1

# STAT2008/STAT6038

Revision of Simple Linear Regression

## The Simple Linear Regression Model and the Parameter Estimates

2

- The **dependent** (or response) variable is the variable we wish to understand or predict
- The **independent** (or predictor) variable is the variable we will use to understand or predict the dependent variable
- **Regression analysis** is a statistical technique that uses observed data to relate the dependent variable to one or more independent variables

## Goal

3

The objective of regression analysis is to build a regression model (or predictive equation) that can be used to describe, predict and control the dependent variable on the basis of the independent variable

## What regression does

4

- Develops an equation which represents the relationship between the variables.
- Simple linear regression – straight line relationship between y and x (i.e. one explanatory variable)
- Multiple linear regression – "straight line" relationship between y and $x_1$, $x_2$, .., $x_k$ where we have *k* explanatory variables
- Non-linear regression – relationship not a "straight line" (i.e. y is related to some function of x, e.g. log(x))

## Goal

5

- To model the relationship between a response variable (*Y*) and an explanatory variable (*X*) using a straight line model of the form.

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- where $\beta_0$ is the intercept, $\beta_1$ is the slope and $\varepsilon$ is a random error
- The interpretation of the intercept $\beta_0$ is that it is the expected value of *Y* when *X* is equal to zero (this may or may not be an interpretable quantity).
- The interpretation of the slope parameter $\beta_1$ is that when *X* increases by 1, then *Y* increases by an amount $\beta_1$.
- The quantities $\beta_0$ and $\beta_1$ are called the parameters of the regression model. They are fixed, unknown quantities that need to be estimated

## Assumptions

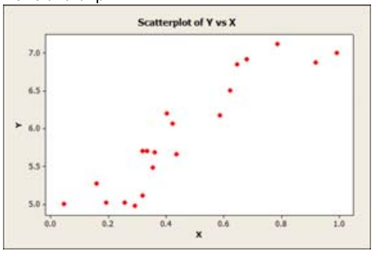
6

- The errors $\varepsilon$ are usually assumed to be independent, zero-mean, constant variance Normal random variables
- so $\varepsilon$ is distributed as a Normal random variate with mean 0 and spread $\sigma$ (standard deviation)

## The key features are:

7

- **Bivariate Fit of Y By X.** A scatterplot – an *X-Y* plot graphically representing the relationship between *X* and *Y*. Here, we are looking for a straight-line relationship




## Key Features

8

- **Linear Fit.** The fitted model is of the form

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

with the two numbers $\hat{\beta}_0$ *and* $\hat{\beta}_1$ given being the estimates of $\beta_0$ and, $\beta_1$ respectively. Note that the fitted line contains no error term $\varepsilon$ since the error random variable is *expected* to be zero.

## Note

9

- In reality predicted values won't be exact
- So, more reasonable is to ask what would be the *expected value* of the dependent variable Y when X=x

## So in general we use

10

- E(Y|X=x)
- If we add in assumption of linearity we get

$$E(Y \mid X = x) = \beta_0 + \beta_1 x$$

- $\beta_0$ and $\beta_1$ are coefficients that determine the straight line

## In general, for any pair of observations

11

$$E(Y_i \mid X = x_i) = \beta_0 + \beta_1 x_i$$

- In practice, the observed value will (almost always) differ from the expected value.
- Denote difference by greek epsilon, $\varepsilon_i$
- Mean of $\varepsilon_i$ will be zero

$$\varepsilon_i = Y_i - E(Y_i \mid X = x_i) = Y_i - (\beta_0 + \beta_1 x_i)$$

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

## $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

12

- Called the population or true regression line
- $\beta_0$ and $\beta_1$ are constants to be estimated
- $\varepsilon_i$ is a random variable with mean = 0
- Interpretation will be in two parts
  - an expectation ($\beta_0 + \beta_1 x_i$) which reflects the systematic relationship, and a discrepancy ($\varepsilon_i$) which represents all the other many factors (apart from X) which may affect Y.

## Estimation

13

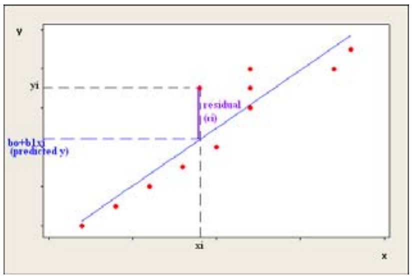
- Is done using a process called "Least Squares Estimation"
- In practice, the population regression line has to be estimated:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
$$\text{is estimated by } \hat{y} = b_0 + b_1 x$$

## How good is the estimate?

14

- Look at the distance between the points $(x_i, y_i)$ and the line.




## Residual

15

- Vertical distance between observed point and fitted line is called the residual.
- That is $r_i = y_i - (b_0 + b_1 x_i)$
- $r_i$ estimates $\varepsilon_i$, the error variable
- Want to determine values of $b_0 + b_1$ that best fit the data – choose the values which minimise the sum of the squared differences between observed and estimated, i.e. choose values of slope and intercept which minimise

$$\sum_{i=1}^{n} (y_i - \hat{y})^2 = \sum_{i=1}^{n} (r_i)^2$$

## This is the least squares method

16

- Choose our estimates of slope and intercept to give the smallest residual sum of squares
- Uses calculus to find estimates

Estimate intercept $\beta_0$ by $b_0$, sometimes use $\hat{\beta}_0$.

Estimate slope $\beta_1$ by $b_1$, sometimes use $\hat{\beta}_1$.

## Estimation

17

- **HOW TO MINIMIZE the sum of the squared errors (SSE)?**

$$choose\ \hat{\beta}_0, \hat{\beta}_1\ \text{to minimise}$$
$$S(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

*Differentiating*

$$A)\quad \frac{\partial S}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)$$

$$B)\quad \frac{\partial S}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i)$$

## Estimation

18

$$\text{At } (\hat{\beta}_0, \hat{\beta}_1), \frac{\partial S}{\partial \beta_0} = \frac{\partial S}{\partial \beta_1} = 0$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$
$$\hat{\beta}_1 = S_{xy} / S_{xx} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

## The regression line

19

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$
$$\overline{Y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{X}$$
$$(\overline{X}, \overline{Y}) \quad \text{is on the fitted line}$$

## The Regression Line

20

$$\text{(ii)}\ \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2}$$

$$r = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 \sum (Y_i - \overline{Y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

$\hat{\beta}_1$ directly reflects how correlated $X$ and $Y$ are!

## Example

21

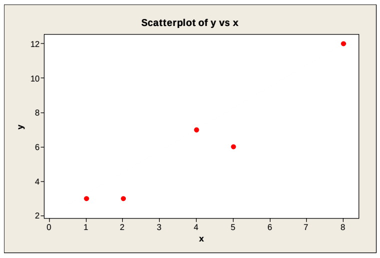
- The investment in certain share portfolios ($x$) and the value after a year ($y$)in \$000 are given in the table below.

| | | | | | |
|---|---|---|---|---|---|
| $x$ | 1 | 2 | 4 | 5 | 8 |
| $y$ | 3 | 3 | 7 | 6 | 12 |

- Fit a regression line to these data by hand

## Best first step – look at a scatterplot!

22




## Want to find $b_0$, $b_1$ for $\hat{y} = b_0 + b_1 x$

23

- Given that cov(x,y)=9.75, var(x)=7.5
- x average=4, y average =6.2

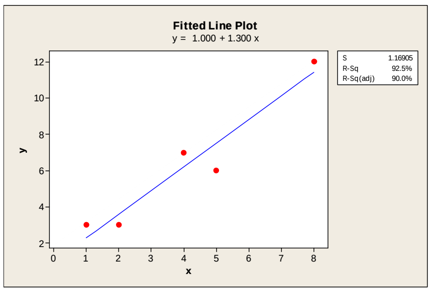
$$\hat{\beta}_1 = b_1 = \frac{\text{cov}(x, y)}{s_x^2} = \frac{9.75}{7.5} = 1.3$$

$$\hat{\beta}_0 = b_0 = \overline{y} - b_1 \overline{x} = 6.2 - 1.3 * 4 = 1$$

## So, least squares regression line is

24

$$\hat{y} = 1 + 1.3x$$



## Can find predicted values and residuals,



| x | y | $\hat{y}_i = 1+1.3x_i$ | $r_i = y_i - \hat{y}_i$ | $(y_i - \hat{y}_i)^2$ |
|---|---|---|---|---|
| 1 | 3 | 2.3 | 0.7 | 0.49 |
| 2 | 3 | 3.6 | -0.6 | 0.36 |
| 4 | 7 | 6.2 | 0.8 | 0.64 |
| 5 | 6 | 7.5 | -1.5 | 2.25 |
| 8 | 12 | 11.4 | 0.6 | 0.36 |

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 4.1$$



- No other choices of $b_0$, $b_1$ will give a smaller SSE – in this sense, it is the "line of best fit"
- Interpretation: $\hat{y} = 1 + 1.3x$
  - x is the investment($000), y is the value after a year ($000)
  - $b_0$=1 is the intercept, $b_1$=1.3 is the slope coefficient
  - Slope: for each extra $1000 invested, value after a year is expected to increase by $1300.
  - Intercept: as model only fitted in range x=1 to x=8, no interpretation can be given (refers to what happens when x=0); model implies that if $0 is invested, value after a year is $1000 – obviously not sensible ➔ demonstrates the danger of extrapolating.
  - General rule – can't determine the value of y for a value of X outside our sample range of x.