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.

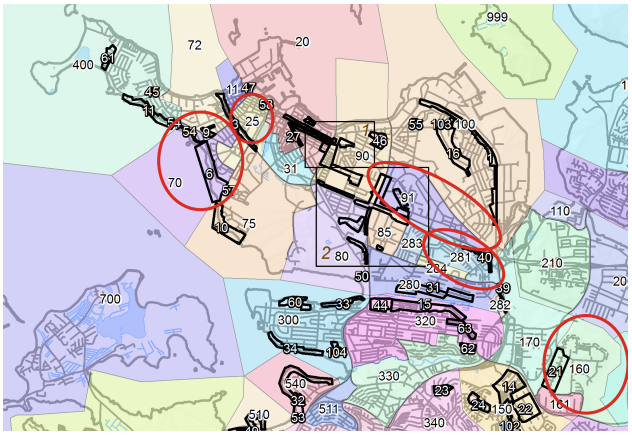


Figure 1: Svæði sem skoðuð voru eru rauðmerkt og nr. 70, 25, 91, 281 og 160.

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) )

# 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
```

Að lokum skiptum við gagnasafninu okkar í þjálfunar- og prófunarsafn með 75% gagnapunkta í því fyrrnefnda og fjórðung í því síðarnefnda, til þess að geta prófað niðurstöður okkar á óháðan máta.

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)
```

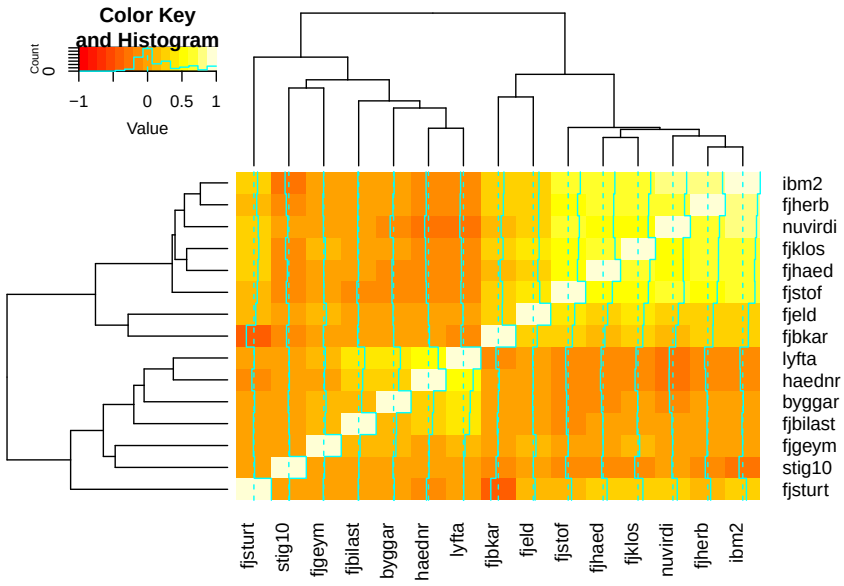
Við sjáum að það fær hærra RMSE en fyrsta líkanið okkar líkt og búast má við. Þetta einfalda líkan fær **8721.10** í RMSE á þjálfunarsetti og **8313.26** í RMSE á prófunarsetti, hvort tveggja mjög hátt.

2. Fækkun breyta

Í lýsingu verkefnisins var tekið fram að breytan LYFTA segði til um það hvort lyfta væri í húsinu eða ekki. Í raunsafninu er sú breyta með gildi sem hlýtur að vísa til fjölda lyfta í húsinu. Við ákváðum að það væri einfaldara að stilla þessa breytu líkt og verkefnislýsingin lýsti, þar sem það jafnar út hlutfallslegan fjölda milli breytanna. Við byrjum því 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]))
```



Myndin að ofan sýnir svokallað hitakort af samspili allra tölulegra breyta í gagnasettinu okkar. Þeim mun líkari sem tvær breytur eru, þeim mun ljósari er liturinn (hann er eðlilega ljósastur á hornalínunni). Við sjáum þyrpingu í efra hægra horninu sem sýnir mikil innbyrðis líkindi milli allra breyta. Samspil þátta í líkaninu sem eru of svipaðir og leggja því sömu þætti til mátsins, getur valdið því að vægi þeirra þátta dreifist óeðlilega milli margra breyta. Þannig getur mikilvægt framlag litið út fyrir að dreifast eins og brauðmylsna milli margra samverkandi þátta í stað þess að koma skýrt fram í einum þeirra, auk þess sem lítil hnik í þessum þáttum geta valdið mikilli skekkju í mælingum. Þar að auki er ennþá líklegra að við ofmátum líkanið að gagnasettinu okkar, ef margar breytur fá að fínstilla þetta samspil fyrir ólíka gagnapunkta, í stað þess að ein breyta reyni að finna meðalveg fyrir þá alla í einu.

Af þessum sökum er mikilvægt að reyna að greina þetta línulega hæði innan breytanna en að því er ekki hlaupið. Við grípum því til þess ráðs að skoða eiginvigra og eigingildi fylkisins.

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

Úr tölulegu flokkunum smíðum við líkan sem mátar núvirðisbreytuna okkar og skoðum fylki skýribreytanna út frá eigingildum og eiginvigrum. Ástandstala þessa fylkis er vísir að því hversu næm niðurstaða líkansins er fyrir litlu hniki. Í þessu tilfelli mælist hún **2.1e+12**, sem er ógnarstór og því vísir að því að fylkið sé sérlega næmt fyrir örsmæðarbreytingum. Þetta svarar til þess að einhverjir dálkvigrar fylkisins séu innbyrðis línulega háðir. Til þess að leita að slíku línulegu hæði væri hægt að skoða allar mögulegar summur innan gilda eiginvigra fylkisins en við erum þegar komin á sporið, þökk sé hitakortinu okkar að ofan.

Þar 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 tökum til skoðunar minnsta eigingildi fylkisins og hefjumst handa við að greina línulegt hæði milli þeirra gilda í tilsvarandi eiginvigri sem svara til breytanna okkar úr hitakortinu. Við sleppum því að skoða núvirðisbreytuna í þessu samhengi, því línulegt samband milli óháðu breytunnar okkar og annarra skýribreyta er jákvætt fyrir líkanið. Við viljum koma auga á línulegt hæði innan skýribreytanna okkar.

```
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)])
```

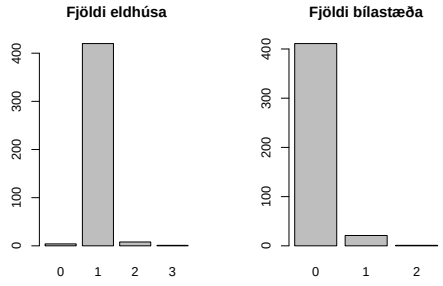
Íbuðartegund	Eignartegund	n
11	Einbylishus	45
11	Ibudareign	19
11	Parhus	6
11	Radhus	13
12	Ibudareign	350

```
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
```

Með því að leggja gildi 5 og 13 í vigrinum við gildi 6, 10 og 12 fæst tala sem liggur verulega nálægt 0. Þetta svarar til þess að þessir tveir hópar séu að leggja svipaðar upplýsingar til líkansins okkar. Ef við skoðum þær breytur sem heyra til þessara gilda, þá skýrist það fyrir okkur að fermetrafjöldi og fjöldi stofa tjá nokkurn veginn sömu upplýsingar og fjöldi herbergja, klósetta og hæða. Við ákveðum því að taka þrjár síðarnefndu út úr líkaninu. Í kjölfarið hækkar RMSE úr mátinu en aðlagað R^2 gerir það sömuleiðis að örlitlu leyti. Á þessu stigi máls er ekki endilega best að elta slíkar lýsistærðir, því seinna meir mun líkanið okkar njóta góðs af því að þessar breytur hafi verið fjarlægðar.

Athugum næst 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.



3. Ítarlegri gagnaúrvinnsla

Athugum nú aðrar breytur, sem ekki er eins augljóst hvort megi fjarlægja eða ekki. Við sjáum að tegund eignar og íbúðartegund kóða fyrir mjög svipuðum eiginleikum fasteignar. Í tegundaflokkinum er parhúsarflokkurinn sá eini sem mælist með almennilega svörun með p-gildi upp á rúmlega 0.33, þó mögulega að undanskildum einbýlishúsaflokknum sem er grunnflokkurinn í líkaninu. 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()
```

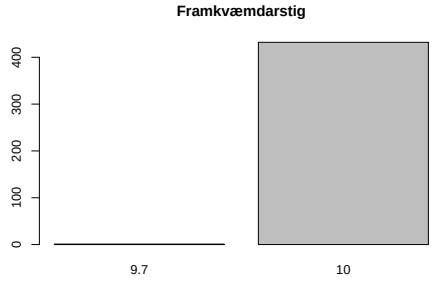
Hér sést að allar íbúðartegundir nr. 12 eru eignartegundin Íbúðareign. Þó eru nokkrar íbúðareignir sem falla í flokk 11 ásamt öllum hinum tegundum. Eftir samanburð á líkönum sem tóku annars vegar íbúðartegund og eignartegund út þá var það metið svo að betra væri að taka íbúðartegund út. RMSE lækkar og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

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

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

Önnur breyta sem mögulega er að rýra líkanið er undirmatssvæði. Skoðum hvernig gildin liggja þar. Taflan að ofan sýnir að langflestir punktarnir falla í undirmatssvæði 0 og 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 Kringlumýrar- og Miklubraut), skipta litlu máli í líkaninu eins og mál standa núna. Á hinn bóginn þá innihalda undirmatssvæði í eðli sínu færri punkta en matssvæðin og vænta má að þau séu valin af góðum ástæðum, til aðgreiningar innan svæða. 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 vegnaði í síðari greiningu á líkönunum.

Síðar - þegar vendipunktar innan skýribreyta voru skoðaðir - kom í ljós 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.




```
td5 = subset(td4, select = -stig10 )
lm.fifth = lm(nuvirdi ~ ., data = td5)
s.fifth <- summary(lm.fifth)
s.fifth$adj.r.squared
```

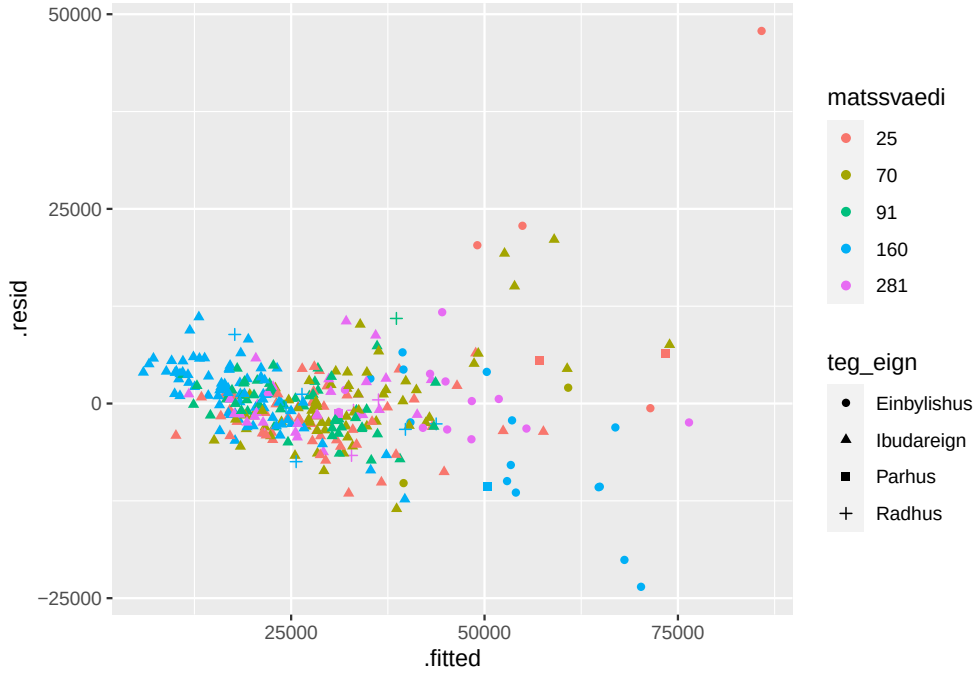
```
## [1] 0.8305896
```

Að þessu loknu stóð líkanið okkar þannig að RMSE á þjálfunarsetti mældist **5784.05** og gildi aðlagaðs R^2 mældist **0.83**.

4. Grunnskoðun á línuleika

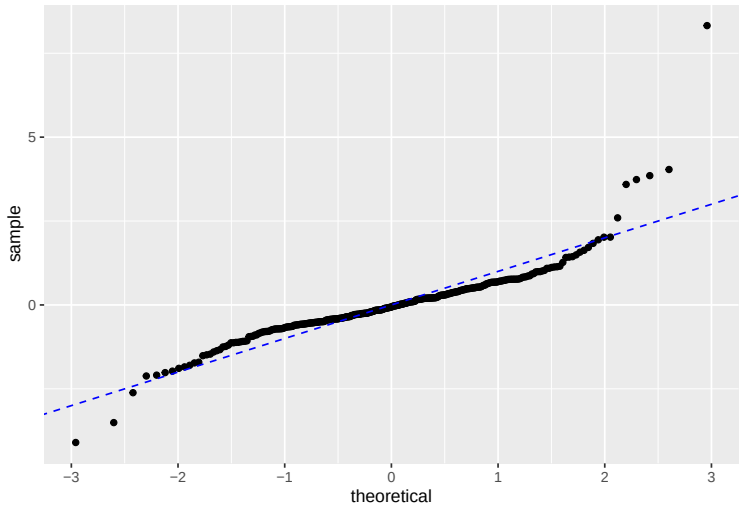
Lengra yrði ekki haldið með líkanið án þess að skoða undirliggjandi hegðun skýribreyta og óháðu breytunnar. Við viljum fella línulegt líkan að sambandi þeirra en sjaldnast er raunin sú að kraftar í okkar umhverfi, hvort sem litið er til eðlisfræði eða efnahagskerfa, hegði sér með fullkomlega línulegum hætti. Við viljum því greina það hvort merki um slíkan ólínuleika sé að finna í gögnunum okkar og eitt besta tól til þess er að plotta leifðina úr líkaninu á móti spágildunum. Ef líkanið heldur vel í við vaxandi eða minnkandi gildi, þá ætti leifðin að vera jöfn yfir svið spágilda.

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



Við sjáum aftur á móti greinileg merki um misdreifni í leifðinni, þar sem hún eykst með veldisfalli eftir því sem gildi núvirðis vex. Við ættum að geta séð þetta vel líka með QQ-plotti af leifðinni, þar sem leifðinni er raðað í stærðarröð og dreifingin ætti að falla að fræðilegum gildum út frá normaldreifingu leifðar út línulegu falli.

```
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 nokkru 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 sem nema hvers konar vörpun sé viðeigandi við y-dreifinguna okkar eru sérstaklega næmar fyrir slíkum frávikum.