Homework # 5

Due Wednesday 4/11

Turn in R code where applicable.

1. *Multinomial in Exponential Family Form:* A multivariate generalization of the exponential dispersion family is

$$f(\mathbf{y}_i|\theta_i,\phi) = \exp\left\{[\mathbf{y}_i^T\theta_i - b(\theta_i)]/a(\phi) + c(\mathbf{y}_i,\phi)\right\},$$

   where θ_i is the natural parameter. Show that a multinomial variate $\mathbf{y}_i$ for a single trial with parameters $\{\pi_j, j = 1,\ldots,J-1\}$ is in the $(J-1)$-parameter exponential family, with baseline-category logits as natural parameters.

2. The `GSS.csv` data on Blackboard contains survey data for studying the effect of gender and race on political party identification.

   (a) Fit a baseline category logit model with Independent as the baseline category, and gender and race as covariates. Report the R output.

   (b) Does the model fit well?

   (c) Interpret the gender effect for Democrat vs. Independent and Republican vs. Independent. Is there are statistically significant gender effect (1) overall or (2) for each baseline logit model?

   (d) Find the predicted probability of being a democrat for black females.

   (e) Without calculating estimated probabilities, explain why the intercept estimates indicate that for black females, $\hat{\pi}_D > \hat{\pi}_I > \hat{\pi}_R$.

   (f) Using the results from the model in part (a), find the prediction equation for $\log\left(\frac{\pi_D}{\pi_R}\right)$.

   (g) Fit a baseline category logit with Republican as the baseline category, and gender and race as covariates. Find the prediction equation for $\log\left(\frac{\pi_D}{\pi_R}\right)$. How does it compare to part (f)?

   (h) Fit the two separate logistic models corresponding to the baseline category logit model fit in part (a). Compare the estimated coefficients and their statistical significance.

   (i) Explain your findings in part (h). Why are they the same/different?

3. The `attitudes.csv` data on Blackboard contains survey data for studying on attitudes on legalizing abortion. The data is clustered, where each subject was asked about their support of legalizing abortion in three different scenarios. The dependent variable for analysis is support for legalized abortion (`response = 1`) and the possible covariates are female (`gender = 1`) and scenario (`question = 1,2,3`).

   (a) Fit a logistic regression model via GEE with an unstructured correlation using `gender` and `question` as covariates. Treat the "time" variable (`question`) as categorical. Report the R output.

   (b) State and interpret the estimated correlation matrix. What does this suggest about a reasonable working correlation structure? Why?

(c) Fit the same model as in part (a) but with an exchangeable correlation structure. State the correlation matrix. How does this compare to the correlation matrix estimated in part (b)?

(d) Interpret the gender effect as estimated in part (c). What are the results from the Wald test of the relation between `response` and `gender`?

(e) Find a 95% confidence interval for $\hat{\beta}_{gender}$ and $e^{\hat{\beta}_{gender}}$.

(f) Find the estimated odds of support for legalized abortion in scenario 2 for a male.

(g) Fit the model in part (c) treating the "time" variable (`question`) as continuous. Compare results to those from part (c). Suggest a questionnaire design that would reasonably warrant treating `question` as continuous.

(h) Fit a logistic regression ignoring the clustered nature of the data. Compare the coefficient estimates and standard errors from this independence model to the exchangeable model in part (c).

(i) Give a heuristic explanation of why the within-subject standard errors are much larger in the GLM than with GEE, yet the between-subject standard error is smaller.

4. *Explaining Q3, Part (i):* A positive correlation results in reduced standard error values for within-cluster estimated effects. What about between-cluster effects? Consider

$$y_{ij} \sim Bernoulli(\pi_i), \text{ where } i = 1, 2 \text{ and } j = 1, \ldots, T.$$

Suppose that for $i = 1, 2$, $corr(y_{ij}, y_{ik}) = \rho$ and $corr(y_{1j}, y_{2k}) = 0$ for all $j \neq k$. Find $SE(\hat{\pi}_1 - \hat{\pi}_2)$ and show that it is larger when $\rho > 0$ than when observations within the two samples are independent.