Splines and Housing

Michael Pena

Introduction

I have sourced this dataset from Kaggle.com, which is based on the 1990 California Census Data. The goal of this analysis is to explore the effects that the median age of homes in an area and the population size have on the median value of California homes. Each observation represents data from a census block group, a geographic unit used by the Census Bureau that typically consists of several hundred to a few thousand people.

In examining the relationship that home age and population density have on home values, I plan to apply both linear regression and spline regression models. Linear regression will help establish a baseline by modeling a straightforward relationship between these predictors and home values. In contrast, spline regression will allow for flexibility in capturing potential non-linear trends in the data, particularly if home value changes in complex ways with varying house age or population density. By comparing these models, I aim to assess whether a more flexible approach improves predictive accuracy and provides deeper insights into the factors influencing home value.

Data Exploration

```r
# import data
df <- read.csv("housing.csv", header = T)
# import libraries
library(splines)
library(GGally)
```

```
Loading required package: ggplot2
```

```
Registered S3 method overwritten by 'GGally':
  method from
  +.gg   ggplot2
```

```r
library(dplyr)
```

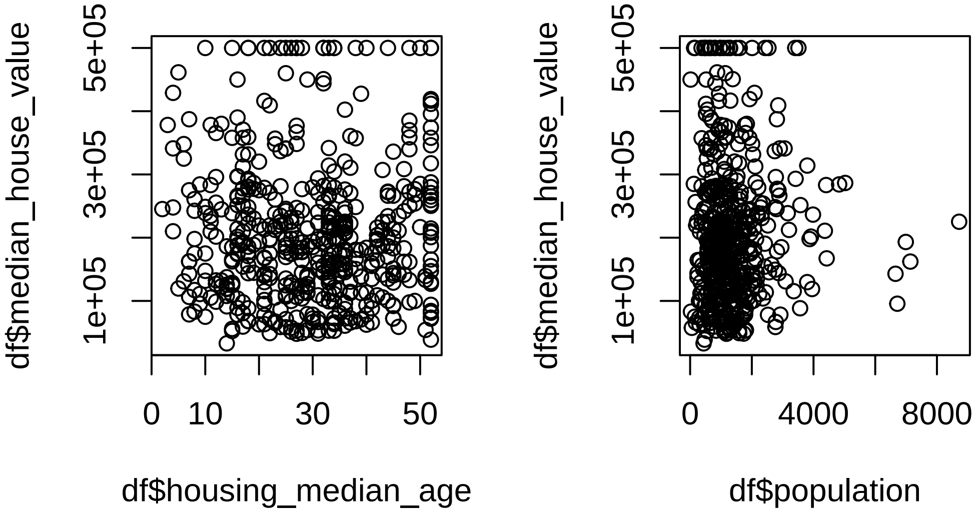
```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
library(npreg)
```

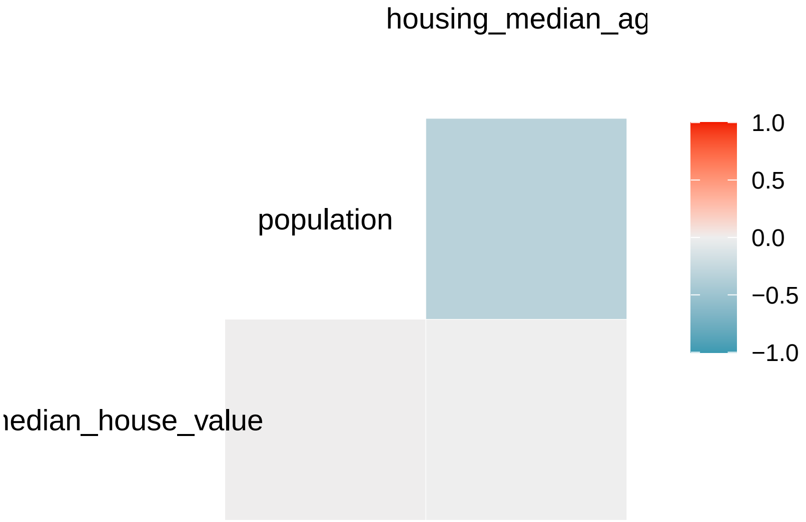
```
Package 'npreg' version 1.1.0
Type 'citation("npreg")' to cite this package.
```

```r
df <- df %>% select(median_house_value,population,housing_median_age)
# sample from the data
set.seed(80808)
k <- sample(1:20640, size = 500, rep = F)
df = df[k,]
```

```r
# need to correlation of these predictors
par(mfrow = c(1,2))
plot(y=df$median_house_value,x=df$housing_median_age)
plot(y=df$median_house_value,x=df$population)
```

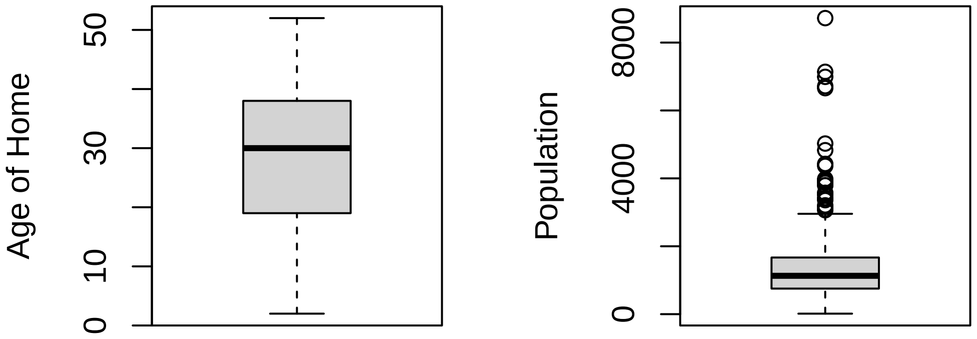


```
ggcorr(df)
```



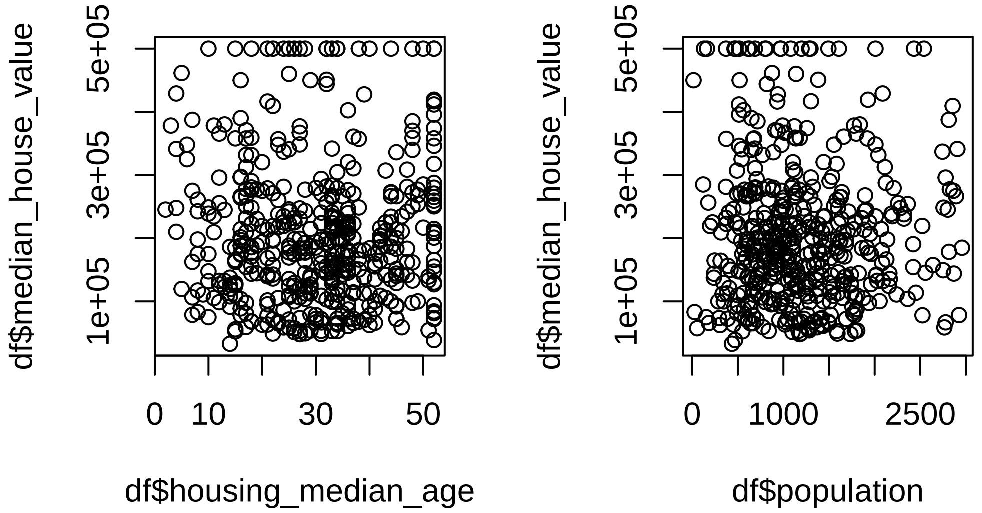
From here we see that population and housing_median_age are not well correlated. Let's also work within bounds of usual data that and we can consider the other data sample that we don't use anomalies.

```
par(mfrow = c(1,2))
boxplot(df$housing_median_age, ylab = "Age of Home") -> bp1
boxplot(df$population, ylab = "Population") -> bp2
```



Let’s remove outliers of the population first.

```r
# remove outliers and plot the graphs again
df = df[df$population < min(bp2$out),]
par(mfrow = c(1,2))
plot(y=df$median_house_value,x=df$housing_median_age)
plot(y=df$median_house_value,x=df$population)
```



Model Fitting

```r
# linear model
prim_lm1 <- lm(df$median_house_value ~ df$population)
prim_lm2 <- lm(df$median_house_value ~ df$housing_median_age)
summary(prim_lm1)
```

```
Call:
lm(formula = df$median_house_value ~ df$population)

Residuals:
    Min      1Q  Median      3Q     Max 
-174848  -80775  -19194   57699  298898 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   209124.749  11246.276   18.59   <2e-16 ***
df$population     -3.158      8.305   -0.38    0.704    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 112300 on 472 degrees of freedom
Multiple R-squared:  0.0003063, Adjusted R-squared:  -0.001812 
F-statistic: 0.1446 on 1 and 472 DF,  p-value: 0.7039
```

```
summary(prim_lm2)
```

```
Call:
lm(formula = df$median_house_value ~ df$housing_median_age)

Residuals:
    Min      1Q  Median      3Q     Max 
-172420  -80723  -19876   58005  294682 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)           2.053e+05  1.354e+04  15.166   <2e-16 ***
df$housing_median_age 2.496e-01  4.135e+02   0.001        1    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 112300 on 472 degrees of freedom
Multiple R-squared:  7.722e-10, Adjusted R-squared:  -0.002119 
F-statistic: 3.645e-07 on 1 and 472 DF,  p-value: 0.9995
```

```
# spline models
 spline1 <- lm(median_house_value ~ ns(population, df = 4), data = df)
 spline2 <- lm(median_house_value ~ ns(housing_median_age, df = 3), data = df)
summary(spline1)
```

```

Call:
lm(formula = median_house_value ~ ns(population, df = 4), data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-184834  -80090  -19041   51118  310785 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)               197216      35312   5.585 3.96e-08 ***
ns(population, df = 4)1    -5382      34239  -0.157    0.875    
ns(population, df = 4)2   -27378      30870  -0.887    0.376    
ns(population, df = 4)3    53890      80524   0.669    0.504    
ns(population, df = 4)4    32656      32129   1.016    0.310    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 111900 on 469 degrees of freedom
Multiple R-squared:  0.01347,   Adjusted R-squared:  0.005052 
F-statistic:   1.6 on 4 and 469 DF,  p-value: 0.173
```

```
summary(spline2)
```

```

Call:
lm(formula = median_house_value ~ ns(housing_median_age, df = 3), 
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-201476  -81512  -15955   50271  312315 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                       240720      27410   8.782   <2e-16 ***
```

```
ns(housing_median_age, df = 3)1   -51988      20503  -2.536   0.0115 *
ns(housing_median_age, df = 3)2   -56198      62374  -0.901   0.3681
ns(housing_median_age, df = 3)3    19728      17820   1.107   0.2688
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 111300 on 470 degrees of freedom
Multiple R-squared:  0.02141,   Adjusted R-squared:  0.01516
F-statistic: 3.428 on 3 and 470 DF,  p-value: 0.01708
```

Compare Models

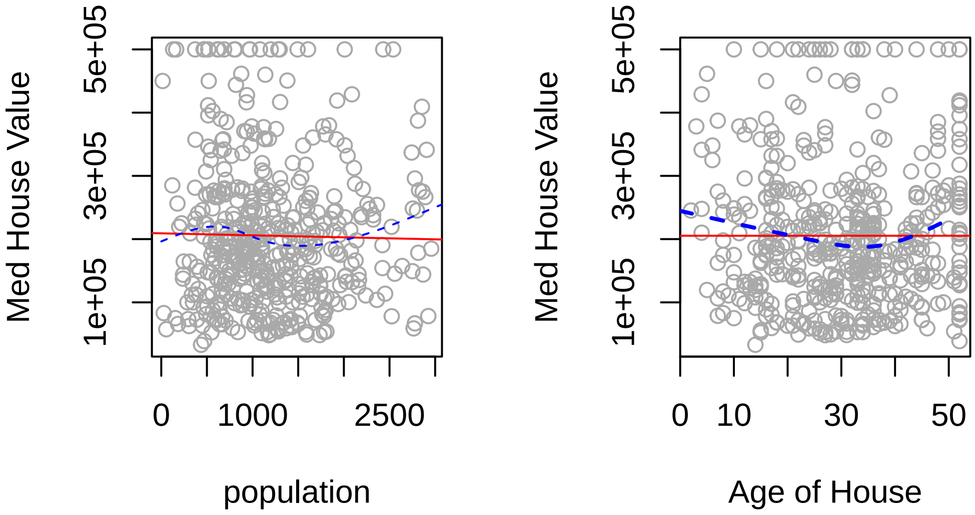
```
# compare the models
models <- list("prim_lm1" = prim_lm1, "prim_lm2" = prim_lm2, 'spline1' = spline1, 'spline2' =
for (name in names(models)) {
  model <- models[[name]]
  cat("Model:", name, "\n")
  cat("AIC:", AIC(model), "\n")
  cat("BIC:", BIC(model), "\n\n")
}
```

```
Model: prim_lm1
AIC: 12373.14
BIC: 12385.63

Model: prim_lm2
AIC: 12373.29
BIC: 12385.77

Model: spline1
AIC: 12372.86
BIC: 12397.83

Model: spline2
AIC: 12367.03
BIC: 12387.84
```

```
# denote predictions
n = length(df[,1])
pseq = seq(0,12000,length = n)
```

```r
aseq = seq(0,50, length = n)
df.seq = as.data.frame(cbind(pseq,aseq))
names(df.seq) = c("population","housing_median_age")
df$pred_spl1 <- predict(spline1, newdata = df.seq)
df$pred_spl2 <- predict(spline2, newdata = df.seq)
```

```r
par(mfrow = c(1,2))
# visualizations
  # population
  plot(df$population ,df$median_house_value, xlab = "population", ylab = "Med House Value",c
  abline(prim_lm1, col = "red")
  lines(pseq ,df$pred_spl1, col = 'blue', lty = 2, )
  # housing median age
  plot(df$housing_median_age ,df$median_house_value, xlab = "Age of House", ylab = "Med Hous
  abline(prim_lm2, col = "red")
  lines(aseq,df$pred_spl2, col = 'blue', lty = 2, lwd = 2)
```



Interpretation

I find that the splines are better at pointing out slight trends when for example we observe the Population plot, we do notice that smaller towns tend to have cheaper homes. This reflects in the real world where the smaller the less populated an area is, it will not be as in demand. However, the splines could possibly be exaggerating these trends and the correct move forward would be to cross validate this.

In the short term, the splines models score lower AIC and BIC than the linear models, but this difference is arguably negligible.

Model	AIC	BIC
linear model (population)	12373.14	12385.63
linear model (house age)	12373.29	12385.77
spline model (population)	12372.86	12397.83
spline model (house age)	12367.03	12387.84

This data seems to be so uncorrelated that while the splines reveals slight trends of up and down, it's not significantly different from what the linear line is already saying. I think if we care about these subtleties, then sure, splines are reveal a swooping trend the Value of the home versus Age; Smaller populated areas have lower home values then densely populated areas. But even for the latter, there is not much change. I think this can only reveal that predicting Home Value based on these two predictors is not time well spent.

In general, splines are probably better than linear regression (excluding polynomial) in regards to curve fitting as splines allow for flexibility which can in-turn lead to more accurate predictions. Note that it is crucial to address over-fitting as this can happen when one mindlessly uses this method. Still, if we are practical and use this on data that has an obvious trend, it will most likely be more efficient than a plain linear line.

Conclusion

The variables selected for this analysis—median age of homes and population size—may not have been optimal predictors of California home prices, as they do not appear to have a strong general effect on housing values. The spline models, while capable of capturing subtle non-linear patterns, did not reveal a meaningful relationship between the predictors and the response variable. In a practical context, the results suggest that these predictors are not significantly associated with home values, and further analysis might benefit from exploring other variables that have a more direct impact on housing prices in California.

Moreover, this analysis did not provide clear evidence that splines outperform a linear approach in this context. Although spline models can reveal subtle trends that linear regression might miss, these trends may be too minor to hold interpretive value here. Additionally, I did not perform cross-validation; implementing a k-fold cross-validation could provide insights into the generalizability and robustness of the spline models compared to linear regression. Overall, splines are best applied to data that exhibit clear non-linear relationships, and in cases where predictors are only weakly correlated with the response variable, a more straightforward modeling approach may be sufficient.