

Stats 101/101G/108 Workshop

Regression and Correlation [RC]

2020

by Leila Boyle



Stats 101/101G/108 Workshops

The Statistics Department offers workshops and one-to-one/small group assistance for Stats 101/101G/108 students wanting to improve their statistics skills and understanding of core concepts and topics.

Leila's website for Stats 101/101G/108 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 303S.288 (second floor of the Science Centre, Building 303S)
l.boyle@auckland.ac.nz; (09) 923-9045; 021 447-018

Want help with Stats?

Stats 101/101G/108 appointments

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

Stats 101/101G/108 Workshops

One computing workshop, four exam prep workshops and four drop-in sessions are held during the second half of the semester.

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!

No booking is required – just turn up to any workshop! You are also welcome to come along virtually on Zoom if you prefer. Search your emails for “Leila” to find the link – email Leila at l.boyle@auckland.ac.nz if you can’t find it.

- **Computer workshop: Hypothesis Tests in SPSS**

www.tinyURL.com/stats-HTS

Computing for Assignment 3 – covers the **computing** you need to do for **Questions 3 and 4** (iNZight plots & SPSS output). There are **six identical** sessions:

- Friday 16 October, 3-4pm
- Monday 19 October, 10-11am
- Monday 19 October, 2-3pm
- Tuesday 20 October, 4-5pm
- Wednesday 21 October, 11am-midday
- Wednesday 21 October, 3-4pm

- **Exam prep workshops**

- **Chi-Square Tests** www.tinyURL.com/stats-CST

Exam revision for Chapter 9 – Saturday 24 October, 1-4pm, LibB15 (useful exam prep and also useful for the **Chapter 9 Quiz** due at 11pm on Wednesday 28 October!)

- **Regression and Correlation** www.tinyURL.com/stats-RC

Exam revision for Chapter 10 – Saturday 31 October, 9.30am-12.30pm, LibB10 (useful exam prep and also useful for the **Chapter 10 Quiz** due at 11pm on Wednesday 4 November!)

- **Hypothesis Tests: Proportions** www.tinyURL.com/stats-HTP

Exam revision for Chapters 6 & 7 (with a focus on proportions) – Tuesday 3 November, 9.30am-12.30pm, LibB10 (useful exam prep)

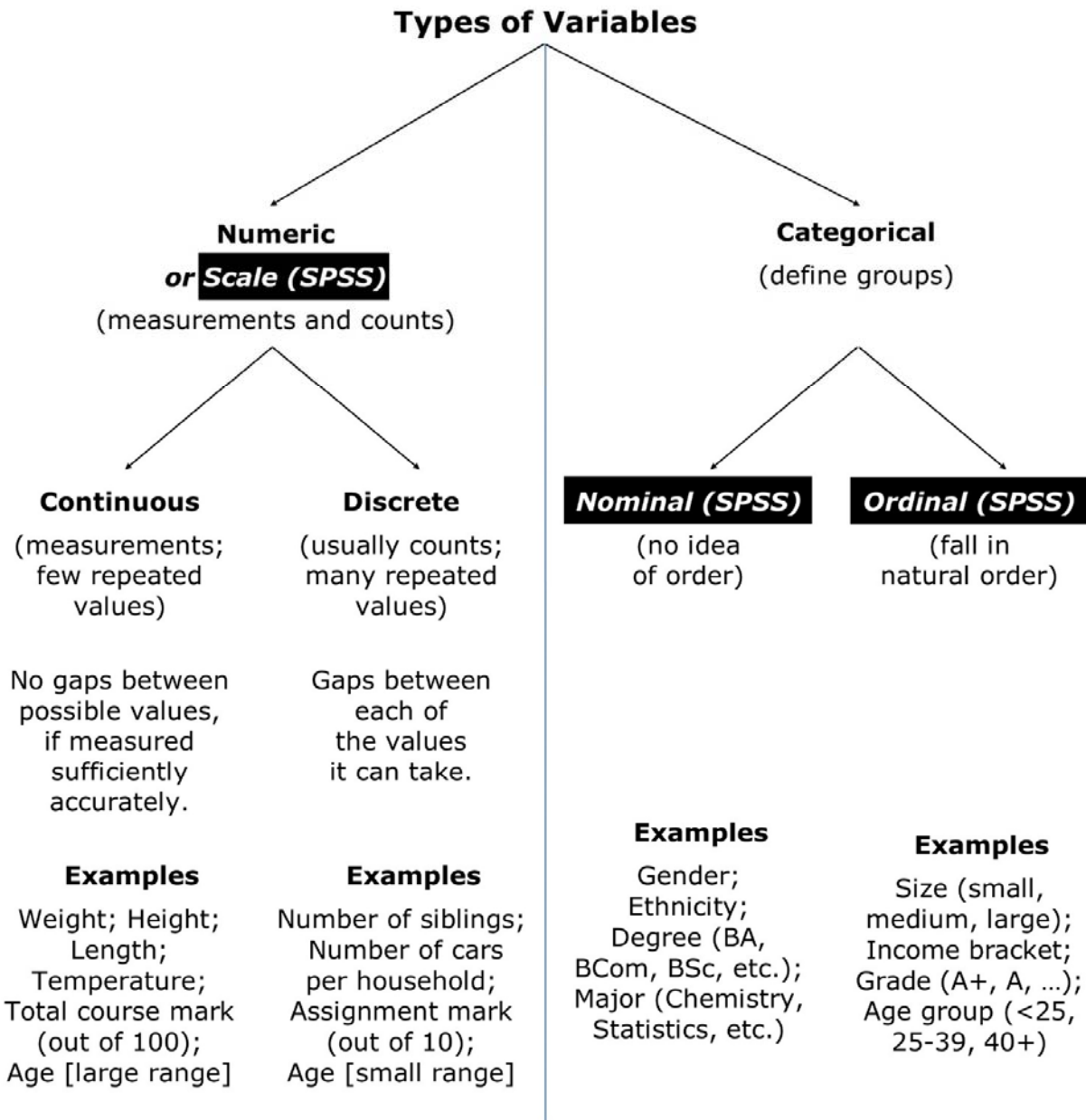
- **Hypothesis Tests: Means** www.tinyURL.com/stats-HTM


Exam revision for Chapter 6, 7 & 8 – Tuesday 3 November, 1-4pm, LibB10 (useful exam prep)

- **Drop-in sessions**

- Saturday 17 October, 9.30am-4pm, LibB10
- Saturday 24 October, 9.30am-12.30pm, LibB15
- Monday 26 October, 9.30am-4pm, LibB10
- Saturday 31 October, 1-4pm, LibB10

Regression and Correlation

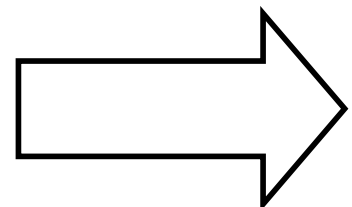


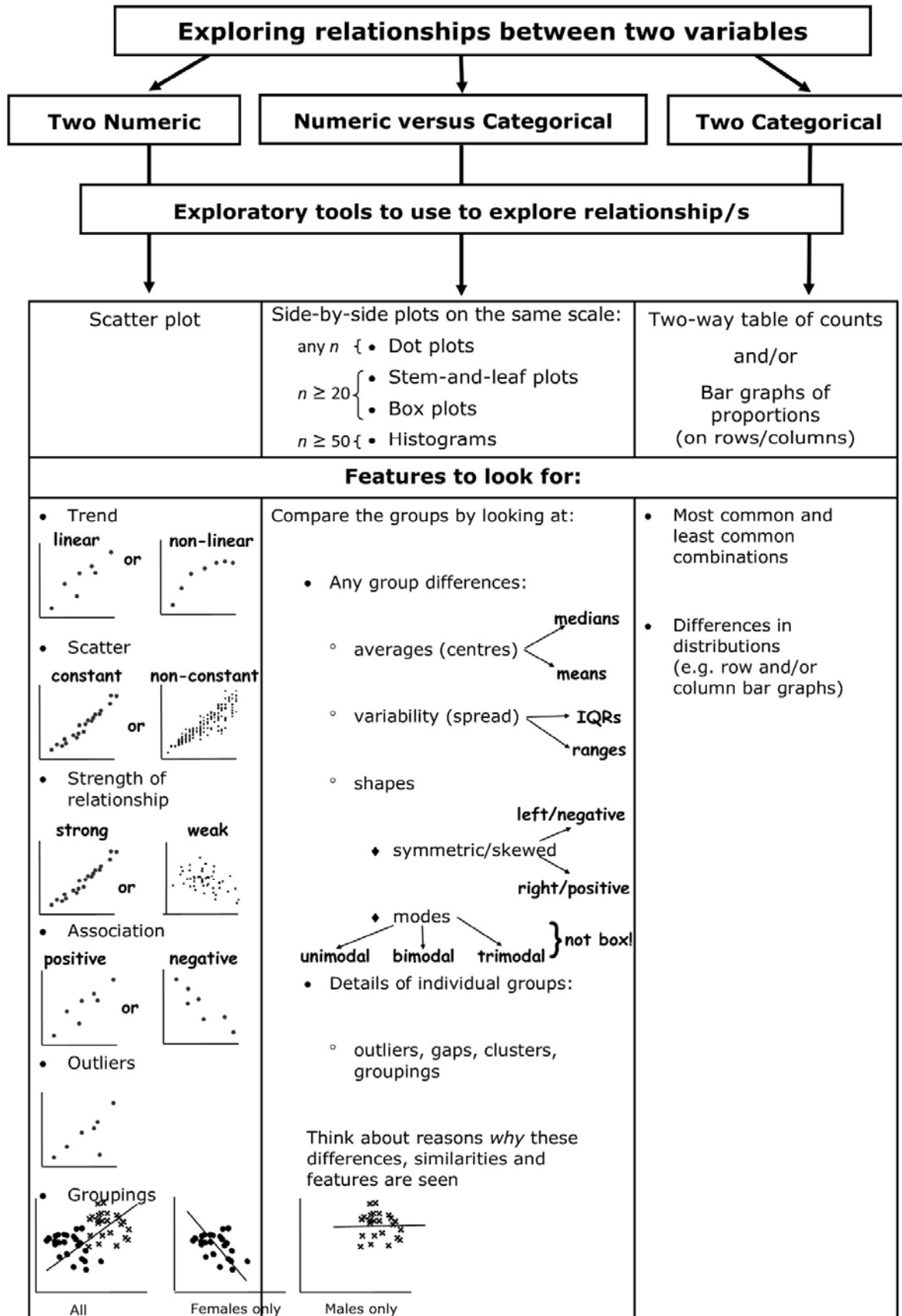
 **Useful reference:** Chance Encounters, pages 40 – 42

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

What to look for in a scatter plot:

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





Regression

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

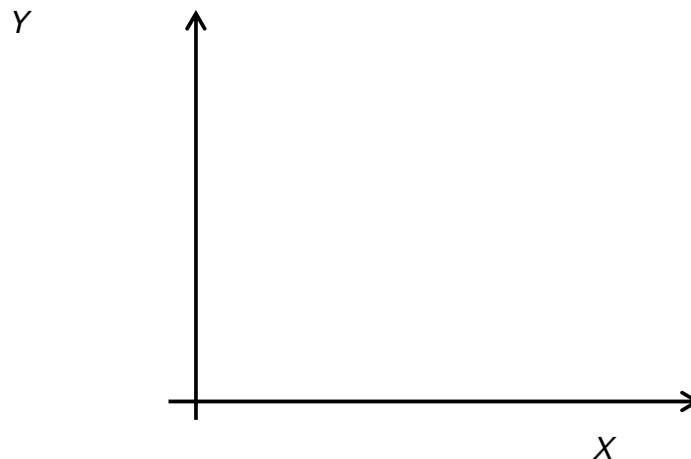
- X is used to **explain** or **predict** the behaviour of Y
- X is the **explanatory** or **independent** variable
- Y is the **dependent** or **response** variable

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 .



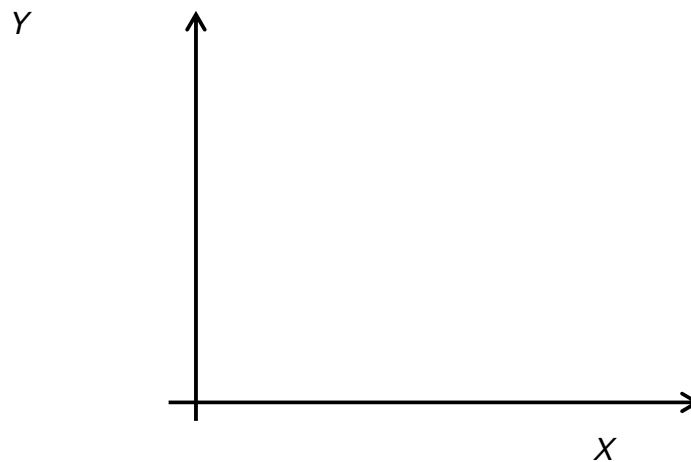
The single **least squares regression line** for each linear regression model:

- minimises the sum of the squared residuals/prediction errors
- has $\sum \text{residuals} = 0$ (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**

Recall the **t**-test?

**see back page
for Formulae Sheet**

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

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	<i>p-value</i>
X-axis_Variable	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	<i>r</i>	t_0	<i>p-value</i>

a Dependent Variable: **Y-axis_Variable**

Recall: The P-value:

- In regression, we are carrying out t -tests, just like in Chapter 7 and 8!
- Therefore, the **P-value**:
 - ✓ is the **conditional** probability of observing a test statistic as extreme as that observed or more so, **given that** the null hypothesis, H_0 , is true.
 - ✓ is the probability that sampling variation would produce an estimate that is at least as far from the hypothesised value than the estimate we obtained from our data, **assuming that** the null hypothesis, H_0 , is true.
 - ✓ measures the strength of evidence **against** H_0 .

- We interpret the P -value as a **description** of the **strength of evidence against the null hypothesis, H_0** . The **smaller** the P -value, the **stronger** the evidence against H_0 :

P -value	Evidence against H_0
> 0.10	None
≈ 0.07	Weak
≈ 0.05	Some
≈ 0.01	Strong
≤ 0.001	Very Strong

- An alternative approach often found in research articles and news items is to describe the test result as (statistically) significant or not significant. A test result is said to be significant when the P -value is "small enough"; usually people 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.

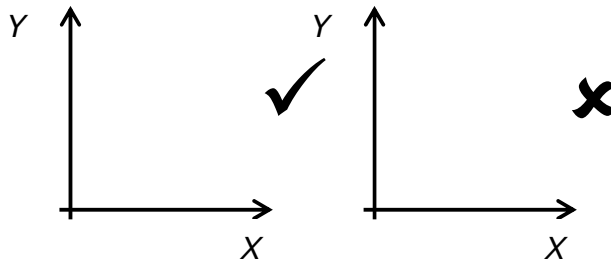
The level of significance can be thought of as a false alarm error rate, i.e. it is the proportion of times that the null hypothesis will be rejected when it is actually true (which can result in action being taken when really no action should be taken).

Thus, a statistically significant result means that a study has produced a "small" P -value (usually $< 5\%$).

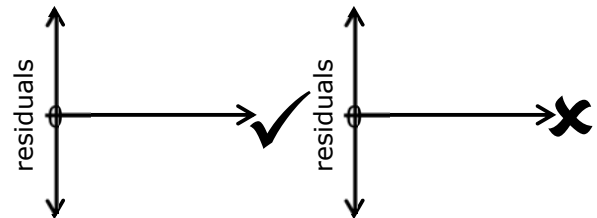
- **Assumptions** of simple linear regression are:
 1. There is a **linear** relationship between X and Y .
 2. Errors are all **independent**.
 3. Errors are **Normally** distributed (with $\mu = 0$).
 4. Errors all have the **same std deviation**, σ , regardless of the value of x .
- **Assumption checking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

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

Scatterplot of data:

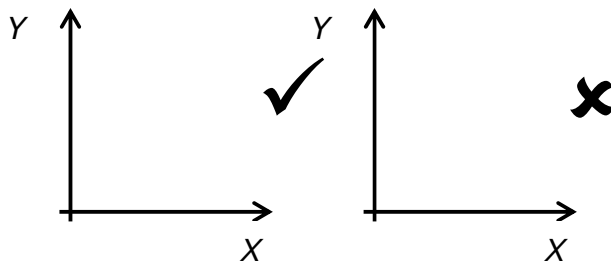


Residual plot:

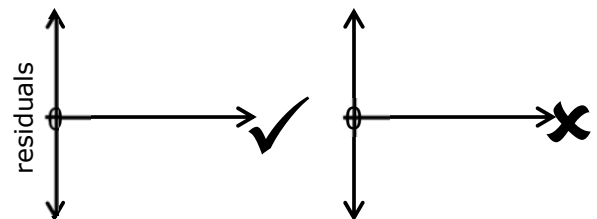


3. Errors are **Normally** distributed (with $\mu = 0$).

Scatterplot of data:

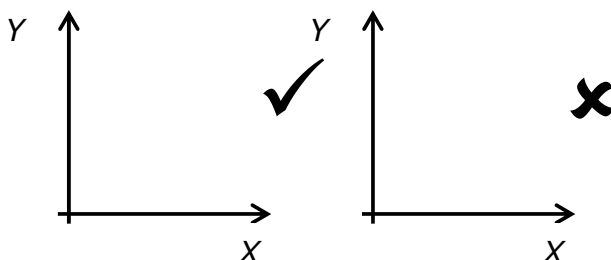


Residual plot:

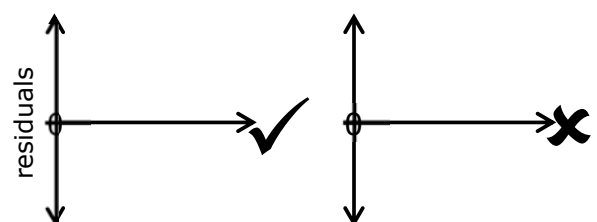


4. Errors all have the **same std deviation**, σ , regardless of the value of x .

Scatterplot of data:

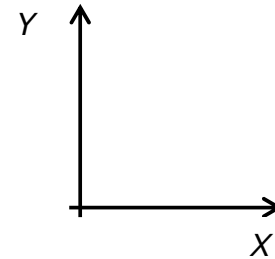


Residual plot:



• **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 .

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

- ✓ **Prediction Interval**

This predicts the Y -value for an **individual** with a specified value of x .

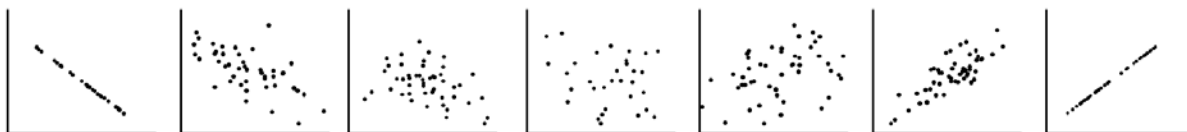
The width of the interval allows for:

- 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:

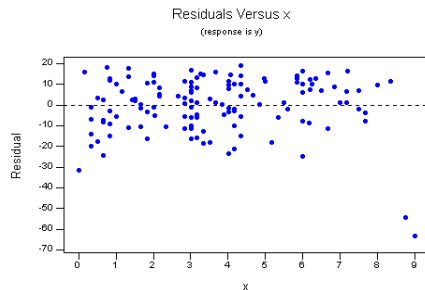


Figure A

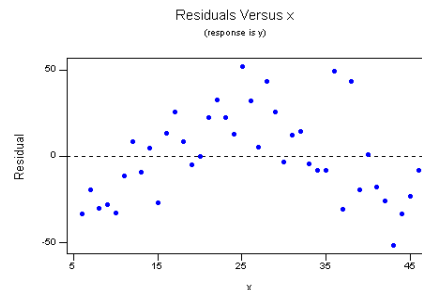


Figure B

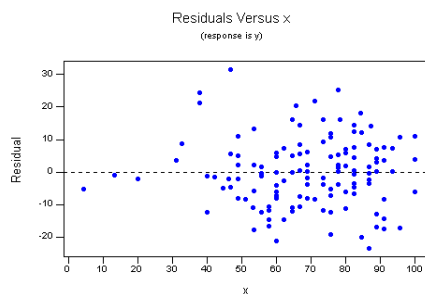


Figure C

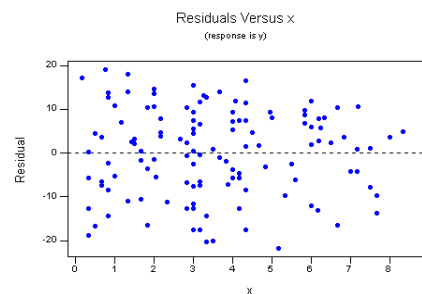


Figure D

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 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.

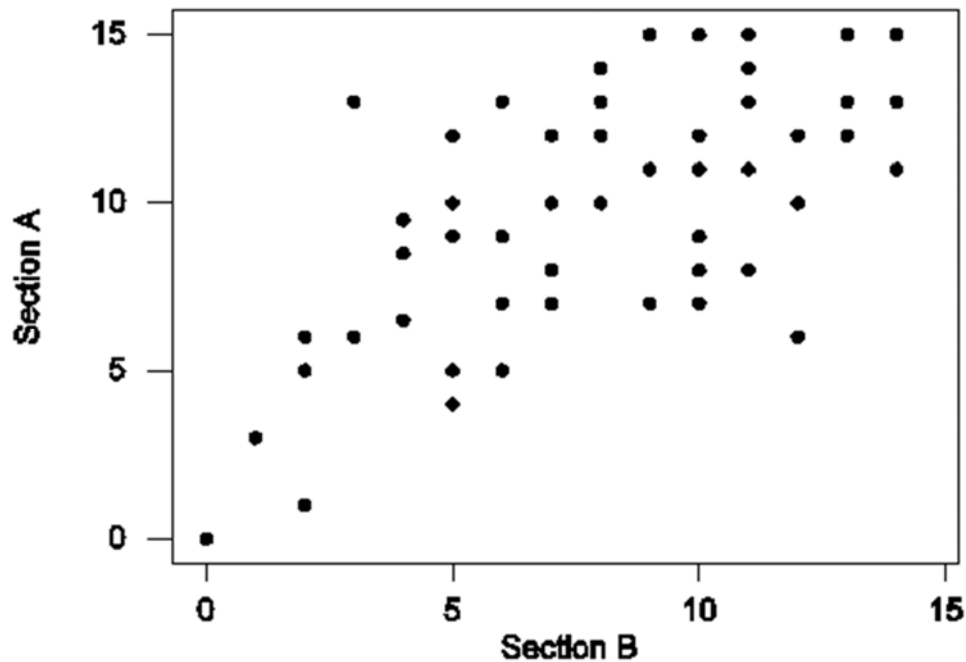


Figure 3: Scatter plot of marks in course test

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

7. 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.

9. 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.

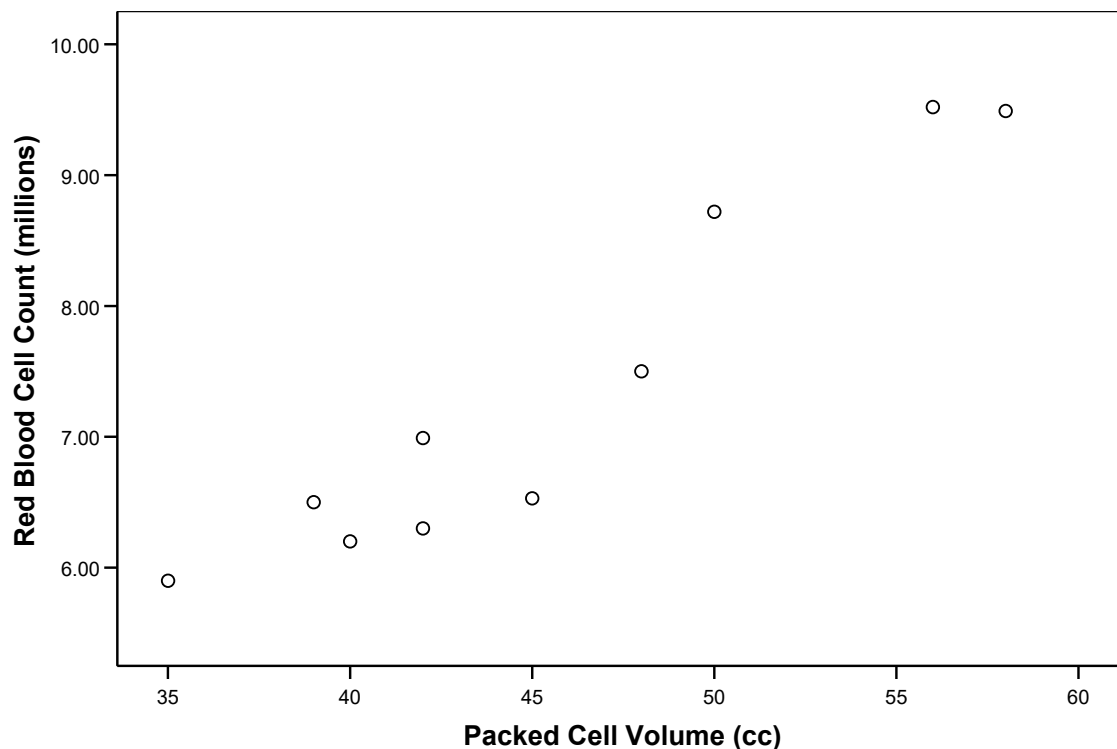
Questions 10 to 17 refer to the following information.

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:

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

The scatter plot, residual plot and SPSS output of these data are given below:

Scatter plot of Red Blood Cell Count versus Packed Cell Volume



Regression

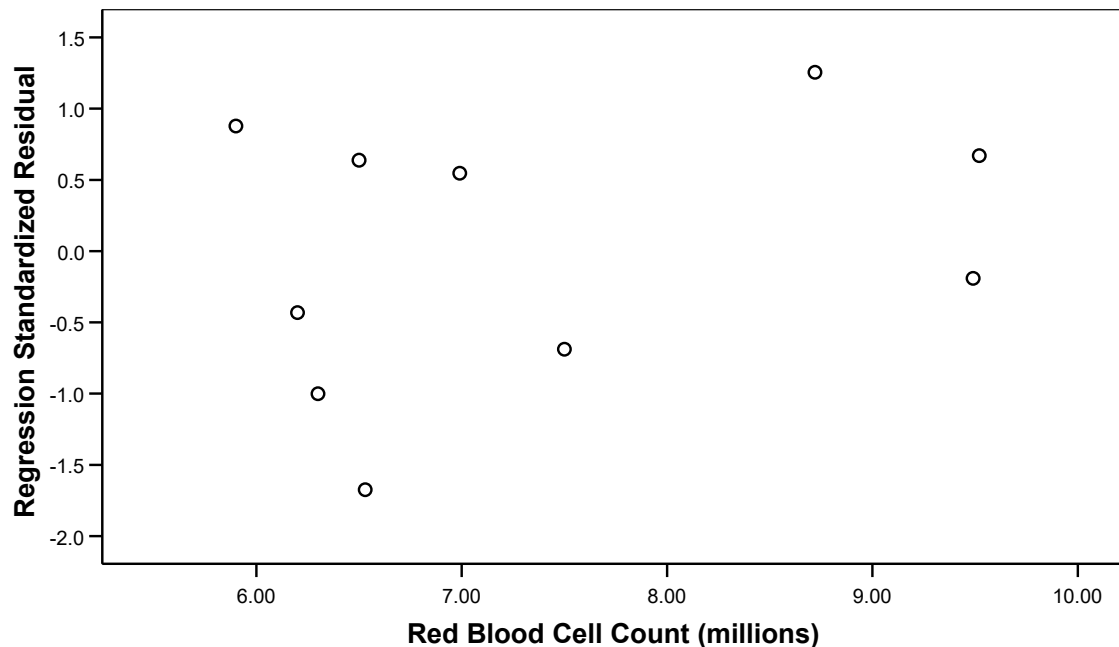
Coefficients(a)

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)

Scatterplot

Dependent Variable: Red Blood Cell Count (millions)



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

- (1) Average Packed Cell Volume = $0.177 \times \text{Red Blood Cell Count} - 0.68$
- (2) Average Red Blood Cell Count = $0.177 - 0.68$
- (3) Average Red Blood Cell Count = $0.177 \times \text{Packed Cell Volume} - 0.68$
- (4) Average Red Blood Cell Count = $0.177 - 0.68 \times \text{Packed Cell Volume}$
- (5) Average Packed Cell Volume = $0.177 - 0.68 \times \text{Red Blood Cell Count}$

11. For the data in the scatter plot on page 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) $r = 0.13$
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:
- (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 $H_1: \hat{\beta}_1 \neq 0$
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.
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

Questions 18 to 23 refer to the following information.

A 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 a random sample of 43 cars with engine sizes under 2500cc and associated plots are shown in Figure 1, Table 1 and Figure 2 (all given below).

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

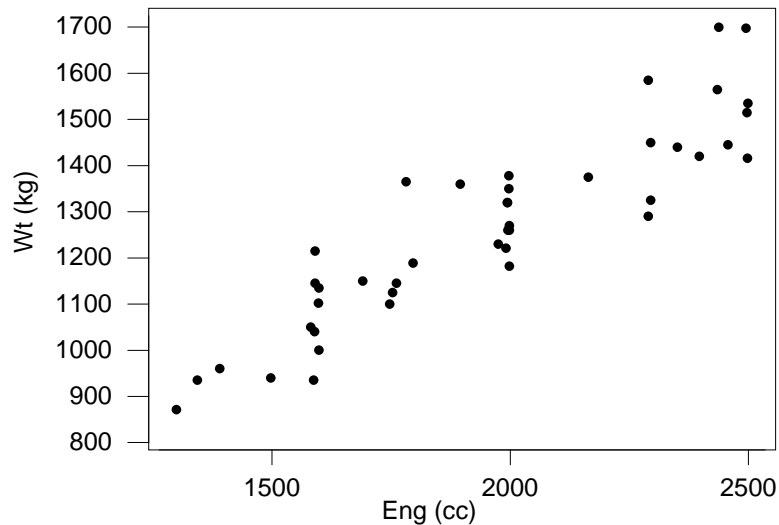


Figure 1: Scatter plot of weight versus engine size for cars with engines smaller than 2500cc

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

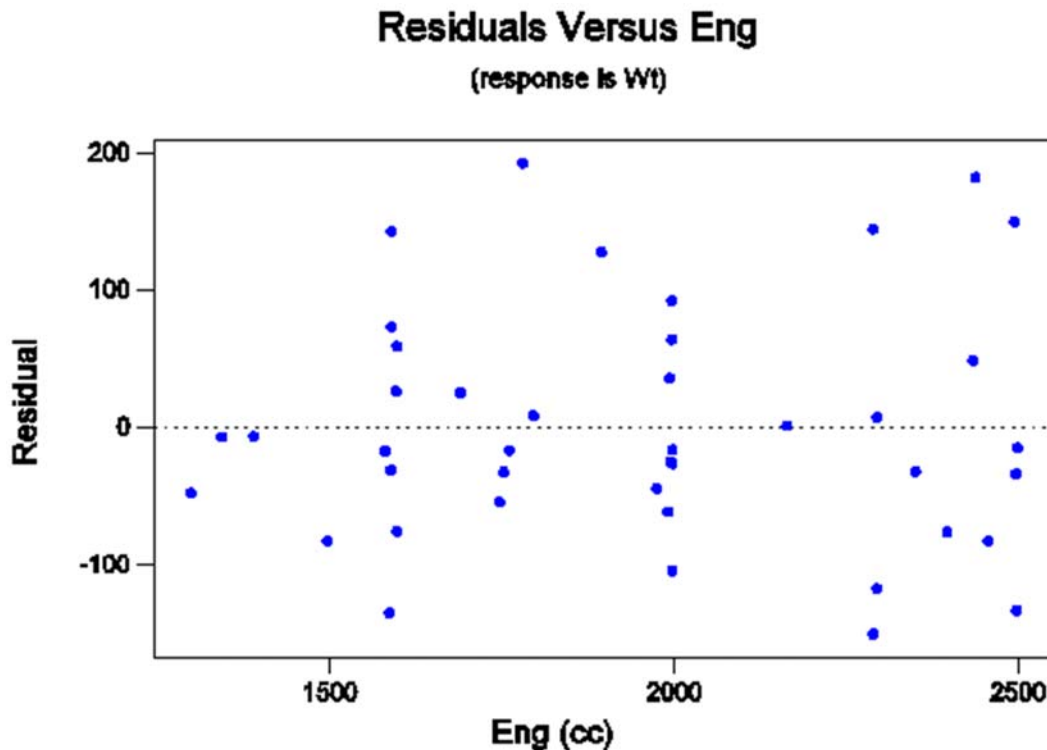


Figure 2: Scatter plot of residuals versus engine size for cars with engines smaller than 2500cc

18. 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

19. 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

20. 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
21. 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 $H_1: \beta_0 \neq 0$
22. You may need to refer to Figure 1 on page 19 and Figure 2 on page 20 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.

23. The researcher also used data from a random sample of 60 cars (irrespective of engine size, recall there were 43 cars previously) to investigate if engine size could be used to predict the weight of a car. The residual plot in Figure 3 (given below) was produced as part of this investigation. The **most** useful information provided by this plot about this linear regression model is that:

- (1) the errors are not independent.
- (2) the errors are not Normally distributed.
- (3) the relationship between weight and engine size is not linear.
- (4) the mean of the errors is not equal to zero.
- (5) all of the assumptions underlying this regression model are satisfied.

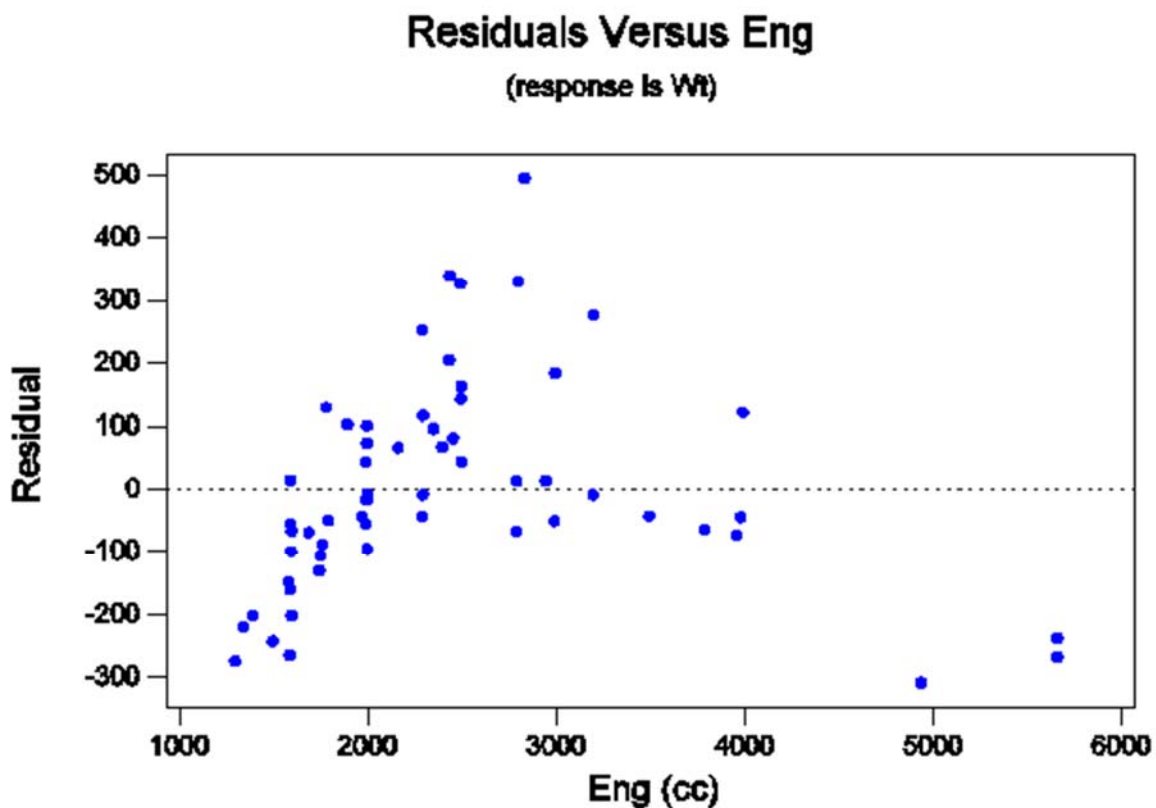


Figure 3: Scatter plot of residuals versus engine size for all cars

Questions 24 to 32 refer to the following information.

Various measurements were made on 312 African elephants, all with known ages up to 15 years.

Three of the variables recorded were:

- Sex** – Male
– Female
- Height** Shoulder height of the elephant in centimetres
- Age** Age of the elephant in years

Assume that these 312 elephants are a random sample of some larger population of the same breed of elephant in the same age range.

A scatter plot of **Height** against **Age** for the male elephants aged between 5 and 15 years is shown in Figure 5 below.

A simple linear regression analysis was carried out on the data from male elephants aged between 5 and 15 years to determine the expected height from age for elephants like these ones. A residual plot of this analysis is shown in Figure 6, page 24, and other regression output is shown in Tables 3 and 4, page 24.

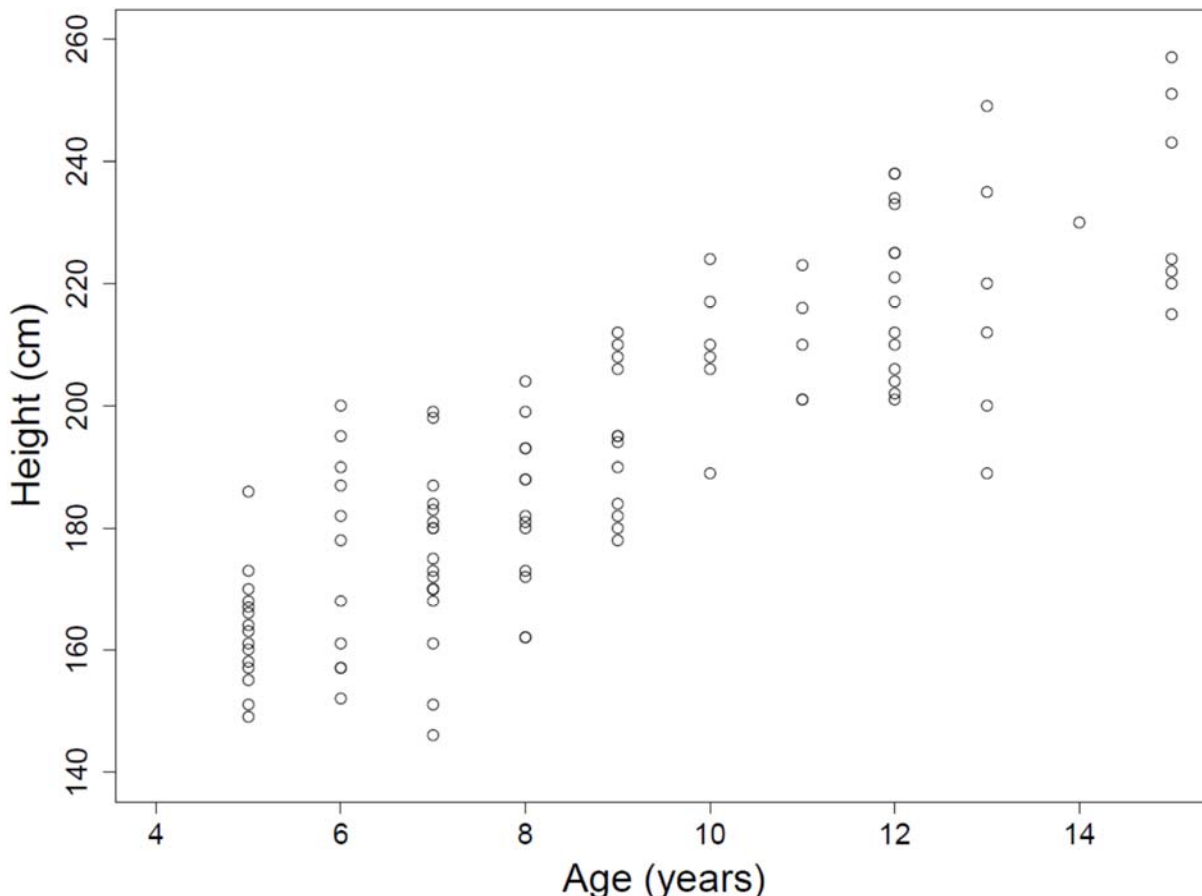


Figure 5: Male elephants, aged 5 years or older

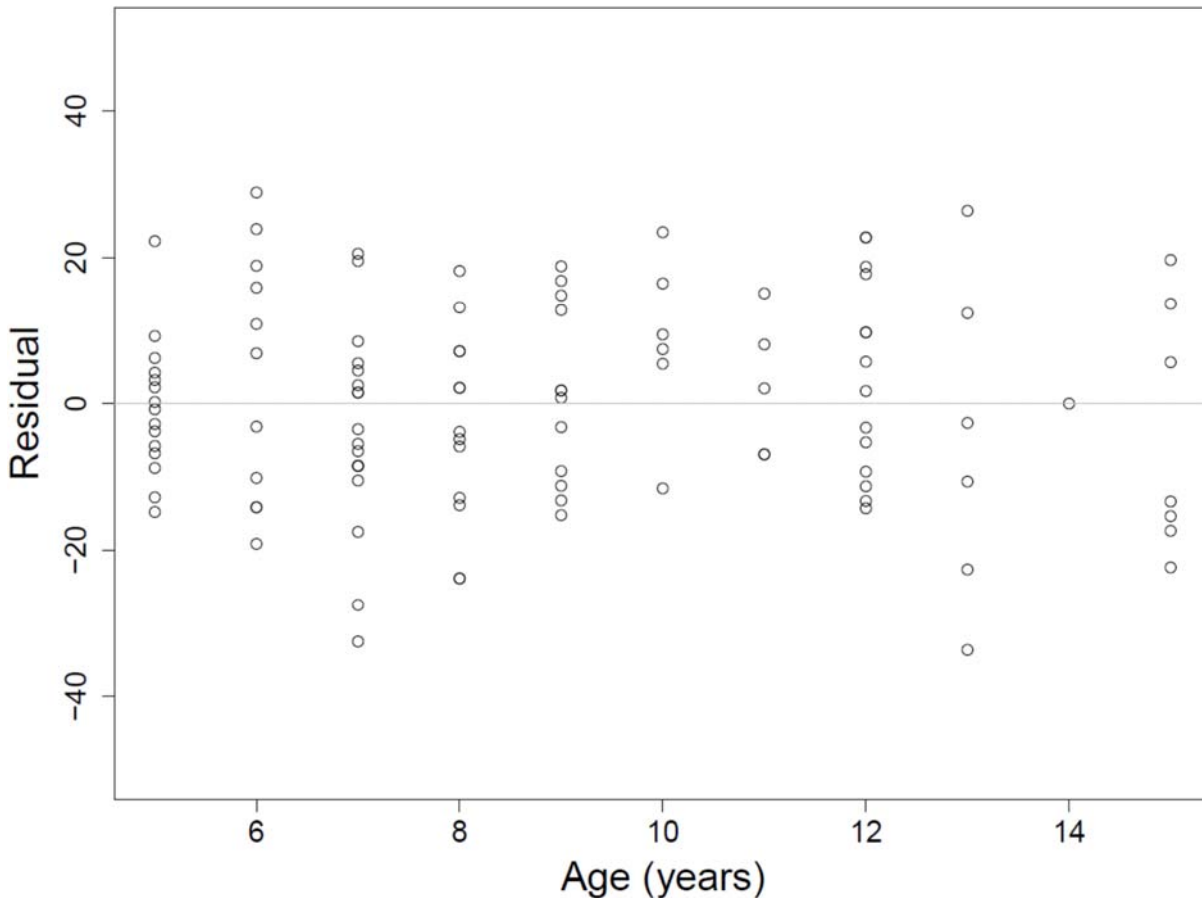


Figure 6: Residual plot for regression of Height against Age

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)	127.005	4.184		30.355	.000	118.710	135.301
	Age	7.357	.447	.848	16.459	.000	6.471	8.243

a. Dependent Variable: Height

Table 3: Simple linear regression output

Sex	Age	Height	LCMI_1	UMCI_1	LICI_1	UICI_1
M	5	161	159.46522	168.11615	136.17266	191.40870
M	6	168	167.48694	174.80854	143.62598	198.66949
M	7	180	175.39612	181.61347	151.05103	205.95856
M	9	193	190.59199	195.84582	165.81551	220.62230
M	12	214	211.47749	219.10266	187.74772	242.83243

Table 4: Simple linear regression prediction output

(Data courtesy of Dr Sam Ferreira, Conservation Ecology Research Unit, University of Pretoria, South Africa.)

Questions 24 and 25 refer to the scatter plot in Figure 5, page 23.

24. Which **one** of the following statements about these male elephants aged between 5 and 15 years is **false**?
- (1) For any pair of these elephants, the older elephant is not necessarily the taller.
 - (2) This plot shows that by the age of 12 years, elephants like these appear to have, on average, stopped growing in height.
 - (3) This plot suggests, on average, a reasonably constant increase in height from one year to the next for elephants like these.
 - (4) Some of the 6-year-old elephants are taller than at least one of the 13-year-old elephants.
 - (5) The shortest elephant is not in the youngest year group, but the tallest elephant is in the oldest year group.
25. Which **one** of the following statements **best** describes the linear relationship between **Height** and **Age**?
- The sample correlation coefficient, r , is closest to:
- (1) 0.95
 - (2) -0.35
 - (3) -0.85
 - (4) 0.85
 - (5) 0.35
26. Based on Figures 5 and 6, pages 23 and 24, which **one** of the following statements is **true**?
- (1) None of the features in these plots raises any concerns about conducting a simple linear regression analysis on these data.
 - (2) The residual plot confirms that the errors are related for male elephants like these and so we should be concerned about conducting a simple linear regression analysis on these data.
 - (3) The plots clearly suggest an increase in spread of the errors with an increase in age for male elephants like these and so we should be concerned about conducting a simple linear regression analysis on these data.
 - (4) The plots provide enough of a hint of the existence of two groups, (5–10 year-olds and 11–15 year-olds) to give us concerns about conducting a simple linear regression analysis on these data.
 - (5) There is a strong suggestion in the plots of a non-linear relationship for male elephants like these and it is strong enough for us to be concerned about using a linear model on these data.

Questions 27 to 32 refer to the regression output in Tables 3 and 4 on page 24 and assume that the simple linear regression model is valid. (Note that this assumption may not be true.)

27. The equation for the least squares regression line for this analysis is:
- (1) Expected **Height** = $127.005 + 7.357 \times \text{Age}$
 - (2) Expected **Height** = $127.005 + 4.184 \times \text{Age}$
 - (3) Expected **Height** = $7.357 + 0.447 \times \text{Age}$
 - (4) Expected **Height** = $7.357 + 0.848 \times \text{Age}$
 - (5) Expected **Height** = $7.357 + 127.005 \times \text{Age}$
28. Under this regression analysis, we would expect the height of an 8-year-old male elephant in this population to be approximately:
- (1) 135 cm
 - (2) 158 cm
 - (3) 178 cm
 - (4) 186 cm
 - (5) 173 cm
29. Which **one** of the following statements is the **best** interpretation of the confidence interval (6.471, 8.243) given in Table 3, page 24?
- With 95% confidence, we estimate that for male elephants in this population aged between 5 and 15 years:
- (1) the y -intercept of the true regression line is somewhere between 6.471 and 8.243.
 - (2) there is, on average, an increase in age of somewhere between 6.471 and 8.243 years associated with each additional centimetre increase in height.
 - (3) there is, on average, an increase in height of 1.772 cm for each additional year in age.
 - (4) the slope of the true regression line is somewhere between 6.471 and 8.243.
 - (5) there is, on average, an increase in height of somewhere between 6.471 cm and 8.243 cm for each additional year in age.

Questions 30 and 31 refer to the t -test for no linear relationship between **Height** and **Age**.

30. In a test for no linear relationship between **Height** and **Age**, the hypotheses are:

- (1) $H_0: \hat{\beta}_0 \neq 0$ $H_1: \hat{\beta}_0 = 0$
- (2) $H_0: \hat{\beta}_1 = 0$ $H_1: \hat{\beta}_1 \neq 0$
- (3) $H_0: \beta_1 = 0$ $H_1: \beta_1 \neq 0$
- (4) $H_0: \beta_1 \neq 0$ $H_1: \beta_1 = 0$
- (5) $H_0: \beta_0 = 0$ $H_1: \beta_0 \neq 0$

31. Which **one** of the following is a **correct** interpretation of the P -value for the test for no linear relationship?

- (1) We have evidence of a very strong positive linear relationship between **Height** and **Age**.
- (2) We have very strong evidence of no linear relationship between **Height** and **Age**.
- (3) We have very strong evidence of a linear relationship between **Height** and **Age**.
- (4) We have evidence of no linear relationship between **Height** and **Age**.
- (5) We have very strong evidence of a very strong positive linear relationship between **Height** and **Age**.

32. Referring to Table 4, page 24, which **one** of the following statements is **true**?

With 95% confidence, we estimate that:

- (1) the mean height of all 9-year-old male elephants in this population is 193.21891 cm.
- (2) the mean height of all 7-year-old male elephants in this population is somewhere between 151 cm and 206 cm.
- (3) a 12-year-old male elephant in this population is somewhere between 211 cm and 219 cm in height.
- (4) a 5-year-old male elephant in this population is 163.79068 cm in height.
- (5) a 6-year-old male elephant in this population is somewhere between 144 cm and 199 cm in height.

33. 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.
34. 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).
35. 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.

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

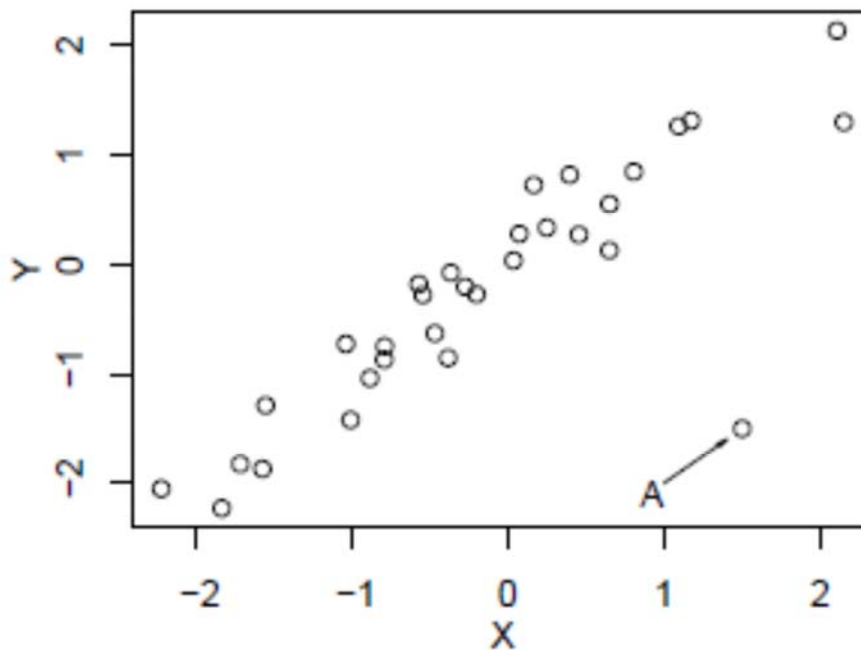


Figure 4: A scatter plot.

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.
37. 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.

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.

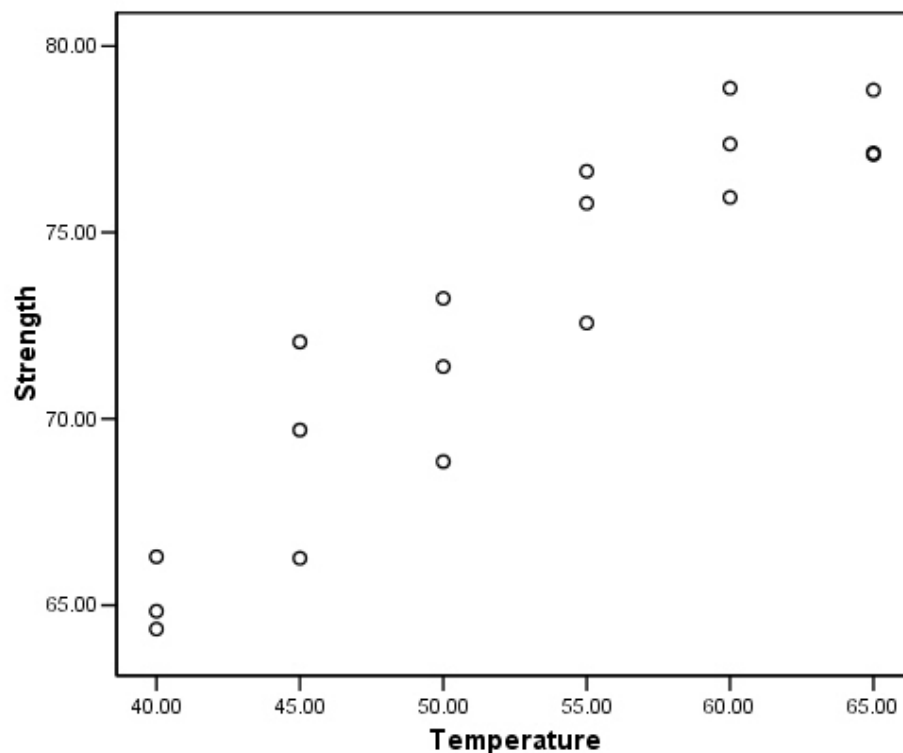
Questions 41 to 49 refer to the following information.

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 ($^{\circ}\text{C}$)		Strength	Temperature ($^{\circ}\text{C}$)
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:

Scatterplot of particle board strength against temperature



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

41. The fitted least squares regression line for these data is:
- (1) Predicted Temperature = $45.454 + 0.518 \times \text{Strength}$
 - (2) Predicted Strength = $2.846 + 0.054 \times \text{Temperature}$
 - (3) Predicted Strength = $45.454 + 0.518 \times \text{Temperature}$
 - (4) Predicted Strength = $0.518 + 45.454 \times \text{Temperature}$
 - (5) Predicted Temperature = $0.518 + 45.454 \times \text{Strength}$
42. For the data in the scatter plot on page 31, which one of the following values would be the closest to the sample correlation coefficient, r ?
- (1) $r = 0.9$
 - (2) $r = 0.6$
 - (3) $r = 0.3$
 - (4) $r = 0.5$
 - (5) $r = 0.1$
43. 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 $H_1 : \hat{\beta}_1 \neq 0$

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

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

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

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

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

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

45. When the hypothesis test referred to in Questions 43 and 44 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.

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

(1) 102.4

(4) 79.9

(2) 73.9

(5) 47.0

(3) 13.8

47. The baking temperature and particle board strength for sample number 11 was 55°C 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
48. The fitted least squares regression line indicates that for each increase of 2.5°C 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.
49. Why is the prediction interval for the strength of particle board baked at a temperature of 55°C more useful than the corresponding point estimate given in Question 46?
- (1) Because the prediction interval is **only one** of the plausible values of that the strength could be when the baking temperature is 55°C .
 - (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°C .
 - (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°C .
 - (5) The prediction interval is not as useful as a corresponding confidence interval.

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.

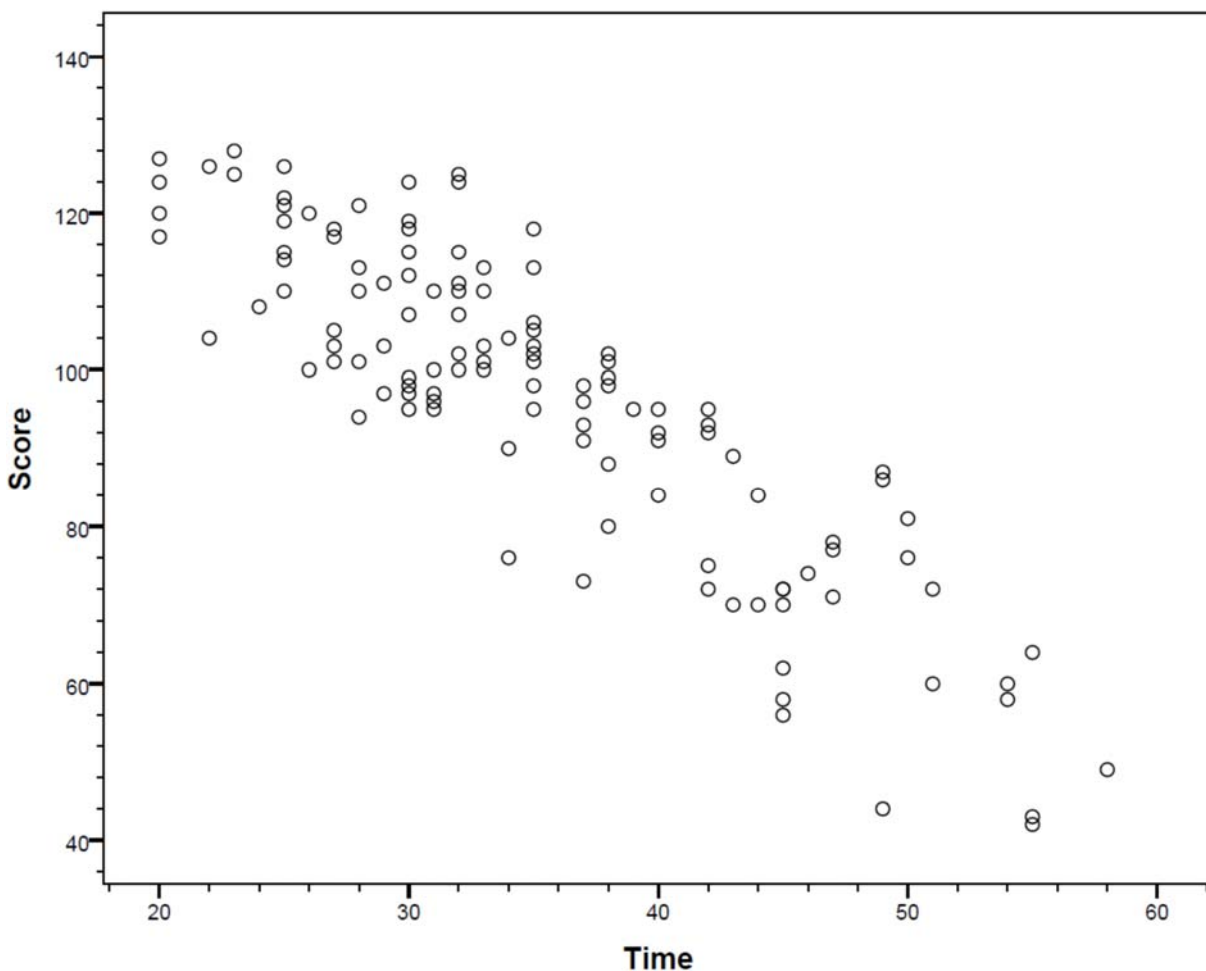


Figure 7: Total psychometric score against walking time (seconds)

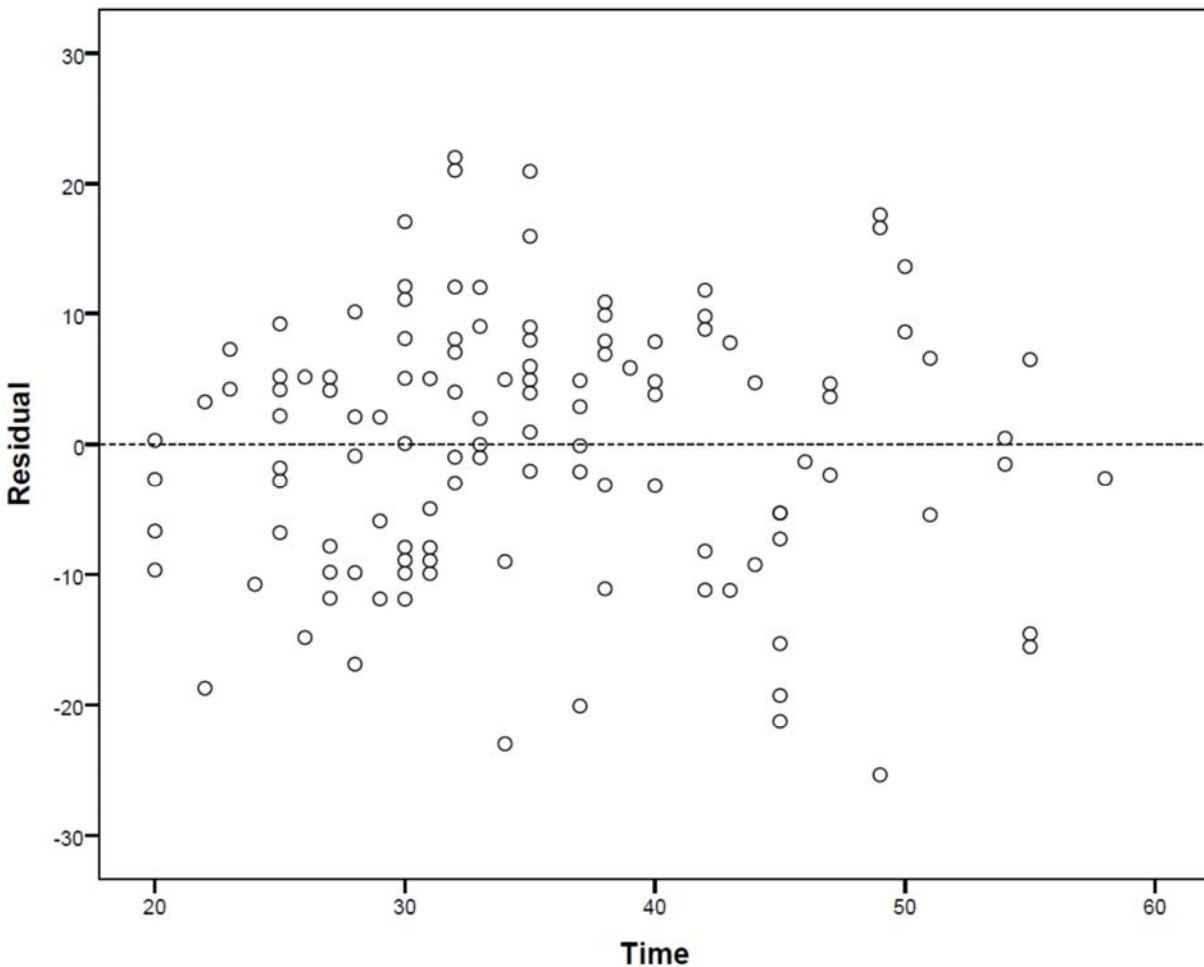


Figure 8: Residual plot

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** = $-1.976 + 0.102 \times \text{Time}$
 - (2) Average **Time** = $-1.976 + 166.194 \times \text{Score}$
 - (3) Average **Score** = $166.194 - 1.976 \times \text{Time}$
 - (4) Average **Score** = $-1.976 + 166.194 \times \text{Time}$
 - (5) Average **Time** = $166.194 - 1.976 \times \text{Score}$
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
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
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

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:

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

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**:

- (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.
- (2) will be centred either higher or lower than the **prediction interval**, depending on the estimates about which each interval is centred.
- (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.
- (4) will be wider and centred higher than the **prediction interval** because it has to account for two sources of variability.
- (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**?

- (1) The least squares regression technique minimises the sum of the squared residuals.
- (2) t -procedures are used to find confidence intervals for the slope of the true line.
- (3) A strong relationship between two numeric variables is indicated when the slope of the line is close to either 1 or -1.
- (4) The X -variable is used to predict or explain the behaviour of the Y -variable.
- (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 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 .

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 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: $estimate \pm t \times se(estimate)$

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

Applications:

1. Single mean μ : $estimate = \bar{x}$; $df = n - 1$
2. Single proportion p : $estimate = \hat{p}$; $df = \infty$
3. Difference between two means $\mu_1 - \mu_2$: (independent samples)
 $estimate = \bar{x}_1 - \bar{x}_2$; $df = \min(n_1 - 1, n_2 - 1)$

4. Difference between two proportions $p_1 - p_2$:

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

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: $f_0 = \frac{s_B^2}{s_W^2}$; $df_1 = k - 1$, $df_2 = n_{tot} - k$

The Chi-square test

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

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

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

Regression

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

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