THE UNIVERSITY OF AUCKLAND

SEMESTER ONE 2018
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 penalised.

SECTION B (76 marks)

- Answer all questions.

Total for both parts: 100 marks.

SECTION A

1. Suppose we fit a model and calculate a 95% confidence interval and a 95% prediction interval for an observation. The confidence interval is $(-1.1, 2.1)$ and the prediction interval is $(-0.8, 1.8)$. Which of these statements is **TRUE**?

   (zz) These intervals are not consistent with statistical theory; they have been calculated incorrectly.

   (1) If we calculate the prediction interval for a number of successive samples from the same population, it will contain the future observation 90% of the time.

   (1) If we calculate the confidence interval for a number of successive samples from the same population, it will contain the true mean 90% of the time.

   (1) If we calculate the prediction interval for a number of successive samples from the same population, it will contain the true mean 95% of the time.

   (1) If we calculate the confidence interval for a number of successive samples from the same population, it will contain the future observation 95% of the time.

2. Consider diagnostics for a linear model. If the constant variance assumption fails, which of the following options is a potential solution?

   (zz) Transform the response and fit the model again, using a Box-Cox plot to choose the transformation.

   (1) Delete influential points and fit the model again.

   (1) Delete the variables that are collinear and fit the model again.

   (1) Use weighted least squares, where the weights are the inverse of the coefficients.

   (1) Use the backwards elimination method.

The next two questions are based on the following scenario.

Suppose we have a response variable Y and an explanatory factor X, which has four levels. We ran the following code in R:

```
> mymodel <- lm(Y ~ X)
> plot(mymodel, which = 1:6)
```

3. We want to test whether the effect of all levels of X is the same. What is the correct code to do this in R?

(zz) `anova(mymodel)`

(1) `anova(submodel, mymodel)`

(1) `summary(mymodel)`

(1) `t.test(mymodel)`

(1) `plot(mymodel)`

4. Which plots are produced by the code?

(zz) Residuals vs Fitted, Normal Q-Q, Scale-Location, Cook's distance, Residuals vs Leverage, Cook's distance vs Leverage.

(1) Residuals vs Fitted, Normal Q-Q, Scale-Location, Residuals vs Leverage.

(1) Residuals vs Fitted, Normal Q-Q, Scale-Location vs Cook's distance, Residuals vs Fitted, Cook's distance vs Leverage.

(1) Residuals vs Fitted, Normal Q-Q, Scale-Location, Cook's distance, Residuals vs Leverage, Fitted values.

(1) Residuals, Normal Q-Q, Scale-Location, Cook's distance, Leverage, Fitted values.

The next two questions are based on the following analysis.

Blackburn Rovers is a football club based in Lancashire, England. One of their fans cross-classified all their league matches over the last five seasons by result (Win, Draw, or Loss) and match location (Home or Away). A ‘Home’ match is played at Ewood Park in Blackburn, while an ‘Away’ match is played at the opposition’s football ground. The data are shown in the following contingency table:

	Win	Draw	Loss
Home	53	35	27
Away	35	40	40

The following code was used to analyse these data:

```
> blackburn.df


  result location count
1    Win     Home    53
2   Draw     Home    35
3   Loss     Home    27
4    Win     Away    35
5   Draw     Away    40
6   Loss     Away    40


> blackburn.fit <- glm(count ˜ result*location, family = "poisson",
                       data = blackburn.df)
> anova(blackburn.fit, 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       9.49
result           2     2.91         3       6.58    0.234
location         1     0.00         2       6.58    1.000
result:location  2     6.58         0       0.00    0.037 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> summary(blackburn.fit)


Call:
glm(formula = count ~ result * location, family = "poisson",
    data = blackburn.df)

Deviance Residuals:
[1]  0  0  0  0  0  0

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)              3.69e+00   1.58e-01   23.33   <2e-16 ***
resultDraw              -6.25e-17   2.24e-01    0.00    1.000
resultWin               -1.34e-01   2.31e-01   -0.58    0.564
locationHome            -3.93e-01   2.49e-01   -1.58    0.115
resultDraw:locationHome  2.60e-01   3.40e-01    0.76    0.445
resultWin:locationHome   8.08e-01   3.31e-01    2.44    0.015 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance:  9.4884e+00  on 5  degrees of freedom
Residual deviance: -1.7764e-15  on 0  degrees of freedom
AIC: 44.81

Number of Fisher Scoring iterations: 3


> coef(blackburn.fit)


            (Intercept)              resultDraw               resultWin
             3.6889e+00             -6.2465e-17             -1.3353e-01
           locationHome resultDraw:locationHome  resultWin:locationHome
            -3.9304e-01              2.5951e-01              8.0799e-01


> exp(coef(blackburn.fit))


            (Intercept)              resultDraw               resultWin
                40.0000                  1.0000                  0.8750
           locationHome resultDraw:locationHome  resultWin:locationHome
                 0.6750                  1.2963                  2.2434
```

5. Which of the following statements is **FALSE**?

   (zz) We estimate that the odds of Blackburn Rovers winning are approximately 0.88 times the odds of them losing.

   (1) There is evidence of an association between the result and the location of a Blackburn Rovers match.

   (1) We estimate that the odds of Blackburn Rovers winning rather than losing are approximately 124% higher if they are playing at home than they are if they are playing away.

   (1) We estimate that the log-odds of Blackburn Rovers winning rather than losing if they are playing at home are approximately 0.81 higher than the log-odds of them winning rather than losing if they are playing away.

   (1) We estimate that the odds of Blackburn Rovers winning rather than losing if they are playing at home are approximately 2.24 times the odds of them winning rather than losing if they are playing away.

6. Which of the following statements is **TRUE**?

   (zz) The residual deviance of the model `blackburn.fit` is zero because we have fitted the saturated model.

   (1) We should drop the `result:location` interaction because the *p*-value for `resultDraw:locationHome` is large.

   (1) The null deviance of the model `blackburn.fit` is smaller than the residual deviance.

   (1) If we dropped the `result:location` interaction, the resulting model would probably have a smaller residual deviance than the model `blackburn.fit`.

   (1) We probably should not trust this analysis, because some of the cells have counts that are too small.

7. Which of the following statements about logistic regression with a logit link function is **FALSE**?

(zz) We assume that the variance of the response variable is constant across all observations.

(1) We assume that the log-odds of a trial being successful is a linear combination of the explanatory variables.

(1) We assume that the observations are independent.

(1) We assume that the response variable comes from a binomial distribution.

(1) By default, R uses the logit link function when a logistic regression model is fitted.

8. A Poisson regression model with a log link funcion was fitted to a response variable Y, using only a single numeric expanatory variable, X. Estimates of the linear predictor's intercept, β_0, and slope, β_1, were obtained. Which of the following is an appropriate interpretation?

(zz) For every one-unit increase in X, we estimate that the expected value of Y is multiplied by $\exp(\hat{\beta}_1)$.

(1) When $x = 0$, we estimate that the expected value of Y is β_0.

(1) For every one-unit increase in X, we estimate that the odds of success are multiplied by $\exp(\hat{\beta}_1)$.

(1) When $x = 0$, we estimate that the odds of success are equal to $\exp(\hat{\beta}_0)/(1+\exp(\hat{\beta}_0))$.

(1) For every one-unit increase in X, we estimate that the expected value of Y increases by $\hat{\beta}_1$.

9. Which of the following statements about the use of offsets in generalised linear models is **FALSE**?

(zz) We estimate a coefficient for an offset, which allows us to interpret the relationship between the offset and the response variable.

(1) An offset adds a fixed value to the linear predictor for each observation.

(1) When we fit a Poisson regression model with a log link function in R, we can use the argument `offset = log(t)` if we think the expected value of the response is directly proportional to the variable `t`.

(1) While offsets are most common for Poisson regression models, we can use them for logistic or standard linear regression models, too.

(1) An explanatory variable's estimated coefficient is very close to 1. The fitted values from our model are unlikely to change much if we use the variable as an offset instead.

10. Which of the following statements about estimated coefficients of linear models (LMs) and generalised linear models (GLMs) is **FALSE**?

(zz) The estimated coefficients of a GLM maximise the deviance.

(1) The estimated coefficients of a LM minimise the residual sum of squares.

(1) The estimated coefficients of a GLM maximise the likelihood.

(1) The estimated coefficients of a GLM maximise the log-likelihood.

(1) The estimated coefficients of a GLM minimise the sum of the squared deviance residuals.

11. Consider a logistic regression model fitted to ungrouped data. The observed responses from the first two observations are given by y_1 and y_2. The first was observed as a success ($y_1 = 1$), and has a fitted probability under the model of $\hat{p}_1 = 0.8$. The second was observed as a failure ($y_2 = 0$), and has a fitted probability under the model of $\hat{p}_2 = 0.7$. Which of the following statements is **TRUE**?

    (zz) The Pearson residual of the second observation is larger in magnitude (i.e., further from zero) than the Pearson residual of the first observation.

    (1) The further an observed value is from its expected value, the closer to zero the deviance residual is.

    (1) The deviance residual of the first observation is negative, and the deviance residual of the second observation is positive.

    (1) Under the fitted model, the expected value of the first observation is 0.2, and the expected value of the second observation is 0.3.

    (1) It is possible for an observation to have a positive deviance residual, but a negative deviance residual.

12. Which of the following statements is **FALSE**?

    (zz) The null deviance is always equal to or smaller than the residual deviance.

    (1) The log-likelihood of the saturated model is always equal to or larger than the log-likelihood of the fitted model.

    (1) The residual deviance is equal to twice the difference between the log-likelihood of the saturated model and the log-likelihood of the fitted model.

    (1) The residual deviance is equal to the sum of the squared deviance residuals.

    (1) The residual deviance of the saturated model is always equal to zero.

SECTION B

13. [**8 marks**] Guess the analysis: choose the most appropriate model to fit for each of the scenarios described below. Different scenarios may have the same answer.

For each scenario, select one of these three possible answers:

(1) Linear regression model

(2) Logistic regression model

(3) Poisson regression model

(a) TVNZ has just released a new TV show. Their market analyst wishes to build a model to predict whether or not specific individuals will enjoy the show. They conduct a survey, collecting variables from participants such as gender, age, income, and occupation. They also asked participants if they enjoyed a pilot episode of the show.

[2 marks]

Logistic regression model.

(b) A STATS 330 lecturer is interested to see if the number of questions posted on Piazza is related to the closeness of an assignment deadline. Each day, they count the number of Piazza questions that were posted, and record the number of days until the next assignment deadline.

[2 marks]

Poisson regression model.

(c) A detective wishes to determine whether or not there is an association between a serial killer's gender (female or male) and their preferred method (poisoning, strangulation, and so on). They cross-classify a sample of convicted serial killers using these two variables.

[2 marks]

When I wrote this question, I intended the answer to be 'Poisson regression model'; however, after submitting the exam I realised that a logistic regression model would work fine too. Both 'Poisson regression model' and 'Logistic regression model' were marked correct.

(d) A University of Auckland empolyee wishes to determine the quickest way to get to work. Each day, they randomly select a transportation method (bus, train, or walk) and a departure time (8:00am, 8:15am, or 8:30am). They record how long their journey took.

[2 marks]

Linear regression model.

14. [**4 marks**] Suppose we wish to predict the expenditure on cancer treatment for a patient at the Auckland hospital. We have collected the following variables from a sample of cancer patients:

    - Stage of cancer
    - Type of cancer
    - Cancer treatment expenditure
    - Age
    - Gender
    - Ethnicity
    - Marital status

    (a) Is it possible to build a predictive model with this information? Explain your answer.

    [2 marks]

    (b) We also want to build an explanatory model to investigate if dietary habits cause cancer. Can we fit a model to do so using only the variables above? Explain your answer.

    [2 marks]

15. [**17 marks**] The data for this question were collected from a sample of 44 male and 51 female athletes at the Australian Institute of Sport. The data set contains the following variables:

sex The athlete's sex, either `female` or `male`.

sport The athlete's sport, either basketball (`BBal`), rowing (`Row`), swimming (`Swim`), or tennis (`Tennis`).

BMI The athletes body mass index, calculated by dividing their weight (in kg) by their height (in m) squared.

X.Bfat The athlete's body fat percentage.

Printed below are the first three observations, and summary statistics for each of the variables:

```
> head(sport.df, 3)


     sex sport   BMI X.Bfat
1 female BBall 20.56  19.75
2 female BBall 20.67  21.30
3 female BBall 21.86  19.88
```

```
> summary(sport.df)


     sex          sport          BMI            X.Bfat     
 female:51   BBall :25   Min.   :17.1   Min.   : 6.16  
 male  :44   Row   :37   1st Qu.:21.3   1st Qu.: 8.92  
             Swim  :22   Median :22.7   Median :12.20  
             Tennis:11   Mean   :22.8   Mean   :13.91  
                         3rd Qu.:24.0   3rd Qu.:18.62  
                         Max.   :26.8   Max.   :28.83  
```

We analysed these data using the following code:

```
> sport.fit <- lm(log(X.Bfat) ~ BMI + sport*sex, data = sport.df)
> reg <- allpossregs(sport.fit)[, -c(1, 2, 3, 4)]
> reg


     AIC    BIC    CV BMI Row Swim Tennis male Row:male Swim:male Tennis:male
1 161.12 166.23 0.467   0   0    0      0    1        0         0           0
2 130.05 137.71 0.380   1   0    0      0    1        0         0           0
3 112.36 122.58 0.330   1   0    1      0    1        0         0           0
4 107.89 120.66 0.315   1   0    1      0    1        0         1           0
5 106.14 121.46 0.313   1   0    1      1    1        0         1           0
6 102.21 120.09 0.303   1   0    1      1    1        0         1           1
7 102.84 123.28 0.306   1   1    1      1    1        0         1           1
8 104.00 126.98 0.308   1   1    1      1    1        1         1           1
```

```
> sport.fit.2 <- lm(log(X.Bfat) ~ sport, data = sport.df)
> anova(sport.fit.2)


Analysis of Variance Table

Response: log(X.Bfat)
          Df Sum Sq Mean Sq F value Pr(>F)
sport      3   1.87   0.625    3.91  0.011 *
Residuals 91  14.54   0.160
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> summary(sport.fit.2)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.5956     0.0799   32.47   <2e-16 ***
sportRow      0.0752     0.1035    0.73    0.469
sportSwim    -0.2835     0.1168   -2.43    0.017 *
sportTennis  -0.1043     0.1446   -0.72    0.472
---
Residual standard error: 0.4 on 91 degrees of freedom
Multiple R-squared:  0.114,Adjusted R-squared:  0.085
F-statistic: 3.91 on 3 and 91 DF,  p-value: 0.0112
```

(a) Write down the mathematical formula for the model `sport.fit`. This should describe the relationship between the explanatory variables and the response, and show its assumptions.

[3 marks]

(b) Based on the results from the `allposregs()` function, which model will return the lowest prediction error? Write down its mathematical formula and the R code that could be used to fit this model.

[3 marks]

(c) Consider the model `sport.fit.2` and the output from the `anova()` function, shown above. What is the null hypothesis, or what are the null hypotheses, associated with the p-value(s) in this output?

[3 marks]

(d) What do you conclude about the hypothesis (or hypotheses) from the output of the `anova()` function?

[2 marks]

(e) Consider the model `sport.fit.2` and the output from the `summary()` function, shown above. What is the null hypothesis, or what are the null hypotheses, associated with the p-value(s) in this output?

[3 marks]

(f) What do you conclude about the hypothesis (or hypotheses) from the output of the `summary()` function?

[3 marks]

16. [**18 marks**] The data for this question are related to a sample of 1599 Portugese red wines. Various physiochemical properties of the wines were measured. Additionally, a panel of judges decided whether or not each wine was of 'good quality'. The data set contains the following variables:

good.quality This variable takes the value 1 if the wine is of 'good quality', and 0 otherwise.

fixed.acidity The fixed concentration of tartaric acid (g per dm^3).

volatile.acidity The volatile concentration of tartaric acid (g per dm^3).

residual.sugar The concentration of residual sugars (g per dm^3).

chlorides The concentration of sodium chloride (g per dm^3).

f.sulfur.dioxide The concentration of free sulfur dioxide (mg per dm^3).

density The density of the wine (g per cm^3).

sulphates The concentration of potassium sulphate (g per dm^3).

alcohol The alcohol level of the wine (percentage alcohol by volume).

The following final model was fitted in R:

```
> wine.fit <- glm(good.quality ~ fixed.acidity + volatile.acidity +
                      residual.sugar + chlorides + t.sulfur.dioxide +
                      I(t.sulfur.dioxide^2) + density + sulphates +
                      I(sulphates^2) + alcohol + I(alcohol^2),
                  family = "binomial", data = wine.df)
```

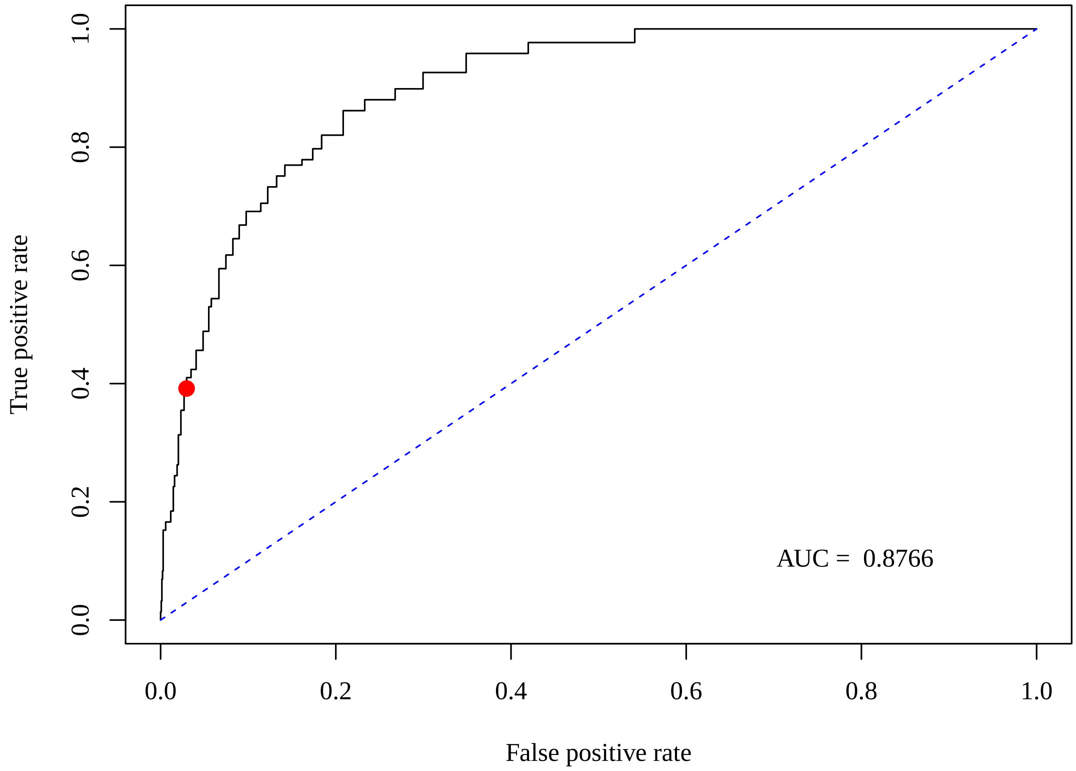
```
> summary(wine.fit)
Call:
glm(formula = good.quality ~ fixed.acidity + volatile.acidity +

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            2.44e+02   1.01e+02    2.42  0.01554 *
fixed.acidity          2.68e-01   8.48e-02    3.17  0.00154 **
volatile.acidity      -2.28e+00   6.66e-01   -3.43  0.00061 ***
residual.sugar         2.24e-01   7.87e-02    2.85  0.00436 **
chlorides             -6.89e+00   3.57e+00   -1.93  0.05378 .
t.sulfur.dioxide      -2.85e-02   7.65e-03   -3.73  0.00020 ***
I(t.sulfur.dioxide^2)  1.17e-04   5.07e-05    2.31  0.02104 *
density               -2.93e+02   1.02e+02   -2.88  0.00399 **
sulphates              2.22e+01   5.06e+00    4.39  1.1e-05 ***
I(sulphates^2)        -1.10e+01   3.25e+00   -3.39  0.00070 ***
alcohol                5.67e+00   1.49e+00    3.81  0.00014 ***
I(alcohol^2)          -2.20e-01   6.51e-02   -3.38  0.00072 ***
---
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1269.92  on 1598  degrees of freedom
Residual deviance:  825.17  on 1587  degrees of freedom
```

```
> ROC.curve(wine.fit)

Area under ROC curve =  0.8766
```



The following code was used to predict which wines in the sample were 'good quality' wines. A probability cutoff of $c = 0.5$ was used.

```
> wine.pred <- predict(wine.fit, type = "response")
> wine.predcode <- ifelse(wine.pred < 0.5, "Not good", "Good")
```

The results of this code have been reformatted to give the following confusion matrix:

		Predicted	
		Good	Not good
Observed	Good	85	132
	Not good	41	1341

(a) Interpret the effects of the following variables on wine quality under the model `wine.fit`:

(i) Concentration of residual sugars.

[2 marks]

Holding all other variables constant, a 1 g per dm^3 increase in the concentration of residual sugars is associated with an increase in the odds of a wine being 'good quality' of between 6% and 45%.

(ii) Volatile concentration of tartaric acid.

[2 marks]

Holding all other variables constant, a 1 g per dm^3 increase in the volatile concentration of tartaric acid is associated with a decrease in the odds of a wine being 'good quality' of between 63% and 97%.

(b) Calculate the following:

(i) In-sample sensitivity.

[2 marks]

$$\frac{85}{85+132} \approx 0.39$$

(ii) In-sample specificity.

[2 marks]

$$\frac{1341}{41+1341} \approx 0.97$$

(iii) In-sample error rate.

[2 marks]

$$\frac{41+132}{85+132+41+1341} \approx 0.11$$

(c) Based on your calculations, Comment on the model's predictive power using a probability cutoff of $c = 0.5$. How well does it predict wines that are good quality? How well does it predict wines that are not good quality? [3 marks]

The model does a good job of predicting which wines are not of good quality. However, it correctly classifies less than half of all good-quality wines.

(d) A chemist wishes to use this model to predict which wines are good quality. However, they want the model to correctly predict 80% of the good-quality wines in the sample. They adjust their probability cutoff c accordingly. Using the ROC curve above, approximately what proportion of wines in the sample that are **not** of good quality will they correctly predict using this adjusted cutoff? Give your answer to one decimal place.

[2 marks]

0.8

(e) Specificity and sensitivity can also be calculated via crossvalidation using the R function `cross.val()`. Would you expect the sensitivity and specificity calculated by `cross.val()` to be higher or lower than your in-sample calculations above? Explain your answer.

[3 marks]

They would probably be lower. In-sample predictions are usually overoptimisitic, because we are testing predictive power using the observations that the model was fitted to. This leads to correct-prediction rates that are too high. Crossvalidation provides an estimate of the out-of-sample prediction rates.

17. **[29 marks]** In 2015, New Zealand's National Institute of Water and Atmospheric Research (NIWA) sampled 486 sites on rivers around the country. At each site, they recorded whether or not various freshwater species were present.

Each site was cross-classified based on these presence/absence data to form a contingency table. The data set `fishy.df` has the following variables:

eel	Presence (`1`) or absence (`0`) of the longfin eel.
koura	Presence (`1`) or absence (`0`) of the koura, a type of crayfish.
bully	Presence (`1`) or absence (`0`) of the upland bully.
trout	Presence (`1`) or absence (`0`) of the brown trout.
count	The number of sites with a particular combination of the above variables.

Freshwater biologists were interested in how the species interact. Is it common to find species at the same site? Or do some species avoid one another?

The data are shown below:

```
> fishy.df


   eel koura bully trout count
1    0     0     0     0   233
2    0     0     0     1    41
3    0     0     1     0    35
4    0     0     1     1    12
5    0     1     0     0    12
6    0     1     0     1     2
7    0     1     1     0     2
8    0     1     1     1     1
9    1     0     0     0    52
10   1     0     0     1    34
11   1     0     1     0     9
12   1     0     1     1    13
13   1     1     0     0    16
14   1     1     0     1     8
15   1     1     1     0     4
16   1     1     1     1    12
```

For example, from the first row, there were 233 sites that had none of the species present. From the fourth row, there were 12 sites at which the upland bully and the brown trout present, but the longfin eel and koura were not present.

```
> fishy.fit.1 <- glm(count ~ eel*koura*bully*trout,
                      family = "poisson", data = fishy.df)
>
> fishy.fit.2 <- glm(count ~ (eel + koura + bully + trout)^2,
                      family = "poisson", data = fishy.df)
```

```
> summary(fishy.fit.2)
```

```
Call:
glm(formula = count ~ (eel + koura + bully + trout)^2, family = "poisson",

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   5.4632     0.0642   85.04  < 2e-16 ***
eel          -1.5204     0.1449  -10.49  < 2e-16 ***
koura        -3.0946     0.2668  -11.60  < 2e-16 ***
bully        -1.9837     0.1711  -11.59  < 2e-16 ***
trout        -1.7869     0.1595  -11.21  < 2e-16 ***
eel:koura     1.8526     0.3251    5.70  1.2e-08 ***
eel:bully     0.2525     0.2749    0.92  0.35833
eel:trout     1.3429     0.2347    5.72  1.1e-08 ***
koura:bully   0.7185     0.3367    2.13  0.03286 *
koura:trout   0.0911     0.3280    0.28  0.78115
bully:trout   0.8875     0.2627    3.38  0.00073 ***
---
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 850.0360  on 15  degrees of freedom
Residual deviance:   3.0544  on  5  degrees of freedom
```

```
> fishy.fit.3 <- glm(count ~ eel + koura + bully + trout + eel:koura +
                         eel:bully + eel:trout + koura:bully + bully:trout,
                     family = "poisson", data = fishy.df)
> summary(fishy.fit.3)
```

```
Call:
glm(formula = count ~ eel + koura + bully + trout + eel:koura +

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   5.4628     0.0643   85.02  < 2e-16 ***
eel          -1.5290     0.1421  -10.76  < 2e-16 ***
koura        -3.0821     0.2626  -11.73  < 2e-16 ***
bully        -1.9869     0.1710  -11.62  < 2e-16 ***
trout        -1.7838     0.1591  -11.22  < 2e-16 ***
eel:koura     1.8773     0.3127    6.00  1.9e-09 ***
eel:bully     0.2463     0.2743    0.90  0.36920
eel:trout     1.3623     0.2241    6.08  1.2e-09 ***
koura:bully   0.7363     0.3305    2.23  0.02588 *
bully:trout   0.8959     0.2609    3.43  0.00059 ***
---
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 850.0360  on 15  degrees of freedom
Residual deviance:   3.1312  on  6  degrees of freedom
```

```
> fishy.fit.4 <- glm(count ˜ eel + koura + bully + trout + eel:koura +
                         eel:trout + koura:bully + bully:trout,
                     family = "poisson", data = fishy.df)
> summary(fishy.fit.4)
Call:
glm(formula = count ˜ eel + koura + bully + trout + eel:koura +

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   5.4569     0.0641   85.09  < 2e-16 ***
eel          -1.4937     0.1355  -11.02  < 2e-16 ***
koura        -3.1160     0.2630  -11.85  < 2e-16 ***
bully        -1.9363     0.1596  -12.13  < 2e-16 ***
trout        -1.8106     0.1583  -11.44  < 2e-16 ***
eel:koura     1.9013     0.3109    6.12  9.7e-10 ***
eel:trout     1.3934     0.2213    6.30  3.0e-10 ***
koura:bully   0.8300     0.3132    2.65  0.00804 **
bully:trout   0.9636     0.2494    3.86  0.00011 ***
---
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 850.0360  on 15  degrees of freedom
Residual deviance:   3.9273  on  7  degrees of freedom
```

```
> confint(fishy.fit.4)


Waiting for profiling to be done...


               2.5 %  97.5 %
(Intercept)  5.32863  5.5801
eel         -1.76581 -1.2340
koura       -3.66910 -2.6328
bully       -2.26073 -1.6339
trout       -2.13076 -1.5094
eel:koura    1.30797  2.5336
eel:trout    0.96218  1.8307
koura:bully  0.20057  1.4340
bully:trout  0.47172  1.4517


> AIC(fishy.fit.1, fishy.fit.2, fishy.fit.3, fishy.fit.4)


            df     AIC
fishy.fit.1 16 101.936
fishy.fit.2 11  94.990
fishy.fit.3 10  93.067
fishy.fit.4  9  91.863
```

(a) What are the assumptions of a Poisson regression model?

[2 marks]

(1) The log of the expected count is a linear combination of the explanatory variables, (2) each observed response is a Poisson random variable, and (3) the responses are independent.

(b) Write an equation to calculate the expected number of sites with a particular combination of the presence/absence variables under the model `fishy.fit.4`. Define any notation you use that is not obvious.

[2 marks]

$$\log(\mu_i) = \beta_0 + \beta_1 e_i + \beta_2 k_i + \beta_3 b_i + \beta_4 t_i + \beta_5 e_i k_i + \beta_6 e_i t_i + \beta_7 k_i b_i + \beta_8 b_i t_i.$$

Here, μ_i is the expected count for the ith combination of explanatory variables, and e_i, k_i, b_i, and t_i are dummy variables indicating presence of longfin eel, koura, upland bully, and brown trout, respectively.

(c) Consider the following code and output:

```
> anova(fishy.fit.2, fishy.fit.1, test = "Chisq")

Analysis of Deviance Table

Model 1: count ~ (eel + koura + bully + trout)^2
Model 2: count ~ eel * koura * bully * trout
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         5       3.05
2         0       0.00  5     3.05     0.69
```

What is the null hypothesis being tested here? Refer to effects estimated by the model `fishy.fit.1` in your answer.

[3 marks]

The null hypothesis is that the submodel `fishy.fit.2` is correct, and that the additional coefficients in `fishy.fit.1` are all zero. In other words, the null hypothesis is that there are no three- or four-way interactions.

(d) What can you conclude from the hypothesis test conducted in question (c)?

[2 marks]

The p-value is large, so the null hypothesis is plausible. There is no evidence to suggest there are three- or four-way interactions. We should accept the simpler model `fishy.fit.2` over the saturated model `fish.fit.1`.

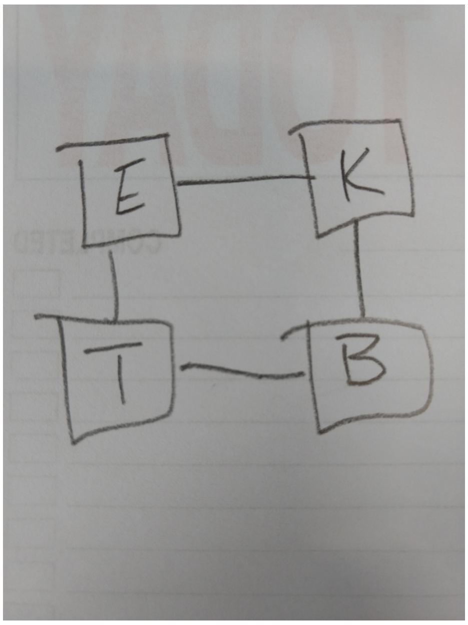
(e) The model `fishy.fit.2` was simplified to `fishy.fit.3`, and then further simplified to `fishy.fit.4`. Briefly state why you think that these were sensible decisions.

[2 marks]

The terms that were dropped were associated with large p-values. Additionally, the AIC improved each time we dropped a term.

(f) Sketch the association graph for the model `fishy.fit.4`.

[3 marks]



Where E, K, T, and B represent the factors `eel`, `koura`, `trout`, and `bully`, respectively.

(g) Describe the relationship between the following pairs of factors under the model `fishy.fit.4`. For each pair, select one of these three possible answers:

(1) Independent

(2) Conditionally independent given other factors

(3) Dependent

If you select option (2), state which other factor(s) the independence is conditional upon. Both pairs may have the same answer.

(i) `bully` and `trout`

[2 marks]

(i) The factors `bully` and `trout` are dependent.

(ii) `bully` and `eel`

[2 marks]

(ii) The factors `bully` and `eel` are conditionally independent, given `koura` and `trout`.

(h) Assume the model `fishy.fit.4` is the correct model. A freshwater biologist is interested in the association between presence of the koura and presence of the brown trout. They wish to simplify the contingency table by collapsing over another factor.

(i) Is it appropriate to collapse over the factor `eel`? Briefly explain your answer.

[2 marks]

(i) No. The factor `eel` has an interaction with both `koura` and `trout`, so it is a confounding variable.

(ii) Is it appropriate to collapse over the factor `bully`? Briefly explain your answer.

[2 marks]

(ii) No. As above.

(i) The freshwater biologist believes that the longfin eel and the brown trout avoid one another. In other words, holding all other variables constant, sites with longfin eel are less likely to have brown trout present than sites without longfin eel. Does the analysis above suggest that the biologist's belief is correct? Explain your answer.

[3 marks]

No. In fact, there is a significant *positive* association between presence of longfin and eel and presence of brown trout, providing evidence for the *opposite* of the biologist's belief.

(j) Provide a 95% confidence interval for the odds ratio that quantifies the association between the presence of koura and the presence of the upland bully, holding all other variables constant.

[2 marks]

A 95% confidence interval for the log of the odds-ratio is $(0.20, 1.43)$, and so a 95% confidence interval for the odds-ratio is $(1.22, 4.20)$.

(k) Write a sentence interpreting your confidence interval from question (j).

[2 marks]

Holding the other factors constant, the odds of koura being present at sites with upland bully are between 1.22 and 4.20 times as high as they are at sites without upland bully. (It is fine to switch "koura" and "upland bully" in this interpretation.)