CPT_S 575 Data Science: Assignment 4

Reet Barik

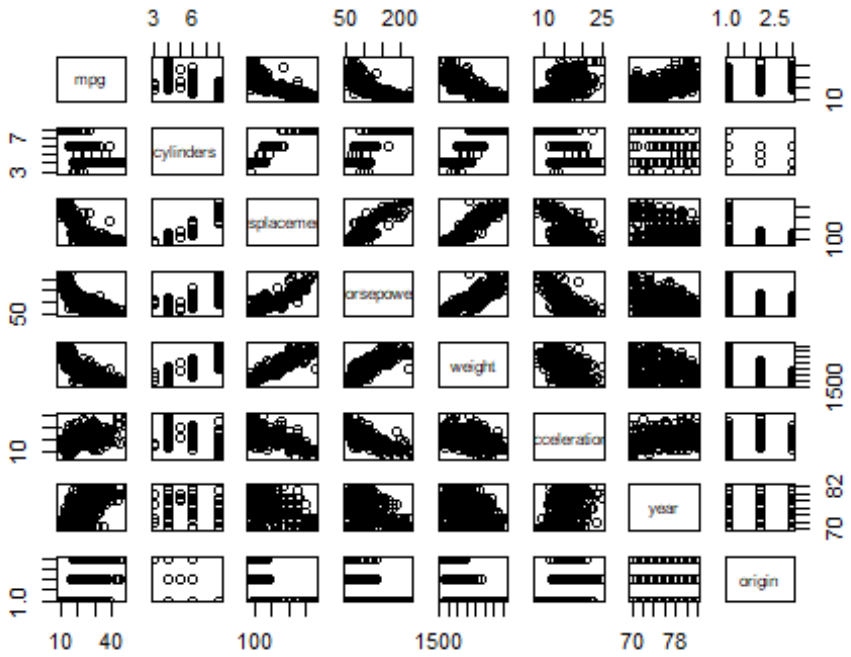
October 10, 2019

Question 1

```
library(latexpdf)
library(MASS)
library(ISLR)
auto_df <- read.csv("https://scads.eecs.wsu.edu/wp-
content/uploads/2017/09/Auto.csv", na.strings = "?")
auto_df <- na.omit(auto_df)
```

(a) Produce a scatterplot matrix which includes all the variables in the data set.

```
pairs(subset(auto_df, select=-c(name)))
```



(b) Compute the matrix of correlations between the variables. You will need to exclude the name variable, which is qualitative.

```r
cor(subset(auto_df, select=-c(name)))
```

```
##                     mpg  cylinders displacement horsepower     weight
## mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442
## cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273
## displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
## horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
## weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000
## acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392
## year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199
## origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054
##              acceleration       year     origin
## mpg             0.4233285  0.5805410  0.5652088
## cylinders      -0.5046834 -0.3456474 -0.5689316
## displacement   -0.5438005 -0.3698552 -0.6145351
## horsepower     -0.6891955 -0.4163615 -0.4551715
## weight         -0.4168392 -0.3091199 -0.5850054
## acceleration    1.0000000  0.2903161  0.2127458
## year            0.2903161  1.0000000  0.1815277
## origin          0.2127458  0.1815277  1.0000000
```

(c) Perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Show a printout of the result (including coefficient, error and t values for each predictor). Comment on the output: i. Which predictors appear to have a statistically significant relationship to the response, and how do you determine this? ii. What does the coefficient for the displacement variable suggest, in simple terms?

```r
auto.fit = lm(mpg~.-name, data=auto_df)
summary(auto.fit)
```

```
##
## Call:
## lm(formula = mpg ~ . - name, data = auto_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
## cylinders     -0.493376   0.323282  -1.526  0.12780
## displacement   0.019896   0.007515   2.647  0.00844 **
## horsepower    -0.016951   0.013787  -1.230  0.21963
## weight        -0.006474   0.000652  -9.929  < 2e-16 ***
## acceleration   0.080576   0.098845   0.815  0.41548
```

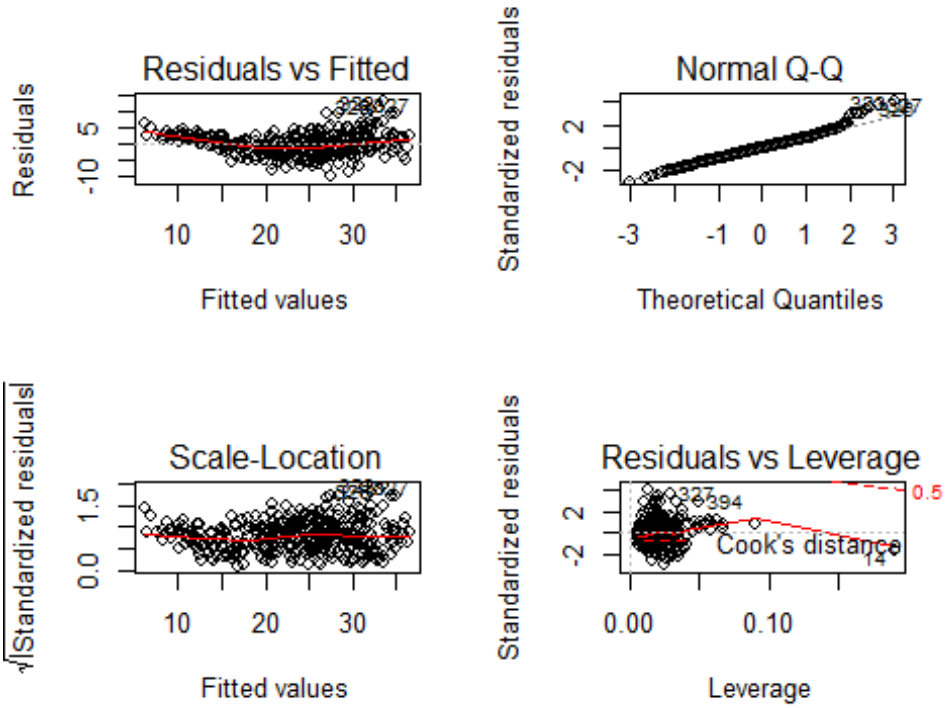
```
## year           0.750773   0.050973  14.729  < 2e-16 ***
## origin         1.426141   0.278136   5.127 4.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.328 on 384 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

(i) Displacement, weight, year and origin are four predictors which have a statistically significant relationship to the response. This is determined by their low p-values (<0.01).

(ii) Assuming that all the predictors are uncorrelated, the coefficient for the displacement variable, in simple terms suggests that for each unit increase in displacement there is 0.019896 unit change in the 'mpg' variable, while all the other variables stay fixed.

(d) Produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

```
par(mfrow=c(2,2))
plot(auto.fit)
```



In the Residuals vs Fitted plot there are a few outliers lying outside the [-3, 3] range on the residual axis (in particular, point 323, 326, and 327 can be considered as unusually high outliers). Point 14 shows high leverage for this model as can be seen in the Residuals vs Leverage plot.

(e) Fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

```
summary(lm(mpg ~ weight*displacement  + (weight*cylinders) +
             cylinders*displacement, data=auto_df))
```

```
##
## Call:
## lm(formula = mpg ~ weight * displacement + (weight * cylinders) +
##     cylinders * displacement, data = auto_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.1599  -2.5204  -0.3546   1.7851  17.8829
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             4.903e+01  6.743e+00   7.271 2.01e-12 ***
## weight                 -8.351e-03  3.026e-03  -2.759  0.00607 **
## displacement           -9.357e-02  3.919e-02  -2.387  0.01746 *
## cylinders               1.851e+00  2.075e+00   0.892  0.37289
## weight:displacement     2.499e-05  8.250e-06   3.029  0.00262 **
## weight:cylinders       -3.801e-04  6.720e-04  -0.566  0.57197
## displacement:cylinders -2.026e-03  3.826e-03  -0.529  0.59682
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.106 on 385 degrees of freedom
## Multiple R-squared:  0.7275, Adjusted R-squared:  0.7232
## F-statistic: 171.3 on 6 and 385 DF,  p-value: < 2.2e-16
```

The interaction between weight and displacement is the only statistically significant 2nd order term in this model. Other interactions like (weight, cylinders) and (displacement, cylinders) are statistically insignificant.

(f) Try transformations of the variables with X^3 and log(X). Comment on your findings.

```
auto.pow_transform_fit=lm(mpg~.,data=subset(auto_df, select=-c(name))^3)
summary(auto.pow_transform_fit)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = subset(auto_df, select = -c(name))^3)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21793  -7195  -1400   5067  62213
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.726e+04  4.812e+03  -5.664 2.90e-08 ***
## cylinders    -1.685e+01  7.298e+00  -2.308 0.021516 *
## displacement  2.938e-04  8.944e-05   3.285 0.001113 **
## horsepower    1.820e-04  6.479e-04   0.281 0.778916
## weight       -2.800e-07  4.837e-08  -5.789 1.47e-08 ***
## acceleration  1.099e+00  2.982e-01   3.686 0.000261 ***
## year          1.030e-01  9.612e-03  10.712  < 2e-16 ***
## origin        3.959e+02  6.383e+01   6.202 1.44e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11080 on 384 degrees of freedom
## Multiple R-squared:  0.5788, Adjusted R-squared:  0.5711
## F-statistic: 75.37 on 7 and 384 DF,  p-value: < 2.2e-16
```

If we construct a multiple linear regression model with each predictor variable X transformed as X^3, the R^2 value decreases which shows a poor fit of this model compared to the original.

```
auto.log_transform_fit=lm(mpg~.,data=log(subset(auto_df, select=-c(name))))
summary(auto.log_transform_fit)
```

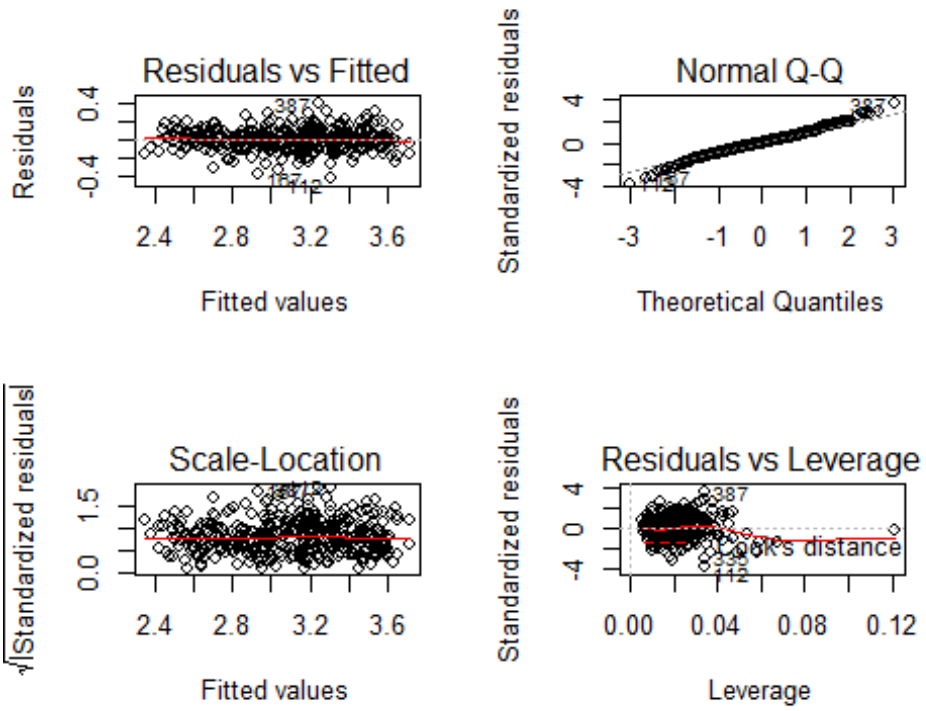
```
##
## Call:
## lm(formula = mpg ~ ., data = log(subset(auto_df, select = -c(name))))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41298 -0.07098  0.00055  0.06150  0.39532
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.155391   0.648230  -0.240  0.81068
## cylinders    -0.082815   0.061429  -1.348  0.17841
```

```
## displacement  0.006625   0.056970   0.116  0.90748
## horsepower   -0.294389   0.057652  -5.106 5.18e-07 ***
## weight       -0.569666   0.082397  -6.914 1.98e-11 ***
## acceleration -0.179239   0.059536  -3.011  0.00278 **
## year          2.243989   0.131661  17.044  < 2e-16 ***
## origin        0.044848   0.018821   2.383  0.01767 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1136 on 384 degrees of freedom
## Multiple R-squared:  0.8903, Adjusted R-squared:  0.8883
## F-statistic: 445.3 on 7 and 384 DF,  p-value: < 2.2e-16
```

Log transform seems better (for multiple linear regression) as it increase the R^2 coefficient. The statistically significant predictors change in this case as compared to the original.

We also look at the residuals vs fitted plot given below to justify that the model with log-transformed variables is a better fit since the trend is almost horizontal.

```
par(mfrow=c(2,2))
plot(auto.log_transform_fit)
```



Question 2

```
suppressMessages(library(MASS))
attach(Boston)
boston_df <- na.omit(Boston)
```

(a) For each predictor, fit a simple linear regression model to predict the response. Include the code, but not the output for all models in your solution. In which of the models is there a statistically significant association between the predictor and the response? Considering the meaning of each variable, discuss the relationship between crim and nox, chas, medv and dis in particular. How do these relationships differ?

```
for (n in names(Boston)){
  lin_reg_models <- list()
  if (n!='crim'){
    lin_reg_models[[n]] <- (lm(crim~get(n), data=Boston))
  }
}
```

All variables other than 'chas' is statistically significant w.r.t. the response variable 'crim' because the p-values for the coefficients of all these variables is significantly low. According to the linear prediction model:

(1) A high coefficient (31.249) for nox predictor implies that the per capita crime rate increases significantly with the increase in nitrogen oxides concentration in the air. Increase in Nitrogen oxide (harmful pollutant) content might lead to adverse living conditions but the relation between crime rate and bad living conditions is not necessarily an established one.

(2) There is no statistically significant relationship between crime rate and Charles River dummy variable. This tracks because both the variables seems unrelated.

(3) The crime rate decreases with the increase in median value of owner-occupied homes as the coefficient for 'medv' variable is statistically significant.

(4) The crime rate increase with the decrease in distance from Boston's employment centers ('dis' variable). This relationship goes against common sense because more employment opportunities tends to result in a decrease in crime rate in general. This suggests that we should consider some interactions among the predictor variables.

(b) Fit a multiple regression model to predict the response using all the predictors. Describe your results. For which predictors can we reject the null hypothesis?

```
summary(lm(crim~., data=Boston))
```

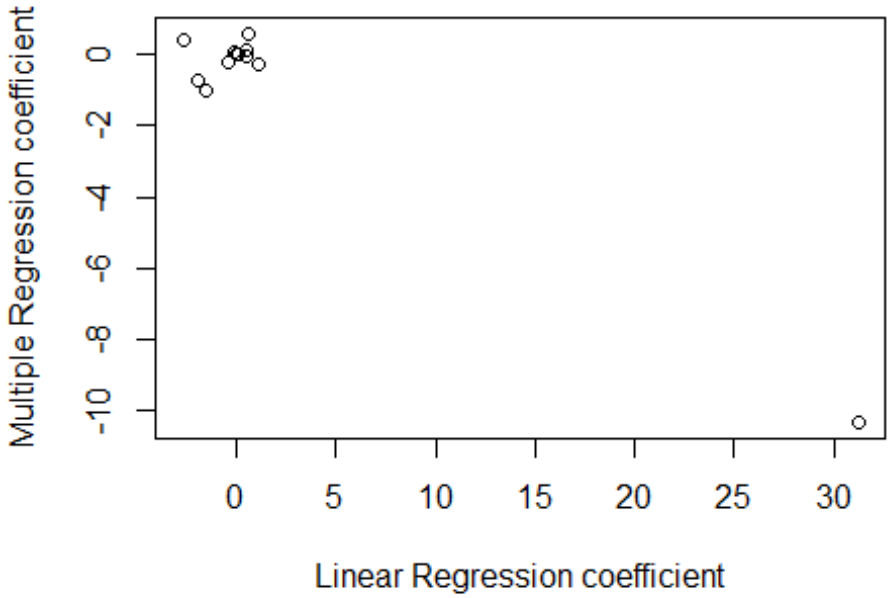
```
##
## Call:
## lm(formula = crim ~ ., data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.924 -2.120 -0.353  1.019 75.051
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.033228   7.234903   2.354 0.018949 *
## zn            0.044855   0.018734   2.394 0.017025 *
## indus        -0.063855   0.083407  -0.766 0.444294
## chas         -0.749134   1.180147  -0.635 0.525867
## nox         -10.313535   5.275536  -1.955 0.051152 .
## rm            0.430131   0.612830   0.702 0.483089
## age           0.001452   0.017925   0.081 0.935488
## dis          -0.987176   0.281817  -3.503 0.000502 ***
## rad           0.588209   0.088049   6.680 6.46e-11 ***
## tax          -0.003780   0.005156  -0.733 0.463793
## ptratio      -0.271081   0.186450  -1.454 0.146611
## black        -0.007538   0.003673  -2.052 0.040702 *
## lstat         0.126211   0.075725   1.667 0.096208 .
## medv         -0.198887   0.060516  -3.287 0.001087 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.439 on 492 degrees of freedom
## Multiple R-squared:  0.454,  Adjusted R-squared:  0.4396
## F-statistic: 31.47 on 13 and 492 DF,  p-value: < 2.2e-16
```

We reject the null hypothesis for 'zn', 'dis', 'rad', 'black', and 'medv' because their coefficients have statistically significant p-values(<0.05).

(c) How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis. What does this plot tell you about the various predictors?

```
linear_coeff <- list()
for(n in names(Boston[-1])){
  linear_coeff[[n]] <- lm(crim~get(n), data=Boston)$coefficients[2]
}
multiple_coeff <- lm(crim~., data=Boston)$coefficients[-1]
```

```
plot(linear_coeff, multiple_coeff,
     xlab='Linear Regression coefficient',
     ylab='Multiple Regression coefficient')
```



The plot suggests following points about the two models:

(1) The magnitude of multiple linear regression coefficients for most of the predictors (nox -> outlier) remains within a small range of their respective linear regression coefficients.

(2) The sign of the coefficients remains same for statistically significant predictors of multiple simple regression model (i.e. 'dis', 'rad', 'black', and 'medv') while the sign of most other predictors change when going from simple to multiple linear regression. This shows that it is important to take into account the interactions within the various predictors.

(d) Is there evidence of non-linear association between any of the predictors and the response?

```
for (n in names(Boston[-1])){
    non_linear_models <- list()
    if (n!='chas'){
      non_linear_models[[n]] <- lm(crim~poly(get(n), 3), data=Boston)
```

```
  }
}
```

There is no non-linear relationship between 'black' variable and the response as signified by the high p-values for 2nd and 3rd order coefficients. Some variables like 'nox', 'indus', 'dis', 'age', 'ptratio' and 'medv' show 3rd order relationship with 'crim' variable while 'zn', 'rad', 'tax', 'rm' and 'lstat' shows 2nd order relationship.

Question 3

(a) What are the issues that could arise in using linear regression (via least squares estimates) when error terms are correlated?

i) Regression coefficient represents the rate of change of one variable as a function of changes in the other. When errors are correlated, the regression coefficients are unbiased but they no longer have minimum variance.

ii) The standard error of the coefficient measures how precisely the model estimates the coefficient's unknown value. A precise estimate has a low standard error. When errors are correlated, standard errors estimated will be far less than they actually are. This will make the results seem more accurate than they really are.

iii) Confidence intervals are a measure of overall quality of regression. These intervals and other tests of significance will no longer be valid as the confidence interval will narrow down. Thus we may have an unwarranted sense of confidence in the model.

(b) What methods can be applied to deal with correlated errors? Mention at least one method.

The major issue in least squares estimates was the standard errors calculation, and hence we could employ other errors like Newey-West standard errors. We could also use other linear estimators (like feasible general least sqaures (f-GLS)) which are better than ordinary least squares. The other way of tackling would be to apply transformation, where we transform the response Y using a concave function such as log Y. This results in shrinkage of the larger response. If we have a good idea of the variance of each response, we can fit our model by weighted least squares. weighted least squares is a simple remedy to fit the model with weights proportional to the inverse variances. This would reduce the variance and correlation among the error terms. Sometimes correlation is due to the omission of a variable from the model. Thus uncovering this variable would solve this issue.