Greining fasteignaverðs

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 ákveðinna svæða Reykjavíkur að gefnum upplýsingum um fasteignir. Við lýsum því hér hvaða skref við tókum í líkanasmíðinni, hvernig við völdum breytur, mátum gæði líkans og bárum saman líkön sem komu til greina. XXvalidate

Gögnin sem unnið var með voru fasteignamat eigna fyrir árið 2017, unnið árið 2016. Einungis var notast við gagnapunkta af fimm svæðum innan Reykjavíkur sem sjá má á korti hér að neðan; miðbæ frá Bræðraborgarstíg að Tjörn, Melar að sjó, Háaleiti og skeifan, Hólar og Berg í Breiðholti og loks Réttarholt. Á þessum svæðum voru í heildina 433 fasteignir með 21 breytum auk núvirðis. Breyturnar 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. Stig framkvæmdar er skali upp á 10 sem metur hvort húsnæðið sé tilbúið. Matssvæði eru hverfin sem sjást á myndinni hér að neðan og undirmatssvæði eru ákveðin svæði innan þeirra.

Til að byrja með skoðuðum við gögnin, meðaltöl, há- og lággildi ásamt því að umbreyta þeim yfir á rétt snið, svo sem kaupdegi yfir á dagsetningaform og flokkunarbreytum yfir å viðeigandi form.

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 veitir þar með engar upplýsingar um gagnapunktana.

```
# 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

#nums <- unlist(lapply(data, is.numeric))
numericNames <- colnames(dplyr::select(data, where(is.numeric)))
```

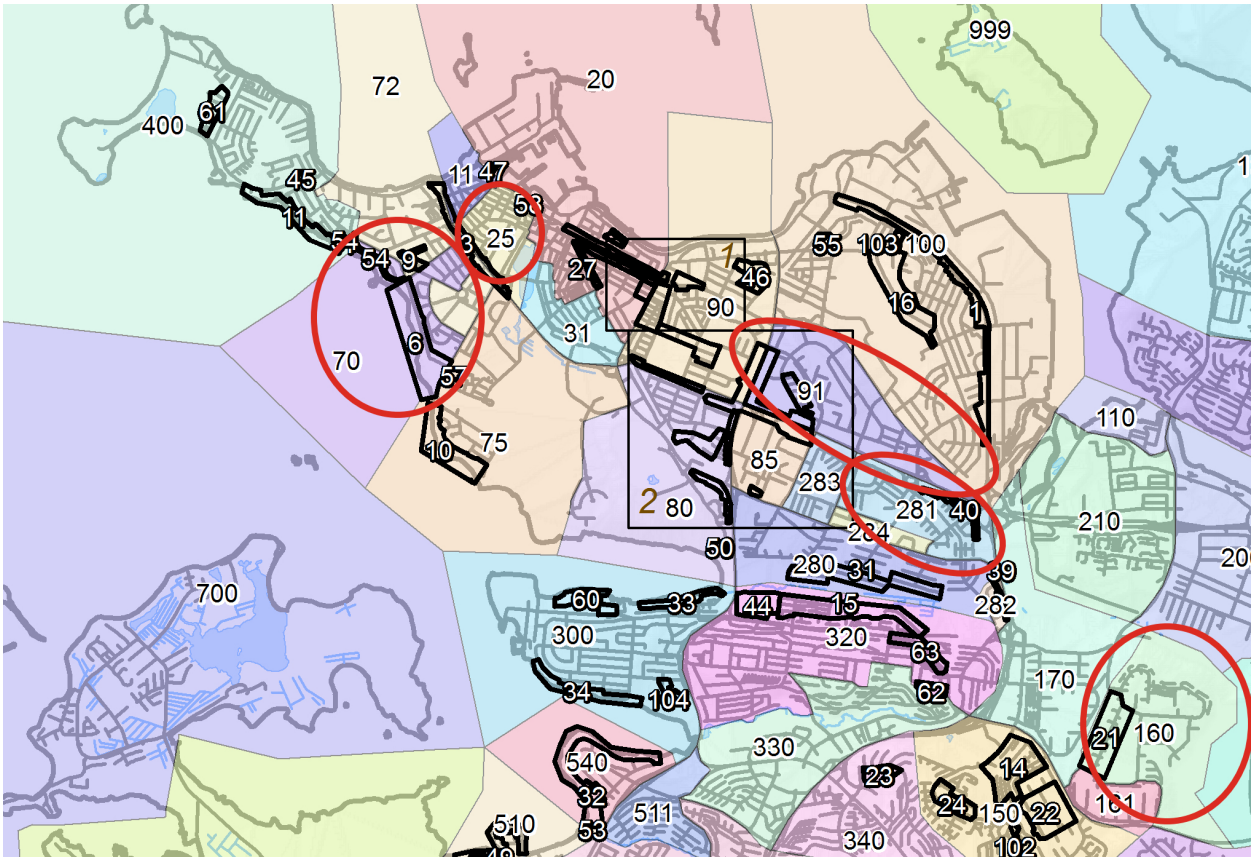


Figure 1: Svæði sem skoðuð voru eru rauðmerkt og nr. 70, 25, 91, 281 og 160.

	Núvirdi	Byggingarár	Nr. hæðar	Fjöldi lyfta	Fermetrafjöldi	Fjöldi hæða	Fjöldi bílastæða
	Min. : 5993	Min. :1901	Min. :0.000	Min. :0.0000	Min. : 21.90	Min. :1.000	Min. :0.00000
	1st Qu.: 20196	1st Qu.:1953	1st Qu.:1.000	1st Qu.:0.0000	1st Qu.: 65.50	1st Qu.:1.000	1st Qu.:0.00000
	Median : 25655	Median :1964	Median :2.000	Median :0.0000	Median : 85.30	Median :1.000	Median :0.00000
	Mean : 29163	Mean :1963	Mean :1.928	Mean :0.1663	Mean : 95.41	Mean :1.201	Mean :0.05312
	3rd Qu.: 33985	3rd Qu.:1974	3rd Qu.:3.000	3rd Qu.:0.0000	3rd Qu.:111.90	3rd Qu.:1.000	3rd Qu.:0.00000
	Max. :133665	Max. :2014	Max. :7.000	Max. :2.0000	Max. :289.30	Max. :3.000	Max. :2.00000

```
summary(data[numericNames]) %>%
  kbl(align = 'c', col.names = c("Núvirdi", "Byggingarár",  "Nr. hæðar",    "Fjöldi lyfta", "Fermetrafj
  kable_styling()
```

```
# library(table1)
# table1::table1(~., data[numericNames])
# Þessi er flottari en það vantar kvantíla
```

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

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 ~ ., data = train_data)
s.first = summary(lm.first)
```

Þessi fyrsta tilraun til að máta gögnin sýnir okkur við hverju má búast, hvaða breytur spila ekki lykilhlutverk og

Þetta líkan fær 5583.9400026 í RMSE og 0.8383239 í aðlagað R^2.

Er við skoðum spágildi líkansins út frá prófunargagnasetti fæst 6088.8611461 í RMSE.

Skoðum til viðbótar annað grunnlíkan sem mátar núvirði eingöngu við fermetraverð. Þetta líkan mætti hugsa sem grunnviðmið parsímóníunnar.

```
# Skoðum einnig annars konar grunnlíkan, sem tekur bara mið af fermetrum:
lm.simple <- lm(nuvirdi ~ ibm2, data = train_data)
s.simple <- summary(lm.simple)
sqrt(mean(residuals(lm.simple)^2))
```

```
## [1] 8721.104
```

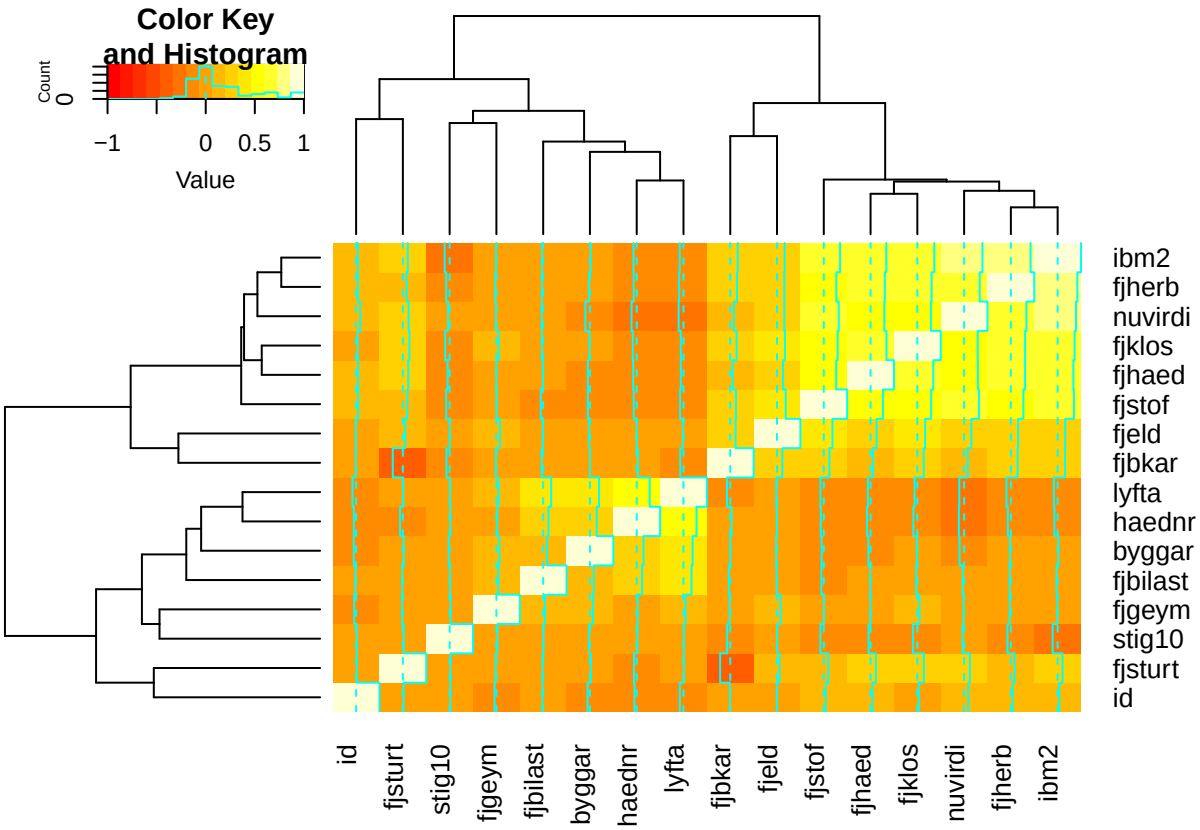
Við sjáum að það fær hærra RMSE en fyrsta líkanið okkar.

Þetta einfalda líkan fær 8313.261402 í RMSE á prófunarsetti, lægra en á þjálfunarsafni en hvort tveggja mjög hátt.

Hvað við tökum út og hvers vegna

Byrjum á að breyta breytunni 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.

```
# Flestar fasteignir hafa ekki lyftu og örfáar hafa fleiri en eina. Við ákváðum að breyta þeirri breytu
data[ ,"lyfta"] <- data[,"lyfta"]>0

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




```
# Sjáum cluster af hópum sem eru mjög líkir. Skoðum eigingildi:
X <- model.matrix(lm(nuvirdi ~ ., data[numericNames]))
eigenX <- eigen(t(X) %*% X)
condNumber <- max(eigenX$values)/min(eigenX$values)
```

Hér sést að það eru hópur af breytum sem sýna mikla fylgni hvor við aðra og við skoðun á eiginfylkjum XX sést að ástandstalan fyrir fylkið er gríðarhá og greinilegt að eitthvað sé á seyði hér. Við skoðum því eiginvigra með sérlega lág eigingildi.

```
eigenX$values
```

```
##  [1] 1.814369e+11 4.156443e+08 8.174698e+05 8.989131e+02 2.827940e+02
##  [6] 1.924734e+02 1.462716e+02 7.943030e+01 5.246434e+01 4.481396e+01
## [11] 3.426761e+01 2.186300e+01 1.833594e+01 1.224375e+01 3.189413e+00
## [16] 7.906379e-04
```

```
# Sjáum að eigingildi 16 er pínkulítið XXX ATHUGA, BREYTA 14 í 16
eigenX$vectors[, 16]
```

```
##  [1]  9.953095e-01 -1.426530e-05  3.698563e-05  1.384371e-04 -6.701872e-06
```

```
##  [6]  3.365934e-04  1.951746e-05  9.428943e-06  3.014166e-04 -5.543314e-04
## [11]  4.840391e-04  5.744766e-05 -1.112204e-04  1.584237e-04 -9.673812e-02
## [16]  5.392918e-09
```

```
colnames(data[numericNames])[c(5, 6, 10, 12, 13)]
```

```
## [1] "ibm2"   "fjhaed" "fjklos" "fjherb" "fjstof"
```

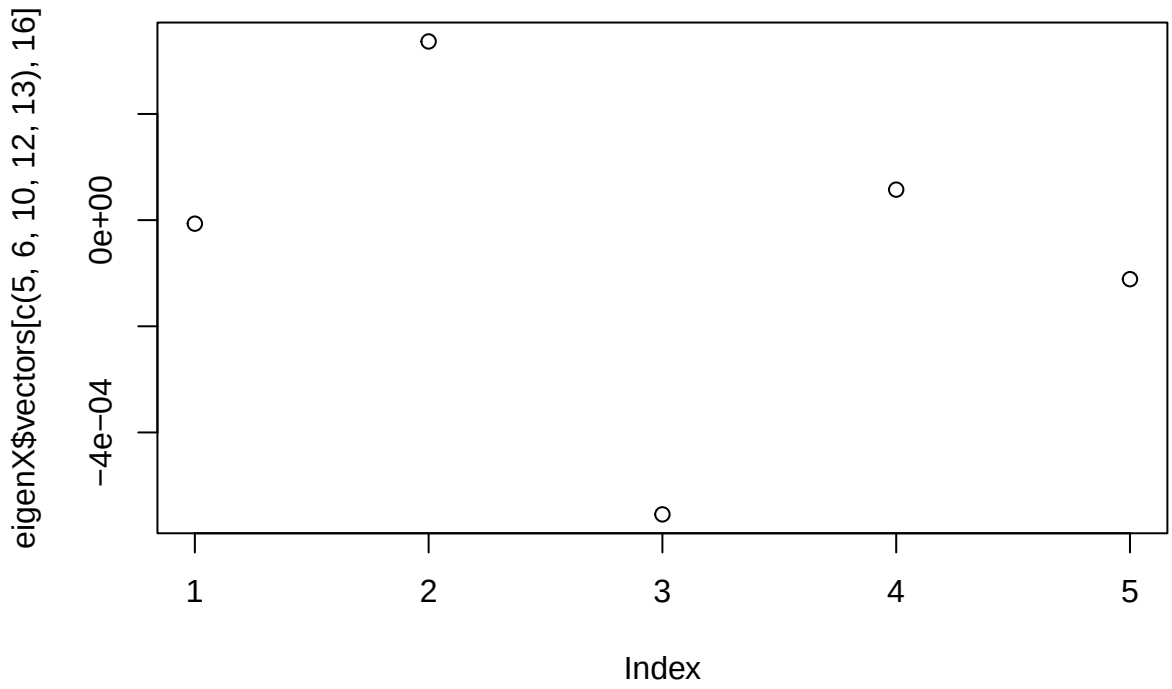
```
sum(eigenX$vectors[c(5, 6, 10, 12, 13), 16])
```

```
## [1] -0.0002782126
```

```
sum(eigenX$vectors[,16])
```

```
## [1] 0.8994271
```

```
plot(eigenX$vectors[c(5, 6, 10, 12, 13),16])
```



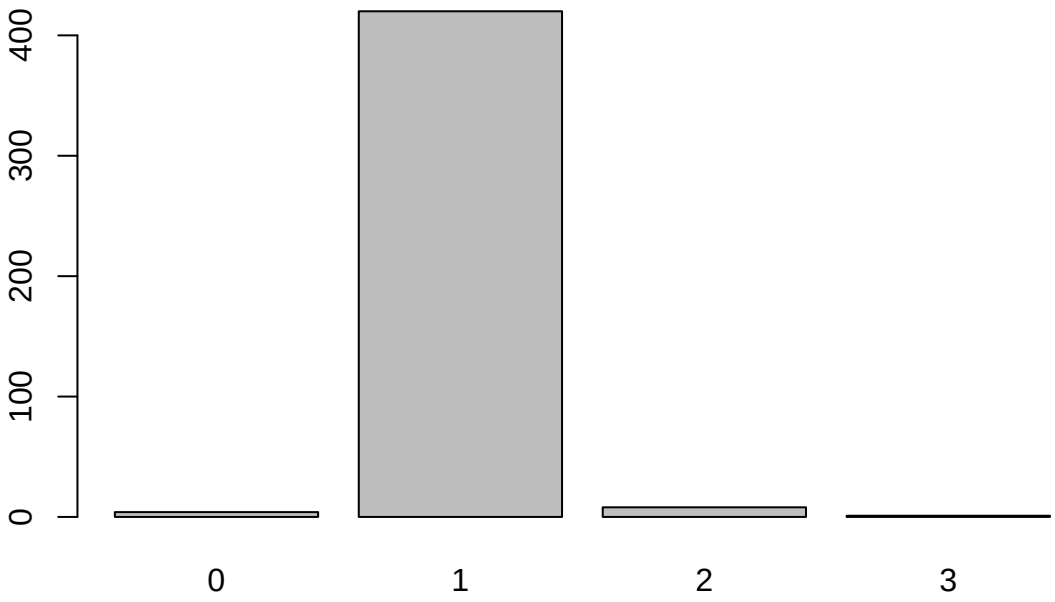

```
bad_actors <- eigenX$vectors[c(5, 6, 10, 12, 13), 16]
sum(bad_actors[c(1,5)]) - sum(bad_actors[c(2,3,4)])
```

```
## [1] 4.236807e-05
```

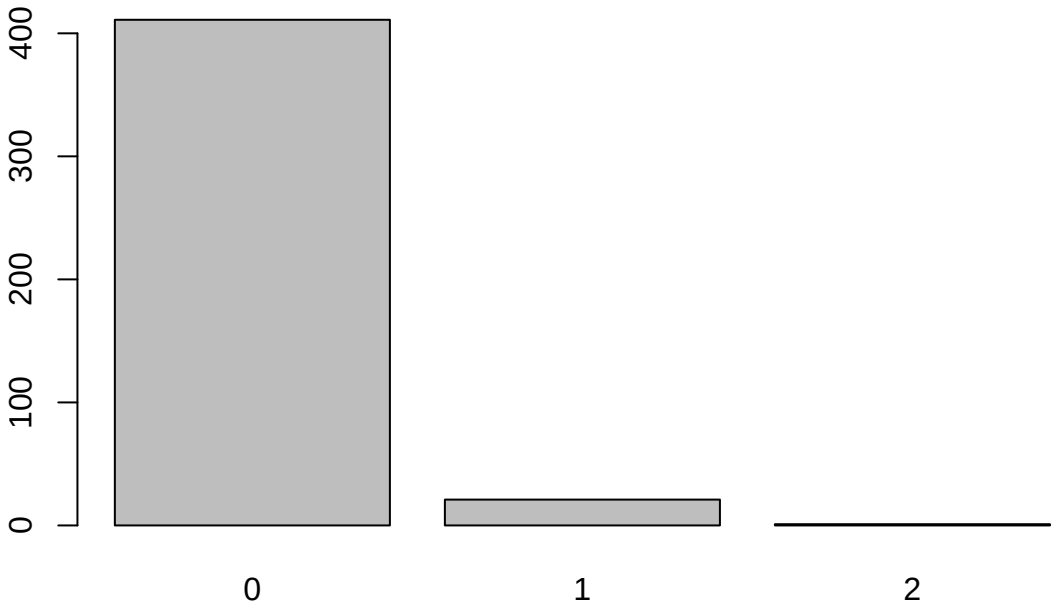
Við sjáum að fermetrafjöldi og fjöldi stofa tjá nokkurn veginn sömu upplýsingar og fjöldi herbergja, klósetta og hæða. Við ákveðum því að taka þrjár síðarnefndu út. Þegar við skoðum líkanið sem út úr því kemur sést að RMSE hækkar en aðlagað R^2 gerir það sömuleiðis að örlitlu leyti.

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.

```
barplot(table(data$fjeld))
```




```
barplot(table(data$fjbilast))
```



STIG 3: GAGNAÚRVINNSLA, MINNA AUGLJÓSIR FLOKKAR FJARLÆGÐIR

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.

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

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

```
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.

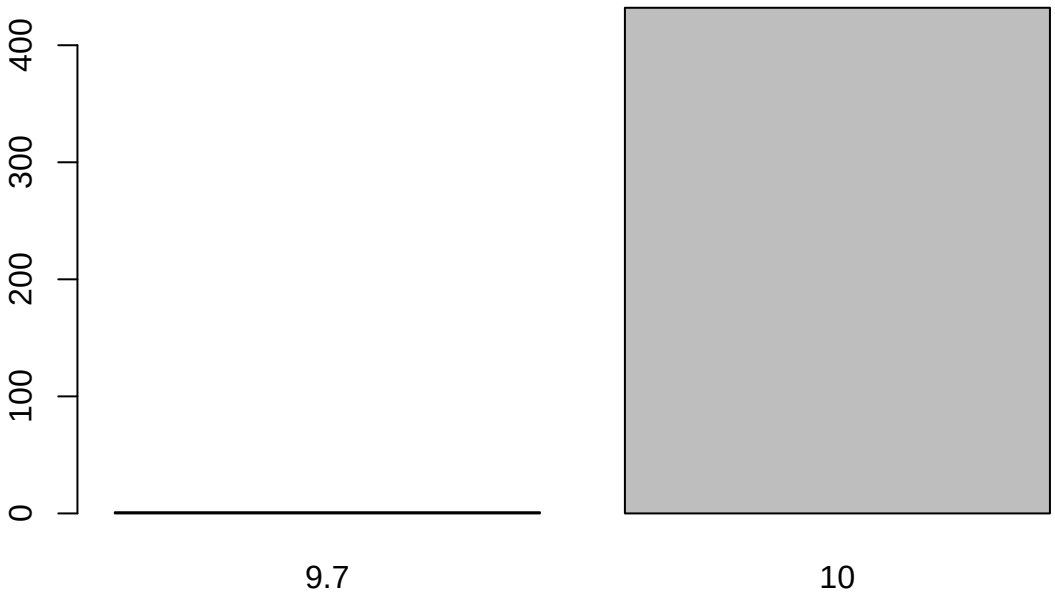
Önnur breyta sem mögulega er að rýra líkanið er undirmatssvæði. Skoðum hvernig gildin liggja þar.

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

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 þessar 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 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.

```
barplot(table(data$stig10))
```




```
td5 = subset(td4, select = -stig10 )
lm.fifth = lm(nuvirdi ~ ., data = td5)
sqrt(mean(residuals(lm.fifth)^2))
```

```
## [1] 5784.049
```

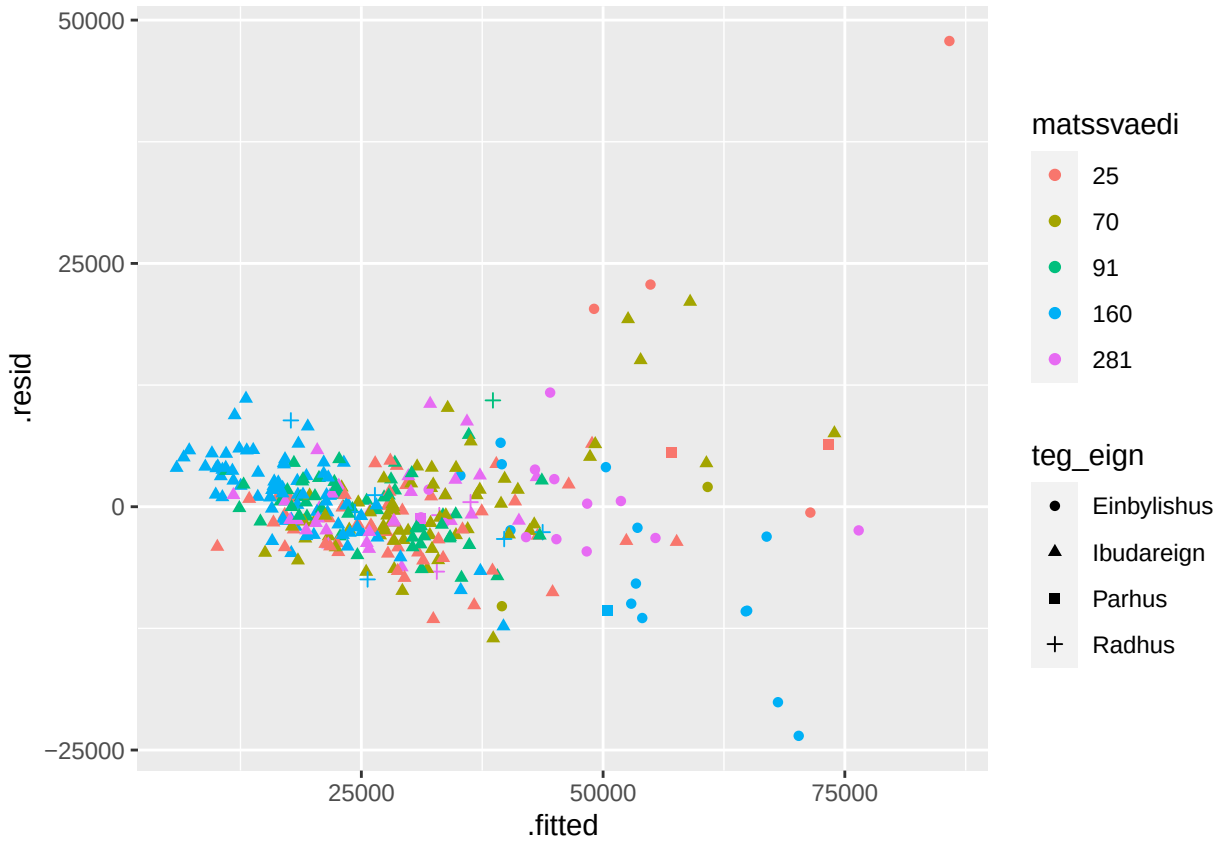
```
s.fifth <- summary(lm.fifth)
s.fifth$adj.r.squared
```

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

STIG 4A: HELDUR LÍNULEIKI? GRUNNSKOÐUN

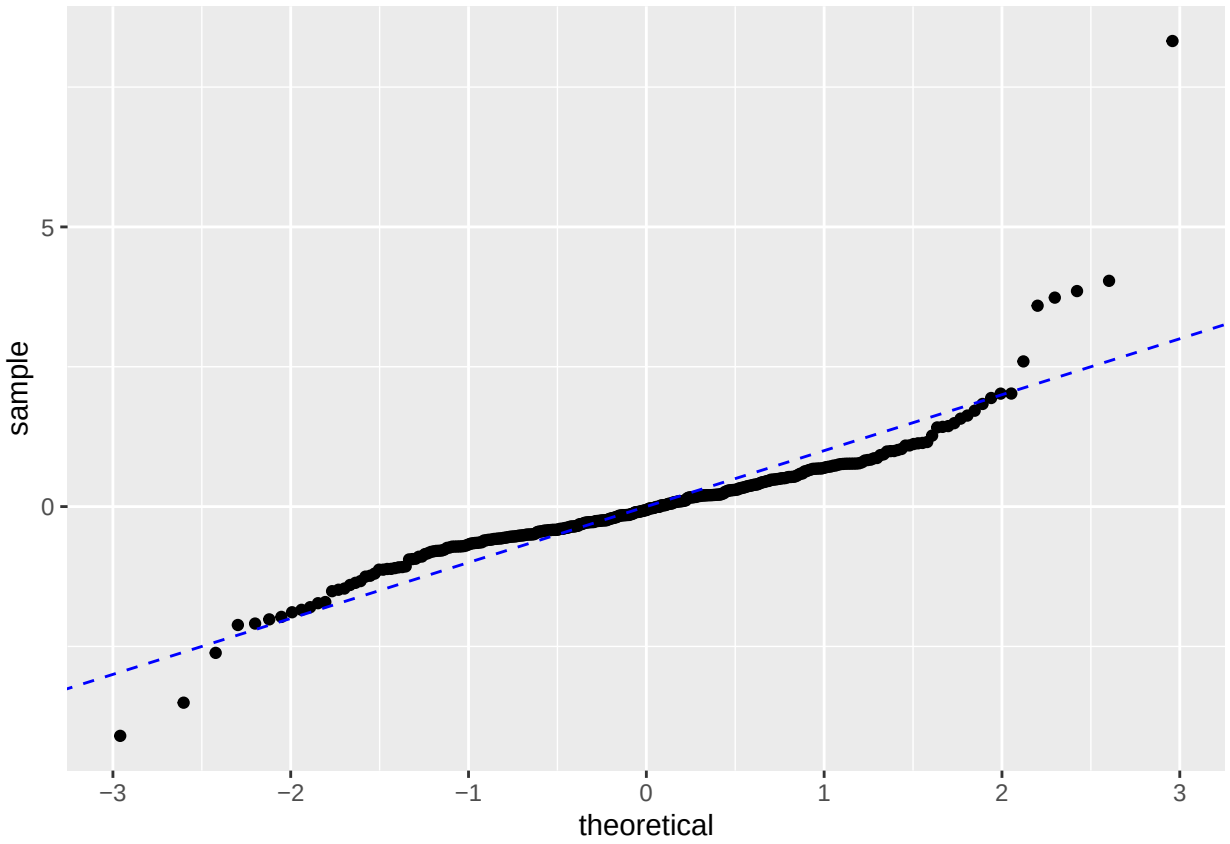
Byrjum á að skoða eitt mikilvægasta plottið, leifð á móti spá.

```
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, XXX MISDREIFNI 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 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 (XX skýribreyta? man ekki orðin). Réttast er þó að skoða fyrst útlaga og áhrifamikla punkta vegna þess að BoxCox og aðrar aðferðir eru sérstaklega næmar fyrir slíku.

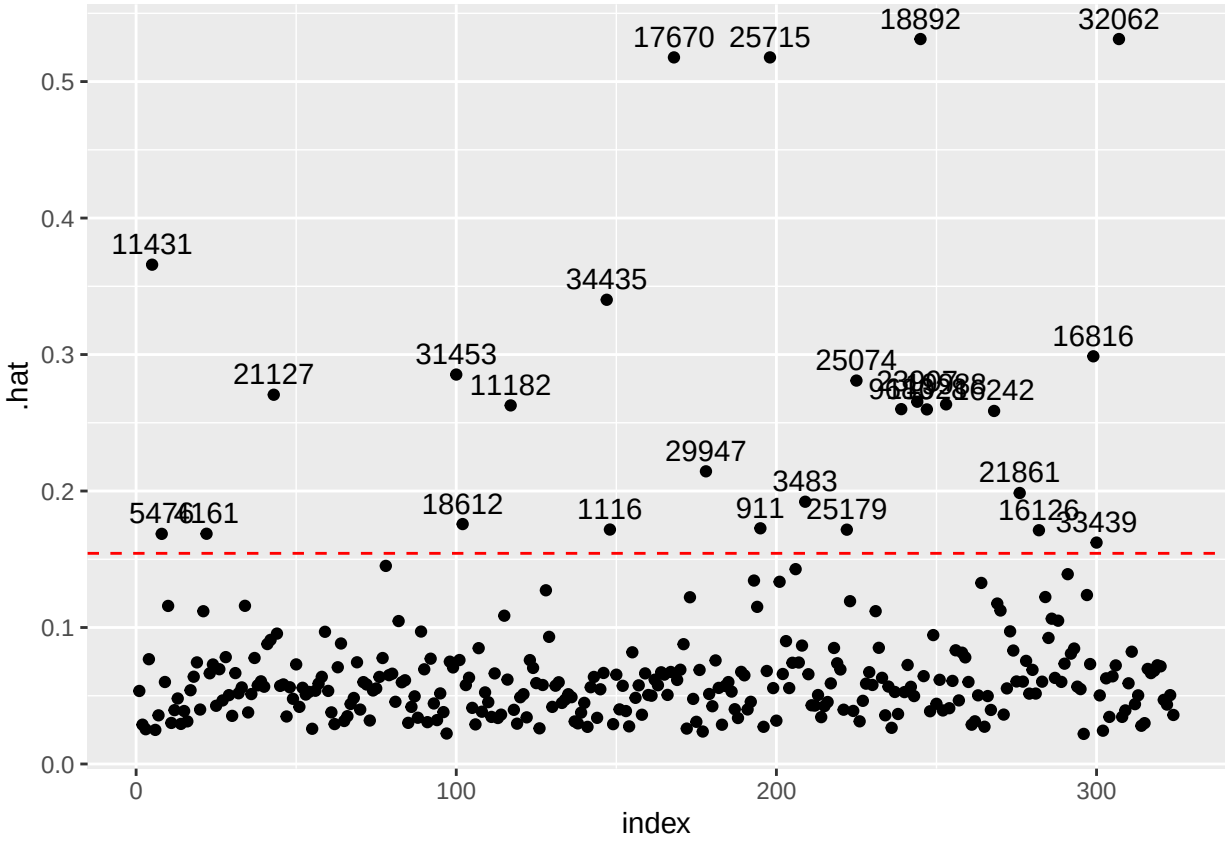
STIG 5: ÚTLAGAR OG MIKILVÆGIR PUNKTAR

Í fyrra kafla greindum við skýr merki um ólínuleika í gögnunum, bæði þegar horft var til QQ-plotts af leifðinni en einnig sást að leifðin jókst eftir því sem gildi núvirðis jókst. Við þessu er til gott ráð en það er að breyta y-breytunni okkar með falli sem gerir dreifinguna línulega.

Slíkar breytingar geta þó verið næmar fyrir útlögum og öðrum vendipunktum í gagnasettinu. Því er ráð að greina gögnin fyrst út frá slíkum punktum áður en við förum að eiga við eðli gagnapunktanna of mikið. Við byrjum á því að skoða vægi punkta út frá Mahalanobis fjarlægð en það var einmitt í gegnum þessa skoðun sem punktur 10944 skar sig frá fjöldanum en gildi hans mældist í þessari skoðun $h_{ii} = 1.0$. Eftirgrennslan leiddi í ljós að hann var eini punkturinn í undirsafninu sem tók ekki gildið 10 í framkvæmdarstigi (flokknum `stig10`) en þar mældist hann með 9.7 í einkunn. Af þessum sökum, var sá flokkur nær alfarið bundinn við þennan eina punkt og því var flokkurinn fjarlægður úr gagnasettinu. Að þessu loknu fengum við skýrari mynd af þessum áhrifamiklu gagnapunktum í skýribreytum okkar, líkt og grafið að neðan sýnir. Gagnapunktar sem skáru sig frá restinni voru skráðir til greiningar síðar meir.

```
fortData$rn <- row.names(fortData)
fortData$index <- 1:nrow(fortData)
fortData$.jackknife <- rstudent(lm.fifth)
n <- nrow(fortData)
p <- nrow(summary(lm.fifth)$coefficients)
```

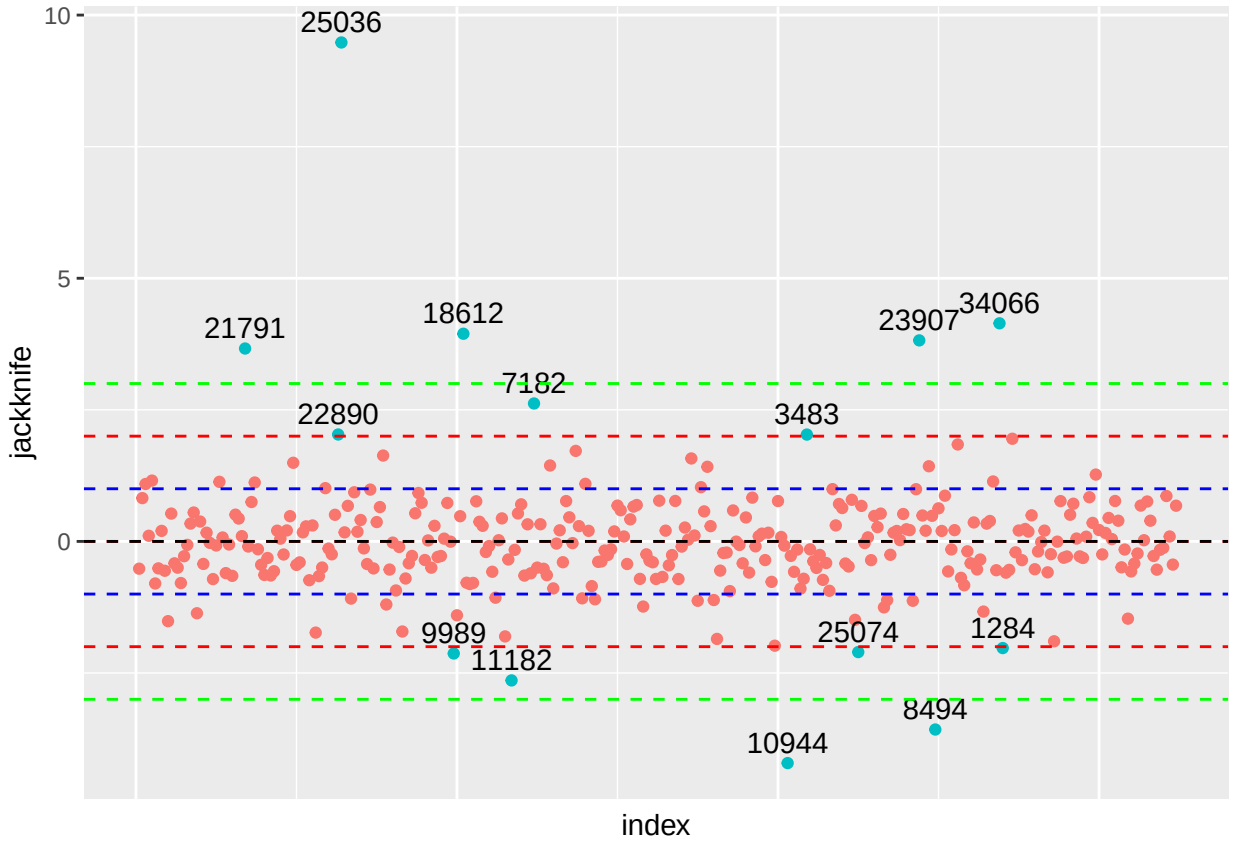
```
fortData %>%
  ggplot(aes(x = index, y = .hat)) +
  geom_point() +
  geom_hline(yintercept = 2*p/n, lty = 2, col = 'red') +
  geom_text(aes(label = ifelse(.hat > 2*p/n, rn, '')), hjust = 0.5, vjust = -0.5)
```



Grafið að ofan er ekki eina leiðin til þess að greina punkta sem skekkja gögnin. Vendipunktar líkt þeim að ofan gefa til kynna afgerandi gildi í skýribreytum, þar sem gildin skera sig frá öðrum innan sama flokks, en slík frávik geta einnig birst í óháðu breytunni okkar. Þar sem samspil skýribreytanna er lagað að því að máta óháðu breytuna okkar, geta slíkir útlagar valdið talsverðri skekkju í heildarmódelinu. Það eru einmitt slíkir öfgapunktar í óháðu breytunni okkar, verðmatinu, sem geta skekkt vörpun okkar á óháðu breytunni. Við greinum punktana annars vegar út frá jackknife firðinni og hinsvegar Cooks fird.

```
fortData %>%
  dplyr::select(.jackknife, rn, index) %>%
  gather(residual, jackknife, -index, -rn) %>%
  mutate(residual = factor(residual,
                           levels = c('.jackknife'))) %>%
  ggplot(aes(x = index, y = jackknife, label=rn)) +
  geom_point(aes(color = ifelse(abs(jackknife)>2, "darkred", "black"))) +
  geom_text(aes(label=ifelse(abs(jackknife)>2,rn,'')),hjust=0.5,vjust=-0.5) +
  geom_hline(yintercept = 0, lty = 2, col = 'red') +
  theme(legend.position="none",
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank()) +
  geom_hline(yintercept = 1, lty = 2, col = 'blue') +
  geom_hline(yintercept = -1, lty = 2, col = 'blue') +
  geom_hline(yintercept = 2, lty = 2, col = 'red') +
```

```
  geom_hline(yintercept = -2, lty = 2, col = 'red') +
  geom_hline(yintercept = 3, lty = 2, col = 'green') +
  geom_hline(yintercept = -3, lty = 2, col = 'green') +
  geom_hline(yintercept = 0, lty = 2) + expand_limits(x = c(0, 282))
```

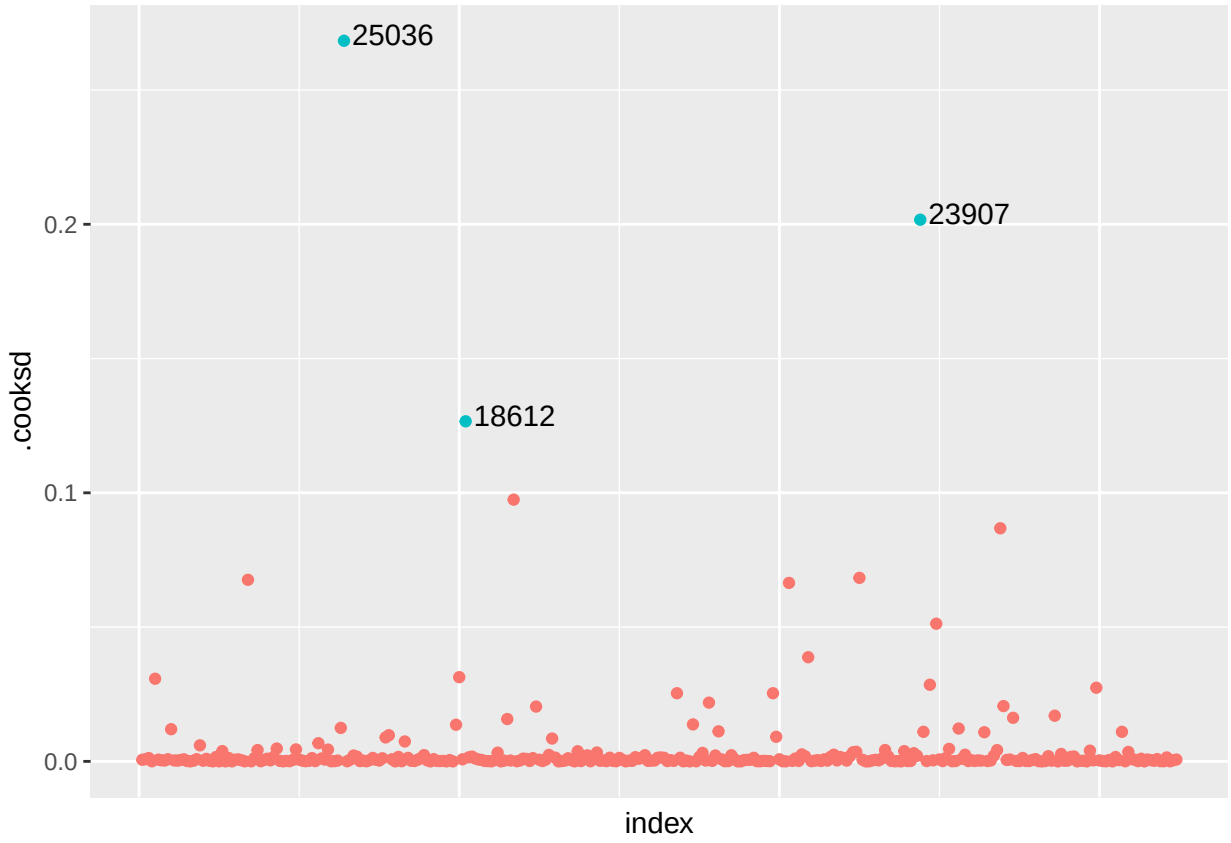


Myndin að ofan sýnir jackknife gildi fyrir alla punkta gagnasettsins, þar sem svarta punktalínan sýnir gildið núll. Bláa línan merkir gildi 1, þarnæst rauða línan gildið 2 og græna línan gildið 3. Við mátum það svo að best væri að skoða einungis þá gagnapunkta sem mældust yfir þremur, þar sem kunnátta okkar um söfnun upplýsinganna og undirliggjandi jöfnur verðmatsins væru ekki nægar til að skera út gagnapunkta, nema augljóst væri að þeir væru vandamál.

Síðasta tólið til að greina frávik í núvirðisflokkinum var að skoða útlaga í gagnasettinu, þá punkta sem mældust nógu langt frá almennri dreifingu að þeir réttlættu frekar skoðun. Plottið að neðan sýnir svokallaða Cooks fird, sem lýsir þessu fyrir bæri á skíran hátt, og aftur sjáum við nokkra skíra punkta sem skera sig frá restinni. Í öllum þessum prófunum voru afgerandi gagnapunktar skráðir og í kjölfarið voru þeir skoðaðir saman út frá öllum tölfræðiverkfærunum sem minnst hefur verið á.

```
alpha <- 0.05
tCrit <- qt(p = 1 - alpha/(2 * n), n - p - 1)
outliers = fortData$rn[c(which(abs(fortData$.jackknife) > tCrit))]

fortData %>%
  ggplot(aes(x = index, y = .cooksd, label=rn)) +
  geom_point(aes(color = ifelse(.cooksd>0.1, "darkred", "black"))) +
  geom_text(aes(label=ifelse(.cooksd>0.1,rn,'')),hjust=-0.1,vjust=0.2) +
  theme(legend.position="none",
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```

rn	highLeverage	outlier	influential	totalMarks
21791	FALSE	FALSE	FALSE	0
25036	FALSE	TRUE	TRUE	2
4216	FALSE	FALSE	FALSE	0
18612	TRUE	TRUE	FALSE	2
11182	TRUE	FALSE	FALSE	1
17670	TRUE	FALSE	FALSE	1
25715	TRUE	FALSE	FALSE	1
23907	TRUE	FALSE	TRUE	2
18892	TRUE	FALSE	FALSE	1
8494	FALSE	FALSE	FALSE	0
34066	FALSE	TRUE	FALSE	1
32062	TRUE	FALSE	FALSE	1



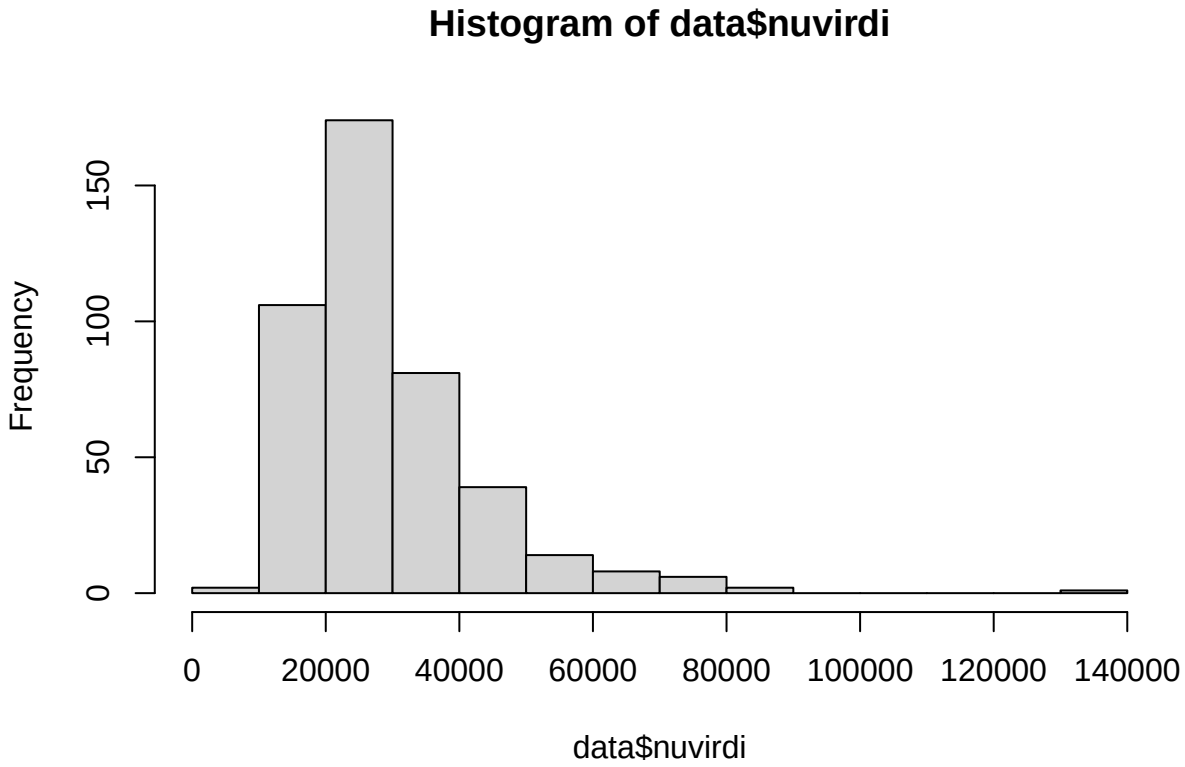

```
theoreticalT <- qt(p = 1 - 0.05/(2 * n), df = n - p - 1)
suspicious <- c(25036, 23907, 21791, 18612, 34066, 11182,
                8494, 4216, 17670, 25715, 18892, 32062)
fortData %>%
  filter(rn %in% suspicious) %>%
  mutate(highLeverage = .hat > 2*p/n,
         outlier = abs(.jackknife) > theoreticalT,
         influential = .cooksd > 0.15) %>%
  dplyr::select(rn, highLeverage, outlier, influential) %>%
  mutate(totalMarks = highLeverage + outlier + influential) %>%
  kbl(align = 'c') %>%
  kable_styling()
```

	nuvirdi	kdagur	teg_eign	byggar	ibm2	matssvaedi	.fitted
25036	133665	2015-06-01	Einbylishus	1927	273.9	25	85813.88
18612	80062	2014-09-25	Ibudareign	1946	176.8	70	59001.53
23907	71862	2015-05-21	Ibudareign	1952	123.2	70	52584.17

Punktunum voru gefin stig, eftir því hversu oft þeir lendu utan við gefin viðmið í skoðununum þremur. Enginn punktur fékk stig úr öllum þremur mælingum en þrír mældust með tvö stig. Af þeim punktum var einn sem sérstaklega skar sig út, gagnapunktur 25036. Við frekari skoðun kom í ljós að hann svarar til dýrustu fasteign í öllu gagnasettinu sem við unnum með (þ.e. innan þeirra undirsvæða sem greiningin tekur til). Sé litið til dreifingar verðs í gagnasettinu sést að þessi fasteign er langtum verðmætari en allar aðrar. Því var það svo álitið að eitthvað sérstakt væri hér í gangi. Við reyndum að komast að því hvaða fasteign þetta væri en fasteignanúmerin úr gagnasettinu eru úreld. Frá og með 8. apríl 2018 var fasteignanúmerum breytt og þau eldri sem við vinnum með ganga ekki í leit á Þjóðskrá. Út frá matssvæðinu er ljóst að eignin liggur á milli Bræðraborgarstígs og Tjarnarinnar og því er allt eins víst að um sé að ræða forsætisráðherrabústaðinn við Tjarnargötu, Skólabæ eða aðra þjóðþekkta eign. Við ákváðum því að fjarlægja punktinn, þar sem hann væri ekki lýsandi fyrir restina af gagnasettinu.

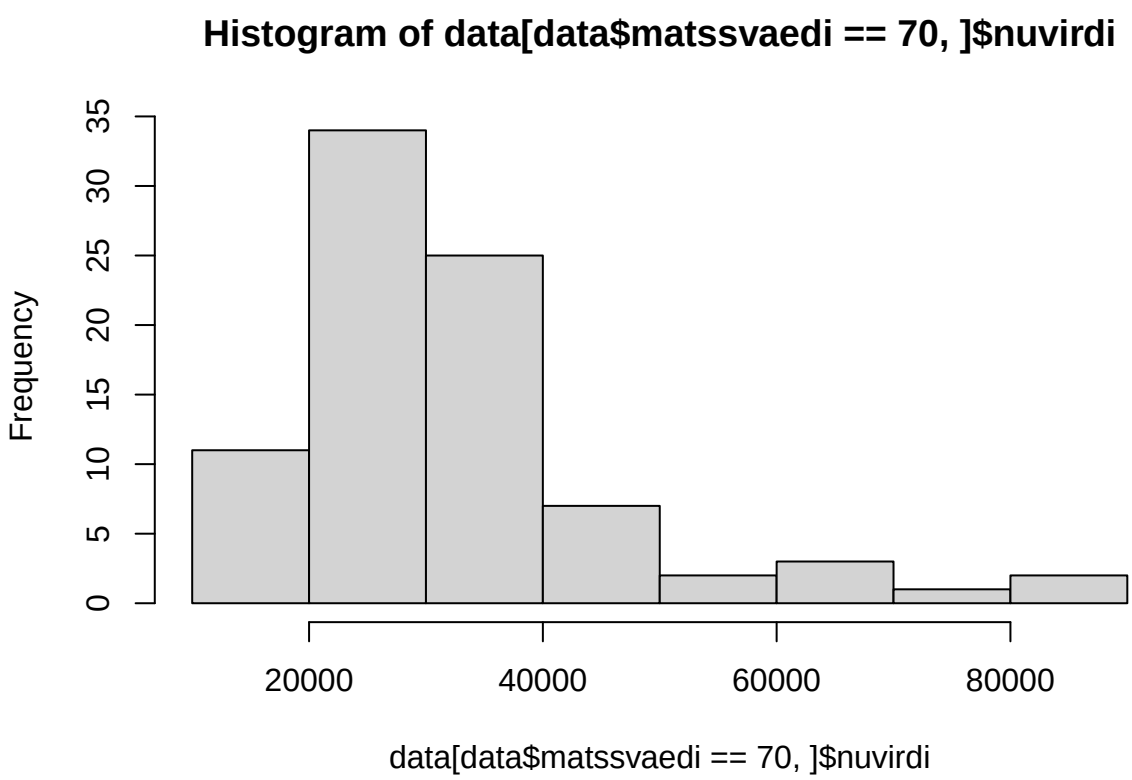
```
possible_rejects = c(23907, 25036, 18612)

fortData[fortData$rn %in% possible_rejects, c(1,2,3,4,7,12,18)] %>%
  kbl(align = 'c') %>%
  kable_styling()
```

```
hist(data$nuvirdi)
```



Hinir tveir voru metnir á miklu lægra verði en þeir ættu að vera. Báðir liggja við Ægissíðu og eru metnir á verð sem kemur heim og saman við aðrar eignir þar en módelið okkar spáði þeim verði u.þ.b. 20 milljónum lægra en raun bar vitni. Skoðun á þessum punktum leiddi ekkert sérstaklega skrítið í ljós og því var ákveðið að hér væri við módelið okkar að sakast, ekki punktana. Því var ekki ákveðið að fjarlægja þessa gagnapunkta úr settinu.

```
hist(data[data$matssvaedi == 70,]$nuvirdi)
```



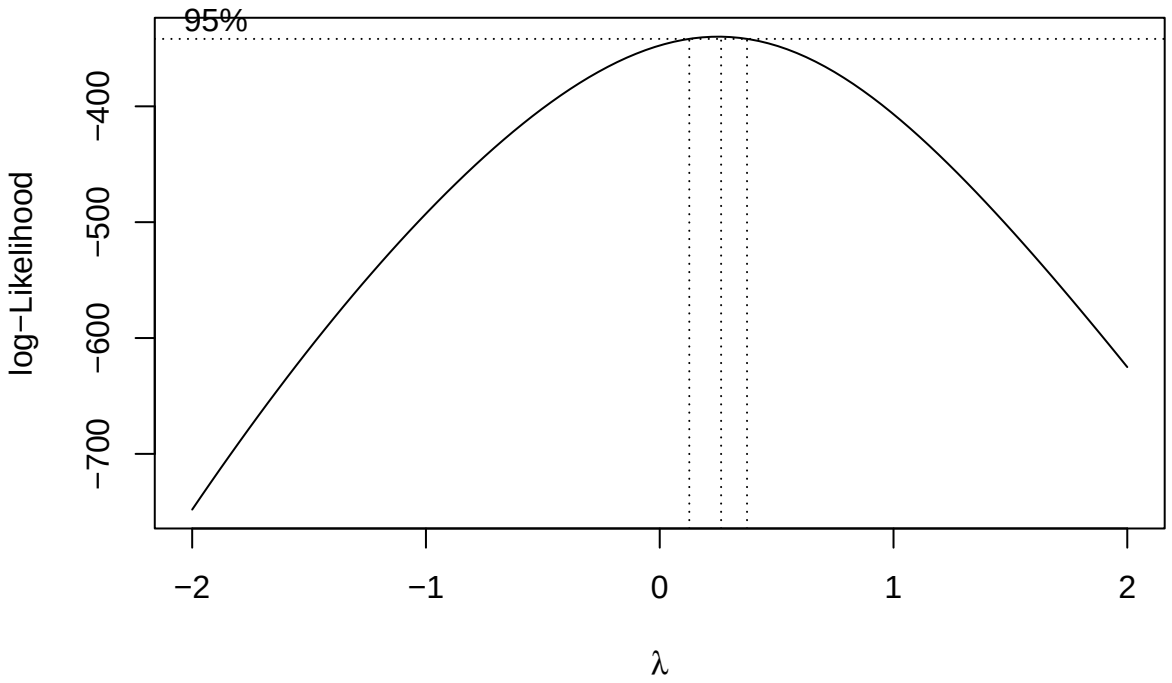

```
td6 <- td5[td5$id != 25036,]
```

STIG 4B: HELDUR LÍNULEIKI? UMBREYTINGAR

Nú þegar gagnapunktar höfðu verið skoðaðir og einn þeirra fjarlægður, auk þess sem matið um framkvæmdarstig íbúða hafði verið fjarlægt, var talið tímabært að umbreyta núvirðisbreytunum til að fella þær betur að normaldreifingu. Almennt fall til þess að umbreyta slíku kallast BoxCox og byggir á einungis einum fasta, λ, sem hægt er að meta með innbyggðu falli í R. Fyrir óháða breytu y er fallið metið sem $boxcox(y)|_{y=0} = \log(y)$ þegar $\lambda = 0$ en annars metið sem

$$boxcox(y)|_{y\neq 0} = \frac{y^\lambda - 1}{\lambda}.$$

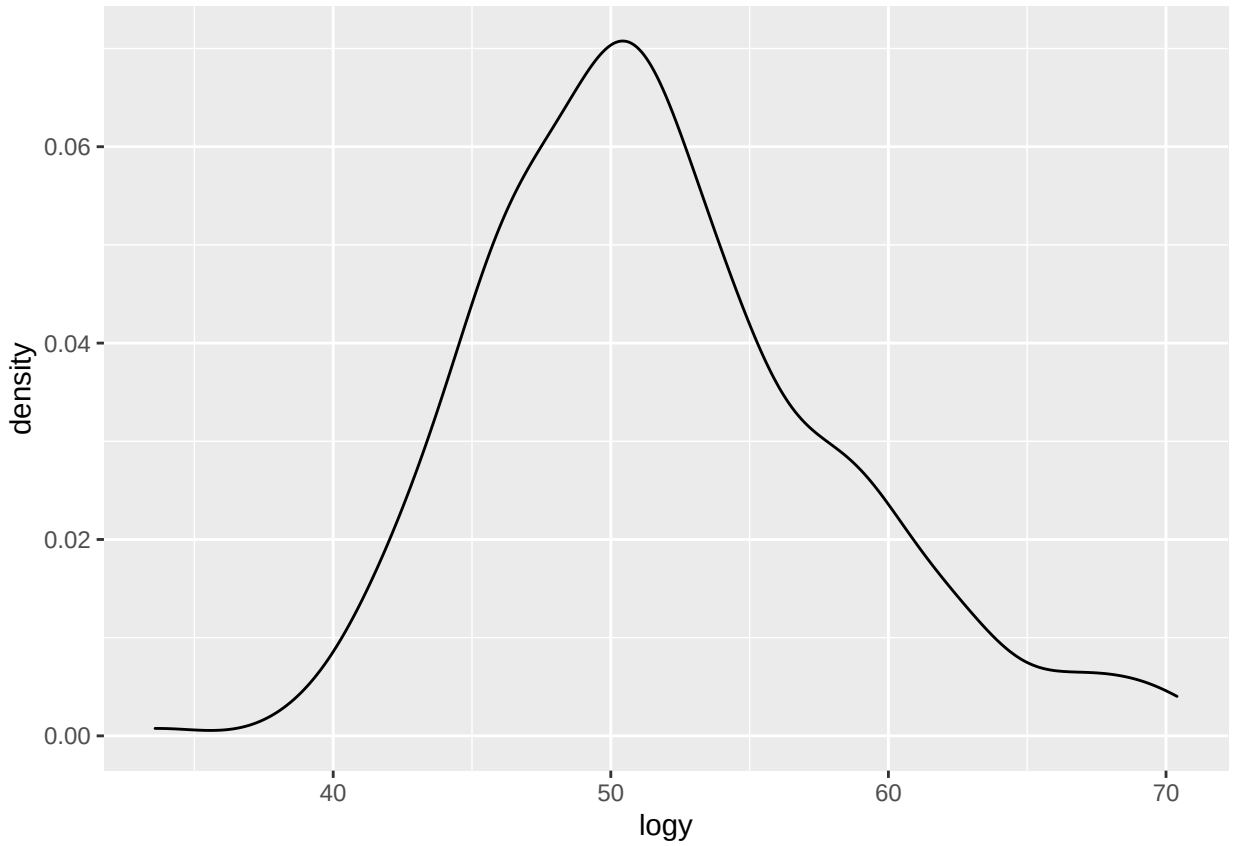
```
bcTest <- boxcox(lm( nuvirdi ~ . -id, td6))
```



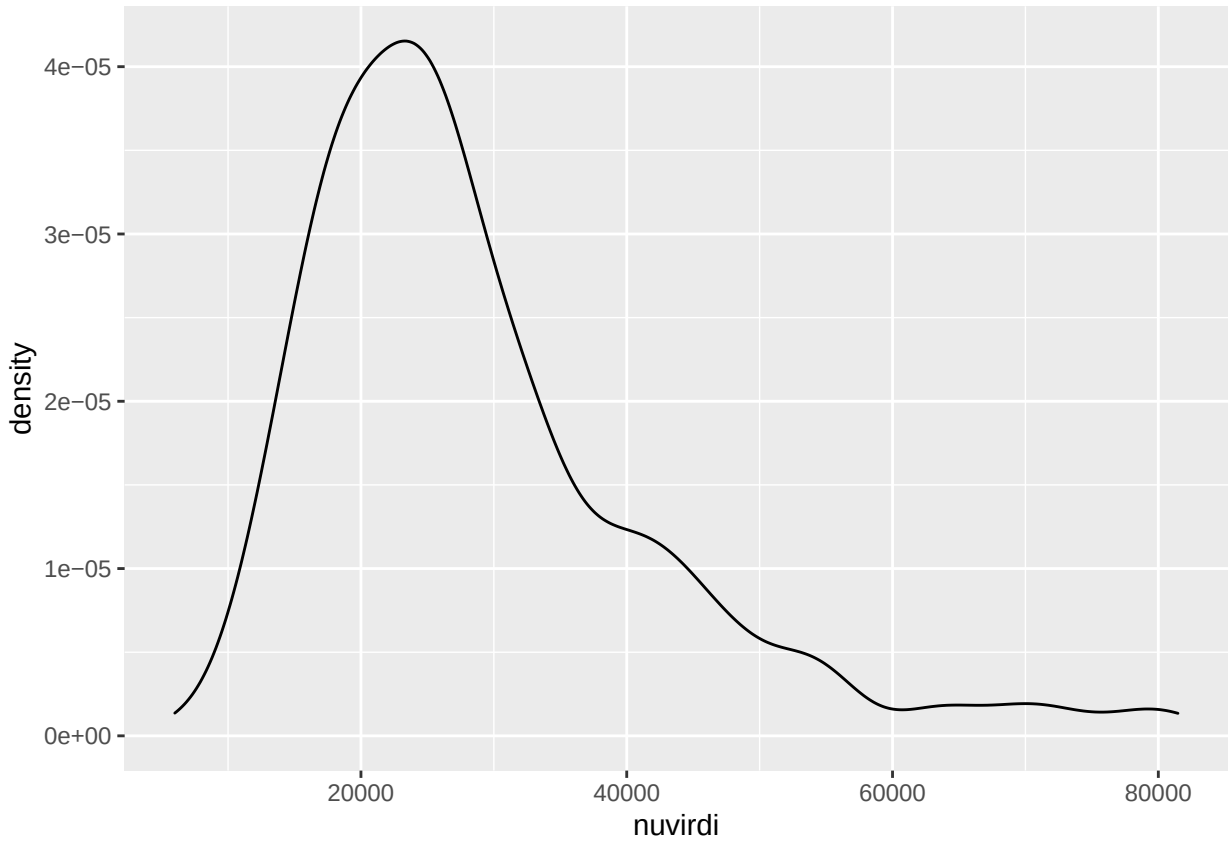

```
indexOfLLPeak <- which.max(bcTest$y)
lambda <- bcTest$x[indexOfLLPeak]
```

Slík greining gaf gildi $\lambda = 26/99$, sem er langt fyrir neðan 5. Í einhverjum tilvikum geta útlagar skekkt matið svo að gildi λ verður mjög stór tala og er það varhugavert að fylgja gildum sem liggja nálægt eða yfir 5. Í þessu tilfelli virðist gildið nokkuð hóflegt. Við umbreytum því breytunni okkar eftir þessum stika og sjáum að munurinn á dreifingunni er ansi mikill, þar sem dreifingin er mun normaldreifðari eftir umbreytinguna.

```
par(mfrow = c(1,2))
td6$logy <- bxcx(td6$nuvirdi, lambda, InverseQ = FALSE, type = "BoxCox")

lm.sixth = lm(logy ~ . -nuvirdi -id, data = td6)
s.sixth <- summary(lm.sixth)

fortLogy = fortify(lm.sixth)
# Sjáum að logy hegðar sér betur (meira línulega) en venjulega núvirðið:
# (sameina á eitt plott með facet wrap eða eitthvað)
fortLogy%>% ggplot(aes(x = logy)) +
  geom_density()
```




```
fortLogy %>% ggplot(aes(x = nuvirdi)) +
  geom_density()
```




```
resid_sixth <- bxcx(residuals(lm.sixth), lambda, InverseQ = TRUE, type = "BoxCox")
```

```
## Warning: modified inverse Box-Cox transformation used
```

Að þessu loknu er réttast að skoða grunnmódelið aftur og kanna hvernig skekkjunni ber saman við gildið áður. Við skiljum nú breytuna núvirði frá gögnunum, þar sem hún er ekki hluti af módelinu lengur. Við fellum módelið að gagnasettinu okkar og fáum sláandi niðurstöðu, RMSE mælist nú 10.1031897, sem er sláandi munur frá fyrri módelum. Við viðurkennum það að það runnu á okkur tvær grímur þegar þessar niðurstöður bárust; annars vegar gleði yfir því að módelið stæði sig svona miklu betur eftir þessa umbreytingu og hins vegar kvíði yfir því að þessar niðurstöður væru tortryggilegar. Frekari rýni leiddi hins vegar ekkert varhugavert í ljós, svo við héldum áfram að reyna að betrumbæta forritið ennþá frekar.

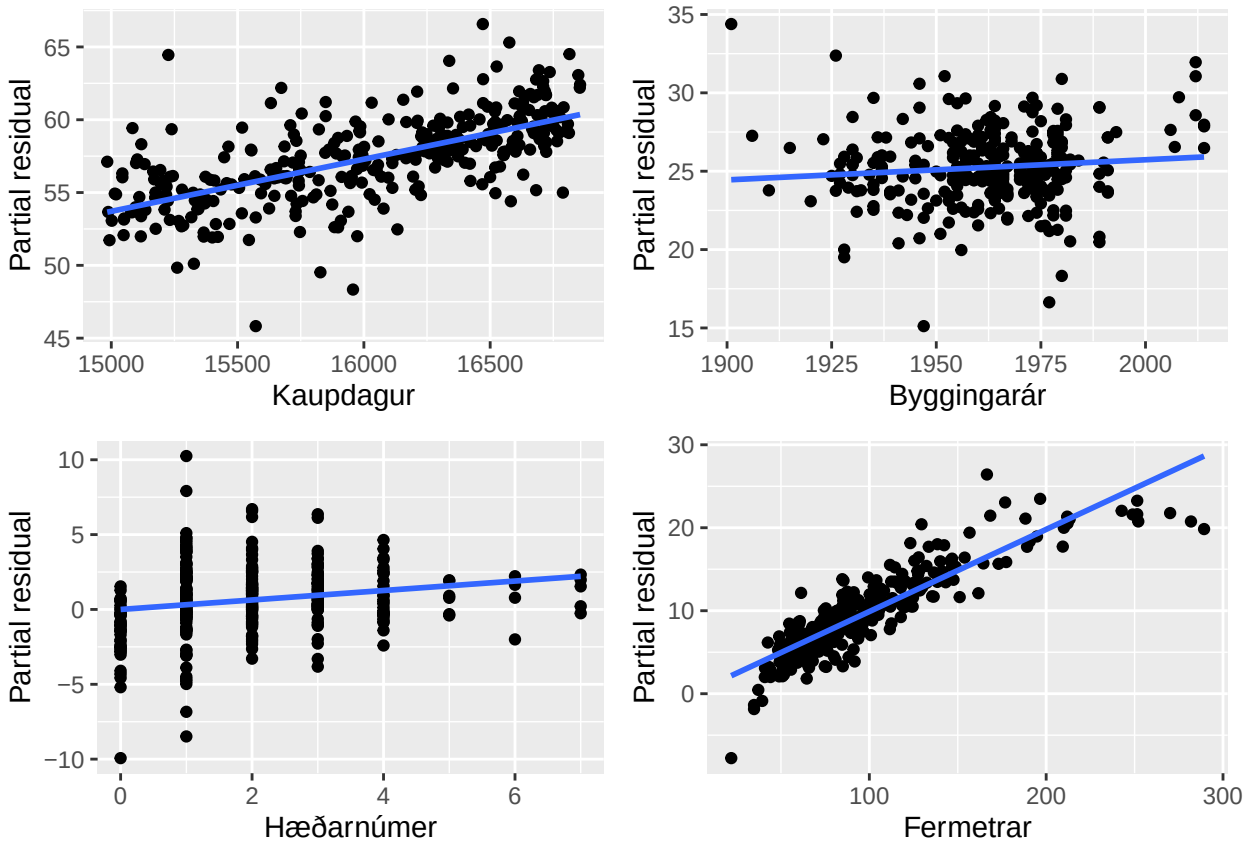
Kafli 5: Varpanir skýribreyta

Næst var litið til skýribreytanna en vera má að samband þeirra við óháðu breytuna sé ekki línulegt. Að neðan má sjá þrjár samfelldar breytur og eina ósamfellda en hún lýsir því á hvaða hæð í húsinu fasteignin er. Í þeirri ósamfelldu má greina nokkra ölduhreyfingu, þar sem fasteignir á miðjuhæðum virðast almennt vera dýrari en þær á grunnhæð eða ofar en fjórða hæð. Þessi alda gefur til kynna ólínulegt samband milli hæðar og fasteignamats, sem einföld lína fangar ekki almennilega. Sömu sögu er að segja um fermetrana en þar er tilhneiging hallatölunnar að minnka jafnt og þétt eftir því sem fermetrar aukast. Þar sem hreyfingin þar er einsleit er líklegt að stuðlar í öðru veldi geti náð utan um þessa breytinguna en í tilfelli ölduhreyfingarinnar fyrir hæðina er líklegra að þriðja veldi breytunnar geti náð utan um hegðun hennar.

```
fortLogy$rn <- row.names(fortLogy)
fortLogy$index <- 1:nrow(fortLogy)
fortLogy$.jackknife <- rstudent(lm.sixth)
n5 <- nrow(fortLogy)
p5 <- nrow(summary(lm.sixth)$coefficients)
```

```
betas <- c("(Intercept", "Kaupdagur", "teg_eignIbudareign", "teg_eignParhus", "teg_eignRadhus",
           "Byggingarár", "Hæðarnúmer", "lyftaTRUE", "Fermetrar", "fjbkar", "fjsturt", "fjstof", "fjgeym
           "stig10", "matssvaedi70", "matssvaedi91", "matssvaedi160", "matssvaedi281",
           "undirmatssvaedi3", "undirmatssvaedi6", "undirmatssvaedi21", "undirmatssvaedi28",
           "undirmatssvaedi40", "undirmatssvaedi48", "undirmatssvaedi54", "id")

beta1 <- summary(lm.sixth)$coefficients[2, 1]
beta2 <- summary(lm.sixth)$coefficients[3, 1]
beta3 <- summary(lm.sixth)$coefficients[4, 1]
beta4 <- summary(lm.sixth)$coefficients[5, 1]
beta5 <- summary(lm.sixth)$coefficients[6, 1]
beta6 <- summary(lm.sixth)$coefficients[7, 1]
beta7 <- summary(lm.sixth)$coefficients[8, 1]
beta8 <- summary(lm.sixth)$coefficients[9, 1]
beta <- c(beta1, beta2, beta3, beta4,beta5, beta6, beta7, beta8)
plots <- list()
for(i in 1:length(beta)) {
  regressor <-  model.matrix(lm.sixth)[, (i + 1)]
  partialRes <- fortLogy$.resid + regressor * beta[i]
  tibble(x = regressor,
         y = partialRes) %>%
    ggplot(aes(x = x, y = y)) +
    geom_point() +
    stat_smooth(method = 'lm', se = F) +
    labs(x = betas[i + 1],
         y = 'Partial residual') -> plots[[i]]
}
cowplot::plot_grid(plots[[1]], plots[[5]], plots[[6]], plots[[8]],
                   nrow = 2, ncol = 2)
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



Við litum einnig til efri breytanna tveggja, kaupdags og byggingarárs. Kaupdagurinn virðist vera nokkuð jafndreifður eða í það minnsta nokkuð vel mátaður af línu. Í byggingarárinu er mögulega hægt að sjá drög að því að fasteignir byggðar á árunum 1950-1980 séu verðminni en eldri og yngri eignir. Slík hegðun gæfi til kynna að annars stigs breyta gæti lýst hluta sambandsins betur en línuleg breyta. Við nánari skoðun sýndu tölfræðimöt ekki fram á að slíkt væri réttlætanlegt en í ljós kom að fermetraverð naut verulegs góðs af annars stigs breytu og þriðja stigs breyta var möguleg. Við létum þó við sitja með annars stigs breytu fyrir fermetra. Fyrir hæðarnúmer skilaði annars stigs breyta litlu nýju en þriðja stigs breytan sýndi fram á mikla umbót. Sökum þess að breytur af lægra stigi má ekki fjarlægja á undan breytum af hærra stigi, voru annars og þriðja stigs breytur settar inn fyrir hæðarnúmer.

Við ákváðum einnig að fjarlægja breytuna fyrir sturtu í íbúðunum, þar sem hún hafði aldrei fengið almennilegt p-gildis mat í neinum útfærslum líkansins og virtist tjá fyrir marga sömu eiginleika og fjöldi baðkara tjáði fyrir um. Eftir þessar umbreytingar sjáum við eftirfarandi breytingar á gröfunum að ofan:

```
lmOrthoPolyIBM2 <- lm(td6$logy ~ poly(td6$ibm2, 3))
summary(lmOrthoPolyIBM2)
```

```
##
## Call:
## lm(formula = td6$logy ~ poly(td6$ibm2, 3))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3765 -2.4599 -0.2134  2.3452 11.4769
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           51.6832     0.1969 262.455  < 2e-16 ***
## poly(td6$ibm2, 3)1  88.9085     3.5391  25.122  < 2e-16 ***
```

```
## poly(td6$ibm2, 3)2 -27.4059     3.5391  -7.744 1.29e-13 ***
## poly(td6$ibm2, 3)3  -8.9189     3.5391  -2.520   0.0122 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.539 on 319 degrees of freedom
## Multiple R-squared:  0.6862, Adjusted R-squared:  0.6832
## F-statistic: 232.5 on 3 and 319 DF,  p-value: < 2.2e-16
```

```
lmOrthoPolyHAEDIR <- lm(td6$logy ~ poly(td6$haednr, 3))
summary(lmOrthoPolyHAEDIR)
```

```
##
## Call:
## lm(formula = td6$logy ~ poly(td6$haednr, 3))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.502  -4.014  -0.555   4.075  17.635
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            51.6832     0.3325 155.418  < 2e-16 ***
## poly(td6$haednr, 3)1 -26.6553     5.9765  -4.460 1.14e-05 ***
## poly(td6$haednr, 3)2  -7.2617     5.9765  -1.215    0.225
## poly(td6$haednr, 3)3  23.9463     5.9765   4.007 7.67e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.977 on 319 degrees of freedom
## Multiple R-squared:  0.105,  Adjusted R-squared:  0.09658
## F-statistic: 12.47 on 3 and 319 DF,  p-value: 9.837e-08
```

```
summary(lm.sixth)
```

```
##
## Call:
## lm(formula = logy ~ . - nuvirdi - id, data = td6)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9326 -1.1715 -0.0659  1.2255  9.9321
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -3.794e+01  1.990e+01  -1.907 0.057524 .
## kdagur              3.581e-03  2.614e-04  13.697  < 2e-16 ***
## teg_eignIbudareign -1.453e+00  6.868e-01  -2.115 0.035258 *
## teg_eignParhus      3.823e+00  1.319e+00   2.898 0.004034 **
## teg_eignRadhus      7.316e-01  9.297e-01   0.787 0.431996
## byggar              1.287e-02  1.018e-02   1.264 0.207072
## haednr              3.162e-01  1.241e-01   2.547 0.011351 *
## lyfta               4.752e-01  4.167e-01   1.140 0.255105
## ibm2                9.900e-02  5.235e-03  18.912  < 2e-16 ***
## fjbkar              5.994e-01  4.445e-01   1.348 0.178539
```

```
## fjsturt            -6.244e-02  3.457e-01  -0.181 0.856784
## fjstof              3.944e-01  3.449e-01   1.144 0.253710
## fjgeym              3.194e-01  2.163e-01   1.477 0.140795
## matssvaedi70       -2.225e-02  4.893e-01  -0.045 0.963768
## matssvaedi91       -3.610e+00  6.438e-01  -5.607 4.68e-08 ***
## matssvaedi160      -6.866e+00  5.806e-01 -11.825  < 2e-16 ***
## matssvaedi281      -1.856e+00  5.500e-01  -3.375 0.000836 ***
## undirmatssvaedi3   -2.142e+00  1.060e+00  -2.021 0.044218 *
## undirmatssvaedi6    2.516e+00  1.262e+00   1.993 0.047209 *
## undirmatssvaedi21   1.665e-01  6.367e-01   0.261 0.793936
## undirmatssvaedi28  -3.244e-01  6.954e-01  -0.467 0.641161
## undirmatssvaedi40  -4.266e-01  1.753e+00  -0.243 0.807909
## undirmatssvaedi48   3.994e+00  1.316e+00   3.035 0.002620 **
## undirmatssvaedi54  -2.764e+00  1.765e+00  -1.566 0.118518
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.399 on 299 degrees of freedom
## Multiple R-squared:  0.8649, Adjusted R-squared:  0.8545
## F-statistic: 83.19 on 23 and 299 DF,  p-value: < 2.2e-16
```

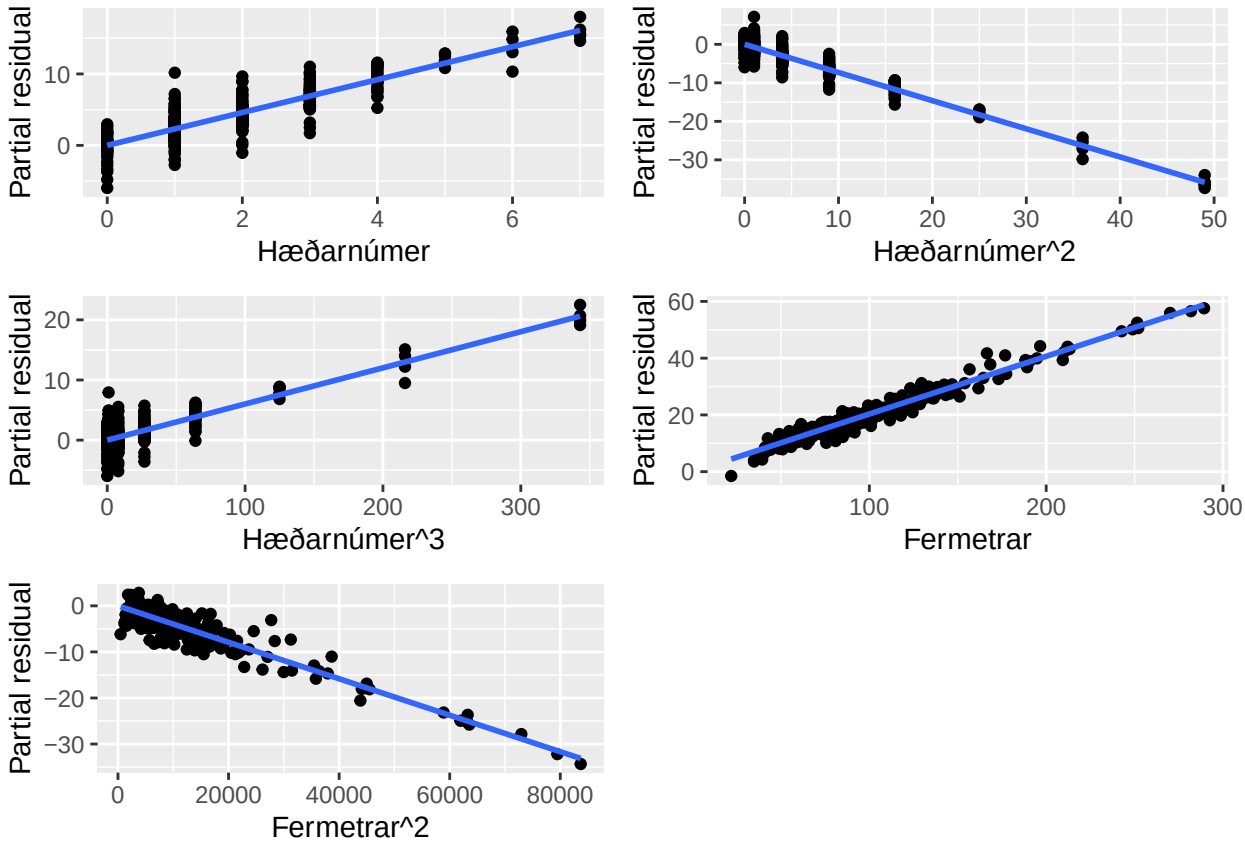
```
td7 <- subset(td6, select = -c(fjsturt))
td7$ibm22 <- td7$ibm2^2
td7$haednr2 <- td7$haednr^2
td7$haednr3 <- td7$haednr^3

lm.seventh <- lm(logy ~ . -nuvirdi -id, data = td7)
summary(lm.seventh)
```

```
##
## Call:
## lm(formula = logy ~ . - nuvirdi - id, data = td7)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9676 -1.1941 -0.0031  1.0690  7.8716
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -3.643e+01  1.659e+01  -2.196 0.028859 *
## kdagur              3.507e-03  2.164e-04  16.208  < 2e-16 ***
## teg_eignIbudareign -2.059e+00  5.603e-01  -3.674 0.000283 ***
## teg_eignParhus      2.350e+00  1.090e+00   2.156 0.031854 *
## teg_eignRadhus     -1.002e+00  7.763e-01  -1.291 0.197628
## byggar              9.844e-03  8.494e-03   1.159 0.247439
## haednr              2.301e+00  4.668e-01   4.930 1.37e-06 ***
## lyfta               1.069e+00  3.535e-01   3.023 0.002724 **
## ibm2                2.032e-01  1.091e-02  18.621  < 2e-16 ***
## fjbkar              1.738e-01  3.142e-01   0.553 0.580511
## fjstof              4.223e-01  2.851e-01   1.481 0.139616
## fjgeym              2.123e-01  1.777e-01   1.195 0.233137
## matssvaedi70       -2.852e-01  4.087e-01  -0.698 0.485900
## matssvaedi91       -4.011e+00  5.354e-01  -7.491 7.87e-13 ***
## matssvaedi160      -6.389e+00  4.779e-01 -13.368  < 2e-16 ***
## matssvaedi281      -2.389e+00  4.552e-01  -5.249 2.92e-07 ***
```

```
## undirmatssvaedi3   -1.451e+00  8.742e-01  -1.660 0.097980 .
## undirmatssvaedi6    2.626e+00  1.038e+00   2.530 0.011938 *
## undirmatssvaedi21  -4.230e-01  5.327e-01  -0.794 0.427732
## undirmatssvaedi28   6.313e-01  5.707e-01   1.106 0.269490
## undirmatssvaedi40  -1.378e+00  1.442e+00  -0.955 0.340152
## undirmatssvaedi48   2.730e+00  1.084e+00   2.517 0.012359 *
## undirmatssvaedi54  -1.770e+00  1.452e+00  -1.219 0.223652
## ibm22              -3.958e-04  3.792e-05 -10.438  < 2e-16 ***
## haednr2            -7.315e-01  1.850e-01  -3.954 9.62e-05 ***
## haednr3             6.009e-02  1.926e-02   3.121 0.001983 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.969 on 297 degrees of freedom
## Multiple R-squared:  0.9095, Adjusted R-squared:  0.9019
## F-statistic: 119.4 on 25 and 297 DF,  p-value: < 2.2e-16
```

```
fortNG = fortify(lm.seventh)
n7 <- nrow(fortNG)
p7 <- nrow(summary(lm.seventh)$coefficients)

calling = c(7,25,26,9,24)
betas7 <- c("Hæðarnúmer","Hæðarnúmer^2", "Hæðarnúmer^3", "Fermetrar", "Fermetrar^2")

beta71 <- summary(lm.seventh)$coefficients[7, 1]
beta72 <- summary(lm.seventh)$coefficients[25, 1]
beta73 <- summary(lm.seventh)$coefficients[26, 1]
beta74 <- summary(lm.seventh)$coefficients[9, 1]
beta75 <- summary(lm.seventh)$coefficients[24, 1]
beta7 <- c(beta71, beta72, beta73, beta74, beta75)
plots7 <- list()
for(i in 1:length(beta)) {
  regressor <-  model.matrix(lm.seventh)[, (calling[i])]
  partialRes <- fortNG$.resid + regressor * beta7[i]
  tibble(x = regressor,
         y = partialRes) %>%
    ggplot(aes(x = x, y = y)) +
    geom_point() +
    stat_smooth(method = 'lm', se = F) +
    labs(x = betas7[i],
         y = 'Partial residual') -> plots7[[i]]
}

cowplot::plot_grid(plots7[[1]], plots7[[2]], plots7[[3]], plots7[[4]], plots7[[5]],
                   nrow = 3, ncol = 2)
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```




```
rev_resid_7 <- bxcx(residuals(lm.seventh), lambda, InverseQ = TRUE, type = "BoxCox")
```

```
## Warning: modified inverse Box-Cox transformation used
```

```
sqrt(mean(rev_resid_7^2))
```

```
## [1] 5.771263
```

```
test_data$logy <- bxcx(test_data$nuvirdi, lambda, InverseQ = FALSE, type = "BoxCox")
test_data$ibm22 <- test_data$ibm2^2
test_data$haednr2 <- test_data$haednr^2
test_data$haednr3 <- test_data$haednr^3

test_resid = (predict(lm.seventh, test_data) - test_data$logy)

rev_resid_test <- bxcx(test_resid, lambda, InverseQ = TRUE, type = "BoxCox")
```

```
## Warning: modified inverse Box-Cox transformation used
```

```
sqrt(mean(rev_resid_test^2))
```

```
## [1] 4.642696
```

Við skoðum hvaða áhrif þetta hefur haft á líkanið okkar. RMSE á þjálfunargagnasettinu okkar mælist núna 5.7712632 sem er rúmur helmingur þess sem mældist áður. Nú er eðlilegt að spyrja sig hvort við séum búnir að sníða þetta líkan nákvæmlega að gagnasettinu okkar og þar með tapa almennu eðli þess sem spálíkani fyrir önnur gagnasett. Vegna þessa skiptum við gagnasetti okkar í tvennt, þar sem fjórðungur var settur til hliðar og hefur aldrei verið meðhöndlaður. Við getum gengið úr skugga um að gildi líkansins haldi með því að prófa að spá fyrir um virði fasteignanna í þessum fjórðungi út frá líkaninu okkar. Við þurfum að uppfæra efra stigs breytur um hæðir og fermetra og síðan umbreyta núvirðisbreytunni okkar með BoxCox fallinu. Við skoðum

leifðina sem spáin skilur eftir og umbreytum henni til baka með andhverfu BoxCox fallsins. Niðurstaðan gefur RMSE 4.642696, sem er jafnvel lægra en á þjálfunargagnasettinu okkar. Á þessum tímapunkti runnu allar þær grímur sem eftir voru á okkur. Við vorum vissulega búnir að meðhöndla margar breytur vel og fara í gegnum sniðug greiningarferli en okkur þótti þessi villa vera mun lægri en við höfðum búist við. Að því sögðu, þá fundum við ekkert sem benti til þess að við værum óafvitandi með einhver brögð í tafli, svo við héldum áfram að þróa líkanið.

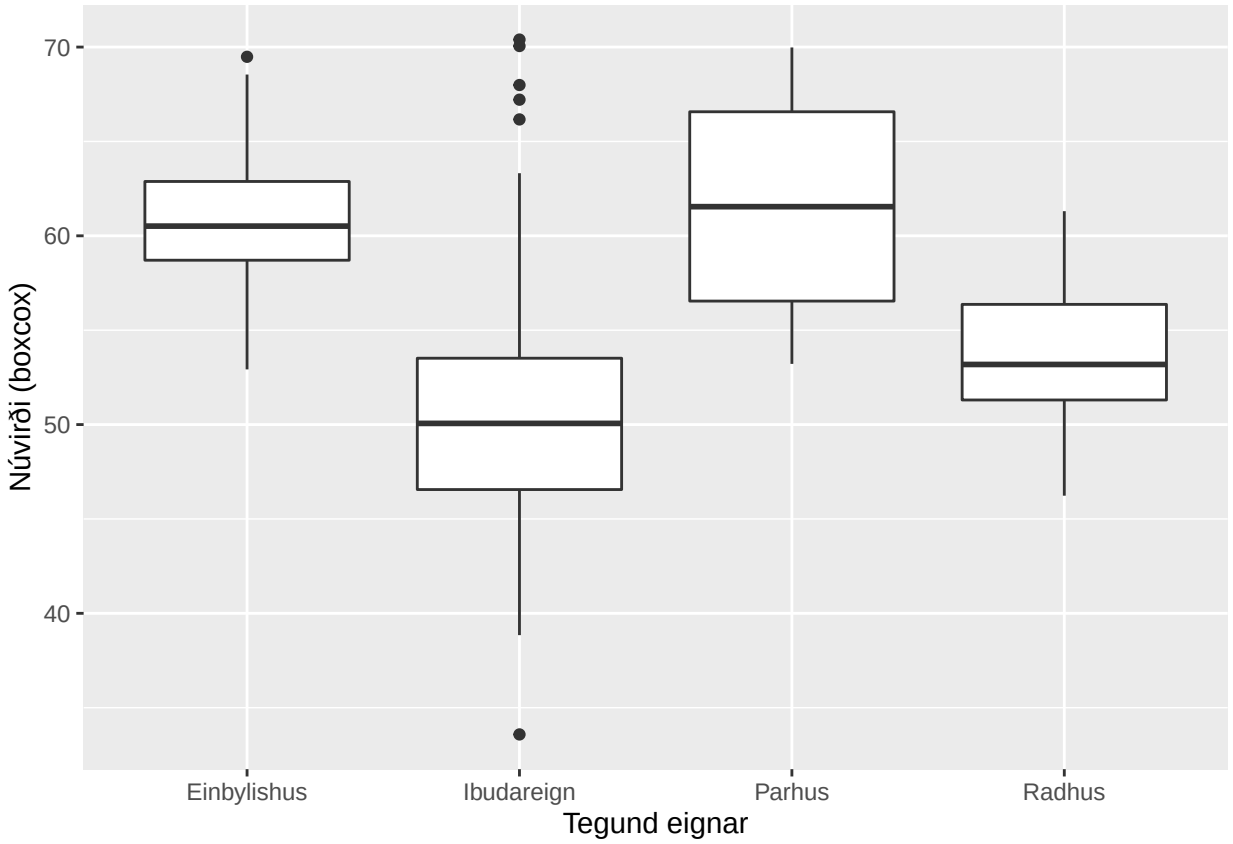
Flokkunarfræðilegar breytur

Ennþá átti eftir að skoða þær breytur líkansins sem ekki voru tölulegar. Við höfðum þegar losað okkur við nokkrar flokkunarfræðilegar breytur en þær sem eftir sitja eru óskoðaðar. Við viljum kanna möguleika á því að fækka þeim eða sleppa með öllu. Við skulum taka hverja breytu fyrir í stuttu máli.

Tegund eignar

Fjórar breytur eru innan flokksins um tegund eignar; parhús, raðhús, einbýlishús og íbúðareign. Fjöldanum er nokkuð misdreift milli flokka en yfir 85% eigna eru skráðar sem íbúðareign en um 1% parhús. Svona ójöfnuður milli flokka er óhentugur, þar sem fá hús í flokki parhúsa fá mikið meira vægi en húsin innan íbúðareigna. Því er æskilegt að kanna hvort hægt sé að sameina flokkana í stærri söfn, ef dreifing gagnanna gefur ástæðu til þess.

```
ggplot(td7, aes(x=teg_eign, y=logy)) + geom_boxplot() + xlab("Tegund eignar") + ylab("Núvirði (boxcox)")
```




```
fit.teg_eign_check <- aov(logy ~ teg_eign, data = td7)
tukeyteg_eign_check <- TukeyHSD(fit.teg_eign_check)

# data.frame(tukeyteg_eign_check[1:1]) %>% kbl(align ='c') %>%
#  kable_styling(latex_options = "HOLD_position")
```

Við skoðuðum þessa flokkunarfræðilegu þætti út frá ANCOVA greiningu og einkum heiðarleikaprófi Tukeys, þar sem líkindi milli flokkanna eru borin saman innbyrðis og p-gildis mati skilað út frá margföldum samanburði. Niðurstöður úr slíku prófi reyndust eftirfarandi:

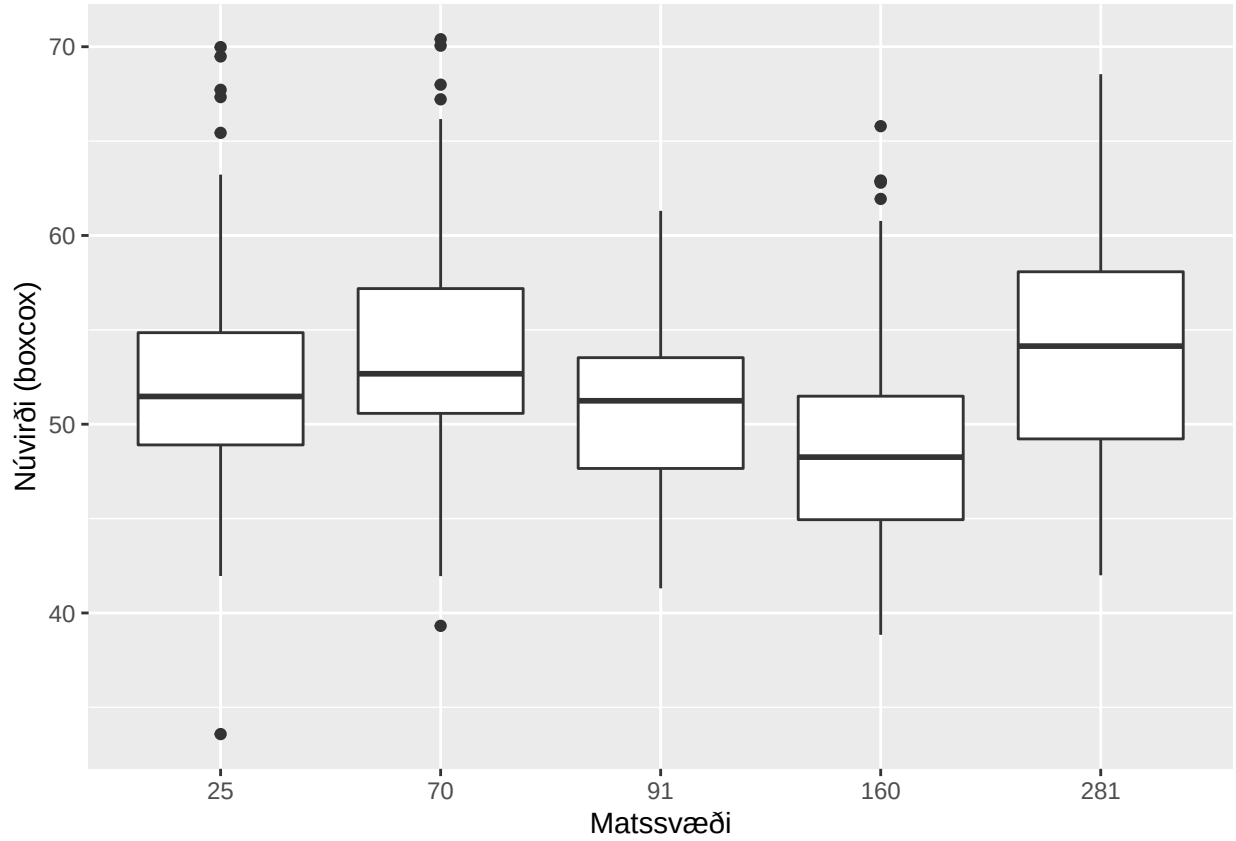
p-gildi samanburðar	Íbúðareign	Einbýlishús	Parhús	Raðhús
Íbúðareign	-	0.0000000	0.0002607	0.2158763
Einbýlishús	-	-	0.9943943	0.0016362
Parhús	-	-	-	0.0665462

Gildi töflunnar gefa til kynna hversu sterkur greinarmunur sé á breytunum með tilliti til núvirðis fasteignar. Gildi samanburðar einbýlishúsa og parhúsa sker sig frá hinum og er næstum því 1, sem gefur til kynna að munurinn sé verulega lítill. Kassaritið að ofan sýnir að þetta ætti að standast og því má vel vera að betra módel sameini þessa tvo flokka í einn, til aðgreiningar frá hinum tveimur.

Matssvæði

Næsti flokkur er matsvæði eignanna, sem eru 5 talsins. Frekar er hægt að lesa sér til um sérhvert matssvæði í upphafi skýrslunnar en kassaritið að neðan sýnir að þau eru ekki sérlega afgerandi sín á milli. Tölfræðileg greining mun gefa okkur skýrari mynd af því hvort og hvernig hægt sé að fækka breytum í þessum flokki.

```
ggplot(td7, aes(x=matssvaedi, y=logy)) + geom_boxplot() + xlab("Matssvæði") + ylab("Núvirði (boxcox)")
```




```
fit.matssv <- aov(logy ~ matssvaedi, data = td7)
tukeymatssv <- TukeyHSD(fit.matssv)

#data.frame(tukeymatssv[1:1]) %>% kbl(align ='c') %>%
#  kable_styling(latex_options = "HOLD_position")
```

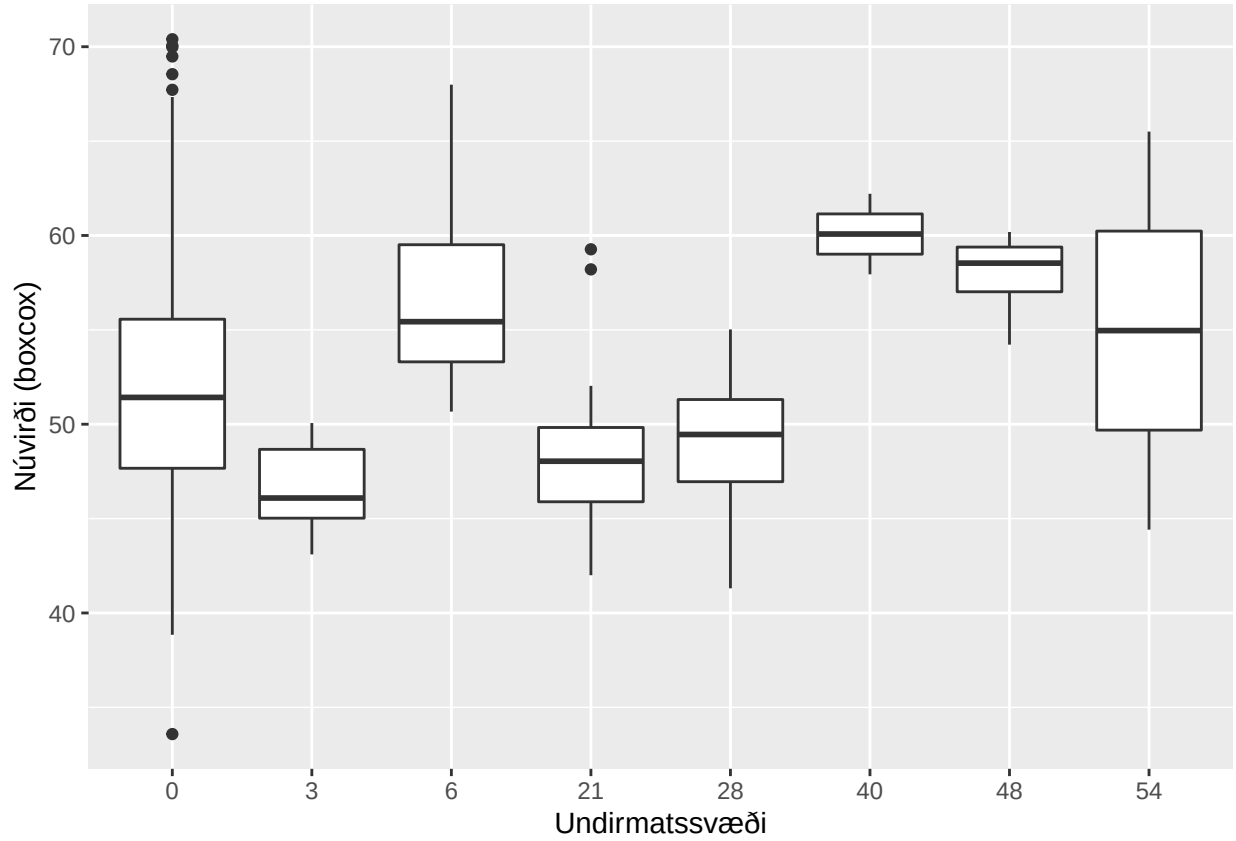
p-gildi samanburðar	Miðbær (25)	Melar að sjó (70)	Háaleiti/Skeifa (91)	Hólar, Berg (160)	Réttarholt (281)
Miðbær (25)	-	0.8103585	0.4713571	0.0075621	0.9017006
Melar að sjó (70)	-	-	0.0449796	0.0000290	0.9999435
Háaleiti/Skeifa (91)	-	-	-	0.5698535	0.1046991
Hólar, Berg (160)	-	-	-	-	0.0003792

Við sjáum að þessi samanburður er ekki eins skýr og sá fyrri en helst er að sjá sterk líkindi milli hópanna 70 og 281, þó rök séu fyrir því að sameina hóp 25 undir sama hatt. Það er samt varhugavert að ætla líkaninu að meta fasteignir í miðbænum á sama máta og í Réttarholti og í Vesturbæ við Ægisíðu.

Undirmatssvæði

Síðasta flokkunarfræðilega breytan tekur til sérstakra undirsvæða innan hvers matsvæðis, þar sem ástæða er talin til að gera sérstaklega greinarmun á íbúðum á undirmatssvæðinu og annarra nærliggjandi fasteigna. Svæðin eru alls 8 og samspil þeirra því ansi flókið.

```
ggplot(td7, aes(x=undirmatssvaedi, y=logy)) + geom_boxplot() + xlab("Undirmatssvæði") + ylab("Núvirði (
```




```
fit.undirmatssvaedi_check <- aov(logy ~ undirmatssvaedi, data = td7)
tukeyundirmatssvaedi_check <- TukeyHSD(fit.undirmatssvaedi_check)

#data.frame(tukeyundirmatssvaedi_check[1:1]) %>% kbl(align ='c') %>%
#  kable_styling(latex_options = "HOLD_position")
```

p-gildi	0	3	6	21	28	40	48	54
0	-	0.3696893	0.6636552	0.1706785	0.1706347	0.5822655	0.5555788	0.9976527
3	-	-	0.1128441	0.9986891	0.9921720	0.1223266	0.0822330	0.6976071
6	-	-	-	0.1262593	0.1550891	0.9996017	1.0000000	0.9998049
21	-	-	-	-	0.9999945	0.1619849	0.0876828	0.8262787
28	-	-	-	-	-	0.1925245	0.1085773	0.8700096
40	-	-	-	-	-	-	0.9998948	0.9906150
48	-	-	-	-	-	-	-	0.9993428

Við sjáum að mörg sterk samspil myndast hér, einkum og sér í lagi milli undirmatssvæða 6 (Ægisíðu, Starhaga og hluta Lynghaga) og 48 (Safamýri). Við ákváðum að þessi breyting væri of afgerandi, einkum þar sem áhrif slíkra þátta geta tekið örum breytingum milli ára.