University of Massachusetts Amherst

Analyzing Housing Data in Boston

Presented by Om, Kirat, John

Objectives of our study

01. Develop a model that predicts the median price of a house in an area

02. Perform exploratory data analysis about features pertaining to crime rate

What are we working with?

The Boston Housing Dataset[1]

- 506 records
- 13 input features
- 1 output feature
- Data drawn from the Boston Standard Metropolitan Statistical Area (SMSA) in 1970
- Each data point represents a census tract in the Boston SMSA for owner-occupied one-family houses

Input features

01 CRIM

the per capita crime rate, by town

02 ZN

the proportion of residential land zoned for lots over 25,000 sq. ft

03 INDUS

the proportion of non-retail business acres per town

04 CHAS

Charles River dummy variable (1 if tract bounds river; 0 otherwise)

Input features

05 NOX

nitric oxides concentration (parts per 10 million)

06 RM

The average number of rooms per dwelling

07 AGE

the proportion of owner-occupied units built prior to 1940

08 DIS

weighted distances to five Boston employment centers

Input features

09 RAD

index of accessibility to radial highways

10 TAX

full-value property-tax rate per $10,000

11 B

1000(Bk - 0.63)^2, Where Bk is the proportion of blacks by the town*

12 PTRATIO

the pupil-teacher ratio by the town

Input features

13 LSTAT

% lower status
of the
population

Output feature

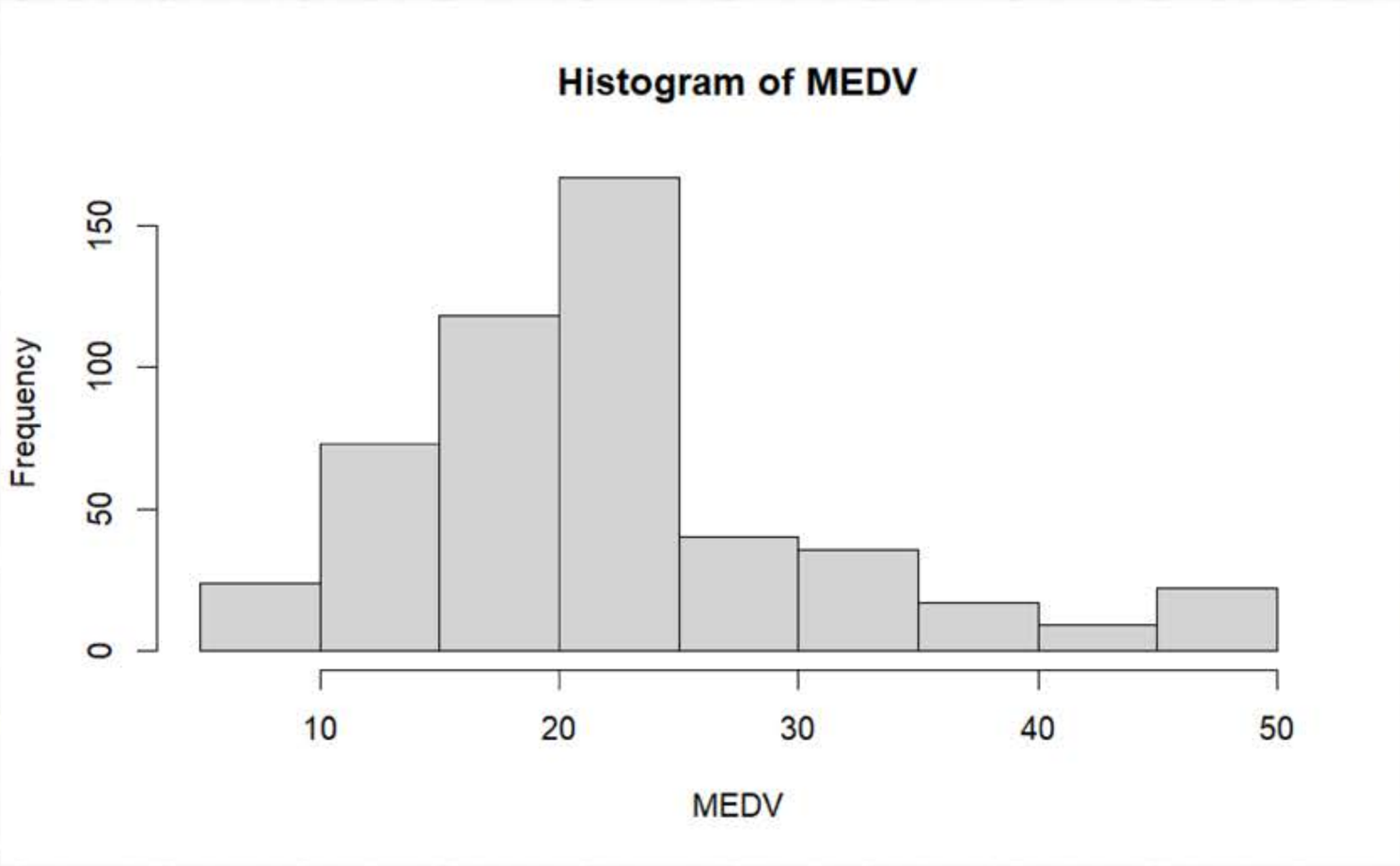
14 MEDV

the median value of owner-occupied homes in $1000s

A note

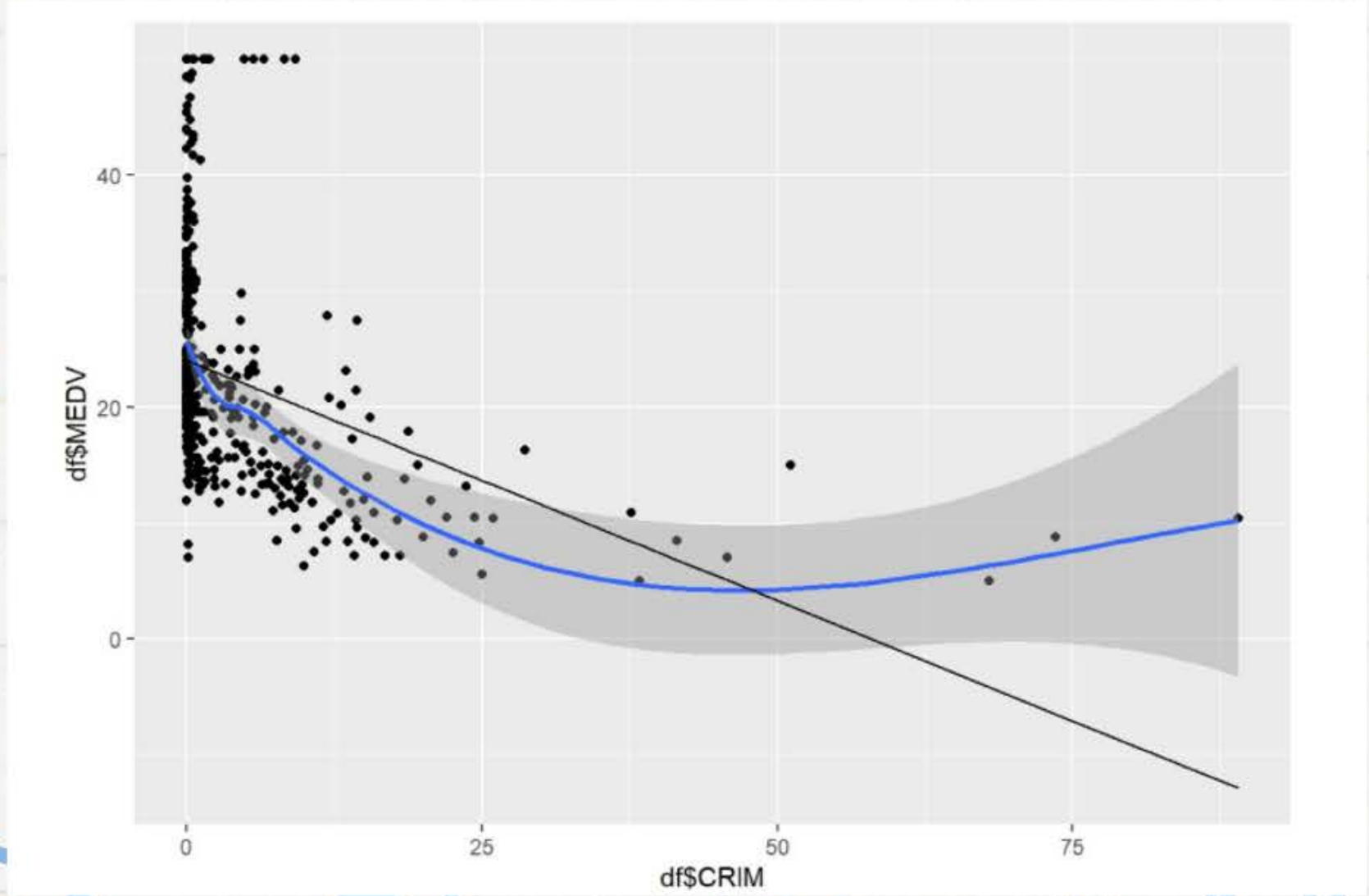
This dataset was developed on the basis of data collected in 1970. Some variables, like input feature#11 pose an ethical problem: as investigated in [2], “B” is a non-invertible variable that was engineered by the authors of this dataset by assuming that racial self-segregation has a positive impact on house prices. We will not make conclusions/inferences about said feature.

Preprocessing and data exploration



Although there is no explicit mention of this, we suspect that MEDV has been cut-off at 50

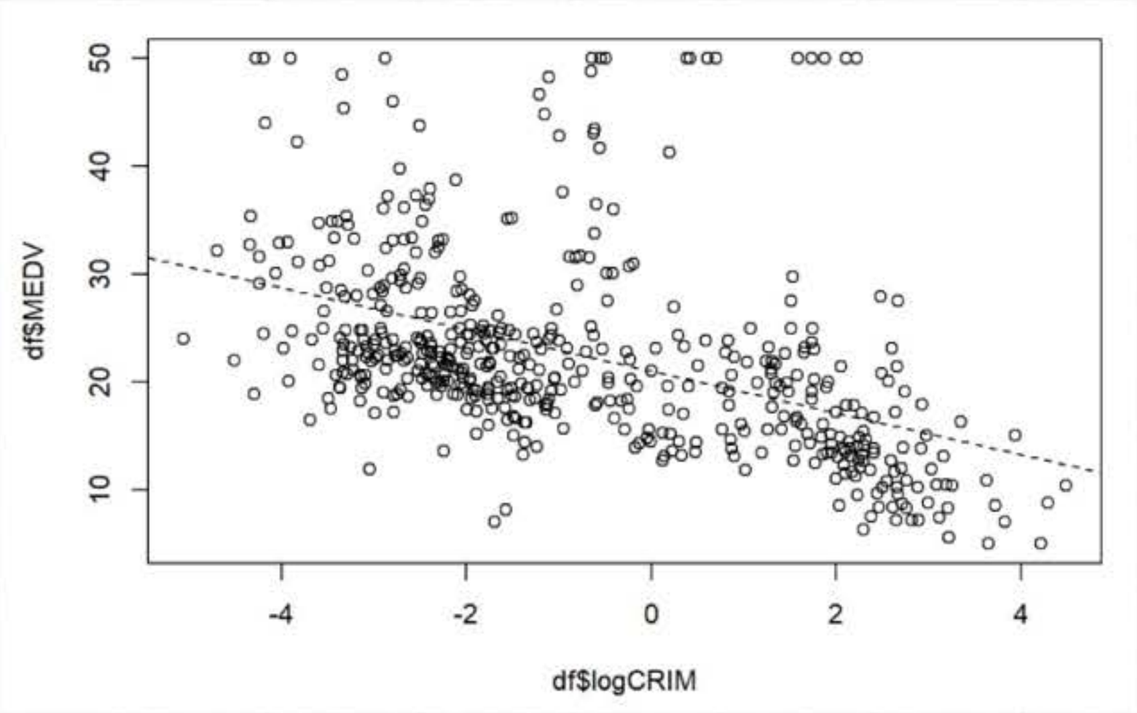
Transforming CRIM



- non-linear relationship
- there are too many points with high CRIM values
- R^2 is relatively low (0.1491): the model doesn't fit well

Transforming CRIM

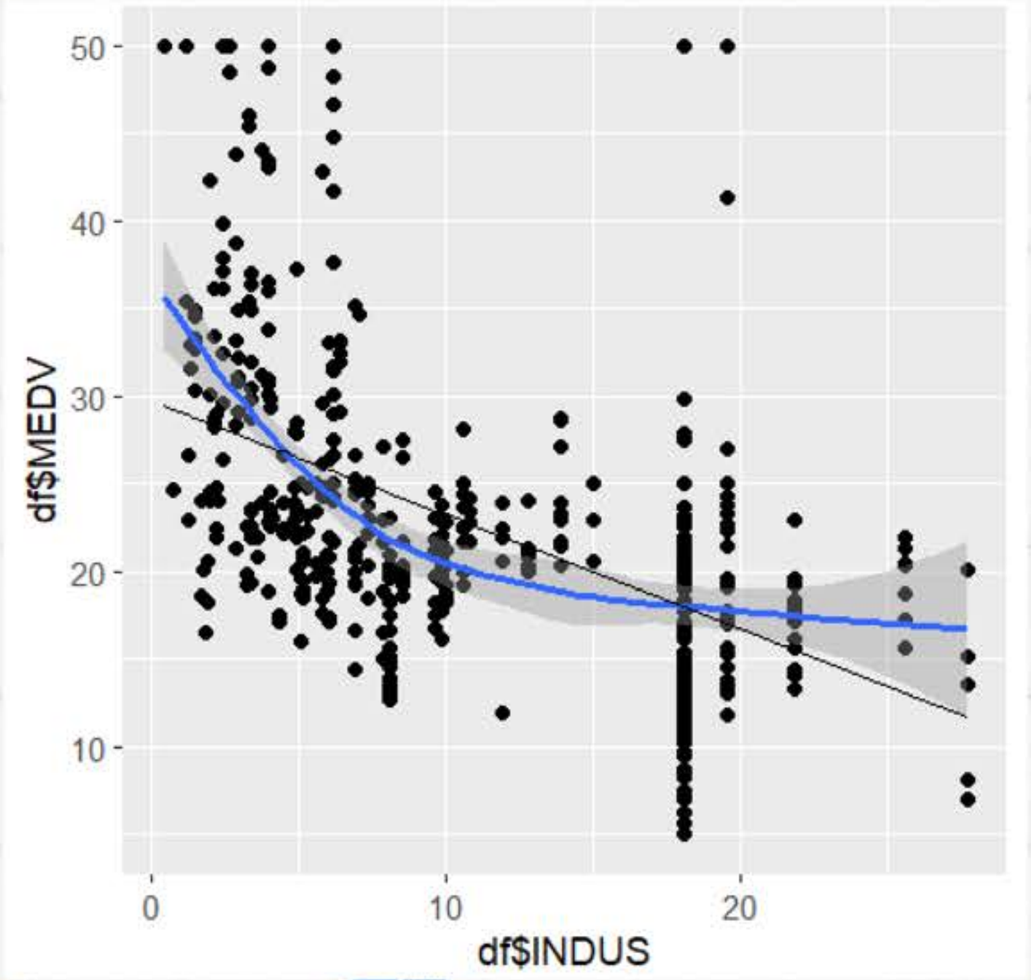
We deal with this problem by using log(CRIM) instead of CRIM.

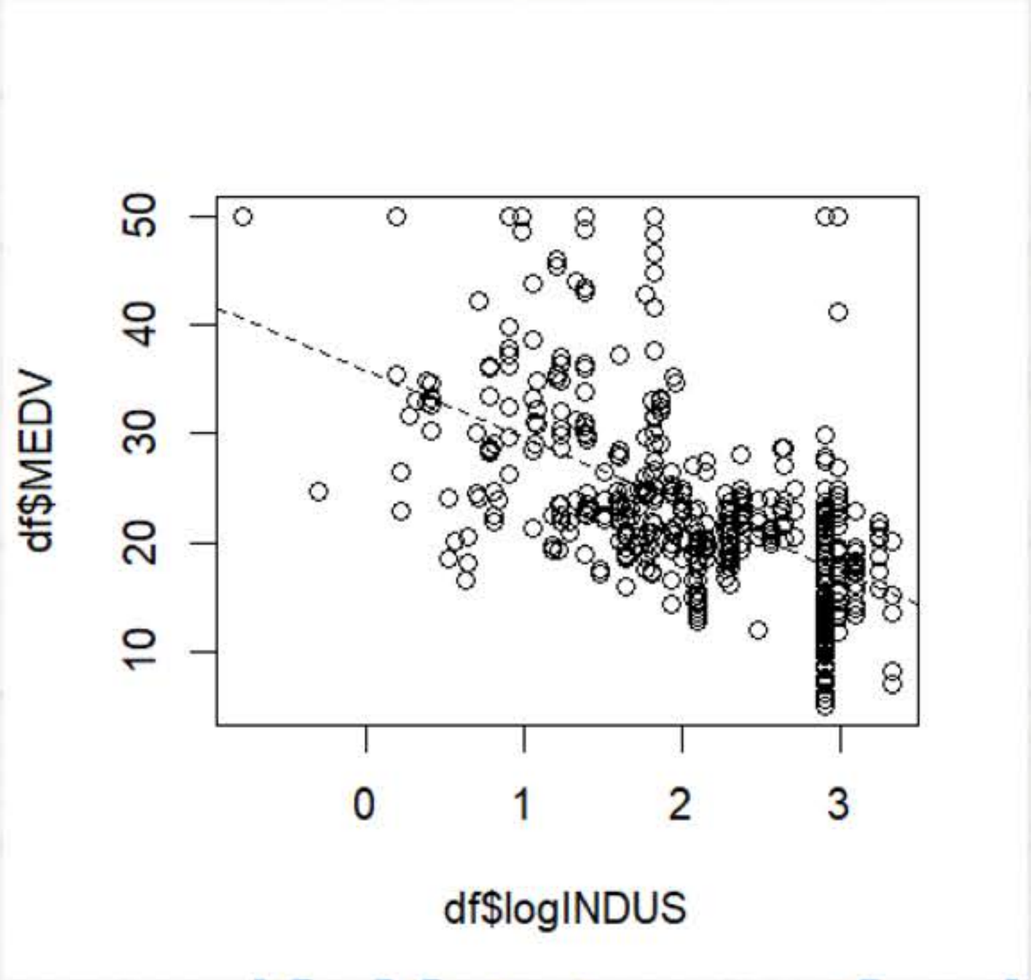



```
##
## Call:
## lm(formula = MEDV ~ logCRIM, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.303  -5.159  -2.427   2.666  33.271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  21.0246     0.3877   54.23   <2e-16 ***
## logCRIM      -1.9325     0.1688  -11.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.201 on 504 degrees of freedom
## Multiple R-squared:  0.2064, Adjusted R-squared:  0.2048
## F-statistic: 131.1 on 1 and 504 DF,  p-value: < 2.2e-16
```

Transforming INDUS

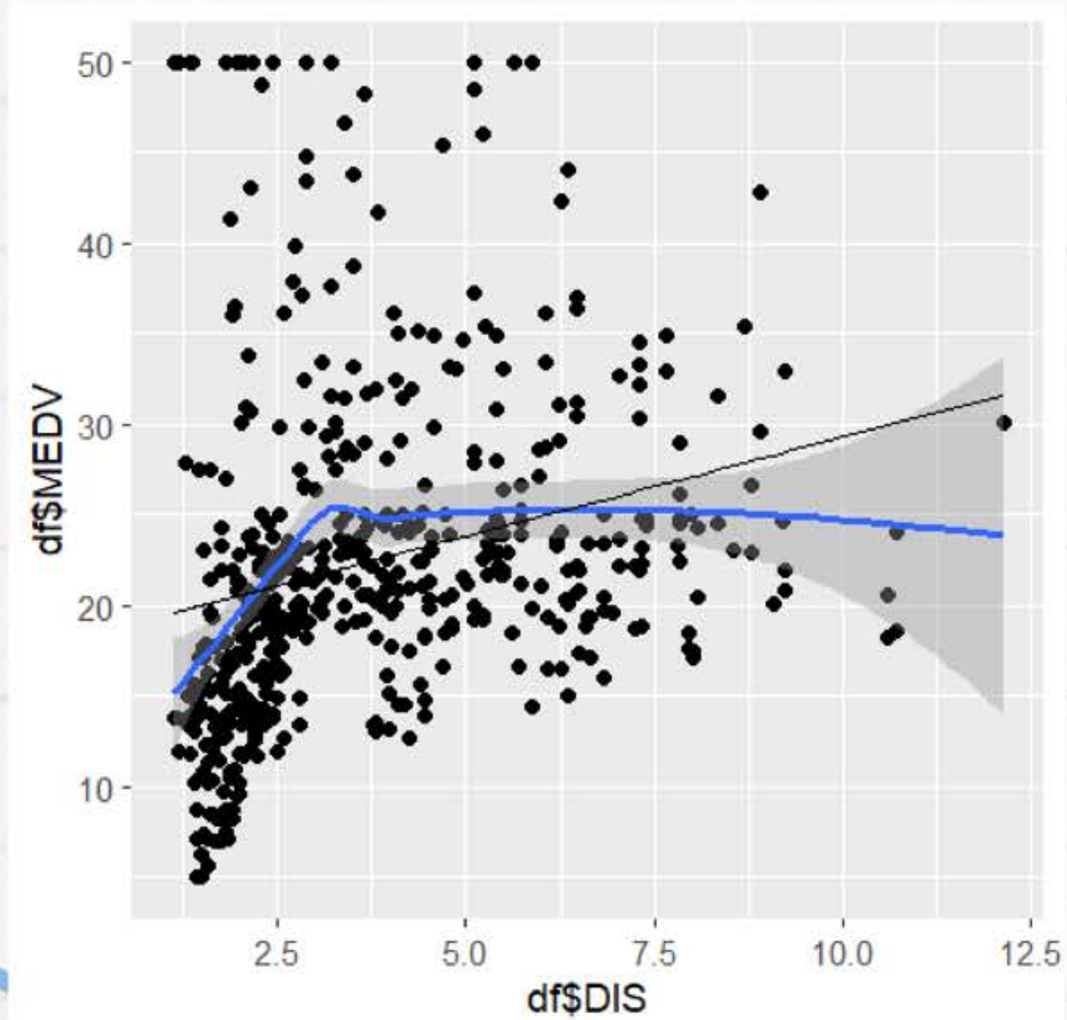
For similar reasons, we utilise log(INDUS) instead of INDUS.

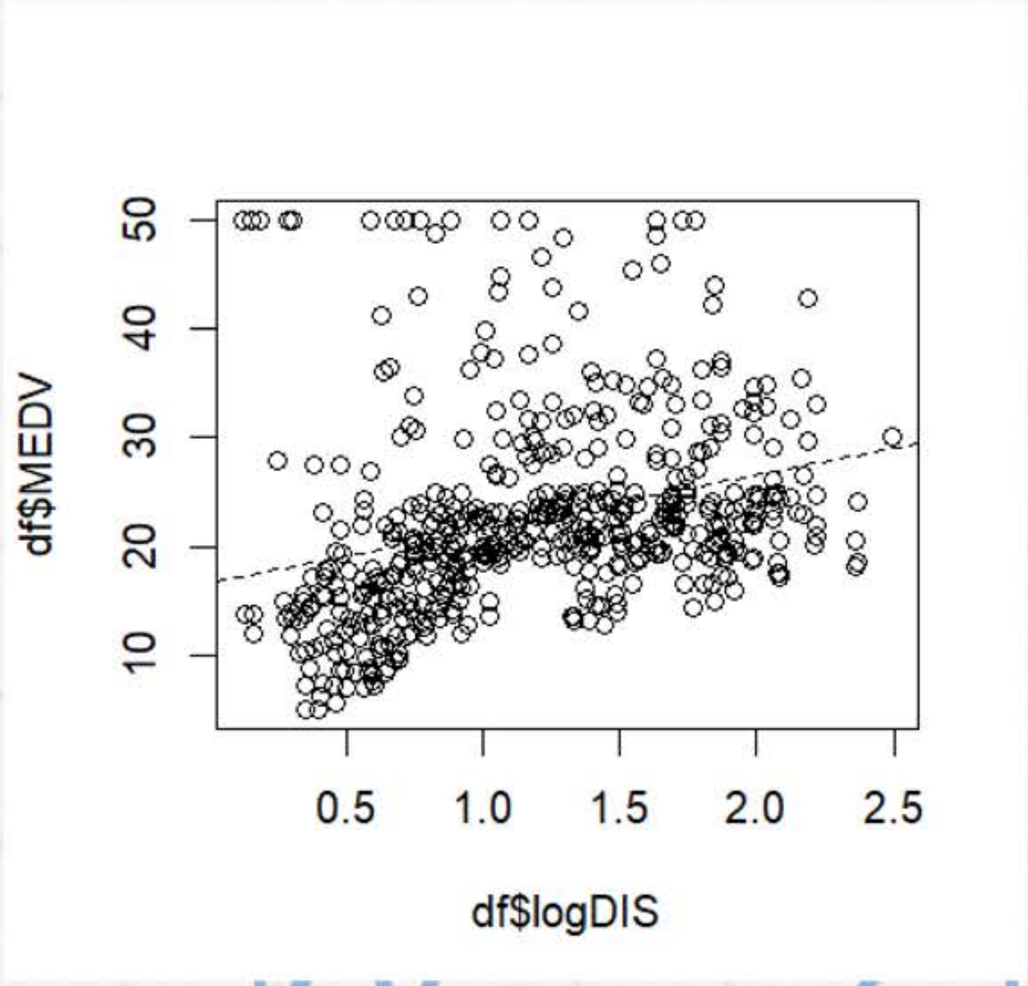




Transforming DIS

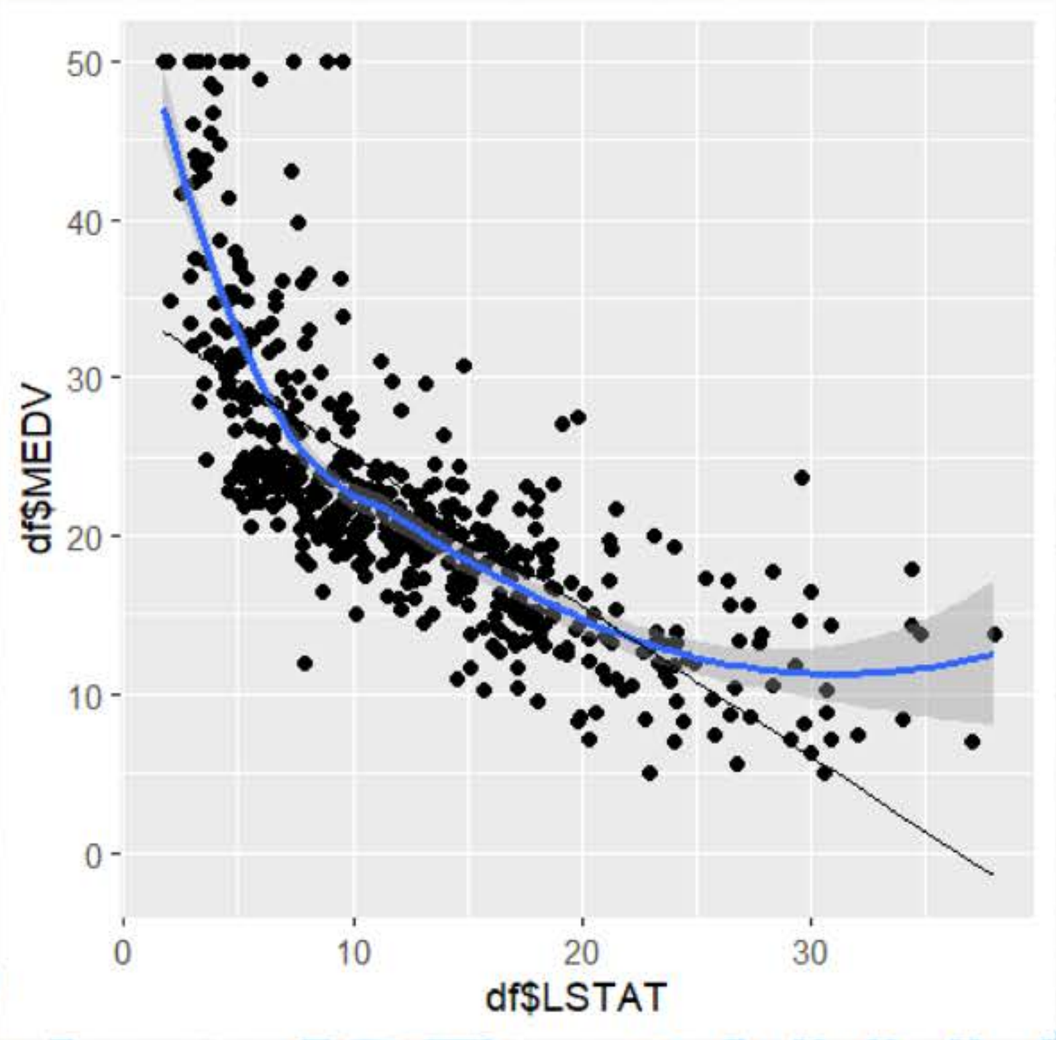
For similar reasons, we utilise log(DIS) instead of DIS

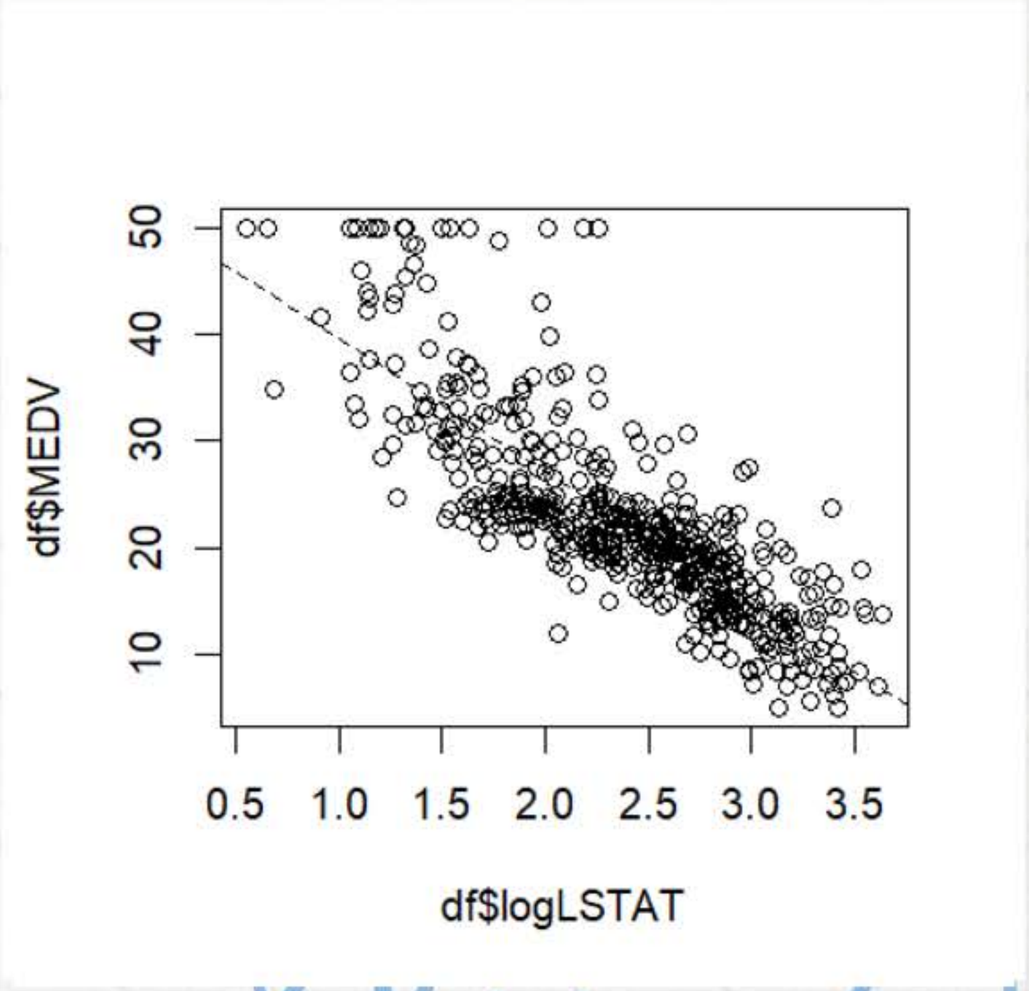




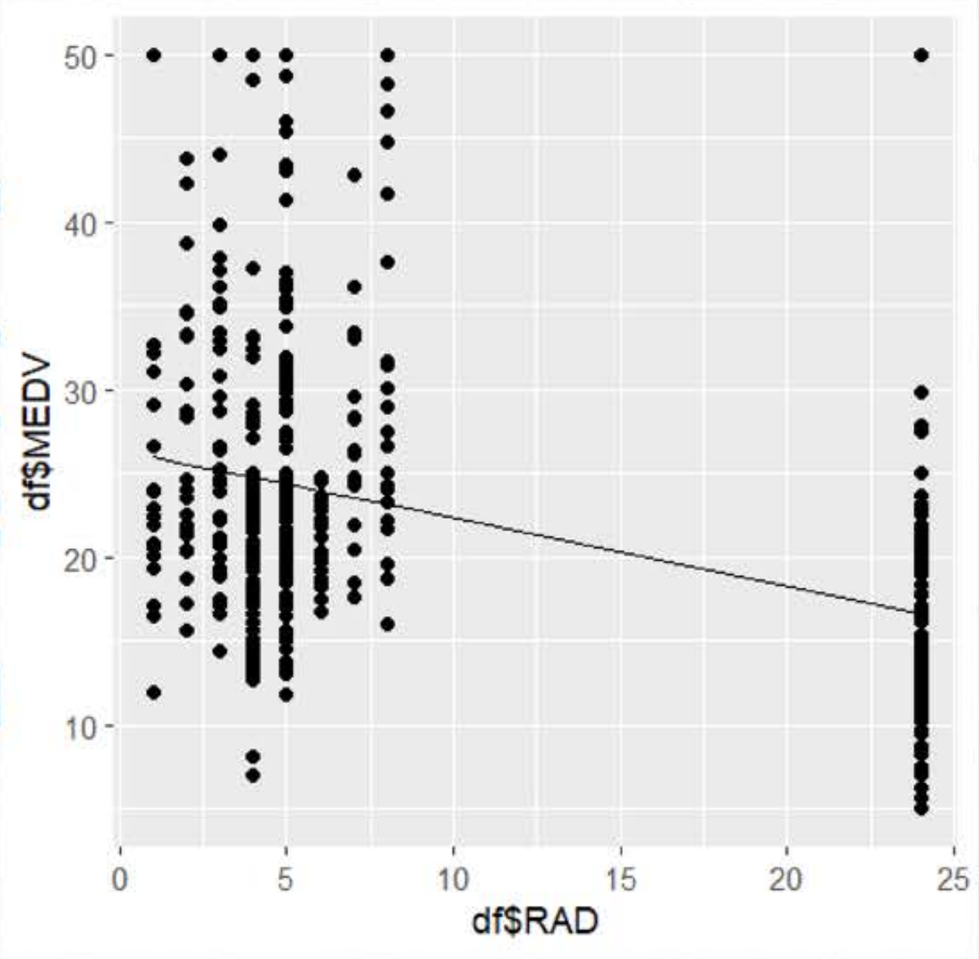
Transforming LSTAT

For similar reasons, we utilise log(LSTAT) instead of LSTAT

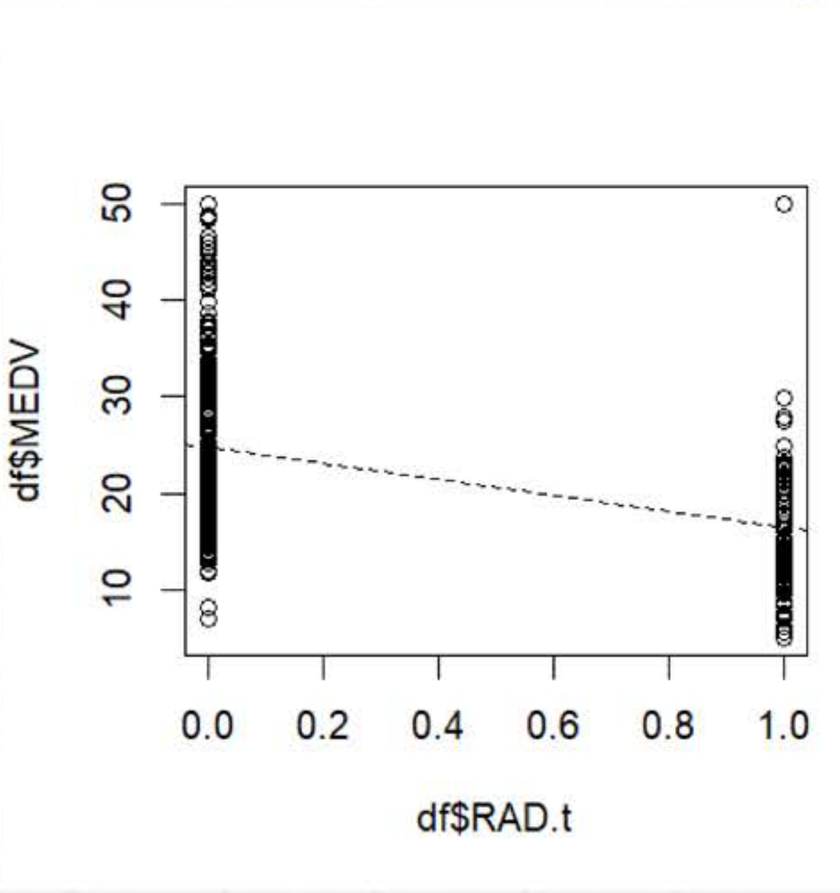




Transforming RAD



- a lot of “within-group” variance
- will lead to bad predictions
- Entries with high RAD make up ~25% of our dataset
- Cannot be ignored!
- We encode RAD as a dummy variable: when RAD = 24, RAD.t = 1



Preliminary work

- analyzed correlation matrix and excluded NOX from every model, it was found to be correlated with logDIS (|correlation| > 0.8)
- makes sense: concentration of NOX in the air decreases as the distance from an employment center increases
- utilized a 80% - 20% training - test split for predictive models

Models

Model 1

MEDV ~ all untransformed variables

R^2 : 0.7119
training error (MSE) : 24.27
test error (MSE) : 20.93

```
Residuals:
     Min       1Q   Median       3Q      Max
-16.9259  -3.0007  -0.6438   1.8681  27.6160

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.257083   4.828341   5.438 9.48e-08 ***
CRIM        -0.089464   0.043782  -2.043  0.04168 *
ZN           0.048421   0.016129   3.002  0.00285 **
INDUS       -0.094590   0.070326  -1.345  0.17940
CHAS         2.693266   1.017172   2.648  0.00843 **
RM           4.005595   0.485520   8.250 2.41e-15 ***
AGE         -0.009707   0.015191  -0.639  0.52319
DIS         -1.139605   0.220153  -5.176 3.62e-07 ***
RAD          0.254382   0.081281   3.130  0.00188 **
TAX         -0.014746   0.004608  -3.200  0.00149 **
PTRATIO     -0.671910   0.148965  -4.511 8.55e-06 ***
LSTAT       -0.568460   0.059334  -9.581  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.001 on 393 degrees of freedom
Multiple R-squared:  0.7198,    Adjusted R-squared:  0.7119
F-statistic: 91.77 on 11 and 393 DF,  p-value: < 2.2e-16
```

Model 2

MEDV ~ all untransformed variables that are significant

R^2 : 0.7117
training error (MSE) : 24.42
test error (MSE) : 20.88

```
Residuals:
     Min       1Q   Median       3Q      Max
-16.9481  -3.0915  -0.6576   1.8787  27.2456

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.414385   4.778433   5.319 1.75e-07 ***
CRIM        -0.085866   0.043737  -1.963 0.050323 .
ZN           0.053498   0.015685   3.411 0.000715 ***
CHAS         2.455485   1.005613   2.442 0.015053 *
RM           4.011499   0.476599   8.417 7.19e-16 ***
DIS         -0.982761   0.189899  -5.175 3.63e-07 ***
RAD          0.287228   0.078128   3.676 0.000269 ***
TAX         -0.018017   0.004013  -4.489 9.40e-06 ***
PTRATIO     -0.684421   0.148270  -4.616 5.30e-06 ***
LSTAT       -0.593845   0.054461 -10.904  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.003 on 395 degrees of freedom
Multiple R-squared:  0.7181,    Adjusted R-squared:  0.7117
F-statistic: 111.8 on 9 and 395 DF,  p-value: < 2.2e-16
```

Model 3

MEDV ~ all transformed variables

R^2 : 0.7648
training error (MSE) : 19.87
test error (MSE) : 17.19

```
Residuals:
     Min       1Q   Median       3Q      Max
-16.0441  -2.5334  -0.3072   2.0251  25.0982

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 50.332705   5.078302   9.911  < 2e-16 ***
ZN          -0.007006   0.014708  -0.476  0.63407
CHAS         2.404570   0.923337   2.604  0.00956 **
RM           2.789881   0.463543   6.019 4.03e-09 ***
AGE          0.001449   0.014557   0.100  0.92078
TAX         -0.010357   0.003892  -2.661  0.00810 **
PTRATIO     -0.580564   0.137353  -4.227 2.95e-05 ***
logINDUS    -1.625644   0.559816  -2.904  0.00389 **
RAD.t        2.517417   1.350327   1.864  0.06302 .
logDIS      -4.276004   0.823230  -5.194 3.30e-07 ***
logLSTAT    -9.507501   0.679755 -13.987  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.519 on 394 degrees of freedom
Multiple R-squared:  0.7706,    Adjusted R-squared:  0.7648
F-statistic: 132.4 on 10 and 394 DF,  p-value: < 2.2e-16
```

Model 4

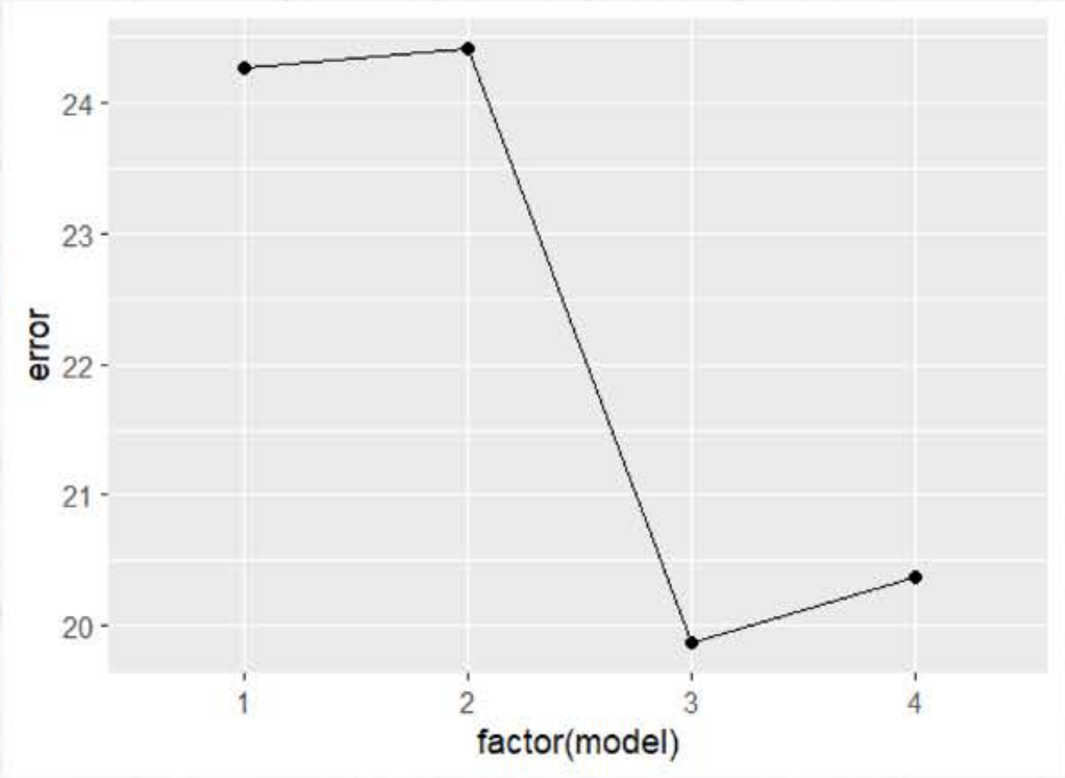
MEDV ~ all transformed variables that are significant

R^2 : 0.7613
training error (MSE) : 20.37
test error (MSE) : 18.17

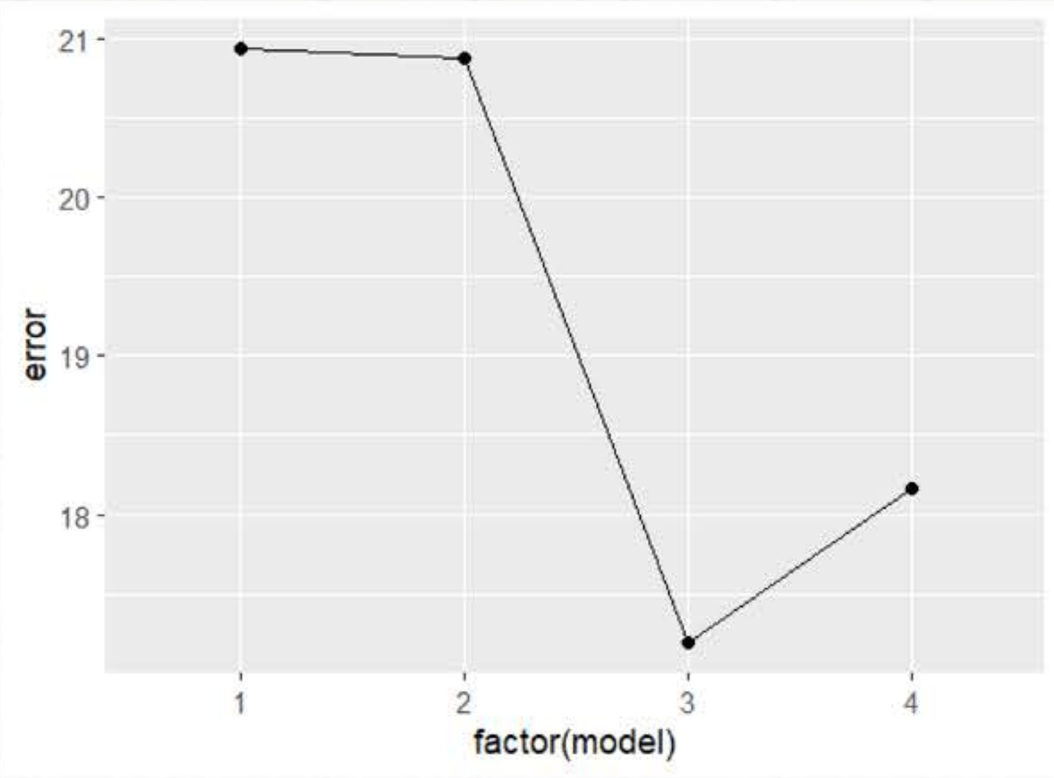
```
Residuals:
     Min       1Q   Median       3Q      Max
-16.610  -2.577  -0.518    2.267  24.870

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  47.9459     4.6726  10.261  < 2e-16 ***
CHAS          2.6347     0.9229   2.855  0.00453 **
RM            2.7776     0.4448   6.244 1.09e-09 ***
PTRATIO      -0.5964     0.1227  -4.860 1.69e-06 ***
logINDUS     -2.0066     0.4906  -4.090 5.22e-05 ***
logDIS       -4.1199     0.6370  -6.468 2.92e-10 ***
logLSTAT     -9.5723     0.6068 -15.776  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.553 on 398 degrees of freedom
Multiple R-squared:  0.7648,    Adjusted R-squared:  0.7613
F-statistic: 215.7 on 6 and 398 DF,  p-value: < 2.2e-16
```

Interpretation

Plots of training and test errors



Training MSE



Test MSE

Interpretation

From the plots

- The test error is less than the training error in each case
- Model 3 has the lowest training AND test errors (seemingly best!)
- Explains ~80% of variance

Interpretation

From model 4

- **Beta_{logLSTAT}: A 10% increase in percentage of “lower status population” in a tract decreases the median price in that tract by approximately $912 on average.**

```
Coefficients:
             Estimate
(Intercept)  47.9459
CHAS          2.6347
RM            2.7776
PTRATIO      -0.5964
logINDUS     -2.0066
logDIS       -4.1199
logLSTAT     -9.5723
```

Interpretation

From model 4

- **Beta_{logINDUS}: A 10% increase in the proportion of non-retail business acres per town decreases the median price in that tract by approximately $392 on average.**

```
Coefficients:
             Estimate
(Intercept)  47.9459
CHAS          2.6347
RM            2.7776
PTRATIO      -0.5964
logINDUS     -2.0066
logDIS       -4.1199
logLSTAT     -9.5723
```

Interpretation

From model 4

- **Beta_{logDIS}: A 10% increase in “weighted distance from five Boston employment centres” in a tract decreases the median price in that tract by, on average, $191.**

```
Coefficients:
             Estimate
(Intercept)  47.9459
CHAS          2.6347
RM            2.7776
PTRATIO      -0.5964
logINDUS     -2.0066
logDIS       -4.1199
logLSTAT     -9.5723
```

Interpretation

From model 4

- **Beta_{CHAS}: If a tract surrounds the Charles river, a house will cost $2635 more on average than when it does not.**

```
Coefficients:
             Estimate
(Intercept)  47.9459
CHAS          2.6347
RM            2.7776
PTRATIO      -0.5964
logINDUS     -2.0066
logDIS       -4.1199
logLSTAT     -9.5723
```

Interpretation

From model 4

- **Beta_{RM}: An increase by one, of the median number of rooms in properties in a tract leads to the increase of the median price of house in that tract by $2777**

```
Coefficients:
             Estimate
(Intercept)  47.9459
CHAS          2.6347
RM            2.7776
PTRATIO      -0.5964
logINDUS     -2.0066
logDIS       -4.1199
logLSTAT     -9.5723
```

Problems

- Our models have low R^2 terms.
- Residuals have approximately mean zero and constant variance, but do not satisfy normality
- Is a linear model even appropriate after having performed so many transformations?
- If you add more covariates to increase R^2, the model will become more specific.
- If we use logisical regression, is MEDV useful to interpret as a binary variable?

Next steps

- Work further on objective 2
- Logistical regression conducted on a more appropriate outcome
- A more appropriate binary outcome is CRIM
- CRIM = (# crimes / population) * 100
- Transform to CRIMLESSONE:
- 1 (safe) if CRIM < 5, 0 (unsafe) otherwise

Next steps

- New goal is the predict the probability that we can secure a house that we can be 95% sure is safe.
- In this case, we can have predictors like MEDV and DIS.
- Given a budget and how close you want to live from work you can figure out whether the house you buy will be safe.
- Results help you adjust you standards.

References

1. Harrison, D. and Rubinfeld, D.L. `Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978.
2. https://medium.com/@docintangible/racist-data-destruction-113e3eff54a8

Questions?

Thank you very much!