Chapter 9 - Multiple and Logistic Regression

Samuel Iguku Kigamba

11/28/2019

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.

$$Weight = minus8.94 * Smoke + 123.05$$

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

```
Babies born to a non-smoker is 123.05 oz and 123.05 ??? 8.94 = 114.11 oz for a smoker mother.
```

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

```
There is a statistically significant relationship between smoking and average birth weight as evic
```

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.

$$absntsm = minus 9.11 eth + 3.1 sex + 2.15 lrn + 18.9$$

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

```
eth: average number of absentee days of non-aboriginal students is 9.11 lower than that of aborigin

sex: average number of absentee days of male students is 3.10 higher than that of female students.

lrn: average number of absentee days of slow learners is 2.15 higer than that of 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.

```r
eth = 0
sex = 1
lrn = 1
total = 18.93 - (9.11*eth) + (3.10*sex) + (2.15*lrn)
2 - total #missed 2 days of school
```

```
## [1] -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.

```r
round(1 - (240.57 / 264.17), 3) # R^2
```

```
## [1] 0.089
```

```r
round(1 - (240.57 / 264.17) * (146 - 1) / (146 - 3 - 1), 3) #Adjusted R^2
```

```
## [1] 0.07
```

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 should remove the lrn 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
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 lower the temperatures the higher 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.

The low p-value means that the relationship between the temperatures and the damaged O-rings has a
The lowest temp recorded is 53 and thus the intercept is outside of this reasonable range.

	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

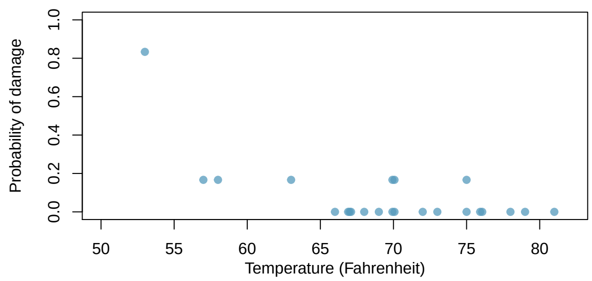
(c) Write out the logistic model using the point estimates of the model parameters.

$$log(p/1minusp) = 11.6630minus0.2162\ddot{O}Temperature$$

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

Yes. We are justified to have a concern over the O-ring due to the low p-value.

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.

```
p51<- exp(11.663-51*.2126)/(1+exp(11.663-51*.2126)) #ambient temperatures 51 degrees Fahrenheit
p51
```

```
## [1] 0.6943212
```

```
p53<- exp(11.663-53*.2126)/(1+exp(11.663-53*.2126)) #ambient temperatures 53 degrees Fahrenheit
p53
```

```
## [1] 0.5975339
```

```
p55<- exp(11.663-55*.2126)/(1+exp(11.663-55*.2126)) #ambient temperatures 55 degrees Fahrenheit
p55
```

```
## [1] 0.4925006
```

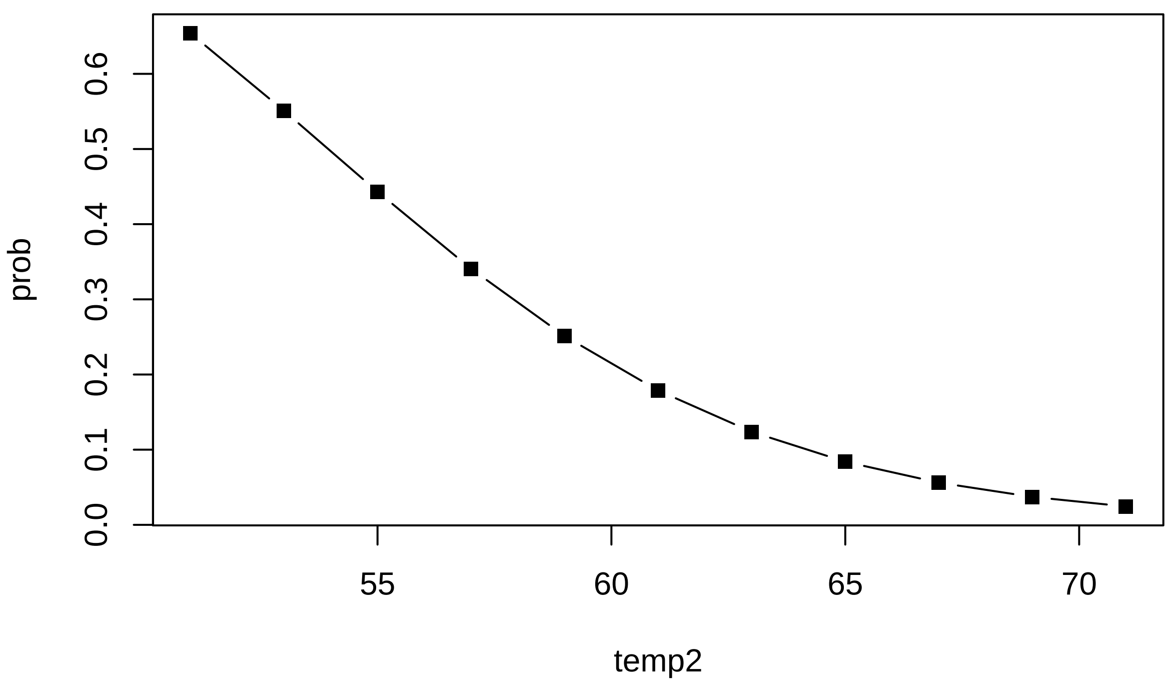
The model-estimated probabilities for several additional ambient temperatures are provided below, where subscripts indicate the temperature:

$\hat{p}_{57} = 0.341$ $\hat{p}_{59} = 0.251$ $\hat{p}_{61} = 0.179$ $\hat{p}_{63} = 0.124$

$\hat{p}_{65} = 0.084$ $\hat{p}_{67} = 0.056$ $\hat{p}_{69} = 0.037$ $\hat{p}_{71} = 0.024$

(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
temp2 <- c(seq(51, 71, 2))
prob <- exp(11.6630 - 0.2162 * temp2) / (1 + exp(11.6630 - 0.2162 * temp2))
plot(data.frame(temp2, prob), type = "b", pch = 15)
```



(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.

Assumptions: all observations are independent. We need to consider all variables that might be responsible to the damage of the O-rings.