Lecture 10 - Multiple Regression

Bill Perry

# Lecture 09: Review

Covered

* Regression T-Test Anova
* Regression Assumptions
* Model II Regression

# Lecture 10: Overview

Multiple Linear Regression model

* Regression parameters
* Analysis of variance
* Null hypotheses
* Explained variance
* Assumptions and diagnostics
* Collinearity
* Interactions
* Dummy variables
* Model selection
* Importance of predictors

# **Lecture 10:** Analyses

What if more than one predictor (X) variable?

* If predictors continuous
* Mix between categorical and continuous
* Can use multiple linear regression

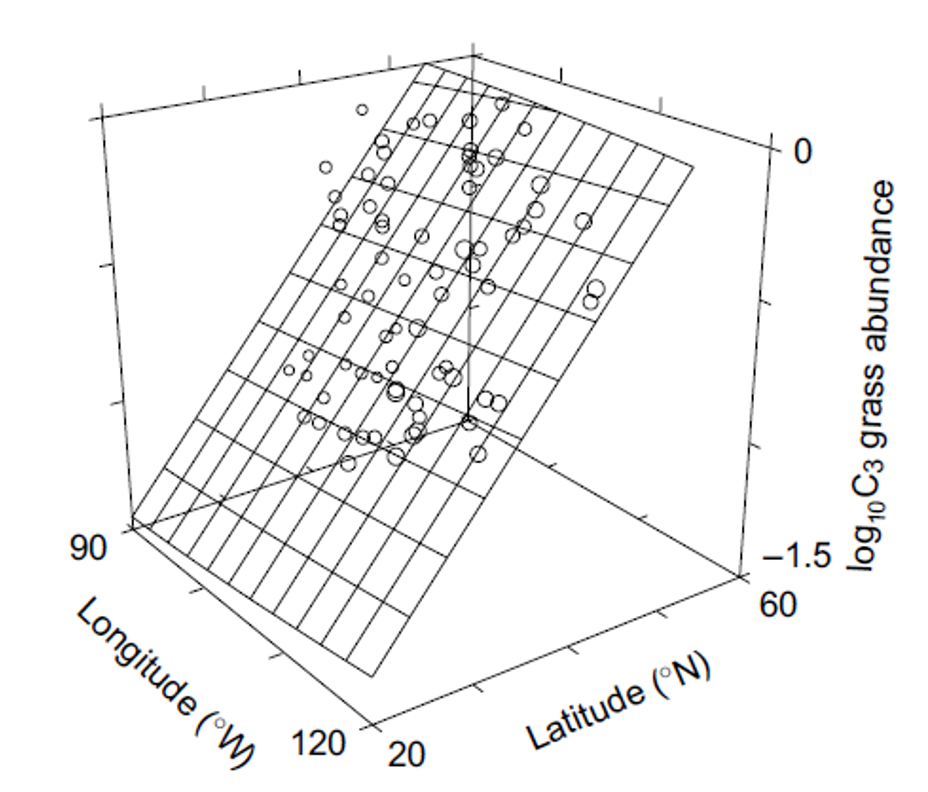
|  | Independent variable |  |
| --- | --- | --- |
| **Dependent variable** | **Continuous** | **Categorical** |
| **Continuous** | Regression | ANOVA |
| **Categorical** | Logistic regression | Tabular |

# **Lecture 10:** Analyses

Abundance of C3 grasses can be modeled as function of

* latitude
* longitude
* both

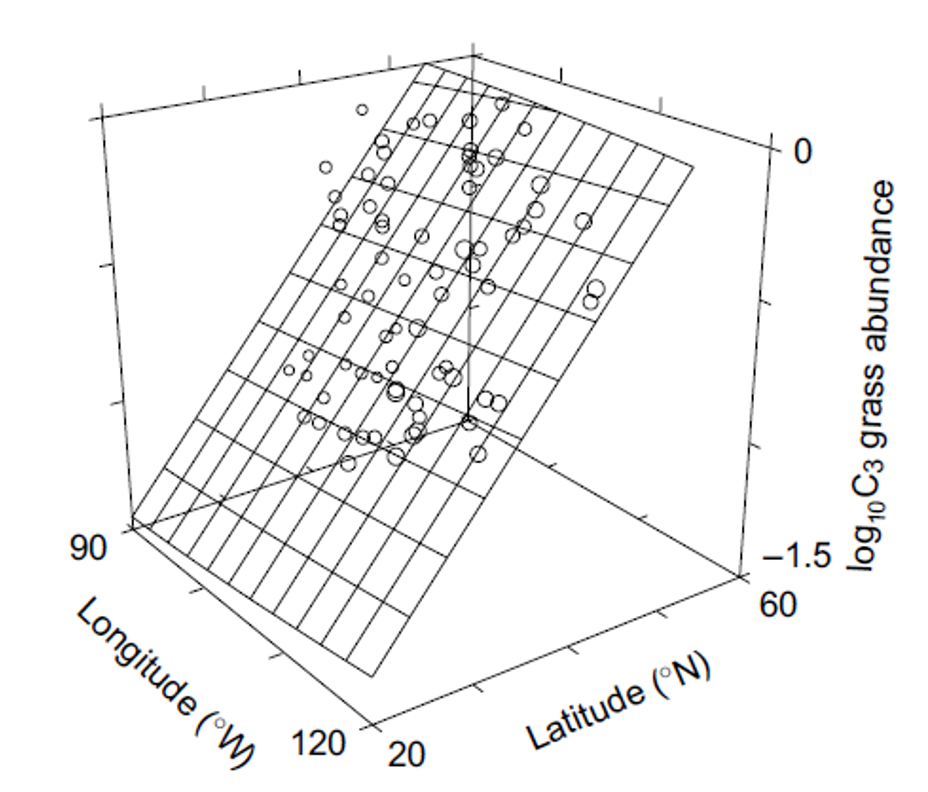
Instead of line, modeled with (hyper)plane



# **Lecture 10:** Analyses

Used in similar way to simple linear regression:

* Describe nature of relationship between Y and X’s
* Determine explained / unexplained variation in Y
* Predict new Ys from X
* Find the “best” model

S

# **Lecture 10:** Analyses

Crawley 2012: “Multiple regression models provide some of the most profound challenges faced by the analyst”:

* Overfitting
* Parameter proliferation
* Multicollinearity
* Model selection



# **Lecture 10:** Analyses

Multiple Regression:

* Set of i= 1 to n observations
* fixed X-values for p predictor variables (X1, X2…Xp)
* random Y-values:
* yi: value of Y for ith observation X1 = xi1, X2 = xi2,…, Xp = xip
* β0: population intercept, the mean value of Y when X1 = 0, X2 = 0,…, Xp = 0

# **Lecture 10:** Multiple linear regression model

Multiple Regression:

* β1: partial population slope, change in Y per unit change in X1 holding other X-vars constant
* β2: partial population slope, change in Y per unit change in X2 holding other X-vars constant
* βp: partial population slope, change in Y per unit change in Xp holding other X-vars constant

# **Lecture 10:** Regression parameters

Multiple Regression:

* εi: unexplained error - difference bw yi and value predicted by model (ŷi)
* NPP = β0 + β1(lat) + β2 (long) + β3 (soil fertility) + εi

# **Lecture 10:** Regression parameters

Multiple Regression:

* Estimate multiple regression parameters (intercept, partial slopes) using OLS to fit the regression line
* OLS minimize ∑(yi-ŷi)2, the SS (vertical distance) between observed yi and predicted ŷi for each xij
* ε estimated as residuals: εi = yi-ŷi
* Calculation solves set of simultaneous normal equations with matrix algebra

# **Lecture 10:** Regression parameters

Regression equation can be used for prediction by subbing new values for predictor (X) variables

* Confidence intervals calculated for parameters
* Confidence and prediction intervals depend on number of observations and number of predictors
  + More observations decrease interval width
  + More predictors increase interval width
* Prediction should be restricted to within range of X variables

# **Lecture 10:** Analyses of variance

Variance - SStotal partitioned into SSregression and SSresidual

* SSregression is variance in Y explained by model
* SSresidual is variance not explained by model

| Source of variation | SS | df | MS | Interpretation |
| --- | --- | --- | --- | --- |
| Regression |  |  |  | Difference between predicted observation and mean |
| Residual |  |  |  | Difference between each observation and predicted |
| Total |  |  |  | Difference between each observation and mean |

# **Lecture 10:** Analyses

SS converted to non-additive MS (SS/df)

* MSresidual: estimate population variance
* MSregression: estimate population variance + variation due to strength of X-Y relationships
* MS do not depend on sample size

| Source of variation | SS | df | MS |
| --- | --- | --- | --- |
| Regression |  |  |  |
| Residual |  |  |  |
| Total |  |  |  |

# **Lecture 10:** Hypotheses

Two Hos usually tested in MLR:

* “Basic” Ho: all partial regression slopes equal 0; β1 = β2 = … = βp = 0
* If “basic” Ho true, MSregression and MSresidual estimate variance and their ratio (F-ratio) = 1
* If “basic” Ho false (at least one β ≠ 0) MSregression estimates variance + partial regression slope and their ratio (F-ratio)
* will be > 1 - F-ratio compared to F-distribution for p-value

# **Lecture 10:** Hypotheses

Also: is any specific β = 0 (explanatory role)?

* E.g., does LAT have effect on NPP?
* These Hs tested through model comparison
* Model w 3 predictors X1, X2,X3 (model 1):
* yi= β0 +β1xi1+β2xi2+β3xi3+ εi
* To test Ho that β1 = 0 compare fit of model 1 to model 2:
* yi= β0 +β2xi2+β3xi3+ εi

# **Lecture 10:** Hypotheses

* If SSregression of mod1=mod2, cannot reject Ho β1 = 0
* If SSregression of mod1 > mod2, evidence to reject Ho β1 = 0
* SS for β1 is SSextraβ1 = Full SSregression - Reduced SSregression
* Use partial F-test to test Ho β1 = 0 :

Can also use t-test (R provides this value)

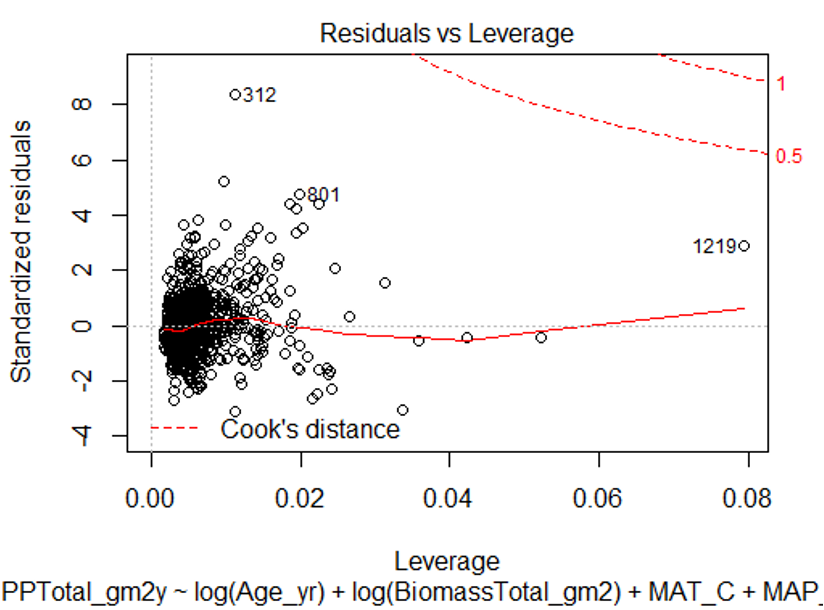
# **Lecture 10:** Explained variance

Explained variance (r2) is calculated the same way as for simple regression:

* r2 values can not be used to directly compare models
* r2 values will always increase as predictors added
* r2 values with different transformation will differ

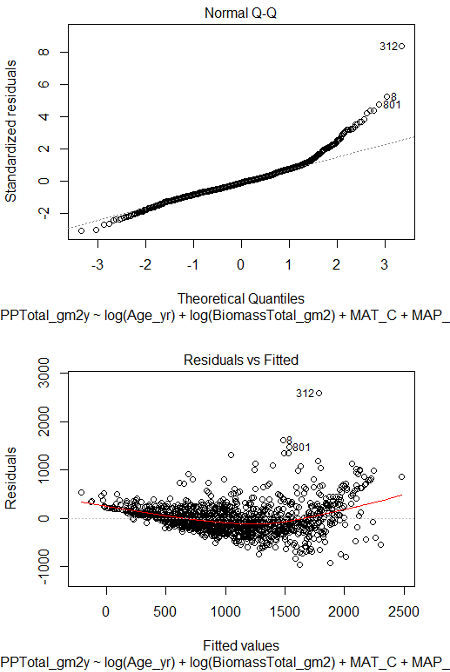
# **Lecture 10:** Assumptions and diagnostics

* Assume fixed Xs; unrealistic in most biological settings
* No major (influential) outliers
* Check leverage, influence- Cook’s Di



# **Lecture 10:** Assumptions and diagnostics

* Normality, equal variance, independence
* Residual QQ-plots, residuals vs. predicted values plot
* Distribution/variance often corrected by transforming Y



# **Lecture 10:** Assumptions and diagnostics

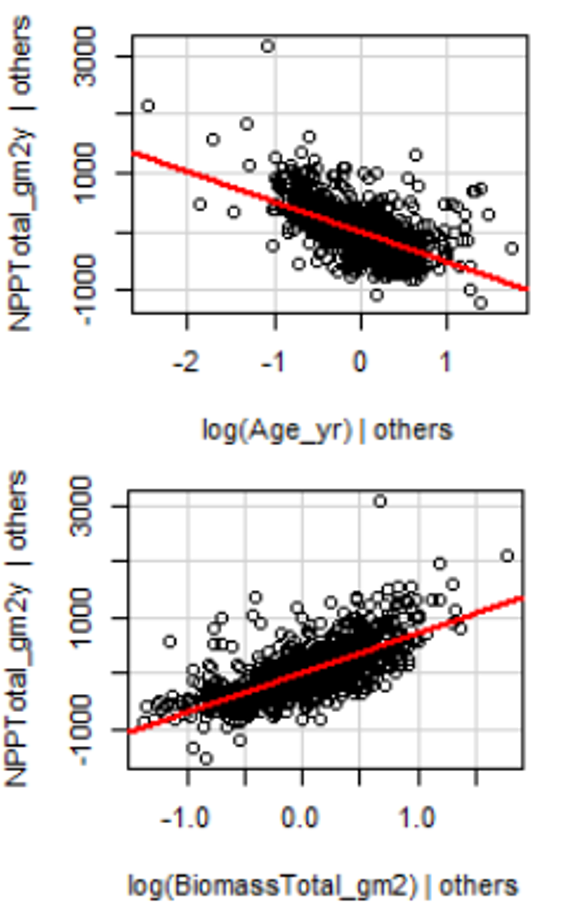
More observations than predictor variables

* Ideally at least 10x observations than predictors to avoid “overfitting”
* Uncorrelated predictor variables (assessed using scatterplot matrix; VIFs)
* Linear relationship between Y and each X, holding others constant (non-linearity assessed by AV plots)

# **Lecture 10:** Analyses

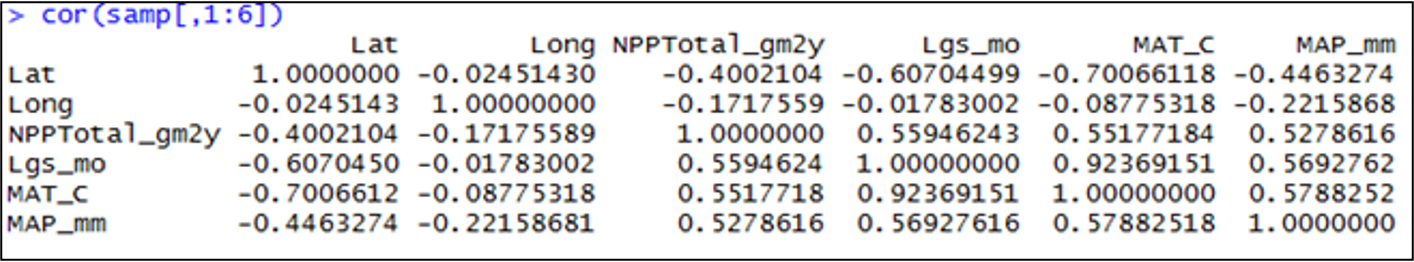
Regression of Y vs. each X does not consider effect of other predictors:

want to know shape of relationship while holding other predictors constant



# **Lecture 10:** Collinearity

* Potential predictor variables are often correlated (e.g., morphometrics, nutrients, climatic parameters)
* Multicollinearity (strong correlation between predictors) causes problems for parameter estimates
* Severe collinearity causes unstable parameter estimates: small change in a single value can result in large changes in βp - estimates
* Inflates partial slope error estimates, loss of power



# **Lecture 10:** Collinearity

Collinearity can be detected by:

* Variance inflation Factors:
  + VIF for Xj=1/ (1-r2 )
  + VIF > 10 = bad
* Best/simplest solution:
  + exclude variables that are highly correlated with other variables
  + they are probably measuring similar
  + thing and are redundant

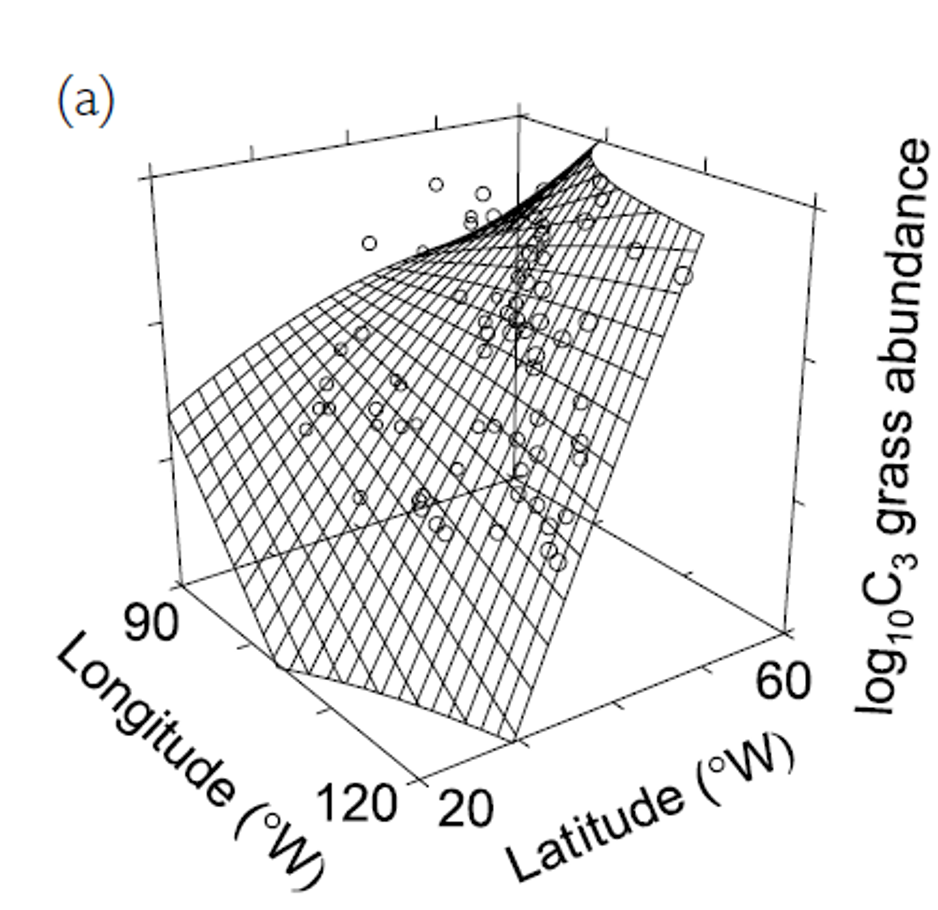
# **Lecture 10:** Interactions

Predictors can be modeled as:

* additive (effect of temp, plus precip, plus fertility) or
* multiplicative (interactive)
* Interaction: effect of Xi depends on levels of Xj
* The partial slope of Y vs. X1 is different for different levels of X2 (and vice versa); measured by β3

“Curvature” of the regression (hyper)plane

# **Lecture 10:** Analyses



# **Lecture 10:** Analyses

Adding interactions:

* many more predictors (“parameter proliferation”):
* 2n; 6 params= 64 terms; 7 params= 128
* interpretation more complex
* When to include interactions? When they make biological sense

# **Lecture 10:** Dummy variables

Multiple Linear Regression accommodates continuous and categorical variables (gender, vegetation type, etc.) Categorical vars as “dummy vars”, n of dummy variables = n-1 categories

**Sex M/F:**

* Need 1 dummy var with two values (0, 1)

**Fertility L/M/H:**

* Need 2 dummy var, each with two values (0, 1): fert1 (0 if L or H, 1 if M), fert2 (1 if H, 0 if L or M)

| Fertility | fert1 | fert2 |
| --- | --- | --- |
| Low | 0 | 0 |
| Med | 1 | 0 |
| High | 0 | 1 |

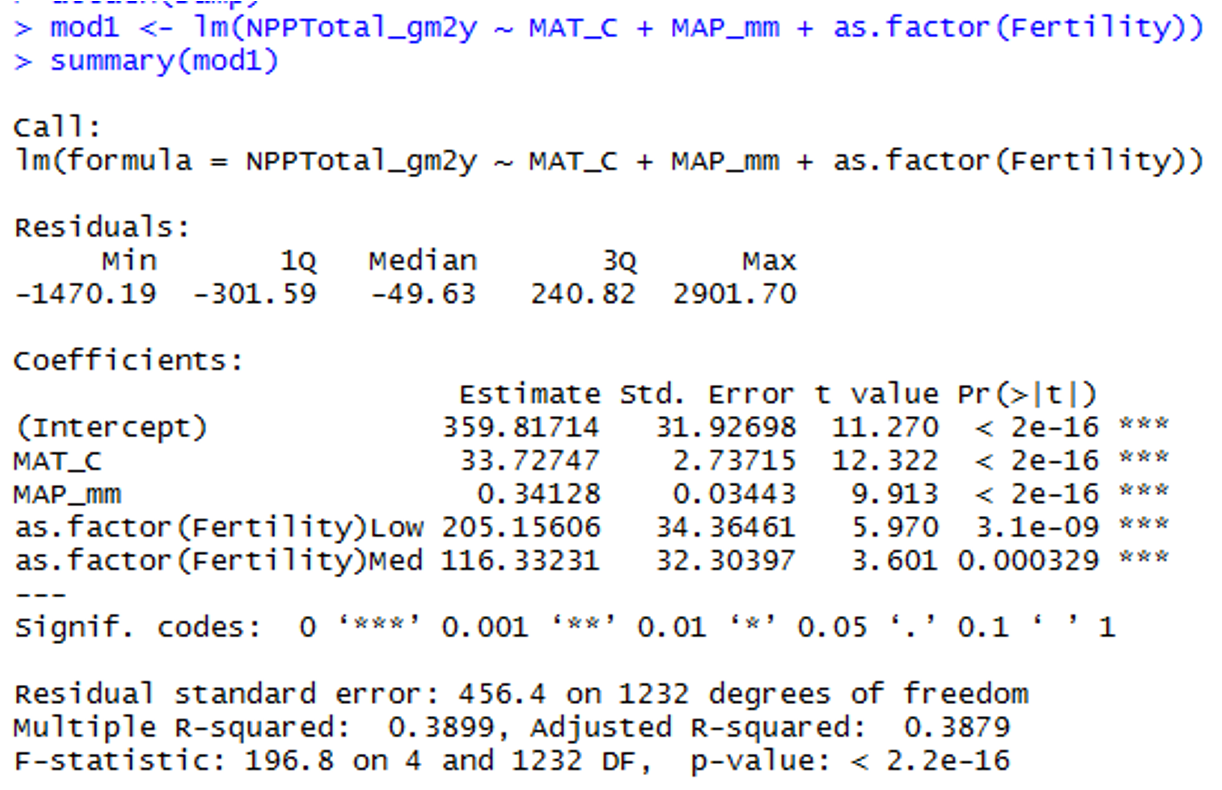
# **Lecture 10:** Analyses

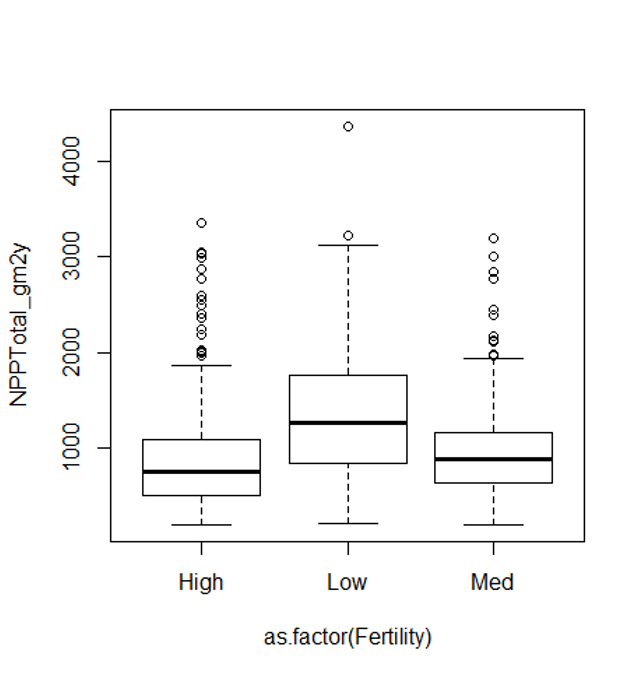
Coefficients interpreted relative to reference condition

* R codes dummy variables automatically
* picks “reference” level alphabetically
* Dummy variables with more than 2 levels add extra predictor variables to model

| Fertility | fert1 | fert2 |
| --- | --- | --- |
| Low | 0 | 0 |
| Med | 1 | 0 |
| High | 0 | 1 |

# **Lecture 10:** Analyses



S

# **Lecture 10:** Comparing models

When have multiple predictors (and interactions!)

* how to choose “best” model?
* Which predictors to include?
* Occam’s razor: “best” model balances complexity with fit to data

To chose:

* compare “nested” models

Overfitting

* getting high r2 just by having more (useless predictors)
* so r2 is not a good way of choosing between nested models

# **Lecture 10:** Comparing models

Need to account for increase in fit with added predictors:

* Adjusted r2
* Akaike’s information criterion (AIC)
* Both “penalize” models for extra predictors
* Higher adjusted r2 and lower AIC are better when comparing models

# **Lecture 10:** Comparing models

But how to compare models?

* Can fit all possible models
  + compare AICs or adj- r2,
  + tedious w lots of predictors
* Automated forward (and backward) stepwise procedures: start w no terms (all terms), add (remove) terms w largest (smallest)
  + partial F statistic

We will use manual form of backward selection

# **Lecture 10:** Analyses

# 

# **Lecture 10:** Predictors

Usually want to know relative importance of predictors to explaining Y

* Three general approaches:
* Using F-tests (or t-tests) on partial regression slopes
* Using coefficient of partial determination
* Using standardized partial regression slopes

# **Lecture 10:** Predictors

Using F-tests (or t-tests) on partial regression slopes:

* Conduct F tests of Ho that each partial regression slope = 0
* If cannot reject Ho, discard predictor
* Can get additional clues from relative size of F-values
* Does not tell us absolute importance of predictor (usually can not directly compare slope parameters)

# **Lecture 10:** Predictors

Using coefficient of partial determination:

* the reduction in variation of Y due to addition of predictor (Xj)

SSextra

* Increased in SSregression when Xj is added to model
* Reduced SSresidual is the unexplained SS from model without Xj

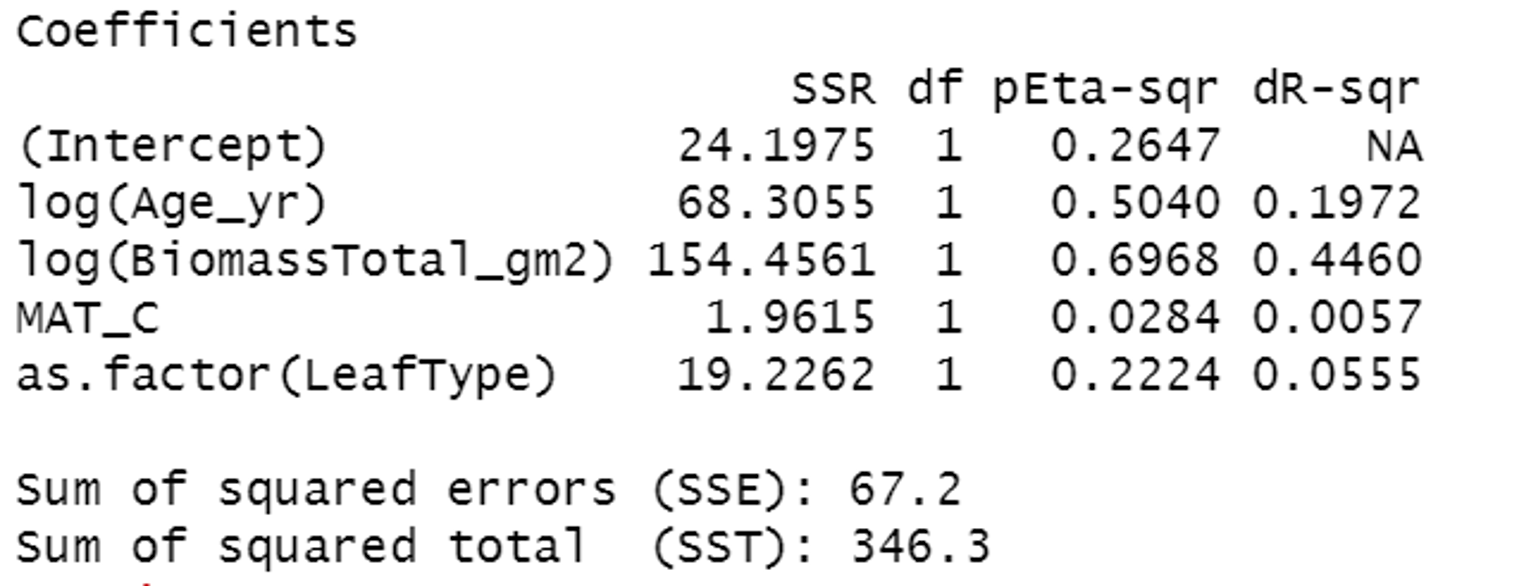
# **Lecture 10:** Predictors

Using standardized partial regression slopes:

* predictors of predictor variables can not be directly compared
* Why?
* Standardize all vars (mean = 0, sd= 1)
* Scales are identical and larger PRS mean more important variable

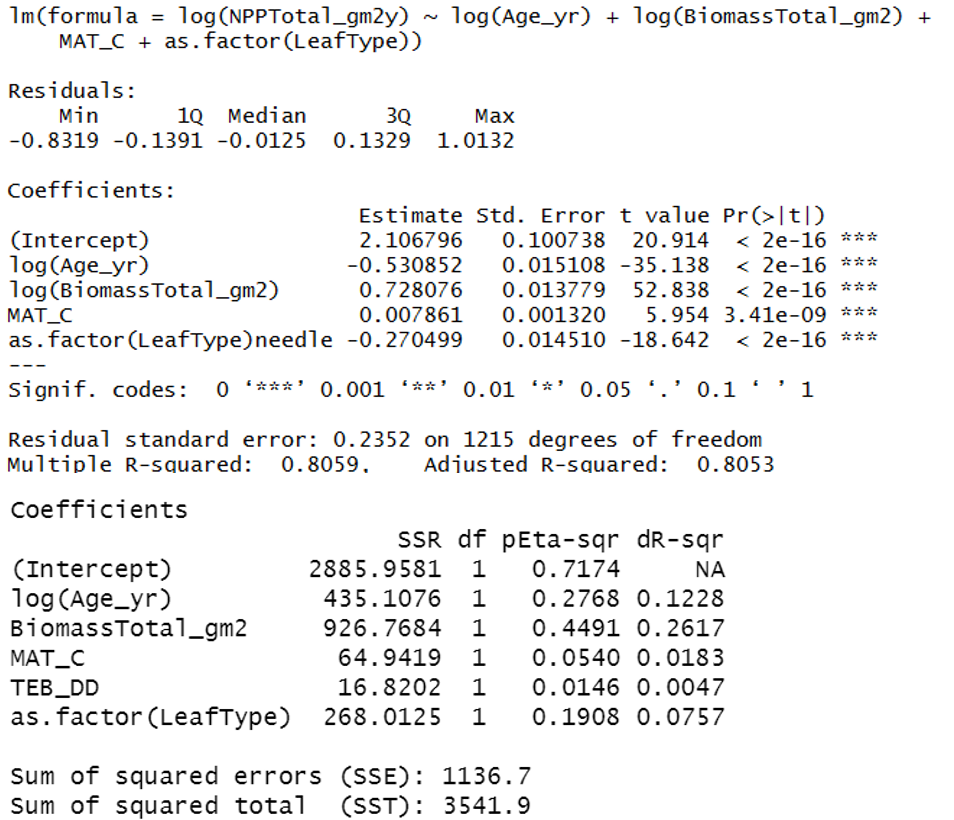
# **Lecture 10:** Predictors

Using partial r2 values:



# **Lecture 10:** Reporting results

Results are easiest to report in tabular format



# **Lecture 10:** Reporting results

Results are easiest to report in tabular format

