Examen Statistische modellen en data-analyse

3de bachelor wiskunde

22 augustus 2014

Vraag 1: Formele vragen

1. Stel $\mathbf{Y} \sim N_n(\mu \mathbf{1}_n; \mathbf{\Sigma})$ waarbij $\mathbf{\Sigma} = (1-\rho)\mathbf{I}_n + \rho \mathbf{J}_n$, met $\mathbf{1}_n$ de kolommatrix die op elk van de n rijen het getal 1 heeft staan en $\mathbf{J}_n = \mathbf{1}_n \mathbf{1}_n'$. Zijn het steekproefgemiddelde en de steekproefvariantie onafhankelijk? Toon aan.

2. Stel $(Y_1, Y_2)' \sim N(\mathbf{0}, \mathbf{I}_2)$ en definieer:

$$W_1 = a(Y_1 + Y_2)^2$$
$$W_2 = b(3Y_1 - 2Y_2)^2$$
$$W_3 = W_1 + W_2$$

   Zoek a en b opdat W_1 en W_2 allebei een χ^2-kwadraat verdeling volgen en ga na of W_3 dan eveneens een χ^2-verdeling volgt of niet.

3. Beschouw de volgende lineaire modellen die het effect van geslacht G en een confounder X op een uitkomst Y beschrijven, waarbij de error termen $\epsilon_\alpha, \epsilon_\beta, \epsilon_\gamma$ met varianties gelijk aan respectievelijk $\sigma^2_\alpha, \sigma^2_\beta, \sigma^2_\gamma$ telkens voldoen aan de standaard voorwaarden voor het lineaire model.

$$Y = \gamma_0 + \varepsilon_\gamma \tag{1}$$
$$Y = \alpha_0 + \alpha_1 G + \varepsilon_\alpha \tag{2}$$
$$Y = \beta_0 + \beta_1 G + \beta_2 X + \varepsilon_\beta \tag{3}$$

   We beschikken over 100 proefpersonen bestaande uit 10 mannen ($G = 0$ voor $i = 1, \ldots, 10$) die een lukrake steekproef vormen van de mannen in de doelpopulatie en de rest vrouwen ($G = 1$ voor $i = 11, \ldots, 100$) die eveneens een lukrake steekproef zijn uit het vrouwelijk deel van de doelpopulatie. De confounder X werd gecentreerd.

   (a) Schat de verwachtingswaarde van de uitkomst voor de ganse doelpopulatie waarin de man/vrouw verhouding 50/50 is, gebruik makend van de maximum likelihoodschatters van de parameters: eerst voor model (1) , dan voor model (2) en tenslotte voor model (3). Geef telkens aan of het gaat om een onvertekende schatter en waarom (niet).

   (b) Geef de verwachtingswaarde van de kleinstekwadratenschatters $\hat{\alpha}_1$ en $\hat{\beta}_1$.

(c) De variantie van $\hat{\alpha}_1$ is gelijk aan $\dfrac{\sigma_\alpha^2}{9}$.

Toon aan dat de variantie van $\hat{\beta}_1$ gelijk is aan $\dfrac{\sigma_\beta^2 \sum_{i=1}^{100} x_i^2}{9\sum_{i=1}^{100} x_i^2 - \left(\sum_{i=11}^{100} x_i\right)^2}$.

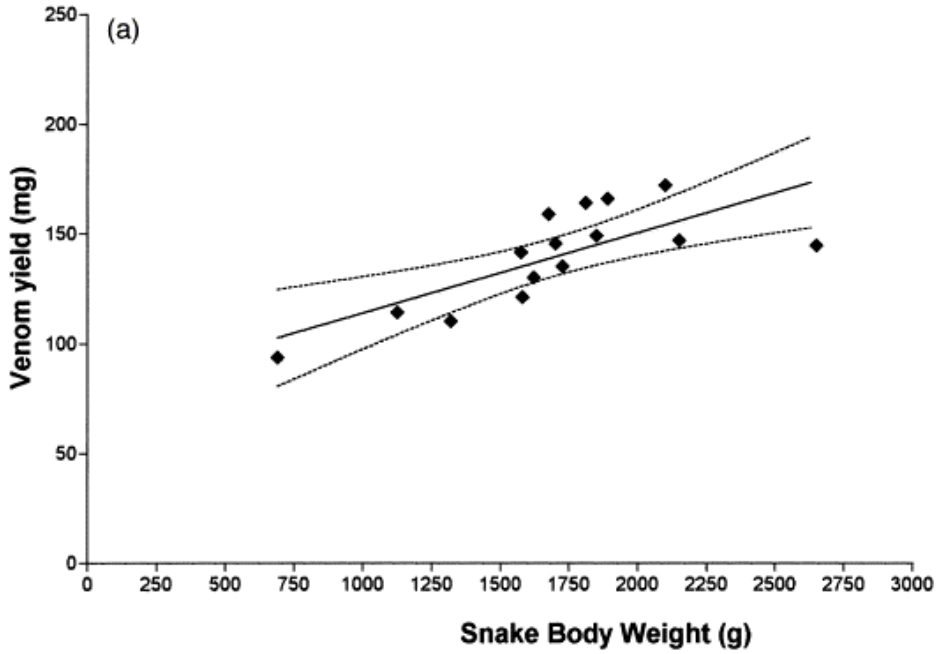
(d) Bespreek onder welk model de geschatte helling voor G de meest precieze schattingen genereert als $\beta_2 \neq 0$ en X en G 'onafhankelijk zijn in de dataset' zodat $\sum_{i=1}^{10} x_i = 0$.

U mag gebruiken dat:

$$\begin{pmatrix} a & b & c \\ b & d & e \\ c & e & f \end{pmatrix}^{-1} = \frac{1}{adf - ae^2 - b^2 f - c^2 d + 2bce} \begin{pmatrix} df - e^2 & ce - bf & be - cd \\ ce - bf & af - c^2 & bc - ae \\ be - cd & bc - ae & ad - b^2 \end{pmatrix}$$

Vraag 2: Giftige slangen door *de Roodt et al.*

In *de Roodt et al.* wordt de gifproductie van twee soorten groefkopadders bestudeerd als een functie van hun gewicht met het oog op de toekomstige gifoogst. De punten tonen het gewicht en de gifproductie voor individuele adders van elk van de twee soorten, respectievelijk, met telkens de kleinste kwadratenlijn en zijn puntsgewijze betrouwbaarheidsintervallen.



1. Eerst zullen we de gifproductie van de C. d. terrificus bestuderen, waarvoor de data getoond worden in bovenstaande figuur (a).

(a) Schrijf het theoretische eenvoudig regressiemodel neer voor de gifproductie van een C. d. terrificus slang in functie van haar lichaamsgewicht. Definieer alle parameters.

(b) Volgens de auteurs zijn de geschatte parameters voor dit model: helling=$0.0635 \pm 0.009(SE)$; intercept =$77.49 \pm 16.8(SE)$; $R^2 = 0.543$; $s = 15.73$. Controleer elk van deze cijfers. Gebruik hiervoor de informatie in de figuur en toon duidelijk aan hoe jij deze (of zeer gelijkaardige) cijfers zou verwerven.

(c) Geef de (theoretische) interpretatie van elk van deze cijfers en bespreek bondig de praktische relevantie van deze interpretatie.

(d) Geef alle veronderstellingen van het eenvoudige lineaire regressiemodel and bespreek hun plausibiliteit voor het model voor de data in figuur 1(a), gebaseerd op de informatie die tot nu toe gegeven werd.

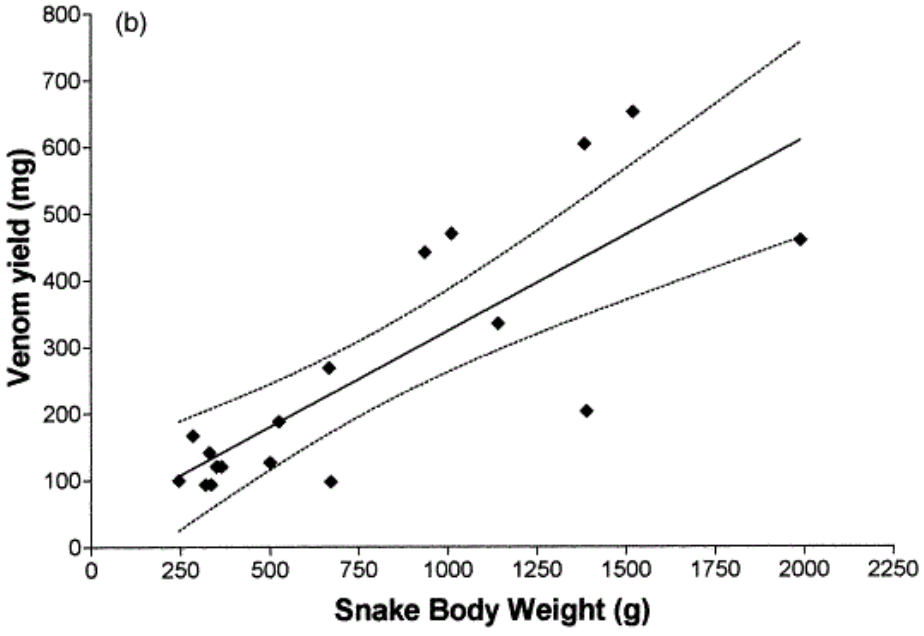
In wat volgt, mag je alle overeenkomstige modelveronderstellingen gebruiken alsof ze voldaan zijn.

(e) Beschrijf en test de nulhypothese (versus de alternatieve hypothese) dat gifproductie lineair onafhankelijk is van het lichaamsgewicht bij de C. d. terrificus.

(f) Bereken het 95% betrouwbaarheidsinterval voor de verwachte gifproductie van de C. d. terrificusslang met een lichaamsgewicht van twee kilogram.

(g) Verklaar de vorm van de lijnen van de betrouwbaarheidsintervallen in de figuur.

(h) Bereken een 95% predictie-interval voor de gifproductie van een C. d. terrificusslang met een lichaamsgewicht van twee kilogram. Vergelijk kort de resultaten van deze vraag en van vraag (f) en leg uit.



2. Bekijk figuur (b) nu goed. Hierna zullen we de gifproductie van zowel C. d. terrificus (figuur (a)) als B. alternatus (figuur (b)) bestuderen.

   (a) Beschrijf op basis van de informatie in beide figuren, een gepast theoretisch meervoudig lineair regressiemodel voor de gifproductie van een groefkopadder als een functie van zijn lichaamsgewicht en zijn soort. Definieer alle parameters.

   (b) Volgens de auteurs zijn de parameters voor het eenvoudig model van B. alternatus de volgende: helling=$0.2884 \pm 0.058(SE)$;
   intercept =$35.74 \pm 48.7(SE)$; $R^2 = 0.70$; $s = 113.1$. Schat de waarden van alle parameters in jouw meervoudig regressiemodel, gebaseerd op deze informatie en de informatie die eerder werd gegeven voor de C. d. terrificus.

   (c) Bespreek hoe je kan toetsen of de relatie tussen lichaamsgewicht en gifproductie dezelfde is voor beide soorten (berekeningen zijn niet nodig).

   (d) Lijst alle vooronderstellingen van het meervoudige lineaire regressiemodel op, en bespreek hun plausibiliteit voor het model in 2(a), gebaseerd op de gegeven informatie. Bespreek de problemen die te verwachten zijn bij de vooronderstellingen die niet voldaan zijn en stel een analysemethode voor om de meest belangrijke schending op te lossen.

   (e) Veronderstel dat je start van een eenvoudige ANOVA waarin de gifproductie voorgesteld wordt als een functie van de soort. Bespreek hoe een partiële residuenplot jou in staat stelt om op grafische wijze te bestuderen of het lichaamsgewicht ook zou moeten toegevoegd worden als voorspeller aan dit model. Welke trends verwacht je te zien? Leg uit!

   (f) Bij B. alternatus zijn de vrouwelijke slangen 3 keer zo groot als de mannelijke slangen van hun soort. Bij C. d. terrificus is deze verhouding slechts 1.2. Bespreek de implicaties van dit feit bij de interpretatie en het praktische gebruik van de bovenvermelde analyse.

Vraag 3: Het schatten van de menselijke leeftijd

In hun artikel uit 2010 'Estimating human age from T-cell DNA rearrangements' stellen Zubakov et al. een nieuwe methode voor om leeftijd te voorspellen op basis van bloedstalen. Ze belanden bij een eenvoudig lineair regressiemodel dat leeftijd voorspelt in functie van het genormaliseerd sjTREC niveau, dCt genoemd. Het intercept wordt α genoemd en de helling β.

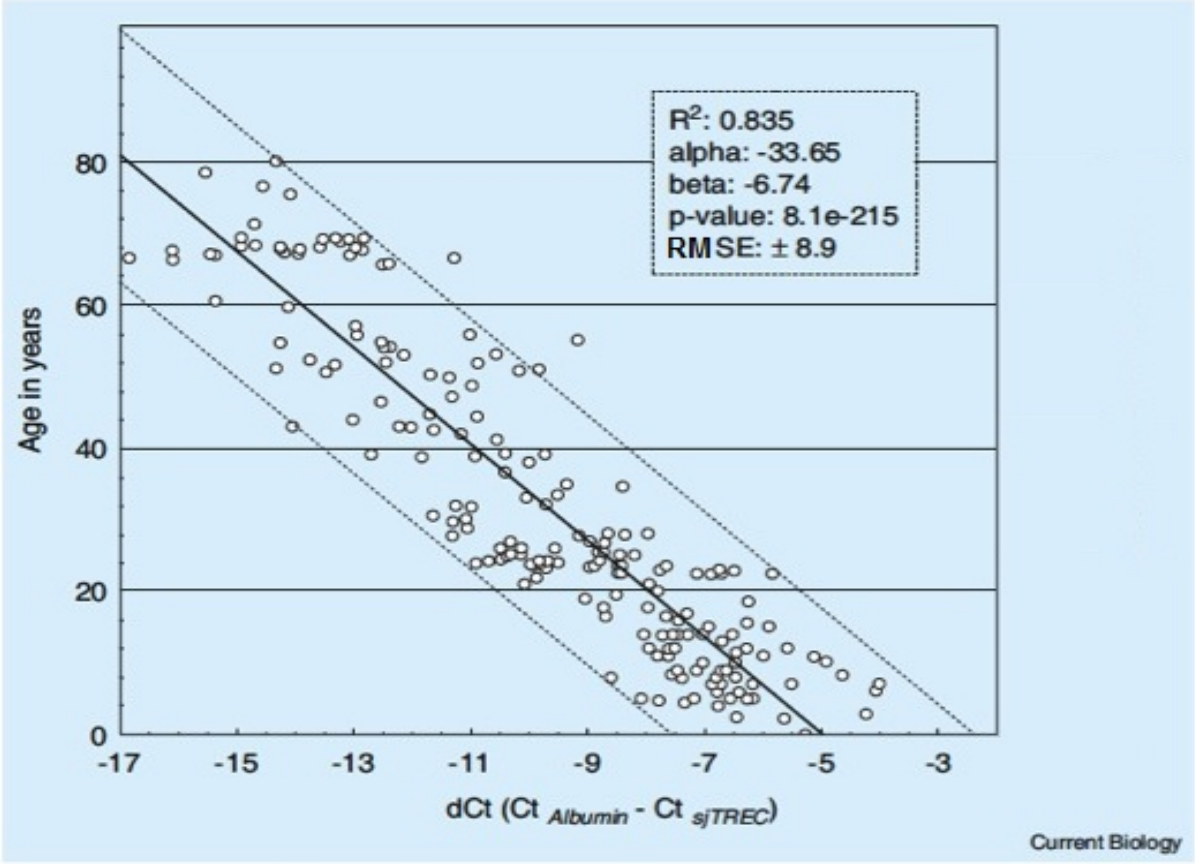


Figure 1. Age prediction from blood-derived DNA samples.
Linear regression of the relationships between human individual age and normalized sjTREC abundance in peripheral blood of 195 individuals (dotted lines correspond to 95% prediction interval).

1. Schrijf dit model neer en geef duidelijk de betekenis van alle gebruikte notatie weer.

2. Leid de standaarderror voor $\hat{\beta}$ af uit de informatie die in de figuur wordt meegegeven. Leg duidelijk uit hoe u uw benadering afgeleid hebt en geef alle relevante formules.

3. Geef de nul- en alternatieve hypothese geassocieerd met de p-waarde voor $\hat{\beta}$, zowel wiskundig als in woorden. Geef ook een duidelijke en correcte interpretatie van die (exacte) p-waarde.

4. Bereken de power die deze studie had om een alternatief te ontdekken dat gelijk is aan -0.5. Definieer alle grootheden die u nodig heeft voor deze berekening. Indien nodige informatie niet zou af te leiden zijn uit de informatie in de figuur, gebruik er dan realistische waarden voor.

5. Bespreek vier mogelijke manieren om **deze** power te verhogen in termen van hun praktische uitvoerbaarheid.

Veel succes!