FISH 604
Module 6:
Generalized Linear Models

Instructor: Franz Mueter
Lena Point, Rm 315
796-5448
fmueter@alaska.edu

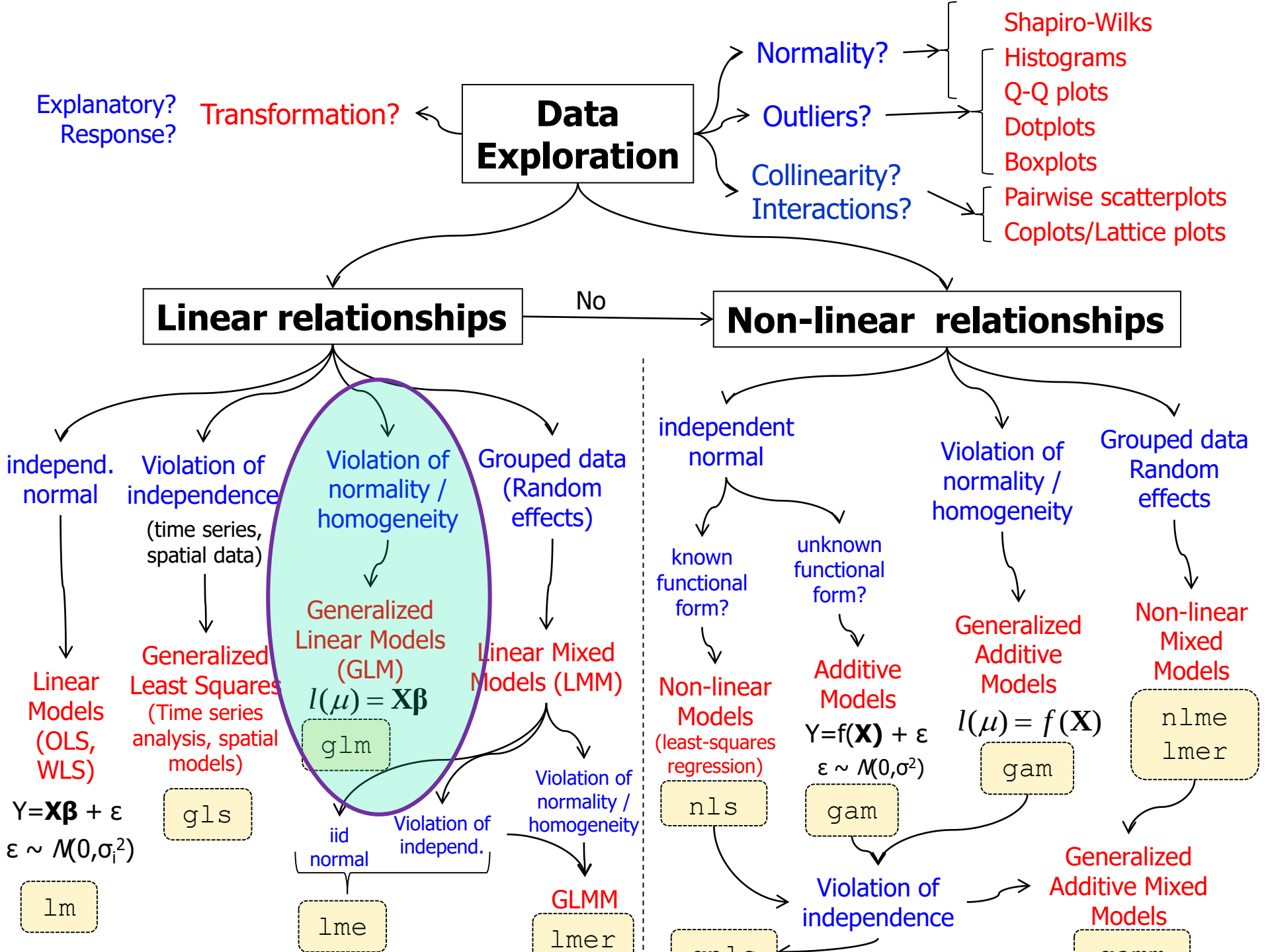
MSL FISH 604

Objectives and Outcomes

- Objectives
  - Introduce Generalized Linear Models to understand their structure and differences
- Outcomes
  - Understand at least 5 types GLMs, know when to use them, and be able to fit them to data
    - Logistic regression (binomial models)
    - Regression for count data
      - Poisson regression and Negative binomial regression
      - Models for zero-inflated count data
    - Multinomial models

Regression models

- Linear Models (LM)
  - Simple / multiple linear regression
  - Analysis of (co)variance (ANO(C)VA)
- Generalized Linear Models (GLM)
  - Binomial models (Logistic regression)
  - Poisson & negative binomial models
  - Multinomial & Zero-inflated models
- Generalized Additive Models (GAM)
  - Non-parametric smoothers
- Mixed-effects models (linear/non-linear)
- Non-linear models (NLM)


Data Exploration
Explanatory? Response?
Transformation?
Normality?
Outliers?
Collinearity? Interactions?
Shapiro-Wilks
Histograms
Q-Q plots
Dotplots
Boxplots
Pairwise scatterplots
Coplots/Lattice plots
Linear relationships
No
Non-linear relationships
independ. normal
Linear Models (OLS, WLS)
Y=Xβ + ε
ε ~ N(0,σi2)
lm
Violation of independence
(time series, spatial data)
Generalized Least Squares
(Time series analysis, spatial models)
gls
Violation of normality / homogeneity
Generalized Linear Models (GLM)
l(μ) = Xβ
glm
Grouped data (Random effects)
Linear Mixed Models (LMM)
iid normal
Violation of independ.
lme
Violation of normality / homogeneity
GLMM
lmer
independent normal
known functional form?
unknown functional form?
Non-linear Models
(least-squares regression)
nls
Additive Models
Y=f(X) + ε
ε ~ N(0,σ2)
gam
Violation of normality / homogeneity
Generalized Additive Models
l(μ) = f(X)
gam
Grouped data Random effects
Non-linear Mixed Models
nlme
lmer
Violation of independence
Generalized Additive Mixed Models

Generalized Linear Models

Why "Generalized"?

- Accomodates non-normal response distribution
  - Family: Binomial, Poisson, Gamma, and other distributions from the exponential family can be chosen
- Accomodates (smooth) transformations of linear predictor:
  - Link: The expected value of the response is a smooth function of a linear predictor ($\mathbf{X}\boldsymbol{\beta}$):

$$E(y) = \mu = \eta(\mathbf{X}\boldsymbol{\beta})$$

The inverse of η is called the link function:

$$l(\mu) = \mathbf{X}\boldsymbol{\beta}$$

Linear predictor

→ The model is linear on the transformed scale

The exponential family

- In linear modeling we apply the Gaussian (=normal) distribution to the responses (residuals)
- Ecological data often do NOT follow a normal distribution, e.g. binomial data, Poisson counts, abundance data (positive only, often right-skewed)
- These, and many other distributions can be written in a general formulation:

$$f(y;\theta,\phi) = e^{\frac{y\cdot\theta - b(\theta)}{a(\phi)} + c(y,\theta)}$$

Data

Distributional parameters

'Scale' parameter

MSL FISH 604

The exponential family

- In linear modeling we apply the Gaussian (=normal) distribution to the responses (residuals)
- Ecological data often do NOT follow a normal distribution, e.g. binomial data, Poisson counts, abundance data (positive only, often right-skewed)
- These, and many other distributions can be written in a general formulation:

$$f(y;\theta,\phi) = e^{\frac{y\cdot\theta - b(\theta)}{a(\phi)} + c(y,\theta)}$$

- For example, show that: $\theta = \log(\lambda); \quad \phi = 1; \quad a(\phi) = 1$

$$b(\theta) = \exp(\theta); \quad c(y,\theta) = \log(y!)$$

results in the Poisson distribution

The exponential family

- Advantage of using the exponential family
  - One set of equations can be used for all distributions in exponential family to estimate parameters via maximum likelihood
  - Mean and variance can be specified by single set of equations using the first and second derivatives of b:

$$E(Y) = b'(\theta)$$

$$Var(Y) = b''(\theta) \cdot a(\phi)$$

GLM vs. LM

"Ordinary" linear models are a special case of GLM with:

- family = Gaussian
- scale parameter = variance (σ^2)
- Link function = identity ($l(\mu) = \mu$)

$$E(y) = \mu = \mathbf{X\beta}$$

$$y = \mathbf{X\beta} + \varepsilon$$

$$y \sim N(\mu, \sigma^2) \Leftrightarrow \varepsilon \sim N(0, \sigma^2)$$

Fitting GLMs

1. Specify linear predictor (as in LM)
2. Specify 'family', based on nature of random variability, e.g.:
   - Gaussian (normal) for continuous data
   - Binomial / multinomial for categorical response
   - Poisson (or negative binomial) for positive counts
3. Specify link function
   - Examples: Identity, log, logit, probit
   - Only certain link function are appropriate for a given distribution

Fitting GLMs

4. Fit model using <u>iterative algorithm</u> (No closed-form solution as in LM)
   - Maximize likelihood (or, equivalently, minimize 'deviance')
     - Usually implemented through "Iterative weighted least-squares" or IWLS algorithm
   - More difficult to fit, may not always work!
   - Provide starting values for parameters if necessary!

Common GLMs

Family:	Gaussian	Binomial	Poisson
Default link:	identity	logit	log
Variance:	Constant σ^2	$\mu(1-\mu)$	μ

Linear model

Variance constant, does not depend on mean!

Logistic regression (binary response)

Poisson regression (response: counts or rates)

Variance depends on mean!

Deviance

- Goodness of fit is measured in deviance instead of RSS based (e.g. R^2):

$$D = -2 \log L$$

  - The likelihood function L depends on the GLM family (error distribution)
  - Statistical programs will typically return the deviance as part of the output
- Deviance plays similar role as residual sum of squares in linear models

Analysis of deviance

- Testing effects of explanatory variables and their interactions
- Equivalent to F-test in linear models
- Instead of F-test, use likelihood ratio test
- Test statistic has a χ^2-distribution
  (D_{full}, $D_{reduced}$ are the deviances of the full and reduced model, respectively, p and q are the corresponding numbers of parameters, and ϕ is the scale parameter)

$$\boxed{\frac{D_{reduced} - D_{full}}{\phi} \sim \chi^2_{p-q}}$$

(equivalent to ratio of respective likelihoods because $D = -2\log L$)

Likelihood ratio test

$$D_{reduced} - D_{full} = -2\log L_{reduced} - (-2\log L_{full})$$
$$= 2*(\log L_{full} - \log L_{reduced})$$
$$= 2*\log\left(\frac{L_{full}}{L_{reduced}}\right)$$

(Likelihood ratio)

- Null-hypothesis H_0:
  - reduced (simpler) model is the "true" model, i.e. any 'extra' parameters in the full model are zero ($\beta_{i(full)} = 0$)
- If difference in deviance is large (i.e. if $L_{full} >> L_{reduced}$) → reject H_0

Example: Effects of a toxin on moths

LOGISTIC REGRESSION

Logistic regression: binomial response example

sex	dose	dead	alive
M	1	1	19
M	2	4	16
M	4	9	11
M	8	13	7
M	16	18	2
M	32	20	0
F	1	0	20
F	2	2	18
F	4	6	14
F	8	10	10
F	16	12	8
F	32	16	4

- Example: Effect of a toxin on moths.
- Batches of 20 moths of each sex were exposed to different doses of a toxin
- Number of animals alive after 3 days were recorded
- Response: dead or alive
- Response variable is binomial (# of "successes", i.e. # dead moths out of total)

Logistic regression: binomial response example

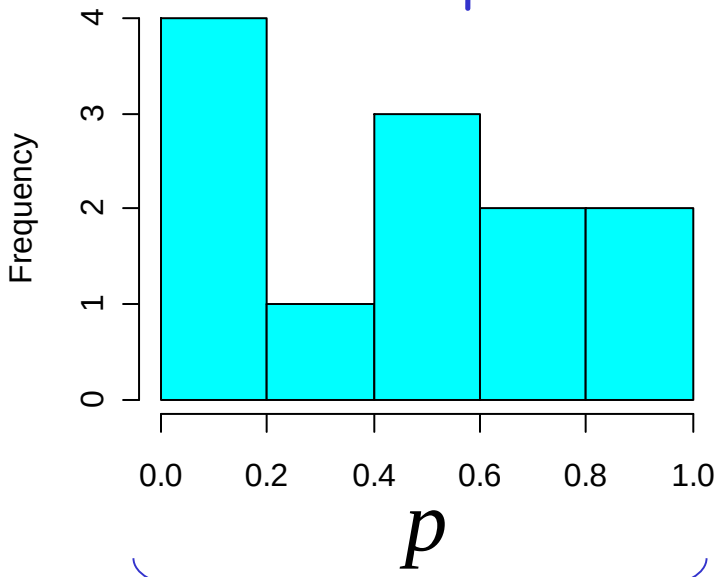
- Response: Number dead / alive per batch
- Explanatory variables: Sex and log(Dose) with interaction (to allow different response by sex)
- Family = binomial
- Link = logit (see below)
- Mean response for a given sex at a given dose (what we actually model!) is the probability of dying (p), which is related to linear predictor through the <u>link function</u>:

$$\text{logit}(p) = \log(\overbrace{p/(1\text{-}p)}^{\text{="odds-ratio"}}) = \underbrace{\boldsymbol{X\beta}}_{\text{="log-odds"}}$$

Logistic regression: binomial response example

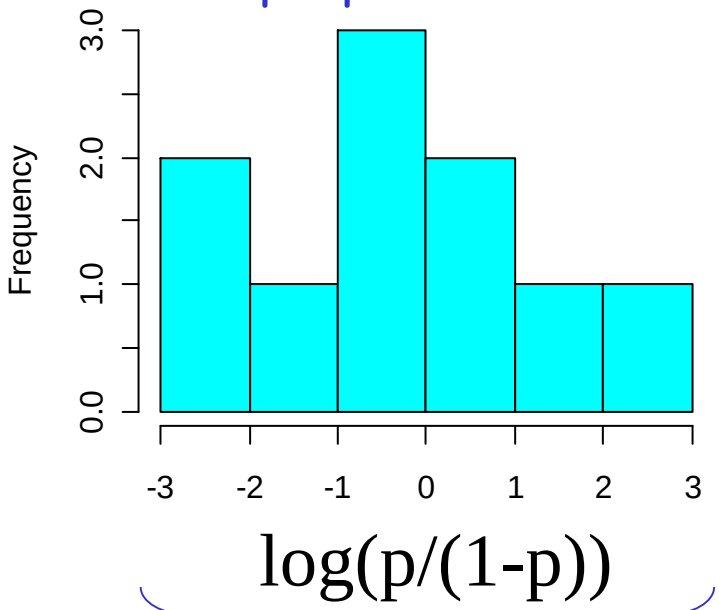
Effect of logit-transformation of proportions:

Observed proportions of dead animals per batch



Range: 0 to 1

Logit-transformed proportions



Range: -∞ to ∞

GLM

Logit model: $$\boxed{\log\left(\frac{p}{1-p}\right) = \mathbf{X\beta}}$$

solve for p!

$$\frac{p}{1-p} = e^{\mathbf{X\beta}}$$

$$p = e^{\mathbf{X\beta}}(1-p)$$

$$p = e^{\mathbf{X\beta}} - e^{\mathbf{X\beta}}p$$

$$p + e^{\mathbf{X\beta}}p = e^{\mathbf{X\beta}}$$

$$p(1+e^{\mathbf{X\beta}}) = e^{\mathbf{X\beta}}$$

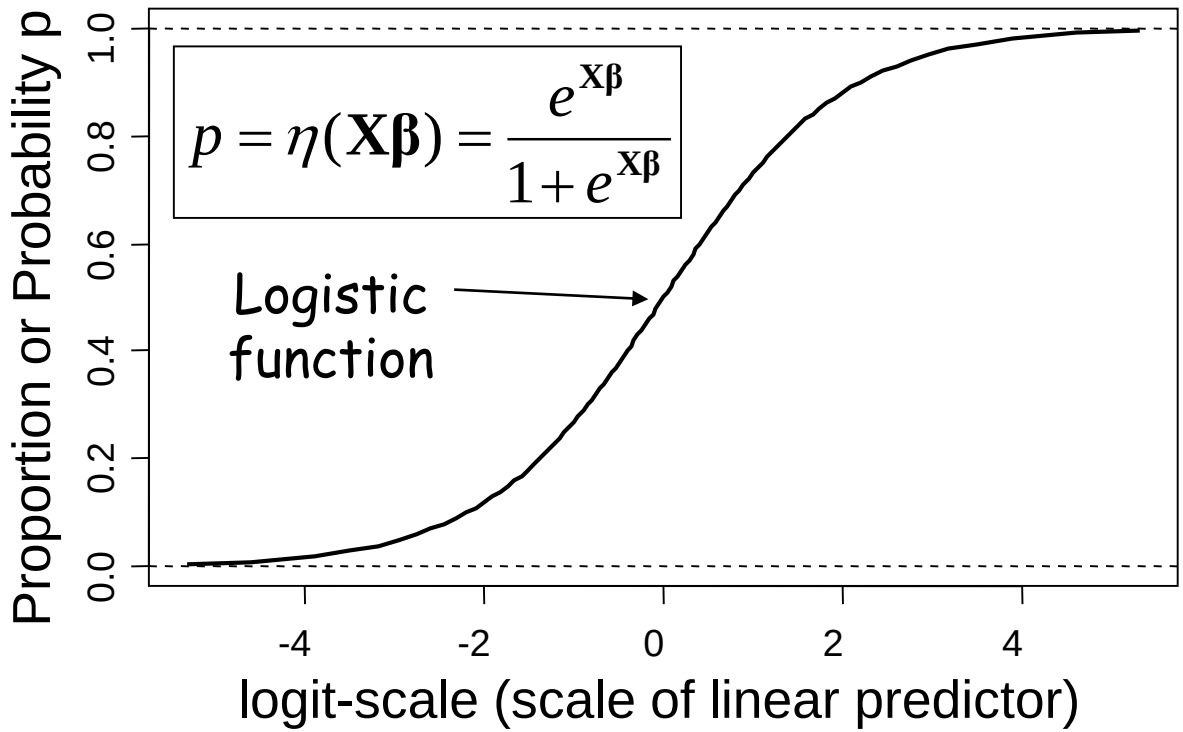
Logistic function: $$\boxed{p = \frac{e^{\mathbf{X\beta}}}{1+e^{\mathbf{X\beta}}} = \frac{1}{e^{-\mathbf{X\beta}}+1}}$$

Logistic regression: binomial response example

Effect of logit-transformation of proportions:

Logistic function is inverse of logit! Back-transformation from logit (p) to p guarantees that:

$$0 < p < 1$$



Logistic regression: $\text{logit}(p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$

Logistic regression: binomial response example

- Likelihood of observing k_i dead moths in a batch of n_i = 20 moths:

$$L_i(k_i) = \binom{n_i}{k_i} p_i^{k_i} (1 - p_i)^{n_i - k_i}$$

- Likelihood of observing $\{k_1, k_2, \ldots\}$ dead moths in batches of n = 20 moths (every batch had 20 moths in the experiment, but n may differ):

$$L(k_1, k_2, k_3, \ldots) = \prod_i \binom{n}{k_i} p_i^{k_i} (1 - p_i)^{n - k_i}$$

Logistic regression: binomial response example

- Probability of a moth in batch i dying due to the toxin is modeled as a function of the dose of the toxin (where l is the logit function):

$$l(p_i) = \alpha + \beta \cdot x_i$$

$$p_i = 1 \big/ (1 + e^{-(\alpha + \beta \cdot x_i)})$$ (=logistic function)

- Hence, find parameters α and β that maximize the likelihood of the observations k_1, k_2, ...:

$$L(k_1, k_2, \ldots) = \prod_i \binom{n}{k_i} \left(\left(1 + e^{-(\alpha + \beta \cdot x_i)}\right)^{-1} \right)^{k_i} \left(1 - \left(1 + e^{-(\alpha + \beta \cdot x_i)}\right)^{-1}\right)^{n - k_i}$$

Logistic regression: binomial response example

Linear predictor ($X\beta$) is handled exactly as in LM:

$$\text{logit}(p) = \boldsymbol{X\beta} = Sex_i + \beta_i * \log_2(\text{dose})$$

Slope by sex

Intercept by sex

(i = male OR female)

Implement through an interaction

R code

```
> ldose <- log2(dose)
> moths.lr <- glm(DA ~ sex*ldose, family = binomial)
> summary(moths.lr)
```

DA is a matrix with two columns:

Dead	Alive
19	1
16	4
11	9
…	…
8	12
4	16

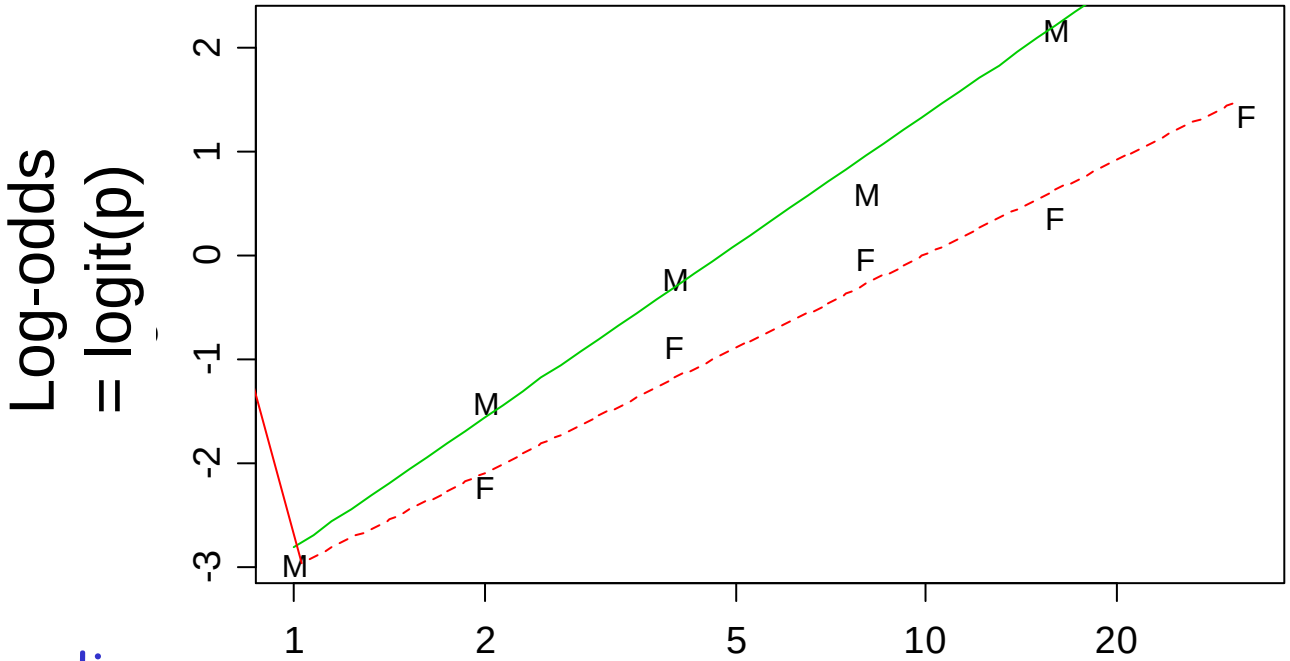
```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.9935     0.5527  -5.416 6.09e-08
sexM          0.1750     0.7783   0.225    0.822
ldose         0.9060     0.1671   5.422 5.89e-08
sexM:ldose    0.3529     0.2700   1.307    0.191

    Null deviance: 124.8756  on 11  degrees of freedom
Residual deviance:   4.9937  on  8  degrees of freedom
AIC: 43.104
```

Based on normal z scores because variance assumed known (binomial)

No overall F-test, but can use χ^2-test {`anova()`} to test for significant reduction in deviance!

Logistic regression: binomial response example

Linear fits on logit (log-odds) scale

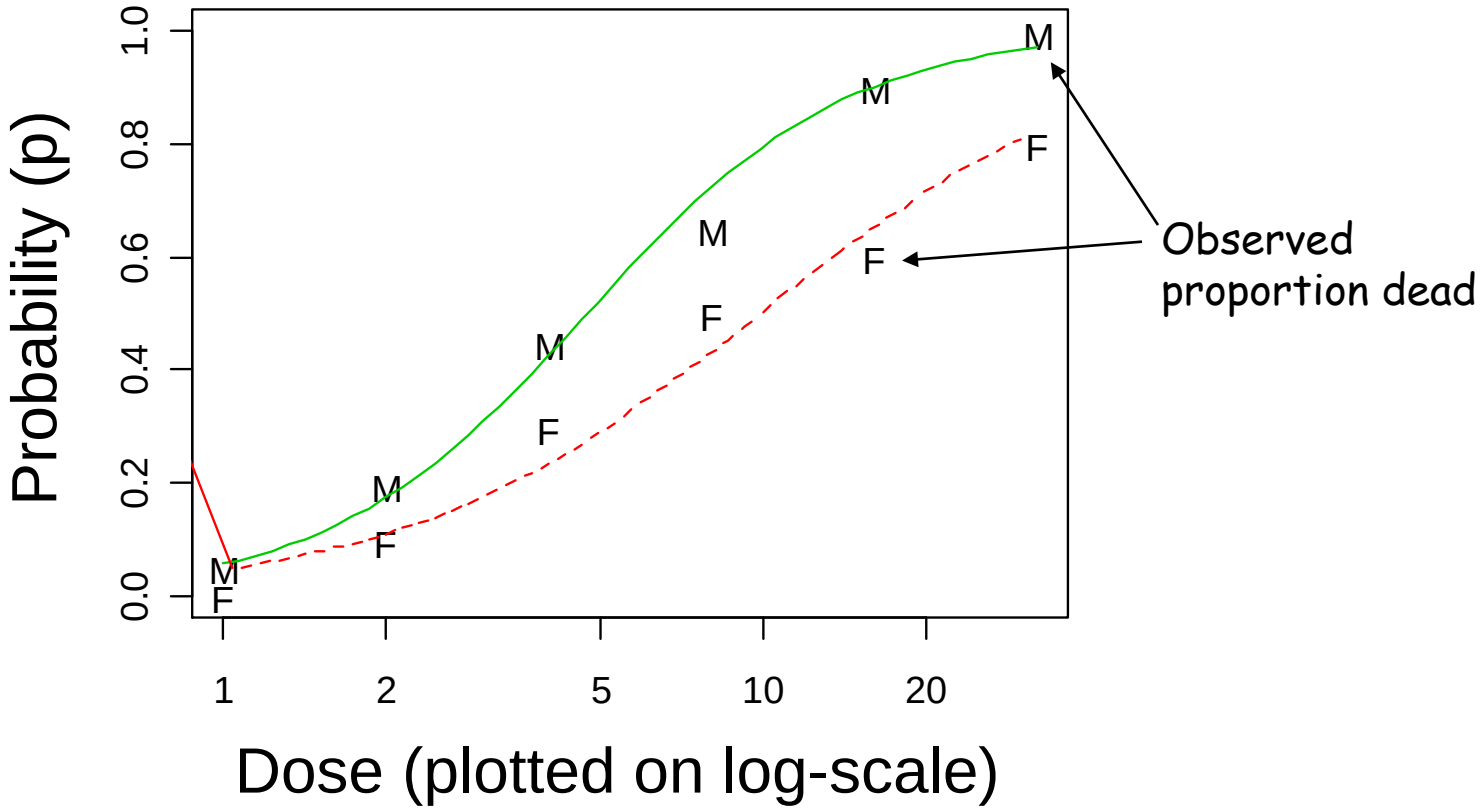


Dose (plotted on log-scale)

Fits are <u>linear</u> on logit scale

Logistic regression: binomial response example

Model fit on "response" scale: estimated probability of dying by sex and dose



Logistic regression: binomial response example

- How to interpret parameters:

R code

```
> moths.lr <- glm(DA ~ sex*log(dose), family = binomial)
> glm(DA ~ sex/log(dose), family = binomial)
> summary(moths.lr)

Coefficients:
                Estimate     ...
(Intercept)     -2.9935      ...
sexM             0.1750      ...
ldose            0.9060      ...
sexM:ldose       0.3529      ...

    Null deviance: 124.8756  on 11  degrees of freedom
Residual deviance:   4.9937  on  8  degrees of freedom
…
```

Equivalent models!

α_F Intercept for females (on logit scale)

α_d Difference in intercept (male - female)

β_F Slope for females (on logit scale)

β_d Difference in slope (male - female)

Logistic regression: binomial response example

Computing predicted values:

To compute predicted values (logit-scale) for **female moths** at different doses:

$$\widehat{logit}(p) = \boxed{\widehat{\alpha_F} + \widehat{\beta_F} * \log_2(dose)}$$

To compute predicted values (logit-scale) for **male moths** at different doses:

$$\widehat{logit}(p) = \boxed{\underbrace{(\widehat{\alpha_F} + \widehat{\alpha_d})}_{\text{Intercept (males)}} + \underbrace{(\widehat{\beta_F} + \widehat{\beta_d})}_{\text{Slope (males)}} * \log_2(dose)}$$

To compute these same values in R:

```
> predict(moths.lr)
```

To compute predicted values on the back-transformed scale:

```
> predict(moths.lr, type = "response")
```

$$\boxed{p = \eta(\mathbf{X\beta}) = \frac{e^{\mathbf{X\beta}}}{1 + e^{\mathbf{X\beta}}}}$$

Logistic regression: binomial response example

- Analysis of deviance:

```
> anova(moths.lr, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: DA

Terms added sequentially (first to last)

          Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                         11    124.876
sex        1    6.077        10    118.799     0.014
ldose      1  112.042         9      6.757 3.499e-26
sex:ldose  1    1.763         8      4.994     0.184
```

similar to p-value from z-test (slide 24)

Logistic regression: binomial response example

- Model comparison:

Re-fit model without interaction:

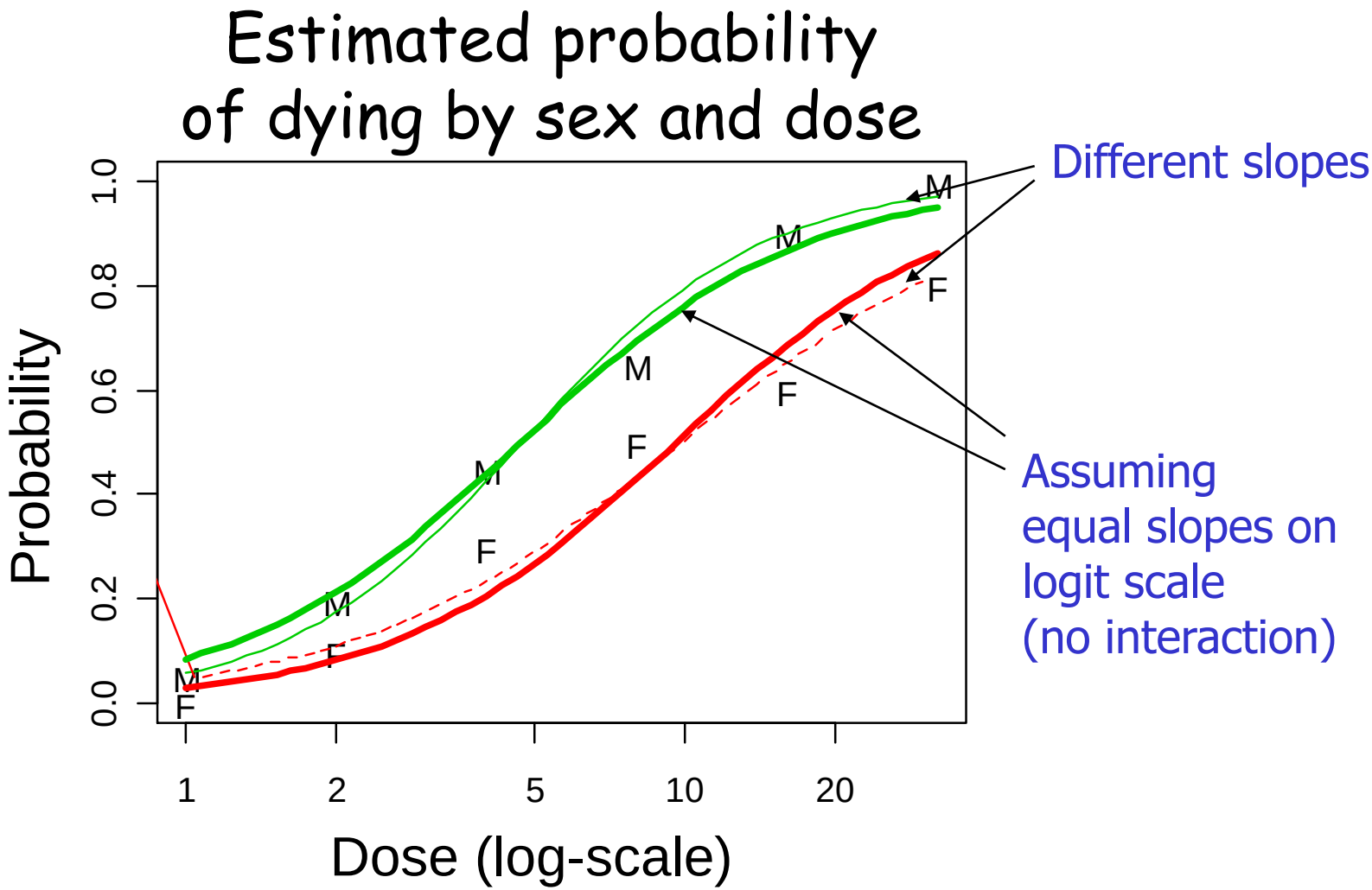
```
> moths.lr2 <- update(moths.lr, ~ . - sex:ldose)
```

Compare models using Analysis of Deviance (likelihood-ratio test):

```
> anova(moths.lr2, moths.lr, test="Chisq")
Analysis of Deviance Table

Model 1: DA ~ sex + ldose
Model 2: DA ~ sex * ldose
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         9     6.7571
2         8     4.9937  1   1.7633    0.1842
```

same as above

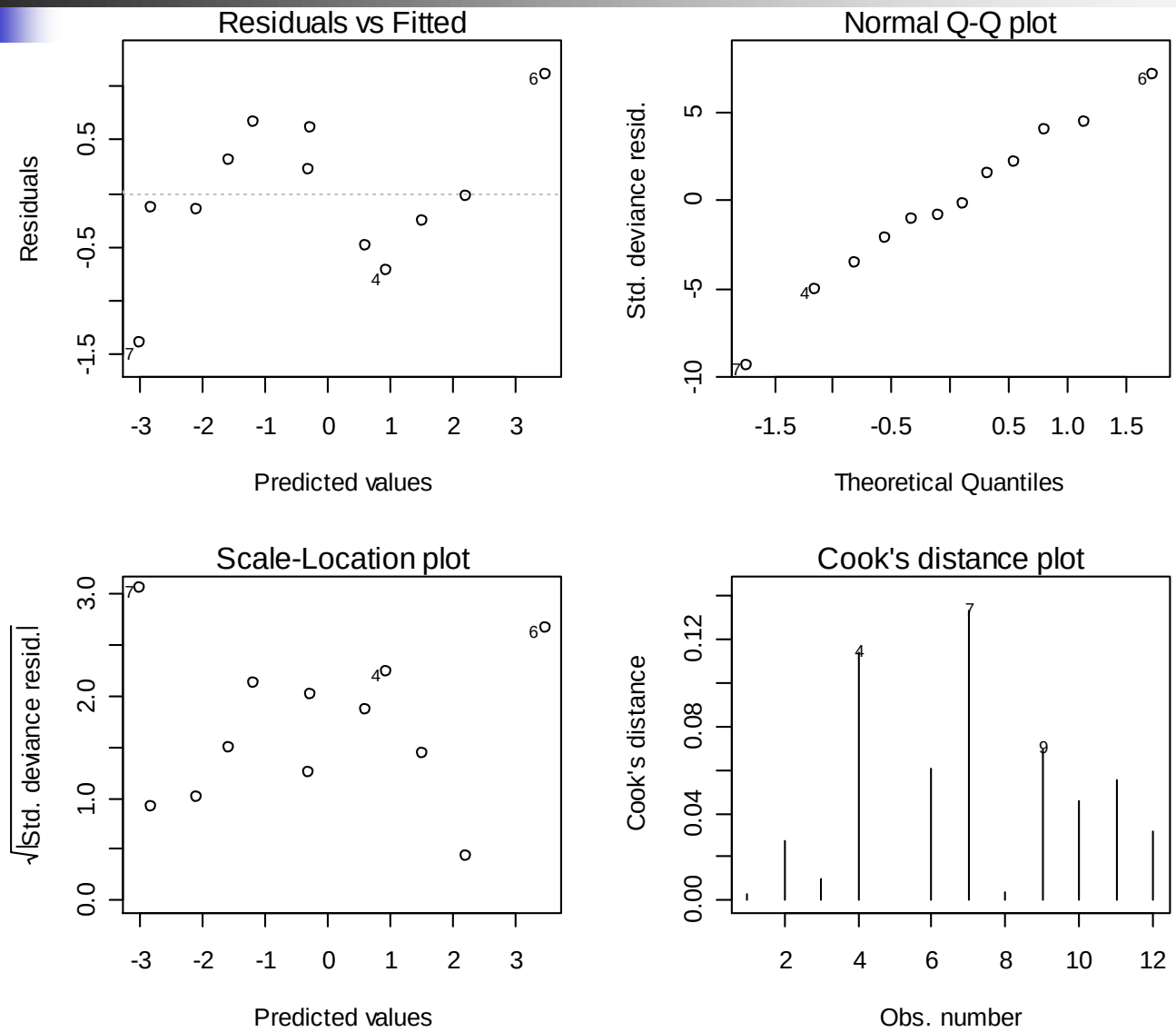
Logistic regression: binomial response example



Model diagnostics

- As for LM, diagnostics use residuals from the fitted model
- For GLM, several different types of residuals are available:
  - Pearson residuals $\boxed{\varepsilon_i = \frac{y_i - \hat{\mu}_i}{\sqrt{Var(\hat{\mu}_i)}}}$ approximately mean 0 and equal variance for all i
    - often badly skewed (asymmetric around 0)
    - measure of goodness of fit for binomial & Poisson models (Sum of squared Pearson residuals = Pearson's χ^2 statistic)
  - Deviance residuals (=contribution of each residual to overall deviance, with appropriate sign, standardized)
    - most useful for diagnostics, default residuals in many R functions (<u>approximately</u> normal, should have equal variances)
  - Response residuals (= observed – predicted response)

Model diagnostics



Deviance residuals

Logistic regression: binary (0/1) response

- Instead of number of "successes" as response variable (binomial response), we use binary response variable (0/1) in the form of a string of zeros and ones corresponding to each "success" or "failure"
- Example: presence/absence of skates in Gulf of Alaska trawl samples
- Goal: Test for differences in the probability of occurrence between years

(Could also analyze moth data that way)

Logistic regression: binary response

- Response: Absence/presence (0/1)
- Explanatory variables: Year (primary variable of interest), Covariates: stratum and /or depth
- Family = binomial
- Link = logit
- What we model: The probability of catching skates (p) in a trawl sample for a given year and stratum and at a given depth; which is related to the linear predictor through the link function:

$$\log(p_{ij}/(1-p_{ij})) = \text{Year}_i + \text{Stratum}_j + \beta * \text{Depth}$$

Logistic regression: binary response

- Model in R:

```
> glm(Skates ~ Year + Stratum
      + Depth, family = binomial)
```

Data:

Binary response!

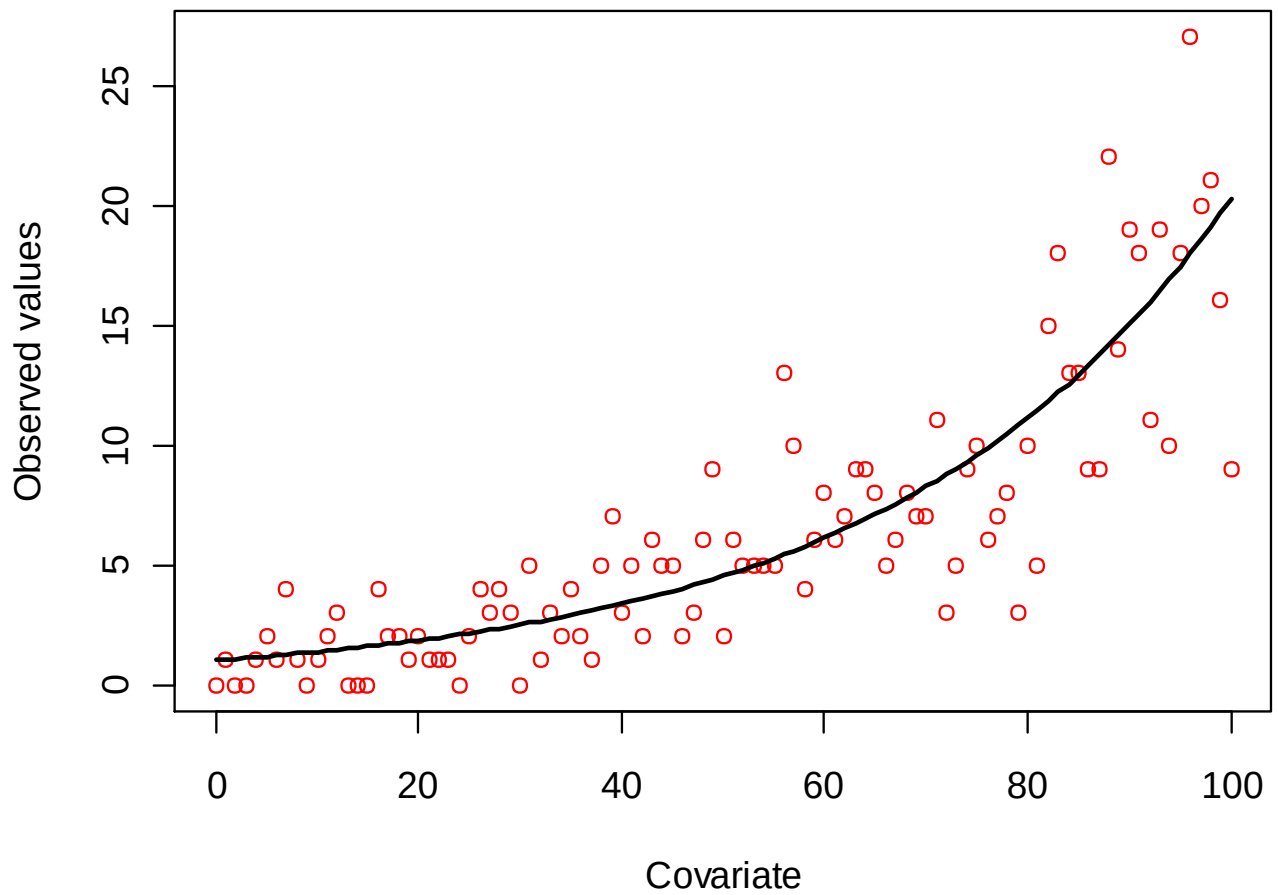
Year	Stratum	Depth	Skates
2001	A	120	0
2001	A	78	0
2001	A	25	0
2001	A	90	1
2001	A	146	0
2001	B	54	1
2001	B	180	0
2001	B	123	0
…			
2003	A	44	1
2003	A	126	0
2003	A	67	0
2003	A	35	1
…			
2003	B	134	0

Regression models

- Linear Models (LM)
  - Simple / multiple linear regression
  - Analysis of (co)variance (ANO(C)VA)
- Generalized Linear Models (GLM)
  - Binomial models (logistic regression)
  - Poisson & negative binomial models
  - Multinomial & Zero-inflated models
- Generalized Additive Models (GAM)
  - Non-parametric smoothers
- Mixed-effects models (linear/non-linear)
- Non-linear models (NLM)

Poisson regression

An artificial example



Mean response:

$\mu_i = \exp(0.01 + 0.03 * X)$

$\log(\mu_i) = 0.01 + 0.03 * X$

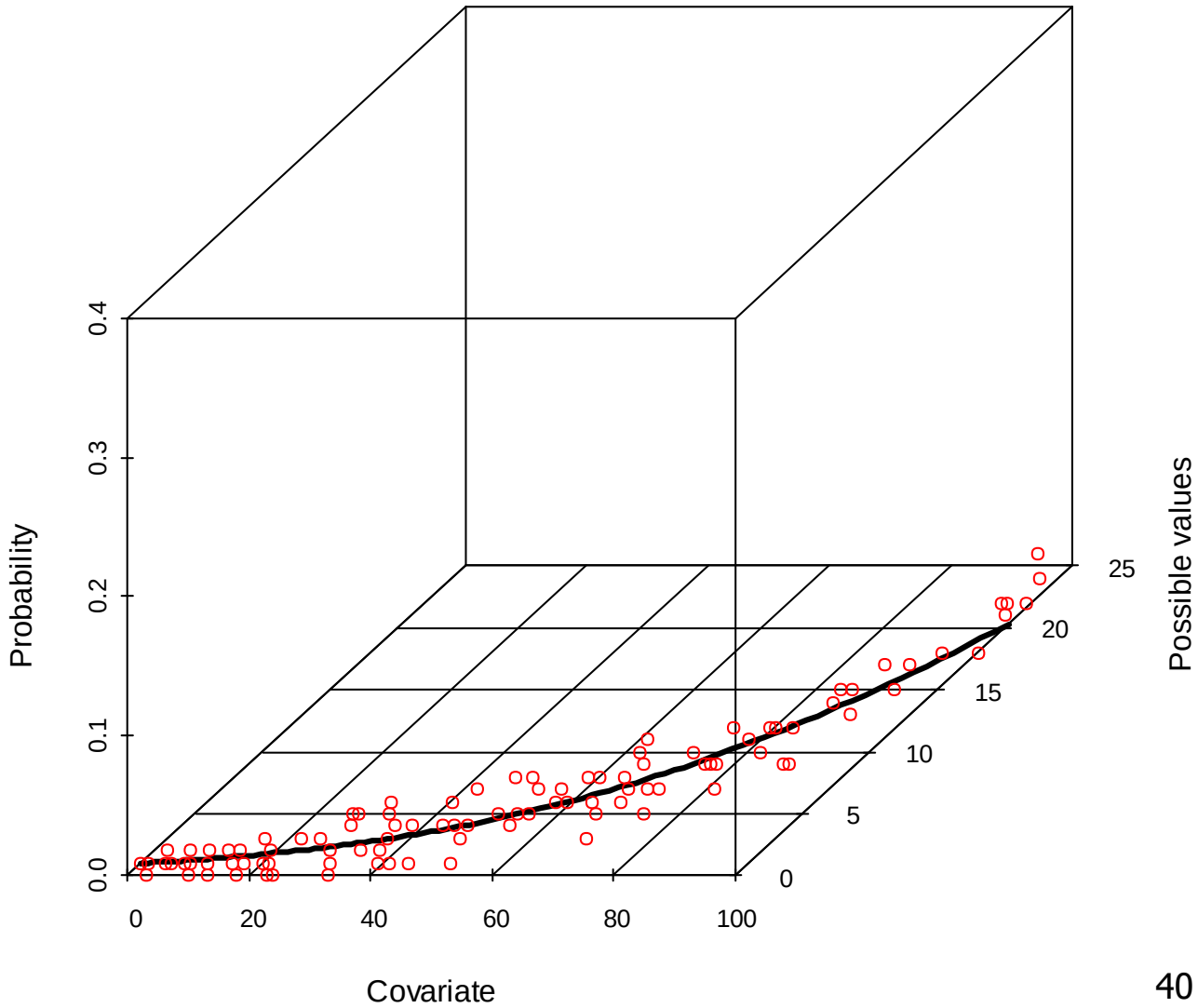
Observations Y_i

$E(Y_i) = \mu_i$

$var(Y_i) = \mu_i$

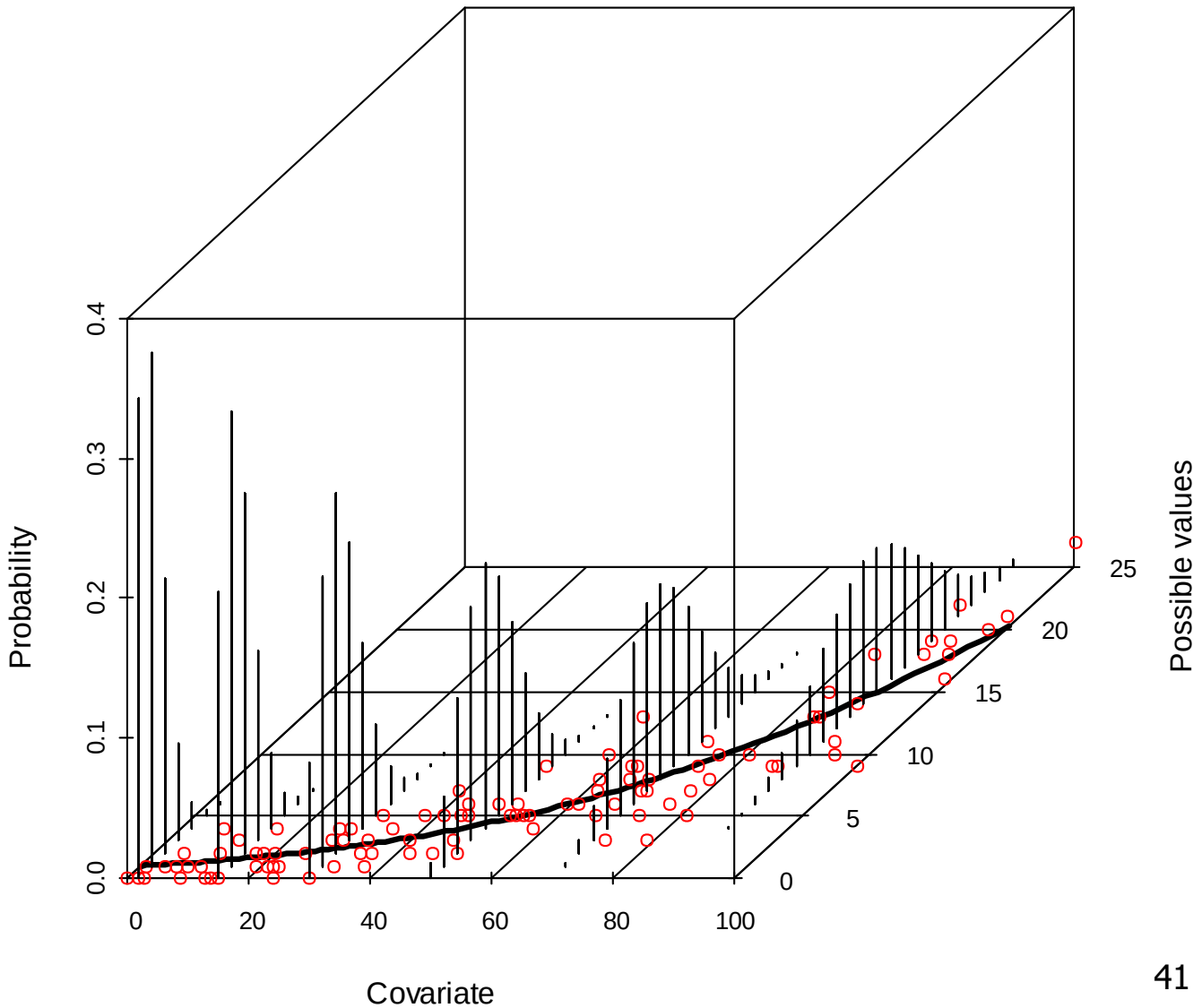
Poisson regression

Same data!
Different view!



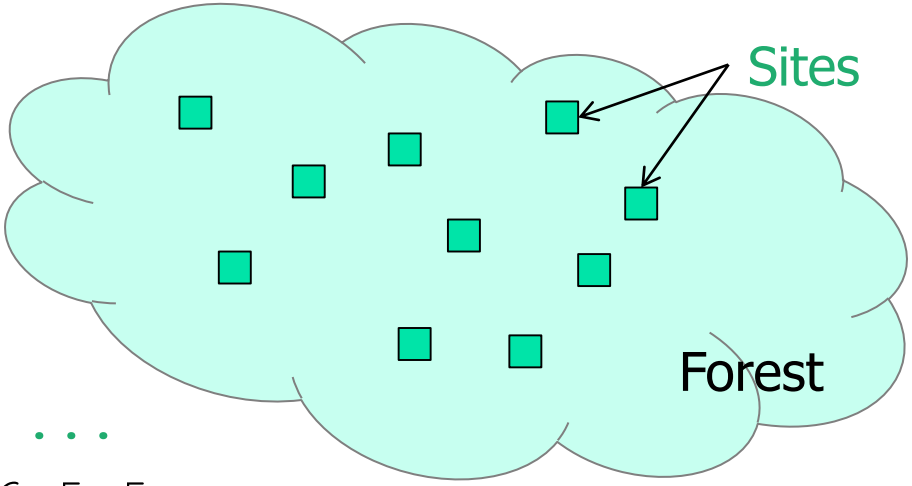
Poisson regression

Probability distribution of observed values at various levels of covariate



Poisson regression example

- Observational study of salamanders. Number of salamanders per unit area at randomly selected sites
- Explanatory variables:
  - Percent cover
  - Forest age




```
SITE : 1 2 3 4 5 6 7 8 9 10 ...
SALAMAN : 13 11 11 9 8 7 6 6 5 5 ...
PCTCOVER : 85 86 90 88 89 83 83 91 88 90 ...
FORESTAGE: 316 88 548 64 43 368 200 71 42 551 ...
```

- Problem: Predict salamander density as a function of %cover and forest age

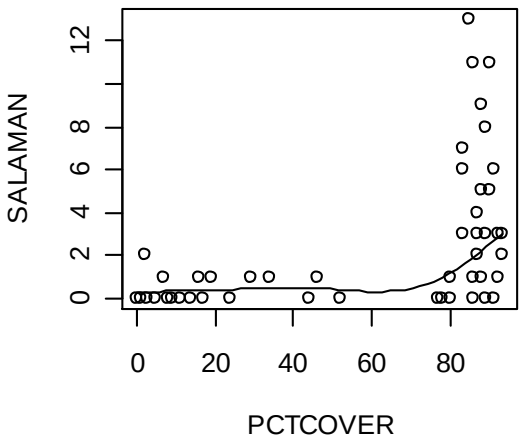
Poisson regression example

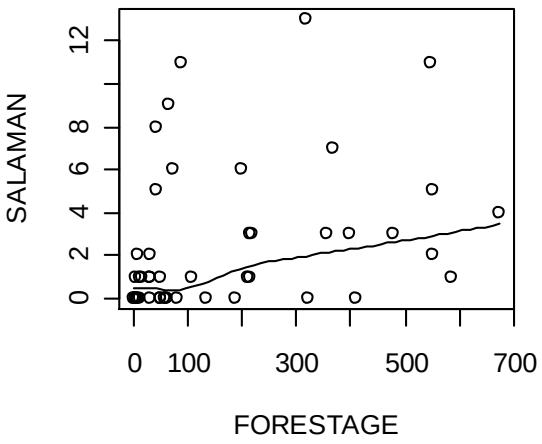
- Response: Counts of salamanders per unit area
- Explanatory variables: Cover and forest age
- Family = poisson
- Link = log ("log-linear model", could also use other links, for example identity)
- What we model: The number of salamanders in a given area μ; which is related to the linear predictor through the link function:

$$\log(\mu) = \alpha + \beta_1 * \text{cover} + \beta_2 * \text{age} + \beta_3 * (\text{age} * \text{cover})$$

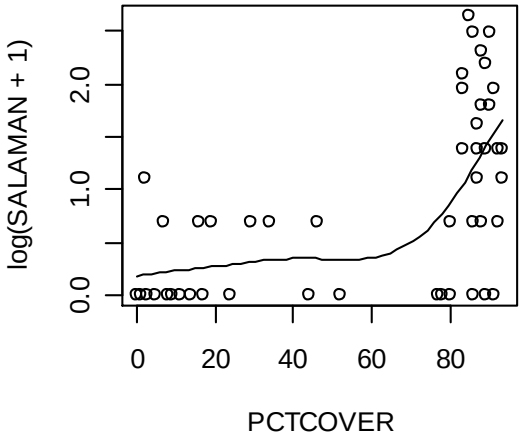
Poisson regression example

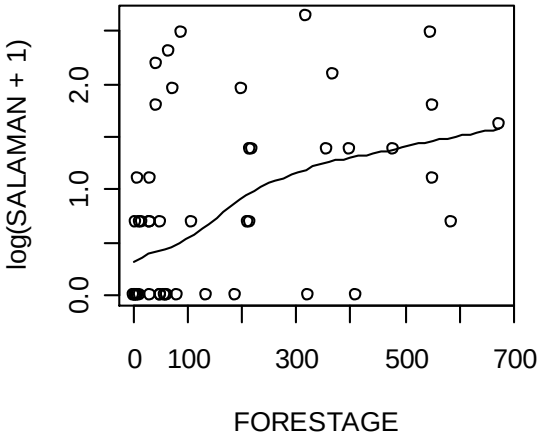
"raw" scale





log scale
log(y+1)





Poisson regression example

Default link: log

```
glm(formula = SALAMAN ~ PCTCOVER * FORESTAGE, family = poisson,
    data = salamander)
...

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.388e+00  5.038e-01  -2.754  0.00588 **
PCTCOVER             3.147e-02  6.145e-03   5.121 3.04e-07 ***
FORESTAGE           -2.812e-03  6.800e-03  -0.414  0.67918
PCTCOVER:FORESTAGE   3.141e-05  7.625e-05   0.412  0.68033
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 190.22  on 46  degrees of freedom
Residual deviance: 121.13  on 43  degrees of freedom
AIC: 214.19

Number of Fisher Scoring iterations: 6
```

Test based on normal distribution (variance assumed fixed!) H_0: $\beta = 0$

Variance fixed (=mean)

Null model (intercept)

This model

"RSS equivalent"

Model selection criterion!

Iterative solution!

No overall F-test, but can use χ^2-test {`anova()`} to test for significant reduction in deviance!

A note on R^2 values

- What is R^2?
  - R^2 as explained variability
  - R^2 squared as improvement from null model to fitted model
  - R^2 as the square of the correlation (between the predicted values and the actual values)
- In OLS, R^2 is based on residual versus total sum of squares:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{\mu}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}$$

→ Not the same interpretation in GLM (maximum likelihood estimation)

A note on R^2 values

- A number of different 'Pseudo-R2' have been proposed for models fit via maximum likelihood

Efron's:

$$Pseudo - R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{\mu}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}$$

McFadden's

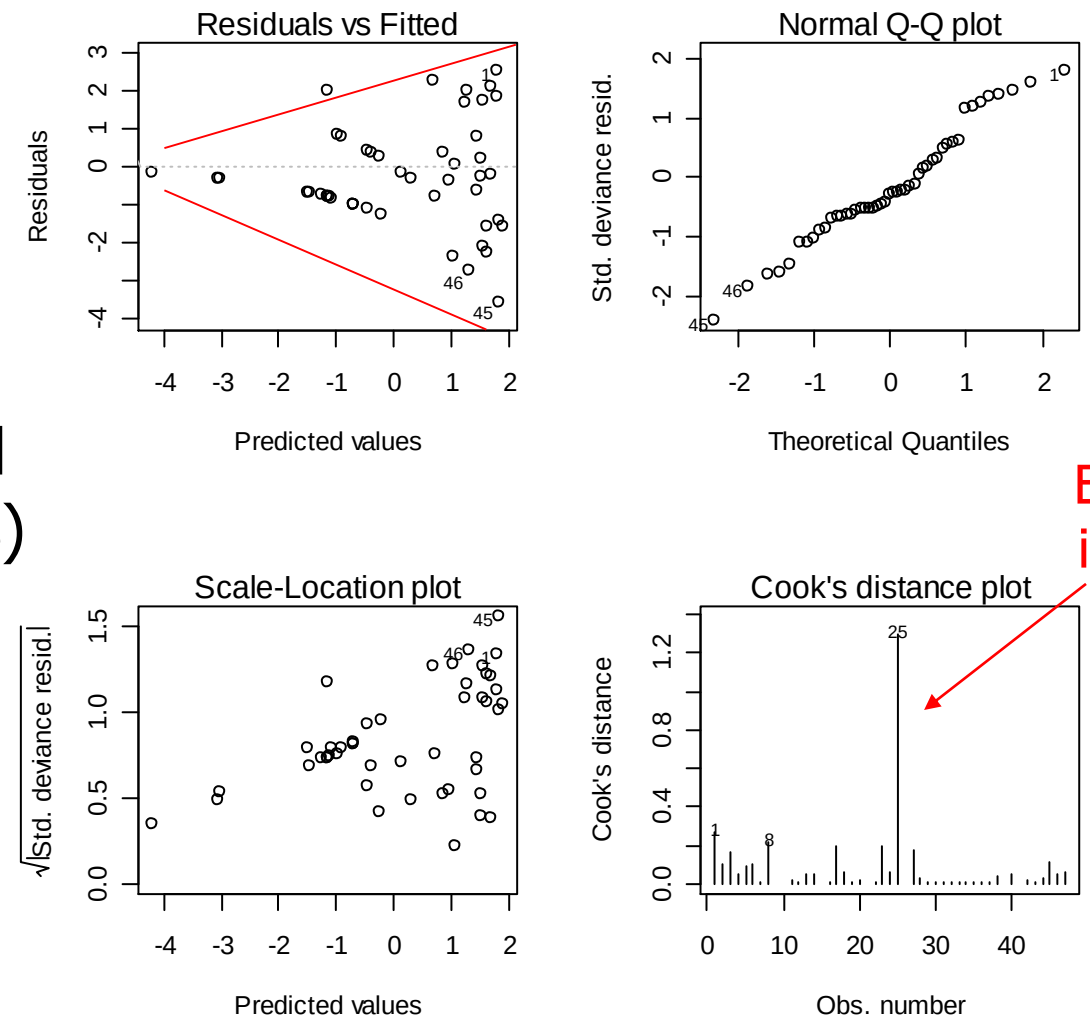
$$Pseudo - R^2 = 1 - \frac{Residual\ Deviance}{Null\ Deviance}$$

See https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/ for a good discussion

Poisson regression example

→ Heteroscedasticity! (Overdispersion?)

<u>Deviance residuals</u> (should be approx. normally distributed with equal variances)



Extremely influential point

Overdispersion

- Binomial and Poisson models assume the following variances:
  - Binomial: $\mathrm{Var}(Y) = n\,p\,(1 - p)$
  - Poisson: $\mathrm{Var}(Y) = \lambda$

  No "variance parameter" (unlike, e.g., normal distr.)
- In practice, data often have some degree of overdispersion relative to the standard assumption, which can be described by an overdispersion parameter ϕ :
  - Binomial: $\mathrm{Var}(Y) = \phi * n\,p\,(1 - p)$
  - Poisson: $\mathrm{Var}(Y) = \phi * \lambda$

Poisson regression example with overdispersion

```
fit <- glm(formula = SALAMAN ~ PCTCOVER * FORESTAGE,
           family = quasipoisson)
> summary(fit)

...
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -1.388e+00  8.507e-01  -1.631   0.1102
PCTCOVER             3.147e-02  1.038e-02   3.033   0.0041 **
FORESTAGE           -2.812e-03  1.148e-02  -0.245   0.8077
PCTCOVER:FORESTAGE   3.141e-05  1.287e-04   0.244   0.8084
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 2.851019)

    Null deviance: 190.22  on 46  degrees of freedom
Residual deviance: 121.13  on 43  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
```

Identical estimates

wider std. err.

$\sqrt{\phi} \cdot SE^*$

SE^* = std.err. estimated without overdispersion (slide 44)

t-tests (var. estimated)

H_0: $\beta = 0$

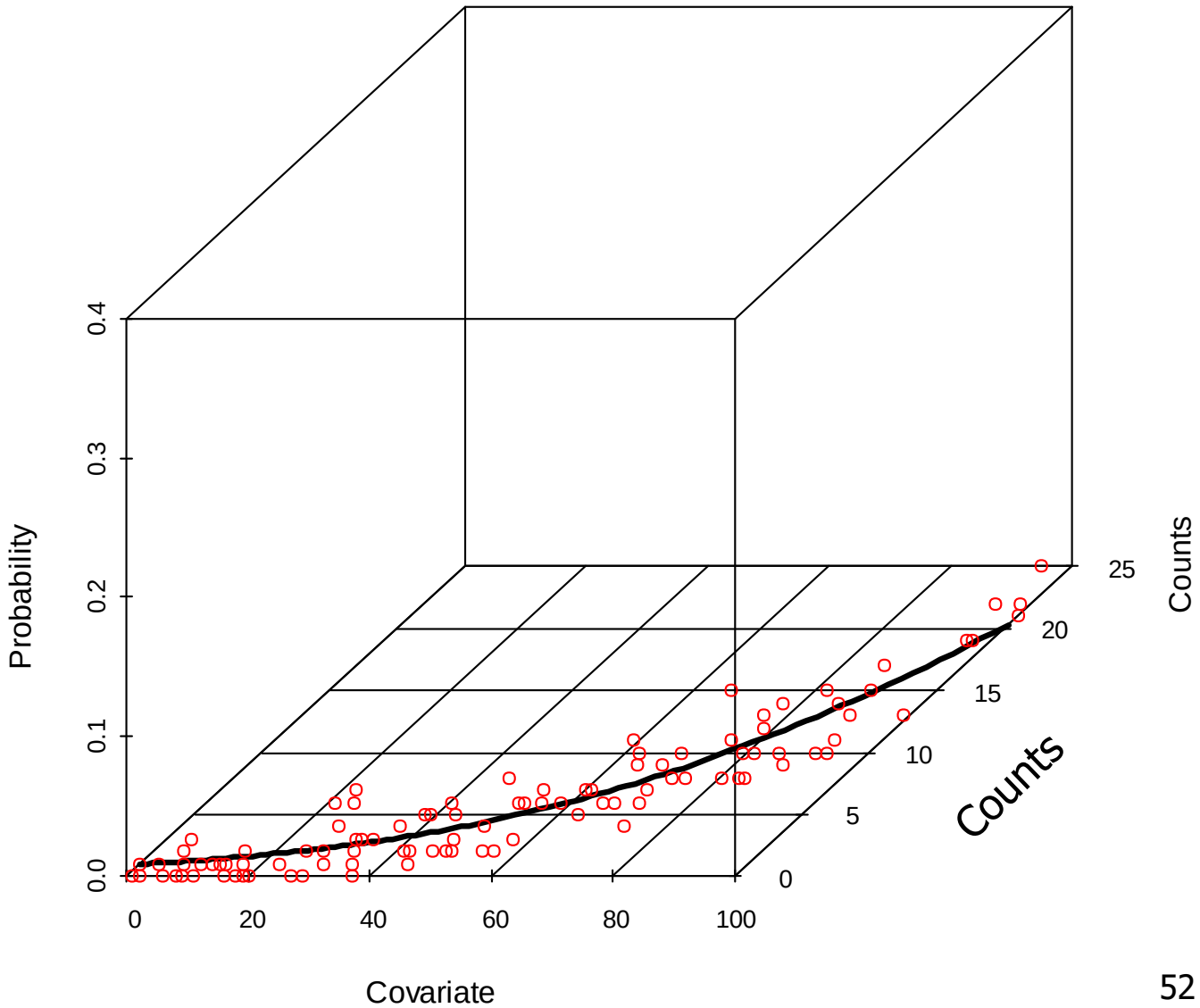
Identical to model without overdisperion!

Overdispersion parameter: ϕ

Negative binomial model

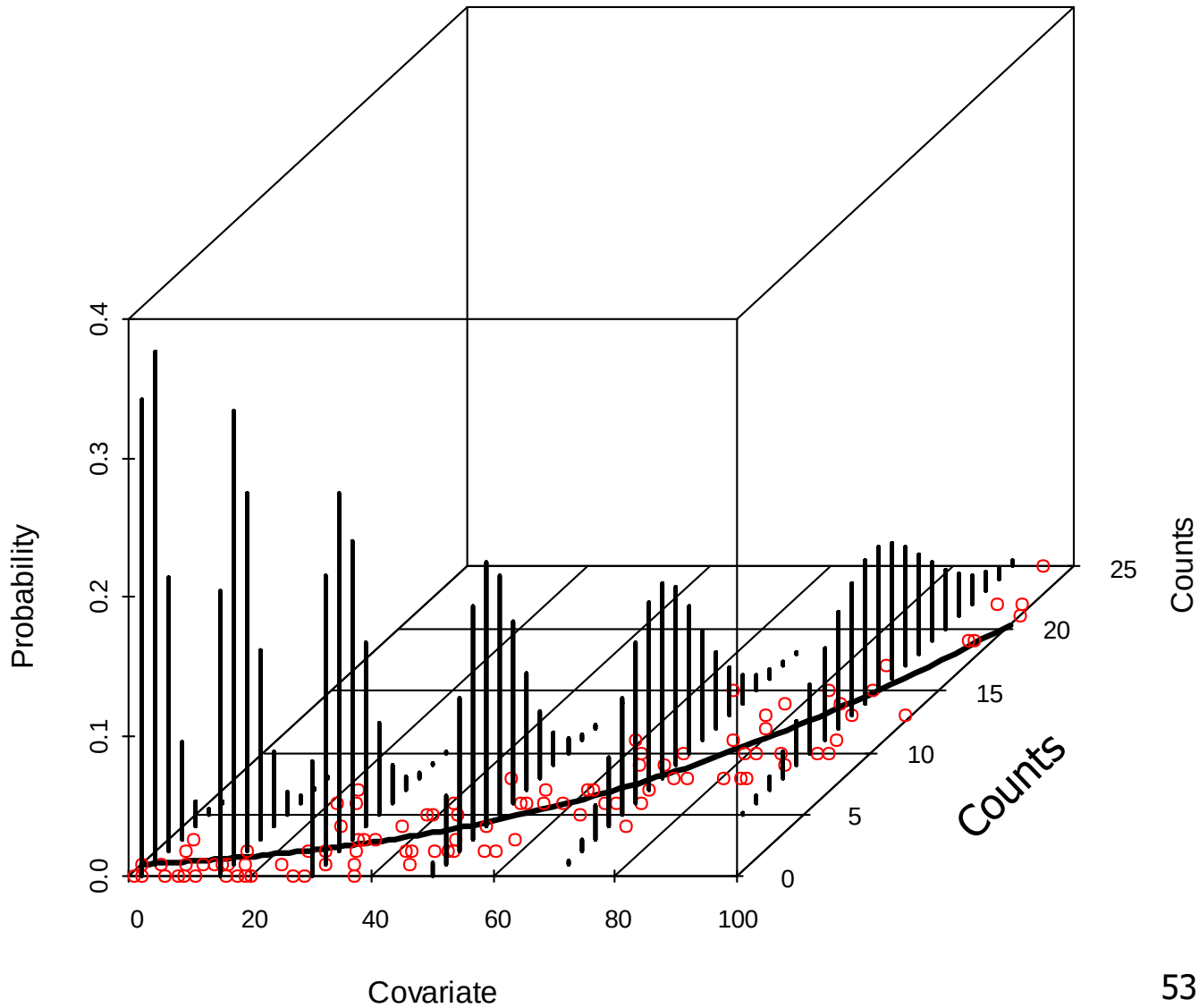
- Alternative to overdispersed Poisson distribution (i.e. when variance is larger than mean)
- Often used for animal counts if individuals have a clustered distribution
  (if cluster size follows logarithmic series, total number of individuals follows negative binomial)
- Many other derivations/interpretations of the negative binomial, for example as a Poisson process with gamma variance at each mean level μ

Negative binomial regression



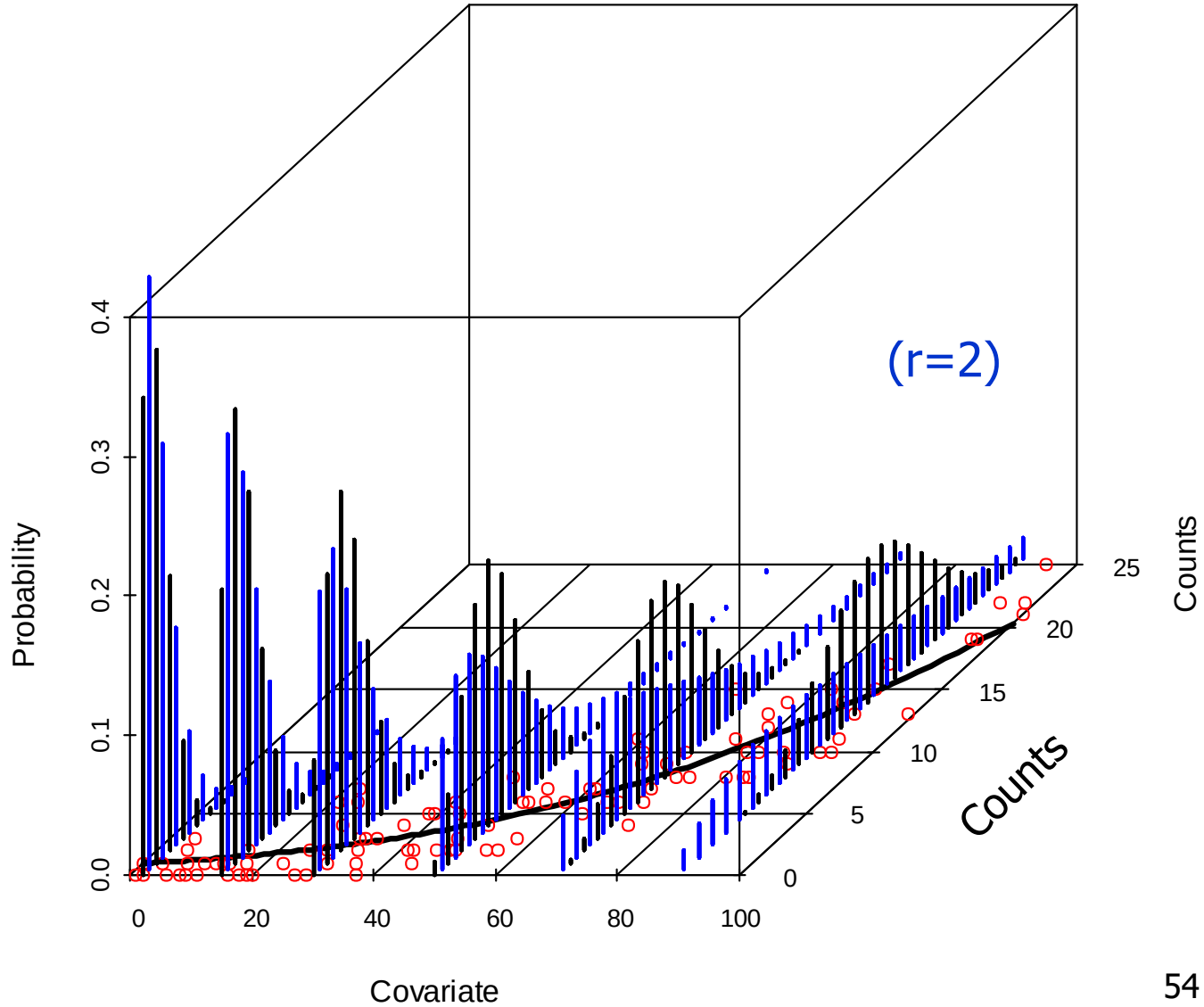
Negative binomial regression

Poisson probability distribution of observed values at various levels of covariate



Negative binomial regression

Poisson & Negative Binomial(2) probability distribution of observed values at various levels of covariate



Negative binomial distribution

- Probability function with parameters "size" (r) and probability (p):

$$\Pr(X = k) = \binom{r+k-1}{k} \mathrm{p}^{\mathrm{r}} (1-\mathrm{p})^{\mathrm{k}} \qquad \mathrm{k} = 0,1,2,\ldots$$

```
neg.binom.pdf <- function(r, p, k) choose(r+k-1, k) * p^r * (1-p)^k
```

Same as:

```
dnbinom(k, size=r, prob = p)
```

OR:

```
dnbinom(k, size=r, mu = m)
```

- Mean and variance (2 parameterizations)

$$E(X) = r\frac{1-p}{p} \qquad \mathrm{var(X)} = r\frac{1-p}{p^2}$$

or:

$$E(X) = \mu \qquad \mathrm{var(X)} = \mu + \mu^2 / r$$

'theta' (θ) parameter in '`glm.nb()`'

Negative binomial distribution

- A third, common parameterization (e.g. SAS, SPSS, Stata) uses 'α' to describe the additional variability:

$$E(X) = \mu_i$$

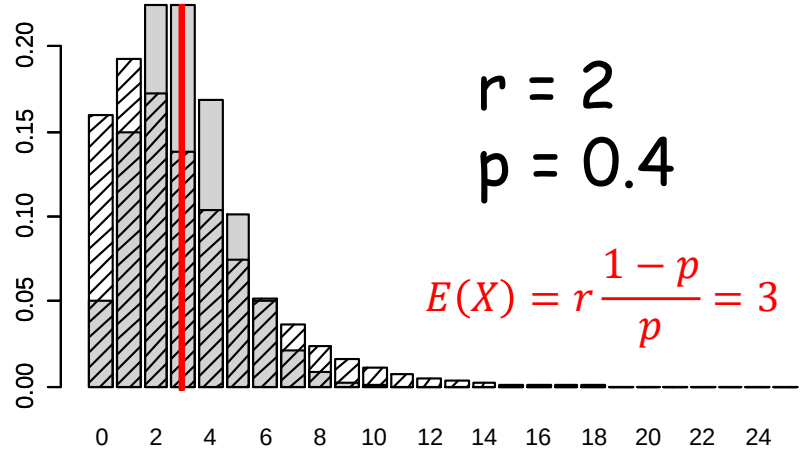
$$\mathrm{var}(X) = \underbrace{\mu_i}_{\text{Poisson variance}} + \underbrace{\alpha \cdot \mu_i^2}_{\text{Additional variance}}$$

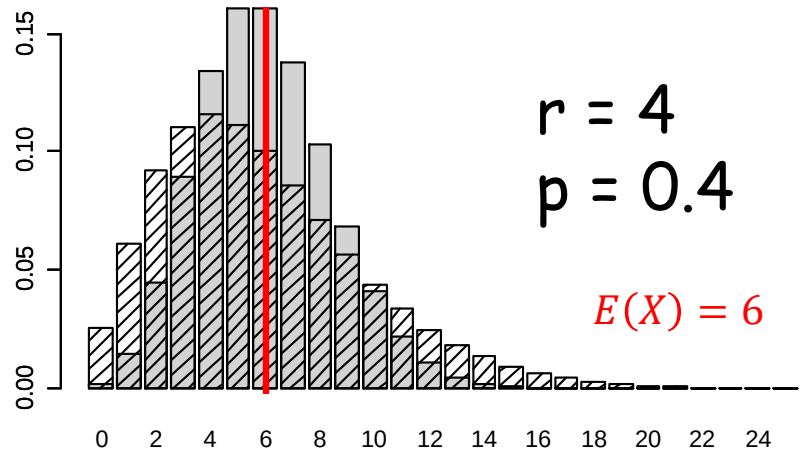
where:
$\alpha = 1/r$

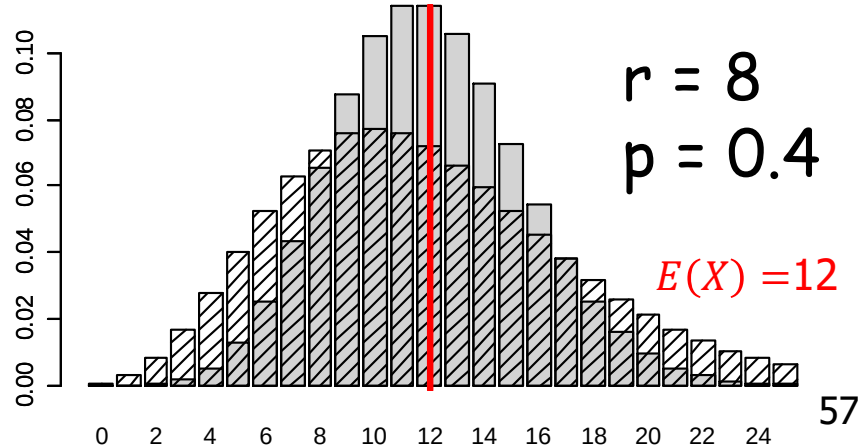
GLM

MSL FISH 604

- Examples of negative binomial probability distribution (hatched bars) relative to Poisson distribution (grey) with the same mean
- Variance increases with smaller p







Negative binomial model

- Implementation in R
  - Function `glm.nb()` in the MASS package
    - Estimates dispersion parameter 'theta' (=inverse of dispersion parameter 'alpha' estimated by some other packages (SAS, SPSS, Stata))
  - `gam()` has families '`nb`' and '`negbin`'
  - Function `gamlss()` in package 'gamlss' with argument: '`family = NBI`'

    (Stasinopoulos DM, Rigby RA (2007). Generalized Additive Models for Location Scale and Shape (GAMLSS) in R." Journal of Statistical Software, 23(7). URL http://www.jstatsoft.org/v23/i07/)

Poisson & NB models

- Poisson regression as special case of negative binomial regression
  - Model based on Poisson distribution is <u>nested</u> within the same model based on the NB distribution
  - Hence significance of over-dispersion can be evaluated via likelihood ratio test!
    (see Lab)

Summary

- Logistic regression for binomial (0/1) data
  - Model probability of "success" (1)
  - Logit as linear function of predictors
- Poisson regression to model counts (0:n)
  - Model mean count or "rate" of a (rare) event
  - ("log-linear models" for contingency tables)
- Overdispersion
  - Quasi-binomial or quasi-Poisson to account for "extra" variation (no true likelihood!)
- Negative binomial models for over-dispersed count data

References

- McCullagh, P., and Nelder, J.A. 1989. Generalized Linear Models. Chapman and Hall, London.
  Classical, comprehensive reference
- Firth, D. 1991. Generalized linear models. In Statistical theory and modelling. Edited by D.V. Hinkley, N. Reid and E.J. Snell. Chapman and Hall, London. pp. 55-82.
  good but technical review paper - see pdf file!
- Faraway, J.J. (2004). Extending the Linear Model with R. Chapman & Hall/CRC. Basic introduction with R examples.
- Dobson, A.J. and Barnett, A. G. (2008). An Introduction to Generalized Linear Models. 3rd edition. CRC, Boca Raton.
- McCulloch, C.E., and Searle, S.R. (2005). Generalized, Linear, and Mixed Models. John Wiley & Sons. (available through UAF Ebook library!)