

Ecological Statistics

Review of Linear Regression

Non-linear module – part 1

C. Staudhammer - Eco Stats (fall 2008)

Regression in ecological work

- Often used to evaluate and/or characterize associations between two continuous variables
- Common in mensurative (empirical observation) studies
- Used to estimate parameters for population models, e.g., growth, survival

C. Staudhammer - Eco Stats (fall 2008)

Outline

- Simple linear regression
 - The least squares method
 - Estimation of parameters and confidence intervals
 - Evaluating goodness of fit
 - Assumptions
- Multiple linear regression
 - Estimation of parameters and confidence intervals
 - Evaluating goodness of fit
 - Additional assumptions and issues
 - Multicollinearity
- Transformations
- Model selection

C. Staudhammer - Eco Stats (fall 2008)

Simple Linear Regression

- Method of finding the “best” linear relationship between x and y

$$E(y_i|x_i) = \beta_0 + \beta_1 x_i$$

Model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ (population)

$y_i = b_0 + b_1 x_i + e_i$ (sample)

Predictions: $\hat{y}_i = b_0 + b_1 x_i$

Errors: $e_i = y_i - \hat{y}_i$

C. Staudhammer - Eco Stats (fall 2008)

Least Squares Method

- To find regression coefficients, minimize sum of squared errors:

$$\min\left(\sum_{i=1}^n e_i^2\right) = \min\left(\sum_{i=1}^n (y_i - (b_0 + b_1 x_i))^2\right)$$

- How do we find minimum?
 - Take partial derivatives w.r.t. b_0, b_1 , set them equal to zero, and solve for b_0, b_1

C. Staudhammer - Eco Stats (fall 2008)

The “normal” equations

$$b_0 = \frac{1}{n} \sum_{i=1}^n y_i - b_1 \frac{1}{n} \sum_{i=1}^n x_i = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n} (\sum x_i)^2} = \frac{SP_{xy}}{SS_x}$$

What are the sums of squares corrected for?

- * $\sum y_i^2$ is called the uncorrected sum of squares of Y; SS_y is the corrected sum of squares of Y
- * $\sum x_i y_i$ is called the uncorrected sum of products vs. SP_{xy} is the corrected sum of products
- * $SP_{xy}/(n-1) = \text{COV}(x, y)$
- It measures the joint variation between x and y
- If $\text{COV}(x, y) = 0$, then x and y are independent

What kind of relationship do we have if $\text{COV}(x, y) = 0$?

C. Staudhammer - Eco Stats (fall 2008)

Properties of a least squares regression line

- Always passes through (\bar{x}, \bar{y})
- Sum of residuals is zero, i.e., $\sum e_i = 0$
- Sum of squared residuals is at a minimum

Even though these properties always hold for a least squares regression line, this does not mean our relationship is good... How do we tell if it is good?!

C. Staudhammer - Eco Stats (fall 2008)

Partitioning the sums of squares

$$SS_Y = \sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2$$

$$= \sum_{i=1}^n (y_i - \bar{y})^2$$

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

$$SS_Y = SS_{RES} + SS_{REG}$$

C. Staudhammer - Eco Stats (fall 2008)

Graphic representation of the partitioning of the sums of squares

C. Staudhammer - Eco Stats (fall 2008)

Sampling distributions of model parameter estimates

Model: $\mu_{y|x_i} = \beta_0 + \beta_1 x_i$

Predictions: $\hat{y}_i = b_0 + b_1 x_i$

b_0	b_1	\hat{y}_k
b_{01}	b_{11}	\hat{y}_{k1}
b_{02}	b_{12}	\hat{y}_{k2}
b_{03}	b_{13}	\hat{y}_{k3}
...
b_{01000}	b_{11000}	\hat{y}_{k1000}

To get the sampling distribution:
 - Take a sample of ten (x_i, y_i) pairs.
 - Compute b_0, b_1 , and the predictions (\hat{y}_i)
 - Do this 1000 times...

For example, I generated data following the model for tree heights:
 $height_i = -4.0 + 0.7 \cdot diameter_i + e_i$
 where: e_i randomly distributed $\sim N(0, 3)$

C. Staudhammer - Eco Stats (fall 2008)

Results of simulation

b_0	b_1
-2.01	0.766
-2.12	0.684
-3.08	0.912
-4.44	0.715
...	...
-5.12	0.560

$\bar{b}_0 = -4.7938, \bar{b}_1 = 0.675$

Note: \hat{y}_i, b_0 , and b_1 are random variables

C. Staudhammer - Eco Stats (fall 2008)

Properties of the LS Estimates

- Expected values:
 - $E(\hat{y}|x_k) = \mu_{y|x_k}$
 - $E(b_0) = \beta_0$
 - $E(b_1) = \beta_1$
- Standard errors:
 - Intercept: $S_{b_0} = \sqrt{\frac{MS_{RES} \sum x_i^2}{n \cdot SS_x}}$
 - Slope: $S_{b_1} = \sqrt{MS_{RES} / SS_x}$
- Estimated y values: $S_{\hat{y}|x_k} = \sqrt{MS_{RES} \left(\frac{1}{n} + \frac{(x_k - \bar{x})^2}{SS_x} \right)}$

If the assumptions of regression are met, \hat{y}_{x_k}, b_0, b_1 are $\sim N$

C. Staudhammer - Eco Stats (fall 2008)

CI for predicted y values

- Where are we most certain about our predictions, i.e., where are we most precise?

C. Staadhämmer - Eco Stats (fall 2008)

Testing the Hypothesis that there is a linear relationship between X and Y

We want to know how much of the variance can be explained by the regression line. We test to see if the variance due to the regression is bigger than that of the residuals:

Hypothesis: $H_0: \sigma^2_{REG} / \sigma^2_{RES} = 1$
 $H_1: \sigma^2_{REG} / \sigma^2_{RES} > 1$

Test: $F_{(1, n-2)} = \frac{MS_{REG}}{MS_{RES}} = \frac{SS_{REG} / (1)}{SS_{RES} / (n-2)}$

Is this a one or two-tailed test??

C. Staadhämmer - Eco Stats (fall 2008)

Hypothesis testing for the intercept

- Is the intercept necessary?
 $H_0: \beta_0 = 0$
 $H_1: \beta_0 \neq 0$
- Level of significance: α
- Critical Value: $\pm t(\alpha/2; n-2)$
- Test statistic: $t_{(n-2)} = \frac{b_0 - 0}{S_{b_0}}$
- Decision: Do not reject null if $t_{(n-2)}$ is in the interval: $[-t_{(\alpha/2; n-2)}, +t_{(\alpha/2; n-2)}]$ Otherwise, reject

So why then does SAS give us an F-test???
 $t_{(\alpha/2; n-2)}^2 = F_{(\alpha; 1, n-2)}$

C. Staadhämmer - Eco Stats (fall 2008)

Hypothesis testing for the slope

- Is the slope necessary?
 $H_0: \beta_1 = 0$
 $H_1: \beta_1 \neq 0$
- Level of significance: α
- Critical Value: $\pm t(\alpha/2; n-2)$
- Test statistic: $t_{(n-2)} = \frac{b_1 - 0}{S_{b_1}}$
- Decision: Do not reject null if $t_{(n-2)}$ is in the interval: $[-t_{(\alpha/2; n-2)}, +t_{(\alpha/2; n-2)}]$; otherwise, reject
- Note: this is equivalent to testing $\sigma^2_{REG} / \sigma^2_{RES} = 1$ because: $(t_{(\alpha/2; n-2)})^2 = F_{(\alpha; 1, n-2)}$

C. Staadhämmer - Eco Stats (fall 2008)

How do we measure how good our regression is?

- Coefficient of determination = r^2
 $r^2 = \frac{SS_{REG}}{SS_Y} = 1 - \frac{SS_{RES}}{SS_Y}$
- $0 \leq r^2 \leq 1$
- r^2 is the proportion of variation of y (in terms of sums of squares) that is due to the independent variable; or, the proportion of change in y explained by the independent variable.

C. Staadhämmer - Eco Stats (fall 2008)

How do we measure how good our regression is? - 2

- Correlation coefficient = r
 $r = \frac{SP_{XY}}{\sqrt{SS_X \cdot SS_Y}}$
- $-1 \leq r \leq 1$ -- it shows direction of relationship
- r can be used to test the significance of the regression, using a special r table. F and r tests are equivalent!.

What does my relationship look like if $r < 0$?

C. Staadhämmer - Eco Stats (fall 2008)

How do we measure how good our regression is? - 3

- Root mean square error (RMSE) – $S_{y \cdot x}$

$$S_{y \cdot x} = SE_e = \sqrt{MS_{RES}} = \sqrt{\frac{SS_{RES}}{n-2}}$$
- $S_{y \cdot x}$ measures the variation of the individual points around the regression line.
- Within ± 1 $S_{y \cdot x}$ of the regression line, we can expect to find 68% of the observations (what assumption do we need for this?)
- Sometimes called square root residual variance, or Standard error of the estimate

C. Staudhammer - Eco Stats (fall 2008)

Assumptions for regression analysis

1. The model is correctly specified, i.e., $\mu_{y|x_i} = \beta_0 + \beta_1 x_i$

For instance, I fit this model

Whereas, I should have fit this model

How can we avoid this?

C. Staudhammer

Assumptions for regression analysis - 2

2. For any given x_i , $y_i \sim N(\mu_{y|x_i}, \sigma^2_{y|x_i})$

C. Staudhammer - Eco Stats (fall 2008)

Assumptions for regression analysis - 3

3. The y_i are homoscedastic, e.g., we don't have this...

- We could test for this

C. Staudhammer - Eco Stats (fall 2008)

Assumptions for regression analysis - 4

- Any observations of y_i is independent from all other observations of y_i
- For a given x_i , the observations of y_i are randomly selected.
- The values of x_i are fixed and/or measured without error.

- Note: The most important assumptions for testing are #2 and #3.

C. Staudhammer - Eco Stats (fall 2008)

Multiple Linear Regression

Y predicted by a series of p regressor variables

$$E(y_i|x_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi}$$

Model: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + \varepsilon_i$ (pop'n)

$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \dots + b_p x_{pi} + e_i$ (sample)

Predictions: $\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \dots + b_p x_{pi}$

Errors: $e_i = y_i - \hat{y}_i$

C. Staudhammer - Eco Stats (fall 2008)

Estimation of parameters

- Least squares estimation can be done using the same procedure as in SLR (minimize the sum of squared errors)
 - Take partial derivatives with respect to each coefficient and set each equal to zero
 - Solve the system of simultaneous equations
 - SAS procedure PROC REG uses this method, as does Excel
- Other estimation techniques include Maximum Likelihood estimation

C. Staadhammer - Eco Stats (fall 2008)

Evaluating Goodness of Fit

- Multivariate r^2 is denoted R^2 and calculated exactly the same way:

$$R^2 = \frac{SS_{REG}}{SS_Y} = 1 - \frac{SS_{RES}}{SS_Y}$$
- Multivariate r is denoted R and is calculated as the positive square root of R^2

Note: As independent variables are added to the regression SS_{REG} cannot decrease. Therefore, more variables will always result in a greater or equal R^2 .

C. Staadhammer - Eco Stats (fall 2008)

Evaluating Goodness of Fit – 2

- The “Adjusted R^2 ” attempts to correct for the insensitivity of R^2 to the number of independent variables by adjusting both the numerator and the denominator by their respective degrees of freedom:

$$R_{adj}^2 = 1 - \frac{(n-1)(1-R^2)}{n-p-1}$$

C. Staadhammer - Eco Stats (fall 2008)

Evaluating Goodness of Fit – 2

- Akaike's Information Criteria

$$AIC = n \ln \left(\frac{SS_{RES}}{n} \right) + 2(p+1)$$

Sometimes written:
 $AIC = -2\ell + 2(p+1)$
 where: ℓ is the log-likelihood

- Bayesian Information Criteria

$$BIC = n \ln \left(\frac{SS_{RES}}{n} \right) + (p+1) \ln(n)$$

Sometimes written:
 $BIC = -2\ell + (p+1)\ln(n)$

- Mallor's C_p

$$C_p = \frac{SS_{RES}}{\hat{\sigma}^2} + 2(p+1) - n$$

where:
 $\hat{\sigma}^2$ is the residual MS after regression with all predictor variables

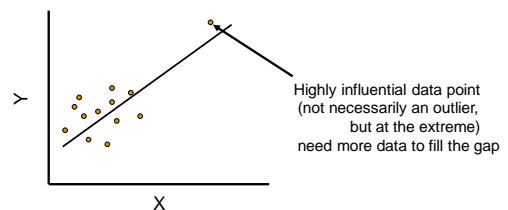
C. Staadhammer - Eco Stats (fall 2008)

Additional issues in MLR

- Estimation requires matrix algebra
 - A linear combination of x variables may collectively predict y
 - You cannot adequately evaluate the fit using one-way plots of y on each x
 - The removal of one x variable could render another as important
- *To aid in model evaluation, analyze residuals and compute influence diagnostics
- If two or more explanatory variables are highly correlated, you will have multicollinearity

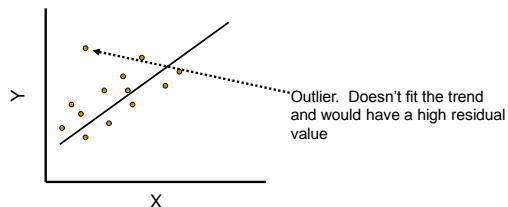
C. Staadhammer - Eco Stats (fall 2008)

Influence Diagnostics



C. Staadhammer - Eco Stats (fall 2008)

Influence Diagnostics - 2



C. Staudhammer - Eco Stats (fall 2008)

Influence Diagnostics - 3

- Influence diagnostics explore how each data point is influencing the MLR
 - In MLR, can't use graphs to "see" the influential points
 - Even when residuals are low, influence can be large on parameter estimates
- Statistical options for evaluating influence in SAS procedure PROC REG
 - INFLUENCE option in MODEL statement: DFFITS, DFBETAS

C. Staudhammer - Eco Stats (fall 2008)

Multicollinearity

- A high degree of linear correlation amongst two or more explanatory variables makes it difficult to separate their effects on the dependent variable
 - Indicated by highly correlated x's
 - e.g.: use of human population size and nutrient discharge to predict impacts to streams
- Linear dependencies among x variables render the model more uncertain
 - May inflate or deflate standard errors around parameter estimates
- Does not matter for prediction, but can greatly influence the interpretation of parameters
- To test, use INFLUENCE option in MODEL statement of SAS procedure PROC REG with VIF keyword

C. Staudhammer - Eco Stats (fall 2008)

Multicollinearity - 2

Variance Inflation Factor (VIF):

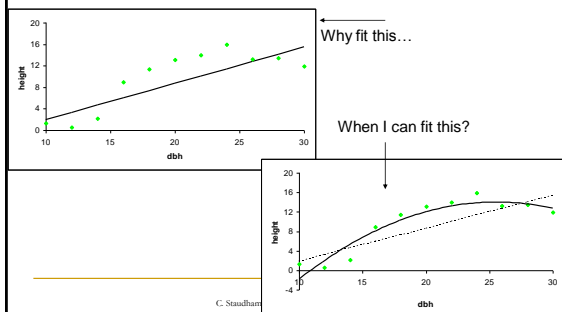
$$VIF = \frac{1}{1 - R_i^2}$$

where: R_i^2 = The R^2 by regressing the set of $p-1$ independent variables against the i th

*VIF < 10 is often considered not too severe (but see Graham, M. H. 2003. Confronting multicollinearity in ecological multiple regression. *Ecology* 84: 2809–2815)

Fall 2007, C. Staudhammer

What if my relationship is **not** linear?



C. Staudham

Why transform variables in regression?

- To meet assumptions of regression
 - To achieve normality
 - To stabilize the variance
 - To obtain a linear relationship between independent and dependent variables

C. Staudhammer - Eco Stats (fall 2008)

What is the impact of transforming variables in regression?

- When units of x changed, e.g., $y_i = b_0 + b_1 \ln(x_i)$
 - Regression coefficients are applied to the transformed value of x
 - No impact on tests, fit statistics
- Units of y are changed, e.g., $\ln(y_i) = b_0 + b_1 \ln(x_i)$
 - Predictions are in the transformed units
 - Parameter estimates and tests of significance are valid for the transformed model**
 - Fit statistics are impacted substantially
 - R^2 as automatically computed is the percentage of variation in the transformed variable that is explained by the x variables
 - RMSE is in transformed units

→ It is inappropriate to compare R^2 from equations of transformed data to those of un-transformed

C. Staadhammer - Eco Stats (fall 2008)

Example

We fit $\sqrt{\text{Height}}$ versus dbh for 20 Cedar trees -we might feel good about this relationship, as our transformed dependent variable is $\sim N$ and it fits the data better than a linear equation

- But... what is this R-squared (0.8996) measuring?

DBH	sqrt(ht)
41.5	5.19
17.4	4.76
32.8	4.90
20.5	4.74
44.3	5.63
10.7	2.83
10.1	2.83
12.1	3.32
8.1	2.35
4.5	1.67
27.4	5.21
32.2	4.69
52.8	5.26
53.5	5.68
63.6	5.39
42.8	5.05
64.2	6.04
50.6	6.02
36.1	5.59
36.4	5.59

C. Staadhammer - Eco Stats (fall 2008)

Example - 2

Examine the regression output from an un-transformed Height model (dbh and dbh-squared are independent variables)

SUMMARY OUTPUT - Height

Regression Statistics	
Multiple R	0.9427
R Square	0.8887
Adjusted R Square	0.8756
Standard Error	3.6764
Observations	20

ANOVA					
	df	SS	MS	F	Significance F
Regression	2	1837.0	918.5	67.9	0.0000
Residual	17	230.0	13.5		
Total	19	2067.1			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-1.9543	2.7097	-0.7212	0.4806	-7.6712	3.7626
DBH	1.2438	0.1848	6.7318	0.0000	0.8540	1.6336
dbh^2	-0.0113	0.0027	-4.2106	0.0006	-0.0169	-0.0056

C. Staadhammer - Eco Stats (fall 2008)

Example - 3

Versus... the regression output from the sqrt(Height) transformed model (dbh and dbh-squared are independent variables)

SUMMARY OUTPUT - Sqrt(Height)

Regression Statistics	
Multiple R	0.9485
R Square	0.8996
Adjusted R Square	0.8878
Standard Error	0.4348
Observations	20

ANOVA					
	df	SS	MS	F	Significance F
Regression	2	28.8	14.4	76.1	0.0000
Residual	17	3.2	0.2		
Total	19	32.0			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	1.3547	0.3203	4.2299	0.0006	0.6790	2.0304
DBH	0.1716	0.0218	7.8571	0.0000	0.1255	0.2177
dbh^2	-0.0017	0.0003	-5.3008	0.0001	-0.0023	-0.0010

C. Staadhammer - Eco Stats (fall 2008)

Comparing model fit statistics with the back-transformation method

- Predictions are transformed *back* into the original units by using the inverse function
 - E.g., if you predict $\sqrt{\text{height}}$, change these predictions into height by squaring them
- Using the predictions in original units, compute residuals in original units
 - Then compute SS_{res} , to get:
 - RMSE = $\sqrt{MS_{res}}$, R-squared = $(1 - SS_{res}/SS_y)$
- This intuitive method is often used in forestry, but its statistical properties have *not* been investigated

C. Staadhammer - Eco Stats (fall 2008)

Example - 4

DBH	sqrt(ht)	pred(sqrt(ht))	res(sqrt(ht))
41.5	5.19	5.59	-0.40
17.4	4.76	3.83	0.93
32.8	4.90	5.18	-0.28
20.5	4.74	4.17	0.58
44.3	5.63	5.67	-0.04
10.7	2.83	3.00	-0.17
10.1	2.83	2.92	-0.09
12.1	3.32	3.19	0.13
8.1	2.35	2.63	-0.29
4.5	1.67	2.09	-0.42
27.4	5.21	4.80	0.41
32.2	4.69	5.14	-0.45
52.8	5.26	5.74	-0.48
53.5	5.68	5.74	-0.05
63.6	5.39	5.49	-0.09
42.8	5.05	5.63	-0.58
64.2	6.04	5.46	0.58
50.6	6.02	5.75	0.28
36.1	5.59	5.36	0.23
36.4	5.59	5.38	0.21

Ssres	3.21
Msres	0.189
RMSE	0.435
Ssy in sqrt units	32.00
rsq	89.96%

Untransformed RMSE = $\sqrt{0.189}$

Untransformed R-sq = $1 - SS_{res}/SS_y$

C. Staadhammer - Eco Stats (fall 2008)

Example - 5

- We add on to this worksheet, back-transforming our predictions into *original units*, and then computing residuals, RMSE, and R-sq in *original units*

DBH	HEIGHT	sqrt(ht)	pred(7ht)	res(7ht)	BACK-transformed	
					pred(ht)	res(ht)
41.5	26.9	5.19	5.59	-0.40	31.24	-4.34
17.4	22.7	4.76	3.93	0.83	14.69	8.01
32.8	24	4.90	5.18	-0.28	26.83	-2.83
20.5	22.5	4.74	4.17	0.58	17.37	5.13
44.3	31.7	5.63	6.67	-0.04	32.12	-0.42
10.7	8	2.83	3.00	-0.17	8.99	-0.99
10.1	8	2.83	2.92	-0.09	8.51	-0.51
12.1	11	3.32	3.19	0.13	10.15	0.85
8.1	5.5	2.35	2.63	-0.29	6.94	-1.44
4.5	2.8	1.67	2.09	-0.42	4.38	-1.58
27.4	27.1	5.21	4.80	0.41	23.02	4.08
32.2	22	4.69	5.14	-0.45	26.44	-4.44
52.8	27.7	5.26	6.74	-0.48	32.98	-5.28
53.5	32.3	5.68	6.74	-0.05	32.92	-0.62
63.6	29.1	5.39	5.49	-0.09	30.13	-1.03
42.8	25.5	5.05	5.63	-0.58	31.68	-6.18
64.2	36.5	6.04	6.46	0.58	29.85	6.55
50.6	36.3	6.02	5.75	0.28	33.02	3.28
36.1	31.3	5.59	6.36	0.23	28.78	2.52
36.4	31.3	5.59	6.38	0.21	28.94	2.36

Sstres	3.21	Sstres	294.26
Mstres	0.189	Mstres	17.309
RMSE	0.435	RMSE	4.160
Ssy in sqrt units	32.00	Ssy in orig units	2067.06
r-sq	89.96%	r-sq	89.76%

C. Staudhammer - Eco Stats (fall 2008)

Take home message for transformations

- When you transform the dependent variable, model results are not directly comparable to those of non-transformed models
 - This is the reason why non-linear regression is recommended over transformation!
- While the back-transformation method is used in forestry applications, it needs further study to evaluate its statistical properties

C. Staudhammer - Eco Stats (fall 2008)

Model Selection

- More x variables cause higher R^2 , but not necessarily a better model
- Strive for model parsimony
 - Optimal model has a low number of parameters, precise parameter estimates, and good prediction confidence
 - Overfitting is a concern

Why?

C. Staudhammer - Eco Stats (fall 2008)

Selection Procedures

- Lots of options (not specific to MLR)
- AIC used often for a variety of models, with some debate
- Objectivity is the goal of model selection procedures
- Canned Procedures
 - Forward
 - Backward
 - Stepwise

C. Staudhammer - Eco Stats (fall 2008)

Forward Selection

- First X chosen to maximize the R^2
- Second X chosen to produce the largest increase in R^2 in the presence of X_1
- Procedure terminates when no additional X values improve the R^2 to a desired level, indicated by the p value

C. Staudhammer - Eco Stats (fall 2008)

Stepwise Selection

- All regressors considered for addition or removal at each step, to maximize the R^2 and the significance of the model
- Pre-selected p_{in} and p_{out}
- Final model chosen after all are considered in the presence of the other X variables

C. Staudhammer - Eco Stats (fall 2008)

Backward Selection

- All X variables entered into model
- Systematically removed if not significant
- Procedure terminates when no additional X values can be removed (i.e., all are significant at the specified p value)

C. Staudhammer - Eco Stats (fall 2008)

Take home messages

- Regression is a very powerful and very common method for describing associations between two (or more) continuous variables
- Always be mindful of the assumptions of regression
- Model selection and fitting involves both statistical tests and common sense evaluation of how your model relates to the data

C. Staudhammer - Eco Stats (fall 2008)