37252 Regression Analysis

Exam instructions

This exam is available during the **15:30 PM – 18:30 PM** window on 5 June. This includes: time taken to solve four questions (around 2 hours), reading time, time required to scan and upload answers to Canvas, any potential internet connectivity issues and any other issues stemming from online assessment. It is advised to start as early as possible during the exam availability window. The answers must be uploaded to Canvas before the deadline 5 June 18:30 PM AEST. **Late submission will not be accepted**.

Scan the answers using your preferred scanning program and upload a **single PDF file named *EXAM_37252_studentnumber***. Make sure you also include your name and student number on the first page of the document.

Students please note

- The exam is worth 50% of your marks for the subject
- There are four questions, each worth 20 marks
- Attempt all questions

Important Notice – Exam Conditions and Academic Integrity

In attempting this examination and submitting an answer file, candidates are undertaking that the work they submit is a result of their own unaided efforts and that they have not discussed the questions or possible answers with other persons during the examination period. Candidates who are found to have participated in any form of cooperation or collusion or any activity which could amount to academic misconduct in the answering of this examination will have their marks withdrawn and disciplinary action will be initiated on a complaint from the Examiner.

Exam answers must be submitted via Turnit. Vivas or other invigilated tasks may be used to verify student achievement of learning outcomes to ensure they have completed the work on their own and to assess their knowledge of the answers they have submitted.

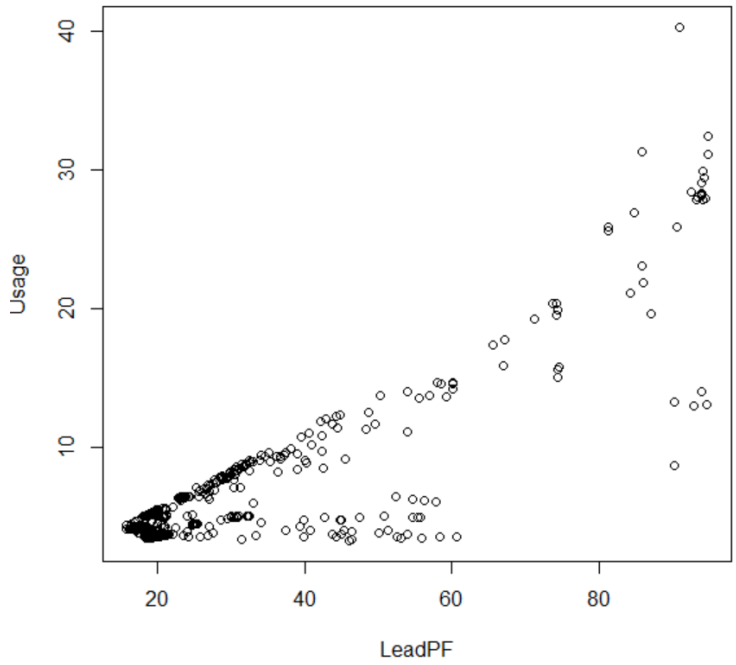
Students must not post any requests for clarification on the Discussion Boards on Canvas or Microsoft Teams. Any requests for clarification should be directed to the subject coordinator Joanna Wang on Joanna.wang@uts.edu.au. Where clarification is required it will be broadcast by email to all students in the exam group.

Question 1 [20 marks].

The steel industry machine equipment energy consumption was collected using a smart meter and additional information about the electricity consumption was collected and stored.

As first step of the analysis, we are interested in modelling industry energy consumption (`Usage`) in KWh as a function of leading current power factor (`LeadPF`) using 500 observations.

A scatter plot is reproduced below.



(a) [4 marks] Based on the scatterplot, describe the type, direction and strength of any relationship between the two variables. Is a simple linear regression model suitable for this data?

A simple linear regression model was built with `LeadPF` as the predictor variable and `Usage` as the response variable. Selected `R` output is displayed below.

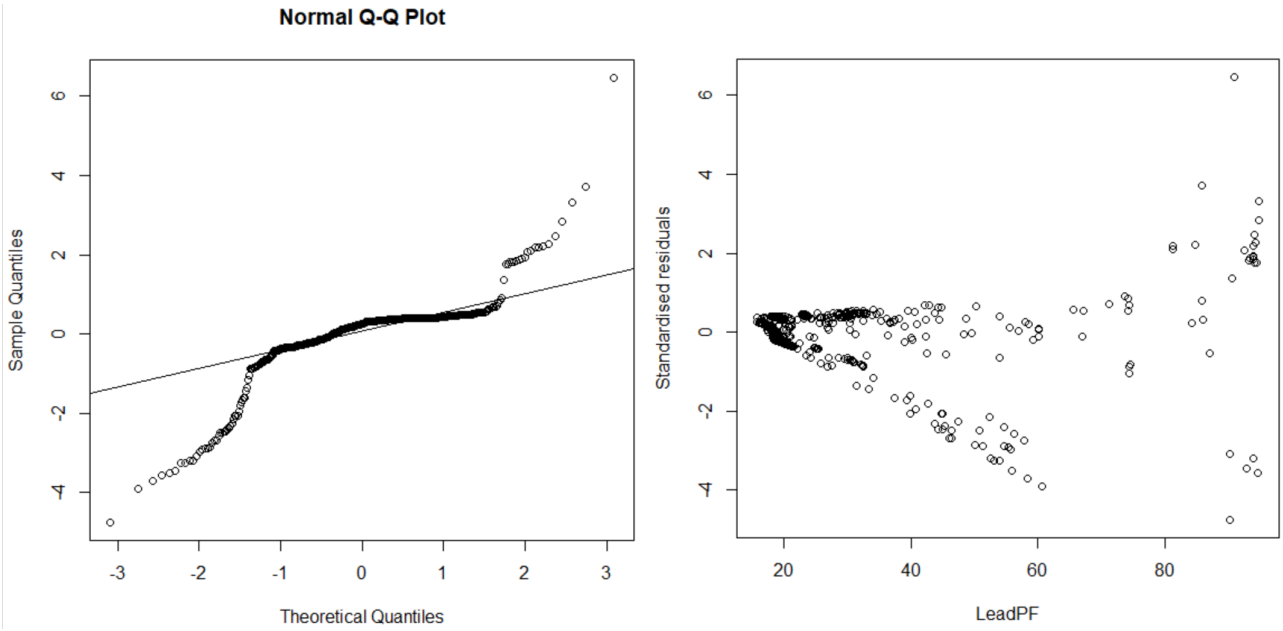
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.721280   0.237501  -3.037  0.00252 **
LeadPF       0.251786   0.006515
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.837 on 498 degrees of freedom
Multiple R-squared:   0.75,   Adjusted R-squared:  0.7495
```

(b) [3 marks] Write down the estimated regression equation and interpret the estimated coefficients.

(c) **[5 marks]** Is the regression model statistically significant at the 0.05 level (the relevant quantiles are $t_{0.90} \approx 1.28$, $t_{0.95} \approx 1.65$, $t_{0.975} \approx 1.96$, $t_{0.995} \approx 2.59$)? Write down the null and alternative hypotheses, calculate the test statistic, report the result of the chosen test and state a conclusion in plain English.

In addition to the output above, R also produced the following residual plots and statistics.




```
Durbin-Watson test

data:  mod1
DW = 0.95677, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0

> cooksD<-cooks.distance(mod1)
> round(summary(cooksD),3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   0.000   0.006   0.000   0.449
```

(d) **[2 marks]** Using the output provided, discuss whether the assumptions of the regression model are satisfied. Decide if the regression model can be relied upon.

(e) **[2 marks]** What problems did you identify in part (d) and what step could you take to rectify the problem?

(f) **[2 marks]** Using the output provided, determine if there is a potential problem with influential point(s). Provide reasons for your answer.

(g) **[2 marks]** Calculate the 90% two-sided confidence interval for the slope parameter β_1 ($t_{0.90} \approx 1.28$, $t_{0.95} \approx 1.65$, $t_{0.975} \approx 1.96$, $t_{0.995} \approx 2.59$).

Question 2 [20 marks].

We now include another predictor variable WD which equals to 1 for a weekday and 0 for weekend, in addition to the variable LeadPF used in Question 1.

A correlation analysis is shown below.

	Usage	LeadPF	WD
Usage	1.000	0.866	0.155
LeadPF	0.866	1.000	0.091
WD	0.155	0.091	1.000

A multiple linear regression model is constructed with LeadPF and WD as predictor variables and Usage as the response. Selected R output is shown below.

```
> summary(aov(mod2))
             Df Sum Sq Mean Sq F value   Pr(>F)
LeadPF        1  12025   12025 1526.87  < 2e-16 ***
WD            1     95      95   12.04 0.000567 ***
Residuals   497   3914       8

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.02310    0.25050  -4.084 5.16e-05 ***
LeadPF       0.24975    0.00647  38.601  < 2e-16 ***
WD           0.88892    0.25623   3.469 0.000567 ***
```

(a) **[3 marks]** Write down the estimated regression equation and provide interpretations of the estimated beta coefficients for LeadPF and WD.

(b) **[2 marks]** Calculate the coefficient of determination R^2 and interpret this number.

(c) **[4 marks]** Calculate a statistical quantity to assess multicollinearity. Does this quantity signal a problem of multicollinearity?

(d) **[2 marks]** Why you would use adjusted R^2 rather than R^2 to compare the fit of different models for the same dependent variable?

(e) **[6 marks]** Perform a hypothesis test at 0.05 level of significance to determine if one unit increase in LeadPF will result in more than 0.2 unit increase in Usage. Some quantiles from the relevant Student's T distribution are ($t_{0.05} \approx -1.65, t_{0.95} \approx 1.65, t_{0.975} \approx 1.96$). Write down the null and alternative hypotheses, calculate the test statistics, report the result of the test and state a conclusion in plain English.

(f) **[3 marks]** Would you include an interaction between LeadPF and WD? Explain your answer. The estimated β coefficient for the interaction between LeadPF and WD is 0.06, interpret this number.

Question 3 [20 marks]

In the last two questions, logistic regression models will be developed to predict whether a patient will survive at least one year following a heart attack, based on 131 observations.

We will start with a simple logistic regression using the variables summarised in the table below.

Name	Role	Description
alive	response	binary dummy for survival ("0" means "No", "1" means "Yes")
age	predictor	Age at heart attack is recoded into quartiles (ageBin): $ageBin = 1$ being the lowest age category and $ageBin = 4$ being the highest age category

Crosstab information of these two variables is displayed below. Expected counts under the null hypothesis of no association are also given. In R, `Q1` represents the first quartile and so on.

```
Cell Contents
|-------------------------|
|                       N |
|              Expected N |
| Chi-square contribution |(removed)
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|

Total Observations in Table:  131
```

datQ3$ageBin	datQ3$alive 0	1	Row Total
Q1	31	4	35
	23.511	11.489	
	0.886	0.114	0.267
	0.352	0.093	
	0.237	0.031	
Q2	23	11	34
	22.840	11.160	
	0.676	0.324	0.260
	0.261	0.256	
	0.176	0.084	
Q3	17	12	29
	19.481	9.519	
	0.586	0.414	0.221
	0.193	0.279	
	0.130	0.092	

Q4	17	16	33
	22.168	10.832	
	0.515	0.485	0.252
	0.193	0.372	
	0.130	0.122	
Column Total	88	43	131
	0.672	0.328	

(a) [4 marks] Calculate the odds of survive at least one year for $ageBin = 1$ and $ageBin = 4$. Hence calculate the odds ratio with $ageBin = 1$ as the reference category and interpret this quantity.

Below is a table of some quantiles from the relevant Chi-square distribution.

$\chi^2_{0.005}$	$\chi^2_{0.01}$	$\chi^2_{0.025}$	$\chi^2_{0.05}$	$\chi^2_{0.1}$	$\chi^2_{0.9}$	$\chi^2_{0.95}$	$\chi^2_{0.975}$	$\chi^2_{0.99}$	$\chi^2_{0.995}$
0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34	12.84

(b) [5 marks] Using a Chi-square test of independence at 0.05 significance level, determine if there is a statistically significant relationship between survive at least one year following a heart attack and age of an individual. Write down the null and alternative hypotheses, calculate the test statistic from the cross tabulation, report the result of the test giving a reason for you answer and state your conclusion in plain English.

A simple logistic regression model was constructed with $alive$ as response and $ageBin$ as predictor. Some R output is shown below.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.0477     0.5313  -3.854 0.000116 ***
ageBinQ2      1.3101     0.6455   2.030 0.042391 *
ageBinQ3      1.6994     0.6515   2.609 0.009092 **
ageBinQ4      1.9871     0.6353   3.128 0.001761 **

Analysis of Deviance Table

Model 1: alive ~ 1
Model 2: alive ~ ageBin
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       130     165.83
2       127     152.74  3   13.091 0.004444 **
```

(c) [3 mark] Write down the estimated logistic regression model in

- **i.** log-odds scale
- **ii.** odds scale
- **iii.** probability scale

(d) [3 marks] Interpret the estimated coefficient for $ageBin = 4$ and hence show that the binary logistic regression model gives estimated values of the odds ratio closely matching those calculated from part (a) of this question.

(e) [2 marks] Does the output from the logistic regression model support a significant relationship between survive at least one year following a heart attack and age quartiles? Justify your answer.

(f) [2 marks] Determine if the model predicts survive at least one year after a heart attack for $ageBin = 2$.

(g) [1 mark] What type of variable is $ageBin$?

Question 4 [20 marks]

The model in Question 3 is extended with inclusion of an additional continuous predictor *lvdd,* which is the left ventricular end-diastolic dimension. This is a measure of the size of the heart at end diastole.

A multiple logistic regression is constructed with some R output shown below.

```
mod4 <- glm(alive ~ ageBin + lvdd, family = "binomial", data = datQ3)


Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.4871            -2.752  0.00593 **
ageBinQ2      1.2803     0.6488   1.973  0.04846 *
ageBinQ3      1.6562     0.6551   2.528  0.01146 *
ageBinQ4      1.9368     0.6381   3.035  0.00240 **
lvdd          0.3081     0.2419   1.274  0.20266

Hosmer and Lemeshow goodness of fit (GOF) test

data:  datQ3$alive, fitted(mod4)
X-squared = 9.3696, df = 8, p-value = 0.3121


> anova(mod_null, mod4, test = "LRT")
Analysis of Deviance Table

Model 1: alive ~ 1
Model 2: alive ~ ageBin + lvdd
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       130     165.83
2       126     151.09  4   14.738 0.005276 **

> mod_lvdd <- glm(alive ~ lvdd, family = "binomial", data = datQ3)
> anova(mod_lvdd, mod4, test = "LRT")
Analysis of Deviance Table

Model 1: alive ~ lvdd
Model 2: alive ~ ageBin + lvdd
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       129     163.27
2       126     151.09  3   12.177   0.0068 **


> exp(confint.default(mod4))
                 2.5 %     97.5 %
(Intercept) 0.002552488  0.3665982
ageBinQ2    1.008692095 12.8308565
ageBinQ3
ageBinQ4    1.986177835 24.2265826
lvdd        0.847127028  2.1861890
```

(a) **[2 marks]** Write down the estimated regression model in odds scale for $ageBin = 2$ and provide an interpretation of the estimated beta coefficient for the binary dummy of $ageBinQ3$ on the log-odds scale.

(b) **[4 marks]** Using 0.05 significance level, assess the adequacy of this model. Write down the null and alternative hypotheses, the test statistic and p-value, the test decision and state a conclusion in plain English.

(c) **[4 marks]** Using 0.05 significance level, determine if the regression model is statistically significant. Write down the null and alternative hypotheses, the test statistic and p-value, report the test result and state a conclusion in plain English. Which two models do we compare for this test?

(d) **[4 marks]** Using 0.05 significance level, test whether the true beta coefficient of $lvdd$ is 0.3. Write down the null and alternative hypotheses, test decision with reason and state a conclusion in plain English.

(e) **[2 marks]** Calculate the standard error for the estimate of the intercept term.

(f) [**2 marks**] Calculate 95% CI for exp $(\beta_{ageBinQ3})$ where $\beta_{ageBinQ3}$ is the beta coefficient of the binary dummy variable for $ageBin = 3$. Note: $Z_{0.975} = 1.96$.

(g) **[2 marks]** If you consider an interaction between $lvdd$ and $ageBin$,

i. write down the multiple logistic regression model with this interaction on the log-odds scale

ii. the estimated beta coefficient for the interaction between $ageBinQ2$ and $lvdd$ is -0.48. Interpret the exponential of this value on the odds scale.