1. Introduction

- **Simple linear regression model (population)**: $Y = \beta_0 + \beta_1 x + \varepsilon$
- **Simple linear regression (sample)**: $\hat{y} = b_0 + b_1 x$
- **Least squares criterion**: line that best fits set of data points is the one that has the smallest sum of squared errors. Errors are vertical distances of data points to line.
- **Sum of squares error**:

$$\sum e_i^2 = \sum \left(y_i - \hat{y}_i\right)^2$$

- **Regression line**: fits set of data points according to least squares criterion.
- **Regression equation**: equation of regression line given n data points:

$$b_1 = \frac{X_{xy}}{S_{xx}} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}, \quad b_0 = \overline{y} - b_1 \overline{x}$$

- For linear regression, interpolation is reasonable but extrapolation may not be.
- **Influential observation**: data point whose removal would cause the regression line to change considerably. It affects the robustness of the model.
- **Total variation** in observed values of response variable:

$$\mathrm{SST} = \sum (y_i - \overline{y})^2$$

- Variation in observed values of response variable explained by regression:

$$\mathrm{SSR} = \sum \left(\hat{y}_i - \overline{y}\right)^2$$

- Variation in observed values of response variable not explained by regression:

$$\mathrm{SSE} = \sum \left(y_i - \hat{y}_i\right)^2$$

- **Coefficient of determination (R^2)**: proportion of variation in observed values of response variable explained by regression:

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = \frac{\mathrm{SST} - \mathrm{SSE}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}$$

  R^2 measures the utility of the regression equation for making a prediction. We have $R^2 \in [0, 1]$, if close to 0 then not useful for predictions, if near 1 then useful for predictions.
- **Notation**:

$$S_{xx} = \sum (x_i - \overline{x})^2 = \sum x_i^2 - n\overline{x}^2$$
$$S_{xy} = \sum (x_i - \overline{x})(y_i - \overline{y}) = \sum x_i y_i - n\overline{x}\,\overline{y}$$
$$S_{yy} = \sum (y_i - \overline{y})^2 = \sum y_i^2 - n\overline{y}^2$$

- For $(x_1, y_1), ..., (x_n, y_n)$, R^2 is the square of the sample correlation coefficient.

- **Adjusted $\boldsymbol{R^2}$**: modification of R^2 which accounts for number of independent variables, k. In simple linear regression, $k = 1$. Adjusted R^2 only increases when a significant related independent variable is added to model.

$$\text{Adjusted } R^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

  This penalises having more predictors.