y="Valu

plotted
```

```
## `geom_smooth()` using formula 'y ~ x'

## `geom_smooth()` using formula 'y ~ x'
```




**Answer: To clean up the graph, we took a sample of 200 from the data set and plotted that sample with the regression lines. This kept the points from cluttering up the graph. We made the points more transparent so that the trend lines would stand out. We also added a title, as well as labels for the X and Y axes.**

### 0.5 Question:

**Load the 'cps_clean.csv' dataset (available on Canvas). Regress income on education and interpret the coefficient.** «««< HEAD

```
install.packages("readr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
library(readr)
cps <- read_csv('cps_clean.csv')
```

```
## New names:
## * `` -> `...1`

## Rows: 5000 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (9): ...1, age, female, married, employed, edu, ftotval, inctot, health
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
new<-lm(inctot~edu, cps)
summary(new)
```

```
##
## Call:
## lm(formula = inctot ~ edu, data = cps)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -105468  -31110  -11435   12596 1071106
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -64942.0     5084.1  -12.77   <2e-16 ***
## edu           8114.7      359.2   22.59   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 70930 on 4998 degrees of freedom
## Multiple R-squared:  0.09265,    Adjusted R-squared:  0.09247
## F-statistic: 510.4 on 1 and 4998 DF,  p-value: < 2.2e-16
```

======= **Answer: $\beta_1$=8114.7 This means that for every 1 year increase in education, the regression predicts and increase in income of $8114.7**

## 0.6 Question:

**Add additional control variables to the regression you estimated above. How does this change your interpretation of the coefficient on education?**

```r
model3<-lm(inctot~edu+age+female+health, cps)
summary(model3)
```

```
##
## Call:
## lm(formula = inctot ~ edu + age + female + health, data = cps)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -119742  -27920   -8100   13004 1061392
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -71318.78    6409.66 -11.127  < 2e-16 ***
## edu           7754.69     358.36  21.639  < 2e-16 ***
## age            948.74      80.08  11.848  < 2e-16 ***
## female      -28583.74    1939.13 -14.741  < 2e-16 ***
## health       -5921.98     982.68  -6.026  1.8e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68450 on 4995 degrees of freedom
## Multiple R-squared:  0.1553, Adjusted R-squared:  0.1546
## F-statistic: 229.5 on 4 and 4995 DF,  p-value: < 2.2e-16
```

**Answer: We added the control variables age, female, and health. With these control variables**

**the new $\beta_1$ for education is 7754.69. This is smaller than our original estimate, suggesting that the control variables were adding a sum upward bias in the original regression.**

# 1 Submission instructions:

1) Make sure the final version of your assignment is uploaded on GitHub in both html and Rmarkdown format.