HW5

Jiankun (Bob) Dong CM3226

2023-12-01

Probelm 1 and 2:

```r
dataSet <- read.csv("./A06.csv")
dataSet$temp_level <- as.numeric(dataSet$temp>=98.6)
dataSet$sex <- as.factor(dataSet$sex)
SexVTempLevel_T <- table(dataSet$temp_level,dataSet$sex)
colnames(SexVTempLevel_T) <- c('Male','Female')
SexVTempLevel_T
```

```
##
##     Male Female
##   0   51     30
##   1   14     35
```

Problem 3: 1.
$H_0 : p_1 = p_2$
$H_1 : p_1 \neq p_2$
$\alpha = 0.05$
2.
z = $\frac{\hat{p_1} - \hat{p_2}}{\hat{p}*(1-\hat{p})*(\frac{1}{n_1}+\frac{1}{n_2})}$
3.
Decision rule: reject H_0 if |z|$\geq$ 1.96
4.

```r
prop.test(c(51,30),c(65,65),conf.level = 0.95, correct = TRUE)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(51, 30) out of c(65, 65)
## X-squared = 13.102, df = 1, p-value = 0.0002951
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.1506100 0.4955439
## sample estimates:
##    prop 1    prop 2
## 0.7846154 0.4615385
```

```r
p1_hat <- 51/65
p2_hat <- 30/65
p_hat <- 81/130
(z<-(p1_hat-p2_hat)/(p_hat*(1-p_hat)*(1/65+1/65)))
```

```
## [1] 44.70899
```

5. Conclusion: reject H_0 since z is greater than 1.96. We reject the hypothesis that the proportion of people having high body temperature is the same across men and women.

Problem 4:

```r
m <- glm(dataSet$temp_level ~ dataSet$sex, family=binomial)
```

1. $H_0 : \beta_1 = 0$
   $H_1 : \beta_1 \neq 0$
   $\alpha = 0.05$

2. $$z = \frac{\beta_1}{SE\_\beta\_1}$$

3. Decision rule: reject H_0 if $|z| \geq 1.96$
4.

```r
summary(m)
```

```
##
## Call:
## glm(formula = dataSet$temp_level ~ dataSet$sex, family = binomial)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.2928     0.3017  -4.285 1.83e-05 ***
## dataSet$sex2   1.4469     0.3911   3.700 0.000216 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 172.26  on 129  degrees of freedom
## Residual deviance: 157.45  on 128  degrees of freedom
## AIC: 161.45
##
## Number of Fisher Scoring iterations: 4
```

z = 1.4469/0.3911 = 3.6995653 > 1.96 5. Reject H_0 because z>1.96. There is evidence of an association between sex and temperature level. The odds ratio for sex is 4.2499193 for change in sex. the associated

95% confidence interval is between 1.9745569 and 9.1472748
Problem 5:

```r
dataSet$male <- ifelse(dataSet$sex == 1, 1, 0)
m2 <- glm(dataSet$temp_level ~ dataSet$male+dataSet$Heart.rate,
          family=binomial)
summary(m2)
```

```
##
## Call:
## glm(formula = dataSet$temp_level ~ dataSet$male + dataSet$Heart.rate,
##     family = binomial)
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -4.56918    2.13930  -2.136 0.032693 *
## dataSet$male       -1.38919    0.39868  -3.484 0.000493 ***
## dataSet$Heart.rate  0.06337    0.02850   2.223 0.026195 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 172.26  on 129  degrees of freedom
## Residual deviance: 152.24  on 127  degrees of freedom
## AIC: 158.24
##
## Number of Fisher Scoring iterations: 4
```

```r
wald.test(b=coef(m2), Sigma=vcov(m2), Terms = 2:3)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 17.3, df = 2, P(> X2) = 0.00017
```

Odds ratio for sex: 0.2492771
Odds ratio for heart rate for 10 beat increase: 1.8845706

```r
dataSet$prob_2 <- predict(m2, type=c("response"))
par(mfrow = c(1,1))
g2 <- roc(dataSet$temp_level ~ dataSet$prob_2)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```r
auc(g2)
```

```
## Area under the curve: 0.7289
```

Problem6:
The c statistic for the first model is:

```r
dataSet$prob_1 <- predict(m, type=c("response"))
par(mfrow = c(1,1))
g1 <- roc(dataSet$temp_level ~ dataSet$prob_1)
```

```
## Setting levels: control = 0, case = 1
```

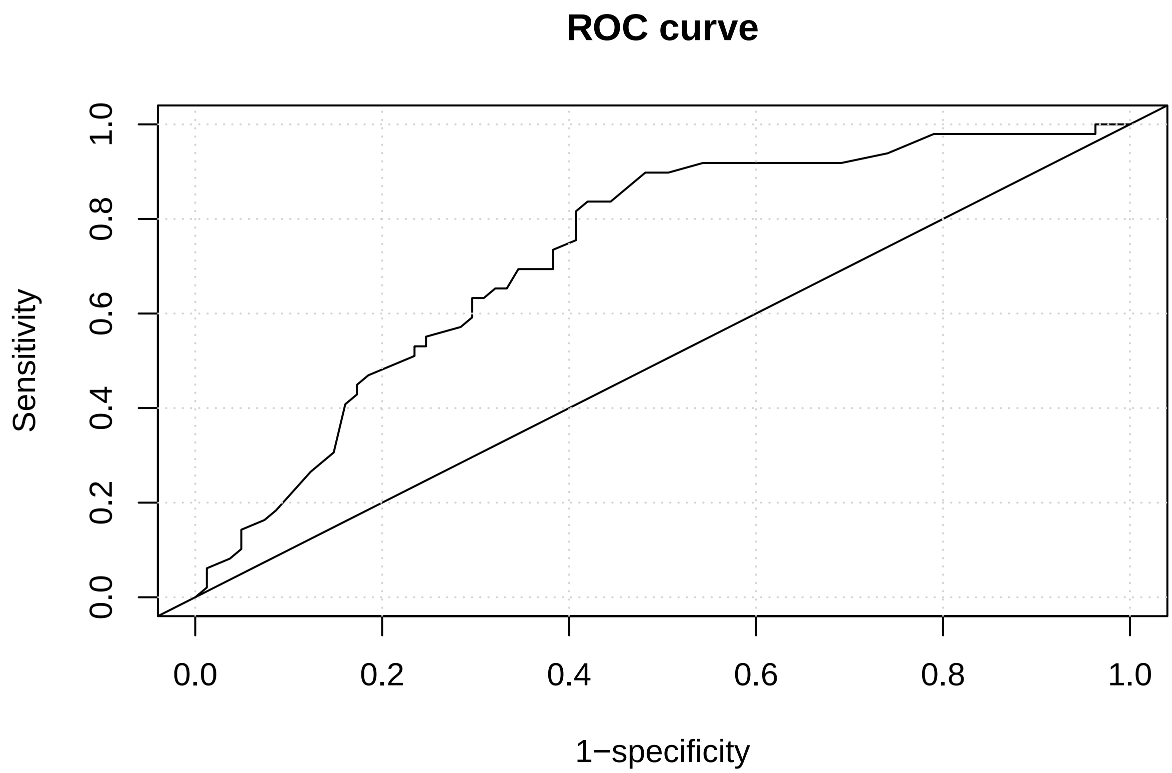
```
## Setting direction: controls < cases
```

```r
auc(g1)
```

```
## Area under the curve: 0.672
```

And that's smaller than the second model. Therefore the second model is the better one with c statistic of 0.7289

```r
plot(1-g2$specificities, g2$sensitivities,
     type="l", xlab="1-specificity",
     ylab="Sensitivity", main="ROC curve")
abline(a=0, b=1)
grid()
```


ROC curve
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
1−specificity