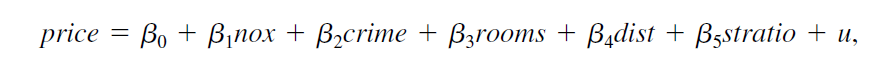
Wooldridge Source: D. Harrison and D.L. Rubinfeld (1978), “Hedonic Housing Prices and the Demand for Clean Air,” by Harrison, D. and D.L.Rubinfeld, Journal of Environmental Economics and Management 5, 81-102. Diego Garcia, a former Ph.D. student in economics at MIT, kindly provided these data, which he obtained from the book Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, by D.A. Belsey, E. Kuh, and R. Welsch, 1990.



The above model used in the **Wooldridge text book** (*Wooldridge, Example 6.1, page: 186*). This dataset (hprice2) contains information on variables such as median housing prices, median income levels, average family size, and so on, for fairly small geographical areas.

**Explain the theory behind my model**

By following the above model, I have designed the following regression equation where, I have omitted just one variable from the original one. But unfortunately, my **dataset doesn’t have any dummy variable**.

Here in the model,

**Dependent variable**

* **price**: median housing price, $

**Independent variables**

* **rooms:** avg number of rooms
* **stratio:** average student-teacher ratio
* **crime:** crimes committed per capita
* **nox:** nit ox concen; parts per 100m

In this model, I am going to estimate the relationship between housing price and independent variables (rooms, stratio, crime, nox) by using OLS method. So, the model can answer the following question:

1. How much housing price increase for additional room?
2. If student-teacher ratio increases, how house price will change?
3. Does crime effect the housing price?
4. Does nox increase or decrease house price?

**Determine the functional form**

To determine the functional form, I am going to use ramsey reset test.

There are multiple ways to test the specification errors. Here, to determine the functional form, I am going to use Ramsey Reset test.

**Assignment 16: The steps involved in RESET are as follows:**

**1.** From the chosen model, obtain the estimated result of *price*, that is *yhat*.

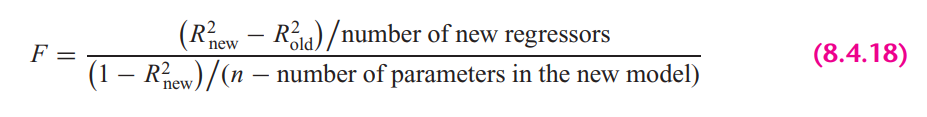




**2.** Rerunning the regression Equation by introducing yhat\_2, yhat\_3, yhat\_4 in some form as an additional regressors. Thus, we run,



**3.** Let obtained R2 from both equation and calculating F value by applying the following formula:



In **STATA**, the calculation looks as follows:



**5.** If the computed F value is significant, at the 5 percent level, one can accept the hypothesis that the model is mis-specified.

**Findings**: From the results we can see that, F value is significant at 5% percent significance level, therefore, we can not reject the hypothesis that the model is misspecified.

**STATA version Ramsey rest** test gives the following result:



It also suggest the same result. Here we cannot reject the null hypothesis (Model has no omitted variables) at 95% confidence level. So, from all of the test, we can conclude that, the chosen model is misspecified.

**Explaining the OLS equation**

In STATA, by the following command, obtained regression result.



**Findings from the result:**

**R-squared:** R-Squared is the proportion of variance in the dependent variable (*price*) which can be predicted from the independent variables (*nox, crime, rooms, stratio*). This value indicates that **61.40%** of the variance in price can be predicted from the variables *nox, crime, rooms, stratio*

**nox:** If nox increase per 100 meters, will expect a decrease in housing price by 1303.7 dollar, holding other variables constant. The variable is statistically significant at 5% significance level.

**crime:** If crime committed per capita increase by 1 more, will expect a decrease in housing price by $ 141, holding other variables constant. The variable is statistically significant at 5% significance level.

**rooms:** If one extra room added in to the house, will expect an increase in housing price by 6944 dollars, holding other variables constant. The variable is statistically significant at 5% significance level.

**stratio:** If average teacher per student increase by 1%, will expect a decrease in housing price by $1050, holding other variables constant. The variable is statistically significant at 5% significance level.

**\_const**: Average housing price is 6014 dollars but it is statistically insignificant at 95% confidence interval

**Heteroskedasticity test**

1. **Graphical Method**

To plot the heteroskedasticity I have followed the following steps:

1. Run the regression equation and obtained the residuals of this regression equation



1. Plotting residuals against the regression fitted values by STATA built in command





Here, if we look at the residual plot against individual explanatory variable, it looks as follows



****

**Findings**

If we look at the residual plot against fitted values, we can see variance is not constant as fitted value increases. In the same way, we can see the relationship between residuals and explanatory variable is not constant as the value of individual variable either increases or decreases or the variance is not constant across observations.

1. **Park test**
2. Run the regression of Equation and obtain the residuals (µi) of this regression equation.



1. Run the following auxiliary regression:





1. To interpret the auxiliary regression result, we need to formulate the null and the alternative hypotheses. The null hypothesis of homoskedasticity is:

Here, the alternative is that at least one of the a’s is different from zero, in this case variables lcrime, lrooms, stratiocoefficientsare different from zero. So, from the hypothesis assumption, we can reject the null hypothesis. Therefore, we can conclude that park test says, there are heteroskedasticity presence in the model.

1. Computing the LM statistics (LM = nR2), where n is the number of observations used in order to estimate the auxiliary regression in Step 2, and R2 is the coefficient of determination of this regression. The LM-statistic is distributed under a chi-square distribution with degrees of freedom equal to the number of slope coefficients included in the auxiliary regression (or k − 1), which in my case is 4 and significance level is 5%.



**Findings from the Park Test**

As we have seen at step 3 that we can reject the null hypothesis and also from the LM test we can see that, NR2 is greater than the critical chi2 value. So, in this case we also can reject the null hypothesis of constant variance (presence of heteroskedasticity).

1. **Glesjer test**

The Glesjer test can be performed in STATA as follows:

1. First, the regression equation model is estimated with OLS, using the predict command is used to obtain the residuals (ei)



1. Run the following auxiliary regression:





1. To interpret the auxiliary regression result, formulate the null and the alternative hypotheses. The null hypothesis of homoskedasticity is:

Here, the alternative is that at least one of the a’s is different from zero, in which case at least one of the variable’s affects the variance of the residuals, which will be different for different i. From the auxiliary regression, we can see explanatory variables, *nox and rooms,* p-value is smaller than the alpha (level of significance) value, which makes the variables statistically significant and different from zero. So, from the hypothesis assumption, we can reject the null hypothesis. Therefore, we can conclude that Glesjer test says, there are heteroskedasticity presence in the model.

1. Computing the LM statistics (LM = nR2), where n is the number of observations used in order to estimate the auxiliary regression in Step 2, and R2 is the coefficient of determination of this regression. The LM-statistic follows the χ2 (chi-square) distribution with p − 1 degrees of freedom.



**Findings from the Glesjer Test**

As we have seen at step 3 that we can reject the null hypothesis and also from the LM test we can see that, NR2 is greater than the critical chi2 value. So, in this case we also can reject the null hypothesis of constant variance, which indicates the evidence of heteroskedasticity.

1. **Gold field Quandt test**

To detect the heteroskedasticity by gold field quandt test, involving the following steps:

1. Sort the data according to the variable *nox*.



1. Breaking the sample into two different sub-samples. To choose the sub samples, the following formula can be applied:



So, from the first and last, sample size is 203 by excluding the middle observations.

1. Now run OLS for both sub-samples in order to obtain the Mean square of residual (RSS/df), using the following commands:





1. Calculating F-statistics for Gold Quandt, F-critical and P-value as follows:



**Findings from the Gold Field Quandt Test**

Final conclusion can be made from the F-statistics and F-critical values. Since F-statistics is greater than the F-critical value, therefore it indicates the evidence heteroskedasticity.

1. **Breusch-Pagan Godfrey test**
2. Estimate Eq. by OLS and obtain the residuals



1. Obtaining variance of the regression by applying the following calculations in STATA



1. Constructing variables Pi defined as



1. Regress Pi thus constructed on the Z’s as



1. Obtaining the ESS from the above regression result and defining theta as follows:



1. Theta follows the chi-square distribution with (K − 1) degrees of freedom, so the chi2 critical values with 4 degrees of freedom and 5% significance level as follows:



**Findings from the Breusch-Pagan Godfrey Test**

If in a model the computed Theta (= χ2) exceeds the critical χ2 value at the chosen level of significance, one can reject the hypothesis of homoscedasticity. Here from the result, we can see that THETA > chi2, therefore it indicates the presence of heteroskedasticity in the model.

1. **White’s general heteroskedasticity test**
2. The regression equation model is estimated with OLS, using the command to obtain the residuals (ui)



1. Run the following auxiliary regression





1. To interpret the auxiliary regression result, formulate the null and the alternative hypotheses. The null hypothesis of homoskedasticity is:

Here, Alternative is that at least one of the a’s is different from zero, in this case variables, From the auxiliary regression, we can see that some of the variables are statistically significant therefore the variables coefficients are different from zero. So, from the hypothesis assumption, we can reject the null hypothesis. Therefore, we can conclude that General white test says, there are heteroskedasticity presence in the model.

1. Computing the LM statistics (LM = nR2), where n is the number of observations used in order to estimate the auxiliary regression in Step 2, and R2 is the coefficient of determination of this regression. The LM-statistic follows the χ2 (chi-square) distribution with p − 1 degrees of freedom.



**Findings from the white LM test**

Final conclusion can be made by comparing the LM statistics and chi square critical value. Reject the null and conclude that there is significant evidence of heteroskedasticity when LM-statistical is greater than the critical value (LM-stat > χ2 7, 0.5).

**Autocorrelation test**

1. **Graphical method to detect Autocorrelation**

First the regression equation model is estimated with OLS, using the following command is used to obtain the residuals (ui)



Because our dataset includes cross-sectional data, we need to generate time variable to plot the residuals against time.



The following command is used to create the lagged series of residuals. Here *lag1\_uhat* is for the lag operator of first order.





**Findings:** Here, it’s hard to find any relationship by looking the scatter plot of against , It seems like they have zero autocorrelation or have very weak positive relationship.

1. **Runs test**

A run is defined to be a succession of one or more identical symbols which are followed and proceeded be a different or no symbol at all.

In the run test the hypothesizes are,

Here, by run test we find out how many times a positive trend became negative and how many times negative trend became positive by crossing mean or median value. For the error term threshold is 0.



To see the visual how many times the error term run positively and negatively across the time.



**Findings**: Here, we can see p-value is less than 0.05 and we reject null hypothesis. So, run test says that the error terms are autocorrelated.

1. **Durbin Watson test**

In STATA, the following two steps required for Durbin Watson test:

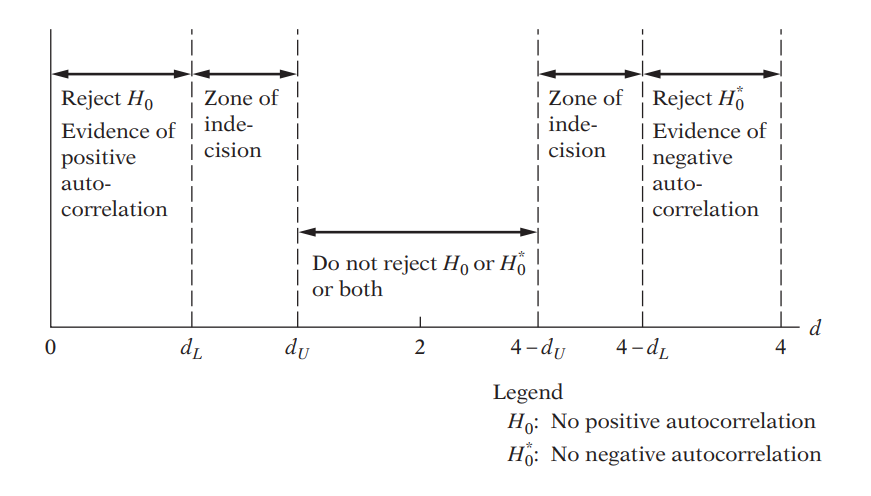
1. Estimate the model by using OLS



1. Estimate DW test value by the following command



**A rule of thumb for d-Watson test:**



**Findings**: In my case, d-statistics is 0.827, In my case total sample size is: 506 and no. of term with intercept is 5, so, in this case, upper limit is 1.80942 and lower limit is 1.72789. so, (4-0.827) = 3.173, which is greater than D upper limit, in this case, we can say, no correlation exists.

1. **Breusch-Godfrey test**

In the BG test, hypothesis are as follows:

Estimating the OLS and obtaining residuals



Here I am using 2 lags of orders to see the autocorrelation residuals and with its previous 2 lags



Moving Average equation looks as follows with 2 lags order:

In Stata, the result looks as follows:



**Findings**: From the result, we can see both lags are statistically significant at 5% significance level means both lags influenced current lag positively.



**Findings from the LM test:** Here we can see LM-vaue > chi2 critical value, therefore we will reject the null hypothesis and can conclude that the model has autocorrelation



**Multicollinearity test**

**11. Looking at the value of R-squared and t value**



**Findings**

In this regression model, R squared is not too high, 61.40%. If we look at t-statistics of the explanatory variables, we can see that all the t value is higher, where all of the variables are statistically significant at 95% confidence interval. In this case, we can say that there are no multicollinearity presents among the explanatory variables.

**12. Pair-wise correlations among regressors**

In STATA, by the following command we can get pair wise correlation value of the variables.



**Findings**

From the results, we can see that all the variables are pair wise correlated. But here none of the variables are highly pair wise correlated. Therefore, we can conclude that there isn’t enough pairwise correlation among regressors which can cause multicollinearity problem.

**13. Auxiliary regression for multicollinearity**

So, first auxiliary regression, where nox is the dependent variable



**Findings**: If multicollinearity were presents, R square from the auxiliary regression would be very high but we can see it’s very low means that the variable is not collinear with other independent variables.

So, second auxiliary regression, where crime is the dependent variable



**Findings**: If multicollinearity were presents, R square from the auxiliary regression would be very high but we can see it’s very low means that the variable is not collinear with other independent variables.

Third auxiliary regression, where stratio is the dependent variable



**Findings**: If multicollinearity were presents, R square from the auxiliary regression would be very high but we can see it’s very low means that the variable is not collinear with other independent variables.



**Findings**: If multicollinearity were presents, R square from the auxiliary regression would be very high but we can see it’s very low means that the variable is not collinear with other independent variables.

**14. Partial Correlations**

In STATA, by the following command we can get partial correlation value of the variables



**Findings**: From the results, we can see that variables are very weekly partially correlated. But here none of the variables are highly partially correlated. Therefore, we can conclude that there isn’t enough partial correlation among regressors which can cause multicollinearity problem.

**15. Condition Index**

In STATA, by the following command, we can obtain the Eigenvalue and corresponding Conditional Index.



**Rule of thumb:** If k is between 100 and 1000 there is moderate to strong multicollinearity and if it exceeds 1000 there is severe multicollinearity. Alternatively, if the (CI = √k) is between 10 and 30, there is moderate to strong multicollinearity and if it exceeds 30 there is severe multicollinearity

**Findings**

From all of the observation, can see that Conditional index, Eigen value, VIF is very low for the explanatory variables. Therefore, we can conclude that, multicollinearity doesn’t exist among the regressors.