11.2 Theory Behind Regression

The idea of regression was developed by:

Sir Francis Galton (1822–1911) studied heredity and how tall or short couples have children, and how the heights of parents affect the height of their children tend to *regress*, or revert to the more typical mean height for people of the same gender.

regress - return to a former or less developed state

The regression line using the population parameters can be seen as:

$$y_i = \beta_0 + \beta_1 x_i$$

Estimates of regression line:

- β_0: population y-intercept parameter
- β_1: population slope parameter

The regression line using the sample estimates can be seen as:

$$y_i = b_0 + b_1 x_i$$

Estimates of regression line:

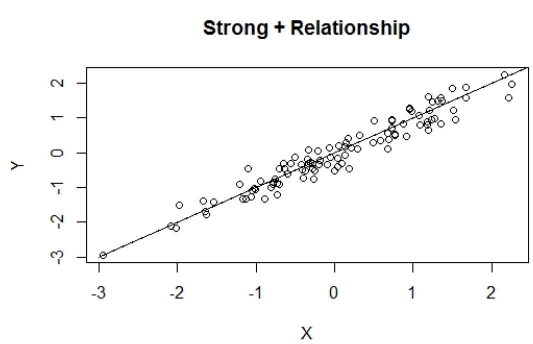
- b_0: sample y-intercept estimate
- b_1: sample slope estimate

How to calculate sample estimate slope b_1:

$$b_1 = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$
$$b_1 = r\frac{s_y}{s_x}$$

How to calculate sample estimate y-intercept b_0:

$$b_0 = \bar{y} - b_1\bar{x}$$



Using the Regression Equation for Predictions

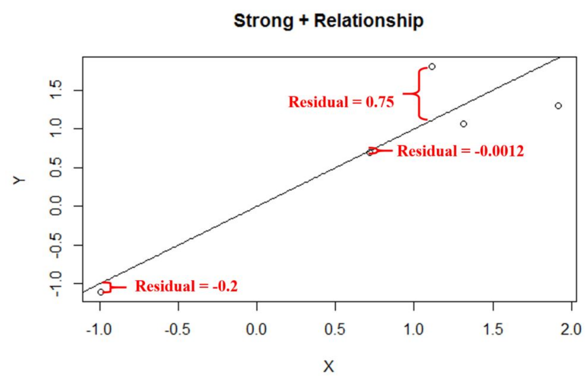
- Use the regression equation for predictions when the regression line on the scatterplot fits the points very closely.
- Use the regression equation for predictions only when the linear correlation coefficient r indicates that there is a linear correlation between the two variables (Reject $H_0 : \rho = 0$)
- Use the regression line for predictions only if the data does not go much beyond the interval of the data.
- If the regression equation does not appear to be useful for making predictions, the best predicted value of a variable is its point estimate, which is its sample mean.

Example 3:

Residuals and the Least-Squares Property

For a sample of data that contains x, the independent variable and y, the dependent variables, the residual is calculated by taking the difference between the observed value y and the predicted y-value, $\hat{y}_i$ using the regression equation.

$$\text{residual}\epsilon_i = y_i - \hat{y}_i$$

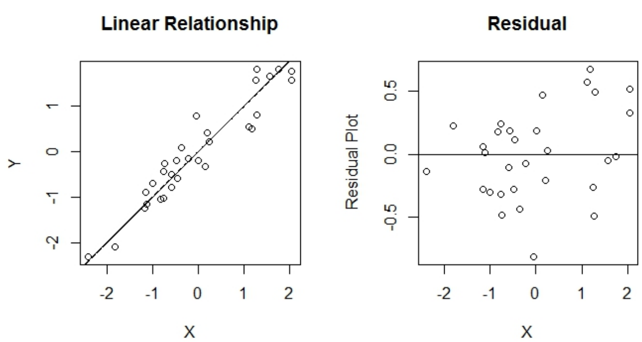


The regression line satisfies the **least-squares** property which is the sum of the squares of the residuals is very small.

Residual Plots

How to make a residual plot

1. Use the same x-axis as the scatterplot
2. use a vertical axis of residual values
3. Draw a horizontal reference line through the residual value of 0



When looking at a residual plot be sure that:

- it should not have any pattern that is not a straight-line pattern
- it should not become thicker (or thinner) when viewed from left to right

Coefficient of Determination, r^2
The correlation squared is called the coefficient of determination, as mentioned before, r^2 is the proportion of the variance explained by the regression line.

- r^2: is the variance explained the regression line
- $1 - r^2$: is the variance that is **not** explained by the regression line

The residual is the unexplained variance within the model. Another way of looking at this is to partition the deviation each value from the mean:

$$y - \bar{y} = y - \bar{y} + \hat{y} - \hat{y}$$
$$y - \bar{y} = (\hat{y} - \bar{y}) + (y - \hat{y})$$

Partition of Variation:

- $\sum y - \bar{y}$: Total Variation
- $\sum(\hat{y} - \bar{y})^2$: Explained Variation
- $\sum(y - \hat{y})^2$: Unexplained Variation

Equation for coefficient of Determination:

$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

Complete Regression Analysis:

- Construct a scatterplot and ensure a linear relationship exists between x and y
- Construct a residual plot and ensure no pattern exists
- Ensure data and residuals are normal*

Example 4:

Review of notation:

- y_i:
- $\hat{y}_i$:
- ϵ_i:
- $\bar{y}$:
- $\bar{x}$:
- x_i:
- r:
- r^2:
- β_0:
- β_1:
- b_0:
- b_1:

Prediction interval for an individual y
A prediction interval is an interval estimate of a predicted value of y.

When an x is used to predict $\hat{y}$ from the regression line an interval can be calculated to a confidence interval for y

$$s_e = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}}$$

$$ME = t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n(\sum x^2) - (\sum x)^2}}$$

where x_0 denotes the given x value, $t_{\alpha/2}$ has $n-2$ degrees of freedom, s_e

Example 5: