ANOVA REGIONI

Per confrontare i voti degli studenti nelle tre diverse regioni, ho fatto un'ANOVA. Ho lavorato con tre diversi modelli, nei quali ho considerato le covariate dello studente oppure no, ho usato il raggruppamento per scuole oppure no.

MODELLO ANOVA 2

Il modello è della seguente forma:

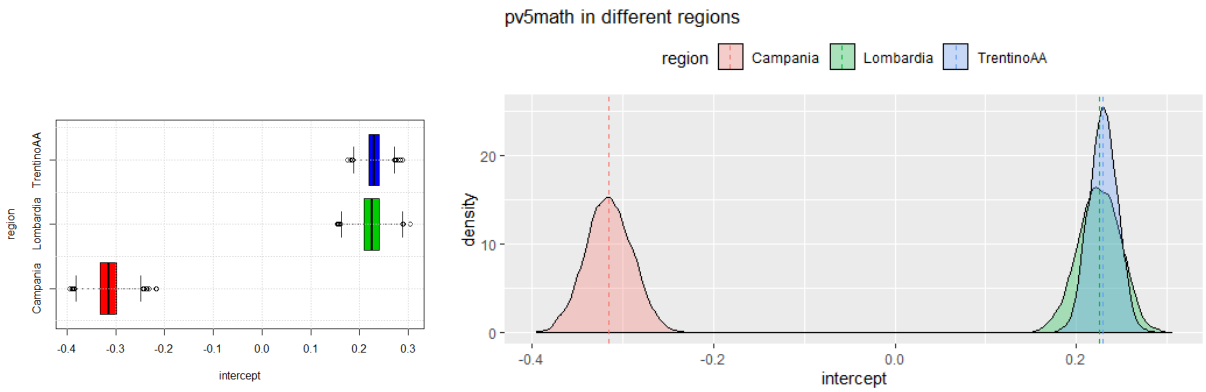
$$
\begin{aligned}
&Y\_ij \mid theta\_j, S \sim N\_2(theta\_j, S^{(-1)}) \\
&theta1, ..., theta3 \sim N\_2(mean0, I\_2 * omega^2) \\
&sigma1 \sim Unif(0,1) \\
&sigma2 \sim Unif(0,1) \\
&rho \sim Unif(-1,1) \\
&S = [sigma2^2, -sigma1*sigma2*rho; \\
&-sigma1*sigma2*rho, sigma1^2] *1/det(S)
\end{aligned}
$$

Ho fissato gli elementi del vettore mean0 come l'intercetta dei modelli lineari con tutte le covariate significative scelte dalla Cross Validation e con risposta, rispettivamente, pv5math e pv5read. Ho fissato omega^2 pari a 5 (giusto per abbondare un po' rispetto alla standardizzazione). Come valori iniziali, ho passato sigma1 e sigma2 come deviazioni standard campionarie di pv5math e pv5read, rho come correlazione tra pv5math e pv5read, theta_j uguale al vettore nullo (0,0).

N iterazioni: 1000 + update 1000 + 11000. Thin: 5

Si arriva a convergenza: bei traceplot, posterior dei parametri regolari, gewake test con buoni p-value. Tuttavia, l'autocorrelazione non va a zero come dovrebbe (sintomo che il modello è, di per sé, troppo privo di informazioni, perché non stiamo dando nessuna covariata dello studente!)

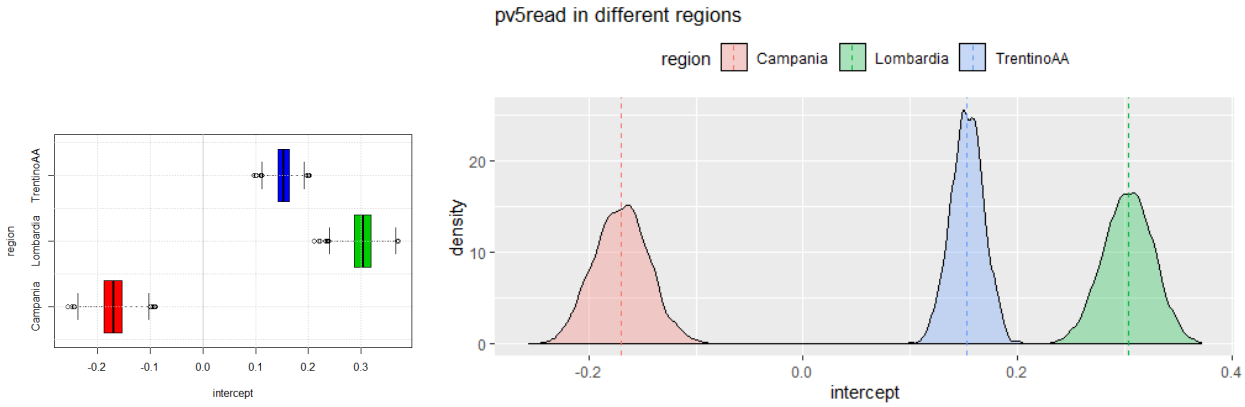
Ho confrontato la media delle risposta pv5math a seconda della regione, ovvero il primo elemento dei parametri theta_j. Qui di seguito, il plot della distribuzione a posteriori dei parametri theta1, theta2 e theta3. Si vede che la Campania ha media negativa ed è molto penalizzata rispetto a Lombardia e TrentinoAA, positive e molto simili tra loro.



In particolare, le medie, deviazioni standard e intervalli di confidenza sono riportati nella seguente tabella. Inoltre, ecco il boxplot con quantili e mediane.

region	mean	Sd	credible interval (con TCL)	Credible interval (con quantili)
Campania -0.3147280 0.02547644 -0.3646618 -0.2647942 -0.3645029 -0.2645475				
Lombardia 0.2250432 0.02308432 0.1797980 0.2702885 0.1784901 0.2676404				
TrentinoAA 0.2293455 0.01566724 0.1986377 0.2600532 0.1987464 0.2592699				

Ho confrontato la media delle risposta pv5read a seconda della regione, ovvero il secondo elemento del parametro theta_j. Qui di seguito, il plot della distribuzione a posteriori dei parametri theta1, theta2 e theta3. Il podio delle regioni per "bravura" vede prima la Lombardia e ultima la Campania. Il Trentino ha comunque media positiva.



In particolare medie, sd e credible intervals sono:

region	Mean	Sd	Credible interval (TCL)	Credible interval (quantiles)
Campania -0.1697985 0.02531866 -0.2194230 -0.1201739 -0.2179756 -0.1199283				
Lombardia 0.3029194 0.02317202 0.2575022 0.3483366 0.2567986 0.3470104				
TrentinoAA 0.1529317 0.01534730 0.1228510 0.1830124 0.1224712 0.1827399				

PERCIÒ, tutto ciò si può interpretare nel modo seguente: se considero i voti di studenti di regioni diverse solamente dipendenti dalla regione stessa e non da qualità dello studente, allora la Campania ha risultati di gran lunga peggiori che la Lombardia e il TrentinoAA. La Lombardia si classifica sempre in testa e ottiene le maggiori valutazioni degli studenti in math e read.

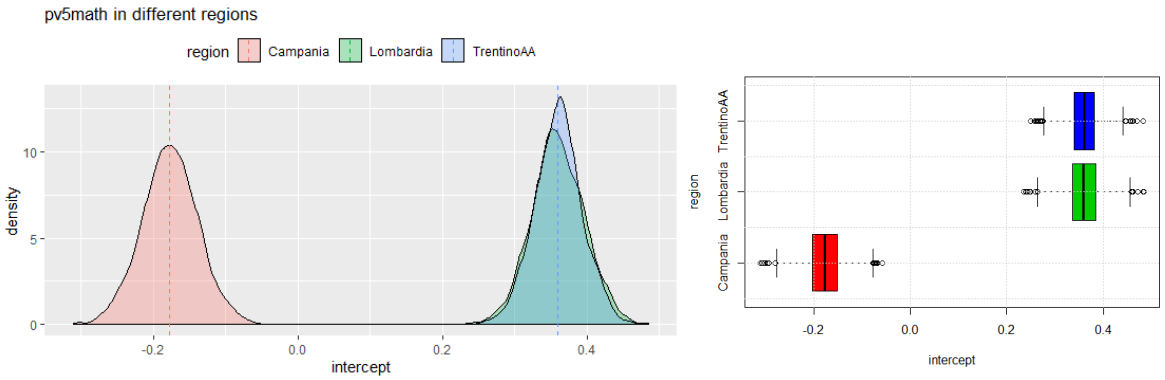
MODELLO ANOVA 1

Y_ij | mu_ij, sigma^2 ~ N_2(mu_ij, S^(-1))
mu_ij = beta0j + beta1*Xi1 + beta2*Xi2 + ... beta_p*Xip
beta1, ..., beta_p ~ N_2(meanp, I_2* tau^2)
beta01, ..., beta03 ~ N_2(mean0, I_2* omega^2)
sigma1 ~ Unif(0,1)
sigma2 ~ Unif(0,1)
rho ~ Unif(-1,1)
S = [sigma2^2, -sigma1*sigma2*rho;
-sigma1*sigma2*rho, sigma1^2] *1/det(S)

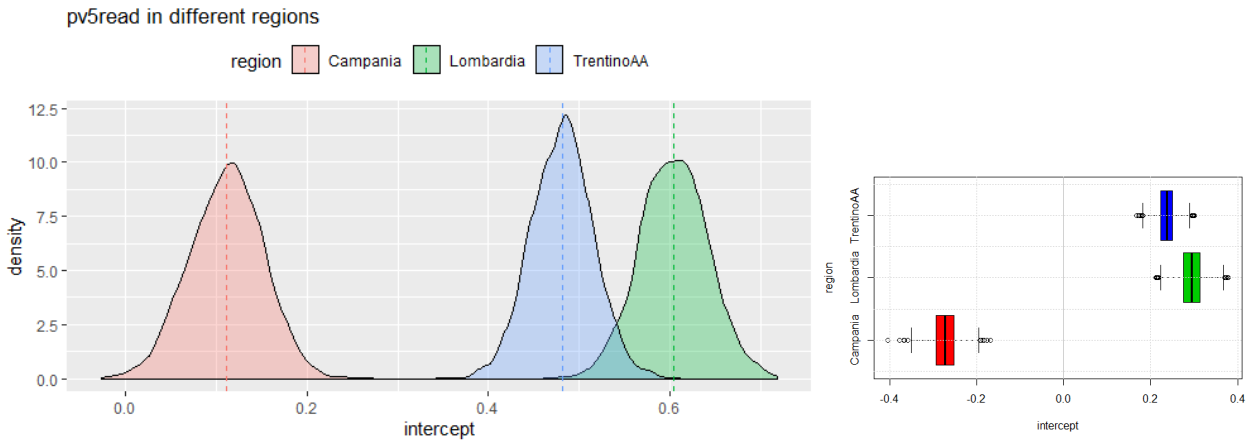
Ho fissato i parametri esattamente come prima. In più questa volta ho fissato meanp come un vettore p-dimensionale di zeri e tau^2 ancora uguale a 5. Stesse iterazioni e thin di sopra.

Si arriva a convergenza: bei traceplot, posterior dei parametri regolari, gewake test con (non sempre ma di solito) buoni p-value. Tuttavia, l'autocorrelazione va a zero più o meno velocemente a seconda dei coefficienti.

Ho confrontato il coefficiente beta0 a seconda della regione. Ovvero, sto confrontando la componente della media di Y che dipende dalla regione, quella parte indipendente dallo studente e le sue covariate (A pari di covariate, lo studente con voti migliori è nella tal regione). Qui di seguito, il plot della distribuzione a posteriori degli elementi dei parametri beta0_1, beta0_2 e beta0_3, prima per pv5math e poi per pv5read.



region	Mean	Sd	Credible interval (TCL)	Credible interval (quantile)
Campania -0.1772858 0.03853565 -0.2528157 -0.1017559 -0.2529625 -0.09892981				
Lombardia 0.3600284 0.03637959 0.2887244 0.4313324 0.2893358 0.43175629				
TrentinoAA 0.3601418 0.03290334 0.2956513 0.4246324 0.2955318 0.42582461				



region	Mean	Sd	Credible interval (TCL)	Credible interval (quantiles)
Campania 0.1113851 0.04011034 0.03276882 0.1900014 0.03352121 0.1870346				
Lombardia 0.6043338 0.03712032 0.53157795 0.6770896 0.53105229 0.6775359				
TrentinoAA 0.4810430 0.03363588 0.41511671 0.5469694 0.41575474 0.5475236				

PERCIÒ, si può concludere che *esattamente come prima,* la Lombardia troneggia sempre, la Campania è la peggiore delle tre e il Trentino si difende bene. DA NOTARE: Mentre prima la Campania era sempre negativa e anche di molto, adesso ha i beta0 molto più spostati verso destra e, addirittura, nel reading ha media positiva. Analogamente, anche Trentino e Lombardia si spostano verso dx. Questo mi fa pensare che, anche se la relazione tra le regioni resta sempre la stessa (il podio è lo stesso), con l'introduzione delle covariate studente il voto viene penalizzato da qualcos'altro piuttosto che dalla regione. Ciò che penalizza è relativo alle componenti dei coefficienti beta1 e beta2 con segno negativo:

pv5math: videogames, text_anxiety, study_before, ISCED
pv5read: gender(penalizza i M), videogames, text_anxiety, study_before, ISCED

CONCLUSIONI

Non so quale dei modelli sia necessario utilizzare per spiegare la differenza tra le regioni. Il risultato evidenzia sempre lo stesso andamento (Campania BUU, Lombardia e Trentino TOP), ma con distribuzioni spostate e a volte significativamente. Onestamente, terrei buoni i confronti dei theta nel modello anova2 e dei beta0 nel modello anova1 seguendo l'interpretazione che ho scritto sopra.

BF and hypothesis testing

In quest'ultima parte, visti i risultati dell'anova, proviamo a fare qualche test di ipotesi per classificare le tre regioni. Siccome sembra più accurato, consideriamo il modello anova1 e confrontiamo i beta0.

Per valutare la veridicità dell'ipotesi nulla, è stato calcolato il Bayes Factor relativo al test come il rapporto tra la posterior odd P(H0|data)/P(H1|data) e la prior odd P(H0)\P(H1). Entrambi i valori sono stati calcolati a partire da un campionamento dei parametri confrontati dalla prior e dalla posterior.

In particolare, beta0_1, beta0_2, beta0_3 ~ N_2((0,0), omega^2 *I_2) iid e perciò sono stati campionati N valori per ciascun coefficiente da tale distribuzione.

Il campionamento dalla legge a posteriori dei beta0_1, beta0_2, beta0_3 è stato effettuato durante l'analisi precedente per valutare il modello anova1. Perciò, sono stati estratti gli N valori per ciascun parametro campionati dalla posterior.

TEST 1:

H0: La campania è peggio di tutte → H0: beta0_1 < beta0_2 & beta0_1 < beta0_3
H1: Non è vero → H1: altrimenti

```
P0 = sum(beta01_prior[,1] < beta03_prior[,1] & beta01_prior[,1] < beta02_prior[,1] &
      beta01_prior[,2] < beta03_prior[,2] & beta01_prior[,2] < beta02_prior[,2])/N
PriorOdd = P0/(1-P0) = 0.1195929

P0 = sum(beta01_post[,1] < beta03_post[,1] & beta01_post[,1] < beta02_post[,1] &
      beta01_post[,2] < beta03_post[,2] & beta01_post[,2] < beta02_post[,2])/N
PostOdd = P0/(1-P0) = INF
```

BF = PostOdd/PriorOdd = INF (siccome H0 si verifica sempre nel mio campione)

Per cui l'evidenza dell'ipotesi nulla è più che chiara

TEST 2:

H0: il trentino è peggio della lombardia → H0: beta0_3 < beta0_2
H1: Non è vero → H1: altrimenti

P0 = sum(beta01_prior[,1] < beta03_prior[,1] & beta01_prior[,1] < beta02_prior[,1] &
beta01_prior[,2] < beta03_prior[,2] & beta01_prior[,2] < beta02_prior[,2])/N
PriorOdd = P0/(1-P0) = 0.3325257

P0 = sum(beta01_post[,1] < beta03_post[,1] & beta01_post[,1] < beta02_post[,1] &
beta01_post[,2] < beta03_post[,2] & beta01_post[,2] < beta02_post[,2])/N
PostOdd = P0/(1-P0) = 0.9625335

BF = PostOdd/PriorOdd = 2.894613
2log(BF) = 2.125703

Per cui c'è evidenza a favore di H0 ma molto debole