ENV 710

Poisson regression

Duke | NICHOLAS SCHOOL OF THE ENVIRONMENT

roadmap

- recap
- install these packages: `faraway`, `AER`, `MASS`, `sjPLot`, `ggplot2`

where we are

interactions
centering/scaling explanatory variables
random effects and mixed models

↓

generalized linear models

- Poisson regression
- logistic regression
- binomial logistic regression

Poisson regression

$$ln(\lambda_i) = \beta_0 + \beta_1 X_1 ... \beta_k X_k$$

response linked to the linear combination by a log link function

one-unit increase in X_i is associated with a multiplicative change in the mean λ_i by a factor of $\exp(\beta_i)$

$$P(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!}$$

$$Y \in \{0, 1, 2, 3, 4, ...\}$$

things to consider

1. Y_i, not expected to be normally distributed
2. Y_i, not expected to be linearly related to continuous predictors
3. overdispersion, variance >> mean
4. compare models with AIC, deviance test, or likelihood ratio test
5. compare model fit with deviance test
6. offsets – if modeling rates

I – species numbers

What geographical attributes determine the number of plants species on an island?

load the data

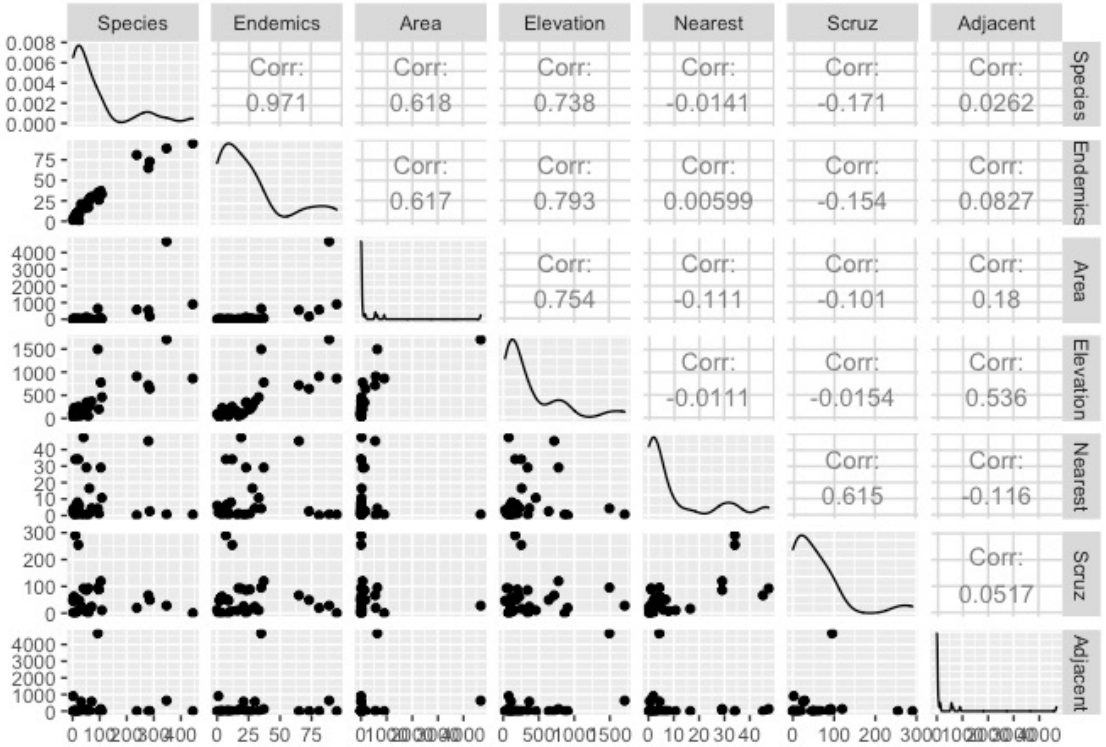
```
require(faraway)
data(gala)
```

I – species numbers

what type of model should be used?
which explanatory variable is likely to have a significant effect on species numbers?

What geographical attributes determine the number of plants species on an island?

- Endemics: the number of endemic species
- Area: the area of the island, km^2
- Elevation: the highest elevation of the island, m
- Nearest: the distance from the nearest island, km
- Adjacent: the area of the adjacent island, km^2



I – species numbers

What geographical attributes determine the number of plants species on an island?

load the data

```
install.packages("faraway")
data(gala)
```

run and reduce the model

```
glm(…, family = poisson, data = gala)
```

check model assumptions and correct

I – species numbers

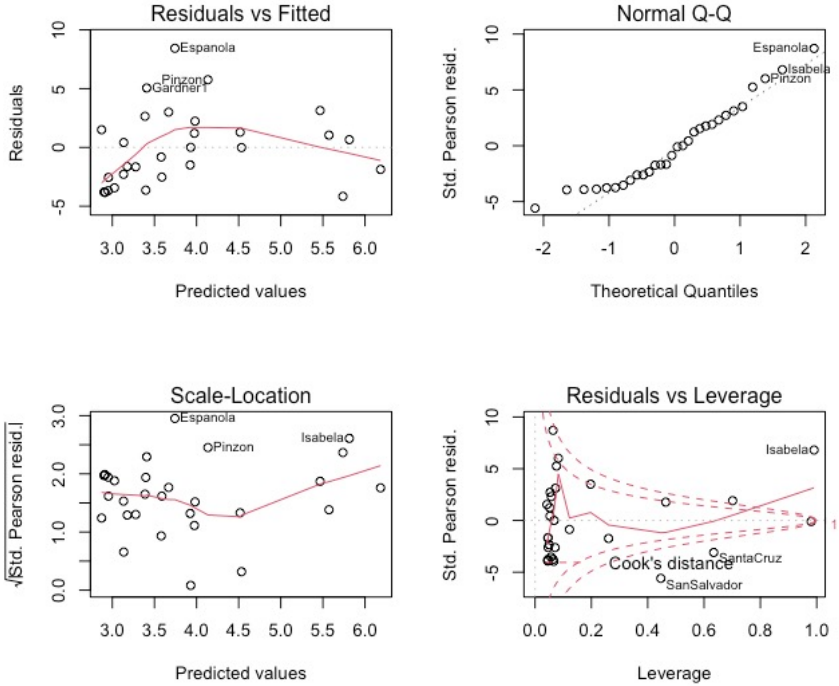
```
sp1 <- glm(Species ~ Endemics + Area + Elevation +
           Nearest + Adjacent, family = poisson,
           data = gala)
 summary(sp1)
```

```
Call:
glm(formula = Species ~ Endemics + Area + Elevation +
    Nearest + Adjacent, family = poisson, data = gala)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.796e+00  5.326e-02  52.489  < 2e-16 ***
Endemics     3.465e-02  1.620e-03  21.387  < 2e-16 ***
Area        -9.892e-05  3.680e-05  -2.688  0.00719 **
Elevation    2.098e-04  1.879e-04   1.117  0.26419
Nearest      9.492e-03  1.389e-03   6.835  8.2e-12 ***
Adjacent     5.018e-05  4.785e-05   1.049  0.29434
---
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  314.77  on 24  degrees of freedom
AIC: 487.6
```

- the full species model with a Poisson probability distribution



I – species numbers

```
sp2 <- update(sp1, .~.-Adjacent)
 summary(sp2)

Call:
glm(formula = Species ~ Endemics + Area + Elevation
    + Nearest, family = poisson, data = gala)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.794e+00  5.332e-02  52.399  < 2e-16 ***
Endemics     3.325e-02  9.164e-04  36.283  < 2e-16 ***
Area        -1.266e-04  2.559e-05  -4.947 7.53e-07 ***
Elevation    3.799e-04  9.432e-05   4.028 5.63e-05 ***
Nearest      9.049e-03  1.327e-03   6.819 9.18e-12 ***
---
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  315.88  on 25  degrees of freedom
AIC: 486.71
```

```
require(AER)
 dispersiontest(sp2)

data:  sp2

z = 3.5213, p-value = 0.0002147

alternative hypothesis: true
dispersion is greater than 1

sample estimates:

dispersion

   10.0066
```

- a reduced model, but what is the overdispersion test telling us?

I – species numbers

```
sp3 <- update(sp2, .~., family = quasipoisson)
 summary(sp3)
```

```
Call:
glm(formula = Species ~ Endemics + Area + Elevation
    + Nearest, family = quasipoisson, data = gala)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.7940464  0.1840214  15.183 3.99e-14 ***
Endemics     0.0332484  0.0031624  10.514 1.16e-10 ***
Area        -0.0001266  0.0000883  -1.433   0.1641
Elevation    0.0003799  0.0003255   1.167   0.2542
Nearest      0.0090490  0.0045798   1.976   0.0593 .
---
(Dispersion parameter for quasipoisson family taken to be
11.90987)

    Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  315.88  on 25  degrees of freedom
AIC: NA
```

- use the quasipoisson to adjust for significant overdispersion
- how does the quasipoisson alter hypothesis testing of the coefficients?

I – species numbers

```
sp4 <- update(sp3, .~.-Elevation)
 summary(sp4)

Call:
glm(formula = Species ~ Endemics + Area +
    Nearest, family = quasipoisson, data = gala)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.868e+00  1.672e-01  17.152 1.08e-15 ***
Endemics     3.551e-02  2.509e-03  14.153 9.96e-14 ***
Area        -4.542e-05  5.370e-05  -0.846   0.4054
Nearest      9.289e-03  4.516e-03   2.057   0.0499 *
---

(Dispersion parameter for quasipoisson family taken to
 be 11.72483)

    Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  330.84  on 26  degrees of freedom
AIC: NA
```

```
sp5 <- update(sp4, .~.-Area)
 summary(sp5)

Call:
glm(formula = Species ~ Endemics + Nearest,
    family = quasipoisson, data = gala)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.881615   0.165742  17.386 3.43e-16 ***
Endemics    0.034452   0.002189  15.738 4.00e-15 ***
Nearest     0.009811   0.004466   2.197   0.0368 *
---

(Dispersion parameter for quasipoisson family taken to
be 11.64264)

    Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  339.47  on 27  degrees of freedom
AIC: NA
```

- reduce the quasipoisson model to find the minimum adequate model

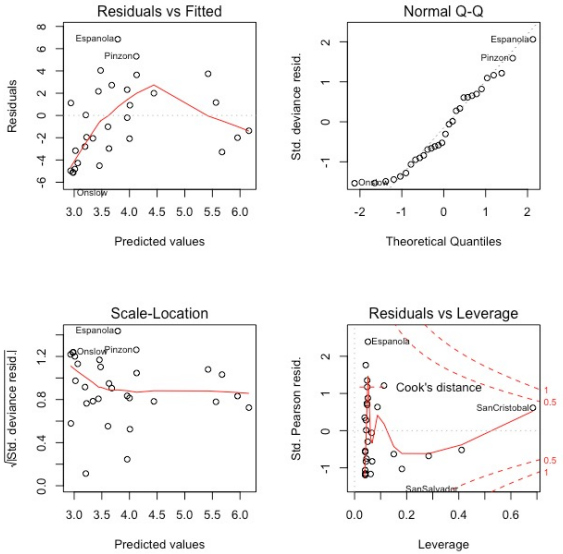
I – species numbers

```
sp5 <- update(sp4, .~.-Area)
 summary(sp5)

Call:
glm(formula = Species ~ Endemics + Nearest,
    family = quasipoisson, data = gala)

    Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  339.47  on 27  degrees of freedom
AIC: NA
```

- fit of sp5 (reduced model with quasipoisson) to the data?
- high mean of species numbers suggests residuals could be distributed normally…




```
mean(gala$Species)
[1] 85.23333

outlierTest(sp5)
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
          unadjusted p Bonferroni p
Espanola 2.124834     0.0336           NA
```

I – species numbers

```
require(MASS)
 sp6 <- glm.nb(Species ~ Endemics + Nearest,
               data = gala)
  summary(sp6)

Call:
glm.nb(formula = Species ~ Endemics + Nearest, data =
gala, init.theta = 2.754871184,
    link = log)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 2.576329   0.183165  14.066   <2e-16 ***
Endemics    0.043453   0.004191  10.367   <2e-16 ***
Nearest     0.005963   0.008098   0.736    0.462
---

(Dispersion parameter for Negative Binomial(2.7549)
family taken to be 1)

    Null deviance: 140.601  on 29  degrees of freedom
Residual deviance:  33.267  on 27  degrees of freedom
AIC: 284.13

              Theta:  2.755
          Std. Err.:  0.811

 2 x log-likelihood:  -276.132
```

- an alternative to the quasipoisson is to fit a model with a negative binomial distribution which includes an extra parameter, θ, the overdispersion parameter
- generalizes Poisson regression, loosening the assumption that the variance is equal to the mean
- negative binomial is nested within the Poisson, so can use AIC or likelihood ratio test to compare models.

```
sp5p <- update(sp5, .~., family = poisson)

AIC(sp5p, sp6)

df      AIC
sp5p  3 506.2993
sp6   4 284.1318

lrtest(sp5p, sp6)

Likelihood ratio test

Model 1: Species ~ Endemics + Nearest
Model 2: Species ~ Endemics + Nearest
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   3 -250.15
2   4 -138.07  1 224.17  < 2.2e-16 ***
```

I – species numbers

```
sp7 <- update(sp6, .~.-Nearest)
 summary(sp7)
```

```
glm.nb(formula = Species ~ Endemics, data = gala,
init.theta = 2.695721183, link = log)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 2.631244   0.164242   16.02   <2e-16 ***
Endemics    0.043785   0.004233   10.34   <2e-16 ***
---
(Dispersion parameter for Negative Binomial(2.6957) family
taken to be 1)

    Null deviance: 137.826  on 29  degrees of freedom
Residual deviance:  33.204  on 28  degrees of freedom
AIC: 282.66

     Theta:  2.696      Std. Err.:  0.789
 2 x log-likelihood:  -276.662
```

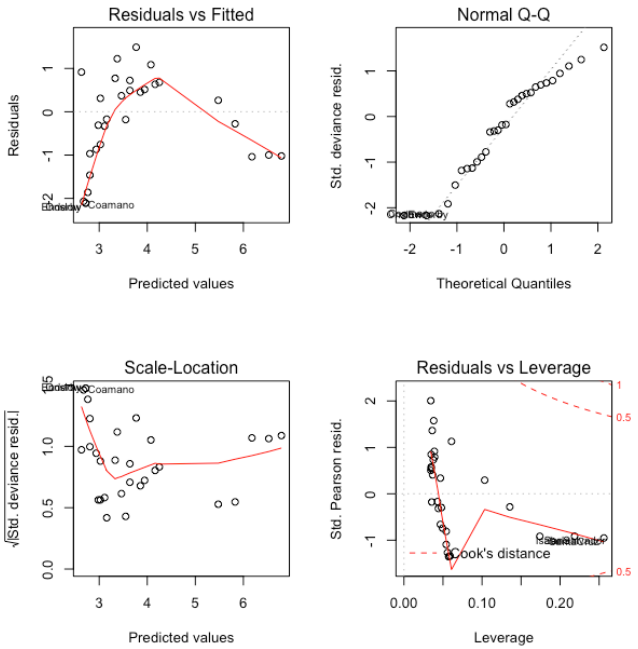
```
AIC(sp6, sp7)

    df      AIC
sp6  4 284.1318
sp7  3 282.6618

pchisq(sp7$deviance, df=sp7$df.residual,
lower.tail=FALSE)

[1] 0.2283267
```

- deviance goodness of fit test: deviance measures how well our model predictions match the observed outcomes and follows a chi-squared distribution, with degrees of freedom equal to the difference in the number of parameters



I – species numbers

```
sp7 <- update(sp6, .~.-Nearest)
 summary(sp7)
```

```
glm.nb(formula = Species ~ Endemics, data = gala,
init.theta = 2.695721183, link = log)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 2.631244   0.164242   16.02   <2e-16 ***
Endemics    0.043785   0.004233   10.34   <2e-16 ***
---
(Dispersion parameter for Negative Binomial(2.6957) family
taken to be 1)

    Null deviance: 137.826  on 29  degrees of freedom
Residual deviance:  33.204  on 28  degrees of freedom
AIC: 282.66

     Theta:  2.696      Std. Err.:  0.789
 2 x log-likelihood:  -276.662
```

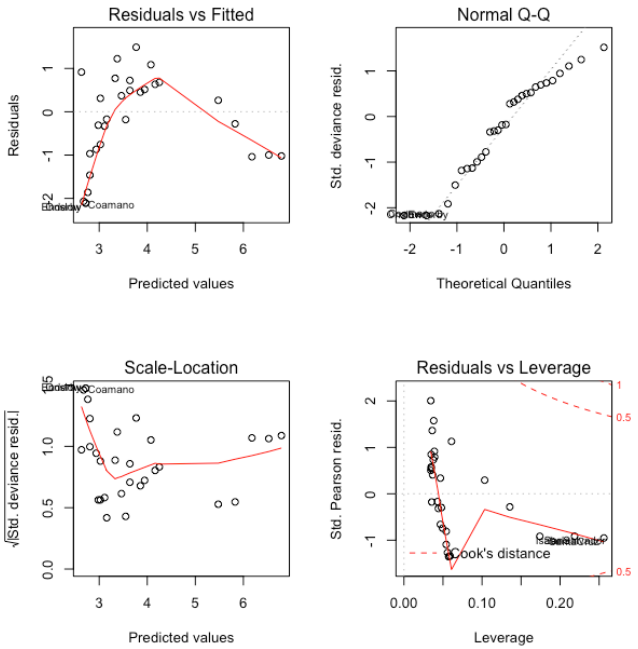
```
AIC(sp6, sp7)

    df      AIC
sp6  4 284.1318
sp7  3 282.6618

pchisq(sp7$deviance, df=sp7$df.residual,
lower.tail=FALSE)

[1] 0.2283267
```

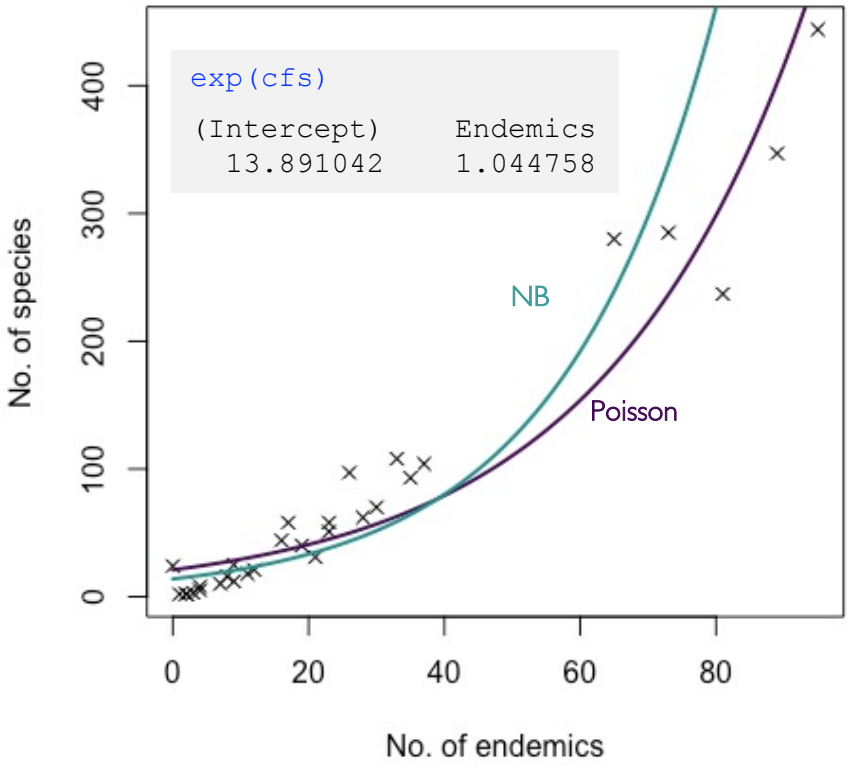
- reduce sp6 by taking out *Nearest*
- the negative binomial is a distribution and therefore has an AIC, which we can use to compare models



I – species numbers

```
my_clr <- viridis(n=3)
cfs7 <- coef(sp7)
cfs1 <- coef(sp11)

with(gala, plot(x = Endemics, y = Species, pch = 4,
                xlab = "No. of endemics",
                ylab = "No. of species"))
x <- with(gala, seq(min(Endemics), max(Endemics),
          length = 100))
curve(exp(cfs11[1] + cfs11[2]*x), add = T, col =
      my_clr[1], lwd = 2)
curve(exp(cfs7[1] + cfs7[2]*x), add = T, col =
      my_clr[2], lwd = 2)
```



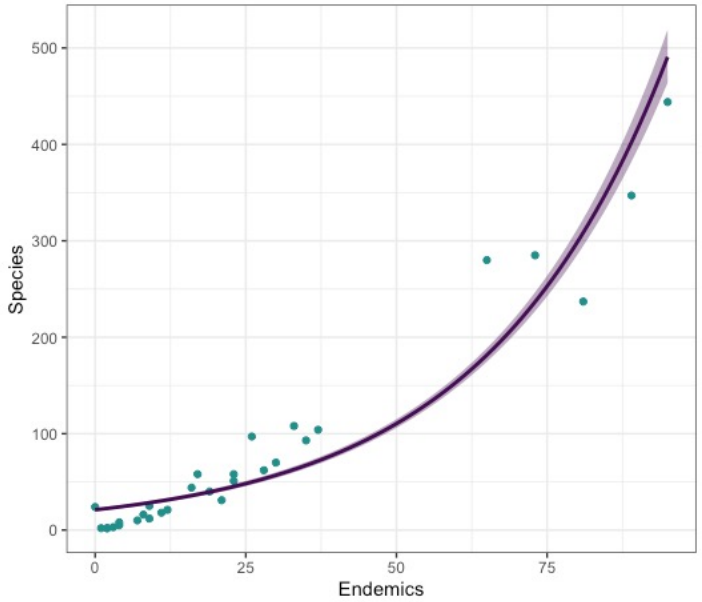
parameter interpretation

- mean number of species is 13.9
- with every addition endemic species, the number of island species increases by 1.04 times or ~4%

I – species numbers

```
library(viridis)
my_clr <- viridis(n=3)

ggplot(gala, aes(x=Endemics, y = Species)) +
       geom_point(color = my_clr[2]) +
       geom_smooth(method = "glm", se = TRUE,
       method.args = list(family = "poisson"),
                          fill = my_clr[1],
                          color = my_clr[1]) +
   theme_bw()
```



parameter interpretation

- mean number of species is 13.9
- with every addition endemic species, the number of island species increases by 1.04 times or ~4%

Questions?