Regression validity

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

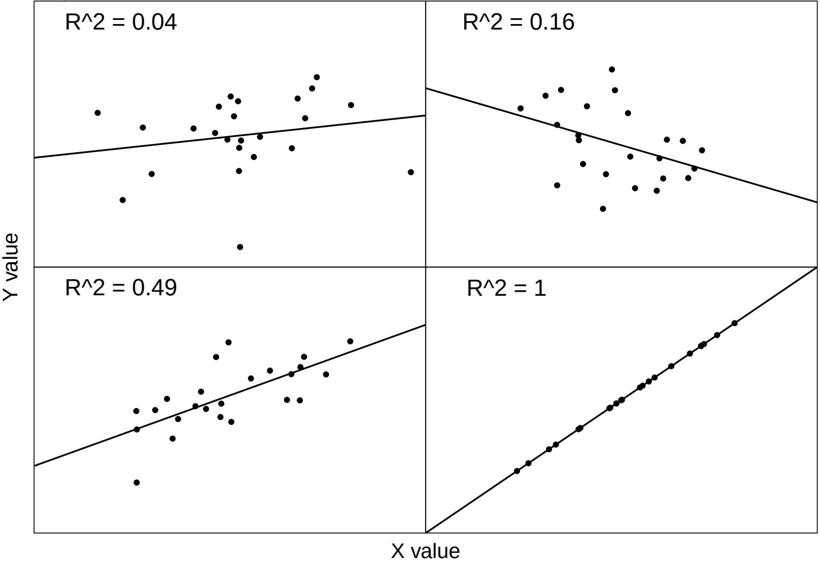

Figure 1: Four sample regressions showing different coefficients of determination

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}).
- Conveniently, R^2 is also just the Pearson product moment correlation (r) squared.

Visualising the coefficient of determination

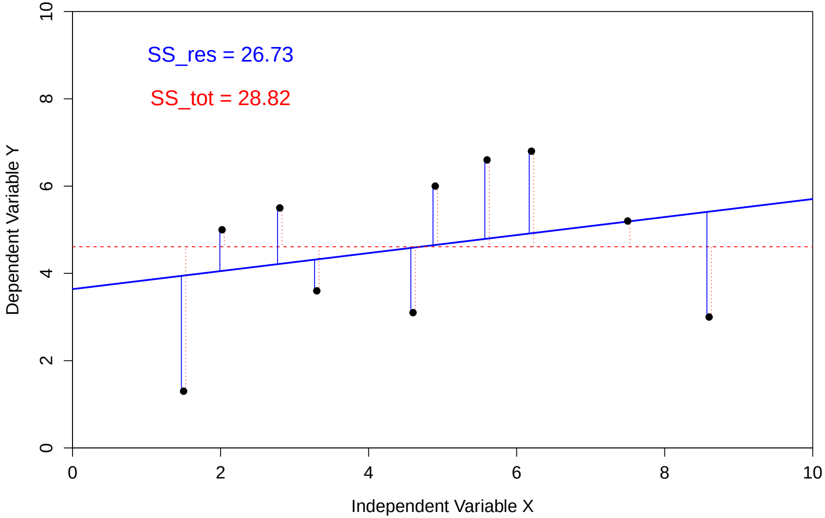

Figure 2: A regression of one dependent variable y against the independent variable x. Blue vertical lines show residuals of the linear model, while red dotted vertical lines show residual deviations of from the mean of y.

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

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 (a and b; recall that $y = a + bx$).
- For both coefficients a and b, 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 SPSS will calculate p-values to test our null hypothesis for both a and b coefficients. Lecture notes include the details of how this is done (you do not need to know these details).

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.

- **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

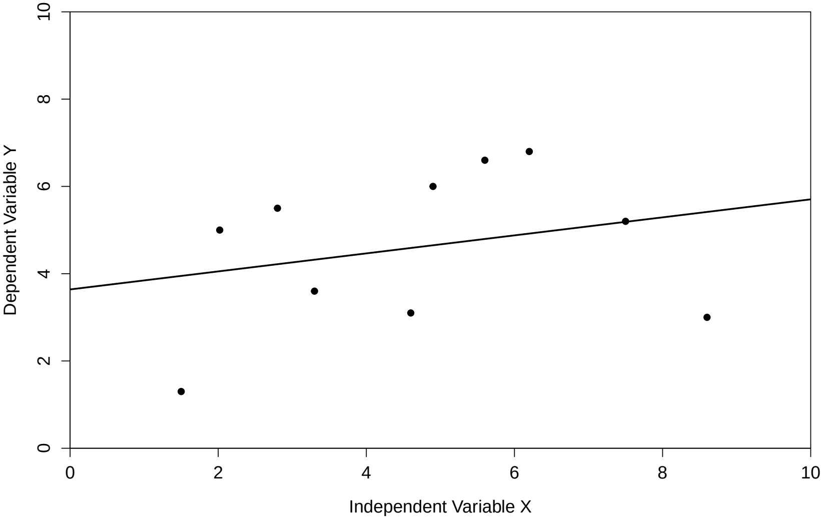

Figure 3: A regression of one dependent variable y against the independent variable x.