Artificial Skill and Overfitting

Timothy DelSole

George Mason University, Fairfax, Va and
Center for Ocean-Land-Atmosphere Studies, Calverton, MD

July 21, 2010

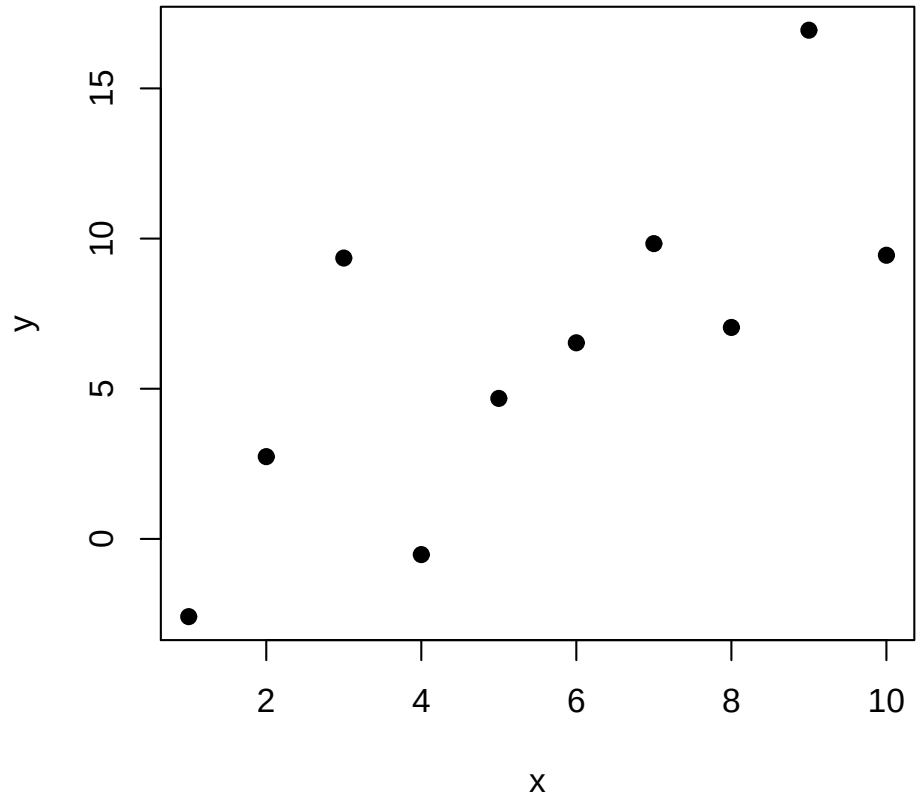
Question From Lecture 1

If 2 predictors are better than one,
then are 10 predictors better than two?

In general, should we use as many predictors as possible?

Fitting Data

What is a "good" model of x and y?



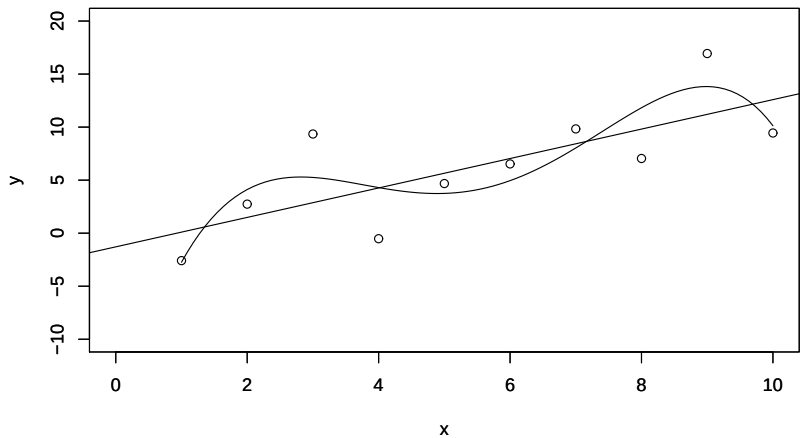
Consider two models of the above data

Model A $y = \beta_0 + \beta_1 x + \epsilon$

Model B $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \beta_5 x^5 + \epsilon$

Which Model is Best?

Is it the model that fits the data with the least error?

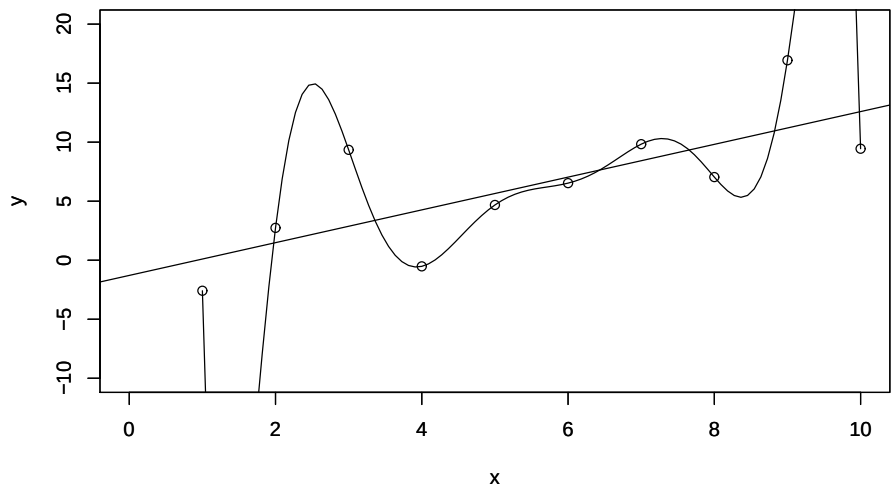


Model	Standard Error
Model A	3.76
Model B	3.01

By this criterion, the "best" model is model B.

Take the Logical Extreme

If the criterion is to select the best fit model, then we should select a 9'th order polynomial, because it fits 10 data points *exactly*.



What is Wrong with Choosing the Best Fit Model?

Logical Extreme of Choosing the Best Fit Model

Model B contains model A (simply set all but β_0 and β_1 to zero).

Model B has more flexibility than model A to capture variability.

Higher order polynomials have increasing flexibility.

If the criterion is to select the model that fits the data the best, then the highest order polynomial will always be selected.

Adding Random Predictor Also Gives Better Fit

Model C $\quad y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \beta_5 x^5 + \beta_6 z + \epsilon$

where z is independent random noise.

Model	Standard Error
Model A	3.76
Model B	3.01
9'th order polynomial	0
Model C	2.98

Can you explain why adding a random predictor improves the fit?

Why Adding a Random Predictor Improves the Fit

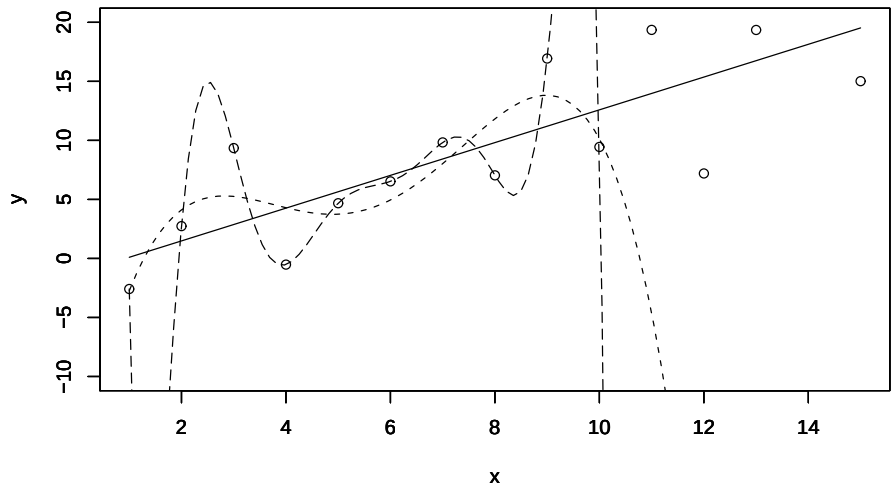
Even though z is independent random noise, the *sample covariance* between y and z does not vanish.

Loosely speaking, the method of least squares interprets the nonzero covariances as real relations, and uses these apparent relations to fit the data as well as possible.

Logical Extreme: you can fit *any* data simply by adding enough random predictors to the model.

Another Problem With Selecting the Best Fit Model: It does not generalize to independent data

Consider "new" data from the **population** $y = ax + \epsilon$.



The curves give the best fit polynomials of order 1, 5, 9.

Only the line fits the new, independent data well. The other polynomials are grossly incorrect.

Distinguish Fitting from Forecasting

Fitting is the process of describing a specific sample (including its random variations) with a model.

Forecasting is the process of predicting independent data (i.e., data that is independent of the sample used to construct the model).

Overfitting

Fitting a model with "too many" parameters is called **overfitting**.

If the true model has noise, e.g., $y = ax + \epsilon$, then fitting the data exactly to a high order polynomial is fitting the model **to noise**.

Unfortunately, we rarely know how many parameters is "too many."

Artificial Skill

The forecast error variance is always greater, on average, than the fit error variance. The difference between the mean error variances of the forecast and fit is called the **artificial skill**.

If the in-sample SSE is used to estimate the prediction error, then the model will appear to be more skillful than it really is on independent data. The enhancement in skill is due solely to overfitting the model and hence is artificial.

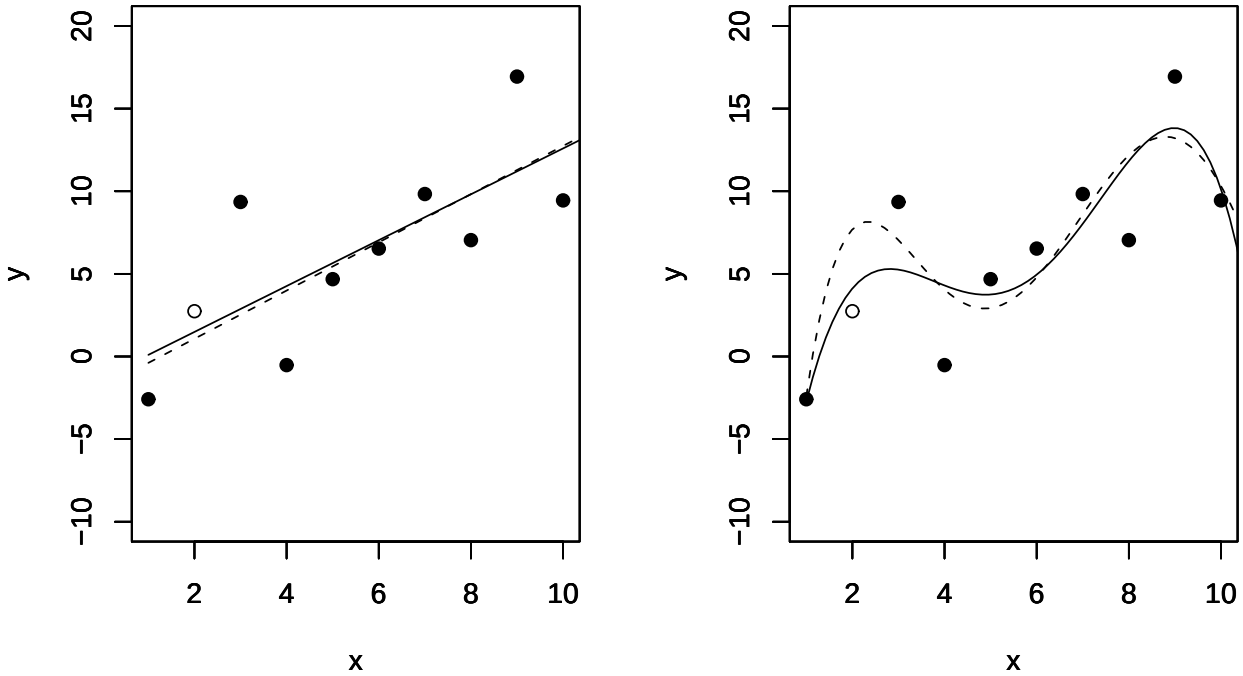
If the goal is to **predict** independent data, then the model should **generalize** to independent data.

How Do You Select a Model that "Generalizes" to other Independent Data?

Leave-One-Out Regression

- Withhold one sample.
- Fit the model to the remaining sample (of size $N - 1$).
- Use the resulting model to predict the withheld sample.
- Measure the error of the resulting prediction.

Result of Leave-One-Out Regression



SOLID: Fit using all data.
DASH: Fit using all but one datum (open circle)

The difference between the withheld sample and the model fit is smaller for the line fit than for the higher order polynomial.

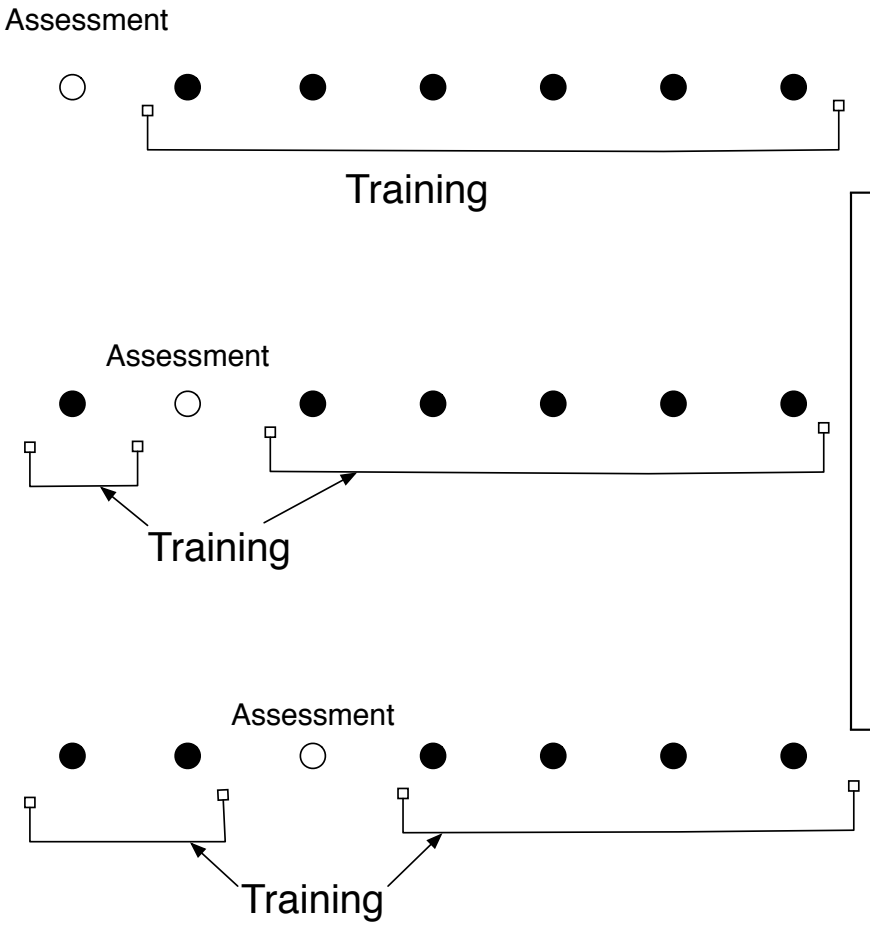
Note: Largest difference occurs near the withheld datum.

Do You See a Problem with Leave-One-Out Regression?

- The withheld sample might be **unrepresentative**.

Can you think of a way to avoid this problem?

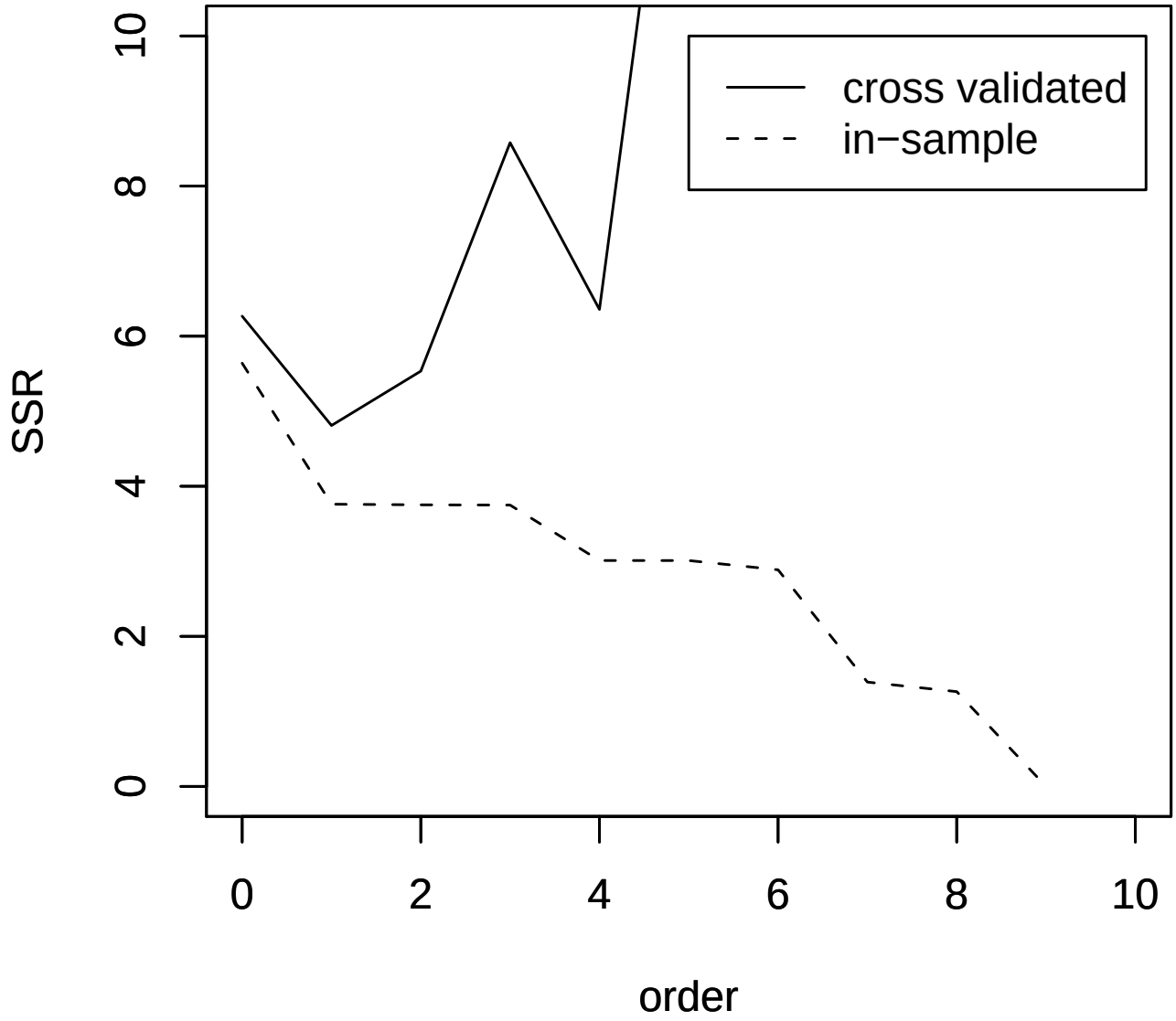
Leave-One-Out Cross Validation



Training Sample: Sample used to estimate model parameters.

Assessment Sample: Sample used to assess/test model predictions.

Cross Validation of Polynomial Fits



Cross Validation of Polynomial Fits

- Cross validated SSE is minimum for first order polynomial.
  - Implies that line fit generalizes to independent data the best.
  - Consistent with the population model that generated the data.
- In-sample SSE decreases monotonically with order.
- Cross validated SSE has relatively large fluctuations.
  - Fluctuations in SSE_{cv} can produce **spurious** minima.

AICC: Another Predictor Selection Method

Use a selection criterion that penalizes against "complexity."

$$AICC_K = \log \hat{\sigma}_K^2 + \frac{N+K}{N-K-2}$$

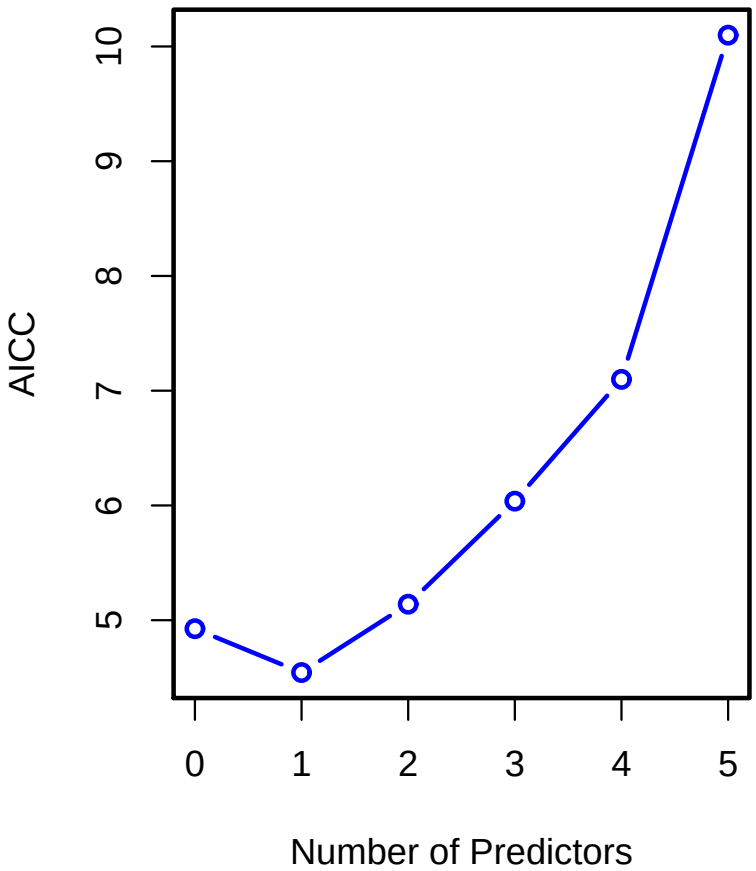
$$SIC_K = \log \hat{\sigma}_K^2 + \frac{K}{N} logN$$

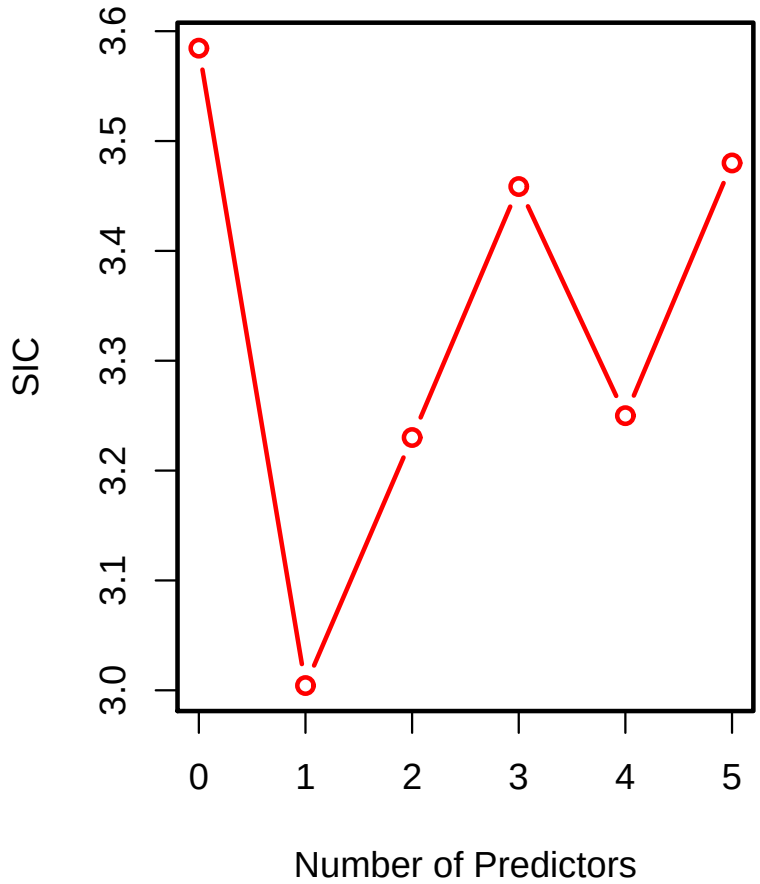
Goodness-of-Fit Penalty Function

where $\hat{\sigma}_K^2 = SSE/N$.

The more predictors K, the stronger the penalty.

Application of AICC with Polynomial Fits





Minimum AICC and SIC occur for one predictor (i.e., a first order polynomial), consistent with the true population model.

Let R do the Work

R has several packages for doing cross validation, AIC, etc. I have broken things step by step for pedagogical reasons.

See http://www.statmethods.net/stats/regression.html for a nice summary of things you can do.

Seasonal Variability Indices

Downloaded from http://www.cpc.ncep.noaa.gov/data/indices/

* North Atlantic Oscillation (NAO)
* East Atlantic (EA.NP)
* East Atlantic/Western Russia (EA.WR)
* Scandinavia (SCA)
* West Pacific (WP)
* East Pacific-North Pacific (EP-NP)
* Pacific/North American (PNA)
* Pacific Decadal Oscillation (PDO)
* El Nino/Southern Oscillation (NINO3.4)
* Quasi-Biennial Oscillation (QBO30)
* Index of zonal averaved 500mb temperature (z500t)
* 850mb Trade Wind Index (EPAC850)
* Average South Atlantic SST (SATL)
* Average North Atlantic SST (NATL)
* year (year)

Goal: We want to predict Indian Monsoon Rainfall

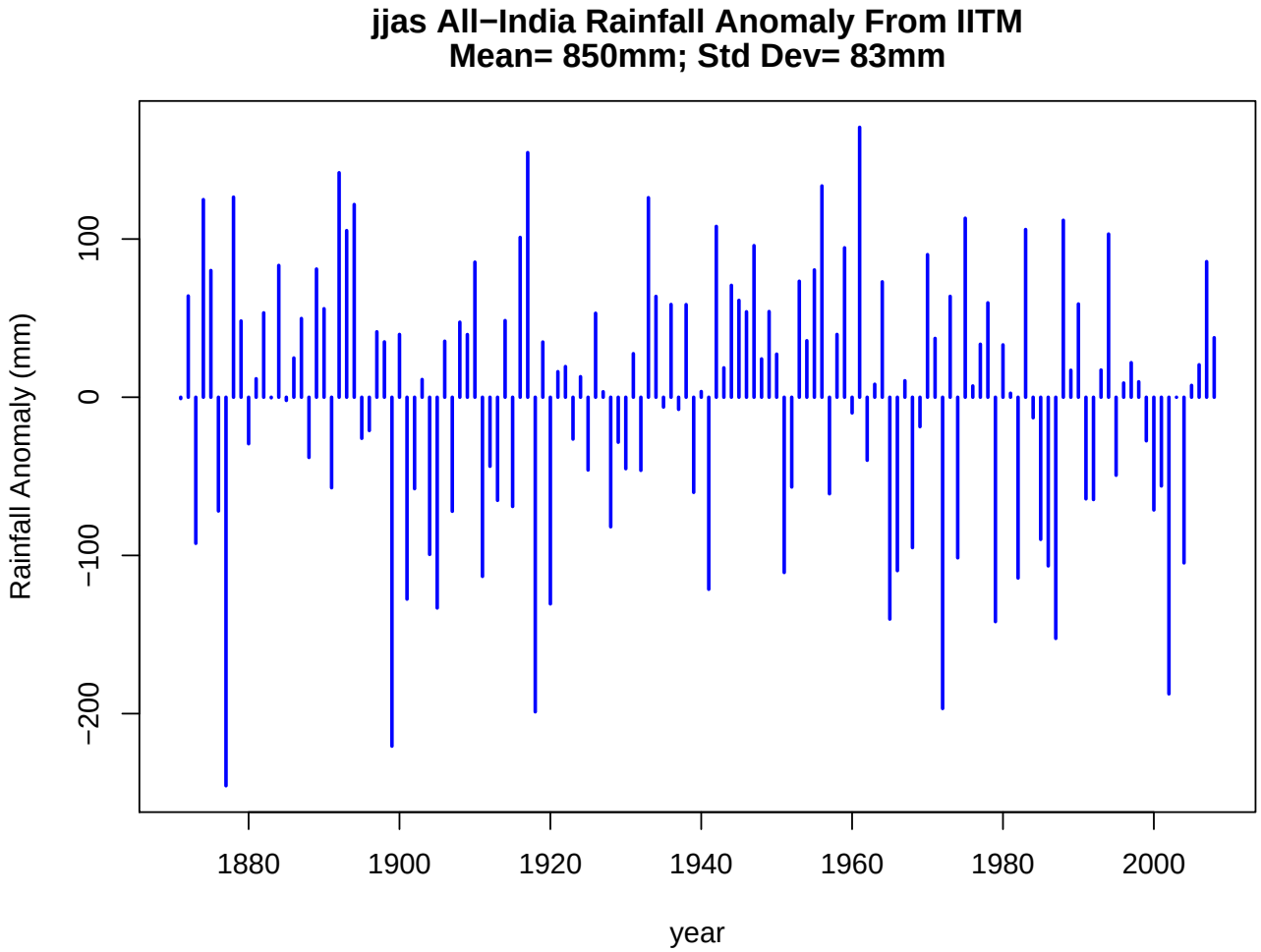


Figure: JJAS Indian Monsoon Rainfall

How to Make Bar Graph of ISMR

```
# make bar graph with labels
plot(year,y-mean(y),type="h",col="blue",
       xlab="year",ylab="Rainfall Anomaly (mm)",lwd=2)

# calculate mean and standard deviation
ismr.mean = mean(y)
ismr.sd   = sd(y)

# make titles
ftitle.top = paste("JJAS All-India Rainfall Anomaly From IITM")
ftitle.bot = paste("Mean= ",round(ismr.mean),"mm;
          Std Dev= ",round(ismr.sd),"mm",sep="")
title(main=ftitle.top,line=2)
title(main=ftitle.bot,line=1)
```

Comparing Correlations in R

```
# x is a matrix of seasonal indices
# y is a vector of Monsoon Rainfall

xy.cor = cor(x,y) ; # correlation between y and columns of x
cor.crit = 2/sqrt(length(y)); # approx. 5% significance level

# identify symmetric limits on x-axis
xrange = max(abs(xy.cor))
xrange = c(-xrange,xrange)

# make bar plot
barplot(xy.cor[,1],horiz=TRUE,las=1,xlim=xrange,
            col="yellow",cex.names=1.5)

# draw significance levels
abline(v= cor.crit ,col="red",lty="dashed",lwd=3)
abline(v=-cor.crit ,col="red",lty="dashed",lwd=3)
```

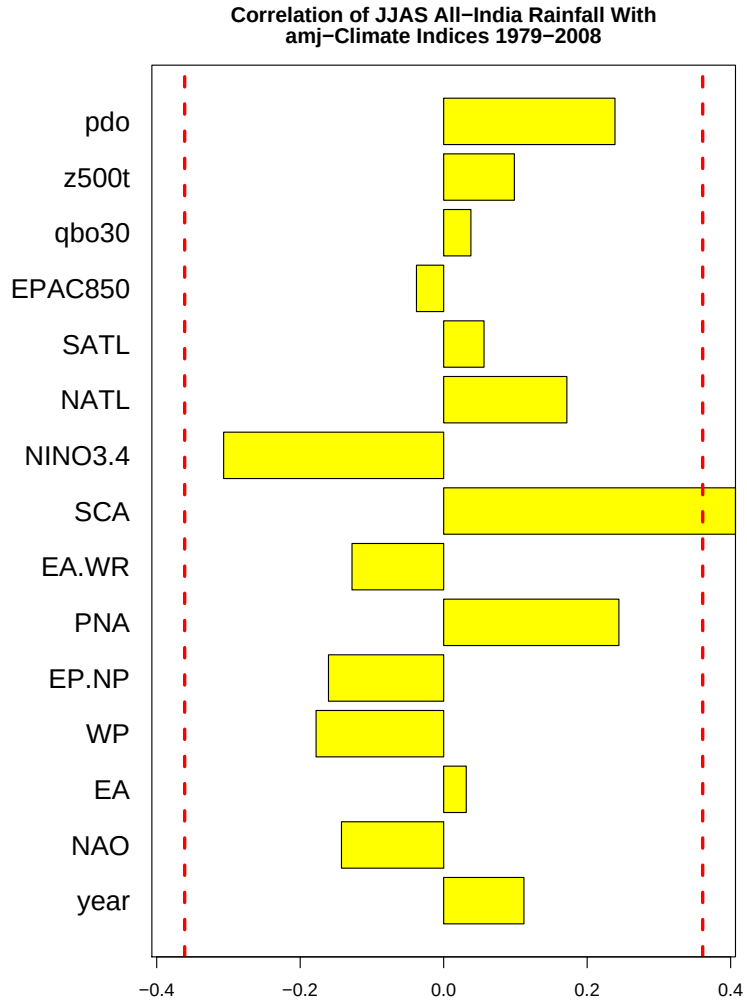


Figure: Correlation between AMJ-climate indices and all-India Monsoon Rainfall for 1979-2008 (30 yrs). Red dashed is the 5% significance level.

Linear Regression With R

```
# fit linear regression model y = ...
y.all = lm ( y ~ year + nao + ea + wp + ep.np
             + pna + ea.wr + sca + nino34 + natl
             + satl + epac850 + qbo + z500 + pdo )

# extract statistical properties
summary(y.all)
```

Linear Regression

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -16671.897   8192.566  -2.035   0.0612 .
year             8.804      4.121   2.137   0.0508 .
nao            -15.351     29.796  -0.515   0.6145
ea             -37.655     30.900  -1.219   0.2431
wp              56.643     44.861   1.263   0.2273
ep.np          -93.421     55.628  -1.679   0.1152
pna             18.757     34.156   0.549   0.5915
ea.wr           -4.303     39.243  -0.110   0.9142
sca             28.581     37.915   0.754   0.4634
nino34         -77.229     38.748  -1.993   0.0661 .
natl           -70.167     67.269  -1.043   0.3146
satl          -106.613     72.611  -1.468   0.1641
epac850         15.575     27.111   0.574   0.5748
qbo             10.908     19.580   0.557   0.5863
z500            23.150     19.478   1.189   0.2544
pdo             41.566     25.116   1.655   0.1202
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 77.64 on 14 degrees of freedom
Multiple R-squared: 0.5292, Adjusted R-squared: 0.0248
F-statistic: 1.049 on 15 and 14 DF,  p-value: 0.4667
```

Conclusions Based on Correlations and Model Fitting

- Individual correlations not significant (except Scandinavia)
- p-value for all predictors vanishing $= 0.47$ (not significant)
- No single coefficient in full model is significant.

So far, no evidence that any of the predictors are related to ISMR.

All Possible Subsets

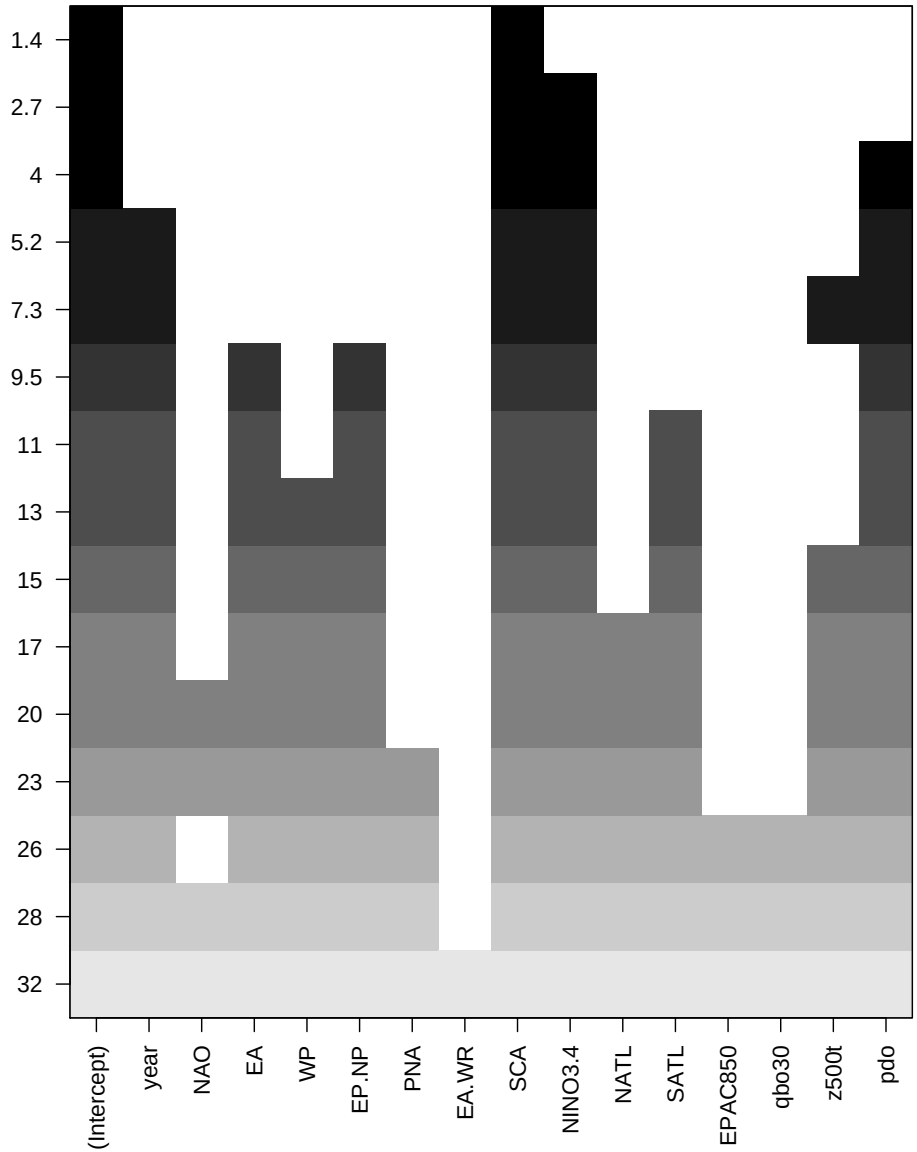
```
# do all possible subsets of predictors
xy.step = regsubsets(x=x,y=y,nvmax=dim(x)[2])

# plot indicator of selected predictors
plot(xy.step)
```

Interpretation of regsubsets

`regsubsets` performs an exhaustive search for the best subsets of the variables in x for predicting y in linear regression.

- Minimum model has only the intercept, by default.
- For more complicated minimum model, use `force.in` to specify terms in the minimum model (a logical vector with TRUE for variables in the minimum model and FALSE for variables not in the minimum model).
- regsubsets() will give you the best model with one variable, the best with two variables, and so on. The object produced by summary() of the regsubsets() has a component $bic that gives the BIC value for each of the best models.
- use plot() to see which variables are in the good models, and which variables tend to occur together or separately

Indicator Plot for All Possible Subset Selection



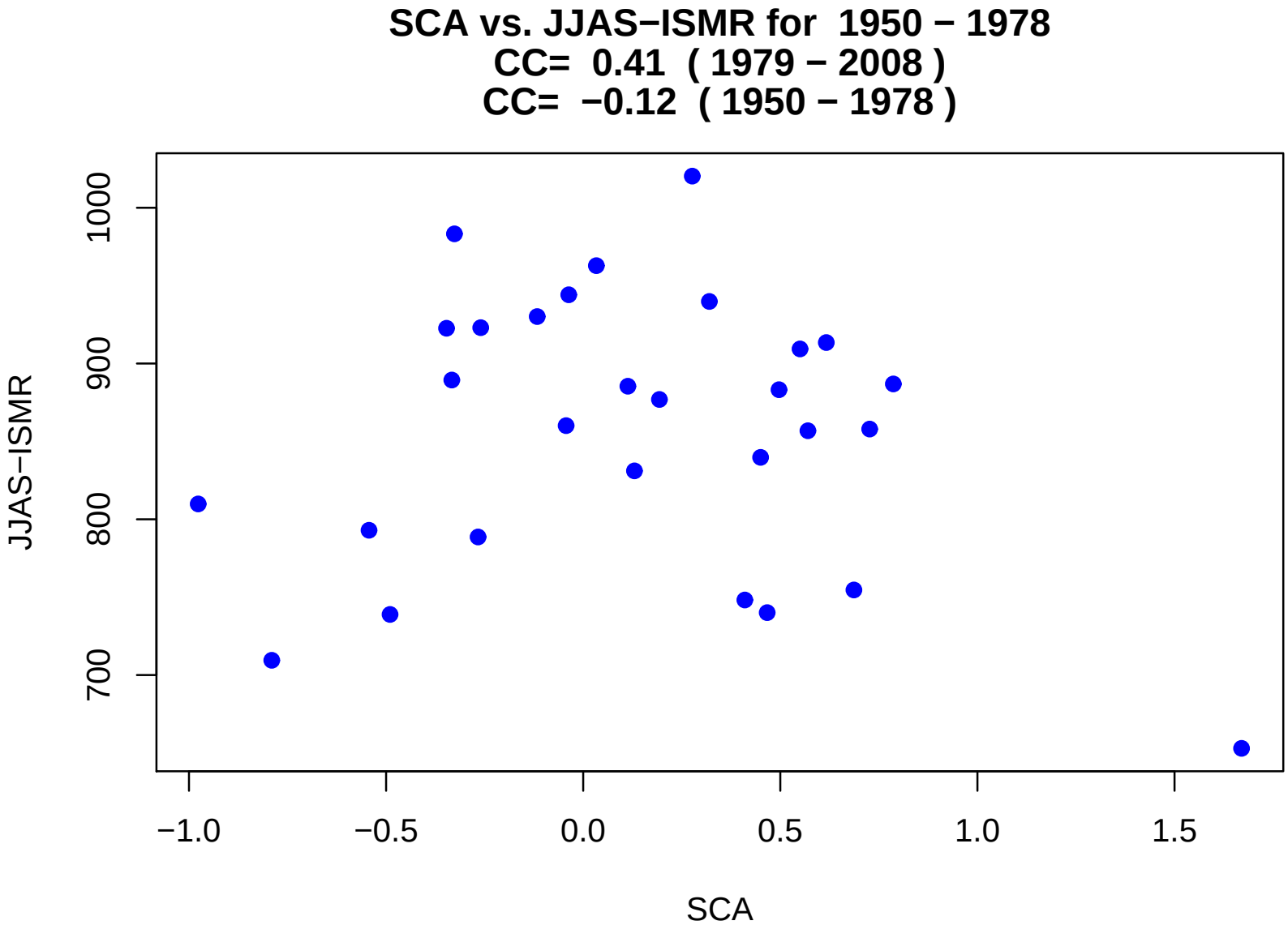
All Possible Subsets

```
# summarize the results
summary(xy.step)$bic
 [1]   1.377163  2.724493  4.020964  5.189060  7.273588
 [6]   9.476993 10.905596 12.860013 14.740451 16.766610
 [11] 19.573594 22.614447 25.610864 28.443235 31.818675
```

Rule: choose the model with the smallest BIC

This rule indicates we should chose SCA for predicting ISMR.

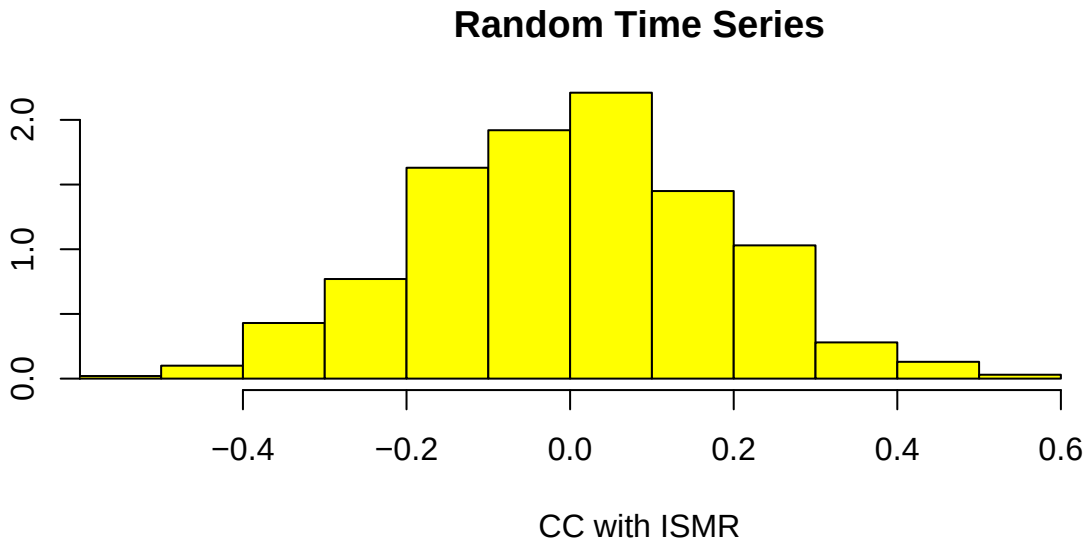
Independent Verification

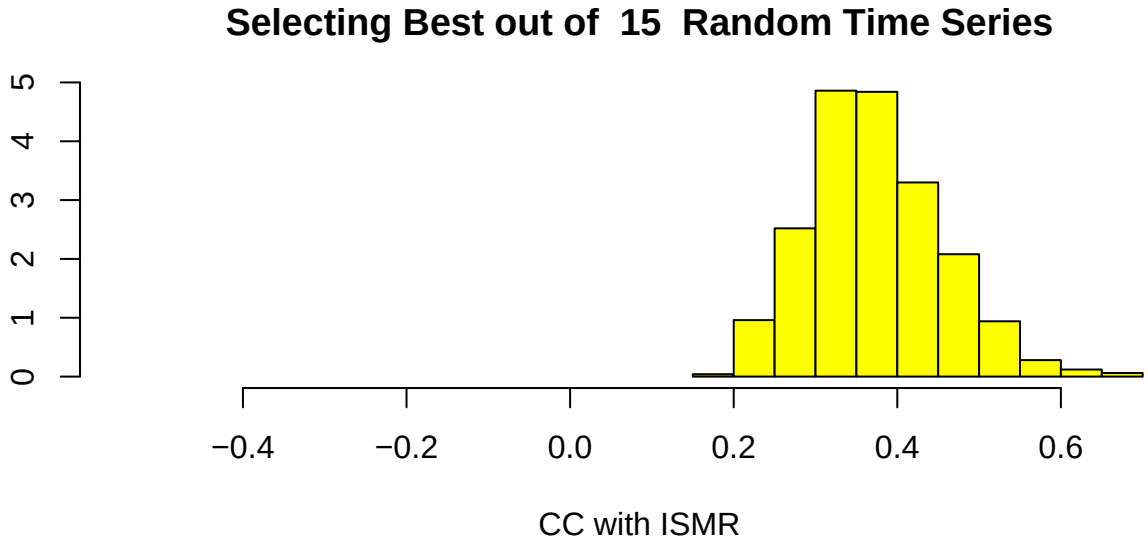


Why is the training correlation so much larger than assessment correlation?

Answer: the way we selected predictors creates a bias).

Correlations When Selection is Taken Into Account





Comments About Overfitting and Artificial Skill

There is no general solution to the problem of overfitting.

Statistical climate prediction is particularly problematic because thousands of candidate predictors need to be considered.

If cross validation is used to select predictors AND assess skill, then the skill will be artificially inflated because the predictors were chosen to optimize skill for the available sample.

When the process of selecting the "best" variables from a pool of hundreds is taken into account, results usually are not significant.

To obtain statistically significant results, initial pool of variables must be restricted based on criteria independent of the data, such as on physics or independent climate simulations.