Statistical foundations

Michael K. Tippett

International Research Institute for Climate and Society
The Earth Institute, Columbia University

ERFS Climate Predictability Tool Training Workshop
May 4-9, 2009

Ideas

- Linear regression.
- Goodness of fit.
- Overfitting.
- Selection bias.
- Cross validation.
- PCA, PCR.
- Multivariate regression and CCA.

Linear regression models

$$\hat{y} = Ax + b$$

- y = the predictand, the quantity to be predicted. E.g., rainfall, temperature.
- $\hat{y}$ = the prediction from the regression model.
- x = the predictor.
- A = regression coefficient(s).
- b = constant = $\langle y \rangle - A\langle x \rangle$

Linear relation between predictor and predictand.

Generalized linear models can be used for nonlinear relations

Linear regression models

Select A so that the sum of squared errors $(y - \hat{y})$ is minimized.

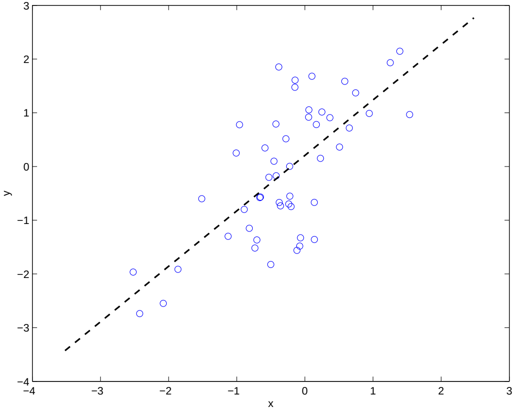
$$\min \left\langle \|y - Ax\|^2 \right\rangle = \min \sum_{i=1}^{n} \|y(t_i) - Ax(t_i)\|^2$$

n samples, for instance, at different times.

Univariate linear regression model

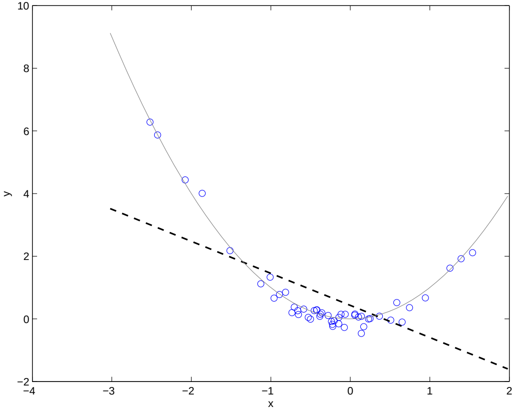
When x and y are scalars,

Regression finds the line $Ax + b$ that minimizes the sum of squared vertical differences between the line and the data.



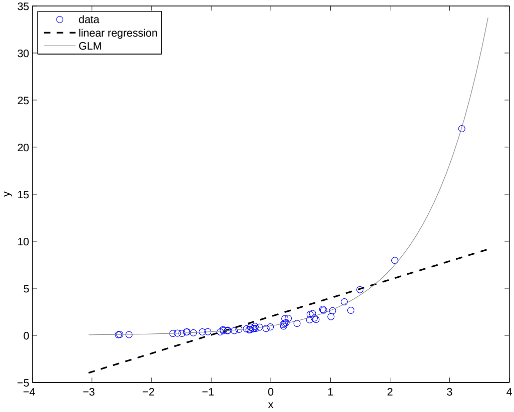
Nonlinear relations.

Line fits the data poorly.



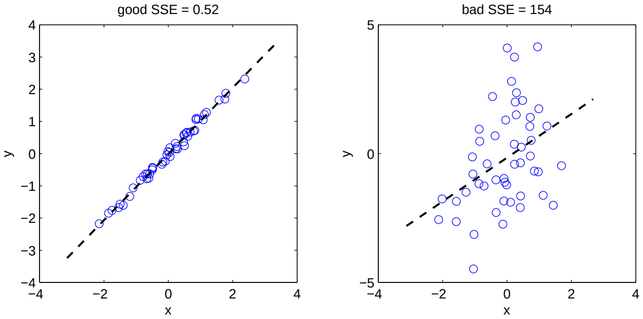
Nonlinear relations.

Line fits the data poorly. GLM (Poisson regression) does well.



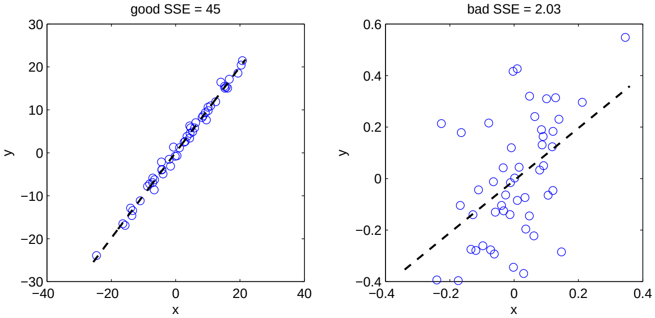
Goodness of fit

SSE = sum of square errors = $\sum(y - ax - b)^2$



Goodness of fit

SSE depends on the magnitude of the data.



Goodness of fit

Normalize SSE by a quantity proportional to the variance of the predictand.

SSA = sum of squared anomalies = $\sum(y-b)^2$

$$\frac{SSE}{SSA} = \text{fraction of variance unexplained} = \text{"noise"}$$

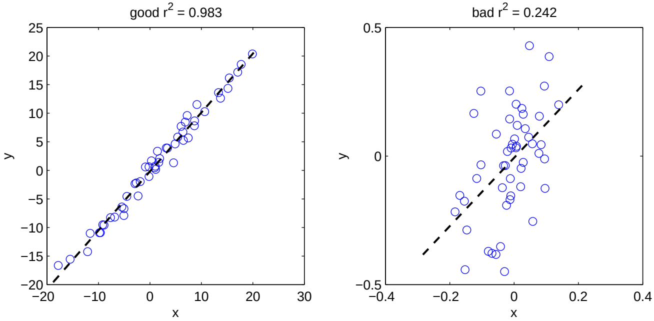
Related to the linear correlation r by

$$\frac{SSE}{SSA} = 1 - r^2$$

$$r^2 = \text{fraction of explained variance} = \text{"signal"}$$

Goodness of fit

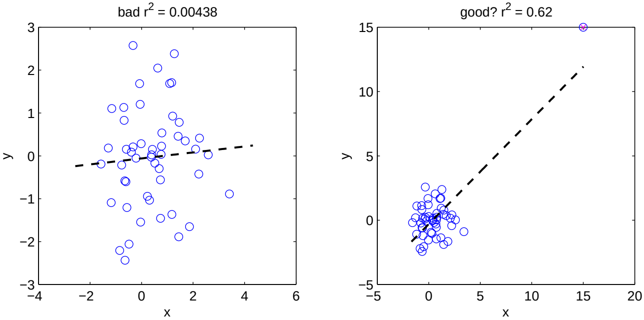
Correlation does not depend on the magnitude of the data.



Goodness of fit

Regression and correlation sensitive to outliers.

Assumption of Gaussian distributions.



Multiple linear regression

$$\hat{y} = a_1 x_1 + a_2 x_2 + \ldots a_m x_m + b$$

- y = the predictand, the quantity to be predicted. E.g., rainfall, temperature.
- $\hat{y}$ = the prediction from the regression model.
- $x_1, x_2, \ldots, x_m$ = the m predictors.
- $a_1, a_2, \ldots, a_m$ = regression coefficients.
- b = constant = $\langle y \rangle - a_1 \langle x_1 \rangle - a_2 \langle x_2 \rangle - \ldots a_m \langle x_m \rangle$

Data and number of predictors

Simple counting arguments show that at least $n = m + 1$ data points are required to estimate the $m + 1$ parameters of the regression model.

But with $n = m + 1$ data points and *any* m predictors the data can fit *perfectly*.

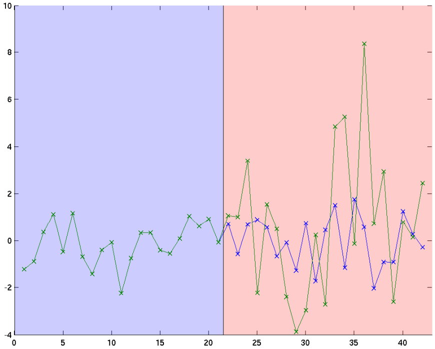
Such a regression, while fitting the data perfectly, would likely have little skill on independent data.

Overfitting. Using a model with so many parameters that random features of the training data are reproduced as the expense of predictive skill in independent data.

Example: Splitting the data to check for overfitting

- m predictors are used to form a regression using $n/2$ data points.
- Apply the regression equations to the $n/2$ data points not used in developing the regression.
- Check goodness of fit in the independent data.

training verification

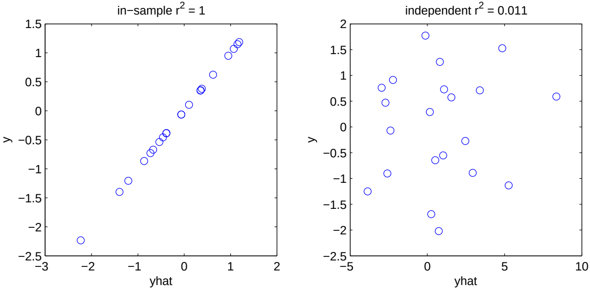


Example

$m = 20$ predictors

$n = 42$ samples

Predictors are unrelated to the predictand. (Random numbers).

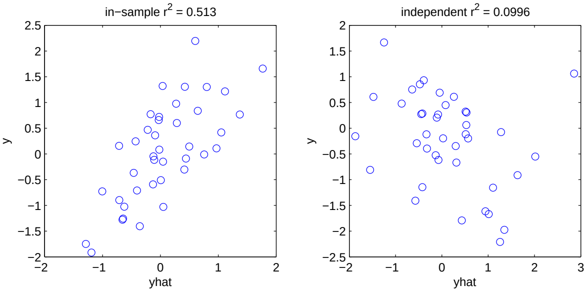


Example

$m = 20$ predictors

$n = 80$ samples

Predictors are unrelated to the predictand.

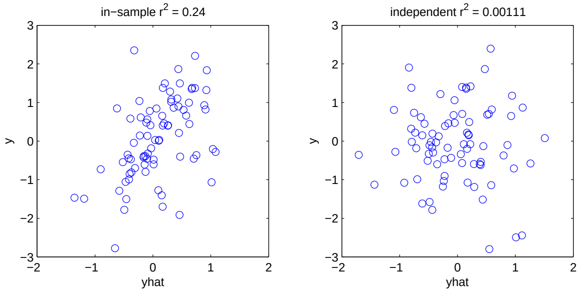


Example

$m = 20$ predictors

$n = 160$ samples

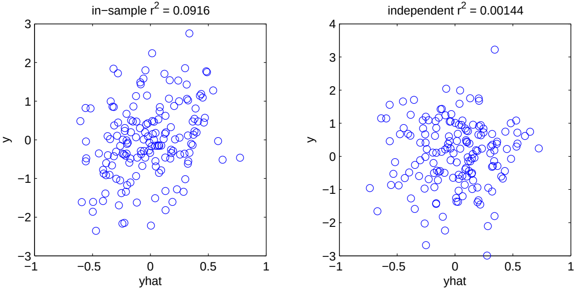
Predictors are unrelated to the predictand.



Example

$m = 20$ predictors
$n = 320$ samples

Predictors are unrelated to the predictand.



Rule of thumb

The more data, the harder to overfit.

5–10 samples per predictor.

$n \approx (5 \times m) - (10 \times m)$

Key points

The skill of a regression based forecast should be computed on independent data.

- Independent of the data used to compute the regression coefficients. Over-fitting.
- Independent of the data used to select the predictors. Selection bias.

Selection Bias: Example

A flawed procedure

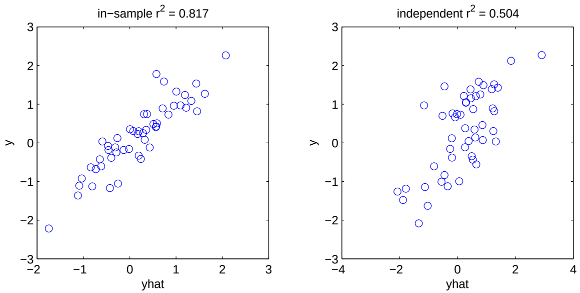
- Choose predictor that are well-correlated with the predictand using the entire data set.
- Compute the regression coefficients using the first half of the data.
- Apply the regression to the second half of the data to compute the skill.

Sounds good, but if the predictors are *selected* from a large pool of predictors using *all* the data, skill appears better than it really is.

Selection Bias: Example

$m = 10$ predictors
$n = 50$ samples

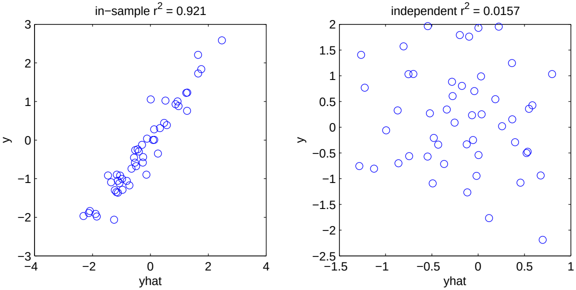
Predictors are unrelated ($r^2 = 0$) to the predictand as before, but were *selected* based on their correlation with the predictand in the *entire* data set.



Selection Bias: Example

$m = 10$ predictors
$n = 50$ samples

Predictors are unrelated ($r^2 = 0$) to the predictand as before, but were *selected* based on their correlation with the predictand in the first *half* of the data set.

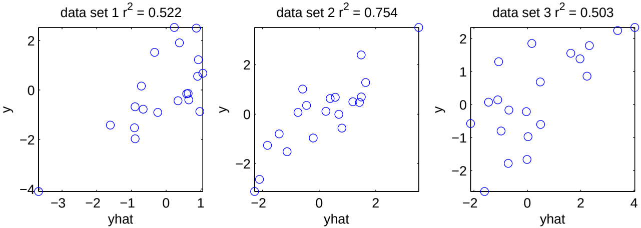


Avoiding selection bias (1)

3 independent data sets:

- Estimate regression models on data set 1.
- Select a “best” model on the basis of skill on data set 2.
- Evaluate skill of “best” model on data set 3.

Example: Select 2 predictors from a pool of 100.
One predictor has $r^2 = 0.5$, the rest are unrelated.
Each data set has 25 samples.
“Best” model $r^2 = 0.45$



Avoiding selection bias (2)

2 independent data sets:

- Estimate regression models on data set 1.
- Select a “best” model on the basis of an in-sample criteria like AIC or BIC.
  - Rewards model fit but penalizes model complexity.
- Evaluate skill of best model on data set 2.

Cross-validation

A method for mimicking real forecasts.

An alternative to splitting the data.

- Remove some number k of samples from the data set.
- Compute the model on the remaining $n - k$ samples.
- Use that model to predict the left-out samples.
    - Sometimes a set of k contiguous in time samples are left out and only the middle one is predicted to deal with temporal correlation.
- Repeat.

Often $k = 1$.

Data reduction and PCA

For climate forecasts, the length of the historical record severely limits the number of predictors

What if the predictors are spatial fields such as the SST or the output of a CGCM?

The number of grid point values (100’s, 1000’s)is large compared to the number time samples (10’s for climate)

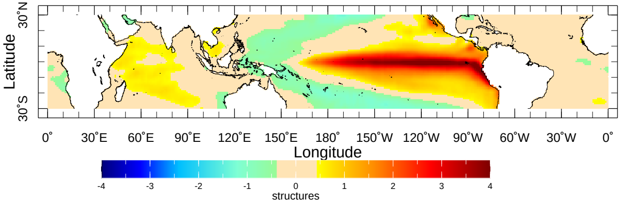
Need to represent the information in the spatial field using fewer numbers.

- Pick a few “representative” grid points
    - Hard to do, Noisy.
- Spatial averages e.g., NINO 3.4. All India Rainfall.
- Principal component analysis (PCA). Also called EOF.
    - Weighted spatial average.
    - Weights are chosen in an optimal manner to maximize explained variance.

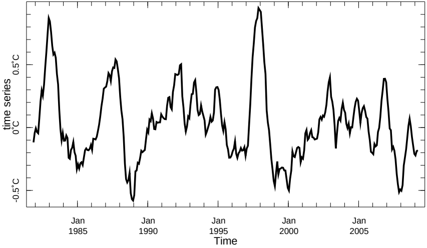
Example: PCA of tropical SST

Weight (EOF) 1

1.



Time series (PC) 1

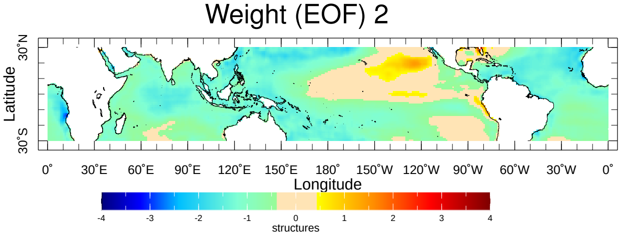


1.

Explains 26% of the total variance of the SST field.

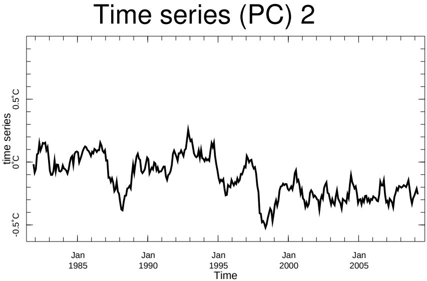
Computed in the Data Library web interface.

Example: PCA of tropical SST



Weight (EOF) 2

2.



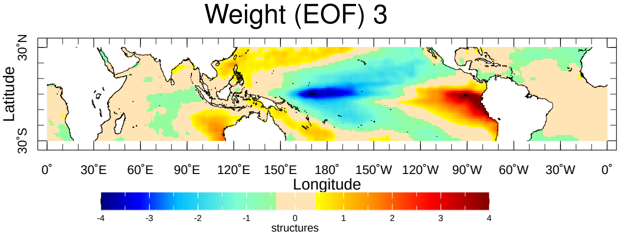
Time series (PC) 2

2.

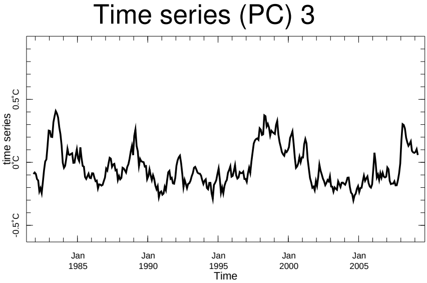
Explains 12.5% of the total variance of the SST field.

Computed in the Data Library web interface.

Example: PCA of tropical SST



3.



3.

Explains 7.4% of the total variance of the SST field.

Computed in the Data Library web interface.

PCR

Principal component regression (PCR).

- $\hat{y} = a_1 x_1 + a_2 x_2 + \ldots a_m x_m + b$
- Predictors x_i are PCs.

Multivariate linear regression

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_l \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & . & . & . & a_{1m} \\ a_{21} & a_{22} & . & . & . & a_{2m} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ a_{l1} & a_{l2} & . & . & . & a_{lm} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_m \end{bmatrix}$$

- l predictands
- m predictors
- $l \times m$ regression coefficients.

  Useful fact. Each row of regression coefficients can be computed separately. Generally not true for the columns.

Multivariate linear regression

How to interpret the regression coefficients?

$$A = \begin{bmatrix} a_{11} & a_{12} & . & . & . & a_{1m} \\ a_{21} & a_{22} & . & . & . & a_{2m} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ a_{l1} & a_{l2} & . & . & . & a_{lm} \end{bmatrix}$$

Which coefficients are important? Which are marginal?

Hard. Unless A is diagonal.

Multivariate linear regression

To interpret the regression coefficients, transform x and y so that the regression matrix is diagonal.

$$y' = A'x'$$

$$A' = \begin{bmatrix} a_{11'} & 0 & 0 & 0 & 0 & 0 \\ 0 & a'_{22} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & . & 0 & 0 \\ 0 & 0 & 0 & 0 & . & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & a'_{lm} \end{bmatrix}$$

Univariate regressions. Easier to understand.

$$y'_i = a'_{ii} x'_i$$

Multivariate linear regression & CCA

Many ways to diagonalize A.

Canonical correlation analysis (CCA).

- the regression coefficients are correlation coefficients,
- the new variables are uncorrelated,

New variables

- maximize correlation
- are linear combinations of old variables,
- correspond to spatial patterns.

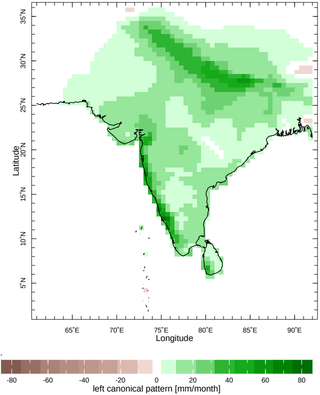
PCR is a special case of CCA.

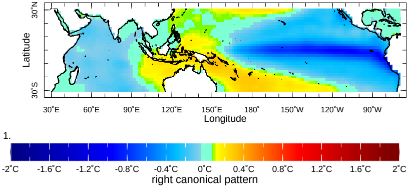
CCA: Example

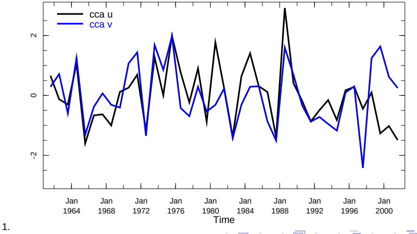
- JJAS rainfall
- JJAS Pacific SST
- Data 1961-2001
- Use 4 rainfall PCs and 3 SST PCs.

CCA: Example

Pair 1







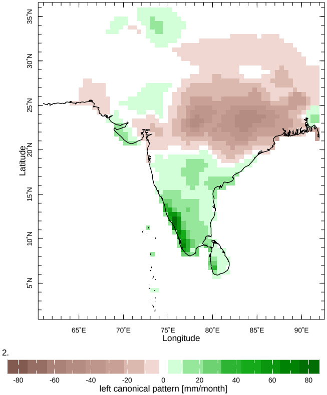
Correlation = 0.56

14%, 36%

Computed in the Data Library web interface.

CCA: Example

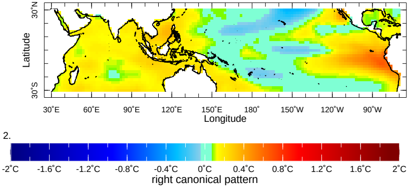
Pair 2

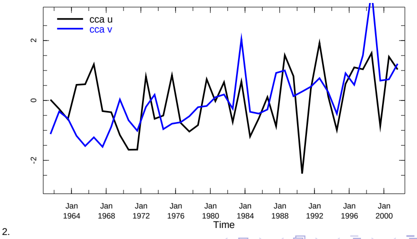


Correlation = 0.40

10%, 18%

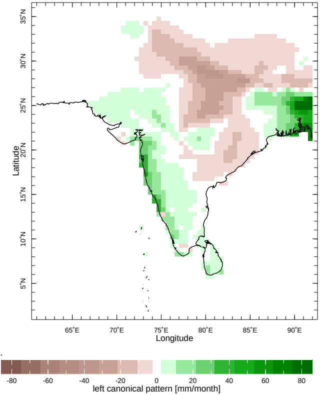
Computed in the Data Library web interface.

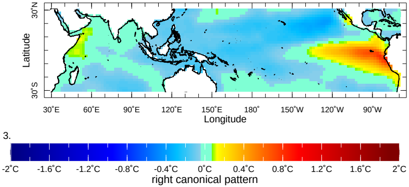




CCA: Example

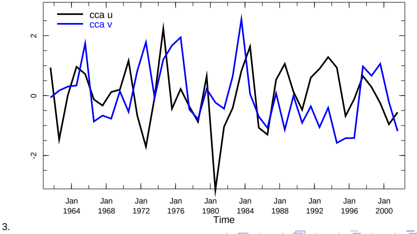
Pair 3





Correlation = 0.11

10%, 12%



Computed in the Data Library web interface.

Summary

- Use regression to model linear relations.
- Minimize squared error.
- Correlation measures variance explained. Goodness of fit.
- Train regression and validate skill in separate data sets.
- Need many more samples than predictors to avoid overfitting.
- Selection bias.
- Cross validation.
- PCA, PCR.
- Multivariate regression and CCA.