Regression Analysis: Linear, Logistic and Mixed Models

Cynthia Lord

Florida Medical Entomology Lab

University of Florida

ESA 2010: MUVE Section Symposium
We are confronted by insurmountable opportunities: Novel statistics for entomologists

Regression analysis

- Predict outcome (dependent variable) from one or more independent variables
- Implies causality
- Used to explore relationships and assess contributions
- Models developed for explicit prediction

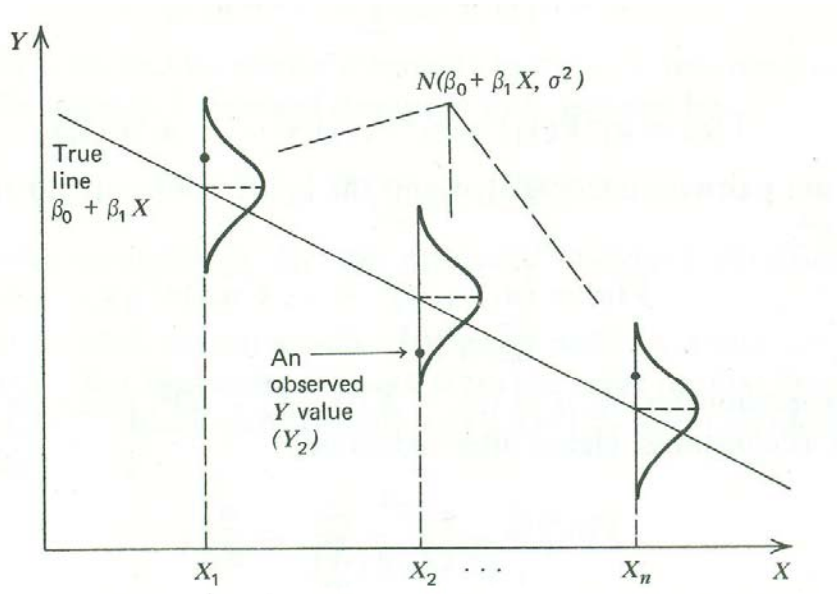
Familiar “types”

- Multiple regression
- Logistic regression
- Stepwise regression
- GLM: generalized linear models
- GLMM: generalized linear mixed models

"Traditional" Linear regression

- Dependent and independent variables
  - Numerical: continuous or ordinal
  - Fixed effects model
    - All levels /entire region of interest

    Included in x range
- Normal distribution for errors
- Single dependent variable
- Can be simple, single variable
  - $Y=b_0 + b_1x + \varepsilon$
- Or complex
  - $Y=b_0 + b_1x_1 + b_2x_2 +\ldots+ b_{ij}x_i*x_j + \varepsilon$
  - Multivariate, interactions



“Traditional”: Criteria for fitting model

- Hypothesis testing approach
  - F test, R^2
  - Are coefficients significantly different from 0?
  - Does variable make a significant contribution to the model?

Model Selection

- Main effects
- Interactions to specified level
  - Power issues
- Full model
- Stepwise selection

Only a single model is ultimately considered

GLM

- Generalized linear models
  - Extensions of fixed effects linear models
  - Used where standard assumptions are violated
    - Normal distributions
  - Use a link function to link a linear model to the mean of Y
    - Based on distribution of Y
  - Common distribution
    - Binomial: logit link

Logistic regression

- Form of dependent variable: Binary
  - Binomial distribution
  - 0 or 1, often coding for qualitative result
    - Infected or not
    - Present or not
- Probability that y=1
  - $p(y=1) = \pi$
- Logit link
  - $\text{Log}(\pi / 1 - \pi) = b_0 + b_1 x_1 + b_2 x_2 + \ldots$

Logistic - interpretation

- Probability of a "1" response given predictor variables
- Expressed as likelihood of an event
- Odds ratios (or log odds ratios)
  - Odds of a "1" response with one X over another
  - More common with levels of X rather than continuous

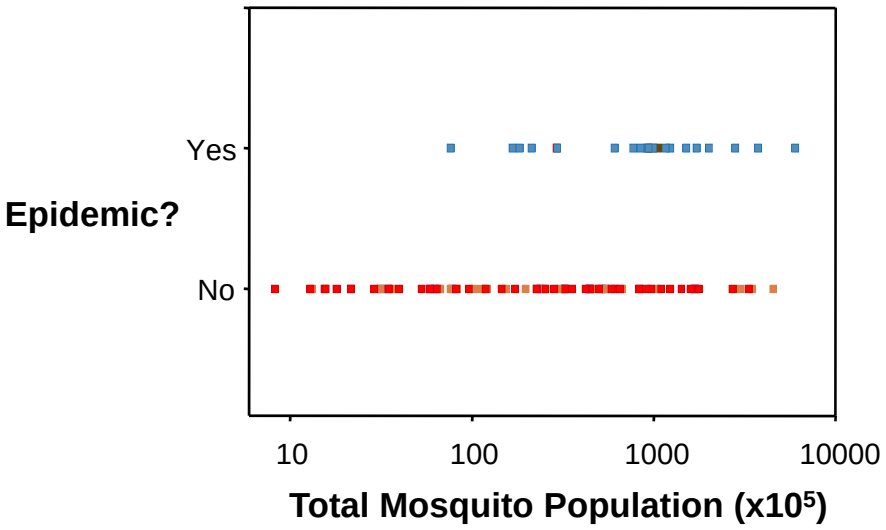
Example: logistic stepwise regression

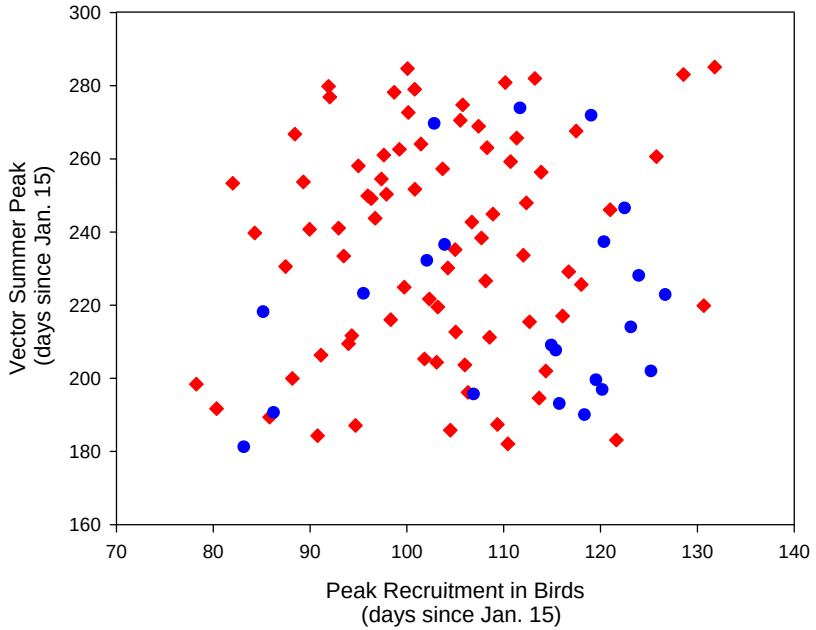
- Data from simulation model: St. Louis Encephalitis virus
- Outcome (dependent variable): epidemic or not
  - Epidemic = 1
  - Not epidemic = 0
- 14 input parameters as independent variables
- Sampled from defined parameter space
  - Fixed effects
- Main effects only, no interactions
  - Power of data with n=100 and many independent variables

Example: Stepwise logistic regression

Logistic Model $R^2 = 0.34$

Mosquito population
Mosquito mortality (baseline)





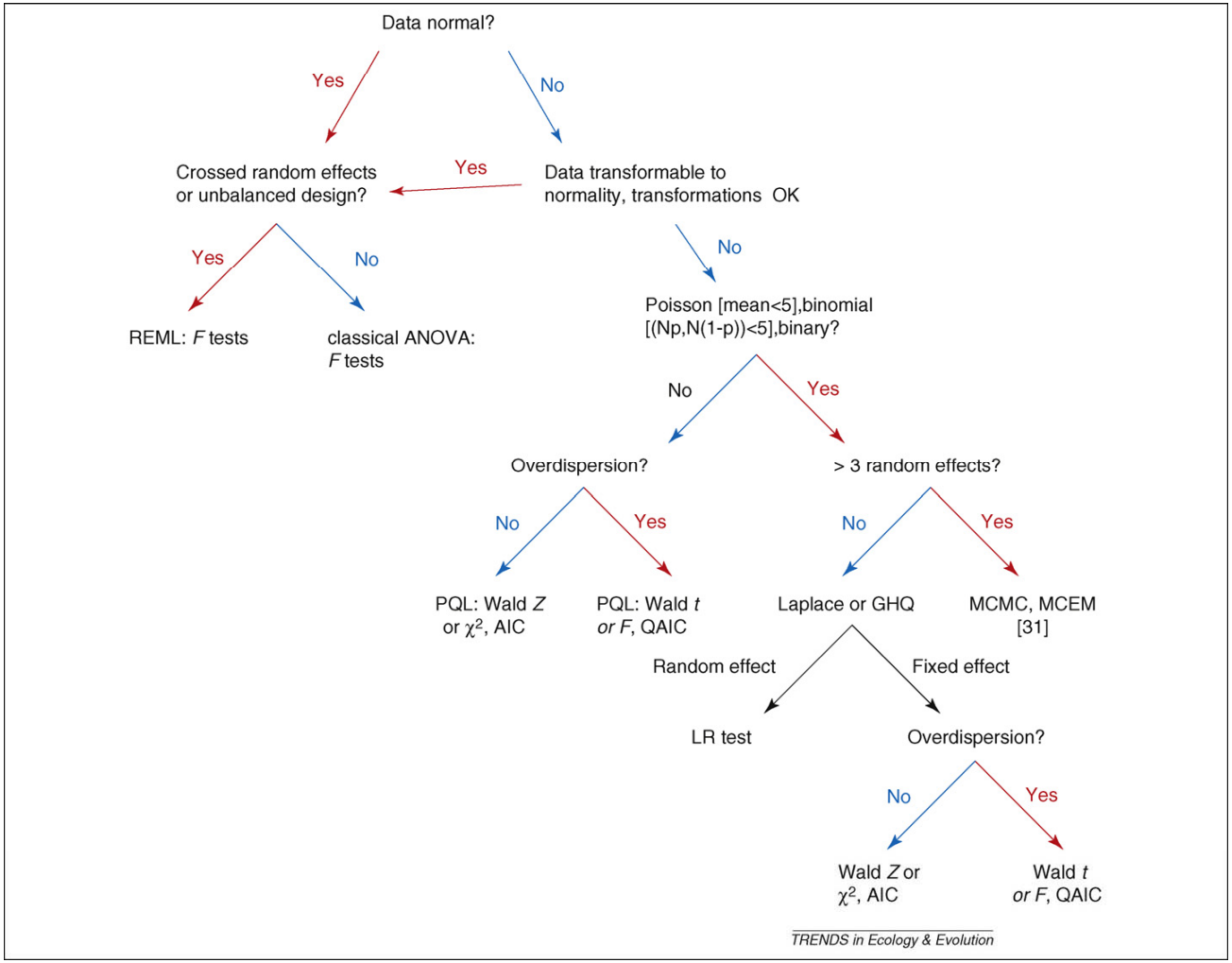
Time of peak bird recruitment
Time of summer peak in mosquitoes

Generalized Linear Mixed Models

- GLM
  - Fixed effects
- Random effects
  - Independent variable is considered random if its levels plausibly represent a larger population with a probability distribution

GLMM models

Depends on:
Distribution of variables
Design of study
Variance structure



Bolker et al. 2008.
TREE 24: 127-135

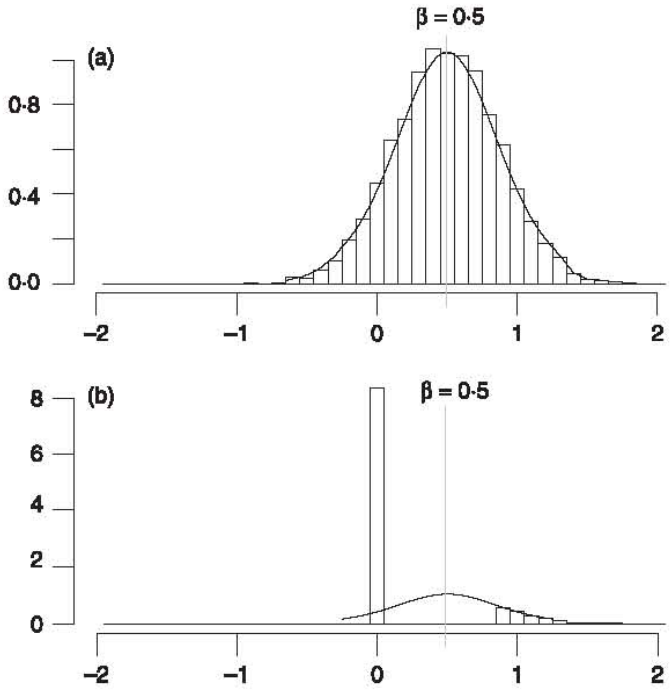
Bias in parameter estimation

Hypothesis testing: If coefficient is not significant, it is treated as 0

Simulation study:
Y=1 + 0.5*x +e
e~N(0,1)
n = 10
Regression model fit

Slope tested using hypothesis testing

Slope = 0 accepted if p<0.05



Distribution of slopes from models fit

Distribution of slopes from hypothesis testing model selection

Whittingham et al. 2006. J. Anim. Ecol. 75: 1182-1189

Stepwise regression

- Entry of independent variables sequentially
- Test for exit of entered variables
- Can be done using hypothesis testing or model selection methods
  - Choice of criteria used
- Penalty for increased complexity
- In hypothesis testing, modified by stringency for entry & exit

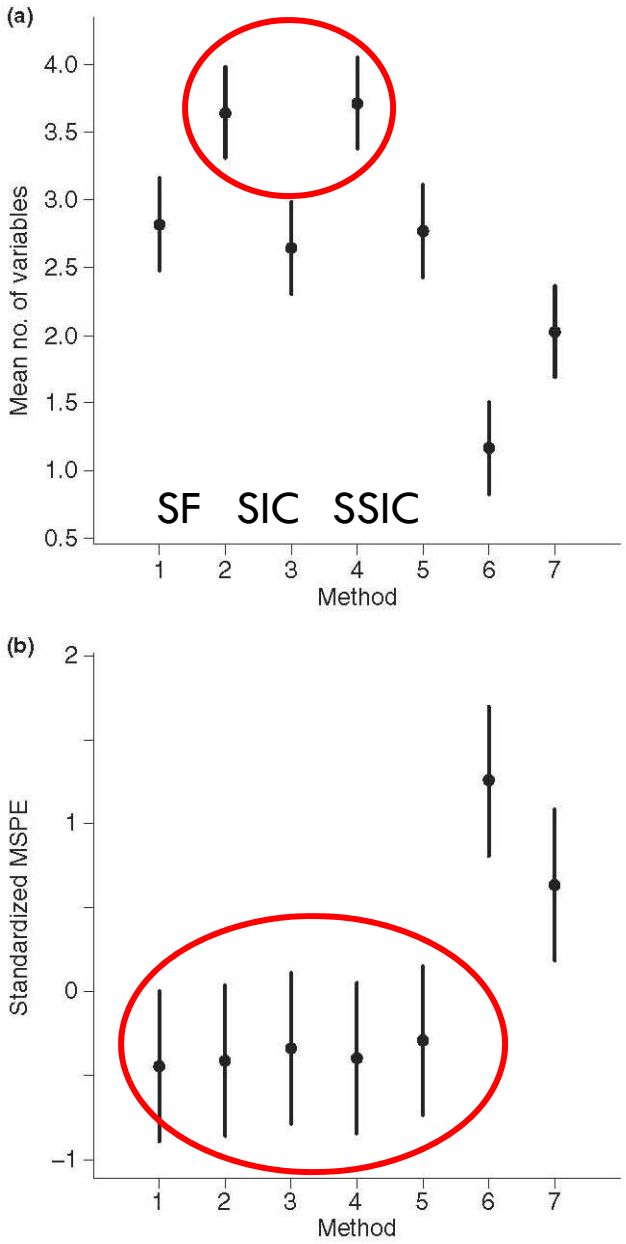
Issues in stepwise regression

- Hypothesis testing
  - Involves many sequential tests
  - Type I (false positive) errors inflated
  - Also affects distribution of F-statistic
  - Overall significance of the final model affected
- Fits single model with no assessment of whether others would have similar predictive ability

Variable selection methods

- Stepwise selection with hypothesis testing
- Stepwise selection with information criteria
  - AIC
  - BIC
- All subsets and examine for fit
- All penalize model for increased complexity
- Do methods bias towards increased complexity?

12 data sets
Reanalyzed using
1: Stepwise F test
2: Stepwise AIC
3: Stepwise BIC
4: All subset AIC
5: All subset BIC
6: Regression tree A
7: Regression tree B



Models produced by IC methods included more variables

Models produced by IC and stepwise methods had similar predictive value

Murtaugh. 2009. Ecology Letters 12:1061-1068

Issues with regression

- Model selection
  - How many models should be considered?
- Hypothesis testing or IC methods
- Multiple dependent variables?
- Biological realism and hypothesis generation

Statistics are a tool

Acknowledgements

- ESA and MUVE for funding and supporting the symposium
- Steve Juliano
- Jonathan Day

ESA Symposium References – Lord presentation

Bolker, B. M., M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J. S. White. 2008. Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol. Evol. 24:127-135.

Littell, R. C., G. A. Milliken, W. W. Stroup, R. D. Wolfinger, and O. Schabenberger. 2006. SAS for mixed models. SAS Institute Inc., Cary, NC.

Murtaugh, P. A. 2009. Performance of several variable-selection methods applied to real ecological data. Ecol. Lett. 12:1061-1068.

Scheiner, S.M. and J. Gurevitch (eds.). 2001. Design and analysis of ecological experiments. Oxford University Press, New York, NY.

Shoukri, M. M. and C. A. Pause. 1999. Statistical methods for health sciences. CRC Press LLC, Boca Raton, FL.

Snedecor, G. W. and W. G. Cochran. 1980. Statistical methods. The Iowa State University Press, Ames, Iowa.

Stephens, P. A., S. W. Buskirk, and C. Martínez del Rio. 2007. Inference in ecology and evolution. Trends Ecol. Evol. 22:192-197.

Whittingham, M. J., P. A. Stephens, R. B. Bradbury, and R. P. Freckleton. 2006. Why do we still use stepwise modelling in ecology and behaviour? J. Anim. Ecol. 75:1182-1189.