# BIOS 765 Homework 2

Leo Li
730031954

**Problem 1(a).**

In Table 1, we present the percentage of shoots that gave zero roots for each photoperiod.

Table 1: Percentages of Shoots Giving Zero Roots for each Photoperiod

| Photoperiod | Number of Zeros | Number of Shoots | Percentage (%) |
|---|---|---|---|
| 8-hour | 2 | 140 | 1.43 |
| 16-hour | 62 | 130 | 47.69 |

**Problem 1(b).**

The corresponding model expression for the zero-inflated Poisson (ZIP) regression model for number of roots as the outcome can be written as the following:

$$y_i = \begin{cases} 0, \text{ with probability } \psi_i \\ g(y_i), \text{ with probability } 1-\psi_i \end{cases}$$

where $g(y_i)$ is the Poisson probability function $g(y_i|\mu_i) = \frac{\exp[-\mu_i]\mu_i^{y_i}}{y_i!}$. In this model, the probability distribution, $P(Y_i = y_i)$, can be expressed as:

$$P(y_i = 0|x_{i1}, x_{i2}, x_{i3}) = \psi_i(\gamma) + [1-\psi_i(\gamma)]\exp(-\mu_i(\lambda)),$$

$$P(y_i|x_{i1}, x_{i2}, x_{i3}) = [1-\psi_i(\gamma)]\frac{\exp[-\mu_i(\lambda)]\mu_i(\lambda)^{y_i}}{y_i!},\ y_i > 0,$$

where,

$$\log(\mu_i) = \lambda_0 + \lambda_1 x_{i1} + \lambda_2 x_{i2} + \lambda_3 x_{i1} x_{i2},$$

$$\log(\frac{\psi_i}{1-\psi_i}) = \gamma_0 + \gamma_1 x_{i1},$$

in which $y_i$ is the number of roots of the shoot $i$; $\mu_i$ is the mean number of roots of the shoots in the "susceptible class"; and $\psi_i$ represents the probability of excess zeros.

Now, we fit the corresponding ZIP regression model as well as the zero-inflated negative binomial (ZINB) regression model. Table 2 presents the regression parameter estimates, standard errors, and p-values for both models.

Table 2: Analysis of Parameter Estimates for ZIP and ZINB Regression Models

| | | | ZIP Regression Model | | | ZINB Regression Model | | |
|---|---|---|---|---|---|---|---|---|
| Model Part | Parameter | Effect | Estimate | SE | p-value | Estimate | SE | p-value |
| Distribution Part | $\lambda_0$ | Intercept | 1.868 | 0.059 | <.001 | 1.867 | 0.071 | <.001 |
| | $\lambda_1$ | x1 | 0.009 | 0.105 | 0.931 | 0.006 | 0.127 | 0.962 |
| | $\lambda_2$ | x2 | 0.092 | 0.042 | 0.028 | 0.091 | 0.052 | 0.079 |
| | $\lambda_3$ | x1x2 | -0.258 | 0.079 | 0.001 | -0.259 | 0.094 | 0.006 |
| | $\alpha$ | Dispersion | | | | 0.070 | 0.025 | |
| Zero-Inflation Part | $\gamma_0$ | Intercept | -4.262 | 0.732 | <.001 | -4.381 | 0.827 | <.001 |
| | $\gamma_1$ | x1 | 4.159 | 0.753 | <.001 | 4.265 | 0.846 | <.001 |

**Problem 1(c).**

Now, we would like to test the dispersion parameter from the ZINB model. We use the non-standard LRT to compare ZIP and ZINB to test $H_0 : \alpha = 0$ vs. $H_1 : \alpha > 0$. We obtain the LRT test statistic as $Q_L = 2(-618.2 + 625.1) = 13.7$, and the corresponding p-value is less than 0.01, so that we reject the null hypothesis that the dispersion parameter equals 0. Then we can interpret the result of the hypothesis testing as that the ZINB regression is the preferred model.

**Problem 1(d).**

Now, we would like to estimate and interpret the exponentiated regression coefficients from both model parts from the ZINB model that we have chosen in part (c). We would also like to report the 95% confidence intervals.

- $\exp(\lambda_1)$ is estimated as 1.01, with 95% CI as (0.78, 1.29). It can be interpreted as the incidence rate ratio (IRR) for the "susceptible" class comparing the 16-hour photoperiod group to the 8-hour photoperiod group, when the BAP equals to $2.2\mu M$.

- $\exp(\lambda_2)$ is estimated as 1.10, with 95% CI as (0.99, 1.21). It can be interpreted as the incidence rate ratio (IRR) for the "susceptible" class when there is one unit increase in log BAP, within the 8-hour photoperiod group.

- $\exp(\lambda_3)$ is estimated as 0.77, with 95% CI as (0.64, 0.93). It can be interpreted as the ratio of two incidence rate ratios (IRRs). Specifically, it can be interpreted as the ratio of the IRR for the "susceptible" class when there is one unit increase in log BAP within the 16-hour photoperiod group and the IRR for the "susceptible" class when there is one unit increase in log BAP within the 8-hour photoperiod group.

- $\exp(\gamma_1)$ is estimated as 71.13, with 95% CI as (13.56, 373.08). It can be interpreted as the odds ratio of being classified as "non-susceptible" class of the 16-hour photoperiod group and the 8-hour photoperiod group.

**Problem 1(e).**

In this question, we would like to fit a marginalized zero-inflated negative binomial (MZINB) regression model as an alternative to the "best" model chosen in part (c).

Unlike the ZINB regression model, the MZINB regression model models the overall mean, $\nu_i$, instead of the mean for the "susceptible class", $\mu_i$, so that the MZINB regression model has a more straightforward interpretation to the regression coefficients. However, they do have the same assumption regarding the underlying data generation process:

$$y_i = \begin{cases} 0, \text{ with probability } \psi_i \\ g(y_i), \text{ with probability } 1-\psi_i \end{cases}$$

where $g(y_i)$ is the negative binomial probability function,

$$g(y_i|\mu_i, \alpha) = \frac{\Gamma(y_i + 1/\alpha)}{\Gamma(y_i+1)\Gamma(1/\alpha)}(\frac{1}{1+\alpha\mu_i})^{1/\alpha}(1-\frac{1}{1+\alpha\mu_i})^{y_i}.$$

In this model, estimation is performed by converting each individual's ZINB likelihood contribution $g(y_i|\mu_i, \psi_i, \alpha)$ to $g(y_i|\nu_i, \psi_i, \alpha)$ via the inversion of $\nu_i = (1-\psi_i)\mu_i$. Then the probability distribution, $P(Y_i = y_i)$, can be expressed as:

$$P(y_i = 0|x_{i1}, x_{i2}, x_{i3}) = \psi_i(\gamma) + [1-\psi_i(\gamma)][1+\alpha\frac{\nu_i(\beta)}{1-\psi_i(\gamma)}]^{-\alpha^{-1}},$$

$$P(y_i|x_{i1}, x_{i2}, x_{i3}) = [1-\psi_i(\gamma)]\frac{\Gamma(y_i+1/\alpha)}{y_i!\Gamma(1/\alpha)}[\frac{1}{1+\alpha\frac{\nu_i(\beta)}{1-\psi_i(\gamma)}}]^{1/\alpha}[1-\frac{1}{1+\alpha\frac{\nu_i(\beta)}{1-\psi_i(\gamma)}}]^{y_i},\ y_i > 0,$$

where,

$$\log(\nu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2},$$

$$\log(\frac{\psi_i}{1-\psi_i}) = \gamma_0 + \gamma_1 x_{i1},$$

in which $y_i$ is the number of roots of the shoot $i$; $\nu_i$ is the overall mean number of roots of the shoots; and $\psi_i$ represents the probability of excess zeros. Then, the fitted model can be written as the following:

$$\log(\hat{\nu_i}) = 1.855 - 0.618 x_{i1} + 0.091 x_{i2} - 0.259 x_{i1} x_{i2},$$

$$\log(\frac{\hat{\psi_i}}{1-\hat{\psi_i}}) = -4.382 + 4.266 x_{i1},$$

and the fitted value for the dispersion parameter is that $\hat{\alpha} = 0.070$. Table 3 presents the 95% confidence intervals along with the estimates resulted from the MZINB regression model.

Then, the estimated exponentiated regression coefficients can be interpreted as the following:

- $\exp(\beta_1)$ can be estimated as 0.539, with 95% CI as (0.400, 0.727). $\exp(\hat{\beta}_1)$ can be interpreted as the estimated incidence rate ratio (IRR) for the overall population comparing the 16-hour photoperiod group to the 8-hour photoperiod group, when the BAP equals to 2.2 $\mu M$.

Table 3: Analysis of Parameter Estimates for MZINB Regression Model

| Model Part | Parameter | Effect | Estimate | 95% CI | |
|---|---|---|---|---|---|
| Distribution Part | $\beta_0$ | Intercept | 1.855 | 1.712 | 1.997 |
| | $\beta_1$ | x1 | -0.618 | -0.917 | -0.319 |
| | $\beta_2$ | x2 | 0.091 | -0.011 | 0.194 |
| | $\beta_3$ | x1x2 | -0.259 | -0.445 | -0.073 |
| | $\alpha$ | Dispersion | 0.070 | 0.022 | 0.118 |
| Zero-Inflation Part | $\gamma_0$ | Intercept | -4.382 | -6.012 | -2.752 |
| | $\gamma_1$ | x1 | 4.266 | 2.599 | 5.932 |

- $\exp(\beta_2)$ can be estimated as 1.096, with 95% CI as (0.989, 1.214). $\exp(\hat{\beta}_2)$ can be interpreted as the estimated incidence rate ratio (IRR) for the overall population when there is one unit increase in log BAP, within the 8-hour photoperiod group.
- $\exp(\beta_3)$ can be estimated as 0.772, with 95% CI as (0.641, 0.929). $\exp(\hat{\beta}_3)$ can be interpreted as the estimated ratio of the IRR for the overall population when there is one unit increase in log BAP within the 16-hour photoperiod group and the IRR for the overall population when there is one unit increase in log BAP within the 8-hour photoperiod group.
- $\exp(\gamma_1)$ can be estimated as 71.215, with 95% CI as (13.453, 376.945). $\exp(\hat{\gamma}_1)$ can be interpreted as the estimated odds ratio of being classified as "non-susceptible" class of the 16-hour photoperiod group and the 8-hour photoperiod group.

**Problem 1(f).**

When BAP level is $2.2\mu M$, we estimate the incidence density ratio of the comparison of the 16-hour photoperiod versus the 8-hour photoperiod as,

$$\exp(\hat{\beta}_1) = 0.539,$$

and the corresponding confidence interval is (0.378, 0.700).

When BAP level is $4.4\mu M$, we estimate the incidence density ratio of the comparison of the 16-hour photoperiod versus the 8-hour photoperiod as,

$$\exp(\hat{\beta}_1 + (\log 4.4 - \log 2.2)\hat{\beta}_3) = 0.450,$$

and the corresponding confidence interval is (0.347, 0.554).

When BAP level is $8.8\mu M$, we estimate the incidence density ratio of the comparison of the 16-hour photoperiod versus the 8-hour photoperiod as,

$$\exp(\hat{\beta}_1 + (\log 8.8 - \log 2.2)\hat{\beta}_3) = 0.376,$$

and the corresponding confidence interval is (0.292, 0.460).

When BAP level is $17.6\mu M$, we estimate the incidence density ratio of the comparison of the 16-hour photoperiod versus the 8-hour photoperiod as,

$$\exp(\hat{\beta}_1 + (\log 17.6 - \log 2.2)\hat{\beta}_3) = 0.314,$$

and the corresponding confidence interval is (0.226, 0.403).

We can interpret the results as that, when the BAP level increases, the intensity density ratio of the comparison of the 16-hour photoperiod versus the 8-hour photoperiod decreases. In other words, the relative incidence of the 16-hour photoperiod group with respect to the incidence of the 8-hour photoperiod group will decrease as the level of BAP increases.

**Problem 2(a).**

We would like to fit the model of no three factor interaction (JM, JC, MC), and we estimate $\pi_0$, the conditional probabilities of falling into the categories represented by the rows of Table A. Then, we fit all of the three 2 two-factor interaction models, and we realize that the model with (JC, MC) is the best 2 two-factor interaction model. In Table 4, we present the observed as well as the model predicted conditional probabilities from models of (JM, JC, MC) and (JC, MC).

Table 4: Observed Probabilities and Model Predicted Probabilities from Models (JM, JC, MC) and (JC, MC)

| System | Count | Observed | Predicted (JM, JC, MC) | Predicted (JC, MC) |
|---|---|---|---|---|
| J only | 77269 | 0.266 | 0.266 | 0.266 |
| M only | 138578 | 0.477 | 0.477 | 0.477 |
| C only | 13391 | 0.046 | 0.046 | 0.043 |
| J and M | 25631 | 0.088 | 0.088 | 0.088 |
| J and C | 3079 | 0.011 | 0.011 | 0.014 |
| M and C | 23876 | 0.082 | 0.082 | 0.085 |
| J, M, and C | 8528 | 0.029 | 0.029 | 0.027 |

Considering the goodness of fit of model (JC, MC) relative to model (JM, JC, MC), we have that $Q_L = 360.6$ which asymptotically follows $\chi^2_1$ distribution, leading to a p-value less than 0.01. Therefore, we conclude that the model (JC, MC) and (JM, JC, MC) are significantly different in goodness of fit. Then, this result supports the full model of (JM, JC, MC).

**Problem 2(b).**

The $X$ matrix for the model (JM, JC, MC) can be given as the following,

$$X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & -1 & -1 & 1 \\ -1 & 1 & 1 & -1 & -1 & 1 \\ -1 & 1 & -1 & -1 & 1 & -1 \\ -1 & -1 & 1 & 1 & -1 & -1 \\ -1 & -1 & -1 & 1 & 1 & 1 \end{bmatrix},$$

where the very last row corresponds to the deleted row, $x'_{222}$, and the first seven rows correspond to $X_0$. The estimate to $\pi_{222}$ is that,

$$\hat{\pi}_{222} = \frac{\exp(x'_{222}\hat{\beta})}{1'_r[\exp(X'\hat{\beta})]} = 0.69.$$

In addition, the estimate to $\gamma$ is that,

$$\hat{\gamma} = \frac{\hat{\pi}_{222}}{1 - \hat{\pi}_{222}} = 2.24.$$

**Problem 2(c).**

In this question, we would like to determine $\hat{n}_{222}$ and $\hat{N}$ where $n_{222}$ is the number of the poor mentally ill not captured, and $N$ is the total population size of the poor mentally ill in King County. We have that,

$$\hat{n}_{222} = n\hat{\gamma} = 648963,$$
$$\hat{N} = n + \hat{n}_{222} = 939315.$$

Note that the answers have been rounded to the nearest integers to represent the number of people.

**Problem 2(d).**

We would like to compute $\widehat{var}[\hat{\pi}_0, \hat{\gamma}]' = BV_{\hat{\beta}}B'$, where,

$$B = D_{\exp[A_3 \log\{A_2[\exp(X\hat{\beta})]\}]} A_3 D^{-1}_{A_2 \exp(X\hat{\beta})} A_2 D_{\exp(X\hat{\beta})} X.$$

After some algebra, we obtain an estimate of $var[\hat{\pi}_0, \hat{\gamma}]'$ as:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 9.816305e-08 | -9.000449e-09 | -1.066143e-09 | -2.699223e-08 | -8.367709e-09 | -4.805179e-08 | -4.684730e-09 | 7.471116e-06 |
| -9.000449e-09 | 2.770270e-07 | -3.219659e-09 | -8.057105e-08 | -2.497792e-08 | -1.451112e-07 | -1.414678e-08 | -8.383622e-06 |
| -1.066143e-09 | -3.219659e-09 | 3.613213e-08 | -9.768130e-09 | -2.993256e-09 | -1.738960e-08 | -1.695335e-09 | -7.779708e-06 |
| -2.699223e-08 | -8.057105e-08 | -9.768130e-09 | 6.707066e-07 | -7.578041e-08 | -4.351685e-07 | -4.242625e-08 | 5.643009e-06 |
| -8.367709e-09 | -2.497792e-08 | -2.993256e-09 | -7.578041e-08 | 2.600274e-07 | -1.349075e-07 | -1.300056e-08 | -8.338632e-06 |
| -4.805179e-08 | -1.451112e-07 | -1.738960e-08 | -4.351685e-07 | -1.349075e-07 | 8.561567e-07 | -7.552810e-08 | 4.044227e-06 |
| -4.684730e-09 | -1.414678e-08 | -1.695335e-09 | -4.242625e-08 | -1.300056e-08 | -7.552810e-08 | 1.514818e-07 | 7.343610e-06 |
| 7.471116e-06 | -8.383622e-06 | -7.779708e-06 | 5.643009e-06 | -8.338632e-06 | 4.044227e-06 | 7.343610e-06 | 3.069431e-03 |

**Problem 2(e).**

Now, we would like to determine $se(\hat{n}_{222})$ and $se(\hat{N})$. We have that,

$$Var(\hat{n}_{222}) = N\hat{\gamma}^2\pi_{222}(1-\pi_{222}) + n^2V_{\hat{\gamma}},$$

$$Var(\hat{N}) = N(1+\hat{\gamma})^2\pi_{222}(1-\pi_{222}) + n^2V_{\hat{\gamma}}.$$

Then the value of $se(\hat{n}_{222})$ and $se(\hat{N})$ are determined as 16117.3 and 16151.3, respectively

**Problem 2(f).**

Here, we would also like to state some assumptions of our analysis:

- We assume that the cell frequencies in the $2^3$ contingency table follow a multinomial distribution. To be more specific, we assume that multivariate responses are independently and identically distributed such that each individual can be considered to have a multinomial distribution with sample size 1 and $\pi' = (\pi_{111}, ..., \pi_{222})$.
- The $2^3$ contingency table follows a log-linear model with no $d^{th}$ order interaction effect. This assumption is required to fit a model to a table with an empty cell. A more parsimonious model may be considered to obtain greater efficiency for an estimator of $N$.

**Problem 3(a).**

In this question, we would like to fit the hierarchical loglinear model containing all two-factor interactions as well as the three-factor interaction EDC. Table 6 in Appendix 1 presents the corresponding model results. The likelihood ratio test results in a test statistic of 155.25, following a $\chi^2_{154}$ distribution, suggesting the goodness of fit. However, there are some cells with very sparse or even zero counts, so that it is not appropriate to use this test to justify the goodness-of-fit of the model in our data.

**Problem 3(b).**

Now, we fit a reduced hierarchical loglinear model by deleting two-factor interaction terms from the model in (a) that are not statistically significant ($p > 0.10$). Table 7 in Appendix 1 presents the corresponding model results. Then, we would like to conduct a chi-square test of goodness-of-fit for the reduced model based upon comparison of deviance to the model in (a). The test statistic is calculated as $\Delta Deviance = Deviance[(b)] - Deviance[(a)] = 172.66 - 155.25 = 17.41$, and degrees of freedom of the asymptotic $\chi^2$ distribution is calculated as $d.f. = 168 - 154 = 14$, resulting in a p-value of 0.235, which implies that the goodness of fit of model in (b) is not significantly different from the model in (a). Since the model in (b) is more parsimonious, we choose the model in (b) over the model in (a).

**Problem 3(c).**

From the model results from (b), we would like to estimate the odds ratio of diet and Medical Care, for each ethnic group. When ethnicity is white ($E = 1$), we have that,

$$\widehat{OR}_{DC|E=1} = \exp(4\hat{\lambda}^{DC}_{27(11)} + 4\hat{\lambda}^{EDC}_{127(111)}) = \exp(4 \times 0.1313 + 4 \times 0.1036) = 2.56.$$

Then, we need to calculate the variance of the log odds ratio estimator:

$$\begin{aligned} Var(4\hat{\lambda}^{DC}_{27(11)} + 4\hat{\lambda}^{EDC}_{127(111)}) &= 16[Var(\hat{\lambda}^{DC}_{27(11)}) + Var(\hat{\lambda}^{EDC}_{127(111)}) + Cov(\hat{\lambda}^{DC}_{27(11)}, \hat{\lambda}^{EDC}_{127(111)})] \\ &= 16(0.0421^2 + 0.0572^2 + 2(-0.000275)) = 0.0719. \end{aligned}$$

The 95% confidence interval for the log odds ratio is,

$$(\log \widehat{OR}_{DC|E=1} - 1.96se(\log \widehat{OR}_{DC|E=1}), \log \widehat{OR}_{DC|E=1} + 1.96se(\log \widehat{OR}_{DC|E=1})) = (0.414, 1.465).$$

Exponentiating we obtain a 95% confidence interval for the conditional odds ratio: (1.51, 4.33).

By following the same reasoning, we can obtain the estimated conditional odds ratio of diet and Medical Care for African American ($E = 2$) and Native American ($E = 3$), respectively. For African American ($E = 2$), we estimate conditional odds ratio of diet and Medical Care as 0.78, and the corresponding confidence interval as (0.45, 1.36). For Native American ($E = 3$), we estimate conditional odds ratio of diet and Medical Care as 2.41, and the corresponding confidence interval as (1.28, 4.54).

Then, we can interpret the above results as the following:

- Among whites ($E = 1$), conditional on exercise, home glucose monitoring, insulin use, foot care, a person who uses diet to manage diabetes is estimated to be 2.56 (with 95% confidence interval of (1.51, 4.33)) times more likely to use medical care than a person who did not use diet as indicated by consuming five servings of fruits and vegetables on at least one day in the past week.

- Among African Americans ($E = 2$), conditional on exercise, home glucose monitoring, insulin use, foot care, a person who uses diet to manage diabetes is estimated to be 0.78 (with 95% confidence interval of (0.45, 1.36)) times more likely to use medical care than a person who did not use diet as indicated by consuming five servings of fruits and vegetables on at least one day in the past week.

- Among Native Americans ($E = 3$), conditional on exercise, home glucose monitoring, insulin use, foot care, a person who uses diet to manage diabetes is estimated to be 2.41 (with 95% confidence interval of (1.28, 4.54)) times more likely to use medical care than a person who did not use diet as indicated by consuming five servings of fruits and vegetables on at least one day in the past week.

**Problem 3(d).**

Table 5: Summary of Estimated Odds Ratios and 95% Confidence Intervals

| Odds Ratio | Estimated OR | 95% CI | |
|---|---|---|---|
| $\widehat{OR}_{XI}$ | 0.712 | 0.503 | 1.007 |
| $\widehat{OR}_{GI}$ | 3.222 | 2.257 | 4.599 |
| $\widehat{OR}_{DF}$ | 1.580 | 1.163 | 2.147 |
| $\widehat{OR}_{XF}$ | 1.538 | 1.133 | 2.088 |
| $\widehat{OR}_{GF}$ | 1.395 | 1.025 | 1.898 |
| $\widehat{OR}_{IC}$ | 1.936 | 1.364 | 2.746 |

In table 5, we present a summary of the odds ratio estimates, 95% confidence intervals corresponding to each two-factor interaction term that does not involve ethnicity.

Now, we can provide the interpretations as the following:

- Conditional on ethnicity, diet, home glucose monitoring, foot care, and medical care, a person who exercises to manage diabetes is estimated to be 0.712 (with 95% confidence interval of (0.503, 1.007)) times more likely to use insulin than a person who did not exercise.

- Conditional on ethnicity, diet, exercise, foot care, and medical care, a person who uses home glucose monitoring to manage diabetes is estimated to be 3.222 (with 95% confidence interval of (2.257, 4.599)) times more likely to use insulin than a person who did not use home glucose monitoring to manage diabetes.

- Conditional on ethnicity, exercise, home glucose monitoring, insulin use, and medical care, a person who uses diet to manage diabetes is estimated to be 1.580 (with 95% confidence interval of (1.163, 2.147)) times more likely to use foot care than a person who did not use diet as indicated by consuming five servings of fruits and vegetables on at least one day in the past week.

- Conditional on ethnicity, diet, home glucose monitoring, insulin use, and medical care, a person who exercises to manage diabetes is estimated to be 1.538 (with 95% confidence interval of (1.133, 2.088)) times more likely to use foot care than a person who did not exercise.

- Conditional on ethnicity, diet, exercise, insulin use, and medical care, a person who uses home glucose monitoring to manage diabetes is estimated to be 1.395 (with 95% confidence interval of (1.025, 1.898)) times more likely to use foot care than a person who did not use home glucose monitoring to manage diabetes.

- Conditional on ethnicity, diet, exercise, home glucose monitoring, foot care, a person who uses insulin to manage diabetes is estimated to be 1.936 (with 95% confidence interval of (1.364, 2.746)) times more likely to use medical care than a person who did not use insulin to manage diabetes.

**Problem 4(a).**

In this question, we would like to derive the asymptotic covariance matrix of $\hat{\beta}$. We start by writing out the likelihood function as,

$$\begin{aligned} L_s &= \prod_{i=1}^{s} \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \\ &= \prod_{i=1}^{s} \frac{e^{-\exp(\alpha+x_i'\beta)}[\exp(\alpha+x_i'\beta)]^{y_i}}{y_i!}. \end{aligned}$$

Note that the notation subscript of the notation $L_s$ indicates that the sample size is $s$. Then, the log likelihood function is that,

$$l_s = \log L_n = -\sum_{i=1}^{s} \exp(\alpha + x_i'\beta) + \sum_{i=1}^{s} y_i(\alpha + x_i'\beta) - \sum_{i=1}^{s} \log(y_i!).$$

Then, we derive the score functions as the following:

$$\frac{\partial}{\partial \alpha} l_s = -\sum_{i=1}^{s} \exp(\alpha + x_i'\beta) + \sum_{i=1}^{s} y_i,$$

$$\frac{\partial}{\partial \beta} l_s = -\sum_{i=1}^{s} x_i \exp(\alpha + x_i'\beta) + \sum_{i=1}^{s} y_i x_i.$$

The MLE $[\hat{\alpha}, \hat{\beta}]'$ can be obtained by setting the score functions above to zero. Then we continue to derive the Fisher Information as the following:

$$-\frac{\partial^2}{\partial \alpha^2} l_s = \sum_{i=1}^{s} \exp(\alpha + x_i'\beta) = \sum_{i=1}^{s} \mu_i = n,$$

$$-\frac{\partial^2}{\partial \beta^2} l_s = \sum_{i=1}^{s} x_i \exp(\alpha + x_i'\beta) x_i' = X' D_\mu X = n X' D_\pi X,$$

$$-\frac{\partial^2}{\partial \alpha \beta} l_s = \sum_{i=1}^{s} x_i \exp(\alpha + x_i'\beta) = X'\mu = n X'\pi,$$

and then, the asymptotic covariance matrix for $(\hat{\alpha}, \hat{\beta})$ is that,

$$Var(\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix}) = \begin{bmatrix} n & n\pi' X \\ nX'\pi & nX' D_\pi X \end{bmatrix}^{-1} = \frac{1}{n} \begin{bmatrix} 1 & \pi' X \\ X'\pi & X' D_\pi X \end{bmatrix}^{-1},$$

which is in the form of the inverse of a block matrix. Now, we would like to calculate the inverse of this block matrix. Here, we need to use the following fact from linear algebra that,

the inverse of a $2 \times 2$ block matrix $M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$ ,where $A$ and $D$ are invertible squared matrix, is that,

$$M^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -D^{-1}C(A - BD^{-1}B)^{-1} & (D - CA^{-1}B)^{-1} \end{bmatrix}.$$

Since $\alpha$ is not of inferential interest (i.e., nuisance parameter), we only focus on the the part of the covariance matrix related to $\beta$. By calculating the inverse of the block matrix that we have obtained earlier, we can see that the lower right component, which is just the asymptotic covariance matrix of $\hat{\beta}$, can be expressed as,

$$\frac{1}{n}(X'D_\pi X - X'\pi\pi'X)^{-1} = \frac{1}{n}(X'(D_\pi - \pi\pi')X)^{-1},$$

as desired.

**Problem 4(b).**

First, let us recap some key points on maximum likelihood estimation and asymptotic variance of the multinomial loglinear model. The likelihood function is given as,

$$L_n = n! \prod_{j=1}^{r} [\frac{\exp(x_j'\beta)}{\sum_{i=1}^{r} \exp(x_i'\beta)}]^{y_j} / y_j!$$

The log likelihood function is given as,

$$l_n = \log n! - \sum_{j=1}^{r} \log y_j! + \sum_{j=1}^{r} y_j x_j'\beta - n \log[\sum_{j=1}^{r} \exp(x_j'\beta)].$$

The score function is then derived as,

$$\frac{\partial}{\partial \beta} l_n = y'X - n\pi'X.$$

The MLE can be obtained by setting the score function to zero, and the Fisher information can be derived as,

$$-\frac{\partial^2}{\partial \beta^2} l_n = nX'(D_\pi - \pi\pi')X.$$

Consequently, the asymptotic covariance matrix for the MLE $\hat{\beta}$ is that,

$$Var(\beta) = \frac{1}{n}(X'(D_\pi - \pi\pi')X)^{-1}.$$

Given that the score equation for $\beta$ are identical for the Poisson loglinear model and the multinomial loglinear model based on single multinomial sampling, there are several implications for the statistical inference:

- The maximum likelihood estimators $\hat{\beta}$ are the same under Poisson loglinear model and under multinomial loglinear model.

- As we have shown in part (a), the asymptotic covariances of $\hat{\beta}$ are the same under two types of loglinear model.

- It is permissible to use the Poisson loglinear model for analyzing contingency table generated from single multinomial sampling, because the single multinomial loglinear model for the cell probability can be expressed as a Poisson loglinear model for the expected cell counts, where the multinomial sample size parameter $n$ relates to the intercept parameter in the Poisson loglinear model for cell counts. Specifically, in the multinomial loglinear model, the probability of the $j^{th}$ cell is

$$\pi_j = \frac{\exp(x_j'\beta)}{\sum_{i=1}^{r}\exp(x_j'\beta)},$$

  and the model can be expressed as a model for cell counts because,

$$\log n\pi = \log\{n\frac{\exp(X\beta)}{1_r'\exp(X\beta)}\},$$

$$\log\mu = X\beta + [\log n - \log\{1_r'\exp(X\beta)\}]1_r = X\beta + 1_r\alpha.$$

- The results based on likelihood ratio tests for the Poisson loglinear model are applicable to data generated from single multinomial sampling. The reason is that, the deviance statistic will be the same under both loglinear models, because the $\alpha$ term, which relates purely to the sample size (which is random in Poisson loglinear model), cancels in multinomial loglinear model.

# Appendix 1: Model Results from Problem 3(a) and 3(b)

Table 6: Analysis of Parameter Estimates from the Hierarchical Loglinear Model in Problem 3(a)

| Parameter | Level | Estimate | SE | Chi-square | p-value |
|---|---|---|---|---|---|
| E | 1 | 0.165 | 0.066 | 6.300 | 0.012 |
| | 2 | 0.039 | 0.065 | 0.360 | 0.548 |
| D | 1 | 0.106 | 0.047 | 5.070 | 0.024 |
| E*D | 1 1 | 0.254 | 0.058 | 18.850 | <.001 |
| | 2 1 | 0.054 | 0.059 | 0.840 | 0.359 |
| X | 1 | -0.104 | 0.045 | 5.240 | 0.022 |
| E*X | 1 1 | 0.113 | 0.054 | 4.350 | 0.037 |
| | 2 1 | -0.028 | 0.058 | 0.230 | 0.633 |
| D*X | 1 1 | -0.043 | 0.041 | 1.090 | 0.297 |
| G | 1 | -0.026 | 0.047 | 0.310 | 0.576 |
| E*G | 1 1 | 0.094 | 0.057 | 2.760 | 0.097 |
| | 2 1 | -0.062 | 0.060 | 1.060 | 0.303 |
| D*G | 1 1 | -0.010 | 0.043 | 0.060 | 0.812 |
| X*G | 1 1 | 0.014 | 0.041 | 0.110 | 0.739 |
| I | 1 | -0.479 | 0.047 | 102.350 | <.001 |
| E*I | 1 1 | -0.182 | 0.065 | 7.800 | 0.005 |
| | 2 1 | 0.107 | 0.065 | 2.660 | 0.103 |
| D*I | 1 1 | 0.034 | 0.048 | 0.510 | 0.476 |
| X*I | 1 1 | -0.079 | 0.047 | 2.840 | 0.092 |
| G*I | 1 1 | 0.295 | 0.047 | 39.970 | <.001 |
| F | 1 | 0.027 | 0.046 | 0.340 | 0.560 |
| E*F | 1 1 | -0.103 | 0.055 | 3.500 | 0.062 |
| | 2 1 | 0.100 | 0.058 | 2.960 | 0.085 |
| D*F | 1 1 | 0.116 | 0.041 | 8.020 | 0.005 |
| X*F | 1 1 | 0.115 | 0.040 | 8.250 | 0.004 |
| G*F | 1 1 | 0.083 | 0.042 | 4.010 | 0.045 |
| I*F | 1 1 | -0.026 | 0.047 | 0.310 | 0.576 |
| C | 1 | -0.086 | 0.047 | 3.330 | 0.068 |
| E*C | 1 1 | -0.124 | 0.059 | 4.490 | 0.034 |
| | 2 1 | 0.163 | 0.060 | 7.480 | 0.006 |
| D*C | 1 1 | 0.120 | 0.043 | 7.860 | 0.005 |
| X*C | 1 1 | 0.002 | 0.041 | 0.000 | 0.952 |
| G*C | 1 1 | 0.059 | 0.042 | 1.980 | 0.160 |
| I*C | 1 1 | 0.146 | 0.047 | 9.770 | 0.002 |
| F*C | 1 1 | 0.064 | 0.041 | 2.450 | 0.118 |
| E*D*C | 1 1 1 | 0.104 | 0.057 | 3.310 | 0.069 |
| | 2 1 1 | -0.193 | 0.059 | 10.830 | 0.001 |

Table 7: Analysis of Parameter Estimates from the Reduced Hierarchical Loglinear Model in Problem 3(b)

| Parameter | Level | Estimate | SE | Chi-square | p-value |
|---|---|---|---|---|---|
| E | 1 | 0.159 | 0.065 | 6.070 | 0.014 |
| | 2 | 0.043 | 0.064 | 0.450 | 0.502 |
| D | 1 | 0.098 | 0.042 | 5.340 | 0.021 |
| E*D | 1 1 | 0.235 | 0.057 | 16.850 | <.001 |
| | 2 1 | 0.068 | 0.059 | 1.360 | 0.244 |
| I | 1 | -0.477 | 0.047 | 103.150 | <.001 |
| E*I | 1 1 | -0.155 | 0.061 | 6.390 | 0.012 |
| | 2 1 | 0.092 | 0.063 | 2.160 | 0.142 |
| X | 1 | -0.095 | 0.044 | 4.620 | 0.032 |
| X*I | 1 1 | -0.085 | 0.044 | 3.690 | 0.055 |
| G | 1 | -0.024 | 0.046 | 0.290 | 0.592 |
| G*I | 1 1 | 0.293 | 0.045 | 41.450 | <.001 |
| F | 1 | 0.017 | 0.040 | 0.170 | 0.681 |
| D*F | 1 1 | 0.114 | 0.039 | 8.550 | 0.004 |
| X*F | 1 1 | 0.108 | 0.039 | 7.620 | 0.006 |
| G*F | 1 1 | 0.083 | 0.039 | 4.480 | 0.034 |
| C | 1 | -0.086 | 0.047 | 3.360 | 0.067 |
| E*C | 1 1 | -0.125 | 0.058 | 4.620 | 0.032 |
| | 2 1 | 0.165 | 0.059 | 7.770 | 0.005 |
| D*C | 1 1 | 0.131 | 0.042 | 9.700 | 0.002 |
| I*C | 1 1 | 0.165 | 0.045 | 13.680 | 0.000 |
| E*D*C | 1 1 1 | 0.104 | 0.057 | 3.270 | 0.070 |
| | 2 1 1 | -0.193 | 0.059 | 10.800 | 0.001 |

# Appendix 2: SAS Code

```
*********************************************
          Question 1
*********************************************;

proc import datafile="C:\Users\haoli\Desktop\Bios
765\homework\homework1\Q2dat.xlsx"
        out=q1
        dbms=xlsx
        replace;
run;

data q1;
  set q1;
  x1 = (photoperiod = 16);
  x2 = log(BAP)-log(2.2);
  x1x2 = x1*x2;
run;

title '1(b) ZIP Regression Model';
proc genmod data = q1;
  model root = x1 x2 x1x2 / link=log dist=zip;
  zeromodel x1;
run;

title '1(b) ZINB Regression Model';
proc genmod data = q1;
  model root = x1 x2 x1x2 / link=log dist=zinb;
  zeromodel x1;
run;

title '1(e) MZINB Prerequisite: NB Regression';
proc genmod data=q1;
  model root = x1 x2 x1x2 / dist=negbin;
run;

title '1(e) MZINB Regression Model';
proc nlmixed data= q1 qpoints=15;
parms a0=-4.3805 a1=4.2645 phi=0.5221
      b0= 1.8761 b1= -0.7059 b2=0.0725 b3=-0.1820;
linpinfl = a0 + a1*x1;
psi = 1/(1+exp(-linpinfl));
nu = exp(b0 + b1*x1 + b2*x2 + b3*x1x2);
mu = nu/(1-psi);
alpha = 1/phi;
theta = 1/(1+(mu/alpha));
if root=0 then loglike =log(psi + (1-psi)*(theta**alpha));
else loglike = log(1-psi) + lgamma(root+alpha) - lgamma(alpha)
      + root*log(1-theta)+alpha*log(theta) - lgamma(root+1);
model root ~ general(loglike);
estimate '2.2' exp(b1);
estimate '4.4' exp(b1+0.693147*(b3));
estimate '8.8' exp(b1+1.386294*(b3));
estimate '17.6' exp(b1+2.079442*(b3));
run;
```

```
*********************************************
           Question 2
*********************************************;

*coding: 1=yes, 2=no;
data q2; input j m c count;
if count = 0 then count =1e-20;
cards;
1 2 2 77269
2 1 2 138578
2 2 1 13391
1 1 2 25631
1 2 1 3079
2 1 1 23876
1 1 1 8528
;
run;

title '2(a) No Three Factor Interaction (JM, JC, MC)';
proc catmod data=q2;
  weight count;
  model j*m*c = _response_ / noprofile noresponse noiter p=prob covb;
  loglin j|m j|c m|c;
run;

title '2(a) 2 Two-Factor Interaction Model (JM, JC)';
proc catmod data=q2;
  weight count;
  model j*m*c = _response_ / noprofile noresponse noiter p=prob;
  loglin j|m j|c;
run;

title '2(a) 2 Two-Factor Interaction Model (JM, MC)';
proc catmod data=q2;
  weight count;
  model j*m*c = _response_ / noprofile noresponse noiter p=prob;
  loglin j|m m|c;
run;

title '2(a) 2 Two-Factor Interaction Model (JC, MC)';
proc catmod data=q2;
  weight count;
  model j*m*c = _response_ / noprofile noresponse noiter p=prob;
  loglin j|c m|c;
run;

*********************************************
           Question 3
*********************************************;

libname q3 "C:\Users\haoli\Desktop\Bios 765\homework\homework2\";

data q3;
  set q3.elders2;
  dummy = 1;
run;
```

```
ods output summary = t;
proc means data=q3 n completetypes;
  class E D X G I F C;
  var dummy;
run;

proc sort data=t;
  by E descending D descending X descending G descending I descending F
descending C;
run;

title '3(a) Fit Full Model';
proc catmod data=t order=data;
  weight dummy_N;
  model E*D*X*G*I*F*C = _response_ /noresponse noiter zero=sampling;
  loglin E|D|X|G|I|F|C@2 E*D*C;
run;

title '3(b) Fit Reduced Model';
proc catmod data=t order=data;
  weight dummy_N;
  model E*D*X*G*I*F*C = _response_ /noresponse noiter zero=sampling covb;
  loglin E|D E|I X|I G|I D|F X|F G|F E|C D|C I|C  E*D*C;
run;
```

# Appendix 3: R Code

```
######## Part (b) #########


x <- matrix(c(1,1,1,1,1,1,
              1,1,-1,1,-1,-1,
              1,-1,1,-1,1,-1,
              1,-1,-1,-1,-1,1,
              -1,1,1,-1,-1,1,
              -1,1,-1,-1,1,-1,
              -1,-1,1,1,-1,-1,
              -1,-1,-1,1,1,1),
            ncol = 6, byrow = T)


beta <- matrix(c(-0.7894, -0.1313, -1.2453,  0.1101, 0.1645, 0.5306), ncol = 1)
uncondpi <- exp(x%*%beta)/sum(exp(x%*%beta))
pi222 <- uncondpi[8]
gamma <- pi222/(1-pi222)


######## Part (c) #########


n <- 77269+138578+13391+25631+3079+23876+8528
n222 <- n*gamma
N = n222+n


######## Part (d) #########


I <- diag(1, 8)
```

```
A2 <- rbind(I, matrix(c(1,1,1,1,1,1,1,0), nrow = 1))

A3 <- cbind(I, matrix(c(-1,-1,-1,-1,-1,-1,-1,-1), nrow = 8))

D1 <- diag(as.numeric(exp(x%*%beta)),8, 8)

D2 <- solve(diag(as.numeric(A2%*%exp(x%*%beta)), 9, 9))

D3 <- diag(as.numeric(exp(A3%*%log(A2%*%exp(x%*%beta)))), 8, 8)

B <- D3%*%A3%*%D2%*%A2%*%D1%*%x

library(readxl)

X2d <- read_excel("C:/Users/haoli/Desktop/Bios 765/homework/homework2/2d.xlsx", col_names =
FALSE)

Vb <- as.matrix(X2d)

Var <- B%*%Vb%*%t(B)


######## Part (e) #########


var.n222 <- N*(gamma^2)*pi222*(1-pi222)+(n^2)*Var[8,8]

se.n222 <- sqrt(var.n222)

var.N <- N*((1+gamma)^2)*pi222*(1-pi222)+(n^2)*Var[8,8]

se.N <- sqrt(var.N)
```