# Homework #5

*Zhijian Liu*

## 14.28

b. Using the groups formed in part (a), conduct a Hosmer-Lemeshow goodness of fit test for the appropriateness of the logistic function; use $\alpha = .05$. State the alternatives, decision rule, and conclusions. What is the P-value of the test?

The model obtained in part (d) of Problem 14.22:

$$log\frac{\pi}{1-\pi} = \beta_0 + \beta_1 \cdot (age) + \beta_2 \cdot (health)$$

```r
# Model & yhat
reg <- glm(formula = Y ~ X1 + X2, family = binomial(link = logit), data = df)
p.hat <- reg$fitted
# Hosmer-Lemeshow procedure
hosmerlem<-function(y, yhat, g = 10){
  # p.hat -> (quantile, quantile)
  cutyhat <- cut(yhat, breaks = quantile(yhat, probs = seq(0, 1, 1/g)),
                 include.lowest = T) # 0.1384917 -> (0.0998,0.149]
  # within each group, count number of success and failure
  obs <- xtabs(cbind(1 - y, y) ~ cutyhat)
  # within each group, calculate expected amount of success and failure
  expect <- xtabs(cbind(1 - yhat, yhat) ~ cutyhat)
  # chi-square statistics
  chisq <- sum((obs - expect)^2/expect)
  # p-value
  P <- 1 - pchisq(chisq, g - 2)
  # outupt
  c("X^2" = chisq, Df = g - 2, "P(>Chi)" = P)
}
# Hosmer-Lemeshow goodness of fit test
hosmerlem(df$Y, p.hat, g=8)
```

```
##         X^2          Df     P(>Chi)
## 11.97045484  6.00000000  0.06263118
```

- $H_0 : log\frac{\pi}{1-\pi} = \beta_0 + \beta_1 \cdot (age) + \beta_2 \cdot (health)$
  $H_A : log\frac{\pi}{1-\pi} \neq \beta_0 + \beta_1 \cdot (age) + \beta_2 \cdot (health)$
- Decision rule:
  If p-value $< 0.05$, reject the null hypothesis and conclude that the model is not a good fit of the data.
  If p-value $> 0.05$, fail to reject the null hypothesis and conclude that the model is a good fit of the data.
- Conclusion: p-value obtained from the Hosmer-Lemeshow goodness of fit test is $0.06263 > 0.05$. So we fail to reject the null hypothesis and conclude that the model is a good fit of the data.

## 14.24

a. Refer to **Toxicity experiment** Problem 14.12. Use the groups given there to conduct a deviance goodness of fit test of the appropriateness of logistic regression model (14.20). Control the risk of a Type I error at .01. State the alternatives, decision rule, and conclusion.

    - In addition, conduct a Pearson Chi-square goodness of fit test at $\alpha = 0.01$.

```
# model
reg <- glm(Y/n~X, family = binomial(link=logit), weight = n, data = df)
# deviance goodness of fit test
c("X^2" = reg$dev,
  Df = nrow(df)-2,
  "P(>Chi)" = 1-pchisq(reg$dev, df=nrow(df)-2))
```

```
##       X^2        Df   P(>Chi)
## 1.4490930 4.0000000 0.8356191
```

- $H_0 : log\frac{\pi}{1-\pi} = \beta_0 + \beta_1 \cdot (dose.lv)$
  $H_A : log\frac{\pi}{1-\pi} \neq \beta_0 + \beta_1 \cdot (dose.lv)$
- Decision rule:
  If p-value $<$ 0.05, reject the null hypothesis and conclude that the model is not a good fit of the data.
  If p-value $>$ 0.05, fail to reject the null hypothesis and conclude that the model is a good fit of the data.
- Conclusion: p-value obtained from the deviance goodness of fit test is 0.8356191 $>$ 0.01. So we fail to reject the null hypothesis and conclude that the model is a good fit of the data.

```
# Pearson Chi-square Procedure(counts data only)
pgof<-function(n, y, pihat, p){
  # observed number of success
  Oy<-y
  # observed number of failure
  On<-n-y
  # expected number of success
  Ey<-n*pihat
  # expected number of failure
  En<-n*(1-pihat)
  # chi-square statistics
  chisq<-sum((Oy-Ey)^2/Ey) + sum((On-En)^2/En)
  # p-value
  pvalue <- 1 - pchisq(chisq, length(n)-p)
  # output
  c("X^2" = chisq, Df = length(n)-p, "P(>Chi)" = pvalue)
}
# Pearson Chi-square goodness of fit test
pgof(df$n, df$Y, reg$fitted, 2)
```

```
##       X^2        Df   P(>Chi)
## 1.4517865 4.0000000 0.8351462
```

The result from Pearson Chi-square goodness of fit test is very closed to that of the deviance goodness of fit test with p-value = 0.8351462. So the Pearson Chi-square goodness of fit test also validates fitness of the model.

### 14.36

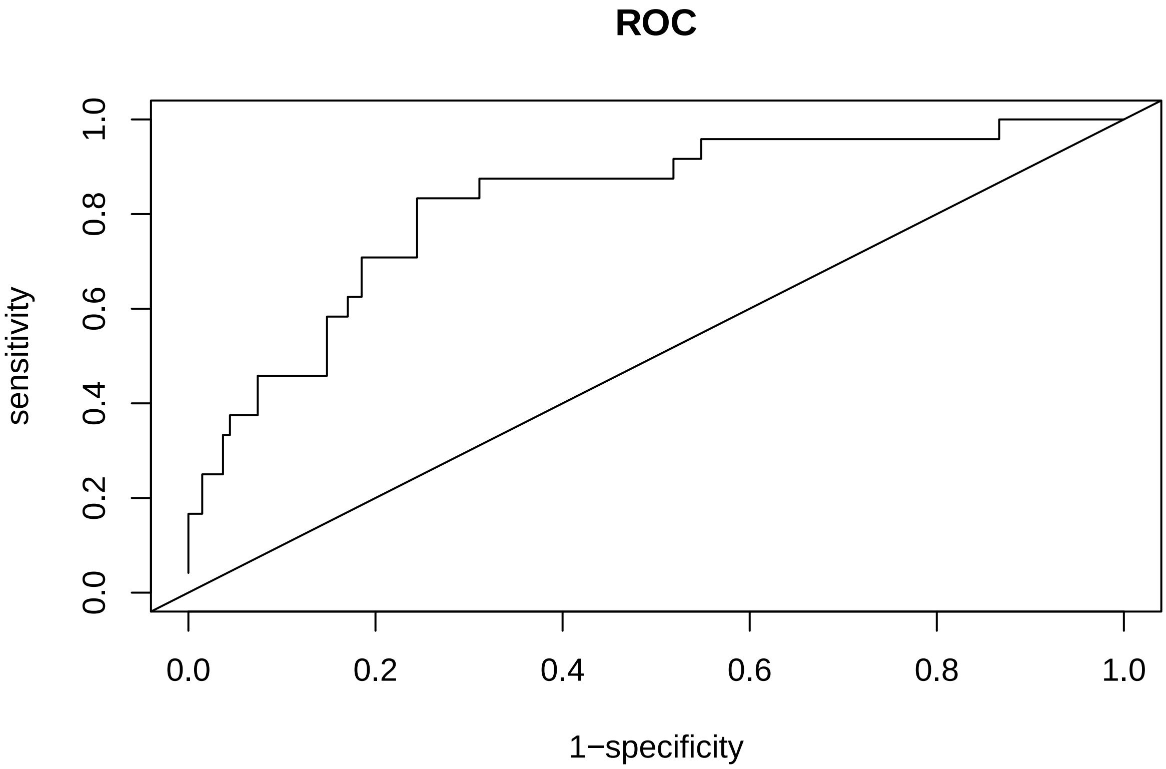
b. A prediction rule is to be based on the fitted regression function in Problem 14.14a. For the sample cases, find the total error rate, the error rate for clients receiving the flu shot, and the error rate for clients not receiving the flu shot for the following cutoffs: .05, .10, .15, .20.

| cutoff | Total.Error.Rate | False.Positive.Rate | False.Negative.Rate |
|---:|---:|---:|---:|
| 0.05 | 0.5408805 | 0.6296296 | 0.0416667 |
| 0.10 | 0.3647799 | 0.4074074 | 0.1250000 |
| 0.15 | 0.2641509 | 0.2814815 | 0.1666667 |
| 0.20 | 0.2012579 | 0.1629630 | 0.4166667 |

The total error rate, the error rate for clients receiving the flu shot (False Positive Rate), and the error rate for clients not receiving the flu shot (False Negative Rate) for 4 different cutoffs are listed as above.

c. Based on your results in part(b), which cutoff minimizes the total error rate? Are the error rates for clients receiving the flu shot and for clients not receiving the flu shot fairly balanced at this cutoff? Obtain the area under the ROC curve to assess the model's predictive power here. What do you conclude?

From part(b), cutoff = 0.2 minimizes the total error rate, but the error rates for clients receiving the flu shot and for clients not receiving the flu shot are not balanced in this situation.



```
##             AUR
## [1,] 0.8223765
```

The ROC curve is shown as above, and the area under the curve is 0.8223765, which indicates a high predictive power.

d. How can you establish whether the observed total error rate for the best cutoff in part (c) is a reliable indicator of the predictive ability of the fitted regression function and the chosen cutoff?

| cutoff | Total.Error.Rate | False.Negative.Rate | False.Positive.Rate |
|---:|---:|---:|---:|
| 0.1648737 | 0.2389937 | 0.2083333 | 0.2444444 |

No, even the best cutoff has a higher total error rate than an unreliable cutoff, 0.2 in part (c) for instance. Moreover, the total error rate cannot show the predictive ability of the fitted model. We use area under ROC instead to show the predictive ability. So, the observed total error rate is not a reliable indicator.