Napovedovanje onesnaženosti zraka

Peter Us

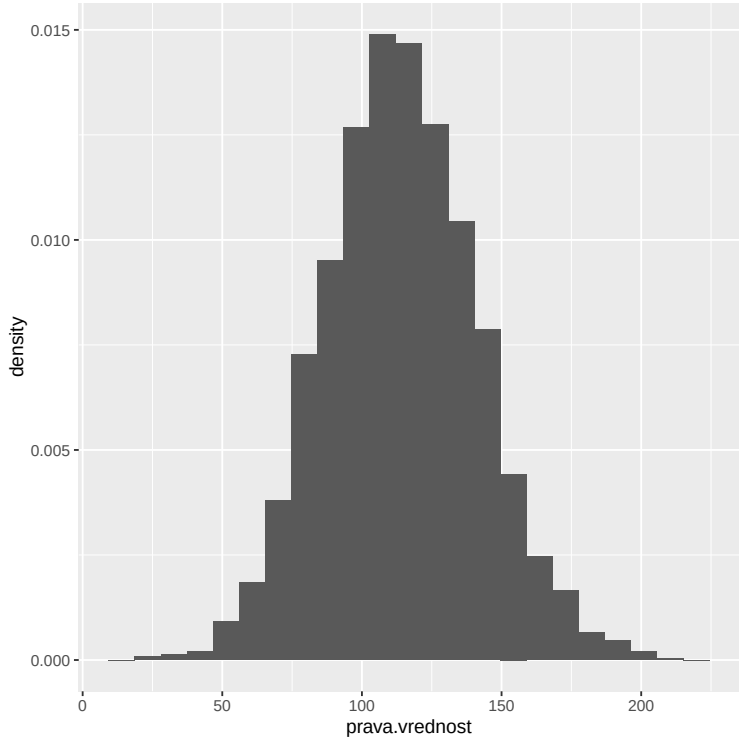
5. junij 2016

Oris vsebine

- Opis problema
- Predprocesiranje podatkov
- Gamma GLM
- Hiearhična linearna regresija

Opis problema

- Napovedovanje koncentracije ozona
  - Med leti 2011 in 2015
  - 8 meteoroloških postaj
  - 2 vrsti napovedi (za trenutni dan in naslednji)
- 8772 primerov
- 114 atributov
- mankajoče vrednosti
- časovni podatki




```
> summary(dataset$prava.vrednost)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  23.30   95.85  113.20  114.20  132.00  210.20      34
```

Predprocesiranje

Atributi z veliko mankajočimi vrednosti

```
> dat <- dat[, -which(colMeans(is.na(dat)) > 0.3)]
```

Visoko korelirani atributi

```
> library(caret)
> removeIdx <- findCorrelation(cor(dat_temp), cutoff = .90)
> dat <- dat[, -removeIdx]
```

Ostane

- 7862 primerov
- 60 atributov

Predprocesiranje

Mankajoče vrednosti, skaliranje, standardizacija

- naredimo v fazi učenja / testiranje modela
- uporabimo vrednosti naučene na testni mnozici
- Primer standardizacije:
  - (učnaMnožica - mean(učnaMnožica)) / sd(učnaMnožica)
  - (testnaMnožica - mean(učnaMnožica)) / sd(učnaMnožica)

Evalvacija modelov

- Podatke razdelimo po letih v 5 skupin
  - Za napovedovanje podatkov nekega leta uporabimo podatke iz vseh prejšnjih let
- Učimo in napovedujemo ločeno za vsako izmed 8 postaj
  - Hrastnik, Iškrba, Koper, Krvavec, Ljubljana, Murska Sobota, Nova Gorica, Otlica
- Dve vrsti napovedi: trenutni dan in naslednji dan

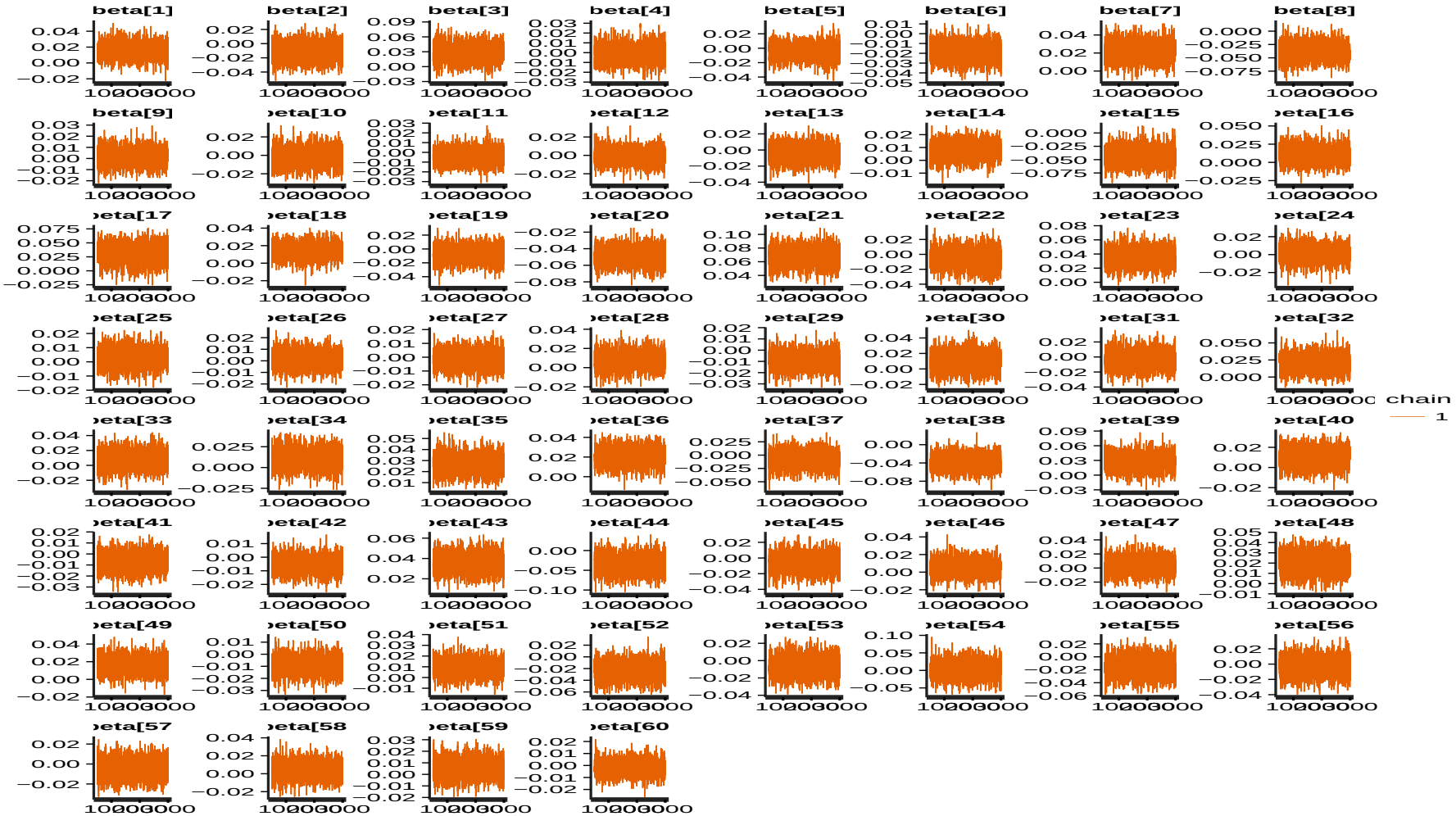
Gamma GLM

Model

```
data {
  int<lower=0> n; // number of samples
  int<lower=0> k; // number of attributes
  matrix[n, k] x; // samples
  vector[n] y; //target
  int<lower=0> n_new; // number of predicting samples
  matrix[n_new, k] x_new;
}

parameters {
  real alpha;
  vector[k] beta;
  real<lower=0.0001> shape;
}

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

generated quantities {
  vector[n_new] y_new;
    for(i in 1:n_new)
      y_new[i] <- gamma_rng(shape, shape / exp(x_new[i,] * beta + alpha));
}
```

Gamma GLM

Traceplot:

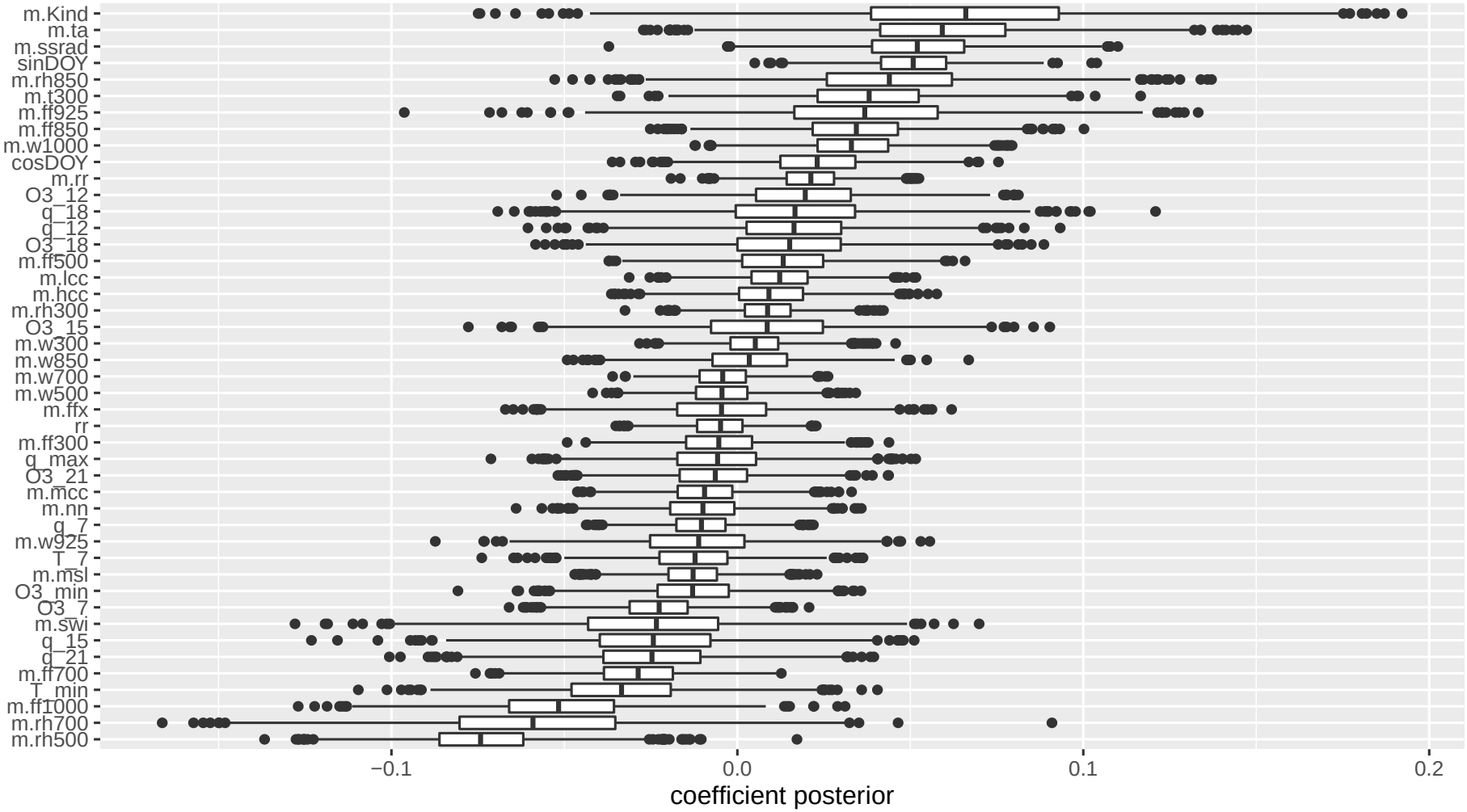


Gamma GLM

```
Inference for Stan model: gamma_reg_all.
1 chains, each with iter=3000; warmup=500; thin=1;
post-warmup draws per chain=2500, total post-warmup draws=2500.
```

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
alpha	4.80	0.00	0.00	4.79	4.79	4.80	4.80	4.80	2500	1
beta[1]	0.02	0.00	0.01	0.00	0.01	0.02	0.02	0.03	2316	1
beta[2]	−0.01	0.00	0.01	−0.03	−0.02	−0.01	0.00	0.01	1976	1
beta[3]	0.03	0.00	0.02	−0.01	0.02	0.03	0.04	0.06	1896	1
beta[4]	0.00	0.00	0.01	−0.02	−0.01	0.00	0.01	0.02	2500	1
beta[5]	0.00	0.00	0.01	−0.02	−0.01	0.00	0.00	0.01	2000	1
beta[6]	−0.02	0.00	0.01	−0.04	−0.03	−0.02	−0.01	0.00	2500	1
beta[7]	0.02	0.00	0.01	0.00	0.01	0.02	0.03	0.04	2500	1
beta[8]	−0.04	0.00	0.02	−0.07	−0.05	−0.04	−0.03	−0.01	1948	1
beta[9]	0.00	0.00	0.01	−0.01	0.00	0.00	0.01	0.02	2500	1
beta[10]	0.00	0.00	0.01	−0.02	−0.01	0.00	0.00	0.02	2117	1
...										
beta[56]	0.00	0.00	0.01	−0.03	−0.01	0.00	0.00	0.02	2500	1
beta[57]	0.00	0.00	0.01	−0.02	−0.01	0.00	0.00	0.01	2500	1
beta[58]	0.00	0.00	0.01	−0.01	0.00	0.01	0.01	0.02	2500	1
beta[59]	0.00	0.00	0.01	−0.01	0.00	0.00	0.01	0.02	2500	1
beta[60]	0.00	0.00	0.01	−0.01	−0.01	0.00	0.00	0.01	2500	1
shape	89.94	0.12	5.83	78.79	85.99	89.88	93.77	101.45	2500	1
lp__	−2033.51	0.22	6.04	−2046.54	−2037.39	−2033.29	−2029.24	−2022.30	766	1

Gamma GLM


m.Kind
m.ta
m.ssrad
sinDOY
m.rh850
m.t300
m.ff925
m.ff850
m.w1000
cosDOY
m.rr
O3_12
q_18
q_12
O3_18
m.ff500
m.lcc
m.hcc
m.rh300
O3_15
m.w300
m.w850
m.w700
m.w500
m.ffx
rr
m.ff300
q_max
O3_21
m.mcc
m.nn
q_7
m.w925
T_7
m.msl
O3_min
O3_7
m.swi
q_15
q_21
m.ff700
T_min
m.ff1000
m.rh700
m.rh500
−0.1
0.0
0.1
0.2
coefficient posterior

Gamma GLM

Rezultati napovedovanja

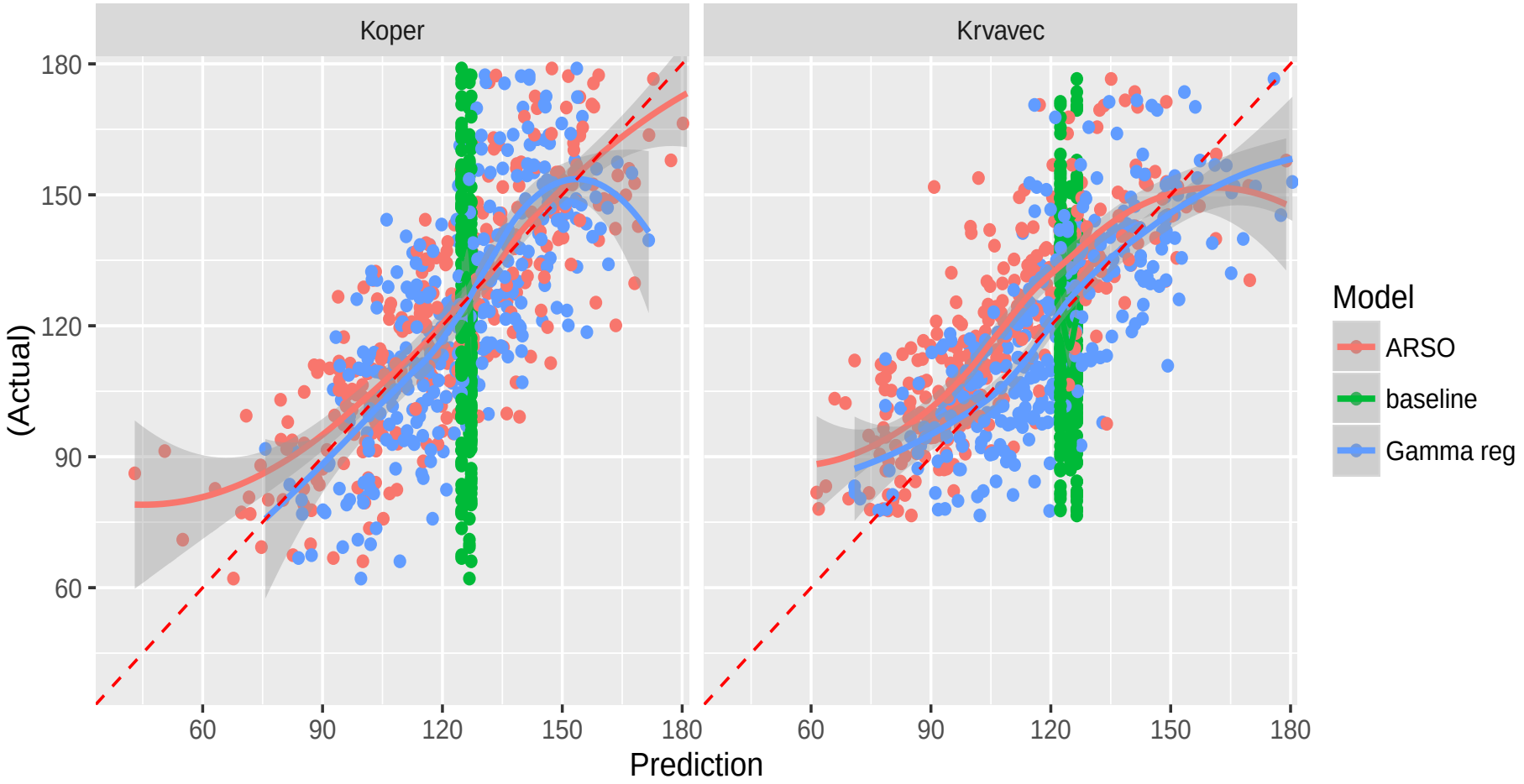
RMSE:

	ARSO	baseline	Lin Reg.	Gamma Reg.
Hrastnik	245.0967	709.8045	313.6543	311.8470
Iskrba	386.8695	661.4325	321.4349	311.4579
Koper	278.9845	773.2572	363.7615	360.3686
Krvavec	320.8344	550.3603	257.5804	237.2069
Ljubljana	333.8019	902.5365	347.4990	338.3196
MurskaSobota	224.0382	646.6000	288.9817	730.6269
NovaGorica	507.3541	1093.1313	381.1029	409.1273
Otlica	335.6851	783.4668	405.7707	379.6134

Standardne napake:

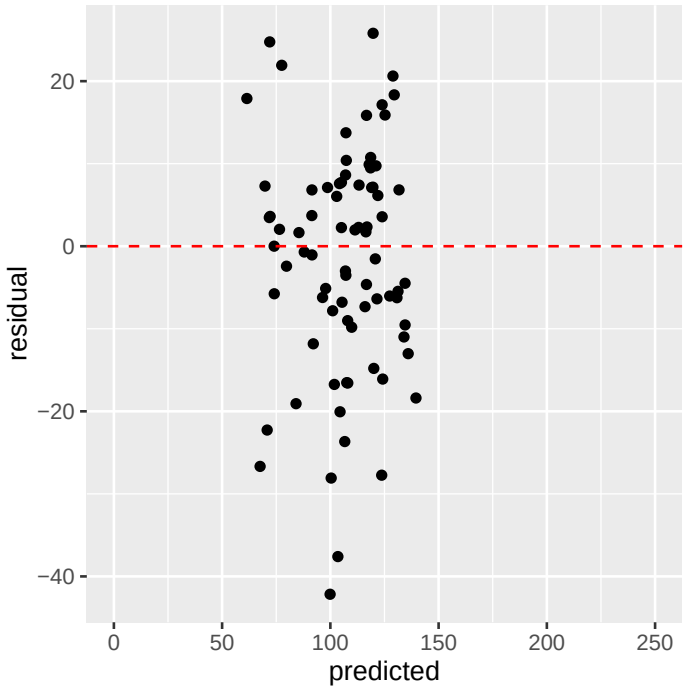
	ARSO	baseline	Lin Reg.	Gamma Reg.
Hrastnik	21.79971	45.65775	27.78950	26.88194
Iskrba	36.68279	55.29940	30.48019	30.04657
Koper	23.15402	58.12643	30.61805	30.88656
Krvavec	27.77972	36.28856	21.28812	20.14176
Ljubljana	36.41336	68.29918	36.46786	33.18792
MurskaSobota	27.59880	45.95760	30.74177	462.48528
NovaGorica	48.30702	74.96148	31.25988	36.91643
Otlica	32.90305	47.48895	26.16724	25.58927

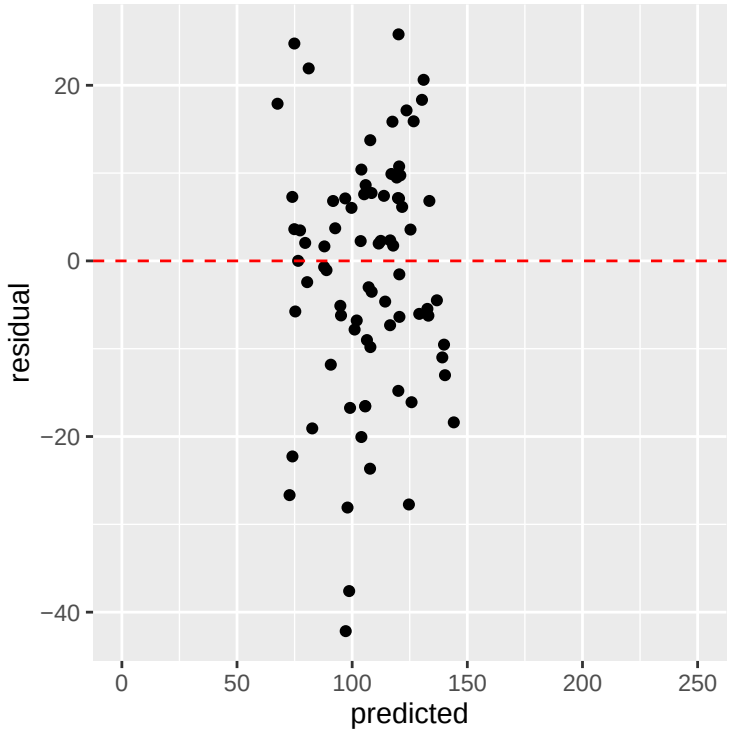
Gamma GLM


Koper
Krvavec
(Actual)
Prediction
180
150
120
90
60
60
90
120
150
180
Model
ARSO
baseline
Gamma reg

Linearna regresija vs Gamma regresija

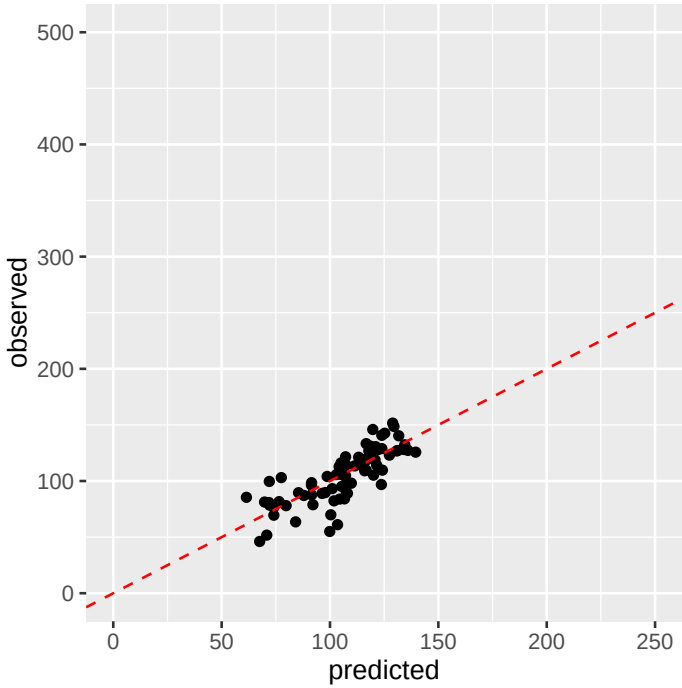
Residuali

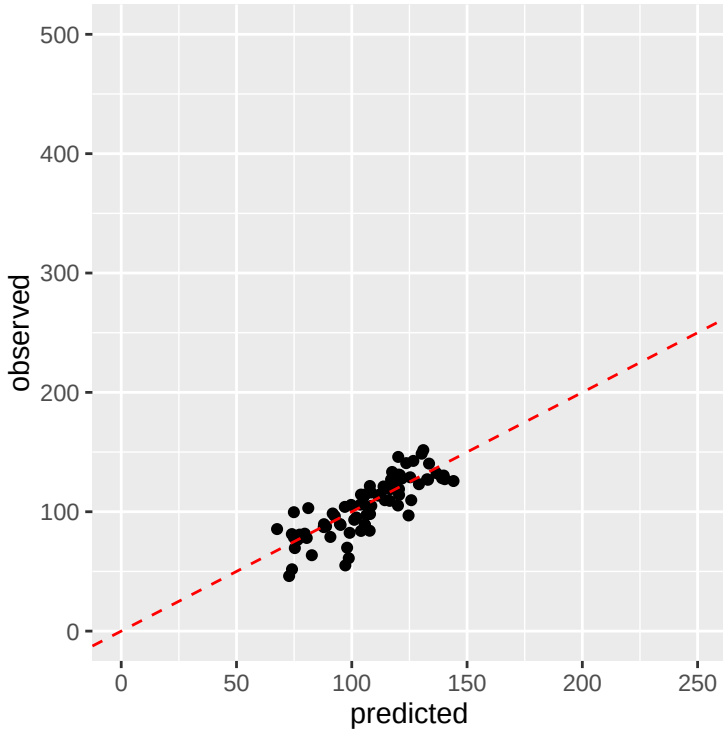




Linearna regresija vs Gamma regresija

Predicted vs observed





Hiearhičen linearen model

- Smiselno je, da pri napovedovanju vrednosti za eno postajo upoštevamo tudi ostale postaje
- Predpostavimo, da so koeficienti β iz multivariatne normalne porazdelitve:
  - multivariatna normalna porazdelitev je skupna vsem postajam
  - Za postajo i: $\beta_i \sim multi\_normal(\mu, \sum)$
- Parametrov skupne multivariatne normalne porazdelitve ne poznamo - jih vključimo v model kot parametre.

Hiearhičen linearen model

Hiper-apriorne porazdelitve:
Za postajo i: $\beta_i \sim multi\_normal(\mu, \sum)$

- $\mu \sim \mathbf{0}$
- $\sum = diag\_matrix(\tau) * \Omega * diag\_matrix(\tau)$
  - Kjer je τ scale vector in Ω korelacijska matrika
  - Lažje razumevanje pomena paremetrov in postavlanje primernih apriornih porazdelitev
  - Priporočljive: $\tau_k \sim Cauchy(0, 2.5)$ in $\Omega \sim LKJcorr(\nu); \nu \geq 1$
    Pod pogojem, da so vhodni podatki standardizirani

Hiearhičen linearen model

```
data {
  int<lower=0> sts; // number of stations
  int<lower=0> n; // number of samples
  int<lower=0> k; // number of attributes
  matrix[n, k] x; // samples
  vector[n] y; // targets
  int idx[n]; // station indexes of samples
  vector[k] zeros;
  }

parameters {
  corr_matrix[k] Omega; // prior correlation
  vector<lower=0>[k] tau; // prior scale
  vector[k] betas[sts]; // ind. coeffs
  vector[sts] alpha; // intercept
  real<lower=0> sigma;
}

model {
  tau ~ cauchy(0, 2.5);
  Omega ~ lkj_corr(1);
  betas ~ multi_normal(beta_mus, quad_form_diag(Omega, tau));

  for (i in 1:n){
    y[i] ~ normal(x[i] * betas[idx[i]] + alpha[idx[i]], sigma);
  }
}
```

Hiearhičen linearen model

```
data {
  int<lower=0> sts; // number of stations
  int<lower=0> n; // number of samples
  int<lower=0> k; // number of attributes
  matrix[n, k] x; // samples
  vector[n] y; // targets
  int idx[n]; // station indexes of samples
  vector[k] zeros;
}

parameters {
  cholesky_factor_corr[k]L_Omega; // prior correlation
  vector<lower=0>[k] tau; // prior scale
  vector[k] betas[sts]; // ind. coeffs
  vector[sts] alpha; // intercept
  real<lower=0> sigma;
}
model {
  tau ~ cauchy(0, 2.5);
  L_Omega ~ lkj_corr_cholesky(2);
  betas ~ multi_normal_cholesky(zeros, diag_pre_multiply(tau, L_Omega));

  for (i in 1:n){
    y[i] ~ normal(x[i] * betas[idx[i]] + alpha[idx[i]], sigma);
  }
}
```

Velja:

$$\Omega = L_\Omega * L_\Omega^T$$

$$diag\_pre\_multiply(a, b) = diag\_matrix(a) * b$$

Hiearhičen linearen model

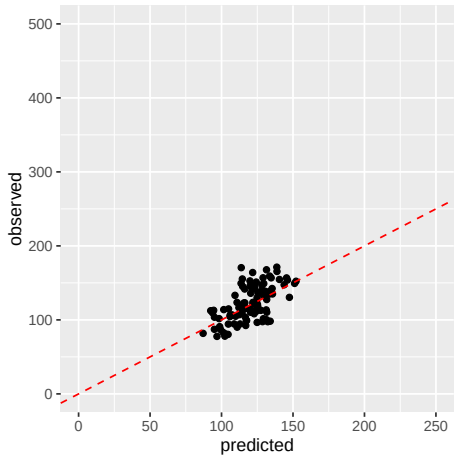
- močno povečano število parametrov
  - število $\beta = k * st$
  - korelacijska matrika dimenzije: $\Omega = \frac{k^2}{2}$ (simetrična)
- Zato prilagoditve v evalvaciji:
  - Izbermo 15 atributov
  - Za vsako izmed postaj napovedujemo samo za leto 2015
  - Učimo na podatkih vseh postaj pred letom 2015

Hiearhičen linearen model

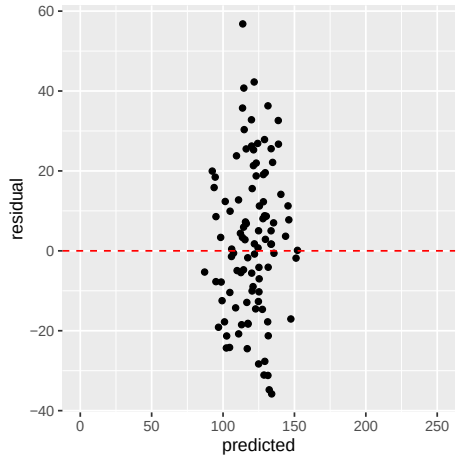
```
Inference for Stan model: hiearhical_reg.
1 chains, each with iter=6000; warmup=500; thin=1;
post-warmup draws per chain=5500, total post-warmup draws=5500.

                 mean se_mean     sd       2.5%        25%      n_eff Rhat
L_Omega[1,1]     1.00    0.00   0.00       1.00       1.00       5500  NaN
L_Omega[1,2]     0.00    0.00   0.00       0.00       0.00       5500  NaN
L_Omega[1,3]     0.00    0.00   0.00       0.00       0.00       5500  NaN
...
L_Omega[15,14]0.06       0.00   0.22      -0.39      -0.09       4058     1
L_Omega[15,15]0.45       0.00   0.14       0.18       0.35       1063     1
tau[1]           2.24    0.01   0.80       1.08       1.69       3107     1
...
tau[15]          1.05    0.02   0.65       0.14       0.56        757     1
betas[1,1]       2.76    0.02   0.93       0.97       2.13       3625     1
...
betas[8,15]     -0.44    0.02   0.80      -2.31      -0.86       2399     1
alpha[1]        106.39    0.03   1.32     103.81     105.52       2336     1
alpha[2]        110.47    0.02   1.18     108.13     109.67       3037     1
alpha[3]        120.61    0.02   1.27     118.12     119.76       3703     1
alpha[4]        115.98    0.06   2.62     110.87     114.20       2100     1
alpha[5]        106.39    0.01   1.05     104.33     105.70       5500     1
alpha[6]        109.02    0.02   1.14     106.73     108.24       5500     1
alpha[7]        119.45    0.04   1.19     117.17     118.66        870     1
alpha[8]        114.82    0.02   1.22     112.41     113.98       3831     1
sigma            19.27    0.00   0.23      18.83      19.12       5500     1
lp__         -12518.34    0.31  11.67  -12541.99  -12525.89       1416     1
```

Hiearhičen linearen model

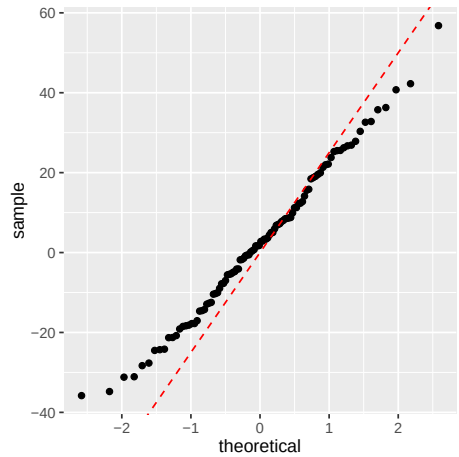
Predicted vs observed



Predicted vs residual



Theoretical vs sample



Rezultati

Rezultati napovedovanja

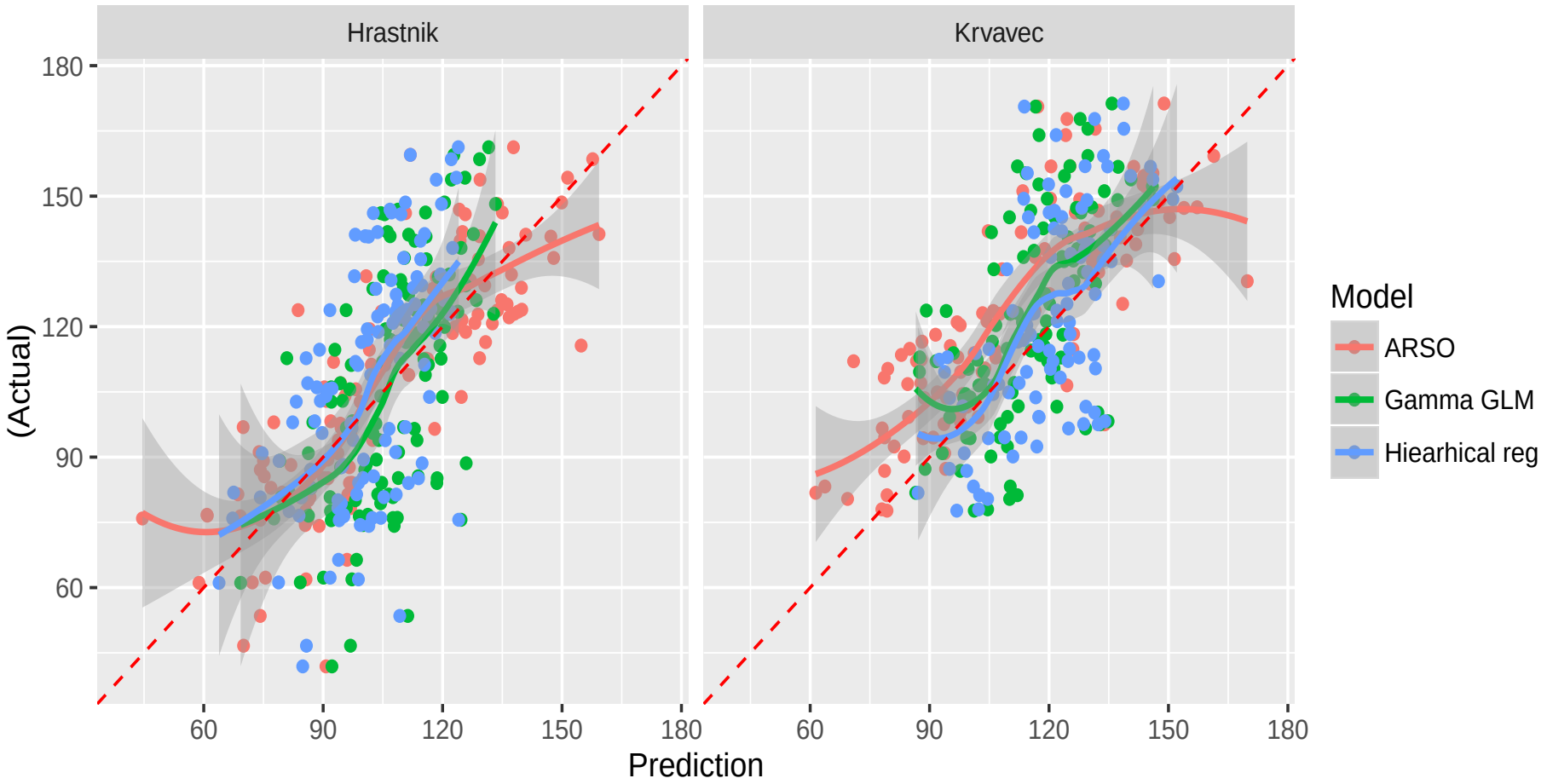
RMSE:

	ARSO	baseline	Gamma GLM	Hiearhical reg
Hrastnik	240.8624	784.0064	474.2788	505.6522
Iskrba	321.6947	659.4305	481.9896	418.4417
Koper	271.8148	1098.592	721.1761	675.9680
Krvavec	368.4219	636.1638	423.5086	356.7094
Ljubljana	231.1643	897.0012	572.3642	491.9442
MurskaSobota	244.7838	625.1524	418.7605	495.1632
NovaGorica	804.4294	1132.483	887.8743	603.4293
Otlica	378.3348	702.5894	460.1057	459.1148

Standardne napake:

	ARSO	baseline	Gamma GLM	Hiearhical reg
Hrastnik	36.10290	85.56374	55.01289	54.40281
Iskrba	49.14359	104.09695	87.90603	68.92420
Koper	37.78400	128.68644	98.91060	89.61301
Krvavec	50.31421	72.02047	54.93803	47.89539
Ljubljana	33.38309	127.52598	85.29842	71.29617
MurskaSobota	32.20261	97.80810	74.63881	74.27205
NovaGorica	112.53522	127.97011	101.33297	67.81641
Otlica	64.52910	69.83299	50.14827	49.71800

Rezultati


Hrastnik
Krvavec
(Actual)
Prediction
180
150
120
90
60
60
90
120
150
180
Model
ARSO
Gamma GLM
Hiearhical reg

Možno nadalnje delo:

- Razširitev hiearhičnega modela
  - Dodati apriorne hiper-porazdelitve na α parameter
  - Iskanje najbolj primernih apriornih porazdelitev
- Temeljita analiza aposteriornih vrednosti modelov
  - npr. Ω in τ pri hiearhičnem modelu (medsebojni vpliv atributov)
- Obširnejša analiza in preprocesiranje atributov
- Uporaba nelinearnih modelov