BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI HYDERABAD CAMPUS

First Semester 2020-21

Econ F241-Econometric Methods

Assignment - II

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. Modify your simple linear regression model (of Assignment-I) by including at least 2- 3 additional independent variables. Also, give justification for the selection of the additional variables.

Ans)

$y = \beta_1 + \beta_2 x_2 + ... + \beta_7 x_7 + e$

The model assumptions remain the same, with the additional requirement that no independent variable is a linear combination of the others.

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

Extending the chosen X or the Independent , Explanatory variable chosen mmax which is the maximum main memory in kilobytes are mmin which is the minimum main memory in kilobytes , cach which is the cache memory in kilobytes , chmin which are the minimum

channels in units , chmax which are maximum channels in unit , myct which is machine cycle time in nanoseconds. So in total there are now 6 independent , explanatory variables.

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(mmax) the Maximum Main Memory which is an 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.

Similarly the independent variable(mmin) the Minimum Main Memory which is an 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.

The independent variable Cache tries to justify the CPU performance as Cache is a small amount of high-speed random access memory (RAM) built directly within the CPU. It is used to temporarily hold data and instructions that the processor is likely to reuse. The bigger its cache, the less time a processor has to wait for instructions to be fetched.Hence better is the performance of the CPU.

The independent variable (chmax) Maximum channels in a unit tries to justify CPU performance as Channel capacity is the tightest upper bound on the rate of information that can be reliably transmitted over a

communications channel. By the noisy-channel coding theorem, the channel capacity of a given channel is the limiting information rate (in units of information per unit time) that can be achieved with arbitrarily small error probability. So more the maximum channels in a unit better the CPU performance.

Similarly the independent variable (chmin) Minimum channels in a unit tries to justify CPU performance as Channel capacity is the tightest upper bound on the rate of information that can be reliably transmitted over a communications channel. By the noisy-channel coding theorem, the channel capacity of a given channel is the limiting information rate (in units of information per unit time) that can be achieved with arbitrarily small error probability. So more the minimum channels in a unit better the CPU performance.
The independent variable (myct) Machine cycle time in nanoseconds tries to justify CPU performance as

CPU Execution Time = CPU clock cycles * Machine cycle time . So less is the machine cycle time , less will be the CPU execution time and better will be the CPU performance.

2. On the selected model, attempt the following (Please note that all the results should be reported in a proper tabular format in R, wherever possible):-

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

Ans)

```
> summary(machine,align="c")
   vendor            model                myct            mmin           mmax           cach
 Length:209        Length:209         Min.   :  17.0   Min.   :   64   Min.   :   64   Min.   :  0.00
 Class :character  Class :character   1st Qu.:  50.0   1st Qu.:  768   1st Qu.: 4000   1st Qu.:  0.00
 Mode  :character  Mode  :character   Median : 110.0   Median : 2000   Median : 8000   Median :  8.00
                                      Mean   : 203.8   Mean   : 2868   Mean   :11796   Mean   : 25.21
                                      3rd Qu.: 225.0   3rd Qu.: 4000   3rd Qu.:16000   3rd Qu.: 32.00
                                      Max.   :1500.0   Max.   :32000   Max.   :64000   Max.   :256.00
    chmin           chmax            prp              erp
 Min.   : 0.000   Min.   :  0.00   Min.   :   6.0   Min.   :  15.00
 1st Qu.: 1.000   1st Qu.:  5.00   1st Qu.:  27.0   1st Qu.:  28.00
 Median : 2.000   Median :  8.00   Median :  50.0   Median :  45.00
 Mean   : 4.699   Mean   : 18.27   Mean   : 105.6   Mean   :  99.33
 3rd Qu.: 6.000   3rd Qu.: 24.00   3rd Qu.: 113.0   3rd Qu.: 101.00
 Max.   :52.000   Max.   :176.00   Max.   :1150.0   Max.   :1238.00
```

The Tidy function works only for double values and we had attributes like model and Vendor which had character values so had to use the summary function.

From this we can comment about the different statistical values of the explanatory variables as well as the target variable .

In myct we can see that the mean is quite higher than the minimum value and the median value denoting a small number of points which have very large numeric values.

Similarly in mmin and mmax we can see that the mean and maximum values differ quite a lot which shows a

fewer number of points with large numeric values.
In cach , chmin and chmax we can see that the minimum value is zero so we can see that there would be CPUs which don't require cache, channels meaning they are not totally necessary for making a proper functioning CPUs and add as an addon to improve the performance. The history of computer evolution totally agrees with this fact as we can see that these were not essentially part of the CPU when the first Von neumann architecture was proposed but were instead later designed when fast , low memory chips which later got the name cache and the information theory was published which identified the need for these components for increasing the performance and functioning of CPUs.

b. Estimate the model & interpret the result:-

Ans)

Table: The multiple regression model

	coefficient	Std. Error	tvalue	p-value
(Intercept)	-55.89393361	8.0450055701	-6.9476563	4.998100e-11
myct	0.04885490	0.0175190946	2.7886658	5.798460e-03
mmin	0.01529257	0.0018268120	8.3711798	9.000000e-15
mmax	0.00557139	0.0006418073	8.6807829	1.000000e-15
cach	0.64140143	0.1395617540	4.5958252	7.585893e-06
chmin	-0.27035755	0.8556704410	-0.3159599	7.523592e-01
chmax	1.48247217	0.2200382480	6.7373385	1.646290e-10

From this model we can see that as the p-values of all the explanatory variables is quite lower than the level of significance hence any hypothesis regarding need to include that explanatory variable or not would be rejected and hence we can say that all the variables that are included have significance and are required for the proper multi linear regression model for the target variable PRP(published relative performance).

Also we can see that chmin has a negative coefficient and hence a higher value of minimum channels in units means a lower performance as the difference between the maximum and minimum channels in units helps in proper

signal noise reduction(Information theory).

The lesser the minimum channels the better as the signal values information passing is not bottlenecked.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.589e+01  8.045e+00  -6.948 5.00e-11 ***
myct         4.885e-02  1.752e-02   2.789   0.0058 **
mmin         1.529e-02  1.827e-03   8.371 9.42e-15 ***
mmax         5.571e-03  6.418e-04   8.681 1.32e-15 ***
cach         6.414e-01  1.396e-01   4.596 7.59e-06 ***
chmin       -2.704e-01  8.557e-01  -0.316   0.7524
chmax        1.482e+00  2.200e-01   6.737 1.65e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

c. Calculate all the model selection criteria for the model. Comment on the goodness-of-fit of the model.:-

Ans)

	c.Rsq..AdjRsq..aic..bic.
R-squared	0.8648907
Adjusted R-squared	0.8608775
AIC	2313.3519686
BIC	2340.0906426

The R-squared and adjusted R-squared values of the

model are pretty high denoting that the model fits the data well.

But there are some other points to keep in mind .

Like if cross-sectional data are being used to estimate a causal effect, then low R^2 's are typical and not necessarily a concern.

What is more important is to avoid omitted variable bias and to have a sample size sufficiently large to get a reliable estimate of the coefficient of interest .

The second problem is one related to predictive models, namely, that comparing models on the basis of R^2 is only legitimate if the models have the same number of explanatory variables.

Adding more variables always increases R^2 even if the variables added have no justification. As variables are added the sum of squared errors SSE goes down and thus R^2 goes up. If the model contains N - 1 variables, then $R^2=1$.

Among several models, the best fit is the one that

maximizes R^2 or adjusted R^2.

On the contrary, the best model must minimize AIC or BIC.

d. Use the RESET test to justify your selected model. If the model is not adequate, then try to adjust your model to build an appropriate model. Please discuss each step involved in the process. No marks will be given without proper interpretation and discussion of the steps:-

Ans)

```
> resettest(mod1, power=2, type="fitted")

        RESET test

data:  mod1
RESET = 173.2, df1 = 1, df2 = 201, p-value < 2.2e-16

> resettest(mod1, power=2:3, type="fitted")

        RESET test

data:  mod1
RESET = 93.488, df1 = 2, df2 = 200, p-value < 2.2e-16
```

In both cases, p-values are quite lower than 0.05, indicating that the model marginally fails the specification test and some higher order terms may be necessary .

This indicates that our model is inadequate.

So we try to improve our model by choosing another functional form.

Tasking suggestions of when we previously modified our simple linear regression model in Assignment 1 to include quadratic terms of mmax(maximum main memory) as we realised that the effect of mmax on the performance is quadratic and thus the value increases in a polynomial way.

In this model also we realise that we should take quadratic forms of mmax and mmin as both indicate similar variables of main memory in kilobytes and are just the maximum and minimum limit values of the same.

So our new adjusted model is

$$\mathbf{prp(Y)=\beta_1+\beta_2(mmax)^2+\beta_3(mmin)^2+\beta_4(cach)+\beta_5(chmax)+\beta_6(chmin)+\beta_7(myct)\ .}$$

The Reset Test value for the adjusted model is:-

```
> mod2 <- lm(prp~I(mmax^2)+I(mmin^2)+cach+myct+chmax+chmin,data = machine)
> resettest(mod2, power=2:3, type="fitted")

        RESET test

data:  mod2
RESET = 2.9444, df1 = 2, df2 = 200, p-value = 0.05492

> resettest(mod2, power=2, type="fitted")

        RESET test

data:  mod2
RESET = 0.38083, df1 = 1, df2 = 201, p-value = 0.5379
```

As the p-value now is greater than the level of significance 0.05 thus we cannot reject the null hypothesis of the reset test and thus our model is adequate.
So From now on all the results and evaluations will be on our adjusted model.

e. Test at least one joint hypothesis with linear combinations of regression coefficients from your model and comment on the results.:-

Ans)

$prp(Y)=\beta_1+\beta_2(mmax)^2+\beta_3(mmin)^2+\beta_4(cach)+\beta_5(chmax)$

+ β_6(chmin) + β_7(myct)

So let our null hypothesis be :-

That when cach=0,chmin=0,chmax=0 and myct=800,mmax=2000kB, mmin=800kB then prp=20.

And as this is happening the condition 2**β_2(mmax) + 2β_3(mmin) = 250.(lets say)**

Table: Joint hypotheses with the 'linearHypothesis' function

res.df	rss	df	sumsq	statistic	p.value
204	2.549336e+16	NA	NA	NA	NA
202	4.301903e+05	2	2.549336e+16	5.985326e+12	0

```
Linear hypothesis test

Hypothesis:
4000 I(mmax^2)  + 1600 I(mmin^2) = 250
(Intercept)  + 4e  + 06 I(mmax^2)  + 640000 I(mmin^2)  + 800 myct = 20

Model 1: restricted model
Model 2: prp ~ I(mmax^2) + I(mmin^2) + cach + myct + chmax + chmin

  Res.Df        RSS Df  Sum of Sq          F    Pr(>F)
1    204 2.5493e+16
2    202 4.3019e+05  2 2.5493e+16 5.9853e+12 < 2.2e-16 ***
```

As we can see that the p-value is so much smaller than the level of significance so we can say that we can reject the

null hypothesis.

f. Test for the presence of collinearity and heteroskedasticity in your model using plots and formal tests and modify the model to control for them. Discuss the steps involved in the process with proper interpretation of the results:-

Ans)

A test that may be useful in detecting collinearity is to calculate the variance inflation factor, VIF, for each regressor. The rule of thumb is that a regressor produces collinearity if its VIF is greater than 10.

Table: Variance inflation factors for the 'machine' regression model

regressor	VIF
I(mmax^2)	2.676208
I(mmin^2)	2.103343
cach	1.744396
myct	1.148116
chmax	2.112609
chmin	1.948235

As we can see that none of the explanatory variables has VIF value greater than 10 so we can say that there is no collinearity.

Now we test for Heteroskedasticity .

First we use the Breusch-Pagan heteroskedasticity test.

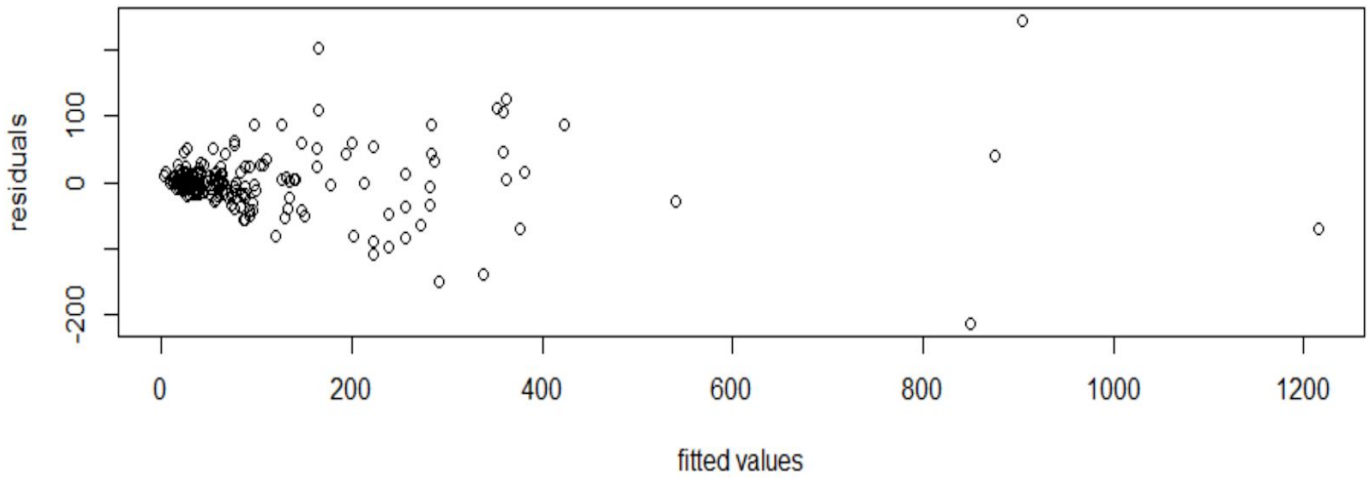
Table: Breusch-Pagan heteroskedasticity test

statistic	p.value	parameter	method
93.57778	5e-18	6	studentized Breusch-Pagan test

Since $\chi^2 > \chi^2_{cr}$, we reject the null hypothesis of homoscedasticity, which means there is heteroskedasticity in our data and model.

Alternatively, we can find the p-value corresponding to the calculated χ and say that as it is lower than the significance level we can reject the null hypothesis .

So our model has heteroskedasticity .



This is the scatter plot between residuals and fitted values. From this also we can see that there is heteroskedasticity in our model as we can see a clear fan or cone shaped pattern where the spread of residuals increases with the increase in fitted values.

If we're willing to accept the fact that ordinary least squares no longer produces the best linear unbiased estimators (BLUE), we can still perform our regression analysis to correct the issue of incorrect standard errors so that our interval estimates and hypothesis tests are valid. We do this by using **heteroskedasticity-consistent standard errors** or simply **robust standard errors.**

Table: Robust (HC1) standard errors in the 'mod2/Machine' equation

term	estimate	std.error	statistic	p.value
(Intercept)	19.8167122	5.7326589	3.4568100	0.0006663
I(mmax^2)	0.0000002	0.0000000	5.7589337	0.0000000
I(mmin^2)	0.0000004	0.0000001	3.2023437	0.0015837
cach	0.9441568	0.2128023	4.4367790	0.0000150
myct	-0.0111583	0.0071225	-1.5666184	0.1187691
chmax	0.4698422	0.3849667	1.2204749	0.2237077
chmin	0.7406909	1.0318778	0.7178088	0.4737045

g. For your estimated model, check whether the assumptions of no autocorrelation & normality of error term are satisfied by using scatter plots. Discuss your findings.:-

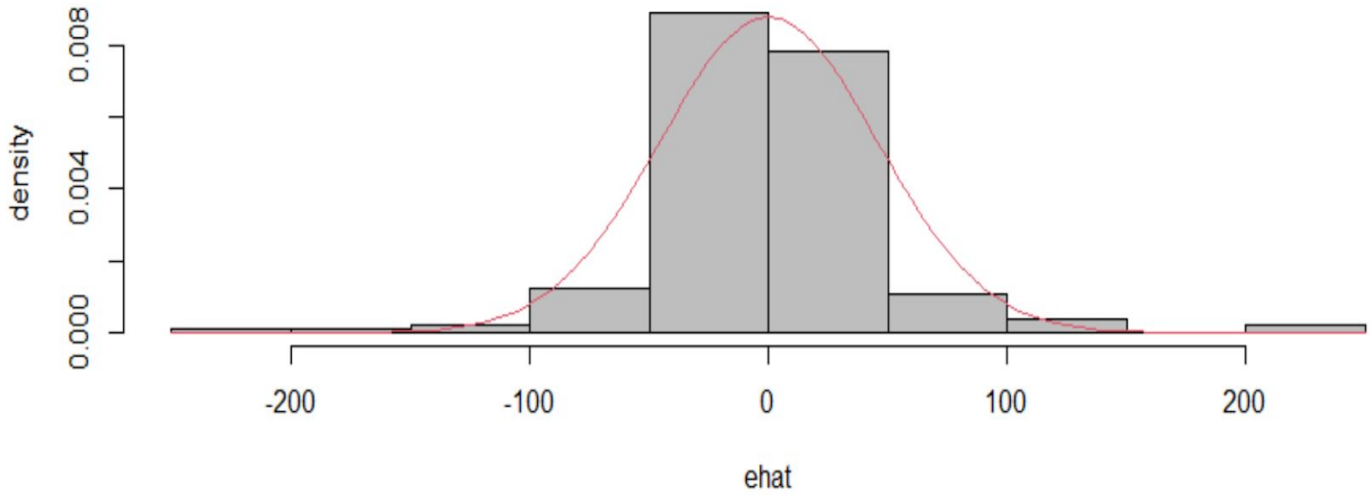
Ans)

For Normality of error term we can perform the Jarque-Bera test, for which the null hypothesis is “Series is normally distributed”.

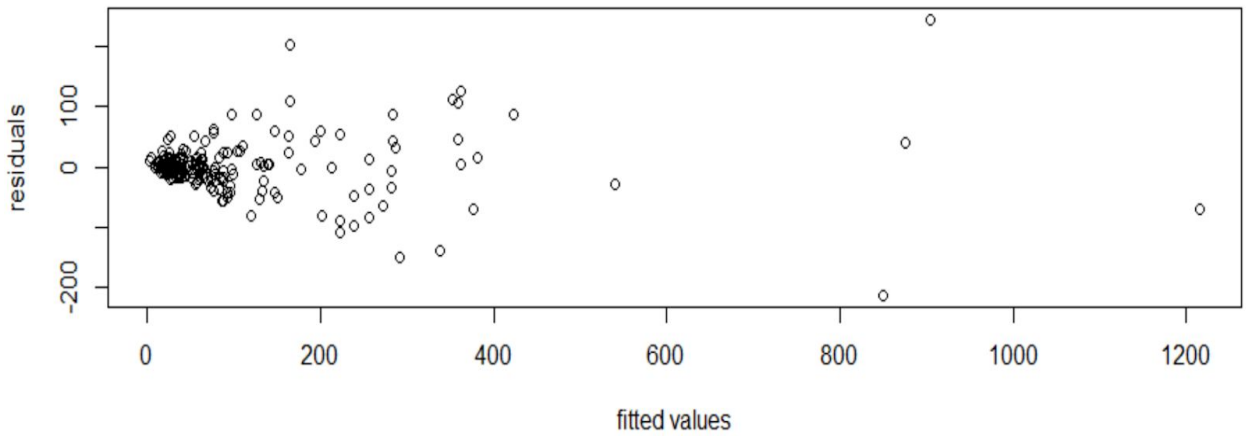
```
> jarque.bera.test(ehat)

        Jarque Bera Test

data:  ehat
X-squared = 591.67, df = 2, p-value < 2.2e-16
```

As the p-value is way smaller than the level of significance we can reject the null hypothesis and thus our model fails the Normality assumption.

Also the test statistic value is higher than the critical value.



From the histogram and superimposed normal distribution also we can say that the normality assumption fails.



Also from looking at the residual scatter plots we can see that there is no autocorrelation in this cross sectional dataset.

So the assumption of no autocorrelation holds.

h. Lastly, make predictions based on your model and comment on the result:-

Ans)

Let's make a prediction with the following values of explanatory variables:-

myct=203 , mmin=2868 , mmax = 11800 , cach = 25 , chmin = 5 , chmax = 20 .

(All values are according to their relevant scalings and

hence are already scaled to be used as it is)

```
> temp <- predict(mod2, newdata=predpoint,interval="confidence")
```

```
> kable(temp,caption="Prediction for the estimated model")
```

Table: Prediction for the estimated model

fit	lwr	upr
79.33756	72.68759	85.98753

As we can see the fitted value of prp (or published relative performance) is 79.33756 at a default confidence level of 95%.

The confidence interval reflects the uncertainty around the mean predictions.

The 95% confidence interval with the given explanatory variables is (72.68759,85.98753).

There is a difference between prediction and forecasting. Prediction meaning the estimation of an expected value of the response and forecasting meaning an estimate of a particular value of the response.

As we are doing prediction that is why we pass interval=”confidence” as this gives us a prediction with its confidence interval.

While forecasting we pass interval=”prediction” as this gives us a forecast with its confidence interval.

**********THE END**********