STA 712 Homework 3

Due: Friday, September 22, 12:00pm (noon) on Canvas.

Instructions: Submit your work as a single PDF, or as two separate PDFs (one for Parts 1 and 2, and another for Part 3). Parts 1 and 2 should be created using LaTeX; see the course website for a homework template file and instructions on getting started with LaTeX and Overleaf. See the Overleaf guide on mathematical expressions to get started writing math in LaTeX. Part 3 should be created using R Markdown or Quarto, so that all code needed to reproduce your results is included in the knitted document.

1 Cumulants and cumulant generating functions

1. Let Y be a random variable, and recall that the *moment generating function* (MGF) of Y is given by
$$M(t) = \mathbb{E}[e^{tY}].$$
We call M the moment generating function because
$$\frac{d^k}{dt^k} M(t)\bigg|_{t=0} = \mathbb{E}[Y^k].$$
We also define the *cumulant generating function* (CGF): $C(t) = \log M(t)$.

   (a) Show that
$$\frac{d}{dt} C(t)\bigg|_{t=0} = \mathbb{E}[Y].$$

   (b) Show that
$$\frac{d^2}{dt^2} C(t)\bigg|_{t=0} = Var(Y).$$

2 GLMs and the canonical link function

2. Suppose we are interested in modeling a response variable Y_i, given explanatory variables X_i. We use the generalized linear model
$$\begin{aligned} Y_i &\sim EDM(\mu_i, \phi) \\ g(\mu_i) &= \beta^T X_i, \end{aligned}$$
where $f(Y_i; \theta, \phi) = a(Y_i, \phi) \exp\left\{\dfrac{Y_i\theta_i - \kappa(\theta_i)}{\phi}\right\}$, and g is the canonical link function (that is, $g(\mu_i) = \theta_i$, the canonical parameter). One reason the canonical link function is nice is that it makes Fisher scoring nice.

   (a) Show that the score function is $U(\beta) = \dfrac{X^T(Y - \boldsymbol{\mu})}{\phi}$, where X is the design matrix, $Y = (Y_1, ..., Y_n)^T$, and $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)^T$.

   (b) Show that the Fisher information is $\mathcal{I}(\beta) = \dfrac{X^T V X}{\phi}$, where $V = \text{diag}(V(\mu_1), ..., V(\mu_n))$, and $V(\mu_i) = Var(Y_i)/\phi$.

3 Practice with Poisson regression

Here we work with data on the number of articles published by biochemistry PhD students in the last three years of their PhD program. You can load the `articles` data into R by

```
library(foreign)
articles <- read.dta("http://www.stata-press.com/data/lf2/couart2.dta")
```

The `articles` dataset contains the following columns:

- `art`: articles published in last three years of Ph.D.
- `fem`: sex (recorded as male or female)
- `mar`: marital status (recorded as married or single)
- `kid5`: number of children under age six
- `phd`: prestige of Ph.D. program
- `ment`: articles published by their mentor in last three years

Research question: We are interested in estimating the relationship between prestige of the PhD program, and the number of articles published, after accounting for sex, marital status, children, and the productivity of their research mentor.

3. Here you will use Poisson regression to investigate this research question.

   (a) Write down a Poisson regression model that will allow you to answer the research question. Describe how you will use the model to answer the research question.

   (b) Fit your model from (a), and report the equation of the fitted model. Interpret any estimated coefficients relevant to the research question.

   (c) Model diagnostics for Poisson regression are similar to diagnostics for logistic and linear regression models.
   - Create quantile residual plots to check the shape assumption for quantitative variables (you may use the `qresid` function in the `statmod` package)
   - Calculate Cook's distance to check for any influential points (use a threshold of 0.5 or 1 to identify influential points)
   - Calculate variance inflation factors to check for multicollinearity (see the `vif` function in the `car` package, and use a threshold of 5 or 10 to identify high multicollinearity).

   (d) Address any violations to the model assumptions (transformations for shape violations; report results with and without influential points; and combine or remove columns for high multicollinearity). If you made any changes to your model, report and interpret your new fitted model here.

   (e) If all model assumptions are satisfied, inference for Poisson regression models is similar to inference for logistic regression models (Wald or likelihood ratio tests can be used, and are calculated the same way). Carry out a hypothesis test to investigate the research question. You should:
   - State the null and alternative hypotheses in terms of one or more βs
   - Calculate a test statistic and p-value
   - Make a conclusion in the context of the original question