- **Types of regression**: Linear (continuous data), Logistic (categorical data), Cox (categorical data in a survival analysis).

- **Explanatory variable (X)**: Also known as a covariate, predictor, or independent variable.

- **Outcome variable (Y)**: Also known as response or dependent variable.

- **Simple Linear Regression (SLR)**: Looks at a relationship between two continuous variables where the relationship between the two variables is approximately a straight line.

- **SLR equation**:

$$Y = \beta_0 + \beta_1 x + e$$

  - This implies that the mean response is related to x by

$$\mu_Y = \beta_0 + \beta_1 x$$

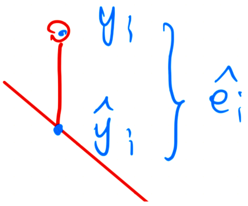
  - **Y** is the numerical outcome variable (continuous or approximately so).
  - **x** is the explanatory variable.
  - β_0 is the intercept or constant (where the line crosses the y-axis).
  - β_1 is the slope of the line.
  - **e** (often denoted ϵ) is the random error or residual term.

- **SLR equation for estimating:**

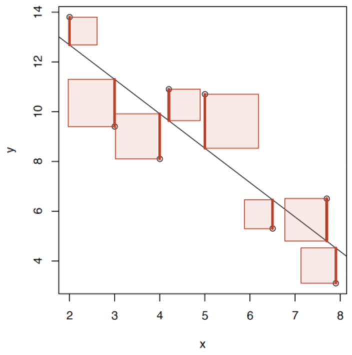
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- **Residual ('estimated error') term**:

$$\hat{e}_i = y_i - \hat{y}_i$$



- **How to find regression line**: The line of best fit minimises the sum of the squares of the residuals.



- **Equation for how to find regression line**:

$$\sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- **How to calculate β_1 and β_0**:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

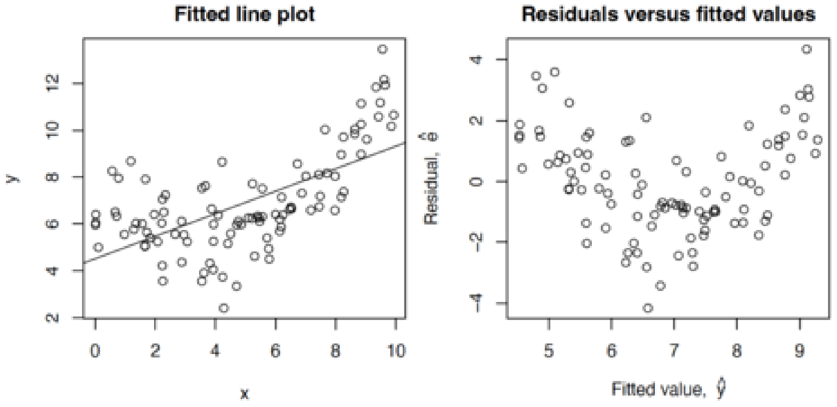
- **Example for how to calculate regression (Stress and Blood Pressure)**:
  - Get n, n = 6
  - Find the explanatory and outcome
  - Calculate β_1 and β_0
  - Get the regression equation
  - Using R for SLR

- **Assumptions for Simple Linear Regression (LINE)**:
  - Linearity: The relationship between the mean response μ_Y and x is described by a straight line.
  - Independence: The responses $Y_1, Y_2, ..., Y_n$ are statistically independent.
  - Normality: The error terms $e_1, e_2, ..., e_n$ come from a normal distribution.
  - Equal variance: The error terms all have the same variance, σ^2 ('homoscedastic').

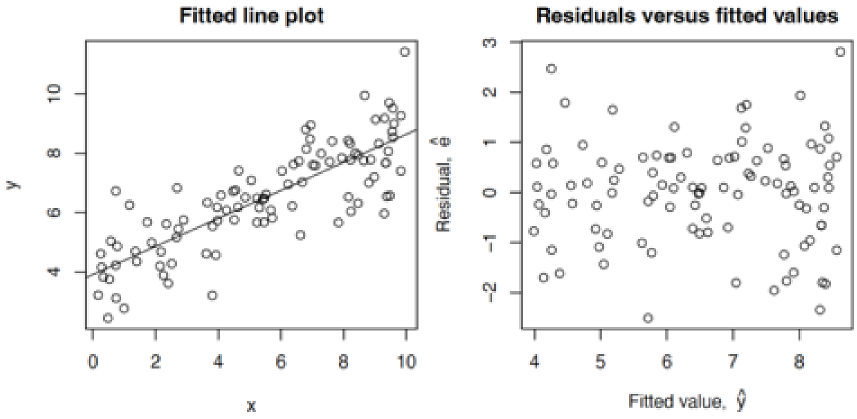
- **What diagram is used for checking linearity**: Residual plot.

$$\hat{e}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

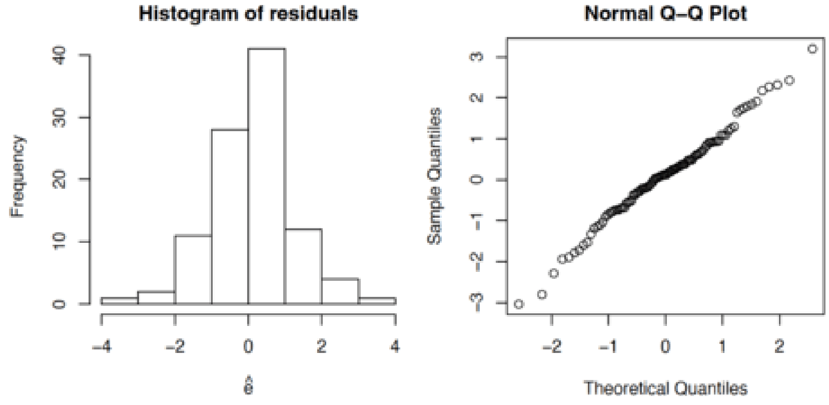
- **Failure of linearity assumption**:



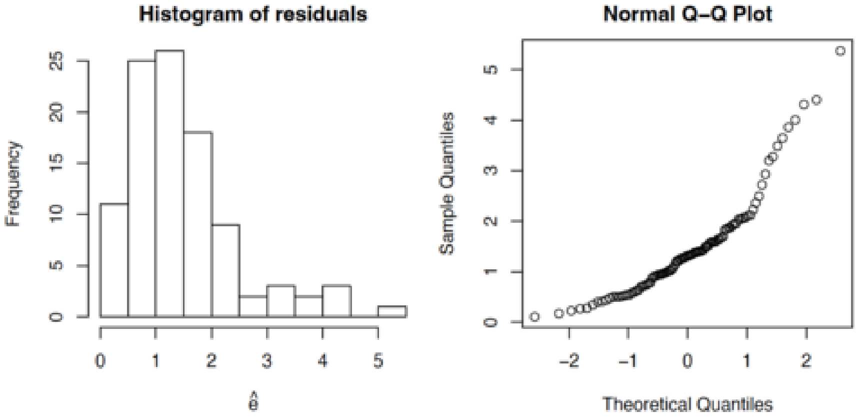
- **Linearity assumption holds**:



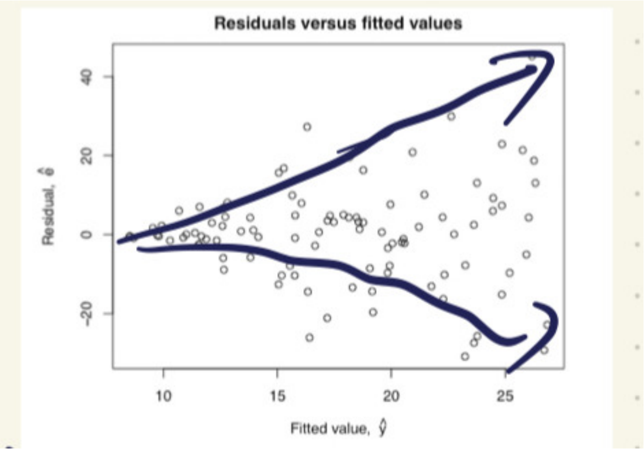
- **Checking independence assumption**: May get insight by thinking about the study design (Ask yourself questions).
- **Plot for Checking the normality assumption**: Q - Q plot.
- **Pass of normality assumption**:



- **Fail of normality assumption**:



- **Checking equal variance assumption (homoscedasticity)**: Pass if the residual plot is not like this.



- **What is the impact if Fail of the linearity assumption**: Critical. If that assumption fails, all conclusions drawn from the model will be invalid.
- **What is the impact if Fail of independence or equal variance assumptions**: Remain valid. However, estimates can be inefficient. Follows that the fitted regression line is useable. Any test results or confidence intervals based on the regression model will be invalid.
- **What is the impact if Fail of normality assumption**: Typically least important. Effects validity of confidence intervals and test results when the sample size n is small.
- **What to do with outliers**: The first thing to do is check that the data are correctly recorded. If data cannot be corrected, try refitting regression with outliers removed, but still investigate the cause of outliers - may be very important.
- **Estimate of error variance**:

$$S_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}(\hat{e}_i^2) = \frac{RSS}{n-2}$$

$$where\ RSS = \sum_{i=1}^{n}\hat{e}_i^2\ is\ the\ residual\ sum\ of\ squares.$$

- **Degree of freedom for SLR's CI**: v = n - 2 because there're two parameters.
- **What is the multiplier for SLR's CI**:

$$t = \frac{estimate - null}{std.error}$$

- **What is the SE for SLR's CI**:

$$s_{\hat{\beta}_1} = \frac{s_e}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

- **Using R to find SLR's CI**:

```
> confint(model1)
                 2.5 %     97.5 %
(Intercept) 24.8300345 62.4009555
X            0.2557407  0.7284774
```

- $\beta_1 = 0$ **indicates what**: That the response is not (linearly) related to the predictor. So the estimated slope will (almost) always be non-zero: $\hat{\beta}_1 \neq 0$.

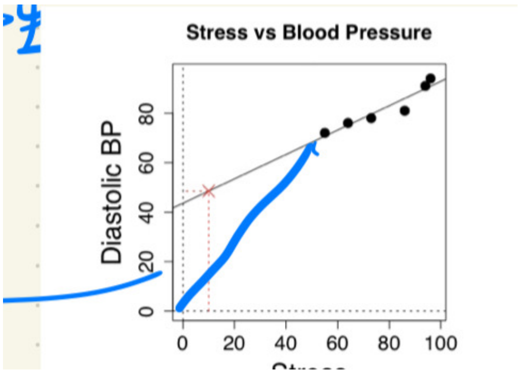
- **Steps to test to assess the strength of evidence in the data for** $\beta_1 \neq 0$:
  - Setting up the hypotheses: $H_0 : \beta_1 = 0$, $H_A : \beta_1 \neq 0$.
  - Calculating The test statistic

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$$

  - Computing the p-value
  - Draw conclusion with rejecting or not H_0

- **When predicting the data**: Ignore e_0.

- **Why not recommend extrapolating when predicting data**: The plot may not be linear.



- **Prediction error**: The prediction error is analogous to a standard error, but takes account of both sources of uncertainty. For prediction at x_0, the prediction error is:

$$PE(\hat{y}_0) = s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

- **Prediction interval formula**:

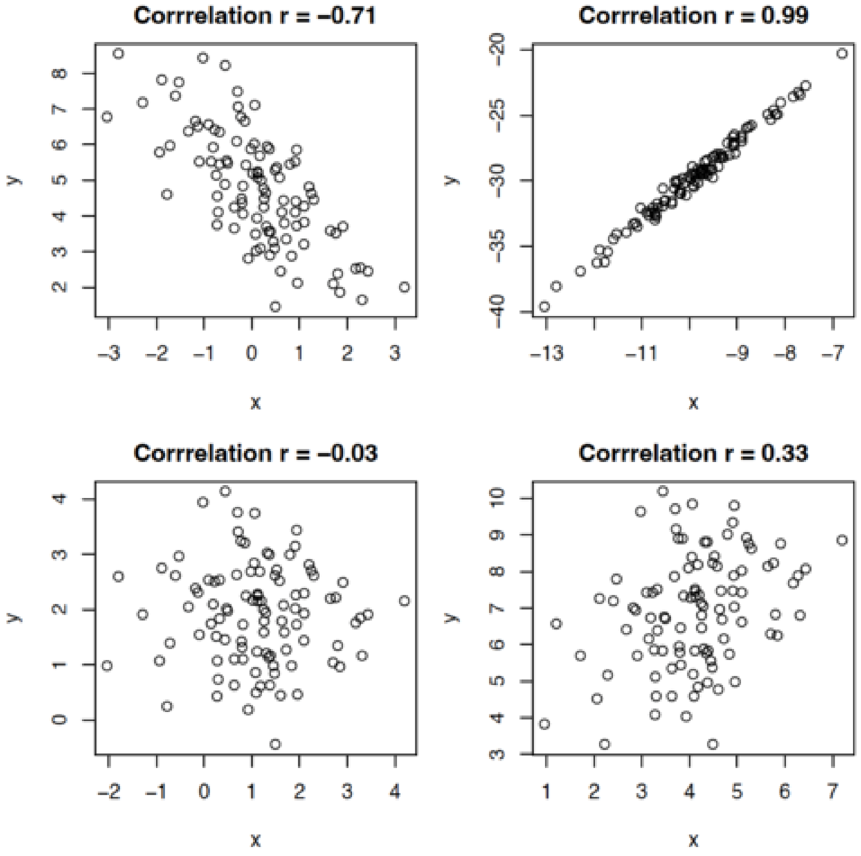
$$\hat{y}_0 \pm t_{(1-\frac{\alpha}{2},n-2)} \times PE(\hat{y}_0)$$

- **Correlation coefficient (r)**: Summarises the strength of a linear relationship between variables. It is a measure of linear association between variables. It describes both the strength and direction of the relationship.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

  - **r** $\in [-1, 1]$. A positive value of r means that Y and X increase together. A negative value of r means that as X increases, Y decreases (and vice-versa).
  - The strength of the linear relationship increases as r tends towards 1 or -1. r = 0 corresponds to no linear relationship between the variables.

- **Scatterplots for r**:



Corrrelation r = –0.71

Corrrelation r = 0.99

Corrrelation r = –0.03

Corrrelation r = 0.33

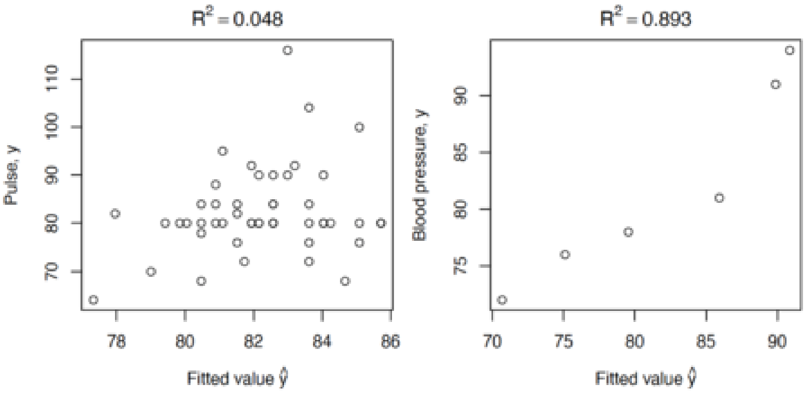
- **Re-write for r**: s_x and s_y are sample standard deviations for x and y variables. s_{xy} is the sample covariance between x and y.

$$= \frac{S_{xy}}{S_x S_y}$$

- **Correlation coefficient versus regression models**: The correlation coefficient is a summary of the data. Unlike linear regression, the correlation coefficient does not specify a model for the data, and cannot (for example) be used for prediction. The correlation coefficient is symmetric in the variables. That is, the correlation between x and y is the same as the correlation between y and x. In regression, the variables are not handled symmetrically. Regression models look at variation in Y for fixed values of x.

- **Coefficient of determination (R^2)**: R^2, is a measure of how well a regression model describes the data. R^2 is the squared correlation between the observed and predicted responses. $R^2 \in [0, 1]$.

- **Meaning for the value of R^2**: A high value of R^2 (close to 1) indicates a regression model that describes the data very well. Conversely, a low value of R^2 (close to 0) indicates a regression that describes the data poorly.



- **What describes the overall variation in the response variable?**: Total sum of squares.
$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

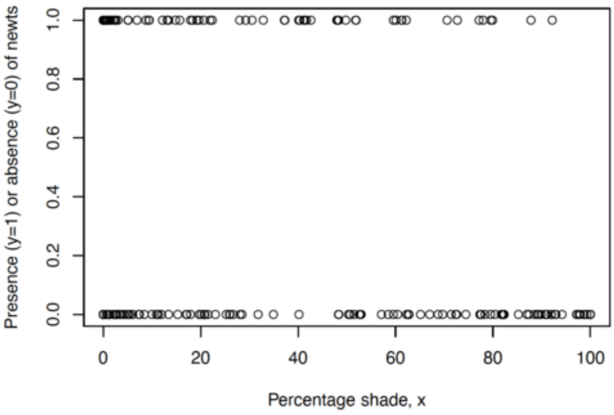
- **What describes the total variation of the data points about the regression line?**: Residual sum of squares (RSS can be thought of as variation not explained by the regression model).
$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- **What describes as the amount of variation in the response that is explained by the regression model?**: Explained sum of squares.
$$ESS = TSS - RSS$$

- **Equation of R^2**:
$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

- **Correlation does not equal causation**: e.g., just because there's more ice cream in the summer and more drowning in the summer doesn't mean there's a link between ice cream and drowning.

- **Logistic regression**: Outcome variable is binary.



- **Equation for logistic regression**:

$$logit(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

  Y is the binary outcome variable, $Y = 1$ or $Y = 0$ for each observation. p is the probability that specified category will occur; i.e. $p = Pr(Y = 1)$. x is the explanatory variable. Parameters β_0, β_1 are the regression coefficients. β_0 is intercept and β_1 slope 'on the logit scale'. In the formula, log is the natural logarithm (log to base e).

- **Which technique do we use when estimating the regression coefficients?**: Maximum likelihood estimation.

- **What will increase x by one unit result in?**: A multiplicative change of e^{β_1} to the odds.

- **Formula for logistic curve for the probability p**:

$$p = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

- **Testing in logistic regression**:
  - Define the hypotheses: $H_0 : \beta_1 = 0$ and $H_A : \beta_1 \neq 0$.
  - The test statistic is:

$$z = x\frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$$

  - Get the corresponding p-value.
  - Reject/not reject H_0.
  - Conclusion.

- **Multiple regression model**:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + e$$

- **Mean value of the Multiple regression model**:

$$\mu_Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

- **Applications of multiple regression**:
  - Adjusting for the effect of confounding variables.
  - Establishing which variables are important in explaining the values of the response variable.
  - Predicting values of the response variable.
  - Describing the strength of the association between the response variable and the explanatory variables.
- **Least squares estimates**:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- **RSS for $\hat{e}_i$**:

$$RSS = \sum_{i=1}^{n} \hat{e}_i^2$$

To estimate the error variance σ_e^2.

- **Usual estimate**:

$$s_e^2 = \frac{RSS}{n-k-1}$$