# Linear Model Assumptions

Sharon G Nielsen
Sharon Nielsen Statistical Consulting and Training
sharon@snstats.com.au

## What is a residual?

Before beginning to think about assumptions, it is important to understand what a residual is and how it is calculated. In the context of simple linear regression, the residual for an observation is the perpendicular distance between the regression line and the observed value. In Figure 1 below we picture a regression line and two of the residuals (as indicated by the red lines). If the observed value is above the regression line, the residual will be positive. If the observed value is below the regression line, the residual will be negative.

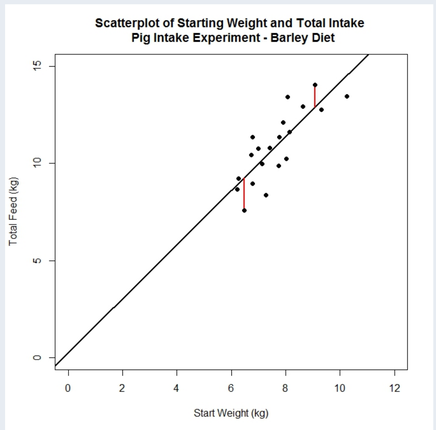



Figure 1: Simple linear regression - with two residuals shown in red

A residual is calculated for each observation by subtracting the observed value from the "fitted value" (the value on the regression line). It is these residual values that we will work with to determine if the model assumptions have been met or not.

## Simple Linear Regression

The assumption for simple linear regression involve the distribution of the residuals. The residuals need to:

1. have a mean of zero
2. be normally distributed
3. have a constant variance
4. be independent

## Checking the model assumptions

1. The residuals will have a mean of zero if the model includes the intercept and the method of least squares regression is employed.
2. The residuals need to be checked for normality. This can be done graphically or using a hypothesis test.
3. A graph is used to determine if the residuals have constant variance.
4. The residuals need to be independent of each other. This can be determined by examining the design of the experiment or study.

## What happens if the model assumptions are not checked?

With regression, the calculation of the regression parameters is not affected if any of the assumptions are violated. However, inferences made when the model assumptions are not met can lead to erroneous conclusions. For example, the conclusion might be made that there is a statistically significant linear relationship between the variables, when in fact there is not. Therefore, always check the model assumptions.

## Residual Plots

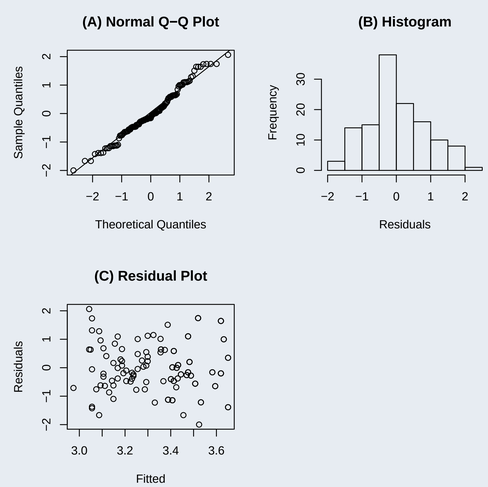



Figure 2: Residual plots: (A) QQ plot (B) Histogram (C) Residual vs fitted plot

## The residuals have a mean of zero

As explained earlier the residuals can be positive or negative. If the intercept is not fitted in the simple linear regression model it is possible that the residuals will not add to zero, therefore the mean will not be zero. If the intercept is fitted in the model and least squares regression is used, as is usually the case when fitting a simple linear regression model, the residuals will have a mean of zero.

## The residuals are normally distributed

It is good to look at your residuals graphically and to use a hypothesis test to assess the normality of the residuals. Graph the residuals using a normal probability plot or QQ plot and also graph the residuals in a histogram (Figure 2 - (A) & (B)). The residuals need to be approximately normal for this assumption to have been met.
The points in the QQ plot should follow the theoretical (diagonal) line if the residuals are normally distributed (see Figure 4 - (A)). The histogram should follow (at least approximately) a normal distribution. A Shapio-Wilks test for normality can be used to determine whether the residuals are normally distributed. The hypotheses for these tests would be:

$H_0$ : The residuals are normally distributed
$H_a$ : The residuals are not normally distributed

## Constant variance of the residuals

This is the most important assumption of all the assumptions. It can only be checked by inspecting the residual versus fitted plot, see section (C) in Figure 2. If your residual versus fitted plot looks something like that shown in Figure 3 you will need to fit a different model to the data. As you look across the graph from left to right the size of the residuals is increasing. The graph is said to be "fan" shaped or you can see a "v" on it's side. If you have a residual plot that shows heterogeneity as in Figure 3 making inferences based on the results from the analysis may be erroneous and should not be used. Either a weighted least squares analysis model or data transformation can be tried to remedy this problem.

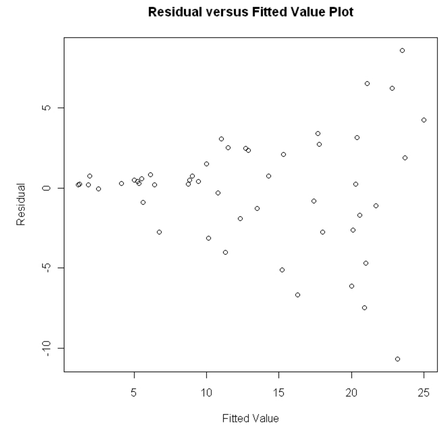



Figure 3: Residual vs fitted plot: showing heterogeneity

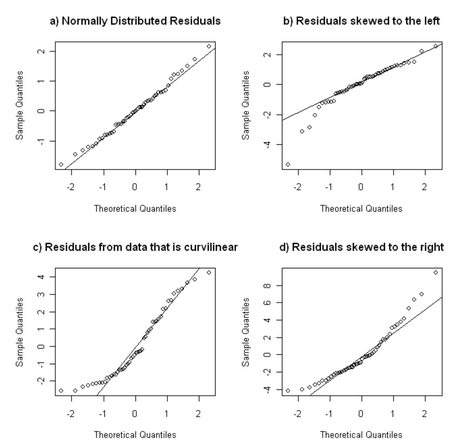



Figure 4: QQ plots departures from normality: (A) Normal Distribution (B) Left Skew (C) Curvilinear (D) Right Skew

# Linear Model Assumptions

Sharon G Nielsen
Sharon Nielsen Statistical Consulting and Training
sharon@snstats.com.au

## Analysis of Variance - ANOVA

Analysis of variance is used when the response variable is is continuous and the predictor variable/s are categorical. Usually the primary question of interest is to determine if there are any differences in the responses among the treatments. The best way to visualise the data at the start of the analysis is through a comparison boxplot, Figure 5.

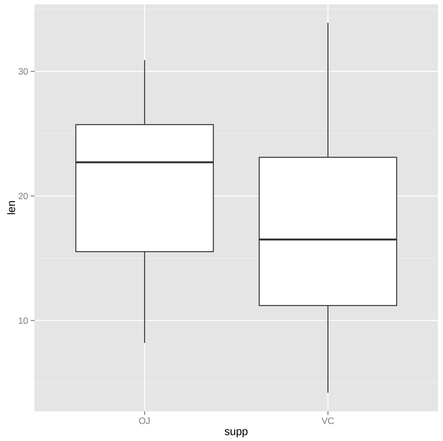



Figure 5: Comparison boxplots - assess the median (middle) line for differences in the response among the treatments and also the distance or spread of each of the boxlots

We can see from Figure 5 that the medians are different, the left-hand boxplot has a higher median than the right. Also, the spread of the right-hand boxplot is greater than that of the left.

## ANOVA - model assumptions

As, ANOVA and regression are related through linear models (that the mathematics behind the two analyses is the same) it is not surprising to find that the assumptions for ANOVA include all of the the assumptions for simple linear regression, plus one additional one.

1. the residuals need to have a mean of zero
2. the residuals need to be normally distributed
3. the residuals need to have a constant variance
4. the residuals need to be independent
5. the variance of each of the factor levels need to be equal

### The data does not have to be normally distributed!

Notice that the assumptions around these models are associated with the residuals **NOT** the data itself. If you think about the situation where you are expecting the treatments you have used in the experiment to have an effect on the results. You could see large differences in the response due to the different treatments. In this situation you would not expect to see that your data was normally distributed. It is for this reason that we work with the residuals not the data when checking the model assumptions.

## Equality of variance of each of the factor levels

This assumption should be tested using any one of a number of similar tests including Levene's test, Brown-Forsyth test, Bartlett's test and Hartley's test. These test check the following hypothesis:

$H_0$ : The variance of the factor levels are equal
$H_a$ : The variance of the factor levels are not equal

This assumption should be checked before the model is fitted to the data. If the assumption is not met, a weighted least squares analysis should be done.

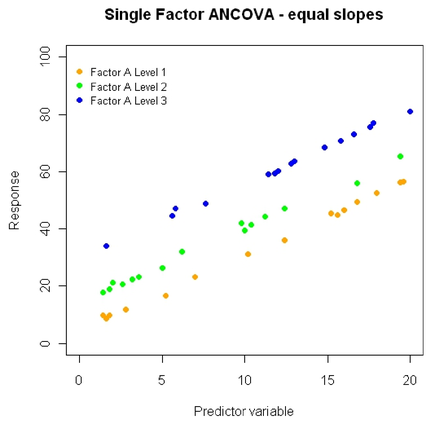
### Why the model assumptions are not met...

There can be many reasons for the model assumptions not being met. It could be that all the important terms are not fitted in the model, or the relationships between the data have not been accommodated by the model. There may be a relationship between the mean of the data and the variance (heterogeneity) or the lack of independence, just to name a few. Every time a model is fitted to the data the model assumptions need to be checked and where they are not met, a different model tested or a transformation of the data done. The new model or the model based on the transformed data will need to have the model assumptions checked again.

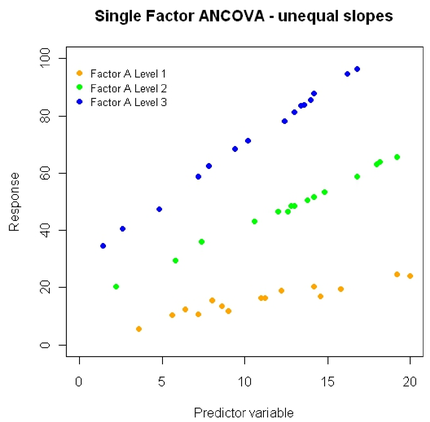
## Analysis of covariance

Two simple ANCOVA models are:

$$y_{ij} = \mu + \alpha_i + \beta(x_{ij} - \bar{x}) + \varepsilon_{ij} \qquad (1)$$




$$y_{ij} = \mu + \alpha_i + \beta(x_{ij} - \bar{x}) + \alpha\beta(x_{ij} - \bar{x}) + \varepsilon_{ij} \qquad (2)$$




## Analysis of covariance continued

The assumptions associated with these models are:

1. the residuals need to have a mean of zero
2. the residuals need to be normally distributed
3. the residuals need to have a constant variance
4. the residuals need to be independent
5. the variance of each of the factor levels need to be equal
6. parallel slopes - if using Equation 1, the slope for each of the factor levels needs to be similar. Otherwise, Equation 2 must be used.
7. If using either Equation 1 or 2, there must be a linear relationship between the covariate and the response.
8. The covariate range must be similar for all factor levels.

### How to handle outliers

Dealing with outliers can always be problematic, should they be left in or removed for the analysis? Firstly, once an possible outlier is identified, always go back and check for typographical errors or whether there is some reason why the observations has caught your attention. If there is a clear reason for excluding the observation (say the experimental unit was treated differently to the rest for some reason), then report the outlier and exclude from the analysis. Otherwise and outlier detection method may be used to justify the exclusion of the observation. I am in favour of analysing the data with and without the "outlier". Then the impact of the outlier can be assessed. What you need to think carefully about is that the observation might be real and part of the normal variability associated with the population of interest. By excluding an "outlier" without justification the results from the analysis can be compromised.