ENV 710: Lecture 16

generalized linear models

Duke | NICHOLAS SCHOOL OF THE ENVIRONMENT

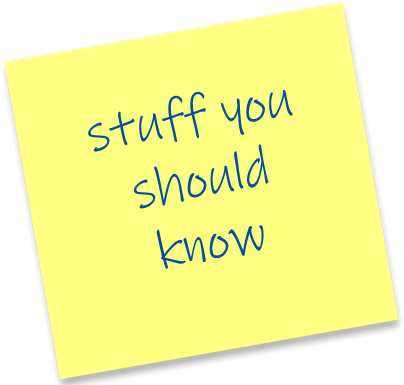
generalized linear models

glm's

learning goals

- generalized linear models

- Poisson regression
  - what is it? when to use it?
  - how does it work?
  - interpretation of coefficients
  - modeling rates
  - overdispersion & quasipoisson



generalized linear models (glm's)

- linear → response is still modeled as a linear combination of predictors (independent variables)

- generalized → link function, g(μ), links the response to the linear predictor ("transforms") *Y*

  - normal: $g(\mu) = \mu$ identity link
  - Poisson: $g(\mu) = log(\mu)$ log-link
  - binomial/logistic: $g(\mu) = log[\mu/(1-\mu)]$ log-odds

- GLMs fit by *iteratively reweighted least squares*

generalized linear model

$$g(\mu) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$$

link: links linear predictor to response through the "link" function

random component: response variable

systematic component: linear predictor

- as in lm's, the relationship between the transformed response and the predictor is assumed to be linear

generalized linear models

Poisson regression

what if we have count data?

- suppose we have aggregate counts of some event
  - number of elephant matings vs. age
  - number of salamanders vs. forest cover

- nature of count data
  - discrete, skewed distribution
  - high proportion of zero outcomes
  - always ≥ 0

- OLS and general linear models don't work
  - relationship between *X* and *Y* is nonlinear
  - counts are heteroscedastic
  - outcomes must be non-negative values

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 $\exp(\beta_i)$

advantages:

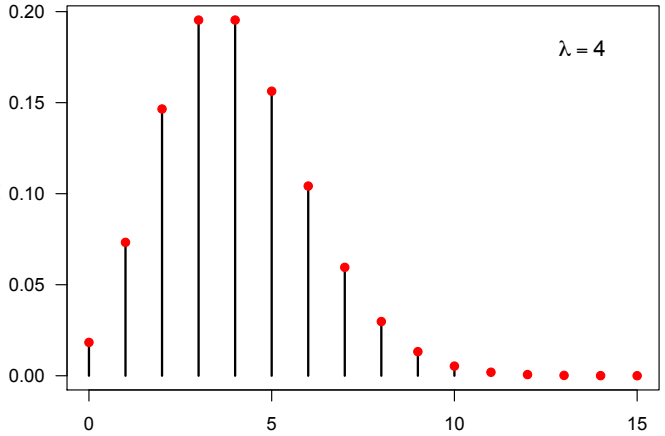
1. do not have to transform the response variable
2. if link produces additive effects, do not need constant variance

Poisson distribution

Y is distributed as a Poisson random variable; λ *is* the mean and variance

$$Y \sim Poisson(\lambda)$$

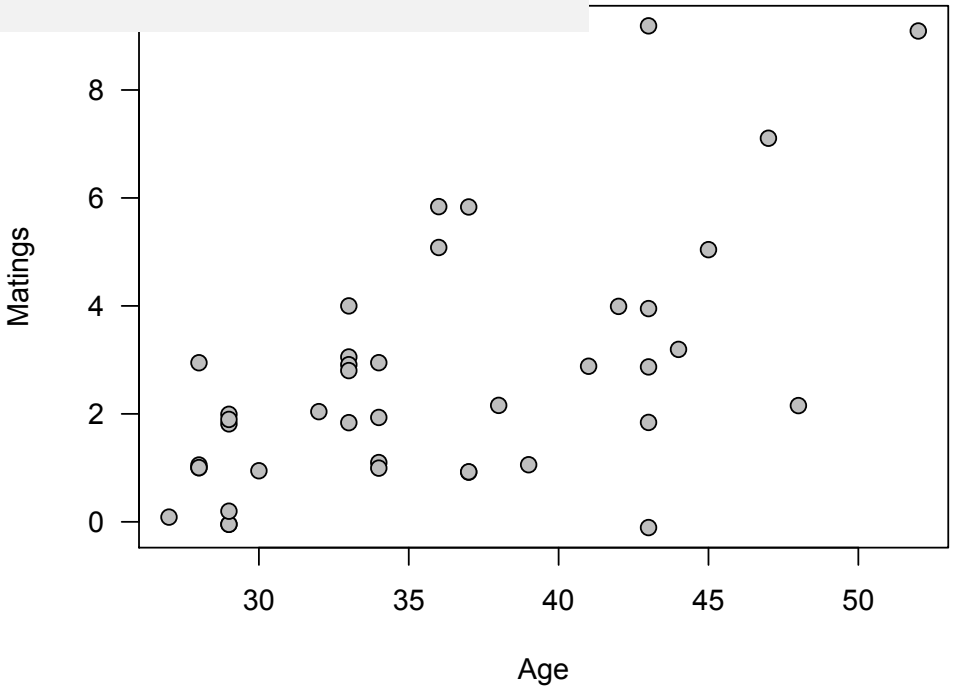
$$P(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!} \qquad Y \in \{0, 1, 2, 3, 4, ...\}$$



example

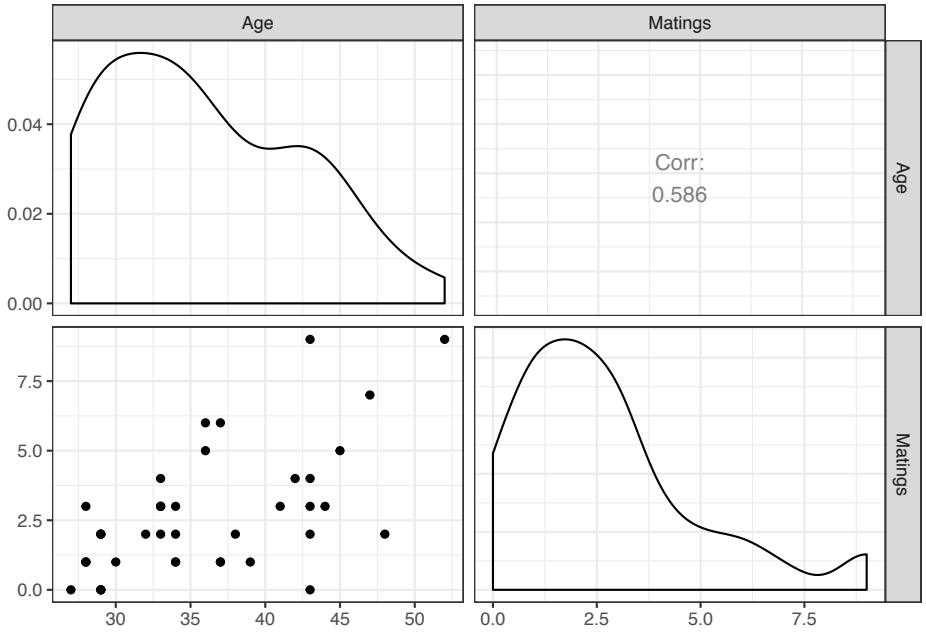
Elephant mating

Young and old male elephants compete for female mates. Because male elephants continue to grow through their lives, older elephants are larger and more successful at mating. What is the relationship between number of matings and elephant age?



example

Elephant mating



1. explore the data
2. fit the model
3. reduce model (using backwards selection)
4. test the goodness of fit of the model
5. test for overdispersion:
6. interpret model coefficients

2 fit the model

```
ele1 <- glm(Matings ~ Age, family = poisson)
 summary(ele1)
```

```
Call:glm(formula = Matings ~ Age, family = poisson)

Coefficients: Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.58201    0.54462  -2.905  0.00368 **
Age          0.06869    0.01375   4.997 5.81e-07 ***

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 75.372  on 40  degrees of freedom
Residual deviance: 51.012  on 39  degrees of freedom
AIC: 156.46         Number of Fisher Scoring iterations: 5
```

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

$$\lambda_i = e^{\beta_0 + \beta_1 X_1 ... \beta_k X_k}$$

parameter interpretation

$$ln(\lambda_i) = \beta_0 + \beta_1 X_{1i}$$

parameters are given on the scale of the link function

interpretation of β_0:

e^{β_0} is the mean of the Poisson distribution when X_1=0

interpretation of β_1:

increasing X_1 by 1-unit has a multiplicative effect on the mean of the Poisson by e^{β_1}

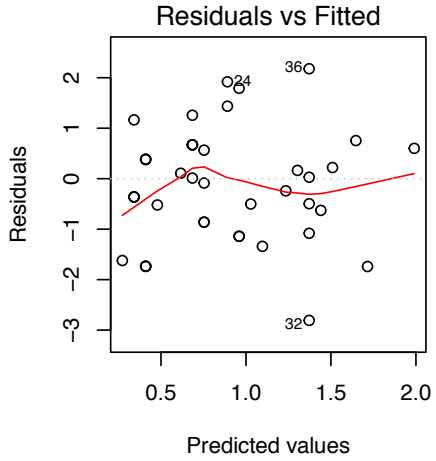
elephant matings

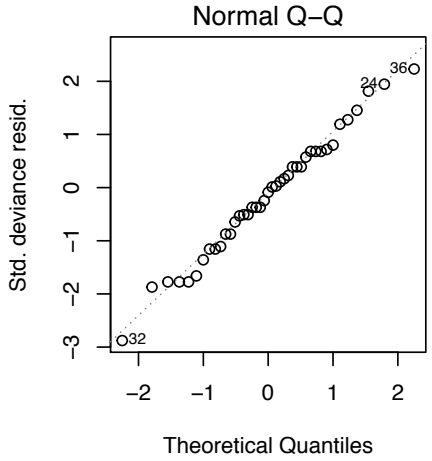
- $e^{-1.58} = 0.21$

mean number of matings for an elephant of age 0

- $e^{0.07} = 1.07$

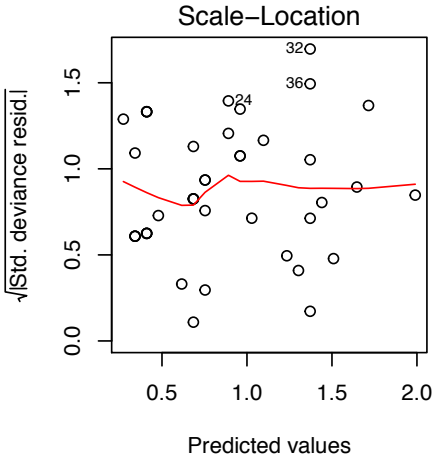
additional 1 year in age increases the expected number of elephant mates by a 1.07 times, or roughly 7%

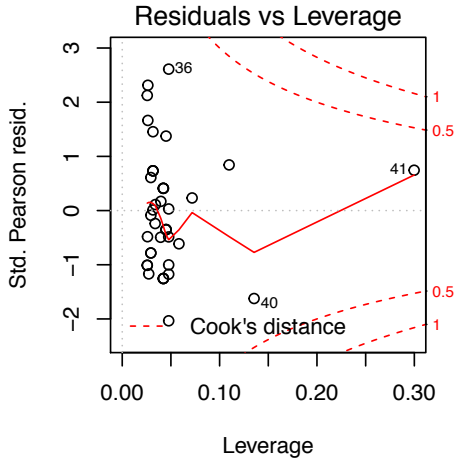





```
plot(ele1)
```

problem if >5% of |residuals| > 2





check for data points with high influence

example

Elephant mating

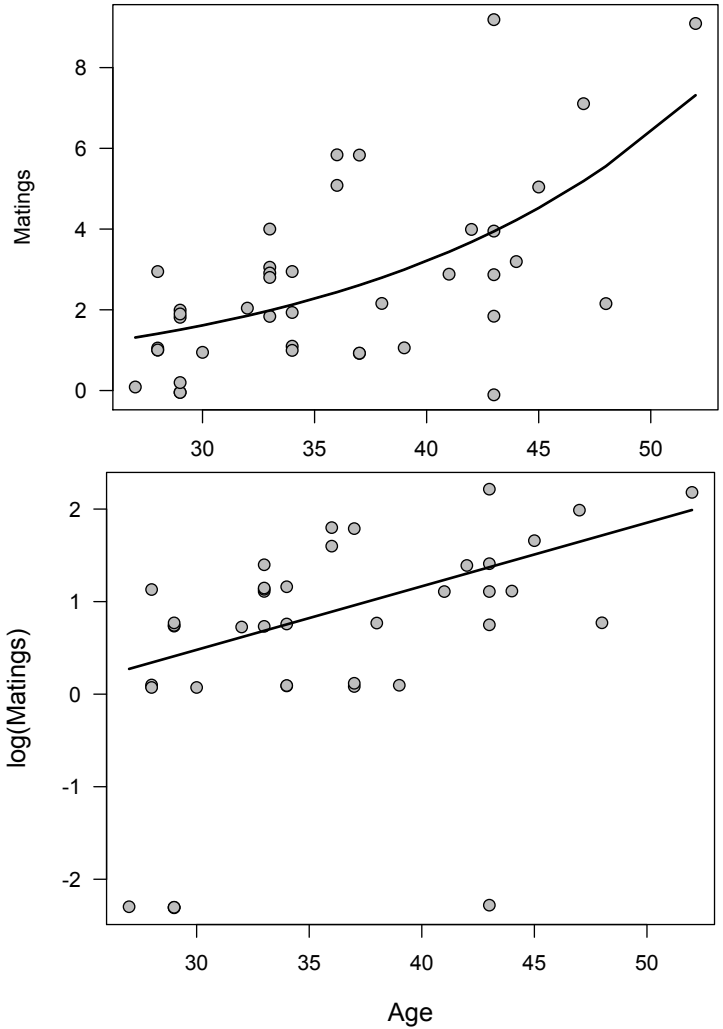
model elephant mating

```
eledat <- case2201
  ele1 <- glm(Matings ~ Age, family = poisson)

  coeff <- coef(ele1)
   xvals <- sort(Age)
    plot(Age, jitter(Matings), las=1, pch=21,
         bg="grey", cex=1.2, ylab="Matings")
    log.means <- coeff[1]+coeff[2]*xvals
  mean.values <- exp(log.means)
lines(xvals, mean.values, lwd=2)
```

mates for 30-year-old elephant?

```
lambda <- exp(coeff[1] + coeff[2]*30)
```

```
(Intercept)
   1.614098
```



goodness of fit – does model fit the data?

- deviance is the likelihood-ratio statistic, $\mathcal{L}$, for comparing model, *M*, to the saturated model, *S*, where the saturated model uses a distinct parameter for each observation
- deviance is a measure of how well the model fits the data (and has an approximately chi-square distribution)
- null deviance is the deviance of the null model with just an intercept
- residual deviance is the deviance of the fitted model
- when means (counts) are small (<5), deviance test does not work well

$$deviance = -2[\mathcal{L}_M - \mathcal{L}_S]$$

- H_0: model, M, fits the data (or is correctly specified)
  - → non-significant p-value indicates that the model fits the data well
  - → significant p-value rejects the H_0 that the model fits the data

goodness of fit

- goodness of fit test for elephant model
- compare model fit to null model
- H_0: model fits the data
  - non-significant test means the model fits the data well

```
Null deviance: 75.372  on 40 degrees of freedom
Residual deviance: 51.012 on 39 degrees of freedom
AIC: 156.46 Number of Fisher Scoring iterations: 5
```

```
pchisq(q=ele1$deviance, df=ele1$df.residual,
lower.tail=FALSE)

[1] 0.09426231

pchisq(q=51.02, df=39, lower.tail=FALSE)
[1] 0.09426231
```

fail to reject the H_0 that our model fits well

overdispersion

- overdispersion, Φ, is the presence of greater variability in a dataset than expected based on a given statistical model

- Poisson model: $Var(Y_i) >> E(Y_i)$

$$z_i = \frac{y_i - \hat{y}_i}{\sigma_{\hat{y}_i}} \qquad \phi = \sum z_i^2 / (n-k)$$

n = number of observations,
k = number of parameters

```
n <- length(eledat$Matings)
 k <- 2
  yhat <- predict(ele1, type = "response")
  z <- (eledat$Matings - yhat)/sqrt(yhat)
 phi <- sum(z^2)/(n-k)
phi

[1] 1.157334
```

```
require(AER)
dispersiontest(ele1)

        Overdispersion test

data:  ele1
z = 0.49631, p-value = 0.3098
alternative hypothesis: true dispersion is
        greater than 1
sample estimates:
dispersion
  1.107951
```

quasipoisson

```
ele1 <- glm(Matings ~ Age, family = poisson)
```

```
Coefficients: Estimate   Std. Error z value  Pr(>|z|)
(Intercept)   -1.58201   0.54462    -2.905   0.00368 **
Age           0.06869    0.01375    4.997    5.81e-07 ***

(Dispersion parameter for poisson family taken to be 1)
Null deviance: 75.372  on 40  degrees of freedom
Residual deviance: 51.012  on 39  degrees of freedom
AIC: 156.46
```

standard errors are scaled by the square root of the overdispersion ratio

```
ele2 <- glm(Matings ~ Age, family = quasipoisson)

Coefficients:  Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.58201   0.58590  -2.700   0.0102 *
Age             0.06869   0.01479   4.645 3.81e-05 ***
(Dispersion parameter for quasipoisson family taken to be 1.157334)
Null deviance: 75.372  on 40  degrees of freedom
Residual deviance: 51.012  on 39  degrees of freedom
AIC: NA
```

comparison of nested models

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

$$-2log_e(\mathcal{L}_r/\mathcal{L}_f) \sim \chi^2$$

- a comparison of *null deviance* and *residual deviance* is used to test the significance of parameters

- Likelihood Ratio Test is used for this nested test, following a X^2 distribution under H_0 being true

```
require(lmtest)
lrtest(ele0, ele1)

Likelihood ratio testModel 1: Matings ~ 1
Model 2: Matings ~ Age

#Df  LogLik Df Chisq Pr(>Chisq)
1  1 -88.409
2  2   2 -76.229  1 24.36  7.991e-07 ***
```

```
ele0 <- glm(Matings ~ 1, family = poisson)
ele1 <- glm(Matings ~ Age, family = poisson)

anova(ele0, ele1, test= "Chisq")
[1] 7.99062e-07

pchisq(ele1$null.deviance-ele1$deviance, df = 1, lower.tail = F)
[1] 7.99062e-07
```

example

Fiji fertility study

What factors influence the number of children born by women in Fiji?

- number of years since marriage (1=0-4, 2=5-9, 3=10-14, 4=15-19, 5=20-24, 6=25-29)
- residence (1=suva, 2=urban, 3=rural)
- education (1=none, 2=lower primary, 3=upper primary, 4=secondary +)

	dur	res	educ	mean	var	n	chd
1	1	1	1	0.50	1.14	8	4
2	1	1	2	1.14	0.73	21	24
3	1	1	3	0.90	0.67	42	38
4	1	1	4	0.73	0.48	51	37
5	1	2	1	1.17	1.06	12	14
6	1	2	2	0.85	1.59	27	23

offsets – modeling rates

- model counts as a fraction of the total population or as a ratio of total effort
- model the rate at which events occur
- the adjustment term, *-log(t)* is called an offset

$$log(\mu/t) = \beta_0 + \beta_1 X_i$$

$$log(\mu) - log(t) = \beta_0 + \beta_1 X_i$$

$$\mu = t \cdot e^{(\beta_0 + \beta_1 X_i)}$$

```
glm(mu ~ factor(X), family = poisson, offset = log(t))
```

example

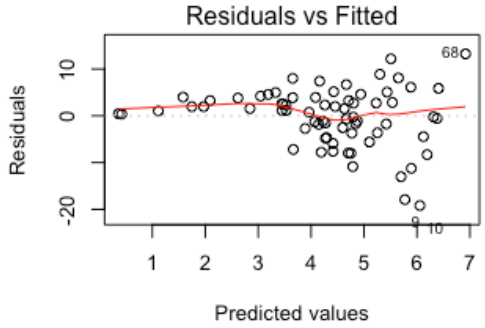
Fiji fertility study

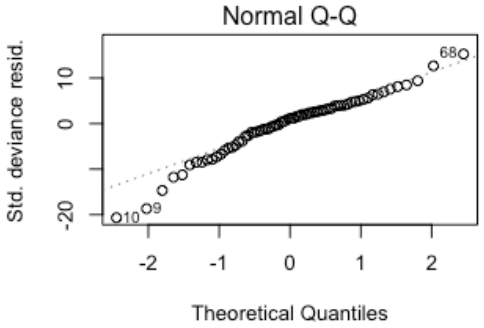
```
my_glm <- glm(chd~factor(res)+factor(educ), offset=log(n), family=poisson, data=dat)
summary(my_glm)

Call:
glm(formula = chd ~ factor(res) + factor(educ),
    family = poisson, data = dat, offset = log(n))

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.58469    0.02832  55.951  < 2e-16 ***
factor(res)2   0.12216    0.03247   3.762 0.000168 ***
factor(res)3   0.06038    0.02820   2.141 0.032252 *
factor(educ)2 -0.21482    0.02183  -9.840  < 2e-16 ***
factor(educ)3 -0.61918    0.02916 -21.231  < 2e-16 ***
factor(educ)4 -1.22427    0.05217 -23.467  < 2e-16 ***
---

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3731.9  on 69  degrees of freedom
Residual deviance: 2646.5  on 64  degrees of freedom
AIC: 3088
```

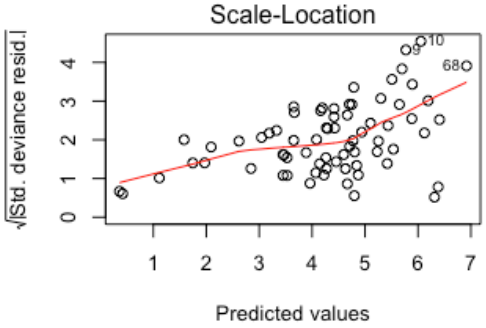
```
pchisq(my_glm$null.deviance-
my_glm$deviance, df = 5,
lower.tail = F)

[1] 1.976057e-232

pchisq(q=2646.5,
df=65,lower.tail=FALSE)

[1] 0

dispersiontest(my_glm)

Overdispersion test

data:  my_glm
z = 5.2745, p-value = 6.654e-08
alternative hypothesis: true
dispersion is greater than 1
sample estimates:
dispersion
  34.17192
```

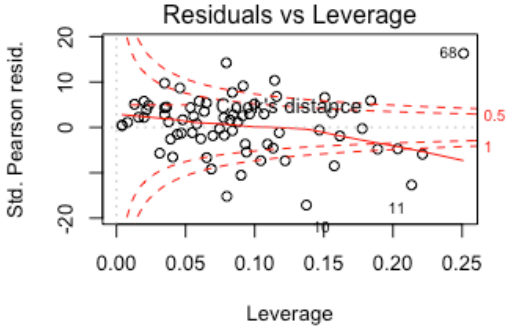
example

Fiji fertility study










```
pchisq(my_glm$null.deviance-
my_glm$deviance, df = 5,
lower.tail = F)
```

```
[1] 1.976057e-232
```

```
pchisq(q=2646.5,
df=65,lower.tail=FALSE)
```

```
[1] 0
```

```
dispersiontest(my_glm)
```

```
Overdispersion test

data:  my_glm
z = 5.2745, p-value = 6.654e-08
alternative hypothesis: true
dispersion is greater than 1
sample estimates:
dispersion
  34.17192
```

Questions?