GR5291 Advanced Data Analysis Problem Set 3

Francis Zhang

September 20, 2024

Question

1.Fit a multiple linear regression model to predict medv (median value of owner-occupied homes in $1000s) using the following set of predictors:

- crim per capita crime rate by town.
- zn proportion of residential land zoned for lots over 25,000 sq.ft.
- indus proportion of non-retail business acres per town.
- nox nitrogen oxides concentration (parts per 10 million).
- rm average number of rooms per dwelling.
- age proportion of owner-occupied units built prior to 1940.

2.State and assess the validity of the underlying assumptions, and suggest remedial measures in case of violations of any of the underlying assumptions

- Linearity/functional form
- Normality
- Homoscedasticity
- Uncorrelated error

3.Repeat (1) using Least Median of Squares Regression and compare the results with those obtained in (1).

Solution

Question 1

```
library(MASS)
names(Boston)
```

```
##  [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"
##  [8] "dis"     "rad"     "tax"     "ptratio" "black"   "lstat"   "medv"
```

```
head(Boston)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
##   medv
```

```
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7
```

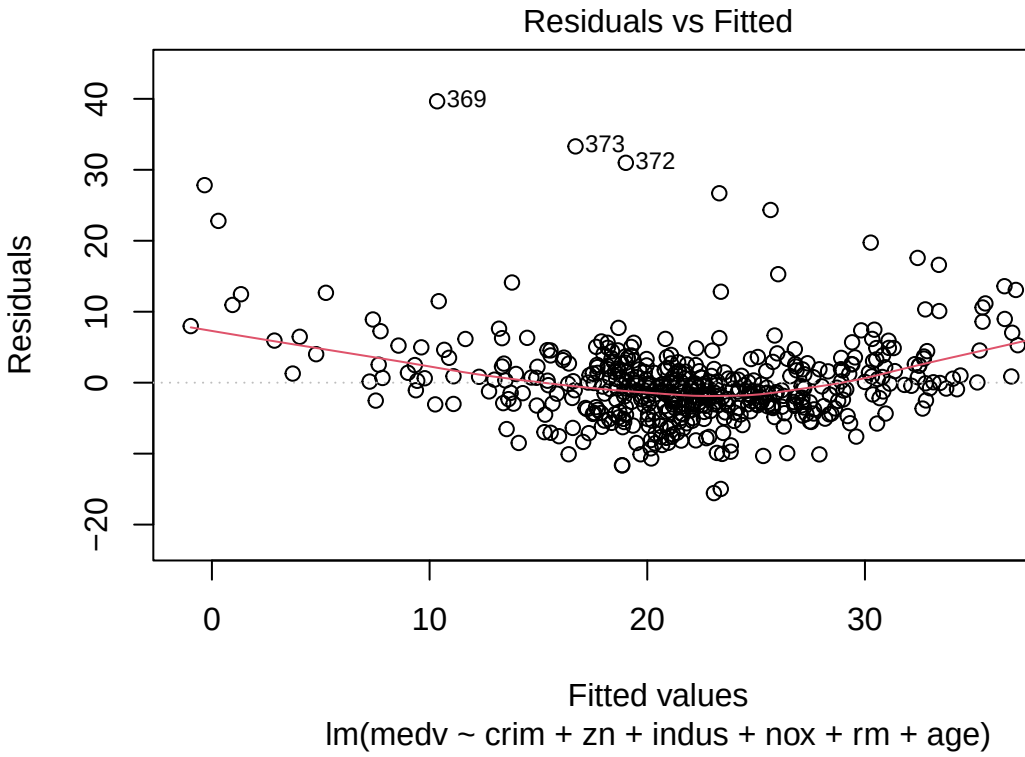
```
lm_model <- lm(medv ~ crim + zn + indus + nox + rm + age, Boston)
summary(lm_model)
```

```
##
## Call:
## lm(formula = medv ~ crim + zn + indus + nox + rm + age, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.797  -3.179  -0.723   2.265  39.652
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -20.15009    3.25112  -6.198 1.20e-09 ***
## crim         -0.18818    0.03515  -5.353 1.32e-07 ***
## zn            0.01439    0.01480   0.972    0.331
## indus        -0.12264    0.06558  -1.870    0.062 .
## nox          -3.24789    4.16231  -0.780    0.436
## rm            7.61122    0.42278  18.003  < 2e-16 ***
## age          -0.02139    0.01504  -1.422    0.156
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.051 on 499 degrees of freedom
## Multiple R-squared:  0.5722, Adjusted R-squared:  0.5671
## F-statistic: 111.3 on 6 and 499 DF,  p-value: < 2.2e-16
```

Our regression model is $medv = -20.150 - 0.188crim + 0.014zn - 0.123indus - 3.248nox + 7.611rm - 0.021age$.

Question 2

```
plot(lm_model, which = 1)
```



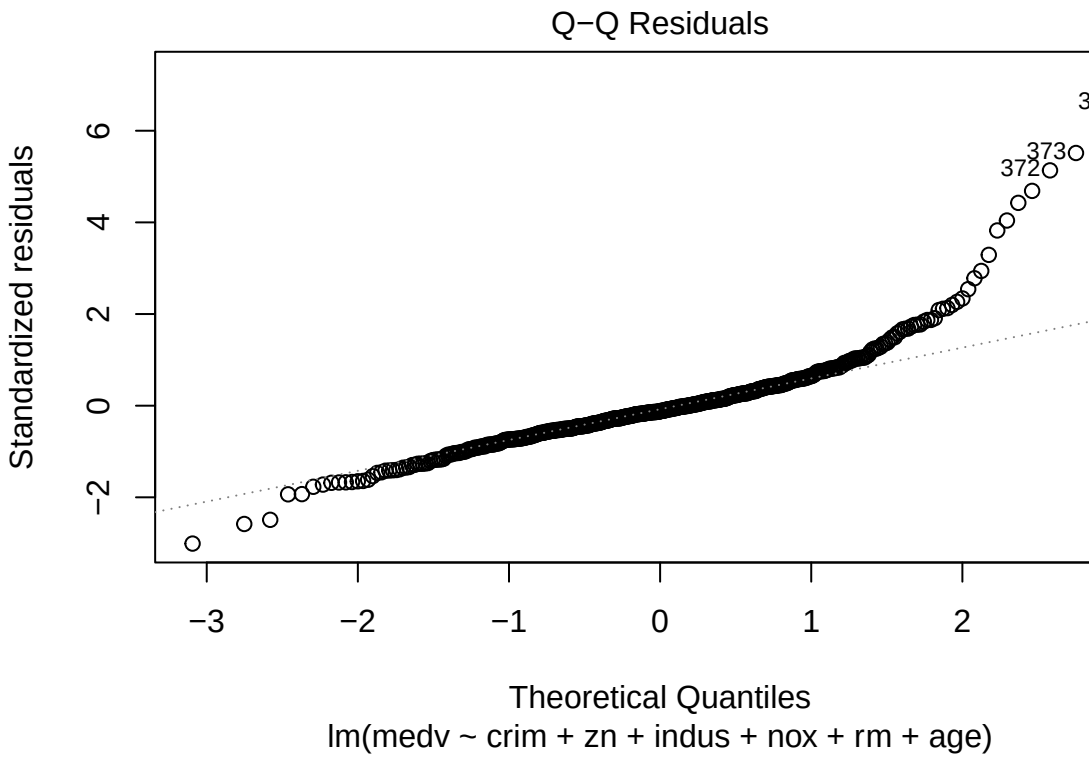
a. Linearity / Functional Form

Assumption: The relationship between the dependent variable (*medv*) and each predictor is assumed to be linear.

Test: A Residuals vs Fitted Plot is used to assess linearity. If the plot shows a random scatter of points without a clear pattern, the linearity assumption holds. If a pattern (such as a curve) appears, the assumption is violated. In this case, the residuals versus fitted plot indicates some potential non-linearity.

Remedial Measure: Add polynomial terms or apply transformations to non-linear predictors.

```
plot(lm_model, which = 2)
```



b. Normality of Residuals

```
shapiro.test(residuals(lm_model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(lm_model)
## W = 0.86686, p-value < 2.2e-16
```

Assumption: The residuals (errors) of the model should be normally distributed.

Test: A Q-Q Plot and Shapiro-Wilk Test can be used to assess normality. The Q-Q plot compares the distribution of residuals to a theoretical normal distribution. In this Q-Q plot, the residuals deviate from the diagonal, and the Shapiro-Wilk test yields a very small p-value ($< 2.2e\text{-}16$), indicating a violation of normality.

Remedial Measure: Transforming the dependent variable (e.g., log transformation) or using robust regression methods like LMS.

```
library(lmtest)
```

c. Homoscedasticity

```
## Loading required package: zoo
```

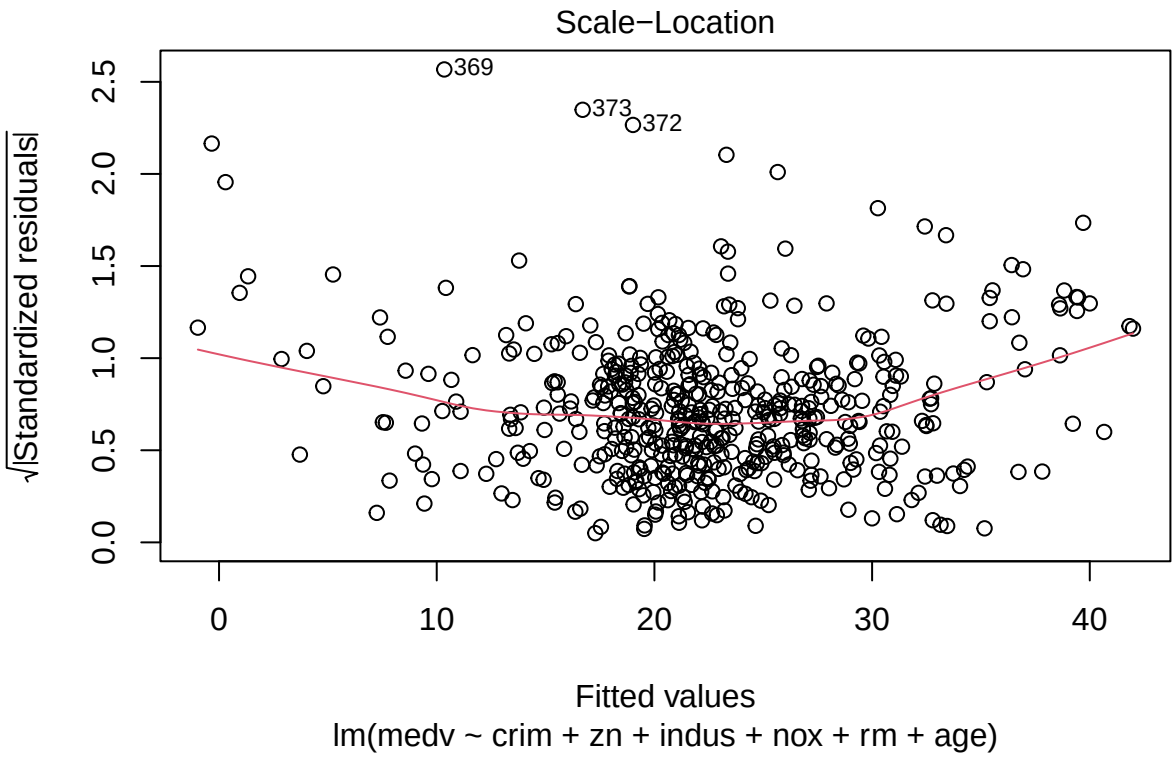
```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
bptest(lm_model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm_model
## BP = 21.012, df = 6, p-value = 0.001825
```

```
plot(lm_model, which = 3)
```



Assumption: The variance of the residuals should remain constant across all levels of the fitted values.

Test: The Breusch-Pagan Test and the Scale-Location Plot are used to test this. In this case, the Breusch-Pagan test gives a p-value of 0.001825, indicating a violation of homoscedasticity (heteroscedasticity is present).

Remedial Measure: Apply transformations (log/square-root) to the response variable, or use Weighted Least Squares (WLS) regression.

```
library(lmtest)
dwtest(lm_model)
```

d. Uncorrelated Errors

```
##
##  Durbin-Watson test
##
## data:  lm_model
## DW = 0.74375, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

Assumption: The residuals should not be correlated with each other. Correlated residuals typically indicate that important predictors have been omitted from the model or that the data has some structure (e.g., time series or spatial structure) that has not been accounted for.

Test: The Durbin-Watson Test is used to check for autocorrelation. In this case, the Durbin-Watson statistic is 0.74375 with a very low p-value (< 2.2e-16), indicating significant positive autocorrelation in the residuals.

Remedial Measure: Use Generalized Estimating Equations (GEE) to account for correlated residuals or include lagged predictors.

Question 3

```
lms_model <-
  lqs(medv ~ crim + zn + indus + nox + rm + age, Boston, method = "lms")

lms_model
```

```
## Call:
## lqs.formula(formula = medv ~ crim + zn + indus + nox + rm + age,
##     data = Boston, method = "lms")
##
## Coefficients:
## (Intercept)         crim           zn        indus          nox           rm
##  -49.349374    -1.010104     0.017715    -0.004084    30.298707    10.091348
##         age
##   -0.099849
##
## Scale estimates 3.583 3.456
```

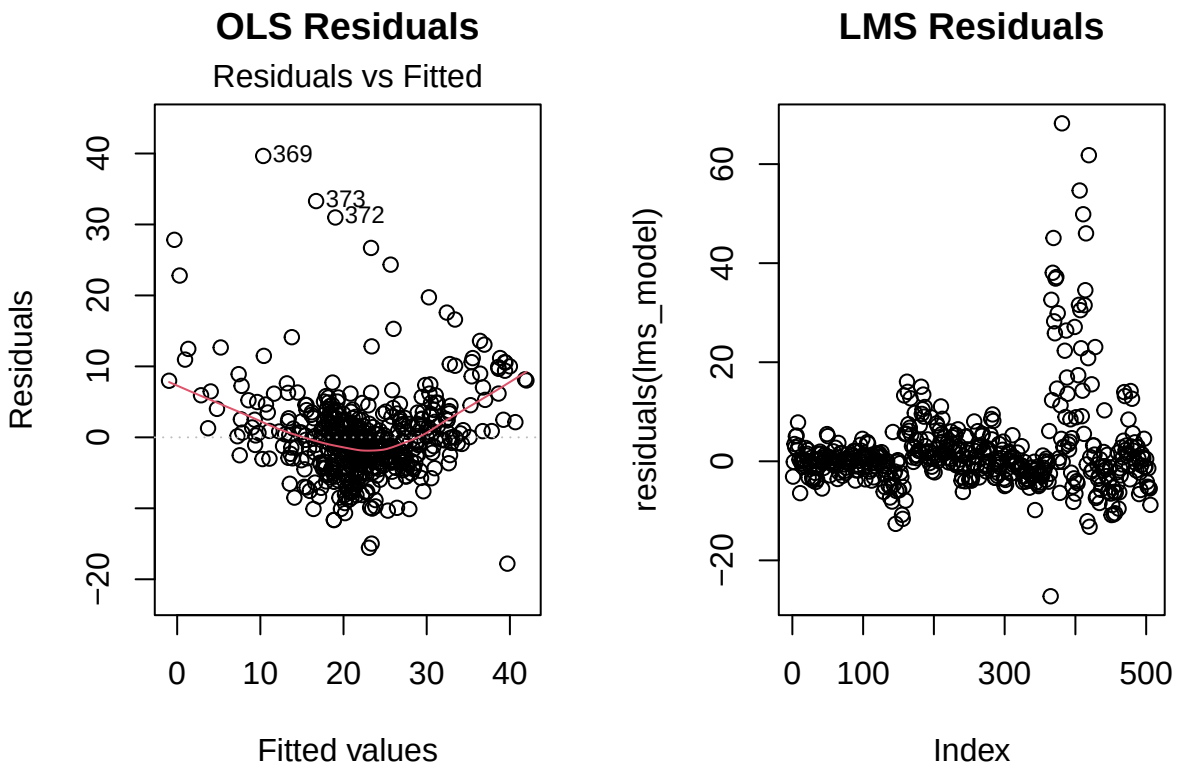
```
# Coefficients of the OLS model
coef(lm_model)
```

```
##  (Intercept)         crim           zn        indus          nox           rm
## -20.15009449  -0.18817525   0.01438957  -0.12264331  -3.24789038   7.61121532
##          age
##  -0.02139131
```

```
# Coefficients of the LMS model
coef(lms_model)
```

```
##   (Intercept)          crim            zn         indus           nox
## -49.349373901  -1.010104306   0.017715228  -0.004083774  30.298707166
##            rm           age
##  10.091348075  -0.099848785
```

```
# Plot residuals for OLS vs LMS
par(mfrow = c(1, 2))
plot(lm_model, which = 1, main = "OLS Residuals")
plot(residuals(lms_model), main = "LMS Residuals")
```



Comparison Between OLS and LMS Regression

Explanation: The Least Median of Squares (LMS) regression minimizes the median of squared residuals, making it more robust to outliers compared to Ordinary Least Squares (OLS), which minimizes the sum of squared residuals. This robustness is especially beneficial in datasets that might contain outliers or influential points that could skew the OLS results.

a. Coefficients Comparison:

• The coefficients in the OLS and LMS models can differ significantly if outliers are present. In the provided results: • The LMS model shows a larger intercept compared to OLS , and some predictors like rm have higher coefficients in LMS than in OLS. • Predictors like nox have coefficients that are substantially different between the two models, suggesting that LMS is handling influential data points differently from OLS.

b. Residuals Comparison:

• The residuals plot comparison between OLS and LMS (as shown in the problem set) highlights that LMS tends to produce more stable residuals, especially in the presence of outliers. • OLS Residuals: The residual plot for OLS shows some larger deviations, indicating that OLS might be affected by extreme values. • LMS Residuals: The residuals in LMS tend to be more concentrated around zero, indicating that LMS is less sensitive to outliers.

c. Impact of Outliers:

• OLS Sensitivity: OLS can be heavily influenced by outliers, leading to biased estimates and increased variance in predictions. If the dataset contains outliers or highly influential points, OLS may not perform well. • LMS Robustness: LMS is more robust to outliers, as it focuses on minimizing the median of squared residuals. This makes LMS a better choice in situations where the data may contain anomalies that could unduly influence the model.

d. Fit of the Models:

• OLS Fit: The R-squared value from OLS is typically higher because OLS optimizes for the sum of squared residuals. However, a higher R-squared does not always indicate a better model when

outliers are present. • LMS Fit: LMS does not provide an R-squared value, but it often gives a better representation of the central trend in the data when outliers exist.

e. Conclusion:

• Use of OLS: OLS is appropriate when the data is clean (free of outliers), and the assumptions of linear regression (normality, homoscedasticity) are met. • Use of LMS: LMS should be used when there are concerns about outliers or influential data points that could unduly impact the OLS results. LMS produces more robust estimates, especially in the presence of data that violates some assumptions of OLS.