Homework 2

Due Thursday, 2/4/21 on Canvas.

"Enzyme.txt" contains the data set for problems 1 and 2. If $Z \in \mathbb{R}$ is a random variable, the notation $Z \sim (\mu, v)$ is such that $\mathbb{E}(Z) = \mu$ and $\text{Var}(Z) = v$.

1. In an enzyme kinetics study the velocity of a reaction (Y) is expected to be related to the concentration (X) as follows:

$$Y_i = \frac{\gamma_0 X_i}{\gamma_1 + X_i} + \epsilon_i, \quad \epsilon_i \overset{i.i.d}{\sim} (0, \sigma^2), \quad i = 1, \ldots, n = 18.$$

   (a) We must first obtain starting points for Gauss-Newton to be able to estimate γ_0 and γ_1. Observe that

$$1/\,\mathbb{E}(Y_i) = (1/X_i)\gamma_1/\gamma_0 + 1/\gamma_0.$$

   Use this to obtain starting points for Gauss-Newton.

   (b) Estimate γ_0 and γ_1 using the starting points obtained in part (a).

2. Refer to the analysis of the enzyme kinetics in problem 1:

   (a) Plot the estimated nonlinear regression function and data on the same graph. Does the fit appear to be adequate?

   (b) Plot the residuals against the fitted values and obtain the normal qq-plot. Comment on the fit of the model.

   (c) Assume that the fitted model is appropriate and that large sample inference can be employed. Report the test statistic and two-sided p-value of the test of $H_0 : \gamma_1 = 20$.

3. Refer to the analysis of the enzyme kinetics in problems 1 and 2. Perform a bootstrap with 1000 samples, and compute 95% percentile confidence intervals for γ_1. Is it close to the confidence interval based on the large sample theory?

4. Consider a random variable Y which, conditional on a known covariate X, has a Bernoulli distribution with

$$logit\,\{P\,(Y = 1 \mid X)\} = \beta_0 + \beta_1 X.$$

   Ideally, we draw a random sample from the population to fit the model, and then do inference for β_1. However, in some applications, the number of cases that $Y = 1$ is small. The enriched study then uses an enriched-sample instead of a random sample, where the probability of an individual being included in the study is greater if the outcome $Y = 1$. Let $Z = 1$ if an individual is included in the enriched study and $Z = 0$ otherwise. We denote that $P(Z = 1 \mid Y = 1) = \gamma_1, P(Z = 1 \mid Y = 0) = \gamma_0$, with $\gamma_1 > \gamma_0 > 0$, where individuals are selected ONLY based on Y and not on X.

(a) Show that

$$logit\left\{P\left(Y=1\mid X,Z=1\right)\right\}=\beta_0^*+\beta_1 X,$$

where $\beta_0^* = \beta_0 + \log(\gamma_1/\gamma_0)$.

(b) Can the estimated effect of X from an enriched study be used to infer the effect in the general population?

(c) Can the estimated probability of $Y = 1$ given $X = x_0$ from an enriched study be used to infer the probability in the general population?

5. (Exponential tilting) Let Y be a random variable and $M(\theta) = \mathbb{E}\left\{e^{\theta Y}\right\}$ be its moment generating function. When appropriate, we will assume that $M(\theta) < \infty$ for all $\theta \in (-\epsilon, +\epsilon)$ for some $\epsilon > 0$, which implies all of the moments of Y exist and $\mathbb{E}\left(Y^k\right) = M^{(k)}(0)$ for all $k = 0, 1, 2, \ldots$. Here you will derive some useful properties of exponential families, which are the building blocks of GLMs.

(a) Define $K(\theta) = \log\{M(\theta)\}$ to be the cumulant generating function. Show that $K'(0) = \mathbb{E}(Y)$ and $K''(0) = \mathrm{Var}(Y)$.

proper density fcn

(b) Let $f_0(y)$ be a density with respect to the usual Lebesgue measure, i.e. $f_0(y) \geq 0$ and $\int f_0(y)dy = 1$ (everything you will show applies to densities with respect to arbitrary measures; Lebesgue is assumed for simplicity). For the remainder of the problem, let $M(\theta) = \int e^{y\theta} f_0(y)dy$ be its moment generating function and $K(\theta) = \log\{M(\theta)\}$. We will assume that 0 lies in the interior of $R = \{\theta : M(\theta) < \infty\}$. Define a family of densities to be

$$f(y;\theta) \propto e^{\theta y} f_0(y), \quad \theta \in R.$$

What is the normalizing constant for $f(y;\theta)$ in terms of θ?

(c) Show that

$$\ell(y;\theta) = \log\{f(y;\theta)\} = h(y) + \theta y - K(\theta), \quad \theta \in R$$

for some function h that only depends on y.

(d) Now let Y have density $f(y;\theta)$. Show that the cumulant generating function of Y, $K_\theta(t)$, is such that

$$K_\theta(t) = K(\theta + t) - K(\theta).$$

Use this to show that $\mathbb{E}(Y) = K'(\theta)$ and $\mathrm{Var}(Y) = K''(\theta)$.

(e) Now suppose $Y_i \sim f\left(y; \boldsymbol{x}_i^T\boldsymbol{\beta}\right)$ for $i = 1, \ldots, n$ and some $\boldsymbol{x}_i, \boldsymbol{\beta} \in \mathbb{R}^p$. Under the assumption that the Y_i's are independent, let $g(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell\left(Y_i; \boldsymbol{x}_i^T\boldsymbol{\beta}\right)$ be the log likelihood. Use parts (c) and (d) to show that

(i) $g(\boldsymbol{\beta})$ is concave.

(ii) The MLE $\hat{\boldsymbol{\beta}}$ satisfies $\boldsymbol{X}^T\left\{\boldsymbol{Y} - \mathbb{E}_{\hat{\boldsymbol{\beta}}}(\boldsymbol{Y})\right\} = \boldsymbol{0}$, where $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}$, $\boldsymbol{Y} = (Y_1, \dots, Y_n)^T$ and $\mathbb{E}_{\hat{\boldsymbol{\beta}}}(\boldsymbol{Y})$ is the expectation of $\boldsymbol{Y}$ under the model $Y_i \sim f\left(y; \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}\right)$ for all $i = 1, \dots, n$.