STAT 8320 Spring 2015 Assignment 4

Peng Shao 14221765

March 16, 2015

▶ **1.** **Solution.** (a). Define

$$\begin{aligned}
\boldsymbol{Y}_i &= (Y_i1, Y_i2)' \\
\boldsymbol{\beta} &= (\beta_0, \beta_0)' \\
\boldsymbol{b}_i &= (b_{0i}, b_{1i})' \\
\boldsymbol{W}_i &= \begin{pmatrix} 1 & W_{i1} \\ 1 & W_{i2} \end{pmatrix} \\
\boldsymbol{\epsilon}_i &= (\epsilon_{i1}, \epsilon_{i2})' \\
\boldsymbol{D} &= \begin{pmatrix} 4 & 1 \\ 1 & 2 \end{pmatrix} \\
\boldsymbol{\Sigma} &= \begin{pmatrix} \sigma^2 & \\ & \sigma^2 \end{pmatrix}
\end{aligned}$$

Then the model can be written as

$$\boldsymbol{Y}_i = \boldsymbol{\beta} + \boldsymbol{W}_i \boldsymbol{b}_i + \boldsymbol{\epsilon}_i$$

The marginal variance/covariance matrix of $\boldsymbol{Y}_i$ is

$$\begin{aligned}
Var(\boldsymbol{Y}_i) &= Var(\boldsymbol{\beta} + \boldsymbol{W}_i \boldsymbol{b}_i + \boldsymbol{\epsilon}_i) = Var(\boldsymbol{W}_i \boldsymbol{b}_i) + Var(\boldsymbol{\epsilon}_i) \\
&= \boldsymbol{W}_i Var(\boldsymbol{b}_i) \boldsymbol{W}_i' + \boldsymbol{\Sigma} \\
&= \boldsymbol{W}_i \boldsymbol{D} \boldsymbol{W}_i' + \boldsymbol{\Sigma} \\
&= \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 4 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} + \begin{pmatrix} 2 & \\ & 2 \end{pmatrix} \\
&= \begin{pmatrix} 10 & 11 \\ 11 & 18 \end{pmatrix}
\end{aligned}$$

(b) The conditional variane/covariance matrix of $\boldsymbol{Y}_i$ is

$$Var(\boldsymbol{Y}_i|\boldsymbol{b}_i) = Var(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

(c). The hypotheses are

$$H_0 : cov(b_{0i}, b_{1i}) = 0 \quad \text{v.s.} \quad H_a : cov(b_{0i}, b_{1i}) = 0$$

Then statistic is

$$\Lambda = -2(\ell(Reduced\ Model) - \ell(Full\ Model)) = 426 - 420 = 6 \sim \chi^2(1)$$

Then the value of statistic is greater than $\chi^2_{0.95}(1) = 3.84$ with the P-value=0.014. Thus, we reject the null hypothesis, that is, we should favor the model for which the random effects parameters are dependent.

▶ **2. Solution.** (a). Because $Y \sim \text{BIN}(50, p_i)$,where p_i is the rate of germination, we consider a Binomial Regression Model. Then

$$f(y_i; 50, p_i) = \binom{50}{y_i} p_i^{y_i}(1-p_i)^{50-y_i}$$

The exponential form of probability mass function is

$$f(y_i; \theta, \phi) = \exp\left[\frac{\theta_i y_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right]$$

where

- $\theta_i = \text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$
- $b(\theta) = 50\log(1 + e^{\theta})$
- $a(\phi) = 1$
- $c(y_i, \phi) = \log\binom{50}{y_i}$

This is the random component of the model.

Since y_i can be treated as a grouped Bernoulli variable, the link function is

$$\eta_i = g(p_i) = \text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$$

where η_i is linear predictor.

Because there are two factor and 4 levels for each, the linear predictor is

$$\eta_i = \mu + \sum_{j=1}^{3} \beta_j X_{ij} + \sum_{k=1}^{3} \gamma_k Z_{ik} + \sum_{j=1}^{3} \sum_{k=1}^{3} \delta_{jk} X_{ij} Z_{ik}$$

where

- $X_{ij} = 1$, if jth level of temperature; $X_{ij} = -1$, if 4th level of temperature; otherwise $X_{ij} = 0$.
- $Z_{ik} = 1$, if kth level of concentration; $Z_{ik} = -1$, if 4th level of temperature; otherwise $Z_{ik} = 0$.

The link function and the linear predictor is the systematic component of the model.

(b) Yes. The degree of freedom of model 4 is 57, and the degree of freedom of model 5 is 48. The hypotheses are

$$H_0: \text{ model 4} \quad \text{v.s.} \quad H_a: \text{ model 5}$$

Because

$$T = \frac{D(model\ 4) - D(model\ 5)}{\phi} = \frac{148.1 - 55.6}{1} = 92.5 > \chi^2_{0.95,57-48} = 16.92$$

then we reject H_0, that is, the interaction term is needed in the model.

(c) The hypotheses are

$$H_0: \text{ model 5 is good enough to fit data} \quad \text{v.s.} \quad H_a: \text{ model 5 is lack of fit}$$

The statistic is

$$D^* = 2(\ell(saturated\, model) - \ell(model\, 5)) = D(model\, 5) = 55.6 < \chi^2_{0.95,64-16} = 65.17077$$

Thus, we fail to reject null hypothesis, i.e., the mode 5 is good enough to fit data. In addition, the dispersion parameter is $\phi = 55.6/48 = 1.158$, which is very close to 1, so the binomial model is appropriate.

To sum up, there is no significant evidence of lack of fit for model 5.

(d) Because

$$\begin{aligned}\operatorname{logit}(\hat{p}_{X_1+1}) - \operatorname{logit}(\hat{p}_{X_1}) &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 \\ &\quad - \beta_0 + \beta_1(X_1+1) + \beta_2 X_2 + \beta_3 (X_1+1) X_2 \\ &= \beta_1 + \beta_3 X_2\end{aligned}$$

Then the odds ratio for X_1 is

$$\frac{\text{odds}[\hat{p}_{X_1+1}]}{\text{odds}[\hat{p}_{X_1}]} = e^{\beta_1 + \beta_3 X_2}$$

The $\exp(\beta_1)$ here is that the odds ratio for X_1 when X_2 equals zero, while $\exp(\beta_1)$ in the models without interaction means odds ratio for X_1 just holding X_2 unchanged. So the meaning is different.

▶ 3. Solution.

(a). From the scatterplots (Figure.1), we can see that the displacement, horsepower, weight and acceleration seem to have effects the level of gas milage because most one of mpg01 distributed in low level of displacement, horsepower and weight, or the high level of acceleration. But these variables have some multicollinearity, so maybe only one or two variable of them will be selected into the model.

From the boxplots (Figure **??**), the boxes of variable acceleration, model year and origin have some overlap, so this means that these variables may not have significant effects on the gas milage, then they may not be selected into the models.

That is all that can be found from the plots.

Figure 1: Scatterplots of Variables

Missing File ./png/scatter.png

(b). Since we do not have too may variables, just to be safe, I will ignore the results from part(a), and still treat all independent variables as potential predictors to select the appropriate ones. I use the 'selection=' option of model statement to preform a predictor selection, as the code below,

```
PROC LOGISTIC DATA=q31 DESCENDING;
class cylinders modelyear origin;
```

```
MODEL mpg01=cylinders displacement horsepower weight
 acceleration modelyear origin/selection=stepwise;
RUN;
```

where I treat "cylinders", "modeler" and "origin" as indictor variables. The variable of selection result is {cylinders, modelyear, weight}, which is inn accordence with what we got from part(a), except "modelyear" becoming a significant predictor. So the model becomes

$$Model\ 1:\ Gas\ Mileage = Overallmean + Weight + Cylinder + Modelyear$$

with the 17 parameters (Figure 2).

Figure 2: Predictors Selected

Missing File ./lst/h3re34.lst

Then I score the training dataset under this model, the result shows that the error rate is only 5.56%, so I have to say that it is really a good prediction model. But I still try the other models,

$$\begin{aligned}
&Model\ 2:\ Gas\ Mileage = Overallmean + Weight + Cylinder + Modelyear + Interaction\\
&Model\ 3:\ Gas\ Mileage = Overallmean + Weight + Cylinder\\
&Model\ 4:\ Gas\ Mileage = Overallmean + Weight + Horsepower + Modelyear
\end{aligned}$$

Here, Model 2 extends Model 1 by adding interaction; Model 3 drop the variable model year, since we did not want to consider this variable from part(a); and Model 4 is obtained from model selection procedure without treating discrete variables as indicator variables (which may make the model hard to interpret).

Figure 3: Score for Training Dataset

Missing File ./lst/h3re38.lst

I scored data under different models. Model 4 and Model 3 has worse performance in predicting, with the error rates about 10% and 12% respectively. So compared to Model 1, they cannot be better model. Model 2, however, has a little lower error rate about 4%, which improves Model 1 slightly. But Model 2 has two problems: 1) the number of parameters become 57. This may cause model overfitted the data because of two many variable. The decision boundary of the model becomes more complicated and make the prediction for new data more uncertain; 2) the interaction term is very hard to be interpreted. For example, it may be a little difficult to understand than why the number of cylinders has some interaction with model year. Thus, I prefer the Model 1 to be the best model for this problem

The confusion matrix of this model is Figure 4. Then

$$TP = 160, \quad FN = 11, \quad FP = 8, \quad TN = 163$$

$$\begin{aligned}
\text{sensitivity} &= \frac{TP}{TP + FN} = \frac{160}{160 + 11} = 93.57\% \\
\text{specificity} &= \frac{TN}{TN + FP} = \frac{163}{163 + 8} = 95.32\% \\
\text{Type I error} &= \frac{FP}{TN + FP} = \frac{8}{163 + 8} = 4.68\% \\
\text{Type II error} &= \frac{FN}{TP + FN} = \frac{11}{160 + 11} = 6.43\%
\end{aligned}$$

The sensitivity and specificity are both very good, which means that the model has a good ability to determine high and low gas mileage.

Figure 4: Confusion Table for Training Dataset

Missing File ./lst/h3re39.lst

Finally, the deviance of this model is 90.7560 with the degree of freedom 318 (Figure 5), so the dispersion parameter is 0.2854. The assumption of this model is the data follows Bernoulli distribution, i.e., ϕ equals 1. Thus there is much less variation than we expected.

Figure 5: Model Fitness Summary

Missing File ./lst/h3re31.lst

(c) The code here is

```
PROC LOGISTIC DATA=q31 DESCENDING outmodel=model;
class cylinders modelyear origin;
MODEL mpg01=cylinders weight modelyear/ scale=none AGGREGATE;
score data=q31 out=score fitstat;
RUN;
PROC LOGISTIC inmodel=model;
score data=q3t1 out=Scoret fitstat;
RUN;
proc freq data=scoret;
tables F_mpg01* I_mpg01/nocol nocum norow nopercent;
run;
```

The output are listed as below (Figure 6 and Figure 7). The overall error rate is 10%. From the confusion matrix (Figure 7), the sensitivity is 83%, and specificity is 96.15%, they are both good, and the ability of determining low gas mileage of this model is better than that of determining high gas mileage.

Figure 6: Score for Test Dataset

Missing File ./lst/h3re40.lst

Figure 7: Confusion for Test Dataset

Missing File ./lst/h3re41.lst

▶ **4. Solution.** (a) We consider a Poisson regression model. The random component of this model is

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - e^{\theta}}{\phi} + c(y,\phi)\right\}$$

the link function is

$$g(\mu) = \log(\mu) = \eta$$

and the linear predictor is

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

where Y is fish count and $X_i = 1$, if ith macrohabitat; otherwise $X_i = 0$.

From SAS output (Figure 8), firstly because the type 1 and type 3 analysis is significant, which means that the factor microhabitat has significant effect. Then we study on the estimate summary table. The intercept is the estimate of the effect of macrohabitat 4, the β_is are the difference between the effect of macrohabitat i and the effect of macrohabitat 4. Because the estimates of β_is are all greater than 0, we can know that the effects of macrohabitat 1,2 and 3 are all greater than that of 4, and the order is $macrohabitat\ 1 > macrohabitat\ 2 > macrohabitat\ 3 > macrohabitat\ 4$. Then since the P-value of the estimate of β_2 is greater than 0.05, there is no significance difference between macrohabitat 2 and macrohabitat 4, which also means no significance difference between macrohabitat 2 and macrohabitat 3 or macrohabitat 3 and macrohabitat 4. However, the P-value of the estimate of β_1 is less than 0.05, which means that there is a significance difference between macrohabitat 1 and macrohabitat 4. Actually, we can construct the models under other coding schemes like reference to "1" for indicator variables, and we can get to know that the significance difference is only between macrohabitat 1 and macrohabitat 4, while there is no significant differences between any other pair of macrohabitats.

Figure 8: Estimate Summary

Missing File ./lst/h3re47.lst

The dispersion parameter is 9.8390, which is much greater than 1. It implies that the data is over dispersion based on this Poisson model. This makes us to think about other models like Extra-Poisson model or zero inflated Poisson model to fix this problem.

(b). Because of the excess of zero (Figure 9), we consider the zero-inflated Poisson model,

$$f(y) = \begin{cases} \omega + (1-\omega)e^{-\lambda} & \text{for } y = 0 \\ (1-\omega)\frac{\lambda^y e^{-\lambda}}{y!} & \text{for } y = 1, 2, ... \end{cases}$$

and

$$h(\omega_i) = \boldsymbol{z_i}\boldsymbol{\gamma}$$
$$g(\lambda_i) = \boldsymbol{x_i}\boldsymbol{\beta}$$

In this problem, $h(\cdot)$ is logit function, $g(\cdot)$ is log function, X is macrohabitat, and Z is gear type. Here, the macrohabitat is treated as count portion of the model, and the gear type is treated as zero-inflation portion in this model. From the output, we can see that P-values of estimates of γ_3 an γ_5 are less that 0.05, which means that type 2 gear and type 4 gear are significantly different from type 5 gear, so we can think that gear type has a significant effect, or we can say the zero-inflated model is significantly better that the pure Poisson regression model. We can also perform a Vuong test (Figure 10, since these two are not nested models which can be tested by full-reduced model test. From the output, we can see that P-values of three statistics are all less than 0.05, which mean one of the two models is closer to the true model and the test prefer the ZIP model.

Figure 9: Histogram of Fish Count

Missing File ./png/h3re9.png

Figure 10: Test Result for ZIP Model

Missing File ./lst/h3re66.lst