Topic: Line of Best Fit

Time: 45 mins Marks: /45 marks

Calculator Assumed

Question One: [2, 2: 4 marks]

a) Find the equation for the least squares regression line of B on A for the data set below.

A	5.2	3.1	6.9	4.3	6.5	8.5	7.5	5.0	5.6
B	32	12	15	18	10	11	9	16	15

b) Read the following description of a data set.

The university administration has received several complaints that Mrs Henderson is a very hard marker. Before taking any action, the Dean of the university wanted to see for herself if there was any truth behind the complaints. She gave Mrs Henderson unmarked copies of assessments previously marked by other examiners at the university. The Dean recorded the mean of the mark given by the other examiners as x and the mark given by Mrs Henderson as y, for each paper.

The least squares regression line of this data set is: $\hat{y} = -1.234x - 20.512$

Comment on the findings.

Question 2: [2, 3, 2, 2, 4, 2, 2: 17 marks]

Data for 10 individuals training for an ultra-marathon was collected. The number of months each person spent training for the ultra-marathon was recorded as was the number of kilometers they were able to complete before needing assistance in the 80 km event.

Months spent training	12	9	7	6	6	5	10	4	8	11
Kilometres achieved	80	69	70	50	67	45	72	30	73	75

An organiser for the event said to participants that many months of training were required to complete the event. He claimed, "The more you train, the more you will be able to complete". He went further to suggest that a minimum of 12 months of training is required in order to be able to complete the full 80 km.

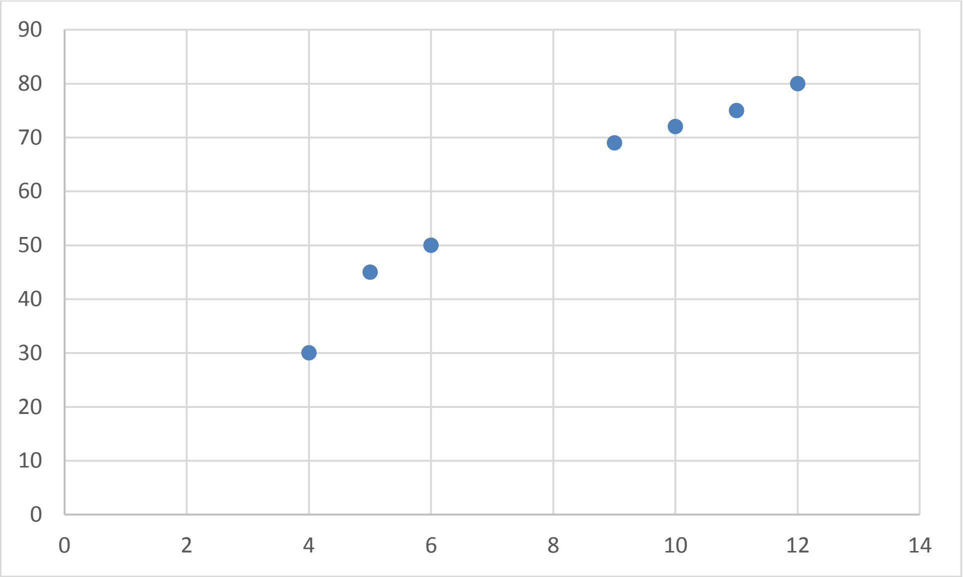
a) State the explanatory and response variable in the claim above.

b) Which statistical measure would be needed in order to assess whether or not this statement is reasonable? Calculate that measure and comment.

c) Calculate the least squares regression line for kilometres achieved (K) on Months spent training (M).

d) Comment on the claim made by the organiser. Consider your calculations in your answer.

The diagram below is the partial scatter plot for the sample in the table on the previous page. Three data points are missing.



d) Label the axes **and** complete the scatter plot by including the three missing data points.

e) Draw the regression line from 2(c) on the graph.

f) Using the line of best fit, calculate the residual for the participant who trained for 6 months and completed 67 km of the course.

Question Three: [2, 6, 3: 11 marks]

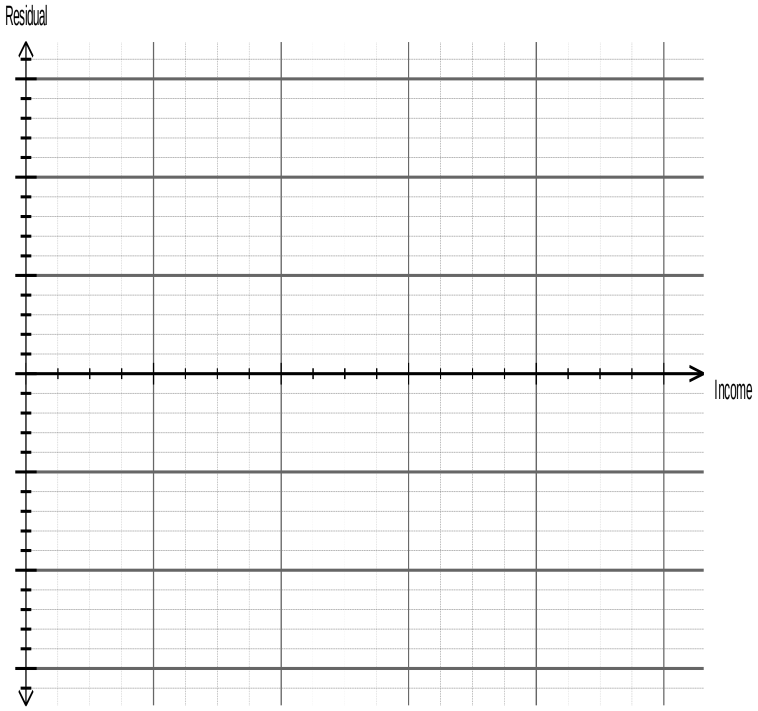
The following data was collected in a national health survey. One of the main focuses of the survey was to ascertain whether or not there was a link present between family income and the percentage of the family classified as overweight.

The table below shows a snap shot of the data collected.

Family income \$1000s (*I*)	25	37	43	51	53	61	65	70	90
% of family overweight (*P*)	3	5	8	9	9	12	10	8	9
residual	-2.3	A	1.0	1.3	B	3.0	1.0	C	D

a) Calculate the least squares regression line for the % of the family being overweight and the family income.

b) Calculate the values of A, B, C and D in the table above and plot the residuals on the graph below.



c) Does the least squares regression line calculated in part a) provide a good model for predicting the percentage of the family who is overweight based on the family income?

Question Four: [3, 3, 4, 1, 2 : 13 marks]

Student loan debt is debt incurred under Higher Education Loans Programmes (HELP), the government education payment scheme, and other government higher education schemes. It also includes debt incurred prior to 2005 under the Higher Education Contributions Scheme (HECS) and the Student Financial Supplement Scheme (SFSS).

Consider the following table relating to HECS debts from 1989 – 2004 and the scatterplot for the Accumulated HECS debt as at 30 June.

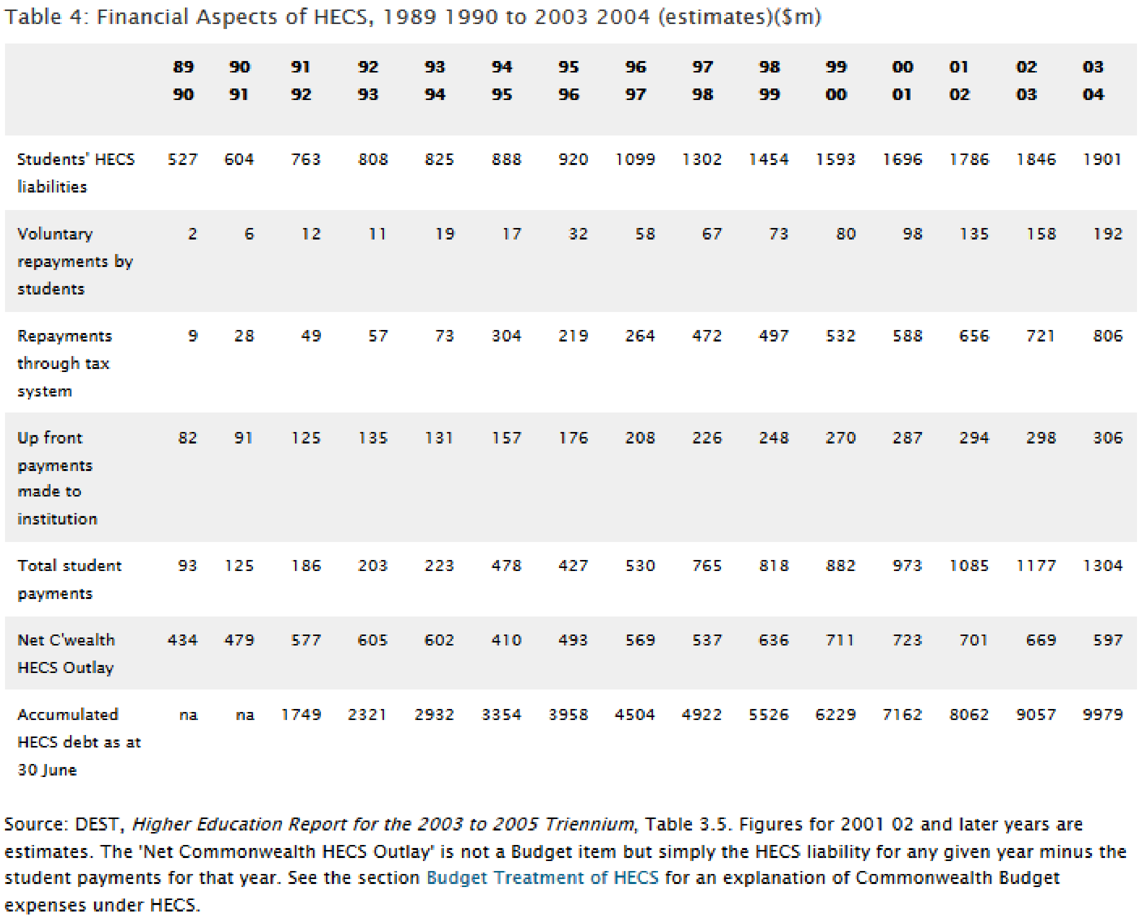
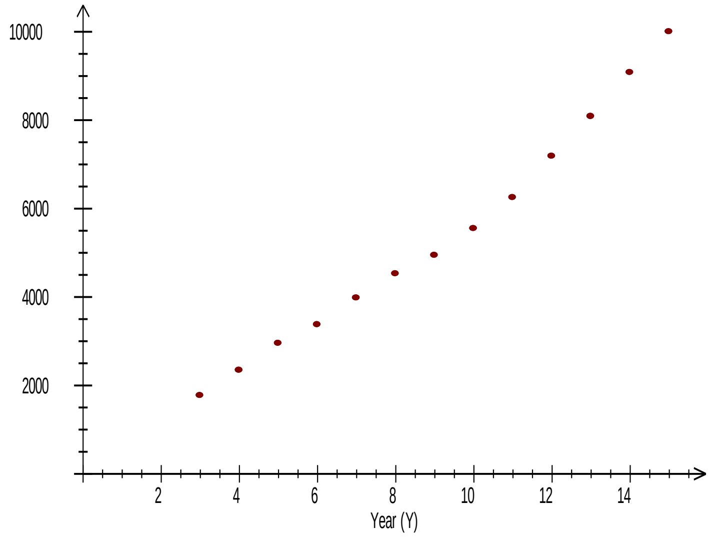


Table 4: Financial Aspects of HECS, 1989 1990 to 2003 2004 (estimates)($m)

	89 90	**90 91**	**91 92**	**92 93**	**93 94**	**94 95**	**95 96**	**96 97**	**97 98**	**98 99**	**99 00**	**00 01**	**01 02**	**02 03**	**03 04**
Students' HECS liabilities	527	604	763	808	825	888	920	1099	1302	1454	1593	1696	1786	1846	1901
Voluntary repayments by students	2	6	12	11	19	17	32	58	67	73	80	98	135	158	192
Repayments through tax system	9	28	49	57	73	304	219	264	472	497	532	588	656	721	806
Up front payments made to institution	82	91	125	135	131	157	176	208	226	248	270	287	294	298	306
Total student payments	93	125	186	203	223	478	427	530	765	818	882	973	1085	1177	1304
Net C'wealth HECS Outlay	434	479	577	605	602	410	493	569	537	636	711	723	701	669	597
Accumulated HECS debt as at 30 June	na	na	1749	2321	2932	3354	3958	4504	4922	5526	6229	7162	8062	9057	9979

Source: DEST, *Higher Education Report for the 2003 to 2005 Triennium*, Table 3.5. Figures for 2001 02 and later years are estimates. The 'Net Commonwealth HECS Outlay' is not a Budget item but simply the HECS liability for any given year minus the student payments for that year. See the section Budget Treatment of HECS for an explanation of Commonwealth Budget expenses under HECS.

The scatterplot below plots Accumulated HECS debt as at 30 June against years 1991 – 2004. Note: 89 – 90 is represented as year 1 and an entry of 'na' in the table cannot be included in the scatterplot.



a) Calculate the least squares regression line and the correlation coefficient for the Accumulated HECS Debt, *H*, over time, *Y*. (Let 89 – 90 be year 1)

b) Describe the relationship between these variables.

c) Use the linear regression to predict the Accumulated HECS debt for 2004/2005 and 2016 - 2017. Comment on the reliability of these predictions.

d) Calculate the coefficient of determination.

e) Describe what the value calculated in part d) means.

education equals

Topic: Line of Best Fit SOLUTIONS

Time: 45 mins Marks: /45 marks

Calculator Assumed

Question One: [2, 2: 4 marks]

a) Find the equation for the least squares regression line of B on A for the data set below.

A	5.2	3.1	6.9	4.3	6.5	8.5	7.5	5.0	5.6
B	32	12	15	18	10	11	9	16	15

$\hat{B} = -1.4274A + 23.6755$ (4 dp) ✓✓

b) Read the following description of a data set.

The university administration has received several complaints that Mrs Henderson is a very hard marker. Before taking any action, the Dean of the university wanted to see for herself if there was any truth behind the complaints. She gave Mrs Henderson unmarked copies of assessments previously marked by other examiners at the university. The Dean recorded the mean of the mark given by the other examiners as x and the mark given by Mrs Henderson as y, for each paper.

The least squares regression line of this data set is: $\hat{y} = -1.234x - 20.512$

Comment on the findings.

For each additional mark in the original score, the least squares regression line predicts that Mrs Henderson would mark the paper down by an additional 1.234 marks.

✓ ✓

Question 2: [2, 3, 2, 2, 4, 2, 2: 17 marks]

Data for 10 individuals training for an ultra-marathon was collected. The number of months each person spent training for the ultra-marathon was recorded as was the number of kilometers they were able to complete before needing assistance in the 80 km event.

Months training	12	9	7	6	6	5	10	4	8	11
Km achieved	80	69	70	50	67	45	72	30	73	75

An organiser for the event said to participants that many months of training were required to complete the event. He claimed, “The more you train, the more you will be able to complete”. He went further to suggest that a minimum of 12 months of training is required in order to be able to complete the full 80 km.

a) State the explanatory and response variable in the claim above.

Explanatory – months training ✓

Response – Km achieved ✓

b) Which statistical measure would be needed in order to assess whether or not this statement is reasonable? Calculate that measure and comment.

$r = 0.8508$

✓

✓ ✓

There is a strong, positive, linear correlation between the variables.

c) Calculate the least squares regression line for kilometres achieved (K) on Months spent training (M).

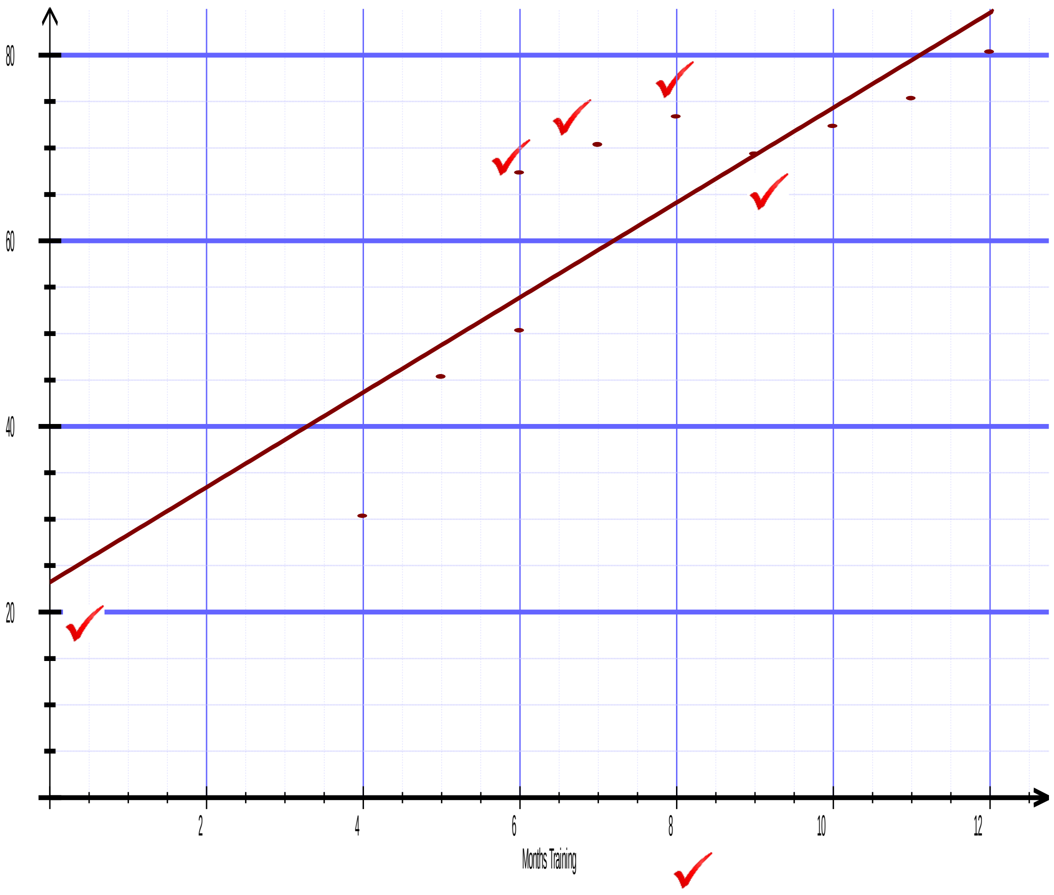
$\hat{K} = 5.1132M + 23.2170\ (4dp)$

✓ ✓

d) Comment on the claim made by the organiser. Consider your calculations in your answer.

While there is a strong correlation, causation cannot be assumed as there may be other factors involved. ✓ ✓

The diagram below is the partial scatter plot for the sample in the table on the previous page. Three data points are missing.



(both labels)

d) Label the axes **and** complete the scatter plot by including the three missing data points.

e) Draw the regression line from 2(c) on the graph.

f) Using the line of best fit, calculate the residual for the participant who trained for 6 months and completed 67 km of the course.

$67 - 53.8962 = 13.104$

Question Three: [2, 6, 3: 11 marks]

The following data was collected in a national health survey. One of the main focuses of the survey was to ascertain whether or not there was a link present between family income and the percentage of the family classified as overweight.

The table below shows a snap shot of the data collected.

Family income $1000s (I)	25	37	43	51	53	61	65	70	90
% of family overweight (P)	3	5	8	9	9	12	10	8	9
residual	-2.3	A	1.0	1.3	B	3.0	1.0	C	D

a) Calculate the least squares regression line for the % of the family being overweight and the family income.

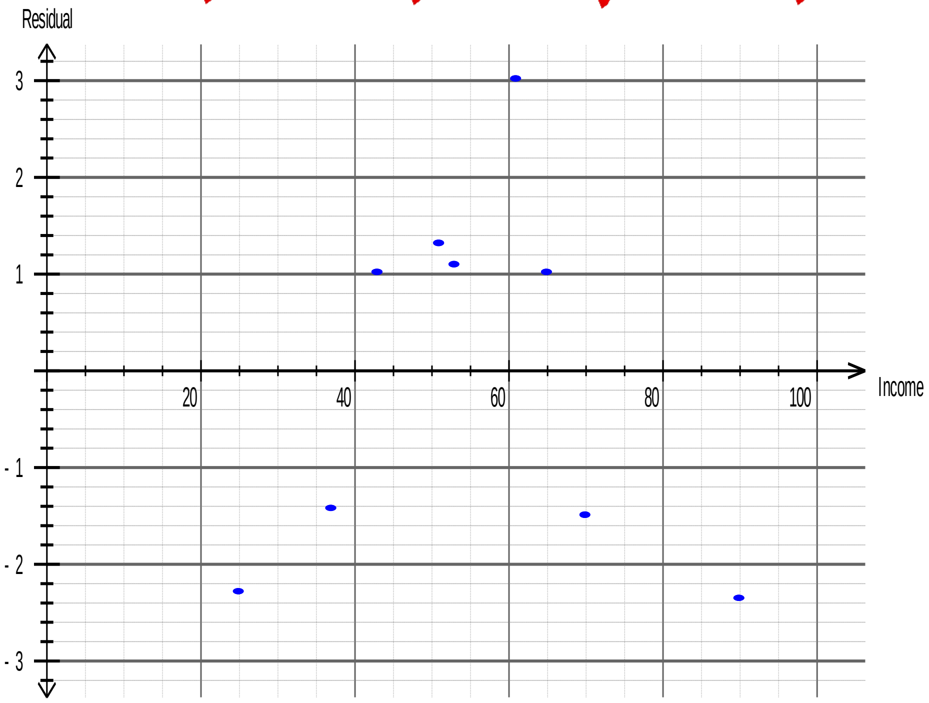
$\hat{P} = 0.9314I + 2.9883\ (4dp)$

✓ ✓

b) Calculate the values of A, B, C and D in the table above and plot the residuals on the graph below.

$A = -1.44$ $B = 1.08$ $C = -1.51$ $D = -2.37$

✓ ✓ ✓ ✓



✓ ✓

c) Does the least squares regression line calculated in part a) provide a good model for predicting the percentage of the family who is overweight based on the family income?

✓ No a linear regression is not a good model. The residuals are not randomly scattered above and below the income axis. They are clustered in the middle. ✓ ✓

Question Four: [3, 3, 4, 1, 2 : 13 marks]

Student loan debt is debt incurred under Higher Education Loans Programmes (HELP), the government education payment scheme, and other government higher education schemes. It also includes debt incurred prior to 2005 under the Higher Education Contributions Scheme (HECS) and the Student Financial Supplement Scheme (SFSS).

Consider the following table relating to HECS debts from 1989 – 2004

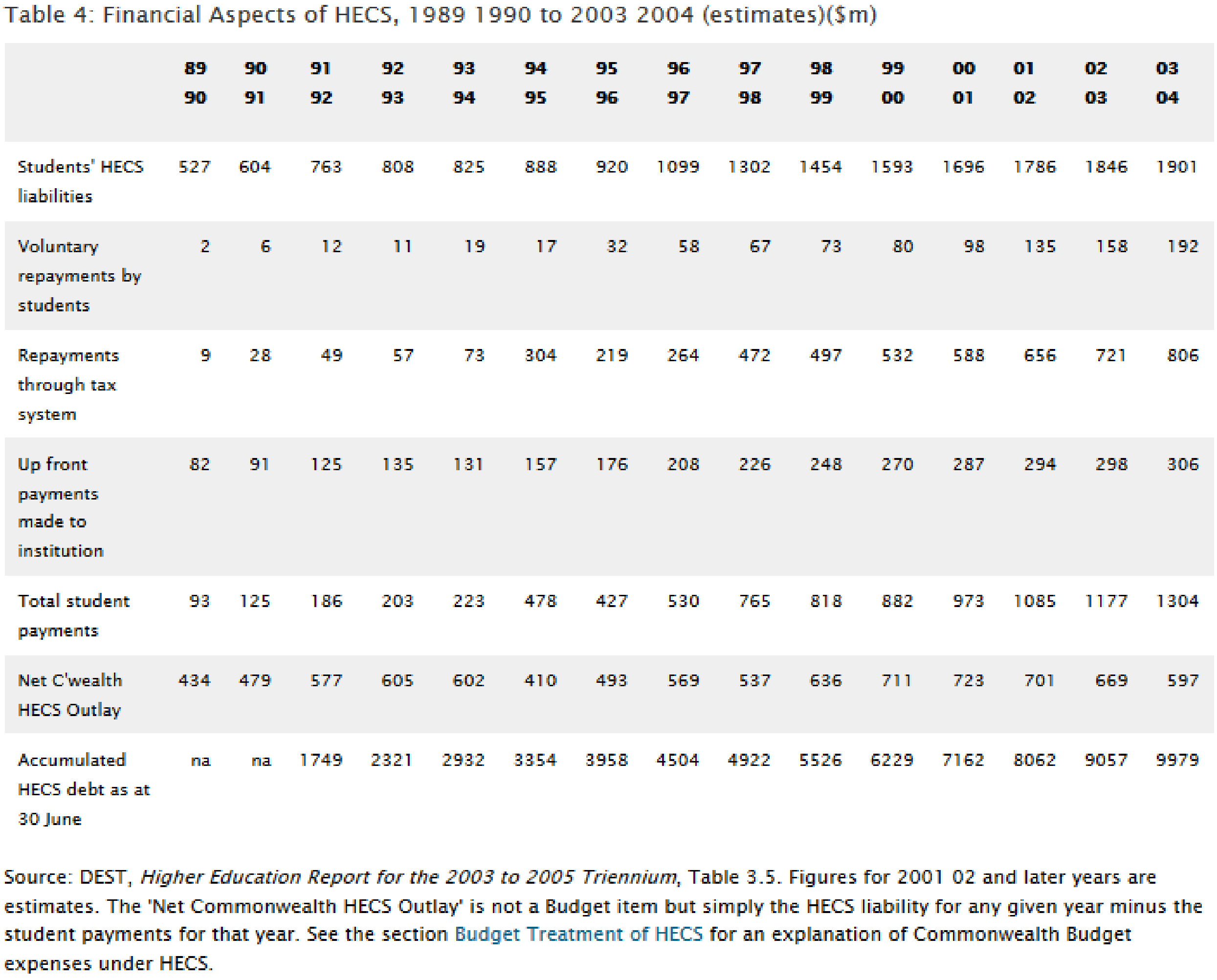
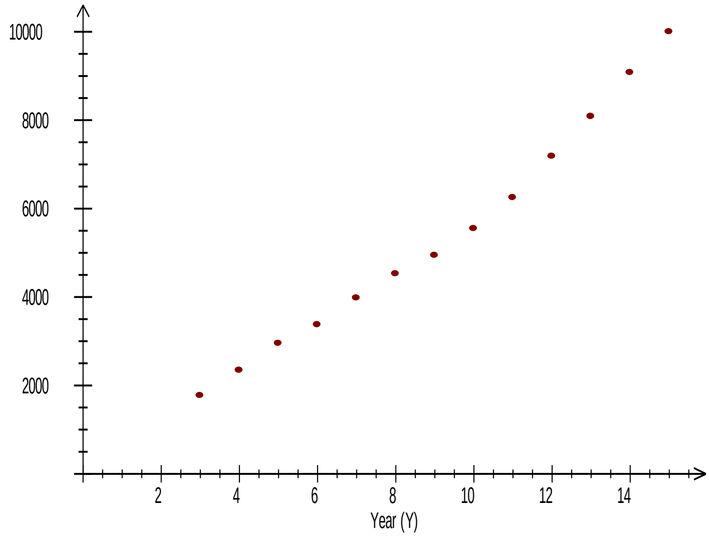


Table 4: Financial Aspects of HECS, 1989 1990 to 2003 2004 (estimates)($m)

	89	90	91	92	93	94	95	96	97	98	99	00	01	02	03
	90	91	92	93	94	95	96	97	98	99	00	01	02	03	04
Students' HECS liabilities	527	604	763	808	825	888	920	1099	1302	1454	1593	1696	1786	1846	1901
Voluntary repayments by students	2	6	12	11	19	17	32	58	67	73	80	98	135	158	192
Repayments through tax system	9	28	49	57	73	304	219	264	472	497	532	588	656	721	806
Up front payments made to institution	82	91	125	135	131	157	176	208	226	248	270	287	294	298	306
Total student payments	93	125	186	203	223	478	427	530	765	818	882	973	1085	1177	1304
Net C'wealth HECS Outlay	434	479	577	605	602	410	493	569	537	636	711	723	701	669	597
Accumulated HECS debt as at 30 June	na	na	1749	2321	2932	3354	3958	4504	4922	5526	6229	7162	8062	9057	9979

Source: DEST, *Higher Education Report for the 2003 to 2005 Triennium*, Table 3.5. Figures for 2001 02 and later years are estimates. The 'Net Commonwealth HECS Outlay' is not a Budget item but simply the HECS liability for any given year minus the student payments for that year. See the section Budget Treatment of HECS for an explanation of Commonwealth Budget expenses under HECS.

The scatterplot below plots Accumulated HECS debt as at 30 June against years 1991 – 2004. Note: 89 – 90 is represented as year 1, an NA in the table cannot be included in the scatterplot.



a) Calculate the least squares regression line and the correlation coefficient for the Accumulated HECS Debt, *H*, over time, *Y*. (Let 89 – 90 be year 1)

$\hat{H} = 662.4615Y - 596.3846\ (4dp)$ $r = 0.9905\ (4dp)$

✓ ✓ ✓

b) Describe the relationship between these variables.

Very strong, positive, linear correlation.

✓ ✓ ✓

c) Use the linear regression to predict the Accumulated HECS debt for 2004/2005 and 2016 - 2017. Comment on the reliability of these predictions.

$2004 - 2005\ Y = 16\ \hat{H} = 10002.9994$
HECS debt is approximately \$1 0003 ✓
This prediction is reliable since the correlation coefficient is very high and the prediction is only one cycle out of the given data. ✓

$2016 - 2017\ Y = 28\ \hat{H} = 17\,952.5374$
HECS debt is approximately \$17 952.54 ✓
This prediction is unreliable since it is an extrapolation. ✓

d) Calculate the coefficient of determination.

$r^2 = 0.9811$ ✓

e) Describe what the value calculated in part d) means.

98% of the variation in HECS debt can be explained by the variation in time.
✓ ✓