5.3 - R: Logistic Regression Case Study

Stat 5100: Dr. Bean

See 1:15-2:10 of www.youtube.com/watch?v=j4JOjcDFtBE
and 3:31-4:22 of www.youtube.com/watch?v=gEjXjfxoNXM
(full text here: http://millercenter.org/scripps/archive/speeches/detail/3413)

The January 18, 1986 explosion of the space shuttle Challenger was investigated by the Presidential Commission on the Space Shuttle Challenger Accident. The Commission's 1986 report attributed the explosion to a burn through of an O-ring seal at a field joint in one of the solid-fuel rocket boosters. This 1986 launch was the 25th space shuttle launch. After each of the previous 24 launches, the solid rocket boosters were inspected. The following data are from the Commission's 1986 report, with the following variables:

Flight	an identifier code for the launch
Temp	temperature (degrees F) at launch
Damage	indicator of damage to the field joint (a missing value is recorded for one launch where the solid rocket boosters were not recovered)

Note that seven of the 24 launches experienced field joint damage but did not explode like the Challenger. The Challenger launch was Flight STS51L (not in these data) and the temperature was 31.

```
library(stat5100)
data(shuttle)

shuttle
```

```
##    Flight Temp Damage
## 1    STS1   66     NO
## 2    STS9   70     NO
## 3  STS51B   75     NO
## 4    STS2   70    YES
## 5  STS41B   57    YES
## 6  STS51G   70     NO
## 7    STS3   69     NO
## 8  STS41C   63    YES
## 9  STS51F   81     NO
## 10   STS4   80   <NA>
## 11 STS41D   70    YES
## 12 STS51I   76     NO
## 13   STS5   68     NO
## 14 STS41G   78     NO
## 15 STS51J   79     NO
## 16   STS6   67     NO
## 17 STS51A   67     NO
## 18 STS61A   75    YES
## 19   STS7   72     NO
## 20 STS51C   53    YES
## 21 STS61B   76     NO
## 22   STS8   73     NO
```

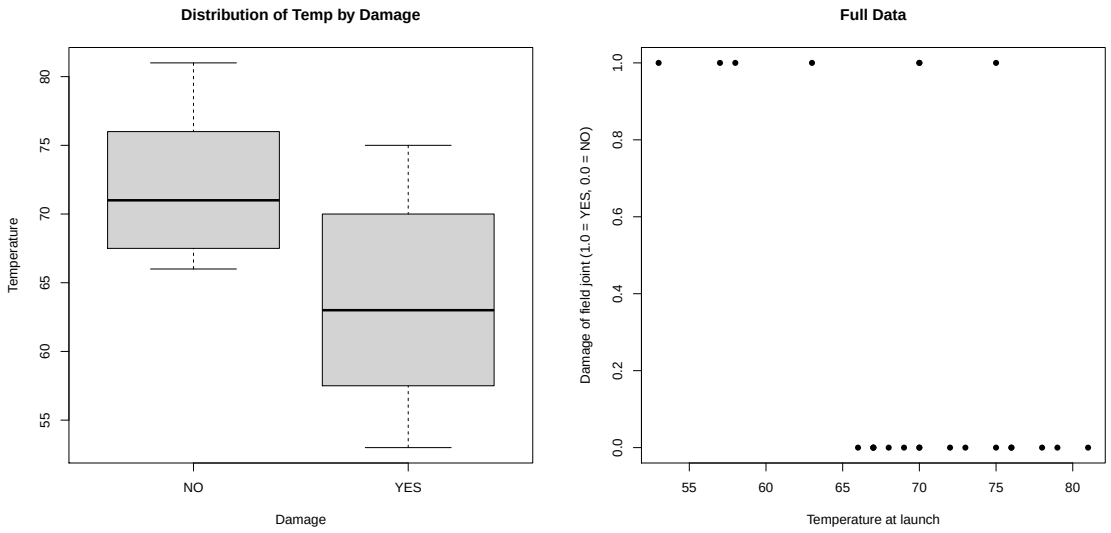
```
## 23 STS51D   67     NO
## 24 STS61C   58    YES
```

We will follow the following steps in this case study:

1. Visualize the data
2. Evaluate the probability of damage based on temperature
3. Check for influential observations and outliers
4. Calculate the probability of damage at temperature 31 (temperature at Challenger launch)
5. How is logistic regression different from ANOVA?

1. Visualize the data

```
# We will look at the distribution of temperature by damage
boxplot(shuttle$Temp ~ shuttle$Damage, main = "Distribution of Temp by Damage",
        xlab = "Damage", ylab = "Temperature")

# Full data, still separated by temperature
damage_numeric <- as.numeric(shuttle$Damage == "YES")
plot(shuttle$Temp, damage_numeric, main = "Full Data", xlab = "Temperature at launch",
     ylab = "Damage of field joint (1.0 = YES, 0.0 = NO)", pch = 16)
```



Based upon the above visualizations, we would conclude that damage mostly occurred to the field joint at lower temperatures.

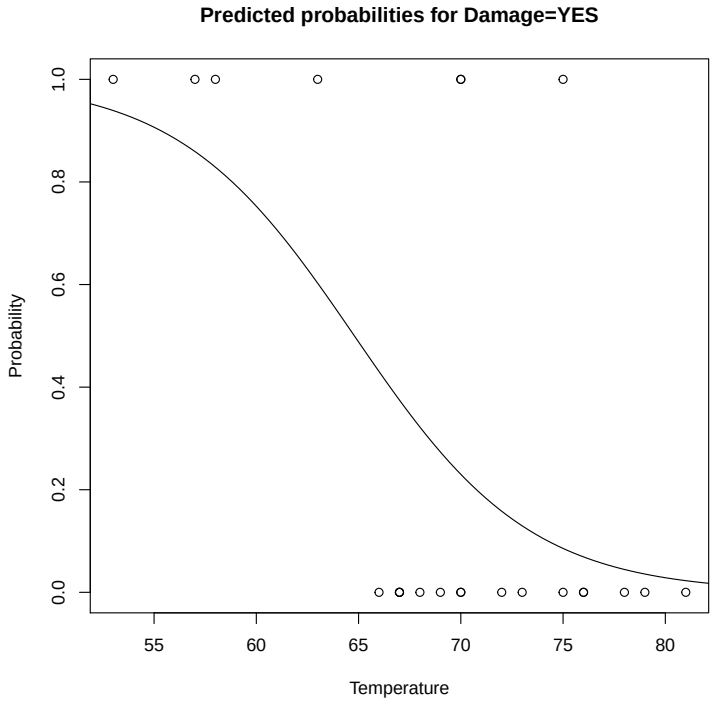
2. Evaluate the probability of damage based on temperature

```
# Convert damage to factor type first
shuttle$Damage <- as.factor(shuttle$Damage)

shuttle_logreg <- glm(Damage ~ Temp, data = shuttle,
                      family = "binomial")
summary(shuttle_logreg)
```

```
##
## Call:
## glm(formula = Damage ~ Temp, family = "binomial", data = shuttle)
##
## 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
##   (1 observation deleted due to missingness)
## AIC: 24.315
##
## Number of Fisher Scoring iterations: 5
```

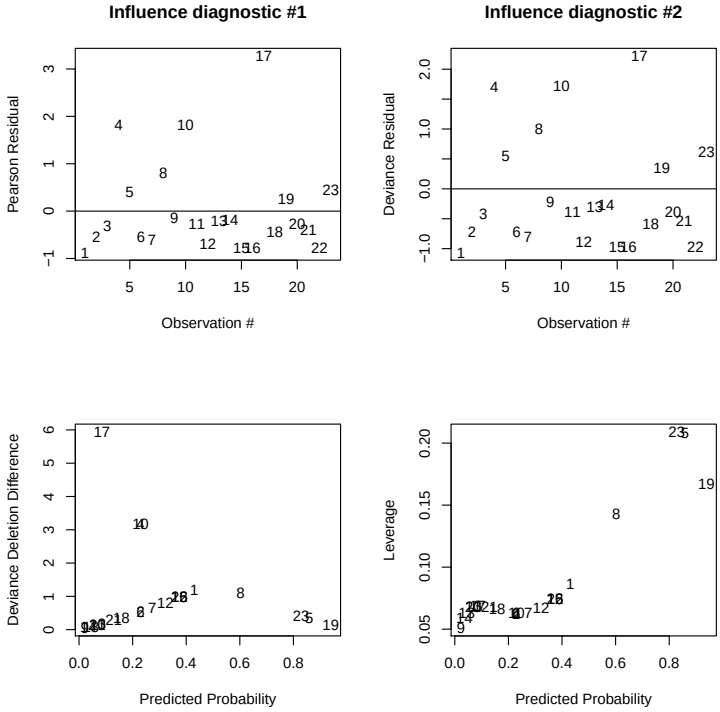
Now, let's create a fit plot showing the probability of damage for various levels of temperature.

```
temp_range <- seq(50, 85, length.out = 100)
pred_prob_damage <- predict(shuttle_logreg, newdata = data.frame(Temp = temp_range),
                            type = "response")

plot(shuttle$Temp, damage_numeric, main = "Predicted probabilities for Damage=YES",
     xlab = "Temperature", ylab = "Probability")
lines(temp_range, pred_prob_damage)
```



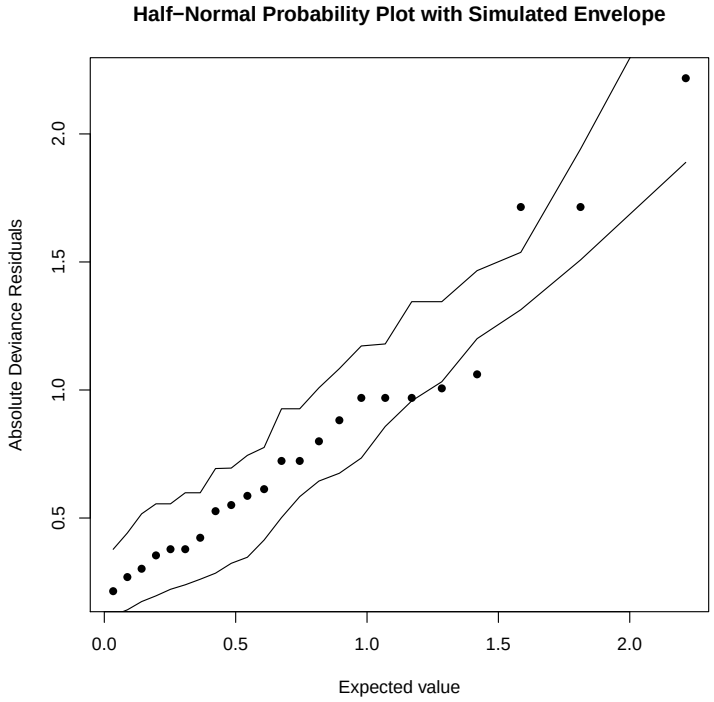
3. Check for influential observations and outliers

```
stat5100::logistic_influence_diagnostics(shuttle_logreg)
```



Now let's do an outlier check using the simulated envelope function. Note, however, that the optimization fails to converge for this example and thus the simulated envelope output is not reliable to use.

```
stat5100::simulated_envelope_logreg(shuttle_logreg)

## Warning:  glm.fit:  algorithm did not converge
## Warning:  glm.fit:  fitted probabilities numerically 0 or 1 occurred
```



Let's print some suspect observations from the influential diagnostics to get a sense for what types of observations they were:

```
shuttle[17, ]

##    Flight Temp Damage
## 17 STS51A   67     NO

shuttle[4, ]

##   Flight Temp Damage
## 4   STS2   70    YES

shuttle[10, ]

##    Flight Temp Damage
## 10   STS4   80   <NA>
```

Let's try refitting a model but excluding those observations.

```
new_shuttle <- shuttle[-c(4, 10, 17), ]
new_shuttle_logreg <- glm(Damage ~ Temp, data = new_shuttle, family = "binomial")
summary(new_shuttle_logreg)

##
## Call:
## glm(formula = Damage ~ Temp, family = "binomial", data = new_shuttle)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -1.0209  -0.6600  -0.3174   0.3151   2.3535
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  16.6681     8.1242   2.052   0.0402 *
## Temp         -0.2583     0.1198  -2.156   0.0311 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 25.127  on 20  degrees of freedom
## Residual deviance: 16.330  on 19  degrees of freedom
## AIC: 20.33
##
## Number of Fisher Scoring iterations: 5
```

Note, however, that perfect separation between the 0's and 1's across temperature causes the maximum likelihood optimization to fail to converge. This causes the standard errors of the coefficients to become very unstable which ruins inference. The solution to this is to use a penalized regression version of OLS. Unfortunately, none of our visual diagnostics work for this alternative form of logistic regression.

```
# Create a design matrix.
form <- Damage ~ Temp
model_frame <- model.frame(form, data = new_shuttle)

# glmnet requires at least two explanatory variables. We have included the
# intercept as the "second" variable but should NEVER do this if we have
# more than one explanatory variable to work with.
# Alpha = 0, makes this akin to "ridge regression"
penalized_logit <- glmnet::glmnet(model.matrix(form, model_frame),
                                  model.response(model_frame),
                                  alpha = 0,
                                  family = "binomial",
                                  lambda = 0.001)
```

4. Calculate the probability of damage at temperature 31

```
predict(penalized_logit, newx = matrix(c(31, 1), ncol = 2), type = "response")

##             s0
## [1,] 0.9999999
```

5. How is logistic regression different from ANOVA?

```
anova(lm(Temp ~ Damage, data = shuttle))

## Analysis of Variance Table
##
## Response: Temp
##           Df Sum Sq Mean Sq F value   Pr(>F)
## Damage     1 344.47  344.47  9.6301 0.005383 **
```

```
## Residuals 21 751.18   35.77
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm(Temp ~ Damage, data = shuttle))
```

```
##
## Call:
## lm(formula = Temp ~ Damage, data = shuttle)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7143  -5.1250  -0.7143   4.8750  11.2857
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   72.125      1.495  48.237  < 2e-16 ***
## DamageYES     -8.411      2.710  -3.103  0.00538 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.981 on 21 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3144,Adjusted R-squared:  0.2818
## F-statistic:  9.63 on 1 and 21 DF,  p-value: 0.005383
```

```
# Plot the linear model:
plot(damage_numeric, shuttle$Temp, xlab = "Damage (1=YES, 0=NO)",
     ylab = "Temperature", main = "Fit plot for Temp")
abline(a = 72.125, b = -8.411)
```