Stat 539 Homework 7

Kenny Flagg

May 3, 2017

1. *Section 9.1.1 on p. 287-288 illustrates an example of the effects of ignoring correlation within subjects. Derive and verify the two expressions for var(b) and var(w) in equation (9.1) on p. 288.*

   For the between-subject effect,

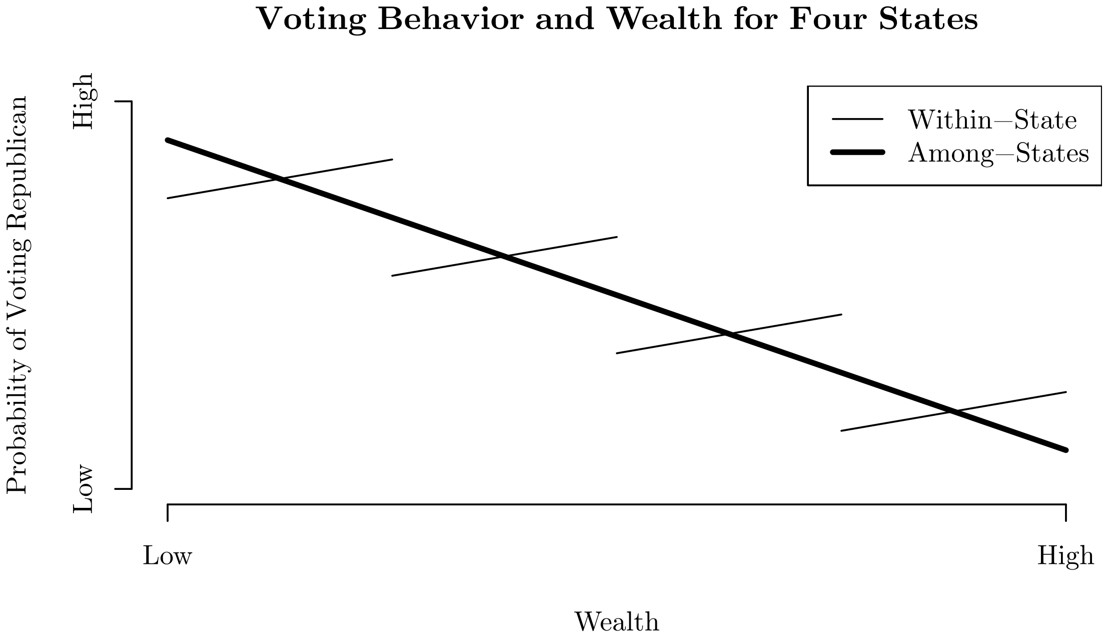
$$
\begin{aligned}
var(b) &= var\left(\frac{\bar{y}_1^A + \bar{y}_2^A}{2} - \frac{\bar{y}_1^B + \bar{y}_2^B}{2}\right) \\
&= \frac{1}{4n^2} var\left(\sum_{i=1}^{n} y_{i1}^A + \sum_{i=1}^{n} y_{i2}^A - \sum_{i=n}^{n} y_{i1}^B - \sum_{i=1}^{n} y_{i2}^B\right) \\
&= \frac{1}{4n^2}\left(var\left(\sum_{i=1}^{n} y_{i1}^A\right) + var\left(\sum_{i=1}^{n} y_{i2}^A\right) + var\left(\sum_{i=n}^{n} y_{i1}^B\right) + var\left(\sum_{i=1}^{n} y_{i2}^B\right)\right. \\
&\qquad + 2cov\left(\sum_{i=1}^{n} y_{i1}^A, \sum_{i=1}^{n} y_{i2}^A\right) - 2cov\left(\sum_{i=1}^{n} y_{i1}^A, \sum_{i=1}^{n} y_{i1}^B\right) - 2cov\left(\sum_{i=1}^{n} y_{i1}^A, \sum_{i=1}^{n} y_{i2}^B\right) \\
&\qquad \left. - 2cov\left(\sum_{i=1}^{n} y_{i2}^A, \sum_{i=1}^{n} y_{i1}^B\right) - 2cov\left(\sum_{i=1}^{n} y_{i1}^A, \sum_{i=1}^{n} y_{i2}^B\right) + 2cov\left(\sum_{i=1}^{n} y_{i1}^B, \sum_{i=1}^{n} y_{i2}^B\right)\right) \\
&= \frac{1}{4n^2}\left(\sum_{i=1}^{n} var(y_{i1}^A) + \sum_{i=1}^{n} var(y_{i2}^A) + \sum_{i=n}^{n} var(y_{i1}^B) + \sum_{i=1}^{n} var(y_{i2}^B)\right. \\
&\qquad + 2\sum_{i=1}^{n}\sum_{j=1}^{n} cov(y_{i1}^A, y_{j2}^A) - 2\sum_{i=1}^{n}\sum_{j=1}^{n} cov(y_{i1}^A, y_{j1}^B) - 2\sum_{i=1}^{n}\sum_{j=1}^{n} cov(y_{i1}^A, y_{j2}^B) \\
&\qquad \left. - 2\sum_{i=1}^{n}\sum_{j=1}^{n} cov(y_{i2}^A, y_{j1}^B) - 2\sum_{i=1}^{n}\sum_{j=1}^{n} cov(y_{i1}^A, y_{j2}^B) + 2\sum_{i=1}^{n}\sum_{j=1}^{n} cov(y_{i1}^B, y_{j2}^B)\right) \\
&= \frac{1}{4n^2}\left(n\sigma^2 + n\sigma^2 + n\sigma^2 + n\sigma^2 + 2\sigma^2\sum_{i=1}^{n} corr(y_{i1}^A, y_{i2}^A) - 0 - 0 - 0 - 0 + 2\sigma^2\sum_{i=1}^{n} corr(y_{i1}^B, y_{i2}^B)\right) \\
&= \frac{1}{4n^2}\left(4n\sigma^2 + 4n\sigma^2\rho\right) \\
&= \frac{\sigma^2(1+\rho)}{n}.
\end{aligned}
$$

For the within-subject effect,

$$
\begin{aligned}
var(w) &= var\left(\frac{\bar{y}_1^A + \bar{y}_1^B}{2} - \frac{\bar{y}_2^A + \bar{y}_2^B}{2}\right) \\
&= \frac{1}{4n^2} var\left(\sum_{i=1}^{n} y_{i1}^A + \sum_{i=1}^{n} y_{i1}^B - \sum_{i=n}^{n} y_{i2}^A - \sum_{i=1}^{n} y_{i2}^B\right) \\
&= \frac{1}{4n^2}\Bigg(var\left(\sum_{i=1}^{n} y_{i1}^A\right) + var\left(\sum_{i=1}^{n} y_{i1}^B\right) + var\left(\sum_{i=n}^{n} y_{i2}^A\right) + var\left(\sum_{i=1}^{n} y_{i2}^B\right) \\
&\qquad + 2cov\left(\sum_{i=1}^{n} y_{i1}^A, \sum_{i=1}^{n} y_{i1}^B\right) - 2cov\left(\sum_{i=1}^{n} y_{i1}^A, \sum_{i=1}^{n} y_{i2}^A\right) - 2cov\left(\sum_{i=1}^{n} y_{i1}^A, \sum_{i=1}^{n} y_{i2}^B\right) \\
&\qquad - 2cov\left(\sum_{i=1}^{n} y_{i1}^B, \sum_{i=1}^{n} y_{i2}^A\right) - 2cov\left(\sum_{i=1}^{n} y_{i1}^B, \sum_{i=1}^{n} y_{i2}^B\right) + 2cov\left(\sum_{i=1}^{n} y_{i2}^A, \sum_{i=1}^{n} y_{i2}^B\right)\Bigg) \\
&= \frac{1}{4n^2}\Bigg(\sum_{i=1}^{n} var(y_{i1}^A) + \sum_{i=1}^{n} var(y_{i1}^B) + \sum_{i=n}^{n} var(y_{i2}^A) + \sum_{i=1}^{n} var(y_{i2}^B) \\
&\qquad + 2\sum_{i=1}^{n}\sum_{j=1}^{n} cov(y_{i1}^A, y_{j1}^B) - 2\sum_{i=1}^{n}\sum_{j=1}^{n} cov(y_{i1}^A, y_{j2}^A) - 2\sum_{i=1}^{n}\sum_{j=1}^{n} cov(y_{i1}^A, y_{j2}^B) \\
&\qquad - 2\sum_{i=1}^{n}\sum_{j=1}^{n} cov(y_{i1}^B, y_{j2}^A) - 2\sum_{i=1}^{n}\sum_{j=1}^{n} cov(y_{i1}^B, y_{j2}^B) + 2\sum_{i=1}^{n}\sum_{j=1}^{n} cov(y_{i2}^A, y_{j2}^B)\Bigg) \\
&= \frac{1}{4n^2}\left(n\sigma^2 + n\sigma^2 + n\sigma^2 + n\sigma^2 + 0 - 2\sigma^2\sum_{i=1}^{n} corr(y_{i1}^A, y_{i2}^A) - 0 - 0 - 2\sigma^2\sum_{i=1}^{n} corr(y_{i1}^B, y_{i2}^B) + 0\right) \\
&= \frac{1}{4n^2}\left(4n\sigma^2 - 4n\sigma^2\rho\right) \\
&= \frac{\sigma^2(1-\rho)}{n}.
\end{aligned}
$$

2. *Agresti Exercise 9.25 (p. 328). For recent US Presidential elections, in each state wealthier voters tend to be more likely to vote Republican, yet states that are wealthier in an aggregate sense are more likely to have more Democrat than Republican votes (Gelman and Hill 2007, Section 14.2). Sketch a plot that illustrates how this instance of Simpson's paradox could occur. Specify a GLMM with random effects for states that could be used to analyze data for a sample of voters using their state of residence, their household income, and their vote in an election. Explain how the model could be generalized to allow the income effect to vary by state, to reflect that Republican-leaning states tend to have stronger associations between income and vote.*

   This could occur if the people with more wealth are more likely to vote Republican than people in their state with less wealth, but the mean wealth differs among states, and people in states that have more wealth on average are generally less likely to vote Republican. The plot below illustrates this for four hypothetical states. The thin lines show the within-state associations for the amounts of wealth present in each state. The thick line shows the among-state association, where states that are centered higher on the wealth axis have lower mean probabilities of voting Republican.



   A GLMM to model this would be

$$\begin{aligned} y_{ij} &\sim \text{Binomial}(1, \pi_{ij}); \\ \text{logit}(\pi_i) &= \beta_0 + \beta_1 x_{ij} + u_{0j}, \\ u_{0j} &\sim \text{N}(0, \sigma_0^2) \end{aligned}$$

   for the ith voter in the jth state, where $y_{ij} = 1$ if they voted Republican or 0 otherwise, x_{ij} is their household income, and u_{0j} is a random intercept for state j.

We can allow the strength of the association to change across states by including a random slope.

$$\begin{aligned} y_{ij} &\sim \text{Binomial}(1, \pi_{ij}); \\ \text{logit}(\pi_i) &= \beta_0 + \beta_1 x_{ij} + u_{0j} + u_{1j} x_{ij}, \\ \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} &\sim \text{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_{12} \\ \sigma_{12} & \sigma_1^2 \end{pmatrix} \right) \end{aligned}$$

3. *Agresti Exercise 9.39 (p. 332).*

(a) *Use the GEE approach to fit the logistic model, assuming an exchangeable working correlation structure for observations within a litter. Show how the empirical sandwich adjustment increases the SE values compared with naive binomial ML. Report the estimated within-litter correlation between the binary responses, and compare with the value of 0.192 that yields the quasi-likelihood results with beta-binomial variance function.*

The model for the jth fetus in the ith litter is

$$E(s_{ij}) = \pi_{ij}, \quad Var(s_{ij}) = \pi_{ij}(1 - \pi_{ij}), \quad Corr(s_{ij}, s_{i,k}) = \alpha;$$
$$\text{logit}(\pi_{ij}) = \beta_0 + \beta_1 z_i + \beta_2 h_i$$

where $s_{ij} = 1$ if the fetus died and 0 otherwise, $z_i = 1$ if the mother of litter i received a placebo and 0 if she received an iron injection, and h_i is the mother's hemoglobin level.

The table below shows that the standard errors from the GEE fit using the robust sandwich are about 1.6 times the GLM standard errors. The within-litter correlation estimated by GEE is $\widehat{\alpha} = 0.1754$, which is similar to (but a tiny bit smaller than) the estimated value from Agresti's beta-binomial model.

	GLM Estimate	GLM SE	GEE Estimate	GEE SE
$\widehat{\beta}_0$	-0.624	0.790	-0.723	1.295
$\widehat{\beta}_1$	2.651	0.482	2.755	0.766
$\widehat{\beta}_2$	-0.187	0.074	-0.176	0.127

(b) *Fit the GLMM that adds a normal random intercept u_i for litter i to the binomial logistic model. Interpret the estimated effects, and explain why they are larger than with the GEE approach.*

The random-intercept model is

$$\begin{aligned} s_{ij}|u_i &\sim \text{Binomial}(1, \pi_{ij}); \\ \text{logit}(\pi_{ij}) &= \beta_0 + \beta_1 z_i + \beta_2 h_i + u_i, \\ u_i &\sim \text{N}(0, \sigma_u^2). \end{aligned}$$

The table below shows the estimated coefficients. The estimated random effect variance is $\widehat{\sigma}_u^2 = 2.36$.

	Estimate	Std. Error
$\widehat{\beta}_0$	-1.280	1.619
$\widehat{\beta}_1$	3.937	1.069
$\widehat{\beta}_2$	-0.182	0.143

$\widehat{\beta}_0$: We estimate that the odds of death are 0.278 for a fetus in a litter from a typical mother who received an iron injection and had a hemoglobin level of zero. (Note that hemoglobin levels of zero were not observed and probably could not occur.)

$\widehat{\beta}_1$: We estimate that the odds of death are 5028.3% higher for a fetus in a typical litter from a mother who received a placebo injection than for a fetus in a typical litter from a mother with the same hemoglobin level who received an iron injection.

$\widehat{\beta}_2$: We estimate that the odds of death are 16.7% lower for a fetus in a typical litter from a mother with a hemoglobin level one unit higher than for a fetus in a typical litter from a mother with the same treatment status and the lower hemoglobin level.

These effect estimates are larger because they are conditional on the random effect, that is, either within one litter observed under different conditions, or between litters that fall in the same position within the distribution of litters of the same treatment status and mother's hemoglobin level. This in in contrast to the marginal effects in part (a) which are more general because they are averaged over all of the possible random effects.

4. *The Skin Cancer Prevention Study was a randomized, double-blind, placebo-controlled clinical trial of beta carotene to prevent non-melanoma skin cancer in high risk subjects (Greenberg et al., 1989, 1990; also see Stuckel, 1993). A total of 1805 subjects were randomized to either placebo or 50 mg of beta carotene per day for five years. Subjects were examined once a year and biopsied if a cancer was suspected to determine the number of new skin cancers occurring since the last exam. The outcome variable is a count of the number of new skin cancers per year. The outcome was evaluated on 1683 subjects comprising a total of 7081 measurements. The main objective of the analyses is to compare the effects of beta carotene on skin cancer rates. Variables are defined as follows:*

ID	Identification number for subject
Center	Center where treated
Age	Age of subject in years
Skin	Skin type (1 = burns; 0 = otherwise)
Gender	1 = male; 0 = female
Exposure	Count of the number of previous skin cancers
Y	Count of the number of new skin cancers per year
Trt	1 = beta carotene; 0 = placebo
Year	Year of follow-up

(We are going to ignore the correlation that may be present with the Center variable; do not use the Center variable in any of the following analyses.)

(a) *Consider a Poisson generalized linear mixed model for the subject-specific log rate of skin cancers with randomly varying intercepts. Fit a model with linear trends for the log rate over time and allow the slopes to depend on the treatment group. Do not put any other covariates (besides Trt and Year) into the model. Report the fitted model equation.*

The estimated model is

$$\begin{aligned} Y_{ij}|u_i &\sim \text{Poisson}(\mu_{ij}); \\ \log(\widehat{\mu}_{ij}) &= -2.38 + 0.0793\text{Trt}_i - 0.000104\text{Year}_{ij} + 0.0348\text{Trt}_i\text{Year}_{ij} + \widehat{u}_i, \\ \widehat{u}_i &\sim \text{N}(0, 2.15). \end{aligned}$$

(b) *Write an interpretation for each of the estimated fixed effects, β_j.*

$\widehat{\beta}_0$: In year zero, for a typical subject receiving the placebo, we estimate the mean number of new skin cancers to be 0.0927. (Note that the first year observed was year 1, so this interpretation is not appropriate.)

$\widehat{\beta}_1$: In year zero, for a typical subject receiving the beta-carotene treatment, we estimate the mean number of new skin cancers to be 8.25% higher than the mean number of new skin cancers for a typical subject receiving the placebo. (Again, it is not appropriate to interpret year zero.)

$\widehat{\beta}_2$: For a subject receiving the placebo, we estimate the mean number of new skin cancers to be 0.0104% lower than the mean number of new skin cancers for that subject in the previous year.

$\widehat{\beta}_3$: For a typical subject receiving the beta-carotene treatment and a typical subject receiving the placebo, we estimate the mean number of new skin cancers for the beta-carotene subject relative to the placebo subject to be 3.54% higher than in the previous year.

(c) *What is the estimate of the standard deviation of the randomly varying intercepts? Give an interpretation of this value in context of the problem.*

The estimated standard deviation of the random intercepts is $\widehat{\sigma}_u = 1.46$. For a given treatment status and year, we estimate that 95% of subjects have expected new skin cancer counts between 94.3% lower and 1664.7% higher than the mean number of new skin cancers for a typical subject.

(d) *What conclusions do you draw about the effect of beta carotene on the log rate of skin cancers using the model in part (a)? Provide results that support your conclusions.*

I fit a reduced model without the interaction term and treatment main effect, so year is the only predictor. The likelihood ratio test provides little to no evidence of an association between the beta-carotene treatment and the log-rate of new cancers after controlling for year ($\chi^2_2 = 3.93$, p-value = 0.14).

	Df	deviance	Chisq	Chi Df	Pr(>Chisq)
Reduced	3	5786.7			
Full	5	5782.7	3.9344	2	0.1399

(e) *Obtain the predicted (empirical BLUP) random effect for each subject. (Only print the first few BLUPs.)*

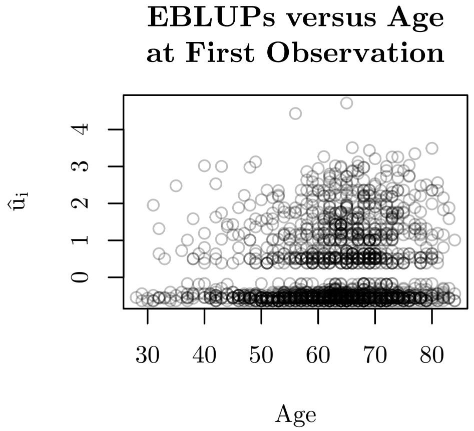
ID	$\widehat{u}_i$
30	-0.565
100012	-0.634
100023	1.743
100034	1.585
100045	-0.565
100056	1.585
100067	-0.565
100078	0.386
100089	0.386
100102	-0.488

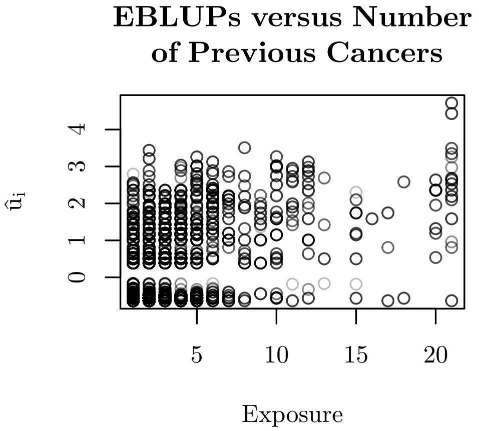
i. *Calculate the sample variance of the predictions. How does it compare to the estimate of the variance of the random intercepts obtained from your fitted model? Why might they differ?*

The sample variance of the EBLUPs is 1.03, much smaller than the parameter estimate $\sigma^2_u = 2.15$. They are expected to differ slightly because the estimated variance is for the population while the sample variance of the EBLUPs depends on the observed counts. A difference this large might be due a poor model fit.

ii. *Plot the predictions against age and the count of the number of previous skin cancers. What do you conclude from these plots?*

The plots below do not show any trends in the random intercepts across age and number of previous cancers, but there are two groups of EBLUP values. We can conclude that the predicted random effects do not follow a normal distribution.





(f) *Fit a marginal model with the same link and fixed effects as your model from part (a), assuming an exchangeable correlation structure. Report the fitted model equation and write an interpretation for each of the estimated fixed effects, β_j.*

The estimated model is

$$\begin{aligned} Y_{ij} &\sim \text{Poisson}(\mu_{ij}), \\ \widehat{Cor}(Y_{ij}, Y_{ik}) &= 0.3781; \\ \log(\widehat{\mu}_{ij}) &= -1.37 + 0.0606\text{Trt}_i + 0.000147\text{Year}_{ij} + 0.0323\text{Trt}_i\text{Year}_{ij}. \end{aligned}$$

$\widehat{\beta}_0$: In year zero, for subjects receiving the placebo, we estimate the mean number of new skin cancers to be 0.2552.

$\widehat{\beta}_1$: In year zero, we estimate the mean number of new skin cancers for subjects receiving the beta-carotene treatment to be 6.25% higher than the mean number of new skin cancers for subjects receiving the placebo.

$\widehat{\beta}_2$: For subjects receiving the placebo, we estimate the mean number of new skin cancers to be 0.0147% higher than the mean number of new skin cancers for subjects receiving the placebo in the previous year.

$\widehat{\beta}_3$: We estimate the mean number of new skin cancers for the subjects receiving the beta-carotene relative to the placebo subjects to be 3.28% higher than in the previous year.

(g) *For the main objective of the analysis, is a GLMM or marginal model more appropriate? Explain.*

For this analysis, marginal models are more appropriate because the goal of the study is to see if the beta-carotene treatment would be effective for the overall population, not for individuals.

(h) *Using the method you chose in part (g), repeat the analysis adjusting for skin type, age, and the count of the number of previous skin cancers. What conclusions do you draw about the effect of beta carotene on the adjusted log rate of skin cancers?*

The model is

$$\begin{aligned} Y_{ij} &\sim \text{Poisson}(\mu_{ij}), \\ Cor(Y_{ij}, Y_{ik}) &= \alpha; \\ \log(\mu_{ij}) &= \beta_0 + \beta_1 \text{Skin}_i + \beta_2 \text{Age}_i + \beta_3 \text{Exposure}_i + \beta_4 \text{Trt}_i + \beta_5 \text{Year}_{ij} + \beta_6 \text{Trt}_i \text{Year}_{ij}. \end{aligned}$$

The table below shows the estimated parameters.

	Estimate	Std. Error
$\widehat{\beta}_0$	-1.280	1.619
$\widehat{\beta}_1$	3.937	1.069
$\widehat{\beta}_2$	-0.182	0.143

I fit a reduced model without the treatment by year interaction and the treatment main effect. The Wald test provides no evidence of an association between the beta-carotene treatment and the log-rate of new cancers after controlling for skin type, age, number of previous cancers, and year ($\chi^2_2 = 1.85$, p-value $= 0.40$).

Df	X2	P(>|Chi|)
2	1.8531	0.3959