Multiple linear regression homework

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.

(a) Write the equation of the regression line.

$$\begin{aligned} sc\hat{o}re &= \hat{\beta}_0 + \hat{\beta}_1 \times parity \\ &= 120.7 - 1.93 \times parity \end{aligned}$$

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

The slope (120.07) indicates the first born (parity = 0) would be predicted to weigh 120.07 ounces. The others born, based on the slope of -1.93 would be (r 120.07 - 1.93) oz.

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

Given the p-value of 0.1052 for the parity parameter, I conclude there is not a statistically significant relationship between average birth weight and parity.

8.4 Absenteeism.

Researchers interested in the relationship between absenteeism from school and certain demographic characteristics of children collected data from 146 randomly sampled stu- dents in rural New South Wales, Australia, in a particular school year. Below are three observations from this data set.

(a) Write the equation of the regression line.

[y = 18.93 - 9.11 x {eth} + 3.10 x {sex} + 2.15 x {lrn}]

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

The slope of eth indicates that, all else being equal, there is a 9.11 day reduction in the predicted absenteeism when the subject is no aboriginal.

The slope of sex indicates that, all else being equal, there is a 3.10 day increase in the predicted absenteeism when the subject is male.

The slope of lrn indicates that, all else being equal, there is a 2.15 day increase in the predicted absenteeism when the subject is a slow learner.

(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.

```
eth <- 0
sex <- 1
lrn <- 1
actualDaysMissed <- 2

aDP <- 18.93 - 9.11 * eth + 3.1 * sex + 2.15 * lrn
aDP

## [1] 24.18

residual <- actualDaysMissed - aDP
residual

## [1] -22.18
```

(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 R2 and the adjusted R2. Note that there are 146 observations in the data set.

```
n <- 146
k <- 3
vRes <- 240.57
vOut <- 264.17

R2 <- 1 - (vRes / vOut)
R2

## [1] 0.08933641

adjR2 <- 1 - (1 - R2) * ( (n-1) / (n-k-1) )
adjR2

## [1] 0.07009704
```

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.

Answer: Based on the adjusted (R^2)=0.0723, the learner status variable, lrn, which is high and should be removed first.

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.

(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.

Answer: After 53 degrees, its either 1 or 0. but i don't see a pattern.

(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. Describe the key components of this summary table in words.

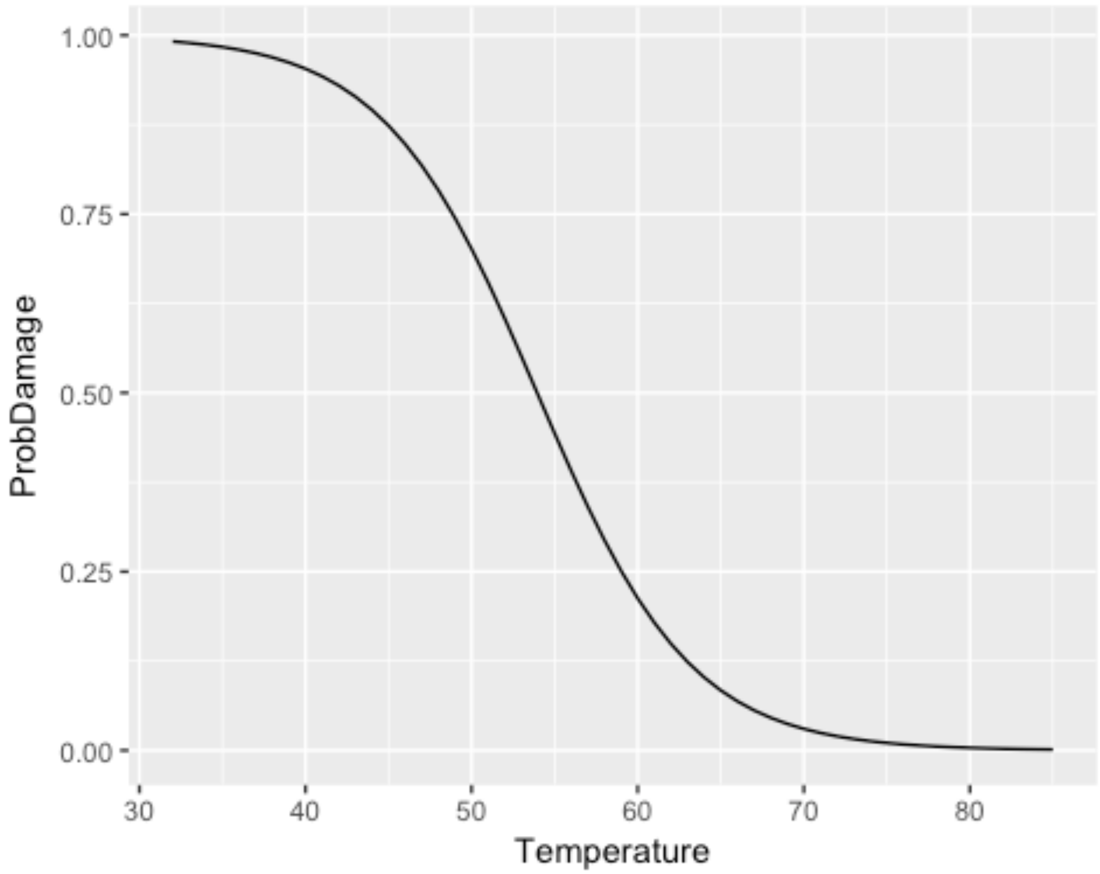
```
dfModel <- data.frame(Estimate=c(11.6630, -0.2162),
                      StdError=c(3.2963, 0.0532),
                      zValue=c(3.54, -4.07),
                      PrAbsZ=c(0.0004, 0.0000))
rownames(dfModel) <- c('Intercept', 'Temperature')
dfModel

##             Estimate StdError zValue PrAbsZ
## Intercept    11.6630   3.2963   3.54  4e-04
## Temperature  -0.2162   0.0532  -4.07  0e+00
```

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

[*e() = 11.6630 - 0.2162 x*{temp}]

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

```
library(ggplot2)

origModel <- function(temp)
{
  right <- 11.6630 - 0.2162 * temp

  prob <- exp(right) / (1 + exp(right))

  return (prob)
}

temps <- seq(32, 85)
dfProbDamage <- data.frame(Temperature=temps,
ProbDamage=origModel(temps))


g1 <- ggplot(dfProbDamage) + geom_line(aes(x=Temperature, y=ProbDamage
))
g1
```



Based on the above graph, high damage has happened below 50 degree. I do think concerns regarding the O-rings are justified.

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.

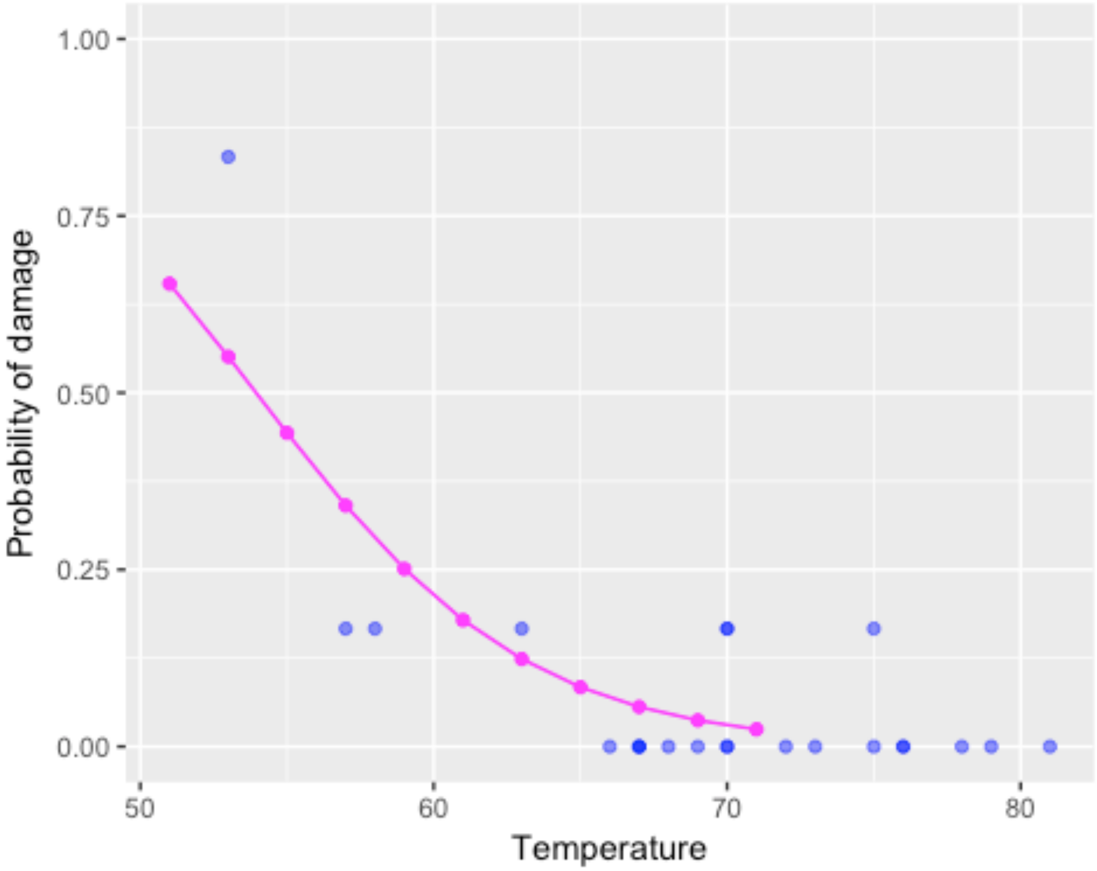
(a) Use the model to calculate the probability that an O-ring will become damaged at each of the following ambient temperatures: 51, 53, and 55 degrees Fahrenheit. The model-estimated probabilities for several additional ambient temperatures are provided below, where subscripts indicate the temperature: pˆ57 = 0.341 pˆ59 = 0.251 pˆ61 = 0.179 pˆ63 = 0.124 pˆ65 = 0.084 pˆ67 = 0.056 pˆ69 = 0.037 pˆ71 = 0.024

```
temps <- c(51,53,55)

dPD <- data.frame(Temperature=temps, ProbDamage=origModel(temps))
dPD

##   Temperature ProbDamage
## 1          51  0.6540297
## 2          53  0.5509228
## 3          55  0.4432456
```

(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.

```
# raw data
dfRaw <- data.frame(Missing=seq(1, 23),
                    Temp=c(53,57,58,63,66,67,67,67,68,69,70,70,70,
                           70,72,73,75,75,76,76,78,79,81),

Damaged=c(5,1,1,1,0,0,0,0,0,0,1,0,1,0,0,0,0,1,0,0,0,0,0),

Undamaged=c(1,5,5,5,6,6,6,6,6,6,5,6,5,6,6,6,6,5,6,6,6,6,6))
dfRaw$ProbDamage <- dfRaw$Damaged / (dfRaw$Damaged + dfRaw$Undamaged)
head(dfRaw)

##   Missing Temp Damaged Undamaged ProbDamage
## 1       1   53       5         1  0.8333333
## 2       2   57       1         5  0.1666667
## 3       3   58       1         5  0.1666667
## 4       4   63       1         5  0.1666667
## 5       5   66       0         6  0.0000000
## 6       6   67       0         6  0.0000000

#model

temps <- seq(51, 71, by=2)
dfProbDamage <- data.frame(Temperature=temps,
ProbDamage=origModel(temps))

#visualization combining the raw data and the model curve
```

```
g1 <- ggplot(dfRaw) +
  geom_point(aes(x=Temp, y=ProbDamage), alpha=0.5, colour="blue") +
  geom_line(data=dfProbDamage, aes(x=Temperature, y=ProbDamage),
colour="magenta") +
  geom_point(data=dfProbDamage, aes(x=Temperature, y=ProbDamage),
colour="magenta") +
  labs(x="Temperature", y="Probability of damage") +
  ylim(0, 1)
g1
```



(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.

Based on the visualization above, the temperature data does appear to have a linear relationship with the probability of damage to O-rings, at least to some significant degree.

Conditions/assumptions required for logistic regression model validity include:

Each predictor (x), is linearly related to logit((p_i)) if all other predictors are help constant. Each outcome (Y_i) is independent of the other outcomes.