|  |  |
| --- | --- |
| **Biostokastikum** | 2014-06-12 |

Overdispersion: Programming code and references

## R code

### Example Leaves

## Observations

Plant <- factor(1:20) ; Plant

Treatment <- factor(c(rep("Active", 10), rep("Control", 10)))

Infested <- c(6, 3, 7, 1, 0, 0, 4, 9, 10, 2, 5, 9, 14, 3, 20, 8, 7, 7, 8, 5)

N <- c(20, 20, 20, 20, 18, 20, 18, 20, 20, 20, 20, 19, 20, 20, 20, 20, 15, 20, 20, 20)

BinMat <- matrix(c(Infested, N - Infested), ncol = 2) ; BinMat

# Alternatively, write BinMat <- cbind(Infested, N - Infested)

## Model with independent Bernoulli observations

Model.1 <- glm(BinMat ~ Treatment, family = binomial)

summary(Model.1) # The data is overdispersed: deviance = 94.5274 on 18 degrees of freedom ;

## Generalized linear mixed model

library(MASS)

Model.Mix <- glmmPQL(BinMat ~ Treatment, random = ~ 1 | Plant, family=binomial(link="logit"))

summary(Model.Mix)

library(lme4)

Model.Mixed <- glmer(BinMat ~ Treatment + (1 | Plant), family = binomial, nAGQ = 25)

Model.Null <- glmer(BinMat ~ (1 | Plant), family = binomial, nAGQ = 25)

summary(Model.Mixed)

anova(Model.Null, Model.Mixed) # P = 0.017

## Adjustment using dispersion factor

Model.Adjusted <- glm(BinMat ~ Treatment, family = quasibinomial)

anova(Model.Adjusted, test = "F") # P = 0.0331

## GEE

library(geepack)

Model.GEE <- geeglm(BinMat ~ Treatment, family = binomial, id = Plant, corstr = "independence")

# There was only one observation per subject (i.e. Plant),

# so it was not necessary to specify any more complicated covariance structure

summary(Model.GEE)

anova(Model.GEE) # P = 0.017

### Example Seed mixes

## Observations

Block <- factor(c(rep(1, 4), rep(2, 4), rep(3, 4), rep(4, 4)))

Mix <- factor(rep(1:4, 4))

Count <- c(24, 12, 8, 13, 9, 9, 9, 18, 12, 8, 44, 0, 8, 12, 25, 0)

Error <- 1:16 ; Error

## Poisson model

Model.1 <- glm(Count ~ Block + Mix , family = poisson)

Model.2 <- glm(Count ~ Block, family = poisson)

anova(Model.2, Model.1, test = "Chisq") # P < 0.0001

summary(Model.1) # Deviance = 90.23 on 9 degrees of freedom. The data is overdispersed.

## Adjustment using dispersion factor

Model.Q1 <- glm(Count ~ Block + Mix , family = quasipoisson)

summary(Model.Q1)

Model.Q2 <- glm(Count ~ Block, family = quasipoisson)

anova(Model.Q2, Model.Q1, test = "F") # P = 0.3749

library(lsmeans)

lsmeans(Model.Q1, "Mix")

### Example Sheep milk

Method <- c(rep("Mechanical", 10), rep("Manual", 10)) ; Method

Count <- c(2966, 569, 59, 1887, 3452, 189, 93, 618, 130, 2493,

186, 107, 65, 126, 123, 164, 408, 324, 548, 139)

obs <- 1:20

## Poisson model

Model.Poisson <- glm(Count ~ Method, family = poisson)

summary(Model.Poisson)

Pearson.resid <- residuals(Model.Poisson, type = "pearson")

plot(as.numeric(Pearson.resid) ~ obs, ylab = "Pearson residuals", xlab = "Observation number", ylim = c(-60, 60),

cex.axis = 1.2, cex.lab = 1.5, pch = 21, cex = 1.5, col = "black", bg = "red")

abline(h = 0, lty = "dashed")

## Negative binomial model

library(MASS)

Model.NegBin <- glm.nb(Count ~ Method)

summary(Model.NegBin)

deviance <- Model.NegBin$deviance ; deviance # Deviance = 22.77

df <- Model.NegBin$df.residual ; df # Residual df = 18

pchisq(deviance, df, lower.tail = FALSE) # P = 0.200

Pearson.residuals <- Model.NegBin$residuals\*Model.NegBin$theta/sqrt(Model.NegBin$w)

plot(Pearson.residuals ~ obs, ylab = "Pearson residuals", xlab = "Observation number", ylim = c(-2, 2),

cex.axis = 1.2, cex.lab = 1.5, cex = 1.5, pch = 21, col = "black", bg = "green")

abline(h = 0, lty = "dashed")

## SAS code

\*\*\* Example Leaves ;

**data** Leaves ;

input Plant Treatment $ Infested N ;

cards ;

1 Active 6 20

2 Active 3 20

3 Active 7 20

4 Active 1 20

5 Active 0 18

6 Active 0 20

7 Active 4 18

8 Active 9 20

9 Active 10 20

10 Active 2 20

11 Control 5 20

12 Control 9 19

13 Control 14 20

14 Control 3 20

15 Control 20 20

16 Control 8 20

17 Control 7 15

18 Control 7 20

19 Control 8 20

20 Control 5 20

;

**run** ;

\* Model with independent Bernoulli observations ;

**proc** **genmod** data = Leaves ;

class Treatment ;

model Infested/N = Treatment / dist = binomial type3 ;

lsmeans Treatment / ilink pdiff ;

**run** ; \* Wald: P < 0.0001, LR-test: P < 0.0001 ;

\* This result is invalid, because the data is overdispersed:

Deviance = 94.5274 on 18 degrees of freedom ;

\* Analysis on Total Counts ;

**proc** **means** data = Leaves ;

var Infested N ;

by Treatment ;

output out = TotalCounts sum = Infested N ;

**run** ;

**proc** **print** data = TotalCounts ;

**run** ;

**proc** **genmod** data = TotalCounts ;

class Treatment ;

model Infested/N = Treatment / dist = binomial type3 ;

lsmeans Treatment / ilink pdiff ;

**run** ; \* P < 0.0001 ;

\* Adjustment using dispersion factor ;

**proc** **genmod** data = Leaves ;

class Treatment ;

model Infested/N = Treatment / dist = binomial type3 pscale;

ods output ModelFit = Deviance ;

lsmeans Treatment / diff ;

**run** ; \* Wald: 0.0243, LR (F-test): 0.0331 ;

\* Estimate Control: -0.2278 ;

\* Estimate Active: -1.2993 ;

\* Estimate Active - Control: -1.07 ;

**proc** **logistic** data = Leaves ;

class Treatment ;

model Infested/N = Treatment / scale = Pearson ;

**run** ; \* Wald: 0.0243, LR (Chi2-test): 0.03 ;

**proc** **genmod** data = Leaves ;

class Treatment ;

model Infested/N = Treatment / dist = binomial type3 dscale;

**run** ; \* Wald: 0.0388, LR (F-test): 0.0484 ;

\* Test of Overdispersion ;

**data** OverdispersionTest ;

set Deviance ;

Pvalue = **1** - probchi(Value, DF) ;

**run** ;

**proc** **print** data = OverdispersionTest ;

**run** ;

\* Generalized linear mixed model ;

**proc** **glimmix** data = Leaves ;

class Plant Treatment ;

model Infested/N = Treatment / dist = binomial ddfm = Sat solution ;

random Plant ;

**run** ; \* P = 0.0325 ;

\* Estimate Control: -0.193 ;

\* Estimate Active: -1.5441 ;

\* Estimate Active - Control: -1.3511 ;

\* GEE ;

**proc** **genmod** data = Leaves ;

class Plant Treatment ;

model Infested/N = Treatment / dist = binomial type3 ;

repeated subject = Plant / type = exch ; \* type = exch is not needed here, becuase the cluster size is one ;

**run** ; \* Wald: P = 0.0173, Score: P = 0.0323 ;

\* Estimate Control: -0.2278 ;

\* Estimate Active: -1.2993 ;

\* Estimate Active - Control: -1.0715 ;

\*\*\* Example Seed mixes ;

\* Source: Dataset 11.6 A in Littell et al. (1996), p. 600 ;

**data** Method1 ;

input Block Mix Count ;

cards ;

1 1 24

1 2 12

1 3 8

1 4 13

2 1 9

2 2 9

2 3 9

2 4 18

3 1 12

3 2 8

3 3 44

3 4 0

4 1 8

4 2 12

4 3 25

4 4 0

;

**run** ;

\*\* Poisson model ;

**proc** **genmod** data = Method1 ;

class Block Mix ;

model Count = Block Mix / dist = poisson type3 ;

ods output ModelFit = Deviance ;

lsmeans Mix / cl ilink ;

**run** ; \* P < 0.0001 ;

**data** OverdispersionTest ;

set Deviance ;

Pvalue = **1** - probchi(Value, DF) ;

**run** ;

**proc** **print** data = OverdispersionTest ;

**run** ;

\*\* Adjustment using dispersion factor ;

**proc** **genmod** data = Method1 ;

class Block Mix ;

model Count = Block Mix / dist = poisson type3 pscale ;

lsmeans Mix / cl ilink ;

**run** ; \* F-test: P = 0.3749 ;

\*\*\* Example sheep milk ;

**data** Somatic ;

length Method $ **10.** ;

input Method $ Count ;

cards ;

Mechanical 2966

Mechanical 569

Mechanical 59

Mechanical 1887

Mechanical 3452

Mechanical 189

Mechanical 93

Mechanical 618

Mechanical 130

Mechanical 2493

Manual 186

Manual 107

Manual 65

Manual 126

Manual 123

Manual 164

Manual 408

Manual 324

Manual 548

Manual 139

;

**run** ;

**proc** **print** ;

**run** ;

\*\* Poisson model ;

**proc** **genmod** data = Somatic plot = reschi ;

class Method ;

model Count = Method / dist = poisson ;

**run** ;

\*\* Negative binomial model ;

**proc** **genmod** data = Somatic plot = reschi ;

class Method ;

model Count = Method / dist = negbin ;

lsmeans Method / pdiff ;

ods output ModelFit = Deviance ;

output out = Reschi reschi = reschi ;

**run** ;

**proc** **print** data = Reschi ;

**run** ;

**data** OverdispersionTest ;

set Deviance ;

Pvalue = **1** - probchi(Value, DF) ;

**run** ;

**proc** **print** data = OverdispersionTest ;

**run** ;

## References

Collett, D. (2003). "Modelling binary data," 2/Ed. Chapman & Hall/CRC, Boca Raton, Florida.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2004). "Applied Longitudinal Analysis," Wiley, Hoboken, New Jersey.

Hilbe, J. M. (2011). "Negative binomial regression," 2/Ed. Cambridge University Press, Cambridge.

Littell, R. C., Milliken, G. A., Stroup, W. W., and Wolfinger, R. D. (1996). "SAS System for mixed models," SAS Institute, Cary, NC.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., and Schabenberger, O. (2006). "SAS for mixed models," SAS Institute, Cary NC.

McCullagh, P., and Nelder, J. (1989). "Generalized linear models," 2/Ed. Chapman and Hall/CRC, Boca Raton.

Morel, J. G., and Neerchal, N. K. (2012). "Overdispersion Models in SAS," SAS Institute, Cary, NC.

Olsson, U. (2002). "Generalized linear models: an applied approach," Studentlitteratur, Lund.

Piepho, H. P. (1999). Analysing disease incidence data from designed experiments by generalized linear models. *Plant Pathology* **48**, 668-674.

Stroup, W. W. (2012). "Generalized linear mixed models: modern concepts, methods and applications," Chapman & Hall/CRC, Boca Raton, Florida.

Zeileis, A., Kleiber, C., and Jackman, S. (2008). Regression models for count data in R. *Journal of Statistical Software* **27**, 1-25.

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., and Smith, G. M. (2009). "Mixed effects models and extensions in ecology with R," Springer, New York.