I3-TD3
Multiple Linear Regression

Problem 1

(Data file: `water`) For this problem, consider the regression problem with response `BSAAM`, and three predictors as regressors given by `OPBPC`, `OPRC`, and `OPSLAKE`.

```
library(alr4)
```

```
## Loading required package: car

## Loading required package: carData

## Loading required package: effects

## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
data(water)

#help("water")
head(water)
```

```
##   Year APMAM APSAB APSLAKE OPBPC  OPRC OPSLAKE  BSAAM
## 1 1948  9.13  3.58    3.91  4.10  7.43    6.47  54235
## 2 1949  5.28  4.82    5.20  7.55 11.11   10.26  67567
## 3 1950  4.20  3.77    3.67  9.52 12.20   11.35  66161
## 4 1951  4.60  4.46    3.93 11.14 15.15   11.13  68094
## 5 1952  7.15  4.99    4.88 16.34 20.05   22.81 107080
## 6 1953  9.70  5.65    4.91  8.88  8.15    7.41  67594
```

1. Examine the scatterplot matrix drawn for these three regressors and the response. What should the correlation matrix look like (i.e., which correlations are large and positive, which are large and negative, and which are small?) Compute the correlation matrix to verify your results.

2. Get the regression summary for the regression of `BSAAM` on these three regressors. Explain what the "t-values" columns of your output means.

Problem 2

Berkeley Guidance Study (Data file: `BGSgirls`) Data from the Berkeley Guidance Study on the growth of boys and girls. We will view body mass index at age 18 `BIM18`, as the response, and weights in kilogram at ages 2, 9, and 18, `WT2`, `WT9`, and `WT18` as predictor.

```
library(alr4)
data("BGSgirls")

#help("BGSgirls")
head(BGSgirls)
```

```
##     WT2  HT2  WT9   HT9  LG9 ST9 WT18  HT18 LG18 ST18 BMI18 Soma
## 67 13.6 87.7 32.5 133.4 28.4  74 56.9 158.9 34.6  143  22.5  5.0
## 68 11.3 90.0 27.8 134.8 26.9  65 49.9 166.0 33.8  117  18.1  4.0
## 69 17.0 89.6 44.4 141.5 31.9 104 55.3 162.2 35.1  143  21.0  5.5
## 70 13.2 90.3 40.5 137.1 31.8  79 65.9 167.8 39.3  148  23.4  5.5
## 71 13.3 89.4 29.9 136.1 27.7  83 62.3 170.9 36.3  152  21.3  4.5
## 72 11.3 85.5 22.8 130.6 23.4  60 47.4 164.9 31.8  126  17.4  3.0
```

1. Obtain the scatterplot matrix for these four variables. Define which predictor variable has the strongest relationship with `BMI18` and what can you say about it. Is transformation neccessary in this case?
2. Comment on the correlation among the predictor variables.
3. Obtain the summary table for the multiple linear regression with the three predictors. Interpret the β_j coefficients obtained from the model. Do the results make sense?
4. The unexpected sign of coefficients may be due to the correlation between the regressors. This is the problem of multicollinearity. In this case, since all the three original regressors measure weight, combining them together is reasonable. Consider a set of linear transformations of the weight variables below:

$$ave = (WT2 + WT9 + WT18)/3$$
$$lin = WT18 - WT2$$

$$quad = WT2 - 2 \times WT9 + WT18$$

Since the three weight variables are approximately equally spaced in time, these three variables correspond to the average weight, a linear component in time, and a quadratic component in time; see Oehlert (2000) or Kennedy and Gentle (1980), for example, for a discussion of orthogonal polynomials.

Fit with these regressors using the girls in the Berkeley Guidance Study data and compare with the results in Problem 4.3.

Problem 3

(Data file: `Transact`) The data in this example consists of a sample of branches of a large Australian bank (Cunningham and Heathcote, 1989). Each branch makes transactions of two types, and for each of the branches we have recorded the number `T1` of type 1 transactions and the number `t2` of type 2 transactions. The response is `time`, the total minutes of labor used by the branch.

```r
library(alr4)
data(Transact)

#head(Transact)

Transact$a = (Transact$t1 + Transact$t2)/2
Transact$d = Transact$t1 - Transact$t2
head(Transact)
```

```
##    t1   t2  time      a     d
## 1   0 1166  2396  583.0 -1166
## 2   0 1656  2348  828.0 -1656
## 3   0  899  2403  449.5  -899
## 4 516 3315 13518 1915.5 -2799
## 5 623 3969 13437 2296.0 -3346
## 6 395 3087  7914 1741.0 -2692
```

Define `a = (t1 + t2)/2` to be the average transaction time, and `d = t1 - t1`, and fit the following four mean functions.

i. M1: $E(time|t1, t2) = \beta_{01} + \beta_{11}t1 + \beta_{21}t2$
ii. M2: $E(time|t1, t2) = \beta_{02} + \beta_{32}a + \beta_{42}d$
iii. M3: $E(time|t1, t2) = \beta_{03} + \beta_{23}t2 + \beta_{43}d$
iv. M4: $E(time|t1, t2) = \beta_{04} + \beta_{14}t1 + \beta_{24}t2 + \beta_{34}a + \beta_{44}d$

1. In the fit of M4, some of the coefficients estimates are labeled as "aliased (NA)" or else they are simply omitted. Explain what this means and why this happens.
2. What aspects of the fitted regressions are the same? What aspects are different?
3. Why is the estimate for $t2$ different in M1 and M3?

Problem 4

Cakes (Data file: `cakes`) Oehlert (2000) provides data from a small experiment with $n = 14$ observations on baking packaged cake mixes. Two factors, X_1 = backing time minutes and X_2 = baking temperature in degrees F, were varied in the experiment. The response Y was the average palatability score of four cakes bakes at baked at a given combination of (X_1, X_2), with higher values desirable.

```r
library(alr4)
data(cakes)

head(cakes)
```

```
##   block X1  X2    Y
## 1     0 33 340 3.89
## 2     0 37 340 6.36
## 3     0 33 360 7.65
## 4     0 37 360 6.79
## 5     0 35 350 8.36
## 6     0 35 350 7.63
```

Suppose we have a model:

$$E(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2.$$

1. Fit the model and verify that the significance levels for the quadratic terms and interaction are all less than 0.005. When fitting the polynomials, tests concerning main effects in models that include a quadratic are generally not of much interest.

2. The cake experiment was carried out in two blocks of seven observations each. It is possible that the response might diff by block. For example, if the blocks were different days, then differences in air temperature or humidity when the cakes were mixed might have some effect on `Y`. We can allow for block effects by adding a factor block to the mean function and possibly allowing for block by regressor interactions. And block effects to the mean function fit in a new model and summarize results. The blocking is indicated by the variable `block` in the data file.

Problem 5

(Data file: `BGsall`) Refer to the Berkeley Guidance study described in Problem 2. Using the data file `BGSall`, consider the regression of `HT18` on `HT9` and the grouping factor `Sex`.

```
library(alr4)
data(BGSall)

head(BGSall)
```

```
##   Sex  WT2  HT2  WT9   HT9  LG9 ST9  WT18  HT18 LG18 ST18 BMI18 Soma
## 1   0 13.6 90.2 41.5 139.4 31.6  74 110.2 179.0 44.1  226  34.4  7.0
## 2   0 12.7 91.4 31.0 144.3 26.0  73  79.4 195.1 36.1  252  20.9  4.0
## 3   0 12.6 86.4 30.1 136.5 26.6  64  76.3 183.7 36.9  216  22.6  6.0
## 4   0 14.8 87.6 34.1 135.4 28.2  75  74.5 178.7 37.3  220  23.3  2.0
## 5   0 12.7 86.7 24.5 128.9 24.2  63  55.7 171.5 31.0  200  18.9  1.5
## 6   0 11.9 88.1 29.8 136.0 26.7  77  68.2 181.8 37.0  215  20.6  3.0
```

1. Draw the scatterplot of `HT18` versus `HT9`, using a different symbol for males and females. Comment on the information in the graph about an appropriate mean function for these data.
2. Obtain the appropriate test for a parallel regression model.
3. Assuming the parallel regression model is adequate, estimate a 95% confidence interval for the difference between males and females. For the parallel regression model, this is the difference in the intercepts of the two groups.

Problem 6

Sex discrimination (Data file: `salary`) The data file concerns salary and other characteristics of all faculty in a small Midwestern college collected in the early 1980s for presentation in legal proceedings for which discrimination against women in salary was at issue. All persons in the data hold tenured or tenure track positions; temporary faculty are not included. The variables include `degree`, a factor with levels Male and Female; `Year`, years in current rank; `ysdeg`, years since highest degree, and `salary`, academic year salary in dollars.

```
library(alr4)
data("salary")

head(salary)
```

```
##    degree rank    sex year ysdeg salary
## 1 Masters Prof   Male   25    35  36350
```

```
## 2 Masters Prof   Male   13    22  35350
## 3 Masters Prof   Male   10    23  28200
## 4 Masters Prof Female    7    27  26775
## 5     PhD Prof   Male   19    30  33696
## 6 Masters Prof   Male   16    21  28516
```

1. Get appropriate graphical summaries of the data and discuss the graphs.
2. Test the hypothesis that the mean salary for men and women is the same. What alternative hypothesis do you think is appropriate?
3. Assuming no interactions between `sex` and the other predictors, obtain a 95% confidence interval for the difference in salary between males and females.
4. Finkelstein (1980), in a discussion of the use of regression in discrimination cases, wrote, "[a] variable may reflect a position or status bestowed by the employer, in which cases if there is discrimination in the award of the position or status, the variable may be 'tainted.'" Thus, for example, if discrimination is at work in promotion of faculty to higher ranks, using rank to adjust salaries before comparing the sexes may be not acceptable to the courts. Exclude the variable `rank`, refit, and summarize. '

Problem 7

This question involves the use of multiple linear regression on the `Auto` data set.

```r
library(ISLR)
data(Auto)
head(Auto)
```

```
##   mpg cylinders displacement horsepower weight acceleration year origin
## 1  18         8          307        130   3504         12.0   70      1
## 2  15         8          350        165   3693         11.5   70      1
## 3  18         8          318        150   3436         11.0   70      1
## 4  16         8          304        150   3433         12.0   70      1
## 5  17         8          302        140   3449         10.5   70      1
## 6  15         8          429        198   4341         10.0   70      1
##                        name
## 1 chevrolet chevelle malibu
## 2         buick skylark 320
## 3        plymouth satellite
## 4             amc rebel sst
## 5               ford torino
## 6          ford galaxie 500
```

1. Produce a scatterplot matrix which includes all of the variables in the data set.
2. Compute the matrix of correlations between the variables using the function `cor()`. You will need to exclude the `name` variable, which is qualitative.
3. Use the `lm()` function to perform a multiple linear regression with `mpg` as the response and all other variables except `name` as the predictors. Use the `summary()` function to print the results. Comment on the output. For instance:

i. Is there a relationship between the predictors and the response?

ii. Which predictors appear to have a statistically significant relationship to the response?
iii. What does the coefficient for the `year` variable suggest?

4. Use the plot() function to produce diagnostic plots of the linear regression ft. Comment on any problems you see with the ft. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

5. Use the * and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

6. Try a few different transformations of the variables, such as $\log(X), \sqrt{X}, X^2$. Comment on your findings.

Problem 8

In this exercise, we will generate simulated data, and will then use this data to perform best subset selection.

1. Use the `norm()`function to generate a predictor X of length $n = 100$, as well as a noise vector ϵ of length $n = 100$.

2. Generate a response vector Y of length $n = 100$ according to the model

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon,$$

   where $\beta_0, \beta_1, \beta_2$ and β_3 are constants of your choice.

3. Use the `regsubsets()` function to perform best subset selection in order to choose the best model containing the predictors $X, X^2, ..., X^n$. What is the best model obtained according to C_p, BIC, and adjust R^2? Show some plots to provided evidence for your answer, and report the coefficients of the best model obtained. Note you will need to use the `data.frame()` function to create a single data set containing both X and Y.

4. Repeat (3), using forward stepwise selection and also using backwards stepwise selection. How does your answer compare to the results in (3)?

5. Now fit a lasso model to the simulated data, again using $X, X^2, ..., X^10$ as predictors. Use cross-validation to select the optimal value of λ Create plots of the cross-validation error as a function of λ. Report the resulting coefficient estimates, and discuss the results obtained.

6. Now generate a response vector Y according to the model.

$$Y = \beta_0 + \beta_7 X^7 + \epsilon,$$

   and perform best subset selection and the lasso. Discuss the results obtained.

Problem 9

In this exercise, we will predict the number of applications received using the other variables in the `College` data set.

1. Split the data set into a training set and a test set.

2. Fit a linear model using least squares on the training set, and report the test error obtained.

3. Fit a ridge regression model on the training set, with λ chosen by cross-validation. Report the test error obtained.
4. Fit a lasso model on the training set, with λ chosen by cross validation. Report the test error obtained, along with the number of non-zero coefficient estimates.
5. Fit a PCR model on the training set, with M chosen by cross validation. Report the test error obtained, along with the value of M selected by cross-validation.
6. Fit a PLS model on the training set, with M chosen by cross validation. Report the test error obtained, along with the value of M selected by cross-validation.
7. Comment on the results obtained. How accurately can we predict the number of college applications received? Is there much difference among the test errors resulting from these five approaches?

Problem 10

We have seen that as the number of features used in a model increases, the training error will necessarily decrease, but the test error may not. We will now explore this in a simulated data set.

1. Generate a data set with $p = 20$ features, $n = 1,000$ observations, and an associated quantitative response vector generated according to the model

$$Y = X\beta + \epsilon$$

   where β has some elements that are exactly equal to zero.
2. Split your data set into a training set containing 100 observations and a test set containing 900 observations.
3. Perform best subset selection on the training set, and plot the training set MSE associated with the best model of each size.
4. Plot the test set MSE associated with the best model of each size.
5. For which model size does the test set MSE take on its minimum value? Comment on your results. If it takes on its minimum value for a model containing only an intercept or a model containing all of the features, then play around with the way that you are generating the data in (1) until you come up with a scenario in which the test set MSE is minimized for an intermediate model size.