STA2201H Winter 2020 Assignment 1

Due: 5pm, 31 January 2020

What to hand in: .Rmd file and the compiled pdf

How to hand in: Submit files via Quercus

1 Exponential family

The random variable Y belongs to the exponential family of distributions if its support does not depend upon any unknown parameters and its density or probability mass function takes the form

$$p(y|\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right)$$

Assume ϕ is known.

a) Show $E\left[\frac{dp}{d\theta}\right] = 0$ and $E\left[\frac{d^2p}{d\theta^2}\right] = 0$
b) Using a) show $E[Y] = b'(\theta)$ and $Var(Y) = \phi b''(\theta)$
c) Denote $\log p(y|\theta,\phi)$ as $\ell(\theta)$. Using b) show $E\left[\ell(\theta)\right] = 0$ and $Var\left[\ell(\theta)\right] = \phi^{-1}b''(\theta)$.

2 Overdispersion

Suppose that the conditional distribution of outcome Y given an unobserved variable θ is Poisson, with a mean and variance θ, so

$$Y|\theta \sim \text{Poisson}(\mu\theta)$$

a) Assume $E(\theta) = 1$ and $Var(\theta) = \sigma^2$. Using the laws of total expectation and total variance, show $E(Y) = \mu$ and $Var(Y) = \mu(1 + \mu\sigma^2)$.
b) Assume θ is Gamma distributed with α and β as shape and scale parameters, respectively. Show the unconditional distribution of Y is Negative Binomial.
c) In order for $E(Y) = \mu$ and $Var(Y) = \mu(1 + \mu\sigma^2)$, what must α and β equal?

3 Simulation

Generate 100 datasets of an explanatory x and Poisson outcome y with the following code:

```
set.seed(123)
X <- matrix(NA, 100, 100)
Y <- matrix(NA, 100, 100)

for(i in 1:100){
  x <- rnorm(100)
  y <- rpois(100,lambda = exp(0.5+1*x+0.2*x^2))
  X[i,] <- x
  Y[i,] <- y
}
```

a) Fit a Poisson GLM to each dataset, with the 'correct' explanatory variables (x and x^2). Store the coefficient estimates and standard errors from each model run (hint: you can get SEs from `sqrt(diag(vcov(mod)))` where `mod` is your model object).

b) Calculate the coverage probability of a 2 standard error confidence interval for the coefficient on x and assess whether this is a useful way to construct 95% confidence intervals.

c) Calculate 100 Wald tests for the coefficient on x against $\beta_0 = 1$ (i.e. the true value). (hint: the `pt()` function will return the P-value of a t-test with a specified degrees of freedom). In how many of the tests is the null rejected?

Now generate 100 new datasets based on the code below:

```
set.seed(321)
X2 <- matrix(NA, 100, 100)
Y2 <- matrix(NA, 100, 100)
for(i in 1:100){
  weights <- ifelse(X[i,]>1, 10, 1)
  probs <- weights/sum(weights)
  to_keep_2 <- sample(1:length(X[i,]), 25, prob = probs)

  x2 <- X[i,to_keep_2]
  y2 <- Y[i,to_keep_2]

  X2[i,] <- x2
  Y2[i,] <- y2
}
```

d) Repeat parts a)-c) on the new data $X2$ and $Y2$.

e) What is happening here? Give a brief description of what you observe. How does this relate to a 'real world' situation of collecting data?

4 Opioid mortality in the US

The following questions relate to the `opioids` dataset, which you can find in the `data` folder of the `applied_stats` repo. It's an RDS file, which you can read in using `read_rds` from the `tidyverse`. There is also a `opioids_codebook.txt` file which explains each of the variables in the dataset.

The data contains deaths due to opioids by US from 2008 to 2017. In addition, there are population counts and a few other variables of interest. The goal is to explore trends and patterns in opioid deaths over time and across geography. The outcome of interest is `deaths`.

Please make sure to clearly explain any findings or observations you make, rather than just handing in code and output. You will be assessed not only on the code but also on how you communicate your findings with a combination of writing and analysis.

a) Perform some exploratory data analysis (EDA) using this dataset, and briefly summarize in words, tables and charts your main observations. You may use whatever tools or packages you wish. You may want to explore the `geofacet` package, which plots US state facets in the correct geographic orientation.

b) Run a Poisson regression using `deaths` as the outcome and `tot_pop` as the offset. (remember to `log` the offset). Include the `state` variable as a factor and change the reference category to be Illinois (you can do this using the `relevel` function). Investigate which variables to include, justifying based on your EDA in part a). Interpret your findings, including visualizations where appropriate. Include an analysis of which states, after accounting for other variables in the model, have the highest opioid mortality.

c) What's an issue with using population as an offset, given the limited information available in this dataset? (hint: the probability of death varies by age).

d) Rerun your Poisson regression using `expected_deaths` as an offset. How does this change the interpretation of your coefficients?

e) Investigate whether overdispersion is an issue in your current model.

f) If overdispersion is an issue, rerun your analysis using negative binomial regression (using the `glm.nb` function in the `MASS` library). Does this change the significance of your explanatory variables? Do a Likelihood Ratio Test to see which is the preferred model.

g) Summarize your findings, giving the key insights into trends in opioid mortality over time and across states, and any factors that may be associated with these changes. What other variables may be of interest to investigate in future?