Cheat Sheet: Modeling and Analysis of Bivariate Data

Essential Concepts

- Bivariate statistics are statistics that measure the relationship between two variables.
- Scatterplots visually show the relationship (or lack of a relationship) between two quantitative variables.
- Trends:
    - A scatterplot shows a positive trend if the response variable (represented on the vertical axis) tends to increase as the explanatory variable (represented on the horizontal axis) increases.
    - If the response variable tends to decrease as the explanatory variable increases, then the scatterplot shows a negative trend.
    - Two variables are not associated if knowing the value of one variable does not give you any information about the other variable.
- The correlation is a measure of the strength of the linear relationship between quantitative variables. A scatterplot with low correlation may still present a strong nonlinear relationship.
- The relationship between two variables is said to be linear when the points on the scatterplot resemble a straight line.
- A statistic for measuring the strength and direction of the linear relationship between two quantitative variables is the Pearson Correlation Coefficient, r.
    - The correlation coefficient, r, is always between -1 and 1
    - The correlation coefficient has no unit and is independent of response and explanatory variables
    - The sign of the correlation coefficient describes the direction of the association
    - The values of the correlation coefficient describes the strength of the linear association between the response and explanatory variables

- The following table illustrates the strength of linear relationships:

Correlation Coefficient, r	General Interpretation
-1 to -0.7	Strong negative linear relationship
-0.7 to -0.3	Moderate negative linear relationship
-0.3 to -0.1	Weak negative linear relationship
-0.1 to 0.1	Negligible or no linear relationship
0.1 to 0.3	Weak positive linear relationship
0.3 to 0.7	Moderate positive linear relationship
0.7 to 1	Strong positive linear relationship

- Association does not imply causation. Do not interpret a high correlation between explanatory and response variables as a cause-and-effect relationship.
- Outliers appear as departures from the general trend, i.e. extreme observations in a bivariate data.
- The explanatory variable (x) is the variable that is thought to explain or predict the response variable of a study.
- The response variable ($\widehat{y}$) measures the outcome of interest in the study. This variable is thought to depend in some way on the explanatory variable. It is often referred to as the "variable of interest" for the researcher. The explanatory variable is used to predict/calculate/determine the response variable.
- The Least Squares Regression (LSR) analysis is a statistical method used to make predictions about missing observations in bivariate data. It can also be described as linear modeling.
- The line of best fit, $\widehat{y} = a + bx$, is simply the best line that describes the data points. The line of best fit is also called the Least Squares Regression Line (LSRL).

- The estimated slope b tells us the predicted change in $(\hat{y})$ given a one-unit increase in the value of the explanatory variable x.
    - Interpretation: For every one (unit) increase in (explanatory variable units), we predict an average increase/decrease of ___ (response variable units) in (response variable).
- Within the range of the explanatory variable, we can use the line of best fit to make predictions.
- Extrapolation is the prediction of a response value using an explanatory variable value that is outside the range of the original data.
- The Coefficient of Determination, R^2 or r^2, is the proportion of the variation in the response variable that can be explained by its linear relationship with the explanatory variable. The coefficient of determination is the square of the correlation coefficient. R^2 should be interpreted as a percentage.
- The residual for a data point is the difference between the observed value of the response variable and the linear model's prediction. It is the difference between the observed and predicted values. It is the vertical distance between the observed and predicted values.
- In fitting a regression line, an outlier can also be an observation that does not fit the trend of the data as well. The influential point is the outlier does not fit the trend of the data.
- The residual standard error, s_e, is a measure of the variability in the residuals, which quantify the spread of the points around the line of best fit on the scatterplot.

Key Equations

estimated slope

$b = r\frac{S_y}{S_x}$, where S_y and S_x are the sample standard deviations for the response and explanatory variables and r is the correlation coefficient for the dataset

estimated y-intercept

$a = \bar{y} - b\bar{x}$, where $\bar{y}$ and $\bar{x}$ are the sample means for the response and explanatory variables

line of best fit

$\hat{y} = a + bx$, where $\hat{y}$ is the general predicted value of the response variable (*pronounced* y-hat), a is the estimated value of the y-intercept, and b is the estimated slope

Pearson correlation coefficient, r

$$r = \frac{\sum(\frac{x-\bar{x}}{s_x})(\frac{y-\bar{y}}{s_y})}{n-1}$$

residual

$y - \hat{y}$

residual standard error

$$s_e = \sqrt{\frac{1}{n-2}(y_i - \hat{y}_i)^2}$$

Glossary

a

The estimated value of the y-intercept of the line best fit.

b

the estimated slope or the constant rate of change of the line best fit

bivariate data

two variables linked because both observations are measured from the same individual or unit

coefficient of determination, R^2 or r^2

proportion of the variation in the response variable that can be explained by its linear relationship with the explanatory variable

explanatory variable

the variable that is thought to explain or predict the response variable of a study

extrapolation

prediction for values of the explanatory variable that fall outside the range of the data

influential point

a point that drastically changes the equation of the line, consequently increasing the values of all of the residuals

linear

points on the scatterplot resemble a straight line

negative trend

if the response variable (represented on the vertical axis) tends to increase as the explanatory variable (represented on the horizontal axis) decreases

non-linear

can appear scattered about a smooth curve or have no patterns at all

outlier

data that appears as departures from the general trend

positive trend

if the response variable (represented on the vertical axis) tends to increase as the explanatory variable (represented on the horizontal axis) increases

r

Pearson Correlation Coefficient of two quantitative variables

residual

the vertical error associated with each data point from the line best fit

residual standard error

the measure of the variability of the residuals

response variable

measures the outcome of interest in the study

scatterplots

show the relationship (or lack of a relationship) between two quantitative variables

x

the explanatory variable of the bivariate data

$\hat{y}$

the predicted value of the response variable