CS555, Data Analysis and Visualization Homework 6

Jefferson Parker, [japarker@bu.edu](mailto:japarker@bu.edu)

See the accompanying R code at the end of this document.

1. The data was saved to a tab-delimited text file and loaded into R. A new dichotomous variable, temp\_level, was created using the ifelse() function to categorize samples by body temperature of 98.6 (above / equal to or below).
2. The table below (Table 1) quantifies the number of subjects by temperature level (0 or 1) and sex (1 = male, 2 = female). As shown by the marginal totals, there are equal numbers of male and female subjects, but the number of female subjects in the high temperature class is proportionally higher than males with high temperature.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | Sex | | Total |
| Male = 1 | Female = 2 |
| Temp\_Level | Low = 0 | 51 (78%) | 30 (46%) | 81 |
| High = 1 | 14 (22%) | 35 (54%) | 49 |
| Total | | 65 | 65 | 130 |

Table 1: Counts of patients by sex and temperature level.

1. The risk difference (proportional difference) between high temp\_level male and female subjects is -0.32, indicating that males have a 32% lower risk of having a high temperature level.

Formally, we are interested in testing if the proportion of people with higher body temperature is different between males and females.

* + Our hypotheses and confidence levels are
    - H0: pmale = pfemale
    - H1: pmale ≠ pfemale
    -  = 0.05
  + Our test statistic is p-value generated by analytical software. Alternatively we could test against either the Z-statistic or 2.
  + Our decision rule is to fail to reject H0 if our ptest > 0.05, and to reject H0 if ptest <= 0.05.
  + Our p-value from testing is 7.2-5 (below)

2-sample test for equality of proportions without continuity correction

data: c(14, 35) out of c(65, 65)

X-squared = 14.444, df = 1, p-value = 7.218e-05

alternative hypothesis: less

5 percent confidence interval:

-1.0000000 -0.4549046

sample estimates:

prop 1 prop 2

0.2153846 0.5384615

* + Our decision is to reject H0, the proportion of people with higher temperature level is actually different between males and females.

To determine if females have a *different average body temperature* than males we can perform a one-way ANOVA, and we can use a pairwise t-test to determine if the mean body temperature of females (overall) is *higher* than males. (NOTE – the sex term needed to be treated as a factor for the execution of ANOVA tests).

Without formally stating the conditions of these two tests, we can state that the average female body temperature is different than the average male body temperature with a p-value of 3.23-5. Furthermore, pairwise t-test of mean female temperature being higher than mean male temperature is significant with p-value 1.6-5. Therefore it is affirmed that the average female body temperature is higher than the average male body temperature.

1. Using logistic regression to test the response of high temp\_level using sex as the predictor, we want to test if there is a difference in the odds of having a temperature above 98.6 between males and females.

* Our hypotheses and confidence level are:
  + H0: 1 = 0
  + H1: 1 ≠ 0
  +  = 0.05
* Our test statistic is , at our critical z = 1.96
* Our decision rule is to fail to reject H0 if |ztest| ≥ 1.96, otherwise we reject H0. Alternatively if the associated test p-value ≤ 0.05 we reject H0.
* The calculated p-value is 2.7-5
* We reject H0 and conclude that the odds of having temperature greater than 98.6 are actually different for males and females.

The odds ratio from the logistic regression model is 4.25, meaning that females are 4.25 times as likely to have a temperature above 98.6. Furthermore, we are 95% confident that the true odds ratio is between 1.97 and 9.15 times increased risk for females.

The c-statistic (ROC AUC) for this model is 0.672.

1. Applying a multiple logistic regression to predict body temperature level from sex and heart rate, without formal testing of the similar hypothesis as described in question 4, we find that both sex and heart rate are significant predictors of body temperature (p-values 0.0005 and 0.03 respectively).

|  |  |  |  |
| --- | --- | --- | --- |
|  | OR | 2.50% | 97.50% |
| sex | 4.01 | 1.84 | 8.76 |
| Heart.rate (10 beats) | 1.88 | 1.08 | 3.29 |

Table 2: Odds Ratio and 95% confidence interval values for multiple logistic regression model.

As seen in Table 2, the risk of having a high body temperature level is 4.01 times higher in women than men, and 1.88 times per 10 units of heart rate (presumably beats per minute) regardless of sex. Furthermore we are 95% confident that the true increased risk of elevated temperature is between 1.84 – 8.76 times for women to have elevated temperature. Finally, we are 95% confident that the true risk of increased body temperature level is between 1.08 and 3.29 times per 10 units of heart rate.

The c-statistic for the multiple regression model is 0.729.

1. The multiple regression model predicting body temperature level from both sex and heart rate fit the data better based on the ROC AUC measure (c-statistic) being higher (0.729 vs. 0.672).



**R Code:**

# CS555 Data Analysis and Visualization

# Homework6.R

# Jefferson Parker, japarker@bu.edu

# 20180806

# Load libraries

library(aod);

library(pROC);

# Load data

inputDir <- 'C:/Users/jparker/Code/Input';

setwd(inputDir);

tempData <- read.table(file = "heartTempBySex.txt", header = TRUE, stringsAsFactors = FALSE, sep = "\t");

# 1. We are interested in the proportion of body temperature greater than 98.6.

# Create a dichotomous body temperature variable to capture this.

tempData$temp\_level <- ifelse(tempData$temp >= 98.6, 1, 0);

# 2. Summarize the data relating body temperature level by sex.

table(tempData$temp\_level, tempData$sex, dnn = c("TempLevel", "Sex"));

# 3. Calculate the risk difference

#

# Proportion(high temp & male) - Proportion(high temp & female);

(14/65) - (35/65);

# Test if the proportion(s) of people with higher temperature (a = 0.05) is

# the same across males and females.

#

# Males is first proportion (1,2).

prop.test(x = c(14, 35), n = c(65, 65), alternative = 'less', conf.level = 0.05, correct = FALSE);

# Do females have higher body temperatures than males?

# This is asking about the actual body temperatures of our sample, not the proportions, so we need

# ANOVA.

summary(aov(tempData$temp ~ as.factor(tempData$sex)));

pairwise.t.test(tempData$temp, as.factor(tempData$sex), p.adjust.method = 'none', alternative = 'greater');

# 4. Perform a logistic regression with sex as the only explanatory variable.

mSex <- glm(temp\_level ~ sex, data = tempData, family = 'binomial');

summary(mSex);

# Odds Ratio and 95% Confidence Interval (1 unit difference)

exp(cbind(OR = coef(mSex), confint.default(mSex)));

# What is the c-statistic of the model?

# Generate the ROC curve to get AUC.

tempData$probSex <- predict(mSex, type = c("response"));

rSex <- roc(tempData$temp\_level ~ tempData$probSex);

rSex;

plot.roc(rSex, print.auc = TRUE);

grid();

# 5. Perform a multiple logistic regression predicting body temperature level from sex and heart rate.

mBoth <- glm(temp\_level ~ sex + Heart.rate, data = tempData, family = 'binomial');

summary(mBoth);

# What is the odds ratio for sex and heart rate (10 beat increase).

# CI = exp(beta +/- z(1 - alpha/2) + SE(beta) \* proportion\_difference)

rbind(

# Sex

cbind(

OR = exp(summary(mBoth)$coefficients[2, 1]),

'2.5 %' = exp((summary(mBoth)$coefficients[2, 1] - qnorm(0.975) \* summary(mBoth)$coefficients[2, 2])),

'97.5 %' = exp((summary(mBoth)$coefficients[2, 1] + qnorm(0.975) \* summary(mBoth)$coefficients[2, 2]))

),

# Heart rate

cbind(

OR = exp(summary(mBoth)$coefficients[3, 1] \* 10),

'2.5 %' = exp((summary(mBoth)$coefficients[3, 1] - qnorm(0.975) \* summary(mBoth)$coefficients[3, 2]) \* 10),

'97.5 %' = exp((summary(mBoth)$coefficients[3, 1] + qnorm(0.975) \* summary(mBoth)$coefficients[3, 2]) \* 10)

)

);

# What is the c-statistic for this model.

tempData$probBoth <- predict(mBoth, type = c("response"));

rBoth <- roc(tempData$temp\_level ~ tempData$probBoth);

rBoth;

plot.roc(rBoth, print.auc = TRUE);

grid();