# 550.413 Assignment 4.5

Fall 2016

**Instruction:** This assignment consists of 4 problems. The assignment is due on **Wednesday, December 7, 2016** at 3pm, in class. If you cannot make it to class, please leave the assignment under the door at Whitehead Hall 306E and email the course instructor. If possible, please type up your assignments, preferably using LaTeX.

## Problem 1: (20pts)

The next problem use the prostate dataset from **faraway**.

1. With lpsa as the response variable and the other variables as a predictor variable, perform variable selection using
   (a) Backward elimination.
   (b) AIC
   (c) Adjusted $R^2$.
   (d) Mallows $C_p$.

   Please provide plots of your variable selection procedure, when necessary.
2. Try out ridge regression with the above dataset. Split your dataset into two parts of roughly the same proportion. Use one part of the data to estimate the coefficients using ridge regression. Compute the fitted values for lpsa based on the other part of the data. Evaluate the mean square error of the fitted values. How did you select the amount of shrinkage ?

## Solution:

```
data(prostate)
library("leaps")
b <- regsubsets(lpsa ~ ., data = prostate)
bs <- summary(b)
bs
```

```
## Subset selection object
## Call: regsubsets.formula(lpsa ~ ., data = prostate)
## 8 Variables  (and intercept)
##         Forced in Forced out
## lcavol      FALSE      FALSE
## lweight     FALSE      FALSE
## age         FALSE      FALSE
## lbph        FALSE      FALSE
## svi         FALSE      FALSE
## lcp         FALSE      FALSE
## gleason     FALSE      FALSE
## pgg45       FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          lcavol lweight age lbph svi lcp gleason pgg45
## 1  ( 1 ) "*"    " "     " " " "  " " " " " " "     " "
## 2  ( 1 ) "*"    "*"     " " " "  " " " " " " "     " "
## 3  ( 1 ) "*"    "*"     " " " "  "*" " " " " "     " "
## 4  ( 1 ) "*"    "*"     " " "*"  "*" " " " " "     " "
## 5  ( 1 ) "*"    "*"     "*" "*"  "*" " " " " "     " "
## 6  ( 1 ) "*"    "*"     "*" "*"  "*" " " " " "     "*"
## 7  ( 1 ) "*"    "*"     "*" "*"  "*" "*" " " "     "*"
## 8  ( 1 ) "*"    "*"     "*" "*"  "*" "*" "*"     "*"

plot(2:9, bs$adjr2, xlab = "# of predictors", ylab = "Adjusted R^2")
plot(2:9, bs$cp, xlab = "# of predictors", ylab = "Mallow's Cp")
plot(2:9, length(prostate$lpsa) * log(bs$rss) + 2 * (2:9), ylab = "AIC", xlab = "# of predic
mod.full <- lm(lpsa ~ ., data = prostate)
summary(mod.full)

##
## Call:
## lm(formula = lpsa ~ ., data = prostate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7331 -0.3713 -0.0170  0.4141  1.6381
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.669337   1.296387   0.516  0.60693
## lcavol       0.587022   0.087920   6.677 2.11e-09 ***
## lweight      0.454467   0.170012   2.673  0.00896 **
## age         -0.019637   0.011173  -1.758  0.08229 .
## lbph         0.107054   0.058449   1.832  0.07040 .
## svi          0.766157   0.244309   3.136  0.00233 **
```

```
## lcp         -0.105474   0.091013  -1.159  0.24964
## gleason      0.045142   0.157465   0.287  0.77503
## pgg45        0.004525   0.004421   1.024  0.30886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7084 on 88 degrees of freedom
## Multiple R-squared:  0.6548,Adjusted R-squared:  0.6234
## F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

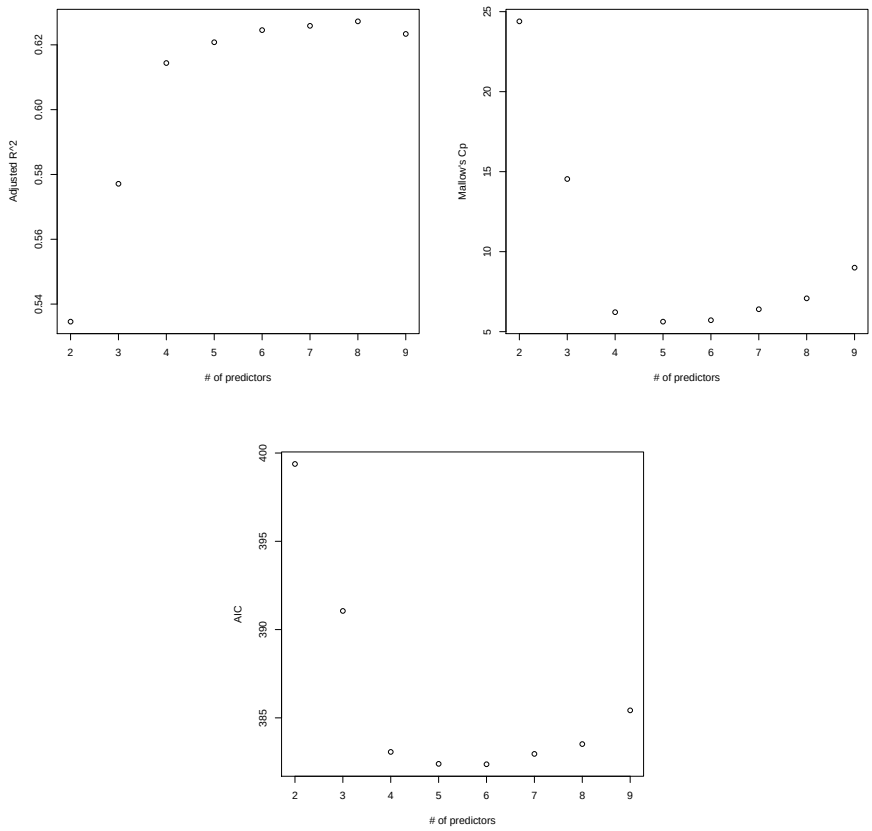
```
mod1 <- update(mod.full, . ~ . - gleason)
summary(mod1)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi + lcp +
##     pgg45, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.73117 -0.38137 -0.01728  0.43364  1.63513
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.953926   0.829439   1.150  0.25319
## lcavol       0.591615   0.086001   6.879 8.07e-10 ***
## lweight      0.448292   0.167771   2.672  0.00897 **
## age         -0.019336   0.011066  -1.747  0.08402 .
## lbph         0.107671   0.058108   1.853  0.06720 .
## svi          0.757734   0.241282   3.140  0.00229 **
## lcp         -0.104482   0.090478  -1.155  0.25127
## pgg45        0.005318   0.003433   1.549  0.12488
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7048 on 89 degrees of freedom
## Multiple R-squared:  0.6544,Adjusted R-squared:  0.6273
## F-statistic: 24.08 on 7 and 89 DF,  p-value: < 2.2e-16
```

```
mod2 <- update(mod1, . ~ . - lcp)
summary(mod2)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi + pgg45,
##     data = prostate)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.77711 -0.41708  0.00002  0.40676  1.59681
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.980085   0.830665   1.180  0.24116
## lcavol       0.545770   0.076431   7.141 2.31e-10 ***
## lweight      0.449450   0.168078   2.674  0.00890 **
## age         -0.017470   0.010967  -1.593  0.11469
## lbph         0.105755   0.058191   1.817  0.07249 .
## svi          0.641666   0.219757   2.920  0.00442 **
## pgg45        0.003528   0.003068   1.150  0.25331
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7061 on 90 degrees of freedom
## Multiple R-squared:  0.6493,Adjusted R-squared:  0.6259
## F-statistic: 27.77 on 6 and 90 DF,  p-value: < 2.2e-16

## Subsequent steps not included
```




Both AIC and Mallow's $C_p$ suggests that the "best" model is the model with five parameters including the intercept and the predictor variables `lcavol`, `lweight`, `lbph` and `svi`. The adjusted $R^2$ criterion is monotone increasing up to 8 parameters (including the intercept); however, we note that the adjusted $R^2$ for the model with 4 parameters and the adjusted $R^2$ for the model with 8 parameters are not very different. In particular, the adjusted $R^2$ for the model with 6 parameters through the model with 8 parameters are almost identical. Therefore, using the adjusted $R^2$ criterion, we can reasonably claim that the "best" model has 6 parameters including the intercept and the predictor variables `lcavol`, `lweight`, `lbph`, `svi` and `age`. We finally do some iterations of backward elimination. If we set the notional significance level of the test to be 0.1, then backward elimination give us the model with 4 parameters including the intercept `lcavol`, `lweight` and `svi`.

For part (b), we fit a ridge regression model as follows.

```
set.seed(1234)  ## Make the result reproducible
train.idx <- sample(1:nrow(prostate), floor(nrow(prostate)/2), replace = FALSE)
```

```r
test.idx <- setdiff(1:nrow(prostate), train.idx)  ## The indices for the test set
y.train <- prostate$lpsa[train.idx]
y.train.centered <- y.train - mean(y.train)
X <- within(prostate, rm(lpsa))
X.train <- as.matrix(X[train.idx, ])
X.train.centered <- X.train - outer(rep(1, length(train.idx)), colMeans(X.train))
g.ridge <- lm.ridge(y.train.centered ~ X.train.centered, lambda = c(seq(from = 0, to = 500,
    by = 0.1)))
matplot(g.ridge$lambda, t(g.ridge$coef), type = "l", lty = 1, xlab = expression(eta), ylab =

select(g.ridge)
```

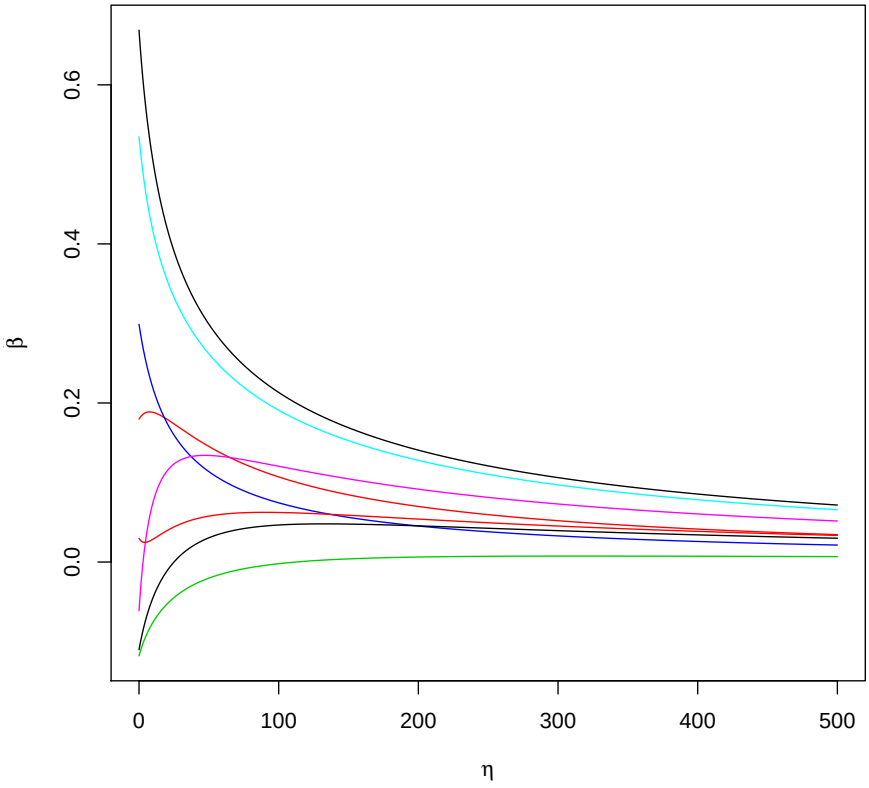
```
## modified HKB estimator is 4.374819
## modified L-W estimator is 3.384372
## smallest value of GCV  at 8.9
```

```r
## We choose \eta via the generalized cross validation criterion
beta.ridge <- coef(g.ridge)[66, -c(1)]  ## Ignore the estimated intercept term
X.test <- as.matrix(X[test.idx, ])
y.test <- prostate$lpsa[test.idx]
yhat.test <- (X.test - outer(rep(1, length(test.idx)), colMeans(X.train))) %*% beta.ridge +
    mean(y.train)
(SSE <- sum((y.test - yhat.test)^2))
```

```
## [1] 23.64871
```

```r
(MSE <- SSE/length(y.test))
```

```
## [1] 0.4826267
```




We choose $\eta$ via generalized cross validation. The value of $\eta$ chosen is 6.5. For this choice of $\eta$, the estimated mean square error is 0.483. We note that this estimated mean square error is quite close to the estimated mean square error for ordinary least square on the full data, which is approximately 0.5. The above code suggests that we have to be a little careful in that after centering the columns of the training data, we have to use the same centering for estimating the fitted values for the test data.

## Problem 2: (20pts)

This problem uses the dataset on the level of the enzyme creatinine kinase in the blood stream and the risk of heart attack. The data is given in the book "Handbook of small datasets" by Hand et al. (1994) and is also available from the link `http://www.stat.ncsu.edu/research/sas/sicl/data/heart.dat`. You might have to manually enter the data by hand into **R**, but since it

is a small dataset ...

The data was recorded in a study in which the level of creatinine kinase (CK) was measured for 360 different patients suspected of suffering from a heart attack. Whether or not each patient had really suffered a heart attack was established much later, after more prolonged medical investigation.

The format of the data is as follows. The first column contains intervals of CK, the second column is the number of patients with CK level in that interval who did suffered a heart attack and the third column is the number of patients with CK level in that interval who had not suffered a heart attack.

We first convert each interval into a number corresponding to the midpoint of that interval; we can replace the interval 40 or below by the number 20 and the interval 480 or above by the number 500.

1. Using the above data, first fit a logistic regression model with the `logit` link, using `CK` as the predictor variable. Look at the diagnostic plots of the deviance. Is the model a good fit to the data ? Why or why not ?
2. Plot the predicted and observed probability of heart attack against `CK` level. Discuss how the coefficient for `CK` relate to the odds or probability of suffering a heart attack ?
3. You may try adding powers of `CK`, e.g., $\texttt{CK}^2$ or $\texttt{CK}^3$ as additional predictor variable. Is the new model "better" than the model in part (1) ?
4. Now fit the model in part (1), but this time using the `probit` link function. Compare the estimated probability of suffering a heart attack for the `logit` link function and the `probit` link function when $\texttt{CK} = 380$ and when $\texttt{CK} = 460$.

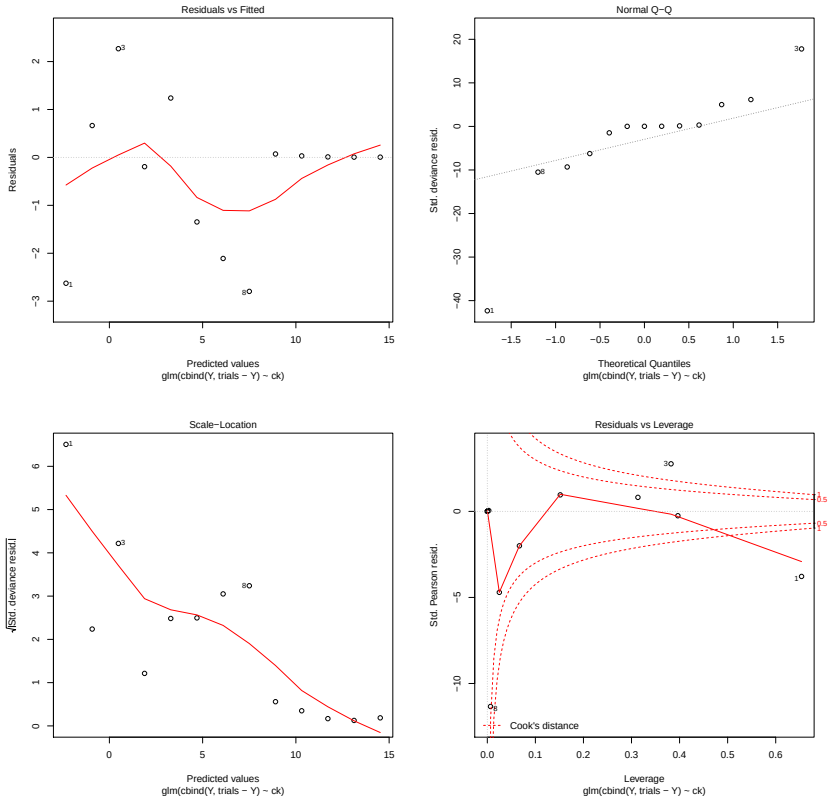
## Solution:

```
ck <- seq(from = 20, to = 500, by = 40)
Y <- c(2, 13, 30, 30, 21, 19, 18, 13, 19, 15, 7, 8, 35)
trials <- c(90, 39, 38, 35, 21, 20, 19, 14, 19, 15, 7, 8, 35)
mod1 <- glm(cbind(Y, trials - Y) ~ ck, family = binomial())
summary(mod1)
```

```
##
## Call:
## glm(formula = cbind(Y, trials - Y) ~ ck, family = binomial())
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.79579  -1.34637   0.00587   0.07173   2.26860
##
## Coefficients:
```

```
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.028360   0.366977  -8.252   <2e-16 ***
## ck           0.035104   0.004081   8.602   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 311.29  on 12  degrees of freedom
## Residual deviance:  28.14  on 11  degrees of freedom
## AIC: 51.596
##
## Number of Fisher Scoring iterations: 6
```

```
plot(mod1)
```

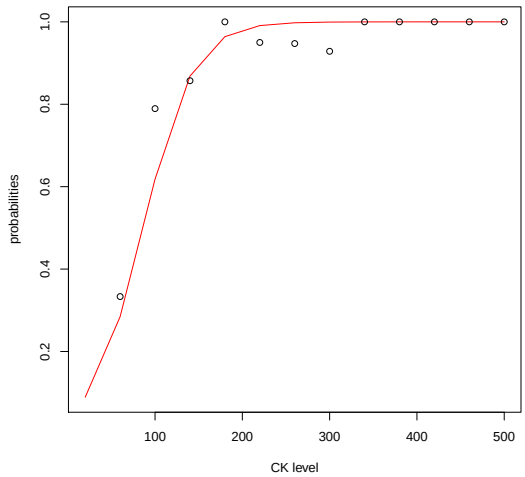



For part (a), we note that the residual deviance is quite larger than the degrees of freedom. We also note that the deviance residuals (this is different from the residual deviance) exhibits a non-constant pattern with respect to fitted value

and that there are several points with large Cook distance. Since there are only a few data points (13 in this case), and several of these data points have empirical failure probability of 0 (for example, at CK level 500, all 35 patients suffered a heart attack), we surmise that it is hard for our model to fit all the data points well and so there will be a lack of fit in our model.

For part (b), we plot the estimated probabilities and empirical probabilities

```
plot(ck, fitted(mod1), xlab = "CK level", type = "l", col = "red", ylab = "probabilities")
points(ck, Y/trials)
```




The estimated coefficient for CK is 0.035. This coefficient is positive, which indicates that higher level of CK is associated with a higher odd of having a heart attack. For our original data, the observations are stratified into different categories based on the CK level. We thus conclude that going from one category to the next category, the odd of having a heart attack increases by a factor of $e^{40\times 0.035} = 4.072$.

For part (c), we fit a model with higher order terms for CK

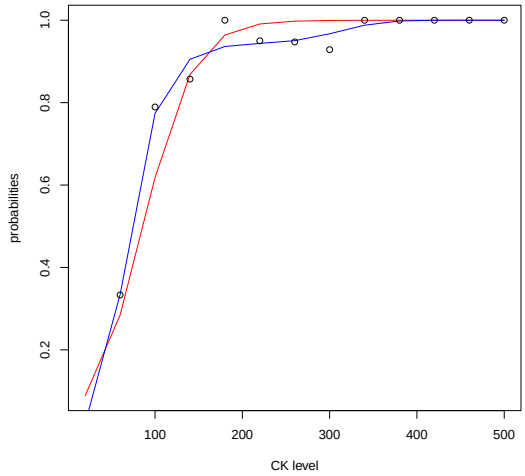
```
mod2 <- update(mod1, ~. + I(ck^2) + I(ck^3))
summary(mod2)

##
## Call:
## glm(formula = cbind(Y, trials - Y) ~ ck + I(ck^2) + I(ck^3),
##     family = binomial())
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
```

```
## -0.91095  -0.01629   0.01678   0.23005   1.66260
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.890e+00  9.977e-01  -5.903 3.57e-09 ***
## ck           1.146e-01  2.453e-02   4.671 2.99e-06 ***
## I(ck^2)     -5.113e-04  1.721e-04  -2.971  0.00297 **
## I(ck^3)      7.744e-07  3.468e-07   2.233  0.02554 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 311.2869  on 12  degrees of freedom
## Residual deviance:   4.6775  on  9  degrees of freedom
## AIC: 32.134
##
## Number of Fisher Scoring iterations: 8
```

```
anova(mod1, mod2, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(Y, trials - Y) ~ ck
## Model 2: cbind(Y, trials - Y) ~ ck + I(ck^2) + I(ck^3)
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        11    28.1402
## 2         9     4.6775  2   23.463 8.038e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(ck, fitted(mod1), xlab = "CK level", type = "l", col = "red", ylab = "probabilities")
lines(ck, fitted(mod2), xlab = "CK level", col = "blue", ylab = "probabilities")
points(ck, Y/trials)
```




We note that the model with a cubic term appears to fit the data much better than our original model. The residual deviance is now smaller than the degrees of freedom, the fitted probabilities and the empirical probabilities are more similar.

For part (d), the probit model gives

```
mod3 <- glm(cbind(Y, trials - Y) ~ ck, family = binomial("probit"))
fitted(mod3)[c(10, 12)]

##        10        12
## 0.9999996 1.0000000

fitted(mod1)[c(10, 12)]

##        10        12
## 0.9999667 0.9999980
```

The estimated probabilities of a heart attack when the `CK` level is 380 and 460 for both the probit and the logit link function is almost identical. This is once again due to the fact that the observations with the largest `CK` level all have empirical failure rate of 0.

## Problem 3: (20pts)

This problem uses the `pima` dataset from the `faraway` package. The data was collected in a study on 768 adult female Pima Indians by the National Institute of Diabetes and Digestive and Kindey Diseases. The goal of the study was to investigate factors related to diabetes.

1. Take a look at the predictor variables. In particular, are 0 values for `diastolic`, `bmi`, `glucose` and so on reasonable ?
2. Remove observations with missing values in the predictor variables; the function `union` in **R** might be useful here. Then fit a logistic regression model with `test`, the result of the diabetes test, as the response variable and all the other variables as predictors. Can you use the deviance as measure of whether the model fits the data ?
3. Do women who test positive for diabetes have higher diastolic blood pressure ? Is the diastolic blood pressure significant in the regression model ? Explain the distinction between the two questions above and discuss why their answers are only apparently contradictory.
4. Compare the above model with the model using only the predictor variables `diastolic`, `bmi` and `age`.
5. What does the coefficients of `diastolic`, `bmi` and `age` meant ? How do they relate to the variable `test` ?
6. Suppose you want to predict the result of the diabetes test using the model with predictor variables `diastolic`, `bmi` and `age` by thresholding the estimated probability. What threshold should you use ?

## Solution:

```
library("faraway")
data(pima)
summary(pima)
```

```
##     pregnant         glucose        diastolic         triceps         insulin
##  Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00   Min.   :  0.0
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00   1st Qu.:  0.0
##  Median : 3.000   Median :117.0   Median : 72.00   Median :23.00   Median : 30.5
##  Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54   Mean   : 79.8
##  3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00   3rd Qu.:127.2
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00   Max.   :846.0
##       bmi           diabetes           age             test
##  Min.   : 0.00   Min.   :0.0780   Min.   :21.00   Min.   :0.000
##  1st Qu.:27.30   1st Qu.:0.2437   1st Qu.:24.00   1st Qu.:0.000
##  Median :32.00   Median :0.3725   Median :29.00   Median :0.000
##  Mean   :31.99   Mean   :0.4719   Mean   :33.24   Mean   :0.349
##  3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00   3rd Qu.:1.000
##  Max.   :67.10   Max.   :2.4200   Max.   :81.00   Max.   :1.000
```

We note that the data, more specifically the variables `glucose`, `bmi`, `triceps`, and `diastolic`, has zero values. Common sense indicates that there must be some sort of confounding between "0" and "N/A" for the missing values. We

thus chose to remove observations with these "0" value from the data (there are other ways to handling missing values, such as imputation, but we will not discuss them herein).

```
good.data <- (pima$glucose != 0) & (pima$diastolic != 0) & (pima$triceps != 0) & (pima$bmi !
    0)
pima.clean <- pima[good.data, ]
## We removed almost 30% of the data
nrow(pima.clean)/nrow(pima)

## [1] 0.6927083

g1.clean <- glm(test ~ ., data = pima.clean, family = "binomial")
summary(g1.clean)

##
## Call:
## glm(formula = test ~ ., family = "binomial", data = pima.clean)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8627  -0.6639  -0.3672   0.6347   2.4942
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.677562   1.005400  -9.626  < 2e-16 ***
## pregnant     0.121235   0.043926   2.760 0.005780 **
## glucose      0.037439   0.004765   7.857 3.92e-15 ***
## diastolic   -0.009316   0.010446  -0.892 0.372494
## triceps      0.006341   0.014853   0.427 0.669426
## insulin     -0.001053   0.001007  -1.046 0.295651
## bmi          0.085992   0.023661   3.634 0.000279 ***
## diabetes     1.335764   0.365771   3.652 0.000260 ***
## age          0.026430   0.013962   1.893 0.058371 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 676.79  on 531  degrees of freedom
## Residual deviance: 465.23  on 523  degrees of freedom
## AIC: 483.23
##
## Number of Fisher Scoring iterations: 5

cor.test(pima.clean$test, pima.clean$diastolic)

##
```

```
##  Pearson's product-moment correlation
##
## data:  pima.clean$test and pima.clean$diastolic
## t = 4.2958, df = 530, p-value = 2.071e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.09998078 0.26432033
## sample estimates:
##       cor
## 0.1834319
```

```
g2.clean <- glm(test ~ diastolic + bmi + age, data = pima.clean, family = "binomial")
summary(g2.clean)
```

```
##
## Call:
## glm(formula = test ~ diastolic + bmi + age, family = "binomial",
##     data = pima.clean)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1802  -0.8305  -0.5281   0.9942   2.2374
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.208322   0.778002  -7.980 1.47e-15 ***
## diastolic   -0.001253   0.009190  -0.136    0.892
## bmi          0.104025   0.016739   6.215 5.14e-10 ***
## age          0.064622   0.010152   6.365 1.95e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 676.79  on 531  degrees of freedom
## Residual deviance: 577.18  on 528  degrees of freedom
## AIC: 585.18
##
## Number of Fisher Scoring iterations: 4
```

```
anova(g2.clean, g1.clean, test = "Chi")
```

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

```
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       528     577.18
## 2       523     465.23  5   111.95 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

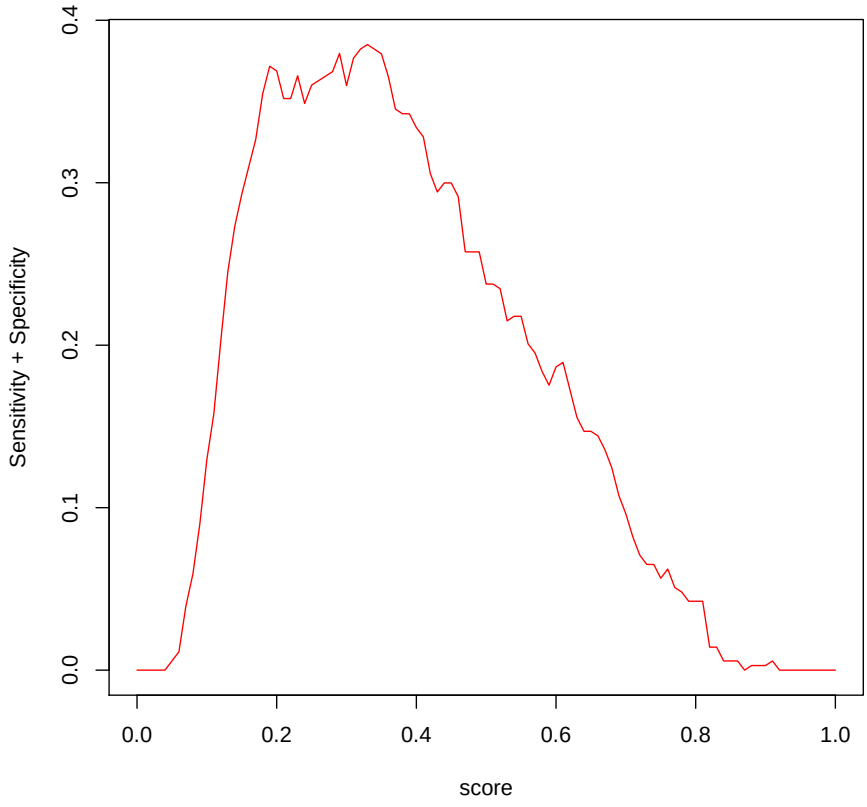
For part (c), we fit a binomial GLM model to the data. The `deviance` is not a meaningful measure of goodness-of-fit for this data, as the response variable is binary, and hence the chi-square approximation for the deviance under the null is no longer valid. We note that the coefficient for `diastolic` is *not* significant in the presence of other predictor variables. Nevertheless, the correlation between `diastolic` blood pressure and `test` positive for diabetes is positive, and a correlation test using the Pearson product-moment correlation indicates that the correlation is highly significant. This apparent "contradiction" is due mostly to the fact that there are other relevant predictor variables in the model.

For part (d), we can compare model `g2`, where only the predictor variables `diastolic`, `bmi` and `age` are used, with the model `g1` by comparing their deviance measure. In this case, the difference between the two deviance is approximated, under the null hypothesis that the two models are the "same", by the chi-square distribution. We note that there is scant evidence that the two models are comparable.

For part (e), the coefficients can be interpreted as follows. The coefficient for `bmi` in model `g2.clean` is 0.104. Now $e^{0.104} = 1.110$ and hence a change in the body mass index by 1 unit (the body mass index usually ranges from a minimum of 20 to a maximum of say 40) while keeping the remaining variables fixed is associated with roughly a 11% increase in the odd of testing positive for diabetes. Similarly, an increase in age by 1 year (while keeping the other variables fixed) is associated with roughly a 6% increase in the odd of testing positive for diabetes.

```
par(mfrow = c(2, 2))

pima.clean$fitted2 <- g2.clean$fitted.values

score.vec <- seq(0, 1, by = 0.01)
tpr.vec <- numeric(101)
fpr.vec <- numeric(101)


for (i in 1:101) {
    s <- score.vec[i]
    true.positive <- sum((pima.clean$fitted2 > s) * pima.clean$test)
    false.positive <- sum((pima.clean$fitted2 > s) * (1 - pima.clean$test))
    true.negative <- sum((pima.clean$fitted2 <= s) * (1 - pima.clean$test))
    false.negative <- sum((pima.clean$fitted2 <= s) * pima.clean$test)
```

```
    tpr.vec[i] <- true.positive/(true.positive + false.negative)
    fpr.vec[i] <- false.positive/(true.negative + false.positive)
}

par(mfrow = c(1, 1))
plot(score.vec, tpr.vec - fpr.vec, type = "l", col = "red", xlab = "score", ylab = "Sensitiv
```




For part (f), we can choose the threshold to maximize the sum of sensitivity and specificity. Equivalently, we choose the threshold to maximize the difference between the true positive rate and the false positive rate. A quick look at the difference between `tpr` and `fpr` against the threshold suggests that we can choose the threshold to be roughly 0.35. We could also choose the threshold to maximize a weighted sum of sensitivity and specificity, depending on how important a false positive is compared to say a false negative.

## Problem 4: (20pts)

The CrabSatellite dataset, available from `https://www.unc.edu/courses/2010fall/ecol/563/001/data/hw/crabs.txt`, records the mating behavior of male and female horseshoe crabs in the Gulf of Mexico. A nesting female horseshoe crab typically has a male crab resident in her nest. In addition, she may have other males, called satellites, residing nearby. These satellites are unattached males and come to the beach to crowd around the nesting couples and compete with the attached males for fertilization. Some females are ignored by satellite males while some attract more satellites than others.

The variables in this data set are as follows

- `color` denote the female crab's color and is ordered from 1 to 5 where 1 is "light" and 5 is "dark" colored. Not all of these categories are represented in this data set.
- `spine` denote the female's spine condition and is ordered from 1 for "good" to 3 for "broken"
- `width` denote female's width (in cm).
- `weight` denote the females's weight (in g)
- `num.satellites` denote the number of nearby satellites.

Using this dataset, fit a model and find an answer to the question "how do unattached males choose among available females" ? Do you think your model is a good fit to the data ? Why or why not ?

## Solution:

```
dat <- read.table("crabs.txt", header = FALSE)
colnames(dat)
```

```
## [1] "V1" "V2" "V3" "V4" "V5"
```

```
colnames(dat) <- c("color", "spine", "width", "satellites", "weight")
```

We first read the data and format the column names. For simplicity, we will represent `color` and `spine` as integers as opposed to factors.

```
g1 <- glm(satellites ~ color + spine + width + weight, data = dat, family = poisson())
summary(g1)
```

```
##
## Call:
## glm(formula = satellites ~ color + spine + width + weight, family = poisson(),
##     data = dat)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0126  -1.8846  -0.5406   0.9448   4.9602
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.3435447  0.9684204  -0.355  0.72278
## color       -0.1849325  0.0665236  -2.780  0.00544 **
## spine        0.0399764  0.0568062   0.704  0.48160
## width        0.0275251  0.0479425   0.574  0.56588
## weight       0.0004725  0.0001649   2.865  0.00417 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 632.79  on 172  degrees of freedom
## Residual deviance: 551.85  on 168  degrees of freedom
## AIC: 917.15
##
## Number of Fisher Scoring iterations: 6
```

```
g2 <- glm(satellites ~ color + weight, data = dat, family = poisson())
anova(g2, g1, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model 1: satellites ~ color + weight
## Model 2: satellites ~ color + spine + width + weight
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       170     552.79
## 2       168     551.85  2  0.94042   0.6249
```

```
summary(g2)
```

```
##
## Call:
## glm(formula = satellites ~ color + weight, family = poisson(),
##     data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9785  -1.9159  -0.5471   0.9181   4.8338
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.614e-01  3.008e-01   0.869  0.38496
```

```
## color       -1.728e-01  6.155e-02  -2.808  0.00499 **
## weight       5.459e-04  6.749e-05   8.088 6.05e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 632.79  on 172  degrees of freedom
## Residual deviance: 552.79  on 170  degrees of freedom
## AIC: 914.09
##
## Number of Fisher Scoring iterations: 6
```

As the number of satellites is a non-negative integer, we try fitting a Poisson linear model with `color`, `spine`, `width` and `weight` as the predictor variables. We see that the coefficients of `spine` and `width` are not statistically significant. It could be the case that `width` is highly correlated with `weight`, and as such, only one of these variables will be statistically significant in the model. The coefficient for `color` is negative, which indicates that female crabs with darker shell color has less number of satellites. The coefficient for `weight` is positive, and since weight is measured in gram, this indicates that heavier female crabs have more satellites. Our answer to the question "how do unattached males choose among available females ?" is therefore "unattached males prefer heavier female with light-colored shell".

The deviance measure is high compared to the degrees of freedom, which indicates a lack of fit for both of our model `g1` and `g2` above.

```
library("pscl")
table(dat$satellites)
```

```
##
##  0  1  2  3  4  5  6  7  8  9 10 11 12 14 15
## 62 16  9 19 19 15 13  4  6  3  3  1  1  1  1
```

```
colSums(predprob(g1))
```

```
##           0           1           2           3           4           5           6
## 15.02507022 31.83257885 37.19534655 32.19628968 23.25669967 14.87228797  8.68229557
##           7           8           9          10          11          12          13
##  4.70953426  2.40979304  1.19010437  0.59222462  0.31817727  0.19687456  0.14025004
##          14          15
##  0.10790294  0.08390591
```

```
colSums(predprob(g2))
```

```
##           0           1           2           3           4           5           6
## 14.78775316 31.73045972 37.42224762 32.51550213 23.42745142 14.85884224  8.57016281
##           7           8           9          10          11          12          13
```

```
##  4.58503432  2.31433352  1.12924933  0.55736534  0.30020997  0.19018949  0.14183900
##          14          15
##  0.11563517  0.09562093
```

We next take a look at the empirical distribution of the number of satellites and compare it against the distribution as specified by our model `g1` and `g2`. We see that the number of female crabs with 0 satellites is 62. Meanwhile, the predicted number of female crabs with 0 satellites is approximately 15 for both model `g1` and `g2`. This indicates that the response variable in the data has many more 0 observations compared to our model. Thus, it might be more appropriate to model the response variable as a zero inflated Poisson model.