# 

**Linear Regression Diagnostics - Stata**

Today we’re talking about linear regression diagnostics, and all you need to know to conduct these assumption checks in Stata.

# Linear Regression Diagnostics

## Assumptions for Linear Regression

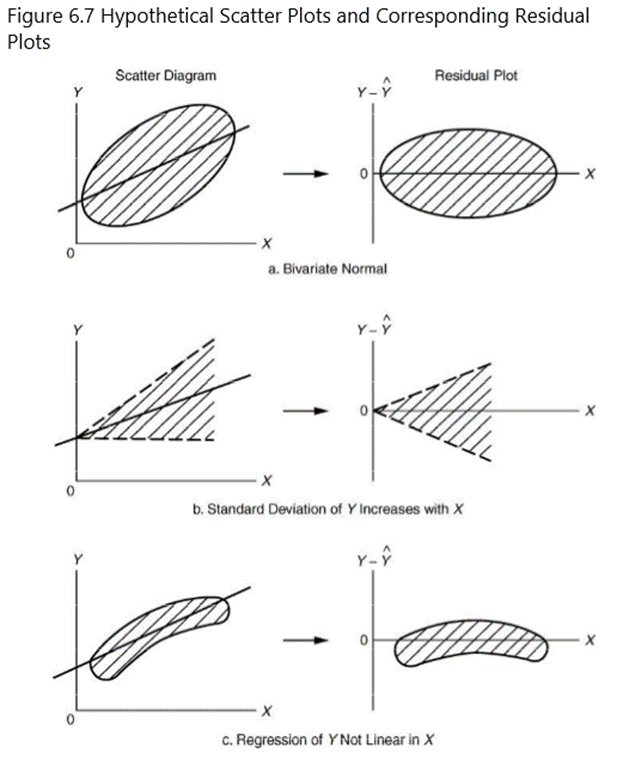
We will consider 4 assumptions to determine if multiple linear regression is a good fit for the structure of our data. Diagnositc analyses are used to determine whether or not these assumptions are valid:

1. Linearity
2. Normality
3. Variance of Y does not depend on the level of X (homoscedasticity)
4. Independent Observations

## Linearity

If you cannot assume linearity, you will have a poor fit of the regression line to the data, estimates of your model will be inaccurate, and you will make inaccurate inferences for your hypothesis. Basically, your data wont fit your model well and its meaningless.

In simple linear regression, plot Y versus X with a fitted regression line and plot the residuals versus each of the independent variables. Your model has only one set of residuals since they incorporate all of the independent variables.



### In Stata

1. Run your model

regress outcomevariable independentvariable

1. Assess the residuals

rvpplot independentvariable

1. Assess the plot, you want a fairly normal scatter of points. If you have a fan shape, bean or linear grouping, you will need to transform your data.

### **To correct for non-linearity**

Transform either your outcome variable Y, or your independent variable X, or both. Make sure to include the transformed variable and the original form of the variable in your model. For example, if you squared height, also include height in your regress command

*Original Command:* **regress fev age height**

*Transformed Command:* **regress fev age height height\*height**

## **Normality**

We assume that the residuals have a normal distribution with mean 0 and variance σ2. If the residuals are not normal, it won’t affect the estimates of your model, but it can dramatically affect their statistical significance. When you have a small sample, this may be more of a problem.

### **In Stata**

1. Run your model

regress outcomevariable independentvariable

1. Assess normality of the residuals

predict bresid, resid

histogram bresid, bin(10) fraction normal

1. Assess the histogram, you are looking for a normal distribution.

### **To correct for non-normality**

Transform either your outcome variable Y, or your independent variable X, or both. Make sure to include the transformed variable and the original form of the variable in your model. For example, if you squared height, also include height in your regress command

*Original Command:* **regress fev age height**

*Transformed Command:* **regress fev age height height\*height**

## **Variance of Y does not depend on the level of X (homoscedasticity)**

We assume the residuals have a constant variance. If we don’t have constant variance, then we could explore other options for a better fit of the data.

### **In Stata**

1. Run your model

regress outcomevariable independentvariable

1. Assess constant variance

rvpplot independentvariable

estat hettest

1. Assess the residual plots for a random pattern. Use the Cook-Weisberg test for heteroscedasticity to assess constant variance. **We want a p-value >0.05.**

If the Cook-Weisberg test is significant, then we **do not** have constant variance.

### **To correct for no constant variance**

Find a transformation that stabilizes the variance. Use an alternative procedure of weighted least squares which weights observations according to their relative importance or degree of variability.

## 

## **Independent Observations**

We should assume that observations are independent of each other. This is a decision you make based off of the study design of the study. If a study is taking multiple observations on the same person, as in a prospective study, or longitudinal analysis, reconsider if your data is independent or not.

## **Quick Review**

Remember, regression diagnostics are something you should do before you interpret your results. All of the four assumptions need to be met in order to make good inferences about your data. Influential observations or outliers in your data can significantly impact your regression diagnostics. Instead of transforming your data, you may consider assessing influential observations. Remember that transforming your independent variables can be difficult to interpret: what does it mean if someone’s height2 increases their chance of respiratory problems? Think critically about how you will interpret the results and how meaningful they are when considering transformations.