STATISTICAL MODELLING Part IIC / Michaelmas 2025
Example Sheet 4 (of 4)

In questions that follow, a generalized linear model (GLM) assumes $Y_i \mid X_i \overset{ind.}{\sim} f(y;\theta_i,\sigma_i^2), i = 1,\ldots,n$, where $f(y;\theta,\sigma^2) = e^{\frac{1}{\sigma^2}\{\theta y - K(\theta)\}} f_0(y;\sigma^2)$ is the density function of an exponential dispersion family with natural parameter $\theta \in \mathbb{R}$, mean parameter $\mu = \mu(\theta) = K'(\theta)$, and cumulant function $K(\theta)$; θ_i is determined by $g(\mu(\theta_i)) = X_i^T\beta$ where the link function g is strictly increasing and twice differentiable; $\sigma_i^2 = \sigma^2/w_i$ where $w_1,\ldots,w_n$ are known.

1. Show that the normal linear model is an instance of GLM. How many iterations of Fisher scoring is requried to find the maximum likelihood estimator $\hat{\beta}$ for the normal linear model? Does your answer depend on the initial values for the algorithm?

2. Let $D(\mu_1,\mu_2)$ denote the deviance between two distributions in the same one-parameter exponential (dispersion) family with mean parameters $\mu_1 \in \mathbb{R}$ and $\mu_2 \in R$ (with dispersion parameter = 1). For vectors $\mu_1 = (\mu_{11},\ldots,\mu_{1n})$ $\mu_2 = (\mu_{21},\ldots,\mu_{2n})$, the total deviance between μ_1 and μ_2 is defined as $D^{(n)}(\mu_1,\mu_2) = \sum_{i=1}^n D(\mu_{1i},\mu_{2i})$.

   (a) Show that $D(\mu_1,\mu_2) = (\mu_1-\mu_2)^2$ for the normal family, then use this to show that the total deviance between Y and the maximum likelihood estimator of μ in the normal linear model is equal to the residual sum of sqaures.

   (b) Show that in a GLM, the maximum likelihood estimator of β is given by

   $$\hat{\beta} = \arg\min_{\beta\in\mathbb{R}^p} D^{(n)}(Y,\mu)$$

   where $\mu = (\mu_1,\ldots,\mu_n)$ and $\mu_i = \mu(X_i^T\beta), i = 1,\ldots,n$.

   (c) 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 in a canonical GLM (dispersion parameter taken as 1) with linear predictors $\eta = X\beta$ and $\eta = X_0\beta_0$. Show that

   $$D^{(n)}(Y,\hat{\mu}_0) = D^{(n)}(Y,\hat{\mu}) + D_+(\hat{\mu},\hat{\mu}_0).$$

   Hint: Show that $D^{(n)}(\hat{\mu},\hat{\mu}_0) = 2\{l(\hat{\beta}) - l(\hat{\beta}_0)\}$ where l is the log-likelihood function.

3. 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 $\pi_i := \mathsf{P}(Y_i = 1) \in (0,1)$,

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

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 $i = 1, \ldots, N$,

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

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

$$\frac{\mathsf{P}(Y_i = 1 \mid Z_i = 1)}{\mathsf{P}(Y_i = 0 \mid Z_i = 1)} = \frac{p_1}{p_0} \exp(\alpha + X_i^T \beta).$$

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

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 fixed 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

$$\mathsf{E}(Y_i) = \Phi(X_i^T \beta)$$

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. 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 $1 \leq i < j \leq 5$. Thus we have 10 observations, which we will assume are realisations of independent binomial random variables Y_{ij} with

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

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. $\mathsf{P}(Z \leq 1.64) \approx 0.95$ 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. $\mathsf{P}(Z \leq 1.96) \approx 0.975$ when $Z \sim N(0,1)$]

7. Let $Y_i \sim \text{Poisson}(\mu_i), i = 1, \ldots, n$ be independent and $Y_+ = \sum_{i=1}^n Y_i$. Show that

$$(Y_1, \ldots, Y_n) \mid Y_+ \sim \text{Multinomial}(Y_+, (\pi_1, \ldots, \pi_n)),$$

where $\pi_i = \mu_i/\mu_+, i = 1, \ldots, n$ and $\mu_+ = \sum_{i=1}^n \mu_i$.

8. 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 education	Father's 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

$$\begin{gathered} 0 \leq \eta < 1, \\ \alpha_i, \beta_j > 0,\ \phi_i \geq 0, \\ \sum_i \phi_i = \sum_i \alpha_i = \sum_j \beta_j = 1. \end{gathered}$$

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. When maximising the likelihood, why might we expect that the positivity constraints will be met even if we don't enforce them?

9. Suppose $Y \sim \text{Bernoulli}(\pi)$ and

$$X \mid Y = 0 \sim \text{N}(\mu_0, \Sigma_0), \quad X \mid Y = 1 \sim \text{N}(\mu_1, \Sigma_1), \quad \Sigma_0 \neq \Sigma_1.$$

Find the Bayes classifier under this model and justify why it is often referred to as *quadratic discriminant analysis*.

10. Show that the log-likelihood for logistic regression (binomial GLM with canonical link) with data $(Y_1, X_1), \ldots, (Y_n, X_n) \in \{0, 1\} \times \mathbb{R}^p$ can be written as

$$-\frac{1}{C} \sum_{i=1}^{n} \phi((2Y_i - 1)X_i^T \beta),$$

where $\phi(s) = \log_2(1 + e^{-s})$ is the "logistic loss" and C is a constant that you must specify. Plot the function ϕ using R and comment on how it compares with $\phi_{\text{zero-one}}(s) = 1_{\{s \leq 0\}}$ and $\phi_{\text{hinge}}(s) = \max\{1 - s, 0\}$.

11. Consider the following two generalized linear models:

$$g(\mathsf{E}(Y \mid X, Z)) = \beta_0 + \beta_1 X + \beta_2 Z,$$
$$g(\mathsf{E}(Y \mid X)) = \alpha_0 + \alpha_1 X.$$

Suppose the first model is correctly specified and $X \in \{0, 1\}$ is binary (so the second model is saturated). We say the link function g is *collapsible* if $\beta_1 = \alpha_1$ when Z is independent of X.

(a) Show that the identity link $g(\mu) = \mu$ and the log link $g(\mu) = \log(\mu)$ are collapsible. *Hint: Write* $\mathsf{E}(Y \mid X = 1, Z) = h_{\beta_1}(\mathsf{E}(Y \mid X = 0, Z))$ *for a characteristic transformation* h_{β_1} *that you must specify, then show that* g *is collapsible when* h_{β_1} *is an affine function.*

(b) Consider the logistic link $g(\mu) = \log(\mu/(1 - \mu))$. Show that the coefficient of X is always attenuated in the second model in the sense that $|\alpha_1| \leq |\beta_1|$. *Hint: Show that the characteristic transformation* h_{β_1} *for the logistic link is concave if* $\beta_1 > 0$ *and convex if* $\beta_1 < 0$.

12. Consider the binary classification problem with the zero-one loss $l(y, \delta(x)) = 1_{\{y \neq \delta(x)\}}$, so the risk of a classifier δ is simply the misclassification probability $R(\delta) = \Pr(Y \neq \delta(X))$.

(a) Show that the risk of the Bayes classifier is $R^* = \mathsf{E}\{\min(\pi(X), 1 - \pi(X)\}$ where $\pi(X) = \Pr(Y = 1 \mid X)$.

(b) Let $R_{k,n}$ denote the risk of the k-nearest neighbour classifier trained with a sample of n i.i.d. data points. Suppose k is odd and π is a continuous function. Show that as $n \to \infty$ we have $R_{k,n} \to R_k$ for some limiting risk that satisfies

$$R^* \leq \cdots \leq R_3 \leq R_1 \leq 2R^*.$$

(c) Describe how you expect $R_{k,n}$ to behave as a function of n and k. Explain why you expect such behaviour and use R simulations to check whether your intuition is correct.