Nogara - Final assignment

Alberto Galdino Nogara

2024-06-21

Point 1.

Introduction

Welcome to this comprehensive statistical analysis of Boston house prices. In this R Markdown document, our aim is to gain valuable insights into the factors influencing housing prices in the city of Boston. Understanding the determinants of house prices is of paramount importance for various stakeholders, including prospective home buyers, real estate developers, and policymakers. By conducting a rigorous statistical analysis, we aim to shed light on the key factors that contribute to variations in housing prices, ultimately providing actionable insights for informed decision-making.

Dataset Overview

The dataset at the heart of our analysis encapsulates a wealth of information related to housing characteristics and prices across different neighborhoods in Boston. As we delve into this rich collection of data, our goal is to employ statistical methods to discern patterns and relationships that can inform our understanding of the housing market dynamics. The provenance of the data is "StatLib - Carnegie Mellon University" (The Boston house-price data of Harrison, D. and Rubinfeld, D.L. 'Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978.); while the dataset has been downloaded from Kaggle. The dataset contains 506 observations and 14 variables (of which 13 can be considered as predictors variables, while the MEDV variable can be considered as the response variable). What follow now is a short description of each variable:

```
# - 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
# - CHAS = Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
# - NOX = nitric oxides concentration (parts per 10 million)
# - RM = average number of rooms per dwelling
# - AGE = proportion of owner-occupied units built prior to 1940
# - DIS = weighted distances to five Boston employment centres
# - RAD = index of accessibility to radial highways
# - TAX = full-value property-tax rate per $10,000
# - PTRATIO = pupil-teacher ratio by town
# - B = B=1000(Bk - 0.63)^2 where Bk is the proportion of African origin people by town
# - LSTAT = % lower status of the population
# - MEDV = Median value of owner-occupied homes in $1000's
```

Point 2.

Analysis Goals

1. **Descriptive Exploration:** Provide an overview of the dataset, examining summary statistics, data distributions, and key characteristics of the variables.
2. **Factors Influencing Prices:** Identify and analyze the key factors influencing housing prices in Boston, exploring relationships between variables and their impact on property values.
3. **Modeling and Prediction:** Explore the potential for modeling house prices using statistical techniques, with an emphasis on predictive analytics to understand future price trends.

More specifically, what follow are the applied goals in this analysis: * **Variable Selection**: Identify the most influential predictors of house prices using variable selection techniques. * **Model Interpretation**: Interpret the coefficients and significance levels of regression models. * **Assumption Checking**: Assess the assumptions of linear regression models. * **Outlier Detection**: Identify and examine potential outliers in the dataset. * **Model Comparison**: Compare different regression models to determine the best fit. * **Heteroscedasticity**: Investigate techniques to address heteroscedasticity in the data.

Now let's look at how the data can address these goals.

Point 3.

```
sum(is.na(data)) #Checking for missing values
```

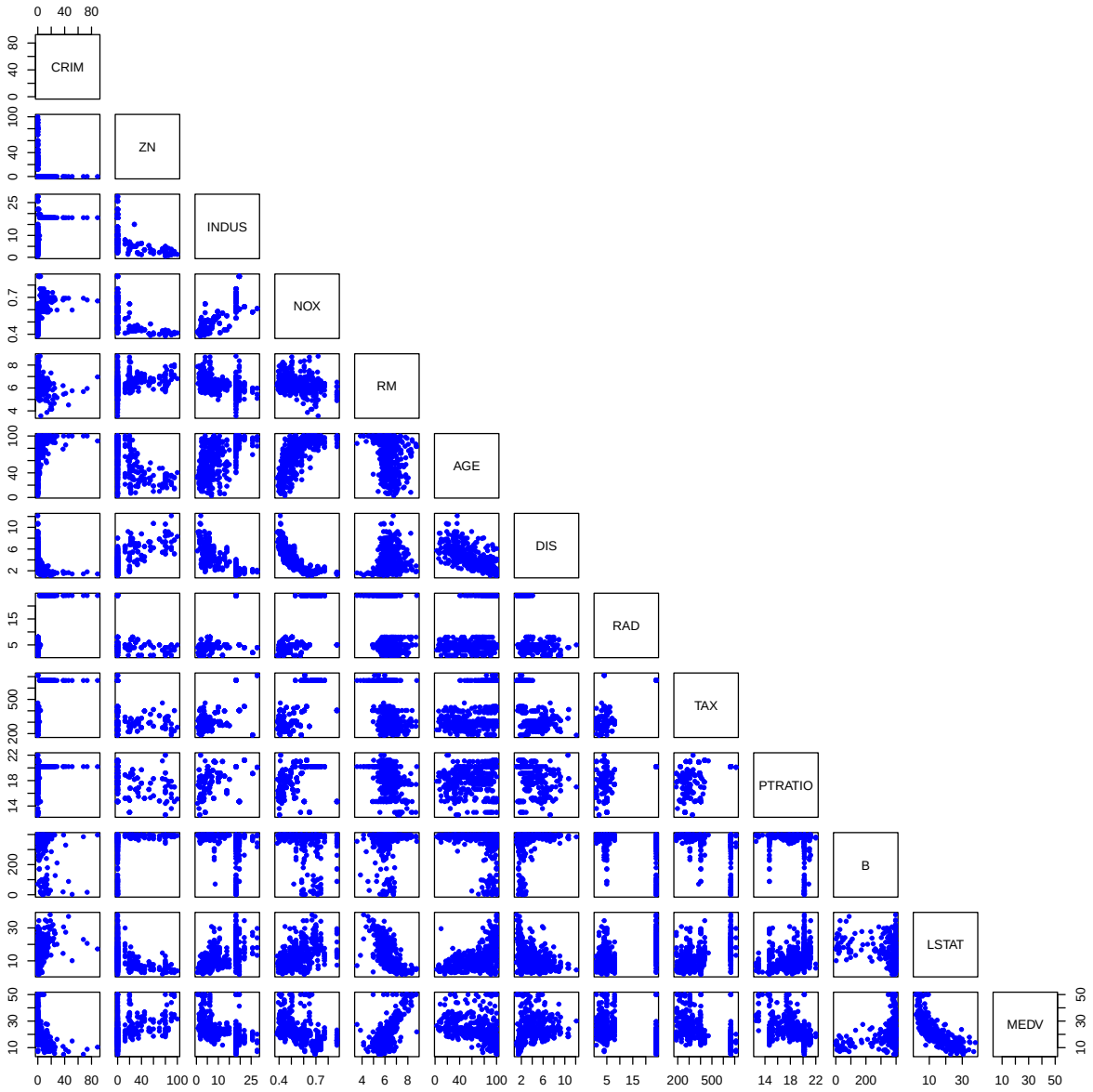
```
## [1] 0
```

```
summary(data) #quick summary of all the variables
```

```
##       CRIM                ZN              INDUS            CHAS        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
##       NOX               RM              AGE              DIS        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
##       RAD              TAX           PTRATIO            B        
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38  
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67  
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23  
```

```
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90
##      LSTAT            MEDV
##  Min.   : 1.73   Min.   : 5.00
##  1st Qu.: 6.95   1st Qu.:17.02
##  Median :11.36   Median :21.20
##  Mean   :12.65   Mean   :22.53
##  3rd Qu.:16.95   3rd Qu.:25.00
##  Max.   :37.97   Max.   :50.00
```

```
# Create scatterplot matrix
pairs(data[, c("CRIM", "ZN", "INDUS", "NOX", "RM", "AGE", "DIS", "RAD", "TAX", "PTRATIO", "B", "LSTAT",
               "MEDV")],
      col = "blue", pch = 20, upper.panel = NULL) #dummy variable CHAS excluded
```



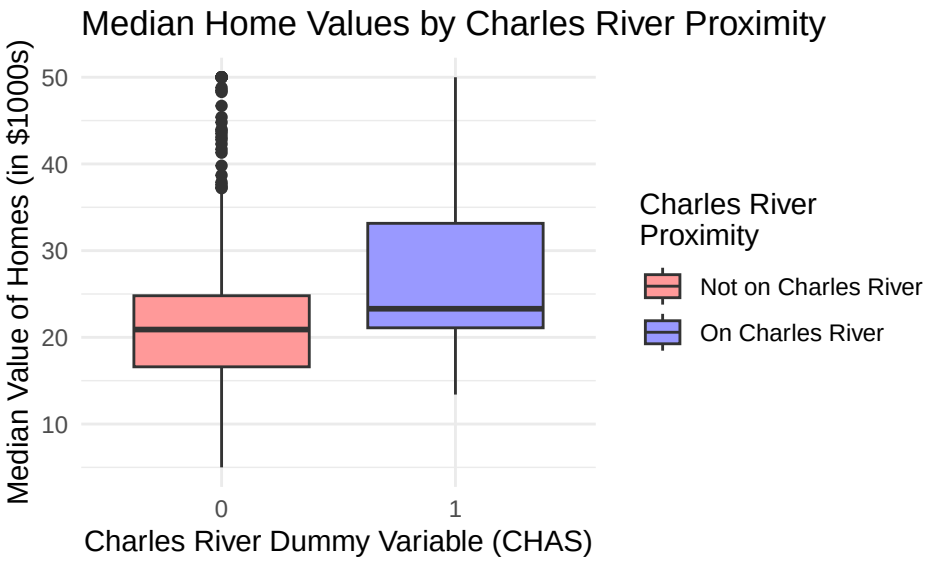
Above we can look at the scatterplox matrix of our dataset, which shows pairwise relationships between the

variables. From the figure we can make some comments:

- **CRIM vs. MEDV:** There seems to be a negative relationship; as crime rate increases, the median value of homes tends to decrease.
- **ZN (proportion of residential land zoned for lots over 25,000 sq.ft.):** Higher values of ZN seem to be associated with higher median values (MEDV), indicating that larger lot sizes might correspond to more expensive houses.
- **INDUS (proportion of non-retail business acres per town) vs. NOX (nitric oxides concentration):** There's a clear positive relationship, suggesting that more industrial areas have higher pollution levels.
- **NOX vs. DIS (weighted distances to five Boston employment centers):** As NOX increases, DIS tends to decrease, indicating that more polluted areas are closer to employment centers.
- **RM vs. MEDV:** A strong positive relationship is observed; more rooms typically indicate higher house values.
- **AGE (proportion of owner-occupied units built prior to 1940) vs. DIS:** Older houses tend to be closer to employment centers, as indicated by the negative relationship.
- **DIS vs. MEDV:** There's a slight positive trend, suggesting that houses further from employment centers are more valuable, possibly due to desirability of less densely populated areas.
- **RAD (accessibility to radial highways) vs. TAX (full-value property-tax rate per $10,000):** There's a clear positive relationship, indicating that properties with better highway access tend to have higher tax rates.
- **PTRATIO (pupil-teacher ratio by town) vs. MEDV:** There's a general negative trend, suggesting that higher values of PTRATIO (more students per teacher) are associated with lower house values.
- **B (1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town) vs. MEDV:** It's difficult to discern a clear pattern from the scatterplot, suggesting a weak relationship.
- **LSTAT vs. MEDV:** There's a strong negative relationship; as the percentage of lower status population increases, the median value of homes tends to decrease significantly.

These interpretations from the scatterplot matrix suggest that variables such as RM, LSTAT, PTRATIO, and CRIM have more evident relationships with MEDV, indicating they may be strong predictors of house value in the Boston area.

```
# Create boxplot of MEDV split by CHAS with color differentiation
ggplot(data, aes(x = as.factor(CHAS), y = MEDV, fill = as.factor(CHAS))) +
  geom_boxplot() +
  scale_fill_manual(values = c("#FF9999", "#9999FF"),
                    labels = c("Not on Charles River", "On Charles River"),
                    name = "Charles River\nProximity") +
  labs(x = "Charles River Dummy Variable (CHAS)",
       y = "Median Value of Homes (in $1000s)",
       title = "Median Home Values by Charles River Proximity") +
  theme_minimal()
```



The comparison indicates that proximity to the Charles River may be associated with higher median home values. The presence of extreme values, particularly for homes not on the river, suggests that there can be high-value properties outside of the river area as well, although they are less common. The absence of similar extreme values for river-adjacent homes may imply a more stable market or less variation in home features that significantly drive up value.

Point 4

Now we are going to consider all the available covariates to explain our response variable (MEDV) and we are going to - initially - assume a linear regression model.

```
ols <- lm(MEDV ~ . , data = data)
summary(ols)
```

```
##
## Call:
## lm(formula = MEDV ~ ., data = data)
##
## 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
## CHAS         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
```

The regsubsets() function included in the leaps library performs best subset selection by identifying the best model that contains a given number of predictors, where best is quantified using RSS. The summary() command outputs the best set of variables for each model size.

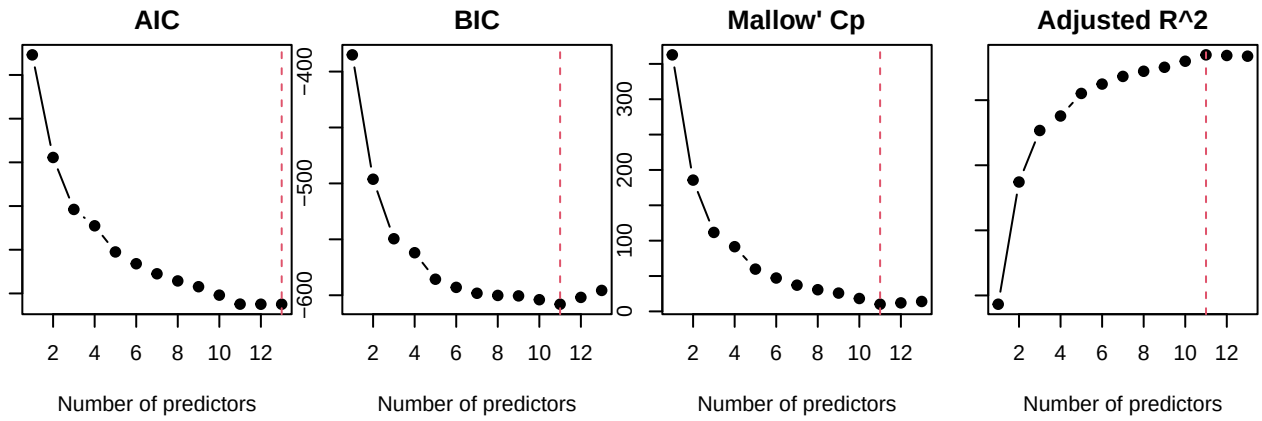
```
ols_best_subsets <- regsubsets(MEDV ~ . ,nvmax = 13 , data = data)
summary(ols_best_subsets)
```

```
## Subset selection object
## Call: regsubsets.formula(MEDV ~ ., nvmax = 13, data = data)
## 13 Variables  (and intercept)
##         Forced in Forced out
## CRIM        FALSE      FALSE
## ZN          FALSE      FALSE
## INDUS       FALSE      FALSE
## CHAS        FALSE      FALSE
## NOX         FALSE      FALSE
## RM          FALSE      FALSE
## AGE         FALSE      FALSE
## DIS         FALSE      FALSE
## RAD         FALSE      FALSE
## TAX         FALSE      FALSE
## PTRATIO     FALSE      FALSE
## B           FALSE      FALSE
## LSTAT       FALSE      FALSE
## 1 subsets of each size up to 13
## Selection Algorithm: exhaustive
##           CRIM ZN  INDUS CHAS NOX RM  AGE DIS RAD TAX PTRATIO B   LSTAT
## 1  ( 1 )  " "  " " " "   " "  " " " " " " " " " " "     " " "*"
## 2  ( 1 )  " "  " " " "   " "  " " "*" " " " " " " "     " " "*"
## 3  ( 1 )  " "  " " " "   " "  " " "*" " " " " " " "*"     " " "*"
## 4  ( 1 )  " "  " " " "   " "  " " "*" " " "*" " " "*"     " " "*"
## 5  ( 1 )  " "  " " " "   " "  "*" "*" " " "*" " " "*"     " " "*"
## 6  ( 1 )  " "  " " " "   "*"  "*" "*" " " "*" " " "*"     " " "*"
## 7  ( 1 )  " "  " " " "   "*"  "*" "*" " " "*" " " "*"     "*" "*"
## 8  ( 1 )  " "  "*" " "   "*"  "*" "*" " " "*" " " "*"     "*" "*"
## 9  ( 1 )  "*"  " " " "   "*"  "*" "*" " " "*" "*" "*"     "*" "*"
## 10  ( 1 ) "*"  "*" " "   " "  "*" "*" " " "*" "*" "*"     "*" "*"
## 11  ( 1 ) "*"  "*" " "   "*"  "*" "*" " " "*" "*" "*"     "*" "*"
## 12  ( 1 ) "*"  "*" "*"   "*"  "*" "*" " " "*" "*" "*"     "*" "*"
## 13  ( 1 ) "*"  "*" "*"   "*"  "*" "*" "*" "*" "*" "*"     "*" "*"
```

An asterisk indicates that a given variable is included in the corresponding model. For instance, this output indicates that the best two-variable model contains only RM and LSTAT.

Point 5

What is the best overall model obtained according to different criteria (e.g. AIC, BIC, adjusted R2, Mallow's Cp) and Cross-validation error? We can get to know that by the following figures:



From the plots above we can infer that the best overall model is the one with 11 predictors - the difference in AIC between the model with 11 predictors and the full one is so minimal that in front of the others plots, we are comfortable in choosing the model with 11 predictors. "Best" in this case means that it offers the best trade-off between predictive accuracy and model complexity. As form the regsubset function above we can observe that the best model with 11 predictors is the one leaving out the independent variables INDUS and AGE.

```
ols_best = lm(MEDV ~ . - INDUS - AGE, data=data)
round(coef(ols_best),4)
```

```
## (Intercept)        CRIM          ZN        CHAS         NOX          RM
##     36.3411     -0.1084      0.0458      2.7187    -17.3760      3.8016
##         DIS         RAD         TAX     PTRATIO           B       LSTAT
##     -1.4927      0.2996     -0.0118     -0.9465      0.0093     -0.5226
```

Cross-validation error

```
hat <- influence(ols_best)$hat
yhat <- fitted(ols_best)
CV = mean((data$MEDV-yhat)^2/(1-hat)^2)
CV
```

```
## [1] 23.51325
```

```
ols_pre = lm(MEDV ~ . - INDUS - AGE - CHAS, data=data)
hat2 <- influence(ols_pre)$hat
yhat2 <- fitted(ols_pre)
CV2 = mean((data$MEDV-yhat2)^2/(1-hat2)^2)
CV2
```

```
## [1] 23.74331
```

```
ols_post = lm(MEDV ~ . - AGE , data=data)
hat3 <- influence(ols_post)$hat
yhat3 <- fitted(ols_post)
CV3 = mean((data$MEDV-yhat3)^2/(1-hat3)^2)
CV3
```

```
## [1] 23.57196
```

We have just used the LOOCV method to compute the cross-validation errors. We have compared the best model with 11 predictors, the model with 10 predictors and the one with 12. This results confirms what we have said before: the model with the smallest cross-validation error is the preferred for our analysis, which is the one with 11 predictors.

Point 6

```
vif(ols_best)
```

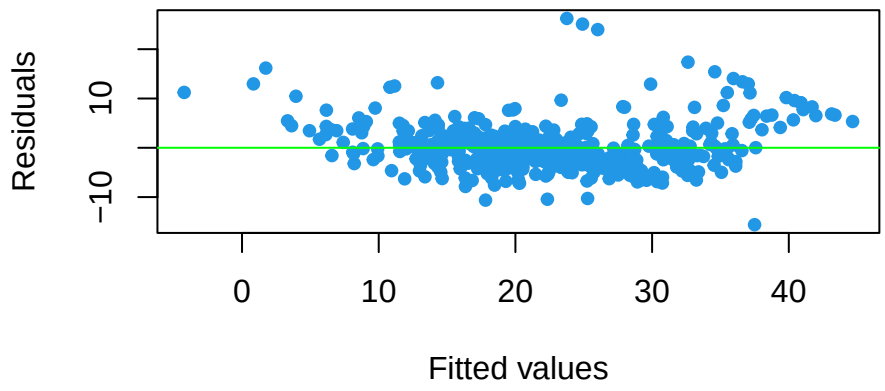
```
##     CRIM       ZN     CHAS      NOX       RM      DIS      RAD      TAX
## 1.789704 2.239229 1.059819 3.778011 1.834806 3.443420 6.861126 7.272386
##  PTRATIO        B    LSTAT
## 1.757681 1.341559 2.581984
```

The output from vif() function displays the Variance Inflation Factors (VIF) for each predictor in our best-fitting ordinary least squares (OLS) regression model. VIF is a measure of how much the variance of an estimated regression coefficient increases because of collinearity. A rule of thumb is that a VIF greater than 10 indicates high collinearity between this predictor and the others, which can affect the reliability of the coefficient estimates. As we notice all the predictors have VIF values below 10, indicating that we do not have problematic levels of collinearity in our model.

Point 7

Costant variance assumptiopn for the errors

```
plot(fitted(ols_best), residuals(ols_best),
     xlab="Fitted values", ylab="Residuals",
     col=4, pch=19, cex=0.8) +
abline(h = 0, col = "green")
```

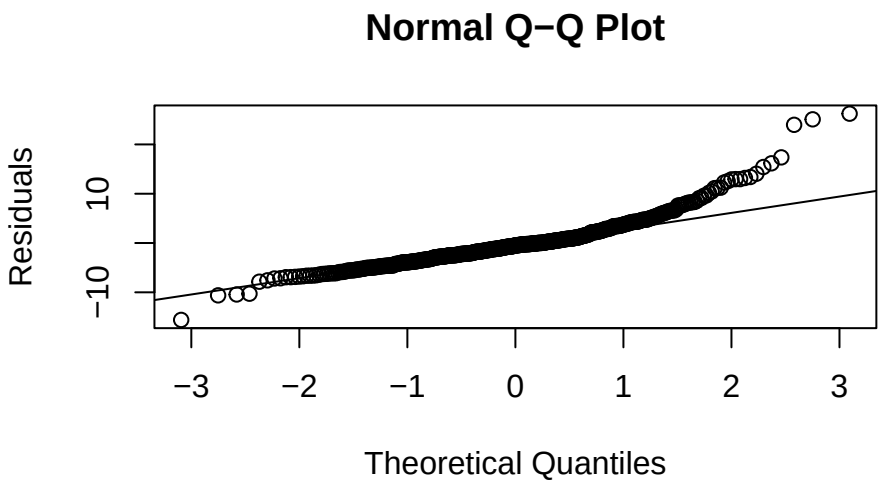



```
## integer(0)
```

The plot is designed to check for non-constant variance and non-linearity. The residuals are centered around the horizontal line at zero, which is good—it suggests that there's no systematic bias in the predictions. Regarding heteroscedasticity we see that there is no shape like a right (or left) opening megaphone or a double outward bow, hence we are in a homoscedastic environment. We can also see that here are no clear patterns provided by the graph in its entirety, hence we do not have a problem of non-linearity.

Normality of the residuals

```
qqnorm (residuals (ols_best), ylab="Residuals")
qqline (residuals (ols_best))
```



Normal residuals should follow the line approximately. Here, most of the residuals look normal even though a slightly long tailed distribution is clearly noticeable. In addition, we use the Shapiro-Wilk normality test:

```
shapiro.test(residuals(ols_best))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(ols_best)
## W = 0.90131, p-value < 2.2e-16
```

The null hypothesis is that the residuals are normal. Since the p-value is less than the significance level of 0.05, we reject the null hypothesis. A Bootstrap or Permutation Test (in a more detailed work) is highly recommended.

High leverage points

High leverage points are observations in a dataset that have unusual predictor (independent variable) values and exert a disproportionate influence on the estimation of a regression model's parameters. In the context of linear regression, these points are outliers in the space of the independent variables, rather than in the space of the dependent variable.

```
infl <- influence(ols_best)
hat <- infl$hat
sum(hat) #We verify that the sum of the leverages is indeed twelve
```

```
## [1] 12
```

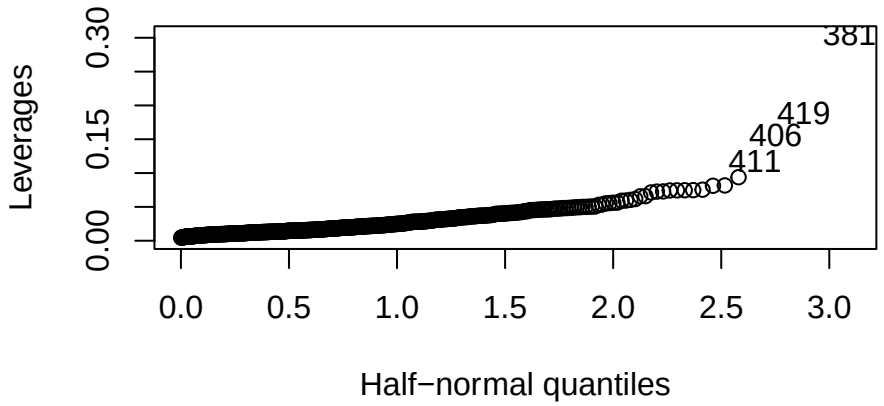
```
          #- the number of parameters in the model
```

```
hat[which(hat>=(2*25/nrow(data)))]
```

```
##       381       406       411       419
## 0.3048409 0.1564260 0.1175336 0.1893709
```

The above numbered observations have an high leverage statistic, as shown in the next figure that displays the half-normal plot on the leverages. In the half-normal plot we are usually not looking for a straight line relationship since we do not necessarily expect a positive normal distribution for the leverages. We are looking for points that diverge substantially from the rest of the data.

```
halfnorm(hat, 4, labs = rownames(data), ylab="Leverages")
```



Outliers

An outlier is a data point that differs significantly from other observations in a dataset. An outlier is one which has a standardized residual $> |3|$.

```
rsta <- rstandard(ols_best)
range(rsta)
```

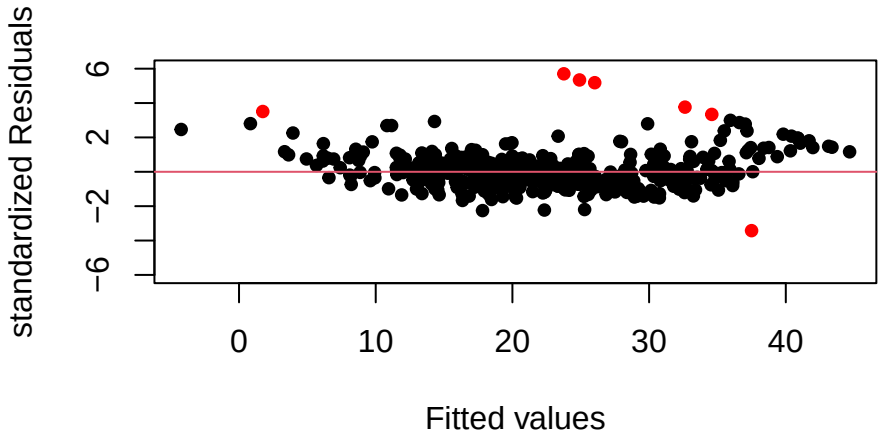
```
## [1] -3.423915  5.702120
```

```
sum(abs(rsta)>3)
```

```
## [1] 7
```

We have 7 outliers. Let's plot them:

```
plot(fitted(ols_best), rsta,
     xlab="Fitted values", ylab="standardized Residuals", pch=19, cex=0.8,
     ylim=c(-6,6),
     col=ifelse(abs(rsta) > 3, "red", "black")) # Color points red if abs(residuals) > 3)
abline(h=0, col=2)
```



Influential points

An influential point is one whose removal from the dataset would cause a large change in the fit. An influential point may or may not be an outlier and may or may not have large leverage but it will tend to have at least one of these two properties. Cook's distance uses both leverage and the size of the residual to determine the influence of an observation. It's calculated for each observation and can be interpreted as the change in the regression coefficients if that observation were excluded from the analysis.

```
cook <- cooks.distance(ols_best)
cook[which.max(cook)]
```

```
##       369
## 0.1612148
```

As a rule of thumb we can say that the observations with a Cook's distance > 1 are considered to be influential points. Here there are none.

Point 8

As evidenced by the diagnostics above we can affirm that our model does not have significant issues regarding the underlying assumptions of it. There is evidence of a short tailed distribution in the QQ-plot - which is made to check the normality of residuals - evidenced by the Shapiro-Wilk test. Since: A. there are no problems of non-constant variance (heteroscedasticity) or non-linearity; B. applying a response transformation does not change the situation in terms of residuals distribution; C. the distribution is the QQ-plot is short tailed; We can move on and ignore the issue. Regarding the outliers it is evident that even if they exists, they are: a. just a few; b. not due to measurement errors or data entry errors, but instead they are genuine "extreme" values. Hence, we treat them non treating them, which means no further action or investigation is required.

Point 9

The summary of the best model is the following:

```r
summary_ols_best <- summary(ols_best)
summary(ols_best)
```

```
##
## Call:
## lm(formula = MEDV ~ . - INDUS - AGE, data = data)
##
## 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 ***
## CHAS          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
```

Coefficients Interpretation:

1. **Intercept (36.34145)**: This is the expected value of `MEDV` when all predictors are held at zero. Given the nature of these variables, this might not be a meaningful interpretation since not all predictors are likely to be zero in any practical scenario.

2. **CRIM (-0.10843)**: On average, a unit increase in per capita crime rate by town (`CRIM`) is associated with a decrease in `MEDV` by 0.10843 thousand dollars, holding all other variables constant.

3. **ZN (0.045845)**: A unit increase in the proportion of residential land zoned for lots over 25,000 square feet (`ZN`) is associated with an increase in `MEDV` by 0.045845 thousand dollars, holding other variables constant.

4. **CHAS (2.718716)**: This coefficient suggests that if a property borders the Charles River (`CHAS`), there is, on average, an increase in `MEDV` by about 2.718716 thousand dollars, all else being equal.

5. **NOX (-17.376023)**: An increase of one unit in the nitric oxides concentration (`NOX`) is associated with a decrease in `MEDV` by 17.376023 thousand dollars, controlling for other factors.

6. **RM (3.801579)**: The average number of rooms per dwelling (`RM`) has a strong positive association with `MEDV`. Each additional room is associated with an increase of 3.801579 thousand dollars in `MEDV`, ceteris paribus.

We can interpet the rest of the coefficient using the same reasoning.

Uncertainties (Standard Errors):

The standard error reflects the level of uncertainty regarding the estimated coefficients. A smaller standard error indicates that we can be more confident about the estimate. For example, the standard error for the coefficient of `RM` is relatively small (0.406316), indicating that we can be quite confident that the true effect of the number of rooms on the median value is close to the estimated coefficient of 3.801579 thousand dollars.

Model Summary:

The summary also provides a residual standard error, which is the average amount that the response will deviate from the true regression line (4.736 thousand dollars in this case). The R-squared value is 0.7406, meaning about 74.06% of the variability in `MEDV` can be explained by the model. The adjusted R-squared, which accounts for the number of predictors in the model, is slightly lower at 0.7348, indicating a very good fit. The F-statistic is significant ($p < 2.2e\text{-}16$). This indicates that the overall model is statistically significant.

Point 10

In a multiple linear regression analysis, hypothesis testing is conducted for each individual coefficient to assess whether it is significantly different from zero. The null hypothesis (H0) typically states that the coefficient is equal to zero, implying that the corresponding predictor has no effect on the response variable. The alternative hypothesis (H1) asserts that the coefficient is not equal to zero, indicating a significant effect. If the confidence interval for the coefficient does not contain zero, it supports the rejection of the null hypothesis.

```
alpha = 0.05
confint(ols_best, level = 1-alpha)
```

```
##                     2.5 %       97.5 %
## (Intercept)  26.384649126  46.29764088
## CRIM         -0.172817670  -0.04400902
## ZN            0.019275889   0.07241397
## CHAS          1.040324913   4.39710769
## NOX         -24.321990312 -10.43005655
## RM            3.003258393   4.59989929
## DIS          -1.857631161  -1.12779176
## RAD           0.175037411   0.42417950
## TAX          -0.018403857  -0.00515209
## PTRATIO      -1.200109823  -0.69293932
## B             0.004037216   0.01454447
## LSTAT        -0.615731781  -0.42937513
```

In summary, all predictor variables included in the model are statistically significant and should be considered when modeling the median value of owner-occupied homes in the Boston area.

Point 11

Now we are going to test a group of regressors (the ones identified by the best overall model techniques used above - like BIC, Mallow Cp's , Adjusted R^2) and all the regressors. To do this we are going to implement the ANOVA test.

```
anova = anova(ols_best,ols)
anova$"Pr(>F)"
```

```
## [1]        NA 0.9443416
```

As shown by the high p-value of the ANOVA test, the best overall model identified using the aforementioned techniques is not statistically different from the full model (the one considering all the regressors). This may be caused by the fact that the initial dataset provenance is from a prestigious university, which may have had already cleaned the original data and prepared it for an accurate analysis from the beginning. Alternatively another explanation may be that the high p-value here indicates that the additional parameters in the fuller model do not provide a statistically significant improvement in explaining the variance in the dependent variable compared to the reduced model. This means that the simpler model might be preferable due to its parsimony.

Point 12

To know how well the model fit the data we are going to use two metrics: R^2 and Adjusted R^2:

```
summary_ols_best$r.squared
```

```
## [1] 0.7405823
```

```
summary_ols_best$adj.r.squared
```

```
## [1] 0.7348058
```

The metrics ranges from 0 to 1. The higher the value the better the fit between the model and the observed data. We can see that approximately 73% of the variability in the response can be explained by the predictors.

Point 13

Prediction of the response and associated uncertainty given a new observation with following regressors.

```
new.data = data.frame("CRIM" = 2.05, "ZN" = 28, "INDUS" = 9.69, "CHAS" = 1,
                      "NOX" = 0.40, "RM" = 4, "AGE" = 68.57, "DIS" = 4, "RAD" = 5,
                      "TAX" = 349, "PTRATIO" = 16, "B" = 160, "LSTAT" = 25)
y_pred = predict(ols_best, newdata = new.data,
                 interval = "prediction", level = 0.95)
y_pred
```

```
##        fit     lwr      upr
## 1 13.07217 3.25058 22.89375
```

The predicted MEDV is 13.07 .

Point 14

```r
simulated_data <- data.frame(
  CRIM = data$CRIM,
  ZN = data$ZN,
  INDUS = data$INDUS,
  CHAS = data$CHAS,
  NOX = data$NOX,
  RM = data$RM,
  AGE = data$AGE,
  DIS = data$DIS,
  RAD = data$RAD,
  TAX = data$TAX,
  PTRATIO = data$PTRATIO,
  B = data$B,
  LSTAT = data$LSTAT
)

predicted_MEDV = predict(ols_best, newdata = simulated_data)

plot(data$MEDV, col = "black", pch = 19, xlab = "Index",
     ylab = "MEDV", main = "Original vs Simulated MEDV values",
     ylim = c(0, 50))
points(predicted_MEDV, col = "pink", pch = 19)
legend("topright", legend = c("Original", "Simulated"), col = c("black", "pink"),
       pch = 19)
```