Blanda modellar i R

Jorunn Slagstad

Universitetet i Bergen

20. desember 2006

Regresjon

Regresjon: ser på endring i ein respons Y for bestemte verdiar av ein variabel X:

$$E(Y) = \beta_0 + \beta \cdot X$$

der

- Y kallast responsvariabel
- X kallast forklaringsvariablel
- β_0 og β kallast parametrar
- **Lineær regresjon:** lineært forholdet mellom *responsen* og *parametrane* til forklaringsvariablane

Eit datasett

Datasettet Orthodont:

- Ser på endring i ortopedisk avstand hos barn i bestemte aldrar.
- Observasjonar av 27 barn, og 4 målingar av avstanden for kvart barn.
- Avstand er responsvariabel,
- og alder og kjønn mulige forklaringsvariablar.

Multivariat regresjonsmodell for Orthodont

På matriseform lik:

$$\begin{bmatrix} y_{1i} \\ y_{2i} \\ y_{3i} \\ y_{4i} \end{bmatrix} = \begin{bmatrix} 1 & 8 & 1_i & 8 \cdot 1_i \\ 1 & 10 & 1_i & 10 \cdot 1_i \\ 1 & 12 & 1_i & 12 \cdot 1_i \\ 1 & 14 & 1_i & 14 \cdot 1_i \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \epsilon_{1i} \\ \epsilon_{2i} \\ \epsilon_{3i} \\ \epsilon_{4i} \end{bmatrix}.$$

- Dette er ein *multivariat regresjonsmodell*.
- Her føl responsvektoren ei *multivariat normalfordeling*.
- $\epsilon_{1i}, \epsilon_{2i}, \ldots$ er støyledd, $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$.
- $E(\boldsymbol{y}_i) = \boldsymbol{X}_i \boldsymbol{\beta}$
- $\text{Var}(\boldsymbol{y}_i) = \text{Var}(\boldsymbol{\epsilon}_i)$

Korrelasjon

- Korleis gjere rede for korrelasjon mellom observasjonar på same individ?
- **Svar:** introdusere parametrar som varierar mellom individ.
- Vi vil då få ein blanda modell.

Blanda modellar

Blanda modell kan formulerast når:

- ein har fleire observasjonar på ulike *eksperimentelle einheitar*
  - eksperimentell einheit: element som medfører at vi forventar mindre variasjon for målingar i same einheit enn for målingar i ulike einheitar.
  - eksempel: eit individ, ei risikogruppe (bilmerke, yrkesgruppe), eit geografisk område
- ein antar at nokre av regresjonsparametrane i modellen varierer mellom einheitar.

Definisjon av ein lineær blanda modell

Definisjon (Ein lineær blanda modell)

I symbol:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n, \tag{1}$$

- $\boldsymbol{y}_i$ *inneheld* p *observasjonar av einheit* i
- $\boldsymbol{Z}_i$ *matrise med delmengd av elementa i matrisa* $\boldsymbol{X}_i$
- $\boldsymbol{b}_i$ *vektor med variable parametrar, der* $\boldsymbol{b}_i \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{\Psi})$
- $\boldsymbol{\epsilon}_i$ *vektor med støyledd, der* $\boldsymbol{\epsilon}_i \sim \mathcal{N}_p(\boldsymbol{0}, \sigma^2\boldsymbol{I})$
- $p \geq q$

Ein random intercept-modell for Orthodont-dataa

Døme Éin variabel parameter:

- Modell med variabel startverdi, legg til variabel parameter b_i:

$$\boldsymbol{y}_i = \boldsymbol{X\beta} + b_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, 27.$$

- der observasjonar på same individ i er korrelerte,
- observasjonar på ulike individ er uavhengige,
- og $b_i \sim \mathcal{N}(0, \sigma_b^2)$
- $b_1 = b_2 = \cdots = 0 \Longrightarrow \sigma_b^2 = 0$. Er $\sigma_b^2 > 0$?
- $E(\boldsymbol{y}_i) = \boldsymbol{X\beta}$, mens

$$\text{Cov}(y_{ij}, y_{ik}) = \begin{cases} \text{Var}(y_{ij}) = \sigma^2 + \sigma_b^2 & \text{dersom} \quad j = k \\ \text{Var}(b_i) = \sigma_b^2 & \text{dersom} \quad i \neq k \end{cases} \tag{2}$$

Blanda modell med to variable parametrar

Døme Variabelt konstantleddet òg stigningstal i modell:

$$\boldsymbol{y}_i = \boldsymbol{X}\boldsymbol{\beta} + \begin{bmatrix} 1 & 8 \\ 1 & 10 \\ 1 & 12 \\ 1 & 14 \end{bmatrix} \begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} + \boldsymbol{\epsilon}_i$$

der

- $\boldsymbol{b}_i \sim \mathcal{N}_2(\boldsymbol{0}, \boldsymbol{\Psi})$,
- og

$$\boldsymbol{\Psi} = \begin{bmatrix} \psi_1 & \psi_{12} \\ \psi_{21} & \psi_2 \end{bmatrix}$$

- er ei symmetrisk, positivt definitt matrise,
- og $\psi_2 = 0 \Longrightarrow \boldsymbol{\Psi} = \psi_1 = \sigma_b^2$.

Generalisert lineær modell

Eksponensialfamilien:

$$f(y) = \exp\left\{\frac{y\theta + b(\theta)}{\phi} + c(y, \phi)\right\},$$

der

- $b(.)$ og $c(.)$ er kjente funksjonar,
- θ kallast kanonisk parameter (eller naturleg parameter)
- ϕ kallast skaleringsparameter
- Forventning:
$$\mu = E(y) = b'(\theta),$$
- Varians:
$$\text{Var}(y) = \phi b''(\theta) = \phi V(\mu)$$

Ein GLM

Generalisert lineær modell, GLM, er definert som:

$$g(\mu_i) = g(E(y_i)) = \boldsymbol{X}_i\boldsymbol{\beta} \quad i = 1, \ldots, n$$

Funksjonen g er linken mellom den forventa responsvektor og lineær prediktor: $g(\mu_i) = \eta_i = \boldsymbol{X}_i\boldsymbol{\beta}$.
Ein GLM er aktuell for forsikringsdata:

- ikkje-normalfordelte observasjonar
  - observasjonar av storleik av krav (log-normalfordeling)
  - tal på krav i ulike poliser (Poisson fordeling)
  - ventetider mellom krav (eksponensialfordeling)
- modellere ein transformasjon av responsen

Kva dersom vi har repeterte målingar?

Korleis kan vi utvide ein GLM til å handtere korrelerte observasjonar?

- **Svar:** Ein generalisert lineær blanda modell, forkorta GLMM!

Generalisert lineær blanda modell

Definisjon (Ein GLMM)

Ein generalisert lineær modell, forkorta GLMM, har forma:

$$g(E(\boldsymbol{y}_i \mid \boldsymbol{b}_i)) = \boldsymbol{\eta}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i, \quad i = 1, \ldots, n,$$

der

- *den betinga fordelinga til responsen er ei fordeling som er medlem av eksponensialfamilien,*
- *komponentane i den lineære prediktoren er både faste og variable,*
- *g er link-funksjon,*
- *og dei variable parametrane antas å ha ei multivariat normalfordeling:*

$$\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}) \quad i = 1, \ldots, n. \tag{3}$$

Aktuelt datasett

Datasett med 931 observasjonar av tal på krav i 133 ulike risikogrupper.

- National Council on Compensation Insurance (New York).
- Klugman (1992).
- Registrert over ein sjuårsperiode.
- Responsvariabel: talet på krav i dei ulike risikogruppene.
- Forklaringsvariablar: år, lønningslister (Payroll).
- Eksperimentell einheit: risikogruppe (yrkesgruppe).

Modell for talet på krav

[Antonio og Beirlant (2006)] har definert og analysert to GLMM-ar for datasettet der

- talet på krav er Poisson fordelt:

$$\boldsymbol{y}_i \mid \boldsymbol{b}_i \sim \text{Poisson}(\boldsymbol{\mu}_i).$$

- logfunksjonen er link-funksjon,
- lønningslista (Payroll) er ein *offset*-parameter,
- og talet på krav aukar med åra.

Modellformulering

[Antonio og Beirlant (2006)] sine modellar:

- Modell med **éin variabel parameter** (random intercept-modell):

$$\log(\boldsymbol{\mu_i}) = \log(\text{Payroll})_i + \beta_0 + \text{Year}_i \cdot \beta_1 + b_i$$
$$b_i \sim \mathcal{N}(0, \sigma_b^2)$$

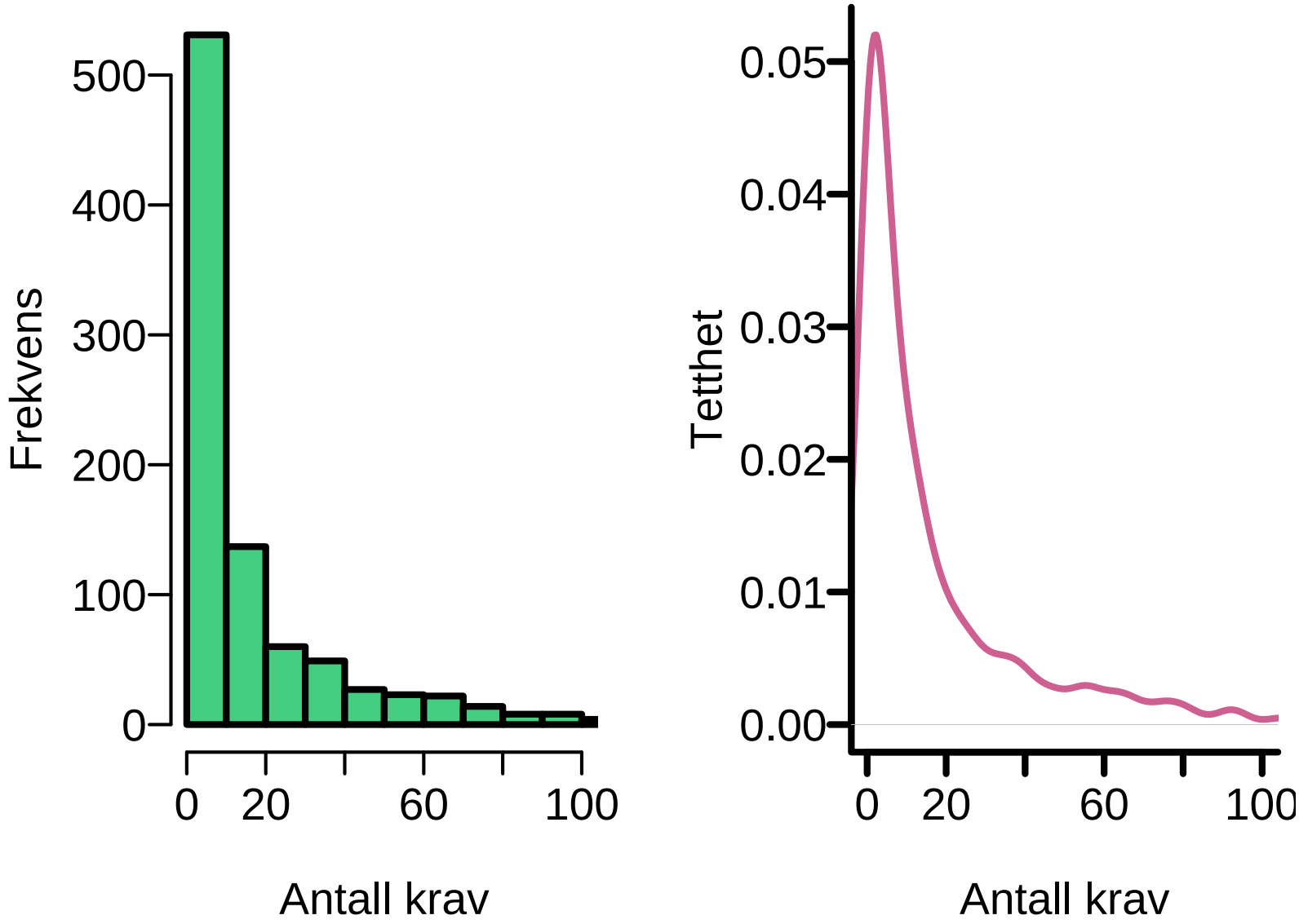
- Fann at $\sigma_b^2 > 0$.
- Modell med **to variable parametrar**:

$$\log(\boldsymbol{\mu_i}) = \log(\text{Payroll})_i + \beta_0 + \text{Year}_i \cdot \beta_1 + b_{i0} + \text{Year}_i \cdot b_{i1}$$
$$\boldsymbol{b}_i \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Psi}), \quad \boldsymbol{\Psi} = \begin{bmatrix} \psi_1 & \psi_{12} \\ \psi_{21} & \psi_2 \end{bmatrix}$$

- Undersøke om $\psi_2 > 0$.

Glatta histogram av talet på krav

Er $E(\boldsymbol{y}_i) = \text{Var}(\boldsymbol{y}_i)$ oppfylt?



Overdispersjon

- Overdispersjon: $E(Y_i) < \text{Var}(Y_i) \Longrightarrow$ forventningsskjeive estimat i modell .

År	1	2	3	4	5	6	7
Snitt	14,9	16,2	16,6	17,5	22,8	17,3	16,7
Varians	690,5	624,6	632,1	790,2	1181,8	720,3	649,1
Høve	46,3	38,5	38,0	45,1	51,8	41,6	38,8

- Klugman-dataa er overdisperserte!
- Årsak: uobserverbare faktorar.
- Korleis gjere rede for overdispersjon?
  1. Anta ei anna fordeling for responsen.
  2. Anta ein underliggande prosess som produserar fleire nullar enn konsistent med Poisson fordelinga.

Negativ binomisk fordeling

Negativ binomisk fordeling:

$$P(Y_i = y_i, \alpha, \mu_i) = \frac{\Gamma(y_i + \alpha)}{\Gamma(\alpha) y_i!} \left(\frac{\alpha}{\mu_i + \alpha}\right)^{\alpha} \left(\frac{\mu_i}{\mu_i + \alpha}\right)^{y_i} \quad y_i = 0,1,2,\ldots \tag{4}$$

- Forventning: $E(Y_i) = \mu_i$
- Varians: $\text{Var}\, Y_i = \mu_i + \frac{1}{\alpha}\mu_i^2$
- Ingen overdispersjon: $\alpha = \infty \Longrightarrow \text{Var}\, Y_i = \mu_i$
- Negativ binomisk GLMM:

$$\boldsymbol{y}_i \mid \boldsymbol{b}_i \sim \text{nbin}(\alpha, \boldsymbol{\mu}_i)$$

Samanlikning av modellar

Ynskjer å undersøke om:

- ei **Poisson fordeling** eller ei **negativ binomisk fordeling** høver best
- éin eller to variable parametrar i modell
- får eg samme resultat som [Antonio og Beirlant (2006)]?

Dataverktøy i R:

- [Pinheiro og Bates (2000)] sin `lmer`-funksjon (Poisson-modellar)
- [Skaug *et al.* (2006)Skaug, Fournier og Nielsen] sin `glmm.admb`-funksjon (Poisson- òg nb-modellar)

Resultat av analyser

Resultata mine:

	$\lvert \boldsymbol{b}_i \rvert$	$\lvert \boldsymbol{\theta} \rvert$	logLik	AIC	`easyFlag`
Poisson GLMM	1	3	-2521.83	5049.7	F
Neg. bin. GLMM	1	4	-2245.02	4898.0	F
Poisson GLMM	2	5	-2521.84	5053.7	F
Neg. bin. GLMM	2	6	-2437.0	4886.0	T

Modell for Klugman-dataa

Kva for modell passar **best**?

- Resultat ved `glmm.admb` høver betre til [Antonio og Beirlant (2006)] sine, enn resultat ved `lmer`.
- «Sikraste» slutning eg kan ta $\Longrightarrow$ modell med éin variabel parameter og negativ binomisk fordeling best.

Eit randproblem

I mi oppgåve: Samanlikning av blanda modellar kan formulerast i hypotesa

$$H_0 : \psi_2 = 0 \quad \text{mot} \quad H_1 : \psi_2 \geq 0$$

- $\mathbf{\Psi}$ positivt definitt $\Longleftrightarrow \psi_2 > 0$.
- ψ_2 på randa under H_0.
- Effekt på testobservatoren vår som er likelihood ratio.

Likelihood ratio

Likelihood ratio-observator:

$$\mathbf{\Lambda} = -2(\log L(\boldsymbol{\theta_0} \mid \text{data}) - \log L(\boldsymbol{\theta_1} \mid \text{data}))$$

Generell teori:

- $\mathbf{\Lambda} \sim \chi^2(p - q)$
- p = dimensjon av $\boldsymbol{\theta}_1$ (# parametrar i generell modell)
- q = dimensjon av $\boldsymbol{\theta}_0$ (# parametrar i spesifikk modell)

Orthodont-modellar:

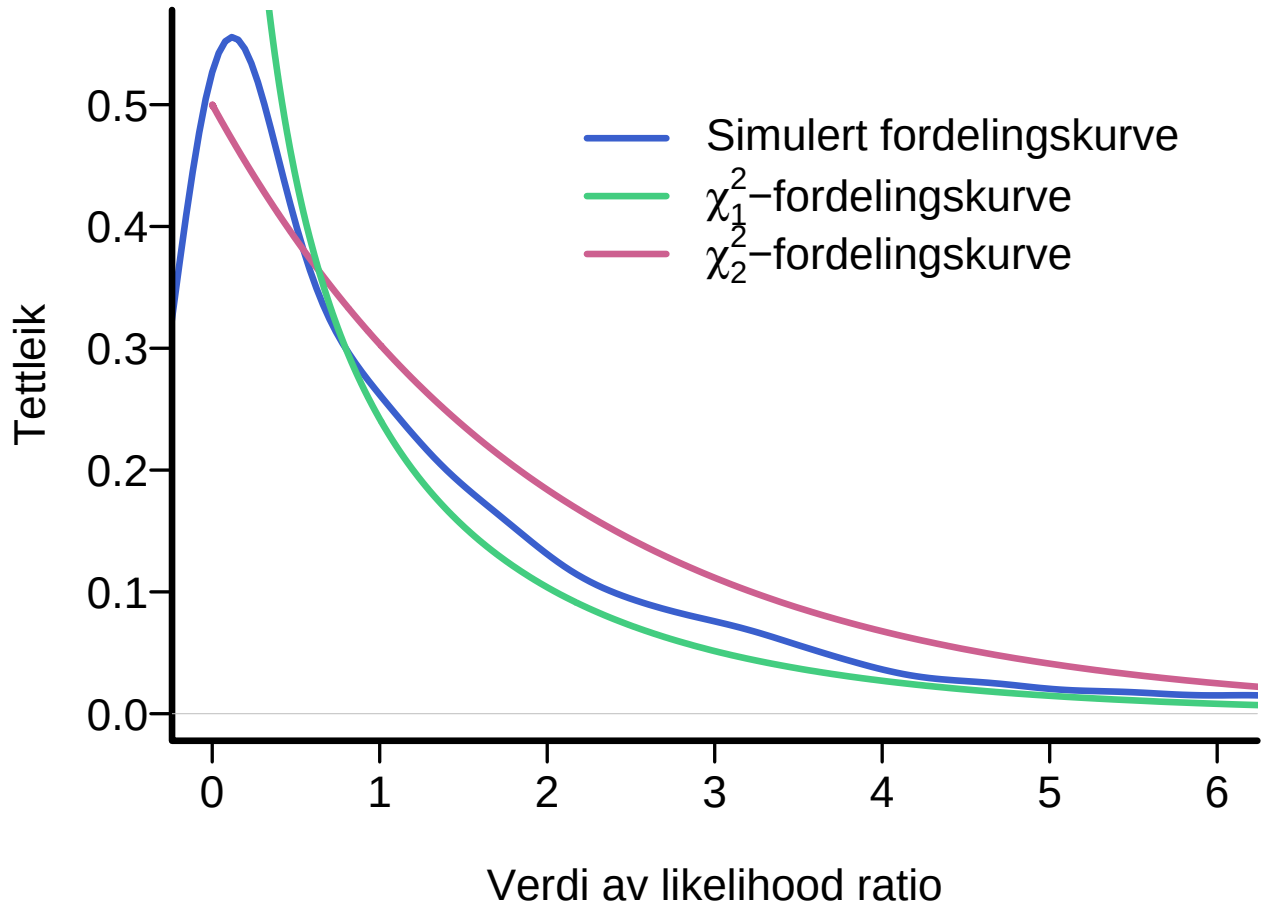
- $\mathbf{\Lambda} \sim \chi^2_2$.

Simulering av likelihood ratio

Er simulerte verdiar χ^2_2-variablar?

- Plottar eit glatta histogram av verdiane,
- og eit glatta histogram av verdiar frå ei χ^2_2-fordeling,
- og eit glatta histogram av verdiar frå ei χ^2_1-fordeling.

Plott



- Kurva fell i mellom dei to fordelingskurvene.

Stram og Lee

[Stram og Lee (1994)] utførte ei liknande simulering for Orthodont-dataa. Deira teori:

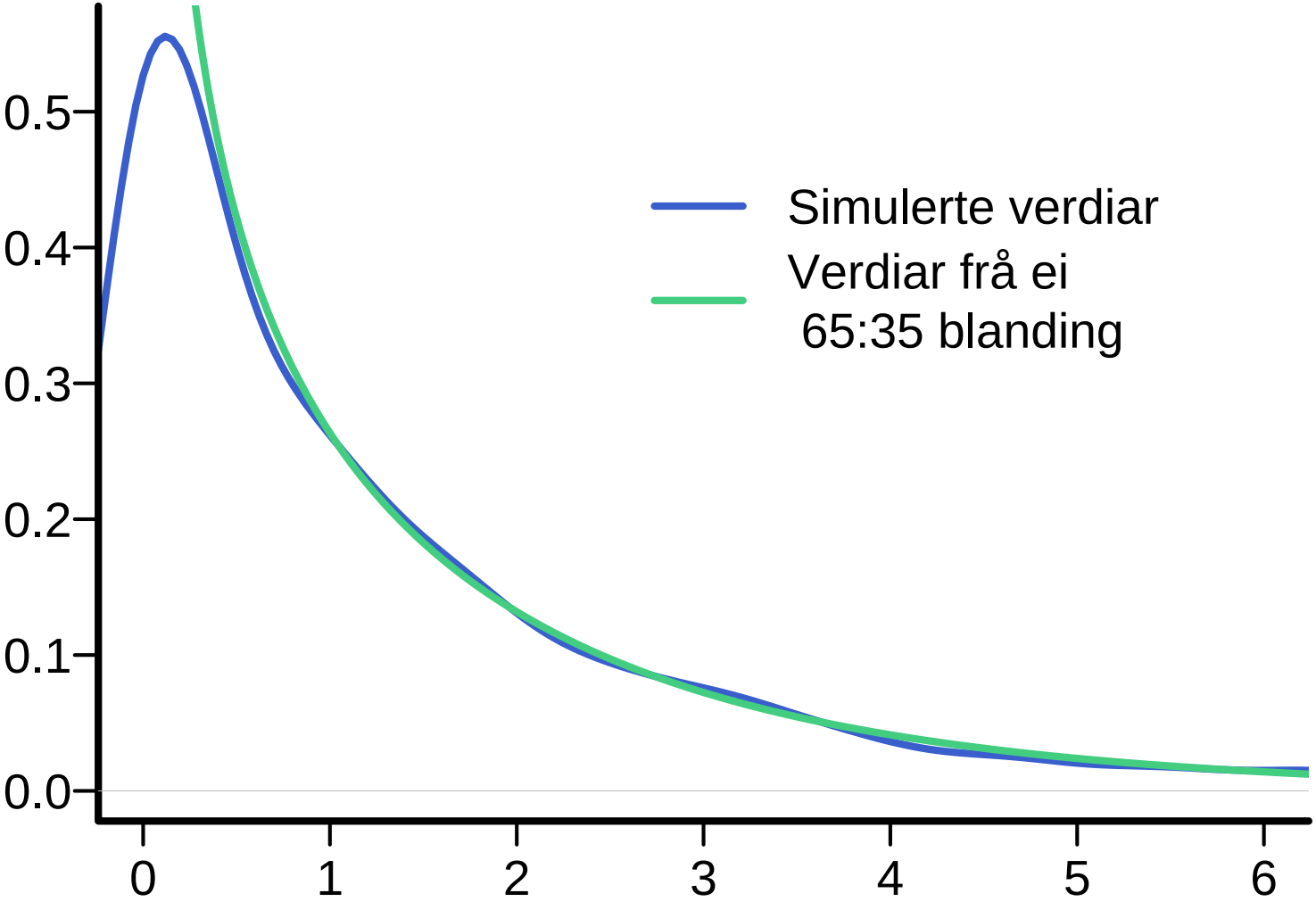
- $\mathbf{\Lambda} \sim 0{,}5\chi^2_k + 0{,}5\chi^2_{k+1}$
- $k = |\,\boldsymbol{\theta}_0\,|$ (# variable parametrar til modellen under H_0

For Orthodont-modellar:

- $\mathbf{\Lambda} \sim 0{,}5\chi^2_1 + 0{,}5\chi^2_2$.
- Stemmer dette godt med plottet?
- Ja, mogelingeins ei litt tyngre vekt på χ^2_1-fordelinga.

Mine resultat

Eg plottar dei simulerte verdiane med «nye» vekter inspirert av [Pinheiro og Bates (2000)]:



Analysar av blanda modellar

For **ordinære** blanda modellar:

- Testobservatoren $\mathbf{\Lambda}$ består av verdiar av element på randa av definisjonsområdet sitt.
- Medfører avvik frå generell teori.
- $\mathbf{\Lambda}$ fordelt ved generell teori gir konservative slutningar.

Og for **GLMM**-ar:

- Ein GLMM gjev fleire mogelegheitar for fordelinga til responsen.
- Gode for longitudinelle skadeforsikringsdata.
- Analysar av GLMM kan vere vanskelege pga dei mange moglegheitane.

Takk

Takk for meg!

Jorunn Slagstad

Litteratur

Antonio K. og Beirlant J. (2006).
«Actuarial statistics with generalized linear mixed models».
`http://www.econ.kuleuven.be/public/NDBAE81/GLMMRevisionIME.pdf`.

Pinheiro J.C. og Bates D.M. (2000).
Mixed-Effects Models in S and S-PLUS.
Statistics and Computing. Springer.

Skaug H., Fournier D. og Nielsen A. (2006).
«glmmADMB: Generalized Linear Mixed Models using AD Model Builder».
`http://otter-rsch.com/admbre/examples/glmmadmb/glmmADMB.html`.

Stram D.O. og Lee J.W. (1994).
«Variance components testing in the longitudinal mixed effects model».
Biometrics, **volum 50**, side 1171–1177.