FinalExamCode

Tristan Tran

12/17/2020

Problem 3

We are interested in finding the important predictors of the houses in Boston, assess their adjusted effect sizes (in direction and magnitude) and use the best linear regression model for interpretation and prediction. We want to analyze the BostonHousingdataset that contains 506 observations on 14 variables. The dataset is a part of the R mlbench package that also provides the necessary variable descriptions (?BostonHousing). Perform all necessary data analysis steps and write a section summarizing the finding

The fields are all continuous except for the two categorical variables chas and rad.

chas : Charles River dummy variable (=1 if tract bounds river; 0 otherwise) (categorical)

rad : index of accessibility to radial highways

```
data("BostonHousing")
base=lm(medv~.,data=BostonHousing)
summary(base)
```

```
##
## Call:
## lm(formula = medv ~ ., data = BostonHousing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.595  -2.730  -0.518   1.777  26.199
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
## crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
## zn           4.642e-02  1.373e-02   3.382 0.000778 ***
## indus        2.056e-02  6.150e-02   0.334 0.738288
## chas1        2.687e+00  8.616e-01   3.118 0.001925 **
## nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
## rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
## age          6.922e-04  1.321e-02   0.052 0.958229
```

```
## dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
## rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
## tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
## ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
## b            9.312e-03  2.686e-03   3.467 0.000573 ***
## lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.745 on 492 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7338
## F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

Our base R-squared is 0.7406.and Adjusted R-Squared of: 0.7338.

```
final = stepAIC(base)
```

```
## Start:  AIC=1589.64
## medv ~ crim + zn + indus + chas + nox + rm + age + dis + rad +
##     tax + ptratio + b + lstat
##
##           Df Sum of Sq   RSS    AIC
## - age      1      0.06 11079 1587.7
## - indus    1      2.52 11081 1587.8
## <none>                 11079 1589.6
## - chas     1    218.97 11298 1597.5
## - tax      1    242.26 11321 1598.6
## - crim     1    243.22 11322 1598.6
## - zn       1    257.49 11336 1599.3
## - b        1    270.63 11349 1599.8
## - rad      1    479.15 11558 1609.1
## - nox      1    487.16 11566 1609.4
## - ptratio  1   1194.23 12273 1639.4
## - dis      1   1232.41 12311 1641.0
## - rm       1   1871.32 12950 1666.6
## - lstat    1   2410.84 13490 1687.3
##
## Step:  AIC=1587.65
## medv ~ crim + zn + indus + chas + nox + rm + dis + rad + tax +
##     ptratio + b + lstat
##
##           Df Sum of Sq   RSS    AIC
## - indus    1      2.52 11081 1585.8
## <none>                 11079 1587.7
## - chas     1    219.91 11299 1595.6
## - tax      1    242.24 11321 1596.6
## - crim     1    243.20 11322 1596.6
## - zn       1    260.32 11339 1597.4
## - b        1    272.26 11351 1597.9
## - rad      1    481.09 11560 1607.2
## - nox      1    520.87 11600 1608.9
## - ptratio  1   1200.23 12279 1637.7
## - dis      1   1352.26 12431 1643.9
```

```
## - rm       1   1959.55 13038 1668.0
## - lstat    1   2718.88 13798 1696.7
##
## Step:  AIC=1585.76
## medv ~ crim + zn + chas + nox + rm + dis + rad + tax + ptratio +
##     b + lstat
##
##           Df Sum of Sq   RSS    AIC
## <none>                 11081 1585.8
## - chas     1    227.21 11309 1594.0
## - crim     1    245.37 11327 1594.8
## - zn       1    257.82 11339 1595.4
## - b        1    270.82 11352 1596.0
## - tax      1    273.62 11355 1596.1
## - rad      1    500.92 11582 1606.1
## - nox      1    541.91 11623 1607.9
## - ptratio  1   1206.45 12288 1636.0
## - dis      1   1448.94 12530 1645.9
## - rm       1   1963.66 13045 1666.3
## - lstat    1   2723.48 13805 1695.0
```

```
summary(final)
```

```
##
## Call:
## lm(formula = medv ~ crim + zn + chas + nox + rm + dis + rad +
##     tax + ptratio + b + lstat, data = BostonHousing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.5984  -2.7386  -0.5046   1.7273  26.2373
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.341145   5.067492   7.171 2.73e-12 ***
## crim         -0.108413   0.032779  -3.307 0.001010 **
## zn            0.045845   0.013523   3.390 0.000754 ***
## chas1         2.718716   0.854240   3.183 0.001551 **
## nox         -17.376023   3.535243  -4.915 1.21e-06 ***
## rm            3.801579   0.406316   9.356  < 2e-16 ***
## dis          -1.492711   0.185731  -8.037 6.84e-15 ***
## rad           0.299608   0.063402   4.726 3.00e-06 ***
## tax          -0.011778   0.003372  -3.493 0.000521 ***
## ptratio      -0.946525   0.129066  -7.334 9.24e-13 ***
## b             0.009291   0.002674   3.475 0.000557 ***
## lstat        -0.522553   0.047424 -11.019  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.736 on 494 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7348
## F-statistic: 128.2 on 11 and 494 DF,  p-value: < 2.2e-16
```

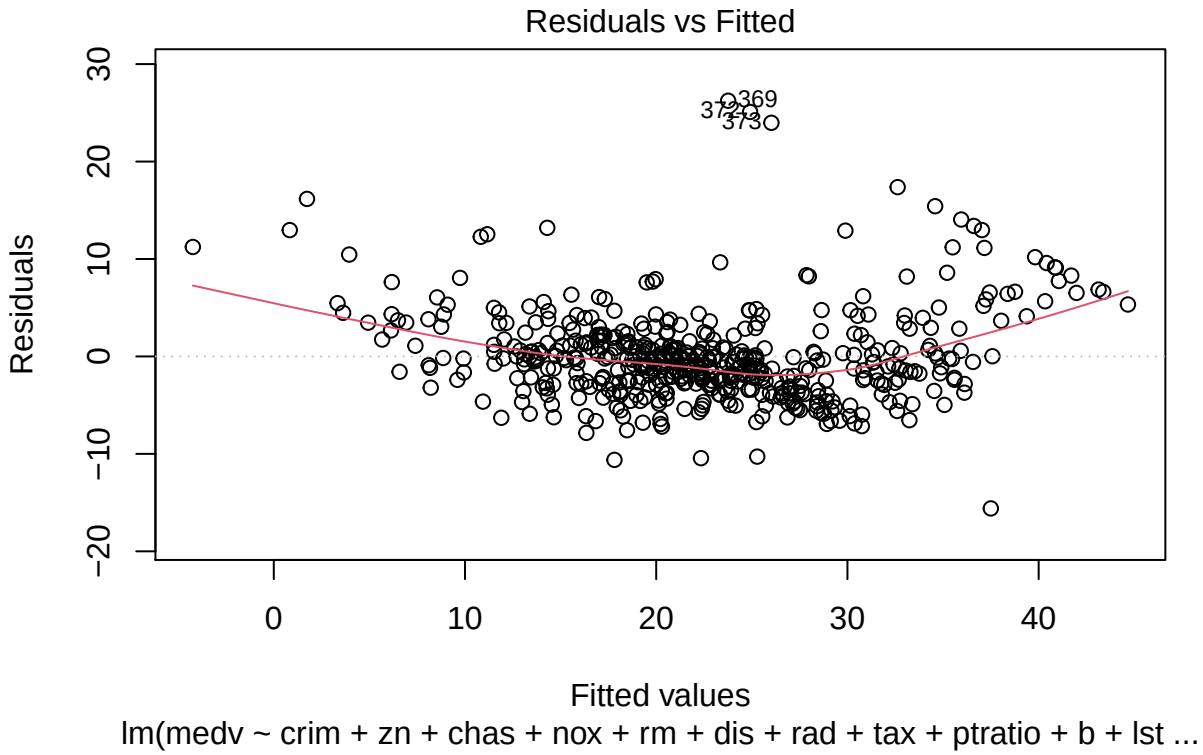
using this data, we can see that all of our covariates are significant. now we need to do residual diagnostics.

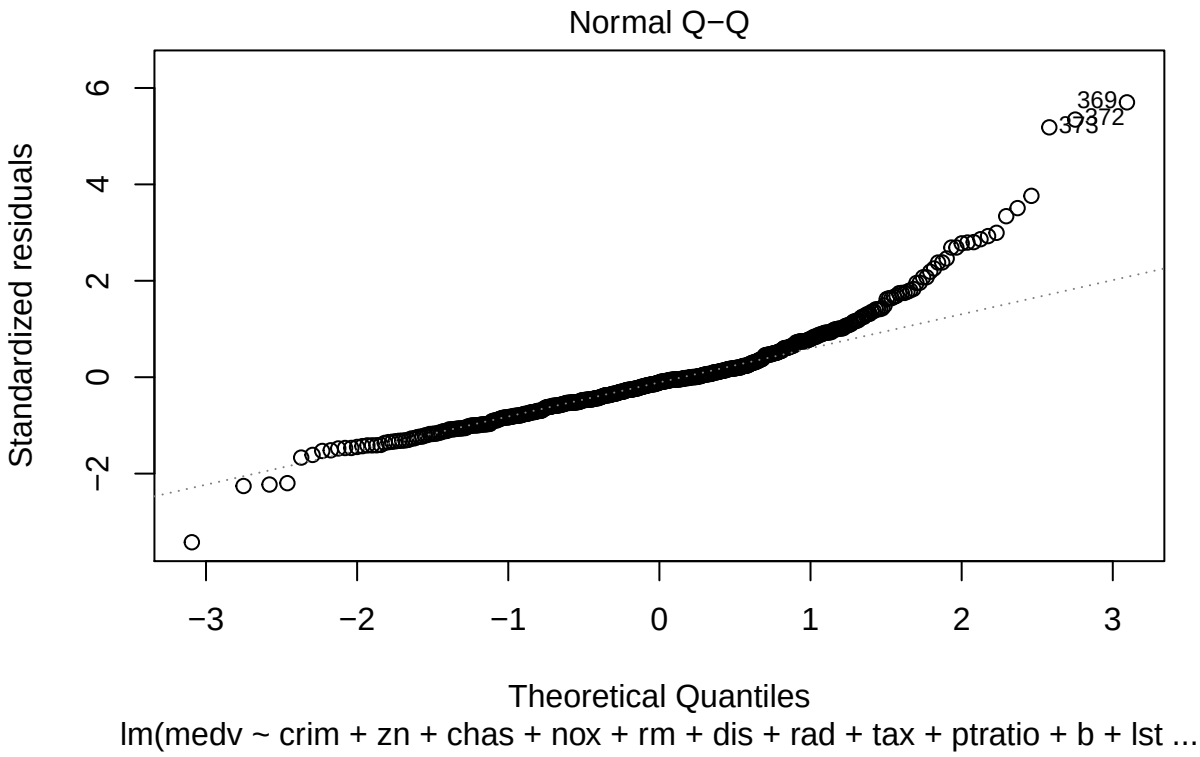
Our data suggests that there is an intercept of 36.34 for the median value of houses. There is a positive effect of the proportion of residential zoned land. The most important factor of median property value is the nitric oxides concentration. This has the largest effect size of all the covariates.

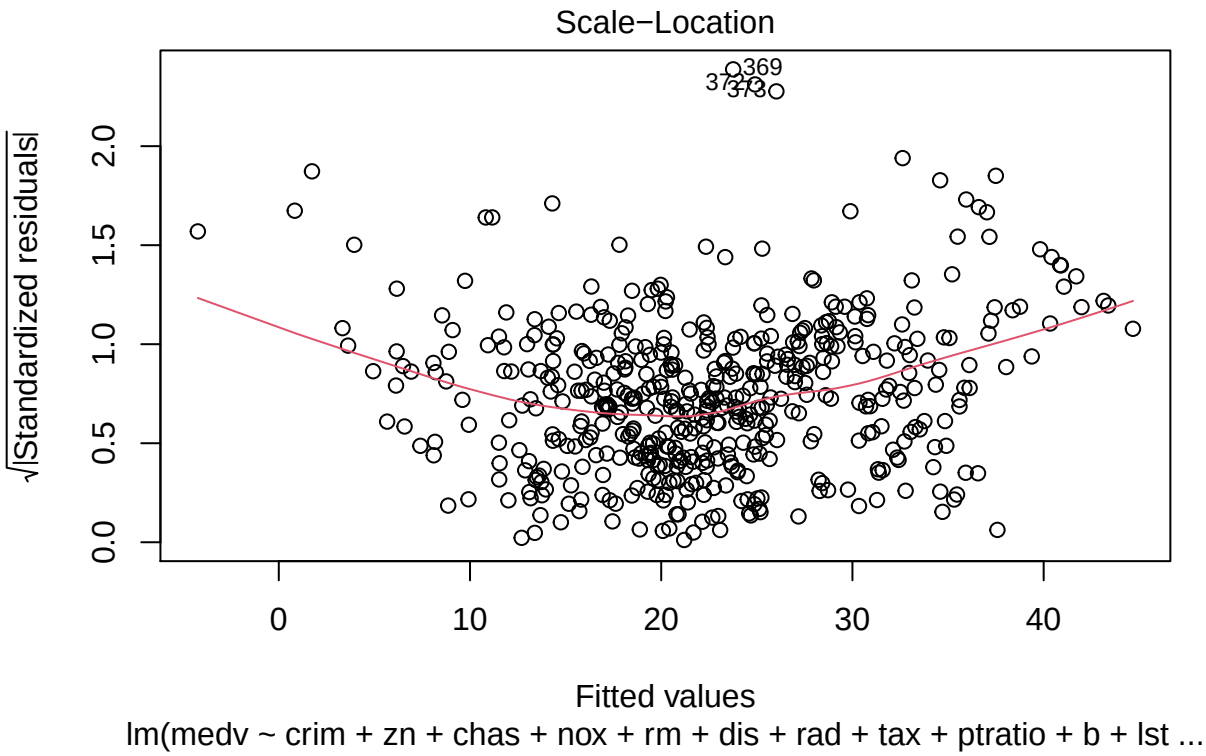
The average number of rooms per dwelling and the chas are also both strongly influential factors. Everything else is pretty insignificant in terms of effect size. the are statistically significant, but don't have large effects

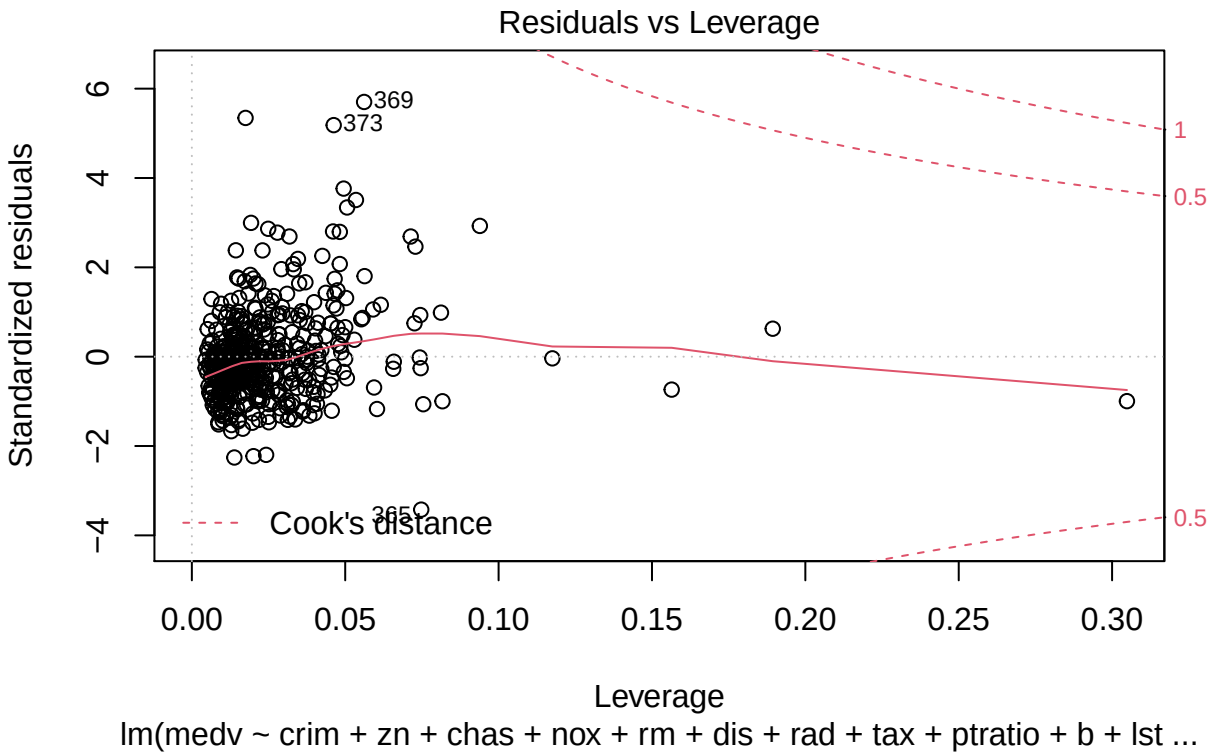
The ptratio is also very strongly correlated with median house value. There are less teachers per student as house value decreases. Every 10% change in the per capita crime rate the there is a negative effect on the property value.

```
plot(final)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(medv ~ crim + zn + chas + nox + rm + dis + rad + tax + ptratio + b + lst ...
369
372
373
−2
0
2
4
6
−3
−2
−1
0
1
2
3


Scale−Location
369
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0
10
20
30
40
Fitted values
lm(medv ~ crim + zn + chas + nox + rm + dis + rad + tax + ptratio + b + lst ...



The residuals vs fitted values plot shows a pretty random scatter. It is close to zero and seems to be pretty flat, so we can probably do better.

Our qq plot is very good for the lower quantities. We do a very good job for our median data. Only our tails have large deviations.

The scale-location looks very random so that is a good thing

In this particular data set the cooks distance is very consistent and nothing is dramatically influential.

Problem 7

The data in Table 6.8 were collected to test two psychological models of numerical cognition. Does the process of numbers depend on the way numbers are presented? Thirty-two subjects were required to make a series of quick numerical judgements represented as two number words or two single Arabic digits. The subjects were asked to respond "same" if the two numbers had the same numerical parity and "different" if the two numbers had a different parity. Half of the subjects were assigned a block of Arabic digit trials, followed by a block of number word trials, and half of the subjects received the blocks of trials in the reverse order. Within each block, the order of "same" and "different" parity trials was randomized for each subject. For each of the four combinations of parity and format, the median reaction times for correct responses were recorded for each subject. Here X_1 = Median reaction time for word format different parity combination X_2 = Median reaction time for word format-same parity combination X_3 = Median reaction time for Arabic format-different parity combinations X_4 = Median reaction time for Arabic format-same parity combinations

A) test for treatment effects using a repeated measures design. We need to test the effects of the different variables. we have four variables with two effects

There are three measures we are testing. First we need to test that their is no effect on parity. that is The difference between the same parity and the different parity.

$\mu_1 + \mu_3 - (\mu_2 + \mu_4)$

We also need to check to see that the word format has no effect so $\mu_1 + \mu_2 - (\mu_3 + \mu_4)$

Now we need to test if there is an effect of the interaction. $\mu_1 + \mu_4 - (\mu_3 + \mu_2) = 0$

Summarized

$\mu_1 - \mu_2 + \mu_3 - \mu_4 = 0$
$\mu_1 + \mu_2 - \mu_3 - \mu_4 = 0$
$\mu_1 - \mu_2 - \mu_3 + \mu_4 = 0$

We will rewrite this in matrix/vector notation with the contrast matrix C

$$C = \begin{pmatrix} 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

$H_0 : C\mu = 0$

$Cx \sim N_4(C\mu, C\Sigma C^T)$ $T^2 = n(C\overline{x})^T(CSC^T)^{-1}(C\overline{x}) \leq \frac{(n-1)(q-1)}{(n-q+1)} F_{q-1,n-q+1}(\alpha)$

```
alpha = 0.05
n = dim(d6_8)[1]
q = dim(d6_8)[2]
xbar = colMeans(d6_8)
S = var(d6_8)
C<- c(1,-1,1,-1,1,1,-1,-1,1,-1,-1,1)
C<- matrix(C,nrow=3, byrow=TRUE)
Cx = C%*%xbar
CSC = C%*%S%*%t(C)
T2 = n* t(Cx) %*% solve(CSC) %*% Cx
reject <- (n-1)*(q-1)/(n-q+1)*qf(1-alpha,n-1,n-q+1)
```

Since $153.7275056 > \frac{(n-1)(q-1)}{(n-q+1)} F_{q-1,n-q+1}(\alpha) = 5.9268314$ we reject the null hypothesis.

B) construct 95% simultaneous confident intervals for the contrasts represeting the number format effect, the parity type effect and the interaction effect. Interpret the resulting intervals.

We should work at making $1 - \alpha$ confidence intervals for Cx.

$T^2 = n(C\overline{x})^T(CSC^T)^{-1}(C\overline{x}) \leq \frac{(n-1)(q-1)}{(n-q+1)} F_{q-1,n-q+1}(\alpha)$

Where c_i is a row of the contrast matrix

$c_i\mu : c_i^T$

```
cont_labels = c("Parity","Form","Interaction")
p = nrow(C)
for (i in 1:p){
  print(cont_labels[i])
  fstat = sqrt((n-1)*(q-1)/(n-q+1)*qf(1-alpha,n-1,n-q+1)* t(C[i,])%*%S%*%C[i,]/n)
  upper_bound = C[i,]%*%xbar + fstat
  lower_bound = C[i,]%*%xbar - fstat
  CI =c(lower_bound,upper_bound)
  names(CI) = c("Lower","Upper")
```

```
  print(CI)
}
```

```
## [1] "Parity"
##    Lower    Upper
## 146.1117 266.5445
## [1] "Form"
##    Lower    Upper
## 220.5598 393.2840
## [1] "Interaction"
##     Lower     Upper
## -65.47410  20.63035
```

C) The absence of interaction supports the M model of numerical cognition, while the presence of interaction supports the C and C model of numerical cognition. Which model is supported in this experiment?

The interaction effect size was -22.421875 which is inside of the confidence interval for the contrast. The M model is a reasonable population for the scores.

D) For each subject, construct three difference scores corresponding to the number format contrast, parity type contrast, and the interaction contrast. Is a multivariate normal distribution a reasonable population model for these data? Explain.

```
X <- data.matrix(d6_8,rownames.force = NA)
difference_scores <- as.data.frame(X%*%t(C))
names(difference_scores) <- c("Parity","Form","Interaction")
mvn(difference_scores,mvnTest="royston",univariateTest ="CVM")
```

```
## $multivariateNormality
##      Test        H    p value MVN
## 1 Royston 6.952016 0.07477815 YES
##
## $univariateNormality
##               Test    Variable Statistic   p value Normality
## 1 Cramer-von Mises   Parity       0.1347    0.0356    NO
## 2 Cramer-von Mises    Form        0.0514    0.4798    YES
## 3 Cramer-von Mises Interaction    0.0742    0.2384    YES
##
## $Descriptives
##              n      Mean  Std.Dev Median    Min   Max   25th    75th      Skew
## Parity      32 206.32812 139.9195 169.25  -34.5 607.0 121.25 283.375 0.8980471
## Form        32 306.92188 200.6721 276.75  -75.0 879.5 192.25 438.500 0.6150826
## Interaction 32 -22.42188 100.0367 -37.50 -217.0 229.5 -82.50  29.750 0.3086301
##               Kurtosis
## Parity       0.6024854
## Form         0.6810366
## Interaction -0.0977505
```

```
mvn(difference_scores, mvnTest="hz",univariateTest = "SW")
```

```
## $multivariateNormality
##            Test      HZ    p value MVN
## 1 Henze-Zirkler 1.02264 0.01106792  NO
##
## $univariateNormality
##           Test    Variable Statistic   p value Normality
## 1 Shapiro-Wilk   Parity       0.9348    0.0535    YES
## 2 Shapiro-Wilk    Form        0.9586    0.2518    YES
## 3 Shapiro-Wilk Interaction    0.9692    0.4763    YES
##
## $Descriptives
##              n      Mean  Std.Dev Median    Min   Max   25th    75th      Skew
## Parity      32 206.32812 139.9195 169.25  -34.5 607.0 121.25 283.375 0.8980471
## Form        32 306.92188 200.6721 276.75  -75.0 879.5 192.25 438.500 0.6150826
## Interaction 32 -22.42188 100.0367 -37.50 -217.0 229.5 -82.50  29.750 0.3086301
##               Kurtosis
## Parity       0.6024854
## Form         0.6810366
## Interaction -0.0977505
```

By the Royston multivariate normality test, the multivariate normal model is reasonable; however, the Heinze-Zirkler test says that a Multivariate normal test is not appropriate.

Problem 8

Consider the air-pollution data listed in Table 1.5. Your job is to summarize these data in fewer than p = 7 dimensions if possible. Conduct a principal component analysis of the data using both the covariance matrix S and the correlation matrix R. What have you learned? Does it make any difference which matrix is chosen for analysis? Can the data be summarized in three or fewer dimensions? Can you interpret the principal components?

```
names(d1_5) <- c("wind","solar","CO","NO","NO2","O3","HC")
#d1_5 <- scale(d1_5,center=TRUE,scale=TRUE)
S=var(d1_5)
R = cor(d1_5,method="pearson",use="complete.obs")
ev_S <- eigen(S)
ev_R <- eigen(R)
d1_5m <- data.matrix(d1_5)
p =ncol(d1_5)
```

Covariance

$cov(x) = \Sigma$

(λ_i, e_i) are the eigenvalue, eigenvector pairs of Σ

$y_i = e_i^T$

```
eig_vecS <-ev_S$vectors
eig_valS <-ev_S$values
pcaS <- data.frame(matrix(ncol = 7, nrow = 42))
var_S <- data.frame(matrix(ncol = 7, nrow = 1))
```

```
for (i in 1:ncol(eig_vecS)){
  pcaS[,i] <-d1_5m%*%eig_vecS[,i]
  var_S[,i] <- eig_valS[i]/p
}
var_S <- var_S/sum(var_S)

print(var_S)
```

```
##         X1         X2         X3          X4          X5          X6
## 1 0.872948 0.08112714 0.03289281 0.007242569 0.003671092 0.001516979
##             X7
## 1 0.0006014096
```

Correlation

```
eig_vecR <-ev_R$vectors
eig_valR <-ev_R$values
pcaR <- data.frame(matrix(ncol = 7, nrow = 42))
var_R <- data.frame(matrix(ncol = 7, nrow = 1))
for (i in 1:ncol(eig_vecR)){
  pcaR[,i] <-d1_5m%*%eig_vecR[,i]
  var_R[,i] <- eig_valR[i]/p
}
var_R <- var_R/sum(var_R)
print(var_R)
```

```
##          X1        X2        X3        X4         X5         X6         X7
## 1 0.3338261 0.1980001 0.1720094 0.1038695 0.09335379 0.07666983 0.02227128
```

When I scale the data, the correlation and covariance matrices are equivalent in principal component analysis. If i do not scale my matrices, I can reduce the data to three principal components. This makes sense if the data is similar in nature, but it would be hard to do meaningful analysis with this data. the correlation matrix is basically a re-scaled variance matrix anyways. The covariance matrix is gives us more information in pca and does reduces the dimensionality further than correlation when unscaled and un centered.