Assumptions for linear regression:

* **linearity and additivity** of the relationship between dependent and independent variables:
  + The expected value of dependent variable is a straight-line function of each independent variable, holding the others fixed.
  + The slope of that line does not depend on the values of the other variables.
  + The effects of different independent variables on the expected value of the dependent variable are additive.
* **statistical independence** of the errors (in particular, no correlation between consecutive errors in the case of time series data)
* **homoscedasticity** (constant variance) of the errors
  + versus time (in the case of time series data)
  + versus the predictions
  + versus any independent variable
* **normality** of the error distribution.

<http://people.duke.edu/~rnau/testing.htm>

UsingR book page 270.

Assumptions for multiple linear regression:

* Linear relationship
* Multivariate normality
* No or little multicollinearity
* No auto-correlation
* Homoscedasticity

! It is also important to check for outliers since multiple linear regression is sensitive to outlier effects.

Pairs plot?

**Residuals vs Fitted** shows whether there is a linear relation between the independent and dependent variables,

<http://i.stack.imgur.com/RU17l.png>

first is linear, second is linear but heteroscedastic, third is not linear.

**Scale-location** also shows whether the data is homo or heteroscedastic, a flat curve means homo, a sloped line means heteroscedastic.

<http://i.stack.imgur.com/nPHfL.png>

**Q-Q Plot** shows linearity of the data, when the points are on a line the data is linear, otherwise there may be no linear relation.  
<http://stats.stackexchange.com/questions/58141/interpreting-plot-lm>

**Cooks distance**

The lower right plot shows the standardized residuals against leverage. Note that the standardized residuals are centered around zero and reach 2-3 standard deviations away from zero, and symmetrically so about zero, as would be expected for a normal distribution. Leverage is a measure of how much each data point influences the regression. Because the regression must pass through the centroid, points that lie far from the centroid have greater leverage, and their leverage increases if there are fewer points nearby. As a result, leverage reflects both the distance from the centroid and the isolation of a point. The plot also contours values of Cook’s distance, which measures how much the regression would change if a point was deleted. Cook’s distance is increased by leverage and by large residuals: a point far from the centroid with a large residual can severely distort the regression. On this plot, you want to see that the red smoothed line stays close to the horizontal gray dashed line and that no points have a large Cook’s distance (i.e, >0.5). Both are true here.

Script:

#generate linear model plots

par(mfrow=c(2,2))

n = 100

b1 = 2.5

sd = 3

x1 = 1:n

eps = rnorm(x1,0,sd)

y = b1 \* x1 + eps

plot(lm(y~x1))

#generate homo/heteroscedastic plot

pl.het = rnorm(x1, 0, sd = seq(1,100,by=(0.125\*sd)))

pl.hom = rnorm(x1, 0, sd)

y = b1 \* x1 + pl.het

plot(b1\*x1+pl.het,x1)

plot(b1\*x1+pl.hom,x1)

I am going to explain which assumptions the linear regression analysis makes about the data, and how to examine if your data fits these assumptions by looking at some plots.

First of all, the data has to have a linear relation between the dependent and the independent variable, if this is not the case another method of analysis probably fits the model better.

Secondly the data needs to be homoscedastic (as opposed to heteroscedastic), which means that the random variables in the data have the same variance. I have an image to display more clearly what this entails. (describe the image).

And finally the error needs to be normally distributed.

When the R-function plot() is used on a linear model, R gives four plots (displayed here on the slide) which can be used to check the validity of the assumptions.

On the top left there is the residuals vs fitted plot, which displays whether there is a linear relation between the dependent and independent variables, this holds for multiple linear regression too. If the data is symmetrical around the 0.

On the top right is the normal Q-Q plot which also displays linearity of the data. When the data fits on the line the data is linear and otherwise it isn’t.

On the bottom left is a scale-location plot. On this plot you can see whether the data is hetero or homoscedastic. When the red line is flat this means the data is homoscedastic, when it is sloped the data is heteroscedastic.

The bottom right is the residuals vs leverage plot, which displays whether there are data points that heavily influence the regression line, as long as all the values on this plot are small (<0.5) this is ok, else there might be some outliers skewing the results.