# HW08_Sampathirao_A

*Anvita Sampathirao*

*7/13/2019*

## R Markdown

#1.1

```
dataset<- read.csv("seatbelts.csv", stringsAsFactors = FALSE)
#head(dataset)
y<- dataset$fatalityrate
x1<- dataset$primary
summary(lm(y~x1))
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## 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 ***
## x1          -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
```

Having the primary law enforced has a significant effect on fatality rate (p-value<= 0.05)

Having the primary law enforced decreases the fatality rate by about 0.17%

R^2 is 0.0083. Thus, only about 0.8% of the variation in fatality rate can be attributed to variation in enforcing the primary law.

#1.2

```
cor(x1,y)
```

```
## [1] -0.09115458
```

For exogeneity condition, our independent variable primary law enforcement is independent of error, i.e. it is caused externally to fatality rate environment.

#1.3

Selected Income to see if an increase in per capita income resulted in people being negligent about the seatbelt laws to study the behavioral impact.

Selected Mean Age to see if there is a relation between age groups and fatality rates and if certain age groups are more prone to fatalities.

```
x2<-dataset$income
x3<-dataset$age
summary(lm(y~x1 + x2))
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.011492 -0.002829 -0.000410  0.002098  0.023253
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.780e-02  6.161e-04  61.355   <2e-16 ***
## x1           7.492e-04  4.935e-04   1.518    0.129
## x2          -9.118e-07  3.354e-08 -27.182   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.004385 on 762 degrees of freedom
## Multiple R-squared:  0.4965, Adjusted R-squared:  0.4952
## F-statistic: 375.7 on 2 and 762 DF,  p-value: < 2.2e-16
```

Having the primary law enforced increases the fatality rate by about 0.075% and the effect is insignificant (p value>= 0.05)

Additionally, a unit increase in income level decreases the fatality rate by 0.00009% and the effect is significant (p value<= 0.05)

The R^2 is 0.4965. Thus, 49.65% of the variation in the fatality can be attributed to variation in having the primary law enforced and income.

Adjusted R^2= 0.4952

F statistic is 375.7 and the corresponding p value is less than 0.05. Thus, the model predicts fatality rate better than the mean of fatality rate.

```
summary(lm(y~x1 + x2 + x3))
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0112366 -0.0028087 -0.0003946  0.0022006  0.0205882
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.037e-02  3.356e-03  15.008  < 2e-16 ***
## x1           7.452e-04  4.892e-04   1.523 0.128148
## x2          -8.562e-07  3.631e-08 -23.578  < 2e-16 ***
## x3          -3.862e-04  1.014e-04  -3.808 0.000151 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.004346 on 761 degrees of freedom
## Multiple R-squared:  0.5059, Adjusted R-squared:  0.504
## F-statistic: 259.8 on 3 and 761 DF,  p-value: < 2.2e-16
```

Having the primary law enforced increases the fatality rate by about 0.074% and the effect is insignificant (p value>= 0.05)

Additionally, a unit increase in income level decreases the fatality rate by 0.00008% and the effect is significant (p value<= 0.05)

Additionally, a unit increase in the mean age decreases the fatality rate by 0.038% and the effect is significant (p value<= 0.05)

The R^2 is 0.5059. Thus, 50.59% of the variation in the fatality can be attributed to variation in having the primary law enforced, income and age. Adjusted R^2 is 0.504

F statistic is 259.8 and the corresponding p value is less than 0.05. Thus, the model predicts fatality rate better than the mean of fatality rate.

#2.1

```
library("readxl")
ndataset<- read_excel("CollegeDistance.xls", col_names = TRUE)
head(ndataset)
```

```
## # A tibble: 6 x 14
##   female black hispanic bytest dadcoll momcoll ownhome urban cue80 stwmfg80
##    <dbl> <dbl>    <dbl>  <dbl>   <dbl>   <dbl>   <dbl> <dbl> <dbl>    <dbl>
## 1      0     0        0   39.2       1       0       1     1   6.2     8.09
## 2      1     0        0   48.9       0       0       1     1   6.2     8.09
## 3      0     0        0   48.7       0       0       1     1   6.2     8.09
## 4      0     1        0   40.4       0       0       1     1   6.2     8.09
## 5      1     0        0   40.5       0       0       0     1   5.6     8.09
## 6      0     0        0   54.7       0       0       1     1   5.6     8.09
## # ... with 4 more variables: dist <dbl>, tuition <dbl>, ed <dbl>,
## #   incomehi <dbl>
```

```
a<- cov(ndataset$dist, ndataset$ed)
b1<- a/(sd(ndataset$dist)^2)
b1
```

```
## [1] -0.07337271
```

```
b0<- mean(ndataset$ed)- (b1*mean(ndataset$dist))
b0
```

```
## [1] 13.95586
```

```r
yhatfun<- function(x){
  yhat<- b0 + (b1*x)
  return(yhat)
}
edfit<- yhatfun(ndataset$dist)
SSE<- sum((ndataset$ed - edfit)^2)
TSS <- sum((ndataset$ed - mean(ndataset$ed))^2)
Rsq <- (TSS -SSE)/TSS
Rsq
```

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

```r
summary(lm(ndataset$ed~ndataset$dist))
```

```
##
## Call:
## lm(formula = ndataset$ed ~ ndataset$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 ***
## ndataset$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
```

Having distance from College has a significant effect on years of education completed (p-value<= 0.05)

Increasing unit distance from College decreases years of education completed by about 7.34% i.e. estimated slope (beta-1)= -0.07337

R^2 is 0.00745. Thus, only about 0.7% of the variation in years of education can be attributed to variation in distance from college.

#2.2

```r
xmat <- as.matrix(cbind(ndataset$dist,
                        ndataset$bytest,
                        ndataset$female,
                        ndataset$black,
                        ndataset$hispanic,
                        ndataset$incomehi,
                        ndataset$ownhome,
                        ndataset$dadcoll,
                        ndataset$momcoll,
                        ndataset$cue80,
```

```
                        ndataset$stwmfg80))
xmat <- cbind(1,xmat)
head(xmat)
```

```
##      [,1] [,2]  [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12]
## [1,]    1  0.2 39.15    0    0    0    1    1    1     0   6.2  8.09
## [2,]    1  0.2 48.87    1    0    0    0    1    0     0   6.2  8.09
## [3,]    1  0.2 48.74    0    0    0    0    1    0     0   6.2  8.09
## [4,]    1  0.2 40.40    0    1    0    0    1    0     0   6.2  8.09
## [5,]    1  0.4 40.48    1    0    0    0    0    0     0   5.6  8.09
## [6,]    1  0.4 54.71    0    0    0    0    1    0     0   5.6  8.09
```

```
Q2Y<- ndataset$ed
betas <- solve( t(xmat) %*% xmat )   %*%   t(xmat) %*% Q2Y
betas
```

```
##              [,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
```

```
edhat<-xmat %*% betas
head(edhat)
```

```
##          [,1]
## [1,] 13.30192
## [2,] 13.40737
## [3,] 13.25198
## [4,] 12.83477
## [5,] 12.46529
## [6,] 13.78308
```

The estimated effect of dist on ed is -0.03080391. i.e. the rate of decrease in years of education with an increase in distance from college has decreased by 0.04.

The estimated parameters differ because of non inclusion of other variables such as the ones included in this model (race, ethnicity, test scores, income, ownership, parents' educational background, county's unemployment rate, state hourly wage)

#2.3

$$H_0 : \beta_i = 0$$
$$H_a : \beta_i \neq 0$$

```r
n <- nrow(xmat) # Number of observations, rows
kPlus1 <- ncol(xmat) # columns of xmat = k + 1
dof<- n-kPlus1 #Degree of freedom
se_y <- sqrt(sum( (Q2Y - edhat)^2 ) / (n - kPlus1) )
se_beta<- se_y * sqrt( diag( solve( t(xmat) %*% xmat )) )
data<- data.frame(betas,se_beta, row.names = c("Intercept",
                                               "Distance",
                                               "bytest",
                                               "female",
                                               "black",
                                               "hispanic",
                                               "incomehi",
                                               "ownhome",
                                               "dadcoll",
                                               "momcoll",
                                               "cue80",
                                               "stwmfg80"))
colnames(data)<- c("beta","betaerror")
data
```

```
##                  beta   betaerror
## Intercept  8.86137322 0.249705370
## Distance  -0.03080391 0.012337745
## bytest     0.09244736 0.003167406
## female     0.14337772 0.050453511
## black      0.35380829 0.071234510
## hispanic   0.40235145 0.074264234
## incomehi   0.36659524 0.060679243
## ownhome    0.14564162 0.066640862
## dadcoll    0.56991528 0.073718170
## momcoll    0.37918361 0.081549788
## cue80      0.02441799 0.009609480
## stwmfg80  -0.05020441 0.019801292
```

```r
variables<- nrow(data)
t_value<- rep(0,variables)
for(i in 1:variables){
  t_value[i]<- data$beta[i]/ data$betaerror[i]
}
t_value # Calculating t values to test beta hypothesis individually
```

```
##  [1] 35.487315 -2.496721 29.187094  2.841779  4.966810  5.417836  6.041526
##  [8]  2.185470  7.731001  4.649719  2.541031 -2.535411
```

```r
t_critical<- qt(0.975, dof)
#Function to perform t test
tTest<- function(t){
  ifelse(abs(t)>=t_critical, "Reject Null ", "Cant reject H0")
}
#checking for each value of beta
for(i in 1:variables){
  print(tTest(t_value[i]))
}
```

```
## [1] "Reject Null "
## [1] "Reject Null "
## [1] "Reject Null "
## [1] "Reject Null "
## [1] "Reject Null "
## [1] "Reject Null "
## [1] "Reject Null "
## [1] "Reject Null "
## [1] "Reject Null "
## [1] "Reject Null "
## [1] "Reject Null "
## [1] "Reject Null "
```

#2.4

```
#R^2 value
tssm2 <- sum((Q2Y - mean(Q2Y))^2)
ssem2 <- sum((Q2Y-edhat)^2)
r2m2 <- (tssm2-ssem2)/tssm2
r2m2
```

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

```
#Adjusted R^2 value
n1 <- length(Q2Y)
k1 <- ncol(xmat)-1
dft1 <- n1 - 1
dfe1 <- n1 - k1 - 1
Adjr2m2<- (tssm2/dft1 - ssem2/dfe1)/ (tssm2/dft1)
Adjr2m2
```

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

Would prefer Adjusted R^2 as a measure of goodness of fit because it avoids overfitting, i.e., with increasing number of variables, adjusted R^2 decreases while R^2 does not.

```
#To verify
summary(lm(ndataset$ed~ndataset$dist +
     ndataset$bytest +
     ndataset$female +
     ndataset$black +
     ndataset$hispanic +
     ndataset$incomehi +
     ndataset$ownhome +
     ndataset$dadcoll +
     ndataset$momcoll +
     ndataset$cue80 +
     ndataset$stwmfg80))
```

```
##
## Call:
## lm(formula = ndataset$ed ~ ndataset$dist + ndataset$bytest +
```

```
##     ndataset$female + ndataset$black + ndataset$hispanic + ndataset$incomehi +
##     ndataset$ownhome + ndataset$dadcoll + ndataset$momcoll +
##     ndataset$cue80 + ndataset$stwmfg80)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2752 -1.1429 -0.2216  1.1733  5.0559
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        8.861373   0.249705  35.487  < 2e-16 ***
## ndataset$dist     -0.030804   0.012338  -2.497  0.01258 *
## ndataset$bytest    0.092447   0.003167  29.187  < 2e-16 ***
## ndataset$female    0.143378   0.050454   2.842  0.00451 **
## ndataset$black     0.353808   0.071235   4.967 7.11e-07 ***
## ndataset$hispanic  0.402351   0.074264   5.418 6.41e-08 ***
## ndataset$incomehi  0.366595   0.060679   6.042 1.67e-09 ***
## ndataset$ownhome   0.145642   0.066641   2.185  0.02892 *
## ndataset$dadcoll   0.569915   0.073718   7.731 1.36e-14 ***
## ndataset$momcoll   0.379184   0.081550   4.650 3.44e-06 ***
## ndataset$cue80     0.024418   0.009609   2.541  0.01109 *
## ndataset$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
```

#2.5

$$\text{Bivariate model}: ed = 13.95 - 0.07 * dist$$

$$\text{Multivariate model}: ed = 8.86 - 0.03 * dist + 0.09 * testscore + 0.14 * gender + 0.35 * race$$
$$+0.40 * ethnicity + 0.36 * income + 0.14 * ownership + 0.57 * dadedu + 0.37 * momedu$$
$$+0.02 * unemprate - 0.05 * statehrlywage$$

```r
#Model 1- Biviriate Regression
yhatfun(2)
```

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

```r
#Model 2- Multivariate Regression
Input<- matrix(c(1, 2, 58, 0, 1, 0, 1, 1, 0, 1, 7.5, 9.75))
bobhat<- data$beta %*% Input
bobhat
```

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

Would prefer the results from Multivariate regression model as it predicts that Bob had an additional year of education after his AA degree as it has a higher adjusted R^2 value.

#2.6

$$H_0 : \text{None of them are significantly different from 0}$$

$$H_a : \text{At least one coefficient is significantly different from 0}$$

```
F_stat<- (r2m2/k1) / ((1-r2m2)/(n1-k1-1))
F_stat
```

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

```
pf(F_stat, k1, (n1-k1-1), lower.tail= F)
```

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

Because the p value is less than 0.05, we can reject the null hypothesis, i.e. all the parameters in the model are not simultaneously equal to 0. Atleast one of them is different than 0.