Bayesi regresszió

Szakdolgozat

Szabó Kinga Eszter

2020. december 28.

Eötvös Loránd Tudományegyetem
Természettudományi Kar
Valószínűségelméleti és Statisztikai Tanszék

Témavezető:

Csiszár Villő
Adjunktus

Szabó Kinga Eszter
Elemző matematika Bsc

Tartalomjegyzék

I. Téma ismertetése **4**

II. Bayesi statisztika **5**

1. Bayes tétel **5**
1.1. Példa Bayes tételére 6

2. Hipotézisvizsgálat folytonos paraméter esetén **8**
2.1. Példa a bayesi hipotézisvizsgálatra: 9

3. A priori eloszlások **10**

III. Regresszió **14**

4. Magyarázó változók mátrixa **14**

5. Klasszikus regresszió bemutatása **15**
5.1. Példa klasszikus regresszióra 16

6. Bayesi regresszió **18**
6.1. Regresszió célja 19
6.2. Szabályozás, méretek csökkentése 20

IV. Megközelítések összehasonlítása **21**

V. Gyakorlati példák, és R-es megoldások **23**

7. I.Feladat **23**
7.1. Klasszikus megközelítés 23
7.2. Bayesi megközelítés 24
7.2.1. Alapértelmezett priorokkal 24
7.2.2. Lapos priorokkal 25

8. II.Feladat **26**
8.1. Klasszikus megközelítés 26
8.2. Bayesi megközelítés 27
8.2.1. Alapértelmezett priorokkal 27
8.2.2. Lapos priorokkal 28

VI. Összefoglalás, eredmények kiértékelése **29**

Köszönetnyilvánítás

Elsősorban szeretném megköszönni témavezetőmnek Dr. Csiszár Villőnek, aki a témaválasztást javasolta, hogy türelmével, segítségével, elképesztő szakmai tudásával, és motivációjával nagyban elősegítette a szakdolgozatom létrejöttét.
Emellett szeretném megköszönni a családomnak, hogy támogattak mentálisan, fizikailag, anyagilag, és igyekeztek minél több terhet levenni a vállamról, hogy legyen elég időm, és energiám ezt a fantasztikus témakört megismerni, elmélyedni benne, feldolgozni, és ezt a szakdolgozatot létrehozni.
És végül de nem utolsó sorban szeretném megköszönni a páromnak, hogy mindvégig mellettem állt, kitartott mellettem amikor már úgy éreztem hogy nem bírom, bátorított, lelkesített, esetenként vigasztalt, és mindig feltöltött annyi erővel, hogy ne adjam fel, és tovább haladhassak a céljaim felé.

I. rész
Téma ismertetése

Szakdolgozatom témája a Bayesi regresszió lesz, ami egy átlagostól eltérő megközelítése a statisztikai következtetéseknek. Lényege, hogy Bayes tételét felhasználva írjuk fel a kezdeti lehetőségeket, mint változó értékeket, amelyeknek tudjuk valamilyen tulajdonságát, például az eloszlását. Ezen témakörnek az ismertetésénél nagyon fontos, hogy megvizsgáljuk az alapvető összefüggéseket, amik segíteni fogják a megértést, mivel ez a típusú statisztikai számítás eltér a hagyományos módszerektől.
A statisztikai következtetésekben mintákat figyelünk meg, és különböző tulajdonságai alapján modellt állítunk fel, következtetünk a jövőre, és emellett feltételezzük, hogy a majdani értékek is ugyanúgy fognak viselkedni, ahogy a mintaelemek alapján várható. De ezzel legtöbbször feltételezni kell azt, hogy a használt minta megismételhető, vagyis általánosítható az egész adatra, elég jól reprezentálja az egész halmaz tulajdonságait.
Ezzel ellentétben a bayesi megközelítés azt használja fel, hogy a modell paramétereire valószínűségi változókként tekintünk. Így ezeknek a változóknak meghatározhatjuk a prior vagyis az előzetes eloszlását, amelyet a minta, vagyis a megfigyelésünk módosíthat. Ebből kapjuk meg a paraméterek utólagos, vagyis a poszteriori eloszlását, ami alapján már adhatunk előrejelzéseket, illetve intervallumot amiben az az eloszlás megtalálható lesz.
A Bayes-féle következtetés központi eleme, a bizonytalanság közvetlen számszerűsítése, ami azt jelenti, hogy elvileg nincs akadálya a sok paraméterrel és bonyolult, többrétegű valószínűségmeghatározással rendelkező modellek felépítésének sem.

A gyakorlatban a problémák többsége az ilyen nagy modellek felállításával és számításával kapcsolatos.
A szakdolgozatom első felében az általános Bayesi gondolkodásmódról lesz szó. Foglalkozok vele, hogy hogyan épül fel, milyen részei vannak, ezek mit jelentenek, és hogyan befolyásolják a megoldást. Láthatóvá válik, hogy milyen esetekben hasznos, és milyen plusz információkkal tud szolgálni.
A második fejezetben a klasszikus, és a Bayesi lineáris regresszióról lesz szó egy kicsit behatóbban is. Először a klasszikus módszert tekintem át, majd a Bayesi módszert részletesebben is megvizsgálom.
Majd a kétféle megközelítést összehasonlítom, hogy láthatóvá váljanak az eltérések, illetve az összefüggések.
Ezután néhány feladaton keresztül R-es programkóddal is bemutatom az eltéréseket a két módszer között, illetve itt már képi megjelenése is lesz az adatoknak, ami szemléletesen mutatja be a feladatok megoldásait.

II. rész
Bayesi statisztika

Ebben fejezetben a Bayesi nézet általános áttekintésével fogok foglalkozni. Ez egy másik irányú gondolkodásmód, mint amivel általában a matematikában és a statisztikában találkozni szoktunk. Itt azt a kérdést tesszük fel, hogy a végeredményt amit kaptunk, mi befolyásolta. Arra gondolhatunk, hogy ha van egy adathalmazunk, és egy ismeretlen paraméterét szeretnénk tudni, akkor a matematika lehetőséget ad arra, hogy kiszámoljuk, hogy ha adunk egy pontos értéket a paraméternek, akkor az mennyire fog jól passzolni a mintához, tehát adott paraméter mellett mennyire lesz valószínű az adott minta. Ez alapján már felállíthatunk egy jobb modellt, ami megadja az utólagos eloszlásokat.

1. Bayes tétel

A névadó Thomas Bayes(1701-1761) angol matematikus, presbiteriánus lelkész.

A problémakör pedig amivel foglalkozott az volt, hogy ha egy dobozban fekete és fehér golyók vannak, akkor mennyi a valószínűsége, hogy feketét húzunk, mi történik ha többször húzunk, hogyan befolyásolja a megoldást, hogy melyik színből mennyi van a dobozban. Megoldása az inverz valószínűségre az "*An Essay towards solving a Problem in the Doctrine of Chances*" című munkájában jelent meg, de ezt csak tanítványa, Richard Price adta ki halála után, 1763-ban. Bayes így definiálja a valószínűséget: "Egy esemény valószínűsége az arány aközött, hogy mi az elképzelés, attól függően, hogy az eseménynek milyen lehetséges kimenetelei lehetnek, valamint a várt dolog értéke között a megfigyelt következtetések alapján." Ezt úgy fordíthatnánk le, hogy az esemény lehetséges kimenetelei az összes eset, a várt dolog értéke a következtetések alapján pedig a jó esetek száma. Tehát a valószínűség: jó eset osztva az összes esettel. Ennek egy lépéssel bonyolultabb változata a feltételes valószínűség, amikor egy másik eset bekövetkezésétől is függ a megfigyelésünk.Ezt a következőképpen fogalmazhatjuk meg:
Legyen A és B két egymástól nem független esemény. Ekkor ha ismerjük A és B események valószínűségét, ezek egyike sem 0, akkor a feltételes valószínűségek definíciójából könnyen adódik, hogy

$$P(AB) = P(A|B)P(B)$$

és természetesen

$$P(BA) = P(B|A)P(A)$$

és mivel az együttes bekövetkezés valószínűsége szimmetrikus, azaz:

$$P(AB) = P(BA)$$

a jobb oldalak egyenlővé tételével

$$P(A|B)P(B) = P(B|A)P(A)$$

Ezt átrendezve kapjuk Bayes tételét.

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Hogyha pedig B egy $B_1, B_2, ..., B_K$ teljes eseményrendszer egyik eseménye, és A valószínűségét a teljes valószínűség tétele szerint felbontjuk, akkor a Bayes-tétel másik alakját kapjuk.
Ez a következőképpen néz ki:

$$\sum_{i=1}^{K} P(A|B_i)P(B_i) = P(A)$$

Ezek mellett a feltételek mellett Bayes tétele:

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{K} P(A|B_i)P(B_i)}$$

ahol $j = 1, .., K$

1.1. Példa Bayes tételére

Tegyük fel, hogy egy betegség kimutatásához elvégeznek egy szűrővizsgálatot. Ez a vizsgálat tudjuk, hogy 90% valószínűséggel tudja kimutatni a betegséget. De megtörténhet, hogy valakit betegnek diagnosztizál, pedig egészséges. Ennek a valószínűsége 3 százalék. A betegség a lakosság 35%-át érinti. Azt szeretnénk tudni, hogy ha egy lakosról a tesztelés után az derül ki, hogy egészséges, mi a valószínűsége annak, hogy tényleg az.
Ha a betegség valószínűsége 35%, akkor annak a valószínűsége, hogy valaki egészséges, 65% lesz. Viszont ha a teszt mondja azt, hogy valaki egészséges, akkor annak nyilván nagyobb a valószínűsége. De persze nem biztos, hiszen a teszt sem tökéletes. De akkor vajon mekkora a valószínűsége annak, hogy valaki tényleg egészséges?

P(egészséges|teszt azt mondja, hogy egészésges)=?

Jelöljük az eseményeket a következőképpen:

B_1 = egészséges

B_2 = beteg

A = a teszt azt mondja, hogy egészséges

Felírhatjuk Bayes tételét:

$$P(B_1|A) = \frac{P(A|B_1)P(B_1)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2)}$$

$P(B_1) = 0.65$
$P(B_2) = 0.35$

Ha valaki egészséges, akkor a teszt 3 % eséllyel téved, és mondja betegnek, tehát 97 % az esély arra, hogy ha valaki egészséges, akkor a teszt is ezt mutatja.

Ezért $P(A|B_1) = 0.97$

Ha pedig valaki beteg, akkor azt 90 %-os valószínűséggel mutatja tényleg betegnek, és 10 %-ban pedig egészségesnek mondja.

Ebből $P(A|B_2) = 0.10$

Ezután már csak be kell helyettesíteni az egyenletbe a megfelelő értékeket, és megkapjuk a keresett valószínűséget.

$$P(B_1|A) = \frac{0.97 \cdot 0.65}{0.97 \cdot 0.65 + 0.1 \cdot 0.35} = 0.9474$$

Ennek alapján azt láthatjuk, hogy annak a valószínűsége, hogy egészségesek vagyunk ha a teszt azt mondja, az csak 94.74 % lesz, tehát ennyire pontos a teszt.

2. Hipotézisvizsgálat folytonos paraméter esetén

A XXI. századi statisztikában, már sok helyen megjelenik a bayes-i gondolkodásmód, lévén sokrétű problémákat kell matematikailag megfogalmazni, amiben nagyon nagy szerepe van a feltételes valószínűségeknek is. Elterjedt a használata hierarchikus modellek felépítésénél. Az a poszteriori eloszlás explicit kiszámítása legtöbbször reménytelen, de hatékony Monte Carlo módszerek állnak rendelkezésre. Ehhez érdemes áttekinteni a fentebb említett összefüggéseket. Amikor egy adat bekövetkezésének a valószínűségére vagyok kíváncsi, akkor minden lehetséges paramétert meg kellene próbálni beilleszteni a modellbe, hogy megkapjuk, melyik a legpontosabb. Diszkrét esetben ez egy $\sum$-ként fog megjelenni a nevezőben, folytonos esetben viszont pontosan a paraméteren vett integrál értékével fog megegyezni.
A hétköznapokban ez egy visszafelé következtetés az események között. Ez a fajta fordított gondolkodás amúgy a statisztikától nem áll olyan távol, elég csak a likelihood módszerre gondolni.
A statisztikai következtetésekről azt tudjuk, hogy minél nagyobb minta áll rendelkezésre, annál pontosabb előrejelzéseket lehet mondani. De figyelembe kell venni, hogy a hétköznapi életben bizonyos esetekben ezek a minták nem csak hogy drágák lehetnek, néha nagyon körülményes a megfigyelésük, sokszor pontatlanok, de egyéb nehezítő körülmények is felmerülhetnek, ami miatt törekedni kell rá, hogy minél kevesebb mintával kapjuk meg az általunk elvárt pontosságú eredményt.
Például egy gyógyszerkísérlet esetén a kezelt betegek túlélési arányát kell összehasonlítani a más kezelésben részesültekkel.

Ezek alapján a bayesi hipotézisvizsgálat lépései a következőképpen alakulnak:
1: Hipotézisek a priori valószínűségének megadása
2: Szabály megalkotása a döntéshez
3: Paraméterek prior sűrűségfüggvényének kiválasztása az egyes hipotézisek esetén
4: Poszterior sűrűségfüggvény kiszámítása az adatok alapján
5: Döntés

1: A Bayes-statisztikában, nincsen olyan nullhipotézisünk amire feltétlenül fókuszálni kellene csak több, egymással egyenrangú hipotézis. Ezért itt nem is kell sokat gondolkodni, hiszen a hipotézisek "számozásának" nincs jelentősége.

2: A legtöbb esetben ez is egyértelmű a bayesiánus módszernél, hiszen a döntés annyit jelent, hogy a legvalószínűbbnek talált hipotézist választjuk, vagyis azt amelyiknek a legnagyobb az a poszteriori valószínűsége.

3: A prior sűrűségfüggvény foglalja össze az előzetes ismereteinket. A lehetséges eloszlásokat a modell vizsgálatában fogom taglalni.

4: A prior sűrűségfüggvény és a minta valószínűsége, azaz a likelihood függvény segítségével létrehozzuk a poszterior sűrűségfüggvényt is. Folytonos esetben ehhez Bayes tételét fogjuk átalakítani:

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{\int_\theta P(y|\theta)P(\theta)d\theta}$$

Az integrál azért került a nevezőbe, mivel a valószínűségeket folytonos esetben a sűrűségfüggvény integrálja adja meg.

2.1. Példa a bayesi hipotézisvizsgálatra:

Egy tusfürdővel foglalkozó cég szeretne betörni egy új piacra, de az eddigi kampányai a népszerűsítésre nem voltak igazán jók: a legutóbbi közvélemény-kutatás szerint csak a lakosság 4 %-a ismerte a terméket név alapján. Emiatt felvettek egy új szakembert, hogy segítsen népszerűsíteni. Az elvárásuk az lesz, hogy legalább 8 %-ra nőjön az ismertség. A következő felmérésben százból kilenc ember ismerte fel a márkát. Sikeres volt-e a reklám, ha a főnökei csak akkor ismerik el eredményesnek a munkáját, ha legalább 90 % biztonsággal 8 %, vagy afölötti tusfürdő ismertsége?

$H_1 : p < 0.08$
$H_2 : p \geq 0.08$
A döntési szabály: $P(H_2) \geq 0.9$
Prior sűrűségfüggvény: $\sim Beta(1, 24)$
Ennek az a priori eloszlásnak a várható értéke 0.04, és eszerint a H_1 hipotézis a priori valószínűsége kb. 0.86, míg a H_2 hipotézisé csak kb. 0.14.
A likelihood függvény, és a poszterior sűrűségfüggvény:

$$P(\frac{9}{100}|p) = \binom{100}{9} p^9(1-p)^{91}$$

$$f(p|\frac{9}{100}) = \frac{P(\frac{9}{100}|p)f(p)}{\int_0^1 P(\frac{9}{100}|p)f(p)dp} = \frac{\binom{100}{9} p^9(1-p)^{91}(1-p)^{23}}{\int_0^1 \binom{100}{9} p^9(1-p)^{91}(1-p)^{23}dp}$$

A nevező egy konstans, tehát az a poszteriori sűrűségfüggvény $p^9(1-p)^{114}$ konstansszorosa. Ez egy olyan Béta eloszlás lesz, aminek a paraméterei 10 és 115. Ezután azt kell meghatároznunk, hogy emellett az a poszteriori eloszlás mellett a 0.08-nál nagyobb paraméterértékek poszterior valószínűsége.

A H_1 hipotézis a poszteriori valószínűsége $Beta(10, 115)$ eloszlás mellett 0.536 lesz, míg a H_2 hipotézisé 0.464 ami messze elmarad az elvárt 0.9-től. Így a következtetésünk az lesz, hogy nem volt elég sikeres a reklámkampány. Természetesen más a priori eloszlás választása esetén más következtetésre is juthattunk volna.

3. A priori eloszlások

Ebben a fejezetben megvizsgáljuk, hogy milyen a priori eloszlásokat szoktak használni a bayesi statisztikában. 2 féle lehetőséget különböztetünk meg:

Van előzetes információnk, tehát viszonylag nagy biztonsággal tudjuk a paraméterek eloszlását. Ekkor informatív priorról beszélünk.

Amikor nincsen előzetes információnk, vagy nagyon kicsi a minta, és ezért nem akarjuk, hogy az előzetes tudás befolyással legyen a megoldásra, akkor nem informatív priorokról beszélünk.

Konjugált priorok

Mivel a poszterior eloszlás egy szorzat, ezért célszerű ekkor úgy választani a prior eloszlást is, hogy megegyezzen likelihood függvény típusával, és ekkor a poszterior eloszlás is ugyanilyen lesz. Ebben az esetben konjugált priorokról beszélünk, amikre pontosan az igaz, hogy ugyanabba az eloszlás csoportba tartoznak, mint a belőlük képzett poszterior.

A konjugált priorok tulajdonságai:
-Egy prior több likelihood függvényre is lehet konjugált, de nem mindig lehet megadni, van olyan likelihood függvény, aminek nincs konjugált prior eloszlása.

Béta konjugált a priori eloszlást a következő esetekben használhatunk:

-A minta eloszlása $Binomialis(m, \theta)$ ahol m ismert, és θ az ismeretlen paraméter
a poszteriori eloszlás:

$$p(\theta|x) \sim Beta(\sum_{i=1}^{n} x_i + \alpha, nm - \sum_{i=1}^{n} x_i + \beta)$$

-A minta eloszlása $Bernoulli(\theta)$ ahol θ az ismeretlen paraméter
a poszteriori eloszlás:

$$p(\theta|x) \sim Beta(\sum_{i=1}^{n} x_i + \alpha, n - \sum_{i=1}^{n} x_i + \beta)$$

-A minta eloszlása $Geometriai(\theta)$ ahol θ az ismeretlen paraméter
a poszteriori eloszlás:

$$p(\theta|x) \sim Beta(\alpha + n, \beta + \sum_{i=1}^{n} x_i)$$

-A minta eloszlása $NegBin(r, \theta)$ ahol r ismert, és θ az ismeretlen paraméter
a poszteriori eloszlás:

$$p(\theta|x) \sim Beta(\alpha + nr, \beta + \sum_{i=1}^{n} x_i)$$

Gamma konjugált a priori eloszlást a következő esetekben használhatunk:

-A minta eloszlása $Gamma(\alpha, \beta)$ ahol α ismert, és az ismeretlen β-ra a prior $Gamma(\alpha_0, \beta_0)$ a poszteriori eloszlás:

$$p(\beta|x) \sim Gamma(\alpha_0 + n\alpha, \beta_0 + \sum_{i=1}^{n} x_i)$$

-A minta eloszlása $Exp(\lambda)$ és a prior λ-ra $Gamma(\alpha_0, \beta_0)$
a poszteriori eloszlás:

$$p(\lambda|x) \sim Gamma(\alpha_0 + n, \beta_0 + \sum_{i=1}^{n} x_i)$$

-A minta eloszlása $Poisson(\lambda)$ és a prior λ-ra $Gamma(\alpha_0, \beta_0)$
a poszteriori eloszlás:

$$p(\lambda|x) \sim Gamma(\alpha_0 + \sum_{i=1}^{n} x_i, \beta_0 + n)$$

-A minta eloszlása $Norm(\mu, \tau = \frac{1}{\sigma^2})$ ahol μ ismert, τ pedig ismeretlen, a prior eloszlás τ-ra $Gamma(\alpha_0, \beta_0)$
a poszteriori eloszlása:

$$p(\tau|x) \sim Gamma(\alpha_0 + \frac{n}{2}, \beta_0 + \frac{1}{2}\sum_{i=1}^{n}(x_i - \tau)^2)$$

Egyéb konjugált priorok

-A minta eloszlása $Norm(\mu, \tau = \frac{1}{\sigma^2})$ ahol τ ismert, μ ismeretlen, és a prior eloszlás μ-re $Norm(\mu_0, \tau_0)$
a poszteriori eloszlás:

$$p(\mu|x) \sim Norm(\frac{\mu_0\tau_0 + \tau\sum_{i=1}^{n} x_i}{\tau_0 + n\tau}, \tau_0 + n\tau)$$

-A minta eloszlása $Norm(\mu, \sigma^2)$, ahol μ ismert, σ^2 ismeretlen, és a prior σ^2-re $InverzGamma(\alpha, \beta)$ eloszlást követ. Az, hogy a σ^2 *inverzgamma* eloszlású azt jelenti, hogy az $\frac{1}{\sigma^2}$ *gamma* eloszlást követ.
a poszteriori eloszlás:

$$p(\sigma^2|x) \sim InverzGamma(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2)$$

-Végül álljon itt a normális eloszlás példája abban az esetben, ha egyik paraméterét sem ismerjük. Itt nem konjugált a priori eloszlást választunk, hanem megadjuk, hogy mi egy gyakori a priori eloszlás ebben az esetben.Amikor a minta $Norm(\mu, \tau)$ de egyik paraméterét sem ismerjük, akkor is fel tudjuk írni a poszterior eloszlást, felhasználva, hogy a prior eloszlás $p(\mu, \tau) \propto \frac{1}{\tau}$

a poszteriori eloszlás:

$$p(\mu,\tau|x) \propto \tau^{\frac{n}{2}-1} exp(-\frac{1}{2\tau}(\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu)^2))$$

Várható érték és szórásnégyzet a poszteriori eloszlása:

Úgy kaphatjuk meg a μ vagy σ^2 sűrűségfüggvényét, hogy az együttes sűrűségfüggvényt kiintegráljuk az egyik változó szerint.
A fenti példák közül illusztrációként levezetjük azt az esetet, amikor a minta normális eloszlású, a szórás ismert, a várható érték pedig ismeretlen paraméter. Azért választottuk ezt a példát, mivel ez a később tárgyalandó regressziós modellnek egy nagyon egyszerű speciális esete. Ebből fakad az is, hogy most a mintaelemeket nem x-ek, hanem y-ok jelölik.

Legyen $y_1, y_2, \ldots, y_n$ független, azonos eloszlású minta, ahol a mintaelemek normális eloszlásúak μ várható értékkel és σ^2 szórásnégyzettel. Az egyszerűség kedvéért feltesszük, hogy σ ismert. Megmutatjuk, hogy ekkor μ-re a konjugált a priori eloszlás normális, azaz ha normális eloszlást választunk a priori eloszlásnak, akkor az a poszteriori eloszlás is normális lesz.

Legyen tehát μ a priori eloszlása $N(\mu_0, \sigma_0^2)$, és $\tau_0 = 1/\sigma_0^2$, azaz

$$p(\mu) \propto \exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\} = \exp\left\{-\frac{\tau_0(\mu-\mu_0)^2}{2}\right\}.$$

Továbbá a minta likelihoodja

$$p(y|\mu) \propto \prod_{i=1}^{n} \exp\left\{-\frac{(y_i-\mu)^2}{2\sigma^2}\right\} = \prod_{i=1}^{n} \exp\left\{-\frac{\tau(y_i-\mu)^2}{2}\right\}.$$

Bayes tétele szerint az a poszteriori eloszlás konstans szorzótól eltekintve megegyezik az a priori sűrűség és a likelihood szorzatával:

$$p(\mu|y) \propto p(\mu)p(y|\mu) \propto \exp\left\{-\left(\frac{\tau_0(\mu-\mu_0)^2}{2} + \sum_{i=1}^{n}\frac{\tau(y_i-\mu)^2}{2}\right)\right\}.$$

A kitevőben kerek zárójelben álló mennyiség a μ-nek másodfokú függvénye. Alakítsuk ezt teljes négyzetté!

$$\frac{\tau_0(\mu-\mu_0)^2}{2} + \sum_{i=1}^{n}\frac{\tau(y_i-\mu)^2}{2} = \frac{\tau_0+n\tau}{2}\mu^2 - (\tau_0\mu_0 + \tau\sum y_i)\mu + \text{konst.}$$

Ebből teljes négyzetet kialakítva a következőt kapjuk:

$$\frac{\tau_0+n\tau}{2}\left(\mu - \frac{\tau_0\mu_0+\tau\sum y_i}{\tau_0+n\tau}\right)^2 + \text{konst.}$$

Ezt visszaírva az eredeti kifejezésbe, kapjuk, hogy

$$p(\mu|y) \propto \exp\left\{-\frac{\tau_0+n\tau}{2}\left(\mu - \frac{\tau_0\mu_0+\tau\sum y_i}{\tau_0+n\tau}\right)^2\right\},$$

ami éppen egy olyan normális eloszlás sűrűségfüggvénye, melynek várható értéke $\frac{\tau_0\mu_0+\tau\sum y_i}{\tau_0+n\tau}$, szórásnégyzetének reciproka pedig $\tau_0 + n\tau$.

Improprius priorok

Improprius priorokról akkor beszélhetünk, hogyha nem negatívak, de az integráljuk $[-\infty, \infty]$ között nem 1, hanem ∞. De ezekből a priorokból ugyanúgy számolhatunk fprmálisan a poszteriori eloszlást, és gyakran az már valódi eloszlás lesz, tehát sokszor még az ilyen kevés információt tartalmazó kezdeti feltétel is nagyon hasznos. Ilyen improprius a priori eloszlás például a lapos (konstans) prior a normális eloszlás várható értékére.

Nem informatív priorok

Amikor sok adat van, és csak néhány paraméter, akkor nagyon hasznos tud lenni a nem informatív prior eloszlás, és elfogadható eredményeket ad úgy, hogy kevesebbet kellett dolgozni, mintha a prior tudást írjuk fel valószínűségként. Kis méret, vagy sok paraméter esetén a likelihood függvény kevésbé élesen csúcsosodik, ezért akkor fontosabb a prior eloszlás, és a hierarchikus modell.

III. rész
Regresszió

Ebben a fejezetben fogom vizsgálni a lineáris regressziót, először a hagyományos, majd a Bayesi megközelítésből is. De előtte még fontos, hogy a magyarázó változók mátrixával megismerkedjünk, mivel ezek fogják az elméleti adatokat összefoglalni, ez segít abban, hogy egy feladatként tekinthessünk az egyenletekre.
A legáltalánosabb esetet rendes lineáris regressziónak nevezzük, ahol a megfigyelt hibák függetlenek, és egyenlő szórásúak. Vektorként felírva, a következőképpen néz ki:

$$y|\beta, \sigma, X \sim N(X\beta, \sigma^2 I)$$

4. Magyarázó változók mátrixa

Tekintsünk egy regressziós problémát, ahol a megjósolni kívánt függő változó nem egyetlen valós értékű skalár, hanem valós számok m hosszúságú vektora. Van n darab megfigyelés, ahol minden egyes megfigyelésnél i darab elem áll, $k-1$ magyarázó változó, csoportosítva egy k hosszú vektorba (ahova egy konstans 1 került még be). Ezt úgy lehet tekinteni, mint m darab kapcsolódó regressziós problémát.

$y_{i,1} = x_i^T \beta_1 + \epsilon_{i,1}$

$y_{i,2} = x_i^T \beta_2 + \epsilon_{i,2}$
...
$y_{i,m} = x_i^T \beta_m + \epsilon_{i,m}$

Ezt tekinthetjük egyetlen regressziós problémának, ahol az eredmény egy sorvektor, és a regressziós együttható vektorok is egymásmellé vannak téve, a következő képpen:

$y_i^T = x_i^T \mathbf{B} + \epsilon_i^T$

A $\mathbf{B}$ együttható mátrix egy olyan mátrix, ahol az egyes regressziós problémákhoz tartozó együttható vektorokat vízszintesen egymásra helyezzük:

$$\mathbf{B} = [(\beta_1)...(\beta_m)] = [\begin{pmatrix} \beta_{1,1} \\ . \\ . \\ . \\ \beta_{k,1} \end{pmatrix} ... \begin{pmatrix} \beta_{1,m} \\ . \\ . \\ . \\ \beta_{k,m} \end{pmatrix}]$$

A teljes problémát mátrixokkal írhatjuk fel:
$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$ Itt az X mátrix k darab magyarázó változó értékeit tartalmazza, a megfigyelések száma pedig n.

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} & ... & x_{1k} \\ 1 & x_{21} & x_{22} & x_{23} & ... & x_{2k} \\ 1 & x_{31} & x_{32} & x_{33} & ... & x_{3k} \\ ... & ... & ... & ... & ... & \\ 1 & x_{n1} & x_{n2} & x_{n3} & ... & x_{nk} \end{pmatrix}$$

5. Klasszikus regresszió bemutatása

A regresszió mint kifejezés az angol *"regression to the mean"* kifejezésből származik, amelynek jelentése, visszatérés az átlaghoz. Első említése Francis Galton-tól származik, aki férfiak, és fiaik magasságának összefüggéseit vizsgálta. Az ő meglátása szerint, a magasabb apáknak a fiaik is magasabbak lesznek, de nem ugyanannyival, és így a gyerekek visszatérnek az átlagos testmagasságához. A regressziószámítás lényege az, hogy megvizsgáljuk, hogy az általunk kiválasztott eredményváltozó hogyan függ a többi, úgynevezett magyarázó változótól. Tehát keressük a változókra vonatkozó adott tulajdonságú függvényeket, kapcsolatokat, és ezeket írjuk fel képlettel. Tehát a regressziószámítás lényege, hogy egy bizonyos változó (az eredményváltozó) hogyan függ más változók alakulásától (a magyarázó változóktól). Ez képlettel a következőképpen néz ki lineáris függvénykapcsolat esetén:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$$

Ekkor ε a hibatag.
y becslése:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + ... + \widehat{\beta}_k x_k$$

A megoldáshoz $\beta_0, \beta_1...$ paraméterek kellenek, amelyeknek kapcsán több kérdés is felmerül. Jól választottuk meg a modellünket? Helyesek a magyarázó változók? Meggyőzően magyarázzák az eredményváltozót? Ezeknek a bizonytalanságoknak a számszerűsítésére használjuk a hibákat. Megvizsgáljuk a reziduálisokat, azaz az illeszkedési hibákat azoknál a pontoknál, amelyek alapján a regressziót felírtuk. Ezek az alábbi formában jelennek meg:

$$e_i = y_i - \widehat{y_i}$$

A reziduálisokból képzett mutató, az eltérés-négyzetösszeg

$$SSE = \sum e_i^2 = \sum (y_i - \widehat{y_i})^2$$

Kiszámoljuk a reziduális szórást

$$s_e = \sqrt{\frac{SSE}{n}} = \sqrt{\frac{\sum e_i^2}{n}}$$

és az R^2 determinációs együtthatót, ami azt mutatja, hogy a modellben használt magyarázó változók, mekkora részét magyarázzák meg az y változó varianciájának.

$$R^2 = \frac{\sum_{i=1}^{n} (\widehat{y_i} - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

Ez az egyik legfontosabb információ a modell helyességéről. Ezután hipotézisvizsgálattal tesztelhetünk minden magyarázó változót, amivel kiszűrhetjük, hogy melyek az igazán fontosak, amik tényleg magyaráznak, és megbízhatóak.
Végül globális F-próbával az egész modellt teszteljük.

5.1. Példa klasszikus regresszióra

Ha túl sokat eszünk, el fogunk hízni. Ennek számszerűsítésére végeznek kísérletet 12 emberrel, akiknek a testtömegük azonos. Egy hónapig vizsgálták a kilók növekedését, ezek lesznek az eredményváltozók, a naponta bevitt energia és a napi mozgás a magyarázó változó, valamint még egy változó, ami 1 vagy 0 attól függően, hogy a vizsgált ember hajlamos-e a hízásra, vagy sem. A feladathoz tartozó adatokat egy mátrixba rendezzük, ahol az oszlopok rendre: szokásos konstans egyes oszlop, napi bevitt energia (kalória), mozgással töltött idő (perc), hízásra való hajlam. Az $\underline{y}$ vektor az egy hónap alatt bekövetkezett változást mutatja.

$$X = \begin{pmatrix} 1 & 2000 & 0 & 0 \\ 1 & 2200 & 15 & 0 \\ 1 & 2500 & 0 & 0 \\ 1 & 2600 & 15 & 0 \\ 1 & 2800 & 20 & 0 \\ 1 & 3000 & 30 & 0 \\ 1 & 2000 & 0 & 1 \\ 1 & 2100 & 15 & 1 \\ 1 & 2500 & 0 & 1 \\ 1 & 2600 & 15 & 1 \\ 1 & 2600 & 20 & 1 \\ 1 & 2800 & 30 & 1 \end{pmatrix} \quad \underline{y} = \begin{pmatrix} 0 \\ -1 \\ 1 \\ 1 \\ 0 \\ 2 \\ 1 \\ -1 \\ 2 \\ 2 \\ 1 \\ 1 \end{pmatrix}$$

Először kiszámítjuk a lineáris regresszió paramétereinek becslését: $\underline{\widehat{\beta}} = (X^T X)^{-1} X^T \underline{y}$

Ehhez szükség van az $(X^T X)^{-1}$ mátrixra, ami a következőképpen néz ki:

$$(X^T X)^{-1} = \begin{pmatrix} 8.8292 & -0.0038 & 0.0679 & -0.4843 \\ -0.0038 & 0.000002 & -0.00004 & -0.0001 \\ 0.0679 & -0.00004 & 0.0001 & -0.0029 \\ -0.4843 & -0.0001 & -0.0029 & 0.3452 \end{pmatrix}$$

$\widehat{\beta}_i$ paraméterei a $\underline{\widehat{\beta}}$ vektorban vannak

$$\underline{\widehat{\beta}} = (X^T X)^{-1} X^T \underline{y} = \begin{pmatrix} -7.7272 \\ 0.0036 \\ -0.0725 \\ 0.8044 \end{pmatrix}$$

Vagyis ezek szerint $\widehat{\beta}_0 = -7.7272$, $\widehat{\beta}_1 = 0.0036$, $\widehat{\beta}_2 = -0.0725$, $\widehat{\beta}_3 = 0.8044$
Ezeket értelmezhetjük:

$\widehat{\beta}_0$: A konstans tag

$\widehat{\beta}_1$: Ez mutatja meg, hogy ha napi egy kalóriával többet viszünk be, akkor az a hónap végén hány kiló pluszt fog okozni.

$\widehat{\beta}_2$: Ebből az látható, hogy napi egy perc mozgás hogyan befolyásolja a súlyt.

$\widehat{\beta}_3$: Azt mutatja, hogy ugyanannyi bevitt kalória és ugyanannyi mozgás mellett a hízásra hajlamos emberek mennyivel többet híznak, mint azok, akik nem hajlamosak hízásra.

Ezután kiszámoljuk $\widehat{y}$ értékét, az egyenletbe x-et írva.

$$\underline{\widehat{y}} = \begin{pmatrix} -0.53 \\ -0.89 \\ 1.27 \\ 0.55 \\ 0.90 \\ 0.89 \\ 0.28 \\ -0.45 \\ 2.07 \\ 1.35 \\ 0.98 \\ 0.98 \end{pmatrix} \quad \underline{y} = \begin{pmatrix} 0 \\ -1 \\ 1 \\ 1 \\ 0 \\ 2 \\ 1 \\ -1 \\ 2 \\ 2 \\ 1 \\ 1 \end{pmatrix}$$

$\overline{y} = 0.75$

Ekkor a determinációs együttható: $R^2 = 0.685$

Ebből azt látjuk, hogy a modell magyarázó ereje 68%.

6. Bayesi regresszió

Ebben a fejezetben a bayesi regresszió kerül a középpontba. Ez a regressziónak egy olyan fajtája, ahol Bayes-következtetéseket építhetünk a statisztikába. Ha a regressziós modell hibái függetlenek, és normális eloszlást követnek, valamint adott egy a priori eloszlás, akkor a modell paramétereinek eloszlására explicit eredményt kaphatunk.

A lineáris regresszió esetén a modellünk:

$$y_i = x_{i1}\beta_1 + \cdots + x_{ik}\beta_k + \varepsilon_i = x_i^\top \beta + \varepsilon_i, \quad i = 1, \ldots, n,$$

ahol $x_i = (x_{i1}, \ldots, x_{ik})^\top$ az i-edik megfigyeléshez tartozó magyarázó változók vektora, $\beta = (\beta_1, \ldots, \beta_k)^\top$ az ismeretlen együtthatók vektora, $\varepsilon_i \sim N(0, \sigma^2)$ pedig független hibák. Ekkor $y_i \sim N(x_i^\top \beta, \sigma^2)$ függetlenek $i = 1, \ldots, n$ esetén, melyet a következő alakban is írhatunk:

$$y|\beta, \sigma, X \sim N(X\beta, \sigma^2 I),$$

ahol I az $n \times n$-es identitás mátrix.

Kérdés, hogy milyen a priori eloszlást válasszunk a (β, σ) paraméterekre. Az egyik leggyakrabban használt (neminformatív) választás a

$$p(\beta, \sigma^2) \propto \sigma^{-2}$$

improprius a priori eloszlás, ez annak felel meg, hogy a $(\beta, \log \sigma)$ vektor sűrűsége konstans. Ekkor explicit formában megadható a paraméterek a poszteriori eloszlása, melyet a következő szorzat-alakban adunk meg:

$$p(\beta, \sigma^2|y) = p(\sigma^2|y)p(\beta|\sigma^2, y).$$

Itt a második tényező az együtthatók feltételes a poszteriori eloszlása σ^2-re, ami egy többdimenziós normális eloszlás:

$$\beta|\sigma^2, y \sim \mathrm{N}(\hat{\beta}, V_\beta \sigma^2),$$

ahol

$$\hat{\beta} = (X^\top X)^{-1} X^\top y, \quad V_\beta = (X^\top X)^{-1}.$$

Az első tényező pedig a σ^2 a poszteriori eloszlása, mely (skálázott) inverz-χ^2 lesz:

$$\sigma^2|y \sim \text{Inv-}\chi^2(n-k, s^2),$$

ahol

$$s^2 = \frac{1}{n-k}(y - X\hat{\beta})^\top (y - X\hat{\beta}).$$

Az $Inv - \chi^2(\nu, s^2)$ eloszlás ugyanaz, mint az $Inv - Gamma(\nu/2, \nu s^2/2)$ eloszlás.

Észrevehető, hogy ennél az a priori eloszlásnál a β paraméterek Bayes-becslése (az a poszteriori várható érték) $\hat{\beta}$, mely megegyezik a klasszikus lineáris regressziónál levezetett becslésekkel. Megjegyezzük, hogy a kapott a poszteriori eloszlás valódi eloszlás lesz (tehát véges az integrálja), amennyiben $n > k$, és az X mátrix rangja k.

Milyen más a priori választás jön még szóba? Itt már jóval nehezebb egy konjugált a priori - a poszteriori eloszláspárt találni a (β, σ^2) paraméterekre, ezért csak az eredményeket foglaljuk össze. Érdekes módon éppen az előbb kapott a poszteriori eloszlásra hasonló a priori lesz a jó választás:

$$p(\beta, \sigma^2) = p(\sigma^2)p(\beta|\sigma^2),$$

ahol σ^2 a priori eloszlása inverz-χ^2, és β feltételes a priori eloszlása σ^2-re többdimenziós normális. Ekkor az a poszteriori eloszlás is pontosan ilyen szorzat alakú, csak az inverz-χ^2, illetve a normális eloszlás paraméterei a minta alapján módosulnak.

Természetesen más a priori eloszlások is szóba jönnek, az R-ben írt rstanarm csomagban választhatunk például t, Cauchy, vagy exponenciális eloszlást is.

6.1. Regresszió célja

A regresszió célja Bayesi esetben az lesz, hogy egy adott paraméter eloszlását megtaláljuk a megfigyelt minta esetében. Itt fontos a paraméterek értéktartományának meghatározása, hogy biztosan a megfigyeléseinket kapjuk.

A következtetés alapkoncepciója a statisztikai következtetéshez hasonló elvekből indul ki. A statisztikai következtetés során a következtetések bizonytalanságának mértéke a valószínűség számszerűsítésével határozható meg. Lényeg megértésének érdekében tekintsük a következő példát: a klasszikus statisztikában nem lenne értelme közvetlen módon hozzárendelni egy eseményhez annak valószínűségét, ha tudjuk, hogy az az esemény csak egyetlen egyszer fordulhat elő (például egyetlen érmefeldobás eredménye).
Ugyanakkor azt érdemes megjegyezni, hogy minél többször ismételjük meg az érmefeldobást, az eredményeink annál jobban fognak az 50%-os (fele-fele) arányhoz közelíteni. Ez a bayesi statisztika fő szála, ez adja a módszer létjogosultságát.

A statisztikai modellek általánosan meghatároznak néhány statisztikai feltételezést és folyamatot, amelyek a minta adatainak generálásának módjából adódnak. Ezen statisztikai modelleknek számos paramétere van, amelyek módosíthatóak. Az érmefeldobós példát tekintve: az kísérlet kimenete felfogható egy Bernoulli-eloszlást követő mintaként, amely két lehetséges kimenetet modellez. A Bernoulli-eloszlásnak egyetlen paramétere van, amely ebben az esetben megfeleltethető az egyik lehetséges kimenetel valószínűségének. A Bayesi következtetés során az adatok elemzéséhez kulcsfontosságú egy megfelelő modell kidolgozása. Az esetek többségében a modellek csak megközelítik a valós folyamatot, de nem vesznek figyelembe bizonyos tényezőket, amelyek befolyásolhatják az adatokat. A Bayesi következtetésekben a valószínűség hozzárendelhető adott modellek paramétereihez, a paraméterek pedig valószínűségi változóként kifejezhetőek. A Bayesi következtetések új bizonyítékok begyűjtését követően a Bayes-tétel segítségével frissítik/aktualizálják a valószínűségi mutatókat, ezzel segítve a pontosabb eredményeket.

6.2. Szabályozás, méretek csökkentése

Akárcsak a klasszikus regresszióban, a bayesi esetben is fontos, hogy a magyarázó változókat jól válasszuk meg.

Ez a gyakorlatban azt jelenti, hogy a nem megfigyelt változók együtthatóit nullává tehetjük. Valamint az előzetes információinkat is bővíthetjük. Sok magyarázó változó esetén, ha ezeknek a valószínűsége irreleváns, akkor a prior eloszlását nulla középpontú t-eloszlásnak vehetjük. Ez alapján modhatjuk azt, hogy a változó most nem fontos, de van egy előzetes információ benne, ami viszont fontos.

Emellett fontos az adatok megfelelő alakba történő átírása. Például gyakran szükséges lehet az adatok átskálázása, hogy megfelelő nagyságrendűek legyenek. A maximum likelihood vagy legkisebb négyzetek becslésénél, vagy ha lineáris regressziót hajtunk végre neminformatív egyenletes a priori eloszlással a β_j együtthatók lineáris transzformációi nincsenek hatással a következtetésekre. Bayes-i környezetben azonban transzformációk - akár lineáris transzformációk is lineáris modellben - fontosak lehetnek.

Három fontos szempont játszik szerepet a Bayesi szabályozásban:

-Az eloszlás helye és nagysága:
Egy koncentráltabb prior mellett a regularizáció fontosabb mint a kevésbé koncentrált esetben.

-Az eloszlás analitikus formája:
A normális eloszlás a becsléseket növelheti, például egy Cauchy eloszlás jobban regularizál az átlag közelében lévő becslésekben és kevésbé a messze lévő becslésekben.

-Hogyan foglaljuk össze a poszteriori következtetést:
Egy olyan poszterior eloszlásnál, amely néhány variációt elhagy, simábban (vagyis jobban szabályozva) nézhet ki, a függvény, mint a teljes poszteriort tekintenénk, így a számunkra érdekesebb információk jobban nyomon követhetővé válnak.

IV. rész
Megközelítések összehasonlítása

Ha általánosságban tekintünk a statisztikára, akkor láthatjuk, hogy nagyon nagy jelentősége van az elégséges statisztika fogalmának. Akkor nevezünk egy statisztikát elégségesnek egy paraméterre, hogyha mindent amit megtudhatunk a paraméterről, azt a statisztikából valóban meg is tudjuk.
Ehhez szorosan kapcsolódik a minták egyik legfontosabb tulajdonsága, a likelihood függvénye, amelyre a következő elvet ismerjük:

Ha két kísérletnek ugyanaz a likelihood függvénye (pozitív konstans szorzótól eltekintve), akkor ugyanazt az evidenciát kell szolgáltatniuk, vagyis ugyanabból a statisztikából vannak.

A következő példa azt mutatja be, hogy a klasszikus statisztikában ez hol sérül.

-Példa a likelihood elv megsértésére

Tegyük fel, hogy független, $0-1$ eseményeket n-szer megfigyelünk. Legyen ez az első kísérlet. Ilyenkor az eloszlás binomiális lesz.

$$P(k|p,n) = \binom{n}{k} p^k (1-p)^{n-k}$$

A második kísérletben azonban megállunk, miután k-szor bekövetkezett az 1-es esemény.
Így ennek az eloszlása negatív binomiális lesz.

$$P(n|p,k) = \binom{n-1}{k-1} p^k (1-p)^{n-k}$$

Ha az első kísérletben k db egyes van, és a második kísérletben pedig n db kísérlet kellett a k db egyeshez, akkor a két kísérlet likelihood függvénye konstans szorzótól eltekintve meg fog egyezni. De mivel a valószínűségek nem egyeznek, ezért az ezekből adódó tesztek vagy konfidencia intervallumok mások lehetnek.

A következő táblázatban összefoglalva is látható a lényegi eltérés a két módszer között.

Tulajdonság	**Klasszikus statisztika**	**Bayesi statisztika**
Paraméter	Rögzített, $P(X\|\theta)$ X függvénye,és θ fix	V.v. $P(X\|\theta)$ θ függvénye, és X fix
Valószínűség	Objektív	Akár szubjektív is lehet
Prior információ	Nincs, vagy csak kevés	Van, és döntően fontos
Minta	Egy van, de többet feltételezünk	Egy van, és csak azt értékeljük
Intervallum	Konfidenciaintervallum	Credible intervallum

Természetesen ez a táblázat még igényel némi magyarázatot.
Az első, és leginkább szembetűnő különbség, hogy míg a klasszikus statisztika rögzített paraméterekkel dolgozik, addig a bayesi valószínűségi változókként tekint rájuk ami lehetővé teszi

több információ bevonását is.

A második sorban mutatkozik meg, hogy a bayesi statisztika sokkal megengedőbb a valószínűségekkel szemben, és jobban figyelembe tudja venni a szubjektív értékeket is. Ezt abból láthatjuk, hogy MI határozzuk meg, milyen a priori eloszlásfüggvényt használjunk. Erre a későbbiekben még ki fogok térni, hogy mikor milyen alakot érdemes használni
Nagyon fontos emellett, hogy itt mutatkozik meg, hogy míg a klasszikus statisztika fix X értékekhez, vagyis a minta elemeihez keres θ paraméter értéket, ami egy intervallumban lesz, addig a Bayesi pont fordítva gondolkodik, azaz fix θ paraméter mellett hogyan alakul az X minta bekövetkezésének a valószínűsége.

A harmadik sorban azt láthatjuk, hogy szintén vannak különbségek, mert míg az első esetben a külső információk nincsenek, és a modellbe nehezen kódolhatóak, addig Bayesi megközelítésben egyéb, a mintától független külső tényezőket is figyelembe tudunk venni az eloszlás meghatározásakor. Pontosan emiatt gondolják sokan szubjektívnek a Bayesi statisztikát, hiszen általunk előre közölt információkat használunk az előrejelzésben, de ha jobban belegondolunk, a valóságban is nagyon ritka az, hogy semmilyen előzetes információval nem rendelkezünk ami elősegíthetné a feladatunkat.

A következő sora a táblázatnak talán a legszemléletesebb bemutatása az eltérésnek. Az általános esetben ugyanis csak egy mintánk van, de úgy tekintünk rá, hogy ez a minta jól reprezentálja a teljes halmazt, és bármikor, akárhányszor megismételve ugyanezt az eloszlást kapnánk. Ez természetesen ellentmond az általános tapasztalatokkal, hiszen a minták létrehozása mindig erőforrásokat igényel.
A Bayesi esetben viszont szintén egy minta van, de csak azt az egyet fogjuk értékelni, és nem feltételezzük a megismételhetőségét. Csak annyit állítunk, hogy erre az egyedi mintára adott prior mellett milyen a poszterior eloszlást kapunk, de ez egyedül nem reprezentálja a teljes egészet, mivel ha más mintával számolok, akkor más eredményt fogok kapni.

Az utolsó sorban láthatjuk, hogy mit kapunk az egyes módszerek végén. Klasszikus esetben, egy konfidencia intervallumot. Ez szó szerinti értelmezésben annyit jelent, hogy bármi is a paraméter értéke, a mintából kiszámolt intervallum adott eséllyel tartalmazni fogja a paramétert. Ellentétben a Bayesi esettel, ahol egy olyan intervallumot kapok, amiben valahány % eséllyel ténylegesen benne van a paraméter. Persze ehhez kell egy a prior eloszlás, vagyis előzetes információkat kell felhasználni.

V. rész
Gyakorlati példák, és R-es megoldások

Ebben a fejezetben két feladaton keresztül fogom az alkalmazásokat bemutatni. Az egyik egyváltozós, diszkrét eset, a másik pedig 3 változós, folytonos eset. Mind a két feladat két részből fog állni, az első a klasszikus, a második pedig a bayesi irány. Így láthatóvá válnak a különbségek, amit az összefoglalásban fogok taglalni.

A megoldáshoz az R programot fogom használni, ami a statisztikában az egyik leggyakrabban használt programozási nyelv. Sokrétű lehetőségeket tartalmaz, amelyeket különböző csomagokban lehet hozzá letölteni, és számtalan matematikai problémára megoldást kínálnak. Számomra most az rstanarm csomag az egyik legfontosabb. Ez a csomag kimondottan a Bayes típusú problémakör megoldására szolgál, beépített függvénykkel, amelyek lehetővé teszik a szemléletes bemutatást is. A konkrét feladatok kódjai a függelékben találhatóak, az adatokat tartalmazó táblázattal együtt.

7. I.Feladat

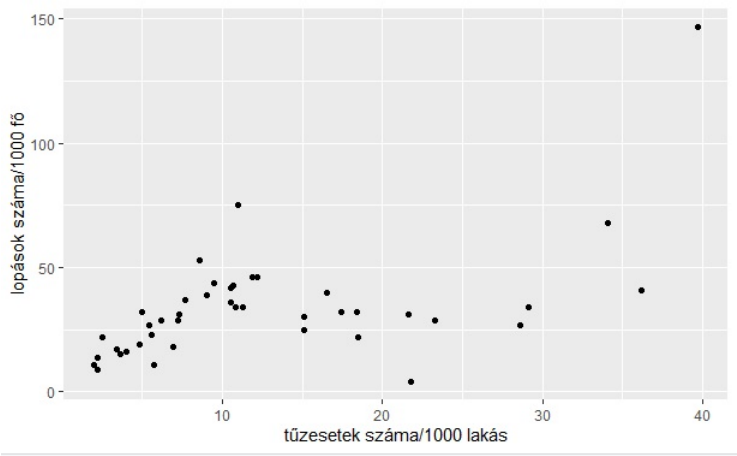
A mintánk Chicago államból van, és az adatok struktúrája az állam irányítószámonkénti felosztását tartalmazza. Az első oszlopban az átlagos tűzesetek számát látjuk 1000 lakásonként, a másodikban pedig az átlagos lopások számát látjuk 1000 főre. Megvizsgálom klasszikus, és Bayesi módszerekkel, hogy milyen összefüggés van a két adat között, tehát hogy hogyan függ össze a tűzesetek száma, és a lopások száma.

7.1. Klasszikus megközelítés

Ebben az esetben feltételezzük, hogy az adataink között lineáris összefüggés van.

A megoldás lépései:
Adatok vizsgálata
Modell felírása
Lineáris modell illesztése
Eredmények értékelése

Az adatainkat a következő grafikonnal szemléltethetjük:



Ebből a grafikonból az olvasható ki első ránézésre, hogy az adatok lineárishoz hasonlóan növekednek amíg a tűzesetek száma nem nő túl nagyra. Amikor viszont a tűzesetek száma megnő, akkor a lopások száma már nem lineárisan fog növekedni, sőt elég nagy visszaesés van. Egy kiugró adatot látunk, ahol a tűzesetek, és a lopások száma is nagyon magas.

A lineáris regressziót az R beépített függvényeként találjuk meg, lm, vagy glm paranccsal, majd a summary utasítással kapunk bővebb információt.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-11.251   -5.292   -3.882    4.122   21.811

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.90749    2.24983   2.181 0.035101 *
Y            0.23125    0.05536   4.177 0.000155 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 66.7027)

    Null deviance: 3832.0  on 41  degrees of freedom
Residual deviance: 2668.1  on 40  degrees of freedom
AIC: 299.55

Number of Fisher Scoring iterations: 2
```

Ez alapján azt láthatjuk, hogy az Y együtthatója szignifikáns, tehát elég jól magyarázza az X változót.

7.2. Bayesi megközelítés

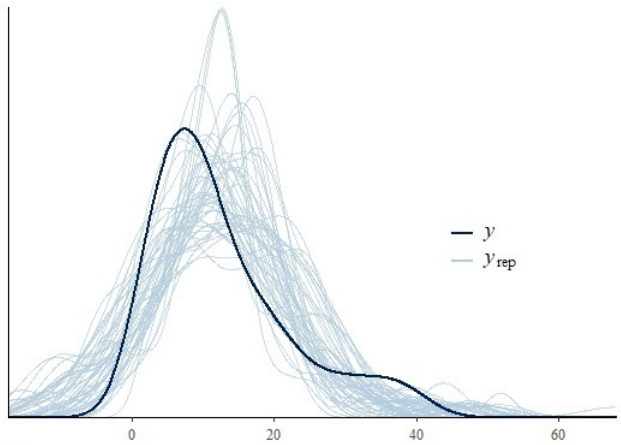
7.2.1. Alapértelmezett priorokkal

Most a bayesi statisztikával vizsgáljuk meg az adatokat. Ehhez az rstanarm csomagon kívül a bayesplot, a ggplot2, és readxl csomagokra is szükség van.

A megoldás lépései:
Adatok vizsgálata
Modell felépítése
Bayes-i modell illesztése
Priorok ellenőrzése
Eredmények értékelése

Az adatok továbbra is ugyanazok mint az előző részben, tehát futtathatjuk a bayes-i regressziót, először alapértelmezett priorral.
Az eredmények a következőképpen néznek ki grafikus ellenőrzéssel együtt:

```
Model Info:
 function:     stan_glm
 family:       gaussian [identity]
 formula:      X ~ Y
 algorithm:    sampling
 sample:       4000 (posterior sample size)
 priors:       see help('prior_summary')
 observations: 42
 predictors:   2

Estimates:
              mean   sd   10%   50%   90%
(Intercept) 4.9    2.3  2.0   5.0   7.9
Y           0.2    0.1  0.2   0.2   0.3
sigma       8.3    1.0  7.2   8.3   9.6

Fit Diagnostics:
           mean   sd   10%   50%   90%
mean_PPD 12.7    1.9 10.3  12.7  15.0

The mean_ppd is the sample average posterior predictive distribution
of the outcome variable (for details see help('summary.stanreg')).

MCMC diagnostics
              mcse Rhat n_eff
(Intercept)   0.0  1.0  3360
Y             0.0  1.0  3348
sigma         0.0  1.0  3106
mean_PPD      0.0  1.0  3623
log-posterior 0.0  1.0  1705
```



Paraméterbecslések pontos értéke

```
(Intercept)           Y
  4.9397363   0.2316333
```

```
Priors for model 'glm_bayXY'
------
Intercept (after predictors centered)
  Specified prior:
    ~ normal(location = 13, scale = 2.5)
  Adjusted prior:
    ~ normal(location = 13, scale = 24)

Coefficients
  Specified prior:
    ~ normal(location = 0, scale = 2.5)
  Adjusted prior:
    ~ normal(location = 0, scale = 1)

Auxiliary (sigma)
  Specified prior:
    ~ exponential(rate = 1)
  Adjusted prior:
    ~ exponential(rate = 0.1)
```

Ebből azt látjuk, hogy ha az illesztett modellből szimulálunk x-adatokat (az adott y-ok mellett), akkor a szimulált adatok nem követik jól az eredetileg megfigyelt adatokat.
A prior eloszlás eredetileg normális volt, és az R program automatikusan átskálázza, az adatok nagyságrendjének, szórásának megfelelően. Az ábrán a "Specified" prior az alapértelmezett, beépített változat, és az "Adjusted prior" pedig az átskálázott prior, amit valójában a program használ.
Hogyha szimulált adataink nem követik az eredetieket, akkor érdemes lehet próbálkozni más a priori eloszlásokkal is.

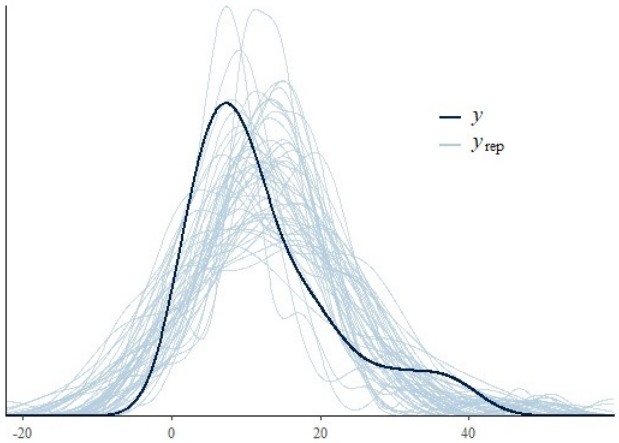
7.2.2. Lapos priorokkal

A konstans tagra, és az y együtthatójára nézve konstans lesz a prior eloszlás.
Ebből a következő adatokat kapjuk:

```
Model Info:
 function:     stan_glm
 family:       gaussian [identity]
 formula:      X ~ Y
 algorithm:    sampling
 sample:       4000 (posterior sample size)
 priors:       see help('prior_summary')
 observations: 42
 predictors:   2

Estimates:
              mean   sd   10%   50%   90%
(Intercept) 4.9    2.3  2.1   4.9   7.8
Y           0.2    0.1  0.2   0.2   0.3
sigma       8.3    1.0  7.2   8.3   9.5

Fit Diagnostics:
           mean   sd   10%   50%   90%
mean_PPD 12.7    1.9 10.3  12.7  15.0

The mean_ppd is the sample average posterior
predictive distribution of the outcome variable

(for details see help('summary.stanreg')).

MCMC diagnostics
              mcse Rhat n_eff
(Intercept)   0.0  1.0  3816
Y             0.0  1.0  3794
sigma         0.0  1.0  3133
mean_PPD      0.0  1.0  3634
log-posterior 0.0  1.0  1602
```

Paraméterbecslések pontos értéke

```
(Intercept)           Y
  4.8758613   0.2324736
```



Priorok ellenőrzése

```
Intercept (after predictors centered)
 ~ flat

Coefficients
 ~ flat

Auxiliary (sigma)
  Specified prior:
    ~ exponential(rate = 1)
  Adjusted prior:
    ~ exponential(rate = 0.1)
```

Az eredmények hasonlóak, de talán kicsit jobb képet mutat a modell diagnosztikája. Itt (csakúgy, mint az előző modellben) a hiba szórására a prior eloszlás exponenciális, és azt láthatjuk, hogy itt a világoskék görbék jobban közelítik a sötétkék görbét, tehát jobb közelítést kaptunk mint a korábbiakban.

8. II.Feladat

A második feladat háttere az, hogy 3 államban méréseket végeznek a radon légköri jelenlétére. Majd ezeket a méréseket kell adatokként felhasználni a regressziónál. A táblázatban meg van adva, hogy az adott államban az emeleten vagy pedig a földszinten végeztek mérést. Ezen adatokra fogom megnézni a klasszikus, és a bayesi statisztikát.

8.1. Klasszikus megközelítés

Itt egy többváltozós lineáris regressziót kell megoldani, ehhez a glm függvényt használhatjuk. Az eredmények a következők lesznek:

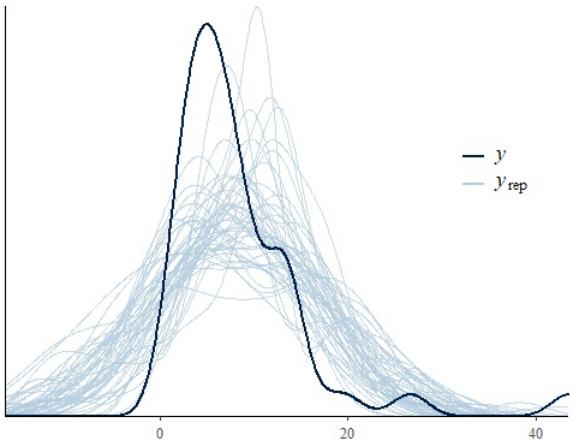
```
Call:
glm(formula = radon ~ emelet + Beallam + Callam, family = gaussian,
    data = radonadat)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-8.223  -4.597  -1.276   3.177  34.003

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.4968     2.2232   4.272  0.00013 ***
emelet       -2.9295     3.1270  -0.937  0.35491
Beallam      -1.4212     3.0145  -0.471  0.64008
Callam        0.2259     3.0424   0.074  0.94122
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 61.2462)

    Null deviance: 2335.8  on 40  degrees of freedom
Residual deviance: 2266.1  on 37  degrees of freedom
AIC: 290.86

Number of Fisher Scoring iterations: 2
```

Ebből azt láthatjuk, hogy a konstans tag az ami szignifikáns az adatainkra nézve, tehát jól magyarázza a függő változót. A többi magyarázó változó együtthatója azonban nem szignifikáns, tehát nem találunk szignifikáns különbséget aszerint, hogy emeleten mérünk-e vagy sem, illetve aszerint sem, hogy melyik államban történt a mérés.

8.2. Bayesi megközelítés

8.2.1. Alapértelmezett priorokkal

```
Model Info:
 function:     stan_glm
 family:       gaussian [identity]
 formula:      radon ~ emelet + Beallam + Callam
 algorithm:    sampling
 sample:       4000 (posterior sample size)
 priors:       see help('prior_summary')
 observations: 41
 predictors:   4

Estimates:
              mean   sd   10%   50%   90%
(Intercept)  9.5    2.3  6.7   9.4  12.4
emelet      -2.9    3.2 -7.0  -2.9   1.2
Beallam     -1.5    3.0 -5.3  -1.5   2.4
Callam       0.3    3.1 -3.7   0.3   4.2
sigma        8.0    1.0  6.8   7.9   9.3

Fit Diagnostics:
            mean   sd   10%   50%   90%
mean_PPD  8.5    1.8  6.3   8.6  10.8

The mean_ppd is the sample average posterior predictive distribution
of the outcome variable   (for details see help('summary.stanreg')).

MCMC diagnostics
              mcse Rhat n_eff
(Intercept)   0.0  1.0  3487
emelet        0.1  1.0  3955
Beallam       0.1  1.0  3442
Callam        0.1  1.0  3236
sigma         0.0  1.0  3105
mean_PPD      0.0  1.0  3984
log-posterior 0.0  1.0  1595
```



Paraméterbecslések pontos értéke

```
(Intercept)      emelet     Beallam      Callam
  9.4493465  -2.8818192  -1.4569308   0.2951751
```

Priorok ellenőrzése

```
Priors for model 'glm_post1'
------
Intercept (after predictors centered)
  Specified prior:
    ~ normal(location = 8.5, scale = 2.5)
  Adjusted prior:
    ~ normal(location = 8.5, scale = 19)

Coefficients
  Specified prior:
    ~ normal(location = [0,0,0], scale = [2.5,2.5,2.5])
  Adjusted prior:
    ~ normal(location = [0,0,0], scale = [47.62,39.79,39.79])

Auxiliary (sigma)
  Specified prior:
    ~ exponential(rate = 1)
  Adjusted prior:
    ~ exponential(rate = 0.13)
```

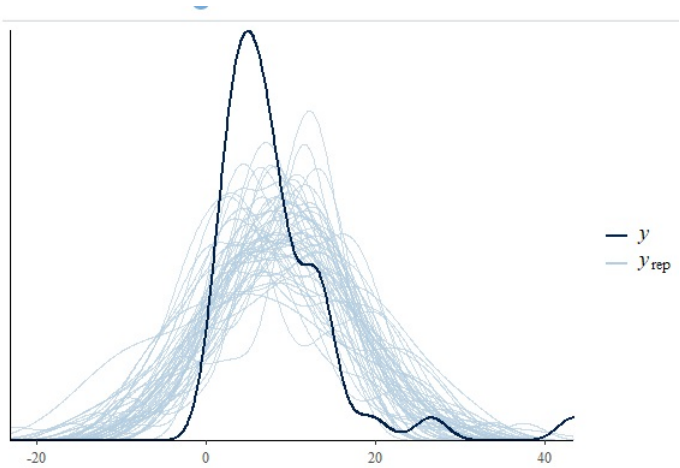
Itt azt láthatjuk, hogy a prior eloszlás normális volt, ezt használja a program is, csak átskálázva. A grafikon alapján nem követi igazán jól a világoskék a sötétkéket, de van néhány olyan görbe, ami egészen hasonló hozzá, tehát alapvetően nem rossz a megközelítés. De azért itt is érdemes lehet megnézni, hogy egy másik prior eloszlással milyen eredményeket kapunk.

8.2.2. Lapos priorokkal

```
Model Info:
 function:     stan_glm
 family:       gaussian [identity]
 formula:      radon ~ emelet + Beallam + Callam
 algorithm:    sampling
 sample:       4000 (posterior sample size)
 priors:       see help('prior_summary')
 observations: 41
 predictors:   4

Estimates:
              mean   sd   10%   50%   90%
(Intercept)  9.5    2.3  6.5   9.5  12.4
emelet      -3.0    3.2 -7.1  -2.9   1.2
Beallam     -1.4    3.1 -5.3  -1.4   2.5
Callam       0.3    3.2 -3.8   0.3   4.4
sigma        8.0    1.0  6.9   7.9   9.3

Fit Diagnostics:
           mean   sd   10%   50%   90%
mean_PPD  8.5    1.9  6.1   8.5  10.8

The mean_ppd is the sample average posterior predictive distribution of
      the outcome variable (for details see help('summary.stanreg')).

MCMC diagnostics
               mcse Rhat n_eff
(Intercept)    0.0  1.0  3191
emelet         0.0  1.0  4310
Beallam        0.1  1.0  3540
Callam         0.1  1.0  2954
sigma          0.0  1.0  3690
mean_PPD       0.0  1.0  3839
log-posterior  0.0  1.0  1712
```

Paraméterbecslések értéke:

```
(Intercept)      emelet     Beallam      Callam
  9.4668357  -2.9280907  -1.3537448   0.2646434
```



Prior ellenőrzése:

```
Priors for model 'glm_post2'
------
Intercept (after predictors centered)
 ~ flat

Coefficients
 ~ flat

Auxiliary (sigma)
  Specified prior:
    ~ exponential(rate = 1)
  Adjusted prior:
    ~ exponential(rate = 0.13)
```

Itt konstans priorral dolgoztunk, és azt láthatjuk, hogy közel sem követi annyira a világoskék a sötétet, mint az előző felírásban. Most σ-ra exponenciális a priori eloszlást vettünk, és a program is ezzel számol. De mivel ez kevésbé jól magyarázza, ezért az első verziót tekinthetjük jobbnak.

VI. rész
Összefoglalás, eredmények kiértékelése

Összességében tekintve elmondható, hogy a klasszikus, és a Bayesi megközelítés sok esetben nagyon hasonló eredményeket hoz ki. De nem elhanyagolható az a különbség, ami a két módszer között van. Pont ez az, amiért mind a két módszernek van létjogosultsága a statisztikában, és vannak olyan esetek, ahol a Bayesi gondolkodásmód jobban leírja a minket körülvevő világot. Sajnos azonban gyakran nem lehet az aposzteriori eloszlást meghatározni. Ennek a kiküszöbölésére használhatjuk a Monte-Carlo-módszert.
Monte-Carlo-módszerek összefoglaló névvel illetünk számos eljárást, technikát, melyek közös jellemzője, hogy véletlenszám-sorozatok generálásán alapulnak. A módszerek népszerűségének oka rendkívül egyszerű: analitikusan követhetetlen feladatok eredményeit vagyunk képesek tetszőleges közelítéssel meghatározni velük. A robbanásszerű elterjedéshez a matematikai alapok lefektetésén kívül szükség volt egy másik összetevőre, a véletlen értékeket generáló, a számításokat gyorsan elvégző számítógépekre.
A Monte-Carlo-módszer egy olyan sztochasztikus szimulációs módszer, amely számítástechnikai eszközök segítségével előállítja egy adott kísérlet végeredményét, ezek után az eredményként kapott numerikus jellemzőket feljegyzik és kiértékelik. Az eredmény hibájának meghatározása szórás kiszámításával történik. Az álvéletlenszámokat, melyek a kísérletekben szereplő valószínűségi változók értékei, számítógép állítja elő. Több programnyelv is tartalmaz ilyen álvéletlenszám-generátort, például a C programnyelv. Hasonló véletlenszámokat lehetne generálni a kaszinók kedvelt játékával, a rulettel is. Ezért nevezte el a módszer kifejlesztője, Neumann János „Monte-Carlo”-módszernek. Felhasználási területe mára már majdnem minden természettudományos diszciplínára kiterjedt. Ennek a módszernek az alapja is a Bayesi következtetési rendszer, a valószínűségi változók, és a szórások.

Ez is jól mutatja, hogy Bayes tétele milyen nagyon sokrétű, és sok helyen használható. Számomra nagyon izgalmas volt, hogy ezzel a gondolkodásmóddal megismerkedhettem, mivel a korábbi tanulmányaim folyamán nem kerültek elő ilyen mélységben a Bayesi módszerek. De ez is csak azt mutatja, hogy a matematikában mindig lesznek olyan területek, amik újabb és újabb lehetőségeket hoznak be a képbe, és minél komolyabb számítógépek állnak rendelkezésre ezek megoldásaihoz, annál több lehetőség van arra, hogy eddig megoldatlan problémákat is megfejthessünk.

Függelék

Az első feladathoz tartozó táblázat:

Tűzesetek száma 1000 lakásonként	Lopások száma 100 főre vetítve	Tűzesetek száma 1000 lakásonként	Lopások száma 100 főre vetítve
6,2	29	18,4	32
9,5	44	36,2	41
10,5	36	39,7	147
7,7	37	18,5	22
8,6	53	23,3	29
34,1	68	12,2	46
11	75	5,6	23
6,9	18	21,8	4
7,3	31	21,6	31
15,1	25	9	39
29,1	34	3,6	15
2,2	14	5	32
5,7	11	28,6	27
2	11	17,4	32
2,5	22	11,3	34
4	16	3,4	17
5,4	27	11,9	46
2,2	9	10,5	42
7,2	29	10,7	43
15,1	30	10,8	34
16,5	40	4,8	19

Az első feladat R kódjai:

```
#I.feladat

adatok_tuz_lopas <- read_excel("   ")
View(adatok_tuz_lopas)
qplot(X,Y,data=adatok_tuz_lopas, xlab="tűzesetek száma/1000 lakás", ylab="lopások száma/1000 fő")

#lineáris regresszió
glm_linXY <- glm(X~Y, data=adatok_tuz_lopas, family=gaussian)
summary(glm_linXY)

#bayes regresszió deafult priorokkal
glm_bayXY<-stan_glm(X~Y, data=adatok_tuz_lopas, family=gaussian)
summary(glm_bayXY) #eredmények
pp_check(glm_bayXY) #grafikus ellenőrzés
prior_summary(glm_bayXY) #prior ellenőrzése
glm_bayXY$coefficients #paraméterbecslések pontos értéke

#másik bayes lapos a priori eloszlással a konstans tagra, és az Y-ra
glm_bay2XY <- stan_glm(X~Y, data=adatok_tuz_lopas, family=gaussian, prior=NULL, prior_intercept=NULL)
summary(glm_bay2XY) #eredmények
pp_check(glm_bay2XY) #grafikus ellenőrzés
prior_summary(glm_bay2XY) #prior ellenőrzése
glm_bay2XY$coefficients #paraméterbecslések pontos értéke
```

A második feladathoz tartozó táblázat

County	Radon koncentrációs eredmények földszint	1.emelet
Blue Earth	5.0, 13.0, 7.2, 6.8, 12.8, 9.5, 6.0, 3.8, 1.8, 6.9, 4.7, 9.5	14.3, 5.8
Clay	12.9, 2.6, 26.6, 1.5, 13.0, 8.8, 19.5, 9.0, 13.1, 3.6,	0.9, 3.5, 2.5, 6.9
Goodhue	14.3, 7.6, 2.6, 43.5, 4.9, 3.5, 4.8, 5.6, 3.5, 3.9, 6.7	6.9, 9.8

A második feladathoz tartozó R kódok

```
#II.feladat

radonadat <- read_excel("C:/Users/szabo/OneDrive/Asztali gép/egyetem/szakdoga/regresszió adatok rendezve.xlsx")
View(radonadat)


# klasszikus lineáris regresszió a feladatra:
glm_fit <- glm(radon~emelet+Beallam+Callam, data=radonadat, family=gaussian)
summary(glm_fit)

# Bayes-i regresszió a default a priori eloszlásokkal:
glm_post1 <- stan_glm(radon~emelet+Beallam+Callam, data=radonadat, family=gaussian) # regresszió futtatása
summary(glm_post1)  # eredmények kiírása
pp_check(glm_post1) # egy grafikus modell-ellenőrző eljárás (diagnosztika): a halvány kékeknek
# illik közel lenni a feketéhez
prior_summary(glm_post1) #prior ellenőrzése
glm_post1$coefficients

# egy másik beayes-i regresszió, itt a konstans tag és a sebesség együtthatójára is "lapos" (flat)
# a priori eloszlást állítottam be:
glm_post2 <- stan_glm(radon~emelet+Beallam+Callam, data=radonadat, family=gaussian, prior=NULL,
                      prior_intercept=NULL)
summary(glm_post2)
pp_check(glm_post2)
prior_summary(glm_post2) # itt lehet ellenőrizni az a priorikat (a szórásra maradt a default exponenciális)
glm_post2$coefficients # így lehet kinyerni az illesztésből a paraméterbecslések pontos értékét
```

Források

A Felhasznált szakirodalmak:

[1] A. Gelman, J.B.carlin, H.S.Stern, D.B.Dunson, A.Vehtari, D.B.Rubin-Bayesian Data Analysis -Third edition (feb.2020)

[2] Hunyadi László - Bayesi gondolkodás a ststisztikában - Statisztikai szemle 89.évfolyam. 10-11.szám (2011)

[3] Cameron Davidson-Pilon - Bayesian Methods for Hackers - Indiana (2015)

[4] Merlise Clyde, Mine Cetinkaya-Rundel, Colin Rundel, David Banks, Christine Chai, Lizzy Huang: An Introduction to Bayesian Thinking; Published with bookdown, 2019

[5] David Hitchcock, University of South Carolina Columbia: STAT 535 (Introduction to Bayesian Data Analysis), Spring 2014

[6] Kiss-Fülöp Zsófia: A Bayes-becslés és a bayesi hipotézisvizsgálat, ELTE Természettudományi Kar, Alkalmazott matematikus BSc, Szakdolgozat 2020

A feladatokhoz használt adatok:

[1] Tűzesetek és lopás:
$https://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/slr/frames/frame.html$
Fire and theft in Chicago

[2] Radon bomlása:
A. Gelman, J.B.carlin, H.S.Stern, D.B.Dunson, A.Vehtari, D.B.Rubin-Bayesian Data Analysis -Third edition (feb.2020)
Feladat:378.oldal 1.feladat
Táblázat:195.oldal "Radon measurement (pCi/L)"