Línuleg líkön

Línuleg tölfræðilíkön

Tölfræðilíkön eru verkfæri sem við notum til að kanna samband *háðra breyta* (dependent variables) og einnar eða fleiri *óháðra breyta* (independent variable). Við munum aðeins fjalla um línuleg líkön með einni háðri breytu. Stundum má fullyrða um orsakasamband á milli breytanna og þá er *svarbreytan* ávalt háða breytan og *skýribreyturnar* þær óháðu. Í kennslubókinni skoðum við aðeins líkön með einni óháðri breytu en í þessari samantekt leyfum við tvær óháðar breytur en almennt geta breyturnar verið fleiri. Líkönin sem við skoðum hér má því skrifa sem:

$$\text{háð breyta} \sim \text{óháð breyta}_1 + \text{óháð breyta}_2$$

Það fer eftir gerð óháðu breytanna að hvaða gerð líkanið okkar er:

- Ef skýribreyturnar eru samfelldar talnabreytur (eða strjálar talnabreytur sem geta tekið mörg gildi) er talað um *aðhvarfsgreiningu* (regression).
- Ef skýribreyturnar eru flokkabreytur er talað um *fervikagreiningu* (ANOVA).
- Ef önnur breytan er samfelld talnabreyta og hin er flokkabreyta er talað um *samvikagreiningu* (ANCOVA).

Í R notum við `lm()`-aðferðina til að meta þessar þrjár gerðir af línulegum líkönum.

Þessi samantekt er eingöngu ætluð til að sýna hvernig framkvæma má ofangreindar aðferðir í R. Áður en þið beitið þessum aðferðum í framtíðinni mæli ég með að þið lesið ykkur vel til um þær, sér í lagi hvaða forsendur þurfa að gilda.

Einþátta fervikagreining sem línulegt líkan

Í kafla 9 í bókinni kynntumst við einföldustu gerð fervikagreiningar, *einhliða fervikagreiningu*, sem einnig má kalla *einþátta fervikagreiningu*. Við höfum séð hvernig einþátta fervikagreining er notuð til að draga ályktanir um meðaltöl tveggja eða fleiri hópa en hana má einnig setja fram sem *línulegt líkan*.

Við beitum einþátta fervikagreiningu ef óháða breytan okkar er aðeins ein og hún er strjál. Í dæminu sem við skoðuðum í bókinni er háða breytan okkar breyting á blóðþrýstingi og óháða breytan okkar er lyf. Við notum bókstafinn a til að tákna fjölda flokka/hópa óháðu breytunnar. Í dæminu okkar er $a = 3$ þar sem við erum með þrjú lyf.

Línulega líkanið má skrifa sem:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

þar sem $i = 1, 2, ..., a$ og $j = 1, 2, ..., n$. Hér gerum við ráð fyrir að við höfum jafn margar mælingar í hverjum flokki/hópi (n).

- y_{ij} er mæling nr. j í hópi/flokki nr. i
- μ er heildarmeðaltalið
- τ_i er frávik flokks nr. i frá heildarmeðaltalinu μ
- ε_{ij} eru frávik mælingar nr. j frá gildinu $\mu + \tau_i$ sem henni tilheyrir, við köllum ε_{ij} *leifar*.

Venja er að tákna óháðu breytuna með grískum bókstaf en líkanið úr dæminu okkar gætum við líka skrifað sem:

$$y_{ij} = \mu + \text{lyf}_i + \varepsilon_{ij}$$

Tilgáturnar má rita sem:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_a$$

á móti gagntilgátunni

$$H_1 : \text{Að minnsta kosti eitt meðaltal er frábrugðið hinum}$$

Jafngildar tilgátur eru:

$$H_0 : \tau_1 = \tau_2 = ... = \tau_a = 0$$

á móti gagntilgátunni

$$H_1 : \text{Að minnsta kosti ein } \tau_i \text{ er ekki núll}$$

Fervikagreiningartöfluna skrifum við eins og áður en hér notum við A í stað Tr í bókinni þar sem við munum bæta annarri háðri breytu inn í líkanið á eftir (hún verður táknuð með B í töflunni).

Fervikasummur	Frígráður	Meðalfervikasummur
SS_A	$a-1$	$MS_A = \frac{SS_A}{a-1}$
SS_E	$N-a$	$MS_E = \frac{SS_E}{N-a}$
SS_T	$N-1$	

Prófstærðin er:

$$F = \frac{SS_A/(a-1)}{SS_E/(N-a)} = \frac{MS_A}{MS_E}.$$

Sé núlltilgátan sönn fylgir prófstærðin F-dreifingu með $a-1$ og $N-a$ fjölda frígráða, eða $F \sim F_{(a-1,N-a)}$, þar sem a er fjöldi flokka og N er heildarfjöldi mælinga. Hafna skal H_0 ef F $> F_{1-\alpha,(a-1,N-a)}$.

Við getum notað `lm()`-aðferðina til að framkvæma fervikagreiningu (einnig sýnt í R frá grunni, þar er úttakið einnig útskýrt):

```
fit1<-lm(blod~lyf,data=dat)
anova(fit1)


## Analysis of Variance Table
##
## Response: blod
##           Df Sum Sq Mean Sq F value  Pr(>F)
## lyf        2 144.57  72.286  9.2175 0.00245 **
## Residuals 15 117.63   7.842
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Það þarf að passa að óháða breytan sé flokkabreyta (mögulega þarf að breyta henni í flokkabreytu með `factor()`) aðferðinni.

Skoðum nú aftur líkanið hér að ofan og hvernig við fáum mötin á stikunum í R. Líkanið má almennt skrifa sem

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

en það má setja líkanið fram á nokkra mismunandi vegu. Sjálfgefna stillingin í R er að matið á μ, $\hat{\mu}$, er reiknað fyrir tiltekinn *viðmiðunarflokk*, en gildin $\hat{\tau}_i$ lýsa frávikum frá meðaltali þess flokks. $\hat{\mu}$ er því ekki matið á heildarmeðaltalinu heldur matið á meðaltalinu í viðmiðunarflokknum.

Mat á meðaltölum fyrir aðra flokka en viðmiðunarflokkinn eru svo fengin með:

$$\hat{\mu}_i = \hat{\mu} + \hat{\tau}_i$$

þar sem i er númer flokksins sem reikna á matið fyrir.

R velur þann flokk sem er fremstur í stafrófinu sem viðmiðunarflokk. Ef flokkarnir voru kóðaðir sem talnabreytur mun R velja flokkinn með lægsta gildið sem viðmiðunarflokk.

Mötin á stikunum má fá fram með að mata `summary()` aðferðina með `lm()`-hlut:

```
summary(fit1)
```

```
##
## Call:
## lm(formula = blod ~ lyf, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8533 -1.8363 -0.1317  2.6462  4.0350
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    8.143      1.143   7.123 3.49e-06 ***
## lyf2          -1.858      1.617  -1.149  0.26840
## lyf3           4.863      1.617   3.008  0.00883 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.8 on 15 degrees of freedom
## Multiple R-squared:  0.5514,Adjusted R-squared:  0.4916
## F-statistic: 9.218 on 2 and 15 DF,  p-value: 0.00245
```

`(Intercept)` gefur matið á meðaltalinu fyrir viðmiðunarflokkinn, stuðlar fyrir aðra flokka eru frávik þeirra meðaltala frá meðaltali viðmiðunarflokksins. Skoðum þetta fyrir gögnin okkar:

Flokkur	Tákn	Reikningur	Útkoma
lyf 1	$\hat{\mu}$	8.143	8.143
lyf 2	$\hat{\mu} + \hat{\tau}_1$	8.143 - 1.858	6.285
lyf 3	$\hat{\mu} + \hat{\tau}_2$	8.143 + 4.863	13.006

Þið hafið nú séð tvenns konar úttök úr línulegum líkönum þar sem `lm()` aðferðin hefur verið notuð til að meta líkanið. Tökum saman hvenær við notum hvort úttak:

- Við skoðum `anova(fit1)` til að framkvæma tilgátuprófið.
  - Núlltilgátan er að meðalgildi allra flokkanna séu jöfn.
  - Ef við höfnum henni fullyrðum við að munur sé á milli flokka.
- Við skoðum `summary(fit1)` til að sjá hvert matið á meðalgildunum er.
  - Það mat er fyrst og fremst áhugavert ef það er munur á flokkunum.

Tveggja þátta fervikagreining

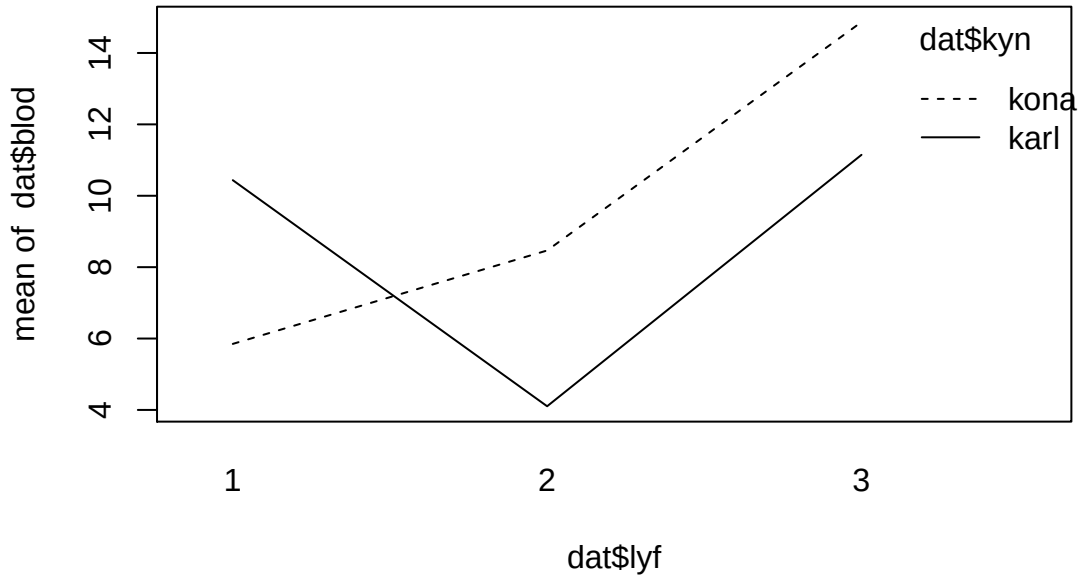
Ef við höfum tvær óháðar breytur í línulegu líkani og þær eru báðar flokkabreytur talað um tveggja þátta fervikagreiningu. Hugsum okkur aftur dæmið hér að ofan og nú með þeim aukaupplýsingum að lyfin voru prófuð á körlum og konum. Nú höfum við tvær flokkabreytur, lyf og kyn, og því getum við notað tveggja þátta fervikagreiningu til að kanna tengslin. Við notum a til að tákna fjölda hópa í fyrri flokkabreytunni (líkt og áður) en b til að tákna fjölda flokka í þeirri seinni. Gögnin okkar eru:

	Lyf		
	1	2	3
konur	4.29	10.32	12.89
	5.37	8.85	15.68
	7.89	6.23	16.03
karlar	11.28	3.23	9.43
	8.10	4.51	12.86
	11.93	4.57	11.15

Hér er $a = 3$, $b = 2$ og $n = 3$. Takið eftir að n táknar hér fjöldi mælinga í hverri "blöndu"af flokkum (t.d. lyf 1 - kona og lyf 3 - karl).

Það má vera að svo kölluð *víxlhrif* (interaction) séu til staðar á milli breytanna tveggja en þá víxlverka þær hvor við aðra. Í okkar dæmi myndi það þýða að áhrif lyfjanna eru ólík eftir kynjum. Gott er að skoða gögnin myndrænt til að kanna hvort víxlhrif séu til staðar. Við gerum það í R með *víxlhrifamynd*:

```
interaction.plot(dat$lyf,dat$kyn,dat$blod)
```



Á myndinni má sjá meðalblóðþrýstingslækkun fyrir "hópana"6. Sjá má á myndinni að lyfin virðast virka mismunandi á blóðþrýsting kvenna og karla.

Línulega líkanið má almennt skrifa sem:

$$y_{ijk} = \mu + \tau_i + \gamma_j + (\tau\gamma)_{ij} + \varepsilon_{ijk}$$

þar sem $i = 1, 2, ..., a$, $j = 1, 2, ..., b$ og $k = 1, 2, ..., n$ og $(\tau\gamma)_{ij}$ eru víxlhrifin milli τ og γ.

Fervikagreiningartaflan er:

Fervikasummur	Frígráður	Meðalfervikasummur
SS_A	$a-1$	$MS_A = \frac{SS_A}{a-1}$
SS_B	$b-1$	$MS_B = \frac{SS_B}{b-1}$
SS_{AB}	$(a-1)\cdot(b-1)$	$MS_{AB} = \frac{SS_{AB}}{(a-1)\cdot(b-1)}$
SS_E	$ab\cdot(n-1)$	$MS_E = \frac{SS_E}{ab\cdot(n-1)}$
SS_T	$N-1$	

Við byrjum á að kanna hvort víxlhrif séu til staðar:

$$H_0 : (\tau\gamma)_{11} = (\tau\gamma)_{12} = ... = (\tau\gamma)_{21} = ... = (\tau\gamma)_{ab} = 0$$

á móti gagntilgátunni

$$H_1 : \text{Að minnsta kosti ein } (\tau\gamma)_{ij} \text{ er ekki 0.}$$

Prófstærðin til að kanna hvort víxlhrif séu til staðar er:

$$F = \frac{SS_{AB}/(a-1)\cdot(b-1)}{SS_E/ab\cdot(n-1)} = \frac{MS_{AB}}{MS_E}.$$

Sé núlltilgátan sönn fylgir prófstærðin F-dreifingu með $(a-1)\cdot(b-1)$ og $ab\cdot(n-1)$ fjölda frígráða, eða $F \sim F_{((a-1)\cdot(b-1),ab\cdot(n-1))}$. Hafna skal H_0 ef F > $F_{1-\alpha,((a-1)\cdot(b-1),ab\cdot(n-1))}$

Skoðum nú hvernig má fá fervikasummurnar og kanna tilgátuna í R:

```
fit2<-lm(blod~lyf + kyn + lyf:kyn, data=dat)
anova(fit2)
```

```
## Analysis of Variance Table
##
## Response: blod
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## lyf        2 144.572  72.286 23.5961 6.942e-05 ***
## kyn        1   6.113   6.113  1.9956  0.183168
## lyf:kyn    2  74.759  37.379 12.2016  0.001283 **
## Residuals 12  36.762   3.063
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Við höfnum núlltilgátunni og ályktum að það séu í raun víxlhrif til staðar (p-gildi = 0.001283).

Séu víxlhrif til staðar prófum við ekki hina þættina í líkaninu. Ef engin víxlhrif eru til staðar þá fjarlægjum við víxlhrifin úr líkaninu, metum það upp á nýtt og prófum hina þættina tvo.

Skoðum nú aftur líkanið hér að ofan og hvernig við fáum mötin á stikunum í R:

$$y_{ijk} = \mu + \tau_i + \gamma_j + (\tau\gamma)_{ij} + \varepsilon_{ijk}$$

Það má setja líkanið fram á fleiri en einn hátt. Eins og áður er matið á μ, $\hat{\mu}$, reiknað fyrir tiltekinn viðmiðunarflokk, og eru þeir flokkar sem eru fremstir í stafrófinu valdir sem viðmiðunarflokkur.

Mötin á stikunum má fá fram með að mata `summary()` aðferðina með `lm()`-hlut:

```
summary(fit2)
```

```
##
## Call:
## lm(formula = blod ~ lyf + kyn + lyf:kyn, data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.337 -1.388  0.395  1.083  2.040
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.850      1.011   5.789 8.62e-05 ***
## lyf2           2.617      1.429   1.831 0.092031 .
## lyf3           9.017      1.429   6.309 3.89e-05 ***
## kynkarl        4.587      1.429   3.209 0.007499 **
## lyf2:kynkarl  -8.950      2.021  -4.428 0.000823 ***
## lyf3:kynkarl  -8.307      2.021  -4.110 0.001446 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.75 on 12 degrees of freedom
## Multiple R-squared:  0.8598,Adjusted R-squared:  0.8014
## F-statistic: 14.72 on 5 and 12 DF,  p-value: 9.192e-05
```

Skoðum þetta fyrir gögnin okkar:

Flokkur	Tákn	Reikningur	Útkoma
lyf 1, konur	$\hat{\mu}$	5.85	5.85
lyf 2, konur	$\hat{\mu} + \hat{\tau}_2$	5.85 + 2.617	8.467
lyf 3, konur	$\hat{\mu} + \hat{\tau}_3$	5.85 + 9.017	14.867
lyf 1, karlar	$\hat{\mu} + \quad \gamma_2$	5.85 + 4.587	10.437
lyf 2, karlar	$\hat{\mu} + \hat{\tau}_2 + \gamma_2 + (\tau\gamma)_{22}$	5.85 + 2.617 + 4.587 - 8.950	4.104
lyf 3, karlar	$\hat{\mu} + \hat{\tau}_3 + \gamma_2 + (\tau\gamma)_{32}$	5.85 + 9.017 + 4.587 - 8.307	11.147

Ef þið farið að vinna með fjölþátta fervikagreiningu í framtíðinni skulið þið einnig skoða `drop1()` skipunina.

Aðhvarfsgreining

Í 10. kafla bókarinnar má lesa um einfalt línulegt aðhvarf, þar sem ein samfelld talnabreyta, eða strjál talnabreyta sem getur tekið mjög mörg gildi, (óháða breytan) er notuð til að spá fyrir um aðra samfellda talnabreytu (háða breytan). Þegar einfalt línulegt aðhvarf er framkvæmt er gert ráð fyrir að samband háðu breytunnar og óháðu breytunnar sé *línulegt* þ.e.a.s. að hægt sé að lýsa sambandi þeirra með jöfnu beinnar línu. *Einfalda aðhvarfsgreiningarlíkanið* (simple linear regression model) má skrifa sem

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

þar sem β_0 og β_1 eru óþekktir stikar, ε er normaldreifð slembistærð með meðaltal 0. Í R-heftinu má sjá hvernig meta má líkanið í R.

Ef við bætum við fleiri óháðum breytum fáum við fjölvíða aðhvarfsgreiningu sem við náum ekki að fjalla um í þessu námskeiði - ég mæli með því að þið farið í framhaldsnámskeið í tölfræði og lærið um hana þar.

Samvikagreining - ANCOVA

ANCOVA greining er blanda af fervikagreiningu og aðhvarfsgreiningu. Við höfum sem áður eina háða samfellda talnabreytu en óháðu breyturnar okkar eru blanda af flokkabreytum og samfelldum talnabreytum (eða strjálum talnabreytum sem geta tekið mörg gildi). Hér munum við skoða tilfellið þegar háðu breyturnar okkar eru ein af hvorri gerð.

Það er yfirleitt svo að við höfum áhuga á að kanna hvort meðaltöl hópa flokkabreytunnar eru mismunandi en til að fá áreiðanlegra próf leiðréttum við fyrir áhrifum talnabreytunnar (af því gefnu að sambandið á milli óháðu talnabreytunnar og háðu breytunnar sé línulegt). Þetta má hugsa sem svo að í staðinn fyrir að lýsa sambandinu á milli háðu og óháðu breytunnar með einni línu, eins og í einföldu línulegu aðhvarfi, verða línurnar eins margar og flokkar flokkabreytunnar okkar eru. Við gerum hér ráð fyrir að hallatölur línanna séu þær sömu en skurðpunktarnir geta verið mismunandi.

Línulega líkanið má skrifa sem:

$$y_{ij} = \tau_i + \beta \cdot x_{ij} + \varepsilon_{ij}$$

Við notum, sem áður, `lm()` aðferðina í R til að meta líkanið. Við mötum aðferðina með:

```
fit_prufa<-lm(nafn_a_hadu_breytu ~ nafn_a_ohadu_talnabreytu
              + nafn_a_ohadu_flokkabreytu,data=nafn_a_gagnasafni)
```

Notumst nú við gögn frá Western Collaborative Group en gögnin má nálgast í R. Við nálgumst gagnasafnið með að sækja `epitools` pakkann og keyra svo:

```
library(epitools)
data(wcgs)
```

Þið getið lesið um rannsóknina með `help(wcgs)`. Í gagnasafninu er m.a breyta sem lýsir efri mörkum blóðþrýstings (`sbp0`) og breyta sem lýsir persónugerð mannanna (`behpat0`). Hún inniheldur fjóra flokka en er skráð sem talnabreyta. Við breytum henni í flokkabreytu með:

```
wcgs$behpat0<-factor(wcgs$behpat0)
```

Við höfum nú áhuga á að kanna hvort efri mörk blóðþrýstings eru mismunandi í hópunum fjórum en þar sem blóðþrýstingur er mjög háður þyngd viljum við leiðrétta fyrir þyngd mannana. Við framkvæmum greininguna með:

```
fit3<-lm(sbp0 ~ weight0 + behpat0, data = wcgs)
```

Til að fá fervikageringartöfluna notum við `aov()` aðferðina:

```
anova(fit3)
```

```
## Analysis of Variance Table
##
## Response: sbp0
##             Df Sum Sq Mean Sq  F value  Pr(>F)
## weight0      1  46216   46216 216.8472 < 2e-16 ***
## behpat0      3   3248    1083   5.0796 0.00165 **
## Residuals 3149 671141     213
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Eins og sjá má er marktækur munur á milli hópanna (p-gildi = 0.00165). Til að fá mötin á stikum líkansins notum við `summary()` aðferðina eins og áður:

```
summary(fit3)


##
## Call:
## lm(formula = sbp0 ~ weight0 + behpat0, data = wcgs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.467 -10.102  -2.430   7.708 100.536
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 98.52304    2.29522  42.925   <2e-16 ***
## weight0      0.17948    0.01234  14.547   <2e-16 ***
## behpat02     0.71530    0.98396   0.727    0.467
## behpat03    -1.36447    0.99150  -1.376    0.169
## behpat04    -1.51271    1.19146  -1.270    0.204
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.6 on 3149 degrees of freedom
## Multiple R-squared:  0.06864,Adjusted R-squared:  0.06746
## F-statistic: 58.02 on 4 and 3149 DF,  p-value: < 2.2e-16
```

Gildið á skurðpunktinum fyrir viðmiðunarhóp flokkabreytunnar má lesa úr `(Intercept)` línunni (98.523). Skurðpunktinn fyrir hóp 2 má reikna með að leggja matið í `behpat02` línunni við matið fyrir viðmiðunarhópinn (98.523 + 0.715 = 99.238) og eins má finna möt skurðpunktana fyrir hina hópana tvo (98.523 + (-1.364) = 97.159) og (98.523 + (-1.513) = 97.01). Matið á hallatölu línanna má lesa út úr `weight0` línunni (0.179).