BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI HYDERABAD CAMPUS

First Semester 2020-21

Econ F241-Econometric Methods

Assignment - I

Bhavish Pahwa - 2018A7PS0168H

Table of Contents

Dataset:-

The dataset chosen was **Computer Hardware Dataset** from UCI Machine Learning Repository.

The dataset has relative CPU Performance Data which is described in terms of cycle time , memory size etc.

The dataset has 10 attributes and 209 observations.

Attribute Information:-

1 Vendor name

2 Model

3 MYCT :- Machine Cycle Time in nanoseconds(integer)

4 MMIN :- Minimum Main Memory in Kilobytes(integer)

5 MMAX :- Maximum Main Memory in Kilobytes(integer)

6 CACH :- Cache Memory in Kilobytes(integer)

7 CHMIN:-Minimum Channels in Units(integer)

8 CHMAX:- Maximum Channels in Units(integer)

9 PRP:- Published Relative Performance(integer)

10 ERP:- Estimated Relative Performance(integer)

1. Start with a simple regression model giving justification for the selection of the dependent & independent variables.:-

Ans)So we start with a simple Linear Regression Model

$y_i = \beta 1 + \beta 2\ x_i + e_i$

The subscript i in the Equation indicates that the relationship applies to each of the 209 observations.

Here Y or the dependent variable is the PRP variable which is the Published relative performance .

The X or the Independent , Explanatory variable chosen is mmax which is the maximum main memory in kilobytes.

The dependent variable is PRP because this is what we want to predict or know about other different CPUs when we are given some internal design information about a CPU and this value clearly depends on the CPU components.

The independent variable is the Maximum Main Memory which is the explanatory variable for this regression as it tries to justify the published relative performance of the CPU as in general a CPU having more main memory will be able to have better performance due to more number of programs being able to run simultaneously on the CPU and by also allowing more memory intensive applications to run smoothly.

2. On the selected model, attempt the following:

a. Calculate some summary statistics of the variables & comment on the same

Ans) Some summary statistics of mmax the explanatory variable.

```
> mean(machine$mmax)
[1] 11796.15
> median(machine$mmax)
[1] 8000
> sd(machine$mmax)
[1] 11726.56
> max(machine$mmax)
[1] 64000
> min(machine$mmax)
[1] 64
>
```

Here mean ,median have their respective meanings . Max and min refer to maximum and minimum value respectively and sd is used to calculate the standard deviation of the variable values . From these statistics of mmax we can see that mmax has a good range of values starting from 64 going up to 64000 which is a quite good range to have in a dataset.
Looking at the mean and median value we can see that the dataset has more values which are low and it has some values which are pretty high

Some summary statistics of PRP the dependent variable.

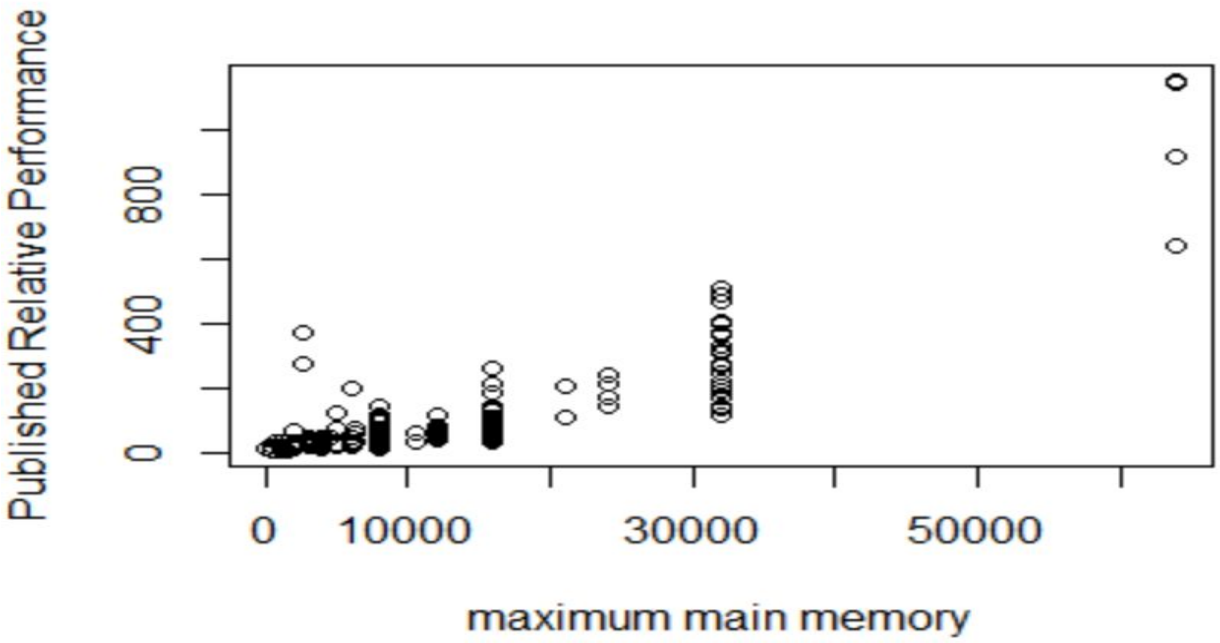
```
> mean(machine$prp)
[1] 105.622
> median(machine$prp)
[1] 50

> sd(machine$prp)
[1] 160.8307
> max(machine$prp)
[1] 1150
> min(machine$prp)
[1] 6
>
```

Here mean ,median have their respective meanings . Max and min refer to maximum and minimum value respectively and sd is used to calculate the standard deviation of the variable values .

b. Make a scatter diagram & comment on the relationship between the variables. Based on the plot, discuss whether it suggests a linear or non-linear relationship and then decide whether you should go with a linear or a non-linear model.

Ans)



As one can observe from the scatter plot there are more sample values of maximum main memory in less than 20000 region than more than 20000 which can be observed from the mean and median value of mmax. There are only 3 CPUs which have maximum main memory value more than 40000 .

By seeing the scatter plot we can also say that there is a positive relationship between mmax and prp but we notice that this relation might not be linear in nature but polynomial as the increase in value of prp with increasing

mmax value does not look constant it looks kinda polynomial or quadratic .

What we want to say in other words is say , we try to imagine the relation between mmax and prp on the basis of scatter plot we can see that as mmax increases so prp increases but the amount of increase is more at higher value of mmax than lower value of mmax .

Thus we can say that the values of prp in relation to mmax increase at an increasing rate .

So justifying this decision on the basis of scatter plot I suggest that there might be a quadratic model which would be better for this case.

The quadratic model requires the square of the independent variable which is mmax .

$y_i = \beta 1 + \beta 2 x_i^2 + e_i$

c. Estimate the model & interpret the result.

Ans)

```
> mod1

Call:
lm(formula = prp ~ I(mmax^2), data = machine)

Coefficients:
(Intercept)    I(mmax^2)
  4.044e+01    2.362e-07


> smod1

Call:
lm(formula = prp ~ I(mmax^2), data = machine)

Residuals:
    Min      1Q  Median      3Q     Max
-371.75  -26.57  -14.22    9.55  325.94


Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.044e+01  5.085e+00   7.952 1.17e-13 ***
I(mmax^2)   2.362e-07  7.516e-09  31.419  < 2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 67.12 on 207 degrees of freedom
Multiple R-squared:  0.8267,    Adjusted R-squared:  0.825
8
F-statistic: 987.2 on 1 and 207 DF,  p-value: < 2.2e-16

> smod1$r.squared
[1] 0.8266575
```

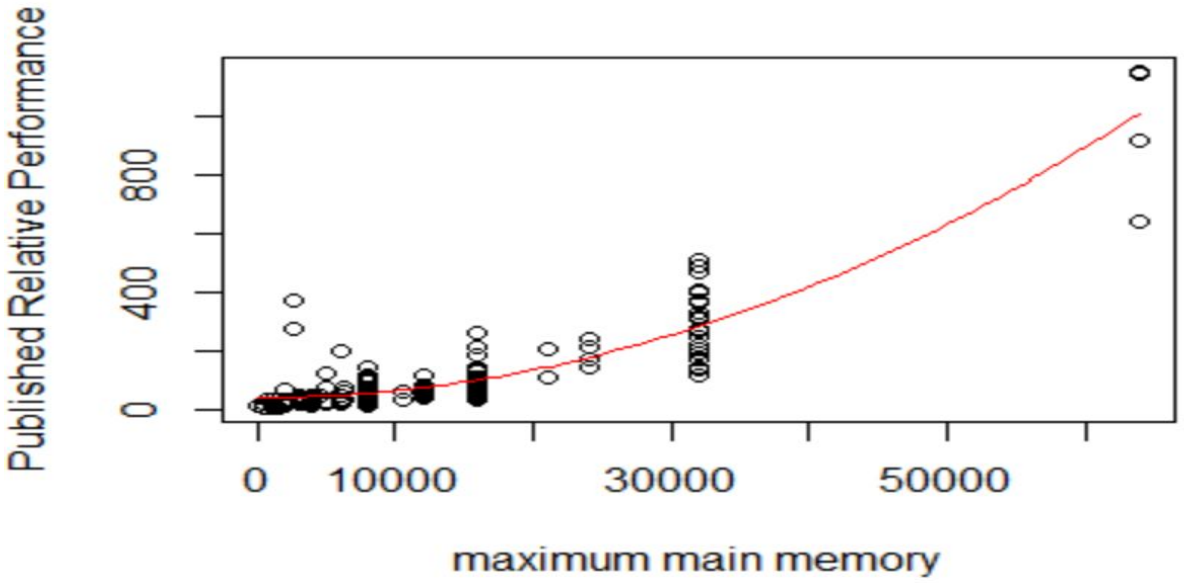
From the results we get from the model we can see that the b2 value is 2.362*10^(-07) and the intercept or b1 value is 40.44 or 4.044*10^01.

We can also see the statistics of residual values where the range of residuals is from -371.5 to 325.94.

We can also observe the p-value of the coefficients . The p-value is very low . The level of significance is denoted by the stars alongside the p-values where the three given stars denote that the level of significance is 0.001. So we can say that the probability of committing type I error in our model is 0.001. Also we can see that the R-squared value is 0.8266575 as we see that this value is quite closer to one , which shows smaller differences between the observed data and the fitted values. Usually, the larger the R-squared , the better the regression model fits your observations. However, this guideline has important caveats . We cannot use R-squared to determine whether the coefficient estimates and predictions are biased, which is why we must assess the residual plots.

d. Fit the regression line on the scatter plot & comment on the same.

Ans)



From the regression curve we can see that the curve provides a good fitting for the lower values of mmax but cannot fit properly at various values of mmax where for the same value of mmax there are various values of prp. Still keeping this in mind that in our dataset as there are multiple values of prp for the same value of mmax the curve has a significantly good fitting which we can also see from the R-squared value . However, the regression

curve consistently under and over-predicts the data along the graph, which is the bias . The Residual vs explanatory variable curve emphasizes this unwanted pattern. An unbiased model has residuals that are randomly scattered around zero. Non-random residual patterns indicate a bad fit despite a high R2.

e. Make a prediction from your model based on the median value of the explanatory variable.

Ans)

As pointed out earlier the median value of mmax is 8000. So we have to predict the value of prp for the mmax value of 8000 based on our model.

```
> medianValue <- median(machine$mmax)
> explanatoryVal <-c(medianValue)
> PredPrpVal <- b1+b2*explanatoryVal^2
> PredPrpVal
[1] 55.55519
```

```
> newx <-data.frame(mmax=c(medianValue))
> yhat <- predict(mod1,newx)
> yhat
       1
55.55519
```

So we apply two methods to find the predicted value and get the result of 55.55519.

f. Run a test of significance on your selected model & discuss the result.

Ans)

Table: Regression output for the 'CPU' model

	Estimate	Std..Error	t.value	Pr...t..
(Intercept)	40.4409355	5.085359	7.952425	0
I(mmax^2)	0.0000002	0.000000	31.419226	0

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.044e+01  5.085e+00   7.952 1.17e-13 ***
I(mmax^2)   2.362e-07  7.516e-09  31.419  < 2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the Test of Significance results we can see that the

p-value for the b2 coefficient is <2*10^(-16) which is less than the corresponding level of significance which is 0.001 so we can reject the null hypothesis(that b2 = 0). Similarly the p-value of b1 coefficient is 1.17*10^(-13) which is less than the corresponding level of significance and thus in this case also we can reject the null hypothesis(that b1=0). After rejecting the NULL hypothesis we can also say that we are rejecting the hypothesis that there is no significant relationship between mmax and prp.

g. Calculate SSR, SSE & SST from your estimated model and comment on the model based upon your calculations.

Ans)

Table: Output generated by the `anova` function

	Df	Sum.Sq	Mean.Sq	F.value	Pr..F.
I(mmax^2)	1	4447613.5	4447613.498	987.1678	0
Residuals	207	932623.6	4505.428	NA	NA

```
> sse <- anov[2,2]
> ssr <- anov[1,2]
> sst <- sse + ssr
> sst
[1] 5380237
> sse
[1] 932623.6
> ssr
[1] 4447613
```

The value of SSR is 4447613.5 which is also known as the sum of squares due to regression .That part of total variation in y, about the sample mean, that is explained by, or due to, the regression. Also known as the "explained sum of squares."

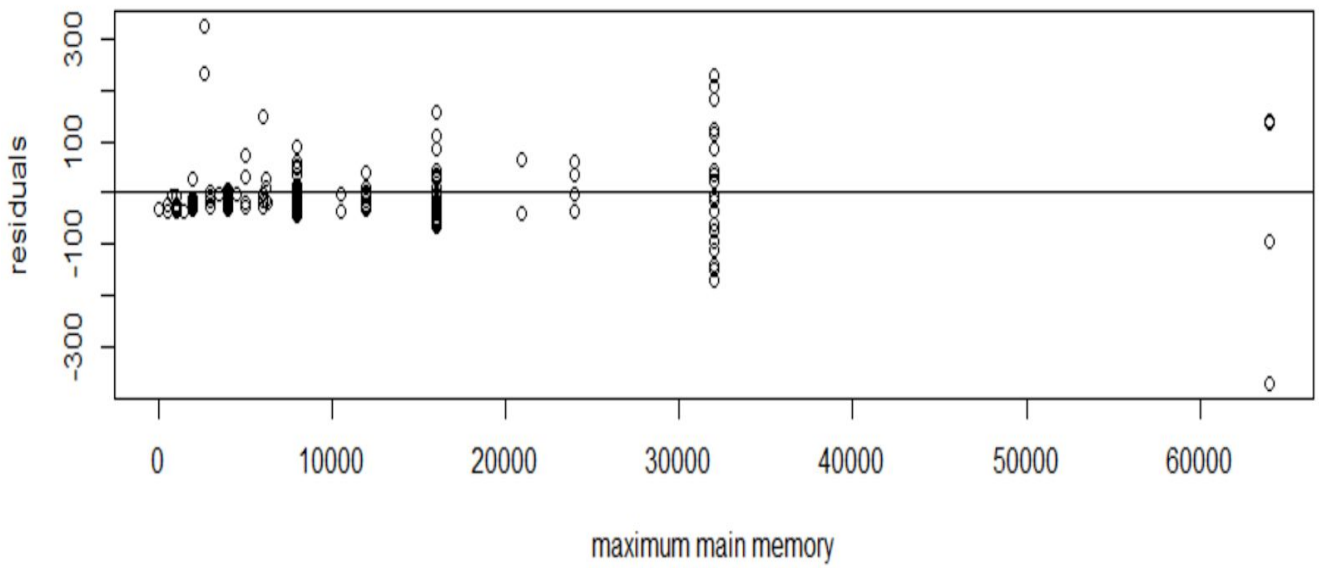
The value of SSE is 932623.6 which is quite small in comparison to the SSR which is the main reason for such a high value of R-squared.SSE is that part of total variation in y about its mean that is not explained by the regression. Also known as the unexplained sum of squares, the residual sum of squares, or the sum of

squared errors.

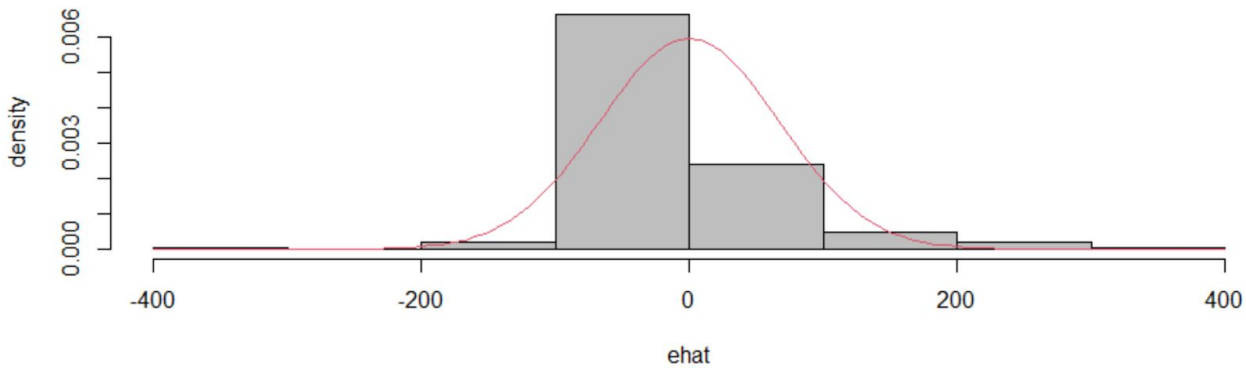
And finally the value of SST or the total sum of squares is 5380237.

h. On your estimated model, check whether the assumptions of homoscedasticity, no autocorrelation & normality of error term is satisfied. Discuss your findings. If the assumptions are not satisfied, suggest some modifications to the model so that the assumptions may get satisfied.

Ans)



As observed from the scatter plot we can see that there is a bowtie residual pattern which is consistent with the variance of error term decreasing and then increasing as x-values increase. So we say that as this residual pattern is associated with Heteroskedasticity the assumption of homoscedasticity is violated.




```
> jarque.bera.test(ehat)

        Jarque Bera Test

data:  ehat
X-squared = 649.24, df = 2, p-value < 2.2e-16
```

From the Jarque Bera Test we can see that as the X-squared value is quite larger than the critical values and thus we can reject the null hypothesis that residual errors are normally distributed.

From the histogram also we can see that there is no normality of ehat .

So we can see that our model has violated the assumptions of homoscedasticity and also violates the assumption of normality of error term. Also the assumption of auto correlation is not violated.

To satisfy the assumptions of homoscedasticity and normality of error term we need to make changes to our model by considering an alternative functional form or transforming the dependent variable. We can choose a better functional form for our model based on the quadratic form by taking into account the Cubic and other higher order polynomial forms.