DE VARIANTIE VAN OP EEN MODEL GEBASEERDE SCHATTERS 

Jeroen Pannekoek  
Kees Zeelenberg

*) Met dank aan Martin Boon en Abby Israëls voor commentaar op eerdere versies.  
De in dit rapport weergegeven opvattingen zijn die van de auteurs en komen niet  
noodzakelijk overeen met het beleid van het Centraal Bureau voor de Statistiek.  

BPA no.: 7459-95-RSM  
13 september 1995  

Proj.RSM-9440-2  
Eerste versie
SAMENVATTING

We onderzochten hoe varianties van schatters die op een regressiemodel gebaseerd zijn, bepaald kunnen worden. Hierbij wordt rekening gehouden met zowel de stochastiek ten gevolge van het steekproefkarakter van de data als de model-stochastiek. We concentreren ons op de ‘fitted value’ van het model en functies ervan. We geven een toepassing op een budgetverdeelmodel.
1. Inleiding

In deze nota onderzoeken we hoe varianties van schatters die op een regressiemodel gebaseerd zijn, bepaald kunnen worden. Hierbij wordt rekening gehouden met zowel de stochastiek ten gevolge van het steekproefkarakter van de data als de model-stochastiek. In § 2 behandelen we de variantie van de 'fitted value', in § 3 de variantie van een lineaire transformatie van de 'fitted value', en in § 4 geven we een toepassing op het budgetverdeelmodel, dat door Boon en De Haan (1995) gebruikt is om ten behoeve van de consumentenprijsindex budgetaandelen voor kleine groepen te berekenen.

2. De variantie van de 'fitted value'

We gaan uit van het lineaire regressiemodel

\[ y_i = x_i' \beta + \epsilon_i, \quad i = 1, 2, \ldots, n, \]  \hspace{1cm} (1)

waarin de index i individu i aangeeft, \( y_i \) de te verklaren variabele is, \( x_i \) een vector met verklarende variabelen en \( \epsilon_i \) een storing met \( E(\epsilon_i) = 0 \). We willen nu de covariantiematrix bepalen van de 'fitted value' \( \hat{y} = Xb \), met b een onder het model zuivere schatter van \( \beta \). We houden er rekening mee dat zowel b als X stochastisch kunnen zijn; de matrix X kan bijvoorbeeld gebaseerd zijn op een steekproef en zal dan variëren tussen steekproeven en dus stochastisch zijn. Om de variantie van het produkt Xb te berekenen gebruiken we de algemene regel dat de variantie van een stochast gelijk is aan de verwachting van de conditionele variantie plus de variantie van de conditionele verwachting:

\[ \text{Var}(a) = E_b[\text{Var}(a|b)] + \text{Var}_b[E(a|b)], \]  \hspace{1cm} (2)

waarin a en b willekeurige stochasten (eventueel vectoren of matrices) zijn; in de literatuur wordt deze regel wel de variatie-analyse-regel genoemd. Toepassen van deze regel met \( a = Xb \) en \( b = X \) geeft

\[ \text{Var}(Xb) = E_X[\text{Var}(Xb|X)] + \text{Var}_X[E(Xb|X)] \]
\[-2\]

\[= E_X (X \text{Var}(b | X) X') + \text{Var}_X (X \beta)\]

\[= E_X (X \text{Var}(b | X) X') + (\beta' \otimes I_n) [\text{Var}_X (\text{vec } X)] (\beta \otimes I_n)\]

\[= E_X (X \text{Var}(b | X) X') + \sum_{k=1}^{K} \sum_{l=1}^{L} \beta_k \beta_l \text{Var}(x_k, x_l), \quad (3)\]

waarin \(x_k = (x_{1k}, x_{2k}, \ldots, x_{nk})'\) de vector met de waarnemingen over variabele \(k\) is. De variantie van \(Xb\) bestaat dus uit twee delen: de eerste term heeft betrekking op de modelvariantie van \(b\) en de tweede term op de variantie van \(X\). Merk op dat de totale variantie van \(Xb\) niet per se kleiner is dan de variantie van \(y\), zelfs niet als de modelvariantie kleiner is dan de variantie van \(y\).

Per definitie is \(X \text{Var}(b | X) X'\) een zuivere schatter voor de eerste term in (3). We kunnen dus \(E_X (X \text{Var}(b | X) X')\) schatten door

\[X \text{Var}(b | X) X'. \quad (4)\]

Afhankelijk van de stochastiek in \(X\) en veronderstellingen over de verdeling van de storingsterm \(\varepsilon\), kan de covariantie-matrix van \(b\) geschat worden. Als bijvoorbeeld \(\text{Var}(\varepsilon) = \sigma^2 I_n\), dan geldt voor de kleinste-kwadratenschatter \(b_{\text{OLS}} = (X'X)^{-1}X'Y\) dat \(\text{Var}(b | X) = \sigma^2 (X'X)^{-1}\), wat geschat kan worden door \(s^2 (X'X)^{-1}\) met \(s^2\) een zuivere schatter van \(\sigma^2\).

3. Variantie van een lineaire transformatie van de 'fitted value'

Variantieformule

We bezien een lineaire transformatie van de 'fitted value', zeg \(Y\), waarin \(Y\) een niet-stochastische \((q \times n)\)-matrix met gewichten is. Voor de variantie van \(Y\) geldt

\[\text{Var}(Y) = \text{Var}(TXb) =
\]

\[= E_X \{TX [\text{Var}(b | X) T X'] + \text{Var}_X (TX \beta)\}
\]
\begin{equation}
= E_{X}\{\Gamma X[\text{Var}(b|X)]X'\Gamma'\} + \sum_{k=1}^{K} \sum_{t=1}^{K} \beta_k \beta_t \Gamma'[\text{Var}(x_k, x_t)]\Gamma'.
\end{equation}

Schatting van \(E_{X}\{\Gamma X[\text{Var}(b|X)]X'\Gamma'\} \)

Voor het geval dat de stochastiek in \(X\) alleen veroorzaakt wordt door het steekproef karakter van \(X\), kunnen we \(E_{X}\) vervangen door \(E_{S}\), waarbij \(E_{S}\) de verwachting over alle steekproeven aangeeft. Per definitie is

\begin{equation}
\Gamma X[\text{Var}(b|S)]X'\Gamma'
\end{equation}

een zuivere schatter van \(E_{S}\{\Gamma X[\text{Var}(b|S)]X'\Gamma'\}\). We kunnen dan \(E_{S}\{\Gamma X[\text{Var}(b|S)]X'\Gamma'\}\) schatten door

\begin{equation}
\Gamma X[\text{Var}(b|S)]X'\Gamma'.
\end{equation}

Modelvariantie van \(b\)

Onder de standaard veronderstellingen: \(E(\varepsilon) = 0\) en \(\text{Var}(\varepsilon) = \sigma^2 I_n\), kan de covariantiematrix van \(b\) geschat worden met

\begin{equation}
\text{Var}(b|S) = (X'X)^{-1}s^2,
\end{equation}

waarin \(s^2 = \sum_{i}(y_i - \hat{y}_i)^2/n\) een schatter is voor \(\sigma^2\).

Een alternatief is om te veronderstellen dat de populatie verdeeld kan worden in strata \(h\) \((h = 1, 2, \ldots, H)\) met \(n_h\) waarnemingen in stratum \(h\), en dat voor stratum \(h\) geldt: \(E(\varepsilon_h) = 0\) en \(\text{Var}(\varepsilon_h) = \sigma^2_h I_{n_h}\). Als \(X\) nu gepartitioneerd wordt als \(X = (X'_1, X'_2, \ldots, X'_H)'\) dan kunnen we schrijven

\begin{equation}
\text{Var}(b|S) = \text{Var}[(X'X)^{-1}X'\varepsilon] = (X'X)^{-1}[\sum_n X'_n \text{Var}(\varepsilon_h)X_h](X'X)^{-1}
\end{equation}
\[ - (X'X)^{-1} \left( \sum_h X_h' X_h \sigma_h^2 \right) (X'X)^{-1}. \] (9)

Een schatter voor bovenstaande covariantiematrix wordt verkregen door voor \( \sigma_h^2 \) de steekproefgrootheid \( s_h^2 = \sum_{i \in h} (y_i - \hat{y}_i)^2 / n_h \) te substitueren.

**Variantie van \( \Gamma' \)**

Om de tweede term in (5) te schatten kunnen \( \beta_k \) en \( \beta_f \) geschat worden met \( b_k \) en \( b_f \). Daarnaast is nog een schatter nodig voor de \( K \times K \) matrix \( \Gamma[\text{Var}(x_k, x_f)]\Gamma' \): de varianties en covarianties van de met \( \Gamma \) gewogen sommen van de \( x \)-variabelen. Voor deze elementen geldt

\[ \Gamma[\text{Var}(x_k, x_f)]\Gamma' = \Gamma E_S [(x_k - E_S x_k)(x_f - E_S x_f)'] \Gamma' \]
\[ - \gamma \Gamma E_S (x_k x_f) - (E_S x_k)(E_S x_f)] \Gamma'. \] (10)

Voor een gestratificeerde steekproef kunnen de elementen tussen rechte haken geschat worden door de steekproefgrootheid

\[ \sum_h (\overline{x}_{hk} \overline{x}_{h'k} - \overline{x}_{hk} \overline{x}_{h'k}) \] (11)

met \( \overline{x}_{hk} = \sum_{i \in h} x_{ik} / n_h \) het kruiselings tweede-orde steekproefmoment in stratum \( h \) en \( \overline{x}_{hk} = \sum_{i \in h} x_{ik} / n_h \) het steekproefgemiddelde in stratum \( h \).

In het geval dat \( x_k \) de indicator-variabele is voor stratum \( h \), geldt:

\[ x_{ik} = 1, \; i \in h ; \; x_{ik} = 0, \; i \notin h ; \] (12)

\[ \overline{x}_{hk} = 1 ; \; \overline{x}_{hk'} = \overline{x}_{h'k}, \] (13)

zodat de variantie- en covariantietermen die corresponderen met een stratificatie-variabele nul zijn.
Relatie met directe schatter voor een gestratificeerde steekproef.

Als alle verklarende variabelen stratum-indicatoren zijn, is de $H \times H$ matrix $X'X$ een diagonaalmatrix met als diagonaal-elementen de steekproefaantallen per stratum, zeg $D(n_h)$. De kolom-vector $X'y$ bevat dan de steekproeftotalen $y_h$ van $y$ per stratum. De geschatte OLS coëfficiënt $b = (X'X)^{-1}X'y$ is dan de $H$-vector met elementen $y_h/n_h$ en de 'fitted value' is

$$\hat{y}_i = X'_i b = y_{h(i)}/n_{h(i)},$$

waarin $h(i)$ het stratum is waartoe individu $i$ behoort. Voor de lineaire combinatie $\Gamma \hat{y}$ geldt dan

$$\Gamma \hat{y} = \sum_h y_h/n_h \gamma_h,$$

waarin $\gamma_h$ de $q$-vector is met als elementen $\gamma_{h,j} = \sum_{i \in h} \gamma_{i,j}$. Omdat $\text{Var}(x_k, x_l) = 0$ in het geval dat $x_k$ en $x_l$ stratum-indicatoren zijn, is $\text{Var}(\Gamma \hat{y})$ volgens (5), (7) en (9)

$$\text{Var}(\Gamma \hat{y}) = \Gamma X(X'X)^{-1} \left[ \sum_h n_h \gamma_h \sigma_h^2 \right] (X'X)^{-1} X' \Gamma'$$

$$= \Gamma XD(n_h)^{-1} D(n_h \sigma_h^2) D(n_h)^{-1} X' \Gamma'$$

$$= \Gamma XD(n_h \sigma_h^2 / n_h) X' \Gamma' = \Gamma B(1_{n_h}, 1_{n_h} \sigma_h^2 / n_h) \Gamma',$$

waarbij $B(1_{n_h}, 1_{n_h} \sigma_h^2 / n_h)$ een blok-diagonale matrix met langs de diagonaal de matrices $1_{n_h} 1_{n_h} \sigma_h^2 / n_h$ en $1_{n_h}$ een $n_h$-vector met enen. Partitioneren van $\Gamma$ als $\Gamma = (\Gamma_1, \ldots, \Gamma_h, \ldots, \Gamma_H)$, waarin $\Gamma_h$ de kolommen van $\Gamma$ die betrekking hebben op stratum $h$, bevat, geeft

$$\text{Var}(\Gamma \hat{y}) = \sum_h \sigma_h^2 / n_h \Gamma_h 1_{n_h} 1_{n_h} \Gamma_h.$$

Stel nu dat $q = 1$, zodat $\Gamma$ een rijvector is, zeg $\gamma$, en dat de elementen van deze rijvector gelijk zijn aan de inverse van de insluitkansen per stratum, zodat $\gamma_i = N_h / n_h$ voor $i \in h$, met $N_h$ het aantal populatie-elementen in stratum $h$, dan gaat (15) over in de Horwitz-Thompson schatter $\gamma^{HT}$ voor het populatie-totaal.
\[ \gamma^T = \sum_h n_h y_h / n_h, \]  

(18)

en, omdat \( \Gamma_h 1_{n_h} = \sum i e_{hi} y_i = n_h, \) (17) in de variantie van \( \gamma^T \)

\[ \text{Var}(\gamma^T) = \sum_h n_h^2 \sigma^2 / n_h. \]

(19)

4. Toepassing op een budgetverdeelmodel


\[ w_{ig} = x_i^T \beta_g + u_{ig}, \quad i = 1, 2, \ldots, n, \quad g = 1, 2, \ldots, G, \]

(20)

waarin de index \( i \) het huishouden aangeeft, de index \( g \) het goed, \( n \) het aantal huishoudens, \( G \) het aantal goederen, \( w_{ig} \) het budgetaanhevel, \( x_i \) een K-vector met verklarende variabelen, \( \beta_g \) een K-vector met parameters en \( u_{ig} \) een storingsterm; per definitie geldt \( e_{ig} = e_i / e_i \) met \( e_i = \sum e_{ig} \) de totale bestedingen door huishouden \( i \). Er wordt aangenomen dat \( E(u_{ig}) = 0, \text{ Cov}(u_{ig}, u_{jh}) = \sigma_{gh} \) en \( \text{Cov}(u_{ig}, u_{ij}) = 0 \) (\( i \neq j \)). Omdat per definitie \( \sum g w_{ig} = 1 \), geldt \( \sum e_{ig} x_i^T \beta_g = 1 \) en \( \sum u_{ig} = 0 \); dit zijn de zogenaamde 'adding-up'-restricties. Omdat per huishouden in (20) voor elk goed de verklarende variabelen dezelfde zijn, geeft gewone kleinste kwadraten dezelfde schattingen als gegeneraliseerde kleinste kwadraten (Theil, 1971, blz. 309-10) en is automatisch aan de 'adding-up'-restricties voldaan (Cramer, 1986, § 7.3). De covariantie-matrix van de OLS-schatter \( \hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_G) \) is

\[ \text{Var}(\hat{\beta}) = \Sigma \otimes (X'X)^{-1}, \]

(21)

waarin \( \Sigma = (\sigma^2_{gh}) \) en \( X = (x_1, x_2, \ldots, x_n)' \). In het hier beschouwde geval met gelijke verklarende variabelen per goed is de kleinste-kwadraten-
schatter zuiver, zodat de analyse van de vorige paragrafen toegepast kan worden.

Door Boon en De Haan (1995) wordt dit model gebruikt om synthetische schatters voor budgetpatronen van kleine groepen huishoudens te bepalen. Het model wordt voor de gehele steekproef geschat, waarna voor een deel-groep de bestedingen geschat worden als de gefitte waarden gewogen met de ophoogfactoren, namelijk $e_{Mg} = \sum_i \theta_i \hat{e}_{ig} = \sum_i \theta_i e_i \hat{u}_{ig}$, en $\theta_i$ de ophoogfactor voor huishouden $i$ is en de sommaties over alle huishoudens in de groep lopen. Boon en De Haan (1995) zijn nu geïnteresseerd in de covariantie tussen de gefitte waarden van de bestedingen per good gewogen met de ophoogfactoren, namelijk $\text{Cov}(e_{Mg}, e_{Mh})$. Deze kunnen als volgt bepaald worden. Model (20) kan geschreven worden als

$$e_{ig} = z_i' \hat{\beta}_g + \epsilon_{ig}. \tag{22}$$

waarin $z_i = x_i e_i$ en $\epsilon_{ig} = u_{ig} e_i$. Merk op dat $e_i$ binnen het model exogene is verondersteld, zodat ook $z_i$ binnen het model exogene is. Derhalve

$$e_{Mg} = \sum_i \theta_i \hat{e}_{ig} = \sum_i \theta_i z_i' \hat{\beta}_g = z_M' \hat{\beta}_g = \sum_k \hat{\beta}_{gk} z_{Mk}. \tag{23}$$

waarin $z_M = \sum_i \theta_i z_i$ de $K$-vector met de geaggregeerde verklarende variabelen is. Uit (23) volgt

$$\text{Cov}(e_{Mg}, e_{Mh}) = \sum_k \sum_l \text{Cov}(\hat{\beta}_{gk} z_{Mk}, \hat{\beta}_{hl} z_{Ml}). \tag{24}$$

Voor vectoren $a$ en $b$ die beide uit 2 elementen bestaan, kan de variantie-analyse-regel (2) geschreven worden als

$$\text{Cov}(a_1, a_2) = E[\text{Cov}(a_1, a_2 | b_1, b_2)] + \text{Cov}[E(a_1 | b_1, b_2), E(a_2 | b_1, b_2)] \tag{25}$$

Toepassen op (24) met $a_1 = \hat{\beta}_{gk} z_{Mk}, a_2 = \hat{\beta}_{hl} z_{Ml}, b_1 = z_{Mk},$ en $b_2 = z_{Ml},$ geeft

$$\text{Cov}(e_{Mg}, e_{Mh}) = \sum_k \sum_l E[\text{Cov}(\hat{\beta}_{gk} z_{Mk}, \hat{\beta}_{hl} z_{Ml} | z_{Mk}, z_{Ml})]$$

$$+ \sum_k \sum_l \text{Cov}(E(\hat{\beta}_{gk} z_{Mk} | z_{Mk}), E(\hat{\beta}_{hl} z_{Ml} | z_{Ml}))$$
\[
= \sum_k \sum_l \mathbb{E}(z_{M_k}z_{M_l} | \text{Cov}(\hat{\beta}_{g,k}, \hat{\beta}_{g,l} | z_{M_k}, z_{M_l})) + \sum_k \sum_l \beta_{g,k}\beta_{g,l} \text{Cov}(z_{M_k}, z_{M_l}).
\]

(26)

Een zuivere schatting voor de eerste term in (26) is

\[
\sum_k \sum_l z_{M_k}z_{M_l} | \text{Cov}(\hat{\beta}_{g,k}, \hat{\beta}_{g,l} | z_{M_k}, z_{M_l}).
\]

(27)

Een consistent schatting van de term tussen rechte haken in (27), de modelcovariantie van de parameters \(\text{Cov}(\hat{\beta}_{g,k}, \hat{\beta}_{g,l} | z_{M_k}, z_{M_l})\), kan verkregen worden door \(\Sigma\) te schatten op basis van de residuen. De tweede term in (26), de covariantie \(\text{Cov}(z_{M_k}, z_{M_l})\) tussen de verklarende variabelen, kan geschat worden op basis van het steekproefontwerp.

Het probleem van Boon en De Haan kan ook in de matrix-notatie van de vorige paragrafen gebracht worden. Model (22) kan geschreven worden als

\[
e_\delta = Z\beta_\delta + \epsilon_\delta,
\]

(28)

waarin \(e_\delta = (e_{1,\delta}, e_{2,\delta}, \ldots, e_{n,\delta})'\) de \(n\)-vector met de bestedingen aan goed \(g\) is, \(Z = (z_1, z_2, \ldots, z_n)'\) de \((n \times K)\)-matrix met de verklarende variabelen, en \(\epsilon_\delta = (\epsilon_{1,\delta}, \epsilon_{2,\delta}, \ldots, \epsilon_{n,\delta})'\) de \(n\)-vector met de storingstermen. Deze modellen per goed kunnen geschreven worden als

\[
y = Z\beta + \epsilon,
\]

(29)

waarin \(y = (e_1', e_2', \ldots, e_G')'\) de \(G\)-vector met de bestedingen is, \(Z = I_G \otimes Z\) de blokdiaogionale matrix met de verklarende variabelen, \(\beta = (\beta_1', \beta_2', \ldots, \beta_G')'\) de vector met parameters, en \(\epsilon = (\epsilon_1', \epsilon_2', \ldots, \epsilon_G')'\) de vector met de storingstermen. De gezoche covarianties zijn nu de elementen van de matrix \(\text{Var}(\Gamma \hat{y})\), waarin \(\Gamma = I_G \otimes \theta'\). Met vergelijking (2) volgt dan

\[
\text{Var}(\Gamma \hat{y}) = E_Z[\text{Var}(\Gamma Z b | Z)] + \text{Var}_Z(\Gamma Z \beta).
\]

(30)

Er geldt \(\Gamma Z = (I_G \otimes \theta' Z) = (I_G \otimes z_\delta')\) en \(\beta = \text{vec}(B)\) met \(B = (\beta_1, \beta_2, \ldots, \beta_G)\) de \((K \times G)\)-matrix met de modelpareters. Derhalve \(\Gamma Z \beta = (I_G \otimes \theta' Z)\beta = B'z_M\), zodat.
\[ \text{Var}(\hat{\Gamma}) = E_Z[\text{Var}(\hat{B}'z_M|Z)] + B'[\text{Var}_Z(z_M)]B, \quad (31) \]

wat overeenkomt met (26). Met behulp van (21) kan de eerste term van (31) vereenvoudigd worden tot

\[ E_Z[(\text{Var}(\hat{B}'z_M|Z)] = E_Z(I_G \otimes z_M')[\text{Var}(\hat{\beta}|Z)](I_G \otimes z_M) \]

\[ = \{E_Z[z_M'(X'X)^{-1}z_M]\} \Sigma; \quad (32) \]

merk op dat \( E_Z[z_M'(X'X)^{-1}z_M] \) een scalar is, waarvoor \( z_M'(X'X)^{-1}z_M \) een zuivere schatter is.

**Literatuur**

