Regression

Introduction to regression

In regression, our objective is to understand some dependent variable Y based on an independent variable X. Regression is a tool for predicting the value of Y as a function of X. We can use this tool to do the following,

Introduction to regression

In regression, our objective is to understand some dependent variable Y based on an independent variable X. Regression is a tool for predicting the value of Y as a function of X. We can use this tool to do the following,

- Support hypotheses of causation of changes in y values due to changes in x values
- Predict y values as a function of x values
- To explain the variation of y values using x values

Introduction to regression

In regression, our objective is to understand some dependent variable Y based on an independent variable X. Regression is a tool for predicting the value of Y as a function of X. We can use this tool to do the following,

- Support hypotheses of causation of changes in y values due to changes in x values
- Predict y values as a function of x values
- To explain the variation of y values using x values

Regression analysis can be used to support causal hypotheses, but it **cannot, by itself be used to determine causality.**

Visualising a regression of y values against x values

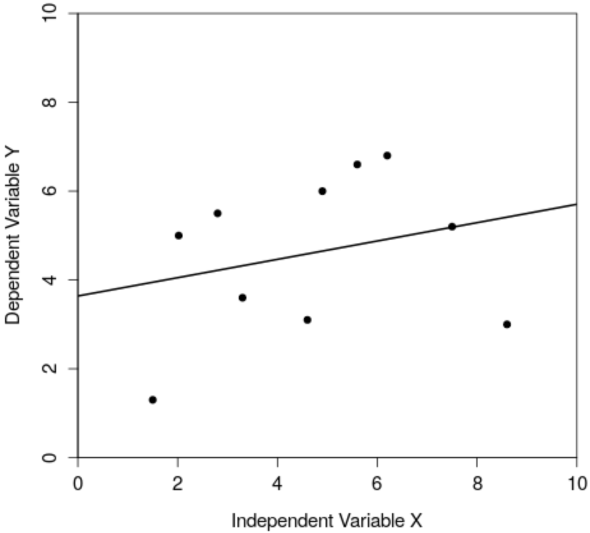

Figure 1: Regression of dependent variable y against independent x.

Regression includes independent and dependent variables

It is critical to correctly distinguish between the independent and dependent variables.

Regression includes independent and dependent variables

It is critical to correctly distinguish between the independent and dependent variables.

- Independent variable (X) is free to vary
- Dependent variable (Y) is predicted to change given a change in the independent variable
- Different results will be obtained if the two variables are confused

Regression includes independent and dependent variables

It is critical to correctly distinguish between the independent and dependent variables.

- Independent variable (X) is free to vary
- Dependent variable (Y) is predicted to change given a change in the independent variable
- Different results will be obtained if the two variables are confused

In an experiment, the independent variable is something that we as researchers have control over (e.g., amount of fertiliser to put down on a field), whereas the dependent variable is something that we would measure when collecting our data (e.g., total crop yield of the field).

Regression line

The line of best fit in a regression can be described mathematically with a simple equation,

$$y = \beta_0 + \beta_1 x.$$

This equation includes the variables x and y, and two coefficients

Regression line

The line of best fit in a regression can be described mathematically with a simple equation,

$$y = \beta_0 + \beta_1 x.$$

This equation includes the variables x and y, and two coefficients

- β_0 is the **intercept**; the value of y that is predicted when $x = 0$
- β_1 is the **slope**; how much y changes for a change one unit of x

Regression line

The line of best fit in a regression can be described mathematically with a simple equation,

$$y = \beta_0 + \beta_1 x.$$

This equation includes the variables x and y, and two coefficients

- β_0 is the **intercept**; the value of y that is predicted when $x = 0$
- β_1 is the **slope**; how much y changes for a change one unit of x

Note that data points rarely will sit right on the regression line. The **residual** is defined by the difference between the measured value of y (i.e., the data point) and the y value predicted by the regression line (i.e., the vertical distance between the data point and the line).

Regression line

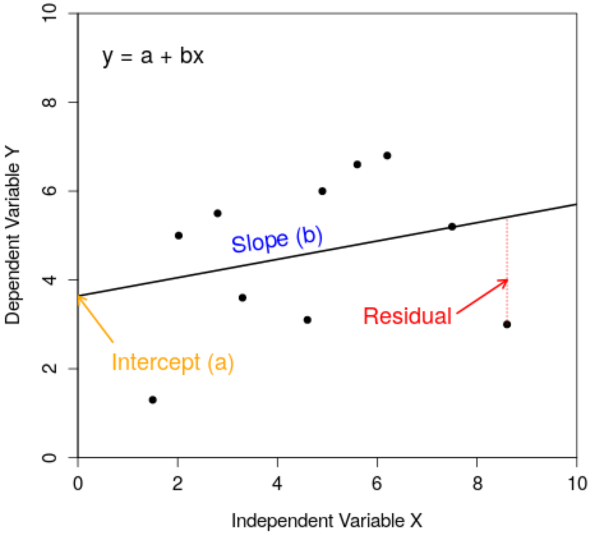

Figure 2: Regression of dependent variable y against independent x.

How do we decide what is the best fit line?

Now we can turn to how we calculate where the regression line should be through our data.

How do we decide what is the best fit line?

Now we can turn to how we calculate where the regression line should be through our data.

- How do we know what our intercept (a) and slope (b) should be?
- Use the method of **least squares regression**
- Minimise the sum of squares of all the residual values

How do we decide what is the best fit line?

Now we can turn to how we calculate where the regression line should be through our data.

- How do we know what our intercept (a) and slope (b) should be?
- Use the method of **least squares regression**
- Minimise the sum of squares of all the residual values

To get an intuitive sense for how the regression line minimises the sum of squares, use **[this interactive application]** to adjust the slope and intercept to try to find the line of best fit (it will turn blue when you succeed).

Assumptions of regression

Regression is a widely used, but also often misused, statistical technique. It is important to be aware of the assumptions underlying linear regression.

Assumptions of regression

Regression is a widely used, but also often misused, statistical technique. It is important to be aware of the assumptions underlying linear regression.

1. **The independent variable X is measured without error**
2. **The relationship between X and Y is linear**
3. **For any value of X, Y is normally distributed**
4. **For all values of X, the variance of the residuals is identical**

Assumptions of regression

Regression is a widely used, but also often misused, statistical technique. It is important to be aware of the assumptions underlying linear regression.

1. **The independent variable X is measured without error**
2. **The relationship between X and Y is linear**
3. **For any value of X, Y is normally distributed**
4. **For all values of X, the variance of the residuals is identical**

Note that even if our assumptions are not perfectly met (indeed, they rarely if ever will be), this does not completely invalidate the method of linear regression. But large violations of one or more of these assumptions might indeed be problematic.

Assumptions of regression

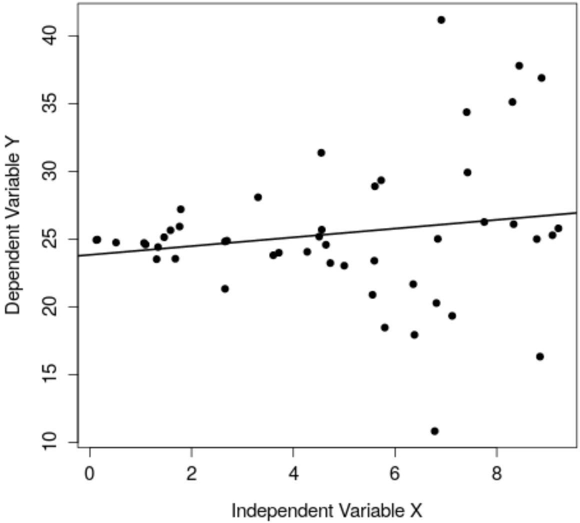

Figure 3: A regression of one dependent variable y against the independent

How good is the fit of our model?

We often want to know how well a regression line fits to the data.

- **Coefficient of determination** (R^2)
- R^2 tells us **how much of the variation in y is explained by the regression equation**

How good is the fit of our model?

We often want to know how well a regression line fits to the data.

- **Coefficient of determination** (R^2)
- R^2 tells us **how much of the variation in y is explained by the regression equation**
- E.g., if $R^2 = 0.83$, then 83% of the variation in y is accounted for by the fitted regression line
- Visually, how tightly the data points in a scatterplot fit to the regression line

How good is the fit of our model?

We often want to know how well a regression line fits to the data.

- **Coefficient of determination** (R^2)
- R^2 tells us **how much of the variation in y is explained by the regression equation**
- E.g., if $R^2 = 0.83$, then 83% of the variation in y is accounted for by the fitted regression line
- Visually, how tightly the data points in a scatterplot fit to the regression line

See the examples below of four different R^2 values to see what this looks like.

Scatterplots of different coefficients of determination

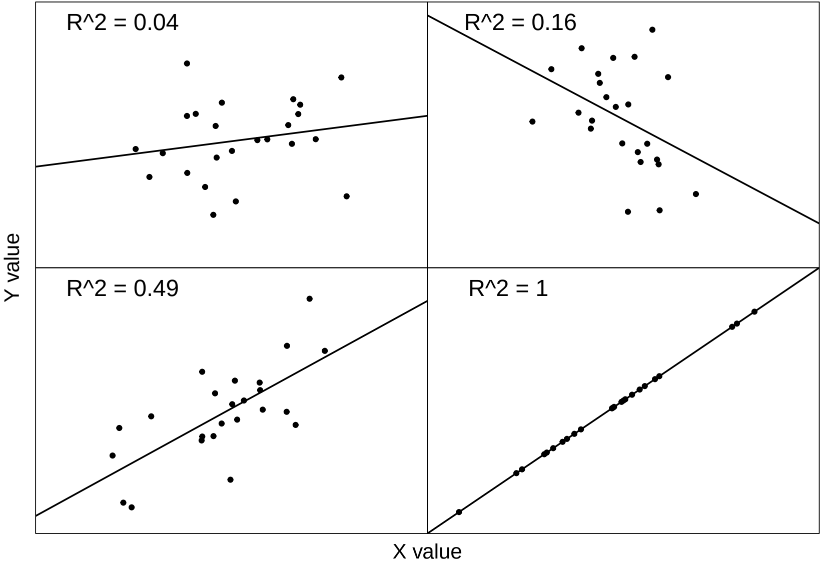

Figure 4: Four different coefficients of determination

More understand of the coefficient of determination

Understanding that **the coefficient of determination tells us how much variation in y is explained by the regression equation** is the important point.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}.$$

- Coefficient of determination compares the sum of squared residuals from the linear model (SS_{res}) to what the sum of squared residuals would be if we had just use the mean value of y (SS_{tot}).

More understand of the coefficient of determination

Understanding that **the coefficient of determination tells us how much variation in y is explained by the regression equation** is the important point.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}.$$

- Coefficient of determination compares the sum of squared residuals from the linear model (SS_{res}) to what the sum of squared residuals would be if we had just use the mean value of y (SS_{tot}).
- Conveniently, R^2 is also just the Pearson product moment correlation (r) squared.

More understand of the coefficient of determination

Understanding that **the coefficient of determination tells us how much variation in y is explained by the regression equation** is the important point.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}.$$

- Coefficient of determination compares the sum of squared residuals from the linear model (SS_{res}) to what the sum of squared residuals would be if we had just use the mean value of y (SS_{tot}).
- Conveniently, R^2 is also just the Pearson product moment correlation (r) squared.

Visualising the coefficient of determination

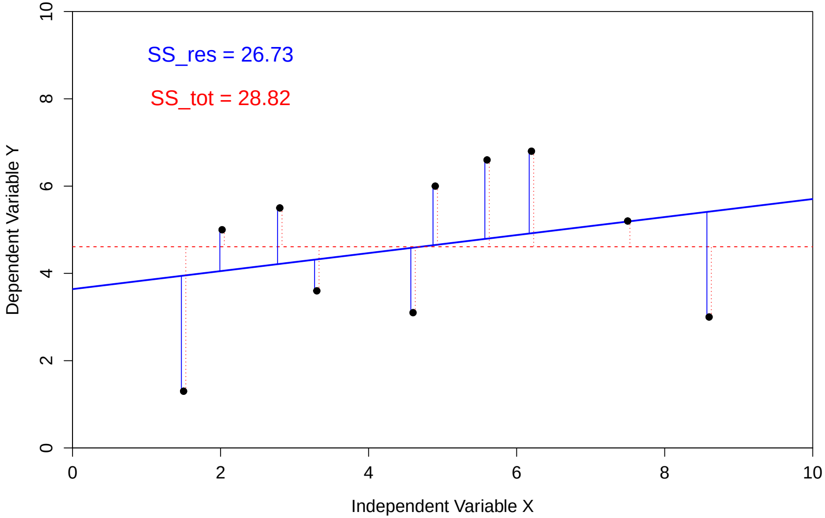

Figure 5: A regression of one dependent variable y against the independent variable x. Blue vertical lines show residuals of the linear model, while red dotted vertical lines show residual deviations of from the mean of y.

The F-test of overall significance

- Test the significance of our overall regression model using an F-test of overall significance
- Determines whether or not our linear regression provides a better fit to the data than a model that does not contain any independent variables

The F-test of overall significance

- Test the significance of our overall regression model using an F-test of overall significance
- Determines whether or not our linear regression provides a better fit to the data than a model that does not contain any independent variables

Hypothesis for F-test of overall significance of a linear model

- **Null:** The model with no independent variables fits the data as well as the linear model
- **Alternative:** The linear model fits the data better than the model with no independent variables

Understand is how to interpret the F-test of overall significance (see below for doing this in practice).

Significance of equation parameters

- Test the significance of individual parameters in the linear model (β_0 and β_1; recall that $y = \beta_0 + \beta_1 x$).
- For both coefficients β_0 and β_1, we can state the null and alternative hypotheses

Hypothesis for coefficient of a linear model

- **Null:** The value of the coefficient equals 0.
- **Alternative:** The value of the coefficient does not equal 0.

Statistical software such as jamovi will calculate p-values to test our null hypothesis for both β_0 and β_1 coefficients.

Assessing the practical validity of regression

The practical validity of the regression model is assessed by comparing the predicted values with the observed data. We can do this in several ways:

Assessing the practical validity of regression

The practical validity of the regression model is assessed by comparing the predicted values with the observed data. We can do this in several ways:

1. Plotting the fitted regression line and checking that the observed data lie close to the line (i.e., high coefficient of determination).

Assessing the practical validity of regression

The practical validity of the regression model is assessed by comparing the predicted values with the observed data. We can do this in several ways:

1. Plotting the fitted regression line and checking that the observed data lie close to the line (i.e., high coefficient of determination).
2. Plotting observed versus predicted values and observing a linear relationship between the independent and dependent variable.

Assessing the practical validity of regression

The practical validity of the regression model is assessed by comparing the predicted values with the observed data. We can do this in several ways:

1. Plotting the fitted regression line and checking that the observed data lie close to the line (i.e., high coefficient of determination).
2. Plotting observed versus predicted values and observing a linear relationship between the independent and dependent variable.
3. Examining the data for large residuals (i.e., outliers), which might be distorting the regression line.

Assessing the practical validity of regression

The practical validity of the regression model is assessed by comparing the predicted values with the observed data. We can do this in several ways:

1. Plotting the fitted regression line and checking that the observed data lie close to the line (i.e., high coefficient of determination).
2. Plotting observed versus predicted values and observing a linear relationship between the independent and dependent variable.
3. Examining the data for large residuals (i.e., outliers), which might be distorting the regression line.
4. Ideally, test the regression model on new observational data to examine how close the predicted values are to the observations

Prediction with linear regression

Regression equations can be used to calculate additional y values when values of x are substituted in a regression equation.

Prediction with linear regression

Regression equations can be used to calculate additional y values when values of x are substituted in a regression equation.

- **Interpolation**: Predictions made within the measurement range of the data
- **Extrapolation**: Predictions made outside the measurement range of the data

Prediction with linear regression

Regression equations can be used to calculate additional y values when values of x are substituted in a regression equation.

- **Interpolation**: Predictions made within the measurement range of the data
- **Extrapolation**: Predictions made outside the measurement range of the data

Care should be taken when extrapolating beyond the measured data because the relationship between the two variables might change.

Predictions with linear regression

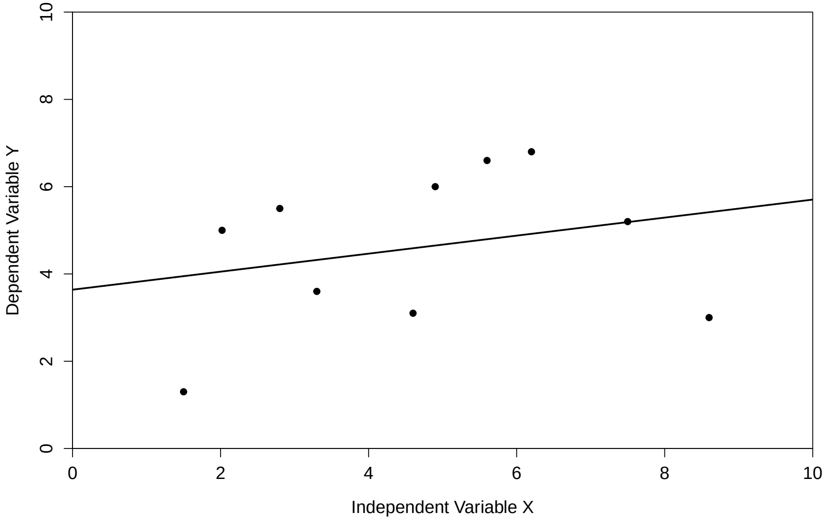

Figure 6: Regression of dependent variable y against independent x.