Statistical Models : Homework 4

2023-02-06

Question 1

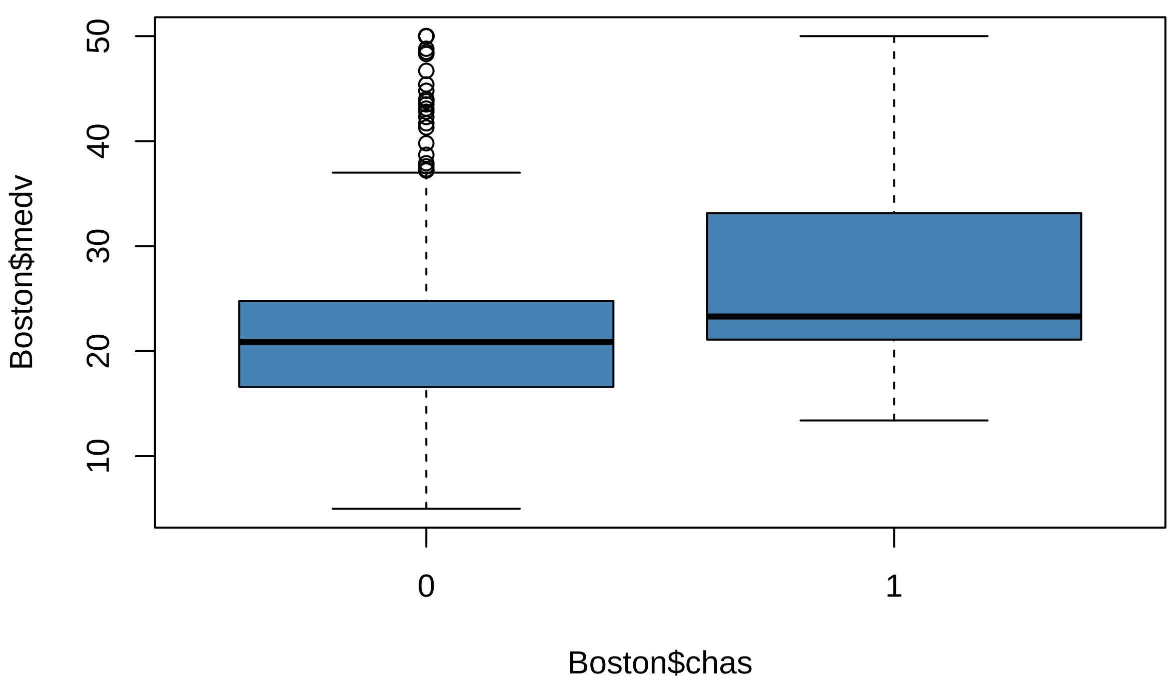
```
library(MASS)
library(ggplot2)
library(quantreg)
```

```
## Loading required package: SparseM

##
## Attaching package: 'SparseM'

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

```
boxplot(Boston$medv ~ Boston$chas,
        col='steelblue')
```



```r
#Fitting  a linear model
model_chas = lm(medv ~ chas, data = Boston)
summary(model_chas)
```

```
##
## 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 ***
## chas          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
```

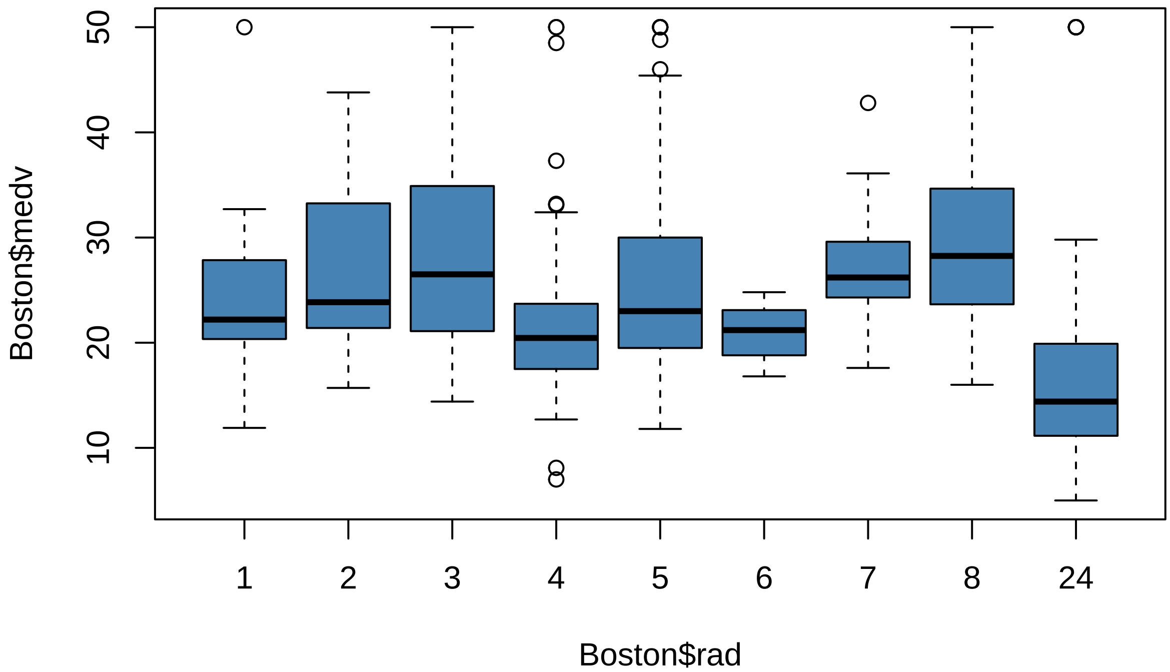
```r
anova(model_chas)
```

```
## 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
```

1(b)

Repeat with rad in place of chas.

```r
boxplot(Boston$medv ~ Boston$rad,
        col='steelblue')
```



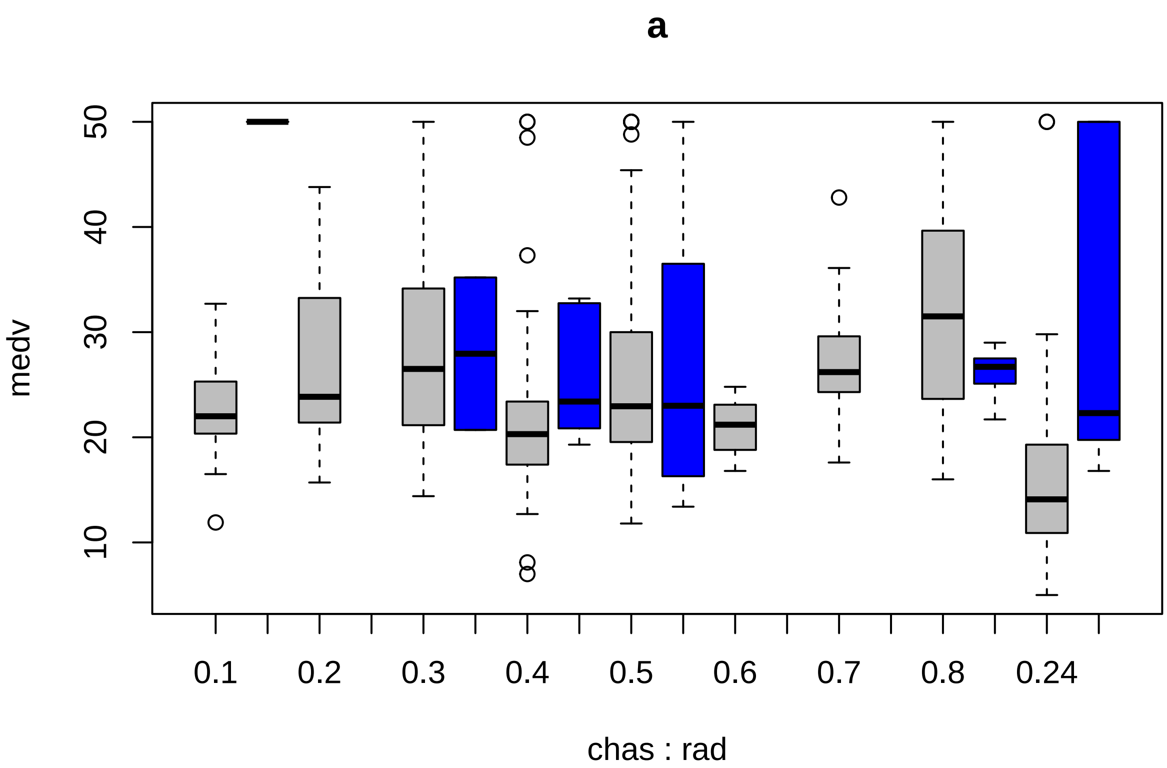

```r
model_rad = lm(medv ~ rad, data = Boston)
summary(model_rad)
```

```
##
## Call:
## lm(formula = medv ~ rad, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.770  -5.199  -1.967   3.321  33.292
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26.38213    0.56176  46.964   <2e-16 ***
## rad         -0.40310    0.04349  -9.269   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.509 on 504 degrees of freedom
## Multiple R-squared:  0.1456, Adjusted R-squared:  0.1439
## F-statistic: 85.91 on 1 and 504 DF,  p-value: < 2.2e-16
```

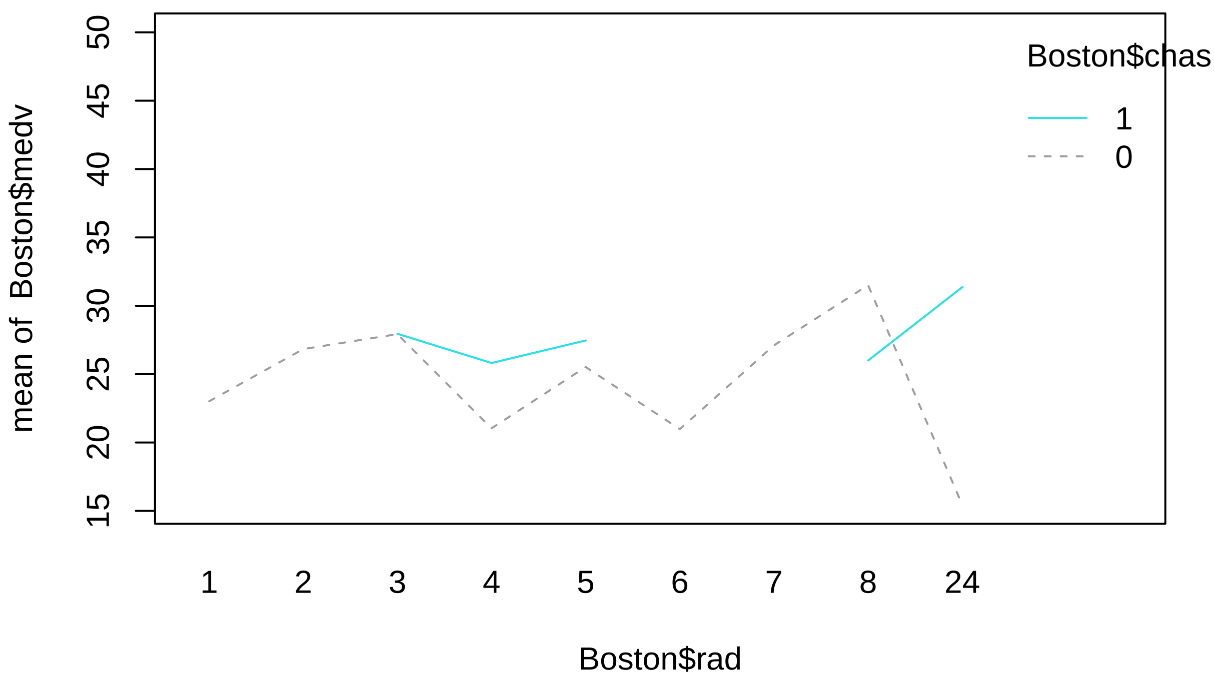
```r
anova(model_rad)
```

```
## Analysis of Variance Table
##
## Response: medv
##            Df Sum Sq Mean Sq F value    Pr(>F)
## rad         1   6221  6221.1  85.914 < 2.2e-16 ***
## Residuals 504  36495    72.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
boxplot(medv ~ chas * rad, data = Boston, main = "a", col=c("grey", "blue"))
```




```
interaction.plot(Boston$rad, Boston$chas, Boston$medv, col = Boston$medv)
```




```
model_combined= lm(medv ~ chas * rad, data = Boston)
summary(model_combined)
```

```
##
## Call:
## lm(formula = medv ~ chas * rad, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.527  -5.127  -1.796   3.548  34.216
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  26.2762     0.5662  46.409  < 2e-16 ***
## chas          0.7775     2.2042   0.353  0.72445
## rad          -0.4372     0.0437 -10.005  < 2e-16 ***
## chas:rad      0.5860     0.1777   3.297  0.00105 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.287 on 502 degrees of freedom
## Multiple R-squared:  0.1929, Adjusted R-squared:  0.188
## F-statistic: 39.98 on 3 and 502 DF,  p-value: < 2.2e-16
```

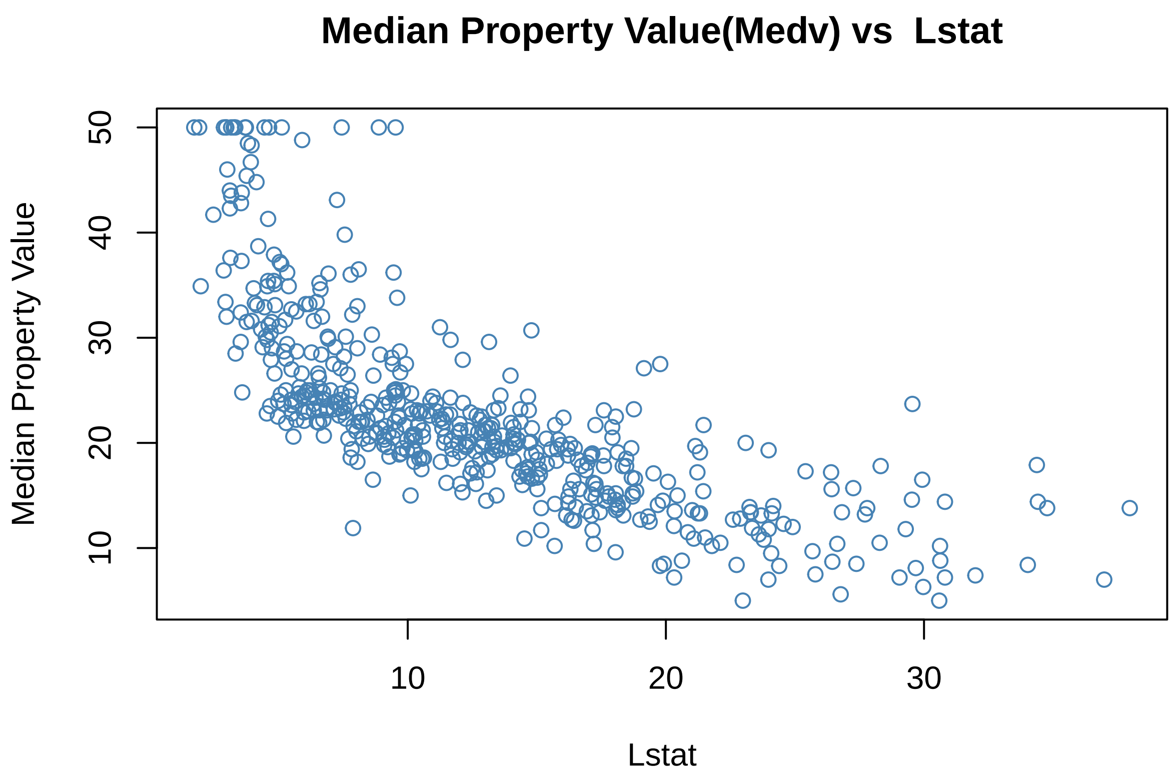
```
anova(model_combined)
```

```
## Analysis of Variance Table
##
## Response: medv
##            Df Sum Sq Mean Sq F value    Pr(>F)
## chas        1   1312  1312.1  19.104 1.505e-05 ***
## rad         1   6179  6179.4  89.972 < 2.2e-16 ***
## chas:rad    1    746   746.5  10.869  0.001047 **
## Residuals 502  34478    68.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

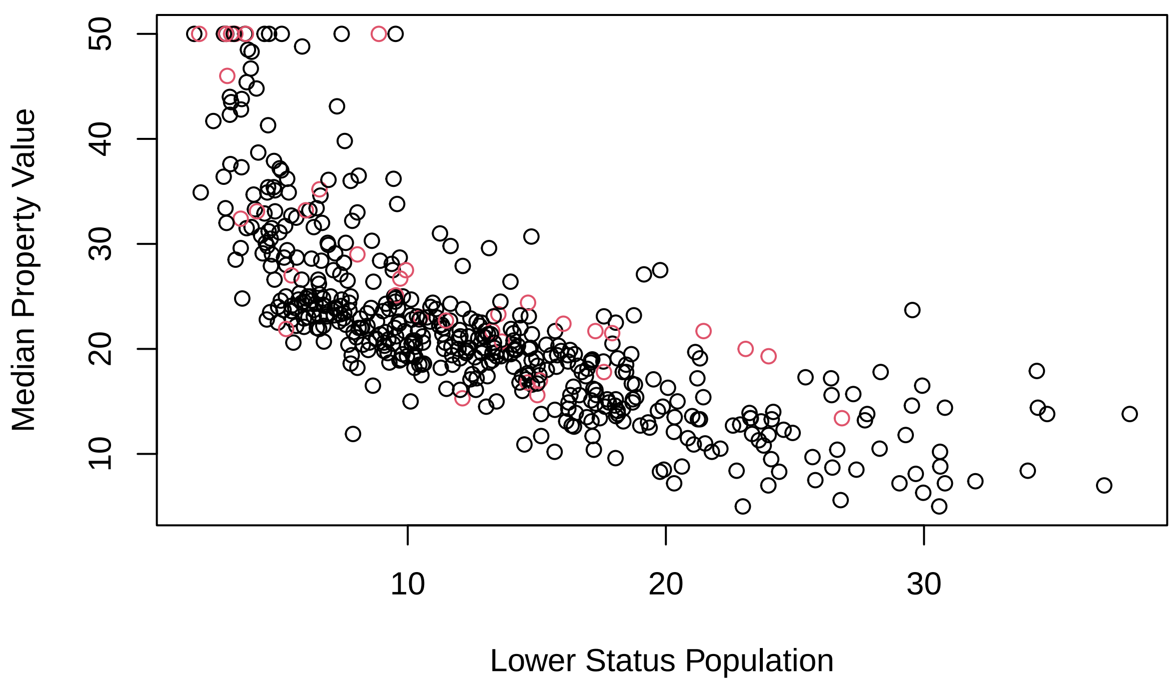
1(d)

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.

```
plot(Boston$medv ~ Boston$lstat,
        col='steelblue', main = "Median Property Value(Medv) vs  Lstat", xlab = "Lstat", ylab = "Median
```

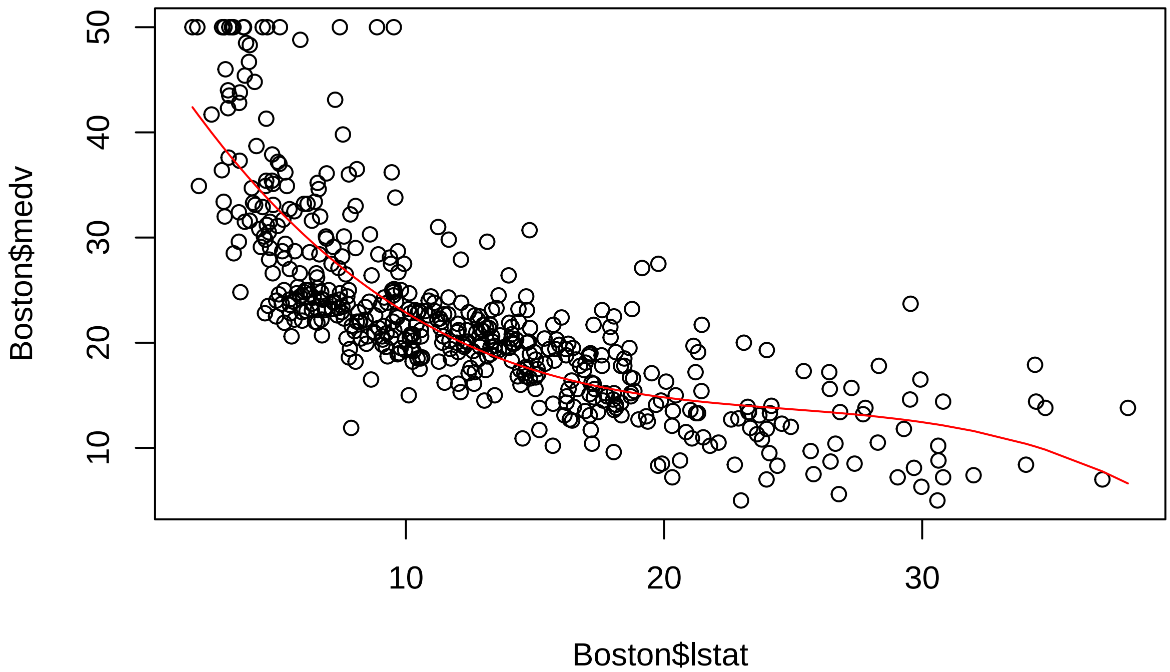



```
plot(Boston$lstat, Boston$medv, xlab = "Lower Status Population", ylab = "Median Property Value", col =
```



Question 2 Consider the same dataset and turn to the problem of fitting a polynomial model explaining medv as a function of lstat.

```
fit <- lm(medv ~ poly(lstat, 3), data = Boston)
plot(Boston$lstat, Boston$medv)
lines(sort(Boston$lstat), predict(fit, newdata = data.frame(lstat = sort(Boston$lstat))), col = "red")
```



Huber's M-estimation:

```
m.huber = rlm(medv ~ poly(lstat, 3),data = Boston, psi = psi.huber)
m.hampel = rlm(medv ~ poly(lstat, 3),data = Boston, psi = psi.hampel)
m.tukey = rlm(medv ~ poly(lstat, 3),data = Boston, psi = psi.bisquare)
fit.lms = lmsreg(medv ~ lstat, data = Boston)
fit.lts = ltsreg(medv ~ lstat, data = Boston)
```

c. Produce a scatterplot and overlay all these fits with different colors and a legend.

Team Contibutions :