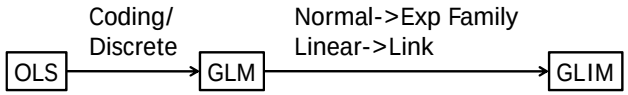
Generalized Linear Model (GLIM)

Road map



1. Linear
2. Error in Y
3. Errors normal
4. Errors indepenent
5. Errors homoscedastic

GLIM

- Developed in 1970's
- A generalization of GLM
- Uses likelihood

To run a GLIM

- Must specify f
  - Called the link function
- Must specify distribution of ε
  - Must be part of a group of distributions known as the exponential family
- A finite list
  - Likelihood can handle any f, ε but GLIM doesn't
- Common choices for link, ε
  - Logit/binomial→logistic
  - Probit/binomial→probit regression
  - 1/x / gamma → gamma regression, michaelis-menton & others
  - Log /gamma → exponential growth
  - Log / Poisson → Poisson regression (contingency tables)

Solving the GLIM

- No formulas
- Have to run an iterative procedure
  - IWLS (iterative weighted least squares)
- Easy to do on a computer
  - Why logistic only became popular in 1970's
- All I'm going to say about innards

5

GLIM in R

- m=glm(dep~indep,family=binomial|poisson|...,data=?)
- Returns coefficients, se, t
- Returns three more things:
  - Null deviance (deviance w/ intercept only or H_0)
  - Model deviance
  - AIC
- Deviance is like Sum Squares
  - Modeldev=-2 log L
  - Nulldev-Modeldev~χ^2_p is actually a Likelihood-Ratio Test
  - R^2=1-Modeldev/Nulldev
- lm(...)=glm(...,family=gaussian)

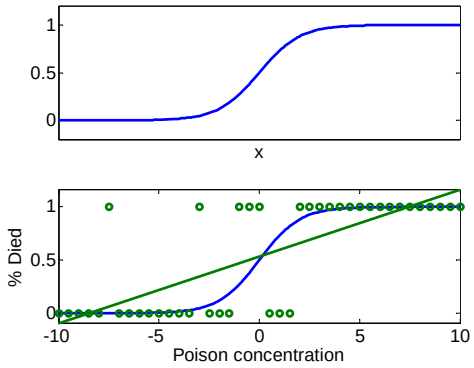
6

Logistic regression

- For binary (true/false or yes/no data)
- Error (e) is binomial
- Link is logistic
  - $\pi=1/(1+\exp(-(b_0+b_1x_1+b_2x_2+...)))$
  - Inverse: $\pi/(1-\pi)= b_0+b_1x_1+b_2x_2+...$
- Final equation is
  - $Y\sim Bernoulli(1/(1+\exp(-(b_0+b_1x_1+b_2x_2+...))))$

7

Logistic regression



8

Running in R, analysis

- m=glm(y~x1+z,family=binomial)
  - Y can be:
    - Boolean variable (0/1, yes/no, true/false, good/bad)
    - Two column matrix [nsuccesses nfailures]
    - Ratio (%success), use weight=n option
    - Also family=binomial(link=probit) or family=binomial(link=loglog)
- summary(m)
  - Looks a lot like lm except has AIC, deviances
- plot(m) gives standard diagnostics
- Don't need anova(m1)
- anova(m1,m2,test="Chisq")
- confidence.ellipse(m) # for GLM, GLIM in library(car)

9

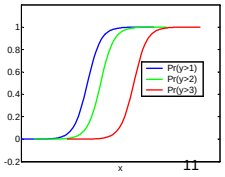
Logistic curves

```
source("budworm.r")
d=budworm()
d #warning ldose is a log2 scale
mfull=glm(cbind(surv, 1-surv)*20 ~ldose*sex,family=binomial,data=d)
mfull=glm(surv~ldose*sex,family=binomial,data=d,weights=rep(20,12))
   #alternative y=%, weights=N
mred=glm(cbind(surv, 1-surv)*20 ~ldose+sex,family=binomial,data=d)
summary(mfull)
summary(mred)
anova(mfull,mred,test="Chisq")
x=seq(0,5,0.1) #log2 dose
#predict as fit on log2-scale
ld.fem=predict(mred,data.frame(ldose=x,sex="F"),type="response")
ld.mal=predict(mred,data.frame(ldose=x,sex="M") ,type="response")
#plot on log2 scale but label
plot(x,ld.fem,type="l",xlab="log2(dose)")
lines(x,ld.mal,lty=2)
text(d$ldose,d$surv,labels=as.character(d$sex)) #points() gives o
```

10

Multinomial logistic

- What if dependent is trinary or more generally discrete with n factors (not binary n=2)
  - A,B,C
- Three options:
  - Baseline: A vs not-A, B vs not-B
    - avoid C: prob(C)=1-prob(A)-prob(B)
    - Coefficients (b) are probability not in baseline
    - Can subtract coefficients for A-B gives probability in A vs B
  - Nested
    - A vs B+C, B vs C
    - Makes most sense when natural nesting
  - Proportional odds
    - A vs BCD, AB vs CD, ABC vs D
    - Makes most sense when ordinal
    - R has function "polr" in library MASS



11

Poisson Regression

- Poisson distribution for counts
- In principle use any time dependent variable is counts (0,1,2,3,4,...)
- In practice, easier to treat as continuous if counts are high
- ct~Poisson($b_0+b_1x_1+b_2x_2+...$)
- m=glm(ct~cont1+disc1+...,family=poisson,data=?)

12

```
d<-read.table("spec_rich.txt" ,h=T)
names(d)
library(lattice)
xyplot(Species~Biomass|pH,data=d)
mint<-glm(Species~Biomass*pH,data=d,poisson)
summary(mint)
Call:
glm(formula = Species ~ Biomass * pH, family = poisson, data = d)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.49779  -0.74845  -0.04023   0.55745   3.22975

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    3.76812    0.06153  61.240  < 2e-16 ***
Biomass       -0.10713    0.01249  -8.577  < 2e-16 ***
pHlow         -0.81557    0.10284  -7.931 2.18e-15 ***
pHmid         -0.33146    0.09217  -3.596 0.000323 ***
Biomass:pHlow -0.15503    0.04003  -3.873 0.000108 ***
Biomass:pHmid -0.03189    0.02308  -1.382 0.166954
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 452.346  on 89  degrees of freedom
Residual deviance:  83.201  on 84  degrees of freedom
AIC: 514.39

Number of Fisher Scoring iterations: 4
```

13

continued

```
msimp<-glm(Species~Biomass+pH,data=d,poisson)
summary(msimp)
anova(mint,msimp,test="Chi")
Analysis of Deviance Table

Model 1: Species ~ Biomass * pH
Model 2: Species ~ Biomass + pH
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        84     83.201
2        86     99.242 -2  -16.040 0.0003288 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

14

Contingency tables

- Counts vs. discrete variables
- One of discrete variables is usually the “dependent” variable
  - Alternatively measure of association (correlation)
- Traditional test is χ^2 but G test is usually preferred
- Poisson regression same as G-test on contingency
  - Also called log-linear due to log link function

	Dioecious	Monoecious
Berry	50	34
Other	27	112

15

Contingency log linear in R

```
d=expand.grid(ecy=c("Dioecious","Monecious"),fruit=c
  ("Berry","Other"),subclass=c("Dicot","Monocot"))
d
ct<-c(50,34,27,112,15,15,15,15)
d=cbind(d,ct)
d
xtabs(ct~ecy+fruit+subclass,data=d) #omit ct also
m=glm(ct~ecy*fruit*subclass,family=poisson,data=d)
summary(m)
mns=glm(ct~ecy*fruit,family=poisson,data=d)
summary(mns)
Anova(m)
```

16

Gamma

- Gamma for asymmetric errors that increase with the mean (heteroscedastic)
- y=exp(at)
  - mg=glm(yn~t,family=Gamma(link=log))
  - But log transformed y is almost as good
- y=ax/(b+x) (Michaelis-Menton)
  - 1/y=b/a*1/x+1/a

17

Summary

- GLIM is generalization of GLM
  - Allows error term to be from "exponential family" instead of just normal
  - Allows a nonlinear "link" function (again from a subset) – but relationship between x_i is still linear
- Logistic/binomial regression for binary dependent variables
- Poisson regression for count dependent variables
  - Especially contingency tables
- Gamma regression also useful

18

Deviance

- Measure of goodness of fit
- $-2(l_{mode}-l_{fulll})$
- =RSS for normal

19

Overdispersion

- Overdispersion = too much variance
  - Impossible for normal since μ & σ are independent
  - Possible for binomial (np(1-p))
  - Common for Poisson $\mu=\sigma=\lambda$
- Diagnosis
  - Rule of thumb – deviance should $\approx$ df
  - If deviance >> df then overdispersion

20

Solutions to overdispersion

- Use negative binomial (has more variance)
  - Can be use in place of Poisson (but not really binomial)
    - library(MASS) glm.nb instead of glm
  - Good approach – still a well-known probability distribution
- Use quasilikelihood
  - var()=ϕ*var

21

Quasibinomial

```
g<-read.table("germination.txt",h=T)
names(g)
y<-g$count/g$sample
y
mint<-
  glm(y~Orobanche*extract,data=g,family=binomial,weigh
  ts=g$sample) #weight=sample – not count
33.278/17
summary(mquas<-
  glm(y~Orobanche*extract,data=g,family=quasibinomial,
  weights=g$sample))
#note, no AIC, report on dispersion parameter, change s
  of significance, use F-statistic to compare
anova(mint,mquas,test="F")
```

22

Zero-inflation

- One causes of overdispersion is too many zeros.
  - Zeros show up for lots of reasons
    - True absence
    - Outside of range
    - Observation error
- Model zeros explicitly

23

ZI models – two approaches

- Two-part/hurdle
  - binomial zero vs. non-zero
  - Poisson or negbin truncated for 1+
- Mixture
  - binomial zero vs. non-zero (same)
  - Poisson or negbin NOT truncated (0-N)
    - Two sources of zero
- Think about causes of zeros
  - If the process that causes 1-N variation can also cause a zero then use mixture

24

In R

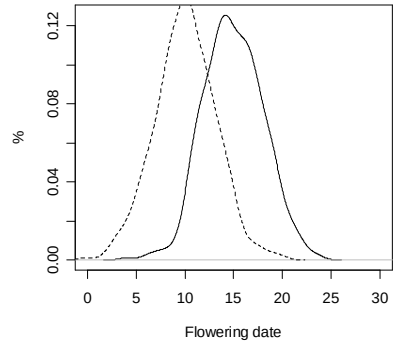
```
library(pscl)
m<-
   zeroinfl(formula(y~x),dist="negbin",link="logit",data
   =…)
m<-
   hurdle(formula(y~x),dist="negbin",link="logit",data=…
   )
```

25

Censored data and event times

26

Time to flower – two populations: Cal & Washington



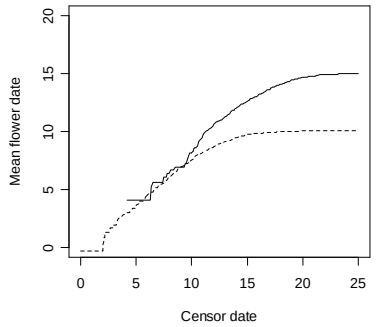
27

Censoring

- Don’t follow all possible points to the end
  - The event is never observed
  - The time to event recorded is just the length of study
  - Often times the long right tail is important

28

Means are a bad way to study this due to censoring



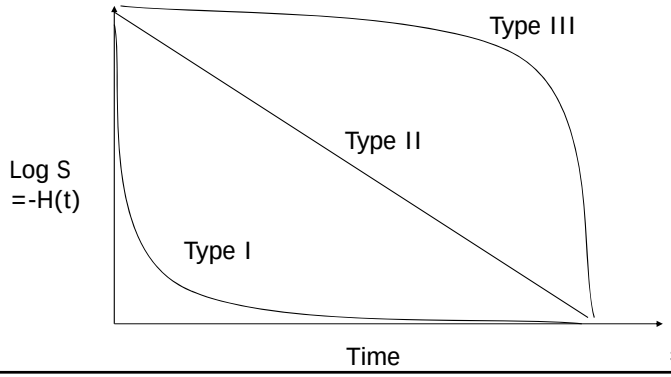
29

More generally

- When we study times to events
  - Already learn that this is modeled by exponential and Weibull
  - We often end up with censoring – "long right tail" omitted from study
- Extremely common in medical
  - Effect of cancer drug on increasing lifespan for a 5 year study – some patients live longer than 5 years
- Important in ecology too:
  - Spermatogenesis in C. elegans decreases life span
  - Do birds have sensecence?
  - Time to mortality of pesticide
  - Other time to event measures (flowering, germination, etc)
  - Could be used in space - dispersal

30

Ecological life history theory



31

Ways to study times to events

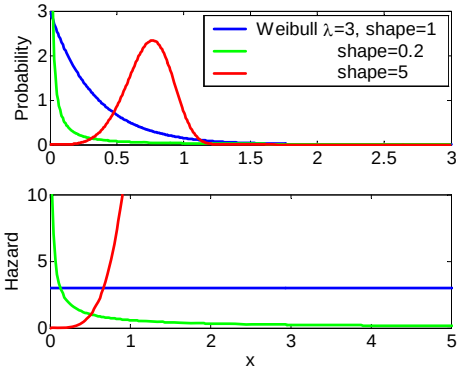
- T a random variable of time to event
- p - The pdf function (% dying in time t-t+dt)
- F - The CDF function (% dead at time t)
- S(t) - The survivorship function (1-F)
- h(t) - The hazard function
  - Fraction of those alive at time t that die
  - P(death at time t|alive at time t)
  - Per capita rate
  - $h(t)=p(t)/S(t)=F'(t)/S(t)=S'(t)/S(t)$
    - Cf with 1/N dN/dt – per capita growth rate
  - For exponential: have constant hazards $h(t)=\lambda$
- H(t) – Cumulative hazard function ($=\int^t h(t)dt$)

32

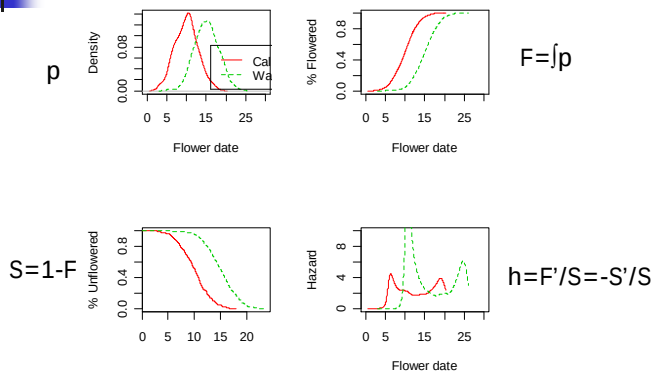
Studying survivor curves

- Usually plot S
  - S estimated using Kaplan-Meier method (non-parametric)
    - Confidence intervals
  - Can compare S (basically a probability distribution) using χ^2
- Usually study effects of independent variables on hazard function:
  - Parametric or proportional hazards h(t) is exponential or Weibull:
    - $h_{regress}(t)=h_{intercept}(t)\exp(\beta x+\varepsilon)$
  - Less parameteric:
    - Same equation but don't need formula for h(t)
    - Cox propotional hazards

Hazard functions



Plots for two populations



In R

```
fd=read.table("flowdate.csv",sep=",",head=T)
summary(fd)
library(survival)
```

- Key is "Surv(time[,censored])" wrapper
  - Censored=0/1 or T/F for "died in study" or 0=right censored
- Plotting

```
plot(survfit(Surv(fd)~pop,data=fd))
```

- Checking for significant differences

```
survdiff(Surv(fd)~pop,data=fd)
```

- Fitting a model

```
(m=survreg(Surv(fd)~pop,data=fd,dist="exponential"))
```

  - Also `coxph`