Mathew Goberdhan - Exam3 - Fall 2021 - Mat 465/565

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

```
KidCreative <- read.delim("~/Documents/KidCreative.txt")
View(KidCreative)
```

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

```
logmodel<-glm(Buy~.,data=KidCreative,family=binomial())
#summary(logmodel)
oddsratio<-exp(logmodel$coefficients)

#Run last two lines together to display both estimates and corresponding odds ratios

logmodel$coefficients
```

```
##     (Intercept)          Income        IsFemale       IsMarried      HasCollege
##   -17.910681740     0.000201561     1.646035848     0.566224252    -0.279359899
##  IsProfessional       IsRetired      Unemployed ResidenceLength      DualIncome
##     0.225320058    -1.158516131     0.988647292     0.024680817     0.451840610
##          Minors             Own           House           White         English
##     1.132877868     1.056442728    -0.926524019     1.863823021     1.530480050
##    PrevChildMag   PrevParentMag
##     1.557247733     0.477731505
```

```
oddsratio
```

```
##     (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.

ANSWER: Scaling Income by dividing it by 10,000 will making results easier to interpret as it will make data more readable. For example, scaling 68,000 to 68 makes data easier to work with.

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.

ANSWER: Before the transformation, the odds ratio of the variable Income was approximately 1 (since the MLE estimate was 'close' to 0). After the transformation, the odds ratio of the new variable myIncome had a 7.5-fold increase.

```
myIncome<-KidCreative$Income / 10000 # scaled income
#myIncome
KidCreative$Income<-myIncome #set variable to new transformation
#KidCreative$Income
fullmod<-glm(Buy~.,data=KidCreative,family=binomial()) # full glm model with myIncome instead of Income
#summary(fullmod)
oddsratiofull<-exp(fullmod$coefficients)

#Run last two lines together to display both estimates and corresponding odds ratios

fullmod$coefficients
```

```
##     (Intercept)          Income        IsFemale       IsMarried       HasCollege
##    -17.91068174      2.01561024      1.64603585      0.56622425      -0.27935990
##  IsProfessional       IsRetired      Unemployed ResidenceLength       DualIncome
##      0.22532006     -1.15851613      0.98864729      0.02468082       0.45184061
##          Minors             Own           House           White          English
##      1.13287787      1.05644273     -0.92652402      1.86382302       1.53048005
##     PrevChildMag   PrevParentMag
##      1.55724773      0.47773151
```

```r
oddsratiofull
```

```
##     (Intercept)          Income        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 works in a similar way in glm()

```r
step(fullmod,direction='backward')
```

```
## Start:  AIC=216.33
## Buy ~ Income + 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
## - Income           1   455.67 487.67
```

```
##
## Step:  AIC=214.38
## Buy ~ Income + 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
## - Income           1   456.12 486.12
##
## Step:  AIC=212.6
## Buy ~ Income + 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
## - Income           1   463.98 491.98
##
## Step:  AIC=210.84
## Buy ~ Income + 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
## - Income           1   476.05 502.05
##
## Step:  AIC=209.3
## Buy ~ Income + 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
## - Income           1   477.45 501.45
##
## Step:  AIC=208.04
## Buy ~ Income + 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
## - Income           1   480.10 502.10
```

```
##
## Call:  glm(formula = Buy ~ Income + IsFemale + IsRetired + ResidenceLength +
##     DualIncome + Minors + Own + House + White + English + PrevChildMag,
##     family = binomial(), data = KidCreative)
```

```
##
## Coefficients:
##     (Intercept)           Income         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
```

```r
simmodel<-glm(formula = Buy ~ Income + IsFemale + IsRetired + ResidenceLength + DualIncome + Minors + Ow

summary(simmodel)
```

```
##
## Call:
## glm(formula = Buy ~ Income + IsFemale + IsRetired + ResidenceLength +
##     DualIncome + Minors + Own + House + White + English + PrevChildMag,
##     family = binomial(), data = KidCreative)
##
## 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 ***
## Income            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
```

```
anova(simmodel,fullmod)
```

```
## Analysis of Deviance Table
##
## Model 1: Buy ~ Income + IsFemale + IsRetired + ResidenceLength + DualIncome +
##     Minors + Own + House + White + English + PrevChildMag
## Model 2: Buy ~ Income + IsFemale + IsMarried + HasCollege + IsProfessional +
##     IsRetired + Unemployed + ResidenceLength + DualIncome + Minors +
##     Own + House + White + English + PrevChildMag + PrevParentMag
##   Resid. Df Resid. Dev Df Deviance
## 1       661     184.04
## 2       656     182.33  5   1.7125
```

```
pchisq(1.7125,5)
```

```
## [1] 0.1126788
```

```
1-pchisq(1.7125,5)
```

```
## [1] 0.8873212
```

D. (10 pts) Once you have your final model in part C, run a 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: simpler model (simmodel)

H_1: fuller model (fullmod)

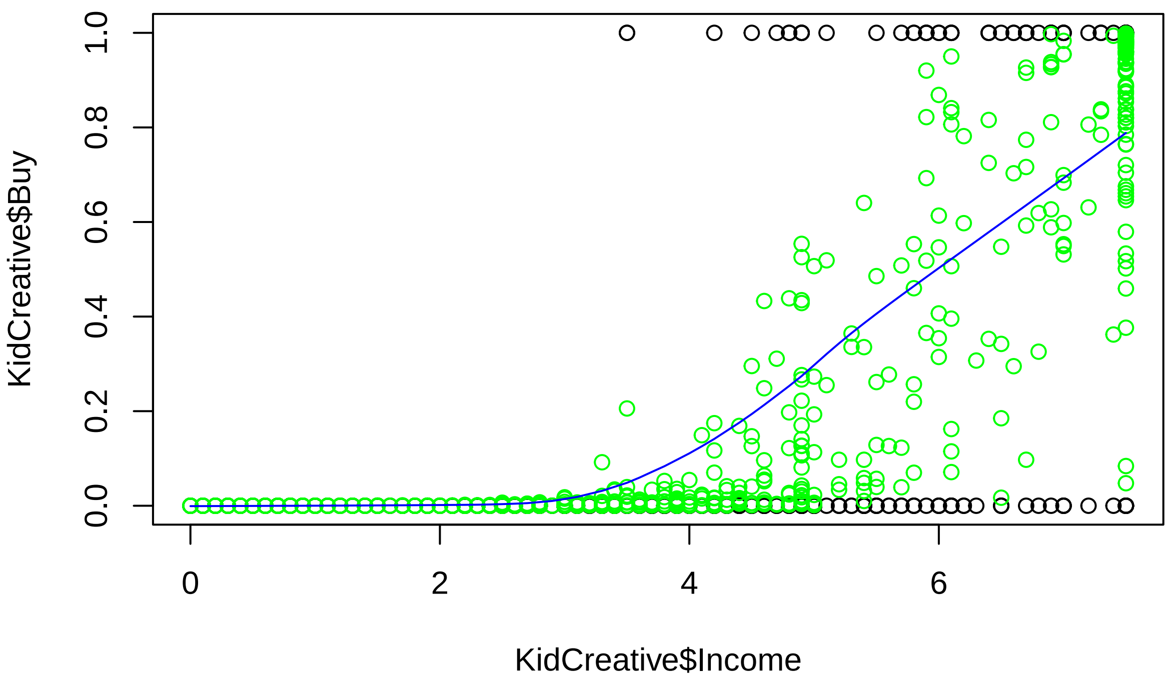
Chi-sq = 1.7125

Conclusion: Our p-value is 0.8873212, so we fail to reject the null hypothesis, therefore we keep our simpler model, i.e, simmodel.

Deviance is calculated as follows: -2([log_like of reduced model] - [log_like of full model]) = reduced deviance - full deviance. The greater the deviance, the worse the model fits in comparison to the full model.

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(KidCreative$Income,KidCreative$Buy)
points(KidCreative$Income,simmodel$fitted.values,col='green')
lines(lowess(KidCreative$Income,simmodel$fitted.values),col='blue')  # enter the fitted values from you
```



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

ANSWER: z-score.

```
cbind(coef=coef(simmodel),oddsratio=exp(simmodel$coefficients),confint(simmodel))
```

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

```
##                        coef    oddsratio        2.5 %       97.5 %
## (Intercept)     -17.69848331 2.058953e-08 -22.41843230 -13.83400389
## Income            1.99158727 7.327155e+00   1.58442076   2.49267671
## IsFemale          1.60535517 4.979628e+00   0.75619031   2.54375342
## IsRetired        -1.24541428 2.878216e-01  -2.93356310   0.40150676
## ResidenceLength   0.02500784 1.025323e+00  -0.00133431   0.05243702
## DualIncome        0.76533711 2.149719e+00  -0.04521326   1.60333998
## Minors            1.20597933 3.340028e+00   0.35967388   2.10985735
## Own               1.24178111 3.461774e+00   0.27736330   2.24966598
## House            -0.93442363 3.928122e-01  -2.15641775   0.26351741
## White             1.86036131 6.426058e+00   0.84708165   2.94593188
## English           1.62269570 5.066730e+00   0.04883304   3.27511257
## PrevChildMag      1.63455938 5.127198e+00   0.29330161   3.09600329
```

G. (5 pts) Write down the equation for the predicted probabilities according to your model.

Answer:

OR = exp(-17.698+1.992(Income)+1.605(IsFemale)-1.245(IsRetired)+0.025(ResidenceLength)+.765(DualIncome)+1.206(Mi
0.934(House)+1.860(White)+1.623(English)+1.635(PrevChildMag))

Predicted Probabilities = OR / (1+OR)

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 amParent Magazine.

Answer:

OR_est = exp(-17.698+1.992(6.8)+1.605(1)+0.025(3)+0.765(1)+1.206(1)+1.860(1)+1.623(1))

Estimated Probablity = OR_est/(1+OR_est)

```
OR_est = exp(-17.698+1.992*(6.8)+1.605*(1)+0.025*(3)+0.765*(1)+1.206*(1)+1.860*(1)+1.623*(1))
OR_est
```

```
## [1] 19.71934
```

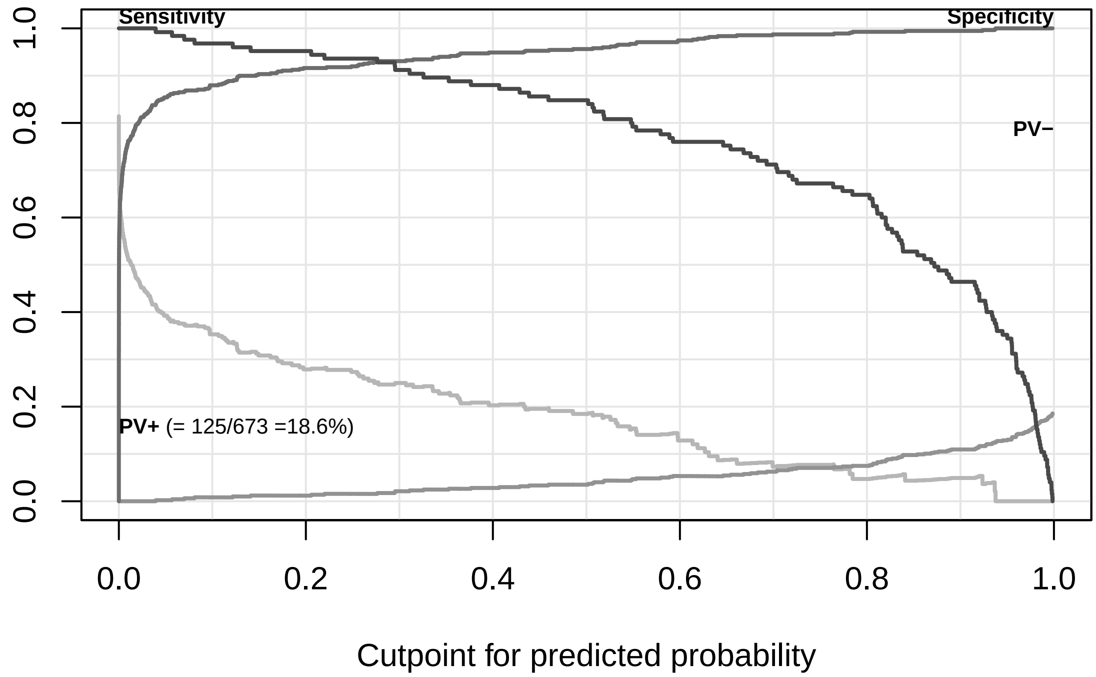
```
OR_est/(1+OR_est)
```

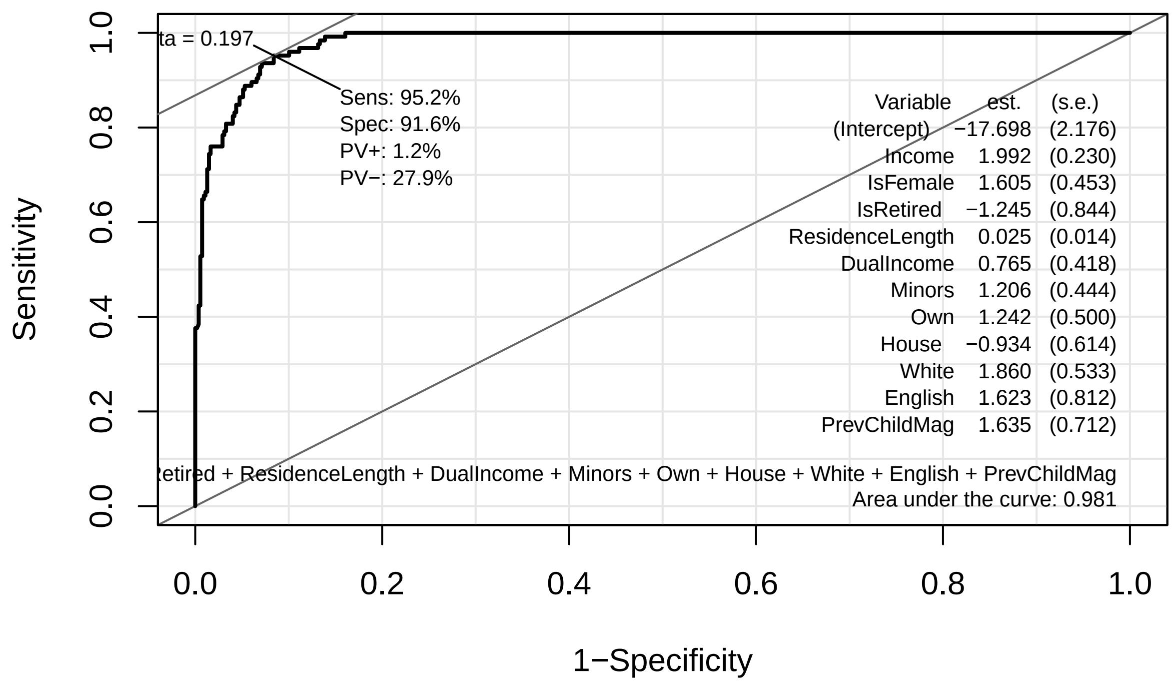
```
## [1] 0.9517359
```

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

H-a: Draw the ROC curve for your model. (Use the ROC curve from the Epi library)

```
library(Epi)
ROC(form=simmodel$formula,data=KidCreative)
```


Sensitivity
Specificity
PV−
PV+ (= 125/673 =18.6%)
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Cutpoint for predicted probability



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.

```
Y<-KidCreative$Buy
model<-simmodel
Ps1=(model$fit>.1)*1
TN1=sum((Ps1==0)*(Y==0))/sum(Y==0)
TP1=sum((Ps1==1)*(Y==1))/sum(Y==1)
Ps2=(model$fit>.2)*1
TN2=sum((Ps2==0)*(Y==0))/sum(Y==0)
TP2=sum((Ps2==1)*(Y==1))/sum(Y==1)
Ps3=(model$fit>.3)*1
TN3=sum((Ps3==0)*(Y==0))/sum(Y==0)
TP3=sum((Ps3==1)*(Y==1))/sum(Y==1)
Ps4=(model$fit>.4)*1
TN4=sum((Ps4==0)*(Y==0))/sum(Y==0)
TP4=sum((Ps4==1)*(Y==1))/sum(Y==1)
Ps5=(model$fit>.5)*1
TN5=sum((Ps5==0)*(Y==0))/sum(Y==0)
```

```
TP5=sum((Ps5==1)*(Y==1))/sum(Y==1)
Ps6=(model$fit>.6)*1
TN6=sum((Ps6==0)*(Y==0))/sum(Y==0)
TP6=sum((Ps6==1)*(Y==1))/sum(Y==1)

sens=cbind(TP1,TP2,TP3,TP4,TP5,TP6)
spec=cbind(TN1,TN2,TN3,TN4,TN5,TN6)

rbind(sens,spec)
```

```
##           TP1       TP2       TP3       TP4       TP5       TP6
## [1,] 0.968000 0.9520000 0.9120000 0.8800000 0.8480000 0.7600000
## [2,] 0.879562 0.9160584 0.9306569 0.9489051 0.9562044 0.9744526
```

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

Answer: Threshold recommended is 0.2

H4. As the threshold increases: sensitivity _______ (decreases) specificity _______ (increases)