Assignment 3 CS 4070

Nirmal Roy, 4724429

December 11, 2017

Question 1

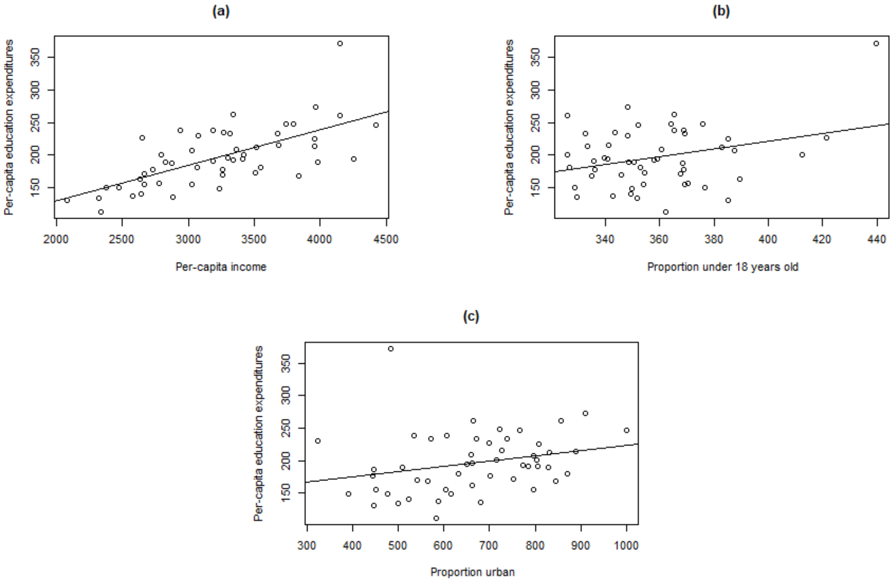


Figure 1: Scatterplots showing relationship between (a) Per-capita education expenditures and per-capita income, (b) Per-capita education expenditures and proportion under 18 years old per 1000, (c) Per-capita education expenditures and proportion urban per 1000 and also the least square lines drawn on them

Income

The least square regression of education expenditure on per capita income gives the values $A = 17.710031$, $B = 0.055376$, $S_E = 34.94$ and $r^2 = 0.4457$. The intercept value A means the value of education expenditure when per capita income is zero and doesn't have any physical significance as we don't have any values where per capita income will be zero. Slope B means the average change of education expenditure is 0.055376 for unit change in per capita income. S_E gives the average error when we try to predict education expenditure from values of per capita income. $S_E = 34.94$ maybe smaller with respect to the data but not negligible. $r^2 = 0.4457$ tells that 44.57% of the variation in education expenditure is captured by its linear regression on per capita income which is a value that is neither too bad nor too good. Hence, the least square line is an average summary of the relationship.

Young

The least square regression of education expenditure on proportion of under 18 gives the values $A = -20.4247$, $B = 0.6039$, $S_E = 44.59$ and $r^2 = 0.0786$. AS explained before, intercept value A means the value of education expenditure when there are no under 18 population and doesn't have any physical significance. Slope B means the average change of education expenditure is 0.6039 for unit change in proportion of under $18.S_E = 44.59$, the average error when we try to predict education expenditure from values of proportion of under 18, is larger than the residual error for the linear regression on income. $r^2 = 0.0786$ tells that only 7.86% of the variation in education expenditure is captured by its linear regression which is a very small proportion and other(or more) variables are needed if we want to predict education expenditure more accurately. The least square line is not a good summary of the relationship between the two variables.

Urban

The least square regression of education expenditure on per capita income gives the values $A = 142.60415$, $B = 0.08083$, $S_E = 45.27$ and $r^2 = 0.05035$. Again, the intercept value A means the value of education expenditure when proportion of urban is zero and doesn't have any physical significance. Slope B gives the average change of education expenditure is 0.08083 for unit change in per capita income. The regression on urban variable gives highest residual error $S_E = 45.27$ and the lowest regression coefficient r^2, which explains only 5.035% of the variation of education expenditure and hence other variables are needed to properly predict its value. The least square line is not a good summary of the relationship between the two variables.

Question 2

Multiple least squares regression

Computing the multiple least squares regression of the response on the explanatory variables gives the values $A = -286.8$, $B_1 = 0.08065$,$B_2 = 0.8173$, $B_3 = -0.1058$, $S_E = 26.69$ and $r^2 = 0.6698$. Following from before, intercept value A means the value of education expenditure when all the explanatory variables are zero and usually of little direct interest, because the fitted value above $X_i = 0$ is rarely important. Slope B_1 means the average change of education expenditure is 0.08065 for unit change in per capita income keeping the other explanatory variables constant. This is a difference between multiple and simple regression, as in the latter, the other variables are simply ignored while calculating coefficients. Similarly, B_2 and B_3 gives the average change in education expenditure for unit change in proportion under 18 and proportion urban respectively, keeping other two variables constant. S_E gives the average error when we try to predict education expenditure from all the explanatory variables combined. As expected, $S_E = 26.69$ is smaller than the residual errors of the least square lines computed above but it's not negligible. r^2 is also higher and multiple least square regression explains 66.98% of the variation of education expenditure. Hence, when we combine the effects of various explanatory variables we get a better estimate of our response variable, per capita education expenditure.

Question 3

3.a

Table 1

Variables	**Coefficient standard errors**	**95% confidence intervals**
[Intercept]	709.1	(-3319.631, -419.0080421)
year	0.3755	(0.2231212, 1.7591416)
tfr	0.004803	(-0.02858911, -0.0089433)
partic	0.02841	(0.01894638 , 0.1351510)
degrees	0.1501	(-0.1077843 , 0.5062294)

Table showing coefficient standard errors and 95-percent confidence intervals for the coefficients

3.b

For the multiple least square regression we have calculated TSS = 8967.031 and RSS = 392.5. Hence, RegSS = TSS - RSS = 8967.031 - 392.5 = 8574.53. Now we construct the Analysis of Variance Table.

Table 2: Anova Table

Source	**Sum of Squares**	**df**	**Mean Sqaure**	**F**
Regression	8574.53	4	2143.632	158.3827
Residual	392.5	29	13.5345	

Mean Square is calculated as (Sum of Squares/df) and then F value is calculated as (Regression Mean Square/ Residual Mean Square). From the F distribution table with degrees of freedom 4 and 29 we get the p-value for F statistics 158.3827 to be $< 2.2e-16$. This value being very close to 0, we can reject the omnibus null hypothesis, that all βs are 0.

3.c

To test the null hypothesis, $H_0 : \beta_1 = 0$ we have to consider the reduced model without the variable year and the full model. Let RSS_1 and RegSS_1 represent, respectively, the residual and regression sums of squares for the full model; similarly, RSS_0 and RegSS_0 are the residual and regression sums of squares for the null model. We have, RSS_1 = 392.50 , RegSS_1= 8574.53, RegSS_0 = TSS - RSS_0 = 8967.031 - 486.79 = 8480.241.

Therefore F-statistics for the null hypothesis is given by,

$$F_0 = \frac{(\text{RegSS}_1 - \text{RegSS}_0)/q}{\text{RSS}_1/(n-k-1)} = \frac{(8574.53 - 8480.241)/1}{392.50/(34-4-1)} = 6.9665 \qquad (1)$$

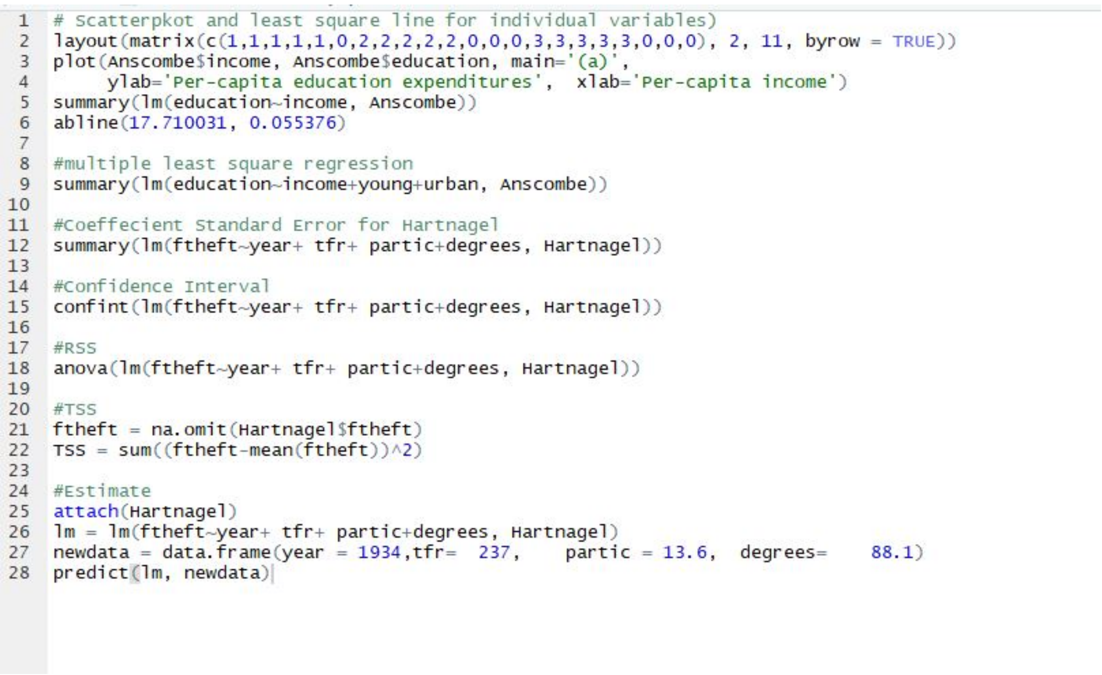
From the F distribution table for 1 and 29 degrees of freedom we have p-value = 0.01323 $<$ 0.05 and hence we can reject the null hypothesis indicating year has a significant influence on the response variable.

Now, for year we have $B_1 = 0.9911$ and $\text{SE}(B_1)$= 0.3755 . Therefore t-value = $B_1/\ \text{SE}(B_1)$ = 2.639. and $2.639^2 = 6.9665 = F$-value as calculated above.

3.d

An estimate for the mean female theft conviction rate per 100, 000 for values of the explanatory variables corresponding to the year 1934 is 15.89. The code is attached in the appendix.

Appendix



```
# Scatterpkot and least square line for individual variables)
layout(matrix(c(1,1,1,1,1,0,2,2,2,2,2,0,0,0,3,3,3,3,3,0,0,0), 2, 11, byrow = TRUE))
plot(Anscombe$income, Anscombe$education, main='(a)',
     ylab='Per-capita education expenditures',  xlab='Per-capita income')
summary(lm(education~income, Anscombe))
abline(17.710031, 0.055376)

#multiple least square regression
summary(lm(education~income+young+urban, Anscombe))

#Coeffecient Standard Error for Hartnagel
summary(lm(ftheft~year+ tfr+ partic+degrees, Hartnagel))

#Confidence Interval
confint(lm(ftheft~year+ tfr+ partic+degrees, Hartnagel))

#RSS
anova(lm(ftheft~year+ tfr+ partic+degrees, Hartnagel))

#TSS
ftheft = na.omit(Hartnagel$ftheft)
TSS = sum((ftheft-mean(ftheft))^2)

#Estimate
attach(Hartnagel)
lm = lm(ftheft~year+ tfr+ partic+degrees, Hartnagel)
newdata = data.frame(year = 1934,tfr=  237,    partic = 13.6,  degrees=    88.1)
predict(lm, newdata)
```

Figure 2: Relevant Codes