Statistische modellen en data-analyse: project 3

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

September 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.)

1 Opgaven

1.1 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. Dit heb ik gedaan door eerst een model op te stellen met alle variabelen als predictors:

```
glm(Stad ~., family=binomial, data=turnhout)
```

Op dat model heb ik dan een stapsgewijze verbetering uitgevoerd gebaseerd op de AIC:

```
step(turnhout.glm)
```

Dit levert het volgende model op:

```
glm(Stad ~ Grijze_druk + Groene_druk + Gemmideld_inkomen_per_aangifte
    + `%_leefloners`, family=binomial, data=turnhout)
```

Als we ook interactiecomponenten in rekening brengen kunnen we het model nog verbeteren maar dit gaat ten koste aan de eenvoud en de interpreteerbaarheid van het model en levert in dit geval onvoldoende verbetering ($\pm 5\%$ minder classificatiefouten).

(Er bestaan betere modelselectiemethoden maar die laten we hier buiten beschouwing, cf. het R pakket 'glmnet')

1.2 Geeft dit regressiemodel een aanvaardbare fit?

Dit gaan we na met een 'Goodness of fit' test waarbij de nul-hypothese betekent dat we het model aanvaarden. De residual deviance van het model is 118.57, dit is kleiner dan $\chi^2_{175,0.05} = 145.4058$ dus op een betrouwbaarheidsniveau van 5% besluiten we dat het model een goede fit geeft.

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

De variabele 'Regio' zit niet in het model, ik heb dus het model dat gebruik maakt van alle predictor variabelen opgesteld om deze vergelijking op te doen.

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, de variabele is niet nuttig.

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

Op de plot van de deviance residuals, figuur 1, is de enige duidelijke uitschieter 'Mesen'. In dit geval betekent dit dat 'Mesen' een twijfelgeval is. Hiervoor is er niet echt een oplossing als we een model willen opstellen voor alle steden en gemeenten in belgië, omdat we niet kunnen bepalen of dit een echte uitschieter is of dat de steekproef een bias heeft.

1.5 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 3 gemeenten geclassificeerd als steden en 23 steden geclassificeerd als gemeentes. Dit in de aanname dat de misclassificatie voor beiden dezelfde kost heeft (threshold 0.5). Wat wel opvalt is dat de meeste steden verkeerd ingedeeld zijn, dit is omdat er veel meer gemeentes dan steden in de steekproef aanwezig zijn, daardoor worden nieuwe observaties sneller ingedeeld als gemeentes.

Voorspelling	Gemeente	Stad
Gemeente	138	3
Stad	23	16

Table 1: Classification table for Logistic Regression

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

Het gebruikte model is gebaseerd op dezelfde variabelen als in het verbeterde logistische model:

```
lda(Stad ~ Grijze_druk + Groene_druk + Gemmideld_inkomen_per_aangifte
    + `%_leefloners`, data = turnhout)
```

Het levert een vergelijkbaar resultaat op, tabel 2, prediction error 15%. 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	140	1
Stad	26	13

Table 2: Classification table for Linear Discriminant Analysis

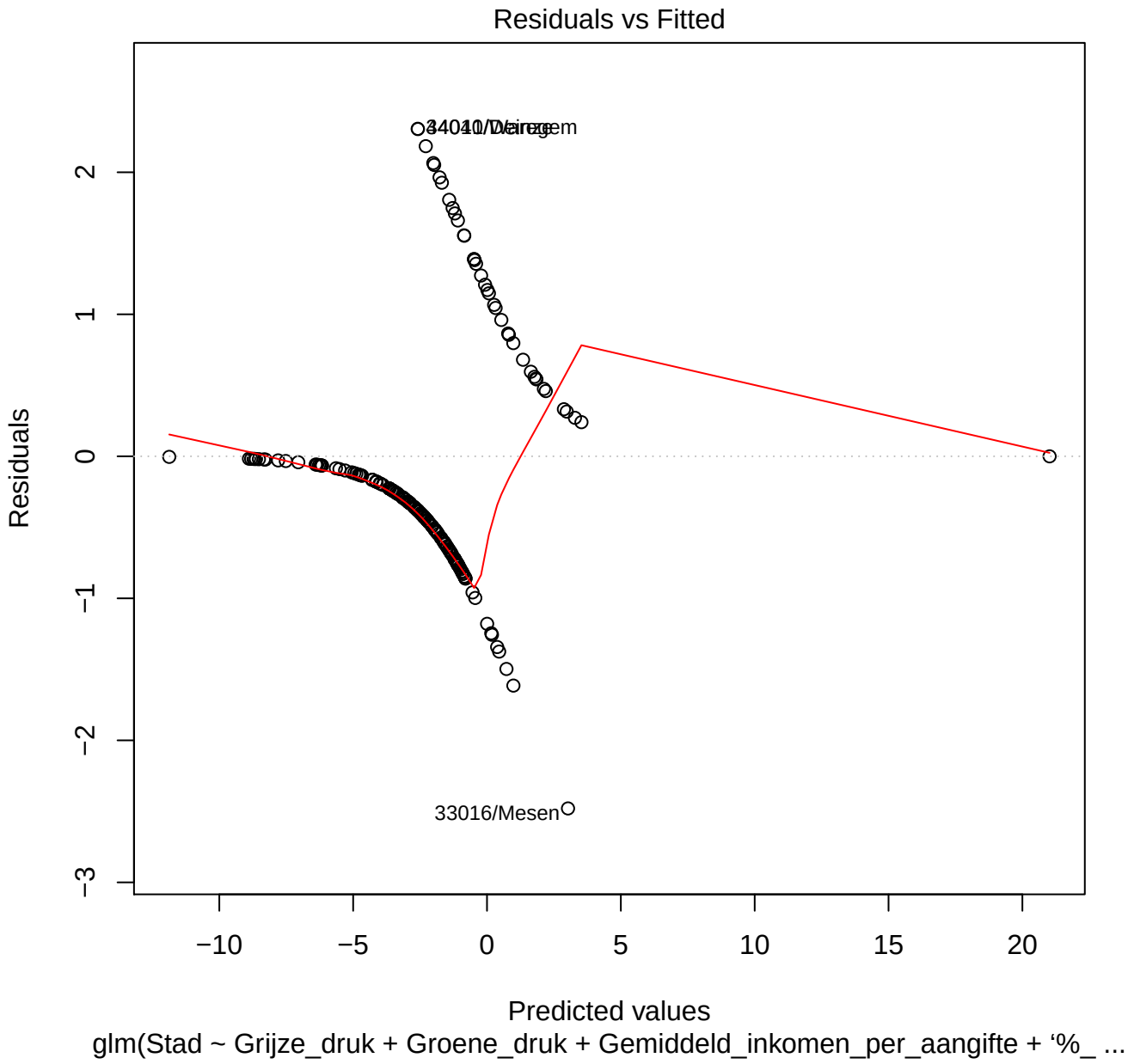

Figure 1: Plot of Deviance Residuals

2 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.
```

```
# Opgaven:
# Het doel is om te onderzoeken of steden van gemeenten kunnen
# onderscheiden worden op basis van de beschikbare variabelen.

library(gdata)
turnhout = read.xls(xls='Regios.xlsx', sheet=1, pattern='druk')
rownames(turnhout) = turnhout[,1]
turnhout = turnhout[,-1]
colnames(turnhout) = c('Migratie_saldo',
                       'Grijze_druk',
                       'Groene_druk',
                       'natuurlijke_loop',
                       '%_aangiften_>_50.000_euro',
                       'Gemiddeld_inkomen_per_aangifte',
                       'werkzaamheidsgraad',
                       'werkloosheidsgraad',
                       '%_leefloners',
                       '%_geboorten_in_kansarme_gezinnen',
                       '%_bejaarden_met_inkomensgarantie',
                       'Gemiddelde_verkoopprijs_van_woonhuizen',
                       'Regio',
                       'Stad')
for(column in 1:ncol(turnhout)) {
    turnhout[,column] = as.numeric(as.character(turnhout[,column]))
}

cat('\n ', 'Summary Regios.xlsx:', '\n',
    '------------------------', '\n')
print(summary(turnhout))
```

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

turnhout.glm = glm (Stad ~., family=binomial, data=turnhout)
turnhout.glms = step(turnhout.glm, ~.)

cat('\n ', 'Summary turnhout.glms:', '\n',
    '-----------------------', '\n')
print(summary(turnhout.glms))

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


# 2. Geeft dit regressiemodel een aanvaardbare fit?

cat('\n ', 'Acceptable fit', '\n',
    '-----------------', '\n')
cat('  ', 'Chi^2_175,0.05: ', qchisq(0.05, 175), '\n') # = 145.4058

# Ja, chi^2_175,0.05 = 145.4058, de model deviance is 118.57 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.


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

##pdf('deviance_residuals.pdf')
#plot(turnhout.glms, which=1) # residuals plot
##dev.off()

# Uit de plot van de deviance residuals is duidelijk dat de enige
# uitschieter `Mesen' is.
```

```
# 5. Hoe goed kan je met dit model steden van gemeenten
#    onderscheiden?

logitpred = predict(turnhout.glms) > 0.5
cat('\n ', 'Classification table:', '\n',
    '-----------------------', '\n')
logt = table(turnhout$Stad, logitpred)
print(logt)
cat('  ', 'Prediction error: ',
    round((logt[1,2]+logt[2,1])/(sum(logt))*100, 2), '%\n')

# Het model heeft een lage (14.4%) prediction error. Van de 180
# observaties worden 3 gemeenten geclassificeerd als steden en 23
# steden geclassificeerd als gemeentes. Dit in de aanname dat
# de misclassificatie voor beiden dezelfde kost heeft
# (threshold 0.5) .

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

library(MASS)
turnhout.lda = lda(Stad
                   ~Grijze_druk
                   +Groene_druk
                   +Gemiddeld_inkomen_per_aangifte
                   +`%_leefloners`, data=turnhout)

cat('\n ', 'Summary turnhout.lda:', '\n',
    '-----------------------', '\n')
print(turnhout.lda)

ldapred = predict(turnhout.lda, turnhout)$class
cat('\n ', 'Classification table:', '\n',
    '-----------------------', '\n')
ldat = table(turnhout$Stad, ldapred)
print(ldat)
cat('  ', 'Prediction error: ',
    round((ldat[1,2]+ldat[2,1])/(sum(ldat))*100, 2), '%\n')

# Het levert een vergelijkbaar resultaat op, prediction error 15%.
# Maar de threshold voor classificatie bij de logistische regressie
# is momenteel niet afgesteld op de data en kan mogelijk nog
# voor verbetering zorgen.
```