THE UNIVERSITY OF AUCKLAND

SEMESTER ONE 2019
Campus: City

STATISTICS

Advanced Statistical Modelling

(Time allowed: TWO hours)

INSTRUCTIONS

- This examination is in two parts, Part A and Part B.
- Attempt **ALL** questions in Part A.
- Attempt **ALL** questions in Part B.
- The total marks for this examination are **100 marks**.

PART A

1. [**33 marks**] When a bank issues a loan, sometimes the customer 'defaults' and fails to make all repayments. A bank would prefer to only approve loans for customers who are unlikely to default. They wish to conduct a statistical analysis to better understand which variables are related to the probability that a customer repays their loan in full. They select a sample of customers whose loan applications have previously been approved. The data set `loans.df` contains the following variables:

`age`	A categorical variable describing the age of the applicant, with levels `A` (less than 30 years), `B` (30–50 years, inclusive), and `C` (greater than 50 years).
`own.house`	A categorical variable describing whether or not the applicant owns a house, with levels `yes` and `no`.
`duration`	The duration of the loan, in months.
`n`	The number of applications with a particular combination of the above variables.
`defaults`	The number of the `n` applicants that defaulted on their loan by failing to pay it back.
`p`	The proportion of applicants who failed to pay back their loan in full, `defaults/n`.

In total there are 30 rows in the data set. The first six rows are displayed below:

```
> head(loans.df)

  defaults   n duration age own.house
1       18  56       12   A        no
2       28 107       12   A       yes
3       11  47       12   B        no
4       23 159       12   B       yes
5        3  16       12   C        no
6        7  48       12   C       yes
```

Consider the model below:

```
> model.A <- glm(cbind(defaults, n - defaults) ~ duration + age + own.house,
                 family = binomial, data = loans.df)
```

```
> summary(model.A)


Call:
glm(formula = cbind(defaults, n - defaults) ~ duration + age +
    own.house, family = binomial, data = loans.df)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.729  -0.652  -0.232   0.412   1.920

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.03093    0.20465   -5.04  4.7e-07 ***
duration     0.03788    0.00616    6.15  7.6e-10 ***
ageB        -0.51713    0.15375   -3.36  0.00077 ***
ageC        -0.47905    0.23786   -2.01  0.04402 *
own.houseyes -0.55122    0.15325   -3.60  0.00032 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 95.764  on 29  degrees of freedom
Residual deviance: 28.366  on 25  degrees of freedom
AIC: 122.7

Number of Fisher Scoring iterations: 4

> 1 - pchisq(deviance(model.A), df.residual(model.A))

[1] 0.29129

> confint(model.A)

Waiting for profiling to be done...

                2.5 %    97.5 %
(Intercept) -1.435745 -0.632692
duration     0.025895  0.050060
ageB        -0.819303 -0.216176
ageC        -0.955403 -0.020827
own.houseyes -0.851081 -0.249891
```

(a) For `model.A`, write equations to define (i) the assumed relationship between the explanatory terms and the probability of a customer defaulting on their loan, and (ii) the assumed distribution of the response variable. [4 marks]

$$\text{logit}(p_i) = \beta_0 + \beta_1 d_i + \beta_2 b_i + \beta_3 c_i + \beta_4 h_i$$
$$Y_i \sim \text{Binomial}(n_i, p_i),$$

where, for the ith group, n_i is the number of customers, Y_i is the number who defaulted on their loan, p_i is the probability of a customer defaulting, d_i is the duration of the loan in months, b_i is a dummy variable equal to 1 if customers are between 30–50 years of age, c_i is a dummy variable equal to 1 if customers are over 50 years of age, and h_i is a dummy variable equal to 1 if customers own their own house.

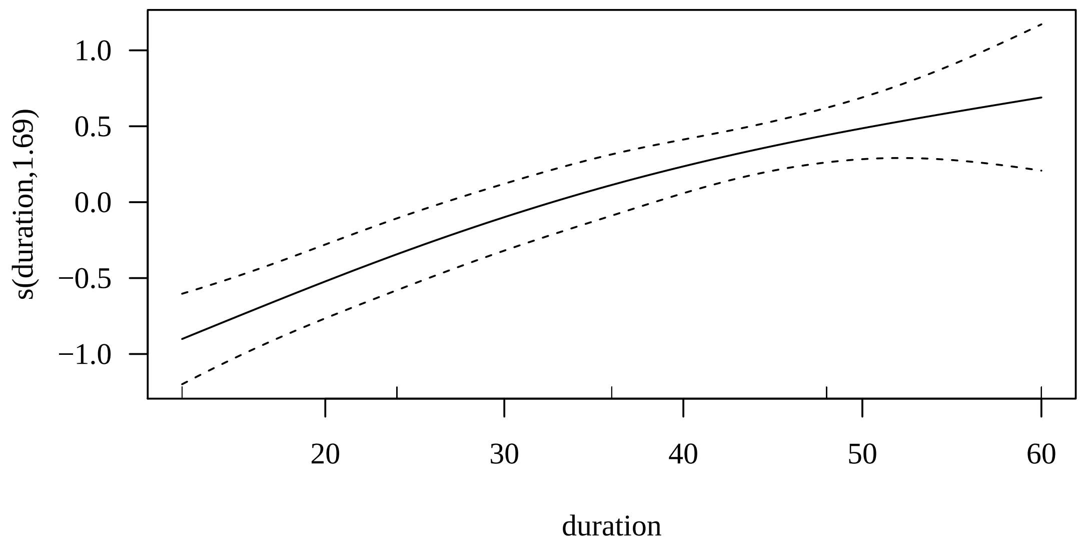
(b) Interpret the effect of loan duration estimated by `model.A`. [3 marks]

Holding all other variables constant, a one-month increase in loan length is associated with an increase in the odds of defaulting of between 2.6 and 5.1%

(c) Based on the model's deviance, is there evidence to suggest the model does not fit the data? Explain your answer. [2 marks]

If we assume that the deviance has a chi-squared distribution under the null hypothesis that the model is correct, then no. A p-value testing this null hypothesis is 0.29, which is large.

```
> library("mgcv")
> model.gam <- gam(cbind(defaults, n - defaults) ~ s(duration, k = 5) +
                      age + own.house,
                  family = binomial, data = loans.df)
> plot(model.gam)
```




```
> model.B <- glm(cbind(defaults, n - defaults) ~ duration + I(duration^2) +
                      age + own.house,
                 family = binomial, data = loans.df)
```

(d) Interpret the GAM plot, in terms of deciding on possible polynomial terms to include in the model. [2 marks]

The GAM plot shows slight curvature. It is probably worth fitting a model with a quadratic term to see how things go, although it is not obvious that this will lead to a better model.

```
> summary(model.B)



Call:
glm(formula = cbind(defaults, n - defaults) ~ duration + I(duration^2) +
    age + own.house, family = binomial, data = loans.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8940  -0.5758  -0.0547   0.5964   1.6770

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.544969   0.358342   -4.31  1.6e-05 ***
duration       0.082261   0.025962    3.17  0.00153 **
I(duration^2) -0.000751   0.000425   -1.77  0.07724 .
ageB          -0.525180   0.154062   -3.41  0.00065 ***
ageC          -0.454662   0.238084   -1.91  0.05618 .
own.houseyes  -0.567527   0.153747   -3.69  0.00022 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 95.764  on 29  degrees of freedom
Residual deviance: 25.261  on 24  degrees of freedom
AIC: 121.6

Number of Fisher Scoring iterations: 4


> library("MuMIn")
> AICc(model.A, model.B)


        df   AICc
model.A  5 125.23
model.B  6 125.28
```

(e) Compare `model.A` and `model.B`. Do you have a preference for which is better? How strong is your preference? Explain your answers. [4 marks]

From the GAM plot, it is not clear whether or not a quadratic term is required. We might prefer `model.A` due to its AICc value being smaller, but is virtually the same as the AICc value for `model.B`; any preference that we have for `model.A` should be extremely slight.

Consider the following models:

```
> model.D <- glm(cbind(defaults, n - defaults) ~ 1, binomial,
                 data = loans.df)
> model.E <- glm(cbind(defaults, n - defaults) ~ as.factor(1:30),
                 binomial, data = loans.df)
> deviance(model.D)

[1] 95.764

> logLik(model.D)

'log Lik.' -90.064 (df=1)

> logLik(model.E)

'log Lik.' -42.182 (df=30)
```

Note that `model.D` estimates one parameter, and `model.E` estimates 30 parameters. Recall there are 30 rows in `loans.df`.

(f) Calculate the following:

i. The deviance of `model.E`. [2 marks]

This is a saturated model. The deviance is 0.

ii. The null deviance. [2 marks]

The null deviance is the deviance of `model.D`, which is 95.76.

iii. The maximized log-likelihood of `model.B`. [4 marks]

$$\begin{aligned} D_B &= 2 \times (L_E - L_B) \\ L_B &= L_E - \frac{D_B}{2} \\ &= -42.18 - \frac{25.26}{2} \\ &= -54.81 \end{aligned}$$

The line of code below carries out a hypothesis test. Under the null hypothesis being tested, the test statistic has a chi-squared distribution.

```
> anova(model.D, model.A, test = "Chisq")
```

(g) What is the null hypothesis being tested? Make your answer specific so that it refers to effects estimated from this particular data set. [2 marks]

The null hypothesis is that loan, age, and house ownership are all unrelated to the probability of a customer defaulting on their loan.

(h) What is the alternative hypothesis? Make your answer specific so that it refers to effects estimated from this particular data set. [2 marks]

At least one of the explanatory variables (loan duration, age, and house ownership) is related to the probability of a customer defaulting on their loan.

(i) Calculate the test statistic, and the degrees of freedom of its chi-squared distribution under the null hypothesis. [4 marks]

The test statistic is the difference in the deviances:

$$95.76 - 28.37 = 67.40.$$

The degrees of freedom is the difference in the number of parameters, which is 4.

(j) Do you think the p-value would be sufficiently small to reject the null hypothesis? Explain your answer. (Hint: the expected value of a chi-squared distribution is equal to its degrees of freedom.) [2 marks]

Yes, because the test statistic is much larger than its expected value. Alternatively, because the output from `summary(model.A)` already provides very strong evidence to suggest that all three variables are related to the probability of a customer defaulting on their loan.

PART B

2. **[9 Marks]** In 2012, BMC Public Health reported the following mean weights by region.

```
North America: 80.7 kg
Oceania, including Australia and NZ: 74.1 kg
Europe: 70.8 kg
Africa: 60.7 kg
Asia: 57.7 kg
```

Suppose that the weights within each region are normally distributed.

Write down R commands to obtain the following answers. Do not try to evaluate their value.

(a) Suppose that the standard deviation for Oceania is 17 kg. What proportion of the population of Oceania is less heavy than the North American mean? [2 marks]

```
> pnorm(80.7, 74.1, 17)

[1] 0.65108
```

(b) What weight does a New Zealander have if he weighs in the top 20% of other New Zealanders? [2 marks]

```
> qnorm(0.80, 74.1, 17)

[1] 88.408
```

(c) In a town just south of the Canadian border the peoples' annual hamburger consumption has a negative binomial distribution with mean 80 and $\theta = 0.472$. What is the probability that somebody randomly chosen from that town consumes 300 or more hamburgers per year? [2 marks]

```
> 1 - pnbinom(300-1, mu = 80, size = 0.472)

[1] 0.055599

> pnbinom(300-1, mu = 80, size = 0.472, lower.tail = FALSE)

[1] 0.055599
```

(d) The percent of Americans weighing 136 kg or more (300 pounds or higher) is around 2.0%. Estimate the standard deviation of the weights of Americans. [3 marks]

```
> (136 - 80.7) / qnorm(1 - 0.02)

[1] 26.926
```

3. **[9 Marks]** A statistician fits the regression model

$$\log y_i = \beta_0 + \beta_1 \log x_i + \varepsilon_i$$

to data (x_i, y_i), $i = 1, \ldots, n$, where $\varepsilon_i \sim N(0, \sigma^2)$ independently.

(a) Give one situation where transforming x_i by taking its logarithm would be a good idea. [1 mark]

If a histogram of x showed right skew.

(b) Given vectors x and y what R command would you type to fit the model? [1 mark]

`lm(log(y) ~ log(x))`

(c) Suppose that $\widehat{\beta}_0 = -0.453$ and $\widehat{\beta}_1 = 0.0163$. For somebody with the value $(x = 2, y = 3)$, what would his/her fitted value be? [3 marks]

The fitted value would be

```
> eta <-  -0.453 + 0.0163 * log(2)
> exp(eta)

[1] 0.64294
```

(d) If x is tripled, what effect does it have on y? That is, if x changes to $3x$, what effect does it have on the mean or median of y? State and prove your result. [4 marks]

The median of y gets multiplied by $3^{\beta_1} \approx 1.018$ when estimates are plugged in. The proof is straightforward; see the notes.

4. **[6 Marks]**

(a) Give a one sentence definition of an explanatory model. [2 marks]

An explanatory model is a statistical model for testing causal theory.

(b) Give a one sentence definition of a predictive model. [2 marks]

A predictive model is a model use for predicting new records/outcomes/-categories.

(c) What effect does measurement error have on the predictive power of a model? [2 marks]

Poor measurement may decrease the predictive power of a model, but the model may still be used for prediction.

5. **[10 Marks]** A simple Poisson regression was fitted to a sample of 2622 female Europeans. The response was `pulse` rate (beats per minute) versus `age` (years). Here is the fitted model:

```
> pfit1


Call:  glm(formula = pulse ~ age, family = poisson, data = eurof)

Coefficients:
(Intercept)          age
   4.270399    -0.000478

Degrees of Freedom: 2621 Total (i.e. Null);  2620 Residual
Null Deviance:     4330
Residual Deviance: 4310  AIC: 20300
```

(a) Of interest to medical researchers and doctors is the age at which the mean pulse is 70 beats per minute. Call this age x_*, say. Show that our estimate from the data is $\widehat{x}_* \approx 45.81$. [2 marks]

```
> pulse.star <- 70
> age.star <- (log(pulse.star) - coef(pfit1)[1]) / coef(pfit1)[2]
> age.star

(Intercept)
     45.813
```

(b) Write several lines of R code to perform nonparametric bootstrapping to obtain an approximate 95% confidence for x_*. [8 marks]

```
> set.seed(123)  # For reproducibility of the results
> n = nrow(eurof)
> Nsim = 1000  # 1e4 is better
> betasBS = matrix(0, Nsim, length(coef(pfit1)))
> for (i in 1:Nsim) {
  # Draw a random sample (with replacement) from these data:
    id = sample(1:n, n, replace = TRUE)
    BS.df = eurof[id, ]  # Bootstrap sample
    mod_i = glm(pulse ~ age, poisson, data = BS.df)
    betasBS[i, ] = coef(mod_i)
  }
>
> pulse.star <- 70
> quantile((log(pulse.star) - betasBS[, 1]) / betasBS[, 2],
           c(0.025, 0.975))

  2.5%  97.5%
26.419 61.820
```

6. **[4 Marks]** Consider the following R code:

```
> fit1 <-  gam(y01 ~ x2 + s(x3) + x4 + I(x4^2), binomial, data = my.df)
> fit2 <- vgam(y01 ~ x2 + s(x3) + x4 + I(x4^2), binomialff, data = my.df)
```

The variable `y01` has values 0 and 1, and variables `x2`–`x4` are numeric. Both `fit1` and `fit2` estimate the same model and are very similar; they come from the `mgcv` and `VGAM` packages respectively.

Write down the formula for the model. Briefly define any notation or quantities used.

The model is a logistic regression. It is logit $\Pr(Y = 1) = \beta_1 + \beta_2 x_2 + f_3(x_3) + \beta_4 x_4 + \beta_5 x_4^2$, where f_3 is an arbitrary smooth function estimated by a smoother such as a spline. The smooth is centred for identifiability. The additive predictor has x_4 entered in as a quadratic.

7. **[4 Marks]** Briefly explain what a test data set is and what purpose it is used for.

Used to estimate the test error curve as a function of model complexity. It is used repeatedly to estimate the right sized complexity—not underfitting or overfitting. The test data is only used once—out of a vault, so to speak. Ideally, training, holdout and test data comprise 1/3 of the total data each, and they are randomly partitioned into such.

8. **[6 Marks]** Suppose we collect data for a group of students in a statistics class with variables X_2 = hours studied, X_3 = undergraduate GPA, and Y = receive an A (1) or not (0). We fit a logistic regression logit $\Pr(Y = 1) = \beta_1 + \beta_2 x_2 + \beta_3 x_3$ and produce estimated coefficients $\widehat{\beta}_1 = -6$, $\widehat{\beta}_2 = 0.03$, $\widehat{\beta}_3 = 0.5$.

   (a) Estimate the probability that a student who studies for 40 hours and has an undergraduate GPA of 6.5 gets an A in the class. [3 marks]

```
> ldata <- as.matrix(data.frame(intercept = 1, hours=40, gpa = 6.5))
> beta.vec <- c(-6, 0.03, 0.5)
> (eta <- c(ldata %*% beta.vec))

[1] -1.55

> (prob <- exp(eta) / (1 + exp(eta)))

[1] 0.17509
```

   (b) According to the fitted model, how many hours would a student with an undergraduate GPA of 6.5 need to study to have exactly a 50% chance of getting an A in the class? [3 marks]

```
> odds  <- 1  # corresponds to prob == 0.5
> (eta <- log(odds))

[1] 0
```

```
> hours <- (eta - sum(ldata[1, -2] %*% beta.vec[-2])) / beta.vec[2]
> hours

[1] 91.667
```

9. **[10 Marks]** The data frame `swiss` from the datasets R package concerns the standardized fertility measure and socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888.

The columns are (in order):

Fertility	a common standardized fertility measure
Agriculture	% of males involved in agriculture as occupation
Examination	% draftees receiving highest mark on army examination
Education	% education beyond primary school for draftees
Catholic	% catholic as opposed to protestant
Infant.Mortality	live births who live less than 1 year.

The following analysis was performed.

```
> data(swiss, package = "datasets")
> full.model <-  lm(Fertility ~ ., data = swiss)
> summary(full.model)



Call:
lm(formula = Fertility ~ ., data = swiss)

Residuals:
    Min      1Q  Median      3Q     Max
-15.274  -5.262   0.503   4.120  15.321

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       66.9152    10.7060    6.25  1.9e-07 ***
Agriculture       -0.1721     0.0703   -2.45   0.0187 *
Examination       -0.2580     0.2539   -1.02   0.3155
Education         -0.8709     0.1830   -4.76  2.4e-05 ***
Catholic           0.1041     0.0353    2.95   0.0052 **
Infant.Mortality   1.0770     0.3817    2.82   0.0073 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.17 on 41 degrees of freedom
Multiple R-squared:  0.707,Adjusted R-squared:  0.671
F-statistic: 19.8 on 5 and 41 DF,  p-value: 5.59e-10


> myvector <- round(diag(solve(cor(swiss[, -1]))), 2)
> myvector
```

```
     Agriculture       Examination         Education          Catholic
            2.28              3.68              2.77              1.94
Infant.Mortality
            1.11


> library("leaps")
> subsets.out <- regsubsets(Fertility ~ ., nbest = 1, data = swiss)
> sso <- summary(subsets.out)
> sso$outmat


         Agriculture Examination Education Catholic Infant.Mortality
1  ( 1 ) " "         " "         "*"       " "      " "
2  ( 1 ) " "         " "         "*"       "*"      " "
3  ( 1 ) " "         " "         "*"       "*"      "*"
4  ( 1 ) "*"         " "         "*"       "*"      "*"
5  ( 1 ) "*"         "*"         "*"       "*"      "*"
```

Based on this output, answer the following questions.

(a) How many possible models could the exhaustive method fit if the intercept term is always included? [3 marks]

Use $2^k - 1$ where $k = 5$. This value is 31.

(b) Comment on the output of `regsubsets()`. [3 marks]

The variable `Education` is always included, so that suggests that that variable is important. It is confirmed by a very low p-value in the full model. `Fertility` is negatively correlated with `Education`.

(c) What is the purpose of computing `myvector`? What does `myvector` say about the data or analysis? [2 marks]

`myvector` computes the VIF, which is a test for multicollinearity. Since no values are greater than 10 then it is not a problem.

(d) If backward elimination was performed starting from the full model, what variable would be removed first? [2 marks]

The variable `Examination`, since it is the only one that is nonsignificant.

10. **[9 Marks]** Consider the topic of causal inference.

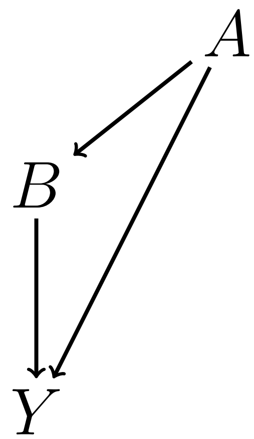
(a) Explain very briefly what an effect modifier is. [3 marks]

We say that M is a modifier of the effect of X on Y when the average causal effect of X on Y varies across levels of M. We handle them by allowing interactions in a statistical model.

(b) State two ways that causal relationships can be investigated. [2 marks]

i. Designed experiments.
ii. Causal analysis of observational data.

(c) Consider the following causal diagram for the remainder of this question.



Suppose the data comprises the following variables inside a data frame called `my.df`:

`y` count response,

`A` categorical variable with 3 levels,

`B` binary variable.

Write down the R command to fit a model to explore all direct causal effects on the response. [2 marks]

```
> glm(y ~ A + B, poisson, data = my.df)  # This
> glm(y ~ A * B, poisson, data = my.df)  # Or this
```

(d) Write down the R command to fit a model to explore all indirect causal effects. [1 mark]

```
> glm(B ~ A, binomial, data = my.df)
```

(e) Write down the R command to fit a model to explore the total effects of A on the response. [1 mark]

```
> glm(y ~ A, poisson, data = my.df)
```