Empirical Project - A Hedonic Housing Model for Midcentury Boston

Khoa Tran & David Jacques

12/1/2020

## Objective

To build a hedonic model to estimate hosuing prices for Boston in 1970, while controlling for structure, neighborhood, accessability, and air pollution

## About the Dataset

Our dataset was originally published by Harrison & Rubenfeld in 1979 in their paper: Hedonic Housing Prices and the Demand for Clean Air. Our dataset differs slightly as it does not contain the PART variable, which is highly correlated to NOX, and our dataset uses a ‘B’ variable as opposed to the original BLK variable to measure proportion of the population that identifies as Black/African American, taking into account market discrimination for neighborhoods that are vast majority white/black.

|  |  |  |
| --- | --- | --- |
| Variable name | Description | Source (Original) |
| Dependent Variable | | |
| MEDV | Median value of owner-occupied homes in $1000's | 1970 U.S. Census |
| Structural Variable | | |
| RM | Average number of rooms per dwelling | 1970 U.S. Census |
| AGE | Proportion of owner-occupied units built prior to 1940 | 1970 U.S. Census |
| Neighborhood Variables | | |
| B | =1000\*(Bk - 0.63)^2 where Bk is the proportion of blacks by town | 1970 U.S. Census |
| LSTAT | % lower status of the population | 1970 U.S. Census |
| CRIM | Crime rate per capita by town | FBI (1970) |
| ZN | Proportion of residential land zoned for lots over 25,000 sqft | Metropolitan Area Planning Commission (1972) |
| INDUS | Proportion of non-retail business acres per town | Vogt, Ivers, and Associates [33] |
| TAX | Full-value property-tax rate per $10,000 | Massachusetts Tax- payers Foundation (1970) |
| PTRATIO | Pupil-teacher ratio by town | Massachusetts Dept. of Education (1971-1972) |
| CHAS | = 1 if tract bounds Charles river; 0 otherwise | 1970 U.S. Census Tract maps |
| Accessibility Variables | | |
| DIS | Weighted distances to five Boston employment centres | Srhnare [29] |
| RAD | Index of accessibility to radial highways | MIT Boston Project |
| Air Pollution Variables | | |
| NOX | Nitric oxides concentration (parts per 10 million) | TASSIM |

data <- read.csv("/Users/davidjacques/Desktop/Empirical Project/realestate.csv")  
#quick look at the dataset  
summary(data)

## CRIM ZN INDUS CHAS   
## Min. : 0.00632 Min. : 0.00 Min. : 0.46 Min. :0.00000   
## 1st Qu.: 0.08232 1st Qu.: 0.00 1st Qu.: 5.19 1st Qu.:0.00000   
## Median : 0.26169 Median : 0.00 Median : 9.69 Median :0.00000   
## Mean : 3.58414 Mean : 11.25 Mean :11.15 Mean :0.06849   
## 3rd Qu.: 3.62118 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.: 45.05 1st Qu.: 2.100   
## Median :0.5380 Median :6.209 Median : 77.30 Median : 3.152   
## Mean :0.5548 Mean :6.288 Mean : 68.62 Mean : 3.784   
## 3rd Qu.:0.6240 3rd Qu.:6.630 3rd Qu.: 94.05 3rd Qu.: 5.118   
## Max. :0.8710 Max. :8.780 Max. :100.00 Max. :12.127   
## NA's :5   
## RAD TAX PTRATIO B   
## Min. : 1.000 Min. :187.0 Min. :12.6 Min. : 0.32   
## 1st Qu.: 4.000 1st Qu.:279.5 1st Qu.:17.4 1st Qu.:374.71   
## Median : 5.000 Median :330.0 Median :19.1 Median :391.34   
## Mean : 9.485 Mean :407.4 Mean :18.5 Mean :356.60   
## 3rd Qu.:24.000 3rd Qu.:666.0 3rd Qu.:20.2 3rd Qu.:396.21   
## Max. :24.000 Max. :711.0 Max. :23.0 Max. :396.90   
##   
## LSTAT MEDV   
## Min. : 1.730 Min. : 5.00   
## 1st Qu.: 7.065 1st Qu.:17.05   
## Median :11.450 Median :21.20   
## Mean :12.880 Mean :22.68   
## 3rd Qu.:17.105 3rd Qu.:25.00   
## Max. :76.000 Max. :67.00   
##

str(data)

## 'data.frame': 511 obs. of 14 variables:  
## $ CRIM : num 0.00632 0.02731 0.02729 0.03237 0.06905 ...  
## $ ZN : num 18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...  
## $ INDUS : num 2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...  
## $ CHAS : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ NOX : num 0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...  
## $ RM : num 6.58 6.42 7.18 7 7.15 ...  
## $ AGE : num 65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...  
## $ DIS : num 4.09 4.97 4.97 6.06 6.06 ...  
## $ RAD : int 1 2 2 3 3 3 5 5 5 5 ...  
## $ TAX : int 296 242 242 222 222 222 311 311 311 311 ...  
## $ PTRATIO: num 15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...  
## $ B : num 397 397 393 395 397 ...  
## $ LSTAT : num 4.98 9.14 4.03 2.94 5.33 ...  
## $ MEDV : num 24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...

#drop all the rows with missing data  
data<-na.omit(data)

## First Regression Models

We begin by running a simple regression model, and determine whether it is more appropriate to have our dependant variable in linear form of logarithmic form.

#first regression  
attach(data)  
ols1<-lm(MEDV~CRIM+ZN+INDUS+CHAS+NOX+RM+AGE+DIS+RAD+TAX+PTRATIO+B+LSTAT,data=data)  
summary(ols1)

##   
## Call:  
## lm(formula = MEDV ~ CRIM + ZN + INDUS + CHAS + NOX + RM + AGE +   
## DIS + RAD + TAX + PTRATIO + B + LSTAT, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -16.889 -2.805 -1.108 1.621 50.555   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 17.358421 6.056032 2.866 0.004331 \*\*   
## CRIM -0.150178 0.040657 -3.694 0.000246 \*\*\*  
## ZN 0.046657 0.017065 2.734 0.006481 \*\*   
## INDUS -0.032804 0.076688 -0.428 0.669016   
## CHAS 3.094171 1.069470 2.893 0.003983 \*\*   
## NOX -16.710092 4.746438 -3.521 0.000471 \*\*\*  
## RM 5.898333 0.479625 12.298 < 2e-16 \*\*\*  
## AGE -0.049467 0.015690 -3.153 0.001716 \*\*   
## DIS -1.702670 0.250146 -6.807 2.92e-11 \*\*\*  
## RAD 0.213582 0.082019 2.604 0.009491 \*\*   
## TAX -0.012069 0.004680 -2.579 0.010205 \*   
## PTRATIO -0.679901 0.158342 -4.294 2.12e-05 \*\*\*  
## B 0.011619 0.003328 3.492 0.000523 \*\*\*  
## LSTAT -0.088308 0.049762 -1.775 0.076577 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.892 on 492 degrees of freedom  
## Multiple R-squared: 0.6268, Adjusted R-squared: 0.617   
## F-statistic: 63.57 on 13 and 492 DF, p-value: < 2.2e-16

#Comparing linear vs log MEDV  
data<-transform(data,MEDV\_log=log(MEDV))  
ols2<-lm(MEDV\_log~CRIM+ZN+INDUS+CHAS+NOX+RM+AGE+DIS+RAD+TAX+PTRATIO+B+LSTAT,data=data)  
summary(ols2)

##   
## Call:  
## lm(formula = MEDV\_log ~ CRIM + ZN + INDUS + CHAS + NOX + RM +   
## AGE + DIS + RAD + TAX + PTRATIO + B + LSTAT, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.78884 -0.10513 -0.02727 0.08653 1.88560   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.3169005 0.2395549 13.846 < 2e-16 \*\*\*  
## CRIM -0.0119873 0.0016082 -7.454 4.12e-13 \*\*\*  
## ZN 0.0011526 0.0006750 1.708 0.088360 .   
## INDUS 0.0002818 0.0030335 0.093 0.926027   
## CHAS 0.1177921 0.0423044 2.784 0.005569 \*\*   
## NOX -0.7333695 0.1877521 -3.906 0.000107 \*\*\*  
## RM 0.1775321 0.0189723 9.357 < 2e-16 \*\*\*  
## AGE -0.0018677 0.0006206 -3.009 0.002752 \*\*   
## DIS -0.0581222 0.0098949 -5.874 7.84e-09 \*\*\*  
## RAD 0.0104854 0.0032444 3.232 0.001312 \*\*   
## TAX -0.0006102 0.0001851 -3.296 0.001051 \*\*   
## PTRATIO -0.0274196 0.0062635 -4.378 1.47e-05 \*\*\*  
## B 0.0005051 0.0001316 3.837 0.000140 \*\*\*  
## LSTAT -0.0111480 0.0019684 -5.664 2.53e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2331 on 492 degrees of freedom  
## Multiple R-squared: 0.6927, Adjusted R-squared: 0.6846   
## F-statistic: 85.32 on 13 and 492 DF, p-value: < 2.2e-16

For both our linear and log models, we see statistical significance for all variables exlcuding INDUS. The log model has a greater R-squared value, signaling that it is the better fitting model for our data, and thus it is the model we will proceed with.

## Further Data Specifications

Next, we introduce further specifications to our data based on economic reasoning and conventions of existing literature.

* The logarithmic form of the LSTAT variable imples that there is greater value to social status in the upper echelons of society, and less so for those in the lower class
* According to traditional theories of urban land rent gradients, DIS and RAD are entered in logarithmic form
* We included a quadratic term RM2 in our model to capture the parabolic relationship between the number of rooms in a house and the price. The reasoning behind this is that the number of rooms in a house may be positively correlated to price until a certain point, at which they become negatively correlated since too many rooms can be seen as burdensome and undesirable

With these specifications in mind, we run our model again to see if these specifications are reflected in our dataset

#Further specifications based on theory  
ols3<-lm(MEDV\_log~CRIM+ZN+INDUS+CHAS+NOX+RM+I(RM^2)+AGE+log(DIS)+log(RAD)+TAX+PTRATIO+B+log(LSTAT),data=data)  
summary(ols3)

##   
## Call:  
## lm(formula = MEDV\_log ~ CRIM + ZN + INDUS + CHAS + NOX + RM +   
## I(RM^2) + AGE + log(DIS) + log(RAD) + TAX + PTRATIO + B +   
## log(LSTAT), data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.76429 -0.09037 -0.01641 0.08349 1.69095   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.726e+00 4.604e-01 12.436 < 2e-16 \*\*\*  
## CRIM -1.291e-02 1.453e-03 -8.880 < 2e-16 \*\*\*  
## ZN 4.005e-05 5.918e-04 0.068 0.946068   
## INDUS 2.049e-04 2.788e-03 0.073 0.941451   
## CHAS 9.581e-02 3.869e-02 2.476 0.013614 \*   
## NOX -7.692e-01 1.778e-01 -4.327 1.83e-05 \*\*\*  
## RM -4.496e-01 1.383e-01 -3.250 0.001232 \*\*   
## I(RM^2) 4.532e-02 1.086e-02 4.172 3.57e-05 \*\*\*  
## AGE -1.150e-03 6.014e-04 -1.912 0.056431 .   
## log(DIS) -2.122e-01 4.074e-02 -5.209 2.80e-07 \*\*\*  
## log(RAD) 7.413e-02 2.243e-02 3.305 0.001020 \*\*   
## TAX -4.791e-04 1.437e-04 -3.335 0.000917 \*\*\*  
## PTRATIO -1.771e-02 5.666e-03 -3.126 0.001875 \*\*   
## B 3.599e-04 1.203e-04 2.991 0.002925 \*\*   
## log(LSTAT) -2.616e-01 2.748e-02 -9.522 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2128 on 491 degrees of freedom  
## Multiple R-squared: 0.7445, Adjusted R-squared: 0.7372   
## F-statistic: 102.2 on 14 and 491 DF, p-value: < 2.2e-16

Taking these adjustments into account, we see an increase in our R-squared values, indicating a better fitting model than our original log model.

## Specifying NOX Variable

Up next, we focus specifically on our NOX variable, which we suspect has a linear relationship with housing price. Based on our economic reasoning, at low levels of Nitric Oxide, the marginal change in NOX should have a relatively mild effect on housing price. One the opposite end of the spectrum, we expect NOX levels to be so high that areas become unsuitable for housing, and thus a decrease in marginal effect. Since we suspect NOX might affect housing prices through the second and third power of NOX, we will compare 3 models (NOX, up to NOX^2 included, and up to NOX^3 included).

#Specificing NOX  
ols4<-lm(MEDV\_log~CRIM+ZN+INDUS+CHAS+NOX+I(NOX^2)+RM+I(RM^2)+AGE+log(DIS)+log(RAD)+TAX+PTRATIO+B+log(LSTAT),data=data)  
summary(ols4)

##   
## Call:  
## lm(formula = MEDV\_log ~ CRIM + ZN + INDUS + CHAS + NOX + I(NOX^2) +   
## RM + I(RM^2) + AGE + log(DIS) + log(RAD) + TAX + PTRATIO +   
## B + log(LSTAT), data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.76558 -0.09169 -0.01649 0.07554 1.67495   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.1845163 0.5654694 9.169 < 2e-16 \*\*\*  
## CRIM -0.0128111 0.0014521 -8.822 < 2e-16 \*\*\*  
## ZN 0.0003789 0.0006257 0.606 0.545080   
## INDUS 0.0001475 0.0027837 0.053 0.957766   
## CHAS 0.1032386 0.0388875 2.655 0.008194 \*\*   
## NOX 1.2140009 1.2197099 0.995 0.320072   
## I(NOX^2) -1.4888235 0.9059381 -1.643 0.100941   
## RM -0.4586605 0.1381933 -3.319 0.000971 \*\*\*  
## I(RM^2) 0.0457227 0.0108464 4.215 2.97e-05 \*\*\*  
## AGE -0.0014373 0.0006253 -2.299 0.021950 \*   
## log(DIS) -0.1878421 0.0432860 -4.340 1.74e-05 \*\*\*  
## log(RAD) 0.0698104 0.0225462 3.096 0.002072 \*\*   
## TAX -0.0004896 0.0001436 -3.410 0.000702 \*\*\*  
## PTRATIO -0.0198167 0.0057991 -3.417 0.000685 \*\*\*  
## B 0.0003534 0.0001202 2.940 0.003435 \*\*   
## log(LSTAT) -0.2614673 0.0274300 -9.532 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2124 on 490 degrees of freedom  
## Multiple R-squared: 0.7459, Adjusted R-squared: 0.7381   
## F-statistic: 95.89 on 15 and 490 DF, p-value: < 2.2e-16

ols5<-lm(MEDV\_log~CRIM+ZN+INDUS+CHAS+NOX+I(NOX^2)+I(NOX^3)+RM+I(RM^2)+AGE+log(DIS)+log(RAD)+TAX+PTRATIO+B+log(LSTAT),data=data)  
summary(ols5)

##   
## Call:  
## lm(formula = MEDV\_log ~ CRIM + ZN + INDUS + CHAS + NOX + I(NOX^2) +   
## I(NOX^3) + RM + I(RM^2) + AGE + log(DIS) + log(RAD) + TAX +   
## PTRATIO + B + log(LSTAT), data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.74324 -0.09507 -0.00716 0.08081 1.64842   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.990e-01 1.574e+00 0.635 0.525978   
## CRIM -1.218e-02 1.458e-03 -8.353 6.94e-16 \*\*\*  
## ZN 1.301e-03 7.007e-04 1.857 0.063848 .   
## INDUS 6.079e-04 2.769e-03 0.220 0.826305   
## CHAS 9.791e-02 3.865e-02 2.533 0.011625 \*   
## NOX 2.207e+01 7.428e+00 2.972 0.003109 \*\*   
## I(NOX^2) -3.574e+01 1.207e+01 -2.962 0.003209 \*\*   
## I(NOX^3) 1.815e+01 6.376e+00 2.846 0.004611 \*\*   
## RM -4.751e-01 1.373e-01 -3.460 0.000587 \*\*\*  
## I(RM^2) 4.761e-02 1.079e-02 4.413 1.25e-05 \*\*\*  
## AGE -1.579e-03 6.228e-04 -2.535 0.011570 \*   
## log(DIS) -1.778e-01 4.312e-02 -4.123 4.40e-05 \*\*\*  
## log(RAD) 7.098e-02 2.239e-02 3.170 0.001617 \*\*   
## TAX -3.614e-04 1.495e-04 -2.418 0.015986 \*   
## PTRATIO -1.866e-02 5.772e-03 -3.232 0.001310 \*\*   
## B 3.734e-04 1.196e-04 3.123 0.001895 \*\*   
## log(LSTAT) -2.555e-01 2.732e-02 -9.352 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2109 on 489 degrees of freedom  
## Multiple R-squared: 0.75, Adjusted R-squared: 0.7419   
## F-statistic: 91.7 on 16 and 489 DF, p-value: < 2.2e-16

Observing these 3 models, we conclude that the model with up to the third power of NOX included is the best fitting model based on the high adjusted R-squared value. Moreover, all variables except INDUS are statistically significant.

## Relaxing INDUS Control

Since the INDUS variable is highly insignificant, we decided to relax this control and observe if this drastically changes our estimates.

#relaxing INDUS  
ols6<-lm(MEDV\_log~CRIM+ZN+CHAS+NOX+I(NOX^2)+I(NOX^3)+RM+I(RM^2)+AGE+log(DIS)+log(RAD)+TAX+PTRATIO+B+log(LSTAT),data=data)  
summary(ols6)

##   
## Call:  
## lm(formula = MEDV\_log ~ CRIM + ZN + CHAS + NOX + I(NOX^2) + I(NOX^3) +   
## RM + I(RM^2) + AGE + log(DIS) + log(RAD) + TAX + PTRATIO +   
## B + log(LSTAT), data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.74453 -0.09389 -0.00637 0.08155 1.64770   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.012e+00 1.572e+00 0.644 0.520032   
## CRIM -1.221e-02 1.450e-03 -8.425 4.03e-16 \*\*\*  
## ZN 1.278e-03 6.919e-04 1.847 0.065306 .   
## CHAS 9.880e-02 3.840e-02 2.573 0.010381 \*   
## NOX 2.199e+01 7.411e+00 2.967 0.003154 \*\*   
## I(NOX^2) -3.558e+01 1.203e+01 -2.957 0.003259 \*\*   
## I(NOX^3) 1.806e+01 6.359e+00 2.841 0.004686 \*\*   
## RM -4.727e-01 1.368e-01 -3.457 0.000594 \*\*\*  
## I(RM^2) 4.738e-02 1.073e-02 4.417 1.23e-05 \*\*\*  
## AGE -1.579e-03 6.222e-04 -2.538 0.011452 \*   
## log(DIS) -1.801e-01 4.172e-02 -4.318 1.90e-05 \*\*\*  
## log(RAD) 6.986e-02 2.178e-02 3.207 0.001427 \*\*   
## TAX -3.491e-04 1.384e-04 -2.522 0.011996 \*   
## PTRATIO -1.857e-02 5.752e-03 -3.228 0.001330 \*\*   
## B 3.724e-04 1.193e-04 3.120 0.001915 \*\*   
## log(LSTAT) -2.553e-01 2.728e-02 -9.359 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2107 on 490 degrees of freedom  
## Multiple R-squared: 0.75, Adjusted R-squared: 0.7424   
## F-statistic: 98 on 15 and 490 DF, p-value: < 2.2e-16

stargazer :: stargazer(ols5,ols6,type="text",title="Table. Relaxing INDUS")

##   
## Table. Relaxing INDUS  
## =====================================================================  
## Dependent variable:   
## -------------------------------------------------  
## MEDV\_log   
## (1) (2)   
## ---------------------------------------------------------------------  
## CRIM -0.012\*\*\* -0.012\*\*\*   
## (0.001) (0.001)   
##   
## ZN 0.001\* 0.001\*   
## (0.001) (0.001)   
##   
## INDUS 0.001   
## (0.003)   
##   
## CHAS 0.098\*\* 0.099\*\*   
## (0.039) (0.038)   
##   
## NOX 22.073\*\*\* 21.989\*\*\*   
## (7.428) (7.411)   
##   
## I(NOX2) -35.735\*\*\* -35.584\*\*\*   
## (12.066) (12.035)   
##   
## I(NOX3) 18.146\*\*\* 18.065\*\*\*   
## (6.376) (6.359)   
##   
## RM -0.475\*\*\* -0.473\*\*\*   
## (0.137) (0.137)   
##   
## I(RM2) 0.048\*\*\* 0.047\*\*\*   
## (0.011) (0.011)   
##   
## AGE -0.002\*\* -0.002\*\*   
## (0.001) (0.001)   
##   
## log(DIS) -0.178\*\*\* -0.180\*\*\*   
## (0.043) (0.042)   
##   
## log(RAD) 0.071\*\*\* 0.070\*\*\*   
## (0.022) (0.022)   
##   
## TAX -0.0004\*\* -0.0003\*\*   
## (0.0001) (0.0001)   
##   
## PTRATIO -0.019\*\*\* -0.019\*\*\*   
## (0.006) (0.006)   
##   
## B 0.0004\*\*\* 0.0004\*\*\*   
## (0.0001) (0.0001)   
##   
## log(LSTAT) -0.255\*\*\* -0.255\*\*\*   
## (0.027) (0.027)   
##   
## Constant 0.999 1.012   
## (1.574) (1.572)   
##   
## ---------------------------------------------------------------------  
## Observations 506 506   
## R2 0.750 0.750   
## Adjusted R2 0.742 0.742   
## Residual Std. Error 0.211 (df = 489) 0.211 (df = 490)   
## F Statistic 91.703\*\*\* (df = 16; 489) 98.004\*\*\* (df = 15; 490)  
## =====================================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

We see that all the remaining variables are significant with little change from the previous model, and the adjusted R-squared remains the same. Thus, we can safely reject the INDUS variable from our model, and have arrived at our final model specification.

## Endogeneity & Heteroskedasticity

After having specified our model, we consider the issues of endogenteity and heteroskedasticity.

Our model assumes that all independent variables are exogenous, so to the best of our knowledge of our data, this condition is satisfied.

Regarding heteroskedasticity, since our data is from a census tract and not individuals, we suspect that there be some degree of heteroskedasticity.

## Breusch-Pagan Test for Heteroskedasticity

#Breusch-Pagan test for heteroskedasticity  
library(lmtest)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

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

bptest(ols6, data=data)

##   
## studentized Breusch-Pagan test  
##   
## data: ols6  
## BP = 68.181, df = 15, p-value = 9.4e-09

Since p-value < 0.01, the error is likely to be heteroskedastic

## White’s Test for Heteroskedasticity

#white's test for heteroskedasticity  
data$residualSQ<-ols6$residuals^2  
data$fit<-ols6$fitted.values  
WhiteOLS<-lm(residualSQ ~fit+I(fit^2), data=data)  
# k+1: the number of explanatory variables + 1 (for intercept)  
Wkplus1<-nrow(summary(WhiteOLS)$coef)  
# obsn: the number of observations  
Wobsn<-WhiteOLS$df.residual+Wkplus1  
Wobsn

## [1] 506

# LM statistics for White test  
WhiteLM<-Wobsn\*summary(WhiteOLS)$r.squared   
WhiteLM

## [1] 17.61794

# Calculating p-value from Chi-squared distribution  
pchisq(WhiteLM, df=2, lower.tail=FALSE)

## [1] 0.0001493874

Based on White’s test, we can also reject the null hypothesis and reach the same conclusion as the BP-test; there is heteroskedasticity in the error of our model. Since it is detectable by the BP-test, this is likely a linear relationship within our error.

## Introducing Robust Standard Error

#robust error  
# HC0 option is for textbook white robust standard error  
library(sandwich)  
robustse<-coeftest(ols6, vcov = vcovHC(ols6, "HC0"))  
robustse

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.0117e+00 1.5150e+00 0.6678 0.5045826   
## CRIM -1.2215e-02 2.4104e-03 -5.0673 5.726e-07 \*\*\*  
## ZN 1.2782e-03 5.0046e-04 2.5540 0.0109495 \*   
## CHAS 9.8800e-02 4.0911e-02 2.4150 0.0161009 \*   
## NOX 2.1989e+01 6.6604e+00 3.3015 0.0010317 \*\*   
## I(NOX^2) -3.5584e+01 1.0776e+01 -3.3022 0.0010293 \*\*   
## I(NOX^3) 1.8065e+01 5.6442e+00 3.2006 0.0014604 \*\*   
## RM -4.7275e-01 1.8350e-01 -2.5763 0.0102783 \*   
## I(RM^2) 4.7383e-02 1.4746e-02 3.2133 0.0013986 \*\*   
## AGE -1.5793e-03 1.1769e-03 -1.3419 0.1802544   
## log(DIS) -1.8014e-01 5.0338e-02 -3.5787 0.0003797 \*\*\*  
## log(RAD) 6.9865e-02 2.3478e-02 2.9758 0.0030669 \*\*   
## TAX -3.4906e-04 1.3937e-04 -2.5045 0.0125859 \*   
## PTRATIO -1.8569e-02 7.1761e-03 -2.5876 0.0099518 \*\*   
## B 3.7236e-04 1.4764e-04 2.5220 0.0119839 \*   
## log(LSTAT) -2.5527e-01 6.7836e-02 -3.7631 0.0001881 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Introducing FWLS/FGLS Method

## FWLS/FGLS  
#step 1. save u\_hat from the original OLS model  
u\_hat<-ols6$residuals  
#step 2. generate log u\_hat  
lu\_hat\_sq<-log(u\_hat^2)  
#step 3. regress log u\_hat on x  
ols62<-lm(lu\_hat\_sq~ CRIM+ZN+CHAS+NOX+I(NOX^2)+I(NOX^3)+RM+I(RM^2)+AGE+log(DIS)+log(RAD)+TAX+PTRATIO+B+log(LSTAT), data=data)  
fv<-ols62$fitted.values  
#step 4. exponentiate the fitted values  
efv<-exp(fv)  
#step 5. WLS  
fwls<-lm(MEDV\_log~CRIM+ZN+CHAS+NOX+I(NOX^2)+I(NOX^3)+RM+I(RM^2)+AGE+log(DIS)+log(RAD)+TAX+PTRATIO+B+log(LSTAT), weights=1/(efv), data=data)  
summary(fwls)

##   
## Call:  
## lm(formula = MEDV\_log ~ CRIM + ZN + CHAS + NOX + I(NOX^2) + I(NOX^3) +   
## RM + I(RM^2) + AGE + log(DIS) + log(RAD) + TAX + PTRATIO +   
## B + log(LSTAT), data = data, weights = 1/(efv))  
##   
## Weighted Residuals:  
## Min 1Q Median 3Q Max   
## -7.1487 -1.2649 -0.1702 1.1057 17.3944   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.404e+00 1.240e+00 1.939 0.053079 .   
## CRIM -1.626e-02 2.977e-03 -5.462 7.48e-08 \*\*\*  
## ZN 5.472e-04 4.031e-04 1.358 0.175232   
## CHAS 5.571e-02 3.248e-02 1.715 0.086909 .   
## NOX 9.912e+00 5.620e+00 1.764 0.078402 .   
## I(NOX^2) -1.588e+01 9.398e+00 -1.689 0.091782 .   
## I(NOX^3) 7.850e+00 5.055e+00 1.553 0.121076   
## RM -3.437e-01 1.499e-01 -2.293 0.022252 \*   
## I(RM^2) 4.468e-02 1.123e-02 3.979 7.97e-05 \*\*\*  
## AGE -1.849e-03 3.742e-04 -4.941 1.07e-06 \*\*\*  
## log(DIS) -1.580e-01 2.812e-02 -5.619 3.23e-08 \*\*\*  
## log(RAD) 4.244e-02 1.543e-02 2.750 0.006171 \*\*   
## TAX -4.393e-04 1.068e-04 -4.115 4.55e-05 \*\*\*  
## PTRATIO -1.900e-02 3.942e-03 -4.820 1.92e-06 \*\*\*  
## B 5.030e-04 1.471e-04 3.418 0.000683 \*\*\*  
## log(LSTAT) -1.217e-01 2.224e-02 -5.475 6.99e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.272 on 490 degrees of freedom  
## Multiple R-squared: 0.7922, Adjusted R-squared: 0.7858   
## F-statistic: 124.5 on 15 and 490 DF, p-value: < 2.2e-16

Observing the tests above, we can see the heteroskedasticity issue has seriously affected our model. After taking these issues into consideration, AGE, ZN, I(NOX^3) all appear to be insignificant.

## Conclusion

Our model expands on the existing work completed by Harrison and Rubenfeld by incorporating the effects of marginal changes in NOX, diminishing returns of more rooms, and marginal impact of socioeconomic status for those at the upper and lower ends of society. While our model is well fitted to the data available, it is susceptible to heteroskedasticity. This could be improved by having more specific observations. Overall, our model provides a good interpretation for the effects these variables have on housing prices.