Linear Regression

Simple Linear Regression

- Consider the following three scenarios:
  - The CEO of the local Tourism Authority would like to know whether a family's annual expenditure on recreation is related to their annual income. This information could be used to tailor marketing campaigns to certain consumer segments.
  - A food company is interested in determining a shelf-life for a new chilled food product and hence they would like to quantify the relationship between microbial activity and time.
  - A car buyer is interested in purchasing a second hand car and would like to ascertain the relationship between a car's age and advertised purchase price.

Simple Linear Regression

- What do these three scenarios have in common?
  - The answer is that they all involve **the quantification** of the relationship between two variables.
  - Deterministic Relation and statistical relationship
  - This relationship is of interest as it allows us
    - To gain an understanding of the problem
    - To make predictions
    - To assess new data in light of the relationship
- Applications:
  - Business
  - Genetics
  - Food Science, etc

Simple Linear Regression

- Data:
  - Suppose you are a consultant to the local Tourism Authority and the CEO of the Authority would like to know whether a family's annual expenditure on recreation is related to their annual income.
  - In addition, if there is a relationship, he would like you to build a statistical model which quantifies the relationship between the two variables
  - A data set consisting of a random sample of 20 families, collected last year is available to help you with the assessment.
  - two quantitative variables → family's annual expenditure on recreation and annual income
  - Dependent Variable (Y) → Expenditure
  - Independent Variable (X) → Income

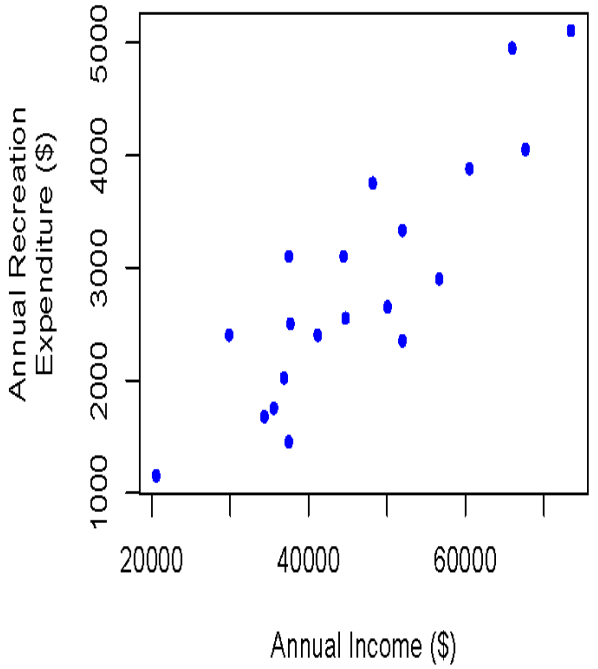
Simple Linear Regression

Expenditure ($)	Income ($)	Exp. ($)	Inc. ($)
2400	41200	1450	37500
2650	50100	2020	36900
2350	52000	3750	48200
4950	66000	1675	34400
3100	44500	2400	29900
2500	37700	2550	44750
5106	73500	3880	60550
3100	37500	3330	52000
2900	56700	4050	67700
1750	35600	1150	20600

Table 9.1: Data on the annual expenditure on recreation and annual income of 20 families. The data can be found in *tourism.csv*.

Simple Linear Regression

- Graphical Summary of Data: Scatter plot → relation between two variables
  - Dependent Variable (Y) → vertical axis
  - Independent Variable (X) → Horizontal Axis
  - A scatter plot is frequently also referred to as a plot of Y versus X.
  - Relation between income and expenditure:
    - Direction: Positive → as income increases so does recreation expenditure;
    - Shape: Roughly linear → the points appear to fall along a straight line
    - Strength: Reasonably strong → there is considerable scatter about a straight line.



Simple Linear Regression

- Numerical summary of the data — Correlation:
  - A numerical summary of the strength of the association between the two variables is often desired. → population correlation coefficient, ρ. → it measures which measures the strength of the linear association between two variables X and Y.
  - Since X and Y are quantitative variables, ρ is also known as Pearson correlation coefficient.
  - The sample correlation coefficient r can be as follows

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

ẋ and ẏ are sample means and s_x and s_y are standard deviations

Simple Linear Regression

- Numerical summary of the data — Correlation:
  - The ρ can take values between −1 and 1 and the interpretation of ρ is as follows
    - A negative value indicates a decreasing relationship between X and Y , that is, as X increases, Y decreases.
    - A positive value indicates an increasing relationship between X and Y , that is, as X increases, so does Y .\
    - A value of 0 indicates that there is no linear relationship between the two variables — this however does not imply that there is no relationship.
    - **The correlation does not give an indication about the value of the slope of any linear relationship**.
  - Whole population can not be measured → only few samples can be measured.

Simple Linear Regression

- The Simple Linear Regression Model:
  - "What is the equation of the linear relationship between the explanatory and response variables?"
  - Quantifying the relationship will allow us to
    - better understand the functional relationship between X and Y → how quickly does Y increase for every unit increase in X
    - **make predictions about Y for a new value of X**.
  - The linear regression has the form:
  - Where Y denotes the dependent variable
  - X denotes the independent variable;
  - β_0 denotes the y-intercept;
  - β_1 denotes the slope of the line; and
  - E denotes a random error.

$$Y_i = \beta_0 + \beta_1 X_i + E_i$$

Simple Linear Regression

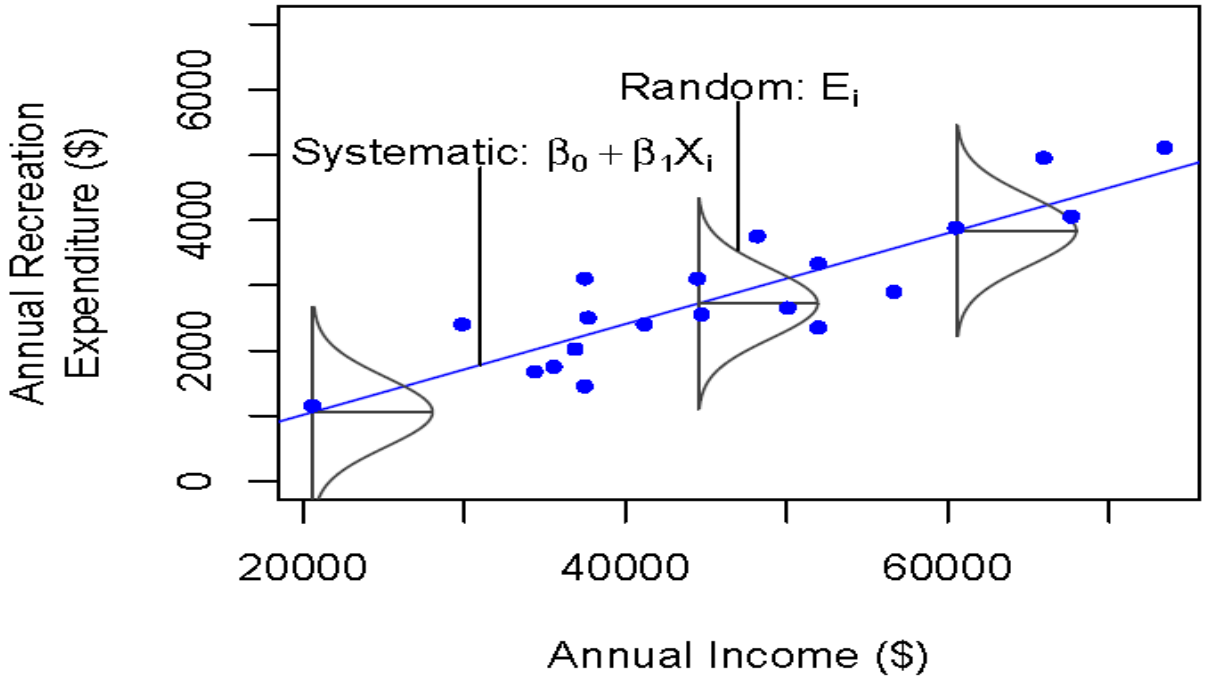


Figure 9.2: Graphical representation of a regression model overlayed on the Tourism data.

Simple Linear Regression

- Any straight line fitted to the data will take the form $y = b_0 + b_1 x$
- where b_0 and b_1 denote the intercept and the slope of the fitted line.
- How good this line fits the data is determined by how close the data points (x_i, y_i) are to the corresponding points on the line, which are given by $\hat{y}_i = b_0 + b_1 x_i$
  - for a given value $\mathbf{x_i}$, the observed value of the dependent variable is $\mathbf{y_i}$ and the corresponding prediction from the line is **ˆyi**
  - The residual is calculated as the difference between the two is $r_i = y_i - \hat{y}_i$
  - If the residual is negative then the observed value lies below the prediction
  - If the residual is positive then the observed value lies above the prediction.
  - The residuals need to be combined in some way to give an overall measure of how well the particular line fits the data
  - positive and negative residuals need add to the lack of fit.

Simple Linear Regression

- A commonly used measure of lack-of-fit is the Residual Sum of Squares, RSS, which is also known as Sum of Squared Errors $\text{RSS} = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
- The best line will result in the smallest RSS from all possible candidate lines $b_0 + b_1x$.
- This smallest RSS indicates that the correspondence between the data and the fitted line is as good as it can possibly be. → this line is called least squares regression (LSR) line and the estimates of the intercept and slope obtained from the LSR line are denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$.
- The RSS is

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - (b_0 + b_1 x_i) \right)^2$$

Simple Linear Regression

- The aim is to find values for b_0 and b_1 such that RSS is minimized. → In order to do this, the partial derivative of RSS with respect to each b_0 and b_1 are taken and equated to zero.

$$\frac{\partial \text{RSS}}{\partial b_0} = \sum_{i=1}^{n} 2\Big(y_i - (b_0 + b_1 x_i)\Big) = 0$$

$$\frac{\partial \text{RSS}}{\partial b_1} = \sum_{i=1}^{n} 2\Big(y_i - (b_0 + b_1 x_i)\Big) x_i = 0$$

- The solution to these equations yields the least squares regression estimates

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

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

- where $\bar{y}$ and $\bar{x}$ denote the mean of y and x.

Linear Regression

- Linear Regression model for the purpose of prediction.
- A form of statistical modeling that attempts to evaluate the relationship between one variable (termed the dependent variable) and one or more other variables (termed the independent variables).
- It is a form of global analysis as it only produces a single equation for the relationship.
- A model for predicting one variable from another.

Linear Regression

- **Example in Class**

Thank you