IS606 - Homework 8

Daniel Dittenhafer

December 3, 2015

8.2 Baby Weights, Part II (p395)

a) Write the equation of the regression line.

$$y = 120.07 - 1.93x_{parity}$$

b) Interpret the slope in the context, and calculate the predicted birth weight of first borns and others. The slope 120.07 indicates the first born (parity = 0) would be predicted to weigh 120.07 ounces. The others born, based on the slope of -1.93 would be 118.14 oz.

c) Is there a statistically significant relationship between the average birth weight and parity? Given the p-value of 0.1052 for the parity parameter, I conclude there is not a statistically significant relationship between average birth weight and parity.

8.4 Absenteeism (p397)

a) Write the equation of the regression line.

$$y = 18.93 - 9.11x_{eth} + 3.10x_{sex} + 2.15x_{lrn}$$

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

- The slope of `eth` indicates that, all else being equal, there is a 9.11 day reduction in the predicted absenteeism when the subject is no aboriginal.
- The slope of `sex` indicates that, all else being equal, there is a 3.10 day increase in the predicted absenteeism when the subject is male.
- The slope of `lrn` indicates that, all else being equal, there is a 2.15 day increase in the predicted absenteeism when the subject is a 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. Using the R code below, we compute the predicted absentee days and further compute the residual against the actual missed days of school.

```r
eth <- 0
sex <- 1
lrn <- 1
actualDaysMissed <- 2

absDaysPred <- 18.93 - 9.11 * eth + 3.1 * sex + 2.15 * lrn
absDaysPred
```

```
## [1] 24.18
```

```
residual <- actualDaysMissed - absDaysPred
residual
```

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

The residual is -22.18 days.

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 adjusted R^2. Note that there are 146 observations in the data set. The following R code computes the R^2 and adjusted R^2:

```
n <- 146
k <- 3
varRes <- 240.57
varOut <- 264.17

R2 <- 1 - (varRes / varOut)
R2
```

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

```
adjustedR2 <- 1 - (1 - R2) * ( (n-1) / (n-k-1) )
adjustedR2
```

```
## [1] 0.07009704
```

8.8 Absenteeism, Part II (p399)

Based on the adjusted R^2=0.0723, the learner status variable, `lrn`, should be removed first.

8.16 Challenger disaster, Part I (p403)

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. Observationally, I see that damaged o-rings are infrequent when temperature is greater than or equal to 66^o. On the other hand, 63^o and below, there is consistent damage to one or more 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. ... Describe the key components of this summary table in words. The model summary is given below:

```
##              Estimate StdError zValue PrAbsZ
## Intercept     11.6630   3.2963   3.54  4e-04
## Temperature  -0.2162   0.0532  -4.07  0e+00
```

The key components are the Intercept and the Temperature values. The Estimate identifies the parameter estimate for the model. The Z value and the P-value aid with distinguishing significant parameters from less significant parameters.

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

$$\log_e(\frac{p_i}{1-p_i}) = 11.6630 - 0.2162x_{temp}$$

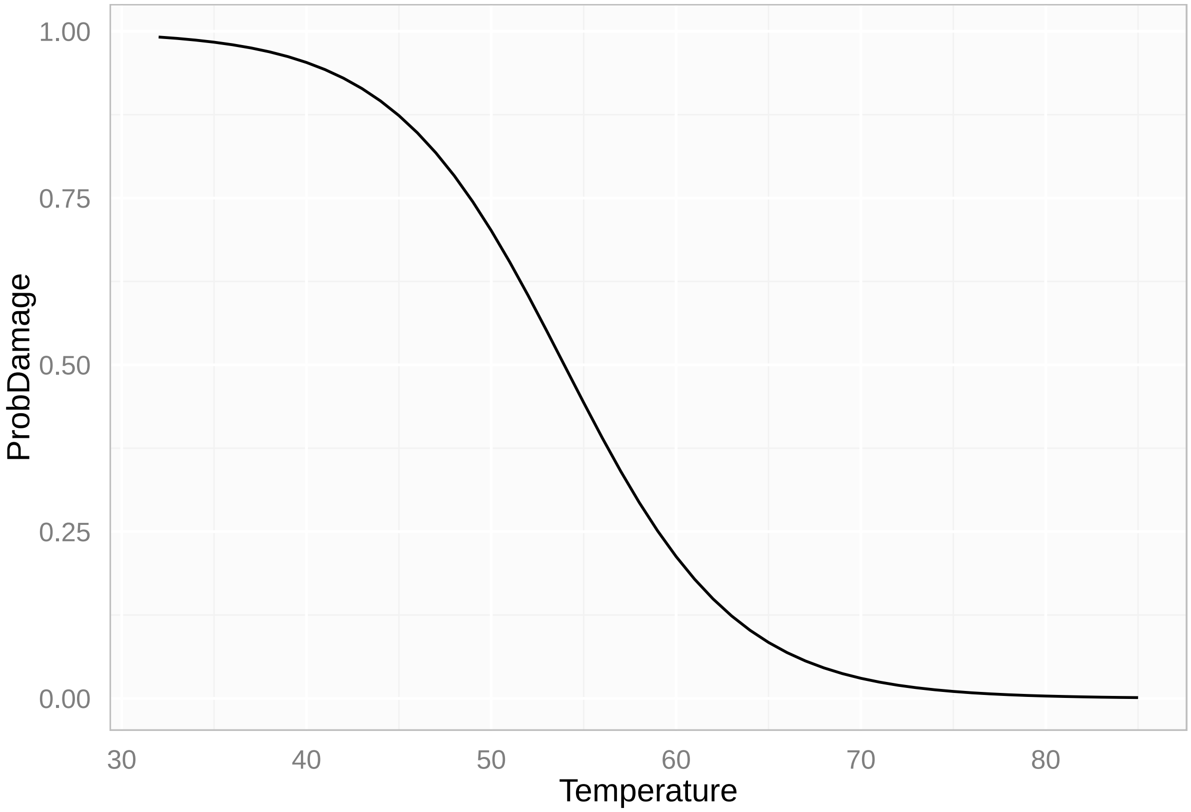
d) Based on the model, do you think concerns regarding O-rings are justified? Explain. First, define a function of the model:

```
oringModel <- function(temp)
{
  right <- 11.6630 - 0.2162 * temp

  prob <- exp(right) / (1 + exp(right))

  return (prob)
}
```

Now, lets look at the model probabilities graphically.



Given the high probabiliy of damage to O-rings under 50^o ($> 70.12\%$) according to the model and the fact that the O-rings are "Criticality 1" components, I do think concerns regarding the O-rings are justified.

"Criticality 1" : meaning that there was no backup if both the primary and secondary O-rings failed, and their failure would destroy the Orbiter and kill its crew. See https://en.wikipedia.org/wiki/Space_Shuttle_Challenger_disaster.

8.18 Challenger disaster, Part II (p404)

```
temps <- c(51,53,55)
dfProbDamage <- data.frame(Temperature=temps, ProbDamage=oringModel(temps))
dfProbDamage
```

a) Use the model to calculate the probability that an O-ring will become damaged at each of the following ambient temperatures: 51, 53, 55 degrees F.

```
##   Temperature ProbDamage
## 1          51  0.6540297
## 2          53  0.5509228
## 3          55  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. First we define the raw data in a data.frame.

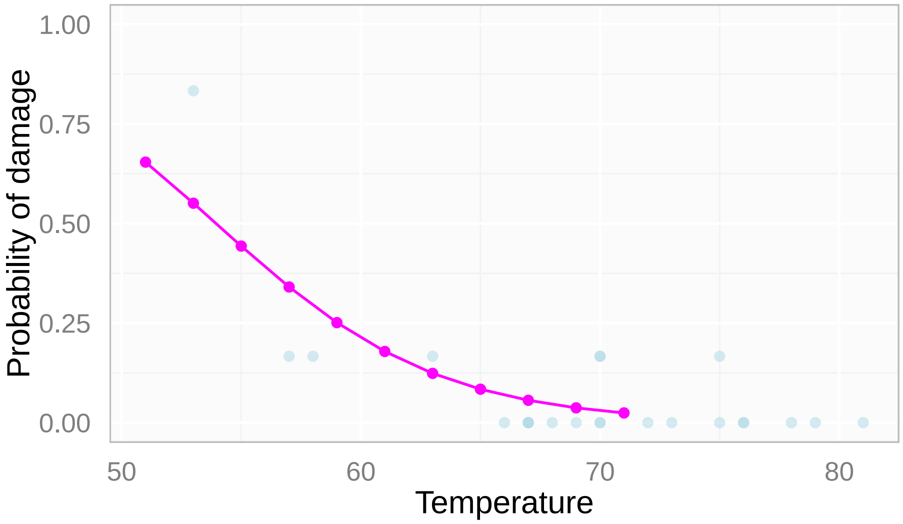
```
dfRaw <- data.frame(Missing=seq(1, 23),
                    Temp=c(53,57,58,63,66,67,67,67,68,69,70,70,70,
                           70,72,73,75,75,76,76,78,79,81),
                    Damaged=c(5,1,1,1,0,0,0,0,0,0,1,0,1,0,0,0,0,1,0,0,0,0,0),
                    Undamaged=c(1,5,5,5,6,6,6,6,6,6,5,6,5,6,6,6,6,5,6,6,6,6,6))
dfRaw$ProbDamage <- dfRaw$Damaged / (dfRaw$Damaged + dfRaw$Undamaged)
head(dfRaw)
```

```
##   Missing Temp Damaged Undamaged ProbDamage
## 1       1   53       5         1  0.8333333
## 2       2   57       1         5  0.1666667
## 3       3   58       1         5  0.1666667
## 4       4   63       1         5  0.1666667
## 5       5   66       0         6  0.0000000
## 6       6   67       0         6  0.0000000
```

Then we create a data.frame for the model-estimated probilibities using our predefined model.

```
temps <- seq(51, 71, by=2)
dfProbDamage <- data.frame(Temperature=temps, ProbDamage=oringModel(temps))
```

Finally, we show the visualization combining the raw data and the model curve.



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. As shown in the visualization above, the logistic regression is a bit slow to respond to the higher damage values. This may be due to a lack of damage data and would likely affect any model. One benefit of using logistic regression is that the probability as output is convenient and easily understood by non-statistians.

Conditions/assumptions required for logistic regression model validity include:

- Each predictor x, is linearly related to $\text{logit}(p_i)$ if all other predictors are help constant.

Based on the visualization of the raw data, the temperature data does appear to have a linear relationship wo the probability of damage to O-rings, at least to some significant degree.

- Each outcome Y_i is independent of the other outcomes.

Given the concept of the damage data being associated with separate shuttle missions after complete refurbishing of the shuttle solid rocket boosters, the independence condition is met.