qbs121_hw3_gibran

Gibran Erlangga

1/25/2022

1 Problems

1.

(a) Write the log likelihood for the logistic regression model $\text{logit}(\Pr[Y|X_1 = x_1, X_2 = x_2]) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

$$$$

(b) Differentiate with respect to β_0.

$$$$

(c) Let f_i be the linear combination $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$ and $p_i = \exp[f_i]/(1 + \exp[f_i])$. Interpret p_i.

$$$$

(d) At the maximum likelihood estimate the derivative above equals zero. Equate the derivative to zero and write in terms of p_i. What does the sum $\sum_{i=1}^{n} p_i$ equal, and what does the mean $\sum_{i=1}^{n} p_i/n$ equal?

(e) How would you describe $\sum(y_i - p_i)^2/n$?

2 Data Analyses

2.1 Analysis of Burn Data

1. Install and utilize the R library *aplore3*. Using the dataset *burn1000* develop a model for predicting death.

```
library(aplore3)

data <- burn1000
head(data, 10)
```

```
##    id facility death  age gender      race tbsa inh_inj flame
## 1   1       11 Alive 26.6   Male     White 25.3      No   Yes
## 2   2        1 Alive  2.0 Female Non-White  5.0      No    No
## 3   3       12 Alive 22.0 Female Non-White  2.0      No    No
## 4   4        1 Alive 37.3   Male     White  2.0      No    No
## 5   5        1 Alive 52.1   Male     White  6.0      No   Yes
## 6   6        6 Alive 50.2   Male     White  7.0      No    No
## 7   7       22 Alive  2.5 Female Non-White  7.0      No    No
## 8   8        1 Alive 53.8 Female     White  0.9      No   Yes
## 9   9        1 Alive 31.9   Male     White  2.0      No    No
## 10 10        1 Alive 41.1   Male     White 22.0      No   Yes
```

```
model_death <- glm(death~tbsa+age+inh_inj+race, family=binomial, data=data)
summary(model_death)
```

```
##
## Call:
## glm(formula = death ~ tbsa + age + inh_inj + race, family = binomial,
##     data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -3.06117  -0.25801  -0.08970  -0.03746   2.52330
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.594684   0.608982 -12.471  < 2e-16 ***
## tbsa         0.090438   0.009088   9.951  < 2e-16 ***
## age          0.084445   0.008484   9.954  < 2e-16 ***
## inh_injYes   1.523055   0.351206   4.337 1.45e-05 ***
## raceWhite   -0.623468   0.298934  -2.086    0.037 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 845.42  on 999  degrees of freedom
## Residual deviance: 339.78  on 995  degrees of freedom
## AIC: 349.78
##
## Number of Fisher Scoring iterations: 7
```

```
#can try lasso / stepwise regression
```

2. Report the C-index (AUROC).

```
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 4.1.1
```

```
## Type 'citation("pROC")' for a citation.
```

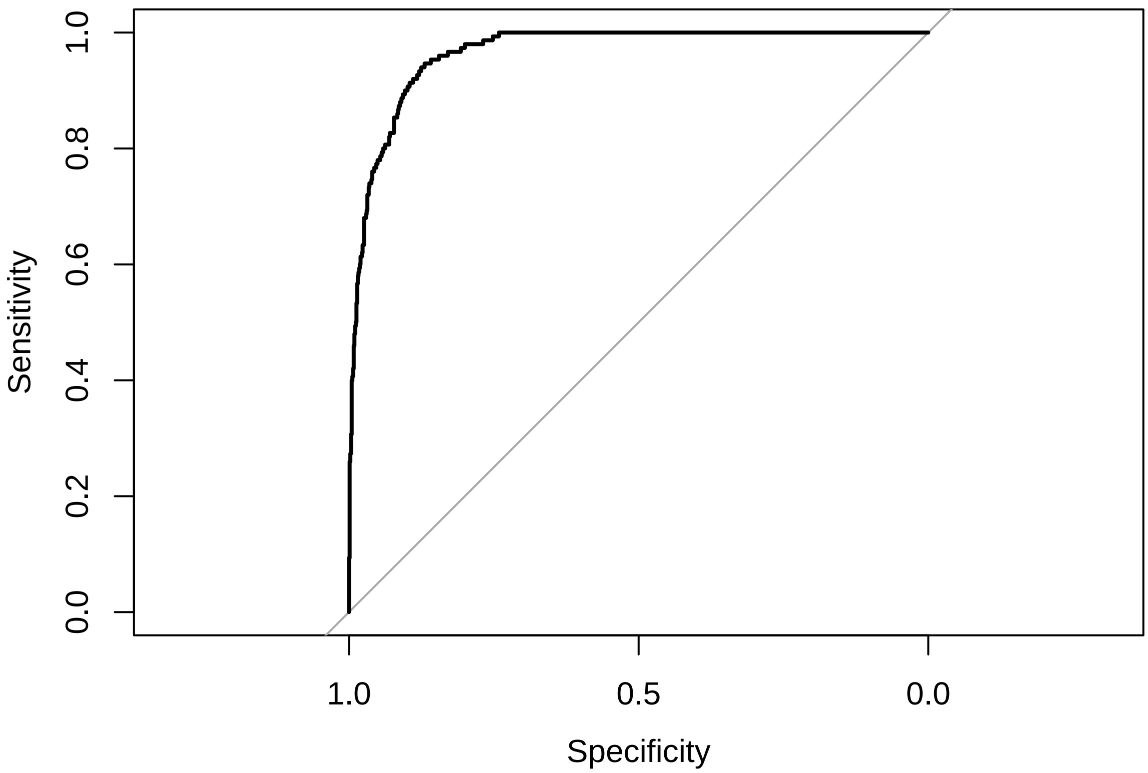
```
##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
roc(data$death, model_death$fitted.values, plot=TRUE)
```

```
## Setting levels: control = Alive, case = Dead

## Setting direction: controls < cases
```




```
##
## Call:
## roc.default(response = data$death, predictor = model_death$fitted.values,     plot = TRUE)
##
## Data: model_death$fitted.values in 850 controls (data$death Alive) < 150 cases (data$death Dead).
## Area under the curve: 0.9661
```

3. Is the effect of *inh_inj* on mortality modified by age?

```r
summary(o <- glm(death~inh_inj + age, family=binomial, data=data))
```

```
##
## Call:
## glm(formula = death ~ inh_inj + age, family = binomial, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5433  -0.4480  -0.2131  -0.1213   3.1485
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.002426   0.338953  -14.76   <2e-16 ***
## inh_injYes   2.923704   0.268678   10.88   <2e-16 ***
## age          0.058782   0.005548   10.60   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 845.42  on 999  degrees of freedom
## Residual deviance: 533.54  on 997  degrees of freedom
## AIC: 539.54
##
## Number of Fisher Scoring iterations: 6
```

The results show a statistically significant result after adding the *age* variable into the equation, we cannot really say anything about the causation between both.

4. Is the effect of age on mortality modified by *inh_inj*?

```r
summary(o <- glm(death~age+inh_inj, family=binomial, data=data))
```

```
##
## Call:
## glm(formula = death ~ age + inh_inj, family = binomial, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5433  -0.4480  -0.2131  -0.1213   3.1485
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.002426   0.338953  -14.76   <2e-16 ***
## age          0.058782   0.005548   10.60   <2e-16 ***
## inh_injYes   2.923704   0.268678   10.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 845.42  on 999  degrees of freedom
```

```
## Residual deviance: 533.54  on 997  degrees of freedom
## AIC: 539.54
##
## Number of Fisher Scoring iterations: 6
```

Similar to the previous question, the results show a statistically significant result after adding the inh_inj variable into the equation, we cannot really say anything about the causation between both.

2.3 Data With a Zero Cell

Create the following dataset consisting of a dependent variable, Success, and two covariates, Treatment and Female, using the following 3 lines of code:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7.9000
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.1.1
```

```
## Warning: package 'tibble' was built under R version 4.1.1
```

```
## Warning: package 'tidyr' was built under R version 4.1.1
```

```
## Warning: package 'readr' was built under R version 4.1.1
```

```
## Warning: package 'stringr' was built under R version 4.1.1
```

```
## Warning: package 'forcats' was built under R version 4.1.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)

Treatment = rep(c(0,1,0,1),each=10)
Female = rep(c(0,1),each=20)
Success = rep(rep(0:1,4), times=c(8,2,5,5,5,5,0,10))

df <- data.frame(Treatment, Female, Success)
head(df, 10)
```

```
##    Treatment Female Success
## 1          0      0       0
## 2          0      0       0
## 3          0      0       0
## 4          0      0       0
## 5          0      0       0
## 6          0      0       0
## 7          0      0       0
## 8          0      0       0
## 9          0      0       1
## 10         0      0       1
```

1. Calculate the success frequency for the 4 combinations of Treatment and Gender.

```
combinations <- list(c(0,0), c(0,1), c(1,0), c(1,1))
success_freq <- data.frame()

for (i in combinations) {
  df_temp <- df %>% filter(Treatment == i[[1]] & Female == i[[2]])
  df_sum <- df_temp %>% group_by(Success) %>% summarise(n=n()) %>% mutate(freq=n/sum(n))
  df_sum$Treatment <- i[[1]]
  df_sum$Female <- i[[2]]
  df_sum <- df_sum[, c(4, 5, 1, 2, 3)]
  success_freq <- rbind(success_freq, df_sum)
}

success_freq
```

```
## # A tibble: 7 x 5
##   Treatment Female Success     n  freq
##       <dbl>  <dbl>   <int> <int> <dbl>
## 1         0      0       0     8   0.8
## 2         0      0       1     2   0.2
## 3         0      1       0     5   0.5
## 4         0      1       1     5   0.5
## 5         1      0       0     5   0.5
## 6         1      0       1     5   0.5
## 7         1      1       1    10   1
```

2. Estimate the odds ratio relating Success to Treatment.

```
model <- glm(Success~Treatment, family=binomial, data=df)
summary(model)
```

```
##
## Call:
## glm(formula = Success ~ Treatment, family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6651  -0.9282   0.7585   0.7585   1.4490
##
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.6190     0.4688  -1.320   0.1867
## Treatment     1.7177     0.6975   2.463   0.0138 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 55.051  on 39  degrees of freedom
## Residual deviance: 48.391  on 38  degrees of freedom
## AIC: 52.391
##
## Number of Fisher Scoring iterations: 4
```

```r
exp(model$coeff['Treatment'])
```

```
## Treatment
##  5.571429
```

3. Estimate the odds ratio relating Success to Gender.

```r
model <- glm(Success~Female, family=binomial, data=df)
summary(model)
```

```
##
## Call:
## glm(formula = Success ~ Female, family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6651  -0.9282   0.7585   0.7585   1.4490
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.6190     0.4688  -1.320   0.1867
## Female        1.7177     0.6975   2.463   0.0138 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 55.051  on 39  degrees of freedom
## Residual deviance: 48.391  on 38  degrees of freedom
## AIC: 52.391
##
## Number of Fisher Scoring iterations: 4
```

```r
exp(model$coeff['Female'])
```

```
##   Female
## 5.571429
```

4. Include in a logistic regression the interaction of Treatment and Gender and comment on its statistical significance and coefficient.

```
summary(glm(Success~Treatment*Female, family=binomial, data=df))
```

```
##
## Call:
## glm(formula = Success ~ Treatment * Female, family = binomial,
##     data = df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.17741  -0.79539   0.00013   1.17741   1.79412
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.3863     0.7906  -1.754   0.0795 .
## Treatment          1.3863     1.0124   1.369   0.1709
## Female             1.3863     1.0124   1.369   0.1709
## Treatment:Female  17.1798  2062.6398   0.008   0.9934
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 55.051  on 39  degrees of freedom
## Residual deviance: 37.734  on 36  degrees of freedom
## AIC: 45.734
##
## Number of Fisher Scoring iterations: 17
```

The coefficient of variable Treatment, Gender and the interaction between both are 1.3863, 1.3863 and 17.1798, respectively. However, none of them shows a statistically significant results. Additionally, the z-score for the interaction variable is extremely small compared to the individual variables.

2.4 Concussion Data

Run the following code to read in and restructure a dataset that recorded concussions in college sports according to sex of athlete, sport and year. The columns in the matrix (data.frame) named Y are the number of athletes with and without concussions respectively.

```
DF <- read.delim("http://users.stat.ufl.edu/~winner/data/concussion.dat", sep="", header=FALSE)
names(DF) <- c("Sex","Sport","Year","Concussion","Count")

DF0 <- DF[DF$Concussion==0, ]
DF1 <- DF[DF$Concussion==1, ]

Cov <- data.frame(DF0[,1:3])
Y <- cbind(CountConc=DF1[,5], CountNoConc=DF0[,5])
```

1. Derive the contingency table of concussion by sex.

```
concussion_1 <- tapply(Y[,1], Cov$Sex, sum)
concussion_0 <- tapply(Y[,2], Cov$Sex, sum)

contingency_table <- rbind(concussion_1, concussion_0)
contingency_table
```

```
##              Female   Male
## concussion_1    304    254
## concussion_0 354049 392966
```

2. Calculate risk (frequency) of concussions by sex, and the risk ratio comparing males to females.

```
risk_ratio_female <- contingency_table[[1]]/(contingency_table[[1]]+contingency_table[[2]])
risk_ratio_male <- contingency_table[[3]]/(contingency_table[[3]]+contingency_table[[4]])
risk_ratio_all <- risk_ratio_male / risk_ratio_female

paste('Risk Ratio Female:', round(risk_ratio_female, 5))
```

```
## [1] "Risk Ratio Female: 0.00086"
```

```
paste('Risk Ratio Male:', round(risk_ratio_male, 5))
```

```
## [1] "Risk Ratio Male: 0.00065"
```

```
paste('Risk Ratio All:', round(risk_ratio_all, 5))
```

```
## [1] "Risk Ratio All: 0.75294"
```

3. Apply Pearson's chi-square test to the contingency table.

```
chisq.test(contingency_table)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  contingency_table
## X-squared = 10.944, df = 1, p-value = 0.0009391
```

The result of Pearson's chi square test is

4. Use logistic regression to test if concussions are equally likely between males and females.

```
summary(glm(Y~Cov$Sex, family=binomial))
```

```
##
## Call:
## glm(formula = Y ~ Cov$Sex, family = binomial)
##
```

```
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -5.619  -2.044  -0.511   2.838   6.335
##
## Coefficients:
##             Estimate Std. Error  z value Pr(>|z|)
## (Intercept) -7.06016    0.05738 -123.045  < 2e-16 ***
## Cov$SexMale -0.28398    0.08504   -3.339  0.00084 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 329.33  on 29  degrees of freedom
## Residual deviance: 318.12  on 28  degrees of freedom
## AIC: 439.2
##
## Number of Fisher Scoring iterations: 5
```

5. Repeat the steps above substituting the variables sports for sex.

```
summary(glm(Y~Cov$Sport, family=binomial))
```

```
##
## Call:
## glm(formula = Y ~ Cov$Sport, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8179  -0.6600  -0.1153   0.6562   1.8720
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -7.39195    0.09094 -81.286  < 2e-16 ***
## Cov$SportGymnastics       -2.11359    1.00416  -2.105   0.0353 *
## Cov$SportLacrosse          0.74177    0.14585   5.086 3.66e-07 ***
## Cov$SportSoccer            1.02668    0.11017   9.319  < 2e-16 ***
## Cov$SportSoftball/Baseball -0.70103    0.13518  -5.186 2.15e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 329.327  on 29  degrees of freedom
## Residual deviance:  42.735  on 25  degrees of freedom
## AIC: 169.81
##
## Number of Fisher Scoring iterations: 6
```

6. Run a multivariable logistic regression of concusions by sex, sports and year.

```r
summary(glm(Y~Cov$Sex+Cov$Sport+Cov$Year, family=binomial))
```

```
##
## Call:
## glm(formula = Y ~ Cov$Sex + Cov$Sport + Cov$Year, family = binomial)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.15064  -0.54657  -0.03252   0.44586   2.43806
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -164.07641  103.00640  -1.593  0.11119
## Cov$SexMale                  -0.27240    0.08534  -3.192  0.00141 **
## Cov$SportGymnastics          -2.17567    1.00491  -2.165  0.03038 *
## Cov$SportLacrosse             0.76735    0.14603   5.255 1.48e-07 ***
## Cov$SportSoccer               1.02734    0.11018   9.324  < 2e-16 ***
## Cov$SportSoftball/Baseball   -0.70705    0.13575  -5.208 1.91e-07 ***
## Cov$Year                      0.07848    0.05155   1.522  0.12791
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 329.33  on 29  degrees of freedom
## Residual deviance:  29.76  on 23  degrees of freedom
## AIC: 160.84
##
## Number of Fisher Scoring iterations: 6
```

7. Report the adjusted odds ratios for sex and sports.

```r
summary(model <- glm(Y~Cov$Sex+Cov$Sport, family=binomial))
```

```
##
## Call:
## glm(formula = Y ~ Cov$Sex + Cov$Sport, family = binomial)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.21117  -0.51645  -0.07118   0.60101   2.05163
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -7.26040    0.09836 -73.817  < 2e-16 ***
## Cov$SexMale                -0.27759    0.08528  -3.255  0.00113 **
## Cov$SportGymnastics        -2.21535    1.00459  -2.205  0.02744 *
## Cov$SportLacrosse           0.76467    0.14603   5.237 1.64e-07 ***
## Cov$SportSoccer             1.02496    0.11017   9.303  < 2e-16 ***
## Cov$SportSoftball/Baseball -0.68881    0.13523  -5.094 3.51e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 329.327  on 29  degrees of freedom
## Residual deviance:  32.085  on 24  degrees of freedom
## AIC: 161.16
##
## Number of Fisher Scoring iterations: 6
```

```
exp(model$coef)
```

```
##                (Intercept)                Cov$SexMale
##               0.0007028248               0.7576045878
##         Cov$SportGymnastics          Cov$SportLacrosse
##               0.1091156798               2.1482782876
##             Cov$SportSoccer Cov$SportSoftball/Baseball
##               2.7869796209               0.5021719697
```

8. Test if there is an interaction of sex and sports.

```
summary(glm(Y~Cov$Sex*Cov$Sport, family=binomial))
```

```
##
## Call:
## glm(formula = Y ~ Cov$Sex * Cov$Sport, family = binomial)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.51537  -0.33429  -0.00009   0.35894   1.67369
##
## Coefficients:
##                                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)                            -7.20070    0.11790 -61.077  < 2e-16
## Cov$SexMale                            -0.41900    0.18525  -2.262   0.0237
## Cov$SportGymnastics                    -2.17574    1.00697  -2.161   0.0307
## Cov$SportLacrosse                       0.35691    0.22891   1.559   0.1190
## Cov$SportSoccer                         1.03907    0.14227   7.303 2.81e-13
## Cov$SportSoftball/Baseball             -0.84340    0.18757  -4.496 6.91e-06
## Cov$SexMale:Cov$SportGymnastics       -14.86887 3411.51584  -0.004   0.9965
## Cov$SexMale:Cov$SportLacrosse           0.72792    0.30407   2.394   0.0167
## Cov$SexMale:Cov$SportSoccer            -0.03789    0.22489  -0.169   0.8662
## Cov$SexMale:Cov$SportSoftball/Baseball  0.32869    0.27290   1.204   0.2284
##
## (Intercept)                            ***
## Cov$SexMale                            *
## Cov$SportGymnastics                    *
## Cov$SportLacrosse
## Cov$SportSoccer                        ***
## Cov$SportSoftball/Baseball             ***
## Cov$SexMale:Cov$SportGymnastics
## Cov$SexMale:Cov$SportLacrosse          *
## Cov$SexMale:Cov$SportSoccer
```

```
## Cov$SexMale:Cov$SportSoftball/Baseball
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 329.327  on 29  degrees of freedom
## Residual deviance:  22.198  on 20  degrees of freedom
## AIC: 159.28
##
## Number of Fisher Scoring iterations: 18
```

3 Simulate and Analyze

2. Explain why the estimate of the coefficient for X in the logistic regression adjusting for covariate Z1 (see below) is significantly different from zero despite the causal effect being zero?

```r
n = 2500
Z1 = rnorm(n)
Z2 = rnorm(n)
X = 0.7*rnorm(n) + 0.7*Z2
Lin = 0*X - 0.0*Z1 + 0.5*Z2 # causal model
Y = runif(n) < 1/(1+exp(-Lin))
summary(glm(Y ~ X + Z1, family=binomial))
```

```
##
## Call:
## glm(formula = Y ~ X + Z1, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7260  -1.1680   0.8167   1.1231   1.6750
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.08051    0.04086   1.971  0.04877 *
## X            0.40197    0.04246   9.467  < 2e-16 ***
## Z1           0.10548    0.04049   2.605  0.00919 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3461.6  on 2499  degrees of freedom
## Residual deviance: 3360.3  on 2497  degrees of freedom
## AIC: 3366.3
##
## Number of Fisher Scoring iterations: 4
```

From the above equation, we can observe that the value of X is determined by Z2 and some noises through rnorm and the value of Y is dependent Z2, multiplied by its coefficient (there are X and Z1 variables but the coefficient is 0). The coefficient estimate of X from the model result is significantly different from zero because both X and Y are dependent on Z2.