Eindopdracht "Statistics for Data Science"

Dennis Fok

13 september 2022

Voor deze eindopdracht zijn er twee mogelijkheden. Hieronder volgt ten eerste de beschrijving van een opdracht gebaseerd op een dataset over universiteiten in de Verenigde Staten. Het is ook mogelijk om de eindopdracht te baseren op een eigen (bedrijfs)dataset. Aan het einde van dit document volgt meer informatie hierover.

Voor beide opdrachten geldt dat de resultaten in drie delen ingeleverd moeten worden, telkens samengevat in een bondig rapport. Hierbij moet voor elke deelanalyse duidelijk zijn wat het doel van de berekening is, wat de uitkomst is en vooral wat de conclusie is. Zorg ervoor dat er voldoende details gegeven zijn, bijvoorbeeld bij het uitvoeren van een statistische toets is het van belang om

1. de keuze van de toets duidelijk te motiveren;
2. de nul- en alternatieve hypothese te beschrijven;
3. de p-waarde te rapporteren;
4. de uiteindelijke conclusie te rapporteren in termen van de toepassing (dus niet alleen "de nul-hypothese wordt verworpen", maar ook wat dit dan betekent).

Voor elk deel van de opdracht moet een rapport in pdf formaat ingeleverd worden, samen met de volledige R code. Het inleveren verloopt via Canvas (zie onder "Assignments"). Het verwerken van de resultaten met behulp van R-Markdown is prima, maar probeer onnodige output te vermijden.

De deadlines zijn als volgt:

- Deel 1: zondag, 25 september (voor het einde van de dag);
- Deel 2: zondag, 16 oktober (voor het einde van de dag);
- Deel 3: zondag, 30 oktober (voor het einde van de dag).

1 Opdracht obv. universiteitsdata

Het databestand `college_statistics.csv` bevat informatie over 777 universiteiten en colleges in de Verenigde Staten in een specifiek jaar. De variabelen in deze dataset zijn:

- Private: Public/private indicator
- Apps: Number of applications received
- Accept: Number of applicants accepted
- Enroll: Number of new students enrolled
- Top10perc: Percentage of new students who were in the top 10% in their high school class
- Top25perc: Percentage of new students who were in the top 25% of their high school class
- F.Undergrad: Number of full-time undergraduates
- P.Undergrad: Number of part-time undergraduates
- Outstate: Out-of-state tuition
- Room.Board: Room and board costs
- Books: Estimated book costs
- Personal: Estimated personal spending
- PhD: Percent of faculty with Ph.D.'s
- Terminal: Percent of faculty with terminal degree
- S.F.Ratio: Student/faculty ratio
- perc.alumni: Percent of alumni who donate
- Expend: Instructional expenditure per student
- Grad.Rate: Graduation rate

In deze opdracht voer je een gedetailleerde statistische analyse uit op deze dataset. Hieronder vind je een lijst van stappen die uitgevoerd moeten worden.

Deel 1

1. Lees de data in als dataframe en zorg ervoor dat de eerste kolom als label voor de rijen wordt gebruikt. Hint: zie `?row.names`.

2. Voer beschrijvende statistiek uit dmv. het maken van een aantal grafieken. Deze grafieken mag je via de "Base R graphics" of mbv. ggplot maken.
   Creëer geschikte grafieken om de volgende vragen te beantwoorden:

   (a) Zijn private universiteiten overwegend kleiner of groter dan publieke universiteiten? Je mag zelf een definitie voor groot/klein definiëren.

   (b) Hoe ziet de verdeling eruit van het acceptatiepercentage? Wat is het acceptatiepercentage voor de meest selectieve universiteit?

   (c) Zijn de meer selectieve universiteiten ook overwegend duurder dan minder selectieve universiteiten? (Je mag zelf bepalen welke kosten je wel/niet mee neemt.)

   (d) Bedenk zelf een extra vraag en creëer een geschikte figuur om deze vraag mee te beantwoorden.

3. Voer hypothesetoetsen uit om de volgende vragen te beantwoorden. Geef telkens duidelijk aan wat de exacte nul- en alternatieve hypothese is die je toetst, motiveer de keuze van de specifieke toets en verwoord duidelijk de conclusie.

   (a) Ontvangen elite scholen een ander aantal aanmeldingen in vergelijking met niet-elite scholen? Definieer "elite-school" als scholen waarvoor geldt dat meer dan 50% van de studenten tot de top 10% van hun high school behoort.

   (b) Is er een verband tussen acceptance rate en graduation rate?

   (c) Bedenk zelf ook een extra hypothese om te toetsen en voer de hypothesetoets uit.

Deel 2

4. Maak een model dat het aantal aanmeldingen kan voorspellen op basis van factoren die voorafgaand aan binnenkomen van de aanmeldingen beschikbaar zijn. Maak daarom bijvoorbeeld **geen** gebruik van de variabelen Accept en Enroll. Deze zijn duidelijk een gevolg (en geen oorzaak) van het aantal aanmeldingen. Volg hierbij de volgende stappen.

   (a) Voer eerst een test uit voor de hypothese dat het aantal aanmeldingen een normale verdeling volgt. Wat is je conclusie? Is deze conclusie van belang voor het verder modelleren van deze variabele?

   (b) Deel de data eerst op willekeurige manier op in een "estimation" en "test" sample. Neem 600 universiteiten in de estimation sample. Zorg ervoor dat deze opdeling reproduceerbaar is. Hint: de R-functies `set.seed` en `sample` kunnen hiervoor gebruikt worden.

(c) Maak eerst een lineair model voor het aantal aanmeldingen. Gebruik hiervoor alleen de estimation sample.

(d) Pas backward elimination toe om het aantal variabelen terug te brengen.

(e) Voer diverse toetsen uit om de aannamen van het lineaire model te testen.

(f) Maak vervolgens een model voor de logaritme van het aantal aanmeldingen (ook weer met backward elimination).

(g) Voer opnieuw de diverse toetsen uit om de aannamen van het model te testen.

(h) Welk van de twee modellen heeft de voorkeur?

(i) Probeer het gekozen model nog verder te verbeteren: denk aan het toevoegen van transformaties van verklarende variabelen.

(j) Hoe interpreteer je de coëfficiënten in het model dat je uiteindelijk hebt gevonden? Wees hierbij heel precies. Welke factoren zijn uiteindelijk het meest van belang?

(k) Gebruik het uiteindelijke model om voorspellingen te maken voor de waarnemingen in de estimation en de test sample.

(l) Vergelijk de voorspelkracht (mbv. mean squared error) van het model op de estimation sample met die op de test sample. Wat concludeer je?

5. In dit onderdeel voer je een ANOVA analyse uit op de relatie tussen de student faculty ratio en het percentage studenten uit de top 25% van de high school. Volg de volgende stappen:

   (a) Maak een factor met drie levels op basis van de variabele Top25perc. De levels zijn: laag (minder dan 20%)/midden/hoog (meer dan 40%).

   (b) Voer de ANOVA analyse uit en geef je conclusie(s). Presenteer de ANOVA resultaten zowel numeriek als grafisch.

   (c) Onderzoek of er outliers zijn. Pas de analyse aan als dat nodig is.

Deel 3

6. Maak een model om de factoren te vinden die bijdragen aan een hoog "slagingssucces".

   (a) Definieer een nieuwe variabele die 1 als het slagingspercentage groter is dan 60% en 0 als dat niet zo is.

   (b) Deel de data opnieuw op in een estimation en een test sample.

   (c) Maak mbv. de estimation data een logit model om de slagingssucces variabele te verklaren. Denk hierbij goed na over transformaties van je variabelen. Bijvoorbeeld heeft het zin om het aantal applicaties, aantal acceptaties, en het aantal enrollments in hetzelfde model op te nemen? Of kunnen sommige van deze variabelen beter als percentages opgenomen worden?

(d) Gebruik wederom backward selection om het aantal verklarende variabelen te verkleinen.

(e) Welke variabelen hebben uiteindelijk een significante invloed?

(f) Bereken het percentage goed voorspelde scholen zowel voor de estimation sample als voor de test sample (maak eerst voorspellingen voor beide datasets en gebruik daarna bijvoorbeeld de functie `confusionMatrix()`).

2 Opdracht obv. eigen (bedrijfs)data

De eindopdracht kan ook gebaseerd worden op eigen (bedrijfs)data. Het idee van de bovenstaande opdracht en de opdeling in 3 delen blijft hierbij gehandhaafd. Meer precies moeten de volgende elementen in de uitgewerkte opdracht aanwezig zijn:

Deel 1

(a) Een uitgebreide analyse op basis van beschrijvende statistiek.
Definieer ongeveer vijf vragen en beantwoord deze met behulp van geschikte grafieken.

(b) Het uitvoeren van hypothesetoetsen.
Definieer ongeveer vier interessante hypothesen en voer de juiste hypothesetoetsen uit.

Deel 2

(a) Het opstellen van een lineair model.
Kies een geschikte afhankelijke variabele en creëer een model om deze te verklaren uit de andere variabelen. Let hierbij goed op de modelspecificatie (heb je wellicht transformaties nodig van de y of de x variabelen?).

(b) Het uitvoeren van een ANOVA of ANCOVA analyse.

Deel 3

(a) Het opstellen van een logit model.
Kies (of creëer) een geschikte binaire variabele en verklaar deze met behulp van een logit model uit de andere variabelen. (Je kunt ook kiezen om een andere niet continue variabele te verklaren mbv. een generalized linear model.)

Als bepaalde onderdelen niet uit te voeren zijn met behulp van de eigen data, dan kan je voor deze onderdelen ook de universiteitsdata gebruiken. Indien je twijfelt of je eigen dataset geschikt is voor deze opdracht, stuur dan een email (dfok@ese.eur.nl).