Homework 5

Charles Hwang

3/31/2022

Charles Hwang

Dr. Whalen

STAT 410-001

31 March 2022

Problem 5.2

```r
rm(list=ls())
sub<-read.table("/Users/newuser/Desktop/Notes/Graduate/STAT 410 - Categorical Data Analysis/Substance2.
alc<-factor(rep(0:1,each=8)) # Coding each of 4 binary explanatory factors
cig<-factor(rep(0:1,2,each=4))
gen<-factor(rep(0:1,8))
rac<-factor(rep(0:1,4,each=2))
sub<-data.frame(alc,cig,gen,rac,sub$count[c(TRUE,FALSE)],sub$count[c(FALSE,TRUE)])
names(sub)<-c("a","c","g","r","Y","N") # Separating marijuana usage and nonusage
step(glm(Y/(Y+N)~a+c+g+r,family=binomial,weights=Y+N,data=sub))
```

```
## Start:  AIC=60.69
## Y/(Y + N) ~ a + c + g + r
##
##         Df Deviance    AIC
## <none>          4.69  60.69
## - r      1      6.89  60.88
## - g      1     15.06  69.05
## - a      1     95.48 149.47
## - c      1    502.23 556.22
```

```
##
## Call:  glm(formula = Y/(Y + N) ~ a + c + g + r, family = binomial, data = sub,
##     weights = Y + N)
##
## Coefficients:
## (Intercept)           a1           c1           g1           r1
##      0.3825      -2.9873      -2.8592       0.3297      -0.2989
##
## Degrees of Freedom: 15 Total (i.e. Null);  11 Residual
## Null Deviance:       860.6
## Residual Deviance: 4.694     AIC: 60.69
```

We can see the AIC stepwise selection process here with the `step()` function.

Problem 5.4

```
S<-read.table("/Users/newuser/Desktop/Notes/Graduate/STAT 410 - Categorical Data Analysis/Students.dat"
S$gender<-factor(S$gender)
S$veg<-factor(S$veg)
S$affil<-factor(S$affil) # 3 levels (1, 2, 3)
# Per Campuswire discussion, it was discovered the textbook considers the ordinal
# variables "ideol" and "relig" to be quantitative (section 11.5.3, pages 315-316).
# Thus, they will not be coded as factors here despite having defined levels.
S$abor<-factor(S$abor)
S$affirm<-factor(S$affirm)
S$life<-factor(S$life)   # 3 levels (1, 2, 3)
i<-glm(abor~ideol,family=binomial,data=S) # Section 5.1.4 (pages 126-127)     # 5.4a
r<-glm(abor~relig,family=binomial,data=S)                        # Step 1
n<-glm(abor~news,family=binomial,data=S)
h<-glm(abor~hsgpa,family=binomial,data=S)
g<-glm(abor~gender,family=binomial,data=S)
1-pchisq(i$null.deviance-i$deviance,i$df.null-i$df.residual) # Running LRTs   # ***
```

```
## [1] 3.268956e-05
```

```
1-pchisq(r$null.deviance-r$deviance,r$df.null-r$df.residual)                  # ***
```

```
## [1] 0.0001434376
```

```
1-pchisq(n$null.deviance-n$deviance,n$df.null-n$df.residual)                  # **
```

```
## [1] 0.006781714
```

```
1-pchisq(h$null.deviance-h$deviance,h$df.null-h$df.residual)                  # " "
```

```
## [1] 0.7863374
```

```
1-pchisq(g$null.deviance-g$deviance,g$df.null-g$df.residual)                  # " "
```

```
## [1] 0.2804659
```

```
irn<-glm(abor~ideol+relig+news,family=binomial,data=S) # Adding ideol, relig, news
step(irn,direction="backward")                                   # Step 2
```

```
## Start:  AIC=37.79
## abor ~ ideol + relig + news
## 
##         Df Deviance    AIC
## <none>       29.791 37.791
## - relig  1   32.014 38.014
## - ideol  1   41.697 47.697
## - news   1   42.522 48.522
```

```
## 
## Call:  glm(formula = abor ~ ideol + relig + news, family = binomial, 
##     data = S)
## 
## Coefficients:
## (Intercept)        ideol        relig         news  
##      3.5205      -1.2515      -0.7198       1.1292  
## 
## Degrees of Freedom: 59 Total (i.e. Null);  56 Residual
## Null Deviance:       62.72
## Residual Deviance: 29.79     AIC: 37.79
```

```
irnh<-glm(abor~ideol+relig+news+hsgpa,family=binomial,data=S)    # Step 3
irng<-glm(abor~ideol+relig+news+gender,family=binomial,data=S)
1-pchisq(irn$deviance-irnh$deviance,irn$df.residual-irnh$df.residual)         # *
```

```
## [1] 0.04485803
```

```
1-pchisq(irn$deviance-irng$deviance,irn$df.residual-irng$df.residual)         # " "
```

```
## [1] 0.8203808
```

```
irnhxxx<-glm(abor~ideol*relig*news*hsgpa,family=binomial,data=S) # Step 4
irnhx<-glm(abor~ideol+relig+news+hsgpa+ideol*relig+ideol*news+ideol*hsgpa+relig*news+relig*hsgpa+news*hs
1-pchisq(irnh$deviance-irnhx$deviance,irnh$df.residual-irnhx$df.residual)     # " "
```

```
## [1] 0.1501267
```

```
1-pchisq(irnh$deviance-irnhxxx$deviance,irnh$df.residual-irnhxxx$df.residual) # " "
```

```
## [1] 0.284446
```

```
irnh # Final model: abor = ideol*(x_i) + relig*(x_r) + news*(x_n) + hsgpa*(x_h)
```

```
##
## Call:  glm(formula = abor ~ ideol + relig + news + hsgpa, family = binomial,
##     data = S)
##
## Coefficients:
## (Intercept)        ideol        relig         news        hsgpa
##     11.8254      -1.4696      -0.7368       1.4015      -2.3686
##
## Degrees of Freedom: 59 Total (i.e. Null);  55 Residual
## Null Deviance:       62.72
## Residual Deviance: 25.77     AIC: 35.77
```

```
library(MASS)                                                                 # 5.4b
stepAIC(glm(abor~gender+age+hsgpa+cogpa+dhome+dres+tv+sport+news+aids+veg+ideol+relig+affirm,family=bino
```

```
## Start:  AIC=51.37
## abor ~ gender + age + hsgpa + cogpa + dhome + dres + tv + sport +
##     news + aids + veg + ideol + relig + affirm
##
##          Df Deviance    AIC
## - sport   1   21.380 49.380
## - gender  1   21.665 49.665
## - age     1   21.752 49.752
## - cogpa   1   22.028 50.028
## - aids    1   22.197 50.197
## - relig   1   22.355 50.355
## - dhome   1   22.466 50.466
## - affirm  1   22.664 50.664
## - dres    1   22.927 50.927
## - tv      1   23.147 51.147
## <none>        21.368 51.368
## - veg     1   23.389 51.389
## - hsgpa   1   24.924 52.924
## - ideol   1   32.261 60.261
## - news    1   34.371 62.371
##
## Step:  AIC=49.38
## abor ~ gender + age + hsgpa + cogpa + dhome + dres + tv + news +
##     aids + veg + ideol + relig + affirm
##
##          Df Deviance    AIC
## - gender  1   21.686 47.686
## - age     1   21.754 47.754
## - aids    1   22.199 48.199
## - cogpa   1   22.261 48.261
## - relig   1   22.397 48.397
## - dhome   1   22.497 48.497
## - affirm  1   22.689 48.689
## - dres    1   22.927 48.927
## - tv      1   23.172 49.172
## <none>        21.380 49.380
## - veg     1   23.778 49.778
## - hsgpa   1   24.990 50.990
## - ideol   1   32.418 58.418
## - news    1   35.239 61.239
##
## Step:  AIC=47.69
## abor ~ age + hsgpa + cogpa + dhome + dres + tv + news + aids +
##     veg + ideol + relig + affirm
##
##          Df Deviance    AIC
## - age     1   22.094 46.094
## - relig   1   22.418 46.418
## - aids    1   22.680 46.680
## - dhome   1   22.713 46.713
## - affirm  1   22.787 46.787
## - dres    1   23.051 47.051
## - cogpa   1   23.200 47.200
## <none>        21.686 47.686
## - veg     1   24.103 48.103
## - tv      1   24.238 48.238
## - hsgpa   1   25.008 49.008
## - ideol   1   33.813 57.813
## - news    1   35.965 59.965
##
## Step:  AIC=46.09
## abor ~ hsgpa + cogpa + dhome + dres + tv + news + aids + veg +
##     ideol + relig + affirm
##
##          Df Deviance    AIC
## - relig   1   22.691 44.691
## - aids    1   22.701 44.701
## - dhome   1   22.740 44.740
## - affirm  1   22.790 44.790
## - dres    1   23.138 45.138
## - cogpa   1   23.553 45.553
## <none>        22.094 46.094
## - veg     1   24.106 46.106
## - tv      1   24.288 46.288
## - hsgpa   1   25.454 47.454
## - ideol   1   33.815 55.815
## - news    1   36.056 58.056
##
## Step:  AIC=44.69
## abor ~ hsgpa + cogpa + dhome + dres + tv + news + aids + veg +
##     ideol + affirm
##
##          Df Deviance    AIC
## - affirm  1   23.286 43.286
## - aids    1   23.371 43.371
## - dhome   1   23.773 43.773
## - veg     1   24.626 44.626
## - cogpa   1   24.653 44.653
## <none>        22.691 44.691
## - dres    1   24.784 44.784
## - tv      1   25.364 45.364
## - hsgpa   1   26.035 46.035
## - news    1   36.921 56.921
## - ideol   1   40.943 60.943
##
## Step:  AIC=43.29
## abor ~ hsgpa + cogpa + dhome + dres + tv + news + aids + veg +
##     ideol
##
##         Df Deviance    AIC
## - aids   1   23.754 41.754
## - dhome  1   23.901 41.901
## - veg    1   24.658 42.658
## - dres   1   24.785 42.785
## - cogpa  1   25.135 43.135
## <none>       23.286 43.286
## - tv     1   25.430 43.430
## - hsgpa  1   26.426 44.426
## - news   1   37.250 55.250
## - ideol  1   41.782 59.782
##
## Step:  AIC=41.75
## abor ~ hsgpa + cogpa + dhome + dres + tv + news + veg + ideol
##
##         Df Deviance    AIC
## - dhome  1   24.266 40.266
## - veg    1   24.712 40.712
## - dres   1   24.790 40.790
## - cogpa  1   25.197 41.197
## - tv     1   25.450 41.450
## <none>       23.754 41.754
## - hsgpa  1   26.694 42.694
## - news   1   37.343 53.343
## - ideol  1   43.856 59.856
##
## Step:  AIC=40.27
## abor ~ hsgpa + cogpa + dres + tv + news + veg + ideol
##
##         Df Deviance    AIC
## - veg    1   25.004 39.004
## - dres   1   25.279 39.279
## - tv     1   25.716 39.716
## - cogpa  1   25.790 39.790
## <none>       24.266 40.266
## - hsgpa  1   27.251 41.251
## - news   1   39.648 53.648
## - ideol  1   46.859 60.859
##
## Step:  AIC=39
## abor ~ hsgpa + cogpa + dres + tv + news + ideol
##
##         Df Deviance    AIC
## - cogpa  1   25.912 37.912
## - tv     1   26.009 38.009
## - dres   1   26.151 38.151
## <none>       25.004 39.004
## - hsgpa  1   27.460 39.460
## - news   1   39.657 51.657
## - ideol  1   50.040 62.040
##
## Step:  AIC=37.91
## abor ~ hsgpa + dres + tv + news + ideol
##
##         Df Deviance    AIC
## - dres   1   27.146 37.146
## - tv     1   27.160 37.160
## - hsgpa  1   27.839 37.839
## <none>       25.912 37.912
## - news   1   40.892 50.892
## - ideol  1   50.933 60.933
##
## Step:  AIC=37.15
## abor ~ hsgpa + tv + news + ideol
##
##         Df Deviance    AIC
## - tv     1   27.944 35.944
## <none>       27.146 37.146
## - hsgpa  1   30.248 38.248
## - news   1   42.143 50.143
## - ideol  1   52.334 60.334
##
## Step:  AIC=35.94
## abor ~ hsgpa + news + ideol
##
##         Df Deviance    AIC
## <none>       27.944 35.944
## - hsgpa  1   32.014 38.014
## - news   1   44.667 50.667
## - ideol  1   54.654 60.654
```

```
##
## Call:  glm(formula = abor ~ hsgpa + news + ideol, family = binomial,
##     data = S)
##
## Coefficients:
## (Intercept)        hsgpa         news        ideol
##      11.287       -2.338        1.291       -1.594
##
## Degrees of Freedom: 59 Total (i.e. Null);  56 Residual
## Null Deviance:       62.72
## Residual Deviance: 27.94     AIC: 35.94
```

```
library(car)                                                                   # 5.4c
yveg<-glm(veg~gender+age+hsgpa+cogpa+dhome+dres+tv+sport+news+aids+ideol+relig+abor+affirm,family=binomi
1-pchisq(yveg$null.deviance-yveg$deviance,yveg$df.null-yveg$df.residual)
```

```
## [1] 0.04481806
```

```
summary(yveg) # Section 5.3.2 (pages 137-138)
```

```
##
## Call:
## glm(formula = veg ~ gender + age + hsgpa + cogpa + dhome + dres +
##     tv + sport + news + aids + ideol + relig + abor + affirm,
##     family = binomial, data = S)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.40618  -0.23680  -0.00534   0.00000   1.90975
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.541e+00  2.902e+03   0.002   0.9985
## gender1     -1.844e+00  1.641e+00  -1.124   0.2612
## age          4.772e-02  7.677e-02   0.622   0.5342
## hsgpa       -3.839e+00  2.885e+00  -1.331   0.1833
## cogpa       -2.937e+00  2.269e+00  -1.294   0.1956
## dhome       -1.123e-03  7.130e-04  -1.575   0.1153
## dres         4.552e-01  2.562e-01   1.776   0.0757 .
## tv          -2.897e-02  1.289e-01  -0.225   0.8221
## sport       -5.885e-01  3.945e-01  -1.492   0.1358
## news         1.344e-01  2.928e-01   0.459   0.6462
## aids         2.069e-01  2.541e-01   0.814   0.4155
## ideol       -2.458e+00  1.469e+00  -1.673   0.0942 .
## relig        1.636e+00  1.071e+00   1.528   0.1266
## abor1       -3.573e+00  3.248e+00  -1.100   0.2713
## affirm1      2.355e+01  2.902e+03   0.008   0.9935
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 50.725  on 59  degrees of freedom
## Residual deviance: 26.645  on 45  degrees of freedom
## AIC: 56.645
```

```
##
## Number of Fisher Scoring iterations: 19
```

We can see that none of the 12 variables are statistically significant at the $\alpha = 0.05$ level using the Wald test, and thus none of them would be selected when manually using the forward selection process. However, we can reason that at least some of these 12 variables are likely statistically significant in the presence of some combinations of each other, which is why the likelihood-ratio test for $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = \beta_{12} = 0$ is statistically significant ($p = 0.0448181$).

Problem 5.7

Problem **5.7a** - We fail to reject H_0 at the $\alpha = 0.05$ level. There is insufficient evidence ($p = 0.2395059$) that the data do not come from a specified distribution.

Problem **5.7b** - We can see the fitted values for AZT usage vs. nonusage, regardless of race, are significantly different (0.1496 vs. 0.2654 for white, 0.1427 vs. 0.2547 for black), which reflects the statistical significance of the AZT variable in the model ($p = 0.00991$). Similarly, we can see the fitted values for race, regardless of AZT usage, are nearly the same (0.1496 vs. 0.1427 for AZT users, 0.2654 vs. 0.2547 for non-AZT users), which reflects the relative unimportance of the race variable in the model ($p = 0.84755$).

Problem **5.7c** - Standardized residuals are by definition adjusted for standard distributions which can make them easier to compare with one another.

Problem 5.12

Problem 5.12a

We can see from the explanation in Section 5.3.1 (pages 136-137) and graph of similar data in Figure 5.2 that the data in this problem have *complete separation* and *perfect discrimination*, which makes $\hat{\beta} = \infty$.

```
x<-c(0,10,20,30,70,80,90,100)
y<-rep(0:1,each=4)
summary(glm(y~x,family=binomial))$coefficients["x",c("Estimate","Std. Error")]
```

```
##    Estimate  Std. Error
##    1.161253 2368.996205
```

```
xb<-c(x,50,50) # Problem 5.12b
yb<-c(y,0,1)
summary(glm(yb~xb,family=binomial))$coefficients["xb",c("Estimate","Std. Error")]
```

```
##   Estimate Std. Error
##   1.011254 528.572070
```

```
# No, I do not believe these are correct. As Section 5.3.1 explains, the data now have
# quasi-complete separation, which still leads software to report inaccurate estimates.
xc<-c(x,50.1,49.9)
summary(glm(yb~xc,family=binomial))
```

```
##
## Call:
## glm(formula = yb ~ xc, family = binomial)
##
## Deviance Residuals:
##       Min        1Q    Median        3Q       Max
## -1.19028  -0.01219   0.00000   0.01219   1.19028
##
## Coefficients:
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -15.1231    34.4257  -0.439     0.66
## xc            0.3025     0.6879   0.440     0.66
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13.8629  on 9  degrees of freedom
## Residual deviance:  2.8434  on 8  degrees of freedom
## AIC: 6.8434
##
## Number of Fisher Scoring iterations: 10
```

We can see the estimates for $\hat{\beta}$ (0.3024623) and $SE_{\hat{\beta}}$ (0.6879377) appear to be more accurate now that there is no perfect discrimination or quasi-complete separation in the data.

Problem 5.17

```
st<-read.table("http://users.stat.ufl.edu/~aa/cat/data/SoreThroat.dat",header=TRUE)
glm(Y~D+T,family=gaussian(link=identity),data=st) # Problem 5.17a
```

```
##
## Call:  glm(formula = Y ~ D + T, family = gaussian(link = identity),
##     data = st)
##
## Coefficients:
## (Intercept)            D            T
##    0.366973     0.009062    -0.319094
##
## Degrees of Freedom: 34 Total (i.e. Null);  32 Residual
## Null Deviance:       8.171
## Residual Deviance: 5.17  AIC: 40.39
```

```
# For every 1 minute increase in the duration of the surgery, the probability of
# a patient experiencing a sore throat upon waking up increases by approximately
# 0.009061879.
# Additionally, the probability of a patient experiencing a sore throat upon waking up
# is approximately 0.3190944 lower when using a tracheal tube to secure the airway
# as opposed to a laryngeal mask.
glm(Y~D+T,family=binomial(link=probit),data=st)   # Problem 5.17b
```

```
##
## Call:  glm(formula = Y ~ D + T, family = binomial(link = probit), data = st)
##
## Coefficients:
## (Intercept)            D            T
##    -0.85469      0.03953     -0.92641
##
## Degrees of Freedom: 34 Total (i.e. Null);  32 Residual
## Null Deviance:       46.18
## Residual Deviance: 30.34     AIC: 36.34
```

```
glm(Y~D+T,family=binomial(link=probit),data=st)$coefficients["D"]
```

```
##          D
## 0.03952956
```

```
abs(glm(Y~D+T,family=binomial(link=probit),data=st)$coefficients["T"])
```

```
##         T
## 0.9264137
```

For every 1 minute increase in the duration of the surgery, the distribution of the latent variable for whether a patient experiences a sore throat upon waking up is shifted higher by approximately 0.0395296 standard deviations.

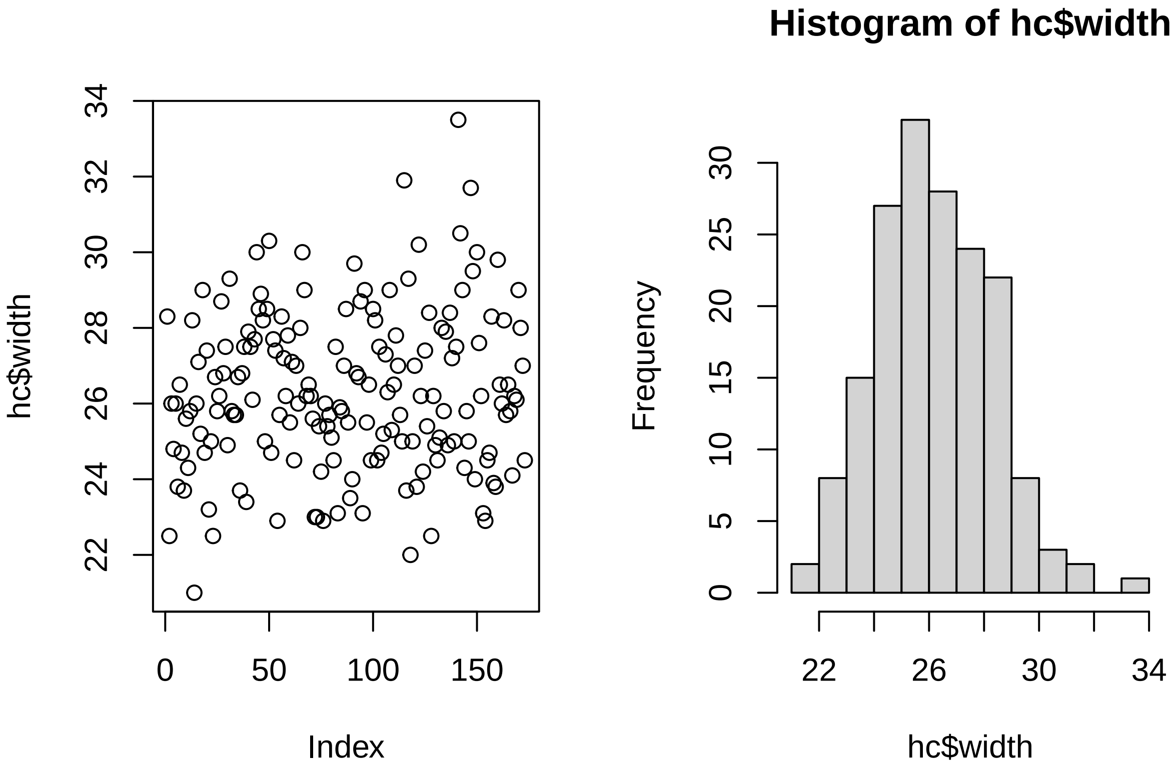
Additionally, the distribution of the latent variable for whether a patient experiences a sore throat upon waking up is shifted lower by approximately 0.9264137 standard deviations when using a tracheal tube to secure the airway as opposed to a laryngeal mask.

Problem 5.20

```
hc<-read.table("http://users.stat.ufl.edu/~aa/cat/data/Crabs.dat",header=TRUE)
rel<-glm(y~width,family=binomial,data=hc)$coefficients
eabx<-as.numeric(exp(rel["(Intercept)"]+rel["width"]*mean(hc$width)))
eabx1<-as.numeric(exp(rel["(Intercept)"]+rel["width"]*(mean(hc$width)+sd(hc$width))))
pib<-as.numeric(eabx/(1+eabx))
pib1<-as.numeric(eabx1/(1+eabx1))
l<-log(pib1/(1-pib1)/(pib/(1-pib)))
d<-1+(1+l^2)*exp(5/4*l^2)/(1+exp(-l^2/4))
ceiling((qnorm(1-0.05)+qnorm(1-0.1)*exp(-l^2/4))^2*(1+2*pib*d)/(pib*l^2)) # Round up
```

```
## [1] 81
```

```
par(mfrow=c(1,2))
plot(hc$width)
hist(hc$width)
```



This result requires the assumption that the variable for width be random and normally distributed, which it appears to be.