Student
Number

Semester 1 Assessment, 2022

School of Mathematics and Statistics

MAST90139 Statistical Modelling for Data Science

This exam consists of 23 pages (including this page)

Authorised materials: printed one-sided copy of the Exam or the Masked Exam made available earlier; any hard-copy or offline e-copy of notes, or offline electronic PDF reader; a Casio FX82 calculator; and blank A4 paper.

Instructions to Students

- During exam writing time you may only interact with the device running the Zoom session with supervisor permission. The screen of any other device must be visible in Zoom from the start of the session.
- If you have a printer, print out the exam single-sided and hand write your solutions into the answer spaces.
- If you do not have a printer, or if your printer fails on the day of the exam,

  (a) download the exam paper to a second device (not running Zoom), disconnect it from the internet as soon as the paper is downloaded and read the paper on the second device;

  (b) write your answers on the Masked Exam PDF if you were able to print it single-sided before the exam day.

  If you do not have the Masked Exam PDF, write single-sided on blank sheets of paper.
- If you are unable to answer the whole question in the answer space provided then you can append additional handwritten solutions to the end of your exam submission. If you do this you MUST make a note in the correct answer space or page for the question, warning the marker that you have appended additional remarks at the end.
- Assemble all the exam pages (or template pages) in correct page number order and the correct way up, and add any extra pages with additional working at the end.
- Scan your exam submission to a single PDF file with a mobile phone or a scanner. Scan from directly above to avoid any excessive keystone effect. Check that all pages are clearly readable and cropped to the A4 borders of the original page. Poorly scanned submissions may be impossible to mark.
- Upload the PDF file via the Canvas Assignments menu and submit the PDF to the GradeScope tool by first selecting your PDF file and then clicking on Upload PDF.
- Confirm with your Zoom supervisor that you have GradeScope confirmation of submission before leaving Zoom supervision.
- You should attempt all questions.
- There are 8 questions with marks as shown. The total number of marks available is 110.

Question 1 (14 marks)

In a clinical study of diabetes, each of 336 female patients was given a test to see whether the patient showed signs of diabetes, with result given in `test` (coded 0 if negative, and 1 if positive). Measurements of variables `glucose`, `bmi`, `pedigree` and `age` were also collected from these patients, where `glucose=` "plasma glucose concentration at 2 hours in an oral glucose tolerance test", `bmi=` "body mass index [weight in kg/(height in metres squared)]", `pedigree=` "diabetes pedigree function", and `age=` "age(years)".
The main purpose of the study was to analyse the relationship between `test` and the other four variables. The data have been stored in the data frame `Q1.dat` with the following information:

```
> head(Q1.dat)

  glucose bmi pedigree age test
1      89  28     0.17  21    0
2      78  31     0.25  26    1
3     197  30     0.16  53    1
4     189  30     0.40  59    1
5     166  26     0.59  51    1
6     103  43     0.18  33    0
```

```
> summary(Q1.dat)

    glucose         bmi           pedigree          age             test
 Min.   : 56   Min.   :18.2   Min.   :0.085   Min.   :21.0   Min.   :0.00
 1st Qu.: 99   1st Qu.:27.8   1st Qu.:0.268   1st Qu.:24.0   1st Qu.:0.00
 Median :119   Median :32.8   Median :0.446   Median :28.0   Median :0.00
 Mean   :122   Mean   :32.3   Mean   :0.519   Mean   :31.8   Mean   :0.33
 3rd Qu.:144   3rd Qu.:36.2   3rd Qu.:0.688   3rd Qu.:38.0   3rd Qu.:1.00
 Max.   :197   Max.   :57.3   Max.   :2.329   Max.   :81.0   Max.   :1.00
```

Answer the following questions in the provided boxes or on plain A4 paper:

(a) The following is R output from fitting a logistic regression model.

```
> Q1mod1 <- glm(test ~ glucose + bmi + pedigree + age, family = binomial,data = Q1.dat)
> summary(Q1mod1)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.81047    1.25381   -8.62  < 2e-16
glucose       0.03639    0.00549    6.62  3.5e-11
bmi           0.08916    0.02430    3.67  0.00024
pedigree      1.05588    0.46598    2.27  0.02346
age           0.05940    0.01451    4.09  4.3e-05

    Null deviance: 426.34  on 335  degrees of freedom
Residual deviance: 291.12  on 331  degrees of freedom
AIC: 301.1
```

(i) What can be said about the adequacy of fit of the model?

(ii) Use an estimate and its associated 95% confidence interval to describe the relationship between `test` and `glucose` in terms of odds ratio.

(iii) For a patient having mean `glucose`, `bmi` and `pedigree` values, estimate the `age` at which or older she will have at least 0.5 probability to be tested positive.

(Question 1 continued on next page)

(Question 1 continued) Answer the questions in the provided boxes or on plain A4 paper:

(b) Patient A has her BMI equal the third quartile of `bmi` while patient B has her BMI equal the first quartile of `bmi`. Values of `glucose`, `pedigree` and `age` do not change between A and B. Use the results from `Q1mod1` to estimate the odds ratio of positive test for A versus B. Also find a 95% confidence interval for this odds ratio.

(Question 1 continued)

(c) Below is R output of variable selection by AIC using backward elimination procedure:

```
> Q1mod2 <- glm(test ~ (glucose + bmi + pedigree + age)^2, family = binomial,data = Q1.dat)
> step(Q1mod2, trace=1)

Start:  AIC=301; test ~ (glucose + bmi + pedigree + age)^2
                   Df Deviance AIC
- glucose:age       1      279 299.1
- bmi:pedigree      1      279 299.2
- pedigree:age      1      279 299.4
- glucose:bmi       1      280 300.5
<none>                     279 301.2
- bmi:age           1      282 302.2
- glucose:pedigree  1      285 305.0

Step:  AIC=299.1
test ~ glucose + bmi + pedigree + age + glucose:bmi + glucose:pedigree + bmi:pedigree + bmi:age + pedigree:age
                   Df Deviance AIC
- bmi:pedigree      1      279 297
- pedigree:age      1      280 298
- glucose:bmi       1      280 298
<none>                     279 299
- bmi:age           1      282 300
- glucose:pedigree  1      286 304

Step:  AIC=297; test ~ glucose + bmi + pedigree + age + glucose:bmi + glucose:pedigree + bmi:age + pedigree:age
                   Df Deviance AIC
- pedigree:age      1      280 296.3
- glucose:bmi       1      280 296.4
<none>                     279 297.0
- bmi:age           1      282 298.1
- glucose:pedigree  1      286 302.2

Step:  AIC=296.3; test ~ glucose + bmi + pedigree + age + glucose:bmi + glucose:pedigree + bmi:age
                   Df Deviance AIC
- glucose:bmi       1      280 294
<none>                     280 296
- bmi:age           1      283 297
- glucose:pedigree  1      287 301

Step:  AIC=294; test ~ glucose + bmi + pedigree + age + glucose:pedigree + bmi:age
                   Df Deviance AIC
<none>                     280 294
- bmi:age           1      284 296
- glucose:pedigree  1      288 300
```

Write down the best logistic model selected by AIC by specifying which predictor terms remain in the best model together with its AIC value.
Also explain how those predictor terms not in the best model are eliminated.

Question 2 (15 marks)

A study was conducted to investigate the toxicity to the tobacco budworm *Heliothis virescens* of doses of the pyrethroid *trans*-cypermethrin to which the moths were beginning to show resistance. Batches of 20 moths of each sex were exposed for three days to the pyrethroid and the number in each batch that were killed was recorded. The results are given below

	Dosage	Number killed (out of 20)	
Dose (μg)	(=$\log_2$(dose))	Males	Females
1	0	1	0
2	1	4	2
4	2	8	6
8	3	13	10
16	4	17	12
32	5	20	16

Five (logistic regression) models (`Q2mod1` to `Q2mod5`) were fitted and resulted in the residual deviances given in the table below, together with the R model specifications. For these models, `sex` is a factor with 2 levels (male; female), `dose.f` refers to dose treated as a factor with 6 levels, `dose` is the dose treated as a (continuous) variable and `dosage` is $\log_2$(`dose`) also continuous.

Model	formula specification	residual deviance	DF
`Q2mod1`	`sex + dose.f`	4.75	5
`Q2mod2`	`sex + dose + I(dose^2)`	14.6	8
`Q2mod3`	`sex*dosage`	5.18	8
`Q2mod4`	`sex + dosage`	6.48	9
`Q2mod5`	`sex + dosage + I(dosage^2)`	5.78	8

The following R output is needed for answering this question.

```
> qchisq(0.95, c(3,4,5,8,9))
[1]  7.81  9.49 11.07 15.51 16.92
```

Answer the following questions:

(a) For each pair of models below explain whether or not they can be compared by a likelihood ratio test. If yes, explain what hypothesis is being tested and provide a conclusion.

(i) model `Q2mod1` and model `Q2mod3`;

(ii) model `Q2mod1` and model `Q2mod5`

(Question 2 continued)

(b) Some R output of model `Q2mod4` is given below

```
> summary(Q2mod4)$coef

            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.405      0.461   -7.39  1.5e-13
sexmale        0.970      0.350    2.77   0.0055
dosage         1.042      0.129    8.10  5.7e-16
```

Use the above output to answer the following questions.

(i) Is the effect of sex significant in the model? Justify your answer.

(ii) Quantify, in terms of odds ratios, the effects of sex and dosage on toxicity.

(iii) Find an estimate of the dosage, that would kill 50% of **male** budworm moths (the LD(50)). State, with reasons, whether the LD(50) for female moths would be greater or less than that for males.

(iv) Obtain an estimate of the probability that a **female** moth will die within three days if subjected to a dose of 16μg of pyrethroid (or a dosage of 4) for three days.

(Question 2 continued)

(c) Which of the 5 models `Q2mod1` to `Q2mod5` listed in the table above is most appropriate for these data? Give details of any tests that you carry out and clearly state the conclusions drawn from each test.

(d) A total of 240 budworms were used in this study. Describe what would have been the same and what would have been different in the output (residual deviances and their degrees of freedom, differences between residual deviances and their degrees of freedom, parameter estimates and their standard errors) had the data been treated as 240 ungrouped, 0–1 observations, rather than as 12 groups of 20 moths.

Question 3 (12 marks)

The following table presents data for 174 poliomyelitis cases from the US. The individuals were categorized based on **age** (A; 6 levels: 2='0-4' years old, 7='5-9', 12='10-14', 17='15-19', 30='20-39', and 50='40+'), **degree of debilitation** (P; 2 levels: `OK`, `Paralyzed`), and whether or not they had been **vaccinated** (V; 2 levels: `Vaccinated`, `Not Vaccinated`).

Table 1: Number of patients disabled, by age and vaccination.

Ages	**Vaccinated**		**Not Vaccinated**	
	OK	Paralyzed	OK	Paralyzed
0-4 (2)	20	14	10	24
5-9 (7)	15	12	3	15
10-14 (12)	3	2	3	2
15-19 (17)	7	4	1	6
20-39 (30)	12	3	7	5
40+ (50)	1	0	3	2

From these data the following residual deviances and residual degrees of freedom were obtained, using log-linear models with Poisson error.

Table 2: Log-linear Model summaries for 174 poliomyelitis cases.

Model	R Formula	Res. Dev.	Res. D.F.
1	A + V + P	34.00	16
2	A + V*P	19.17	15
3	V + A*P	25.42	11
4	P + A*V	28.88	11
5	A*V + V*P	14.05	10
6	A*P + V*P	10.59	10
7	A*V + A*P	20.29	6
8	A*V + V*P + A*P	2.71	5

The following 95 percentiles of χ^2 distributions may be required in answering questions:

```
> qchisq(0.95, df = c(5, 6, 10, 11, 15, 16))

[1] 11.071  12.592  18.307  19.675  24.996  26.296
```

(Question 3 continued on next page)

(Question 3 continued) Answer the following questions:

(a) Give an interpretation to each of the following models.

(i) `A + V + P`

(ii) `A*V + V*P + A*P`

(b) (i) Test the hypothesis that there is no association between vaccination and paralysis when age is given. State the model under test, the test statistic value, and a conclusion at the 5% significance level.

(ii) Test the hypothesis that there is no association between age and paralysis when vaccination status is given. State the model under test, the test statistic value, and a conclusion at the 5% significance level.

(Question 3 continued)

(c) Find the "best" log-linear model based on the results in Table 2. Justify your finding using necessary analysis of deviance tests and model adequacy tests.

Question 4 (12 marks)

Refer to the same data in Table 1 in Question 3. Suppose we want to collapse the table over age to investigate the relation between vaccination and paralysis. The parameter estimates in Model 2 in Table 2 are given in the R output below:

```
> summary(logMod2)

Call:
glm(formula = freq ~ factor(A) + V * P, family = poisson, data = polio)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.70192  -0.69864   0.04334   0.45244   1.69779

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.3563     0.2145  10.987  < 2e-16 ***
factor(A)7   -0.4128     0.1922  -2.148 0.031685 *
factor(A)12  -1.9169     0.3387  -5.660 1.51e-08 ***
factor(A)17  -1.3291     0.2651  -5.014 5.32e-07 ***
factor(A)30  -0.9237     0.2275  -4.061 4.89e-05 ***
factor(A)50  -2.4277     0.4259  -5.701 1.19e-08 ***
V             0.7646     0.2330   3.282 0.001031 **
P             0.6931     0.2357   2.941 0.003274 **
V:P          -1.1982     0.3184  -3.764 0.000168 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 129.139  on 23  degrees of freedom
Residual deviance:  19.173  on 15  degrees of freedom
AIC: 118.29

Number of Fisher Scoring iterations: 5
```

(Question 4 continued on next page)

(Question 4 continued) Answer the following questions:

(a) Is it reasonable to collapse the table over age? Explain based on the results in Table 2.

(b) Give the value of the residual deviance that will be obtained if the 'no association' model is fitted to the collapsed table. Is the 'no association' model adequate?

(c) A saturated log-linear model has been fitted to the collapsed table. Give the values of the coefficients of `V`, `P` and `V:P` terms in the fitted saturated model. Also give the associated standard errors of these coefficient estimates.

Question 5 (12 marks)

In this question we continue to use the data and model results provided in Questions 3 and 4. We now treat `P` as the response variable, and then fit the frequencies of `Paralyzed` and `OK` to `V` by the logistic regression model $\mathcal{M}$: $\text{logit}(p) = \beta_0 + \beta_1 \texttt{V}$, where p is the probability of `Paralyzed` for an individual. Use the equivalence between logistic and log-linear models for certain contingency table data to answer the following questions:

(a) Give the null deviance value for model $\mathcal{M}$ and explain.

(b) Find the residual deviance value of model $\mathcal{M}$. Then test its model adequacy.

(c) Give the maximum likelihood estimates of β_0 and β_1 together with their standard errors.

Question 6 (15 marks)

A survey of terrace house residents in Copenhagen classified the householders according to the degree of `Contact` they had with other residents, their feeling of `Influence` on apartment management and their level of `Satisfaction` with their housing conditions. The data are summarized in the following table and `R` data.frame `satis`.

Contact		Low			High		
Influence		Low	Med.	High	Low	Med.	High
	Low	18	15	7	57	31	5
Satisfaction	Medium	6	13	5	23	21	6
	High	7	13	11	13	13	13

```
> satis

   Freq   Satis Contact Influ
1    18    1Low     Low   Low
2     6 2Medium     Low   Low
3     7   3High     Low   Low
4    15    1Low     Low   Mid
5    13 2Medium     Low   Mid
6    13   3High     Low   Mid
7     7    1Low     Low  High
8     5 2Medium     Low  High
9    11   3High     Low  High
10   57    1Low    High   Low
11   23 2Medium    High   Low
12   13   3High    High   Low
13   31    1Low    High   Mid
14   21 2Medium    High   Mid
15   13   3High    High   Mid
16    5    1Low    High  High
17    6 2Medium    High  High
18   13   3High    High  High
```

Treating the level of `Satisfaction` as a nominal categorical response variable, a multi-categorical logit model has been fitted resulting in the following `R` output:

```
> satis.mod <- multinom(Satis~Contact+Influ, data=satis, weights=Freq, Hess=T)
> summary(satis.mod)

Call:
multinom(formula = Satis ~ Contact + Influ, data = satis, weights = Freq,
    Hess = T)

Coefficients:
        (Intercept) ContactHigh InfluMid InfluHigh
2Medium       -0.96      0.0129     0.65     0.866
3High         -1.05     -0.3770     0.70     1.934

Std. Errors:
        (Intercept) ContactHigh InfluMid InfluHigh
2Medium       0.328       0.320    0.317     0.476
3High         0.341       0.324    0.355     0.439

Residual Deviance: 555
AIC: 571
```

(Question 6 continued on next page)

(Question 6 continued) Answer the following questions:

(a) Write down the model fitted in the above R output. You need to define the response variable and predictors for the model. Also, you need to specify the probability distribution of the response variable and estimates of all parameters in the model.

(b) Provide an interpretation for the coefficient estimate 1.934. Then calculate an approximate 95% confidence interval for the odds ratio of high level of `Satisfaction` versus lower level of `Satisfaction` for householders having high `Influence` feeling against those having low `Influence` feeling on management.

(Question 6 continued)

(c) Estimate the probability for each of the 3 levels of `Satisfaction` for a householder having low degree of `Contact` and high `Influence` feeling.

Question 7 (15 marks)

Refer to the satis data in Q6. Note that the level of Satisfaction is ordinal by nature, thus we treat it as an ordinal categorical response variable. A cumulative proportional odds model is fitted to the satis data, producing the following R output (Note Coefficients Values need to change sign for being used in the model):

```
> satis$SatisO=as.ordered(satis$Satis)
> Osatis.mod=polr(SatisO~Contact + Influ, data=satis,weights=Freq, Hess=T,
                                    method="logistic")
> summary(Osatis.mod)

Call:
polr(formula = SatisO ~ Contact + Influ, data = satis, weights = Freq,
    Hess = T, method = "logistic")

Coefficients:
             Value Std. Error t value
ContactHigh -0.242      0.246  -0.984
InfluMid     0.603      0.256   2.357
InfluHigh    1.591      0.342   4.646

Intercepts:
              Value  Std. Error t value
1Low|2Medium   0.234  0.259      0.905
2Medium|3High  1.496  0.275      5.434

Residual Deviance: 556.90
AIC: 566.90

> solve(Osatis.mod$Hessian)

              ContactHigh InfluMid InfluHigh 1Low|2Medium 2Medium|3High
ContactHigh       0.06072   0.0081    0.0126       0.0457      -0.00093
InfluMid          0.00808   0.0654    0.0350       0.0374       0.00218
InfluHigh         0.01260   0.0350    0.1172       0.0396       0.00656
1Low|2Medium      0.04574   0.0374    0.0396       0.0671      -0.00365
2Medium|3High    -0.00093   0.0022    0.0066      -0.0036       0.01123
```

(Question 7 continued on next page)

(Question 7 continued) Answer the following questions:

(a) Write down the model fitted in the above `R` output. You need to define the response variable and predictors for the model. Also, you need to specify the probability distribution of the response variable and estimates of all parameters in the model.

(b) Estimate the odds ratio of high level of `Satisfaction` versus the other levels of `Satisfaction` for a householder who has *low* degree of `Contact` but *high* `Influence` feeling against another householder who has *high* degree of `Contact` but *low* `Influence` feeling. Also calculate an approximate 95% confidence interval for this odds ratio.

(Question 7 continued)

(c) Estimate the probability for each of the 3 levels of `Satisfaction` for a householder having low degree of `Contact` and high `Influence` feeling.

Question 8 (15 marks)

The toenail data comes from a multi-center study comparing two oral treatments for toenail infection. Patients were evaluated for the degree of separation of the nail. A total of 294 patients were randomised into two treatments and were followed over seven visits: four in the first year and yearly thereafter. Some of the patients did not attend all seven visits, thus only a total of 1908 visits were observed. The patients had not been treated prior to the first visit so this should be regarded as the baseline.
The variables available in the data are

outcome:	$0 =$ none or mild separation, $1 =$ moderate or severe separation
ID:	ID of patient
treatment:	the treatment; $A = 0$ or $B = 1$
month:	time of the visit, in months, from the first visit
visit:	the number of the visit

The purpose of this study is to see how toenail infection responds to the treatments and progresses over time. Some analysis has been done to the data in R, producing the following output.

```
> toenail[1:14,]

   ID outcome treatment  month visit
1   1       1         1  0.000     1
2   1       1         1  0.857     2
3   1       1         1  3.536     3
4   1       0         1  4.536     4
5   1       0         1  7.536     5
6   1       0         1 10.036     6
7   1       0         1 13.071     7
8   2       0         0  0.000     1
9   2       0         0  0.964     2
10  2       1         0  2.000     3
11  2       1         0  3.036     4
12  2       0         0  6.500     5
13  2       0         0  9.000     6
14  3       0         0  0.000     1
```

```
> tail(toenail)

      ID outcome treatment month visit
1903 383       1         1  0.00     1
1904 383       1         1  1.04     2
1905 383       1         1  2.04     3
1906 383       1         1  3.29     4
1907 383       0         1  7.29     5
1908 383       0         1 10.79     6
```

```
> str(toenail)

'data.frame':   1908 obs. of  5 variables:
 $ ID       : int  1 1 1 1 1 1 1 2 2 2 ...
 $ outcome  : int  1 1 1 0 0 0 0 0 0 1 ...
 $ treatment: int  1 1 1 1 1 1 1 0 0 0 ...
 $ month    : num  0 0.857 3.536 4.536 7.536 ...
 $ visit    : int  1 2 3 4 5 6 7 1 2 3 ...
```

```
> library(geepack}
fit.exch <- geeglm(outcome~treatment+month, family=binomial(link="logit"),
data=toenail, id=ID, corstr = "exchangeable", std.err="san.se");  summary(fit.exch)

Call:
geeglm(formula = outcome ~ treatment + month, family = binomial(link = "logit"),
    data = toenail, id = ID, corstr = "exchangeable", std.err = "san.se")

 Coefficients:
            Estimate Std.err  Wald Pr(>|W|)
(Intercept)  -0.6104  0.1777 11.80  0.00059 ***
treatment     0.0402  0.2532  0.03  0.87388
month        -0.2051  0.0259 62.66  2.4e-15 ***
---

Estimated Scale Parameters:
            Estimate Std.err
(Intercept)     1.09   0.423

Correlation: Structure = exchangeable  Link = identity

Estimated Correlation Parameters:
      Estimate Std.err
alpha    0.424   0.182
Number of clusters:   294   Maximum cluster size: 7

> summary(fit.exch)$cov.unscaled

         [,1]      [,2]      [,3]
[1,]  0.03159 -0.031374 -0.001395
[2,] -0.03137  0.064120 -0.000546
[3,] -0.00139 -0.000546  0.000671
```

Use the above output to answer the following questions.

(Question 8 continued on next page)

(Question 8 continued) Answer the following questions:

(a) Let y_{it} be the response value `outcome` of patient i during `visit` t. Write down the model involved in the analysis, including the mean, variance and correlation coefficient of y_{it}'s. Give the estimates of the parameters appearing in the model.

(b) Write down the model's design matrix for data where `ID`=383.

(Question 8 continued)

(c) Estimate the odds ratio of toe infection of a patient with treatment B versus with treatment A at a given value of `month`. Calculate an approximate 95% confidence interval for this odds ratio.

(d) Estimate the probability of toe infection in the first month from the first visit for a patient using treatment A. Also compute an approximate 95% confidence interval for this probability.

End of Exam—Total Available Marks = 110