Ugochinyereec 3523799 MATH215 Assignment 6 Noella Echezona Math 215 Assignment 6

Introduction to Statistics (Athabasca University)

Excellent!
100%

(Revision 10)

Assignment 6

Overview

Total marks: / 62

This assignment covers content from Unit 6. It assesses your knowledge of correlational analysis and regression analysis used to examine the relationship between two quantitative variables.

Instructions

- Show all your work and justify all of your answers and conclusions, except for the True/False questions.
- Keep your work to 4 decimals, unless otherwise stated.
- ***Note:*** Finishing a test of hypotheses with a statement like "reject H_0" or "do not reject H_0" will be insufficient for full marks. You must also provide a written concluding statement in the context of the problem itself. For example, if you are testing hypotheses about the effectiveness of a medical treatment, you must conclude with a statement like, "we can conclude that the treatment is effective" or "we cannot conclude that the treatment is effective."

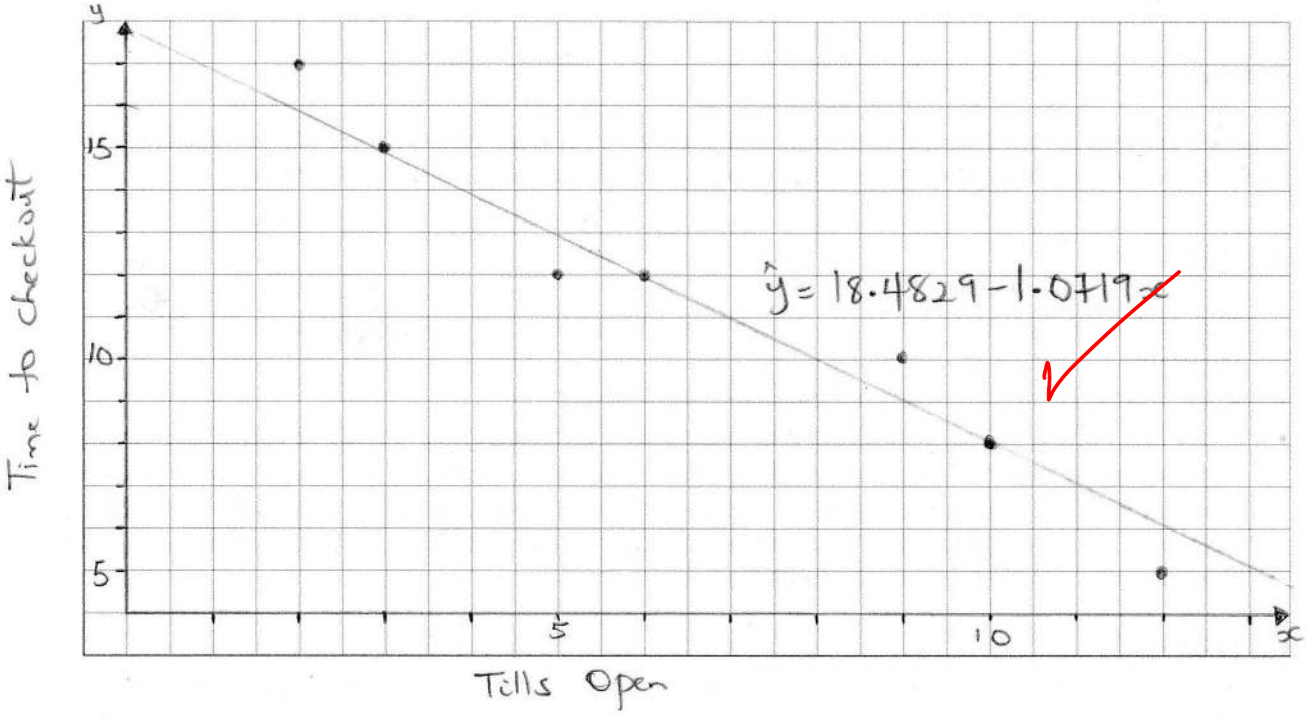
(43 total marks)

1. A large warehouse superstore is interested in optimizing its customers' shopping experiences and, as such, wants to ensure that it is able to staff the store properly during peak hours. The store management is interested in studying the relationship between the number of tills or checkouts that are open in the store and the amount of time it takes for a customer to check out (that is, the time it takes from when they get in line to when they complete their purchase). The data in the following table were collected from a random sample of 7 customers:

Tills Open (x)	*Time to Checkout (minutes)* (y)	xy	x^2	y^2
2	17	34	4	289
9	10	90	81	100
12	5	60	144	25
5	12	60	25	144
3	15	45	9	225
10	8	80	100	64
6	12	72	36	144
$\Sigma x = 47$	$\Sigma y = 79$	$\Sigma xy = 441$	$\Sigma x^2 = 399$	$\Sigma y^2 = 991$

(4 marks)

a. Construct a scatter diagram for these data with "Tills Open" on the horizontal (x) axis, and "Time to Checkout" on the vertical (y) axis. Note: Try to make relatively full use of the graph paper provided.



(2 marks)

b. Describe the general pattern of relationship between the two variables within the context of this question.

In the scatter diagram, there is a negative correlation between the two variables, thus as x (number of tills open) increases, y (amount of time to checkout) decreases.

(11 marks) (Revision 10)

c. Calculate the least squares regression line using "Time to Checkout" as the dependent variable and "Tills Open" as the independent variable.

$\Sigma x = 47, \quad \bar{x} = \Sigma x / n = \frac{47}{7} = 6.714285714$

$\Sigma y = 79. \quad \bar{y} = \Sigma y / n = \frac{79}{7} = 11.28571429$

$SS_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} =$

$= 441 - \frac{(47)(79)}{7} = 441 - 530.4285714$

$= -89.4285714$

$SS_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 399 - \frac{(47)^2}{7} = 399 - 315.5714286$

$= 83.4285714$

$b = \frac{SS_{xy}}{SS_{xx}} = \frac{-89.4285714}{83.4285714} = -1.071917808$ ✓

$a = \bar{y} - b\bar{x} = 11.28571429 - (-1.071917808)(6.714285714)$

$= 11.28571429 - (-7.19716245)$

$= 18.48287672$ ✓

$\hat{y} = a + bx \Rightarrow \hat{y} = 18.4829 - 1.0719x$ ✓

(3 marks)

d. Calculate predicted values for $x = 3$ and $x = 10$. Use these values to help plot the regression line on the scatter diagram you constructed in part a. above.

$x=3$; $\hat{y} = 18.4829 - 1.0719(3) = 15.2672$

$x=10$: $\hat{y} = 18.4829 - 1.0719(10) = 7.7639$

(10 marks)

e. Can it be concluded that the slope of the regression line is negative? Formulate and test the appropriate hypotheses at the 5% significance level. Use the critical value approach. Clearly state and explain your conclusion within the context of the problem.

$SS_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n} = 991 - \frac{(79)^2}{7} = 991 - 891.5714286$

$= 99.4285714$

$S_e = \sqrt{\frac{SS_{yy} - b\,SS_{xy}}{n-2}} = \sqrt{\frac{99.428514 - (-1.0719)(-89.4285714)}{7-2}}$

$= 0.844988559$

$S_b = \frac{S_e}{\sqrt{SS_{xx}}} = \frac{0.844988559}{\sqrt{83.4285714}} = 0.092511011$

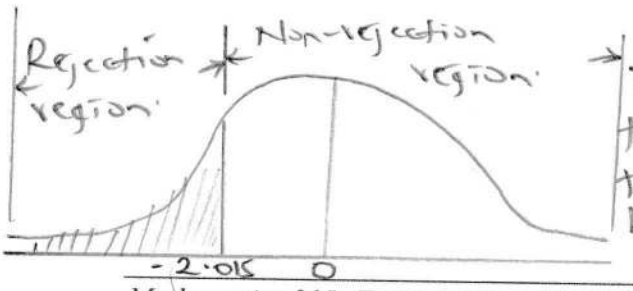
$H_0 : B = 0$; Area in left tail of t distribution $\alpha = 0.05$

$H_1 : B < 0$ $df = n - 2 = 7 - 2 = 5$

Critical value of t for 5 df and 0.05 area in left tail $= -2.015$

$t = \frac{b - B}{S_b} = \frac{(-1.0719) - 0}{0.092511011} = -11.586728\,85$

≈ -11.5867



The value of the test statistic, -11.5867 is less than the critical value, -2.015, thus it falls in the rejection region. Hence, we reject the null hypothesis and conclude that the slope of the regression is negative.

(4 marks)

f. Construct a 95% confidence interval for β.

$n = 7,\ \alpha = 95\% = 0.95\ ;\ CL = 1 - 0.95\ ;\ b = -1.0719$

$df = n - 2 = 7 - 2 = 5$

$S_b = 0.092511011 \simeq 0.0925$

$\alpha/2 = \frac{(1 - 0.95)}{2} = 0.025$

5df and $0.025 = 2.571$

$b \pm t_{sb} = -1.0719 \pm 2.571(0.0925)$

$= -1.3097$ to -0.8341 ✓

(2 marks)

g. Interpret the value of b in the sample regression line. What does it mean in the context of this question?

The value of b, -0.10719 is negative which means that for every 1 unit increase in value of 'tills open'(x), the value of time to checkout in minutes(y) decreases by 1.0719 and vice versa. Thus x and y have a negative correlation between them. ✓

(2 marks)

h. One of the store managers regularly likes to keep 8 tills open on Saturdays. Use the equation of the regression line to provide the manager with the predicted time to check out if 8 tills are open.

$\hat{y} = a + bx \Rightarrow \hat{y} = 18.4829 - 1.0719(8) = 9.9077$

$x = 8$

Predicted time to check out if 8 tills are open ✓

$= 9.9077$ minutes

(1 mark)

i. Which of the following ***cannot*** be answered from the regression equation? Clearly circle only one response.

A. A prediction of the value of y at a particular value of x.

B. An estimate of the slope between y and x.

C. An estimate of whether the linear association between variables is positive or negative.

(D.) An estimate of whether the association between variables is linear or non-linear

(2 marks)

j. In a sentence or two, describe what information the standard deviation of errors provides.

The standard deviation of errors provides information about the cumulative accuracy of an estimate excluding the predicted value.

(2 marks)

k. When the correlation between x and y is -1.0, what will the standard deviation of errors be? Why is this?

The standard deviation of errors between x and y = -1.0 is Zero(0).
This is because a correlation of -1.0 between variables indicates a perfect negative correlation and y values can be perfectly predicted by x values.

(19 total marks)

2. Does the amount of education you have predict your salary? To answer this question, data from a random sample of 8 working adults was collected. Each participant answered the number of years of post-secondary education they have as well as their annual salary (in thousands of dollars). The data was as follows:

Post-Secondary Education (in years) (x)	*Annual Salary (in thousands of dollars)* (y)
4	70
0	55
5	40
8	80
10	125
4	95
2	85
6	60

You may use the following sums and sums of squares and cross products for the questions below.

$\Sigma x = 39$ $\Sigma y = 610$ $SS_{xx} = 70.875$ $SS_{yy} = 4887.5$ $SS_{xy} = 306.25$ $n = 8$

(5 marks)

a. Calculate the least squares regression line using "Annual Salary" as the dependent variable and "Post-Secondary Education" as the independent variable.

$\bar{x} = \frac{\Sigma x}{n} = \frac{39}{8} = 4.875 \quad \bar{y} = \frac{\Sigma y}{n} = \frac{610}{8} = 76.25$

$b = \frac{SS_{xy}}{SS_{xx}} = \frac{306.25}{70.875} = 4.320987654 \approx 4.3211$ ✓

$a = \bar{y} - b\bar{x} = 76.25 - (4.3211)(4.875)$

$= 55.1846375$

≈ 55.1846 ✓

$\hat{y} = 55.1846 + 4.3211x$ ✓

(2 marks)

b. Interpret the value of b in the sample regression line. What does it mean in the context of this question?

Because the value of b, 4.3211 is positive, there is a positive correlation between the variables x and y, thus if post-secondary education increases, the annual salary would also increase.

(2 marks)

c. Compute the linear correlation between "Post-Secondary Education" and "Annual Salary". Express your answer to 4 decimal places of accuracy.

$$\text{Linear Correlation } r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}$$

$$= \frac{306.25}{\sqrt{(70.875)(4887.5)}}$$

$$= 0.520338758 \approx 0.5203$$

The linear correlation btw post-secondary education and annual salary is 0.5203

(2 marks)

d. What percentage of variation in annual salary is explained by its linear relationship with post-secondary education?

$$r^2 = (\text{Linear Correlation})^2 = (0.5203)^2$$

$$= 0.27071209 \approx 0.2707$$

$$0.2707 \times 100 = 27.07\%$$

Thus, 27.07% variation in annual salary is explained by its linear relationship with post-secondary education.

(8 marks) (Revision 10)

e. At the 2.5% significance level, can it be concluded that the correlation between post-secondary education and annual salary is positive? Formulate and test the appropriate hypotheses. Use the critical value approach. Clearly state and explain your conclusion within the context of the question.

$$S_e = \sqrt{\frac{SS_{yy} - b\,SS_{xy}}{n-2}} = \sqrt{\frac{4887.5 - (4.3211)(306.25)}{8-2}}$$

$$= 24.37267297$$

$$S_b = \frac{S_e}{\sqrt{SS_{xx}}} = \frac{24.37267297}{\sqrt{70.875}} = 2.89505371$$

$H_0 : B = 0$

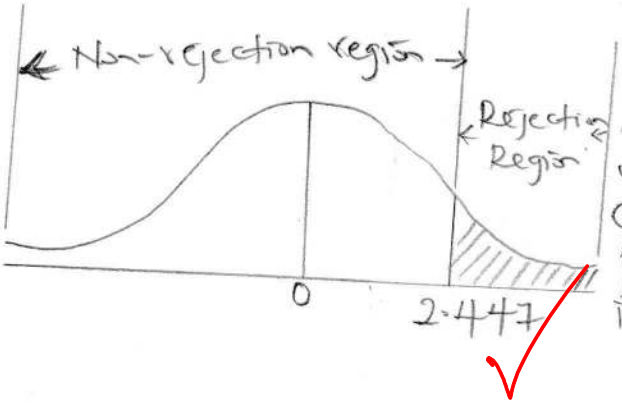
$H_1 : B > 0$

Area in right tail of t distribution $\alpha = 0.025$

$df = n - 2 = 8 - 2 = 6$

Critical value of t for 6df and 0.025 area in right tail = 2.447

$$t = \frac{b - B}{S_b} = \frac{(4.3211) - 0}{2.89505371} = 1.492580253 \approx 1.4926$$



The value of the test statistic, 1.4926 is less than the critical value, 2.447, thus it falls in the non-rejection region. Hence, we would not reject the null hypothesis as we cannot conclude that the correlation between post-secondary education and annual salary is positive.