An Introduction to Machine Learning with Stata

Achim Ahrens

Public Policy Group, ETH Zürich

Presented at the
XVI Italian Stata Users Group Meeting
Florence, 26-27 September 2019

The plan for the workshop

Preamble: What is Machine Learning?

- Supervised vs unsupervised machine learning
- Bias-variance trade-off

Session I: Examples of Machine Learners

- Tree-based methods, SVM
- Using Python for ML in with Stata
- Cluster analysis

Session II: Regularized Regression in Stata

- Lasso, Ridge and Elastic net, Logistic lasso
- `lassopack` and Stata 16's `lasso`

Session III: Causal inference with Machine Learning

- Post-double selection
- Double/debiased Machine Learning
- Other recent developments

Let's talk terminology

Machine learning constructs algorithms that can learn from the data.

Statistical learning is branch of Statistics that was born in response to Machine learning, emphasizing statistical models and assessment of uncertainty.

Robert Tibshirani on the difference between ML and SL (jokingly):

Large grant in Machine learning:	$1,000,000
Large grant in Statistical learning:	$50,000

Let's talk terminology

Artificial intelligence deals with methods that allow systems to interpret & learn from data and achieve tasks through adaption.

This includes robotics, natural language processing. ML is a sub-field of AI. . . .

Data science is the extraction of knowledge from data, using ideas from mathematics, statistics, machine learning, computer programming, data engineering, etc.

Deep learning is a sub-field of ML that uses artificial neural networks (not covered today).

Let's talk terminology

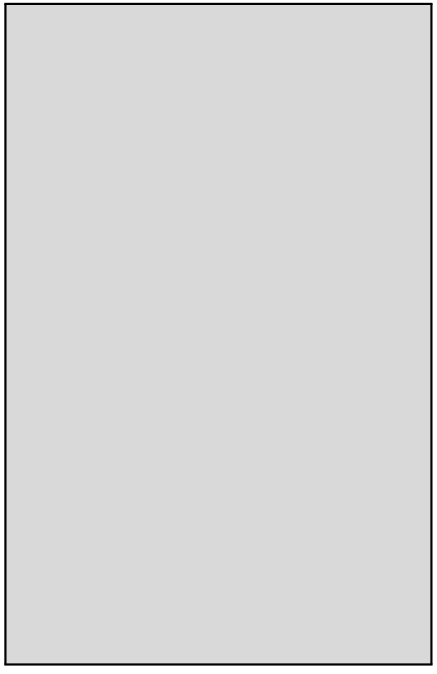
Big data is not a set of methods or a field of research. Big data can come in two forms:

Wide ('high-dimensional') data

Many predictors (large p) and *relatively* small N.

Typical method:
Regularized regression

Tall or long data



Many observations, but only few predictors.

Typical method:
Tree-based methods

Let's talk terminology

Supervised Machine Learning:

- You have an outcome Y and predictors X.
- Classical ML setting: independent observations.
- You fit the model Y want to predict (classify if Y is categorical) using unseen data X_0.

Unsupervised Machine Learning:

- No output variable, only inputs.
- Dimension reduction: reduce the complexity of your data.
- Some methods are well known: Principal component analysis (PCA), cluster analysis.
- Can be used to generate inputs (features) for supervised learning (e.g. Principal component regression).

Econometrics vs Machine Learning

Econometrics

- Focus on parameter estimation and *causal* inference.
- Forecasting & prediction is usually done in a parametric framework (e.g. ARIMA, VAR).
- *Methods:* Least Squares, Instrumental Variables (IV), Generalized Methods of Moments (GMM), Maximum Likelihood.
- *Typical question:* Does x have a causal effect on y?
- *Examples:* Effect of education on wages, minimum wage on employment.
- *Procedure:*
  - Researcher specifies model using diagnostic tests & theory.
  - Model is estimated using the full data.
  - Parameter estimates and confidence intervals are obtained based on large sample asymptotic theory.
- *Strengths:* Formal theory for estimation & inference.

Econometrics vs Machine Learning

Supervised Machine Learning

- Focus on prediction & classification.
- *Wide set of methods:* regularized regression, random forest, regression trees, support vector machines, neural nets, etc.
- General approach is 'does it work in practice?' rather than 'what are the formal properties?'
- *Typical problems:*
  - Netflix: predict user-rating of films
  - Classify email as spam or not
  - Genome-wide association studies: Associate genetic variants with particular trait/disease
- *Procedure:* Algorithm is trained and validated using 'unseen' data.
- *Strengths:* *Out-of-sample* prediction, high-dimensional data, data-driven model selection.

Motivation I: Model selection

The standard linear model

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \varepsilon_i.$$

Why would we use a fitting procedure other than OLS?

Model selection.

We don't know the true model. Which regressors are important?

Including too many regressors leads to **overfitting**: good in-sample fit (high R^2), but bad *out-of-sample* prediction.

Including too few regressors leads to **omitted variable bias**.

Motivation I: Model selection

The standard linear model

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \varepsilon_i.$$

Why would we use a fitting procedure other than OLS?

Model selection.

Model selection becomes even more challenging when the data is **high-dimensional**.

If p is close to or larger than n, we say that the data is high-dimensional.

- If $p > n$, the model is not identified.
- If $p = n$, perfect fit. Meaningless.
- If $p < n$ but large, overfitting is likely: Some of the predictors are only significant by chance (false positives), but perform poorly on new (unseen) data.

Motivation I: Model selection

The standard approach for model selection in econometrics is (arguably) hypothesis testing.

Problems:

- Our standard significance level only applies to *one* test.
- Pre-test biases in multi-step procedures. This also applies to model building using, e.g., the *general-to-specific approach*.
- Especially if p is large, inference is problematic. Need for false discovery control (multiple testing procedures)—rarely done.
- 'Researcher degrees of freedom' and 'p-hacking': researchers try many combinations of regressors, looking for statistical significance (Simmons et al., 2011).

Researcher degrees of freedom

"it is common (and accepted practice) for researchers to explore various analytic alternatives, to search for a combination that yields 'statistical significance,' and to then report only what 'worked.'" Simmons et al., 2011

Motivation II: High-dimensional data

The standard linear model

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \varepsilon_i.$$

Why would we use a fitting procedure other than OLS?

High-dimensional data.

Large p is often not acknowledged in applied work:

- The true model is unknown *ex ante*. Unless a researcher runs one and only one specification, the low-dimensional model paradigm is likely to fail.
- The number of regressors increases if we account for non-linearity, interaction effects, parameter heterogeneity, spatial & temporal effects.

Example: Cross-country regressions, where we have only small number of countries, but thousands of macro variables.

Motivation III: Prediction

The standard linear model

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \varepsilon_i.$$

Why would we use a fitting procedure other than OLS?

Bias-variance-tradeoff.

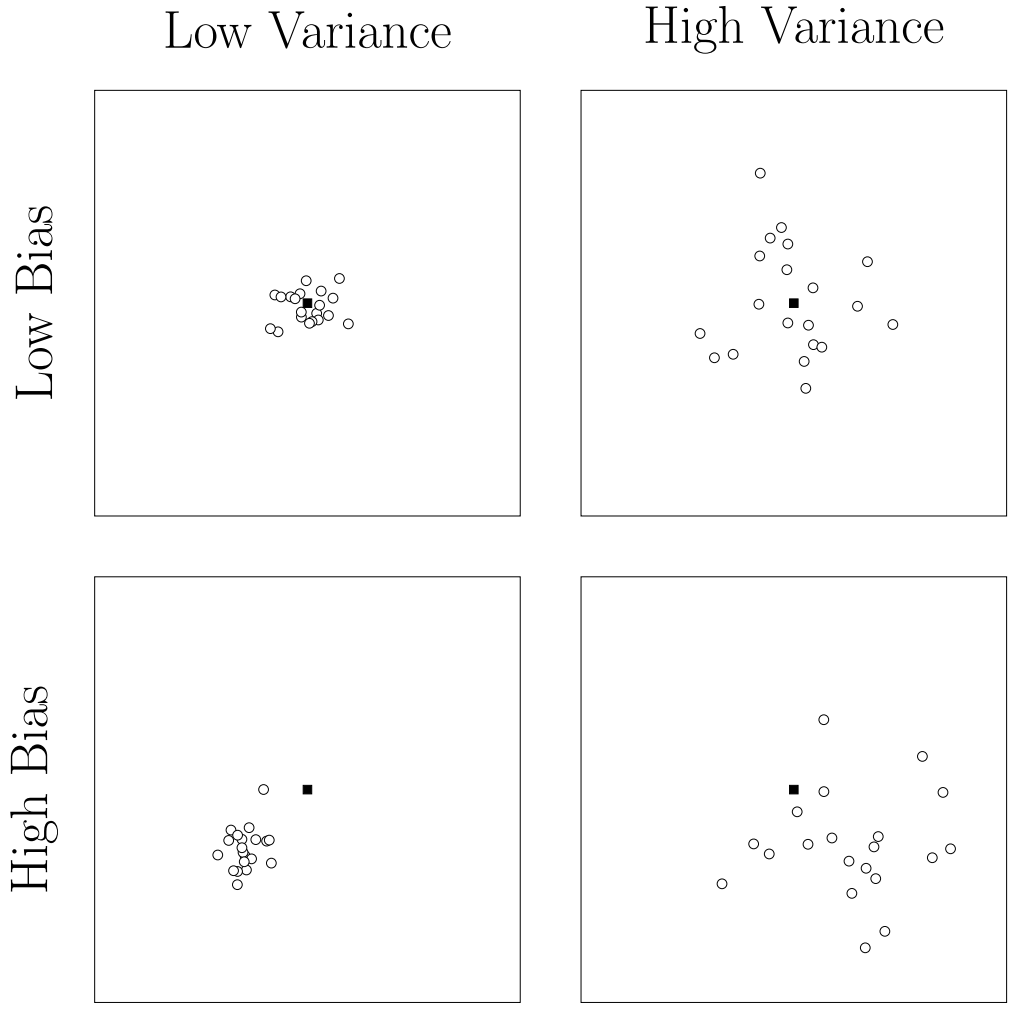
OLS estimator has zero bias, but not necessarily the best *out-of-sample* predictive accuracy.

Suppose we fit the model using the data $i = 1, \ldots, n$. The prediction error for y_0 given x_0 can be decomposed into

$$PE_0 = E[(y_0 - \hat{y}_0)^2] = \sigma_\varepsilon^2 + Bias(\hat{y}_0)^2 + Var(\hat{y}_0).$$

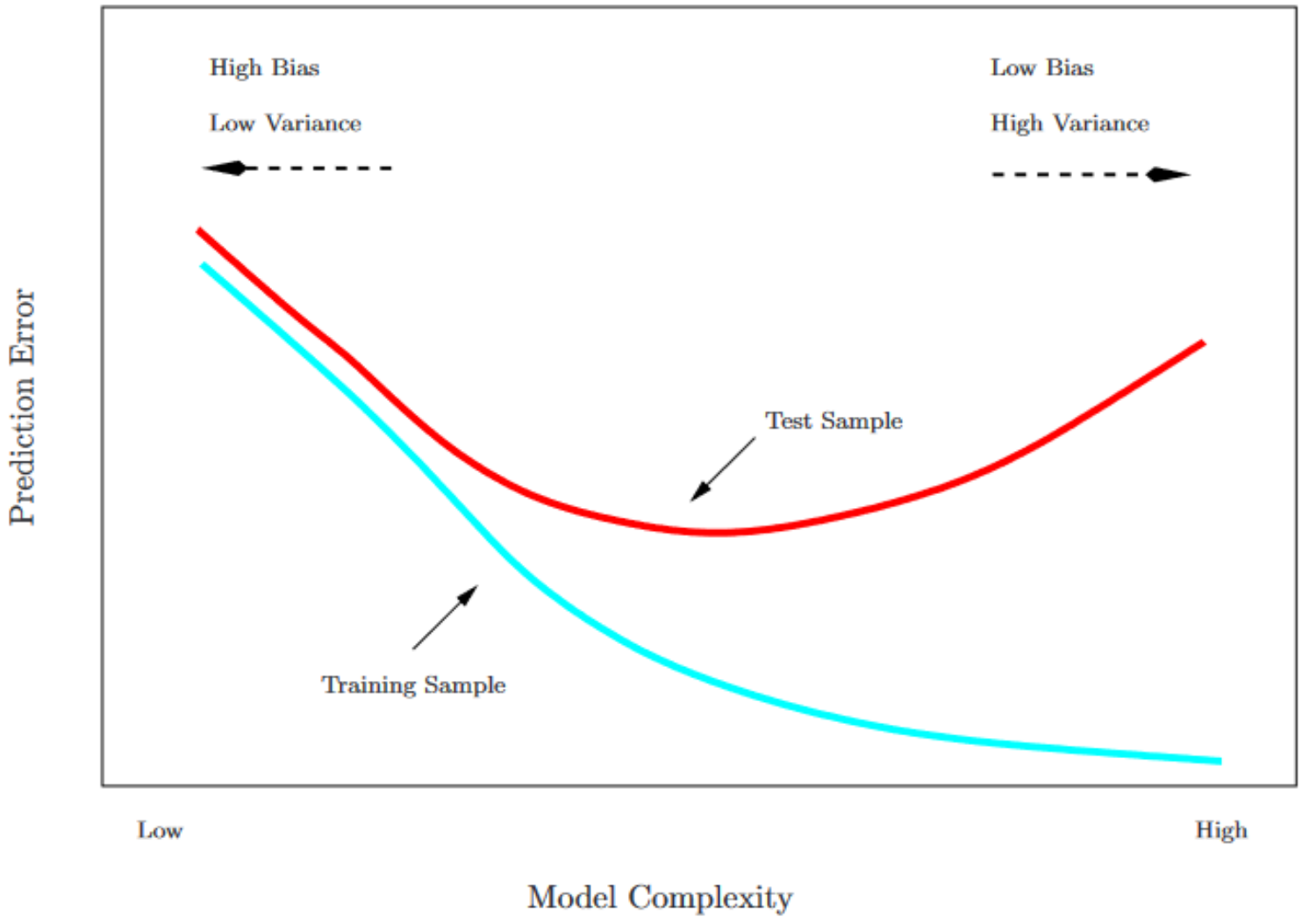
In order to minimize the expected prediction error, we need to select low variance and low bias, but not necessarily zero bias!

Motivation III: Prediction



The squared points ('■') indicate the true value and round points ('○') represent estimates. The diagrams illustrate that a high bias/low variance estimator may yield predictions that are on average closer to the truth than predictions from a low bias/high variance estimator.

Motivation III: Prediction



Source: Tibshirani/Hastie

Motivation III: Prediction

A full model with all predictors (**'kitchen sink approach'**) will have the lowest bias (OLS is unbiased) and R^2 (in-sample fit) is maximised. However, the kitchen sink model likely suffers from **overfitting**.

Removing some predictors from the model (i.e., forcing some coefficients to be zero) induces bias. On the other side, by removing predictors we also reduce model complexity and variance.

The optimal prediction model rarely includes all predictors and typically has a non-zero bias.

Important: High R^2 does not translate into good out-of-sample prediction performance.

How to find the best model for prediction? — This is one of the central questions of ML.

Demo: Predicting Boston house prices

For demonstration, we use house price data available on the StatLib archive.

Number of observations: 506 census tracts
Number of variables: 14

Dependent variable: median value of owner-occupied homes (`medv`)

Predictors: crime rate, environmental measures, age of housing stock, tax rates, social variables. (See Descriptions.)

Demo: Predicting Boston house prices

We divide the sample in half (253/253). Use first half for estimation, and second half for assessing prediction performance.

Estimation methods:

- 'Kitchen sink' OLS: include all regressors
- Stepwise OLS: begin with general model and drop if p-value > 0.05
- 'Rigorous' LASSO with theory-driven penalty
- LASSO with 10-fold cross-validation
- LASSO with penalty level selected by information criteria

Demo: Predicting Boston house prices

We divide the sample in half (253/253). Use first half for estimation, and second half for assessing prediction performance.

	OLS	Stepwise	`rlasso`	`cvlasso`	`lasso2` AIC/AICc	`lasso2` BIC/EBIC$_1$
`crim`	1.201*	1.062*		0.985	1.053	
`zn`	0.0245			0.0201	0.0214	
`indus`	0.01000					
`chas`	0.425			0.396	0.408	
`nox`	-8.443	-8.619*		-6.560	-7.067	
`rm`	8.878***	9.685***	8.681	8.925	8.909	9.086
`age`	-0.0485***	-0.0585***	-0.00608	-0.0470	-0.0475	-0.0335
`dis`	-1.120***	-0.956***		-1.025	-1.057	-0.463
`rad`	0.204			0.158	0.171	
`tax`	-0.0160***	-0.0121***	-0.00267	-0.0148	-0.0151	-0.00925
`ptratio`	-0.660***	-0.766***	-0.417	-0.660	-0.659	-0.659
`b`	0.0178***	0.0175***	0.000192	0.0169	0.0172	0.0110
`lstat`	-0.115*		-0.124	-0.113	-0.113	-0.109
Selected predictors	13	8	6	12	12	7
in-sample RMSE	3.160	3.211	3.656	3.164	3.162	3.279
out-of-sample RMSE	17.42	15.01	7.512	14.78	15.60	7.252

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Constant omitted.

Demo: Predicting Boston house prices

- OLS exhibits lowest in-sample RMSE, but worst out-of-sample prediction performance. Classical example of overfitting.
- Stepwise regression performs slightly better than OLS, but is known to have many problems: biased (over-sized) coefficients, inflated R^2, invalid p-values.
- In this example, AIC & AICc and BIC & EBIC_1 yield the same results, but AICc and EBIC are generally preferable for large-p-small-n problems.
- LASSO with 'rigorous' penalization and LASSO with BIC/EBIC_1 exhibit best out-of-sample prediction performance.

Motivation III: Prediction

There are cases where ML methods can be applied 'off-the-shelf' to policy questions.

Kleinberg et al. (2015) and Athey (2017) provide examples:

- Predict patient's life expectancy to decide whether hip replacement surgery is beneficial.
- Predict whether accused would show up for trial to decide who can be let out of prison while awaiting trial.
- Predict loan repayment probability.

But: in most cases, ML methods are not directly applicable for research questions in econometrics and allied fields, especially when it comes to causal inference.

Motivation III: Prediction

Another example: 'Improving refugee integration through data-driven algorithmic assignment'

Bansak, Ferwerda, Hainmueller, Dillon, Hangartner, Lawrence, and Weinstein, 2018, *Science*

- Refugee integration on settlement location, personal characteristics and synergies between the two.
- For example, the ability to speak French results is expected to lead to higher employment chances in French-speaking cantons of Switzerland.
- Host countries rarely take these synergies into account. Assignment procedures are usually based on capacity considerations (US) or random (Switzerland).

Motivation III: Prediction

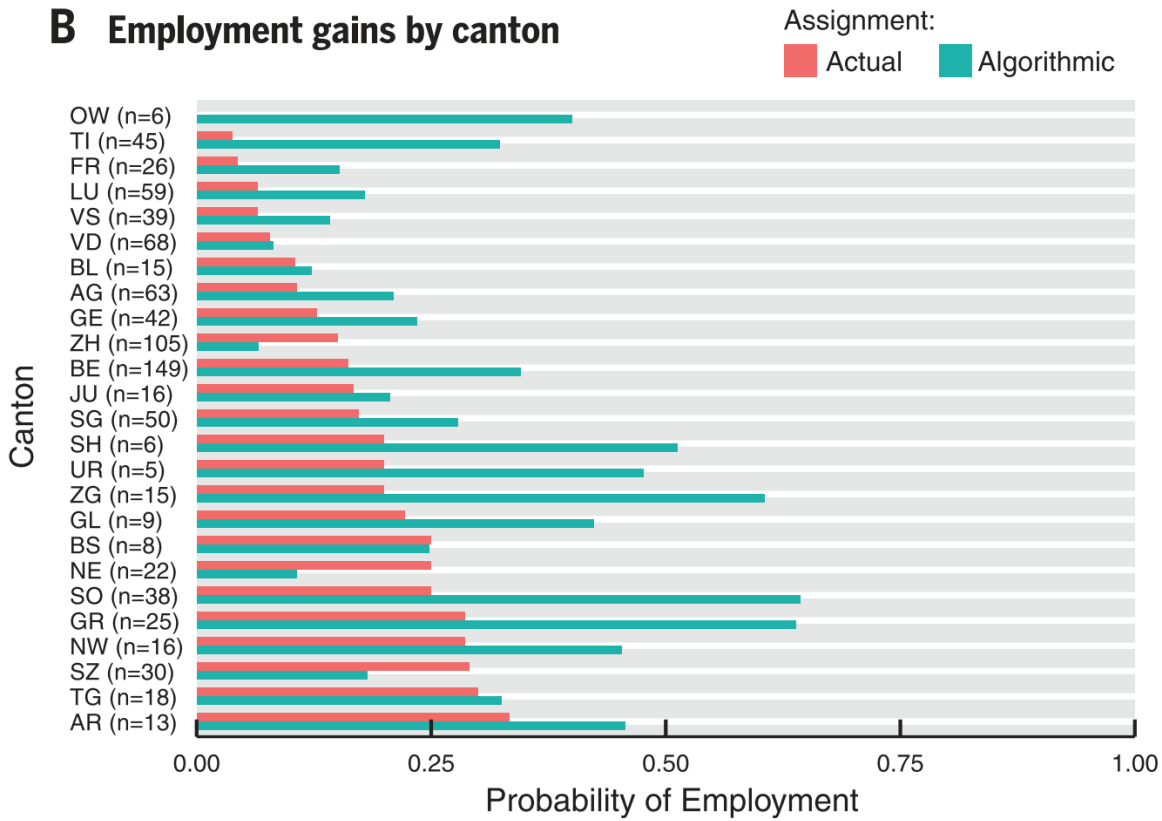
The proposed method proceeds in three steps:

1. *predict* the expected success, e.g. of finding a job using supervised ML
2. *mapping* from individuals to cases, i.e., family units
3. *matching*: assigning each case to a specific location (under constraints, e.g. proportionality)

Note that the first step is a prediction problem, that doesn't require us to make causal statements about the effect of X on Y. That's why ML is so suitable.

Motivation III: Prediction

The refugee allocation algorithm has the potential to lead to employment gains. Predicted vs actual employment shares for Swiss cantons:



Motivation IV: Causal inference

Machine learning offers a set of methods that outperform OLS in terms of *out-of-sample* prediction.

But: in most cases, ML methods are not directly applicable for research questions in econometrics and allied fields, especially when it comes to causal inference.

So how can we exploit the strengths of supervised ML (automatic model selection & prediction) for causal inference?

Motivation IV: Causal inference

Two very common problems in applied work:

- **Selecting controls** to address omitted variable bias when many potential controls are available
- **Selecting instruments** when many potential instruments are available.

Motivation IV: Causal inference

A motivating example is the *partial linear model*:

$$y_i = \underbrace{\alpha d_i}_{\text{aim}} + \underbrace{\beta_1 x_{i,1} + \ldots + \beta_p x_{i,p}}_{\text{nuisance}} + \varepsilon_i.$$

The causal variable of interest or "treatment" is d_i. The xs are the set of potential controls and not directly of interest. We want to obtain an estimate of the parameter α.

The problem is the controls. We want to include controls because we are worried about omitted variable bias – the usual reason for including controls.

But which ones do we use?

Motivation IV: Causal inference

A motivating example is the *partial linear model*:

$$y_i = \underbrace{\alpha d_i}_{\text{aim}} + \underbrace{\beta_1 x_{i,1} + \ldots + \beta_p x_{i,p}}_{\text{nuisance}} + \varepsilon_i.$$

The model corresponds to a setting we often encounter in applied research:

- there is set of regressors which we are primarily interested in and which we expect to be related to the outcome, but...
- we are unsure about which other confounding factors are relevant.

The setting is more general than it seems:

- The controls could include spatial or temporal effects.
- The above model could also be a panel model with fixed effects.
- We might only have a few observed elementary controls, but use a large set of transformed variables to capture non-linear effects.

Example: The role of institutions

Aim: Estimate the effect of institutions on output following Acemoglu et al. (2001, *AER*). Discussion here follows BCH (2014a).

Endogeneity problem: better institutions may lead to higher incomes, but higher incomes may also lead to the development of better institutions.

Identification strategy: use of mortality rates for early European settlers as an instrument for institution quality.

Underlying reasoning: Settlers set up better institutions in places where they are more likely to establish long-term settlements; and institutions are highly persistent.

low death rates	$\rightarrow$	colony attractive, build institutions
high death rates	$\rightarrow$	colony not attractive, exploit

Example: The role of institutions

Argument for instrument exogeneity: disease environment (malaria, yellow fever, etc.) is exogenous because diseases were almost always fatal to settlers (no immunity), but less serious for natives (some degree of immunity).

Major concern: Need to control for other highly persistent factors that are related to institutions & GDP.

In particular: geography. AJR use latitude in the baseline specification, and also continent dummy variables.

High-dimensionality: We only have 64 country observations. BCH (2014a) consider 16 control variables (12 variables for latitude and 4 continent dummies) for geography. So the problem is somewhat 'high-dimensional'.

Example: The role of institutions

This problem can now be solved in Stata.

We first ignore the endogeneoity of institutions and focus on the selection of controls:

```
. clear
. use https://statalasso.github.io/dta/AJR.dta

. pdslasso logpgp95 avexpr ///
        (lat_abst edes1975 avelf temp* humid* steplow-oilres), ///
        robust
```

Example: The role of institutions

OLS using CHS lasso-orthogonalized vars

logpgp95	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
avexpr	.4262511	.0540552	7.89	0.000	.3203049	.5321974

OLS using CHS post-lasso-orthogonalized vars

logpgp95	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
avexpr	.391257	.0574894	6.81	0.000	.2785799	.503934

OLS with PDS-selected variables and full regressor set

logpgp95	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
avexpr	.3913455	.0561862	6.97	0.000	.2812225	.5014684
edes1975	.0091289	.003184	2.87	0.004	.0028883	.0153694
avelf	-.9974943	.2474453	-4.03	0.000	-1.482478	-.5125104
zinc	-.0079226	.0280604	-0.28	0.778	-.0629201	.0470748
_cons	5.764133	.3773706	15.27	0.000	5.024501	6.503766

Standard errors and test statistics valid for the following variables only:
avexpr

Example: The role of institutions

We can do valid inference with the variable of interest (here `avexpr`) and obtain estimates that are robust to misspecification issues (omitting confounders or including the wrong controls).

The same result can be achieved using Stata 16's new `dsregress`.

Example: The role of institutions

The model:

$$\begin{aligned}
\log(\text{GDP per capita})_i &= \alpha \cdot \text{Expropriation}_i + \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i \\
\text{Expropriation}_i &= \pi_1 \cdot \text{Settler Mortality}_i + \boldsymbol{x}_i'\boldsymbol{\pi}_2 + \nu_i \\
\text{Settler Mortality}_i &= \boldsymbol{x}_i'\boldsymbol{\gamma} + u_i
\end{aligned}$$

In summary, we have one endogenous regressor of interest, one instrument, but 'many' controls.

The method:

1. Use the LASSO to regress log(GDP per capita) against controls,
2. use the LASSO to regress Expropriation against controls,
3. use the LASSO to regress Settler Mortality against controls.
4. Estimate model with union of controls selected by Step 1-3.

Example: The role of institutions

LASSO selects Africa dummy (in Step 1 and 3).

Specification	*Controls*	$\hat{\alpha}$ (SE)	*First-stage F*
IV AJR	Latitude	0.97 (0.19)	15.9
IV DS LASSO	Africa	0.77 (0.18)	11.8
'Kitchen Sink' IV	All 16	0.99 (0.61)	1.2

Double-selection LASSO results somewhat weaker (smaller coefficients, first stage F-statistics smaller), but AJR results basically sustained.

Double-selection LASSO performs much better than the 'kitchen sink' approach (using all controls), where the model is essentially unidentified as indicated by first stage F-statistic.

Motivation IV: Causal inference

This is an active and exciting area of research in econometrics. Probably the most exciting area (in my biased view).

Research is lead by (among others):

- Susan Athey (Standford)
- Guido Imbens (Standford)
- Victor Chernozhukov (MIT)
- Christian Hansen (Chicago)

Susan Athey:

> *'Regularization/data-driven model selection will be the standard for economic models'* (AEA seminar)

Hal Varian (Google Chief Economist & Berkeley):

> *'my standard advice to graduate students [in economics] these days is to go to the computer science department and take a class in machine learning.'* (Varian, 2014)

Some key concepts

Bias-variance-tradeoff: Model complexity (e.g., more regressors) implies less bias, but higher variance.

Validation: The model is assessed using unseen data and some loss function (e.g. mean-squared error). *Cross-validation* is a generalisation where we the data is iteratively split in training and validation sample.

Sparse vs. dense problems: Theoretical and practical considerations depend on whether we assume the underlying true data-generating process to be sparse (few relevant predictors) or dense (many predictors).

Tuning parameters: Again and again, we will see tuning parameters. These allow to reduce complex model selection problems into one (or multi)-dimensional problems, where we only need to select the tuning parameter.

New ML features in Stata (incomplete list)

- Lasso and elastic net in **lassopack** & **pdslasso** as well as Stata 16's `lasso`; including lasso for causal inference!
- `randomforest` by Zou/Schonlau (on SSC).
- `svmachines` by Guenter/Schonlau (on SSC) for support vector machines.

A big novelty of Stata 16 is the Python integration which allows to make use of the extensive ML packages of Python (Scikit-learn).

Similarly, we can call R using Haghish's `rcall` (available on github).

New ML features in Stata: Python integration

Random forest in Stata with a few lines (using Boston house price data set).

```
ds crim-lstat
local xvars = r(varlist)

python:

from sfi import Data
import numpy as np
from sklearn.ensemble import RandomForestRegressor

X = np.array(Data.get("‘xvars’"))
y = np.array(Data.get("medv"))

rf = RandomForestRegressor(n_estimators = 1000, random_state = 42)
rf.fit(X,y)
xbhat = rf.predict(X)

Data.addVarFloat(’xbhat’)
Data.store(’xbhat’, None, xbhat)

end
```

Summary I

Machine learning/Penalized regression

- ML provides wide set of flexible methods focused on prediction and classification problems.
- ML outperforms OLS in terms of prediction due to *bias-variance-tradeoff*.

Causal inference in the partial linear model

- Distinction between *parameters of interest* and *high-dimensional set of controls/instruments*.
- General framework allows for causal inference with low-dimensional parameters robust to misspecification; and avoids problems associated with model selection using significance testing.
- But there's a price: the framework is designed for inference on low-dim parameters only.

Summary II

Machine learning/Penalized regression

- Stata has now extensive and powerful features for prediction and causal inference with lasso & friends.
- Other ML methods are less well developed, e.g., random forest.
- But: the ability to call R (via `rcall`) and Python (in Stata 16) makes it relatively easy to access R/Python's ML programs. User-friendly wrapper programs are likely to be developed.

Reference for the lasso:

Ahrens, A., Hansen, C. B., & Schaffer, M. E. (2019). lassopack: Model selection and prediction with regularized regression in Stata. Retrieved from `http://arxiv.org/abs/1901.05397`