Homework 1

Sarah Adilijiang

Problem 3

(a)

The generalized linear model (GLM) for binomial response here is:
likelihood: $P(damage_i = y_i|p_i) = p_i^{y_i}(1-p_i)^{m_i-y_i},\ y_i = 0, 1, ..., m_i$
linear predictor: $\eta_i = \beta_0 + \beta_1 temp_i + \epsilon_i$
link function (logit): $\eta_i = log\frac{p_i}{1-p_i}$

```
# data in R
library(faraway)
data(orings)
model.glm = glm(cbind(damage,6-damage)~temp, family=binomial, data=orings)
sumary(model.glm)
```

```
##              Estimate Std. Error z value  Pr(>|z|)
## (Intercept) 11.662990   3.296263  3.5382 0.0004028
## temp        -0.216234   0.053177 -4.0663 4.777e-05
##
## n = 23 p = 2
## Deviance = 16.91228 Null Deviance = 38.89766 (Difference = 21.98538)
```

```
# data from the paper
temp2 = c(66,70,69,68,67,72,73,70,57,63,70,78,67,53,67,75,70,81,76,79,75,76,58)
damage2 = c(0,1,0,0,0,0,0,0,1,1,1,0,0,2,0,0,0,0,0,0,0,2,0,1)
paper = data.frame(temp2, damage2)
paper = paper[order(temp2), ]

# check if the value of temperature are the same
sum(orings$temp!=paper$temp2)
```

```
## [1] 0
```

```
# check if the number of damage are the same
sum(orings$damage!=paper$damage2)
```

```
## [1] 2
```

```
orings[which(orings$damage!=paper$damage2), ]
```

```
##    temp damage
## 1    53      5
## 18   75      1
```

```
paper[which(orings$damage!=paper$damage2), ]
```

```
##    temp2 damage2
## 14    53       2
## 21    75       2
```

```
# use the paper data to fit the model
model.glm2 = glm(cbind(damage2,6-damage2)~temp2, family=binomial, data=paper)
sumary(model.glm2)
```

```
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.084977   3.052474  1.6659  0.09574
## temp2       -0.115601   0.047024 -2.4584  0.01396
##
## n = 23 p = 2
## Deviance = 18.08633 Null Deviance = 24.23036 (Difference = 6.14404)
```

Answer:

Using the data from the paper, we get $\hat{\beta}_0 = 5.084977$, $\hat{\beta}_1 = -0.115601$, which are the same as shown in the paper.

Using the data from the Faraway package in R, we get $\hat{\beta}_0 = 11.662990$, $\hat{\beta}_1 = -0.216234$. They are different from the paper. Comparing the two datasets, we find that all the temperature values are the the same, however, there are two observations of number of damages are different at temperature 53 and 75. In R, these two data points are (53,5) and (75,1), but in the paper, they are (53,2) and (75,2).

(b)

The generalized linear model (GLM) for binary response here is:
likelihood: $P(event_i = y_i|p_i) = p_i^{y_i}(1-p_i)^{1-y_i},\ y_i = 0\ or\ 1$
linear predictor: $\eta_i = \beta_0 + \beta_1 temp_i + \epsilon_i$
link function (logit): $\eta_i = log\frac{p_i}{1-p_i}$

```
orings$event = as.numeric(orings$damage>0)
model.glm3 = glm(event~temp, family=binomial, data=orings)
sumary(model.glm3)
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 15.04290    7.37863  2.0387  0.04148
## temp        -0.23216    0.10824 -2.1450  0.03196
##
## n = 23 p = 2
## Deviance = 20.31519 Null Deviance = 28.26715 (Difference = 7.95196)
```

Answer:

The binary analysis using the data from the Faraway package in R, we get $\hat{\beta}_0 = 15.04290$, $\hat{\beta}_1 = -0.23216$, which are the same as shown in the paper. Because the binary response here is 1 if there was at one O-ring incident and 0 otherwise. For the two different data points discussed above, they now both become the same data points in the two datasets, which are (53,1) and (75,1). Thus the binary analysis results are the same.

(c)

```
summary(as.factor(orings$damage))
```

```
##  0  1  5
## 16  6  1
```

Answer:

There is also a sparsity problem in the original binomial data. The possible outcomes of binomial responses are (0,1,2,3,4,5,6), however, our data only has 23 data points and most of them (16 data points) have response "0". Within the other 7 data points, 6 of them have response "1", only 1 of them has response "5". Therefore, not only that there are much less data points in response spcace (1,2,3,4,5,6) than "0", but also that there are no data points at all in response spcace (2,3,4,6). Thus the original binomial dataset with mostly zeros also has a sparsity problem.

(d)

```
summary(as.factor(orings$temp))
```

```
## 53 57 58 63 66 67 68 69 70 72 73 75 76 78 79 81
##  1  1  1  1  1  3  1  1  4  1  1  2  2  1  1  1
```

Answer:

With sparse data in logistic GLM models, there might be Hauck-Donner effect. To remedy the sparsity problem in binary data, the first option is to collect more data at each temperature so that there may be more "1"s in the data set. If it's impossible to get more data, we can also try parametric boostrap to generate more data from the origianl data set.

On the other hand, let's look at the temperature variables in the dataset, we see that there are some repeated temperature points, for example, there are 4 points at temperature 70. We can combine these data together and make less "0"s at each temperature, which will also turn the binary data into binomial counts data.

(e)

Answer:

Set the probability of damage in the binomial case as p, and set the probability of event in the binary case as p^*. Since Y=0 iff X=0 in both binary and binomial cases, we have the relationship: $(1-p)^6 = 1-p^* \quad \Rightarrow p = 1-(1-p^*)^{1/6}$

In the binary model, we can get the fitted parameters $\hat{\beta}_0$ and $\hat{\beta}_1$, so $\hat{\eta}_i = \hat{\beta}_0 + \hat{\beta}_1 temp_i$, and $\hat{p^*} = \frac{e^{\hat{\eta}}}{1+e^{\hat{\eta}}}$

Thus, we get $\hat{p} = 1-(1-\hat{p^*})^{1/6}$

Since number of damage: $damage \sim Binomial(6,p)$, so $\hat{damage} = E[Binomial(6,p)] = 6\hat{p}$, thus we can get the expected number of damage from the probability estimated from the binary model.