PHP 2517
Homework 3
Due Wednesday, March 20

The data and the code you need are attached on the canvas site. In this exercise, you will fit several models and then answer questions about them. The R code is already written for you, so you mostly need to provide interpretations as indicated in each question. The CTQ data have been organized as follows:

- We are only considering outcomes from weeks 5 to 12. The variable `postweek` in the R code is coded from 0 to 8, with `postweek = 0` corresponding to week 5, `postweek = 1` corresponding to week 6, and so forth.
- The Fagerstrom score has been centered at its mean, which is 6.29. The variable `totfager.cent` therefore has mean zero when averaged over all observations. Note that Fagersgtrom score is only measured at baseline.
- The variable Z is treatment indicator (1 = exercise, 0 = control).

With regard to notation,

$$\begin{aligned} t_j &= 0, 1, \ldots, 8 \\ &= \text{measurement week as coded in 'postweek'} \\ Y_{ij} &= 1 \text{ if individual } i \text{ quit smoking during week } t_j, \text{ 0 if not} \\ F_i &= \text{centered Fagerstrom score for individual } i \\ Z_i &= 1 \text{ if exercise, 0 if control} \end{aligned}$$

Based on this, please answer the following questions. You *do not* need to hand in R output or graphs.

1. Model M1 in the R code fits the following multilevel model, where $\boldsymbol{X}_{ij} = (1, t_j, F_i, Z_i)$ and $\pi_{ij} = E(Y_{ij} | \boldsymbol{X}_{ij}, \alpha_i)$.

   Level 1:

   $$\begin{aligned} Y_{ij} &\sim \text{Ber}(\pi_{ij}) \\ \text{logit}(\pi_{ij}) &= \alpha_i + \beta_1 t_j + \beta_2 F_i + \beta_3 Z_i \end{aligned}$$

   Level 2:

   $$\alpha_i \sim \mathcal{N}(\beta_0, \tau^2)$$

   (a) Using the output from Model M1, construct a table with estimates and standard errors for each of the parameters in the model. For the τ parameter, you do not have to write the standard error.

(b) Using the model output, provide an estimate of smoking cessation at week 5 for an individual having $\alpha_i = 0$, $F = 0$, and $Z = 0$. Do the same for an individual having $\alpha_i = 0$, $F = 0$, and $Z = 1$.

(c) The R program contains code to construct two histograms, one of the $\widehat{\alpha}_i$ and another of a variable $\widehat{p}_i$. What quantities are being depicted in each of these histograms? Please be specific with your answer, referring to relevant populations defined by the covariates as appropriate.

(d) Histograms like the ones in the previous question can sometimes be helpful in assessing whether the normality assumption makes sense for the random intercept. Based on these histograms, give your assessment of the validity of the normality assumption, and provide a brief justification.

(e) Provide an interpretation of β_3, the coefficient of Z from this model.

(f) Provide an interpretation of the coefficient β_0.

(g) Using the approximation discussed in class, convert the coefficients of Z, F, and t to their 'population-averaged' counterparts.

2. Models G1.indep, G1.exch, and G1.unst fit generalized linear models to the same data using different assumptions about the correlation. The mean function in this model is $\mu_{ij} = E(Y_{ij}|\boldsymbol{X}_{ij})$. For each model, the specification of mean and variance are given by

$$\begin{aligned} \text{logit}(\mu_{ij}) &= \gamma_0 + \gamma_1 t_j + \gamma_2 F_i + \gamma_3 Z_i \\ \text{var}(Y_{ij}|X_{ij}) &= \phi\mu_{ij}(1-\mu_{ij}) \end{aligned}$$

The elements ρ_{jk} of the correlation matrix are as follows:

Model	$\text{corr}(Y_{ij}, Y_{ik})$
G1.indep	$\rho_{jk} = 0$
G1.exch	$\rho_{jk} = \rho$ (all correlations equal)
G1.unst	ρ_{jk} distinct for each (j,k) pair

(a) Which of the GLM correlation structures in the same as for the multilevel model?

(b) Construct a table where the row entries are the regression parameters and the scale parameter ϕ and the columns correspond to each model. Fill the table with the estimate and standard error of each parameter.

(c) What is the interpretation of γ_3 in model G1.unst?

(d) Which of these models gives an estimate of γ_3 that is closest to the *marginalized* estimate of β_3 from the multilevel model?

(e) Construct another table where the row entries correspond to estimates of $P(Y_{ij} = 1)$ when (a) $t_j = 0$, $F = 0$ and $Z = 0$; and (b) $t_j = 0$, $F = 0$ and $Z = 1$. The column entries correspond to each model.

(f) Provide an explanation as to why these estimates are different from each other.

(g) If you were reporting the results of this study for a journal, which model(s) would you report and why, and what would be your conclusion about treatment effect?