Stats 111 Homework 4

Owen Lin

1. a. The null deviance is 75.791, the residual deviance is 73.594. There isn't much difference between these two, that signifies Sex as a covariate does not provide much information on the acceptance status

```
mcat = read.table("D:\\Coding\\Stats111\\Data\\MedGPA.txt", fill=TRUE, header=TRUE)
logit1a = glm(Acceptance~Sex, family = binomial(link = "logit"), data = mcat)
summary(logit1a)
```

```
##
## Call:
## glm(formula = Acceptance ~ Sex, family = binomial(link = "logit"),
##     data = mcat)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.435  -1.084   0.940   0.940   1.274
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.5878     0.3944   1.490    0.136
## SexM         -0.8109     0.5528  -1.467    0.142
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 75.791  on 54  degrees of freedom
## Residual deviance: 73.594  on 53  degrees of freedom
## AIC: 77.594
##
## Number of Fisher Scoring iterations: 4
```

b. Null deviance is the same because it compares the null model (with only intercept) to the saturated model, it doesn't matter what we fit in our model. On the other hand, residual deviance compares our model to the saturated model, so it changes with respect to our covariate. There is a big difference between the two (larger than 10 compare to the small difference in part a). So MCAT is a better indicator of the acceptance status.

```
logit1b = glm(Acceptance~MCAT, family = binomial(link = "logit"), data = mcat)
summary(logit1b)
```

```
##
## Call:
## glm(formula = Acceptance ~ MCAT, family = binomial(link = "logit"),
##     data = mcat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7878  -1.0330   0.4256   0.9225   1.6601
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.71245    3.23645  -2.692  0.00710 **
## MCAT         0.24596    0.08938   2.752  0.00592 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 75.791  on 54  degrees of freedom
## Residual deviance: 64.697  on 53  degrees of freedom
## AIC: 68.697
##
## Number of Fisher Scoring iterations: 4
```

c. To create a test statistic for 1b, we would use Null Deviance - Residual Deviance. H0: null model holds (fit the data well enough) Ha: The proposed model is better than the null model test statistic: 75.791 - 64.697 = 11.094 ~ chi_squared(1)

d. $\ln \frac{p}{1-p} = -6.18 - 7.12 * I(Sex = Male) + 0.19 * MCAT + 0.17 * I(Sex = Male) * MCAT$

For male, one unit increase in MCAT leads to $e^{0.19+0.17}$ times estimated odds of getting accepted.

For female, one unit increase in MCAT leads to $e^{0.19}$ times estimated odds of getting accepted.

```
logit1d = glm(Acceptance ~ Sex + MCAT + Sex*MCAT, family = binomial(link = "logit"), data = mcat)
summary(logit1d)
```

```
##
## Call:
## glm(formula = Acceptance ~ Sex + MCAT + Sex * MCAT, family = binomial(link = "logit"),
##     data = mcat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7857  -0.9770   0.3549   0.9417   2.0304
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -6.1804     4.3247  -1.429    0.153
## SexM         -7.2122     7.1083  -1.015    0.310
## MCAT          0.1887     0.1212   1.557    0.119
## SexM:MCAT     0.1697     0.1946   0.872    0.383
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 75.791  on 54  degrees of freedom
## Residual deviance: 60.924  on 51  degrees of freedom
## AIC: 68.924
##
## Number of Fisher Scoring iterations: 5
```

e. The 95% confidence interval for the estimated odds ratio from a 1 unit increase in MCAT score for a male is (1.06, 1.93)
We are 95% confident that the true odds ratio for 1 unit increase in MCAT score is between 1.06 and 1.93.
The 95% confidence interval for the estimated odds ratio from a 5 unit increase in MCAT score for a male is (1.35, 26.68)

```
linContr.glm(c("MCAT","SexM:MCAT"), c(1,1), model=logit1d)
linContr.glm(c("MCAT","SexM:MCAT"), c(5,5), model=logit1d)
```

```
##
## Test of H_0: exp( 1*MCAT + 1*SexM:MCAT ) = 1 :
##
##   exp( Est )    se.est    zStat       pVal  ci95.lo ci95.hi
## 1    1.43098 0.1522573 2.353646 0.01859033 1.061774 1.92857
##
## Test of H_0: exp( 5*MCAT + 5*SexM:MCAT ) = 1 :
##
##   exp( Est )    se.est    zStat       pVal ci95.lo  ci95.hi
## 1   6.000237 0.7612867 2.353646 0.01859033 1.34946 26.67944
```

f. H0: $\beta_1 = \beta_3 = 0$ Ha: H0 is not true p-value: 0.1516 conclusion: Fail to reject the null, so Sex should not be included as a covariate.

```
lrtest(logit1b, logit1d)
```

```
##
## Assumption: Model 1 nested within Model 2
```

```
##   Resid. Df Resid. Dev Df Deviance pValue
## 1        53     64.697
## 2        51     60.924  2    3.773 0.1516
```

2. a. $\ln \frac{p}{1-p} = -1.31 + 1.83 * sqft - 3.43 * lot + 1.40 * Pool$
The difference between the null deviance and the residual deviance is large.

```
MidwestSales = read.table("D:\\Coding\\Stats111\\Data\\MidwestSales.txt", fill=TRUE, header=FALSE)
names(MidwestSales)=c("id","price","sqft","bed","bath","ac","garage","pool","year","quality","style","lot","hwy")
logit2a = glm(ac~sqft+lot+pool, family=binomial(link="logit"), data=MidwestSales)
summary(logit2a)
```

```
##
## Call:
## glm(formula = ac ~ sqft + lot + pool, family = binomial(link = "logit"),
##     data = MidwestSales)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3262   0.1710   0.4158   0.6822   1.5197
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.309e+00  5.784e-01  -2.263  0.02361 *
## sqft         1.828e-03  3.097e-04   5.901 3.62e-09 ***
## lot         -3.431e-05  9.396e-06  -3.652  0.00026 ***
## pool         1.397e+00  1.037e+00   1.347  0.17792
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 473.59  on 521  degrees of freedom
## Residual deviance: 407.26  on 518  degrees of freedom
## AIC: 415.26
##
## Number of Fisher Scoring iterations: 6
```

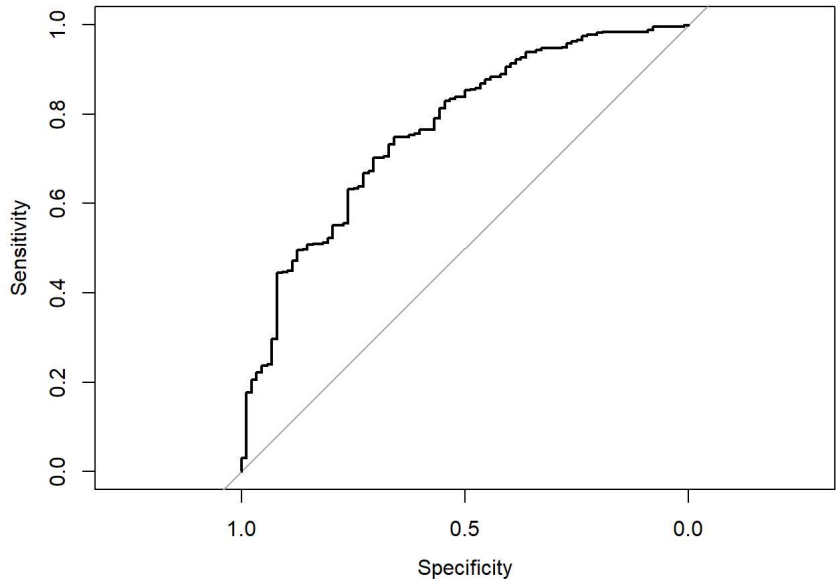
b. The area under the curve is 0.7649. The model is doing a good job on predicting ac because the area is higher than 0.5.

```
roc.curve = roc(MidwestSales$ac~fitted(logit2a))
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
plot(roc.curve)
```



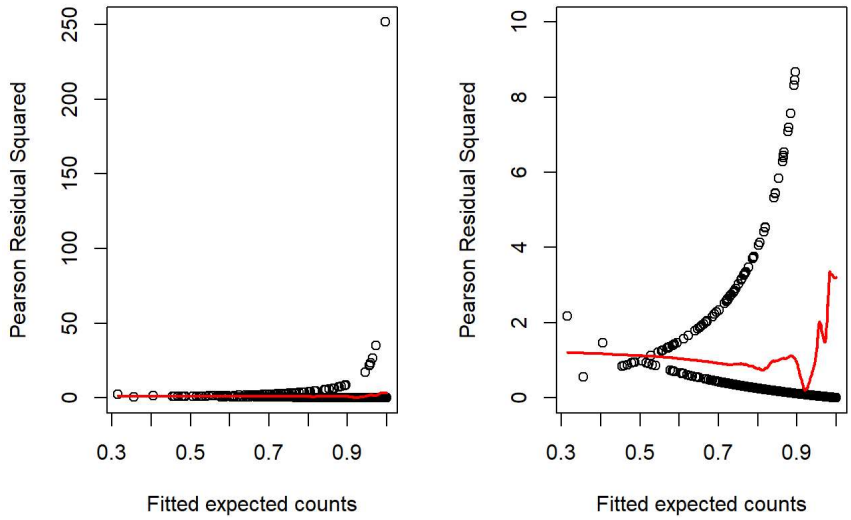

```
roc.curve
```

```
##
## Call:
## roc.formula(formula = MidwestSales$ac ~ fitted(logit2a))
##
## Data: fitted(logit2a) in 88 controls (MidwestSales$ac 0) < 434 cases (MidwestSales$ac 1).
## Area under the curve: 0.7649
```

c. The variance specification is not appropriate: heavy heteroskedasticity in Pearson residual.

```
par(mfrow=c(1,2))
presids = residuals(logit2a, type="pearson")
muhat = fitted(logit2a)
plot(muhat, presids^2, xlab="Fitted expected counts", ylab="Pearson Residual Squared")
sfit = supsmu(muhat, presids^2)
lines(sfit$x[order(sfit$x)] , sfit$y[order(sfit$x)], col="red", lwd=2)

plot(muhat, presids^2, xlab="Fitted expected counts", ylab="Pearson Residual Squared", ylim=c(0,10))
sfit = supsmu(muhat, presids^2)
lines(sfit$x[order(sfit$x)] , sfit$y[order(sfit$x)], col="red", lwd=2)
```



d. The observation with the highest leverage has a sqft of 1550 and a lot size of 14998 which are way below the average house. Also, it has a pool when the average house in the sample don't.

```
MidwestSales[which(hatvalues(logit2a) == max(hatvalues(logit2a))),]
```

```
##      id  price sqft bed bath ac garage pool year quality style   lot hwy
## 394 394 232900 1550   4    2  1      2    1 1962       3     2 14998   0
```

e. We are 95% confident that the true odds ratio(large to small) for ac comparing two houses that differ in sqft by 500 and lot size by 1500 is between 1.76 and 3.19.

```
linContr.glm(c("sqft" , "lot") , c(500,1500) , model=logit2a)
```

```
##
## Test of H_0: exp( 500*sqft + 1500*lot ) = 1 :
```

```
##   exp( Est )    se.est    zStat  pVal  ci95.lo  ci95.hi
## 1   2.368631 0.1515497 5.689963 1e-08 1.759941 3.187842
```

3. Holding other the same, younger people below the age of 35 has $e^{-1.32}$ times the estimated odds of using oral contraceptives than people who is 35 or older.
   Holding other the same, White has $e^{0.622}$ times the estimated odds of using oral contraceptives than non-White.
   Holding other the same, people who get more than 1 year of college education has $e^{0.501}$ times the estimated odds than people who has fewer than 1 year of college education.
   Holding other the same, people who are married has $e^{-0.46}$ times the estimated odds than people who are not married.

4. a. See below

```
nhanes = read.table("D:\\Coding\\Stats111\\Data\\nhaneshw.txt", header=TRUE)
nhanes$agegrp = cut(nhanes$age, breaks=c(0,30,40,50,60,71), right=FALSE)
lapply(split(nhanes, nhanes$male), summary)
```

```
## $`0`
##       age              wt              male        htn              rtid
##  Min.   :20.00   Min.   : 35.90   Min.   :0   Min.   :0.0000   Min.   :   2.0
##  1st Qu.:30.00   1st Qu.: 61.90   1st Qu.:0   1st Qu.:0.0000   1st Qu.: 870.5
##  Median :42.00   Median : 72.14   Median :0   Median :0.0000   Median :1764.0
##  Mean   :43.11   Mean   : 75.71   Mean   :0   Mean   :0.1751   Mean   :1758.4
##  3rd Qu.:56.00   3rd Qu.: 85.05   3rd Qu.:0   3rd Qu.:0.0000   3rd Qu.:2645.5
##  Max.   :70.00   Max.   :191.10   Max.   :0   Max.   :1.0000   Max.   :3528.0
##      agegrp
##  [0,30) :449
##  [30,40):401
##  [40,50):357
##  [50,60):278
##  [60,71):394
##
##
## $`1`
##       age              wt              male        htn              rtid
##  Min.   :20.00   Min.   : 42.70   Min.   :1   Min.   :0.0000   Min.   :   1.0
##  1st Qu.:32.00   1st Qu.: 72.10   1st Qu.:1   1st Qu.:0.0000   1st Qu.: 897.2
##  Median :44.00   Median : 82.00   Median :1   Median :0.0000   Median :1769.0
##  Mean   :45.12   Mean   : 84.84   Mean   :1   Mean   :0.2067   Mean   :1772.5
##  3rd Qu.:60.00   3rd Qu.: 94.50   3rd Qu.:1   3rd Qu.:0.0000   3rd Qu.:2648.8
##  Max.   :70.00   Max.   :193.30   Max.   :1   Max.   :1.0000   Max.   :3529.0
##      agegrp
##  [0,30) :319
##  [30,40):332
##  [40,50):330
##  [50,60):254
##  [60,71):415
##
```

b. $\ln \frac{p}{1-p} = -4.37 + 0.81 * I(30 <= age < 40) + 1.58 * I(40 <= age < 50) + 2.02 * I(50 <= age < 60) + 2.70 * I(60 <= age < 71) + 0.$
People who's in the age group of 30 - 40 has 2.2562 times the estimated odds than people who's below age of 30 or who's above 71.
H0: The null model with just an intercept fit the data well enough
Ha: The proposed model is better
test statistic: 387.337~chi_squared(4)
p-value: ~= 0
conclusion: we reject the null and conclude that the proposed model with age group is better.

```
fit1.full = glm( htn ~ factor(agegrp) + wt + male, family=binomial, data=nhanes )
glmCI( fit1.full, transform = F)
```

```
##                           Est ci95.lo ci95.hi  z value Pr(>|z|)
## (Intercept)           -4.3655 -4.8764 -3.8546 -16.7476   0.0000
## factor(agegrp)[30,40)  0.8137  0.3735  1.2538   3.6234   0.0003
## factor(agegrp)[40,50)  1.5765  1.1658  1.9873   7.5230   0.0000
## factor(agegrp)[50,60)  2.0198  1.6087  2.4310   9.6283   0.0000
## factor(agegrp)[60,71)  2.6991  2.3114  3.0867  13.6461   0.0000
## wt                     0.0153  0.0108  0.0198   6.6467   0.0000
## male                  -0.0311 -0.2154  0.1532  -0.3309   0.7407
```

```
fit1.red = glm( htn ~ wt + male, family=binomial, data=nhanes )
lrtest( fit1.red, fit1.full )
```

```
##
## Assumption: Model 1 nested within Model 2
```

```
##   Resid. Df Resid. Dev Df Deviance pValue
## 1      3526   3377.187
## 2      3522   2989.850  4  387.337      0
```

c. $\ln \frac{p}{1-p} = -4.97 + 0.77 * I(30 <= age < 40) + 2.10 * I(40 <= age < 50) + 2.69 * I(50 <= age < 60) + 3.49 * I(60 <= age < 71) + 0.$
For a male, being in the age group of 30-40 gives an additional $\exp 0.02$ times the estimated odds of hypertension than males who's below the age of 30 or who's above 71 in the model without interaction.
H0: All coefficients for the interaction terms
Ha: H0 does not hold
test statistic: 32.897 ~ chi_squared(4)
p-value: ~= 0
conclusion: We reject the null and conclude that at least one coefficient for the interaction term is nonzero.

```
fit2 = glm(htn ~ factor(agegrp) + wt + male + factor(agegrp)*male, family=binomial, data=nhanes)
glmCI(fit2, transform=FALSE)
```

```
##                                Est ci95.lo ci95.hi  z value Pr(>|z|)
## (Intercept)                -4.9745 -5.6985 -4.2505 -13.4667   0.0000
## factor(agegrp)[30,40)       0.7761  0.0029  1.5492   1.9673   0.0491
## factor(agegrp)[40,50)       2.1062  1.4184  2.7940   6.0022   0.0000
## factor(agegrp)[50,60)       2.6912  2.0070  3.3754   7.7093   0.0000
## factor(agegrp)[60,71)       3.4904  2.8310  4.1499  10.3733   0.0000
## wt                          0.0157  0.0112  0.0202   6.7781   0.0000
## male                        1.0054  0.2356  1.7751   2.5599   0.0105
## factor(agegrp)[30,40):male  0.0206 -0.9236  0.9648   0.0428   0.9659
## factor(agegrp)[40,50):male -0.9513 -1.8154 -0.0871  -2.1576   0.0310
## factor(agegrp)[50,60):male -1.2402 -2.1058 -0.3746  -2.8080   0.0050
## factor(agegrp)[60,71):male -1.4633 -2.2839 -0.6428  -3.4953   0.0005
```

```
lrtest(fit1.full, fit2)
```

```
##
## Assumption: Model 1 nested within Model 2
```

```
##   Resid. Df Resid. Dev Df Deviance pValue
## 1      3522   2989.850
## 2      3518   2956.953  4   32.897      0
```

d.
1. The estimated odds is 0.549, with a 95% CI of (0.45, 0.67)
2. The probability of hypertension for this same person is 0.35
3. The estimated odds ratio for hypertension is 2.39, with a 95% CI of (1.70, 3.37)
4. The estimated odds ratio is 3.99, with a 95% CI of (2.82, 5.65)

```
# aggregate(nhanes$wt, list(nhanes$male, nhanes$agegrp), mean)
linContr.glm(contr.names=c("(Intercept)", "factor(agegrp)[60,71)", "wt", "male", "factor(agegrp)[60,71):male"), contr.coef=c
(1,1,85.543,1,1), model=fit2)

linContr.glm(contr.names=c("(Intercept)", "factor(agegrp)[60,71)", "wt", "male", "factor(agegrp)[60,71):male"), contr.coef=c
(1,1,85.543,1,1), model=fit2, transform=FALSE)
exp(-0.599)/(1+exp(-0.599))

linContr.glm( contr.names=c( "factor(agegrp)[40,50)", "factor(agegrp)[60,71)", "factor(agegrp)[40,50):male", "factor(agegrp)
[60,71):male"), contr.coef=c(-1,1,-1,1), model=fit2 )

linContr.glm(contr.names=c( "factor(agegrp)[40,50)", "factor(agegrp)[60,71)"), contr.coef=c(-1,1), model=fit2)
```

```
##
## Test of H_0: exp( 1*(Intercept) + 1*factor(agegrp)[60,71) + 85.543*wt + 1*male + 1*factor(agegrp)[60,71):male ) = 1 :
##
##   exp( Est )    se.est     zStat  pVal   ci95.lo   ci95.hi
## 1  0.5491759 0.1033585 -5.798616 1e-08 0.4484691 0.6724972
##
## Test of H_0: 1*(Intercept) + 1*factor(agegrp)[60,71) + 85.543*wt + 1*male + 1*factor(agegrp)[60,71):male = 0 :
##
##          Est    se.est     zStat  pVal    ci95.lo    ci95.hi
## 1 -0.5993364 0.1033585 -5.798616 1e-08 -0.8019154 -0.3967574
## [1] 0.3545725
##
## Test of H_0: exp( -1*factor(agegrp)[40,50) + 1*factor(agegrp)[60,71) + -1*factor(agegrp)[40,50):male + 1*factor(agegrp)[6
0,71):male ) = 1 :
##
##   exp( Est )    se.est    zStat    pVal  ci95.lo  ci95.hi
## 1    2.39212 0.1750172 4.983398 6.2e-07 1.697494 3.370993
##
## Test of H_0: exp( -1*factor(agegrp)[40,50) + 1*factor(agegrp)[60,71) ) = 1 :
##
##   exp( Est )   se.est    zStat pVal  ci95.lo  ci95.hi
## 1     3.9918 0.177032 7.819164    0 2.821492 5.647532
```

e. Both age group and gender along with the interactions seem to be significant predictors of hypertension.