```
---
title: "Chapter 9 - Multiple and Logistic Regression"
author: "Leticia Salazar"
output:
  html_document:
    df_print: paged
  pdf_document:
    extra_dependencies:
    - geometry
    - multicol
    - multirow
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```


```{r}
library("DATA606")
```


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.

\begin{center}
\begin{tabular}{rrrrr}
  \hline
           & Estimate & Std. Error & t value   & Pr($>$$|t|$) \\
  \hline
(Intercept) & 123.05   & 0.65   & 189.60   & 0.0000 \\
smoke   & -8.94   & 1.03   & -8.65   & 0.0000 \\
  \hline
\end{tabular}
\end{center}

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.

Intercept = 123.05; slope=-8.94

y = mx + b
```

babyweight = 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 indicated the estimated weight of babies born to mothers who smoke vs. the mothers who don't smoke. For mothers who smoke the predicted birth weight of babies is 114.11 oz while for the non smoker mothers the predicted baby weight is 123.05 oz.

```{r}
# smoker
smoke_baby_wgt <−123.05−8.94*1
smoke_baby_wgt

# non−smoker
non_smoke_baby_wgt <− 123.05−8.94*0
non_smoke_baby_wgt
```

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

Being that the p−value is closer to 0 we can reject the null hypothesis where Ho: B1 is equal to 0 and Ha: B1 is not equal to 0. There is a negative correlation between the babies weight born from smoker or non−smoker mothers.

\clearpage

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.

\begin{center}
\begin{tabular}{r c c c c}
  \hline
    & eth & sex & lrn & days \\
  \hline
1 & 0   & 1   & 1   &   2 \\
2 & 0   & 1   & 1   & 11 \\
$\vdots$ & $\vdots$ & $\vdots$ & $\vdots$ & $\vdots$ \\
146 & 1   & 0   & 0   & 37 \\
  \hline
\end{tabular}

\end{center}

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).

\begin{center}
\begin{tabular}{rrrrr}
  \hline
           & Estimate & Std. Error & t value   & Pr($>$$|t|$) \\
  \hline
(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 \\
  \hline
\end{tabular}
\end{center}

(a) Write the equation of the regression line.

y=18.93−9.11*eth+3.10*sex+2.15*lrn

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

eth: predicts the absenteeism for non−aboriginal students

sex: predicts the average number of days absent by male students

lrn: predicts the average number of days absent by slow 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.

The residual for the first observation of the given student is negative 22.18.

```{r}
eth <−0

# male only
sex <−1
lrn <−1
days_missed <−2
prediction <−18.93−9.11*eth+3.1*sex+2.15*lrn

# residuals
first_observ <−days_missed−prediction
first_observ
```

(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.

The r-squared is 0.0893 and the adjusted r-squared is 0.0701.

```{r}
# variance residual and outcome
residual <-240.57
outcome <-264.17
n <- 146
k <-3

# $R^2$
r_squ <-1-(residual / outcome)
r_squ

# adjusted $R^2$
adju_r2 <-1-(residual / outcome)*((n-1)/(n-k-1))
adju_r2
```

\clearpage

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.

\begin{center}
\begin{tabular}{rlr}
  \hline
  & Model & Adjusted R^2 \\
  \hline
1 & Full model & 0.0701 \\
2 & No ethnicity & -0.0033 \\
3 & No sex & 0.0676 \\
4 & No learner status & 0.0723 \\
  \hline
\end{tabular}
\end{center}

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

We could eliminate the `no ethnicity` since it has a negative value.

\clearpage

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.

\begin{center}
\begin{tabular}{l rrrrr rrrrr rrr}
\hline
\vspace{-3.1mm} \\
Shuttle Mission   & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\
\hline
\vspace{-3.1mm} \\
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 \\
\hline
\\
\cline{1-12}
\vspace{-3.1mm} \\
Shuttle Mission   & 13 & 14 & 15 & 16 & 17 & 18 & 19 & 20 & 21 & 22 & 23 \\
\cline{1-12}
\vspace{-3.1mm} \\
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 \\
\cline{1-12}
\end{tabular}
\end{center}

(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.

Each column above represents a different shuttle mission with the data collected being an observation in respect to the temperature and damaged / undamaged O-rings. As the temperature increases the frequency of the damaged O-rings decreases.

(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.

With a p-value of 0 there's meaning to the relationship between the temperatures and damaged O-rings.

\begin{center}
\begin{tabular}{rrrrr}
  \hline
           & Estimate & Std. Error & z value   & Pr($>$$|z|$) \\
  \hline
(Intercept) & 11.6630 & 3.2963   & 3.54   & 0.0004 \\
Temperature & -0.2162 & 0.0532   & -4.07   & 0.0000 \\
  \hline
\end{tabular}
\end{center}

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

log(p/(1-p)) = 11.6630-0.2162xTemperature

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

Based on the model I do think the concerns about the O-rings are justified. The p-value justifies a strong correlation since it is of a low value.

\clearpage

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}

```{r, echo=FALSE, message=FALSE, warning=FALSE, fig.show="hold",
out.width="50%", fig.height=4}
library(openintro)
# load data ----------------------------------------------------------
if(!file.exists('orings.rda')) {
 download.file('https://github.com/jbryer/DATA606Fall2019/blob/master/
course_data/orings.rda?raw=true',
      'orings.rda')
}
load("orings.rda")
set.seed(17)
# plot probability of damage vs. temperature -----------------------
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)
# probability calculations -----------------------------------------
temperature <-c(51, 53, 55)
logitp <-11.6630 -0.2162 * temperature
p <-exp(logitp) / (1+exp(logitp))
```
\end{center}

(a) The data provided in the previous exercise are shown in the plot. The logistic model fit to these data may be written as
\begin{align*}
\log\left(\frac{\hat{p}}{1-\hat{p}} \right) = 11.6630-0.2162\times Temperature
\end{align*}

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{align*}
&\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{align*}

The probability of O-ring damage at 51 degrees Fahrenheit is 69.43%, at 53 degrees Fahrenheit is 59.75% and at 55 degrees Fahrenheit is 49.25%.

```{r}
# temperature 51

p_51 <-exp(11.663-51*.2126) / (1 + exp(11.663-51*.2126))
p_51

# temperature 53
p_53 <-exp(11.663-53*.2126) / (1 + exp(11.663-53*.2126))
p_53

# temperature 55
p_55 <-exp(11.663-55*.2126) / (1 + exp(11.663-55*.2126))
p_55
```

(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.

Plot

```{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.

Some of my assumptions are that the observations appear to be independent from each other but we'd have to consider all the variables more than 23 missions.