204: Regression Analysis

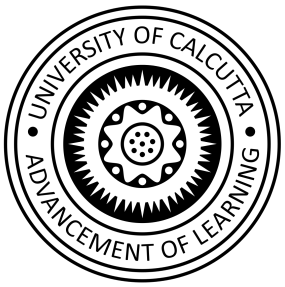
Instructor: Prof. Sugata Sen Roy

Report Card
Data Analysis with Regression

Working with Heteroscedastisity

Submitted by: Soumarya Basak

Submitted on: June 22, 2022



University of Calcutta

Department of Statistics

Contents

Problem 1

1.1 About the Practical

We have a data on consumption expenditure and income of 20 families (in 000). The data is very likely to show heteroscedasticity as income-expenditure data may exhibit heteroscedasticity. Here our task is to check the heteroscedasticity for the data and fit suitable regression in this context.

1.2 Analysis

It is quite obvious that *Consumption-expenditure* depends on *Income* of a person, so for this data, "*Income*"(X) is our regressor and "*Consumption-Expenditure*"(y) is our response variable.

Analysis Required :

1. To fit a regression Model
2. Study the residuals
3. Test for heteroscedasticity
4. Fit a suitable regression in the presence of heteroscedasticity.

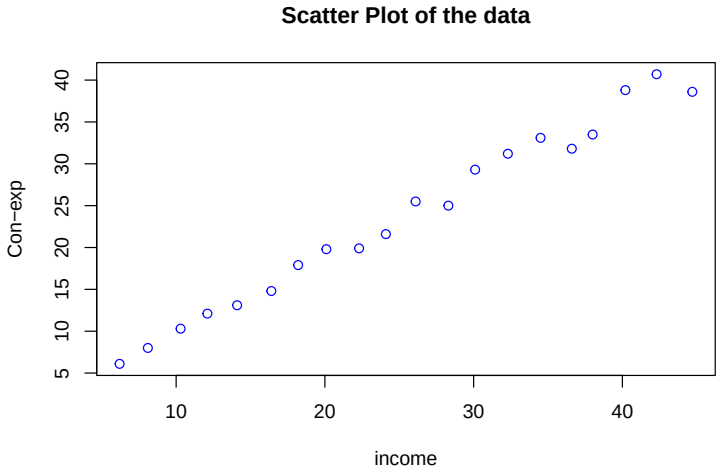
All the analysis required for the date is performed in lab, on suitable software and the result is discussed below.

1.2.1 Fitting Regression Model

Before fitting a model it is always better to have a visualization to the data so we create a scatter plot to observe the pattern in the data.

From the plot it is quite clear that there might be a linear relationship between the variables so now we tried to fit a linear regression model on the data.

Here, we fit a simple linear regression to the data, as



Output:

Parameters	***Estimate***	***Std. Error***	$t-value$	$Pr(>\|t\|)$
Intercept	0.84705	0.70335	1.204	0.244
x_var	0.89932	0.02531	35.534	<2e-16

Residual standard error: 1.314 on 18 degrees of freedom
Multiple R-squared: 0.9859
Adjusted R-squared:0.9852
F-statistic: 1263 on 1 and 18 DF
p-value: < 2.2e-16

Observation :

Here the F-test is significant for the model and has a high R^2 value, but the coefficient for intercept terms is statistically insignificant.

Conclusion 1 :

There are several reason for this result: there may be some influential observations, or there may present heteroscedasticity, there may be auto-correlation etc.
Let work on this consequences one by one.

1.2.2 Detecting Influential Observations

After doing necessary diagnostic we have that, there are 2 observations which may be treated as influential — those are the 6 th and 7th observation.
Since we don't have any knowledge about the experiment, we can't cross check those, so we need to delete those observations from the data and refit the model.

Output:

Parameters	***Estimate***	***Std. Error***	$t-value$	$Pr(>\|t\|)$
Intercept	0.46505	0.75112	0.619	0.545
x_var	0.92057	0.02752	33.452	3.08e-16

Residual standard error: 1.216 on 16 degrees of freedom
Multiple R-squared: 0.9859
Adjusted R-squared:0.985
F-statistic: 1119 on 1 and 16 DF
p-value: 3.082e-16

Observation :

Observe that the above fit is worse than the first fit. As the adjusted R^2 doesn't increases perhaps decreases, intercept is still insignificant and the corresponding p values for F statistic and slop parameter also increases.

Conclusion 2 :

So removing the influential observation doesn't change the model, so it's better to work with the full data set as removing observations causes loss of information.

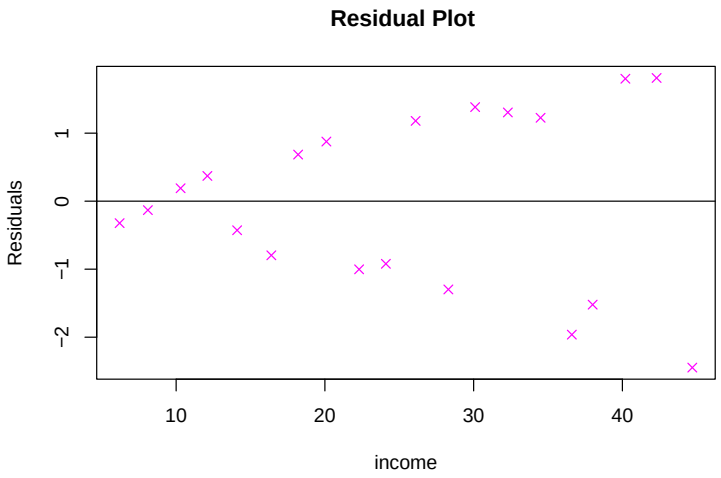
1.2.3 Detecting Heteroscedasticity

As the first reason for a insignificant model fit, we will check for the next reason – presence of heteroscedasticity. As we were sticking with first model, it's residual plot will give us the idea about the variability of the model.

Residual Plot

Observation :

The plot of residuals with respect to the income variable shows that as income increases the variability of the error residuals also increases so the assumption of considering constant variability is the wrong assumption for the model.



Conclusion 3 :

From the residual plot we can suspect the presence of heteroscedasticity. But to be sure about about it we need to perform a statistical test. Here we perform Goldfeld Quandt test.

Goldfeld Quandt Test

Performing Goldfeld Quandt test against the alternative of greater than type we have a **p value of 0.006** which is less than 0.05, so at 5% level of significant we have to reject the null hypothesis that the model is homoscedastic.

Conclusion 4 :

Since there present heteroscedasticity OLS estimator doesn't work here. We need to find the GLS estimators for the coefficients.

1.2.4 Obtaining GLS Estimator

To find GLS estimators of the coefficients we need to find the error variances of the model. Here we are considering a simple model to estimate the error-variances based on the regressor.
We are considering the following model to estimate the error variance:

$$\widehat{\sigma_i^2} = -1.37033 + 0.11580 \times Income$$

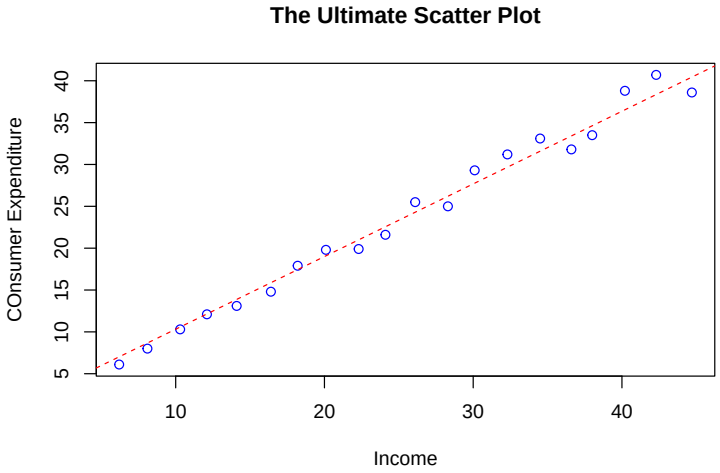
Based on the above equation, we obtain the estimated error-variance-covariance matrix $(\widehat{\Omega})$ and hence we obtain the estimated GLS estimator of the coefficients as:

Intercept : 1.638355
Income : 0.867986

1.2.5 The Final Model

So, the final model, which we can suggest is:

$$\textit{Con\_exp} = 1.638355 + 0.867986 \times \textit{Income}$$



1.3 Summary of Conclusions

1. As we fit the data the intercept comes out to be insignificant. There are several reasons for this — there may be some influential observations, the error-variances may not be equal, there may be some autocorrelation within the error.

2. We find that the 6th and 7th observations are influential observations. But dropping those observations doesn't change the model. So we stick to our original model as removing observation me causes loss of information.

3. Now to look for unequal error variances we make a residual plot with respect to our regressor variable 'income' and the plot shows there may be some heteroscedasticity in the data. Further we go with Goldfeld Quandt test against the alternative of greater than type and the p value comes to be < 0.05, which indicates that the data has heteroscadasticity.

4. In presence of heteroscedasticity OLS estimator doesn't work better so we will have to look for GLS. We estimate the error variance using a simple model based on income and obtain the EGLS estimators for the data.

Problem 2

2.1 About the Practical

We have a data on speed of cars and distance it covers to come to a standstill, we have to fit a suitable regression model and test for heteroscedasticity.

2.2 Analysis

It is quite obvious that *distance covered* will depend on *speed*, so `speed`(X) is the regressor and distance(Y) is the response variable.

Analysis Required :

1. To fit a regression Model
2. Study the residuals
3. Test for heteroscedasticity
4. Fit a suitable regression in the presence of heteroscedasticity.

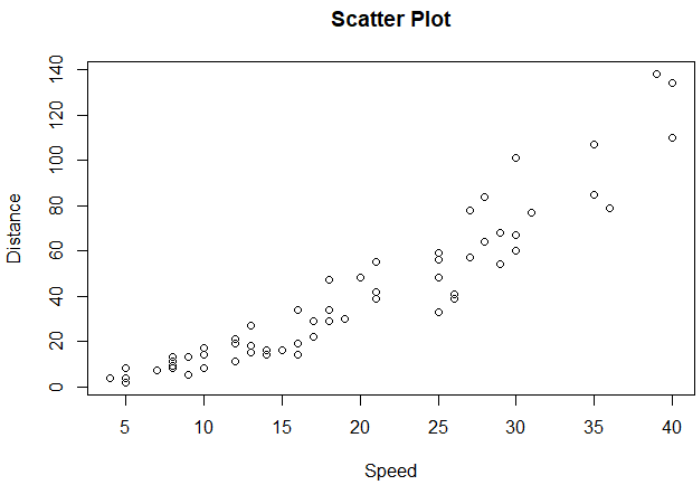
All the analysis required for the date is performed in lab, on suitable software and the result is discussed below.

2.2.1 Fitting Regression Model

Before fitting a model it is always better to have a visualization to the data so we create a scatter plot to observe the pattern in the data.

From the plot it is quite clear that there might be a linear relationship between the variables so now we tried to fit a linear regression model on the data.

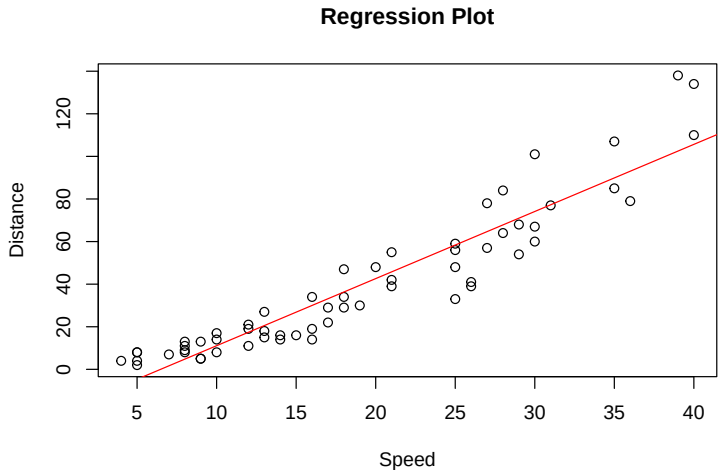
Here, we fit a simple linear regression to the data, as



Output :

Parameters	***Estimate***	***Std. Error***	$t - value$	$Pr(>\|t\|)$
Intercept	-20.4174	3.3446	-6.105	9.16e-08
x_var	3.1515	0.1559	20.213	< 2e-16

Residual standard error: 11.95 on 58 degrees of freedom
Multiple R-squared: 0.8757
Adjusted R-squared: 0.8735
F-statistic: 408.6 on 1 and 58 DF
p-value: < 2.2e-16



Observation :

From the fitting it can be shown that F test is statistically significant and the coefficients are also significant at 5% level, the adjusted R^2 is high we can say the data fit the model fit the data well if the model doesn't avoid the assumptions.

Conclusion 1 :

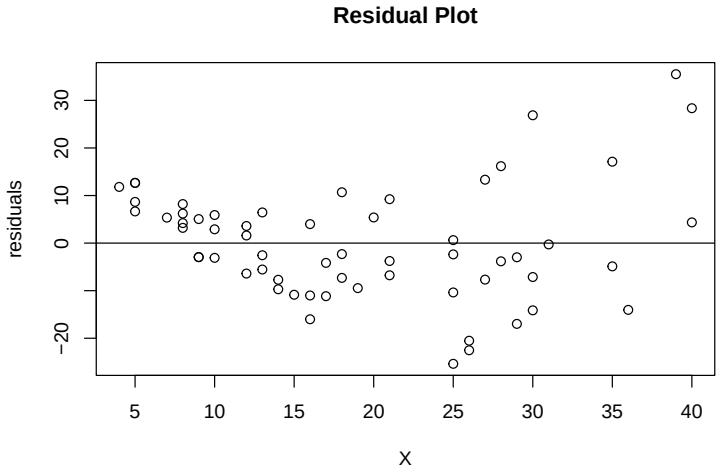
The model will be good enough if we can analyse if the data satisfies the assumptions of a linear model. For a linear model the error-variances must be constant and the errors component must be uncorrelated. If any of these assumptions of linear model is avoided, then the OLS estimator doesn't work better. Then we need to look for the GLS estimators. In the further steps we will try to find out whether the data satisfies the assumptions of the linear model or not.

2.2.2 Detecting Heteroscedasticity

The residual plot of the fit gives us an idea about the variability of the residual. Note that for practical purpose the residuals are treated as an estimator of error component.

Residual Plot

From the residual plot of the model it can be observed that as the x variable increases the variability of the residuals also increases. Space so we can't say that the error variances are constant that is the data has heteroscedasticity.



A Cue for Glejser's Test

The residual plot of regression model gives us a cue of presence of heteroscedasticity. But to check weather there heteroscedasticity present or not we will have to go for a statistical test.

Handling Influential observation

The presence of influential observation makes some high residuals so we check for influential observations in the data and remove them from the data and refit the model but removing this observation doesn't change the model much, add also the residual plot doesn't indicates constant variability after removing observations. So it's better to stick with the full data set as removing observation causes loss of information

Glejser's Test

As we perform Glejser's Test on the data with alternative hypothesis of greater than type, the **p value comes out to be 0.0004** which is less than 0.05 so we have to reject the null hypothesis of homoscedasticity.

Conclusion 2 :

As the data has heteroscedasticity OLS estimator doesn't perform well here. We need to look for GLS estimator of the parameters. Here we will try to estimate the parameters assuming the error variance to be a linear function of speed with an intercept term.

2.2.3 Obtaining GLS Estimator

To find gls estimator of the parameters we assume the error variance to be a linear function of speed with an intercept term.
Based on the data we are considering the following function to estimate the error variance is:

$$\widehat{\sigma_i^2} = -73.672 + 11.123 \times Speed$$

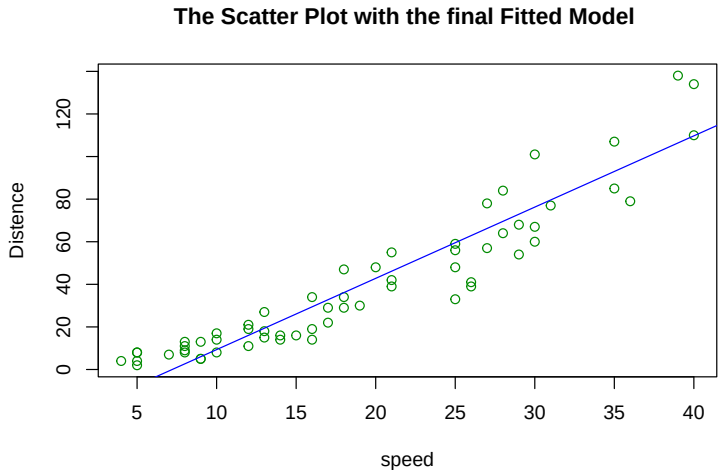
Using the above equation we obtain the estimated error-variance-covariance matrix $(\widehat{\Omega})$ and hence we obtain the estimated GLS estimator of parameters as:

Intercept : -24.118582
Speed : 3.345985

2.2.4 The Final Model

So, the final model, which we can suggest is:

$$\textit{Dis} = -24.118582 + 3.345985 \times \textit{Speed}$$



2.3 Summary of Conclusions

1. We fit a simple linear regression on the data and the F statistic along with the parameters comes out significant so the fitting will be good enough if the data satisfies the assumption of a linear model that is error-variances are constant and the errors are uncorrelated. If any of these assumption does not satisfied by the data OLS estimator doesn't perform well we need to look for GLS estimators.

2. As we plot the residual plot of the fit, we can suspect that there must be unequal residual variance. The presence of influential observations may make some residuals higher. So we check for influential observation and remove them from the data and refit it. But the later model doesn't change much or doesn't resolve heteroscedasticity. So we stick to our first model.

3. To test statistically whether there is heteroscedasticity or not we use Glejser test and value comes out to be 0.0004 which is less than 0.05 so we need to say the data has heteroscedasticity.

4. In presence of heteroscedasticity OLS estimator doesn't work better so we will have to look for GLS. We estimate the error-variance using a linear function of speed with an intercept term an based on these estimated error-variance we get the GLS estimators.

Problem 3

3.1 About the Practical

We have a data consist of the rate of change of stock price and consumer price in 20 countries. We have to fit a suitable regression model and further we need to check whether there heteroscedasticity is present or not. If yes then to get suitable estimators of parameters.

3.2 Analysis

Here, `the rate of change of stock price'(Y) is the response variable and the `Consumer price'(X) is the regressor.

Analysis Required :

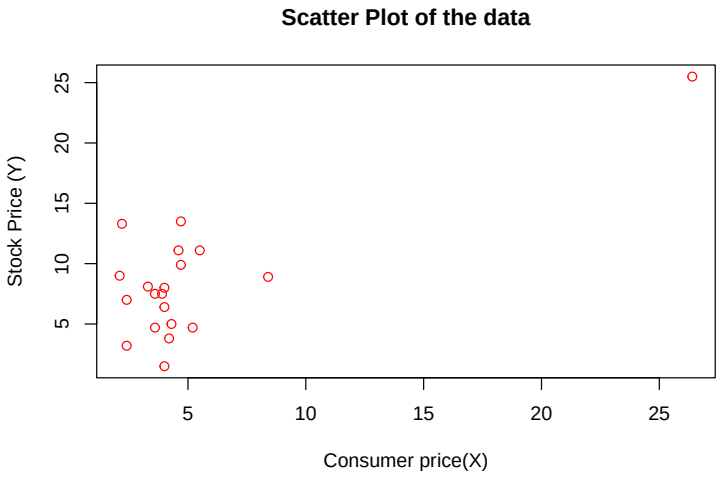
1. To fit a regression Model
2. Test for heteroscedasticity
3. Fit a suitable regression in the presence of heteroscedasticity.

3.2.1 Fitting Regression Model

Before fitting a model it is always better to have a visualization to the data so we create a scatter plot to observe the pattern in the data.

From the plot it is difficult to assume the relationship between the variables.

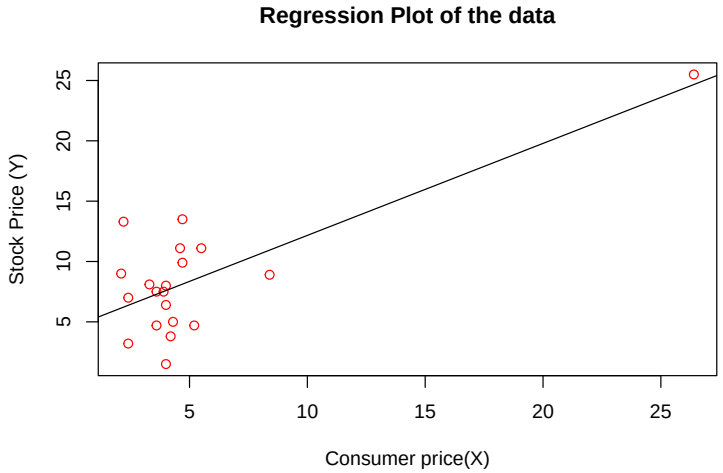
So, we fit a simple linear regression to the data, as:



Output:

Parameters	***Estimate***	***Std. Error***	$t - value$	$Pr(>\|t\|)$
Intercept	4.5400	1.0799	4.204	0.000533
x_var	0.7623	0.1493	5.108	7.36e-05

Residual standard error: 3.375 on 18 degrees of freedom
Multiple R-squared: 0.5917
Adjusted R-squared: 0.569
F-statistic: 26.09 on 1 and 18 DF
p-value: 7.361e-05



Observation :

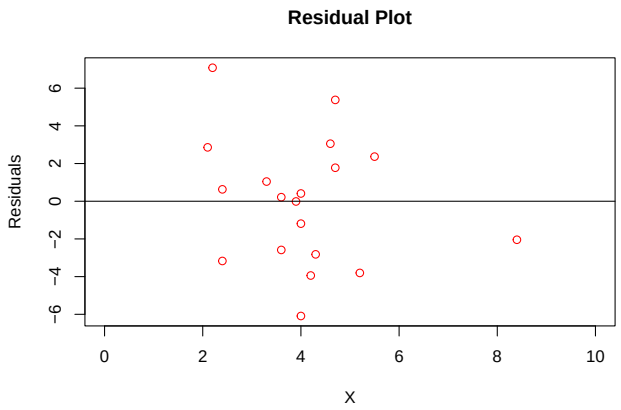
From the fitting it can be shown that F test is statistically significant and the coefficients are also significant at 5% level. So,we can say the data fit the model fit the data well if the model doesn't avoid the assumptions.

Conclusion 1 :

The model will be good enough if we can analyse if the data satisfies the assumptions of a linear model. For a linear model the error-variances must be constant and the errors component must be uncorrelated. If any of these assumptions of linear model is avoided, then the OLS estimator doesn't work better. Then we need to look for the GLS estimators. In the further steps we will try to find out whether the data satisfies the assumptions of the linear model or not.

3.2.2 Detecting Heteroscedasticity

The residual plot of the fit gives us an idea about the variability of the residual. Note that for practical purpose the residuals are treated as an estimator of error component.



Residual Plot

It's difficult to comment about the variability of the residuals seeing the residual plot so it's better if we go for statistical test.

Handling Influential observation

Presence of influential observation me effect residuals so it's better to check the influential observations.
From analysis we get that the 5th and 9th observation are influential and as we fit the model, the model comes to be significant but they **Adjusted** R^2 drop down to **0.02**, that is the model explains almost nothing to the data.
So it's better to use the first model for the analysis.

Golfeld Quandt Test

To test for heteroscedasticity in the data we've performed Goldfeld Quandt Test against the alternative of greater than type and get a p value of 0.48, which is > 0.05. So at 5% level of significant we have to accept that the data is homoscedastic.

Conclusion 2 :

Since the Goldfeld Quandt Test indicates the data is homoscedastic so the OLS estimators will perform best for the model.

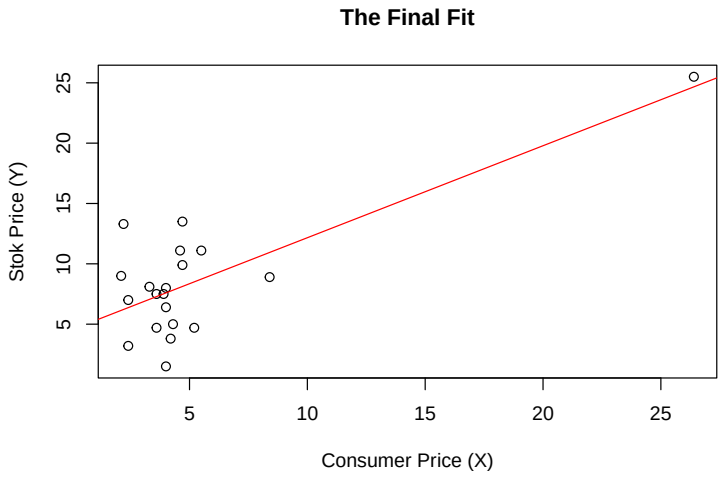
3.2.3 Final Model

Our first model has significant if statistic with deep di with the intercept and slope parameter significant at 5% level and Adjusted R^2 value of 0.56. So the model fits good by explaining 56% of the overall variability. And the estimator of parameters are given by:

Intercept : 4.5400
Speed : 0.7623

So, the final model, which we can suggest is:

$$\textit{Stock\_price} = 4.5400 + 0.7623 \times \textit{Consumer\_Price}$$



3.3 Summary of Conclusions

1. We fit a simple linear regression on the data and the F statistic along with the parameters comes out significant so the fitting will be good enough if the data satisfies the assumption of a linear model that is error-variances are constant and the errors are uncorrelated. If any of these assumption does not satisfied by the data OLS estimator doesn't perform well we need to look for GLS estimators.
2. from the residual plot of the fit it's difficult to comment anything about the variability of residuals. That's why we take the help of statistical test to test for heteroscedasticity in the data.
3. Presence of influential observation may also affect the residuals so we find the influential observations and refit the model after removing those but we get a significant model with Adjusted R^2 of 0.02 which indicates the model explain almost nothing about the data so it's better to stick with the original model with full data set.
4. As we perform Goldfeld Quandt test with alternative of greater than type the test comes to be insignificant so you can see the data is homoscedastic and the OLS estimators work best here.
5. As we get a significant fitting with the first fit we'll stick to that fitting and suggest that for further references.

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