R Notebook

Problem 1: Seat belts 1. For this problem we are trying to test the effectiveness of mandatory seat belts usage laws in reducing traffic mortality. The independent variable (Y) is fatalityrate. Run all your regressions using the lm parameter. You need to download the seatbelts dataset to complete this part.

```
seat_belts_data <- as.data.frame(read.csv(file = "~/Downloads/seatbelts/seatbelts.csv", header = T, str
str(seat_belts_data)
```

```
## 'data.frame':    765 obs. of  14 variables:
##  $ state       : chr  "AK" "AK" "AK" "AK" ...
##  $ year        : int  1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 ...
##  $ fips        : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ vmt         : int  3358 3589 3840 4008 3900 3841 3887 3979 4021 3841 ...
##  $ fatalityrate: num  0.0447 0.0373 0.0331 0.0252 0.0195 ...
##  $ sb_useage   : num  NA NA NA NA NA NA NA 0.45 0.66 0.66 ...
##  $ speed65     : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ speed70     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ drinkage21  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ ba08        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ income      : int  17973 18093 18925 18466 18021 18447 19970 21073 21496 22073 ...
##  $ age         : num  28.2 28.3 28.4 28.4 28.5 ...
##  $ primary     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ secondary   : int  0 0 0 0 0 0 0 0 1 1 ...
```

1.1 Run an interpret the bivariate regression of fatalityrate on primary (this is a binary variable that indicates the primary enforcement of seat belt laws).

```
lm1 <- lm(formula = fatalityrate ~ primary, data = seat_belts_data)
summary(lm1)
```

```
##
## Call:
## lm(formula = fatalityrate ~ primary, data = seat_belts_data)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0133714 -0.0040909 -0.0003789  0.0032309  0.0237715
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0216986  0.0002372  91.468   <2e-16 ***
## primary     -0.0017203  0.0006804  -2.528   0.0117 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.00615 on 763 degrees of freedom
## Multiple R-squared:  0.008309,   Adjusted R-squared:  0.007009
## F-statistic: 6.393 on 1 and 763 DF,  p-value: 0.01166
```

Interpretation: From the bivariate regression of fatalityrate on primary we can see that the primary enforce-

Table 1: Correlation Matrix

	year	fips	vmt	fatalityrate	sb_useage	speed65	speed70	drinkage21	
year	1.0000000	0.0000000	0.1232604	-0.5590983	0.4855162	0.6718047	0.3827189	0.4968923	
fips	0.0000000	1.0000000	-0.0695592	-0.0873099	0.0293668	0.0281335	-0.0022405	-0.0421936	
vmt	0.1232604	-0.0695592	1.0000000	-0.1613661	0.1864113	0.0822517	0.1196134	0.1079103	
fatalityrate	-0.5590983	-0.0873099	-0.1613661	1.0000000	-0.2797353	-0.2818366	-0.0764641	-0.2937547	-
sb_useage	0.4855162	0.0293668	0.1864113	-0.2797353	1.0000000	0.2382442	0.1953355	0.1904076	
speed65	0.6718047	0.0281335	0.0822517	-0.2818366	0.2382442	1.0000000	0.2041189	0.4782063	
speed70	0.3827189	-0.0022405	0.1196134	-0.0764641	0.1953355	0.2041189	1.0000000	0.0993594	
drinkage21	0.4968923	-0.0421936	0.1079103	-0.2937547	0.1904076	0.4782063	0.0993594	1.0000000	
ba08	0.2500599	0.0841268	0.0977319	-0.1698376	0.2076247	0.1920302	0.2183439	0.1308183	
income	0.7814340	-0.1456004	0.2061512	-0.7035575	0.4876521	0.3616334	0.2090382	0.4155422	
age	0.3704737	0.0102775	0.0834575	-0.3754130	0.1148603	0.1889589	0.0296954	0.2034638	-
primary	0.1360999	0.0149645	0.1365343	-0.0911546	0.3873824	-0.0339131	-0.0088191	0.1341232	
secondary	0.5567256	-0.0362128	0.1656742	-0.3222588	0.2171536	0.5371275	0.2168779	0.3408620	

ment of seat belt laws has a negative effect on the fatalityrate. In this, the F statistic is not significant (p = 0.0117) which indicates that the fit of intercept of the model and that of the current model is same.

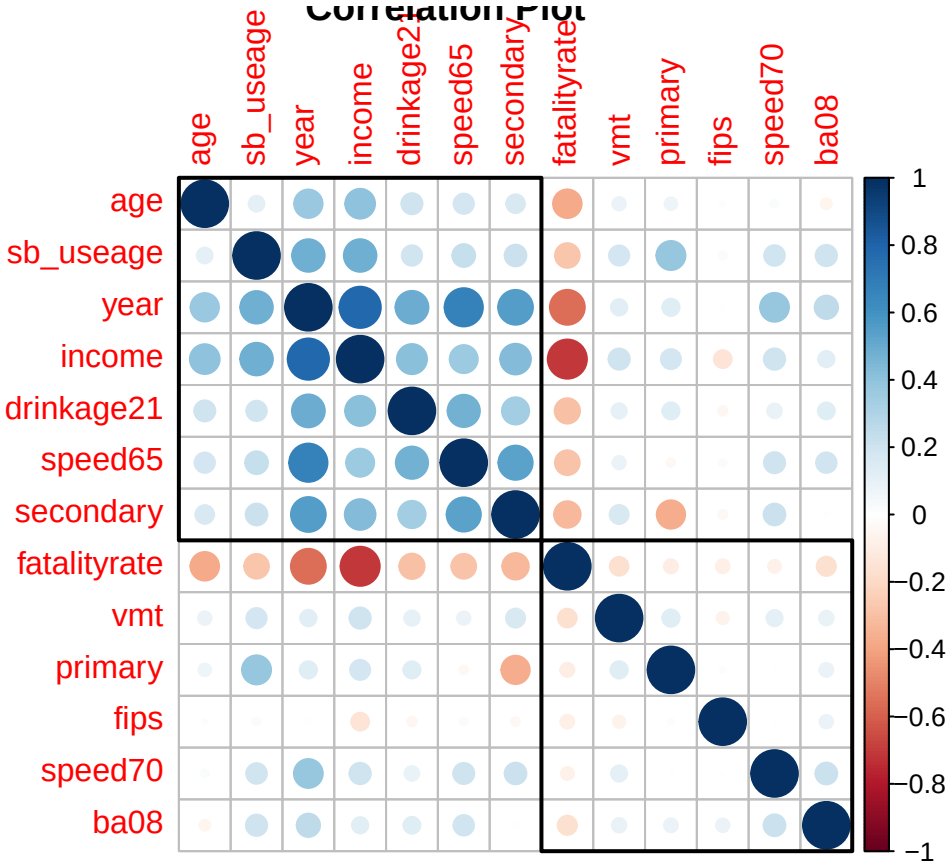
1.2 Create a correlation matrix for the entire dataset using the cor command - exclude non-numeric variables -. Do you think that the exogeneity assumption may not be satisfied for the previous regression? (Explain)

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
library(kableExtra)
seat_belts_data$sb_useage[is.na(seat_belts_data$sb_useage)] <- mean(seat_belts_data$sb_useage[!is.na(sea
seatbelts_numeric <- seat_belts_data[ ,sapply(seat_belts_data, is.numeric)]
cor_plot <- cor(seatbelts_numeric, method = "pearson")
kable(cor_plot, caption = "Correlation Matrix")
```

```
corrplot(cor_plot, order = "hclust", addrect = 2, title = "Correlation Plot")
```



Interpretation: Exogenity is an assumption made in regression analysis wherein it states that an independent variable X is not dependant on the dependent variable Y. However this does not indicate that there is no connection. Since Y is a dependent variable it is dependent to the variable X and the error term (Source: https://www.statisticshowto.datasciencecentral.com/exogeneity/). Exogenity is required because if the independent variable is not independent of the error term and the dependant variable Y, then the regression coefficients are not consistant. Hence, exogenity assumption may not be satisfied for the previous regression.

1.3 Using the dataset provided, run a set of 3 additional multiple regressions by sequentially adding other variables that you think are relevant in the model. For each regression (1) Argue why you add the particular additional variable, (2) interpret the parameters, (3) the R2 and adjusted R2, and, (4) the F-statistic.

```
mr1 <- lm(formula = fatalityrate ~ sb_useage + speed65 + ba08 + income, data = seat_belts_data)
summary(mr1)
```

```
##
## Call:
## lm(formula = fatalityrate ~ sb_useage + speed65 + ba08 + income,
##     data = seat_belts_data)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0118297 -0.0027847 -0.0003746  0.0021477  0.0227571
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.650e-02  6.927e-04  52.693  < 2e-16 ***
## sb_useage    4.422e-03  1.257e-03   3.518 0.000461 ***
```

```
## speed65     -2.875e-04  3.561e-04  -0.808 0.419623
## ba08        -1.879e-03  5.049e-04  -3.722 0.000212 ***
## income      -9.418e-07  3.896e-08 -24.171  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.004329 on 760 degrees of freedom
## Multiple R-squared:  0.5104, Adjusted R-squared:  0.5078
## F-statistic: 198.1 on 4 and 760 DF,  p-value: < 2.2e-16
```

Interpretation: In this multiple regression, I chose seat belt usage, blood alcholo limit, income and speed of 65 mile per hour limit. Based on the correlation plot, these variables are positively correlated R2 and adjusted R2: The R2 is 0.5104 and the adjusted R2 is 0.5078. From this we can say that there is 51% of variation in the fatality rate and by adding the 4 other variables has caused the variation of adjusted R2 to 50.78% F-statistic: The p-value : $<$ 2.2e-16 tells us that the model is better than the intercept only model and that one of the variables has an effect on fatality rate.

2. College on educational attainment. For this problem we are going to explore the effect of distance from college on educational attain- ment. The independent variable (Y) is years of completed education ed. All the estimated regression parameters for this part should be computed using linear algebra - see lesson 8.2 -. Also, any statistic (F-statistic or R2) should be computed manually and without the use of the lm command - you can use the command to verify your work -. You need to download the collegeDistance dataset to complete this part.

```
library("readxl")
college_dist <- as.data.frame(read_excel("~/CollegeDistance.xls"))
str(college_dist)
```

```
## 'data.frame':    3796 obs. of  14 variables:
##  $ female  : num  0 1 0 0 1 0 1 1 0 1 ...
##  $ black   : num  0 0 0 1 0 0 0 0 0 0 ...
##  $ hispanic: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ bytest  : num  39.1 48.9 48.7 40.4 40.5 ...
##  $ dadcoll : num  1 0 0 0 0 0 0 0 1 0 ...
##  $ momcoll : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ownhome : num  1 1 1 1 0 1 1 1 1 1 ...
##  $ urban   : num  1 1 1 1 1 1 0 0 1 1 ...
##  $ cue80   : num  6.2 6.2 6.2 6.2 5.6 5.6 7.2 7.2 5.9 5.9 ...
##  $ stwmfg80: num  8.09 8.09 8.09 8.09 8.09 8.09 8.85 8.85 8.09 8.09 ...
##  $ dist    : num  0.2 0.2 0.2 0.2 0.4 0.4 0.4 0.4 3 3 ...
##  $ tuition : num  0.889 0.889 0.889 0.889 0.889 ...
##  $ ed      : num  12 12 12 12 13 12 13 15 13 15 ...
##  $ incomehi: num  1 0 0 0 0 0 0 0 0 0 ...
```

2.1 Run an interpret the bivariate regression of ed on dist (distance to college). What's the estimated slope?

```
mat1 <- cbind(rep(1, nrow(college_dist)), college_dist$dist)
#(X X) - 1 X Y =
b1 <- solve(t(mat1) %*% mat1 )  %*%  t(mat1) %*% college_dist$ed
b1
```

```
##              [,1]
## [1,] 13.95585611
## [2,] -0.07337271
```

From the above derivation, the slope of the bivariate regression of ed on dist is -0.0733

2.2 Now, run a multiple regression of ed on dist but also include: bytest, female, black, hispanic, incomehi,

ownhome, dadcoll, momcoll, cue80, and, stwmfg80. What is the estimated effect of ed on dist? Compare your result to the previous estimation. Explain why the effects may differ.

```
mat2 <- as.matrix(cbind(college_dist$dist, college_dist$bytest, college_dist$female, college_dist$black
mat2 <- cbind(1, mat2)
#(X X) - 1 X Y =
b2 <- solve(t(mat2) %*% mat2 )  %*%  t(mat2) %*% college_dist$ed
b2
```

```
##              [,1]
##  [1,]  8.86137322
##  [2,] -0.03080391
##  [3,]  0.09244736
##  [4,]  0.14337772
##  [5,]  0.35380829
##  [6,]  0.40235145
##  [7,]  0.36659524
##  [8,]  0.14564162
##  [9,]  0.56991528
## [10,]  0.37918361
## [11,]  0.02441799
## [12,] -0.05020441
```

```
mr2 <- lm(formula = ed ~ dist + bytest + female + black + hispanic + incomehi + ownhome + dadcoll + mom
summary(mr2)
```

```
##
## Call:
## lm(formula = ed ~ dist + bytest + female + black + hispanic +
##     incomehi + ownhome + dadcoll + momcoll + cue80 + stwmfg80,
##     data = college_dist)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2752 -1.1429 -0.2216  1.1733  5.0559
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.861373   0.249705  35.487  < 2e-16 ***
## dist        -0.030804   0.012338  -2.497  0.01258 *
## bytest       0.092447   0.003167  29.187  < 2e-16 ***
## female       0.143378   0.050454   2.842  0.00451 **
## black        0.353808   0.071235   4.967 7.11e-07 ***
## hispanic     0.402351   0.074264   5.418 6.41e-08 ***
## incomehi     0.366595   0.060679   6.042 1.67e-09 ***
## ownhome      0.145642   0.066641   2.185  0.02892 *
## dadcoll      0.569915   0.073718   7.731 1.36e-14 ***
## momcoll      0.379184   0.081550   4.650 3.44e-06 ***
## cue80        0.024418   0.009609   2.541  0.01109 *
## stwmfg80    -0.050204   0.019801  -2.535  0.01127 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.538 on 3784 degrees of freedom
## Multiple R-squared:  0.2829, Adjusted R-squared:  0.2809
## F-statistic: 135.7 on 11 and 3784 DF,  p-value: < 2.2e-16
```

Interpretation: Based on the previous estimation, the effects may differ due to the addition of several variables that contravene the exogenity assumption wherein these variables can cause variations on the dependent variable Y.

2.3 Compute the R2 and the adjusted R2 for both regressions and interpret its significance. Which measure of goodness of fit you prefer in each regression? R2

$$R^2 = \frac{TSS - SSE}{TSS}$$

$$TSS = \sum_i (y_i - \bar{y})^2$$

$$SSE = \sum_i (y_i - \bar{y}_i)^2$$

Adjusted R2

$$adjustedR^2 = \frac{TSS/df_t - SSE/df_e}{TSS/df_t}$$

```
#Model 1
#R2
lm2 <- lm(formula = ed ~ dist, data = college_dist)
ypred <- predict(lm2)
y <- college_dist$ed
tss <- sum((y - mean(y)) ^ 2)
sse <- sum((y - ypred) ^ 2)
r1 <- (tss-sse) / tss
r1
```

```
## [1] 0.007449574
```

```
#adjusted R2
n <- length(y)
k <- 1
dft <- n - 1
dfe <- n - k - 1
(tss / dft - sse / dfe) / (tss / dft)
```

```
## [1] 0.007187963
```

```
#Model 2
#R2
lm3 <- lm(formula = ed ~ dist + bytest + female + black + hispanic + incomehi + ownhome + dadcoll + mom
ypred <- predict(lm3)
y <- college_dist$ed
tss <- sum((y - mean(y)) ^ 2)
sse <- sum((y - ypred) ^ 2)
r2 <- (tss-sse) / tss
r2
```

```
## [1] 0.2829346
```

```
#adjusted R2
n <- length(y)
k1 <- ncol(mat2)-1
dft <- n - 1
dfe <- n - k1 - 1
(tss / dft - sse / dfe) / (tss / dft)
```

```
## [1] 0.2808501
```

Interpretation: Adjusted R2 is a better measure of goodness of fit since it controls the number of variables in this model. Unlike bivariate model where the use of R2 or adjusted R2 does not matter, in multivariate models we need to depend on the other independent variables too.As the number of variables increases, the adjustred R2 is a better measure of goodness of fit. However, in this case, we can use both.

2.4 Bob is a non-hispanic black male. His high school was 20 miles from the nearest college. His base year composite score (bytest) was 58. His family income in 1980 was $26, 000, and his family owned a house. His mother attended college, but his father did not. The unemployment rate in his county was 7.5%, and the state average manufacturing hourly wage was $9.75. Predict Bob's years of completed schooling using both regressions and compare the results. Which result you prefer? (Explain)

```
n1 <- c(1, 2) %*% b1
n1
```

```
##          [,1]
## [1,] 13.80911
```

```
data1 <- c(2, 58, 0, 1, 0, 1, 1, 0, 1, 7.5, 9.75)
c(1, data1) %*% b2
```

```
##          [,1]
## [1,] 15.10058
```

Interpretation: From the bivariate model it predicts that Bob will complete his schooling in 13.8 years and in the multivariate model it predicts that he will complete his schooling in 15 years. I prefer the second model since the R2 is higher.

2.5 Test if all the parameters of the model are simultaneously equal to zero.

$$F - test : F = \frac{R^2/k}{(1-R^2)/n-k-1}$$

```
f <- (r2 / k1) / ((1 - r2) / (n - k1 - 1))
f
```

```
## [1] 135.7331
```

```
#p-value
pf(f, k1, (n - k1 - 1), lower.tail = F)
```

```
## [1] 1.916483e-263
```

Interpretation: The p-value is less than 0.05 we reject the null hypothesis where all the parameters of the model are simultaneously equal to zero