Stata Lab 5.

Multivariate Regression 1

Dr. Adam William Chalmers

Table of Contents

[Part 1. Multivariate Regression Analysis 2](#_Toc97967273)

[Step 1. Do the IVs and the DV covary? 3](#_Toc97967274)

[Step 2. Checking for multicollinearity (correlation between IVs); pre-estimation. 6](#_Toc97967275)

[Multivariate regression analysis: estimation and interpretation. 8](#_Toc97967276)

[Stata Exercises. 1 10](#_Toc97967277)

[Part 2. Extensions of the ‘regress’ command 11](#_Toc97967278)

[Part 3. Increment in R-squared: Semipartial correlations 13](#_Toc97967279)

[Stata Exercises. 2 14](#_Toc97967280)

[Part 4. Regression Diagnostics & Postestimation Functions: Gauss-Markov Conditions 15](#_Toc97967281)

[Assumption 1. Normal Distribution of the DV 15](#_Toc97967282)

[Visual Inspection 15](#_Toc97967283)

[Assessing skewness and kurtosis 16](#_Toc97967284)

[Significance of Skewness and Kurtosis 16](#_Toc97967285)

[Dealing with Non-Normal Distribution of your DV 17](#_Toc97967286)

[Log Transformation and Normalizing Distribution 17](#_Toc97967287)

[Assumption 2. Normal Distribution of Residuals 19](#_Toc97967288)

[Significance of Skewness and Kurtosis for Residuals 20](#_Toc97967289)

[Dealing with Non-Normal Distribution of Residuals: Robust Regression 20](#_Toc97967290)

[Assumption 3. Multicollinearity (post-estimation) 21](#_Toc97967291)

[Dealing with multicollinearity 22](#_Toc97967292)

[Stata Exercises. 3 23](#_Toc97967293)

[Stata Advanced Tasks 24](#_Toc97967294)

[Appendix. How to Present the Results of a Multivariate Regression Analysis with multiple models 26](#_Toc97967295)

# Part 1. Multivariate Regression Analysis

We only rarely carry out bivariate regression analyses. In fact, we know that the social and political world is a complex place and very rarely does one IV explain one DV. Recall our example bivariate regression analysis from our last Stata Lab. We found that Democracy had an important impact on a country’s GDP, but according to the r-squared statistic, it only explained about ~25% of the variation in the DV. What about the other 75% that is not explained? We know that this means that other variables are also doing quite a bit of the “causal work” when it comes to explaining a country’s GDP.

Multivariate regression analysis, as its name implies, allows us to run a regression analysis with multiple independent variables. The basic Stata syntax is this:

. regress dv *iv1 iv2 iv3 …, options*

In theory, you could include as many independent variables as you like. However, we need to also have a good, logical reason for including variables in our model. In fact, our selection of independent variables needs to have some basis in (existing) theory.

**Example 1**. Going back to our last Stata lab, we are still interested in the relationship between (1) countries being democratic, and (2) their wealth (GDP). We can even test the same hypothesis:

H1 *the more democratic a country, the greater its GDP*

However, I now also want to be able to test alternative explanations and/or include some control variables in my analysis. This serves the purpose of improving the overall explanatory power of my regression model and getting a more accurate picture of the impact of the IV (here Democracy) on the DV.

To this end, I include some additional independent variables in my regression analysis. I have selected these additional IVs based on my reading and knowledge of the “existing literature” (studies that also try to explain the determinants of GDP). Each has its own credible causal story (i.e., there is a logical reason to include; they are not just randomly selected). This includes:

Political corruption (**vdem\_corr**)

* Existing studies suggests that the lower the levels of political corruption in a country, the higher its GDP; corruption impedes economic development.

Trade Freedom (**hf\_trade**)

* Existing studies suggest that the higher the trade freedom in a country, the higher its GDP; trade freedom enhances economic development.

Government expenditure on education as %GDP (**wdi\_expedu**)

* Existing studies suggest that higher levels of education are linked to higher levels of GDP

Long term unemployment as %population (**imf\_ue**)

* Existing studies suggest that lower levels of unemployment are linked to higher levels of GDP

**\*\*Pro-tip**: I find the value labels used in the Quality of Government data to be confusing or, at a minimum, unhelpful. Before I conduct my multivariate analysis, I am going to rename all of these variables like this. Note that I have made up the new variable names myself.

. rename wef\_gdpc GDP

. rename eiu\_iod Democracy

. rename vdem\_corr Political\_Corruption

. rename hf\_trade Trade\_Freedom

. rename wdi\_expedu Education\_expenditure

. rename imf\_ue Unemployment

### Step 1. Do the IVs and the DV covary?

Step 1 is to examine whether our IVs covary with our DV. This is one of the four hurdles of causation we learned about in week 1. We can examine covariation both **visually** and **statistically** (by using the pwcorr command to estimate correlation coefficients). We can evaluate covariation using the charts below.

**Chart, scatter chart

Description automatically generated**

Chart

Description automatically generated with low confidence

#### Visual Inspection

There are two different ways to visually inspect covariation. The first way is to create individual graphs for each IV on the DV. Here are the Stata commands and, below, the output.

. graph twoway (scatter GDP Democracy)(lfit GDP Democracy)

. graph twoway (scatter GDP Political\_Corruption)(lfit GDP Political\_Corruption)

. graph twoway (scatter GDP Trade\_Freedom)(lfit GDP Trade\_Freedom)

. graph twoway (scatter GDP Education\_expenditure)(lfit GDP Education\_expenditure)

. graph twoway (scatter GDP Unemployment)(lfit GDP Unemployment)

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*Interpreting the results*: All five IVs are either positively or negatively correlated with our DV. They also covary in the directed predicted in my hypotheses. There are no clear cases of ‘no variation’.

The second way to visually inspect covariation between the IVs and the DV is to create a scatterplot matrix. In this case, we are unable to include a fitted regression line. However, we should be able to interpret covariation in a scatter plot without the fitted regression line. Here’s the Stata command.

. graph matrix GDP Democracy Political\_Corruption Trade\_Freedom Education\_expenditure Unemployment



*Interpreting the results*. This type of graph can be confusing. However, if we orient ourselves, we can see that a scatterplot matrix graph can be very helpful. In this case, since we are only interested in covariation between the IVs and the DV, we only need to inspect the top row of graph (e.g., those in the red box). If you take a close look, you will notice that these are essentially the same graphs that we produced individually using the graph twoway command. There are, as already noted, no fitted regression lines. However, we can still interpret the results. From left to right, the first box is a scatterplot for GDP and Democracy, the second box is for GDP and Political corruption, the third box is for GDP and Trade freedom, the fourth box is for GDP and Education expenditure, and the fifth box is for GDP and Unemployment. The key thing is to look for instances of ‘no correlation’. Again, it is clear from visual inspection that the IVs are either positively or negatively correlated with the DV.

#### Statistical Inspection

A statistical inspection of covariation between IVs and the DV involves using the pwcorr command and an interpretation of correlation coefficients and p-values. The strength of covariation is assessed by interpreting the correlation coefficient. The absence of covariation is indicated when p>0.05.

Table

Description automatically generated

*Interpreting the results*: this correlation matrix gives us correlation coefficients for all variables. However, for right now we are just interested in the covariation between the IVs and the DV. Hence, we can focus our attention on the results in the red box.

We can see moderate correlations, either negative or positive, between our DV and

Political Corruption (-0.6373), Democracy (0.4718) and Trade Freedom (0.4643). All of these results are statistically significant with a p-value <0.001. Education (0.2406) and Unemployment (-0.3097) are weak positive or negative, with lower but still statistically significant p-values of <0.01..

**Conclusion**: All of our IVs covary with our DV. We also know that there are strong theoretical reasons to include these variables in our multivariate regression (e.g., they are derived from the existing literature). As such, we will not omit any of these variables based on our inspection of covariation with the DV.

### Step 2. Checking for multicollinearity (correlation between IVs); pre-estimation.

Step 2 is to examine covariation between our IVs. Our IVs or ‘independent variables’ need to be **independent** from each other. That means that they cannot be (1) explaining / causing each other and (2) redundant or explaining the same thing or part of the same phenomenon. High correlation between IVs is called multicollinearity. Multicollinearity is a problem for regression analysis because it will limit our ability to assess the independent effects of each IV on the DV.

There are different ways to examine multicollinearity, both pre-estimation and post-estimation. Pre-estimation refers to tests we conduct **before** doing regression analyses. Post-estimation refers to tests we conduct **after** running regression analysis. In this instance, we will examine multicollinearity pre-estimation tests. Pre-estimation examination of multicollinearity is conducted both visually and statistically.

#### Visual Inspection for multicollinearity

The easiest way to visually inspect for multicollinearity is to use the scatterplot matrix. This time we exclude the DV.

. graph matrix Democracy Political\_Corruption Trade\_Freedom Education\_expenditure Unemployment



*Interpreting the results*. In contrast to our examination of the covariation of the DV with our IVs, where we wanted to see positive and negative correlations, we now want to see ‘no correlation’. E.g., no correlation means that we do not have a problem of multicollinearity. A quick look at the matrix suggests that there **might** be some degree of multicollinearity for a number of our IVs. The most obvious example is Democracy on Political corruption and potentially Trade freedom as well as Education expenditure (indicated by the red box). But there might also be an issue of multicollinearity between Political corruption and Trade freedom (green box) and a few other IVs.

To say something more definite about multicollinearity we need to conduct a statistical inspection using correlation coefficients. There are no exact rules for establishing multicollinearity using pre-estimation methods. However, for our purposes, we will say that we can observe multicollinearity when the correlation coefficient is strong positive or strong negative and p<0.05. If we obtain these results then we can say that there is a problem with multicollinearity and will need to either drop the offending variable from our regression, find another way to measure the variable (a different indicator), or run multiple and separate regression models.

#### Statistical inspection (pre-estimation)

We use the pwcorr command to conduct a statistical pre-estimation inspection of multicollinearity. We include all IVs but exclude the DV.

. pwcorr Democracy Political\_Corruption Trade\_Freedom Education\_expenditure Unemployment, sig

Table

Description automatically generated

*Interpreting the results*. In our correlation matrix we can see one major offending correlation, namely between Democracy and Policy Corruption. The correlation coefficient is -0.7823 or strong negative and statistically significant with a p-value of <0.001. This is a major cause for concern. In this case, multicollinearity also makes sense logically: Democracy and Political corruption are likely part of the same larger phenomenon (e.g., they are explaining the same thing). As such, these may be redundant IVs. Hence, the case for multicollinearity between Democracy and Political corruption is strong and **we will omit** this variable from our regression analysis.

**Conclusion**: Based on the results, I will **omit** Political corruption from the regression analysis. You may ask: why are we omitting Political corruption and not Democracy? The reason is that our original intention was to examine the impact of Democracy on GDP. Democracy is our IV of interest and is the basis of our central hypothesis (H1).

### Multivariate regression analysis: estimation and interpretation.

As you can see, there is a lot of prep-work that we have to do before we run our multivariate regression analysis. But we are now ready. Here’s the Stata command. Recall that we have decided to omit Political Corruption because we have found some evidence of multicollinearity between Political corruption and Democracy.

. regress GDP Democracy Trade\_Freedom Education\_expenditure Unemployment

Table

Description automatically generated

*Interpreting the results*:

* Effects of Democracy on GDP are positive and statistically significant with a p-value <0.001. With each one unit increase in Democracy, we can expect GDP to increase by $7511.714.
* Results for Trade Freedom and Education Expenditure all have p-values >0.05. Hence, the results are not statistically significant and we therefore do not interpret the results for the coefficients (i.e., we do not interpret them because they are unreliable).
* The effects for Unemployment are negative and statistically significant with a p-value <0.01. For each one unit increase in Unemployment we can expect a decrease in GDP by $910.2451.
* The r-squared is 0.5015, which means that Democracy, Trade Freedom, Education expenditure and Unemployment explain 50.15% of variance in our DV. Recall that in our bivariate regression model that our r-squared was about 25%. As such, we have increased the explanatory power of our regression model by about 25%.
* It is important to note, however, that our “Number of Observations” has been reduced quite a bit. We can see that we only have 71 observations. This is the result of missing observations in our new variables. In real terms, this means that our sample includes far fewer countries than in our first bivariate analysis from the previous Stata lab. This may be cause for concern. We would need to investigate further and see if there is any systematic exclusion of countries that might make it difficult for us to test our hypotheses. For instance, it would be problematic if all of the missing data were from non-democracies. This would mean that variation on our main IV was reduced in a way that makes it difficult to talk about countries across the spectrum of the Democracy variable (from very low levels of democracy at 0, to very high levels at 10). It is up to the researcher to investigate these issues and to really ‘know’ the data.

## **Stata Exercises. 1**

**Task 1**

You want to test the following hypothesis about a country’s life expectancy (wef\_lifexp) and its air quality (epi\_ehair). Your IV is air quality, and your DV is life expectancy. Note that life expectancy is measured in years.

H1. The **higher** a country’s air quality, the **higher** its life expectancy.

To test this hypothesis, you need to also account for the following alternative explanations that we know from the scientific literature have an impact on our DV.

* Ethnical Fractionalistion (al\_ethnic)
* Social capital (bti\_sc)
* GDP (wef\_gdpc)

Complete the following sub-tasks:

1. Rename your variables using the following information:

a. wef\_lifexp a "Life\_Expectancy"

b. epi\_ehair as "Air\_Quality"

c. al\_ethnic as "Ethnical\_Fractionalistion"

d. bti\_sc as "Social\_capital"

e. wef\_gdpc as "GDP"

1. Write the Stata command to visually examine covariation between your IVs and your DV
   1. Are there any instances in matrix where there appears to be ‘no correlation’?
2. Write the Stata command to examine covariation between your IVs and your DV in terms of correlation coefficients and p-values.
   1. Are there any indications in your results where there appears to be ‘no correlation’?
3. Write the Stata command to visually inspect your IVs for issues of multicollinearity.
4. Write the Stata command to examine multicollinearity between your IVs in terms of correlation coefficients and p-values.
   1. Using our threshold of a correlation coefficient of 0.7 and -0.7 and p<0.5, are there any indications in your results where there appears multicollinearity?

**Task 2.**

Run a multivariate regression to test your hypothesis about the impact of Air quality on Life expectancy. Include the following IVs in your regression: Ethnical\_Fractionalistion, Social\_capital and GDP.

1. interpret the magnitude and statistical significance of the effects of all your IVs as well as the explanatory power of the model.
2. Do we find support for our hypothesis in the multivariate regression analysis?
3. Interpret the r-squared?

# Part 2. Extensions of the ‘regress’ command

There are many optional commands that you can use to better specify your regression model. One important option is the ‘if’ command. It can help you run your regression for specific sub-groups of your dataset. The basic Stata syntax is this.

. regress dv *iv1 iv2 iv3,* if var == n

We can add multiple ‘if’ modifiers either with | and & as well as combinations as seen in the table below.

Example:

. regress dv *iv1 iv2 ivn* if var == n | var == n

. regress dv *iv1 iv2 ivn* if var == n & var == n

Text

Description automatically generated

**Example**. I want to re-run my multivariate regression, as above, but this time only for countries that have ‘Emerging’ economies. To do this, I will first need to create a new variable that will allow me to make a distinction between levels of development per country. I will call this new variable ‘Levelofdevelopment’. Note that I have created categories in a less than scientific manner. This is only for demonstration purposes.

. recode GDP (0/1561 = 1 "poor")(1562/5873 = 2 "developing")(5873.1/19018 = 3 "Emerging")(19018.1/117207 = 4 "developed"), gen(Levelofdevelopment)

Here is what my new variable ‘Levelofdevelopment’ looks like, with value labels (left) and numeric labels (right)

|  |  |
| --- | --- |
| A screenshot of a cell phone  Description automatically generated | A screenshot of a cell phone  Description automatically generated |

With my new variable, I can now use the ‘if’ option.

. regress GDP Democracy Trade\_Freedom Education\_expenditure Unemployment if Levelofdevelopment == 3

Table

Description automatically generated

*Interpreting the results*: If we only look at so-called Emerging countries (Levelofdevelopment == 3), then the effects of Democracy are no longer statistically significant. We can see that the p-value for Democracy is >0.05. However, we can also see that our Number of Observations is 23. This is too low and likely means that the entire model is unreliable. One very general rule of thumb is that **30 observations** is the minimum for running a regression analysis.

**Example**. We want to re-run our regression, but this time including countries that we have labelled ‘Developed’ as well as ‘Emerging’.

. regress GDP Democracy Trade\_Freedom Education\_expenditure Unemployment if Levelofdevelopment == 3 | Levelofdevelopment == 4

Table

Description automatically generated

*Interpreting the results*: Including the two sub-categories, Developed and Emerging, has resulted in a higher Number of Observations (=51). We can now see that the effects of Democracy are once again statistically insignificant (p-value <0.05). We can also now interpret the regression coefficient. With each one unit increase in Democracy, we can expect an increase of $6222.055 of GDP. As with our earlier multivariate regression model, we do not see any statistically significant differences for Trade Freedom or Education expenditure. Unemployment, by contrast, is statistically significant with a p-value of <0.01. We can also say with each one unit increase in Unemployment that GDP decreases by $1353.1. Our r-squared suggests that our multivariate regression model explains 42.93% of variance in our DV.

# Part 3. Increment in R-squared: Semipartial correlations

As we already know, our r-squared statistic tells us how much variation in our DV is explained by the IVs in our model. In the previous example multivariate regression the r-squared was 0.4293 and hence we can say that the IVs in our regression model explain 42.93% of variance in our DV. This is a very useful statistic. However, we often want to know how each individual IV impacts the overall explanatory power of our model. To do this we need to calculate the **increments of the r-squared statistic**. The basic command for the increment of r-squared is this:

. pcorr dv *iv1 iv2 iv3 … ivn*

**Example.** We want to estimate the increments of r-squared for our analysis of countries’ GDP (as above).

pcorr GDP Democracy Trade\_Freedom Education\_expenditure Unemployment if Levelofdevelopment == 3 | Levelofdevelopment == 4

A picture containing text, receipt, screenshot

Description automatically generated

*Interpreting the Results*: We are only interested in the row with results for **Semipartial Corr.^2** (the area in the red box), which shows how much each variable contributes uniquely to the explanatory power of the model. The results are partial r-squares, and hence tell us about the explanatory power of each individual variable. Our results show that Democracy has the largest unique impact on the r-squared of 14.12%. Trade Freedom is 1.46%, Education Expenditure is 0.05% and Unemployment is 4.51%. The Semipartial Corr.^2 is a conservative estimate of the effect of each variable because it measures only how much the r-squared increases when that variable is entered into the regression model only after all the other variables are already in the model. You will notice that the sum of the Semipartial Corr.^2 is much lower than our actual r-squared. This is the case because some of our IVs are explaining the **same part** of our DV and our Semipartial Corr.^2 only estimates each IVs unique effect.

## **Stata Exercises. 2**

Task 1.

We are actually only interested in results for democratic countries. In this case, we will use a binary democracy measure (bmr\_demmis), where democracy = 1.

Interpret the magnitude and statistical significance of the effects of all your IVs as well as the explanatory power of the model.

1. Write the Stata command to execute the multivariate regression described above.
2. Interpret the magnitude and statistical significance of the effects of all your IVs as well as the explanatory power of the model.

Task 2

We now want to know the unique contribution of each IV on your DV (life expectancy). Calculate semipartial correlations for your IVs and answer these questions.

1. Which IV contributes most to explaining variation in your DV?
2. What percent of the DV does this variable explain?
3. Which IV contributes least to explaining variation in your DV and what percent is this?

# Part 4. Regression Diagnostics & Postestimation Functions: Gauss-Markov Conditions

It is so easy to fit a bivariate or multivariate regression model using modern statistical software that we tend to forget that there are several critical assumptions behind OLS that, if they don’t hold true, will lead to questionable or even erroneous results. These assumptions are referred to as Gauss-Markov Assumptions.

## Assumption 1. Normal Distribution of the DV

A key assumption of OLS is that our DV is based on continuous data and that it has a normal or Gaussian distribution. To examine whether or not our DV is normally distributed we need to do two things: (1) visually inspect the distribution as a histogram (hist var, freq normal); and (2) assess its skewness and kurtosis using the command *sum var, detail.*

### Visual Inspection

Visual inspection involves creating a histogram of our DV and comparing the distribution with the normal distribution.

. hist GDP, frequency normal



Interpreting the results:

* We can see that our data is not normally distributed. In fact, there is very large amount of data at the low end of the GDP scale. It has a very pronounced positive skew (see the table below for the different types of skewness we can have). We can also see that there might be some outliers on the far-right side of the x-axis.

Types of Skewness

A close up of a device

Description automatically generated

### Assessing skewness and kurtosis

After our visual inspection, we can get even more detail by looking at the detailed summary statistics of our DV. This brings us back to one of our very first commands: summarize.

. sum GDP, detail

![A screenshot of a cell phone

Description automatically generated]()

*Interpreting the results:*

* **Skewness**: Skewness is a measure of whether a distribution trails off in one direction or another. A normal distribution has a skewness of 0. Skewness greater than 0 = positiveness skewness; Skewness less than 0 = negative skewness.
* **Kurtosis**: Kurtosis measures the thickness of the tails of a distribution. When a distribution has thick tails, it will generally have too few observations in the middle of the distribution. Kurtosis can also be about having very thin tails, and then the problem is having too many observations in the middle of the distribution. In our case in the histogram above, the tail on the far left does seem to be thick. A normal distribution will have a kurtosis of 3.00. A value of less than 3.00 means that the tails are too think (hence, too flat in the middle); a value of greater than 3.00 means that the tails are too thin (hence, too peaked in the middle).

### Significance of Skewness and Kurtosis

We can now go one step further and see if our values for skewness (2.17) and kurtosis (8.011) are statistically significant. This will tell us definitively if we have a problem with normal distribution. In other words, if our values for skewness and kurtosis statistically significant (p-value <0.05) then our variable is not normally distributed.

The Stata commands is this:

. sktest DV

And here is our example:

. sktest GDP

![A screenshot of a cell phone

Description automatically generated]()

*Interpreting the results*: We can see that our joint p-value (Prob>chi2) is <0.001. Hence, the distribution is not normal. To report our results, we need to look at our table of summary statistics and this ‘sktest’ output. The conventional method for reporting these results is this:

Skewness = 2.17, p<0.001 and kurtosis = 8.011, p<0.001. Both tested jointly have a p-value <0.001.

Hence, we can say that our **distribution is not normal**.

### Dealing with Non-Normal Distribution of your DV

What to do if you encounter a non-normally distributed DV? There are a few options.

1. One (radical) solution is to simply not use the offending variable. I.e., if the DV is very problematic, then maybe you should search for another indicator. The Quality of Government database has numerous different indicators for GDP. However, this is not always an option.
2. Barring option 1, you might want to think about identifying and then dropping outliers. We will learn how to do this below in the section on Outliers. Be careful not to reduce the N too much.
3. Finally, there is a third option which entails ‘transforming’ your variable to normalize its distribution. You can do this with your DV or any variable. Doing so, however, will complicate the interpretation of your regression results. There are multiple ways to normalize the distribution of your DV. One common way is logarithmic transformation or log transformation.

### Log Transformation and Normalizing Distribution

There are a few ways to transform variables to normalize their distribution. We will learn just one, the **log-transformation**. In this transformation, Stata uses the natural logarithm of the values of your variable to normalize the distribution. The Stata command for this is:

. gen newvarname = ln(var)

Example. We clearly have a problem with the distribution of our DV (GDP). We want to try to normalise the distribution using the log-transformation technique.

. gen log\_GDP = ln(GDP)

We can see our results by comparing histograms of GDP & log\_GDP

|  |  |
| --- | --- |
|  |  |
| Original Distribution | Distribution after Log Transformation |

**Example 1**. We want to now re-run our multivariate regression with our more normally distributed DV. Note that I am still using the renamed variables from an earlier task.

. regress log\_GDP Democracy Trade\_Freedom Ethnic\_Fractionalistion Education\_expenditure Unemployment

Table

Description automatically generated

*Interpreting the results*. For the most part, we interpret these results the way we normally do for OLS. The only difference is in the DV. Remember our DV was log-transformed to normalise distribution. This makes interpretation a little bit more difficult. One solution is to present the results the way they are: for each one unit increase in Democracy, we can expect an increase of about **0.3699773 of the logarithm of GDP**. But what does this mean in real-world terms? If we wanted something a bit more understandable, we could reverse the log transformation by exponentiating the coefficient. The formula for this is: exponentiate the coefficient, subtract one from this number, and then multiply by 100. **This will tell us the percent increase in the DV with each one increase in the DV**.

This is easy in Excel where you can use the = EXP(value) function. Here are the steps using excel.

Step 1: = EXP(0.27762)

= 1.44770175

Step 2: 1.44770175-1

= 0.44770175

Step 3: 0.44770175\*100

= 44.770175

Hence, we can say that with each one unit increase in Democracy, we can expect a %44.770175 increase in GDP.

**\*\*Pro-tip**: Be careful when performing a log-transformation if any of the values in your distribution are ‘0’. You cannot transform these values and the results of your log-transformation will be very unreliable. To eliminate zeros, add 1 to all values to your indictor and then perform the log-transformation.

## Assumption 2. Normal Distribution of Residuals

A second important assumption in OLS is that the residuals (error terms) are normally distributed. This means that we are as likely to overestimate a value as we are to underestimate it.

Examining the distribution of residuals is a two-step process in Stata. You need to first run your regression model and then run a post-estimation test. The Stata commands look like this.

. regress dv iv iv iv iv

. predict newvar, residual

. summarize newvar, detail

Example: We want to test if our residuals are normally distributed in our main regression example used above. Let’s also use our original GDP variable, not the log-transformed version.

. regress GDP Democracy Trade\_Freedom Ethnic\_Fractionalistion Education\_expenditure Unemployment

. predict res, residual

. sum res, detail

A picture containing text, receipt, screenshot

Description automatically generated

*Interpreting the results*: As with our examination above, we see that skewness is greater than 0 (positive skewness) and kurtosis is greater than 3 (tails are too thin). This is consistent with our examination of frequency distribution, above.

**\*\*Pro-tip**: You will notice that in this post-estimation calculation that we generated a new variable, which we called ‘res’. If you were to conduct the same post-estimation technique for residuals, be sure to note that you will need to generate a new variable with a name that is different from ‘res’. You could call it res1, or anything you want. You could also drop ‘res’.

### Significance of Skewness and Kurtosis for Residuals

As with our tests for our distribution, we can run a further post-estimation test to see if our values for skewness (1.135824) and kurtosis (4.074947) are statistically significant. This will tell us definitively if we have a problem with normal distribution. The Stata commands is this:

. sktest var

And here is our example:

. sktest res

Table

Description automatically generated

*Interpreting the results*: To report our results, we need to look at our table of summary statistics and this sktest output. Using the conventional method, we can say that

Skewness = 1.135824, p<0.001 and kurtosis = 4.074947, p>0.05. Both tested jointly have a p-value <0.01.

Hence, our **distribution is not normal**.

### Dealing with Non-Normal Distribution of Residuals: Robust Regression

First, we should be a bit cautious in analyzing the results of our sktest for the distribution of our residuals. Larger N datasets tend to be sensitive to showing statistically positive results. However, we should be more concerned when we are using lower-N data (under 300 observations). When you do encounter a problem with the distribution of your residuals, one solution is to run a **robust regression**. Robust regression treats the standard errors (residuals) in a unique way. The Stata command for robust regression is this:

. regress dv iv iv iv ivn, vce(robust)

**Example 1:** We now know that we have an issue with non-normally distributed residuals in our multivariate regression model. In this case, we use robust regression.

. regress GDP Democracy Trade\_Freedom Ethnic\_Fractionalistion Education\_expenditure Unemployment, vce(robust)

Table

Description automatically generated

*Interpreting the results*: The main thing to note is the difference between this robust regression and our original regression. You will notice that, on balance, the reported p-values (statistical significance) are all a bit lower.

## Assumption 3. Multicollinearity (post-estimation)

As discussed above, when we conduct a multivariate regression analysis it is important to test for multicollinearity. This is about the potential of a high level of correlation between our various IVs. IVs or independent variables need to be independent from each other. A high level of correlation between IVs can give an exaggerated picture of the impact of each variable on the DV as well as the overall explanatory power of our model.

We have already learned one way to examine multicollinearity, namely the through visual inspection and the pwcorr command. However, there is an important post-estimation technique that can give us a very clear picture of multicollinearity. This is a post-estimation test for Variance Inflation Factors (VIF) for each IV.

As with all post-estimation tests, you first run your regression and then run the test, which in this case is just the command ‘vif’ (which stands for Variance Inflation Factors).

. regress GDP Democracy Trade\_Freedom Ethnic\_Fractionalistion Education\_expenditure Unemployment

. vif

Table

Description automatically generated

*Interpreting the results*: The general rule of thumb is that **VIFs >10 indicate multicollinearity**. As we can see, all of our VIF values are <10. As such, we do not have any issues with multicollinearity. If we did find a VIF greater than 10, we would drop the offending variable from our regression analysis and then, following the new regression, estimate and interpret the new VIF score.

### Dealing with multicollinearity

As noted above, one good way of dealing with multicollinearity is to drop the offending variable from the regression model. You could see if you can replace that offending variable with another (different) proxy for the concept you want to measure. Alternatively, you could run a series of regression models where the two variables that have a high level of multicollinearity are analysed separately.

## **Stata Exercises. 3**

**Task 1.**

You need to run diagnostic tests for all of variables used to test your hypothesis about life expectancy and air quality (regress Life\_Expectancy Air\_Quality Ethnic\_Fractionalistion Social\_Capital GDP).

Use histograms with a normal curve to visually inspect the skewness of the distribution of each variable. Which variables can be described as:

1. Positively skewed
2. Negatively skewed
3. Normal (symmetrical) distribution

**Task 2.**

Moving on from our visual inspection, we now want to say something a bit more substantive about the distribution for our variables. Complete the following sub-tasks:

1. What is value for skewness for Ethnical fractionalization? Report its value and describe its skewness (negative, positive, symmetrical).
2. What is the value reported for kurtosis for ethnical fractionalization? Describe kurtosis referring to the ‘tails’.
3. Use a ‘sktest’ and report results for the variable life expectancy. Write the results using the conventional method.
4. Is our distribution normal?

**Task 3.**

We now need to assess assumptions about the normal distribution of residuals for our multivariate regression model. Complete the following tasks:

1. What is the value reported for skewness for our residuals? Describe it.
2. What is the value reported for kurtosis for our residuals? Describe it.
3. Using the conventional method, report the results of a ‘sktest’ of the residuals.
4. Do we have normal distribution?

**Task 4.**

Run your multivariate regression model again, but this time also calculate variance inflation factors.

1. Which variables (if any) pose problems of multicollinearity?

## Stata Advanced Tasks

**Task 1**.

You are interested in examining the determinants of sustainability using the ‘Sustainability’ index (bti\_su). The scientific literature tells us education should have a positive impact on sustainability. You measure education using the 'Education\_Provision' Index (gov\_ixeducindex). This is you IV of interest. However, you also want to include other variables in your multivariate analysis. This includes: (1) the 'Transparency Index' (diat\_ti), (2) 'Oil rents' as % of GDP (wdi\_oilrent), and (3) 'Oil production' (ross\_oil\_prod). As a first step, rename all of the variables using the new variables names indicated above (e.g., rename bti\_su Sustainability\_Index)

Step 1. Run a multivariate regression using the variables listed above. Complete the following:

1. Interpret each coefficient and note its statistical significance.
2. Interpret the r-squared.
3. How much variation is NOT explained in your model?

Step 2. Use pre-estimation techniques (visual inspection and statistical inspection) to examine the extent to which your regression model might be subject to multicollinearity. Can you identify any potential issues with multicollinearity?

Step 3. Use a post-estimation command to examine the extent to which your model might be subject to multicollinearity. Is there any evidence for multicollinearity?

Step 4. Use the command for calculating semi-partial correlations for increments of r-squared on the same multivariate regression model.

1. Which IV contributes the most explanatory power to the model? What is the %.
2. Which IV contributes the least explanatory power to the model? What is the %.

Step 5. Visually inspect the distribution of all your variables (DV and IVs) in the multivariate regression model conducted in step 1. Describe each variable in terms of skewness.

Step 6. Assess the distribution of you DV?

1. Using the conventional method, report on skewness and kurtosis of your DV as well as p-valued for both skewness and kurtosis of your DV.
2. Is the distribution of our DV (Sustainability\_Index) normally distributed?

Step 7. Assess the distribution of residuals in your model (regress Sustainability\_Index Education\_Provision Transparency\_Index Oil\_Rents Oil\_Production).

1. Using the conventional method, report on skewness and kurtosis of your residuals as well as p-valued for both skewness and kurtosis.

Step 8. We are now only interested in countries that have a minimum of 137 Animals in them (bi\_a\_total). Rerun the the same regression model (regress Sustainability\_Index Education\_Provision Transparency\_Index Oil\_Rents Oil\_Production) but only for countries with 137 or more Animals. Report on the coefficients and statistical significance of each IV and interpret the r-squared.

## Appendix. How to Present the Results of a Multivariate Regression Analysis with multiple models

Unfortunately, Stata does not produce very nice tables. Part of your job is to use the output that Stata generates and produce your own table in Word. **Yes, this means cutting and pasting from what you see in Stata to a Word doc. DO NOT copy the raw Stata output.**

|  |  |  |
| --- | --- | --- |
| **Table 2.** OLS Regression Analysis of the Effects of Democracy on GDP | | |
|  | (Model 1) | (Model 2) |
| Democracy | 4887.5\*\*\* | 453.5 |
|  | (766.5) | (1038.9) |
| Corruption |  | -34279.6\*\*\* |
|  |  | (7157.2) |
| Trade |  | 422.4\*\* |
|  |  | (154.0) |
|  |  |  |
| Constant | -13669.0\*\* | -5178.4 |
|  | (4782.1) | (13398.1) |
| N | 144 | 137 |
| R2 | 0.217 | 0.437 |

Notes: \* p < 0.05, \*\* p < 0.01, \*\*\* p < 0.001