HW 8 - Multiple Regression

Shyam BV

December 4, 2016

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.

```
babies <- read.csv("https://raw.githubusercontent.com/jbryer/DATA606Fall2016/master/Data/Data%20from%20
```

(a) Write the equation of the regression line.

```
babies_lm <- lm(weight ~ parity ,babies)

summary(babies_lm)
```

```
##
## Call:
## lm(formula = weight ~ parity, data = babies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.739 -14.739  -2.739  10.261 120.261
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 129.7390     0.7008  185.12  < 2e-16 ***
## parity       -4.2953     1.3766   -3.12  0.00185 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.9 on 1198 degrees of freedom
##   (36 observations deleted due to missingness)
## Multiple R-squared:  0.008061,   Adjusted R-squared:  0.007233
## F-statistic: 9.735 on 1 and 1198 DF,  p-value: 0.001851
```

$$\widehat{weight} = \hat{\beta}_0 + \hat{\beta}_1 \times parity$$

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

Minimum weight of the babies is 129.73.

```
#firstborn

parity =0

129.73-4.2953*0
```

```
## [1] 129.73
```

```
#otherwise
129.73-4.2953*1
```

```
## [1] 125.4347
```

First born babies are having higher weight than other babies. The estimated weight of other born babies is 4.29 lower than first born babies.

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

From the numbers we calculated, the p-value is very less. Hence we reject null hypothesis and conclude that there is an association between parity and birth weight.

8.4 Absenteeism.

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 SouthWales, Australia, in a particular school year. Below are three observations from this data set.

(a) Write the equation of the regression line.

days = 18.93 -9.11 * eth + 3.10 * sex + 2.15 * lrn

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

1. The model predicts an 9.11 decrease in absenteeis if the ethnicity is not aboriginal.
2. Model predicts an increase in 3.10 days if the the child is male.
3. Model predicts an increase in 2.15 days if the child is slow learner.

(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.

Actual days missed = 2

$e_1 = y_1$ - $\hat{y_1}$

```
y_1_hat <- 18.93-9.11*0+3.10*1+2.15*1


residual = 2 - y_1_hat
```

Our model predicts the absenteeis is around 24.18. So the residual 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-squared and the adjusted R-squared. Note that there are 146 observations in the data set.

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

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

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?

The p-value of `learners status` is 0.4177. So if we remove it from the model, we can get better adjusted R-squared.

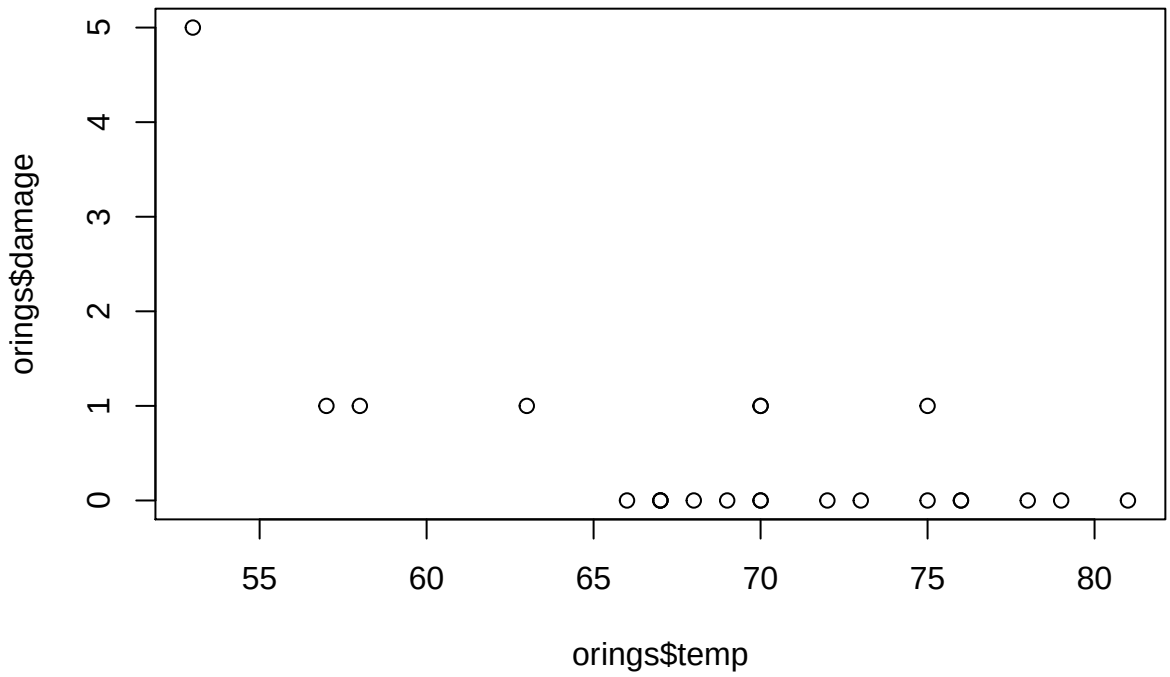
8.16 Challenger disaster,

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.

```
orings <- read.delim("https://raw.githubusercontent.com/jbryer/DATA606Fall2016/master/Data/Data%20from%
```

(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.

```
plot(orings$damage~orings$temp)
```




```
origns_lm <- lm(damage ~ temp,orings)
summary(origns_lm)
```

```
##
## Call:
```

```
## lm(formula = damage ~ temp, data = orings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8271 -0.6207 -0.1421  0.3961  2.9007
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.28571    1.79956   4.049 0.000578 ***
## temp        -0.09786    0.02574  -3.801 0.001043 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8521 on 21 degrees of freedom
## Multiple R-squared:  0.4076, Adjusted R-squared:  0.3794
## F-statistic: 14.45 on 1 and 21 DF,  p-value: 0.001043
```

Plot shows that there are one outlier. Other values are in a specific range.

The damage to the o-rings happens when there is an decrease in -0.09786 temperature.

(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.

Intercept: The number of damaged o-rings when the temperature is 0.

Slope: The damage to the o-rings happens when there is an decrease in -0.2162 temperature.

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

From the above dataset below is the logistic model

$log(\hat{p}/(1\hat{-}p)) = 7.28571$ -0.09786 * temperature

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

Yes. There were an outlier on the normal model. But when the log is performed, the outliers are reduced.

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.

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

$log(\hat{p}/(1\hat{-}p)) = 11.6630$ -0.2162 * temperature

```
temperature = 55

p1=exp( 11.6630 -0.2162 * temperature)

p=p1/(1+p1)
p
```

```
## [1] 0.4432456
```

$\hat{p51} = 0.6540297$

$\hat{p53} = 0.5509228$

$\hat{p55} = 0.4432456$

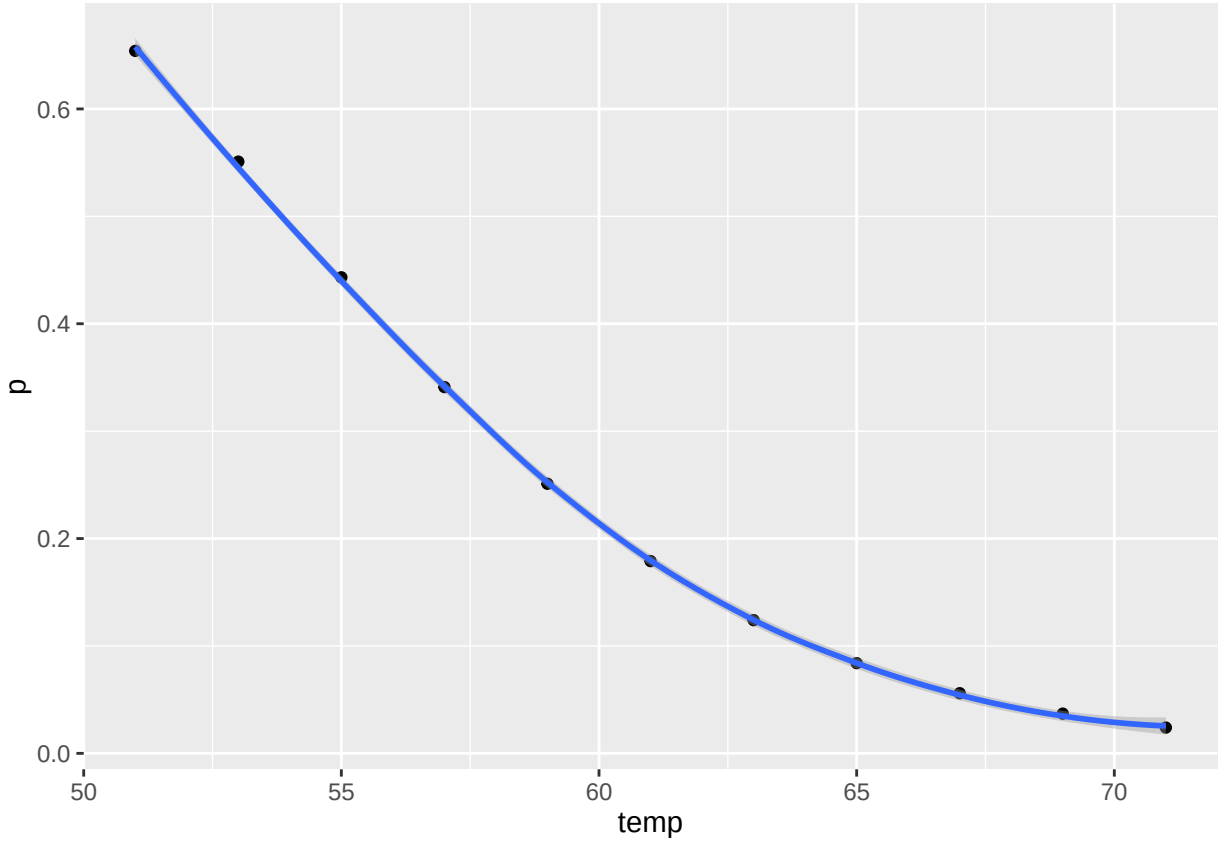
(b) Add the model-estimated probabilities from part (a) on the plot, then connect these dots using a smooth curve to represent the model-estimated probabilities.

```
p <- c(0.6540297,0.5509228,0.4432456,0.341,0.251,0.179,0.124,0.084,0.056,0.037,0.024)

temp <- c(51,53, 55,57, 59,61, 63, 65, 67,69, 71)

orings_prob <- data.frame(p,temp)


ggplot(orings_prob,aes(temp ,p)) + geom_point() + geom_smooth(method="loess")
```



(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.

Below are the conditions to apply lgistic regression model

1. Each predictor x_i is linearly related to logit(pi) if all other predictors are held constant.
2. Each outcome Y_i is independent of the other outcomes.