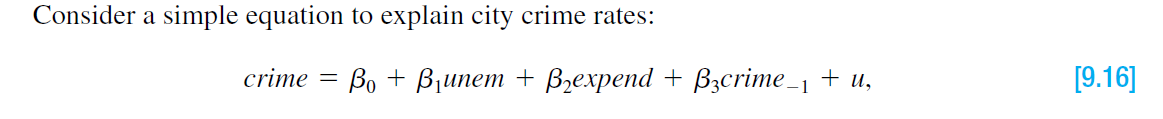
**Source:** These data were collected by David Dicicco, a former MSU undergraduate, for a final project. They came from various issues of the County and City Data Book, and are for the years 1982 and 1985. The following model has found on Introductory economics text book by Wooldridge (*Wooldridge, pages: chapter 9, equation 9.16, page: 303*)



**Explaining the theory behind my model**

Based on the above regression equation, I have come up with the following model. Unfortunatey, my dataset (crime2) doesn’t include any dummy variable. I haven’t found any dummy variable in the above model so I added a dummy variable called *nrtheast*.

Here in the model,

**Dependent variable**

* **crimes:** total number index crimes

**Independent variables**

* **nrtheast:** =1 if city in NE
* **lawexpc:** law enforce. expend. pc, $
* **unem:** unemployment rate
* **cimest-1**: no. of crime at t-1

The theory behind of my model is here I am going to estimate the linear relationship between *crimes* (total number index crimes), based on *northeast, lawexpc, unem, crimest-1* by using OLS method. For example, does crimes increase if unemployment rate goes high or not? Is there any relationship between previous crimes index to current crimes index? What if law enforce expenditure increase, do crime decrease or not? Is crime higher or lower in the northeast area than other area? By using this model, we can answer this question or we can estimate crimes index based on these independent variables.

**Determining the functional form**

One of the assumptions of the classical linear regression model (CLRM), Assumption 9, 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. So, my regression equation looks as follows:

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

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

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

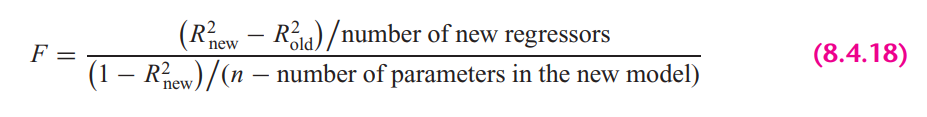




**Step 2.** Rerunning the regression Equation by introducing yhat\_2, yhat\_3, yhat\_4 in some form as an additional regressors. Thus, we run,



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



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



**Rule:** 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 insignificant at 5% percent significance level, therefore, we can reject the hypothesis that the model is misspecified and can say our model is correctly specified

**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 not 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 (*crimes*) which can be esitmated from the independent variables (*unem, crimes\_lag1, lawexpc, northeast*). This value indicates that **37.16%** of the variance in crimes can be predicted from the variables *unem, crimes\_lag1, lawexpc, northeast*

**unem:** If unemployment rate increase by 1%, total number of crimes no. increase by 1664, holding other variables constant. The variable is statistically significant at 5% significance level.

**Crimes\_lag1:** Current period crime rate has positive relationship with previous period crime rate. As per result, if previous period crimes no. was 10, then the expected crime will be 10\*.473 in the current period, holding other variables constant. The variable is statistically significant at 5% significance level.

**lawexpc:** If law enforce expenditure per capita increase by $1, crimes no. will be higher 19.19, holding other variables constant. The variable is statistically significant at 5% significance level. Although, it doesn’t make sense.

**nrtheast:** Crimes no. in northeast is lower by 15474 than any other regions, holding other variables constant. The variable is statistically significant at 5% significance level.

**const**: When all the explanatory variables value is 0, in average *crimes no.* is -8114

**Heteroskedasticity test**

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:

Therefore, having an equal variance means that the disturbances are homoscedastic. However, it is quite common in regression analysis for this assumption to be violated. In this assignment, I am going to test heteroskedasticity by applying different methods.

1. **Graphical Method**

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

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:

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





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 all the explanatory variables 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 no variables are significant. So, from the hypothesis assumption, we cannot reject the null hypothesis. Therefore, we can conclude that park test says, there are homoskedasticity 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 cannot reject the null hypothesis and also from the LM test we can see that, NR2 is smaller than the critical chi2 value. So, in this case we also cannot reject the null hypothesis of constant variance (absence 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 command and predict (gen) command is used 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, 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, . 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 *unem*.



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 37 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 (ei)



1. Run the following auxiliary regression

Here, I have omitted nrtheast2, since urban is a dummy variable, adding square term of a dummy variable can create multicollinearity.





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. In this case variables unem *and unem\_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.



**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**

One of the assumptions of the CLRM states that the covariances and correlations between different disturbances are all zero. 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.

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 *err\_lag1* is for the lag operator of first order.







**Findings:** By looking the scatter plot of against , It seems like they have zero autocorrelation.

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 greater than 0.05 and we cannot reject the null hypothesis of no autocorrelation. Therefore, run test says that the error terms are not autocorrelated.

1. **Durbin Watson test**

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

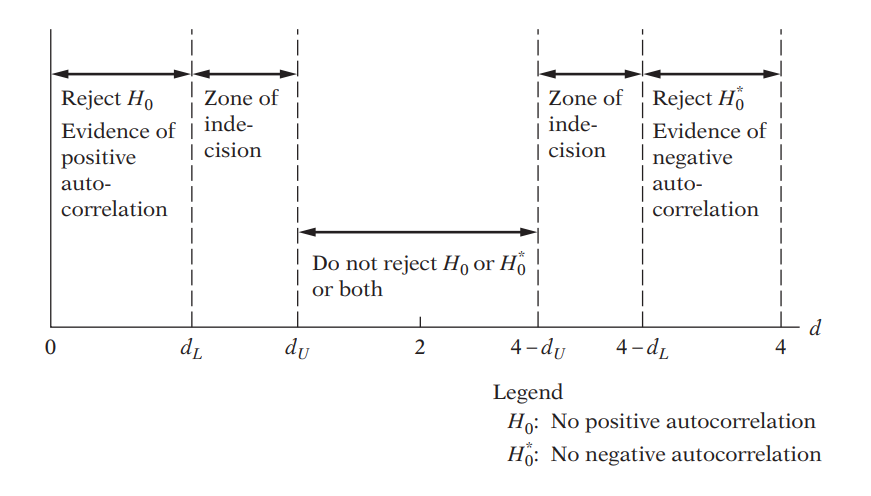
1. Estimate the model by using OLS



1. Estimate DW test value by the following command



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



**Findings**: In my case, d-statistics is 1,875, which is close to 2. Therefore, according to Durbin Watson test, most likely this is the area of zone of indecision or zero autocorrelation.

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 lag is statistically insignificant at 5% significance level means no previous 2 error term influence current error term.



**Findings from the LM test:** Here we can see LM stat < chi2 critical value, therefore we cannot reject the null hypothesis and can conclude that the model has no 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, 37.16%. 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**

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 *unem* 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 unem is not collinear with other independent variables.

So, second auxiliary regression, where crimes\_lag1 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 crimes\_lag1 is not collinear with other independent variables

Third auxiliary regression, where nrtheast 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 nrtheast 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 lawexpc is not collinear with other independent variables

**14. Partial Correlations**

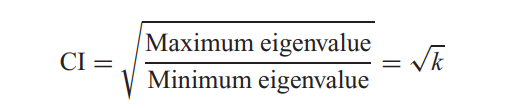
In STATA, by the following command we can get partial correlation value of the variables



**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**

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.