Naam:

Rijksuniversiteit Groningen, Faculteit der Letteren
Vak: Statistiek II, LIX002X05, cursusjaar 2014/2015, 2e semester 2e blok
Docent: Wilbert Heeringa
Plaats en tijd: dinsdag 26 mei 2015 om 13 uur, Offerhauszaal, Academiegebouw

Belangrijke instructies

1. Schrijf uw naam en studentnummer hierboven, schrijf uw initialen op alle bladzijden.
2. U hebt 3 uur de tijd.
3. Het gebruik van boek, syllabus of aantekeningen is niet toegestaan.
4. Geef op alle vragen een antwoord.
5. Let op dat sommige vragen zwaarder tellen dan andere (er zijn in totaal 100 punten).

Opgave 1, basisbegrippen, 'W'/'O' voor waar/onwaar (40 punten)

1.	U bestudeert de invloed van leeftijd op kans van herstel van afasie die als gevolg van een beroerte optreedt. U gebruikt logistische regressie. De residuen van de regressie blijken normaal verdeeld te zijn, maar helaas groot. Logistische regressie is hier niet bruikbaar.	W	O
2.	Het centreren van variabelen is zinvol voor een mixed model, maar is ook zinvol voor een lineaire regressie analyse.	W	O
3.	Het voordeel van ANOVA is dat deze toets niet gevoelig is voor verschillen tussen de standaarddeviaties van de groepen.	W	O
4.	In ANOVA meet de toetsingsgrootheid F de verhouding tussen de variatie binnen de groepen ten opzichte van de variatie tussen de groepen. Dus: F = MSE / MSG.	W	O
5.	Het specifieke voordeel van een mixed-effect model is dat numerieke en categorische variabelen in combinatie met elkaar gebruikt mogen worden.	W	O
6.	U gebruikt ANOVA om de verwerkingstijden tussen vier implementaties van één algoritme) te vergelijken. Nadat U de tijden van 800 proeflopen (200 per implementatie) heeft verzameld, blijken de deelpopulaties sterk niet-normaal verdeeld te zijn, maar (rechts)scheef. Sommige looptijden zijn dus heel lang. Stelling: Men mag ANOVA verder gebruiken omdat de aantallen groot genoeg zijn.	W	O
7.	Voor de ANOVA in (6) geldt: DFG=4, DFE=795 en DFT=799.	W	O
8.	Contrasten worden gebruikt als je vooraf geen idee hebt over de relatie tussen de groepen.	W	O
9.	Gegeven F = 8.96 met df=2,40. De kritieke waarde is 4.76 (alpha=0.05). De alternatieve hypothese moet worden verworpen.	W	O
10.	Een één-factor ANOVA-toets geeft een F met df=1,30. Men had hier net zo goed een t-toets kunnen gebruiken.	W	O
11.	Proefpersonen lezen “woorden” en geven d.m.v. een toetsenknop hun oordeel of wat ze zien een echt woord is (dus *hond* ‘ja’, en *nhdu* ‘nee’). U onderzoekt de	W	O

verhouding tussen de reactietijd van de proefpersonen enerzijds en zowel de lengte als de frequentie van de woorden anderzijds. Stelling: ANOVA is een geschikte analyse voor deze gegevens.

12. In een interactie-grafiek zie je een interactie als de lijnen evenwijdig aan elkaar lopen. Een zuivere interactie wordt echter nooit bereikt. W O

13. U bestudeert het effect van sekse enerzijds en burgerlijke staat anderzijds op de acceptatie van web-sites voor “e-shopping” (winkelen). Omdat mogelijk interactie bestaat tussen deze factoren (sekse en burgerlijke staat), is een meervoudige ANOVA de aangewezen opzet voor de analyse van gegevens. W O

14. Voor de berekening van de Pearson’s correlatiecoëfficïent is het niet nodig om te onderscheiden tussen de te verklaren variabelen en de verklarende variabele. W O

15. In een meervoudige regressie kunnen variabelen die alleen (d.w.z. in enkelvoudige regressiemodellen) niet-significante voorspellers zijn in complexere modellen toch significantie bereiken. W O

16. Omgekeerd kan het zijn dat variabelen die in enkelvoudige regressieanalyses wel significante voorspellers zijn in complexere modellen geen effect meer hebben. W O

17. Eén twee-factor ANOVA is slimmer dan twee één-factor ANOVA’s, omdat je dan minder proefpersonen nodig hebt. W O

18. Een mixed-effects model heeft dezelfde assumpties als een lineair regressiemodel. W O

19. Een ANCOVA test combineert ANOVA met lineaire regressie analyse. W O

20. De effect size in een mixed model wordt gemeten met de determinatiecoëfficient R^2. W O

Opgave 2 (15 punten)

Een adverteerder wil weten of zijn advertenties op bepaalde websteks frequenter worden gelezen. Men kan daar tot op zekere hoogte achter komen door verschillende soorten websteks met elkaar te vergelijken. De adverteerder heeft 60 websteks aselect gekozen en drie versies gemaakt. 20 websteks heeft hij voorzien van veel metadata, 20 heeft hij voorzien van veel animaties, en 20 heeft hij voorzien van zowel metadata als animaties. De adverteerder stelt dan de vraag of deze verschillen in ontwerpen tot verschillen in bezoekersaantallen leiden.

1. Stel een statistische toets voor de analyse van deze gegevens voor. Onderbouw uw keuze.

2. Formuleer de nulhypothese en de alternatieve hypothese.

3. Noem de belangrijkste voorwaarden die u moet controleren om te weten of de door u voorgestelde toets toegepast mag worden.

4. Welke terugvaloptie hebt u als uw gegevens niet aan alle voorwaarden voldoen?

5. Als de waarschijnlijkheid van de toetsingsgrootheid $p = 0.02$ is, wat valt dan te concluderen?

6. Stel dat men vooraf aanwijzingen heeft dat de combinatie van animaties en metadata de meeste bezoekers trekt. Hoe onderzoek je of de bezoekersaantallen van websteks met metadata én animaties hoger zijn dan die van de websteks met alleen veel animaties en alleen metadata?

7. Geef voor de vorige vraag ook de H0 en Ha.

Opgave 3 (15 punten)

Achteraf bezien realiseerde de adverteerder (zie vorige opgave) dat hij het experiment ook anders had kunnen uitvoeren, door namelijk slechts 20 websteks aselect te kiezen en van elke webstek drie versies te maken, één met veel metadata, één met veel animaties, en één met metadata én animaties.

1. Welke toets zou je in deze opzet gebruiken?

2. Heeft de adverteerder gelijk? Waarom zou deze nieuwe opzet wel of geen verbetering zijn?

3. Welke aanname wordt specifiek door deze toets gemaakt?

4. Hoe test je of die aanname voor de data aangenomen mag worden?

Opgave 4 (30 punten)

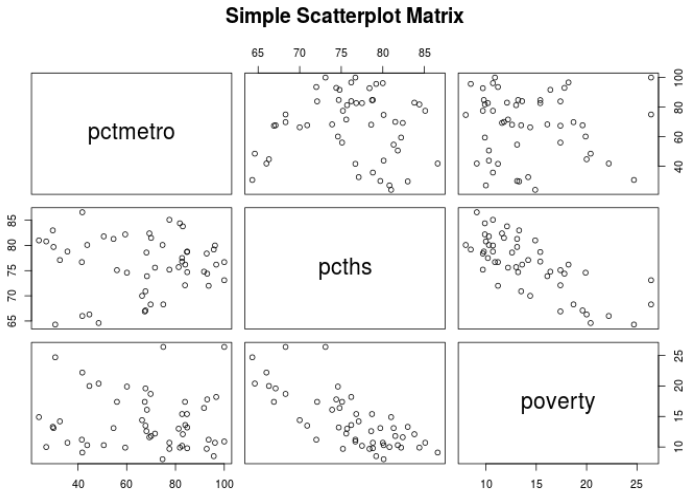
We gebruiken de 'crime dataset' uit *Statistical Methods for Social Sciences, Third Edition* van Alan Agresti and Barbara Finlay (Prentice Hall, 1997). Voor elk van de 50 staten in de USA vinden we onder andere de volgende variabelen: pctmetro (the percent of the population living in metropolitan areas), *pcths* (percent of population with a high school education or above), *poverty* (percent of population living under poverty line) en *crime* (number of violent crimes per 100,000 people). De onderzoekers willen weten of *pctmetro*, *pcths* en *poverty* de mate van criminaliteit in een staat voorspellen.

(Voorbeeld ontleend aan: Introduction to SAS. UCLA: Statistical Consulting Group, http://www.ats.ucla.edu/stat/sas/notes2/, versie 19 mei 2015).

1. Formuleer de nulhypothese en de alternative hypothese.

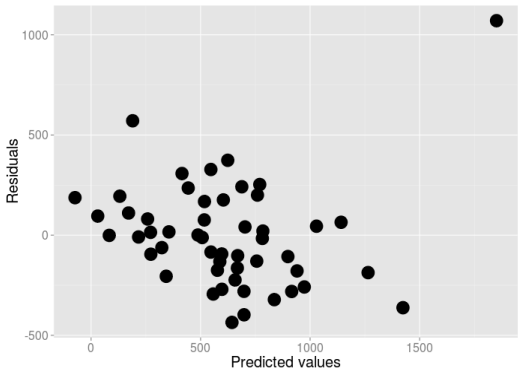
2. De onderzoekers willen graag een multiple regression analyse doen. Is dat inderdaad de juiste toets? Motiveer je antwoord.

3. Tussen de predictoren blijken volgende correlaties te bestaan: -0.00397742, -0.06053852 en -0.7439382. De correlaties tussen de predictoren zijn gevisualiseerd in de volgende spreidingsdiagrammen:



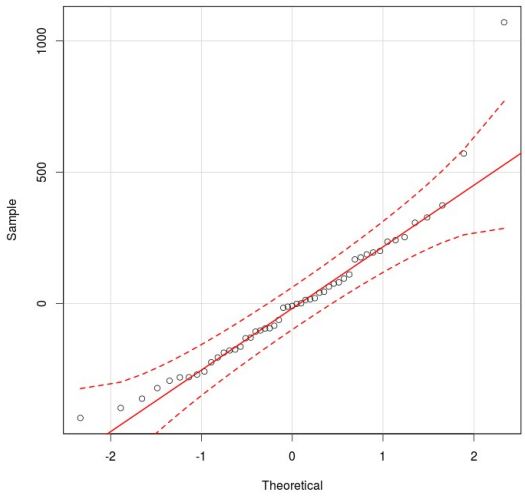
Kunnen de onderzoekers gezien deze correlaties inderdaad een meervoudige regressie analyse uitvoeren? Waarom wel, of waarom niet?

4. De onderzoekers maakten een residuenplot die er zo uitziet:



Welke twee assumpties worden getest met behulp van deze grafiek? Wordt aan beide assumpties voldaan? Waarom wel, of waarom niet?

5. De onderzoekers maakten ook de volgende grafiek:



Hoe heet dit type grafiek? Welke assumptie kun je hiermee toetsen? Met welke test kun je diezelfde assumptie ook toetsen? Wordt – gezien de grafiek - aan de assumptie voldaan? Hoe zie je dat precies?

6. De onderzoekers doen de toets en krijgen als output onder andere:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3334.774    946.024  -3.525 0.000956 ***
pctmetro       11.928      1.759   6.780 1.76e-08 ***
pcths          26.866     10.318   2.604 0.012301 *
poverty        76.864     12.609   6.096 1.93e-07 ***
```

Wat wordt gerepresenteerd door de getallen onder 'Estimate”?

7. Wat concludeer je met betrekking tot de hypothesen bij 1.?

8. Verder wordt nog als output verkregen:

```
Multiple R-squared:  0.6428,    Adjusted R-squared:   0.62
```

Hoe wordt de 'multiple R-squared' berekend? Wat vertelt het getal '0.6428' je precies? Welke maat heeft de voorkeur, de 'multiple R-squared' of de 'adjusted R-squared'?