STT 864 Homework 3

Due on Mar. 4th

1. The table below contains information on 23 (out of 24) pre-Challenger space shuttle flights. (On one flight, the solid rocket motors were lost at sea and so no data are available.) Provided are launch temperatures, t (in ^oF), and a 0-1 response, y, indicating whether there was post-launch evidence of a field joint primary O-ring incident. (O-ring failure was apparently responsible for the tragedy.) $y = 1$ indicates that at least one of the 6 primary O-rings showed evidence of erosion.

Temperature	O-ring Incident	Temperature	O-ring Incident
66	0	67	0
70	1	53	1
69	0	67	0
68	0	75	0
67	0	70	0
72	0	81	0
73	0	76	0
70	0	79	0
57	1	75	1
63	1	76	0
70	1	58	1
78	0		

Treat the response variables, Y_i, as Bernoulli distributed independent launch to launch. Note $\mu_i = E(Y_i) = p_i$ here. We'll model

$$g(\mu_i) = \log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 t_i \text{ for } i = 1, \cdots, n.$$

This is a generalized linear model with Bernoulli response and the "logit" link and it is also known as a logistic regression model. We may use R's glm function to analyze these data. (You may learn about the function in the usual way, by typing ?glm.) Load the MASS package.

Read in the data for this problem by typing

```
> temp<-c(66,70,69,68,67,72,73,70,57,63,70,78,67,53,67,75,70,81,
76,79,75,76,58)
> incidents<-c(0,1,0,0,0,0,0,0,1,1,1,0,0,1,0,0,0,0,0,0,1,0,1)
```

And create and view a logical variable equivalent to the 0-1 response variable by typing

```
> indicate<-(incidents>0)
> indicate
```

(a) We may then fit and summarize a generalized linear model here by typing

```
> shuttle.out<-glm(indicate~temp,family=binomial)
> summary(shuttle.out)
```

The logit is the default link for binomial responses, so we don't need to specify that in the function call. Notice that the $\beta_1 < 0$ is the case where low temperature launches are more dangerous than warm day launches. NASA managers ordered the launch after arguing that these and other data data showed no relationship between temperature and O-ring failure. Was their claim correct? Explain.

(b) glm will provide estimated mean responses (and corresponding standard errors) for values of the explanatory variable(s) in the original data set. To see estimated means $\hat{\mu}_i = \hat{p}_i = \{1 + \exp(-\hat{\beta}_0 - \hat{\beta}_1 t_i)\}^{-1}$ and corresponding standard errors, type

```
> shuttle.fits<-predict.glm(shuttle.out,type="response",
se.fit=TRUE)
> shuttle.fits$fit
> shuttle.fits$se.fit
```

Plot estimated means versus t. Connect those with line segments to get a rough plot of the estimated relationship betweent t and p. Plot "2 $\times$ standard error" bands around that response function as a rough indication of the precision with which the relationship betweent t and p could be known from the pre- Challenger data. Apply the Delta method to obtain the standard error for $\hat{p}_i$. Do you obtain the same standard errors as the above output?

The temperature at Cape Canaveral for the last Challenger launch was 31 ^oF. Of course, hind-sight is always perfect, but what does your analysis here say might have been expected in terms of O-ring performance at that temperature? You can get an estimated 31 ^oF mean and the corresponding standard error by typing

```
> predict.glm(shuttle.out,data.frame(temp=31),se.fit=TRUE,
type="response")
```

2. A Ni-Cad battery manufacturer was interested in finding what set of process conditions produces the smallest fraction of cells with "shorts." For $2 \times 2 \times 2 = 8$ different process set-ups, counts were made of batteries with shorts. The sample sizes (the number of experiments) varied set-up to set-up (from 80 to 100). Factors and their levels in the study were

A- Nylon Sleeve	1-no vs 2-yes
B- Rolling Direction	1-lower edge first vs 2-upper edge first
C- Rolling Order	1-negative first vs 2-positive first

The following R code implements two (binomial) GLM analyses of the manufacturer's data. (The data matrix M has counts of shorts in column 1 and counts of nonshorts in column 2.) The first of the two analyses was done using a logit link and the second used a probit link.

```
> A<-c(1,1,1,1,2,2,2,2)
> B<-c(1,1,2,2,1,1,2,2)
> C<-c(1,2,1,2,1,2,1,2)
> AA<-as.factor(A)
> BB<-as.factor(B)
> CC<-as.factor(C)
> shorts<-c(1,8,0,2,0,1,0,0)
> nonshorts<-c(79,80,90,98,90,89,90,90)
```

```
> M<-matrix(c(shorts,nonshorts),nrow=8)
> contrasts=c("constr.sum","contr.sum")
## This command sets sum contrasts for both main effects
> cbind(M,AA,BB,CC)
> glmout.1<-glm(M~1+AA+BB+CC,family=binomial)
> summary(glmout.1)
> glmout.2<-glm(M~1+AA+BB+CC,family=binomial(link="probit"))
> summary(glmout.2)
```

(a) Does it appear from these data and analyses that any of these 3 factors influence the rate of short production? Is it plausible that the differences between observed rates is "just noise"? Explain. How would you test for it?

Using the following output to answer questions (b) and (c).

```
> predict.glm(glmout.1,type="response",se.fit=TRUE)
> predict.glm(glmout.2,type="response",se.fit=TRUE)
```

(b) Notice that there were 4 of 8 different process set-ups which produced no observed shorts. Which of these would you recommend for future running of the production process? Why?

(c) Give approximate 95% confidence limits for the rate of shorts associated with your choice from part b). Do the estimation first based on the logit, then based on the probit link. How much difference is there between the two set of limits?

3. An engineering student group worked on a project aimed at reducing jams on a large collating machine. They ran the machine at 3 "Air Pressure" settings and 2 "Bar Tightness" conditions and observed

$$y = \text{the number of machine jams experienced in } k \text{ seconds of machine run time.}$$

Their results are below.

Air Pressure	Bar Tightness	y, Jams	k, Run Time
1(low)	1(tight)	27	295
2(medium)	1	21	416
3(high)	1	33	308
1	2(loose)	15	474
2	2	6	540
3	2	11	498

Motivated perhaps by a model that says times between jams under a given machine set-up are independent and exponentially distributed, we will consider an analysis of these data based on a model that says the jam counts are independent Poisson variables. For

$$\mu_{IJ} = \text{the mean count at air pressure } i \text{ and bar tightness } j.$$

suppose that

$$\log \mu_{ij} = \mu + \alpha_i + \beta_j + \log(k_{ij}) \tag{1}$$

Notice that this says

$$\mu_{ij} = k_{ij} \exp(\mu + \alpha_i + \beta_j)$$

If waiting times between jams are independent exponential random variables, the mean number of jams in a period should be a multiple of the length of the period, hence the multiplication here by k_{ij} is completely sensible. Notice that equation (1) is a special case ($\gamma = 1$) of the relationship

$$\log \mu_{ij} = \mu + \alpha_i + \beta_j + \gamma \log(k_{ij})$$

which is a Poisson regression with link function $g(\mu) = \log(\mu)$. As it turns out, glm will fit a relationship like (1) for a Poisson mean that includes an "offset" term($\log k_{ij}$ here). Enter the data for this problem and set things up by typing

```
> A<-c(1,2,3,1,2,3)
> B<-c(1,1,1,2,2,2)
> y<-c(27,21,33,15,6,11)
> k<-c(295,416,308,474,540,498)
> AA<-as.factor(A)
> BB<-as.factor(B)
> options(contrasts=c("contr.sum","contr.sum"))
## This command sets sum contrasts for both main effects
```

(a) Fit and view some summaries for the Poisson generalized linear model (with log link and offset) by typing

```
> collator.out<-glm(y~AA+BB,family=poisson,offset=log(k))
> summary(collator.out)
```

The log link is the default for Poisson observations, so one doesn't have to specify it in the function call. Does it appear that there are statistically detectable Air Pressure and Bar Tightness effects in these data? Explain. If one wants small numbers of jams, which levels of Air Pressure and Bar Tightness does one want?

(b) Notice that estimated "per second jam rates" are given by

$$\exp(\mu + \alpha_i + \beta_j)$$

Give estimates of all 6 of these rates based on the fitted model.

(c) One can get R to find estimated means corresponding to the 6 combinations of Air Pressure and Bar Tightness for the corresponding values of k. This can either be done on the scale of the observations or on the log scale. To see these first of these, type

```
> collator.fits<-predict.glm(collator.out,type="response",
se.fit=TRUE)
> collator.fits$fit
> collator.fits$se
```

How are the "fitted values" related to your values from (b)? To see estimated/fitted log means and standard errors for those, type

```
> lcollator.fits<-predict.glm(collator.out,se.fit=TRUE)
> lcollator.fits$fit
> lcollator.fits$se
```