Homework 4

Charles Hwang

3/2/2022

Charles Hwang

Dr. Whalen

STAT 410-001

2 March 2022

Problem 4.4

```
rm(list=ls())
o0<-(exp(-3.866+0.397*0)/(1+exp(-3.866+0.397*0)))/(1-exp(-3.866+0.397*0)/(1+exp(-3.866+0.397*0)))
o2<-(exp(-3.866+0.397*2)/(1+exp(-3.866+0.397*2)))/(1-exp(-3.866+0.397*2)/(1+exp(-3.866+0.397*2)))
o4<-(exp(-3.866+0.397*4)/(1+exp(-3.866+0.397*4)))/(1-exp(-3.866+0.397*4)/(1+exp(-3.866+0.397*4)))
o5<-(exp(-3.866+0.397*5)/(1+exp(-3.866+0.397*5)))/(1-exp(-3.866+0.397*5)/(1+exp(-3.866+0.397*5)))
or<-data.frame(c(o2/o0,o4/o0,o5/o0),c("",o4/o2,o5/o2),c("","",o5/o4))
names(or)<-c("0","2","4")
row.names(or)<-c("2","4","5")
or
```

```
##          0               2               4
## 2 2.212228
## 4 4.893951 2.21222766265879
## 5 7.279047 3.29036993267909 1.48735593005131
```

We can see the odds of heart disease increases with snoring level, as previously expected. Specifically, those who "occasionally" snore are approximately 2.2122277 times more likely to have heart disease than those who "never" snore, those who snore "nearly every night" are approximately 4.8939512 times more likely to have heart disease than those who "never" snore, those who snore every night are approximately 7.2790474 times more likely to have heart disease than those who "never" snore, those who snore "nearly every night" are approximately 2.21222766265879 times more likely to have heart disease than those who "occasionally" snore, those who snore "every night" are approximately 3.29036993267909 times more likely to have heart disease than those who "occasionally" snore, and those who snore "every night" are approximately 1.48735593005131 times more likely to have heart disease than those who snore "nearly every night".

Problem 4.8

```
hc<-read.table("http://users.stat.ufl.edu/~aa/cat/data/Crabs.dat",header=TRUE)
cw<-glm(y~weight,family=binomial(link="logit"),data=hc)
exp(cw$coefficients["weight"]*0.1)                    # Problem 4.8(b)
```

```
##   weight
## 1.199032
```

```
# For every 0.1 kilogram increase in weight, the odds of a crab having at
# least one satellite crab are multiplied by approximately 1.199032 times.
wt<-xtabs(~hc$y+as.numeric(cw$fitted.values>mean(hc$y))) # Problem 4.8(c)
wt
```

```
##     as.numeric(cw$fitted.values > mean(hc$y))
## hc$y  0  1
##    0 45 17
##    1 43 68
```

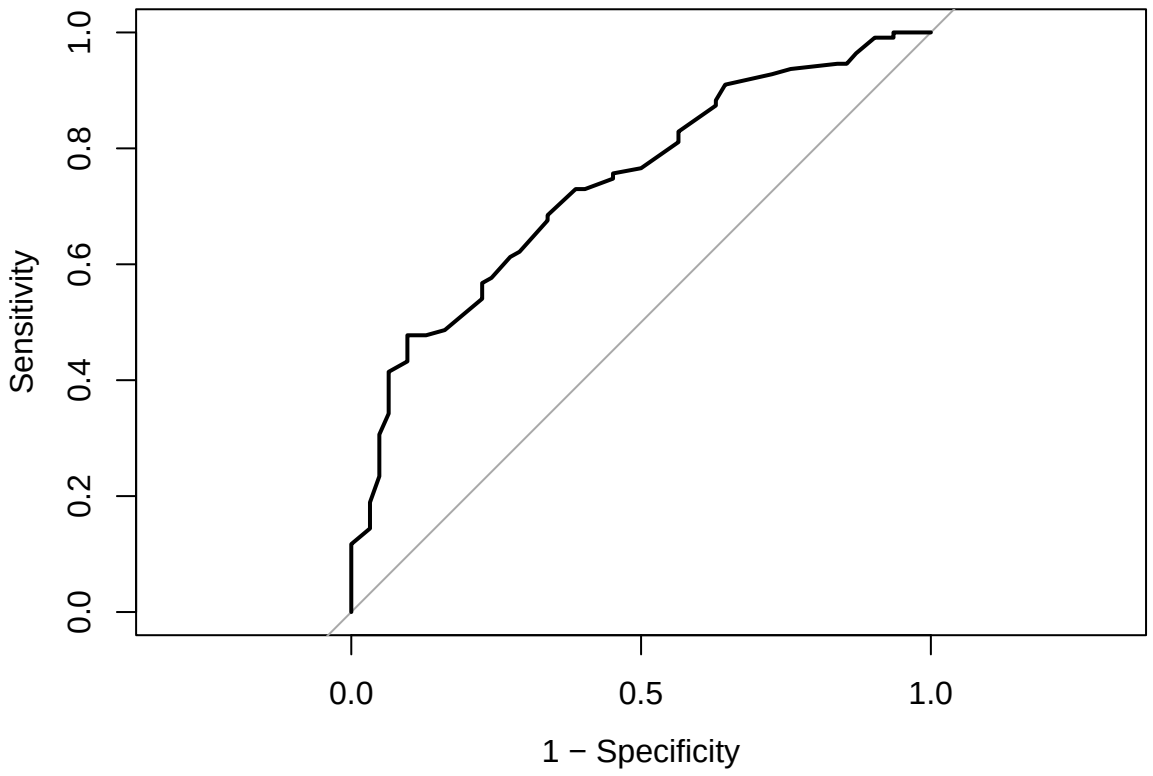
```
wt["1","1"]/sum(wt["1",])
```

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

```
wt["0","0"]/sum(wt["0",])
```

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

```
# The probability of a crab having at least one satellite crab given that this model
# predicted it to have at least one satellite crab is approximately 0.6126126. The
# probability of a crab not having at least one satellite crab given that this model
# predicted it to not have at least one satellite crab is approximately 0.7258065.
library(pROC)                                                    # Problem 4.8(d)
plot.roc(roc(y~cw$fitted.values,data=hc),legacy.axes=TRUE)
```




```
auc(roc(y~cw$fitted.values,data=hc))
```

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

We can see the area under the curve is approximately 0.7379396. There is approximately a 73.7939552 percent probability the predictions and outcomes are concordant.

Problem 4.9

```
cc<-glm(y~factor(color,levels=c(4,1:3)),family=binomial(link="logit"),data=hc) # Problem 4.9(a)
cc$coefficients
```

```
##                         (Intercept) factor(color, levels = c(4, 1:3))1
##                           -0.7621401                           1.8607523
## factor(color, levels = c(4, 1:3))2 factor(color, levels = c(4, 1:3))3
##                            1.7381500                           1.1298648
```

```
# logit(y-hat) = -0.7621401 + 1.8607523(c_1) + 1.73815(c_2) - 1.1298648(c_3)
# The variables for the first color would be c_1=1 and c_2=c_3=0, and the resulting
# probability is logit(y-hat)=-0.7621401+1.8607523(1)+1.73815(0)-1.1298648(0)=1.0986123 ->
# e^1.0986123 / (1 + e^1.0986123) = 0.75.
# The variables for the fourth color would be c_1 = c_2 = c_3 = 0, and the resulting
# probability is logit(y-hat)=-0.7621401+1.8607523(0)+1.73815(0)-1.1298648(0)=-0.7621401 ->
# e^-0.76214 / (1 + e^-0.76214) = 0.3182 = 7/22.
exp(cc$coefficients["factor(color, levels = c(4, 1:3))1"])
```

```
## factor(color, levels = c(4, 1:3))1
##                           6.428571
```

```
# The odds of a medium light crab has at least one satellite crab are approximately
# 6.428571 times of the odds of a dark crab having at least one satellite crab.
anova(cc,test="LRT")                                                    # Problem 4.9(b)
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: y
##
## Terms added sequentially (first to last)
##
##
##                                  Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                                                172     225.76
## factor(color, levels = c(4, 1:3))  3   13.698       169     212.06 0.003347 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# We reject H0 at the alpha = 0.05 level. There is sufficient evidence (p = 0.003346919)
# that color as a nominal-scale factor has a statistically significant effect in the model.
ccq<-glm(y~color,family=binomial(link="logit"),data=hc)                 # Problem 4.9(c)
ccq$coefficients
```

```
## (Intercept)       color
##   2.3634527  -0.7146794
```

```
# logit(y-hat) = 2.3634527 - 0.7146794(c), where c = 1 (ml), 2 (m), 3 (md), 4 (d)
exp(ccq$coefficients["color"])
```

```
##    color
## 0.489349
```

```
# For each darker level of color, the odds are multiplied by approximately 0.489349 times.
anova(ccq,test="LRT")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: y
##
## Terms added sequentially (first to last)
##
##
##       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                    172     225.76
## color  1   12.461       171     213.30 0.0004156 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# We reject H0 at the alpha = 0.05 level. There is sufficient evidence (p = 0.0004156143)
# that color as a quantitative variable has a statistically significant effect in the model.
# One advantage for this particular model with the quantitative color          # Problem 4.9(d)
# variable is that having only one degree of freedom "shrinks" the chi-square distribution
# so that the critical value for the test statistic is lower and the null hypothesis for the
# variable having no effect is more easily rejected. However, one disadvantage is that since
# there are more than two levels for color, it is unknown whether the designated values
# b_c = 1, 2, 3, 4 are statistically accurate for a single coefficient. In other words, the
# distance between the values b_c = 1, 2, 3, 4 are the same, implying the differences in
# probabilities between each level are the same, when that is not necessarily the case.
cwc<-glm(y~weight+color,family=binomial(link="logit"),data=hc)                # Problem 4.9(e)
sd(hc$weight)*cwc$coefficients["weight"]
```

```
##    weight
## 0.9538525
```

```
sd(hc$color)*cwc$coefficients["color"]
```

```
##      color
## -0.4123381
```

We can see that a one-s_w increase in weight is estimated to have approximately 2.3132773 times the effect of a one-s_c increase in color, adjusting for the other variable.

Problem 4.12

```
MBTI<-read.table("http://users.stat.ufl.edu/~aa/cat/data/MBTI.dat",header=TRUE)
glm(drink/n~EI+SN+TF+JP,family=binomial(link="logit"),weights=n,data=MBTI)$coefficients
```

```
## (Intercept)         EIi         SNs         TFt         JPp
##  -2.1140470  -0.5550115  -0.4291508   0.6873349   0.2022295
```

$\text{logit}(\hat{\pi}(x)) = -2.114047 - 0.5550115\beta_{EI} - 0.4291508\beta_{SN} + 0.6873349\beta_{TF} + 0.2022295\beta_{JP}$

$\hat{\pi}(x)$: probability of drinking alcohol frequently

β_{EI}: binary variable for **E**xtroversion/**I**ntroversion (0 for E, 1 for I)

β_{SN}: binary variable for **S**ensing/i**N**tuition (0 for N, 1 for S)

β_{TF}: binary variable for **T**hinking/**F**eeling (0 for F, 1 for T)

β_{JP}: binary variable for **J**udging/**P**erceiving (0 for J, 1 for P)

Looking at the model and the signs of the coefficients, we can see that in order to maximize $\hat{\pi}(x)$, the combination of ENTP should be chosen and thus it has the highest estimated probability of drinking alcohol frequently.

Problem 4.16

```
st<-read.table("http://users.stat.ufl.edu/~aa/cat/data/SoreThroat.dat",header=TRUE)
t<-glm(Y~D*T,family=binomial(link="logit"),data=st) # Problem 4.16(a)
t$coefficients
```

```
## (Intercept)           D           T         D:T
##  0.04978674  0.02847802 -4.47224144  0.07460127
```

```
# logit(y-hat) = 0.04978674 + 0.02847802(b_D) - 4.47224144(b_T) + 0.07460127(b_D)(b_T)
# logit(y-hat|t = 1) = 0.04978674+0.02847802(b_D)-4.47224144(1)+0.07460127(b_D)(1)=  # (i)
# logit(y-hat|t = 1) = -4.4224547 + 0.10307929(b_D)
exp(sum(t$coefficients[c("D","D:T")]))
```

```
## [1] 1.108579
```

```
# For every additional 1 minute increase in the duration of the surgery, the odds of
# the patient experiencing a sore throat upon waking up are multiplied by approximately
# 1.108579 times, given that a tracheal tube was used to secure the airway.
# logit(y-hat|t = 0) = 0.04978674+0.02847802(b_D)-4.47224144(0)+0.07460127(b_D)(0)=  # (ii)
# logit(y-hat|t = 0) = 0.04978674 + 0.02847802(b_D)
exp(sum(t$coefficients["D"]))
```

```
## [1] 1.028887
```

```
# For every additional 1 minute increase in the duration of the surgery, the odds of
# the patient experiencing a sore throat upon waking up are multiplied by approximately
# 1.028887 times, given that a laryngeal mask airway was used to secure the airway.
summary(t)$coefficients
```

```
##               Estimate Std. Error     z value   Pr(>|z|)
## (Intercept)  0.04978674 1.46940126  0.03388233 0.97297098
## D            0.02847802 0.03428577  0.83060747 0.40619541
## T           -4.47224144 2.46706836 -1.81277565 0.06986643
## D:T          0.07460127 0.05776816  1.29139077 0.19656821
```

```
# We fail to reject H0 at the alpha = 0.05 level. There is insufficient
# evidence (p = 0.19656821) that an interaction term is significant in the model. We can
# also see the weight of the coefficient for the interaction term is relatively small.
cor(st$Y,t$fitted)                                    # Problem 4.16(b)
```

```
## [1] 0.6598764
```

```
cor(st$Y,glm(Y~D+T,family=binomial(link="logit"),data=st)$fitted)
```

```
## [1] 0.6528899
```

We can see the correlation of the models with and without the interaction term $\beta_D\beta_T$ are nearly the same. The correlation is improved by approximately 0.0069865 in the model with the interaction terms.

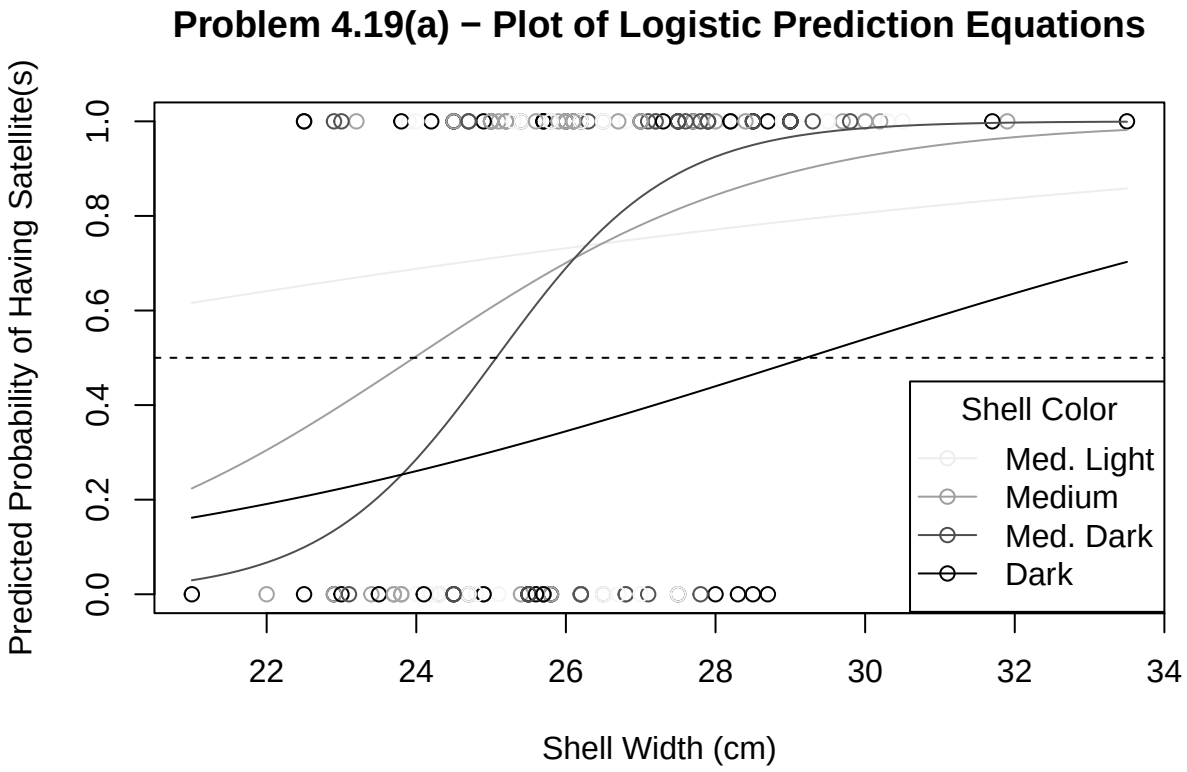
Problem 4.19

```
# If using the equation in the textbook answers (page 354), the code would instead be:
# glm(y~width*factor(color,levels=c(4,1:3)),family=binomial(link="logit"),data=hc)
```

```r
wc<-glm(y~width*factor(color),family=binomial(link="logit"),data=hc) # Problem 4.19(a)
wc$coefficients
```

```
##          (Intercept)                width       factor(color)2
##          -1.75260875           0.10600046          -8.28735421
##       factor(color)3       factor(color)4 width:factor(color)2
##         -19.76545392          -4.10122117           0.31287057
## width:factor(color)3 width:factor(color)4
##           0.75236820           0.09442916
```

```r
# logit(y-hat) = -1.75260875 + 0.10600046(b_w) - 8.28735421(b_2) - 19.76545392(b_3)
# - 4.10122117(b_4) + 0.31287057(b_w)(b_2) + 0.75236820(b_w)(b_3) + 0.09442916(b_w)(b_4)
# logit(y-hat|c = 1) = -1.75260875 + 0.10600046(b_w)
a1<-wc$coefficients["(Intercept)"]
# logit(y-hat|c = 2) = -1.75260875 + 0.10600046(b_w) - 8.28735421(1) + 0.31287057(b_w)(1)
# logit(y-hat|c = 2) = -10.03996 + 0.41887103(b_w)
a2<-sum(wc$coefficients[c("(Intercept)","factor(color)2")])
b2<-wc$coefficients[c("width","width:factor(color)2")]
# logit(y-hat|c = 3) = -1.75260875 + 0.10600046(b_w) - 19.76545392(1) + 0.75236820(b_w)(1)
# logit(y-hat|c = 3) = -21.51806267 + 0.85836866(b_w)
a3<-sum(wc$coefficients[c("(Intercept)","factor(color)3")])
b3<-wc$coefficients[c("width","width:factor(color)3")]
# logit(y-hat|c = 4) = -1.75260875 + 0.10600046(b_w) - 4.10122117(1) + 0.09442916(b_w)(1)
# logit(y-hat|c = 4) = -5.85382992 + 0.20042962(b_w)
a4<-sum(wc$coefficients[c("(Intercept)","factor(color)4")])
b4<-wc$coefficients[c("width","width:factor(color)4")]
plot(c(hc[hc$color==1,"width"],hc[hc$color==2,"width"],hc[hc$color==3,"width"],hc[hc$color==4,"width"])
curve(exp(a1+wc$coefficients["width"]*x)/(1+exp(a1+wc$coefficients["width"]*x)),col="gray93",add=TRUE)
curve(exp(a2+sum(b2)*x)/(1+exp(a2+sum(b2)*x)),col="gray62",add=TRUE)
curve(exp(a3+sum(b3)*x)/(1+exp(a3+sum(b3)*x)),col="gray31",add=TRUE)
curve(exp(a4+sum(b4)*x)/(1+exp(a4+sum(b4)*x)),add=TRUE)
abline(h=0.5,lty=2)
legend(30.6,0.45,title="Shell Color",c("Med. Light","Medium","Med. Dark","Dark"),col=c("gray93","gray62
```




```
exp(wc$coefficients["width"])
```

```
##    width
## 1.111822
```

```
# For every 1 centimeter increase in shell width, the odds of a medium light crab
# having at least one satellite crab are multiplied by approximately 1.111822 times.
exp(sum(b2))
```

```
## [1] 1.520244
```

```
# For every 1 centimeter increase in shell width, the odds of a medium crab having
# at least one satellite crab are multiplied by approximately 1.520244 times.
exp(sum(b3))
```

```
## [1] 2.359309
```

```
# For every 1 centimeter increase in shell width, the odds of a medium dark crab
# having at least one satellite crab are multiplied by approximately 2.359309 times.
exp(sum(b4))
```

```
## [1] 1.221928
```

```
# For every 1 centimeter increase in shell width, the odds of a dark crab having
# at least one satellite crab are multiplied by approximately 1.221928 times.
cor(hc$y,wc$fitted)                                                    # Problem 4.16(b)
```

```
## [1] 0.472076
```

```
cor(hc$y,glm(y~width+factor(color),family=binomial(link="logit"),data=hc)$fitted)
```

```
## [1] 0.4522131
```

We can see the correlation of the models with and without the interaction terms $\beta_w\beta_2$, $\beta_w\beta_3$, and $\beta_w\beta_4$ are

about the same. The correlation is improved by approximately 0.0198629 in the model with the interaction terms.

Problem 4.21

Table 4.4 (page 110) can be constructed using R:

```
wc$fitted.values[6]<-0.50000001 # One crab in Table 4.4 was misfitted for some reason
Actual<-factor(hc$y,levels=1:0)
"Prediction, pi_0=0.50"<-factor(as.numeric(wc$fitted.values>0.5),levels=1:0)
xtabs(~Actual+`Prediction, pi_0=0.50`)
```

```
##       Prediction, pi_0=0.50
## Actual  1  0
##      1 96 15
##      0 31 31
```

```
xtabs(~Actual+`Prediction, pi_0=0.50`)["1","1"]/sum(xtabs(~Actual+`Prediction, pi_0=0.50`)["1",])
```

```
## [1] 0.8648649
```

```
xtabs(~Actual+`Prediction, pi_0=0.50`)["0","0"]/sum(xtabs(~Actual+`Prediction, pi_0=0.50`)["0",])
```

```
## [1] 0.5
```

The probability of a crab having at least one satellite crab given that this model predicted it to have at least one satellite crab (sensitivity) is approximately 0.8648649. The probability of a crab not having at least one satellite crab given that this model predicted it to not have at least one satellite crab (specificity) is approximately 0.5.