

Opgave 1.

Beschrijf een methode voor het schatten van de trend in een tijdreeks.

Opgave 2.

Het is een algemeen gegeven dat mannen in het algemeen langer zijn dan vrouwen. Ook wordt vaak gezegd dat het Nederlandse volk één van de langste volken ter wereld is. Om te onderzoeken welke invloed 'ras' op de lichaamslengte heeft, werd de lengte van 1000 willekeurig gekozen mensen, zowel mannen als vrouwen en van alle rassen, gemeten in meters. Daarbij werd genoteerd of de persoon van het mannelijke of vrouwelijke geslacht was en of de persoon tot het blanke, zwarte, Aziatische of een ander ras behoorde.

Definieer Y_i als de lichaamslengte van de i -de persoon.

a. Beschrijf een additief variantie-analyse model zoals beschreven staat in het diktaat voor dit probleem.

Een ander mogelijk model voor bovenstaand probleem is het volgende:

$$Y_i = \mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{3i} + \beta x_{4i} + e_i, \quad i = 1, \dots, 1000,$$

met $e_i, i = 1, \dots, 1000$, normaal verdeeld, $Ee_i = 0$ en $\text{Cov}(e_i, e_j) = 0$ als $i \neq j$ en $\text{Cov}(e_i, e_j) = \sigma^2$ als $i = j$. De variabelen x_{1i}, x_{2i}, x_{3i} en x_{4i} zijn hier als volgt gedefinieerd:

$$\begin{aligned} x_{1i} &= \begin{cases} 1 & \text{individu } i \text{ is van het blanke ras,} \\ 0 & \text{anders,} \end{cases} \\ x_{2i} &= \begin{cases} 1 & \text{individu } i \text{ is van het zwarte ras,} \\ 0 & \text{anders,} \end{cases} \\ x_{3i} &= \begin{cases} 1 & \text{individu } i \text{ is van het Aziatische ras,} \\ 0 & \text{anders,} \end{cases} \\ x_{4i} &= \begin{cases} 1 & \text{individu } i \text{ is van het mannelijke geslacht,} \\ 0 & \text{anders.} \end{cases} \end{aligned}$$

Merk op dat als individu i niet tot het blanke, zwarte of Aziatische ras hoort, dan geldt $x_{1i} = x_{2i} = x_{3i} = 0$.

- b. Wat is het verschil tussen bovenstaand model en het model dat je onder a. beschreven hebt?
- c. Het variantie-analyse model dat je onder a. beschreven hebt, bevat vier parameters die met de variabele ‘ras’ corresponderen. Zoals in het diktaat, nemen we voor dit model aan dat de som van deze vier parameters gelijk aan nul is. Waarom doen we dit en waarom veronderstellen we in bovenstaand model niet dat $\alpha_1 + \alpha_2 + \alpha_3 = 0$?
- d. Geef de kwalitatieve betekenis van een schatting van een parameter in bovenstaand model en van de parameters in je variantie-analyse model van onderdeel a.
- e. In bovenstaand model vinden we de schattingen $\hat{\mu} = 3.0, \hat{\alpha}_1 = 3.0, \hat{\alpha}_2 = 1.0, \hat{\alpha}_3 = -2.0, \hat{\beta} = 4.0$. Welke schattingen van de parameters verwacht je voor het variantie-analyse model van onderdeel a.?

Opgave 3.

Beschrijf hoe twee geneste gegeneraliseerde lineaire modellen met elkaar kunnen worden vergeleken.

Opgave 4.

Zij

$$Y_i = f(x_i, \theta) + \varepsilon_i, \quad i = 1, \dots, n$$

een niet-lineair regressiemodel met $Y_1, \dots, Y_n$ onafhankelijke stochastische responsvariabelen, x_i een k -dimensionale vector van verklarende variabelen behorende bij Y_i ($i = 1, \dots, n$), ε_i een stochastische meetfout met verwachting 0 en variantie σ^2 behorende bij Y_i ($i = 1, \dots, n$), θ een onbekende p -dimensionale parametervector en f een niet-lineaire functie van θ .

- a. Hoe is de kleinste kwadratenschatter $\hat{\theta}$ van θ gedefinieerd?
- b. Formuleer de drie stappen van het bootstrapschema dat nodig is om een bootstrap betrouwbaarheidsinterval met bij benadering 95% betrouwbaarheid voor θ_l gegeven door

$$\left[2\hat{\lambda} - \lambda_{((1-0.025)B)+1}^*, 2\hat{\lambda} - \lambda_{([0.025B]+1)}^* \right] \quad (1)$$

te verkrijgen. (Je kunt gebruiken dat $\tilde{\varepsilon}_i$ het gecentreerde residu is, d.w.z. $\tilde{\varepsilon}_i = \hat{\varepsilon}_i - 1/n \sum_{i=1}^n \hat{\varepsilon}_i$ met $\hat{\varepsilon}_i = R_i(\hat{\theta}) = Y_i - f(x_i, \hat{\theta})$.)

- c. Men heeft de beschikking over twee data sets waarvan men vermoedt dat hetzelfde niet-lineaire model van toepassing is. Het idee is dat de parameterwaarden voor de beide sets hetzelfde zijn, behalve die van θ_1 . Men bepaalt voor de beide data sets betrouwbaarheidsintervallen met bij benadering 95% betrouwbaarheid voor θ_1 . Men vindt voor de eerste en de tweede data set (6.32, 6.68), respectievelijk, (6.59, 6.19) als het betrouwbaarheidsinterval gebaseerd op de klassieke theorie, en (6.41, 6.80), respectievelijk, (6.83, 6.90) als het bootstrap betrouwbaarheidsinterval. Welke conclusie(s) zou je trekken op grond van deze intervallen? Beargumenteer je antwoord.

Opgave 5.

Hieronder worden enkele praktijksituaties geschetst. Kies voor elke situatie één of meer geschikte modellen. Kies hierbij uit: lineair regressiemodel, variantie-analyse model, niet-lineair model, gegeneraliseerd lineair model, tijdreeksmodel. Beargumenteer je antwoord. Beschrijf het gekozen model zo volledig mogelijk. Vermeld hierbij in elk geval wat de respons- en verklarende variabelen zijn.

- a. Bij een onderzoek naar drugsgebruik onder jongeren worden 1000 mensen tussen de 12 en 16 jaar ondervraagd. Naast het antwoord op de vraag of en zo ja hoeveel drugs (in grammen per maand) ze gebruiken, wordt de leeftijd, het geslacht en de woonplaats genoteerd. In eerste instantie is de onderzoeker geïnteresseerd in de afhankelijkheid tussen het al dan niet gebruiken van drugs en de leeftijd, het geslacht en de grootte van de woonplaats. De woonplaatsen worden hierbij op grootte ingedeeld (klein, middel, groot).
- b. Tot de schrik van de onderzoeker uit onderdeel a. zijn er meer 12-jarige drugsgebruikers dan zij had verwacht. Dit is een reden om de groep met 12-jarige drugsgebruikers nader te bestuderen. Met name wordt nu onderzocht hoe de hoeveelheid drugs die de 12-jarigen per maand tot zich nemen, afhangt van het geslacht en de grootte van de woonplaats. (Opnieuw is de grootte van de woonplaats gedefinieerd als in onderdeel a.)
- c. Om te onderzoeken om de hoeveel tijd en hoeveel van een narcotiserend middel het best zou moeten worden toegediend tijdens een operatie, wordt bij verschillende vrijwilligers een zelfde hoeveelheid van het middel geïnjecteerd en gedurende enige tijd daarna elk uur de hoeveelheid van het middel in het bloed gemeten. Het idee is dat de afname exponentieel is en dat de meetfouten relatief klein zijn. Er wordt een model opgesteld dat dient ter analyse van de gegevens van één persoon.

Normering:

1: 3	2.a: 4	3: 4	4.a: 1	5.a: 3
	2.b: 2		4.b: 2	5.b: 3
	2.c: 3		4.c: 4	5.c: 3
	2.d: 2			
	2.e: 2			

Eindcijfer tentamen = totaal/4 + 1