Chapter 9 - Multiple and Logistic Regression

Chunjie Nan

**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.

The variability within the smokers and nonsmokers 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.)

1. Write the equation of the regression line.

Answer:

smoke<-c(1,0)  
Y = 123.05 - 8.94 \* smoke  
Y

## [1] 114.11 123.05

1. Interpret the slope in this context, and calculate the predicted birth weight of babies born to smoker and nonsmoker mothers.

Answer: “The predicted wieght for smokers is 114.11 ounces” “The predicted wieght for nonsmokers is 123.05 ounces”

1. Is there a statistically significant relationship between the average birth weight and smoking?

Answer: Since the p value is zero, the smoking and weight is statistically significant.

**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.

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).

1. Write the equation of the regression line.

y=18.93−9.11×eth+3.10×sex+2.15×lrn

1. Interpret each one of the slopes in this context.

The slope of eth shows that all others equal, a 9.11 day reduction in the predicted absenteeism when the subject is NO aboriginal.

The slope of sex shows that all others equal, a 3.10 day increase in the predicted absenteeism when the subject is male.

The slope of lrn shows that all else being equal, a 2.15 day increase in the predicted absenteeism when the subject is a slow learner.

1. 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.

eth <- 0   
sex <- 1  
lrn <- 1   
missed\_days <- 2  
  
predicted\_days <- 18.93 - 9.11 \* eth + 3.1 \* sex + 2.15 \* lrn  
  
residual <- missed\_days - predicted\_days  
residual # the residual is -22.18

## [1] -22.18

1. 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 and the adjusted . Note that there are 146 observations in the data set.

n <- 146   
k <- 3   
var\_residual <- 240.57   
var\_students <- 264.17   
R2 <- 1 - (var\_residual/ var\_students) # R2  
adj\_R2 <- 1 - (var\_residual / var\_students) \* ( (n-1) / (n-k-1) )   
  
R2

## [1] 0.08933641

adj\_R2

## [1] 0.07009704

**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 Rsquared for the model as well as adjusted Rsquared values for all models we evaluate in the first step of the backwards elimination process.

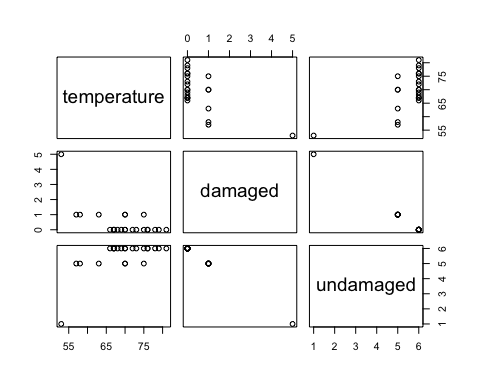
Which, if any, variable should be removed from the model first?

According to the adjusted R2 (0.0723)above, we can remove the lrn first among the variables.

**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.

1. 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.

temperature <- c(53,57,58,63,66,67,67,67,68,69,70,70,70,70,72,73,75,75,76,76,78,79,81)  
damaged <- c(5,1,1,1,0,0,0,0,0,0,1,0,1,0,0,0,0,1,0,0,0,0,0)  
undamaged <- c(1,5,5,5,6,6,6,6,6,6,5,6,5,6,6,6,6,5,6,6,6,6,6)  
  
Shuttle\_Mission <- data.frame(temperature, damaged, undamaged)  
plot(Shuttle\_Mission)

 Higher damaged O rings are lower temperatures were recorded. Less damaged O rings are when higher temperatures were recorded.

1. 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 Intercept and the second one is the Temperature values. The Estimate identifies the parameter estimate for the model. The Z value and the P value help to see the significancy in the model.

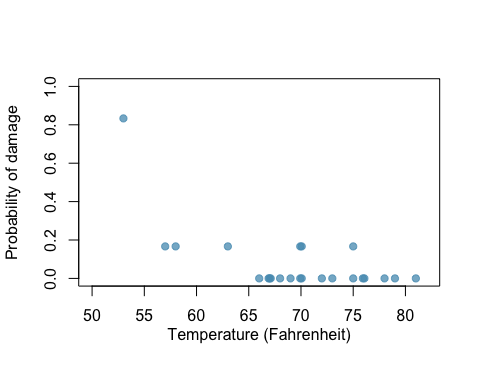
1. Write out the logistic model using the point estimates of the model parameters.

log(p/(1 minus p))=11.6630 minus 0.2162×Temperature

1. Based on the model, do you think concerns regarding O rings are justified? Explain.

The 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.

\begin{center}  \end{center}

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

where 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:

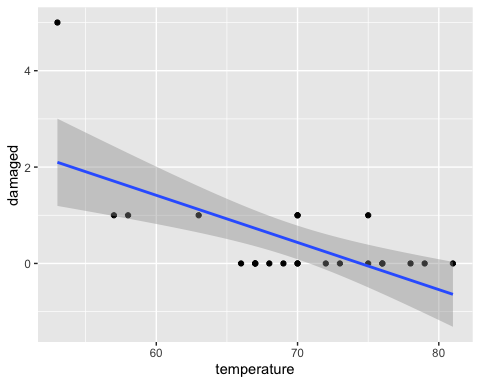
1. Add the model estimated probabilities from part~(a) on the plot, then connect these dots using a smooth curve to represent the model estimated probabilities.

library(ggplot2)  
g<-ggplot(Shuttle\_Mission,aes(x=temperature,y=damaged))   
g<-g + geom\_point()   
g<-g+stat\_smooth(method = 'glm', family = 'binomial')

## Warning: Ignoring unknown parameters: family

g

## `geom\_smooth()` using formula 'y ~ x'

 (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.

Each predictor is linearly related to logit if all other predictors are held constant. Each outcome is independent of the other outcomes.

We assume that those our model met the above conditions.