Seminarska naloga

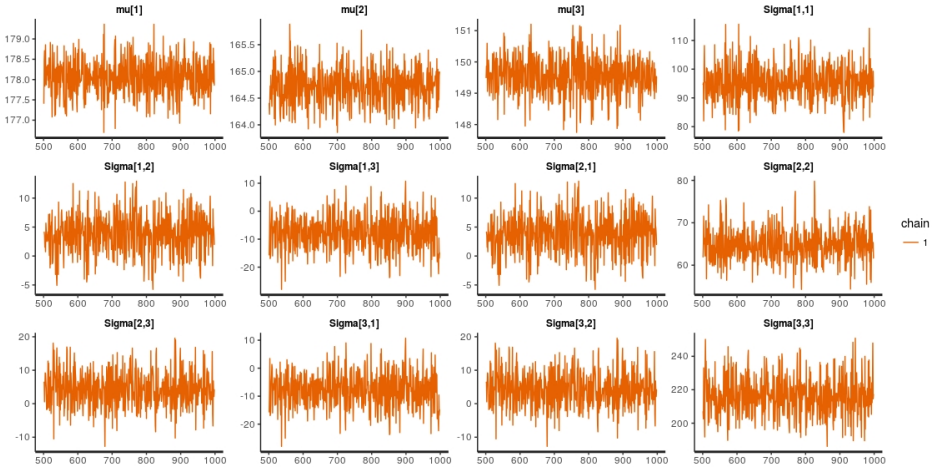
Bogdan Golobič

1 Multivariatna normalna

Za prvi statistični model sem izbral multivariatni normalni. V ta namen sem generiral podatke višin treh skupin: mož, žena in otrok (za vsako skupino 500 podatkov). Pri tem je višina možev porazdeljena normalno z upanjem 178 in standarnim odklonom 10, žene z upanjem 165 in standardnim odklonom 8 in otroci z upanjem 150 in standardnim odklonom 15.
Postavimo naslednji model:

- Podatki: $y_i = (y_{i,1}, y_{i,2}, ..., y_{i,k})$; $i \in \{1, 2, 3\}, k = 500$
- Verjetje: $Y_i|\mu, \Sigma \sim_{iid} N_k(\mu, \Sigma)$
- Apriorna: $\mu \sim N_k(\mu_0, \Lambda_0), \quad \Sigma \sim IW(\alpha_0, S_0)$

Za Stan kodo sem vzel iz predavanj, kjer smo gledali ocene študentov pri predmetih (Dodatki 3.1).
V naslednjem koraku moramo diagnosticirati konvergenco.

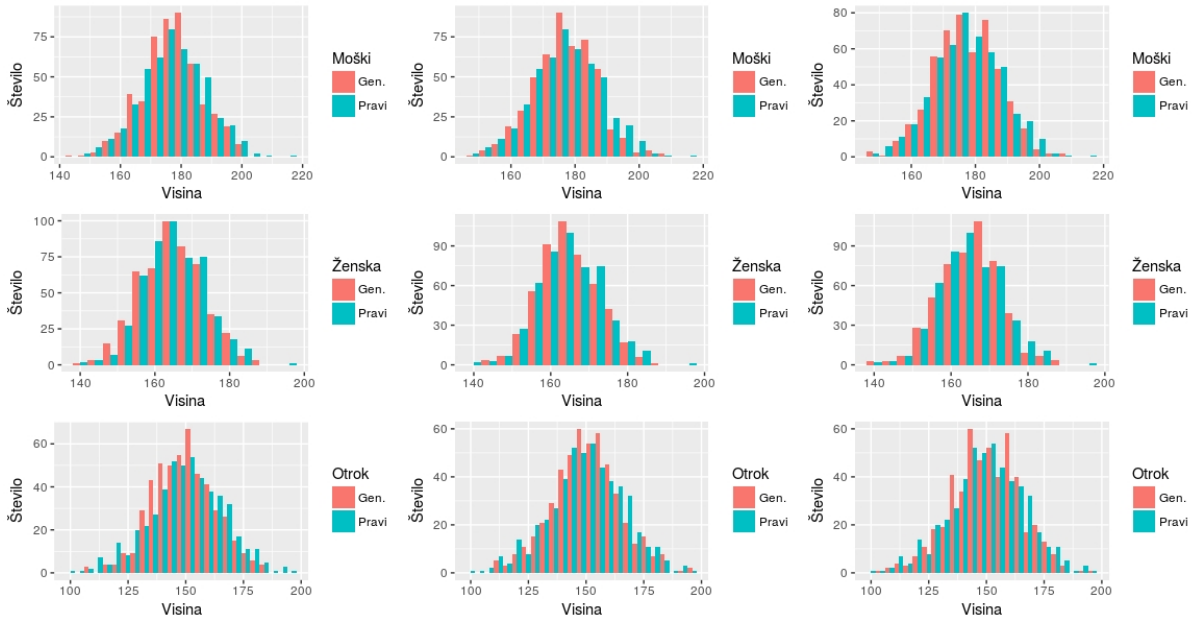


Slika 1: Aposteriorna mvn modela

Vidimo, da ni očitnih znakov, da ni konvergiralo. Zdaj lahko nadaljujemo z analizo. Za začetek pogledamo povzetek vzorčenja.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff
mu[1]	177.30	0.02	0.46	176.38	176.97	177.30	177.59	178.21	500
mu[2]	164.86	0.02	0.35	164.19	164.65	164.84	165.08	165.61	500
mu[3]	149.90	0.03	0.75	148.42	149.41	149.89	150.43	151.29	500
Sigma[1,1]	115.82	0.33	7.37	101.67	110.80	115.29	120.82	132.31	500
Sigma[1,2]	−5.35	0.16	3.57	−11.93	−7.93	−5.04	−2.83	1.54	500
Sigma[1,3]	11.11	0.34	7.59	−3.87	6.53	10.97	15.96	25.69	500
Sigma[2,1]	−5.35	0.16	3.57	−11.93	−7.93	−5.04	−2.83	1.54	500
Sigma[2,2]	65.03	0.19	4.15	57.58	62.10	64.78	68.08	73.45	500
Sigma[2,3]	−12.40	0.27	5.65	−23.50	−16.11	−12.61	−8.74	−1.84	439
Sigma[3,1]	11.11	0.34	7.59	−3.87	6.53	10.97	15.96	25.69	500
Sigma[3,2]	−12.40	0.27	5.65	−23.50	−16.11	−12.61	−8.74	−1.84	439
Sigma[3,3]	260.79	0.67	14.89	234.23	249.18	261.01	271.35	289.84	500
tmpD[1]	10.76	0.02	0.34	10.08	10.53	10.74	10.99	11.50	500
tmpD[2]	8.06	0.01	0.26	7.59	7.88	8.05	8.25	8.57	500
tmpD[3]	16.14	0.02	0.46	15.30	15.79	16.16	16.47	17.02	500

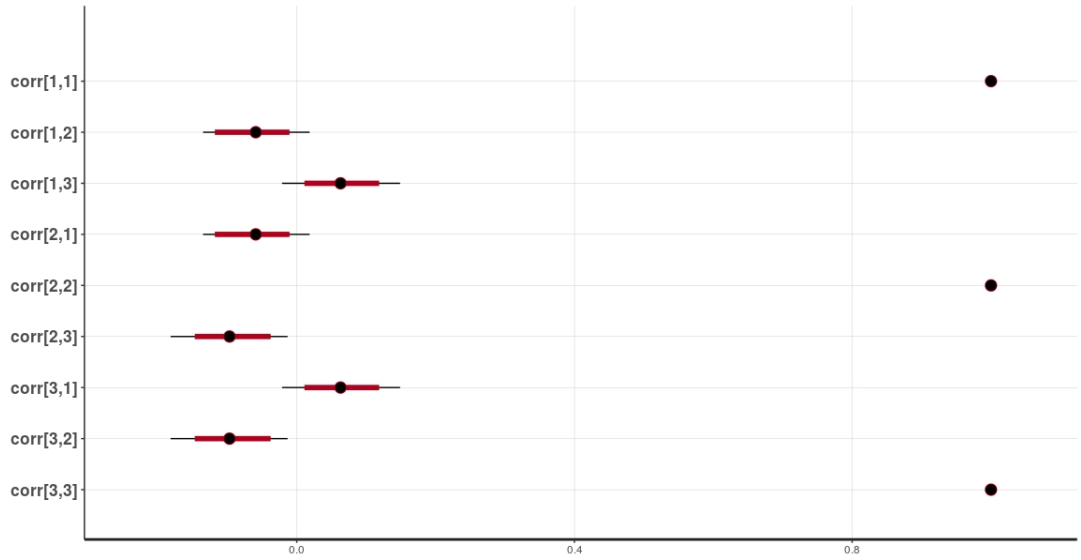
Vidimo, da so upanja (mu) in standardna deviacije (tmpD) zelo blizu tistim, ki smo jih uporabili za generiranje podatkov. V ta namen naredimo še vizualni predictive check.



Slika 2: Primerjava dobljenega modela s podanimi podatki

Kot pričakovano se podatki lepo prilegajo med sabo. To je logično, ker jemljemo podatke kar direktno iz normalne porazdelitve.

Pogledamo še, ali je kakšna korelacija med višinami moža, žene in otroka.



Slika 3: Korelacije med možem, ženo in otrokom. Tukaj moramo gledati le vrstice corr[1,2] (mož in žena), corr[1,3] (mož in otrok) in corr[2,3] (žena in otrok).

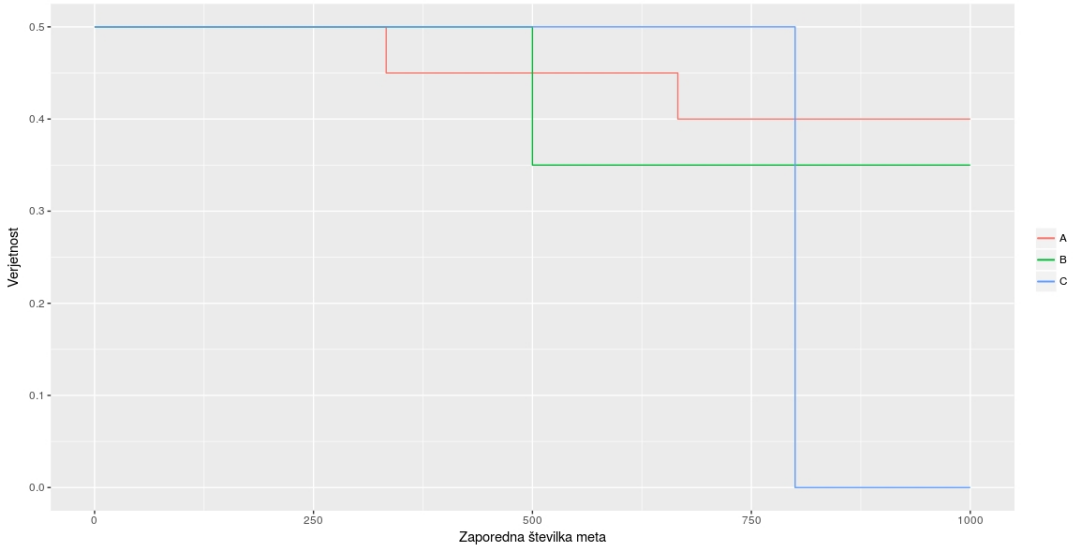
Povezave praktično ni (zelo blizu 0), ker so bili podatki naključno generirani (pri generiranju nisem zagotovil, da bi na primer imeli višji starši višjega otroka).

2 Logistična regresija

Za naslednji model sem izbral logistično regresijo. Za ta namen sem si izmislil situacijo treh metalcev kovancev. Vsak meče kovanec, in ob določenem času zamenja pravični kovanec za nepravičnega. Vprašanje je, kateremu bi naslednjo igro bolj zaupali.
Generirani podatki so naslednji (vsak meče kovanec 1000-krat):

- A metalec ima lažni kovanec z verjetnostjo, da pade čifra"(0) 80%: po 300 metih začne uporabljati lažni kovanec in po 150 metih zamenja nazaj za pravičnega. Zadnjih 150 metov pa uporabi spet lažnega.
- B metalec ima lažni kovanec z verjetnostjo, da pade čifra"(0) 60%: po 500 metih začne uporabljati lažni kovanec.
- C metalec ima lažni kovanec z verjetnostjo, da pade čifra"(0) 100%: Vsak četrti met uporabi lažen kovanec.

Za lažjo predstavo sem dodal še graf sprememb verjetnosti.



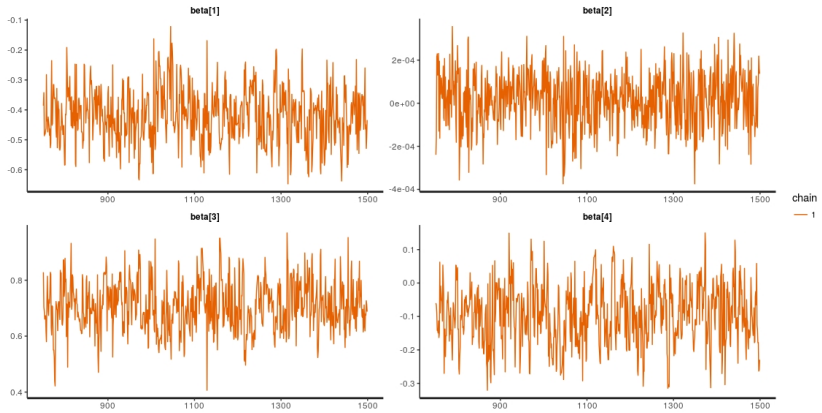
Slika 4: Spremembe verjetnosti, da pade glava pri metalcih A, B in C

Model je naslednji:

- Podatki: $y_i = (k, i, y_{i,k})$; $i \in \{1, 2, 3\}, k = 1000$
- Verjetje: $Y_i|\Theta \sim_{iid} Bernoulli(\Theta)$

Stan kodo sem povzel iz predavanj (Dodatki 3.2).

Pregledamo ali ta model divergira. Iz grafov je razvidno, da ni očitnih znakov, da model ne konvergira.

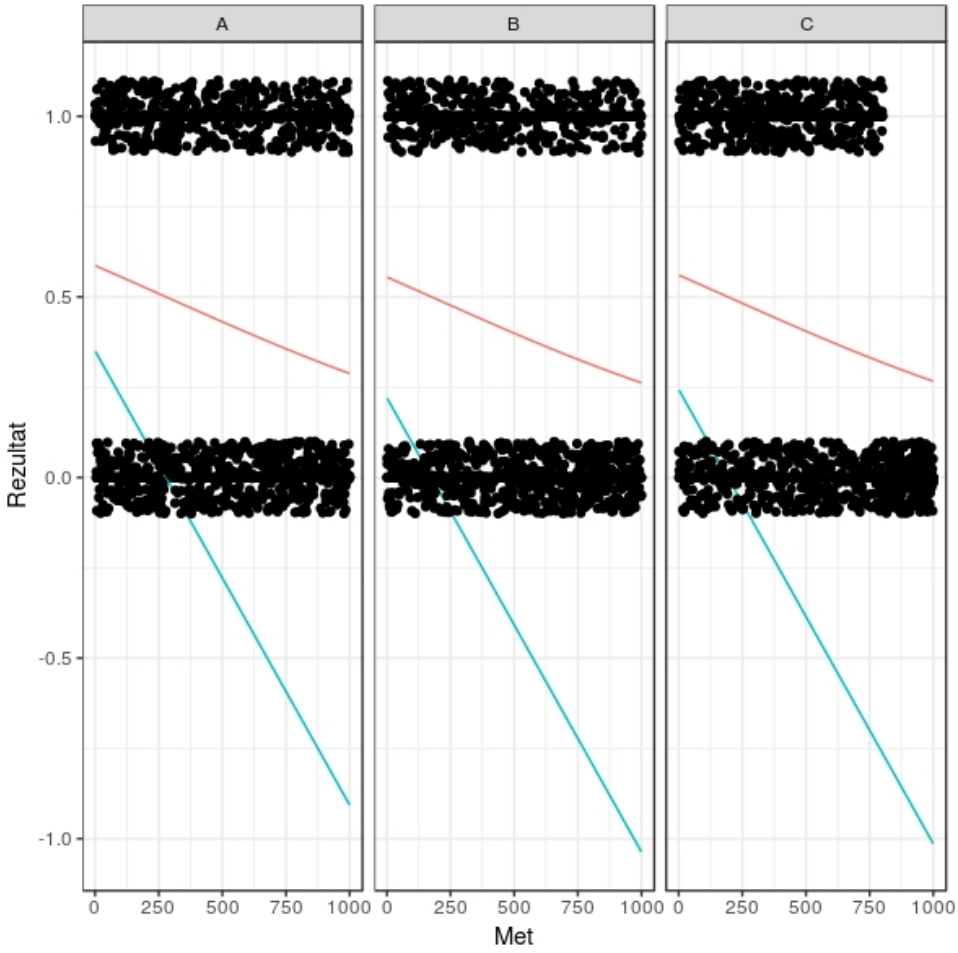


Slika 5: Konvergiranje modela logistične regresije

Pogledamo še povzetek rezultata vzorčenja.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff
beta[1]	0.35	0.00	0.09	0.18	0.29	0.35	0.40	0.53	298
beta[2]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	471
beta[3]	−0.13	0.00	0.08	−0.29	−0.19	−0.13	−0.07	0.03	377
beta[4]	−0.11	0.01	0.09	−0.29	−0.16	−0.11	−0.05	0.06	272

Pogledamo kateremu bi najbolj zaupali. Vidimo, da bi najbolj zaupal metalcu A. Enak rezultat bi lahko

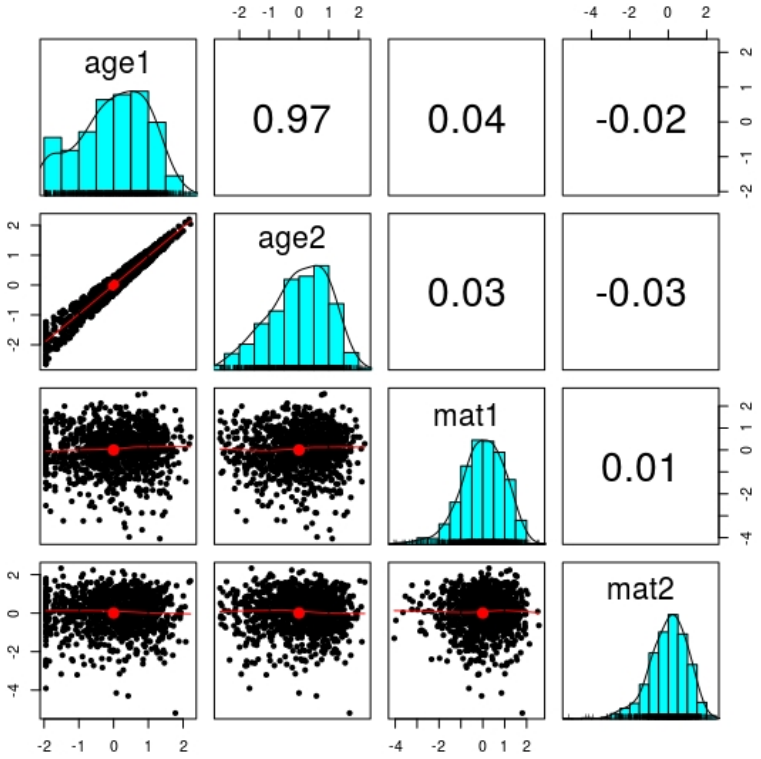


Slika 6: Napovedovanje meta kovancev s pomočjo logistične regresije

dobili, če bi izračunali verjetnosti iz genereranih podatkov.

3 Poissonova regresija

Za tretji model sem izbral Poissonovo regresijo.
Za ta namen sem generiral podatke števila otrok, kjer imamo podatke o starosti partnerja 1 in partnerja 2, ter podatke o oceni na maturi obeh partnerjev.
Starost prvega partnerja (age1) je porazdeljena normalno z upanjem 40 in standardno deviacijo 15 (pri tem sem še poskrbel, da starost ni manjša od 18). Starost drugega partnerja (age2) pa je porazdeljena zvezno enakomerno s spodnjo mejo -5 in 5, ta dobljena vrednost se prišteje starosti prvemu partnerju. Ocena mature (mat1, mat2) obeh partnerjev pa je porazdeljena normalno z upanjem 18 in standardno deviacijo 3.8. Te podatke sem zatem logaritmiral in normaliziral.
Za lažjo predstavo generiranih podatkov pomaga naslednji graf.

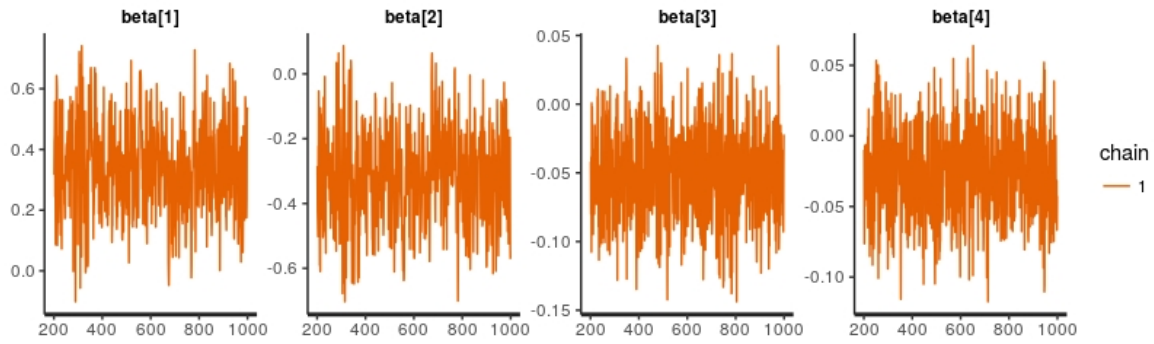


Slika 7: Prikaz generiranih podatkov in povezav med njimi

Model je naslednji:

- Podatki: $Y_i, x_i = (x1_i, x2_i, x3_i, x4_i)$; $i \in \{1, 2, ..., 1000\}$
- Verjetje: $Y_i|\alpha, \beta, x_i \sim_{iid} Poisson(e^{x_i * \alpha + \beta})$

Stan kodo sem povzel iz predavanj (Dodatki 3.3). Pogledamo ali model divergira. Vidimo, da ni očitnih

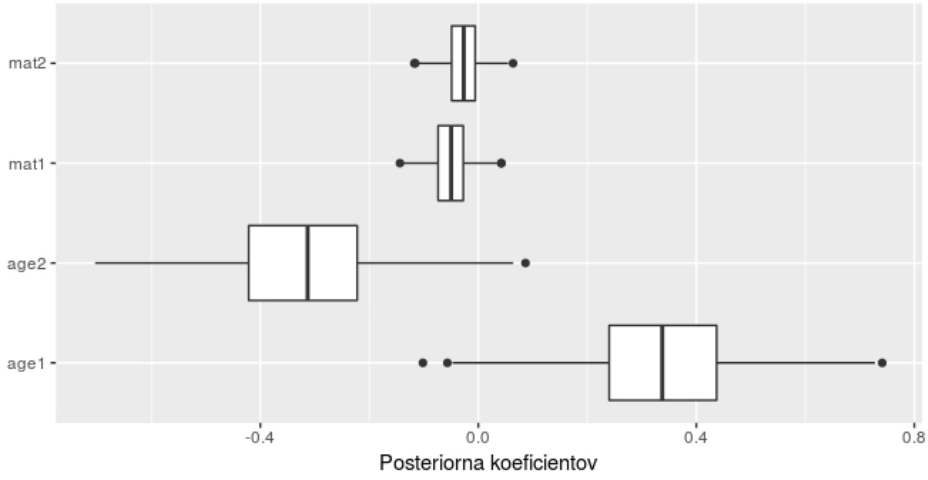


Slika 8: Konvergenca modela Poissonove regresije

znakov, da ni konvergiralo. Pogledamo še povzetek rezultata simulacije.

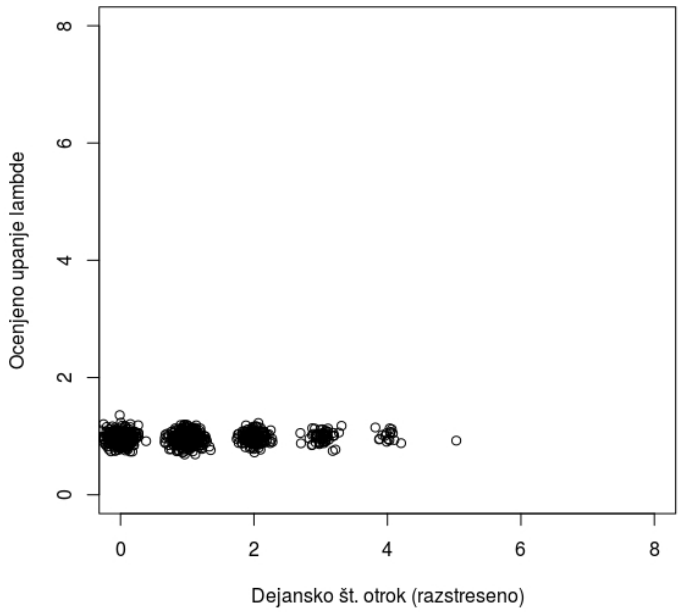
	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff
beta[1]	0.34	0.01	0.14	0.06	0.24	0.34	0.44	0.62	288
beta[2]	−0.32	0.01	0.14	−0.60	−0.42	−0.31	−0.22	−0.03	286
beta[3]	−0.05	0.00	0.03	−0.12	−0.07	−0.05	−0.03	0.01	800
beta[4]	−0.03	0.00	0.03	−0.08	−0.05	−0.03	−0.01	0.04	798
alpha	−0.04	0.00	0.03	−0.10	−0.06	−0.04	−0.02	0.02	800

Za lažjo predstavo sem dodal še graf porazdelitev bet.



Slika 9: Aposteriorna porazdelitev bet

Pogledamo še kako naš model napoveduje število otrok.



Slika 10: Napovedano število otrok proti dejanskim

Vedno napoveduje enega otroka. Razlog je pri naključnih generiranih podatkih. Primernejše generirani podatki bi bili na primer, da bi bolj izobražena in mlajša starša imela manjše število otrok.

4 Dodatki

4.1 MVN model

```
data {
  int<lower=1> d;
  int<lower=0> n;
  matrix[n,d] y;
}
parameters {
  cov_matrix[d] Sigma;
  vector[d] mu;
}
model {
  mu ~ multi_normal(rep_vector(0, d), diag_matrix(rep_vector(1000,d)));
```

```
  Sigma ~ inv_wishart(6, 2 * diag_matrix(rep_vector(1,d)));
  for (i in 1:n) {
    y[i] ~ multi_normal(mu, Sigma);
  }
}
generated quantities {
  corr_matrix[d] corr;
  matrix[d,d] tmp;
  vector[d] tmpD;

  tmpD = diagonal(Sigma);
  for (i in 1:d) tmpD[i] = sqrt(tmpD[i]);
  tmp = inverse(diag_matrix(tmpD));
  corr = tmp * Sigma * tmp;
}
```

4.2 Logistična regresija model

```
data {
  int<lower=0> n;
  int<lower=0> m;
  int<lower=0,upper=1> y[n];
  matrix[n,m] x;
}
parameters {
  vector[m] beta;
}
model {
  for (i in 1:n) {
    y[i] ~ bernoulli_logit(x[i]*beta);
  }
}
generated quantities {
  real theta[n];
  real eta[n];

  for (i in 1:n) {
    eta[i] = x[i]*beta;
    theta[i] = inv_logit(eta[i]);
  }
}
```

4.3 Poissonova regresija model

```
data {
```

```
  int<lower=0> n;
  int<lower=0> m;
  matrix[n,m] x;
  int<lower=0> y[n];
}

parameters {
  real alpha;
  vector[m] beta;
}

model {
  for (i in 1:n) {
    y[i] ~ poisson(exp(x[i] * beta + alpha));
  }
}

generated quantities {
  real<lower=0> lambda[n];
  int<lower=0> pred[n];
  for (i in 1:n) {
    lambda[i] <- exp(x[i] * beta + alpha);
    pred[i] <- poisson_rng(lambda[i]);
  }
}
```