Part II

Statistical Modelling

Year

Paper 1, Section I

5J Statistical Modelling

Consider a possibly biased coin. Suppose the probability of flipping a head is $0 < p < 1$ and p is unknown. Let $r > 0$ be given. In a sequence of flips, let X be the total number of tails when r heads are reached. Show that

$$\mathbb{P}(X = x) = \binom{x+r-1}{x}(1-p)^x p^r, \; x = 0, 1, \ldots .$$

Show that this is a one-parameter exponential family. Find its natural parameter, sufficient statistic, and cumulant function, and compute the mean and variance of X in terms of p.

Paper 2, Section I

5J Statistical Modelling

Explain the following R commands in words, then write down the model that is being fitted.

```
> n <- 100
> p <- 2
> X <- matrix(rnorm(n * p), nrow = n, ncol = p)
> Y <- rbinom(n, size = 1, prob = 0.5)
> sum(Y)
[1] 48
> fit1 <- glm(Y ~ X, family = binomial)
> sum(predict(fit1, type = "response"))
[1] 48
```

Explain why the output of the last command should be exactly the same as the output of `sum(Y)` by writing down the likelihood function of the model.

Do you expect the following command to output exactly 48, too? If not, do you expect it to be very different from 48? Justify your answers.

```
> fit2 <- glm(Y ~ X, family = binomial(probit))
> sum(predict(fit2, type = "response"))
```

Paper 3, Section I

5J Statistical Modelling

Write down the density function of a one-parameter exponential family with natural parameter θ and sufficient statistic Y. Define the *deviance* $D(\theta_1, \theta_2)$ from θ_1 to θ_2, and show that it is equal to

$$D(\theta_1, \theta_2) = 2\{(\theta_1 - \theta_2)\mu_1 - K(\theta_1) + K(\theta_2)\},$$

where μ_1 is the mean parameter corresponding to θ_1 and $K(\cdot)$ is the cumulant function of the exponential family.

Derive the deviance from the Poisson distribution with mean μ_1 to the Poisson distribution with mean μ_2, and find the second order Taylor approximation of the deviance as $\mu_2 \to \mu_1$. [*Hint: Recall that if Y follows a Poisson distribution with mean μ, then* $\mathbb{P}(Y = k) = \mu^k e^{-\mu}/k!$, $k = 0, 1, \ldots$.]

Paper 4, Section I

5J Statistical Modelling

Below is a simplified 1993 dataset of US cars. The columns list 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):

```
> 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  29.1  26     5    180    67   3375
          ...               ...                       ...
91 Volkswagen Corrado  23.3  25     4    159    66   2810
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 are independent. How would you instruct R to model the logarithm of the price as a linear combination of an error term and

(i) an intercept;

(ii) an intercept and all other quantitative properties of the cars;

(iii) an intercept, all other quantitative properties of the cars, and the make of the cars?

Suppose the fitted models are assigned to objects `fit1`, `fit2`, and `fit3`, respectively. Suppose R provides the following analysis of variance table for these models:

```
> anova(fit1, fit2, fit3)

[...]

  Res.Df    RSS Df Sum of Sq       F    Pr(>F)
1     92 8584.0
2     87 3349.1  5    5234.9 69.7334 < 2.2e-16 ***
3     56  840.8 31    2508.3  5.3891 2.541e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

What are your conclusions about the statistical models in `fit1`, `fit2` and `fit3` based on this table? Explain how you can determine the number of unique car manufacturers in this dataset from this table.

Paper 1, Section II

13J Statistical Modelling

Let X be a fixed $n \times p$ design matrix with full column rank. Let H be the projection matrix onto the column space of X. Suppose the n-vector of response Y satisfies $Y \sim \mathrm{N}(\mu, \sigma^2 I_n)$ where the n-vector μ is fixed but unknown. Let $Y^* \sim \mathrm{N}(\mu, \sigma^2 I_n)$ be another random vector that has the same distribution as Y but is independent of Y.

(i) Show that
$$\mathbb{E}(\|HY - Y^*\|^2) = \|(I - H)\mu\|^2 + (n + p)\sigma^2.$$
Explain why the above identity is an example of the bias-variance tradeoff. [You may use without proof the fact that H is a projection matrix with rank p.]

(ii) Suppose σ^2 is known. Show that Mallows' C_p, given by
$$C_p = \|Y - HY\|^2 + 2p\sigma^2,$$
is an unbiased estimator of $\mathbb{E}(\|HY - Y^*\|^2)$.

For the rest of this question, suppose $\mu = X\beta$ for some unknown p-vector β and σ^2 is unknown.

(iii) Write down the $(1 - \alpha)$-level confidence ellipsoid for β.

(iv) Recall *Cook's distance* for the observation (X_i, Y_i) (where X_i^T is the ith row of X) is a measure of the influence of (X_i, Y_i) on the fitted values. Give the precise definition of Cook's distance and give its interpretation in terms of the confidence ellipsoid for β.

(v) In the model above with $n = 100$ and $p = 4$, you notice 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* $\mathbb{P}(Z \leqslant 1.06) = 0.1$, $\mathbb{P}(Z \leqslant 7.78) = 0.9$.
2. *If* $Z \sim F_{4,96}$, *then* $\mathbb{P}(Z \leqslant 0.26) = 0.1$, $\mathbb{P}(Z \leqslant 2.00) = 0.9$.
3. *If* $Z \sim F_{96,4}$, *then* $\mathbb{P}(Z \leqslant 0.50) = 0.1$, $\mathbb{P}(Z \leqslant 3.78) = 0.9$.]

Paper 4, Section II

13J Statistical Modelling

The data frame `worldcup22` contains information about the matches played in a sports competition, including for each team in the match the starting formations (indicated by letters A-L), the expected goals (xg) and the actual goals. In the questions below we will assume that the match results are independent.

```
> worldcup22
     team1   team2 team1_xg team2_xg team1_form team2_form team1_goal team2_goal
1    Qatar Ecuador      0.3      1.2          I          H          0          2
2  England IR Iran      2.1      1.4          E          J          6          2
        ...                      ...                                          ...
63 Croatia Morocco      0.7      1.2          E          F          2          1
64   Japan  France      3.3      2.2          F          E          3          3
> fit1 <- glm(team1_goal ~ team1_form + team2_form, worldcup22,
              family = poisson)
```

(i) Let Y denote the response vector and X denote the design matrix for `fit1`. Write down the likelihood function that is maximized by the command above. [Recall that if Y follows a Poisson distribution with mean μ, then $\mathbb{P}(Y = k) = \mu^k e^{-\mu}/k!$, $k = 0, 1, \ldots$.]

(ii) Comment on the following abbreviated summary of `fit1`. Is there enough information to conclude that the formation of `team1` does not affect its goals? If not, what is the name of the statistical procedure you can use to test this hypothesis?

```
> summary(fit1)
               Estimate Std. Error z value Pr(>|z|)
(Intercept)       1.890      0.581     3.3    0.001 **
team1_formB      -0.672      0.595    -1.1    0.259
team1_formC     -17.865   2446.075     0.0    0.994
team1_formD       0.595      1.293     0.5    0.645
team1_formE      -0.361      0.441    -0.8    0.413
team1_formF      -0.098      0.414    -0.2    0.812
team1_formG      -1.120      1.089    -1.0    0.304
team1_formH      -0.332      0.490    -0.7    0.498
team1_formI      -1.855      1.104    -1.7    0.093 .
team1_formJ       0.285      0.830     0.3    0.731
team2_formK     -18.831   3467.859     0.0    0.996
team2_formB      -1.199      0.565    -2.1    0.034 *
team2_formC      -1.792      1.080    -1.7    0.097 .
team2_formL      -0.905      0.558    -1.6    0.105
team2_formE      -1.482      0.478    -3.1    0.002 **
team2_formF      -1.464      0.504    -2.9    0.004 **
team2_formH      -0.728      0.494    -1.5    0.140
team2_formI      -0.980      0.588    -1.7    0.095 .
team2_formJ      -0.143      0.612    -0.2    0.816
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

[QUESTION CONTINUES ON THE NEXT PAGE]

(iii) Expected goals (xg) is a new metric in sports analytics that computes the number of goals a team should have scored based on the quality of the chances created. State the following two hypotheses mathematically: (a) H_1: `team1_goal` has mean `team1_xg`; (b) H_2: `team1_goal` follows a Poisson distribution with mean `team1_xg`. Then name the result in probability theory that suggests `team1_goal` should approximately follow a Poisson distribution.

(iv) An analyst fitted the following model to test H_1. Does the model fit suggest evidence against H_1? Give one reason why we should be skeptical about the standard errors in the table.

```
> fit2 <- lm(team1_goal ~ team1_xg - 1, worldcup22)
> summary(fit2)
Coefficients:
        Estimate Std. Error t value Pr(>|t|)
team1_xg  1.15790    0.08643    13.4   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(v) The analyst then fitted the following model and computed the 95% confidence interval for the coefficients. Explain why the observation that the confidence interval for `log(team1_xg)` contains 1 does not directly imply that H_2 cannot be rejected at the 5% significance level.

```
> fit3 <- glm(team1_goal ~ log(team1_xg), worldcup22, family = poisson)
> confint(fit3)
                   2.5 %    97.5 %
(Intercept)  -0.1387542 0.3836497
log(team1_xg)  0.6691166 1.3395731
```

Paper 1, Section I

5J Statistical Modelling

Let Y_μ be the Poisson distribution with mean μ. Show that the transformation $g(y) = 2\sqrt{y}$ is "variance stabilising" for Y_μ in the sense that the variance of $g(Y_\mu)$ is approximately 1 when μ is large.

Suppose we fit a linear model to the transformed response $\sqrt{Y}$. How does this differ from using the square root link in the Poisson regression?

Paper 2, Section I

5J Statistical Modelling

(a) Give the definition of an *exponential family* of probability distributions. [You may assume the natural parameter is one-dimensional.]

(b) Suppose $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} f(y;\theta)$ where $f(y;\theta)$ is the density function of an exponential family with natural parameter θ and sufficient statistic Y. Show that $\bar{Y} = \sum_{i=1}^n Y_i/n$ is a sufficient statistic for θ.

(c) In the setting above, show that the maximum likelihood estimator of θ is given by setting the theoretical mean $\mu(\theta) = \mathbb{E}_\theta(Y_1)$ to the empirical mean $\bar{Y}$.

Paper 3, Section I

5J Statistical Modelling

The density function of the Laplace distribution $\text{Laplace}(\mu, \sigma)$ with mean μ and scale parameter σ is given by

$$f(y;\mu,\sigma) = (2\sigma)^{-1} \exp\left\{-\frac{|y-\mu|}{\sigma}\right\}.$$

Briefly comment on why the Laplace distribution cannot be written in exponential dispersion family form.

Consider the linear model where $(X_i, Y_i), i = 1, \ldots, n$ are assumed independent and

$$Y_i \mid X_i \sim \text{Laplace}(X_i^T\beta, \sigma)\,.$$

Show that the maximum likelihood estimator $\hat{\beta}$ of β is obtained by minimising

$$S(\beta) = \sum_{i=1}^n |Y_i - X_i^T\beta|\,.$$

Obtain the maximum likelihood estimator of σ in terms of $S(\hat{\beta})$.

Paper 4, Section I

5J Statistical Modelling

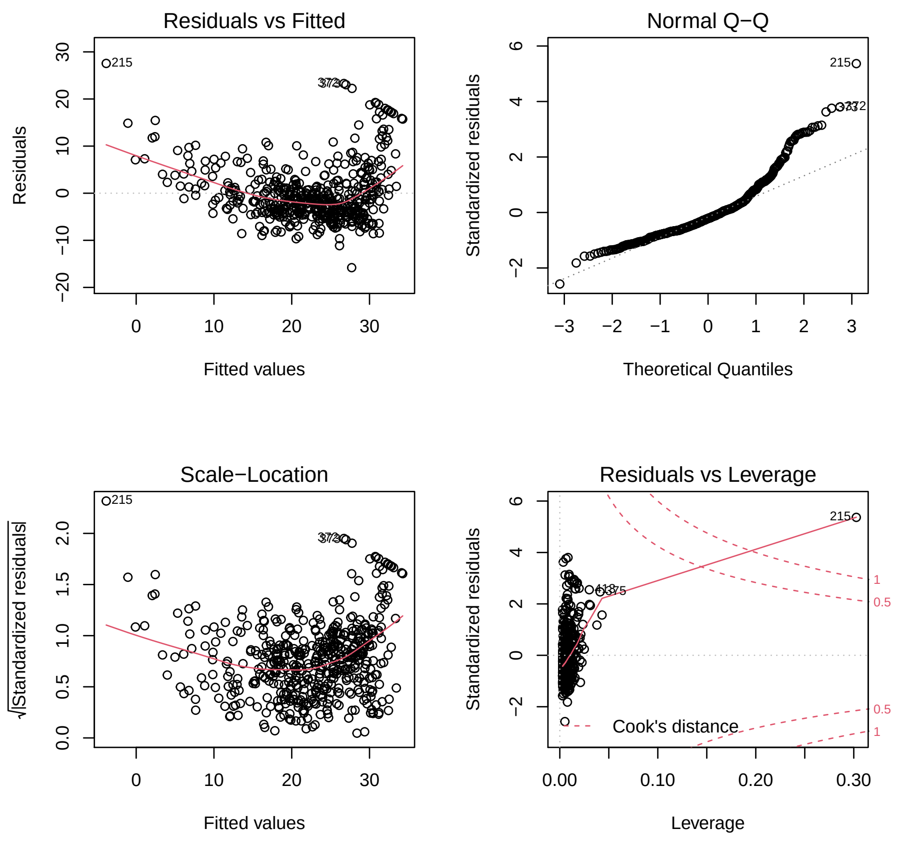
The `Boston` dataset records `medv` (median house value), `age` (average age of houses), `lstat` (percent of households with low socioeconomic status), and other covariates for 506 census tracts in Boston.

```
> head(Boston[, c("medv", "age", "lstat")])
  medv  age lstat
1 24.0 65.2  4.98
2 21.6 78.9  9.14
3 34.7 61.1  4.03
4 33.4 45.8  2.94
5 36.2 54.2  5.33
6 28.7 58.7  5.21
```

Describe the mathematical model fitted in the `R` code below and give three observations from the output of the code that you think are the most noteworthy.

```
> summary(fit <- lm(medv ~ lstat * age , data = Boston))

Residuals:
    Min      1Q  Median      3Q     Max
-15.806  -4.045  -1.333   2.085  27.552

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.0885359  1.4698355  24.553  < 2e-16 ***
lstat       -1.3921168  0.1674555  -8.313 8.78e-16 ***
age         -0.0007209  0.0198792  -0.036   0.9711
lstat:age    0.0041560  0.0018518   2.244   0.0252 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.149 on 502 degrees of freedom
Multiple R-squared:  0.5557,    Adjusted R-squared:  0.5531
F-statistic: 209.3 on 3 and 502 DF,  p-value: < 2.2e-16

>
> par(mfrow = c(2, 2))
> plot(fit)
```

[QUESTION CONTINUES ON THE NEXT PAGE]


Residuals vs Fitted
Residuals
Fitted values
215
372
−20
−10
0
10
20
30
Normal Q−Q
Standardized residuals
Theoretical Quantiles
215
372
−3
−2
−1
0
1
2
3
−2
0
2
4
6
Scale−Location
√|Standardized residuals|
Fitted values
215
372
0.0
0.5
1.0
1.5
2.0
Residuals vs Leverage
Standardized residuals
Leverage
215
418
275
Cook's distance
1
0.5
0.5
1
0.00
0.10
0.20
0.30

Paper 1, Section II

13J Statistical Modelling

The following dataset contains information about some of the passengers on RMS *Titanic* when it sank on 15th April, 1912.

```
> head(titanic)
  Survived Pclass    Sex Age SibSp Parch    Fare Cabin Embarked
1        0      3   male  22     1     0  7.2500  <NA>        S
2        1      1 female  38     1     0 71.2833   C85        C
3        1      3 female  26     0     0  7.9250  <NA>        S
4        1      1 female  35     1     0 53.1000  C123        S
5        0      3   male  35     0     0  8.0500  <NA>        S
6        0      3   male  NA     0     0  8.4583  <NA>        Q
> nrow(titanic)
[1] 889
```

We would like to predict which passengers were more likely to survive (`Survived`, 0 = No, 1 = Yes) using the other covariates, including ticket class (`Pclass`, 1 = 1st, 2 = 2nd, 3 = 3rd), sex (`Sex`), age (`Age`), number of siblings/spouses aboard (`SibSp`), number of parents/children aboard (`Parch`), passenger fare (`Fare`), cabin number (`Cabin`), port of embarkation (`Embarked`, C = Cherbourg, Q = Queenstown, S = Southampton).

(a) Describe what the following chunk of `R` code does.

```
> apply(titanic, 2, function(x) sum(is.na(x)))
Survived   Pclass      Sex      Age    SibSp    Parch     Fare    Cabin
       0        0        0      177        0        0        0      687
Embarked
       0
> titanic$Cabin <- NULL
> titanic$Age[is.na(titanic$Age)] <- mean(titanic$Age, na.rm = TRUE)
```

(b) Write down the generalised linear model fitted (including the likelihood function maximised) by the code below. Define *Akaike's information criterion* (AIC) and explain, in words, how you can use the backward stepwise algorithm and AIC to select a model.

```
> summary(fit <- glm(Survived ˜ ., family = binomial, data = titanic))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6445  -0.5907  -0.4227   0.6214   2.4432
```

[QUESTION CONTINUES ON THE NEXT PAGE]

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept)  5.284055   0.564696   9.357  < 2e-16 ***
Pclass      -1.100033   0.143530  -7.664 1.80e-14 ***
Sexmale     -2.718736   0.200779 -13.541  < 2e-16 ***
Age         -0.039885   0.007855  -5.078 3.82e-07 ***
SibSp       -0.325732   0.109368  -2.978   0.0029 ** 
Parch       -0.092470   0.118702  -0.779   0.4360    
Fare         0.001919   0.002376   0.808   0.4192    
EmbarkedQ   -0.035043   0.381920  -0.092   0.9269    
EmbarkedS   -0.418564   0.236788  -1.768   0.0771 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1182.82  on 888  degrees of freedom
Residual deviance:  784.21  on 880  degrees of freedom
AIC: 802.21

Number of Fisher Scoring iterations: 5
```

(c) The model summary above says “Dispersion parameter for binomial family taken to be 1”. Do you think that is reasonable based on the model summary? Justify your answer. You might find the following information useful.

```
> qnorm(0.25) # 25th-percentile of the standard normal distribution
[1] -0.6744898
```

(d) Give an estimator of the dispersion parameter in this model when it is not fixed at 1.

Paper 4, Section II
13J Statistical Modelling

Consider the following R code:

```
> n <- 1000000
> sigma_z <- 1; sigma_x1 <- 0.5; sigma_x2 <- 1; sigma_y <- 2; beta <- 2
> Z <- sigma_z * rnorm(n)
> X1 <- Z + sigma_x1 * rnorm(n)
> X2 <- Z + sigma_x2 * rnorm(n)
> Y <- beta * Z + sigma_y * rnorm(n)
> lm(Y ~ Z)

Call:
lm(formula = Y ~ Z)

Coefficients:
(Intercept)            Z
  -0.003089     1.999780

> lm(Y ~ X1)

Call:
lm(formula = Y ~ X1)

Coefficients:
(Intercept)           X1
  -0.002904     1.600521

> lm(Y ~ X2)

Call:
lm(formula = Y ~ X2)

Coefficients:
(Intercept)           X2
  -0.002672     0.997499
```

Describe the phenomenon you see in the output above, then give a mathematical explanation for this phenomenon. Do you expect the slope coefficient in the second model to be generally smaller than that in the first model? Do you think modifying (for example, doubling) the value of `sigma_y` will substantially alter the slope coefficient in the second model? Justify your answer.

Paper 1, Section I

5J Statistical Modelling

Let $\mu > 0$. The probability density function of the inverse Gaussian distribution (with the shape parameter equal to 1) is given by

$$f(x;\mu) = \frac{1}{\sqrt{2\pi x^3}} \exp\left[-\frac{(x-\mu)^2}{2\mu^2 x}\right].$$

Show that this is a one-parameter exponential family. What is its natural parameter? Show that this distribution has mean μ and variance μ^3.

Paper 2, 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* (without the dispersion parameter) with logistic and probit link functions. Which of these is the canonical link function?

Paper 3, Section I

5J Statistical Modelling

Consider the normal linear model $Y \mid X \sim \mathrm{N}(X\beta, \sigma^2 I)$, where X is a $n \times p$ design matrix, Y is a vector of responses, I is the $n \times n$ identity matrix, and β, σ^2 are unknown parameters.

Derive the maximum likelihood estimator of the pair β and σ^2. What is the distribution of the estimator of σ^2? Use it to construct a $(1-\alpha)$-level confidence interval of σ^2. [You may use without proof the fact that the "hat matrix" $H = X(X^TX)^{-1}X^T$ is a projection matrix.]

Paper 4, Section I

5J Statistical Modelling

The data frame `data` contains the daily number of new avian influenza cases in a large poultry farm.

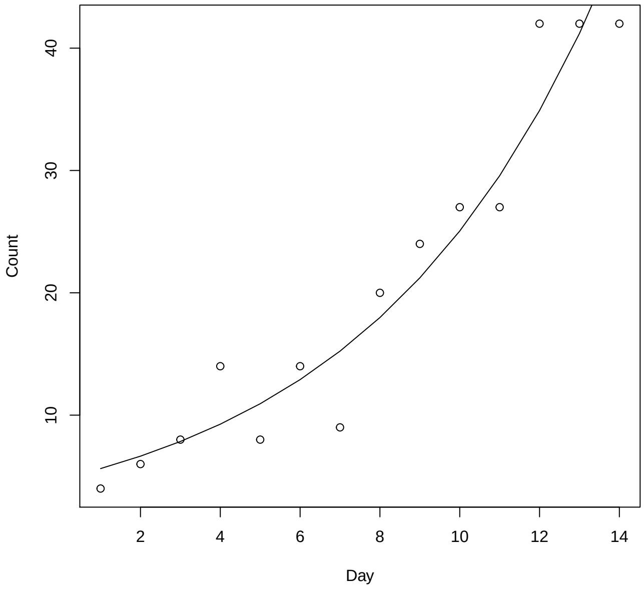
```
> rbind(head(data, 2), tail(data, 2))
   Day Count
1    1     4
2    2     6
13  13    42
14  14    42
```

Write down the model being fitted by the `R` code below. Does the model seem to provide a satisfactory fit to the data? Justify your answer.

The owner of the farm estimated that the size of the epidemic was initially doubling every 7 days. Is that estimate supported by the analysis below? [You may need $\log 2 \approx 0.69$.]

```
> fit <- glm(Count ~ Day, family = poisson, data)

> summary(fit)

Call:
glm(formula = Count ~ Day, family = poisson, data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7298  -0.6639   0.0897   0.4473   1.4466

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.5624     0.1759   8.883   <2e-16 ***
Day           0.1658     0.0166   9.988   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 122.9660  on 13  degrees of freedom
Residual deviance:   9.9014  on 12  degrees of freedom

> pchisq(9.9014, 12, lower.tail = FALSE)
[1] 0.6246105
> plot(Count ~ Day, data)
> lines(data$Day, predict(fit, data, type = "response"))
```

[QUESTION CONTINUES ON THE NEXT PAGE]


Count
10
20
30
40
Day
2
4
6
8
10
12
14

Paper 1, Section II

13J Statistical Modelling

The following data were obtained in a randomised controlled trial for a drug. Due to a manufacturing error, a subset of trial participants received a low dose (LD) instead of a standard dose (SD) of the drug.

```
> data
  treatment  outcome count
1   Control   Better  5728
2   Control    Worse   101
3        LD   Better  1364
4        LD    Worse     3
5        SD   Better  4413
6        SD    Worse    27
```

(a) Below we analyse the data using Poisson regression:

```
> fit1 <- glm(count ~ treatment + outcome, family = poisson, data)
> fit2 <- glm(count ~ treatment * outcome, family = poisson, data)
> anova(fit1, fit2, test = "LRT")
Analysis of Deviance Table

Model 1: count ~ treatment + outcome
Model 2: count ~ treatment * outcome
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1         2      44.48
2         0       0.00  2    44.48 2.194e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(i) After introducing necessary notation, write down the Poisson models being fitted above.

(ii) Write down the corresponding multinomial models, then state the key theoretical result (the “Poisson trick”) that allows you to fit the multinomial models using Poisson regression. [You do not need to prove this theoretical result.]

(iii) Explain why the number of degrees of freedom in the likelihood ratio test is 2 in the analysis of deviance table. What can you conclude about the drug?

(b) Below is the summary table of the second model:

[QUESTION CONTINUES ON THE NEXT PAGE]

```
> summary(fit2)
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                8.65312    0.01321 654.899  < 2e-16 ***
treatmentLD               -1.43494    0.03013 -47.628  < 2e-16 ***
treatmentSD               -0.26081    0.02003 -13.021  < 2e-16 ***
outcomeWorse              -4.03800    0.10038 -40.228  < 2e-16 ***
treatmentLD:outcomeWorse -2.08156    0.58664  -3.548 0.000388 ***
treatmentSD:outcomeWorse -1.05847    0.21758  -4.865 1.15e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(i) *Drug efficacy* is defined as one minus the ratio of the probability of worsening in the treated group to the probability of worsening in the control group. By using a more sophisticated method, a published analysis estimated that the drug efficacy is 90.0% for the LD treatment and 62.1% for the SD treatment. Are these numbers similar to what is obtained by Poisson regression? [*Hint:* $e^{-1} \approx 0.37$*,* $e^{-2} \approx 0.14$*, and* $e^{-3} \approx 0.05$*, where e is the base of the natural logarithm.*]

(ii) Explain why the information in the summary table is not enough to test the hypothesis that the LD drug and the SD drug have the same efficacy. Then describe how you can test this hypothesis using analysis of deviance in R.

Paper 4, Section II

13J Statistical Modelling

Let X be an $n \times p$ non-random design matrix and Y be a n-vector of random responses. Suppose $Y \sim \mathrm{N}(\mu, \sigma^2 I)$, where μ is an unknown vector and $\sigma^2 > 0$ is known.

(a) Let $\lambda \geqslant 0$ be a constant. Consider the ridge regression problem

$$\hat{\beta}_\lambda = \arg\min_\beta \|Y - X\beta\|^2 + \lambda\|\beta\|^2 \,.$$

Let $\hat{\mu}_\lambda = X\hat{\beta}_\lambda$ be the fitted values. Show that $\hat{\mu}_\lambda = H_\lambda Y$, where

$$H_\lambda = X(X^T X + \lambda I)^{-1} X^T \,.$$

(b) Show that

$$\mathbb{E}(\|Y - \hat{\mu}_\lambda\|^2) = \|(I - H_\lambda)\mu\|^2 + \big\{n - 2\,\mathrm{trace}(H_\lambda) + \mathrm{trace}(H_\lambda^2)\big\}\sigma^2.$$

(c) Let $Y^* = \mu + \epsilon^*$, where $\epsilon^* \sim \mathrm{N}(0, \sigma^2 I)$ is independent of Y. Show that $\|Y - \hat{\mu}_\lambda\|^2 + 2\sigma^2 \mathrm{trace}(H_\lambda)$ is an unbiased estimator of $\mathbb{E}(\|Y^* - \hat{\mu}_\lambda\|^2)$.

(d) Describe the behaviour (monotonicity and limits) of $\mathbb{E}(\|Y^* - \hat{\mu}_\lambda\|^2)$ as a function of λ when $p = n$ and $X = I$. What is the minimum value of $\mathbb{E}(\|Y^* - \hat{\mu}_\lambda\|^2)$?

Paper 1, Section I

5J Statistical Modelling

Consider a generalised linear model with full column rank design matrix $X \in \mathbb{R}^{n\times p}$, output variables $Y = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$, link function g, mean parameters $\mu = (\mu_1, \ldots, \mu_n)$ and known dispersion parameters $\sigma_i^2 = a_i\sigma^2, i = 1, \ldots, n$. Denote its variance function by V and recall that $g(\mu_i) = x_i^T\beta, i = 1, \ldots, n$, where $\beta \in \mathbb{R}^p$ and x_i^T is the i^{th} row of X.

(a) Define the *score function* in terms of the log-likelihood function and the *Fisher information matrix*, and define the update of the Fisher scoring algorithm.

(b) Let $W \in \mathbb{R}^{n\times n}$ be a diagonal matrix with positive entries. Note that X^TWX is invertible. Show that

$$\operatorname{argmin}_{b\in\mathbb{R}^p}\left\{\sum_{i=1}^n W_{ii}(Y_i - x_i^Tb)^2\right\} = (X^TWX)^{-1}X^TWY.$$

[*Hint: you may use that* $\operatorname{argmin}_{b\in\mathbb{R}^p}\left\{\|Y - X^Tb\|^2\right\} = (X^TX)^{-1}X^TY$.]

(c) Recall that the score function and the Fisher information matrix have entries

$$\begin{aligned} U_j(\beta) &= \sum_{i=1}^n \frac{(Y_i - \mu_i)X_{ij}}{a_i\sigma^2 V(\mu_i)g'(\mu_i)} \qquad j = 1, \ldots, p, \\ i_{jk}(\beta) &= \sum_{i=1}^n \frac{X_{ij}X_{ik}}{a_i\sigma^2 V(\mu_i)\{g'(\mu_i)\}^2} \qquad j,k = 1, \ldots, p. \end{aligned}$$

Justify, performing the necessary calculations and using part (b), why the Fisher scoring algorithm is also known as the iterative reweighted least squares algorithm.

Paper 2, Section I

5J Statistical Modelling

The data frame WCG contains data from a study started in 1960 about heart disease. The study used 3154 adult men, all free of heart disease at the start, and eight and a half years later it recorded into variable chd whether they suffered from heart disease (1 if the respective man did and 0 otherwise) along with their height and average number of cigarettes smoked per day. Consider the R code below and its abbreviated output.

```
> data.glm <- glm(chd~height+cigs, family = binomial, data = WCG)
> summary(data.glm)
...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.50161    1.84186  -2.444   0.0145
height       0.02521    0.02633   0.957   0.3383
cigs         0.02313    0.00404   5.724 1.04e-08
...
```

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

(b) Interpret the effect on heart disease of a man smoking an average of two packs of cigarettes per day if each pack contains 20 cigarettes.

(c) Give an alternative latent logistic-variable representation of the model. [*Hint: if F is the cumulative distribution function of a logistic random variable, its inverse function is the logit function.*]

Paper 3, Section I

5J Statistical Modelling

Suppose we have data $(Y_1, x_1^T), \ldots, (Y_n, x_n^T)$, where the Y_i are independent conditional on the design matrix X whose rows are the $x_i^T, i = 1, \ldots, n$. Suppose that given x_i, the true probability density function of Y_i is f_{x_i}, so that the data is generated from an element of a model $\mathcal{F} := \{(f_{x_i}(\cdot\,;\theta))_{i=1}^n\,, \theta \in \Theta\}$ for some $\Theta \subseteq \mathbb{R}^q$ and $q \in \mathbb{N}$.

(a) Define the *log-likelihood function* for $\mathcal{F}$, the *maximum likelihood estimator* of θ and *Akaike's Information Criterion* (AIC) for $\mathcal{F}$.

From now on let $\mathcal{F}$ be the normal linear model, i.e. $Y := (Y_1, \ldots, Y_n)^T = X\beta + \varepsilon$, where $X \in \mathbb{R}^{n \times p}$ has full column rank and $\varepsilon \sim N_n(0, \sigma^2 I)$.

(b) Let $\hat{\sigma}^2$ denote the maximum likelihood estimator of σ^2. Show that the AIC of $\mathcal{F}$ is

$$n(1 + \log(2\pi\hat{\sigma}^2)) + 2(p+1).$$

(c) Let χ^2_{n-p} be a chi-squared distribution on $n-p$ degrees of freedom. Using any results from the course, show that the distribution of the AIC of $\mathcal{F}$ is

$$n\log(\chi^2_{n-p}) + n(\log(2\pi\sigma^2/n) + 1) + 2(p+1).$$

[*Hint:* $\hat{\sigma}^2 := n^{-1}\|Y - X\hat{\beta}\|^2 = n^{-1}\|(I-P)\varepsilon\|^2$, *where* $\hat{\beta}$ *is the maximum likelihood estimator of* β *and* P *is the projection matrix onto the column space of* X.]

Paper 4, Section I

5J Statistical Modelling

Suppose you have a data frame with variables `response`, `covar1`, and `covar2`. You run the following commands on `R`.

```
model <- lm(response ~ covar1 + covar2)
summary(model)
...
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.1024     0.1157 -18.164   <2e-16
covar1        1.6329     2.6557   0.615    0.542
covar2        0.3755     2.5978   0.145    0.886
...
```

(a) Consider the following three scenarios:

(i) All the output you have is the abbreviated output of `summary(model)` above.

(ii) You have the abbreviated output of `summary(model)` above together with

```
Residual standard error: 0.8097 on 47 degrees of freedom
Multiple R-squared:  0.8126, Adjusted R-squared:  0.8046
F-statistic: 101.9 on 2 and 47 DF, p-value: < 2.2e-16
```

(iii) You have the abbreviated output of `summary(model)` above together with

```
Residual standard error: 0.9184 on 47 degrees of freedom
Multiple R-squared:  0.000712, Adjusted R-squared:  -0.04181
F-statistic: 0.01674 on 2 and 47 DF, p-value: 0.9834
```

What conclusion can you draw about which variables explain the response in each of the three scenarios? Explain.

(b) Assume now that you have the abbreviated output of `summary(model)` above together with

```
anova(lm(response ~ 1), lm(response ~ covar1), model)
...
  Res.Df     RSS Df Sum of Sq        F Pr(>F)
1     49 164.448
2      ?  30.831  ?   133.618        ? <2e-16
3      ?  30.817  ?     0.014        ?      ?
...
```

What are the values of the entries with a question mark? [You may express your answers as arithmetic expressions if necessary].

Paper 1, Section II

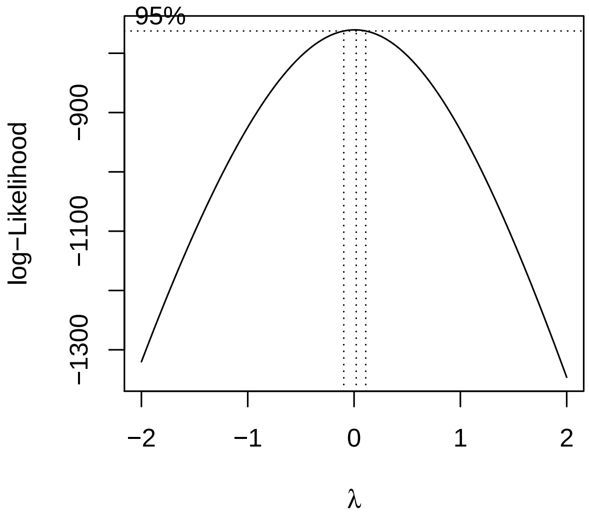
13J Statistical Modelling

We consider a subset of the data on car insurance claims from Hallin and Ingenbleek (1983). For each customer, the dataset includes total payments made per policy-year, the amount of kilometres driven, the bonus from not having made previous claims, and the brand of the car. The amount of kilometres driven is a factor taking values $1, 2, 3, 4$, or 5, where a car in level $i+1$ has driven a larger number of kilometres than a car in level i for any $i = 1, 2, 3, 4$. A statistician from an insurance company fits the following model on `R`.

```
> model1 <- lm(Paymentperpolicyyr ~ as.numeric(Kilometres) + Brand + Bonus)
```

(i) Why do you think the statistician transformed variable `Kilometres` from a factor to a numerical variable?

(ii) To check the quality of the model, the statistician applies a function to `model1` which returns the following figure:



What does the plot represent? Does it suggest that `model1` is a good model? Explain. If not, write down a model which the plot suggests could be better.

[QUESTION CONTINUES ON THE NEXT PAGE]

(iii) The statistician fits the model suggested by the graph and calls it `model2`. Consider the following abbreviated output:

```
> summary(model2)
...
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             6.514035   0.186339  34.958  < 2e-16 ***
as.numeric(Kilometres)  0.057132   0.032654   1.750  0.08126 .
Brand2                  0.363869   0.186857   1.947  0.05248 .
...
Brand9                  0.125446   0.186857   0.671  0.50254
Bonus                  -0.178061   0.022540  -7.900 6.17e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.7817 on 284 degrees of freedom
...
```

Using the output, write down a 95% prediction interval for the ratio between the total payments per policy year for two cars of the same brand and with the same value of `Bonus`, one of which has a `Kilometres` value one higher than the other. You may express your answer as a function of quantiles of a common distribution, which you should specify.

(iv) Write down a generalised linear model for `Paymentperpolicyyr` which may be a better model than `model1` and give two reasons. You must specify the link function.

Paper 4, Section II

13J Statistical Modelling

(a) Define a *generalised linear model* (GLM) with design matrix $X \in \mathbb{R}^{n\times p}$, output variables $Y := (Y_1, \ldots, Y_n)^T$ and parameters $\mu := (\mu_1, \ldots, \mu_n)^T$, $\beta \in \mathbb{R}^p$ and $\sigma_i^2 := a_i\sigma^2 \in (0, \infty), i = 1, \ldots, n$. Derive the moment generating function of Y, i.e. give an expression for $\mathbb{E}\left[\exp\left(t^T Y\right)\right], t \in \mathbb{R}^n$, wherever it is well-defined.

Assume from now on that the GLM satisfies the usual regularity assumptions, X has full column rank, and σ^2 is known and satisfies $1/\sigma^2 \in \mathbb{N}$.

(b) Let $\tilde{Y} := \left(\tilde{Y}_1, \ldots, \tilde{Y}_{n/\sigma^2}\right)^T$ be the output variables of a GLM from the same family as that of part (a) and parameters $\tilde{\mu} := (\tilde{\mu}_1, \ldots, \tilde{\mu}_{n/\sigma^2})^T$ and $\tilde{\sigma}^2 := (\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_{n/\sigma^2}^2)$. Suppose the output variables may be split into n blocks of size $1/\sigma^2$ with constant parameters. To be precise, for each block $i = 1, \ldots, n$, if $j \in \{(i-1)/\sigma^2 + 1, \ldots, i/\sigma^2\}$ then

$$\tilde{\mu}_j = \mu_i \qquad \text{and} \qquad \tilde{\sigma}_j^2 = a_i$$

with $\mu_i = \mu_i(\beta)$ and a_i defined as in part (a). Let $\bar{Y} := (\bar{Y}_1, \ldots, \bar{Y}_n)^T$, where $\bar{Y}_i := \sigma^2 \sum_{k=1}^{1/\sigma^2} \tilde{Y}_{(i-1)/\sigma^2+k}$.

(i) Show that $\bar{Y}$ is equal to Y in distribution. [*Hint: you may use without proof that moment generating functions uniquely determine distributions from exponential dispersion families.*]

(ii) For any $\tilde{y} \in \mathbb{R}^{n/\sigma^2}$, let $\bar{y} = (\bar{y}_1, \ldots, \bar{y}_n)^T$, where $\bar{y}_i := \sigma^2 \sum_{k=1}^{1/\sigma^2} \tilde{y}_{(i-1)/\sigma^2+k}$. Show that the model function of $\tilde{Y}$ satisfies

$$f\left(\tilde{y}; \tilde{\mu}, \tilde{\sigma}^2\right) = g_1\left(\bar{y}; \tilde{\mu}, \tilde{\sigma}^2\right) \times g_2\left(\tilde{y}; \tilde{\sigma}^2\right)$$

for some functions g_1, g_2, and conclude that $\bar{Y}$ is a sufficient statistic for β from $\tilde{Y}$.

(iii) For the model and data from part (a), let $\hat{\mu}$ be the maximum likelihood estimator for μ and let $D(Y; \mu)$ be the deviance at μ. Using (i) and (ii), show that

$$\frac{D(Y; \hat{\mu})}{\sigma^2} =^d 2\log\left\{\frac{\sup_{\tilde{\mu}' \in \widetilde{\mathcal{M}}_1} f(\tilde{Y}; \tilde{\mu}', \tilde{\sigma}^2)}{\sup_{\tilde{\mu}' \in \widetilde{\mathcal{M}}_0} f(\tilde{Y}; \tilde{\mu}', \tilde{\sigma}^2)}\right\},$$

where $=^d$ means equality in distribution and $\widetilde{\mathcal{M}}_0$ and $\widetilde{\mathcal{M}}_1$ are nested subspaces of $\mathbb{R}^{n/\sigma^2}$ which you should specify. Argue that $\dim(\widetilde{\mathcal{M}}_1) = n$ and $\dim(\widetilde{\mathcal{M}}_0) = p$, and, assuming the usual regularity assumptions, conclude that

$$\frac{D(Y; \hat{\mu})}{\sigma^2} \to^d \chi_{n-p}^2 \qquad \text{as } \sigma^2 \to 0,$$

stating the name of the result from class that you use.

Paper 4, Section I

5J Statistical Modelling

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* χ^2_{n-p} *distribution.*]

Paper 3, Section I

5J Statistical Modelling

(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 $\sigma_1^2 = \cdots = \sigma_n^2 = \sigma^2 > 0$. Assume σ^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

$$\sigma^{-2} X^T W X,$$

for a diagonal matrix W depending on the means μ. Compute the entries of W in terms of μ.

Paper 2, Section I

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.

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

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.66937    0.07965   8.404  < 2e-16 ***
miles       -0.04917    0.01839  -2.674  0.00761 **
history      1.48954    0.04818  30.918  < 2e-16 ***

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

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.09539    0.06960   1.371    0.171
incentiveYes  0.91387    0.06504  14.051   <2e-16 ***
history       1.46806    0.04346  33.782   <2e-16 ***

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

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.47050    0.11682  12.588  < 2e-16 ***
incentiveYes  1.73282    0.10917  15.872  < 2e-16 ***
history       0.47322    0.07294   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.

Paper 1, Section I

5J Statistical Modelling

The Gamma distribution with shape parameter $\alpha > 0$ and scale parameter $\lambda > 0$ has probability density function

$$f(y; \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\lambda y} \qquad \text{for } y > 0$$

where Γ 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 α and λ.

Paper 4, Section II

13J Statistical Modelling

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,

$$\mathbb{P}(y_{i,j} = 1) = \frac{e^{\theta_{i,j}}}{1 + e^{\theta_{i,j}}}, \quad 1 \leqslant i < j \leqslant m,$$

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 $\ell \neq k$ (ℓ 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 β if and only if $t(y) = t(y')$; then, T is a minimal sufficient statistic for β.]

Paper 1, Section II

13J 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.

```
> 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?

Paper 4, Section I

5J Statistical Modelling

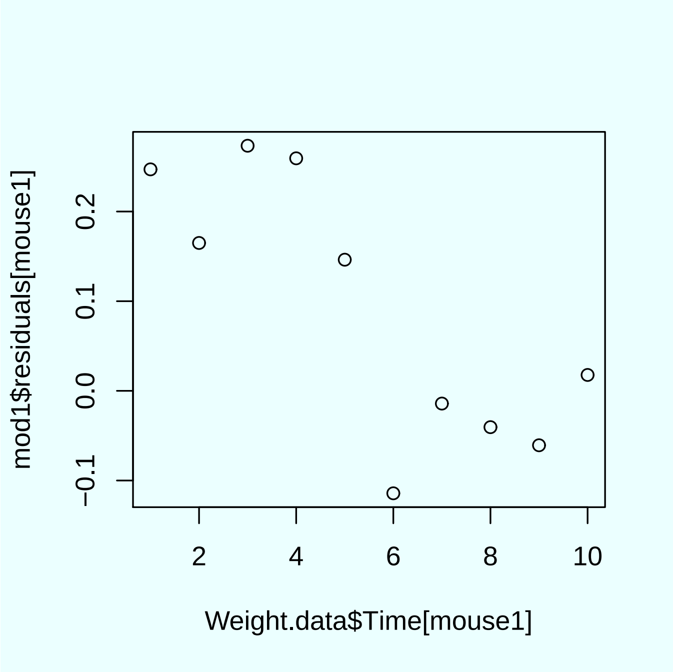
A scientist is studying the effects of a drug on the weight of mice. Forty mice are divided into two groups, control and treatment. The mice in the treatment group are given the drug, and those in the control group are given water instead. The mice are kept in 8 different cages. The weight of each mouse 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.

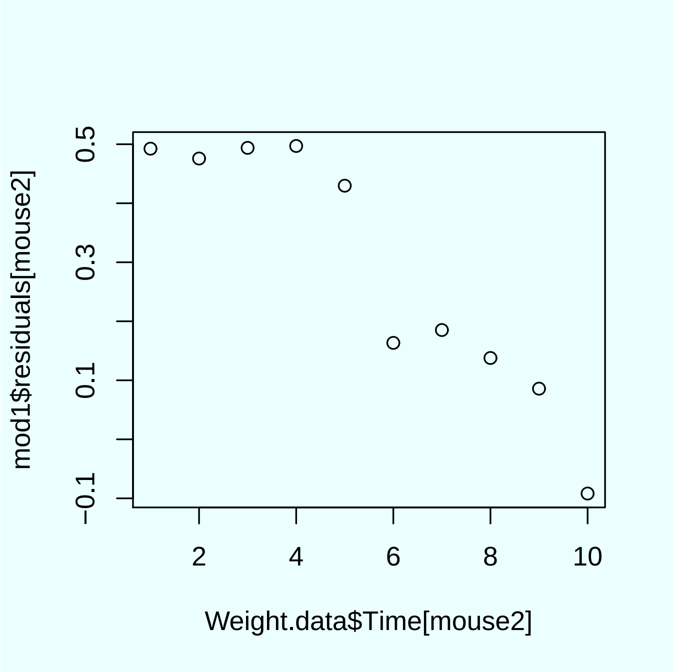
```
> head(Weight.data)
  Time   Group Cage Mouse   Weight
1    1 Control    1     1 24.77578
2    2 Control    1     1 24.68766
3    3 Control    1     1 24.79008
4    4 Control    1     1 24.77005
5    5 Control    1     1 24.65092
6    6 Control    1     1 24.38436
> mod1 = lm(Weight ~ Time*Group + 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
```

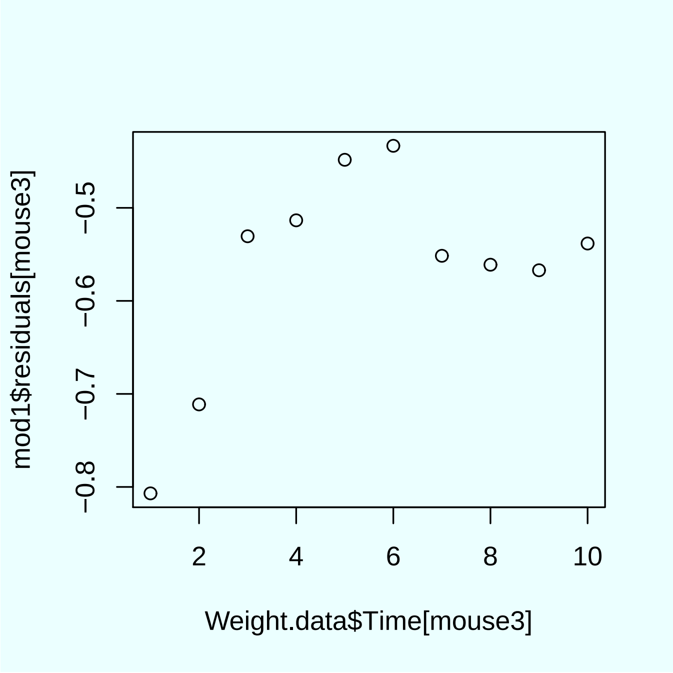
Which parameters describe the rate of weight loss with time in each group? According to the R output, is there a statistically significant weight loss with time in the control group?

Three diagnostic plots were generated using the following R code.

```
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.

Paper 3, Section I

5J Statistical Modelling

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.

```
> 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 χ^2_p distribution with a range of degrees of freedom p. How would you explain the discrepancy between their conclusions?

Degrees of freedom	99th percentile	Degrees of freedom	99th percentile
1	6.63	21	38.93
2	9.21	22	40.29
3	11.34	23	41.64
4	13.28	24	42.98
5	15.09	25	44.31
6	16.81	26	45.64
7	18.48	27	46.96
8	20.09	28	48.28
9	21.67	29	49.59
10	23.21	30	50.89
11	24.72	31	52.19
12	26.22	32	53.49
13	27.69	33	54.78
14	29.14	34	56.06
15	30.58	35	57.34
16	32.00	36	58.62
17	33.41	37	59.89
18	34.81	38	61.16
19	36.19	39	62.43
20	37.57	40	63.69

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.

Paper 1, Section I

5J Statistical Modelling

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
```

```
> mod2 = glm(Ambulance.requests ~ Winter+Weekend, data=Ambulance, family=poisson)
> summary(mod2)

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4480  -0.8544  -0.1153   0.7689   3.5903

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.97077    0.02163  137.34   <2e-16 ***
WinterYes    0.55586    0.03268   17.01   <2e-16 ***
WeekendYes  -0.60371    0.03813  -15.84   <2e-16 ***
---
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: 316.39  on 197  degrees of freedom
AIC: 1267.9
```

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.

```
> 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 m members. The outcomes are modelled by

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

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 β_1, β_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 β, 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 β 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

$$h_i(t) = \frac{f_i(t)}{1 - F_i(t)} \quad \text{for } t \geqslant 0.$$

We shall assume that $h_i(t) = \lambda(t)\exp(\beta^\top x_i)$, 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 β in terms of λ, β, x_i and Y_i only.

(c) Show that the maximum likelihood estimate of β can be obtained through a surrogate Poisson generalised linear model with an offset.

Paper 1, Section I

5J Statistical Modelling

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.

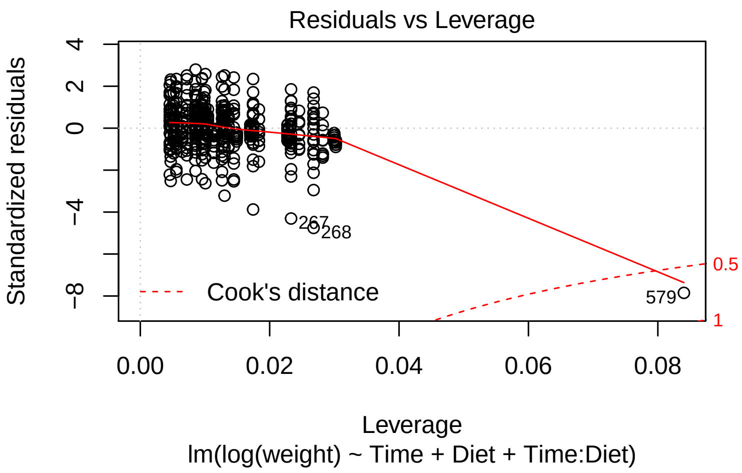
```
attach(ChickWeight)
fit1 = lm(weight˜ Time+Diet)
fit2 = lm(log(weight)˜ Time+Diet)
fit3 = lm(log(weight)˜ Time+Diet+Time:Diet)
```

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

```
> anova(fit2,fit3)
Analysis of Variance Table

Model 1: log(weight) ˜ Time + Diet
Model 2: log(weight) ˜ Time + Diet + Time:Diet
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    574 34.381
2    571 31.589  3    2.7922 16.824 1.744e-10 ***
```

(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?



Paper 2, Section I

5J Statistical Modelling

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 β_j for $j > 0$.

Paper 3, Section I

5J Statistical Modelling

For Fisher's method of Iteratively Reweighted Least-Squares and Newton–Raphson optimisation of the log-likelihood, the vector of parameters β 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.

Paper 4, Section I

5J Statistical Modelling

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)
```

Paper 1, Section II

12J Statistical Modelling

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

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

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 $\tilde{P}_{ab} = \mathbb{E}\tilde{y}_{ab}/\tilde{n}_{ab}$ and $g(\tilde{P}_{ab}) = \beta_a - \beta_b$, where the β_a 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.]

Paper 4, Section II

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.

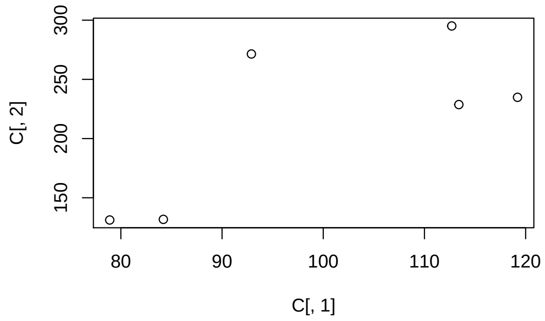
(b) What is the difference between the deviance of the models `fit.2` and `fit.3`?

(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 0.189481
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])
```



Paper 2, Section I

5K Statistical Modelling

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

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

Paper 4, Section I

5K Statistical Modelling

(a) Let $Y_i = x_i^\intercal \beta + \varepsilon_i$ where ε_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 ε_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 $\beta^{(t)}$:

$$\ell(\beta; Y) \approx \ell(\beta^{(t)}; Y) + (\beta - \beta^{(t)})^\intercal D\ell(\beta^{(t)}; Y) + (\beta - \beta^{(t)})^\intercal D^2\ell(\beta^{(t)}; Y)(\beta - \beta^{(t)}),$$

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?

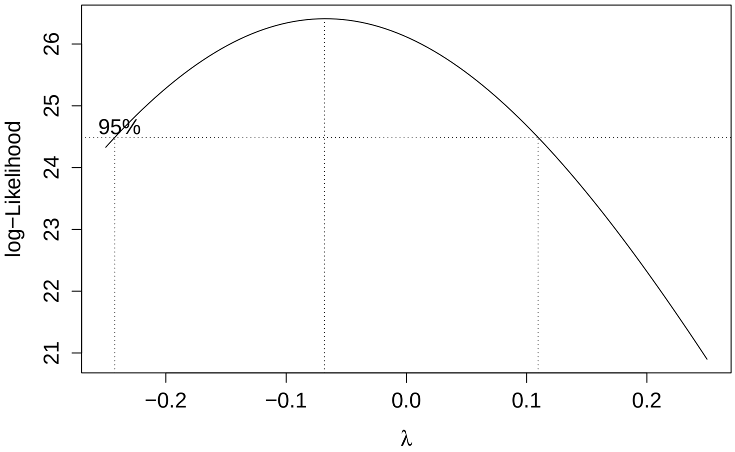
Paper 3, Section I

5K Statistical Modelling

The R command

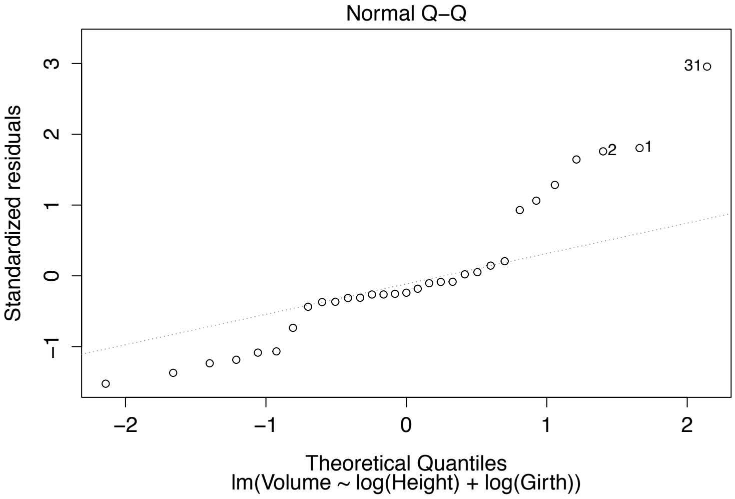
```
> boxcox(rainfall ~ month+elnino+month:elnino)
```

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



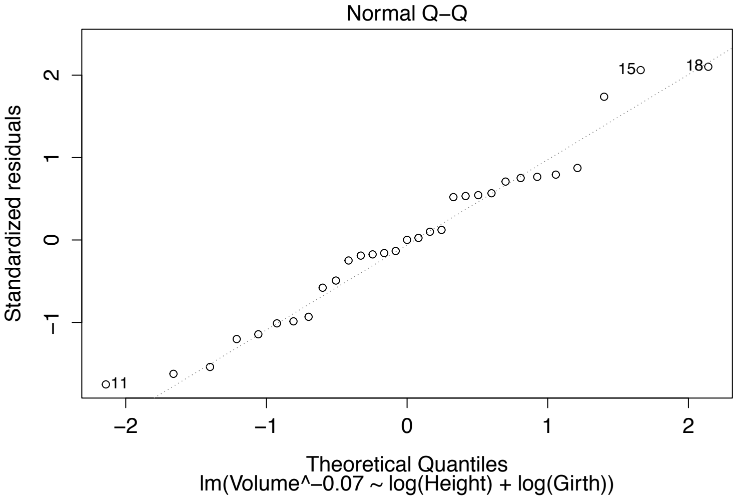
We fit two linear models and obtain the Q–Q plots for each fit, which are shown below in no particular order:

```
> 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)
```



This question continues on the next page

5K Statistical Modelling (continued)



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 ith 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?

Paper 1, Section I

5K Statistical Modelling

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

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

fit the models $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$ and $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$, respectively, with $\varepsilon \sim N(0, \sigma^2)$ in each case.

Setting the parameters raw to FALSE:

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

fits the models $Y = \beta_0 + \beta_1 P_1(X) + \beta_2 P_2(X) + \varepsilon$ and $Y = \beta_0 + \beta_1 P_1(X) + \beta_2 P_2(X) + \beta_3 P_3(X) + \varepsilon$, with $\varepsilon \sim N(0, \sigma^2)$. The function P_i is a polynomial of degree i. Furthermore, the design matrix output by the function poly with raw=F satisfies:

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

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

```
> 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?

Paper 4, Section II

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 $j \geqslant 2$, 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):

```
> 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:

```
> 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?

Paper 1, Section II

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

$$s_i = \frac{Y_i - x_i^\intercal \hat{\beta}}{\tilde{\sigma}\sqrt{1-p_i}},$$

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

$$\tilde{\sigma}^2 = \frac{\|Y - X\hat{\beta}\|_2^2}{(n-p)}.$$

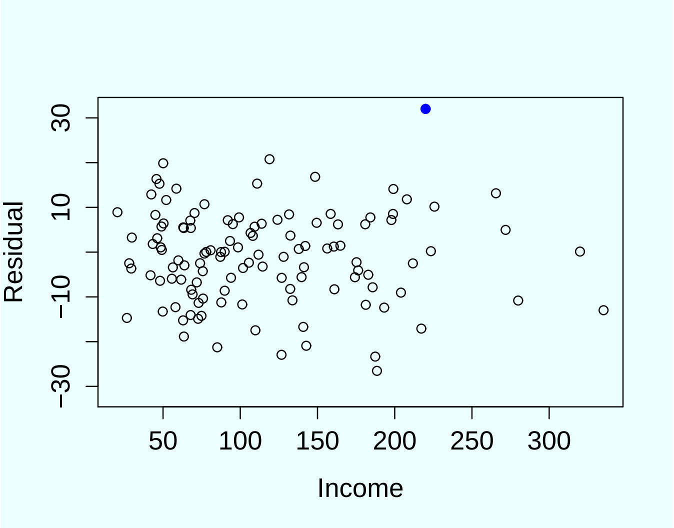
Prove that the joint distribution of $s = (s_1, \ldots, s_n)^\intercal$ is the same in the following two models: (i) $\varepsilon \sim N(0, \sigma I)$, and (ii) $\varepsilon \mid \sigma \sim N(0, \sigma I)$, with $1/\sigma \sim \chi^2_\nu$ (in this model, $\varepsilon_1, \ldots, \varepsilon_n$ are identically t_ν-distributed). [*Hint: A random vector Z is spherically symmetric if for any orthogonal matrix H, $HZ \overset{d}{=} Z$. 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*,

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

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%?



Paper 4, Section I

4J Statistical Modelling

Data on 173 nesting female horseshoe crabs record for each crab its colour as one of 4 factors (simply labelled $1, \ldots, 4$), 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.

```
> 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.

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

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

```
> 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.]

Paper 3, Section I

4J Statistical Modelling

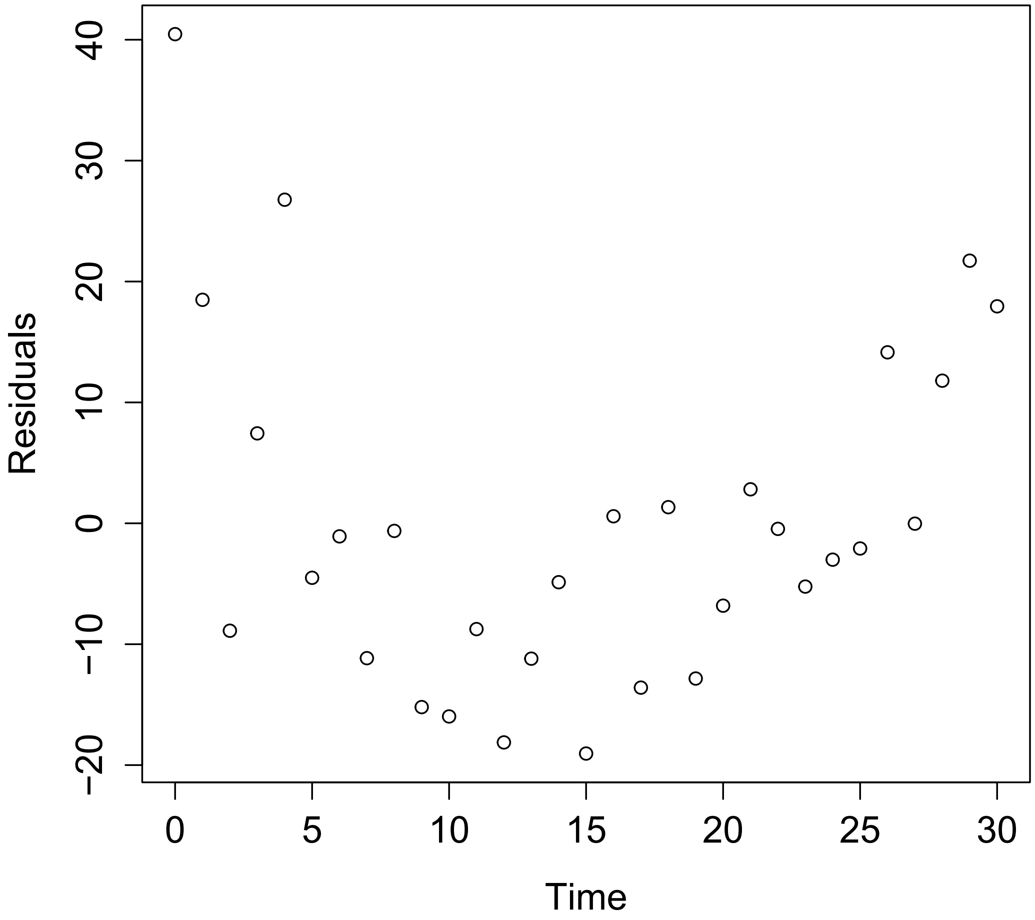
Data are available on the number of counts (atomic disintegration events that take place within a radiation source) recorded with a Geiger counter at a nuclear plant. The counts were registered at each second over a 30 second period for a short-lived, man-made radioactive compound. The first few rows 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.



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 β. Find the log-likelihood for β.

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

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

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 ε_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

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

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

How can we recover μ and β from the parameters of the probit regression model?

Paper 4, Section II

10J Statistical Modelling

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 $p \geqslant 3$ and $\beta \in \mathbb{R}^p$ is an unknown vector of regression coefficients. For $j \in \{1, \ldots, p\}$, denote the jth column of X by X_j, and let X_{-j} be X with its jth column removed. Suppose $X_1 = 1_n$ where 1_n is an n-vector of 1's. Denote the maximum likelihood estimate of β 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

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

[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 Π_V and Π_W denote orthogonal projections onto V and W respectively, then for all $v \in \mathbb{R}^n$, $\|\Pi_W v\|^2 \geqslant \|\Pi_V v\|^2$.]

This question continues on the next page

10J Statistical Modelling (continued)

By considering the fitted values $X\hat{\beta}$, explain why if, for any $j \geqslant 2$, a constant is added to each entry in the jth column of X, then $\hat{\beta}_j$ will remain unchanged. Let $\bar{X}_j = \sum_{i=1}^n X_{ij}/n$. Why is $(*)$ also true when all instances of X_j and X_k are replaced by $X_j - \bar{X}_j 1_n$ and $X_k - \bar{X}_k 1_n$ 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   0.331224 0.3267138
Maths_test 0.3312240   1.000000 0.9371630
Stat_test  0.3267138   0.937163 1.0000000
```

Paper 1, Section II

10J Statistical Modelling

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.

```
> 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.

```
> 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`.

```
> 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.

```
> 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.

Paper 4, Section I

5K Statistical Modelling

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 β.

Define the *Cook's distance* for the observation (Y_i, x_i) where x_i^T is the ith row of X, and give its interpretation in terms of confidence sets for β.

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* $\mathbb{P}(Z \leqslant 1.06) = 0.1$, $\mathbb{P}(Z \leqslant 7.78) = 0.9$.
2. *If* $Z \sim F_{4,96}$, *then* $\mathbb{P}(Z \leqslant 0.26) = 0.1$, $\mathbb{P}(Z \leqslant 2.00) = 0.9$.
3. *If* $Z \sim F_{96,4}$, *then* $\mathbb{P}(Z \leqslant 0.50) = 0.1$, $\mathbb{P}(Z \leqslant 3.78) = 0.9$.]

Paper 3, Section I

5K Statistical Modelling

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`.

```
> 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
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     38 49.815
2     24 35.480 14    14.335 0.6926 0.7599
```

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. $\|X(X^TX)^{-1}X^TY\|^2$
3. $\|(I-P_0)Y\|^2 - \|(I-P)Y\|^2$
4. $\dfrac{\|(P-P_0)Y\|^2/14}{\|(I-P)Y\|^2/24}$
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?

Paper 2, Section I

5K Statistical Modelling

Define the concept of an *exponential dispersion family*. Show that the family of scaled binomial distributions $\frac{1}{n}\mathrm{Bin}(n,p)$, 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?

Paper 1, Section I

5K Statistical Modelling

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.

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

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

```
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 $Z \sim$ Pois(μ) then*

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

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

Paper 4, Section II

13K Statistical Modelling

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:

```
  disease nondisease gender       food
1      77        381    Boy Bottle-fed
2      19        128    Boy Supplement
3      47        447    Boy Breast-fed
4      48        336   Girl Bottle-fed
5      16        111   Girl Supplement
6      31        433   Girl Breast-fed
```

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* $\mathbb{P}(Z \geqslant 1.96) = 0.025$.]

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?

Paper 1, Section II

13K Statistical Modelling

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 $n \geqslant p+3$, and $\beta \in \mathbb{R}^p$ is an unknown vector of regression coefficients. Let X_0 be the matrix formed from the first $p_0 < p$ columns of X, and partition β as $\beta = (\beta_0^T, \beta_1^T)^T$ 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

$$\mathbb{E}_{\beta,\sigma^2}(F) = \frac{(n-p)(\tau^2+1)}{n-p-2},$$

where $\tau^2 = \dfrac{\|(P - P_0)X\beta\|^2}{(p - p_0)\sigma^2}$.

[*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 Π satisfies $\|\Pi\varepsilon\|^2 \sim \sigma^2\chi^2_\nu$, where $\nu = \text{rank}(\Pi)$; and if $Z \sim \chi^2_\nu$ then $\mathbb{E}(Z^{-1}) = (\nu - 2)^{-1}$ when $\nu > 2$.]

Paper 4, Section I

5J Statistical Modelling

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.

```
> 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

           Df  Sum Sq  mean Sq  F value    Pr(>F)
a           3   71.51    23.84    17.79   1.34e-8  ***
b           1  105.11   105.11    78.44  6.91e-13  ***
Residuals  67   89.56     1.34
```

Paper 3, Section I

5J Statistical Modelling

Consider the linear model $Y = X\beta + \epsilon$ where $Y = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^{\mathrm{T}}$, 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 $\widehat{\beta}$ and $\widehat{\sigma}^2$ of β and σ^2 respectively, and state their joint distribution. Show that $\widehat{\beta}$ is unbiased whereas $\widehat{\sigma}^2$ is biased.

Suppose that a new variable Y^* is to be observed, satisfying the relationship

$$Y^* = x^{*\mathrm{T}}\beta + \epsilon^* \,,$$

where x^* $(p \times 1)$ is known, and $\epsilon^* \sim N(0, \sigma^2)$ independently of ϵ. We propose to predict Y^* by $\widetilde{Y} = x^{*\mathrm{T}}\widehat{\beta}$. Identify the distribution of

$$\frac{Y^* - \widetilde{Y}}{\tau\,\widetilde{\sigma}},$$

where

$$\begin{aligned} \widetilde{\sigma}^2 &= \frac{n}{n-p}\widehat{\sigma}^2 \,, \\ \tau^2 &= x^{*\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x^* + 1 \,. \end{aligned}$$

Paper 2, Section I

5J Statistical Modelling

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

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 β_3 is orthogonal to $(\beta_0, \beta_1, \beta_2)^{\mathrm{T}}$ and β_1 is orthogonal to $(\beta_0, \beta_2)^{\mathrm{T}}$.

Let $\widetilde{\beta} = (\beta_0, \beta_2, \beta_1, \beta_3)^{\mathrm{T}}$. Describe the matrix $\widetilde{X}$ such that $Y = \widetilde{X}\widetilde{\beta} + \epsilon$. Show that $\widetilde{X}^{\mathrm{T}}\widetilde{X}$ is block diagonal. Assuming further that this matrix is non-singular, show that the least-squares estimators of β_1 and β_3 are, respectively,

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2} \qquad \text{and} \qquad \widehat{\beta}_3 = \frac{\sum_{i=1}^n P_3(x_i) Y_i}{\sum_{i=1}^n P_3(x_i)^2} \,.$$

Paper 1, Section I

5J Statistical Modelling

Variables $Y_1, \ldots, Y_n$ are independent, with Y_i having a density $p(y \mid \mu_i)$ governed by an unknown parameter μ_i. Define the *deviance* for a model M that imposes relationships between the (μ_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 $\widehat{\mu}_i$ be the maximum likelihood estimate of μ_i under model M. Show that the deviance for this model is given by

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

Now suppose that, under M, $\log \mu_i = \beta^{\mathrm{T}} x_i$, $i = 1, \ldots, n$, where $x_1, \ldots, x_n$ are known p-dimensional explanatory variables and β is an unknown p-dimensional parameter. Show that $\widehat{\mu} := (\widehat{\mu}_1, \ldots, \widehat{\mu}_n)^{\mathrm{T}}$ satisfies $X^{\mathrm{T}} y = X^{\mathrm{T}} \widehat{\mu}$, where $y = (y_1, \ldots, y_n)^{\mathrm{T}}$ and X is the $(n \times p)$ matrix with rows $x_1^{\mathrm{T}}, \ldots, x_n^{\mathrm{T}}$, and express this as an equation for the maximum likelihood estimate $\widehat{\beta}$ of β. [You are not required to solve this equation.]

Paper 4, Section II

13J 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.

A random variable Y is said to have an *inverse Gaussian distribution*, with parameters ϕ and λ (both positive), if its density function is

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

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 $\mathrm{Var}(Y)$ in terms of ϕ and λ. 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 $g(\mu_i)=x_i^{\mathrm{T}}\beta$, where $\beta=(\beta_1,\ldots,\beta_p)^{\mathrm{T}}$ is the vector of unknown regression coefficients and $x_i=(x_{i1},\ldots,x_{ip})^{\mathrm{T}}$ 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 β. Explain how these can be used to compute the maximum likelihood estimate $\widehat{\beta}$ of β.

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):

```
Day  Wind  Level   Stumps
1    10    Medium  26
2    8     Medium  37
⋮    ⋮     ⋮       ⋮
50   12    Medium  32
51   7     Fast    31
⋮    ⋮     ⋮       ⋮
120  3     Fast    28
121  5     Spin    35
⋮    ⋮     ⋮       ⋮
150  6     Spin    31
```

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):

```
>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`?

Paper 4, Section I

5K Statistical Modelling

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 $\mathrm{Var}(Y)$ when $Y \sim \mathrm{Pois}(\lambda)$. What is the canonical link function in this case?

Paper 3, Section I

5K Statistical Modelling

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 ε_i are independent and identically distributed with $N(0, \sigma^2)$ distribution. What does it mean for the pair β_1 and β_2 to be *orthogonal*? What does it mean for all the three parameters β_0, β_1 and β_2 to be *mutually orthogonal*? Give necessary and sufficient conditions on $(x_{i1})_{i=1}^n, (x_{i2})_{i=1}^n$ so that β_0, β_1 and β_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 $\hat{\beta}_2$.

Paper 2, Section I

5K Statistical Modelling

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' (`thorax`) which is known to be positively correlated with lifespan. Data was collected on the following variables:

`longevity` lifespan in days

`thorax` (body) length in mm

`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 `thorax` as the covariate, `treat` as a factor (having the above 5 levels) and `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 `treat10` and `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 λ is shown in the right panel of Figure 1. What transformation should be used to improve the fit and yet retain some interpretability?

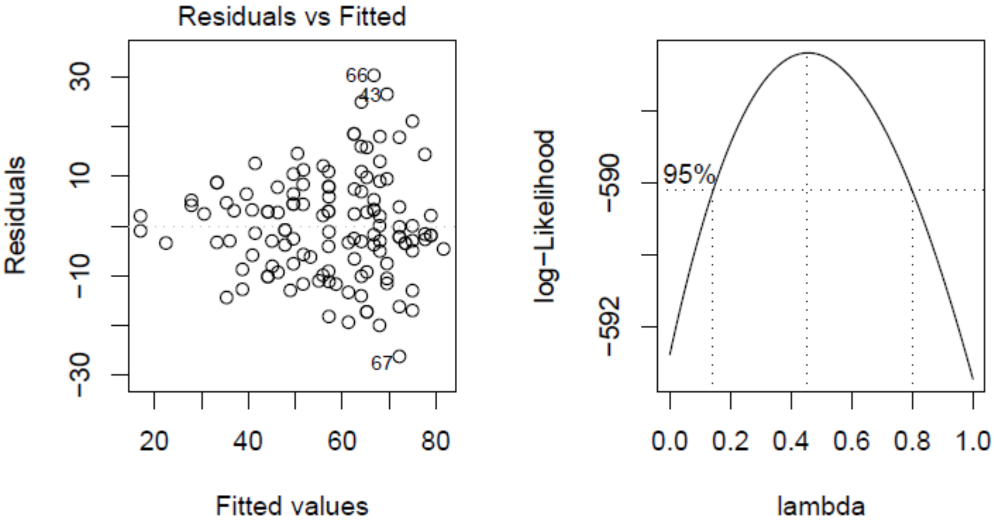


Figure 1: Residual-Fitted plot on the left and Box-Cox plot on the right

Paper 1, Section I

5K Statistical Modelling

Let $Y_1, \ldots, Y_n$ be independent with $Y_i \sim \frac{1}{n_i}\text{Bin}(n_i, \mu_i)$, $i = 1, \ldots, n$, and

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = x_i^\top \beta\,, \tag{1}$$

where x_i is a $p \times 1$ vector of regressors and β 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 μ, 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.

Paper 4, Section II

13K Statistical Modelling

Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be jointly independent and identically distributed with $X_i \sim N(0,1)$ and conditional on $X_i = x$, $Y_i \sim N(x\theta, 1)$, $i = 1, 2, \dots, n$.

(a) Write down the likelihood of the data $(X_1, Y_1), \dots, (X_n, Y_n)$, and find the maximum likelihood estimate $\hat{\theta}$ of θ. [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 θ 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 $X_1, X_2, \dots, X_n$ and use the following property of conditional expectation: for U, V random vectors, any suitable function g, and $x \in \mathbb{R}$,*

$$P\{g(U,V) \leqslant x\} = E[P\{g(U,V) \leqslant x | V\}].]$$

Paper 1, Section II

13K Statistical Modelling

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.

`aged` The patient's age, split into less than 60 (=0) and 60 or over (=1).

`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.

`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.

`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.

`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:
     Min       1Q   Median     3Q     Max
  -2.332  -0.665   -0.300  0.639  2.150

Coefficients:
            Estimate  Std. Error  t value  Pr(>|z|)
(Intercept)   -3.079       0.987    -3.12    0.0018
aged          -0.292       0.754    -0.39    0.6988
grade          0.872       0.816     1.07    0.2850
stage          1.373       0.784     1.75    0.0799
xray           1.801       0.810     2.22    0.0263
acid           1.684       0.791     2.13    0.0334

(Dispersion parameter for binomial family taken to be 1)

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 $\hat{\beta} - \hat{\beta}_{(i)}$ 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?

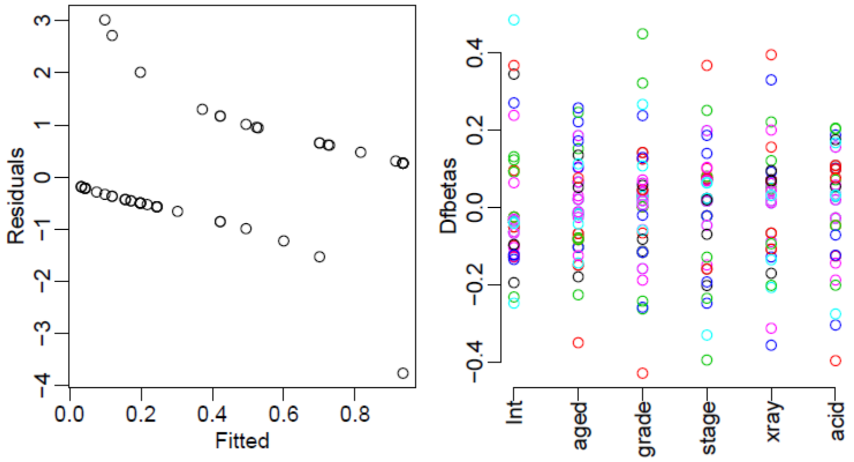


Figure 1: The plot on the left shows the Pearson residuals and the fitted values. The plot on the right shows the changes in the regression coefficients when a single point is omitted for each predictor.

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_θ and Var_θ 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 θ, then its variance satisfies $\mathrm{Var}_\theta(T) \geqslant (n\mathrm{Var}_\theta(U_1(\theta)))^{-1}$. [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?

Paper 4, Section I

5J Statistical Modelling

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

The data look like this:

```
   count  n freq loc sex   age
1     68 31    F  NB   M 15-19
2     14  4    F  NB   F 15-19
3     35 12    F  NB   M 20-24
4     16 11    F  NB   F 20-24
[...]
23     5 15   NF   B   M 25-29
24     6  6   NF   B   F 25-29
```

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

We fit the following Poisson model:

$$\log(\mathbb{E}(\texttt{count}_j)) = \log(n_j\mu_j) = \log(n_j) + \mathbf{x}_j\beta,$$

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):

```
> 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:

```
  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

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

provides the following (slightly abbreviated) summary:

```
[...]
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.992039   4.302734  -2.090   0.0447 *
dist         6.217956   0.601148  10.343 9.86e-12 ***
climb        0.011048   0.002051   5.387 6.45e-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?

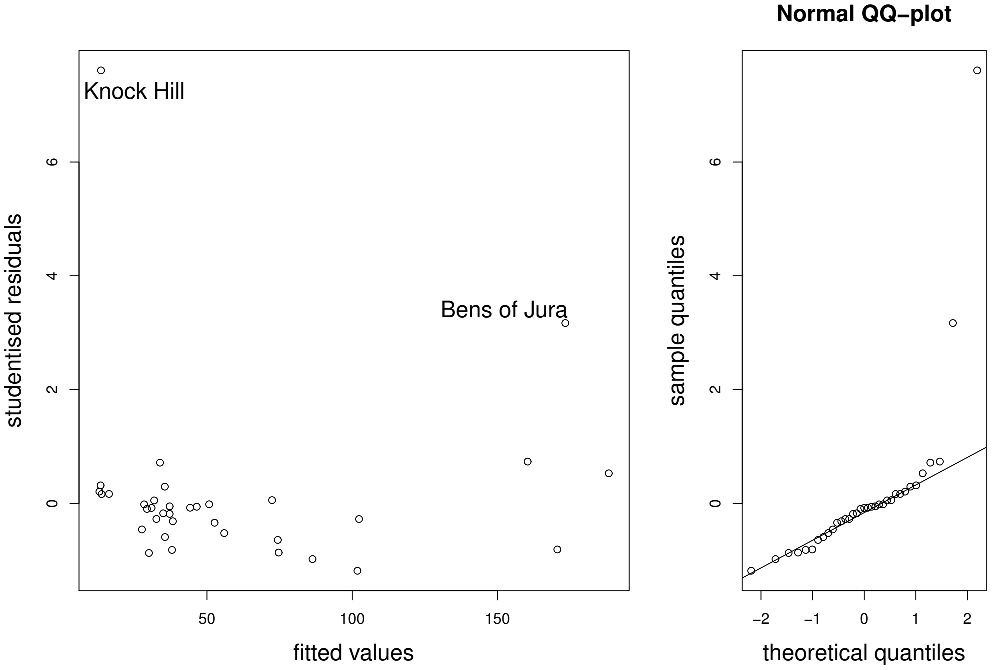


Figure 1: Hills data: diagnostic plots

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 $p \leqslant n$, and where ϵ has a multivariate normal distribution with mean zero and covariance matrix $\sigma^2 I_n$. Write down the likelihood function for β, σ^2 and derive the maximum likelihood estimators $\hat{\beta}, \hat{\sigma}^2$ of β, σ^2. Find the distribution of $\hat{\beta}$. Show further that $\hat{\beta}$ and $\hat{\sigma}^2$ are independent.

Paper 1, Section I

5J Statistical Modelling

Consider a binomial generalised linear model for data $y_1, ..., y_n$ modelled as realisations of independent $Y_i \sim \mathrm{Bin}(1, \mu_i)$ 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 β. Find the log-likelihood for β, and the likelihood equation that must be solved to find the maximum likelihood estimator $\hat{\beta}$ of β. Compute the second derivative of the log-likelihood for β, and explain the algorithm you would use to find $\hat{\beta}$.

Paper 2, Section I

5J Statistical Modelling

Suppose you have a parametric model consisting of probability mass functions $f(y;\theta)$, $\theta \in \Theta \subset \mathbb{R}$. Given a sample $Y_1, ..., Y_n$ from $f(y;\theta)$, define the maximum likelihood estimator $\hat{\theta}_n$ for θ 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 θ. If $Y_1, ..., Y_n$ are independent and identically distributed random variables having a Poisson distribution with parameter θ, 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 θ. Without calculating the variance of S, show that there is no reason to prefer S over $\overline{Y}$.

[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.]

Paper 3, Section I

5J Statistical Modelling

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 σ^2. Without using Cochran's theorem, show carefully that $\hat{\sigma}^2$ is biased. Suggest another estimator $\tilde{\sigma}^2$ for σ^2 that is unbiased.

Paper 4, Section I

5J Statistical Modelling

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):

```
> 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  29.1  26     5    180    67   3375
4      Audi      100  37.7  26     6    193    70   3405
5       BMW     535i  30.0  30     4    186    69   3640
        ...               ...                        ...
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:

```
> summary(fit)

[...]

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.8751080  0.7687276   5.041 2.50e-06 ***
mpg         -0.0109953  0.0085475  -1.286 0.201724
psngr       -0.1782818  0.0290618  -6.135 2.45e-08 ***
length       0.0067382  0.0032890   2.049 0.043502 *
width       -0.0517544  0.0151009  -3.427 0.000933 ***
weight       0.0008373  0.0001302   6.431 6.60e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[...]
```

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?

Paper 1, Section II

13J Statistical Modelling

Consider a generalised linear model with parameter $\beta^\top$ partitioned as $(\beta_0^\top, \beta_1^\top)$, where β_0 has p_0 components and β_1 has $p - p_0$ components, and consider testing $H_0 : \beta_1 = 0$ 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 $\eta = (\eta_1, ..., \eta_n)^T$ given by $\eta_i = x_i^\top \beta$, $i = 1, ..., n$, where $x_{i1} = 1$ for every i. Derive the deviance for this model, and argue that it may be approximated by Pearson's χ^2 statistic.

Paper 4, Section II

13J Statistical Modelling

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 ith 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 β. Find the log-likelihood for β, and the likelihood equations that must be solved to find the maximum likelihood estimator $\hat{\beta}$ of β.

Compute the first and second derivatives of the log-likelihood for β, 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 λ 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 β is partitioned into k sets as follows: $\beta^\top = (\beta_1^\top \ \cdots \ \beta_k^\top)$. What does it mean for a pair of sets β_i, β_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 β_0, β_1 and β_2 to be mutually orthogonal.

If β_0, β_1 and β_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 $\hat{\beta}_2$?

Paper 4, Section I

5I Statistical Modelling

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):

```
> 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.

```
> 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:

```
> 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?

```
> sqrt(sum(resid(fit)^2)/fit$df)
```

Paper 1, Section II

13I Statistical Modelling

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:

```
> cancer
    age malignant survive die
1   <50        no      77  10
2   <50       yes      51  13
3 50-69        no      51  11
4 50-69       yes      38  20
5   70+        no       7   3
6   70+       yes       6   3
```

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

```
> 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.

```
> 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:

```
> 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 $\varepsilon \sim N_n(0, \sigma^2 I)$ and X is an $n \times p$ matrix of full rank $p < n$. Find the form of the maximum likelihood estimator $\hat{\beta}$ of β, and derive its distribution assuming that σ^2 is known.

Assuming the prior $\pi(\beta, \sigma^2) \propto \sigma^{-2}$ find the joint posterior of (β, σ^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.

1/I/5J **Statistical Modelling**

Consider the following Binomial generalized linear model for data $y_1, \dots, y_n$, with logit link function. The data $y_1, \dots, y_n$ are regarded as observed values of independent random variables $Y_1, \dots, Y_n$, where

$$Y_i \sim \mathrm{Bin}(1, \mu_i), \quad \log \frac{\mu_i}{1-\mu_i} = \beta^\top x_i, \quad i = 1, \dots, n,$$

where β is an unknown p-dimensional parameter, and where $x_1, \dots, x_n$ are known p-dimensional explanatory variables. Write down the likelihood function for $y = (y_1, \dots, y_n)$ 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 $x_1^\top, \dots, x_n^\top$, and where $\hat{\mu} = (\hat{\mu}_1, \dots, \hat{\mu}_n)$, 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.

1/II/13J **Statistical Modelling**

Consider performing a two-way analysis of variance (ANOVA) on the following data:

```
> Y[,,1]             Y[,,2]                  Y[,,3]
      [,1] [,2]          [,1]    [,2]             [,1]   [,2]
 [1,] 2.72 6.66     [1,] -5.780  1.7200     [1,] -2.2900 0.158
 [2,] 4.88 5.98     [2,] -4.600  1.9800     [2,] -3.1000 1.190
 [3,] 3.49 8.81     [3,] -1.460  2.1500     [3,] -2.6300 1.190
 [4,] 2.03 6.26     [4,] -1.780  0.7090     [4,] -0.2400 1.470
 [5,] 2.39 8.50     [5,] -2.610 -0.5120     [5,]  0.0637 2.110
   .    .    .        .     .       .         .      .       .
   .    .    .        .     .       .         .      .       .
   .    .    .        .     .       .         .      .       .
```

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.

```
> 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
          Df Sum Sq Mean Sq F value    Pr(>F)
a          1 178.98  178.98  129.83 3.272e-16 ***
b          2 494.39  247.19  179.31 < 2.2e-16 ***
Residuals 56  77.20    1.38
```

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.

```
> fit2 <- lm(y ~ a * b)
> anova(fit2)
Response: y
          Df Sum Sq Mean Sq  F value    Pr(>F)
a          1 178.98  178.98 125.6367 1.033e-15 ***
b          2 494.39  247.19 173.5241 < 2.2e-16 ***
a:b        2   0.27    0.14   0.0963    0.9084
Residuals 54  76.93    1.42
```

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

```
> 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)
```

2/I/5J **Statistical Modelling**

Suppose that we want to estimate the angles α, β and γ (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 σ^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 α 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, β is an unknown p-dimensional parameter, and $\varepsilon \sim N_n(0, \sigma^2 I)$, with σ^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 β. Write in terms of X an expression for the projection matrix $P = (p_{ij} : 1 \leqslant i, j \leqslant n)$ 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 ε is to check whether the studentized fitted residuals

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

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

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

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?

4/I/5J **Statistical Modelling**

A long-term agricultural experiment had $n = 90$ grassland plots, each 25m × 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.

```
> 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.

```
> 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?

```
> 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?

```
> mu.hat <- exp(predict(fit2))

> -2 * (sum(dpois(Species, mu.hat, log = TRUE)) - sum(dpois(Species,

+     Species, log = TRUE)))
```

4/II/13J **Statistical Modelling**

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

$$Y_i \sim \mathrm{ED}(\mu_i, \sigma_i^2), \quad g(\mu_i) = x_i^\top \beta, \quad \sigma_i^2 = \sigma^2 a_i,$$

Here, g is a given link function, β and σ^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 θ.]

[You may use without proof the facts that, for such a random variable Y,

$$E(Y) = K'(\theta(\mu)), \quad \mathrm{var}(Y) = \sigma^2 K''(\theta(\mu)) \equiv \sigma^2 V(\mu).]$$

Show that the score vector and Fisher information matrix have entries:

$$U_j(\beta) = \sum_{i=1}^n \frac{(y_i - \mu_i) x_{ij}}{\sigma_i^2 V(\mu_i) g'(\mu_i)}, \qquad j = 1, \ldots, p,$$

and

$$i_{jk}(\beta) = \sum_{i=1}^n \frac{x_{ij} x_{ik}}{\sigma_i^2 V(\mu_i) (g'(\mu_i))^2}, \qquad j, k = 1, \ldots, p.$$

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 β.

1/I/5I Statistical Modelling

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.

number	fuel	strain
82	0	1
58	0	0
45	0.5	1
27	0.5	0
29	0.75	1
15	1.25	1
6	1.25	1
8	1.5	0
4	1.75	0
.	.	.
.	.	.

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.

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

> summary(fit1)

Coefficients:

             Estimate Std. Error z value Pr(>|z|)

(Intercept)  4.14443    0.05101  81.252  < 2e-16 ***

fuel        -1.47253    0.07007 -21.015  < 2e-16 ***

strain       0.33667    0.06704   5.022 5.11e-07 ***

fuel:strain -0.12534    0.09385  -1.336    0.182
```

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 the summary from the fit of the one above.

```
> fit2 <- glm(number ~ fuel + strain, family = poisson)

> fit3 <- glm(number ~ fuel, family = poisson)
```

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.

```
> 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 $\mu_i = x_i^{\mathrm{T}}\beta$. Here x_i is a vector of known explanatory variables and β is a vector of unknown regression coefficients.

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

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

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

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

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

Briefly comment on why the Laplace distribution cannot be written in exponential dispersion family form.

3/I/5I **Statistical Modelling**

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 β an unknown (possibly vector) parameter. Write $\hat{\beta}$ for the maximum likelihood estimator (m.l.e.) of β 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 β^* for the m.l.e. of β 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.

4/I/5I **Statistical Modelling**

Consider the normal linear model $Y = X\beta + \varepsilon$ in vector notation, where

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_1^{\mathrm{T}} \\ \vdots \\ x_n^{\mathrm{T}} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \quad \varepsilon_i \sim \text{i.i.d. } N(0, \sigma^2),$$

where $x_i^{\mathrm{T}} = (x_{i1}, \ldots, x_{ip})$ is known and X is of full rank ($p < n$). Give expressions for maximum likelihood estimators $\hat{\beta}$ and $\hat{\sigma}^2$ of β and σ^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

$$y^* = x^{*\mathrm{T}}\beta + \varepsilon^*, \qquad \text{where} \qquad \varepsilon^* \sim N(0, \sigma^2).$$

We suppose that x^* is known, and estimate y^* by $\tilde{y} = x^{*\mathrm{T}}\hat{\beta}$. State the distribution of

$$\frac{\tilde{y} - y^*}{\tilde{\sigma}\tau}, \qquad \text{where} \qquad \tilde{\sigma}^2 = \frac{n}{n-p}\hat{\sigma}^2 \qquad \text{and} \qquad \tau^2 = x^{*\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x^* + 1.$$

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

4/II/13I **Statistical Modelling**

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 $\mathbb{E}[Y]$ and $\mathrm{Var}[Y]$ in terms of α and λ. 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 $g(\mu_i) = \eta_i = x_i^{\mathrm{T}}\beta$, where $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is the vector of unknown regression coefficients and $x_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ is the vector of known values of the explanatory variables for the ith 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 β.

[*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 $\hat{\mu} = X\hat{\beta}$).

1/I/5I **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 β.

Define Cook's distance for the observation (x_i, Y_i), where x_i^T is the ith row of X. Give its interpretation in terms of confidence sets for β.

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* $\mathbb{P}(Z \leqslant 0.21) = 0.1$, $\mathbb{P}(Z \leqslant 1.39) = 0.5$ *and* $\mathbb{P}(Z \leqslant 4.61) = 0.9$.
(ii) *If* $Z \sim F_{2,48}$, *then* $\mathbb{P}(Z \leqslant 0.11) = 0.1$, $\mathbb{P}(Z \leqslant 0.70) = 0.5$ *and* $\mathbb{P}(Z \leqslant 2.42) = 0.9$.
(iii) *If* $Z \sim F_{48,2}$, *then* $\mathbb{P}(Z \leqslant 0.41) = 0.1$, $\mathbb{P}(Z \leqslant 1.42) = 0.5$ *and* $\mathbb{P}(Z \leqslant 9.47) = 0.9$.]

1/II/13I **Statistical Modelling**

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.

Age	AB	Hits	Average
19	10	2	0.200
20	92	29	0.315
21	136	37	0.272
22	123	40	0.325
23	317	95	0.300
24	432	139	0.322
25	457	172	0.376
26	540	204	0.378
27	406	128	0.315
28	522	205	0.393
29	529	200	0.378
30	359	134	0.373
31	495	184	0.372
32	540	192	0.356
33	536	173	0.323
34	499	172	0.345
35	518	186	0.359
36	534	199	0.373
37	457	156	0.341
38	459	138	0.301
39	365	105	0.288
40	72	13	0.181

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.

```
> 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:

```
> 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 β. Find the log-likelihood for β.

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

3/I/5I **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

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

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.

Year	Storms	Temp
1944	11	0.165
1945	11	0.080
1946	6	0.000
1947	9	-0.024
⋮	⋮	⋮
2001	15	0.592
2002	12	0.627
2003	16	0.608
2004	15	0.546

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 β is an unknown vector and $\epsilon \sim N_n(0, \sigma^2 I)$. Derive an expression for the maximum likelihood estimator $\hat{\beta}$ of β, and write down its distribution.

Find also the maximum likelihood estimator $\hat{\sigma}^2$ of σ^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* $n - p$.]

1/I/5I **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 $\mathbb{E}(Y_i) = \mu_i$ 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 β is an unknown p-dimensional parameter. Show that $\mathbb{E}(Y_i) = b'(\theta_i)$ and that, if $\ell(\beta, \phi)$ is the log-likelihood function from the observations $(y_1, \ldots, y_n)$, then

$$\frac{\partial \ell(\beta, \phi)}{\partial \beta} = \sum_1^n \frac{(y_i - \mu_i)x_i}{g'(\mu_i)V_i} ,$$

where V_i is to be defined.

1/II/13I **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'.

Algarve	8.00	0.50	3.50	3.00	4.00	100.00
CostaDelSol	6.95	1.30	4.10	12.30	4.10	130.85
Majorca	10.25	1.45	5.35	6.15	3.30	122.20
Tenerife	12.30	1.25	4.90	3.70	2.90	130.85
Florida	15.60	1.90	5.05	5.00	2.50	114.00
Tunisia	10.90	1.40	5.45	1.90	2.75	218.10
Cyprus	11.60	1.20	5.95	3.00	3.60	149.45
Turkey	6.50	1.05	6.50	4.90	2.85	263.00
Corfu	5.20	1.05	3.75	4.20	2.50	137.60
Sorrento	7.70	1.40	6.30	8.75	4.75	215.40
Malta	11.20	0.70	4.55	8.00	4.80	87.85
Rhodes	6.30	1.05	5.20	3.15	2.70	261.30
Sicily	13.25	1.75	4.20	7.00	3.85	174.40
Madeira	10.25	0.70	5.10	6.85	6.85	153.70

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.

```
> 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.

```
> 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, ϵ 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 σ^2.

(a) Let $\hat{\beta}$ be the least-squares estimator of β. 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.

```
> 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
```

4/II/13I **Statistical Modelling**

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

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

where

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

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

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

(ii) Find the equation for $\hat{\beta}$, the maximum likelihood estimator of β, 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 ν, where a, b, c are defined by

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