HW4

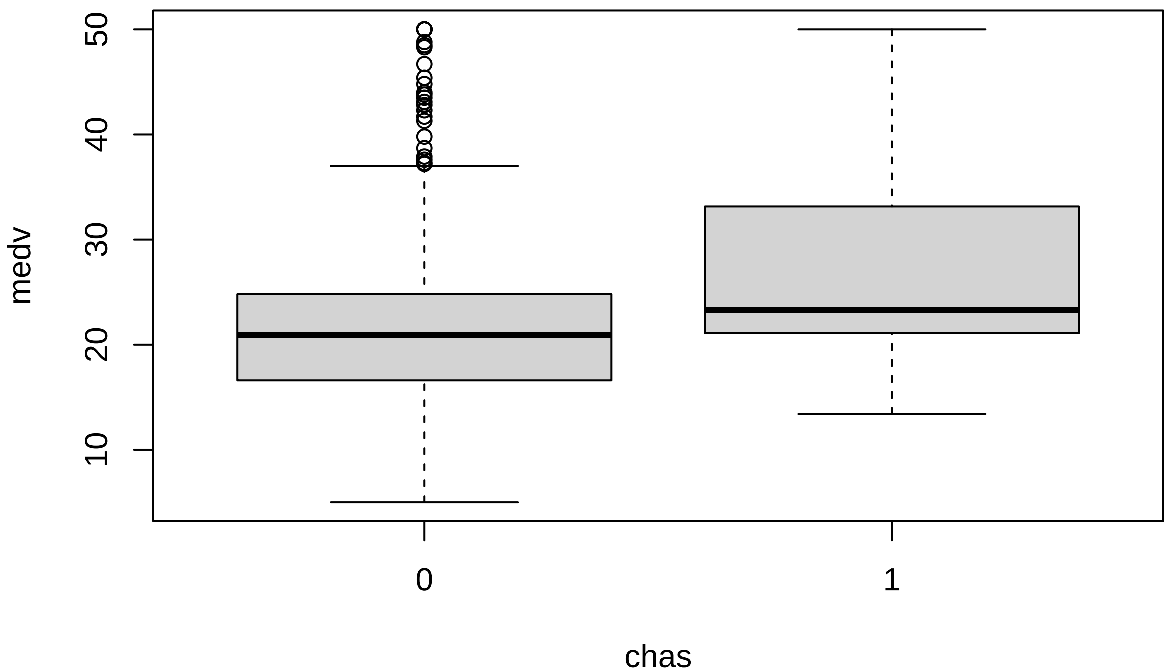
Sandra Villamar and Shobhit Dronamraju

2023-02-09

```
library(MASS)
data(Boston)
```

Problem 1A

```
boxplot(medv ~ chas, data=Boston)
```



From the boxplot, we see that for houses close to the Charles River (i.e. chas = 1),the overall distribution of median house values is higher than for houses not close to the Charles River (i.e. chas = 0). Although there are observations of median house values when chas = 0 that are just as high as when chas = 1, all these observations are outliers to the overall chas = 0 distribution.

```r
Boston$chas = as.factor(Boston$chas)
fit = lm(medv ~ chas, data=Boston)
anova(fit)
```

```
## 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
```

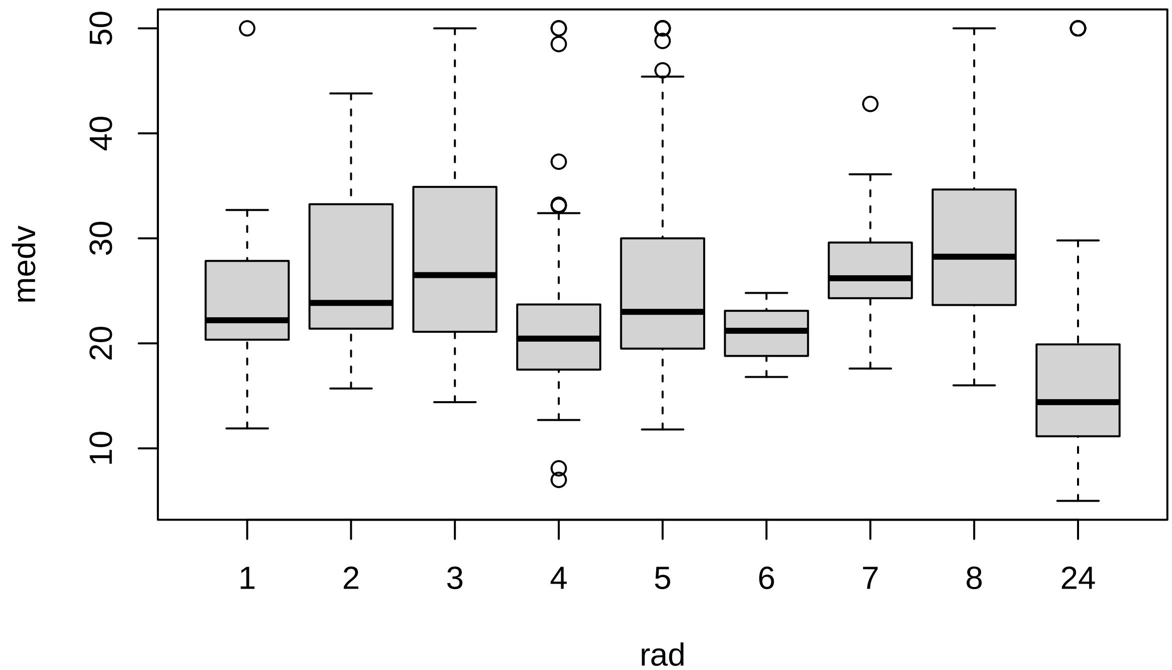
The F-test tests the null hypothesis that all coefficients of the predictor variables in the linear model are equal to 0. If the p-value associated with the F-test is less than a significance level, it means that at least one of the predictor variables is significant in explaining the variance of the response variable.

In this case, the F-test indicates that the predictor variable *chas* is significant in explaining the variance in the median property value *medv*. The p-value is 7.391e-05 which is quite small. This suggests that access to the Charles River is an important factor in determining the median property value in the Boston area.

The result of the F-test is consistent with the boxplots. We saw a different distribution of median house values when separating the observations into their respective *chas* category. This indicates that knowing *chas* helps in predicting median house value.

Problem 1B

```r
boxplot(medv ~ rad, data=Boston)
```



From the boxplot, we see that some values for *rad* produce about the same distribution of median house values such as rad = 3 and rad = 8. On the other hand, other values for *rad* produce very different distributions such as rad = 6 versus rad = 24. In particular, houses with rad = 24 produce a distribution much lower than the others which makes sense as easy highway access is normally seen as an advantage.

```
Boston$rad = as.factor(Boston$rad)
fit = lm(medv ~ rad, data=Boston)
anova(fit)
```

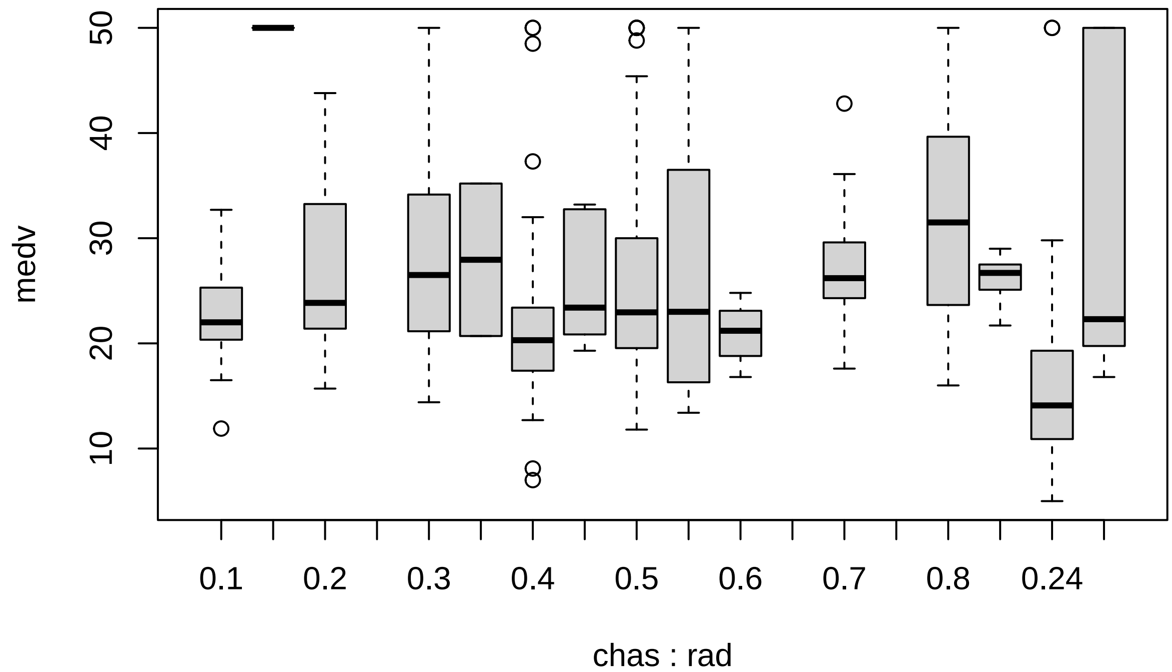
```
## Analysis of Variance Table
##
## Response: medv
##            Df Sum Sq Mean Sq F value    Pr(>F)
## rad         8   9767  1220.9  18.416 < 2.2e-16 ***
## Residuals 497  32949    66.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this case, the F-test indicates that the predictor variable *rad* is significant in explaining the variance in the median property value *medv*. The p-value is < 2.2e-16 which is quite small. This suggests that index of accessibility to radial highways is an important factor in determining the median property value in the Boston area.

The result of the F-test is consistent with the boxplots. We saw a different distribution of median house values when separating the observations into their respective *rad* category. This indicates that knowing *rad* helps in predicting median house value.

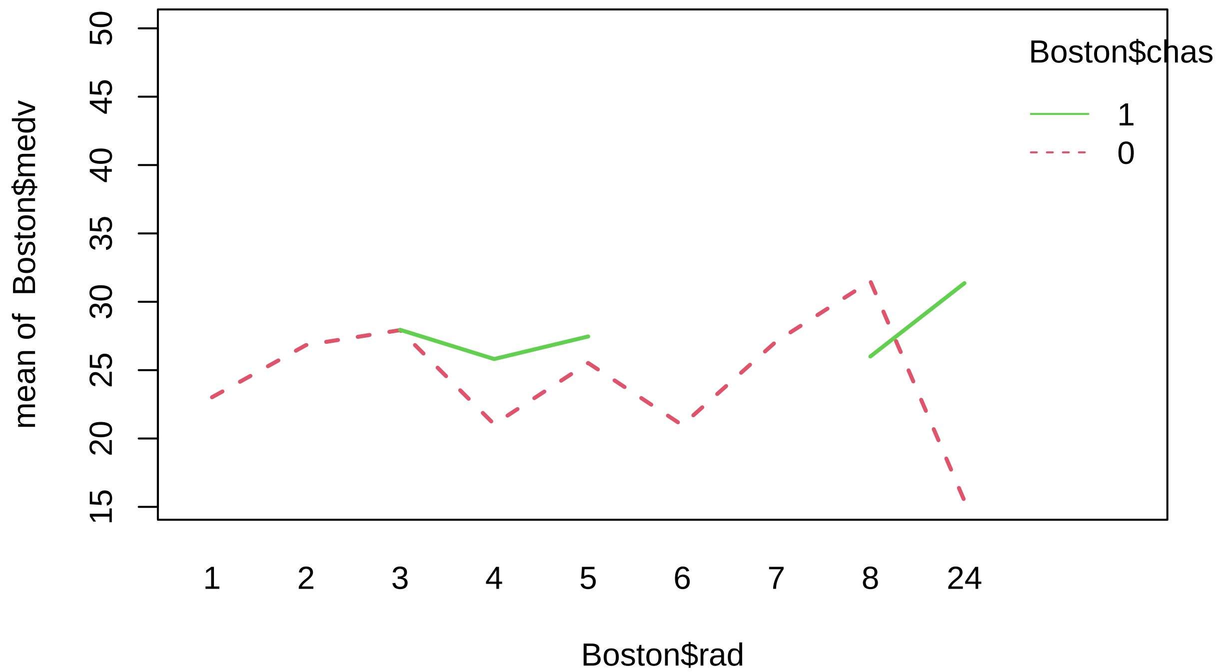
Problem 1C

```
boxplot(medv ~ chas + rad, data=Boston)
```



Obviously, we can not compare every situation as there are not observations for chas = 1 in every category of rad. For those categories of rad that have observations for both categories of chas, we see a variety of things: some pairs have about the same distribution, some have a lower distribution for chas = 0, and some have a lower distribution for chas = 1. We can especially see a big difference in rad when separating out by chas in the categories rad = 4 and rad = 24.

```
interaction.plot(Boston$rad, Boston$chas, Boston$medv, col=2:4, lwd=2, cex.axis=1, cex.lab=1)
```




```r
fit = lm(medv ~ chas * rad, data = Boston)
anova(fit)
```

```
## 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 F-test is testing the following models:

- 1st row: **1** versus **chas**
- 2nd row: **chas** versus **chas + rad**
- 3rd row: **chas + rad** versus **chas + rad + chas*rad**

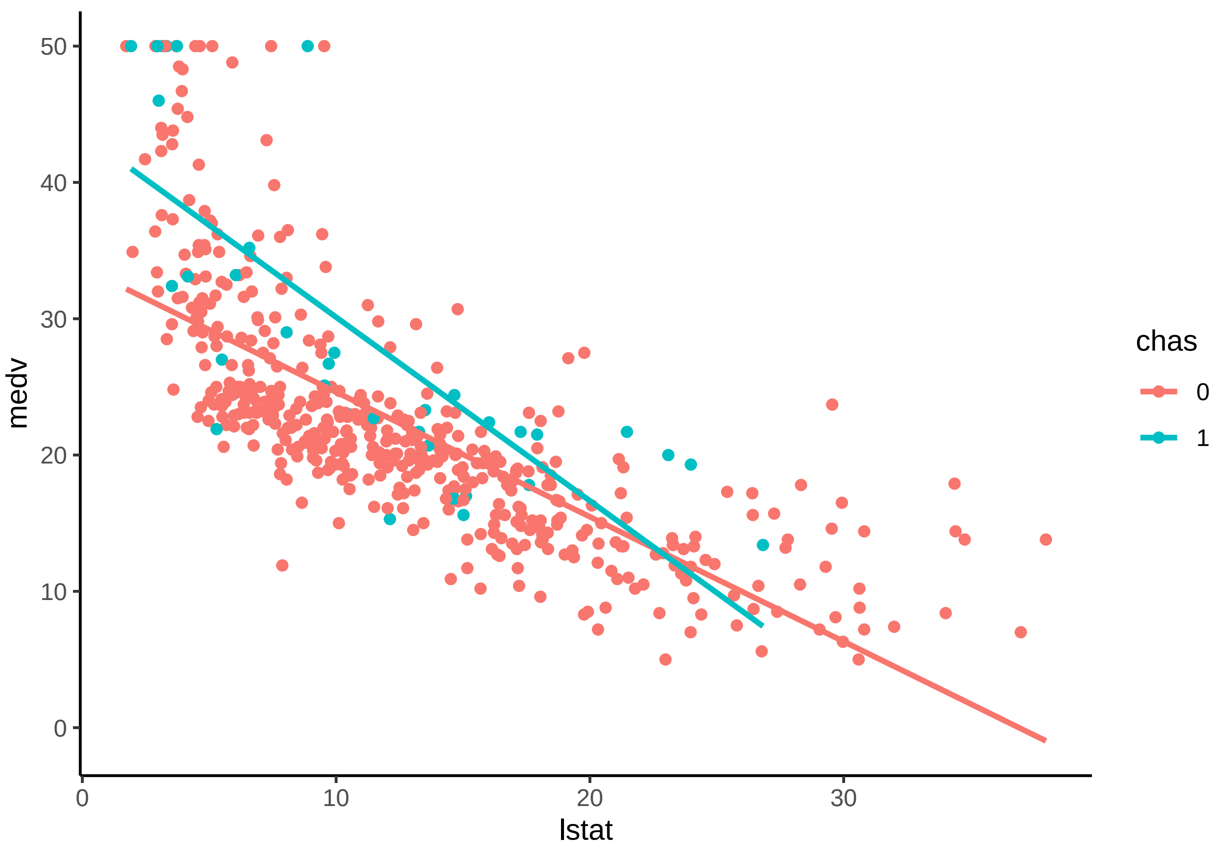
We see that each of the corresponding p-values are quite small, so we conclude that it is a good idea to use the full model of **chas + rad + chas*rad**.

The results of the test are consistent with the boxplots and the interactions plots because we see that looking at both rad and chas together yields much more insight into the median house value. Especially in the interactions plot, we see that for $rad \in \{4, 8, 24\}$, the chas categories produce very different median house values.

Problem 1D

```
library(ggplot2)
ggplot(Boston, aes(x = lstat, y = medv, color = factor(chas))) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "lstat", y = "medv", color = "chas") + theme_classic()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



From the scatterplot, we can see that the rate of decrease in median property value with respect to the percentage of lower status population appears to be steeper for properties that have chas = 0 than those with chas = 1. This indicates that whether a house borders the Charles River has an influence on the rate of decrease.

```
fit = lm(medv ~ chas * lstat, data = Boston)
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: medv
##              Df   Sum Sq Mean Sq  F value    Pr(>F)
## chas          1   1312.1  1312.1  35.7618 4.238e-09 ***
## lstat         1 22718.2 22718.2 619.2025 < 2.2e-16 ***
```

```
## chas:lstat   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 corresponds to the following hypotheses tests:

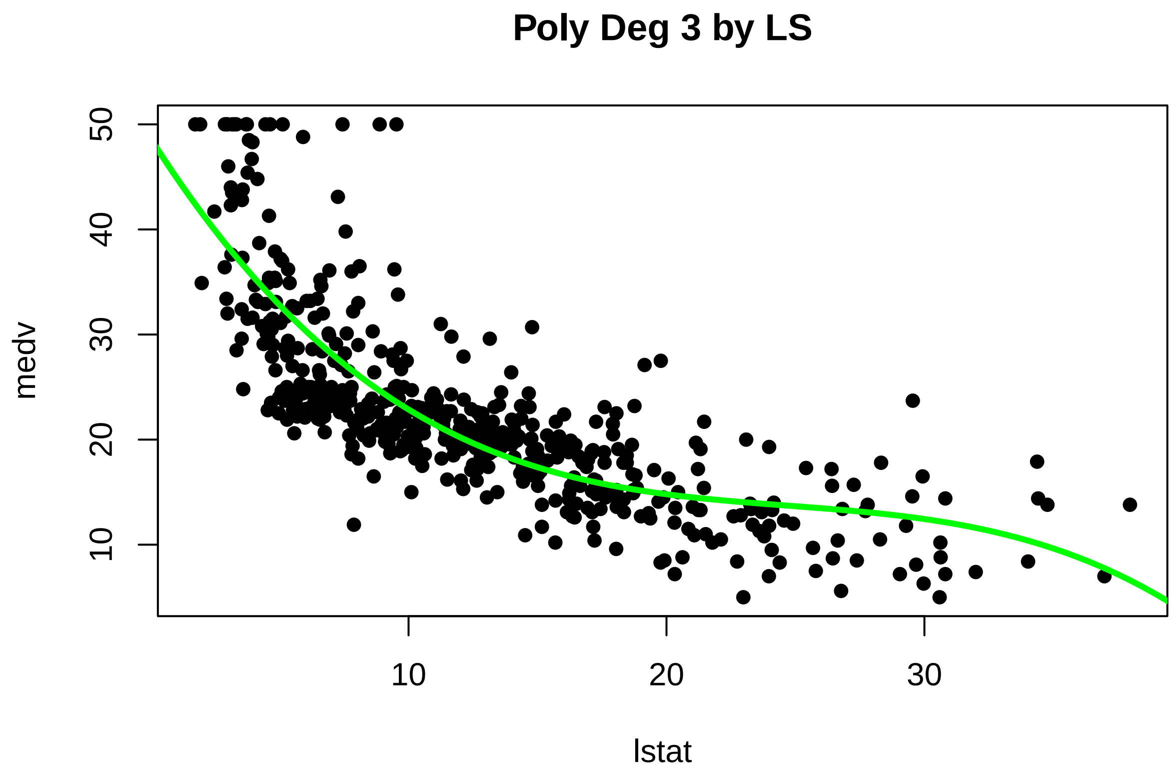
- 1st row: **1** (H0) versus **chas** (H1)
- 2nd row: **chas** (H0) versus **chas + lstat** (H1)
- 3rd row: **chas + lstat** (H0) versus **chas + lstat + chas*lstat** (H1)

Hence, we see that the variables chas and lstat are significant in predicting medv. The interaction term chas*lstat produces a p-value that is not extremely small, but still small enough that the null hypothesis should be rejected in favor of the alternate hypothesis. This means that chas, lstat, and the combined effect of chas and lstat are important in explaining the variance in the median house value.

Problem 2A

```r
library(quantreg)
library(MASS)
attach(Boston)

# fitting polynomial of degree 3 by LS
fit.ls = lm(medv ~ poly(lstat, 3, raw = TRUE))
plot(lstat, medv, pch = 16, main="Poly Deg 3 by LS")
pts = seq(0, 40, len=100)
preds = predict(fit.ls, data.frame(lstat = pts))
lines(pts, preds, col="green", lwd = 3)
```



Problem 2BC

```
attach(Boston)

# fitting polynomial of degree
formula = medv ~ poly(lstat, 3, raw = TRUE)
colors = c("green", "blue", "red", "orange", "purple", "pink", "yellow")
plot(lstat, medv, pch = 16)

# LS
preds = predict(fit.ls, data.frame(lstat = pts))
lines(pts, preds, col=colors[0], lwd = 3)

# L1 regression
fit.l1 = rq(formula, data=Boston)
preds = predict(fit.l1, data.frame(lstat = pts))
lines(pts, preds, col=colors[1], lwd = 3)

# Huber
fit.huber = rlm(formula, data = Boston, maxit=50, psi = psi.huber)
preds = predict(fit.huber, data.frame(lstat = pts))
lines(pts, preds, col=colors[2], lwd = 3)

# Hampel
```

```r
fit.hampel = rlm(formula, data = Boston, maxit=50, psi = psi.hampel)
preds = predict(fit.hampel, data.frame(lstat = pts))
lines(pts, preds, col=colors[3], lwd = 3)

# Tukey
fit.tukey = rlm(formula, data = Boston, maxit=50, psi = psi.bisquare)
preds = predict(fit.tukey, data.frame(lstat = pts))
lines(pts, preds, col=colors[4], lwd = 3)

# Least Median of Squares
fit.lms = lmsreg(formula, data=Boston)
preds = predict(fit.lms, data.frame(lstat = pts))
lines(pts, preds, col=colors[5], lwd = 3)

# Least Trimmed Sum of Squares
fit.lts = ltsreg(formula, data=Boston)
preds = predict(fit.lts, data.frame(lstat = pts))
lines(pts, preds, col=colors[6], lwd = 3)

legend('topright', c('LS','L1', 'Huber', 'Hampel', 'Tukey', "LMS", "LTS"),
       col=colors, lwd=3, bg='white')
```