STAT 444 Project 2 – Smoothing Model

Tianyi Wu

- UW ID: 20934015
- Kaggle public score: 0.13812
- Kaggle private score: ???
- Kaggle submission count/times: 13

Summary

Preprocessing

Transformation (if any, delete if none)

- price: performed box-cox transformation on the response variate price.
- saledate: transformed as the number of days since 1970/01/01, then take the power of 5
- landarea: take log
- stories: take log
- rmdl_diff: log(rmdl_diff + 1) because some of them are zero
- all categorical variables is treated as factors

New Variables (if any, delete if none)

- if_rmdl : indicator whether the house is remodeled
- rmdl_diff: numerical rep the difference between remodel date and sale date; 0 if remodel is after sale
- saleyear: year of the sale
- buy_first: indicator whether the house is build first or buy first
- total_bath: combine number of bathroom and number of half-bathrooms (bathrm+0.5hf_bathrm)

Missing data handling

- yr_rmdl: missing yr_rmdl is recoded as 0 (we use if_rmdl and rmdl_diff instead of yr_rmdl in the model)
- ayb: missing ayb is recoded as eyb - avg_gap between ayb and eyb
- quadrant: missing quadrant is recoded as "NW" bc "NW" is most popular
- kitchen and stories: omitted

Model Building

Main package used: `mgcv::gam`

- fitted all predictors naively
- deleted insignificant ones and adjust number of knots for `ayb`

- forward selection to examine significance of interaction
- added interaction (`ti`) between continous predictors
- added interaction (`by` in `s()`) between continous and categorical

Final Model

- The final model is $\frac{(price^{\lambda}-1)}{\lambda} \sim$ s(rooms) +s(total_bath)+s(rmdl_diff) +s(bedrm)+s(ayb, k = 20, by = cndtn)+s(eyb)+s(saledate)+s(gba) +fireplaces+s(stories) +s(landarea)+s(latitude)+s(longitude) +heat+ac+style+grade+cndtn+roof+intwall+kitchens+nbhd+ward +quadrant+if_rmdl+buy_first + ti(eyb, ayb) + ti(gba,landarea)+ti(longitude,gba) + ti(longitude, ayb) +ti(longitude, eyb)+ti(saledate,latitude)

1.Preprocessing

1.1 Loading data

```r
load("smooth.Rdata")
```

examine categorical predictors

```r
table(dtrain$heat)
```

```
##
##     Air Exchng        Air-Oil  Elec Base Brd       Evp Cool     Forced Air
##              1              2              5              1           1597
## Gravity Furnac  Hot Water Rad        Ht Pump        No Data   Wall Furnace
##              4           1368             53              2              3
##      Warm Cool Water Base Brd
##           1958              6
```

```r
table(dtrain$ac)
```

```
##
##    N    Y
##  803 4197
```

```r
table(dtrain$style)
```

```
##
##         1 Story   1.5 Story Fin 1.5 Story Unfin         2 Story   2.5 Story Fin
##             523             317              13            2867             941
## 2.5 Story Unfin         3 Story         4 Story        Bi-Level     Split Foyer
##              98             143               3               1              36
##     Split Level
##              58
```

```r
table(dtrain$grade)
```

```
##
## Above Average       Average     Excellent Exceptional-A Exceptional-B
##          1320           705           315            71            43
## Exceptional-C Exceptional-D  Fair Quality  Good Quality   Low Quality
##             4             8            13          1365             2
##      Superior     Very Good
##           191           963
```

```r
table(dtrain$cndtn)
```

```
##
##   Average Excellent      Fair      Good      Poor Very Good
##      1631        70        20      2704         6       569
```

```r
table(dtrain$extwall)
```

```
##
##       Aluminum   Brick Veneer   Brick/Siding    Brick/Stone   Brick/Stucco
##             83             66            458             55             75
##   Common Brick       Concrete Concrete Block        Default     Face Brick
##           2680              6              5              2              9
##      Hardboard   Metal Siding        Shingle          Stone   Stone Veneer
##             13              4            133             68             11
##   Stone/Siding   Stone/Stucco         Stucco   Stucco Block   Vinyl Siding
##             53             28            326              1            412
##    Wood Siding
##            512
```

```r
table(dtrain$roof)
```

```
##
##       Built Up      Clay Tile   Comp Shingle Composition Ro  Concrete Tile
##            184             62           2893              1              2
##     Metal- Cpr     Metal- Pre     Metal- Sms        Neopren          Shake
##              3              1             99              6             92
##        Shingle          Slate        Typical
##             52           1600              5
```

```r
table(dtrain$intwall)
```

```
##
##        Carpet  Ceramic Tile       Default      Hardwood Hardwood/Carp
##           112             3             2          4163           559
##   Lt Concrete       Parquet    Wood Floor
##             6             1           154
```

```r
table(dtrain$ward)
```

```
##
## Ward 1 Ward 2 Ward 3 Ward 4 Ward 5 Ward 6 Ward 7 Ward 8
##     17     69   1890   1453    658     26    691    196
```

```r
table(dtrain$quadrant)
```

```
##
##   NE   NW   SE   SW
## 1016 3333  613   11
```

1.2 Missing data handling

Check missing values for each predictor:

```r
colSums(is.na(dtrain))
```

```
##     bathrm  hf_bathrm       heat         ac      rooms      bedrm        ayb
##          0          0          0          0          0          0         14
##    yr_rmdl        eyb    stories   saledate      price        gba      style
##       1999          0          4          0          0          0          0
##      grade      cndtn    extwall       roof    intwall   kitchens fireplaces
##          0          0          0          0          0          1          0
##   landarea   latitude  longitude       nbhd       ward   quadrant
##          0          0          0          0          0         27
```

```r
colSums(is.na(dtest))
```

```
##         Id     bathrm  hf_bathrm       heat         ac      rooms      bedrm
##          0          0          0          0          0          0          0
##        ayb    yr_rmdl        eyb    stories   saledate        gba      style
##          3        411          0          0          0          0          0
##      grade      cndtn    extwall       roof    intwall   kitchens fireplaces
##          0          0          0          0          0          0          0
##   landarea   latitude  longitude       nbhd       ward   quadrant
##          0          0          0          0          0          5
```

So far we don't deal with missing values in `yr_rmdl`, we will add two new variables later to explain it so `yr_rmdl` won't be used directly in the model.

```r
# =========== train ===============
avg_gap_train <- mean(dtrain$eyb-dtrain$ayb, na.rm = T)
# missing ayb is recoded as eyb - avg_gap between ayb and eyb
dtrain$ayb <- ifelse(is.na(dtrain$ayb), dtrain$eyb-avg_gap_train, dtrain$ayb)
# missing quadrant is recoded as "NW" bc "NW" is most popular
dtrain$quadrant <- ifelse(is.na(dtrain$quadrant), "NW", dtrain$quadrant)


#=========== test ============
avg_gap_test <- mean(dtest$eyb-dtest$ayb, na.rm = T)
# missing ayb is recoded as eyb - avg_gap between ayb and eyb
dtest$ayb <- ifelse(is.na(dtest$ayb), dtest$eyb-avg_gap_test, dtest$ayb)
# missing quadrant is recoded as "NW" bc "NW" is most popular
dtest$quadrant <- ifelse(is.na(dtest$quadrant), "NW", dtest$quadrant)
```

```r
colSums(is.na(dtrain))
```

```
##     bathrm  hf_bathrm       heat         ac      rooms      bedrm        ayb
##          0          0          0          0          0          0          0
##    yr_rmdl        eyb    stories   saledate      price        gba      style
##       1999          0          4          0          0          0          0
##      grade      cndtn    extwall       roof    intwall   kitchens fireplaces
##          0          0          0          0          0          1          0
##   landarea   latitude  longitude       nbhd       ward   quadrant
##          0          0          0          0          0          0
```

1.3 new variable

```r
# binary variable check whether the house is remodeled
dtrain$if_rmdl <- ifelse(is.na(dtrain$yr_rmdl), 0, 1)
dtrain$if_rmdl <- as.factor(dtrain$if_rmdl)

# year of the house sold
dtrain$saleyear<-as.numeric(substr(dtrain$saledate, 1, 4))

# the difference between sale year and the remodel year, if remodel is after sale
# then 0
for (i in seq(nrow(dtrain))) {
  if (is.na(dtrain$yr_rmdl[i])) {
    dtrain$rmdl_diff[i] <- 0
  } else if (dtrain$saleyear[i] <= dtrain$yr_rmdl[i]) {
    dtrain$rmdl_diff[i] <- 0
  } else if (dtrain$saleyear[i] > dtrain$yr_rmdl[i])
    dtrain$rmdl_diff[i] <- dtrain$saleyear[i] - dtrain$yr_rmdl[i]
}


# combine bathroom and half_bathroom
dtrain$total_bath <- dtrain$bathrm+0.5*dtrain$hf_bathrm

# whether it's sold first or build first
dtrain$buy_first <- as.factor(as.numeric(dtrain$saleyear < dtrain$ayb))
```

```r
# fill the na for yr_rmdl as 0
dtrain$yr_rmdl <- ifelse(is.na(dtrain$yr_rmdl), 0, dtrain$yr_rmdl)
# omit other na
dtrain_full <- na.omit(dtrain)
```

Now, `dtrain_full` is the complete data frame we will be working with.

1.4 data transformation

```r
# the number of days since 1970/01/01
dtrain_full$saledate <- as.numeric(as.Date(dtrain_full$saledate))
```

```r
# to factor
dtrain_full$heat <- as.factor(dtrain_full$heat)
dtrain_full$ac <- as.factor(dtrain_full$ac)
dtrain_full$style <- as.factor(dtrain_full$style)
dtrain_full$grade <- as.factor(dtrain_full$grade)
dtrain_full$cndtn <- as.factor(dtrain_full$cndtn)
dtrain_full$extwall <- as.factor(dtrain_full$extwall)
dtrain_full$roof <- as.factor(dtrain_full$roof)
dtrain_full$intwall <- as.factor(dtrain_full$intwall)
dtrain_full$ward <- as.factor(dtrain_full$ward)
dtrain_full$quadrant <- as.factor(dtrain_full$quadrant)
```

```r
dtrain_full$gba <- log(dtrain_full$gba)
dtrain_full$landarea <- log(dtrain_full$landarea)
dtrain_full$saledate <- dtrain_full$saledate^5

dtrain_full$stories <- log(dtrain_full$stories)
dtrain_full$rmdl_diff <- log(dtrain_full$rmdl_diff + 1)
```

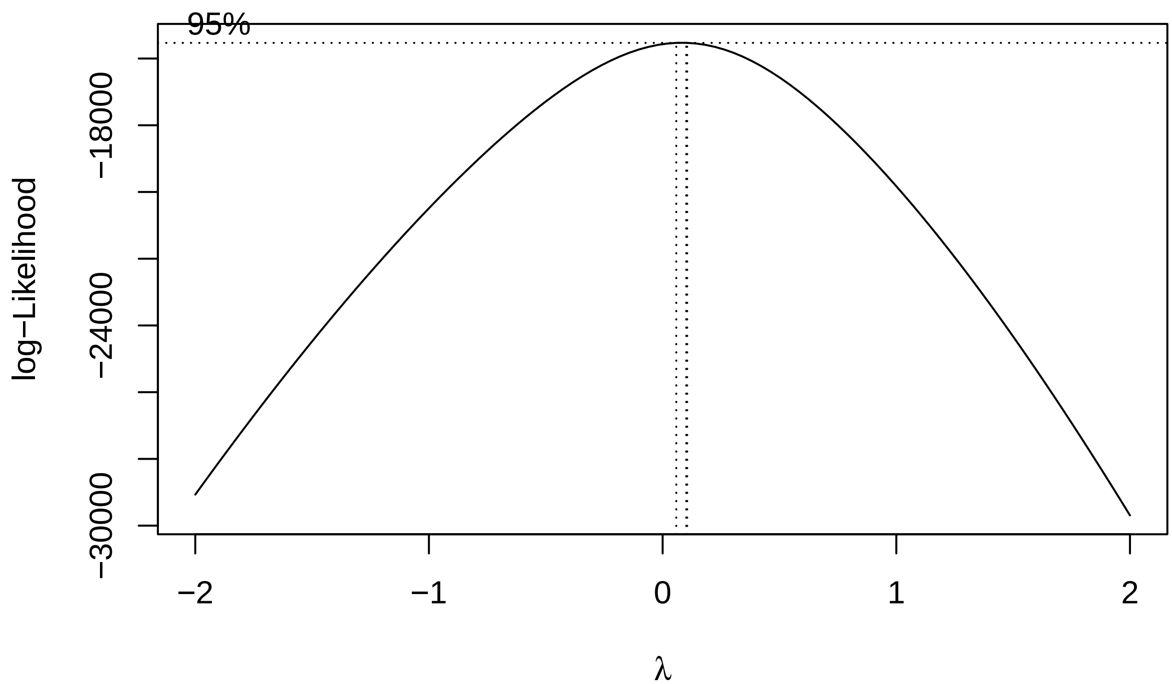
2. Model building

Note: all the output for model summary is hided because too large.

Perform box-cox trnasformation to transform `price`

```r
naive_lm <- lm(price ~ bathrm + rooms+bedrm+ayb+yr_rmdl+eyb+saledate+gba+landarea+
                 latitude+longitude, data = dtrain_full)

library(MASS)
boxcox_tr <- boxcox(naive_lm)
```




```
lambda <- boxcox_tr$x[which.max(boxcox_tr$y)]
```

2.1 The naive model

```
library(mgcv)
```

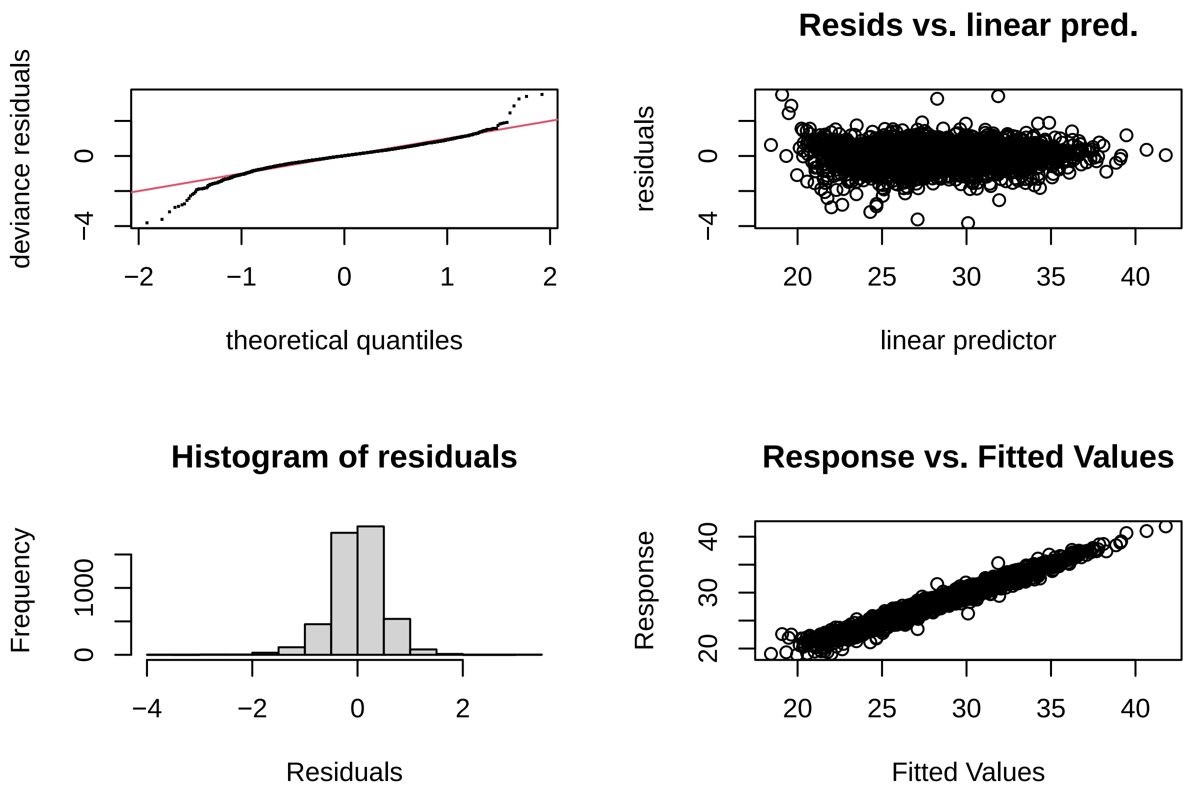
```
## Loading required package: nlme
```

```
## This is mgcv 1.9-1. For overview type 'help("mgcv-package")'.
```

```
naive_md <- gam(((price^lambda-1)/lambda) ~ s(rooms) + s(bathrm)
                +s(total_bath)+s(rmdl_diff)
                +s(bedrm)+s(ayb)+s(eyb)+s(saledate)+s(gba)
                +fireplaces+s(stories)
                +s(landarea)+s(latitude)+s(longitude)
                + heat+ac+style+grade+cndtn+extwall+roof+intwall+kitchens+nbhd+ward
                +quadrant+if_rmdl+buy_first
                ,data = dtrain_full)

summary(naive_md)
```

```r
par(mfrow= c(2,2))
gam.check(naive_md)
```



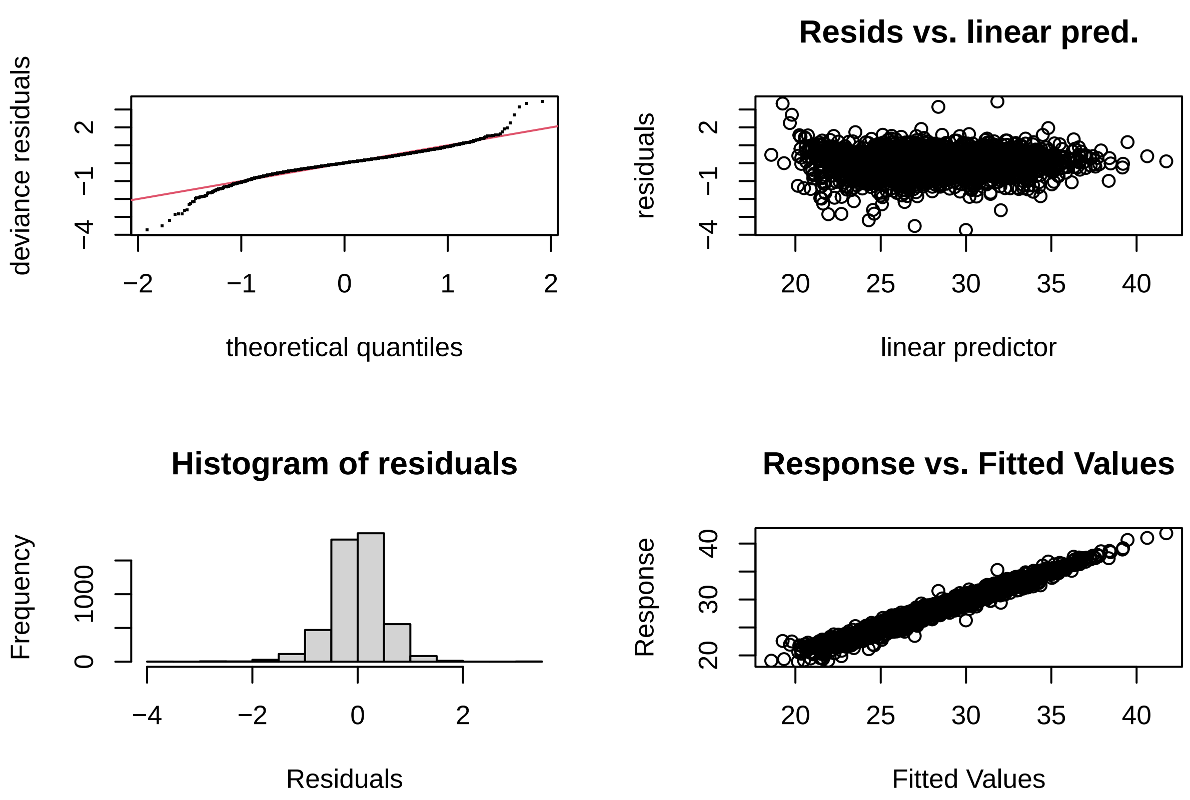

```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 34 iterations.
## The RMS GCV score gradient at convergence was 1.775004e-07 .
## The Hessian was positive definite.
## Model rank =  259 / 261
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##                 k'  edf k-index p-value
## s(rooms)      9.00 8.33    1.01    0.79
## s(bathrm)     9.00 1.00    1.01    0.75
## s(total_bath) 9.00 2.07    1.00    0.49
## s(rmdl_diff)  9.00 5.51    0.98    0.07 .
## s(bedrm)      9.00 8.63    0.99    0.14
## s(ayb)        9.00 7.93    0.98    0.07 .
## s(eyb)        9.00 3.11    1.00    0.57
## s(saledate)   9.00 8.98    1.00    0.57
## s(gba)        9.00 6.42    1.00    0.59
## s(stories)    9.00 1.07    1.00    0.49
## s(landarea)   9.00 8.21    1.00    0.47
```

```
## s(latitude)   9.00 8.23    1.02    0.92
## s(longitude)  9.00 8.25    1.02    0.93
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
m2 <- gam(((price^lambda-1)/lambda) ~ s(rooms)
                +s(total_bath)+s(rmdl_diff)
                +s(bedrm)+s(ayb, k = 20)+s(eyb)+s(saledate)+s(gba)
                +fireplaces
                +s(landarea)+s(latitude)+s(longitude)
                + heat+ac+style+grade+cndtn+extwall+roof+intwall+kitchens+nbhd+ward
                +if_rmdl+buy_first
                ,data = dtrain_full)

# Summary of the model
summary(m2)
```

```r
par(mfrow= c(2,2))
gam.check(m2)
```




```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 42 iterations.
## The RMS GCV score gradient at convergence was 6.433102e-07 .
## The Hessian was positive definite.
```

```
## Model rank =  248 / 250
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##                  k'   edf k-index p-value
## s(rooms)       9.00  8.31    1.01    0.84
## s(total_bath)  9.00  1.95    1.01    0.59
## s(rmdl_diff)   9.00  5.39    0.98    0.09 .
## s(bedrm)       9.00  8.49    0.99    0.17
## s(ayb)        19.00 17.50    0.99    0.18
## s(eyb)         9.00  3.15    1.01    0.67
## s(saledate)    9.00  8.99    1.00    0.56
## s(gba)         9.00  6.66    1.00    0.51
## s(landarea)    9.00  4.31    1.00    0.45
## s(latitude)    9.00  8.24    1.02    0.94
## s(longitude)   9.00  8.00    1.02    0.91
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(naive_md,m2)
```

```
##                df      AIC
## naive_md 220.7362 7784.692
## m2       220.9828 7754.079
```

2.2 interaction

Forward selection to find meaningfull interaction terms (output ignored because too large)

```
included <- c(
  "bathrm",  "ac", "rooms", "bedrm", "ayb", "eyb",
  "stories", "saledate", "price", "gba", "kitchens", "fireplaces", "landarea",
  "latitude", "longitude", "quadrant")


match_index <- match(included, names(dtrain_full))
train <- dtrain_full[,match_index]

null_model <- lm(log(price) ~ 1, data=train)  # medv is the dependent variable; adjust accordingly
full_model <- lm(log(price) ~ (.)^2, data=train)

forward_selection_model <- step(null_model,
                                scope=list(lower=null_model, upper=full_model),
                                direction="forward")
```

```
summary(forward_selection_model)
```

```
m3 <- gam(((price^lambda-1)/lambda) ~ s(rooms)
                +s(total_bath)+s(rmdl_diff)
                +s(bedrm)+s(ayb, k = 20)+s(eyb)+s(saledate)+s(gba)
                +fireplaces+s(stories)
```

```
                +s(landarea)+s(latitude)+s(longitude)
                + heat+ac+style+grade+cndtn+roof+intwall+kitchens+nbhd+ward
                +quadrant+if_rmdl+buy_first
                + ti(eyb, ayb) + ti(gba,bathrm)+ti(gba,landarea)+ti(longitude,gba)
                +ti(gba,latitude)
                ,data = dtrain_full)

# Summary of the model
summary(m3)
```

```
m4 <- gam(((price^lambda-1)/lambda) ~ s(rooms)
                +s(total_bath)+s(rmdl_diff)
                +s(bedrm)+s(ayb, k = 20, by = cndtn)+s(eyb)+s(saledate)+s(gba)
                +fireplaces+s(stories)
                +s(landarea)+s(latitude)+s(longitude)
                + heat+ac+style+grade+cndtn+roof+intwall+kitchens+nbhd+ward
                +quadrant+if_rmdl+buy_first
                + ti(eyb, ayb) + ti(gba,landarea)+ti(longitude,gba)
                +ti(saledate,latitude)
                #+ ti(longitude, ayb)
                #+ti(longitude, eyb)
                ,data = dtrain_full)

summary(m4)
```

```
AIC(m3, m4)
```

```
##          df      AIC
## m3 233.7068 7538.334
## m4 241.5650 7410.936
```

```
# combine style
dtrain_full$style_com <- as.character(dtrain_full$style)

for (i in seq(nrow(dtrain_full))) {
  if (dtrain_full$style_com[i] == "1.5 Story Fin" |
      dtrain_full$style_com[i] == "1.5 Story Unfin") {
    dtrain_full$style_com[i] <- "1.5 Story"
  } else if (dtrain_full$style_com[i] == "2.5 Story Fin" |
             dtrain_full$style_com[i] == "2.5 Story Unfin"){
    dtrain_full$style_com[i] <- "2.5 Story"
  }
}
dtrain_full$style_com <- as.factor(dtrain_full$style_com)
```

```
m5 <- gam(((price^lambda-1)/lambda) ~ s(rooms)
             +s(total_bath)+s(rmdl_diff)
             +s(bedrm)+s(ayb, k = 20, by = cndtn)+s(eyb)+s(saledate)+s(gba)
             +fireplaces
             +s(landarea)+s(latitude)+s(longitude)
             + heat+ac+grade+cndtn+roof+kitchens+nbhd+ward
             +quadrant+if_rmdl+style_com+intwall+buy_first
```

```r
             + ti(eyb, ayb) + ti(gba,landarea)+ti(longitude,gba)
             + ti(longitude, ayb)+ti(saledate,latitude)
             ,data = dtrain_full)
```

```r
AIC(m3, m4, m5)
```

```
##          df      AIC
## m3 233.7068 7538.334
## m4 241.5650 7410.936
## m5 249.9181 7369.730
```

```r
m6 <- gam(((price^lambda-1)/lambda) ~ s(rooms)
             +s(total_bath)+s(rmdl_diff)
             +s(bedrm)+s(ayb, k = 20, by = cndtn)+s(eyb)+s(saledate)+s(gba)
             +fireplaces
             +s(landarea)+s(latitude)+s(longitude)
             + heat+ac+style+grade+cndtn+roof+intwall+kitchens+nbhd+ward
             +quadrant+if_rmdl+buy_first
             + ti(eyb, ayb) + ti(gba,landarea)+ti(longitude,gba)
             + ti(longitude, ayb)
             +ti(longitude, eyb)+ti(saledate,latitude)
             ,data = dtrain_full)
```

```r
AIC(m6)
```

```
## [1] 7355.683
```

2.3 final model

```r
final <- gam(((price^lambda-1)/lambda) ~ s(rooms)
             +s(total_bath)+s(rmdl_diff)
             +s(bedrm)+s(ayb, k = 20, by = cndtn)+s(eyb)+s(saledate)+s(gba)
             +fireplaces
             +s(landarea)+s(latitude)+s(longitude)
             + heat+ac+style+grade+cndtn+roof+intwall+kitchens+nbhd+ward
             +quadrant+if_rmdl+buy_first
             + ti(eyb, ayb) + ti(gba,landarea)+ti(longitude,gba)
             + ti(longitude, ayb)
             +ti(longitude, eyb)+ti(saledate,latitude)
             ,data = dtrain_full)
```