Caijun Qin
12/02/2021

3.
a. For the ith coin, π_i is the number of times heads come up divided by 5 flips.
0.4, 0.8, 0.2, 0.6, 0.6, 1.0, 0.8, 0.4, 0.6, 0.2

b. $\alpha = 0.259$ and $\hat{\sigma} = 0.557$. The predicted values are 0.52, 0.63, 0.46, 0.57, 0.57, 0.68, 0.63, 0.52, 0.57, 0.46.

c. The average absolute distances are 0.22 and 0.08 for the two previous parts.

4.
a. The coefficient estimates for the intercept, alcohol indicator, and marijuana indicator are 1.5047, 2.3947, and -2.1693, respectively. Since cigarettes is given as the first factor level, it is the base. The intercept estimate, 1.5047, therefore relays the log odds of using cigarettes over using no substance at all. Analogously, the coefficient estimates 2.3947 and -2.1693 relay the log odds of using alcohol over no substance and using marijuana over no substance, respectively.

Testing marginal homogeneity on $H_0: \beta_0 = \beta_1 = \beta_2$ rejected the null hypothesis. With an LRT statistic approximately equivalent to 1646.5 as a chi-square estimate and a miniscule p-value of less than 2.2e-16, there exists very strong evidence that at least one of the estimated coefficients is necessary to make the model explanatory.

R code
```
fit <- lme4::glmer(
  formula = ifelse(test = Use == 'yes', yes = 1, no = 0) ~ (1 | Student) + factor(x = Substance, levels = c('cigarettes', 'alcohol', 'marijuana')),
  data = long.sub,
  family = binomial(link = 'logit')
)
summary(fit)

fit.null <- lme4::glmer(
  formula = ifelse(test = Use == 'yes', yes = 1, no = 0) ~ (1 | Student),
  data = long.sub,
  family = binomial(link = 'logit')
)

anova(fit, fit.null)
lmtest::lrtest(fit, fit.null)
```

b. The estimated variance, $\hat{\sigma}$, is 9.141, which is printed out by summary(fit) above.
i. Such a large variance implies high variability of subjects (students) amongst clusters and high correlation between subjects within the same cluster (type of substance use).

Caijun Qin
12/02/2021

ii. A large positive u_i for some student using substance (category) *i* means that there will be a large probability of that student using said substance regardless of how the odds of using substance *i* compare to other substance use.

c. The model from 9.1 assumes a population-average model, while this model recognizes a random component that may differ between clusters (types of substances). The coefficients of this model share the correct signs but with greater magnitude than the model from 9.1, and the std errors for this model are also very likely to be smaller. The log odds from this model are computed within the context of a specific cluster rather than the whole population (or sample for the study), inducing the observed differences in coefficients.

e.
(i) Using GLMM, the following coefficient estimates have been computed.
Intercept = 1.60083
Alcohol = 2.55968
Marijuana = -2.45493
Race = -0.92417
Gender = -0.02119
Alcohol:gender = -0.28681
Marijuana:Gender = 0.52475

(ii) Coefficients in this model have larger magnitudes than their counterparts from 9.3. This model also includes a random component for each subject (just call fitted function on GLMM model).

R code

```
fit1 <- lme4::glmer(
  formula = ifelse(test = Use == 'yes', yes = 1, no = 0) ~ (1 | Student) + factor(x = Substance, levels = c('cigarettes', 'alcohol', 'marijuana')) + Race + Gender + Gender:factor(x = Substance, levels = c('cigarettes', 'alcohol', 'marijuana')),
  data = long.sub1,
  family = binomial(link = 'logit')
)

summary(fit1)
```

6. From the given GEE model, the negative coefficients for severity and drug indicate decrease in log odds while the positive coefficients for time and interaction between drug and time indicate increase in log odds when their respective indicators are on. The difference with the GLMM model is that the intercept for the GLMM can be split into a fixed value and a cluster-specific random component.