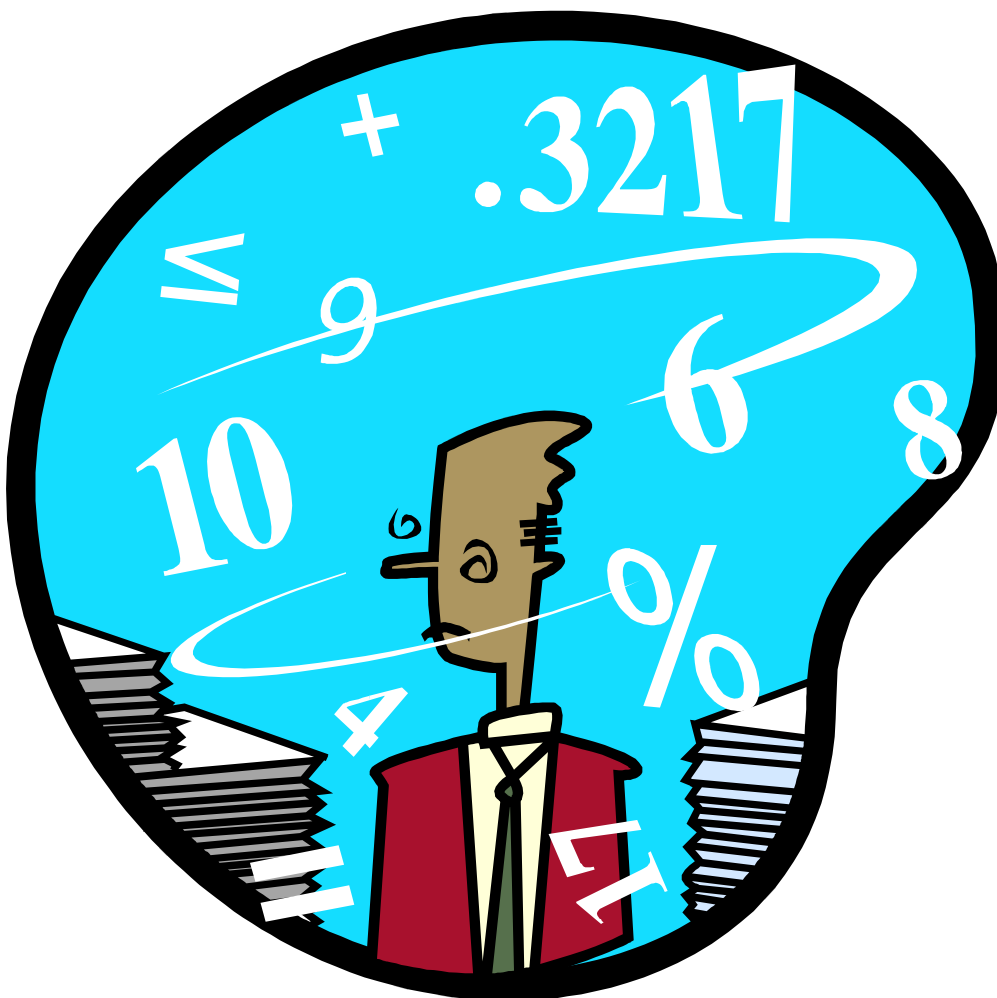


STATS 10X WORKSHOP

EXAM PREP 6: CHAPTER 12

MON 25 OCT & MON 1 Nov 2010



Students **MUST REGISTER** for all workshops with
The Student Learning Centre, 3rd Floor, Information Commons

Statistical help available at the SLC

The Student Learning Centre (SLC) offers help for STATS 10x by offering:

- one-on-one tutoring help, and
- a number of workshops

One-on-one help over S2 2010 including exam period

One-on-one assistance for STATS 10x is available at the SLC. Check appointment availability and book at SLC reception in person (third floor, Information Commons building) or by calling 373-7599 ext. 88850.

Note: SLC tutors are not allowed to help students complete their assignments.

SLC STATS 10x Exam Prep Workshops

Any questions regarding STATS 10x workshops should be forwarded to:

Leila Boyle; SLC Statistics Co-ordinator: l.boyle@auckland.ac.nz

These twelve workshops (six different sessions, each repeated twice) are held prior to the exam, from Saturday 2 October until Monday 1 November 2010 (inclusive).

These workshops concentrate on questions reviewing the **basic concepts**, rather than questions on finer details. They are designed to assist students to achieve a pass and **don't cover all material**.

The timetable for these workshops is available at this workshop, at SLC Reception and on Leila's website. Please enrol in each of your preferred workshops by EITHER:

- ***Dropping by the SLC Reception to enrol in person (Room 320, Level 3, Information Commons Building, 11 Symonds Street) OR***
- ***Emailing slc@auckland.ac.nz with your name, ID number, and the name, date and time of the workshop/s you wish to attend OR***
- ***Calling the SLC Reception on 373-7599 ext. 88850 and book over the phone.***

Useful Websites

- SLC webpage: www.slc.auckland.ac.nz
- Cecil: <https://cecil.auckland.ac.nz>
- **Leila's website for STATS 10x SLC workshop handouts & information:** www.stat.auckland.ac.nz/~leila

Revision Notes

Chapter 12 – Simple Linear Regression

Look at blue pages for good notes and test/exam questions for practice

The main tool for comparing two quantitative variables is the scatter plot.

What to look for in a scatter plot:

- Trend (pattern)
- Scatter
- Outliers
- Association
- Strength of the relationship
- Groupings

Regression

Regression looks at the relationship between two quantitative variables where the two variables take on special roles:

- X is used to **explain** or **predict** the behaviour of Y
- X is the **explanatory** or **independent** variable
- Y is the **dependent** or **response** variable

Two main components of the regression model are:

- **trend** and
- **scatter**.

We use a **least squares regression line** $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ fitted by the computer / calculator to estimate the unknown population parameters β_0 and β_1

The single **least squares regression line** for each linear regression model:

- minimises the sum of the squared residuals/prediction errors
- has $\sum \text{residuals} = 0$ (but so do many other lines)
- has (\bar{x}, \bar{y}) lying on it

• Residuals

- Errors, residuals or prediction errors are all terms for the same thing.
- A residual is the (vertical) distance between the **actual observed value** y_i and the **expected estimated value** \hat{y}_i , i.e.:

$$\text{Errors} = \text{observed} - \text{expected} \quad (\hat{u}_i = y_i - \hat{y}_i)$$



• **Hypotheses**

$H_0: \beta_1 = 0$ (there **is no** linear relationship)

$H_1: \beta_1 \neq 0$ (there **is a** linear relationship)

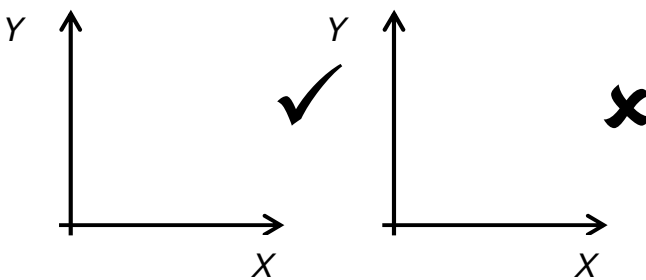
• **Assumptions** of simple linear regression are:

1. There is a **linear** relationship between X and Y .
2. Errors are **Normally** distributed (with $\mu = 0$).
3. Errors all have the **same std deviation**, σ , regardless of the value of x .
4. Errors are all **independent**.

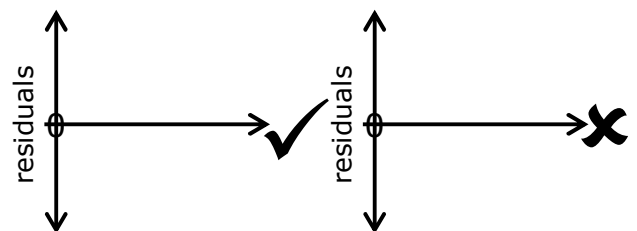
• **Assumption checking using plots of the data and residual plots**

1. There is a **linear** relationship between X and Y .

Scatterplot of data:

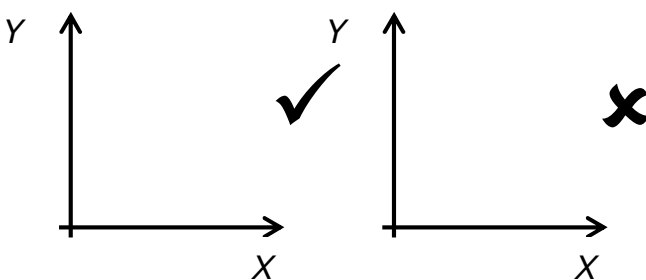


Residual plot:

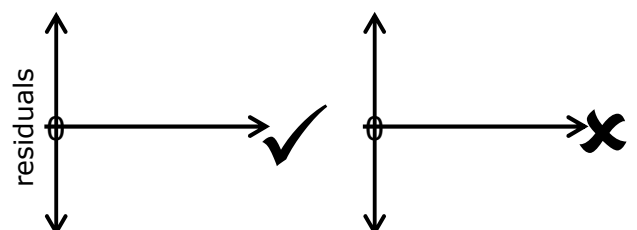


2. Errors are **Normally** distributed (with $\mu = 0$).

Scatterplot of data:

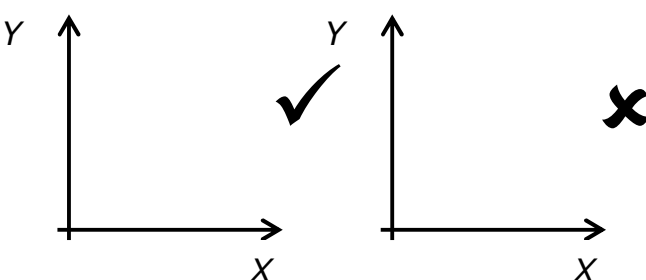


Residual plot:

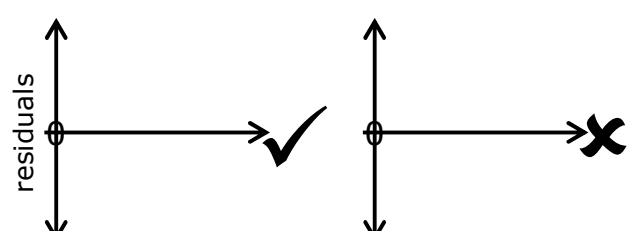


3. Errors all have the **same std deviation**, σ , regardless of the value of x .

Scatterplot of data:



Residual plot:





- **Degrees of Freedom** $df = n - 2$

- **Estimating / Predicting**
 - ✓ Within the range of our observed X -values this can be done with confidence. Predicting outside the range of our observed X -values is dangerous. A relationship that fits the data well may not extend outside that range.

✓ Confidence Interval (for the mean) This estimates the mean Y-value at a specified value of x. The width of the interval allows for:<ul style="list-style-type: none">○ uncertainty about the values of β_0 and β_1.	✓ Prediction Interval This predicts the Y-value for an individual with a specified value of x. The width of the interval allows for:<ul style="list-style-type: none">○ uncertainty about the values of β_0 and β_1 and○ uncertainty due to the random scatter about the line.
--	---

 - ✓ For a given value of x , the **95% prediction interval** is **always wider** than the **95% confidence interval for the mean**.

 - ✓ **The Sample Correlation Coefficient, r**
 - ✓ r measures the **strength** and **direction** of the **linear** association between **two quantitative variables**
 - ✓ The value of r is the same if the axes are swapped around – it doesn't matter which variable is X and which one is Y as r **treats both variables equally**
 - ✓ r measures how close the points in the scatter plot of Y against X (or vice versa) **come to lying on a straight line**
 - $r = 1$, then X and Y have a **perfect positive linear relationship**
 - $r = -1$, then X and Y have a **perfect negative linear relationship**
 - $r = 0$, then X and Y have **no linear** relationship but they may have **some other non-linear relationship**
 - ✓ r has no units – a computer / calculator can give you the value of r
 - ✓ **Correlation DOES NOT imply causation**



Chapter 12 – Questions

1. 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

2. Which one of the following statements is **false**?
 - (1) A relationship between two quantitative variables may look weak because it has been plotted over only a limited range of x -values.
 - (2) When exploring the relationship between two quantitative variables, precise prediction cannot be made from a weak relationship.
 - (3) If we wish to explore the relationship between a qualitative and a quantitative variable, we plot the values of the quantitative variable for each group against the same scale.
 - (4) Cross-tabulation is a process of recording count data when we have two qualitative variables.
 - (5) In regression the explanatory variable is the variable explained by the response variable.

3. Which one of the following statements about simple linear regression analysis is **false**?
 - (1) The least-squares regression line is found by choosing the line that minimises the sum of the squared prediction errors.
 - (2) When a least-squares regression line is fitted to the data, the sum of the prediction errors is zero.
 - (3) For a particular x -value, the 95% prediction interval for the next actual Y -value is generally narrower than the 95% confidence interval for the mean of Y .
 - (4) For a particular x -value, the standard error used to calculate the prediction interval for Y allows for uncertainty about the true values of the intercept and the slope of the line, as well as the uncertainty due to random scatter about the line.
 - (5) When data from a well designed, well executed, controlled experiment indicate a strong relationship between the two variables, we could have reliable evidence of causation.



4. Which **one** of the following statements is **false**?
- (1) A prediction interval for another observation whose x-value is well outside the range of observed values is potentially unreliable.
 - (2) For weak relationships, the width of 95% prediction intervals will be so large that the intervals are of little practical use.

Section Marks in 340 Test

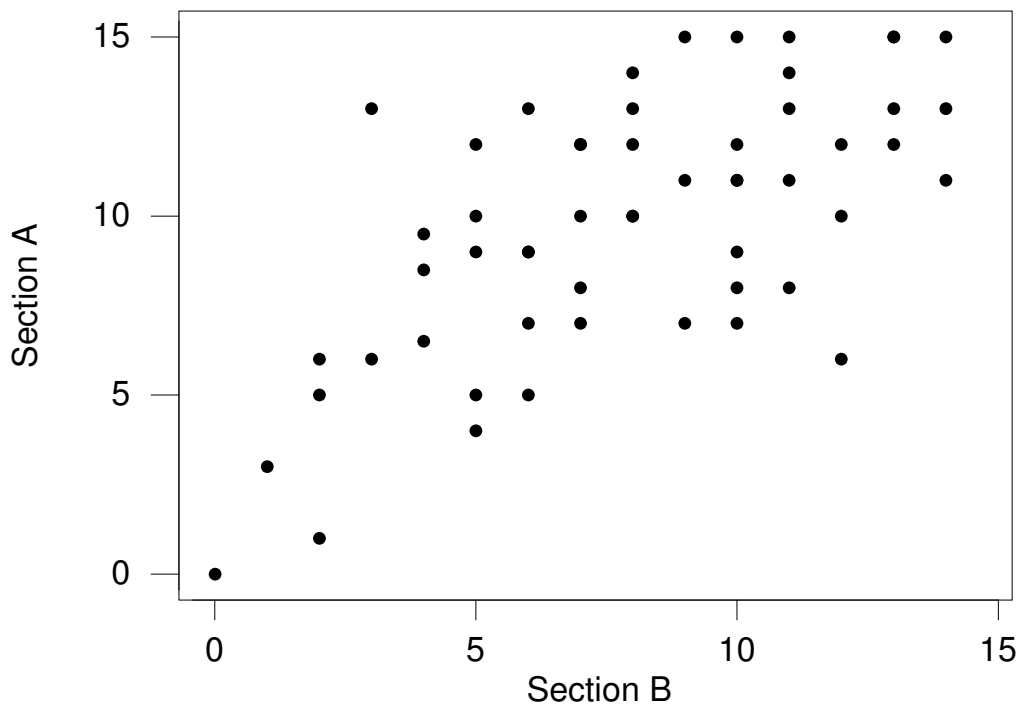


Figure 3: Scatter plot of marks in STATS 340 test



5. 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.
 - (5) The linear relationship between Section A marks and Section B marks is negative and weak to moderate.
6. 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.
- 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.
 - (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) recalculated sample correlation coefficient would decrease because the data would more closely fit a straight line with a negative slope.



Questions 7 to 10 refer to the following set of plots:

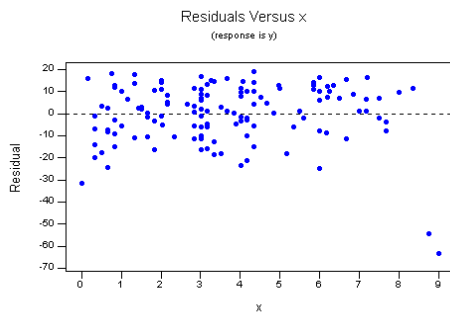


Figure A

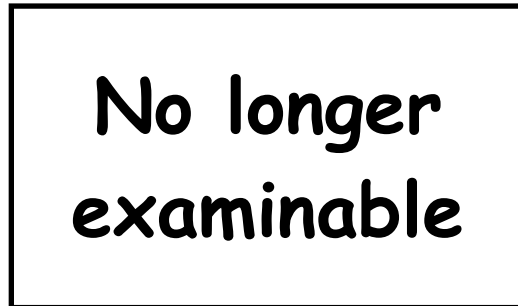


Figure B

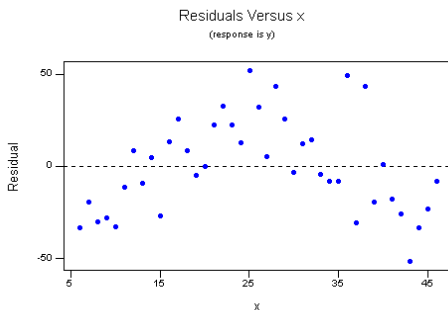


Figure C

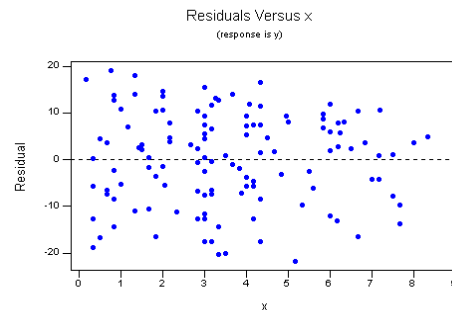


Figure D

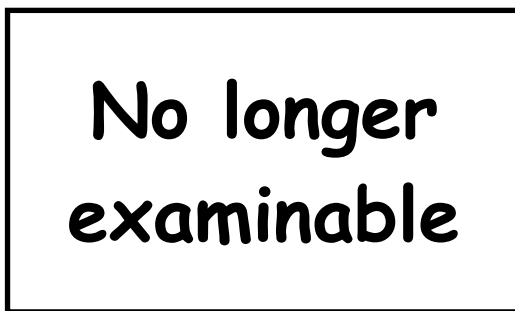


Figure E

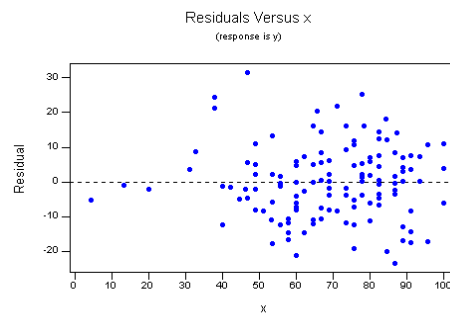


Figure F

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



7. 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

8. 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

9. 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

10. Figure F:

- (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



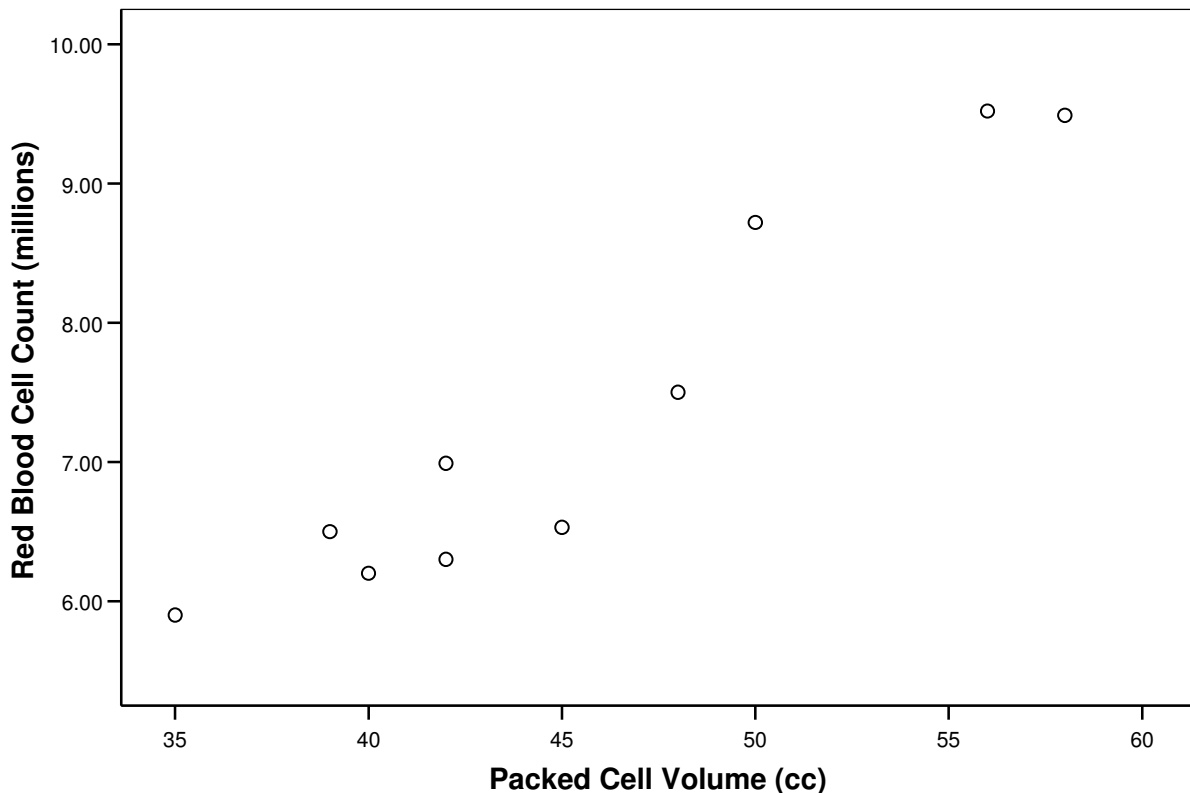
Questions 11 to 18 refer to the following information.

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:

Packed cell volume (cc)	Red blood cell count (millions)
45	6.53
42	6.30
56	9.52
48	7.50
42	6.99
35	5.90
58	9.49
40	6.20
39	6.50
50	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





Regression

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-.680	.918		-.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)

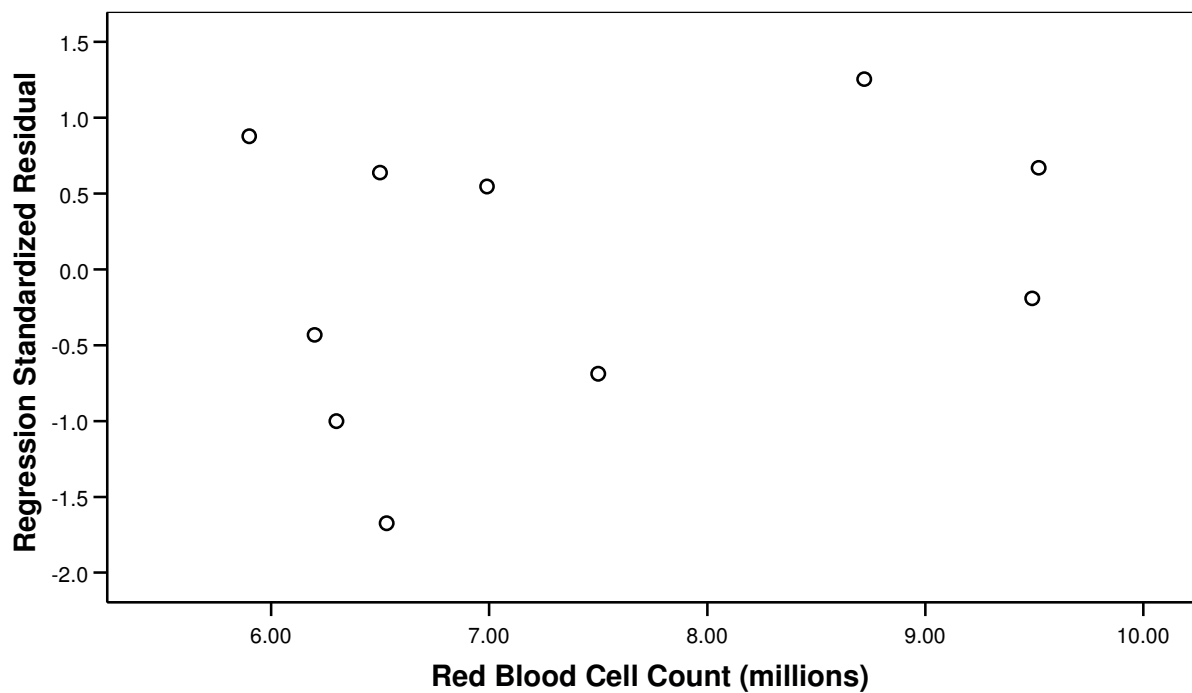
Correlations

		Packed Cell Volume (cc)	Red Blood Cell Count (millions)
Packed Cell Volume (cc)	Pearson Correlation	1	.953(**)
	Sig. (2-tailed)		.000
	N	10	10
Red Blood Cell Count (millions)	Pearson Correlation	.953(**)	1
	Sig. (2-tailed)	.000	
	N	10	10

** Correlation is significant at the 0.01 level (2-tailed).

Scatterplot

Dependent Variable: Red Blood Cell Count (millions)





11. The fitted least squares regression line for these data is:
- (1) $y = 0.177x - 0.68$
 - (2) $\hat{y} = 0.177 - 0.68$
 - (3) $\hat{y} = 0.177x - 0.68$
 - (4) $\hat{y} = 0.177 - 0.68x$
 - (5) $y = 0.177 - 0.68x$
12. The sample correlation coefficient for these data is:
- (1) $r = 0.18$
 - (2) $r = 0.92$
 - (3) $r = 0.95$
 - (4) $r = 0.48$
 - (5) $r = 0.13$
13. The correct null and alternative hypotheses to test that there is no linear relationship between red blood cell count and packed cell volume are:
- (1) $H_0: \hat{\beta}_0 = 1$ and $H_1: \hat{\beta}_0 \neq 1$
 - (2) $H_0: \beta_1 = 0$ and $H_1: \beta_1 \neq 0$
 - (3) $H_0: \beta_1 = 1$ and $H_1: \beta_1 \neq 1$
 - (4) $H_0: \beta_0 = 0$ and $H_1: \beta_0 \neq 0$
 - (5) $H_0: \hat{\beta}_1 = 0$ and $H_1: \hat{\beta}_1 \neq 0$
14. 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 cell blood 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.



15. The fitted least squares regression line can be used to predict the red cell blood count. Dogs with a packed cell volume of 50 cubic centimetres have a predicted red cell blood count of approximately:
- (1) 5.03 million
 - (2) 33.82 million
 - (3) 34.18 million
 - (4) 8.17 million
 - (5) 9.53 million
16. The packed cell volume and red cell blood count 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
 - (2) 0.55
 - (3) 0.81
 - (4) -0.55
 - (5) -0.81
17. Which **one** of the following statements is **false**?
- (1) From the data, we have very strong evidence against there being no relationship between packed cell volume and red cell blood count.
 - (2) From the data, we have very strong evidence of a linear association between packed cell volume and red cell blood count.
 - (3) From the data, we have very strong evidence that there is no linear relationship between packed cell volume and red cell blood count.
 - (4) From the data, we have very strong evidence of a linear relationship between packed cell volume and red cell blood count.
 - (5) From the data, we have very strong evidence of a positive linear relationship between packed cell volume and red cell blood count.



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

- (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 cell blood count of between .66 and 1.12 million.
- (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 cell blood count of between .66 and 1.12 million.
- (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 cell blood count of between .66 and 1.12 million.
- (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 cell blood count of between 1.31 and 2.23 million.
- (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 cell blood count of between 1.31 and 2.23 million.



19. 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 relationship between two variables.
 - (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 .
20. 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.
21. 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 of Y at each value of x .
 - (5) There is a linear relationship between x and the mean value of Y at $X = x$.

Questi

The re
sized e
results
shown

Scatter Plot of Wt versus Eng (Eng less than 2500cc)

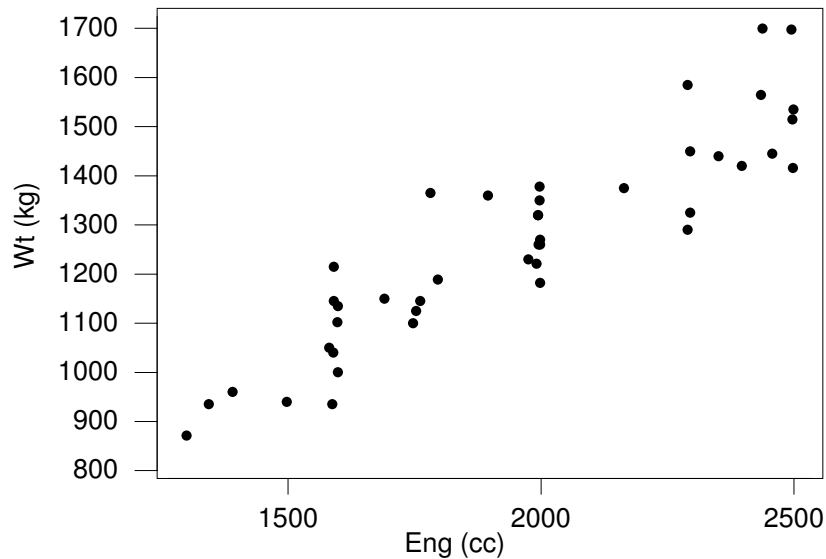


Figure 9: Scatter plot of weight versus engine size for cars with engines smaller than 2500cc

Regression

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients Beta	t	Sig.
		B	Std. Error			
1	(Constant)	235.41	73.68		3.19	.003
	Eng	0.52594	0.03710	.862	14.18	.000

a. Dependent Variable: Wt (kg)

Correlations

		Eng	Wt (kg)
Eng	Pearson Correlation	1	.862 (**)
	Sig. (2-tailed)		.000
	N	43	10
Wt (kg)	Pearson Correlation	.862 (**)	1
	Sig. (2-tailed)	.000	
	N	43	43

** Correlation is significant at the 0.01 level (2-tailed).

Table 14: SPSS output, linear regression analysis of the relationship between weight and engine size

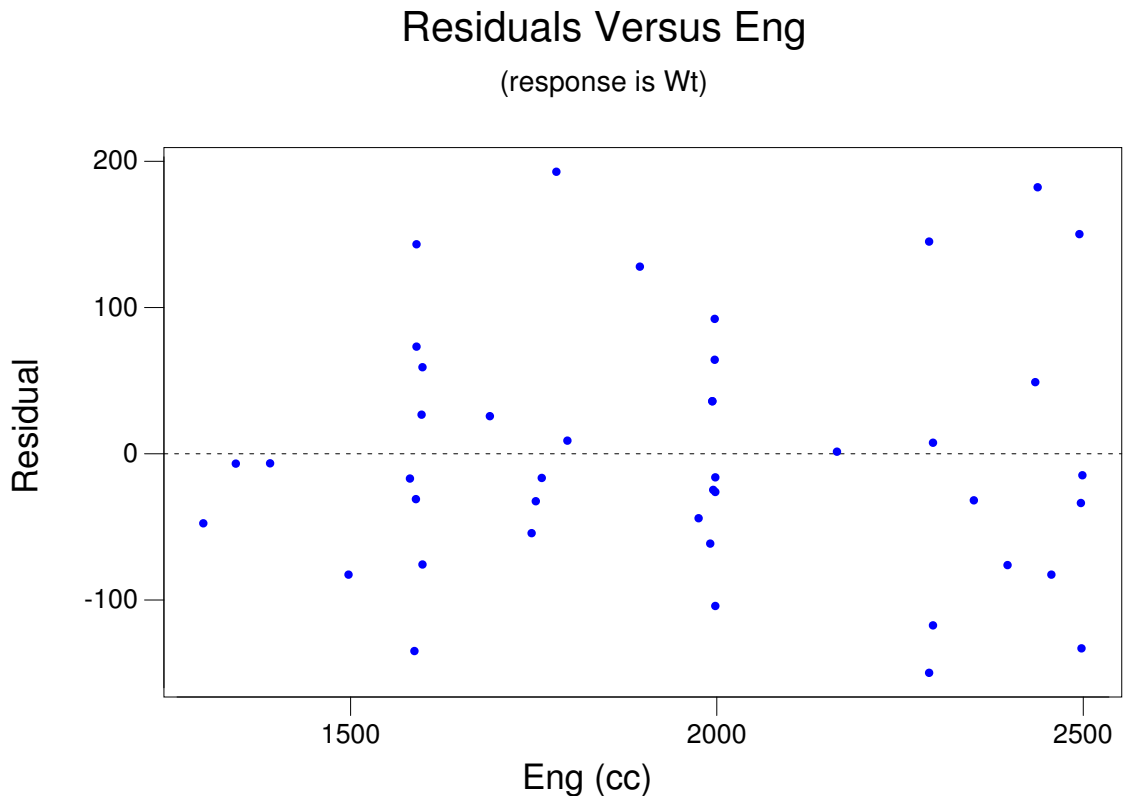


Figure 10: Scatter plot of residuals versus engine size for cars with engines smaller than 2500cc

22. One of the cars in the sample has an engine size of 1590cc and a weight of 1215kg. If a new car has an engine size of 1590cc, the regression equation predicts the car's weight to be approximately:
- (1) 1215kg
 - (2) 1826kg
 - (3) 836kg
 - (4) 1321kg
 - (5) 1072kg
23. Another of the cars in the sample has an engine size of 1497cc and a weight of 940kg. Based on the regression equation, the residual for this car is approximately:
- (1) -83kg
 - (2) 83kg
 - (3) 767kg
 - (4) 1023kg
 - (5) -767kg



24. Suppose that the engine sizes of two cars differ by 500cc. The regression equation predicts that the difference in the weights of these two cars will be:
- (1) 498kg
 - (2) 139kg
 - (3) 263kg
 - (4) 117.5kg
 - (5) 504kg
25. In a test for no linear relationship between engine size and weight the hypotheses are:
- (1) $H_0: \beta_0 \neq 0$ and $H_1: \beta_0 = 0$
 - (2) $H_0: \hat{\beta}_0 = 0$ and $H_1: \hat{\beta}_0 \neq 0$
 - (3) $H_0: \hat{\beta}_1 = 0$ and $H_1: \hat{\beta}_1 \neq 0$
 - (4) $H_0: \beta_1 = 0$ and $H_1: \beta_1 \neq 0$
 - (5) $H_0: \beta_0 = 0$ and $H_1: \beta_0 \neq 0$
26. You may need to refer to Figure 9 and Figure 10 to help answer this question. Which **one** of the following statements about this linear regression analysis is **false**?
- (1) It is reasonable to assume that the error terms have a constant underlying standard deviation.
 - (2) It would be difficult to have faith in a 95% prediction interval for an engine size of 2150cc because there are so few observations with a similar engine size.
 - (3) Engine size is a quantitative variable and weight is a continuous random variable.
 - (4) It would be unwise to use this data to predict the weight of a car with a 3000cc engine.
 - (5) It is believable that the error terms are Normally distributed with a mean of zero.

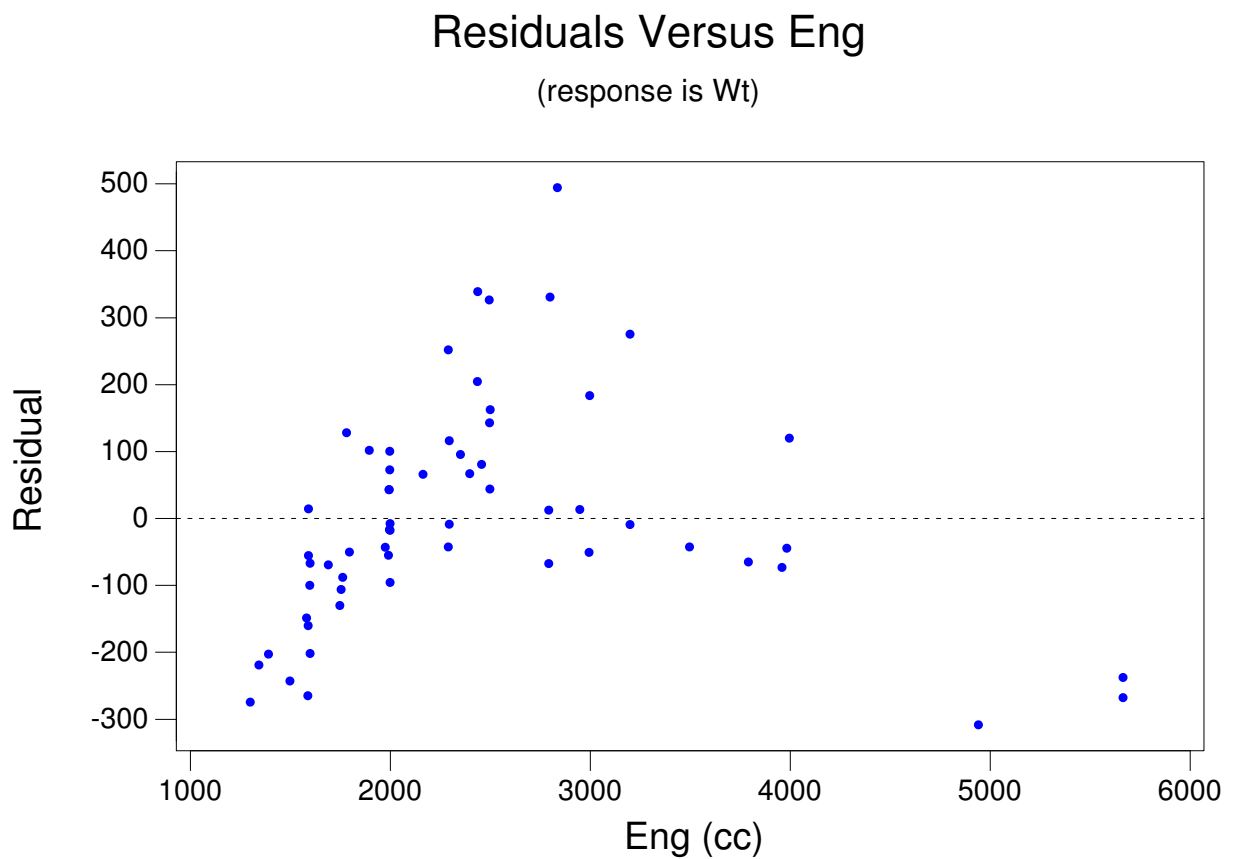


Figure 11: Scatter plot of residuals versus engine size for all cars

ANSWERS

- | | | | | | | | | | | | |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 1. | (4) | 2. | (5) | 3. | (3) | 4. | (5) | 5. | (3) | 6. | (4) |
| 7. | (2) | 8. | (3) | 9. | (1) | 10. | (4) | 11. | (3) | 12. | (3) |
| 13. | (2) | 14. | (4) | 15. | (4) | 16. | (2) | 17. | (3) | 18. | (2) |
| 19. | (3) | 20. | (3) | 21. | (4) | 22. | (5) | 23. | (1) | 24. | (3) |
| 25. | (4) | 26. | (2) | 27. | (3) | | | | | | |