Statistische modellen en data-analyse: project 3

Toon Nolten, r0258654
<toon.nolten@student.kuleuven.be>

Juni 2014

Inleiding

De dataset "Regios.xlsx" bevat data over 180 steden en gemeenten. De opgave bestaat erin een regressiemodel op te stellen om te voorspellen of een observatie een stad of gemeente is. Dit model wordt dan vergeleken met een discriminant analyse voor classificatie.

Ik heb één waarde gecorrigeerd in de dataset, namelijk voor het 'Migratie saldo' van de gemeente Meerhout, dit was 23 wat erg opvalt in vergelijking met de andere waarden, dit moest dan ook verbeterd worden naar 0.23. (Thomas Neven heeft hierop gewezen.)

Opgave 1: Gebruik een regressiemodel om de respons 'Stad' te verklaren aan de hand van de andere beschikbare variabelen.

Omdat de respons een binaire variabele is heb ik logistische regressie gebruikt om een model op te stellen.

Opgave 2: Geeft dit regressiemodel een aanvaardbare fit?

De residual deviance van het model is 114.85, dit is kleiner dan $\chi^2_{167,0.05}$ dus geeft het model een goede fit.

Opgave 3: Is de variabele 'Regio' nuttig in het regressiemodel of kan deze variabele beter weggelaten worden?

De likelihood ratio test toont aan dat de variabele Regio niet belangrijk is in het model. $Partialdeviance = 0.0079 << \chi^2_{1,0.95} = 3.84$ dus we aanvaarden H_0.

Opgave 4: Zijn er invloedrijke observaties of andere afwijkingen die best opgelost worden om een beter model te bekomen?

Uit de plot van de deviance residuals, figuur 1, is duidelijk dat de enige uitschieter 'Mesen' is.

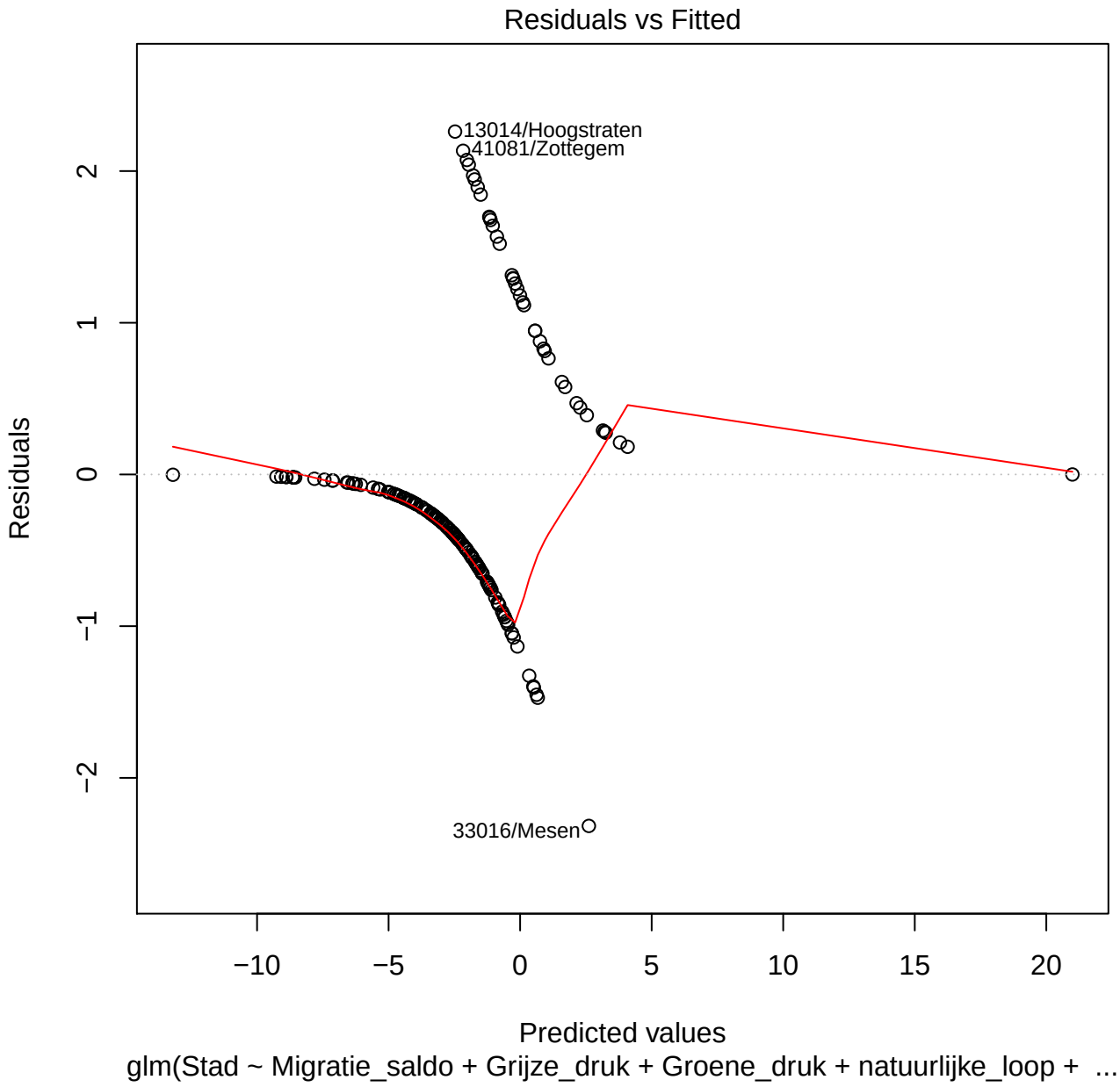

Figure 1: Plot of Deviance Residuals

Opgave 5: Hoe goed kan je met dit model steden van gemeenten onderscheiden?

Het model heeft een lage (14.4%) prediction error, tabel 1. Van de 180 observaties worden 4 gemeenten geclassificeerd als steden en 22 steden geclassificeerd als gemeentes. Dit in de aanname dat de misclassificatie voor beiden dezelfde kost heeft (threshold 0.5).

Voorspelling	Gemeente	Stad
Gemeente	137	4
Stad	22	17

Table 1: Classification table for Logistic Regression

Opgave 6: Via discriminantanalyse kan je ook steden van gemeenten proberen te onderscheiden. Levert dit een beter resultaat op dan met voorgaand model?

Het levert een vergelijkbaar resultaat op, tabel 2, prediction error 15.6%. Maar de threshold voor classificatie bij de logistische regressie is momenteel niet afgesteld op de data en kan mogelijk nog voor verbetering zorgen.

Voorspelling	Gemeente	Stad
Gemeente	136	5
Stad	23	16

Table 2: Classification table for Linear Discriminant Analysis

Appendix: Code

```
# De dataset regios.xlsx bevat voor de steden en gemeenten uit 8
# Vlaamse regio's gegevens voor de volgende variabelen in het jaar
# 2010:

# +----------------+-------------------------------------------------+
# |                | Relatief migratie saldo (% van totale bevolking) |
# | Demografisch   | Grijze druk (% 65+ t.o.v. 20-64)                |
# |                | Groene druk (% -20 t.o.v. 20-64)                |
# |                | natuurlijke loop (+1 bij aangroei, -1 bij daling)|
# +----------------+-------------------------------------------------+
# |                | % belastingsaangiften > 50.000 euro             |
# |                | Gemiddeld inkomen per aangifte                  |
# |                | werkzaamheidsgraad (% werkenden van (18-64)     |
# | Economisch     |                          bevolking)             |
# |                | werkloosheidsgraad (% werkzoekenden van (18-64) |
# |                |                          bevolking)             |
# |                | Gemiddelde verkoopprijs van woonhuizen          |
# +----------------+-------------------------------------------------+
# |                | % leefloners t.o.v. totale bevolking            |
# | Armoede        | % geboorten in kansarme gezinnen                |
# |                | % bejaarden met inkomensgarantie (i.e. uitkering)|
# +----------------+-------------------------------------------------+
# | Geografisch    | Regio                                           |
# |                | Stad                                            |
# +----------------+-------------------------------------------------+

# De variabele 'Regio' is een indicator die aangeeft welke steden en
# gemeenten tot eenzelfde regio behoren. De variabele 'Stad' geeft
# weer of de observatie een stad (1) of gemeente (0) is.
# Je beantwoordt de onderstaande vragen door individueel gepaste
# analyses uit te voeren met R. De bespreking van de resultaten en
# de nodige figuren verwerk je in een schriftelijk rapport dat
# maximaal uit 6 bladzijden mag bestaan (12pt lettergrootte). In een
# appendix voeg je de gebruikte R-code toe. Rapporteer enkel
# resultaten en interpretaties, herhaal geen theorie uit de cursus!
# Het rapport dien je in pdf vorm in ten laatste op *12 juni 2014*
# via Toledo. Dit project telt mee voor 5 punten van het
# eindresultaat.
```

```
40 # Opgaven:
41 # Het doel is om te onderzoeken of steden van gemeenten kunnen
42 # onderscheiden worden op basis van de beschikbare variabelen.
43
44 library(gdata)
45 turnhout = read.xls(xls='Regios.xlsx', sheet=1, pattern='druk')
46 rownames(turnhout) = turnhout[,1]
47 turnhout = turnhout[,-1]
48 colnames(turnhout) = c('Migratie_saldo',
49                        'Grijze_druk',
50                        'Groene_druk',
51                        'natuurlijke_loop',
52                        '%_aangiften_>_50.000_euro',
53                        'Gemiddeld_inkomen_per_aangifte',
54                        'werkzaamheidsgraad',
55                        'werkloosheidsgraad',
56                        '%_leefloners',
57                        '%_geboorten_in_kansarme_gezinnen',
58                        '%_bejaarden_met_inkomensgarantie',
59                        'Gemiddelde_verkoopprijs_van_woonhuizen',
60                        'Regio',
61                        'Stad')
62 for(column in 1:ncol(turnhout)) {
63     turnhout[,column] = as.numeric(as.character(turnhout[,column]))
64 }
65
66 cat('\n ', 'Summary Regios.xlsx:', '\n',
67     '------------------------------', '\n')
68 print(summary(turnhout))
```

```r
# 1. Gebruik een regressiemodel om de respons 'Stad' te verklaren
#    aan de hand van de andere beschikbare variabelen.

turnhout.glm = glm(Stad
                   ~Migratie_saldo
                   +Grijze_druk
                   +Groene_druk
                   +natuurlijke_loop
                   +`%_aangiften_>_50.000_euro`
                   +Gemiddeld_inkomen_per_aangifte
                   +werkzaamheidsgraad
                   +werkloosheidsgraad
                   +Gemiddelde_verkoopprijs_van_woonhuizen
                   +`%_leefloners`
                   +`%_geboorten_in_kansarme_gezinnen`
                   +`%_bejaarden_met_inkomensgarantie`
                   +Regio, family=binomial, data=turnhout)
cat('\n ', 'Summary turnhout.glm:', '\n',
    '--------------------------', '\n')
print(summary(turnhout.glm))

cat('\n ', 'Anova turnhout.glm:', '\n',
    '----------------------', '\n')
print(anova(turnhout.glm, test="Chisq"))


# 2. Geeft dit regressiemodel een aanvaardbare fit?

cat('\n ', 'Acceptable fit', '\n',
    '----------------', '\n')
cat('  ', 'Chi^2_167,0.05: ', qchisq(0.05, 167), '\n') # = 138.1184

# Ja, chi^2_167,0.05 = 138.1184, de model deviance is 114.85 dus
# we aanvaarden het model.

# 3. Is de variabele 'Regio' nuttig in het regressiemodel of kan
#    deze variabele beter weggelaten worden?

turnhout.noregio = update(turnhout.glm, .~.-Regio)

cat('\n ', 'Likelihood ratio test for Regio', '\n',
    '-----------------------------------', '\n')
cat('  ', 'Partial Deviance: ',
    turnhout.noregio$deviance - turnhout.glm$deviance, '\n')
cat('  ', 'Chi^2_1,0.95: ', qchisq(0.95, 1), '\n')

# De likelihood ratio test toont aan dat de variabele Regio niet
# belangrijk is in het model. 0.0079 << 3.84 dus we aanvaarden H_0.
```

```
120 # 4. Zijn er invloedrijke observaties of andere afwijkingen die best
121 #    opgelost worden om een beter model te bekomen?
122
123 ##pdf('deviance_residuals.pdf')
124 #plot(turnhout.noregio, which=1) # residuals plot
125 ##dev.off()
126
127 # Uit de plot van de deviance residuals is duidelijk dat de enige
128 # uitschieter 'Mesen' is.
129
130 # 5. Hoe goed kan je met dit model steden van gemeenten
131 #    onderscheiden?
132
133 logitpred = predict(turnhout.noregio) > 0.5
134 cat('\n ', 'Classification table:', '\n',
135     '-------------------------', '\n')
136 print(table(turnhout$Stad, logitpred))
137 cat('  ', 'Prediction error: ',
138     round((22+4)/(137+4+22+17)*100, 2), '%\n')
139
140 # Het model heeft een lage (14.4%) prediction error. Van de 180
141 # observaties worden 4 gemeenten geclassificeerd als steden en 22
142 # steden geclassificeerd als gemeentes. Dit in de aanname dat
143 # de misclassificatie voor beiden dezelfde kost heeft
144 # (threshold 0.5) .
```

```
146 # 6. Via discriminantanalyse kan je ook steden van gemeenten
147 #    proberen te onderscheiden. Levert dit een beter resultaat op dan
148 #    met voorgaand model?
149
150 library(MASS)
151 turnhout.lda = lda(Stad
152                    ~Migratie_saldo
153                    +Grijze_druk
154                    +Groene_druk
155                    +natuurlijke_loop
156                    +`%_aangiften_>_50.000_euro`
157                    +Gemiddeld_inkomen_per_aangifte
158                    +werkzaamheidsgraad
159                    +werkloosheidsgraad
160                    +Gemiddelde_verkoopprijs_van_woonhuizen
161                    +`%_leefloners`
162                    +`%_geboorten_in_kansarme_gezinnen`
163                    +`%_bejaarden_met_inkomensgarantie`
164                    +Regio, data=turnhout)
165
166 cat('\n ', 'Summary turnhout.lda:', '\n',
167     '------------------------', '\n')
168 print(turnhout.lda)
169
170 ldapred = predict(turnhout.lda, turnhout)$class
171 cat('\n ', 'Classification table:', '\n',
172     '------------------------', '\n')
173 print(table(turnhout$Stad, ldapred))
174 cat('  ', 'Prediction error: ',
175     round((23+5)/(137+4+22+17)*100, 2), '%\n')
176
177 # Het levert een vergelijkbaar resultaat op, prediction error 15.6%.
178 # Maar de threshold voor classificatie bij de logistische regressie
179 # is momenteel niet afgesteld op de data en kan mogelijk nog
180 # voor verbetering zorgen.
```