Kvantitatiivne andmeanalüüs

Marko Sõmer

2022-02-18

Contents

Chapter 1

Sissejuhatus

Siia on koondatud õppematerjalid kursusele “Kvantitatiivne andmeanalüüs II”.

Chapter 2

Lineaarne regressioon

2.1 Lihtne lineaarne regressioon

Lihtne lineaarne regressioon (*simple linear regression*) on statistiline meetod mis võimaldab hinnata ja kvantifitseerida kahe arvtunnuse vahelist suhet. Regressioonsuhte puhul eeldatakse, et üks tunnustest oleks nn sõltuv tunnus ja teine sõltumatu[1], kus sõltuva tunnuse väärtus on mõjutatud (sõltub) sõltumatu tunnuse väärtusest. Kui sõltumatuid tunnuseid on rohkem kui üks, on tegemist mitmese regressiooniga (sellest hiljem), ühe sõltumatu tunnuse korral nn "lihtsa" regressiooniga (*simple linear regression*). Keskendume esialgu "lihtsale" variandile.

Kasutame näitena Piaaci andmestikku. Tõmbame andmestiku sisse ja uurime graafiliselt sissetuleku (*sissetulek*) ning matemaatilise kirjaokuse (*numeracy*) vahelist seost.

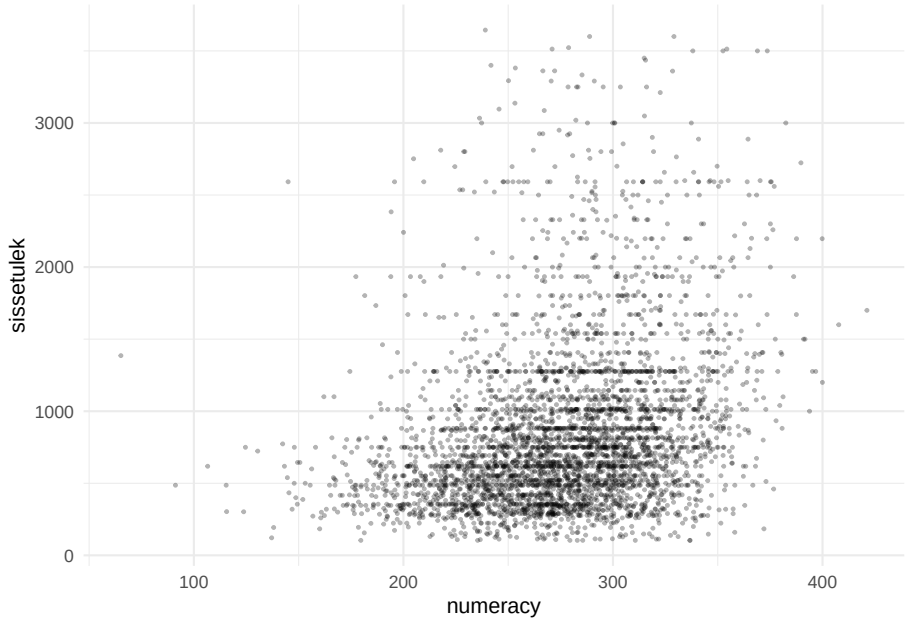
```
# Loeme kõigepealt sisse vajalikud paketid
library(dplyr)
library(ggplot2)
library(readr)

# Tõmbame  sisse andmestiku
piaac <- read_csv("data/piaac.csv")

piaac %>%
  ggplot(aes(x = numeracy, y = sissetulek))+
```

[1] Inglisekeelses terminoloogias kasutatakse sõltuva tunnuse puhul peale *dependent variable* ka nimetusi *response* või *outcome variable* ja sõltumatu tunnuse puhul peale *independent variable* ka *predictor* või *explanatory variable*. Prediktor on ka eesti keeles kasutusel.

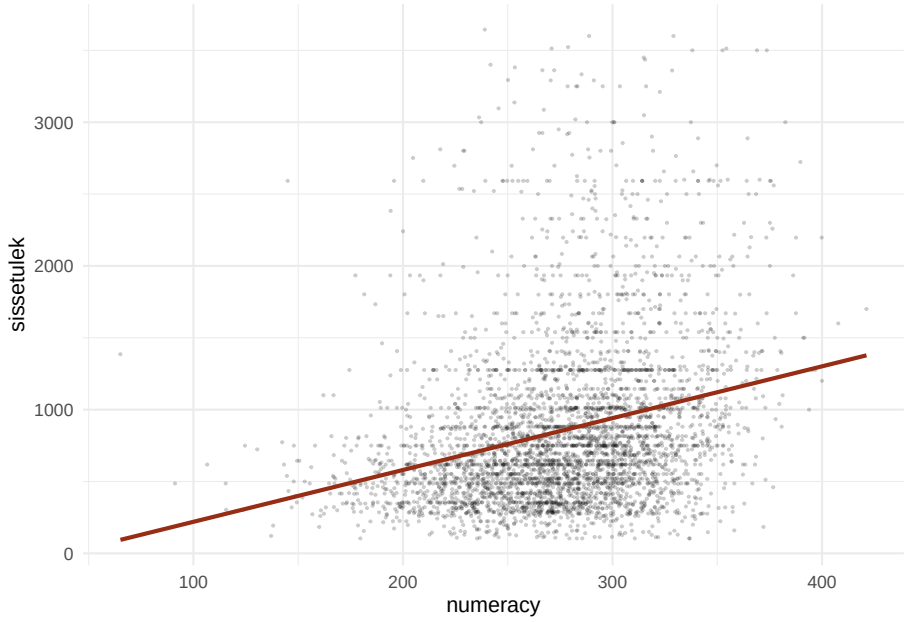
```
  geom_point(size = 0.5, alpha = 0.3)+
  theme_minimal()
```



Tundub, et nende kahe tunnuse vahel on seos olemas. Mida kõrgem on matemaatilise kirjaoskuse skoor, seda kõrgem on sissetulek. Me saame selle suhte kokku võtta regressioonisirge abil. ggplotis on olemas vastav funktsioon `geom_smooth()`, mis selle joone meile graafikule paneb. Kuna me tahame saada lineaarse regressiooni sirget, siis peame `geom_smooth`'is kasutama argumenti `method = "lm"`[2]

```
piaac %>%
  ggplot(aes(x = numeracy, y = sissetulek))+
  geom_point(size = 0.3, alpha = 0.2)+
  geom_smooth(method = "lm", se = F, color = "#972D15")+
  theme_minimal()
```

[2] *Defaultis* annab `geom_smooth` meile mittelineaarse regressioonijoone (vastavalt sellele palju vaatlusi on, kas *gam* või *loess*), mis üritab tunnustevahelist suhet andmete kõikides punktides võimalikult täpselt kirjeldada.



Regressioonisirge on väljendatav tavalise joone võrrandiga:

$$Y = a + bX \tag{2.1}$$

kus a on vabaliige (*intercept*) ja b on sirge tõus (*slope*). Regressiooni kontekstis kutsutakse seda sirge tõusu regressioonikoefitsiendiks või regressioonikordajaks. Vabaliige tähistab Y väärtust juhul kui X on 0 (sirge lõikumine y-teljega) ja sirge tõus ühikulist muutust Y väärtuses kui X väärtus muutub ühe ühiku võrra. Eelneva näite puhul oleks vabaliige võrdne sissetulekuga (Y) juhul kui matemaatilise kirjaoskuse tase (X) oleks 0 ja sirge tõus võrdne keskmise sissetuleku muutusega, mis lisandub iga matemaatilise kirjaoskuse punktiga. Kui sirge tõus on positiivne, siis X'i väärtuse kasvades Y väärtus suureneb, kui negatiivne, siis kahaneb. Kui sirge tõus on aga 0, siis seos kahe tunnuse vahel puudub (iga X väärtuse korral on keskmine Y sama).

Linaarse regressioonanalüüsi eesmärgiks ongi leida parim võimalik sirge (st leida vabaliige ja regressioonikoefitsient, mis seda sirget määratlevad) tunnustevahelise lineaarse suhte kirjeldamiseks. Parim võimalik tähendab siinjuures seda, et see sirge läheb punktiparvest läbi võimalikult keskelt, st kirjeldab kõiki punkte võimalikult hästi.

2.2 Regressioon R-is

R-is käib lihtsa regressioonimudeli tegemine `lm()` (*linear model*) funktsiooniga. Loomulikult on ka teisi funktsioone, mis regressiooni jooksutamisega hakkama saavad ja hea tahtmise korral võib vastava funktsiooni ka mõningase vaevaga ise valmis kirjutada. Kuid jätame teised variandid hetkel kõrvale.

`lm()` funktsioonis tuleb defineerida regressioonivõrrand. Selleks peame määratlema sõltuva tunnuse, seejärel kasutama tildet (`~`) ning seejärel määratlema sõltumatu(d) tunnuse(d): `sõltuv_tunnus ~ sõltumatu_tunnus`[3]. Võtame eelpool toodud näite sissetuleku ja matemaatilise kirjaoskuse seosest ning defineerime regressioonimudeli, millega hindame matemaatilise kirjaoskuse mõju sissetulekule[4]:

```
lm(sissetulek ~ numeracy, data = piaac)
```

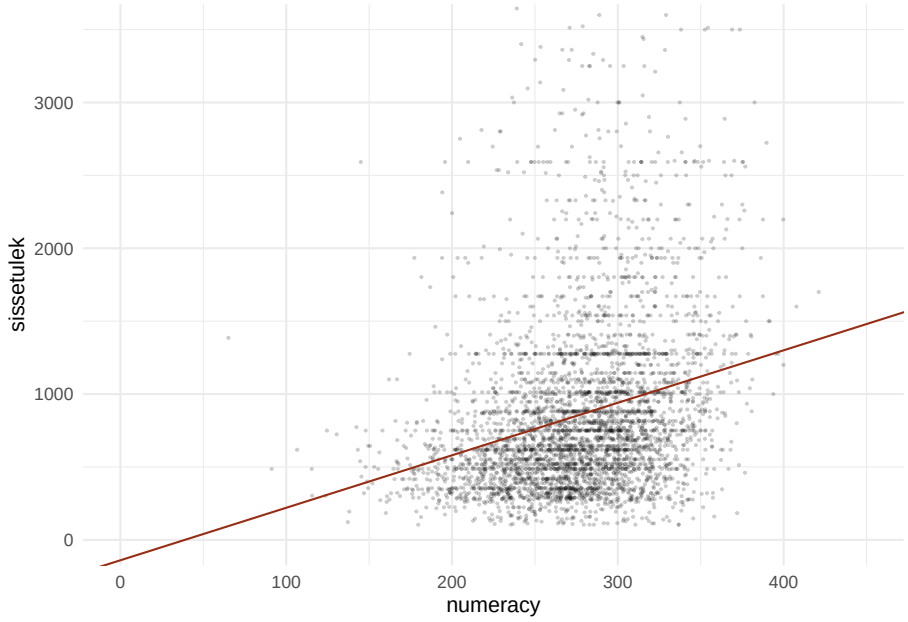
```
##
## Call:
## lm(formula = sissetulek ~ numeracy, data = piaac)
##
## Coefficients:
## (Intercept)     numeracy
##    -140.887        3.606
```

Lihtsalt `lm()` funktsiooni jookustades saame kaks numbrit - vabaliikme (*intercept*), mis antud näite puhul on -140, ja regressioonikoefitsiendi (*regression coefficient*), mis antud näite puhul on 3.6. Mida need meile ütlevad? Nagu eelnevalt juttu oli, siis vabaliige on Y väärtus kui X on 0, ehk siis inimesel, kelle matemaatilise kirjaoskuse skoor on 0, peaks meie mudeli kohaselt sissetulek olema -140. Regressioonikoefitsient aga annab meile teada kui palju Y muutub, kui X muutub ühe ühiku võrra, ehk siis kui matemaatilise kirjaoskuse skoor tõuseb ühe punkti võrra, tõuseb sissetulek keskmiselt 3.6 euro võrra. Nüüd, kui teame mudeli parameetreid, saame nende abil regressioonijoone graafikule kanda ka ilma `geom_smooth`'ita:

```
piaac %>%
  ggplot(aes(x = numeracy, y = sissetulek))+
  geom_point(size = 0.3, alpha = 0.2)+
  geom_abline(slope = 3.6, intercept = -140, color = "#972D15")+
  coord_cartesian(xlim = c(0,450), ylim = c(0,3500))+
  theme_minimal()
```

[3] Hiljem, kui meil on mitu sõltumatut tunnust, eristame tunnused plussiga: `sõltuv_tunnus ~ sõltumatu_tunnus_1 + sõltumatu_tunnus_2 + ...`

[4] Tegelikult ei ole selline mudel korrektne. Sissetuleku jaotus ei vasta hästi regressiooni nõuetele. Miks ei vasta ja kuidas see vastama panna, sellest natuke hiljem. Kuid hetkel kasutame seda puhtalt didaktilistest kaalutustest lähtuvalt.



Kui me teame regressioonisirge tõusu ehk regressioonikoefitsienti ja vabaliiget, siis lähtuvalt sõltumatu tunnuse väärtustest saame prognoosida sõltuva tunnuse väärtuse:

$$\hat{y}_i = b_0 + b_1 x_i \tag{2.2}$$

$\hat{y}_i$ antud võrrandis tähistab hinnatud või prognoositud y väärtust (sellest ka see müts y peal) vaatlusele i. Kui meil on regressioonivõrrand $\hat{y}_i = -140 + 3.6x_i$ ja meil on mingi vaatlus i, kelle x väärtus on näiteks 200, siis saame sellele vaatlusele prognoosida y väärtuseks $-140 + 3.6 \times 200 = 580$. Ehk siis inimesel, kelle matemaatilise kirjaoskuse skoor on 200, peaks meie mudeli järgi sissetulek olema *ca* 580 eurot. Inimesel, kelle matemaatilise kirjaoskuse skoor on 400, peaks sissetulek olema keskmiselt $-140 + 3.6 \times 400 = 1300$ eurot

2.3 Regressiooni jäägid

Samas on muidugi võimatu ühe sirgega kõiki punkte ideaalselt kirjeldada. Iga punkti ja sirge vahele jääb alati mingi viga või teisisõnu, kõik punktid (või vähemalt enamus neist) hälbivad suuremal või vähemal määral regressioonisirgest.

Mida suuremad need hälbed on, seda vähem suudab on meie mudel (regressioonisirge) kirjeldada sõltuva tunnuse variatsiooni ja seda suurem on vea määr meie mudelis. Neid hälbeid kutusutakse **regressiooni jääkideks** (*regression residuals*).

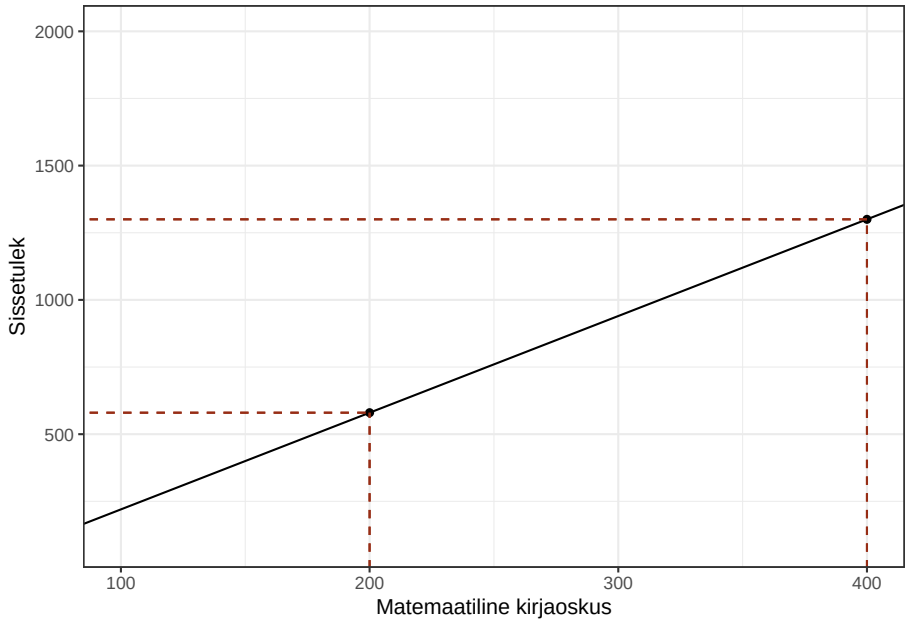


Figure 2.1: Prognoosimine y väärtust kui x on 200

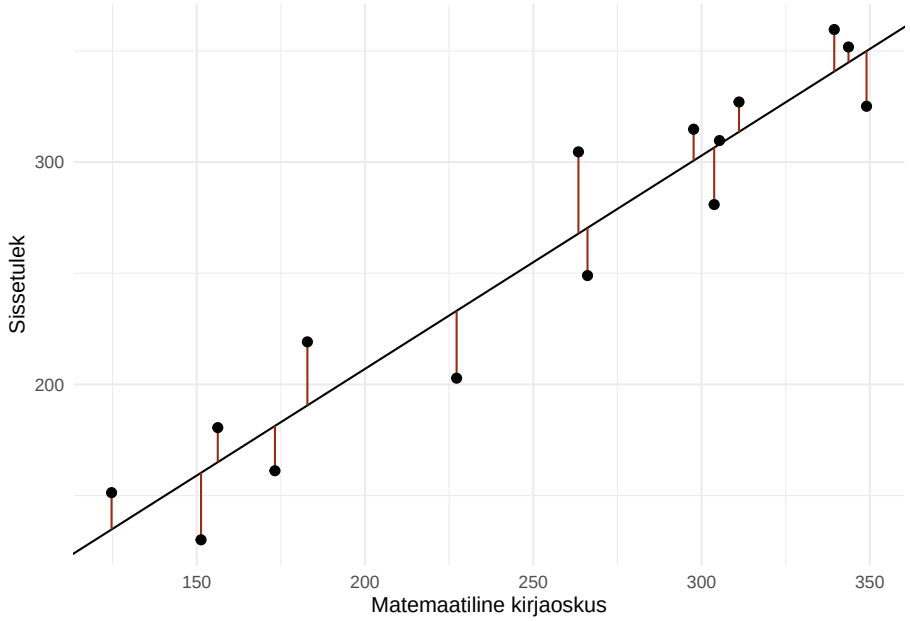


Figure 2.2: Regressiooni jäägid

Ehk siis iga kord, kui prognoosime $\hat{y}_i = \beta_0 + \beta_1 x_i$ abil y_i väärtust, teeme me mingi vea[5]. Seetõttu tuleb regressioonivõrrandile lisada vea komponent (ϵ) ning võrrand ise muutub vastavalt:

$$\hat{y_i} = \beta_0 + \beta x_i + \epsilon \tag{2.3}$$

Kõige parem regressioonisirge annab joon, mille puhul jäägid on minimaalsed, ehk siis joon, mille puhul kõikide vaatluste jääkide summa oleks võimalikult väike. Kuna me ei saa jääke kokku võtta neid lihtsa kokku liites (*ca* pooled jäägid on väiksemad kui regressioonijoon ja *ca* pooled suuremad, seega nende summa oleks 0), siis tuleb nad enne liitmist ruutu panna. Ja meie eesmärgiks on nüüd leida regressioonisirge, mis minimeeriks **ruutjääkide summa** (*residual sum of squares* ehk RSS) ehk siis regressioonisirge, mille puhul RSS oleks võimalikult väike[6].

Eelnevast lähtuvalt on ka küllaltki loogiline, et meetodit, millega RSS minimeeritakse ja regressioonisirge ning vastavad koefitsiendid leitakse, nimetatakse **vähimruutude meetodiks**.

- Kasutades ggplot'i ja tehke punktdiagramm `geom_point()` matemaatilise kirjaoskuse (*numeracy*) ja funktsionaalse lugemisoskuse (*literacy*) vahelisest seosest. Pange *numeracy* x-teljele ja *literacy* y-teljele.

- Kasutades `geom_abline()`'i, lisage joonisele lineaarne regressioonijoon (seega peate eelnevalt `lm()` funktsiooniga leidma regressioonijoone vabaliikme ja regressioonikoefitsiendi)

2.4 Regressioonimudeli sobitumine

Olles leidnud joone, mis kirjeldab kahe tunnuse vahelist seost kõige paremini, võiks ju eeldada, et ülesanne on täidetud. Aga kas ikka on? Ükskõik, millisest punktiparvest võib regressioonijoone läbi panna. Kuid tulenevalt regressioonijääkide (vaatluste hälbed regressioonijoonest) suurusest saame selle joone kohta teha väga erinevaid järeldusi. Kui jäägid on väikesed, siis võime suhteliselt täpselt prognoosida sõltuva tunnuse väärtust või teha järeldusi seose kohta. Kuid mida suuremad on jäägid, seda ebatäpsem on ka meie prognoos/järeldus.

Üldjuhul kasutame regresioonanalüüsi, et teha valimi baasil järeldusi mingi üldkogumi kohta. Meid huvitab, kas see seos, mida näeme oma valimi andmete põhjal, kehtib ka üldkogumis. Saame küll eeldada, et valimipõhiselt leitud regressioonisirge on suhteliselt sarnane üldkogumi sirgele (sirge, mille me saaksime,

[5] Mida saab väljendada kui $\epsilon_i = y_i - \hat{y}_i$

[6] $RSS == e_1^2 + e_2^2 + ... + e_n = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

kui kaasaksime analüüsi kõik üldkogumi liikmed), aga kui sarnane, seda me ei tea. Kui me võtaksime samast üldkogumist teise valimi, siis juhul, kui mõlemad valimid on võetud korrektselt[7] ja valimid on piisavalt suured, siis peaksid nende põhjal leitud regressioonisirged olema suhteliselt sarnased, aga identsed ei ole nad praktiliselt kunagi. Kõikide võimalike valimite puhul me mingil määral alahindame või ülehindame tegelikku, populatsiooni regressioonikoefitsienti (ja ka vabaliiget). Seega, et saada aimu valimipõhise hinnangu täpsusest (vastavusest tegelikule tegelikule üldkogumi parameetrile), peaksime kuidagi välja selgitama valimi kasutamisest tuleneva vea võimaliku suuruse.

Et hinnata mudeli sobivust andmetega ja sellega leitud hinnagute täpsust, vajame mudeli kohta täiendavat infot. Eelnevalt regressioonimudelit `lm()` funktsiooniga jooksutades oli väljund väga lakooniline. Saime teada ainult vabaliikme ja regressioonikoefitsinedi väärtused. Tegelikult on `lm()` tulem muidugi märksa põhjalikum. Muule mudeliga kaasnevale infole saame ligi kui salvestame mudeli esmalt mingisse andmeobjekti ja kasutame selle andmeobjekti peal `summary()` käsku[8].

```
mudel1 <- lm(sissetulek ~ numeracy, data = piaac)
summary(mudel1)
```

```
##
## Call:
## lm(formula = sissetulek ~ numeracy, data = piaac)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1016.7  -351.5  -129.1   179.4  2923.4
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -140.887     56.510  -2.493   0.0127 *
## numeracy       3.606      0.202  17.849   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 555.8 on 3982 degrees of freedom
```

[7] Korrektse valimi võtmise all peame siinkohal silmas eelkõige juhuvalikut. Kõikidel populatsiooni liikmetel/elementidel peab olema võrdne võimalus valimisse sattuda. Kui üldpopulatsiooniks on Eesti elanikkond, aga valimisse võtaksime ainult Tallinna elanikud, siis antud valimi põhjal tehtavad järeldused ei oleks kuidagi üldistatavad kõigile Eesti elanikele, vaid ikkagi ainult tallinnlastele. Lisaks juhuvalimile on veel terve rida spetsiifilisemaid valimidisaine (stratifitseeritud valim, klastervalim jne) mida me hetkel ei käsitle. Kuid tuleb meeles pidada, et keerulisemate valimidisainide puhul tuleb hilisemas analüüsis ja järelduste tegemise käigus valimi moodustamise loogikat arvesse võtta.

[8] Ka `summary()` ei anna välja kogu mudeliobjektis sisalduvat infot. Et näha mida mudeliobjekt veel sisaldab, võib kasutada `str(mudeliobjekt)` käsku.

```
##   (3648 observations deleted due to missingness)
## Multiple R-squared:  0.07408,    Adjusted R-squared:  0.07385
## F-statistic: 318.6 on 1 and 3982 DF,  p-value: < 2.2e-16
```

```
# Kui me ei taha mudelit salvestada, siis saab ka nii:
summary(lm(numeracy ~ literacy, data = piaac))
```

Nüüd näeme juba märksa põhjalikumat väljundit. Vaatame mis seal kirjas on ja kuidas seda tõlgendada. Käime väljundi sektsioonide kaupa läbi (v.a. esimene rida, mis on vist niigi suht selge)

2.4.1 Jääkide jaotus

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1016.7  -351.5  -129.1   179.4  2923.4
```

Väljundis on kirjeldatud regressiooni jääkide (*residuals*) jaotus. Enne nägime, et regressiooni jäägid on regressioonijoone ja tegelike, vaadeldud väärtuste vahe. Mida väiksemad on jäägid, seda täpsemini kirjeldab regressioonijoon andmete vahelist seost. Nägime ka, et pooled jäägid peaksid ideaalis olema suuremad (positiivse märgiga) kui regressioonisirge ja pooled väiksemad (negatiivse märgiga). Seega peaks jääkide keskmine olema ligikaudu 0 ning jääkide jaotus normaaljaotuse sarnane, kus esimene ja kolmas kvartiil, aga ka maksimum ja miinimum, on keskväärtusest umbes sama kaugel. Hiljem vaatame jääkide jaotust ka graafiliselt, mis on märksa mõistlikum viis neid uurida, kuid esmase mulje saab ka siit kätte.

2.4.2 Regressioonikoefitsiendid ja nende olulisus

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -140.887     56.510  -2.493   0.0127 *
## numeracy       3.606      0.202  17.849   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Koefitsientide sektsioonis on esitatud mudeli oluliseim info. ***Estimate*** on hinnang mudeliga leitud regressioonikoefitsientidele. Lihtsa regressiooni puhul on meil ainult vabaliige ja ühe sõltumatu tunnuse koefitsient. Hiljem, mitmese regressiooni kontekstis, on neid koefitsiente rohkem. Vabaliikmeid on aga mudeli kohta alati üks.

Tulbas ***Std. Error*** on toodud koefitsientide standardvead. Standardviga kirjeldab meie mudeli hinnangus sisalduvat määramatust. Me kasutame regressioonikoefitsientide leidmiseks üldjuhul valimipõhiseid andmeid, kuigi tegelikult huvitavad meid ju üldkogums esinevad seosed. Valimipõhine hinnang peaks piisavalt suure valimi korral olema tõenäoliselt küllaltki sarnane üldkogumi vastavale parameetrile, kuid väikese valimi korral puhta juhuse läbi sellest arvestatavalt erineda. Standardviga näitabki kui kindlad me oma mudeli hinnangus olla saame. Mida väiksem on standardviga (võrreldes hinnangu endaga), seda kindlamad võime olla ka oma hinnangus. Standardvea suurs sõltub eelkõige jääkide hajuvusest ja valimi suurusest. Mida väiksemad on jäägid ja mida suurem on valim, seda väiksem on ka standardviga.

Standardvea abil saame t-testi abil testida, kas regressioonikoefitsient erineb oluliselt nullist (kui koefitsient on null, siis seos tunnuste vahel puudub). t-testi tulemust näitab veerg ***t value***. t-väärtus ütleb meile kui mitme standardvea kaugusel meie regressioonikoefitsient 0-st on. Kui on piisavalt kaugel, siis saame järeldada, et leitud koefitsient on ka üldkogumis 0-st erinev. Kui kaugel on aga piisavalt kaugel? See sõltub sellest, kui suurt vea tõenäosust me oleme valmis tolereerima (mingi vea tõenäosus jääb seejuures alati). Üldjuhul valitakse selleks tõenäosuseks 5% (ütleme, et regressioonikoefitsient on statistiliselt oluline usaldusnivool 95% või olulisusnivool $p < 0.05$), aga see võib olla ka 1% või 10%. Siin tegelikult ei ole mingit väga konkreetset piirmäära, millest juhinduda. Kui me aga lepime kokku, et võimaliku vea tõenäosusena aktsepteerime 5-te protsenti, siis peab t-väärtus olema suurem kui *ca* ± 2 (täpne väärtus sõltub vaatluste arvust). Antud näite puhul on t-väärtused -2.5 ja 17.8, ehk siis mõnevõrra suuremad kui ± 2 ja me võime järeldada, et nii vabaliige kui regressioonikoefitsient erinevad olulisusnivool 95% oluliselt nullist (kuigi jah, vabaliige on suhteliselt piiri peal).

Õnneks ei pea me seda täpset t-väärtuse piirmäära ise välja nuputama. R arvutab meile automaatselt võimaliku vea tõenäosuse konkreetse t-väärtuse kohta. See tõenäosus on ära toodud veerus ***Pr(>\|t\|)*** ja seda nimetatakse p-väärtuseks. p-väärtuse tõlgendus on: kui tõenäoline on, et me saaksime niivõrd suure või suurema t-väärtuse nagu me saime, kui regressioonikoefitsient oleks üldkogumis tegelikult 0. Seega kui p-väärtus on näiteks 0.04, siis oleks tõenäosus, et me saaksime sellise regressioonikoefitsiendi, juhul kui üldkogumis oleks regressioonikoefitsient tegelikult 0 (ehk tunnuste vahe seost ei oleks), 0.04 ehk 4% või väiksem. Üldjuhul tahaksime näha p-väärtust, mis on väiksem kui 0.05. Sellisel juhul oleks koefitsient statistiliselt oluline usaldusnivool 95%. Antud näites on meil regressioonikoefitsiendi puhul tegemist väga väikeste p väärtustega (<2e-16 tähendab väiksem kui 2×10^{-16}) ja me võime olla päris kindlad, et koefitsient erineb nullist. Vabaliikme p-väärtus on aga 0.012, ehk kui me kasutaksime usaldusnivood 99% (mille puhul p-väärtus peaks olema väiksem kui 0.01), siis me ei saaks järeldada, et see on statistiliselt oluliselt erinev nullist. Lisaks kuvab R iga p-väärtuse taha ka tärnid, mis indikeerivad selle väärtuse suurust lähtuvalt allolevast legendist.

Miks meil on üldse vaja teada kas koefitsiendid erinevad oluliselt nullist? Aga sellepärast, et kui regressioonisirge oleks 0, siis meie tunnuste vahel ei oleks seost (kui X muutub 1 ühiku võrra, siis Y muutub 0 ühiku võrra, ehk siis Y väärtus ei sõltu X'i väärtusest). Aga kuidas on lood vabaliikmega? Kas ka see peab erinema nullist, et meie mudelist mingit tolku oleks? Tegelikult ju ei pea. Võib täitsa vabalt juhtuda, et regressioonisirge lähebki läbi X ja Y telgede ristumiskoha (Y on 0 kui X on 0). Sellisel juhul oleks vabaliikme t-väärtus väiksem kui 2 ja p-väärtus suurem kui 0.05, kuid mudeli tõlgendust see ei mõjutaks. Ehk siis tavaliselt meid vabaliikme p ja t väärtused väga ei huvita. Küll aga peaks jälgima, et standardviga väga suur (võrreldes vabaliikme endaga) ei oleks.

2.4.3 Jääkide standardviga

```
## Residual standard error: 555.8 on 3982 degrees of freedom
##   (3648 observations deleted due to missingness)
```

Kuidas hinnata regressiooniprognoosi täpsust, ehk siis seda kui hästi regressioonimudel sobitub andmetega (*model fit*)? Üheks võimaluseks on lähtuda samast loogikast mida kasutame tunnuse keskväärtuse täpsuse hindamisel. Ehk kui palju vaatlused keskmiselt erinevad keskväärtusest. Regressioonijoone puhul ei ole meil ühte keskväärtust, mille suhtes vaatluste hälbimist määrata. Kuid iga vaatluse sõltumatu tunnuse väärtuse x kohta on meil "hinnatud" sõltuva tunnuse väärtus $\hat{y}$. Seega tuleb meil lihtsalt vaadata kui palju vaatluste y ja $\hat{y}$ väärtused keskmiselt erinevad, ehk kui suur on keskmine viga meie mudelis. Regressioonanalüüsi kontekstis kutsutakse seda vaatluste varieeruvuse näitajat keskmiseks ruutveaks (*mean squared error*) ehk lühidalt MSE[9]. Kuna aga MSE väärtus on ruudus, siis on seda keeruline interpreteerida (samamoodi nagu ka dispersiooni). Kui me võtame ruutjuure MSE'st, $\sqrt{MSE}$, saame regressiooni jääkide standardhälbe, mida nimetatakse **jääkide standardveaks** (*residual standard error* ehk RSE). Mida väiksem on mudeli RSE, seda paremini mudel andmetega sobitub (seda vähem hälbivad vaatlused regressioonijoonest ehk seda väiksemad on regresiooni jäägid). See, kui väike peaks RSE väärtus hea mudeli korral olema, sõltub eelkõige kontekstist ja sõltuva tunnuse skaalast (samamoodi nagu keskväärtuse standardhälve). Mingeid konkreetseid piirväärtusi siinkohal tuua ei ole võimalik.
Lisaks on siin ära toodud ka *degrees of freedom* ehk vabadusastmete arv jääkide standardvea arvutamisel. Sisuliselt on siin kirjas analüüsi kaasatud vaatluste arv (miinus regressioonikordajate arv, siinses mudelis 2). Ära on toodud ka analüüsist välja jäetud vaatluste arv. Need on need, kellel puudus väärtus vähemalt ühe analüüsitava tunnuse jaoks.

[9] $MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$