CUNY SPS DATA606 HW9 - Multiple and Logistic Regression

Chinedu Onyeka

October 28th, 2021

Problem 1:

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.
(b) Interpret the slope in this context, and calculate the predicted birth weight of babies born to smoker and non-smoker mothers.
(c) Is there a statistically significant relationship between the average birth weight and smoking?

Solution 1:

1a)

The equation of the regression line is given by:
$babyweight = 123.05 - 8.94 * smoke$

1b)

The slope in this context is the amount that the baby_weight will decrease by if the mother is a smoker.

```
smoke_yes <- 1
smoke_no <- 0
baby_weight <- function(smoke){
  weight <- 123.05 - 8.94*smoke
  return(weight)
}
paste0("The predicted birth weight of babies born to smoker is ", baby_weight(smoke_yes))
```

```
## [1] "The predicted birth weight of babies born to smoker is 114.11"
```

```
paste0("The predicted birth weight of babies born to non-smoker is ", baby_weight(smoke_no))
```

```
## [1] "The predicted birth weight of babies born to non-smoker is 123.05"
```

1c)

$H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$
T = -8.65 and the p-value is 0. Therefore, we reject the null hypothesis that the true slope is zero. Hence, we can conclude that there is statistically significant relationship between the average birth weight and smoking habit of mothers.

Problem 2:

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

Solution 2:

2a)

The equation of the regression line is given by:
$absenteeism = 18.93 - 9.11 * eth + 3.10 * sex + 2.15 * lrn$

2b)

first slope: $-9.11 * eth$: The average number days of absenteeism will decrease by 9.11 if the ethnic background of the student is not aboriginal assuming that other demographic characteristics are held constant.
second slope: $3.10 * sex$: The average number of days of absenteeism will increase by 3.10 if the student is male compared to when the student is female assuming that other demographic characteristics are held constant.
third slope: $2.15 * lrn$: The average number of days of absenteeism will increase by 2.15 if the student is a slow learner compared to an average learner assuming that other demographic characteristics are held constant.

2c)

```
# define the absenteeism model as a function
absenteeism <- function(eth, sex, lrn){
  absent <- 18.93 -9.11*eth + 3.10*sex + 2.15*lrn
```

```
  return(absent)
}

# Residuals for the student with details below:
eth <- 0
sex <- 1
lrn <- 1
absenteeism_predicted <- absenteeism(eth, sex, lrn)
absenteeism_actual <- 2
residual <- absenteeism_actual - absenteeism_predicted
paste0("The residual for a student who is aboriginal, male, a slow learner,",
       "and missed 2 days of school is: ",residual)
```

```
## [1] "The residual for a student who is aboriginal, male, a slow learner,and missed 2 days of school i
```

2d)

$R^2 = 1 - \frac{Var(residuals)}{Var(outcome)}$

$R^2_{adj} = 1 - \frac{Var(residuals)}{Var(outcome)} * \frac{n-1}{n-k-1}$

```
var_residual <- 240.57
var_outcome <- 264.17
n <- 146
k <- 3
r_squared <- round((1 -(var_residual/var_outcome)),4)
r_squared_adj <- round((1 - (var_residual/var_outcome)*((n - 1)/(n - k - 1))),4)
paste0("The R-squared is: ", r_squared)
```

```
## [1] "The R-squared is: 0.0893"
```

```
paste0("The R-squared-adjusted is: ", r_squared_adj)
```

```
## [1] "The R-squared-adjusted is: 0.0701"
```

Problem 3:

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?

Solution 3:

Using the process of backward elimination, we should remove the variable with the lowest adjusted R^2. Hence, remove the "No ethnicity" variable first since it has the lowest adjusted R^2.

Problem 4:

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.

(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

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

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

Solution 4:

4a)

For the most part, as the temperature increases, the number of damages O-rings tend to decrease. Hence, there appear to be a negative relationship between temperature and damaged O-rings.

4b)

The framework of the logistic regression is similar to that of multiple linear regression. Therefore, the intercept of the logit function will be 11.6630 while the coefficient of temperature is -0.2162

4c)

The logit model using the point estimates of the model is given by:
$logit(p_i) = log_e(\frac{p_i}{1-p_i}) = 11.6630 - 0.2162 * temperature$

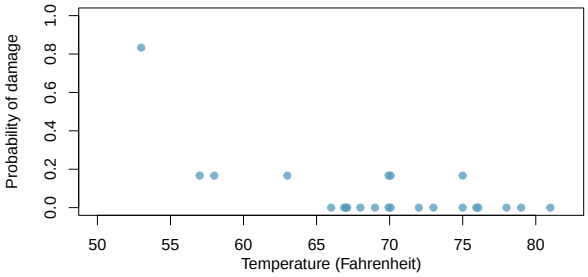
4d)

$H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$

The p-value is approximately 0 and less than 0.05. Therefore, we reject the null hypothesis that the true slope is zero. Hence, we can conclude that there is statistically significant relationship between temperature and damaged O-rings. Hence, the concerns regarding O-rings are justified.

Problem 5:

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:

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

(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.

(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.

Solution 5:

5a)

$\hat{p} = \frac{e^{11.6630-0.2162*temperature}}{1+e^{11.6630-0.2162*temperature}}$

```
# define a function to compute the probabilities
prob_func <- function(temperature){
  logit_p <- 11.6630 - 0.2162 * temperature
  p_hat <- exp(logit_p) / (1+exp(logit_p))
  return(round(p_hat,4))
```

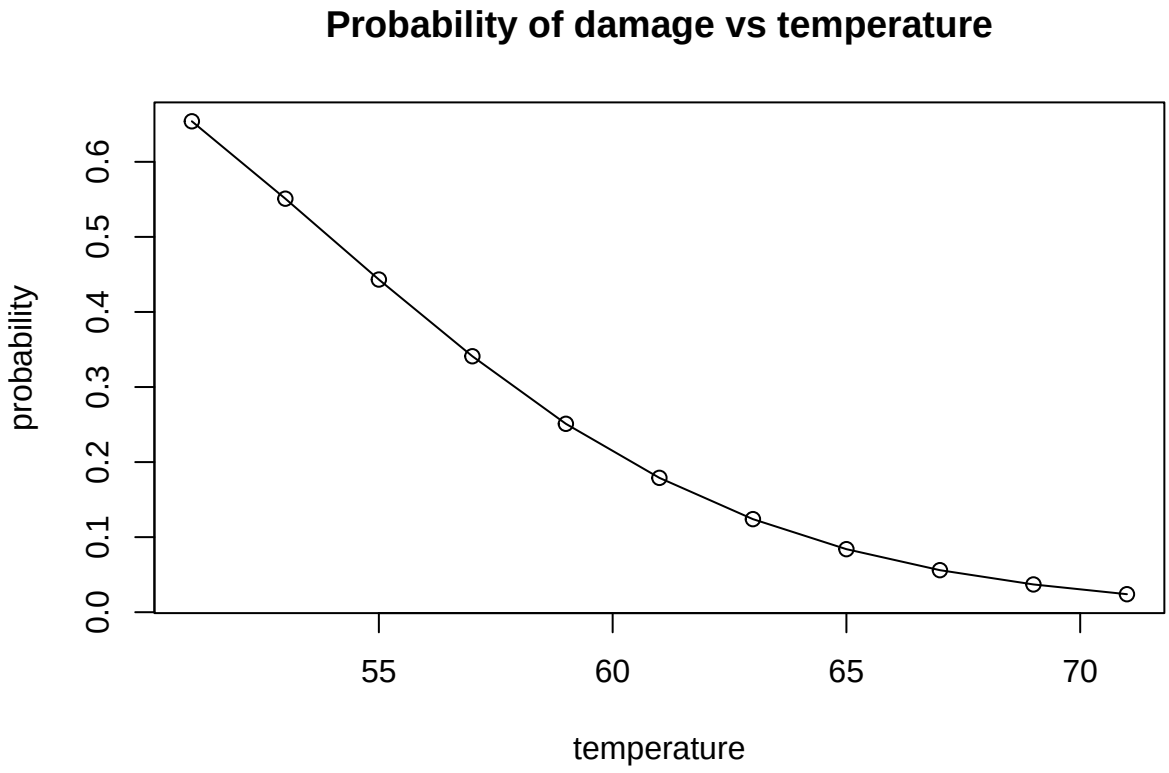
```
}
temp_samp <- c(51, 53, 55)
paste0("The probability that an O-ring will be damaged at temperature ", temp_samp,
       " degrees Fahrenheit is: ", prob_func(temp_samp))
```

```
## [1] "The probability that an O-ring will be damaged at temperature 51 degrees Fahrenheit is: 0.654"
## [2] "The probability that an O-ring will be damaged at temperature 53 degrees Fahrenheit is: 0.5509"
## [3] "The probability that an O-ring will be damaged at temperature 55 degrees Fahrenheit is: 0.4432"
```

5b)

```
temperature <- seq(51, 71, by = 2)
probability <- c(prob_func(temp_samp), 0.341, 0.251, 0.179, 0.124, 0.084, 0.056, 0.037, 0.024)

plot(temperature, probability, main = "Probability of damage vs temperature")
lines(temperature, probability)
```



5c)

The major concern I will have in applying logistic regression in this application is the sample size. I think that the sample size in this case is too small and not enough to get an optimal model. Two major conditions are required to apply Logistic Regression and they are:
i) Each outcome Y_i is independent of the other outcomes.
ii) Each predictor x_i is linearly related to $logit(p_i)$ if all other predictors are held constant.