Exam3 - Fall 2020 - Mat 465/565

Theory problem:

(10 pts) MLE for estimating the probability π in a null model.

In the notes I have expalined how the parameters for the model (the β's) are obtained by maximizing the log likelihood function.

a) Write down the log likelihood for the null model, where all the probabilities are the same: $\pi_i = \pi = \frac{e^{\beta_0}}{1+e^{\beta_0}}$. That is, follow the notes and replace the π_i's by π.

Answer:

b) Find $\frac{\partial \log L}{\partial \pi}$ and use it to find the estimator for π. Hint: follow the notes.

Answer:

If you are at a loss with this problem, ask for help.

Coding problem:

Here are the variables that MZines4You.com has on each customer from third-party sources:

- Household Income (Income; rounded to the nearest \$1,000.00)
- Gender (IsFemale = 1 if the person is female, 0 otherwise)
- Marital Status (IsMarried = 1 if married, 0 otherwise)
- College Educated (HasCollege = 1 if has one or more years of college education, 0 otherwise)
- Employed in a Profession (IsProfessional = 1 if employed in a profession, 0 otherwise)
- Retired (IsRetired = 1 if retired, 0 otherwise)
- Not employed (Unemployed = 1 if not employed, 0 otherwise)
- Length of Residency in Current City (ResLength; in years)
- Dual Income if Married (Dual = 1 if dual income, 0 otherwise)
- Children (Minors = 1 if children under 18 are in the household, 0 otherwise)
- Home ownership (Own = 1 if own residence, 0 otherwise)
- Resident type (House = 1 if residence is a single family house, 0 otherwise)
- Race (White = 1 if race is white, 0 otherwise)
- Language (English = 1 is the primary language in the household is English, 0 otherwise)

Your task is to develop such an equation for one magazine ("Kid Creative") whose target audience are children between the ages of 9 and 12. In the process of sending out the "experimental" e-mails, the ad for "Kid Creative" was shown in 673 e-mails to customers and the purchase behavior recorded.

In addition to the variables for each customer listed above (the ones obtained from 3rd party sources), Mzines4You.com has the following variables from their own databases:

- Previously purchased a parenting magazine (PrevParent = 1 if previously purchased a parenting magazine, 0 otherwise).
- Previously purchased a children's magazine (PrevChild = 1 if previously purchased a children's magazine)

The dependent variable comes from the "experiment;" that is, from the 763 e-mails to customers containing the ad for "Kid Creative" and whether or not the customer purchased the magazine. That is, the dependent variable is

- Purchased "Kid Creative" (Buy = 1 if purchased "Kid Creative," 0 otherwise)

A. Load the dataset KidCreative.txt or KidCreative.xlsx

```r
kid <- read.csv("C:\\Users\\marsh\\Documents\\school\\amat565\\exam3\\KidCreative.csv", header = TRUE)
```

B. (10 pts) a. Obtain the MLE estimates for the coefficients of the logistic model and well as the corresponding odds ratios.

```r
kid.model <- glm(Buy ~ Income + IsFemale + IsMarried + HasCollege + IsProfessional + IsRetired + Unempl
summary(kid.model)
```

```
##
## Call:
## glm(formula = Buy ~ Income + IsFemale + IsMarried + HasCollege +
##     IsProfessional + IsRetired + Unemployed + ResidenceLength +
##     DualIncome + Minors + Own + House + White + English + PrevChildMag +
##     PrevParentMag, family = binomial, data = kid)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.36655  -0.08416  -0.00955  -0.00149   2.49038
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.791e+01  2.223e+00  -8.058 7.74e-16 ***
## Income           2.016e-04  2.359e-05   8.545  < 2e-16 ***
## IsFemale         1.646e+00  4.651e-01   3.539 0.000401 ***
## IsMarried        5.662e-01  5.864e-01   0.966 0.334272
## HasCollege      -2.794e-01  4.437e-01  -0.630 0.528962
## IsProfessional   2.253e-01  4.650e-01   0.485 0.627981
## IsRetired       -1.159e+00  9.323e-01  -1.243 0.214015
## Unemployed       9.886e-01  4.690e+00   0.211 0.833030
## ResidenceLength  2.468e-02  1.380e-02   1.788 0.073798 .
## DualIncome       4.518e-01  5.215e-01   0.866 0.386279
## Minors           1.133e+00  4.635e-01   2.444 0.014521 *
## Own              1.056e+00  5.594e-01   1.888 0.058976 .
## House           -9.265e-01  6.218e-01  -1.490 0.136238
## White            1.864e+00  5.454e-01   3.417 0.000632 ***
## English          1.530e+00  8.407e-01   1.821 0.068678 .
## PrevChildMag     1.557e+00  7.119e-01   2.188 0.028704 *
## PrevParentMag    4.777e-01  6.240e-01   0.766 0.443900
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 646.05  on 672  degrees of freedom
## Residual deviance: 182.33  on 656  degrees of freedom
## AIC: 216.33
##
## Number of Fisher Scoring iterations: 9
```

```r
exp(coef(kid.model))
```

```
##     (Intercept)          Income        IsFemale       IsMarried      HasCollege
##    1.665290e-08    1.000202e+00    5.186379e+00    1.761603e+00    7.562677e-01
##  IsProfessional       IsRetired      Unemployed ResidenceLength      DualIncome
##    1.252724e+00    3.139517e-01    2.687596e+00    1.024988e+00    1.571201e+00
##          Minors             Own           House           White         English
##    3.104578e+00    2.876122e+00    3.959276e-01    6.448342e+00    4.620394e+00
##    PrevChildMag   PrevParentMag
##    4.745742e+00    1.612413e+00
```

Should you keep the variable Income in this scale or should you scale it by dividing by 10,000's? Explain.

You should scale it by 10,000 because very large values will cause the sigmoid function to be very small. This can be a problem because the other coefficients will have to compensate for these large values.

b. Transform the variable Income by dividing it by 10,000. Call it myIncome Obtain the MLE estimated for the coefficients of the new logistic model and well as the corresponding odds ratios. Explain the effect of a unit change in the new variable income has on the odds ratio.

We can see the estimates of the coefficients have become larger and closer to zero.

```
myIncome<-kid$Income / 10000         # scaled income
fullmod<-glm(Buy ~ myIncome + IsFemale + IsMarried + HasCollege + IsProfessional + IsRetired + Unemploye
summary(fullmod)
```

```
##
## Call:
## glm(formula = Buy ~ myIncome + IsFemale + IsMarried + HasCollege +
##     IsProfessional + IsRetired + Unemployed + ResidenceLength +
##     DualIncome + Minors + Own + House + White + English + PrevChildMag +
##     PrevParentMag, family = binomial, data = kid)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.36655  -0.08416  -0.00955  -0.00149   2.49038
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -17.91068    2.22267  -8.058 7.74e-16 ***
## myIncome          2.01561    0.23588   8.545  < 2e-16 ***
## IsFemale          1.64604    0.46510   3.539 0.000401 ***
## IsMarried         0.56622    0.58643   0.966 0.334272
## HasCollege       -0.27936    0.44372  -0.630 0.528962
## IsProfessional    0.22532    0.46499   0.485 0.627981
## IsRetired        -1.15852    0.93233  -1.243 0.214015
## Unemployed        0.98865    4.68961   0.211 0.833030
## ResidenceLength   0.02468    0.01380   1.788 0.073798 .
## DualIncome        0.45184    0.52152   0.866 0.386279
## Minors            1.13288    0.46351   2.444 0.014521 *
## Own               1.05644    0.55945   1.888 0.058976 .
## House            -0.92652    0.62185  -1.490 0.136238
## White             1.86382    0.54540   3.417 0.000632 ***
## English           1.53048    0.84068   1.821 0.068678 .
## PrevChildMag      1.55725    0.71188   2.188 0.028704 *
## PrevParentMag     0.47773    0.62398   0.766 0.443900
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 646.05  on 672  degrees of freedom
## Residual deviance: 182.33  on 656  degrees of freedom
## AIC: 216.33
##
## Number of Fisher Scoring iterations: 9
```

```
exp(coef(fullmod))
```

```
##     (Intercept)        myIncome        IsFemale       IsMarried      HasCollege
##    1.665290e-08    7.505306e+00    5.186379e+00    1.761603e+00    7.562677e-01
##  IsProfessional       IsRetired      Unemployed ResidenceLength      DualIncome
##    1.252724e+00    3.139517e-01    2.687596e+00    1.024988e+00    1.571201e+00
##          Minors             Own           House           White         English
##    3.104578e+00    2.876122e+00    3.959276e-01    6.448342e+00    4.620394e+00
##    PrevChildMag   PrevParentMag
##    4.745742e+00    1.612413e+00
```

C. (10 pts) Run a Backwards selection procedure to simplify the model according to the AIC. Drop one variable at a time. You can use:

- drop1(model,IC="AIC")
- or simply step(, direction="backward") See how it was done is model selection files for regression. It is done in the similar way in glm()

```
step(fullmod, direction="backward")
```

```
## Start:  AIC=216.33
## Buy ~ myIncome + IsFemale + IsMarried + HasCollege + IsProfessional +
##     IsRetired + Unemployed + ResidenceLength + DualIncome + Minors +
##     Own + House + White + English + PrevChildMag + PrevParentMag
##
##                   Df Deviance    AIC
## - Unemployed       1   182.38 214.38
## - IsProfessional   1   182.56 214.56
## - HasCollege       1   182.73 214.73
## - PrevParentMag    1   182.91 214.91
## - DualIncome       1   183.08 215.08
## - IsMarried        1   183.27 215.27
## - IsRetired        1   183.89 215.89
## <none>                 182.33 216.33
## - House            1   184.56 216.56
## - ResidenceLength  1   185.60 217.60
## - English          1   185.71 217.71
## - Own              1   185.92 217.92
## - PrevChildMag     1   187.48 219.48
## - Minors           1   188.73 220.73
## - White            1   195.34 227.34
## - IsFemale         1   197.10 229.10
## - myIncome         1   455.67 487.67
##
## Step:  AIC=214.38
## Buy ~ myIncome + IsFemale + IsMarried + HasCollege + IsProfessional +
##     IsRetired + ResidenceLength + DualIncome + Minors + Own +
##     House + White + English + PrevChildMag + PrevParentMag
##
##                   Df Deviance    AIC
## - IsProfessional   1   182.60 212.60
## - HasCollege       1   182.76 212.76
## - PrevParentMag    1   182.96 212.96
## - DualIncome       1   183.13 213.13
## - IsMarried        1   183.30 213.30
## - IsRetired        1   183.95 213.95
## <none>                 182.38 214.38
## - House            1   184.59 214.59
## - ResidenceLength  1   185.67 215.67
## - English          1   185.79 215.79
## - Own              1   185.94 215.94
## - PrevChildMag     1   187.52 217.52
## - Minors           1   188.84 218.84
## - White            1   195.43 225.43
## - IsFemale         1   197.22 227.22
## - myIncome         1   456.12 486.12
##
## Step:  AIC=212.6
## Buy ~ myIncome + IsFemale + IsMarried + HasCollege + IsRetired +
##     ResidenceLength + DualIncome + Minors + Own + House + White +
##     English + PrevChildMag + PrevParentMag
##
##                   Df Deviance    AIC
## - HasCollege       1   182.84 210.84
## - PrevParentMag    1   183.10 211.10
## - DualIncome       1   183.46 211.46
## - IsMarried        1   183.46 211.46
## <none>                 182.60 212.60
## - IsRetired        1   184.87 212.87
## - House            1   184.94 212.94
## - ResidenceLength  1   185.76 213.76
## - Own              1   186.35 214.35
## - English          1   186.55 214.55
## - PrevChildMag     1   187.71 215.71
## - Minors           1   188.87 216.87
## - White            1   195.43 223.43
## - IsFemale         1   197.23 225.23
## - myIncome         1   463.98 491.98
##
## Step:  AIC=210.84
## Buy ~ myIncome + IsFemale + IsMarried + IsRetired + ResidenceLength +
##     DualIncome + Minors + Own + House + White + English + PrevChildMag +
##     PrevParentMag
##
##                   Df Deviance    AIC
## - PrevParentMag    1   183.30 209.30
## - DualIncome       1   183.63 209.63
## - IsMarried        1   183.71 209.71
## <none>                 182.84 210.84
## - House            1   185.06 211.06
## - IsRetired        1   185.18 211.18
## - ResidenceLength  1   186.03 212.03
## - Own              1   186.37 212.37
## - English          1   186.62 212.62
## - PrevChildMag     1   188.20 214.20
## - Minors           1   189.58 215.58
## - White            1   195.98 221.98
## - IsFemale         1   197.67 223.67
## - myIncome         1   476.05 502.05
## 
## Step:  AIC=209.3
## Buy ~ myIncome + IsFemale + IsMarried + IsRetired + ResidenceLength + 
##     DualIncome + Minors + Own + House + White + English + PrevChildMag
## 
##                   Df Deviance    AIC
## - IsMarried        1   184.04 208.04
## - DualIncome       1   184.33 208.33
## <none>                 183.30 209.30
## - House            1   185.67 209.67
## - IsRetired        1   185.80 209.80
## - ResidenceLength  1   186.56 210.56
## - English          1   187.03 211.03
## - Own              1   187.14 211.14
## - PrevChildMag     1   188.79 212.79
## - Minors           1   189.93 213.93
## - White            1   196.71 220.71
## - IsFemale         1   197.98 221.98
## - myIncome         1   477.45 501.45
## 
## Step:  AIC=208.04
## Buy ~ myIncome + IsFemale + IsRetired + ResidenceLength + DualIncome + 
##     Minors + Own + House + White + English + PrevChildMag
## 
##                   Df Deviance    AIC
## <none>                 184.04 208.04
## - IsRetired        1   186.24 208.24
## - House            1   186.38 208.38
## - DualIncome       1   187.46 209.46
## - ResidenceLength  1   187.50 209.50
## - English          1   188.12 210.12
## - PrevChildMag     1   189.83 211.83
## - Own              1   190.45 212.45
## - Minors           1   191.98 213.98
## - White            1   197.48 219.48
## - IsFemale         1   198.68 220.68
## - myIncome         1   480.10 502.10

## 
## Call:  glm(formula = Buy ~ myIncome + IsFemale + IsRetired + ResidenceLength + 
##     DualIncome + Minors + Own + House + White + English + PrevChildMag, 
##     family = binomial, data = kid)
## 
## Coefficients:
##     (Intercept)         myIncome         IsFemale        IsRetired  
##       -17.69848          1.99159          1.60536         -1.24541  
## ResidenceLength       DualIncome           Minors              Own  
##         0.02501          0.76534          1.20598          1.24178  
##           House            White          English     PrevChildMag  
##        -0.93442          1.86036          1.62270          1.63456  
```

```
##
## Degrees of Freedom: 672 Total (i.e. Null);  661 Residual
## Null Deviance:       646.1
## Residual Deviance: 184   AIC: 208
```

```
reducedmod <- glm(formula = Buy ~ myIncome + IsFemale + IsRetired + ResidenceLength +
    DualIncome + Minors + Own + House + White + English + PrevChildMag,
    family = binomial, data = kid)
summary(reducedmod)
```

```
##
## Call:
## glm(formula = Buy ~ myIncome + IsFemale + IsRetired + ResidenceLength +
##     DualIncome + Minors + Own + House + White + English + PrevChildMag,
##     family = binomial, data = kid)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.35528  -0.08724  -0.01059  -0.00176   2.54322
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -17.69848    2.17596  -8.134 4.17e-16 ***
## myIncome          1.99159    0.23011   8.655  < 2e-16 ***
## IsFemale          1.60536    0.45310   3.543 0.000396 ***
## IsRetired        -1.24541    0.84408  -1.475 0.140088
## ResidenceLength   0.02501    0.01363   1.835 0.066575 .
## DualIncome        0.76534    0.41801   1.831 0.067116 .
## Minors            1.20598    0.44406   2.716 0.006611 **
## Own               1.24178    0.50045   2.481 0.013089 *
## House            -0.93442    0.61377  -1.522 0.127903
## White             1.86036    0.53274   3.492 0.000479 ***
## English           1.62270    0.81172   1.999 0.045599 *
## PrevChildMag      1.63456    0.71167   2.297 0.021630 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 646.05  on 672  degrees of freedom
## Residual deviance: 184.04  on 661  degrees of freedom
## AIC: 208.04
##
## Number of Fisher Scoring iterations: 8
```

```
library(car)
```

```
## Loading required package: carData
```

```
diffdev<-summary(reducedmod)$deviance -summary(fullmod)$deviance
print(paste('Difference in deviances: ',diffdev))
```

```
## [1] "Difference in deviances:  1.71246691045314"
```

```
diffdf<-summary(reducedmod)$df.residual -summary(fullmod)$df.residual
print(paste('Difference in degrees of freedom: ',diffdf))
```

```
## [1] "Difference in degrees of freedom:  5"
```

```
print(paste('p-value: ',1-pchisq(diffdev,diffdf)))
```

```
## [1] "p-value:  0.887325412717683"
```

D. (10 pts) Once you have your final model in part C, run a Wald test (deviance test) to compare the full model to your new simplified model. State the null hypothesis and the alternative hypothesis of this test. Explain how deviance is calculated and how this test works.

Answer:

H_0: simplified model, B_k = 0 where B_k is in the full model but not the simplified model.

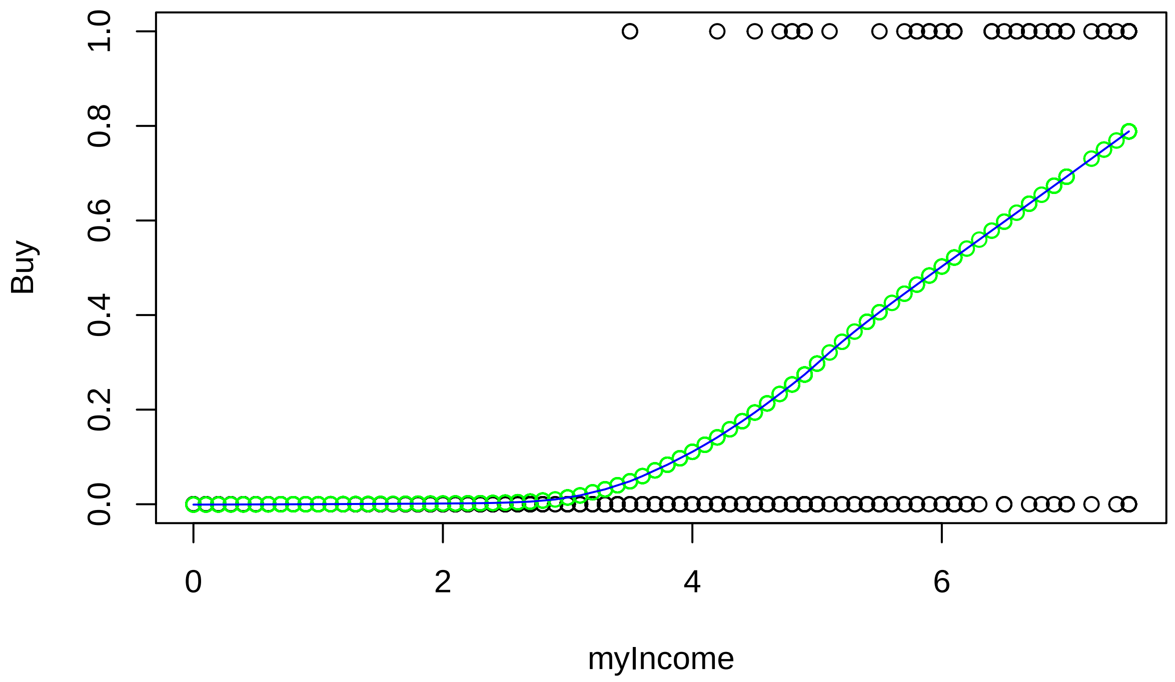
H_1: full model, B_k != 0

Chi-sq = 1.71246691045314

Conclussion: We reject the alternative hypothesis, since the p-value is large and the extra variables in the full model are not significant.

E. (5 pts) Make a scatterplot of the response variable on myIncome, with the fitted logistic response function from the model you obtained in D, together with a lowess smooth superimposed.

```
plot(Buy ~ myIncome, data=kid)
points(lowess(myIncome,reducedmod$fitted),col='green')
lines(lowess(myIncome,reducedmod$fitted),col='blue')  # enter the fitted values from your model
```



F. (5 pts) Obtain a 95% confidence interval for the coefficient of Income as well as for its exponentiated value (odds ratio). State what is the statistic of this test.

```
confint(reducedmod, parm=('myIncome'))
```

```
## Waiting for profiling to be done...
```

```
##    2.5 %   97.5 %
## 1.584421 2.492677
```

```
exp(confint(reducedmod, parm=('myIncome')))
```

```
## Waiting for profiling to be done...
```

```
##     2.5 %    97.5 %
##  4.876466 12.093604
```

G. (5 pts) Write down the equation for the predicted probabilities according to your model. What is the estimated probability that a female with an income of 68,000 will buy the Kids Creative magazine if: she is Married, has College education, is not Professional, is not Retired, is not Unemployed, has lived 3 years in he current city, rents an apartment, her home has Dual Income, has one child, she is White, speaks English, has never bought a Previous Child Magazine not a Parent Magazine.

Answers:

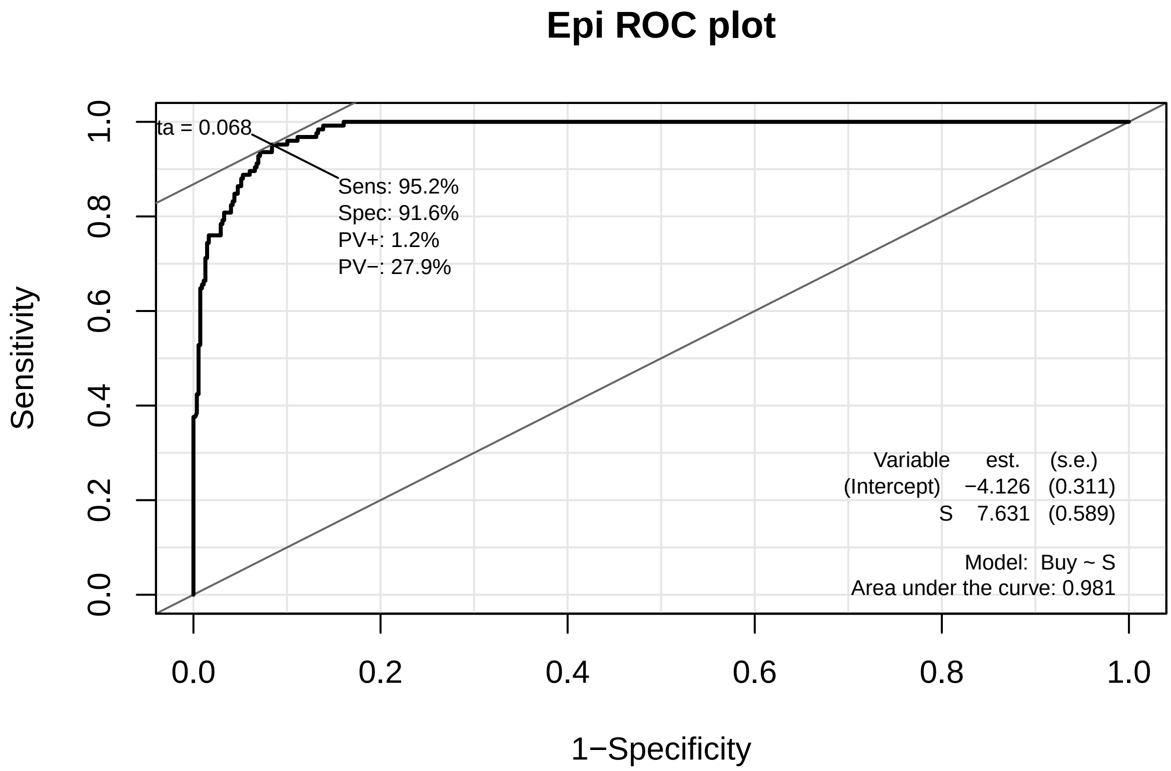
```
x = c(6.8, 1, 0, 3, 1, 1, 0, 0, 1, 1, 0)
pred = 1 / (1 + exp(-17.69848 + 1.99159*x[1] + 1.60536*x[2] + -1.24541*x[3] + 0.02501*x[4] + 0.76534*x[5
pred
```

```
## [1] 0.04837895
```

H. (15 pts) A prediction rule is to be developed.

H-a: Draw the ROC curve for your model.

```
library(Epi)
S<-predict(reducedmod,type='response')
ROC(form=Buy~S,plot="ROC",PV=TRUE,MX=TRUE,AUC=TRUE,data=kid,main="Epi ROC plot")
```



H-b. Find the sensitivity and specificity for the cutoffs: .1, .2, .3, .4, .5, .6

The following computes sensitivity and specificity for the predictions from a logistic model, at a threshold s:

```
Ps=(model$fit>s)*1
TN=sum((Ps==0)*(Y==0))/sum(Y==0)     #specificity
TP=sum((Ps==1)*(Y==1))/sum(Y==1)     #sensitivity
```

Modify that code as needed to do your computations.

```
Ps=(reducedmod$fit>0.1)*1
TN=sum((Ps==0)*(kid$Buy==0))/sum(kid$Buy==0)     #specificity
TP=sum((Ps==1)*(kid$Buy==1))/sum(kid$Buy==1)     #sensitivity
print('cutoff = 0.1')
```

```
## [1] "cutoff = 0.1"
```

```
print(paste('Sensitivity: ',TP))
```

```
## [1] "Sensitivity:  0.968"
```

```
print(paste('Specificity: ',TN))
```

```
## [1] "Specificity:  0.87956204379562"
```

```
Ps=(reducedmod$fit>0.2)*1
TN=sum((Ps==0)*(kid$Buy==0))/sum(kid$Buy==0)     #specificity
TP=sum((Ps==1)*(kid$Buy==1))/sum(kid$Buy==1)     #sensitivity
print('cutoff = 0.2')
```

```
## [1] "cutoff = 0.2"
```

```r
print(paste('Sensitivity: ',TP))
```

```
## [1] "Sensitivity:  0.952"
```

```r
print(paste('Specificity: ',TN))
```

```
## [1] "Specificity:  0.916058394160584"
```

```r
Ps=(reducedmod$fit>0.3)*1
TN=sum((Ps==0)*(kid$Buy==0))/sum(kid$Buy==0)    #specificity
TP=sum((Ps==1)*(kid$Buy==1))/sum(kid$Buy==1)    #sensitivity
print('cutoff = 0.3')
```

```
## [1] "cutoff = 0.3"
```

```r
print(paste('Sensitivity: ',TP))
```

```
## [1] "Sensitivity:  0.912"
```

```r
print(paste('Specificity: ',TN))
```

```
## [1] "Specificity:  0.930656934306569"
```

```r
Ps=(reducedmod$fit>0.4)*1
TN=sum((Ps==0)*(kid$Buy==0))/sum(kid$Buy==0)    #specificity
TP=sum((Ps==1)*(kid$Buy==1))/sum(kid$Buy==1)    #sensitivity
print('cutoff = 0.4')
```

```
## [1] "cutoff = 0.4"
```

```r
print(paste('Sensitivity: ',TP))
```

```
## [1] "Sensitivity:  0.88"
```

```r
print(paste('Specificity: ',TN))
```

```
## [1] "Specificity:  0.948905109489051"
```

```r
Ps=(reducedmod$fit>0.5)*1
TN=sum((Ps==0)*(kid$Buy==0))/sum(kid$Buy==0)    #specificity
TP=sum((Ps==1)*(kid$Buy==1))/sum(kid$Buy==1)    #sensitivity
print('cutoff = 0.5')
```

```
## [1] "cutoff = 0.5"
```

```r
print(paste('Sensitivity: ',TP))
```

```
## [1] "Sensitivity:  0.848"
```

```r
print(paste('Specificity: ',TN))
```

```
## [1] "Specificity:  0.956204379562044"
```

```r
Ps=(reducedmod$fit>0.6)*1
TN=sum((Ps==0)*(kid$Buy==0))/sum(kid$Buy==0)    #specificity
TP=sum((Ps==1)*(kid$Buy==1))/sum(kid$Buy==1)    #sensitivity
print('cutoff = 0.6')
```

```
## [1] "cutoff = 0.6"
```

```r
print(paste('Sensitivity: ',TP))
```

```
## [1] "Sensitivity:  0.76"
```

```
print(paste('Specificity: ',TN))
```

```
## [1] "Specificity:  0.974452554744526"
```

H-c. Combining this information with the ROC curve abouve, which threshold is recommended?

Answer: 0.2

H4. As the threshold increases: sensitivity **decreases**_ (increases/decreases) specificity **increases**_ (increases/decreases)