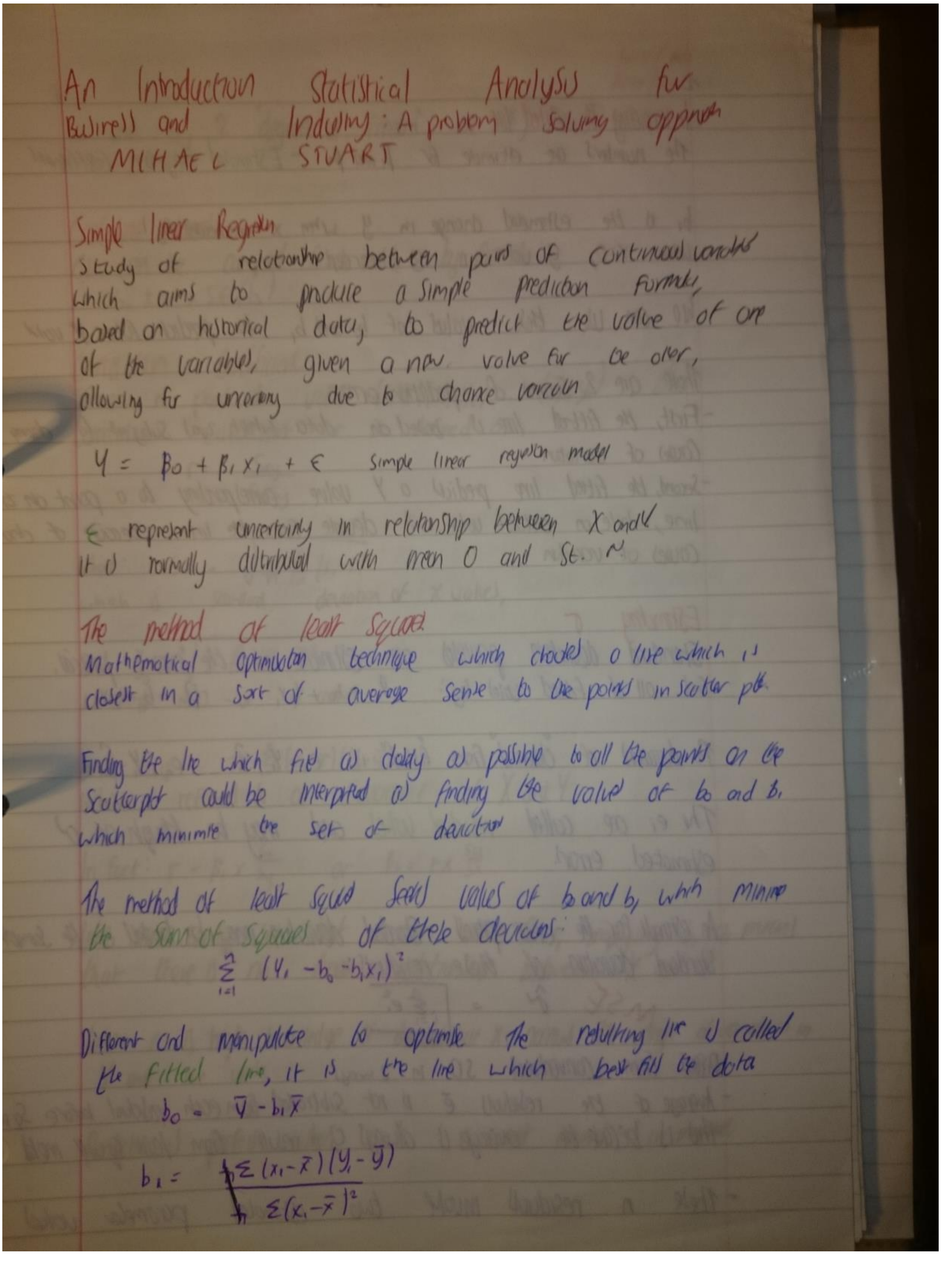

An Introduction Statistical Analysu fur Buiness and Indusmy : A problem Solving approach

MICHAEL STUART

Simple liner Regression

Study of relationship between pairs of continuous variables which aims to produce a simple prediction formula, based on historical data, to predict the value of one of the variables, given a new value for the other, allowing for uncertainty due to chance variation

$Y = \beta_0 + \beta_1 x_1 + \epsilon$ Simple linear regression model

ϵ represent uncertainty in relationship between X and Y
It is normally distributed with mean 0 and St. ~

The method of least Squares

Mathematical optimisation technique which choose a line which is closest in a sort of average sense to the points in scatter plot

Finding the line which fits as closely as possible to all the points on the Scatterplot could be interpreted as finding the value of b_0 and b_1 which minimise the set of deviation

The method of least squares seeks values of b_0 and b_1 which minimise the Sum of Squares of these deviations:

$$\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Differentiate and manipulate to optimise the resulting line is called the Fitted line, it is the line which best fits the data

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2}$$

Interpreting the fitted line

The numbers are estimate for β_1 and β_0 - Estimated Regression Coefficients

b_1 is the estimated change in y when x increased by 1.
When $x=0$, y value is y intercept.

We can use these values of b_0 and b_1 to predict future y values

There are 2 sources of prediction error.

- First, the fitted line is based on data which was subject to chance causes of variation
- Second, the fitted line predicts a Y value corresponding to a point on the line, while an actual value will deviate from the line because of chance causes of variation

Estimating σ

Standard deviation provides an estimate of the "error" process.
Get all of fitted values by: $\hat{y}_1 = b_0 + b_1 x_1 \;\ldots\; \hat{y}_n = b_0 + b_1 x_n$

The deviations of "observed" from "fitted" is $e_1 = Y_1 - \hat{y}_1 \;\ldots\; e_n = Y_n - \hat{y}_n$

The e_i are called residual values and may be thought of as estimated errors.

An estimate for the error process Standard deviation may be based on the sample Standard deviation of these residuals:

$$MSE \quad \hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}$$

Differs from conventional SD in 2 ways

- Average of the residuals $\bar{e}$ is not subtracted from each residual before squaring
This is because the average is always 0, a result from least squares method
- These n residuals involve two estimated parameter values

14-11/[illegible]
14:28/[illegible]

b_0 and b_1, 2 "degrees of freedom" are lost, hence divisor of $n-2$ not n.

Confidence Interval:

CI can be computed to b_0 and b_1 by formula:

$$b_1 \pm 2 \times S.E(b_1)$$

Correlation Coefficient:

Given n pairs of values $(x_1,y_1)(x_2,y_2)\ldots(x_n,y_n)$ the correlation coefficient r is defined as:

$$r = \frac{S_{xy}}{S_x S_y}$$

Where: $S_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$
which is standard deviation of X values,

and $S_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}$
which is S.D. of Y values.

and $S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$
which is referred to as Covariance of X and Y

In fact: $r = \beta_1 \times \frac{S_x}{S_y}$ or $\beta_1 = r \times \frac{S_y}{S_x}$

We immediately see that a zero value for r, implying $b_1=0$, means that there is no linear relationship between X and Y

When r is high, knowledge of the value of X narrows the range of variation in Y and vice versa

Prediction is perfect when $r=1$. r cannot exceed 1 in magnitude