

Applied Biostatistics

- Statistical modeling overview
- Exponential family
- Generalized linear models (GLM)
- Horseshoe crabs
- Analysis of horseshoe crab data using logistic regression
- Odds, odds ratio interpretation of logistic regression
- Horseshoe crab logistic regression model : 1 variable
- Inference for logistic regression
 - CI/test for *coefficients*
 - CI for *probabilities*
- Multiple logistic regression
- Logistic regression with indicators
- Assessing model fit
- Count data and Poisson regression
- Comparing models

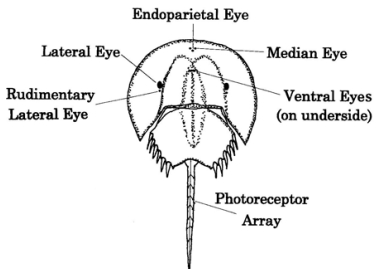
Horseshoe crabs

- Very old (\sim 450 million years), so sometimes called 'living fossils'
- 4 species
- Not actually 'crabs', they are arachnids (like spiders)
- Females \sim 30% bigger than males
- Few survive into adulthood
- Important in biomedical research – their blood has good anti-bacterial properties and is used in developing vaccines and endotoxin testing



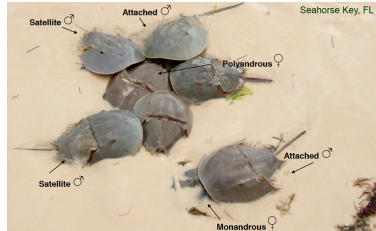
Mating affected by male's condition

- Males are either *attached* or unattached : *satellites* or more distant
 - Attached males are :
 - lighter in color
 - more slime
 - less fouling
 - carapace, eyes and spine in better condition
 - younger
- than unattached males



Sexual biology of horseshoe crabs

- Migrate for spawning in shallow water
- Nesting is synchronized and seasonal
- Tend to nest in (small number of) protected areas
- Reproductive competition in male *Limulus polyphemus* horseshoe crabs
- Operational sex ratio is usually male-biased :
competitive males per female $\sim 1 - 6$



Scientific aim

- Suppose now that we are interested in investigating whether a female horseshoe crab has a satellite or not
- This is a *binary* response
- **Activity** : think about how you might do this and what information (variables) you could collect to study this _____

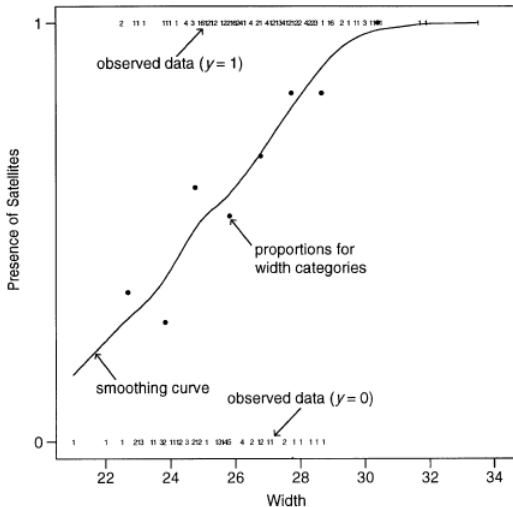
Data for the study

Data on $n = 173$ female horseshoe crabs.

- C = color (1,2,3,4=light medium, medium, dark medium, dark).
 - S = spine condition (1,2,3=both good, one worn or broken, both worn or broken).
 - W = carapace width (cm).
 - Wt = weight (kg).
 - Sa = number of satellites (additional male crabs besides her nest-mate husband) nearby.
-
- **BUT** : what are we going to do with this information ??
 - \Rightarrow need a (statistical) *model*

Exploring the data : carapace width

- Let's first focus on the simplest case where there is only a single variable : carapace width



Statistical modeling

- Goal : to capture important characteristics of the *relationship* between one (or several) explanatory

- Many models are of the form :

$$g(Y) = f(\mathbf{x}) + \text{erreur}$$

- *Differences* between models : the forms of g , f and distributional assumptions about the error term

- Examples of models :

- Linear : $Y = \beta_0 + \beta_1 x + \epsilon$

- Linear $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$

- (Intrinsically) nonlinear : $Y = \alpha x_1^\beta x_2^\gamma x_3^\delta + \epsilon$

- Generalized linear model (e.g. Binomial) :

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x + \beta_2 x_2$$

- Cox proportional hazards model (used in survival analysis) : $h(t) = h_0(t) \exp(\beta x)$

Linear models

- A simple model : $E(Y) = \beta_0 + \beta_1 x$
- Gaussian measurement model : $Y = \beta_0 + \beta_1 x + \epsilon, \epsilon \sim N(0, \sigma^2)$
- More generally : $Y = X\beta + \epsilon$, où Y est $n \times 1$, X is $n \times p$, β is $p \times 1$, ϵ is $n \times 1$, often supposed $N(0, \sigma^2 I_{n \times n})$
- Important application : analysis of designed experiments :
 - a design matrix X such that for the response variable Y : $E(Y) = X\beta$,
where β is a vector of *parameters* (ou contrastes)
 - There are several ways to specify the matrix X for a specific design (this corresponds to the parameterization of the model)
 - \Rightarrow ANOVA

Linear regression model (again)

- For all the linear models that we have seen this semester, the *response variable* has been modeled as a *Normal RV* :

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

- Equally :

$$Y \sim N(\mu, \sigma^2), \quad \mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

- Suitable for a *continuous* response
- **NOT** for a *binary* response
- *Generalized linear models (GLMs)* : generalization of linear models for modeling non-normal response variables
- We will study *logistic regression* for a *binary response variable*

Modification of the response

- Instead of modeling the response directly, could instead model the *probability* of obtaining the value '1' ('success') (that is, *the expected value of the response*)
- Problems :
 - could lead to fitted values outside of *outside of* $[0, 1]$
 - normality assumption on errors is *false*
- Instead of modeling the expected response *directly* as a linear function of the predictors, model a *suitable transformation*
- For binary data, this is generally taken to be the *logit* (or *logistic*) transformation

Generalized linear model : theory

- GLMs allow unified treatment of statistical methods for several important classes of models
- The distribution of the response Y is supposed to belong to an *exponential family* : $f(x | \eta) = h(x) \exp[\eta^T T(x) - A(\eta)]$.
- (Many distributions can be respresented in this form, including the binomial, Normal, Poisson, exponential)
- GLMs are formed from *three components* :
 - **random component** : the *reponse variable* Y , a random component whose distribution belongs to the exponential family
 - **deterministic component** : the *linear predictor*
 $\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$
 - **link function** : describes the *functional relation* between the linear predictor and the mathematical expectation of the response variable Y

Linear models : a new view

- For a linear model :

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \epsilon, \text{ where } \epsilon \sim N(0, \sigma^2)$$

- The *expected response* is $E[Y | x] = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$
- Let η be the *linear predictor* $\eta = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$
- For the (ordinary) linear model : $E[Y | x] = \eta$
- For a *generalized linear model*, there is a *link function* g that relates η with the expected response : $g(E[Y | x]) = \eta$
- For the (ordinary) linear model, $g(y) = y$ (*link = identity*)

Link function

- Generally more clear when we consider the *inverse of the link function* :

$$E[Y|x] = g^{-1}(\eta)$$

- For a binary response (values 0 or 1), then

$$E[Y | x] = P(Y = 1 | x)$$

- In this case, a practical function is

$$E[Y | x] = P(Y = 1 | x) = \frac{e^\eta}{1 + e^\eta}$$

- The corresponding link functions (that is, the inverse of this function) is called the *logit*

- $\text{logit}(x) = \log\left(\frac{x}{1-x}\right)$

- The *logistic regression* models the logit as a function of the predictor variables

Logit transformation

- $\text{logit}(\pi(x)) = \log \text{odds}(\pi(x)) = \log \frac{\pi(x)}{1 - \pi(x)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$

- Then, $\pi(x_1, \dots, x_k) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)}$

- Parameter estimation by *maximum likelihood*

- *Interpretation* : the parameter β_k is such that $\exp(\beta_k)$ is the *OR* (odds ratio) that the response takes value 1 when x_k goes up by 1, when the remaining variables are constant
 $\Rightarrow \beta = \log \text{OR}$

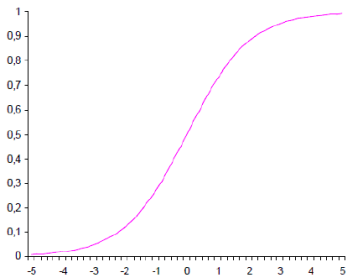
- For example, for binary X , we have

$$\text{OR} = \frac{\left(\frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)} \right) / \left(1 - \frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)} \right)}{\left(\frac{\exp \beta_0}{1 + \exp \beta_0} \right) / \left(1 - \frac{\exp \beta_0}{1 + \exp \beta_0} \right)} =$$

Logistic regression

- Logistic regression is a natural choice for a *binary response*
- Denote one of the 2 possibilities 'success', or $Y = 1$
- We look for a model for estimating the *probability of success* as a function of the explanatory variables
- When using the *logit* transformation, la probabilité of 'success' is of the form :

$$E[Y | x] = P(Y = 1 | x) = \frac{e^\eta}{1 + e^\eta}$$



Logistic modeling of horseshoe crab data : results 1

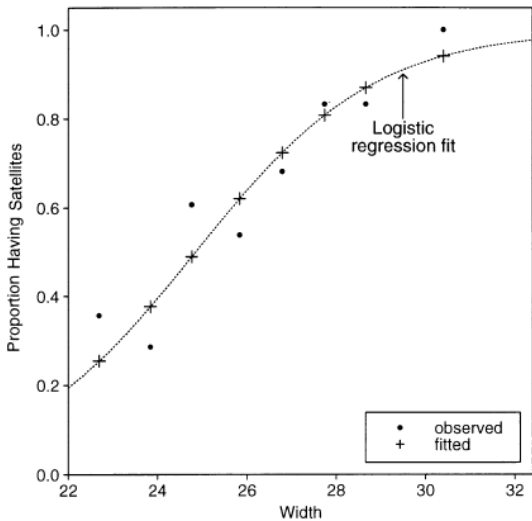


Figure 4.3. Observed and fitted proportions of satellites, by width of female crab.

Logistic modeling of horseshoe crab data : results 2

Table 4.2. Computer Output for Logistic Regression Model with Horseshoe Crab Data

	Log Likelihood				-97.2263	
		Standard Error	Likelihood Ratio	Wald		
Parameter	Estimate	Error	95% Conf. Limits	Chi-Sq	Pr > ChiSq	
Intercept	-12.3508	2.6287	-17.8097 -7.4573	22.07	<.0001	
width	0.4972	0.1017	0.3084 0.7090	23.89	<.0001	

- Now let's estimate $\pi(x)$ = probability (depending on x) of a female crab having a satellite
- Based on the output and the inverse logit function, we have :

$$\pi(x) = \frac{\exp(-12.351 + 0.497 \times x)}{1 + \exp(-12.351 + 0.497 \times x)}$$

- For the minimum sample value (21.0cm), $\pi(x) =$ _____
- For the maximum sample value (33.5cm), $\pi(x) =$

Odds and the OR

- For a probability p , the *odds* is defined as :

$$\text{odds}(p) = \frac{p}{1-p}$$

- For just one *binary* variable X , the *odds ratio (OR)* is the ratio of the odds :

$$OR = \frac{P(Y = 1 | X = 1)/(1 - P(Y = 1 | X = 1))}{P(Y = 1 | X = 0)/(1 - P(Y = 1 | X = 0))}$$

- 3 cases :
 - $OR = 1$: Y is independent of X
 - $OR > 1$: the condition represented by Y is more frequent for individuals with $X = 1$
 - $OR < 1$: the condition represented by Y is more frequent for individuals with $X = 0$

Analogous to linear regression

- The logit function g possesses many of the *same good properties* of the linear regression model
- Mathematically convenient and *flexible* – can include covariates in the model
- Can meaningfully interpret parameters
- **Linear in the parameters**
- A *difference* : Error distribution is *binomial* (not Normal)

Model fitting

- For linear regression, typically fitting is done by the method of *least squares*
- But when *the response is binary*, the 'good' statistical properties of the resulting estimators no longer hold
- The general method that leads us to least squares (for normally distributed errors) is our friend (!!) *maximum likelihood*

Revision : binomial distribution

- Logistic regression is related to the *binomial distribution*
- If there are multiple observations with the same value(s) of the explanatory variable(s), then the individual responses can be added and this sum has a binomial distribution
- Binomial mass function : $P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$
- For a binomial RV with parameters n and p , then the expected value is $\mu = np$ and the variance is $\sigma^2 = np(1 - p)$
- Logistic regression belongs to the 'binomial family' of GLMs

Maximum likelihood estimation

- Likelihood : $f(x_i) \propto \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i}$
- For independent observations, the likelihood is :
$$L(\beta) = \prod_{i=1}^n f(x_i)$$
- log likelihood :
$$l(\beta) = \log[L(\beta)] = \sum_{i=1}^n (\log(\pi(x_i)) + (1 - y_i) \log(1 - \pi(x_i)))$$
- Find the β_i that maximize the log likelihood by differentiating with respect to each β_i and setting all derivatives = 0
- For *linear regression*, these equations are *simple to solve*
- On the other hand, for *logistic regression* the equations are *nonlinear* and *do not have an analytic solution*
- They are solved using a *numerical algorithm* (notably Newton-Raphson)

Confidence intervals

- From the estimated parameters $\hat{\beta}_i^{MLE}$, we obtain the MLE of the linear predictor :

$$\hat{\eta}_{MLE} = \hat{\beta}_0^{MLE} + \sum_{i=1}^p \hat{\beta}_i^{MLE} x_i$$

- In addition, due to the invariance of the MLE, we obtain the MLE of the probability of 'success' :

$$\widehat{\pi(x)} = \frac{e^{\hat{\eta}}}{1 + e^{\hat{\eta}}}$$

- We use the asymptotic normality of the MLE in order to make a CI at $100(1 - \alpha)\%$ for $\eta : \hat{\eta} \pm z_{1-\alpha/2} \times ES(\hat{\eta}) = (J, S)$
- The $100(1 - \alpha)\%$ CI for $\pi(x)$ is thus : $\left(\frac{e^J}{1 + e^J}, \frac{e^S}{1 + e^S} \right)$

BREAK

Model fitting and checking

- For the standard (*fixed effects*) linear model, estimation is usually by *least squares*
- Can be more complicated with *random effects* or when x -variables are subject to measurement error as well
- Checking model : examination of *residuals*
 - Normality
 - Time effects
 - Nonconstant variance
 - Curvature
- Detection of *influential observations*

Linear regression model (again)

- Linear model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

- Another way to write this :

$$Y \sim N(\mu, \sigma^2), \quad \mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

- Suitable for a *continuous* response
- **NOT** suitable for a *binary* response
- **NOT** suitable for a *count* data

Modified model

- Instead of modeling the response directly, could instead model some *function* of the response
- i.e., Instead of modeling the expected response *directly* as a linear model, model a *suitable transformation*
- For binary data, it is convenient to use the *logit* function
- For count data, this is often taken to be the *log* transformation

Modified model for binary data

- Instead of modeling the 0/1 response directly, could instead model the *probability* of '1'
- Problems :
 - could lead to fitted values outside of $[0, 1]$
 - normality assumption on errors is wrong
- Instead of modeling the expected response *directly* as a linear function of the predictors, model a *suitable transformation*
- For binary data, this is generally taken to be the *logit* (or *logistic*) transformation

Logit transformation

- $\text{logit}(p) = \log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$
- Therefore,

$$p(x_1, \dots, x_k) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)}$$

- The parameter β_k is such that $\exp(\beta_k)$ is the *odds ratio* that the response takes value 1 when x_k increases by one, when the remaining variables are constant
- *i.e.* β_k is a *log-odds ratio* (log OR)
- Estimate parameters by *maximum likelihood* rather than least squares

Generalized linear model

- In a standard linear model, the *response variable* is modeled as a *normally distributed*
- However, if the response variable is *dichotomous* or a *count*, it does not make sense to model the outcome as normal
- Generalized linear models (GLMs) are an extension of linear models to model non-normal response variables
- A GLM consists of three components :
 - A *random component*, specifying the conditional distribution of the response variable, Y_i , given the values of the explanatory variables in the model
 - A *linear predictor*
 - A smooth and invertible linearizing *link function*
- We consider *logistic regression* for a count response
- We can consider *Poisson regression* for a count response

Generalized linear models : some theory

- Allows unified treatment of statistical methods for several important classes of models
- Response Y assumed to have *exponential family distribution* :

$$f(y) = \exp[a(y)b(\theta) + c(\theta) + d(y)]$$

- For a standard linear model

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \epsilon, \text{ with } \epsilon \sim N(0, \sigma^2)$$

- The *expected response* is $E[Y | x] = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$
- Let η denote the *linear predictor* $\eta = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$
- For a standard linear model, $E[Y | x] = \eta$
- In a *generalized linear model*, there is a *link function* g between η and the expected response :

$$g(E[Y | x]) = \eta$$

- For a standard linear model, $g(y) = y$ (*identity link*)

Link function

- When the response variable is binary (with values coded as 0 or 1), then $E[Y | x] = P(Y = 1 | x)$
- A convenient function in this case is

$$E[Y | x] = P(Y = 1 | x) = \frac{e^\eta}{1 + e^\eta}$$

- The corresponding link function (inverse of this function) is called the *logit*
- $\text{logit}(x) = \log \frac{x}{1-x}$
- Regression using this model is called *logistic regression*

Link function : examples

Link	Family Name				
	binomial	Gamma	gaussian	inverse.gaussian	poisson
logit	D				
probit	•				
cloglog	•				
identity		•	D		•
inverse		D			
log		•			D
1/mu ²				D	
sqrt					•

Analogous to linear regression

- The logit function g has many of the desirable properties of a linear regression model :
 - Mathematically convenient and flexible
 - Can meaningfully interpret parameters
 - Linear in the parameters
- A difference : Error distribution is binomial (**not** normal)

Inference : tests for coefficients

- *Wald test* statistics are simple ; for 'sufficiently large' samples :

$$z = \frac{\hat{\beta}}{SE(\hat{\beta})} \sim N(0, 1)$$

- Although the Wald test is adequate for large samples, the *likelihood ratio test (LRT)* is more powerful and more reliable for sample sizes often used in practice
- The LRT test statistic compares the maximum L_H of the likelihood function when $\beta = 0$ to the maximum L_A of the likelihood function for unrestricted β :

$$\lambda = -2 \log \frac{L(\hat{\theta}_{MLE}^H)}{L(\hat{\theta}_{MLE}^A)},$$

- Under certain regularity conditions, when H is true $\lambda \sim \chi_p^2$, where $p =$ number of constraints imposed by H (= difference in the number of parameters estimated under the 2 models)

Inference : CI for probabilities

- For simple logistic regression, the estimated (predicted) probability at a fixed x value is given by :

$$P(Y = 1 | x) = \hat{\pi}(x) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x}}$$

- **Activity** : Estimate the probability of a satellite for female crabs of width $x = 26.5\text{cm}$...
- From software, a 95% CI for the true probability $\pi(26.5)$ is (0.61, 0.77)

Why use a model to estimate probabilities ?

- Instead of finding $\hat{\pi}(x)$ using the model fit, as we just did at $x = 26.5$, why not simply use the sample proportion to estimate the probability ??
- For width = 26.5, 4/6 had satellites, so the sample proportion estimate at $x = 26.5$ is $p = 4/6 = 0.67$ (similar to the model-based estimate)
- A small sample exact (binomial) 95% CI is (0.22, 0.96) : much larger than the model-based CI
- When the logistic regression model holds, the model-based estimator of $\hat{\pi}(x)$ is *much better* than that of the sample proportion because it uses *all the data* rather than *only* the data at the fixed x value, giving a more precise estimate
- For example, at $x = 26.5$, software reports a SE = 0.04 for the model-based estimate 0.695
- By contrast, the SE for the sample proportion of 0.67 with only six observations is : _____

Indicator (dummy) predictors

- Let's go back to analyzing our Horseshoe crab data, but instead of only using carapace width as a predictor, let's also include color.
- Color is a *categorical* (factor) variable with five categories : light, medium light, medium, medium dark, dark
- Color is a surrogate for age, since older crabs tending to have darker shells
- The sample contained no light crabs, so we use only the other four categories
- In order to include categorical / factor explanatory variables in a LM or GLM, we need to use *indicator* (sometimes called *dummy*) variables
- The number of dummy variables to include is the number of categories minus 1

Multiple logistic regression

- To incorporate color into the model, we need to introduce 3 indicator variables for the 4 categories
- The model is now

$$\text{logit}[P(Y = 1)] = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \beta_3 c_3 + \beta_4 x$$

where x denotes width and

$c_1 = 1$ for color = medium light, 0 otherwise

$c_2 = 1$ for color = medium, 0 otherwise

$c_3 = 1$ for color = medium dark, 0 otherwise

- Crab color is dark when $c_1 = c_2 = c_3 = 0$

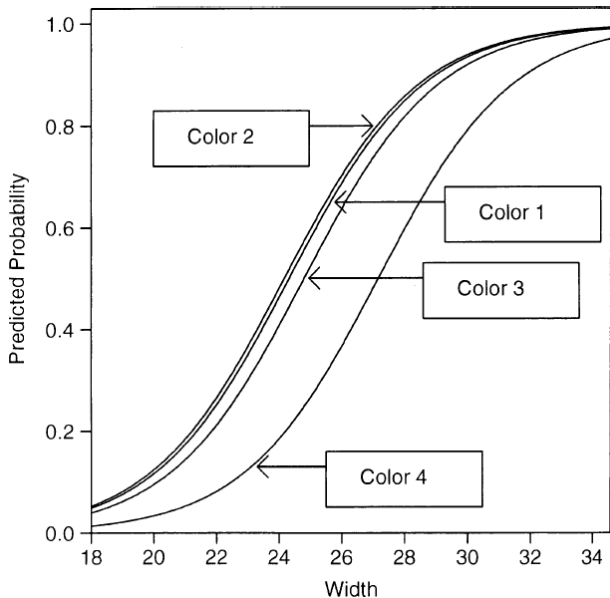
Multiple logistic modeling with width and color : results 1

Table 4.6. Computer Output for Model for Horseshoe Crabs with Width and Color Predictors

Parameter	Estimate	Std. Error	Like. Ratio Confidence	95% Limits	Chi Square	Pr > ChiSq
intercept	-12.7151	2.7618	-18.4564	-7.5788	21.20	<.0001
c1	1.3299	0.8525	-0.2738	3.1354	2.43	0.1188
c2	1.4023	0.5484	0.3527	2.5260	6.54	0.0106
c3	1.1061	0.5921	-0.0279	2.3138	3.49	0.0617
width	0.4680	0.1055	0.2713	0.6870	19.66	<.0001

LR Statistics			
Source	DF	Chi-Square	Pr > ChiSq
width	1	24.60	<.0001
color	3	7.00	0.0720

Multiple logistic modeling with width and color : results 2



Some interpretation

- The model assumes *no interaction* between color and width \Rightarrow width has the same effect (coefficient 0.468) for all colors
- This implies that the shapes of the four curves relating width to $P(Y = 1)$ (for the four colors) are identical
- For each color, a 1 cm increase in width has a multiplicative effect of $e^{0.468} = 1.60$ on the odds that $Y = 1$
- Each curve is the same as any other curve, only shifted to the left or right
- The parallelism of curves in the horizontal dimension implies that two curves never cross
- At all width values, for example, color 4 (dark) has a lower estimated probability of a satellite than the other colors

Let's have some fun !!

- What is the estimated probability for a medium-light crab of average width (26.3 cm) ?? for a dark crab ??
- What are the estimated odds for a medium-light crab ?? for a dark crab ??
- The exponentiated difference between two color parameter estimates is an odds ratio comparing those colors. What is the estimated odds ratio comparing medium-light and dark crabs ?? Interpret.

Evaluation of the fitted model

- In linear regression, ANOVA consists in the decomposition of the total sum of squares of the observations around their mean (SST) :
 - SSE , error sum of squares (residuals = observed - predicted)
 - SSR , regression sum of squares (of the model)
- Large values of SSR suggest the importance of the explanatory variable(s)
- We use the *principle* for logistic regression : comparison of the observed response to the predicted response by the models with//without the explanatory variable(s)
- This comparison is made based on the *log likelihood*

Deviance

- For (ordinary) linear models, parameter estimation by least squares (minimize the sum of squared residuals)
- (Equivalent to ML for the Normal model)
- For GLMs, estimation is by ML
- The *deviance* is (proportional to) $2 \times \ell$
- (Analogous to SSE)
- Obtaining an 'absolute' measure of the quality of model fit (goodness-of-fit) depends on certain assumptions, often not satisfied in practice
- Thus typically focus rather on the *comparison* of competing models
- If the models are *nested* (that is, one model is a sub-model of the other), we can carry out a LRT

Test of goodness-of-fit ('global' test)

- Or rather test of NONgoodness-of-fit (!!)
- Test based on the deviance D of the model
- We reject H : the data conform to the model, for *large values of D (residuals)*
- Under A , there is a parameter for each observation (*saturated model*)
- It is often said – **BUT NOT TRUE!!!!** – that under H , $D(\text{residuals}) \sim \chi^2$ with $\text{df} = \text{df error}$
- (The problem : the asymptotic result for χ^2 does not hold if the number of parameters is not finite, and since the saturated model has one parameter for each of the n observations, then if $n \rightarrow \infty$ the number of parameters is not finite)
- For samples of moderate size, it is not the worst thing in the world to assume this asymptotic distribution

Model comparison

- Linear regression : a coefficient is (statistically) significant if its standardized value $\hat{\beta}/SE(\hat{\beta})$ is 'large'
- We can use this same reasoning for logistic regression (z-test = *Wald test*), but this approach is problematic (lacks power)
- Preferred approach : *likelihood ratio test (LRT)*

- Deviance $D = -2 \left(\sum_{i=1}^n y_i \log \left(\frac{\hat{p}_i}{y_i} \right) + (1 - y_i) \log \left(\frac{1 - \hat{p}_i}{1 - y_i} \right) \right)$

- Comparison of models : calculate the statistic

$$G^2 = D(\text{sub-model}) - D(\text{bigger model})$$

- Under H (the sub-model is sufficient), $G^2 \sim \chi^2$ with degrees of freedom (df) = difference in the number of estimated parameters

Summary : Tests for coefficients

■ One coefficient :

- 1 parameter = β_i , the coefficient of variable x_i in the logistic regression model in the population
- 2 $H : \beta_i = 0$; $A : \beta_i \neq 0$
- 3 TS : • Wald : $z_{obs} = \frac{\hat{\beta}_i}{ES(\hat{\beta}_i)}$ • LRT : $G^2 = -2 \log \frac{L_H}{L_A}$
- 4 p_{obs} : • Wald : $2P(Z > | z_{1-\alpha/2} |)$ • LRT : $P(X^2 > \chi_1^2)$

■ Several coefficients :

- 1 parameters = β_j, \dots, β_k (= q coefficients), of variables x_j, \dots, β_k in the logistic regression model in the population
- 2 $H : \beta_j = \dots = \beta_k = 0$; $A : \text{at least one } \beta_i \neq 0, q \leq i \leq k$
- 3 TS : • LRT : $G^2 = -2 \log \frac{L_H}{L_A}$
- 4 p_{obs} : • LRT : $P(X^2 > \chi_q^2)$

- (Here, we consider the RV $X^2 \sim \chi^2$)

PAUSE

DNA sequencing (optional)

- (Automated) Sanger sequencing
 - ‘first-generation’ technology
 - F. Sanger, 1977
- Process :
 - bacterial cloning or PCR
 - template purification
 - labelling of DNA fragments using the chain termination method with energy transfer, dye-labelled dideoxynucleotides and a DNA polymerase
 - capillary electrophoresis
 - fluorescence detection
- Data : four-colour plots that reveal the DNA sequence

Next-generation sequencing

- Several newer sequencing technologies
 - ‘Next-generation sequencing’ (NGS data)
 - ‘Ultra high-throughput sequencing’ (UHTS data)
- These newer technologies use various strategies that rely on a combination of template preparation, sequencing and imaging, and genome alignment and assembly methods
- Data : four-colour plots that reveal the DNA sequence
- Major advance : ability to produce a *large amount* of data relatively *cheaply*
- Expands experimental possibilities beyond just determining the order of bases

Applications of NGS

- Sequence assembly (original application)
- Resequencing : The sequencing of part of an individual's genome in order to detect sequence differences between the individual and the standard genome of the species
- Gene expression : RNA-Seq
- SNP discovery and genotyping
- Variant discovery and quantification
- Transcription factor binding sites : ChIP-Seq
- Measuring DNA methylation

NGS data generation

- Sequencing technologies incorporate methods that we can class as
 - template preparation
 - sequencing and imaging
 - data analysis
- Combination of specific protocols distinguishes different technologies
- Major technologies :
 - Illumina HiSeq (older : Solexa)
 - 454 (Roche)
 - Applied Biosciences SOLiD
 - Pacific Biosciences SMRT (single molecule real-time)

Data analysis pipeline

- Data are *counts* of short sequences (called 'reads')
- Quality control of data
- Match to reference sequence, read mapping
- Count/summarize number of reads per feature
- Statistical analysis (depends on the specific application)

Sequence data

- Sequence data are *counts*
- DNA sample \implies *population of cDNA fragments*
- Each genomic feature \implies species for which the population size is to be estimated
- Sequencing a DNA sample \implies random sampling of each of these species
- *Aim* : to estimate the relative abundance of each species in the population

Poisson model

- If we assume :
 - each cDNA fragment has the *same chance* of being selected for sequencing
 - the fragments are selected independently
- Then : the number of read counts for a given genomic feature should follow a *Poisson variation law* across repeated sequence runs of the same cDNA sample
- The Poisson model implies that the *mean equals the variance*
- (This relationship has been validated in an early RNA-Seq study using the same initial source of RNA distributed across multiple lanes of an Illumina GA sequencer)

Single gene model

- DNA sample \implies 'library'
- Contains genes $1, \dots, g, \dots$
- For a given gene g in library i , Y_{gi} = number of reads for gene g in library i
- $Y_{gi} \sim \text{Bin}(M, p_{gi})$, where p_{gi} is the proportion of the total number of sequences M in library i that are gene g
- M large, p_{gi} small $\implies Y_{gi} \sim \text{Pois}(\mu_{gi} = Mp_{gi})$
(approximately)

Technical vs. biological replicates

- For the Poisson model, the *variance* is equal to the *mean*
- With *technical replicates*, this relation holds fairly well
- With *biological replicates*, the variance is typically *larger* than expected using the Poisson model
- There are a few different approaches for accounting for this additional variability (overdispersion)

Link function for count data

- We can model the count data $Y_i \sim \text{Pois}(\mu_i)$, $i = 1, \dots, n$
- Want to relate the mean μ_i to one or more *covariates* (for example, treatment/control status)
- A convenient link function in this case is the log :

$$\log \mu_i = \eta = x_i^T \beta$$

- Using a log link ensures that the fitted values of μ_i will remain in the parameter space $[0, \infty)$
- A Poisson model with a log link is sometimes called a *log-linear model*

Variance function for the Poisson model

- The Poisson distributions are a discrete family with probability function indexed by the rate parameter $\mu > 0$:

$$p(y) = \frac{e^{-\mu} \mu^y}{y!}, \quad y = 0, 1, 2, \dots$$

- Under the Poisson model : $E[Y_i] = \text{Var}(Y_i) = \mu_i$
- General form of the relationship between the variance of the response variable and its mean is : $\text{Var}(\text{response}) = \phi V(\mu)$, with ϕ a constant scale factor
 - **Normal** : $V(\mu) = 1$, $\phi = \sigma^2$ (the variance does not depend on the mean)
 - **Binomial** : $V(\mu) = \mu(1 - \mu)$ $\phi = 1$
 - **Poisson** : $V(\mu) = \mu$ $\phi = 1$
- Real data are often *overdispersed*, exhibiting more variation than allowed by the Poisson model

Detecting and handling overdispersion

- When fitting a GLM with binomial or Poisson errors, can often detect overdispersion by *comparing the residual deviance to its degrees of freedom*
- For a well-fitting model, these should be approximately equal
- Overdispersion usually handled with an alternative model :
 - **Quasi-Poisson Model** : Assume $Var(Y_i) = \phi \mu_i$ and estimating the *scale parameter* ϕ
 - *Zero-Inflated Poisson Model* : for modeling the case when there are too many '0' values
 - *Negative Binomial Model* : Can arise from a two-stage model :

$$Y_i \sim Pois(\mu_i^*) \quad \mu_i^* \sim \Gamma(\mu_i/\omega, \omega)$$

Then $Y_i \sim NegBin$, with $E[Y_i] = \mu_i$ and $Var(Y_i) = \mu_i + \mu_i^2/\omega$

Differential gene expression for NGS data

- Several BioConductor (R) packages for identifying differential expression from NGS data
- These mostly use the negative binomial model, since the counts are typically over-dispersed compared to the Poisson model
- The [edgeR](#) package uses an overdispersed Poisson model to account for both biological and technical variability, and uses empirical Bayes methods to moderate the degree of overdispersion across transcripts

Assessing model fit

- In linear regression, an anova table partitions SST , the total sum of squared deviations of observations about their mean, into two parts :
 - SSE , or residual (observed - predicted) sum of squares
 - SSR , or regression sum of squares
- Large SSR suggests the explanatory variable(s) is(are) important
- In linear regression, diagnostics are built around residuals and SSR
- For GLMs, there are a few different kinds of residuals :
Pearson residuals and *deviance* residuals
- Pearson residual for an observation is obtained by subtracting the mean (predicted value) for that observation and dividing by the (estimated) SD
- Deviance residuals are based on the contribution of each point to the likelihood

Deviance

- In standard linear models, estimate parameters by minimizing residual sum of squares
- (Equivalent to ML for normal model)
- In GLM, estimate parameters by ML
- The *deviance* is (proportional to) $2 \times l$
- (Analogous to SSE)
- Obtaining 'absolute' measure of goodness of fit depends on some assumptions that may not be satisfied in practice
- Usually focus on comparing competing models
- When the models are *nested*, can carry out likelihood ratio test

Comparing models

- In linear regression, consider coefficient significant if (squared) standardized value $\hat{\beta}/SE(\hat{\beta})$ is 'large'
- Can also do this for logistic regression (Wald test), but there are some problems with it
- Preferred approach : likelihood ratio test
- Deviance $D = -2 \sum_{i=1}^n y_i \log \left(\frac{\hat{p}_i}{y_i} \right) + (1 - y_i) \log \left(\frac{1 - \hat{p}_i}{1 - y_i} \right)$
- To compare models, compute $G = D(\text{submodel}) - D(\text{bigger model})$
- Under the null (*i.e.* the submodel), $G \sim \chi^2$ with df = difference in the number of estimated parameters

Variance inflation factors

- The meaning of a variance inflation factor is essentially equivalent for linear models and GLMs
- We can use the VIF to look for multicollinearity
- R function `vif` from the `car` package
- Also look at correlation matrix for the data matrix X

Summary

- Residuals are certainly less informative for GLMs than for linear regression
- Issues of outliers and influential observations just as relevant for GLMs as for linear regression : look at Cook's distance plot
- Usually a good idea to *start with simple models* and gradually add in complexity