Homework 9 : Introduction to Linear Regression

Ramnivas Singh

05/02/2021

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?

Answer:

(a) birth weight= 123.05 - 8.94*smoke

(b) If the mother is a smoker, smoke= 1, and if the mother is not, smoke = 0. Smoker mothers have babies with birth weight 8.94 oz less than non-smokers.

(c) H0: Birthweight= 0 There is no association with birth weights between smoker moms and non-smoker moms HA: Birthweight=/= 0 There is association with birth weights between smoker moms and non-smoker moms

Since this p value is small than 0.05 we can reject the null hypothesis. So there is a significant relationship between the average birth weight and smoking. We can concluded that smoking does affect the lowered birth weights.

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.

Answer :

(a) The average number of days absent= 18.93 - (9.11eth) + (3.10sex) + (2.15*lrn)

(b) eth: the average number of days absentee of non-aboriginal students is 9.11 lower than aboriginal students. sex: the average number of days absentee of male students is 3.10 higher than female students. lrn: the average number of days absentee of slow learners is 2.15 higer than average learners.

(c)

```
eth<-0
sex<-1
lrn<-1
total<-18.93 - (9.11*eth) + (3.10*sex) + (2.15*lrn)
2-total
```

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

(d) R^2 given below

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

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

Adjusted R^2 is given below

```
n<-146
k<-3
total1<-(240.57/264.17)*(n-1)/(n-k-1)
R_adj<-1-total1
R_adj
```

```
## [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?

Answer : Learner status to remove, because this has the highest Adjusted R^2 value (0.07230).

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.

Answer:

(a) From the table we can see

Higher number of damaged O-rings are observed at lower temperature Less number of damaged O-rings are observed at higher temperature

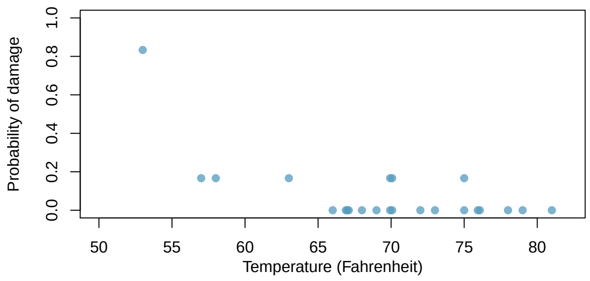
(b) This model is represented by two components:

Intercept Temperature The estimate parameter estimate for the model. The z-value and the p-value help with distinguishing important information and shows how good the variables predict this model.

(c) Logistic model:

(d) yes, based on the model we can say that concerns regarding O-rings are justified. There are more damaged O-rings at lower temperatures and O-rings are significant components.

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:

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

(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.

Answer

(a)

```
p <- function(temp)
{
  damaged_Oring <- 11.6630 - 0.2162 * temp

  pi <- exp(damaged_Oring) / (1 + exp(damaged_Oring))

  return (round(pi,5))
}

p51 <- p(51)
paste0("p51: ",p51)
```

```
## [1] "p51: 0.65403"
```

```
p53 <- p(53)
paste0("p53: ",p53)
```

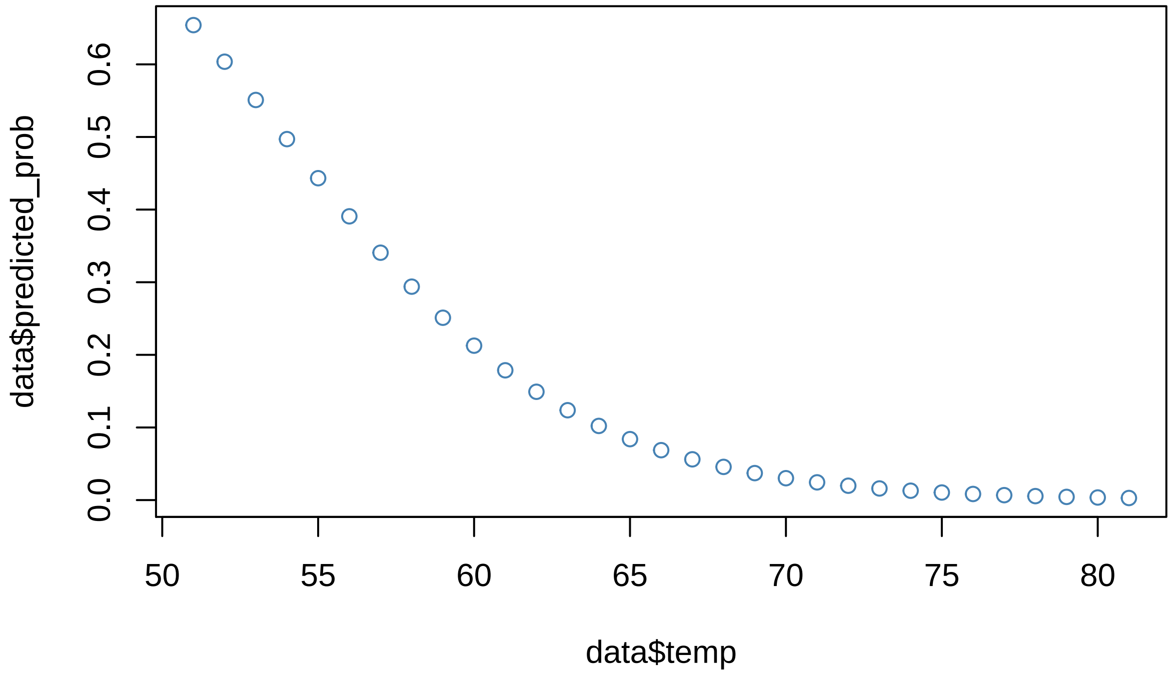
```
## [1] "p53: 0.55092"
```

```
p55 <- p(55)
paste0("p55: ",p55)
```

```
## [1] "p55: 0.44325"
```

(b)

```
  temp <- seq(from = 51, to = 81, by = 1)
  predicted_prob <- exp(11.6630-(0.2162*temp))/(1+exp(11.6630-(0.2162*temp)))
  data <- as.data.frame(cbind(temp, predicted_prob))
  graphics::plot(data$temp, data$predicted_prob,col = "steelblue")
```



(c) Key conditions for fitting in Logistic Regression model:

We can tell each observation is independent. Also the sample size came from only 23 missions which is relatively a small sample size. Aside from the temperature, there might be other variables that could've contributed to the damage of the O-rings. To improve in our future models, we can should include the other variables.