NLNP Praktikum 6

Robin Baudisch, Merlin Kopfmann, Maximilian Neudert

Inhaltsverzeichnis

A1

a)

```
load("awards.RData")
load("DebTrivedi.RData")
plm = glm(num_awards ~ prog + math, data = awards, family = poisson)
lambda_voc <- predict.glm(plm, data.frame(prog = "Vocational",
    math = 60), type = "response")
lambda_acd <- predict.glm(plm, data.frame(prog = "Academic",
    math = 60), type = "response")
lambda_gen <- predict.glm(plm, data.frame(prog = "General", math = 60),
    type = "response")
p1 = round(1 - sum(dpois(0:2, lambda_gen)), 4)
p2 = round(1 - sum(dpois(0:2, lambda_voc)), 4)
p3 = round(1 - sum(dpois(0:2, lambda_acd)), 4)
p = c(p1, p2, p3)
names = c("Vocational", "Academic", "General")
df = data.frame(class = names, chance = p)
df
```

class	chance
Vocational	0.0057
Academic	0.0154
General	0.0891

b)

```
plm_without = glm(num_awards ~ math, data = awards, family = poisson)
lmtest::lrtest(plm, plm_without)
```

#Df	LogLik	Df	Chisq	Pr(>Chisq)
4	-182.7523	NA	NA	NA
2	-190.0381	-2	14.57168	0.0006852

```
AIC(plm_without, plm)
```

	df	AIC
plm_without	2	384.0762
plm	4	373.5045

Um zu überprüfen, ob der Ausbildungstyp einen signifikanten Einfluss auf die Anzahl an Awards hat, wurden zwei Poissonregressionen gefittet (1. num_awards ~ prog + math; 2. num_awards ~ math) und anschließend mittels AIC und Likelihood-Ratio-Test miteinander verglichen.

Beide Vergleichsmethoden kommen zum Ergebnis, dass der Ausbildungsyp („prog“) einen Einfluss auf die Zielgröße („num_awards“) hat.

c)

```
anova <- aov(num_awards ~ prog + math, data = awards)
summary(anova)
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)
prog          2  30.30  15.150   18.62 3.94e-08 ***
math          1  30.88  30.877   37.96 4.03e-09 ***
Residuals   196 159.44   0.813
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Um den Einfluss der Variablen auf die Zielgröße zu überprüfen, wurde eine Varianzanalyse (Anova) durchgeführt. Laut dieser haben sowohl math als auch prog einen signifikaten Einfluss auf num_awards.

```
chisq.test(awards$math, awards$num_awards)
```

```
    Pearson's Chi-squared test

data:  awards$math and awards$num_awards
X-squared = 421.09, df = 234, p-value = 7.602e-13
```

Ein χ^2-Test liefert für math zu $\alpha = 0.05$ das ein signifikantes Ergebnis. Die Nullhypothese, dass math und num_awards unabhängig sind wird verworfen und man kann davon ausgehen, dass math einen signifikanten Einfluss auf num_awards hat.

```
chisq.test(awards$prog, awards$num_awards)
```

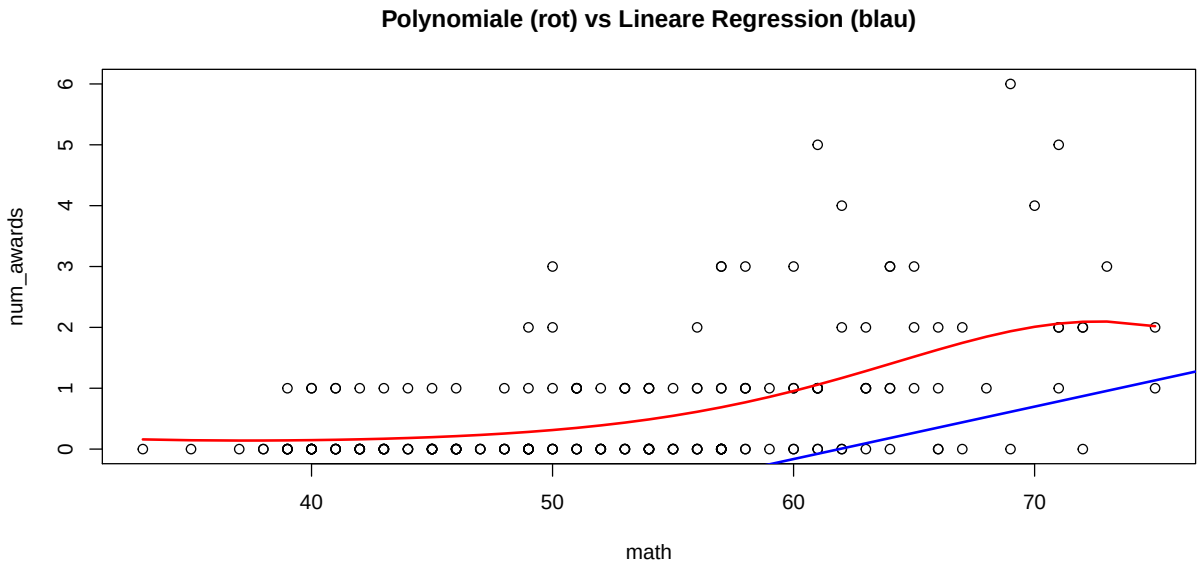
```
    Pearson's Chi-squared test

data:  awards$prog and awards$num_awards
X-squared = 31.46, df = 12, p-value = 0.001676
```

Ein χ^2-Test liefert für prog zu $\alpha = 0.05$ das ein signifikantes Ergebnis. Die Nullhypothese, dass prog und num_awards unabhängig sind wird verworfen und man kann davon ausgehen, dass prog einen signifikanten Einfluss auf num_awards hat.

d)

```
poly <- glm(num_awards ~ math + I(math^2) + I(math^3), data = awards,
    family = "poisson")
title = "Polynomiale (rot) vs Lineare Regression (blau)"
plot(num_awards ~ math, data = awards, main = title)
lines(sort(awards$math), fitted(poly)[order(awards$math)], col = "red",
    lw = 2)
abline(plm_without, col = "blue", lw = 2)
```




```
AIC(poly, plm_without)
```

	df	AIC
poly	4	385.5319
plm_without	2	384.0762

```
anova(poly, plm_without, test = "Chisq")
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
196	201.4771	NA	NA	NA
198	204.0213	-2	-2.544244	0.2802363

Laut AIC ist das polynomiale Modell 3. Grades nicht signifikant besser als das lineare Modell (beide mit `math` als einziger Kovariate).

A2

a)

```
dev = c(dev_model, dev_manual)
names = c("model", "manual")
df = data.frame(method = names, deviance = dev)
df
```

method	deviance
model	24178.54
manual	24178.51

b)

```
summary(poisreg)
```

```
Call:
glm(formula = ofp ~ health + numchron + hosp + married + medicaid,
    family = "poisson", data = DebTrivedi)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.2623  -2.0484  -0.6898   0.7949  16.1776

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.490539   0.013363 111.538   <2e-16 ***
healthpoor       0.338277   0.016536  20.457   <2e-16 ***
healthexcellent -0.372756   0.030241 -12.326   <2e-16 ***
numchron         0.095171   0.006124  15.541   <2e-16 ***
hosp             0.500663   0.013982  35.809   <2e-16 ***
marriedyes      -0.019759   0.012860  -1.537    0.124
medicaidyes      0.032753   0.021101   1.552    0.121
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 26943  on 4405  degrees of freedom
Residual deviance: 24179  on 4399  degrees of freedom
AIC: 36968

Number of Fisher Scoring iterations: 5
```

```
pchisq(poisreg$deviance, df = poisreg$df.residual)
```

```
[1] 1
```

```
nrow(DebTrivedi)
```

```
[1] 4406
```

Um die Abweichung als Gütemetrik zu nutzen, müssen wir unter der Annahme, dass unser Modell korrekt ist, herausfinden, wie viel Variation wir bei den beobachteten Ergebnissen um ihre vorhergesagten Mittel herum erwarten würden.

Da die Abweichung als Likelihood-Ratio-Test zum Vergleich des aktuellen Modells mit dem gesättigten Modell abgeleitet werden kann, wird vermutet, dass (vorausgesetzt das Modell ist korrekt spezifiziert) die Abweichung einer Chi-Quadrat-Verteilung folgt, deren Freiheitsgrade der Differenz in der Anzahl der Parameter entsprechen. Das gesättigte Modell kann als ein Modell betrachtet werden, das für jede Beobachtung einen eigenen Parameter verwendet und somit n Parameter hat. Wenn unser Modell p-Parameter hat, bedeutet dies, dass die Abweichung mit einer Chi-Quadrat-Verteilung auf n-p-Parameter verglichen wird.

Die Abweichung wird hier von der glm-Funktion als „residual deviance“ bezeichnet, hier 24179. Es gibt 4406 Beobachtungen, und unser Modell hat sechs Parameter, so dass die Freiheitsgrade 4399 sind, angegeben durch df.residual. Um den p-Wert für die Varianzgüte des Fit-Tests zu berechnen, berechnen wir einfach die Wahrscheinlichkeit rechts neben dem Varianzwert für die Chi-Quadrat-Verteilung auf 4399 Freiheitsgrade.

Die Nullhypothese ist, dass unser Modell korrekt spezifiziert ist. Ein p-Wert von 1 spricht für ein gut gefittetes Modell.

c)

```
poisreg2 <- glm(ofp ~ health + numchron + hosp + married + medicaid,
    data = DebTrivedi, family = "quasipoisson")
summary(poisreg2)
```

```
Call:
glm(formula = ofp ~ health + numchron + hosp + married + medicaid,
    family = "quasipoisson", data = DebTrivedi)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.2623  -2.0484  -0.6898   0.7949  16.1776

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.49054    0.03558  41.898  < 2e-16 ***
healthpoor      0.33828    0.04402   7.684 1.88e-14 ***
healthexcellent -0.37276    0.08050  -4.630 3.76e-06 ***
numchron        0.09517    0.01630   5.838 5.67e-09 ***
hosp            0.50066    0.03722  13.451  < 2e-16 ***
marriedyes     -0.01976    0.03423  -0.577    0.564
medicaidyes     0.03275    0.05617   0.583    0.560
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 7.086866)

    Null deviance: 26943  on 4405  degrees of freedom
Residual deviance: 24179  on 4399  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

Bei der Poissonregression treffen wir eine starke Modellannahme:

Da bei der Poissonverteilung λ gleich der Erwartungswert, ALS AUCH der Varianz ist, nehmen wir dies auch für die Verteilung in unserem Modell an. Dies ist häufig nicht der Fall.

Überdispersion ist ein Problem, wenn die bedingte Varianz größer ist als der bedingte Mittelwert. Um den Überdispersionparameter zu schätzen, fitten wir ein Quasi-Poisson-Modell auf unsere Daten.

Laut dem neuen Modell ist der geschätzte Überdispersionsparameter bei ~ 7. Das heißt, die bedingte Varianz ist 7-mal größer als der bedingte Mittelwert.

d)

```r
summary(poisreg)
```

```
Call:
glm(formula = ofp ~ health + numchron + hosp + married + medicaid,
    family = ”poisson”, data = DebTrivedi)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.2623  -2.0484  -0.6898   0.7949  16.1776

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     1.490539   0.013363 111.538   <2e-16 ***
healthpoor      0.338277   0.016536  20.457   <2e-16 ***
healthexcellent -0.372756   0.030241 -12.326   <2e-16 ***
numchron        0.095171   0.006124  15.541   <2e-16 ***
hosp            0.500663   0.013982  35.809   <2e-16 ***
marriedyes     -0.019759   0.012860  -1.537    0.124
medicaidyes     0.032753   0.021101   1.552    0.121
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 26943  on 4405  degrees of freedom
```

```
Residual deviance: 24179  on 4399  degrees of freedom
AIC: 36968

Number of Fisher Scoring iterations: 5
```

```
summary(poisreg2)
```

```
Call:
glm(formula = ofp ~ health + numchron + hosp + married + medicaid,
    family = ”quasipoisson”, data = DebTrivedi)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.2623  -2.0484  -0.6898   0.7949  16.1776

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.49054    0.03558  41.898  < 2e-16 ***
healthpoor       0.33828    0.04402   7.684 1.88e-14 ***
healthexcellent -0.37276    0.08050  -4.630 3.76e-06 ***
numchron         0.09517    0.01630   5.838 5.67e-09 ***
hosp             0.50066    0.03722  13.451  < 2e-16 ***
marriedyes      -0.01976    0.03423  -0.577    0.564
medicaidyes      0.03275    0.05617   0.583    0.560
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 7.086866)

    Null deviance: 26943  on 4405  degrees of freedom
Residual deviance: 24179  on 4399  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

Es ändern sich nur die p-Werte der Koeffizienten. Dies rührt von der Veränderung der Verteilungsannahme (von Poisson zu Quasi-Poisson).

e)

	hosp.0	hosp.1
poor	2578	2351
average	14014	5323
excellent	1018	158

	hosp.0	hosp.1
poor	7.032564	11.814785
average	4.920448	8.266407
excellent	3.236101	5.436685

chisq
44.29864

- H_0 : {Erwartete und beobachtete Häufigkeiten sind gleichverteilt}
- H_1 : {Erwartete und beobachtete Häufigkeiten sind nicht gleichverteilt}

Der errechnete Wert des Chi-Quadrat Anpassungstests liegt über 11.07 und damit im Ablehnungsbereich. H0 kann also verworfen und H1 angenommen werden.

Da die Koeffizienten beider Modelle identisch sind, liefern die Predictions auch identische Werte. Der Anpassungstest weist folglich nicht auf eine verbesserte Anpassung des erweiterten Modells hin.