2. Linear Regression Model Assumptions

How do we evaluate a model? How do we know if the model we are using is good? One way to consider these questions is to assess whether the assumptions underlying the simple linear regression model seem reasonable when applied to the dataset in question. Since the assumptions relate to the (population) prediction errors, we do this through the study of the (sample) estimated errors, the residuals.

Assumptions on the Regression Model

We have already highlighted the set of conditions that comprises simple linear regression model as:

$$y_i = \alpha + \beta x_i + \varepsilon_i \qquad \varepsilon_i \sim iiN(0, \sigma^2)$$

- The mean of *Y*, at each value of *X*, is a **Linear function** of *x*.
- The errors are **Independent**.
- The errors at each value of *X* are **Normally distributed**.
- The errors, at each value of *X*, have **Equal variances**.

The four conditions of the model can tell us what can go wrong with our model.

It is clear why we have to evaluate any regression model that we formulate and subsequently estimate.

- All of the estimates, intervals, and hypothesis tests arising in a regression analysis have been developed assuming that the model is correct. That is, all the formulae depend on the model being correct.
- If the model is incorrect, then the formulae and methods we use are at risk of being incorrect.

The good news is that some of the model conditions are more forgiving than others. So, we really need to learn when we should worry the most and when it is possible to be more carefree about model violations.

- All tests and intervals are very sensitive to even minor departures from independence.
- All tests and intervals are sensitive to moderate departures from equal variance.
- The hypothesis tests and confidence intervals for α and β are fairly "**robust**" against departures from normality.
- Prediction intervals are quite sensitive to departures from normality.

The important thing to remember is that the severity of the consequences is always related to the severity of the violation. And, how much you should worry about a model violation depends on how you plan to use your regression model. For example, if all you want to do with your model is test for a relationship between x and y, *i.e.* test that the slope β is 0, you should be alright even if it appears that the normality condition is violated. On the other hand, if you want to use your model to predict a future response, then you are likely to get inaccurate results if the error terms are not normally distributed.

Consequences of Invalid Assumptions

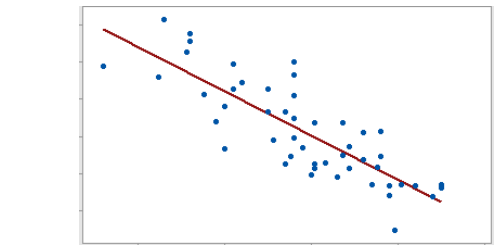
- Using the wrong equation (such as using a straight line for curved data) is a disaster. Predicted values will be wrong in a biased manner, meaning that predicted values will systematically miss the true pattern of the mean of y (as related to the x-variables).
- It is not possible to check the assumption that the overall mean of the errors is equal to 0 because the least squares process causes the residuals to sum to 0. However, if the wrong equation is used and the predicted values are biased, the sample residuals will be patterned so that they may not average 0 at specific values of x.
- The principal consequence of non-constant variance (heterosceasticity) is prediction intervals for individual y values will be wrong because they are determined assuming constant variance. There is a small effect on the validity of t-test and F-test results, but generally regression inferences are robust with regard to the variance issue.
- If the errors do not have a normal distribution, it usually is not particularly serious. Simulation results have shown that regression inferences tend to be robust with respect to normality (or non-normality of the errors). In practice, the residuals may appear to be non-normal when the wrong regression equation has been used.

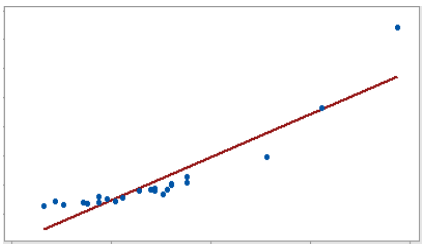
Diagnosing Validity of Assumptions

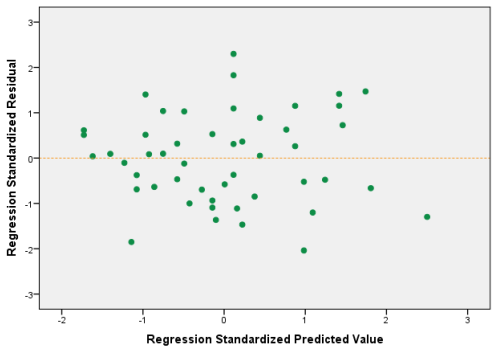
1. *Diagnosing Whether the Right Type of Equation Was Used*

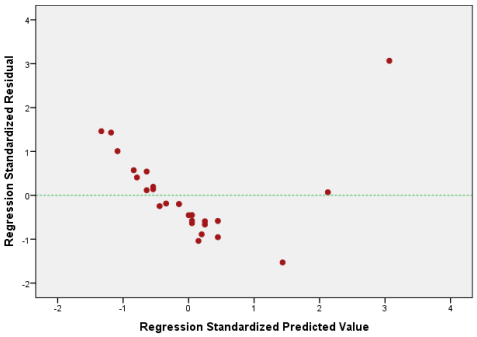
   Examine a plot of (standardised) residuals versus fits (predicted values). A curved pattern for the residuals versus fits plot indicates that the wrong type of equation has been used.

Plot of y against x



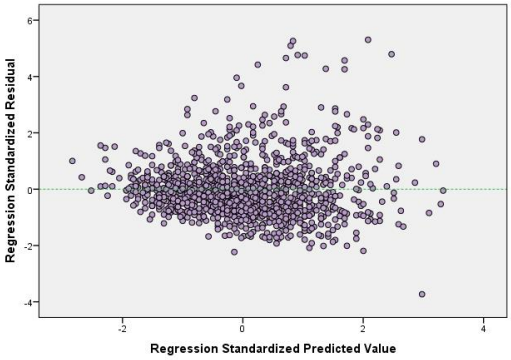






2. *Diagnosing Whether the Variance is Constant or Not*

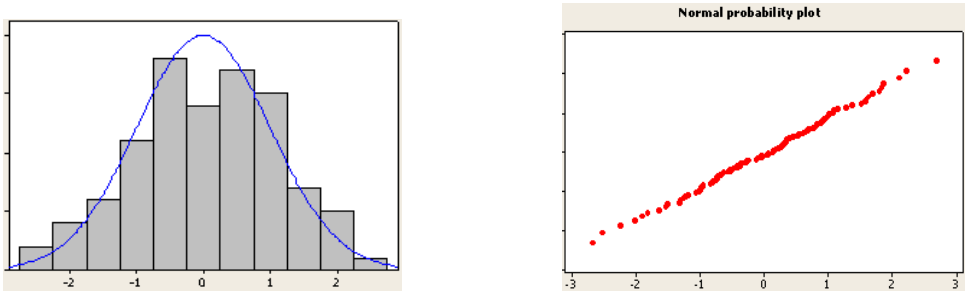
   - Examine a plot of (standardised) residuals versus fits. Obvious differences in the vertical spread of the residuals indicate non-constant variance. The most typical pattern for non-constant variance is a plot of residuals versus fits with a pattern that resembles a sideways cone.



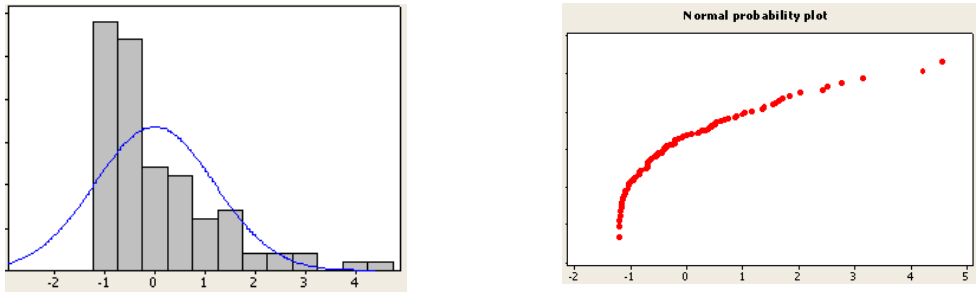
- Do a hypothesis test to test the null hypothesis that the variance of the errors is the same for all values of the *x*-variable(s). There are various statistical tests that can be used, such as the modified Levene's test. In practice, these tests are not used very often because non constant variance tends to be obvious from the plot of residuals versus fits plot.
- One other way to determine whether the variance is stable is to compute the Spearman rank correlation coefficient between the absolute value of the residuals and the predicted values. This statistic measures the correlation between the size of the ordered predicted values and the absolute value of their associated residuals. If this quantity is close to zero, it means there is no correlation between the size of the predicted value and the magnitude of the residual, indicating that the variances are equal. Positive values mean that the magnitude of the residuals increases as the predicted values increase, indicating that the variance increases as the mean increases. Negative values indicate that the variance decreases as the mean increases.

3. *Diagnosing Whether the Errors Have a Normal Distribution*

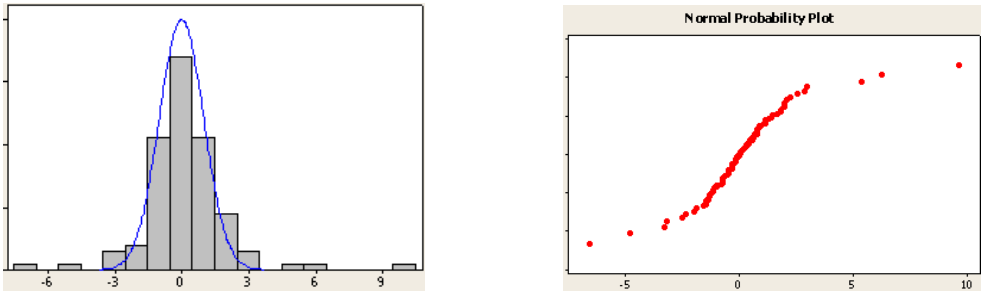
- Examine a histogram of the residuals, overlaid by normal curve to see if it appears to be bell-shaped). The difficulty is that the shape of a histogram may be difficult to judge unless the sample size is large enough.
- Examine P-P (Q-Q) plots of the residuals. Essentially, the ordered (standardised) residuals are plotted against theoretical expected values for a sample from a standard normal curve population. A straight-line pattern indicates that the assumption of normality is reasonable.



Normally distributed errors



Non-normal errors (positively skewed)

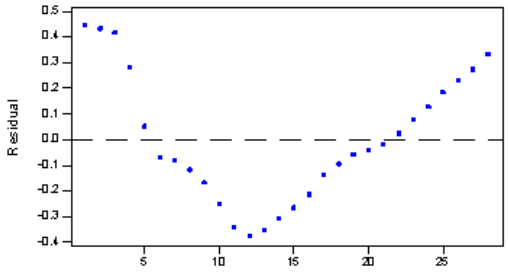


Non-normal errors(too many extreme positive and negative residuals)

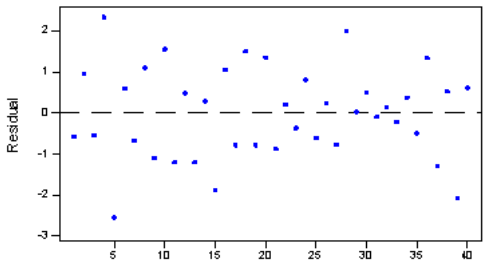
- Do a hypothesis test in which the null hypothesis is that the errors have a normal distribution. Failure to reject this null hypothesis is a good result. It means that it is reasonable to assume that the errors have a normal distribution.
  Kolmogorov-Smirnov Test and **Shapiro-Wilk** are two of such tests.

4. *Diagnosing Independence of the Error Terms*
   Knowing how your data is generated helps evaluate the assumption of independence. Correlated error terms can arise from data from a complex survey design, from repeated measures on a given subject, any type of clustered data, or data gathered over time.

- Usually, experiments are constructed in such a way that independence of the observations will be assumed.
- plots of the residuals versus time (order) to examine if there seems to be any positive or negative autocorrelations.



Positive autocorrelation



Negative autocorrelation

- Durbin-Watson statistic or the first-order autocorrelation. The first-order autocorrelation is defined as $\varepsilon_t = \rho\varepsilon_{t-1} + \nu_t$.
- The Durbin-Watson statistic is $d = \dfrac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$ which, can be shown, is also $d \simeq 2(1-\rho)$.

Remedial Measures

1. When the relationship between the dependent variable and one or more predictor variables does not follow a linear relationship, you might consider transforming the predictor variables to obtain the linearity. Sometimes a polynomial regression model might be a better solution. For applications in physical fields such as chemistry, and biology, the relationship might not be linear in terms of parameters, in which case you might want to fit a nonlinear regression model. When the parametric form of the relationship is difficult or impossible to define, you might want to fit a nonparametric regression model.

2. If your data appears to show heterosceasticity (non-constant variance), you can transform the dependent variable to stabilise the variance, or use different tools to model the non-constant variance. These can include Generalised Least Square method.

3. When the normality assumption is violated, you can transform the dependent variable to normalise the distribution. Alternatively, a non-normal regression procedure can be used.

4. Correlated errors can affect the standard errors of the parameter estimates and therefore affect the confidence intervals and the significance tests for the parameters. You should use other appropriate procedures to model data with correlated errors.

Practical Week 2

Retrieve the file **trig.sav**. The data for this example are from a study of obese patients who took a drug and followed a weight reduction programme for 8 weeks. One objective of the study was to see whether weight loss leads to reductions in triglyceride levels.

The weight loss and reduction in triglyceride levels-together with few other variables- for 34 patients were recorded.

1. Plot the scatter diagram of reduction in triglyceride against weight-loss, add the regression line and comment.

Carry out the regression of the reduction in triglyceride levels on weight-loss.

2. How large/small is R-squared?
3. Is the regression statistically significant?
4. What is the estimate of σ^2?
5. Write down the regression equation.
6. Check the validity of assumptions of the regression as described in the lecture.