Stats 10x Workshop

Regression and Correlation [RC]

2018

by Leila Boyle

Stats 10x Workshops

The Statistics Department offers workshops and one-to-one/small group assistance for Stats 10x students wanting to improve their statistics skills and understanding of core concepts and topics.

Leila's website for Stats 10x workshop hand-outs and information is here: **www.tinyURL.com/stats-10x**

Resources for this workshop, including pdfs of this hand-out and Leila's scanned slides showing her working for each problem are available here: **www.tinyURL.com/stats-RC**

Want to get in touch with Leila?

Leila Boyle
Undergraduate Statistics Assistance, Department of Statistics
Room 303.320 (third floor of the Science Centre, Building 303)
l.boyle@auckland.ac.nz; (09) 923-9045; 021 447-018

Want help with Stats 10x?

Stats 10x appointments

Book your preferred time with Leila here: **www.tinyurl.com/appt-stats**, or contact her directly (see above for her contact details).

Stats 10x Workshops

Workshops are run in a relaxed environment, and allow plenty of time for questions. In fact, this is encouraged ☺

Please make sure you bring your calculator with you to all of these workshops!

- **Preparation at the beginning of the semester:**

Multiple identical sessions of a preparation workshop are run at the beginning of the semester to get students off to a good start – come along to whichever one suits your schedule!

  - Basic Maths and Calculator skills for Statistics **www.tinyURL.com/stats-BM**

- **First half of the semester**

Five theory workshops are held during the first half of the semester:

  - Exploratory Data Analysis **www.tinyurl.com/stats-EDA**
  - Proportions and Proportional Reasoning **www.tinyURL.com/stats-PPR**
  - Observational Studies, Experiments, Polls and Surveys **www.tinyURL.com/stats-OSE**
  - Confidence Intervals: *Means* **www.tinyURL.com/stats-CIM**
  - Confidence Intervals: *Proportions* **www.tinyURL.com/stats-CIP**

Useful Computer Resource:

If you haven't used SPSS before, try working your way through this self-paced tutorial: **www.tinyURL.com/stats-IS**

Second half of the semester

Four theory workshops and one computing workshop are held during the second half of the semester:

- **Statistics Theory Workshops**
  - Hypothesis Tests: *Proportions* **www.tinyURL.com/stats-HTP**
  - Hypothesis Tests: *Means* **www.tinyURL.com/stats-HTM**
  - Chi-Square Tests **www.tinyURL.com/stats-CST**
  - Regression and Correlation **www.tinyURL.com/stats-RC**
- **Computer Workshop:** Hypothesis Tests in *SPSS* **www.tinyURL.com/stats-HTS**

Regression and Correlation

Types of Variables

Numeric
or Scale (SPSS)
(measurements and counts)

Continuous
(measurements; few repeated values)

No gaps between possible values, if measured sufficiently accurately.

Examples
Weight; Height; Length; Temperature; Total course mark (out of 100); Age [large range]

Discrete
(usually counts; many repeated values)

Gaps between each of the values it can take.

Examples
Number of siblings; Number of cars per household; Assignment mark (out of 10); Age [small range]

Categorical
(define groups)

Nominal (SPSS)
(no idea of order)

Examples
Gender; Ethnicity; Degree (BA, BCom, BSc, etc.); Major (Chemistry, Statistics, etc.)

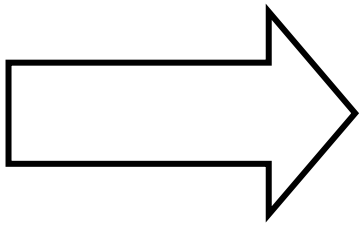
Ordinal (SPSS)
(fall in natural order)

Examples
Size (small, medium, large); Income bracket; Grade (A+, A, …); Age group (<25, 25-39, 40+)

📖**Useful reference:** Chance Encounters, pages 40 – 42

The main tool for comparing two numeric variables is the scatter plot.
What to look for in a scatter plot:

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



Exploring relationships between two variables

Two Numeric	Numeric versus Categorical	Two Categorical
Exploratory tools to use to explore relationship/s		
Scatter plot	Side-by-side plots on the same scale: any n { • Dot plots $n \geq 20$ { • Stem-and-leaf plots • Box plots $n \geq 50$ { • Histograms	Two-way table of counts and/or Bar graphs of proportions (on rows/columns)
Features to look for:		
• Trend linear or non-linear • Scatter constant or non-constant • Strength of relationship strong or weak • Association positive or negative • Outliers • Groupings All / Females only / Males only	Compare the groups by looking at: • Any group differences: ◦ averages (centres) → medians / means ◦ variability (spread) → IQRs / ranges ◦ shapes ♦ symmetric/skewed → left/negative / right/positive ♦ modes → unimodal / bimodal / trimodal (not box!) • Details of individual groups: ◦ outliers, gaps, clusters, groupings Think about reasons *why* these differences, similarities and features are seen	• Most common and least common combinations • Differences in distributions (e.g. row and/or column bar graphs)

Regression

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

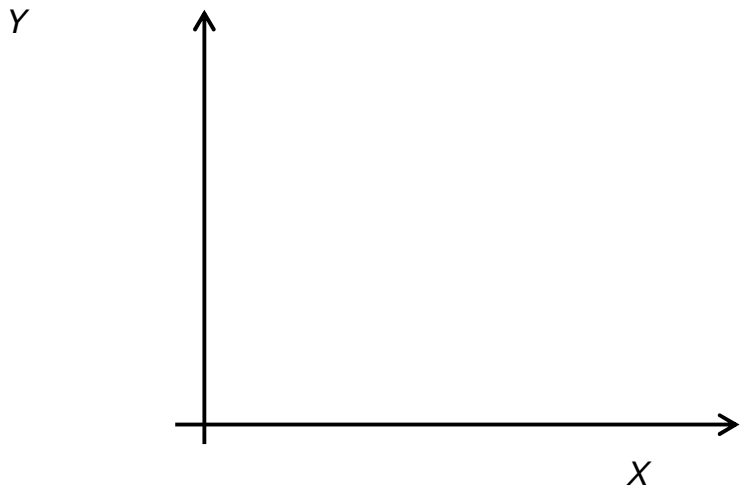
- X is used to **explain** or **predict** the behaviour of Y
- X is the **explanatory** or **independent** variable
- Y is the **dependent** or **response** variable

The two main components of the regression model are:

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

see back page for Formulae Sheet

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.

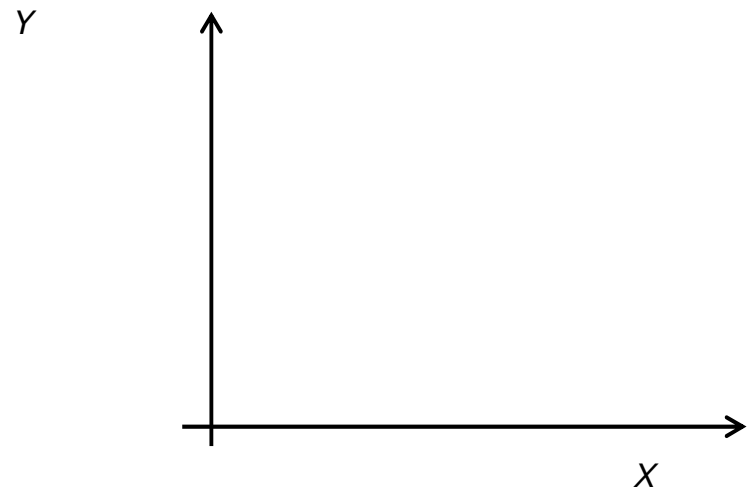


The single **least squares regression line** for each linear regression model:

- minimises the sum of the squared residuals/prediction errors
- has $\sum$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.:

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



- **Hypotheses**

see back page for Formulae Sheet

Recall the *t*-test?

Use: $t_0 = \dfrac{estimate - hypothesised\ value}{std\ error}$

where in this case the degrees of freedom are: $df = n - 2$

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

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

In SPSS, the output always comes out in the same way:

Regression

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	$\hat{\beta}_0$	$se(\hat{\beta}_0)$		t_0	***p-value***
	***X*-axis_Variable**	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	***r***	t_0	***p-value***

a Dependent Variable: ***Y*-axis_Variable**

Correlations

		***X*-axis_Variable**	***Y*-axis_Variable**
***X*-axis_Variable**	Pearson Correlation	1	***r***
	Sig. (2-tailed)		***p-value***
	N	***n***	***n***
***Y*-axis_Variable**	Pearson Correlation	***r***	1
	Sig. (2-tailed)	***p-value***	
	N	***n***	***n***

Recall: The P-value:

- In regression, we are carrying out *t*-tests, just like in Chapter 7.
- Therefore, the *P-value* is the conditional probability of observing a *t*-test statistic as extreme as that observed or more so, **given that** the null hypothesis is true.
- The ***P-value*** measures the **strength of evidence against** the null hypothesis, H_0.
- The **smaller** the *P-value*, the **stronger** the evidence against H_0.

- There are two ways of interpreting the *P-value*:

  1. As a description of the strength of evidence against H_0:

P-value	Evidence against H_0
> 0.10	None
≈ 0.10	Weak
≈ 0.05	Some
≈ 0.01	Strong
< 0.001	Very Strong

  2. As a description of the test result as (statistically) significant or nonsignificant.

     A test result is significant when the *P-value* is "small enough"; usually we say a *P-value* is small enough if it is less than 0.05 (5%):

     Testing at a 5% level of significance:

P-value	Test result	Action
< 0.05	Significant	Reject H_0 in favour of H_1
> 0.05	Nonsignificant	Do not reject H_0

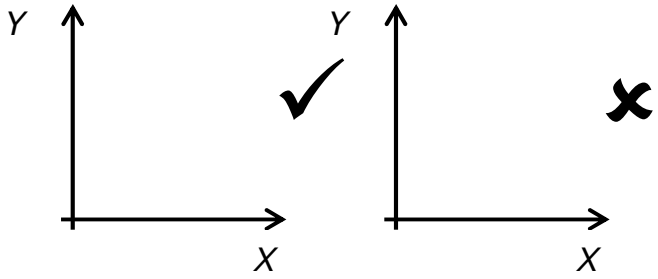
     Testing can be done at any level of significance; 1% is common but 5% is what most researchers use.

     The level of significance is an error rate; and can be thought of as a *false alarm* rate: i.e. it is the proportion of the time that a true null hypothesis will be rejected.

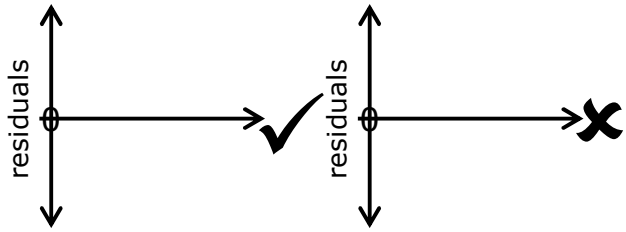
- **Assumptions** of simple linear regression are:
  1. There is a **linear** relationship between X and Y.
  2. Errors are all **independent**.
  3. Errors are **Normal**ly distributed (with $\mu = 0$).
  4. Errors all have the **same std deviation**, σ, regardless of the value of x.
- **Assumption checking using plots of the data and residual plots**

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

Scatterplot of data:

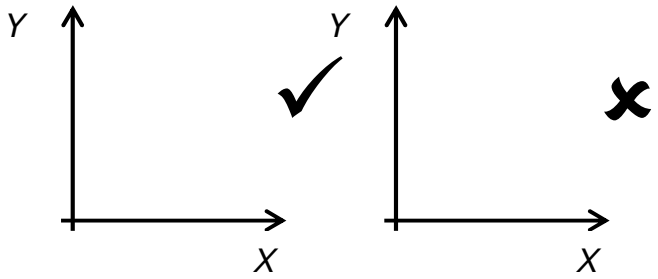


Residual plot:

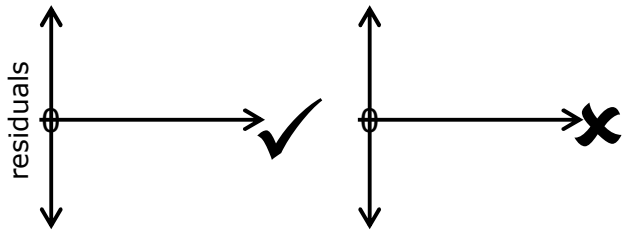


3. Errors are **Normal**ly distributed (with $\mu = 0$).

Scatterplot of data:

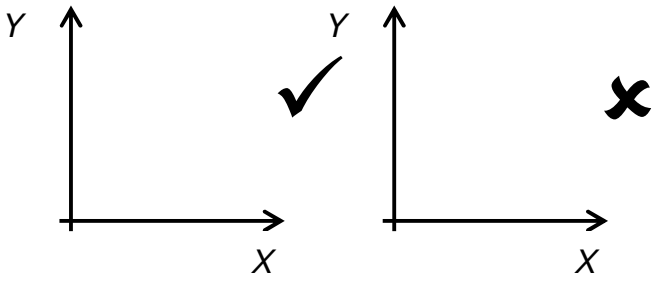


Residual plot:

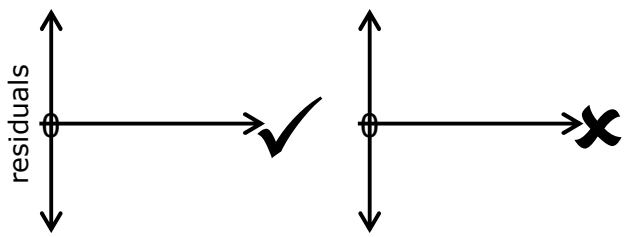


4. Errors all have the **same std deviation**, σ, regardless of the value of x.

Scatterplot of data:

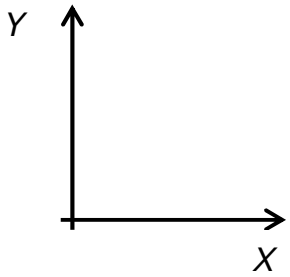


Residual plot:



- **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:
  - uncertainty about the values of β_0 and β_1.

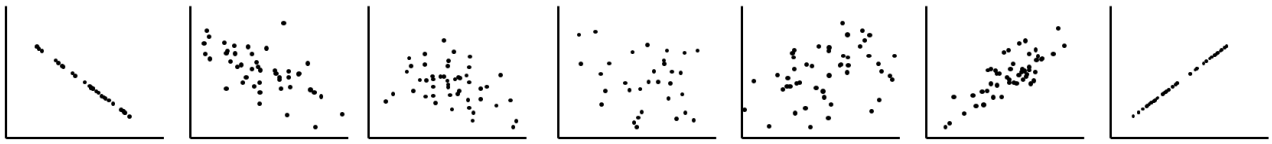
  $\boxed{estimate \pm t \times se(estimate)}$

  ✓ **Prediction Interval**
  This predicts the *Y*-value for **an individual** with a specified value of *x*.
  The width of the interval allows for:
  - 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*** has a value between -1 and +1:



r = -1 *r* = -0.7 *r* = -0.4 *r* = 0 *r* = 0.3 *r* = 0.8 *r* = 1

  - ***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*** = 1, then *X* and *Y* have a **perfect positive linear relationship**

  ✓ ***r*** measures the **strength** and **direction** of the **linear** association between **two numeric variables**

  ✓ ***r*** measures how close the points **come to lying on a straight line**

  ✓ 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*

  ✓ ***r*** has no units → a computer / calculator can give you the value of ***r***

  ✓ **Correlation DOES NOT imply causation**

Practice Questions

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

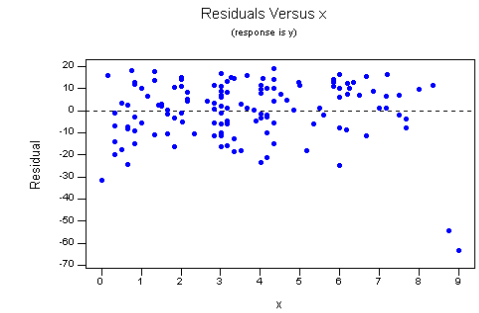


Figure A

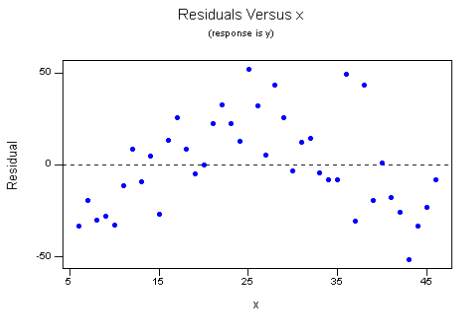


Figure B

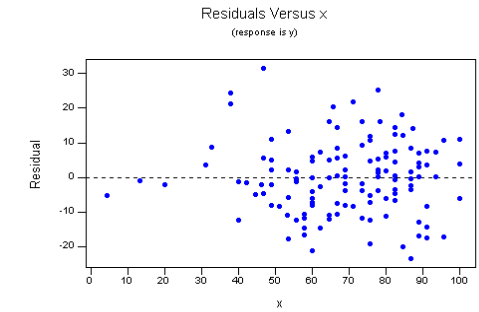


Figure C

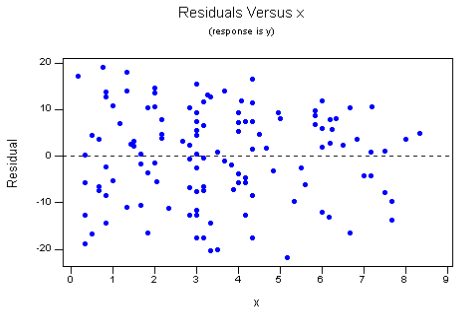


Figure D

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

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

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

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

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

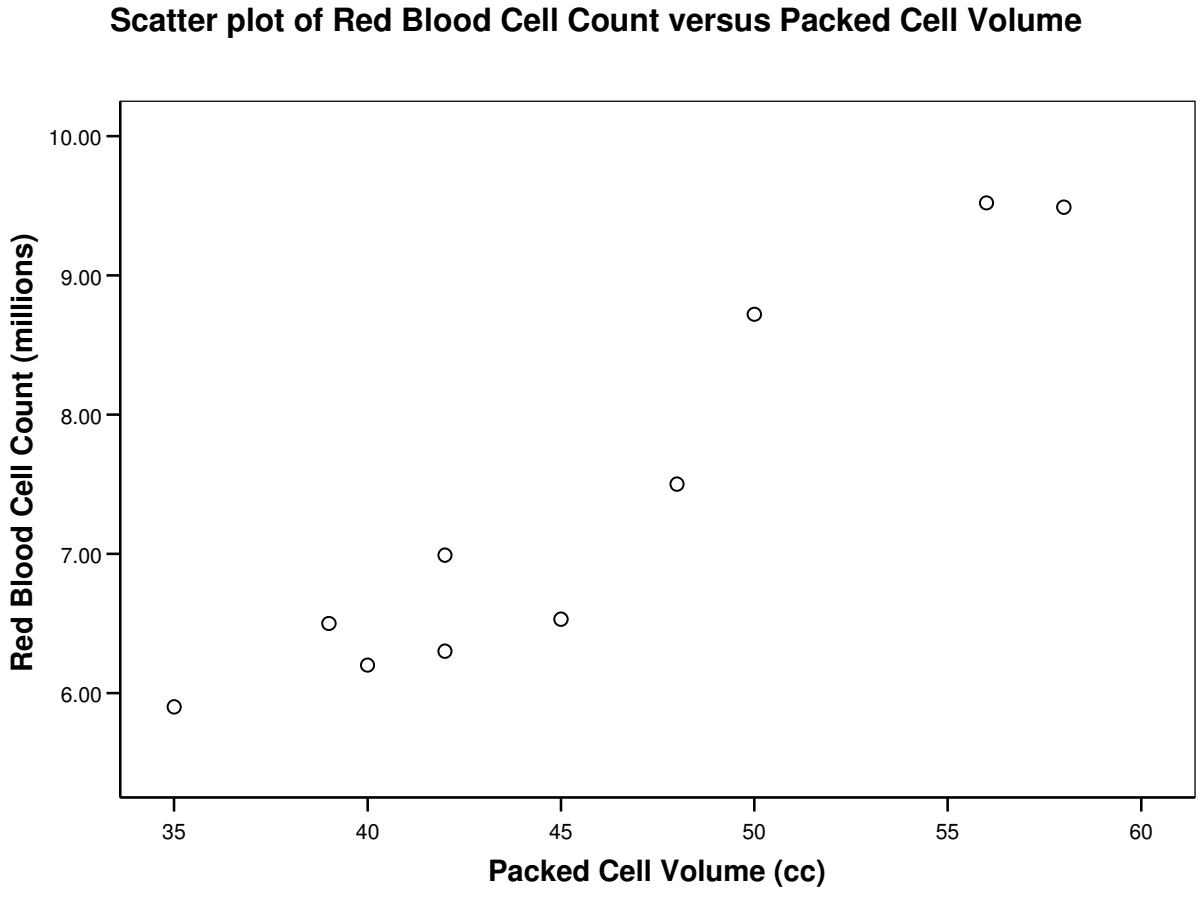
6. Which **one** of the following statements is **false**?
   (1) A relationship between two numeric variables may look weak because it has been plotted over only a limited range of *x*-values.
   (2) When exploring the relationship between two numeric variables, precise prediction cannot be made from a weak relationship.
   (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.
   (4) Cross-tabulation is a process of recording count data when we have two categorical variables.
   (5) In regression the explanatory variable is the variable explained by the response variable.

Questions 7 to **14 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:



Regression

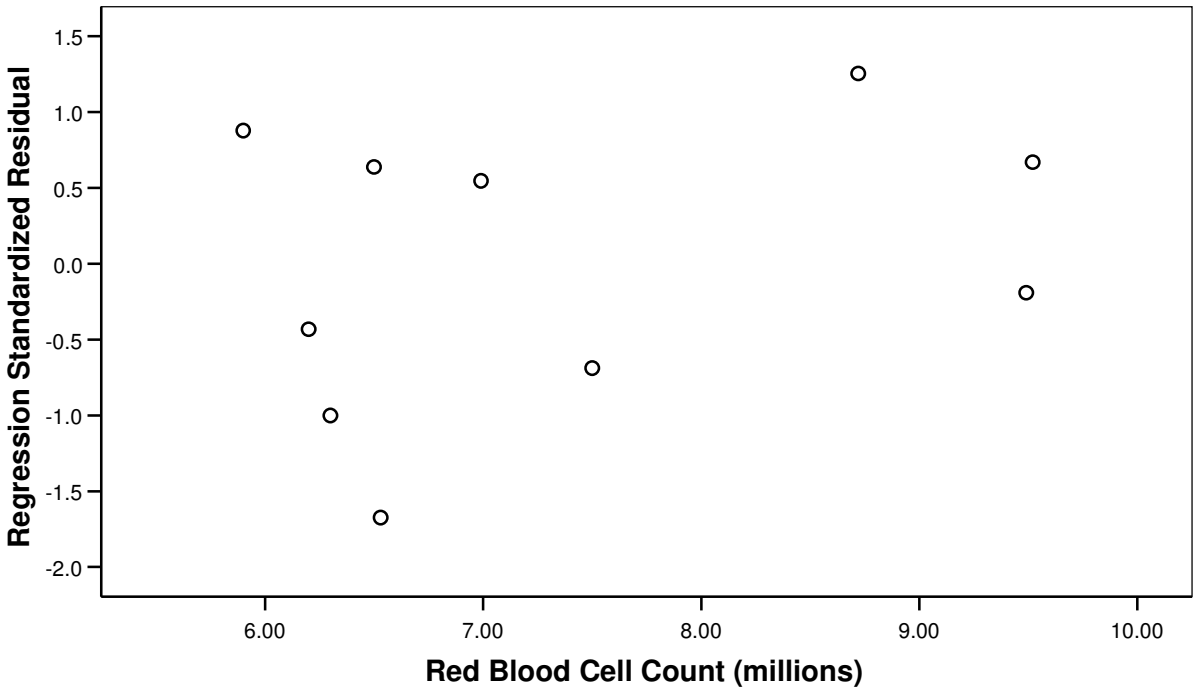
Coefficients(a)

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)

Scatterplot

Dependent Variable: Red Blood Cell Count (millions)



7. 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) $y = 0.177 - 0.68x$

8. For the data in the scatter plot on page 13, 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) $r = 0.13$

9. 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: $\hat{\beta}_0 = 1$ vs H_1: $\hat{\beta}_0 \neq 1$

   (2) H_0: $\beta_1 = 0$ vs H_1: $\beta_1 \neq 0$

   (3) H_0: $\beta_1 = 1$ vs H_1: $\beta_1 \neq 1$

   (4) H_0: $\beta_0 = 0$ vs H_1: $\beta_0 \neq 0$

   (5) H_0: $\hat{\beta}_1 = 0$ vs H_1: $\hat{\beta}_1 \neq 0$

10. 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.

11. 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

12. The packed cell volume and red blood cell 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 (3) 0.81 (5) −0.81

(2) 0.55 (4) −0.55

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

(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 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.

14. 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 blood cell 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 blood cell 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 blood cell 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 blood cell 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 blood cell count of between 1.31 and 2.23 million.

Questions 15 to **20** refer to the following information.

The researcher believes that the engine size of cars with small to moderate sized engines (under 2500cc) could be used to predict the weight of a car. The results of an SPSS linear regression analysis of 43 cars with engine sizes under 2500cc and associated plots are shown in Figure 9, Table 14 and Figure 10 (all given below).

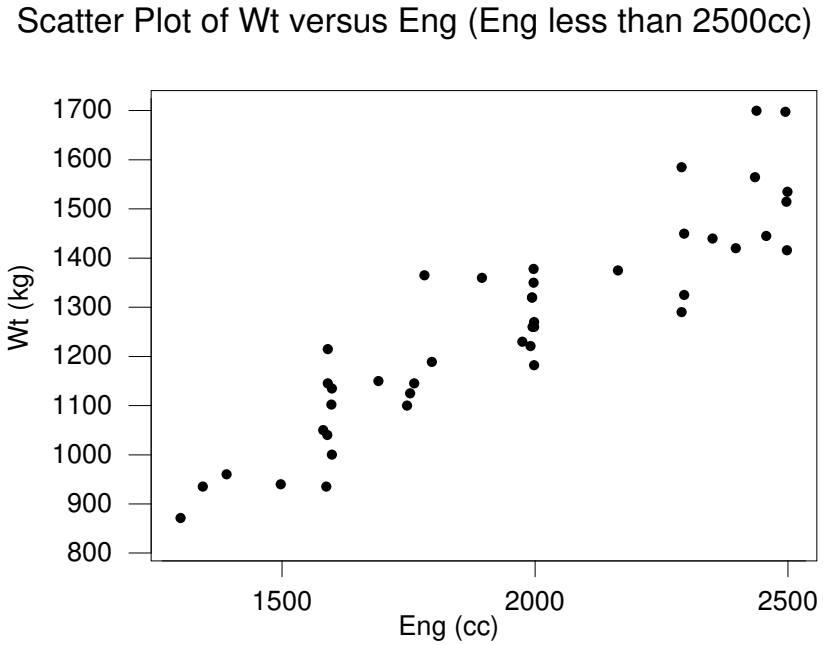


Figure 9: Scatter plot of weight versus engine size for cars with engines smaller than 2500cc

Regression

Coefficients(a)

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

a Dependent Variable: Wt (kg)

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

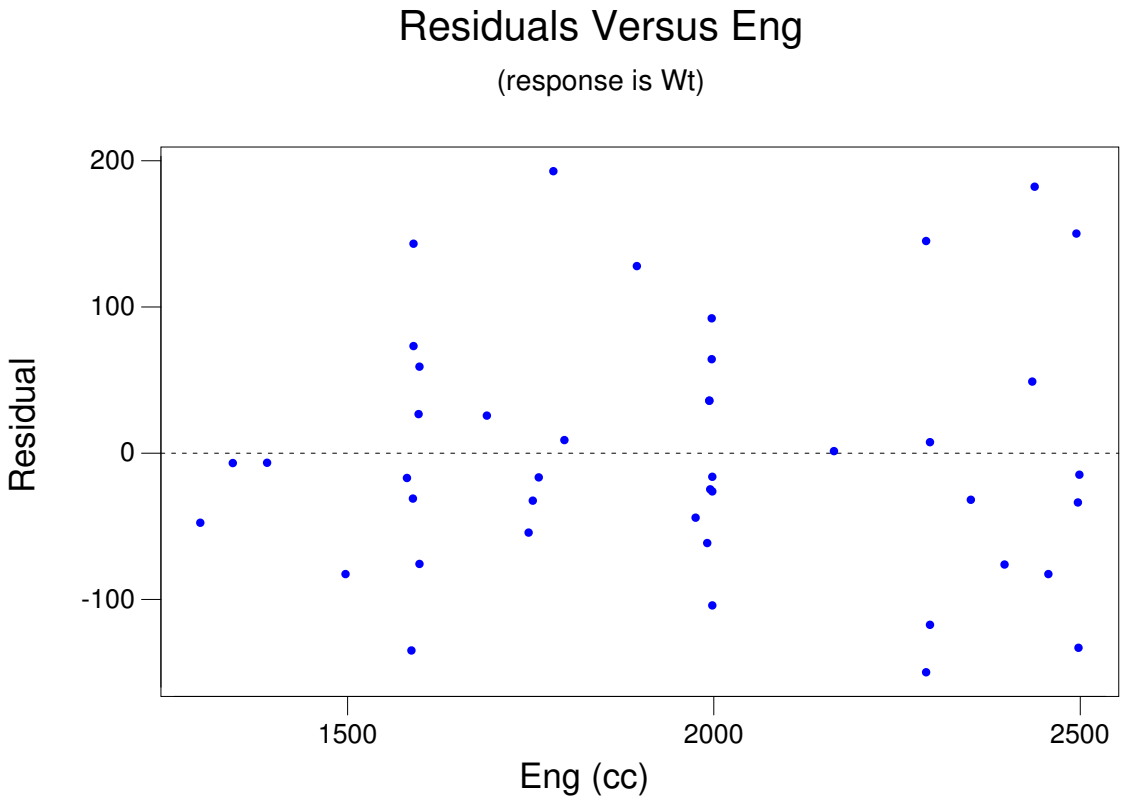


Figure 10: Scatter plot of residuals versus engine size for cars with engines smaller than 2500cc

15. 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

16. 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

17. 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

18. 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: $\beta_0 \neq 0$

19. 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 numeric 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.

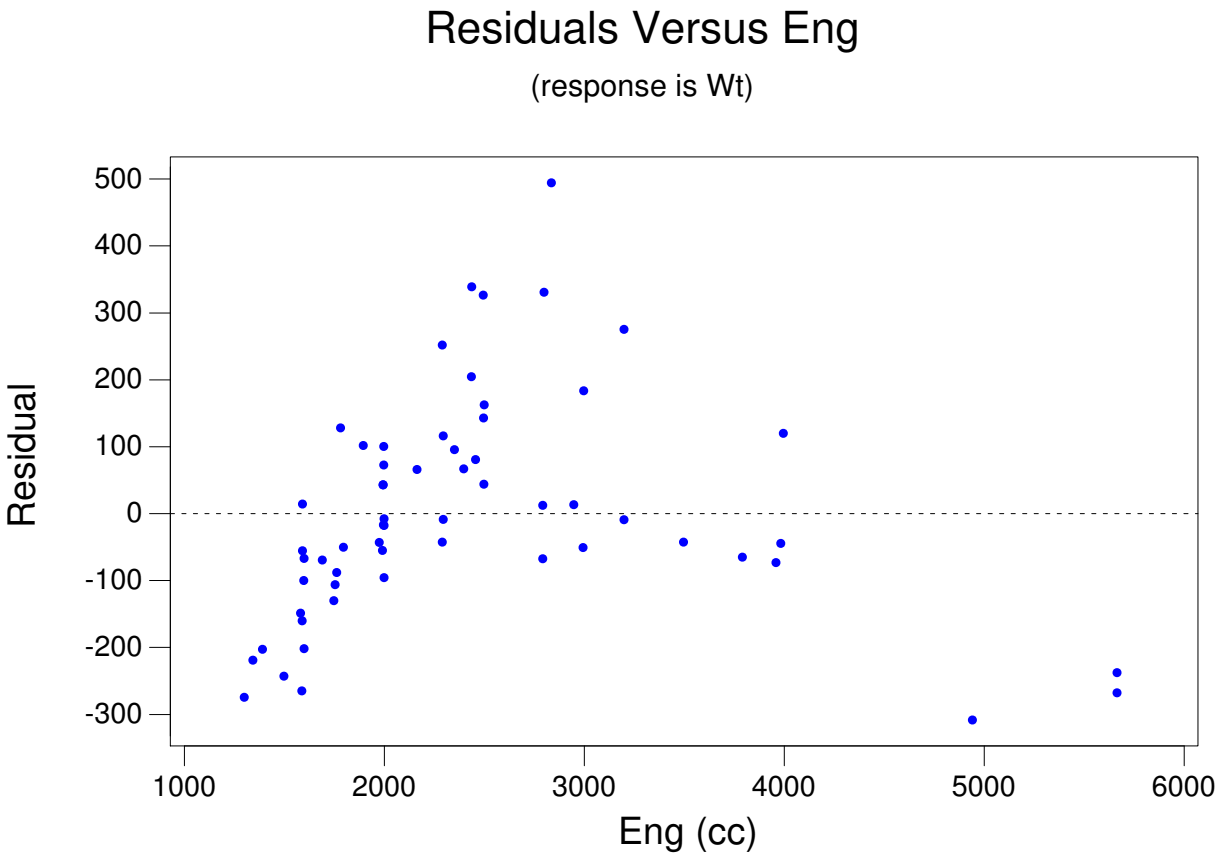


Figure 11: Scatter plot of residuals versus engine size for all cars

Questions 21 and **22** 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.

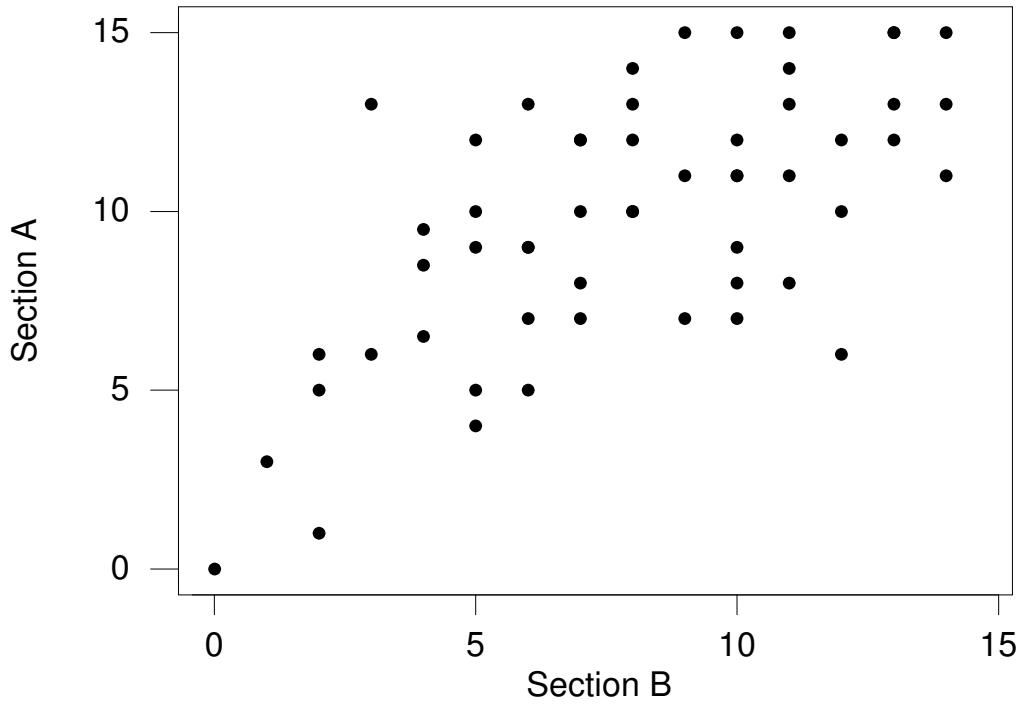


Figure 3: Scatter plot of marks in course test

21. 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.

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

23. 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.

    (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.

    (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.

    (5) For a specified value of x, the 95% confidence interval for the mean is wider than the associated 95% prediction interval.

24. Suppose we wish to test for no linear relationship between heart rate and temperature. We find the *P-value* is less than 1%. We can **correctly conclude** that the *P-value*:

    (1) indicates strong evidence against the null hypothesis, therefore the data contains strong evidence that a perfect linear relationship exists between heart rate and temperature.

    (2) indicates strong evidence against the null hypothesis, therefore the data contains strong evidence that a linear relationship exists between heart rate and temperature.

    (3) indicates strong evidence against the null hypothesis. However, this tells us nothing about whether or not a linear relationship exists between heart rate and temperature.

    (4) indicates strong evidence against the null hypothesis, therefore the data contains strong evidence that a causal linear relationship exists between heart rate and temperature.

    (5) is very small, therefore the data contain no evidence that a linear relationship exists between heart rate and temperature.

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

    (1) A single outlier can have a large influence on the value of the sample correlation coefficient.

    (2) For a least squares regression line, if you add up all the residuals then the total is zero.

    (3) The *Y*-variable is called the independent or explanatory variable and *X*-variable is called the dependent or response variable.

    (4) For a simple linear regression, the (average) pattern seen in the scatter plot must be a straight line.

    (5) The two important components of a regression are the average pattern (trend) and the deviation of the observations from that pattern (scatter about the trend).

26. Which **one** of the following statements is **not** a reason for fitting a linear regression model to the data?

    (1) To estimate parameters in a theoretical model.

    (2) To make predictions.

    (3) To understand a relationship better.

    (4) To find the trend line.

    (5) To conclusively establish the cause of an effect.

27. Consider the point labelled "A" in Figure 4.

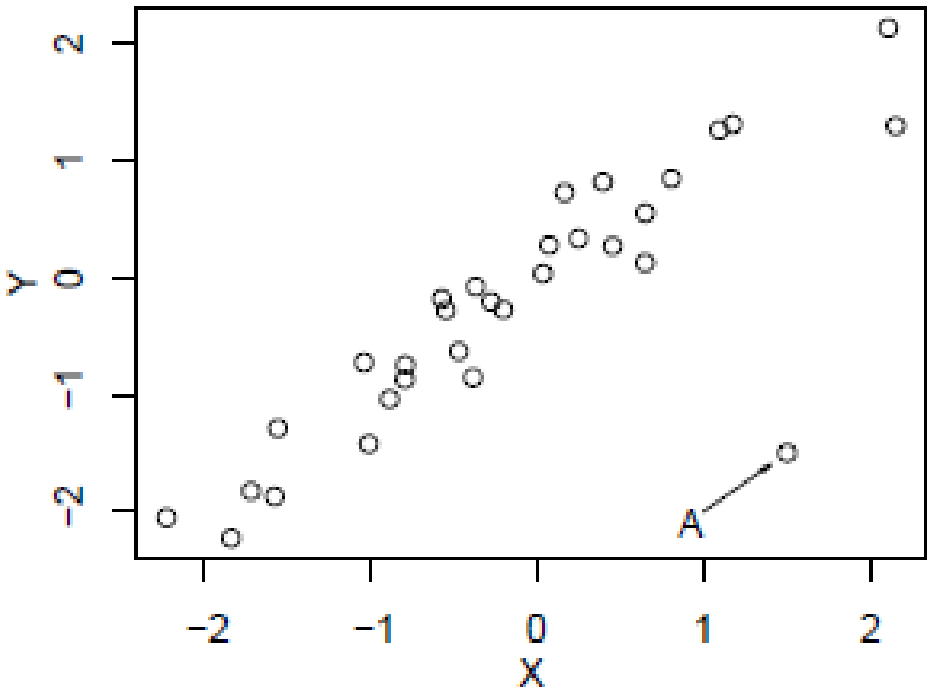


Figure 4: A scatter plot.

Which **one** of the following statements about point "A" is **true**?

(1) The point is an outlier in *X*.

(2) The point is an outlier in *Y*.

(3) It is impossible to tell whether the point is an outlier without looking at a plot of the residuals.

(4) The point should be removed from the analysis.

(5) The point is an outlier because it lies much further from the linear trend than the other points.

28. Which **one** of the following assumptions of the simple linear model **cannot** be checked in a residual plot?

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

(2) The random errors are Normally distributed.

(3) The random errors have the same standard deviation regardless of the value of *x*.

(4) There is a linear relationship between *x* and E(*Y*).

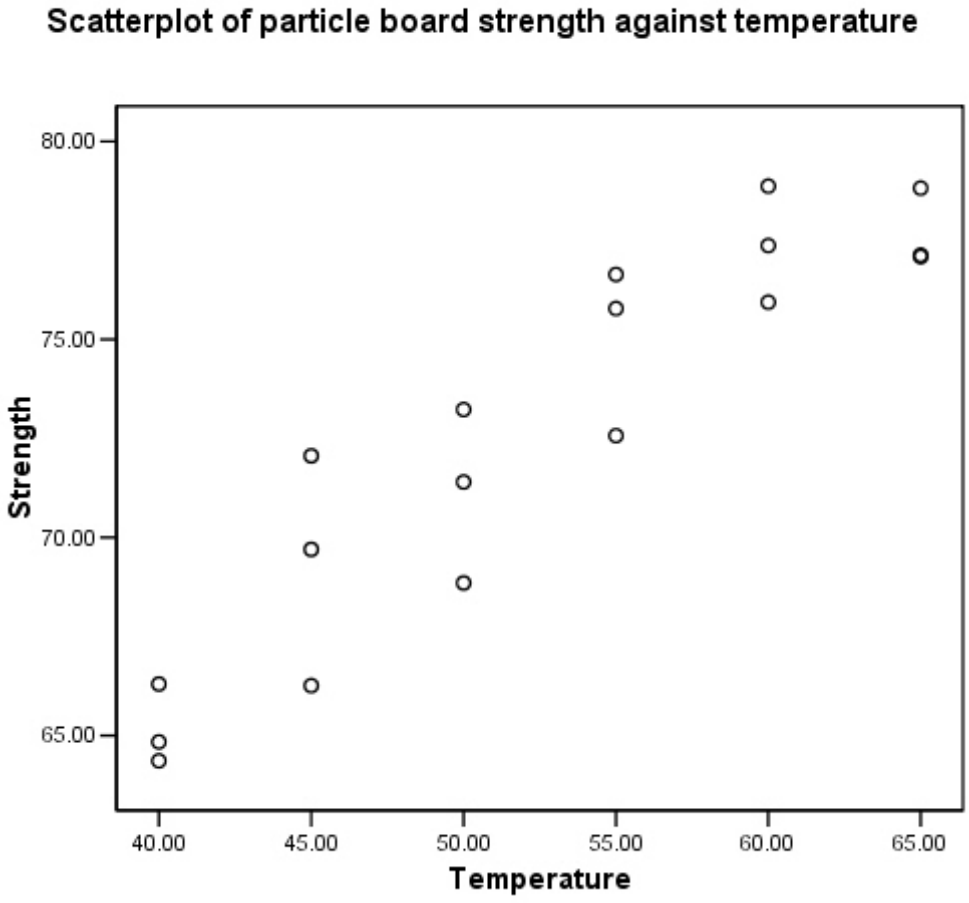
(5) The observations are independent.

Questions 29 to **37 refer to the following information.**

The production of particle boards involves a step in which the boards are baked. A manufacturer of particle boards investigated the effect of the baking temperature (*X*) on the strength of particle boards (*Y*). A total of 18 particle boards were baked using 6 different temperatures (3 boards were baked at each temperature) and the strength of these boards was measured. All aspects of the process other than the baking temperature were kept as similar as possible. The assignment of temperatures to boards and the order of production were determined using random processes. The data follows:

	Strength	Temperature (°*C*)		Strength	Temperature (°*C*)
1	66.30	40	10	75.78	55
2	64.84	40	11	72.57	55
3	64.36	40	12	76.64	55
4	69.70	45	13	78.87	60
5	66.26	45	14	77.37	60
6	72.06	45	15	75.94	60
7	73.23	50	16	78.82	65
8	71.40	50	17	77.13	65
9	68.85	50	18	77.09	65

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



Regression

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	45.454	2.846		15.972	.000
	Temperature	.518	.054	.924	9.672	.000

a Dependent Variable: Strength

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

(1) $y = 45.454 + 0.518x$

(2) $\hat{y} = 45.454 + 0.518$

(3) $\hat{y} = 45.454 + 0.518x$

(4) $\hat{y} = 0.518 + 45.454x$

(5) $y = 0.518 + 45.454x$

30. For the data in the scatter plot on page 25, which one of the following values would be the closest to the sample correlation coefficient, r?

(1) $r = 0.9$

(2) $r = 0.7$

(3) $r = 0.3$

(4) $r = 0.5$

(5) $r = 0.1$

31. The correct null and alternative hypotheses to test that there is no linear relationship between particle board strength and baking temperature 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 $H_1 : \hat{\beta}_1 \neq 0$

32. To test the hypotheses from the previous question, the test statistic would be:

(1) $t_0 = \dfrac{72.62}{2.846}$

(2) $t_0 = \dfrac{0.518}{2.846}$

(3) $t_0 = \dfrac{45.454}{2.846}$

(4) $t_0 = \dfrac{0.518}{0.054}$

(5) $t_0 = \dfrac{45.454}{0.054}$

33. When the hypothesis test referred to in Questions 31 and 32 is conducted, it is found that there is very strong evidence against the hypothesis of no linear relationship between baking temperature and particle board strength. Which **one** of the following statements is **true** for this investigation?

(1) The results of this hypothesis test can be taken as evidence that changes in baking temperature *cause* changes in particle board strength since very strong evidence of a relationship always implies causation.

(2) The results of this hypothesis test cannot be taken as evidence that changes in baking temperature *cause* changes in particle board strength since there may be factors other than baking temperature which affect the strength of particle boards.

(3) The results of this hypothesis test cannot be taken as evidence that changes in baking temperature *cause* changes in particle board strength since strong evidence of a relationship does not necessarily mean the relationship is causal.

(4) The results of this hypothesis test cannot be taken as evidence that changes in baking temperature *cause* changes in particle board strength since the scatter plot shows that for each baking temperature there is a substantial amount of variability in the strength of particle boards.

(5) The results of this hypothesis test can be taken as evidence that changes in baking temperature *cause* changes in particle board strength since a random process was used to assign boards to temperatures.

34. The fitted least squares regression line can be used to predict the strength of particle board. Boards baked at a temperature of 55^oC have a predicted strength of approximately:

(1) 102.4

(2) 73.9

(3) 13.8

(4) 79.9

(5) 47.0

35. The baking temperature and particle board strength for sample number 11 was 55^oC and 72.57 units respectively. Under the fitted least squares line, the value of the residual for this sample is approximately:

    (1) 2.30

    (2) 0.65

    (3) 1.33

    (4) −0.65

    (5) −1.33

36. The fitted least squares regression line indicates that for each increase of 2.5^oC in baking temperature we expect that, on average, the particle board strength will:

    (1) increase by approximately 45.45 units.

    (2) decrease by approximately 0.52 units

    (3) increase by approximately 0.52 units.

    (4) increase by approximately 1.30 units.

    (5) decrease by approximately 1.30 units.

37. Why is the prediction interval for the strength of particle board baked at a temperature of 55^oC more useful than the corresponding point estimate given in Question 34?

    (1) Because the prediction interval is **only one** of the plausible values of that the strength could be when the baking temperature is 55^oC.

    (2) The prediction interval is more useful than the point estimate as it estimates (with a certain level of confidence) a range of plausible values the strength could be when the baking temperature is 55^oC.

    (3) Because the prediction interval is smaller than a corresponding confidence interval and therefore can capture the true value more accurately.

    (4) The prediction interval is more useful than the point estimate as it always captures the true estimate (with a certain level of confidence) in a range of plausible values the strength could be when the baking temperature is 55^oC.

    (5) The prediction interval is not as useful as a corresponding confidence interval.

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

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

    (2) The random errors are all independent.

    (3) The random errors follow a linear trend.

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

    (5) The random errors are Normally distributed.

39. Which **one** of the following statements about correlation is **false**?

    (1) Correlation can be used only if both variables are numeric.

    (2) A scatter plot of the data should be examined before looking at correlation.

    (3) Outliers can deflate the value of the sample correlation coefficient *r*.

    (4) Correlation should not be used when, in a scatter plot, the relationship between two variables appears non-linear.

    (5) Correlation can be used as proof of a causal relationship regardless of how the data were collected.

40. Which **one** of the following statements about a 95% prediction interval and the corresponding 95% confidence interval for the mean is **true**?

    The prediction interval:

    (1) can be either narrower or wider than the confidence interval for the mean, depending on the estimated variability of the values for the slope and intercept of the line.

    (2) is always narrower than the confidence interval for the mean because it only takes into account the uncertainty about the values of the slope and intercept of the line and not the random scatter about the line.

    (3) is always wider than the confidence interval for the mean because as well as taking into account the uncertainty about the values of the slope and intercept of the line, it also takes into account the uncertainty due to the random scatter about the line.

    (4) is always narrower than the confidence interval for the mean because it only takes into account the uncertainty about the value of the slope of the line and not the value of the intercept of the line.

    (5) is always wider than the confidence interval for the mean because as well as taking into account the uncertainty about the value of the slope of the line, it also takes into account the uncertainty about the value of the intercept of the line.

41. 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.

42. 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.

43. 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 $X = x$.

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

(1) $\hat{\beta}_0$ and $\hat{\beta}_1$ are sample estimates of the parameters β_0 and β_1.

(2) The regression model consists of a linear trend plus random scatter.

(3) A correlation coefficient, which is close to 0, indicates no linear relationship between *X* and *Y*.

(4) The least squares estimates minimise the square of the difference between $\hat{\beta}_0$ and $\hat{\beta}_1$.

(5) The errors from a regression analysis are assumed to be Normally distributed with a mean of zero.

45. Which **one** of the following statements about simple linear regression is **false**?

(1) The least squares regression line always passes through the point which is the mean of the values for the *X*-variable and the mean of the values for the *Y*-variable.

(2) Residuals illustrate the scatter about a fitted line.

(3) When $x = 0$, the *y*-value of the point on the least squares regression line estimates the *y*-intercept of the true line in the underlying population.

(4) The fitted line with the smallest total when all of the residuals are squared and then added up is called the least squares regression line.

(5) The residual for an observation is the observed *x*-value minus the predicted *x*-value.

ANSWERS

1. **(2)**	2. **(3)**	3. **(4)**	4. **(1)**	5. **(4)**	6. **(5)**
7. **(3)**	8. **(3)**	9. **(2)**	10. **(4)**	11. **(4)**	12. **(2)**
13. **(3)**	14. **(2)**	15. **(5)**	16. **(1)**	17. **(3)**	18. **(4)**
19. **(2)**	20. **(3)**	21. **(3)**	22. **(4)**	23. **(5)**	24. **(2)**
25. **(3)**	26. **(5)**	27. **(5)**	28. **(5)**	29. **(3)**	30. **(1)**
31. **(2)**	32. **(4)**	33. **(5)**	34. **(2)**	35. **(5)**	36. **(4)**
37. **(2)**	38. **(3)**	39. **(5)**	40. **(3)**	41. **(3)**	42. **(3)**
43. **(4)**	44. **(4)**	45. **(5)**			

FORMULAE

Confidence intervals and *t*-tests

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

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

Applications:

1. Single mean $\boldsymbol{\mu}$: $estimate = \overline{x}$; $df = n - 1$
2. Single proportion $\boldsymbol{p}$: $estimate = \widehat{p}$; $df = \infty$
3. Difference between two means $\boldsymbol{\mu_1 - \mu_2}$: (independent samples)
   $estimate = \overline{x}_1 - \overline{x}_2$; $df = \min(n_1 - 1, n_2 - 1)$
4. Difference between two proportions $\boldsymbol{p_1 - p_2}$:
   $estimate = \widehat{p}_1 - \widehat{p}_2$; $df = \infty$
   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: $f_0 = \dfrac{s_B^2}{s_W^2}$; $df_1 = k - 1,\ df_2 = n_{\text{tot}} - k$

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: $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$

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