

Inleveropgave 3: Regressie en toetskeuze

Statistiek 2022 – 2023

Deadline: zondag 11 december 23:59u

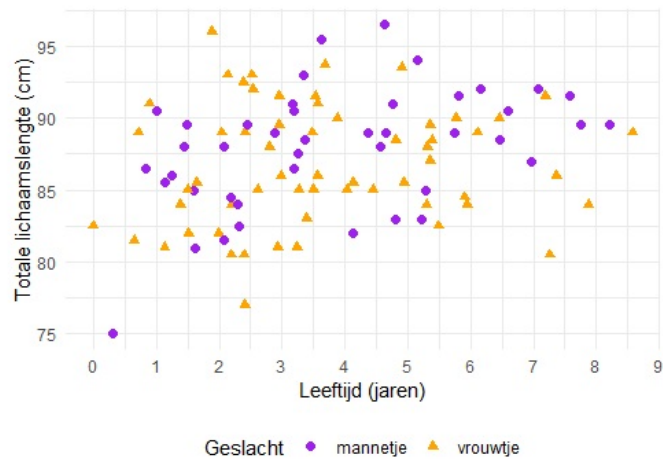
- Tip: lees eerst alle deelvragen goed door voordat je begint aan een opdracht, en bedenk of jouw antwoord aansluit bij de (deel)vraag.
- Maak de opdrachten van de inleveropgaven individueel en geef antwoord in je eigen woorden.
- Bij rekenvragen is alleen het antwoord niet voldoende; laat zien hoe je aan je antwoord komt. Het gebruik van een rekenmachine is toegestaan.
- Voor iedere vraag is (tussen haakjes) het maximum aantal te behalen punten aangegeven. Je kunt punten verliezen wanneer je iets opschrijft dat irrelevant of incorrect is.
- Lever een PDF met je uitwerkingen in via Brightspace, gemaakt met je favoriete tekstverwerker (geen scans van handgeschreven uitwerkingen).
- Let op: zonder digitale ontvangstbevestiging van Turnitin (meteen in beeld en tevens per e-mail) is je uitwerking niet correct ingezonden.

Veel succes!

1. De bergkoesoe is een klimbuideldier dat voorkomt in de bergen van Victoria in Zuid-Australië. Biologen hebben daar in totaal 102 bergkoesoes gevangen en verschillende kenmerken genoteerd. In de grafiek hieronder zie je van elk individu het geslacht, de leeftijd, en totale lichaamslengte weergegeven in een puntgrafiek. Ook is een tabel gegeven die een deel van de uitvoer van een regressieanalyse op deze dataset weergeeft, met X_1 = leeftijd in jaren, X_2 = geslacht en Y = totale lichaamslengte in centimeters.



een bergkoesoe



Variable	Mean	St.dev.
lengte	88.1	4.4
leeftijd	3.8	2.0
geslacht_man	0.58	0.49

Variable	Estimate	Std. Error
(Constant)	87.7	1.05
leeftijd	0.55	0.21
geslacht_man	-3.0	0.82

- (1) (a) Geef (formeel) de geschatte voorspellingsvergelijking en interpreteer de geschatte intercept en helling.
 - (1) (b) Leg uit hoe deze geschatte intercept en helling bepaald worden in de regressieanalyse. Gebruik hierbij de concepten kleinste-kwadraten-schatting (Engels: *least square estimation*) en *SSE*.
 - (1) (c) Stel dat bovenstaande analyse ook geeft: $SSE = 1632$. Bereken een schatting van de conditionele standaarddeviatie en vergelijk deze met de marginale standaarddeviatie(s) van de relevante variabele(n).
 - (2) (d) Bereken voor zowel b_1 als b_2 de teststatistiek behorend bij $H_0 : b_i = 0$ en trek voor zowel X_1 als X_2 een conclusie of er een associatie met totale lichaamslengte is of niet bij een α -niveau van 0.05. Ga uit van een tweezijdige alternatieve hypothese.
2. Collega's van de onderzoekers uit vraag 1 merkten op dat er misschien wel een interactie-effect bestaat tussen geslacht en leeftijd. Dit onderzoeken we verder.
- (1) (a) Geef formeel het populatiemodel en de voorspellingsvergelijking met interactieterm tussen leeftijd en geslacht.
 - (2) (b) In de tabel hieronder zie je de TSS en SSE voor het model zonder interactieterm

('reduced model') en het model met interactieterm ('complete model'). Gebruik de juiste statistische toets om te bepalen of het model met interactieterm significant beter is met $\alpha = 0.05$.

Model	TSS	SSE
Complete	1996	1609
Reduced	1996	1632

(1) (c) Wat betekent je conclusie bij de vorige vraag voor de lichaamslengte van bergkoesoes?

3. Op dit moment bevat de Google Play Store meer dan 3,5 miljoen apps, en dagelijks komen daar zo'n 3800 apps bij. Om onderzoek te doen naar apps die nu beschikbaar zijn heb je een dataset waarin van 10000 apps de volgende variabelen beschikbaar zijn:

- Naam: de naam van de app
- Categorie: de categorie waarin de app beschikbaar is (33 opties)
- Rating: het aantal sterren dat de app krijgt van reviews (tussen 0-5, op 1 decimaal)
- Grootte: de benodigde ruimte voor de app (in megabyte)
- Downloads: het aantal keer dat de app is gedownload
- Type: gratis of betaald
- Prijs: voor betaalde apps, de prijs van de app (in euro)
- Leeftijdsklasse: de leeftijd waarvoor de app geschikt is (iedereen/10+/12+/18+)

Geef voor elk van de onderstaande onderzoeksvragen: 1) de verklarende variabele en de responsvariabele (Engels: *explanatory and response variable*); 2) de meetschaal van ieder van deze variabelen; 3) de toets die je zou kiezen om de onderzoeksvraag te beantwoorden (noteer waar relevant ook: gepaard/ongepaard, voor één of twee groepen); en tot slot 4) de hypothesen. Noteer je antwoorden kort en bondig, licht je antwoorden niet toe.

- ($\frac{3}{4}$) (a) Is de verdeling over de verschillende leeftijdsklassen hetzelfde voor apps uit de verschillende categorieën?
- ($\frac{3}{4}$) (b) Bij de betaalde apps, worden goedkopere apps vaker gedownload dan duurdere?
- ($\frac{3}{4}$) (c) Nemen apps uit de categorie *Kaarten & Navigatie* gemiddeld meer dan 100 megabyte opslagruimte in?
- ($\frac{3}{4}$) (d) Is de proportie gratis apps in de categorie *Entertainment* anders dan in de categorie *Dating*?