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

K.C. Tobin, Weixing Sun, Winston Lin

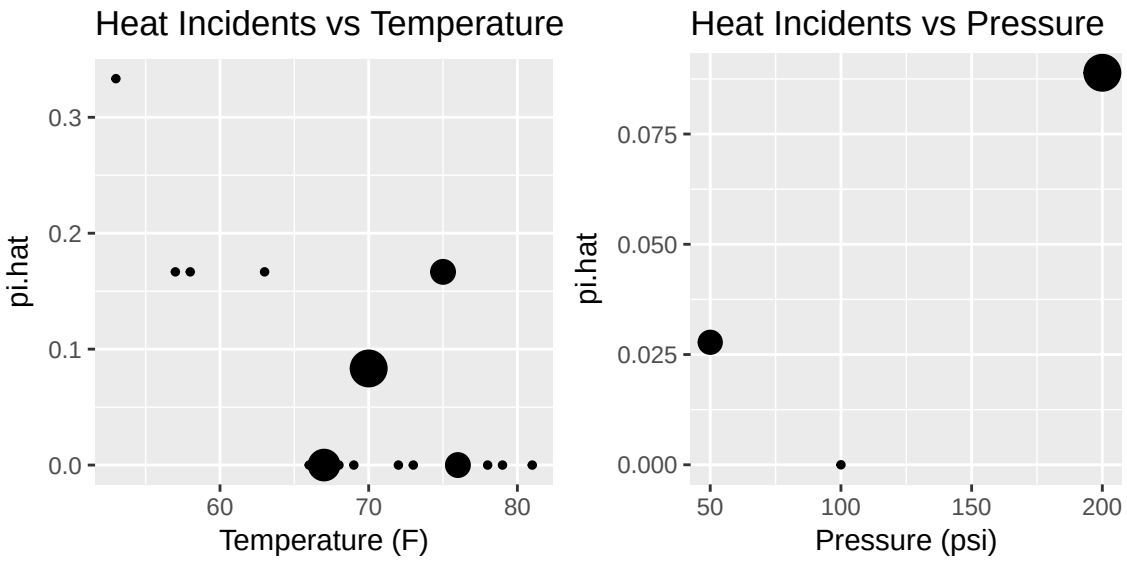
June 15, 2017

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.



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 Poisson 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.

```
# binary response model
dt[, fail := as.numeric(O.ring > 0)]
mod.fit.lr = glm(formula = fail ~ Temp + Pressure, family = binomial(link = logit), data = dt)
summary(mod.fit.lr)
```

```
##
## 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.

```
library(car)
Anova(mod.fit.lr, test = 'LR')
```

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

LRT shows that the effect of pressure is not 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. The pressure variable measured the pressure at which the O-rings were tested 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 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.

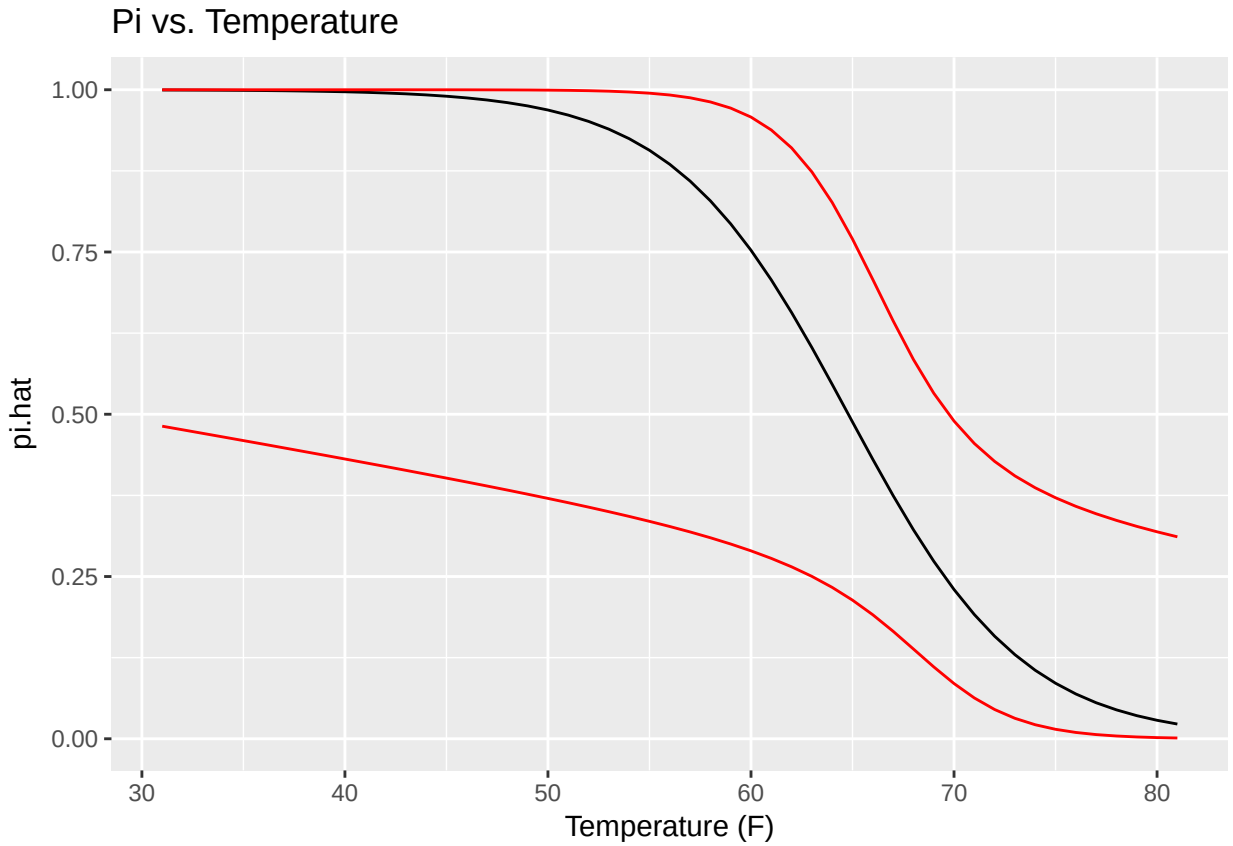
```
# binary response model
mod.fit.lr2 = glm(formula = fail ~ Temp, family = binomial(link = logit), data = dt)
summary(mod.fit.lr2)
```

```
##
## 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°.

```
# 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 = exp(logit$fit) / (1 + exp(logit$fit))

# Wald confidence interval
alpha = 0.05
# 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.upper = logit$fit + qnorm(p = 1 - alpha/2)*logit$se.fit
ci.logit.lower = logit$fit + qnorm(p = alpha/2)*logit$se.fit
ci.pi.upper = exp(ci.logit.upper) / (1 + exp(ci.logit.upper))
ci.pi.lower = exp(ci.logit.lower) / (1 + exp(ci.logit.lower))

# data table with predicted values
pred = 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, aes(x = Temp, y = pi.hat)) +
        geom_line(data = pred, aes(x = Temp, y = ci.pi.upper), color = 'red') +
        geom_line(data = pred, aes(x = Temp, y = ci.pi.lower), color = 'red') +
        labs(title = 'Pi vs. Temperature', x = 'Temperature (F)') +
        theme(legend.position = 'none')
```




```
# plot of expected incidents vs. temperature
plt2 = ggplot() +
        geom_line(data = pred, aes(x = temp, y = n.hat)) +
        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 width is wider for lower temperatures

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.

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°.

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

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.

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