Logistic Regression Oring

Brandon Tao

October 12, 2018

0. Set up

```
library(faraway)
data(orings)
```

1. Background: Space shuttle O-ring seal damage

Study of possible causes of the Challenger explosion shortly after launch in January 1986 drew attention to the O-ring seals in the rocket boosters. The temperature at launch was much lower than at previous shuttle launches. At lower temperatures rubber becomes brittle and less effective as a sealant. Might this have been the cause of the explosion? Should the danger have been understood by the engineers?

Edward Tufte said that the failure to appreciate the danger was strongly associated with the poor graphics used to look at the data.
See his discussion and the graphics in Edward Tufte. 1997. Visual Explanations pp. 39-53

The goal here to model the data as presented by Julian Faraway. Faraway refers to Dala, Fowlkes and Hoadly (1989)for a more detailed description.

This dataset combines observations from 23 previous shuttle missions. Each shuttle has two boosters, each with three O-rings. The dependent variable is the number of the 6 O-rings that show evidence of damage due to blow by and erosion.

A simple model is that the observations from each shuttle is binomial dat. The sample size is six, the number of trials, n. The random variable, x, is the number of damaged O-rings. The 23 shuttles provide 23 experiments and are interested in estimating probability of damage as a function of the independent variable, temperature.

2. Look at the data

```
orings

##    temp damage
## 1    53      5
## 2    57      1
## 3    58      1
## 4    63      1
## 5    66      0
## 6    67      0
## 7    67      0
## 8    67      0
## 9    68      0
## 10   69      0
## 11   70      1
## 12   70      0
## 13   70      1
## 14   70      0
## 15   72      0
## 16   73      0
## 17   75      0
## 18   75      1
## 19   76      0
## 20   76      0
## 21   78      0
## 22   79      0
## 23   81      0
```

3. Regression Models for Binomial Data

With a simple regression model in mind a first thought might be to consider a model of the form:

1) p[i] = a + b * x[i] + e [i] for the ith mission where
2) p[i] is the probability of damage,
3) x[i] is the temperature at launch, and
4) e[i] is the error.

The model could lead to a nonsensical result for some temperatures because we know that probabilities are limited to values in the closed interval [0 1], but the model does not enforce this constraint.

Instead if we use the model
1. p[i] = exp(a + b * x[i])/ (1 + exp(a + b * x[i]), the probabilities are forced to be between 0 and 1.

Let z[i] = a + b*x[i]
1. The right hand side is of form exp(z[i]) / (1+exp(z[i]))

We omit the i subscript and consider the function p(z) = exp(z)/(1+exp(z)).

Since exp(z) >= 0 we can see p(z) is non-negative and less than 1 for finite values of z.

What happens when z approaches -infinity? 1) For negative z exp(z)is the same as 1/exp(|z|) 2)As |z| get large this approaches 0.
3) The denominator, 1 + 1/exp(|z|) approaches 1 4) The ratio approaches 0/1 = 0.

What happens when z approaches infinity? Consider adding and subtracting 1 from the numerator.

1) (1+exp(z) -1)/(1+exp(z)) = 2) 1 -1/(1+exp(z)) 3) The second term goes to 0. The ratio approaches 1.

Hence 0 <= p(z) <= 1

Solving p = exp(z)/(1+exp(z)) for z yields 1) z = log(p/(1-p)) 2) The right hand side is the log of the odds ratio p/(1-p).

Returning to the data at hand we have 1) a + b * x[i] = log(p[i]/(1-p[i]) The simple linear combination is effectively modeling the log odds ratio which is free to range from -infinity to infinity.

A host of other models for p[i] could be used. Any continuous monotone increasing cumulative distribution function with support from -infinity to infinity will work.

You might ask what about using the cumulative normal distribution? Yes, this works and the model has a name, the probit model.

The logit and probit models are often used.

A third sometimes-used model is the complementary log-log function. This has form z = log(-log(1-p))

4. Estimating model parameters

We want to model the probability p, (a parameter in the binomial family of probability distributions), as a function of a linear combination of predictor variables. For several families of probability distributions we can use the generalized linear model algorithm implemented in the glm() function.

Briefly, since this is binomial data and we will be assuming independent observations given the data. It is straight forward to write the log likelihood. The standard approach of obtaining parameter estimates by maximizing the log likelihood works. We can obtain estimates for a, b, and p. The theory also supports estimating standard errors for the maximum likelihood estimates. The p-values are derived from z-scores = estimate/standardError

3 5. Fitting three models

We started with a linear model and the use two general linear models.

5.1 Naive linear model

```
linearModel = lm(damage/6 ~ temp,data=orings)
summary(linearModel)
```

```
##
## Call:
## lm(formula = damage/6 ~ temp, data = orings)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.13786 -0.10345 -0.02369  0.06601  0.48345
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.21429    0.29993   4.049 0.000578 ***
## temp        -0.01631    0.00429  -3.801 0.001043 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.142 on 21 degrees of freedom
## Multiple R-squared:  0.4076, Adjusted R-squared:  0.3794
## F-statistic: 14.45 on 1 and 21 DF,  p-value: 0.001043
```

5.2 Logit model

```
logitModel =
  glm(cbind(damage,6-damage) ~ temp,
     family=binomial, data = orings)
summary(logitModel)
```

```
##
## Call:
## glm(formula = cbind(damage, 6 - damage) ~ temp, family = binomial,
##     data = orings)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9529  -0.7345  -0.4393  -0.2079   1.9565
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 11.66299    3.29626   3.538 0.000403 ***
## temp        -0.21623    0.05318  -4.066 4.78e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 38.898  on 22  degrees of freedom
## Residual deviance: 16.912  on 21  degrees of freedom
## AIC: 33.675
##
## Number of Fisher Scoring iterations: 6
```

5.3 Probit model

```
probitModel =
  glm(cbind(damage,6-damage) ~ temp,
     family=binomial(probit),data=orings)
summary(probitModel)
```

```
##
## Call:
## glm(formula = cbind(damage, 6 - damage) ~ temp, family = binomial(pr
obit),
##     data = orings)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0134  -0.7761  -0.4467  -0.1581   1.9983
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.59145    1.71055   3.269  0.00108 **
## temp        -0.10580    0.02656  -3.984 6.79e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 38.898  on 22  degrees of freedom
## Residual deviance: 18.131  on 21  degrees of freedom
## AIC: 34.893
##
## Number of Fisher Scoring iterations: 6
```

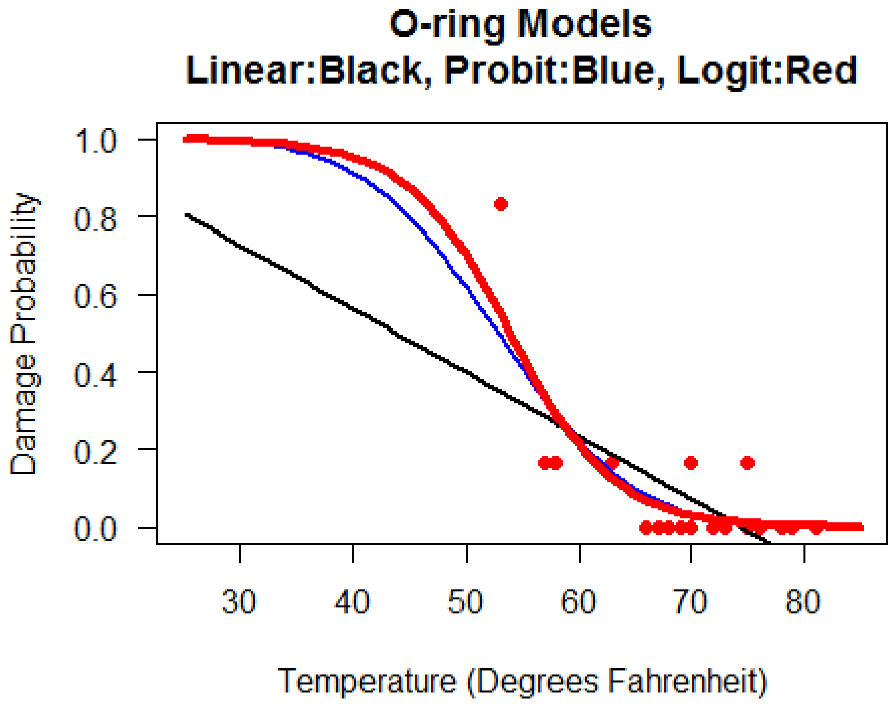
6. Plots of fitted values

```
plot(damage/6~temp,orings,
  xlim = c(25,85), ylim = c(0,1),las = 1,
  xlab = "Temperature (Degrees Fahrenheit)",
  ylab = "Damage Probability",
  pch = 21,bg = "red",col = "red",
  main = "O-ring Models\nLinear:Black, Probit:Blue, Logit:Red")

tempGrid = 25:85
a = coef(linearModel)
lines(tempGrid,a[1] + a[2] * tempGrid,col = "black",lwd = 2)

a = coef(probitModel)
lines(tempGrid,pnorm(a[1] + a[2]* tempGrid),col = "blue",lwd = 2)

a = coef(logitModel)
lines(tempGrid,ilogit(a[1] + a[2] * tempGrid),col = "red",lwd = 4)
```



7. Comment

The predicted temperature for the launch time was 26 to 29 degrees Fahrenheit. The plot from Section 6 shows extrapolated probabilities of damage for this interval. The probabilities for the two generalized linear models are about 1 for this interval. Even the high probabilities from the simple linear model would be cause for concern by the engineers, administrators and crew.

(Yes, the plot in Section 6 could be redesigned to draw more visual attention to this low temperature interval.)

Without the modeling, Tufte's position-along-a-scale plot of damage versus temperature would have caused concern. Yes, entertainment graphics gathered by Tufte (and likely shown in class) can fail to bring out the patterns in the data. In some cases this can lead to disastrous decisions.

Of course it is easy to be critical in hindsight. Nonetheless there is reason to push for quality graphics and for more education so a simple scatterplot is not considered too advanced for the general public.