Bayesiaanse Variantieanalyse

Andrew Gelman*

November 30, 1999

Samenvatting

Variantieanalyse (ANOVA) is een heel belangrijke statistische methode. Jammer genoeg is het in complexe problemen moeilijk om de correcte ANOVA uit te voeren. In klassieke ANOVA is er geen automatische methode voor het vinden van de correcte variantiecomponent voor de noemer van de F-grootheid. Wij kunnen dit probleem oplossen met Bayesiaanse hiërarchische regressie.

1 Inleiding

Wat betekent ANOVA? De econometristen zeggen dat het een speciaal geval van lineaire regressie is. Dus een econometrist heeft ANOVA niet nodig. De Bayesianen zeggen dat ANOVA F-toetsen betekent. Bayesianen houden niet van toetsen, dus houden zij niet van ANOVA.

Wij vinden dat ANOVA geen speciaal geval is van klassieke regressie maar van Bayesiaanse regressie. Voor eenvoudige problemen is klassieke regressie OK, maar voor hiërarchische ontwerpen hebben wij hiërarchische modellen met twee of meer variantiecomponenten nodig.

2 ANOVA met lineaire regressie: goed nieuws

Je wil een ANOVA berekenen. Hoe kun je een lineair regressiemodel dat equivalent is aan je ANOVA-model vinden? Wij zullen dit met enkele voorbeelden uitleggen.

2.1 Bijvoorbeeld: proefopzet volgens latijns vierkant

Om de algemene idee te illustreren, beschouwen wij eerst een 5×5 latijns vierkant.

*Department of Statistics, Columbia University, New York, USA, gelman@stat.columbia.edu. Dank aan Iwin Leenen en Iven Van Mechelen voor de bespreking, aan Rianne Janssen en de redacteurs voor de belangrijke hulp met het Nederlands, aan de Onderzoeksraad van de Katholieke Universiteit Leuven voor krediet F/96/9, en aan de U.S. National Science Foundation voor Young Investigator Award DMS-9796129 en subsidie SBR-97008424.

A	B	C	D	E
B	A	D	E	C
E	C	B	A	D
D	E	A	C	B
C	D	E	B	A

Je kan de ANOVA doen als een lineaire regressie van de 25 gegevens op de volgende variabelen:

- 1 constante,
- 12 indicatorvariabelen:
  - 4 voor de rijen,
  - 4 voor de kolommen,
  - 4 voor de behandelingen.

(Zoals in een ANOVA-tabel moet je 4 variabelen in elke groep hebben, want als je voor de 5 niveaus ook 5 variabelen zou nemen dan zouden de variabelen collineair zijn.) Voor elk van de 3 groepen van variabelen bereken je eerst de variantie van de coëfficiënten, die je dan met de schattingsvariantie vergelijkt. Deze ANOVA gelijkheid is toepasbaar op elke groep van J variabelen:

$$\begin{aligned} \text{variantie van de geschatte parameters } \hat{\beta}_j &= \text{variantie van de echte parameters } \beta_j + \text{schattingsvariantie} \\ \mathrm{E}(\mathrm{var}_{j=1}^{J} \hat{\beta}_j) &= \mathrm{var}_{j=1}^{J} \beta_j + \mathrm{E}(\mathrm{var}(\hat{\beta}_j | \beta_j)) \\ V(\hat{\beta}) &= V(\beta) + V_{\text{schatting}} \end{aligned} \tag{1}$$

Op basis van de lineaire regressie kun je de schattingen van $V(\hat{\beta})$ en $V_{\text{schatting}}$ berekenen. Daarmee kun je dan $V(\hat{\beta})/V_{\text{schatting}}$ berekenen: dit is de F-statistiek voor ANOVA. Ook kan je het verschil $V(\hat{\beta}) - V_{\text{schatting}}$ bepalen voor de schatting van de variantie $V(\beta)$ van de random effecten in de ANOVA.

In deze analyse komen 3 groepen van regressiecoëfficiënten voor, dus krijgen wij 3 F-toetsen of 3 geschatte random effect varianties, zoals in klassieke ANOVA.

2.2 Vergelijking van twee groepen

Klassieke lineaire regressie werkt ook met eenvoudige voorbeelden. Neem een vergelijking van twee groepen, met 10 eenheiden in elke groep. Dit is een klassieke eenwegs ANOVA. Ook hier kan je dezelfde conclusie verkrijgen met lineaire regressie. Je hebt 20 gegevens en 2 predictorvariabelen: 1 constante en 1 voor behandeling (of geen constante en 2 voor behandeling). Je hebt 18 residuele vrijheidsgraden, net zoals in de klassieke eenwegs ANOVA.

2.3 Gepaarde proefopzet

Voor de regressieanalyse van de gepaarde proefopzet moet je de volgende idee gebruiken: in de analyse moet je alle informatie uit het onderzoeksontwerp omvatten. Neem een proefopzet met 10 paren: in elk paar is een behandelde eenheid en een controle eenheid. Je hebt nog 20 gegevens maar nu 11 predictorvariabelen:

- 1 constante,
- 1 indicatorvariabele voor behandeling,
- 9 indicatorvariabelen voor groepen,

en 9 residuele vrijheidsgraden. Als je klassieke lineaire regressie toepast, dan is de coëfficiënt voor behandeling dezelfde als in de normale schatting met ANOVA en is het kwadraat van zijn t-statistiek gelijk aan de ANOVA F-statistiek.

3 ANOVA met klassieke lineaire regressie: slecht nieuws

3.1 ANOVA voor de hiërarchische proefopzet

Nu komen wij aan een voorbeeld waar klassieke regressie niet het correcte ANOVA-antwoord geeft. Bijvoorbeeld, een experimentator heeft 4 behandelingen voor een industrieel proces en hij zal 6 afmetingen nemen in elk van 5 machines. Dit is dus een eenvoudig hiërarchische proefopzet: 4 behandelingen $\times$ 5 machines $\times$ 6 afmetingen, met de volgende ANOVA-tabel:

Variabelen	vg
behandeling	3
behandeling $\times$ machine	16
behandeling $\times$ machine $\times$ afmeting	100
totaal	119

Er zijn geen variabelen voor uitsluitend "machine" of "afmeting" want het onderzoeksontwerp is genest.

Als je hiërarchische ANOVA niet kent, dan is het niet zo gemakkelijk om de standaardfout van de behandelingcoëfficiënten te schatten. Je kan de behandelingcoëfficiënten op twee manieren schrijven:

$$\bar{y}_{i..} = \frac{1}{30} \sum_{jk} y_{ijk} \tag{2}$$

of

$$\bar{y}_{i..} = \frac{1}{5} \sum_{j} \bar{y}_{ij.} \tag{3}$$

Formule (2) gebruikt alle gegevens en suggereert een standaardfout met 29 vrijheidsgraden voor elke behandeling, maar deze formule negeert de geneste aard van het ontwerp. Formule (3) volgt het ontwerp, maar misschien gebruikt hij niet alle informatie, met slechts 4 vrijheidsgraden per behandeling.

Formules (2) en (3) geven dezelfde schattingen voor de behandelingresultaten maar geven verschillende ANOVA F-toetsen. De analyse van dit ontwerp is niet automatisch, omdat je de correcte noemer voor de F-toets moet gebruiken. Voor de variabele behandeling moet je noemer de variantie van "behandeling × machine" zijn en niet de variantie van "behandeling × machine × afmeting".

3.2 Klassieke regressie voor de hiërarchische proefopzet

Kunnen wij dit probleem met klassieke regressie oplossen? Laten wij over een aantal regressiemodellen voor deze 120 gegevens nadenken. De eenvoudigste regressie heeft slechts 4 predictorvariabelen:

- 1 constante,
- 3 indicatorvariabelen voor behandeling,

en 116 residuele vrijheidsgraden. Dit model geeft de verkeerde residuelevariantie. Wij hebben de machinevariantie nodig, niet de afmetingvariantie.

Omdat wij de machines in de proefopzet gebruiken, moeten wij de machines ook in de analyse gebruiken. Dus hebben wij de volgende predictorvariabelen:

- 1 constante,
- 3 indicatorvariabelen voor behandeling,
- 20 indicatorvariabelen voor machines,

en 96 residuele vrijheidsgraden. Maar deze predictorvariabelen zijn collineair. We moeten vier indicatorvariabelen elimineren en nu hebben wij:

- 1 constante,
- 3 indicatorvariabelen voor behandeling,
- 16 indicatorvariabelen voor machines,

en 100 residuele vrijheidsgraden. Deze regressie heeft indicatorvariabelen voor machines, maar nog steeds heeft hij de residuelevariantie van de afmetingen. Dit is nog steeds verkeerd: de correcte residuelevariantie komt van de machines.

4 ANOVA met hiërarchische Bayesiaanse lineaire regressie: goed nieuws

Nu weten wij dat de klassieke regressie niet werkt voor de hiërarchische proefopzet. Wij kunnen hier de klassieke ANOVA gebruiken maar dat gaat niet automatisch—met de ANOVA moeten wij de correcte residuelevariantie weten.

Maar met de Bayesiaanse lineaire regressie kunnen wij de correcte antwoorden automatisch vinden. Je moet deze regels volgen:

1. In het regressiemodel gebruik je alle variabelen die in de proefopzet zijn gebruikt.
2. Maak hiërarchische modellen voor alle groepen van coëfficiënten in het model. Voor elke lijn in de ANOVA-tabel is er een "groep."

Wij geven twee voorbeelden.

4.1 Hiërarchisch ontwerp

Kijk naar het voorbeeld van Sectie 3.1. Wij hebben 120 gegevens en het regressiemodel $y_{ijk} \sim \mathrm{N}((X\beta)_{ijk}, \sigma^2)$, met de volgende coëfficiënten β:

- 1 constante,
- 4 behandelingcoëfficiënten (met een $\mathrm{N}(0, \tau_1^2)$ priorverdeling),
- 20 machinecoëfficiënten (met een $\mathrm{N}(0, \tau_2^2)$ priorverdeling).

Maak je geen zorgen over vrijheidsgraden of over collineairiteit omdat je geen platte *a priori* verdelingen hebt. (Als $\tau_1 = \infty$ of $\tau_2 = \infty$ dan heb je collineairiteit en een ongeschikte posteriorverdeling.) Met Bayesiaanse methoden kun je σ, τ_1, τ_2 schatten en dan conclusies trekken over de β's (met Bayesiaanse inkrimping; bijvoorbeeld, Gelman et al., 1995, en Carlin en Louis, 1996).

De Bayesiaanse conclusie geeft de correcte schattingen voor de variantiecomponenten: σ van de variantie binnen machines, τ_1 voor de variantie tussen machines maar binnen behandelingsgroepen, en τ_2 voor variantie tussen behandelingsgroepen. De posterior varianties voor de behandelingcoëfficiënten komen automatisch van τ_2 (zoals in de correcte klassieke ANOVA) niet van σ (zoals in de verkeerde klassieke regressie). Bijvoorbeeld, als de echte σ nul is, dan naderen in onze Bayesiaanse regressie de posterior varianties van de behandelingcoëfficiënten nul niet.

4.2 Splitplot proefontwerp

4.2.1 Klassieke ANOVA

In moeilijkere problemen kunnen wij ook dezelfde methode voor automatische ANOVA met hiërarchische regressie toepassen. Neem bijvoorbeeld een splitplot latijns vierkant. De klassieke ANOVA voor dit probleem is moeilijk omdat je verschillende variantiecomponenten moet gebruiken voor de verschillende rijen in de ANOVA-tabel (bv., Snedecor en Cochran, 1989, en Kirk, 1995). Neem bijvoorbeeld een ANOVA-tabel voor een 5×5 latijns vierkant met 5 tussengroepsbehandelingen (A, B, C, D, E) en 2 binnengroepsbehandelingen (1, 2):

Variabelen	vg
rijen	4
kolommen	4
(A, B, C, D, E)	4
residu	12
$(1, 2)$	1
$(1, 2) \times$ rijen	4
$(1, 2) \times$ kolommen	4
$(1, 2) \times (A, B, C, D, E)$	4
$(1, 2) \times$ residu	12
totaal	49

Je moet de tussengroepsvariabelen met de tussengroeps residuelevariantie vergelijken en de binnengroepsvariabelen met de binnengroeps residuelevariantie (genoemd "$(1, 2) \times$ residu" in de ANOVA-tabel).

4.2.2 Bayesiaanse analyse met hiërarchische regressie

Als je met klassieke lineaire regressie de splitplot gegevens analyseert, dan krijg je de verkeerde variantieschattingen voor de tussengroepscoëfficiënten, net zoals in Sectie 3. In het hiërarchische model kan je voor elke groep van variabelen een variantiecomponent gebruiken:

Variabelen	vg	variantie
constante	1	τ_0^2
rijen	5	τ_1^2
kolommen	5	τ_2^2
(A, B, C, D, E)	5	τ_3^2
residu	25	τ_4^2
$(1, 2)$	2	τ_5^2
$(1, 2) \times$ rijen	5	τ_6^2
$(1, 2) \times$ kolommen	5	τ_7^2
$(1, 2) \times (A, B, C, D, E)$	5	τ_8^2
$(1, 2) \times$ residu	25	τ_9^2

De posteriorvarianties voor de tussengroeps- en binnengroepsvariabelen komen automatisch.

4.3 Een paradox?

Nu hebben wij een paradox. In de klassieke analyse moeten wij de ontwerpinformatie weten om de correcte residuelevariantie te kiezen: dus met het splitplot ontwerp moeten wij "binnen-" en "tussengroeps" variantiecomponenten bepalen. De Bayesiaanse methode vindt automatisch de correcte varianties zonder bijkomende instructies voor de analyse te moeten geven. Hoe "weet" de Bayesiaanse analyse de ontwerpinformatie?

Hier is de oplossing van de paradox: de ontwerpinformatie voor de hiërarchische proefopzet ligt in de X-matrix van de regressie. De klassieke analyse gebruikt deze informatie niet omdat die een platte priorverdeling ($p(\beta) \propto 1$; i.e., $\tau = \infty$) heeft.

5 Conclusies

Econometristen zeggen dat ANOVA een speciaal geval van klassieke lineaire regressie is (zie, e.g., Kleinbaum et al., 1988, blz. 104). Dat is waar met eenvoudige onderzoeksontwerpen, maar met hiërarchische ontwerpen moet de regressie Bayesiaans zijn om de correcte variantiecomponenten te krijgen, en is dat niet zo gemakkelijk (zie Samuels, Casella, en McCabe, 1991). Deze Bayesiaanse methode is automatischer dan de klassieke ANOVA—wij hebben dat in het splitplot voorbeeld gezien. Misschien zal deze methode in de toekomst in ANOVA computerprogramma's automatisch gebruikt worden.

Andere voordelen van de hiërarchische Bayesiaanse analyse zijn betere schattingen voor de coëfficiënten (zie Gelman et al., 1995, en Carlin en Louis, 1996), vooral als een van de variantiecomponenten dicht bij nul is (zie Rubin, 1981, en Gelman, 1996). Het meest klaarblijkelijke voordeel van de Bayesiaanse methode is dat de conclusies de onzekerheid in de variantiecomponenten bevatten. Dit is belangrijk als je veel coëfficiënten schat omdat dan de klassieke puntschattingen te veranderlijk zijn. (Je kan dat in (1) zien: de verwachte waarde van de variantie van de geschatte parameters $\hat{\beta}_j$ is groter dan de variantie van de echte parameters β_j.) Als een hiërarchische variantiecomponent τ dicht bij nul is, dan is er geen goede puntschatting $\hat{\tau}$: voor een correcte conclusie over de parameters $\hat{\beta}_j$ heb je de onzekerheid in τ nodig.

Misschien maak je je zorgen over de Bayesiaanse methoden omdat ze gebruik maken van een specifiek model. Dit is een echt discussiepunt. Dit is ons antwoord: de Bayesiaanse hiërarchische regressie kan dezelfde conclusies als klassieke ANOVA geven (maar automatischer) voor problemen waar de klassieke regressie niet werkt.

Referenties

Carlin, B. P., en Louis, T. A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis.* Londen: Chapman & Hall.

Samuels, M. L., Casella, G., and McCabe, G. P. (1991). Interpreting blocks and random factors (met discussie). *Journal of the American Statistical Association* **86**, 798–821.

Gelman, A. (1996). Discussie van "Hierarchical generalized linear models," van Y. Lee en J. A. Nelder. *Journal of the Royal Statistical Society B.*

Gelman, A., Carlin, J. B., Stern, H. S., en Rubin, D. B. (1995). *Bayesian Data Analysis.* Londen: Chapman & Hall.

Kirk, R. E. (1995). *Experimental Design: Procedures for the Behavioral Sciences*, derde editie. Pacific Grove, Calif.: Brooks/Cole.

Kleinbaum, D. G., Kupper, L. L., Muller, K. E., and Nizam, A. (1998). *Applied Regression Analysis and Other Multivariate Methods*, derde editie. Pacific Grove, Calif.: Brooks/Cole.

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics* **6**, 377–401.

Snedecor, G. W., en Cochran, W. G. (1989). *Statistical Methods*, achtste editie. Iowa State University Press.

Ontvangen: 25 mei 1999
Geaccepteerd: 12 oktober 1999