OLLSCOIL TEICNEOLAÍOCHTA BHAILE ÁTHA CLIATH
TECHNOLOGICAL UNIVERSITY DUBLIN

TECHNOLOGICAL UNIVERSITY DUBLIN – CITY CAMPUS

School of Mathematical Sciences

TU248 PG Cert Applied Statistics

Year 1

SUMMER EXAMINATIONS 2020/2021

MATH 9903: REGRESSION MODELS

DR J CONDON

DR C HILLS

DR P MURPHY

Answer three questions

All questions carry equal marks

1. A group of environmental scientists study the relationship between the average daily concentration of tropospheric ozone (μg/m^3) at ten locations in a city and the average wind speed (m/sec) and temperature (°C) at those locations. The data are held in an R data frame called `ozone`, a portion of which is shown below.

```
> head(ozone)
  ozone speed temp
1    97     6   32
2   105     8   32
3   106     9   32
4   102     7   37
5   104     8   32
6    96     6   32
```

The following analysis is performed on these data.

```
> summary(fit)
...

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  80.4603    16.1856   4.971 0.001617
speed         3.4727     0.5905   5.881 0.000611
temp         -0.1359     0.4547  -0.299 0.773744
---

Residual standard error: 2.376 on 7 degrees of freedom
Multiple R-squared:  0.8407,  Adjusted R-squared:  0.7952
F-statistic: 18.47 on 2 and 7 DF,  p-value: 0.001613
```

a) Outline the logic of using analysis of variance (ANOVA) to test the standard global null hypothesis after fitting a multiple regression model. Explain the basis of the comparison of the mean squares for error and the mean squares for model under such a null hypothesis. (6)

b) Give the R code for fitting the model shown. (4)

c) Give the complete ANOVA table for the model fitted. State the default null and alternative hypotheses for this table and state your conclusions. (6)

d) Calculate the estimated ozone level for a day with the following environmental conditions: speed=11, temp=29. (5)

question continues overleaf...

e) Give the R code that would be used to calculate a 95% confidence interval for the estimate calculated in part **d)**.

(5)

f) Discuss the evidence from the output given that temperature is related to ozone level. Explain your conclusion by specifying the relevant null and alternative hypotheses, the test statistic and associated p-value.

(7)

[33]

2. A study was conducted to investigate the relationship between the performance of elderly subjects on a memory task and certain risk factors for Alzheimer's disease. The dataset consisted of the following variables:

score	The score obtained on a standardised memory task (lower scores indicated increasing memory impairment).
age	The age in years of the subject.
status	A 'health status' variable coded as follows: `z1Gene`: The subject has the z1 gene which it is believed may be related to Alzheimer's disease. `Prior_Bypass`: The subject does not have the z1 gene, but has undergone a prior coronary by-pass operation. `normal`: The subject has neither the z1 gene nor has undergone a prior coronary by-pass operation.

The R data frame holding these data was called `memory_data` and is partially displayed below.

```
> head(memory_data)
  score age        status
1   9.3  75        z1Gene
2   4.9  83        z1Gene
3  11.1  79        z1Gene
4   6.4  81        z1Gene
5  10.1  71 Prior_Bypass
6   9.4  82 Prior_Bypass
```

a) Describe how dummy variables are used in multiple regression to model categorical predictors. (6)

b) A model is fitted to these data with the following R code:

```
fit1=lm(score~age+factor(status),data=memory_data)
```

Describe in detail the model being fitted by this code, including the assumed geometric relationship between the predictors and the response. (9)

question continues overleaf...

c) A further analysis (`fit2`) are carried out on these data. The code and edited output from this analysis are shown here:

```
> fit2=update(fit1,.~.+age:factor(status),data=memory_data)
> drop1(fit2,test='F')

Model:
...
                    Df Sum of Sq    RSS    AIC F value Pr(>F)
age:factor(status)  2    5.9108 19.857 13.507  1.4834 0.2903
> summary(fit2)

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      29.8325    10.1895   2.928   0.0221
age                              -0.2444     0.1363  -1.793   0.1161
factor(status)Prior_Bypass      -13.5634    16.7184  -0.811   0.4439
factor(status)z1Gene             24.9975    21.5428   1.160   0.2839
age:factor(status)Prior_Bypass    0.1589     0.2181   0.729   0.4898
age:factor(status)z1Gene         -0.3456     0.2748  -1.258   0.2488
---
```

i) For the `fit2` model, predict the score for a subject with status '`normal`' and an age of 82 years. (4)

ii) For the `fit2` model, predict the score for a subject with status '`z1Gene`' and an age of 77 years. (4)

iii) Explain the difference between the `fit1` and `fit2` models. Discuss the evidence for/against the `fit2` model. (10)

[33]

3. a) Define what is meant by underfitting and overfitting in the context of multiple regression and outline the consequences of both for model fitting and statistical inference. (8)

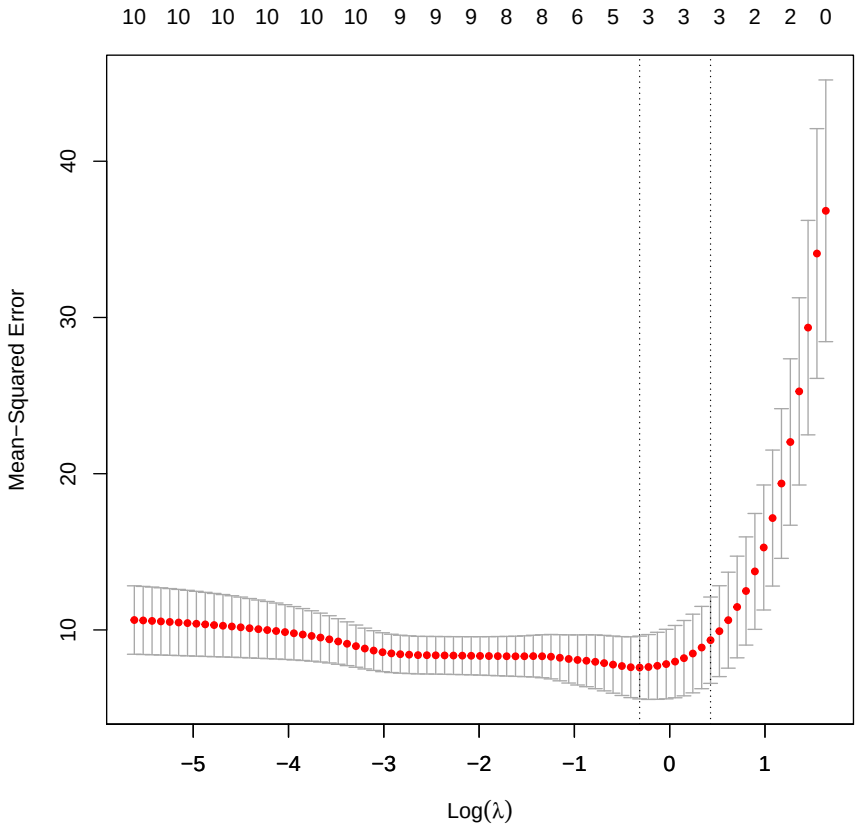
b) The LASSO algorithm is used to fit data where the response variable is the fuel consumption (miles per gallon 'mpg') of 32 different car models. There are 10 predictors which are car engine design characteristics (e.g. weight, horsepower, Number of cylinders etc.). The response is in a vector called `mpg` and the 10 predictors are in a matrix called `preds`.

The glmnet package in R is used and the following code produced the plot shown below.

```
cv=cv.glmnet(preds,mpg,family="gaussian",alpha=1)
plot(cv)
```

i) Describe how the LASSO algorithm penalises model complexity when used in fitting a multiple regression model. (9)

ii) Briefly outline the steps of the cross-validation algorithm used to produce the plot below and describe what this `cv` plot is displaying with regard to an optimal value of the LASSO shrinkage parameter. (7)



iii) From plot, give an approximate value for the λ satisfying the "1-SE" rule and state how many predictors will have non-zero coefficients if this value of λ is used as the shrinkage parameter. (5)

iv) The `cv` plot shown above, displays increasing trends in mean-squared error on both the left and right sides of the plot window. Discuss possible explanations for these trends. (4)

[33]

4. Customer 'churn' is sometimes defined as the loss of a customer to a competitor. A medical device manufacturer is conducting an analysis of their customer churn using historic data. All customer accounts are examined and any without a purchase within the last 6 months is defined to be a 'churn'. They analyse the data using **logistic regression**. The response being modelled is the probability that a customer is a 'churn', with the variable `churn` coded as 1 if the customer has churned, and coded 0 otherwise. The following predictors:

product_range:	Which one of two product ranges that customer had purchased over the preceding 2 years, coded 'a' or 'b'.
country:	Home country of the customer. One of: Ireland, U.K., Canada, U.S.
years:	No. of years since first becoming a customer.

A portion of the analysis and output from R is given in box below.

```
> fit=glm(churn~factor(product_range)+factor(country)+years,family=
   binomial,data=churn)
> summary(fit)
Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)               4.0104     1.7622   2.276  0.02286
factor(product_range)b   -0.3080     0.6570  -0.469  0.63919
factor(country)Ireland   -2.6244     1.1328  -2.317  0.02051
factor(country)UK        -2.8205     1.1557  -2.441  0.01466
factor(country)US        -1.2112     0.7667  -1.580  0.11416
years                    -0.5175     0.2000  -2.588  0.00966
```

a) Give the general formulation of the logistic regression model for binary data, explaining the terms used and making explicit reference to this particular example. (5)

b) Find the estimated odds ratio that an Irish customer will churn over a U.K. customer - all other variables being equal. (5)

c) Discuss the evidence for the predictor 'product_range' being related to the response based on the output given. (7)

d) Predict the probability that a customer with the following values of the predictors are a churn: product_range=a, country=U.S., years=5. (8)

e) Describe how a 95% confidence interval for the fitted probability predicted in part **d)** could be calculated using R code. (8)

[33]