Australian National University

Venue _____

Student Number |__|__|__|__|__|__|__|__|__|

Research School of Finance, Actuarial Studies & Statistics

EXAMINATION SAMPLE QUESTIONS

Semester 2 - End of Semester, 2017

STAT3008/7001 Applied Statistics

Examination Duration: 180 minutes
Reading Time: 15 minutes

Exam Conditions:
Central Examination
Students must return the examination paper at the end of the examination
This examination paper is NOT available to the ANU Library archives

Materials Permitted in The Exam Venue:
(No electronic aids are permitted e.g. laptops, phones)
Calculator (non-programmable)
Unannotated paper-based dictionary (no approval required)
Two A4 pages with notes on both sides.

Materials to Be Supplied to Students:
1 x 20 page plain
Scribble Paper x2

Instructions to Students:

These exam sample questions are revised based on past exam questions, the quiz, Assignment 1 and some tutorial problems. Note that the final exam will cover the material from Week 1 to Week 12. However, these sample questions only cover a very small part. The purpose of these sample questions is to show what the final exam questions may look like.

Maximum points: 100. Please exactly follow the instructions given for each question. Note that you do not need to copy the question themselves in the answer book and please do not write down any irrelevant results. Please mark the number of the question that you are attempting in the answer book very clearly, such that the instructor and tutors can understand which question you are answering.

The significance level is set to be 0.05 throughout the exam.

The following R code and corresponding output may be required in order to answer the questions that follow:

```
> round(qnorm(0.975),4)
[1] 1.96
> round(qnorm(0.95),4)
[1] 1.6449
> round(qt(0.975,493),4)
[1] 1.9648
> round(qt(0.95,493),4)
[1] 1.648
> round(qt(0.975,494),4)
[1] 1.9648
> round(qt(0.95,494),4)
[1] 1.6479
> round(qt(0.975,495),4)
[1] 1.9648
> round(qt(0.95,495),4)
[1] 1.6479
```

where for example, the R function "round(qnorm(0.975),4)" rounds the value in its first argument "qnorm(0.975)" to 4 decimal places.

Question 1

Answer each question "TRUE" or "FALSE". In each case, write the whole word. It is **not** acceptable to write only "T" or "F", and answers presented in this form **will be graded incorrect**.

a) The estimated mean response for the regression $\mu(Y \mid X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \times X_2)$ corresponding to a particular set of explanatory variable values $X_1 = 3, X_2 = 2$ is 15. Based on this information we would estimate that there is more than a 50% chance that the response, given $X_1 = 3, X_2 = 2$, would take on a value greater than 17.

FALSE

b) Modelling marital status (Single, Married, Divorced) as a categorical explanatory variable in a Poisson log-linear regression model will require three parameters to be estimated, not including the intercept, the other variables and interactions included in the model.

FALSE

c) A fitted linear regression model with 5 explanatory variables based on 500 observations returns the least squares estimation $b_5 = 0.21, \mathrm{SE}(b_5) = 0.06.$ You are given two 90% confidence intervals for β_5 (a) (0.11,0.31) and (b) (0.13,0.42) that have been computed based on the fitted regression model. One of the intervals was computed by using the confidence interval formula for multiple linear regression model, the other was computed based on the bootstrap method. We can make a conclusion that interval (b) (0.13,0.42) is the confidence interval computed by using the confidence interval formula for multiple linear regression model with small rounding errors.

FALSE

d) Removing missing values from a dataset before proceeding with an analysis may lead to biased results.

TRUE – for example what if the missing values all corresponded to observations with a certain characteristic such as a small response value.

e) Including a continuous explanatory variable (X_1) in a regression model as a categorical variable with three categories would be a sensible modelling decision.

FALSE

f) The multiple linear regression $\mu(Y/X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$ will have a larger R^2 than the multiple linear regression $\mu(Y/X) = \beta_0 + \beta_1 X_4 + \beta_2 X_5$.

FALSE

g) A multiple linear regression model should only include explanatory variables that have a normal distribution.

FALSE – for example indicator variables are clearly non-normal.

h) Eliminating an explanatory variable to a multiple linear regression model cannot increase the significance, as measured by the t-test, for another explanatory variable that is in the model.

FALSE

Question 2

This question involves a 1992 dataset published by Forbes magazine. The dataset contains information on the wealth, age, and geographic region (Asia, Europe, Middle East, United States, and other) for 225 billionaires. To investigate the relationship between wealth, age and geographic region (a categorical variable with five categories) a linear regression model relating 1/wealth to age and geographic region was fitted. Information regarding the fitted regression models is provided below. In fitting the linear regression model "Middle East" was taken as the baseline value for geographic region. This data was taken from *"The Data and Story Library"* http://lib.stat.cmu.edu/DASL/DataArchive.html

```
> summary(wealth1.reg)

Call: lm(formula = (1/wealth) ~ age + as.factor(region) + age * as.factor(region))

Residuals:

     Min      1Q   Median     3Q    Max

 -0.6334 -0.2126 0.008569 0.1881 0.4843


Coefficients:

                        Value Std. Error t value Pr(>|t|)

          (Intercept)  0.5989  0.0948     6.3163  0.0000
```

```
                  age -0.0008  0.0015    -0.5241  0.6008
   as.factor(region)1  0.0938  0.1534     0.6111  0.5417
   as.factor(region)2  0.1804  0.0820     2.1998  0.0289
   as.factor(region)3 -0.0523  0.0688    -0.7613  0.4473
   as.factor(region)4 -0.0333  0.0422    -0.7890  0.4310
ageas.factor(region)1 -0.0009  0.0024    -0.3715  0.7107
ageas.factor(region)2 -0.0028  0.0012    -2.2396  0.0261
ageas.factor(region)3  0.0009  0.0011     0.8168  0.4150
ageas.factor(region)4  0.0005  0.0006     0.7974  0.4261

Residual standard error: 0.2568 on 215 degrees of freedom
Multiple R-Squared: 0.04388
F-statistic: 1.096 on 9 and 215 degrees of freedom, the p-value is 0.3667



> summary(wealth2.reg)

Call: lm(formula = (1/wealth) ~ age + as.factor(region))
Residuals:
     Min      1Q   Median  3Q   Max
 -0.5388 -0.2211 0.008263 0.2 0.498

Coefficients:
                     Value Std. Error t value Pr(>|t|)
       (Intercept)  0.6167  0.0840     7.3435  0.0000
               age -0.0010  0.0013    -0.8177  0.4144
as.factor(region)1  0.0378  0.0259     1.4584  0.1462
as.factor(region)2  0.0021  0.0203     0.1023  0.9186
as.factor(region)3  0.0029  0.0137     0.2112  0.8329
as.factor(region)4 -0.0001  0.0080    -0.0147  0.9883

Residual standard error: 0.2585 on 219 degrees of freedom
Multiple R-Squared: 0.01276
F-statistic: 0.5663 on 5 and 219 degrees of freedom, the p-value is 0.7258
```

a) Draw a clearly labelled plot depicting the regression lines corresponding to Asia and the United States. You must clearly note the intercepts and slopes on the plot that you produce. Note: Asia corresponds to *as.factor(region)1* and United States to *as.factor(region)2.*

Similar to page 23 in Lecture Notes 5. However, you can only draw by hand but not R in the exam.

b) Based on *wealth1.reg*, can we conclude that the intercepts of the regression lines corresponding to a billionaire located in the United States and the Middle East are not the same?

Yes. The test-statistic and p-value for this test are:
2.2 and 0.03

c) Based on *wealth2.reg*, what would you predict the wealth of a 50-year-old billionaire located in the United States to be?

$$0.6167 - 50 * 0.001 + 0.0021$$

d) If possible, perform a hypothesis test to determine whether the interaction terms between age and region are important variables. It is sufficient to report the test-statistic, what distribution it should be compared to, and its degrees of freedom.

The test statistic is ((0.2585^2)*219-(0.2568^2)*215)/4/0.2568^2=1.73. The distribution that it should be compared to is F distribution, with numbers of degrees of freedom 4 and 215.

Question 3

Question 2 in Tutorial 9.

Question 4

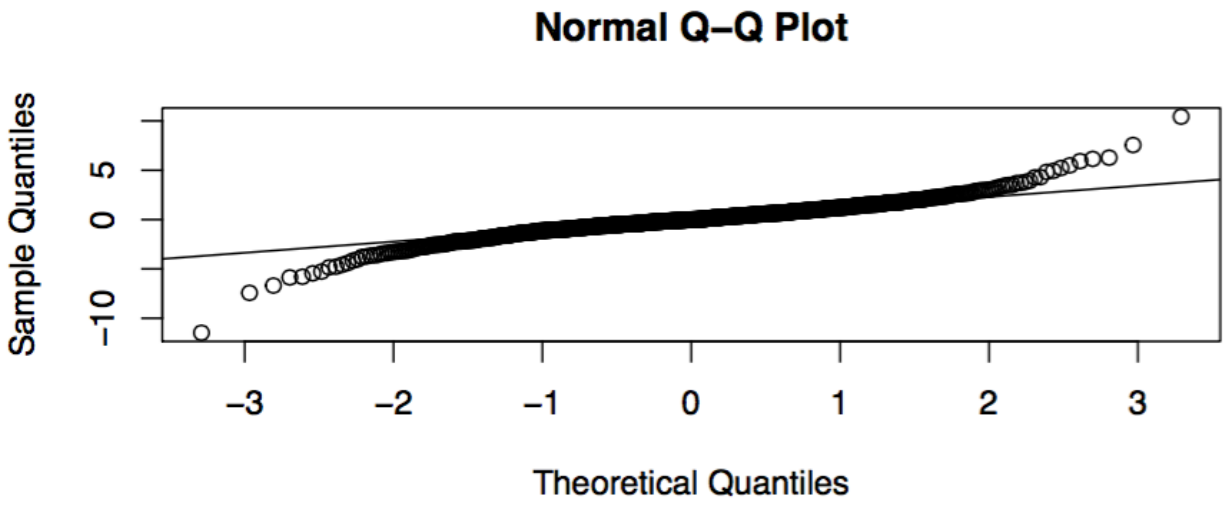
The following are the R codes for a simulation experiment:

```
X=1:100
n=length(X)
Y=rep(0,n)
numsamp=1000
CIl=rep(0,numsamp)
CIu=rep(0,numsamp)
Xnew=data.frame(X=2.5)
set.seed(1)
for(i in 1:numsamp) {
  errors=rnorm(n)
  Y=2+1*X+errors
  SLRfit=lm(Y~X)
  CI=predict(SLRfit,Xnew,interval='confidence',level=0.95)
  CIl[i]=CI[2]
  CIu[i]=CI[3]
}
MeanResponse=2+1*Xnew$X
Count=ifelse(CIl<=MeanResponse & CIu>=MeanResponse, 1,0)
sum(Count)
```

a) Please guess the R output of "sum(Count)". Please give the reason for your guess.

Answer: 950. Because the 95% confidence interval for the mean of response means that if we use the formula in lectures to obtain 1000 confidence intervals for 1000 repeated samples, then around 950 confidence intervals will cover the mean of response.

b) What conclusion can you obtain based on the following Q-Q plot?



The plot just shows that the observations have heavy tails.

Question 5: All quiz questions.

END OF EXAMINATION