Linear Regression Assumptions

1. Independence of observations
2. No Hidden or Missing Variables
3. Linear relationship
4. Normality of the residuals
5. No or little Multicollinearity
6. Homoscedasticity
7. All independent variables are uncorrelated with the error term
8. Observations of the error term are uncorrelated with each other

Linear Regression Assumption 1 — Independence of observations

The first assumption of linear regression is the **independence of observations**. Independence means that there is no relation between the different examples. This is not something that can be deduced by looking at the data: **the data collection** process is more likely to give an answer to this.

A clear case of **dependent observations** *(which we don't want!)* can occur when you are using **time series**. Imagine a daily data measurement of a certain value. In this case, the value of today is closer to the value of yesterday than the value of a long time ago.

A clear case of **independent observations** *(which we do want!)* are **experimental studies** in which participants are randomly assigned to treatment groups. In this case, it is the fact that **assignment is random and forced** that makes sure that there are no hidden relationships between observations.

Linear Regression Assumption 2 — No Hidden or Missing Variables

The second assumption of the linear regression model is that you have used **all relevant explanatory variables** in your model. If you do not do this, you end up with a wrong model, as the model will try to assign coefficients to the variables that do exist in your data set. This is often referred to as **misspecification** of a model.

If adding a variable to the model would make a whole lot of difference, it means that the model is incorrect and useless without it. The only thing you can do in this case is to **get back to your data collection** to find the necessary data.

Linear Regression Assumption 3 — Linear relationship

The third assumption of Linear Regression is that relations between the independent and dependent variables must be linear.

Although this assumption is not always cited in the literature, it is logical and important to check for it. After all, if your relationships are not linear, you should not use a linear model, but rather a **non-linear model** of which plenty exist.

You can check for linear relationships easily by making a scatter plot for each independent variable with the dependent variable. You can use the following R and Python code to do so.

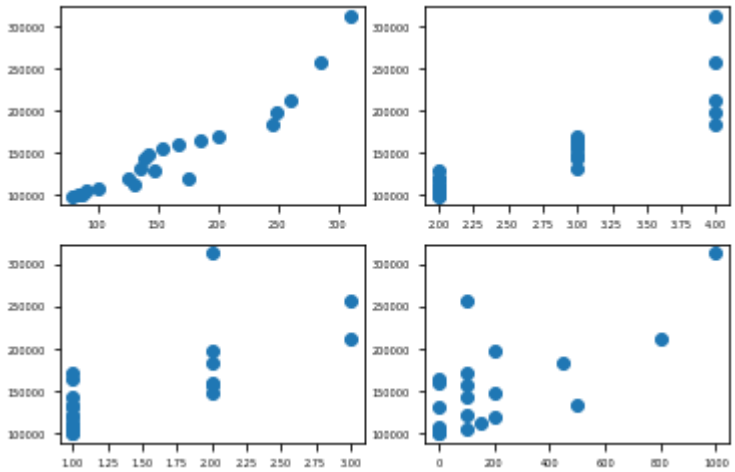
Checking for linear relationships with Python

Assumptions of Linear Regression — Linear Relationships — Python

Checking for linear relationships with R

Assumptions of Linear Regression — Linear Relationships — R

Although there are many other ways to do scatterplots, this approach is simple and good enough for checking assumptions.



The scatterplots resulting from the Python code

To find out whether the 1-on-1 relationships are linear, you need to judge whether the data points are more or less on or around a straight line. Clear antipatterns are when you see curves, parabolas, exponentials, or basically any shape that is recognizable as not a straight line.

The plots do not show perfect straight lines, but this is not a problem. There also isn't any clear non-linear pattern and a linear model may work well on this.

Linear Regression Assumption 4 — Normality of the residuals

The fourth assumption of Linear Regression is that the residuals should follow a normal distribution. Once you obtain the residuals from your model, this is relatively easy to test using either a histogram or a QQ Plot.

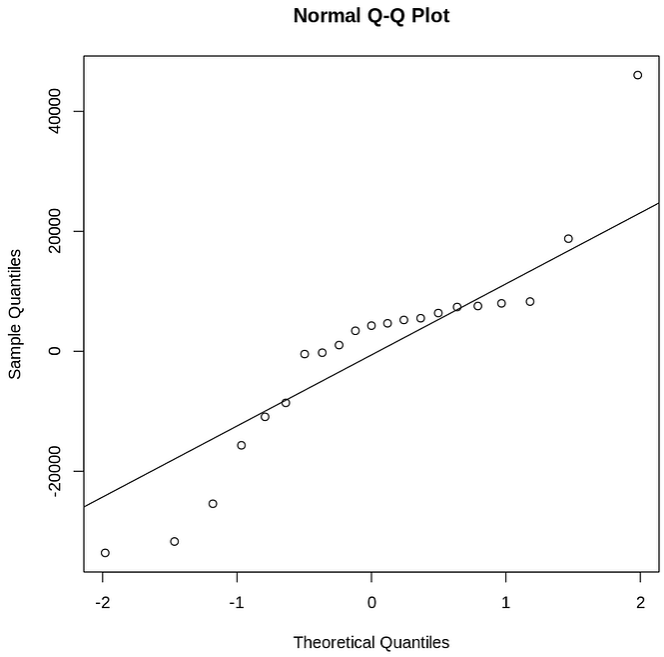
QQ Plots are a bit harder to read but they are more precise to interpret so let's see how to make QQ Plots of the residuals using R and Python.

Checking for normality of the residuals using Python:

Assumptions of Linear Regression — Normal distribution in the residuals — Python

Checking for normality of the residuals using R:

Assumptions of Linear Regression — Normal distribution in the residuals — R



The QQ Plot resulting from the R code

What you need to look at in QQ Plots is whether the points are on the straight line going from bottom left to top right. When deviations occur, they are often located at the lower or higher end of the line, whereas deviations in the middle are less likely.

If you see any type of an S form, an exponential curve, or another shape than a straight line, this means you have a problem: your model is probably not correctly specified. Probably you are missing some variables, or maybe your relationships are not actually linear! You may want to try out nonlinear

models or other specifications of the linear model (using different variables or different preparation of the variables).

In the current example there is clearly an inverted S form meaning that something is probably wrong with the model.

Linear Regression Assumption 5 — No or little Multicollinearity

The fifth assumption of linear regression is that there is no or little multicollinearity. Multicollinearity is the phenomenon when a number of the explanatory variables are strongly correlated.

So why do we want to have strong correlations between each independent variable and the dependent variable, but no correlation between independent variables? The reason is that if two independent variables are correlated, they explain the same information. The model will not be able to know which of the two variables is actually responsible for a change in the dependent variable.

You can test for multicollinearity problems using the Variance Inflation Factor, or VIF in short. The VIF indicates for an independent variable how much it is correlated to the other independent variables. You can compute VIF in R and Python with the following code.

Checking for multicollinearity using R:

Assumptions of Linear Regression — Checking for Multicollinearity using VIF — R

Checking for multicollinearity using Python:

Assumptions of Linear Regression — Checking for Multicollinearity using VIF — Python

```
{'BathRooms': 2.8893491708390076,
 'BedRooms': 6.228346746623228,
 'SquareMeterGarden': 1.7855234347718028,
 'SquareMeterHouse': 5.187199379128779}
```

The resulting VIFs from Python

VIF starts from 1 and has no upper limit. A VIF of 1 is the best you can have as this indicates that there is no multicollinearity for this variable. A VIF of higher than 5 or 10 indicates that there is a problem with the independent variables in your model.

In the current model, there is definitely a problem with the variables BathRooms, BedRooms and SquareMeterHouse. They seem very correlated between each other and it would be necessary to inspect which of those variables are actually needed to explain SellPrice.

Linear Regression Assumption 6 — Homoscedasticity

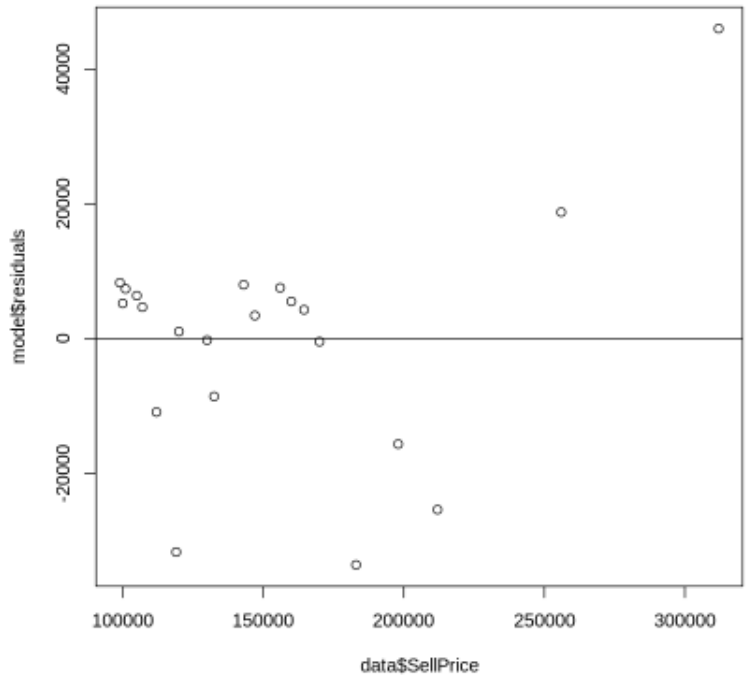
The sixth assumption of linear regression is homoscedasticity. Homoscedasticity in a model means that the error is constant along the values of the dependent variable. The best way for checking homoscedasticity is to make a scatterplot with the residuals against the dependent variable. You can do this with the following R and Python code.

Checking for homoscedasticity using R:

Assumptions of Linear Regression — Homoscedasticity — R

Checking for homoscedasticity using Python:

Assumptions of Linear Regression — Homoscedasticity — Python



Assumptions of Linear Regression — Homoscedasticity plot

Homoscedasticity means a constant error, you are looking for a constant deviation of the points from the zero-line. In the current case, you clearly see two outliers on the top right. In the rest of the points, you also see more points to the top and less to the bottom. This clearly does not look like a constant variance around the zero-line.

If you violate homoscedasticity, this means you have heteroscedasticity. You may want to do some work on your input data: maybe you have some variables to add or remove. Another solution is to do transformations, like applying a logistic or square root transformation to the dependent variable.

If this doesn't change anything, you can also switch to the **weighted least squares model.** Weighted least squares is a model that c*an deal with unconstant variances* and heteroscedasticity is therefore not a problem.

Linear Regression Assumption 7 — All independent variables are uncorrelated with the error term

The seventh diagnostical check of your linear regression model serves to check whether there is correlation between any of the independent variables and the error term. If this happens, it is likely that you have a case of a misspecified model. You may have forgotten an important explanatory variable.

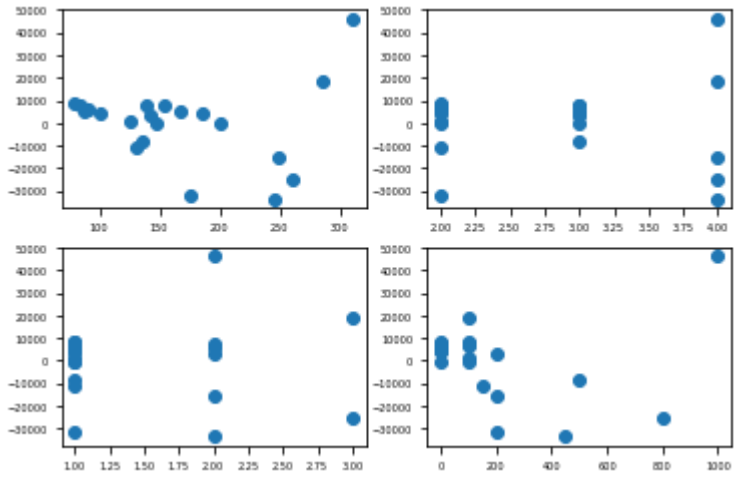
You can obtain the scatter plots using the following R and Python code:

Checking for correlation between the IVs and residuals in Python:

Assumptions of Linear Regression — No correlation between independent variables and residuals — Python

Checking for correlation between the IVs and residuals in R:

Assumptions of Linear Regression — No correlation between independent variables and residuals — R



Assumptions of Linear Regression — No correlation between independent variables and residuals

In those scatter plots, we do not see any clear correlation. The right bottom plot may be a disputable case, yet it is not very clear and convincing of a problem neither.

Linear Regression Assumption 8 — Observations of the error term are uncorrelated with each other

The last model diagnostic that we're going to look at is whether there is a correlation inside the observations of the error term. If this happens, you definitely violate assumption 1: the observations are not drawn randomly.

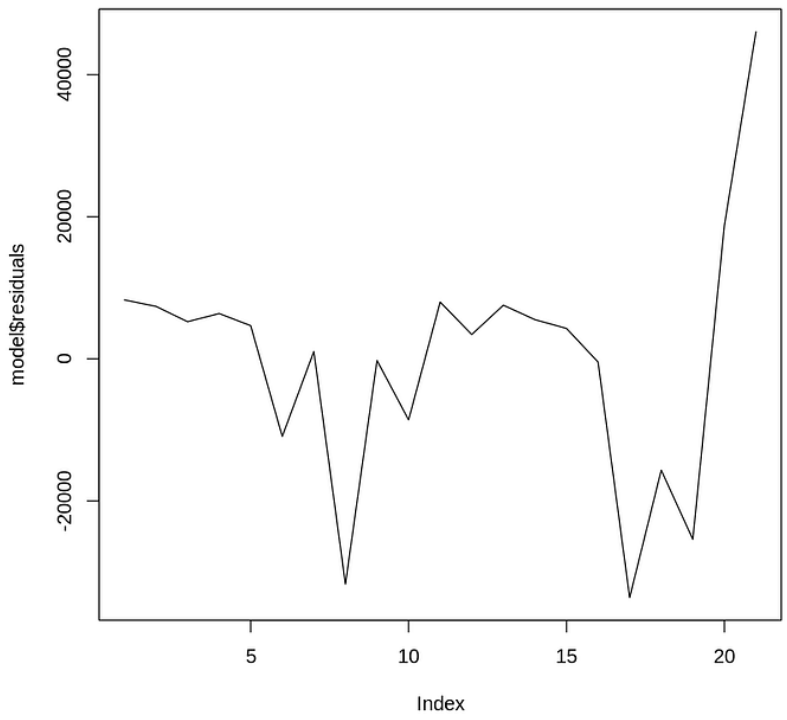
You can do a visual check by plotting the residuals against the order of the residuals. The following code snippets allow you to do this:

Checking for residual autocorrelation in Python:

Assumptions of Linear Regression — Autocorrelation in the residuals — Python

Checking for residual autocorrelation in R:

Assumptions of Linear Regression — Autocorrelation in the residuals — R



Assumptions of Linear Regression — No autocorrelation in the residuals

If a pattern occurs, it is likely that you have a case of a misspecified model. You may have forgotten an important explanatory variable. Or you might be better of using another family of models. If you have autocorrelation you may want to look into time series models like Auto-Regression Moving Average or ARMA.