Statistical Methods for Discrete Response, Time Series, and Panel Data (W271): Lab 1

Professor Jeffrey Yau

Instructions:

- **Due Date: Monday Week 4**
- This lab can be done as an individual project or a group project in a group of up to 3 students from your session
- **Page limit of the pdf report: 20, which does not include the table of content page, if you hav**
- Use the margin, linespace, and font size specified below:
  - fontsize=11pt
  - margin=1in
  - line_spacing=single
- Submission:
  - Each group only needs to make one submission to our course's GitHub repo; please have one of your team members made the submission
  - Submit 2 files:
    1. A pdf file including the summary, the details of your analysis, and all the R codes used to produce the analysis. Please do not suppress the codes in your pdf file.
    2. R markdown file used to produce the pdf file
  - Use the following file-naming convensation; fail to do so will receive 10% reduction in the grade:
    - FirstNameLastName1_FirstNameLastName2_FirstNameLastName3_LabNumber.fileExtension
    - For example, if you have three students in the group for Lab 1, and their names are Gerard Kelley, Steve Yang, and Jeffrey Yau, then you should name your file the following
      - GerardKelley_SteveYang_JeffreyYau_Lab1.Rmd
      - GerardKelley_SteveYang_JeffreyYau_Lab1.pdf
  - Although it sounds obvious, please write the name of each members of your group on page 1 of your pdf and Rmd files.
  - This lab can be completed in a group of up to 3 students in your session. Students are encouraged to work in a group for the lab.
- Other general guidelines:
  - For statistical methods that we cover in this course, use only the R libraries and functions that are covered in this course. If you use libraries and functions for statistical modeling that we have not covered, you have to provide (1) explanation of why such libraries and functions are used instead and (2) reference to the library documentation. Lacking the explanation and reference to the documentation will result in a score of zero for the corresponding question.

  - Your report needs to include
    * A thorough analysis of the given dataset, which includ examiniation of anomalies, missing values, potential of top and/or bottom code, and other potential anomalies, in each of the variables.
    * An introduction section that summarize the key question being asked, the methodology employed (including the final model specification), and a highlight of the main result.
    * A comprehensive Exploratory Data Analysis (EDA) analysis, which includes both graphical and tabular analysis, as taught in this course. Output-dump (that is, graphs and tables that don't come with explanations) will result in a very low, if not zero, score. Since the report has a page-limit, you will have to be selective when choosing visuals to illustarte your key points, assocaited with a concise explanation of the visuals. Please do not ramble. Please remember that your report will have to "walk me through" your analysis.
  - A modeling section that include a detailed narrative. Make sure that your audience (in this case, the professors and your classmates) can easily follow the logic of your analysis that leads to your final model.
    * The rationale of decisions made in your modeling, supported by sufficient empirical evidence. Use the insights generated from your EDA step to guide your modeling step, as we discussed in live sessions.
    * All the steps used to arrive at your final model; these steps must be clearly shown and explained.
  - A conclusion that summarize the final result with respect to the question(s) being asked and key takeaways from the analysis.
- Other requirements:
- Students are expected to act with regards to UC Berkeley Academic Integrity.

Investigation of the 1989 Space Shuttel Challenger Accident

1. Carefullly read the Dala et al (1989) paper (Skip Section 5).
2. Answer question 4 and 5 on Chapter 2 (page 129 and 130) of Bilder and Loughin's *"Analysis of Categorical Data with R"*
3. In addition to the questions in Question 4 and 5, answer the following questions:
    a. Interpret the main result of your final model in terms of both odds and probability of failure
    b. With the same set of explanatory variables in your final model, estimate a linear regression model. Explain the model results; conduct model diagnostic; and assess the validity of the model assumptions. Would you use the linear regression model or binary logistic regression in this case. Please explain.

Dataset Analysis

```
library(car)
```

```
## Loading required package: carData
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:car':
##
##     recode
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(Hmisc)
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:dplyr':
##
```

```
##     src, summarize
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
# Set working directory
setwd("/Users/gurditchahal/f18-GurditC/labs/lab1/")
wd <- getwd()
wd
```

```
## [1] "/Users/gurditchahal/f18-GurditC/labs/lab1"
```

```
df <- read.csv(file="challenger.csv", header=TRUE, sep=",")
```

Introduction to problem and model

EDA

Model Selection and Rationale

Conclusion

4.The failure of an O-ring on the space shuttle Challenger's booster rockets led to its destruction in 1986. Using data on previous space shuttle launches, Dalal et al. (1989) examine the probability of an O-ring failure as a function of temperature at launch and combustion pressure. Data from their paper is included in the challenger.csv file.

Below are the variables: • Flight:Flightnumber • Temp:Temperature(F) at launch • Pressure: Combustion pressure (psi) • O.ring:Number of primary field O-ringfailures • Number:TotalnumberofprimaryfieldO-rings(six total,three eachfor the two booster rockets) The response variable is O.ring,and the explanatory variables are Temp and Pressure. Complete the following:

(a)The authors use logistic regression to estimate the probability an O-ring will fail.In order to use this model, the authors needed to assume that each O-ring is independent for each launch. Discuss why this assumption is necessary and the potential problems with it. Note that a subsequent analysis helped to alleviate the authors' concerns about independence.

This independence assumption is necessary for deriving the likelihood-based solution as we can take products of the probabilities. Potential issues is that the quality/durability of the O-ring might be dependent on the factory or even batch that it came from (clustering) and could interfere with producing a more accurate estimate when left unaccounted for.

(b)Estimate the logistic regression model using the explanatory variables in a linear form.

```
df$O.ring[df$O.ring==2]<-1 # need to reformulate to at least one failure
mod.fit1<-glm(formula=O.ring~Pressure+Temp,data=df,family=binomial(link = logit))
summary(mod.fit1)
```

```
##
## Call:
```

```
## glm(formula = O.ring ~ Pressure + Temp, family = binomial(link = logit),
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1993  -0.5778  -0.4247   0.3523   2.1449
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 13.292360   7.663968   1.734   0.0828 .
## Pressure     0.010400   0.008979   1.158   0.2468
## Temp        -0.228671   0.109988  -2.079   0.0376 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28.267  on 22  degrees of freedom
## Residual deviance: 18.782  on 20  degrees of freedom
## AIC: 24.782
##
## Number of Fisher Scoring iterations: 5
```

(c)Perform LRTs to judge the importance of the explanatory variables in the model.

```
Anova(mod.fit1,Test='LRT')
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: O.ring
##          LR Chisq Df Pr(>Chisq)
## Pressure   1.5331  1   0.215648
## Temp       7.7542  1   0.005359 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(d)The authors chose to remove Pressure from the model based on the LRTs. Based on your results, discuss why you think this was done. Are there any potential problems with removing this variable?

In terms of statistical significance, pressure wasn't found to be statistically significant. Potential problems could be losing precision on temp as well as in probability, especially for edge cases.

5.Continuing Exercise 4, consider the simplified model

$$logit(\pi) = Beta_0 + Beta_1 * Temp$$

, where

$$\pi$$

is the probability of an O-ring failure. Complete the following:

(a)Estimate the model.

```
mod.fit2<-glm(formula=O.ring~Temp,data=df,family=binomial(link = logit))
summary(mod.fit2)
```

```
##
## Call:
## glm(formula = O.ring ~ Temp, family = binomial(link = logit),
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0611  -0.7613  -0.3783   0.4524   2.2175
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  15.0429     7.3786   2.039   0.0415 *
## Temp         -0.2322     0.1082  -2.145   0.0320 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28.267  on 22  degrees of freedom
## Residual deviance: 20.315  on 21  degrees of freedom
## AIC: 24.315
##
## Number of Fisher Scoring iterations: 5
```

(b)Construct two plots: (1)

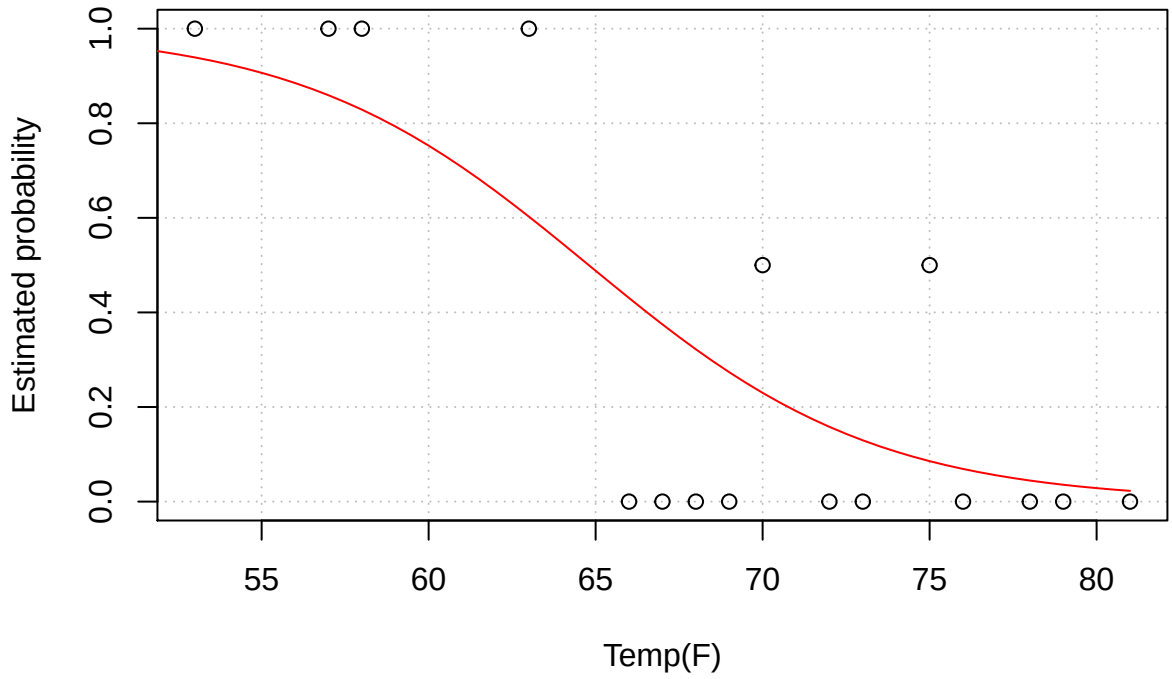
$$\pi$$

vs.Temp and (2) Expected number of failures vs.Temp. Use a temperature range of 31 to 81 on the x-axis even though the minimumtemperature in the data set was 53.

```
w<-aggregate(formula=O.ring~Temp,data=df,FUN=sum)
n<-aggregate(formula=O.ring~Temp,data=df,FUN=length)
w.n<-data.frame(Temperature=w$Temp,Failure=w$O.ring,trials = n$O.ring, proportion = round(w$O.r
head(w.n)
```

```
##   Temperature Failure trials proportion
## 1          53       1      1          1
## 2          57       1      1          1
## 3          58       1      1          1
## 4          63       1      1          1
## 5          66       0      1          0
## 6          67       0      3          0
```

```
plot(x=w$Temp,y=w$O.ring/n$O.ring,xlab="Temp(F)", ylab = "Estimated probability", panel.first =
curve(expr=predict(object=mod.fit2,newdata=data.frame(Temp = x), type = "response"), col = "red
```




```
#Todo:Expected number of failures vs.Temp.
```

(c)Include the 95% Wald confidence interval bands for

$$\pi$$

on the plot. Why are the bands much wider for lower temperatures than for higher temperatures?

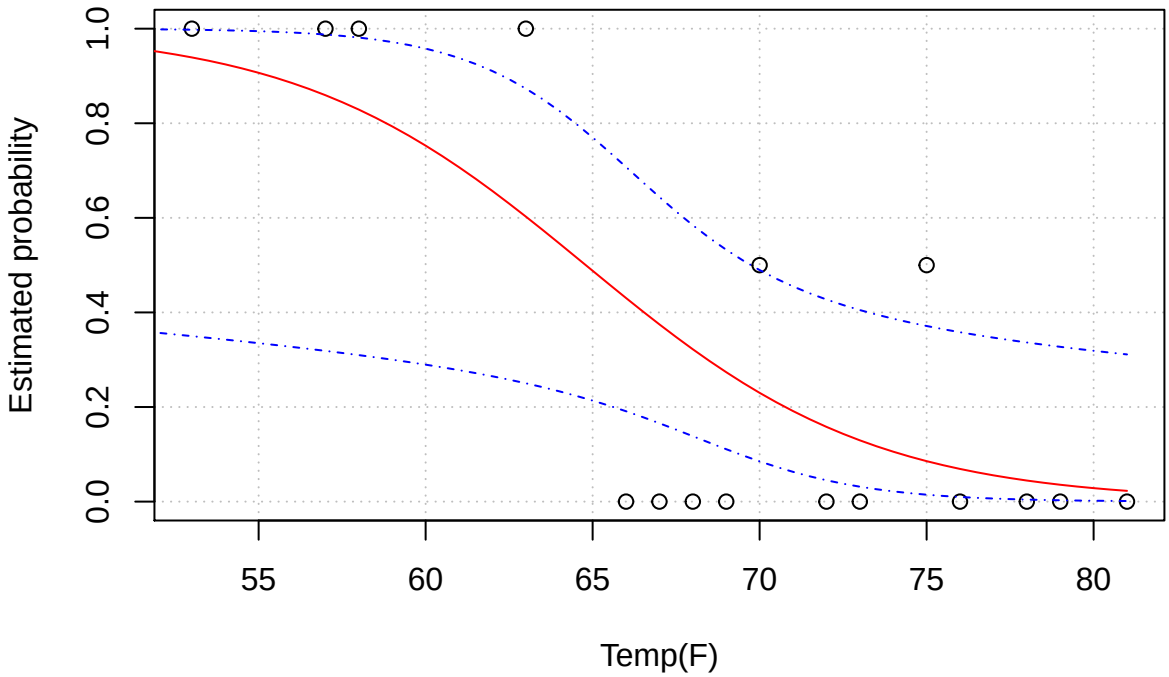
```
plot(x=w$Temp,y=w$O.ring/n$O.ring,xlab="Temp(F)", ylab = "Estimated probability", panel.first =
curve(expr=predict(object=mod.fit2,newdata=data.frame(Temp = x), type = "response"), col = "red

ci.pi<-function(newdata,mod.fit.obj,alpha){linear.pred <- predict(object = mod.fit.obj, newdata
CI.lin.pred.lower <- linear.pred$fit - qnorm(p =1-alpha/2)*linear.pred$se
CI.lin.pred.upper <- linear.pred$fit + qnorm(p =1-alpha/2)*linear.pred$se

CI.pi.lower <- exp(CI.lin.pred.lower) / (1 +exp(CI.lin.pred.lower))
CI.pi.upper <- exp(CI.lin.pred.upper) / (1 +exp(CI.lin.pred.upper))
list(lower = CI.pi.lower, upper = CI.pi.upper)}

curve(expr=ci.pi(newdata=data.frame(Temp=x),mod.fit.obj = mod.fit2, alpha = 0.05)$lower, col =

curve(expr=ci.pi(newdata=data.frame(Temp=x),mod.fit.obj = mod.fit2, alpha = 0.05)$upper, col =
```



Bands are wider due to change in probability across temperature gradient. Much steeper drop in temperature below and above 65 (similar to complete separation problem). Less of a drastic change in higher temperatures due to two "middle" values between 70 and 75.

(d)The temperature was 31 at launch for the Challenger in 1986. Estimate theprobability of an O-ring failure using this temperature, and compute a corresponding confidence interval. Discuss what assumptions need to be made in order to apply the inference procedures.

```
alpha=0.05
predict.data <- data.frame(Temp=31) #data to predict on

linear.pred=predict(object = mod.fit2, newdata = predict.data, #linear part of model
                    type = "link", se = TRUE)

pi.hat = exp(linear.pred$fit)/(1+exp(linear.pred$fit)) #estimated probability
pi.hat
```

```
##         1
## 0.9996088
```

```
CI.lin.pred = linear.pred$fit + qnorm(p = c(alpha/2, 1-alpha/2))*linear.pred$se #confidence in



CI.pi = exp(CI.lin.pred)/(1+exp(CI.lin.pred)) #actual interval
CI.pi
```

```
## [1] 0.4816106 0.9999999
```

(e)Rather than using Wald or profile LR intervals for the probability of failure, Dalalet al. (1989) use a parametric bootstrap to compute intervals. Their process was to (1) simulate a large number

of data sets (n= 23for each) from the estimated model of

$$logit(\pi) = Beta_0 + Beta_1 * Temp$$

;(2)estimate new models for each dataset,say

$$logit(\pi) = Beta_0 + Beta_1 * Temp$$

;and (3)compute at a specific temperature of interest. The authors used the 0.05 and 0.95 observed quantiles from the simulated distribution as their 90% confidence interval limits. Using the parametric bootstrap, compute 90% confidence intervals separately at temperatures of 31 and 72.27

```
#Todo: Bootstrapping CI
```

(f)Determine if a quadratic term is needed in the model for the temperature.

```
mod.fit.Ha<-glm(formula=O.ring~Temp+I(Temp^2),data=df,family=binomial(link = logit))
anova(mod.fit2,mod.fit.Ha,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: O.ring ~ Temp
## Model 2: O.ring ~ Temp + I(Temp^2)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        21     20.315
## 2        20     19.389  1  0.92649   0.3358
```

Quadratic term fails to produce significant effect in change in residual deviance and so we fail to reject that the coefficient is actually 0 for the quadratic term.