Homework 8

CUNY MSDS DATA 606

Duubar Villalobos Jimenez mydvtech@gmail.com

March 7, 2017

Chapter 8 - Multiple and Logistic Regression

```
Practice: 8.1, 8.3, 8.7, 8.15, 8.17
Graded: 8.2, 8.4, 8.8, 8.16, 8.18
```

8.2 Baby weights

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.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	120.07	0.60	199.94	0.0000
parity	-1.93	1.19	-1.62	0.1052

Figure 1:

Answer:

(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 slope indicates that for each one parsity increase, there will be a loss of 1.93 ounces in the baby's weight.

Since the parity varies, below are the two functions for parity.

```
weight <- function(parity){
    yhat <- 120.07 - 1.93 * parity
    return(yhat)
}
```

- If baby is first born: `parity` = 0; the baby's weight will be 120.07 ounces.
- If baby is **NOT** first born: `parity` = 1; the baby's weight will be 118.14 ounces.

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

Since the parity parameter's $p - value = 0.1052$; we can conclude that there is NOT a statistically significant relationship between average birth weight and parity.

8.4 Absenteeism, Part I.

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.

	eth	sex	lrn	days
1	0	1	1	2
2	0	1	1	11
⋮	⋮	⋮	⋮	⋮
146	1	0	0	37

Figure 2:

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).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	18.93	2.57	7.37	0.0000
eth	-9.11	2.60	-3.51	0.0000
sex	3.10	2.64	1.18	0.2411
lrn	2.15	2.65	0.81	0.4177

Figure 3:

Answer:

(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.

- 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 # Aboriginal
sex <- 1 # Male
lrn <- 1 # Slow Learner
missedActualDays <- 2
```

```
predictedDays <- 18.93 - 9.11 * eth + 3.1 * sex + 2.15 * lrn

residual <- missedActualDays - predictedDays
```

The `residual` for the above supplied information will be: -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 R^2 and the adjusted R^2 . Note that there are 146 observations in the data set.

```
n <- 146 # Number os cases to fit the model
k <- 3   # number of predictor variables in the model
varResidual <- 240.57 # Variance of residual
varAllStudents <- 264.17      # Variance for all students

R2 <- 1 - (varResidual / varAllStudents)  # R2

adjustedR2 <- 1 - (varResidual / varAllStudents) * ( (n-1) / (n-k-1) ) # Adjusted R2
```

$$R^2 = 0.0893364$$

$$R^2_{Adjusted} = 0.070097$$

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.

	Model	Adjusted R^2
1	Full model	0.0701
2	No ethnicity	-0.0033
3	No sex	0.0676
4	No learner status	0.0723

Figure 4:

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

Answer:

Based on the Adjusted $R^2 = 0.0723$ the `lrn` variable 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.

Shuttle Mission	1	2	3	4	5	6	7	8	9	10	11	12
Temperature	53	57	58	63	66	67	67	67	68	69	70	70
Damaged	5	1	1	1	0	0	0	0	0	0	1	0
Undamaged	1	5	5	5	6	6	6	6	6	6	5	6

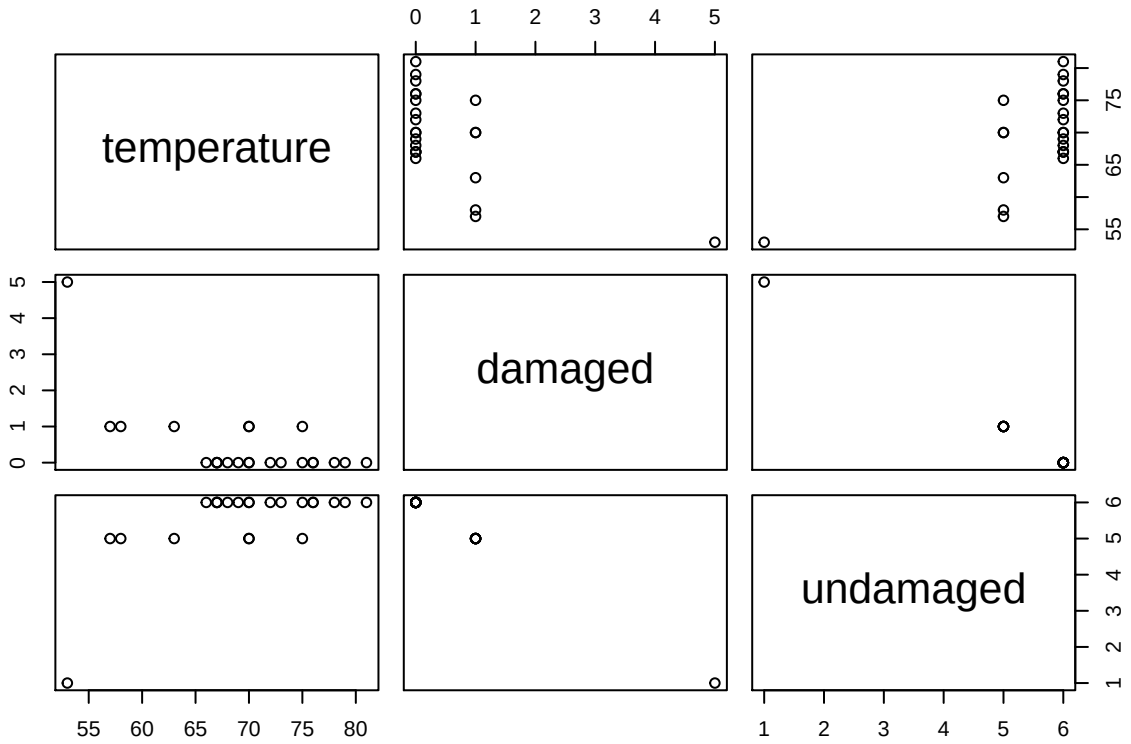
Shuttle Mission	13	14	15	16	17	18	19	20	21	22	23
Temperature	70	70	72	73	75	75	76	76	78	79	81
Damaged	1	0	0	0	0	1	0	0	0	0	0
Undamaged	5	6	6	6	6	5	6	6	6	6	6

Figure 5:

Answer:

(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.

```
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)

ShuttleMission <- data.frame(temperature, damaged, undamaged)

plot(ShuttleMission)
```



By observing the above plot, we can find an interesting observation as follows:

- Higher number of damaged O-rings are observed when lower temperatures were recorded.
- Less number of damaged O-rings are observed when higher temperatures were recorded.

(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.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	11.6630	3.2963	3.54	0.0004
Temperature	-0.2162	0.0532	-4.07	0.0000

Figure 6:

This model is represented by two components: One, is the Intercept and the second one is the `Temperature` values. The Estimate identifies the parameter estimate for the model. The **Z** value and the **P-value** help with distinguishing important information from less significant parameters by telling us how good the variables predict this model.

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

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

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

Based on the collected data, we can deduct a high probability of damage to O-rings under 50^o. Also, since O-rings are *"Critical"* components, 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.

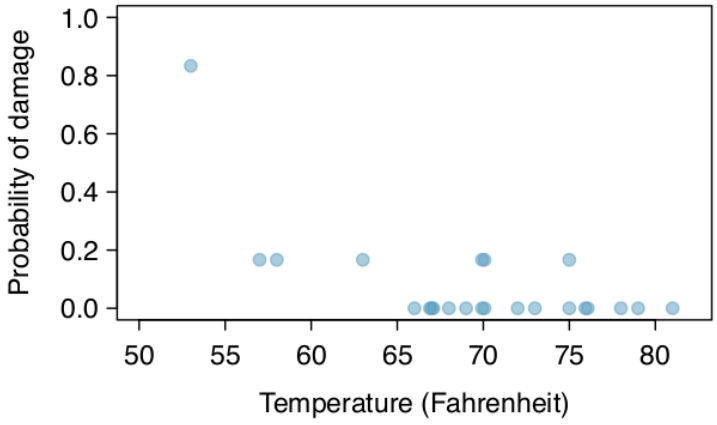


Figure 7:

Answer:

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

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

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

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

Figure 8:

By solving the equation

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

in terms of p, we obtain

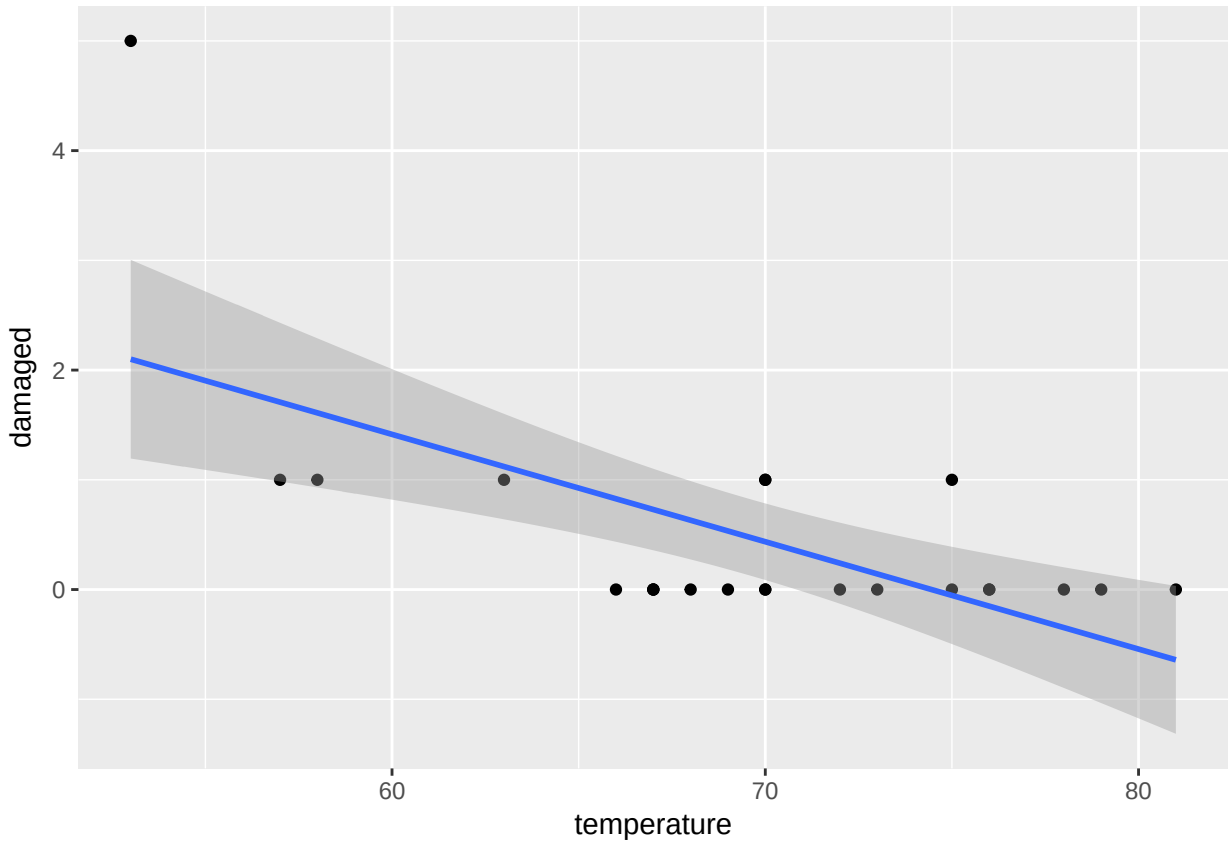
$$\hat{p} = \frac{e^{11.6630 - 0.2162 \times Temperature}}{1 + e^{11.6630 - 0.2162 \times Temperature}}$$

```
p <- function(temp)
{
  damagedOring <- 11.6630 - 0.2162 * temp

  phat <- exp(damagedOring) / (1 + exp(damagedOring))

  return (round(phat*100,2))
}

# Testing final formula to provided results.
p57 <- p(57)
p59 <- p(59)
p61 <- p(61)

# Finding probabilities for 51, 53, and 55 temperatures.
p51 <- p(51)
p53 <- p(53)
p55 <- p(55)
```

- The probability that an O-ring will become damaged at 51 degrees Fahrenheit ambient temperatures is: 65.4%.
- The probability that an O-ring will become damaged at 53 degrees Fahrenheit ambient temperatures is: 55.09%.
- The probability that an O-ring will become damaged at 55 degrees Fahrenheit ambient temperatures is: 44.32%.

(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.

```
ggplot(ShuttleMission,aes(x=temperature,y=damaged)) + geom_point() +
  stat_smooth(method = 'glm', family = 'binomial')
```



(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.)

Logistic regression conditions:

There are two key conditions for fitting a logistic regression model:

1. Each predictor x_i is linearly related to logit(p_i) if all other predictors are held constant.
2. Each outcome Y_i is independent of the other outcomes.

Based on that definition, we have assumed that those conditions are met.