0.1 Residual

Residual (or error) represents unexplained (or residual) variation after fitting a regression model. It is the difference (or left over) between the observed value of the variable and the value suggested by the regression model.

- The $epsilon_i$ values are called **emphresiduals** (they are the errors made using the regression model).
- A residual is positive when the observed value y_i is greater than $\hat{y}_i$, the value predicted using the model.
- A residual is negative when the observed value y_i is less than $\hat{y}_i$.

0.2 Assumptions for Linear Regression

- The general assumptions underlying the regression analysis model presented in this chapter are that (1) the dependent variable is a random variable, and (2) the independent and dependent variables are linearly associated.
- Assumption (1) indicates that although the values of the independent variable may be controlled, the values of the dependent variable must be obtained through the process of random sampling.
- If interval estimation or hypothesis testing is done in the regression analysis, three additional required assumptions are that (3) the variances of the conditional distributions of the dependent variable, given different values for the independent variable, are all equal, (4) the conditional distributions of the dependent variable, given different values for the independent variable, are all normally distributed in the population of values, and (5) the observed values of the dependent variable are independent of each other.
- When you choose to analyse your data using linear regression, part of the process involves checking to make sure that the data you want to analyse can actually be analysed using linear regression.
- You need to do this because it is only appropriate to use linear regression if your data is appropriate for six assumptions that are required for linear regression to give you a valid result.
- In practice, checking for these six assumptions just adds a little bit more time to your analysis, requiring you to click a few more buttons in `R` when performing your analysis, as well as think a little bit more about your data, but it is not a difficult task.
- Often when analysing your own data using `R`, one or more of these assumptions is violated (i.e., not met). This is not uncommon when working with real-world data rather than textbook examples, which often only show you how to carry out linear regression when everything goes well. However, even when your data fails certain assumptions, there is often a solution to overcome this.
- First, lets take a look at these six assumptions:
- **Assumption 1**: Your two variables should be measured at the interval or ratio level (i.e., they are continuous).

  Examples of variables that meet this criterion include revision time (measured in hours), intelligence (measured using IQ score), exam performance (measured from 0 to 100), weight (measured in kg), and so forth.

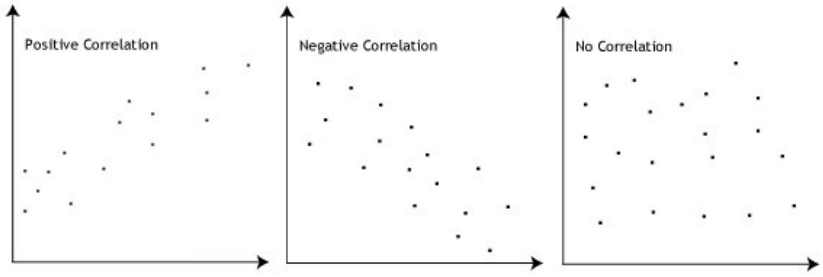

Figure 1:

- **Assumption 2**: There needs to be a linear relationship between the two variables. Whilst there are a number of ways to check whether a linear relationship exists between your two variables, we suggest creating a scatter-plot using `R`, where you can plot the dependent variable against your independent variable, and then visually inspect the scatter-plot to check for linearity. Your scatter-plot may look something like one of the following:

  If the relationship displayed in your scatterplot is not linear, you will have to either run a non-linear regression analysis or ***transform*** your data, which you can do using `R`.

  It is important to learn how to:

  (a) create a scatterplot to check for linearity when carrying out linear regression using `R`;

  (b) interpret different scatterplot results;

  (c) transform your data using `R` if there is not a linear relationship between your two variables.

- **Assumption 3**: There should be no significant outliers. Outliers are simply single data points within your data that do not follow the usual pattern (e.g., in a study of 100 students IQ scores, where the mean score was 108 with only a small variation between students, one student had a score of 156, which is very unusual, and may even put her in the top 1% of IQ scores globally). The following scatterplots highlight the potential impact of outliers:

  The problem with outliers is that they can have a negative effect on the regression equation that is used to predict the value of the dependent (outcome) variable based on the independent (predictor) variable. This will change the output that `R` produces and reduce the predictive accuracy of your results. Fortunately, when using `R` to run linear regression on your data, you can easily include criteria to help you detect possible outliers.

- **Assumption 4**: You should have independence of observations, which you can easily check using the Durbin-Watson statistic, which is a simple test to run using `R`. We explain how to interpret the result of the Durbin-Watson statistic later.

- **Assumption 5**:Your data needs to show ***homoscedasticity***, which is where the variances along the line of best fit remain similar as you move along the line. Whilst we explain more about what this means and how to assess the homoscedasticity of your data in the linear regression line, take a look at the two scatter-plots below, which provide two simple examples: one of data that meets this assumption and one that fails the assumption:

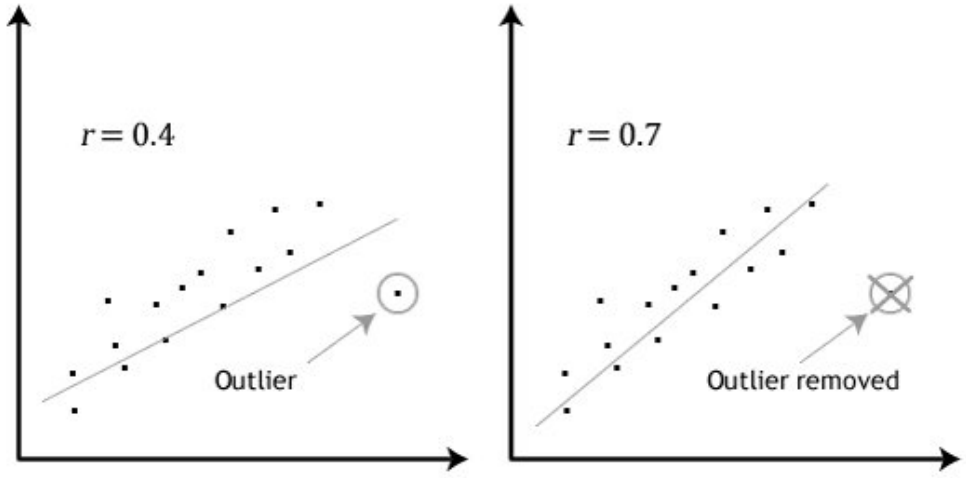

Figure 2: Effect of an Outlier

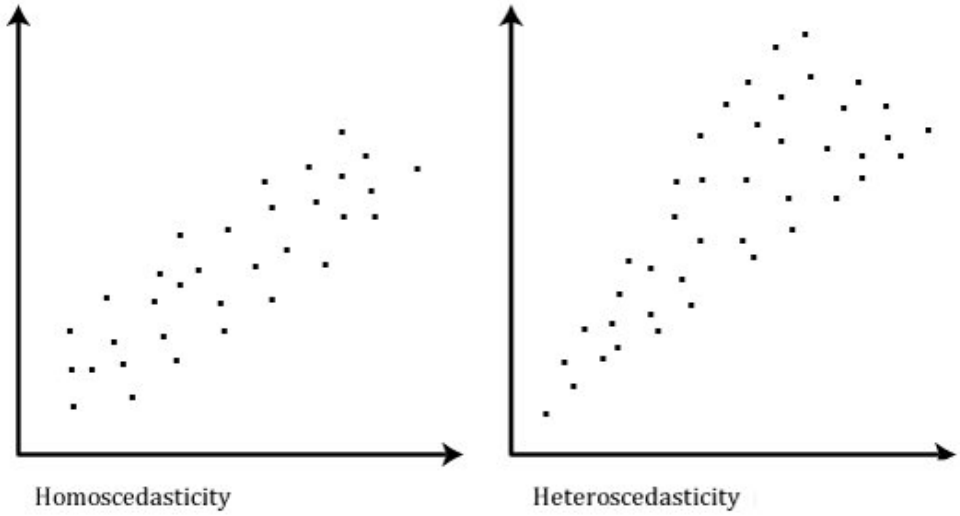

Figure 3: Constant Variance

When you analyse your own data, you will be lucky if your scatterplot looks like either of the two above. Whilst these help to illustrate the differences in data that meets or violates the assumption of homoscedasticity, real-world data is often a lot more messy.

Therefore, in our enhanced linear regression guide, we explain: (a) some of the things you will need to consider when interpreting your data; and (b) possible ways to continue with your analysis if your data

fails to meet this assumption.

- **Assumption 6**: Finally, you need to check that the residuals (errors) of your two variables are approximately normally distributed.

  Two common graphical methods to check this assumption include using either a histogram (with a superimposed normal curve) or by using a Normal P-P Plot. You may also use the Shapiro Wilk Test for normality.

You can check assumptions all assumptions except no.1 using `R`. It is recommended to test these assumptions in this order because it represents an order where, if a violation to the assumption is not correctable, you will no longer be able to use a single linear regression (although you may be able to run another statistical test on your data instead). Just remember that if you do not run the statistical tests on these assumptions correctly, the results you get when running a linear regression might not be valid.

- Linear regression is the most basic and commonly used predictive analysis. Regression estimates are used to describe data and to explain the relationship between one dependent variable and one or more independent variables.

- At the center of the regression analysis is the task of fitting a single line through a scatter plot. The simplest form with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent, c = constant, b = regression coefficients, and x = independent variable.

- Sometimes the dependent variable is also called a criterion variable, endogenous variable, prognostic variable, or regressand. The independent variables are also called exogenous variables, predictor variables or regressors.

- However linear regression analysis consists of more than just fitting a linear line through a cloud of data points. It consists of 3 stages (1) analyzing the correlation and directionality of the data, (2) estimating the model, i.e., fitting the line, and (3) evaluating the validity and usefulness of the model.

- There are 3 major uses for regression analysis (1) causal analysis, (2) forecasting an effect, (3) trend forecasting. Other than correlation analysis, which focuses on the strength of the relationship between two or more variables, regression analysis assumes a dependence or causal relationship between one or more independent and one dependent variable.

- Firstly, it might be used to identify the strength of the effect that the independent variable(s) have on a dependent variable. Typical questions are what is the strength of relationship between dose and effect, sales and marketing spend, age and income.

- Secondly, it can be used to forecast effects or impacts of changes. That is regression analysis helps us to understand how much will the dependent variable change, when we change one or more independent variables. Typical questions are how much additional Y do I get for one additional unit X.

- Thirdly, regression analysis predicts trends and future values. The regression analysis can be used to get point estimates. Typical questions are what will the price for gold be in 6 month from now? What is the total effort for a task X?

0.3 Assumptions for Linear Regression

- The general assumptions underlying the regression analysis model presented in this chapter are that (1) the dependent variable is a random variable, and (2) the independent and dependent variables are linearly associated.
- Assumption (1) indicates that although the values of the independent variable may be controlled, the values of the dependent variable must be obtained through the process of random sampling.
- If interval estimation or hypothesis testing is done in the regression analysis, three additional required assumptions are that (3) the variances of the conditional distributions of the dependent variable, given different values for the independent variable, are all equal, (4) the conditional distributions of the dependent variable, given different values for the independent variable, are all normally distributed in the population of values, and (5) the observed values of the dependent variable are independent of each other.

0.3.1 Assumptions

- When you choose to analyse your data using linear regression, part of the process involves checking to make sure that the data you want to analyse can actually be analysed using linear regression.
- You need to do this because it is only appropriate to use linear regression if your data is appropriate for six assumptions that are required for linear regression to give you a valid result.
- In practice, checking for these six assumptions just adds a little bit more time to your analysis, requiring you to click a few more buttons in `R` when performing your analysis, as well as think a little bit more about your data, but it is not a difficult task.
- Often when analysing your own data using `R`, one or more of these assumptions is violated (i.e., not met). This is not uncommon when working with real-world data rather than textbook examples, which often only show you how to carry out linear regression when everything goes well. However, even when your data fails certain assumptions, there is often a solution to overcome this.
- First, lets take a look at these six assumptions:
- **Assumption 1**: Your two variables should be measured at the interval or ratio level (i.e., they are continuous).

  Examples of variables that meet this criterion include revision time (measured in hours), intelligence (measured using IQ score), exam performance (measured from 0 to 100), weight (measured in kg), and so forth.
- **Assumption 2**: There needs to be a linear relationship between the two variables. Whilst there are a number of ways to check whether a linear relationship exists between your two variables, we suggest creating a scatter-plot using `R`, where you can plot the dependent variable against your independent variable, and then visually inspect the scatter-plot to check for linearity. Your scatter-plot may look something like one of the following:

  If the relationship displayed in your scatterplot is not linear, you will have to either run a non-linear regression analysis or ***transform*** your data, which you can do using `R`.

  It is important to learn how to:

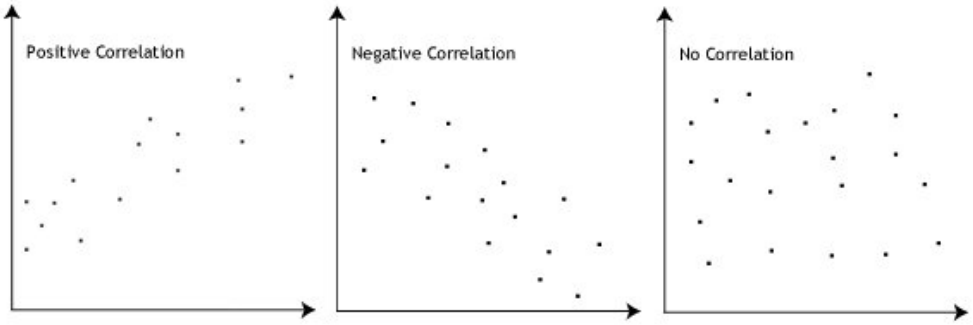

Figure 4: Types of Linear Relationship

(a) create a scatterplot to check for linearity when carrying out linear regression using `R`;

(b) interpret different scatterplot results;

(c) transform your data using `R` if there is not a linear relationship between your two variables.

- **Assumption 3**: There should be no significant outliers. Outliers are simply single data points within your data that do not follow the usual pattern (e.g., in a study of 100 students IQ scores, where the mean score was 108 with only a small variation between students, one student had a score of 156, which is very unusual, and may even put her in the top 1% of IQ scores globally). The following scatterplots highlight the potential impact of outliers:

  The problem with outliers is that they can have a negative effect on the regression equation that is used to predict the value of the dependent (outcome) variable based on the independent (predictor) variable. This will change the output that `R` produces and reduce the predictive accuracy of your results. Fortunately, when using `R` to run linear regression on your data, you can easily include criteria to help you detect possible outliers.

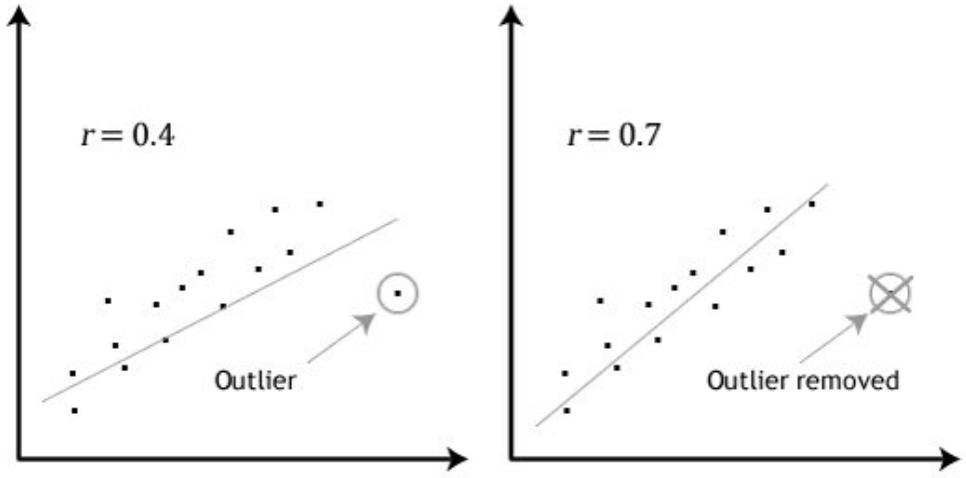

Figure 5: Effect of an Outlier

- **Assumption 4**: You should have independence of observations, which you can easily check using the Durbin-Watson statistic, which is a simple test to run using `R`. We explain how to interpret the result of the Durbin-Watson statistic later.
- **Assumption 5**:Your data needs to show ***homoscedasticity***, which is where the variances along the line of best fit remain similar as you move along the line. Whilst we explain more about what this means and how to assess the homoscedasticity of your data in the linear regression line, take a look at the two scatter-plots below, which provide two simple examples: one of data that meets this assumption and one that fails the assumption:

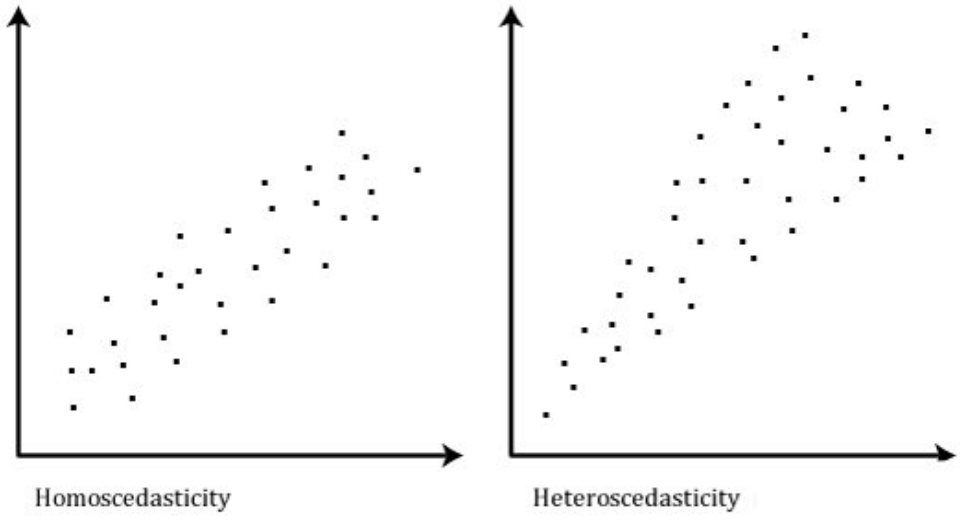

Figure 6: Constant Variance

When you analyse your own data, you will be lucky if your scatterplot looks like either of the two above. Whilst these help to illustrate the differences in data that meets or violates the assumption of homoscedasticity, real-world data is often a lot more messy.

Therefore, in our enhanced linear regression guide, we explain: (a) some of the things you will need to consider when interpreting your data; and (b) possible ways to continue with your analysis if your data fails to meet this assumption.

- **Assumption 6**: Finally, you need to check that the residuals (errors) of your two variables are approximately normally distributed.

  Two common graphical methods to check this assumption include using either a histogram (with a superimposed normal curve) or by using a Normal P-P Plot. You may also use the Shapiro Wilk Test for normality.

You can check assumptions all assumptions except no.1 using R. It is recommended to test these assumptions in this order because it represents an order where, if a violation to the assumption is not correctable, you will no longer be able to use a single linear regression (although you may be able to run another statistical test on your data instead). Just remember that if you do not run the statistical tests on these assumptions correctly, the results you get when running a linear regression might not be valid.