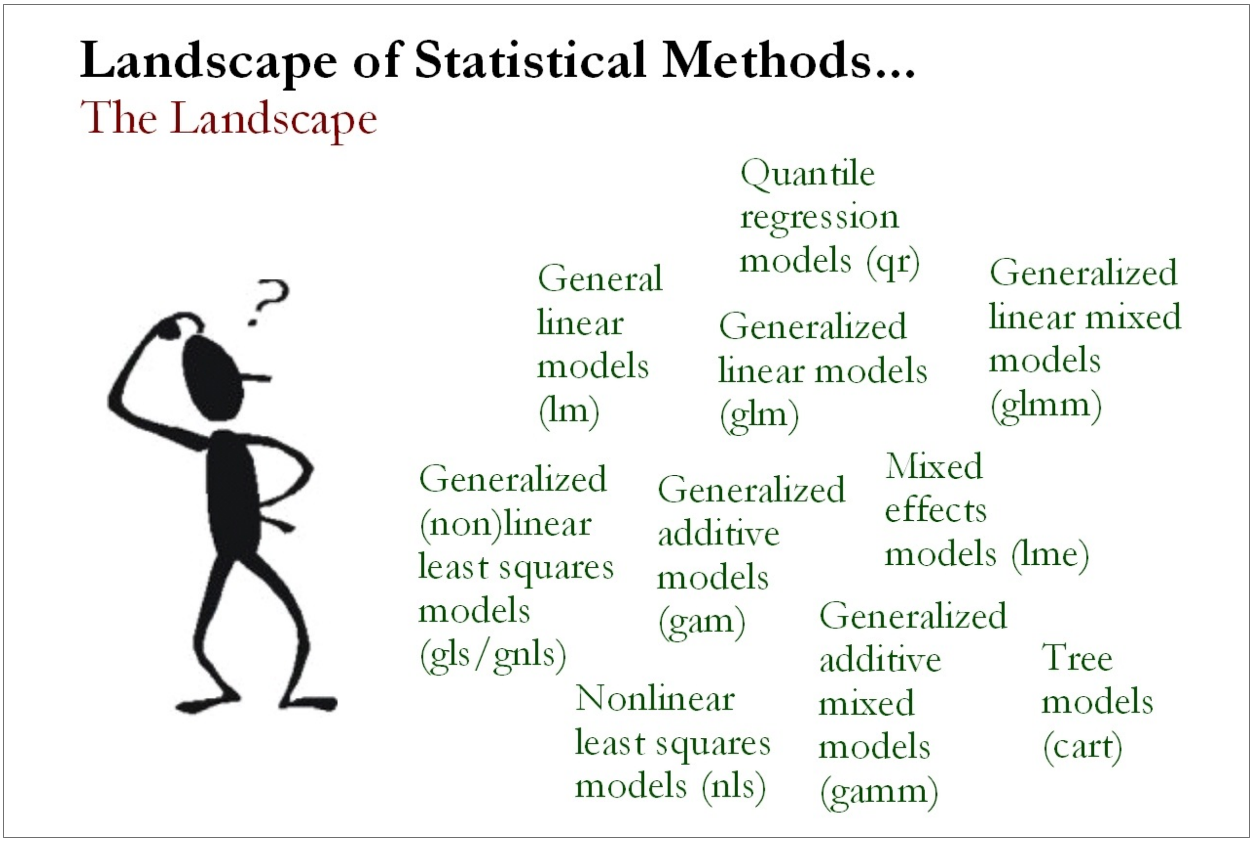
Design and Analysis of Ecological Data
Landscape of Statistical Methods: Part 1

*Much of the material in this section is taken from Bolker (2008) and Zur et al. (2009)



Landscape of Statistical Methods…

The Landscape

Quantile regression models (qr)

General linear models (lm)

Generalized linear models (glm)

Generalized linear mixed models (glmm)

Generalized (non)linear least squares models (gls/gnls)

Generalized additive models (gam)

Mixed effects models (lme)

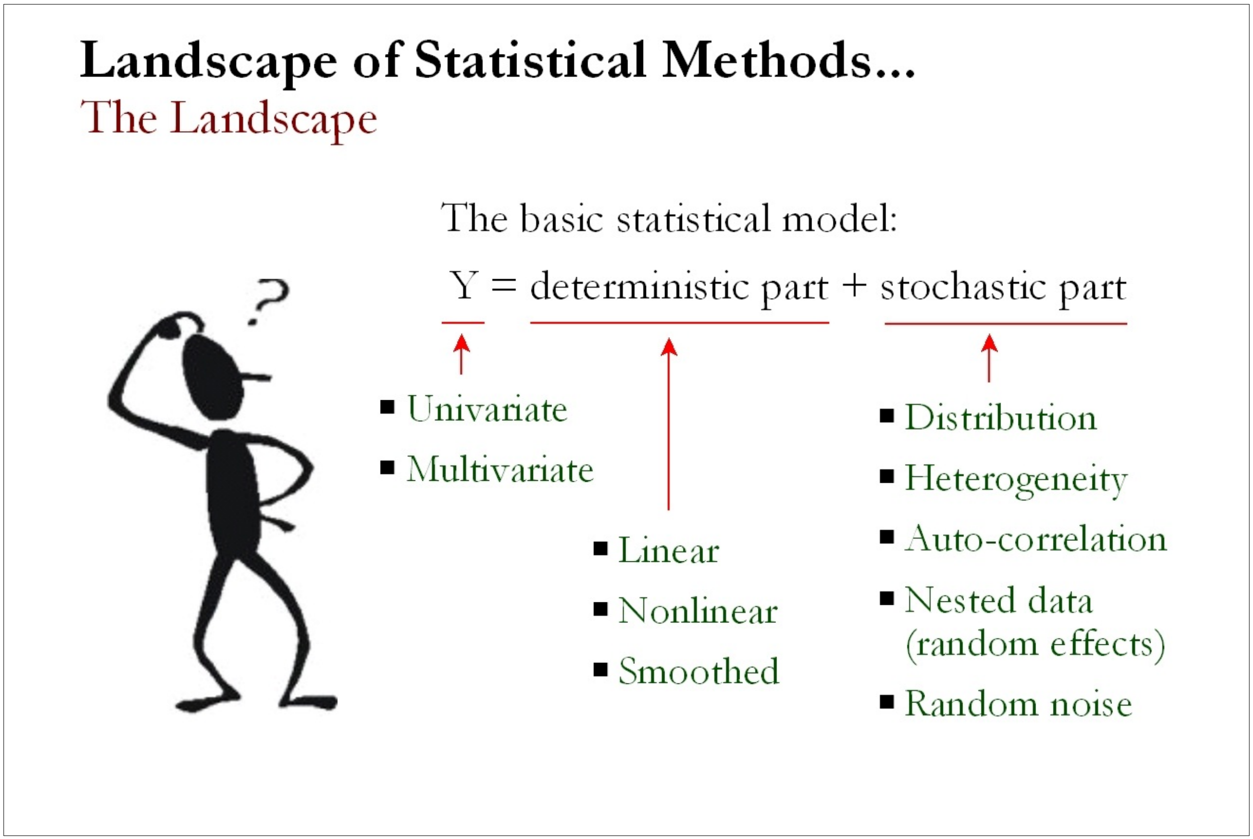
Nonlinear least squares models (nls)

Generalized additive mixed models (gamm)

Tree models (cart)

1. The landscape of statistical methods

The field of ecological modeling has grown amazingly complex over the years. There are now methods for tackling just about any problem. One of the greatest challenges in learning statistics is figuring out how the various methods relate to each other and determining which method is most appropriate for any particular problem. Unfortunately, the plethora of statistical methods defy simple classification. Instead of trying to fit methods into clearly defined boxes, it is easier and more meaningful to think about the factors that help distinguish among methods. In this final section, we will briefly review these factors with the aim of portraying the "landscape" of statistical methods in ecological modeling. Importantly, this treatment is not meant to be an exhaustive survey of statistical methods, as there are many other methods that we will not consider here because they are not commonly employed in ecology. In the end, the choice of a particular method and its interpretation will depend heavily on whether the purpose of the analysis is descriptive or inferential, the number and types of variables (i.e., dependent, independent, or interdependent) and the type of data (e.g., continuous, count, proportion, binary, time at death, time series, circular). We will not review these issues again here, but take the time to refresh your memory by rereading the first two chapters of this course.



Landscape of Statistical Methods...

The Landscape

The basic statistical model:

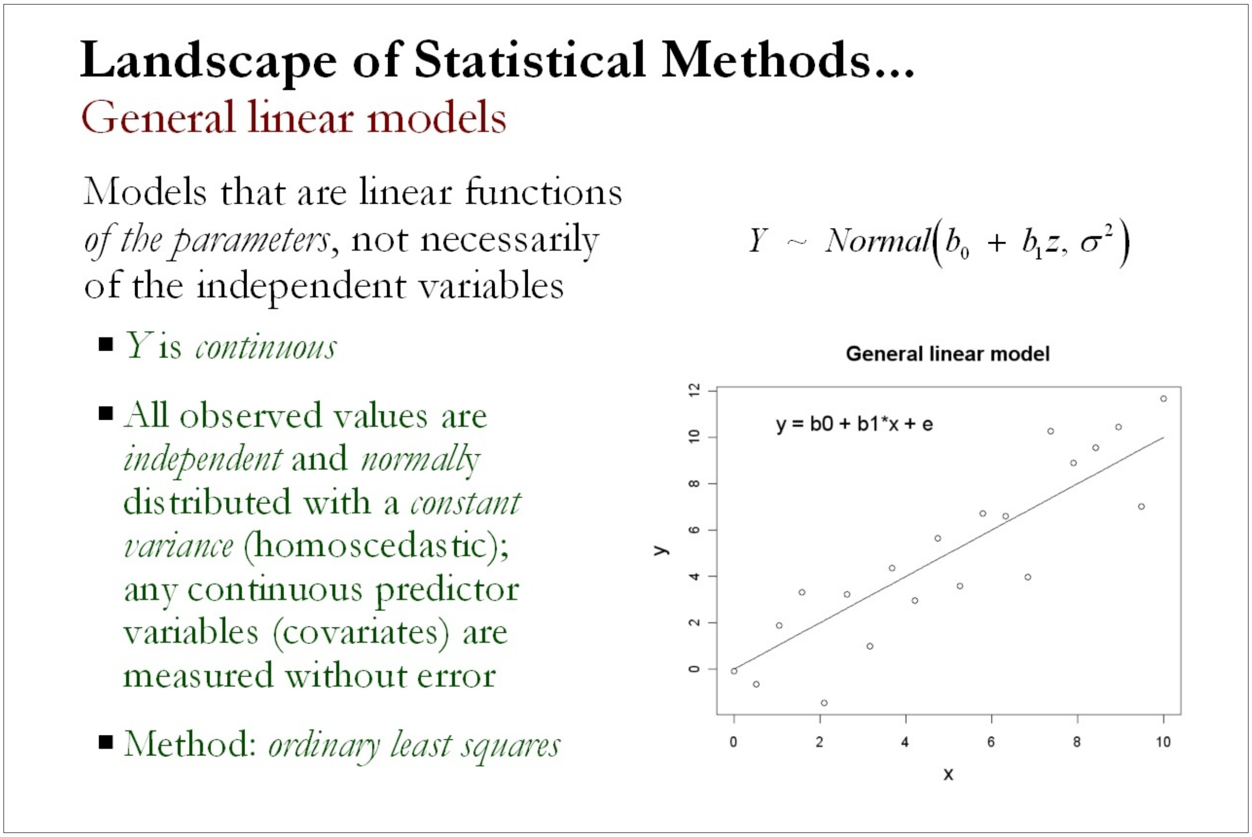
$$\underbrace{Y}_{} = \underbrace{\text{deterministic part}}_{} + \underbrace{\text{stochastic part}}_{}$$

- Univariate
- Multivariate

- Linear
- Nonlinear
- Smoothed

- Distribution
- Heterogeneity
- Auto-correlation
- Nested data (random effects)
- Random noise

The plethora of statistical methods available to ecologists derives from the fact that ecological data is complex – no single method can accommodate the myriad problems we encounter with ecological data. Recall from part one of this course that statistical models typically consist of two parts, a deterministic component and a stochastic (or error) component. Most of the differences among methods are due to differences in assumptions about either the response variable, the deterministic model, or the error model, and most are extensions or modification to the basic general linear model that expresses Y as linear function of X where all observed values are independent and normally distributed with a constant variance. For example, multivariate methods deal with models containing more than one response variables (or a single set of interdependent - presumed response - variables); additive models and nonlinear least squares models deal with nonlinear deterministic models; generalized linear models deal with nonlinear deterministic models in combination with nonnorma error distributions; generalized least squares models deal with heterogeneous (i.e., non-constant) errors; mixed effects models deal with spatially and/or temporally nested error structures; auto-regressive models deal with temporally and/or spatially correlated errors (non-independence); and so on. Non-parametric methods, such as tree-based models (e.g., classification and regression trees) and quantile regression, make no assumptions about the errors. Ultimately, there are almost as many methods as there are problems since analytical methods can be customized to almost any problem. Indeed, the basis for model-based statistics is that the problem drives the construction of the statistical model, the method of estimation is then determined by the statistical model in combination with a selected inference framework (e.g., frequentist versus Bayesian).



Landscape of Statistical Methods…

General linear models

Models that are linear functions *of the parameters*, not necessarily of the independent variables

$$Y \sim Normal\left(b_0 + b_1 z, \sigma^2\right)$$

- *Y* is *continuous*
- All observed values are *independent* and *normally* distributed with a *constant variance* (homoscedastic); any continuous predictor variables (covariates) are measured without error
- Method: *ordinary least squares*

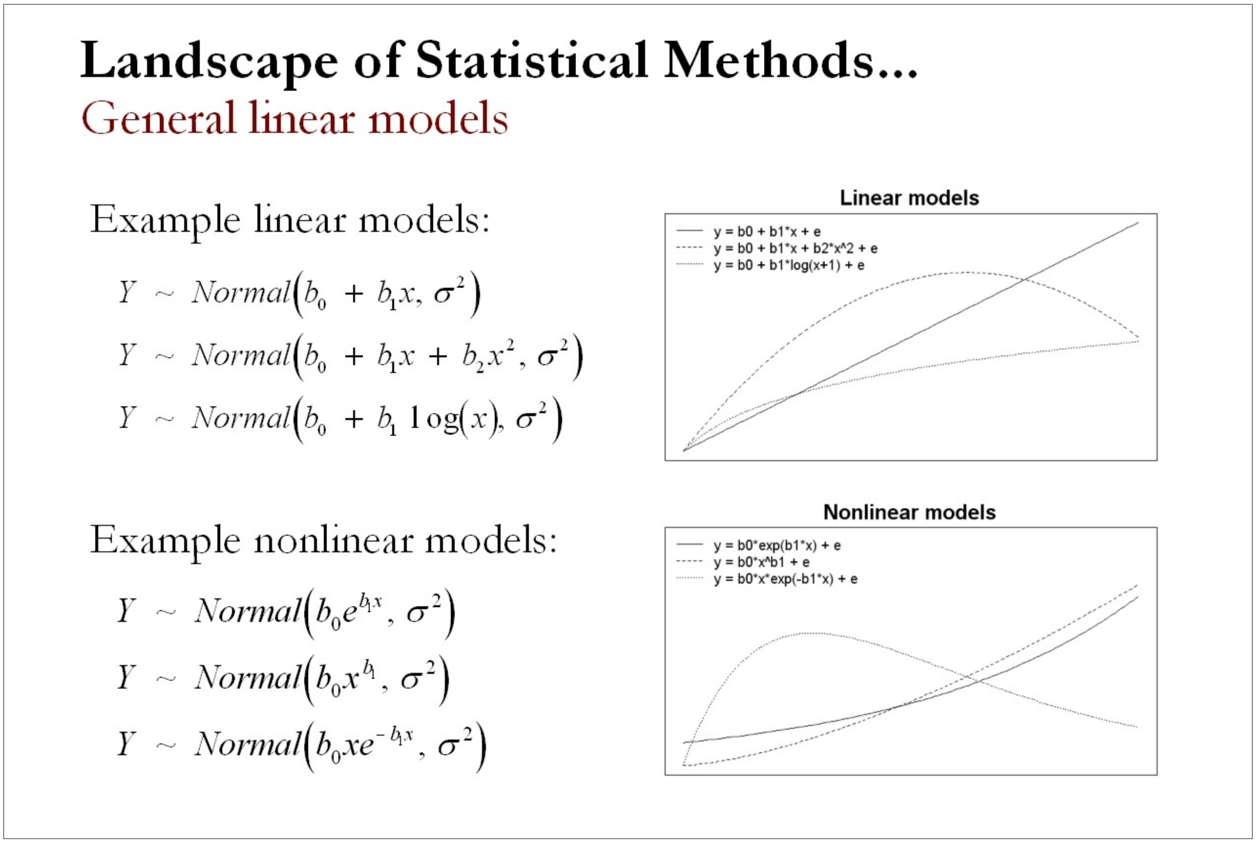
General linear model

2. General linear models

General linear models (sometimes referred to as simply 'linear models', and not to be confused with 'generalized' linear models below) include simple and multiple linear regression, one-way and multiway analysis of variance (ANOVA), and analysis of covariance (ANCOVA). While regression, ANOVA and ANCOVA are often handled differently, and they are usually taught differently in introductory statistics classes, they are all variants of the same basic model. R uses the function lm() for all of these procedures which is based on the method of ordinary least squares.

The assumptions of the general linear model are that all observed values are *independent* and *normally distributed* with a *constant variance* (homoscedastic), and that any continuous predictor variables (covariates) are measured without error. Remember that the assumption of normality applies to the variations around the expected value – the residuals – not to the whole data set.

The 'linear' part of 'general linear model' means that the models are linear functions *of the parameters*, not necessarily of the independent variables. In other words, 'linear' does not mean that the relationship between *Y* and *X* is linear; it simply means that *Y* can be expressed as a linear function of *X*.



Landscape of Statistical Methods...

General linear models

Example linear models:

$$Y \sim Normal\left(b_0 + b_1 x, \sigma^2\right)$$
$$Y \sim Normal\left(b_0 + b_1 x + b_2 x^2, \sigma^2\right)$$
$$Y \sim Normal\left(b_0 + b_1 \log(x), \sigma^2\right)$$

Example nonlinear models:

$$Y \sim Normal\left(b_0 e^{b_1 x}, \sigma^2\right)$$
$$Y \sim Normal\left(b_0 x^{b_1}, \sigma^2\right)$$
$$Y \sim Normal\left(b_0 x e^{-b_1 x}, \sigma^2\right)$$

For example, the following models are all linear regression models:

$$Y \sim Normal\left(b_0 + b_1 x, \sigma^2\right)$$
$$Y \sim Normal\left(b_0 + b_1 x + b_2 x^2, \sigma^2\right)$$
$$Y \sim Normal\left(b_0 + b_1 \log(x), \sigma^2\right)$$

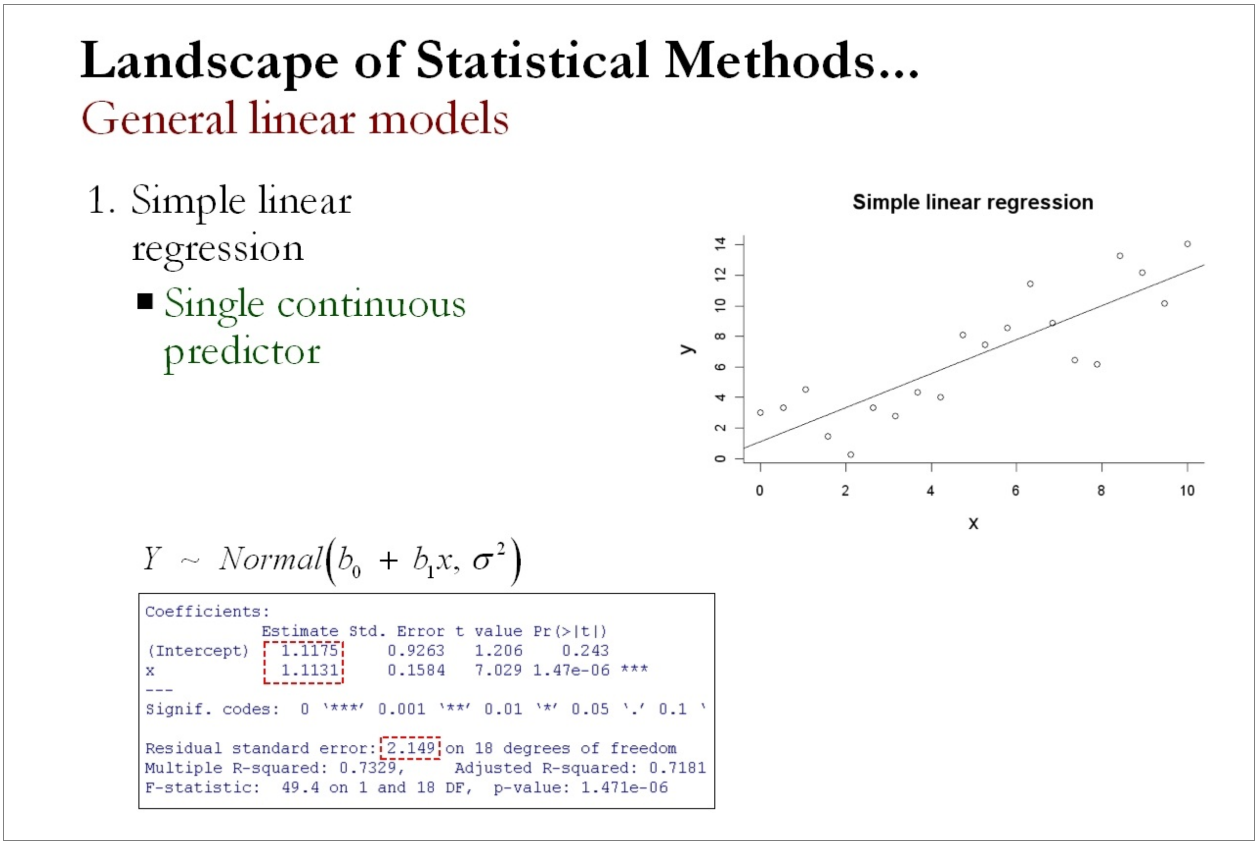
In all these models, we can define a new explanatory variable *z*, such that we have a model of the form:

$$Y \sim Normal\left(b_0 + b_1 z, \sigma^2\right)$$

which is clearly linear in the parameters even though the relationship between *Y* and *X* is not linear except in the first model given above, in which the relationship is expressed as a straight line. However, the following models (exponential, power law, and Ricker) are all nonlinear regression models:

$$Y \sim Normal\left(b_0 e^{b_1 x}, \sigma^2\right)$$
$$Y \sim Normal\left(b_0 x^{b_1}, \sigma^2\right)$$
$$Y \sim Normal\left(b_0 x e^{-b_1 x}, \sigma^2\right)$$

In all of these cases, the models are linear with respect to b_0 but nonlinear with respect to b_1.



Landscape of Statistical Methods…

General linear models

1. Simple linear regression
   - Single continuous predictor

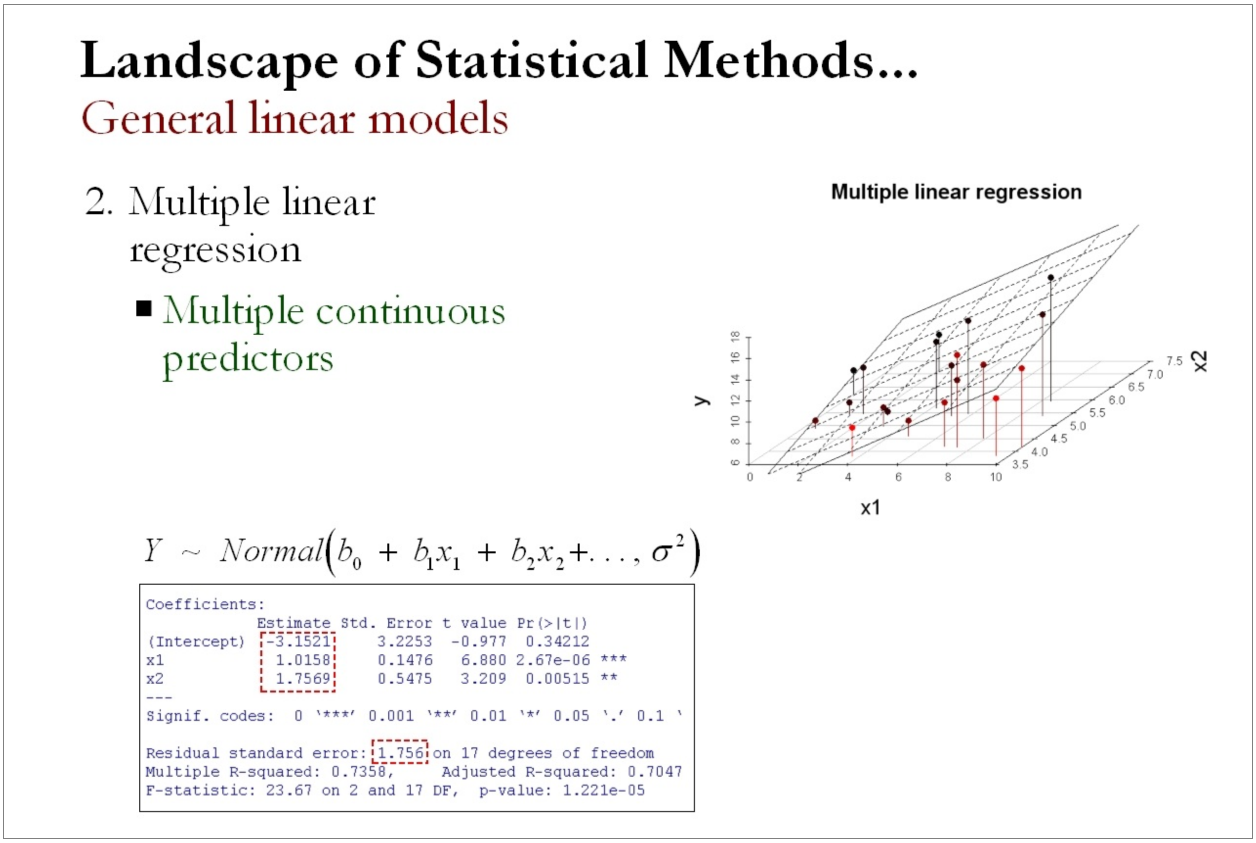
$$Y \sim Normal\left(b_0 + b_1 x,\ \sigma^2\right)$$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.1175     0.9263   1.206    0.243
x            1.1131     0.1584   7.029 1.47e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 2.149 on 18 degrees of freedom
Multiple R-squared: 0.7329,     Adjusted R-squared: 0.7181
F-statistic:  49.4 on 1 and 18 DF,  p-value: 1.471e-06
```

Simple linear regression
Simple, or ordinary, linear regression predicts *y* as a function of a single continuous covariate *x*. The model is:

$$Y \sim Normal\left(b_0 + b_1 x,\ \sigma^2\right)$$

Note, this model assumes the relationship between Y and X is linear and can be expressed as a straight line in the plot of Y against X. If the relationship is not linear, and thus cannot be expressed as a straight line, there are a number of options discussed below.



Landscape of Statistical Methods…

General linear models

2. Multiple linear regression
   - Multiple continuous predictors

$$Y \sim Normal\left(b_0 + b_1x_1 + b_2x_2 + \ldots, \sigma^2\right)$$

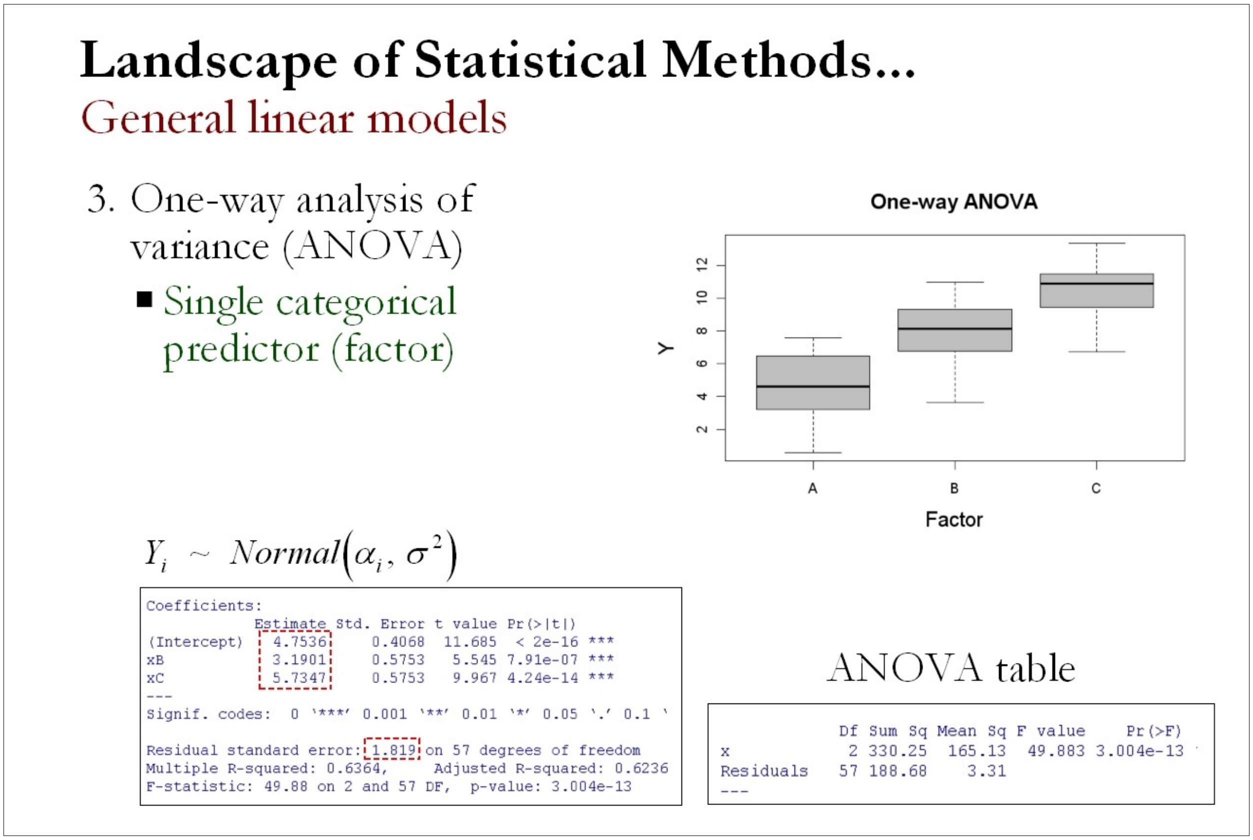
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.1521     3.2253  -0.977  0.34212
x1            1.0158     0.1476   6.880 2.67e-06 ***
x2            1.7569     0.5475   3.209  0.00515 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 1.756 on 17 degrees of freedom
Multiple R-squared: 0.7358,     Adjusted R-squared: 0.7047
F-statistic: 23.67 on 2 and 17 DF,  p-value: 1.221e-05
```

Multiple linear regression
The simple linear regression model can be extended to multiple continuous predictor variables (covariates), as follows:

$$Y \sim Normal\left(b_0 + b_1x_1 + b_2x_2 + \ldots, \sigma^2\right)$$

In addition, we can add interactions among covariates, testing whether the slope with respect to one covariate changes linearly as a function of another covariate, e.g.:

$$Y \sim Normal\left(b_0 + b_1x_1 + b_2x_2 + b_{12}x_1x_2, \sigma^2\right)$$



Landscape of Statistical Methods…

General linear models

3. One-way analysis of variance (ANOVA)
   - Single categorical predictor (factor)

One-way ANOVA
Y
2
4
6
8
10
12
A
B
C
Factor

$$Y_i \sim Normal\left(\alpha_i, \sigma^2\right)$$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.7536     0.4068  11.685  < 2e-16 ***
xB            3.1901     0.5753   5.545 7.91e-07 ***
xC            5.7347     0.5753   9.967 4.24e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 1.819 on 57 degrees of freedom
Multiple R-squared: 0.6364,     Adjusted R-squared: 0.6236
F-statistic: 49.88 on 2 and 57 DF,  p-value: 3.004e-13
```

ANOVA table

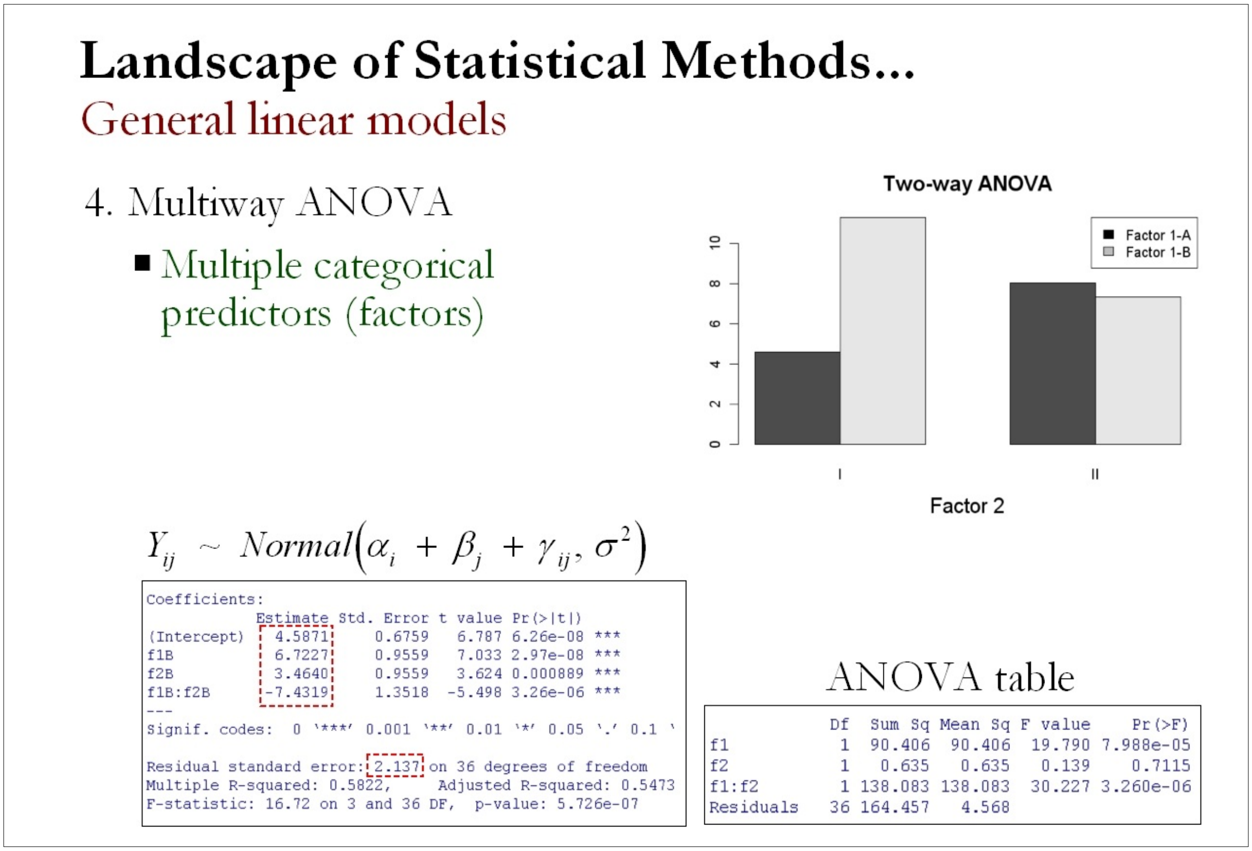
```
          Df Sum Sq Mean Sq F value    Pr(>F)
x          2 330.25  165.13  49.883 3.004e-13
Residuals 57 188.68    3.31
---
```

One-way analysis of variance (ANOVA)

If the predictor variables are discrete (factors) rather than continuous (covariate), the general linear model becomes an analysis of variance. The basic model is:

$$Y_i \sim Normal\left(\alpha_i, \sigma^2\right)$$

In this model, the factor α is a categorical variable (typically defining levels of a treatment) with *i* levels. When fitting regression models, the parameters of the model are easy to interpret – they're just the intercept and the slopes with respect to the covariates. When you have factors, as in ANOVA, the parameterization becomes tricker. By default, the model is parameterized in terms of the differences between the first group and subsequent groups (treatment contrasts) rather than in terms of the mean of each group, although the latter can easily be requested. In ANOVA, the basic purpose is to determine whether the expected value of Y differs between or among levels of the factor.



Landscape of Statistical Methods…

General linear models

4. Multiway ANOVA

- Multiple categorical predictors (factors)

$$Y_{ij} \sim Normal\left(\alpha_i + \beta_j + \gamma_{ij}, \sigma^2\right)$$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.5871     0.6759   6.787 6.26e-08 ***
f1B           6.7227     0.9559   7.033 2.97e-08 ***
f2B           3.4640     0.9559   3.624 0.000889 ***
f1B:f2B      -7.4319     1.3518  -5.498 3.26e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 2.137 on 36 degrees of freedom
Multiple R-squared: 0.5822,     Adjusted R-squared: 0.5473
F-statistic: 16.72 on 3 and 36 DF,  p-value: 5.726e-07
```

ANOVA table

```
          Df  Sum Sq Mean Sq F value    Pr(>F)
f1         1  90.406  90.406  19.790 7.988e-05
f2         1   0.635   0.635   0.139    0.7115
f1:f2      1 138.083 138.083  30.227 3.260e-06
Residuals 36 164.457   4.568
```

Multiway ANOVA

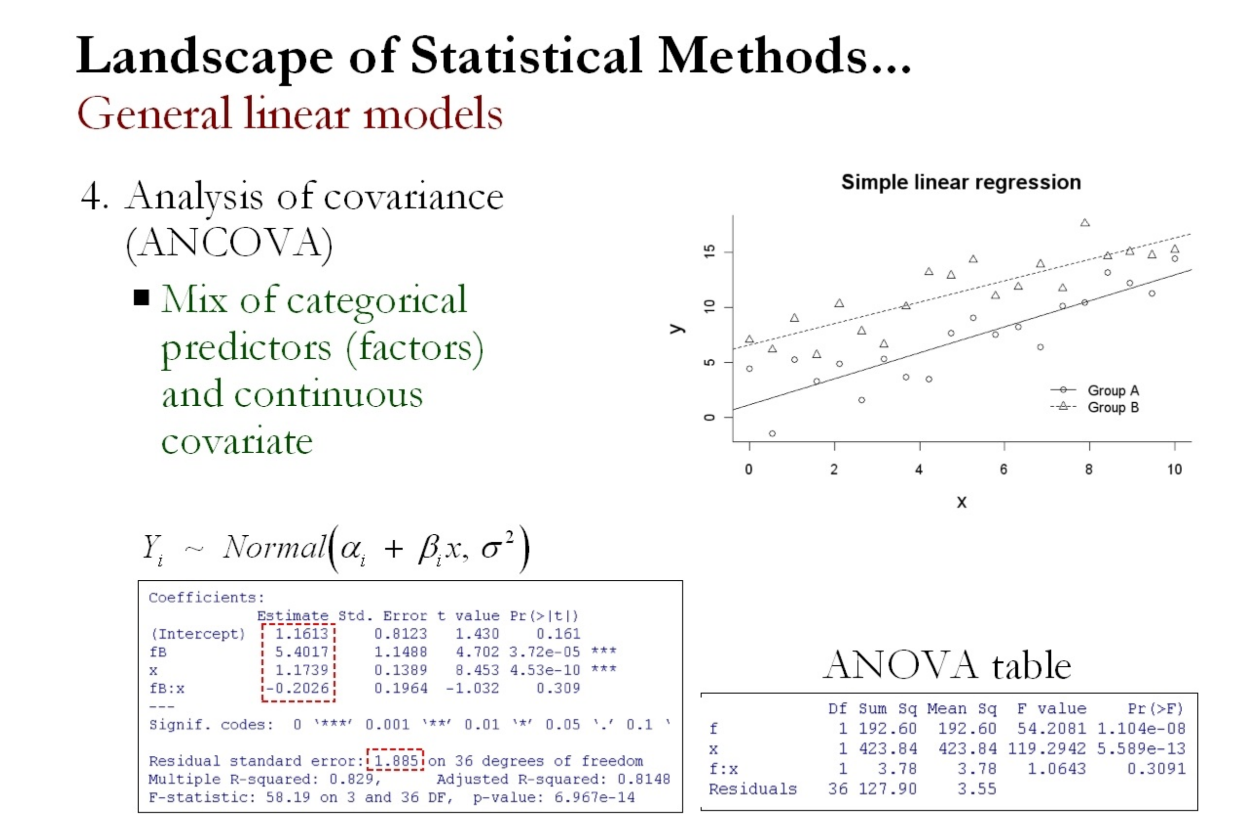
Just as the simple regression model can be extended to include multiple covariates, the ANOVA model can be extended to include multiple factors. For example, the full model for two-way ANOVA (i.e., two factors) with an interaction is:

$$Y_{ij} \sim Normal\left(\alpha_i + \beta_j + \gamma_{ij}, \sigma^2\right)$$

where *i* is the level of the first factor, and *j* is the level of the second factor. The parameters are again defined in terms of contrasts. In the above model consider the case in which the two factors each have two levels:

	Factor 2	
Factor 1	level I	level II
level A	mAI	mAII
level B	mBI	mBII

then there will be four parameters: the first ("intercept") parameter will be the mean of mAI; the second parameter will be the difference between mBI and mAI; the third will be the difference between mAII and mAI; and the fourth, the interaction term, will be the difference between mBII and its expectation if the effects of the two factors were additive, mAI+(mAII-mAI)+(mBI-mAI), which equals mBII-mAII-mBI+mAI. Similar to one-way ANOVA, the basic purpose is to determine whether the expected value of Y differs among levels of the main factors independently or interactively.



Landscape of Statistical Methods…

General linear models

4. Analysis of covariance (ANCOVA)
   - Mix of categorical predictors (factors) and continuous covariate

$$Y_i \sim Normal\left(\alpha_i + \beta_i x, \sigma^2\right)$$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.1613     0.8123   1.430    0.161
fB           5.4017     1.1488   4.702 3.72e-05 ***
x            1.1739     0.1389   8.453 4.53e-10 ***
fB:x        -0.2026     0.1964  -1.032    0.309
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 1.885 on 36 degrees of freedom
Multiple R-squared: 0.829,      Adjusted R-squared: 0.8148
F-statistic: 58.19 on 3 and 36 DF,  p-value: 6.967e-14
```

ANOVA table

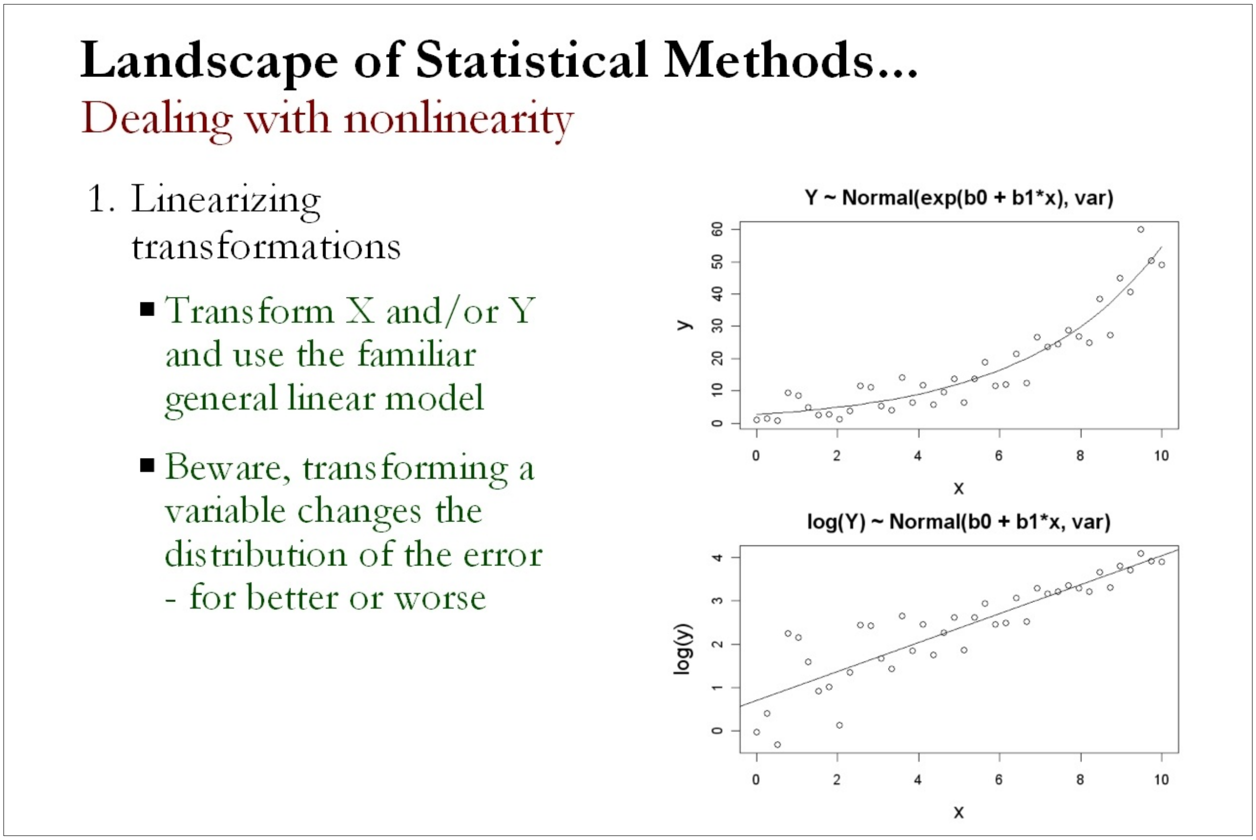
```
          Df Sum Sq Mean Sq  F value    Pr(>F)
f          1 192.60  192.60  54.2081 1.104e-08
x          1 423.84  423.84 119.2942 5.589e-13
f:x        1   3.78    3.78   1.0643    0.3091
Residuals 36 127.90    3.55
```

Analysis of covariance (ANCOVA)

Analysis of covariance defines a statistical model that allows for different intercepts and slopes with respect to a covariate *x* in different groups:

$$Y_i \sim Normal\left(\alpha_i + \beta_i x, \sigma^2\right)$$

The parameters are now the intercept for the first factor level; the differences in the intercepts for each factor level other than the first; the slope with respect to *x* for the first factor level; and the differences in the slopes for each factor level other than the first.



Landscape of Statistical Methods…

Dealing with nonlinearity

1. Linearizing transformations
   - Transform X and/or Y and use the familiar general linear model
   - Beware, transforming a variable changes the distribution of the error - for better or worse

3. Nonlinearity

One of the common situations we deal with in ecological data is that relationships between *Y* and *X* are more often than not nonlinear. There are a variety of methods for dealing with nonlinearity in regression problems.

Linearizing transformations

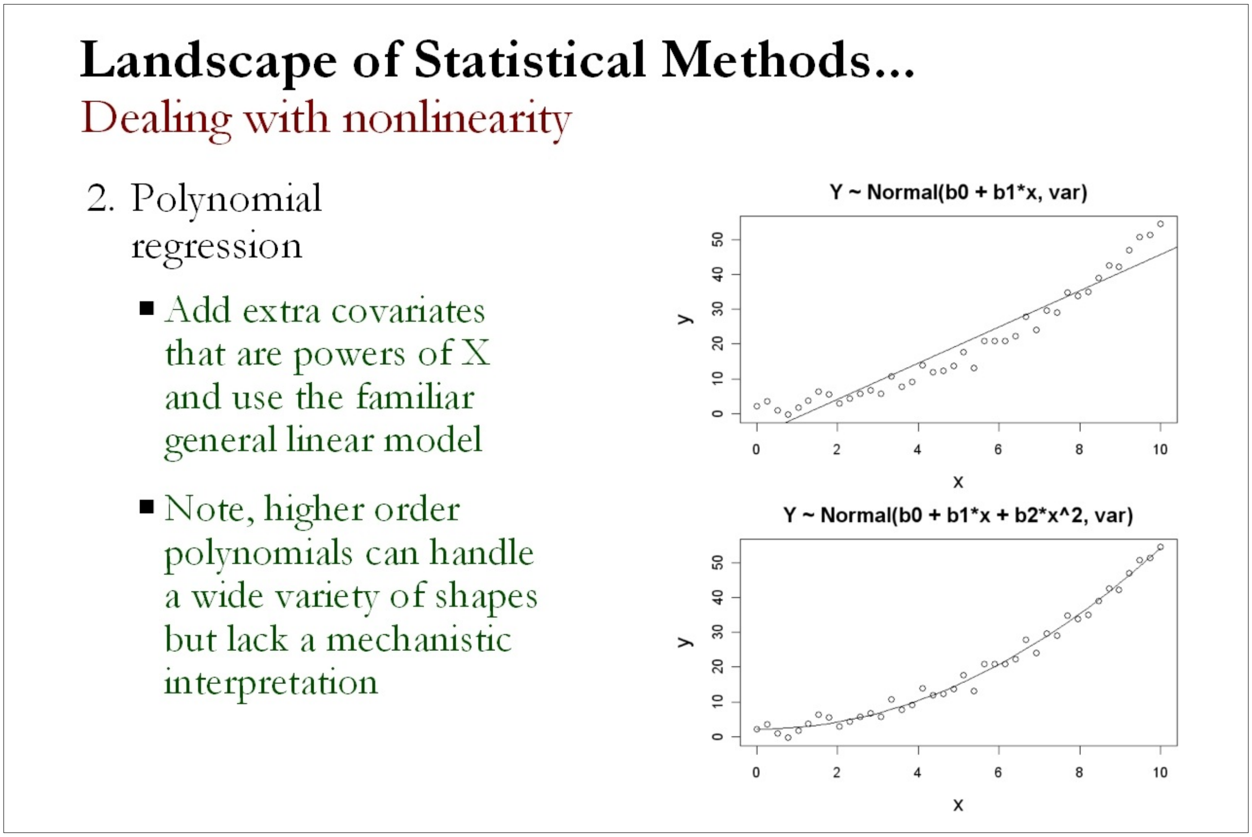
One option is to try and find a transformation of the parameters that linearizes the relationship between *Y* and *X* and then use the familiar general linear model (above). For example, if *Y* increases exponentially as function of *X*:

$$Y \sim Normal\left(e^{b_0+b_1x}, \sigma^2\right)$$

we could log transform both sides of the equation and treat the model as a simple linear regression, as follows:

$$\log(Y) \sim Normal\left(b_0 + b_1x, \sigma^2\right)$$

Before computers were ubiquitous, a lot of ingenuity went into developing transformation methods to linearize common functions. However, transforming variables changes the distribution of the error as well as the shape of the dependence of *Y* on *X*. Ideally we'd like to find a transformation that simultaneously produces a linear relationship and makes the errors normally distributed with constant variance, but these goals are often incompatible. If the errors are normal with constant variance, they won't be after the transformation to linearize *f(x)*. Thus, linearizing transformations are not the ideal solution and other methods (below) are generally preferred.



Landscape of Statistical Methods…

Dealing with nonlinearity

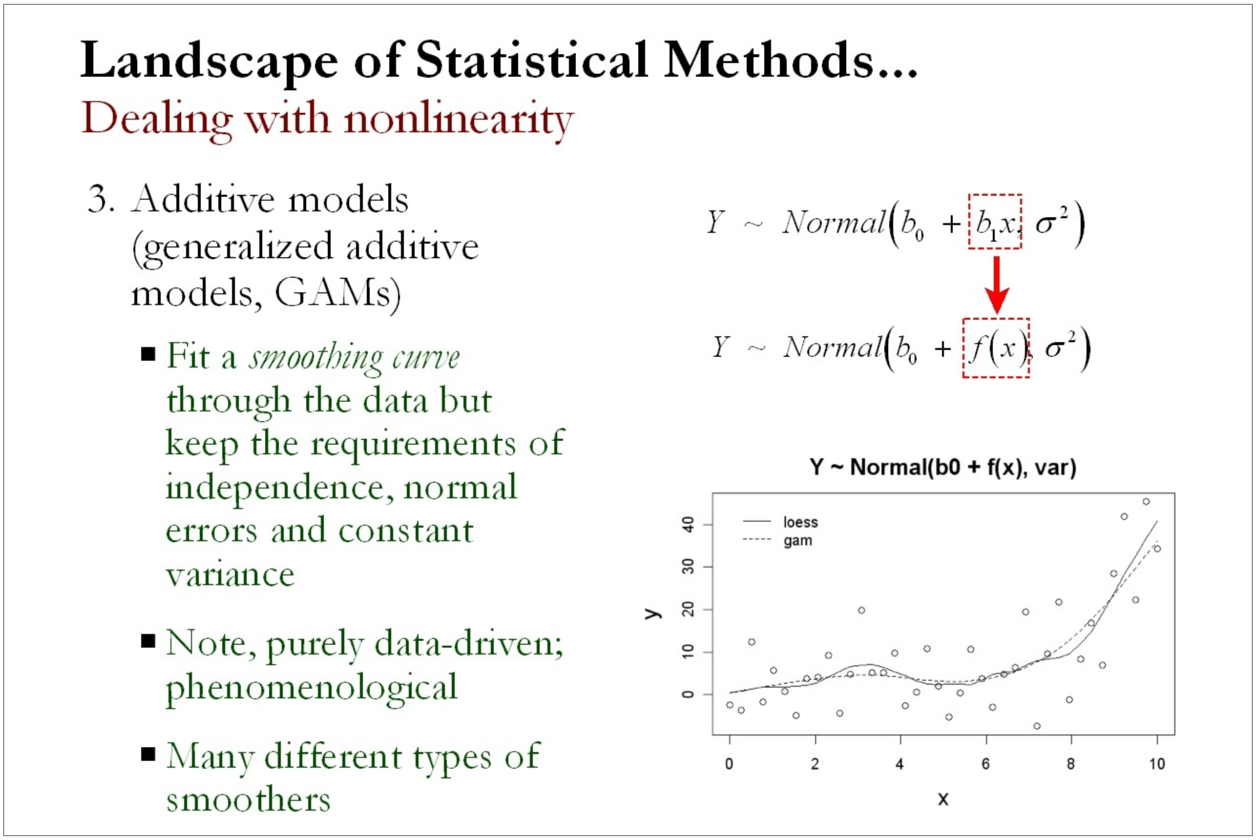
2. Polynomial regression
   - Add extra covariates that are powers of X and use the familiar general linear model
   - Note, higher order polynomials can handle a wide variety of shapes but lack a mechanistic interpretation

Polynomial regression

Another option is to fit a polynomial regression by adding extra covariates that are powers of the original variable (x^2, x^3, ...): *quadratic* if just the x^2 term is added, *cubic* if both the x^2 and x^3 are added, and so on for higher-order terms. For example, the quadratic regression model is:

$$Y \sim Normal\left(b_0 + b_1x + b_2x^2, \sigma^2\right)$$

Note, the polynomial regression model allows nonlinear relationships between Y and X to be expressed, but it is still a general linear model, since it is linear in the parameters, and has the usual requirements of independence, normal errors and constant variance. Polynomial regression is quite flexible in fitting a wide variety of shapes: quadratic terms allow for humps, cubic terms allow for inflections, and powers of 4 allow for local maxima. However, polynomial models rarely allow for a mechanistic interpretation of the parameters and are thus typically used phenomenologically to fit nonlinear patterns in the data.



Landscape of Statistical Methods...

Dealing with nonlinearity

3. Additive models (generalized additive models, GAMs)
   - Fit a *smoothing curve* through the data but keep the requirements of independence, normal errors and constant variance
   - Note, purely data-driven; phenomenological
   - Many different types of smoothers

$$Y \sim Normal\left(b_0 + \boxed{b_1 x}, \sigma^2\right)$$

$$\downarrow$$

$$Y \sim Normal\left(b_0 + \boxed{f(x)}, \sigma^2\right)$$

Additive models
Another option is to fit an additive model (or more generally, a generalized additive model, or GAM), which fits a smoothing curve through the data but keeps the requirements of independence, normal errors and constant variance. Additive models are purely phenomenological since the fits are purely data-driven and not based on any mechanistic understanding of the underlying population. In R there are two main packages for GAM: the gam package and the mgcv package. The mgcv package is more advanced and allows for cross-validation and generalized additive mixed modeling (GAMM) including spatial and temporal correlations as well as nested data and various heterogeneity patterns. Cross-validation is a process that automatically determines the optimal amount of smoothing.

The basic GAM model is:

$$Y \sim Normal\left(b_0 + f(x), \sigma^2\right)$$

Note that the only difference between the simple regression model presented above and the GAM model is the replacement of b_1x by the smoothing curve *f(x)*. The linear regression model gives us a formula and the relationship between Y and X is quantified by an estimated regression parameter plus confidence intervals. In a GAM, we do not have such an equation. We have a smoother, and the only thing we can do with it is plot it. This does not mean that we cannot predict from this model; we can, but not with a simple equation.

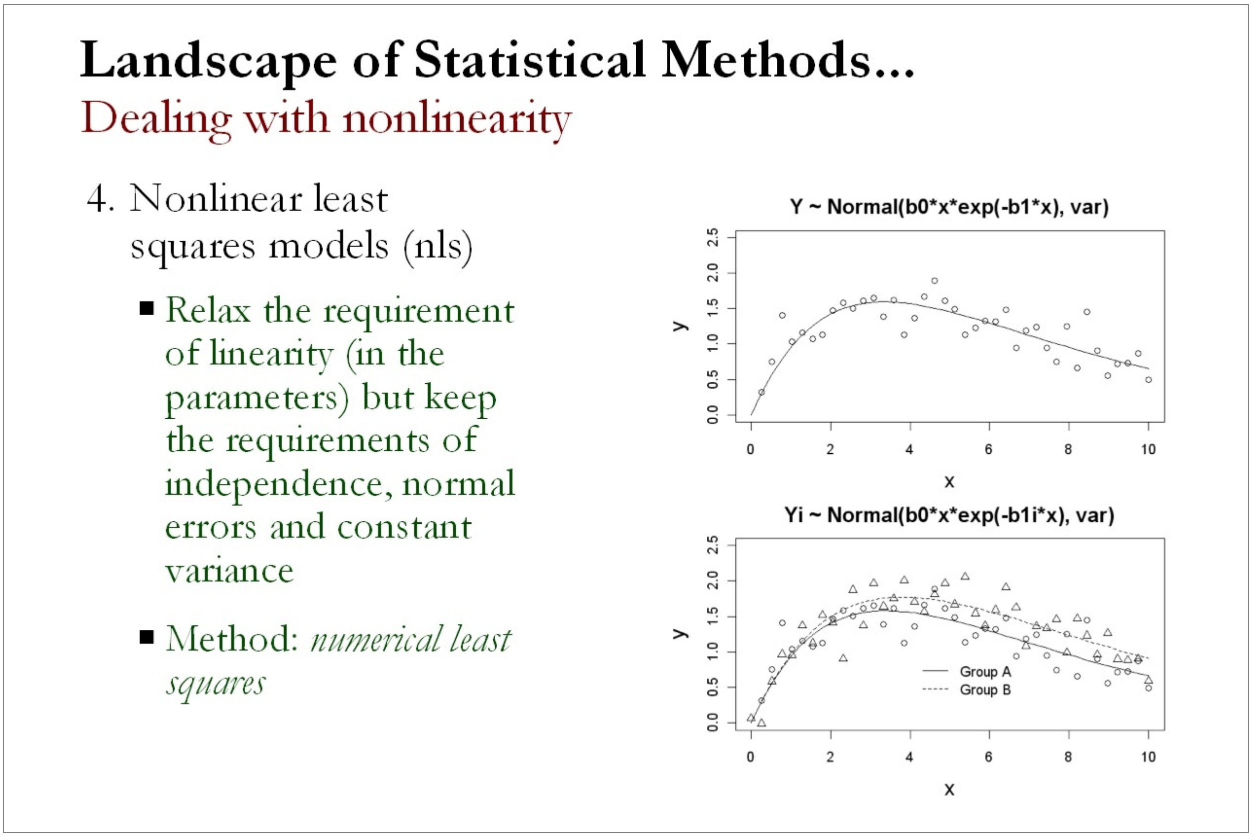
There are many different types of smoothers in GAM. The simplest smoother is referred to as LOESS smoothing, which involves applying local linear regression. Briefly, for each target value of *X*, a window around this value is chosen and all points in this window are used in a local linear regression analysis to predict the value of *X*. This process is repeated for every target value of X, essentially by moving the window along the *X* gradient and getting a new predicted value for each target value of *X*. The curve connecting all of the predicted values is called LOESS smoothing. Alternatively, we can apply a weighted linear regression, where the weights are determined by the distance (along the *X* axis) of the observations from the target value, which is called LOWESS smoothing. It is also possible to use polynomial models of any order; although typically the order is two (quadratic), which may be referred to as local polynomial regression. One major difficulty with these types of smoothers is finding the optimal span width, or window size, which is a matter of bias-variance tradeoff: the narrower the window, the better the fit (i.e., less bias), but the greater the variance or uncertainty (resulting in wide confidence bands). This can be done subjectively based on visual inspection of the smoother and the residuals. Another option is to use AIC to pick the "best" span.

The mgcv package has a wide variety of smoothers plus a built-in cross-validation procedure for choosing the optimal amount of smoothing. The most commonly used smoother in mgcv is a cubic regression spline, although several others are available. Briefly, for the cubic regression spline method, the *X* gradient is divided into a certain number of intervals. In each interval, a cubic polynomial regression is fitted, and the fitted values per segment are then glued together to form the smoothing curve. The point where the intervals connect are called knots. To obtain a smooth connection at the knots, certain conditions are imposed, which we need not worry about here. Thus, for a cubic regression spline, the smooth function in the basic GAM model above is:

$$f(X) = b_0 + b_1 x + b_2 x^2 + b_3 x^3$$

The problem boils down to finding the values of b_0, b_1, b_2, and b_3 for each segment, which can easily be done using ordinary least squares, and finding the optimal number of intervals. The details of how to find the optimal number of intervals goes beyond the scope of this chapter but essentially involves minimizing an objective function that includes a penalty for adding each additional knot so that the number of intervals is kept at a minimum to avoid overfitting the data. The objective function is based on a jackknife cross-validation procedure that involves leaving each observation out in turn, estimating the smoother using the remaining n-1 observations, predicting the value of the held-out observation using the estimated smoother, computing the difference between the predicted value and real value, and summing the squared differences across all observations. Essentially, the number is knots is selected such that the sums of squared prediction residuals is minimized subject to a penalty for adding each additional knot.

Note, the basic GAM model can easily be extended to include additional explanatory variables, including both covariates and factors, as well as interactions between explanatory variables.



Landscape of Statistical Methods…

Dealing with nonlinearity

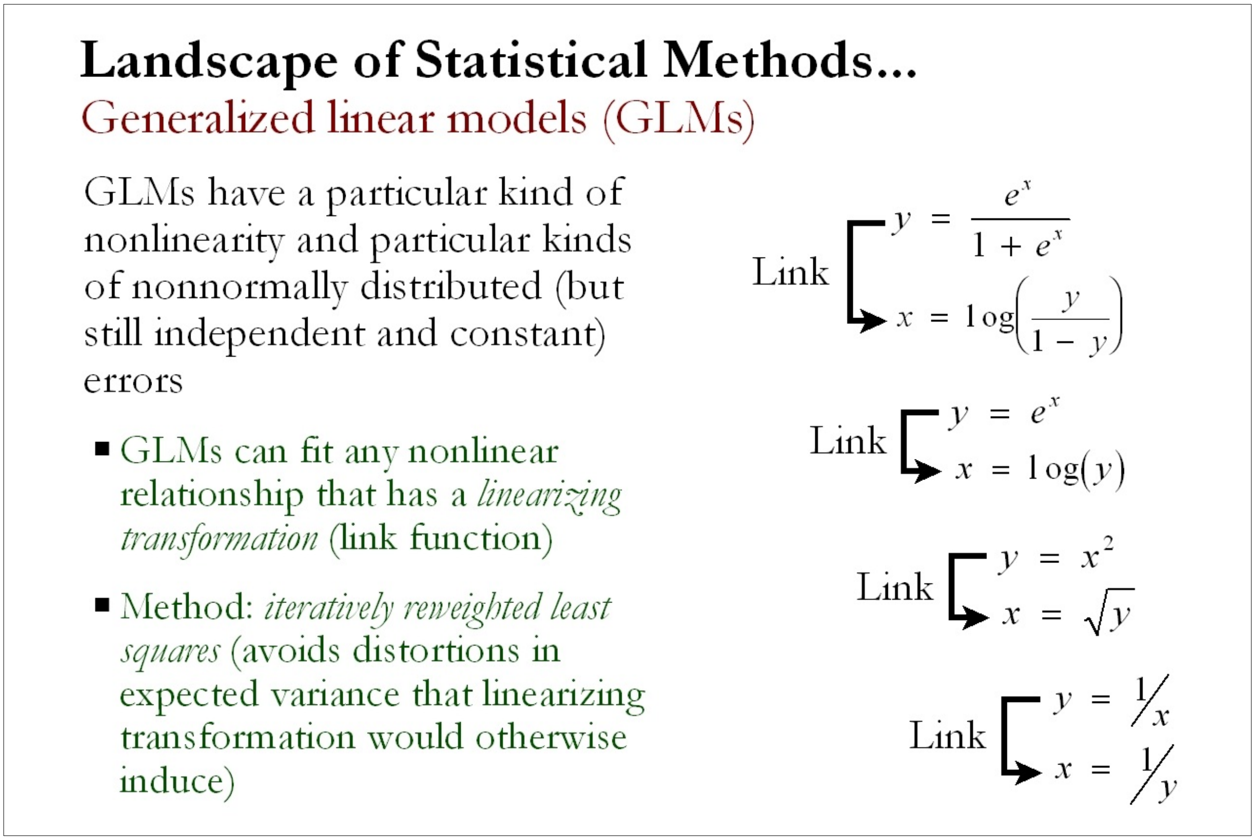
4. Nonlinear least squares models (nls)
   - Relax the requirement of linearity (in the parameters) but keep the requirements of independence, normal errors and constant variance
   - Method: *numerical least squares*

Nonlinear least squares
A final option is to fit a nonlinear regression model using the procedure of nonlinear least squares (NLS). R uses function nls() for this procedure. Nonlinear least squares models relax the requirement of linearity (in the parameters) but keep the requirements of independence, normal errors and constant variance. An example is the Ricker model with normal errors:

$$Y \sim Normal\left(b_0 x e^{-b_1 x}, \sigma^2\right)$$

Note, the Ricker model is linear with respect to b_0 but nonlinear with respect to b_1; thus, not only is the relationship between *Y* and *X* nonlinear, the model itself is nonlinear since it cannot be expressed as a linear function of the parameters. The modern way to solve nonlinear models such as this is to minimize the sums of squares (equivalent to minimizing the negative log-likelihood) computationally, since an analytical solution is not possible. Restricting the variance model to normally distributed errors with constant variance allows the use of specific numeric optimization methods that are more powerful and stable than the generalized algorithms that can accommodate other error distributions (such as generalized nonlinear least squares, gnls).

Fitting models with both nonlinear covariates and categorical variables (the nonlinear analogue of ANCOVA) is more difficult. Two functions from the nlme package, nlsList() and gnls() can handle such models, the latter being much more flexible.



Landscape of Statistical Methods...

Generalized linear models (GLMs)

GLMs have a particular kind of nonlinearity and particular kinds of nonnormally distributed (but still independent and constant) errors

- GLMs can fit any nonlinear relationship that has a *linearizing transformation* (link function)
- Method: *iteratively reweighted least squares* (avoids distortions in expected variance that linearizing transformation would otherwise induce)

Link

$$y = \frac{e^x}{1 + e^x}$$

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

Link

$$y = e^x$$

$$x = \log(y)$$

Link

$$y = x^2$$

$$x = \sqrt{y}$$

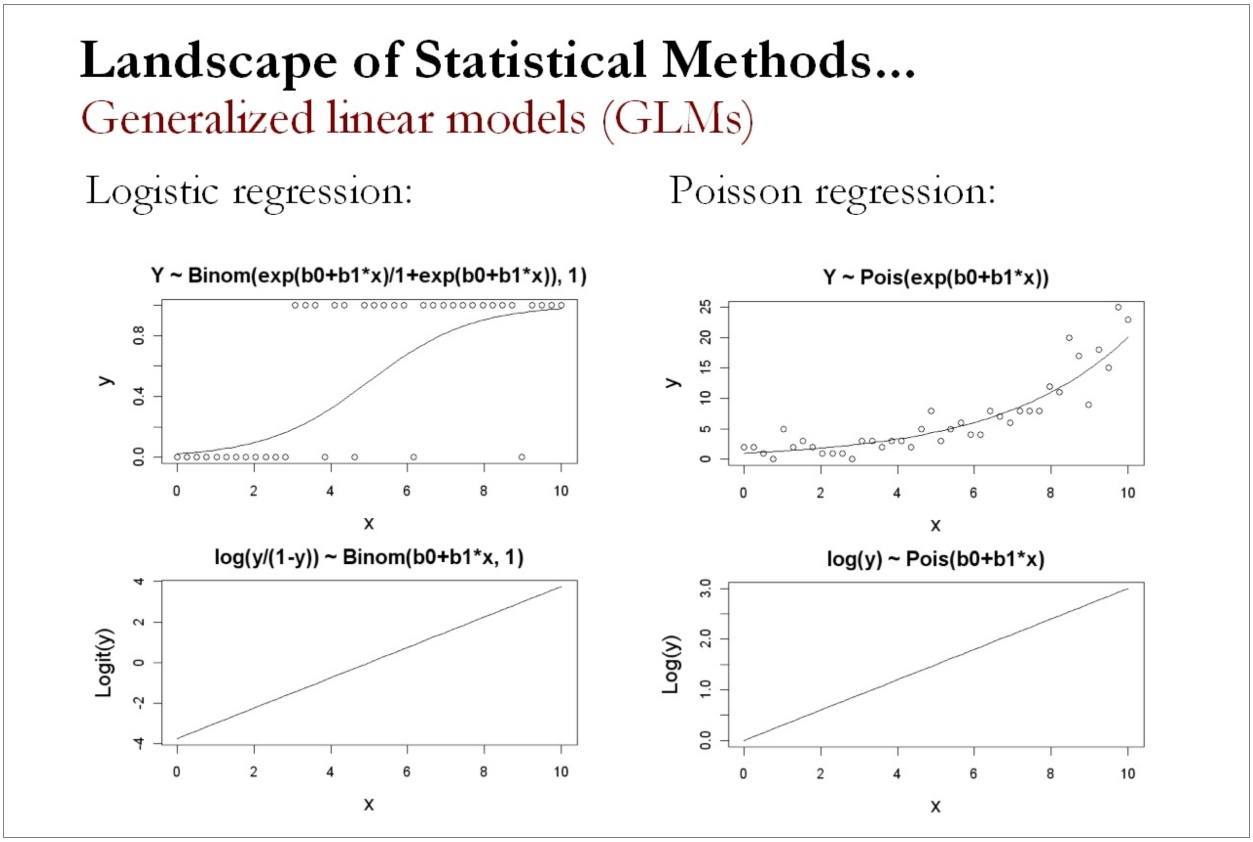
Link

$$y = 1/x$$

$$x = 1/y$$

4. Nonlinearity and nonnormal errors (generalized linear models)

One of the major assumptions of the general linear model and the modifications we discussed above for dealing with nonlinear deterministic relationships (additive models and nonlinear least squares) is that the errors are normally distributed. Not surprisingly, in ecological data this assumption is not met more often than it is met, so we need alternative methods that can handle nonlinear models with nonnormal errors. *Generalized linear models* (not to be confused with general linear models) allow us to analyze models that have a particular kind of nonlinearity and particular kinds of nonnormally distributed (but still independent and constant) errors. Generalized linear models (GLMs) can fit any nonlinear relationship that has a *linearizing transformation*. That is, if y = f(x), there must be some function F such that F(f(x)) is a linear function of x. Previously we said that linearizing transformations were challenging because they change the distribution of the errors – often in bad ways. GLMs are special in this regard because they use the function F to fit the data on the linearized scale (F(y) = F(f(x))) while calculating the expected variance on the untransformed scale in order to correct for the distortions that linearization would otherwise induce. In GLM jargon F is call the *link* function.



Landscape of Statistical Methods…

Generalized linear models (GLMs)

Logistic regression:

Poisson regression:

There are a variety of link functions for some of the most common nonlinearity problems. In each case, the link function serves to linearize the corresponding nonlinear function:

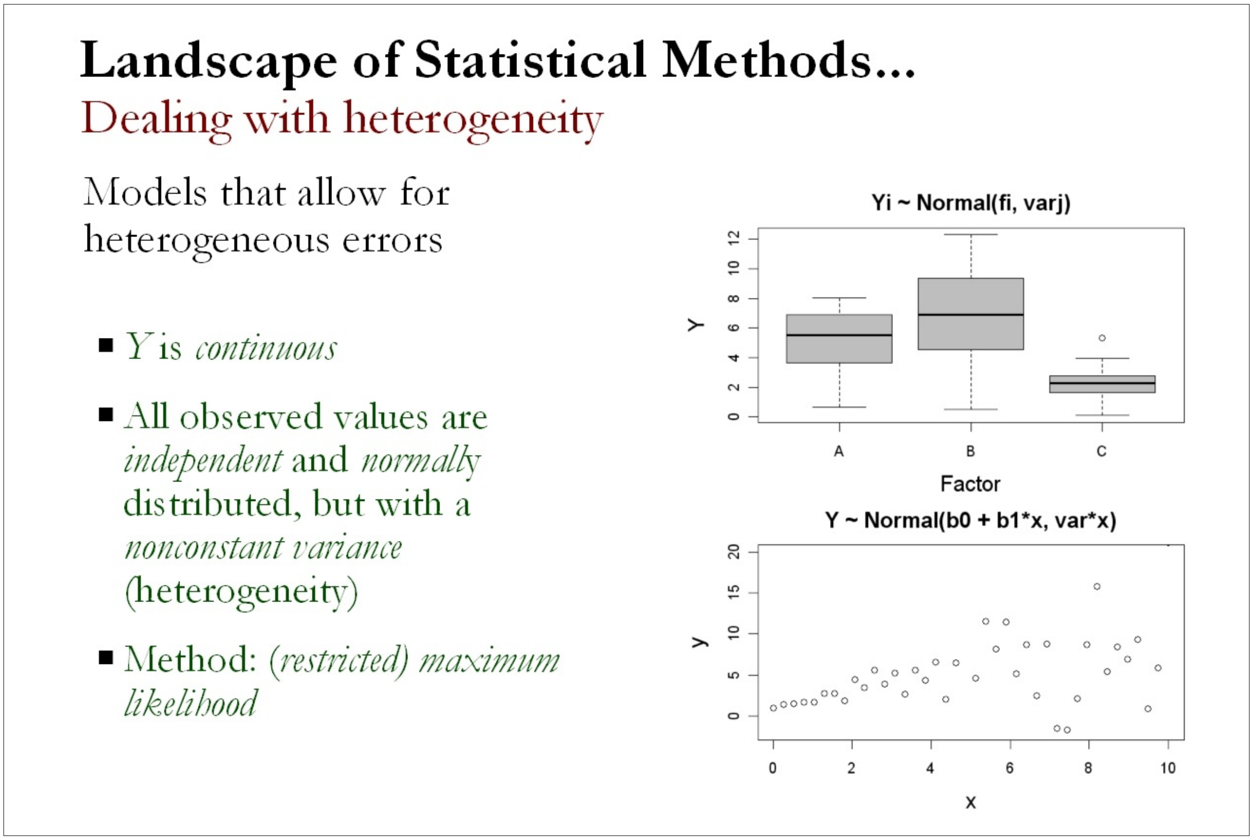
Link function		Nonlinear function
logit	$x = 1\text{og}\left(\frac{y}{1 - y}\right)$	$y = \frac{e^x}{1 + e^x}$
log	$x = 1\text{og}(y)$	$y = e^x$
square root	$x = \sqrt{y}$	$y = x^2$
inverse	$x = {}^{1}\!/_{y}$	$y = {}^{1}\!/_{x}$

The class of nonnormal errors that GLMs can handle is called the *exponential family*. It includes the poisson, binomial, gamma and normal distributions. Each distribution has a standard link function that makes sense for the typical situation. For example, the logit link is standard for a binomial distribution, which in combination is commonly referred to as logistic regression. The logit transformation converts unconstrained values into values between 0 and 1, which are appropriate as probabilities in a binomial model. This is quite useful when dealing with binary response data (0 or 1 data) or proportional response data (0-1), both quite common in ecology. Similarly, the log link is standard for a poisson distribution, which in combination is commonly referred as a log-linear model or simply poisson regression. This is quite useful when dealing with count data (non-negative integer values), which is also quite common in ecology.

GLMs are fit by a numerical optimization process called *iteratively reweighted least squares*, which overcomes the basic problem that transforming the data to make them linear also changes the variance. The key to this procedure is that given an estimate of the regression parameters, and knowing the relationship between the variance and the mean for a particular distribution (which we do for all of the exponential family distributions), one can calculate the variance associated with each point. With this variance estimate, one reestimates the regression parameters weighted each data point by the inverse of its variance; the new estimate gives new estimates of the variance; and so on. This procedure quickly and reliably fits the models, without the user needing to specify starting values for the parameters.

GLMs combine a range of nonnormal error distributions with the ability to work with some reasonable nonlinear functions. Importantly, they also use the same basic model specification framework as general linear models (Lms).

Just as linear models can be morphed into generalized linear models to accommodate nonlinear relationships and nonnormal errors, additive models can be modified into generalized additive models (GAMs) to accommodate nonnormal errors in a similar fashion.



Landscape of Statistical Methods...

Dealing with heterogeneity

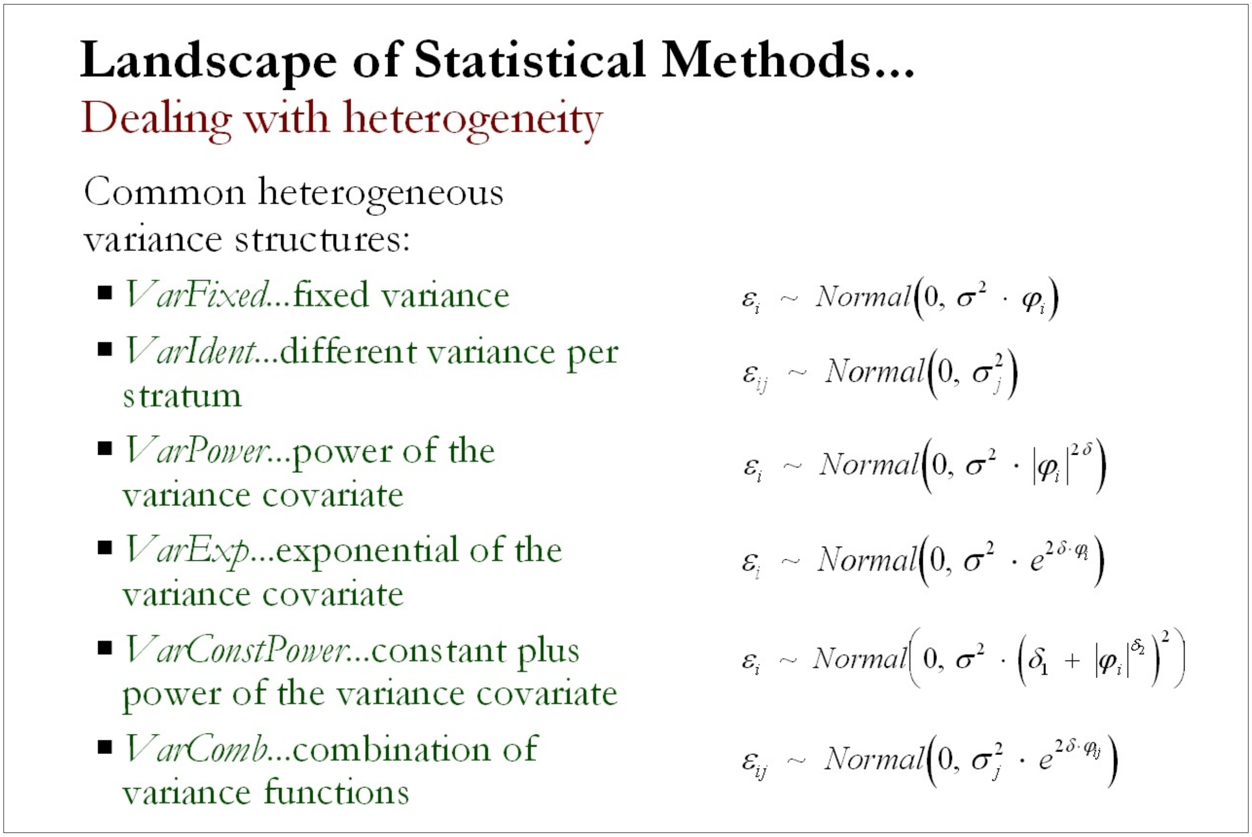
Models that allow for heterogeneous errors

- *Y* is *continuous*
- All observed values are *independent* and *normally* distributed, but with a *nonconstant variance* (heterogeneity)
- Method: *(restricted) maximum likelihood*

5. Heterogeneous errors

One of the most important assumptions of linear models (LMs), generalized linear models (GLMs), additive models (GAMs), and nonlinear least squares (NLS) is that the variance is constant – the homogeneity of variance or homoscedastic variance assumption. Not surprisingly, in ecological data we often encounter situations were this assumption is not met – a problem we refer to as *heterogeneity*. Failure to meet this assumption in the above models may result in parameter estimates with incorrect standard errors and test statistics such as the F statistic and t statistic no longer following their respective distributions, which invalidates the use of these test statistics for assessing statistical significance.

One approach for dealing with heterogeneity is to try a data transformation, such as a log transformation, which sometimes can work to address the issue but also can be risky because data transformations change the distribution of the error, e.g., it may make normally distributed errors become nonnormal after the transformation. Heterogeneity is often interesting ecological information that we may not want to throw away with a data transformation just because it is statistically inconvenient. With a little bit of extra mathematical effort, heterogeneity can be incorporated into the models and can provide extra ecological information.



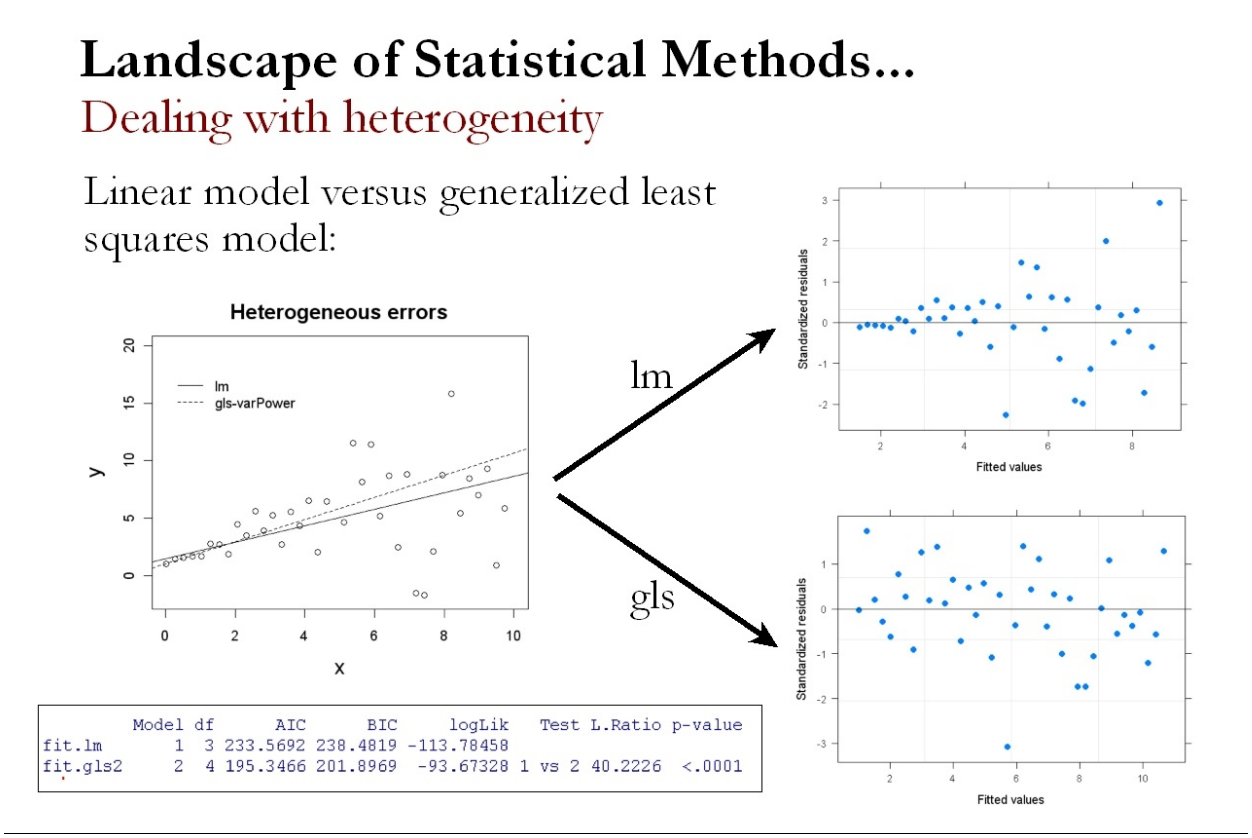
Landscape of Statistical Methods...

Dealing with heterogeneity

Common heterogeneous variance structures:

- *VarFixed*...fixed variance $\varepsilon_i \sim Normal\left(0,\ \sigma^2 \cdot \varphi_i\right)$
- *VarIdent*...different variance per stratum $\varepsilon_{ij} \sim Normal\left(0,\ \sigma_j^2\right)$
- *VarPower*...power of the variance covariate $\varepsilon_i \sim Normal\left(0,\ \sigma^2 \cdot |\varphi_i|^{2\delta}\right)$
- *VarExp*...exponential of the variance covariate $\varepsilon_i \sim Normal\left(0,\ \sigma^2 \cdot e^{2\delta\cdot\varphi_i}\right)$
- *VarConstPower*...constant plus power of the variance covariate $\varepsilon_i \sim Normal\left(0,\ \sigma^2 \cdot \left(\delta_1 + |\varphi_i|^{\delta_2}\right)^2\right)$
- *VarComb*...combination of variance functions $\varepsilon_{ij} \sim Normal\left(0,\ \sigma_j^2 \cdot e^{2\delta\cdot\varphi_{ij}}\right)$

There are a number of options for modeling the variance structure. One option is to use one of the common heterogeneous variance structures. One of these common structures is called the *fixed variance*; it assumes that the variance is proportional to the value of a covariate. For example, the variance may increase as the predictor variable *x* increases, which involves adding no additional parameters to the model. There are several other common structures, including those shown here.



Landscape of Statistical Methods...

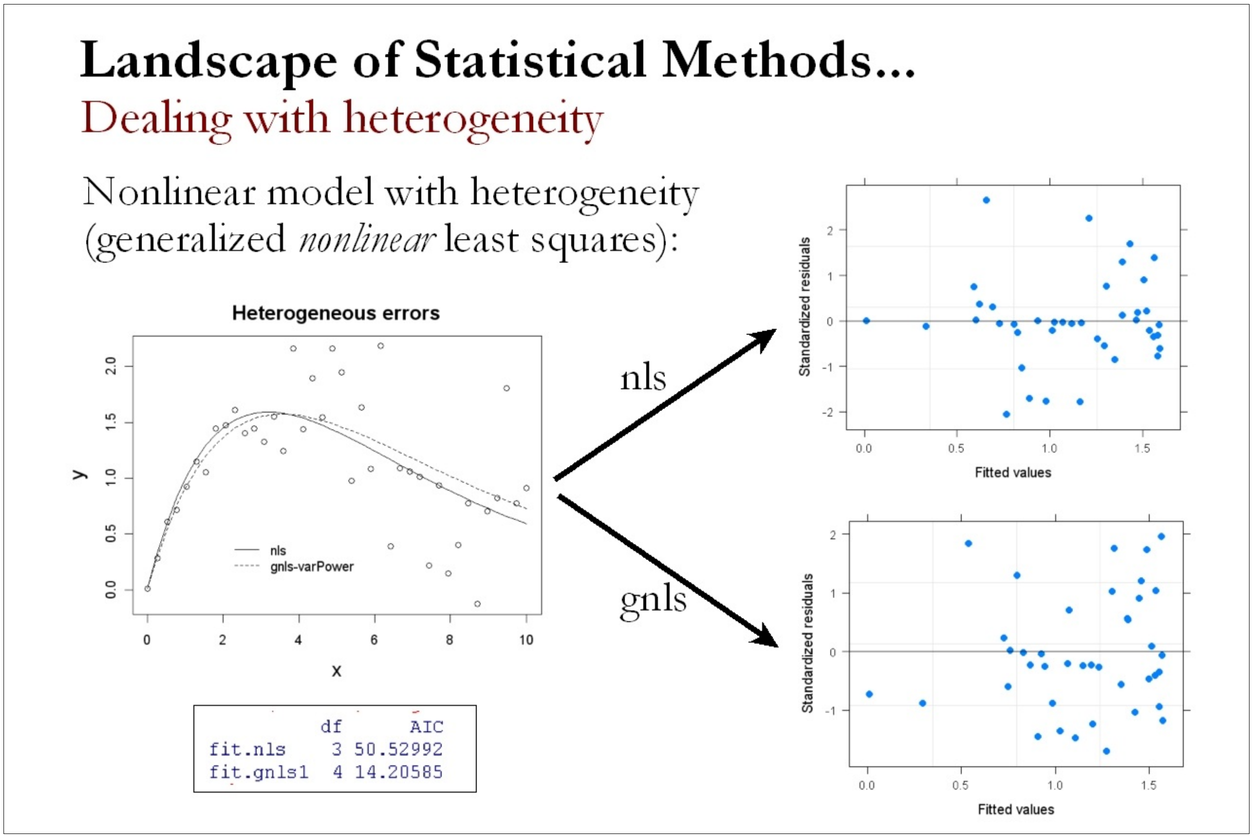
Dealing with heterogeneity

Linear model versus generalized least squares model:

```
             Model df      AIC      BIC    logLik   Test L.Ratio p-value
fit.lm           1  3 233.5692 238.4819 -113.78458
fit.gls2         2  4 195.3466 201.8969  -93.67328 1 vs 2 40.2226  <.0001
```

Generalized least squares (GLS)

A linear model using any one of these common variance structures can be fit with method of generalized least squares (GLS). In the example shown here, the simple linear regression model with heterogeneous errors is fit using the standard linear model (LM), which assumes constant errors, and using generalized least squares with a power variance structure. The variance power structure allows the variance to increase as a power of a covariate, in this case *x*. The residual plots reveal that the LM model results in nonconstant residuals, as expected, whereas the GLS model deals with this issue effectively. In addition, we can use likelihood ratio tests (in this case because the models are nested) and/or AIC to compare the models. In this case, the AIC is considerably smaller for the GLS model and the LRT indicates that it is in fact significantly better than the lm model.



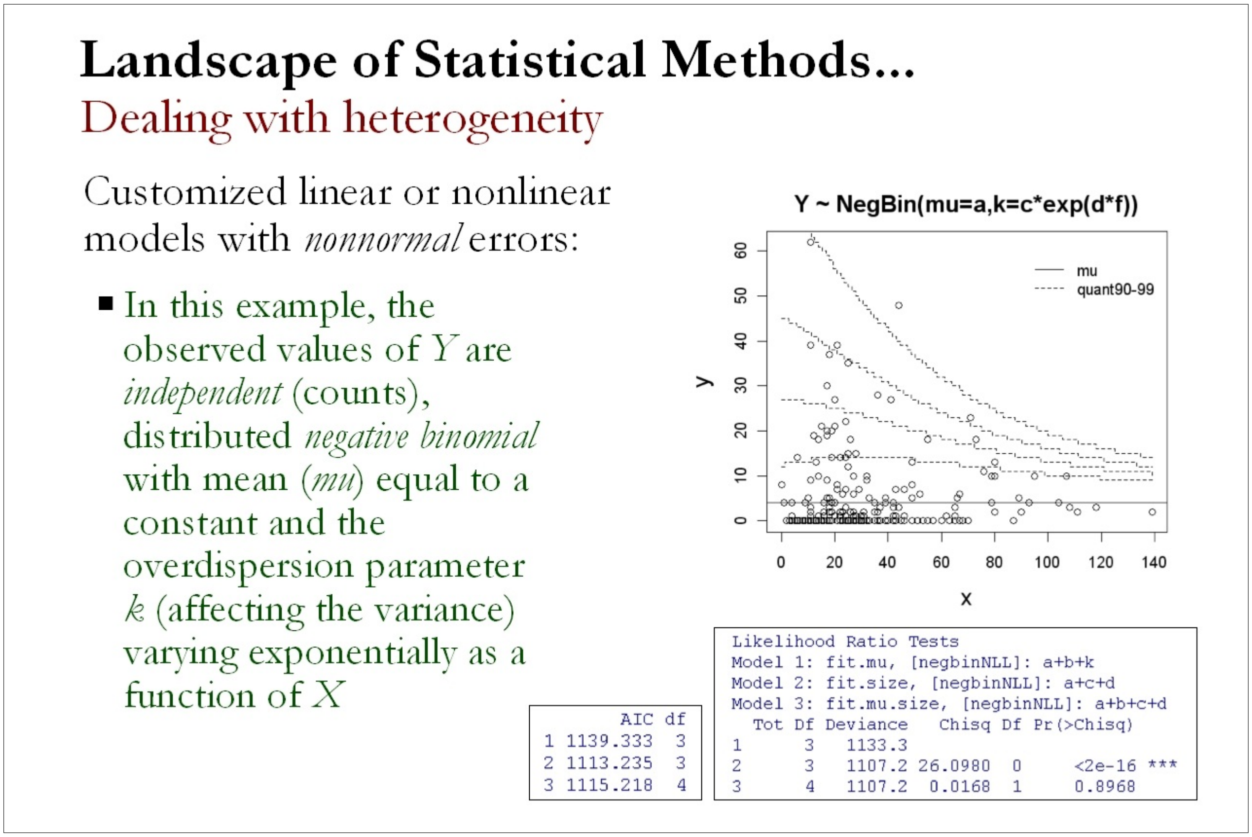
Landscape of Statistical Methods…

Dealing with heterogeneity

Nonlinear model with heterogeneity (generalized *nonlinear* least squares):

```
          df      AIC
fit.nls    3 50.52992
fit.gnls1  4 14.20585
```

Generalized nonlinear least squares (GNLS)
We can account for heterogeneous errors in nonlinear models as well. In the example shown here, a Ricker model is used to produce a nonlinear relationship between *Y* and *X*. In addition to the obvious nonlinear pattern, which cannot easily be linearized, there is a clear indication that the variance is not constant over the range of *X*; it appears to increase in relation to *X*. As in the previous example, we can account for the heterogeneity in the model. If the errors are normally distributed (but heterogeneous), the model can be fit with the method of generalized nonlinear least squares (GNLS), which effectively combines the nonlinear modeling approach of nonlinear least squares (NLS) with the heterogeneous variance modeling approach of GLS. In the example shown here, the GNLS model is clearly a better model based on AIC, even though the fit doesn't appear to be all that different and the residual plots appear to be only marginally better than the NLS model.



Landscape of Statistical Methods…

Dealing with heterogeneity

Customized linear or nonlinear models with *nonnormal* errors:

- In this example, the observed values of *Y* are *independent* (counts), distributed *negative binomial* with mean (*mu*) equal to a constant and the overdispersion parameter *k* (affecting the variance) varying exponentially as a function of *X*

```
    AIC df
1 1139.333  3
2 1113.235  3
3 1115.218  4
```

```
Likelihood Ratio Tests
Model 1: fit.mu, [negbinNLL]: a+b+k
Model 2: fit.size, [negbinNLL]: a+c+d
Model 3: fit.mu.size, [negbinNLL]: a+b+c+d
  Tot Df Deviance   Chisq Df Pr(>Chisq)
1      3   1133.3
2      3   1107.2 26.0980  0      <2e-16 ***
3      4   1107.2  0.0168  1      0.8968
```

Customized models for nonlinear errors

If the errors are normally distributed, but heterogeneous, then the previous methods of GLS (for linear models) and GNLS (for nonlinear models) can be used. However, if the errors are nonnormally distributed, then an alternative method must be used. Fortunately, we can analyze just about any customized model using maximum likelihood (or Bayesian) estimation methods. Indeed, it is rather simple to incorporate changes in variance into an ecological model. All we have to do is define a sensible model that describes how a variance parameter changes as a function of one or more predictor variables. In the example shown here, the data show a typical triangular, or "factor-ceiling" or "limiting factor" profile of many ecological data sets. The triangular distribution is often caused by an environmental variable that sets an upper limit on an ecological response rather than determining its precise value. In this example, the data represent counts and thus a poisson or negative binomial error distribution is appropriate. Here, I fit a negative binomial model with several options: 1) mean as a either a linear function of *x* or an exponentially decreasing function of *x*, and the overdispersion parameter *k* equal to a constant; 2) mean and *k* as exponential functions of *x*; and 3) mean as a constant and *k* as an exponential function of *x*. The results shown here indicate that the third option is the best.

Note, the same general strategy applies for the variance parameter of other error distributions such as the variance of a normal distribution, the shape parameter of the Gamma distribution, or the overdispersion parameter of the beta-binomial distribution. Just as with deterministic models for the mean value, the variance might differ among different groups or treatment levels, might change as a function of a continuous covariate as in the example above, or might depend on the interactions of factors and covariates (i.e., different dependence of variance on the covariate in different groups). Just the variance, or both the mean and the variance, could differ among groups. There are no bounds the variations that can be analyzed, and the Likelihood Ratio Tests (LRT) or AIC values can be used to choose among the alternative models.