Greining fasteignamats

Elías Bjartur Einarsson (ebe19) og Þórhallur Auður Helgason (thh114)

21/10/2021

Inngangur

Í þessu verkefni er markmið okkar að skapa líkan sem nálgar fasteignamat eigna á ákveðnum svæðum Reykjavíkur að gefnum upplýsingum um viðeigandi fasteignir. Við lýsum því hér hvaða skref við tókum í líkanasmíðinni, hvernig við völdum og höfnuðum breytum, mátum gæði líkana og bárum saman þau líkön sem komu til greina.

Gögnin sem unnið var með voru fasteignamöt eigna fyrir árið 2017, unnin árið áður. Einungis var notast við gagnapunkta af fimm svæðum innan Reykjavíkur sem sjá má á korti hér að neðan; miðbær frá Bræðraborgarstíg að Tjörn, Melar að sjó, Háaleiti og skeifan, Hólar og Berg í Breiðholti og loks Réttarholtið. Á þessum svæðum voru í heildina 433 fasteignir með 21 breytum auk núvirðis, óháðu breytunnar. Skýribreyturnar voru eftirfarandi: Fastanúmer íbúðar, Kaupdagur, Tegund eignar, Svæðisnúmer, Byggingarár, Hæð íbúðar, Fjöldi lyfta, Fermetrafjöldi, Fjöldi hæða, Fjöldi bílastæða, Fjöldi baðkara, Fjöldi sturta, Fjöldi klósetta, Fjöldi eldhúsa, Fjöldi herbergja, Fjöldi stofa, Fjöldi geymsla, Stig framkvæmdar, Matssvæði, Undirmatssvæði og Tegund íbúðar. Þær breytur sem ekki segja sig sjálfar eru; Tegund eignar, Svæðisnúmer, Stig framkvæmdar, Matssvæði, Undirmatssvæði og Tegund íbúðar. Tegund eignar skiptist í fjóra flokka; Einbýlishús, parhús, íbúð og raðhús. Tegund íbúðar aðgreinir sérbýlishús frá fjölbýlishúsum, svæðisnúmer er auðkenni sveitarfélags og stig framkvæmdar er skali frá 0 upp í 10 sem metur hvort húsnæðið sé tilbúið. Matssvæði eru hverfin sem sjást á myndinni hér að neðan og undirmatssvæði eru ákveðin svæði innan þeirra.

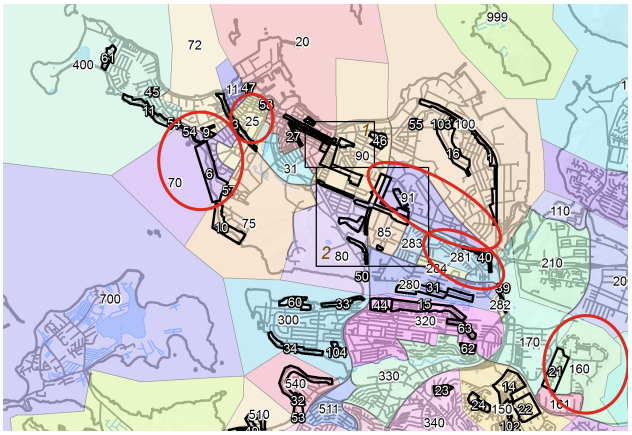


Figure 1: Svæði sem skoðuð voru eru rauðmerkt og nr. 70, 25, 91, 281 og 160.

Til að byrja með skoðuðum við gögnin, meðaltöl breyta, kvantíla og há- og lággildi ásamt því að umbreyta þeim yfir á rétt snið, svo sem kaupdegi yfir á dagsetningaform og flokkunarbreytum yfir í flokka. Einnig athugum við hvort einhver auð gildi eru til staðar. Við fjarlægjum strax fastanúmer sem breytu þar sem hún er einungis auðkenni fasteignar og inniheldur ekki upplýsingar um hana. Sömuleiðis fjarlægjum við svæðisnúmerið þar sem allar breyturnar deila svæðisnúmeri Reykjavíkur og það veitir þar með engar upplýsingar um gagnapunktana.

```r
# Fjarlægjum breytur sem augljóslega skipta ekki máli:
data <- subset( data, select = -c(svfn,rfastnum) )

# Athugum hvort auð gildi séu til staðar
if (sum(apply(data,2,is.nan))){
  print("Athuga auð gildi")
}

# Skilgreinum tegundir breyta:
data[ ,"kdagur"] <- as.Date(data[ ,"kdagur"]) # Kaupdagur sem dagsetning
data[ ,"teg_eign"] <- as.factor(data[ ,"teg_eign"]) # Tegund eignar sem flokkur

data[ ,"matssvaedi"] <- as.factor(data[ ,"matssvaedi"]) # Staðsetning sem flokkur
data[ ,"undirmatssvaedi"] <- as.factor(data[ ,"undirmatssvaedi"]) # Undirstaðsetning sem flokkur
data[ ,"ibteg"] <- as.factor(data[ ,"ibteg"]) # Tegund íbúðar sem flokkur

summary(data)
```

```
##      kdagur              nuvirdi             teg_eign       byggar
##  Min.   :2011-01-10   Min.   :  5993   Einbylishus: 45   Min.   :1901
##  1st Qu.:2012-08-17   1st Qu.: 20196   Ibudareign :369   1st Qu.:1953
##  Median :2013-11-11   Median : 25655   Parhus     :  6   Median :1964
##  Mean   :2013-10-14   Mean   : 29163   Radhus     : 13   Mean   :1963
##  3rd Qu.:2015-02-06   3rd Qu.: 33985                     3rd Qu.:1974
##  Max.   :2016-02-25   Max.   :133665                     Max.   :2014
##
##      haednr          lyfta             ibm2            fjhaed
##  Min.   :0.000   Min.   :0.0000   Min.   : 21.90   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:0.0000   1st Qu.: 65.50   1st Qu.:1.000
##  Median :2.000   Median :0.0000   Median : 85.30   Median :1.000
##  Mean   :1.928   Mean   :0.1663   Mean   : 95.41   Mean   :1.201
##  3rd Qu.:3.000   3rd Qu.:0.0000   3rd Qu.:111.90   3rd Qu.:1.000
##  Max.   :7.000   Max.   :2.0000   Max.   :289.30   Max.   :3.000
##
##     fjbilast           fjbkar          fjsturt           fjklos
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.00000   1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:1.000
##  Median :0.00000   Median :1.0000   Median :0.0000   Median :1.000
##  Mean   :0.05312   Mean   :0.8083   Mean   :0.4088   Mean   :1.187
##  3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.000
##  Max.   :2.00000   Max.   :2.0000   Max.   :2.0000   Max.   :3.000
##
##      fjeld           fjherb           fjstof          fjgeym
##  Min.   :0.000   Min.   : 0.000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:1.000   1st Qu.: 1.000   1st Qu.:1.000   1st Qu.:0.0000
##  Median :1.000   Median : 2.000   Median :1.000   Median :0.0000
##  Mean   :1.014   Mean   : 2.448   Mean   :1.293   Mean   :0.5912
##  3rd Qu.:1.000   3rd Qu.: 3.000   3rd Qu.:2.000   3rd Qu.:1.0000
##  Max.   :3.000   Max.   :13.000   Max.   :4.000   Max.   :4.0000
##
##      stig10       matssvaedi undirmatssvaedi ibteg          id
##  Min.   : 9.700   25 : 79    0      :354     11: 83   Min.   :    2
##  1st Qu.:10.000   70 : 85    28     : 31     12:350   1st Qu.: 9063
##  Median :10.000   91 : 67    21     : 27              Median :17670
```

```
##  Mean   : 9.999   160:136    3       :  7              Mean   :17672
##  3rd Qu.:10.000   281: 66    48      :  5              3rd Qu.:26541
##  Max.   :10.000              6       :  4              Max.   :34864
##                              (Other):  5
```

Við skiptum gagnasafninu okkar í þjálfunar- og prófunarsafn með 75% gagnapunkta í því fyrrnefnda og fjórðung í því síðarnefnda.

1. Fyrsta líkan

Að svo stöddu erum við tilbúnir að máta fyrsta líkanið okkar og greina það. Í fyrstu mátum við núvirði við allar þær breytur sem eftir standa.

```
# Splittum datasetti í þjálfun og prófun:
sizeTraining = floor(0.75 * nrow(data))
trainingSampleRowId <- sample(1:nrow(data), size = sizeTraining, replace = F)
train_data <- data[trainingSampleRowId, ]
test_data <- data[-trainingSampleRowId, ]

# Fittum fyrsta líkan, án nokkurrar vinnslu:
lm.first = lm(nuvirdi ~ . -id, data = train_data)
s.first = summary(lm.first)
```

Þessi fyrsta tilraun til að máta gögnin gefur okkur líkan til að miða við héðan af, það er bara upp á við eftir þetta. Þetta líkan fær **5659.86** í RMSE og **0.83** í aðlagað R^2. Er við skoðum spágildi líkansins út frá prófunargagnasetti fæst **6023.51** í RMSE. Skoðum til viðbótar annað grunnlíkan sem mátar núvirði eingöngu við fermetraverð. Þetta líkan mætti hugsa sem grunnviðmið parsímóníunnar.

```
lm.simple <- lm(nuvirdi ~ ibm2, data = train_data)
s.simple <- summary(lm.simple)
```

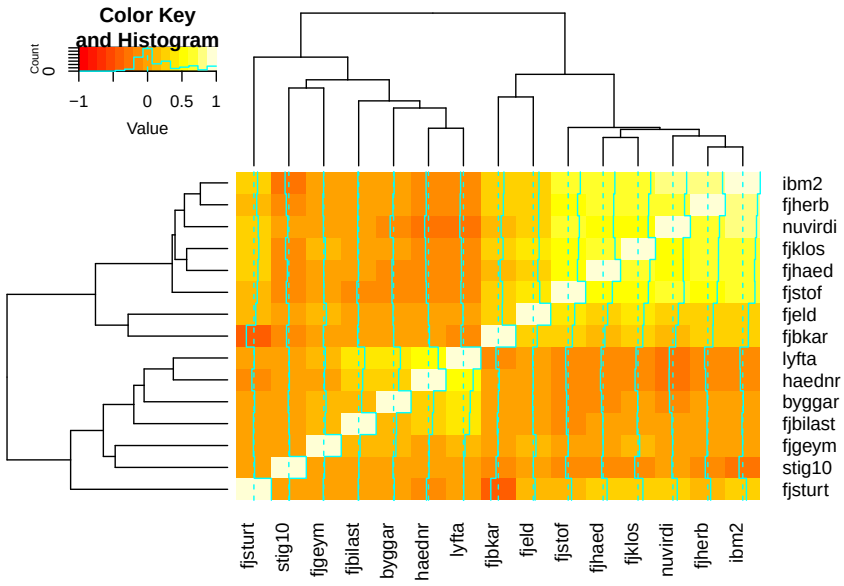
Við sjáum að það fær hærra RMSE en fyrsta líkanið okkar líkt og búast má við. Þetta einfalda líkan fær **8721.10** í RMSE á þjálfunarsetti og **8313.26** í RMSE á prófunarsetti, hvort tveggja mjög hátt.

2. Fækkun breyta

Byrjum á að breyta fjölda lyfta í tvíundarbreytu sem segir til um hvort það sé lyfta eður ei. Skoðum svo hvaða breytur eru línulega háðar og mega missa sín.

```
data[ ,"lyfta"] <- data[,"lyfta"]>0

# Skoðum breytur sem eru of líkar, multiple collinearity:
library(gplots)
heatmap.2(cor(data[numericNames]))
```



Við sjáum þyrpingu sem sýnir mikil líkindi og skoðum eigingildin

```
X <- model.matrix(lm(nuvirdi ~ ., data[numericNames]))
eigenX <- eigen(t(X) %*% X)
condNumber <- max(eigenX$values)/min(eigenX$values)
condNumber
```

```
## [1] 2.11158e+12
```

Hér sést að þessi hópur af breytum sem sýna mikla fylgni hvor við aðra eru fermetrar, fjöldi herbergja, núvirði (óháða breytan okkar), fjöldi klósetta, fjöldi hæði og fjöldi stofa. Við skoðun á breytufylkinu sést að ástandstalan er gríðarhá og greinilegt að eitthvað sé á seyði hér. Við skoðum því þá eiginvigra með sérlega lág eigingildi.

```
eigenX$values
```

```
##  [1] 1.672145e+09 8.215950e+05 9.069496e+02 2.831165e+02 1.934069e+02
##  [6] 1.463075e+02 7.969591e+01 5.246949e+01 4.484771e+01 3.466555e+01
## [11] 2.186847e+01 1.843604e+01 1.224871e+01 3.196135e+00 7.918929e-04
```

Við sjáum að eigingildi 15 er pínkulítið, skoðum það betur.

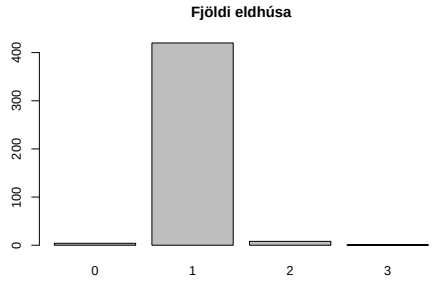
```
tiny <- eigenX$vectors[, 15]
# colnames(data[numericNames])[c(5, 6, 10, 12, 13)]
# sum(tiny[c(5, 6, 10, 12, 13)])
# sum(tiny)
# plot(tiny[c(5, 6, 10, 12, 13)])
bad_actors <- tiny[c(5, 6, 10, 12, 13)]
sum(bad_actors[c(1,5)]) - sum(bad_actors[c(2,3,4)])
```

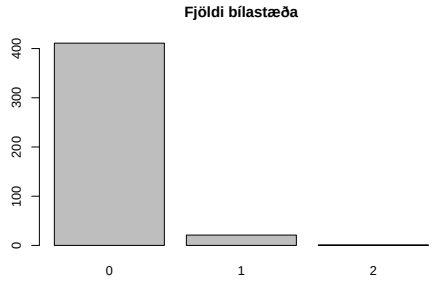
```
## [1] 4.060155e-05
```

Við sjáum að fermetrafjöldi og fjöldi stofa tjá nokkurn veginn sömu upplýsingar og fjöldi herbergja, klósetta og hæða. Við ákveðum því að taka þrjár síðarnefndu út. Þegar við skoðum líkanið sem út úr því kemur sést að RMSE hækkar en aðlagað R^2 gerir það sömuleiðis að örlitlu leyti.

Athugum svo að breyturnar fyrir fjölda eldhúsa og fjölda bílastæða taka nánast sömu gildi í öllum gagnapunktum. Þar að auki eru þær með mjög há p-gildi og við metum það svo að þær megi báðar fjúka. Við það breytist RMSE lítið sem ekkert en aðlagað R^2 hækkar.

Íbuðartegund	Eignartegund	n
11	Einbylishus	45
11	Ibudareign	19
11	Parhus	6
11	Radhus	13
12	Ibudareign	350





3. Ítarlegri gagnaúrvinnsla

Athugum nú aðrar breytur þar sem ástæður til að fjarlægja þær blasa ekki jafnvel við. Við sjáum að tegund eignar og íbúðartegund kóða fyrir mjög svipuðum eiginleikum fasteignar og að parhúsarflokkurinn er sá eini í tegund eignar sem mælist með almennilega svörun, mögulega að undanskildum einbýlishúsaflokknum sem er grunnflokkurinn. Þó eru einungis 6 gagnapunktar í parhúsaflokknum og því mögulega ástæða til að fella tegund eignar inn í íbúðartegund. Skoðum hvernig gögnin liggja í þeim flokkum.

```
group_by(data, Íbuðartegund = ibteg, Eignartegund = teg_eign) %>%
  count() %>%
  kbl() %>%
  kable_styling()
```

Hér sést að allar íbúðartegundir nr. 12 eru eignartegundin Íbúðareign. Þó eru nokkrar íbúðareignir sem falla í flokk 11 ásamt öllum hinum tegundum. Eftir samanburð á líkönum sem tóku annars vegar íbúðartegund og eignartegund út þá var það metið svo að betra væri að taka íbúðartegund út. RMSE lækkar og R^2 hækkar, alveg eins og við viljum.

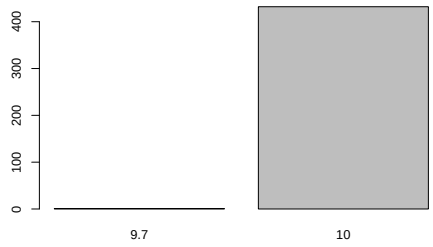
Önnur breyta sem mögulega er að rýra líkanið er undirmatssvæði. Skoðum hvernig gildin liggja þar.

```
group_by(data, undirmatssvaedi) %>%
  count() %>%
  kbl() %>%
  kable_styling()
```

undirmatssvaedi	n
0	354
3	7
6	4
21	27
28	31
40	2
48	5
54	3

Hér sést að langflestir punktarnir falla í undirmatssvæði 0 og að flestir flokkar innihalda einungis örfáar fasteignir. Einu flokkarnir sem fá lágt p-gildi eru 3 og 6, sem eru Ægissíða og Vesturbær NA við Hringbraut, en einungis 7 og 4 stök falla þar undir. Þar að auki virðast fjölmennustu flokkarnir, nr. 21 og 28 (Vesturberg í Breiðholti og Blokkir við Kringlumýra- og Miklubraut), skipta litlu máli. Á hinn bóginn þá innihalda undirmatssvæði í eðli sínu færri punkta en matssvæðin og væntanlega valin af góðum ástæðum. Það er eitt að búa í Vesturbænum en annað að búa á Ægissíðunni með útsýni yfir hafið. Af þessum ástæðum ákváðum við að halda í þessa breytu örlítið lengur og sjá hvernig henni vegnar í síðari greiningu á líkönunum.

Síðar kom í ljós, eftir að veigamiklir punktar voru skoðaðir, að punktur 10944 var óeðlilega áhrifamikill og var það sökum þess að eini breytileikinn í framkvæmdarstigsbreytunni kom frá honum. Allir punktar höfðu gildi 10 í þeirri breytu nema þessi eini. Af þeim völdum fjarlægðum við þá breytu úr líkaninu. Þetta hefði mátt gerast fyrr í ferlinu.



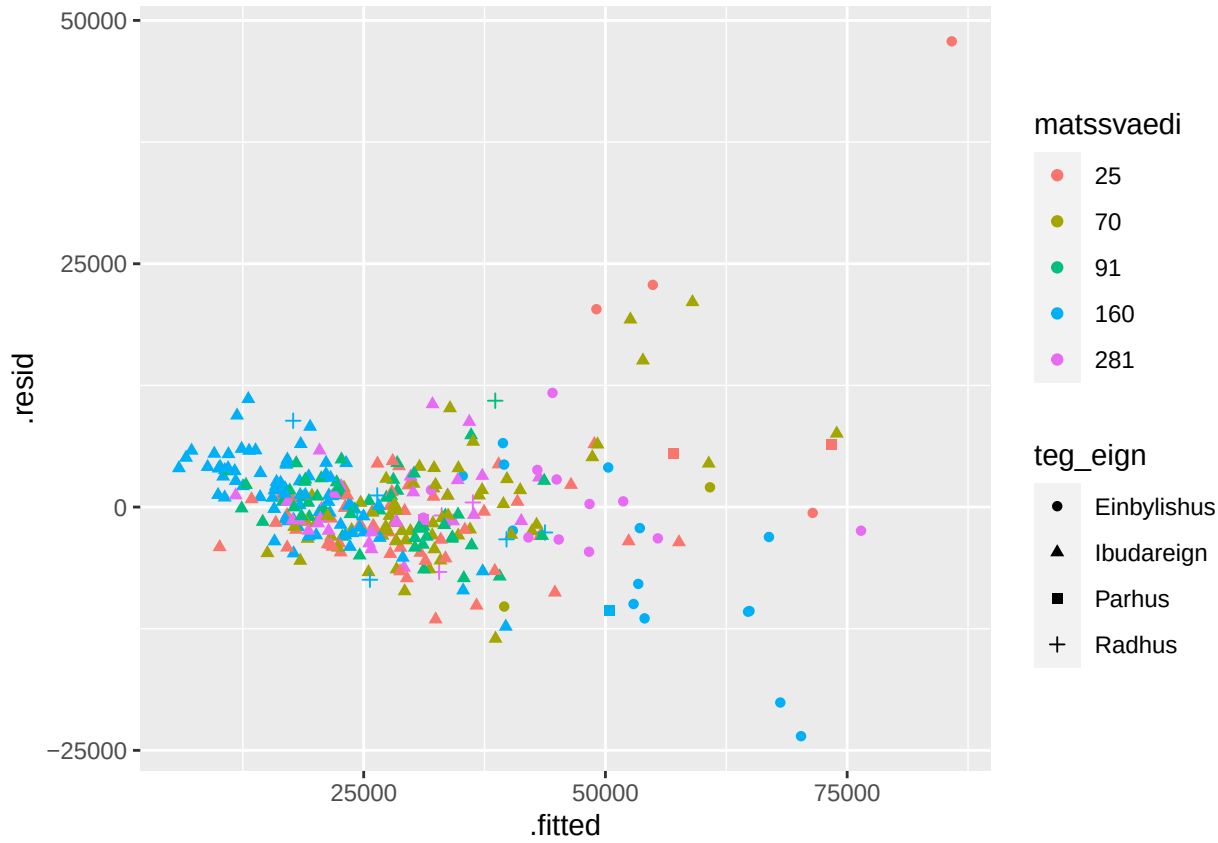

```
## [1] 0.8305896
```

Þá er líkan með **5784.05** í RMSE og **0.83** í aðlagað R^2.

4. Grunnskoðun á línuleika

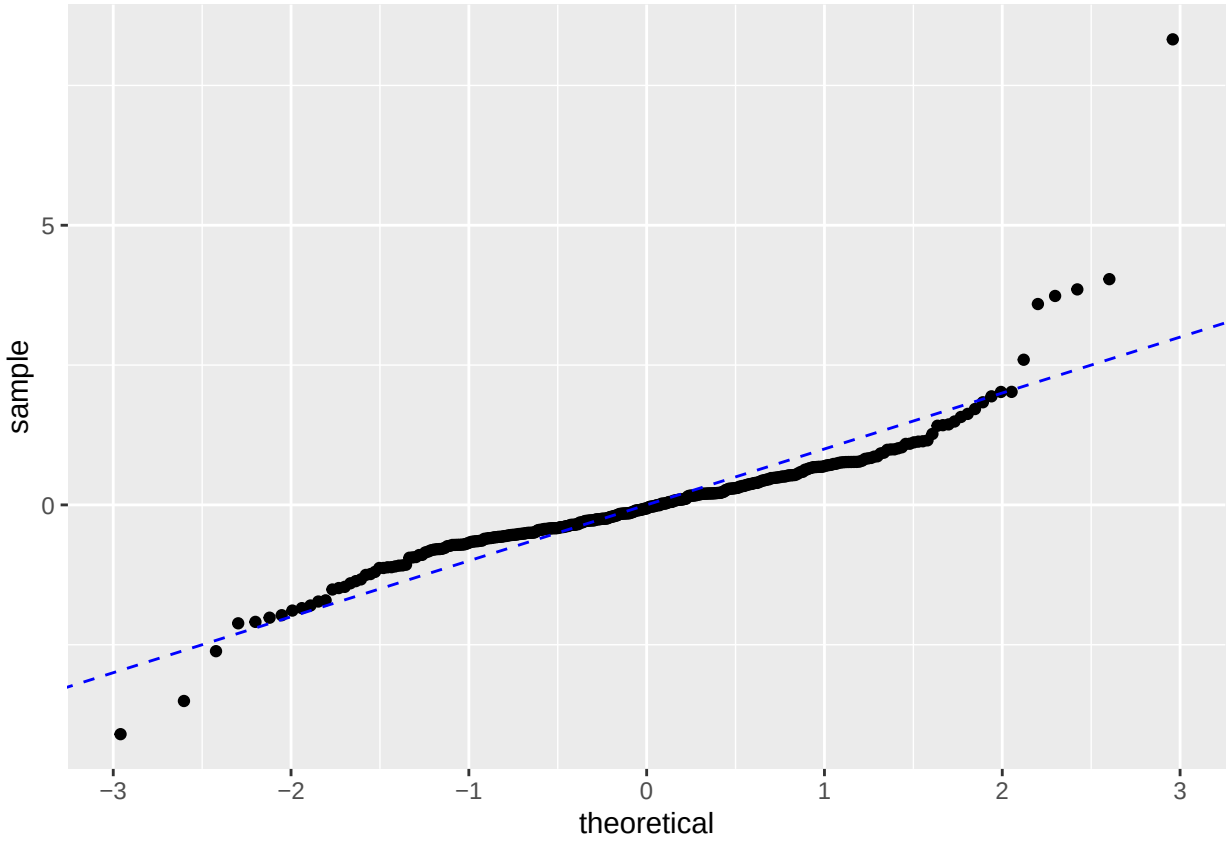
Byrjum á að skoða eitt mikilvægasta plottið, leifð á móti spágildi.

```
fortData <- fortify(lm.fifth)
fortData %>%
  ggplot(aes(x = .fitted, y = .resid, color = matssvaedi, shape = teg_eign)) +
  geom_jitter(width = 0.25)
```



Hér er augljóslega tilfelli af hederoskedaticity, þ.e. leifðin eykst með hærra spágildi. Við ættum að geta séð þetta vel líka með QQ-plotti af leifðinni:

```
tibble(Normal = fortData$.stdresid) %>%
  gather(type, val) %>%
  ggplot(aes(sample = val)) +
  stat_qq() +
  geom_abline(slope = 1, intercept = 0, lty = 2, col = 'blue')
```



Gögnin halda að miklu leyti í við y = x línuna en þó er einn punktur alveg kú-kú og tilhneigingin á báðum endum er samhverf sem bendir til þess að hér sé eitthvað annað en línulegt í gangi.

Af þessu tvennu að ofan drögum við þá ályktun að líklegt sé að við viljum umbreyta y breytunni okkar. Réttast er þó að skoða fyrst útlaga og áhrifamikla punkta vegna þess að BoxCox og aðrar aðferðir eru sérstaklega næmar fyrir slíku.