04b Sample Examination Problems Chapter 11
SOLUTIONS

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

Refer to Chapter 10 Q1 (a)

(b) The table below shows the population of England and Wales in millions for years in the 19th century. Test the null hypothesis that the population regression slope is 0.21.

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

Refer to Chapter 10 Q1 (b) we came at the following fitted line:

$\widehat{Y}_i = -347.83 + 0.198x_i$

$H_0: \beta = 0.21$

$H_1: \beta \neq 0.21$

The test statistic : TS = $\dfrac{b-\beta}{SE(b)} \sim t_{n-2}$

Where the estimated standard error , SE(b) = $\dfrac{S}{\sqrt{(n-1)S_x^2}}$

Where S^2 is the estimated residual variance : $S^2 = \dfrac{(n-1)S_y^2(1-r_{xy}^2)}{n-2}$

r_{xy}^2 : the correlation coefficient

$S^2 = \dfrac{7(4.8583)^2(1-0.9969^2)}{8-2} = 0.16998 \Rightarrow S = \sqrt{0.16998} = 0.4129$

SE(b) = $\dfrac{0.4129}{\sqrt{7(24.49)^2}} = 0.00636$, TS = $\dfrac{b-\beta}{SE(b)} = \dfrac{0.198-0.21}{0.00636} = -1.929$

Let α = 0.05 , Two tailed test critical values : $\pm t_{\alpha/2,n-2} = \pm t_{0.025,6} = \pm 2.447$

Since $|-1.929| \prec 2.447$, it doesn't fall within the rejection region , we do not reject H_0 and therefore ,the true β is not significantly different from 0.21.

2. (a) Show that the least squares estimators of intercept and slope are unbiased estimators of the corresponding population parameters.

We need to show : E[B] = β and E[A] = α

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

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

$$\text{E[B]} = \text{E}\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] = \frac{\sum_{i=1}^{n}(x_i - \bar{x})E(Y_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Now $E(Y_i)$ = E [$\alpha + \beta x_i + \varepsilon_i$] = E [α] + $E[\beta x_i]$ + $E[\varepsilon_i]$

= $\alpha + \beta x_i + 0$

since $\varepsilon_i \sim \text{N}(0, \sigma^2)$, $E[\varepsilon_i] = 0$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\alpha + \beta x_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\alpha\sum_{i=1}^{n}(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} + \frac{\beta\sum_{i=1}^{n}x_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

But $\bar{x} = \frac{\sum_{i=1}^{n}x_i}{n} \Rightarrow \sum_{i=1}^{n}x_i - n\bar{x} = 0$ i.e. $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$

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

Now E[A] = $E[\bar{Y_i} - B\bar{x_i}]$ = $E(\bar{Y}_i)$ - E[$B\bar{x_i}$]

= E[$\alpha + \beta\bar{x} + \bar{\varepsilon}$] – $\bar{x}$ E[B]

= E[α] + $E[\beta\bar{x}]$ + 0] – $\bar{x}$ β

= α + $\beta\bar{x}$ - $\bar{x}$ β = α

(b) The table below shows heights in cm of male children on their fourth and fifth birthdays.

Child	1	2	3	4	5
Fourth Birthday	100.0	95.1	103.3	98.2	98.8
Fifth Birthday	105.5	101.5	110.0	104.5	104.8
Child	6	7	8	9	10
Fourth Birthday	103.0	98.6	97.5	95.3	97.7
Fifth Birthday	109.0	105.5	102.5	100.4	103.6

i. Find the least squares fit (i.e. intercept and slope) of a regression model for response variable height at fifth birthday and explanatory variable height at fourth birthday, and interpret your fitted line.

Model: $Y_i = \alpha + \beta x_i + \varepsilon_i$

$$A = \overline{Y_i} - B\overline{x_i}$$

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

x_i = height at fourth birthday , Y_i = height at fifth birthday: n = 10

$\sum x_i = 98.75$, $\sum Y_i = 1047.3$, $\sum x_i Y_i = 103494.68$

$\sum x_i^2 = 97584.17$, $\bar{x} = 98.75$, $\bar{Y} = 104.73$

$$B = \frac{\sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = = 1.077 \quad , \quad A = \bar{Y} - B\bar{x} = -1.598$$

The fitted line : $\hat{Y} = -1.598 + 1.077x$

Interpretation : For each additional cm of height at fourth birthday , the child is Expected to be 1.077 cm taller later year.

For x = 0 , the height will be negative which is not sensible to fit the model through the origin.

ii. Give a 90% confidence interval for the mean height on the fifth birthday for a height on fourth birthday of 98 cm.

The fitted model : $\widehat{Y} = -1.598 + 1.077x$, The fitted value when x = 98

$\widehat{Y} = -1.598 + 1.077(98) = 103.948$ this the Estimate

$$\widehat{\sigma}^2 = \frac{(n-1)S_y^2(1-r_{xy}^2)}{n-2} = \frac{9(3.031^2)(1-0.890^2)}{10-2} = 0.4015$$

The estimated standard error is given by : $\sqrt{\widehat{\sigma}^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)S_x^2}\right)}$

$$= \sqrt{(0.4015)\left(\frac{1}{10} + \frac{(98-98.75)^2}{9(2.760)^2}\right)} = 0.2084$$

90 % C.I. , α = 0.1 , $t_{\alpha/2,n-2} = t_{0.05,8} = 1.860$

Estimate $\pm t_{\alpha/2,n-2} \times$ Standard Error

= 103.984 $\pm$ 1.860 (0.2084) = (103.596 , 104.371)

iii. Test the null hypothesis that the population regression slope is 0.1.

Apparently there is an error in the question since a slope of 0.1 would make the height of the child smaller in his next birthday , so we'll assume $\beta = 1$

H_0 : $\beta = 1$

H_1 : $\beta \neq 1$, The test statistic : TS = $\frac{b-\beta}{SE(b)} \sim t_{n-2}$

Where the estimated standard error , SE(b) = $\frac{S}{\sqrt{(n-1)S_x^2}}$

Where S^2 is the estimated residual variance : $S^2 = \frac{(n-1)S_y^2(1-r_{xy}^2)}{n-2}$ = 0.4015

SE(b) = $\frac{\sqrt{0.4015}}{\sqrt{9(2.760)^2}} = 0.077$, TS = $\frac{b-\beta}{SE(b)} = \frac{1.077-1}{0.077} = 1$

Let α = 0.05 , Two tailed test critical values : $\pm t_{\alpha/2,n-2} = \pm t_{0.025,8} = \pm 2.306$

Since $1 \prec 2.306$, it doesn't fall within the rejection region , we do not reject H_0 and therefore the true β is not significantly different from 1.

3. (a) Derive from first principles the variance of the estimator of slope for a simple linear regression.

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

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

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

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

$Var[\varepsilon_i] = \sigma^2$ **= 0 + 0 +** σ^2

$$Var(B) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^4} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{(n-1)S_x^2} \text{ since } S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

(b) The table below shows Regional Manufacturing Capital Stock estimates in millions of pounds sterling at 1950 prices in the West Midlands and in the East Midlands.

Year	1950	1951	1952	1953	1954	1955	1956	1957	1958
West Midlands	2649	2742	2834	2918	3001	3114	3246	3385	3495
East Midlands	1748	1810	1854	1903	1944	1982	1991	2012	2028

i. Find the least squares fit of a regression model for response variable East Midlands Capital Stock and explanatory variable West Midlands Capital Stock.

Model: $Y_i = \alpha + \beta x_i + \varepsilon_i$

x_i = West Midland , Y_i = East Midland: n = 9

$\sum x_i = 27384$, $\sum Y_i = 17272$, $\sum x_i Y_i = 52767280$

$\sum x_i^2 = 83994808$, $\bar{x} = 3042.6666$, $\bar{Y} = 1919.1111$

$B = = = 0.31861$, $A = \bar{Y} - B\bar{x}$ = 949.67309

The fitted line : $\hat{Y}_i = 949.67 + 0.32x_i$

ii. Give the analysis of variance table for this regression.

Regression ANOVA table

Source	D.F.	Sum squares SS	MSE	F-value
Regression	k-1	(n-1) $S_y^2 r_{xy}^2$	$\frac{\text{Reg.SS}}{k-1}$	$\frac{\text{Reg.MS}}{\text{Res.MS}} = \frac{S_\beta^2}{S^2}$
Residual	n-k	(n-1) S_y^2(1- r_{xy}^2)	$\frac{\text{Res.SS}}{n-k}$	
Total	n-1	(n-1) S_y^2		

k = the number of regression coefficients = 2 (simple linear model)

Total SS = (n-1) S_y^2 = 8(9488.86) = 75910.89

Residual SS = (n-1) S_y^2(1- r_{xy}^2) = 75910.89(1-0.9497²)= 7446.56

Regression SS = Total SS – Res.SS = 75910.89 - 7446.56 = 68464.33

Reg. MSE = $\frac{\text{Reg.SS}}{k-1}$ = 68464.33/1 = 68464.33

Res. MSE = $\frac{\text{Res.SS}}{n-k} = \frac{7446.56}{7} = 1063.79$

F = $\frac{\text{Reg.MS}}{\text{Res.MS}} = \frac{68464.33}{1063.79} = 64.3588$

Source	D.F.	Sum squares SS	MSE	F-value
Regression	1	68464.33	68464.33	64.36
Residual	7	7446.56	1063.79	
Total	8	75910.89		

iii. Test the null hypothesis that the population regression slope is 0.

If we won't be able to reject H_0 , then the true value of β is 0 and therefore x is not An explanatory variable of Y (i.e. has no effect on Y)

$H_0: \beta = 0$

$H_1: \beta \neq 0$

The test statistic : TS = $\frac{b-\beta}{SE(b)} \sim t_{n-2}$

Where the estimated standard error , SE(b) = $\frac{S}{\sqrt{(n-1)S_x^2}}$

Where S^2 is the estimated residual variance = **1063.79**

SE(b) = $\frac{\sqrt{1063.79}}{\sqrt{8(290349)^2}} = 0.0397$, TS = $\frac{b-\beta}{SE(b)} = \frac{0.31861-0}{0.0397} = 8.022$

Let α = 0.05 , Two tailed test critical values : $\pm t_{\alpha/2,n-2} = \pm t_{0.025,7} = \pm 2.365$

Since $8.022 \succ 2.365$, it does fall within the rejection region , we reject H_0 and therefore the true β is significantly different from 0.

---

iv. Are the usual assumptions for inference on a regression model satisfied in this case?

We assume Y_i observations are :

- Independent :
- Normal
- Constant variance

Not valid because not likely to be independent as the value of some years are affected by the previous years values so dependent.