ROLL NO.- A21

SAP NO. -40525220022

CLASS- MA ECONOMICS

SUBJECT- ECONOMETRICS WITH R, PART 1

SEMISTER- 1

ECONOMETRICS PROJECT

1)Dataset: wine. Wooldridge Source: These data were reported in a New York Times article, December 28, 1994.

A data frame of 21 observations (country wise) on 5 variables of which two variables taken are alcohol (litres of alcohol consumed from wine, per capita) and liver (liver disease death per 100,000).

OUTPUT:

Correlation:

```
> cor(wine$liver,wine$alcohol)
[1] 0.7371928
```

Regression:

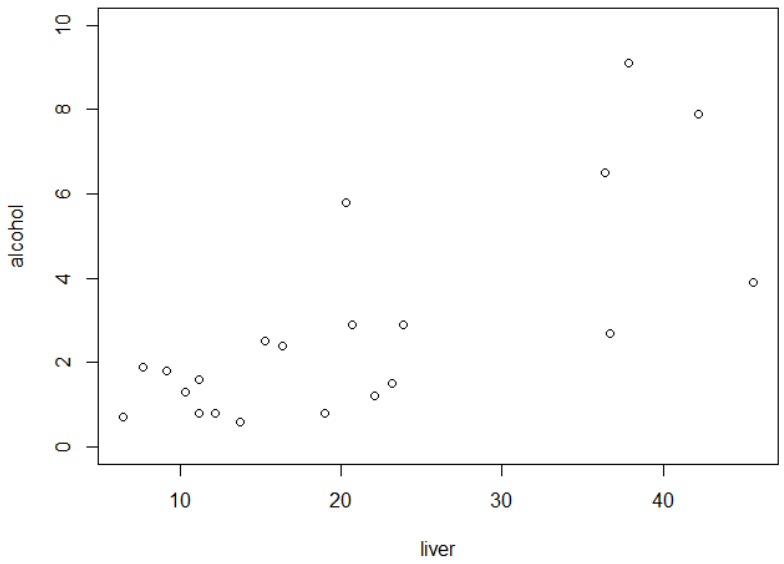
```
Call:
lm(formula = liver ~ alcohol, data = wine)

Residuals:
    Min      1Q  Median      3Q     Max
-11.356  -5.393  -1.524   3.013  20.758

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.8548     2.8024   3.873 0.001023 **
alcohol       3.5864     0.7541   4.756 0.000138 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.29 on 19 degrees of freedom
Multiple R-squared:  0.5435,    Adjusted R-squared:  0.5194
F-statistic: 22.62 on 1 and 19 DF,  p-value: 0.0001375
```

SCATTER PLOT:



INTERPRETATION 1:

Using the data mentioned above, the OLS regression model for 2 variables is given as

$Y_i = B0 + B1X_i + e_i$.

Here,

Y_i = independent variable, represents alcohol consumption in litres. (only wine).

X_i = dependent variable, represents death caused by liver disease.

B0 = intercept, i.e value of Yi if Xi is 0.

B1 is the regression coefficient, i.e change in Xi for 1 unit change in Yi.

H0: There is no effect of consumption of alcohol on deaths caused by liver disease.

H1: There is an effect of alcohol consumption on deaths caused by liver disease.

Correlation coefficient: 0.73

There is significant and positive correlation between the two variables meaning that an increase or decrease in Yi will lead to increase or decrease in Xi respectively.

Residuals: minimum value= -11.36, maximum value= 20.76

The range between minimum and maximum values of the residuals is too high that is the difference between the predicted value of Yi and observed value of Yi is too high. That means that the model predicts certain points that fall far away from the actual observed points. Hence, the model is not a good fit.

B0 value= 10.85 B1 value= 3.59

F- statistic: 22.62

The further away the F- statistic is from 1, the stronger the relationship between the predictor variable (alcohol) and response variable (liver disease deaths). For this model the f- statistic is significantly greater than 1, so it is sufficient to reject the null hypothesis.

P-value: 0.000137

If p- value is less than 5%, we can reject the null hypothesis. In this case the P-value is significantly less.

Hence, we reject the null hypothesis.

Conclusion: There is a significant effect of alcohol consumption (wine) in the 21 countries mentioned in the data on deaths caused by liver disease in those countries.

2) Dataset: gpa1. Wooldridge Source: Christopher Lemmon, a former MSU undergraduate, collected these data from a survey he took of MSU students in Fall 1994.

A data frame with 141 observations on 29 variables of which the 2 variables used were skipped (average lectures missed per week) and colgpa (student's GPA at MSU).

OUTPUT:

Correlation:

```
> cor(fertil1$educ, fertil1$kids)
[1] -0.2230473
```

Regression:

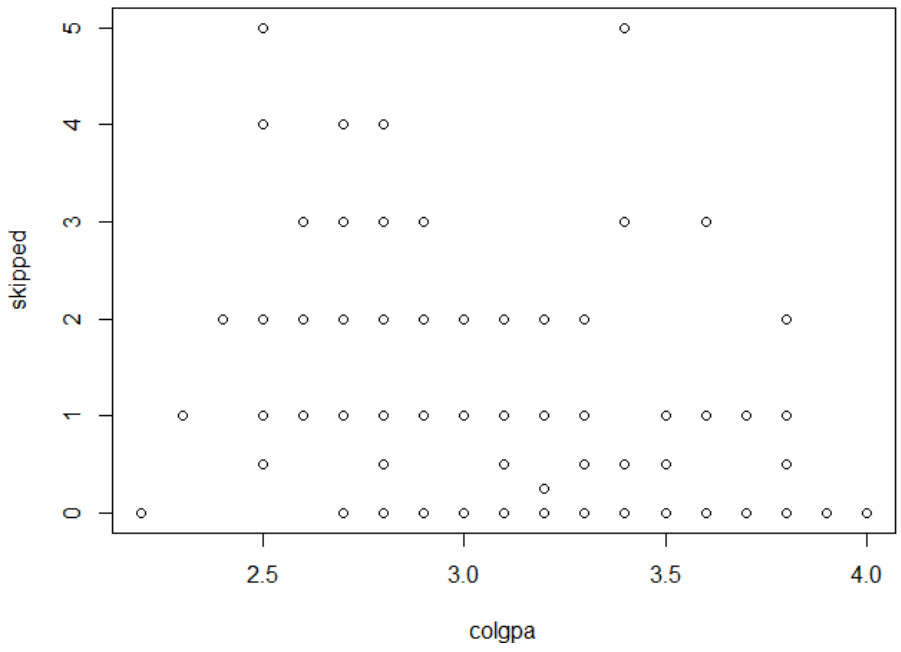
```
Call:
lm(formula = colgpa ~ skipped, data = gpa1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.95308 -0.25308 -0.06356  0.24692  0.84692

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.15308    0.04278  73.713  < 2e-16 ***
skipped     -0.08952    0.02799  -3.198  0.00171 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3606 on 139 degrees of freedom
Multiple R-squared:  0.06855,   Adjusted R-squared:  0.06185
F-statistic: 10.23 on 1 and 139 DF,  p-value: 0.001712
```

SCATTER PLOT:



INTERPRETATION 2:

Using the data mentioned above, the OLS regression model for 2 variables is given as

$Y_i = B0 + B1X_i + e_i$.

Here,

Y_i = independent variable, represents the number of lectures skipped in one week.

X_i = dependent variable, represents a student's GPA in college.

B0 = intercept, i.e value of Yi if Xi is 0.

B1 is the regression coefficient, i.e change in Xi for 1 unit change in Yi.

H0: The number of lectures skipped in a week has a positive relationship with student's GPA.

H1: The number of lectures skipped in a week has a negative relationship with student's GPA.

Correlation coefficient= -0.22

There is a week and negative correlation between the two variables meaning that a greater number of lectures skipped will lead to the students GPA going down and a smaller number of lectures skipped will lead to the GPA increasing.

Residuals: minimum value=-0.95, maximum value= 0.85

The range between minimum and maximum values of the residuals is low that is the difference between the predicted value of Yi and observed value of Yi is close to 0. That means that the model predicts points that fall close to the actual observed points. Hence, the model is a good fit.

B0 value= 3.15 B1 value= 0.090

F- statistic: 10.32

For this model the f- statistic is greater than 1, so it is sufficient to reject the null hypothesis.

P-value: 0.0017

If p- value is less than 5%, we can reject the null hypothesis. In this case the P-value is significantly less.

Hence, we reject the null hypothesis.

Conclusion: The number of lectures a student at MSU skips in a week has a negative effect on their GPA.

3)Dataset: fertil1, Wooldridge Source: W. Sander, “The Effect of Women’s Schooling on Fertility,” Economics Letters 40, 229-233.Professor Sander kindly provided the data, which are a subset of what he used in his article. He compiled the data from various years of the National Opinion Resource Centre’s General Social Survey.

A data frame with 1129 observations on 27 variables of which 2 variables used were educ (years of schooling), and kids (number of kids born).

OUTPUT

Coefficient:

```
> cor(gpa1$skipped,gpa1$colGPA)
[1] -0.26182
```

Regression:

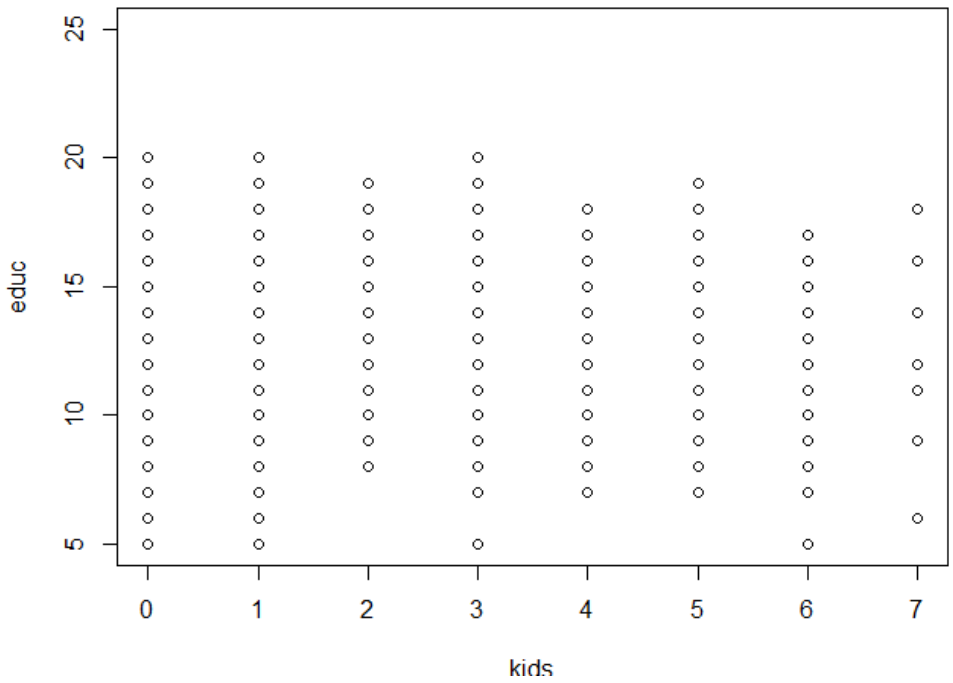
```
Call:
lm(formula = kids ~ educ, data = fertil1)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3766 -0.8397  0.0206  1.1603  4.9987

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.51632    0.23578  19.155  < 2e-16 ***
educ        -0.13972    0.01819  -7.681  3.4e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.613 on 1127 degrees of freedom
Multiple R-squared:  0.04975,   Adjusted R-squared:  0.04891
F-statistic:    59 on 1 and 1127 DF,  p-value: 3.404e-14
```

SCATTER PLOT



INTERPRETATION 3:

Using the data mentioned above, the OLS regression model for 2 variables is given as

$Yi = B0 + B1Xi + ei.$

Here,

Yi = independent variable, represents the number of years of schooling.

Xi = dependent variable, represents the number of children a woman has.

B0 = intercept, i.e value of Yi if Xi is 0.

B1 is the regression coefficient, i.e change in Xi for 1 unit change in Yi.

H0: The number of years of schooling has a positive relationship with the number of kids a woman has.

H1: The number of years of schooling has a negative relationship with the number of kids a woman has.

Correlation coefficient= -0.27

There is a week and negative correlation between the two variables meaning more number of years of schooling will lead to women having slightly less children.

Residuals: minimum value=-4.38, maximum value= 4.99

The range between minimum and maximum values of the residuals is high. Hence, the model is not a good fit.

B0 value= 4.5 B1 value= -0.14

F- statistic: 59

For this model the f- statistic is significantly greater than 1, so it is sufficient to reject the null hypothesis.

P-value: 3.404e-14

If p- value is less than 5%, we can reject the null hypothesis. In this case the P-value is significantly less.

Hence, we reject the null hypothesis.

Conclusion: The number of years a woman goes to school has a negative effect on the number of children she has.