204: Regression Analysis

Instructor: Prof. Sugata Sen Roy

Report Card
Data Analysis with Regression

Auto-Correlation

Submitted by: Soumarya Basak

Submitted on: June 4, 2022

University of Calcutta

Department of Statistics

Contents

Problem 1

1.1 About the Practical

We have a data on *output*(*O*), *labour input*(*L*) and *capital*(*K*) of a firm over 10 years in monetary units. Need to fit a suitable linear model on the data and try to check whether the assumptions of linear model is valid or not.If any assumption of linear model is disobeyed then we need to obtain a better estimator of the model parameters.

1.2 Analysis

Note that, output always depends on labour input and capital input so *output*(*O*) is the response variable and *labour input*(*L*) and *capital*(*K*) are the regressors.

In practice input or output are hardly linearly related so before going in depth of analysis we will see how the variables are dependent.

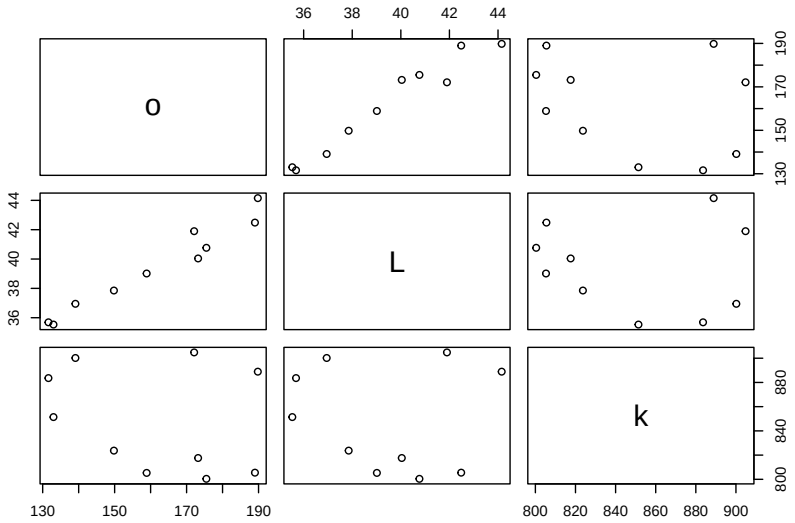


Figure 1.1: Scatter Plot of the Variables

It is clear that the output and the capital input variables are not linearly related so it will be better if we proceed with *Cob-Doglas* model. i.e.

$$O_t = \alpha L_t^{\beta_1} k_t^{\beta_2} u_t$$

$$logO_t = \alpha^* + \beta_1 logL_t + \beta_2 logK_t + v_t$$

Let's Fit the model.
All the analysis required for the date is performed on suitable software and the result is

discussed below.
Output:

Parameters	***Estimate***	***Std. Error***	$t-value$	$Pr(>\|t\|)$
Intercept	2.22246	0.72109	3.082	0.01776
$log(L)$	1.77558	0.06505	27.298	2.27e-08
$log(k)$	-0.54416	0.09704	-5.607	0.00081

Residual standard error: 0.01447 on 7 degrees of freedom
Multiple R-squared: 0.9915
Adjusted R-squared: 0.9891
F-statistic: 410.5 on 2 and 7 DF
p-value: 5.557e-08

Conclusion 1 :

The above model has p value less than 0.05 for the F test, so the model is significant. Also note that the parameters are individually significant at 5% level. So the model is good enough if the data satisfies the assumptions of linear models.

1.2.1 Checking for Constant Error-Variance

As the data is too small it is very much difficult to study the nature (variability) of residuals for the model. As par the residual plot we can consider it a homoscedastic data set.

1.2.2 Checking for Auto correlation

To check the assumption that the errors are uncorrelated we use *Durbin Watson* test on the residuals.
On performing Durbin Watson test we come up with **p value of 0.002** with Durbin Watson statistic d equals to 0.8869685.

Conclusion 2 :

At 95% confidence level we have to accept that the data has autocorrelation. So the OLS estimators will not perform well and we have to workout with this problem of autocorrelation on data to obtained a better GLS estimators of parameters.

1.2.3 Obtaining the GLS estimators

To obtain GLS estimators for the parameters, when the data has auto-correlation, we use the **Cochran-Orcutt Method**. And the estimators of the parameters comes out to be,

Intercept:	1.894243
$log(L)$**:**	1.759270
$log(k)$**:**	-0.136710

1.2.4 The Final Model

So, the final model which will work best for the data is given by,

$$\textbf{[log}(O_t)\textbf{ =1.894243 + (1.759270 } \times \textbf{ log}(L_t)\textbf{)-(0.136710 } \times \textbf{ log}(K_t)\textbf{)}$$

1.3 Summary of Conclusions

1. the variable of the data was not linearly dependent so instead of taking a simple linear model we considered the *Cobb Douglas* model here.
2. Based on the Cobb Douglas model we fit linear regression model and the OLS estimator of parameters comes out to be significant with the intercept term.
3. These OLS estimators will perform well if our data satisfy the assumptions of linear model. But the data exhibits autocorrelation among the residuals. So OLS estimators are not the best here.
4. To workout the problem of Auto correlation we find the GLS estimators of the parameters by *Cochran-Orcutt method*, which are the best for the model.

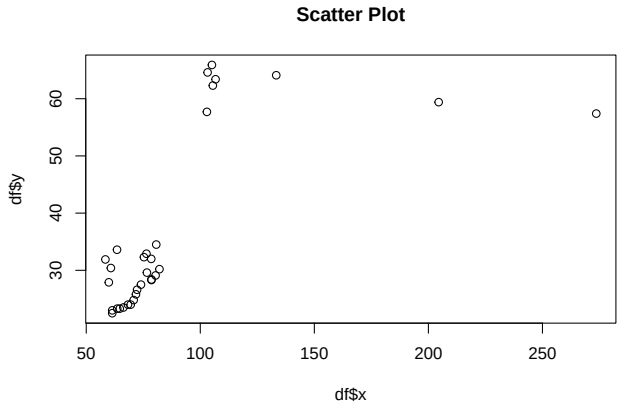
Problem 2

2.1 About the Practical

We have a data on the 'price of oil' and the 'price of bituminous coal'. We want to find whether the price of oil influences the price of coal and fit a suitable model to the data.

2.2 Analysis

For our problem the 'price of oil (x)' is the regressor end the 'price of bituminous coal (y)' is the response variable.
From the scatter plot of the variables it can be easily notice that there is linear relationship between the variables so we doesn't further necessary to take any transformations.
We will feed a simple linear regression model to the data and the output as shown below,



Output:

Parameters	***Estimate***	***Std. Error***	$t-value$	$Pr(>\|t\|)$
Intercept	15.97256	4.70852	3.392	0.00196
X	0.23088	0.04783	4.827	3.8e-05

Residual standard error: 11.72 on 30 degrees of freedom
Multiple R-squared: 0.4372
Adjusted R-squared: 0.4184
F-statistic: 23.3 on 1 and 30 DF
p-value: 3.795e-05

Observation

The above model fits well the data as the parameters statistically significant and the model explains 41% variability off response variable.

Conclusion 1

The above model fits well the data if the data doesn't consist any influential observation and obeys the assumptions of a linear model.

2.2.1 Checking for Influential Observation

From your analysis we find that the data has 4 influential measures those are 26th, 29th, 31st, 32nd observations. As we don't have the knowledge about the genesis of the data, we can't cross verify those.
Since influential observation me affect the model very much we'll fit another model by deleting the 4 influential observations.
The model after deleting the 4 observations are,

Output:

Parameters	***Estimate***	***Std. Error***	$t-value$	$Pr(>\|t\|)$
Intercept	-17.86915	5.16723	-3.458	0.00189
X	0.66155	0.06598	10.027	2e-10

Residual standard error: 5.84 on 26 degrees of freedom
Multiple R-squared: 0.7945
Adjusted R-squared: 0.7866
F-statistic: 100.5 on 1 and 26 DF
p-value: 2.005e-05

Observation

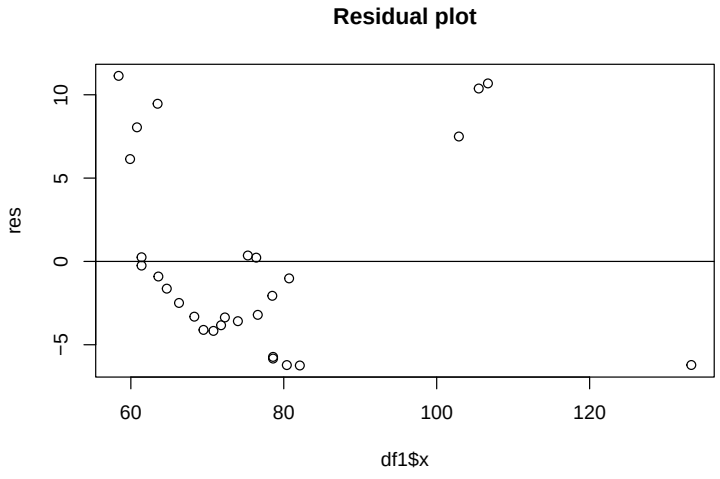
If we compare the new model with the previous one we can easily see that the residual standard error decreases and also the Adjusted R^2 increases to a value for the new model with significant parameters. Deity's the new model explain 78.6% of the variability of response variable which is quite high than the previous model. Which indicates that we come up with a better model after removing the 4 observations.

2.2.2 Checking for Heteroscedasticity

Forever further analysis we will use the new model which we have fitted after deleting the 4 observations.

Residual Plot

From the residual plot with respect of X variable it is very difficult to comment anything about the variability of the residuals. So to check for heteroscedasticity of the data set we have to perform a statistical test.



Goldfeld Quandt Test

Here we perform **Goldfeld Quandt test** against the alternative of greater than type as the plot indicates a increase in variable as X increase. On performing Goldfeld Quandt test on the residuals, we came up with a **p value of 0.06**.

Conclusion 2

At 95% significant level the data is homoscedastic, so the assumption of constant error-variance holds here.

2.2.3 Checking for Auto-correlation

Now to check whether the assumption of uncorrelated errors obeyed — we perform ***Durbin Watson test*** on the residuals, as the residuals can be treated as an estimator of errors.
On performing Durbin-Watson test we come up with a **p value very near to 0**.

Conclusion 3

So we missed that at 95% confidence level the data has autocorrelation. So the OLS estimators will not be the best perhaps we will further looks for GLS estimators.

2.2.4 Obtaining the GLS Estimators

Since the data has autocorrelation the GLS estimators will be the best for the model. We will prefer **Cochran-Orcutt** method to find the GLS estimators of the parameters which are given in the following table after performing the algorithm.

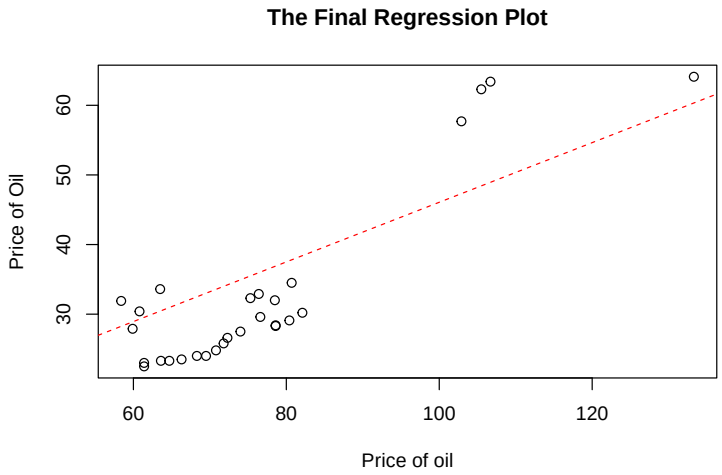
Intercept:	3.235024
X:	0.428367

2.2.5 The Final Model

So, we would prefer the following model,

$$y = 3.235024 + 0.428367 \times x$$

where $y =$ Price of Bituminous coal, and $x =$ Price of oil



2.3 Summary of Conclusions

1. we want to know weather the price of bituminous cold is influenced by the price of oil and to fit a suitable model to the data based on these variables.
2. We fit a simple linear regression and the model comes out to be significant with all parameters. The model would be good if the data satisfies the assumptions of linear model.
3. The very first work we have done is to find the influential observation and it comes out that the 26th, 29th, 31st, 32nd observations are influential observations for the model. We remove these influential observations and repeat the model and get a better model with higher Adjusted R^2.
4. Then to check the presence of heteroscedasticity we perform Goldfeld Quandt test on the residuals obtained from the later model and at 95% level accept the fact that the data is homoscedastic.
5. Then we perform Durbin Watson test to check whether there is autocorrelation or not. The p value near 0 indicates the data has autocorrelation. Therefore we need to obtain the GLS estimators of the parameters.
6. buy Cochran Orcutt method we obtain the GLS estimators of the parameter and obtain the best model.

Appendix

The overall analysis is performed in R software, and the necessary output is discussed here. The data set and the codes can be found in the following Github repository.

Source code : https://github.com/SoumaryaBasak/Regression_Analysis_1.git