Questions 50 to **59** refer to the following information.

A study by Jabourian *et al.* (2014) explored whether walking time might provide an early signal of abnormal cognitive performance. One section of the study looked at 137 healthy adults aged between 50 and 65 years old. Walking time and cognitive performances were assessed.

A total psychometric score (**Score**) was recorded. This was the composite of four psychometric test scores and had a possible range of scores between 0 and 132. A score lower than 87 was considered to be clearly abnormal. Subjects were required to walk with their regular shoes on level ground for 50 metres. From an initial line and standing still, the given order was: "Walk as fast as you can, without running, touch the wall and come back." The time taken to complete this task (**Time**) was also recorded.

A simple linear regression analysis was carried out to investigate the relationship between **Score** and **Time**. The results of this analysis and associated plots are shown in Figure 7 below and in Figure 8 and Table 6, page 34.

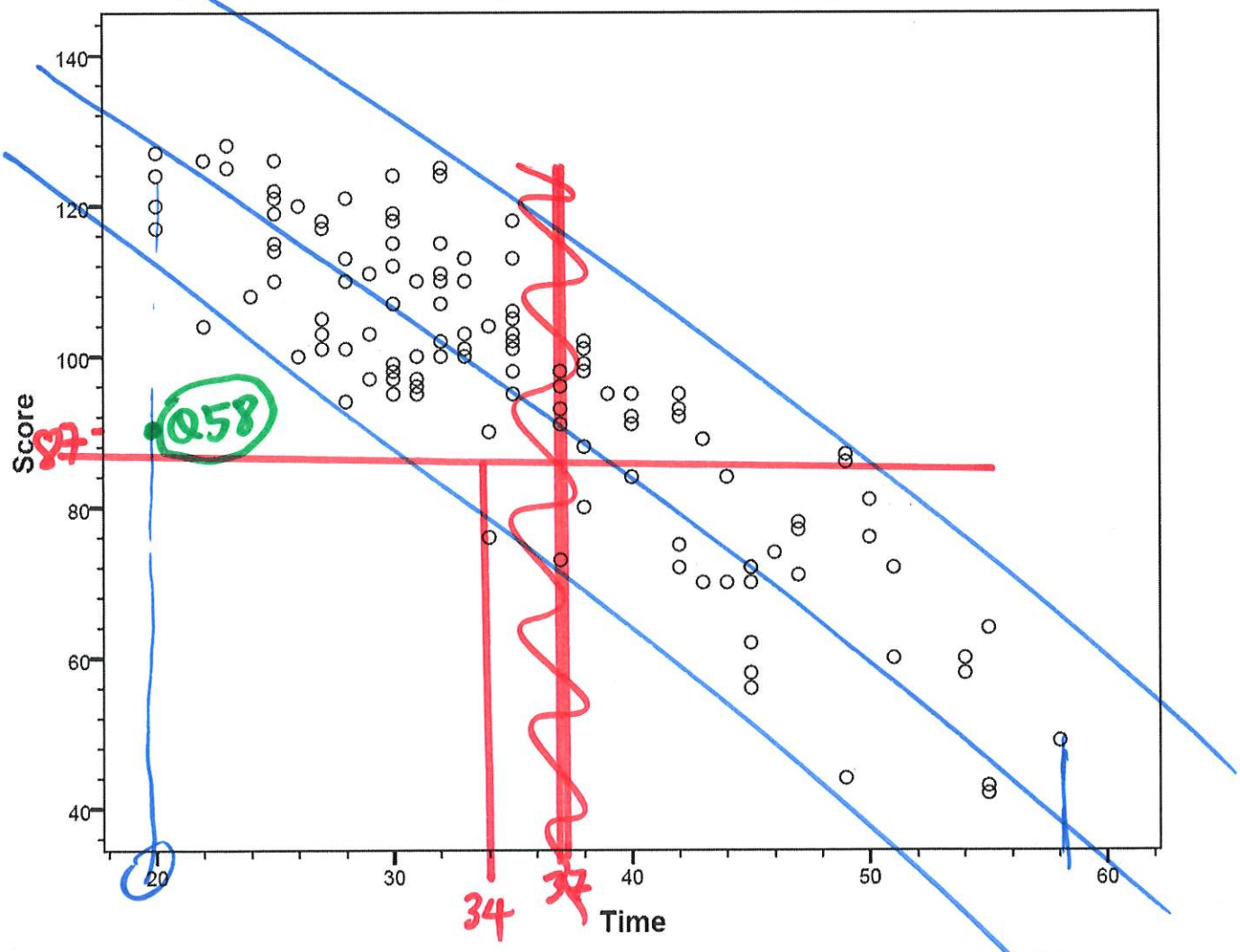


Figure 7: Total psychometric score against walking time (seconds)

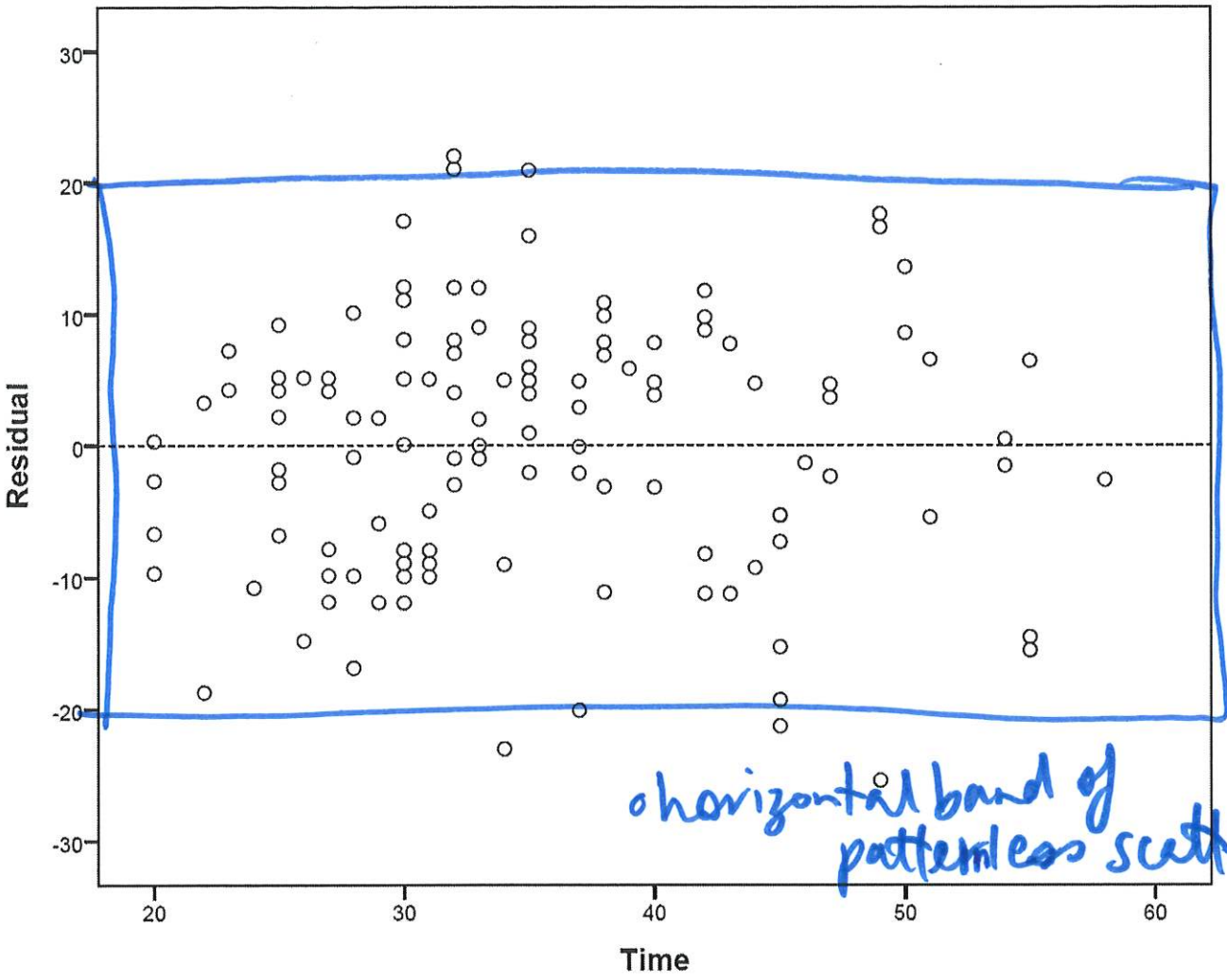


Figure 8: Residual plot

$\hat{y} = 166.194 - 1.976 \times Time$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	166.194	3.730		44.552	.000	158.807	173.582
	Time	-1.976	.102	-.873	-19.318	.000	-2.178	-1.773

a. Dependent Variable: Score

Table 6: Simple linear regression output

50. Based on Figure 7, page 35, which **one** of the following statements is **false**?

(1) All the subjects with a 'clearly abnormal' total psychometric score (< 87) had walking times of at least 34 seconds.

(2) There is a moderately strong relationship and a negative association between total psychometric score and walking time.

(3) All the subjects whose total psychometric scores were **not** considered to be 'clearly abnormal' (≥ 87) had walking times of no more than 34 seconds.

(4) The time taken to complete the walking task for the slowest subject is nearly three times longer than that for the fastest.

(5) The subjects with a relatively long (slow) walking time tend to have a relatively low total psychometric score, and the subjects with a relatively short (fast) walking time tend to have a relatively high total psychometric score.

51. Based on Figure 7, page 35, and Figure 8, page 36, which **one** of the following statements is **false**?

(1) Although the scatter plot of **Score** vs **Time** suggests a linear association, the Residual plot strongly indicates we should be concerned about the assumption of linearity.

(2) The residual plot does not provide us with information about the assumption of independence of the random errors.

(3) The scatter plot of **Score** vs **Time** does not provide us with information about the assumption of independence of the random errors.

(4) The plots do not suggest that we should be concerned with the assumption of constant spread of the random errors.

(5) The plots do not suggest that we should be concerned with the Normality assumption of the random errors.

Questions 52 to **57** assume that the simple linear regression analysis on Score and Time is valid.
(Note this may not be true.)

52. Referring to the simple linear regression output in Table 6, page 36, which **one** of the following statements is **true**?

(1) There is no evidence of a linear relationship between **Score** and **Time**.

(2) There is very strong evidence of a linear relationship between **Score** and **Time**.

(3) There is no evidence of a very strong linear relationship between **Score** and **Time**.

(4) There is very strong evidence that β_0 is zero.

(5) There is very strong evidence that the intercept of the least squares regression line is zero.

53. The equation for the least squares regression analysis is:

(1) Average **Score** = −1.976 + 0.102 × **Time**

(2) Average **Time** = −1.976 + 166.194 × **Score**

(3) Average **Score** = 166.194 − 1.976 × **Time**

(4) Average **Score** = −1.976 + 166.194 × **Time**

(5) Average **Time** = 166.194 − 1.976 × **Score**

see pg 36 for colour coding!

54. For a subject with a walking time of 40 seconds, we would estimate their total psychometric score to be:

(1) 2.1

(2) −79.0

(3) 79.0

(4) 87.2

(5) 166.2

$\hat{y} = 166.194 - 1.976 \times 40$

$= 87.154$

55. One of the subjects in this study had a walking time of 50 seconds and a total psychometric score of 77. Under this regression analysis the residual for this person is:

(1) 9.61

(2) −9.61

(3) −1.98

(4) 1.98

(5) 27

$\text{res} = \text{obs} - \text{exp} = y - \hat{y}$

$= 77 - 67.394$

$= 9.606$

$\hat{y} = 166.194 - 1.976 \times 50$

$= 67.394$

56. For two subjects whose walking times differ by 4 seconds, we would predict their total psychometric score to differ by:

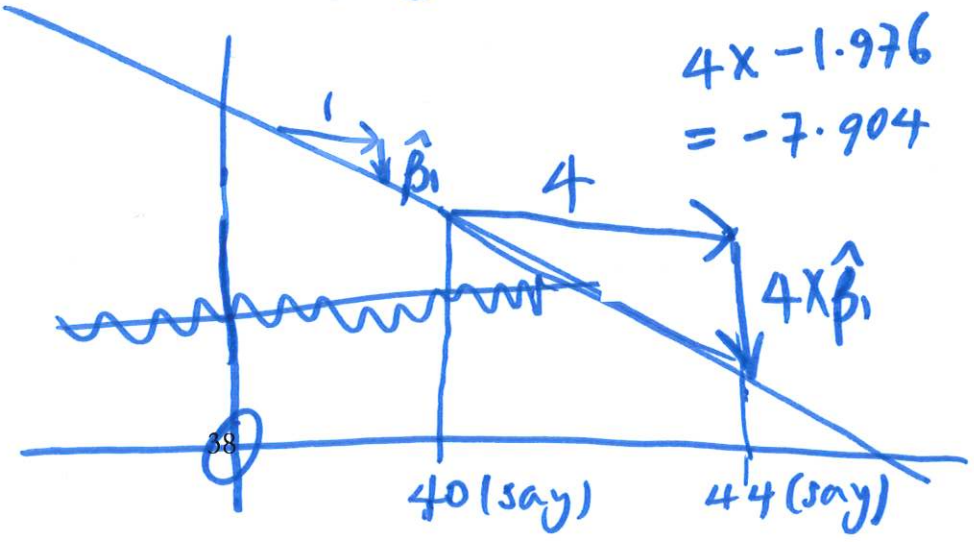
(1) 166.2

(2) 7.9

(3) 2.0

(4) 4.0

(5) 0.4



$4x - 1.976$

$= -7.904$

57. Which **one** of the following statements is the **best** interpretation of the confidence interval (−2.178, −1.773) given in Table 6, page 36?

With 95% confidence we estimate that:

F (1) each additional second taken to walk the 50 metres is associated with an average decrease in total psychometric score of somewhere between 88.7 and 108.9.

F (2) the y-intercept of the true regression line is somewhere between −2.2 and −1.8.

T (3) the slope of the true regression line is somewhere between −2.2 and −1.8.

T (4) each additional second taken to walk the 50 metres is associated with an average decrease in total psychometric score of somewhere between 1.8 and 2.2.

F (5) when the total psychometric score is zero, the time taken to walk the 50 metres is, on average, somewhere between 88.7 and 108.9 seconds.

58. Suppose another participant who completed the task in 20 seconds and had a total psychometric score of 90 was added to the study. When conducting a linear regression analysis, consider the effect of including this participant on the estimates of the slope and the correlation co-efficient. In this context, an increase in the slope means the slope is steeper and a decrease means it is flatter.

Any effect would be to slightly:

(1) increase the slope and decrease the magnitude (size) of the correlation co-efficient.

(2) increase the slope and increase the magnitude (size) of the correlation co-efficient.

(3) decrease the slope and the magnitude (size) of the correlation co-efficient.

(4) decrease the slope but keep the same correlation co-efficient.

(5) decrease the magnitude (size) of the correlation co-efficient but keep the same slope.

incr. scatter → decr. r

flatten line → decr. slope

59. Define

the **confidence interval for the mean** as the 95% confidence interval that results from estimating the mean score when the time taken to complete the task is 30 seconds

and

the **prediction interval** as the 95% prediction interval that results from estimating an individual's score when the time taken to complete the task is 30 seconds.

Which **one** of the following statements about the **confidence interval for the mean** and the **prediction interval** is **true**?

The **confidence interval for the mean**:

F (1) will be wider than the **prediction interval** because it has to take into account the uncertainty about the values of the slope and the intercept of the line whereas the **prediction interval** only takes into account the variation among people whose time to complete the task is 30 seconds.

F (2) will be centred either higher or lower than the **prediction interval**, depending on the estimates about which each interval is centred.

F (3) will be narrower than the **prediction interval** because it only takes into account the variation among people whose time to complete the task is 30 seconds.

F (4) will be wider and centred higher than the **prediction interval** because it has to account for two sources of variability.

T (5) will be narrower than the **prediction interval** because it only takes into account the uncertainty of the fitted line, whereas the prediction interval also takes into account the variation among people whose time to complete the task is 30 seconds.

60. Which **one** of the following statements regarding linear regression analysis is **false**?

T (1) The least squares regression technique minimises the sum of the squared residuals.

T (2) *t*-procedures are used to find confidence intervals for the slope of the true line.

F (3) A strong relationship between two numeric variables is indicated when the slope of the line is close to either 1 or -1.

T (4) The *X*-variable is used to predict or explain the behaviour of the *Y* -variable.

T (5) The two main components of a regression relationship are trend and scatter.

61. Which **one** of the following statements about a simple linear regression analysis is **false**?

(1) It is unwise to make predictions outside the range of the explanatory variable.

(2) The two variables take on special roles – one of the variables is viewed as an explanatory variable and the other as a response variable.

(3) The fitted line, used to summarise the relationship between the two variables, is chosen to maximise the overall size of the residuals.

(4) A single outlier can have a large influence on the position of the least squares regression line.

(5) The scatter plot should show a linear trend.

62. Which **one** of the following is **not** an assumption of the simple linear regression model?

(1) The random errors are Normally distributed with a mean of zero.

(2) The random errors have the same standard deviation, regardless of the value of x.

(3) ~~Each value of Y is independent of its value at X.~~

(4) There is a linear relationship between Y and X.

(5) The random errors are all independent.

63. Which **one** of the following statements is **false**?

(1) The prediction interval for a particular value will always be wider than the confidence interval for the mean.

(2) The estimated slope and intercept from a regression of Y on X will not necessarily be the same as the estimated slope and intercept from a regression of X on Y.

(3) A correlation coefficient, r, of zero indicates that there is no linear relationship between two variables. (but there may be a non-linear relationship).

(4) It is unsafe to predict values outside the range of the observed data.

(5) In a straight line graph, y changes by a fixed amount with each unit change in x.

64. Which **one** of the following statements about checking the assumptions of the simple linear model is **false**?

(1) A scatter plot of Y versus X is useful for checking whether the assumption of a linear relationship between x and E(Y) is reasonable.

(2) A scatter plot of Y versus X is useful for checking for the presence of outliers.

(3) A residual plot is useful for checking whether the assumption of independence of the errors is reasonable.

(4) A residual plot is useful for checking whether the assumption of a linear relationship between x and E(Y) is reasonable.

(5) A residual plot is useful for checking whether the assumption that the random errors all have the same standard deviation, regardless of the value of x, is reasonable.

65. Which **one** of the following statements about the assumptions in the simple linear model is **false**?

(1) The random errors are Normally distributed.

(2) The random errors are all independent.

(3) The random errors have a mean of zero.

(4) There is a linear relationship between x and the ~~standard deviation~~ mean value of Y at each value of x.

(5) There is a linear relationship between x and the mean value of Y at $X = x$.

ANSWERS

1. **(2)**	2. **(3)**	3. **(4)**	4. **(1)**	5. **(4)**	6. **(5)**
7. **(3)**	8. **(4)**	9. **(5)**	10. **(3)**	11. **(3)**	12. **(2)**
13. **(4)**	14. **(4)**	15. **(2)**	16. **(3)**	17. **(2)**	18. **(5)**
19. **(1)**	20. **(3)**	21. **(4)**	22. **(2)**	23. **(3)**	24. **(2)**
25. **(4)**	26. **(1)**	27. **(1)**	28. **(4)**	29. **(5)**	30. **(3)**
31. **(3)**	32. **(5)**	33. **(2)**	34. **(3)**	35. **(5)**	36. **(5)**
37. **(5)**	38. **(3)**	39. **(5)**	40. **(3)**	41. **(3)**	42. **(1)**
43. **(2)**	44. **(4)**	45. **(5)**	46. **(2)**	47. **(5)**	48. **(4)**
49. **(2)**	50. **(3)**	51. **(1)**	52. **(2)**	53. **(3)**	54. **(4)**
55. **(1)**	56. **(2)**	57. **(4)**	58. **(3)**	59. **(5)**	60. **(3)**
61. **(3)**	62. **(3)**	63. **(3)**	64. **(3)**	65. **(4)**	

WHAT SHOULD I DO NEXT?

- Do all the problems in this workshop handout and mark them. If you get a question wrong, have a look at the working on Leila's scanned slides at www.tinyURL.com/stats-RC to see how she did it.
- Go through the Chapter 10 blue pages. This includes:
  - the *notes* on pages 17 to 19,
  - the *glossary* on page 20,
  - the *true/false statements* on page 21,
  - the *Sample Exam Questions* on pages 22 & 23,
  - the *true/false statements* on page 24,
  - the *Sample Exam Questions* on pages 25 & 26,
  - the *Short Response Questions* and *Final Challenge* on page 27, and
  - the *tutorial* material on pages 28 & 29.
- Attend the optional Chapter 10 tutorial.
- Try the **PRACTICE Ch10 Quiz**.
- Do three attempts of the **Chapter 9 Quiz** (due at 11pm on Wed 4 Nov 2020).
- Do the Chapter 10 parts of the Revision Assignment. Get this from Canvas under *Assignments*. **Note that this assignment is not one of the formal assessments for your final mark & grade so it is not to be handed in!**
- Try Chapter 10 questions from three of the past five exams on Canvas (get them from *Modules* → *Past Tests and Exams (with answers)* and use the *Exam questions index* document from there to identify the Chapter 10 questions!)
- If you get anything wrong and don't know why, get some help. You can post a question on Piazza (search first as it may have already been asked!), or talk to someone about it (your lecturer, an Assistance Room tutor or Leila).

FORMULAE

Confidence intervals and t-tests

Confidence interval: $\quad estimate \pm t \times \text{se}(estimate)$

t-test statistic: $\quad t_0 = \dfrac{estimate - hypothesised\ value}{standard\ error}$

Applications:

1. Single mean μ: $\quad estimate = \overline{x}; \quad df = n - 1$
2. Single proportion p: $\quad estimate = \widehat{p}; \quad df = \infty$
3. Difference between two means $\mu_1 - \mu_2$: (independent samples)

   $estimate = \overline{x}_1 - \overline{x}_2; \quad df = \min(n_1 - 1, n_2 - 1)$
4. Difference between two proportions $p_1 - p_2$:

   $estimate = \widehat{p}_1 - \widehat{p}_2; \quad df = \infty$

   Situation (a): *Proportions from two independent samples*

   Situation (b): *One sample of size n, several response categories*

   Situation (c): *One sample of size n, many yes/no items*

The F-test (ANOVA)

F-test statistic: $f_0 = \dfrac{s_B^2}{s_W^2}; \quad df_1 = k - 1, \ df_2 = n_{\text{tot}} - k$

The Chi-square test

Chi-square test statistic: $\chi_0^2 = \displaystyle\sum_{\text{all cells in the table}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$

Expected count in cell $(i, j) = \dfrac{R_i C_j}{n}$

$df = (I - 1)(J - 1)$

Regression

Fitted least-squares regression line: $\quad \widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$

Inference about the intercept, β_0, and the slope, β_1: $\quad df = n - 2$