STATISTICAL MODELLING Part IIC / Michaelmas 2024
Example Sheet 4 (of 4)

1. Show that the log-likelihood for binomial regression with data $(y_1, x_1), \ldots, (y_n, x_n) \in \{0,1\} \times \mathbb{R}^p$ when the response is binary and the canonical link function is used can be written as
$$-\sum_{i=1}^{n} \log(1 + \exp(-\tilde{Y}_i X_i^T \beta)),$$
where $\tilde{Y}_i = 2Y_i - 1$.

2. Let $Y_1, \ldots, Y_n$ be independent with $Y_i \sim N(\mu_i, \sigma^2)$ and $\mu_i = X_i^T \beta$, for $i = 1, \ldots, n$. Show that only one iteration of the Fisher scoring method is required to attain the maximum likelihood estimator $\hat{\beta}$, regardless of the initial values for the algorithm. What feature of the log-likelihood function ensures that this is the case?

3. Let the design matrix X have i^{th} row X_i^T for $i = 1, \ldots, n$. Consider the generalized linear model for data $(X_1^T, Y_1), \ldots, (X_n^T, Y_n)$ with link function $g(\cdot)$ and dispersion parameter $\sigma_i^2 = \sigma^2 / w_i$ for observation i, where $w_1, \ldots, w_n$ are given data weights.

   (a) Use the chain rule to show that the score equations for β may be written as
   $$\sum_{i=1}^{n} \frac{(Y_i - \mu_i) X_{ir}}{\mathrm{Var}_{\beta,\sigma^2}(Y_i) \cdot g'(\mu_i)} = 0, \quad r = 1, \ldots, p,$$
   where $\mathrm{Var}_{\beta,\sigma^2}(Y_i)$ is the variance of Y_i in the exponential dispersion family with mean parameter $\mu_i = g^{-1}(X_i^T \beta)$ and dispersion parameter σ_i^2.

   (b) Show that the Fisher information matrix for the parameters (β, σ^2) takes the block-diagonal form
   $$I(\beta, \sigma^2) = \begin{pmatrix} I_{\beta\beta}(\beta, \sigma^2) & 0 \\ 0 & I_{\sigma^2\sigma^2}(\beta, \sigma^2) \end{pmatrix},$$
   where $I_{\beta\beta}(\beta, \sigma^2)$ is a $p \times p$ matrix. Show that $I_{\beta\beta}(\beta, \sigma^2)$ can be expressed as $\sigma^{-2} X^T W X$ where W is a diagonal matrix in which the i-th diagonal entry of W depends on w_i and μ_i. (You need not specify $I_{\sigma^2\sigma^2}(\beta, \sigma^2)$, and you may assume $\partial^2 \ell / \partial \beta_j \partial \sigma^2 = \partial^2 \ell / \partial \sigma^2 \partial \beta_j$ for all j).

   (c) Suppose the link function g is canonical. How do the expressions in (a) and (b) simplify in this case? Show that the observed information matrix of β (the (β, β)-block in the Hessian matrix of the log-likelihood function) is non-random given X, and use this to conclude that the Newton-Raphson algorithm is equivalent to the Fisher scoring algorithm when the link function is canonical.

4. Suppose that for some strictly increasing function f, we have
$$Y_i^* = f(X_i^T \beta^* + \varepsilon_i), \qquad i = 1, \ldots, n,$$
where $\varepsilon \sim N_n(0, \sigma^2 I)$, and the X_i are covariates in $\mathbb{R}^p$ with first component equal to 1. Suppose that for some constant c, we observe
$$Y_i := \mathbb{1}_{\{Y_i^* > c\}}.$$
Show that $Y_1, \ldots, Y_n$ are independent and
$$\mathbb{E}(Y_i) = \Phi(X_i^T \beta)$$
for some β that you should specify, where Φ is the c.d.f. of the standard normal distribution. This is often called the probit regression model.

5. Load the `Cycling` dataset using the `R` code below:

```
> file_path <- "http://www.statslab.cam.ac.uk/~rds37/teaching/statistical_modelling/"
> Cycling <- read.csv(paste(file_path, "Cycling.csv", sep =""))
> str(Cycling) # You can see which variables are factors and how many levels they have
```

These data were collected by Prof. Ian Walker from the University of Bath. He used an instrumented bicycle to gather proximity data from overtaking motorists when cycling. Recorded in the data is the distance from kerb when a car passed, the type of road that he was cycling on, which city he was in, whether or not a helmet was being worn and other variables. The goal of this data collection was to determine whether wearing a cycle helmet affects how close motorists pass by cyclists.

(a) Fit a normal linear model to the data with `passing.distance` as the response and all other variables as explanatory variables. Under what conditions on the distribution of the data will this model be correct? Probe these conditions by examining the diagnostic plots using `plot.lm`.

(b) Now fit a probit regression where the response is 1 if the passing distance is less than 1 metre, 0 otherwise (the family argument of `glm` will need to be given as `binomial(link = probit)`—this is how non-canonical links are specified). Under what conditions on the distribution of the original data will the probit model be correct? Compare them to your answer in part (a).

6. For an exponential family of distributions $\{f(\cdot;\theta) \mid \theta \in \Theta \subseteq \mathbb{R}\}$, the deviance of $\theta_1 \in \Theta$ from $\theta_2 \in \Theta$ is defined as $D(\theta_1,\theta_2) = 2\mathbb{E}_{\theta_1}\{\log f(Y;\theta_1) - \log f(Y;\theta_2)\}$, where $\mathbb{E}_{\theta_1}$ means the expectation is taken over $Y \sim f(\cdot;\theta_1)$. With an abuse of notation, we often use the mean value parametrisation $\mu_1 = \mathbb{E}_{\theta_1}(Y)$, $\mu_2 = \mathbb{E}_{\theta_2}(Y)$ and write $D(\theta_1,\theta_2)$ as $D(\mu_1,\mu_2)$.

   (a) Show that $D(\theta_1,\theta_2) = 2\{(\theta_1-\theta_2)\mu_1 - K(\theta_1) + K(\theta_2)\}$, where $K(\cdot)$ is the cumulant function of the exponential family. How does this formula change when we include a dispersion parameter in the distribution?

   (b) When $\mu = (\mu_1,\ldots,\mu_n)$ and $\tilde{\mu} = (\tilde{\mu}_1,\ldots,\tilde{\mu}_n)$ are vectors, the total deviance of μ from $\tilde{\mu}$ is defined as $D^{(n)}(\mu,\tilde{\mu}) = \sum_{i=1}^n D(\mu_i,\tilde{\mu}_i)$. Consider the normal linear model: let $Y_1,\ldots,Y_n$ be independent with $Y_i \sim N(\mu_i, 1)$ and $\mu_i = X_i^T\beta$, for $i = 1,\ldots,n$. Show that the total deviance of Y from the maximum likelihood estimator of μ is equal to the residual sum of squares.

   (c) Consider a generalized linear model as set up in Question 3 with the canonical link function. Show that the maximum likelihood estimator of β is given by $\hat{\beta} = \arg\min_\beta D^{(n)}(Y,\mu)$ where μ_i is the mean parameter corresponding to the natural parameter $\theta_i = X_i^T\beta$.

   (d) In the same setting, consider a partitioning $X = (X_0, X_1)$, $\beta = (\beta_0^T, \beta_1^T)^T$ where $X_0 \in \mathbb{R}^{n\times p_0}$, $X_1 \in \mathbb{R}^{n\times(p-p_0)}$, $\beta_0 \in \mathbb{R}^{p_0}$, and $\beta_1 \in \mathbb{R}^{p-p_0}$. Let $\hat{\mu}$ and $\hat{\mu}_0$ be the maximum likelihood estimator of μ under the model $\theta = X\beta$ and $\theta = X_0\beta$. Show that $D_+(Y,\hat{\mu}_0) = D_+(Y,\hat{\mu}) + D_+(\hat{\mu},\hat{\mu}_0)$.
   Hint: By using the score equation for $\hat{\beta}$, show that $D_+(\hat{\mu},\hat{\mu}_0) = 2\{l(\hat{\mu}) - l(\hat{\mu}_0)\}$ where $l(\mu)$ is the log-likelihood function for mean μ. Then consider a special choice of X.

   (e) Do the conclusions in (c) and (d) still hold if we use a non-canonical link function?

7. You see below the results of using `glm` to analyse data from Agresti (1996) on tennis matches between 5 top women tennis players (1989–90). We let Y_{ij} be the number of wins of player i against player j, and let n_{ij} be the total number of matches of i against j, for $1 \le i < j \le 5$. Thus we have 10 observations, which we will assume are realisations of independent binomial random variables Y_{ij} with

$$Y_{ij} \sim \mathrm{Bin}(n_{ij}, \mu_{ij})$$

and

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \alpha_i - \alpha_j.$$

This is known as the Bradley-Terry model and the parameter α_i represents the quality of player i. The data are tabulated in R as follows

```
wins tot sel graf saba navr sanc
   2   5   1   -1    0    0    0
   1   1   1    0   -1    0    0
   3   6   1    0    0   -1    0
   2   2   1    0    0    0   -1
   6   9   0    1   -1    0    0
   3   3   0    1    0   -1    0
   7   8   0    1    0    0   -1
   1   3   0    0    1   -1    0
   3   5   0    0    1    0   -1
   3   4   0    0    0    1   -1
```

Thus for example, the first row tells us that Seles played Graf five times and won on two occasions. We perform the following R commands (the output has been slightly abbreviated).

```
> fit <- glm(wins/tot ~ sel + graf + saba + navr - 1, binomial, weights=tot)
> summary(fit, correlation=TRUE)
Coefficients:
     Estimate Std. Error z value Pr(>|z|)
sel    1.5331     0.7871   1.948  0.05142 .
graf   1.9328     0.6784   2.849  0.00438 **
saba   0.7309     0.6771   1.079  0.28042
navr   1.0875     0.7237   1.503  0.13289
---

    Null deviance: 16.1882  on 10  degrees of freedom
Residual deviance:  4.6493  on  6  degrees of freedom

Correlation of Coefficients:
     sel  graf saba
graf 0.59
saba 0.46 0.60
navr 0.63 0.54 0.49
```

(a) What is the meaning of the `-1` in the model formula and why do you think it was included?

(b) Why is Sánchez (`sanc`) not included in the model formula?

(c) Can we confidently (at the 5% level) say that Graf is better than Sánchez?

(d) Can we confidently (at the 5% level) say that Graf is better than Seles? [Use the correlation matrix and a calculator or R, writing out your calculations. $\mathbb{P}(Z \le 1.64) \approx 0.95$ when $Z \sim N(0,1)$.]

(e) What is your estimate of the probability that Sabatini (`saba`) beats Sánchez, in a single match? Give a 95% confidence interval for this probability. [Use a calculator or R. $\mathbb{P}(Z \le 1.96) \approx 0.975$ when $Z \sim N(0,1)$]

8. (Long Tripos 2005/4/13I)

(a) Suppose that $Y_1, \dots, Y_n$ are independent random variables, and that Y_1 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

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

and $x_1, \dots, x_n$ are given p-dimensional vectors, and ν is known.
Show that $\mathbb{E}(Y_i) = \mu_i$ and that $\mathrm{var}(Y_i) = \mu_i^2/\nu$.

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

(c) 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

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

9. We wish to study how various explanatory variables may contribute to the development of asthma in children. One way to do this would be to randomly select n newborn babies and then study them for the first 5 years, measuring the values of the relevant covariates and noting down whether they develop asthma or not within the study period. However, this sort of experiment may be too expensive to carry out, and instead, we acquire the medical records of some children who developed asthma within the first five years of their life, and some children who did not. Luckily the medical records contain all the covariates we intended to measure.

We can imagine that the records we obtain are a sample from a large collection of data

$$(Y_1, X_1), \ldots, (Y_N, X_N) \in \{0, 1\} \times \mathbb{R}^p,$$

where each Y_i indicates the development of asthma and can be considered as a realisation of a Bernoulli random variable Y_i with $\pi_i := \mathbb{P}(Y_i = 1) \in (0, 1)$,

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \alpha + X_i^T \beta,$$

and all the Y_i are independent. Let Z_i indicate whether (Y_i, X_i) is in our sample: 1 if it is, 0 if not. Suppose that for all $i = 1, \ldots, N$,

$$\mathbb{P}(Z_i = 1 \mid Y_i = 1) = p_1, \quad \text{and} \quad \mathbb{P}(Z_i = 1 \mid Y_i = 0) = p_0,$$

where $p_1, p_0 > 0$ are unknown, and further that the (Y_i, Z_i) are all independent. Show that

$$\frac{\mathbb{P}(Y_i = 1 \mid Z_i = 1)}{1 - \mathbb{P}(Y_i = 1 \mid Z_i = 1)} = \frac{p_1}{p_0} \exp(\alpha + X_i^T \beta).$$

Conclude that it is possible to estimate β from our medical records data, but not α.

10. Agresti (1990) gives the table below, relating mothers' education to fathers' education for a sample of eminent black Americans (defined as persons having a biographical sketch in the publication *Who's Who Among Black Americans*).

Mother's	Father's education			
education	1	2	3	4
1	81	3	9	11
2	14	8	9	6
3	43	7	43	18
4	21	6	24	87

The categories 1–4 indicate increasing levels of education. We wish to model the entries Y_{ij} as components of a multinomial random vector with corresponding probabilities p_{ij} where

$$p_{ij} = \begin{cases} \eta\phi_i + (1 - \eta)\alpha_i\beta_j, & \text{for } i = j \\ (1 - \eta)\alpha_i\beta_j, & \text{for } i \neq j, \end{cases}$$

and

$$0 \leq \eta < 1,$$
$$\alpha_i, \beta_j > 0, \ \phi_i \geq 0,$$
$$\sum_i \phi_i = \sum_i \alpha_i = \sum_j \beta_j = 1.$$

Give an interpretation of this model. Why might we expect that $\eta > 0$ for our data?

Now model the Y_{ij} as independent Poisson random variables with means $\mu_{ij} = \exp(\alpha + x_{ij}^T\theta)$. We wish to choose the covariates x_{ij} such that if we maximise the Poisson likelihood, with non-negativity constraints on some components of θ, we obtain an estimate $\hat{\theta}$ which yields fitted values $\hat{\mu}_{ij} = \exp(\hat{\alpha} + x_{ij}^T\hat{\theta})$ equal to those from the multinomial model above. Describe how the x_{ij} can be chosen, and what non-negativity constrains should be applied.