Chapter 9 - Multiple and Logistic Regression

Donald Butler

11/29/2021

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.

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

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

The smoke variable has only two values, 0 for non-smokers and 1 for smokers. The x-axis will only have these two value, so the slope is equal to the estimate for the smoke variable, -8.94.

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

Yes, with a p-value about zero, the smoke attribute is a significant predictor of birth weight.

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.

$$\widehat{days} = 18.93 - 9.11 \times ethnicity + 3.10 \times sex + 2.15 \times learner$$

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

Non-aboriginal students will have 9.11 fewer absences than aboriginal students. Male students will have 3.10 more absences than female students. Slow learners will have 2.15 more absences 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.

```
2 - (18.93 - 9.11 * (0) + 3.10 * (1) + 2.15 * (1))
```

```
## [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.

```
residual_var <- 240.57
days_var <- 264.17
n <- 146
k <- 3
r2 <- 1 - (residual_var / days_var)
r_adj <- 1 - (residual_var / days_var) * ((n-1) / (n-k-1))
```

$R^2 = 0.0893364$
$R^2_{adj} = 0.070097$

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?

The learner status variable should be removed first because the R^2_{adj} value is higher without it.

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.

9 of the 12 damaged o-rings occurred at 63 degrees and colder which seems to indicate that the temperature is a predictor for the failure of an o-ring.

(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 intercept of 11.6630 indicates the expected number of damaged o-rings at a temperature of 0°. For each degree increase in temperature, the expected number of o-ring failures decreases by 0.2162.

(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.

With a p-value of about zero, temperature is a significant predictor of o-ring failures.

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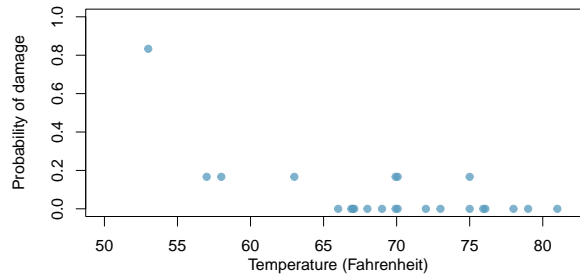
(a) The data provided in the previous exercise are shown in the plot. The logistic model fit to these data may be written as

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

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:

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

```
p_damage <- function(temp) {
  damaged <- 11.6630 - .2162 * temp
  return(exp(damaged)/(1+exp(damaged)))
}

p51 <- p_damage(51)
p53 <- p_damage(53)
p55 <- p_damage(55)
```

$p_{51} = 0.6540297$
$p_{53} = 0.5509228$
$p_{55} = 0.4432456$

(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.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
temp <- seq(50,80,1)
orings_est <- data.frame(temp = temp,pct = p_damage(temp))

orings_est %>%
  ggplot(aes(temp,pct)) +
  geom_point() +
  geom_smooth(formula = y ~ x, method = loess, se = FALSE)
```
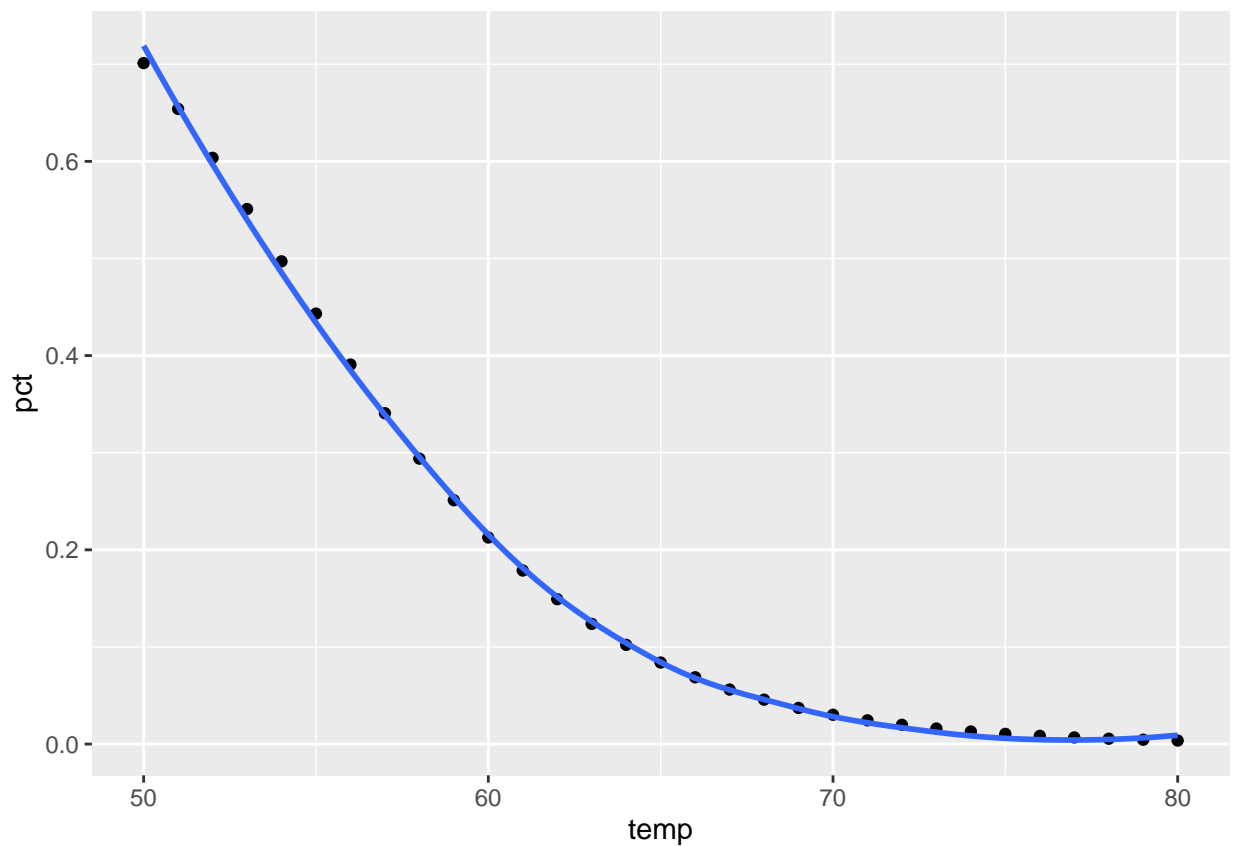

(c) Describe any concerns you may have regarding applying logistic regression in this application, and note any assumptions that are required to accept the model's validity.

1. Each outcome Y_i is independent of the other outcomes.

The shuttle missions are independent events.

2. Each predictor x_i is linearly related to $\text{logit}(p_i)$ if all other predictors are held constant.