HW4_1830

Yue Zhang

2024-10-14

#4.4

```
exp(0.397)
```

```
## [1] 1.487356
```

For each additional unit increase in the snoring score, the odds of heart disease increase by approximately 48.7%.

#4.5 (a)

```
Shuttle = read.table("E:/Biostat/Biostatistics/PH 1830/Shuttle.dat")
colnames(Shuttle) = c("Ft", "Temp", "TD")
Shuttle = Shuttle[-1, ]
str(Shuttle)
```

```
## 'data.frame':    23 obs. of  3 variables:
##  $ Ft  : chr  "1" "2" "3" "4" ...
##  $ Temp: chr  "66" "70" "69" "68" ...
##  $ TD  : chr  "0" "1" "0" "0" ...
```

```
Shuttle$TD = as.numeric(Shuttle$TD)
Shuttle$Temp = as.numeric(Shuttle$Temp)
Shuttle$Ft = as.numeric(Shuttle$Ft)
Shuttle_fit = glm(TD ~ Temp, family = binomial(link = "logit"),
    data = Shuttle)
summary(Shuttle_fit)
```

```
##
## Call:
## glm(formula = TD ~ Temp, family = binomial(link = "logit"), data = Shuttle)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  15.0429     7.3786   2.039   0.0415 *
## Temp         -0.2322     0.1082  -2.145   0.0320 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 28.267  on 22  degrees of freedom
## Residual deviance: 20.315  on 21  degrees of freedom
## AIC: 24.315
##
## Number of Fisher Scoring iterations: 5
```

```r
1 - exp(-0.2322)
```

```
## [1] 0.2072125
```

The equation is $\text{logit}[\pi(\text{x})] = \text{-}0.2322\text{x} + 15.0429$. For every one degree of Fahrenheit increase in temperature, the odds of thermal distress decrease by approximately 20.7%. As the p-value of Temp(0.0320) is smaller than 0.05, we can state that the effect of temperature on the probability of thermal distress is statistically significant.

(b)

```r
temp_31 = data.frame(Temp = 31)
predict_shuttle = predict(Shuttle_fit, temp_31, type = "response")
predict_shuttle
```

```
##         1
## 0.9996088
```

(c)

```r
# For estimated probability = 0.5
log(0.5/(1 - 0.5))
```

```
## [1] 0
```

```r
(0 - 15.0429)/(-0.2322)
```

```
## [1] 64.78424
```

```r
-0.2322 * 0.5 * (1 - 0.5)
```

```
## [1] -0.05805
```

At 64.78 degrees of Fahrenheit the estimated probability is 0.5, and it has slope of -0.05805.

(d) As stated in part a, for every one degree of Fahrenheit increase in temperature, the odds of thermal distress decrease by approximately 20.7%.

(e)

```r
# Null hypothesis: temperature has no effect Alternative
# hypothesis: temperature has an effect

# Wald test
Z_squared = (-0.2322/0.1082)^2
Z_squared
```

```
## [1] 4.605427
```

```r
# Likelihood ratio test
Shuttle_fit_0 = glm(TD ~ 1, family = binomial(link = "logit"),
    data = Shuttle)
lrtest(Shuttle_fit, Shuttle_fit_0)
```

```
## Likelihood ratio test
##
## Model 1: TD ~ Temp
## Model 2: TD ~ 1
##   #Df  LogLik Df Chisq Pr(>Chisq)
## 1   2 -10.158
## 2   1 -14.134 -1 7.952   0.004804 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The Wald statistic Z squared is 4.6054 with p-value equal to 0.0320, the likelihood ratio test statistic is 7.952 with a p-value equal to 0.0048. Both p-values are smaller than 0.05. Thus we can reject the null hypothesis.

#4.8 (a)

```r
Crab = read_table("https://users.stat.ufl.edu/~aa/cat/data/Crabs.dat")
Crab_fit = glm(y ~ weight, family = binomial(link = "logit"),
    data = Crab)
summary(Crab_fit)
```

```
##
## Call:
## glm(formula = y ~ weight, family = binomial(link = "logit"),
##     data = Crab)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.6947     0.8802  -4.198 2.70e-05 ***
## weight        1.8151     0.3767   4.819 1.45e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 225.76  on 172  degrees of freedom
## Residual deviance: 195.74  on 171  degrees of freedom
## AIC: 199.74
##
## Number of Fisher Scoring iterations: 4
```

```r
# For x = 2.437 (i)
predict_crab_1 = predict(Crab_fit, data.frame(weight = 2.437),
    type = "response")
predict_crab_2 = predict(Crab_fit, data.frame(weight = 3.437),
    type = "response")
predict_crab_2 - predict_crab_1
```

```
##         1
## 0.2526267
```

```r
# (ii)
predict_crab_3 = predict(Crab_fit, data.frame(weight = 2.537),
    type = "response")
predict_crab_3 - predict_crab_1
```

```
##         1
## 0.0385229
```

```r
# (iii)
predict_crab_4 = predict(Crab_fit, data.frame(weight = 3.017),
    type = "response")
predict_crab_4 - predict_crab_1
```

```
##         1
## 0.1813523
```

The equation is logit[π(x)] = 1.8151x - 3.6947. For a 1-kg increase in weight, the probability of a satellite will increase by approximately 25.26%. For a 0.10-kg increase, the probability of a satellite will increase by approximately 3.85%. For a standard deviation increase, the probability of a satellite will increase by approximately 18.14%

(b)

```r
crab_mfx = logitmfx(Crab_fit, atmean = FALSE, data = Crab)
crab_mfx
```

```
## Call:
## logitmfx(formula = Crab_fit, data = Crab, atmean = FALSE)
##
## Marginal Effects:
##           dF/dx Std. Err.     z     P>|z|
## weight 0.349534  0.094088 3.715 0.0002032 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
crab_mfx$mfxest["weight", "dF/dx"] * 0.1
```

```
## [1] 0.03495343
```

For every 0.10-kg increase in weight, the probability of a satellite will increase by approximately 3.50%.

(c)

```
# Sample proportion of 1's for y variable
prop = sum(Crab$y)/nrow(Crab)

# Predict y = 1 when est. > prop
predict_crab_5 = as.numeric(fitted(Crab_fit) > prop)

# Classification table with sample proportion cutoff
xtabs(~Crab$y + predict_crab_5)
```

```
##       predict_crab_5
## Crab$y  0  1
##      0 45 17
##      1 43 68
```

```
Sensitivity = 68/(68 + 43)
Sensitivity
```

```
## [1] 0.6126126
```

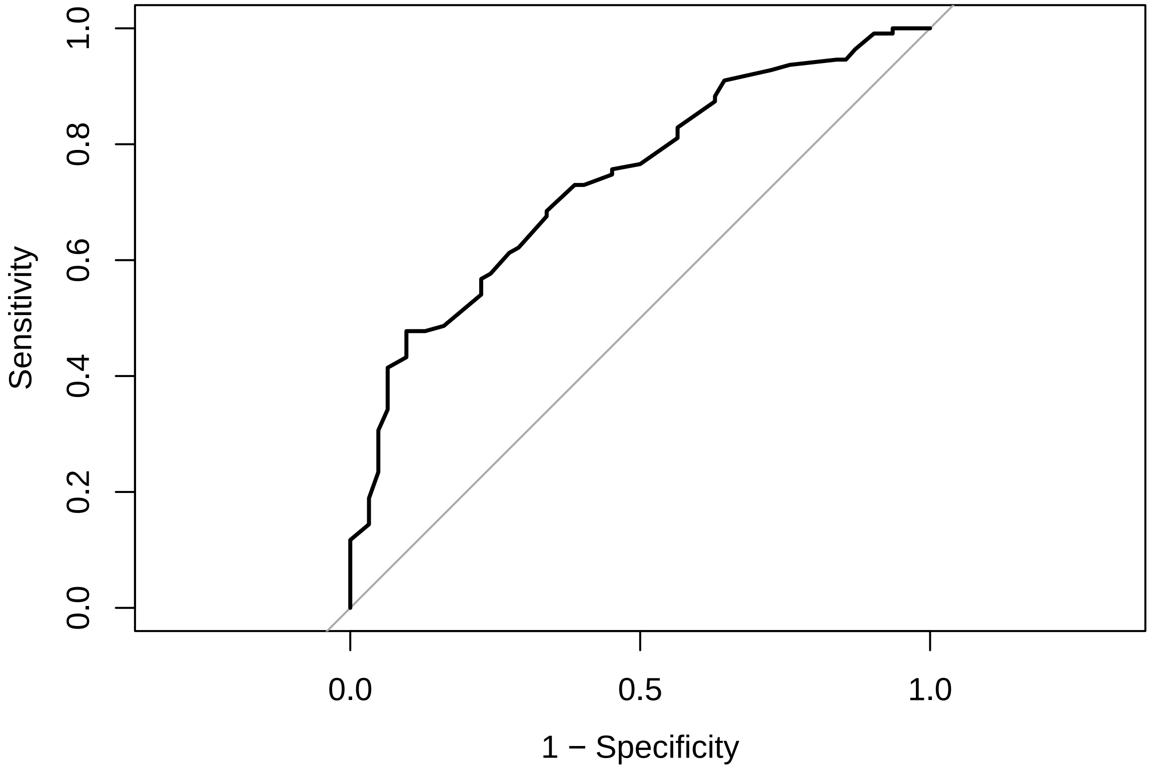
```
Specificity = 45/(45 + 17)
Specificity
```

```
## [1] 0.7258065
```

The sensitivity is 61.26% and the specificity is 72.58% which means that the model is missing around 38.74% of the actual satellites and although it correctly predicting about 72.58% of the crabs that don't have satellites, it still missclassifies around 27.43%.

(d)

```
rocplot = roc(y ~ fitted(Crab_fit), data = Crab)
plot.roc(rocplot, legacy.axes = TRUE)
```



```
auc(rocplot)
```

```
## Area under the curve: 0.7379
```

As the AUC is 0.7379 which means that there is a 73.79% chance that the model will correctly distinguish between a crab that has a satellite and one that doesn't have, we can state that the model is reasonably effective.

#4.9 (a)

```
crab_fit2 = glm(y ~ factor(color), family = binomial(link = "logit"),
    data = Crab)
summary(crab_fit2)
```

```
##
## Call:
## glm(formula = y ~ factor(color), family = binomial(link = "logit"),
##     data = Crab)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.0986     0.6667   1.648   0.0994 .
## factor(color)2  -0.1226     0.7053  -0.174   0.8620
## factor(color)3  -0.7309     0.7338  -0.996   0.3192
## factor(color)4  -1.8608     0.8087  -2.301   0.0214 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 225.76  on 172  degrees of freedom
## Residual deviance: 212.06  on 169  degrees of freedom
## AIC: 220.06
##
## Number of Fisher Scoring iterations: 4
```

```r
exp(-1.8608)
```

```
## [1] 0.1555481
```

The equation is $logit[\pi(x)] = -0.1226 * c_2 - 0.7309 * c_3 - 1.8608 * c_4 + 1.0986$. Crabs with color 4 are about 15.5% as likely to have a satellite compared to crabs with color 1.

(b)

```r
# Null hypothesis: color has no effect Alternative
# hypothesis: color has an effect
crab_fit0 = glm(y ~ 1, family = binomial(link = "logit"), data = Crab)
summary(crab_fit0)
```

```
##
## Call:
## glm(formula = y ~ 1, family = binomial(link = "logit"), data = Crab)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.5824     0.1585   3.673 0.000239 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 225.76  on 172  degrees of freedom
## Residual deviance: 225.76  on 172  degrees of freedom
## AIC: 227.76
##
## Number of Fisher Scoring iterations: 4
```

```r
lrtest(crab_fit2, crab_fit0)
```

```
## Likelihood ratio test
##
## Model 1: y ~ factor(color)
## Model 2: y ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   4 -106.03
## 2   1 -112.88 -3 13.698   0.003347 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The Chisq is 13.698 with df = 3 and p-value equal to 0.003 which is smaller than 0.05. Thus, we can reject the null hypothesis and state that color do have an effect on the probability of a satellite.

(c)

```r
crab_fit3 = glm(y ~ color, family = binomial(link = "logit"),
    data = Crab)
summary(crab_fit3)
```

```
##
## Call:
## glm(formula = y ~ color, family = binomial(link = "logit"), data = Crab)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.3635     0.5551   4.257 2.07e-05 ***
## color        -0.7147     0.2095  -3.412 0.000645 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 225.76  on 172  degrees of freedom
## Residual deviance: 213.30  on 171  degrees of freedom
## AIC: 217.3
##
## Number of Fisher Scoring iterations: 4
```

```r
# Null hypothesis: color has no effect Alternative
# hypothesis: color has an effect Wald test
Z_squared2 = (-0.7147/0.2095)^2
Z_squared2
```

```
## [1] 11.63803
```

```r
# Likelihood ratio test
lrtest(crab_fit3, crab_fit0)
```

```
## Likelihood ratio test
##
## Model 1: y ~ color
## Model 2: y ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   2 -106.65
## 2   1 -112.88 -1 12.461  0.0004156 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The equation is $logit[\pi(x)] = -0.7147 * c + 2.3635$. The z squared is 11.63 and the chisq is 12.461 with df = 1. Both p-values are less than 0.05 which means that we can reject the null hypothesis and state that color has an effect.

(d) Treating color as a quantitative variable will reduce the degrees of freedom which could increase statistical power. However, it is useful especially when the relationship is linear or monotonic. This could lead to model misspecification and poor fit.

#4.10 (a) Null hypothesis: AZT and race have no effect on the probability of developing AIDS symptoms. The deviance decreases increasingly(from 8.3499 to 1.3835) meaning that the model with AZT and race fits the data significantly better than the null model. However, the mode shows that race is not statistically significant as the p-value(0.84755) is larger than 0.10 while AZT's p-value is smaller than 0.1 if we set $\alpha = 0.1$. This means that the improvement in model fit is likely due to AZT rather than race.

(b) For this model, it set AZT = no and Race = Black as the baseline. Therefore for the indicator variable setup, aztyes = 1 for AZT = "yes", aztyes = 0 for AZT = "no"; racewhite = 1 for Race = "White" and racewhite = 0 for Race = "Black".