INSTRUCTIONS: Total marks 160.
Answer ALL questions. Note that the questions are not worth equal marks. Question 1 contains short answer questions worth a total of 65 marks. Questions 2-7 are longer and are worth a total of 95 marks. Show all relevant working.

PLEASE NOTE
Examination candidates may only bring authorised materials into the examination room. If a supervisor finds, during the examination, that you have unauthorised material, in whatever form, in the vicinity of your desk or on your person, whether in the examination room or the toilets or en route to/from the toilets, the matter will be reported to the head of school and disciplinary action will normally be taken against you. This action may result in your being deprived of any credit for this examination or even, in some cases, for the whole unit. This will apply regardless of whether the material has been used at the time it is found.

Therefore, any candidate who has brought any unauthorised material whatsoever into the examination room should declare it to the supervisor immediately. Candidates who are uncertain whether any material is authorised should ask the supervisor for clarification.
1. [65 marks]

(a)  (i) Discuss in not more than a page the purpose of exploratory data analysis, and the common techniques used for continuous data.  

(ii) A market researcher obtains data on which one of six cleaners shoppers use. Suggest one graph and one summary statistic that should be produced as part of the exploratory data analysis.  

Solution.

(i) The purpose of exploratory data analysis is to uncover the key and salient features of the data, such as central location, spread, skewness, outliers and modality. There are two aspects of exploratory data analysis: graphical analysis and summary statistics. Common techniques used for exploratory analysis of continuous data are:
1. histogram, to obtain an idea of central location, spread, skewness, modality and outliers;
2. mean and median, which are measures of central location;
3. variance or standard deviation, which measure spread;
4. upper and lower quartiles and skewness, which allow us to determine the skewness of the data.

(ii) An appropriate summary statistic is the proportion of shoppers who use each of the cleaners, and these can also be displayed as a pie or a bar chart.

(b) If \( P(A) = 0.5, \ P(A \cup B) = 0.8, \) and \( A \) and \( B \) are independent, find \( P(B) \).

Solution. \( P (A \cup B) = P (A) + P (B) - P (A \cap B) \), and independence of \( A \) and \( B \) implies that \( P (A \cap B) = P (A) \times P (B) = 0.5P (B) \). Then substituting values in the first equation gives \( 0.8 = 0.5 + P(B) - 0.5P(B) \) which gives \( P(B) = 0.6 \).

(c) A toy manufacturer buys pre-assembled robotic arms from two different suppliers — 60% of the total order come from supplier 1, and the rest from supplier 2. Past data has shown that the quality control standards of the suppliers are different. One percent of the arms produced by supplier 1 are defective, while supplier 2 produces defective arms at the rate of 3%.

QUESTION 1(c) CONTINUES OVER THE PAGE
(c) (Continued)

Let $S_i$ be the event that an arm came from supplier $i$, $i = 1, 2$, and let $D$ be the event that an arm is defective.

(i) Draw a tree diagram to model this situation. [2 marks]

(ii) What percentage of the arms in the manufacturer's inventory are non-defective? [2 marks]

(iii) If an arm is found to be defective, what is the probability that it came from supplier 1? [2 marks]

Solution.

(i) The tree diagram for the problem is given below.

(ii) $P(D) = 0.6 \times 0.99 + 0.4 \times 0.97 = 0.982$, that is, 98.2%.

(iii) $P(S_1 \mid D) = P(S_1 \cap D) / P(D) = 0.006 / (1 - 0.982) = 1/3$.

(d) The business manager of a production company assesses that the joint probability mass function of the number of items ($N$) produced per week and profit ($P$) per item is as shown in the table below.

<table>
<thead>
<tr>
<th>$P$</th>
<th>20,000</th>
<th>25,000</th>
<th>30,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>$3$</td>
<td>0.0</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>$3.50$</td>
<td>0.2</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>$4$</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

(i) Find the expected profit per item in the next week. [2 marks]

(ii) Find the expected profit per item if 25,000 items are produced per week. [4 marks]
1(d) (Continued)

(iii) What is the expected total profit from sales of this product in the next week? [4 marks]

Solution.

(i) The distribution of profit is given below.

<table>
<thead>
<tr>
<th>p</th>
<th>$3</th>
<th>$3.50</th>
<th>$4</th>
</tr>
</thead>
<tbody>
<tr>
<td>pP(d)</td>
<td>0.3</td>
<td>0.4</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Then $E(P) = 3 \times 0.3 + 3.50 \times 0.4 + 4 \times 0.3 = 3.50$.

(ii) Now we need the conditional distribution of profit given that 25,000 items are produced in a week. This is simply the column corresponding to $N = 25,000$, divided by the column total, and is given below.

<table>
<thead>
<tr>
<th>p</th>
<th>$3</th>
<th>$3.50</th>
<th>$4</th>
</tr>
</thead>
<tbody>
<tr>
<td>pP</td>
<td>N(p</td>
<td>n=25,000)</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Then the expected profit per item when 25,000 items are produced per week is $3 \times 0.6 + 3.50 \times 0.4 + 0 = 3.20$.

(iii) The profit is given by $P \times N$, so we need

$$E(PN) = 3 (25,000 \times 0.15 + 30,000 \times 0.15)$$
$$+ 3.50 (20,000 \times 0.2 + 25,000 \times 0.1 + 30,000 \times 0.1)$$
$$+ 4 (20,000 \times 0.3) = 82,000.$$ 

(e) The distance (in 1,000 km) that a courier travels per week is a continuous random variable $X$ with probability density function

$$f_X(x) = \begin{cases} \frac{2}{3} + \frac{2x}{3} & \text{if } 0 < x < 1, \\ 0, & \text{otherwise.} \end{cases}$$

What is the probability that the courier travels more than 500 km next week? [4 marks]

Solution. The graph of the pdf is given below, and the required probability is given by the shaded area, that is, the area of a trapezium.
(f) Suppose the income of executives is normally distributed with the incomes of men and women independent of each other. The means and standard deviations (in $10,000) respectively are 20, 2.5 for men and 15, 2.0 for women.

(i) What is the probability that a randomly chosen female executive earns more than a randomly chosen male executive? \([4 \text{ marks}]\)

(ii) Find the probability that an executive couple has a combined income of more than $400,000. \([3 \text{ marks}]\)

(iii) Ten executive couples are selected at random. What is the probability that at least one of the couples have a combined income of more than $400,000? \([3 \text{ marks}]\)

(iv) A researcher selects 400 executive couples at random. Use a suitable approximation to estimate the probability that at most 20 of them have a combined income of more than $400,000. \([5 \text{ marks}]\)

**Solution.**

(i) Let the random variables \(M\) and \(F\) respectively denote the income of the male and female executive. Then \(M \sim N(20, 2.5^2), F \sim N(15, 2.0^2)\).

Put \(D = F - M, \text{ so } D \sim N(-5, 10.25)\). Then the required probability...
1(f)(i) (Continued) is

\[ P(D > 0) = P\left(Z > \frac{0 - (-5)}{\sqrt{10.25}}\right) = P(Z > 1.56) = 1 - 0.9406 = 0.0594. \]

(ii) Put \( T = F + M \sim N(35, 10.25) \). We need

\[ P(T > 40) = P\left(Z > \frac{40 - 35}{\sqrt{10.25}}\right) = P(Z > 1.56) = 0.0594. \]

(iii) Let the random variable \( X \) denote the number of executive couples out of 10 who have a combined income of more than $400,000. Then \( X \sim Bin(10, 0.0594) \). We need

\[ P(X \geq 1) = 1 - P(X = 0) = 1 - (1 - 0.0594)^{10} = 0.4579. \]

(iv) Let the random variable \( Y \) denote the number of couples out of 400 who have a combined income of more than $400,000. Then \( Y \sim Bin(400, 0.0594) \), and \( E(X) = 400 \times 0.0594 = 23.76, Var(X) = 400 \times 0.0594 \times (1 - 0.0594) = 22.35 \). Note that here \( n = 400 > 30, np = 23.76 > 5 \) and \( n(1 - p) = 376.24 > 5 \), so the binomial distribution can be approximated by the \( N(23.76, 22.35) \) distribution. Then

\[ P(X \leq 20) \approx P\left(Z < \frac{20.5 - 23.76}{\sqrt{22.35}}\right) = P(Z < -0.69) = 1 - 0.7549 = 0.2451. \]

(g) A 95% confidence interval for a population mean based on a sample size of 200 is (35, 40). Find the value of the observed sample mean and the sample standard deviation. [2 marks]

\[ \bar{x} = (35 + 40)/2 = 37.5. \] Then the half width of the confidence interval is \( 37.5 - 35 = 2.5 = z_{0.025} \times s/\sqrt{200} \), so \( s = 2.5 \times \sqrt{200}/1.96 = 18.04. \]

(h) Discuss the three factors that affect the width of a confidence interval for a population proportion, and the effect of each. [6 marks]

Solution.
1(h) (Continued)

(i) Sample size — as sample size increases, the width of the interval decreases.

(ii) Sample proportion — as sample proportion moves away from 0.5, the width of the interval decreases.

(iii) Confidence level — as the confidence level increases, the width of the interval increases.

(i) A test of hypothesis for a population mean fails to reject the null hypothesis at the 5% level of significance. Does this mean that the null hypothesis is true? Explain with reference to Type I and/or Type II errors. [3 marks]

Solution. Failure to reject the null hypothesis does not mean that the hypothesis is true. A Type II error could have occurred, that is, the null hypothesis may be false, but the data does not provide sufficient evidence against it at the 5% level of significance.

(j) A survey of 200 stockholders of a corporation was taken to determine whether they favoured distribution of the profits as dividends or re-investing the profits in the firm. The table below presents the data, categorising the opinion with respect to the number of shares owned.

<table>
<thead>
<tr>
<th>No. of Shares Owned</th>
<th>Pay dividends</th>
<th>Reinvest</th>
<th>Indifferent</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–99</td>
<td>35</td>
<td>17</td>
<td>8</td>
<td>60</td>
</tr>
<tr>
<td>100–249</td>
<td>31</td>
<td>23</td>
<td>6</td>
<td>60</td>
</tr>
<tr>
<td>250 or more</td>
<td>34</td>
<td>30</td>
<td>16</td>
<td>80</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>100</strong></td>
<td><strong>70</strong></td>
<td><strong>30</strong></td>
<td><strong>200</strong></td>
</tr>
</tbody>
</table>

The corporation wants to determine if the opinion of the investor is independent of the number of shares owned.

(i) State the appropriate null and alternative hypotheses to be tested here. [2 marks]

(ii) If the null hypothesis is true, find the expected frequency of large investors (that is, owning 250 or more shares) favouring re-investment of profits. [2 marks]

(iii) What is the distribution of the test statistic assuming the null hypothesis is true (include the value of any parameters)? [2 marks]
1(j) (Continued)

(iv) The hypothesis test is performed in Excel, giving a p-value of 0.2573. Determine using this information if the opinion of the investors is independent of the number of shares owned. Justify your answer.

Solution.

(i) \( H_0 \): The opinion of the investor is independent of the number of shares owned.
\( H_1 \): \( H_0 \) is false.
(ii) \( \frac{80 \times 70}{200} = 28 \).
(iii) \( \chi^2 \).
(iv) \( p\)-value > 0.05, so the data provides insufficient evidence to reject the null hypothesis at the 5% level of significance. We conclude that the opinion of investors is independent of the number of shares owned by them.
2. [15 marks] The manager of an automotive company has been asked to determine if the new design of their car has increased the length of time before the car has a major problem. If this is true then the company will extend the new vehicle warranty in the hope of attracting more customers. Prior to the design changes, the cars lasted on average 43 months before having a major problem. Data on the length of time before a major problem for a random sample of 50 cars with the new designs gave a mean of 44 months and a standard deviation of 2 months. The manager decides to conduct a hypothesis test as part of her analysis. Let $\mu$ denote the mean length of time (months) before the car with the new design has a major problem.

(a) State the null and alternative hypotheses that the manager is testing.

(b) Give a formula for the test statistic and state its distribution under the null hypothesis, specifying the value of any parameters. State any assumption(s) required for this distribution to be valid.

(c) Find the observed value of the test statistic.

(d) Test the hypotheses at the 2.5% level of significance. On the basis of this analysis, should the warranty of the car be extended? Justify your answer.

(e) What are the consequences of a Type I error here?

(f) How might your analysis change if the sample size was 25? Give reasons for your answer.

Solution.

(a) $H_0 : \mu = 43 \quad H_1 : \mu > 43.$

(b) $T = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim t_{49},$ by CLT, since the sample size is greater than 30. No other assumptions are required.

(c) $t_{obs} = \frac{44 - 43}{2/\sqrt{50}} = 3.5355.$

(d) Critical value $t_{0.025}^{49} \approx 2.009$ (the $t_{0.025}^{50}$ value), and $t_{obs}$ is in the critical region, so the data provides sufficient evidence to reject the null hypothesis at the 2.5% level of significance. Thus we conclude that the mean time before a major problem for the new design is greater than 43 months, that is, there is an improvement in the time for the new design. The warranty should be extended.
2 (Continued)

(e) If a Type I error is made then the null hypothesis would be rejected even though it is true. This would cause the warranty to be extended unjustifiably, and incur a cost to the manufacturer in repairs.

(f) Now the sample size \( n < 30 \), so the test statistic is \( T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \), which has an \( t_{24} \) distribution. Now we need the assumption that the time before a major problem occurs is normally distributed.

3. [10 marks] In order to determine the effect of advertising in the Yellow Pages, Telstra took a sample of 25 retail stores that did not advertise in the Yellow Pages last year but did so this year. The annual sales (in thousands of dollars) for the stores showed a mean increase of 19.75 with a standard deviation of the difference being 30.63.

Can we infer that advertising in the Yellow Pages increased sales? As part of your answer state the appropriate hypotheses and a p-value. State any assumptions required in your analysis. [10 marks]

Solution. Let the random variable \( D \) denote the difference between this year’s and last year’s sales. Then we assume that \( D \sim N(\mu, \sigma^2) \). The hypotheses of interest here are

\[ H_0 : \mu = 0 \text{ (No change in mean sales.)} \]
\[ H_1 : \mu > 0 \text{ (The mean sales have increased.)} \]

The test statistic is

\[ T = \frac{\bar{T} - \mu}{S/\sqrt{25}} \sim t_{24}. \]

The observed value of the test statistic is

\[ t_{obs} = \frac{19.75 - 0}{30.63/\sqrt{25}} = 3.224. \]

The 2.5% level critical value for the \( t_{24} \) distribution is 2.064, so the CR is \( T > 2.064 \). Since \( t_{obs} \) is in the CR, the data provides sufficient evidence to reject the null hypothesis at the 2.5% level of significance. We conclude that advertising in the yellow pages has significantly increased sales.
4. [15 marks] In checking the reliability of a bank’s records, auditing firms sometimes ask a sample of the bank’s customers to confirm the accuracy of their savings account balances as reported by the bank. A bank claims that the true fraction of accounts on which there is a disagreement is no more than 0.05. You as an auditor doubt this claim. Of 400 savings account customers questioned by you, 35 said that their balances disagreed with that reported by the bank.

Does this data provide evidence that the true fraction of accounts subject to disagreement exceeds 0.05? To answer this question you perform a hypothesis test.

Let \( p \) denote the proportion of savings accounts on which there is a disagreement.

(a) State the null and alternative hypotheses that you are testing. [2 marks]

(b) Give an expression for the test statistic and state its distribution under the null hypothesis. [2 marks]

(c) Find the observed value of the test statistic. [2 marks]

(d) Conduct the hypothesis test at the 2.5% level of significance. On the basis of your analysis decide if the bank’s claim is justified. [5 marks]

(e) Find a 95% confidence interval for the proportion of savings accounts that have mis-reported balances. [3 marks]

(f) Verify any assumptions required in the above analysis. [1 mark]

Solution.

(a) \( H_0 : p = 0.05 \quad H_1 : p > 0.05. \)

(b) \( Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1), \) where \( \hat{p} \) is the sample proportion of accounts on which there is a disagreement.

(c) \( \hat{p}_{obs} = 35/400 = 0.0875, \) so \( z_{obs} = \frac{0.0875 - 0.05}{\sqrt{(0.05)(0.95)/400}} = 3.44. \)

(d) \( p-value = P(Z > 3.44) = 1 - 0.9997 = 0.0003 < 0.025, \) so the data provides strong evidence against the null hypothesis at the 2.5% level of significance.

We conclude that the bank’s claim is not justified, and the data indicates that the actual proportion of accounts for which there are disagreements with the balance reported by the bank is more than 0.05.

(e) 95% CI for \( p = 0.0875 \pm 1.96 \sqrt{\frac{(0.0875)(0.9125)}{400}} = 0.0875 \pm 0.0277, \) so 95% CI for \( p = (0.0598, 0.1152). \)

QUESTION 4 CONTINUES OVER THE PAGE
5. [15 marks] The manager of a company that manufactures automobile air conditioners is considering switching his supplier of condensers. Supplier A, the current supplier, prices its product 5% higher than Supplier B does. To maintain the company’s reputation for quality, the manager wants to make sure that Supplier B’s condensers last at least as long as those of Supplier A. To test this, 50 mid-size cars were selected at random. Of these 25 at random were equipped with air conditioners using type A condensers, while the others were equipped with type B condensers. The distance (in thousands of km) travelled by each car before the condenser broke down was recorded. The summary statistics are given below.

<table>
<thead>
<tr>
<th></th>
<th>Type A</th>
<th>Type B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size</td>
<td>$n_1 = 25$</td>
<td>$n_2 = 25$</td>
</tr>
<tr>
<td>Sample mean</td>
<td>$\bar{x}_1 = 115.5$</td>
<td>$\bar{x}_2 = 109.4$</td>
</tr>
<tr>
<td>Sample standard deviation</td>
<td>$s_1 = 21.7$</td>
<td>$s_2 = 22.4$</td>
</tr>
</tbody>
</table>

[Information: The test statistic for a hypothesis test for a difference of two population means based on two independent samples is

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1 + n_2 - 2},$$

where

$$S_p^2 = \frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2}$$
is the pooled variance.]

You are asked to analyse this data, and as part of your analysis you decide to conduct a test of hypothesis.

Let $\mu_1$ and $\mu_2$ be the mean distances travelled before failure for Type A and Type B condensers respectively.

(a) State the relevant null and alternative hypotheses. [2 marks]

(b) State any assumptions required in the analysis. [2 marks]
5 (Continued)
(c) Find the observed value of the test statistic. [3 marks]
(d) Perform the hypothesis test using a significance level of 2.5%. [3 marks]
(e) Write a short statement with a recommendation to the manager based on your analysis. [2 marks]
(f) Find a 95% confidence interval for the difference in mean distances travelled before failure for the the condenser types. [3 marks]

Solution.
(a) \( H_0 : \mu_1 - \mu_2 = 0 \quad H_1 : \mu_1 - \mu_2 > 0 \).
(b) The distances travelled before the condensers breaks down are are normally distributed, with a common variance.
(c) We first compute the pooled variance: \( s_p^2 = \frac{24 (21.7)^2 + 24 (22.4)^2}{48} = 486.325 \), so then \( t_{obs} = \frac{(115.5 - 109.4) - 0}{\sqrt{486.325} \sqrt{\frac{1}{25} + \frac{1}{25}}} = 0.978 \).
(d) The 2.5% critical level for the \( t_{48} \) distribution is approximately 2.009 (the \( t_{50} \) value), so the CR is \( T > 2.009 \). Since \( t_{obs} \) is not in the CR, the data does not provide sufficient evidence to reject the null hypothesis at the 2.5% level of significance. We conclude that the mean distances travelled before breakdown of the condenser are not different for the two suppliers.
(e) The quality of the the two suppliers is the same, so we should switch suppliers.
(f) 
95% CL for \( \mu_1 - \mu_2 = (\bar{x}_1 - \bar{x}_2) \pm t_{0.025} \times s_p \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 6.1 \pm 12.54 \)
so a 95% CI is \((-6.4, 18.6)\).

6. [20 marks] An important quality characteristic of a certain new product is a measure of strength that can only be tested by a rather costly destructive test. Three machines, A, B and C are used to produce the material in a factory. To test if the material produced by each machine is of the same strength, random samples of ten items from each machine were tested. An ANOVA was performed on the data and the results, along with some diagnostics, are given below. 

**QUESTION 6 CONTINUES OVER THE PAGE**
Let $\mu_A$, $\mu_B$ and $\mu_C$ denote the mean strength of the material produced by machines A, B and C respectively.

(a) State the hypotheses that are being tested here. [2 marks]

(b) State the assumptions of the analysis and use the output to check if they hold. [6 marks]

(c) Find the values of the F-Ratio and the Total Sum of Squares in the ANOVA table. [2 marks]

(d) Perform a test of the hypotheses and state your conclusion. [2 marks]

(e) Find 95% confidence intervals for $\mu_A - \mu_B$ and $\mu_A - \mu_C$. Comment on the implications of these intervals. [6 marks]

(f) Write a short statement to the production manager, informing her of the results of your analysis. [2 marks]
Solution.

(a) $H_0 : \mu_A = \mu_B = \mu_C \quad H_1 : H_0$ is false.

(b) (i) **Residuals are normal.** The histogram of residuals is not too different from that expected for a normal distribution. We conclude that the residuals are normal.

(ii) **Residuals have a constant variance.** The scatterplot of residuals against fitted values shows some change in spread. In particular, machine C seems to have a smaller spread. However, the change is not too great, so we conclude that the constant variance assumption is satisfied.

(iii) **Residuals are independent.** The scatterplot of residuals against fitted values shows some pattern, but this is mainly a change in spread, so we conclude that there is no reason to doubt the independence assumption.

(c) F-Ratio $= \frac{60790}{14974.074} = 4.0597$. Degrees of freedom are 2 for SSB, 27 for SSW and 29 to SST. Then

$$SSB = 2 \times 60790 = 121,580, SSW = 27 \times 14974.074 = 404,300,$$

so SST $= 121,580 + 404,300 = 525,880$.

(d) $p$-value $= 0.0287 < 0.05$, so the data provides sufficient evidence to reject the null hypothesis at the 5% level of significance. We conclude that the strength of the material produced by each machine is not the same on average.

(e) The 2.5% critical value for the $t_{27}$ distribution is 2.052, so 95% CL for $\mu_A - \mu_B = (1301 - 1305) \pm 2.052 \sqrt{\frac{14974.0774}{120} + \frac{1}{10}} = -4 \pm 112.30$, so a 95% CI for $\mu_A - \mu_B = (-116.30, 108.30)$. Similarly, 95% CL for $\mu_A - \mu_B = (1301 - 1438) \pm 112.30$, so a 95% CI for $\mu_A - \mu_C = (-249.30, -24.30)$.

The first CI includes 0, so at the 5% level of significance we conclude that the mean strength of materials produced by machines A and B are not different. The second CI does not include 0, so at the 5% level of significance we conclude that the mean strength of the materials produced by machines A and C are different. In fact, this CI lies below 0, indicating that machine C produces materials of higher mean strength than machine A.

(f) The data indicates that the mean strength of materials produced by machines A and B are the same, and these are both lower than the mean strength of materials produced by machine C.
7. [20 marks] The owner of the Original Italian Pizza restaurant wants to investigate if the sales of his speciality, deep dish pizza, are affected by the amount spent on advertising per month. Data on sales were collected from each of his 15 outlets over a month, and a simple linear regression analysis was performed in Excel. Some of the output is given below.

**SUMMARY OUTPUT**

Regression Statistics

<table>
<thead>
<tr>
<th>Multiple R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Standard Error</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9339944</td>
<td>0.87234554</td>
<td>0.862525966</td>
<td>4882.513719</td>
<td>15</td>
</tr>
</tbody>
</table>

ANOVA

<table>
<thead>
<tr>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Significance F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>2117789777</td>
<td>2.12E+09</td>
<td>88.83741</td>
</tr>
<tr>
<td>Residual</td>
<td>13</td>
<td>309906222.7</td>
<td>23838940</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>2427696000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Coefficients

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-32655.0449</td>
<td>10304.60514</td>
<td>-3.168976</td>
<td>0.007396</td>
<td>-54 916.79079 -10393.3</td>
</tr>
<tr>
<td>Advertising Expenditures</td>
<td>1.754764166</td>
<td>0.186174761</td>
<td>9.42536</td>
<td>3.56E-07</td>
<td>1.352558048 2.15697</td>
</tr>
</tbody>
</table>

**Number Sold against Advertising Expenditure**

$$y = 1.7548x - 32655$$

**$R^2 = 0.8723$**

**Normal Probability Plot**

**Residuals**

**Advertising Expenditures Residual Plot**

**Predicted Number Sold**

**Residuals**

**QUESTION 7(a) CONTINUES OVER THE PAGE**
Semester 2 Examinations  
STAT106  
November 2007

7(a) (Continued)  
(a) State the linear regression model and explain each term in the model.  
[4 marks]

(b) Is there a significant positive linear relationship between Number Sold and Advertising Expenditure? As part of your answer state the relevant hypotheses and test them using the output provided.  
[4 marks]

(c) State the estimated equation of regression between Number Sold and Advertising Expenditure.  
[2 marks]

(d) If an extra $1000 dollars is spent on advertising, how will this affect the sales?  
[1 mark]

(e) State and check the assumptions of the linear regression model using the output provided.  
[8 marks]

(f) What will average sales be for a month in which no money is spent on advertising?  
[1 mark]

Solution.  
(a) Number Sold = β₀ + β₁ Advertising Expenditure + residual, where β₀ is the intercept, β₁ is the slope and residual is the random variation term.

(b) The relevant hypotheses are

\[ H₀ : β₁ = 0 \quad \text{vs} \quad H₁ : β₁ > 0. \]

The p-value for this test is \( \frac{1}{2} \times 3.559 \times 10^{-7} = 1.779 \times 10^{-7} \) < 0.025, so there is strong evidence to reject the null hypothesis at the 2.5% level of significance. We conclude that the data provides conclusive evidence that there is a strong positive linear relationship between Advertising Expenditure and Number Sold.

(c) Number Sold = -32655.04 + 1.75 Advertising Expenditure.

(d) The Number Sold will increase by \( 1.745 \times 1000 = 1745 \).

(e) (i) A linear relationship is appropriate. The scatterplot of the data shows that a linear relationship is appropriate. Further, the plot of residuals against fitted values and residuals against Advertising Expenditure are pattern-less, confirming that a linear model is appropriate.

(ii) The residuals are normal. The histogram is inconclusive, as there are not enough data points. The normal probability plot is not too different

QUESTIONS 7(e)(ii) CONTINUES OVER THE PAGE

\[ QUESTION 7(e)(ii) \]
7(e)(ii) (Continued)

from a straight line. The departure is not severe, and we conclude that the normality assumption is not violated.

(iii) **The residuals have constant variance.** The scatterplot of residuals against fitted values shows a constant spread. A similar pattern is observed in the plot of residuals against Advertising Expenditure. Thus we conclude that the constant variance assumption is satisfied.

(iv) **The residuals are independent.** The scatterplot of residuals against fitted values is fairly random; a similar pattern is observed in the plot of residuals against Advertising Expenditure. Thus we conclude that the residuals are independent.

(f) zero.