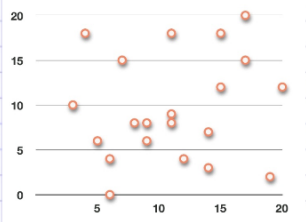
Section 8: REGRESSION

Scatterplots and regression

Scatterplot, scattergraph, scatter diagram: A plot of the data points in a set that compares two variables about the data at the same time



Approximating curve: The curve that approximates the shape of the points in the scatterplot

Curve fitting: The process of finding the equation of the approximating curve

Regression line, best-fit line, line of best fit, least-squares line:
The line that best shows the trend in the data given in a scatterplot

$$\hat{y} = a + bx$$

$$b = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

$$a = \frac{\Sigma y - b\Sigma x}{n}$$

Positive linear relationship: When the regression line has a positive slope

Negative linear relationship: When the regression line has a negative slope

Strong linear relationship: When the data is tightly clustered around the regression line

Weak linear relationship: When the data is loosely clustered around the regression line

Regression: The process of estimating the value of the dependent variable from a given value of the independent variable

Correlation coefficient and the residual

Correlation coefficient: Tells us the strength of the relationship between two variables. The correlation coefficient will have a value on the interval $[-1, 1]$; a value close to $r=-1$ indicates a strong negative relationship, a value close to $r=1$ indicates a strong positive relationship, and a value close to $r=0$ indicates no linear relationship.

$$r = \frac{1}{n-1}\Sigma\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

Residual: The residual e is the actual and predicted (based on the regression line) values for the same data point

Coefficient of determination and RMSE

Coefficient of determination, r^2: Measures the percentage of error we eliminated by using least-squares regression instead of just $\bar{y}$

Root-mean-square error (RMSE), root-mean-square deviation (RMSD):
The standard deviation of the residual

Chi-square tests

Types of χ^2-tests: A χ^2-test for homogeneity, a χ^2-test for association/independence, and a χ^2 goodness-of-fit test

Conditions for inference for a χ^2-test: Random, large counts, independent, categorical