Clustered Data Models: DATA.STAT.750

Exercises 02

Question 1

Derivation of Mean and Variance of Beta-Binomial Distribution

Given:

$$
\begin{aligned}
E(y) &= E[E(y \mid p)] \\
\operatorname{Var}(y) &= E[\operatorname{Var}(y \mid p)] + \operatorname{Var}[E(y \mid p)]
\end{aligned}
$$

Beta Distribution:

- Mean: $E(p) = \frac{\alpha}{\alpha+\beta}$
- Variance: $\operatorname{Var}(p) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Binomial Distribution: Given p, the distribution of y is binomial: $y \mid p \sim \text{Binomial}(n, p)$.
Step 1: Derive the Mean $E(y)$

$$
\begin{aligned}
E(y \mid p) &= np \\
E(y) &= E[E(y \mid p)] = E[np] = nE(p) \\
E(y) &= n\frac{\alpha}{\alpha+\beta}
\end{aligned}
$$

Step 2: Derive the Variance $\operatorname{Var}(y)$

$$
\operatorname{Var}(y) = E[\operatorname{Var}(y \mid p)] + \operatorname{Var}[E(y \mid p)]
$$

1.$E[\operatorname{Var}(y \mid p)]$**:

$$
\begin{aligned}
\operatorname{Var}(y \mid p) &= np(1-p) \\
E[\operatorname{Var}(y \mid p)] &= nE[p(1-p)] = n\left(E(p) - E(p^2)\right)
\end{aligned}
$$

Where:

$$
\begin{aligned}
E(p^2) &= \operatorname{Var}(p) + [E(p)]^2 \\
E(p^2) &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} + \left(\frac{\alpha}{\alpha+\beta}\right)^2 \\
E[\operatorname{Var}(y \mid p)] &= n\left(\frac{\alpha}{\alpha+\beta} - \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} - \left(\frac{\alpha}{\alpha+\beta}\right)^2\right)
\end{aligned}
$$

2. Var$[E(y \mid p)]$**:

$$\text{Var}[E(y \mid p)] = \text{Var}(np) = n^2 \text{Var}(p)$$
$$\text{Var}[E(y \mid p)] = n^2 \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Finally, combine

$$\text{Var}(y) = n\left(\frac{\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)}\right)$$

- **Mean of the Beta-Binomial Distribution**:

$$E(y) = n\frac{\alpha}{\alpha+\beta}$$

- **Variance of the Beta-Binomial Distribution**:

$$\text{Var}(y) = n\left(\frac{\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)}\right)$$

Question 4

```
model <- glm(count ~ race, family = poisson, data = homic)
summary(model)


Call:
glm(formula = count ~ race, family = poisson, data = homic)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.38321    0.09713  -24.54   <2e-16 ***
race         1.73314    0.14657   11.82   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 962.80  on 1307  degrees of freedom
Residual deviance: 844.71  on 1306  degrees of freedom
AIC: 1122

Number of Fisher Scoring iterations: 6
```

Model Fit: Null deviance: 962.80, with 1307 degrees of freedom.
Residual deviance: 844.71, with 1306 degrees of freedom.
The model with the race predictor fits better than the null model AIC (Akaike Information Criterion): 1122.
A lower AIC value indicates a better fit of the model, though it is primarily used for comparing different models.

b)

```
 model_nb <- glm.nb(count ~ race, data = homic)
> summary(model_nb)

Call:
glm.nb(formula = count ~ race, data = homic, init.theta = 0.2023119205,
    link = log)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.3832     0.1172 -20.335  < 2e-16 ***
race          1.7331     0.2385   7.268 3.66e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.2023) family taken to be 1)

    Null deviance: 471.57  on 1307  degrees of freedom
Residual deviance: 412.60  on 1306  degrees of freedom
AIC: 1001.8

Number of Fisher Scoring iterations: 1


              Theta:  0.2023
          Std. Err.:  0.0409

 2 x log-likelihood:  -995.7980
>
```

The significant Theta value (0.2023) and the better model fit (shown by lower residual deviance and AIC) WHich indicates that the data has more variability than the Poisson model. This means the Negative Binomial model is a better.

1 4. C)

Confidence Interval Comparison The Wald 95% confidence interval for the ratio of means for blacks and whites is (4.2, 7.5) for the Poisson GLM and (3.5, 9.0) for the Negative Binomial GLM. The Negative Binomial model is more reliable because it accounts for overdispersion in the data, which is better.