Chapter 9: Multiple and Logistic Regression

Alexis Mekueko

11/26/2020

Contents

Github Link: https://github.com/asmozo24/DATA606_Homework9

Web link: https://rpubs.com/amekueko/696316

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.

Answer: Y = ax + b, a = slope = -8.94, b = intercept = 123.05, Y = Average birth weight of babies

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

Answer: the slope in this case means the ratio of the average birth weight of babies born to the mother health type (smoker, non-smoker). The predicted birth weight of babies born to smoker = -8.94 * 1 + 123.05 = 114.11 . The predicted birth weight of babies born to non-smoker = -8.94 * 0 + 123.05 = 123.05

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

Answer: Ye! Just by looking at the predictor with two levels, there is a difference of abs(114.11 - 123.05) = 8.94 in the average birth weight of babies born between mother smoking and those that don't. Since this study is based on baby born weight in ounce, 8.9 ounce is 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 studentsin rural New South Wales, Australia, in a particular school year. Below are three observations from this dataset.

	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.

Answer: there are 03 predictors, 03 slopes and 01 intercept.

```
y = a1x1 + a2x2 + a3x3 + b
b = 18.93
a1 = -9.11, x1 = eth( ethical background)
a2 = 3.10, x2 = sex
a3 = 2.15, x3 = lrn (learner status)
```

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

Answer: a1 is the ratio of the average number of days absent to students with ethnic background of aboriginal and not aboriginal.

```
a2 is the ratio of the average number of days absent to male and femalee students .
a3 is the ratio of the average number of days absent to slow and average learner 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.

Answer: residual = actual - predicted

```
the actual missed number of day of school  is 2.
the predicted missed number of day of schoo is -9.11 * 0 + 3.10 * 1 + 2.15 * 1 + 18.93 = 24.18
residual = 2 - 24.18 = - 22.18 (so, the predicted is below the regression line )
```

(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 Rsquare and the adjusted Rsquare. Note that there are 146 observationsin the data set.

Answer: Rsquare = 1 - RSS/TSS, Rsquare = coefficient of determination , RSS = sum of squares of residuals, TSS= total sum of squares. Adjusted Rsquare = 1 - ((1-Rsquare)(N-1)/(N-M-1)), where M = number of predictors, N = total sample size

```
## The calculated Rsquare is  8.93 %
```

```
## The calculated Adjusted Rsquare is  7.01 %
```

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?

Answer: Adjusted R-squared increases only when the predictor is significant and affects response. In another words, Adjusted R-squared is used to determine how reliable the correlation is and how much is determined by the addition of independent variables. Adjusted Rsquare will always be less than or equal to Rsquare. The full model Adjusted Rsquare is 7.01%. No learner status adjusted Rsquare is 7.23% , which is overfitting. the lrn variable should be removed from the model first.

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.

Answer: We observe that the least temperature created more damages to 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.

Answer: we have a negative slope (-0.2162) which means from inception , the shuttle misison carried out a temperature likely to cause damage in the O-rings. The intercept (11.6630) means the shuttle mission had high number of damaged O-rings from ground zero.

	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

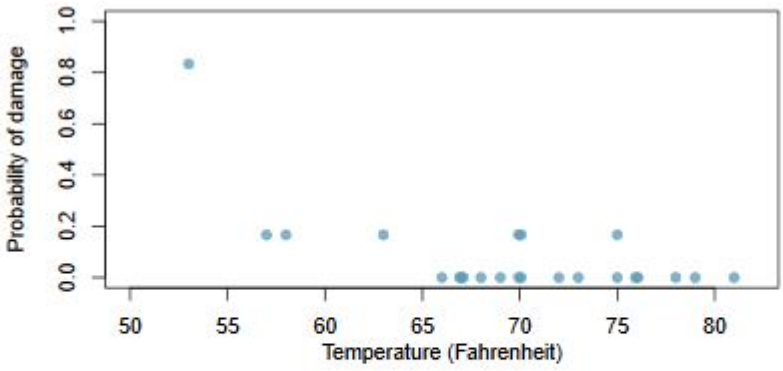
(c) Write out the logistic model using the point estimates of the model parameters.

Answer: ln(P/(1-P)) = ax + b , where P is probability of success, 1-P is probability of failure, b is the intercept, b is the intercept. ln(P/(1-P)) = -0.2162 * x + 11.6630

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

Answer: Yes, the concern regarding 0-rings are justified. The trend observed in the data showed a highly corelation between temperature and damaged O-rings.

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 t 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 the probability that an O-ring will become damaged at each of the following ambient temperature and 55 degrees Fahrenheit. The model-estimated probabilities for several additional ambient tem are provided below, where subscripts indicate the temperature:

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

(b) Add the model-estimated probabilities from part~(a) on the plot, then connect these dot smooth curve to represent the model-estimated probabilities.

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

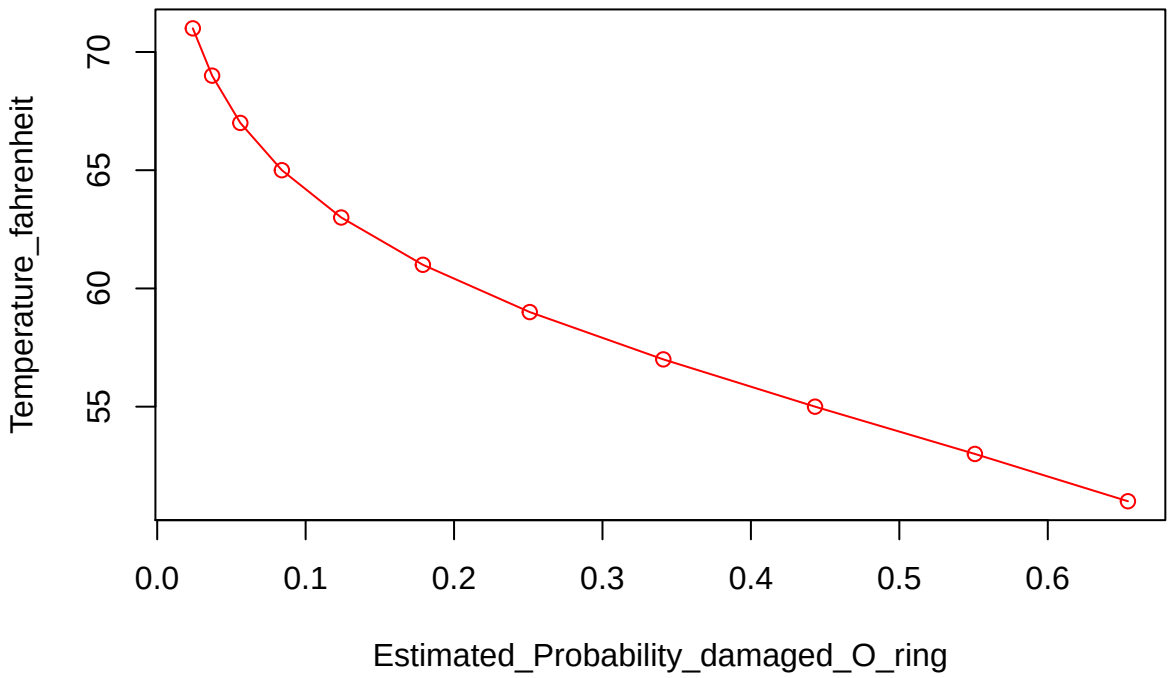
(a) Use the model to calculate the probability that an O-ring will become damaged at each of the following ambient temperatures: 51, 53,and 55 degrees Fahrenheit. The model-estimated probabilities for several additional ambient temperaturesare provided below, where subscripts indicate the temperature:

```
## The probability that an O-ring will become damaged at ambient temperatures of 51 Fahrenheit is  0.654
```

```
## The probability that an O-ring will become damaged at ambient temperatures of 53 Fahrenheit is  0.550
```

```
## The probability that an O-ring will become damaged at ambient temperatures of 55 Fahrenheit is  0.443
```

(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.


Temperature_fahrenheit
55
60
65
70
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Estimated_Probability_damaged_O_ring

Answer: