Олон хэмжээст өгөгдлийн статистик шинжилгээ хичээлийн бие даалтын үлгэрчилсэн жишээ

Ургийн өсөлт дэх тамхидалтын нөлөө

Г.Махгал

2021 оны 4 сарын 26

Агуулга

Удиртгал

Үүгээр шинэ төрсөн хүүхдийн биеийн жин болон өндөр буюу ургийн өсөлтөд эцэг эхийнх нь тамхидалт нөлөөтэй эсэхийг шинжлэн тогтоосон судалгааны үр дүнг танилцуулна. Шинжилгээнд олон хэмжээст хэвийн тархалт сэдвээр үзсэн олон хувьсагчийн регрессийн шугаман загвар улмаар шатлах алгоритм бүхий кластерын шинжилгээ болон хи-квадрат шинжүүр зэрэг зүйлс ашиглана.

1 Өгөгдөл

‘’Birthweight_reduced_kg_R.csv” [2] файлд буй өгөгдлийг X гэсэн датафрейм хэлбэртэй объект байдлаар ачаалав.

```
X <- read.csv(file = "Birthweight_reduced_kg_R.csv")
```

X датафрейм 42 мөр болон ID, Length, Birthweight, Headcirc, Gestation, smoker, mage, mnocig, mheight, mppwt, fage, fedyrs, fnocig, fheight, lowbwt, mage35 гэсэн 16 баганатай. Эдгээр багана буюу хувьсагчдын тайлбар болон хэмжээсийн төрөл зэрэг мэдээллийг хүснэгт 1 дээр харуулав.

Өгөгдөлд чанарын гурван хувьсагч байгаа бөгөөд эдгээрээс зайшгүй шаардлагатай нэг хувьсагчийг фактор төрөлд дараах байдлаар хувиргав.

```
X$smoker <- factor(x = X$smoker, levels = 1:0, labels = c("Yes", "No"), ordered = FALSE)
```

2 Шинжилгээний хэсэг

Эхлээд шинэ төрсөн хүүхдийн биеийн жин болон хэмжээг тамхидалттай шууд холбож үзье. Хүүхдийн биеийн жин болон хэмжээг Length, Birthweight, Headcirc, Gestation хувьсагчид илэрхийлэх тул эдгээр хувьсагчдыг ашиглана. Манай тохиолдолд хүүхдийн эх болон эцэг тамхи татдаг эсэхээс үүдсэн дөрвөн хувилбар байна.

```
table(
  X$smoker,
  factor(x = X$fnocig > 0, levels = c("TRUE", "FALSE"), labels = c("Yes", "No"), ordered = FALSE)
```

	Хувьсагчийн нэр	Тайлбар	Хэмжээсийн төрөл
1	ID	ялгах дугаар	дискрет
2	Length	хүүхдийн өндөр (см)	тасралтгүй
3	Birthweight	хүүхдийн жин (кг)	тасралтгүй
4	Headcirc	толгойн тойрог (см)	тасралтгүй
5	Gestation	тээсэн хугацаа (долоо хоног)	дискрет
6	smoker	эх нь тамхи татдаг бол 1 үгүй бол 0	нэрлэсэн
7	mage	эхийн нас	дискрет
8	mnocig	эхийн нь өдөрт татах тамхины тоо	дискрет
9	mheight	эхийн өндөр (см)	тасралтгүй
10	mppwt	эхийн жирэмслэлтээс өмнөх жин (кг)	тасралтгүй
11	fage	эцгийн нас	дискрет
12	fedyrs	эцгийн боловсролд зарцуулсан жил	дискрет
13	fnocig	эцгийн нь өдөрт татах тамхины тоо	дискрет
14	fheight	эцгийн өндөр (см)	тасралтгүй
15	lowbwt	дутуу жинтэй төрсөн бол 1 үгүй бол 0	нэрлэсэн
16	mage35	эх нь 35-аас дээш настай бол 1 үгүй бол 0	нэрлэсэн

Хүснэгт 1: Өгөгдөл дэх хувьсагчдын нэр, тайлбар болон хэмжээсийн төрөл

```
) ->
  freq
print(freq)
```

```
##
##       Yes No
##   Yes  18  4
##   No    9 11
```

Энэ нь ургийн өсөлтөд ямар нөлөө үзүүлдэг нь одоохондоо тодорхойгүй юм. Иймд кластерын тоо тодорхойгүй байх үед хийдэг шатлах алгоритм бүхий кластерын шинжилгээ [4, §5.2.1] хийнэ. Гэвч эх болон эцгийн тамхидалт хамааралгүй буюу тус тусдаа хүчин гэдэг нь батлагдахгүй байна. Үүнийг хи-квадрат шинжүүрийн [3, §13.1.3] тусламжтай тогтоосон.

```
chisq.test(freq, simulate.p.value = TRUE)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  freq
## X-squared = 6.1855, df = NA, p-value = 0.02549
```

Тодруулбал тамхи их татдаг ба бага эсвэл огт татдаггүй гэсэн хоёр бүлэгт эцэг эхчүүдийг хуваахад бүлгүүд хамааралгүй гэсэн тэг таамаглал үл няцаагдаж байна.

> Иймд эх болон эцгийн тамхидалттай холбогдуулан олон кластерт хуваах шаардлагагүй бололтой.

Түүнчлэн Length, Birthweight, Headcirc, Gestation хувьсагчдын утга агуулга үүнийгээ дагаад масштаб нь харилцан адилгүй байгаа тулд мастабын нөлөөг арилгахын тулд хувьсагчид дээр стандарт хувиргалт хийсний дараа зайн матриц олов.

```
d <- dist(scale(X[2:5]))
```

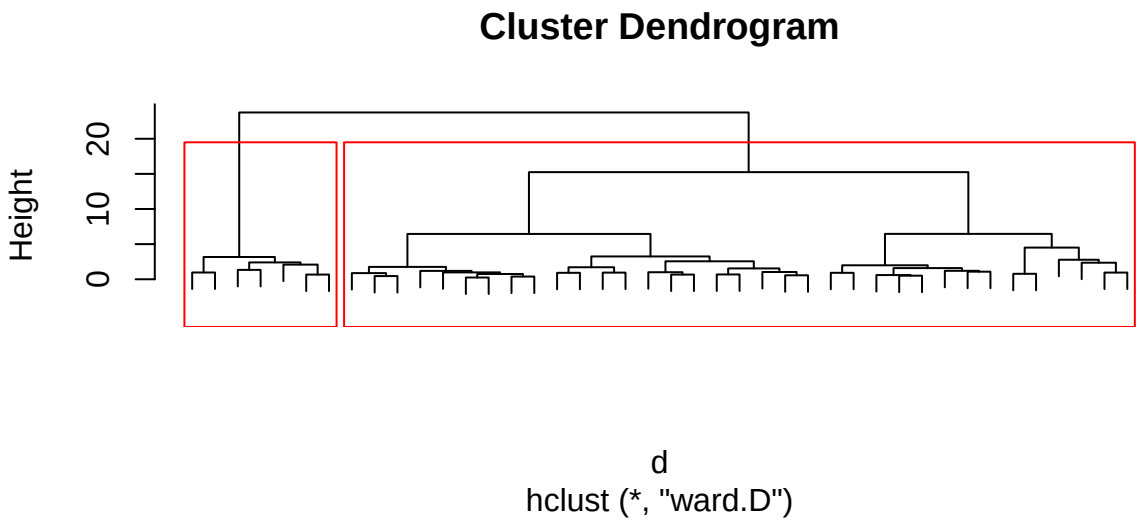
Кластер хоорондын зайн хувьд ‘’single’‘,’‘average” зэрэг нь утга учиргүй кластер хуваалт гаргаж байсан

тул ашиглаагүй. Харин ‘’complete” болон ‘’Ward” зай ашигласан үед бараг ижил кластер хуваалт гарсан. Иймд ‘’Ward” зай ашиглана.

```
cl <- hclust(d, method = "ward.D")
```

Шинжилгээгээр тод ялгарал бүхий кластерууд ажиглагдаж байгааг зураг 1 дээрх дендрограммаас харж болно.

```
plot(cl, labels = FALSE)
rect.hclust(cl, k = 2, border = "red")
```



Зураг 1: Шинэ төрсөн хүүхдийн биеийн хэмжээг хоёр кластерт хуваааж ангилсан нь

Дендрограммыг ажиглавал кластерын тоог 2 эсвэл 3 хэмээн авч болохоор ажээ.

Шатлах алгоритм бүхий кластерын шинжилгээгээр тод ялгарал бүхий кластерууд гарсан ч эдгээр нь хүүхдийн эцэг эхийн тамхидалттай ямар ч холбоогүй байна. Тухайлбал кластерын тоо 2 байх тохиолдлыг авч үзье.

```
groups <- cutree(cl, k = 2)
```

Энэ тохиолдолд тус хоёр кластер эхийн тамхи татах байдалтай хамааралтай болох нь батлагдахгүй байгааг дараах байдлаар тогтоож болно.

```
chisq.test(table(groups, X$smoker), simulate.p.value = TRUE)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  table(groups, X$smoker)
## X-squared = 1.2218, df = NA, p-value = 0.3973
```

Хи-квадрат шинжүүрийн хувьд тэг таамаглал нь чанар хувьсагчид хамааралгүй гэсэн байдгийг санууллъя. Цаашилбал эх эцэг нь тамхи их татдаг эсэхтэй ч холбогдолгүй байгааг дараах үр дүнгээс харна уу.

```
chisq.test(table(groups, X$fnocig > 25), simulate.p.value = TRUE)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
```

```
##
## data:  table(groups, X$fnocig > 25)
## X-squared = 3.0882, df = NA, p-value = 0.1229
```

> Иймд шинэ төрсөн хүүхдийн биеийн хэмжээ буюу ургийн өсөлтийг эцэг эхийн тамхидалтаар шууд тайлбарлах боломжгүй ажээ.

Хүүхдийн өндөр, жин зэрэгт эх, эцгийн өндөр, жин, нас зэрэг хүчин зүйлс нөлөөтэй байж болох тул эдгээрийн нөлөөг тэдгээр хувьсагчдаас зайлуулсаны дараа тамхидалтын нөлөөг сонирхож үзье. Үүний тулд дараах хоёр бүлэг хувьсагчдыг ялгаж авлаа.

```
X2 <- X[2:5]
X1 <- X[c(7,9:11,14)]
```

Ийнхүү

$$X_2 = (\text{Length, Birthweight, Headcirc, Gestation})$$

санамсаргүй векторооc

$$X_1 = (\text{mage, mheight, mppwt, fage, fheight})$$

санамсаргүй векторын нөлөөг зайлуулахаар

$$\begin{aligned} X_2 &= E(X_2|X_1) + U \\ &= \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(X_1 - \mu_1) + U \\ &= \beta_0 + BX_1 + U \end{aligned}$$

олон хувьсагчтай регрессийн шугаман загвар ашиглана [4, §2.8.1]. Энд $B = \Sigma_{21}\Sigma_{11}^{-1}$, $\beta_0 = \mu_2 - B\mu_1$ хэмээн үзсэн билээ. Цаашилбал $\mu_1 = E(X_1)$, $\mu_2 = E(X_2)$ харин $\Sigma_{21} = \text{cov}(X_2, X_1)$, $\Sigma_{11} = \text{cov}(X_1)$ байна. B матриц болон β_0 векторыг түүврээс

```
B <- cov(X2,X1) %*% solve(cov(X1))
b0 <- colMeans(X2) - B %*% colMeans(X1)
```

байдлаар үнэлбэл дараах үр дүн гарна.

$$B = \begin{pmatrix} -0.15 & 0.22 & 0.01 & 0.19 & 0.06 \\ -0.06 & 0.03 & 0.01 & 0.05 & -0.00 \\ -0.18 & 0.17 & -0.03 & 0.25 & 0.01 \\ -0.18 & 0.05 & 0.05 & 0.19 & 0.08 \end{pmatrix}$$

$$\beta_0 = \begin{pmatrix} 3.02 \\ -2.19 \\ 3.46 \\ 12.61 \end{pmatrix}$$

Улмаар

$$U = X_2 - (\beta_0 + BX_1)$$

байдлаар X_2 дээрх X_1 санамсаргүй векторын нөлөөг зайлуулна. Үүний тулд дараах хэлбэртэй код бичиж ажиллуулна.

```
U <- X2 - t(replicate(n = nrow(X2), expr = as.vector(b0)) + B %*% t(X1))
```

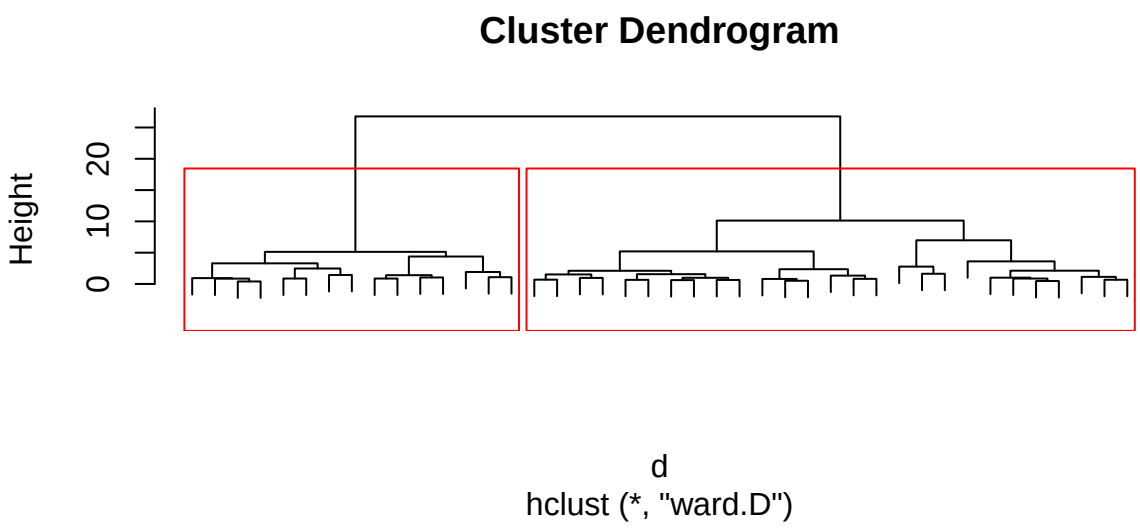
Ийнхүү X_2 буюу Length, Birthweight, Headcirc, Gestation хувьсагчдаас X_1 буюу mage, mheight, mppwt, fage, fheight хувьсагчдын нөлөөг зайлуулсанаа U буюу R програмын U объектод хадгаллаа. Мөн U нь датафрейм хэлбэртэй байна.

Ийнхүү тамхидалтын хажуугаар нөлөөлж буй хөндлөнгийн зарим хүчин зүйлсийн нөлөөг зайлуулсан тул өмнө хийсэн кластерын шинжилгээгээ ахин шинээр хийх боломжтой боллоо. Энэ тохиолдолд U датафреймд буй өгөгдөл дээр шинжилгээ хийх нь мэдээжийн хэрэг юм. Мөн энэ удаад ч ‘’Ward” зай тохиромжтой байна.

```
d <- dist(scale(U))
cl <- hclust(d, method = "ward.D")
```

Кластерын шинжилгээний үр дүнг зураг 2 дээрх дендрограммаас харна уу.

```
plot(cl, labels = FALSE)
rect.hclust(cl, k = 2, border = "red")
```



Зураг 2: Тамхидалтын хажуугаар нөлөөлөх хөндлөнгийн зарим хүчин зүйлсийн нөлөөг зайлуулсаны дараа шинэ төрсөн хүүхдийн биеийн хэмжээг хоёр кластерт хувааж ангилсан нь

Мөн энд олсон хоёр кластер эцэг эхийн тамхидалттай холбогдолтой болох нь хи-квадрат шинжүүрийн тусламжтай батлагдаж байна.

```
groups <- cutree(cl, k = 2)
chisq.test(table(groups, X$fnocig > 25), simulate.p.value = TRUE)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  table(groups, X$fnocig > 25)
## X-squared = 11.543, df = NA, p-value = 0.001999
```

Бид дээр хүүхдийн эцэг тамхи их татдаг эсэх нь эхийн тамхи их татдаг эсэхтэй холбоотойг тогтоосноо эргэн санана уу.

> Ийнхүү шинэ төрсөн хүүхдийн биеийн жин болон хэмжээнд эцэг эхийнх нь тамхидалт нөлөөтэй гэсэн үр дүнд хүрлээ.

Үүнийг бас эхийн тамхидалттай шууд холбосон байдлаар хийж гүйцэтгэх боломжтой. Нөхөр өдөрт 25-аас олон янжуур татдаг байх нь эхнэр 7 буюу үүнээс олон янжуур татахтай холбоо хамааралтай болохыг дараах байдлаар харж болно.

```
chisq.test(table(X$fnocig > 25, X$mnocig >= 7), correct = FALSE, simulate.p.value = TRUE)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
```

```
## data:  table(X$fnocig > 25, X$mnocig >= 7)
## X-squared = 5.5588, df = NA, p-value = 0.04748
```

Харин хүүхдийн эх өдөрт 6-аас олон янжуур татдаг байх нь бидний олсон хоёр кластертай холбогдолтой байж болохоор байгааг хи-квадрат шинжүүрийн үр дүнгээс харж болно.

```
chisq.test(table(groups, X$mnocig >= 7), correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(groups, X$mnocig >= 7)
## X-squared = 2.5926, df = 1, p-value = 0.1074
```

Түүнчлэн жирэмсэн эх тамхиа хаяж чадахгүй байлаа гэхэд тамхины хэрэглээгээ өдөрт 6 буюу үүнээс цөөн янжуур болтол бууруулах шаардлагатай гэсэн судалгааны дүгнэлт [1] байгаа нь бидний гаргасан үр дүн үндэслэлтэй болохыг нотолсон явдал юм.

3 Дүгнэлт

1. Шинэ төрсөн хүүхдийн биеийн жин болон өндөр буюу ургийн өсөлтөд эцэг эхийн тамхидалт нөлөөлж байна.
2. Тамхидалтын нөлөөг илрүүлэхийн тулд эцэг эхийн өндөр, биеийн жин зэрэг хүчин зүйлсийг тооцох шаардлагатай.
3. Жирэмсэн эх ургийн өсөлтдөө мэдэгдэхүйц дарамт учруулахгүй гэвэл тамхиа хаях буюу хэрэглээгээ өдөрт 6 янжуураас ихгүй болтол бууруулах хэрэгтэй.

Ашигласан материал

[1] Mariana Caricati Kataoka andothers. “Smoking during pregnancy and harm reduction in birth weight: a cross-sectional study”. in: BMC Pregnancy and Childbirth 18.1 (march 2018), page 67. ISSN: 1471-2393. DOI: 10.1186/s12884-018-1694-4. URL: https://doi.org/10.1186/s12884-018-1694-4.

[2] Ellen Marshall. Birthweight reduced data set. URL: https://www.sheffield.ac.uk/mash/statistics/datasets.

[3] Г.Махгал Б.Жамъяншарав. Магадлал статистикийн удиртгал. МУИС Пресс, 2017. ISBN: 9789997858320.

[4] Г.Махгал Ш.Мөнгөнсүх. Олон хэмжээст өгөгдлийн статистик шинжилгээ. 2017. ISBN: 9789997816481.