47/50

MATH-661: Generalized Linear Models

Midterm Exam

Due Tuesday March 27, 2018

INSTRUCTIONS

No collaboration or discussion is permitted on the midterm exam. If you need clarifications, you can contact me but you are not allowed to ask anyone else. Please note that clarifications are limited to ambiguities in the wording of questions. This exam is intended to demonstrate your grasp of the material, so no help will be provided.

Please fill your name and sign the following honor pledge:

I, MICHAEL LEIBERT, pledge that I have not violated the Georgetown University honor code (see http://gervaseprograms.georgetown.edu/honor/). The work I am submitting for this exam is completely my own. I have not communicated with anyone and have not allowed any other student to use or borrow portions of my work. I understand that if I violate this honesty pledge, I will be reported for academic dishonesty to the Honor Council.

Signature : ____________

- Show the details of your work in order to get full credit for correct answers, and partial credit for incorrect answers if you are on the right track.
- Provide interpretations and conclusions in the context of the problem.
- Include the relevant R code and output for each question, when applicable.
- The exam must be e-mailed to `mgt26@georgetown.edu` by 11:59 pm on Tuesday, March 27, 2018.

Michael Leibert
Math 661
Midterm

Part I: Aspirin and heart attack [20 points]

A study investigating the association between heart attacks and the use of aspirin is conducted. Age is a potential confounder and is also considered. The following indicator variables are defined:

$$Y = \begin{cases} 1 & \text{if heart attack} \\ 0 & \text{if no heart attack} \end{cases} \qquad \text{Aspirin} = \begin{cases} 1 & \text{if aspirin} \\ 0 & \text{if placebo} \end{cases}$$

$$\text{Age1} = \begin{cases} 1 & \text{if age is } 40-50 \\ 0 & \text{otherwise} \end{cases} \qquad \text{Age2} = \begin{cases} 1 & \text{if age is} > 50 \\ 0 & \text{otherwise} \end{cases}$$

The following table shows the results of fitting logistic regression models for $P(Y=1)$:

Model	Covariates	Estimate $\hat{\beta}$	Standard Error	log-likelihood
1	None	-2.99	0.19	-116.54
2	Aspirin	-0.82	0.41	-114.41
3	Age1	-0.19	0.47	-116.27
	Age2	0.17	0.45	
4	Aspirin	-0.82	0.41	-114.14
	Age1	-0.18	0.47	
	Age2	0.19	0.45	
5	Aspirin	-0.65	0.63	-113.83
	Age1	-0.22	0.59	
	Age2	0.39	0.54	
	(Age1)*Aspirin	0.10	0.97	
	(Age2)*Aspirin	-0.68	1.03	

(a) Test the null hypothesis of constant aspirin effect on the risk of heart attack across age groups (i.e., no interaction between aspirin and age).

To test the null hypothesis of constant aspirin effect on the risk of heart attack across age groups we first look at the equation for the interaction model,

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \beta_0 + \beta_1 ASP + \beta_2 AGE_1 + \beta_3 AGE_2 + \beta_4 ASP \cdot AGE_1 + \beta_5 ASP \cdot AGE_2.$$

When Aspirin is a placebo, the equation reduces to:

$$\begin{aligned}\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) &= \beta_0 + \beta_1(0) + \beta_2 AGE_1 + \beta_3 AGE_2 + \beta_4(0) \cdot AGE_1 + \beta_5(0) \cdot AGE_2 \\ &= \beta_0 + \beta_2 AGE_1 + \beta_3 AGE_2.\end{aligned}$$

And when Aspirin is truly Aspirin, the equation can be written as:

$$\begin{aligned}\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) &= \beta_0 + \beta_1(1) + \beta_2 AGE_1 + \beta_3 AGE_2 + \beta_4(1)\cdot AGE_1 + \beta_5(1)\cdot AGE_2 \\ &= \beta_0 + \beta_1 + (\beta_2+\beta_4)\, AGE_1 + (\beta_3+\beta_5)\, AGE_2.\end{aligned}$$

To detect the presence of interaction effects we can test H_0: $\beta_4 = 0$ and H_0: $\beta_5 = 0$.

-0.5
If you want to use the Wald test, you should still test jointly (not each separately)
H0: beta_4 = beta_5 = 0

$$z = \left(\frac{0.1}{0.97}\right) = 0.1030928$$

```
.1/.97

## [1] 0.1030928

2* ( 1-pnorm(.1/.97) )

## [1] 0.9178893
```

$$z = \left(-\frac{0.68}{1.03}\right)^2 = 0.4358563$$

```
(-.68/1.03)^2

## [1] 0.4358563

 ( 1-pchisq( (-.68/1.03)^2 , 1) )

## [1] 0.5091292
```

Both p-values are large, so we can individually say: we fail to reject H_0: $\beta_4 = 0$, constant aspirin effect on the risk of heart attack for age group 40-50; and we fail to reject H_0: $\beta_5 = 0$, constant aspirin effect on the risk of heart attack for age group > 50.

However, we wish to detect aspirin effect across all age groups, so we test H_0: $\beta_4 = \beta_5 = 0$. We can do so with a likelihood ratio test. Recall from above that model 3 is nested within the saturated model (model 5).

The likelihood-ratio statistic is $G^2 = -2\,(\ell_0 - \ell_1)$ with $3 - 0 = 0$ degrees of freedom.

-1.5
To compare constant effect of aspirin across age groups, i.e., no interaction effect, we compare model 4 to model 5

$$\begin{aligned}G^2 &= -2(\ell_0 - \ell_1) \\ &= -2(-116.27 + 113.83) \\ &= 4.88\end{aligned}$$

```
-2 * ( -116.27 - -113.83      )

## [1] 4.88

1-pchisq(4.88 ,3)

## [1] 0.180798
```

The test statistic yields a p-value $= P\left(\chi^2_3 \geq 4.88\right) = 0.180798$. At $\alpha = 0.05$ we fail to reject H_0 and conclude that there is constant aspirin effect on the risk of heart attack for all age groups.

(b) Based on the model with additive/main effects for age and aspirin (Model 4)

Equation of the model:

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \beta_0 + \beta_1 ASP + \beta_2 AGE_1 + \beta_3 AGE_2.$$

i. Calculate the MLE of the odds ratio of aspirin use on heart attack, adjusting for age. Provide a 95% confidence interval for this odds ratio and interpret it in context.

$$\text{Odds Ratio } = \exp\left(\hat{\beta}_1\right) \Longrightarrow \exp(-0.82) = 0.4404317$$

$$\begin{aligned} 95\% \text{ CI} &\Longrightarrow \Big(\exp\left(-.82 - 1.96 \cdot 0.41\right), \exp\left(-.82 + 1.96 \cdot 0.41\right)\Big) \\ &\Longrightarrow \Big(\exp\left(-1.6236\right), \exp\left(-0.0164\right)\Big) \\ &\Longrightarrow \Big(0.1971875, 0.9837337\Big) \end{aligned}$$

We can conclude with approximately 95% confidence that the odds ratio is between 0.1971875 and 0.9837337. The CI does not contain $\exp(0) = 1$ and reject H_0: $\beta_1 = 0$. The odds ratio tells us that with aspirin use the odds of a heart attack are 0.44 times lower, after controlling for age.

ii. Perform a Wald test of the null hypothesis that there is no effect of aspirin on the risk of heart attack, controlling for age. What do you conclude?

$$\text{Test } H_0\text{: } \beta_1 = 0 \text{ vs } H_1\text{: } \beta_1 \neq 0$$

$$z = \left(-\frac{0.82}{0.41}\right)^2 = 4 \sim \chi^2_1$$

```
1-pchisq(4,1)
## [1] 0.04550026
```

Reject H_0 and conclude there is a significant effect on aspirin after controlling for age.

iii. Perform a likelihood ratio test of the null hypothesis that there is no effect of age on the risk of heart attack, controlling for aspirin use. State your conclusion.

If we want to evaluate the effect of age across all levels,

$$H_0 : \beta_2 = \beta_3 = 0.$$

The likelihood-ratio statistic is $G^2 = -2\left(\ell_0 - \ell_1\right)$ with $4 - 2 = 2$ degrees of freedom.

$$
\begin{aligned}
G^2 &= -2(\ell_0 - \ell_1) \\
&= -2(-114.41 + 114.14) \\
&= 0.54
\end{aligned}
$$

The test statistic yields a p-value $= P\left(\chi^2_2 \geq 0.54\right) = 0.7633795$. At $\alpha = 0.05$ we fail to reject H_0 and conclude that there is no evidence that age is associated with the probability of having a heart attack, adjusting for aspirin use.

```
-2 * ( -114.41 - -114.14)
## [1] 0.54
1-pchisq(-2 * ( -114.41 - -114.14),2)
## [1] 0.7633795
```

(c) Evaluate the deviance of each model provided in the table and assess its goodness-of fit. Which models do not provide adequate fit to the data?

Model	Logit(π_i)	*df*	G^2	p
null	β_0	5	5.42	0.367
ASP	$\beta_0 + \beta_1 ASP$	4	1.16	0.885
AGE	$\beta_0 + \beta_2 AGE_1 + \beta_3 AGE_2$	3	4.88	0.181
$ASP + AGE$	$\beta_0 + \beta_1 ASP + \beta_2 AGE_1 + \beta_3 AGE_2$	2	0.62	0.733
Saturated	$\beta_0 + \beta_1 ASP + \beta_2 AGE_1 + \beta_3 AGE_2 + \beta_4 ASP \cdot AGE_1 + \beta_5 ASP \cdot AGE_2$	0	0	-

The AGE and null models have relatively ~~low~~ p's and therefore ~~are poor~~ fits to the data.

```
logL<-c(-116.54,-114.41,-116.27,-114.14,-113.83)
degF<-c(5,4,3,2,0)

2*(-113.83 - logL)

## [1] 5.42 1.16 4.88 0.62 0.00

1-pchisq( 2*(-113.83 - logL) ,  degF )

## [1] 0.3667979 0.8846394 0.1807980 0.7334470 1.0000000
```

(d) Perform model selection using analysis-of-deviance. Make sure you describe all the steps to arrive at your final model.

Table 1: H_0: Smaller model fits well.

	ΔG	Δdf	p-value
null vs. ASP	4.26	1	0.039
null vs. AGE	0.54	2	0.763
ASP vs. ASP+ AGE	0.54	2	0.763
AGE vs. ASP+ AGE	4.26	1	0.039
$ASP + AGE$ vs. $ASP \cdot AGE$	0.62	2	0.733

```
#G^2 DEVIANCES
2*(-113.83 - -116.54)   #null

## [1] 5.42

2*(-113.83 - -114.41)   #ASP

## [1] 1.16

2*(-113.83 - -116.27)   #AGE

## [1] 4.88

2*(-113.83 - -114.14)   #AGE+ASP

## [1] 0.62

2*(-113.83 - -113.83)   #AGE*ASP

## [1] 0

AOD<-data.frame(c(
5.42-1.16,  5.42-4.88,  1.16-0.62,  4.88-0.62,  0.62
),c(
1-pchisq(5.42-1.16, 1),   1-pchisq(5.42-4.88, 2),   1-pchisq(1.16-0.62, 2),
1-pchisq(4.88-0.62, 1),   1-pchisq( 0.62, 2)))

rownames(AOD)<-c("null vs ASP","null vs AGE","ASP vs ASP+AGE","AGE vs ASP+AGE","ASP+AGE vs ASP*AGE");
names(AOD)<-c("dG","pvalue");AOD

##                      dG     pvalue
## null vs ASP        4.26 0.03901992
## null vs AGE        0.54 0.76337949
## ASP vs ASP+AGE     0.54 0.76337949
## AGE vs ASP+AGE     4.26 0.03901992
## ASP+AGE vs ASP*AGE 0.62 0.73344696
```

AGE does better than the null model, but it was not a good fit to begin with. We are left with the *ASP* model and the main effects model, both of which do better than the model they are nested under. Because the *ASP* model is a simpler model than the main effects one, and also appears to fit better, we will select that as our final model.

Part II: Credit risks for bank loan [30 points]

Banks want to reduce the rate of loan defaults. Loan officers want to be able to identify characteristics that are indicative of people who are likely to default on loans, and then use those characteristics to identify good and bad credit risks.

Financial and demographic information are collected on 850 past and prospective customers. Of these, 700 are customers who were previously given loans and 150 are prospective customers that the bank needs to classify as good of bad credit risks. The data are saved in `Bank_loan.txt` and contain the following variables:

age	age in years
ed	highest level of education
	1: did not complete high school; 2: high school degree
	3: some college; 4: college degree; 5: post-bachelor degree
employ	years with current employer
address	years at current address
income	household income in thousands
debtinc	debt to income ratio ($\times$100)
creddebt	credit card debt in thousands
othdebt	other debt in thousands
default	previously defaulted – 0: No, 1: Yes, NA: prospective customers

1. **Exploratory data analysis & data processing**

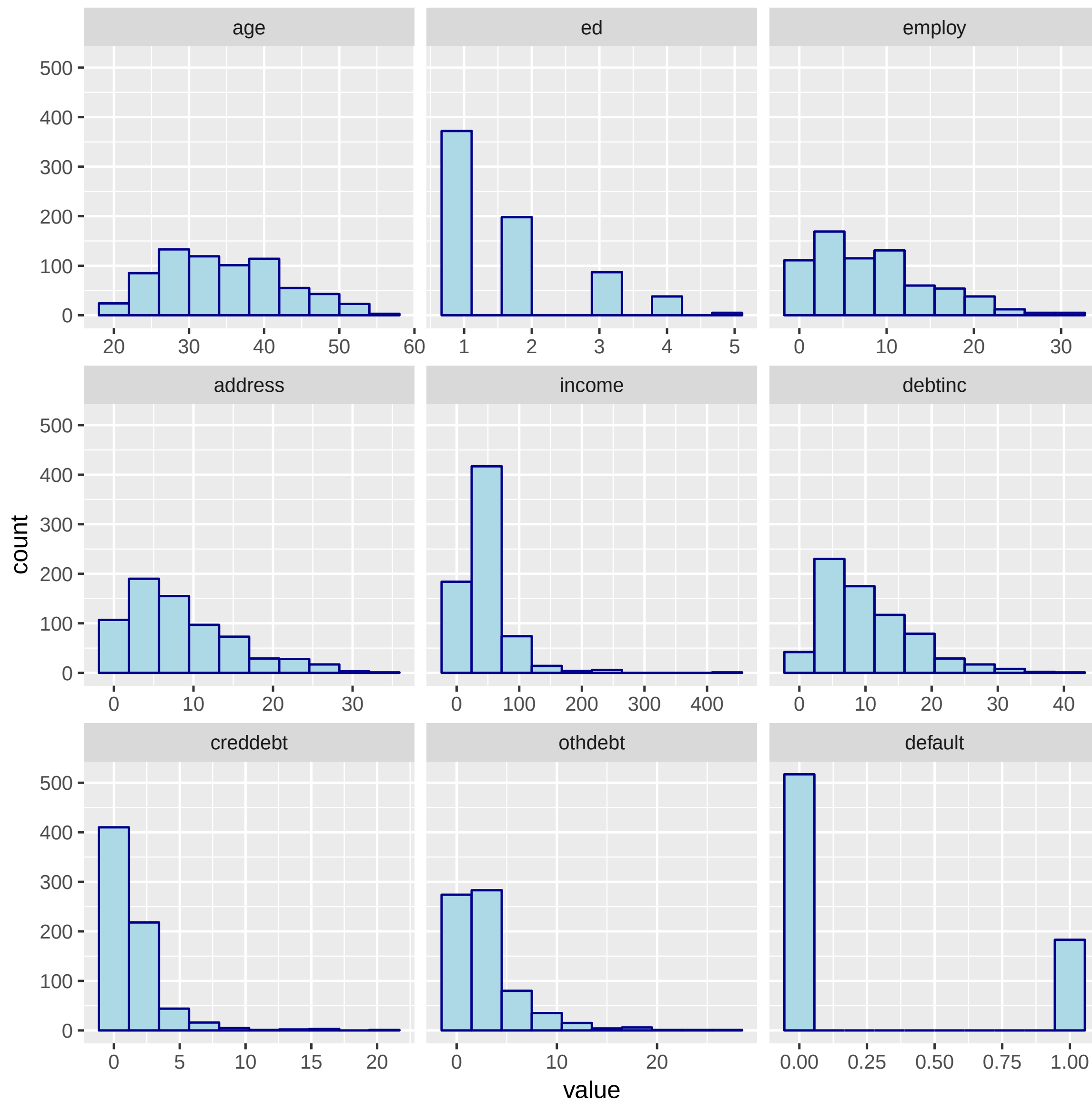
   (a) Provide appropriate summary statistics and graphical displays for the variables in the data. Discuss their distributions.

   Nearly all the distributions are right-skewed. In general the sample has more younger people than older, more lower educated than higher educated, most of the years at the the current employer and current address are under 10, average income is around 47 with few people over 100, most people have low debt to income ratio, low credit card debt, and low other debt. Also, most people have not defaulted.

```
dat<-read.table("Bank_loan.txt",header=T)
library(ggplot2)
library(reshape2)

 ggplot(data = melt(dat[!(is.na(dat[,ncol(dat)])),]),
  mapping = aes(    x = value))   +
        geom_histogram( bins = 10, color="darkblue", fill="lightblue") +
        facet_wrap(~variable, scales = 'free_x')
## No id variables; using all as measure variables
```

-1
It would be helpful to provide measures of center and spread for the continuous variables, and proportions for the categorical variables.




```
dat$ed<-as.factor(dat$ed)
summary(dat)
```

```
##       age          ed          employ          address
##  Min.   :20.00   1:460   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:29.00   2:235   1st Qu.: 3.000   1st Qu.: 3.000
##  Median :34.00   3:101   Median : 7.000   Median : 7.000
##  Mean   :35.03   4: 49   Mean   : 8.566   Mean   : 8.372
##  3rd Qu.:41.00   5:  5   3rd Qu.:13.000   3rd Qu.:12.000
##  Max.   :56.00           Max.   :33.000   Max.   :34.000
##
##      income          debtinc         creddebt          othdebt
##  Min.   : 13.00   Min.   : 0.10   Min.   : 0.0117   Min.   : 0.04558
##  1st Qu.: 24.00   1st Qu.: 5.10   1st Qu.: 0.3822   1st Qu.: 1.04594
##  Median : 35.00   Median : 8.70   Median : 0.8851   Median : 2.00324
##  Mean   : 46.68   Mean   :10.17   Mean   : 1.5768   Mean   : 3.07879
##  3rd Qu.: 55.75   3rd Qu.:13.80   3rd Qu.: 1.8984   3rd Qu.: 3.90300
##  Max.   :446.00   Max.   :41.30   Max.   :20.5613   Max.   :35.19750
##
##     default
```

```
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.2614
##  3rd Qu.:1.0000
##  Max.   :1.0000
##  NA's   :150
```

(b) Since there are few observations with post-bachelor degree (`ed` = 5), combine these with the group with college degree (`ed` = 4). You will be using education with these 4 levels in subsequent analyses.

```
a<-which(dat$ed == 5)
dat[which(dat$ed == 5),2]<-4
dat[a, ];rm(a)

##     age ed employ address income debtinc creddebt   othdebt default
## 367  36  4      5      12     20     8.1 0.729000  0.891000       0
## 385  52  4      9       0     70     9.4 1.329160  5.250840       1
## 388  46  4     15       0    126     3.1 0.476532  3.429468       0
## 457  37  4      9      16    177     5.9 0.887655  9.555345       0
## 503  42  4      6      23    190     7.8 3.156660 11.663340       0
```

(c) Separate the 150 prospective customers for whom credit risk is to be predicted from the 700 past customers.

```
prosp<-dat[is.na(dat[,ncol(dat)]),]
dat<-dat[!(is.na(dat[,ncol(dat)])),]
nrow(dat);nrow(prosp)

## [1] 700
## [1] 150

tail(prosp)

##     age ed employ address income debtinc creddebt  othdebt default
## 845  23  1      3       4     13     3.1 0.045539 0.357461      NA
## 846  34  1     12      15     32     2.7 0.239328 0.624672      NA
## 847  32  2     12      11    116     5.7 4.026708 2.585292      NA
## 848  48  1     13      11     38    10.8 0.722304 3.381696      NA
## 849  35  2      1      11     24     7.8 0.417456 1.454544      NA
## 850  37  1     20      13     41    12.9 0.899130 4.389870      NA
```

2. **Model building & diagnostics** – use the 700 past customers for this task.

(a) Perform stepwise selection.

```
fit.null<-glm(default~1,family=binomial, data=dat)
fit.sat<-glm(default~. ,family=binomial, data=dat)
step(fit.null, scope=list(lower=fit.null, upper=fit.sat), direction="both" ,trace=0)

##
## Call:  glm(formula = default ~ debtinc + employ + creddebt + address +
##     age, family = binomial, data = dat)
```

```
##
## Coefficients:
## (Intercept)      debtinc       employ     creddebt      address
##    -1.63128      0.08926     -0.26076      0.57265     -0.10365
##         age
##     0.03256
##
## Degrees of Freedom: 699 Total (i.e. Null);  694 Residual
## Null Deviance:     804.4
## Residual Deviance: 553.2  AIC: 565.2
```

i. Provide the equation of the selected model.

$$\text{logit}(\hat{\pi}_i) = -1.63128 + 0.08926\ debtinc - 0.26076\ employ + 0.57265\ creddebt - 0.10365\ address + 0.03256\ age$$

ii. Interpret the effect of each of the covariates in the selected model.

- Controlling for all other variables, a 1-unit increase in debt to income ratio ($\times 100$) is associated with a 0.08926 increase in log-odds of defaulting.

- Controlling for all other variables, a 1-unit increase in years at current employer is associated with a 0.26076 decrease in log-odds of defaulting.

- Controlling for all other variables, a 1-unit increase in credit card debt (in thousands) is associated with a 0.57265 increase in log-odds of defaulting.

- Controlling for all other variables, a 1-unit increase in years at current address is associated with a 0.10365 decrease in log-odds of defaulting.

- Controlling for all other variables, a 1-unit increase in age is associated with a 0.03256 decrease in log-odds of defaulting.

iii. Assess the goodness-of-fit of the selected model.

We can the Hosmer-Lemeshow goodness-of-fit test to assess the model fit.

H_0: the model fits the data well.

```
stepFit<-glm(formula = default ~ debtinc + employ + creddebt + address +
    age, family = binomial, data = dat)

res<-hoslem.test(stepFit$y,fitted(stepFit))
res
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  stepFit$y, fitted(stepFit)
## X-squared = 3.6418, df = 8, p-value = 0.8879
cbind(res$observed,res$expected)                                  #£
##                      y0 y1    yhat0      yhat1
## [0.000457,0.00907] 70  0 69.70880  0.2912023
```

```
## (0.00907,0.0313]   68  2 68.66537  1.3346344
## (0.0313,0.0627]    66  4 66.80522  3.1947796
## (0.0627,0.116]     63  7 63.79967  6.2003317
## (0.116,0.178]      64  6 59.99545 10.0045462
## (0.178,0.252]      52 18 54.96048 15.0395172
## (0.252,0.37]       49 21 48.32948 21.6705232
## (0.37,0.486]       41 29 40.80865 29.1913485
## (0.486,0.675]      31 39 30.43524 39.5647641
## (0.675,0.999]      13 57 13.49165 56.5083530
```

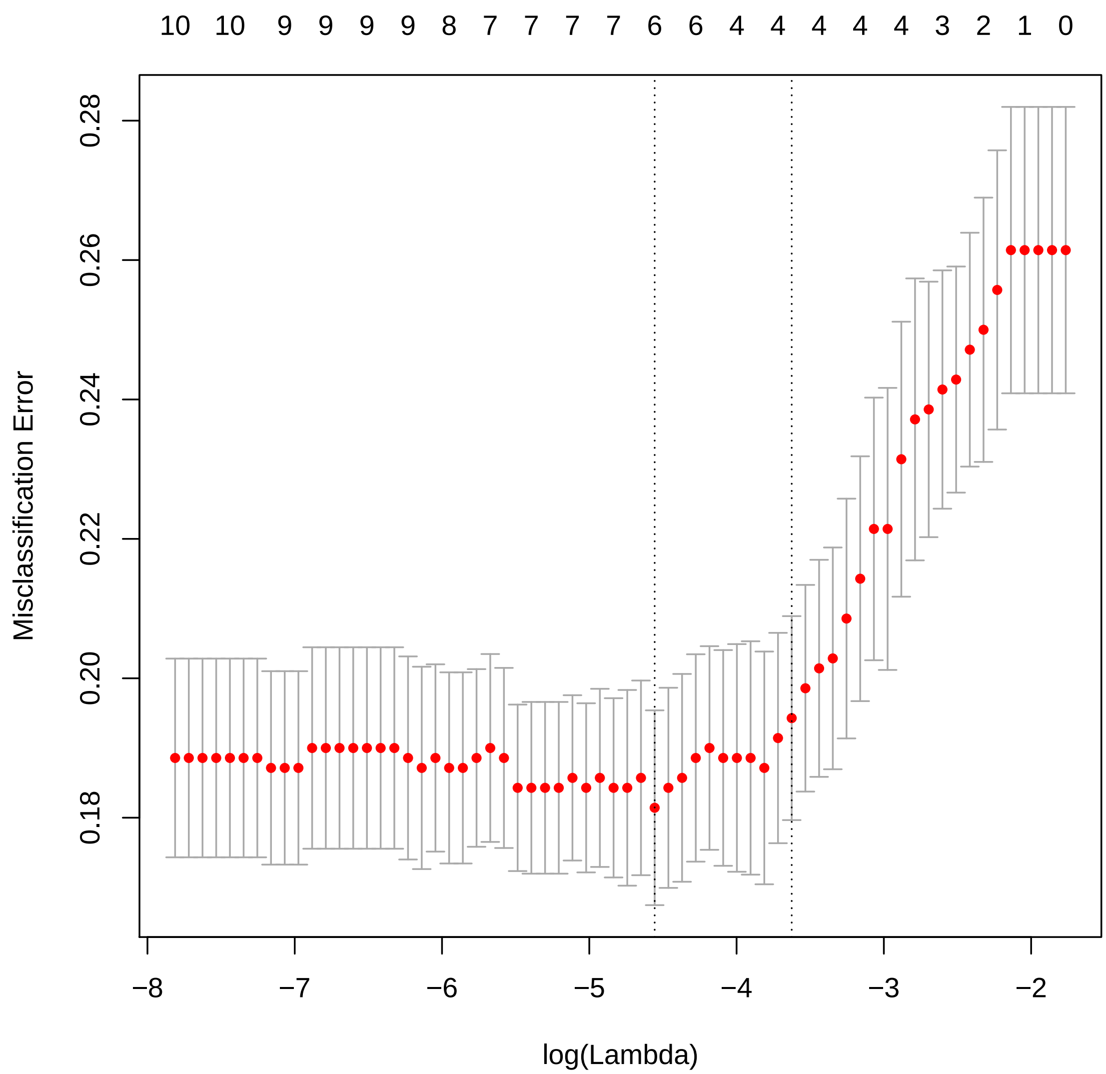
We fail to reject at $\alpha = 0.05$. Thus, there is not sufficient evidence to suggest that the model does not provide adequate fit to the data. Given the large p-value, we can declare a well fitting model.

(b) Perform a lasso variable selection using the misclassification error as criterion for choosing λ.

```
library("mlbench");library(glmnet)

X = model.matrix(default ~ . , data=dat)
Y = as.numeric(dat$default  )

cvfit = cv.glmnet(x=X[,-1], y=Y, family="binomial", type.measure="class")
plot(cvfit   )
```




```
lambda_1se = cvfit$lambda.1se
coef(cvfit, s=lambda_1se)
```

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                       1
## (Intercept) -1.12204726
## age           .
## ed2           .
## ed3           .
## ed4           .
## ed5           .
## employ      -0.13739321
## address     -0.03108643
## income        .
## debtinc      0.07646120
## creddebt     0.27531415
## othdebt       .
```

```
lassoFit<-glm(formula = default ~ debtinc + employ + creddebt + address ,
        family = binomial, data = dat)
        summary(lassoFit)
```

```
##
## Call:
## glm(formula = default ~ debtinc + employ + creddebt + address,
##     family = binomial, data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4482  -0.6392  -0.3111   0.2582   2.8495
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.79107    0.25154  -3.145  0.00166 **
## debtinc      0.08827    0.01854   4.760 1.93e-06 ***
## employ      -0.24260    0.02806  -8.646  < 2e-16 ***
## creddebt     0.57300    0.08727   6.566 5.18e-11 ***
## address     -0.08125    0.01960  -4.145 3.39e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 804.36  on 699  degrees of freedom
## Residual deviance: 556.73  on 695  degrees of freedom
## AIC: 566.73
##
## Number of Fisher Scoring iterations: 6
```

i. Compare the models selected using `lambda.1se` to the stepwise selected model in (a). Perform a likelihood ratio test to choose the preferred model between the two at $\alpha = 0.05$.

H_0: Lasso model fits as well as the stepwise model.

We can use a likelihood ratio test to compare this four covariates model (lasso) to the five covariates model (stepwise), which will have an approximate chi-square distribution with $df = 695 - 694 = 1$. This is equivalent to evaluating the change in deviance between the two models:

```
deviance(lassoFit)-deviance(stepFit)
## [1] 3.555619
1-pchisq(deviance(lassoFit)-deviance(stepFit), 1)
## [1] 0.05934418
```

We fail to reject H_0 at $\alpha = 0.05$, thus the model with five covariates does not provide a better fit compared to the model with four covariates.

(c) For the model selected based on lasso

i. Identify observations with unusual/outlying standardized residuals. How do the predictions for these individuals, based on their fitted values $\hat{\pi}_i$, compare to their observed default status?

```
dat$pii<-predict(lassoFit,   type="response")            #£

h<-hatvalues(lassoFit)
```

```r
e<-resid(lassoFit,type="pearson")
e.std<-e/sqrt(1-h)
hist(e.std)
```

```r
dat[as.numeric( which( abs(e.std) > 2 ) ), c(6,3,7,4,9,10) ]
##     debtinc employ creddebt address default        pii
## 10     19.7      0 2.777700      13       0 0.81505701
## 16      8.6      9 0.817516       6       1 0.09670408
## 26     17.6      0 2.140160       2       0 0.86131340
## 36     26.0      6 6.048900       7       0 0.95005352
## 53     12.9     16 3.032016      18       1 0.03700212
## 62      7.4     13 1.457652       1       1 0.07325835
## 69      8.2      8 1.492154       3       1 0.19832433
## 107    11.2      9 2.016000      18       1 0.09169451
## 152     6.1     13 2.151104      23       1 0.01725101
## 185    24.2      0 1.424654       7       0 0.83098020
## 187    15.0     14 2.792850      21       1 0.04883231
## 193    11.2     10 0.815360       0       1 0.14663235
## 202     6.1      8 0.284504      10       1 0.05505157
## 214     5.4      5 0.581418       3       1 0.19185483
## 219     4.0      7 0.447600       8       1 0.07380372
## 232     8.0      3 0.563200      11       1 0.20041281
## 264    19.2      9 0.801792      11       1 0.15264483
## 281     4.8     11 0.345408       6       1 0.03470598
## 295     9.8     10 3.236548      12       1 0.18657131
## 299     5.0      4 0.549450       5       1 0.19599444
## 320     3.1      4 0.283960       1       1 0.19680727
## 332     2.4      3 0.259200       3       1 0.19746142
## 356    14.5      7 0.373520       6       1 0.18499966
## 385     9.4      9 1.329160       0       1 0.20050560
## 492    25.2     13 2.316132      13       1 0.19006687
## 515     8.6     14 1.201248       1       1 0.05618503
## 577     5.6      8 0.569296       0       1 0.12882892
## 618     5.1      6 0.449820       7       1 0.10837307
## 651     6.4      6 0.594432       7       1 0.12899668
## 662     0.9      0 0.118017      13       1 0.15443981
## 678     2.1      6 0.390852       9       1 0.07119029
## 696     4.6      6 0.262062      15       1 0.05170308
```

Because the data are not grouped, looking at plots does not help with identifying unusual/outlying standardized residuals. However we know the standardized residuals $r_i \sim N(0,1)$. According to Agresti, absolute values of the r_i's larger than about 2 or 3 provide evidence of lack of fit.

To be conservative we will look at the r_i's with an absolute value larger than 2. It is clear from the subsetted data that the predictions for defaulting were off. There are some cases where $\hat{\pi}_i < 0.1$ and the individual still defaulted. Conversely, we see individuals with a $\hat{\pi}_i > 0.8$ that paid back their loan. There is a guy with a $\hat{\pi}_i = 0.95005352$ who does not default and someone with a $\hat{\pi}_i = 0.01725101$ who does. From what we see here, it is much more common that the model predicts that a person will not default and they end up not paying back their loan.

for the outliers, but not overall

ii. Using a cut-off of 0.3 for predicting whether a person defaults or not on a loan

- what proportion of the 700 customers would have been predicted as defaulting (and thus would have been denied a loan)?

Of the 700 customers, approximately 34.14% would have been denied a loan.

```
nrow(   dat[which(dat$pii > .3),]   ) / 700              #£
## [1] 0.3414286
```

- what would be the misclassification rate?

The misclassification would have been approximately 21.857%.

```
table( dat[ ( which( dat$pii < .3) ),]$default )
##
##   0   1
## 415  46
table( dat[ ( which( dat$pii > .3) ),]$default )
##
##   0   1
## 102 137
(46 + 107 ) / 700
## [1] 0.2185714
```

ok given typo

- what would be the misclassification rate among the defaulters?

Among the defaulters, the misclassification rate would have been 25.14%.

```
defaulters<-dat[which(dat$default == 1),]
nrow( defaulters[which(defaulters$pii < .3) ,] ) / nrow(defaulters)
## [1] 0.2513661
```

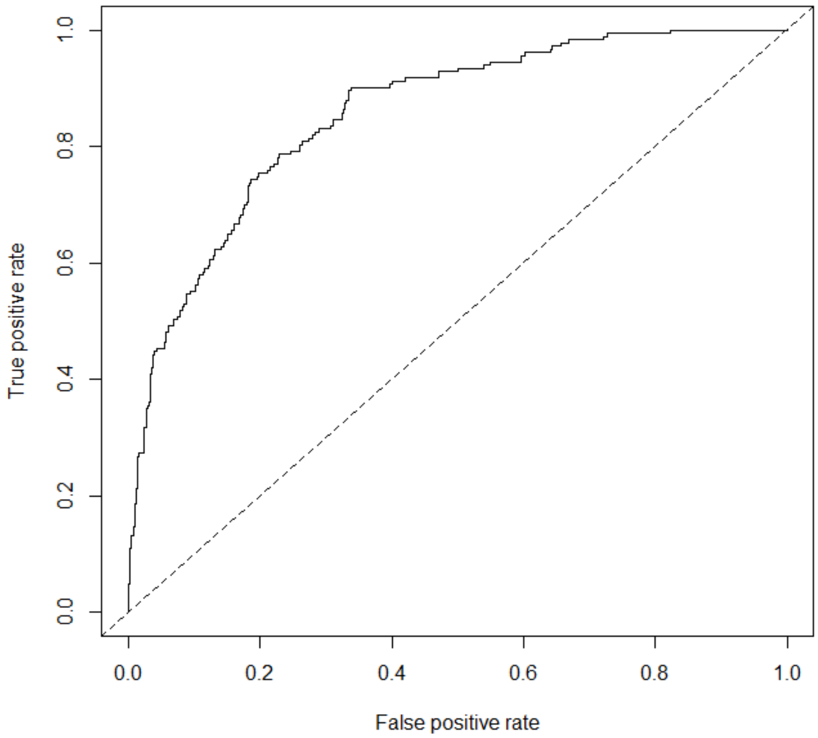
- what would be the misclassification rate among the non-defaulters?

Among the non-defaulters, the misclassification rate would have been 19.7%.

```
nondef<-dat[which(dat$default == 0),]
nrow( nondef[which(nondef$pii > .3) ,] ) / nrow(nondef)
## [1] 0.1972921
```

iii. Provide the ROC curve and the area under the ROC curve for the selected model.

```
library(ROCR)
pred = prediction(fitted(lassoFit) ,   dat$default  )
perf = performance(pred, "tpr", "fpr")
plot(perf)
abline(a=0, b=1, lty=2)
> auc.perf = performance(pred, "auc")
> auc.perf@y.values
[[1]]
[1] 0.8556088
```



3. **Prediction for future customers** – consider the model selected based on lasso.

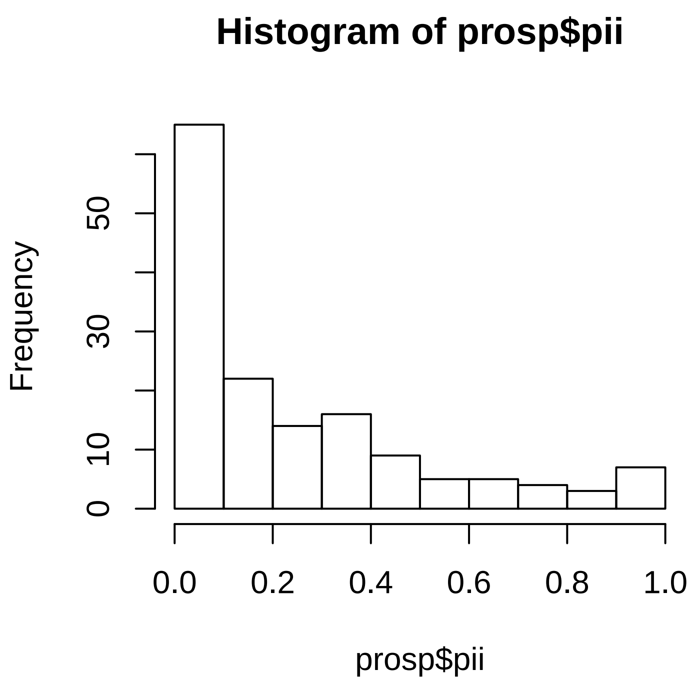
   (a) Calculate the predicted probabilities of loan default for the 150 prospective customers.

```
prosp$pii<-predict(lassoFit,  prosp, type="response")  ; prosp$pii
##   [1] 0.0114875528 0.0726114465 0.6437104595 0.0854981599 0.3597747991
##   [6] 0.4328725167 0.3614165848 0.8716077403 0.0971204932 0.1315346906
##  [11] 0.0134482517 0.0243808005 0.0016053977 0.0025643147 0.2186331609
##  [16] 0.3598649771 0.9687999028 0.0277146526 0.4139720393 0.0125046673
##  [21] 0.2646521753 0.0443285369 0.1048735141 0.0001419559 0.2769337701
##  [26] 0.1271427609 0.0326302918 0.0540236499 0.0018985252 0.1002703243
##  [31] 0.1384650936 0.0068804278 0.6685185768 0.0343296919 0.0264113321
##  [36] 0.1682668280 0.3374621297 0.0931446622 0.4721521600 0.2190248507
##  [41] 0.6082237807 0.0268481737 0.0013388184 0.0726857581 0.0432188672
##  [46] 0.8314669941 0.2227004169 0.0024924677 0.1346764664 0.0010756380
##  [51] 0.0020901174 0.2522315788 0.0824303600 0.0073251128 0.3011591947
##  [56] 0.0145657837 0.5526108954 0.4699194441 0.4105053948 0.4746639252
##  [61] 0.0625583701 0.1826263109 0.0950475595 0.9006884561 0.0693678625
##  [66] 0.0358082501 0.1599615643 0.0697350368 0.0466037460 0.1848182213
##  [71] 0.7425257823 0.0063498847 0.3112947494 0.5938474590 0.0001165731
##  [76] 0.3024542212 0.3813462403 0.5435158612 0.1976373491 0.0011859021
##  [81] 0.1252826664 0.6855740801 0.3236856401 0.5772005330 0.8121272695
##  [86] 0.2061148050 0.0049611778 0.2767424134 0.0312164736 0.7405561471
##  [91] 0.0453884086 0.3306800249 0.0059025287 0.3775281534 0.0239120469
##  [96] 0.0203284142 0.4451871649 0.3950614422 0.0021567854 0.4496990669
## [101] 0.3724864044 0.7651680990 0.3890330088 0.9954646739 0.0742908287
## [106] 0.1180902422 0.0216915796 0.9359528815 0.0633402310 0.0789806962
```

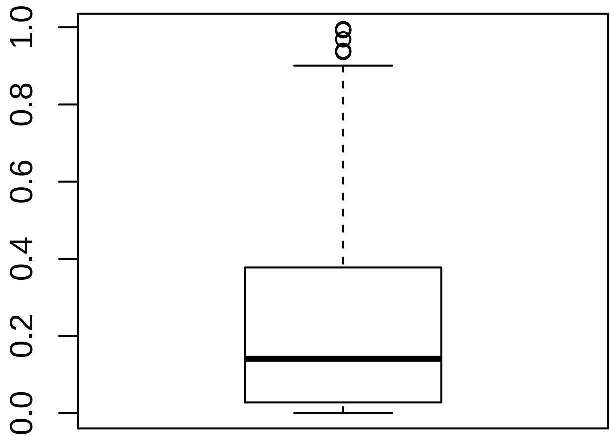
```
## [111] 0.9931636329 0.0129674025 0.0517447074 0.2935913712 0.0204681616
## [116] 0.3809051519 0.1213206525 0.1829436048 0.4363489134 0.2724861734
## [121] 0.3229106408 0.1814589396 0.6923599858 0.1905555018 0.2581449128
## [126] 0.5144986894 0.7019998342 0.0108133460 0.1937207647 0.0331280205
## [131] 0.0211468106 0.0166307959 0.0073074109 0.1457160024 0.1257581647
## [136] 0.0114857883 0.9392996713 0.0805382525 0.0094941535 0.2534422601
## [141] 0.9924487592 0.0549102350 0.0108346330 0.2434409874 0.1759367178
## [146] 0.0105036658 0.1436218639 0.0301374981 0.2690034510 0.0063978129
```

(b) Provide a histogram and a boxplot of the predicted probabilities.

```
hist(prosp$pii)
```




```
boxplot(prosp$pii)
```



(c) Using a cut-off of 0.3, how many of the 150 prospective customers would be expected to default on a loan?

Approximately 33% would be expected to default on a loan.

```
nrow( prosp[which(prosp$pii > .3),]  )  / nrow(prosp)   #£
## [1] 0.3266667
```