Psych 253
Advanced Statistical Modeling

Modeling: Basic framework and strategy

Daniel Yamins
Wu Tsai Neurosciences Institute
Departments of Psychology and Computer Science
Stanford Artificial Intelligence Laboratory
Stanford University

Russ Poldrack
Department of Psychology
Stanford University

Science is ultimately about measurement and modeling

Here we will ask:

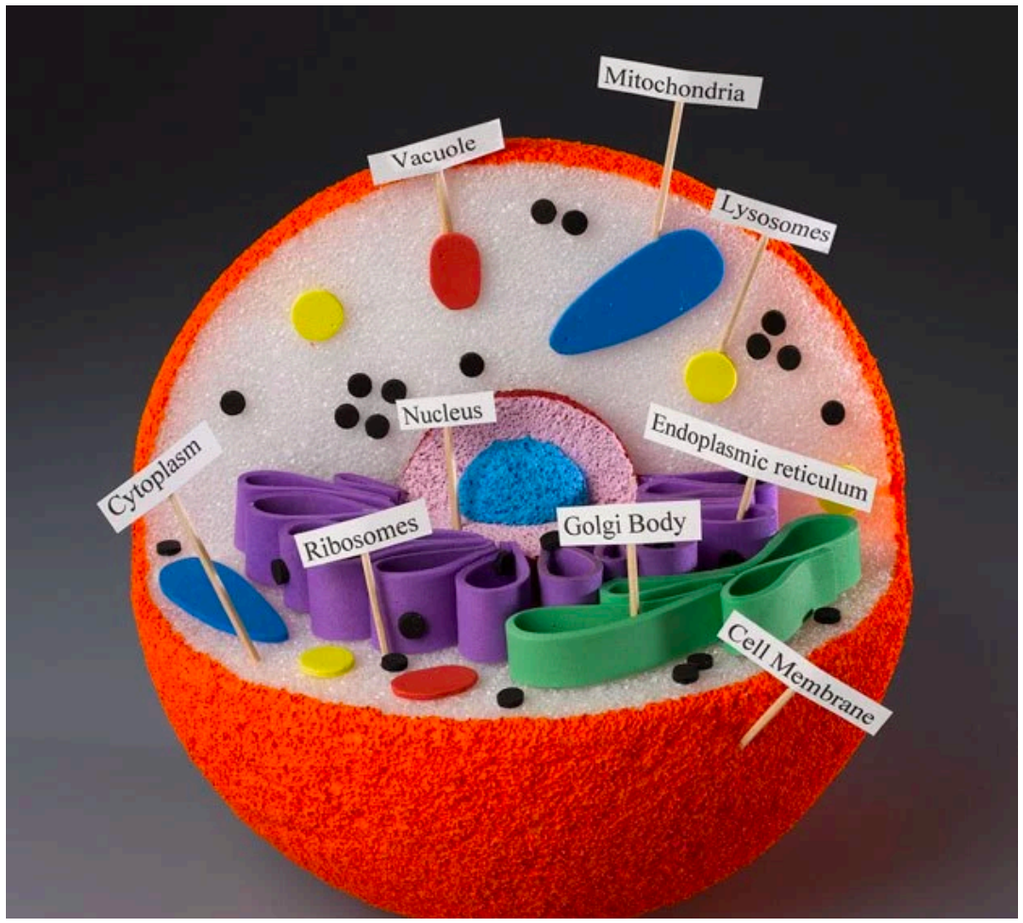
what is a "model", and what makes one model "better" than another?

What is a model?

A simplified description of a system of interest

"all models are wrong, but some are useful” - G.E.P. Box

Models simplify the world for us



The basic statistical model

$$y = f(x) + \epsilon$$

what we actually observe (the data)

what we expect to observe (our prediction) based on known features x

difference between expected and observed (error)

The basic statistical model

$$y = f(x) + \epsilon$$

Our goal is to learn *f(x)* - that is, to approximate the function best relates X and Y

How do we decide which possible *f(x)* is "best"?

Two aspects of a statistical model

$$y = f(x, \theta) + \epsilon$$

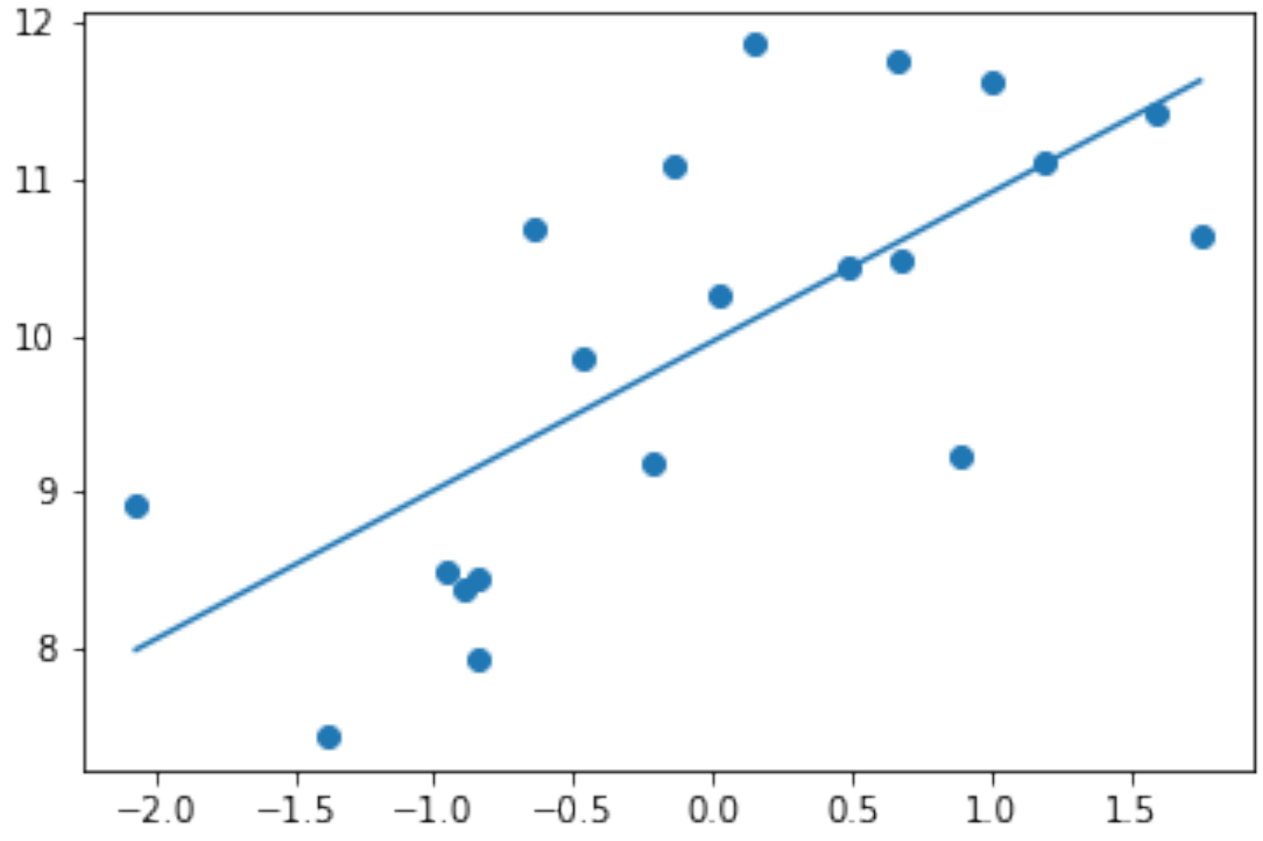
known features

estimated parameters

- The structure of $f(x, \theta)$
  - A description of the process by which X leads to y
  - We generally determine this *a priori* or through model search/comparison
- The particular parameter values θ
  - We learn these from data through optimization

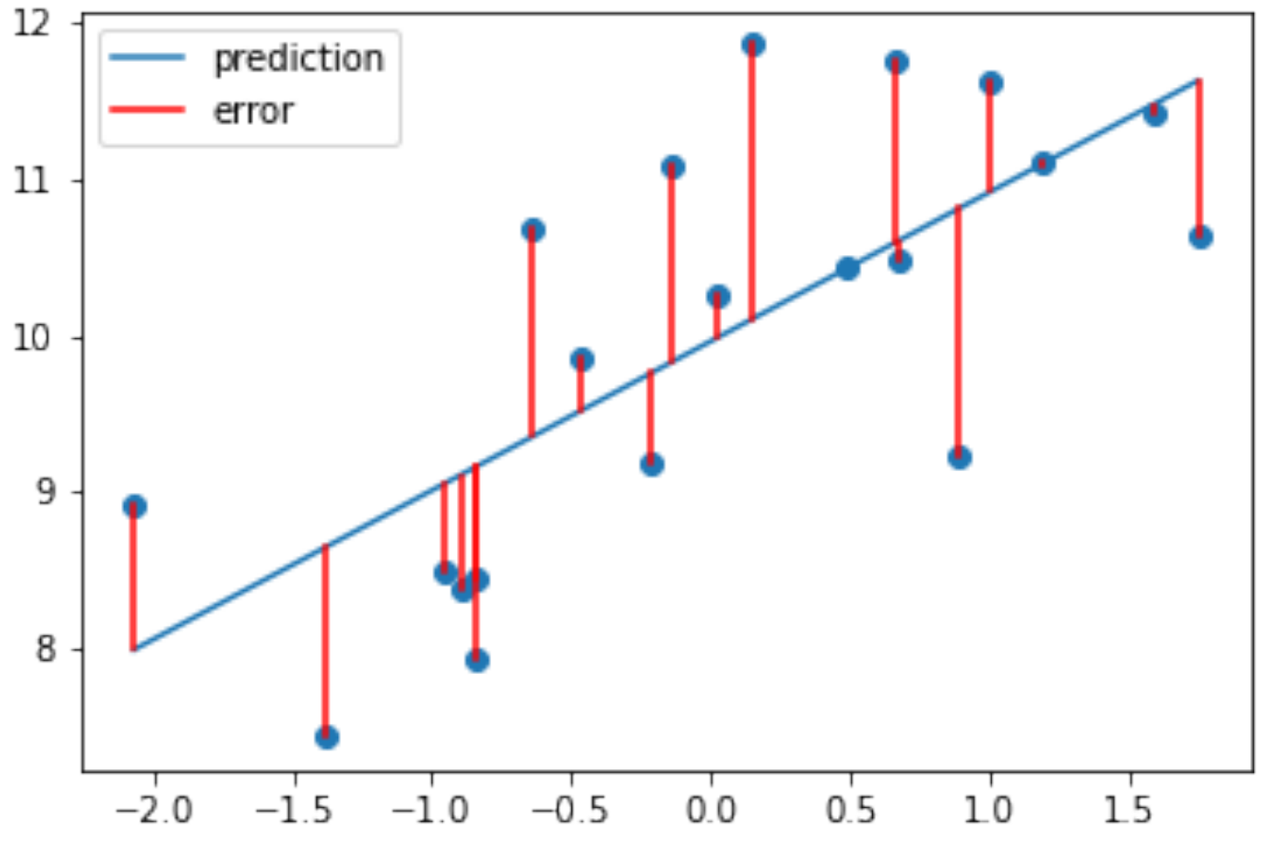
Quantifying error in modeling

$$y = f(x, \theta) + \epsilon$$



Quantifying error in modeling

$$y = f(x, \theta) + \epsilon$$



How can we summarize the error in a single value?

Loss function

The loss quantifies how well our model fits the data (subject to whatever constraints we want to place on it)

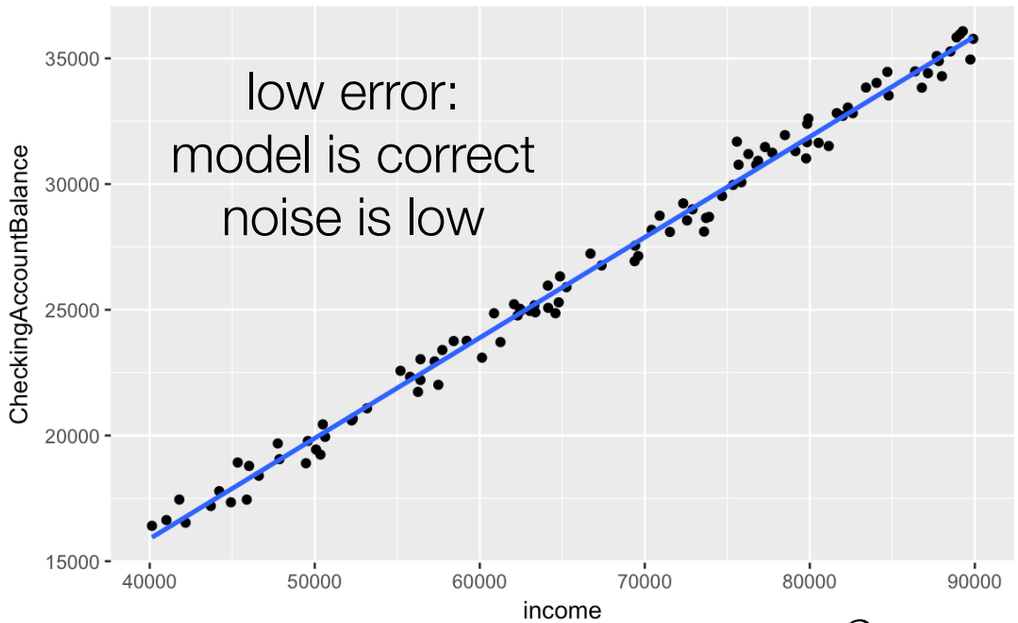
commonly used loss functions:

$$\text{mean squared error} = \frac{\sum_i^N (y_i - \hat{y}_i)^2}{N}$$

$$\text{mean absolute error} = \frac{\sum_i^N |y_i - \hat{y}_i|}{N}$$

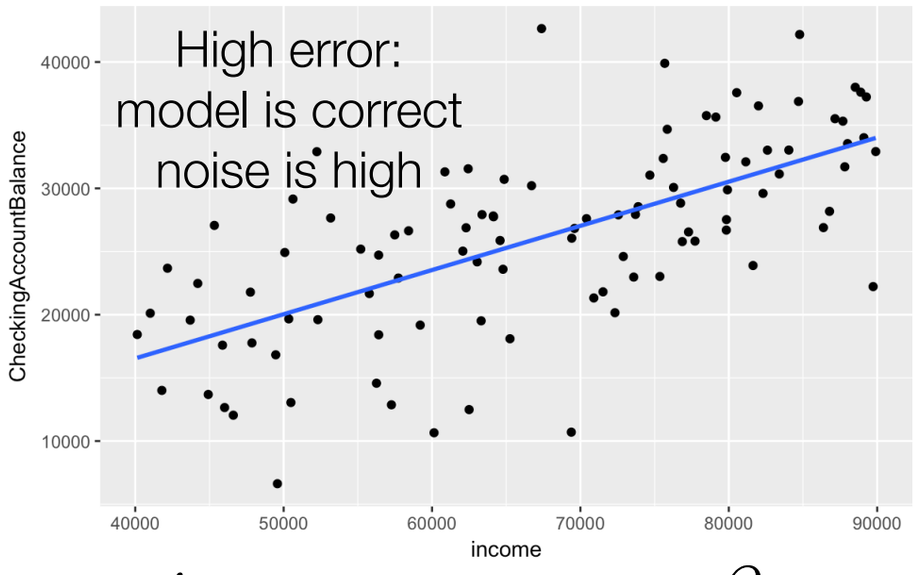
We use optimization to find the parameter values θ that minimize our loss function

Sources of error in modeling



$$true : y = x * \beta$$

$$model : y = x * \beta$$

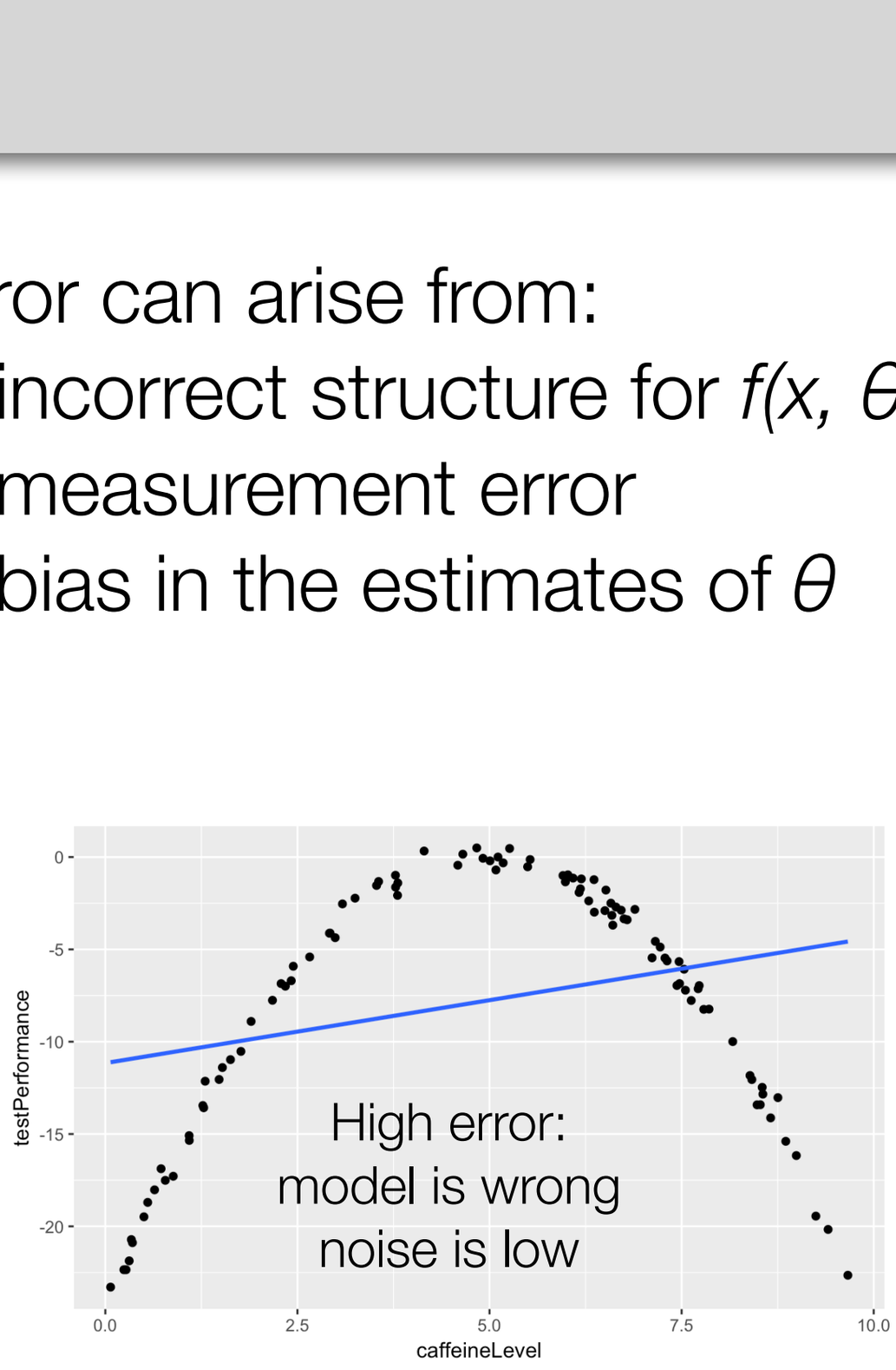


$$true : y = x * \beta$$

$$model : y = x * \beta$$

Error can arise from:

- incorrect structure for *f(x, θ)*
- measurement error
- bias in the estimates of *θ*

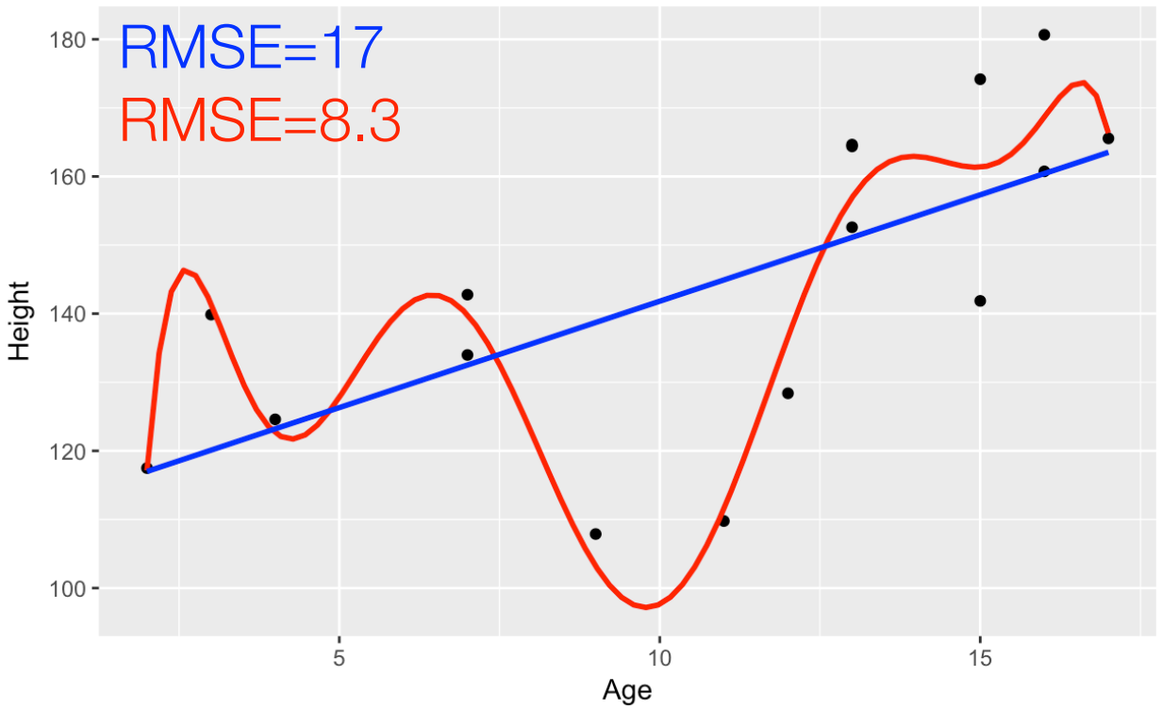


$$true : y = x^2 * \beta$$

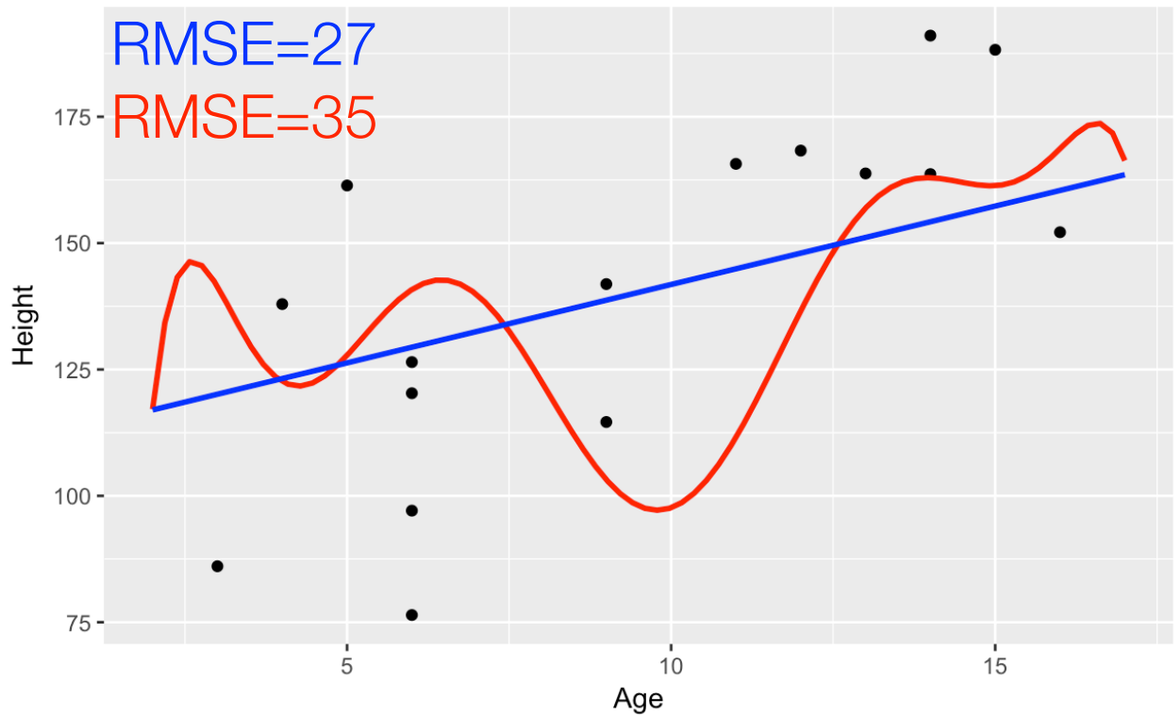
$$model : y = x * \beta$$

What makes a model "good"?

Low error on original sample



Low error on new sample



- A more complex model will always fit the data better
  - The model increasing "overfits" the random noise in the data as it becomes more complex
- The model whose parameters match the true population structure should provide the best prediction to new samples from the same population
  - How can we determine that?

Cross Validation

https://en.wikipedia.org/wiki/Cross-validation_(statistics)

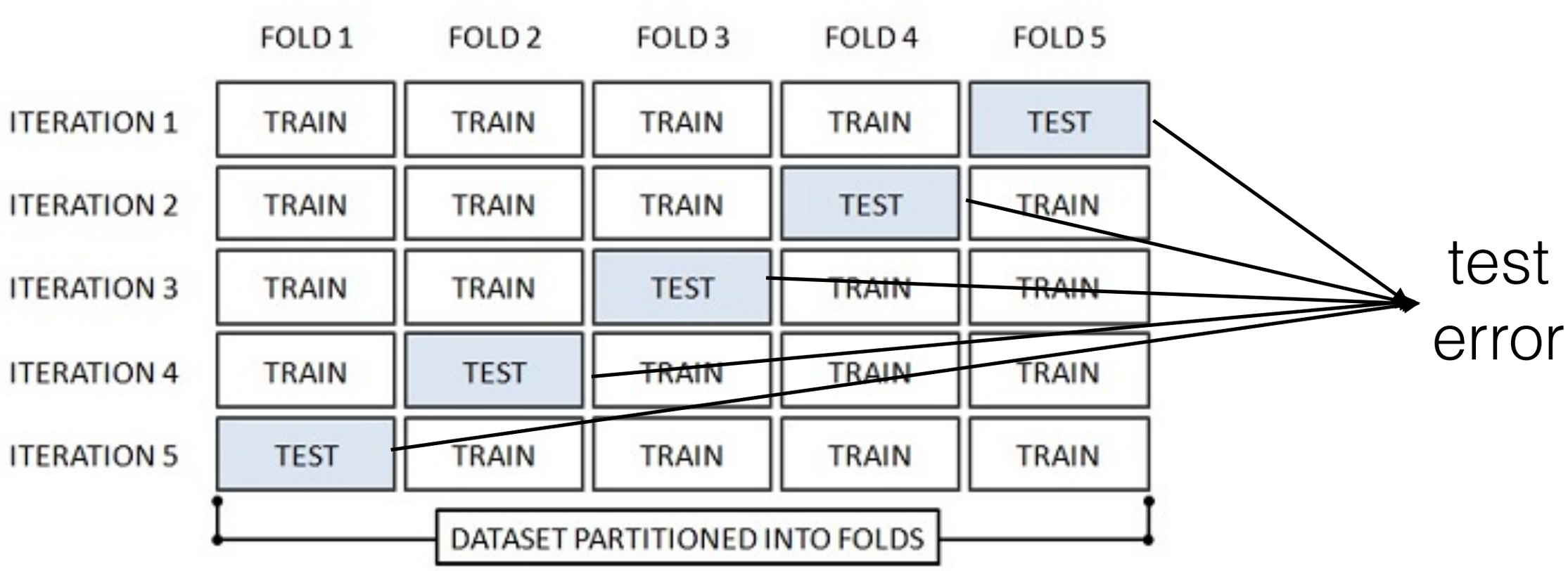
"Cross-validation is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set."

Cross Validation

https://en.wikipedia.org/wiki/Cross-validation_(statistics)

"Cross-validation is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set."

```
Procedure:
    1) estimate model params on training data
    2) evaluate performance on testing data
```



https://www.dummies.com/programming/big-data/data-science/resorting-cross-validation-machine-learning/

Cross Validation

https://en.wikipedia.org/wiki/Cross-validation_(statistics)

"Cross-validation is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set."

```
Procedure:
    1) estimate model params on training data
    2) evaluate performance on testing data
```

Q: Why are we doing this?

Cross Validation

https://en.wikipedia.org/wiki/Cross-validation_(statistics)

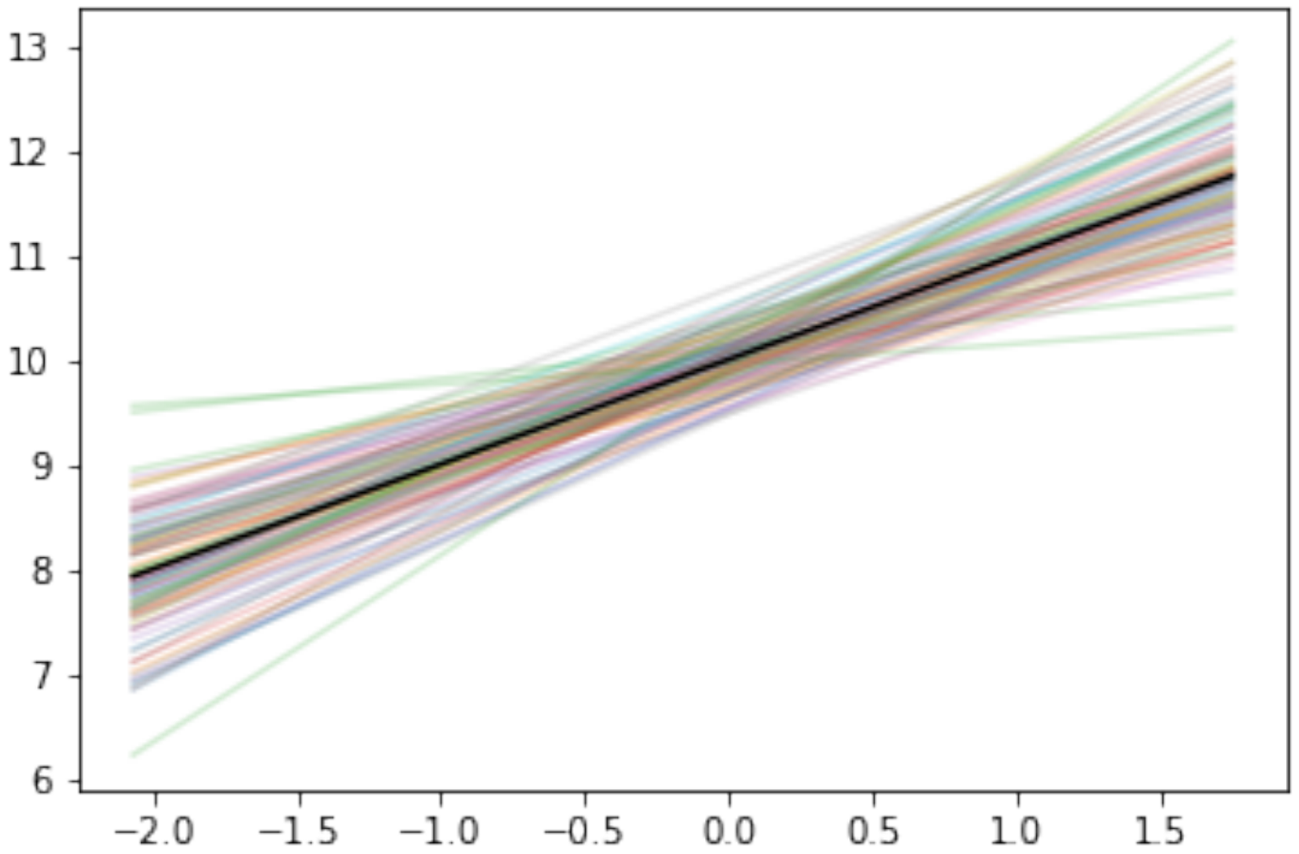
"Cross-validation is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set."

```
Procedure:
    1) estimate model params on training data
    2) evaluate performance on testing data
```

Q: Why are we doing this?

A: To prevent overfitting (fooling ourselves)

Noise in model fitting



- The fit of any model includes contributions from the true underlying signal as well as from measurement noise

J. R. Statist. Soc. B (1983),
45, *No.* 3, *pp.* 311–354

Regression, Prediction and Shrinkage

By J. B. COPAS

University of Birmingham, UK

[*Read before the* Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, January 12th, 1983*, Professor R. N. Curnow *in the Chair*]

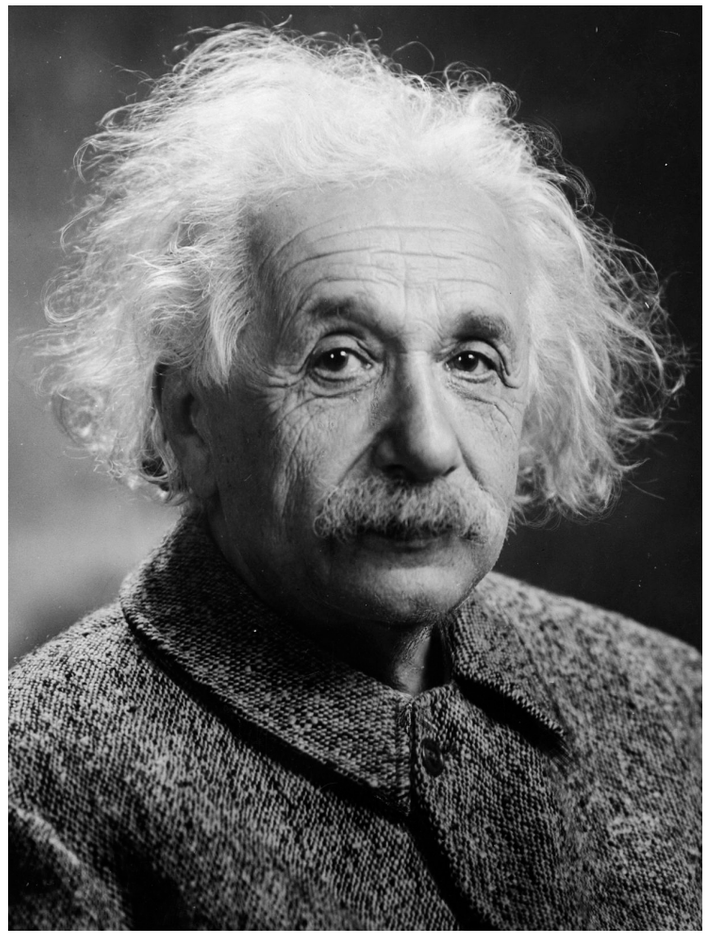
SUMMARY

The fit of a regression predictor to new data is nearly always worse than its fit to the original data. Anticipating this shrinkage leads to Stein-type predictors which, under certain assumptions, give a uniformly lower prediction mean squared error than least squares. Shrinkage can be particularly marked when stepwise fitting is used: the shrinkage is then closer to that expected of the full regression rather than of the subset regression actually fitted. Preshrunk predictors for selected subsets are proposed and tested on a number of practical examples. Both multiple and binary (logistic) regression models are considered.

“Since any assessment of retrospective fit "uses the data twice", it is obvious that it gives too optimistic a picture of the validation fit likely to be obtained on new data.”

The principle of parsimony

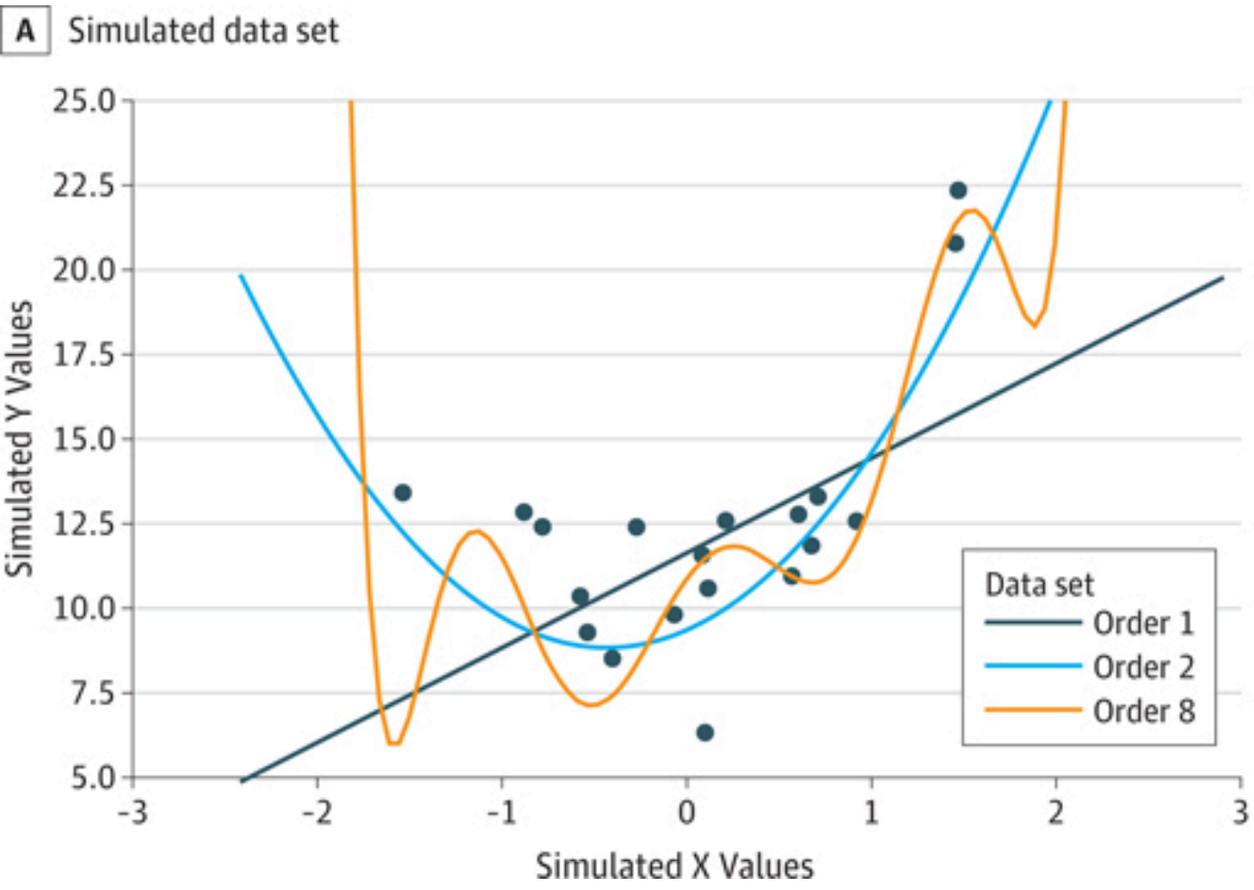
- "It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience."
  - Albert Einstein, 1933
- Paraphrased as "everything should be as simple as it can be, but not simpler"

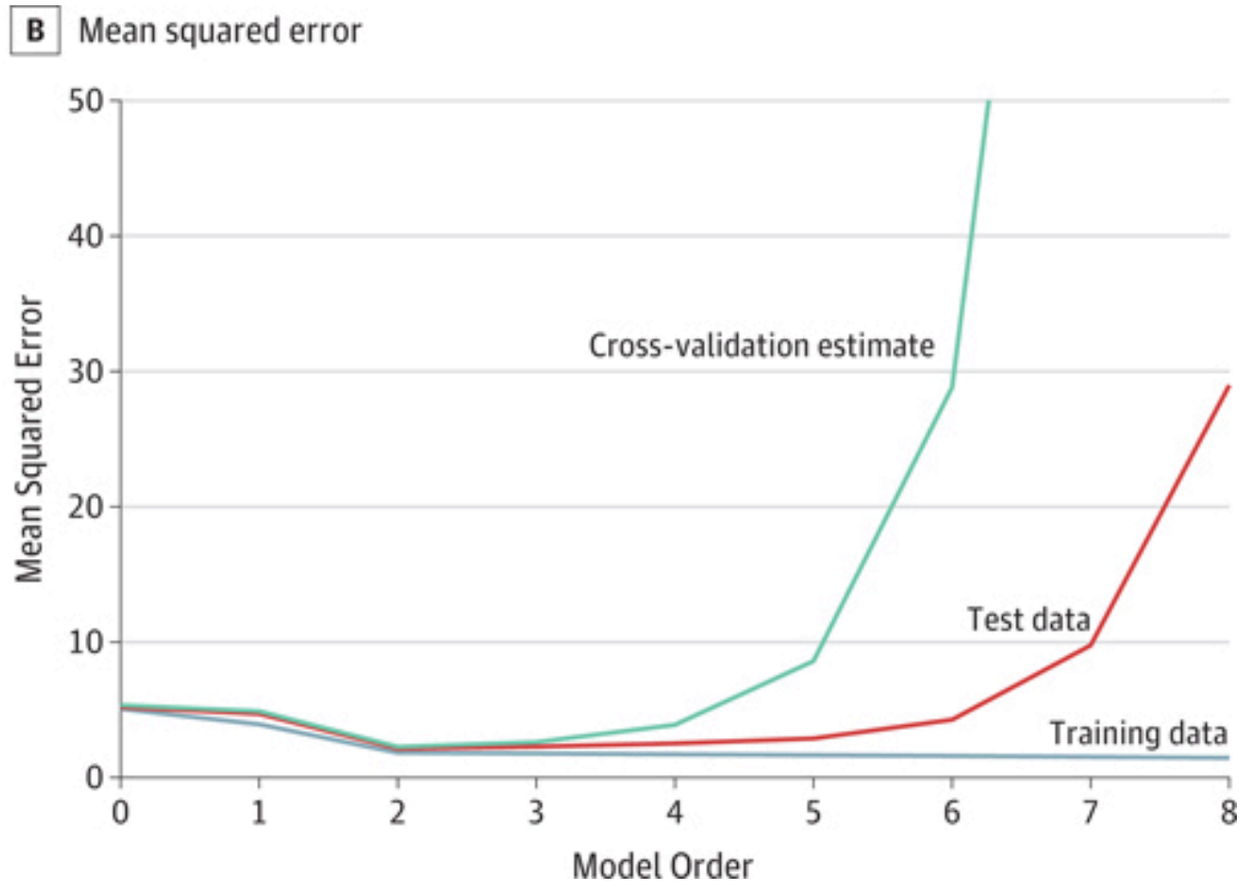


Selecting models using cross-validation

Data generated from quadratic model

Error estimated using cross-validation





Test error is minimized for true model order (2) for both CV and new sample

Poldrack et al., 2020, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7250718/

Cross Validation

Q: How do you get the train/test splits?

1. Leave-***p***-out: train on all-but-***p***, test on ***p*** for all subsets of ***p*** stimuli; average results

Cross Validation

Q: How do you get the train/test splits?

1. Leave-***p***-out: train on all-but-***p***, test on ***p*** for all subsets of ***p*** stimuli; average results

2. ***k***-fold: partition data into ***k*** equal-sized subsets; train on all but on subset; test on held-out slice; average results

Cross Validation

Q: How do you get the train/test splits?

1. Leave-***p***-out: train on all-but-***p***, test on ***p*** for all subsets of ***p*** stimuli; average results

2. ***k***-fold: partition data into ***k*** equal-sized subsets; train on all but on subset; test on held-out slice; average results

3. random subsampling (aka Shuffle-split): randomly choose training subset and non-overlapping testing subset; average results over many random choices

Cross Validation

Q: How do you get the train/test splits?

1. Leave-***p***-out: train on all-but-***p***, test on ***p*** for all subsets of ***p*** stimuli; average results

2. ***k***-fold: partition data into ***k*** equal-sized subsets; train on all but on subset; test on held-out slice; average results

3. random subsampling (aka Shuffle-split): randomly choose training subset and non-overlapping testing subset; average results over many random choices

Ensuring that cross validation works properly

test data cannot be used in any part of the analysis
prior to assessment of test accuracy

An example of "double dipping" in cross-validation

Individual Classification of Mild Cognitive Impairment Subtypes by Support Vector Machine Analysis of White Matter DTI

S. Haller, P. Missonnier, F.R. Herrmann, C. Rodriguez, M.-P. Deiber, D. Nguyen, G. Gold, K.-O. Lovblad, and P. Giannakopoulos

AJNR Am J Neuroradiol 34:283–91 Feb 2013

MATERIALS AND METHODS: Sixty-six prospective participants were included: 18 with sd-aMCI, 13 with sd-fMCI, and 35 with md-aMCI. Statistics included group comparisons using TBSS and individual classification using SVMs.

The individual-level SVM analysis provided discrimination between the MCI subtypes with accuracies around 97%.

The analysis included 2 steps. In the first step, we performed a feature selection. The rationale behind this step is that not all voxels discriminate between groups… The second step consisted of the "actual" classification analyses for each comparison by using the SVM algorithm.

Table 4: Individual SVM classification based on DTI FA TBSS

	md-aMCI versus sd-fMCI	md-aMCI versus sd-aMCI	sd-fMCI versus sd-aMCI
Number of subjects	34/11	34/15	11/15
Chance rate	0.76	0.69	0.58
SVM analysis			
Accuracy	98.40 (5.90)	97.70 (6.61)	99.67 (3.33)
TP rate	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
FP rate	0.06 (0.23)	0.07 (0.20)	0.01 (0.05)
TN rate	0.94 (0.23)	0.94 (0.20)	1.00 (0.05)
FN rate	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)

Note:—Accuracy, true-positive (TP), false-positive (FP), true-negative (TN), and false-negative (FN) rates for individual classifications using a SVM classifier. Note that the accuracy is calculated as average accuracy of 10 repetitions using 10-fold cross-validation (average and standard deviation).

An example of "double dipping" in cross-validation

The "Peeking" Effect in Supervised Feature Selection on Diffusion Tensor Imaging Data

S. Diciotti et al.

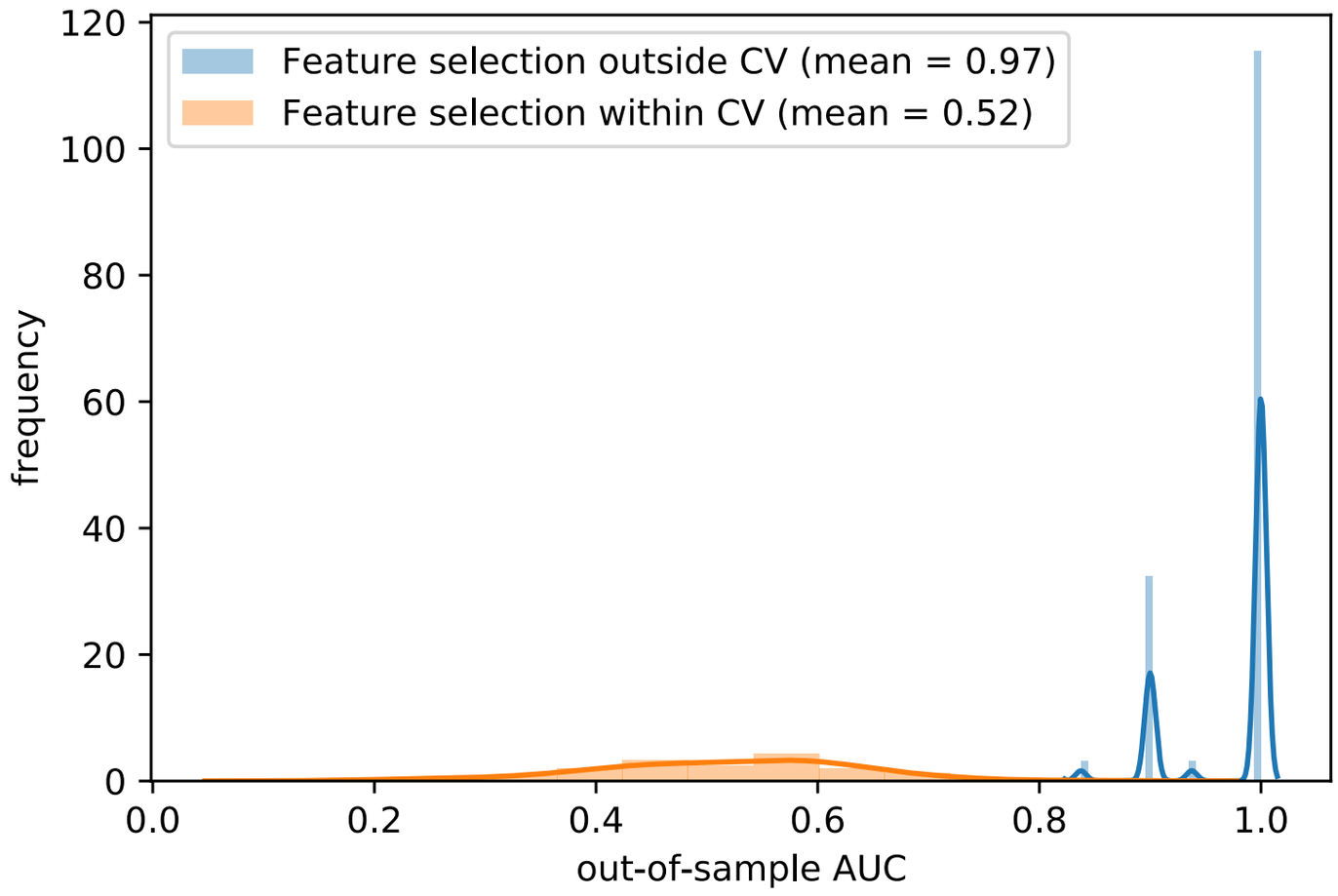
AJNR Am J Neuroradiol ●:● ● 2013

The above-mentioned study presents a questionable use of supervised feature selection, which was performed on the entire dataset (ie, on both training and test data) instead of only on the training set of each partition generated during the cross-validation procedure.

We attempted to discriminate between 30 patients with amnesic MCI and 21 with mild AD by using the processing pipeline and the same type of data used by Haller et al. We repeated the analysis by using either incorrect cross-validation (ie, feature selection on the entire dataset followed by classification in cross-validation, as carried out by Haller et al[1]) or correct cross-validation (feature selection within each training set of the cross-validation).

In the former analysis, patients with mild AD were classified with 80.0% sensitivity and 96.7% specificity, while in the latter analysis, results dropped to 45.3% sensitivity and 67.3% specificity.

Simulation: Feature section inside vs outside of CV loop



Ensuring that cross validation works properly

test data cannot be used in any part of the analysis
prior to assessment of test accuracy

cross validation should *always* encompass the entire
analysis pipeline