Chapter 8 HW

Kishore Prasad

```
library(ggplot2)
```

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`.

	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

(a) Write the equation of the regression line.

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

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

a) The equation for baby weight is:

$\hat{Baby weight}$ = 120.07 - 1.93 × parity

b) The slope (120.07) means that a first born (with a parity of 0) would be predicted to weigh 120.07 (120.07 - 1.93 × 0) ounces.

The others (with a parity of 1) would be predicted to weigh 118.14 (120.07 - 1.93 × 1) ounces

c) A p-value of 0.1052 for parity indicates that there is no 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

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).[18]

	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

(a) Write the equation of the regression line.

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

(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.

(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.

a) The equation for absenteeism is given by:

$\hat{absenteeism}$ = 18.93 - 9.11 × eth + 3.10 × sex + 2.15 × lrn

b) The following are the interpretation of slope for each variable:

- All else being equal, there is a decrease of 9.11 days in predicted absenteeism when the eth value is "No" (not aboriginal) in an observation.
- All else being equal, there is an increase of 3.1 days in predicted absenteeism when the sex is male in an observation.
- All else being equal, there is a decrease of 2.15 days in predicted absenteeism when the lrn value is "slow learner" in an observation.

c) We first need to calculate the predicted days missed as below:

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

```
actual <- 2

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

```
## [1] 24.18
```

```
residual <- actual - predicted
residual
```

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

Therefore the residual is -22.18

d) The following is the calculation:

```
n <- 146
k <- 3
var_residuals <- 240.57
var_absentdays <- 264.17

R2 <- 1 - (var_residuals / var_absentdays)
R2
```

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

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

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

Therefore, $R^2 = 0.0893364$ and adjusted $R^2 = 0.070097$

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

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

a) The variable with the highest adjusted R^2 is lrn (0.0723). Hence this variable 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.

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

(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.

(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

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

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

a) Based on a visual inspection of the table, I can find that damages to O-ring is more pronounced for temperatures less than 66 farenheit. There seem to be less damage when the temperature is equal to or above this.

b) The key components are described as follows:

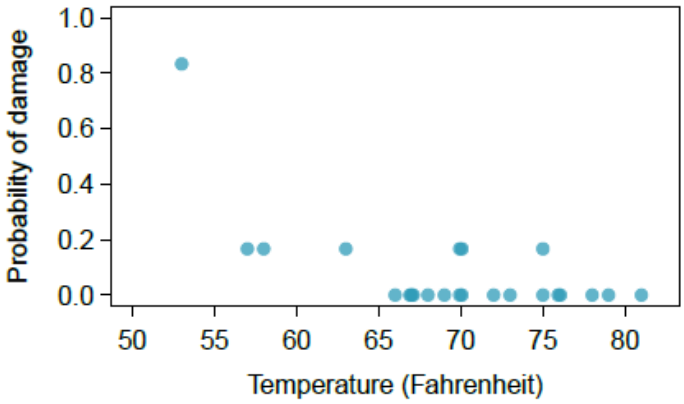
- Intercept - The intercept estimate of 11.6630 indicates what the failure value will be if the temperature was 0.
- Slope - The slope estimate of -0.2162 indicates that the failure decreases by 0.2162 as the temperature increases by 1 degree.
- The z value and P value help in identifying the strength of the relationship

c) The following is the equation for the logistic regression:

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

d) I would say that the concerns regarding the O-rings are justified. The p-value close to zero.It seems to indicate that O-ring failure is strongly correlated to temperature.

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

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

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:

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

(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.

(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.

a) The following are the calculations for 51, 53 and 55 degrees:

```
temp = 51
p51 <- exp(11.6630 - (0.2162 * temp)) / (1 + exp(11.6630 - (0.2162 * temp)))
p51
```

```
## [1] 0.6540297
```

```
temp = 53
p53 <- exp(11.6630 - (0.2162 * temp)) / (1 + exp(11.6630 - (0.2162 * temp)))
p53
```

```
## [1] 0.5509228
```

```
temp = 55
p55 <- exp(11.6630 - (0.2162 * temp)) / (1 + exp(11.6630 - (0.2162 * temp)))
p55
```

```
## [1] 0.4432456
```

Therefore the probabilities for 51, 53 and 55 are 0.6540297, 0.5509228 and 0.4432456 respectively.

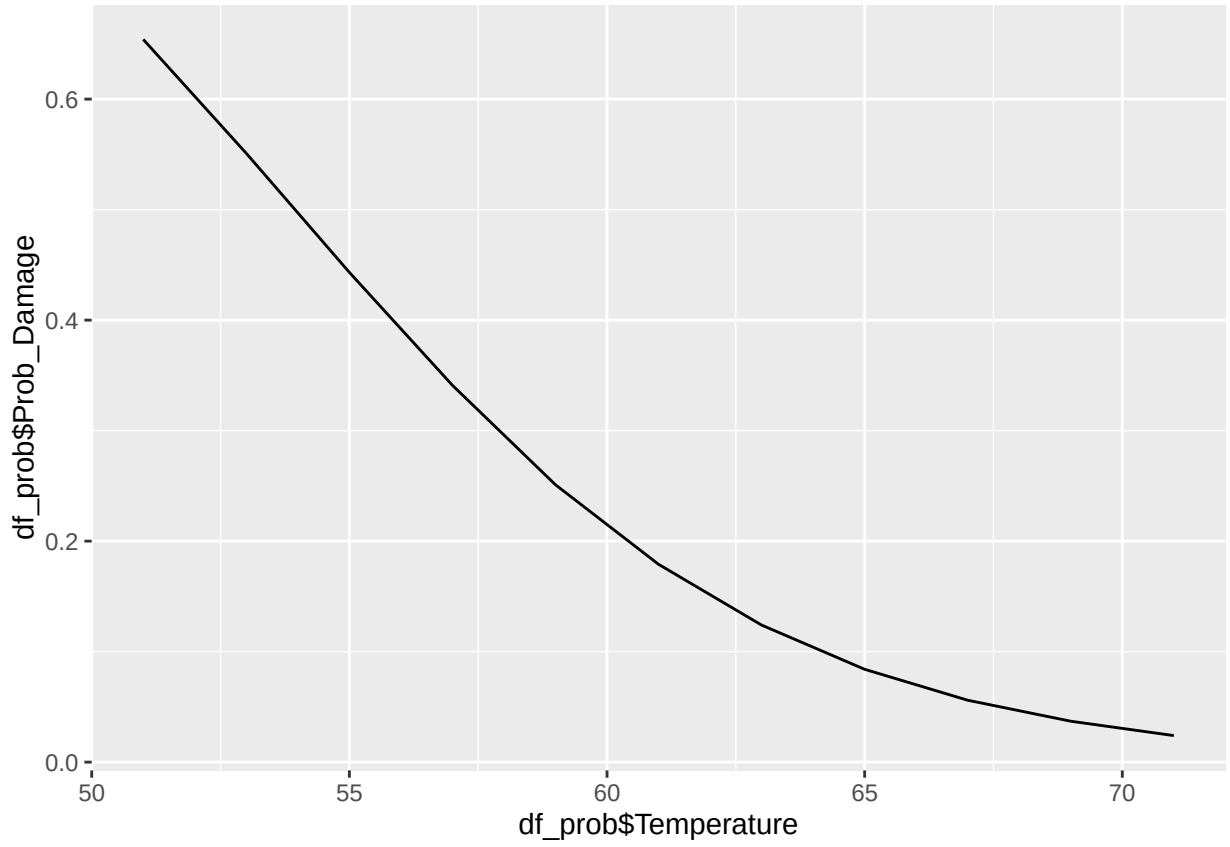
b) Lets first create a data frame based on the values provided:

```
temps <- seq(51, 71, by=2)
probs <- c(p51, p53, p55, 0.341, 0.251, 0.179, 0.124, 0.084, 0.056, 0.037, 0.024)

df_prob <- data.frame(Temperature=temps, Prob_Damage=probs)
```

We now plot the values as a smooth curve:

```
g1 <- ggplot(df_prob) + geom_line(aes(x=df_prob$Temperature, y=df_prob$Prob_Damage ))
g1
```



c) I would say that we do not have sufficient data to learn from a logistic model. The following are the conditions / assumptions for a logistic model:

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. Based on the visualization above, temperature does seem to have a linear relationship to the probability of damage.

Each outcome (Y_i) is independent of the other outcomes. I am assuming that the O-rings are completely replaced for each mission. Given this assumption, I would say that the independence condition is met.