STATISTICAL MODELLING Part IIC / Michaelmas 2022
Example Sheet 3 (of 4)

1. Look at the `cabbages` data in the `library(MASS)` package (use `?cabbages` to find out about the dataset). Investigate whether the planting date has a significant effect on the weight of the cabbage head. Write out the models you have fitted and explain any conclusions you come to.

2. Download the Cambridge Colleges data with

```
> path <- "http://www.statslab.cam.ac.uk/~rds37/teaching/statistical_modelling/"
> Colleges <- read.csv(file.path(path, "Colleges.csv"))
```

   Fit a linear model with the percentage of firsts as the response and the logarithm of the wine budget as a covariate. Pick a college (possibly your own) and test whether it is an outlier. Looking at a plot of the data, what appears to be the most outlying college? Note you can add the names of the colleges to the plot by issuing

```
text(log(WineBudget), PercFirsts, rownames(Colleges), cex=0.6, pos=3)
```

   after plotting the data (provided the data frame `Colleges` is attached). What is the issue with using your test to now determine whether this college is an outlier?

3. Suppose $Y_1, \ldots, Y_n$ is an i.i.d. sample from $N(\mu, 1)$. What is the asymptotic distribution of the maximum likelihood estimator of $\mathbb{P}(Y_1 < 0)$?

4. Show the following families of distributions are (possibly multi-parameter) exponential families; all parameters are unknown unless noted otherwise. Then find the corresponding natural parameters, sufficient statistics, and cumulant functions.

   (a) The normal distribution, $\mathrm{N}(\mu, \sigma^2)$:

   $$f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, y \in \mathbb{R}.$$

   (b) The Gamma distribution, $\mathrm{Gamma}(\alpha, \beta)$:

   $$f(y; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}, \ y > 0.$$

   (c) The negative binomial distribution (number of failures until k successes are reached in Bernoulli trials), $\mathrm{NegBin}(k, p)$ with fixed k:

   $$f(y; p) = \binom{y+k-1}{y} p^k (1-p)^y, \ y = 0, 1, 2, \ldots$$

5. Let $Y \in \mathbb{R}$ be a random variable whose moment generating function is finite on an open interval containing zero. Show that the first three cumulants are $\kappa_1 = \mathbb{E}(Y)$, $\kappa_2 = \mathrm{Var}(Y)$, $\kappa_3 = \mathbb{E}(Y - \kappa_1)^3$, respectively.

6. Derive the mean and variance of the negative binomial distribution $\mathrm{NegBin}(k, p)$ in terms of k and p.

7. Suppose $\mu \sim \pi(\cdot)$ where $\pi(\cdot)$ is an unknown density function on $\mathbb{R}$. Suppose $Y \mid \mu \sim N(\mu, \sigma^2)$ and σ^2 is known. Derive Tweedie's formula

   $$\mathbb{E}(\mu \mid Y) = Y + \sigma^2 \cdot \frac{f'(Y)}{f(Y)},$$

   where $f(y) = \int \pi(\mu) f(y; \mu, \sigma^2) d\mu$ is the marginal density of Y, and $f(y; \mu, \sigma^2)$ is the density of $N(\mu, \sigma^2)$. *Hint: The posterior distribution of θ given Y is an exponential family.*

8. Suppose $Y_1, \ldots, Y_n$ is an i.i.d. sample from a regular exponential family with natural parameter θ (regular means that the natural parameter space is open), sufficient statistic $T(y) = y$, cumulant function $K(\cdot)$, mean μ, and variance V.

   (a) Show that the distribution of $\bar{Y} = \sum_{i=1}^n Y_i/n$ is in an exponential family with natural parameter $\theta^{(n)} = n\theta$ and cumulant function $K^{(n)}(\theta^{(n)}) = nK(\theta^{(n)}/n)$. What is the mean and variance of $\bar{Y}$? *Hint: What is the joint density of* $Y_1, \ldots, Y_n$*?*

   (b) The deviance of θ_1 from θ_2 is defined as

$$D(\theta_1, \theta_2) = 2\mathbb{E}_{\theta_1}\left\{\log \frac{f(Y;\theta_1)}{f(Y;\theta_2)}\right\}.$$

   Show that the deviance in the exponential family for $\bar{Y}$ of natural parameter $\theta_1^{(n)} = n\theta_1$ from $\theta_2^{(n)} = n\theta_2$ is $nD(\theta_1, \theta_2)$. Denote this as $D^{(n)}(\theta_1, \theta_2)$.

   (c) With an abuse of notation, we also denote $D(\theta_1, \theta_2)$ as $D(\mu_1, \mu_2)$ where μ_1 and μ_2 are the mean parameters corresponding to θ_1 and θ_2 in the exponential family. Similarly, $D^{(n)}(\mu_1, \mu_2) = nD(\mu_1, \mu_2)$. The deviance residual is defined as

$$R = \text{sign}(\bar{Y} - \mu) \cdot \sqrt{D^{(n)}(\bar{Y}, \mu)}.$$

   Show that $R^2 \xrightarrow{d} \chi_1^2$ as $n \to \infty$. You need not specify the regularity conditions for any asymptotic results used. *Hint: Use Wilk's theorem.*

9. Suppose $Y_1, \ldots, Y_n$ is an i.i.d. sample from an non-degenerate exponential dispersion family:

   (a) Compute the cumulant generating function of Y_1 and deduce expressions for the mean μ and variance V of Y_1.

   (b) Show that the MLE of μ is given by the sample mean $\bar{Y} = \sum_{i=1}^n Y_i/n$.

10. We say Y has the inverse Gaussian distribution with parameters ϕ and λ, and write $Y \sim IG(\phi, \lambda)$ if its density is

$$f_Y(y;\phi,\lambda) = \frac{\sqrt{\lambda}}{\sqrt{2\pi}y^{3/2}} e^{\sqrt{\lambda\phi}} \exp\Big\{-\frac{1}{2}\Big(\frac{\lambda}{y} + \phi y\Big)\Big\},$$

    $y \in (0, \infty)$, $\lambda \in (0, \infty)$, $\phi \in (0, \infty)$. Compute the cumulant generating function of Y, and hence find its mean and variance.

    Show that the family of inverse Gaussian densities above is an exponential dispersion family. Specify its natural and dispersion parameters, and the spaces they belong to. Find the mean function μ, mean space $\mathcal{M}$, canonical link function and variance function $V(\mu)$. *Hint: Find a reason to guess that* σ^2 *is a function of* λ *alone, then find* V *as a function of* ϕ *and* λ*, and derive the expression for* σ^2 *in terms of* λ*.*

11. Let $Y_1, \ldots, Y_n$ be independent Poisson random variables with mean μ. Compute the maximum likelihood estimator $\hat{\mu}$. By considering $n\hat{\mu}$, write down the distribution of $\hat{\mu}$ and deduce its asymptotic distribution directly. Verify that this asymptotic distribution agrees with that predicted by the general asymptotic theory for maximum likelihood estimators.

12. Let the design matrix X have i^{th} row X_i^T for $i = 1, \ldots, n$. Consider the generalised 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 likelihood equations for β may be written as

$$\sum_{i=1}^n \frac{(Y_i - \mu_i)X_{ir}}{\text{Var}_{\beta,\sigma^2}(Y_i) \cdot g'(\mu_i)} = 0, \quad r = 1, \ldots, p,$$

    where $\mu_i = g^{-1}(X_i^T\beta)$.

(b) Show that the Fisher information matrix for the parameters (β, σ^2) takes the form

$$i(\beta, \sigma^2) = \begin{pmatrix} i_{\beta\beta}(\beta, \sigma^2) & 0 \\ 0 & i_{\sigma^2\sigma^2}(\beta, \sigma^2) \end{pmatrix},$$

Show that $i_{\beta\beta}(\beta, \sigma^2)$ can be expressed as $\sigma^{-2} X^T W X$ where W is a diagonal matrix with

$$W_{ii} = \frac{1}{w_i V(\mu_i)\{g'(\mu_i)\}^2},$$

(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) How do the expressions in (a) and (b) simplify when $g(\mu_i)$ is the canonical link function?