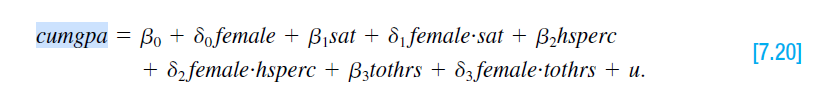
**Source:** Wooldridge text book.



The above model used in the **Wooldridge text book** (*Wooldridge, pages: chapter 4, c11, page: 237*).

**Explaining the theory behind my model**

By following the above model, I have conducted the following regression equation.

Here,

**Dependent variable**

* **cumgpa:** cumulative GPA

**Independent variables**

* **hsperc:** percentile in h.s.
* **sat**: SAT score
* **tothrs:** total hours prior to term
* **female:** =1 if female

By using OLS method, I am going estimate the linear relationship between cumgpa and hsperc, sat, tothrs, female. Here by this regression, we can estimate, if total hours prior term increase, how cumulative GPA changes and wheather a female has more cumgpa than male or not. If someone had good marks at hsperc and sat, are there cumgpa is likely to increase or not.

**Determining the functional form**

Among OLS assumptions, one of the assumptions of the classical linear regression model (CLRM), is that the regression model used in the analysis is “correctly” specified: If the model is not “correctly” specified, encounter the problem of model specification error or model specification bias. It can be happened for various reason. There are multiple ways to test the specification errors. Here, I am going to use **Ramsey Reset test**, to determine the functional form

**Assignment 16: Ramsey RESET test**

1. From the selected model, obtain the estimated result of *cumgpa*, as *yhat*.

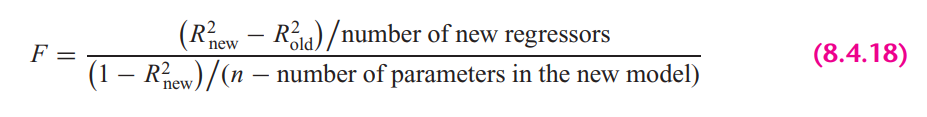




**2.** Rerunning the regression Equation by introducing yhat\_2, yhat\_3, yhat\_4 in some form as an additional regressors. Thus, we run,



**3.** To obtain R2 from both equation and calculating F value by applying the following formula: *from Gujrati book*



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.

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

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



Here, we cannot reject the null hypothesis (Model has no omitted variables) at 95% confidence level. Therefore, the selected model is not misspecified.

**Explaining the OLS equation**

The following command used in STATA to obtain regression result.



**Explanations of the regression result:**

**R-squared:** R-Squared is the proportion of variance in the dependent variable (*educ*) which can be predicted from the independent variables (*hsperc, sat, tothrs, female*). This value indicates that **24.05%** of the variance in *cumgpa* can be predicted from the variables *hsperc, sat, tothrs, female***.**

**hsperc:** The coefficient is -.0068 means that If mother *hs* mark increase by 1 percent, will expect a decrease in *cumgpa* (educ) by -.0068 point, holding other variables constant. The variable is statistically significant at 5% significance level.

**sat:** The coefficient is .0009 means that If sat score increase by 1 point, will expect an increase in *cumgpa* by .0009 point, holding other variables constant. The variable is statistically significant at 5% significance level.

**tothrs:** The coefficient is .0101 means that if total hours of study in prior term increase by 1 hour, will expect an increase in *cumgpa* by .0101 point, holding other variables constant. The variable is statistically significant at 5% significance level.

**female:** The coefficient is .168 means for female students estimated cumgpa is higher by .168 point than the male students, holding other variables constant. The variable is statistically significant at 5% significance level.

**const**: When all the explanatory variables value is 0, in average *cumgpa* is .84

**Heteroskedasticity test**

Having an equal variance means that the disturbances are homoscedastic. However, it is quite common in regression analysis for this assumption to be violated.

A sequence of random variables is homoscedastic if all its random variables have the same finite variance. This is also known as homogeneity of variance. The complementary notion is called heteroscedasticity, where:

**Assignment 1: Graphical Method**

Heteroskedasticity can be detected by the scatter plot. Graphical method is also known as the informal way, is by inspection of different graphs.

To plot the heteroskedasticity, the following steps are as follows:

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





To plot the explanatory variables against residuals:





**Interpretation**

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.

**Assignment 2: Park test**

1. Run the regression of Equation and obtain the residuals (µi) of this regression equation.



1. Run the following auxiliary regression:

Since, log of 0, can make a variable undefined. Therefore, I didn’t transform female variable to log(female).





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

From the auxiliary regression, we can see explanatory variables, *lhsperc and ltothrs,* p-value is greater than the alpha (level of significance) value, which makes the variables statistically insignificant. In this case we do not reject the null hypothesis or . Here, the alternative is that at least one of the a’s is different from zero, in this case variables *lsat and female* coefficientsare 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. To estimate the chi2 distribution, in this case df is 4 and significance level is 5%.



**Interpretation**

From the 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).

**Assignment 3: 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 to obtain the residuals (ei)



1. Run the following auxiliary regression:





1. In this case, 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, *sat and female,* p-value is greater than the alpha (level of significance) value, which makes the variables statistically insignificant. In this case we do not reject the null hypothesis or . Here, the alternative is that at least one of the a’s is different from zero, in this case variables hsperc *and tothrs* coefficientsare 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.



**Interpretation from the Glesjer Test**

From the step 3 that we can reject the null hypothesis and also from the LM test we can see that, LM statistics 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.

**Assignment 4: 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 *sat*.



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 293 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:



**Interpretation from the Gold Field Quandt Test**

From the F-statistics and F-critical values. Since F-statistics is lower than the F-critical value, and we can see the corresponding p value is greater than 0.05, which make F value insignificant, therefore it indicates the no heteroskedasticity if we sort it by sat score.

**Assignment 5: Breusch-Pagan Godfrey test**

1. Estimating 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:



**Interpretation 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.

**Assignment 6: White’s general heteroskedasticity test**

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



1. Run the following auxiliary regression

Here, I have omitted female2, since female is a dummy variable, adding square term of a dummy variable can create multicollinearity.



Run the auxiliary regression by the following command:



1. Formulating the null and the alternative hypotheses. The null and alternative hypothesis of white test is as follows:

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 some of the explanatory variable’s p-value is greater than the alpha (level of significance) value, which makes the variables statistically insignificant. But the alternative is that at least one of the a’s is different from zero, in this case variables sat, tothrs, sat\_sq *and tothrs\_sq* coefficientsare 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.



**Interpretation from the white LM test**

From the LM statistics and chi square critical value, we can 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**

Assumptions from the of the CLRM states that the covariances and correlations between different disturbances are all zero.

This assumption states that the error terms and are independently distributed, termed serial independence. If this assumption is no longer true then the disturbances are not pairwise independent, but are pairwise autocorrelated.

In this situation:

which means that an error occurring at period i may be correlated with one at period j.

**Assignment 7: Graphical method to detect Autocorrelation**

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



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 *error\_lag1* 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.

**Assignment 8: 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.





**Interpretation**: 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.

**Assignment 9: Durbin Watson test**

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

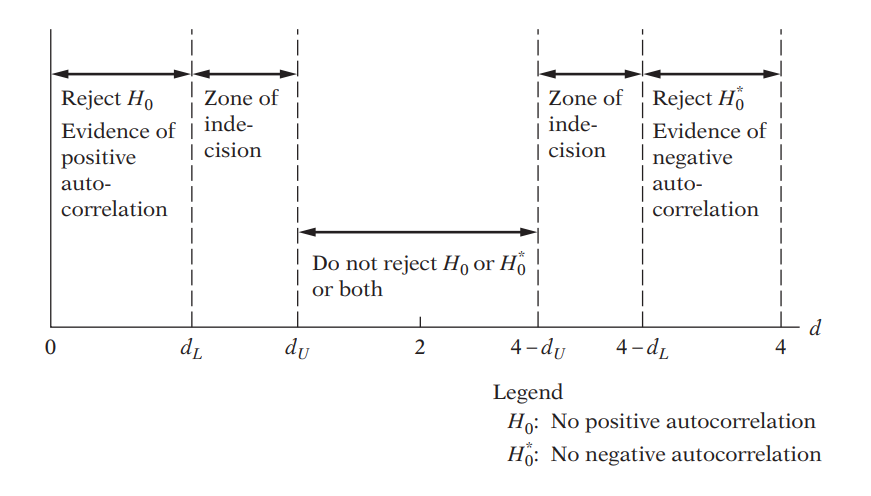
1. Estimate the model by using OLS



1. Estimate DW test value by the following command



**A rule of thumb for d-Watson test: from Gujrati book**



Here,

**Findings**: In my case, d-statistics is 2. Therefore, according to Durbin Watson test, error terms are not autocorrelated.

**Assignment 10: Breusch-Godfrey test**

BG test’s 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 lag is statistically insignificant at 5% significance level means previous up to 2 lags has no correlation with current error term.



**Findings from the LM test:** Here we can see NR2 < chi2 critical value, therefore we cannot reject the null hypothesis and can conclude that the model has **no autocorrelation**.

The following test can be obtained by STATA.



Same results found.

**Multicollinearity test**

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



**Interpretation**

In this regression model, R squared is not too high, 24.05%. 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.

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

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



**Interpretation**

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.

**Assignment 13. Auxiliary regression for multicollinearity**

Since multicollinearity arises because one or more of the regressors are exact or approximately linear combinations of the other regressors, one way of finding out which X variable is related to other X variables is to regress each Xi on the remaining X variables and compute the corresponding R2, which we designate as R2i.

So, first auxiliary regression, where ***hsperc*** is the dependent variable



**Interpretation**: In general, if multicollinearity were presents, R square from the auxiliary regression would be very high but we can see it’s significantly high. From that we can say that hsperc does not causes to create collinearity with other independent variables.

So, second auxiliary regression, where ***sat*** is the dependent variable



**Interpretation**: In general, if multicollinearity were presents, R square from the auxiliary regression would be very high but we can see it’s significantly high. From that we can say that sat does not causes to create collinearity with other independent variables.

Third auxiliary regression, where ***female*** is the dependent variable



**Interpretation**: In general, if multicollinearity were presents, R square from the auxiliary regression would be very high but we can see it’s significantly high. From that we can say that female does not causes to create collinearity with other independent variables.

Fourth auxiliary regression, where ***tothrs*** is the dependent variable



**Interpretation**: In general, if multicollinearity were presents, R square from the auxiliary regression would be very high but we can see it’s significantly high. From that we can say that tothrs does not causes to create collinearity with other independent variables.

**Assignment 14: Partial Correlations**

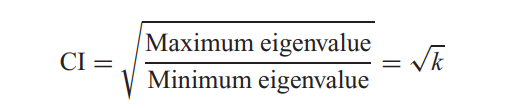
In STATA, by the following command we can get partial correlation value of the variables



**Interpretation**: 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.

**Assignment 15: Condition Index**

Conditional Index can be calculated as follows:



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.