INSH5301 Intro Computational Statistics

Ali Banijamali

03/16/2020

PROBLEM 1: Seat belts

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.

```
# Reading the data:
seat.belt <- read.csv(
  paste('C:/Users/alibs/Google Drive/Courses/INSH5301 Intro Computational Statistics/',
  'Module 8 - Multiple Regression 1/seatbelts/seatbelts.csv', sep=''),
  stringsAsFactors = F)
# paste is for the path to be inside borders when printed as pdf.
```

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

ANS.

```
f_rate.vs.primary <- lm(fatalityrate ~ primary, data=seat.belt)
summary(f_rate.vs.primary)
```

```
##
## Call:
## lm(formula = fatalityrate ~ primary, data = seat.belt)
##
## 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
```

The result show that the primary has a negative effect on the fatality rate (-0.0017203). The p-value is not great (0.0117) neither is the effect of primary on the fatality rate (-0.0017203). R^2 is very small too (0.008309). Only %0.8 of the variation in fatality rate is explained by primary.

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)

ANS.

```
# There are NA cells in the 'sb_useage' column. I am replacing them with the average
# value in that column.
seat.belt[is.na(seat.belt$sb_useage), 'sb_useage'] <- mean(seat.belt$sb_useage, na.rm=T)

# The rest of the non-binary columns like age, income, ... don't have NAs and Nothing
# can be done about the binary items (setting as average value is not meaningful there)
# Just to make sure, we remove NAs using na.omit, which doesn't change the number of rows
# meaning that no more NA is remained in the table
seat.belt_na.dropped <- na.omit(seat.belt)

cor(seat.belt[, -1]) # I removed the state column because it was non-numeric
```

```
##                     year        fips         vmt fatalityrate   sb_useage
## year          1.0000000  0.00000000  0.12326043  -0.55909831  0.48551618
## fips          0.0000000  1.00000000 -0.06955920  -0.08730994  0.02936682
## vmt           0.1232604 -0.06955920  1.00000000  -0.16136608  0.18641134
## fatalityrate -0.5590983 -0.08730994 -0.16136608   1.00000000 -0.27973533
## sb_useage     0.4855162  0.02936682  0.18641134  -0.27973533  1.00000000
## speed65       0.6718047  0.02813346  0.08225171  -0.28183657  0.23824423
## speed70       0.3827189 -0.00224051  0.11961343  -0.07646409  0.19533553
## drinkage21    0.4968923 -0.04219360  0.10791028  -0.29375465  0.19040758
## ba08          0.2500599  0.08412680  0.09773188  -0.16983761  0.20762467
## income        0.7814340 -0.14560039  0.20615121  -0.70355753  0.48765212
## age           0.3704737  0.01027750  0.08345748  -0.37541297  0.11486032
## primary       0.1360999  0.01496448  0.13653429  -0.09115458  0.38738238
## secondary     0.5567256 -0.03621285  0.16567423  -0.32225884  0.21715360
##                  speed65      speed70  drinkage21         ba08     income
## year          0.67180466  0.382718930  0.49689226  0.250059948  0.7814340
## fips          0.02813346 -0.002240510 -0.04219360  0.084126801 -0.1456004
## vmt           0.08225171  0.119613435  0.10791028  0.097731880  0.2061512
## fatalityrate -0.28183657 -0.076464087 -0.29375465 -0.169837613 -0.7035575
## sb_useage     0.23824423  0.195335533  0.19040758  0.207624667  0.4876521
## speed65       1.00000000  0.204118915  0.47820628  0.192030237  0.3616334
## speed70       0.20411891  1.000000000  0.09935943  0.218343882  0.2090382
## drinkage21    0.47820628  0.099359428  1.00000000  0.130818347  0.4155422
## ba08          0.19203024  0.218343882  0.13081835  1.000000000  0.1218024
## income        0.36163338  0.209038166  0.41554216  0.121802409  1.0000000
## age           0.18895887  0.029695419  0.20346384 -0.054660435  0.4075273
## primary      -0.03391311 -0.008819134  0.13412318  0.089580841  0.1840774
## secondary     0.53712751  0.216877909  0.34086197  0.007397379  0.4328184
##                     age      primary    secondary
## year          0.37047374  0.136099910  0.556725590
## fips          0.01027750  0.014964483 -0.036212848
## vmt           0.08345748  0.136534289  0.165674229
## fatalityrate -0.37541297 -0.091154579 -0.322258843
## sb_useage     0.11486032  0.387382377  0.217153600
## speed65       0.18895887 -0.033913106  0.537127509
## speed70       0.02969542 -0.008819134  0.216877909
## drinkage21    0.20346384  0.134123179  0.340861965
```

```
## ba08         -0.05466044  0.089580841  0.007397379
## income        0.40752735  0.184077428  0.432818383
## age           1.00000000  0.073052548  0.160011863
## primary       0.07305255  1.000000000 -0.368623307
## secondary     0.16001186 -0.368623307  1.000000000
```

When we are talking about exogeneity of a variable, we mean that this variable is independent of other variables in the system. In other words, it is not affected by other variables. By looking at the correlation matrix above, we can see that primary doesn't have a strong correlation with other variables, except for sb_usage with a correlation of 0.387382377 and secondary with a correlation of -0.368623307. So, although the correlation is weak, there is still some correlation between primary and these two variables. Therefore, exogeneity will not be satisfied for primary.

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.

ANS.

We'll start by the one variable checked in part 1 and then add more variables:

```
f_rate.vs.primary <- lm(fatalityrate ~ primary, data=seat.belt)
summary(f_rate.vs.primary)
```

```
##
## Call:
## lm(formula = fatalityrate ~ primary, data = seat.belt)
##
## 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
```

Adding 1st variable: (speed70)

```
f_rate.vs.primary <- lm(fatalityrate ~ primary + speed70, data=seat.belt)
summary(f_rate.vs.primary)
```

```
##
## Call:
## lm(formula = fatalityrate ~ primary + speed70, data = seat.belt)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0135043 -0.0040833 -0.0003528  0.0033030  0.0236385
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0218316  0.0002446  89.248   <2e-16 ***
## primary     -0.0017332  0.0006788  -2.553   0.0109 *
## speed70     -0.0018606  0.0008660  -2.148   0.0320 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.006135 on 762 degrees of freedom
## Multiple R-squared:  0.01428,    Adjusted R-squared:  0.01169
## F-statistic: 5.519 on 2 and 762 DF,  p-value: 0.00417
```

The new variable doesn't seem to have more value than the previous value. Let's add the second variable: (income)

```
f_rate.vs.primary <- lm(fatalityrate ~ primary + speed70 + income, data=seat.belt)
summary(f_rate.vs.primary)
```

```
##
## Call:
## lm(formula = fatalityrate ~ primary + speed70 + income, data = seat.belt)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0114463 -0.0027320 -0.0005345  0.0021973  0.0233904
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.805e-02  6.189e-04  61.473  < 2e-16 ***
## primary      8.194e-04  4.918e-04   1.666  0.09606 .
## speed70      1.829e-03  6.306e-04   2.901  0.00383 **
## income      -9.330e-07  3.418e-08 -27.301  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.004364 on 761 degrees of freedom
## Multiple R-squared:  0.502,  Adjusted R-squared:  0.5001
## F-statistic: 255.7 on 3 and 761 DF,  p-value: < 2.2e-16
```

Finally, the third variable to be added is age:

```
f_rate.vs.primary <- lm(fatalityrate ~ primary + speed70 + income + age, data=seat.belt)
summary(f_rate.vs.primary)
```

```
##
## Call:
## lm(formula = fatalityrate ~ primary + speed70 + income + age,
##     data = seat.belt)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0112056 -0.0027348 -0.0004371  0.0022451  0.0208319
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.005e-02  3.345e-03  14.962  < 2e-16 ***
```

```
## primary      8.101e-04  4.878e-04   1.661 0.097220 .
## speed70      1.687e-03  6.268e-04   2.691 0.007273 **
## income      -8.782e-07  3.708e-08 -23.684  < 2e-16 ***
## age         -3.692e-04  1.012e-04  -3.649 0.000281 ***
## ---
## 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:  0.5106, Adjusted R-squared:  0.508
## F-statistic: 198.2 on 4 and 760 DF,  p-value: < 2.2e-16
```

To sum up: primary and speed have a positive effect, meaning that when there was a primary enforcement and the speed was above 70, the fatality rate has increased. The other two variables have the oppoite relationship (negative) which means that As people have more income and their age increase, the fatality rate decreases. These all make sense: If you have higher speed, the chance of a deadly crash increases, or young people tend to drive more carelessly. With higher income, people afford to buy more expensive cars which are equiped with more safety features and are generally considered safer.
Although all of these variables have their own effect on the fatality rate, we should note that the effect of Age and Income is much more prominent! Notice that when we added them, the R^2 immediately increased. and the other interesting point is that when we added Income after Age, R^2 didn't change that much. This shows that they are related somehow. It makes sense, as you grow older, your income generally increases. Also, if you look at the p-values, income has the lowest p-value. So we can safely say between these variables, the most important variable is income. We can even say that, if you have more money, you probably have a better car and you are more inclined to drive at higher speeds because more expensive cars are more stable in higher speeds and people can easily drive at higher speeds without feeling that they are going too fast.

PROBLEM 2. College on educational attainment

For this problem we are going to explore the effect of distance from college on educational attainment. 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.

```
# Reading the data:
col.dist <- read.csv(
  paste('C:/Users/alibs/Google Drive/Courses/INSH5301 Intro Computational Statistics/',
  'Module 8 - Multiple Regression 1/collegedistance/CollegeDistance.csv', sep=''),
  stringsAsFactors = F)
# paste is for the path to be inside borders when printed as pdf.
```

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

ANS.
We will go by the following formula:
$Y = X\beta + \epsilon$
Multiplying both sides by X^T (X transpose):
$X^TY = X^TX\beta + X^T\epsilon$
Knowing that $X^T\epsilon = 0$, We then multiply both sides by $(X^TX)^{-1}$, we will have:
$(X^TX)^{-1}X^TY = (X^TX)^{-1}X^TX\beta$

Again knowing that $A^{-1}A = 1$, we will have:
$\beta = (X^T X)^{-1} X^T Y$
Now we can find βs using this formula:

```
# Since this is a bivariate regression between distance and education:
X.1 <- as.matrix(cbind(1, col.dist$dist))
Y.1 <- as.matrix(col.dist$ed)

beta.1 <- solve(t(X.1) %*% X.1) %*% (t(X.1) %*% Y.1)
print(paste('Intercept = ', beta.1[1], sep=''))
```

```
## [1] "Intercept = 13.9558561147894"
```

```
print(paste('Slope = ',     beta.1[2], sep=''))
```

```
## [1] "Slope = -0.0733727071292289"
```

```
# For checking and comparison:
ed.vs.dist <- lm(ed ~ dist, data=col.dist)
summary(ed.vs.dist)
```

```
##
## Call:
## lm(formula = ed ~ dist, data = col.dist)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9559 -1.8091 -0.6624  2.0515  4.4844
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.95586    0.03772 369.945   <2e-16 ***
## dist        -0.07337    0.01375  -5.336    1e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.807 on 3794 degrees of freedom
## Multiple R-squared:  0.00745,    Adjusted R-squared:  0.007188
## F-statistic: 28.48 on 1 and 3794 DF,  p-value: 1.004e-07
```

The negative slope means that as the distance to the college increase, the years of complete education decrease. However, the R-squared shows that this variable, explains only %0.7 of the variation in education level.

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.

ANS.

```
# Deleting these columns:'urban', 'tuition', 'ed':
X.2 <- as.matrix(cbind(1, col.dist[, -c(8, 12, 13)]))
Y.2 <- as.matrix(col.dist$ed)

beta.2 <- solve(t(X.2) %*% X.2) %*% (t(X.2) %*% Y.2)
beta.2
```

```
##                 [,1]
## 1         8.86137322
## female    0.14337772
## black     0.35380829
## hispanic  0.40235145
## bytest    0.09244736
## dadcoll   0.56991528
## momcoll   0.37918361
## ownhome   0.14564162
## cue80     0.02441799
## stwmfg80 -0.05020441
## dist     -0.03080391
## incomehi  0.36659524
```

```
# For checking and Comparison:
ed.vs.vars <- lm(ed ~ female+black+hispanic+bytest+dadcoll+momcoll+
                  ownhome+cue80+stwmfg80+dist+incomehi, data=col.dist)
summary(ed.vs.vars)
```

```
##
## Call:
## lm(formula = ed ~ female + black + hispanic + bytest + dadcoll +
##     momcoll + ownhome + cue80 + stwmfg80 + dist + incomehi, data = col.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 ***
## 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 ***
## bytest       0.092447   0.003167  29.187  < 2e-16 ***
## dadcoll      0.569915   0.073718   7.731 1.36e-14 ***
## momcoll      0.379184   0.081550   4.650 3.44e-06 ***
## ownhome      0.145642   0.066641   2.185  0.02892 *
## cue80        0.024418   0.009609   2.541  0.01109 *
## stwmfg80    -0.050204   0.019801  -2.535  0.01127 *
## dist        -0.030804   0.012338  -2.497  0.01258 *
## incomehi     0.366595   0.060679   6.042 1.67e-09 ***
## ---
## 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
```

As we can see, the model now has a larger R-squared, meaning that the model now explains %28 of the variations in the dependent variable (compared to %0.7 when we had only distance variable in the model.) Also the overal p-value has decreased with the addition of new variables.

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?

ANS.

```
# 1. Computing R-square & adjusted R-squared for the regression model 1:
# Calculating the predicted Y:
Y.1_pred <- X.1 %*% beta.1
Y.1      <- as.matrix(col.dist$ed)

# Calculating R-squared:
TSS <- sum( (Y.1 - mean(Y.1))^2 )
SSE <- sum( (Y.1 - Y.1_pred)^2 )
R2  <- (TSS-SSE)/TSS
cat('R-squared for model 1 =', R2)
```

```
## R-squared for model 1 = 0.007449574
```

```
# Calculating adjusted R-squared:
n <- nrow(X.1)   # n is no. of observations
k <- ncol(X.1)-1 # k is no. of variables

dft <- n-1
dfe <- n-k-1

adj_R2 <- (TSS/dft - SSE/dfe) / (TSS/dft)
cat('Adjusted R-squared for model 1 =', adj_R2)
```

```
## Adjusted R-squared for model 1 = 0.007187963
```

```
# 2. Computing R-square & adjusted R-squared for the regression model 2:
# Calculating the predicted Y:
Y.2_pred <- X.2 %*% beta.2
Y.2      <- as.matrix(col.dist$ed)

# Calculating R-squared:
TSS <- sum( (Y.2 - mean(Y.2))^2 )
SSE <- sum( (Y.2 - Y.2_pred)^2 )
R2  <- (TSS-SSE)/TSS
cat('R-squared for model 2 =', R2)
```

```
## R-squared for model 2 = 0.2829346
```

```
# Calculating adjusted R-squared:
n <- nrow(X.2)   # n is no. of observations
k <- ncol(X.2)-1 # k is no. of variables

dft <- n-1
dfe <- n-k-1

adj_R2 <- (TSS/dft - SSE/dfe) / (TSS/dft)
cat('Adjusted R-squared for model 2 =', adj_R2)
```

```
## Adjusted R-squared for model 2 = 0.2808501
```

Generally, the adjusted R-squared is a better measure because it is less prone to overfitting when we have a lot of parameters explaining the dependent variable. However, in the first model, since it is being explained by only one parameter, the mentioned case doesn't really apply because with only one variable we shouldn't

be worried about overfitting. In the second model however, we must definiitely pay more attention to the adjusted R-squared.

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)

ANS.

```
# 1.First let's fill Bob's info in a matrix, they should be in the following order:
colnames(X.2)
```

```
##  [1] "1"        "female"   "black"    "hispanic" "bytest"   "dadcoll"
##  [7] "momcoll"  "ownhome"  "cue80"    "stwmfg80" "dist"     "incomehi"
```

```
Bob.info.1 <- c(1, 20)
Bob.info.2 <- c(1, 0, 1, 0, 58, 0, 1, 1, 7.5, 9.75, 20, 1)

# Years of completed school based on model 1:
Bob.Y1 <- Bob.info.1 %*% beta.1
cat("Bob's years of completed school based on model 1 =", Bob.Y1)
```

```
## Bob's years of completed school based on model 1 = 12.4884
```

```
# Years of completed school based on model 2:
Bob.Y2 <- Bob.info.2 %*% beta.2
cat("Bob's years of completed school based on model 2 =", Bob.Y2)
```

```
## Bob's years of completed school based on model 2 = 14.54611
```

As we calculated in the previous part, both R-squared and the adjusted R-squared are greater for the second model, the second model incorporates more parameters and takes into account more variables (of all the information that we have about Bob, the first model only uses the college's distance). The second model explains more of the variation in years of education according to it's R-squared/adjusted R-squared.

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

ANS.
We have to use F-statistics. F-statistics is defined by:
$F - statistics = \frac{R^2/k}{(1-R^2)/(n-k-1)}$

```
# Calculating the F-statistic for model 2:
n <- nrow(X.2)
k <- ncol(X.2)-1

Y.2_pred <- X.2 %*% beta.2
Y.2      <- as.matrix(col.dist$ed)

TSS <- sum( (Y.2 - mean(Y.2))^2 )
SSE <- sum( (Y.2 - Y.2_pred)^2 )
R2  <- (TSS-SSE)/TSS

# Calculating F-statistic:
```

```r
F.stat <- (R2/k) / ((1-R2)/(n-k-1))
F.stat
```

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

```r
# Calculating p-value:
pf(F.stat, k, (n-k-1), lower.tail=F)
```

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

F-test is another way to test the significance of the model. It basically tests if the parameters of the model (βs) are significantly different from zero:
$H_0 = None\ of\ \beta s\ are\ significantly\ different\ from\ zero$ ($The\ model\ is\ entirely\ insignificant.$)
The very small p-value (1.916483e-263) shows that all of the parameters put together can't be by chance and we can reject the null hypothesis.