Chapter 9 - Multiple and Logistic Regression

Jose A. Mawyin

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.

average_birth_weight = 123.05 - 8.94 * (smoke)

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

The slope appears to be negative but only modifies the intercept if the mother smokes.

average_birth_weight (smoker_mom) = 123.05 - 8.94 * (1) = 114.11 ounces

average_birth_weight (non_smoker_mom) = 123.05 - 8.94 * (0) = 123.05 ounces

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

Yes, the p-value of (negative) 8.65 indicates a high statistically significant relationship between the average birth weight and smoking.

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.

average_absent_days = 18.93 - 9.11 * eth + 3.10 * sex + 2.15 * lrn

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

The ethnic background slope indicates that being "not aboriginal" decreases the absent days by an average of 9.11 days. Being "female" increases the absent days by an average of 3.10. Being a "slow learner" adds in average 2.15 days to the absent days.

(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.

average_absent_days = 18.93 - 9.11 * eth + 3.10 * sex + 2.15 * lrn

estimated_value = 18.93 -9.11 * 0 + 3.10 * 1 + 2.15 * 1 = 24.18

Residuals = 24.18 - 2 = 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.

For R^2:

$$R^2 = 1 - \frac{\text{SS}_{res}}{\text{SS}_{tot}}$$

```r
var.residual <- 240.57
var.total <- 264.17
R.squared <- 1 - (var.residual)/(var.total)
```

R^2: 0.0893

For the adjusted R^2:

$$R^2_{adjusted} = 1 - \frac{n-1}{n-m}\left(1 - R^2\right)$$

>where n is the number of observations and m is the number of predictors in the model.

```r
n <- 146
m <- 3
R.squared.adjusted <- 1 - ( (n-1) / (n-m)) * (1-R.squared)
```

R^2: 0.0766

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?

> We could remove the "No learner status" as the removal of this variable seems to improve the fit of the linear model as it shows an increase in the R^2 value. The linear regression model without this variable better explains the variability in the collected data.

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
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.

The low temperatures increases the number of 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

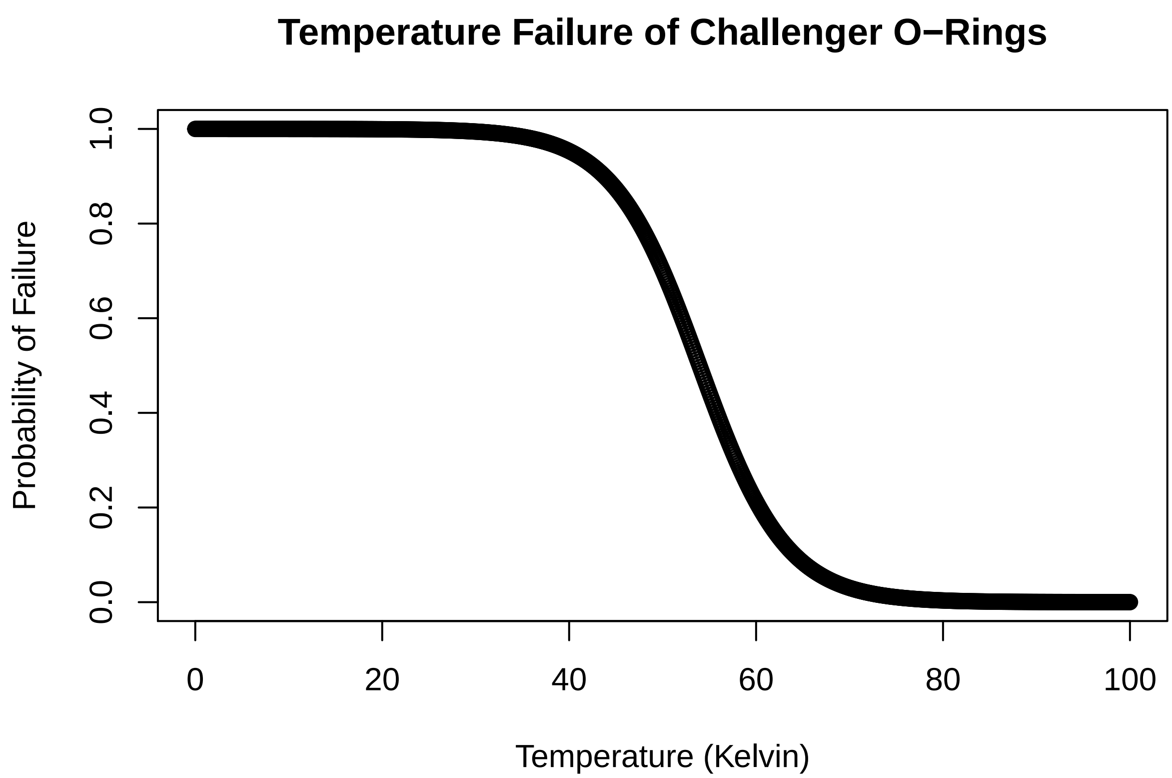
$$p(x) = \sigma(t) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

The key components are first the coefficients in the logistic regression above. The Intercept (11.6630) gives us the value of coefficient β_0 while the Temperature value (-0.2162) gives the coefficient $beta_1$. We also see that the z-value of 4.07 tell us that Temperature is statistically signficiant in the failure determination of the O-Rings.

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

$$p(x) = \sigma(t) = \frac{1}{1 + e^{-(11.6630 - 0.2162x)}}$$

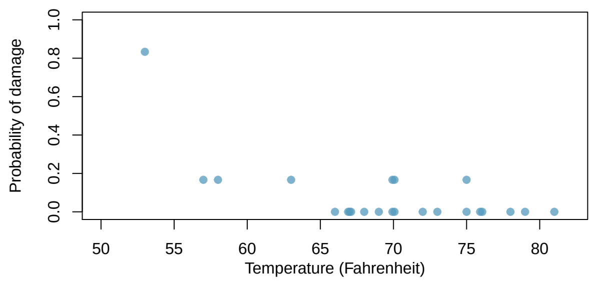
```r
B.0 <- 11.6630
B.1 <- -0.2162
x <- seq(0,100,0.1)
plot(x, 1/(1+exp(-(11.6630-0.2162*x))),main="Temperature Failure of Challenger O-Rings",
ylab="Probability of Failure", xlab="Temperature (Kelvin)")
```



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

I agree that temperature is a significant predictor in the failure of O-Rings. We see how the failure probability increases from 0.2 to 0.8 when temperature drops from 60K to 40K.

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:

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

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

$$\left(\frac{\hat{p}}{1-\hat{p}}\right) = e^{(11.6630 - 0.2162 \times T)}$$

$$\hat{p} = (1-\hat{p})e^{(\ldots)}$$

$$\hat{p} = e^{(\ldots)} - \hat{p}e^{(\ldots)}$$

$$\hat{p} + \hat{p}e^{(\ldots)} = e^{(\ldots)}$$

$$\hat{p}(1+e^{(\ldots)}) = e^{(\ldots)}$$

$$\hat{p} = \frac{e^{(\ldots)}}{1+e^{(\ldots)}}$$

When T = 51K

```
Temp <- 51
e.value <- 11.6630 - 0.2162*Temp
P.hat <- ( exp(e.value) ) / ( 1 + exp(e.value) )
cat("The model-estimated probabilities for failure at T = ", Temp,"K is", P.hat )
```

```
## The model-estimated probabilities for failure at T =  51 K is 0.6540297
```

When T = 53K

```
Temp <- 53
e.value <- 11.6630 - 0.2162*Temp
P.hat <- ( exp(e.value) ) / ( 1 + exp(e.value) )
cat("The model-estimated probabilities for failure at T = ", Temp,"K is", P.hat )
```

```
## The model-estimated probabilities for failure at T =  53 K is 0.5509228
```

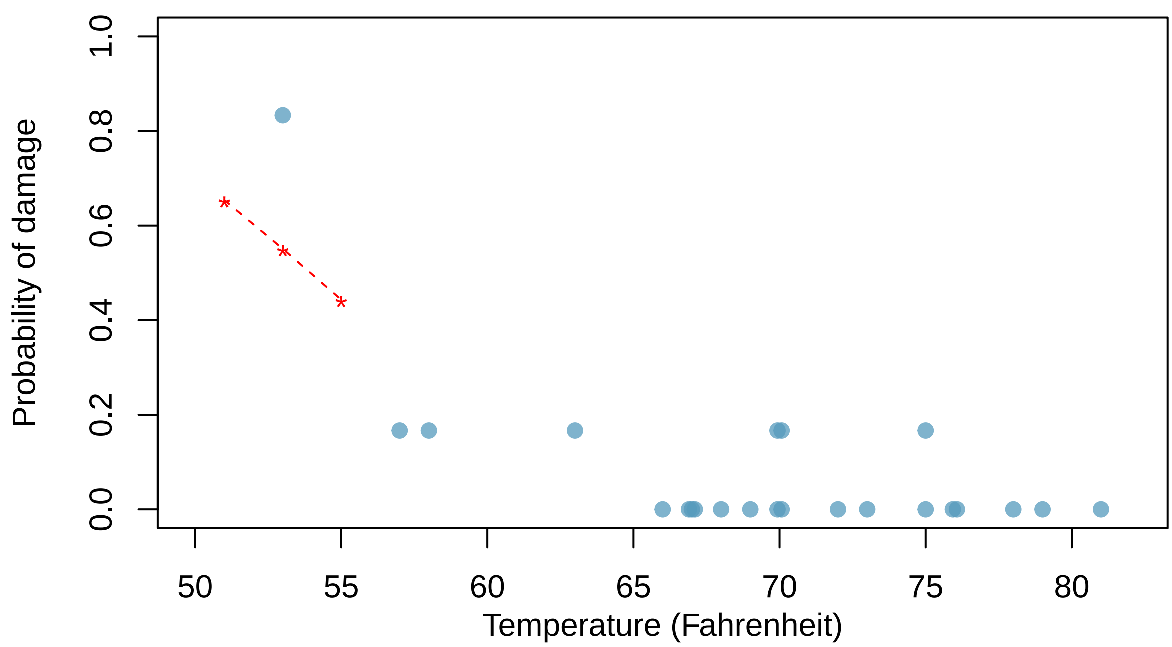
When T = 55K

```
Temp <- 55
e.value <- 11.6630 - 0.2162*Temp
P.hat <- ( exp(e.value) ) / ( 1 + exp(e.value) )
cat("The model-estimated probabilities for failure at T = ", Temp,"K is", P.hat )
```

```
## The model-estimated probabilities for failure at T =  55 K is 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.

```r
Prob <- c(0.6540297, 0.5509228, 0.4432456)
Temp <- c(51, 53, 55)
these <- orings[,1] %in% c(67, 70, 76)
plot(orings[,1] +
       c(rep(0, 5), c(-0.1, 0, 0.1), 0, 0, -0.07, -0.07, 0.07, 0.07,
         rep(0, 4), -0.07, 0.07, 0, 0, 0),
     orings[,2]/6,
     xlab = "", ylab = "Probability of damage",
     xlim = c(50, 82), ylim = c(0,1),
     col = COL[1,2], pch = 19)
mtext("Temperature (Fahrenheit)", 1, 2)
points(Temp, Prob, col="red", pch="*")
lines(Temp, Prob, col="red",lty=2)
```



(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.

Temperature has strong statistical signifiance in determining wether an O-Ring will fail or not. However, we can see how few data points are in the low temperature range when O-Rings typically fail. For example, we were able to estimate the failure probability for the temperatures of 51, 53, and 55 degrees Fahrenheit. The failure probability was generated using all the data points in the study. However, there is only one observation in this range of temperatures.