Chapter 9 - Multiple and Logistic Regression

Zachary Palmore

```
library(DATA606)
```

```
## Loading required package: shiny

## Loading required package: openintro

## Loading required package: airports

## Loading required package: cherryblossom

## Loading required package: usdata

## Loading required package: OIdata

## Loading required package: RCurl

## Loading required package: maps

## Loading required package: ggplot2

## Loading required package: markdown

##
## Welcome to CUNY DATA606 Statistics and Probability for Data Analytics
## This package is designed to support this course. The text book used
## is OpenIntro Statistics, 3rd Edition. You can read this by typing
## vignette('os3') or visit www.OpenIntro.org.
##
## The getLabs() function will return a list of the labs available.
##
## The demo(package='DATA606') will list the demos that are available.

##
## Attaching package: 'DATA606'

## The following objects are masked from 'package:openintro':
##
##     calc_streak, present, qqnormsim
```

```
## The following object is masked from 'package:utils':
##
##     demo
```

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.

```
data(babies)
summary(lm(babies$bwt ~ babies$smoke))
```

```
##
## Call:
## lm(formula = babies$bwt ~ babies$smoke)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -68.05 -11.05   0.89  10.95  52.95
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   123.047      0.649 189.597   <2e-16 ***
## babies$smoke   -8.938      1.033  -8.653   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.68 on 1224 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.05764,    Adjusted R-squared:  0.05687
## F-statistic: 74.87 on 1 and 1224 DF,  p-value: < 2.2e-16
```

y = -8.94x + 123.05 OR y = 123.05 - 8.94x

Where y = weight of a baby in ounces and x = status of the mother as smoker or non-smoker.

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

The slope is the average expected change in baby weights given the status of the mother as a smoker or not. Given the negative trend in the slope, we can predict that babies born to smoker mothers weight less than non-smoker mothers.

Using the equation of the regression line we can predict the average weight of a baby in ounces born to mothers who smoke and those who do not given their categorical status of 1 and 0 respectively.

```
# Baby born to smoker mother
123.05 - 8.94*1
```

```
## [1] 114.11
```

```
# Baby born to non-smoker mother
123.05 - 8.94*0
```

```
## [1] 123.05
```

We can see that, based on this analysis, the average baby weight of smoker mothers is 8.94 ounces less than the babies born from those who are non-smokers.

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

If our null hypothesis is that there the difference in average baby weight of babies born to smoking and non-smoking mothers is zero and the alternative hypothesis is that there is a difference in the average weight of babies born to smoking and non-smoking mothers then the results are statistically significant and we reject the null hypothesis in favor of the alternative. It appears there is a significant negative correlation between the average baby weight and the status of the mother as a smoker or non-smoker.

In short, yes, it is statistically significant beyond the 0.001 level.

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.

$$y = 18.98 - 9.11a + 3.10b + 2.15c$$

Where y is the number of days absent, a is the variable for ethnicity, b is the sex, and c is the learning status.

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

The slope of the ethnicity variable indicates that the average number of days absent for non-aboriginal students is 9.11 days lower than that of aboriginal students

The slope of the variable for sex indicates that the average number of days absent for male students is 3.10 days higher than that of female students

The slope of the status for learning indicates that the average number of days absent for slow learning students is 2.15 days higher than that of normal students.

(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.

```
a <- 0 # aboriginal
b <- 1 # male
c <- 1 # slow learner
total <- 18.93 - (9.11*a) + (3.10*b) + (2.15*c) # find total est days
2 - total # missed 2 days of school
```

```
## [1] -22.18
```

The residual for the first observation who is aboriginal, male, a slow learner, and missed 2 days of school is -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.

We can calculate the R^2 as:

```
1 - (240.57 / 264.17)
```

```
## [1] 0.08933641
```

The adjusted R^2 value is found by including the sample size demonstrated by:

```
1 - (240.57 / 264.17) * (146 - 1) / (146 - 3 - 1)
```

```
## [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 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?

Learner status should be removed because it improves the adjusted r-squared values slightly.

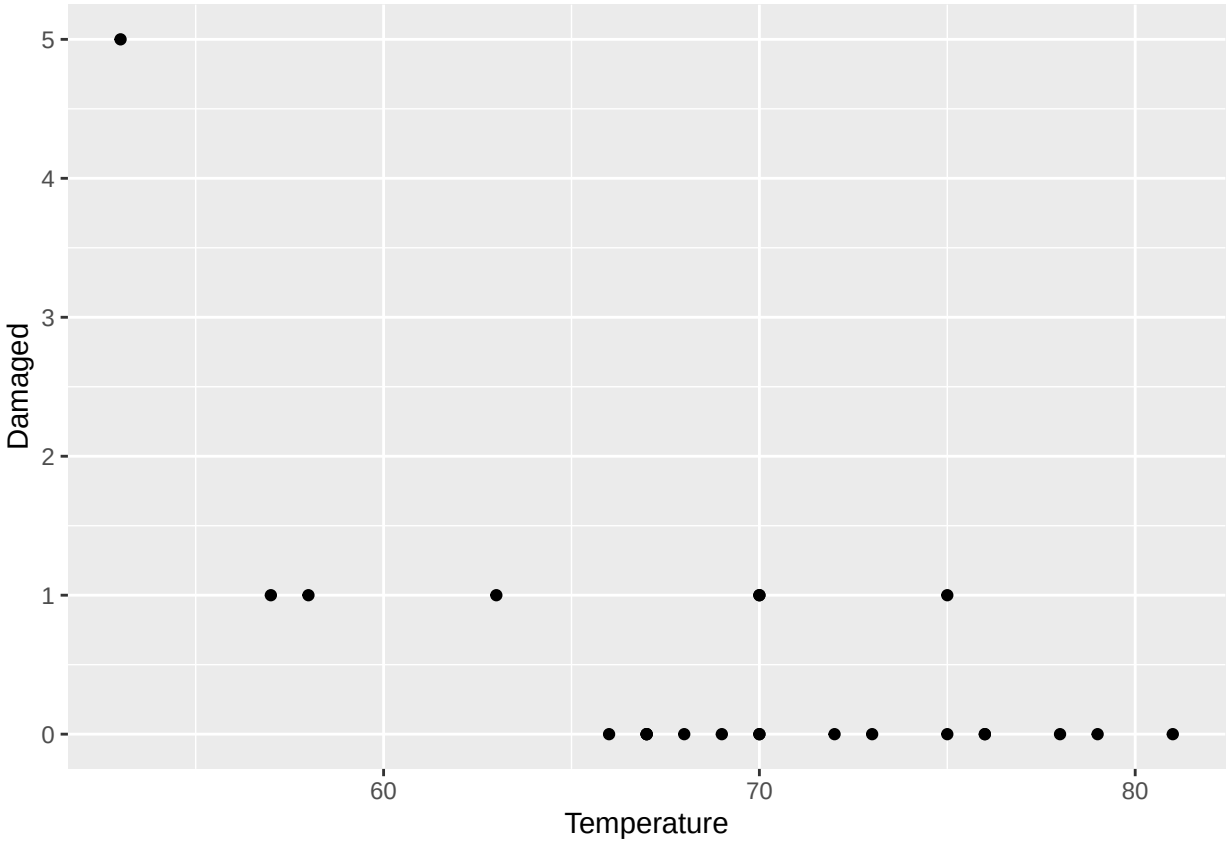
Challenger disaster, Part I. (9.16, p. 380) On January 28, 1986, a routine launch was anticipated for the Challenger space shuttle. Seventy-three seconds into the flight, disaster happened: the shuttle broke apart, killing all seven crew members on board. An investigation into the cause of the disaster 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

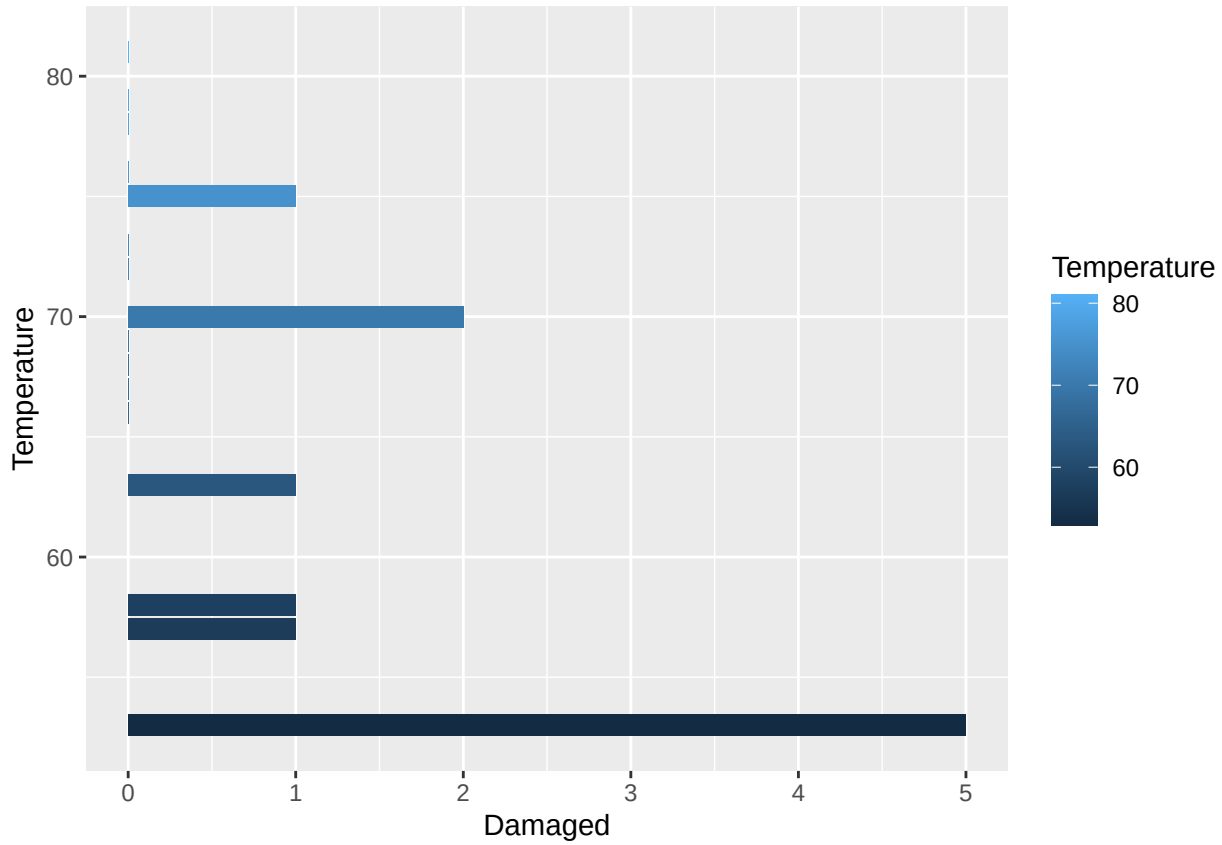
Shuttle Mission	13	14	15	16	17	18	19	20	21	22	23
Temperature	70	70	72	73	75	75	76	76	78	79	81
Damaged	1	0	0	0	0	1	0	0	0	0	0
Undamaged	5	6	6	6	6	5	6	6	6	6	6

(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.

```
Temperature <-  c(53, 57, 58, 63, 66, 67, 67, 67, 68, 69, 70, 70, 70, 70, 72, 73, 75, 75, 76, 76, 78, 79
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)
oringstats <- data.frame(cbind(Temperature, Damaged, Undamaged))
ggplot(oringstats, aes(Temperature, Damaged)) + geom_point()
```

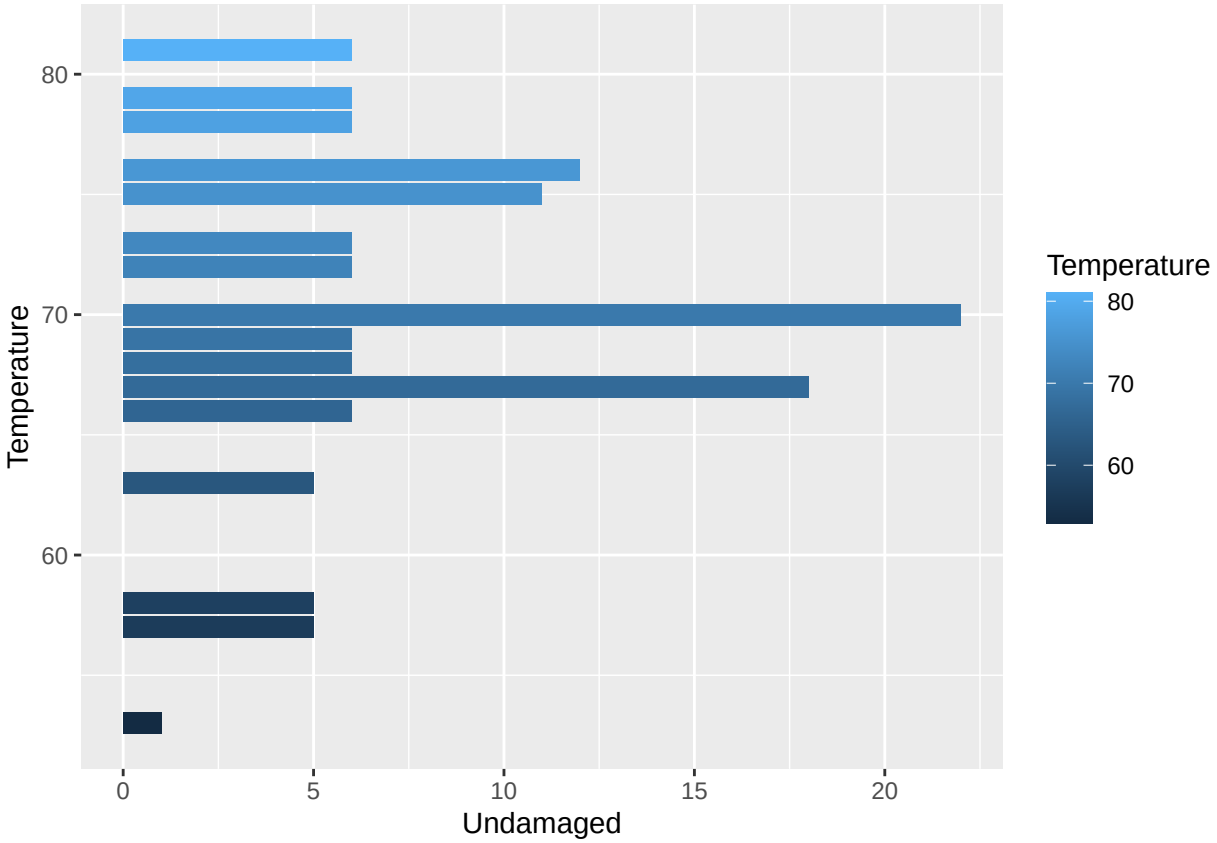



```
ggplot(oringstats, aes(Temperature, Damaged)) + geom_col(aes(fill = Temperature)) + coord_flip()
```



It appears as though higher temperatures resulted in fewer damaged orings. The reverse also appears to be true given the temperature of most undamaged orings. Specifically, that lower temperatures resulted in greater chance of damaged orings.

```
ggplot(oringstats, aes(Temperature, Undamaged)) + geom_col(aes(fill = Temperature)) + coord_flip()
```



(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

Given the slope of the regression equation, there is a negative trend in damaged orings as temperature increases. This suggests higher temperatures result in a lower chance of damaging the orings.

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

Given the slope and intercept we can write the logistic model as:

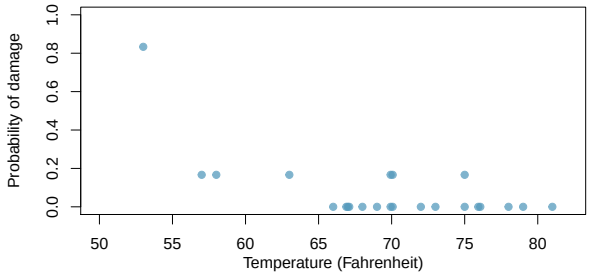
$$log(\frac{\hat{p}}{1-\hat{p}}) = 11.6630 - 0.2162(t)$$

Where "t" is the temperature and $\hat{p}$ is the probability that the oring will be damaged in this model.

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

Given the significance of these results, yes, I think the concerns regarding o-rings are justified since it is very unlikely they would have occurred by chance. The results are signifcant at the .001 alpha level.

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:

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

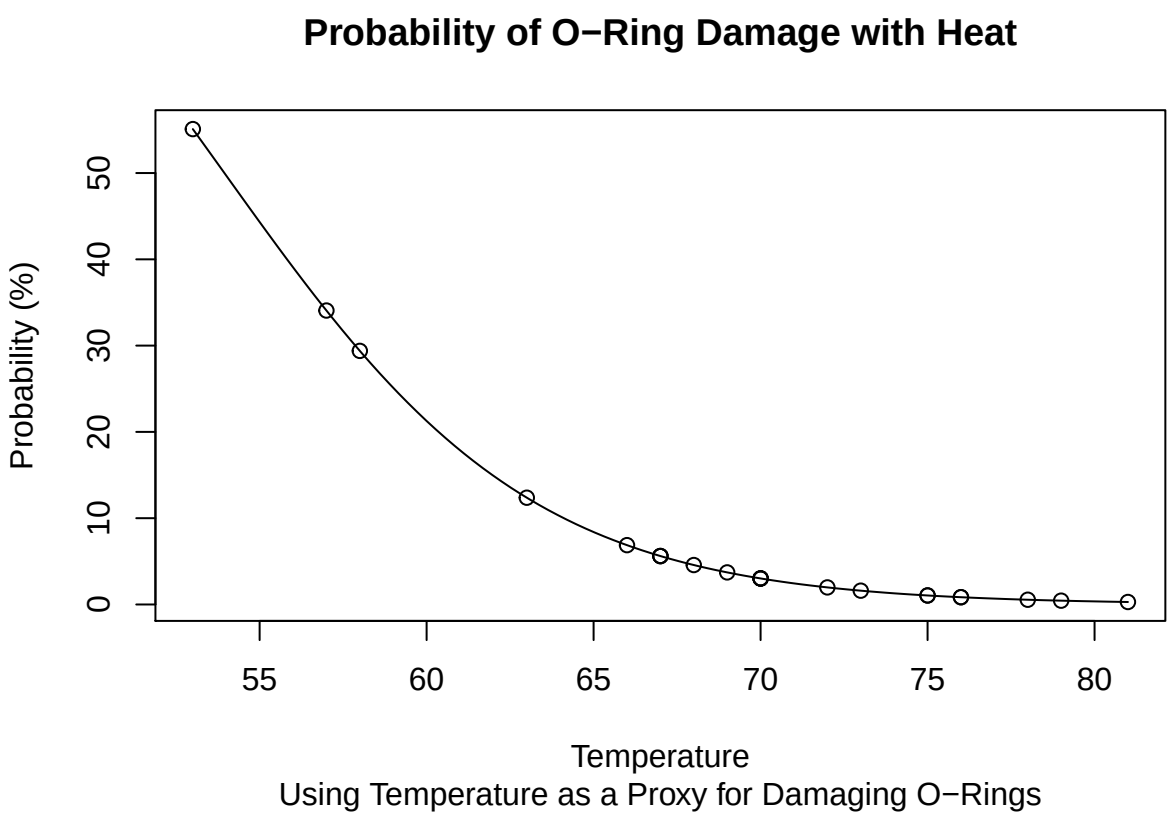
```
# Using a loop
t = c(51,53,55) # add selected temperatures
fun <- function(t){
  o  <- 11.6630 - 0.2162 * t  # use model paramters
  Poi <- 100*(exp(o) / (1+exp(o))) # calculate probability
  return(Poi)
}
sapply(t, fun) # apply to the selected temperatures
```

```
## [1] 65.40297 55.09228 44.32456
```

(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.

```
# Applying the function to the all Temperatures
# (This should show repeats of model-est probabilities given above)
meps <- sapply(Temperature, fun) # Store as model - estimated probabilities - "meps"
```

```r
# Then plot them over the change in temperature
plot(y=meps, x=Temperature, ylab="Probability (%)", main = "Probability of O-Ring Damage with Heat", su
curve(fun(x), add=TRUE)
```



(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.

We assume that the result of each probability is independent of the other results and that the predictors are linearly related. This second condition is difficult to validate since there is a small sample size. If the sample size were larger, this would be easier to validate. It would also help if more variables were reviewed to exclude other factors that may have caused (or will cause) damage to O-Rings.