STATISTICAL MODELLING Part IIC / Michaelmas 2023
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 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 be a real-valued 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. Use these to find the mean and variance of the negative binomial distribution $\mathrm{NegBin}(k, p)$ (in terms of k and p).

6. 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.*

7. Suppose $Y_1, \ldots, Y_n$ is an i.i.d. sample from a regular exponential family with natural parameter $\theta \in \Theta \subseteq \mathbb{R}$ (regular means Θ is open), sufficient statistic $T(y) = y$, cumulant function $K(\theta)$, mean parameter $\mu = \mu(\theta)$, and variance $V(\theta) > 0$.

(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) Show likelihood ratio statistic for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, after a monotone transformation, is given by $D^{(n)}(\hat{\theta}, \theta_0)$. What is the limiting distribution of this statistic when $n \to \infty$? Justify your answer using the approximation $D(\theta_1, \theta_2) \approx I^{(1)}(\theta_2)(\theta_1 - \theta_2)^2$ when $\theta_1 \approx \theta_2$, where $I^{(1)}(\theta_2)$ is the Fisher information of one observation from the the distribution $f(\cdot;\theta_2)$.

(d) 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

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

What is the limiting distribution of R when $n \to \infty$?

8. Suppose $Y_1, \ldots, Y_n$ is an i.i.d. sample from an exponential dispersion family:

$$f(y;\theta,\sigma^2) = e^{\{\theta y - K(\theta)\}/\sigma^2} f_0(y;\sigma^2),$$

where $\theta, \sigma \in \mathbb{R}$ and $f_0(y;\sigma^2)$ is some density function. Suppose the distribution is non-degenerate in the sense that $\text{Var}(Y_1) > 0$.

(a) Compute the cumulant generating function of Y_1 and use it to show that $\mathbb{E}(Y_1) = K'(\theta)$ and $\text{Var}(Y_1) = \sigma^2 K''(\theta)$.

(b) Show that the MLE of $\mathbb{E}(Y_1)$ is given by the sample mean $\bar{Y} = \sum_{i=1}^n Y_i/n$.

9. The probability density function of the Gamma distribution with shape parameter $\alpha > 0$ and rate parameter $\beta > 0$ is given by

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

Show that this is an exponential dispersion family by finding the natural and dispersion parameters. *Hint: It may be useful to know that the mean and variance of the Gamma distribution are* α/β *and* α/β^2*, respectively.*

10. 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.

11. Consider a generalised linear model with vector of responses $Y = (Y_1, \ldots, Y_n)^T$ and design matrix X with i^{th} row X_i^T. Write $\hat{\mu}_i = g^{-1}(X_i^T\hat{\beta})$ where $\hat{\beta}$ is the maximum likelihood estimate of the vector of regression coefficients. Show that if the link function g is the canonical link, the dispersion parameter $\sigma^2 = 1$, and the weight $w_i = 1$, then

$$X^T Y = X^T \hat{\mu}.$$

Conclude that if an intercept term is included in X, then

$$\sum_{i=1}^n \hat{\mu}_i = \sum_{i=1}^n Y_i.$$

12. In this question, we will compare deviance residuals with Pearson's residuals and explore Bartlett's correction (Bartlett's correction is not examinable but provides useful motivation to consider exponential families). Consider the distribution Gamma(α, β) in Question 9.

    (a) Show that if $\alpha > 0$ is fixed, this is an exponential family with natural parameter $\theta = -\beta$ and mean parameter $\mu = \alpha/\beta$.

    (b) Suppose α is a positive integer. Let $Y_1, \ldots, Y_\alpha \sim \text{Gamma}(1, \beta)$ be independent. Show that $\sum_{i=1}^{\alpha} Y_i \sim \text{Gamma}(\alpha, \beta)$. *Hint: A probability distribution is uniquely determined by its moment generating function.*

    (c) Read the next chunk of `R` code and explain what it does. Execute the code in your `R` console and plot the histogram of `Y`. What do you see from the histogram?

```
alpha <- 5
beta <- 1
mu <- alpha / beta
deviance.gamma <- function(mu1, mu2, alpha) {
  2 * alpha * (log(mu2 / mu1) + mu1 / mu2 - 1)
}
Y <- rgamma(10000, alpha, 1)
dev <- deviance.gamma(Y, mu, alpha)
resid.dev <- sign(Y - mu) * sqrt(dev)
resid.pearson <- (Y - mu) / sqrt(alpha/beta^2)
```

    (d) Compare the distribution of `resid.dev` and `resid.pearson`. Which one is closer to $\text{N}(0, 1)$? You may find the function `qqnorm` useful.

    (e) The skewness and kurtosis of a probability distribution are defined as, respective,

$$\gamma = \kappa_3/\kappa_2^{3/2} = \frac{\mathbb{E}\{(Y - E(Y))^3\}}{\text{Var}(Y)^{3/2}}, \quad \delta = \kappa_4/\kappa_2^2 = \frac{\mathbb{E}\{(Y - E(Y))^4\}}{\text{Var}(Y)^2},$$

    where κ_r is the rth cumulant of the distribution and Y is a random variable that follows that distribution. It is known that the skewness of kurtosis of the Gamma(α, β) distribution is given by

$$\gamma_\alpha = 2/\sqrt{\alpha}, \quad \delta_\alpha = 6/\alpha + 3.$$

    Estimate γ and δ using your `Y` and compare your estimates with the above theoretical values.

    (f) It is noted in the lectures that the deviance residual approximately follows a normal distribution even when the sample size α is relatively small. The mean of this normal distribution is roughly $-\gamma_1/(6\sqrt{\alpha})$ (see Theorem 1.4 in Bradley Efron, *Exponential Families in Theory and Practice*, CUP). Is this similar to what your `resid.dev` shows?