Problem Set 1

Professor: Dr. Claire Duquennois

Group Member 1: Chean Shea

Group Member 2: Mayowa Idowu

Group Member 3: Milan Stefanelli

Group Member 4: Kiersten Kochanowski

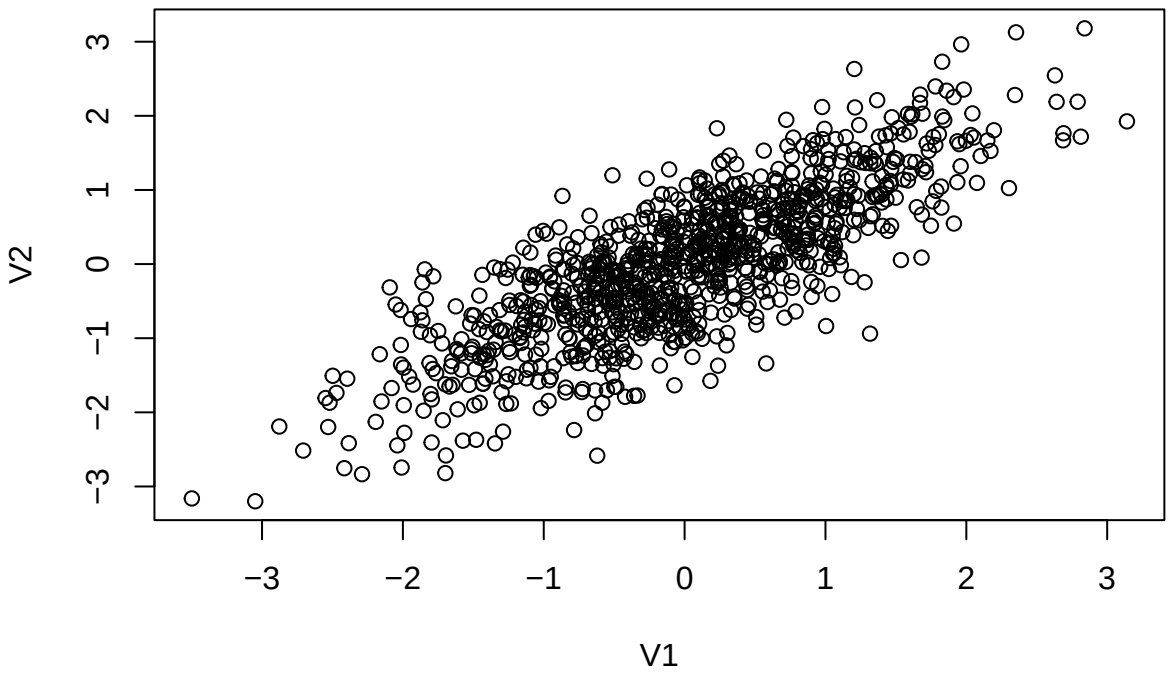
Omitted Variable Bias Simulation (Group excercise)

```
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.

Question 1: Write a chunk in which you regress Y on V_1 and V_2. Are your estimates of β_1 and β_2 biased?

Answer: No, our data was randomly generated (within parameters) so their coefficients are not bias. This is also affirmed by the fact that the V1 and V2 coefficients in our regression model are almost identical to the one B1 and B2 generated above.

```
View(out)
M1 <- lm(data=out, Y ~ V1 + V2)

stargazer(M1, type="text", header=FALSE, title="Regression Coefficients",
```

```
                      style="qje",
                      covariate.labels = c("B1", "B2", "Intercept"),
                      column.labels = c("Model #1"))
```

```
##
## Regression Coefficients
## ================================================
##                                   Y
##                                Model #1
## ------------------------------------------------
## B1                             2.984***
##                                (0.052)
##
## B2                             6.057***
##                                (0.052)
##
## Intercept                       0.039
##                                (0.031)
##
## N                               1,000
## R2                              0.987
## Adjusted R2                     0.987
## Residual Std. Error        0.983 (df = 997)
## F Statistic            38,488.890*** (df = 2; 997)
## ================================================
## Notes:              ***Significant at the 1 percent level.
##                      **Significant at the 5 percent level.
##                      *Significant at the 10 percent level.
```

Question 2: Write a chunk in which you regress Y on V_1 only. Is your estimate of β_1 biased?

Answer: Yes, our beta 1 (V1) coefficient is biased due to omitted variable bias. By omitting V2, our beta 1 coefficient has upward bias (larger than it would be with V2 in the model: 7.81 vs 2.93).

```
M2 <- lm(data=out, Y~V1)

library(stargazer)
stargazer(M2, M1, type="text", header=FALSE, title="Comparison of Regression Coefficients (M2 & M1)",
                      style="qje",
                      covariate.labels = c("B1", "B2", "Intercept"),
                      column.labels = c("Model #2", "Model #1"))
```

```
##
## Comparison of Regression Coefficients (M2 & M1)
## ============================================================================
##                                          Y
##                             Model #2                  Model #1
##                               (1)                       (2)
## ----------------------------------------------------------------------------
## B1                          7.830***                  2.984***
##                             (0.119)                   (0.052)
##
## B2                                                    6.057***
```

```
##                                                    (0.052)
##
## Intercept                     0.039                 0.039
##                              (0.119)               (0.031)
##
## N                             1,000                 1,000
## R2                            0.812                 0.987
## Adjusted R2                   0.812                 0.987
## Residual Std. Error      3.767 (df = 998)      0.983 (df = 997)
## F Statistic         4,317.193*** (df = 1; 998) 38,488.890*** (df = 2; 997)
## ============================================================================
## Notes:                              ***Significant at the 1 percent level.
##                                      **Significant at the 5 percent level.
##                                      *Significant at the 10 percent level.
```

Question 3: 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 β_1 biased? Can you explain why/why not?

Answer: No, beta 1 is not biased in this model. We can see that the coefficient is very similar to model #1 where beta two was also included (3.01 vs 2.93). We can also affirm this mathematically given the relationship below

Y-B2V2 = B1V1 (Model #3), which is equal to Y=B1V1 + B2V2 (Model #1)

```
Y_adj <- ((out$Y)-(B2*out$V2))
M3 <- lm(data=out, Y_adj ~ V1)
stargazer(M3, M1, type="text", header=FALSE, title="Comparison of Regression Coefficients (M3 & M1)",
                      style="qje",
                      covariate.labels = c("B1", "B2", "Intercept"),
                      column.labels = c("Model #3", "Model #1"))
```

```
##
## Comparison of Regression Coefficients (M3 & M1)
## ============================================================================
##                               Y_adj                        Y
##                              Model #3                  Model #1
##                                (1)                       (2)
## ----------------------------------------------------------------------------
## B1                           3.030***                  2.984***
##                              (0.031)                   (0.052)
##
## B2                                                     6.057***
##                                                        (0.052)
##
## Intercept                     0.039                     0.039
##                              (0.031)                   (0.031)
##
## N                             1,000                     1,000
## R2                            0.905                     0.987
## Adjusted R2                   0.905                     0.987
## Residual Std. Error      0.983 (df = 998)          0.983 (df = 997)
## F Statistic         9,482.001*** (df = 1; 998) 38,488.890*** (df = 2; 997)
## ============================================================================
## Notes:                              ***Significant at the 1 percent level.
```

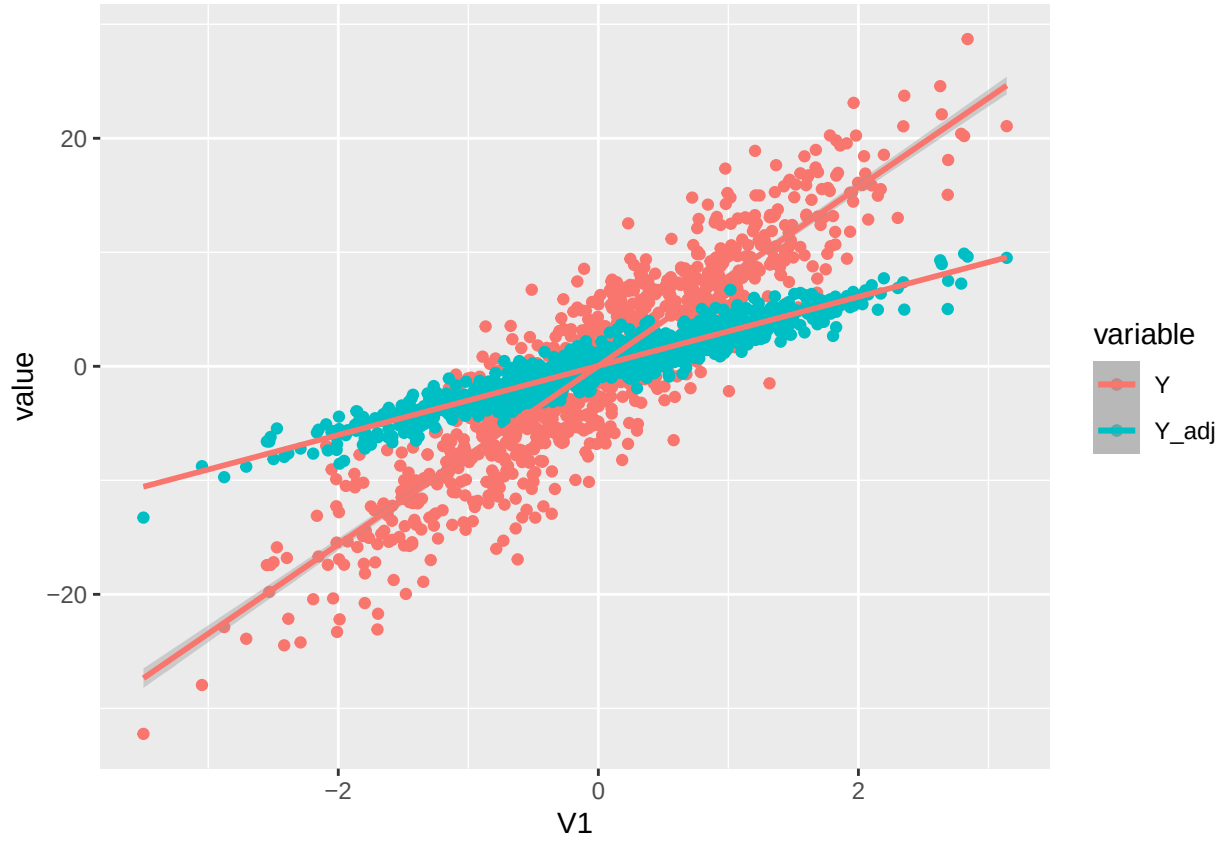
```
##                                    **Significant at the 5 percent level.
##                                    *Significant at the 10 percent level.
```

Question 4: 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

```
plotted<-ggplot(out, aes(V1, y = value, color = variable)) +
   geom_point(aes(y = Y, col = "Y")) +
   geom_point(aes(y = Y_adj, col = "Y_adj"))+
   geom_smooth(method='lm', aes(y = Y, col = "Y"))+
   geom_smooth(method='lm', aes(y = Y_adj, col = "Y"))

plotted
```

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

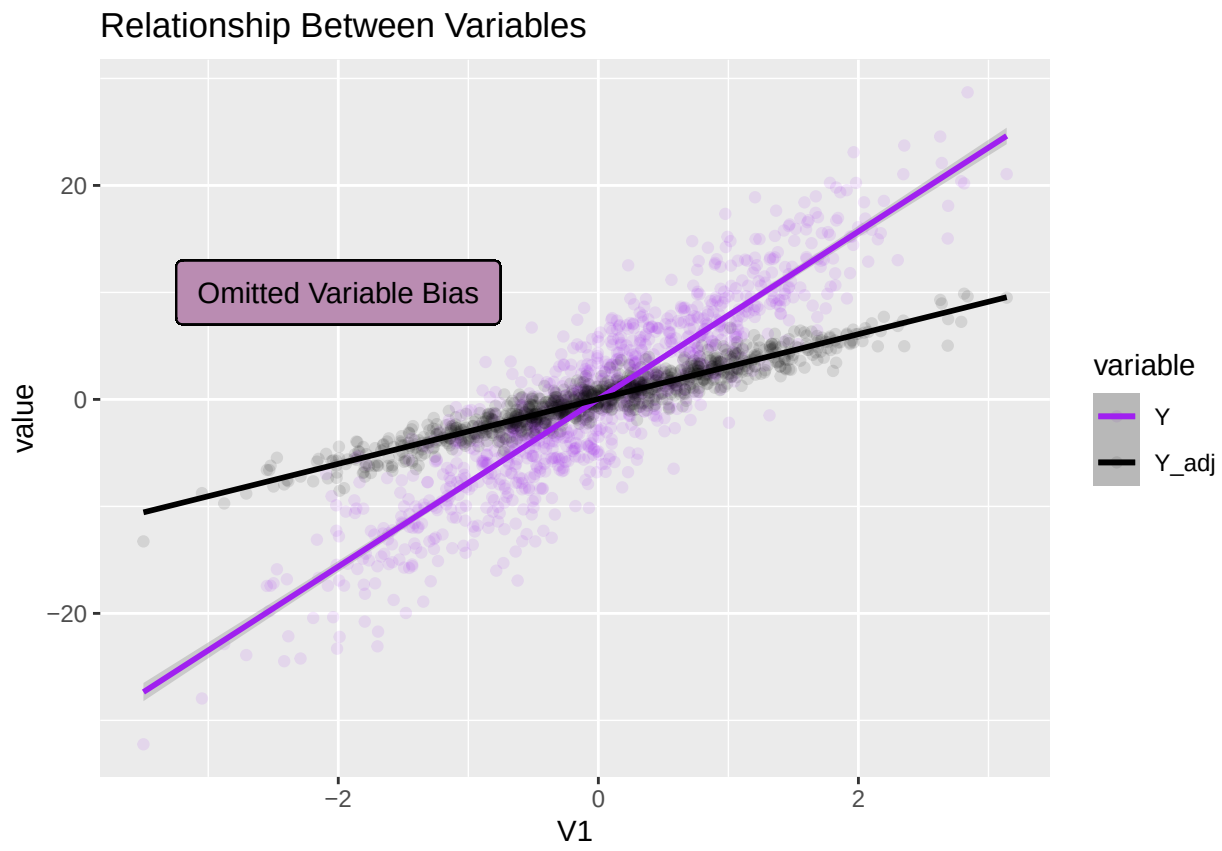


New Data Visualizations We interpreted the provided visual to be describing the relationship between V1 and Y with and without omitted variable bias (OVB). Our first visual makes this argument more clear by indicating which LOBF was influenced by OVB, using the same scatter plot method.

```
plotted<-ggplot(out, aes(V1, y = value, color = variable)) +
   geom_point(aes(y = Y, col = "Y"), alpha = .1) +
   geom_point(aes(y = Y_adj, col = "Y_adj"), alpha = .1)+
   geom_smooth(method='lm', aes(y = Y, col = "Y"))+
   geom_smooth(method='lm', aes(y = Y_adj, col = "Y_adj")) +
   scale_color_manual(values=c("Purple","Black")) +
   labs(title = "Relationship Between Variables") +
   geom_label(
    label="Omitted Variable Bias",
    x=-2,
    y=10,
    label.padding = unit(0.55, "lines"), # Rectangle size around label
    label.size = 0.35,
    color = "black",
    fill="#BA8CB2"
  )

plotted
```

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



Our second visual takes an even simpler approach by illustrating the difference in the V1 coefficient (and its relationship to Y) when it is and isn't biased by missing variable V2.

```
data_vis <- plot_coefs(M2, M3, omit.coefs =c("(Intercept)","V2"), model.names=c("OVB", "No OVB"))
data_vis2 <- data_vis + coord_flip()
data_vis2
```