HW8

Ahsanul Choudhury

November 28, 2016

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.

$$\hat{y} = 120.07 - 1.93 \times parity$$

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

The estimated birth weight of babies not first borns is 1.93 ounces less than baby first borns weight of 120.07.

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

The p-value = 0.1052, which is greater than 0.05 so we can conclude that we cannot reject the null hypothesis, there is no statistically significant relationship between birthweight 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 students in rural New South Wales, Australia, in a particular school year. Below are three observations from this data set.

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: 0 - aboriginal, 1 - not aboriginal), sex (sex: 0 - female, 1 - male), and learner status (lrn: 0 - average learner, 1 - slow learner).

(a) Write the equation of the regression line.

$$\hat{y} = 18.93 - 9.11 \times eth + 3.10 \times sex + 2.15 \times lrn$$

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

- ***Slope for variable `eth` represents the average number of absences that would be reduced when the subject are not aborigenes.***
- ***Slope for variable `sex` represents the increase in average number of absences when the subject is male.***
- ***Slope for variable `lrn` represents the increase in average number of absences when the suject is identified as 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

pre_ab_days <- 18.93 - 9.11*eth + 3.1*sex + 2.15*lrn
missed_days <- 2

residual <- missed_days - pre_ab_days

residual
```

```
## [1] -22.18
```

residual = -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
res_var <- 240.57
ab_var <- 264.17

R2 <- 1 - (res_var / ab_var)
R2
```

```
## [1] 0.08933641
```

```
adjustedR2 <- 1 - (1 - R2) * ( (n-1) / (n-k-1) )
adjustedR2
```

```
## [1] 0.07009704
```

R2 = 0.0893364, adjusted R2 = 0.070097.

8.8 Absenteeism, Part II. Exercise 8.4 considers a model that predicts the number of days bsent 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.

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

Based on the adjusted R2 data, variable `lrn` should be removed from the model 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.

From the data it looks like more O-ring damage happens at lower temparature then higher temparature

(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.

The Estimate identifies the parameter estimate for the model. The z value and the p-value help us identify significant parameters and less significant parameters.

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

$$log(\frac{p_i}{1-p_i}) = 11.6630 - 0.2162 \times Temperature$$

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

Based on the model I think concerns regarding o-rings are justified. We have a low p-value to back it up.

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

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

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

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

model <- function(temp)
{
  form <- 11.6630 - 0.2162 * temp

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

  return (prob)
}

df <- data.frame(temperature=temps, prob_of_damage=model(temps))
knitr::kable(df)
```

temperature	prob_of_damage
51	0.6540297
53	0.5509228
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.

```
temperature <- 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)

raw_data <- data.frame(temperature = temperature, damaged = damaged,
                       undamaged = undamaged)
raw_data$prob_of_damage <- model(temperature)
raw_data
```

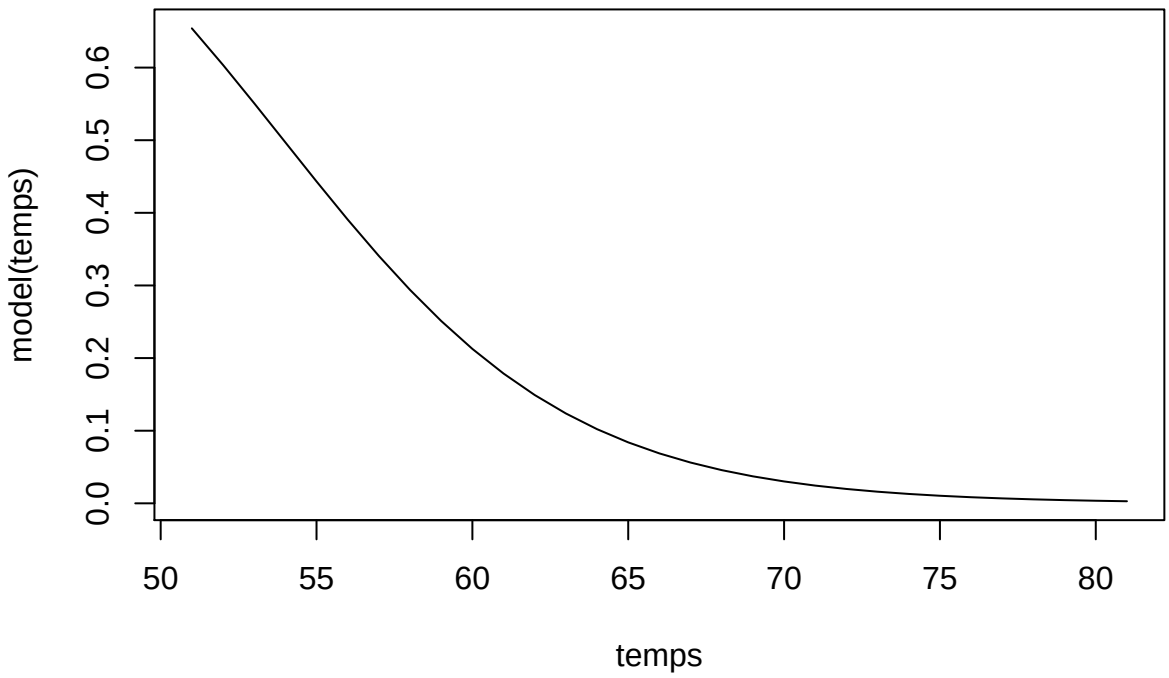
```
##    temperature damaged undamaged prob_of_damage
## 1           53       5         1    0.550922830
## 2           57       1         5    0.340649763
## 3           58       1         5    0.293882838
## 4           63       1         5    0.123727017
## 5           66       0         6    0.068740463
## 6           67       0         6    0.056125657
## 7           67       0         6    0.056125657
## 8           67       0         6    0.056125657
## 9           68       0         6    0.045712204
## 10          69       0         6    0.037154789
## 11          70       1         5    0.030148728
## 12          70       0         6    0.030148728
## 13          70       1         5    0.030148728
## 14          70       0         6    0.030148728
## 15          72       0         6    0.019774295
## 16          73       0         6    0.015991141
## 17          75       0         6    0.010436032
## 18          75       1         5    0.010436032
## 19          76       0         6    0.008424090
## 20          76       0         6    0.008424090
## 21          78       0         6    0.005483026
## 22          79       0         6    0.004421698
## 23          81       0         6    0.002873921
```

```
temps <- seq(51, 81)
df_prob_damage <- data.frame(Temperature=temps, ProbDamage=model(temps))
df_prob_damage
```

```
##    Temperature  ProbDamage
## 1           51 0.654029738
## 2           52 0.603626816
```

```
## 3           53 0.550922830
## 4           54 0.497050034
## 5           55 0.443245647
## 6           56 0.390740650
## 7           57 0.340649763
## 8           58 0.293882838
## 9           59 0.251091387
## 10          60 0.212654228
## 11          61 0.178697069
## 12          62 0.149135196
## 13          63 0.123727017
## 14          64 0.102128054
## 15          65 0.083938432
## 16          66 0.068740463
## 17          67 0.056125657
## 18          68 0.045712204
## 19          69 0.037154789
## 20          70 0.030148728
## 21          71 0.024430239
## 22          72 0.019774295
## 23          73 0.015991141
## 24          74 0.012922227
## 25          75 0.010436032
## 26          76 0.008424090
## 27          77 0.006797363
## 28          78 0.005483026
## 29          79 0.004421698
## 30          80 0.003565071
## 31          81 0.002873921
```

```
plot(temps, model(temps), type = "l")
```



(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.

Our model depends largely on data at 53 degree fahrenheit where 5 out of 6 O-rings were damaged, if this outlier is because of another factor then our model's validity will be seriously compromised, fot the validity of the model we are assuming there are no other variables.