Tues, 4-23-19
Validating the Regression Model

- Feedback on A11

1. **Skill score and the RE statistic**
2. **Validation strategies**
3. **Error bars for predictions**
4. **Interpolation vs extrapolation**

Read notes_12.pdf

A11 Feedback

1. Download A11x.pdf from D2L
2. Automatic points, for running assignment and having uploaded by due time, is already marked in parentheses at top of first page
3. Each assignment has maximum possible 10 points; if you make no deductions, score is 10/10
4. A11x is color coded for points; purple=1; yellow=0.5; blue=0.5
5. Open your copy of the same assignment pdf you uploaded
6. In Acrobat Reader, using “Add text box,” mark in right margin for deductions only, with deduction and segment reference : (eg., -0.5 A); round to tenths in deductions (e.g., no -0.25)
7. At top of your pdf, mark grade like this : 9.5/10
8. If necessary, put any comments at top near the grade
9. Upload your self-graded pdf to folder A11_graded in D2L

Calibration vs Validation

Calibration

1. Fitting the model to the data
2. "calibration", "construction", "estimation" data
3. Accuracy statistics:
   {R^2, R_a^2,}
   SSE_c
   MSE_c
   $RMSE_c$

Validation

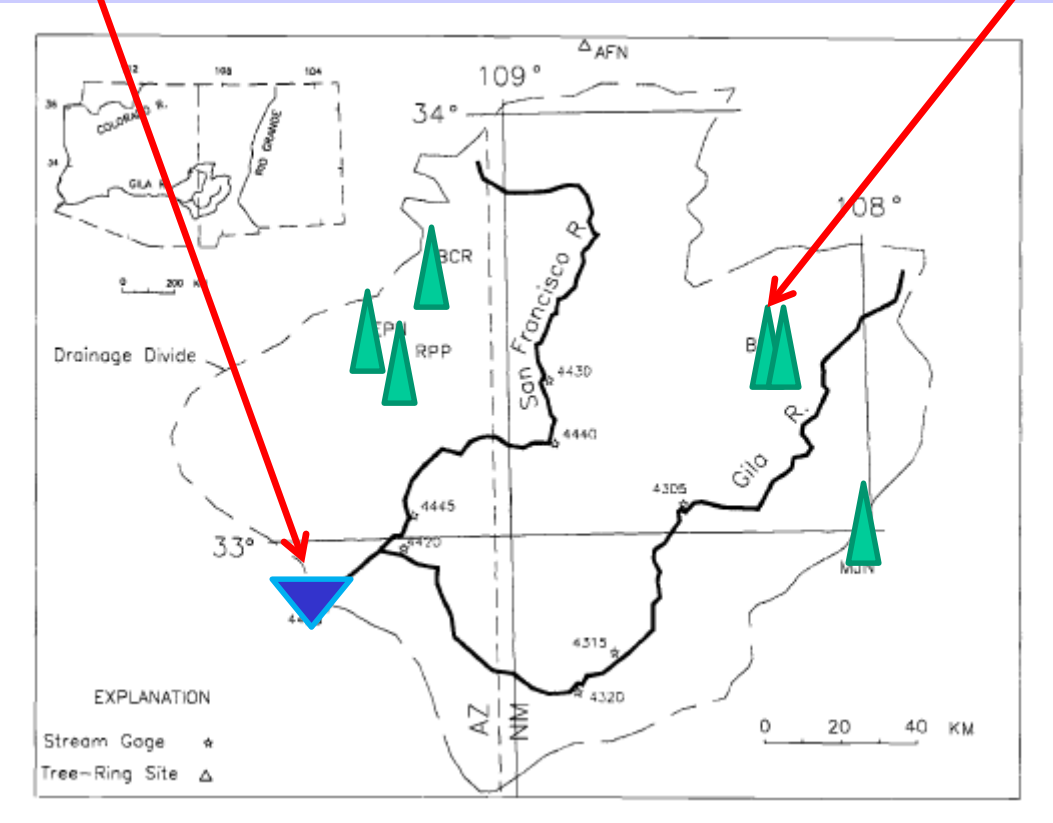
1. Testing the model on data not used to fit the model
2. "validation", "verification", "independent" data
3. Accuracy statistics:
   RE
   SSE_v
   MSE_v
   $RMSE_v$

Validation: example

Gila River, AZ, reconstructed annual flow from tree rings

Stream gauge

Tree-ring site

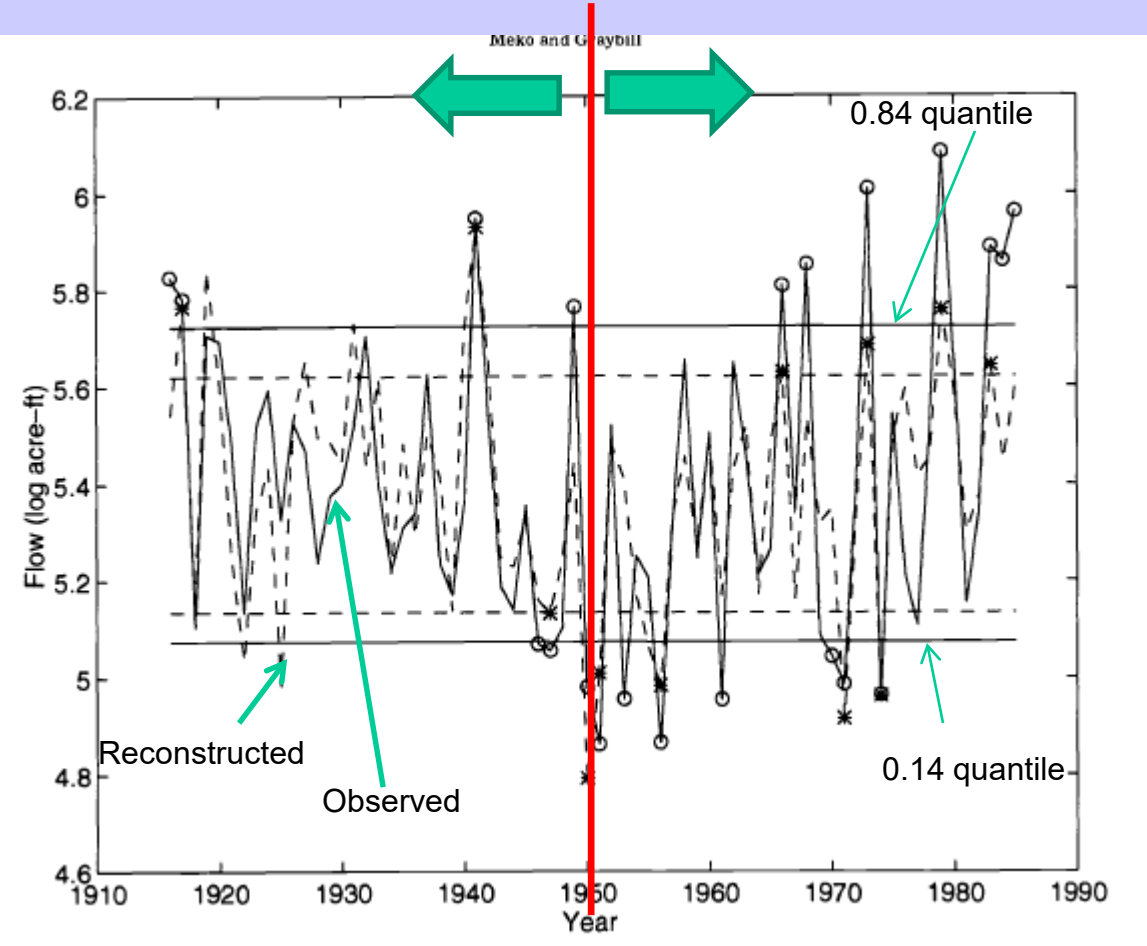


Split-Sample Validation

- Fit model on 1916-50 and validate on 1951-85
- Fit model on 1951-85 and validate on 1916-50

Meko and Graybill, 1995. Water Resources Bulletin 31(4):

Example (cont.)



- Imperfect skill (observed different than reconstructed)
- Compressed variance (quantile thresholds for dry and wet years narrower for reconstructed than for observed)
- Can derive quantitative “skill score” to summarize validation skill

General Definition of a Skill Score*

$$\text{Skill Score} = \frac{A - A_{ref}}{A_{perf} - A_{ref}}$$

Measures skill relative to some reference prediction (or reconstruction)

A = accuracy of prediction

A_{perf} = accuracy for "perfect" prediction

A_{ref} = accuracy of some "reference" prediction

*Wilks, D.S., 1995, Statistical methods in the atmospheric sciences: Academic Press, 467 p.

Skill Score for "R^2-type" accuracy measure

$$A = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}$$

"Accuracy" by regression R^2 (see class notes)

$$A_{\mathrm{perf}} = 1$$

$$A_{\mathrm{ref}} = 1 - \frac{\mathrm{SSE}_{\mathrm{ref}}}{\mathrm{SST}}$$

$$\mathrm{Skill} = \frac{A - A_{\mathrm{ref}}}{A_{\mathrm{perf}} - A_{\mathrm{ref}}} = \frac{\left(1 - \frac{\mathrm{SSE}}{\mathrm{SST}}\right) - \left(1 - \frac{\mathrm{SSE}_{\mathrm{ref}}}{\mathrm{SST}}\right)}{1 - \left(1 - \frac{\mathrm{SSE}_{\mathrm{ref}}}{\mathrm{SST}}\right)}$$

$$= \left(\frac{\mathrm{SSE}_{\mathrm{ref}} - \mathrm{SSE}}{\mathrm{SST}}\right) \Big/ \left(\frac{\mathrm{SSE}_{\mathrm{ref}}}{\mathrm{SST}}\right) = 1 - \frac{\mathrm{SSE}}{\mathrm{SSE}_{\mathrm{ref}}}$$

Reduction of Error (RE) Statistic

$$\text{Skill} = 1 - \frac{\text{SSE}}{\text{SSE}_{\text{ref}}}$$

y_t = observed predictand

$\hat{y}_t$ = reconstructed predictand

$\overline{y}_c$ = calibration-period mean of y_t

Define the reference reconstruction to be $\overline{y}_c$

$$\text{SSE} = \sum_v \left(y_t - \hat{y}_t \right)^2 \quad = \quad \text{error sum-of-square for validation period}$$

$$\text{SSE}_{\text{ref}} = \sum_v \left(y_t - \overline{y}_c \right)^2 \quad = \quad \text{reference error sum-of-square for validation period}$$

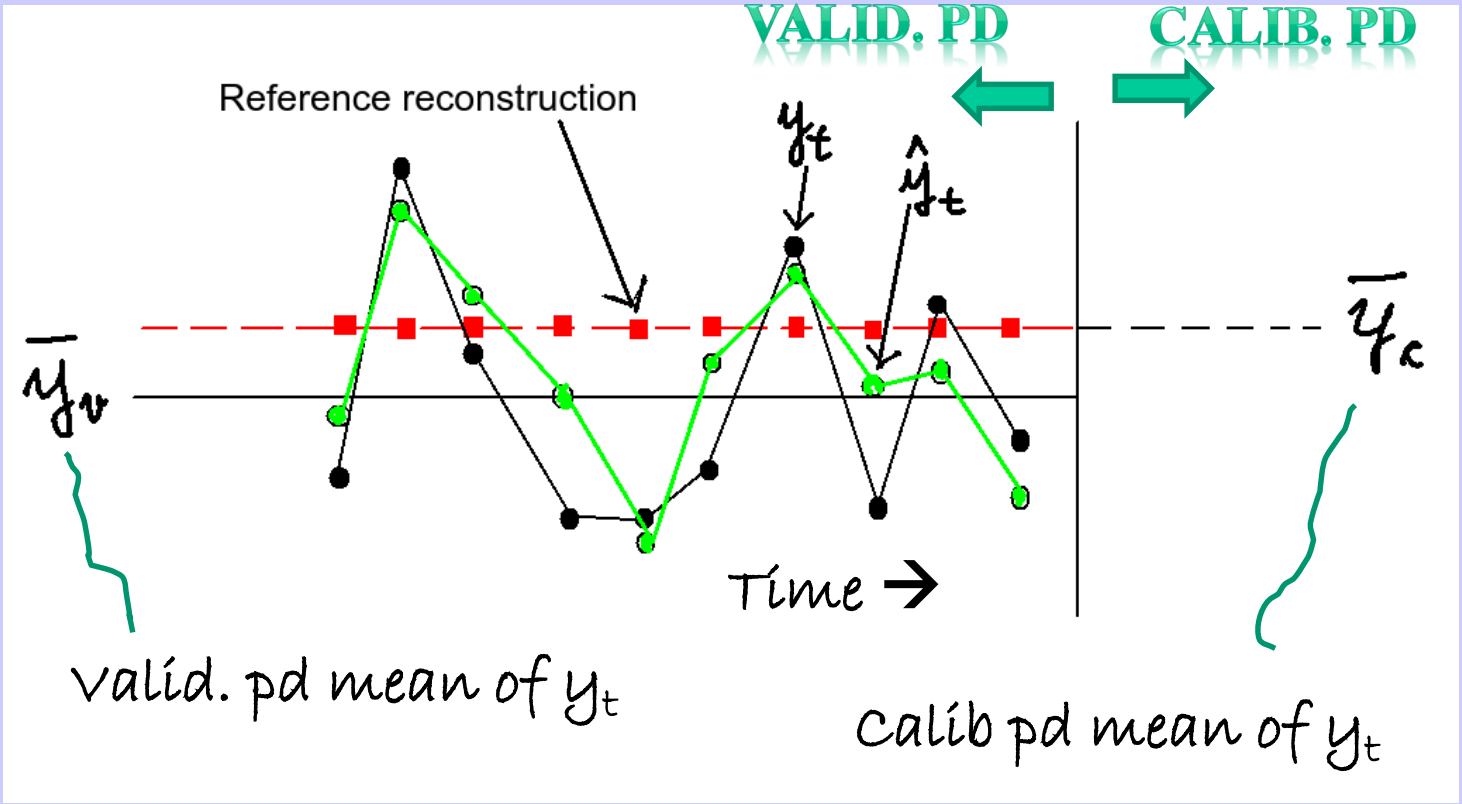
$$\text{Skill} = 1 - \frac{\text{SSE}}{\text{SSE}_{\text{ref}}} \equiv \text{RE}$$

$-\infty < \text{RE} \leq 1$

"Some" skill: RE>0

RE Statistic: the "reference" reconstruction

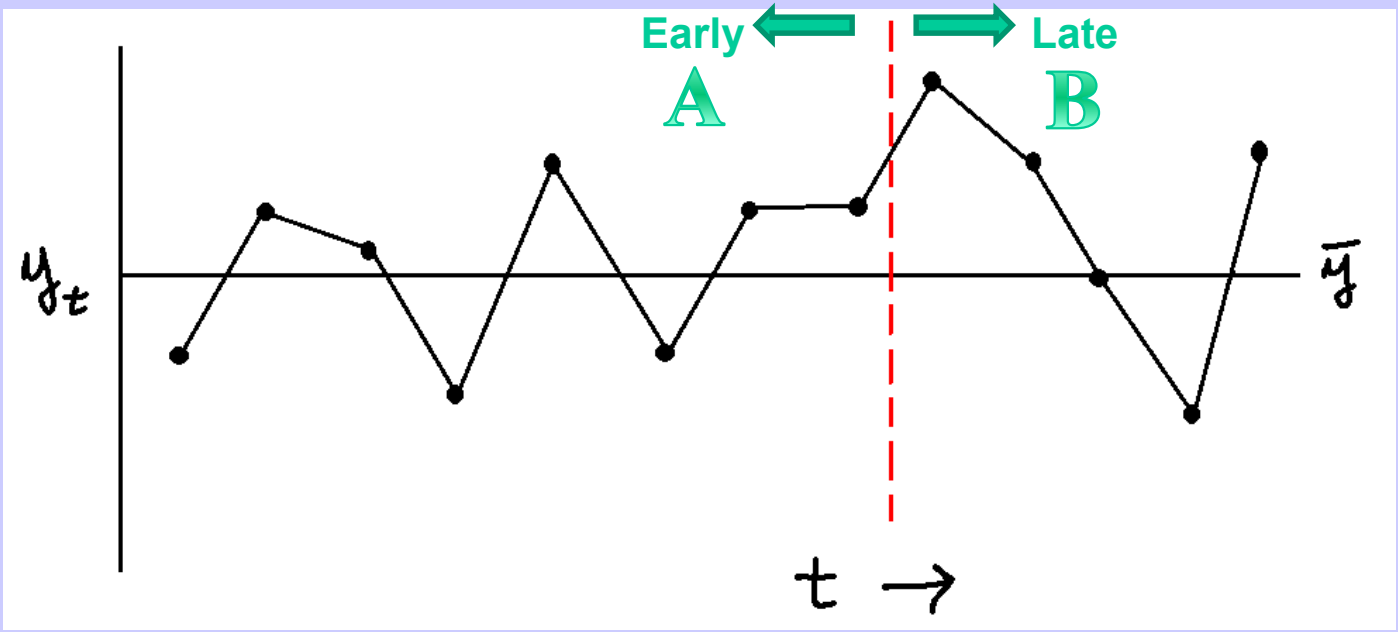
RE: is the reconstruction better than a null or reference reconstruction defined as a constant value at the calibration-period mean of the observed predictand for each year of the validation period?



RE Statistic usually evaluated by one or both of two methods:

- Split-sample validation
- Cross-validation

Split-sample validation: setup



$$\mathrm{RMSE}_B = \sqrt{\frac{1}{N_B}\sum_B \hat{e}^2}$$

1) Calibrate on A, validate on B → RE_B, $RMSE_B$
2) Calibrate on B, validate on A → RE_A, $RMSE_A$

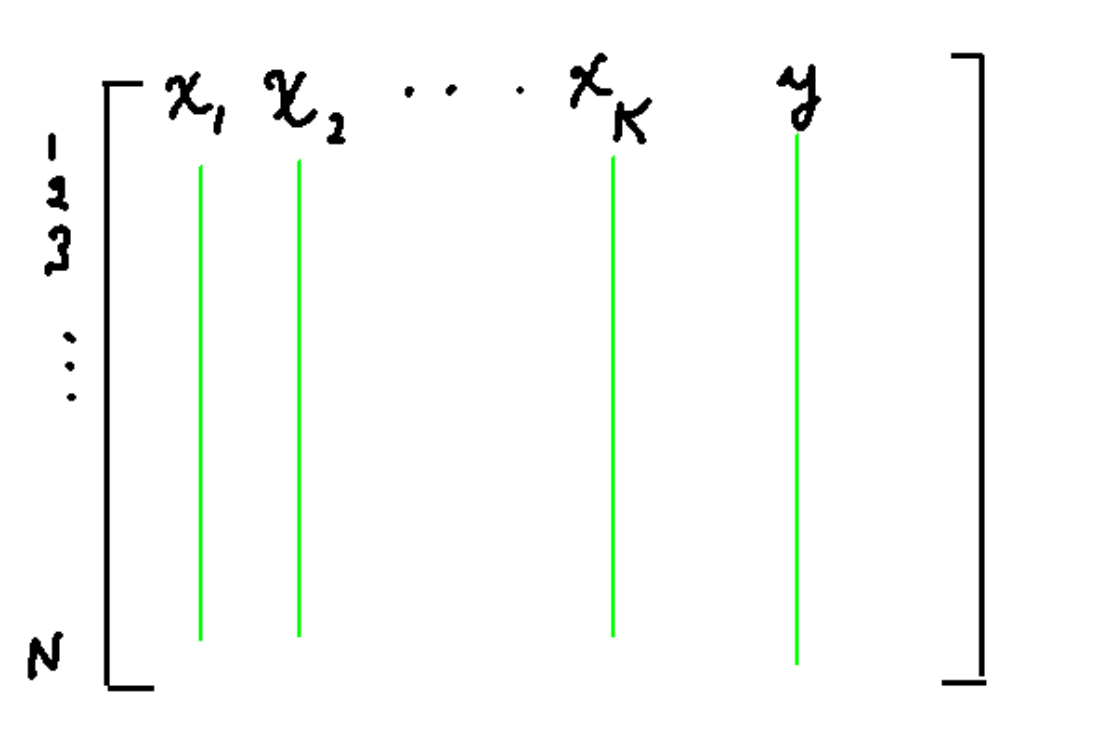
Split-sample validation: usual steps

1) Calibrate on full (A+B) period (e.g., by stepwise)
2) Use the identified predictors and go through the split-sample validation (1&2, previous slide)
3) Defend skill with RE_A, RE_B
4) Use full-period model (A+B) for long-term reconstruction

Weak points

- Ideally, want long time series (e.g., NA+NB>100)
- Model "validated" not same as model used for predictions (reconstructions)
- Using A+B initially is "peeking" at the validation data

Cross-validation: setup



- Given N years of overlap of predictand y and predictors x
- Calibrate and validate N different models, each time leaving out 1 observation and validating on the remaing N-1 observations

Cross-validation: steps

1) Omit i^{th} row, and calibrate on remaining $N-1$ rows

2) Predict $\hat{y}_i$, the "left out" observation of the predictand

3) Compute and store "deleted residual", $\hat{e}_{(i)}$

4) Repeat 1-3 for all observations -->$\hat{e}_{(i)}$, $i = 1, 2, \ldots N$

5) Compute RMSEV, the root-mean-square error of validation

$$\mathrm{RMSE_v} = \sqrt{\frac{\sum_N \hat{e}_{(i)}^2}{N}}, \qquad \text{and } \mathrm{RE_v}$$

SSV vs CV: strong points and weak points
Split-sample validation vs cross-validation

1) SSV requires long period of overlap (e.g., 100 obs) of predictor and predictand time series; CV can get by with much shorter overlap

2) Models "validated" in SSV often differ greatly from model actually used to generate predictions (reconstructions)

3) "Leave-1-out" CV may not be appropriate when lags complicate relationship between y and x; may need to omit more than 1 obs each step (to ensure "independence" of calibration and validation data)

4) SSV can give explicit information on ability of the predictors to track low-frequency variations in the predictand (true only if the overlap of x and y data actually contains appreciable low-frequency variation). CV does not give this information directly, but extended analysis, such as cross-spectral analysis of observed and predicted y can ge used to get it. This still requires, however, that the overlap of y and x contain low-frequency variation.

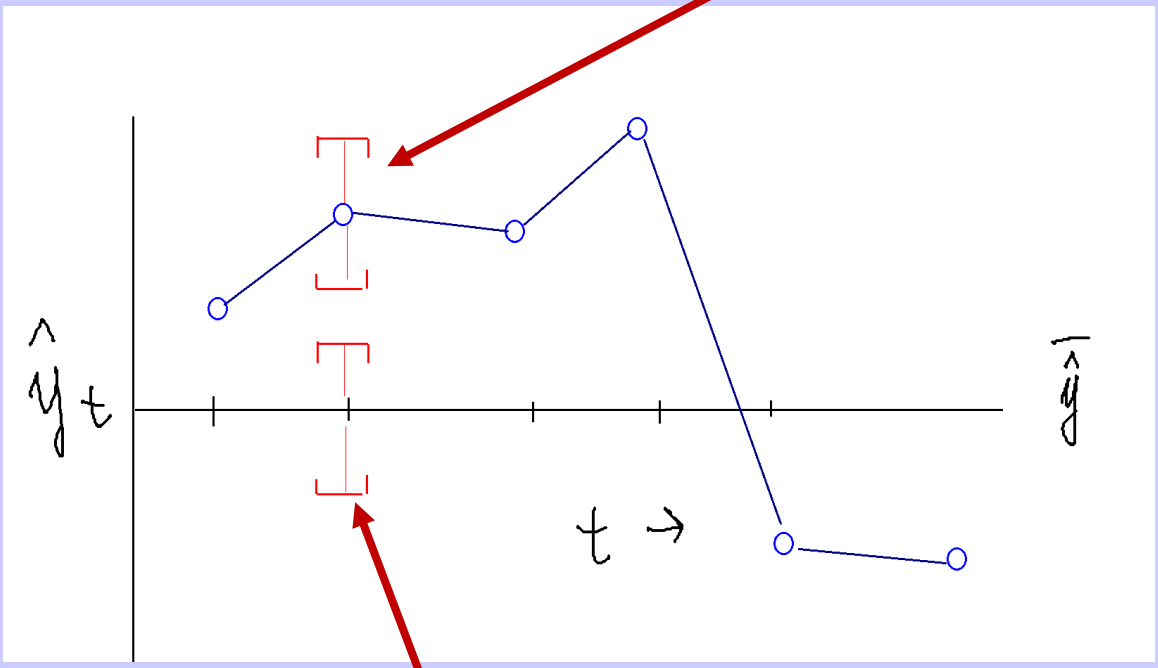
Error bars on reconstructions – alternative plotting positions

Assume reconstruction $\hat{y}_t$

and an estimated 95% confidence interval $\pm \Delta y$

Option 1: plot CI around the estimate:

True (unkown) value of predictand has 95% of falling in interval



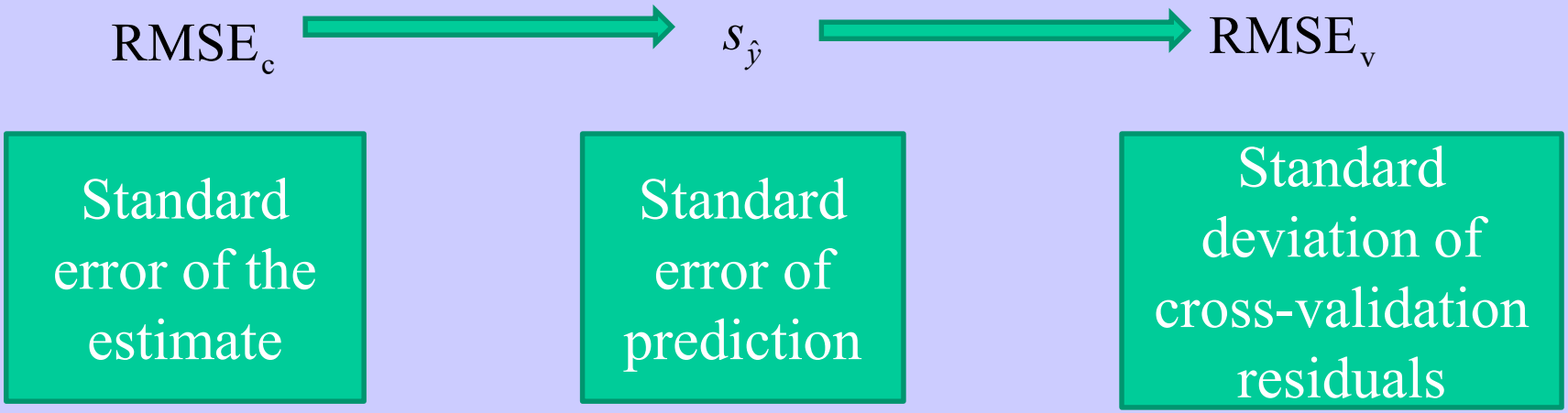
As long as the confidence interval is symmetrical, either way of plotting the CI is acceptable. Both directly show whether the reconstructed value is significantly different than the mean of the predictand in the particular year

Option 2: plot CI around the calibration-period mean of $\hat{y}_t$

- Less messy looking plot than option 1
- Less informative than option 1 about plausible values of unknown y_t

Error bars: how to estimate?

Three alternatives



1) Error bars from standard error of the estimate

(Standard error of estimate is same thing as root-mean-square error of calibration)

Assume:

Regression model with k predictors

N observations in calibration period

$$\hat{e}_t = y_t - \hat{y}_t$$

Mean square error of calibration

$$\text{SSE} = \sum_{t=1}^{N} \hat{e}_t^2$$

$$\text{MSE} = \text{SSE} / \left(N - k - 1 \right)$$

$$s_e \equiv \text{RMSE}_c = \sqrt{\text{MSE}}$$

Resulting 95% CI is ±2 RMSE, *and is a scalar*

2) Error bars from standard error of prediction

$$s_{\hat{y}} = s_e \left[1 + \frac{1}{N} + \frac{(x_* - \bar{x})^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \right]^{1/2}$$

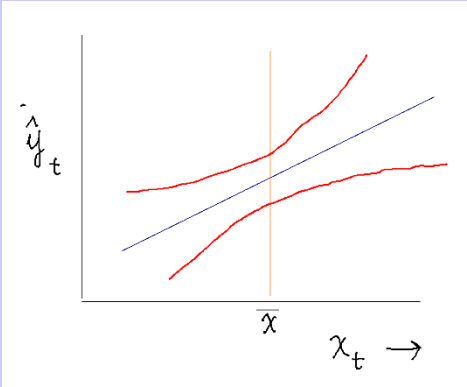
s_e = standard error of the estimate (previous slide)

N = number of calibration-period observations

x_* = value of the predictor in some year outside the calibration period

$x_i, i = 1, N$ values of calibration-period observations

- *Always greater than standard error of estimate*
- *Depends on how far the predictor is from its calibration-period mean*
- *Not a scalar: differs for each year of the reconstruction*
- *This equation for simple linear regression only (see notes for more complicated equation that applies to MLR)*



Flaring of CI around estimate as predictor differs more and more from its calibration mean

3) Error bars from cross-validation residuals

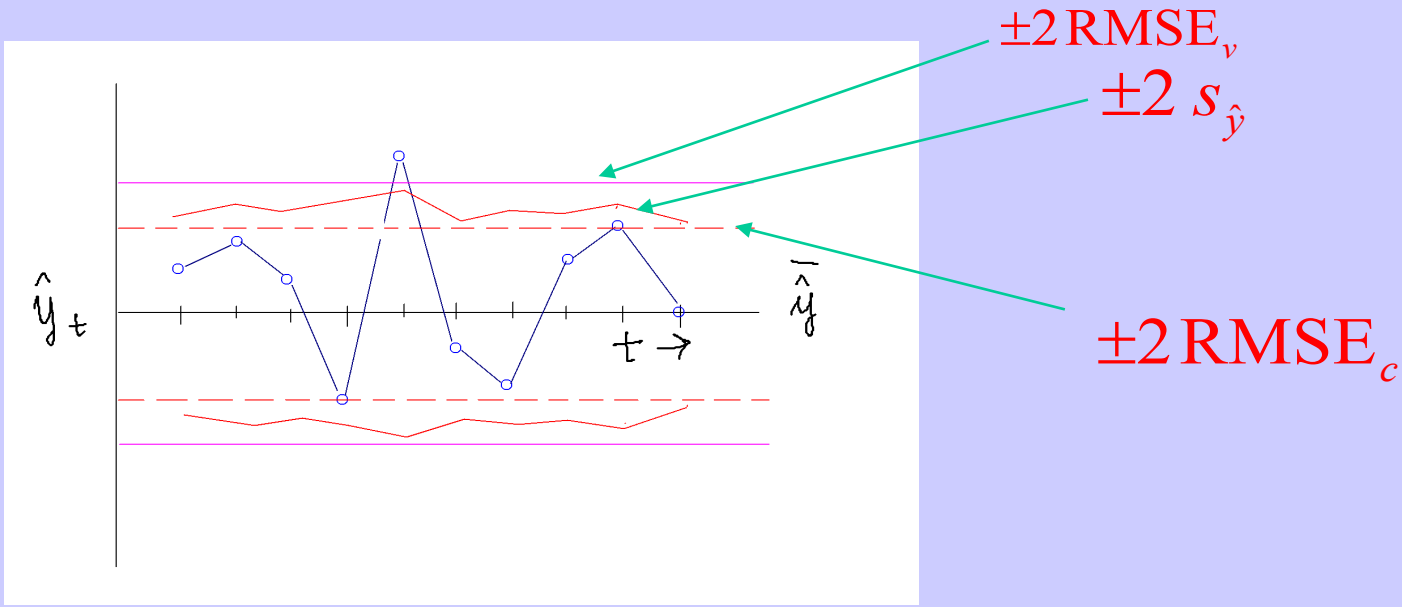
Resulting 95% CI is ±2 $\mathrm{RMSE_v}$, *and is a scalar*

$$\mathrm{RMSE_v} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\hat{e}_{(i)}^2}$$

$$\hat{e}_{(i)} = \text{deleted residual}$$

$$\mathrm{PRESS} = \sum_{i=1}^{N}\hat{e}_{(i)}^2 \quad \equiv \quad \text{"PRESS statistic"}$$

Error bars - comparative



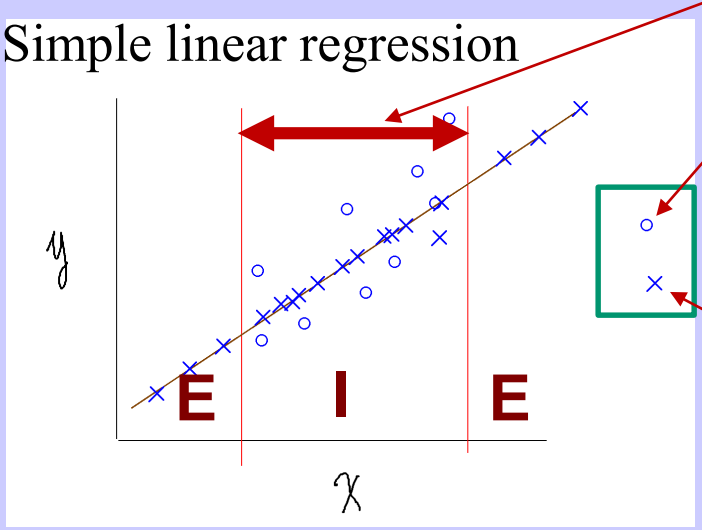
Given RMSE_c, $s_{\hat{y}}$, or RMSE_v and an assumed distribution of errors, can get CI

Assume $e_t \sim \mathrm{N}\left(0, \mathrm{RMSE}_v^2\right)$, say

95% CI is $\simeq \hat{y}_t \pm 2\mathrm{RMSE}_v$, if around data

$\simeq \bar{\hat{y}}_t \pm 2\mathrm{RMSE}_v$, if around calibration mean

Extrapolation vs Interpolation (E vs I)

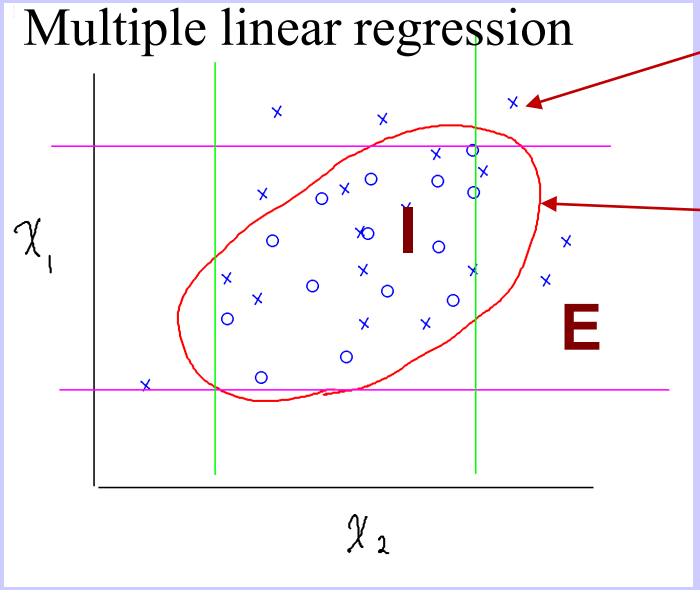


Calibration period range of x

Calibration data

Predictions

Predictions are defined as E vs I depending on whether the predictors for a given predicted value are inside or outside their "multivariate space" defined by the calibration period



Predictions from these x are extrapolations

Predictions from x inside this bivariate calibration-period space are interpolations