Chapter 9 - Multiple and Logistic Regression

Baby weights, Part I. (9.1, p. 350) The Child Health and Development Studies investigate a range of topics. One study considered all pregnancies between 1960 and 1967 among women in the Kaiser Foundation Health Plan in the San Francisco East Bay area. Here, we study the relationship between smoking and weight of the baby. The variable *smoke* is coded 1 if the mother is a smoker, and 0 if not. The summary table below shows the results of a linear regression model for predicting the average birth weight of babies, measured in ounces, based on the smoking status of the mother.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	123.05	0.65	189.60	0.0000
smoke	-8.94	1.03	-8.65	0.0000

The variability within the smokers and non-smokers are about equal and the distributions are symmetric. With these conditions satisfied, it is reasonable to apply the model. (Note that we don't need to check linearity since the predictor has only two levels.)

(a) Write the equation of the regression line.

An equation for this regression line would be:

$\hat{weight} = 123.05 - 8.94 \times smoke$

Given that the table above in the problem has calculated the slope and y-intercept, we were able to plug these into an equation.

(b) Interpret the slope in this context, and calculate the predicted birth weight of babies born to smoker and non-smoker mothers.

Since we know that the variable smoke is coded 1 if a mother is a smoker, and a 0 if not, we can utilize this knowledge to plug this into our equation. From the equation above, the slope of the model predicts a baby will be 8.94 ounces lighter for an average birth weight for mothers that are smokers.

The calculation for the predicted birth weight of babies born to a smoker:

$babywei\hat{g}ht_{smoker} = 123.05 - (8.94 \times 1) = 114.11 ounces$

The calculation for the predicted birth weight of babies born to a non-smoker:

$babyweigh\hat{t}_{nonsmoker} = 123.05 - (8.94 \times 0) = 123.05 ounces$

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

After looking at the table above, we can see that there is a statistically significant relationship between the average birth weight and smoking. with a p-value less than 0.001, this shows that the slope of β_1 is not zero.

Absenteeism, Part I. (9.4, p. 352) 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).

	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.

An equation of the regression line based on the table above would look like:

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

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

The slopes of the regression line equation can be interpreted based on the following:

- **With all other variables being constant, students that have an ethnicity of aboriginal, on average, miss 9.11 more days than students that do not have an ethnicity of aboriginal.**
- **With all other variables being constant, male students, on average miss 3.10 more days than female students.**
- **With all other variables being constant, students identified as slow learners, on average, miss 2.15 more days than average learners.**

(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.

To calculate the residual, we can utlize the equation we build in part (a):

```r
observed <- 2
eth <- 0
sex <- 1
lrn <- 1
intercept <- 18.93

daysAbsent <- intercept - (9.11 * eth) + (3.10 * sex) + (2.15 * lrn)

paste0('The prediction is ', daysAbsent, ' days absent.')
```

```
## [1] "The prediction is 24.18 days absent."
```

```
paste0('The observed value is ', observed, ' days absent.')
```

```
## [1] "The observed value is 2 days absent."
```

```
paste0('The residual for the first observation is ', observed - daysAbsent, '.')
```

```
## [1] "The residual for the first observation is -22.18."
```

We can see that the residual for the first observation is -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.

We can calculate the R^2 value by utilizing the following equation:

$R^2 = 1 - \frac{Var(e_i)}{Var(y_i)}$

```
var_res <- 240.57
var_outcome <- 264.17
R_squared <- 1 - (var_res/var_outcome)

paste0('The R-squared value is ', format(R_squared, digits = 5), '.')
```

```
## [1] "The R-squared value is 0.089336."
```

Now, we can calculate the *adjusted* R^2 value by utilizing the following equation:

$R^2_{adj} = 1 - \frac{s^2_{residuals}/(n-k-1)}{s^2_{outcome}/(n-1)} = \frac{s^2_{residuals}}{s^2_{outcome}} \times \frac{n-1}{n-k-1}$

```
n <- 146
k <- 3 # number of predictors

R_squared_adj <- 1 - (var_res/var_outcome) * ((n - 1)/(n-k-1))
paste0('The R-squared adjusted value is ', format(R_squared_adj, digits = 5), '.')
```

```
## [1] "The R-squared adjusted value is 0.070097."
```

Absenteeism, Part II. (9.8, p. 357) Exercise above 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?

From the table, it looks like the largest adjusted R^2 value is generated when the learner status variable is removed from the full model ($R^2_{adj} = 0.0723$). Therefore, I would argue, without being given any additional information, that the learner status variable ('lrn') should be removed from the model first.

Challenger disaster, Part I. (9.16, p. 380) 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	13	14	15	16	17	18	19	20	21	22	23
Temperature	53	57	58	63	66	67	67	67	68	69	70	70	70	70	72	73	75	75	76	76	78	79	81
Damaged	5	1	1	1	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0
Undamaged	1	5	5	5	6	6	6	6	6	6	5	6	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.

From my observations of the data, it looks like the higher the number of damaged O-rings, the cooler the temperature in Fahrenheit at the time of launch. Conversely, the lower the number of damaged O-rings, the higher the temperature in Fahrenheit at the time of launch.

(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

The logistic regression model was created based on the function:

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

The key components of the summary table are:

- **We can see that the key components of the summary table are the y-intercept, meaning that if the temperature were 0 degrees, the probability of damaged O-rings would be:**

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

which would yield a probability of:

```
prob_0F <- exp(11.663)/(1+exp(11.663))
paste0('The probability of a damaged O-ring at 0F is ', format(prob_0F, digits = 10), '.')
```

```
## [1] "The probability of a damaged O-ring at 0F is 0.9999913936."
```

- **We can also see that the slope is equal to -0.2162, indicating that for every degree F increase, we'll see an adjustment in the probability of a damaged O-ring based on multiplication of the slope by the temperature. In other words, as the temperature increases, the probability of a damaged O-ring decreases.**

```
prob_1F <- exp(11.663-0.2162)/(1+exp(11.663-0.2162))
paste0('The probability of a damaged O-ring at 1F is ', format(prob_1F, digits = 10), '.')
```

```
## [1] "The probability of a damaged O-ring at 1F is 0.9999893165."
```

```
prob_45F <- exp(11.663-0.2162*45)/(1+exp(11.663-0.2162*45))
paste0('The probability of a damaged O-ring at 45F is ', format(prob_45F, digits = 10), '.')
```

```
## [1] "The probability of a damaged O-ring at 45F is 0.8736914987."
```

```
prob_60F <- exp(11.663-0.2162*60)/(1+exp(11.663-0.2162*60))
paste0('The probability of a damaged O-ring at 60F is ', format(prob_60F, digits = 10), '.')
```

```
## [1] "The probability of a damaged O-ring at 60F is 0.2126542284."
```

```
prob_80F <- exp(11.663-0.2162*80)/(1+exp(11.663-0.2162*80))
paste0('The probability of a damaged O-ring at 80F is ', format(prob_80F, digits = 10), '.')
```

```
## [1] "The probability of a damaged O-ring at 80F is 0.003565070533."
```

- **And finally, since the p-value is very small, less than 0.001, we can conclude that the slope of the line is not equal to zero, indicating that there is a relationship between the probability of a damaged O-ring and temperature.**

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

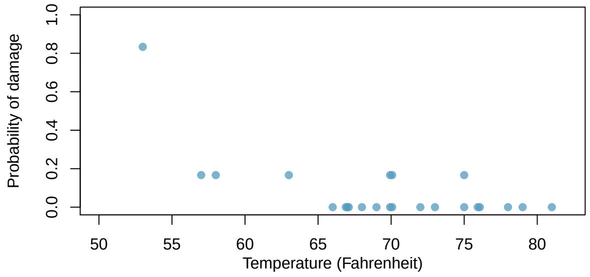
The logistic model using the point estimates of the model parameters can be written as:

$log(\frac{p_i}{1-p_i}) = 11.663 - 0.2162 \times temperature$

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

Yes, based on the model, I do think concerns regarding O-rings and temperature are justified. As we saw from part(b), when we inputted higher temperatures into the model, we had a much lower probability of a damaged O-ring. As an example, at 45 degrees F, the probability of a damaged O-ring was roughly 87%. However, when the temperature inputted was 60 degrees F, the probability of a damaged O-ring dropped drastically to roughly 21%. At 80 degrees F, the probability of a damaged O-ring was a staggering 0.4%.

Challenger disaster, Part II. (9.18, p. 381) Exercise above 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:

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

Similar to the last question, I'll calculate the model-estimated probabilities of a damaged O-ring at 51, 53, and 55 degrees F:

```r
prob_51F <- exp(11.663-0.2162*51)/(1+exp(11.663-0.2162*51))
prob_53F <- exp(11.663-0.2162*53)/(1+exp(11.663-0.2162*53))
prob_55F <- exp(11.663-0.2162*55)/(1+exp(11.663-0.2162*55))
```

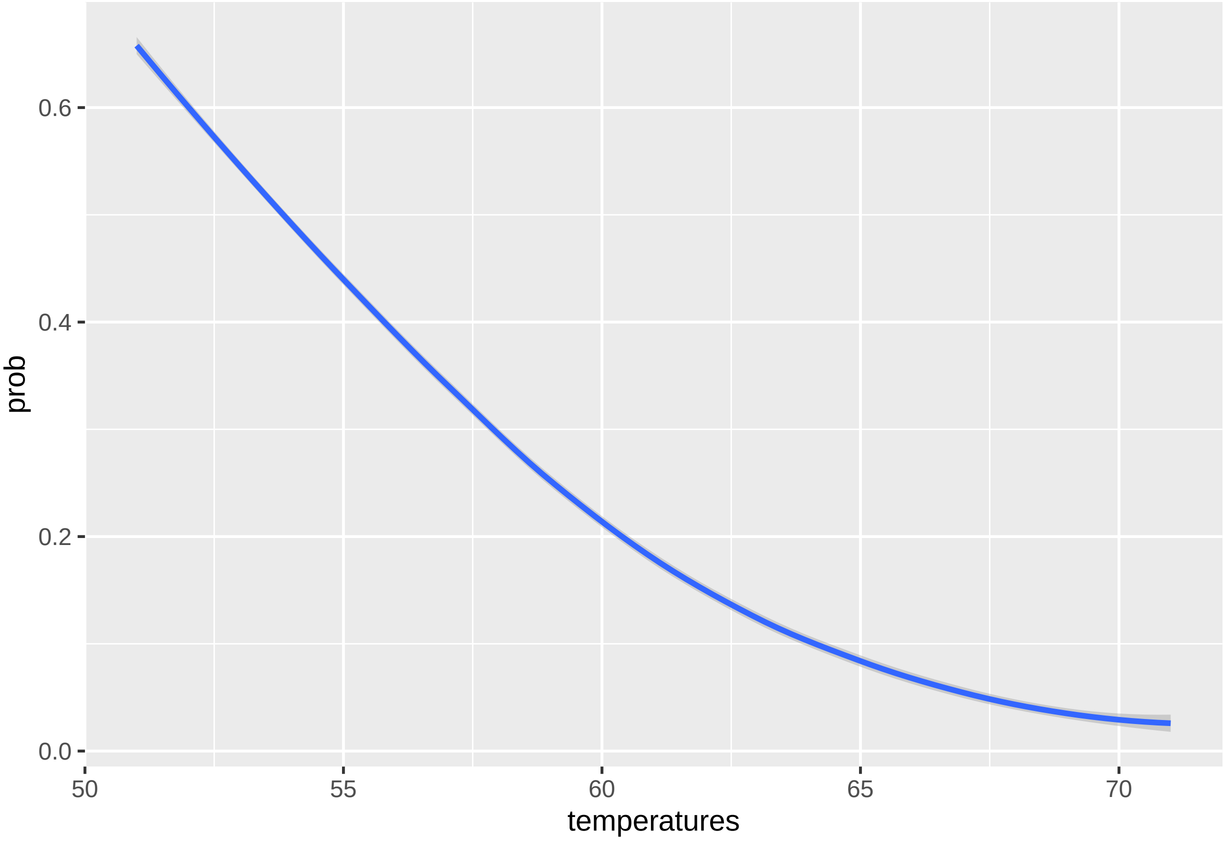
The calculated probabilities are:

- $\hat{p}_{51} = 0.654$
- $\hat{p}_{53} = 0.551$
- $\hat{p}_{55} = 0.443$

(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.

```r
library(ggplot2)
temperatures <- c(51,53,55,57,59,61,63,65,67,69,71)
prob <- exp(11.663-0.2162*temperatures)/(1+exp(11.663-0.2162*temperatures))
plot_df <- data.frame(temperatures, prob)

qplot(temperatures, prob, geom = "smooth")
```



(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 few concerns that I have regarding applying logistic regression to this application, are that:

- **There are many other factors other than temperature that could lead to damage of an O-ring. For instance, we aren't accounting for the age of the O-ring, whether it had been used in past shuttles or space missions, and/or the manufacturer of the O-ring.**
- **Additionally, logistic regression implies that an O-ring will either be damaged or not, but there is a lot of subjectivity in that conclusion. What constitutes an O-ring being damaged? Also, if an O-ring is used multiple times, and all times except for one the O-ring is not damaged, does this impact the factors of the model at all?**