Chapter 1

Two-Variable Quantitative Data

A **response** (*dependent*) **variable** measures the outcome while a **explanatory** (*independent*) **variable** explains the data.
The term **bivariate data** refers to data with two variables that are recorded for the same set of *individuals*.
A **scatterplot** is, as the name suggests, a set of points, each representing an individual, scattered about a grid with x and y axes that correspond to the response and explanatory variables respectively.[1] They are the most effective way for the relationship between two quantitative variables measured on the same individuals.

When storing data within lists, it is crucial that each element x_n and y_n correspond to the same individual.

A scatterplot can be described by its *form*, *direction*, and *strength*.
Form describes the nature (shape) of the variable's relationships (linear, polynomial, root, exponential, logarithmic, sinusoidal, etc.). More generally, form can be described as linear or nonlinear. **Strength** describes how strongly correlated the variables are.
Association direction describes whether a positive change in the explanatory variable results in an increase (positive) or decrease (negative) in the explanatory variable.

1.1 Correlation

The **correlation coefficient** $\boldsymbol{r}$ is equal to the sum of the products of the z-scores of each variable for each individual divided by one less than the number of individuals.

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

The magnitude of r corresponds to strength while its sign indicates the *association direction*.

[1] To plot the data of two lists as a scatterplot, `stat plot` (`2nd/stat plot`)) can be pressed, and then a plot selected and toggled `On`, its first option selected for type, and the `X` and `Ylist` set to their corresponding lists. To cycle through individuals in `Xlist` index order, it `trace` can be selected.

The following should be noted regarding r:

- It is the same for x against y and y against x (making it *commutative*).
- It requires that both variables be quantitative.
- It is independent of the units of the variables.
- Its magnitude cannot exceed 1.
- It only measures the strength of *linear* relationships.

  The correlation may be 0 even when there is a strong nonlinear pattern.

- It is not *resistant*, as its formula includes non-resistant means and standard deviations, meaning that it is prone to being affected by outliers.
- It does not completely describe bivariate data.

  A strong correlation alone is not enough to ensure linearity.

- It is not affected by *linear transformations* of either variable, even if they are not transformed correspondingly.

1.2 Regression and Residuals

The **line of best fit** of a scatterplot is the line that best predicts the data. A method for finding a linear line of best fit is **least-squares regression**.[2] The line generated by this method is referred to as the **least-squares regression line** or **LSRL**.

The LSRL enables predictions to be made regarding how the response variable when the explanatory variable is changed. It is defined as the line the minimizes the sum of the squares of the **residuals** (denoted $\boldsymbol{e}$), the differences between the predicted and actual values of the response variable.[3]

$$e_i - y_i - \hat{y}_i$$

A **residual plot** shows the residuals of each data point. A random pattern suggests linearity.[4]

The regression line of a residual plot is always $\hat{y} = 0$, as the sum of all residuals is equal to 0. This also means that the mean of the residuals is equal to 0.

Due to the fact that it uses a sum of every point's values, the LSRL is not resistant to **influential observations**. Influential observations are often outliers in the y direction, especially in smaller data sets. To identify whether a point is an outlier, a modified box plot can be created. A point is influential if removing it markedly changes r.

The predicted value for the explanatory variable is denoted by putting a "hat" over the variable. The LSRL can be described as a linear transformation that maps x onto y:

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

[2]On a calculator, a linear regression can be carried out by entering the data into two corresponding lists. From there, `LinReg(ax+b)` (`stat/CALC/4`) with the appropriate `X` and `Ylist`s entered. This also calculates the r value. To store the equation, `Store RegEQ` can be set to $\texttt{Y}_n$. Before a regression is run, though, `DiagnosticOn` (`2nd/catalog/DiagnosticOn`) should be selected.

[3]A list of residuals can be created by setting $\texttt{L}_n$ to `RESID` (`2nd/list/7`), though the regression should first be run.

[4]This can be found on a calculator by setting `Ylist` to `RESID` (`2nd/list/7`) when creating a scatterplot. This should be done after running a regression.

Points with large residuals are *outliers*.
Predictions made within the data are **interpolations** while those made outside the bounds of the data are **extrapolations**.

Extrapolation should be avoided, as it is more likely not to be accurate than interpolation.

a and b can be calculated as such:[5]

$$a = \bar{y} - bx \qquad\qquad b = r\frac{s_y}{s_x}$$

The definition of a means that the LSRL always passes through $(\bar{x}, \bar{y})$.
The **coefficient of determination r^2** is the percentage of the change/variation in the response variable that can be explained by the LSRL relating the response variable to the explanatory variable. It is maximized by the LSRL.[6]
Clusters are groups of points that are similar.
To add a categorical variable, the data can be displayed using different symbols for each group.

1.3 Transformations to Achieve Linearity

The form of a relationships is not always linear, but the LSRL calculates a linear line of best fit. To resolve this, x and/or y can be transformed using powers, roots, or logarithms. This process is known as **linearization**.
The axes of the scatterplot are changed in accordance with the linearization.
An **exponential model** takes the form of an exponential function.

$$\hat{y} = ab^x$$

[5] a and b can be derived by differentiating the sum of the squares of the residuals.

$$E = \sum(y_i - \hat{y})^2 = \sum(y - a - bx_i)^2$$

Because the error is being minimized, its derivatives are equal to 0.

$$\begin{aligned}
\partial_a E &= \partial_a \sum(y_i - a - bx_i)^2 \\
0 &= \sum 2(y_i - a - bx_i)(\partial_a[y_i - a - bx_i]) \\
&= \sum(y_i - a - bx_i)(-1) \\
&= \sum y_i - \sum a - \sum bx_i = \sum y_i - an - b\sum x_i \\
a &= \frac{\sum y_i - b\sum x_i}{n} = \bar{y} - b\bar{x}
\end{aligned}$$

$$\begin{aligned}
\partial_b E &= \partial_b \sum(y_i - a - bx_i)^2 \\
0 &= \sum 2(y_i - a - bx_i)(\partial_b[y_i - a - bx_i]) \\
&= \sum(y_i - a - bx_i)(-x_i) = \sum(x_i y_i - x_i\bar{y} + b\bar{x}x_i - bx_i^2) \\
&= \sum(x_i y_i - x_i\bar{y}) - b\sum(x_i^2 - \bar{x}x_i) \\
b &= \frac{\sum(x_i y_i - \bar{y}x_i)}{\sum(x_i^2 - \bar{x}x_i)} = \sum\left(\frac{y_i - \bar{y}}{x_i - \bar{x}}\right) = \sum\left(\frac{(x_i - \bar{x})(y_i - \bar{y})}{(x_i - \bar{x})^2}\right) \\
&= \frac{1}{s_x(n-1)}\sum\left(\frac{(x_i - \bar{x})(y_i - \bar{y})}{s_x}\right) \\
&= \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) \times \frac{s_y}{s_x} = r\frac{s_y}{s_x}
\end{aligned}$$

[6] Given no information regarding x, the predicted value of y is simply $\bar{y}$, so the residual will be the difference between y and $\bar{y}$. The sum of all residuals must therefore be 0.

$$\sum e_i = \sum(y_i - \bar{y}) = \sum y_i - \sum \bar{y} = \sum y_i - n\bar{y} = \sum y_i - \frac{n\sum y_i}{n} = -\sum y_i - \sum y_i = 0$$

Because the sum of the residuals is 0, the sum of the squares of the residuals must instead be used to compare errors between approximations. r^2 is equal to the percentage of error removed by using the LSRL rather than $\hat{y} = \bar{y}$.

A **logarithmic model** takes the form of a natural logarithmic function.

$$\hat{y} = a + b \ln x$$

Exponentiating one variable and taking the logarithm of the other are equivalent operations.

A **power model** takes the form of a polynomial with direct variation.

$$\hat{y} = ax^b$$

Taking the logarithm of both sides simply results in a power model after it is re-expressed.

$$\begin{aligned} \ln \hat{y} &= a + b \ln x \\ e^{\ln \hat{y}} &= e^{a + b \ln x} \\ \hat{y} &= e^a e^{\ln x^b} = cx^b \end{aligned}$$

The model that bests fits the data is the one that minimizes s, the standard deviation of the residuals (the **root-mean-square deviation or RMSD**:[7]

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

[8]

[7]The formula for the standard deviation of the residuals can be derived as such:

$$s_e = \sqrt{\frac{\sum (e_i - \bar{e})^2}{n-1}} = \sqrt{\frac{\sum (y_i - \hat{y}_i - 0)^2}{n-1}}$$

[8]The following Minitab table is often provided and can be interpreted as such:

Predictor	Coef	SE Coef	T	P
Constant	a	$s_a = \sqrt{s\left(\frac{1}{n} + \frac{\bar{x}^2}{s_x\sqrt{n-1}}\right)}$	$t_a = \frac{a}{s_a}$	$P(\lvert T\rvert > \lvert t_a\rvert)$
[Explanatory Variable]	b	$s_b = \frac{s}{s_x\sqrt{n-1}}$	$t_b = \frac{b}{s_b}$	$P(\lvert T\rvert > \lvert t_b\rvert)$
S = s	R-Sq = r^2		R-Sq(adj) = r^2 excluding influential points	

Information regarding the third, fourth, and fifth columns is found in Chapter 12.

Chapter 2

Two-Variable Categorical Data

Segmented bar graphs have one bar per variable. Each bar is divided into segments according to their *relative frequency*. The height of each bar is equal to 1.
Side-by-side bar graphs show the distribution of one categorical variable for each value of another. The grouping of the bars is based on one of the categorical variables while the bars themselves show the frequencies of the values of the other.
Mosaic plots are variants of segmented bar graphs that display the number of individuals in a category by making each bar's width proportional to it.
A **two-way table** shows data on the relationship between two categorical variables for some group of individuals.
There are 3 types of *relative frequencies*:

1. **Marginal relative frequency** is the proportion of individuals with a specific value for one categorical variable.
$$\text{marginal relative frequency} = \frac{n}{N}$$

2. **Joint relative frequency** is the proportion of individuals with a specific value for one categorical variable as well as a specific value for another categorical variable.
$$\text{joint relative frequency} = \frac{n_{A \wedge B}}{N}$$

3. **Conditional relative frequency** is the proportion of individuals that have a specific value for one variable that also have a specific value for another variable.
$$\text{conditional relative frequency} = \frac{n_{A \wedge B}}{n_B}$$

If knowing the value of one variable helps in predicting that of another, there is an *association* between them. To test for association, the marginal and conditional relative frequencies can be compared. If they vary *significantly*, there is an association.