Q1-6

Practice Questions

Questions 1 to **4** refer to the following set of plots:

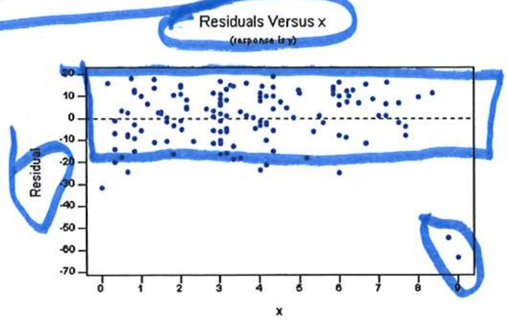


Figure A

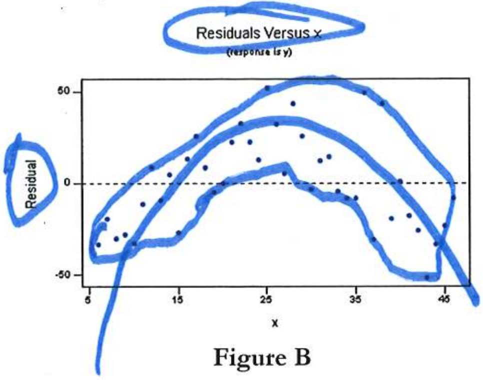


Figure B

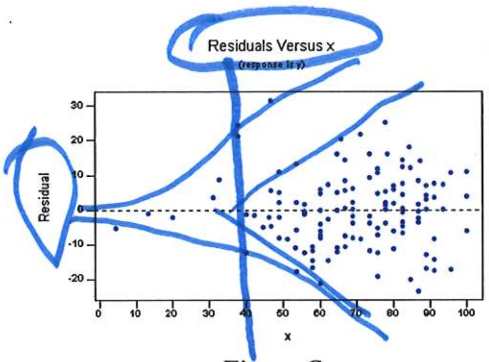


Figure C

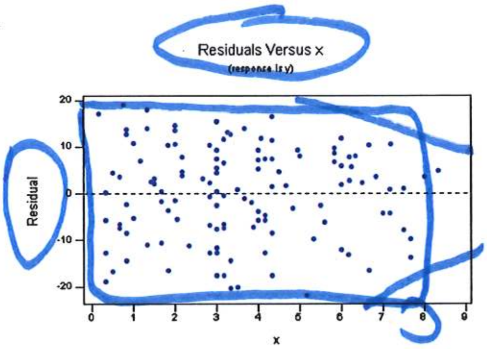


Figure D

In each of the above plots determine whether or not there any problems with the assumptions underlying linear regression model:

1. Figure A:

   (1) No problems – there is roughly a horizontal patternless band

   [✓ (2)] Normality; Outliers

   (3) Non-linear

   (4) Non-constant scatter – this implies that the error variability is not independent of x

   ~~(5)~~ Observations are not independent can't be checked!

2. Figure B:

   (1) No problems – there is roughly a horizontal patternless band

   (2) Normality; Outliers

   [✓ (3)] Non-linear

   (4) Non-constant scatter – this implies that the error variability is not independent of x

   ~~(5)~~ Observations are not independent ditto

3. Figure C:

   (1) No problems – there is roughly a horizontal patternless band

   (2) Normality; Outliers

   (3) Non-linear

   [✓ (4)] Non-constant scatter – this implies that the error variability is not independent of x

   ~~(5)~~ Observations are not independent ditto

4. Figure D:

   [✓ (1)] No problems – there is roughly a horizontal patternless band

   (2) Normality; Outliers

   (3) Non-linear

   (4) Non-constant scatter – this implies that the error variability is not independent of x

   ~~(5)~~ Observations are not independent ditto

5. The type of plot used to analyse variables in a regression model is a:

   (1) Side-by-side dot plot

   (2) Side-by-side box plot

   (3) Table of counts

   [✓ (4)] Scatterplot

   (5) Histogram

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

   T (1) A relationship between two numeric variables may look weak because it has been plotted over only a limited range of *x*-values.

   T (2) When exploring the relationship between two numeric variables, precise prediction cannot be made from a weak relationship.

   T (3) If we wish to explore the relationship between a categorical and a numeric variable, we plot the values of the numeric variable for each group against the same scale.

   T (4) Cross-tabulation is a process of recording count data when we have two categorical variables.

   F [✓ (5)] In regression the explanatory variable is the variable explained by the response variable.

Questions 7 and 8 are about the following information.

The course lecturers for a university course wanted to investigate the strength of the linear relationship between marks of the two sections of the test, Section A and Section B. Figure 3 below shows a scatter plot of the data.

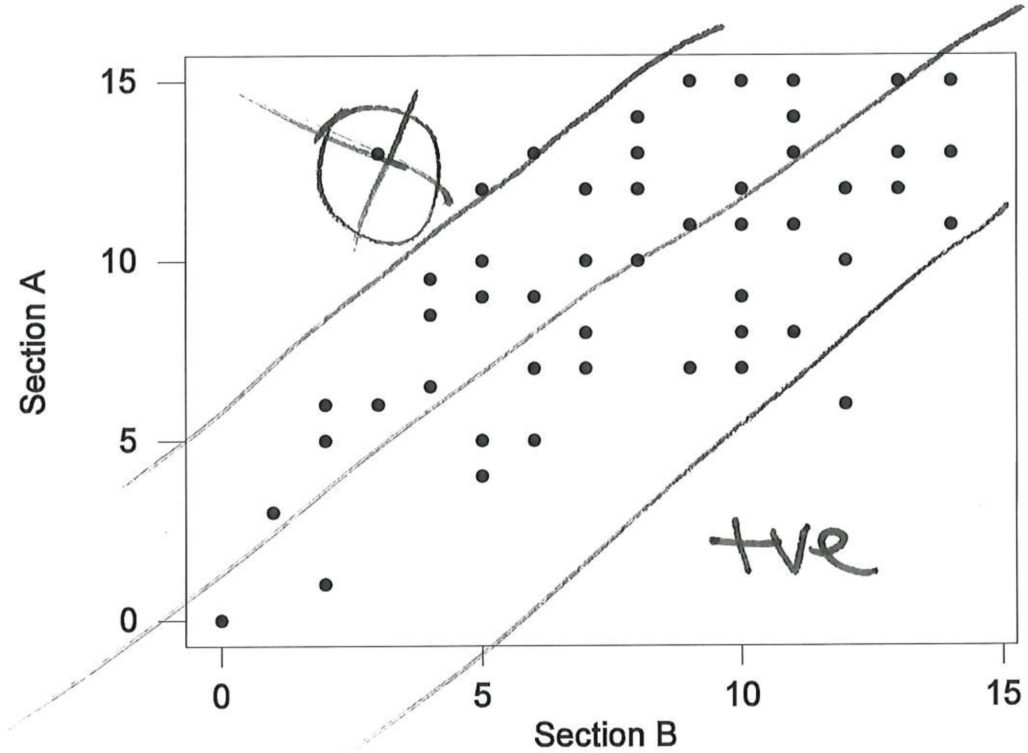


Figure 3: Scatter plot of marks in course test

moderate strength

7. The sample correlation coefficient for the relationship between Section A marks and Section B marks is $r = 0.653$. Which **one** of the following statements is the correct interpretation of this value of r?

(1) The linear relationship between Section A marks and Section B marks is so weak it is not worth studying.

(2) The linear relationship between Section A marks and Section B marks is positive and very strong.

[✓ (3)] The linear relationship between Section A marks and Section B marks is positive and weak to moderate.

(4) Each increase of one mark in Section B is associated with an increase of 0.653 marks in Section A. Regression model has the slope!

(5) The linear relationship between Section A marks and Section B marks is negative and weak to moderate.

(not correlation...)

8. Suppose that on further investigation it was found that the student who scored 13 marks in Section A and 3 marks in Section B was ill during the test and had to leave without completing Section B. It was decided to remove this observation from the analysis and recalculate the sample correlation coefficient. → stronger relationship outlier/outside value!

Which **one** of the following statements is **true**?

(1) It is impossible to determine how the recalculated sample correlation coefficient would compare with the original value of 0.653.

(2) The recalculated sample correlation coefficient would increase because the ~~slope~~ of the new fitted line would be greater than the slope of the original fitted line. → regression!

(3) The recalculated sample correlation coefficient would decrease because the ~~slope~~ of the new fitted line would be less than the slope of the original fitted line.

[✓ (4)] The recalculated sample correlation coefficient would increase because the data would more closely fit a straight line with a positive slope.

(5) The recalculated sample correlation coefficient would decrease because the data would more closely fit a straight line with a negative slope.

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

T (1) A prediction interval for another observation whose x-value is well outside the range of observed values is potentially unreliable.

T (2) For weak relationships, the width of 95% prediction intervals will be so large that the intervals are of little practical use.

T (3) For a specified value of x, the width of a confidence interval for the mean allows for uncertainty in the estimates of β_0 and β_1.

T (4) For a specified value of x, the width of a prediction interval for another observation allows for uncertainty in the estimates of β_0 and β_1 and for uncertainty due to the random scatter about the line.

F [✓ (5)] For a specified value of x, the 95% confidence interval for the mean is ~~wider~~ narrower than the associated 95% prediction interval.

8Qs ≈ 24mins

Questions 10 to **17 refer to the following information.**

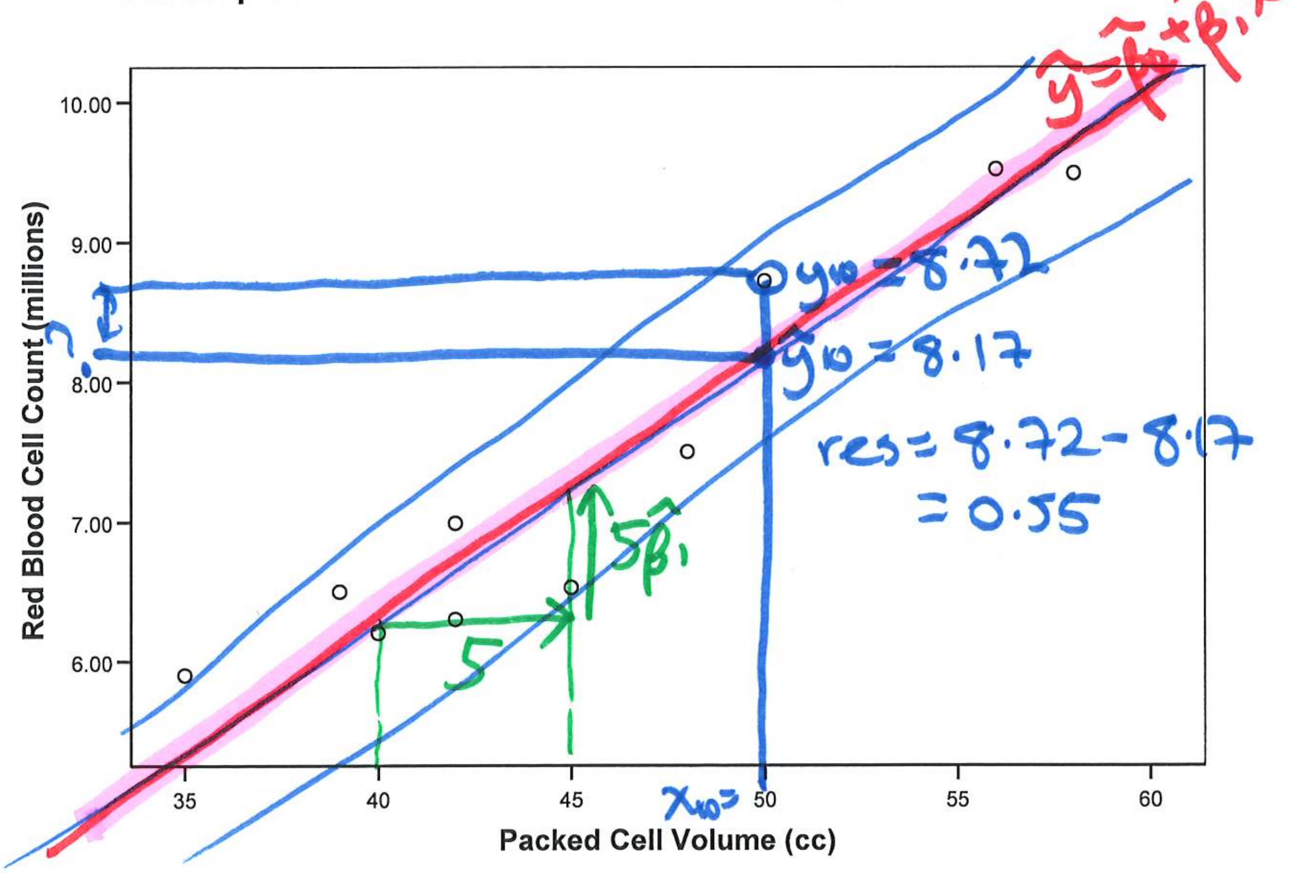
Counting the number of red blood cells in a sample of blood using a microscope is a difficult and time-consuming task. However, the packed cell volume is much easier to measure. To find a possible relationship between these two variables, blood samples are taken for 10 dogs. The following data was obtained:

N X Y

Packed cell volume (cc)	Red blood cell count (millions)
$x_1 =$ 45	$y_1 =$ 6.53
$x_2 =$ 42	$y_2 =$ 6.30
56	9.52
48	7.50
42	6.99
35	5.90
58	9.49
40	6.20
39	6.50
$x_{10} =$ 50	$y_{10} =$ 8.72

A scatter plot and some computer output of these data are given below:

Scatter plot of Red Blood Cell Count versus Packed Cell Volume



1cc: (.131, .223)
5cc: (.655, 1.115)
10cm: (1.31, 2.23)

Regression

Coefficients(a)

Model $\hat{\beta}$s		Unstandardized Coefficients		Standardized Coefficients	t_0	p-val Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-.680	.918	r	-.741	.480	-2.796	1.436
	Packed Cell Volume (cc)	.177	.020	.953	8.871	.000	(.131 ,	.223)

a Dependent Variable: Red Blood Cell Count (millions)

β_0, y-intercept

β_1, slope

x

y

$se(\hat{\beta}_1)$

$t_0 = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)} = \frac{.177}{.02} = 8.85$

$\boxed{\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x}$ $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$

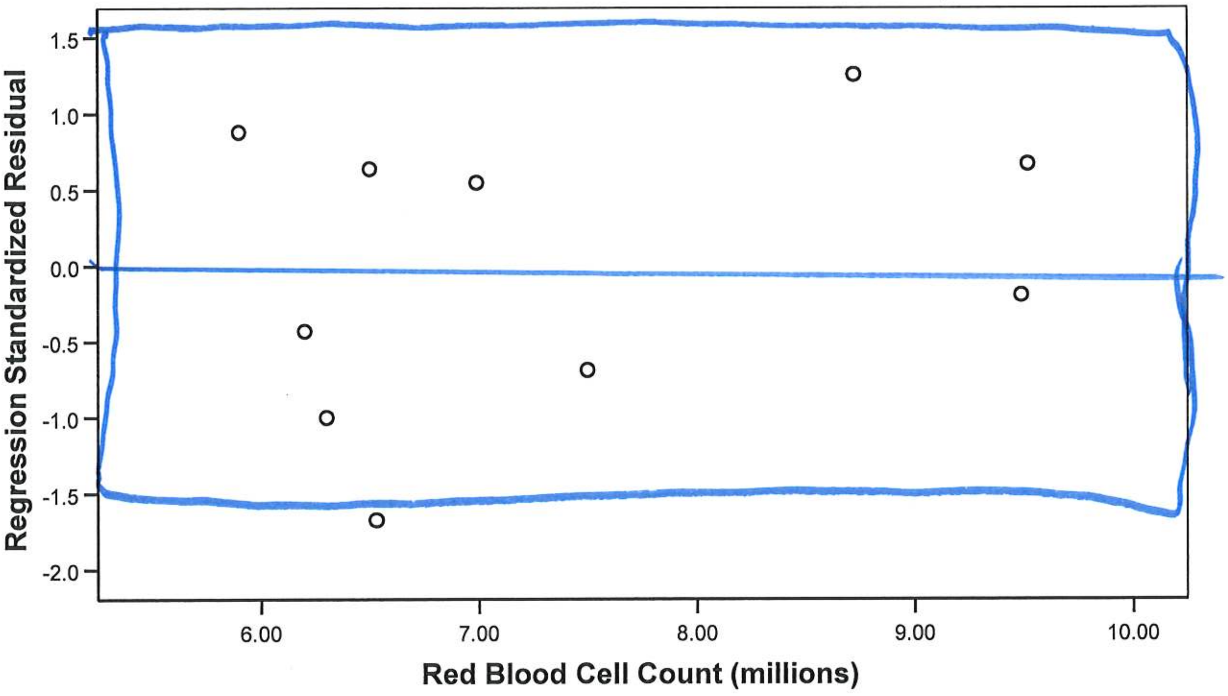
pred. est. exp. av. red blood cell count = -.68 + .177 × packed cell vol.

p-val = .000 → v. st. ev. against H_0!

$\boxed{df = n-2} = 10-2 = 8$

Scatterplot

Dependent Variable: Red Blood Cell Count (millions)



10. The fitted least squares regression line for these data is:

(1) Average Packed Cell Volume = 0.177 × Red Blood Cell Count − 0.68

(2) Average Red Blood Cell Count = 0.177 − 0.68

[✓ (3)] Average Red Blood Cell Count = 0.177 × Packed Cell Volume − 0.68

(4) Average Red Blood Cell Count = 0.177 − 0.68 × Packed Cell Volume

(5) Average Packed Cell Volume = 0.177 − 0.68 × Red Blood Cell Count

11. For the data in the scatter plot on page ~~13~~ 15, which **one** of the following values would be the **closest** to the sample correlation coefficient, *r*?

(1) ~~$r = 0.18$~~
(2) $r = 0.70$
[✓ (3)] $r = 0.95$
(4) ~~$r = 0.48$~~
(5) ~~$r = 0.13$~~

not much scatter

→ st. linear assoc.!

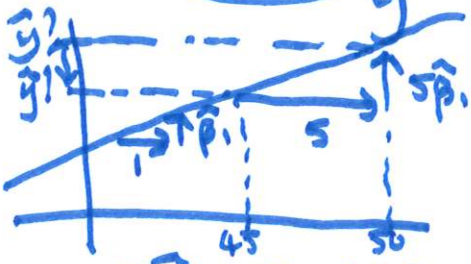
12. The correct null and alternative hypotheses to test that there is no linear relationship between red blood cell count and packed cell volume are:

slope! β_1

~~(1)~~ ~~H_0: $\hat{\beta}_0 = 1$ vs H_1: $\hat{\beta}_0 \neq 1$~~
[✓ (2)] H_0: $\beta_1 = 0$ vs H_1: $\beta_1 \neq 0$
(3) H_0: $\beta_1 = \cancel{1}$ vs H_1: $\beta_1 \neq \cancel{1}$
(4) H_0: $\cancel{\beta_0} = 0$ vs H_1: $\cancel{\beta_0} \neq 0$
~~(5)~~ ~~H_0: $\hat{\beta}_1 = 0$ vs H_1: $\hat{\beta}_1 \neq 0$~~

13. The fitted least squares regression line indicates that for each increase of 5 cubic centimetres in packed cell volume we expect that, on average, the red blood cell count will:

~~(1)~~ decrease by approximately .68 million.
~~(2)~~ decrease by approximately 3.4 million.
~~(3)~~ increase by approximately 3.4 million.
[✓ (4)] increase by approximately .89 million.
(5) decrease by approximately .89 million.



$5\hat{\beta}_1 = 5 \times .177$
$= .885$

14. The fitted least squares regression line can be used to predict the red blood cell count. Dogs with a packed cell volume of 50 cubic centimetres have a predicted red blood cell count of approximately:

(1) 5.03 million
(2) 33.82 million
(3) 34.18 million
[✓ (4)] 8.17 million
(5) 9.53 million

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$= -.68 + .177 \times 50$

$= 8.17$ million

15. The packed cell volume x_{10} and red blood cell count y_{10} for dog number 10 was 50 cubic centimetres and 8.72 million respectively. Under the fitted least squares line, the value of the residual for this dog is approximately:

(1) 2.30 (3) 0.81 (5) −0.81

[✓ (2)] 0.55 → (4) −0.55

$$res = obs - exp = y - \hat{y} = 8.72 - 8.17 = .55$$

16. Which **one** of the following statements is **false**? $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$ } p-val = .000 v. st. ev. against H_0

(1) From the data, we have very strong evidence against there being no linear relationship between packed cell volume and red blood cell count.

(2) From the data, we have very strong evidence of a linear association between packed cell volume and red blood cell count.

[✓ (3)] From the data, we have very strong evidence that there is ~~no~~ a linear relationship between packed cell volume and red blood cell count.

(4) From the data, we have very strong evidence of a linear relationship between packed cell volume and red blood cell count.

(5) From the data, we have very strong evidence of a (positive) linear relationship between packed cell volume and red blood cell count.

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

T (1) With 95% confidence, we estimate from the data that, on average, an increase of 5 cubic centimetres in the packed cell volume is associated with an increase in red blood cell count of between .66 and 1.12 million.

[✓ F ✗ (2)] With 95% confidence, we estimate from the data that an increase of 5 cubic centimetres in the packed cell volume is associated with an increase in red blood cell count of between .66 and 1.12 million.

T (3) With 95% confidence, we estimate from the data that an increase of 5 cubic centimetres in the packed cell volume is associated with an increase in the mean red blood cell count of between .66 and 1.12 million.

T (4) With 95% confidence, we estimate from the data that, on average, an increase of 10 cubic centimetres in the packed cell volume is associated with an increase in red blood cell count of between 1.31 and 2.23 million.

T (5) With 95% confidence, we estimate from the data that an increase of 10 cubic centimetres in the packed cell volume is associated with an increase in the mean red blood cell count of between 1.31 and 2.23 million.