HOMEWORK 3 - Exercises in Chapter 2

STOR 590, FALL 2020

Rui Li

8/27/2020

Instructions

Chapter 2, questions 2 and 4. Hand in on gradescope.

Exercises

2. The National Institute of Diabetes and Digestive and Kidney Diseases conducted a study on 768 adult female Pima Indians living near Phoenix. The purpose of the study was to investigate factors related to diabetes. The data may be found in the dataset *pima*.

(a) Create a factor version of the test results and use this to produce an interleaved histogram to show how the distribution of insulin differs between those testing positive and negative. Do you notice anything unbelievable about the plot?

```
#Insulin: 2-Hour serum insulin (mu U/ml)
#Test: coded 0 if negative, 1 if positive
#Import and summary dataset pima
library("faraway")

## Warning: package 'faraway' was built under R version 3.6.3

summary(pima)

##     pregnant         glucose        diastolic         triceps
##  Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
##  Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
##  Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
##  3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
##     insulin           bmi           diabetes           age
##  Min.   :  0.0   Min.   : 0.00   Min.   :0.0780   Min.   :21.00
##  1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437   1st Qu.:24.00
##  Median : 30.5   Median :32.00   Median :0.3725   Median :29.00
##  Mean   : 79.8   Mean   :31.99   Mean   :0.4719   Mean   :33.24
##  3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00
##  Max.   :846.0   Max.   :67.10   Max.   :2.4200   Max.   :81.00
##       test
##  Min.   :0.000
##  1st Qu.:0.000
```

```
##  Median :0.000
##  Mean   :0.349
##  3rd Qu.:1.000
##  Max.   :1.000
```

```
head(pima)
```

```
##   pregnant glucose diastolic triceps insulin  bmi diabetes age test
## 1        6     148        72      35       0 33.6    0.627  50    1
## 2        1      85        66      29       0 26.6    0.351  31    0
## 3        8     183        64       0       0 23.3    0.672  32    1
## 4        1      89        66      23      94 28.1    0.167  21    0
## 5        0     137        40      35     168 43.1    2.288  33    1
## 6        5     116        74       0       0 25.6    0.201  30    0
```

```
#Factor version as test results
pima$test = as.factor(pima$test)
levels(pima$test) = c("negative","positive")

#Interleaved historgram
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

```
ggplot(pima) +
  geom_histogram(aes(x=insulin, y=..density.., fill=test), position="dodge",b
inwidth = 50) +
  ggtitle("Insulin Difference on Test Results")+
  theme(plot.title = element_text(size=15, face="bold", hjust=0.5))
```

Answers: The insulin density, which equals to zero, is extremely high. It is unusual and may indicate some missing datas.

(b) Replace the zero values of insulin with the missing value code NA. Recreate the interleaved histogram plot and comment on the distribution.

```
#Replace zero with NA
pima$insulin[pima$insulin==0]=NA

#Interleaved historgram
library(ggplot2)
ggplot(pima) +
  geom_histogram(aes(x=insulin, y=..density.., fill=test), position="dodge",b
inwidth = 50) +
  ggtitle("Insulin Difference on Test Results")+
  theme(plot.title = element_text(size=15, face="bold", hjust=0.5))

## Warning: Removed 374 rows containing non-finite values (stat_bin).
```

Answers: According to the plot above, the negative results have a relatively larger density in the low insulin level than the positive results. Therefore, the positive test results generally have a higher insulin standard than the negative test results do.

(c) Replace the incredible zeroes in other variables with the missing value code. Fit a model with the result of the diabetes test as the response and all the other variables as predictors. How many observations were used in the model fitting? Why is this less than the number of observations in the data frame.

```
#Overview pima
summary(pima)
```

```
##     pregnant         glucose       diastolic         triceps
##  Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
##  Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
##  Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
##  3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
##
##     insulin           bmi          diabetes          age
##  Min.   : 14.00   Min.   : 0.00   Min.   :0.0780   Min.   :21.00
##  1st Qu.: 76.25   1st Qu.:27.30   1st Qu.:0.2437   1st Qu.:24.00
##  Median :125.00   Median :32.00   Median :0.3725   Median :29.00
##  Mean   :155.55   Mean   :31.99   Mean   :0.4719   Mean   :33.24
##  3rd Qu.:190.00   3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00
##  Max.   :846.00   Max.   :67.10   Max.   :2.4200   Max.   :81.00
```

```
##  NA's   :374
##        test
##  negative:500
##  positive:268
##
##
##
##
##
```

```
#help(pima)
```

According to the help manual:

glucose: Plasma glucose concentration at 2 hours in an oral glucose tolerance test.

diastolic: Diastolic blood pressure (mm Hg).

triceps: Triceps skin fold thickness (mm).

bmi: Body mass index (weight in kg/(height in metres squared)).

These variables are unusual to have zero value, and should be replaced with missing value NA.

```
#Replace with missing value
pima$glucose[pima$glucose==0]=NA
pima$diastolic[pima$diastolic==0]=NA
pima$triceps[pima$triceps==0]=NA
pima$bmi[pima$bmi==0]=NA

#Fit a model
full.glm = glm(test ~ .,family = "binomial", data= pima)
summary(full.glm)
```

```
##
## Call:
## glm(formula = test ~ ., family = "binomial", data = pima)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7823  -0.6603  -0.3642   0.6409   2.5612
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.004e+01  1.218e+00  -8.246  < 2e-16 ***
## pregnant     8.216e-02  5.543e-02   1.482  0.13825
## glucose      3.827e-02  5.768e-03   6.635 3.24e-11 ***
## diastolic   -1.420e-03  1.183e-02  -0.120  0.90446
## triceps      1.122e-02  1.708e-02   0.657  0.51128
## insulin     -8.253e-04  1.306e-03  -0.632  0.52757
## bmi          7.054e-02  2.734e-02   2.580  0.00989 **
```

```
## diabetes     1.141e+00  4.274e-01   2.669  0.00760 **
## age          3.395e-02  1.838e-02   1.847  0.06474 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 344.02  on 383  degrees of freedom
##   (376 observations deleted due to missingness)
## AIC: 362.02
##
## Number of Fisher Scoring iterations: 5
```

```
#Extract the Number of Observations from the Fit.
library("stats")
print(paste0("The observations of model are ", nobs(full.glm),"."))
```

```
## [1] "The observations of model are 392."
```

```
print(paste0("The observations of dataframe are ", nrow(pima),"."))
```

```
## [1] "The observations of dataframe are 768."
```

Answers: As from the result above, 392 observations are used in the model fitting. The number is much smaller compared to the sample size of the dataframe, because the samples that exist missing value cannot be employed in the model fitting.

(d) Refit the model but now without the insulin and triceps predictors. How many observations were used in fitting this model? Devise a test to compare this model with that in the previous question.

```
#Refit the model
re.glm = glm(test ~ pregnant + glucose + diastolic + bmi + diabetes + age, fa
mily = "binomial", data = pima)
summary(re.glm)
```

```
##
## Call:
## glm(formula = test ~ pregnant + glucose + diastolic + bmi + diabetes +
##     age, family = "binomial", data = pima)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8062  -0.7229  -0.4049   0.7173   2.3959
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.962146   0.820892 -10.918  < 2e-16 ***
## pregnant     0.117863   0.033418   3.527  0.00042 ***
## glucose      0.035194   0.003605   9.763  < 2e-16 ***
```

```
## diastolic   -0.008916   0.008618  -1.035  0.30084
## bmi          0.090926   0.015740   5.777 7.61e-09 ***
## diabetes     0.960515   0.306415   3.135  0.00172 **
## age          0.016944   0.009834   1.723  0.08489 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 931.94  on 723  degrees of freedom
## Residual deviance: 672.86  on 717  degrees of freedom
##   (44 observations deleted due to missingness)
## AIC: 686.86
##
## Number of Fisher Scoring iterations: 5
```

```
#Extract the Number of Observations from the Fit.
library("stats")
print(paste0("The observations of refit model are ", nobs(re.glm),"."))
```

```
## [1] "The observations of refit model are 724."
```

Answers: According to the result above, the observations of the refit model are 724, which is different to the observations of the previous model. Therefore, we need to remove samples with missing value to have the same observations on two models.

```
#Remove samples with missing value
new.pima = na.omit(pima)

#Compare two models
full.glm = glm(test ~ ., family = "binomial", data = new.pima)
re.glm = glm(test ~ pregnant + glucose + diastolic + bmi + diabetes + age, fa
mily = "binomial", data = new.pima)
anova(re.glm, full.glm, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model 1: test ~ pregnant + glucose + diastolic + bmi + diabetes + age
## Model 2: test ~ pregnant + glucose + diastolic + triceps + insulin + bmi +

##     diabetes + age
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       385     344.88
## 2       383     344.02  2  0.85931   0.6507
```

Answers: The analysis of deviance table displays that p-value is greater than 0.05. Thus, *insulin*, and *triceps* are not significant in the model. We can exclude them from the full model, and the refit model performs better.

(e) Use AIC to select a model. You will need to take account of the missing values. Which predictors are selected? How many cases are used in your selected model?

```r
#Create a full model beforre tested with AIC
newfull.glm = glm(test ~ pregnant + glucose + diastolic + bmi + diabetes + ag
e, family = "binomial", data = new.pima)

#Select a model
select.glm = step(newfull.glm, trace = 0)
summary(select.glm)

##
## Call:
## glm(formula = test ~ pregnant + glucose + bmi + diabetes + age,
##     family = "binomial", data = new.pima)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8827  -0.6535  -0.3694   0.6521   2.5814
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.992080   1.086866  -9.193  < 2e-16 ***
## pregnant     0.083953   0.055031   1.526 0.127117
## glucose      0.036458   0.004978   7.324 2.41e-13 ***
## bmi          0.078139   0.020605   3.792 0.000149 ***
## diabetes     1.150913   0.424242   2.713 0.006670 **
## age          0.034360   0.017810   1.929 0.053692 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 344.89  on 386  degrees of freedom
## AIC: 356.89
##
## Number of Fisher Scoring iterations: 5

#Extract the Number of Observations from the Fit.
library("stats")
print(paste0("The observations of selected model are ", nobs(re.glm),"."))

## [1] "The observations of selected model are 392."
```

Answers: According to the results above, the model selects five predictors, which are *pregnant, glucose, bmi, diabetes*, and *age*. There are 392 cases in the selected model.

(f) Create a variable that indicates whether the case contains a missing value. Use this variable as a predictor of the test result. Is missingness associated with the test result? Refit the selected model, but now using as much of the data as reasonable. Explain why it is appropriate to do this.

```r
#Create variable to indicate missing value
pima$`miss` = ifelse(rowSums(is.na(pima))>0,1,0)

#Association between missingness and test result
miss.glm = glm(test ~ miss, family = "binomial", data = pima)
summary(miss.glm)
```

```
##
## Call:
## glm(formula = test ~ miss, family = "binomial", data = pima)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9564  -0.9564  -0.8977   1.4159   1.4857
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.7008     0.1073  -6.533 6.47e-11 ***
## miss          0.1558     0.1515   1.028    0.304
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 993.48  on 767  degrees of freedom
## Residual deviance: 992.43  on 766  degrees of freedom
## AIC: 996.43
##
## Number of Fisher Scoring iterations: 4
```

```r
#Check significance
anova(miss.glm, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: test
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                   767     993.48
## miss  1   1.0579       766     992.43   0.3037
```

```r
#Refit selected mode
fit.glm = glm(test ~ pregnant + glucose + bmi + diabetes + age, family = "bin
omial", data = new.pima)
summary(fit.glm)
```

```
##
## Call:
## glm(formula = test ~ pregnant + glucose + bmi + diabetes + age,
##     family = "binomial", data = new.pima)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8827  -0.6535  -0.3694   0.6521   2.5814
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.992080   1.086866  -9.193  < 2e-16 ***
## pregnant     0.083953   0.055031   1.526 0.127117
## glucose      0.036458   0.004978   7.324 2.41e-13 ***
## bmi          0.078139   0.020605   3.792 0.000149 ***
## diabetes     1.150913   0.424242   2.713 0.006670 **
## age          0.034360   0.017810   1.929 0.053692 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 344.89  on 386  degrees of freedom
## AIC: 356.89
##
## Number of Fisher Scoring iterations: 5
```

Answers: According to the results above, the p-value of missingnes is 0.304 > 5%, indicating that missingness is not significant, and does not associate with the test result. Therefore, we can exclude all samples with missing value, which does not have significant effect on our selected model. My final fitted model is *glm(test ~ pregnant + glucose + bmi + diabetes + age, family = "binomial", data = new.pima).*

(g) Using the last fitted model of the previous question, what is the difference in the odds of testing positive for diabetes for a woman with a BMI at the first quartile compared with a woman at the third quartile, assuming that all other factors are held constant? Give a confidence interval for this difference.

```r
#Get BMI value for Q1 and Q3
bmi.25th = quantile(new.pima$bmi, 0.25)
bmi.75th = quantile(new.pima$bmi, 0.75)

#BMI Coefficent
bmi.coef = coef(fit.glm)[4]

#Difference in odds
diff = bmi.coef*(bmi.25th-bmi.75th)
diff.odds = exp(diff)/(1+exp(diff))
print(diff.odds)
```

```
##       bmi
## 0.3363045

#Confidence Interval
bmi.conf = confint(fit.glm,'bmi')
odds_ratio = exp(bmi.conf*(bmi.25th-bmi.75th))
odds_ratio/(1+odds_ratio)

##     2.5 %    97.5 %
## 0.4165102 0.2606397
```

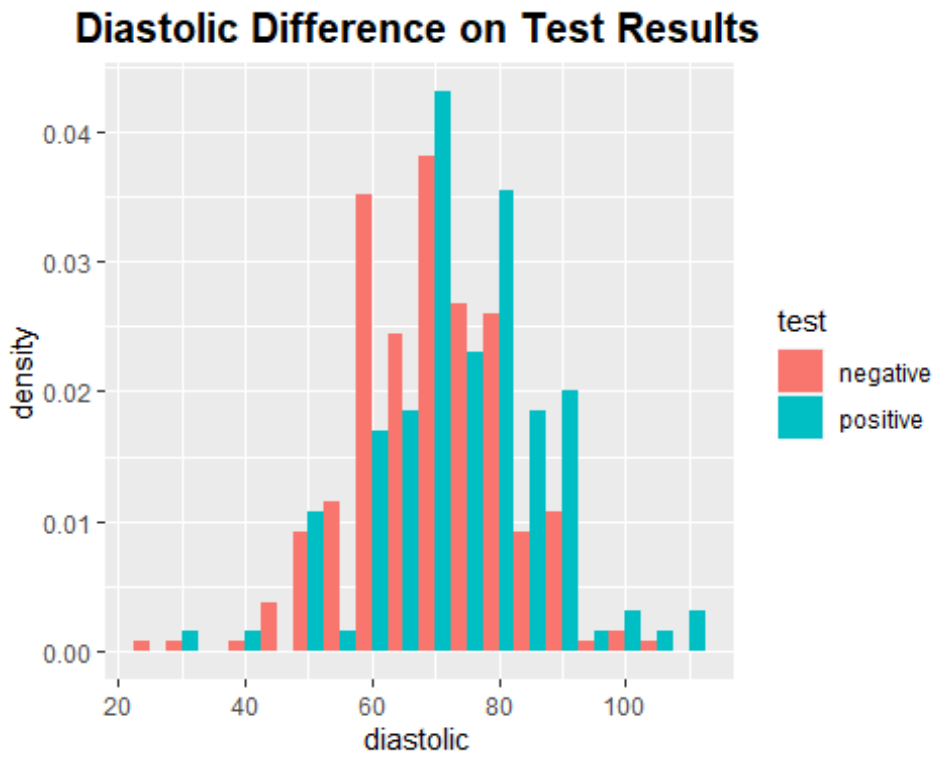
Answers: According to the results above, with other factors constant the difference in the odds of testing positive for diabetes for a woman with a BMI at the first quartile compared with a woman at the third quartile is 0.336, with a 95% interval between 0.26 and 0.42.

(h) Do women who test positive have higher diastolic blood pressures? Is the diastolic blood pressure significant in the regression model? Explain the distinction between the two questions and discuss why the answers are only apparently contradictory.

```
#Relationship between diastolic and test results
cor(new.pima$diastolic,as.numeric(new.pima$test))

## [1] 0.1926733

#Interleaved historgram
library(ggplot2)
ggplot(new.pima) +
  geom_histogram(aes(x=diastolic, y=..density.., fill=test), position="dodge",
binwidth = 5) +
  ggtitle("Diastolic Difference on Test Results")+
  theme(plot.title = element_text(size=15, face="bold", hjust=0.5))
```




```
#Regression model
mol = glm(test ~ pregnant + glucose + diastolic + bmi + diabetes + age + insu
lin + triceps,family = "binomial", data=pima)
summary(mol)

##
## Call:
## glm(formula = test ~ pregnant + glucose + diastolic + bmi + diabetes +
##     age + insulin + triceps, family = "binomial", data = pima)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7823  -0.6603  -0.3642   0.6409   2.5612
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.004e+01  1.218e+00  -8.246  < 2e-16 ***
## pregnant     8.216e-02  5.543e-02   1.482  0.13825
## glucose      3.827e-02  5.768e-03   6.635 3.24e-11 ***
## diastolic   -1.420e-03  1.183e-02  -0.120  0.90446
## bmi          7.054e-02  2.734e-02   2.580  0.00989 **
## diabetes     1.141e+00  4.274e-01   2.669  0.00760 **
## age          3.395e-02  1.838e-02   1.847  0.06474 .
## insulin     -8.253e-04  1.306e-03  -0.632  0.52757
## triceps      1.122e-02  1.708e-02   0.657  0.51128
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 344.02  on 383  degrees of freedom
##   (376 observations deleted due to missingness)
## AIC: 362.02
##
## Number of Fisher Scoring iterations: 5
```

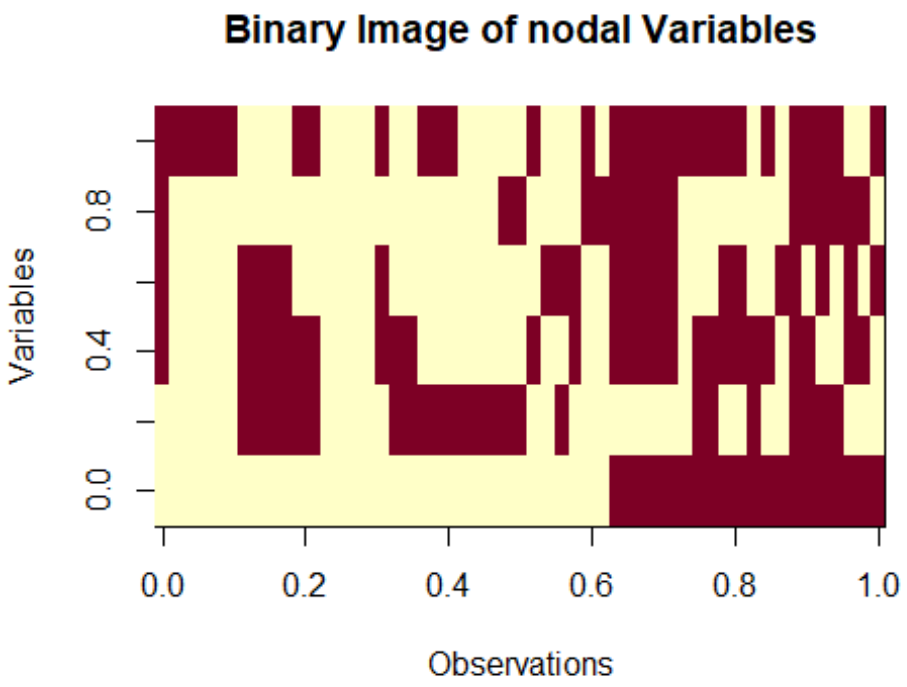
Answers: According to the results above, *diastolic* has a postive correlation with *test* results. However, from the histogram, we can see that the positive and negative cases have similar distribution in diastolic level. Also, according to our glm model, *diastolic* is not significant enough for the test response, since the p-value is 0.9 much larger than 5%. Therefore, two questions have contradictory answers.

4. Treatment of prostate cancer depends on whether the cancer has spread to the surrounding lymph nodes. This can be determined using a surgical procedure but it would be better if noninvasive methods could be used. Load in the data and learn about the variables by:

```
data(nodal, package="boot")
help(nodal, package="boot")
```

(a) A plot consisting of a binary image of the data can be constructed as:

```
nodal$m <- NULL
image(as.matrix(nodal[order(nodal$r),]), xlab = "Observations", ylab = "Varia
bles", main = "Binary Image of nodal Variables")
```



Improve this plot by ordering the cases on the response and labeling the axes informatively using the axis command.

Answers: The binary image above shows that each column represents a sample, and each row represents a variable. The response is at the bottom row, and the value of each cell is plot by binary color.

(b) Fit an appropriate model with nodal outcome as the response and the other five variables as predictors. Is there evidence that at least some of the five predictors are related to the response?

```
#Model with nodal outcome
full.glm = glm(r ~ ., family = "binomial", data = nodal)
summary(full.glm)

##
## Call:
## glm(formula = r ~ ., family = "binomial", data = nodal)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3317  -0.6653  -0.2999   0.6386   2.1502
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.0794     0.9868  -3.121   0.0018 **
## aged         -0.2917     0.7540  -0.387   0.6988
```

```
## stage         1.3729     0.7838   1.752   0.0799 .
## grade         0.8720     0.8156   1.069   0.2850
## xray          1.8008     0.8104   2.222   0.0263 *
## acid          1.6839     0.7915   2.128   0.0334 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 70.252  on 52  degrees of freedom
## Residual deviance: 47.611  on 47  degrees of freedom
## AIC: 59.611
##
## Number of Fisher Scoring iterations: 5
```

Answers: According to the result above, *xray* and *acid* have p-value < 0.05, and *stage* has p-value < 0.1, indicating that these three predictors have significant effect on the response.

(c) Fit a smaller model that removes aged and grade from the model. Can this smaller model be used in preference to the larger model?

```
#Fit a smaller model
small.glm = glm(r ~ xray + acid + stage, family = "binomial", data = nodal)
summary(small.glm)
```

```
##
## Call:
## glm(formula = r ~ xray + acid + stage, family = "binomial", data = nodal)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1231  -0.6620  -0.3039   0.4710   2.4892
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.0518     0.8420  -3.624  0.00029 ***
## xray          1.9116     0.7771   2.460  0.01390 *
## acid          1.6378     0.7539   2.172  0.02983 *
## stage         1.6453     0.7297   2.255  0.02414 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 70.252  on 52  degrees of freedom
## Residual deviance: 49.180  on 49  degrees of freedom
## AIC: 57.18
##
## Number of Fisher Scoring iterations: 5
```

```r
#Compare two models
anova(small.glm, full.glm, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model 1: r ~ xray + acid + stage
## Model 2: r ~ aged + stage + grade + xray + acid
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        49     49.180
## 2        47     47.611  2   1.5696   0.4562
```

Answers: According to the result above, the p-value is 0.4562 > 0.1, indicating that the deleted predictors do not have significant impact on the model. Therefore, the smaller model has better performance than the full model.

(d) How much does having a serious x-ray result increase the odds of nodal involvement compared to a nonserious result? (Use the smaller model.) Give a 95% confidence interval for the odds.

```r
#Odds Ratio with 95% confidence
xray.conf=confint(small.glm,'xray')
diff.xray= exp(xray.conf)/(1+exp(xray.conf))
diff.xray
```

```
##     2.5 %    97.5 %
## 0.6123535 0.9725957
```

Answers: The result shows that the 95% confidence of the odds ratio is between 0.61 to 0.97.

(e) Fit a model with all five predictors and all their two-way interactions. Explain why the standard errors of the coefficients are so large.

```r
#Full model with interactions
int.full.glm = glm(r ~ .^2, family = "binomial", data = nodal)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
summary(int.full.glm)
```

```
##
## Call:
## glm(formula = r ~ .^2, family = "binomial", data = nodal)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.89302  -0.00016   0.00000   0.00015   1.89302
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -38.55   31256.83  -0.001    0.999
## aged          -54.81   32392.28  -0.002    0.999
```

```
## stage          74.21   11937.81   0.006    0.995
## grade          38.55   31256.83   0.001    0.999
## xray           17.18   16447.57   0.001    0.999
## acid           36.94   31256.83   0.001    0.999
## aged:stage     18.46    7928.05   0.002    0.998
## aged:grade     31.97   38433.53   0.001    0.999
## aged:xray      56.96   17366.75   0.003    0.997
## aged:acid      36.35   31407.10   0.001    0.999
## stage:grade   -92.34   14996.24  -0.006    0.995
## stage:xray    -17.06   28388.60  -0.001    1.000
## stage:acid    -72.60   11937.81  -0.006    0.995
## grade:xray     36.50   31778.73   0.001    0.999
## grade:acid     54.48   31841.24   0.002    0.999
## xray:acid     -35.70    8157.65  -0.004    0.997
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 70.252  on 52  degrees of freedom
## Residual deviance: 29.542  on 37  degrees of freedom
## AIC: 61.542
##
## Number of Fisher Scoring iterations: 20
```

Answers: The standard errors are so large because we include all the two-way interactions and make multicollinearity appear.

(f) Use the bias-reduced model fitting method described in the chapter to fit the model of the previous question. Which interaction is largest?

```
#Use the bias-reduced model to fit the full interaction model
library(brglm)
full.bglm <- brglm(r ~ .^2, family = "binomial", data = nodal)
summary(full.bglm)

##
## Call:
## brglm(formula = r ~ .^2, family = "binomial", data = nodal)
##
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.6438     1.6586  -1.594    0.111
## aged         -0.6995     1.8290  -0.382    0.702
## stage         2.3156     1.8196   1.273    0.203
## grade         2.6023     1.9506   1.334    0.182
## xray          0.6900     2.0954   0.329    0.742
## acid          1.2729     1.8124   0.702    0.483
## aged:stage    0.3728     1.8119   0.206    0.837
## aged:grade   -1.5480     1.8594  -0.833    0.405
## aged:xray     1.4587     1.9780   0.737    0.461
```

```
## aged:acid     0.3634     1.7722   0.205    0.838
## stage:grade  -2.8219     1.8258  -1.546    0.122
## stage:xray    0.8375     2.2944   0.365    0.715
## stage:acid   -0.9608     1.6387  -0.586    0.558
## grade:xray   -0.1890     2.2626  -0.084    0.933
## grade:acid    0.5575     1.8918   0.295    0.768
## xray:acid    -0.2560     1.8611  -0.138    0.891
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 45.450  on 52  degrees of freedom
## Residual deviance: 38.997  on 37  degrees of freedom
## Penalized deviance: 45.20152
## AIC:  70.997
```

Answers: From the result above, we can see that the intercation between *stage* and *grade* is the largest, and has the biggest absolute coefficient of -2.82.

(g) If the predicted response probability exceeds 0.5, the case is classified positively and, if not, negatively. Use the bias-reduced model to classify the cases in the dataset. Compare these to the actual classifications. How many were wrongly classified? Repeat this comparison for the model in (b). Do you think these misclassification rates are a reasonable estimate of how these models will perform in the future?

```
#Predicted response based on bias-reduced model
pre.reponse = predict(full.bglm, type = "response")

#Make classification
nodal$`predict` = ifelse(pre.reponse>0.5,1,0)

#Compare to actual classifications
class.bias.error = sum(nodal$r!=nodal$predict)
error.rates = class.bias.error/nrow(nodal)
print(class.bias.error)

## [1] 8

print(error.rates)

## [1] 0.1509434

#Predicted response based on bias-reduced model
pre.reponse = predict(full.glm, type = "response")

#Make classification
nodal$`predict` = ifelse(pre.reponse>0.5,1,0)

#Compare to actual classifications
class.full.error = sum(nodal$r!=nodal$predict)
error.rates = class.full.error/nrow(nodal)
print(class.full.error)
```

```
## [1] 10
```

```
print(error.rates)
```

```
## [1] 0.1886792
```

Answers: From the results above, the bias-reduced model has 8 misclassification with 15% error rates, while the model in (b) has 10 misclassification with 18% error rates. Since the error rates of the two models are so closed, we can conclude that misclassification rates are a reasonable estimate of how these models will perform in the future.