Werkcollege opgaven
Statistiek

Werkcollege 13 & 14: Inleveropgave

Inleiding

Bij deze laatste inleveropgave is het de bedoeling een echte dataset te analyseren. Dit is een wat vrijere opdracht, waarbij een verslagje van de analyse en conclusies ingeleverd moet worden. Deze opdracht telt twee keer zo zwaar als eerdere inleveropdrachten. Deze opdracht mag alleen of per tweetal gemaakt worden.

Bij het schrijven van het verslag houd je in gedachten dat een opdrachtgever je heeft benaderd met onderstaande tekst.

01-02-1987

Beste statisticus,

Bijgaand treft u een dataset met gegevens van 52 Amerikaanse bedrijven uit de Forbes500 lijst van beste bedrijven in de VS over het jaar 1986. Hierin vindt u van elk van deze bedrijven de omzet, het kapitaal en de marktwaarde. Verder is per bedrijf aangegeven tot welke van de volgende vier sectoren het behoort: energie, financieel, productie of detailhandel.

Graag zou ik u willen vragen te onderzoeken hoe de omzet door deze factoren beïnvloed wordt, zodat ik daar rekening mee kan houden in mijn bussinessplan. In afwachting van uw spoedige reactie verblijf ik.

Met vriendelijke groet,

B. Gates

Suggesties

Hieronder staan enkele hints en suggesties voor de analyse. De bedoeling is dat je ook zelf nadenkt wat zinvol is om te doen, argumenten geeft voor gemaakte keuzes en uitlegt wat je gedaan hebt. Je verslag is een samenhangend verhaal waarin je je bevindingen presenteert. In het bijzonder wordt afgeraden om je te beperken tot het puntsgewijs afwerken van onderstaand lijstje. Je mag gebruik maken van functies die in `R` beschikbaar zijn, maar je opdrachtgever heeft wellicht geen specifieke kennis van `R`. Wel heeft hij tijdens zijn studie het vak Statistiek gevolgd, waarbij gebruik gemaakt werd van een boek van Van der Vaart, Bijma en Jonker. Je kunt dus zinnen gebruiken als "We fitten een lineair model met ... als onafhankelijke variabelen en berekenen de kleinste kwadraten schatter voor ...."

Voor deze opdracht word je geacht je methoden en conclusies uit te leggen, omdat wij ook willen kunnen controleren wat je precies gedaan hebt. In het echt zou dat misschien niet nodig zijn, of misschien zelfs niet de bedoeling. Ook conclusies die je normaal gesproken niet zou rapporteren aan je opdrachtgever kunnen in dit geval van belang zijn.

1. De variabelen in de dataset heten `sales`, `assets`, `marketval` en `sector`.
2. Begin met het maken van relevante plotjes om een indruk te krijgen van de data.
3. Stel jezelf als doel om een goed lineair regressiemodel te fitten.
4. Overweeg om de financiële variabelen niet direct in het model op te nemen, maar eerst de logaritme te nemen. Dit kun je ook doen bij de afhankelijke variabele.

5. Het meest basale model heeft 6 regressieparameters en een parameter σ. Voeg interactie-termen toe met de categorische variabele `sector`.

6. Overweeg of het zinvol is om de variantie per categorie te schatten. Hoe ziet je model er dan uit? Probeer het compact op te schrijven. Om te toetsen of $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4$ kun je gebruiken dat

$$\frac{\mathrm{SS}_{\mathrm{res}}}{\sigma^2} \sim \chi^2(n-k)$$

in een model met k regressieparameters. De volgende toetsingsgrootheid is geschikt:

$$T = (n-k)\log\left(\frac{\mathrm{SS}_{\mathrm{res}}}{n-k}\right) - \sum_{i=1}^{4}(n_i - k_i)\log\left(\frac{\mathrm{SS}_{\mathrm{res}}^{(i)}}{n_i - k_i}\right).$$

Hierbij is n de steekproefomvang, k het totaal aantal regressieparameters, en k_i het aantal voor categorie i. De verdeling van T kun je simuleren.

7. Je kunt je afvragen of de continue variabelen van significante invloed zijn, en of deze invloed afhankelijk is van de sector. Voor dit soort vragen is een F-toets geschikt.