Part II

Statistical Modelling

Year
2019
2018
2017
2016
2015
2014
2013
2012
2011
2010
2009
2008
2007
2006
2005
In a normal linear model with design matrix $X \in \mathbb{R}^{n \times p}$, output variables $y \in \mathbb{R}^n$ and parameters $\beta \in \mathbb{R}^p$ and $\sigma^2 > 0$, define a $(1 - \alpha)$-level prediction interval for a new observation with input variables $x^* \in \mathbb{R}^p$. Derive an explicit formula for the interval, proving that it satisfies the properties required by the definition.

[You may assume that the maximum likelihood estimator $\hat{\beta}$ is independent of $\sigma^{-2}\|y - X\hat{\beta}\|_2^2$, which has a $\chi^2_{n-p}$ distribution.]

(a) For a given model with likelihood $L(\beta), \beta \in \mathbb{R}^p$, define the Fisher information matrix in terms of the Hessian of the log-likelihood.

Consider a generalised linear model with design matrix $X \in \mathbb{R}^{n \times p}$, output variables $y \in \mathbb{R}^n$, a bijective link function, mean parameters $\mu = (\mu_1, \ldots, \mu_n)$ and dispersion parameters $\sigma^2_1 = \cdots = \sigma^2_n = \sigma^2 > 0$. Assume $\sigma^2$ is known.

(b) State the form of the log-likelihood.

(c) For the canonical link, show that the Fisher information matrix is equal to

$$\sigma^{-2}X^TWX,$$

for a diagonal matrix $W$ depending on the means $\mu$. Compute the entries of $W$ in terms of $\mu$. 

Part II, 2019 List of Questions
5J Statistical Modelling

The cycling data frame contains the results of a study on the effects of cycling to work among 1,000 participants with asthma, a respiratory illness. Half of the participants, chosen uniformly at random, received a monetary incentive to cycle to work, and the other half did not. The variables in the data frame are:

- miles: the average number of miles cycled per week
- episodes: the number of asthma episodes experienced during the study
- incentive: whether or not a monetary incentive to cycle was given
- history: the number of asthma episodes in the year preceding the study

Consider the R code below and its abbreviated output.

```r
> lm.1 = lm(episodes ~ miles + history, data=cycling)
> summary(lm.1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)  
(Intercept)   0.6694    0.0796   8.404  < 2e-16 ***
miles         -0.0492    0.0184  -2.674  0.0076 **
history       1.4895    0.0482  30.918  < 2e-16 ***

> lm.2 = lm(episodes ~ incentive + history, data=cycling)
> summary(lm.2)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)  
(Intercept)   0.0954    0.0696   1.371  0.171
incentiveYes  0.9139    0.0650  14.051  < 2e-16 ***
history       1.4681    0.0435  33.782  < 2e-16 ***

> lm.3 = lm(miles ~ incentive + history, data=cycling)
> summary(lm.3)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)  
(Intercept)  1.4705    0.1168  12.588  < 2e-16 ***
incentiveYes 1.7328    0.1092  15.872  < 2e-16 ***
history      0.4732    0.0729   6.487  1.37e-10 ***
```

(a) For each of the fitted models, briefly explain what can be inferred about participants with similar histories.

(b) Based on this analysis and the experimental design, is it advisable for a participant with asthma to cycle to work more often? Explain.
The Gamma distribution with shape parameter $\alpha > 0$ and scale parameter $\lambda > 0$ has probability density function

$$f(y; \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\lambda y}$$

for $y > 0$

where $\Gamma$ is the Gamma function. Give the definition of an *exponential dispersion family* and show that the set of Gamma distributions forms one such family. Find the cumulant generating function and derive the mean and variance of the Gamma distribution as a function of $\alpha$ and $\lambda$. 
A sociologist collects a dataset on friendships among \(m\) Cambridge graduates. Let \(y_{i,j} = 1\) if persons \(i\) and \(j\) are friends 3 years after graduation, and \(y_{i,j} = 0\) otherwise. [You may assume that \(y_{i,j} = y_{j,i}\) and \(y_{i,i} = 0\).] Let \(z_i\) be a categorical variable for person \(i\)'s college, taking values in the set \(\{1, 2, \ldots, C\}\). Consider logistic regression models,

\[
P(y_{i,j} = 1) = \frac{e^{\theta_{i,j}}}{1 + e^{\theta_{i,j}}}, \quad 1 \leq i < j \leq m,
\]

with parameters either

(i) \(\theta_{i,j} = \beta_{z_i, z_j}\); or,

(ii) \(\theta_{i,j} = \beta_{z_i} + \beta_{z_j}\); or,

(iii) \(\theta_{i,j} = \beta_{z_i} + \beta_{z_j} + \beta_0 \delta_{z_i, z_j}\), where \(\delta_{z_i, z_j} = 1\) if \(z_i = z_j\) and 0 otherwise.

(a) Write down the likelihood of the models.

(b) Show that the three models are nested and specify the order. Suggest a statistic to compare models (i) and (iii), give its definition and specify its asymptotic distribution under the null hypothesis, citing any necessary theorems.

(c) Suppose persons \(i\) and \(j\) are in the same college \(k\); consider the number of friendships, \(M_i\) and \(M_j\), that each of them has with people in college \(\ell \neq k\) (\(\ell\) and \(k\) fixed). In each of the models above, compare the distribution of these two random variables. Explain why this might lead to a poor quality of fit.

(d) Find a minimal sufficient statistic for \(\beta = (\beta_k)_{k=0,1,\ldots,C}\) in model (iii). [You may use the following characterisation of a minimal sufficient statistic: let \(f(\beta; y)\) be the likelihood in this model, where \(y = (y_{i,j})_{i,j=1,\ldots,m}\); suppose \(T = t(y)\) is a statistic such that \(f(\beta; y)/f(\beta; y')\) is constant in \(\beta\) if and only if \(t(y) = t(y')\); then, \(T\) is a minimal sufficient statistic for \(\beta\).]
13.1 Statistical Modelling

The `ice_cream` data frame contains the result of a blind tasting of 90 ice creams, each of which is rated as poor, good, or excellent. It also contains the price of each ice cream classified into three categories. Consider the R code below and its output.

```r
> table(ice_cream)
   score
   price  excellent good poor
   high    12     8    10
   low      7     9   14
   medium  12    11     7
>
> ice_cream.counts = as.data.frame(xtabs(Freq ~ price + score, data=table(ice_cream)))
> glm.fit = glm(Freq ~ price + score, data=ice_cream.counts, family="poisson")
> summary(glm.fit)

Call:  
glm(formula = Freq ~ price + score - 1, family = "poisson", data = ice_cream.counts)

Deviance Residuals:  
1 2 3 4 5 6 7 8 9  
0.5054 -1.1019  0.5054  -0.4475  -0.1098  0.5304  -0.1043  1.0816  -1.1019

Coefficients:  
   Estimate Std. Error z value Pr(>|z|)
pricehigh 2.335e+00  2.334e-01  10.01   <2e-16 ***
pricelow  2.335e+00  2.334e-01  10.01   <2e-16 ***
pricemedium 2.335e+00  2.334e-01 10.01   <2e-16 ***
scoregood -1.018e-01  2.607e-01  -0.39  0.696
scorepoor  3.892e-14  2.540e-01  0.00  1.000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 257.2811 on 9 degrees of freedom
Residual deviance: 4.6135 on 4 degrees of freedom
AIC: 51.791

(a) Write down the generalised linear model fitted by the code above.

(b) Prove that the fitted values resulting from the maximum likelihood estimator of the coefficients in this model are identical to those resulting from the maximum likelihood estimator when fitting a multinomial model which assumes the number of ice creams at each price level is fixed.

(c) Using the output above, perform a goodness-of-fit test at the 1% level, specifying the null hypothesis, the test statistic, its asymptotic null distribution, any assumptions of the test and the decision from your test.

(d) If we believe that better ice creams are more expensive, what could be a more powerful test against the model fitted above and why?
5J Statistical Modelling

A scientist is studying the effects of a drug on the weight of mice. Forty mice are divided into two groups, control and treatment. The mice in the treatment group are given the drug, and those in the control group are given water instead. The mice are kept in 8 different cages. The weight of each mouse is monitored for 10 days, and the results of the experiment are recorded in the data frame `Weight.data`. Consider the following R code and its output.

```r
> head(Weight.data)
   Time Group Cage Mouse Weight
1    1   Control  1     1  24.77578
2    2   Control  1     1  24.68766
3    3   Control  1     1  24.79008
4    4   Control  1     1  24.77005
5    5   Control  1     1  24.65092
6    6   Control  1     1  24.38436

> mod1 = lm(Weight ~ Time*Group + Cage, data=Weight.data)

> summary(mod1)

Call:
lm(formula = Weight ~ Time * Group + Cage, data = Weight.data)

Residuals:
Min       1Q   Median       3Q      Max
-1.36903 -0.33527 -0.01719  0.38807  1.24368

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.534771  0.100336 244.525  < 2e-16 ***
Time        -0.006023  0.012616  -0.477  0.63334
GroupTreatment  0.321837  0.121993   2.638  0.00867 **
Cage2       -0.400228  0.095875  -4.174  3.68e-05 ***
Cage3        0.286941  0.102494   2.800  0.00537 **
Cage4        0.007535  0.095875   0.079  0.93740
Cage6        0.124767  0.125530   0.994  0.32087
Cage8       -0.295168  0.125530  -2.351  0.01920 *
Time:GroupTreatment -0.173515  0.017842  -9.725  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5125 on 391 degrees of freedom
Multiple R-squared: 0.5591, Adjusted R-squared: 0.55
F-statistic: 61.97 on 8 and 391 DF, p-value: < 2.2e-16
```

Which parameters describe the rate of weight loss with time in each group? According to the R output, is there a statistically significant weight loss with time in
the control group?

Three diagnostic plots were generated using the following R code.

```r
mouse1 = (Weight.data$Mouse==1)
plot(Weight.data$Time[mouse1],mod1$residuals[mouse1])
mouse2 = (Weight.data$Mouse==2)
plot(Weight.data$Time[mouse2],mod1$residuals[mouse2])
mouse3 = (Weight.data$Mouse==3)
plot(Weight.data$Time[mouse3],mod1$residuals[mouse3])
```
Based on these plots, should you trust the significance tests shown in the output of the command `summary(mod1)`? Explain.
The data frame `Cases.of.flu` contains a list of cases of flu recorded in 3 London hospitals during each month of 2017. Consider the following R code and its output.

```r
> table(Cases.of.flu)
         Hospital
    Month    A    B    C
  April   10  40  27
  August   9  34  19
 December 24 129  81
 February 49 134  74
 January  45 138  78
  July    5  47  35
  June    11  36  22
 March    20  82  41
  May     5  43  23
 November 17  82  62
 October   6  26  19
 September 6  40  21

> Cases.of.flu.table = as.data.frame(table(Cases.of.flu))
> head(Cases.of.flu.table)
     Month Hospital Freq
     1  April     A   10
     2  August    A   9
     3 December   A  24
     4 February   A  49
     5 January    A  45
     6  July      A  5

> mod1 = glm(Freq ~., data=Cases.of.flu.table, family=poisson)
> mod1$dev
[1] 28.51836
> levels(Cases.of.flu$Month)
[1] "April"  "August"  "December"  "February"  "January"  "July"
[7] "June"  "March"  "May"  "November"  "October"  "September"
> levels(Cases.of.flu$Month) <- c("Q2","Q3","Q4","Q1","Q1","Q3",
  
  "Q2","Q1","Q2","Q4","Q4","Q3")
> Cases.of.flu.table = as.data.frame(table(Cases.of.flu))
> mod2 = glm(Freq ~., data=Cases.of.flu.table, family=poisson)
> mod2$dev
[1] 17.9181
```

Describe a test for the null hypothesis of independence between the variables `Month` and `Hospital` using the deviance statistic. State the assumptions of the test.

Perform the test at the 1% level for each of the two different models shown above. You may use the table below showing 99th percentiles of the \( \chi^2 \) distribution with a range of degrees of freedom \( p \). How would you explain the discrepancy between their conclusions?
Paper 2, Section I
5J Statistical Modelling

Consider a linear model $Y = X\beta + \sigma^2 \varepsilon$ with $\varepsilon \sim N(0, I)$, where the design matrix $X$ is $n$ by $p$. Provide an expression for the $F$-statistic used to test the hypothesis $\beta_{p_0+1} = \beta_{p_0+2} = \cdots = \beta_p = 0$ for $p_0 < p$. Show that it is a monotone function of a log-likelihood ratio statistic.

<table>
<thead>
<tr>
<th>Degrees of freedom</th>
<th>99th percentile</th>
<th>Degrees of freedom</th>
<th>99th percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.63</td>
<td>21</td>
<td>38.93</td>
</tr>
<tr>
<td>2</td>
<td>9.21</td>
<td>22</td>
<td>40.29</td>
</tr>
<tr>
<td>3</td>
<td>11.34</td>
<td>23</td>
<td>41.64</td>
</tr>
<tr>
<td>4</td>
<td>13.28</td>
<td>24</td>
<td>42.98</td>
</tr>
<tr>
<td>5</td>
<td>15.09</td>
<td>25</td>
<td>44.31</td>
</tr>
<tr>
<td>6</td>
<td>16.81</td>
<td>26</td>
<td>45.64</td>
</tr>
<tr>
<td>7</td>
<td>18.48</td>
<td>27</td>
<td>46.96</td>
</tr>
<tr>
<td>8</td>
<td>20.09</td>
<td>28</td>
<td>48.28</td>
</tr>
<tr>
<td>9</td>
<td>21.67</td>
<td>29</td>
<td>49.59</td>
</tr>
<tr>
<td>10</td>
<td>23.21</td>
<td>30</td>
<td>50.89</td>
</tr>
<tr>
<td>11</td>
<td>24.72</td>
<td>31</td>
<td>52.19</td>
</tr>
<tr>
<td>12</td>
<td>26.22</td>
<td>32</td>
<td>53.49</td>
</tr>
<tr>
<td>13</td>
<td>27.69</td>
<td>33</td>
<td>54.78</td>
</tr>
<tr>
<td>14</td>
<td>29.14</td>
<td>34</td>
<td>56.06</td>
</tr>
<tr>
<td>15</td>
<td>30.58</td>
<td>35</td>
<td>57.34</td>
</tr>
<tr>
<td>16</td>
<td>32.00</td>
<td>36</td>
<td>58.62</td>
</tr>
<tr>
<td>17</td>
<td>33.41</td>
<td>37</td>
<td>59.89</td>
</tr>
<tr>
<td>18</td>
<td>34.81</td>
<td>38</td>
<td>61.16</td>
</tr>
<tr>
<td>19</td>
<td>36.19</td>
<td>39</td>
<td>62.43</td>
</tr>
<tr>
<td>20</td>
<td>37.57</td>
<td>40</td>
<td>63.69</td>
</tr>
</tbody>
</table>
The data frame `Ambulance` contains data on the number of ambulance requests from a Cambridgeshire hospital on different days. In addition to the number of ambulance requests on each day, the dataset records whether each day fell in the winter season, on a weekend, or on a bank holiday, as well as the pollution level on each day.

```
> head(Ambulance)

Winter Weekend Bank.holiday Pollution.level Ambulance.requests
1 Yes Yes No High 16
2 No Yes No Low 7
3 No No No High 22
4 No Yes No Medium 11
5 Yes Yes No High 18
6 No No No Medium 25
```

A health researcher fitted two models to the dataset above using R. Consider the following code and its output.

```
> mod1 = glm(Ambulance.requests ~ ., data=Ambulance, family=poisson)
> summary(mod1)

Call:
glm(formula = Ambulance.requests ~ ., family = poisson, data = Ambulance)

Deviance Residuals:
  Min       1Q   Median       3Q      Max
-3.2351  -0.8157  -0.0982   0.7787   3.6568

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.968477   0.036770  80.732  < 2e-16 ***
WinterYes    0.547756   0.033137  16.530  < 2e-16 ***
WeekendYes   -0.607910   0.038184 -15.921  < 2e-16 ***
Bank.holidayYes  0.165684   0.049875   3.322   0.000894 ***
Pollution.levelLow -0.032739   0.042290 -0.774    0.438846
Pollution.levelMedium -0.001587   0.040491 -0.039    0.968734
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 818.08 on 199 degrees of freedom
Residual deviance: 304.97 on 194 degrees of freedom
AIC: 1262.4
```
Define the two models fitted by this code and perform a hypothesis test with level 1% in which one of the models is the null hypothesis and the other is the alternative. State the theorem used in this hypothesis test. You may use the information generated by the following commands.

```r
> qchisq(0.01, df=2, lower.tail=FALSE)
[1] 9.21034
> qchisq(0.01, df=3, lower.tail=FALSE)
[1] 11.34487
> qchisq(0.01, df=4, lower.tail=FALSE)
[1] 13.2767
> qchisq(0.01, df=5, lower.tail=FALSE)
[1] 15.08627
```
Paper 4, Section II
13J Statistical Modelling

Bridge is a card game played by 2 teams of 2 players each. A bridge club records the outcomes of many games between teams formed by its members. The outcomes are modelled by

\[
P(\text{team } \{i, j\} \text{ wins against team } \{k, \ell\}) = \frac{\exp(\beta_i + \beta_j + \beta_{\{i,j\}} - \beta_k - \beta_{\{k,\ell\}})}{1 + \exp(\beta_i + \beta_j + \beta_{\{i,j\}} - \beta_k - \beta_{\{k,\ell\}})},
\]

where \(\beta_i \in \mathbb{R}\) is a parameter representing the skill of player \(i\), and \(\beta_{\{i,j\}} \in \mathbb{R}\) is a parameter representing how well-matched the team formed by \(i\) and \(j\) is.

(a) Would it make sense to include an intercept in this logistic regression model? Explain your answer.

(b) Suppose that players 1 and 2 always play together as a team. Is there a unique maximum likelihood estimate for the parameters \(\beta_1, \beta_2\) and \(\beta_{\{1,2\}}\)? Explain your answer.

(c) Under the model defined above, derive the asymptotic distribution (including the values of all relevant parameters) for the maximum likelihood estimate of the probability that team \(\{i, j\}\) wins a game against team \(\{k, \ell\}\). You can state it as a function of the true vector of parameters \(\beta\), and the Fisher information matrix \(i_N(\beta)\) with \(N\) games. You may assume that \(i_N(\beta)/N \to I(\beta)\) as \(N \to \infty\), and that \(\beta\) has a unique maximum likelihood estimate for \(N\) large enough.

---

Paper 1, Section II
13J Statistical Modelling

A clinical study follows a number of patients with an illness. Let \(Y_i \in [0, \infty)\) be the length of time that patient \(i\) lives and \(x_i \in \mathbb{R}^p\) a vector of predictors, for \(i \in \{1, \ldots, n\}\). We shall assume that \(Y_1, \ldots, Y_n\) are independent. Let \(f_i\) and \(F_i\) be the probability density function and cumulative distribution function, respectively, of \(Y_i\). The hazard function \(h_i\) is defined as

\[
h_i(t) = \frac{f_i(t)}{1 - F_i(t)} \quad \text{for } t \geq 0.
\]

We shall assume that \(h_i(t) = \lambda(t)\exp(\beta^T x_i)\), where \(\beta \in \mathbb{R}^p\) is a vector of coefficients and \(\lambda(t)\) is some fixed hazard function.

(a) Prove that \(F_i(t) = 1 - \exp(-\int_0^t h_i(s)ds)\).

(b) Using the equation in part (a), write the log-likelihood function for \(\beta\) in terms of \(\lambda, \beta, x_i\) and \(Y_i\) only.

(c) Show that the maximum likelihood estimate of \(\beta\) can be obtained through a surrogate Poisson generalised linear model with an offset.
The dataset ChickWeights records the weight of a group of chickens fed four different diets at a range of time points. We perform the following regressions in R.

```r
attach(ChickWeight)
fit1 = lm(weight ~ Time+Diet)
fit2 = lm(log(weight) ~ Time+Diet)
fit3 = lm(log(weight) ~ Time+Diet+Time:Diet)
```

(i) Which hypothesis test does the following command perform? State the degrees of freedom, and the conclusion of the test.

```r
> anova(fit2,fit3)
```

Analysis of Variance Table

<table>
<thead>
<tr>
<th>Model 1: log(weight) ~ Time + Diet</th>
<th>Model 2: log(weight) ~ Time + Diet + Time:Diet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Res.Df</td>
<td>RSS</td>
</tr>
<tr>
<td>----------</td>
<td>-------</td>
</tr>
<tr>
<td>1</td>
<td>574</td>
</tr>
<tr>
<td>2</td>
<td>571</td>
</tr>
</tbody>
</table>

(ii) Define a diagnostic plot that might suggest the logarithmic transformation of the response in fit2.

(iii) Define the dashed line in the following plot, generated with the command `plot(fit3)`.
What does it tell us about the data point 579?

![Residuals vs Leverage](image)
A statistician is interested in the power of a $t$-test with level 5% in linear regression; that is, the probability of rejecting the null hypothesis $\beta_0 = 0$ with this test under an alternative with $\beta_0 > 0$.

(a) State the distribution of the least-squares estimator $\hat{\beta}_0$, and hence state the form of the $t$-test statistic used.

(b) Prove that the power does not depend on the other coefficients $\beta_j$ for $j > 0$.

For Fisher’s method of Iteratively Reweighted Least-Squares and Newton–Raphson optimisation of the log-likelihood, the vector of parameters $\beta$ is updated using an iteration

$$\beta^{(m+1)} = \beta^{(m)} + M(\beta^{(m)})^{-1}U(\beta^{(m)})$$

for a specific function $M$. How is $M$ defined in each method?

Prove that they are identical in a Generalised Linear Model with the canonical link function.
A Cambridge scientist is testing approaches to slow the spread of a species of moth in certain trees. Two groups of 30 trees were treated with different organic pesticides, and a third group of 30 trees was kept under control conditions. At the end of the summer the trees are classified according to the level of leaf damage, obtaining the following contingency table.

```
> xtabs(count~group+damage.level,data=treeConditions)

          damage.level
    group               Severe.Damage Moderate.Damage Some.Damage
    Control              22                 5                 3
    Treatment 1          18                 4                 8
    Treatment 2          14                 3                13
```

Which of the following Generalised Linear Model fitting commands is appropriate for these data? Why? Describe the model being fit.

(a) `fit <- glm(count~group+damage.level,data=treeConditions,family=poisson)`

(b) `fit <- glm(count~group+damage.level,data=treeConditions,family=multinomial)`

(c) `fit <- glm(damage.level~group,data=treeConditions,family=binomial)`

(d) `fit <- glm(damage.level~group,data=treeConditions,family=binomial, weights=count)`
The Cambridge Lawn Tennis Club organises a tournament in which every match consists of 11 games, all of which are played. The player who wins 6 or more games is declared the winner.

For players $a$ and $b$, let $n_{ab}$ be the total number of games they play against each other, and let $y_{ab}$ be the number of these games won by player $a$. Let $\tilde{n}_{ab}$ and $\tilde{y}_{ab}$ be the corresponding number of matches.

A statistician analysed the tournament data using a Binomial Generalised Linear Model (GLM) with outcome $y_{ab}$. The probability $P_{ab}$ that $a$ wins a game against $b$ is modelled by

$$\log \left( \frac{P_{ab}}{1 - P_{ab}} \right) = \beta_a - \beta_b,$$

with an appropriate corner point constraint. You are asked to re-analyse the data, but the game-level results have been lost and you only know which player won each match.

We define a new GLM for the outcomes $\tilde{y}_{ab}$ with $\tilde{P}_{ab} = E\tilde{y}_{ab}/\tilde{n}_{ab}$ and $g(\tilde{P}_{ab}) = \beta_a - \beta_b$, where the $\beta$s are defined in (*). That is, $\beta_a - \beta_b$ is the log-odds that $a$ wins a game against $b$, not a match.

Derive the form of the new link function $g$. [You may express your answer in terms of a cumulative distribution function.]
12J Statistical Modelling

The dataset diesel records the number of diesel cars which go through a block of Hills Road in 6 disjoint periods of 30 minutes, between 8AM and 11AM. The measurements are repeated each day for 10 days. Answer the following questions based on the code below, which is shown with partial output.

(a) Can we reject the model fit.1 at a 1% level? Justify your answer.

(b) What is the difference between the deviance of the models fit.2 and fit.3?

(c) Which of fit.2 and fit.3 would you use to perform variable selection by backward stepwise selection? Why?

(d) How does the final plot differ from what you expect under the model in fit.2? Provide a possible explanation and suggest a better model.

> head(diesel)
  period num.cars day
1   1      69 1
2   2      97 1
3   3     103 1
4   4      99 1
5   5      67 1
6   6      91 1
> fit.1 = glm(num.cars~period,data=diesel,family=poisson)
> summary(fit.1)

Deviance Residuals:
  Min 1Q Median 3Q Max
-4.0188 -1.4837 -0.2117 1.6257 4.5965

Coefficients:
  Estimate Std. Error z value Pr(>|z|)
(Intercept) 4.628535 0.029288 158.035 <2e-16 ***
period -0.006073 0.007551 -0.804 0.421
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 ? 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 262.36 on 59 degrees of freedom
Residual deviance: 261.72 on 58 degrees of freedom
AIC: 651.2

> diesel$period.factor = factor(diesel$period)
> fit.2 = glm(num.cars~period.factor,data=diesel,family=poisson)
> summary(fit.2)

Coefficients:
  Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.36818  0.03560  122.698  <  2e-16  ***  
period.factor2  0.35655  0.04642  7.681  1.58e-14  ***  
period.factor3  0.41262  0.04590  8.991  <  2e-16  ***  
period.factor4  0.36274  0.04636  7.824  5.10e-15  ***  
period.factor5  0.06501  0.04955  1.312  0.18948  
period.factor6  0.16334  0.04841  3.374  0.000741  ***  
---  
Signif. codes:  0 ‘***’  0.001 ‘**’  0.01 ‘*’  0.05 ‘.’  0.1 ‘ ’  1  

> fit.3 = glm(num.cars~(period>1)+(period>2)+(period>3)+(period>4)+(period>5),  
  data=diesel,family=poisson)  
> summary(fit.3)

Coefficients:  
    Estimate Std. Error z value Pr(>|z|)  
(Intercept)  4.36818  0.03560  122.698  <  2e-16  ***  
period > 1TRUE  0.35655  0.04642  7.681  1.58e-14  ***  
period > 2TRUE  0.05607  0.04155  1.350  0.1771  
period > 3TRUE -0.04988  0.04148  -1.202  0.2292  
period > 4TRUE -0.29773  0.04549  -6.545  5.96e-11  ***  
period > 5TRUE  0.09833  0.04758  2.066  0.0388  *  
---  
Signif. codes:  0 ‘***’  0.001 ‘**’  0.01 ‘*’  0.05 ‘.’  0.1 ‘ ’  1  

> C = matrix(nrow=6,ncol=2)  
> for (period in 1:6) {  
    nums = diesel$num.cars[diesel$period == period]  
    C[period,] = c(mean(nums),var(nums))  
  }  
> plot(C[,1],C[,2])

Part II, 2017 List of Questions  
[TURN OVER
Paper 2, Section I

5K Statistical Modelling

Define an exponential dispersion family. Prove that the range of the natural parameter, \( \Theta \), is an open interval. Derive the mean and variance as a function of the log normalizing constant.

\[ \text{Hint: Use the convexity of } e^x, \text{ i.e. } e^{px + (1-p)y} \leq pe^x + (1-p)e^y \text{ for all } p \in [0, 1]. \]

Paper 4, Section I

5K Statistical Modelling

(a) Let \( Y_i = x_i^T \beta + \varepsilon_i \) where \( \varepsilon_i \) for \( i = 1, \ldots, n \) are independent and identically distributed. Let \( Z_i = I(Y_i < 0) \) for \( i = 1, \ldots, n \), and suppose that these variables follow a binary regression model with the complementary log-log link function \( g(\mu) = \log(-\log(1 - \mu)). \) What is the probability density function of \( \varepsilon_1 \)?

(b) The Newton–Raphson algorithm can be applied to compute the MLE, \( \hat{\beta} \), in certain GLMs. Starting from \( \beta^{(0)} = 0 \), we let \( \beta^{(t+1)} \) be the maximizer of the quadratic approximation of the log-likelihood \( \ell(\beta; Y) \) around \( \beta^{(t)} \):

\[
\ell(\beta; Y) \approx \ell(\beta^{(t)}; Y) + (\beta - \beta^{(t)})^T D\ell(\beta^{(t)}; Y) + (\beta - \beta^{(t)})^T D^2\ell(\beta^{(t)}; Y)(\beta - \beta^{(t)}),
\]

where \( D\ell \) and \( D^2\ell \) are the gradient and Hessian of the log-likelihood. What is the difference between this algorithm and Iterative Weighted Least Squares? Why might the latter be preferable?
The R command

```r
> boxcox(rainfall ~ month+elnino+month:elnino)
```

performs a Box–Cox transform of the response at several values of the parameter $\lambda$, and produces the following plot:

![Box-Cox plot](image)

We fit two linear models and obtain the Q–Q plots for each fit, which are shown below in no particular order:

```r
> fit.1 <- lm(rainfall ~ month+elnino+month:elnino)
> plot(fit.1, which=2)
> fit.2 <- lm(rainfall^-0.07 ~ month+elnino+month:elnino)
> plot(fit.2, which=2)
```

![Q-Q plots](image)

This question continues on the next page
Define the variable on the y-axis in the output of `boxcox`, and match each Q–Q plot to one of the models.

After choosing the model `fit.2`, the researcher calculates Cook’s distance for the $i$th sample, which has high leverage, and compares it to the upper 0.01-point of an $F_{p,n-p}$ distribution, because the design matrix is of size $n \times p$. Provide an interpretation of this comparison in terms of confidence sets for $\hat{\beta}$. Is this confidence statement exact?
The body mass index (BMI) of your closest friend is a good predictor of your own BMI. A scientist applies polynomial regression to understand the relationship between these two variables among 200 students in a sixth form college. The R commands

```r
> fit.1 <- lm(BMI ~ poly(friendBMI, 2, raw=T))
> fit.2 <- lm(BMI ~ poly(friendBMI, 3, raw=T))
```

fit the models

\[ Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon \]

and

\[ Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon, \]

respectively, with \( \epsilon \sim N(0, \sigma^2) \) in each case.

Setting the parameters `raw` to `FALSE`:

```r
> fit.3 <- lm(BMI ~ poly(friendBMI, 2, raw=F))
> fit.4 <- lm(BMI ~ poly(friendBMI, 3, raw=F))
```

fits the models

\[ Y = \beta_0 + \beta_1 P_1(X) + \beta_2 P_2(X) + \epsilon \]

and

\[ Y = \beta_0 + \beta_1 P_1(X) + \beta_2 P_2(X) + \beta_3 P_3(X) + \epsilon, \]

with \( \epsilon \sim N(0, \sigma^2) \). The function \( P_i \) is a polynomial of degree \( i \). Furthermore, the design matrix output by the function `poly` with `raw=F` satisfies:

```r
> t(poly(friendBMI, 3, raw=F))%*%poly(a, 3, raw=F)
```

\[
\begin{array}{ccc}
1 & 2 & 3 \\
1 & 1.000000e+00 & 1.288032e-16 & 3.187554e-17 \\
2 & 1.288032e-16 & 1.000000e+00 & -6.201636e-17 \\
3 & 3.187554e-17 & -6.201636e-17 & 1.000000e+00 \\
\end{array}
\]

How does the variance of \( \hat{\beta} \) differ in the models `fit.2` and `fit.4`? What about the variance of the fitted values \( \hat{Y} = X \hat{\beta} \)? Finally, consider the output of the commands

```r
> anova(fit.1, fit.2)
> anova(fit.3, fit.4)
```

Define the test statistic computed by this function and specify its distribution. Which command yields a higher statistic?
12K Statistical Modelling

For 31 days after the outbreak of the 2014 Ebola epidemic, the World Health Organization recorded the number of new cases per day in 60 hospitals in West Africa. Researchers are interested in modelling \( Y_{ij} \), the number of new Ebola cases in hospital \( i \) on day \( j \geq 2 \), as a function of several covariates:

- **lab**: a Boolean factor for whether the hospital has laboratory facilities,
- **casesBefore**: number of cases at the hospital on the previous day,
- **urban**: a Boolean factor indicating an urban area,
- **country**: a factor with three categories, Guinea, Liberia, and Sierra Leone,
- **numDoctors**: number of doctors at the hospital,
- **tradBurials**: a Boolean factor indicating whether traditional burials are common in the region.

Consider the output of the following R code (with some lines omitted):

```r
> fit.1 <- glm(newCases ~ lab + casesBefore + urban + country + numDoctors + tradBurials,
+ data=ebola, family=poisson)
> summary(fit.1)

Coefficients:

|                | Estimate | Std. Error | z value | Pr(>|z|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 0.094731 | 0.050322   | 1.882   | 0.0598   |
| labTRUE        | 0.011298 | 0.049498   | 0.228   | 0.8195   |
| casesBefore    | 0.324744 | 0.007752   | 41.891  | < 2e-16  ***
| urbanTRUE      | -0.091554| 0.088212   | -1.038  | 0.2993   |
| countryLiberia | 0.088490 | 0.034119   | 2.594   | 0.0095   **
| countrySierra Leone | -0.197474 | 0.036969 | -5.342  | 9.21e-08 ***
| numDoctors     | -0.020819| 0.004658   | -4.470  | 7.83e-06 ***
| tradBurialsTRUE| 0.054296 | 0.031676   | 1.714   | 0.0865   .
```

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(a) Would you conclude based on the \( z \)-tests that an urban setting does not affect the rate of infection?

(b) Explain how you would predict the total number of new cases that the researchers will record in Sierra Leone on day 32.

We fit a new model which includes an interaction term, and compute a test statistic using the code:

```r
> fit.2 <- glm(newCases ~ casesBefore + country + country:casesBefore + numDoctors,
+ data=ebola, family=poisson)
> fit.2$deviance - fit.1$deviance
[1] 3.016138
```

(c) What is the distribution of the statistic computed in the last line?

(d) Under what conditions is the deviance of each model approximately chi-squared?
12K Statistical Modelling

(a) Let \( Y \) be an \( n \)-vector of responses from the linear model \( Y = X\beta + \varepsilon \), with \( \beta \in \mathbb{R}^p \). The *internally studentized residual* is defined by

\[
s_i = \frac{Y_i - x_i^\top \hat{\beta}}{\tilde{\sigma} \sqrt{1 - p_i}},
\]

where \( \hat{\beta} \) is the least squares estimate, \( p_i \) is the leverage of sample \( i \), and

\[
\tilde{\sigma}^2 = \frac{\|Y - X\hat{\beta}\|^2}{(n - p)}
\]

Prove that the joint distribution of \( s = (s_1, \ldots, s_n)^\top \) is the same in the following two models: (i) \( \varepsilon \sim \mathcal{N}(0, \sigma I) \), and (ii) \( \varepsilon | \sigma \sim \mathcal{N}(0, \sigma I) \), with \( 1/\sigma \sim \chi^2_\nu \) (in this model, \( \varepsilon_1, \ldots, \varepsilon_n \) are identically \( t_\nu \)-distributed). [Hint: A random vector \( Z \) is spherically symmetric if for any orthogonal matrix \( H \), \( HZ \) has the same distribution as \( Z \). If \( Z \) is spherically symmetric and a.s. nonzero, then \( Z/\|Z\|_2 \) is a uniform point on the sphere; in addition, any orthogonal projection of \( Z \) is also spherically symmetric. A standard normal vector is spherically symmetric.]

(b) A social scientist regresses the income of 120 Cambridge graduates onto 20 answers from a questionnaire given to the participants in their first year. She notices one questionnaire with very unusual answers, which she suspects was due to miscoding. The sample has a leverage of 0.8. To check whether this sample is an outlier, she computes its *externally studentized residual*,

\[
t_i = \frac{Y_i - x_i^\top \hat{\beta}}{\tilde{\sigma}_{(i)} \sqrt{1 - p_i}} = 4.57,
\]

where \( \tilde{\sigma}_{(i)} \) is estimated from a fit of all samples except the one in question, \( (x_i, Y_i) \). Is this a high leverage point? Can she conclude this sample is an outlier at a significance level of 5%?

(c) After examining the following plot of residuals against the response, the investigator calculates the externally studentized residual of the participant denoted by the black dot, which is 2.33. Can she conclude this sample is an outlier with a significance level of 5%?
Data on 173 nesting female horseshoe crabs record for each crab its colour as one of 4 factors (simply labelled 1, ..., 4), its width (in cm) and the presence of male crabs nearby (a 1 indicating presence). The data are collected into the R data frame `crabs` and the first few lines are displayed below.

```r
> crabs[1:4,]
colour width males
1 2  28.3 1
2 3  22.5 0
3 1  26.0 1
4 4  21.0 0
```

Describe the model being fitted by the R command below.

```r
> fit1 <- glm(males ~ colour + width, family = binomial, data=crabs)
```

The following (abbreviated) output is obtained from the `summary` command.

```r
> summary(fit1)

Coefficients:  
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -11.38      2.873  -3.962   7.43e-05 ***
colour2       -0.07       0.740   0.098     0.922
colour3       -0.22       0.777  -0.288     0.773
colour4       -1.32       0.853  -1.560     0.119
width         -0.46       0.106   4.434    9.26e-06 ***
```

Write out the calculation for an approximate 95% confidence interval for the coefficient for `width`. Describe the calculation you would perform to obtain an estimate of the probability that a female crab of colour 3 and with a width of 20cm has males nearby. [You need not actually compute the end points of the confidence interval or the estimate of the probability above, but merely show the calculations that would need to be performed in order to arrive at them.]
Data are available on the number of counts (atomic disintegration events that take place within a radiation source) recorded with a Geiger counter at a nuclear plant. The counts were registered at each second over a 30 second period for a short-lived, man-made radioactive compound. The first few rows of the dataset are displayed below.

```
> geiger[1:3,]
   Time Counts
1   0  750.0
2   1  725.2
3   2  695.0
```

Describe the model being fitted with the following R command.

```
> fit1 <- lm(Counts ~ Time, data=geiger)
```

Below is a plot against time of the residuals from the model fitted above.

![Residuals vs Time Plot]

Referring to the plot, suggest how the model could be improved, and write out the R code.
for fitting this new model. Briefly describe how one could test in R whether the new model is to be preferred over the old model.

**Paper 2, Section I**

**4J Statistical Modelling**

Let $Y_1, \ldots, Y_n$ be independent Poisson random variables with means $\mu_1, \ldots, \mu_n$, where $\log(\mu_i) = \beta x_i$ for some known constants $x_i \in \mathbb{R}$ and an unknown parameter $\beta$. Find the log-likelihood for $\beta$.

By first computing the first and second derivatives of the log-likelihood for $\beta$, describe the algorithm you would use to find the maximum likelihood estimator $\hat{\beta}$. [Hint: Recall that if $Z \sim \text{Pois}(\mu)$ then

$$
P(Z = k) = \frac{\mu^k e^{-\mu}}{k!}
$$

for $k \in \{0, 1, 2, \ldots\}$].]

**Paper 1, Section I**

**4J Statistical Modelling**

The outputs $Y_1, \ldots, Y_n$ of a particular process are positive and are believed to be related to $p$-vectors of covariates $x_1, \ldots, x_n$ according to the following model

$$
\log(Y_i) = \mu + x_i^T \beta + \varepsilon_i.
$$

In this model $\varepsilon_i$ are i.i.d. $N(0, \sigma^2)$ random variables where $\sigma > 0$ is known. It is not possible to measure the output directly, but we can detect whether the output is greater than or less than or equal to a certain known value $c > 0$. If

$$
Z_i = \begin{cases} 
1 & \text{if } Y_i > c \\
0 & \text{if } Y_i \leq c,
\end{cases}
$$

show that a probit regression model can be used for the data $(Z_i, x_i), i = 1, \ldots, n$.

How can we recover $\mu$ and $\beta$ from the parameters of the probit regression model?
Consider the normal linear model where the $n$-vector of responses $Y$ satisfies $Y = X\beta + \varepsilon$ with $\varepsilon \sim N_n(0, \sigma^2 I)$. Here $X$ is an $n \times p$ matrix of predictors with full column rank where $p \geq 3$ and $\beta \in \mathbb{R}^p$ is an unknown vector of regression coefficients. For $j \in \{1, \ldots, p\}$, denote the $j$th column of $X$ by $X_j$, and let $X_{-j}$ be $X$ with its $j$th column removed. Suppose $X_1 = 1_n$ where $1_n$ is an $n$-vector of 1’s. Denote the maximum likelihood estimate of $\beta$ by $\hat{\beta}$. Write down the formula for $\hat{\beta}_j$ involving $P_{-j}$, the orthogonal projection onto the column space of $X_{-j}$.

Consider $j, k \in \{2, \ldots, p\}$ with $j < k$. By thinking about the orthogonal projection of $X_j$ onto $X_k$, show that

$$\text{var}(\hat{\beta}_j) \geq \frac{\sigma^2}{\|X_j\|^2} \left(1 - \left(\frac{X_k^T X_j}{\|X_k\|^2 \|X_j\|}\right)^2\right)^{-1}. \quad (*)$$

[You may use standard facts about orthogonal projections including the fact that if $V$ and $W$ are subspaces of $\mathbb{R}^n$ with $V$ a subspace of $W$ and $\Pi_V$ and $\Pi_W$ denote orthogonal projections onto $V$ and $W$ respectively, then for all $v \in \mathbb{R}^n$, $\|\Pi_W v\|^2 \geq \|\Pi_V v\|^2$.]

This question continues on the next page.
10J Statistical Modelling (continued)

By considering the fitted values \(X\hat{\beta}\), explain why if, for any \(j \geq 2\), a constant is added to each entry in the \(j\)th column of \(X\), then \(\hat{\beta}_j\) will remain unchanged. Let \(\bar{X}_j = \frac{\sum_{i=1}^{n} X_{ij}}{n}\). Why is (*) also true when all instances of \(X_j\) and \(X_k\) are replaced by \(X_j - \bar{X}_j\) and \(X_k - \bar{X}_k\) respectively?

The marks from mid-year statistics and mathematics tests and an end-of-year statistics exam are recorded for 100 secondary school students. The first few lines of the data are given below.

```
> exam_marks[1:3,]
       Stat_exam Maths_test Stat_test
1        83       94       92
2        76       45       27
3        73       67       62
```

The following abbreviated output is obtained:

```
> summary(lm(Stat_exam ~ Maths_test + Stat_test, data=exam_marks))

Coefficients:
            Estimate Std. Error t value Pr(>|t|) 
(Intercept)   25.0342    8.2694   3.027   0.00316 **
Maths_test     0.2782    0.3708   0.750   0.45503
Stat_test      0.1643    0.3364   0.488   0.62641
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

F-statistic: 6.111 on 2 and 97 DF,  p-value: 0.003166
```

What are the hypothesis tests corresponding to the final column of the coefficients table? What is the hypothesis test corresponding to the final line of the output? Interpret the results when testing at the 5% level.

How does the following sample correlation matrix for the data help to explain the relative sizes of some of the \(p\)-values?

```
> cor(exam_marks)
        Stat_exam Maths_test Stat_test 
Stat_exam 1.0000000 0.3312240 0.3267138 
Maths_test 0.3312240 1.0000000 0.9371630 
Stat_test  0.3267138 0.9371630 1.0000000
```
An experiment is conducted where scientists count the numbers of each of three different strains of fleas that are reproducing in a controlled environment. Varying concentrations of a particular toxin that impairs reproduction are administered to the fleas. The results of the experiment are stored in a data frame `fleas` in R, whose first few rows are given below.

```r
> fleas[1:3, ]
   number conc strain
1     81 0.250     0
2     93 0.250     2
3    102 0.875     1
```

The full dataset has 80 rows. The first column provides the number of fleas, the second provides the concentration of the toxin and the third specifies the strain of the flea as factors 0, 1 or 2. Strain 0 is the common flea and strains 1 and 2 have been genetically modified in a way thought to increase their ability to reproduce in the presence of the toxin.

This question continues on the next page
10J Statistical Modelling (continued)

Explain and interpret the R commands and (abbreviated) output below. In particular, you should describe the model being fitted, briefly explain how the standard errors are calculated, and comment on the hypothesis tests being described in the summary.

```r
> fit1 <- glm(number ~ conc*strain, data=fleas, family=poisson)
> summary(fit1)
```

Coefficients:

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|----------|
| (Intercept) 4.47171 | 0.03849 | 116.176 | < 2e-16 *** |
| conc -0.28700 | 0.06727 | -4.266 | 1.99e-05 *** |
| strain1 0.09381 | 0.05483 | 1.711 | 0.087076 . |
| strain2 0.12157 | 0.05591 | 2.175 | 0.029666 * |
| conc:strain1 0.34215 | 0.09178 | 3.728 | 0.000193 *** |
| conc:strain2 0.02385 | 0.09789 | 0.244 | 0.807510 |

Explain and motivate the following R code in the light of the output above. Briefly explain the differences between the models fitted below, and the model corresponding to fit1.

```r
> strain_grp <- fleas$strain
> levels(strain_grp)
[1] "0" "1" "2"
> levels(strain_grp) <- c(0, 1, 0)
> fit2 <- glm(number ~ conc + strain + conc:strain_grp, + data=fleas, family=poisson)
> fit3 <- glm(number ~ conc*strain_grp, data=fleas, family=poisson)
```

Denote by $M_1, M_2, M_3$ the three models being fitted in sequence above. Explain the hypothesis tests comparing the models to each other that can be performed using the output from the following R code.

```r
> c(fit1$dev, fit2$dev, fit3$dev)
[1] 56.87 56.93 76.98
> qchisq(0.95, df = 1)
[1] 3.84
```

Use these numbers to comment on the most appropriate model for the data.
Consider the normal linear model where the \( n \)-vector of responses \( Y \) satisfies
\[
Y = X\beta + \varepsilon
\]
with \( \varepsilon \sim N_n(0, \sigma^2 I) \) and \( X \) is an \( n \times p \) design matrix with full column rank. Write down a \((1 - \alpha)\)-level confidence set for \( \beta \).

Define the Cook’s distance for the observation \((Y_i, x_i)\) where \( x_i^T \) is the \( i \)th row of \( X \), and give its interpretation in terms of confidence sets for \( \beta \).

In the model above with \( n = 100 \) and \( p = 4 \), you observe that one observation has Cook’s distance 3.1. Would you be concerned about the influence of this observation? Justify your answer.

[Hint: You may find some of the following facts useful:

1. If \( Z \sim \chi^2_4 \), then \( P(Z \leq 1.06) = 0.1, \ P(Z \leq 7.78) = 0.9 \).
2. If \( Z \sim F_{4,96} \), then \( P(Z \leq 0.26) = 0.1, \ P(Z \leq 2.00) = 0.9 \).
3. If \( Z \sim F_{96,4} \), then \( P(Z \leq 0.50) = 0.1, \ P(Z \leq 3.78) = 0.9 \).]
In an experiment to study factors affecting the production of the plastic polyvinyl chloride (PVC), three experimenters each used eight devices to produce the PVC and measured the sizes of the particles produced. For each of the 24 combinations of device and experimenter, two size measurements were obtained.

The experimenters and devices used for each of the 48 measurements are stored in R as factors in the objects `experimenter` and `device` respectively, with the measurements themselves stored in the vector `psize`. The following analysis was performed in R.

```r
> fit0 <- lm(psize ~ experimenter + device)
> fit <- lm(psize ~ experimenter + device + experimenter:device)
> anova(fit0, fit)

Analysis of Variance Table

Model 1: psize ~ experimenter + device
Model 2: psize ~ experimenter + device + experimenter:device

<table>
<thead>
<tr>
<th>Res.Df</th>
<th>RSS</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>F</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>38</td>
<td>49.815</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>35.480</td>
<td>14</td>
<td>14.335</td>
<td>0.6926</td>
<td>0.7599</td>
</tr>
</tbody>
</table>
```

Let \( X \) and \( X_0 \) denote the design matrices obtained by `model.matrix(fit)` and `model.matrix(fit0)` respectively, and let \( Y \) denote the response `psize`. Let \( P \) and \( P_0 \) denote orthogonal projections onto the column spaces of \( X \) and \( X_0 \) respectively.

For each of the following quantities, write down their numerical values if they appear in the analysis of variance table above; otherwise write ‘unknown’.

1. \( \| (I - P)Y \|_2^2 \)
2. \( \| X(X^T X)^{-1} X^T Y \|_2^2 \)
3. \( \| (I - P_0)Y \|_2^2 - \| (I - P)Y \|_2^2 \)
4. \( \frac{\| (P - P_0)Y \|_2^2 / 14}{\| (I - P)Y \|_2^2 / 24} \)
5. \( \sum_{i=1}^{48} Y_i / 48 \)

Out of the two models that have been fitted, which appears to be the more appropriate for the data according to the analysis performed, and why?
Define the concept of an exponential dispersion family. Show that the family of scaled binomial distributions $\frac{1}{n}\text{Bin}(n, p)$, with $p \in (0, 1)$ and $n \in \mathbb{N}$, is of exponential dispersion family form.

Deduce the mean of the scaled binomial distribution from the exponential dispersion family form.

What is the canonical link function in this case?

Write down the model being fitted by the following R command, where $y \in \{0, 1, 2, \ldots\}^n$ and $X$ is an $n \times p$ matrix with real-valued entries.

```r
fit <- glm(y ~ X, family = poisson)
```

Write down the log-likelihood for the model. Explain why the command

```r
sum(y) - sum(predict(fit, type = "response"))
```

gives the answer 0, by arguing based on the log-likelihood you have written down.

[Hint: Recall that if $Z \sim \text{Pois}(\mu)$ then

$$P(Z = k) = \frac{\mu^k e^{-\mu}}{k!}$$

for $k \in \{0, 1, 2, \ldots\}$.]
In a study on infant respiratory disease, data are collected on a sample of 2074 infants. The information collected includes whether or not each infant developed a respiratory disease in the first year of their life; the gender of each infant; and details on how they were fed as one of three categories (breast-fed, bottle-fed and supplement). The data are tabulated in R as follows:

<table>
<thead>
<tr>
<th>number</th>
<th>disease</th>
<th>nondisease</th>
<th>gender</th>
<th>food</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>77</td>
<td>381</td>
<td>Boy</td>
<td>Bottle-fed</td>
</tr>
<tr>
<td>2</td>
<td>19</td>
<td>128</td>
<td>Boy</td>
<td>Supplement</td>
</tr>
<tr>
<td>3</td>
<td>47</td>
<td>447</td>
<td>Boy</td>
<td>Breast-fed</td>
</tr>
<tr>
<td>4</td>
<td>48</td>
<td>336</td>
<td>Girl</td>
<td>Bottle-fed</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
<td>111</td>
<td>Girl</td>
<td>Supplement</td>
</tr>
<tr>
<td>6</td>
<td>31</td>
<td>433</td>
<td>Girl</td>
<td>Breast-fed</td>
</tr>
</tbody>
</table>

Write down the model being fit by the R commands on the following page:
> total <- disease + nondisease
> fit <- glm(disease/total ~ gender + food, family = binomial,
+ weights = total)

The following (slightly abbreviated) output from R is obtained.

> summary(fit)

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.6127     0.1124  -14.347  < 2e-16 ***
genderGirl           -0.3126     0.1410   -2.216   0.0267 *
foodBreast-fed       -0.6693     0.1530   -4.374  1.22e-05 ***
foodSupplement       -0.1725     0.2056   -0.839   0.4013
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 26.37529 on 5 degrees of freedom
Residual deviance: 0.72192 on 2 degrees of freedom

Briefly explain the justification for the standard errors presented in the output above.

Explain the relevance of the output of the following R code to the data being studied, justifying your answer:

> exp(c(-0.6693 - 1.96*0.153, -0.6693 + 1.96*0.153))
[1] 0.3793940 0.6911351

[Hint: It may help to recall that if \( Z \sim N(0, 1) \) then \( P(Z \geq 1.96) = 0.025 \).]

Let \( D_1 \) be the deviance of the model fitted by the following R command.

> fit1 <- glm(disease/total ~ gender + food + gender:food,
+ family = binomial, weights = total)

What is the numerical value of \( D_1 \)? Which of the two models that have been fitted should you prefer, and why?
Consider the normal linear model where the \( n \)-vector of responses \( Y \) satisfies
\[
Y = X\beta + \varepsilon \sim N_n(0, \sigma^2 I).
\]
Here \( X \) is an \( n \times p \) matrix of predictors with full column rank where \( n \geq p + 3 \), and \( \beta \in \mathbb{R}^p \) is an unknown vector of regression coefficients.

Let \( X_0 \) be the matrix formed from the first \( p_0 \) columns of \( X \), and partition \( \beta \) as
\[
\beta = \begin{pmatrix} \beta_0^T \\ \beta_1^T \end{pmatrix}^T
\]
where \( \beta_0 \in \mathbb{R}^{p_0} \) and \( \beta_1 \in \mathbb{R}^{p - p_0} \). Denote the orthogonal projections onto the column spaces of \( X \) and \( X_0 \) by \( P \) and \( P_0 \) respectively.

It is desired to test the null hypothesis \( H_0 : \beta_1 = 0 \) against the alternative hypothesis \( H_1 : \beta_1 \neq 0 \). Recall that the \( F \)-test for testing \( H_0 \) against \( H_1 \) rejects \( H_0 \) for large values of
\[
F = \frac{\| (P - P_0) Y \|^2 / (p - p_0)}{\| (I - P) Y \|^2 / (n - p)}.
\]

Show that \( (I - P)(P - P_0) = 0 \), and hence prove that the numerator and denominator of \( F \) are independent under either hypothesis.

Show that
\[
E_{\beta, \sigma^2}(F) = \frac{(n - p)(\tau^2 + 1)}{n - p - 2},
\]
where \( \tau^2 = \frac{\| (P - P_0) X \beta \|^2}{(p - p_0)\sigma^2} \).

[In this question you may use the following facts without proof: \( P - P_0 \) is an orthogonal projection with rank \( p - p_0 \); any \( n \times n \) orthogonal projection matrix \( \Pi \) satisfies \( \| \Pi \varepsilon \|^2 \sim \sigma^2 \chi^2_\nu \), where \( \nu = \text{rank}(\Pi) \); and if \( Z \sim \chi^2_\nu \) then \( E(Z^{-1}) = (\nu - 2)^{-1} \) when \( \nu > 2 \).]
The output $X$ of a process depends on the levels of two adjustable variables: $A$, a factor with four levels, and $B$, a factor with two levels. For each combination of a level of $A$ and a level of $B$, nine independent values of $X$ are observed.

Explain and interpret the R commands and (abbreviated) output below. In particular, describe the model being fitted, and describe and comment on the hypothesis tests performed under the summary and anova commands.

```r
> fit1 <- lm(x ~ a+b)
> summary(fit1)

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 2.5445 | 0.2449 | 10.39 | 6.66e-16 *** |
| a2 | -5.6704 | 0.4859 | -11.67 | < 2e-16 *** |
| a3 | 4.3254 | 0.3480 | 12.43 | < 2e-16 *** |
| a4 | -0.5003 | 0.3734 | -1.34 | 0.0923 |
| b2 | -3.5689 | 0.2275 | -15.69 | < 2e-16 *** |

> anova(fit1)

Response: x

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>3</td>
<td>71.51</td>
<td>23.84</td>
<td>17.79</td>
</tr>
<tr>
<td>b</td>
<td>1</td>
<td>105.11</td>
<td>105.11</td>
<td>78.44</td>
</tr>
<tr>
<td>Residuals</td>
<td>67</td>
<td>89.56</td>
<td>1.34</td>
<td></td>
</tr>
</tbody>
</table>
```
Consider the linear model \( Y = X\beta + \epsilon \) where \( Y = (Y_1, \ldots, Y_n)^T, \beta = (\beta_1, \ldots, \beta_p)^T, \) and \( \epsilon = (\epsilon_1, \ldots, \epsilon_n)^T, \) with \( \epsilon_1, \ldots, \epsilon_n \) independent \( N(0, \sigma^2) \) random variables. The \( (n \times p) \) matrix \( X \) is known and is of full rank \( p < n. \) Give expressions for the maximum likelihood estimators \( \hat{\beta} \) and \( \hat{\sigma}^2 \) of \( \beta \) and \( \sigma^2 \) respectively, and state their joint distribution. Show that \( \hat{\beta} \) is unbiased whereas \( \hat{\sigma}^2 \) is biased.

Suppose that a new variable \( Y^* \) is to be observed, satisfying the relationship
\[
Y^* = x^T\beta + \epsilon^*,
\]
where \( x^* \) \((p \times 1)\) is known, and \( \epsilon^* \sim N(0, \sigma^2) \) independently of \( \epsilon. \) We propose to predict \( Y^* \) by \( \tilde{Y} = x^T\hat{\beta}. \) Identify the distribution of \( Y^* - \tilde{Y} \)
\[
\frac{\tau^2}{\tilde{\sigma}^2},
\]
where
\[
\tilde{\sigma}^2 = \frac{n-n}{n-p} \hat{\sigma}^2,
\]
\[
\tau^2 = x^T(X^TX)^{-1}x^* + 1.
\]

Consider a linear model \( Y = X\beta + \epsilon, \) where \( Y \) and \( \epsilon \) are \((n \times 1)\) with \( \epsilon \sim N_n(0, \sigma^2I), \beta \) \((p \times 1)\), and \( X \) \((n \times p)\) of full rank \( p < n. \) Let \( \gamma \) and \( \delta \) be sub-vectors of \( \beta. \) What is meant by orthogonality between \( \gamma \) and \( \delta? \)

Now suppose
\[
Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 P_3(x_i) + \epsilon_i \quad (i = 1, \ldots, n),
\]
where \( \epsilon_1, \ldots, \epsilon_n \) are independent \( N(0, \sigma^2) \) random variables, \( x_1, \ldots, x_n \) are real-valued known explanatory variables, and \( P_3(x) \) is a cubic polynomial chosen so that \( \beta_3 \) is orthogonal to \( (\beta_0, \beta_1, \beta_2)^T \) and \( \beta_1 \) is orthogonal to \( (\beta_0, \beta_2)^T. \)

Let \( \tilde{\beta} = (\beta_0, \beta_2, \beta_1, \beta_3)^T. \) Describe the matrix \( \tilde{X} \) such that \( Y = \tilde{X}\tilde{\beta} + \epsilon. \) Show that \( \tilde{X}^T\tilde{X} \) is block diagonal. Assuming further that this matrix is non-singular, show that the least-squares estimators of \( \beta_1 \) and \( \beta_3 \) are, respectively,
\[
\tilde{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} \quad \text{and} \quad \tilde{\beta}_3 = \frac{\sum_{i=1}^n P_3(x_i) y_i}{\sum_{i=1}^n P_3(x_i)^2}.\]
Variables $Y_1, \ldots, Y_n$ are independent, with $Y_i$ having a density $p(y | \mu_i)$ governed by an unknown parameter $\mu_i$. Define the *deviance* for a model $M$ that imposes relationships between the $(\mu_i)$.

From this point on, suppose $Y_i \sim \text{Poisson}(\mu_i)$. Write down the log-likelihood of data $y_1, \ldots, y_n$ as a function of $\mu_1, \ldots, \mu_n$.

Let $\hat{\mu}_i$ be the maximum likelihood estimate of $\mu_i$ under model $M$. Show that the deviance for this model is given by

$$2 \sum_{i=1}^{n} \left\{ y_i \log \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i) \right\}.$$

Now suppose that, under $M$, $\log \mu_i = \beta^T x_i$, $i = 1, \ldots, n$, where $x_1, \ldots, x_n$ are known $p$-dimensional explanatory variables and $\beta$ is an unknown $p$-dimensional parameter. Show that $\hat{\mu} := (\hat{\mu}_1, \ldots, \hat{\mu}_n)^T$ satisfies $X^T y = X^T \hat{\mu}$, where $y = (y_1, \ldots, y_n)^T$ and $X$ is the $(n \times p)$ matrix with rows $x_1^T, \ldots, x_n^T$, and express this as an equation for the maximum likelihood estimate $\hat{\beta}$ of $\beta$. [You are not required to solve this equation.]
Let $f_0$ be a probability density function, with cumulant generating function $K$. Define what it means for a random variable $Y$ to have a model function of exponential dispersion family form, generated by $f_0$.

A random variable $Y$ is said to have an inverse Gaussian distribution, with parameters $\phi$ and $\lambda$ (both positive), if its density function is

$$f(y; \phi, \lambda) = \frac{\sqrt{\lambda}}{\sqrt{2\pi y^3}} e^{\sqrt{\lambda \phi}} \exp \left\{ -\frac{1}{2} \left( \frac{\lambda}{y} + \phi y \right) \right\} \quad (y > 0).$$

Show that the family of all inverse Gaussian distributions for $Y$ is of exponential dispersion family form. Deduce directly the corresponding expressions for $E(Y)$ and $\text{Var}(Y)$ in terms of $\phi$ and $\lambda$. What are the corresponding canonical link function and variance function?

Consider a generalized linear model, $M$, for independent variables $Y_i$ ($i = 1, \ldots, n$), whose random component is defined by the inverse Gaussian distribution with link function $g(\mu) = \log(\mu)$: thus $g(\mu_i) = x_i^T \beta$, where $\beta = (\beta_1, \ldots, \beta_p)^T$ is the vector of unknown regression coefficients and $x_i = (x_{i1}, \ldots, x_{ip})^T$ is the vector of known values of the explanatory variables for the $i$th observation. The vectors $x_i$ ($i = 1, \ldots, n$) are linearly independent. Assuming that the dispersion parameter is known, obtain expressions for the score function and Fisher information matrix for $\beta$. Explain how these can be used to compute the maximum likelihood estimate $\hat{\beta}$ of $\beta$. 
Paper 1, Section II
13J Statistical Modelling

A cricket ball manufacturing company conducts the following experiment. Every
day, a bowling machine is set to one of three levels, “Medium”, “Fast” or “Spin”, and
then bowls 100 balls towards the stumps. The number of times the ball hits the stumps
and the average wind speed (in kilometres per hour) during the experiment are recorded,
yielding the following data (abbreviated):

<table>
<thead>
<tr>
<th>Day</th>
<th>Wind</th>
<th>Level</th>
<th>Stumps</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>Medium</td>
<td>26</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>Medium</td>
<td>37</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>50</td>
<td>12</td>
<td>Medium</td>
<td>32</td>
</tr>
<tr>
<td>51</td>
<td>7</td>
<td>Fast</td>
<td>31</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>120</td>
<td>3</td>
<td>Fast</td>
<td>28</td>
</tr>
<tr>
<td>121</td>
<td>5</td>
<td>Spin</td>
<td>35</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>150</td>
<td>6</td>
<td>Spin</td>
<td>31</td>
</tr>
</tbody>
</table>

Write down a reasonable model for $Y_1, \ldots, Y_{150}$, where $Y_i$ is the number of times the ball
hits the stumps on the $i^{th}$ day. Explain briefly why we might want to include interactions
between the variables. Write R code to fit your model.

The company’s statistician fitted her own generalized linear model using R, and
obtained the following summary (abbreviated):

```r
>summary(ball)
Coefficients:
Estimate  Std. Error   z value  Pr(>|z|)
(Intercept)  -0.37258  0.05388  -6.916  4.66e-12 ***
Wind        0.09055  0.01595   5.676  1.38e-08 ***
LevelFast   -0.10005  0.08044  -1.244  0.213570
LevelSpin  0.29881  0.08268   3.614  0.000301 ***
Wind:LevelFast  0.03666  0.02364   1.551  0.120933
Wind:LevelSpin -0.07697  0.02845  -2.705  0.006825 **
```

Why are LevelMedium and Wind:LevelMedium not listed?

Suppose that, on another day, the bowling machine is set to “Spin”, and the
wind speed is 5 kilometres per hour. What linear function of the parameters should
the statistician use in constructing a predictor of the number of times the ball hits the
stumps that day?

Based on the above output, how might you improve the model? How could you fit
your new model in R?
Define the concepts of an exponential dispersion family and the corresponding variance function. Show that the family of Poisson distributions with parameter $\lambda > 0$ is an exponential dispersion family. Find the corresponding variance function and deduce from it expressions for $E(Y)$ and $\text{Var}(Y)$ when $Y \sim \text{Pois}(\lambda)$. What is the canonical link function in this case?

Consider the linear model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i,$$

for $i = 1, 2, \ldots, n$, where the $\varepsilon_i$ are independent and identically distributed with $N(0, \sigma^2)$ distribution. What does it mean for the pair $\beta_1$ and $\beta_2$ to be orthogonal? What does it mean for all the three parameters $\beta_0, \beta_1$ and $\beta_2$ to be mutually orthogonal? Give necessary and sufficient conditions on $(x_{i1})_{i=1}^n, (x_{i2})_{i=1}^n$ so that $\beta_0, \beta_1$ and $\beta_2$ are mutually orthogonal. If $\beta_0, \beta_1, \beta_2$ are mutually orthogonal, find the joint distribution of the corresponding maximum likelihood estimators $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$. 
The purpose of the following study is to investigate differences among certain treatments on the lifespan of male fruit flies, after allowing for the effect of the variable ‘thorax length’ (\textit{thorax}) which is known to be positively correlated with lifespan. Data was collected on the following variables:

\textit{longevity}  
\textit{lifespan in days}

\textit{thorax}  
\textit{(body) length in mm}

\textit{treat}  
a five level factor representing the treatment groups. The levels were labelled as follows: “00”, “10”, “80”, “11”, “81”.

No interactions were found between thorax length and the treatment factor. A linear model with \textit{thorax} as the covariate, \textit{treat} as a factor (having the above 5 levels) and \textit{longevity} as the response was fitted and the following output was obtained. There were 25 males in each of the five groups, which were treated identically in the provision of fresh food.

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | -49.98   | 10.61      | -4.71   | 6.7e-06  |
| treat10        | 2.65     | 2.98       | 0.89    | 0.37     |
| treat11        | -7.02    | 2.97       | -2.36   | 0.02     |
| treat80        | 3.93     | 3.00       | 1.31    | 0.19     |
| treat81        | -19.95   | 3.01       | -6.64   | 1.0e-09  |
| thorax         | 135.82   | 12.44      | 10.92   | <2e-16   |

Residual standard error: 10.5 on 119 degrees of freedom
Multiple R-Squared: 0.656, Adjusted R-squared: 0.642
F-statistics: 45.5 on 5 and 119 degrees of freedom, p-value: 0

(a) Assuming the same treatment, how much longer would you expect a fly with a thorax length 0.1mm greater than another to live?

(b) What is the predicted difference in longevity between a male fly receiving treatment \textit{treat10} and \textit{treat81} assuming they have the same thorax length?

(c) Because the flies were randomly assigned to the five groups, the distribution of thorax lengths in the five groups are essentially equal. What disadvantage would the investigators have incurred by ignoring the thorax length in their analysis (i.e., had they done a one-way ANOVA instead)?

(d) The residual-fitted plot is shown in the left panel of Figure 1 overleaf. Is it possible to determine if the regular residuals or the studentized residuals have been used to construct this plot? Explain.

(e) The Box–Cox procedure was used to determine a good transformation for this data. The plot of the log-likelihood for \( \lambda \) is shown in the right panel of Figure 1. What transformation should be used to improve the fit and yet retain some interpretability?
Let $Y_1, \ldots, Y_n$ be independent with $Y_i \sim \binom{n_i}{\mu_i}, i = 1, \ldots, n$, and
\[ \log \left( \frac{\mu_i}{1 - \mu_i} \right) = x_i^\top \beta, \] (1)
where $x_i$ is a $p \times 1$ vector of regressors and $\beta$ is a $p \times 1$ vector of parameters. Write down the likelihood of the data $Y_1, \ldots, Y_n$ as a function of $\mu = (\mu_1, \ldots, \mu_n)$. Find the unrestricted maximum likelihood estimator of $\mu$, and the form of the maximum likelihood estimator $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_n)$ under the logistic model (1).

Show that the deviance for a comparison of the full (saturated) model to the generalised linear model with canonical link (1) using the maximum likelihood estimator $\hat{\beta}$ can be simplified to
\[ D(y; \hat{\mu}) = -2 \sum_{i=1}^n \left[ n_i y_i x_i^\top \hat{\beta} - n_i \log(1 - \hat{\mu}_i) \right]. \]

Finally, obtain an expression for the deviance residual in this generalised linear model.
Let \((X_1, Y_1), \ldots, (X_n, Y_n)\) be jointly independent and identically distributed with \(X_i \sim N(0, 1)\) and conditional on \(X_i = x\), \(Y_i \sim N(x\theta, 1)\), \(i = 1, 2, \ldots, n\).

(a) Write down the likelihood of the data \((X_1, Y_1), \ldots, (X_n, Y_n)\), and find the maximum likelihood estimate \(\hat{\theta}\) of \(\theta\). [You may use properties of conditional probability/expectation without providing a proof.]

(b) Find the Fisher information \(I(\theta)\) for a single observation, \((X_1, Y_1)\).

(c) Determine the limiting distribution of \(\sqrt{n}(\hat{\theta} - \theta)\). [You may use the result on the asymptotic distribution of maximum likelihood estimators, without providing a proof.]

(d) Give an asymptotic confidence interval for \(\theta\) with coverage \((1 - \alpha)\) using your answers to (b) and (c).

(e) Define the observed Fisher information. Compare the confidence interval in part (d) with an asymptotic confidence interval with coverage \((1 - \alpha)\) based on the observed Fisher information.

(f) Determine the exact distribution of \(\left(\sum_{i=1}^n X_i^2\right)^{1/2}(\hat{\theta} - \theta)\) and find the true coverage probability for the interval in part (e). [Hint. Condition on \(X_1, X_2, \ldots, X_n\) and use the following property of conditional expectation: for \(U, V\) random vectors, any suitable function \(g\), and \(x \in \mathbb{R}\),
\[
P\{g(U, V) \leq x\} = E[P\{g(U, V) \leq x|V\}].
\]
The treatment for a patient diagnosed with cancer of the prostate depends on whether the cancer has spread to the surrounding lymph nodes. It is common to operate on the patient to obtain samples from the nodes which can then be analysed under a microscope. However it would be preferable if an accurate assessment of nodal involvement could be made without surgery. For a sample of 53 prostate cancer patients, a number of possible predictor variables were measured before surgery. The patients then had surgery to determine nodal involvement. We want to see if nodal involvement can be accurately predicted from the available variables and determine which ones are most important. The variables take the values 0 or 1.

- \( r \) An indicator 0=no/1=yes of nodal involvement.
- \( \text{aged} \) The patient’s age, split into less than 60 (=0) and 60 or over (=1).
- \( \text{stage} \) A measurement of the size and position of the tumour observed by palpation with the fingers. A serious case is coded as 1 and a less serious case as 0.
- \( \text{grade} \) Another indicator of the seriousness of the cancer which is determined by a pathology reading of a biopsy taken by needle before surgery. A value of 1 indicates a more serious case of cancer.
- \( \text{xray} \) Another measure of the seriousness of the cancer taken from an X-ray reading. A value of 1 indicates a more serious case of cancer.
- \( \text{acid} \) The level of acid phosphatase in the blood serum where 1=high and 0=low.

A binomial generalised linear model with a logit link was fitted to the data to predict nodal involvement and the following output obtained:

![Deviance Residuals](image)

![Coefficients](image)

Null deviance: 70.252 on 52 degrees of freedom  
Residual deviance: 47.611 on 47 degrees of freedom  
AIC: 59.61

Number of Fisher Scoring iterations: 5
(a) Give an interpretation of the coefficient of xray.

(b) Give the numerical value of the sum of the squared deviance residuals.

(c) Suppose that the predictors, stage, grade and xray are positively correlated. Describe the effect that this correlation is likely to have on our ability to determine the strength of these predictors in explaining the response.

(d) The probability of observing a value of 70.252 under a Chi-squared distribution with 52 degrees of freedom is 0.047. What does this information tell us about the null model for this data? Justify your answer.

(e) What is the lowest predicted probability of the nodal involvement for any future patient?

(f) The first plot in Figure 1 shows the (Pearson) residuals and the fitted values. Explain why the points lie on two curves.

(g) The second plot in Figure 1 shows the value of $\hat{\beta} - \hat{\beta}_i(i)$ where $(i)$ indicates that patient $i$ was dropped in computing the fit. The values for each predictor, including the intercept, are shown. Could a single case change our opinion of which predictors are important in predicting the response?

![Figure 1](image-url)
Paper 1, Section I

5J Statistical Modelling

Let \( Y_1, \ldots, Y_n \) be independent identically distributed random variables with model function \( f(y, \theta) \), \( y \in \mathcal{Y}, \theta \in \Theta \subseteq \mathbb{R} \), and denote by \( E_\theta \) and \( \text{Var}_\theta \) expectation and variance under \( f(y, \theta) \), respectively. Define \( U_n(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(Y_i, \theta) \). Prove that \( E_\theta U_n(\theta) = 0 \). Show moreover that if \( T = T(Y_1, \ldots, Y_n) \) is any unbiased estimator of \( \theta \), then its variance satisfies \( \text{Var}_\theta(T) \geq n \text{Var}_\theta(U_1(\theta))^{-1} \). [You may use the Cauchy–Schwarz inequality without proof, and you may interchange differentiation and integration without justification if necessary.]

Paper 2, Section I

5J Statistical Modelling

Let \( f_0 \) be a probability density function, with cumulant generating function \( K \). Define what it means for a random variable \( Y \) to have a model function of exponential dispersion family form, generated by \( f_0 \). Compute the cumulant generating function \( K_Y \) of \( Y \) and deduce expressions for the mean and variance of \( Y \) that depend only on first and second derivatives of \( K \).

Paper 3, Section I

5J Statistical Modelling

Define a generalised linear model for a sample \( Y_1, \ldots, Y_n \) of independent random variables. Define further the concept of the link function. Define the binomial regression model with logistic and probit link functions. Which of these is the canonical link function?
The numbers of ear infections observed among beach and non-beach (mostly pool) swimmers were recorded, along with explanatory variables: frequency, location, age, and sex. The data are aggregated by group, with a total of 24 groups defined by the explanatory variables:

- **freq** \( F = \text{frequent}, \ NF = \text{infrequent} \)
- **loc** \( \NB = \text{non-beach}, \ B = \text{beach} \)
- **age** \( 15-19, \ 20-24, \ 24-29 \)
- **sex** \( F = \text{female}, \ M = \text{male} \)
- **count** the number of infections reported over a fixed time period
- **n** the total number of swimmers

The data look like this:

<table>
<thead>
<tr>
<th>count</th>
<th>n</th>
<th>freq</th>
<th>loc</th>
<th>sex</th>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>68</td>
<td>F</td>
<td>NB</td>
<td>M</td>
<td>15-19</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
<td>F</td>
<td>NB</td>
<td>F</td>
<td>15-19</td>
</tr>
<tr>
<td>3</td>
<td>35</td>
<td>F</td>
<td>NB</td>
<td>M</td>
<td>20-24</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>F</td>
<td>NB</td>
<td>F</td>
<td>20-24</td>
</tr>
<tr>
<td>[...]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>5</td>
<td>NF</td>
<td>B</td>
<td>M</td>
<td>25-29</td>
</tr>
<tr>
<td>24</td>
<td>6</td>
<td>NF</td>
<td>B</td>
<td>F</td>
<td>25-29</td>
</tr>
</tbody>
</table>

Let \( \mu_j \) denote the expected number of ear infections of a person in group \( j \). Explain why it is reasonable to model \( \text{count}_j \) as Poisson with mean \( n_j \mu_j \).

We fit the following Poisson model:

\[
\log(\mathbb{E}(\text{count}_j)) = \log(n_j \mu_j) = \log(n_j) + x_j \beta,
\]

where \( \log(n_j) \) is an offset, i.e. an explanatory variable with known coefficient 1.

R produces the following (abbreviated) summary for the main effects model:

```
Call: glm(formula = count ~ freq + loc + age + sex, family = poisson, offset = log(n))

Coefficients:      Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.48887   0.12271  3.984 6.78e-05 ***
freqNF -0.61149   0.10500 -5.823 5.76e-09 ***
locNB 0.53454   0.10668  5.011 5.43e-07 ***
age20-24 -0.37442   0.12836 -2.917 0.00354 **
age25-29 -0.18973   0.13009 -1.458 0.14473
sexM -0.08985   0.11231 -0.800 0.42371
```

Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
Why are expressions freqF, locB, age15-19, and sexF not listed?

Suppose that we plan to observe a group of 20 female, non-frequent, beach swimmers, aged 20-24. Give an expression (using the coefficient estimates from the model fitted above) for the expected number of ear infections in this group.

Now, suppose that we allow for interaction between variables age and sex. Give the R command for fitting this model. We test for the effect of this interaction by producing the following (abbreviated) ANOVA table:

| Resid. Df | Resid. Dev | Df | Deviance | P(>|Chi|) |
|-----------|------------|----|----------|---------|
| 1         | 18         | 51.714 |
| 2         | 16         | 44.319 | 2       | 7.3948  | 0.02479 * |

---

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Briefly explain what test is performed, and what you would conclude from it. Does either of these models fit the data well?
Paper 1, Section II

13J Statistical Modelling

The data consist of the record times in 1984 for 35 Scottish hill races. The columns list the record time in minutes, the distance in miles, and the total height gained during the route. The data are displayed in R as follows (abbreviated):

```r
> hills
          dist  climb  time
Greenmantle 2.5  650    16.083
Carnethy    6.0 2500   48.350
Craig Dunain 6.0  900   33.650
Ben Rha     7.5  800   45.600
Ben Lomond  8.0 3070   62.267
[...]
Cockleroi   4.5  850   28.100
Moffat Chase 20.0 5000 159.833
```

Consider a simple linear regression of time on dist and climb. Write down this model mathematically, and explain any assumptions that you make. How would you instruct R to fit this model and assign it to a variable `hills.lm1`?

First, we test the hypothesis of no linear relationship to the variables `dist` and `climb` against the full model. R provides the following ANOVA summary:

```r
              Res.Df    RSS  Df Sum of Sq    F     Pr(>F)
1             34 85138
2             32 6892  2   78247 181.66 < 2.2e-16 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
```

Using the information in this table, explain carefully how you would test this hypothesis. What do you conclude?

The R command

```r
summary(hills.lm1)
```

provides the following (slightly abbreviated) summary:

```r
[...]
Coefficients:               Estimate  Std. Error   t value  Pr(>|t|)
(Intercept)    -8.9920390  4.3027340   -2.0900  0.04472 *
dist            6.2179558  0.6011478   10.3429 9.861e-12 ***
climb           0.0110476  0.0020511    5.3870 6.450e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
[...]
```
Carefully explain the information that appears in each column of the table. What are your conclusions? In particular, how would you test for the significance of the variable climb in this model?

Figure 1: Hills data: diagnostic plots

Finally, we perform model diagnostics on the full model, by looking at studentised residuals versus fitted values, and the normal QQ-plot. The plots are displayed in Figure 1. Comment on possible sources of model misspecification. Is it possible that the problem lies with the data? If so, what do you suggest?

Paper 4, Section II
13J Statistical Modelling

Consider the general linear model $Y = X\beta + \epsilon$, where the $n \times p$ matrix $X$ has full rank $p \leq n$, and where $\epsilon$ has a multivariate normal distribution with mean zero and covariance matrix $\sigma^2I_n$. Write down the likelihood function for $\beta, \sigma^2$ and derive the maximum likelihood estimators $\hat{\beta}, \hat{\sigma}^2$ of $\beta, \sigma^2$. Find the distribution of $\hat{\beta}$. Show further that $\hat{\beta}$ and $\hat{\sigma}^2$ are independent.
Consider a binomial generalised linear model for data $y_1, \ldots, y_n$ modelled as realisations of independent $Y_i \sim \text{Bin}(1, \mu_i)$ and logit link $\mu_i = e^{\beta x_i} / (1 + e^{\beta x_i})$ for some known constants $x_i$, $i = 1, \ldots, n$, and unknown scalar parameter $\beta$. Find the log-likelihood for $\beta$, and the likelihood equation that must be solved to find the maximum likelihood estimator $\hat{\beta}$ of $\beta$. Compute the second derivative of the log-likelihood for $\beta$, and explain the algorithm you would use to find $\hat{\beta}$.

Suppose you have a parametric model consisting of probability mass functions $f(y; \theta)$, $\theta \in \Theta \subset \mathbb{R}$. Given a sample $Y_1, \ldots, Y_n$ from $f(y; \theta)$, define the maximum likelihood estimator $\hat{\theta}_n$ for $\theta$ and, assuming standard regularity conditions hold, state the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$.

Compute the Fisher information of a single observation in the case where $f(y; \theta)$ is the probability mass function of a Poisson random variable with parameter $\theta$. If $Y_1, \ldots, Y_n$ are independent and identically distributed random variables having a Poisson distribution with parameter $\theta$, show that $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$ and $S = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$ are unbiased estimators for $\theta$. Without calculating the variance of $S$, show that there is no reason to prefer $S$ over $\bar{Y}$.

[You may use the fact that the asymptotic variance of $\sqrt{n}(\hat{\theta}_n - \theta)$ is a lower bound for the variance of any unbiased estimator.]

Consider the linear model $Y = X\beta + \varepsilon$, where $Y$ is a $n \times 1$ random vector, $\varepsilon \sim N_n(0, \sigma^2 I)$, and where the $n \times p$ nonrandom matrix $X$ is known and has full column rank $p$. Derive the maximum likelihood estimator $\hat{\sigma}^2$ of $\sigma^2$. Without using Cochran’s theorem, show carefully that $\hat{\sigma}^2$ is biased. Suggest another estimator $\tilde{\sigma}^2$ for $\sigma^2$ that is unbiased.
Below is a simplified 1993 dataset of US cars. The columns list index, make, model, price (in $1000), miles per gallon, number of passengers, length and width in inches, and weight (in pounds). The data are displayed in R as follows (abbreviated):

```r
> cars
   make model price mpg psngr length width weight
 1 Acura Integra 15.9 31 5 177  68  2705
 2 Acura Legend 33.9 25 5 195  71  3560
 3 Audi 90 90.1 29 5 180  67  3375
 4 Audi 100 37.7 26 6 193  70  3405
 5 BMW  535i 30.0 30 4 186  69  3640
 6 ... ...
 92 Volvo 240 22.7 28 5 190  67  2985
 93 Volvo 850 26.7 28 5 184  69  3245
```

It is reasonable to assume that prices for different makes of car are independent. We model the logarithm of the price as a linear combination of the other quantitative properties of the cars and an error term. Write down this model mathematically. How would you instruct R to fit this model and assign it to a variable “fit”?

R provides the following (slightly abbreviated) summary:

```r
> summary(fit)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.8751 0.7687  5.041  2.50e-06 ***
mpg          -0.0109 0.0085 -1.286  2.01e-02
psngr        -0.1783 0.0291 -6.135  2.45e-08 ***
length       -0.0067 0.0033 -2.049  0.0435 *
width         0.0139 0.0155  0.913  0.3621
weight        0.0202 0.0001  0.202  0.0060 ***
```

Briefly explain the information that is being provided in each column of the table. What are your conclusions and how would you try to improve the model?
Consider a generalised linear model with parameter $\beta^\top$ partitioned as $(\beta_0^\top, \beta_1^\top)$, where $\beta_0$ has $p_0$ components and $\beta_1$ has $p - p_0$ components, and consider testing

$$H_0 : \beta_1 = 0 \text{ against } H_1 : \beta_1 \neq 0.$$ 

Define carefully the deviance, and use it to construct a test for $H_0$.

[You may use Wilks’ theorem to justify this test, and you may also assume that the dispersion parameter is known.]

Now consider the generalised linear model with Poisson responses and the canonical link function with linear predictor $\eta = (\eta_1, \ldots, \eta_n)^T$ given by $\eta_i = x_i^\top \beta$, $i = 1, \ldots, n$, where $x_{i1} = 1$ for every $i$. Derive the deviance for this model, and argue that it may be approximated by Pearson’s $\chi^2$ statistic.
Every day, Barney the darts player comes to our laboratory. We record his facial expression, which can be either “mad”, “weird” or “relaxed”, as well as how many units of beer he has drunk that day. Each day he tries a hundred times to hit the bull’s-eye, and we write down how often he succeeds. The data look like this:

```
Day Beer Expression BullsEye
1 3 Mad 30
2 3 Mad 32
60 2 Mad 37
61 4 Weird 30
110 4 Weird 28
111 2 Relaxed 35
150 3 Relaxed 31
```

Write down a reasonable model for \( Y_1, \ldots, Y_n \), where \( n = 150 \) and where \( Y_i \) is the number of times Barney has hit bull’s-eye on the \( i \)th day. Explain briefly why we may wish initially to include interactions between the variables. Write the R code to fit your model.

The scientist of the above story fitted her own generalized linear model, and subsequently obtained the following summary (abbreviated):

```
> summary(barney)

Coefficients:
              Estimate Std. Error  z value Pr(>|z|)
(Intercept)  -0.37258  0.05388  -6.916  4.66e-12 ***
Beer         -0.09055  0.01595  -5.676  1.38e-08 ***
ExpressionWeird -0.10005  0.08044  -1.244   0.213570
ExpressionRelaxed  0.29881  0.08268   3.614  0.000301 ***
Beer:ExpressionWeird  0.03666  0.02364   1.551   0.120933
Beer:ExpressionRelaxed -0.07697  0.02845  -2.705  0.006825 **
```

Why are ExpressionMad and Beer:ExpressionMad not listed? Suppose on a particular day, Barney’s facial expression is weird, and he drank three units of beer. Give the linear predictor in the scientist’s model for this day.

Based on the summary, how could you improve your model? How could one fit this new model in R (without modifying the data file)?
Paper 1, Section I
5I Statistical Modelling
Consider a binomial generalised linear model for data $y_1, \ldots, y_n$, modelled as realisations of independent $Y_i \sim \text{Bin}(1, \mu_i)$ and logit link, i.e. $\log \frac{\mu_i}{1-\mu_i} = \beta x_i$, for some known constants $x_1, \ldots, x_n$, and an unknown parameter $\beta$. Find the log-likelihood for $\beta$, and the likelihood equations that must be solved to find the maximum likelihood estimator $\hat{\beta}$ of $\beta$.

Compute the first and second derivatives of the log-likelihood for $\beta$, and explain the algorithm you would use to find $\hat{\beta}$.

Paper 2, Section I
5I Statistical Modelling
What is meant by an exponential dispersion family? Show that the family of Poisson distributions with parameter $\lambda$ is an exponential dispersion family by explicitly identifying the terms in the definition.

Find the corresponding variance function and deduce directly from your calculations expressions for $\mathbb{E}(Y)$ and $\text{Var}(Y)$ when $Y \sim \text{Pois}(\lambda)$.

What is the canonical link function in this case?

Paper 3, Section I
5I Statistical Modelling
Consider the linear model $Y = X \beta + \varepsilon$, where $\varepsilon \sim N_n(0, \sigma^2 I)$ and $X$ is an $n \times p$ matrix of full rank $p < n$. Suppose that the parameter $\beta$ is partitioned into $k$ sets as follows: $\beta^T = (\beta_1^T \cdots \beta_k^T)$. What does it mean for a pair of sets $\beta_i, \beta_j$, $i \neq j$, to be orthogonal? What does it mean for all $k$ sets to be mutually orthogonal?

In the model
$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$
where $\varepsilon_i \sim N(0, \sigma^2)$ are independent and identically distributed, find necessary and sufficient conditions on $x_{11}, \ldots, x_{n1}, x_{12}, \ldots, x_{n2}$ for $\beta_0, \beta_1$ and $\beta_2$ to be mutually orthogonal.

If $\beta_0, \beta_1$ and $\beta_2$ are mutually orthogonal, what consequence does this have for the joint distribution of the corresponding maximum likelihood estimators $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$?
Sulphur dioxide is one of the major air pollutants. A dataset by Sokal and Rohlf (1981) was collected on 41 US cities/regions in 1969–1971. The annual measurements obtained for each region include (average) sulphur dioxide content, temperature, number of manufacturing enterprises employing more than 20 workers, population size in thousands, wind speed, precipitation, and the number of days with precipitation. The data are displayed in R as follows (abbreviated):

```r
> usair
region so2 temp manuf pop wind precip days
1 Phoenix 10 70.3 213 582 6.0 7.05 36
2 Little Rock 13 61.0 91 132 8.2 48.52 100
... ... ... ... ...
41 Milwaukee 16 45.7 569 717 11.8 29.07 123
```

Describe the model being fitted by the following R commands.

```r
> fit <- lm(log(so2) ~ temp + manuf + pop + wind + precip + days)
```

Explain the (slightly abbreviated) output below, describing in particular how the hypothesis tests are performed and your conclusions based on their results:

```r
> summary(fit)
```

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.2532456 1.4483686 5.008 1.68e-05 ***
temp -0.0599017 0.0190138 -3.150 0.00339 **
manuf 0.0012639 0.0004820 2.622 0.01298 *
pop -0.0007077 0.0004632 -1.528 0.13580
wind -0.1697171 0.0555563 -3.055 0.00436 **
precip 0.0173723 0.0111036 1.565 0.12695
days 0.0004347 0.0049591 0.088 0.93066

Residual standard error: 0.448 on 34 degrees of freedom
```

Based on the summary above, suggest an alternative model.

Finally, what is the value obtained by the following command?

```r
> sqrt(sum(resid(fit)^2)/fit$df)
```
A three-year study was conducted on the survival status of patients suffering from cancer. The age of the patients at the start of the study was recorded, as well as whether or not the initial tumour was malignant. The data are tabulated in R as follows:

<table>
<thead>
<tr>
<th></th>
<th>cancer</th>
<th>age</th>
<th>malignant</th>
<th>survive</th>
<th>die</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>cancer</td>
<td>&lt;50</td>
<td>no</td>
<td>77</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>cancer</td>
<td>&lt;50</td>
<td>yes</td>
<td>51</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>cancer</td>
<td>50-69</td>
<td>no</td>
<td>51</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>cancer</td>
<td>50-69</td>
<td>yes</td>
<td>38</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>cancer</td>
<td>70+</td>
<td>no</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>cancer</td>
<td>70+</td>
<td>yes</td>
<td>6</td>
<td>3</td>
</tr>
</tbody>
</table>

Describe the model that is being fitted by the following R commands:

```r
> total <- survive + die
> fit1 <- glm(survive/total ~ age + malignant, family = binomial, + weights = total)
```

Explain the (slightly abbreviated) output from the code below, describing how the hypothesis tests are performed and your conclusions based on their results.

```r
> summary(fit1)
```

**Coefficients:**

|             | Estimate | Std. Error | z value | Pr(>|z|) |
|-------------|----------|------------|---------|----------|
| (Intercept) | 2.0730   | 0.2812     | 7.372   | 1.68e-13 *** |
| age50-69    | -0.6318  | 0.3112     | -2.030  | 0.0424 *  |
| age70+      | -0.9282  | 0.5504     | -1.686  | 0.0917 .   |
| malignantyes| -0.7328  | 0.2985     | -2.455  | 0.0141 *  |

----

Null deviance: 12.65585 on 5 degrees of freedom
Residual deviance: 0.49409 on 2 degrees of freedom
AIC: 30.433

Based on the summary above, motivate and describe the following alternative model:

```r
> age2 <- as.factor(c("<50", "<50", "50+", "50+", "50+", "50+"))
> fit2 <- glm(survive/total ~ age2 + malignant, family = binomial, + weights = total)
```

This question continues on the next page
Based on the output of the code that follows, which of the two models do you prefer? Why?

> summary(fit2)

Coefficients:

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|----------|
| (Intercept) | 2.0721 | 0.2811 | 7.372 | 1.68e-13 *** |
| age250+ | -0.6744 | 0.3000 | -2.248 | 0.0246 * |
| malignantyes | -0.7310 | 0.2983 | -2.451 | 0.0143 * |

---

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Null deviance: 12.656 on 5 degrees of freedom
Residual deviance: 0.784 on 3 degrees of freedom
AIC: 28.723

What is the final value obtained by the following commands?

> mu.hat <- inv.logit(predict(fit2))
> -2 * (sum(dbinom(survive, total, mu.hat, log = TRUE)
+ - sum(dbinom(survive, total, survive/total, log = TRUE)))

Paper 4, Section II

13I Statistical Modelling

Consider the linear model $Y = X\beta + \varepsilon$, where $\varepsilon \sim N_n(0, \sigma^2I)$ and $X$ is an $n \times p$ matrix of full rank $p < n$. Find the form of the maximum likelihood estimator $\hat{\beta}$ of $\beta$, and derive its distribution assuming that $\sigma^2$ is known.

Assuming the prior $\pi(\beta, \sigma^2) \propto \sigma^{-2}$ find the joint posterior of $(\beta, \sigma^2)$ up to a normalising constant. Derive the posterior conditional distribution $\pi(\beta|\sigma^2, X, Y)$.

Comment on the distribution of $\hat{\beta}$ found above and the posterior conditional $\pi(\beta|\sigma^2, X, Y)$. Comment further on the predictive distribution of $y^*$ at input $x^*$ under both the maximum likelihood and Bayesian approaches.
Consider the following Binomial generalized linear model for data \( y_1, \ldots, y_n \), with logit link function. The data \( y_1, \ldots, y_n \) are regarded as observed values of independent random variables \( Y_1, \ldots, Y_n \), where

\[
Y_i \sim \text{Bin}(1, \mu_i), \quad \log \frac{\mu_i}{1 - \mu_i} = \beta^\top x_i, \quad i = 1, \ldots, n,
\]

where \( \beta \) is an unknown \( p \)-dimensional parameter, and where \( x_1, \ldots, x_n \) are known \( p \)-dimensional explanatory variables. Write down the likelihood function for \( y = (y_1, \ldots, y_n) \) under this model.

Show that the maximum likelihood estimate \( \hat{\beta} \) satisfies an equation of the form

\[
X^\top y = X^\top \hat{\mu},
\]

where \( X \) is the \( p \times n \) matrix with rows \( x_1^\top, \ldots, x_n^\top \), and where \( \hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_n) \), with \( \hat{\mu}_i \) a function of \( x_i \) and \( \hat{\beta} \), which you should specify.

Define the deviance \( D(y; \hat{\mu}) \) and find an explicit expression for \( D(y; \hat{\mu}) \) in terms of \( y \) and \( \hat{\mu} \) in the case of the model above.
Consider performing a two-way analysis of variance (ANOVA) on the following data:

\[
\begin{array}{ccc}
[1,] & 2.72 & 6.66 \\
[2,] & 4.88 & 5.98 \\
[3,] & 3.49 & 8.81 \\
[4,] & 2.03 & 6.26 \\
[5,] & 2.39 & 8.50 \\
\end{array}
\quad
\begin{array}{ccc}
[1,] & -5.780 & 1.7200 \\
[2,] & -4.600 & 1.9800 \\
[3,] & -1.460 & 2.1500 \\
[4,] & -1.780 & -0.5120 \\
[5,] & -2.610 & 0.0637 \\
\end{array}
\quad
\begin{array}{ccc}
[1,] & 1.7200 & 0.158 \\
[2,] & 1.9800 & 1.190 \\
[3,] & 2.1500 & 1.190 \\
[4,] & 0.7090 & 1.470 \\
[5,] & -0.5120 & 2.110 \\
\end{array}
\]

Explain and interpret the R commands and (slightly abbreviated) output below. In particular, you should describe the model being fitted, and comment on the hypothesis tests which are performed under the `summary` and `anova` commands.

```r
> K <- dim(Y)[1]
> I <- dim(Y)[2]
> J <- dim(Y)[3]
> c(I,J,K)
[1] 2 3 10
> y <- as.vector(Y)
> a <- gl(I, K, length(y))
> b <- gl(J, K * I, length(y))
> fit1 <- lm(y ~ a + b)
> summary(fit1)

Coefficients:

|            | Estimate | Std. Error | t value | Pr(>|t|) |
|------------|----------|------------|---------|----------|
| (Intercept)| 3.7673   | 0.3032     | 12.43   | < 2e-16  ***|
| a2         | 3.4542   | 0.3032     | 11.39   | 3.27e-16 ***|
| b2         | -6.3215  | 0.3713     | -17.03  | < 2e-16  ***|
| b3         | -5.8268  | 0.3713     | -15.69  | < 2e-16  ***|
```

> anova(fit1)
Response: y

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1</td>
<td>178.98</td>
<td>178.98</td>
<td>129.83</td>
<td>3.272e-16 ***</td>
</tr>
<tr>
<td>b</td>
<td>2</td>
<td>494.39</td>
<td>247.19</td>
<td>179.31</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>Residuals</td>
<td>56</td>
<td>77.20</td>
<td>1.38</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The following R code fits a similar model. Briefly explain the difference between this model and the one above. Based on the output of the `anova` call below, say whether you prefer this model over the one above, and explain your preference.

```r
> fit2 <- lm(y ~ a * b)
> anova(fit2)
```

Response: y

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1</td>
<td>178.98</td>
<td>178.98</td>
<td>125.6367</td>
<td>1.033e-15 ***</td>
</tr>
<tr>
<td>b</td>
<td>2</td>
<td>494.39</td>
<td>247.19</td>
<td>173.5241</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>a:b</td>
<td>2</td>
<td>0.27</td>
<td>0.14</td>
<td>0.0963</td>
<td>0.9084</td>
</tr>
<tr>
<td>Residuals</td>
<td>54</td>
<td>76.93</td>
<td>1.42</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Finally, explain what is being calculated in the code below and give the value that would be obtained by the final line of code.

```r
> n <- I * J * K
> p <- length(coef(fit2))
> p0 <- length(coef(fit1))
> PY <- fitted(fit2)
> P0Y <- fitted(fit1)
> ((n - p)/(p - p0)) * sum((PY - P0Y)^2)/sum((y - PY)^2)
```

Part II 2008

2008
2/I/5J Statistical Modelling

Suppose that we want to estimate the angles $\alpha$, $\beta$ and $\gamma$ (in radians, say) of the triangle $ABC$, based on a single independent measurement of the angle at each corner. Suppose that the error in measuring each angle is normally distributed with mean zero and variance $\sigma^2$. Thus, we model our measurements $y_A, y_B, y_C$ as the observed values of random variables

$$Y_A = \alpha + \varepsilon_A, \quad Y_B = \beta + \varepsilon_B, \quad Y_C = \gamma + \varepsilon_C,$$

where $\varepsilon_A, \varepsilon_B, \varepsilon_C$ are independent, each with distribution $N(0, \sigma^2)$. Find the maximum likelihood estimate of $\alpha$ based on these measurements.

Can the assumption that $\varepsilon_A, \varepsilon_B, \varepsilon_C \sim N(0, \sigma^2)$ be criticized? Why or why not?

3/I/5J Statistical Modelling

Consider the linear model $Y = X\beta + \varepsilon$. Here, $Y$ is an $n$-dimensional vector of observations, $X$ is a known $n \times p$ matrix, $\beta$ is an unknown $p$-dimensional parameter, and $\varepsilon \sim N_n(0, \sigma^2 I)$, with $\sigma^2$ unknown. Assume that $X$ has full rank and that $p \ll n$. Suppose that we are interested in checking the assumption $\varepsilon \sim N_n(0, \sigma^2 I)$. Let $\hat{Y} = X\hat{\beta}$, where $\hat{\beta}$ is the maximum likelihood estimate of $\beta$. Write in terms of $X$ an expression for the projection matrix $P = (p_{ij} : 1 \leq i, j \leq n)$ which appears in the maximum likelihood equation $\hat{Y} = X\hat{\beta} = PY$.

Find the distribution of $\hat{\varepsilon} = Y - \hat{Y}$, and show that, in general, the components of $\hat{\varepsilon}$ are not independent.

A standard procedure used to check our assumption on $\varepsilon$ is to check whether the studentized fitted residuals

$$\hat{\eta}_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - p_{ii}}}, \quad i = 1, \ldots, n,$$

look like a random sample from an $N(0, 1)$ distribution. Here,

$$\hat{\sigma}^2 = \frac{1}{n - p} ||Y - X\hat{\beta}||^2.$$

Say, briefly, how you might do this in R.

This procedure appears to ignore the dependence between the components of $\hat{\varepsilon}$ noted above. What feature of the given set-up makes this reasonable?
Statistical Modelling

A long-term agricultural experiment had \( n = 90 \) grassland plots, each \( 25m \times 25m \), differing in biomass, soil pH, and species richness (the count of species in the whole plot). While it was well-known that species richness declines with increasing biomass, it was not known how this relationship depends on soil pH. In the experiment, there were 30 plots of “low pH”, 30 of “medium pH” and 30 of “high pH”. Three lines of the data are reproduced here as an aid.

```r
> grass[c(1,31, 61), ]
   pH  Biomass Species
 1 high 0.4692972 30
 31 mid 0.1757627 29
 61 low 0.1008479 18
```

Briefly explain the commands below. That is, explain the models being fitted.

```r
> fit1 <- glm(Species ~ Biomass, family = poisson)
> fit2 <- glm(Species ~ pH + Biomass, family = poisson)
> fit3 <- glm(Species ~ pH * Biomass, family = poisson)
```

Let \( H_1, H_2 \) and \( H_3 \) denote the hypotheses represented by the three models and fits. Based on the output of the code below, what hypotheses are being tested, and which of the models seems to give the best fit to the data? Why?

```r
> anova(fit1, fit2, fit3, test = "Chisq")
Analysis of Deviance Table
Model 1: Species ~ Biomass
Model 2: Species ~ pH + Biomass
Model 3: Species ~ pH * Biomass
        Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1          88      407.67
2          86      99.24   2   308.43 1.059e-67
3          84      83.20   2   16.04 3.288e-04
```

Finally, what is the value obtained by the following command?

```r
> mu.hat <- exp(predict(fit2))
> -2 * (sum(dpois(Species, mu.hat, log = TRUE)) - sum(dpois(Species, + Species, log = TRUE)))
```
Consider the following generalized linear model for responses $y_1, \ldots, y_n$ as a function of explanatory variables $x_1, \ldots, x_n$, where $x_i = (x_{i1}, \ldots, x_{ip})^\top$ for $i = 1, \ldots, n$. The responses are modelled as observed values of independent random variables $Y_1, \ldots, Y_n$, with
\[ Y_i \sim ED(\mu_i, \sigma_i^2), \quad g(\mu_i) = x_i^\top \beta, \quad \sigma_i^2 = \sigma^2 a_i, \]
Here, $g$ is a given link function, $\beta$ and $\sigma^2$ are unknown parameters, and the $a_i$ are treated as known.

[Hint: recall that we write $Y \sim ED(\mu, \sigma^2)$ to mean that $Y$ has density function of the form
\[ f(y; \mu, \sigma^2) = a(\sigma^2, y) \exp \left\{ \frac{1}{\sigma^2} [\theta(\mu) y - K(\theta(\mu))] \right\} \]
for given functions $a$ and $\theta$.]

You may use without proof the facts that, for such a random variable $Y$,
\[ E(Y) = K'(\theta(\mu)), \quad \text{var}(Y) = \sigma^2 K''(\theta(\mu)) \equiv \sigma^2 V(\mu). \]

Show that the score vector and Fisher information matrix have entries:
\[ U_j(\beta) = \sum_{i=1}^n \frac{(y_i - \mu_i)x_{ij}}{\sigma_i^2 V(\mu_i) g'(\mu_i)}, \quad j = 1, \ldots, p, \]
and
\[ i_{jk}(\beta) = \sum_{i=1}^n \frac{x_{ij}x_{ik}}{\sigma_i^2 V(\mu_i) (g'(\mu_i))^2}, \quad j, k = 1, \ldots, p. \]

How do these expressions simplify when the canonical link is used?

Explain briefly how these two expressions can be used to obtain the maximum likelihood estimate $\hat{\beta}$ for $\beta$. 

Part II 2008

Part II 2008
According to the *Independent* newspaper (London, 8 March 1994) the Metropolitan Police in London reported 30475 people as missing in the year ending March 1993. For those aged 18 or less, 96 of 10527 missing males and 146 of 11363 missing females were still missing a year later. For those aged 19 and above, the values were 157 of 5065 males and 159 of 3520 females. This data is summarised in the table below.

```
age  gender still  total
 1  Kid    M    96  10527
 2  Kid    F   146  11363
 3 Adult  M   157  5065
 4 Adult  F   159  3520
```

Explain and interpret the R commands and (slightly abbreviated) output below. You should describe the model being fitted, explain how the standard errors are calculated, and comment on the hypothesis tests being described in the summary. In particular, what is the worst of the four categories for the probability of remaining missing a year later?

```
> fit <- glm(still/total ~ age + gender, family = binomial,
+             weights = total)
> summary(fit)
```

Coefficients:

|                     | Estimate | Std. Error | z value | Pr(>|z|) |
|---------------------|----------|------------|---------|----------|
| (Intercept)         | -3.06073 | 0.07216    | -42.417 | < 2e-16  *** |
| ageKid              | -1.27079 | 0.08698    | -14.610 | < 2e-16  *** |
| genderM             | -0.37211 | 0.08671    | -4.291  | 1.78e-05 *** |

Residual deviance: 0.06514 on 1 degrees of freedom

For a person who was missing in the year ending in March 1993, find a formula, as a function of age and gender, for the estimated expected probability that they are still missing a year later.
1/II/13I Statistical Modelling

This problem deals with data collected as the number of each of two different strains of *Ceriodaphnia* organisms are counted in a controlled environment in which reproduction is occurring among the organisms. The experimenter places into the containers a varying concentration of a particular component of jet fuel that impairs reproduction. Hence it is anticipated that as the concentration of jet fuel grows, the mean number of organisms should decrease.

The table below gives a subset of the data. The full dataset has \( n = 70 \) rows. The first column provides the number of organisms, the second the concentration of jet fuel (in grams per litre) and the third specifies the strain of the organism.

<table>
<thead>
<tr>
<th>number</th>
<th>fuel</th>
<th>strain</th>
</tr>
</thead>
<tbody>
<tr>
<td>82</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>58</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>45</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>27</td>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>29</td>
<td>0.75</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>1.25</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1.25</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>1.5</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1.75</td>
<td>0</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

Explain and interpret the R commands and (slightly abbreviated) output below. In particular, you should describe the model being fitted, explain how the standard errors are calculated, and comment on the hypothesis tests being described in the summary.

```r
> fit1 <- glm(number ~ fuel + strain + fuel:strain, family = poisson)
> summary(fit1)
```

<table>
<thead>
<tr>
<th>Coefficients:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate Std. Error z value Pr(</td>
</tr>
<tr>
<td>(Intercept) 4.14443 0.05101 81.252 &lt; 2e-16 ***</td>
</tr>
<tr>
<td>fuel -1.47253 0.07007 -21.015 &lt; 2e-16 ***</td>
</tr>
<tr>
<td>strain 0.33667 0.06704 5.022 5.11e-07 ***</td>
</tr>
<tr>
<td>fuel:strain -0.12534 0.09385 -1.336 0.182</td>
</tr>
</tbody>
</table>

The following R code fits two very similar models. Briefly explain the difference between these models and the one above. Motivate the fitting of these models in light of
Denote by $H_1$, $H_2$, $H_3$ the three hypotheses being fitted in sequence above.

Explain the hypothesis tests, including an approximate test of the fit of $H_1$, that can be performed using the output from the following R code. Use these numbers to comment on the most appropriate model for the data.

```r
> c(fit1$dev, fit2$dev, fit3$dev)
[1] 84.59557 86.37646 118.99503
> qchisq(0.95, df = 1)
[1] 3.841459
```

2/I/5I Statistical Modelling

Consider the linear regression setting where the responses $Y_i$, $i = 1, \ldots, n$ are assumed independent with means $\mu_i = x_i^T \beta$. Here $x_i$ is a vector of known explanatory variables and $\beta$ is a vector of unknown regression coefficients.

Show that if the response distribution is Laplace, i.e.,

$$ Y_i \sim f(y_i; \mu_i, \sigma) = (2\sigma)^{-1} \exp \left\{ -\frac{|y_i - \mu_i|}{\sigma} \right\}, \quad i = 1, \ldots, n; \quad y_i, \mu_i \in \mathbb{R}; \quad \sigma \in (0, \infty); $$

then the maximum likelihood estimate $\hat{\beta}$ of $\beta$ is obtained by minimising

$$ S_1(\beta) = \sum_{i=1}^{n} |Y_i - x_i^T \beta|. $$

Obtain the maximum likelihood estimate for $\sigma$ in terms of $S_1(\hat{\beta})$.

Briefly comment on why the Laplace distribution cannot be written in exponential dispersion family form.
Consider two possible experiments giving rise to observed data $y_{ij}$ where $i = 1, \ldots, I, j = 1, \ldots, J$.

1. The data are realizations of independent Poisson random variables, i.e.,

$$ Y_{ij} \sim f_1(y_{ij}; \mu_{ij}) = \frac{\mu_{ij}^{y_{ij}}}{y_{ij}!} \exp\{-\mu_{ij}\} $$

where $\mu_{ij} = \mu_{ij}(\beta)$, with $\beta$ an unknown (possibly vector) parameter. Write $\hat{\beta}$ for the maximum likelihood estimator (m.l.e.) of $\beta$ and $\hat{y}_{ij} = \mu_{ij}(\hat{\beta})$ for the $(i, j)$th fitted value under this model.

2. The data are components of a realization of a multinomial random ‘vector’

$$ Y \sim f_2((y_{ij}); n, (p_{ij})) = n! \prod_{i=1}^{I} \prod_{j=1}^{J} \frac{p_{ij}^{y_{ij}}}{y_{ij}!} $$

where the $y_{ij}$ are non-negative integers with

$$ \sum_{i=1}^{I} \sum_{j=1}^{J} y_{ij} = n \quad \text{and} \quad p_{ij}(\beta) = \frac{\mu_{ij}(\beta)}{n}. $$

Write $\beta^*$ for the m.l.e. of $\beta$ and $y^*_{ij} = np_{ij}(\beta^*)$ for the $(i, j)$th fitted value under this model.

Show that, if

$$ \sum_{i=1}^{I} \sum_{j=1}^{J} \hat{y}_{ij} = n, $$

then $\hat{\beta} = \beta^*$ and $\hat{y}_{ij} = y^*_{ij}$ for all $i, j$. Explain the relevance of this result in the context of fitting multinomial models within a generalized linear model framework.
Consider the normal linear model \(Y = X\beta + \varepsilon\) in vector notation, where \(Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}\), \(X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix}\), \(\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}\), \(\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}\), \(\varepsilon_i \sim \text{i.i.d. } N(0, \sigma^2)\), where \(x_i^T = (x_{i1}, \ldots, x_{ip})\) is known and \(X\) is of full rank \((p < n)\). Give expressions for maximum likelihood estimators \(\hat{\beta}\) and \(\hat{\sigma}^2\) of \(\beta\) and \(\sigma^2\) respectively, and state their joint distribution.

Suppose that there is a new pair \((x^*, y^*)\), independent of \((x_1, y_1), \ldots, (x_n, y_n)\), satisfying the relationship \(y^* = x^T\hat{\beta} + \varepsilon^*\), where \(\varepsilon^* \sim N(0, \sigma^2)\).

We suppose that \(x^*\) is known, and estimate \(y^*\) by \(\tilde{y} = x^T\hat{\beta}\). State the distribution of \(\frac{\tilde{y} - y^*}{\hat{\sigma}}\), where \(\hat{\sigma}^2 = \frac{n}{n-p} \hat{\sigma}^2\) and \(\tau^2 = x^T(X^TX)^{-1}x^* + 1\).

Find the form of a \((1 - \alpha)\)-level prediction interval for \(y^*\).

Let \(Y\) have a Gamma distribution with density
\[
f(y; \alpha, \lambda) = \frac{\lambda^\alpha y^{\alpha-1}}{\Gamma(\alpha)} e^{-\lambda y}.
\]
Show that the Gamma distribution is of exponential dispersion family form. Deduce directly the corresponding expressions for \(E[Y]\) and \(\text{Var}[Y]\) in terms of \(\alpha\) and \(\lambda\). What is the canonical link function?

Let \(p < n\). Consider a generalised linear model (g.l.m.) for responses \(y_i, i = 1, \ldots, n\) with random component defined by the Gamma distribution with canonical link \(g(\mu)\), so that \(g(\mu_i) = \eta_i = x_i^T\beta\), where \(\beta = (\beta_1, \ldots, \beta_p)^T\) is the vector of unknown regression coefficients and \(x_i = (x_{i1}, \ldots, x_{ip})^T\) is the vector of known values of the explanatory variables for the \(i\)th observation, \(i = 1, \ldots, n\).

Obtain expressions for the score function and Fisher information matrix and explain how these can be used in order to approximate \(\hat{\beta}\), the maximum likelihood estimator (m.l.e.) of \(\beta\).

[Use the canonical link function and assume that the dispersion parameter is known.]

Finally, obtain an expression for the deviance for a comparison of the full (saturated) model to the g.l.m. with canonical link using the m.l.e. \(\hat{\beta}\) (or estimated mean \(\hat{\mu} = X\hat{\beta}\)).
Statistical Modelling

Assume that observations \( Y = (Y_1, \ldots, Y_n)^T \) satisfy the linear model

\[
Y = X\beta + \epsilon,
\]

where \( X \) is an \( n \times p \) matrix of known constants of full rank \( p < n \), where \( \beta = (\beta_1, \ldots, \beta_p)^T \) is unknown and \( \epsilon \sim N_n(0, \sigma^2 I) \). Write down a \((1 - \alpha)\)-level confidence set for \( \beta \).

Define Cook’s distance for the observation \((x_i, Y_i)\), where \( x_i^T \) is the \( i \)th row of \( X \). Give its interpretation in terms of confidence sets for \( \beta \).

In the above model with \( n = 50 \) and \( p = 2 \), you observe that one observation has Cook’s distance 1.3. Would you be concerned about the influence of this observation?

[You may find some of the following facts useful:

(i) If \( Z \sim \chi^2_2 \), then \( P(Z \leq 0.21) = 0.1, P(Z \leq 1.39) = 0.5 \) and \( P(Z \leq 4.61) = 0.9 \).

(ii) If \( Z \sim \chi^2_{2,48} \), then \( P(Z \leq 0.11) = 0.1, P(Z \leq 0.70) = 0.5 \) and \( P(Z \leq 2.42) = 0.9 \).

(iii) If \( Z \sim \chi^2_{48,2} \), then \( P(Z \leq 0.41) = 0.1, P(Z \leq 1.42) = 0.5 \) and \( P(Z \leq 9.47) = 0.9 \).]
The table below gives a year-by-year summary of the career batting record of the baseball player Babe Ruth. The first column gives his age at the start of each season and the second gives the number of ‘At Bats’ (AB) he had during the season. For each At Bat, it is recorded whether or not he scored a ‘Hit’. The third column gives the total number of Hits he scored in the season, and the final column gives his ‘Average’ for the season, defined as the number of Hits divided by the number of At Bats.

<table>
<thead>
<tr>
<th>Age</th>
<th>AB</th>
<th>Hits</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>10</td>
<td>2</td>
<td>0.200</td>
</tr>
<tr>
<td>20</td>
<td>92</td>
<td>29</td>
<td>0.315</td>
</tr>
<tr>
<td>21</td>
<td>136</td>
<td>37</td>
<td>0.272</td>
</tr>
<tr>
<td>22</td>
<td>123</td>
<td>40</td>
<td>0.325</td>
</tr>
<tr>
<td>23</td>
<td>317</td>
<td>95</td>
<td>0.300</td>
</tr>
<tr>
<td>24</td>
<td>432</td>
<td>139</td>
<td>0.322</td>
</tr>
<tr>
<td>25</td>
<td>457</td>
<td>172</td>
<td>0.376</td>
</tr>
<tr>
<td>26</td>
<td>540</td>
<td>204</td>
<td>0.378</td>
</tr>
<tr>
<td>27</td>
<td>406</td>
<td>128</td>
<td>0.315</td>
</tr>
<tr>
<td>28</td>
<td>522</td>
<td>205</td>
<td>0.393</td>
</tr>
<tr>
<td>29</td>
<td>529</td>
<td>200</td>
<td>0.378</td>
</tr>
<tr>
<td>30</td>
<td>359</td>
<td>134</td>
<td>0.373</td>
</tr>
<tr>
<td>31</td>
<td>495</td>
<td>184</td>
<td>0.372</td>
</tr>
<tr>
<td>32</td>
<td>540</td>
<td>192</td>
<td>0.356</td>
</tr>
<tr>
<td>33</td>
<td>536</td>
<td>173</td>
<td>0.323</td>
</tr>
<tr>
<td>34</td>
<td>499</td>
<td>172</td>
<td>0.345</td>
</tr>
<tr>
<td>35</td>
<td>518</td>
<td>186</td>
<td>0.359</td>
</tr>
<tr>
<td>36</td>
<td>534</td>
<td>199</td>
<td>0.373</td>
</tr>
<tr>
<td>37</td>
<td>457</td>
<td>156</td>
<td>0.341</td>
</tr>
<tr>
<td>38</td>
<td>459</td>
<td>138</td>
<td>0.301</td>
</tr>
<tr>
<td>39</td>
<td>365</td>
<td>105</td>
<td>0.288</td>
</tr>
<tr>
<td>40</td>
<td>72</td>
<td>13</td>
<td>0.181</td>
</tr>
</tbody>
</table>
Explain and interpret the R commands below. In particular, you should explain the model that is being fitted, the approximation leading to the given standard errors and the test that is being performed in the last line of output.

```r
> Mod <- glm(Hits/AB~Age+I(Age^2),family=binomial,weights=AB)
> summary(Mod)
```

Coefficients:

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|----------|
| (Intercept) -4.5406713 | 0.8487687 | -5.350 | 8.81e-08 *** |
| Age 0.2684739 | 0.0565992 | 4.743 | 2.10e-06 *** |
| I(Age^2) -0.0044827 | 0.0009253 | -4.845 | 1.27e-06 *** |

Residual deviance: 23.345 on 19 degrees of freedom

Assuming that any required packages are loaded, draw a careful sketch of the graph that you would expect to see on entering the following lines of code:

```r
> Coef <- coef(Mod)
> Fitted <- inv.logit(Coef[1]+Coef[2]*Age+Coef[3]*Age^2)
> plot(Age,Average)
> lines(Age,Fitted)
```

2/I/5I Statistical Modelling

Let $Y_1, \ldots, Y_n$ be independent Poisson random variables with means $\mu_1, \ldots, \mu_n$, for $i = 1, \ldots, n$, where $\log(\mu_i) = \beta x_i$, for some known constants $x_i$ and an unknown parameter $\beta$. Find the log-likelihood for $\beta$.

By first computing the first and second derivatives of the log-likelihood for $\beta$, explain the algorithm you would use to find the maximum likelihood estimator, $\hat{\beta}$.
Statistical Modelling

Consider a generalized linear model for independent observations $Y_1, \ldots, Y_n$, with $\mathbb{E}(Y_i) = \mu_i$ for $i = 1, \ldots, n$. What is a linear predictor? What is meant by the link function? If $Y_i$ has model function (or density) of the form

$$f(y_i; \mu_i, \sigma^2) = \exp \left[ \frac{1}{\sigma^2} \left\{ \theta(\mu_i) y_i - K(\theta(\mu_i)) \right\} \right] a(\sigma^2, y_i),$$

for $y_i \in \mathcal{Y} \subseteq \mathbb{R}$, $\mu_i \in \mathcal{M} \subseteq \mathbb{R}$, $\sigma^2 \in \Phi \subseteq (0, \infty)$, where $a(\sigma^2, y_i)$ is a known positive function, define the canonical link function.

Now suppose that $Y_1, \ldots, Y_n$ are independent with $Y_i \sim \text{Bin}(1, \mu_i)$ for $i = 1, \ldots, n$. Derive the canonical link function.
4/I/5I Statistical Modelling

The table below summarises the yearly numbers of named storms in the Atlantic basin over the period 1944–2004, and also gives an index of average July ocean temperature in the northern hemisphere over the same period. To save space, only the data for the first four and last four years are shown.

<table>
<thead>
<tr>
<th>Year</th>
<th>Storms</th>
<th>Temp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1944</td>
<td>11</td>
<td>0.165</td>
</tr>
<tr>
<td>1945</td>
<td>11</td>
<td>0.080</td>
</tr>
<tr>
<td>1946</td>
<td>6</td>
<td>0.000</td>
</tr>
<tr>
<td>1947</td>
<td>9</td>
<td>-0.024</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2001</td>
<td>15</td>
<td>0.592</td>
</tr>
<tr>
<td>2002</td>
<td>12</td>
<td>0.627</td>
</tr>
<tr>
<td>2003</td>
<td>16</td>
<td>0.608</td>
</tr>
<tr>
<td>2004</td>
<td>15</td>
<td>0.546</td>
</tr>
</tbody>
</table>

Explain and interpret the R commands and (slightly abbreviated) output below.

```
> Mod <- glm(Storms~Temp,family=poisson)
> summary(Mod)
```

Coefficients:

|                | Estimate | Std. Error | z value | Pr(>|z|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 2.26061  | 0.04841    | 46.697  | < 2e-16  *** |
| Temp           | 0.48870  | 0.16973    | 2.879   | 0.00399  ** |

Residual deviance: 51.499 on 59 degrees of freedom

In 2005, the ocean temperature index was 0.743. Explain how you would predict the number of named storms for that year.
4/II/13I Statistical Modelling

Consider a linear model for \( Y = (Y_1, \ldots, Y_n)^T \) given by

\[
Y = X\beta + \epsilon,
\]

where \( X \) is a known \( n \times p \) matrix of full rank \( p < n \), where \( \beta \) is an unknown vector and \( \epsilon \sim N_n(0, \sigma^2 I) \). Derive an expression for the maximum likelihood estimator \( \hat{\beta} \) of \( \beta \), and write down its distribution.

Find also the maximum likelihood estimator \( \hat{\sigma}^2 \) of \( \sigma^2 \), and derive its distribution.

[You may use Cochran’s theorem, provided that it is stated carefully. You may also assume that the matrix \( P = X(X^TX)^{-1}X^T \) has rank \( p \), and that \( I - P \) has rank \( n - p \).]
Statistical Modelling

Suppose that $Y_1, \ldots, Y_n$ are independent random variables, and that $Y_i$ has probability density function

$$f(y_i | \theta_i, \phi) = \exp \left( \frac{(y_i \theta_i - b(\theta_i))}{\phi} + c(y_i, \phi) \right).$$

Assume that $E(Y_i) = \mu_i$ and that there is a known link function $g(.)$ such that

$$g(\mu_i) = \beta^T x_i,$$

where $x_1, \ldots, x_n$ are known $p$-dimensional vectors and $\beta$ is an unknown $p$-dimensional parameter. Show that $E(Y_i) = b'(\theta_i)$ and that, if $\ell(\beta, \phi)$ is the log-likelihood function from the observations $(y_1, \ldots, y_n)$, then

$$\frac{\partial \ell(\beta, \phi)}{\partial \beta} = \sum_{i=1}^{n} \frac{(y_i - \mu_i) x_i}{g'(\mu_i) V_i},$$

where $V_i$ is to be defined.

Statistical Modelling

The Independent, June 1999, under the headline ‘Tourists get hidden costs warnings’ gave the following table of prices in pounds, called ‘How the resorts compared’.

<table>
<thead>
<tr>
<th>Resort</th>
<th>1/2005</th>
<th>1/3</th>
<th>1/4</th>
<th>1/5</th>
<th>1/6</th>
<th>1/7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algarve</td>
<td>8.00</td>
<td>0.50</td>
<td>3.50</td>
<td>3.00</td>
<td>4.00</td>
<td>100.00</td>
</tr>
<tr>
<td>CostaDelSol</td>
<td>6.95</td>
<td>1.30</td>
<td>4.10</td>
<td>12.30</td>
<td>4.10</td>
<td>130.85</td>
</tr>
<tr>
<td>Majorca</td>
<td>10.25</td>
<td>1.45</td>
<td>5.35</td>
<td>6.15</td>
<td>3.30</td>
<td>122.20</td>
</tr>
<tr>
<td>Tenerife</td>
<td>12.30</td>
<td>1.25</td>
<td>4.90</td>
<td>3.70</td>
<td>2.90</td>
<td>130.85</td>
</tr>
<tr>
<td>Florida</td>
<td>15.60</td>
<td>1.90</td>
<td>5.05</td>
<td>5.00</td>
<td>2.50</td>
<td>114.00</td>
</tr>
<tr>
<td>Tunisia</td>
<td>10.90</td>
<td>1.40</td>
<td>5.45</td>
<td>1.90</td>
<td>2.75</td>
<td>218.10</td>
</tr>
<tr>
<td>Cyprus</td>
<td>11.60</td>
<td>1.20</td>
<td>5.95</td>
<td>3.00</td>
<td>3.60</td>
<td>149.45</td>
</tr>
<tr>
<td>Turkey</td>
<td>6.50</td>
<td>1.05</td>
<td>6.50</td>
<td>4.90</td>
<td>2.85</td>
<td>263.00</td>
</tr>
<tr>
<td>Corfu</td>
<td>5.20</td>
<td>1.05</td>
<td>3.75</td>
<td>4.20</td>
<td>2.50</td>
<td>137.60</td>
</tr>
<tr>
<td>Sorrento</td>
<td>7.70</td>
<td>1.40</td>
<td>6.30</td>
<td>8.75</td>
<td>4.75</td>
<td>215.40</td>
</tr>
<tr>
<td>Malta</td>
<td>11.20</td>
<td>0.70</td>
<td>4.55</td>
<td>8.00</td>
<td>4.80</td>
<td>87.85</td>
</tr>
<tr>
<td>Rhodes</td>
<td>6.30</td>
<td>1.05</td>
<td>5.20</td>
<td>3.15</td>
<td>2.70</td>
<td>261.30</td>
</tr>
<tr>
<td>Sicily</td>
<td>13.25</td>
<td>1.75</td>
<td>4.20</td>
<td>7.00</td>
<td>3.85</td>
<td>174.40</td>
</tr>
<tr>
<td>Madeira</td>
<td>10.25</td>
<td>0.70</td>
<td>5.10</td>
<td>6.85</td>
<td>6.85</td>
<td>153.70</td>
</tr>
</tbody>
</table>
Here the column headings are, respectively: Three-course meal, Bottle of Beer, Suntan Lotion, Taxi (5km), Film (24 exp), Car Hire (per week). Interpret the R commands, and explain how to interpret the corresponding (slightly abbreviated) R output given below. Your solution should include a careful statement of the underlying statistical model, but you may quote without proof any distributional results required.

```r
> price = scan("dresorts") ; price
> Goods = gl(6,1,length=84); Resort=gl(14,6,length=84)
> first.lm = lm(log(price) ~ Goods + Resort)
> summary(first.lm)

Coefficients:

Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.8778 0.1629 11.527 < 2e-16
Goods2 -2.1084 0.1295 -16.286 < 2e-16
Goods3 -0.6343 0.1295 -4.900 6.69e-06
Goods4 -0.6284 0.1295 -4.854 7.92e-06
Goods5 -0.9679 0.1295 -7.476 2.49e-10
Goods6 2.8016 0.1295 21.640 < 2e-16
Resort2 0.4463 0.1978 2.257 0.02740
Resort3 0.4105 0.1978 2.076 0.04189
Resort4 0.3067 0.1978 1.551 0.12584
Resort5 0.4235 0.1978 2.142 0.03597
Resort6 0.2883 0.1978 1.458 0.14963
Resort7 0.3457 0.1978 1.748 0.08519
Resort8 0.3787 0.1978 1.915 0.05993
Resort9 0.0943 0.1978 0.477 0.63508
Resort10 0.5981 0.1978 3.025 0.00356
Resort11 0.3281 0.1978 1.659 0.10187
Resort12 0.2525 0.1978 1.277 0.20616
Resort13 0.5508 0.1978 2.785 0.00700
Resort14 0.4590 0.1978 2.321 0.02343

Residual standard error: 0.3425 on 65 degrees of freedom
Multiple R-Squared: 0.962
```
2/I/5I Statistical Modelling

You see below three R commands, and the corresponding output (which is slightly abbreviated). Explain the effects of the commands. How is the deviance defined, and why do we have d.f.=7 in this case? Interpret the numerical values found in the output.

```r
> n = scan()
3 5 16 12 11 34 37 51 56
> i = scan()
1 2 3 4 5 6 7 8 9
> summary(glm(n~i,poisson))
deviance = 13.218
d.f. = 7
Coefficients:
     Value Std.Error
(intercept) 1.363 0.2210
     i     0.3106 0.0382
```

3/I/5I Statistical Modelling

Consider the model \(Y = X\beta + \epsilon\), where \(Y\) is an \(n\)-dimensional observation vector, \(X\) is an \(n \times p\) matrix of rank \(p\), \(\epsilon\) is an \(n\)-dimensional vector with components \(\epsilon_1, \ldots, \epsilon_n\), and \(\epsilon_1, \ldots, \epsilon_n\) are independently and normally distributed, each with mean 0 and variance \(\sigma^2\).

(a) Let \(\hat{\beta}\) be the least-squares estimator of \(\beta\). Show that
\[(X^T X) \hat{\beta} = X^T Y\]
and find the distribution of \(\hat{\beta}\).

(b) Define \(\hat{Y} = X \hat{\beta}\). Show that \(\hat{Y}\) has distribution \(N(X\beta, \sigma^2 H)\), where \(H\) is a matrix that you should define.

[You may quote without proof any results you require about the multivariate normal distribution.]
4/I/5I Statistical Modelling

You see below five R commands, and the corresponding output (which is slightly abbreviated). Without giving any mathematical proofs, explain the purpose of these commands, and interpret the output.

```r
> Yes = c(12, 27,11,24)
> Total = c(117,170,52,118)
> Sclass = c("a","a","b","b")
> Sclass = factor(Sclass)
> summary(glm(Yes/Total~ Sclass, binomial, weights=Total))

Coefficients:

    Estimate Std. Error z value
(Intercept)  -1.8499   0.1723  -10.739
Sclassb       0.4999   0.2562   1.951

Residual deviance: 1.9369 on 2 degrees of freedom
Number of Fisher Scoring iterations: 4
```
(i) Suppose that $Y_1, \ldots, Y_n$ are independent random variables, and that $Y_i$ has probability density function

$$f(y_i|\beta, \nu) = \left(\frac{\nu y_i}{\mu_i}\right)^\nu e^{-\frac{y_i\nu}{\mu_i}} \frac{1}{\Gamma(\nu)} \frac{1}{y_i} \text{ for } y_i > 0$$

where

$$1/\mu_i = \beta^T x_i, \text{ for } 1 \leq i \leq n,$$

and $x_1, \ldots, x_n$ are given $p$-dimensional vectors, and $\nu$ is known.

Show that $\mathbb{E}(Y_i) = \mu_i$ and that $\text{var}(Y_i) = \mu_i^2/\nu$.

(ii) Find the equation for $\hat{\beta}$, the maximum likelihood estimator of $\beta$, and suggest an iterative scheme for its solution.

(iii) If $p = 2$, and $x_i = \begin{pmatrix} 1 \\ z_i \end{pmatrix}$, find the large-sample distribution of $\hat{\beta}_2$. Write your answer in terms of $a, b, c$ and $\nu$, where $a, b, c$ are defined by

$$a = \sum \mu_i^2, \quad b = \sum z_i \mu_i^2, \quad c = \sum z_i^2 \mu_i^2.$$