Exam 3 - Take Home Exam- DANA YOUNG

A. Load the dataset KidCreative.txt

B. (10pts) i. Obtain the MLE estimates for the coefficients of the logistic model and well as the corresponding odds ratios. Should you keep the variable Income in this scale of should you scale it by dividing by 10,000's? Explain.

```
#checking for NA's
print(paste0("Amount of NA's:  ",sum(is.na(KidCreative))))
```

```
## [1] "Amount of NA's:  0"
```

```
#next we will create a full model
logmod<- glm(Buy~., family=binomial(), data=KidCreative)
cbind(MLE.estimates= coef(logmod), odds.ratio= exp(logmod$coef))
```

```
##                  MLE.estimates   odds.ratio
## (Intercept)     -17.910681740 1.665290e-08
## Income            0.000201561 1.000202e+00
## IsFemale          1.646035848 5.186379e+00
## IsMarried         0.566224252 1.761603e+00
## HasCollege       -0.279359899 7.562677e-01
## IsProfessional    0.225320058 1.252724e+00
## IsRetired        -1.158516131 3.139517e-01
## Unemployed        0.988647292 2.687596e+00
## ResidenceLength   0.024680817 1.024988e+00
## DualIncome        0.451840610 1.571201e+00
## Minors            1.132877868 3.104578e+00
## Own               1.056442728 2.876122e+00
## House            -0.926524019 3.959276e-01
## White             1.863823021 6.448342e+00
## English           1.530480050 4.620394e+00
## PrevChildMag      1.557247733 4.745742e+00
## PrevParentMag     0.477731505 1.612413e+00
```

In this case we should scale the variable income by diving by 10,000. We would do this because the other variables are either 0 or 1 and income is in thousands, this causes the coefficent for income to become very tiny compared to the other variables. Re-scaling allows it to be more inline with the other variables.

ii. Transform the variable Income accordingly. 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.

```
KidCreative$Income <- KidCreative$Income/10000
attach(KidCreative)
```

```
## The following objects are masked from KidCreative (pos = 3):
##
##     Buy, DualIncome, English, HasCollege, House, Income, IsFemale,
##     IsMarried, IsProfessional, IsRetired, Minors, Own,
##     PrevChildMag, PrevParentMag, ResidenceLength, Unemployed,
##     White
```

```
logmod<- glm(Buy~., family=binomial(), data=KidCreative)
#summary(logmod)
#Below is a table with MLE estimates and odds ratio
cbind(MLE.estimates= coef(logmod), odds.ratio= exp(logmod$coef))
```

```
##                 MLE.estimates   odds.ratio
## (Intercept)      -17.91068174 1.665290e-08
## Income             2.01561024 7.505306e+00
## IsFemale           1.64603585 5.186379e+00
## IsMarried          0.56622425 1.761603e+00
## HasCollege        -0.27935990 7.562677e-01
## IsProfessional     0.22532006 1.252724e+00
## IsRetired         -1.15851613 3.139517e-01
## Unemployed         0.98864729 2.687596e+00
## ResidenceLength    0.02468082 1.024988e+00
## DualIncome         0.45184061 1.571201e+00
## Minors             1.13287787 3.104578e+00
## Own                1.05644273 2.876122e+00
## House             -0.92652402 3.959276e-01
## White              1.86382302 6.448342e+00
## English            1.53048005 4.620394e+00
## PrevChildMag       1.55724773 4.745742e+00
## PrevParentMag      0.47773151 1.612413e+00
```

The unit change in the variable Income allows the variable's odds ratio to be more in line with the other variables odds ratios.

```
summary(logmod)
```

```
##
## Call:
## glm(formula = Buy ~ ., family = binomial(), data = KidCreative)
##
## 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 ***
## Income            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
```

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")

```
stepAIC(logmod, 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
```

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

```
##
## 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
```

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.

```
anova(newlog, logmod, test= "Chisq")
```

```
## 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 Pr(>Chi)
## 1       661     184.04
## 2       656     182.33  5   1.7125   0.8873
```

```
AIC(newlog,logmod)
```

```
##        df      AIC
## newlog 12 208.0408
## logmod 17 216.3284
```

```
chisquare<- 184.04-182.33
chisquare
```

```
## [1] 1.71
```

```
df<- 5
1-pchisq(chisquare,df)
```

```
## [1] 0.8876375
```

For a deviance test the hypothesis are Test:

H0: Null model P(Y=1)=e^(β_0)/(1+e^(β_0)) (new simplified model)

H1: Model with variables P(Y=1)=e^(β_0+β_1 X_1+⋯+β_r X_r)/(1+e^(β_0+β_1 X_1+⋯+β_r X_r)) (full model)

Test statistic: (Simplified model deviance) – (Full model deviance)
=-2(LogL(Simplified model) – LogL(Full model deviance))
= 2(LogL(Full model deviance) – LogL(Simplified model)) (1)

Test statistic is approximately Chi-square with df=number of variables added p-value: P(Chi-sq>value we got in (1))

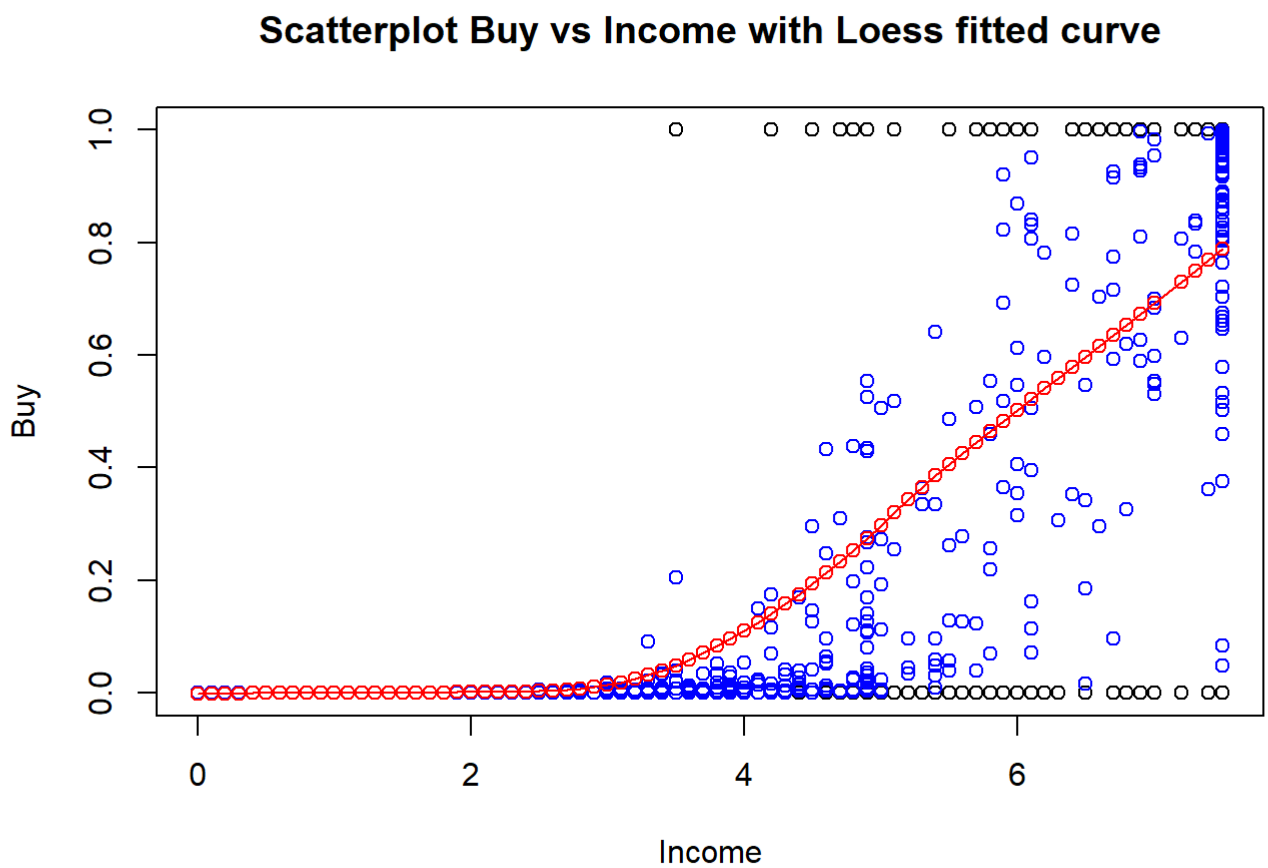
The deviance is calculated from Deviance= −2logL(model)

This test is used to determine which model is the better model.

The chi-square of 1.71 with 5 degrees of freedom and an associated p-value of 0.8876 tells us that our reduced model fits better then our full model.

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

```
#attach(KidCreative)
plot(Buy~Income, data=KidCreative, main="Scatterplot Buy vs Income with Loess fitted cur
ve")
points(Income, newlog$fitted, col='blue')
points(lowess(KidCreative$Income,newlog$fitted), col ='red')
lines(lowess(KidCreative$Income,newlog$fitted), col ='red')
```



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.

```
cbind(coefficient.Income = confint(newlog,"Income"), oddsratio= exp(confint(newlog,"Inco
me")))
```

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

```
##        coefficient.Income oddsratio
## 2.5 %            1.584421  4.876466
## 97.5 %           2.492677 12.093604
```

The statistic for this test is the profile likelihood.

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 her 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 nor a Parent Magazine.

```
summary(newlog)
```

```
##
## 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
```

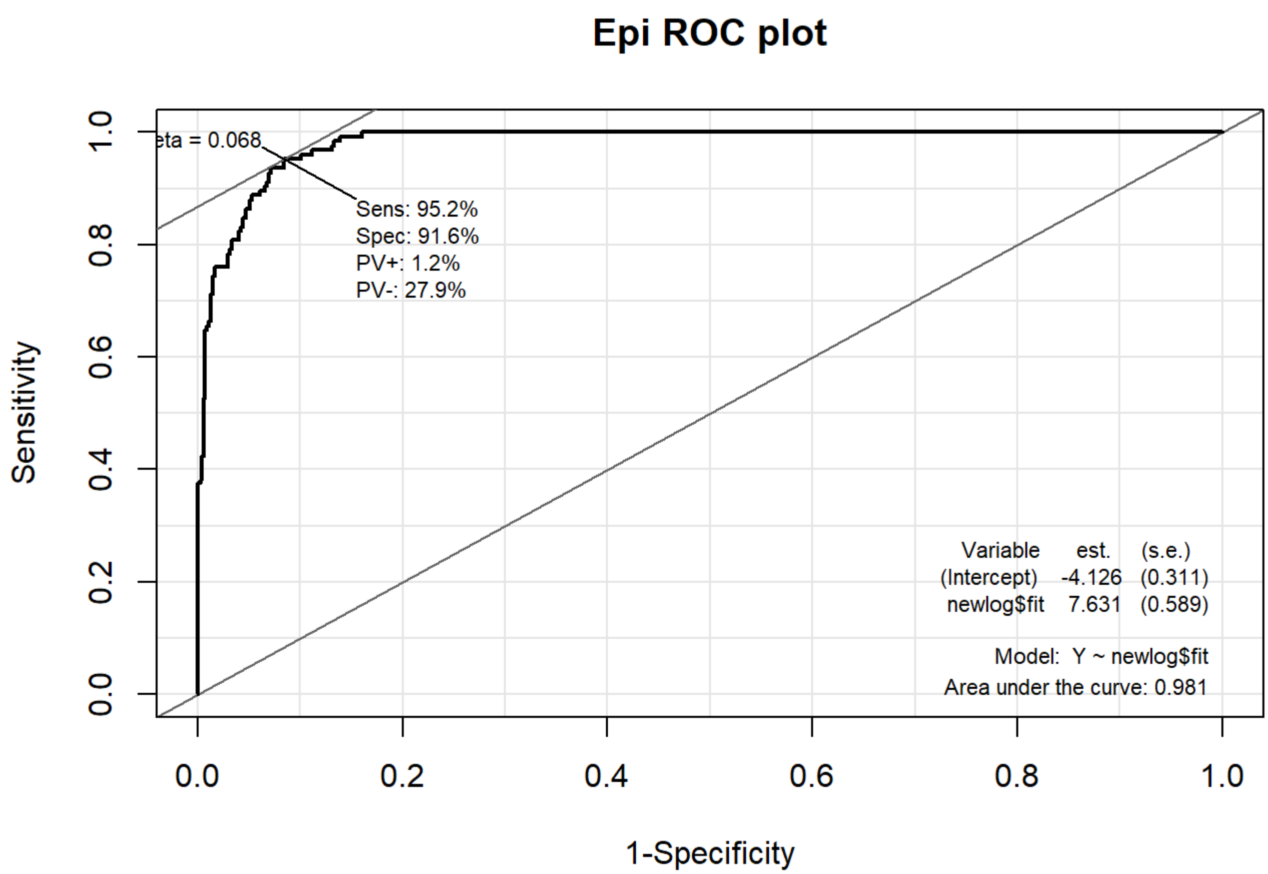
The equation for the predicted probabilites according to my model is P(y=1|X=x1,…,x11)= exp(intercept+ Income| 5/ MJi q epi X2+IsRetired| 7/ Vi whi r xpi r kx/ x4+DualIncome| 9/ Q mnswx x6+Own| ; / L sywi x8+White| = / I r kpwt x10+PrevChildMag| 55-35/ i | t ,mxi vgi t x/ Mgsq i x1+IsFemale\ 6/ MVi xwi h x3+Residentlength| 8/ HyepMgsq i x5+Minors| : / S{ r x7+House| </ [I mi x9 +English| 54/ Tvi zGI mpQ ek x11))

```
j<- exp(1.99159*6.8+ 1.60536*1+3*0.02501+0.76534*1+1.20598*1+1.86036*1 +1.62270*1)
Prob<- j/(1+j)
print(Prob, digits = 10)
```

```
## [1] 0.999999999
```

H. (10 pts) A prediction rule is to be developed. Draw the ROC curve for your model. Do the dynamic plotting at the end of the file ROC curves to get a better idea of what happens with the different cutoff (threshold) values which you will see at the bottom of the graph on the left side. Find the total error rates (number of misclassified observations/total number of observations) for cutoffs: .1, .2, .3, .4, .5, .6 Make a decision rule based on your findings, that is, decide on a cutoff value for prediction. Explain why you chose such value.

```
#Y values {0 or 1}
# S the predicted values from logistic model
S<- predict(newlog)
Y<- Buy
library(Epi)
ROC(form=Y~newlog$fit,plot="ROC",PV=TRUE,MX=TRUE,AUC=TRUE,data=KidCreative ,main="Epi RO
C plot")
```




```
library(ROCR)
```

```
## Loading required package: gplots
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
roc.curve=function(s,print=FALSE){
  # s is the threshold
  Ps=(S>s)*1
  FP=sum((Ps==1)*(Y==0))/sum(Y==0)
  TP=sum((Ps==1)*(Y==1))/sum(Y==1)
  if(print==TRUE){
    print(table(Observed=Y,Predicted=Ps))
  }
  vect=c(FP,TP)
  names(vect)=c("FPR","TPR")
  return(vect)
}
```

```
## cutoff .1
```

```
##         Predicted
## Observed   0   1
##        0 526  22
##        1  24 101
```

```
##        FPR        TPR
## 0.04014599 0.80800000
```

```
## prediction error rate: 0.1040119
```

```
## cutoff .2
```

```
##         Predicted
## Observed   0   1
##        0 530  18
##        1  26  99
```

```
##        FPR        TPR
## 0.03284672 0.79200000
```

```
## prediction error rate: 0.07726597
```

```
## cutoff .3
```

```
##         Predicted
## Observed   0   1
##        0 532  16
##        1  27  98
```

```
##        FPR        TPR
## 0.02919708 0.78400000
```

```
## prediction error rate: 0.07280832
```

```
## cutoff .4
```

```
##         Predicted
## Observed   0   1
##        0 534  14
##        1  30  95
```

```
##        FPR        TPR
## 0.02554745 0.76000000
```

```
## prediction error rate: 0.06389302
```

```
## cutoff .5
```

```
##         Predicted
## Observed   0   1
##        0 536  12
##        1  30  95
```

```
##        FPR        TPR
## 0.02189781 0.76000000
```

```
## prediction error rate: 0.06389302
```

```
## cutoff .6
```

```
##         Predicted
## Observed   0   1
##        0 539   9
##        1  30  95
```

```
##        FPR        TPR
## 0.01642336 0.76000000
```

```
## prediction error rate: 0.0653789
```

Notice above that a cutoff of .1 has the highest probability of detecting True Positives. On the other had it also has a high proportion of False Postives. Cutoff .3 has half the rate of FP and a decent TPR and a better prediction error.

so, Based on my findings I would choose a cutoff value of .3