Assignment ECON 204 A

Pakhi Chachra

2024-01-17

Problem 1

Use the data set `Growth` and carry out the following exercises-

i. Construct a table that shows the sample mean, standard deviation, and minimum and maximum values for the series *Growth*, *Trade-Share*, *YearsSchool*, *Oil*, *Rev_Coups*, *Assassinations*, and *RGDP60*. Include the appropriate units for all entries.

ii. Run a regression of *Growth* on *TradeShare*, *YearsSchool*, *Rev_Coups*, *Assassinations*, and RGDP60. What is the value of the coefficient on *Rev_Coups*? Interpret the value of this coefficient.

iii. Use the regression to predict the average annual growth rate for a country that has average values for all regressors.

iv. Repeat (iii) but now assume that the country's value for *TradeShare* is one standard deviation above the mean.

Ans.

i)

To begin with the question, we first load the database into the R environment using the following code chunk:

```
library(readxl)
Growth <- read_excel("Growth.xlsx")
attach(Growth)
```

After doing the above, we now calculate the required data employing the following functions:

- `mean()` : To calculate the average values
- `sd()` : To calculate the standard deviation
- `min()` : To calculate the minimum values
- `max()` : To calculate the maximum values

The above functions are used to then calculate the sample mean, standard deviation, and minimum and maximum values for the series Growth, Tradeshare, YearsSchool, Oil, Rev_Coups, Assassinations, and RGDP60 in the following way:

```
Growth_mean <- colMeans(Growth[2:8])
Growth_sd <- sapply(Growth[2:8], sd)
Growth_min <- sapply(Growth[2:8], min)
Growth_max <- sapply(Growth[2:8], max)
table_Growth <- data.frame(Growth_mean, Growth_sd, Growth_min, Growth_max)
row.names(table_Growth) <- c("Average annual growth (in %)",
                             "Oil", "GDP per capita in 1960 (in $)",
                             "Share of Trade in economy (proportion)",
                             "Years of Schooling (in years)",
                             "Average Annual number of Revolutions",
                             "Average no. of Annual Political Assassinations")
colnames(table_Growth) <- c("Mean", "std dev", "Min value", "Max Value")
units_growth <- c("percentage", "binary (0 or 1)","dollars", "proportion",
           "years", "count", "count")
table_Growth$units <- units_growth
table_Growth
```

```
##                                                      Mean      std dev
## Average annual growth (in %)                    1.8691197    1.8161889
## Oil                                             0.0000000    0.0000000
## GDP per capita in 1960 (in $)                3130.8125339 2522.9786371
## Share of Trade in economy (proportion)          0.5423919    0.2283326
## Years of Schooling (in years)                   3.9592187    2.5534647
## Average Annual number of Revolutions            0.1700666    0.2254557
## Average no. of Annual Political Assassinations  0.2819010    0.4941590
##                                                  Min value    Max Value
## Average annual growth (in %)                    -2.811944    7.1568546
## Oil                                              0.000000    0.0000000
## GDP per capita in 1960 (in $)                  366.999939 9895.0039062
## Share of Trade in economy (proportion)           0.140502    1.1279370
## Years of Schooling (in years)                    0.200000   10.0699997
## Average Annual number of Revolutions             0.000000    0.9703704
## Average no. of Annual Political Assassinations   0.000000    2.4666667
##                                                        units
## Average annual growth (in %)                      percentage
## Oil                                           binary (0 or 1)
## GDP per capita in 1960 (in $)                        dollars
## Share of Trade in economy (proportion)            proportion
## Years of Schooling (in years)                          years
## Average Annual number of Revolutions                   count
## Average no. of Annual Political Assassinations         count
```

Hence the above table showcases the sample mean, standard deviation, and minimum and maximum values for the series Growth, Tradeshare, YearsSchool, Oil, Rev_Coups, Assassinations, and RGDP60.

ii)

Ans. We use the function `lm()` to run a regression of the dependent varibale (Growth) on the dependent variables (TradeShare, YearsSchool, Rev_Coups, Assassinations, and RGDP60). This Multiple linear Regression Model can be written in the following manner:

$$growth = \beta_0 + \beta_1 tradeshare + \beta_2 yearsschool + \beta_3 rev_coups + \beta_4 assasinations + \beta_5 rgdp60e + \mu$$

Now, we run the regression in the following manner:

```
growth_lm <- lm(growth~tradeshare+yearsschool+rev_coups+assasinations+rgdp60)
growth_lm
```

```
##
## Call:
## lm(formula = growth ~ tradeshare + yearsschool + rev_coups +
##     assasinations + rgdp60)
##
## Coefficients:
##   (Intercept)     tradeshare    yearsschool      rev_coups  assasinations
##     0.6268915      1.3408193      0.5642445     -2.1504256      0.3225844
##        rgdp60
##    -0.0004613
```

The value of the coefficient on rev_coups is `-2.1504256`. The value of this coefficient can be estimated as follows:

Interpretation of β_3: Keeping the other variables of trashare, assasinations and rgdp60 constant, with a unit increase in the average annual number of revolutions, insurrections (successful or not) and coup d'etats in any particular country from 1960 to 1995, then, on average, the average annual percentage growth of real GDP from 1960 to 1995 of that particular country would have decreased by **2.1504256 %**.

iii)

Ans. In order to predict the average annual growth rate for a country that has average values for all regressors, we will substitute the mean values of the regressors calculated in part (i).

We know that:

$$growth = \beta_0 + \beta_1 tradeshare + \beta_2 yearsschool + \beta_3 rev_coups + \beta_4 assasinations + \beta_5 rgdp60e + \mu$$

Substituting the values of coefficients:

$$growth = 0.627 + 1.341 tradeshare + 0.564 yearsschool - 2.150 rev_coups + 0.282 assasinations - 0.00046 rgdp60 + \mu$$

Substituting the values of the coefficents:

$$growth = 0.627 + 1.341 * 0.542 + 0.564 * 3.959 - 2.150 * 0.17006 + 0.282 * 0.3225 - 0.00046 * 3130.8125$$

$$growth = 1.869085876$$

Thus, the predicted average annual growth rate for a country that has average values for all regressors is **1.869085876 %.**

iv)

Ans From the previous part, we know that:

$$growth = 0.627 + 1.341 tradeshare + 0.564 yearsschool - 2.150 rev_coups + 0.282 assasinations - 0.00046 rgdp60 + \mu$$

However, now the regressor `tradeshare` changes as we take the value "one standard deviation above the mean". Hence, the change in the regressor `tradeshare` will be as follows:

$$tradeshare = mean(tradeshare) + sd(tradeshare)$$

$$tradeshare = 0.5423919 + 0.2283326$$

$$tradeshare = 0.7707245$$

Substituting the new value in the regression equation:

$$growth = 0.627 + 1.341 * 0.7707 + 0.564 * 3.959 - 2.150 * 0.17006 + 0.282 * 0.3225 - 0.00046 * 3130.8125$$

$$growth = 2.175239$$

Thus, the predicted average annual growth rate will now be **2.175239 %.**

Problem 2

Data on the weekly sales of a major brand of canned tuna by a supermarket chain in a large midwestern U.S. city during a mid-1990s calendar year are contained in the file `tuna`. There are 52 observations on the variables:

- SAL1 = unit sales of brand no. 1 canned tuna
- APR1 = price per can of brand no. 1 canned tuna
- APR2, APR3 = price per can of brands nos. 2 and 3 of canned tuna
- DISP = an indicator variable that takes the value one if there is a store display for brand no. 1 during the week but no newspaper ad; zero otherwise
- DISPAD = an indicator variable that takes the value one if there is a store display and a newspaper ad during the week; zero otherwise

i. Estimate, by least squares, the log-linear model

$$ln(SALI) = \beta_0 + \beta_1 APR1 + \beta_2 APR2 + \beta_3 APR3 + \beta_4 DISP + \beta_5 DISPAD + e$$

ii. Discuss and interpret the estimates of β_1, β_2, and β_3.

iii. Are the signs and relative magnitudes of the estimates of β_4 and β_5 consistent with economic logic?

iv. Test, at the $\alpha = 0.05$ level of significance, each of the following hypotheses:

   a. $H_0 : \beta_4 = 0, H_1 : \beta_4 \neq 0$
   b. $H_0 : \beta_5 = 0, H_1 : \beta_5 \neq 0$
   c. $H_0 : \beta_4 = 0, \beta_5 = 0, H_1 : \beta_4 or \beta_5 \neq 0$
   d. $H_0 : \beta_5 \leq \beta_4, H_1 : \beta_5 > \beta_4$

v. Discuss the relevance of the hypothesis tests in (iv) for the supermarket chain's executives.

i)

Ans. In order to run a regression, we first load the database in our R environment

```r
library(readxl)
tuna <- read_excel("tuna.xlsx")
attach(tuna)
```

After doing so, we run a regression of log of (SAL1) (dependent variable/regressand) on APR1, APR2, APR3, DISP and DISPAD (independent variables/ regressors) using the function `lm()`.

```r
sal_reg <- lm(log(sal1)~apr1+apr2+apr3+disp+dispad)
summary(lm(log(sal1)~apr1+apr2+apr3+disp+dispad))
```

```
##
## Call:
## lm(formula = log(sal1) ~ apr1 + apr2 + apr3 + disp + dispad)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70001 -0.21573 -0.03785  0.26241  0.74457
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.9848     0.6464  13.900  < 2e-16 ***
## apr1         -3.7463     0.5765  -6.498 5.17e-08 ***
## apr2          1.1495     0.4486   2.562 0.013742 *
## apr3          1.2880     0.6053   2.128 0.038739 *
## disp          0.4237     0.1052   4.028 0.000209 ***
## dispad        1.4313     0.1562   9.165 6.04e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3397 on 46 degrees of freedom
## Multiple R-squared:  0.8428, Adjusted R-squared:  0.8257
## F-statistic: 49.33 on 5 and 46 DF,  p-value: < 2.2e-16
```

Hence, the model can be estimated in the following manner:

$$\log(\text{sal1}) = 8.9848 - 3.7463 \times \text{apr1} + 1.1495 \times \text{apr2} + 1.2880 \times \text{apr3} + 0.4237 \times \text{disp} + 1.4313 \times \text{dispad}$$

ii)

The estimated value of β_1= `-3.7463`

Interpretation of β_1 : β_1 shows the partial elasticity of Y with respect to APR1. This indicates that with a unit increase in price per can of brand no. 1 canned tuna, on average, the unit sales of brand no. 1 canned tuna decreases by 3.7463 %.

The estimated value of β_2= 1.1495

Interpretation of β_2 : β_2 shows the partial elasticity of Y with respect to APR2. This indicates that with a unit increase in price per can of brand no. 2 canned tuna, on average, the unit sales of brand no. 1 canned tuna increases by 1.1495 %.

The estimated value of β_3= 1.2880

Interpretation of β_3 : β_3 shows the partial elasticity of Y with respect to APR3. This indicates that with a unit increase in price per can of brand no. 3 canned tuna, on average, the unit sales of brand no. 1 canned tuna increases by 1.2880 %.

iii)

Ans. The positive signs for both β_4 and β_5 align with economic logic. It implies that having a store display (DISP) and both a store display and a newspaper ad (DISPAD) are associated with higher unit sales of brand no. 1 canned tuna since more is the advertising, more people would get to know about the brand and hence, visit the store to buy the canned tuna more.

The magnitudes suggest that $\beta_5 = 1.4313$ (coefficient for DISPAD) has a larger effect than $\beta_4 = 0.4237$ (coefficient for DISP) on SAL1. This is also consistent with economic logic since in case of DISPAD, both displays and newspaper ads are being used which is likely to reach more potential customers as compared to DISP.

iv)

Ans. Given: $\alpha = 0.05$

a.

$$\text{Null Hypothesis } (H_0): \quad H_0 : \beta_4 = 0 \tag{1}$$
$$\text{Alternative Hypothesis } (H_1): \quad H_1 : \beta_4 \neq 0 \tag{2}$$
$$\tag{3}$$

$$t = \frac{\beta_4 - 0}{\text{SE}(\beta_4)} \tag{4}$$
$$t = \frac{0.4237 - 0}{0.1052} \tag{5}$$
$$t = 4.027567 \tag{6}$$

Thus, p value:

```
summary(sal_reg)$coefficients["disp", "Pr(>|t|)"]
```

```
## [1] 0.0002089866
```

Here p value = `0.0002089866` is < 0.05

Hence, we reject the H_o and we can say that there is statistical evidence to support that β_4 (i.e, store display for tuna without newspaper ad)is statistically different from 0 and affects SAL1 (i.e., unit sales of brand no. 1 canned tuna).

b

$$\text{Null Hypothesis } (H_0): \quad H_0 : \beta_5 = 0 \tag{7}$$
$$\text{Alternative Hypothesis } (H_1): \quad H_1 : \beta_5 \neq 0 \tag{8}$$
$$\tag{9}$$

$$t = \frac{\beta_5 - 0}{\text{SE}(\beta_5)} \tag{10}$$

$$t = \frac{1.4313 - 0}{0.1562} \tag{11}$$

$$t = 9.163892 \tag{12}$$

Thus, p value:

```
summary(sal_reg)$coefficients["dispad", "Pr(>|t|)"]
```

```
## [1] 6.03864e-12
```

Here p value = `6.03864e-12` is < 0.05

Hence, we reject the H_o and we can say that there is statistical evidence to support that β_5 (i.e, store display for tuna with newspaper ad)is statistically different from 0 and affects SAL1 (i.e., unit sales of brand no. 1 canned tuna).

c.

$$\text{Null Hypothesis } (H_0): \quad H_0 : \beta_4 = 0, \beta_5 = 0 \tag{13}$$

$$\text{Alternative Hypothesis } (H_1): \quad H_1 : \beta_4 or \beta_5 \neq 0 \tag{14}$$

$$\tag{15}$$

Since we are studying joint effects, we will use F statistic to test our hypothesis. we will compare the sum of square residuals in a restricted model (when β_4 and $\beta_5 = 0$) and to an unrestricted model where, β_4 and β_5 affect sal_1.

Taking linear regression model `sal_reg` and calculating its SSR,

```
ssr_ur_salreg <- sum(residuals(sal_reg)^2)
ssr_ur_salreg
```

```
## [1] 5.307252
```

The SSR for the unrestricted model is 5.307252.

Now to calculate the SSR for restricted model, we will put β_4 and $\beta_5 = 0$.

The SSR for restricted regression model can be calculated as:

```
sal_reg_rest <- lm(log(sal1)~apr1+apr2+apr3)
ssr_r_salreg <- sum(residuals(sal_reg_rest)^2)
ssr_r_salreg
```

```
## [1] 15.00229
```

The SSR for the unrestricted model is 15.00229.

Now to calculate F statistic, we use the formula:

$$F = \frac{(\text{SSR}_{\text{restricted}} - \text{SSR}_{\text{unrestricted}}) \div q}{\text{SSR}_{\text{unrestricted}} \div (n - k - 1)}$$

$$\mathrm{F} = \frac{(0.5591 - 0.3397)/2}{0.3397/52 - 5 - 1} \tag{16}$$

$$\mathrm{F} = 0.1097/0.00738478 \tag{17}$$

$$\mathrm{F} = 14.8548772 \tag{18}$$

$$\tag{19}$$

Thus, F stat = 14.8548772 whereas, the F critical = 3.2

Hence, F stat is greater than F critical and the H_o can be **rejected**. We can say that there is statistical evidence to support that β_4 and β_5 (i.e, store display for tuna without newspaper ad and with newspaper ads) are jointly statistically different from 0 and affect SAL1 (i.e., unit sales of brand no. 1 canned tuna).

d.

$$\text{Null Hypothesis } (H_0): \quad H_0 : \beta_5 - \beta_4 <= 0 \tag{20}$$

$$\text{Alternative Hypothesis } (H_1): \quad H_1 : \beta_5 - \beta_4 > 0 \tag{21}$$

$$\tag{22}$$

$$t = \frac{\hat{\beta}_5 - \hat{\beta}_4}{\mathrm{SE}(\beta_5 - \beta_4)} \tag{23}$$

$$t = \frac{1.4313 - 0.4237}{\mathrm{SE}(\beta_5 - \beta_4)} \tag{24}$$

$$\tag{25}$$

$$\mathrm{SE}(\beta_5 - \beta_4) = \sqrt{\mathrm{var}(\beta_5) + \mathrm{var}(\beta_4) - 2\mathrm{cov}(\beta_5, \beta_4)} \tag{26}$$

To calculate SE, kindly check the following code:

```
tuna_unrest <-lm(log(sal1)~apr1+apr2+apr3+disp+dispad)
tuna_res <-lm(log(sal1)~apr1+apr2+apr3)
se_disp <- summary(tuna_unrest)$coefficients["disp", "Std. Error"]
se_dispad <- summary(tuna_unrest)$coefficients["dispad", "Std. Error"]
var_disp <- se_disp^2
var_dispad <- se_dispad^2
cov_matrix_tuna <- vcov(tuna_unrest)
cov_b4_b5 <- cov_matrix_tuna["disp", "dispad"]
se_b4_b5 <- sqrt(var_disp+var_dispad-(2*cov_b4_b5))
se_b4_b5
```

```
## [1] 0.1469155
```

Thus,

$$\mathrm{SE}(\beta_5 - \beta_4) = 0.1469155 \tag{27}$$

Using (27) in (24)

$$t = \frac{\hat{\beta}_5 - \hat{\beta}_4}{\mathrm{SE}(\beta_5 - \beta_4)} \tag{28}$$

$$t = \frac{1.4313 - 0.4237}{0.1469155} \tag{29}$$

$$t = 6.858364 \tag{30}$$

Thus $t = 6.858364$.

The corresponding p-value is:

```
p_value <- (1 - pt(abs(6.858364), 46))
p_value
```

```
## [1] 7.432872e-09
```

The p-value 7.432872e-09 is less than α value of 0.05. Hence, **we reject the H_o.** We can say that there is statistical evidence to support that β_5 (i.e, store display for tuna with newspaper ads) is not less than or equal to β_4 (i.e, store display for tuna without newspaper ad).

v)

The hypothesis tests conducted in (iv) are quite relevant for supermarket chains' executives as these tests tell us what is more efficient in the sales of brand no.1 canned tuna, in terms of advertising only through display or with a combination of diplay and newspaper ads. The following findings are significant:

- Advertising only with store displays causes a difference in sales of branded no.1 canned tuna (it is statistically different from 0 moreover coefficient is positive so it will cause an increase)
- Advertising with store displays + newspaper ads causes a difference in sales of branded no.1 canned tuna (it is statistcally different from 0 moreover coefficient is positive so it will cause an increase)
- The variables disp and dispad jointly affect the sales of brand no.1 canned tuna
- Lastly, there is some statistical evidence that advertising with newspapers and store displays both is more effective than advertising only through store displays.

All of these points are important for supermarket chains' executives in order to reform their marketing strategies better and reach more potential customers.

Problem 3

The file `cocaine` contains 56 observations on variables related to sales of cocaine powder in northeastern California over the period 1984–1991. The data are a subset of those used in the study Caulkins, J. P. and R. Padman (1993), "Quantity Discounts and Quality Premia for Illicit Drugs," *Journal of the American Statistical Association*, 88, 748–757. The variables are:

- PRICE = price per gram in dollars for a cocaine sale
- QUANT = number of grams of cocaine in a given sale
- QUAL = quality of the cocaine expressed as percentage purity
- TREND = a time variable with 1984 = 1 up to 1991 = 8

Consider the regression model-

$$PRICE = \beta_0 + \beta_1 QUANT + \beta_2 QUAL + \beta_3 TREND + e$$

i. What signs would you expect on the coefficients β_1, β_2, and β_3?

ii. Estimate the given regression equation. Report the results and interpret the coefficient estimates. Have the signs turned out as you expected?

iii. What proportion of variation in cocaine price is explained jointly by variation in quantity, quality, and time?

iv. It is claimed that the greater the number of sales, the higher the risk of getting caught. Thus, sellers are willing to accept a lower price if they can make sales in larger quantities. Set up H_0 and H_1 that would be appropriate to test this hypothesis. Carry out the hypothesis test.

v. Test the hypothesis that the quality of cocaine has no influence on price against the alternative that a premium is paid for better-quality cocaine.

vi. What is the average annual change in the cocaine price? Can you suggest why price might be changing in this direction? *(Hint: focus on the TREND variable)*

i)

Ans. Before starting, we load the dataset in R environment:

```
library(readxl)
cocaine <- read_excel("cocaine.xlsx")
```

The estimated model is given as:

$$PRICE = \beta_0 + \beta_1 QUANT + \beta_2 QUAL + \beta_3 TREND + e$$

We expect all the signs to be positive because: - The higher the quantity is will lead to higher average price demanded by the sellers.(positive relationship) - The higher the quality is, the higher will be the cost of production and hence it will be sold at a higher price - Due to inflation, over the years, price levels generally increase. Hence, as more time passes by, the price of cocaine is also expected to go up in order for sellers to keep up with the price of the market.

ii)

We can estimate linear regression using the function `lm()`.

```
attach(cocaine)
lm_cocaine <- lm(price~ quant+qual+trend)
lm_cocaine
```

```
##
## Call:
## lm(formula = price ~ quant + qual + trend)
##
## Coefficients:
## (Intercept)        quant         qual        trend
##    90.84669     -0.05997      0.11621     -2.35458
```

Thus, the model can be written as follows:

$$PRICE = 90.84669 - 0.05997 QUANT + 0.11621 QUAL - 2.35458 TREND$$

Interpretation of β_1: With a 1 gram increase in the quantity of cocaine, on average, the price per gram in dollars for a cocaine sale decreases by **$0.05997**.

Interpretation of β_2: With a 1 purity percentage increase in the quality of cocaine, on average, the price per gram in dollars for a cocaine sale increases by **$0.11621**.

Interpretation of β_3: With a 1 year increase in the quantity of cocaine, on average, the price per gram in dollars for a cocaine sale decreases by **$2.35458**.

The signs are not as expected because dealing a larger quantity induces a risk of getting caught by law enforcers in the sellers. hence, this can explain the negative relationship between prices and quantity.

For trend, the negative sign might indicate a decreasing trend in cocaine prices over the specified period. Increased law enforcement and success in combating drug trafficking might be leading to a decrease in prices over time. This could be a result of disruptions to the supply chain and explain the negative relationship.

iii)

The proportion of variation in cocaine price is explained jointly by variation in quantity, quality, and time can be explained by R^2. $\mathbf{R^2}$ determines the fit of the model and accounts for the variation happening in Y due to independent variables.

```
cocaine_r <- summary(lm_cocaine)$r.squared
cocaine_r
```

```
## [1] 0.50965
```

Hence, **50.96%** of the variation in price per ram of cocaine is explained by variation in quantity, quality, and time.

iv)

$$\text{Null Hypothesis } (H_0): \quad H_0 : \beta_1 >= 0 \tag{31}$$
$$\text{Alternative Hypothesis } (H_1): \quad H_1 : \beta_1 < 0 \tag{32}$$
$$\tag{33}$$

- Null Hypothesis H_o: Quantity and price move in same directions
- Alternate Hypothesis H_1 : Quantity and price move in opposite directions.

$$t = \frac{\beta_1 - 0}{\text{SE}(\beta_1)} \tag{34}$$
$$t = \frac{-0.05997 - 0}{0.01018} \tag{35}$$
$$t = 5.890963 \tag{36}$$

Thus, p value:

```
p_cocaine_quant <- 2 * (1 - pt(abs(5.890963), 52))
p_cocaine_quant
```

```
## [1] 2.860797e-07
```

Thus, p value is less than α value of 0.05

Hence, we reject the H_o. This indicates that it is statistically significant to show that there is some statistical support that sellers are willing to accept a lower price if they can make sales in larger quantities.

v)

Given:

$$\text{Null Hypothesis } (H_0): \quad H_0 : \beta_2 = 0 \tag{37}$$
$$\text{Alternative Hypothesis } (H_1): \quad H_1 : \beta_2 > 0 \tag{38}$$
$$\tag{39}$$

- Null Hypothesis H_o: Quality has no effect on price
- Alternate Hypothesis H_1 : Quality has a positive effect on price

$$t = \frac{\beta_2 - 0}{\text{SE}(\beta_2)} \tag{40}$$
$$t = \frac{0.11621 - 0}{0.20326} \tag{41}$$
$$t = 0.571730788 \tag{42}$$

The corresponding p value will be:

```
p_cocaine_qual <- 2*(1-pt(0.571730788, 52))
p_cocaine_qual
```

```
## [1] 0.5699677
```

Here the p-value is greater than α value of 0.05.

Thus, we fail to reject the H_o.

This indicates that it is is statistically significant that we fail to reject that quality does not have an effect on price per gram of cocaine.

vi)

Ans. The average annual change in the price of cocaine is -2.35458 (a decrease of $2.35458 annually). A possible reason behind this can be changes in drug policies, both domestically and internationally, may influence the drug trade. For example, policies aimed at reducing drug demand, would force sellers also to sell cocaine at a lower price.

Problem 4

Use the data in `ATTEND` from *Wooldridge* package for this exercise.

i. Obtain the minimum, maximum, and average values for the variables *atndrte*, *priGPA*, and *ACT*.

ii. Estimate the given model and write the results in equation form. Interpret the intercept. Does it have a useful meaning?

$$atndrte = \beta_0 + \beta_1 priGPA + \beta_2 ACT + u$$

iii. Discuss the estimated slope coefficients. Are there any surprises?

iv. What is the predicted *atndrte* if $priGPA = 3.65$ and $ACT = 20$? What do you make of this result? Are there any students in the sample with these values of the explanatory variables?

v. If Student A has $priGPA = 3.1$ and $ACT = 21$ and Student B has $priGPA = 2.1$ and $ACT = 26$, what is the predicted difference in their attendance rates?

i)

Ans. The minimum, maximum nd average values for the variables atndrte, priGPA, and ACT can be found using the following functions:

- min() : minimum values
- max(): maximum values
- mean() : average values

```
library(wooldridge)
attach(attend)
min_values <- c(min(atndrte), min(priGPA), min(ACT))
max_values <- c(max(atndrte), max(priGPA), max(ACT))
mean_values <- c(mean(atndrte), mean(priGPA), mean(ACT))

#A table can be constructed using this data
summary_table <- data.frame(
  Variable = c("atndrte", "priGPA", "ACT"),
  Minimum = min_values,
  Maximum = max_values,
  Mean = mean_values
)

print(summary_table)
```

```
##   Variable Minimum Maximum      Mean
## 1  atndrte   6.250  100.00 81.709559
## 2   priGPA   0.857    3.93  2.586775
## 3      ACT  13.000   32.00 22.510294
```

ii)

Ans. The model can be estimated using the function `lm()`. It is given as follows:

```
lm(atndrte~priGPA+ACT)
```

```
##
## Call:
## lm(formula = atndrte ~ priGPA + ACT)
##
## Coefficients:
## (Intercept)       priGPA          ACT
##      75.700       17.261       -1.717
```

Thus, the estimated model will look like:

$$atndrte = 75.7 + 17.261 \times priGPA - 1.717 \times ACT$$

Interpretation of the intercept β_0: When the priGPA (cumulative GPA prior to term) and ACT (ACT score) is equal to 0, then on average, the atndrte (percent classes attended) is 75.7 %.

The intercept does not have a meaningful interpretation because ACT is graded on a scale of 1 to 36. Hence, achieving a score of 0 on the ACT is not possible.

iii)

Ans.

Interpretation of the β_1: The coefficient for priGPA is positive (17.261), indicating that as the cumulative GPA prior to the term (priGPA) increases by one unit, the predicted percentage of classes attended (atndrte) is expected to increase by 17.261 %, assuming ACT scores remain constant.

Interpretation of the β_2: The coefficient for ACT is negative (-1.717), suggesting that as the ACT score increases by one unit, the predicted percentage of classes attended is expected to decrease by 1.717 %, holding priGPA constant.

The negative coefficient for ACT suggests that, on average, as the ACT score increases, the percentage of classes attended tends to decrease. This seems counterintuitive, as one might assume that students with higher standardized test scores would attend classes more regularly. This can be an indication that the model does not include other factors influencing this relationship.

iv)

Given:

- priGPA= 3.65
- ACT = 20

$$atndrte = 75.7 + 17.261 \times 3.65 - 1.717 \times 20$$

$$atndrte = 75.7 + 17.261 \times 3.65 - 1.717 \times 20$$

$$atndrte = 104.36265$$

This result fails to make sense since highest attendance is capped at 100%. Values above this limit do not make sense. It could be an indication of limitations in the model or issues related to extrapolation beyond the observed range of the data.

To check the data, with priGPA= 3.65 & ACT = 20 we can us ethe function subset() and nrow() to count the number of students with thes evalue sof the explanatory variable:

```
nrow(subset(attend, ACT ==20 & priGPA== 3.65))
```

```
## NULL
```

Thus, there is no student in the sample with these explanatory variables and this makes sense as well since, it is not possible to have 104.36265% attendance.

v)

Given:

Student A

- priGPA= 3.1
- ACT = 21

Student B

- priGPA= 2.1
- ACT = 26

Attendance rate of Student A:

$$atndrte_A = 75.7 + 17.261 \times 3.1 - 1.717 \times 21$$

$$atndrte_A = 93.1521$$

Attendance rate of Student B:

$$atndrte_B = 75.7 + 17.261 \times 2.1 - 1.717 \times 26$$

$$atndrte_B = 67.3061$$

Thus, the difference between Student A and Student B's attendance rate is:

$$atndrte_d = atndrte_A - atndrte_B$$
$$atndrte_d = 93.1521 - 67.3061$$

$$atndrte_d = 25.846$$

Thus, the difference between the attendance rate of Student A and B is **25.846 %**.

Problem 5

Consider the following regression equation-

$$\widehat{Price} = 109.7 + 0.567BDR + 26.9Bath + 0.239Hsize + 0.005Lsize + 0.1Age - 56.9Poor \quad (43)$$
$$(22.1) \quad (1.23) \quad (9.76) \quad (0.021) \quad (0.00072) \quad (0.23) \quad (12.23) \quad (44)$$
$$R^2 = 0.85, SER = 45.8 \quad (45)$$

Where:

- Price: the selling price (in $1000)

- BDR: the number of bedrooms
- Bath: the number of bathrooms
- Hsize: the size of the house (in square feet)
- Lsize: the lot size (in square feet)
- Age: the age of the house (in years)
- Poor: a binary variable that is equal to 1 if the condition of the house is reported as "poor."

i. Is the coefficient on *BDR* statistically significantly different from zero?

ii. Typically four-bedroom houses sell for more than three-bedroom houses. Is this consistent with your answer to (i), and with the regression in general?

iii. Lot size is measured in square feet. Do you think that another scale might be more appropriate? Why or why not?

iv. The *F*-statistic for omitting *BDR* and *Age* from the regression is $F = 2.38$. Are the coefficients on *BDR* and *Age* statistically different from zero at the 10% level?

i)

Ans. Given:

- Assuming level of significance (α) = 0.05

$$\widehat{Price} = 109.7 + 0.567BDR + 26.9Bath + 0.239Hsize + 0.005Lsize + 0.1Age - 56.9Poor \quad (46)$$

$$(22.1) \quad (1.23) \quad (9.76) \quad (0.021) \quad (0.00072) \quad (0.23) \quad (12.23) \quad (47)$$

$$R^2 = 0.85, SER = 45.8 \quad (48)$$

To test: - If β_1 is significant or not (coefficient on BDR)

Sol.

$$\text{Null Hypothesis } (H_0): \quad H_0 : \beta_1 = 0 \quad (49)$$

$$\text{Alternative Hypothesis } (H_1): \quad H_1 : \beta_1 \neq 0 \quad (50)$$

$$(51)$$

$$t = \frac{\beta_1 - 0}{\text{SE}(\beta_1)} \quad (52)$$

$$t = \frac{0.567 - 0}{1.23} \quad (53)$$

$$t = 0.46097561 \quad (54)$$

The corresponding p-value for this test will be (using `pt()` function):

```
p_val_bdr <- 2 * (1 - pt(abs(0.46097561), df= Inf))
p_val_bdr
```

```
## [1] 0.6448161
```

Here, p_value (0.6448161) is greater than the level of significance α (0.05).

Thus, we fail to reject the H_o.

This indicates that the coefficient on BDR is not statistically significant and we fail to reject the assumption that number of bedrooms might not have any effect on the selling price of the house.

ii)

Ans. To determine whether four-bedroom houses sell for more than three-bedroom houses, we can check the change in price with respect to change in bedrooms (while keeping the other variables constant in the regression model).

$$Price = 0.567 \times 4$$

$$Price = 2.268$$

In a model where price only depends on number of bedrooms, the price of a 4-bedroom house would be \$2268.

For a 3 bedroom house:

$$Price = 0.567 \times 3$$

$$Price = 1.701$$

In a model where price only depends on number of bedrooms, the price of a 4-bedroom house would be \$1701.

Thus, this statement is consistent with the regression as the number of rooms increase, on average, the price also increases, hence, four-bedroom houses sell for more than three-bedroom houses.

However, from the results in part (i), coefficient on BDR is not statistically significant and we cannot confidently say it's different from zero, interpreting the impact of the number of bedrooms on the selling price might be challenging.

The comparison made in part (ii) assumes a positive impact of bedrooms on price, which may not be consistent with the result from part (i). To maintain consistency, it might be more appropriate to say that, based on the available evidence, we don't have sufficient confidence to claim a statistically significant positive impact of the number of bedrooms on the selling price.

iii)

The coefficient for "Lsize" in the regression equation is given as 0.005. This means that, on average, a one-unit increase in lot size is associated with a 0.005-unit increase in the predicted selling price. This change is quite small and seems a little insignificant. However, if we scale Lsize in thousand of square feet then,

$$Lsize = \frac{Lsize}{1000}$$

This will push up the coefficent in the following manner:

$$\widehat{Price} = 109.7 + ... + 0.005 \times 1000 Lsize + ...$$

Hence, the transformed equation becomes:

$$\widehat{Price} = 109.7 + ... + 5 \times Lsize + ...$$

This makes the explanatory variable more eaisly interpretable and the change in selling price of house also seems more significant.

iv)

Given:

- $F_{stat} = 2.38$ (when β_1 (coeff for BDR) and β_5= 0 (coeff for AGE))
- $\alpha = 0.1$
- df= 2 (numerator) and infinite (denominator)

Sol.

$$\text{Null Hypothesis } (H_0): \quad H_0 : \beta_1 = 0, \beta_5 = 0 \tag{55}$$

$$\text{Alternative Hypothesis } (H_1): \quad H_1 : \beta_1 or \beta_5 \neq 0 \tag{56}$$

$$\tag{57}$$

Using F statistic,

$$F = \frac{(\text{SSR}_{\text{restricted}} - \text{SSR}_{\text{unrestricted}}) \div q}{\text{SSR}_{\text{unrestricted}} \div (n - k - 1)}$$

We know, $F_{stat} = 2.38$

Now calculating F critical at df= 2 (numerator) and infinite (denominator).

Thus $F_{critical}$ at $\alpha = 0.1$ is found to be 2.3

Thus, $F_{stat} > F_{critical}$. Hence, we **reject the** H_o**.**

This indicates that there is statistical evidence to support that β_1 (coefficient for number of bedrooms) and β_5 (coefficient for age) are jointly statistically different at 10% level from 0 and have an effect on the selling price of the house.