Project - Inference

Camden Jones

2023-04-28

Data

The data in this report can be found at this link and in the table below: Eruptions

and

```
library(readr)
table<-matrix(c(27,39,216,228,42,21,144,117),ncol=2,byrow=TRUE)
colnames(table)<-c('Cases, n=300','Controls, n=300')
rownames(table)<-c('High income','Home was owned','Underground wiring','Father living in the home')
knitr::kable(table[1:4,1:2], format="markdown")
```

	Cases, n=300	Controls, n=300
High income	27	39
Home was owned	216	228
Underground wiring	42	21
Father living in the home	144	117

Study

Problem 1

If we wanted to know whether high income was a factor in the proportion of children with brain cancer, we could perform the hypothesis test of H0: p1 = p2 vs. H1: p1 =/ p2 (does not equal) where p1 is the proportion of cases with high income and p2 is the proportion of controls with high income. The same could be done for each of the other characteristics.

a) Perform a hypothesis test for each characteristic to determine whether the proportion of cases differs from the proportion of controls. State whether each characteristic is significant at the alpha = 0.05 level.
b) Because multiple hypothesis tests are being performed, an adjustment to the p-values is warranted. Describe the adjustment and state the updated p-values.
c) For which characteristic(s) does there appear to be evidence of a difference between the proportions of cases and controls?

Problem 2

The measurements include the duration of the eruption (Eruption), the duration of the dormant period immediately before the eruption (Dormant.Before), the duration of the dormant period immediately after the eruption (Dormant.After), and the height of the eruption (Height). All times are in minutes and the height is in meters.

a) Use R to construct a multiple regression model for predicting dormant time immediately after an eruption based on the duration of the eruption, the dormant time before eruption, and the height.

b) Determine if there are variables that do not significantly contribute to the prediction. If so, eliminate them and construct a new model.
c) Use the model from part b) to predict the duration of the dormant period immediately after an eruption if the duration of the eruption is 3.2 minutes, the duration of the dormant period before the eruption is 75 minutes, and the height of the eruption is 42 meters.

Problem 1

a) Perform a hypothesis test for each characteristic to determine whether the proportion of cases differs from the proportion of controls. State whether each characteristic is significant at the alpha = 0.05 level.

```
highincome<-c(27,39)
samples<-c(300,300)
prop.test(highincome,samples,alternative="two.sided")
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  highincome out of samples
## X-squared = 2.0599, df = 1, p-value = 0.1512
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.09330281  0.01330281
## sample estimates:
## prop 1 prop 2
##   0.09   0.13
```

```
homeowned<-c(216,228)
prop.test(homeowned,samples,alternative="two.sided")
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  homeowned out of samples
## X-squared = 1.0482, df = 1, p-value = 0.3059
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.11345514  0.03345514
## sample estimates:
## prop 1 prop 2
##   0.72   0.76
```

```
undergroundwiring<-c(42,21)
prop.test(undergroundwiring,samples,alternative="two.sided")
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  undergroundwiring out of samples
## X-squared = 7.0941, df = 1, p-value = 0.007734
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.0179296 0.1220704
## sample estimates:
```

```
## prop 1 prop 2
##   0.14   0.07
```

```
fatherinhome<-c(144,117)
prop.test(fatherinhome,samples,alternative="two.sided")
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  fatherinhome out of samples
## X-squared = 4.5841, df = 1, p-value = 0.03227
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.007657993 0.172342007
## sample estimates:
## prop 1 prop 2
##   0.48   0.39
```

The assumptions for a test for difference between two proportions are met as we have:

1. Independent random samples of children who have brain cancer and children who do not have brain cancer.

2. The populations of children who have brain cancer and who do not have brain cancer are sufficiently large: at least 20 times the size of the samples n=300.

3. Both samples contain at least 10 individuals in each category.

The results of each hypothesis test comparing the proportions of cases and controls for children with each characteristic are as follows at the alpha = 0.05 level:

1. Children in families with high income. p-value = 0.1512 : NOT significant.

2. Children in families who owned the home. p-value = 0.3059 : NOT significant.

3. Children living near underground wiring. p-value = 0.007734 : SIGNIFICANT.

4. Children whose father was living in the home. p-value = 0.03227 : SIGNIFICANT.

b) Because multiple hypothesis tests are being performed, an adjustment to the p-values is warranted. Describe the adjustment and state the updated p-values.

Because multiple hypothesis tests are being performed, we will perform a Bonferroni adjustment in which we multiply each p-value by the number of tests performed, in this case, by 4. The resultant adjusted p-values and new determinations of significance at alpha = 0.05 are as follows:

1. Children in families with high income. p-value = 0.6048 : still NOT significant.

2. Children in families who owned the home. p-value = 1.2236 : still NOT significant.

3. Children living near underground wiring. p-value = 0.030936 : still SIGNIFICANT.

4. Children whose father was living in the home. p-value = 0.12908 : NO LONGER significant.

c) For which characteristic(s) does there appear to be evidence of a difference between the proportions of cases and controls?

After the Bonferroni adjustment to the p-values, which is appropriate for this scenario given multiple hypothesis tests being performed, the only characteristic in which there appears to be evidence of a difference between the proportions of cases and controls is that of living in

neighborhoods in which the electrical wiring was underground, given the adjusted p-value of 0.030936 which is still less than 0.05. All other p-values after adjustment were greater than 0.05, indicating no difference between proportions for those characteristics. Given the nature of the data which shows twice as many cases as controls (double the proportion of the sample) for children living in neighborhoods with underground electrical wiring, this difference indicates a higher likelihood of brain cancer in children living in neighborhoods where the electrical wiring is underground.

Problem 2

```
eruptions<-read.csv("https://www.dropbox.com/s/dy0ibm9t0mjv646/OldFaithful.csv?dl=1")
attach(eruptions)
```

a) Use R to construct a multiple regression model for predicting dormant time immediately after an eruption based on the duration of the eruption, the dormant time before eruption, and the height.

```
model.dormantafter<-lm(Dormant.After~Dormant.Before+Eruption+Height)
model.dormantafter
```

```
##
## Call:
## lm(formula = Dormant.After ~ Dormant.Before + Eruption + Height)
##
## Coefficients:
##    (Intercept)  Dormant.Before        Eruption          Height
##      124.54439        -0.72660         0.54362        -0.09428
```

Shown above is the model constructed to predict dormant time immediately after an eruption based on duration, height, and dormant time immediately before eruption. The coefficients can be used to construct an equation to solve for this prediction as shown below:

Dormant.After= 124.54439 - 0.72660(Dormant.Before) + 0.54362(Eruption) - 0.09428(Height)

b) Determine if there are variables that do not significantly contribute to the prediction. If so, eliminate them and construct a new model.

To determine this, we examine a summary of the model.dormantafter.

```
summary(model.dormantafter)
```

```
##
## Call:
## lm(formula = Dormant.After ~ Dormant.Before + Eruption + Height)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.391  -8.281  -1.403   8.235  21.169
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    124.54439   10.71293  11.626  < 2e-16 ***
## Dormant.Before  -0.72660    0.18261  -3.979 0.000201 ***
## Eruption         0.54362    2.46115   0.221 0.825988
## Height          -0.09428    0.16516  -0.571 0.570395
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.976 on 56 degrees of freedom
## Multiple R-squared:  0.475,  Adjusted R-squared:  0.4469
## F-statistic: 16.89 on 3 and 56 DF,  p-value: 6.188e-08
```

This summary calculates p-values to determine if the Beta of each variable is non-zero, with p-values below 0.05 indicating that the given variable is likely to have a non-zero Beta, meaning it likely contributes to the prediction. If the p-value for a given variable is greater than 0.05, the variable does not likely have a non-zero Beta or contribute to the prediction. This summary shows that both Eruption/duration (p-value=0.825988) and Height (p-value=0.570395) have p-values greater than 0.05. They are not likely to have non-zero Betas and hence do not significantly contribute to the prediction. However, Dormant.Before has a p-value=0.000201, indicating a non-zero Beta and a contribution to the prediction. Given this information, we will construct a new model using only the Dormant.Before variable while eliminating the Eruption and Height variables.

```
newmodel.dormantafter<-lm(Dormant.After~Dormant.Before)
newmodel.dormantafter
```

```
##
## Call:
## lm(formula = Dormant.After ~ Dormant.Before)
##
## Coefficients:
##    (Intercept)  Dormant.Before
##       119.1420         -0.6833
```

```
summary(newmodel.dormantafter)
```

```
##
## Call:
## lm(formula = Dormant.After ~ Dormant.Before)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.0589  -8.3380  -0.9005   7.8995  20.8911
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    119.14197    6.86189  17.363  < 2e-16 ***
## Dormant.Before  -0.68333    0.09515  -7.181 1.45e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.843 on 58 degrees of freedom
## Multiple R-squared:  0.4707, Adjusted R-squared:  0.4615
## F-statistic: 51.57 on 1 and 58 DF,  p-value: 1.453e-09
```

This new model indicates an equation for the prediction of dormant time immediately after an eruption as follows:

Dormant.After = 119.1420 - 0.6833(Dormant.Before)

As shown by the summary, this new model has an adjusted R-squared of 0.4615, which is greater than that of the previous model's 0.4469. This increase in the adjusted R-squared

indicates that the new model does a better job of predicting the dormant time immediately following an eruption.

c) Use the model from part b) to predict the duration of the dormant period immediately after an eruption if the duration of the eruption is 3.2 minutes, the duration of the dormant period before the eruption is 75 minutes, and the height of the eruption is 42 meters.

Using the model from part b), we only need the duration of the dormant period before the eruption, which is 75 minutes. Plugging this into the equation from part b) we get:

```
119.1420-(.6833*75)
```

```
## [1] 67.8945
```

We can also use the following code to even more accurately calculate the prediction using the model's coefficients:

```
predictiondata<-c(1,75)
sum(coefficients(newmodel.dormantafter)*predictiondata)
```

```
## [1] 67.89224
```

Either way, if the duration of the dormant period before an eruption is 75 minutes, the predicted duration of the dormant period immediately after said eruption is about 67.89 minutes.