MichaelY - HW8 - Multiple and Logistic regression

Michael Y.

May 5th, 2019

```
###setwd("c:/Users/Michael/DROPBOX/priv/CUNY/MSDS/201902-Spring/DATA606-Jason/Homework")
```

Homework - Chapter 8 - Multiple and Logistic Regression (pp.372-394)

Practice : 8.1, 8.3, 8.7, 8.15, 8.17 (pp.395-404)

Datasets:

8.1 - babies

8.3 - babies

8.7 - babies

8.15 - possum

8.17 - possum

Exercises: 8.2, 8.4, 8.8, 8.16, 8.18 (pp.395-404)

Datasets:

8.2 - babies

8.4 - absenteeism

8.8 - absenteeism

8.16 - orings

8.18 - orings

#.#

Exercise 8.2 Baby weights, Part II.

Exercise 8.1 introduces a data set on birth weight of babies.

Another variable we consider is parity, which is 0 if the child is the first born, and 1 otherwise.

The summary table below shows the results of a linear regression model for predicting the average birth weight of babies, measured in ounces, from parity.

```
# look at the exact data set referenced
data("babies")
summary(babies)
```

```
##       case             bwt          gestation         parity            age
##  Min.   :   1.00   Min.   : 55.00   Min.   :148.00   Min.   :0.00000   Min.   :15.000
##  1st Qu.: 309.75   1st Qu.:108.75   1st Qu.:272.00   1st Qu.:0.00000   1st Qu.:23.000
##  Median : 618.50   Median :120.00   Median :280.00   Median :0.00000   Median :26.000
##  Mean   : 618.50   Mean   :119.58   Mean   :279.34   Mean   :0.25485   Mean   :27.255
##  3rd Qu.: 927.25   3rd Qu.:131.00   3rd Qu.:288.00   3rd Qu.:1.00000   3rd Qu.:31.000
##  Max.   :1236.00   Max.   :176.00   Max.   :353.00   Max.   :1.00000   Max.   :45.000
##                                     NA's   :13                         NA's   :2
##      height          weight           smoke
##  Min.   :53.000   Min.   : 87.00   Min.   :0.00000
##  1st Qu.:62.000   1st Qu.:114.75   1st Qu.:0.00000
##  Median :64.000   Median :125.00   Median :0.00000
##  Mean   :64.048   Mean   :128.63   Mean   :0.39478
##  3rd Qu.:66.000   3rd Qu.:139.00   3rd Qu.:1.00000
##  Max.   :72.000   Max.   :250.00   Max.   :1.00000
##  NA's   :22       NA's   :36       NA's   :10
```

Model

```
p802model <- lm(bwt ~ parity, data=babies)
p802model
```

```
##
## Call:
## lm(formula = bwt ~ parity, data = babies)
##
## Coefficients:
## (Intercept)       parity
##    120.0684      -1.9287
```

```
summary(p802model)
```

```
##
## Call:
## lm(formula = bwt ~ parity, data = babies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.068 -11.068   0.396  10.932  57.860
```

```
##
## Coefficients:
##              Estimate Std. Error  t value            Pr(>|t|)
## (Intercept) 120.06840    0.60052 199.9422 <0.0000000000000002 ***
## parity       -1.92872    1.18954  -1.6214              0.1052
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.224 on 1234 degrees of freedom
## Multiple R-squared:  0.0021259,  Adjusted R-squared:  0.0013173
## F-statistic:  2.629 on 1 and 1234 DF,  p-value: 0.10519
```

(a) Write the equation of the regression line.

$$\widehat{BirthWeight} = 120.06840391 - 1.92872137 \times parity$$

(b) Interpret the slope in this context, and calculate the predicted birth weight of first borns and others.

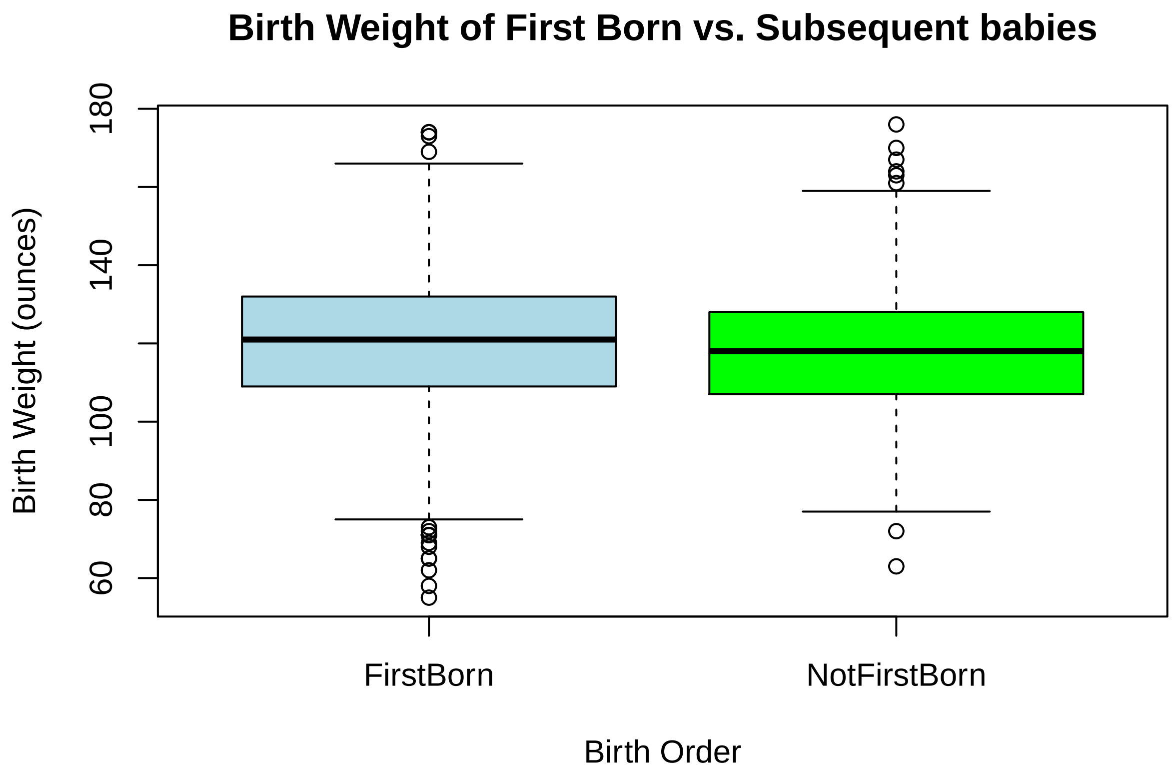
The "slope" functions to separate the average birthweight into two horizontal lines:

For babies who are firstborn, their average birth weight is 120.06840391 ounces.

For babies who are not firstborn, their average birth weight is 118.13968254 ounces.

As such, the "slope" simply adjusts the intercept into separate values for firstborn vs. not firstborn.

```
babies$par="FirstBorn"
babies$par[babies$parity==1]="NotFirstBorn"
babies$par2=as.factor(babies$par)
boxplot(babies$bwt~babies$par2,
        main="Birth Weight of First Born vs. Subsequent babies",
        xlab="Birth Order",
        ylab="Birth Weight (ounces)",
        col=c("lightblue","green"),
        show.names=T)
```



(c) Is there a statistically significant relationship between the average birth weight and parity?

No, because the p-value is .1052, the relationship is not considered statistically significant.

#·#

Exercise 8.4 Absenteeism.

Researchers interested in the relationship between absenteeism from school and certain demographic characteristics of children collected data from 146 randomly sampled students in rural New SouthWales, Australia, in a particular school year.

Below are three observations from this data set.

```
# look at the exact data set referenced

#install.packages("MASS")
#data(package="MASS")
data(quine, package="MASS")
summary(quine)
```

```
##  Eth     Sex     Age      Lrn          Days
##  A:69    F:80    F0:27    AL:83    Min.   : 0.000
##  N:77    M:66    F1:46    SL:63    1st Qu.: 5.000
```

```
##                F2:40           Median :11.000
##                F3:33           Mean   :16.459
##                                3rd Qu.:22.750
##                                Max.   :81.000
```

```
head(quine,n=2)
```

```
##   Eth Sex Age Lrn Days
## 1   A   M  F0  SL    2
## 2   A   M  F0  SL   11
```

```
tail(quine,n=1)
```

```
##     Eth Sex Age Lrn Days
## 146   N   F  F3  AL   37
```

The summary table below shows the results of a linear regression model for predicting the average number of days absent based on

ethnic background (eth: A - aboriginal, N - not aboriginal),

sex (sex: F - female, M - male), and

learner status (lrn: AL - average learner, SL - slow learner).

```
p804model <- lm(Days ~ Eth + Sex + Lrn, data=quine)
p804model
```

```
##
## Call:
## lm(formula = Days ~ Eth + Sex + Lrn, data = quine)
##
## Coefficients:
## (Intercept)          EthN          SexM         LrnSL
##     18.9318       -9.1122        3.1043        2.1542
```

```
summary(p804model)
```

```
##
## Call:
## lm(formula = Days ~ Eth + Sex + Lrn, data = quine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.1903 -10.0780  -4.9279   5.7680  59.9140
##
## Coefficients:
##             Estimate Std. Error t value         Pr(>|t|)
## (Intercept)  18.9318     2.5704  7.3652 0.00000000001319 ***
## EthN         -9.1122     2.5989 -3.5063        0.0006087 ***
## SexM          3.1043     2.6371  1.1771        0.2411076
## LrnSL         2.1542     2.6505  0.8127        0.4177317
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 15.673 on 142 degrees of freedom
## Multiple R-squared:  0.089331,   Adjusted R-squared:  0.070092
## F-statistic: 4.6431 on 3 and 142 DF,  p-value: 0.003967
```

(a) Write the equation of the regression line.

```
p804coefs <- p804model$coefficients
p804coefs
```

```
## (Intercept)        EthN        SexM       LrnSL
##  18.9318482  -9.1122415   3.1042554   2.1541571
```

$$\widehat{Days} = 18.93184821 + (-9.11224147) I_{Eth_{(0:Aboriginal)}^{(1:NotAboriginal)}} + (3.10425539) I_{Sex_{(0:female)}^{(1:male)}} + (2.15415712) I_{Lrn_{(0:AverageLearner)}^{(1:SlowLearner)}}$$

```
Q... = mean(quine$Days)
print("Eth:")
Q0.. = mean((subset(quine, Eth=="A" ))$Days)
Q1.. = mean((subset(quine, Eth=="N" ))$Days)
c(Q0.. , Q1.., -Q0.. + Q1..)

print("Sex:")
Q.0. = mean((subset(quine, Sex=="F" ))$Days)
Q.1. = mean((subset(quine, Sex=="M" ))$Days)
c(Q.0. , Q.1., -Q.0. + Q.1.)

print("Lrn:")
Q..0 = mean((subset(quine, Lrn=="AL" ))$Days)
Q..1 = mean((subset(quine, Lrn=="SL" ))$Days)
c(Q..0 , Q..1, -Q..0 + Q..1)

Q000 = mean((subset(quine, Eth=="A" & Sex=="F" & Lrn=="AL"))$Days)
Q001 = mean((subset(quine, Eth=="A" & Sex=="F" & Lrn=="SL"))$Days)
Q010 = mean((subset(quine, Eth=="A" & Sex=="M" & Lrn=="AL"))$Days)
Q011 = mean((subset(quine, Eth=="A" & Sex=="M" & Lrn=="SL"))$Days)
Q100 = mean((subset(quine, Eth=="N" & Sex=="F" & Lrn=="AL"))$Days)
Q101 = mean((subset(quine, Eth=="N" & Sex=="F" & Lrn=="SL"))$Days)
Q110 = mean((subset(quine, Eth=="N" & Sex=="M" & Lrn=="AL"))$Days)
Q111 = mean((subset(quine, Eth=="N" & Sex=="M" & Lrn=="SL"))$Days)
```

(b) Interpret each one of the slopes in this context.

Eth(N):

All other things equal, children who are not aboriginal (Eth=N)are expected to be absent on fewer days than children who are aboriginal (Eth=A); the difference is -9.11224147 days

Sex(M):

All other things equal, children who are MALE (Sex=M) are expected to be absent on 3.10425539 more days than children who are FEMALE (Sex=M)

Lrn(SL):

All other things equal, children who are Slow Learners (Lrn=SL) are expected to be absent on 2.15415712 more days than children who are Average Learners (Lrn=AL)

(c) Calculate the residual for the first observation in the data set: a student who is aboriginal, male, a slow learner, and missed 2 days of school.

```
p804coefs
```

```
## (Intercept)        EthN        SexM       LrnSL
##  18.9318482  -9.1122415   3.1042554   2.1541571
```

```
FirstStudent <- quine[1,]
FirstStudent
```

```
##   Eth Sex Age Lrn Days
## 1   A   M  F0  SL    2
```

```
# identify which coefficients are to be used, i.e.,
# (1 for intercept, 1 if NotAboriginal, 1 if Male, 1 if Slow Learner)

# This observation:
# (1 for intercept, 0 for Aboriginal, 1 for Male, 1 for Slow Learner)
q = c(1,0,1,1)
q
```

```
## [1] 1 0 1 1
```

```
# Compute coefficients are to be used (Aboriginal is set to zero, since it's base case)
p804coefs * q
```

```
## (Intercept)        EthN        SexM       LrnSL
##  18.9318482   0.0000000   3.1042554   2.1541571
```

```
# compute the dot product
FirstStudentEstimatedDays <- as.numeric(p804coefs %*% q)
print(paste("FirstStudentEstimatedDays = ", FirstStudentEstimatedDays))
```

```
## [1] "FirstStudentEstimatedDays =  24.1902607210836"
```

```
# display the actual number of days absent
FirstStudentActualDays <- FirstStudent$Days
print(paste("FirstStudentActualDays = ", FirstStudentActualDays))
```

```
## [1] "FirstStudentActualDays =  2"
```

```
# compute the residual: Actual minus estimated
ManualFirstStudentResidual <- FirstStudentActualDays - FirstStudentEstimatedDays
ManualFirstStudentResidual
```

```
## [1] -22.190261
```

```r
# compare vs. the value returned by the model
ModelFirstStudentResidual <- p804model$residuals[1]
ModelFirstStudentResidual
```

```
##          1
## -22.190261
```

For the first student:

The estimated number of days absent is 24.19026072

The actual number of days absent was 2

The residual for the first student is -22.19026072 .

This matches the number returned by the model, which is -22.19026072 .

(d) The variance of the residuals is 240.57, and the variance of the number of absent days for all students in the data set is 264.17. Calculate the R^2 and the adjusted R^2. Note that there are 146 observations in the data set.

```r
ResidVariance <- var(p804model$residuals)
ResidVariance
```

```
## [1] 240.56881
```

```r
AbsentDaysVariance <- var(quine$Days)
AbsentDaysVariance
```

```
## [1] 264.16726
```

```r
n = dim(quine)[1]
n
```

```
## [1] 146
```

```r
# calculate the R2 manually
manual_R2 <- 1 - ResidVariance/AbsentDaysVariance
manual_R2
```

```
## [1] 0.089331491
```

```r
# check vs. the value returned by the regression
model_R2 <- summary(p804model)$r.squared
model_R2
```

```
## [1] 0.089331491
```

```r
# calculate the adjR2 manually
manual_adjR2 <- 1 - ResidVariance/AbsentDaysVariance * (n-1)/(n-3-1)
manual_adjR2
```

```
## [1] 0.070092016
```

```
# check vs. the value returned by the regression
model_adjR2 <- summary(p804model)$adj.r.squared
model_adjR2
```

```
## [1] 0.070092016
```

The R^2 is 0.08933149 and the $Adjusted R^2$ is 0.07009202 .

These match the values returned by the regression model: 0.08933149 and 0.07009202 .

#·#

Exercise 8.8 Absenteeism, Part II.

Exercise 8.4 considers a model that predicts the number of days absent using three predictors: ethnic background (eth), gender (sex), and learner status (lrn).

The table below shows the adjusted R-squared for the model as well as adjusted R-squared values for all models we evaluate in the first step of the backwards elimination process.

```
tabl=data.frame(matrix(
c(
c("Fullmodel", 0.0701),
c("Noethnicity", -0.0033),
c("Nosex", 0.0676),
c("No learner status", 0.0723)
),
nrow = 4, ncol = 2, byrow = T,
dimnames=list(
c(1,2,3,4),
c("Model","Adjusted R2")
)),stringsAsFactors = F)
tabl
```

```
##               Model Adjusted.R2
## 1         Fullmodel      0.0701
## 2       Noethnicity     -0.0033
## 3             Nosex      0.0676
## 4 No learner status      0.0723
```

Which, if any, variable should be removed from the model first?

The variable “Learner Status” (`lrn`) should be dropped first, because doing so increases the adjusted-R^2 the most.

Confirm by running backward-stepwise algorithm:

```
# look at the exact data set referenced
require(olsrr)
```

```
## Loading required package: olsrr
```

```
## Warning: package 'olsrr' was built under R version 3.5.3
##
## Attaching package: 'olsrr'
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
ols_step_backward_p(p804model,prem = .3,details = T)
```

```
## Backward Elimination Method
## ---------------------------
##
## Candidate Terms:
##
## 1 . Eth
## 2 . Sex
## 3 . Lrn
##
## We are eliminating variables based on p value...
##
## - Lrn
##
## Backward Elimination: Step 1
##
##  Variable Lrn Removed
##
##                          Model Summary
## ----------------------------------------------------------------
## R                       0.292       RMSE                 15.655
## R-Squared               0.085       Coef. Var            95.114
## Adj. R-Squared          0.072       MSE                 245.068
## Pred R-Squared          0.046       MAE                  11.743
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## --------------------------------------------------------------------
##                 Sum of
##                Squares         DF    Mean Square      F        Sig.
## --------------------------------------------------------------------
## Regression     3259.515          2       1629.757     6.65    0.0017
## Residual      35044.739        143        245.068
## Total         38304.253        145
## --------------------------------------------------------------------
##
##                                  Parameter Estimates
## ----------------------------------------------------------------------------------------
##       model      Beta    Std. Error    Std. Beta      t        Sig       lower     upper
## ----------------------------------------------------------------------------------------
## (Intercept)    19.984         2.218                 9.010    0.000     15.600    24.368
##        EthN    -9.065         2.595       -0.279    -3.493    0.001    -14.194    -3.935
##        SexM     2.778         2.603        0.085     1.067    0.288     -2.368     7.923
```

```
## ----------------------------------------------------------------------------------------
##
##
##
## No more variables satisfy the condition of p value = 0.3
##
##
## Variables Removed:
##
## - Lrn
##
##
## Final Model Output
## ------------------
##
##                          Model Summary
## ----------------------------------------------------------------
## R                       0.292       RMSE                15.655
## R-Squared               0.085       Coef. Var           95.114
## Adj. R-Squared          0.072       MSE                245.068
## Pred R-Squared          0.046       MAE                 11.743
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## ---------------------------------------------------------------------
##                  Sum of
##                 Squares         DF    Mean Square      F        Sig.
## ---------------------------------------------------------------------
## Regression     3259.515          2       1629.757     6.65    0.0017
## Residual      35044.739        143        245.068
## Total         38304.253        145
## ---------------------------------------------------------------------
##
##                                  Parameter Estimates
## ----------------------------------------------------------------------------------------
##       model      Beta    Std. Error    Std. Beta      t        Sig       lower     upper
## ----------------------------------------------------------------------------------------
## (Intercept)    19.984         2.218                9.010    0.000     15.600    24.368
##        EthN    -9.065         2.595       -0.279   -3.493    0.001    -14.194    -3.935
##        SexM     2.778         2.603        0.085    1.067    0.288     -2.368     7.923
## ----------------------------------------------------------------------------------------

##
##
##                            Elimination Summary
## ------------------------------------------------------------------------
##         Variable                 Adj.
## Step    Removed     R-Square    R-Square     C(p)        AIC        RMSE
## ------------------------------------------------------------------------
##    1    Lrn           0.0851      0.0723    2.6605    1222.5231    15.6547
## ------------------------------------------------------------------------
```

Confirmed; `lrn` is removed first.

#.#

Exercise 8.16 Challenger disaster, Part I.

On January 28, 1986, a routine launch was anticipated for the Challenger space shuttle. Seventy-three seconds into the flight, disaster happened: the shuttle broke apart, killing all seven crew members on board.

An investigation into the cause of the disaster focused on a critical seal called an O-ring, and it is believed that damage to these O-rings during a shuttle launch may be related to the ambient temperature during the launch.

The table below summarizes observational data on O-rings for 23 shuttle missions, where the mission order is based on the temperature at the time of the launch.

Temp gives the temperature in Fahrenheit,

Damaged represents the number of damaged O-rings, and

Undamaged represents the number of O-rings that were not damaged.

```r
# look at the exact data set referenced
data("orings")
summary(orings)
```

```
##       temp            damage
##  Min.   :53.000   Min.   :0.00000
##  1st Qu.:67.000   1st Qu.:0.00000
##  Median :70.000   Median :0.00000
##  Mean   :69.565   Mean   :0.47826
##  3rd Qu.:75.000   3rd Qu.:1.00000
##  Max.   :81.000   Max.   :5.00000
```

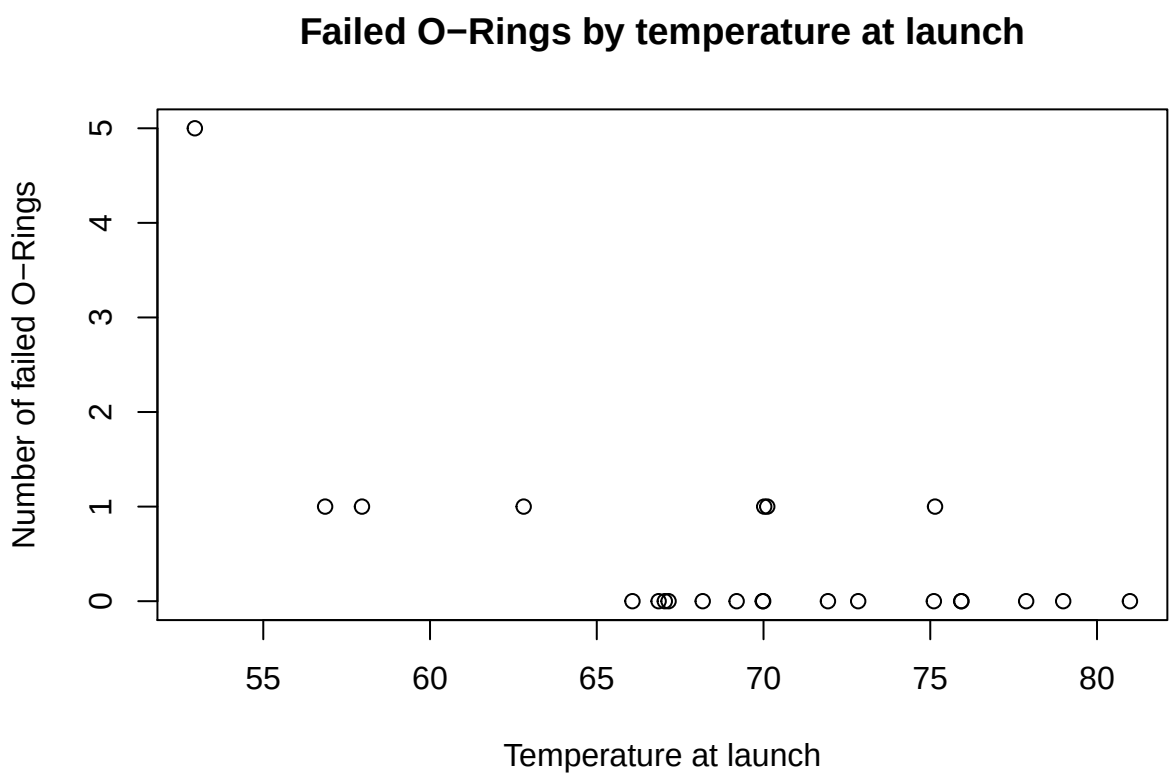
```r
orings$undamaged = 6 - orings$damage
orings
```

```
##    temp damage undamaged
## 1    53      5         1
## 2    57      1         5
## 3    58      1         5
## 4    63      1         5
## 5    66      0         6
## 6    67      0         6
## 7    67      0         6
## 8    67      0         6
## 9    68      0         6
## 10   69      0         6
## 11   70      1         5
## 12   70      0         6
## 13   70      1         5
## 14   70      0         6
```

```
## 15   72       0          6
## 16   73       0          6
## 17   75       0          6
## 18   75       1          5
## 19   76       0          6
## 20   76       0          6
## 21   78       0          6
## 22   79       0          6
## 23   81       0          6
```

(a) Each column of the table above represents a different shuttle mission. Examine these data and describe what you observe with respect to the relationship between temperatures and damaged O-rings.

```
plot(orings$damage~jitter(orings$temp),
     xlab="Temperature at launch",
     ylab="Number of failed O-Rings",
     main="Failed O-Rings by temperature at launch")
```



For higher temperatures, no o-rings are damaged.

As temperature decreases, there are cases where 1 o-ring is damaged.

When temperature is quite low (53 degrees), most of the o-rings are damaged.

(b) Failures have been coded as 1 for a damaged O-ring and 0 for an undamaged O-ring, and a logistic regression model was fit to these data. A summary of this model is given below.

```
## a messy way to create a 138x2 matrix of (temperature, failure or success)
## necessary for the regression
failarray138=array(0,dim=138)
temperaturearray=array(rep(orings[,"temp"],6),dim=c(1,138))  # 138 entries

failarray23=orings[,2] # 5 1 1 1 0 0 0 0 0 0 1 0 1 0 0 0 0 1 0 0 0 0 0
fail1=array(unlist(lapply(X = failarray23,FUN = min,1)),dim=23) # 1 1 1 1 0 0 0 0 0 0 1 0 1 0 0 0 0 1 0
failarray23 = failarray23 - fail1 # 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
fail2=array(unlist(lapply(X = failarray23,FUN = min,1)),dim=23)
failarray23 = failarray23 - fail2 # 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
fail3=array(unlist(lapply(X = failarray23,FUN = min,1)),dim=23)
failarray23 = failarray23 - fail3 # 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
fail4=array(unlist(lapply(X = failarray23,FUN = min,1)),dim=23)
failarray23 = failarray23 - fail4 # 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
fail5=array(unlist(lapply(X = failarray23,FUN = min,1)),dim=23)
failarray23 = failarray23 - fail5 # 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
fail6=array(unlist(lapply(X = failarray23,FUN = min,1)),dim=23)
failarray138 = array(c(fail1,fail2,fail3,fail4,fail5,fail6),dim=c(1,138))
failsarray=t(rbind(temperaturearray,failarray138))
colnames(failsarray)=c("temp","failure")
fails=data.frame(failsarray)

failsbytemp=fails[order(fails$temp),]
failsarraybytemp=cbind(failsbytemp$temp,failsbytemp$failure)
colnames(failsarraybytemp)=c("temp","failure")
failsbytemp=data.frame(failsarraybytemp)
```

Match the model shown in the text

```
fail_model=glm(failure~temp, data=failsbytemp,family = binomial)
summary(fail_model)
```

```
##
## Call:
## glm(formula = failure ~ temp, family = binomial, data = failsbytemp)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.26457  -0.33952  -0.24715  -0.12991   3.02159
##
## Coefficients:
##              Estimate Std. Error z value   Pr(>|z|)
## (Intercept) 11.662990   3.296157  3.5384  0.0004026 ***
## temp        -0.216234   0.053175 -4.0665 0.00004773 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 76.7448  on 137  degrees of freedom
## Residual deviance: 54.7594  on 136  degrees of freedom
## AIC: 58.7594
##
## Number of Fisher Scoring iterations: 6
```

```
logit_predictions=t(t(predict(fail_model)))
prob_predictions=logistic(logit_predictions)
predicts = cbind(fails,prob_predictions)
```

Describe the key components of this summary table in words.

The model estimates the logit of the probability that an o-ring will fail.

A positive value for the logit corresponds to a probability greater than 0.5, while a negative value for the logit corresponds to a probability of failure lower than 0.5 .

The logistic function converts from the logit to the probability, using the formula

$$\hat{p_i} = \frac{e^{logit(\hat{p_i})}}{1 + e^{logit(\hat{p_i})}}$$

The value `11.6630` for the intercept indicates the value of the logit if the temperature were zero.

This would correspond to a probability of failure of 0.99999139 , which is near certainty.

The value `-0.2162` measures the estimated reduction in the value of the logit for each degree increase in temperature above zero, where the scale is in fahrenheit.

This means that at a temperature of about 54 degrees the logit would be zero, which would correspond to a probabilty of failure of 0.5 . The higher the temperature, the smaller the value of the logit, which corresponds to a successive reduction in the probability of failure.

The standard error of 3.2963 measures the dispersion of the estimate of the intercept,

while the standard error of 0.0352 measures the dispersion of the estimated coefficient on the temperature.

Given that there are a total of n=138 observations (there are 23 missions, with 6 o-rings on each mission, for a total of 6*23 = 138), the standard error is calculated as

$$std.error = \frac{std.deviation}{\sqrt{138}}$$

The Z-value is simply the estimate divided by the standard error, or

$$z.value = \frac{estimate}{std.error}$$

which measures how many standard deviations the estimate is away from the zero.

In this case, the estimate for the intercept is $\frac{11.6630}{3.2963} = 3.54$ **standard deviations above zero, while the estimate for the temperatire is** $\frac{3.2963}{0.0532} = -4.07$ **standard deviations below zero .**

The Pr(>|z|), also known as the p-value, is the measure of the area under the tail from the Z.value to the corresponding end of the distribution. In each case, the value is close to zero. This indicates that the result is statistically significant.

(c) Write out the logistic model using the point estimates of the model parameters.

$$logit(\hat{p_i}) = ln\left(\frac{\hat{p_i}}{1-\hat{p_i}}\right) = 11.6630 - 0.2162 \times temperature_i$$

(d) Based on the model, do you think concerns regarding O-rings are justified? Explain.

Yes, at lower temperatures, O-rings appear to be more likely to fail.

#.#

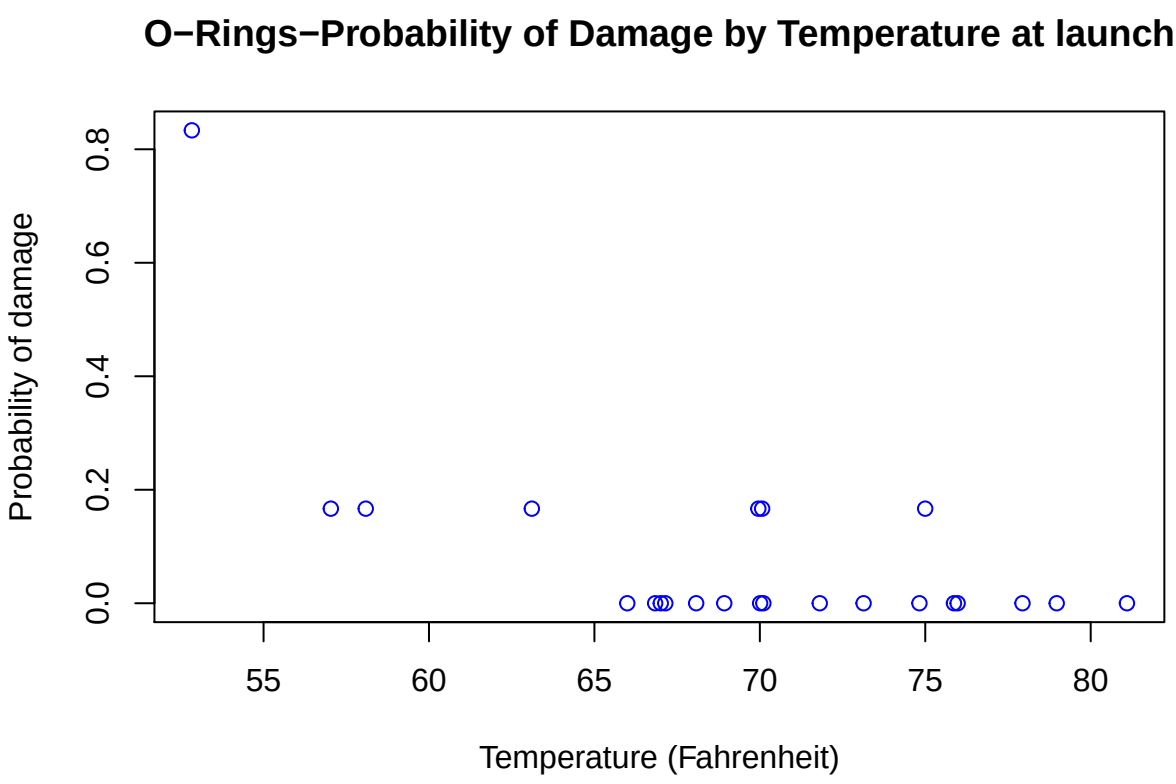
Exercise 8.18 Challenger disaster, Part II.

Exercise 8.16 introduced us to O-rings that were identified as a plausible explanation for the breakup of the Challenger space shuttle 73 seconds into takeoff in 1986.

The investigation found that the ambient temperature at the time of the shuttle launch was closely related to the damage of O-rings, which are a critical component of the shuttle.

See this earlier exercise if you would like to browse the original data.

```
orings$failure=orings$damage/6.0
plot(orings$failure~jitter(orings$temp),  col="blue",
     xlab="Temperature (Fahrenheit)",
     ylab="Probability of damage",
     main="O-Rings-Probability of Damage by Temperature at launch"
)
```



(a) The data provided in the previous exercise are shown in the plot. The logistic model fit to these data may be written as

$$log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \quad 11.6630 \quad - \quad (0.2162)Temperature$$

where $\hat{p}$ is the model-estimated probability that an O-ring will become damaged.

Use the model to calculate the probability that an O-ring will become damaged at each of the following ambient temperatures: 51, 53, and 55 degrees Fahrenheit.

```
# make a temperature list of the three temperatures requested
temperature_list = seq(51,55,2)
temperature_list
```

```
## [1] 51 53 55
```

```
# predict the logit for each of the three temperatures
logit_preds = predict.glm(fail_model,newdata=data.frame(temp=temperature_list))
names(logit_preds) = temperature_list
logit_preds
```

```
##          51          53          55
##  0.63507282  0.20260550 -0.22986183
```

```
# The desired probabilities for the three temperatures
prob_preds = logistic(logit_preds)
```

```
print("The desired probabilities are:")
```

```
## [1] "The desired probabilities are:"
```

```
print(t(t(prob_preds)))
```

```
##          [,1]
## 51 0.65363882
## 53 0.55047882
## 55 0.44278623
```

The desired probabilities at temperatures (51, 53, 55) are (0.65363882, 0.55047882, 0.44278623) .

```
# compute the probabilities from the equation rather than the model

manual_probs = logistic(11.6630 - (0.2162) * temperature_list)
manual_probs
```

```
## [1] 0.65402974 0.55092283 0.44324565
```

The model-estimated probabilities for several additional ambient temperatures are provided below, where subscripts indicate the temperature:

$$\hat{p}_{57} = 0.341$$
$$\hat{p}_{59} = 0.251$$
$$\hat{p}_{61} = 0.179$$
$$\hat{p}_{63} = 0.124$$
$$\hat{p}_{65} = 0.084$$
$$\hat{p}_{67} = 0.056$$
$$\hat{p}_{69} = 0.037$$
$$\hat{p}_{71} = 0.024$$

Repeat the above, using the entire range of applicable temperatures:

```
# make a temperature list of the three temperatures requested
temperature_list = seq(51,81,2)
temperature_list
```

```
##  [1] 51 53 55 57 59 61 63 65 67 69 71 73 75 77 79 81
```

```
# predict the logit for each of the three temperatures
logit_preds = predict.glm(fail_model,newdata=data.frame(temp=temperature_list))
names(logit_preds) = temperature_list
logit_preds
```

```
##          51          53          55          57          59          61          63          65
##  0.63507282  0.20260550 -0.22986183 -0.66232916 -1.09479649 -1.52726381 -1.95973114 -2.39219847
##          67          69          71          73          75          77          79          81
## -2.82466579 -3.25713312 -3.68960045 -4.12206778 -4.55453510 -4.98700243 -5.41946976 -5.85193709
```

```r
# The desired probabilities for the three temperatures
prob_preds = logistic(logit_preds)

print("The desired probabilities are:")
```

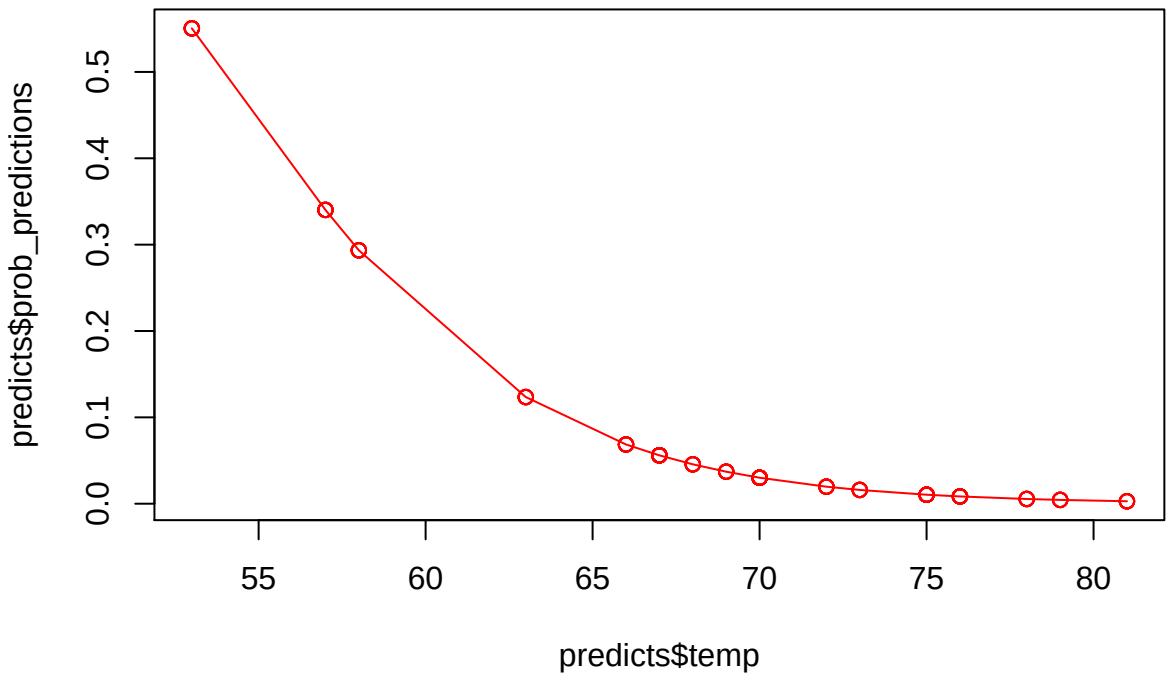
```
## [1] "The desired probabilities are:"
```

```r
print(t(t(prob_preds)))
```

```
##            [,1]
## 51 0.6536388178
## 53 0.5504788165
## 55 0.4427862349
## 57 0.3402165925
## 59 0.2507161453
## 61 0.1783943753
## 63 0.1234961473
## 65 0.0837695402
## 67 0.0560057456
## 69 0.0370714129
## 71 0.0243730934
## 73 0.0159523561
## 75 0.0104098839
## 77 0.0067798159
## 79 0.0044099608
## 81 0.0028660879
```

plot the predicted values from the model

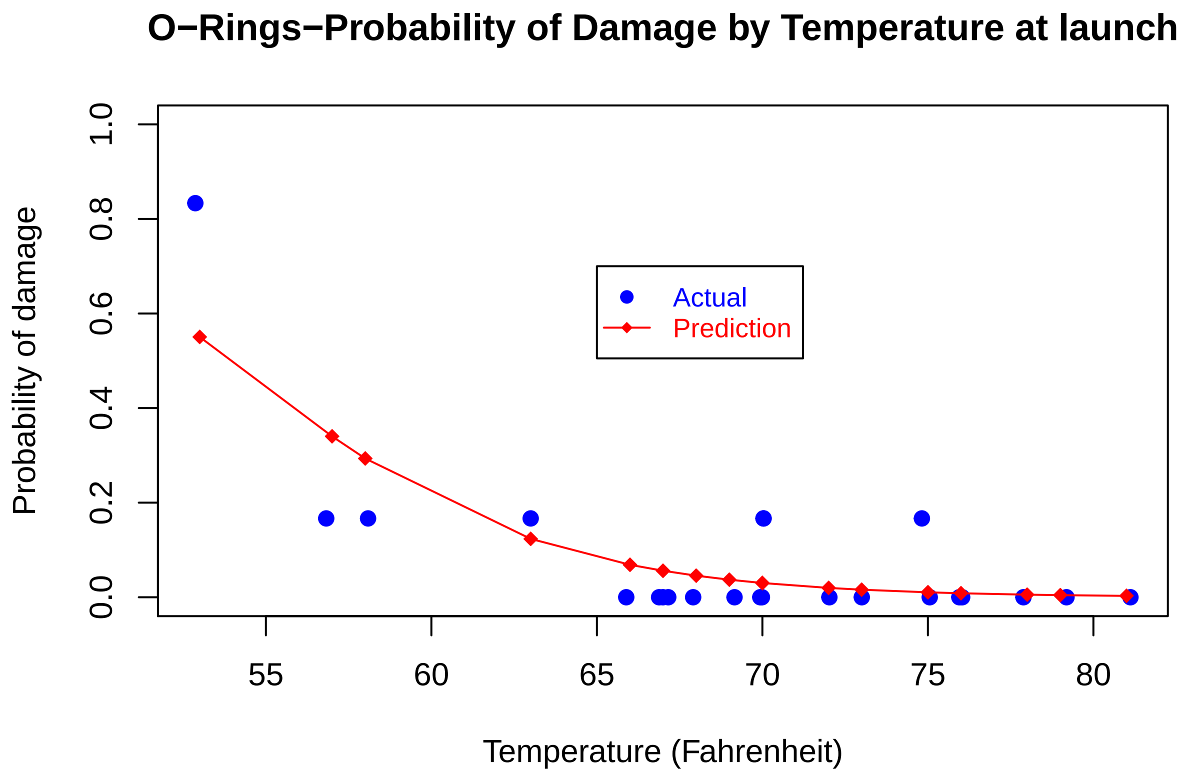
```r
logit_predictions=t(t(predict(fail_model)))
prob_predictions=logistic(logit_predictions)
predicts = cbind(failsbytemp,prob_predictions)
plot(predicts$temp,predicts$prob_predictions,col="red")
lines(predicts$temp,predicts$prob_predictions,type="l",col="red")
```



(b) Add the model-estimated probabilities from part (a) on the plot, then connect these dots using a smooth curve to represent the model-estimated probabilities.

Here's a plot incorporating both the probability points (from top, in blue) as well as the predictions (in red) on a single graph:

```
#plot.new()
plot(orings$failure~jitter(orings$temp),  col="blue",
     xlab="Temperature (Fahrenheit)",
     ylab="Probability of damage",ylim=c(0,1),
     main="O-Rings-Probability of Damage by Temperature at launch",
     pch=19
)
points(predicts$temp,predicts$prob_predictions,col="red",pch=18)
lines(predicts$temp,predicts$prob_predictions,col="red",lty=1)
legend(x=65, y=0.7, legend=c("Actual", "Prediction"),
       cex=0.8,
       pch = c(19, 18), lty = c(NA, 1),
       col = c("blue", "red"), text.col = c("blue", "red"))
```



(c) Describe any concerns you may have regarding applying logistic regression in this application, and note any assumptions that are required to accept the model's validity.

The use of Logistic Regression does seem reasonable to illustrate the relation of damage to O-rings vs. the temperature at shuttle launch.

Here it was necessary to expand the 23-row table of launches, with damaged and undamaged O-ring counts, into a table of 6*23 = 128 rows, indicating the damage or non-damage to each O-ring. The assumption of independence of the observations is thus compromised, as each group of 6 O-rings were on an individual shuttle launch. Thus it is to be expected that the results on each O-ring within a group would be similar.

Of course, there may be other reasons, apart from temperature, which could cause damage to an O-ring, but those are not considered here. The upshot is that a shuttle which suffered damage to just a single O-ring was not impacted in the same way as a shuttle which lost 5 O-rings, which resulted in the tragic loss of Challenger in 1986.

#.#