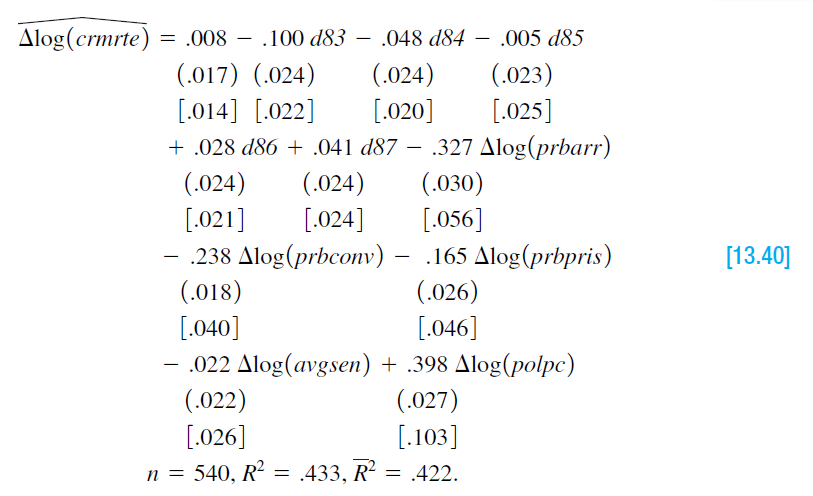
Wooldridge Source: From C. Cornwell and W. Trumball (1994), “Estimating the Economic Model of Crime with Panel Data,” Review of Economics and Statistics 76, 36 -366. Professor Cornwell kindly provided the data.



The above model used in the **Wooldridge text book** (*Wooldridge, Chapter-3, page: 430*).

**Explain the theory behind my model**

By following the above model, I want to estimate a simpler version of their model. The regression equation of my model looks as follows:

Here in the model,

**Dependent variable**

* **crmrte**: log of hourly wage, $

**Independent variables**

* **polpc:** police per capita
* **prbconv:** ’probability’ of conviction
* **prbarr:** ’probability’ of arrest
* **west:** =1 if in western N.C.

From the model, we can estimate the relationship between crime rate and independent variables (polpc, prbconv prbarr) by using OLS method. So, the model can answer the following question:

1. Does increase in police decrease the crime rate?
2. If probability of conviction and arrest increases, does crime rate goes up and down?
3. Does crime rate in the west is higher than other areas?

**Determine the functional form**

There are multiple ways to test the specification errors. I am going to use Ramsey Reset test.

**Ramsey Reset test (Assignment: 16)**

First running the regression equation and estimating, crmrte\_hat by the following command

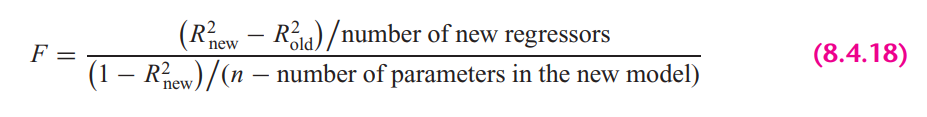




Rerun the regrussion with the yhat order.



Obtaining R2 from both equation and calculating F value by applying the following formula:



In STATA, the calculation looks as follows:



In this case, if the computed F value is significant, at the 5 percent level, one can accept the hypothesis that the model is mis-specified otherwise model is correctly specified.

**Findings**: From the results we can see that, F value is significant at 95% percent confidence interval, therefore, we can not reject the hypothesis that the model is misspecified.

STATA version Ramsey rest test gives the following result:



Here we cannot reject the null hypothesis (Model has no omitted variables) at 5% significance level. So, from all of the test, we can say that the 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 (*crmrte*) which can be predicted from the independent variables (*prbarr, prbconv, polpc, west*). This value indicates that **40.26%** of the variance in *crmrte* can be predicted from the variables *prbarr, prbconv, polpc, west*

**prbarr:** 1% increase in the probability of arrest is predicted to lower the crime rate by about -.047%., holding other variables constant. The variable is statistically significant at 5% significance level.

**prbconv:** 1% increase in the probability of conviction is predicted to lower the crime rate by about -.0032%., holding other variables constant. The variable is statistically significant at 5% significance level.

**polpc:** 1% increase in police per capita increases crime rates by about 3.1 %, holding other variables constant. The variable is statistically significant at 5% significance level.

**west:** Crime rate is lower in west by .014%. The variable is statistically significant at 5% significance level.

**const**: Average crime rate is .0458, when other variables are zero.

# Heteroskedasticity test

## Graphical Method

To plot the heteroskedasticity I have followed the following steps:

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



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





**Findings**

From the residual plot against fitted values, we can see variance is not constant as fitted value increases. From that we can say that heteroskedasticity could be present in the model.

## Park test

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



1. Run the following auxiliary regression:





The hypothesis are:

Here, the alternative is that at least one of the a’s is different from zero, in this case all the variable’scoefficientsare 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.

## Glesjer test

The Glesjer test can be performed in STATA as follows:

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



Run the following auxiliary regression:





The hypothesis are:

Here, the alternative is that at least one of the a’s is different from zero, in this case all the variable’scoefficientsare 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.

## 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 *earns*.



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 252 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**

By comparing F-statistics and F-critical values, we can say that F-statistics is smaller than the F-critical value, therefore it indicates the evidence homoskedasticity if we sort it by earns. We might get different result; we sort the data by different variables.

## Breusch-Pagan Godfrey test

Estimate Eq. by OLS and obtain the residuals



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



Constructing variables Pi and Regress Pi thus constructed on the Z’s as



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



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

If the computed THETA (= χ2) exceeds the critical chi2 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.

Alternatively in STATA, we can test the same result by the following command,



## White’s general heteroskedasticity test

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



Run the following auxiliary regression





In this case,

Here, the alternative is that at least one of the a’s is different from zero, in this case almost all the variable’scoefficientsare 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.



From the LM test, we can see, LM stat > chi2 crit value, in this case, we reject the null hypothesis of constant variance.

Alternatively in STATA: also gives the same result



# Autocorrelation test

## Graphical method to detect Autocorrelation

First the regression equation model is estimated with OLS



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:** By looking the scatter plot of against , It seems like they have positive auto correlation

## Runs test

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.





**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.

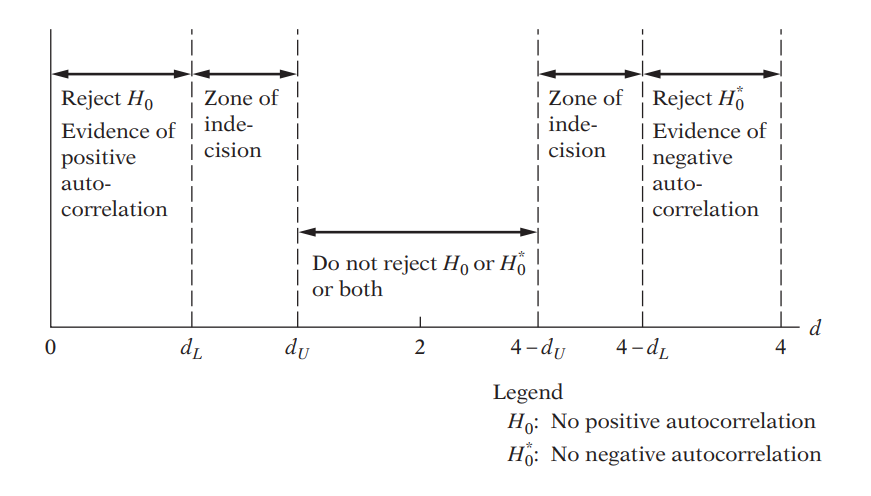
## Durbin Watson test

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

1. Estimate the model by using OLS
2. Estimate DW test value by the following command



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



**Findings**: In my case, d-statistics is close to 0. Therefore, It seems like error terms are positively autocorrelated.

## Breusch-Godfrey test

In this case the 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 first lag is statistically significant at 5% significance. Which means previous error term influence the current error term positively. Here, 2nd lag is insignificant.

**Alternative way in stata**



# Multicollinearity test

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



**Findings**

In this regression model, R squared is not too high,40.26%. 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

## 

**Findings:** First auxiliary regression, where prbarr is the dependent variable and second auxiliary regression, where prbconv is the dependent variable. Here 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**: Third auxiliary regression, where polpc is the dependent variable and fourth auxiliary regression, where west is the dependent variable and. Same conclusion can be drawn that, 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.