Strategies for Variable Selection

- 2 main reasons for reducing the number of explanatory variables in the model:
  - Philisophical: Simplicity is preferable to complexity
    - KISS principle
    - Occam's Razor
  - Practical: unnecessary terms yield less precise estimates, particularly if explanatory variables are highly correlated to each other

Loss of Precision

- Simple linear regression: $Var(\hat{\beta}_j) = \frac{\sigma^2}{(n-1)s_j^2}$
- Multiple regression: $Var(\hat{\beta}_j) = \frac{\sigma^2}{(n-1)s_j^2(1-R_j^2)}$
- s_j^2 is the sample variance of X_j
- R_j^2 is the R-squared term for the regression of X_j on all the other explanatory variables
- As R_j^2 increases variance is inflated! A problem when X's are highly correlated.

Multicollinearity

- Multicollinearity occurs when at least one of the explanatory variables can be predicted well from some combination of the other explanatory variables
- $(1 - R_j^2)$ also called tolerance is the proportion of X_j 's variance not explained by the other X variables
- Potential for problems when correlation among parameter estimates is higher than 0.9 in absolute value

Effects of Multicollinearity

- Inflated SE's of regression coefficients
  - coefficients expected to vary greatly across samples
  - wider confidence intervals
  - diminished ability of tests to detect significant effects
  - unexpected signs of coefficients:

    for 2 collinear X's that are both positively correlated with Y, one of the multiple regression coefficients may be negative!

Can estimate joint effects but not separate effects

Steps for Analysis with Many Variables

- specify objectives of analysis
- screen to eliminate redundant variables
- perform graphical exploratory analysis, transform to linearize
- fit a BIG model, check residuals, transform again if needed, screen for outliers and influential cases, refit if needed
- find a suitable subset of X's to answer objectives
- make inferences based on selected subset

Finding the "Best" subsets

Sequential Selection of Variables

- Forward - start small and add variables
- Backward - start big and remove variables that are not significant
- Stepwise (alternate forward and backwards steps)

F-to-enter and F-to-remove

River Nitrogen - Ex 12.21

- The rise in abundance of algae is thought to be due to the increase in nutrients such as nitrate and other forms of nitrogen
- It is theorized that excessive amounts of nitrate are due to human influences
- Researchers gathered data from around the world to gauge the evidence that nitrates in the discharge of rivers are associated with human population density

Data

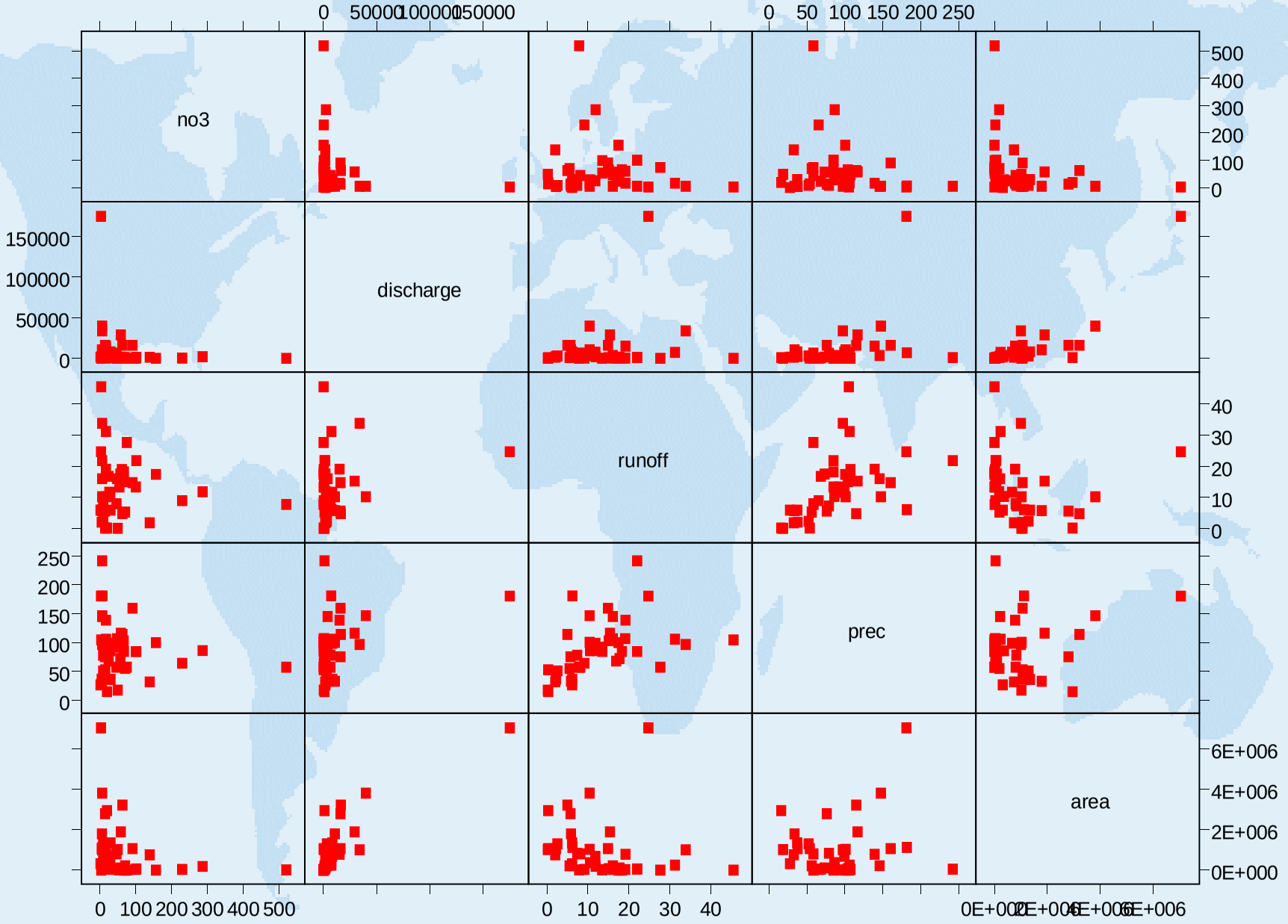
- Response: nitrate concentration NO3
- Explanatory variables:
  - discharge of river into ocean
  - runoff (from the watershed)
  - precipitation
  - area of watershed
  - density (people/square km)
  - nitrate deposition
  - nitrate precipitation nprec

Questions

- Is NO3 associated with deposition or nitrate precipitation, after accounting for discharge, runoff, precipitation, and area?

- Is there a further effect of population density on NO3, after accounting for the effects of nitrogen deposition and nitrogen precipitation?

Scatterplot



Log Transformations