

## 04b Sample Examination Problems Chapter 10 SOLUTIONS

---

1. (a) Derive from first principles the least squares estimator of slope for a simple linear regression.

We seek  $\beta$  from first principles

Simple linear model:  $Y_i = \alpha + \beta x_i + \varepsilon_i$

$Y_i$  : single response variable ,  $\alpha$  : intercept

$\beta$  : single slope on a single explanatory variable  $x_i$

$\varepsilon_i$  : residual error

$\alpha$  and  $\beta$  unknown constants ,we need to estimate them.

The question asks to estimate the slope, i.e.  $\beta$

Let the fitted values :  $\hat{Y}_i = A + B x_i$

A is the estimate of  $\alpha$  , B is the estimate of  $\beta$

The residuals are :  $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$

We seek A and B such that to minimize the sum of squares of the residuals.

$$\sum_{i=1}^n \hat{\varepsilon}_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - A - Bx_i)^2 \text{ is minimum}$$

$$\Rightarrow \sum_{i=1}^n (Y_i - A - Bx_i)^2 \leq \sum_{i=1}^n (Y_i - \alpha - \beta x_i)^2$$

$$A = \hat{\alpha} , B = \hat{\beta}$$

Minimise  $S = \sum_{i=1}^n (Y_i - \alpha - \beta x_i)^2$  using partial derivatives ,

$$S_{\alpha} = -2 \sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta} x_i) = 0 \Rightarrow \sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta} x_i) = 0 \Rightarrow \sum_{i=1}^n (Y_i - A - Bx_i) = 0 \quad (1)$$

$$S_{\beta} = -2 \sum_{i=1}^n x_i (Y_i - \hat{\alpha} - \hat{\beta} x_i) = 0 \Rightarrow \sum_{i=1}^n x_i (Y_i - \hat{\alpha} - \hat{\beta} x_i) = 0$$

$$\Rightarrow \sum_{i=1}^n x_i (Y_i - A - Bx_i) = 0 \quad (2) , \text{ which is the same as letting}$$

$$(1) \sum_{i=1}^n \hat{\varepsilon}_i = 0 \text{ and } (2) \sum_{i=1}^n \hat{\varepsilon}_i x_i = 0$$

$$(1) \sum_{i=1}^n \hat{\varepsilon}_i = 0 \Rightarrow \sum_{i=1}^n (Y_i - A - Bx_i) = 0 \text{ dividing both sides by } n : \bar{Y} - A - B\bar{x} = 0$$

$$\Rightarrow A = \bar{Y} - B\bar{x} \text{ substitute this in (2) :}$$

$$(2) \sum_{i=1}^n \hat{\varepsilon}_i x_i = 0 \Rightarrow \sum_{i=1}^n x_i (Y_i - A - Bx_i) = 0$$

$$\Rightarrow \sum_{i=1}^n x_i (Y_i - \bar{Y} + B\bar{x} - Bx_i) = 0 \Rightarrow \sum_{i=1}^n x_i (Y_i - \bar{Y}) = B \sum_{i=1}^n x_i (x_i - \bar{x})$$

$$\Rightarrow B = \frac{\sum_{i=1}^n x_i (Y_i - \bar{Y})}{\sum_{i=1}^n x_i (x_i - \bar{x})} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

The proof of the last result:

$$\frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i (Y_i - \bar{Y}) - \sum_{i=1}^n \bar{x} (Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})}$$

$$= \frac{\sum_{i=1}^n x_i (Y_i - \bar{Y}) - \bar{x} \sum_{i=1}^n (Y_i - \bar{Y})}{\sum_{i=1}^n x_i (x_i - \bar{x}) - \bar{x} \sum_{i=1}^n (x_i - \bar{x})}$$

$$\text{but } \bar{Y} = \frac{\sum_{i=1}^n Y_i}{n} \Rightarrow \sum_{i=1}^n Y_i - n\bar{Y} = 0 \text{ i.e. } \sum_{i=1}^n (Y_i - \bar{Y}) = 0$$

$$= \frac{\sum_{i=1}^n x_i (Y_i - \bar{Y}) - \bar{x}(0)}{\sum_{i=1}^n x_i (x_i - \bar{x}) - \sum_{i=1}^n \bar{x} (x_i - \bar{x})} = \frac{\sum_{i=1}^n x_i (Y_i - \bar{Y})}{\sum_{i=1}^n x_i (x_i - \bar{x}) - \bar{x} \sum_{i=1}^n (x_i - \bar{x})}$$

$$= \frac{\sum_{i=1}^n x_i (Y_i - \bar{Y})}{\sum_{i=1}^n x_i (x_i - \bar{x}) - (0)} = \frac{\sum_{i=1}^n x_i (Y_i - \bar{Y})}{\sum_{i=1}^n x_i (x_i - \bar{x})}$$

- (b) The table below shows the population of England and Wales in millions for years in the 19th century.

Year	1801	1811	1821	1831	1841	1851	1861	1871
Popn.	8.89	10.16	12.00	13.90	15.91	17.93	20.07	22.71

- Find the least squares fit of a regression model for response variable population and explanatory variable year. Give the intercept and slope of the fitted line.
- Should you fit a straight line through (0, 0) to these data rather than allowing an arbitrary intercept?
- How would your fitted regression line change if the population were measured in thousands?

- i. We need a fitted line, i.e. to estimate the intercept and the slope

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

$$A = \bar{Y} - B\bar{x}$$

$$B = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

You may use stats1 formula : 
$$B = \frac{\sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$$

$x_i = \text{Year}$ ,  $Y_i = \text{population}$  :  $n = 8$

$$\sum x_i = 14688, \quad \sum Y_i = 121.57, \quad \sum x_i Y_i = 224032.97$$

$$\sum x_i^2 = 26971368, \quad \bar{x} = 1836, \quad \bar{Y} = 15.19625$$

$$B = \frac{\sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{224032.97 - 8(1836)(15.19625)}{26971368 - 8(1836)^2} = 0.1977$$

$$A = \bar{Y} - B\bar{x} = 15.19625 - 0.1977(1836) = -347.8290$$

The fitted line :  $\hat{Y} = -347.83 + 0.2x$

ii. For the fitted model in (i), if  $x = 0$  (i.e. in Year 0),

The predict value for population :  $\hat{Y} = -347.83$

If we force the line of best fit through the origin i.e. a predicted population value of 0 which doesn't make sense.

But consider the dangers of Extrapolation since we have data only for the period 1801 – 1871 which exhibits a linear relation during this period.

You can see this by a simple scatter plot diagram.

iii. Changing the unit of measurement of the variable does not change the fit of the regression.

2. (a) Find from first principles the least squares estimator for the slope of a line through the origin fitted to  $n$  pairs of values  $(x_i, Y_i)$ .

**Model:**  $Y_i = \beta x_i + \varepsilon_i$ , **intercept is zero**

We seek  $B$  such that to minimize the sum of squares of the residuals.

$$\sum_1^n \hat{\varepsilon}_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - Bx_i)^2 \text{ is minimum}$$

$$\Rightarrow \sum_{i=1}^n (Y_i - Bx_i)^2 \leq \sum_{i=1}^n (Y_i - \beta x_i)^2$$

$$\text{Minimize } S = \sum_{i=1}^n (Y_i - \beta x_i)^2, \frac{dS}{d\beta} = -2 \sum_{i=1}^n (Y_i - \beta x_i) = 0$$

$$\sum_{i=1}^n (Y_i - \beta x_i) = 0 \text{ i.e. } \sum_1^n \hat{\varepsilon}_i x_i = 0 \Rightarrow \sum_1^n x_i (Y_i - Bx_i) = 0$$

$$\Rightarrow \sum_1^n x_i Y_i - \sum_{i=1}^n Bx_i^2 = 0 \Rightarrow \sum_1^n x_i Y_i - B \sum_{i=1}^n x_i^2 = 0$$

$$\Rightarrow B = \frac{\sum_1^n x_i Y_i}{\sum_{i=1}^n x_i^2}$$

(b) The table below shows Regional Manufacturing Capital Stock Estimates in millions of pounds sterling at 1970 prices in Wales and in Scotland.

- i. Find the least squares fit of a regression model for response variable Scotland Capital Stock and explanatory variable Wales Capital Stock.
- ii. Interpret your regression line.

Year	1950	1951	1952	1953	1954	1955	1956	1957	1958
Wales	1116	1162	1219	1256	1316	1381	1426	1500	1563
Scotland	1746	1815	1868	1918	1958	2011	2066	2110	2153

i. **Model 1:**  $Y_i = \alpha + \beta x_i + \varepsilon_i$

$$A = \bar{Y} - B\bar{x}$$

$$B = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$$

$x_i$  = Wales Capital Stock ,  $Y_i$  = Scotland Capital Stock :  $n = 9$

$$\sum x_i = 11939 , \sum Y_i = 17645 , \sum x_i Y_i = 23573840$$

$$\sum x_i^2 = 16024659 , \bar{x} = 1326.56 , \bar{Y} = 1960.6$$

$$B = \frac{\sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{224032.97 - 8(1836)(15.19625)}{26971368 - 8(1836)^2} = 0.897$$

$$A = \bar{Y} - B\bar{x} = 776.96$$

The fitted line :  $\hat{Y} = 776.97 + 0.897x$

$$\text{Model 2 : } Y_i = \beta x_i + \varepsilon_i , B = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2} = \frac{23573840}{16024659} = 1.47$$

The fitted line :  $\hat{Y} = 1.47x$

ii. Model 1 :

For every 1 million increase in Wales , Scotland increases by 0.897 units  
(0.897 million)

If  $x = 0$  then scotland increase by 776.97

Model 2 :

For every 1 million increase in Wales , Scotland increases by 1.47 units  
(1.47 million)

- 
3. (a) Derive from first principles the least squares estimators of intercept and slope for a simple linear regression model.

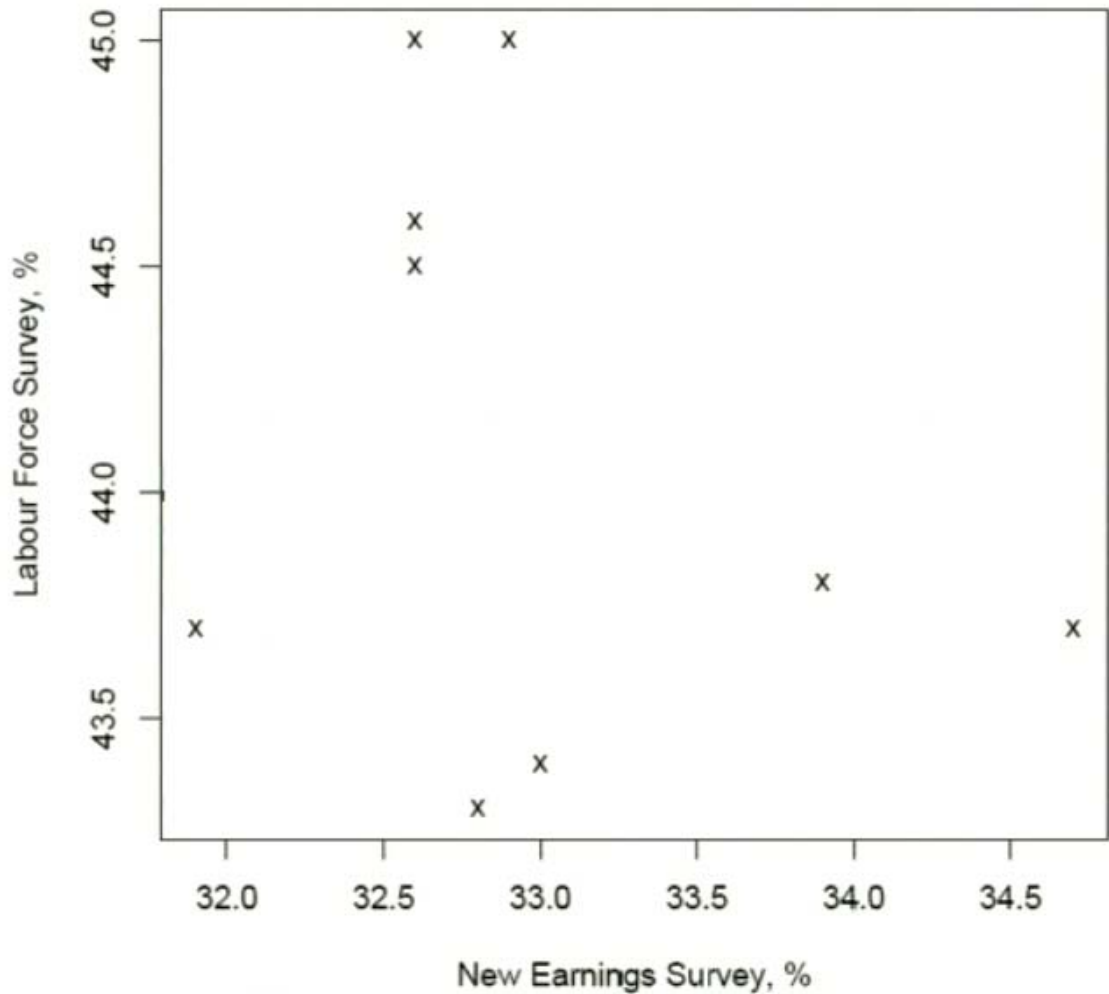
**Refer to problem 1(a)**

- (b) The following table shows the proportions of part-time women employees in Great Britain according to the New Earnings Survey (NES) and the Labour Force Survey (LFS), over several recent years.
- Make a scatter diagram for these data.
  - Fit a regression model with response variable the LFS percentages, and explanatory variable the NES percentages.
  - Is your fitted model sensible?

Year	NES %	LFS %
1985	32.6	44.6
1986	32.9	45.0
1987	32.6	45.0
1988	32.6	44.5
1989	31.9	43.7
1990	32.8	43.3
1991	33.0	43.4
1992	33.9	43.8
1993	34.7	43.7

- i. From the scatter diagram, there is no clear linear relationship between the variables.

Scatter plot of part-time women employees by survey



ii.  $\sum x_i = 297$  ,  $\sum Y_i = 397$  ,  $\sum x_i Y_i = 13099.44$

$\sum x_i^2 = 9806.44$  ,  $\bar{x} = 33$  ,  $\bar{Y} = 44.1111$

$$B = \frac{\sum_{i=1}^n x_i Y_i - n \bar{x} \bar{Y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} = -0.2132$$

$A = \bar{Y} - B \bar{x} = 51.1479$

The fitted line :  $\hat{Y} = 51.1479 - 0.2132x$

- iii. The negative slope suggests : as x increases the Y decreases , since we are measuring surveys for the same variables , they are negatively related , if x = 0 then Y = 51% which means one survey is suggesting that 0 % is the rate of employment and the other is suggesting 51% and therefore the model is not suitable.