w271_ChristianMillsop_TennisonYu_Lab1

Christian Millsop, Tennison Yu

January 22, 2019

Question 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. + Rocket motors are potentially re-used. Are O-rings replaced each time? + There are 3 primary O-rings per rocket. Operational anomalies that occur in a rocket likely affect all o-rings on the rocket. + https://stats.stackexchange.com/questions/259704/is-there-i-i-d-assumption-on-logistic-regression

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

```
data = read.csv(file="./challenger.csv", header=TRUE)
data$Prob = data$O.ring/data$Number
```

```
fit.a = glm(formula = Prob ~ Temp + Pressure, family=binomial(link=logit),data = data)
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial
## glm!
```

```
summary(fit.a)
```

```
##
## Call:
## glm(formula = Prob ~ Temp + Pressure, family = binomial(link = logit),
##     data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.42299  -0.26267  -0.21671  -0.06634   0.95601
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.520195   8.540918   0.295    0.768
## Temp        -0.098297   0.109959  -0.894    0.371
## Pressure     0.008484   0.018806   0.451    0.652
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4.0384  on 22  degrees of freedom
## Residual deviance: 2.7576  on 20  degrees of freedom
## AIC: 9.2286
```

```
##
## Number of Fisher Scoring iterations: 6
```

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

```
fit.a.null = glm(formula = Prob ~ 1, family=binomial(link=logit),data = data)
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial
## glm!
```

```
fit.a.reduced = glm(formula = Prob ~ Temp, family=binomial(link=logit),data = data)
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial
## glm!
```

```
anova(fit.a.null, fit.a.reduced, fit.a, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Prob ~ 1
## Model 2: Prob ~ Temp
## Model 3: Prob ~ Temp + Pressure
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        22     4.0384
## 2        21     3.0144  1  1.02401   0.3116
## 3        20     2.7576  1  0.25678   0.6123
```

- (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 is less important than Temperature. Neither explanatory variable is statistically significant by LRT though.
- Parsimony is desirable.
- PV = nRT, we expect some relationship or interaction between pressure and temperature.
- High pressure can cause blow holes, which expose the O-ring to high temperature

Question 5

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

```
fit.b = glm(formula = Prob ~ Temp, family=binomial(link=logit),data = data)
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial
## glm!
```

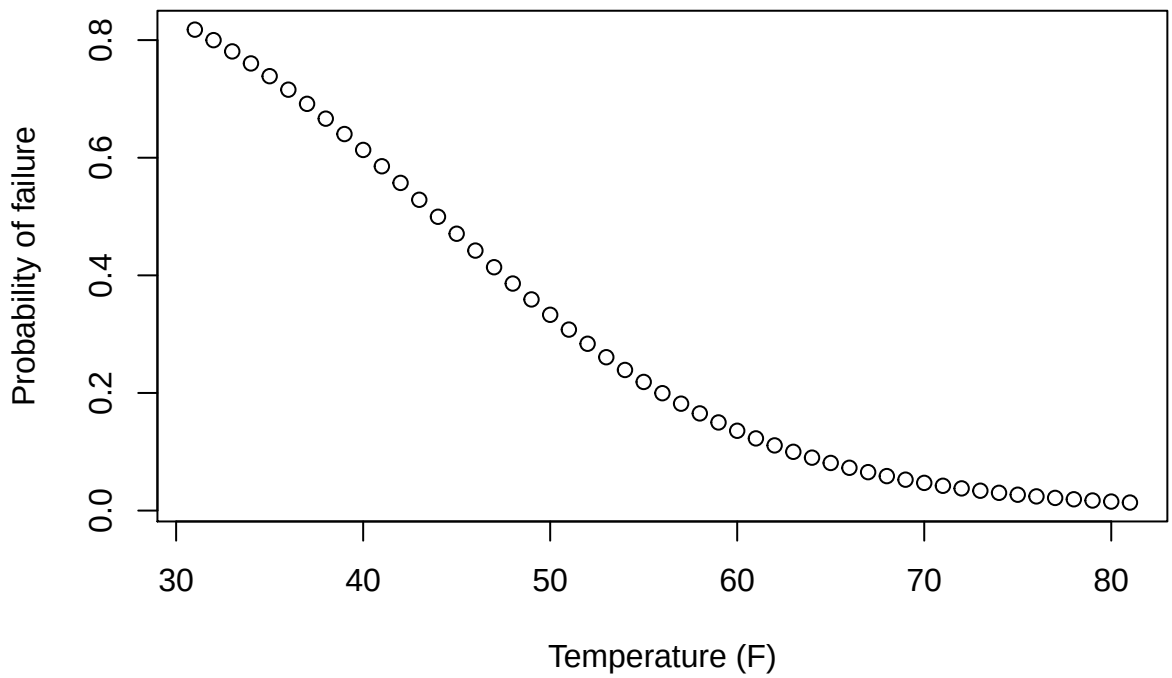
```
summary(fit.b)
```

```
##
## Call:
## glm(formula = Prob ~ Temp, family = binomial(link = logit), data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.38876  -0.31965  -0.22093  -0.01788   1.08248
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
```

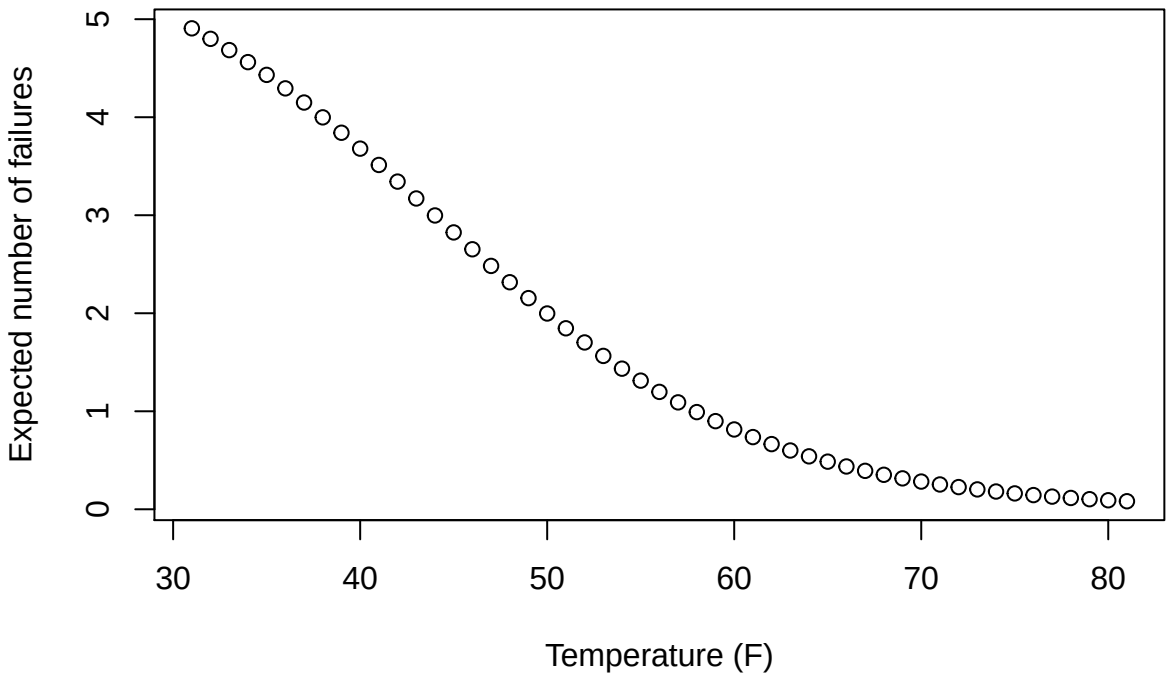
```
## (Intercept)   5.0850     7.4770   0.680    0.496
## Temp         -0.1156     0.1152  -1.004    0.316
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4.0384  on 22  degrees of freedom
## Residual deviance: 3.0144  on 21  degrees of freedom
## AIC: 7.2026
##
## Number of Fisher Scoring iterations: 6
```

- (b) Construct two plots: (1) π vs. $Temp$ and (2) Expected number of failures vs. $Temp$. Use a temperature range of 31ř to 81ř on the x-axis even though the minimum temperature in the data set was 53ř.
- The first plot, π vs. $Temp$, is the probability that any given o-ring will fail at a certain temperature.
- The second plot is the sum of the probabilities for all 6 of the o-rings to fail at a certain temperature.

```
plotrange = seq(31, 81, 1)
pred_odds = predict(fit.b, newdata=data.frame(Temp=plotrange), type="link")
plot(plotrange, exp(pred_odds)/(1+exp(pred_odds)), xlim=c(31,81), ylab="Probability of failure", xlab="
```




```
plot(plotrange, exp(pred_odds)/(1+exp(pred_odds))*6, xlim=c(31,81), ylab="Expected number of failures",
```

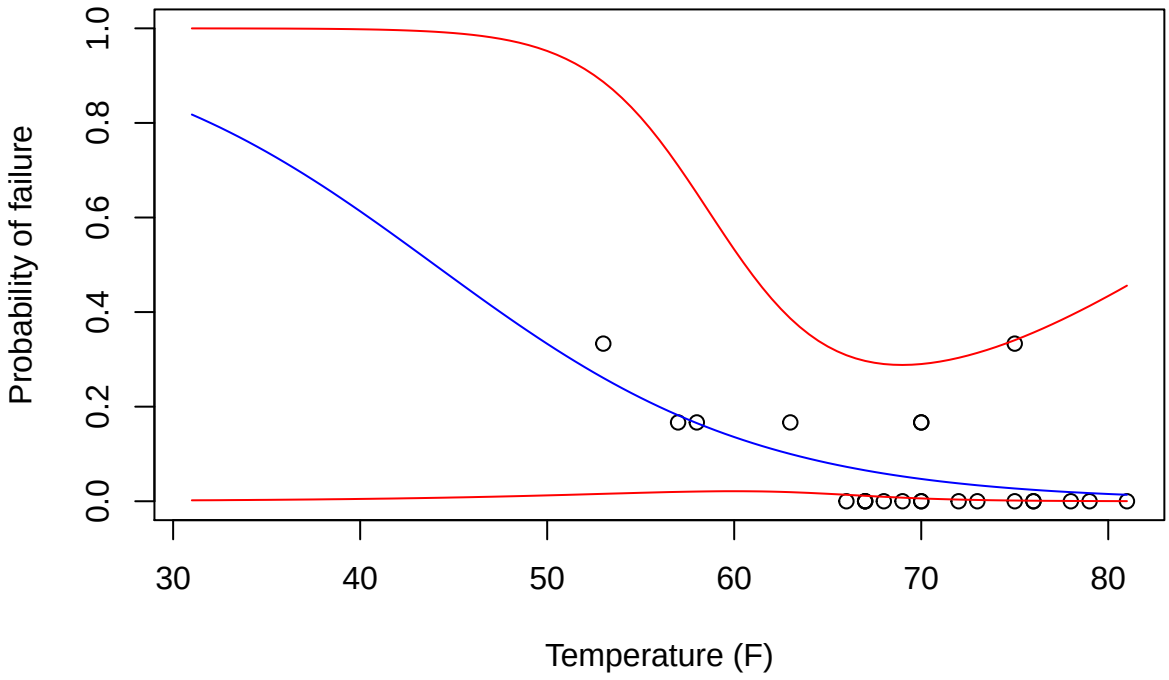


- (c) Include the 95% Wald confidence interval bands for π on the plot. Why are the bands much wider for lower temperatures than for higher temperatures?

- The CI bands are wider at lower temperatures because there is no data below 53ř.

```
ci.plot = function(model,xdata, alpha) {
  z = qnorm(1-alpha/2)
  predictions = predict(model, newdata=data.frame(Temp=xdata), type="link", se.fit=TRUE)
  lower_ci = predictions$fit - z*predictions$se.fit
  upper_ci = predictions$fit + z*predictions$se.fit
  list(lower=exp(lower_ci)/(1+exp(lower_ci)),upper=exp(upper_ci)/(1+exp(upper_ci)))
}

plot(data$Temp, data$Prob, xlim = c (31, 81), ylim=c(0,1), xlab="Temperature (F)", ylab="Probability of
curve (expr = predict (object = fit.b, newdata = data.frame(Temp = x), type = "response"), col = "blue"
curve (expr = ci.plot(fit.b,x,0.05)$lower, col = "red", add = TRUE, xlim = c (31, 81), ylim=c(0,1))
curve (expr = ci.plot(fit.b,x,0.05)$upper, col = "red", add = TRUE, xlim = c (31, 81), ylim=c(0,1))
```



- (d) The temperature was 31ř 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.

```
predicted_odds = predict(fit.b, newdata=data.frame(Temp=31), type="link", se.fit = TRUE)
predicted_prob = exp(predicted_odds$fit)/(1+exp(predicted_odds$fit))
ci_odds = c(predicted_odds$fit - qnorm(1-.05/2)*predicted_odds$se.fit, predicted_odds$fit + qnorm(1-.05,
ci = exp(ci_odds)/(1+exp(ci_odds))
"Predicted probability of at least one O-ring failure with 95% confidence interval:"
```

```
## [1] "Predicted probability of at least one O-ring failure with 95% confidence interval:"
```

```
paste(round(ci[1],5), "<", round(predicted_prob,5), "<", round(ci[2],5))
```

```
## [1] "0.00194 < 0.81777 < 0.9999"
```

- (e) Rather than using Wald or profile LR intervals for the probability of failure, Dalal et al. (1989) use a parameteric 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 $logit(\pi) = \beta_0 + \beta_1 Temp$; (2) estimate new models for each data set, say $logit(\hat{\pi}^*) = \hat{\beta}_0^* + \hat{\beta}_1^* Temp$; 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 31ř and 72ř.

```
bootstrap =function(test_values){
  samples = data[sample(nrow(data), size=23, replace=TRUE), ]
  fit.boot = suppressWarnings(glm(Prob ~ Temp, family=binomial(link=logit),data = samples, control = li
  predicted_odds = predict(fit.boot, newdata=data.frame(Temp=test_values), type="link")
```

```
  predicted_prob = exp(predicted_odds)/(1+exp(predicted_odds))
}
x = replicate(10000,bootstrap(c(31,72)))
xt = aperm(x,c(2,1))

paste("90% CI for 31F:")
```

```
## [1] "90% CI for 31F:"
```

```
quantile(xt[,1], probs=c(.05,.95))
```

```
##        5%       95%
## 0.1810934 0.9913927
```

```
paste("90% CI for 72F:")
```

```
## [1] "90% CI for 72F:"
```

```
quantile(xt[,2], probs=c(.05, .95))
```

```
##          5%         95%
## 0.006116177 0.079398505
```

- (f) Determine if a quadratic term is needed in the model for the temperature.
- Adding a quadratic term for temperature didn't improve the model very much.
- The residual deviance decreased slightly from 3.01 to 2.93 (~3%).
- The AIC improved from 7.2 to 9.2
- LRT doesn't indicate that temperature^2 is important
- Is there a physical reason why we would suspect temperature is quadratic?

```
library(car)
```

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

```
fit.c = glm(formula = Prob ~ Temp + I(Temp^2), family=binomial(link=logit),data = data)
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial
## glm!
```

```
summary(fit.c)
```

```
##
## Call:
## glm(formula = Prob ~ Temp + I(Temp^2), family = binomial(link = logit),
##     data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.34423  -0.29551  -0.25303  -0.00545   1.02920
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 22.126148  58.284482   0.380    0.704
## Temp        -0.650885   1.814484  -0.359    0.720
## I(Temp^2)    0.004141   0.013943   0.297    0.766
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4.0384  on 22  degrees of freedom
## Residual deviance: 2.9319  on 20  degrees of freedom
## AIC: 9.2373
##
## Number of Fisher Scoring iterations: 6
```

```
Anova(fit.c, test="LR")
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial
## glm!

## Warning in eval(family$initialize): non-integer #successes in a binomial
## glm!
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Prob
##           LR Chisq Df Pr(>Chisq)
## Temp       0.11980  1     0.7293
## I(Temp^2)  0.08245  1     0.7740
```

```
anova(fit.b, fit.c, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Prob ~ Temp
## Model 2: Prob ~ Temp + I(Temp^2)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        21     3.0144
## 2        20     2.9319  1  0.08245    0.774
```

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

    - Odds: The coefficient for temperature is -0.1156. The odds of failure decrease by $\exp(c\beta_1)$ for each unit increase in c. For example, a 1 degree increase in temperature yields a 0.89 decrease in the odds of failure.
    - Probability: The probability of failure are $\hat{\pi} = \frac{\exp(-0.1156Temp+5.0850)}{1+\exp(-0.1156Temp+5.0850)}$. Increase in temperature results in reduced probability of failure.
        - Does this need a confidence interval?

    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.