THE UNIVERSITY OF AUCKLAND

SEMESTER ONE 2017
Campus: City

STATISTICS

Advanced Statistical Modelling

(Time allowed: THREE hours)

INSTRUCTIONS

SECTION A: Multiple Choice (24 marks)

- Answer **ALL 12** questions on the coloured teleform sheet provided.
- To answer, fill in the appropriate box on the teleform sheet.
- Use pencil only. To change an answer, erase the original answer completely and fill in a new answer.
- If you give more than one answer to any question, you will receive zero marks for that question.
- All questions carry the same mark value.
- All questions have a single correct answer.
- Incorrect answers are not penalized.

SECTION B (76 marks)

- Answer all questions.

Total for both parts: 100 marks.

SECTION A

1. Which one of the following statements about smoothing is TRUE?

   (zz) The functions `ns()` and `bs()` create regression splines and reside in the R package splines. They are used within modelling functions such as `lm()` and `glm()`, and appear on the right hand side of the formula.

   (1) When smoothing the choice of its smoothing parameter is unimportant for avoiding underfitting and overfitting.

   (1) Fitting a quadratic polynomial in x, compared to fitting a cubic polynomial in x, will result in a residual sum of squares that is lower or equal in value.

   (1) The mgcv package can fit additive models, and its smoothing parameter selection is extremely reliable.

   (1) Additive models are model-driven rather than data-driven, and work best when there are no interactions.

2. Which one of the following statements is FALSE?

   (zz) `contrasts(as.factor(letters[5:9]))` returns a 4×5 matrix and whose first row are all zeroes.

   (1) The `poly()` function creates orthogonal polynomials, which results in computations that are more numerically stable than if ordinary polynomials are used.

   (1) The central concept of smoothing is localness, also known as a neighbourhood, and smoothing forms the basic idea behind additive models.

   (1) The function `anova()`, when applied to a `"lm"` object, tests the statistical significance of adding the terms in a sequential manner.

   (1) The fitted values of a 1-way ANOVA do not depend on whether `contr.treatment()`, `contr.SAS()` or `contr.helmert()` is used, and the first of these functions is the default in R for unordered factors.

3. Consider a Poisson regression model where μ represents the expected value of the response. Which of the following statements is TRUE?

   (zz) $\log(\mu)$ is a linear combination of the explanatory variables.

   (1) μ is a linear combination of the explanatory variables.

   (1) $\exp(\mu)$ is a linear combination of the explanatory variables.

   (1) $\log(\mu/(1-\mu))$ is a linear combination of the explanatory variables.

   (1) $\mu/(1-\mu)$ is a linear combination of the explanatory variables.

4. Consider two logistic regression models fitted to the same data, Model A and Model B. Model B is a submodel of Model A, and let Model B be the correct model. Let D_A be the deviance of Model A and D_B be the deviance of Model B. Which of the following statements is FALSE?

   (zz) If the models were fitted to **ungrouped** data and there were a large number of observations, then $D_B - D_A$ **does not** have an approximate chi-squared distribution.

   (1) If the models were fitted to **grouped** data, and the number of trials associated with each observation was large, then D_B **does** have an approximate chi-squared distribution.

   (1) If the models were fitted to **grouped** data and there were a large number of total trials, then $D_B - D_A$ **does** have an approximate chi-squared distribution.

   (1) The null deviance does not depend on which model is being considered.

   (1) $D_A \leq D_B$.

5. Consider prediction for logistic regression models, where cases with estimated success probabilities of c or greater are predicted as 'successes'. Which of the following statements is TRUE?

   (zz) Increasing the threshold c leads to a decrease in sensitivity.

   (1) If the area under the ROC curve is 0.5 or greater, then the model has good predictive power.

   (1) Adding additional explanatory variables to a model will always increase specificity and sensitivity.

   (1) A model that is good for prediction has high specificity and low sensitivity.

   (1) The area under the ROC curve depends on the value of c that is selected.

6. Consider a Poisson regression model for which there is evidence of overdispersion. Which of the following statements is FALSE?

   (zz) If a quasi-Poisson model were fitted to the data then we would expect both the estimated coefficients and their standard errors to remain the same.

   (1) A potential remedy for the overdispersion is to fit a quasi-Poisson model.

   (1) If a quasi-Poisson model were fitted to the data, we would expect the estimate of the dispersion parameter to be greater than 1.

   (1) A potential remedy for the overdispersion is to fit a negative binomial regression.

   (1) If a quasi-Poisson model were fitted to the data then we would expect the p-values listed in the rightmost column of the `summary()` output to increase.

The next two questions are based on the following data. In total, 180 beer drinkers were given identical bottles of beer to drink. They were split into three groups of 60: the first was told that the price of the beer was high, the second was told that the price was medium, and the third was told that the price of the beer was low. Each beer drinker categorised the quality of the beer as either 'poor' or 'good'. It was of interest to determine whether the perceived price of a beer was related to the reported quality. The data are shown in the following contingency table:

		Quality	
		Poor	Good
Price	Low	24	36
	Med	22	38
	High	12	48

The following code was used to analyse these data:

```
> beer.df


   price quality count
1   high    poor    12
2    low    poor    24
3 medium    poor    22
4   high    good    48
5    low    good    36
6 medium    good    38


> fit.beer.1 <- glm(count ~ quality * price, poisson, data = beer.df)
> fit.beer.2 <- glm(count ~ quality + price, poisson, data = beer.df)
> anova(fit.beer.1, test = "Chisq")


Analysis of Deviance Table

Model: poisson, link: log

Response: count

Terms added sequentially (first to last)



              Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                              5       29.9
quality        1     23.3         4        6.6  1.4e-06 ***
price          2      0.0         2        6.6    1.000
quality:price  2      6.6         0        0.0    0.037 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


> exp(confint(fit.beer.1))


Waiting for profiling to be done...
```

```
                            2.5 %   97.5 %
(Intercept)             25.48403 49.07326
qualitypoor              0.39304  1.11070
pricemedium              0.66843  1.67003
pricehigh                0.86790  2.06677
qualitypoor:pricemedium  0.41399  1.81488
qualitypoor:pricehigh    0.16150  0.83524
```

7. For the following statement, fill in the blank (????) to give the best interpretation:

   The estimated odds of a high-price drinker rating the beer as 'good' are approximately ???? times the odds of a medium-price drinker rating the beer as 'good'.

   (zz) 2.32

   (1) 0.43

   (1) 2.67

   (1) 2.49

   (1) 1.15

8. Which of the following statements is FALSE?

   (zz) We have evidence to suggest that the odds of a low-price drinker rating the beer as 'good' are different to the odds of both medium- and high-price drinkers rating the beer as 'good', because neither of the corresponding confidence intervals for these odds ratios contain 0.

   (1) We have evidence against the null hypothesis that the perceived price of a beer is independent of its reported quality; it is rejected at the 5% level of significance.

   (1) The deviance of model `fit.beer.2` is approximately 6.6.

   (1) The deviance of model `fit.beer.1` is 0.

   (1) The model `fit.beer.2` assumes that reported beer quality and perceived beer price are independent.

The next two questions are based on votes in the US Senate on a bill regarding the Corporate Average Fuel Economy (CAFE) standard. The bill was widely held to be beneficial to automotive manufacturers, as a vote of NO would have forced them to increase fuel economy across their fleets. Along with their vote on the bill (YES or NO), each senator's party affiliation (Democrat or Republican) was recorded, as well as their total lifetime dollar amount contributed to them by the automotive industry. The data are as follows:

Party	Contributions	Yes	No
D	0	8	21
D	1	2	7
D	2	7	2
D	3	2	1
R	0	3	3
R	1	17	1
R	2	13	1
R	3	10	1

Here, 'D' refers to the Democrat party, and 'R' refers to the Republican party. The 'Contributions' variable is the lifetime dollar amount contributed to a senator by the automotive industry in tens of thousands of dollars (so a value of 2 indicates US$20 000 in contributions). The columns 'Yes' and 'No' indicate the numbers of senators who voted YES and NO on the bill for each unique combination of the 'Party' and 'Contributions' variables. The code below analyses these data.

```
> vote.df


  party contributions yes no
1     D             0   8 21
2     D             1   2  7
3     D             2   7  2
4     D             3   2  1
5     R             0   3  3
6     R             1  17  1
7     R             2  13  1
8     R             3  10  1


> fit.int <- glm(cbind(yes, no) ~ party * contributions,
                 binomial, data = vote.df)
> fit.add <- glm(cbind(yes, no) ~ party + contributions,
                 binomial, data = vote.df)
> anova(fit.add, fit.int, test = "Chisq")


Analysis of Deviance Table

Model 1: cbind(yes, no) ~ party + contributions
Model 2: cbind(yes, no) ~ party * contributions
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         5       6.56
2         4       6.56  1  0.00418     0.95
```

```
> summary(fit.add)



Call:
glm(formula = cbind(yes, no) ~ party + contributions, family = binomial,
    data = vote.df)

Deviance Residuals:
      1        2        3        4        5        6        7        8
 0.3333  -1.2784   0.9995  -0.4782  -1.0737   1.3192   0.0673  -0.8323

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.105      0.381   -2.90  0.00370 **
partyR           1.996      0.555    3.60  0.00032 ***
contributions    0.802      0.280    2.86  0.00421 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 43.8617  on 7  degrees of freedom
Residual deviance:  6.5633  on 5  degrees of freedom
AIC: 30.64

Number of Fisher Scoring iterations: 4
```

9. Which of the following statements is TRUE?

(zz) The model `fit.add` has a smaller AIC than `fit.int`.

(1) The output from `anova()` indicates that we should prefer the model with the interaction term.

(1) In the `glm()` functions above, a 'success' is defined as a senator voting NO on the bill.

(1) The deviance of the model `fit.int` is 0.

(1) One of the deviance residuals of the model `fit.add` is surprisingly large.

10. This question involves interpretation of the output from `summary(fit.add)`. Interpret this output, regardless of whether or not you think `fit.add` is a suitable model. For the following statement, fill in the blank (????) to give the best interpretation:

    The estimated probability of voting YES on the bill for a Republican who has US$10 000 in contributions is approximately ????.

    (zz) 0.845

    (1) 0.425

    (1) 0.709

    (1) 1.000

    (1) The `summary(fit.add)` table does not provide enough information to calculate the estimated probability.

11. We can use a case-control study to test if two factors are independent. Which of the following statements about case-control studies is TRUE?

    (zz) If the prevalence of one of the factor's levels is low, then the standard error of an odds-ratio estimate is likely to be smaller if we use case-control sampling instead of prospective sampling.

    (1) Prospective sampling results in very biased estimates of odds ratios when the prevalence of one of the factor's levels is low; case-control sampling fixes this problem.

    (1) The standard error of an odds-ratio estimate from a case-control study is biased, because we have not sampled at random from the population.

    (1) An odds-ratio estimate from a case-control study is biased, because we have not sampled at random from the population.

    (1) Case-control sampling involves sampling from the entire population completely at random.

12. Consider fitting a regression tree to data that are assumed to have a normal distribution. Let the number of terminal nodes (or ‘leaves’) be k. Which of the following statements is TRUE?

(zz) At each stage of growing a regression tree, we create a split to maximise the reduction in the residual sum of squares.

(1) The tree with the largest number of leaves is the best.

(1) A regression tree fits the expected response value as a smooth function of the covariates.

(1) In general, we can only use regression trees if we assume that the response has a normal distribution.

(1) One way to choose which tree to use is to consider a penalty function approach: we can compute $RSS + \alpha \times k$ for each candidate tree, and select the tree with the largest value of this criterion.

SECTION B

13. [**7 Marks**] Consider the topic of linear models.

(a) What are the quantities AIC, BIC and Cp used for? [2 marks]

They are used for model selection. Given k variables, these quantities measure how good the model fits the data. The best models are those with smaller values.

(b) The quantities in (a) are examples of the penalty function approach. Describe this approach by writing down the single underlying formula behind them all and briefly explaining the logic behind it. Remember to define any symbols that you use. [5 marks]

We balance 2 opposing quantities by solving the minimization problem

$$\min_{\boldsymbol{\theta}} \quad A + \lambda \, B \tag{1}$$

where $\boldsymbol{\theta}(\lambda)$ *is the vector of parameters to be estimated and* λ (≥ 0) *is the balancing or trade-off parameter. By 'opposing', it is meant that A and B go in opposite directions as a model becomes more complex. For the 3 quantities considered here, B is proportional to the number of parameters. And often A is a goodness of fit measure such as a scaled residual sum of squares.*

14. [**6 Marks**] Consider the topic of missing values.

(a) What is meant by an available case analysis? Give a simple example. [4 marks]

Retain the full data set and for subsequent analyses use those observations that have no missing values for the subset of variables required for that analysis. For example, if a data frame `ldata` has variables `x1` to `x10` then `lm(y ~ x1 + x2, ldata)` will only delete cases with missing values in `x1` and `x2`.

(b) A vector `x` has missing values. Write one line of R code to replace the missing values by the mean of all the non-missing `x` values. [2 marks]

```
x[is.na(x)] <- mean(x, na.rm = TRUE)
```

15. [**13 Marks**] The following data come from a medical study of the factors affecting patterns of insulin-dependent diabetes mellitus in 43 children. The purpose is to investigate the dependence of the level of serum C-peptide on various other factors in order to understand the patterns of residual insulin secretion. The response variable is the logarithm of C-peptide concentration (pmol/ml) at diagnosis, and the explanatory variables are age (in years) and base deficit (a measure of acidity).

```
> dfit <- gam(log(Cpeptide) ~ s(age) + s(baseDeficit), data = diabetes)
> plot(dfit, resid = TRUE, pch = 19)
```

This produces Figure 1.

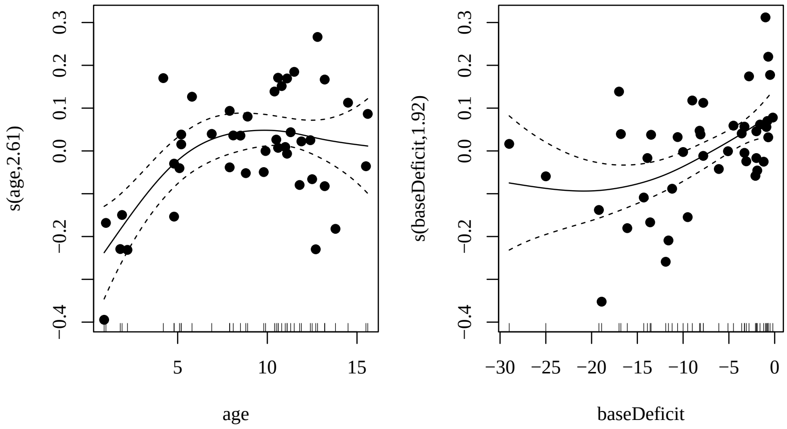


Figure 1: Component functions from an additive model fitted to the `diabetes` data. The points are the partial residuals.

(a) Comment on the function in the right-hand plot of Figure 1 with respect to influential points. [2 marks]

There are 2 influential observations on the LHS. The result of these 2 values is to swing the smoother upwards.

(b) Letting $X_1 =$ `age` and $X_2 =$ `baseDeficit`, write down the mathematical equation that `dfit` is fitting. Define any other symbols you use. [3 marks]

Let $\log Y$ *be* ***`log(Cpeptide)`****. Then we are fitting* $\log Y_i = \alpha + g_1(X_{i1}) + g_2(X_{i2}) + \varepsilon_i$ *where the errors are* $N(0,\ \sigma^2)$ *independently and* $i = 1, \ldots, 43 = n$*. The* g_k *are (centred) smooth functions estimated by a smoother such as a spline.*

(c) Suppose that the fitted intercept is 1.545 and that the smooths are centred to have mean 0 (for identifiability). Given a 3 year old child whose base deficit is -20 pmol/ml, give an approximate point estimate for his C-peptide level. [4 marks]

```
> child.df <- data.frame(age = 3, baseDeficit = -20)
> predict(dfit, child.df)  # On a log scale

     1
1.3371

> exp(predict(dfit, child.df))  # Answer (median C-peptide actually)

     1
3.8078
```

(d) A totally parametric model might be used rather than the additive model. Write down the equation of what might be a reasonable totally parametric model. [4 marks]

Function g_1 *might be proportional to* ***`sqrt(age)`****, or possibly* ***`log(age)`*** *would be good. Or possibly a quadratic in* ***`age`*** *might be okay too. Function* g_2 *might be replaced by a quadratic in* ***`baseDeficit`*** *[a linear function might be okay too]. Hence something like* $\log Y_i = \beta_0 + \beta_1 \cdot X_{i1}^{1/2} + \beta_2 \cdot X_{i2} + \beta_3 \cdot X_{i2}^2 + \varepsilon_i,\ \varepsilon_i \sim N(0,\ \sigma^2)$ *independently, might be a plausible model.*

16. [**10 Marks**] The data frame `divusa` from the faraway R package allows the divorce rates in the USA from 1920–1996 to be modelled using 6 covariates. All variables are numeric and are described as follows.

yr	the year from 1920–1996
divorce	divorce per 1000 women aged 15 or more
unem	unemployment rate
fmlb	percent female participation in labor force aged 16+
marr	marriages per 1000 unmarried women aged 16+
brth	births per 1000 women aged 15–44
milt	military personnel per 1000 population

```
> full.model <-  lm(divorce ~ ., data = divusa)
> summary(full.model)



Call:
lm(formula = divorce ~ ., data = divusa)

Residuals:
   Min     1Q Median     3Q    Max
-2.909 -0.921 -0.093  0.745  3.469

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 380.1476    99.2037    3.83  0.00027 ***
yr           -0.2031     0.0533   -3.81  0.00030 ***
unem         -0.0493     0.0538   -0.92  0.36217
fmlb          0.8079     0.1149    7.03  1.1e-09 ***
marr          0.1498     0.0238    6.29  2.4e-08 ***
brth         -0.1169     0.0147   -7.96  2.2e-11 ***
milt         -0.0428     0.0137   -3.12  0.00265 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.51 on 70 degrees of freedom
Multiple R-squared:  0.934,Adjusted R-squared:  0.929
F-statistic:  166 on 6 and 70 DF,  p-value: <2e-16


> allpossregs(full.model, Cp.plot = FALSE, dp = 2)


    rssp sigma2 adjRsq     Cp    AIC    BIC    CV yr unem fmlb marr brth milt
1 418.10   5.57   0.83 109.70 186.70 191.38 39.86  0    0    1    0    0    0
2 304.38   4.11   0.87  62.00 139.00 146.03 29.95  0    0    1    0    1    0
3 209.84   2.87   0.91  22.69  99.69 109.07 21.62  0    0    1    1    1    0
4 183.08   2.54   0.92  13.00  90.00 101.72 19.95  1    0    1    1    1    0
5 162.12   2.28   0.93   5.84  82.84  96.90 18.00  1    0    1    1    1    1
6 160.20   2.29   0.93   7.00  84.00 100.41 18.22  1    1    1    1    1    1
```

Based on this output, answer the following questions.

(a) Backward elimination applied to `full.model` would probably delete the `unem` variable first—TRUE or FALSE? Give a reason for your answer. [2 marks]

True. The ***summary()*** *p-value is large for this variable... in fact, it is the only nonsignificant variable. So it is likely to be dropped first.*

(b) Which model or models would be the best choice if one wanted to predict the divorce rate for the year 1997? Briefly give a reason for your answer. [3 marks]

The best choice is to look at the ***CV*** *column. It is minimized by the 5 variable model. An alternative is to look at the* ***adjRsq*** *column—it another predictive criterion—the 2 models at the bottom are best since this is maximized (They cannot be distinguished because only 2 decimal places are used). Incidentally, the differences in* ***adjRsq*** *between the first few models from the bottom is slight.*

(c) Suppose I want to choose a model to explain what the data says to a sociologist colleague. Which model or models would be the best choice? Briefly give a reason for your answer. [3 marks]

The AIC, BIC, etc. criteria all suggest that the 5 covariate model (2nd from the bottom) is best. These criteria are to be minimized.

(d) Suppose the model

```
model2 <-  lm(divorce ~ poly(yr, 2) + poly(unem, 2) +
                        poly(fmlb, 2) + poly(marr, 2) +
                        poly(brth, 2) + poly(milt, 2),
               data = divusa)
```

was fitted. Then what is the output from typing `length(coef(model2))`? [2 marks]

Answer:

```
> length(coef(model2))

[1] 13
```

17. [**20 Marks**]

(a) What are the assumptions of a Poisson regression model? [2 marks]

Answer: (1) The log of the expected response is a linear combination of the explanatory variables. (2) Each observed response is a Poisson random variable. (3) The responses are independent.

The rest of this question refers to data that come from a study investigating a particular type of minor damage caused by waves to the forward sections of ships' hulls. In total, 60 ships were inspected for hull damage, and the number of damage incidents were recorded from each. Hull construction engineers are interested in determining if the design of the hull is related to the number of observed damage incidents. Hull designs vary across manufacturers, and potentially improve from year to year. The variables recorded are as follows:

incidents The number of damage incidents detected on the boat.

company The company that constructed the boat; either A, B, C, or D.

year The year of construction; either 8, 9, or 10, representing 2008, 2009, and 2010, respectively.

service The number of months the boat had been in service.

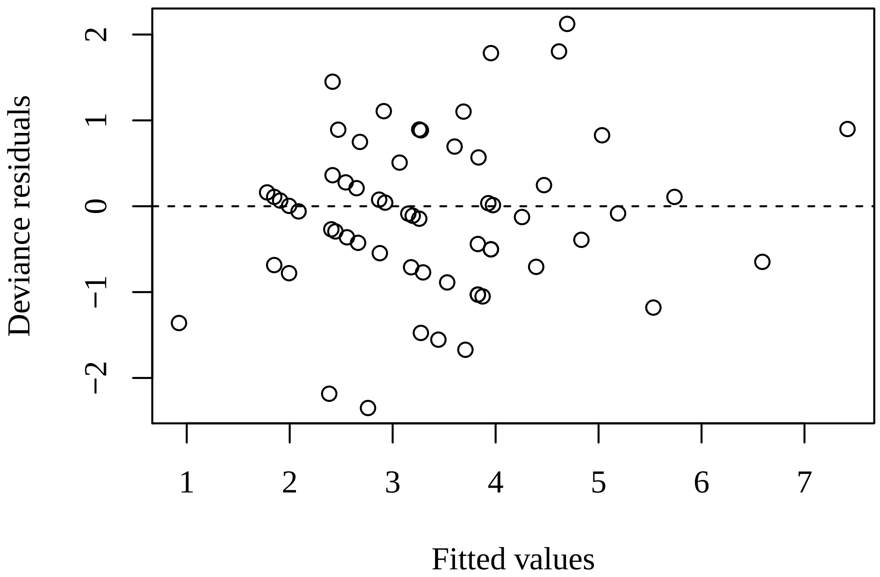
The data are stored in the data frame `ship.df`. Below is some R code to analyse the data.

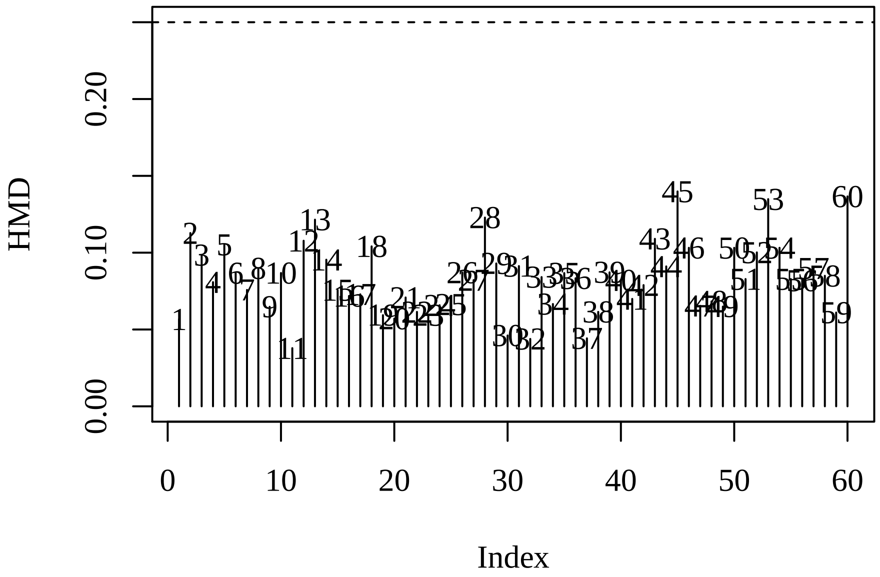
```
> head(ship.df, 10)

   incidents year company service
1          2    8       A       7
2          4    8       B      13
3          2    8       C      13
4          2    8       D       9
5          2    8       A      13
6          1    8       B      10
7          2    8       C      10
8          3    8       D      10
9          0    8       A       8
10         5    8       B      10

> fit.ship.1 <- glm(incidents ~ company * as.factor(year),
                    offset = log(service), family = "poisson",
                    data = ship.df)
> anova(fit.ship.1, test = "Chisq")

Analysis of Deviance Table

Model: poisson, link: log

Response: incidents
```

```
Terms added sequentially (first to last)


                        Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                                       59       65.4
company                  3     9.96        56       55.5    0.019 *
as.factor(year)          2     6.00        54       49.5    0.050 *
company:as.factor(year)  6     3.05        48       46.4    0.802
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> fit.ship.2 <- glm(incidents ~ company + as.factor(year),
                    offset = log(service), family = "poisson",
                    data = ship.df)
> fit.ship.3 <- glm(incidents ~ company + year,
                    offset = log(service), family = "poisson",
                    data = ship.df)
> anova(fit.ship.3, fit.ship.2, test = "Chisq")

Analysis of Deviance Table

Model 1: incidents ~ company + year
Model 2: incidents ~ company + as.factor(year)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        55       51.1
2        54       49.5  1     1.59     0.21

> summary(fit.ship.3)


Call:
glm(formula = incidents ~ company + year, family = "poisson",
    data = ship.df, offset = log(service))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3497  -0.6901  -0.0849   0.3984   2.1242

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.2871     0.8168    0.35    0.725
companyB      0.0942     0.1919    0.49    0.624
companyC     -0.4770     0.2177   -2.19    0.028 *
companyD      0.0697     0.1848    0.38    0.706
year         -0.1872     0.0887   -2.11    0.035 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 65.425  on 59  degrees of freedom
Residual deviance: 51.066  on 55  degrees of freedom
AIC: 233.4

Number of Fisher Scoring iterations: 4
```

```
> exp(confint(fit.ship.3))

Waiting for profiling to be done...

              2.5 %  97.5 %
(Intercept) 0.26379 6.50631
companyB    0.75306 1.60107
companyC    0.40143 0.94544
companyD    0.74646 1.54338
year        0.69743 0.98764

> deviance(fit.ship.3)

[1] 51.066

> df.residual(fit.ship.3)

[1] 55

> 1 - pchisq(deviance(fit.ship.3), df.residual(fit.ship.3))

[1] 0.62566
```

```
> plot(fitted(fit.ship.3), residuals(fit.ship.3, type = "deviance"),
       xlab = "Fitted values", ylab = "Deviance residuals")
> abline(h = 0, lty = 2)
> HMD <- hatvalues(fit.ship.3)
> big.HMD <- 3*5/60
> plot(HMD, type = "h", ylim = c(0, big.HMD))
> text(HMD)
> abline(h = big.HMD, lty = 2)
```





(b) Write down an expression for the expected number of damage incidents modelled by `fit.ship.3` (i.e., $\mu = \cdots$). Make sure to define any notation you use. [3 marks]

Answer: $\mu = \exp(\beta_0 + \beta_1 \times C_B + \beta_2 \times C_C + \beta_3 \times C_D + \beta_4 \times y + \log(s))$, *where* C_J *is a dummy variable for Company* J, y *is the year of construction, and*

s is the length of service in months.

(c) We sometimes use offsets when we fit generalised linear models. What is an offset? Why have we used `offset = log(service)` above? [3 marks]

Answer: An offset is a variable added to the linear predictor, but its coefficient is fixed at 1. Offsets are particularly useful for Poisson regression models when we are concerned with rates of occurance of some event, for example, per unit space or time, or per capita. In this case, it is sensible to assume that the expected number of damage incidents is proportional to the time the boat was in service: keeping all else constant, a boat in service for twice as long is likely to have twice as many damage incidents. Setting the offset $\log(s)$ *achieves this:*

$$\mu = \exp(\eta + \log(s)) = \exp(\eta) \times s, \tag{2}$$

and so now $\exp(\eta)$ *provides the damage incident rate per month, where* η *is the rest of the linear predictor without the offset term.*

(d) Consider the choice between the models `fit.ship.2` and `fit.ship.3`. What are the advantages of fitting year as a numeric variable? What are the advantages of turning it into a factor? [3 marks]

Answer: If we fit year as a numeric variable then we only need a single parameter to model its effect. We can also make predictions of damage incident rates for years other than those in our data (although extrapolating beyond the range of our data is dangerous). If we fit year as a factor, we no longer need to assume that year has a linear relationship with the log of the damage incidence rate. However, we must estimate an additional parameter, and we can only estimate/predict damage incidence rates for the years we have in our data.

(e) Comment on the adequacy of the model `fit.ship.3`, based on the description of the data and what you can see in the output and plots above. [3 marks]

Answer: It is unlikely that there are any problems with the independence assumption; damage to one boat is unlikely to be related to damage to any other. The model deviance does not provide any evidence to suggest lack of fit. The hat matrix diagonals do not suggest that there are any points with high leverage. The residual plot does not show any stong pattern. There are three large residuals with magnitudes of greater than two, although we would probably expect this given that we have 60 observations. We could possibly try removing these points, but this is unlikely to change much due to their low leverage. Overall, based on what we can tell from the output, the model looks appropriate.

(f) Is there evidence to suggest that newer boats have lower damage incident rates? Interpret the effect of the year of manufacture as estimated by the model `fit.ship.3`. [3 marks]

Answer: Yes, there is evidence to suggest that newer boats have lower damage incident rates; perhaps this could be due to improved hull designs. Holding all else constant and with 95% confidence, each one year increase in the date of manufacture is associated with a decrease in the expected number of damage incidents per month of between 1.2% and 30.2%. Note that students may also make a multiplicative interpretation: each one year increase in the data of manufacture multiplies the expected number of damage incidents by between 0.697 and 0.988.

(g) Let β_{2009} and β_{2010} be the coefficients of the dummy variables for years 2009 and 2010, respectively, in the model `fit.ship.2`. In the `anova(fit.ship.3, fit.ship.2, test = "Chisq")` output above, what null hypothesis is the p-value testing? Write your answer in terms of β_{2009} and β_{2010}. [3 marks]

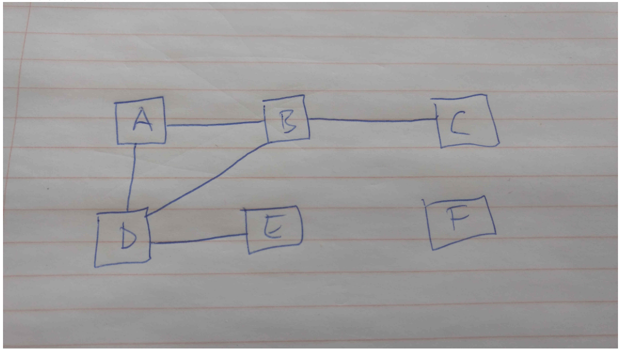
Answer: The model `fit.ship.3` is a submodel of `fit.ship.2` for which $\beta_{2010} - \beta_{2009} = \beta_{2009}$, or $\beta_{2010} = 2\beta_{2009}$. That is, `fit.ship.3` restricts the difference in the log of expected damage incidents between the years 2010 and 2009 to be the same as the difference between 2009 and 2008. Thus, the null hypothesis tested by the `anova()` command is $H_0 : \beta_{2010} = 2\beta_{2009}$.

18. **[20 Marks]** Consider the analysis of a six-dimensional contingency table, with observations cross-classified by factors A, B, C, D, E, and F. Let the number of observations falling into each unique combinaton of levels be stored in the vector `counts`. Say the following code is used to fit a Poisson regression model to these data:

```
> fit <- glm(counts ~ A * B * D + D * E + C * B + F, family = "poisson")
```

(a) Sketch the association graph associated with this model. [2 marks]

Answer:



(b) Describe the relationship between the following pairs of factors under this model: [2 marks]

(i) A and F.

Answer: Factors A and F are independent.

(ii) C and E.

Answer: Factors C and E are conditionally independent, given B and D.

(c) Assume the model above is correct. It is of interest to investigate the relationship between factors B and E, and in doing so we wish to simplify the contingency table by collapsing over another factor. [2 marks]

(i) Is it appropriate to collapse over factor D?

Answer: No. Factor D is a confounder. Its interactions with both B and E are nonzero. We should hold D constant when investigating the relationship between B and E.

(ii) Is it appropriate to collapse over factor C?

Answer: Yes. The C:E interaction is zero, so C is not a confounder for B and E.

The rest of this question involves the following data set. A sample of 4295 soldiers who fought in the American Civil War were cross-classified by the following factors:

Rank Either ‘Private’ or ‘Higher Rank’.

Infantry Either ‘Yes’ or ‘No’, indicating whether or not the soldier was in the infantry.

Fate This is ‘Survived’ if the soldier survived, otherwise ‘Illness’, ‘Injury’, or ‘Other’, indicating the cause of death.

The data were loaded into R. The data set includes the column `counts`, which gives the total number of soldiers with each unique combination of the above factors. The data were analysed using the following code:

```
> infantry.df

      rank infantry     fate counts
1  private      yes survived   2367
2  private       no survived   1124
3   higher      yes survived    262
4   higher       no survived     95
5  private      yes  illness    285
6  private       no  illness     50
7   higher      yes  illness     13
8   higher       no  illness      2
9  private      yes   injury     60
10 private       no   injury     15
11  higher      yes   injury      7
12  higher       no   injury      4
13 private      yes    other      8
14 private       no    other      2
15  higher      yes    other      1
16  higher       no    other      0

> fit.war.1 <- glm(counts ˜ rank*infantry*fate, family = "poisson",
                    data = infantry.df)
> fit.war.2 <- glm(counts ˜ (rank + infantry + fate)ˆ2,
                    family = "poisson", data = infantry.df)
> fit.war.3 <- glm(counts ˜ rank*fate + infantry*fate,
                    family = "poisson", data = infantry.df)
```

```
> anova(fit.war.1, test = "Chisq")

Analysis of Deviance Table

Model: poisson, link: log

Response: counts

Terms added sequentially (first to last)


                   Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                                 15      12644
rank                1     3367        14       9276  < 2e-16 ***
infantry            1      701        13       8576  < 2e-16 ***
fate                3     8503        10         72  < 2e-16 ***
rank:infantry       1        3         9         69   0.0869 .
rank:fate           3       13         6         56   0.0041 **
infantry:fate       3       53         3          3  1.6e-11 ***
rank:infantry:fate  3        3         0          0   0.4551
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> AIC(fit.war.1, fit.war.2, fit.war.3)

          df    AIC
fit.war.1 16 110.66
fit.war.2 13 107.27
fit.war.3 12 109.27

> summary(fit.war.2)


Call:
glm(formula = counts ~ (rank + infantry + fate)^2, family = "poisson",
    data = infantry.df)

Deviance Residuals:
      1        2        3        4        5        6        7        8
-0.0404   0.0587   0.1218  -0.2004   0.0103  -0.0244  -0.0478   0.1261
      9       10       11       12       13       14       15       16
 0.2538  -0.4819  -0.6762   1.1989  -0.0532   0.1100   0.1599  -0.5504

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)                4.5744     0.1000   45.76  < 2e-16 ***
rankprivate                2.4485     0.1041   23.53  < 2e-16 ***
infantryyes                0.9864     0.1164    8.47  < 2e-16 ***
fateillness               -3.9717     0.2967  -13.39  < 2e-16 ***
fateinjury                -3.8539     0.3730  -10.33  < 2e-16 ***
fateother                 -6.4616     1.2016   -5.38  7.6e-08 ***
rankprivate:infantryyes   -0.2391     0.1215   -1.97   0.0490 *
rankprivate:fateillness    0.8643     0.2704    3.20   0.0014 **
rankprivate:fateinjury    -0.3391     0.3279   -1.03   0.3012
rankprivate:fateother      0.0531     1.0507    0.05   0.9597
infantryyes:fateillness    0.9891     0.1544    6.41  1.5e-10 ***
```

```
infantryyes:fateinjury    0.4841     0.2624    1.85   0.0650 .
infantryyes:fateother     0.7365     0.7827    0.94   0.3467
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 12643.6309  on 15  degrees of freedom
Residual deviance:     2.6137  on  3  degrees of freedom
AIC: 107.3

Number of Fisher Scoring iterations: 4

> 1 - pchisq(deviance(fit.war.2), df.residual(fit.war.2))

[1] 0.45509
```

(d) We have selected the model `fit.war.2` here. What steps lead to this decision? [3 marks]

Answer: There was no evidence to suggest a three-way interaction, so it was dropped. We tried dropping the ***`rank:infantry`*** *interaction, but this worsened the AIC. The deviance of* ***`fit.war.2`*** *does not provide any evidence for lack-of-fit.*

(e) The type of model fitted in `fit.war.2` has a common name. What is it? Briefly explain what this implies regarding the relationships between rank, infantry status, and fate. [3 marks]

Answer: This is the homogeneous association model. None of the variables are independent of any other; however, odds ratios for any pair of the factors do not depend on the level of the third.

(f) From the output above, how do the estimated odds of a private dying from illness rather than surviving compare to that of a soldier of a higher rank? Provide and interpret a 95% confidence interval to aid your explanation. [4 marks]

Answer: A 95% confidence interval for the relevant odds ratio is given by $\exp(0.8643 \pm 1.96 \times 0.2704) = (1.40, 4.03)$. *The interpretation is as follows: With 95% confidence, we estimate that the odds of a private dying of illness rather than surviving are between 1.40 and 4.03 times those of a soldier of a higher rank. Students may also give a percentage interpretation, with a CI of (40, 303)%.*

(g) How do the odds of death from illness rather than survival compare across infantry and non-infantry soldiers? Make sure to interpret the relevent coefficient from the output. There is no need to calculate a 95% confidence interval for this question. [2 marks]

Answer: There is strong evidence to suggest that the odds of death from illness rather than survival are different for infantry and non-infantry soldiers. We estimate that the odds of death from illness rather than survival for infantrymen are about 2.68 times that of soldiers who are not in the infantry. Students may instead give a percentage interpretation: the odds of death from illness rather than survival for infantrymen are about 168% higher than soldiers who are not in the infantry. They may also give the reverse interpretations: the odds of death from illness rather than survival for soldiers not in the infantry are 0.37 times that of soldiers who are in the infantry, and so on.

(h) Note that some of the cell counts are small—particularly for the '`other`' level of the `fate` variable. Why might this be concerning? How might we investigate to what extent this impacts our analysis? [2 marks]

Answer: Various test statistics we have looked at (e.g., the deviance) only have an approximate chi-squared distribution if the expected cell counts are large. If this is not the case then some of our p-value calculations and subsequent interpretations may not be accurate. One way to establish how closely a test statistic is approximated by a chi-squared distribution is to simulate under the model we have fitted to get an empirical estimate of the test statistic's distribution.