Homework\_8

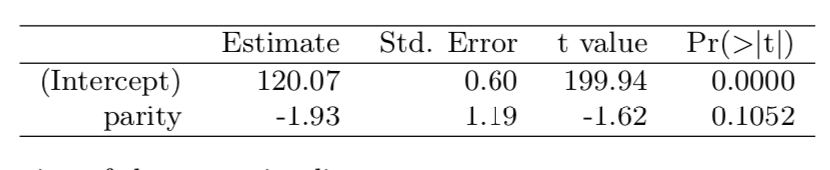
Priya Shaji

4/25/2019

## Multiple and Logistic Regression

### Q\_8.2

Baby weights, Part II. Exercise 8.1 introduces a data set on birth weight of babies. Another variable we consider is parity, which is 0 if the child is the first born, and 1 otherwise. The summary table below shows the results of a linear regression model for predicting the average birth weight of babies, measured in ounces, from parity.



1. Write the equation of the regression line.

Ans a)

Equation is y = b0 + b1x Weight = 120.07 - 1.93 \* Parity

1. Interpret the slope in this context, and calculate the predicted birth weight of first borns and others.

Ans b)

Slope of the regression line in this context says that, for every child who is not a first born will predicted to be have 1.93 less than birth weight of the first child. So the firstborn weighs 120.07 ounces and every other child afterwords would be 120 - 1.93 \* 1 = 118.07

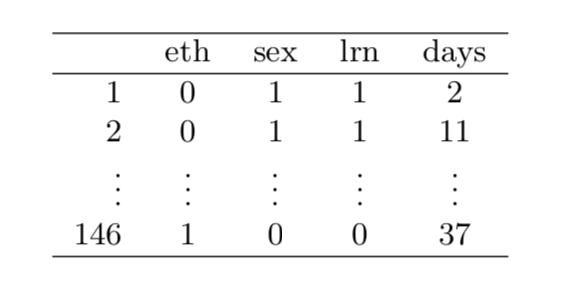
1. Is there a statistically significant relationship between the average birth weight and parity?

Ans c)

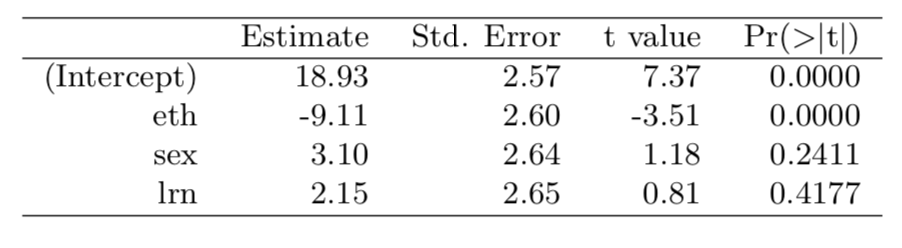
The p value of the parity is 0.1052 which is greater than 0.05. So there is no statistically significant relationship between avarage birth weight and parity.

### Q\_8.4

Absenteeism. Researchers interested in the relationship between absenteeism from school and certain demographic characteristics of children collected data from 146 randomly sampled stu- dents in rural New South Wales, Australia, in a particular school year. Below are three observations from this data set.



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).



1. Write the equation of the regression line.

Ans a)

Equation is: Absenteeism = 18.93 - 9.11 \* (ethnic background) + 3.10 \* (sex) + 2.15 \* (learner status)

1. Interpret each one of the slopes in this context.

Ans b)

Ethnic background: The model predicts a 9.11 (absent) days decrease in non-aboriginal children, all else held constant.

Sex: The model predicts a 3.10 (absent) days increase in males over females, all else held constant.

Learner status: The model predicts a 2.15 (absent) days increase in slow learners over average learners, all else held constant.

1. 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.

Ans c)

Absent <- 18.93 - 9.11 \* (0) + 3.10 \* (1) + 2.15 \* (1)  
Residual <- 2 - Absent  
paste("Residual for this student: ", Residual)

## [1] "Residual for this student: -22.18"

1. 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 R2 and the adjusted R2. Note that there are 146 observations in the data set.

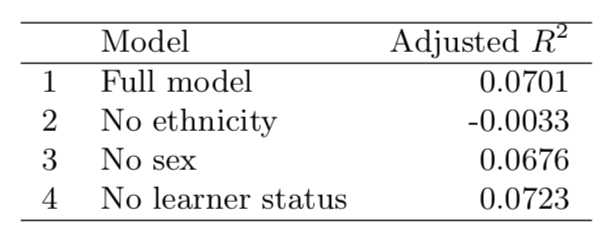
Ans d)

## [1] "R-squared: 0.0893"

## [1] "R-squared adjusted: 0.0701"

### Q\_8.8

Absenteeism, Part II. Exercise 8.4 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.



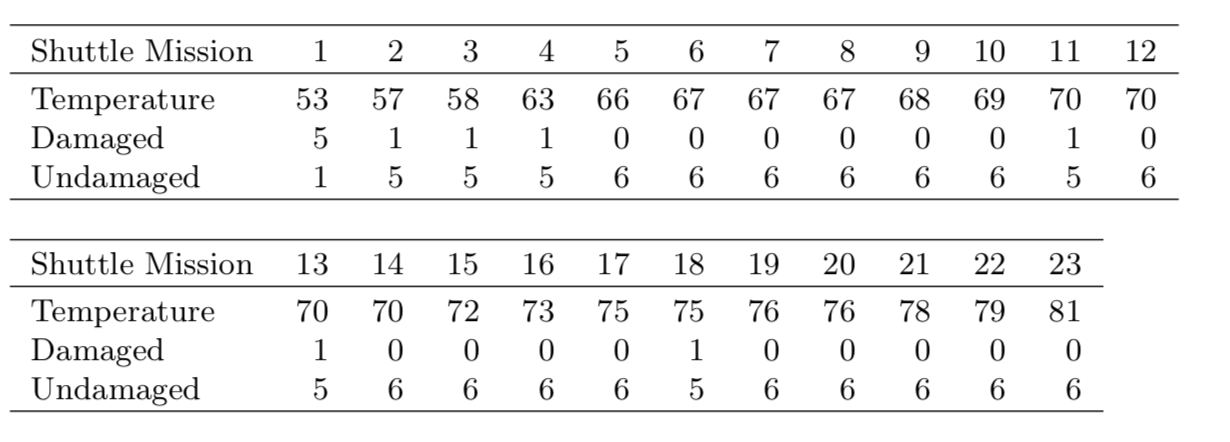
Which, if any, variable should be removed from the model first?

Ans 8.8)

The learner status should be removed from the model first, since we get a better adjusted R2

### Q\_8.16

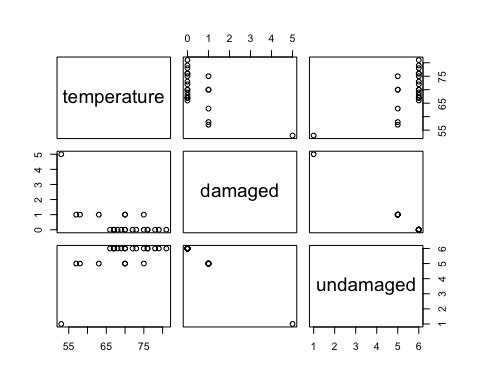
Challenger disaster, Part I. On January 28, 1986, a routine launch was anticipated for the Challenger space shuttle. Seventy-three seconds into the flight, disaster happened: the shuttle broke apart, killing all seven crew members on board. An investigation into the cause of the disaster focused on a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



1. 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.

Ans a)

temperature <- c(53,57,58,63,66,67,67,67,68,69,70,70,70,70,72,73,75,75,76,76,78,79,81)  
  
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)  
  
ShuttleMission <- data.frame(temperature, damaged, undamaged)  
  
plot(ShuttleMission)

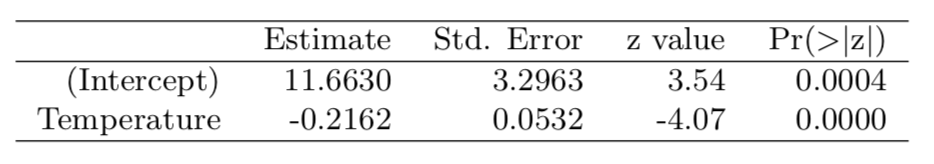


According to the above data,

Higher number of damaged O-rings are observed when lower temperatures were recorded.

Less number of damaged O-rings are observed when higher temperatures were recorded.

1. 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.



Ans b)

This model is represented by two components: the Intercept and Temperature values. The Estimate identifies the parameter estimate for the model. The Z value and the P-value help with distinguishing important information from less significant parameters by telling us how good the variables predict this model.

1. Write out the logistic model using the point estimates of the model parameters.

Ans c)

loge(pi/(1???pi))=11.6630???0.2162×Temperature

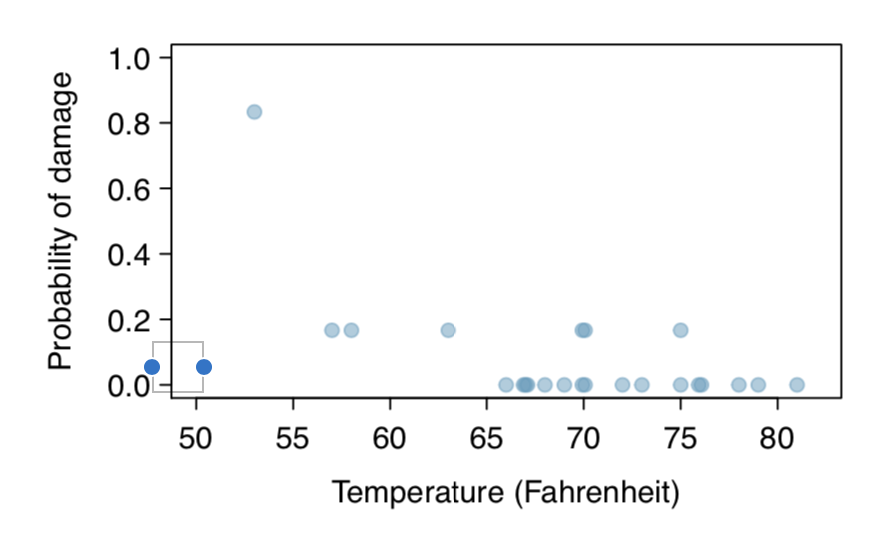
1. Based on the model, do you think concerns regarding O-rings are justified? Explain.

Ans d)

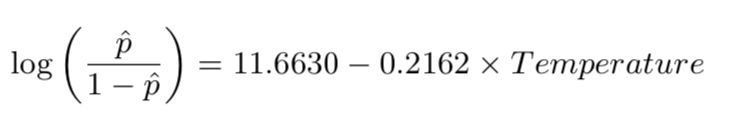
According to the data, we can deduct a high probability of damage to O-rings under 50 degree.

### Q\_8.18

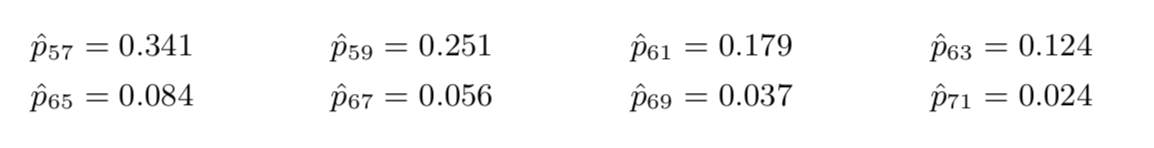
Challenger disaster, Part II. Exercise 8.16 introduced us to O-rings that were identified as a plausible explanation for the breakup of the Challenger space shuttle 73 seconds into takeoff in 1986. The investigation found that the ambient temperature at the time of the shuttle launch was closely related 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.



1. The data provided in the previous exercise are shown in the plot. The logistic model fit to these data may be written as



where pˆ 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:



Ans a)

p <- function(temp)  
{  
 damagedOring <- 11.6630 - 0.2162 \* temp  
   
 phat <- exp(damagedOring) / (1 + exp(damagedOring))  
   
 return (round(phat\*100,2))  
}

Testing final formula to provided results. p57 <- p(57)

p59 <- p(59)

p61 <- p(61)

Finding probabilities for 51, 53, and 55 temperatures. p51 <- p(51)

p53 <- p(53)

p55 <- p(55)

The probability that an O-ring will become damaged at 51 degrees Fahrenheit ambient temperatures is: 65.4%.

The probability that an O-ring will become damaged at 53 degrees Fahrenheit ambient temperatures is: 55.09%.

The probability that an O-ring will become damaged at 55 degrees Fahrenheit ambient temperatures is: 44.32%.

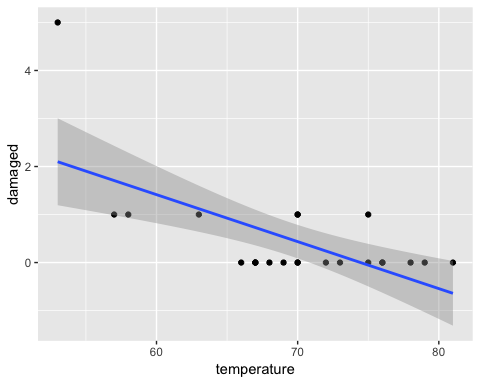
1. Add the model-estimated probabilities from part (a) on the plot, then connect these dots using a smooth curve to represent the model-estimated probabilities.

Ans b)

library(ggplot2)

ggplot(ShuttleMission,aes(x=temperature,y=damaged)) + geom\_point() +   
 stat\_smooth(method = 'glm', family = 'binomial')

## Warning: Ignoring unknown parameters: family



1. 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.

Ans c)

Logistic regression conditions: There are two key conditions for fitting a logistic regression model:

Each predictor x\_i is linearly related to logit(p\_i ) if all other predictors are held constant.

Each outcome Y\_i is independent of the other outcomes.

Based on that definition, we have assumed that those conditions are met.