Statistical Methods - HW4

2023-02-09

Problem 1

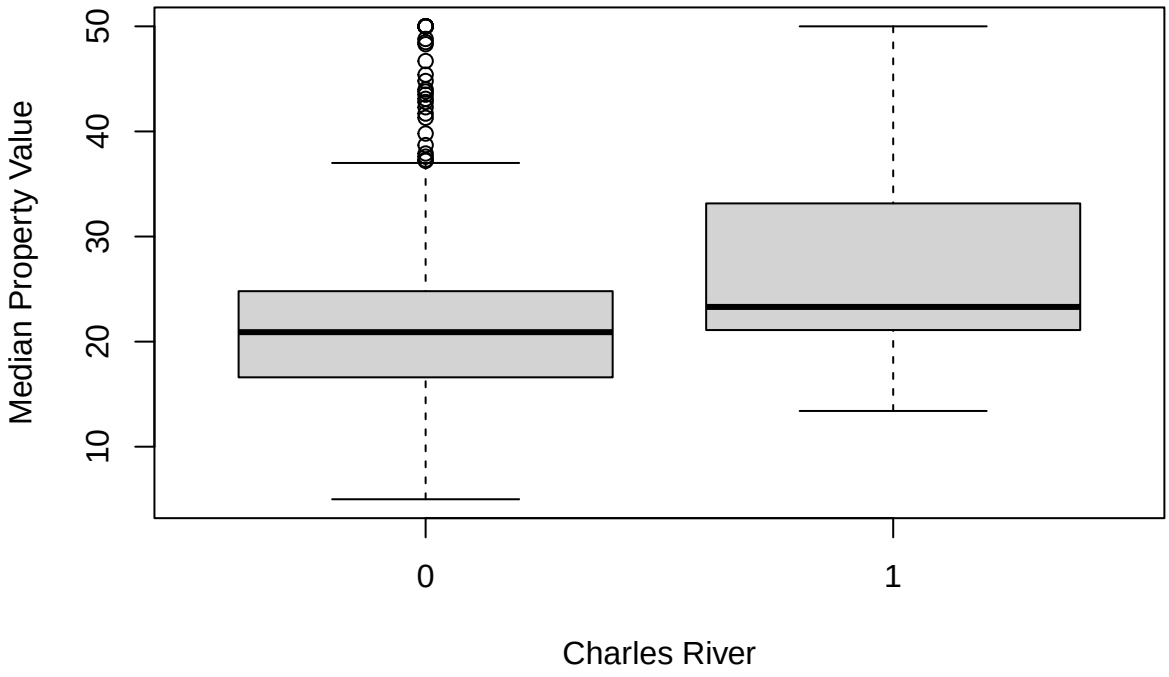
In this problem, you practice working with predictor variables that are discrete. Consider the Boston dataset in the package MASS. Take as response the median property value.

(a) Look at side-by-side boxplots for medv where the groups are defined by chas. Comment on what you observe. In particular, compare the different groups visually. Then fit a model explaining medv as a function of chas. Output an ANOVA table. What is the F-test testing? Is the result consistent with the boxplots?

The following code is to plot the side-by-side boxplots of median property value "medv" based on the variable "chas":

```
# Load data
library(MASS)
data(Boston)

# Convert chas to categorical variable
Boston$chas = as.factor(Boston$chas)
# Plot the box plot of medv based on chas
boxplot(medv ~ chas, data = Boston, main = "Median Property Value vs Charles River",
        xlab = "Charles River", ylab = "Median Property Value")
```




```
# Fit the model
model1 <- lm(medv ~ chas, data = Boston)
# Print summary
summary(model1)
```

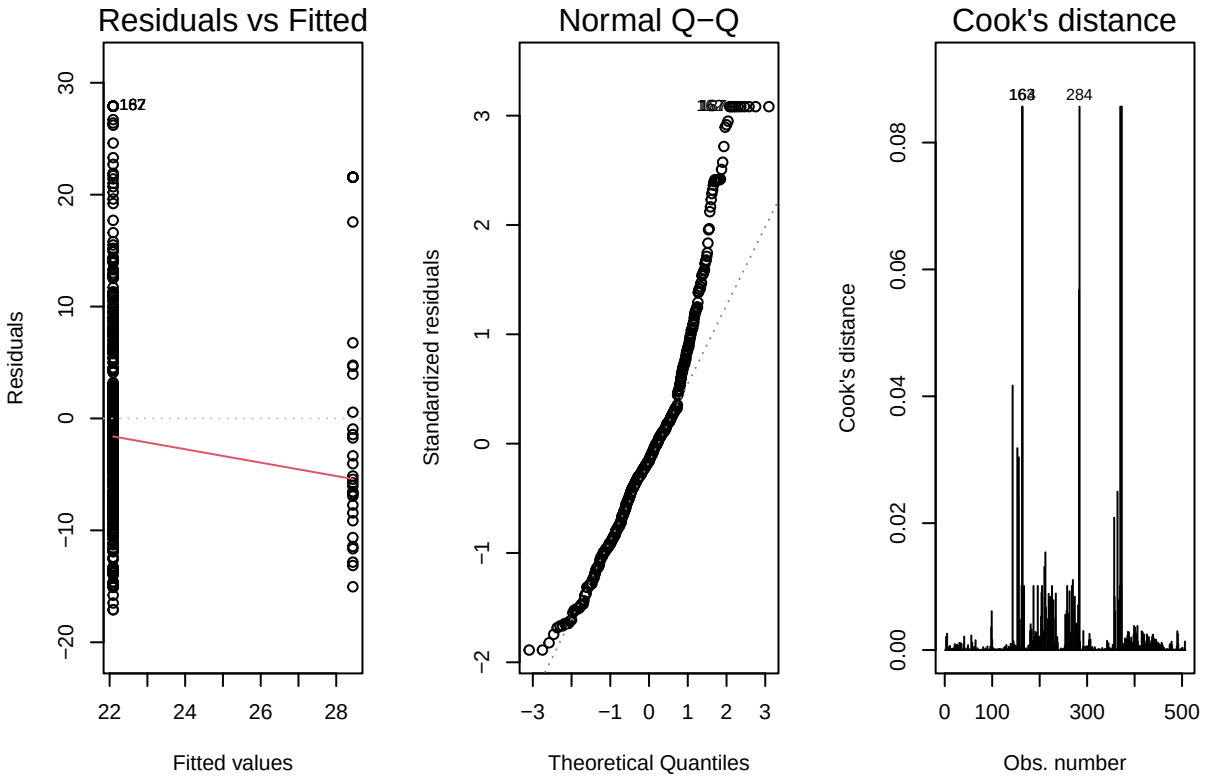
```
##
## Call:
## lm(formula = medv ~ chas, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.094  -5.894  -1.417   2.856  27.906
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.0938     0.4176  52.902  < 2e-16 ***
## chas1         6.3462     1.5880   3.996 7.39e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.064 on 504 degrees of freedom
## Multiple R-squared:  0.03072,    Adjusted R-squared:  0.02879
## F-statistic: 15.97 on 1 and 504 DF,  p-value: 7.391e-05
```

The above boxplot show that the median property value for properties that border the Charles River (chas = 1) is almost same to properties that don't (chas = 0). Due to this, the chas value seems to not have significant impact on the medv values.

We can also see that the variance of the medv value is lower for properties that border the Charles River (chas = 1) and higher for properties that don't (chas = 0).

In conclusion, chas = 0 has wider variability and large range. Also it has many outliers.

```
# Check whether the fitted model follows the general assumptions
par(mfrow=c(1,3))
plot(model1, which = c(1,2,4))
```



As seen in the above plots, it can be inferred that the general assumptions doesn't hold in this case i.e., the residuals doesn't have mean = 0, and are not normally distributed.

```
# ANOVA table
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: medv
##            Df Sum Sq Mean Sq F value    Pr(>F)
## chas        1   1312 1312.08  15.972 7.391e-05 ***
```

```
## Residuals 504  41404   82.15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The null hypothesis here is that the chas values has no significant impact on the median property value.

The F-test in this ANOVA table tests the null hypothesis that the means of the median property value is the same for areas that border the Charles River and areas that do not.

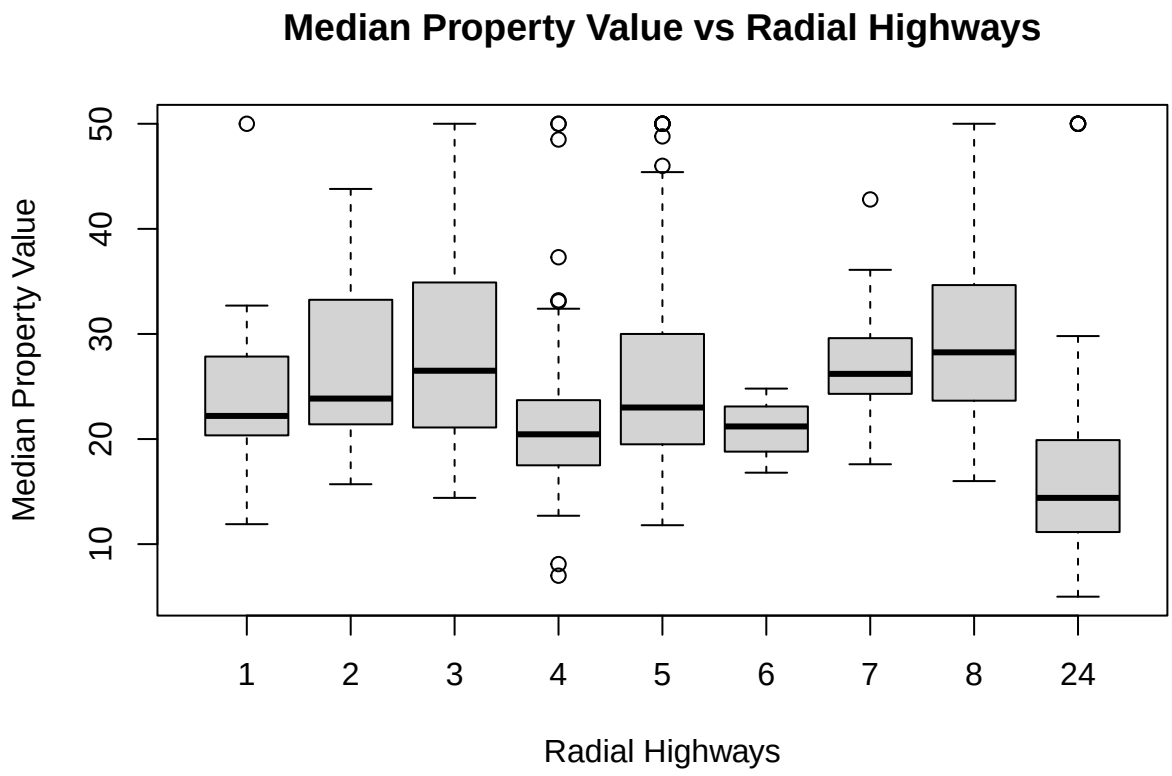
The F-value of this test (15.972) and low p-value (7.391e-05) indicate that this hypothesis can be rejected, and therefore we can conclude that there is a significant difference in the median property value for areas that border the Charles River (chas = 1) compared to areas that do not (chas = 0).

Eventhough the general assumptions doesn't hold, very low p-value of the order 10^-5 indicates that that the null hypothesis can be rejected.

Thus, chas can have a significant impact on the medv values which is different from the observations of the boxplot we made.

(b) Repeat with rad in place of chas.

```
# Convert rad to categorical variable
Boston$rad = as.factor(Boston$rad)
# Plot the box plot of medv based on rad
boxplot(medv ~ rad, data = Boston, main = "Median Property Value vs Radial Highways",
        xlab = "Radial Highways", ylab = "Median Property Value")
```



```r
# Fit the model
model2 <- lm(medv ~ rad, data = Boston)
# Print summary
summary(model2)
```

```
##
## Call:
## lm(formula = medv ~ rad, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.387  -5.280  -1.732   3.175  33.596
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   24.365      1.821  13.383  < 2e-16 ***
## rad2           2.468      2.465   1.001   0.3172
## rad3           3.564      2.249   1.584   0.1137
## rad4          -2.978      1.979  -1.504   0.1331
## rad5           1.342      1.973   0.680   0.4966
## rad6          -3.388      2.422  -1.399   0.1624
## rad7           2.741      2.686   1.020   0.3080
## rad8           5.993      2.465   2.431   0.0154 *
## rad24         -7.961      1.954  -4.075 5.36e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.142 on 497 degrees of freedom
## Multiple R-squared:  0.2287, Adjusted R-squared:  0.2162
## F-statistic: 18.42 on 8 and 497 DF,  p-value: < 2.2e-16
```

As seen in the above plot, there isn't much variance of medv value for rad = 1 ,4, 6, 7 and 24, which indicates that the median property value isn't much affected by these radial highways. Whereas, the other rad values have slightly different medv values indicating that they might impact the medv values.

Also we can see that, rad = 2, 3, 5, and 8 has wider variability and large range compared to the other rad values.

The median property values are almost similar for rad = 1, 2, 4, 5, and 6 and also for rad = 3, 7, and 8 which is slightly higher. Also, rad = 24 has very low median property value compared to all the others.

```r
# ANOVA table
anova(model2)
```

```
## Analysis of Variance Table
##
## Response: medv
##            Df Sum Sq Mean Sq F value    Pr(>F)
## rad         8   9767  1220.9  18.416 < 2.2e-16 ***
## Residuals 497  32949    66.3
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The null hypothesis here is that the rad values has no significant impact on the median property value.

The F-test in this ANOVA table tests the null hypothesis that the means of the median property value is the same for houses that have different rad values.

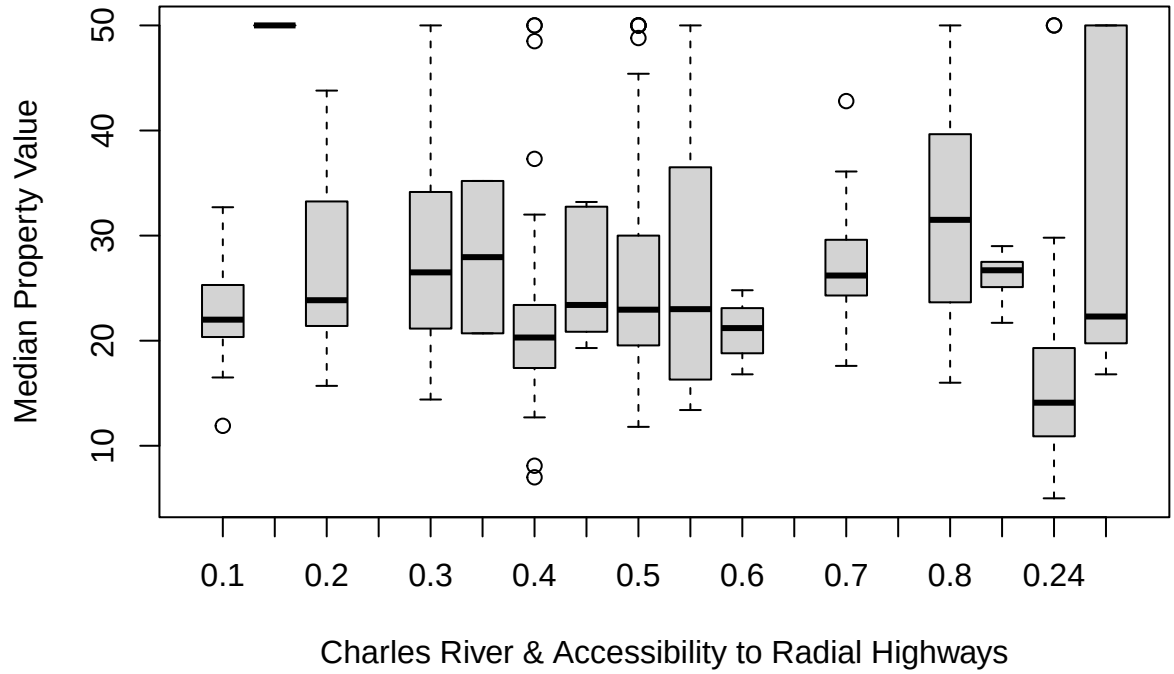
The F-value of this test (18.42) and low p-value (< 2.2e-16) indicate that this hypothesis can be rejected, and therefore we can conclude that there is a significant difference in the median property value for houses that have different rad values.

Eventhough the general assumptions doesn't hold, very low p-value of the order 10^-16 indicates that that the null hypothesis can be rejected.

Thus, rad can have a significant impact on the medv values.

(c) Produce a nice boxplot display of medv where the groups are defined by chas and rad jointly. Comment on what you observe. Then look at an interaction plot. Then fit a model explaining medv as a function of chas and rad with interactions. Output an ANOVA table. What are the different F-tests testing? Compare with the previous F-test as appropriate. Are the results of these tests consistent with the plots you just looked at?

```
# Plot the boxplot pf medv based on chas and rad jointly
boxplot(medv ~ chas + rad, data = Boston, main = "Median Property Value vs Charles River & Accessibilit
        xlab = "Charles River & Accessibility to Radial Highways", ylab = "Median Property Value")
```




```
# Fit the model explaining medv as a function of chas and rad without interactions
model_wi <- lm(medv ~ chas + rad, data = Boston)
# Print summary
```

```
summary(model_wi)
```

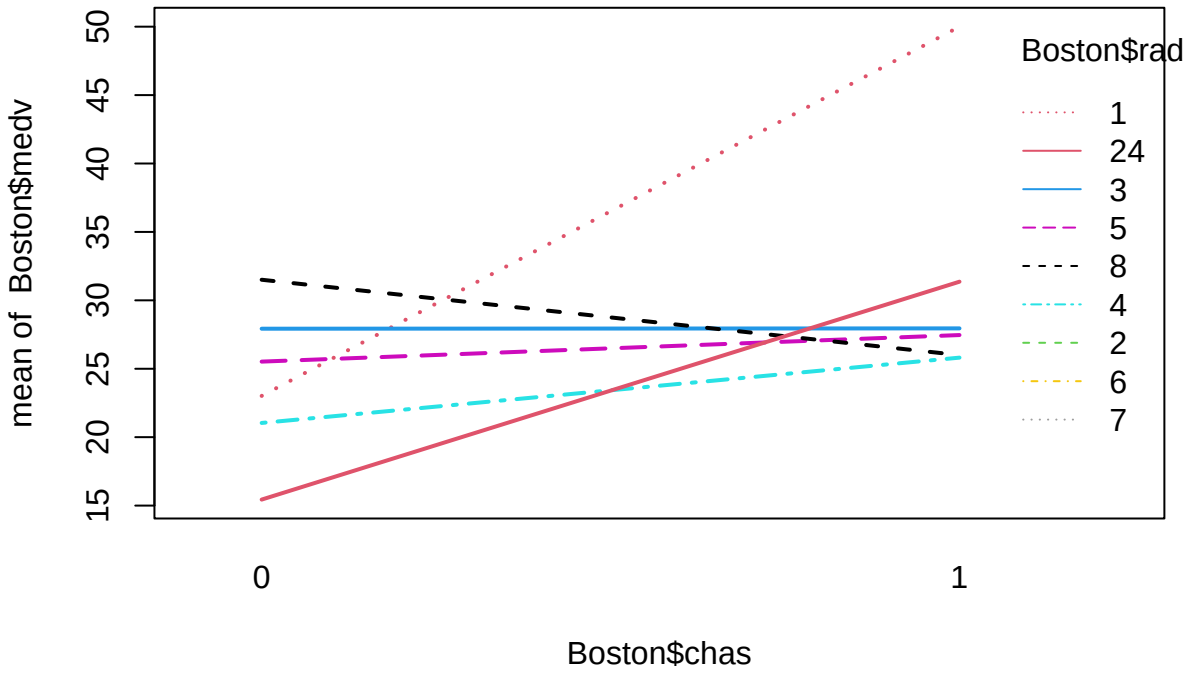
```
##
## Call:
## lm(formula = medv ~ chas + rad, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.396  -5.163  -1.573   3.372  33.937
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   24.084      1.796  13.410  < 2e-16 ***
## chas1          5.627      1.426   3.947 9.07e-05 ***
## rad2           2.750      2.431   1.131    0.259
## rad3           3.549      2.217   1.601    0.110
## rad4          -3.106      1.951  -1.592    0.112
## rad5           1.085      1.945   0.558    0.577
## rad6          -3.107      2.388  -1.301    0.194
## rad7           3.022      2.648   1.141    0.254
## rad8           5.102      2.440   2.091    0.037 *
## rad24         -8.021      1.926  -4.165 3.67e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.025 on 496 degrees of freedom
## Multiple R-squared:  0.2521, Adjusted R-squared:  0.2386
## F-statistic: 18.58 on 9 and 496 DF,  p-value: < 2.2e-16
```

The boxplot above doesn't give a legible intuition on whether the chas and rad values jointly have an impact on the medv values or not.

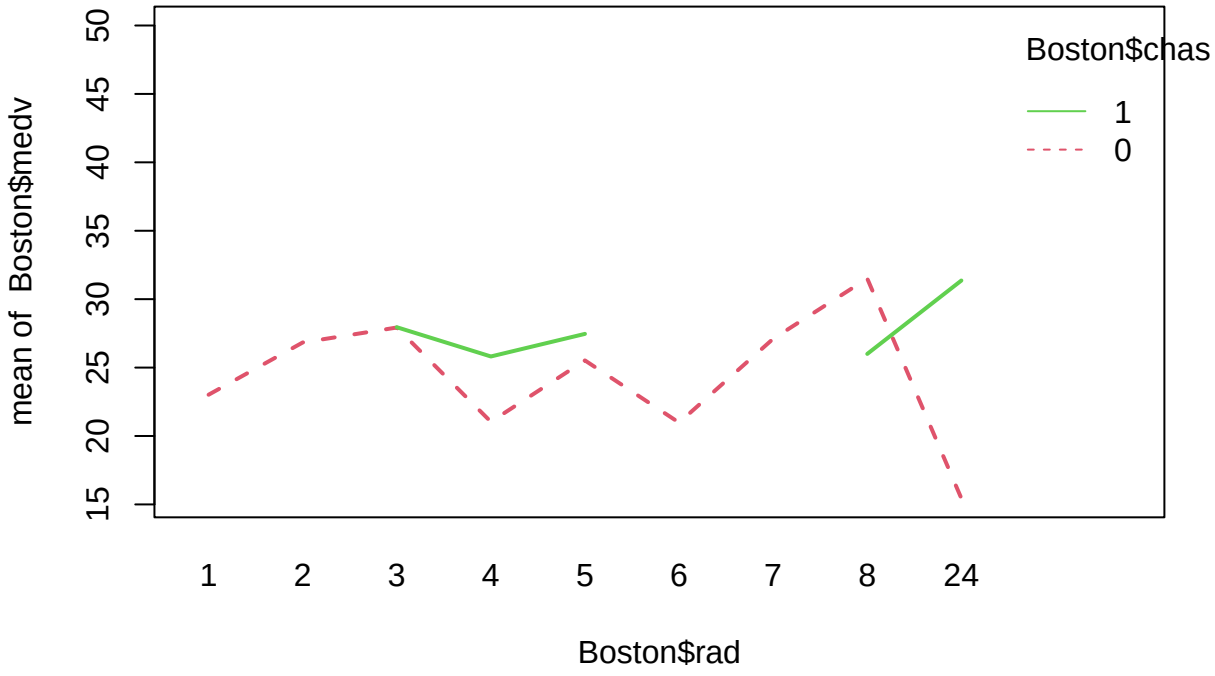
As seen in the above plot, for rad = 3 and rad = 5, the median medv values are almost similar and doesn't make any difference with respect to the bordering of Charles River. Whereas, for the other rad values, we can't conclude about the Charles River bordering's impact on Accessibility to Radial Highways since there isn't much similarity.

We can create an interaction plot to see if there is an interaction between chas and rad in explaining medv. We can use the interaction.plot function in the graphics library in R to create this plot.

```
library(graphics)
# Plot the interaction plot of chas, rad and medv
interaction.plot(Boston$chas, Boston$rad, Boston$medv, col=2:10, lwd=2, cex.axis=1, cex.lab=1)
```




```
interaction.plot(Boston$rad, Boston$chas, Boston$medv, col=2:3, lwd=2, cex.axis=1, cex.lab=1)
```



From the above plots, it can be the case where there might be a slight interaction for lines where values of rad equals to 3, 5, 8, 4 with values of chas being either 0 or 1.

However, we can't conclude much about the interactions between chas and rad from the above interaction plots, as most of the lines are not parallel or similar to any of the other lines.

```
# Fit the model explaining medv as a function of chas and rad with interactions
model_i <- lm(medv ~ chas * rad, data = Boston)
# Print summary
summary(model_i)
```

```
##
## Call:
## lm(formula = medv ~ chas * rad, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.505  -5.089  -1.339   3.512  34.561
##
## Coefficients: (3 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   23.016      1.794  12.829  < 2e-16 ***
## chas1         26.984      8.023   3.363 0.000830 ***
## rad2           3.818      2.401   1.590 0.112534
## rad3           4.912      2.217   2.215 0.027209 *
## rad4          -1.976      1.954  -1.011 0.312483
## rad5           2.505      1.951   1.284 0.199702
## rad6          -2.039      2.360  -0.864 0.388087
## rad7           4.090      2.611   1.567 0.117833
## rad8           8.489      2.537   3.346 0.000882 ***
## rad24         -7.577      1.927  -3.933  9.6e-05 ***
## chas1:rad2        NA         NA      NA       NA
## chas1:rad3   -26.962      9.831  -2.743 0.006318 **
## chas1:rad4   -22.212      8.521  -2.607 0.009422 **
## chas1:rad5   -25.042      8.397  -2.982 0.003005 **
## chas1:rad6        NA         NA      NA       NA
## chas1:rad7        NA         NA      NA       NA
## chas1:rad8   -32.489      8.934  -3.637 0.000305 ***
## chas1:rad24  -11.060      8.515  -1.299 0.194581
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.82 on 491 degrees of freedom
## Multiple R-squared:  0.2971, Adjusted R-squared:  0.2771
## F-statistic: 14.82 on 14 and 491 DF,  p-value: < 2.2e-16
```

```
# ANOVA table
anova(model_i)
```

```
## Analysis of Variance Table
##
## Response: medv
##            Df  Sum Sq Mean Sq F value    Pr(>F)
```

```
## chas        1  1312.1 1312.08 21.4563 4.642e-06 ***
## rad         8  9458.3 1182.29 19.3339 < 2.2e-16 ***
## chas:rad    5  1920.6  384.13  6.2816 1.156e-05 ***
## Residuals 491 30025.2   61.15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The above model is fit to explain medv as a function of chas and rad with interactions.

The F-test above is testing if the model has an interaction between chas and rad variables. The F-value of 6.28 and a low p-value of 1.156e-05 indicates that there is an interaction between chas and rad variables.

```
# ANOVA table to compare the F-tests
anova(model_i, model_wi)
```

```
## Analysis of Variance Table
##
## Model 1: medv ~ chas * rad
## Model 2: medv ~ chas + rad
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    491 30025
## 2    496 31946 -5   -1920.6 6.2816 1.156e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

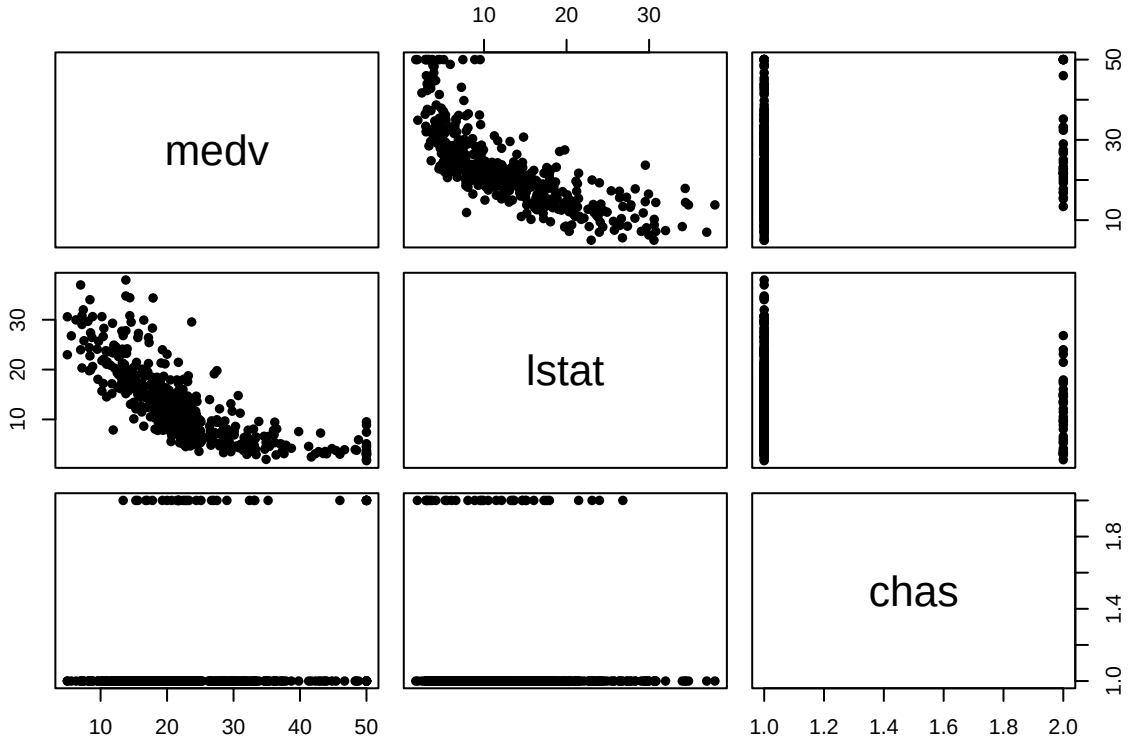
The difference is F-test values between the previously observed models and this model with interactions indicates that there is a significant impact of the interactions on the medv values.

As seen from the previous interaction plot, we were not able to conclude about the interaction but the results of these tests imply that there is an interaction.

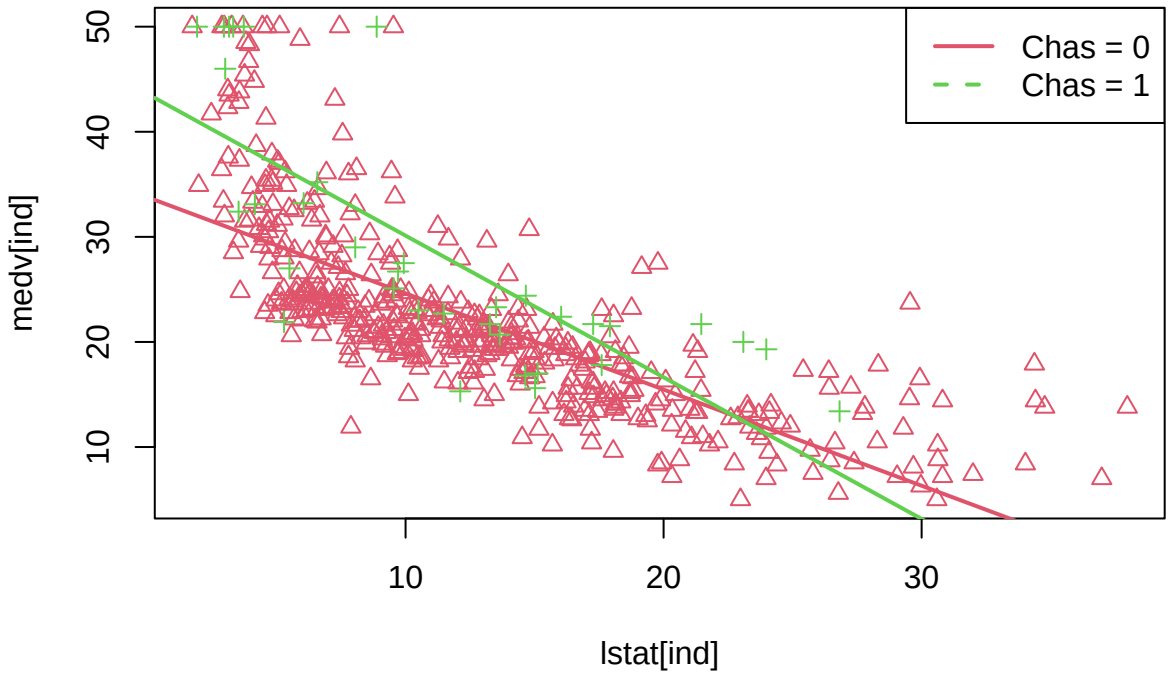
(d) It makes sense that median property value decreases with the percentage of lower status population lstat, and this is indeed what is observed here. Does the rate of decrease depend on whether the area borders the Charles River? Produce a plot that helps answer that question. Then formulate that into a hypothesis testing problem and perform an appropriate test.

To see if the rate of decrease of median property value decreases with the percentage of lower status population lstat depends on whether the area borders the Charles River or not, we can produce a scatter plot of medv against lstat colored by the variable chas. This will help us visualize if the relationship between medv and lstat changes based on the value of chas.

```
pairs(medv ~ lstat + chas, data = Boston, pch = 16)
```




```
ind = (Boston$chas==0)
plot(medv[ind] ~ lstat[ind], data = Boston, col=2, pch=2)
fit1 = lm(medv[ind] ~ lstat[ind], data = Boston)
abline(fit1, col=2, lwd=2)
ind = (Boston$chas==1)
points(medv[ind] ~ lstat[ind], data = Boston, col=3, pch=3)
fit2 = lm(medv[ind] ~ lstat[ind], data = Boston)
abline(fit2, col=3, lwd=2)
legend("topright", c("Chas = 0", "Chas = 1"),
       col = c(2, 3),
       lwd = 2, lty = 1:2)
```




```
summary(fit1)
```

```
##
## Call:
## lm(formula = medv[ind] ~ lstat[ind], data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.657  -3.667  -1.221   1.557  24.953
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.76716    0.55719   60.60   <2e-16 ***
## lstat[ind]  -0.91498    0.03808  -24.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.919 on 469 degrees of freedom
## Multiple R-squared:  0.5517, Adjusted R-squared:  0.5508
## F-statistic: 577.2 on 1 and 469 DF,  p-value: < 2.2e-16
```

```
summary(fit2)
```

```
##
## Call:
## lm(formula = medv[ind] ~ lstat[ind], data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.562  -5.526  -2.083   6.749  18.377
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   43.592      2.593  16.814  < 2e-16 ***
## lstat[ind]    -1.348      0.199  -6.775 1.01e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.757 on 33 degrees of freedom
## Multiple R-squared:  0.5817, Adjusted R-squared:  0.5691
## F-statistic:  45.9 on 1 and 33 DF,  p-value: 1.006e-07
```

From the plot, we can see that the relationship between medv and lstat is similar i.e., medv is inversely proportional to lstat for both groups (chas = 0 and chas = 1), but the slope is steeper(i.e., rate of decrease is higher) for the group chas = 1 which borders the Charles River.

We can perform a hypothesis test to determine if the rate of decrease of median property value decreasing with the percentage of lower status population lstat are significantly different. The null hypothesis is that the rate of decrease of median property value decreasing with the percentage of lower status population lstat is same for both chas values.

```
model4 <- lm(medv ~ lstat * chas, data = Boston)
summary(model4)
```

```
##
## Call:
## lm(formula = medv ~ lstat * chas, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.657  -3.787  -1.289   1.644  24.953
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.76716    0.57018  59.222  < 2e-16 ***
## lstat       -0.91498    0.03897 -23.478  < 2e-16 ***
## chas1        9.82513    2.10320   4.672 3.84e-06 ***
## lstat:chas1 -0.43288    0.16017  -2.703  0.00711 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.057 on 502 degrees of freedom
## Multiple R-squared:  0.5688, Adjusted R-squared:  0.5663
## F-statistic: 220.8 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
anova(model4)
```

```
## Analysis of Variance Table
##
## Response: medv
##             Df  Sum Sq Mean Sq  F value    Pr(>F)
## lstat        1 23243.9 23243.9 633.5325 < 2.2e-16 ***
## chas         1   786.3   786.3  21.4318 4.675e-06 ***
## lstat:chas   1   268.0   268.0   7.3044  0.007112 **
## Residuals  502 18418.1    36.7
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table of the model medv ~ lstat * chas includes an F-test for the interaction term lstat:chas. This F-test tests whether the interaction between lstat and chas is significantly different from zero, implying that the relationship between medv and lstat depends on chas.

The F-statistic for the interaction term lstat:chas is 7.3044 and the corresponding p-value is 0.007112, which is less than 0.05. Therefore, we reject the null hypothesis and conclude that the relationship between medv and lstat for the two groups (chas = 0 and chas = 1) are significantly different.

This result is consistent with what we observed from the scatter plot, as the rate of decrease of median property value decreasing with the percentage of lower status population lstat was seen to be higher for the group chas = 1 which borders the Charles River.

Problem 2

Consider the same dataset and turn to the problem of fitting a polynomial model explaining medv as a function of lstat.

(a) Fit a polynomial model of degree 3 by least squares.

To fit a polynomial model of degree 3 by least squares, we can use the lm function in R.

```r
# Fit polynomial model of degree 3 by least squares
fit.ls <- lm(medv ~ poly(lstat, 3), data = Boston)
summary(fit.ls)
```

```
##
## Call:
## lm(formula = medv ~ poly(lstat, 3), data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.5441  -3.7122  -0.5145   2.4846  26.4153
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       22.5328     0.2399  93.937  < 2e-16 ***
## poly(lstat, 3)1 -152.4595     5.3958 -28.255  < 2e-16 ***
## poly(lstat, 3)2   64.2272     5.3958  11.903  < 2e-16 ***
## poly(lstat, 3)3  -27.0511     5.3958  -5.013 7.43e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.396 on 502 degrees of freedom
## Multiple R-squared:  0.6578, Adjusted R-squared:  0.6558
## F-statistic: 321.7 on 3 and 502 DF,  p-value: < 2.2e-16
```

(b) Repeat with each robust method covered in the lecture notes/slides.

To fit a polynomial model using robust methods, we can use the rlm function from the MASS package.

We can also fit a polynomial model using the lmsreg and ltsreg function.

```
# Fit robust models
# L1 regression
require(quantreg)
```

```
## Loading required package: quantreg
```

```
## Loading required package: SparseM
```

```
##
## Attaching package: 'SparseM'
```

```
## The following object is masked from 'package:base':
##
##     backsolve
```

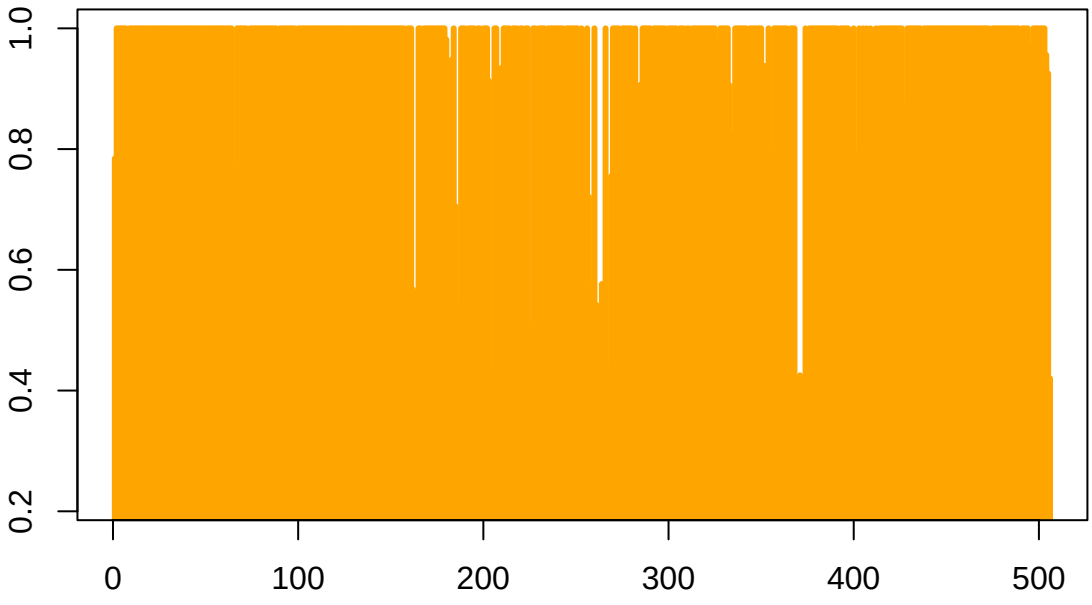
```
fit.l1 = rq(medv ~ poly(lstat,3),data = Boston)
summary(fit.l1)
```

```
##
## Call: rq(formula = medv ~ poly(lstat, 3), data = Boston)
##
## tau: [1] 0.5
##
## Coefficients:
##                 coefficients lower bd   upper bd
## (Intercept)       21.44583     21.10727   22.06917
## poly(lstat, 3)1 -137.62436   -152.23491 -123.42122
## poly(lstat, 3)2   44.03688     38.05664   65.28219
## poly(lstat, 3)3  -16.30155    -26.38196   -2.35653
```

```
# Fit a polynomial model of degree 3 using Huber weights
fit.huber <- rlm(medv ~ poly(lstat, 3), data = Boston, maxit=50)
summary(fit.huber)
```

```
##
## Call: rlm(formula = medv ~ poly(lstat, 3), data = Boston, maxit = 50)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.8373  -3.0157  -0.2219   2.7329  26.8409
##
## Coefficients:
##                 Value     Std. Error t value
## (Intercept)       21.9774    0.2037   107.8773
## poly(lstat, 3)1 -143.0378    4.5827   -31.2126
## poly(lstat, 3)2   55.9111    4.5827    12.2005
## poly(lstat, 3)3  -21.7848    4.5827    -4.7537
##
## Residual standard error: 4.323 on 502 degrees of freedom
```

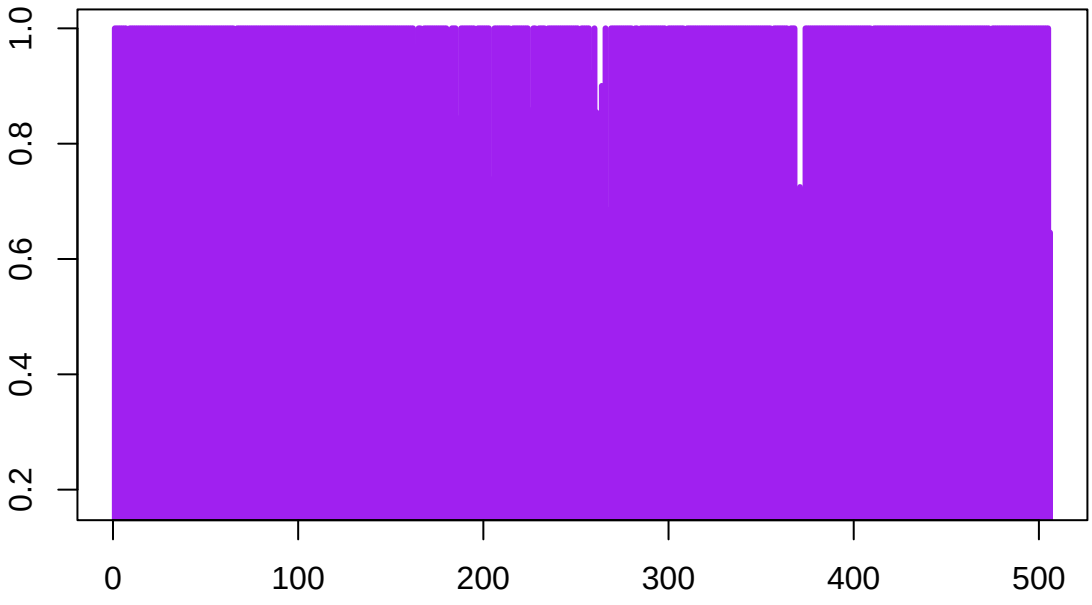
```
plot(fit.huber$w, ylab="", main="Huber weights", type='h', lwd=3, col="orange")
```




```
# Fit a polynomial model of degree 3 using Hampel weights
fit.hampel = rlm(medv ~ poly(lstat, 3), data = Boston, maxit=50, psi = psi.hampel)
summary(fit.hampel)
```

```
##
## Call: rlm(formula = medv ~ poly(lstat, 3), data = Boston, maxit = 50,
##     psi = psi.hampel)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.0819  -3.3423  -0.3086   2.6869  26.7536
##
## Coefficients:
##                  Value     Std. Error t value
## (Intercept)       22.2032    0.2161   102.7330
## poly(lstat, 3)1 -147.0125    4.8616   -30.2395
## poly(lstat, 3)2   60.7791    4.8616    12.5019
## poly(lstat, 3)3  -25.7454    4.8616    -5.2957
##
## Residual standard error: 4.547 on 502 degrees of freedom
```

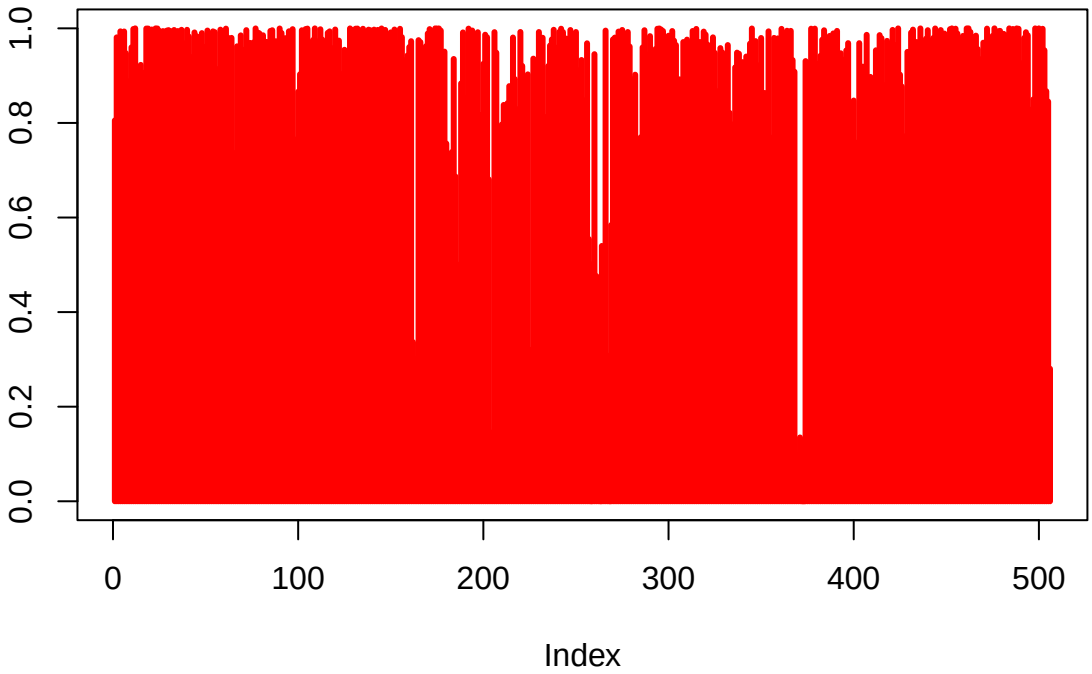
```
plot(fit.hampel$w, ylab="", main="Hampel weights", type='h', lwd=3, col="purple")
```




```
# Fit a polynomial model of degree 3 using Tukey weights
fit.tukey = rlm(medv ~ poly(lstat, 3), data = Boston, maxit=50, psi = psi.bisquare)
summary(fit.tukey)
```

```
##
## Call: rlm(formula = medv ~ poly(lstat, 3), data = Boston, maxit = 50,
##     psi = psi.bisquare)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.4084  -2.7037  -0.1377   2.9575  27.0035
##
## Coefficients:
##                  Value     Std. Error t value
## (Intercept)       21.6329    0.1962   110.2426
## poly(lstat, 3)1 -134.8470    4.4141   -30.5493
## poly(lstat, 3)2   48.3681    4.4141    10.9577
## poly(lstat, 3)3  -15.4625    4.4141    -3.5030
##
## Residual standard error: 4.172 on 502 degrees of freedom
```

```
plot(fit.tukey$w, ylab="", main="Tukey weights", type='h', lwd=3, col="red")
```




```
# Least median square model
fit.lms = lmsreg(medv ~ poly(lstat, 3), data = Boston)
fit.lms$coefficients
```

```
##     (Intercept) poly(lstat, 3)1 poly(lstat, 3)2 poly(lstat, 3)3
##        19.79435      -107.15055       -11.89300        23.39231
```

```
# Least median trimmed mean model
fit.lts = ltsreg(medv ~ poly(lstat, 3), data = Boston)
fit.lts$coefficients
```

```
##     (Intercept) poly(lstat, 3)1 poly(lstat, 3)2 poly(lstat, 3)3
##        19.43441      -105.33433       -11.85334        30.73339
```

The weights seen in the above plots indicate how much weight is assigned to each point in the model. The outliers have lower weights.

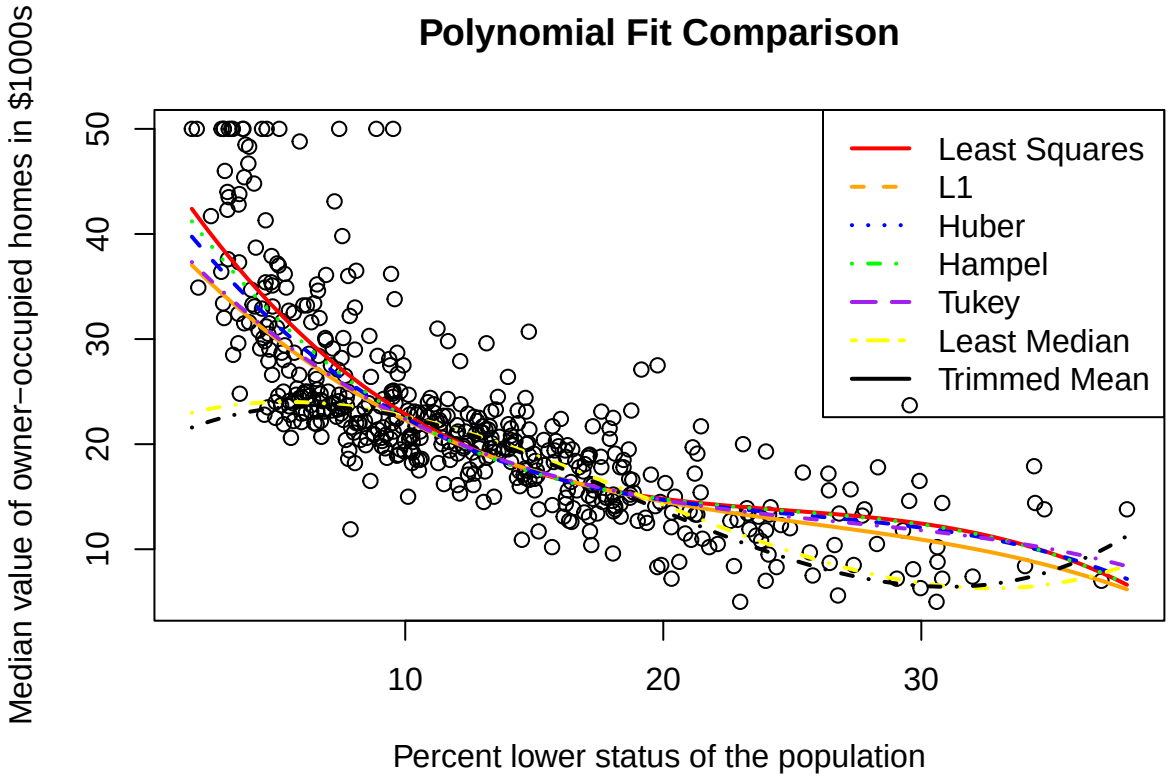
(c) Produce a scatterplot and overlay all these fits with different colors and a legend. [HINT: use the function predict.]

Finally, we plot a scatterplot of medv versus lstat and overlay all the fits using the predict() function with different colors and a legend.

```
# Create scatterplot with all fits
x <- seq(min(Boston$lstat), max(Boston$lstat), length.out = 100)
y_ls <- predict(fit.ls, newdata = data.frame(lstat = x))
y_l1 <- predict(fit.l1, newdata = data.frame(lstat = x))
y_huber <- predict(fit.huber, newdata = data.frame(lstat = x))
y_hampel <- predict(fit.hampel, newdata = data.frame(lstat = x))
y_tukey <- predict(fit.tukey, newdata = data.frame(lstat = x))
```

```r
y_lms <- predict(fit.lms, newdata = data.frame(lstat = x), raw = TRUE)
y_lts <- predict(fit.lts, newdata = data.frame(lstat = x), raw = TRUE)

plot(Boston$lstat, Boston$medv, main = "Polynomial Fit Comparison",
     xlab = "Percent lower status of the population",
     ylab = "Median value of owner-occupied homes in $1000s")
lines(x, y_ls, col = "red", lwd = 2, lty = 1)
lines(x, y_l1, col = "orange", lwd = 2, lty = 1)
lines(x, y_huber, col = "blue", lwd = 2, lty = 2)
lines(x, y_hampel, col = "green", lwd = 2, lty = 3)
lines(x, y_tukey, col = "purple", lwd = 2, lty = 4)
lines(x, y_lms, col = "yellow", lwd = 2, lty = 4)
lines(x, y_lts, col = "black", lwd = 2, lty = 4)
legend("topright", c("Least Squares", "L1", "Huber", "Hampel", "Tukey", "Least Median", "Trimmed Mean")
       col = c("red", "orange", "blue", "green", "purple", "yellow", "black"),
       lwd = 2, lty = 1:7)
```



The fit of different robust methods are shown in the above plot.

LMS and LTS are high breakdown point methods which are more robust to outliers.

Huber, Hampel and Tukey are low breakdown point methods which are less robust to outliers compared to the LMS and LTS methods.

Contribution Statement:

1. Swetha Arunraj (PID: A59019948):

- Coded Problem 1
- Added comments to the code of Problem 1
- Added description/comments to Problem 1
- Contributed to embed the codes in R Markdown and create the final PDF

2. Harin Raja Radha Krishnan (PID: A59019874):

- Coded Problem 2
- Added comments to the code of Problem 2
- Added description/comments to Problem 2
- Contributed to embed the codes in R Markdown and create the final PDF