Data 606 - HW8

Dhananjay Kumar

December 10, 2016

Question 8.2, Baby Weights, Part II

Exercise 8.1 indtroduces a data set on birth weight and babies. Another variable we consider is parity, which is 0 if a 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.

a. Write the equation of the regression line:

birthweight = 120.07 - 1.93 × parity

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

The slope in this context is the amount of ounces less a baby weighs if it is first born than if it were not first born. It is a straight line between the mean of the birthweights at x = 0 (if the child is first born) and the mean of the birthweights at x = 1 (if the child is not).

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

It looks like there is not. The p-value for parity is above 0.5, at 0.1052.

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 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 regressionmodel for predicting the average number of days absent based on ethnic background, sex, and learner status.

a. Write the equation of the regression line:

absent days = 18.93 - 9.11 × eth + 3.10 × sex + 2.15 × lrn

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

With multiple variables the slope is similar to the single variable case, however, the slope now refers to the change in number of days absent due to a certain variable, given that all other variables are held constant, or controled for.

c. Calculate the residual for the first observation in the data set: a student who is aboriginal, male, a slow learner, and missed two days of school.

Lets calculate what our model would expect for these variables:

```r
library(ggplot2)
eth <- 0
sex <- 1
lrn <- 1

predictdays <- 18.93 - 9.11*eth + 3.1*sex + 2.15*lrn
days <- 2

resid <- days - predictdays

resid
```

```
## [1] -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 R^2 and the adjusted R^2. Note that there are 146 observations in the data set.

```r
varresid <- 240.57
varabs <- 264.17
n <- 146
k <- 3

R2 <- 1-(varresid/varabs)
R2a <- R2*((n-1)/(n-k-1))

R2
```

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

```r
R2a
```

```
## [1] 0.0912238
```

8.8, Absenteeism, Part II

Exercise 8.4 considers a model that predicts the number of days absent using three predictors: eth, sex, and lrn. The table below shows the adjusted R-squared for the model as well as the 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?

It looks like learner status should be removed from the model first, since we get a better adjusted R^2

8.16, Challenger Disaster, Part I

a. Each column of theh table above represents a different shutttle mission. Examine these data and describe what you observe with respect to the relationship between temperature and damaged O-rings

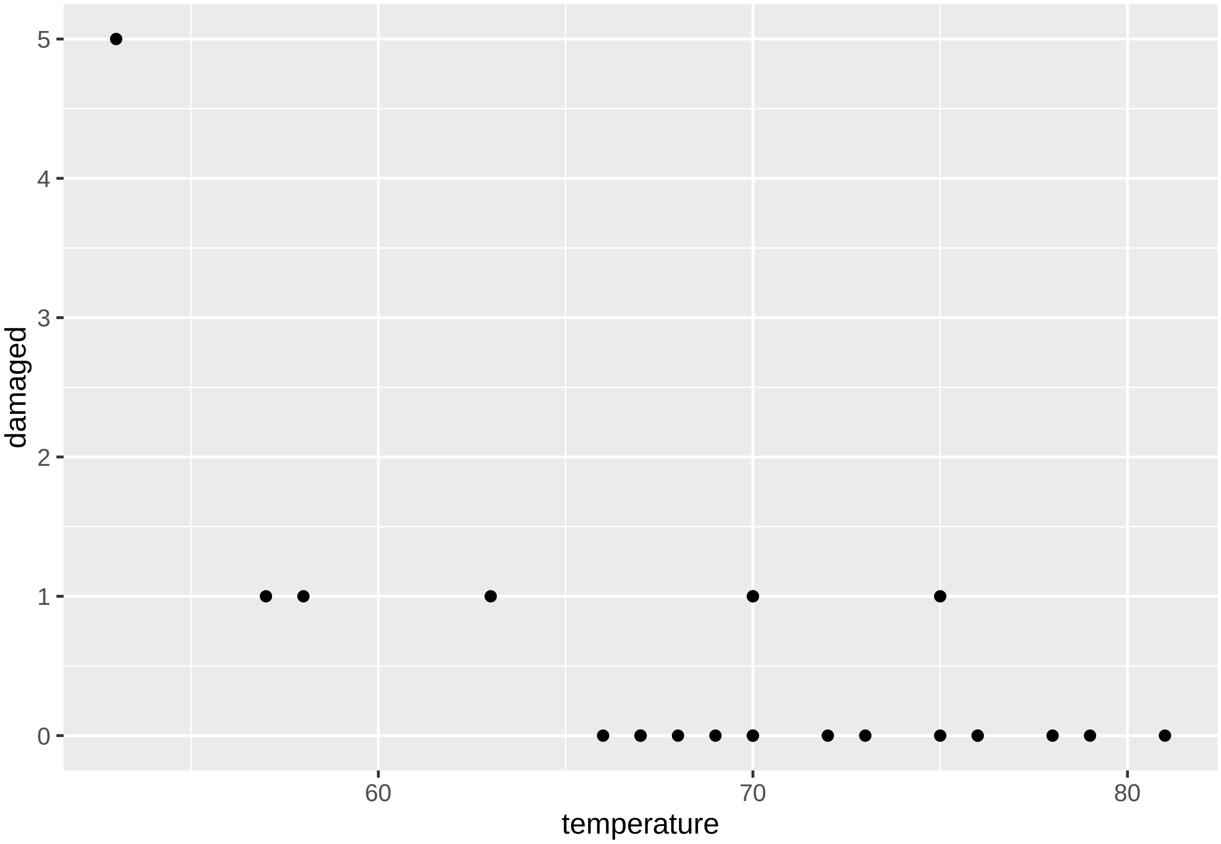
```r
temperature <- c(53,57,58,63,66,67,67,67,68,69,70,70,70,70,72,73,75,75,76,76,
                 78,79,81)

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)

data <- data.frame(temperature = temperature, damaged = damaged,
                   undamaged = undamaged)

ggplot(data,aes(x=temperature,y=damaged)) + geom_point()
```



It does seem like low temperatutures are more likely to cause damaged O-Rings. Though there are some high temperatures that result in damage, and overall there is more undamaged O-rings than damaged O-Rings, it looks like below a certain temperature there are no un-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.

Here, we have a slightly different equation than a normal linear regression. We have the form:

$$log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_1 + ...$$

Our goal is to find out P_i, but otherwise, the intercepts and parameters mean the same thing as in linear regression. Similarly for our z-value and p-tests: these tell how good the variable predicts our model.

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

using the form above:

$$log(\frac{p_i}{1-p_i}) = 11.6630 - 0.2162 \times Temperature$$

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

It looks like the concerns are valid, since we have such low P-values. Given our evidence, it would appear that low temperatures are more likely to cause damaged O-rings.

8.18, Challenger disaster, Part II

a. The data provided in the previous exercise are shown in the plot, with the logistic model given above. 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.

```
challengermodel <- function(temp){
  return(exp(11.663 - 0.2162*temp)/(1 + exp(11.663 - 0.2162*temp)))
}

challengermodel(51)
```

```
## [1] 0.6540297
```

```
challengermodel(53)
```

```
## [1] 0.5509228
```

```
challengermodel(55)
```

```
## [1] 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.

To more easily do this with R, I used the stat smooth function instead of manally plotting out the datapoints

In making this stat-smooth, I ended up disaggregating the 5 damaged o-rings at 53 degrees, and added that as 5 different datapoints for a single damaged o-ring at 53 degrees (this is how I assume a logistic regression was arrived at when we had 5 damaged o-rings):

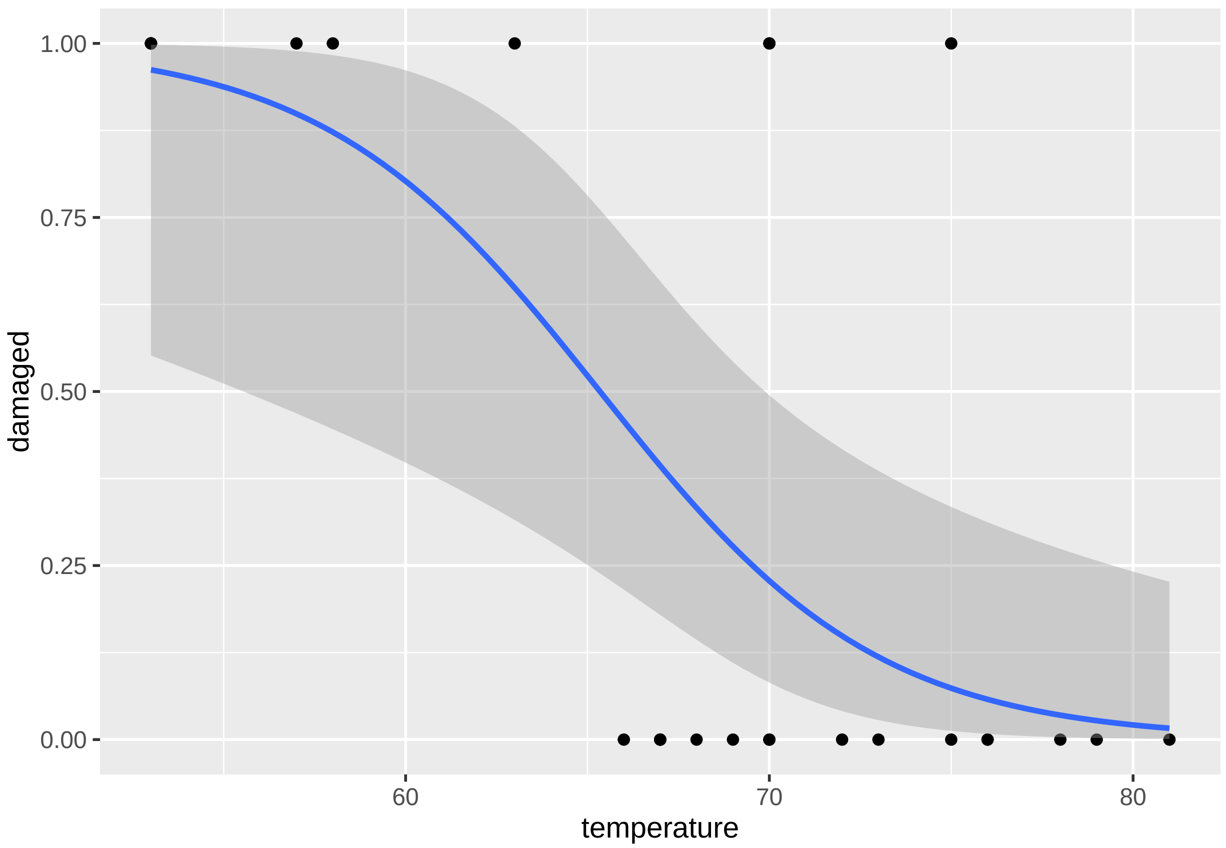
```
temperature <- c(53,53,53,53,53,57,58,63,66,67,67,67,68,69,70,70,70,70,72,73,
                 75,75,76,76,78,79,81)

damaged <- c(1,1,1,1,1,1,1,1,0,0,0,0,0,0,1,0,1,0,0,0,0,1,0,0,0,0,0)

undamaged <- c(0,0,0,0,0,5,5,5,6,6,6,6,6,6,5,6,5,6,6,6,6,5,6,6,6,6,6)

data <- data.frame(temperature = temperature, damaged = damaged,
                   undamaged = undamaged)

ggplot(data,aes(x=temperature,y=damaged)) + geom_point() + stat_smooth(method = 'glm', method.args = li
```



I would be slightly concerned using this model to predict O-ring failures, and would like to find more variables. There are a few high temperatures where failures occurred. This would be a good way to come up with a "threshold". For example, we could make sure no launches happen when the weather is below 65 degrees, when there was guarenteed to be at least one o-ring failure.

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.

I noted my concerns above. Logistic regression tends to find a "boundary" for flipping between categories, and works best in situations where those boundaries are clearly defined. If there are more confounding variables, those should be discovered and added to the model before trusting it. In this case, it seems obvious there are other variables causing o-ring failures at higher temperatures, and if anything this model should be a signal that more research is needed.