STAT 239 Homework 2

Tarek Tohme

April 29th, 2018

1 Challenger Disaster

Question 1 A logistic model was used to predict the probability of a failure as a function of temperature and pressure. The response is modified to be true if at least one o-ring failed, and false otherwise.

logReg1 is the model with only temperature as a predictor, and logReg2 the one with both temperature and pressure. For logReg1 the estimator was -0.2322 0.1082, which means for every unit increase in temperature, the response decreases by 0.2322, which means the odds of failure get divided by about 1.25. For logReg 2 the estimate was -0.2415 0.1097.

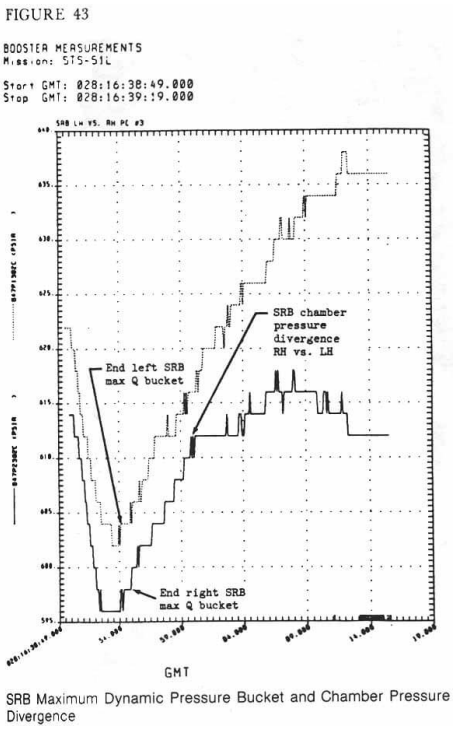

Figure 1: Telemetry data from NASA showing internal pressures of both SRBs during the flight. According to NASA telemetry data, pressure was 610 psi at time of o-ring failure. (source: www.spaceflightnow.com)

There is little difference between both models in terms of deviance and temperature estimates. This shows that pressure doesn't have a strong effect on the model's fit. The responses for some relevant temperatures and pressures are shown in Figure 1: the estimated probability of a failure is 0.999997 according to logReg2, and 0.999609 as predicted by logReg1. According to NASA telemetry data, the internal pressure of the right

SRB (the one that got fractured and caused the explosion) was 610 psi just before the pressure started to drop abnormally. However, even with 50 psi as a predictor, the estimated probability is still 0.999357.

Question 2 The following figures show the predicted probabilities of failure and success for LDA and QDA.

```
> lda1
Call:
lda(fail ~ temp + pres, data = train2)

Prior probabilities of groups:
        0         1
0.93478261 0.06521739

Group means:
      temp     pres
0 69.96899 143.7984
1 63.77778 172.2222

Coefficients of linear discriminants:
             LD1
temp -0.138210810
pres  0.007756029
> predict(lda1, data.frame(temp = 31, pres = 610), type = "response")
$class
[1] 1
Levels: 0 1

$posterior
            0         1
1 0.001584903 0.9984151
```

Figure 2: LDA

```
> qda1
Call:
qda(fail ~ temp + pres, data = train2)

Prior probabilities of groups:
         0         1
0.93478261 0.06521739

Group means:
      temp     pres
0 69.96899 143.7984
1 63.77778 172.2222
> predict(qda1, data.frame(temp = 31, pres = 610), type = "response")
$class
[1] 1
Levels: 0 1

$posterior
             0         1
1 2.720642e-05 0.9999728
```

Figure 3: QDA

According to all three models, the chance of failure given the conditions of the launch were overwhelmingly high.

2 Titanic Disaster

Question 1 Figure 4 below shows four separate tables where each corresponds to a combination of sex and age, and the ratios of surv vs m against class for each table . By visual comparison, the predicted values don't match the prior probabilities very well, except for adult females. Also, in the frequencies plot in Figure [ ], we see that sex and age are noticeably correlated: there is a difference in the way sex affects the frequency of survival when age is varied, holding class fixed. For children, sex doesn't affect survival at all, but for adults it does. The range of survival frequencies for adult males was 0.1 to 0.35 whereas for adult females it was 0.45 to 0.95. Therefore, this model isn't very adequate without at least one interaction parameter.

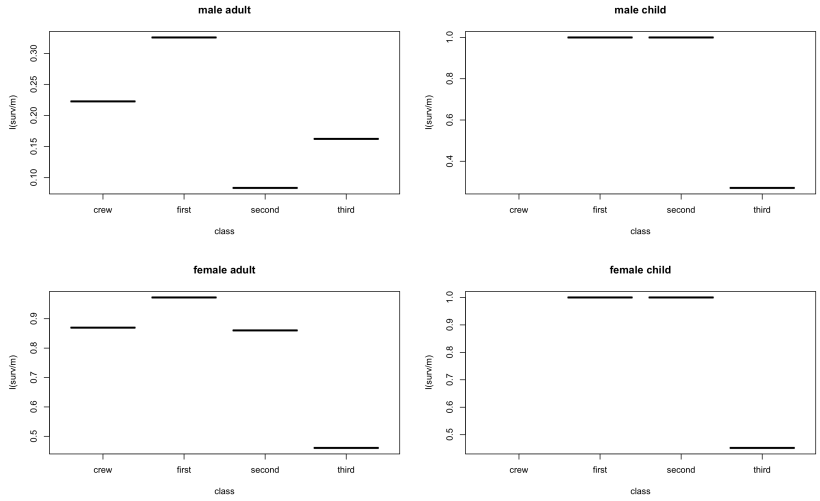

Figure 4: Prior probabilities of each combination of predictors

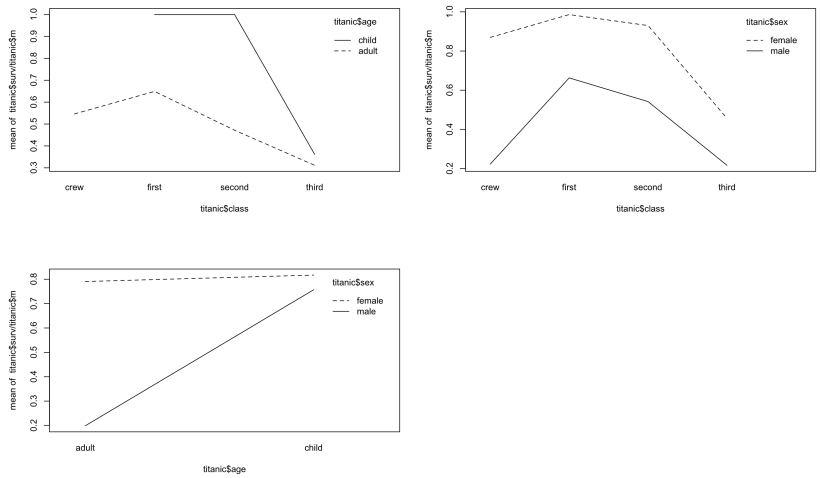

Figure 5: Interaction plots

Figure 5 shows interaction plots of all variables. They clearly show an interaction between class and age, and age and sex variables. This suggests that a model which contains only class, age and sex as predictors is unsuitable for this dataset.

Question 2 logModel2 is a logistic regression model that contains class, age, sex along with all three two-factor interactions.

```
> Anova(logModel2, type=2)
Analysis of Deviance Table (Type II tests)

Response: cbind(surv, m - surv)
            LR Chisq Df Pr(>Chisq)
classf        120.73  3  < 2.2e-16 ***
agef           20.34  1  6.486e-06 ***
sexf          359.37  1  < 2.2e-16 ***
classf:sexf    65.01  3  4.984e-14 ***
classf:agef    37.26  2  8.101e-09 ***
agef:sexf       1.69  1     0.1942
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 6: Analysis of Deviance for logModel2

An ANOVA test on logModel2 shows that removing the age:sex interaction term from the model results in a residual deviance of 1.69, which still indicates a good fit. Dropping class:age produces a residual deviance of 44.2, and dropping class:sex produces a residual deviance of 66.7. This shows that compared to the other interaction terms, age:sex doesn't contribute much to the log-likelihood function of the model. The command used was Anova(logModel2, type=2), which computes a Likelihood Ratio chi-squared test for each addition of a new predictor, respecting the principle of marginality (adding each predictor before any interaction terms it appears in). The p-value of the LR chi-squared test of age:sex was 0.1942, which means we can safely reject the alternative hypothesis (the age:sex coefficient is non-zero) at a threshold of 5%.

A stepwise comparison of the models was performed using the command step(logModel2). The results show that dropping age:sex decreases the AIC of the model from 70.6 to 70.3. This isn't a drastic improvement, but it confirms that this interaction term doesn't improve the model significantly.

The speculation in part 1 about age and sex being correlated may be justified if the variation in the age:sex interaction term was affected by other variables. In this case, the added variable plot of age:sex would be very spread out, indicating that the variation in the response not due to the other predictors is poorly explained by the variation of age:sex not due to other predictors. In other words, including the effect of other predictors on the variation of age:sex might better explain the variation of the response. This is confirmed in figure 7.

From the estimates obtained in the final model (figure 8), we see that first class passengers had a log-odds ratio of surviving of 1.658, and those in third class had a log-odds ratio of -2.115. This means that the odds of surviving among first class passengers was 5.25 times that of the other classes, while those of third class were 0.12 times that of the other classes. These odds translate to a 13% chance of survival for third class, and 84% chance for first class passengers, knowing that the prior probability of survival for the entire population was 32%.

According to the model, children seem to have been equally likely to survive between first and third classes. The coefficient for the firstclass/child dummy variable has a p-value in excess of 0.99, so does the secondclass/child dummy variable. This is confirmed in an ANOVA test using the ungrouped data: removing the class:age variable from the model results in a deviance of only 0.483. Note that class:age was a meaningful predictor with the ungrouped dataset. This is because in the ungrouped dataset, age is a continuous variable, whereas it was a categorical variable in the grouped dataset. While in both the grouped and ungrouped dataset the coefficients for class:age were close to zero and their p-values were high (0.998 and 0.488 resp.), keeping the interaction between class and age in the grouped dataset yields a better model fit. Looking at

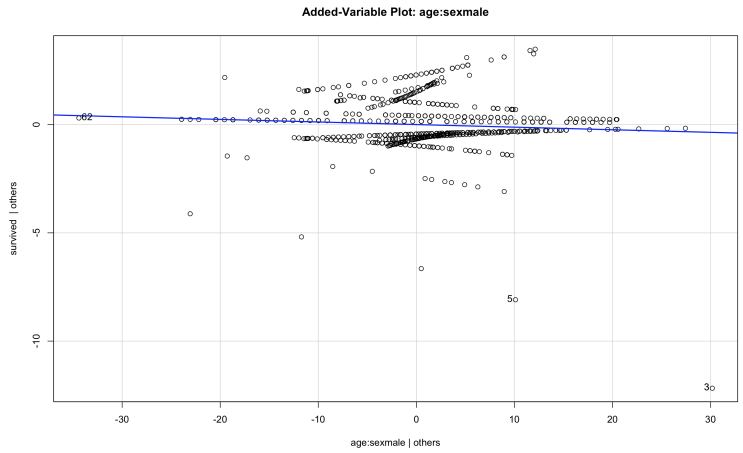

Figure 7: Added variable plot for age:sex interaction variable

```
Coefficients: (1 not defined because of singularities)
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            1.897e+00  6.191e-01   3.064 0.002183 **
classffirst            1.658e+00  8.003e-01   2.072 0.038264 *
classfsecond          -8.004e-02  6.876e-01  -0.116 0.907325
classfthird           -2.115e+00  6.370e-01  -3.319 0.000902 ***
agefchild              3.379e-01  2.692e-01   1.255 0.209391
sexfmale              -3.147e+00  6.245e-01  -5.039 4.68e-07 ***
classffirst:sexfmale  -1.136e+00  8.205e-01  -1.385 0.166162
classfsecond:sexfmale -1.068e+00  7.466e-01  -1.431 0.152539
classfthird:sexfmale   1.762e+00  6.516e-01   2.704 0.006860 **
classffirst:agefchild  2.242e+01  1.650e+04   0.001 0.998915
classfsecond:agefchild 2.442e+01  1.301e+04   0.002 0.998502
classfthird:agefchild         NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 671.9622  on 13  degrees of freedom
Residual deviance:   1.6854  on  3  degrees of freedom
AIC: 70.306
```

Figure 8: Summary of final model, with only class:age and class:sex interactions

the prior probabilities shows clearly that children in third class had a much lower survival rate than the rest, but this pattern isn't captured by the model.

Theoretical Questions

4.2 – a) The LDA decision rule is defined as

$$G(x) = \text{argmax}_k \, \delta_k(x)$$

$$\delta(x) = x^T \hat{\Sigma}^{-1} \hat{\mu}_k - \frac{1}{2} \hat{\mu}_k^T \hat{\Sigma}^{-1} \hat{\mu}_k + \log \pi_k \qquad (4.10 \text{ in ESL})$$

Therefore $G(x) = 2$ iff. $\text{argmax}_k (\delta_1, \delta_2) = 2$

$$\Leftrightarrow \delta_2 > \delta_1$$

$$\Leftrightarrow x^T \hat{\Sigma}^{-1} (\hat{\mu}_2 - \hat{\mu}_1) - \frac{1}{2} (\hat{\mu}_2 - \hat{\mu}_1)^T \hat{\Sigma}^{-1} (\hat{\mu}_2 - \hat{\mu}_1) + \log \left(\frac{N_2/N}{N_1/N} \right) > 0$$

$$\Leftrightarrow x^T \hat{\Sigma}^{-1} (\hat{\mu}_2 - \hat{\mu}_1) > \frac{1}{2} (\mu_2 - \mu_1)^T \hat{\Sigma}^{-1} (\hat{\mu}_2 - \hat{\mu}_1) + \log \left(\frac{N_2}{N_1} \right)$$

b) $\hat{\beta} = (X^T X)^{-1} X^T Y$

$$X^T X = \begin{pmatrix} 1_{N_1}^T & 1_{N_2}^T \\ X_1^T & X_2^T \end{pmatrix} \begin{pmatrix} 1_{N_1} & X_1 \\ 1_{N_2} & X_2 \end{pmatrix} = \begin{pmatrix} N_1 + N_2 & N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2 \\ N_1 \hat{\mu}_1^T + N_2 \hat{\mu}_2^T & X_1^T X_1 + X_2^T X_2 \end{pmatrix}$$

$$\sum_{k=1}^{K} \sum_{g_i = k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T = N_1 \hat{\mu}_1 \hat{\mu}_1^T + N_2 \hat{\mu}_2 \hat{\mu}_2^T$$

c) $(\hat{\mu}_2 - \hat{\mu}_1)^T$ is a $1 \times p+1$ vector and β is a $(p+1) \times 1$ vector, therefore $(\hat{\mu}_2 - \hat{\mu}_1)^T \beta$ is a scalar. Therefore $\hat{\Sigma}_B \beta$ is proportional to $(\hat{\mu}_2 - \hat{\mu}_1)$.

d)

4.3 - LDA using $\hat{Y}$ is identical to LDA in the original space

iff. $G_k(x) = G_k(y) \quad \forall k \in K$

Let k be some class in the set K, and let

$$\delta_k(x) = x^T \Sigma^{-1} \hat{\mu}_k - \frac{1}{2} \hat{\mu}_k^T \Sigma^{-1} \hat{\mu}_k$$

be the discriminant function for LDA with X as a predictor.

Want to prove that

$$\delta_k(x) \text{ is maximal} \iff \delta_k(y) \text{ is maximal} \quad (*)$$

$\boxed{\Rightarrow}$

Let $\ell \in K$ s.t $\ell \neq k$. We have then

$$\delta_k(x) > \delta_\ell(x)$$

$$\iff x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k > x^T \Sigma^{-1} \mu_\ell - \frac{1}{2} \mu_\ell^T \Sigma^{-1} \mu_\ell + \log \pi_\ell$$

$$\iff \log\left(\frac{\pi_k}{\pi_\ell}\right) - \frac{1}{2}(\hat{\mu}_k + \hat{\mu}_\ell)^T \Sigma^{-1}(\hat{\mu}_k - \hat{\mu}_\ell) + x^T \Sigma^{-1}(\hat{\mu}_k - \hat{\mu}_\ell) > 0$$

Since $\hat{\mu}_i = 1_{N_i}^T X_i / N_i$, the sample means $\hat{\mu}_i$ can be transformed by B, just like any other point in $\mathbb{R}^p$.

$$\iff -\frac{1}{2}(\hat{\mu}_k + \hat{\mu}_\ell)^T \Sigma^{-1}(\hat{\mu}_k - \hat{\mu}_\ell) + x^T \Sigma^{-1}(\hat{\mu}_k - \hat{\mu}_\ell) > 0$$

$$\iff x^T \Sigma^{-1}(\hat{\mu}_k - \hat{\mu}_\ell) > \frac{1}{2}(\hat{\mu}_k + \hat{\mu}_\ell)^T \Sigma^{-1}(\hat{\mu}_k - \hat{\mu}_\ell)$$

~~since~~ multiplying both sides by $(\hat{\mu}_k - \hat{\mu}_\ell)^{-1} \Sigma$

$$\iff x^T > \frac{1}{2}(\hat{\mu}_k + \hat{\mu}_\ell)^T$$

$$\iff x^T \hat{B} > \frac{1}{2}(\hat{\mu}_k + \hat{\mu}_\ell)^T \hat{B}$$

$$\iff \hat{y} > \frac{1}{2}(\hat{\mu}_k' + \hat{\mu}_\ell')^T$$

where μ_i' is the average of the transformed $\hat{y}_i$: ~~vector~~ sample:

$$\mu_i' = \left[1_N^T X_i / N_i\right] \hat{B} = \frac{1}{N_i} 1_N^T Y$$

The other direction of (*) can be obtained by applying a similar procedure as for $\boxed{\Rightarrow}$.