BIOS 765 Homework 1

Leo Li
730031954

Problem 1(a).

In this question, we would like to construct an analysis of deviance table that include the following four models:

- Model [0]: Saturated model;
- Model [2]: m_cat, b_cat, mb;
- Model [3*]: m_cat, b_cat;
- Model [4]: m_cat, b.

As in Table 1, we include the likelihood ratio tests and the corresponding p-values for first-order differences (i.e., comparison of deviances for pairs of models from consecutive rows).

Table 1: Analysis of Deviance

model	df	deviance	Δdf	Δdev	p-value
Model [0]: Saturated model	0	0			
			19	11.23	0.916
Model [2]: m_cat, b_cat, mb	19	11.23			
			1	4.30	0.038
Model [3*]: m_cat, b_cat	20	15.53			
			3	2.57	0.463
Model [4]: m_cat, b	23	18.10			

Problem 1(b).

In this question, we would like to present an argument for the case that model [4] is the "best" model. Since in our table, there are some cells with sparse counts, the chi-square approximation to the deviance may not be very good for assessing goodness of fit. Comparing nested model with change in deviance is more justifiable. Let $\Delta Deviance = Deviance[4] - Deviance[3] = 18.1 - 15.53 = 2.57$, and the corresponding p-value is 0.463 which is non-significant, leading to the choice of model 4 as the "best" model.

Now, we would like to construct the likelihood ratio test for the second-order comparison of model [2] versus model [4]. The LRT test statistics is calculated as $LRT = 2 \times (9463.26 - 9459.83) = 6.87$, which asymptotically follows the χ^2 distribution with 4 degrees of freedom, leading to a p-value of 0.143. Since the p-value is less than the significance level of 0.05, we can interpret the result as that the goodness of fit of model [2] is not significantly different from model [4], so that model [4], which is more parsimonious, would suffice to fit the data.

Problem 1(c).

By using β-notation and setting "Under 20" as the reference group for maternal age, we can write the model expression for model [4] as the following:

$$\log(\lambda_i) = \alpha + \beta_1 x_{i1} + ... + \beta_5 x_{i5} + \beta_6 x_{i6}$$

where λ_i represent the number of cases found per number of live births; $x_{i1}, ..., x_{i5}$ are dummy variables for maternal age for "20-24", "25-29", "30-34", "35-39", and "40 and Over", respectively, with "Under 20" as the reference; and x_{i6} is linear birth order (1, 2, 3, 4, or 5). In addition, the fitted model equation for model [4] can be written as:

$$\log(\hat{\lambda}_i) = -7.72 + 0.02x_{i1} + 0.25x_{i2} + 0.79x_{i3} + 1.90x_{i4} + 3.09x_{i5} - 0.03x_{i6}$$

Then, in Table 2, we also present the regression parameter estimates, estimated standard errors, z-scores, and p-values for model [4], in which the definition of each parameter follows from the β-notation for the model expression for model [4], as discussed above.

Table 2: Analysis of Parameter Estimates

Parameter	Estimate	Estimated SE	z-score	p-value
α	-7.724	0.089	-87.275	<.001
β_1	0.022	0.100	0.216	0.829
β_2	0.252	0.102	2.471	0.013
β_3	0.786	0.104	7.544	<.001
β_4	1.899	0.103	18.489	<.001
β_5	3.093	0.106	29.204	<.001
β_6	-0.029	0.016	-1.804	0.072

Problem 1(d).

We would like to compute 95% confidence intervals for incidence density ratios describing the effect on Down Syndrome of each maternal age interval relative to "Under 20", and for one unit increase in birth order. Generally speaking, if we write the Poisson regression model in the general form such that $\log(\lambda_i) = x_i^T \beta$, in which x_i is the 7×1 design matrix for i^{th} cell including the intercept, and β is the corresponding regression coefficients. If we would like to compute the incidence density ratio of cell A relative to cell B, we have that,

$$\text{Incidence Density Ratio} = \frac{\lambda_A}{\lambda_B} = \exp(\log(\lambda_A) - \log(\lambda_B)) = \exp((x_A - x_B)^T \beta).$$

Then, we can use the above reasoning to compute the 95% CI for each incidence density ratio, as presented in Table 3.

Table 3: 95% Confidence Intervals (CI) for Incidence Density Ratios of Different Maternal Age Groups and of One Unit Increase in Birth Order

Incidence Density Ratio	Parameter	Point Estimate	95% CI	
"20-24" Relative to "Under 20"	$\exp(\beta_1)$	1.02	0.84	1.24
"25-29" Relative to "Under 20"	$\exp(\beta_2)$	1.29	1.05	1.57
"30-34" Relative to "Under 20"	$\exp(\beta_3)$	2.19	1.79	2.69
"35-39" Relative to "Under 20"	$\exp(\beta_4)$	6.68	5.46	8.17
"40 and Over" Relative to "Under 20"	$\exp(\beta_5)$	22.04	17.91	27.12
One Unit Increase in Birth Order	$\exp(\beta_6)$	0.97	0.94	1.00

Then, we can interpret the confidence intervals in Table 3 as the following:

- If we assume that the birth orders are held as constant, then the women in the age interval of 20-24 are estimated to have 1.02 times as many Down Syndrome birth as women under 20. The corresponding confidence interval is (0.84, 1.24), which means that if we repeat the study for multiple times, then there is a 95% probability that the confidence intervals constructed in this way will cover the population true incidence density ratio of maternal age interval of 20-24 relative to maternal age of under 20.

- If we assume that the birth orders are held as constant, then the women in the age interval of 25-29 are estimated to have 1.29 times as many Down Syndrome birth as women under 20. The corresponding confidence interval is (1.05, 1.57), which means that if we repeat the study for multiple times, then there is a 95% probability that the confidence intervals constructed in this way will cover the population true incidence density ratio of maternal age interval of 25-29 relative to maternal age of under 20.

- If we assume that the birth orders are held as constant, then the women in the age interval of 30-34 are estimated to have 2.19 times as many Down Syndrome birth as women under 20. The corresponding confidence interval is (1.79, 2.69), which means that if we repeat the study for multiple times, then there is a 95% probability that the confidence intervals constructed in this way will cover the population true incidence density ratio of maternal age interval of 30-34 relative to maternal age of under 20.

- If we assume that the birth orders are held as constant, then the women in the age interval of 35-39 are estimated to have 6.68 times as many Down Syndrome birth as women under 20. The corresponding confidence interval is (5.46, 8.17), which means that if we repeat the study for multiple times, then there is a 95% probability that the confidence intervals constructed in this way will cover the population true incidence density ratio of maternal age interval of 35-39 relative to maternal age of under 20.

- If we assume that the birth orders are held as constant, then the women 40 and over are estimated to have 20.04 times as many Down Syndrome birth as women under

20. The corresponding confidence interval is (17.91, 27.12), which means that if we repeat the study for multiple times, then there is a 95% probability that the confidence intervals constructed in this way will cover the population true incidence density ratio of maternal age of 40 and over relative to maternal age of under 20.

- If we assume that the maternal age intervals are held as constant, then one unit increase in birth order is estimated to result in 0.97 times as many Down Syndrome birth. The corresponding confidence interval is (0.94, 1.00), which means that if we repeat the study for multiple times, then there is a 95% probability that the confidence intervals constructed in this way will cover the population true incidence density ratio for a one unit increase in birth order.

Problem 1(e).

We would like to strengthen the argument such that model [4] is the "best" model by determining the observation with the largest standardized Pearson residual among all observations. Observation #24 (m_cat = 35-39, b = 4) has the largest standardized Pearson residual in absolute value, which equals to 1.63, so there are no outliers under model [4], which implies that model [4] fits well.

The largest standardized Pearson residual from model [3] from Notes 3 is 1.58, which is not very different from that from model [4].

Problem 2(a).

In this question, we use Table 4 to present the mean and variance of the number of roots and sample size for each combination of photoperiod and BAP. From Table 4, we observe that the value of the variance is generally larger than the value of the mean, especially in the situation of 8-hour photoperiod and 2.2 μM of BAP, as well as all four situations under 16-hour photoperiod.

Table 4: Sample Size (Number of Shoots), Mean, and Variance of the Number of Root for Each Combination of Photoperiod and BAP

Photoperiod	8-hour				16-hour			
BAP (μM)	2.2	4.4	8.8	17.6	2.2	4.4	8.8	17.6
n	30	30	40	40	30	30	30	40
mean	5.83	7.77	7.50	7.15	3.27	2.73	3.13	2.45
variance	14.14	7.56	8.46	8.75	16.55	14.75	13.50	8.51

Problem 2(b).

Table 5: Analysis of Parameter Estimates

Effect	Parameter Estimate	Standard Error	p-value
Intercept	1.826	0.088	<.001
log BAP	0.069	0.042	0.097
I(photoperiod = 16)	-0.571	0.157	<.001
logBAP×I(photoperiod = 16)	-0.178	0.077	0.022

We fit the Poisson regression model to the $n = 270$ observations, and in Table 5, we present the parameter estimates, standard errors and p-values. In this situation, the Pearson chi-square goodness-of-fit test is not valid because of excessive zeros.

Problem 2(c).

We fit the same Poisson regression model to the data set with eight observations corresponding to the combinations of photoperiod and BAP, and we use the total number of roots as the outcome and the natural log of the number of shoots as offset. The model is valid considering the sample size, because the counts are sufficiently large. The Pearson chi-square goodness-of-fit test statistics is computed as 9.533, which asymptotically follow a χ^2_4 distribution. Note that the goodness-of-fit statistic is not close to 1, suggesting overdispersion. In other words, the Pearson chi-square goodness-of-fit test suggests a poor fitting of the model. The parameter estimates and standard errors are identical to those model results in part (b), as shown in Table 5.

Problem 2(d).

In this question, we would like to find the "best" negative binomial regression model. We start by fitting a negative binomial regression model where the number of roots is the outcome, and explanatory variables are an indicator variable for photoperiod = 16, the natural log of BAP, and their interaction. However, we later find out that the interaction term in the model is not significant at the two-sided 0.05 level: The Wald test statistic is 1.73, which asymptotically follows a χ^2_1 distribution, resulting in a p-value of 0.189, which is larger than the significance level.

Then, we move on to fit a model without the interaction term. Specifically,

$$\log(\mu_i) = \beta_0 + \sum_{k=1}^{2} \beta_k x_{ik},$$

where μ_i is the expected number of roots for the i^{th} shoot, and x_{i1} and x_{i2} are the corresponding indicator variable for photoperiod = 16 and the natural log of BAP, respectively. The number of roots for the i^{th} shoot, y_i, is assumed to follow $NB(\mu_i, \alpha)$ distribution. Then, the fitted model equation is,

$$\log(\hat{\mu}_i) = 1.976 - 0.910x_{i1} - 0.008xi2,$$

and the dispersion parameter α is estimated as 0.530. Table 6 is used to present the regression parameter estimates, standard errors, Wald chi-square values, and p-values.

Table 6: Analysis of Parameter Estimates

Effect	Parameter	Estimate	SE	Wald Chi-Square	p-value
Intercept	β_0	1.976	0.152	168.26	<.001
I(photoperiod = 16)	β_1	-0.910	0.108	71.34	<.001
log BAP	β_2	-0.008	0.069	0.01	0.905
Dispersion	α	0.530	0.083		

Problem 2(e).

The 95% CI for the dispersion parameter α is estimated as (0.389, 0.721), which contains values substantially different from zero, indicating extra-Poisson variation. We can test the overdispersion by using the likelihood ratio test comparing the negative binomial to the Poisson model with the same covariates: $H_0 : \alpha = 0$ vs $H_1 : \alpha > 0$. The test statistic is calculated as $Q_L = 2(785.83-701.93) = 167.80$, which follows a 50:50 mixture of distribution which is $\frac{1}{2}(\chi^2_0+\chi^2_1)$, and the corresponding p-value is less than 0.01, which lead us to reject the null hypothesis that the dispersion parameter equals to zero. The result can be interpreted as that the effect of overdispersion is significant in the model, so that the Poisson model provides a poor fit.

Problem 2(f).

The estimation of incidence ratio follow the similar reasoning from 1(d), the incidence density ratio for the effects of photoperiod on the mean number of roots can be estimated as $\exp(\hat{\beta}_1) = \exp(-0.910) = 0.40$, and the corresponding 95% confidence interval is (0.33, 0.50). The incidence density ratio can be interpreted as the following: if we assume that the value of BAP is held as constant, then the number of roots for the shoots where the photoperiod equals 16 is estimated to be 0.40 times as those where the photoperiod equals 8. The corresponding 95% confidence interval is estimated to be (0.33, 0.50), which means that if we repeat the study for multiple times, then there is a 95% probability that the confidence intervals constructed in this way will cover the population true incidence density ratio for the effects of photoperiod on the mean numer of roots.

Problem 3(a).

In this question, we would like to derive a matrix expression for the asymptotic variance of the predicted survival rates from the piecewise exponential regression model. From the piecewise exponential regression model, we are given that,

$$\hat{S} = \exp\{A_1 \otimes I_4 \exp(X\hat{\beta}_A)\}$$

Since $\hat{S}$ is comprised of a sequence of functions, then we have that,

$$Var(\hat{S}) = [H(\beta_A)]Var(\hat{\beta}_A)[H(\beta_A)]',$$

where $H(\beta)$ is a product of the first derivative matrices $H_k(\beta)$, where k indicates the k^{th} operation in accordance with chain rule. Then, we define the following functions:

- $f_1 = X\hat{\beta}_A$
- $f_2 = \exp(f_1) = \exp(X\hat{\beta}_A)$
- $f_3 = A_1 \otimes I_4 f_2 = A_1 \otimes I_4 \exp(X\hat{\beta}_A)$
- $f_4 = \exp(f_3) = \exp(A_1 \otimes I_4 \exp(X\hat{\beta}_A))$

Then, $\hat{S} = f_4(f_3(f_2(f_1)))$. By chain rule,

$$\begin{aligned} H(\beta_A) &= H_4(\beta_A)H_3(\beta_A)H_2(\beta_A)H_1(\beta_A) \\ &= D_{\exp(A_1 \otimes I_4 \exp(X\beta_A))} A_1 \otimes I_4 D_{\exp(X\beta_A)} X. \end{aligned}$$

Therefore, an estimator of the asymptotic variance matrix of the predicted survival rate is that,

$$\hat{Var}(\hat{S}) = [D_{\exp(A_1 \otimes I_4 \exp(X\hat{\beta}_A))} A_1 \otimes I_4 D_{\exp(X\hat{\beta}_A)} X] Var(\hat{\beta}_A) [D_{\exp(A_1 \otimes I_4 \exp(X\hat{\beta}_A))} A_1 \otimes I_4 D_{\exp(X\hat{\beta}_A)} X]',$$

where

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix},$$

and

$$\hat{\beta}_A = \begin{bmatrix} -5.228 \\ -0.805 \\ -0.726 \end{bmatrix},$$

and

$$A_1 = \begin{bmatrix} -6 & 0 & 0 \\ -6 & -18 & 0 \\ -6 & -18 & -36 \end{bmatrix},$$

and

$$Var(\hat{\beta}_A) = \begin{bmatrix} 0.0280 & -0.0270 & -0.0069 \\ -0.0270 & 0.0349 & -0.0003 \\ -0.0069 & -0.0003 & 0.0488 \end{bmatrix},$$

and I_4 is the 4 by 4 identity matrix, and D is the diagonal matrix with the corresponding values on the subscript on the diagonal.

Problem 3(b).

In this question, we would like to determine the numerical value of the estimated asymptotic variance matrix of the predicted survival rates for the duodental ulcer data, by following the formula developed in part (a). The estimated asymptotic variance matrix, $\hat{Var}(\hat{S})$, is,

0.0000272	0.0000273	0.0000274	0.0000272	0.0000273	0.0000274	0.0000101	0.0000062	-0.0000012	0.0000272	0.0000273	0.0000274
0.0000273	0.0000418	0.0000669	0.0000273	0.0000418	0.0000669	0.0000059	0.0000036	-0.0000007	0.0000273	0.0000418	0.0000669
0.0000274	0.0000669	0.0001358	0.0000274	0.0000669	0.0001358	-0.0000016	-0.0000010	0.0000002	0.0000274	0.0000669	0.0001358
0.0000272	0.0000273	0.0000274	0.0000272	0.0000273	0.0000274	0.0000101	0.0000062	-0.0000012	0.0000272	0.0000273	0.0000274
0.0000273	0.0000418	0.0000669	0.0000273	0.0000418	0.0000669	0.0000059	0.0000036	-0.0000007	0.0000273	0.0000418	0.0000669
0.0000274	0.0000669	0.0001358	0.0000274	0.0000669	0.0001358	-0.0000016	-0.0000010	0.0000002	0.0000274	0.0000669	0.0001358
0.0000101	0.0000059	-0.0000016	0.0000101	0.0000059	-0.0000016	0.0000148	0.0000255	0.0000457	0.0000101	0.0000059	-0.0000016
0.0000062	0.0000036	-0.0000010	0.0000062	0.0000036	-0.0000010	0.0000255	0.0000534	0.0001057	0.0000062	0.0000036	-0.0000010
-0.0000012	-0.0000007	0.0000002	-0.0000012	-0.0000007	0.0000002	0.0000457	0.0001057	0.0002183	-0.0000012	-0.0000007	0.0000002
0.0000272	0.0000273	0.0000274	0.0000272	0.0000273	0.0000274	0.0000101	0.0000062	-0.0000012	0.0000272	0.0000273	0.0000274
0.0000273	0.0000418	0.0000669	0.0000273	0.0000418	0.0000669	0.0000059	0.0000036	-0.0000007	0.0000273	0.0000418	0.0000669
0.0000274	0.0000669	0.0001358	0.0000274	0.0000669	0.0001358	-0.0000016	-0.0000010	0.0000002	0.0000274	0.0000669	0.0001358

Then, the standard error estimates of the maximum likelihood model predicted survival rate can be obtained by taking square roots of the diagonal entries of $\hat{Var}(\hat{S})$ presented above. Specifically, we have that,

$$\hat{se}(\hat{S}) = \begin{bmatrix} 0.00522 \\ 0.00646 \\ 0.01166 \\ 0.00522 \\ 0.00646 \\ 0.01166 \\ 0.00385 \\ 0.00731 \\ 0.01478 \\ 0.00522 \\ 0.00646 \\ 0.01166 \end{bmatrix}.$$

Problem 4(a).

Since we have that $y_{ij} \sim Poisson(\mu_{ij})$, then the characteristics function for y_{ij} can be written as

$$\phi_{y_{ij}}(t) = \exp[\mu_{ij}(e^{it} - 1)].$$

For $i = 1, ..., s$, and $j = 1, ..., r$, y_{ij}'s are independent Poisson random variables. Then we have that,

$$\phi_{n_{i.}}(t) = \phi_{\sum_{j=1}^{r} y_{ij}}(t) = \prod_{j=1}^{r} \phi_{y_{ij}}(t) = \prod_{j=1}^{r} \exp[\mu_{ij}(e^{it} - 1)] = \exp[(e^{it} - 1) \sum_{j=1}^{r} \mu_{ij}],$$

which is the characteristics function for the random variable that follows $Poisson(\sum_{j=1}^{r} \mu_{ij})$. Then, we have that $n_{i.} = \sum_{j=1}^{r} y_{ij} \sim Poisson(\sum_{j=1}^{r} \mu_{ij})$. Since for $i = 1, ..., s$, and $j = 1, ..., r$, y_{ij}'s are independent, then $n_{i.}$'s are independent as well. Therefore, we have that $n_{i.}$ for $i = 1, ..., s$ are independent Poisson random variables with means $\mu_{i.} = \sum_{j=1}^{r} \mu_{ij}$.

Problem (b).

We can start by considering the random variable $(y_{11}, ..., y_{1r}|n_{1.})$, we have that,

$$f(y_{11}, ..., y_{1r}|n_{1.}) = \frac{f(y_{11}, ..., y_{1r}, n_{1.})}{f(n_{1.})} = \frac{\prod_{j=1}^{r} \frac{\mu_{1j}^{y_{1j}} e^{-\mu_{1j}}}{y_{1j}!}}{\frac{\mu_{1.}^{n_{1.}} e^{\sum_{j=1}^{r} \mu_{1j}}}{n_{1.}!}} = \frac{[\prod_{j=1}^{r} \mu_{1j}^{y_{1j}}] n_{1.}!}{[\prod_{j=1}^{r} y_{1j}!] \mu_{1.}^{\sum_{j=1}^{r} y_{1j}}} = \frac{n_{1.}! \prod_{j=1}^{r} \pi_{1j}^{y_{1j}}}{\prod_{j=1}^{r} y_{1j}!}.$$

By using the similar reasoning, we can generalize the results to any $i = 1, ..., s$, such that,

$$f(y_{i1}, ..., y_{ir}|n_{i.}) = \frac{n_{i.}! \prod_{j=1}^{r} \pi_{ij}^{y_{ij}}}{\prod_{j=1}^{r} y_{ij}!}.$$

Since for $i = 1, ..., s$, and $j = 1, ..., r$, y_{ij}'s are independent, $n_{i.}$'s are independent, and in addition, for $i = 1, ..., s$ and $k = 1, ..., s$, $n_{i.}$ is also independent to any y_{kj} where $i \neq k$. Therefore, we have that,

$$f(y_{11}, ..., y_{1r}, ..., y_{s1}, ..., y_{sr}|n_{1.}, ..., n_{s.}) = \prod_{i=1}^{s} f(y_{i1}, ..., y_{ir}|n_{i.}) = \prod_{i=1}^{s} \frac{n_{i.}! \prod_{j=1}^{r} \pi_{ij}^{y_{ij}}}{\prod_{j=1}^{r} y_{ij}!},$$

as desired.

Appendix: SAS Code

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

proc format;
  value birthf
   1 = '1'
   2 = '2'
   3 = '3'
   4 = '4'
   5 = '5+'
;
  value agef
   1 = 'Under 20'
   2 = '20-24'
   3 = '25-29'
   4 = '30-34'
   5 = '35-39'
   6 = '40 and over';
run;

data A;
   input m_age b_order count risk;
   lnrisk = log(risk);
   m_age2=m_age**2;
   b_ord2=b_order**2;
   mb = m_age*b_order;
   m2b= m_age2*b_order;
   mb2= m_age*b_ord2;
   m2b2=m_age2*b_ord2;
cards;
1 5   0    327
1 4   1   2293
1 3   3  15050
1 2  25  72202
1 1 107 230061
2 1 141 329449
2 2 150 326701
2 3  71 175702
2 4  26  68800
2 5   8  30666
3 1  60 114920
3 2 110 208667
3 3 114 207081
3 4  64 132424
3 5  63 123419
4 1  40  39487
4 2  84  83228
4 3 103 117300
4 4  89  98301
4 5 112 149919
5 1  39  14208
5 2  82  28466
```

```
5 3 108  45026
5 4 137  46075
5 5 262 104088
6 1  25   3052
6 2  39   5375
6 3  75   8660
6 4  96   9834
6 5 295  34392
;
run;

title 'Saturated model [0]';
proc genmod data=a order=data;
  format m_age agef. b_order birthf.;
  class m_age b_order;
  model count = m_age b_order m_age*b_order/dist = poisson link=log
offset=lnrisk;
run;

title 'Model [2]';
proc genmod data=a order=data;
  format m_age agef. b_order birthf.;
  class m_age b_order;
  model count = m_age b_order mb/dist = poisson link=log offset=lnrisk;
run;

title 'Model [3*]';
proc genmod data=a order=data;
  format m_age agef. b_order birthf.;
  class m_age b_order;
  model count = m_age b_order/dist = poisson link=log offset=lnrisk;
run;

title 'Model [4]';
proc genmod data=a order=data;
  format m_age agef. b_order birthf.;
  class m_age (ref="Under 20");
  model count = m_age b_order/dist = poisson link=log offset=lnrisk;
  output out=resid4 pred=pred xbeta=xbeta stdxbeta=std reschi=reschi
resdev=resdev stdreschi=stdreschi;
run;

title 'residuals for model [4]';
proc print data = resid4;
  var pred xbeta std reschi resdev stdreschi;
run;

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

proc import datafile="C:\Users\haoli\Desktop\Bios
765\homework\homework1\Q2dat.xlsx"
        out=q2
        dbms=xlsx
        replace;
run;
```

```
proc means data=q2 mean var;
  class photoperiod bap;
  var root;
run;

data q2;
  set q2;
  logBAP = log(BAP);
run;

title '2(b) Poisson Regression Model';
proc genmod data = q2;
  class photoperiod(ref='8')/param=ref;
  model root = logBAP photoperiod logBAP*photoperiod / d=p link=log;
run;

data q2short;
input root shoots photoperiod bap;
datalines;
175 30 8 2.2
233 30 8 4.4
300 40 8 8.8
286 40 8 17.6
98 30 16 2.2
82 30 16 4.4
94 30 16 8.8
98 40 16 17.6
;

data q2short;
  set q2short;
  logbap = log(bap);
  logshoots = log(shoots);
run;

title '2(c) Poisson Regression Model';
proc genmod data = q2short;
  class photoperiod(ref='8');
  model root = logBAP photoperiod logBAP*photoperiod / d=p link=log
offset=logshoots;
run;

title '2(d) Negative Binomial Regression with Interaction';
proc genmod data = q2;
  class photoperiod(ref='8');
  model root = logBAP photoperiod logBAP*photoperiod / d=negbin link=log;
run;

title '2(d) Negative Binomial Regression without Interaction';
proc genmod data = q2;
  class photoperiod(ref='8');
  model root = logbap photoperiod / d=negbin link=log;
run;

title '2(e) Corresponding Poisson Regression to Test Overdispersion';
proc genmod data = q2;
```

```
  class photoperiod(ref='8');
  model root = logBAP photoperiod / d=p link=log;
run;

proc format;
  value operf
   1 = 'V + D'
   2 = 'V + A'
   3 = 'V + H'
   4 = 'GR';
  value timef
   1 = '0-6  '
   2 = '7-24 '
   3 = '25-60';run;
data A;
   input oper time count risk;
   risk3=risk/1000;
   lnrisk=log(risk);
   obsrate3 = count/risk3 /* observed rate */;
   if time=2 or time=3 then time23=1; else time23=0;
   if oper=3 then VH=1; else VH=0;
cards;
1 1 10 1962
1 2 13 5445
1 3 26 9252
2 1 9  1932
2 2 16 5427
2 3 18 9468
3 1 9  2016
3 2 5  5724
3 3 10 10440
4 1 9 2025
4 2 15 5688
4 3 24 9810
; run;

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

proc IML;
 beta_hat = {-5.2275,-0.8046 , -0.7263};
 A = I(4)@{-6 0 0 ,-6 -18 0, -6 -18 -36};
 X_A = J(12,1,1)||{0 0,1 0, 1 0,0 0,1 0, 1 0,0 1,1 1, 1 1,0 0,1 0, 1 0 };
 f1 = X_A * beta_hat;
 f2 = A * exp(f1);
 f3 = exp(f2);
 f2_star = A*diag(exp(f1));
 f3_star = diag(f3);
 Var_hat_beta = {0.028 -0.02698 -0.006904, -0.02698 0.03490 -0.000309, -
0.006904 -0.000309 0.04881};
 Var_s = f3_star*f2_star*X_A*Var_hat_beta*X_A`*f2_star`*f3_star`;
 diag_var_s = diag(Var_s);
 print Var_s;
quit;
```