Wooldridge Source: Professor Daniel Hamermesh, at the University of Texas, compiled these data from the May 1991 Current Population Survey. Professor Hamermesh kindly provided these data.

The above model used in the **Wooldridge text book** (*Wooldridge, Example 6.1, page: 599*).

**Explain the theory behind my model**

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

Here in the model,

**Dependent variable**

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

**Independent variables**

* hispanic: =1 if wife hispanic
* exper: age - educ - 6
* earns: wife’s weekly earnings
* educ: wife’s yrs schooling

In this model, I am going to estimate the relationship between hourly wage and independent variables (Hispanic, exper, earns, educ) by using OLS method. So, the model can answer the following question:

1. If wife is Hispanic, how wage changes?
2. How much wage varies in terms of experience?
3. Does total weekly earnings reflect the hourly wage?
4. If someone is more educated, how much extra wage she earns?

**Determine the functional form**

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

**Ramsey Reset test**

First run the regression and estimating, lwage\_hat

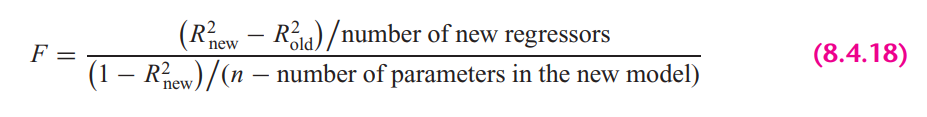




Rerun the regrussion with newly introduced hat variables



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



In STATA, the calculation looks as follows:



Here the rule is 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 suggests the same 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 (*lwage*) which can be predicted from the independent variables (educ, earns, exper, hispanic). This value indicates that **65.93%** of the variance in price can be predicted from the variables educ, earns, exper, hispanic.

**educ:** If wife’s years of schooling increases by 1 year, we will expect an increase in hourly wage by 2.8%, holding other variables constant. The variable is statistically significant at 5% significance level.

**earns:** If wife’s weekly average income increases by $1, will expect an increase in hourly wage by 0.16, holding other variables constant. The variable is statistically significant at 5% significance level.

**exper:** If experience increase by 1 year, expected hourly wage will increase by .14%, holding other variables constant. The variable is statistically significant at 5% significance level.

**hispanic:** If the wife is Hispanic, expected income will decrease by 6.1% than the mean value, holding other variables constant. The variable is statistically significant at 5% significance level.

**\_const**: Average hourly wage is 1.17

# **Heteroskedasticity test**

1. **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**

If we look at the residual plot against fitted values, we can see variance is not constant as fitted value increases. From that we can say, this is possible that my model has heteroskedasticity.

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



1. 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 some of 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.

1. **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:





In this case,

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

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 *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 1315 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.

1. **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 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.

Alternatively in STATA



1. **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 some of 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.

Alternatively in STATA: also gives the same result



# **Autocorrelation test**

1. **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 *epsilon\_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.

1. **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.

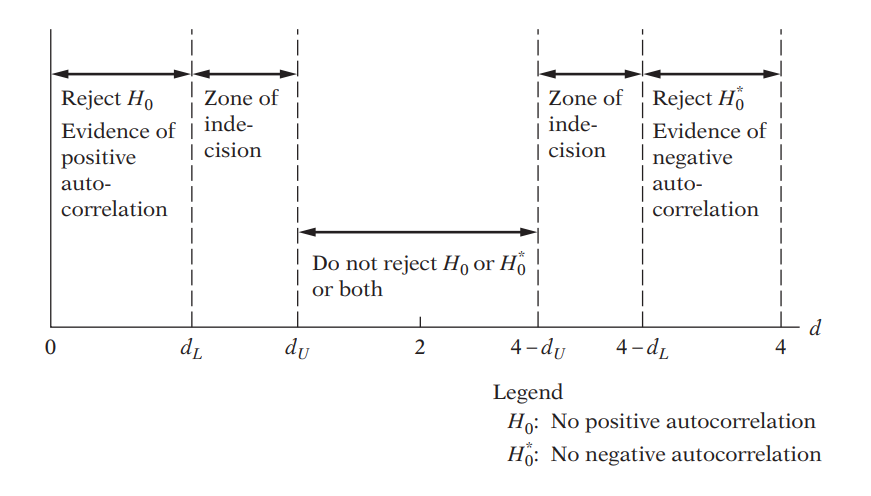
1. **Durbin Watson test**

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 2. Therefore we cannot reject null hypothesis, which says there are no autocorrelation in the model

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 insignificant at 5% significance. Which means previous error term doesn’t influence the current error term.

**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, 65.93%. 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 educ 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 earns 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 exper 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.