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

K.C. Tobin, Weixing Sun, Winston Lin

June 18th, 2017

Investigation of the 1989 Space Shuttle Challenger Accident

1. Read the Dala et al (1989) paper (attached in this zip file).

- Central question: What is the probability of catastrophic field-joint failure if we launch tomorrow morning at 31°F?
- Analysis of flights before Challenger launch showed that field joint O-ring reliability was in doubt due to joint rotation, ring compression, and ring erosion during ignition. Effects were more severe at low temperatures.
- Binomial model (n = 6) for probability of thermal distress per field joint included only temperature. Pressure was dropped from model after testing hypothesis of no pressure effect, which found 90% CIs overlapped significantly (Model 3.2). Model 3.2 assumes independence of thermal distress across 6 joints. (p.5)
- Binary model, probability of thermal distress on at least one joint, was shown to follow binomial model closely. Independence assumption not required.
- Models estimated that 4-5 O-rings would be damaged at 31°F.
- Bootstrap CIs for betas and probabilities were estimated using binomial model instead of actual data. CIs for $t < 65$ were wide due to lack of data in that range. 90% CI for $t = 30$ was (1, 6).
- Sensitivity analysis conducted by fitting models for all 23 points with each point held out and comparing standardized residuals of between coefficients of hold out and full models. Holding out Point 21 resulted in large residual, but did not invalidate model.
- Linear relation between logit model and temperature was confirmed by adding quadratic term and performing LRT. Linearity also confirmed by estimating nonparametric relationship based on smoothing of data. Model was plotted against smoothed values and fit well. Standardized residuals between smoothed and fitted values were mostly random versus temperature except for 2 outliers (points 9 and 21).
- Binary model (n = 2) for probability of thermal distress per nozzle joint included both temperature and pressure. Pressure was found to have a greater effect than temperature. (pp.9-10)

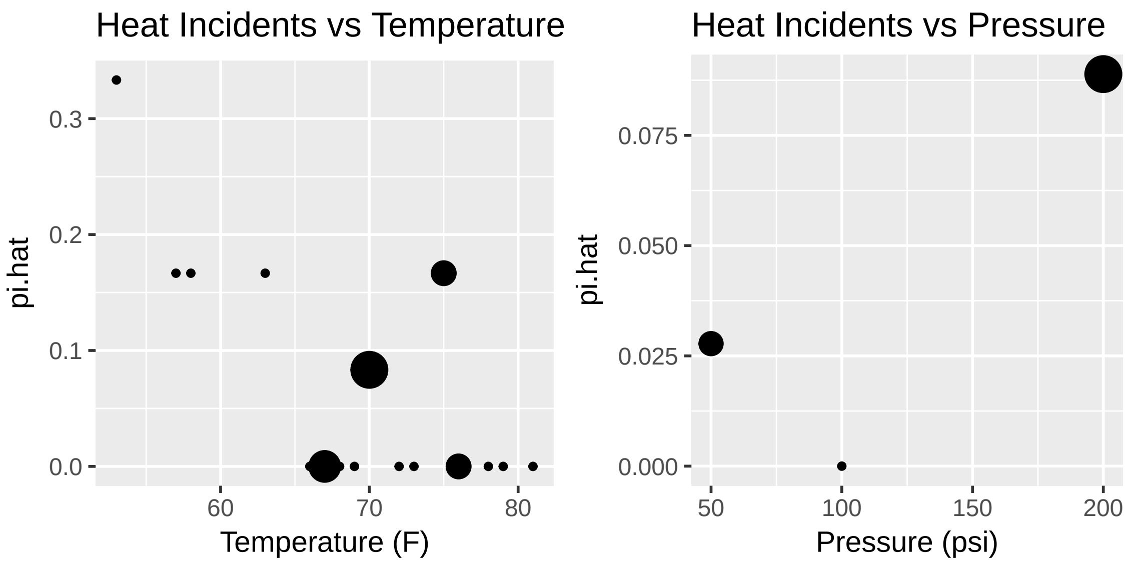
2. Conduct a thorough analysis and EDA of the given dataset "challenger.csv", as we did in live session 2 and 3. Pay attention to the instructions given above.

The challenger data includes five variables of 23 observations of previous shuttle launches prior to the incident: - Flight: a unique identifier for the shuttle launch - Temp: a continuous variable representing joint temperature at launch in Farenheit - Pressure: a discrete variable taking on values of 50, 100, and 200 representing the test pressure for for the leak test port - O.ring: a discrete value representing the number of observed o-rings thats had some level of failure upon examination post launch - Number: the number of total o-rings (this is always 6)

In reviewing the data we can see that the decision to increase the test pressure sensor to 200 psi happened quite early so most observations are at this 200 level. We can also see from the temperature data that most observations for temperature occur between 65 and 75 degrees which given Florida climate knowledge makes sense. There are only 3 observations below 60 degrees with the lowest being 53 degrees.

In analyzing the temperature vs o-ring failure bivariate relationship there does appear to be some correlation between lower temperature and a higher number of incidents with an interesting data point at 75 degrees which bucks the trend. This observation was noted in the paper and even removed from some models.

The relationship between o-ring failure and pressure is much less defined. We see a low incidence rate at 50 psi and no failures at 100 psi (though there are only 2 recorded flights). The majority of the data at 200 psi does not appear to indicate a high incident rate (under 10%) and is clearly the majority of the data.



3. Answer question 4 and 5 on Chapter 2 (page 129 and 130) of Bilder and Loughin's *"Analysis of Categorical Data with R"*

2.4. 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: Flight number
- Temp: Temperature (F) at launch
- Pressure: Combustion pressure (psi)
- O.ring: Number of primary field O-ring failures
- Number: Total number of primary field O-rings (six total, three each for 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.

- The assumption of independence between O-ring failures ensured that probability of failure could be modeled using a binomial distribution, which assumes independence between trials. The potential problem with this is that O-ring failures might not be completely independent, and thus, the probability model would have limited explanatory or predictive power in physical reality.

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

```r
# fit binomial response model
mod.fit.lr1 = glm(formula = O.ring/Number ~ Temp + Pressure, weights=Number,
                  family = binomial(link = logit), data = dt)
summary(mod.fit.lr1)
```

```
##
## Call:
## glm(formula = O.ring/Number ~ Temp + Pressure, family = binomial(link = logit),
##     data = dt, weights = Number)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0361  -0.6434  -0.5308  -0.1625   2.3418
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.520195   3.486784   0.723   0.4698
## Temp        -0.098297   0.044890  -2.190   0.0285 *
## Pressure     0.008484   0.007677   1.105   0.2691
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 24.230  on 22  degrees of freedom
## Residual deviance: 16.546  on 20  degrees of freedom
## AIC: 36.106
##
## Number of Fisher Scoring iterations: 5
```

```r
# binarize outcome: fail = 1 if O.ring > 0, else 0
dt[, fail := as.numeric(O.ring > 0)]

# fit binary response model
mod.fit.lr1b = glm(formula = fail ~ Temp + Pressure, family = binomial(link = logit), data = dt)
summary(mod.fit.lr1b)
```

```
##
## Call:
## glm(formula = fail ~ Temp + Pressure, family = binomial(link = logit),
##     data = dt)
##
## 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 .
## Temp        -0.228671   0.109988  -2.079   0.0376 *
## Pressure     0.010400   0.008979   1.158   0.2468
## ---
## 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.

```r
library(car)

# LRTs for binary model
Anova(mod.fit.lr1, test = 'LR')
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: O.ring/Number
##          LR Chisq Df Pr(>Chisq)
## Temp       5.1838  1     0.0228 *
## Pressure   1.5407  1     0.2145
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#anova(mod.fit.lr1, test = 'Chisq')
```

- LRT shows strong evidence that the effect of temperature is statistically significant (p-value < 0.05). There is no strong evidence that the effect of pressure is statistically significant (p-value $>> 0.05$).

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?

- Pressure was removed from the model because it was shown to have little explanatory power in the LRTs (p-value > 0.05). The problem with the pressure variable was that it was measured during tests of O-rings before launch, not the actual pressure on the joint during the launch. Thus, removing the pressure variable given is reasonable. However, actual pressure during the launch would probably affect the probability of O-ring failure given the physical relationship between temperature and pressure in gasses.

2.5. Continuing Exercise 4, consider the simplified model $\text{logit}(\pi) = \beta_0 + \beta_1 \text{Temp}$, where π is the probability of an O-ring failure. Complete the following:

a. Estimate the model.

```r
# fit binomial response model
mod.fit.lr2 = glm(formula = O.ring/Number ~ Temp, weights=Number,
                  family = binomial(link = logit), data = dt)
summary(mod.fit.lr2)
```

```
##
## Call:
## glm(formula = O.ring/Number ~ Temp, family = binomial(link = logit),
##     data = dt, weights = Number)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.95227  -0.78299  -0.54117  -0.04379   2.65152
##
## Coefficients:
```

```
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.08498    3.05247   1.666   0.0957 .
## Temp        -0.11560    0.04702  -2.458   0.0140 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 24.230  on 22  degrees of freedom
## Residual deviance: 18.086  on 21  degrees of freedom
## AIC: 35.647
##
## Number of Fisher Scoring iterations: 5
```

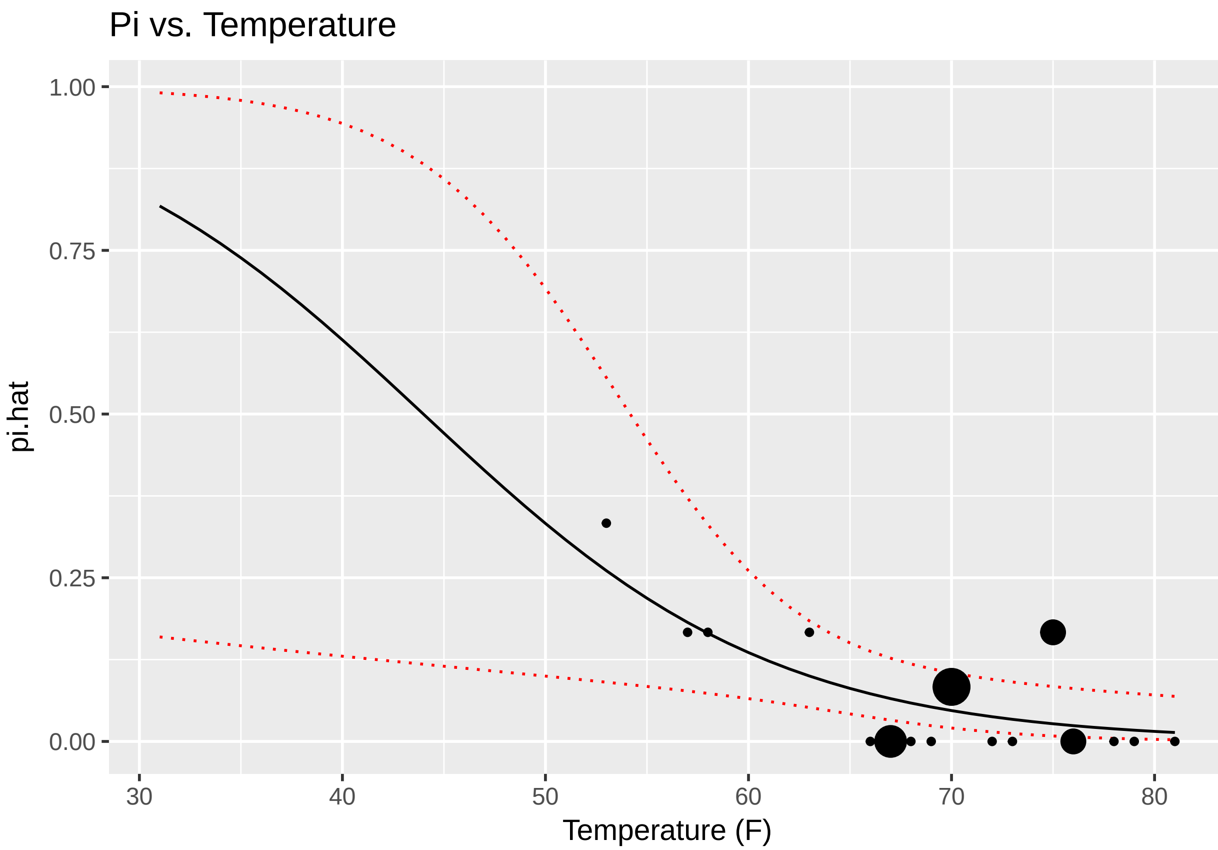
```r
# fit binary response model
mod.fit.lr2b = glm(formula = fail ~ Temp, family = binomial(link = logit), data = dt)
summary(mod.fit.lr2b)
```

```
##
## Call:
## glm(formula = fail ~ Temp, family = binomial(link = logit), data = dt)
##
## 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
```

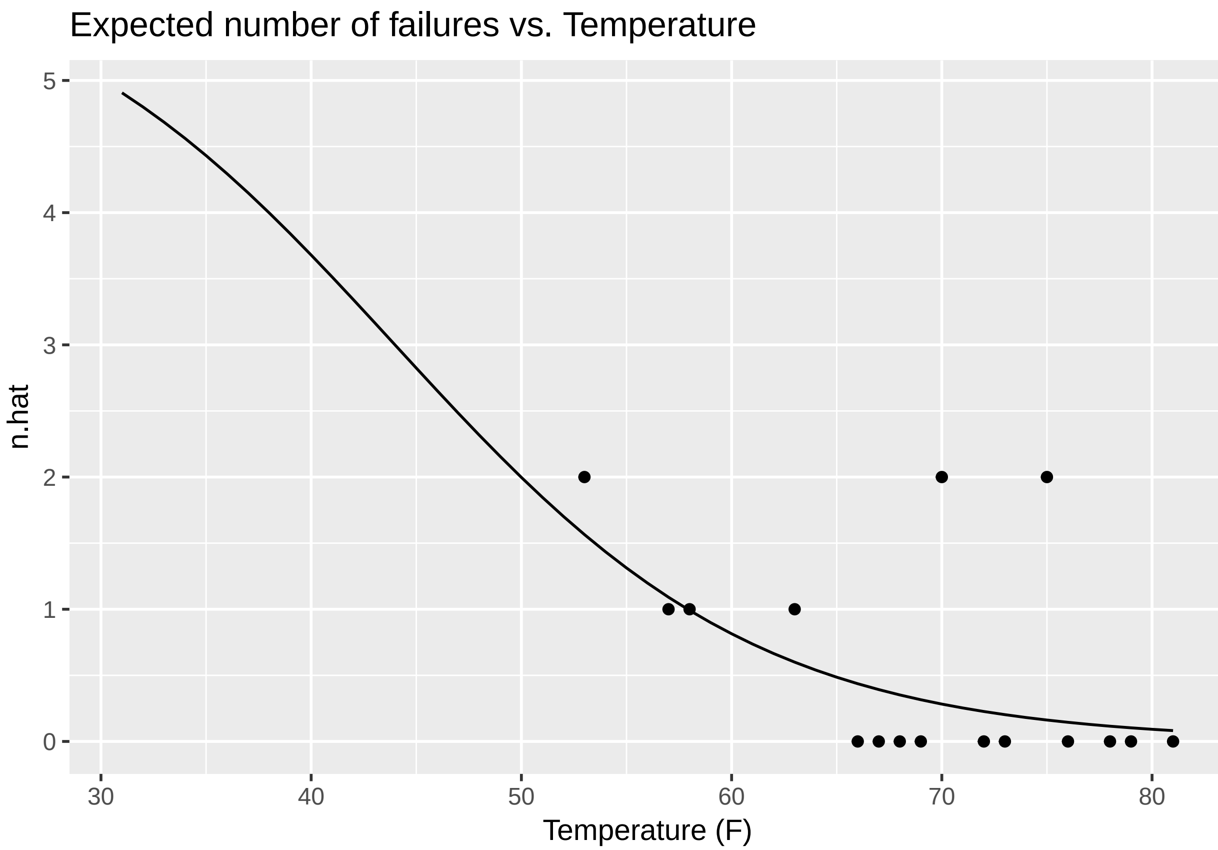
b. Construct two plots: (1) π vs. Temp and (2) Expected number of failures vs. Temp. Use a temperature range of 31° to 81° on the x-axis even though the minimum temperature in the data set was 53°.

```r
# calcs pi from logit value
pi.logit = function(log.odds) {
        exp(log.odds) / (1 + exp(log.odds))
}

# predicted values with temperature only model
pred.x = data.frame(Temp = 31:81)
logit = predict(object = mod.fit.lr2, newdata = pred.x, type = 'link', se = TRUE)
pi.hat = pi.logit(logit$fit)

# Wald confidence interval for pi
alpha = 0.05
```

```r
# *** multiplying vector of scores only works for single prediction, not vector...
# ci.logit = logit$fit + qnorm(p = c(alpha/2, 1 - alpha/2))*logit$se.fit
# ci.pi = exp(ci.logit) / (1 + exp(ci.logit))

ci.logit.lower = logit$fit - qnorm(p = 1 - alpha/2)*logit$se.fit
ci.pi.lower = pi.logit(ci.logit.lower)
ci.logit.upper = logit$fit + qnorm(p = 1 - alpha/2)*logit$se.fit
ci.pi.upper = pi.logit(ci.logit.upper)

# data table with predicted values
pred.results = data.table(pred.x,
                  pi.hat = pi.hat,
                  n.hat = pi.hat * 6,
                  ci.pi.lower = ci.pi.lower,
                  ci.pi.upper = ci.pi.upper)

# plots of pi.hat and expected incidents vs. temperature
par(mfrow = c(2,2))

# plot of pi.hat vs. temperature
plt1 = ggplot()
plt1 + geom_line(data = pred.results, aes(x = Temp, y = pi.hat)) +
        geom_line(data = pred.results, aes(x = Temp, y = ci.pi.upper), color = 'red',
                  linetype = 'dotted') +
        geom_line(data = pred.results, aes(x = Temp, y = ci.pi.lower), color = 'red',
                  linetype = 'dotted') +
        geom_point(data = dtt, aes(x = Temp, y = pi.hat, size = n)) +
        labs(title = 'Pi vs. Temperature', x = 'Temperature (F)') +
        theme(legend.position = 'none')
```




```r
# plot of expected incidents vs. temperature
plt2 = ggplot()
plt2 + geom_line(data = pred.results, aes(x = Temp, y = n.hat)) +
        geom_point(data = dtt, aes(x = Temp, y = w)) +
        labs(title = 'Expected number of failures vs. Temperature', x = 'Temperature (F)') +
        theme(legend.position = 'none')
```



c. Include the 95% Wald confidence interval bands for π on the plot. Why are the bands much wider for lower temperatures than for higher temperatures?

- The confidence interval is wider for lower temperatures because there is no data on incidents below 53°, and there are higher variance observations at lower temperatures. The confidence intervals narrows at temperatures of 70° and above, where the bulk of our data are. Intuitively, we are less confident predictions made in ranges where there is little or no data to inform them.

d. The temperature was 31° at launch for the Challenger in 1986. Estimate the probability 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.

```r
# prediction at 31 degrees
pred31.logit = predict(object = mod.fit.lr2, newdata = data.frame(Temp = 31), type = "link",
                  se.fit = TRUE)
pred31 = pi.logit(pred31.logit$fit)


# 95% Wald CI
ci.logit31 = pred31.logit$fit + qnorm(p = c(alpha/2, 1-alpha/2))*pred31.logit$se.fit
ci.pi31 = pi.logit(ci.logit31)
# round(c(lower = ci.pi.lower[1], upper = ci.pi.upper[1]), 4)
ci.pi31
```

```
## [1] 0.1596025 0.9906582
```

- $\hat{\pi} = 0.8178$, 95% Wald CI = 0.1596, 0.9907. In order to perform inference with our model, we assume trials are independent to apply a binomial distribution. We also assume there is no complete separation of data, which would cause extreme values of coefficients that would not converge easily.

e. Rather than using Wald or profile LR intervals for the probability of failure, Dalal et al. (1989) use a parametric bootstrap to compute intervals. Their process was to (1) simulate a large number of data sets (n = 23 for each) from the estimated model of $\text{logit}(\hat{\pi}) = \hat{\beta}_0 + \hat{\beta}_1\text{Temp}$; (2) estimate new models for each data set, say $\text{logit}(\hat{\pi}^*) = \hat{\beta}_0^* + \hat{\beta}_1^*\text{Temp}$; and (3) compute $\hat{\pi}^*$ at a specific temperature of interest. The authors used the 0.05 and 0.95 observed quantiles from the $\hat{\pi}^*$ simulated distribution as their 90% confidence interval limits. Using the parametric bootstrap, compute 90% confidence intervals separately at temperatures of 31° and 72°.

```r
### simulates failures based on estimated prob distribution of failures
bootstrap <- function(pi.hat, data){
        # simulate failures
        failures <- rbinom(n = 23, size = 6, pi.hat)
        data$O.ring <- failures

        # fit binary model
        mod.fit.bs = glm(formula = O.ring/Number ~ Temp, weights = Number,
                         family = binomial(link = logit), data = data)

        # pred31 <- unlist(predict(mod, data.frame(Temp = 31), type = "response"))
        # pred72 <- unlist(predict(mod, data.frame(Temp = 72), type = "response"))

        # make predictions
        pred31 <- predict(object = mod.fit.bs, newdata = data.frame(Temp = 31),
                          type = "response")
        pred72 <- predict(object = mod.fit.bs, newdata = data.frame(Temp = 72),
                          type = "response")
        return(list(pred.31 = pred31, pred.72 = pred72))
}

# estimate prob distribution for bootstrapping
pi.hat <- predict(object = mod.fit.lr2, newdata = data.frame(Temp = dt$Temp),
                  type = "response")

# bootstrap 10000 samples and return predictions as sorted list
preds <- replicate(10000, bootstrap(pi.hat, dt))
preds.31 <- sort(unlist(preds[1,]))
preds.72 <- sort(unlist(preds[2,]))

# take 0.05 and 0.95 quantiles of observed outcomes as 90% CI
ci.pi31 = quantile(preds.31, c(.05, .95))
ci.pi31
```

```
##        5%       95%
## 0.1810167 0.9882131
```

```r
ci.pi72 = quantile(preds.72, c(.05, .95))
ci.pi72
```

```
##         5%        95%
## 0.01127681 0.06930766
```

90% Wald CI:

Temp = 31: 0.181, 0.988

Temp = 72: 0.011, 0.069

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

- ANOVA tests show that the quadratic term for temperature does not add explanatory power to the original model and can be removed. In fact our LRT and Wald tests show a reduction in the explanatory power of temperature as a result of adding the quadratic term.

```
# model with quadratic term for temperature
mod.fit.lr3 = glm(formula = O.ring/Number ~ Temp + I(Temp^2), weights = Number, family = binomial(link =
                  data = dt)
summary(mod.fit.lr3)
```

```
##
## Call:
## glm(formula = O.ring/Number ~ Temp + I(Temp^2), family = binomial(link = logit),
##     data = dt, weights = Number)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.84320  -0.72385  -0.61980  -0.01335   2.52101
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 22.126148  23.794426   0.930    0.352
## Temp        -0.650885   0.740756  -0.879    0.380
## I(Temp^2)    0.004141   0.005692   0.727    0.467
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 24.230  on 22  degrees of freedom
## Residual deviance: 17.592  on 20  degrees of freedom
## AIC: 37.152
##
## Number of Fisher Scoring iterations: 5
```

```
# LRTs
Anova(mod.fit.lr3, test = 'LR')
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: O.ring/Number
##           LR Chisq Df Pr(>Chisq)
## Temp       0.71878  1     0.3965
## I(Temp^2)  0.49470  1     0.4818
```

```
anova(mod.fit.lr2, mod.fit.lr3, test = 'Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: O.ring/Number ~ Temp
## Model 2: O.ring/Number ~ Temp + I(Temp^2)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        21     18.086
## 2        20     17.592  1   0.4947   0.4818
```

4. 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.

- The final model only contains Temp as the explanatory variable. Because p-values of pressure and temp quadratic terms are much larger than 0.05, it means that these parameters are not statistically significant in the model.
- For each degree that temperature decreases, the odds of any joint failure occurring will increase by 1.12 times under our binary model. At 31°, the predicted probability of any joint failing occurring would be 82%, the probability of at least one joint experiencing failure would be effectively 100% so flights at such temperatures should be avoided.

```r
library(stargazer)
```

```
##
## Please cite as:
##  Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2. http://CRAN.R-project.org/package=stargazer
```

```r
stargazer(mod.fit.lr1, mod.fit.lr2, mod.fit.lr3, type = "text")
```

```
##
## ===============================================
##                        Dependent variable:
##                   -----------------------------
##                          O.ring/Number
##                      (1)       (2)       (3)
## -----------------------------------------------
## Temp              -0.098**  -0.116**   -0.651
##                    (0.045)   (0.047)   (0.741)
##
## Pressure            0.008
##                    (0.008)
##
## I(Temp2)                                0.004
##                                        (0.006)
##
## Constant            2.520    5.085*    22.126
##                    (3.487)   (3.052)  (23.794)
##
## -----------------------------------------------
## Observations         23        23        23
## Log Likelihood     -15.053   -15.823   -15.576
## Akaike Inf. Crit.  36.106    35.647    37.152
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

```r
# odds for binomial model
summary(mod.fit.lr2)
```

```
##
## Call:
## glm(formula = O.ring/Number ~ Temp, family = binomial(link = logit),
##     data = dt, weights = Number)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.95227  -0.78299  -0.54117  -0.04379   2.65152
##
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.08498    3.05247   1.666   0.0957 .
## Temp        -0.11560    0.04702  -2.458   0.0140 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 24.230  on 22  degrees of freedom
## Residual deviance: 18.086  on 21  degrees of freedom
## AIC: 35.647
##
## Number of Fisher Scoring iterations: 5
```

```r
1 / exp(mod.fit.lr2$coefficients[2])
```

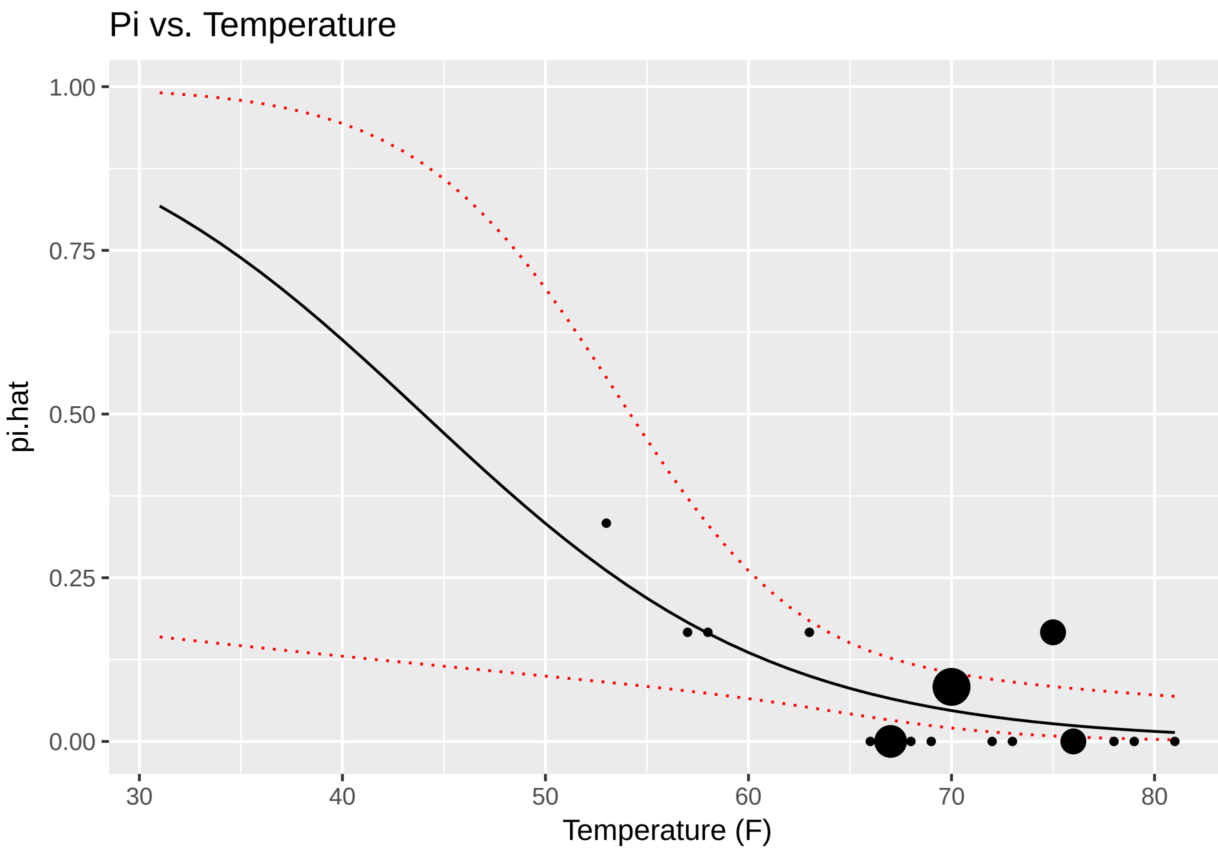
```
##     Temp
## 1.122548
```

```r
pred31.logit = predict(object = mod.fit.lr2, newdata = data.frame(Temp = 31), type = "link",
                 se.fit = TRUE)
pred31 = pi.logit(pred31.logit$fit)
# pi.hat at 31 degreeF in our final model
round(pred31,4)
```

```
##      1
## 0.8178
```

b. Plot the main effect of your final model with the y-axis being probability of failure and x-axis being *temperature*.

```r
# plot of pi.hat vs. temperature
plt1 = ggplot()
plt1 + geom_line(data = pred.results, aes(x = Temp, y = pi.hat)) +
        geom_line(data = pred.results, aes(x = Temp, y = ci.pi.upper), color = 'red',
                  linetype = 'dotted') +
        geom_line(data = pred.results, aes(x = Temp, y = ci.pi.lower), color = 'red',
                  linetype = 'dotted') +
        geom_point(data = dtt, aes(x = Temp, y = pi.hat, size = n)) +
        labs(title = 'Pi vs. Temperature', x = 'Temperature (F)') +
        theme(legend.position = 'none')
```


Pi vs. Temperature
1.00
0.75
0.50
0.25
0.00
pi.hat
30
40
50
60
70
80
Temperature (F)