STAT 477/STAT 577
HW 8 - Solutions

Every year, graduating seniors submit applications for medical school. There are many factors medical schools use to select applicants. The data file **Med.csv** contains data on 55 seniors who applied to medical school, all from the same liberal arts college in the Midwest. For each applicant, the variables GPA, MCAT, Gender, Apps, and Acceptance were collected. Here is some information on these five variables:

- GPA - Cumulative undergraduate grade point average
- MCAT - Medical School Admission Test, a standardized test that measures aptitude and achievement for medical school
- Sex.1.0 - 1 = Female, 0 = Male
- Apps - Number of medical schools to which the student applied
- Acceptance - 1 = Accepted, 0 = Denied

Read in the data:

```
med.data<- read.csv(file.choose(), header = T)
```

1. Fit a logistic regression model for predicting the log odds of being accepted into medical school from the variables MCAT and GPA. Use this logistic regression to answer the following questions.

   Fit the model:

```
med1.model<- glm(Acceptance ~ MCAT + GPA, data = med.data,
                 family = binomial(link = "logit"))
summary(med1.model)
```

```
##
## Call:
## glm(formula = Acceptance ~ MCAT + GPA, family = binomial(link = "logit"),
##     data = med.data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -22.3727     6.4538  -3.467 0.000527 ***
## MCAT          0.1645     0.1032   1.595 0.110786
## GPA           4.6765     1.6416   2.849 0.004389 **
## ---
## 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: 54.014  on 52  degrees of freedom
## AIC: 60.014
##
## Number of Fisher Scoring iterations: 5
```

(a) (4 pts) Give the equation for predicting the log odds of acceptance from the variables MCAT and GPA.

Using the model output above, the predicted log odds of acceptance is:

$$\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = -22.3727 + 0.1645 * \text{MCAT}_i + 4.6765 * \text{GPA}_i$$

(b) (4 pts) Give the equation for predicting the probability of acceptance from the variables MCAT and GPA. Use this equation to predict the probability of acceptance for a student with a GPA of 3.54 and a MCAT score of 38.

The predicted probability of acceptance is:

$$\hat{p}_i = \frac{e^{-22.3727+0.1645*\text{MCAT}_i+4.6765*\text{GPA}_i}}{1+e^{-22.3727+0.1645*\text{MCAT}_i+4.6765*\text{GPA}_i}}$$

We can use this equation to predict the probability of acceptance when GPA = 3.54 and MCAT = 38 as:

$$\hat{p}_i = \frac{e^{-22.3727+0.1645*38+4.6765*3.54}}{1+e^{-22.3727+0.1645*38+4.6765*3.54}} = 0.6066$$

(c) (4 pts) Find and interpret a 95% confidence interval for the probability of acceptance for a student with a GPA of 3.54 and a MCAT score of 38.

```
student<- data.frame(MCAT = 38, GPA = 3.54)
glm.prob.ci(med1.model, newdata = student, 0.95)

## [[1]]
##         2.5      97.5
## 1 0.4159114 0.7695156
```

The confidence interval is from 0.4159 to 0.7695. This means we are 95% confident a student from this population with a GPA of 3.54 and a MCAT of 38 has a probability of being accepted to medical school between 0.4159 and 0.7695.

(d) (12 pts) Test for the significance of the overall model using the likelihood ratio test.

The null and alternative hypotheses for this test are:

$$H_0 : \beta_1 = \beta_2 = 0$$

$$H_a : \text{at least one } \beta_j \neq 0$$

We will first need to fit the model with just the intercept and then use the `anova()` function to obtain the test statistic and p-value for the test.

```
med0.model<- glm(Acceptance ~ 1, data = med.data,
                 family = binomial(link = "logit"))
anova(med0.model, med1.model, test = "Chisq")

## Analysis of Deviance Table
##
## Model 1: Acceptance ~ 1
## Model 2: Acceptance ~ MCAT + GPA
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        54     75.791
## 2        52     54.014  2   21.777 1.867e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The test statistic is 21.777 with a p-value < 0.0001. We conclude we have extremely strong evidence that at least one of the explanatory variables is important in explaining the probability of acceptance to medical school in this population of students.

(e) (12 pts - 6 each) Test for the significance of GPA and MCAT scores separately using the Wald test.

The Wald test information is contained in the summary output of the model. For MCAT:

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

Test Statistic: $z = 1.595$
p-value: 0.110786
Conclusion: We have little evidence that adding the MCAT score to the model that already contains GPA helps to explain the probability of acceptance to medical school in this population of students.

For GPA:

$$H_0 : \beta_2 = 0$$

$$H_a : \beta_2 \neq 0$$

Test Statistic: $z = 2.849$
p-value: 0.004389
Conclusion: We have strong evidence that adding GPA to the model that already contains MCAT score helps to explain the probability of acceptance to medical school in this population of students.

2. Now fit a logistic regression model for predicting the log odds of being accepted into medical school from the variables MCAT, GPA and Sex. Use this logistic regression to answer the following questions.

   Fit model:

```
med2.model<- glm(Acceptance ~ MCAT + GPA + Sex.1.0, data = med.data,
                 family = binomial(link = "logit"))
summary(med2.model)

##
## Call:
## glm(formula = Acceptance ~ MCAT + GPA + Sex.1.0, family = binomial(link = "logit
##     data = med.data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -25.2431     7.2019  -3.505 0.000456 ***
## MCAT          0.1809     0.1080   1.675 0.093946 .
## GPA           5.1392     1.8508   2.777 0.005491 **
## Sex.1.0       1.2580     0.7303   1.723 0.084965 .
## ---
## 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: 50.786  on 51  degrees of freedom
## AIC: 58.786
##
## Number of Fisher Scoring iterations: 5
```

   (a) (8 pts) Give the equation for predicting the log odds of acceptance from the variables MCAT and GPA for Females and the equation for predicting the log odds of acceptance from the variables MCAT and GPA for Males.

   Using the output above and the coding Sex = 1 for Females and Sex = 0 for Males we have the two equations below.

For Females:

$$\begin{aligned}\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) &= -23.9851 + 0.1809 * \text{MCAT}_i + 5.1392 * \text{GPA}_i + 1.2580 * (1) \\ &= -22.7271 + 0.1809 * \text{MCAT}_i + 5.1392 * \text{GPA}_i\end{aligned}$$

For Males:

$$\begin{aligned}\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) &= -23.9851 + 0.1809 * \text{MCAT}_i + 5.1392 * \text{GPA}_i + 1.2580 * (0) \\ &= -23.9851 + 0.1809 * \text{MCAT}_i + 5.1392 * \text{GPA}_i\end{aligned}$$

(b) (11 pts) Give the value of the coefficient for Sex in the model. Calculate and interpret a 95% confidence interval for the associated odds ratio.
The value of $\hat{\beta}_3 = 1.2580$. The 95% confidence interval is calculated as:

```
exp(confint(med2.model)[4,])

## Waiting for profiling to be done...

##      2.5 %     97.5 %
##  0.8949833 16.5025204
```

The confidence interval is from 0.8950 to 16.5025. This means we are 95% confident the odds of acceptance to medical school for females in this population of students is between 0.8950 to 16.5025 times the odds of acceptance to medical school for males in this population of students.

(c) (6 pts) Test for the significance of the variable Sex in the model with GPA and MCAT using the Wald test.
The Wald test is contained in the summary model output.

$$H_0 : \beta_3 = 0$$

$$H_a : \beta_3 \neq 0$$

Test Statistic: $z = 1.723$
p-value: 0.084965
Conclusion: We have weak evidence that the variable Sex helps to explain the probability of acceptance to medical school for student in this population in the model that already contains MCAT and GPA.

3. Add an interaction term between GPA and MCAT to the logistic regression model from Problem 1 above. Use this logistic regression to answer the following questions.

Fit the model:

```
med3.model<- glm(Acceptance ~ MCAT + GPA + MCAT:GPA, data = med.data,
                 family = binomial(link = "logit"))
summary(med3.model)
```

```
##
## Call:
## glm(formula = Acceptance ~ MCAT + GPA + MCAT:GPA, family = binomial(link = "logi
##     data = med.data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  17.0632    34.1200   0.500    0.617
## MCAT         -0.9359     0.9737  -0.961    0.336
## GPA          -6.6350    10.0837  -0.658    0.511
## MCAT:GPA      0.3154     0.2864   1.101    0.271
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 75.791  on 54  degrees of freedom
## Residual deviance: 53.186  on 51  degrees of freedom
## AIC: 61.186
##
## Number of Fisher Scoring iterations: 5
```

(a) (4 pts) Give the equation for predicting the log odds of acceptance from the variables MCAT and GPA and their interaction.

Using the model output above, the predicted log odds of acceptance is:

$$\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = 17.0632 - 0.9359 * \text{MCAT}_i - 6.6350 * \text{GPA}_i + 0.3154 * \text{MCAT}_i * \text{GPA}_i$$

(b) (5 pts) Give the equation for predicting the probability of acceptance from the variables MCAT and GPA and their interaction. Use this equation to predict the probability of acceptance for a student with a GPA of 3.54 and a MCAT score of 38. How does this value compare to the one you calculated in Problem 1, part (b)?

$$\hat{p}_i = \frac{e^{17.0632 - 0.9359 * \text{MCAT}_i - 6.6350 * \text{GPA}_i + 0.3154 * \text{MCAT}_i * \text{GPA}_i}}{1 + e^{17.0632 - 0.9359 * \text{MCAT}_i - 6.6350 * \text{GPA}_i + 0.3154 * \text{MCAT}_i * \text{GPA}_i}}$$

Substituting MCAT = 38 and GPA = 3.54 gives

$$\hat{p}_i = \frac{e^{17.0632 - 0.9359 * 38 - 6.6350 * 3.54 + 0.3154 * 38 * 3.54}}{1 + e^{17.0632 - 0.9359 * 38 - 6.6350 * 3.54 + 0.3154 * 38 * 3.54}} = 0.6080$$

The two predicted probabilities are very close.

(c) (4 pts) Find and interpret a 95% confidence interval for the probability of acceptance for a student with a GPA of 3.54 and a MCAT score of 38. How does this confidence interval compare to the one you calculated in Problem 1, part (c)?

```
glm.prob.ci(med3.model, newdata = student, 0.95)

## [[1]]
##        2.5     97.5
## 1 0.405431 0.777963
```

The confidence interval is from 0.4054 to 0.7780. This means we are 95% confident a student from this population with a GPA of 3.54 and a MCAT of 38 has a probability of being accepted to medical school between 0.4054 and 0.7780.

The two confidence intervals are very similar.

(d) (6 pts) Test for the significance of the interaction term in this model using the Wald test.

The Wald test is contained in the summary model output.

$$H_0 : \beta_3 = 0$$

$$H_a : \beta_3 \neq 0$$

Test Statistic: $z = 1.101$
p-value: 0.271
Conclusion: We have little to no evidence the interaction between MCAT and GPA helps to explain the probability of acceptance to medical school in this population of students.

4. (20 pts; 10 pts for model; 10 pts for summary of fit) Use the **step** function in R and the criteria AIC and BIC to find a good model to predict the probability of acceptance from the possible explanatory variables MCAT, GPA, Sex and Apps. Explain your selection process. Once you have found a good model, summarize the fit of the model using pseudo R^2, the Hosmer-Lemeshow goodness of fit test, a confusion table and associated statistics, and the ROC Curve and area under this curve.

First fit the model with all four explanatory variables:

```
med4.model<- glm(Acceptance ~ MCAT + GPA + Sex.1.0 + Apps, data = med.data,
                 family = binomial(link = "logit"))
summary(med4.model)


##
## Call:
## glm(formula = Acceptance ~ MCAT + GPA + Sex.1.0 + Apps, family = binomial(link =
##     data = med.data)
##
```

```
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -26.17567    7.60382  -3.442 0.000577 ***
## MCAT          0.18514    0.10911   1.697 0.089739 .
## GPA           5.28407    1.89209   2.793 0.005227 **
## Sex.1.0       1.23803    0.73275   1.690 0.091110 .
## Apps          0.03310    0.07475   0.443 0.657948
## ---
## 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: 50.590  on 50  degrees of freedom
## AIC: 60.59
##
## Number of Fisher Scoring iterations: 5
```

Next we can use the `step()` function to find a model using the AIC criterion.

```
medAIC.model<- step(med4.model, trace = 0)
summary(medAIC.model)


##
## Call:
## glm(formula = Acceptance ~ MCAT + GPA + Sex.1.0, family = binomial(link = "logit
##     data = med.data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -25.2431     7.2019  -3.505 0.000456 ***
## MCAT          0.1809     0.1080   1.675 0.093946 .
## GPA           5.1392     1.8508   2.777 0.005491 **
## Sex.1.0       1.2580     0.7303   1.723 0.084965 .
## ---
## 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: 50.786  on 51  degrees of freedom
## AIC: 58.786
##
## Number of Fisher Scoring iterations: 5
```

Next we will use the same function, but will use the BIC criterion.

```
medBIC.model<- step(med4.model, k = log(55), trace = 0)
summary(medBIC.model)
```

```
##
## Call:
## glm(formula = Acceptance ~ GPA, family = binomial(link = "logit"),
##     data = med.data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -19.207      5.629  -3.412 0.000644 ***
## GPA            5.454      1.579   3.454 0.000553 ***
## ---
## 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: 56.839  on 53  degrees of freedom
## AIC: 60.839
##
## Number of Fisher Scoring iterations: 4
```

The `medAIC.model` has three explanatory variables: GPA, MCAT, and Sex while the `medBIC.model` has only one explanatory variable: GPA. At this point, I would like to look at a Likelihood ratio test between the two models to see if adding MCAT and Sex to the model that already includes GPA is helpful in explaining the probability of acceptance to medical school in this population of students.

```
anova(medBIC.model, medAIC.model, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Acceptance ~ GPA
## Model 2: Acceptance ~ MCAT + GPA + Sex.1.0
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        53     56.839
## 2        51     50.786  2   6.0526   0.0485 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since this test has a test statistic 6.0526 and p-value = 0.0485, we have moderate evidence these two variables are helpful in explaining the probability of acceptance to medical school in this population of students even after taking into account the student's GPA.

My decision is to use the `medAIC.model` with all three variables. You can certainly use the smaller model with only GPA. I think it is clear that GPA should be in any model you choose and Apps should not be in any of your models.

McFadden's R^2:

```
McFR2(medAIC.model)

## [1] 0.3299149
```

Hosmer-Lemeshow's Goodness of Fit Test:

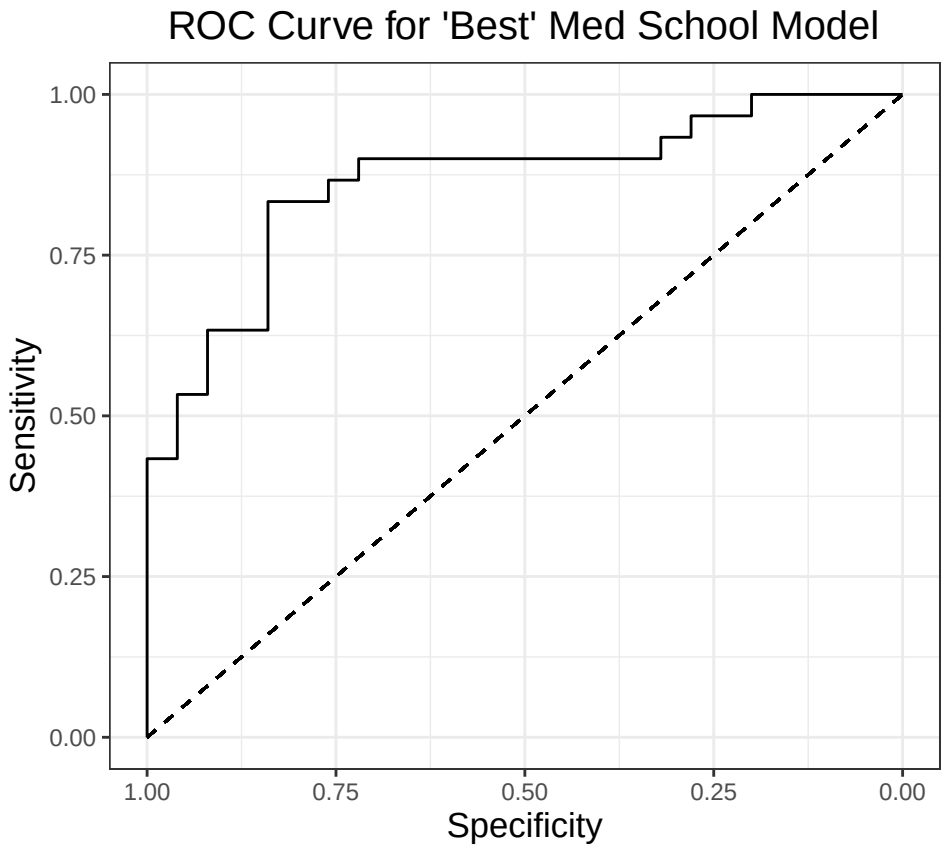
```
hoslem.test(med.data$Acceptance, medAIC.model$fitted.values, g = 5)

##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  med.data$Acceptance, medAIC.model$fitted.values
## X-squared = 7.7273, df = 3, p-value = 0.052
```

Confusion Table and Related Statistics:

```
confusion.glm(medAIC.model)

## $`Confusion Table`
##          predicted
## observed  0  1
##        0 21  4
##        1  5 25
##
## $Agreement
## [1] 0.8363636
##
## $Sensitivity
##         1
## 0.8333333
##
## $Specificity
##    0
## 0.84
```

Roc Curve:

```
med.roc<- roc(med.data$Acceptance ~ medAIC.model$fitted.values)

## Setting levels:  control = 0, case = 1
## Setting direction:  controls < cases

ggroc(med.roc)+
  theme_bw()+
  theme(axis.title.y = element_text(size = rel(1.2)),
        axis.title.x = element_text(size = rel(1.2)),
        axis.text.x = element_text(size = rel(1)),
        axis.text.y = element_text(size = rel(1)),
        plot.title = element_text(hjust=0.5, size = rel(1.4)))+
  geom_segment(aes(x = 1, xend = 0, y = 0, yend = 1),
               linetype="dashed")+
  labs(x = "Specificity",
       y = "Sensitivity",
       title = "ROC Curve for 'Best' Med School Model")
```



Area Under Roc Curve:

```
auc(med.roc)

## Area under the curve: 0.8653
```

Summary: Given the model fit and prediction checks above, the model is a good fit for the data. It does a good job in predicting the observations, and about the same in separately predicting the successes and failures. The ROC curve looks good and the area under the ROC curve is a high value.