Examen Statistiek III

Bachelor Wiskunde, 20 januari 2021

Naam: .....

U krijgt 3 uur om aan dit open boek examen te werken. Dit is niet ruim begroot, maar volstaat indien u de cursus goed beheerst. Laat u echter niet afschrikken: het examen ziet er lang uit, maar dat is vooral door de vele output.

Probeer maximaal 2u aan de eerste vraag te besteden (reken dus maximum 30 minuten telkens voor deel a, b, c en d), 15 min aan de tweede vraag en 45 min aan de derde vraag. De score-verdeling is evenredig met de tijd die nodig geacht wordt; dus 2/3 van de punten staan op vraag 1. Bij eventuele deliberatie speelt het in uw voordeel als u ook van vraag 2 of 3 een onderdeel goed kon beantwoorden (in vergelijking met slechts de eerste vraag).

Maak telkens uw redenering duidelijk. Een eindresultaat, zelfs indien correct, krijgt pas punten als duidelijk is hoe u ertoe kwam. **Gelieve deze opgave samen met uw oplossingen in te dienen.**

1. Op basis van een onderdeel van de 2007-2008 National Health and Nutrition Examination Survey (NHANES) in de VS onderzoeken we het effect van deelname aan een programma voor schoolmaaltijden op het BMI. De volgende variabelen zijn hiervoor ter beschikking:

   - BMI: Body mass index
   - School_meal: Deelname aan schoolmaaltijdprogramma's (1: Ja, 0: Nee)
   - age: Leeftijd kind
   - ChildSex: Geslacht kind (1: Mannelijk, 0: Vrouwelijk)
   - black: ras (1: Zwart, 0: anders)
   - mexam: ras (1: Spaanse origine: 0: anders)
   - pir200_plus: Gezin boven 200% van het federale armoedeniveau (1: Ja, 0: Nee)
   - WIC: Deelname aan een speciaal aanvullend voedingsprogramma (1: Ja, 0: Nee)
   - Food_Stamp: Deelname aan voedselstempelprogramma (1: Ja, 0: Nee)
   - fsdchbi: Voedselzekerheid voor kind (1: veilig, 0: onzeker)
   - AnyIns: Beschikt men over een verzekering van om het even welke aard (1: Ja, 0: Nee)
   - RefSex: Geslacht van de volwassen respondent (die de vragenlijst invulde) (1: man, 0: vrouw)
   - RefAge: Leeftijd van de volwassen respondent (die de vragenlijst invulde)

   (a) Beschouw eerst de volgende output:

   ```
   MODEL 1

   Call:
   lm(formula = BMI ~ School_meal, data = nhanes_bmi)
   ```

```
Residuals:
   Min     1Q Median     3Q    Max
-7.850 -4.106 -1.530  2.584 28.120

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.8162     0.1675 118.272   <2e-16 ***
School_meal   0.5339     0.2257   2.366   0.0181 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.419 on 2328 degrees of freedom
Multiple R-squared:  0.002398,Adjusted R-squared:  0.001969
F-statistic: 5.596 on 1 and 2328 DF,  p-value: 0.01809
```

i. Interpreteer de coëfficiënt bij `School_meal` in MODEL 1.

ii. De R^2 van dit MODEL 1 is zeer laag. Kunnen we hieruit besluiten dat dit model foutief is? Leg uit.

iii. Kunt u achterhalen hoeveel procent van de kinderen in deze studie deelgenomen hebben aan het programma voor schoolmaaltijden? Het volstaat om de variantie van de variabele `School_meal` te berekenen, en vervolgens uit te leggen hoe hieruit kan afgeleid worden (mogelijks niet op unieke manier) hoeveel procent van de kinderen in deze studie deelgenomen hebben aan het programma voor schoolmaaltijden; deze laatste berekening hoeft u niet uit te voeren.

(b) We voegen nu ook alle andere predictoren toe aan het model. We beschouwen daarbij 2 mogelijkheden: in MODEL 2 wordt `age` enkel als hoofdeffect opgenomen; in MODEL 3 wordt de leeftijd (die gehele waarden van 4 tot en met 17 aanneemt) daarnaast ook als categorische variabele opgenomen. Het model bevat veel predictoren. **Bij interpretatie van coëfficiënten hoeft u daarom uiteraard niet alle predictoren op te schrijven, maar mag u verkorten tot bvb 'leeftijd, ... (alle overige predictoren in het model)'.**

```
MODEL 2:
Call:
lm(formula = BMI ~ . , data = nhanes_bmi)

Residuals:
    Min      1Q  Median      3Q     Max
-9.3813 -3.0313 -0.8922  1.9492 23.9685

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.7338066  0.5544644  22.966   <2e-16 ***
School_meal  0.0612479  0.2208829   0.277   0.7816
```

```
age          0.7296864  0.0256259  28.475   <2e-16 ***
ChildSex    -0.2084058  0.1886977  -1.104   0.2695
black        0.4411737  0.2390988   1.845   0.0651 .
mexam        0.6190676  0.2428658   2.549   0.0109 *
pir200_plus -0.4673003  0.2322279  -2.012   0.0443 *
WIC         -0.0007535  0.2612967  -0.003   0.9977
Food_Stamp  -0.1014021  0.2488077  -0.408   0.6836
fsdchbi     -0.1835593  0.2324985  -0.790   0.4299
AnyIns      -0.3367494  0.2833039  -1.189   0.2347
RefSex      -0.4910218  0.2002450  -2.452   0.0143 *
RefAge       0.0173300  0.0101659   1.705   0.0884 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.544 on 2317 degrees of freedom
Multiple R-squared:  0.3018,Adjusted R-squared:  0.2982
F-statistic: 83.46 on 12 and 2317 DF,  p-value: < 2.2e-16

MODEL 3:
Call:
lm(formula = BMI ~ . + factor(age), data = nhanes_bmi)

Residuals:
    Min      1Q  Median      3Q     Max
-9.4149 -2.9978 -0.9129  1.9127 23.0577

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   13.64975    0.64374  21.204  < 2e-16 ***
School_meal    0.19490    0.23292   0.837  0.40280
age            0.63313    0.04100  15.442  < 2e-16 ***
ChildSex      -0.22265    0.18870  -1.180  0.23817
black          0.45223    0.23914   1.891  0.05874 .
mexam          0.59805    0.24312   2.460  0.01397 *
pir200_plus   -0.38174    0.23292  -1.639  0.10135
WIC           -0.03040    0.26307  -0.116  0.90802
Food_Stamp    -0.12256    0.24906  -0.492  0.62272
fsdchbi       -0.19142    0.23298  -0.822  0.41140
AnyIns        -0.33538    0.28371  -1.182  0.23727
RefSex        -0.51645    0.20019  -2.580  0.00995 **
RefAge         0.01797    0.01016   1.769  0.07708 .
factor(age)5  -0.39229    0.45953  -0.854  0.39338
factor(age)6  -0.77703    0.44978  -1.728  0.08420 .
```

```
factor(age)7  -1.20885    0.42955  -2.814  0.00493 **
factor(age)8  -0.34991    0.41781  -0.837  0.40241
factor(age)9  -0.09451    0.43055  -0.220  0.82627
factor(age)10 -0.12928    0.43173  -0.299  0.76462
factor(age)11 -0.09074    0.42511  -0.213  0.83099
factor(age)12  0.60178    0.48217   1.248  0.21214
factor(age)13  0.51774    0.51518   1.005  0.31502
factor(age)14  1.14913    0.48027   2.393  0.01681 *
factor(age)15  0.67329    0.53136   1.267  0.20525
factor(age)16  0.16850    0.52666   0.320  0.74905
factor(age)17       NA         NA      NA       NA
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.533 on 2305 degrees of freedom
Multiple R-squared:  0.3087,Adjusted R-squared:  0.3015
F-statistic: 42.88 on 24 and 2305 DF,  p-value: < 2.2e-16
```

i. Interpreteer de coëfficiënt bij `factor(age)10` in MODEL 3.

ii. Leg uit waarom volgens u de coëfficiënt bij `factor(age)17` in MODEL 3 niet geschat werd ('NA' staat voor 'not available').

iii. Ziet u op basis van de output voor MODEL 2 en/of 3 mogelijkheden om de nulhypothese te toetsen dat het gemiddelde BMI lineair varieert met leeftijd (conditioneel op de overige variabelen in het model)? Zo ja, voer deze toets uit. Zo nee, toon met R-code hoe u dit zou kunnen toetsen indien u toegang had tot de data.

(c) Het onderstaande model werd bekomen via double selection (voor het effect van deelname aan schoolmaaltijdprogramma's op het BMI). De uitkomst `logBMI` in dit model werd daarbij bekomen als de natuurlijke logaritme van `BMI`.

```
MODEL 4

Call:
lm(formula = logBMI ~ School_meal + black + mexam + pir200_plus +
    RefSex + RefAge + WIC + Food_Stamp + fsdchbi + factor(age),
    data = nhanes_bmi)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4912 -0.1406 -0.0351  0.1127  0.6885

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.7591322  0.0228246 120.884  < 2e-16 ***
School_meal    0.0081416  0.0102028   0.798  0.42497
```

```
black          0.0193500  0.0104750   1.847  0.06484 .
mexam          0.0314553  0.0105485   2.982  0.00289 **
pir200_plus   -0.0198450  0.0100687  -1.971  0.04885 *
RefSex        -0.0218838  0.0087676  -2.496  0.01263 *
RefAge         0.0007401  0.0004448   1.664  0.09628 .
WIC            0.0006060  0.0115232   0.053  0.95806
Food_Stamp    -0.0062360  0.0108331  -0.576  0.56491
fsdchbi       -0.0063355  0.0101784  -0.622  0.53371
factor(age)5   0.0124964  0.0208036   0.601  0.54811
factor(age)6   0.0274126  0.0210114   1.305  0.19214
factor(age)7   0.0388173  0.0206631   1.879  0.06043 .
factor(age)8   0.1174713  0.0204457   5.746 1.04e-08 ***
factor(age)9   0.1605844  0.0211311   7.599 4.30e-14 ***
factor(age)10  0.1884998  0.0211911   8.895  < 2e-16 ***
factor(age)11  0.2227784  0.0208402  10.690  < 2e-16 ***
factor(age)12  0.2778976  0.0227992  12.189  < 2e-16 ***
factor(age)13  0.3027875  0.0238608  12.690  < 2e-16 ***
factor(age)14  0.3597645  0.0218363  16.475  < 2e-16 ***
factor(age)15  0.3704443  0.0231564  15.997  < 2e-16 ***
factor(age)16  0.3771960  0.0222122  16.982  < 2e-16 ***
factor(age)17  0.3970411  0.0233378  17.013  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1986 on 2307 degrees of freedom
Multiple R-squared:  0.3438,Adjusted R-squared:  0.3375
F-statistic: 54.94 on 22 and 2307 DF,  p-value: < 2.2e-16
```

i. Interpreteer de coëfficiënt bij `School_meal` in MODEL 4. Let daarbij op de definitie van de uitkomst

   ```
   > nhanes_bmi$logBMI = log(nhanes_bmi$BMI)
   ```

   Geef bij voorkeur de interpretatie in termen van het geometrische gemiddelde van BMI (d.i. de exponentiële van het gemiddelde van de logaritme van het BMI).

ii. Bereken een 95% betrouwbaarheidsinterval voor het effect dat u in de vorige vraag geïnterpreteerd heeft.

iii. In deze vraag vergelijken we bovenstaande double selection procedure met single selection (d.i. modelbouw enkel op basis van het model voor de uitkomst). Welke aanpak biedt volgens u meest garanties op

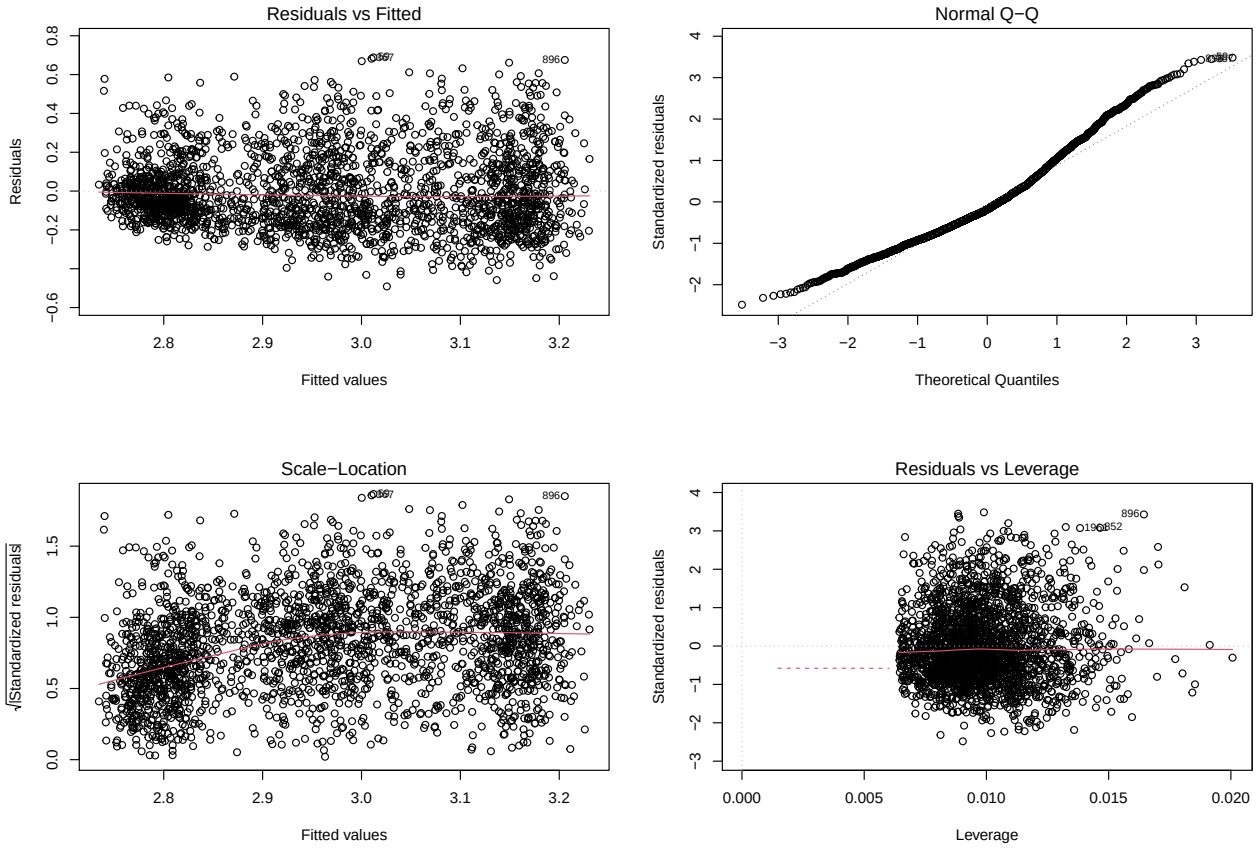
   A. smalle betrouwbaarheidsintervallen voor de coëfficiënt van `School_meal`? single selection / double selection

   B. een betrouwbaarheidsinterval voor het effect van schoolmaaltijden dat met 95% kans het echte effect bevat? single selection / double selection

   C. nauwkeurige predicties voor BMI? single selection / double selection

Motiveer telkens uw antwoord.

iv. Hieronder ziet u residuplots voor MODEL 4. Als we een accuraat 95% betrouwbaarheidsinterval voor de coëfficiënt van `School_meal` willen bekomen, zou u dan gebruik maken van de standaard intervallen die op basis van bovenstaande output te bekomen zijn, of ziet u op basis van de residuplots redenen en mogelijkheden om het beter te doen, hetzij door aanpassingen aan het model, hetzij door aanpassingen aan de berekening van het betrouwbaarheidsinterval. Stel hoogstens 1 aanpassing voor, en verduidelijk uw redenering op basis van de residuplots.



(d) Tot slot breiden we het finale model uit met drie termen (zie onderaan in de output):

```
MODEL 5

Call:
lm(formula = lBMI ~ School_meal * black + School_meal * mexam +
    School_meal * pir200_plus + RefSex + RefAge + WIC + Food_Stamp +
    fsdchbi + factor(age), data = nhanes_bmi)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.4790 -0.1396 -0.0340  0.1125  0.6901

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              2.7587782  0.0235759 117.017  < 2e-16 ***
School_meal              0.0089927  0.0160355   0.561  0.57499
black                    0.0448834  0.0159993   2.805  0.00507 **
mexam                    0.0434599  0.0166664   2.608  0.00918 **
pir200_plus             -0.0381395  0.0139209  -2.740  0.00620 **
RefSex                  -0.0212830  0.0087619  -2.429  0.01521 *
RefAge                   0.0007921  0.0004451   1.780  0.07524 .
WIC                     -0.0002889  0.0115335  -0.025  0.98002
Food_Stamp              -0.0047817  0.0108527  -0.441  0.65955
fsdchbi                 -0.0076842  0.0101839  -0.755  0.45060
factor(age)5             0.0152250  0.0208082   0.732  0.46443
factor(age)6             0.0307435  0.0210500   1.460  0.14429
factor(age)7             0.0437140  0.0207881   2.103  0.03559 *
factor(age)8             0.1216305  0.0205604   5.916 3.80e-09 ***
factor(age)9             0.1660987  0.0212632   7.812 8.51e-15 ***
factor(age)10            0.1927270  0.0212679   9.062  < 2e-16 ***
factor(age)11            0.2263730  0.0209607  10.800  < 2e-16 ***
factor(age)12            0.2816984  0.0229408  12.279  < 2e-16 ***
factor(age)13            0.3070438  0.0239177  12.837  < 2e-16 ***
factor(age)14            0.3618884  0.0218532  16.560  < 2e-16 ***
factor(age)15            0.3740492  0.0231816  16.136  < 2e-16 ***
factor(age)16            0.3799751  0.0222412  17.084  < 2e-16 ***
factor(age)17            0.3995594  0.0233313  17.125  < 2e-16 ***
School_meal:black       -0.0470607  0.0209785  -2.243  0.02497 *
School_meal:mexam       -0.0255905  0.0214081  -1.195  0.23207
School_meal:pir200_plus  0.0371777  0.0187042   1.988  0.04697 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1984 on 2304 degrees of freedom
Multiple R-squared:  0.3463,Adjusted R-squared:  0.3392
F-statistic: 48.82 on 25 and 2304 DF,  p-value: < 2.2e-16
```

i. Interpreteer de coëfficiënt bij `School_meal:black` in MODEL 5. Let daarbij op de definitie van de uitkomst.

ii. Interpreteer de 3 cijfers in volgende output

```
> exp(predict(model, newdata = data.frame(School_meal = 1, age = 12,
  black = 1, mexam = 0, pir200_plus = 0, WIC = 0, Food_Stamp = 0,
  fsdchbi = 1, RefSex = 0, RefAge = 40), interval = "prediction"))
       fit      lwr      upr
1 21.56979 14.58021 31.91008
```

waarbij het object `model` MODEL 5 bevat.

iii. Wat kunt u besluiten uit de p-waarde in de output

```
Analysis of Variance Table

Model 1: logBMI ~ School_meal + black + mexam + pir200_plus + RefSex
   + RefAge + WIC + Food_Stamp + fsdchbi + factor(age)
Model 2: logBMI ~ School_meal * black + School_meal * mexam + School_meal *
    pir200_plus + RefSex + RefAge + WIC + Food_Stamp + fsdchbi +
    factor(age)
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1   2307 91.037
2   2304 90.692  3   0.34521 2.9233 0.03273 *
---
Signif. codes:
0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

iv. Interpreteer de 2 schattingen ('Estimate') in de volgende output:

```
> model = glm(logBMI ~ School_meal*black + School_meal*mexam
   + School_meal*pir200_plus + RefSex + RefAge + WIC + Food_Stamp
   + fsdchbi +  factor(age),data = nhanes_bmi)
> summary(stdGlm(model,X="School_meal",data = nhanes_bmi,
   subsetnew = (nhanes_bmi$School_meal == 1)))

  Estimate Std. Error lower 0.95 upper 0.95
0     2.99     0.0106       2.97       3.01
1     2.98     0.0068       2.97       3.00
```

De R-help file vermeldt het volgende rond het argument `subsetnew`: *an optional logical statement specifying a subset of observations to be used in the standardization. This set is assumed to be a subset of the subset (if any) that was used to fit the regression model.*

2. Stel dat we een lineair regressiemodel met 2 predictoren fitten. Verduidelijk waarom het voor onderzoek van heteroscedasticiteit belangrijk is om zich te baseren op figuren van de *residu's* in functie van elk van de predictoren, eerder dan figuren van de *uitkomsten* in functie elk van de predictoren.

3. Beschouw een uitkomst Y en twee predictoren X en Z, waarbij X en Z onafhankelijk en gemiddeld 0 zijn.

(a) Is het model

$$E(XZ|X,Z) = \alpha_0 + \alpha_1 X + \alpha_2 Z$$

volgens u correct gespecificeerd? Leg uit.

(b) Toon aan dat de OLS schatter voor $(\alpha_0, \alpha_1, \alpha_2)$ naar $(0,0,0)$ convergeert. Het is niet nodig om hiertoe matrices te berekenen en te inverteren.

(c) Stel dat in werkelijkheid

$$E(Y|X,Z) = \omega_1(X) + \omega_2(Z) + \gamma XZ,$$

voor bepaalde functies $\omega_1(.)$ en $\omega_2(.)$. Gebruik dan het resultaat uit de vorige vraag om aan te tonen dat de OLS schatter voor β_3 in model

$$E(Y|X,Z) = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ$$

asymptotisch onvertekend is voor γ.

(d) Wat leert u uit de eigenschap van OLS schatters in de vorige vraag?