
## FACULTY OF SCIENCE AND ENGINEERING

### DEPARTMENT OF MATHEMATICS AND STATISTICS

# MID-TERM ASSESSMENT EXAMINATION 1

MODULE TITLE: Applied Statistic for Administration DURATION OF EXAM: 45 minutes

LECTURER: Mr. Kevin O'Brien GRADING SCHEME: 15 marks

### INSTRUCTIONS TO CANDIDATES

- This exam will start at 12:05, and will last 45 minutes.
- Each question will be worth either 1 or 2 Marks. There are 15 Marks worth of questions.
- All questions must be attempted (LENS students please see below)
- Write all of your answers in the exam script. Write the script number on any other documents you submit.
- It is your responsibility to return the script to collection box. An audit of scripts will take place immediately after the exam. If your script is account for in that audit, you are deemed to be absent, and will receive no marks.
- **IMPORTANT for LENS Student:** Specifically approved LENS students have to answer any selection of questions that have an aggregate mark of 12 Marks.
  - They may skip any three of the 1-Mark Questions
  - OR - They may skip a 1-Mark Question and a 2-Mark Question
  - The mark will be rescaled by 125 %.
  - They are advised to skip questions that are indicated by an asterisk symbol ("$*$"), but it is not compulsory that they do so.

# Attempt ALL questions

Explain why an adjusted R2 value is often preferred to R2 when comparing models.

Usual assumptions are that the residual (error) terms should be independent, identically distributed, have zero mean and constant variance, and, if the usual inferences and tests are to be made, be Normally distributed.

### Question 3 Part B (6 Marks)

Suppose we have a regression model, described by the following equation

$$\hat{y} = 28.81 + 6.45x_1 + 7.82x_2$$

We are given the following pieces of information.

- The standard deviation of the response variance $y$ is 10 units.
- There are 53 observations.
- The *Coefficient of Determination* (also known as the *Multiple R-Squared*) is 0.75.

Complete the *Analysis of Variance* Table for a linear regression model. The required values are indicated by question marks.

| | DF | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| Regression | ? | ? | ? | ? | $< 2.2e^{-16}$ |
| Error | ? | ? | ? | | |
| Total | ? | ? | | | |

## Question 1 Inference Procedures

### Question 1 Part A (5 Marks)

Numeric Transformations, such as logarithmic transformation, are often used in statistical analysis as an approach for dealing with non-normal data.

(i.) (1 Mark) Describe the purpose of Tukey's Ladder (referencing direction and relative strength).

(ii.) (2 Marks) Give two examples of a transformation for various types of skewed data (i.e. an example for both types of skewness).

(iii.) (1 Mark) Discuss the limitations of numeric transformations.

### Question 1 Part B (5 Marks)

The typing speeds for one group of 12 Engineering students were recorded both at the beginning of year 1 of their studies. The results (in words per minute) are given below:

| | | | | | |
|---|---|---|---|---|---|
| 149 | 146 | 112 | 142 | 168 | 153 |
| 137 | 161 | 156 | 165 | 170 | 159 |

Use the Dixon Q-test to determine if the lowest value (118) is an outlier. You may assume a significance level of 5%.

(i.) (1 Mark) State the Null and Alternative Hypothesis for this test.

(ii.) (2 Marks) Compute the test statistic

(iii.) (1 Mark) State the appropriate critical value.

(iv.) (1 Mark) What is your conclusion to this procedure.

**Question 1 Part C (5 Marks)**

(i.) (3 Marks) Provide a brief description for three tests from the family of Grubb's Outliers Tests. Include in your description a statement of the null and alternative hypothesis for each test

(ii.) (2 Marks) Describe any required assumptions for tests, and the limitations of these tests.

**Question 1 Part D (10 Marks)**

Assume that the diameter of a critical component is normally distributed with a mean of 250mm and a standard deviation of 15mm. You are required to estimate the approximate probability of the following measurements occurring on an individual component.

(i.) (3 Mark) Greater than 245mm.

(ii.) (3 Marks) Less than 265mm.

(iii.) (4 Marks) Between 245mm and 265mm.

Use the normal tables to determine the probabilities for the above exercises. You are required to show all of your workings.

# Question 4. Linear Models (25 Marks)

## Question 4 Part A (12 Marks)

The mercury level of several tests of sea-water from costal areas was determined by atomic-absorption spectrometry. The results obtained are as follows

The analysis of the relationship between concentration and absorbance is obtained in R and presented below.

```
x<-seq(0,100,by=10)
y<- c(0.321, 0.834, 1.254, 1.773, 2.237, 2.741, 3.196, 3.678,
4.217, 4.774, 5.261)
model<- lm(y~x)
summary(model)

Call:
lm(formula = y ~ x)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.2933636  0.0234754   12.50 5.45e-07
x           0.0491982  0.0003968  123.98 7.34e-16
---

Residual standard error: 0.04162 on 9 degrees of freedom
Multiple R-squared: 0.9994,     Adjusted R-squared: 0.9993
F-statistic: 1.537e+04 on 1 and 9 DF,  p-value: 7.337e-16

confint(model)
2.5 %     97.5 %
(Intercept) 0.24025851 0.34646876
x           0.04830054 0.05009582
```

(i) (2 marks) Determine and interpret the slope and the intercept of the regression line.

(ii) (2 marks) State the 95% confidence interval for the slope and the intercept coefficients. Interpret this intervals with respect to any relevant hypothesis tests

(iii) (2 marks) Explain in which way is the prediction intervals different from the confidence intervals for fitted values in linear regression?

(iv) (2 Marks) The following piece of `R` code gives us a statistical metric. What is this metric? What is it used for? How should it be interpreted.

```
> AIC(model)
[1] -34.93389
```

**Question 4 Part B (12 Marks)**

Given the AIC for each candidate model, use ***Backward Selection*** to determine the optimal model for predicting values of $y$ with predictor variables $x_1$, $x_2$,$x_3$ and $x_4$.

Suppose we have 5 predictor variables. Use **Forward Selection** and **Backward Selection** to choose the optimal set of predictor variables, based on the AIC measure.

| Variables | AIC | Variables | AIC |
|---|---|---|---|
| $\emptyset$ | 200 | x1, x2, x3 | 74 |
| | | x1, x2, x4 | 75 |
| x1 | 150 | x1, x2, x5 | 79 |
| x2 | 145 | x1, x3, x4 | 72 |
| x3 | 135 | x1, x3, x5 | 85 |
| x4 | 136 | x1, x4, x5 | 95 |
| x5 | 139 | x2, x3, x4 | 83 |
| | | x2, x3, x5 | 82 |
| x1, x2 | 97 | x2, x4, x5 | 78 |
| x1, x3 | 81 | x3, x4, x5 | 85 |
| x1, x4 | 94 | | |
| x1, x5 | 88 | x1, x2, x3, x4 | 93 |
| x2, x3 | 87 | x1, x2, x3, x5 | 120 |
| x2, x4 | 108 | x1, x2, x4, x5 | 104 |
| x2, x5 | 87 | x1, x3, x4, x5 | 101 |
| x3, x4 | 105 | x2, x3, x4, x5 | 89 |
| x3, x5 | 82 | | |
| x4, x5 | 86 | x1, x2, x3, x4, x5 | 100 |

**Question 3 Part B (6 Marks)**

Suppose we have a regression model, described by the following equation

$$\hat{y} = 28.81 + 6.45x_1 + 7.82x_2$$

We are given the following pieces of information.

- The standard deviation of the response variance $y$ is 10 units.
- There are 53 observations.
- The *Coefficient of Determination* (also known as the *Multiple R-Squared*) is 0.75.

Complete the *Analysis of Variance* Table for a linear regression model. The required values are indicated by question marks.

| | DF | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| Regression | ? | ? | ? | ? | $< 2.2e^{-16}$ |
| Error | ? | ? | ? | | |
| Total | ? | ? | | | |

# Polynomial Regression

- Polynomial regression models are useful in situations where the analyst knows that ***curvilinear effects*** are present in the response variable.
- Polynomial models are also useful as approximating functions to unknown and possible very complex nonlinear relationship.

In the context of Statistical Modelling, what is means the "Law of Parsimony".

What is the Akaike information criterion? How would you use it in statical modelling? How does it differ from other metrics such as the coefficient of determination. How would you interpret an AIC value.

compare and contrast forward selection and backward selection as variable selection procedures.

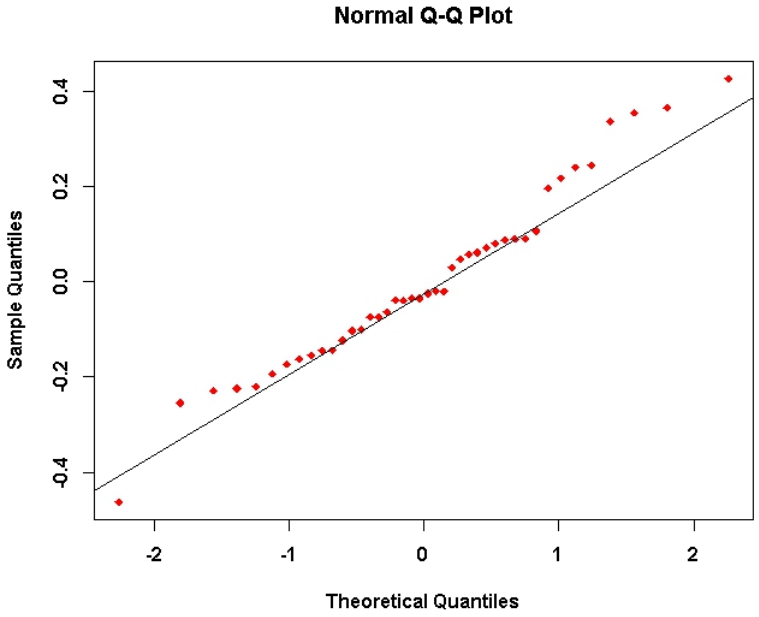
In the context of Statistical Modelling,What is meant by Stepwise regression?

Explain why an adjusted R2 value is often preferred to R2 when comparing models.

Usual assumptions are that the residual (error) terms should be independent, identically distributed, have zero mean and constant variance, and, if the usual inferences and tests are to be made, be Normally distributed.

## Q2. Testing Normality (3 Marks)

A graphical procedure was carried out to assess whether or not this assumption of normality is valid for data set `Y`. Consider the Q-Q plot in the figure below.




i. (1 Mark) Provide a brief description on how to interpret this plot.

ii. (1 Mark) What is your conclusion for this procedure? Justify your answer.

## Q3. Testing Normality (4 Marks)

Consider the following inference procedure performed on data set $X$.

```
> shapiro.test(X)

Shapiro-Wilk normality test

data:  X
W = 0.8914, p-value = 0.07047
```

i. (1 Mark) Describe what is the purpose of this procedure.

ii. (1 Mark) What is the null and alternative hypothesis?

iii. (1 Mark) Write the conclusion that follows from it.

iv. (1 Mark) Tests for Normality are known to be susceptible to low power. Discuss what is meant by this.

## Q4. Dixon Q Test For Outliers (4 Marks)

The typing speeds for one group of 12 Engineering students were recorded both at the beginning of year 1 of their studies. The results (in words per minute) are given below:

| | | | | | | |
|---|---|---|---|---|---|---|
| 121 | 146 | 150 | 149 | 142 | 170 | 153 |
| 137 | 161 | 156 | 165 | 137 | 178 | 159 |

Use the Dixon Q-test to determine if the lowest value (121) is an outlier. You may assume a significance level of 5%.

i. (1 Mark) Formally state the null hypothesis and the alternative hypothesis.

ii. (1 Mark) Compute the Test Statistic.

iii. (2 Mark) By comparing the Test Statistic to the appropriate Critical Value, state your conclusion for this test.

## Q1. Dixon Q Test For Outliers (4 Marks)

The typing speeds for one group of 12 Engineering students were recorded both at the beginning of year 1 of their studies. The results (in words per minute) are given below:

| | | | | | |
|---|---|---|---|---|---|
| 118 | 146 | 149 | 142 | 170 | 153 |
| 137 | 161 | 156 | 165 | 178 | 159 |

Use the Dixon Q-test to determine if the lowest value (118) is an outlier. You may assume a significance level of 5%.

i. (1 Mark) State the Null and Alternative Hypothesis for this test.

ii. (1 Marks) Compute the test statistic

iii. (1 Mark) State the appropriate critical value.

iv. (1 Mark) What is your conclusion to this procedure

# Formulae and Tables

## Critical Values for Dixon Q Test

| N | $\alpha = 0.10$ | $\alpha = 0.05$ | $\alpha = 0.01$ |
|---|---|---|---|
| 3 | 0.941 | 0.970 | 0.994 |
| 4 | 0.765 | 0.829 | 0.926 |
| 5 | 0.642 | 0.710 | 0.821 |
| 6 | 0.560 | 0.625 | 0.740 |
| 7 | 0.507 | 0.568 | 0.680 |
| 8 | 0.468 | 0.526 | 0.634 |
| 9 | 0.437 | 0.493 | 0.598 |
| 10 | 0.412 | 0.466 | 0.568 |
| 11 | 0.392 | 0.444 | 0.542 |
| 12 | 0.376 | 0.426 | 0.522 |
| 13 | 0.361 | 0.410 | 0.503 |
| 14 | 0.349 | 0.396 | 0.488 |
| 15 | 0.338 | 0.384 | 0.475 |
| 16 | 0.329 | 0.374 | 0.463 |

## Critical Values for Chi Square Test