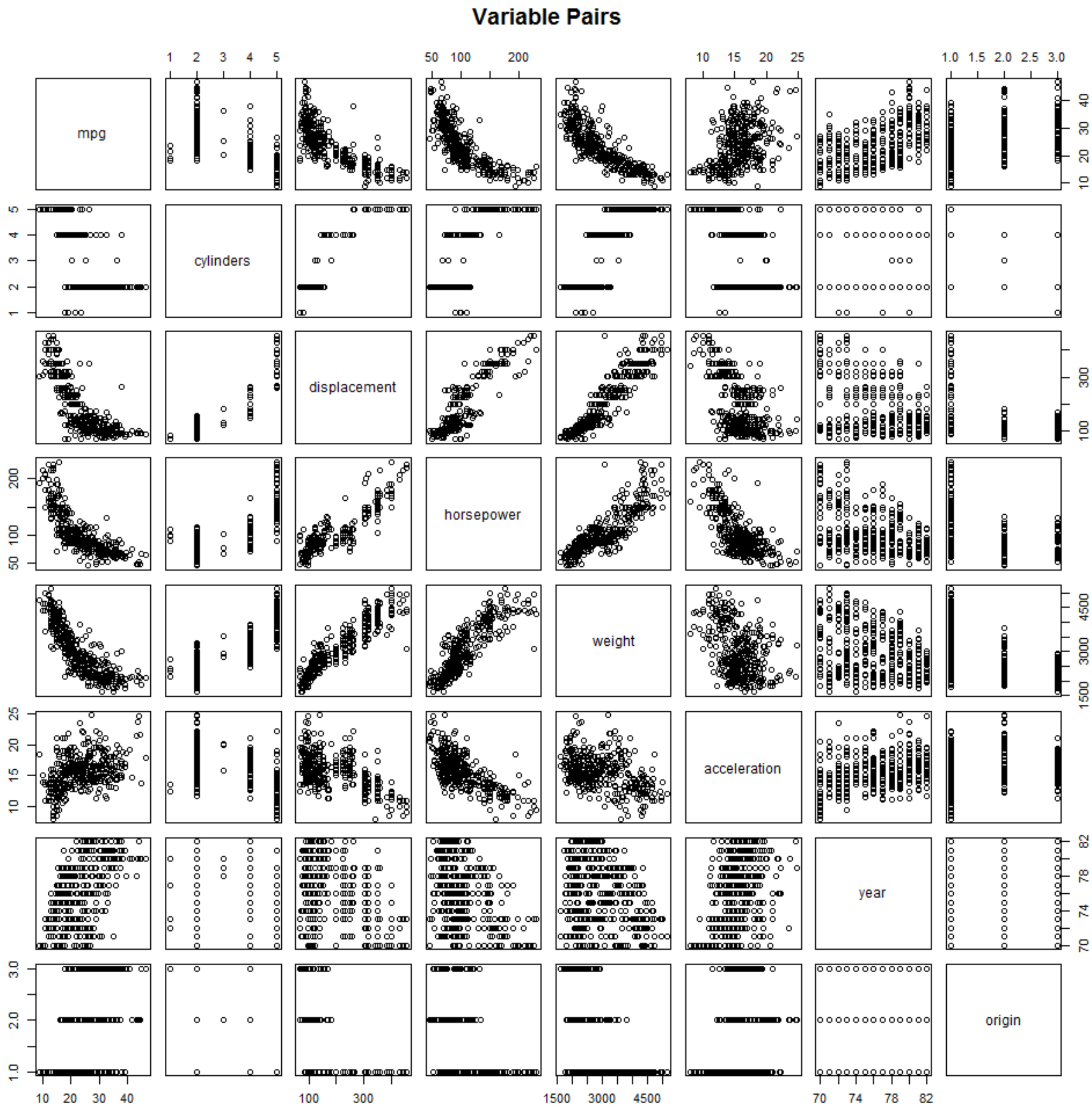
Ryan Torelli
CptS 483-04
Assignment 3
October 2, 2017

```
# 1.
# Read Auto.csv
Auto <- read.csv("Auto.csv", header=TRUE, colClasses=c("name"="character"), na.strings="?")

# Omit missing data
dim(Auto)
Auto <- na.omit(Auto)
dim(Auto)
<output omitted>

# Show variables and name
names(Auto)
<output omitted>
```

ISLR defines numeric variables as quantitative and categorical variables as qualitative. For the Auto dataset introduced in chapter 1 and worked with as a lab in chapter 2, ISLR treats cylinders as qualitative and origin as quantitative. Cylinders is numeric, discrete, ordered, and has range [4,8]. Origin is numeric, discrete, unordered, and has range [1,3]. In this assignment for the Auto dataset, I treat names as qualitative and all other variables as quantitative.

```
# (a)
# Plot variables by scatterplot
pairs(subset(Auto, select=-name), main="Variable Pairs")
```



# **(b)**
# Compute correlations
cor(subset(Auto, select=-name))

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

(c)
Regress mpg on all variables

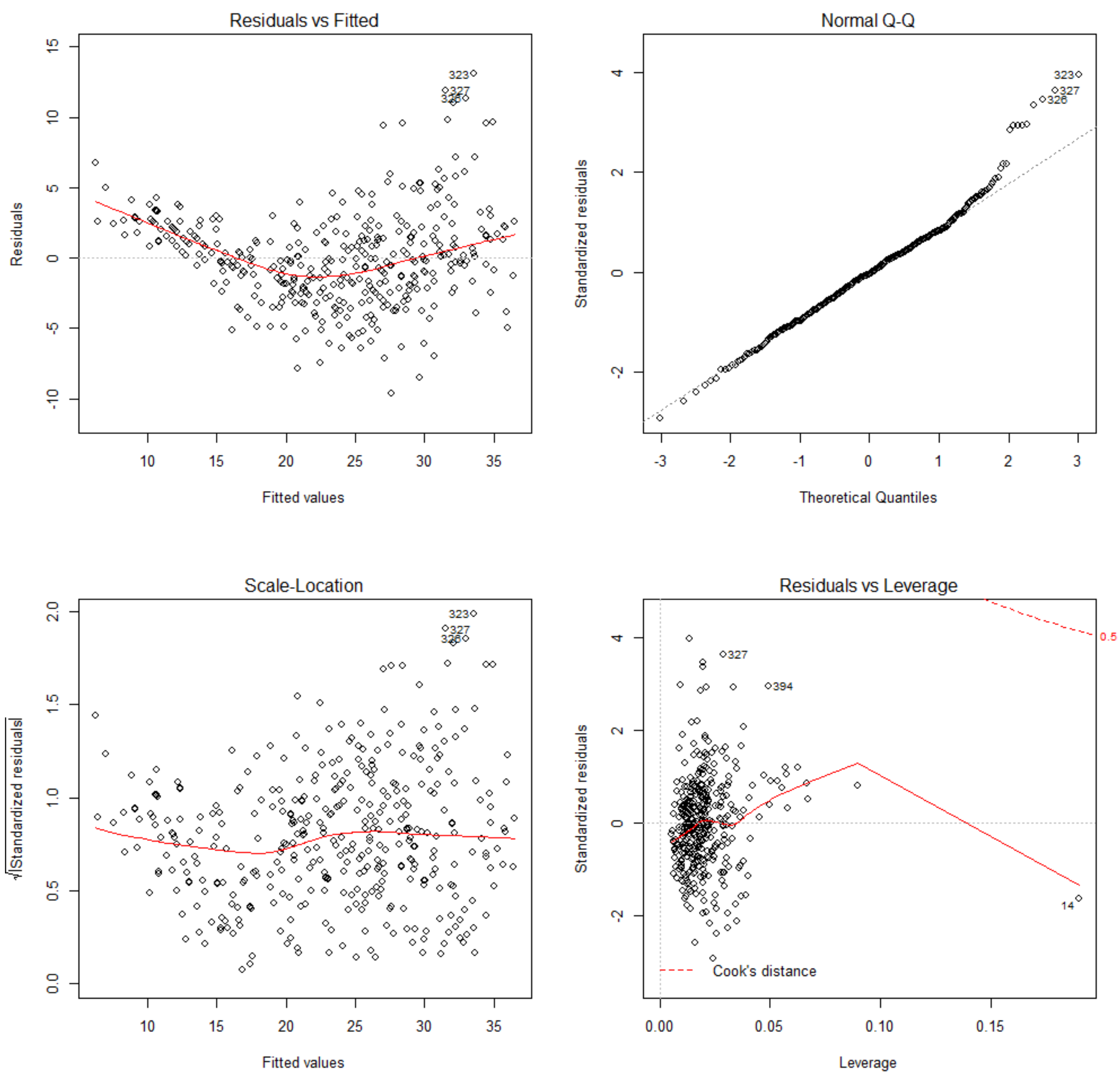
```
mpg.regression <- lm(mpg~cylinders+displacement+horsepower+weight+acceleration+
                     year+origin, data=Auto)
summary(mpg.regression)
```

```
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.
The probability of observing a value equal to or greater than the absolute value of the t-statistic indicates the significance variable-response relationship. The column labeled Pr(>|t|) lists these probabilities, called p-values, for each variable. For a threshold of 0.05, the significant variables are displacement, weight, year, and origin.

ii.
The coefficient of year indicates change in mpg per year. From one year to the next, mpg increases by 0.75 according to the model. Fuel efficiency is increasing.

```
# (d)
# Plot diagnostics of regression
par(mfrow=c(2,2))
plot(mpg.regression)
```



The Residuals v Fitted plot distinguishes linear from non-linear residual patterns. The plot is mildly non-linear with greater dispersion at larger fitted values. There are about five to ten residuals with absolute value of ten or more.

The Normal Q-Q plot distinguishes normal from non-normal residual distributions. The plot deviates from normal toward the greater positive quantiles.

The Scale-Location plot shows variance among variables. The plot is largely horizontal. There is a homogenous spread of residuals along fitted values.

The Residuals v Leverage plot identifies extreme values that adversely impact fit of regression. All data is within Cook's distance but one data point, labeled 14, has leverage > 0.15.

(e)
There are seven variables that are not mpg, meaning there are 7!/(2!(7-2)!) pairs and as many interaction terms to check. Instead, select the four statistically significant variables from 1(c) and calculate the interaction term for the pair most highly correlated.

```
# Correlate displacement, weight, year, and origin
cor(subset(Auto, select=c(displacement,weight,year,origin)))
```

```
             displacement     weight       year     origin
displacement    1.0000000  0.9329944 -0.3698552 -0.6145351
weight          0.9329944  1.0000000 -0.3091199 -0.5850054
year           -0.3698552 -0.3091199  1.0000000  0.1815277
origin         -0.6145351 -0.5850054  0.1815277  1.0000000
```

```
# Regress mpg on four variables and one interaction term
mpg.regression <- lm(mpg~displacement*weight+year+origin, data=Auto)
summary(mpg.regression)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-10.6119  -1.7290  -0.0115   1.5609  12.5584 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)         -8.007e+00  3.798e+00  -2.108   0.0357 *  
displacement        -7.148e-02  9.176e-03  -7.790 6.27e-14 ***
weight              -1.054e-02  6.530e-04 -16.146  < 2e-16 ***
year                 8.194e-01  4.518e-02  18.136  < 2e-16 ***
origin               3.567e-01  2.574e-01   1.386   0.1666    
displacement:weight  2.104e-05  2.214e-06   9.506  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.016 on 386 degrees of freedom
Multiple R-squared:  0.8526,	Adjusted R-squared:  0.8507 
F-statistic: 446.5 on 5 and 386 DF,  p-value: < 2.2e-16
```

Displacement and weight have a statistically significant interaction term using threshold $p<0.05$.

(f)
There are seven variables that are not mpg and two transforms under consideration. For simplicity, select and transform one variable from among the four statistically significant variables in 1(c).

```
# Regress mpg on weight
mpg.regression <- lm(mpg~weight, data=Auto)

# Regress mpg on weight + weight^2
mpg.regression.squared <- lm(mpg~weight+I(weight^2), data=Auto)

# Regress mpg on weight + log(weight)
mpg.regression.log <- lm(mpg~weight+log(weight), data=Auto)

# Compare models by ANOVA
anova(mpg.regression, mpg.regression.squared, mpg.regression.log)
```

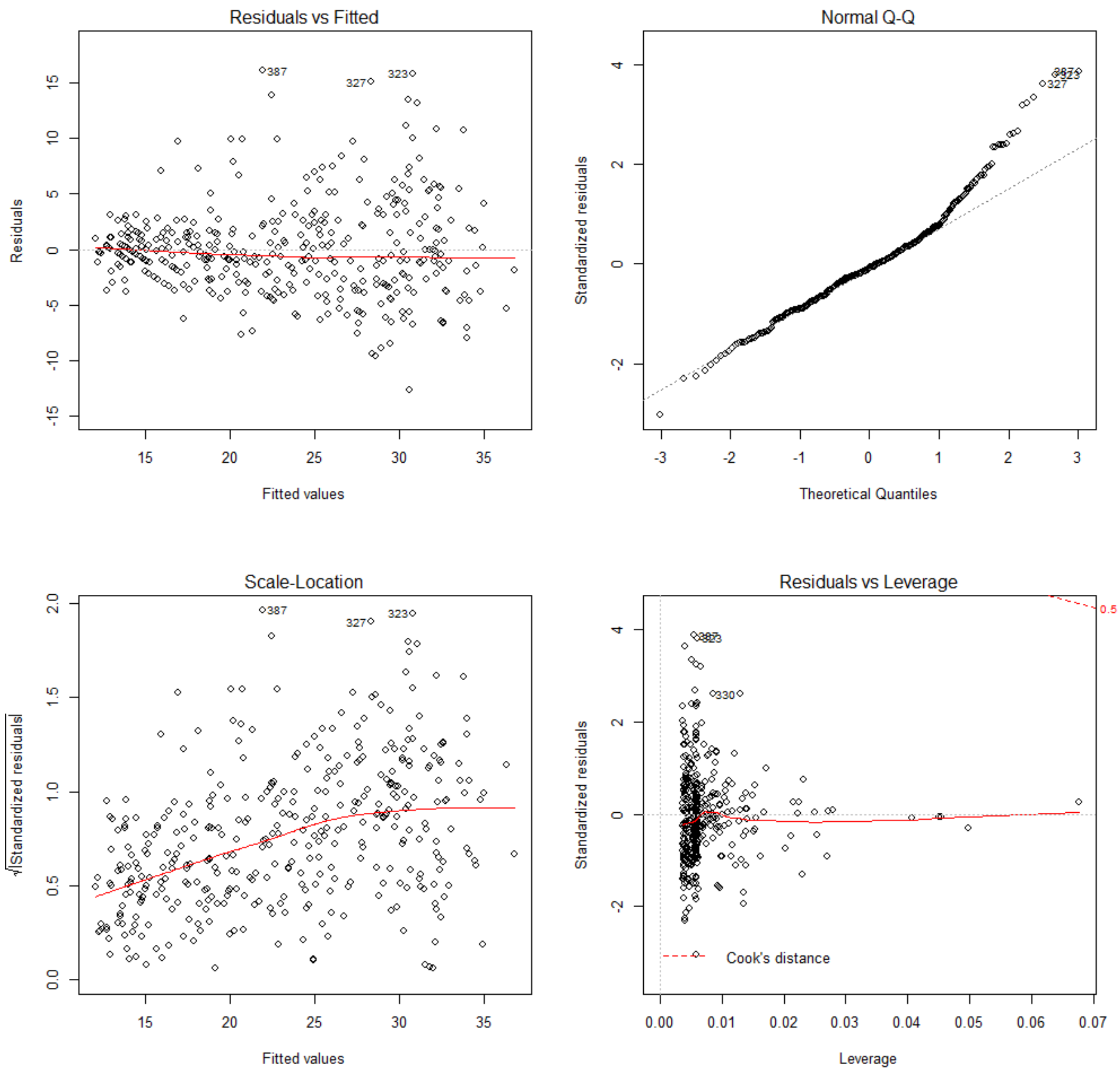
```
Analysis of Variance Table

Model 1: mpg ~ weight
Model 2: mpg ~ weight + I(weight^2)
Model 3: mpg ~ weight + log(weight)
Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1    390 7321.2
2    389 6784.9  1    536.34 30.75 5.429e-08 ***
3    389 6812.2  0    -27.29
---
```

The model with weight and weight^2 is superior to a model with only weight and a model with weight + log(weight).

```
# Summarize fit of superior model
summary(mpg.regression.squared)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-12.6246  -2.7134  -0.3485   1.8267  16.0866

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.226e+01  2.993e+00  20.800  < 2e-16 ***
weight      -1.850e-02  1.972e-03  -9.379  < 2e-16 ***
I(weight^2)  1.697e-06  3.059e-07   5.545 5.43e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.176 on 389 degrees of freedom
Multiple R-squared:  0.7151,	Adjusted R-squared:  0.7137
F-statistic: 488.3 on 2 and 389 DF,  p-value: < 2.2e-16
```

# Plot diagnostics of regression
par(mfrow=c(2,2))
plot(mpg.regression.squared)



The Residuals v Fitted plot distinguishes linear from non-linear residual patterns. The plot is linear.
The Normal Q-Q plot distinguishes normal from non-normal residual distributions. The plot deviates from normal toward the positive quantiles.
The Scale-Location plot shows variance among variables. The plot is not horizontal. There is a non-homogenous spread of residuals along fitted values.
The Residuals v Leverage plot identifies extreme values that adversely impact fit of regression. All data is within Cook's distance. The data point with greatest leverage has leverage of 0.07.

```
# 2.
# Load library that contains Boston dataset
library(MASS)

# View variables and dimensions
names(Boston)
dim(Boston)
<output omitted>

# (a)
#Regress crim on other variables
summary(lm(crim~zn))
summary(lm(crim~indus))
summary(lm(crim~chas))
summary(lm(crim~nox))
summary(lm(crim~rm))
summary(lm(crim~age))
summary(lm(crim~dis))
summary(lm(crim~rad))
summary(lm(crim~tax))
summary(lm(crim~ptratio))
summary(lm(crim~black))
summary(lm(crim~lstat))
summary(lm(crim~medv))
<output omitted>

# Summarize chas and view its values
summary(chas)
chas
<output omitted>
```

Each one of the variables but for chas fits a statistically significant linear model for p-value < 0.01. The variable chas takes values 0 or 1 and encodes information about the Charles River. It is not quantitative.

The relationships between crim and zn, nox, or ptratio differ by coefficient and intercept. The variable zn has negative coefficient, whereas nox and ptratio have positive coefficient. As zn increases, crim decreases and, as nox or ptratio increase, crim increases. The absolute value of the nox coefficient is greater than that of ptratio and ptratio greater than zn. For an increase of one unit, nox impacts crim most, ptratio second-most, and zn least. The variable zn has positive intercept, whereas nox and ptratio have negative intercept. Since crim is never negative, nox and ptratio have minimum value greater than zero. All three relationships are significant for p-value < 5e-06.

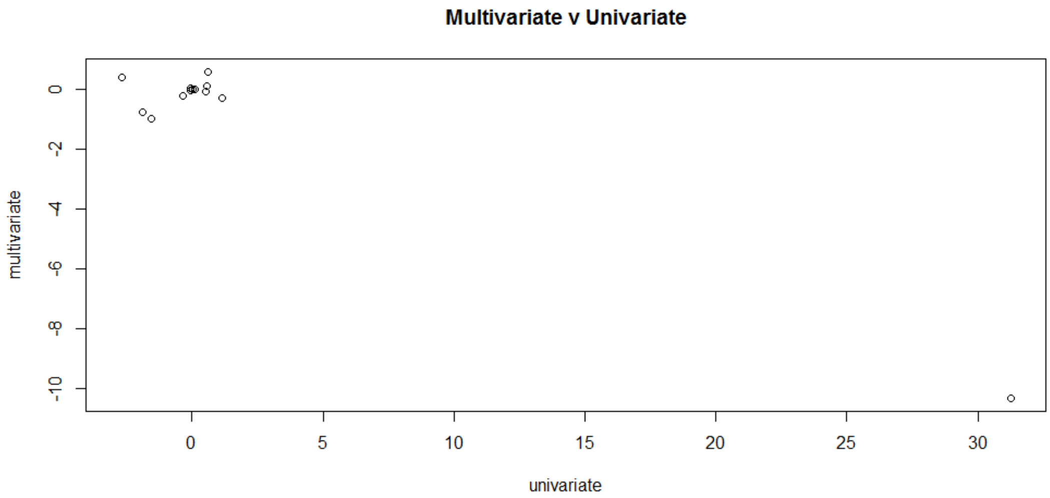
# **(b)**
# Regress crim on all variables
summary(lm(crim~., data=Boston))

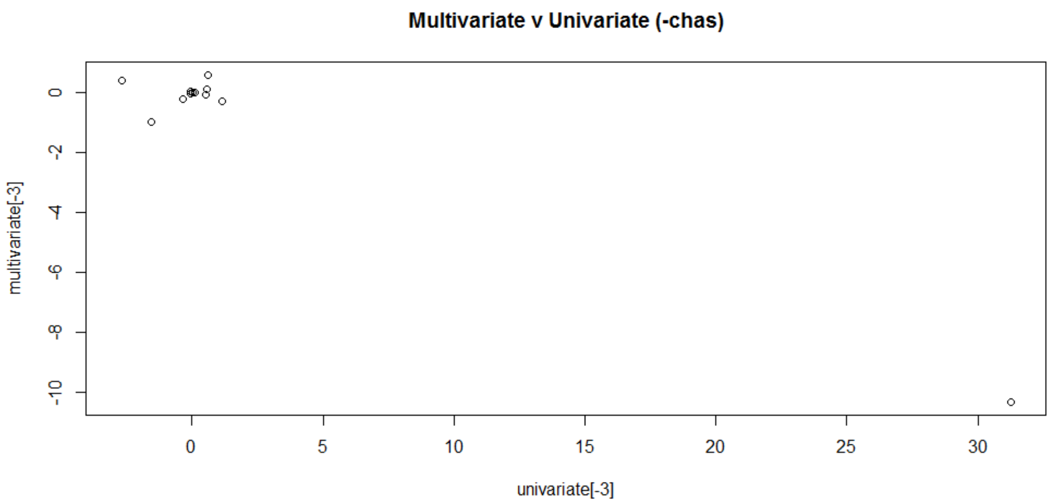
```
Residuals:
   Min     1Q Median     3Q    Max
-9.924 -2.120 -0.353  1.019 75.051

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
```

For threshold p<0.05, we reject the null hypothesis for variables zn, dis, rad, black, and medv.

(c)

```
# Plot univariate regression coefficients on x-axis and multivariate on y-axis
univariate <- c(coef(lm(crim~zn))[2], coef(lm(crim~indus))[2], coef(lm(crim~chas))[2],
                coef(lm(crim~nox))[2], coef(lm(crim~rm))[2], coef(lm(crim~age))[2],
                coef(lm(crim~dis))[2], coef(lm(crim~rad))[2], coef(lm(crim~tax))[2],
                coef(lm(crim~ptratio))[2], coef(lm(crim~black))[2], coef(lm(crim~lstat))[2],
                coef(lm(crim~medv))[2])
multivariate <- coef(lm(crim~., data=Boston))[2:14]
par(mfrow=c(2,1))
plot(univariate, multivariate, main="Multivariate v Univariate")
plot(univariate[-3], multivariate[-3], main="Multivariate v Univariate (-chas)")
```





Six of 13 coefficients change sign from univariate to multivariate. The greatest absolute change occurs with the variable nox, which changes from 31 to -10. The greatest relative change occurs with variable age, which changes from 0.1 to 0.001.

(d)
Regress crim on other variables as third-order polynomial

```
summary(lm(crim~poly(zn,3)))
summary(lm(crim~poly(indus,3)))
summary(lm(crim~poly(nox,3)))
summary(lm(crim~poly(rm,3)))
summary(lm(crim~poly(age,3)))
summary(lm(crim~poly(dis,3)))
summary(lm(crim~poly(rad,3)))
summary(lm(crim~poly(tax,3)))
summary(lm(crim~poly(ptratio,3)))
summary(lm(crim~poly(black,3)))
summary(lm(crim~poly(lstat,3)))
summary(lm(crim~poly(medv,3)))
<output omitted>
```

Yes, there is evidence of non-linear association between variables and crim. The polynomial regression for variables zn, indus, nox, rm, age, dis, rad, tax, ptratio, lstat, and medv have at least one statistically significant polynomial coefficient for p-value < 0.05.

3.

```
# Install and load ISLR package
install.packages("ISLR")
library(ISLR)

# View variables and dimensions of Weekly dataset
names(Weekly)
dim(Weekly)
<output omitted>
```

The Direction variable in column nine is qualitative.

(a)

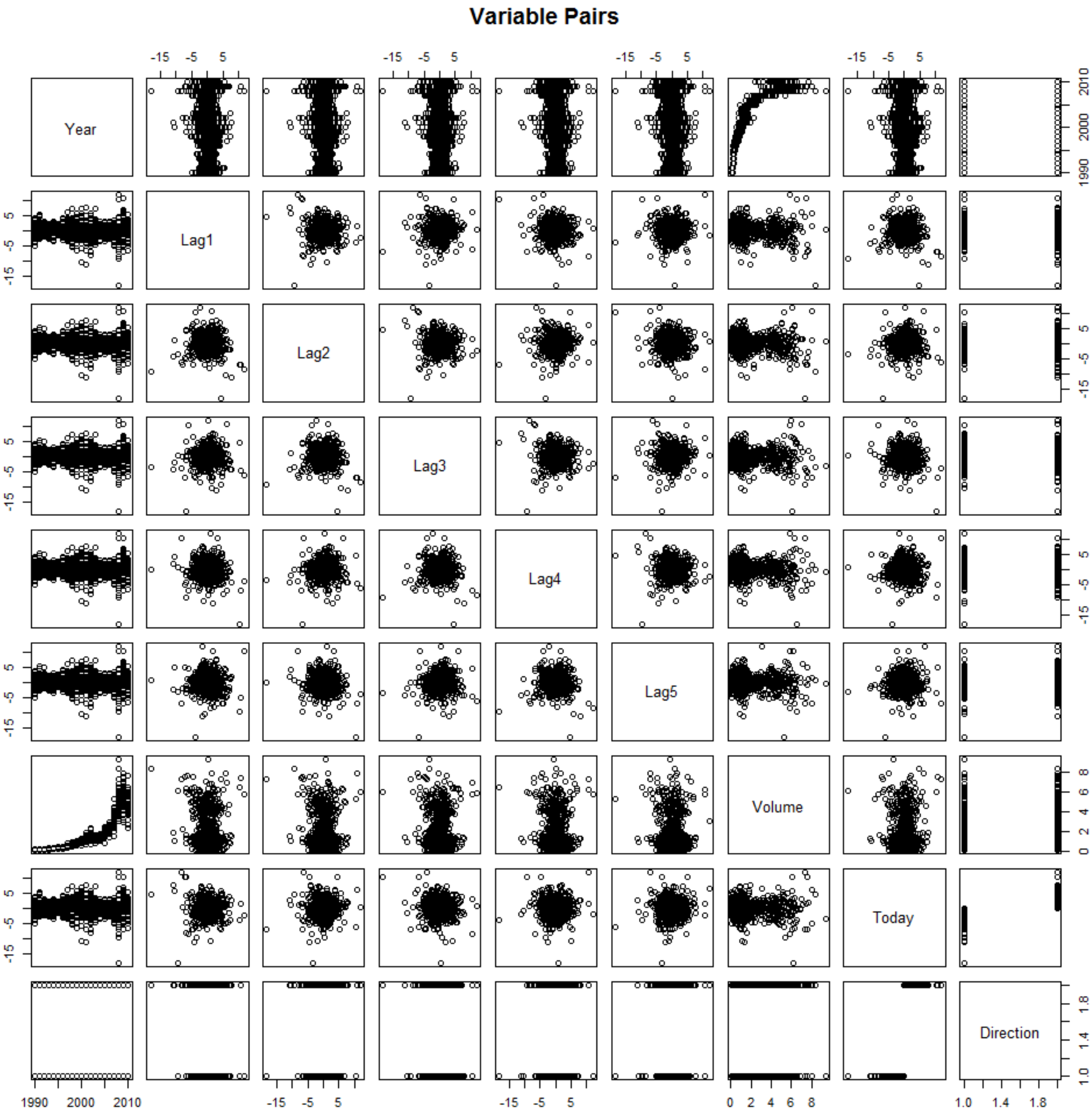
```
# Summarize Weekly
summary(Weekly)
```

```
     Year          Lag1              Lag2
 Min.   :1990   Min.   :-18.1950   Min.   :-18.1950
 1st Qu.:1995   1st Qu.: -1.1540   1st Qu.: -1.1540
 Median :2000   Median :  0.2410   Median :  0.2410
 Mean   :2000   Mean   :  0.1506   Mean   :  0.1511
 3rd Qu.:2005   3rd Qu.:  1.4050   3rd Qu.:  1.4090
 Max.   :2010   Max.   : 12.0260   Max.   : 12.0260
     Lag3              Lag4              Lag5
 Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950
 1st Qu.: -1.1580   1st Qu.: -1.1580   1st Qu.: -1.1660
 Median :  0.2410   Median :  0.2380   Median :  0.2340
 Mean   :  0.1472   Mean   :  0.1458   Mean   :  0.1399
 3rd Qu.:  1.4090   3rd Qu.:  1.4090   3rd Qu.:  1.4050
 Max.   : 12.0260   Max.   : 12.0260   Max.   : 12.0260
    Volume           Today          Direction
 Min.   :0.08747   Min.   :-18.1950   Down:484
 1st Qu.:0.33202   1st Qu.: -1.1540   Up  :605
 Median :1.00268   Median :  0.2410
 Mean   :1.57462   Mean   :  0.1499
 3rd Qu.:2.05373   3rd Qu.:  1.4050
 Max.   :9.32821   Max.   : 12.0260
```

```
# Correlate quantitative variables
cor(Weekly[-9])
```

```
            Year         Lag1          Lag2        Lag3         Lag4          Lag5         Volume       Today
Year   1.00000000 -0.032289274 -0.03339001 -0.03000649 -0.031127923 -0.030519101  0.84194162 -0.032459894
Lag1 -0.03228927  1.000000000 -0.07485305  0.05863568 -0.071273876 -0.008183096 -0.06495131 -0.075031842
Lag2  -0.03339001 -0.074853051  1.00000000 -0.07572091  0.058381535 -0.072499482 -0.08551314  0.059166717
Lag3  -0.03000649  0.058635682 -0.07572091  1.00000000 -0.075395865 0.060657175 -0.06928771 -0.071243639
Lag4 -0.03112792 -0.071273876  0.05838153 -0.07539587  1.000000000 -0.075675027 -0.06107462 -0.007825873
Lag5   -0.03051910 -0.008183096 -0.07249948  0.06065717 -0.075675027 1.000000000 -0.05851741  0.011012698
Vol 0.84194162 -0.064951313 -0.08551314 -0.06928771 -0.061074617 -0.058517414  1.00000000 -0.033077783
Today -0.03245989 -0.075031842  0.05916672 -0.07124364 -0.007825873 0.011012698 -0.03307778  1.000000000
```

```
# Plot pairs of Weekly
pairs(Weekly, main="Variable Pairs")
```



The strongest correlation is between Year and Volume. Volume grows with time.

(b)
Regress Direction on Lag1-5 and Volume

```
dir.regression < - glm(Direction~Lag1+Lag2+Lag3+Lag4+Lag5+Volume, data=Weekly,
                        family=binomial)
summary(dir.regression)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6949  -1.2565   0.9913   1.0849   1.4579

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.26686    0.08593   3.106   0.0019 **
Lag1        -0.04127    0.02641  -1.563   0.1181
Lag2         0.05844    0.02686   2.175   0.0296 *
Lag3        -0.01606    0.02666  -0.602   0.5469
Lag4        -0.02779    0.02646  -1.050   0.2937
Lag5        -0.01447    0.02638  -0.549   0.5833
Volume      -0.02274    0.03690  -0.616   0.5377
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1496.2  on 1088  degrees of freedom
Residual deviance: 1486.4  on 1082  degrees of freedom
AIC: 1500.4

Number of Fisher Scoring iterations: 4
```

The regression coefficient of Lag2 is statistically significant for p-value < 0.05.

(c)
Predict Direction on probability threshold 0.5;
"Up" on prob > 0.5; otherwise, "Down"

```
probabilities = predict(dir.regression, type = "response")
predictions = rep("Down", length(probabilities))
predictions[probabilities > 0.5] = "Up"
table(predictions, Weekly$Direction)
mean(predictions == Weekly$Direction)
```

```
predictions Down  Up
       Down   54  48
       Up    430 557
[1] 0.5610652
```

The prediction accuracy is the percentage of correct predictions from among all predictions and is 56%. If "Up" is predicted, the prediction is correct (557)/(48+557) or 92% of the time. Whereas, if "Down" is predicted, the prediction is correct 54/(54+430) or 11% of the time. The model predicts "Up" with greater accuracy than "Down" for the training set.

(d)

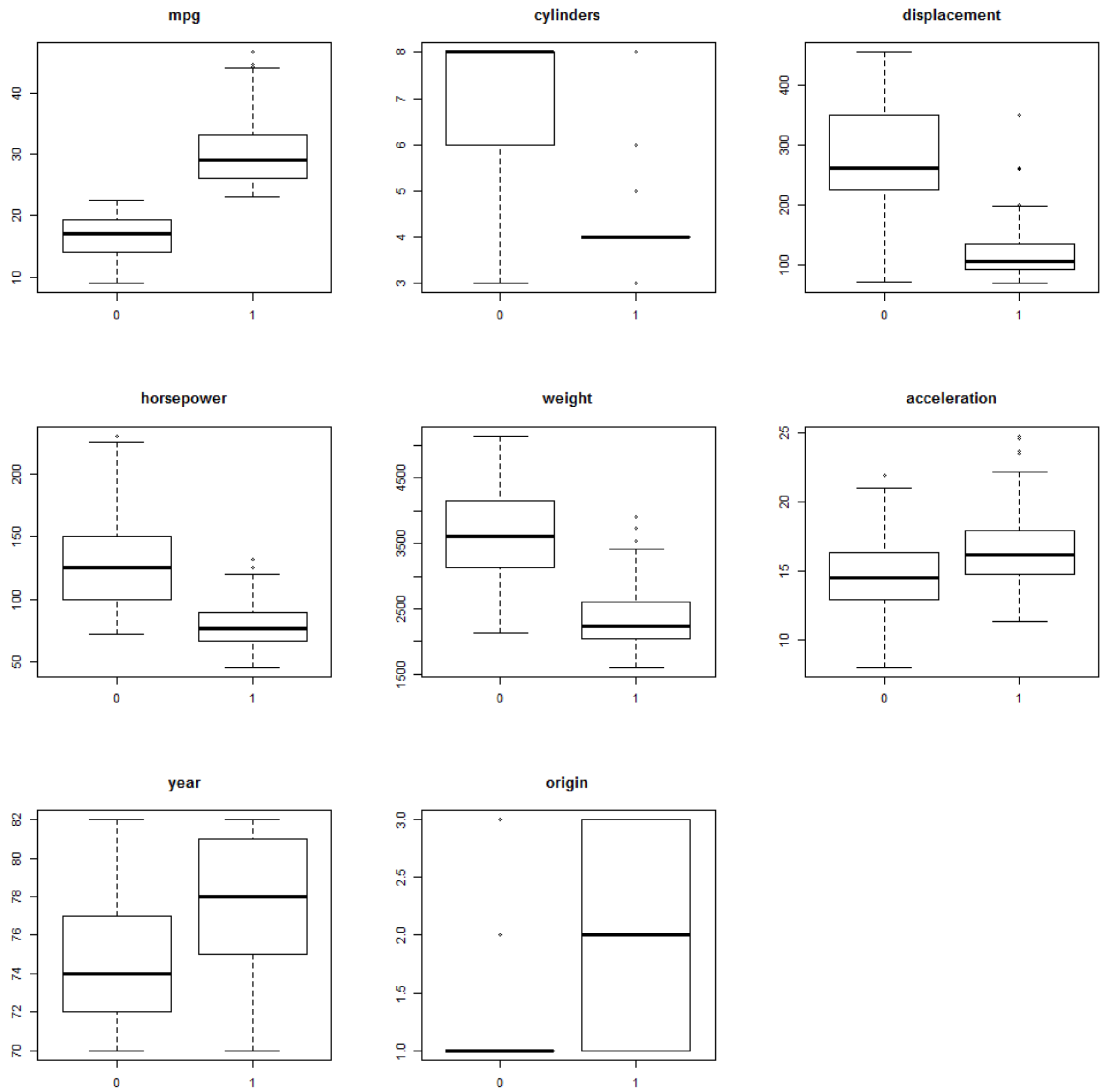
```
# Regress Direction on Lag2 for 1990-2007
train <- Weekly$Year < 2008
dir.regression <- glm(Direction~Lag2, data=Weekly, family=binomial, subset=train)

# Predict Direction for 2008-2010 on probability threshold 0.5;
# "Up" on prob > 0.5; otherwise, "Down"
test <- Weekly[!train, ]
probabilities <- predict(dir.regression, test, type = "response")
predictions <- rep("Down", length(probabilities))
predictions[probabilities > 0.5] <- "Up"
table(predictions, Weekly$Direction[!train])
mean(predictions == Weekly$Direction[!train])
```

```
predictions Down Up
       Down    7  5
       Up     65 79
[1] 0.5512821
```

This model was trained on data from 1990-2007 and tested on held-out data from 2008-2010. Its prediction accuracy is worse than the model in 3(c) that trained on the entire dataset and tested on the same.

```r
# 4.
# Read Auto.csv
Auto <- read.csv("Auto.csv", header=TRUE, colClasses=c("name"="character"), na.strings="?")

# Omit missing data
dim(Auto)
Auto <- na.omit(Auto)
dim(Auto)
<output omitted>

# Show variables and name
names(Auto)
<output omitted>

# (a)
# Create binary variable mpg01 that takes 1 if mpg > median(mpg), 0 otherwise
mpg01 <- rep(0, length(Auto$mpg))
median.mpg <- median(Auto$mpg)
mpg01[Auto$mpg > median.mpg] <- 1

# Add mpg01 to Auto
Auto$mpg01 <- mpg01
dim(Auto)
<output omitted>
```

(b)
```
# Plot mpg01 as factor against other variables
mpg01 <- as.factor(Auto$mpg01)
par(mfrow=c(3,3))
plot(mpg01, Auto$mpg, main="mpg")
plot(mpg01, Auto$cylinders, main="cylinders")
plot(mpg01, Auto$displacement, main="displacement")
plot(mpg01, Auto$horsepower, main="horsepower")
plot(mpg01, Auto$weight, main="weight")
plot(mpg01, Auto$acceleration, main="acceleration")
plot(mpg01, Auto$year, main="year")
plot(mpg01, Auto$origin, main="origin")
```



The variables most likely to predict mpg01 are those with values that separate along a threshold value for co-occurrence of mpg values above or below the median mpg. More simply, variables that bin on mpg01 with clear separation are predictors. Excluding mpg, the variables cylinders, displacement, horsepower, and weight have boxplots on mpg01 that separate best.

```
# Compute correlation of mpg01 and other variables
cor(subset(Auto, select=-name))[,9]
```

```
             mpg  cylinders displacement horsepower     weight acceleration      year    origin     mpg01
mpg01  0.8369392 -0.7591939  -0.7534766 -0.6670526 -0.7577566    0.3468215 0.4299042 0.5136984 1.0000000
```

```
# (c)
# Regress mpg01 on cylinders, displacement, horsepower, and weight for rows [1,80]
train <- rep(FALSE, length(Auto$mpg01))
train[1:80] <- TRUE
mpg01.regression <- glm(mpg01~cylinders+displacement+horsepower+weight, data=Auto,
                        family=binomial, subset=train)

# Predict mpg01 on probability threshold 0.5; 1 on prob > 0.5; otherwise, 0
test <- Auto[!train, ]
probabilities <- predict(mpg01.regression, test, type = "response")
predictions <- rep(0, length(probabilities))
predictions[probabilities > 0.5] <- 1
table(predictions, Auto$mpg01[!train])
mean(predictions == Auto$mpg01[!train])
```

```
predictions   0   1
          0 129  64
          1   9 110
[1] 0.7660256
```

The model was trained on cylinders, displacement, horsepower, and weight from rows [1,80] and tested on rows [81,392]. Its prediction accuracy is 0.766. Its test error is 1-0.766 or 0.234.