Analysis of Cross-Sectional Data

Kevin Sheppard
https://kevinsheppard.com/teaching/mfe/

Modules

Overview

- Introduction to Regression Models
- Parameter Estimation and Model Fit
- Properties of OLS Estimators
- Hypothesis Testing
- Hypothesis Testing in Regression Models
- Wald and t-Tests
- Lagrange Multiplier and Likelihood Ratio Tests
- Heteroskedasticity
- Specification Failures
- Model Selection
- Checking for Specification Errors
- Machine Learning Approaches to Regression

Course Structure

- Course presented through two overlapping channel:
  1. In-person lectures
  2. Notes that accompany the lecture content
     - Read before or after the lecture or when necessary for additional background
- Slides are primary – material presented during lecturers is examinable
- Notes are secondary and provide more background for the slides
- Slides are derived from notes so there is a strong correspondence

Monitoring Your Progress

- Self assessment
  - Review questions in printer-friendly version of slides
    - Self-assessment
  - Multiple choice questions on Canvas made available each week
    - Answers available immediately
  - Long-form problem distributed each week
    - Answers presented in a subsequent class
- Marked Assessment
  - Empirical projects applying the material in the lectures
  - Each empirical assignment will have a written and code component

Introduction to Regression Models

Basic Notation

$$Y_i = \beta_1 X_{1,i} + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i} + \epsilon_i,$$

- Y_i: Regressand, Dependent Variable, LHS Variable
- $X_{j,i}$: Regressor, also Independent Variable, RHS Variable, Explanatory Variable
- ϵ_i: Innovation, also Shock, Error or Disturbance
- n observations, indexed $i = 1, 2, \ldots, n$
- k regressors, indexed $j = 1, 2, \ldots, k$

Usually use matrix notation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- $\mathbf{y}$: $n \times 1$
- $\mathbf{X}$: $n \times k$
- $\boldsymbol{\beta}$: $k \times 1$
- $\boldsymbol{\epsilon}$: $n \times 1$

More Notation

Row form:

$$Y_i = \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i$$

Column form:

$$\mathbf{y} = \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \ldots + \beta_k \mathbf{x}_k + \boldsymbol{\epsilon}_i$$

Throughout the notes and slides:

- Standard $math\ notation$ indicates a scalar: $y_i, x_i, \beta, \epsilon_i$
- Scalar random variables are upper case: Y_i, X_i, Z_i
- Lower case **bold math** indicates a vector: $\mathbf{y}, \mathbf{x}_i, \boldsymbol{\epsilon}, \boldsymbol{\beta}$
- Upper case **bold math** indicates a matrix: $\mathbf{X}, \mathbf{A}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}$

Introduction

What is a linear regression?

Many specifications can be examined using the tools of linear regression

$$Y_i = \beta X_i + \epsilon_i$$

- Two key requirements
  - Additive error
  - One multiplicative parameter per term

Examples:

- Polynomials $$Y_i = \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$$
- Level shifts $$Y_i = \beta_1 X_i + \beta_2 X_i I_{[X_i > \kappa]} + \epsilon_i$$
  - $I_{[X_i > \kappa]}$ is an indicator variable that takes the value 1 or 0
  - $I_{[X_i > \kappa]} = 1$ if $X_i > \kappa$
- "Non-linear" relationships

$$Y_i = \beta_1 \sin X_i + \beta_2 \ln X_i + \epsilon_i$$

What *cannot* be analyzed as a linear regression?

- Non-separable parameters

$$Y_i = \beta_1 X_i^{\beta_2} + \epsilon_i$$

  - Lots of solutions: Non-linear least squares, Maximum Likelihood, GMM

- ARCH

$$Y_t = \sqrt{\sigma_t^2}\epsilon_t$$

$$\sigma_t^2 = \omega + \alpha Y_{t-1}^2$$

- Some models can be transformed into a LR

$$Y_i = \beta_1 X_i^{\beta_2}\epsilon_i \Rightarrow \ln Y_i = \ln \beta_1 + \beta_2 \ln X_i + \ln \epsilon_i$$

$$\tilde{Y}_i = \tilde{\beta}_1 + \beta_2 \tilde{X}_i + \tilde{\epsilon}_i$$

  - Requires non-negativity of Y_i and X_i

Regression Coefficient Interpretation

- *Ceteris Paribus*
  - Not usually applicable
- Holding other (included) variables constant
  - More reasonable

On average

$$Y_i = \beta_1 X_{1,i} + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i}$$

$$\beta_k \approx \frac{\partial Y_i}{\partial X_{k,i}}$$

More complicated when model nonlinear in X_i

$$\ln Y_i = \beta \ln X_i$$

$$\beta \approx \frac{\partial Y_i}{\partial X_i} \frac{X_i}{Y_i} = E_{y,x}$$

$$Y_i = \beta_1 X_i + \beta_2 X_i^2$$

$$\beta_1 + 2\beta_2 X_i \approx \frac{\partial Y_i}{\partial X_i}$$

What is a model?

An important but challenging question

- Two competing views
- Data generating process (DGP)
  - Model taken as literal
  - Simpler to think about
  - Implausible for nearly everything we do
- Approximation to probability law (a.k.a. distribution)
  - All models are misspecified, but...
  - Even misspecified models can aid in understanding important relationships
  - Reduces reality to tractable problem
  - Some caution is needed
- My favorite example: GARCH model

$$Y_t = \sqrt{\sigma_t^2}\epsilon_t$$

$$\sigma_t^2 = \omega + \alpha Y_{t-1}^2 + \beta\sigma_{t-1}^2$$

  - Relates today's variance to yesterday's variance and the squared return

Example: Approximate Factor Models

- Factor models are widely used in finance
  - Capital Asset Pricing Model (CAPM)
  - Arbitrage Pricing (APT)
  - Risk Exposure
- Basic specification

$$R_i = \mathbf{f}_i \boldsymbol{\beta} + \epsilon_i$$

  - R_i: Return on dependent asset, often *excess* (R_i^e)
  - $\mathbf{f}_i$: $1 \times k$ vector of factor innovations
  - ϵ_i innovation, corr($\epsilon_i, F_{j,i}$)=0, $j = 1, 2, \ldots, k$
- Special Case: CAPM

$$R_i - R_i^f = \beta(R_i^m - R_i^f) + \epsilon_i$$

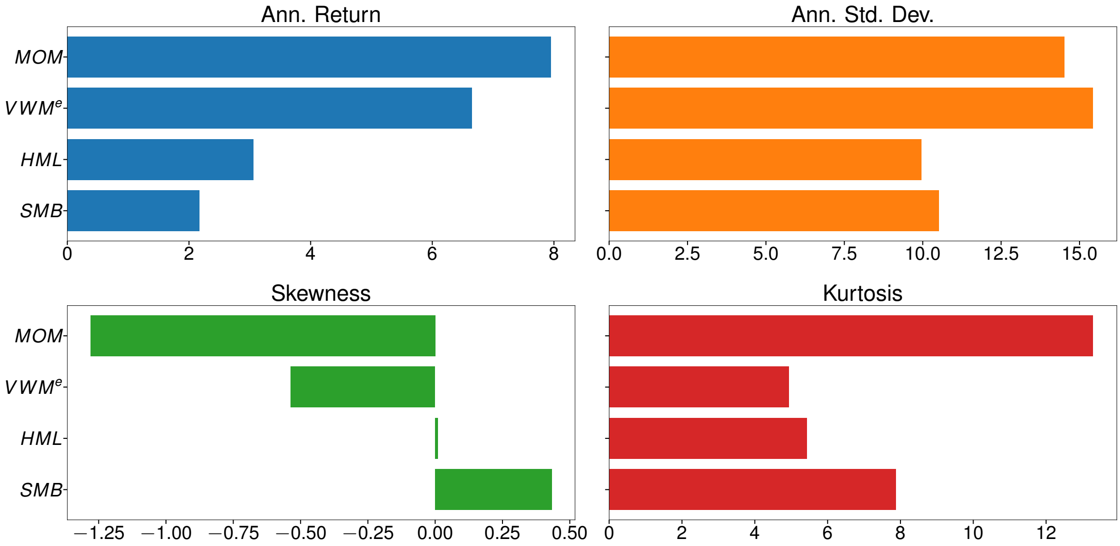
$$R_i^e = \beta R_i^{me} + \epsilon_i$$

Empirical Illustration

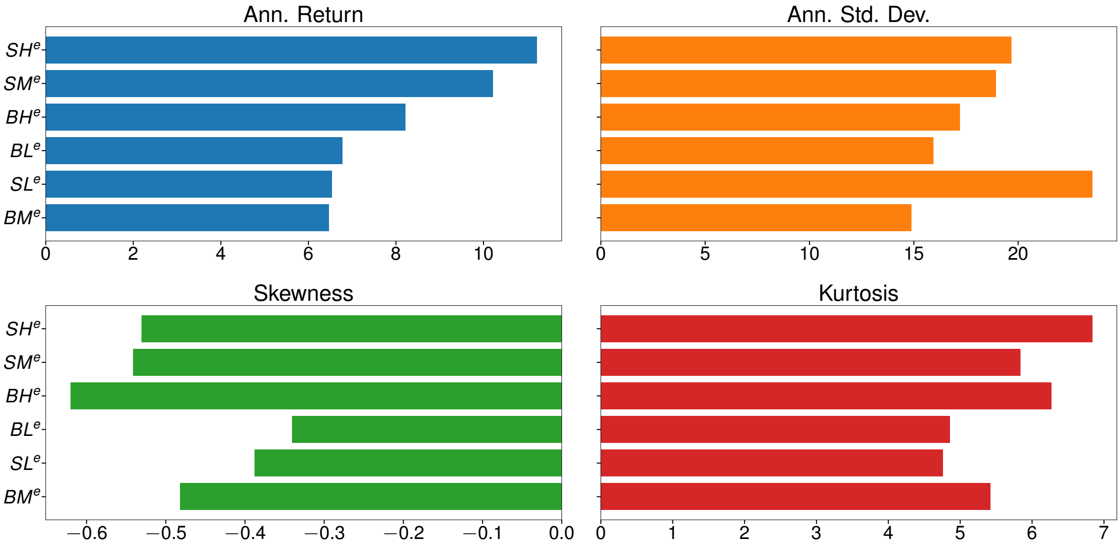
Data

- Fama-French 3 factors + Momentum
  - VWM^e - Excess return on Value-Weighted-Market
  - SMB - Return on the size portfolio
  - HML - Return on the value portfolio
  - MOM - Return on the momentum portfolio
- Size-Value sorted portfolio return data
  - Size
    - S: Small
    - B: Big
  - Value
    - H: High BE/ME
    - M: Middle BE/ME
    - L: Low BE/ME
- 49 Industry Portfolios
- All returns excess except SMB, HML, MOM

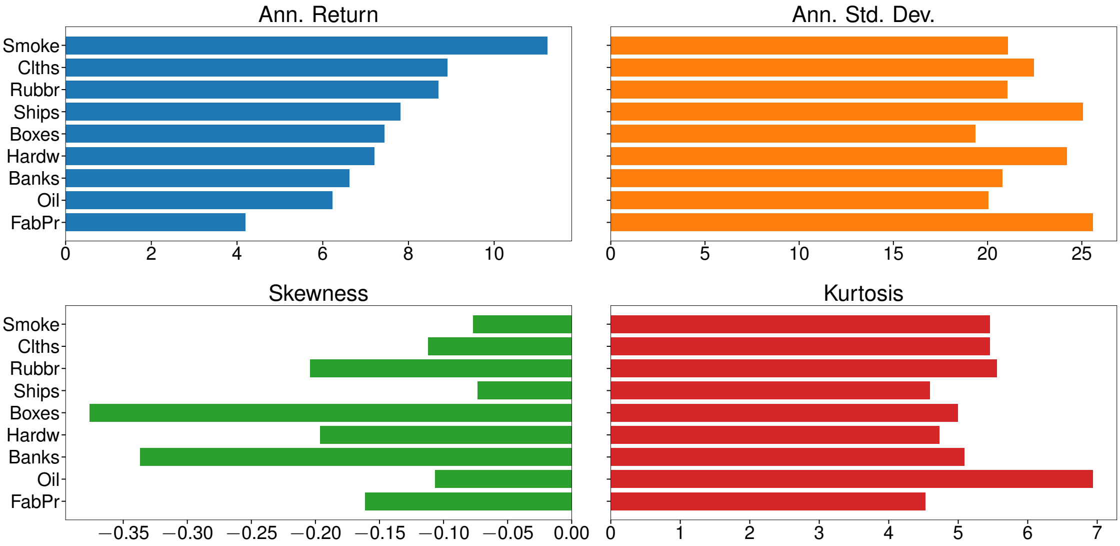
Factor Portfolio Summary Statistics



Component Portfolio Summary Statistics



Industry Portfolio Summary Statistics



Common Variable Transformations

Dummy Variables

Definition (Dummy Variable)

A dummy variable is a variable that takes the value 0 or 1.

- Value depends on the value of some X variable(s)
- Denoted

$$I_{[f(\mathbf{x})\leq c]}$$

  - $f(\mathbf{x})$ is some function of the regressors
  - c is an arbitrary constant
  - $\leq$ could be anything that would produce a *logical* expression ($\neq$, $>$)
  - *Cannot* depend on y_i
- Dummies in finance
  - Asymmetries: $I_{[X_i<0]}$
  - Calendar effects: $I_{[X_i=1]}$ whereX_i is the month or day of the week
  - Structural breaks: $I_{[X_i>1987]}$ where X_i is the year

Variable Interactions

- Nonlinearities often introduced through interactions

$$X_{1,i}^2 \text{ , } X_{1,i}X_{2,i} \text{ or } X_{1,i}^2 X_{2,i}$$

- Interactions can include dummy variables

$$X_{1,i}I_{[X_{1,i}<0]} - \text{Asymmetric slope coefficient}$$

$$X_{1,i}X_{2,i}I_{[X_{1,i}<0]}I_{[X_{2,i}<0]} - \text{Asymmetric slope coefficient in (-,-) quadrant}$$

- Interactions, particularly dummy interactions can capture important highly-linear features

Kinked lines	$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i I_{[X_i<0]} + \epsilon_i$
Jumps in lines	$Y_i = \beta_1 + \beta_2 I_{[X_i<0]} + \beta_3 X_i + \epsilon_i$
Piece-wise linear splines	$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i I_{[X_i>c]} + (\beta_1 + \beta_2 c - \beta_3 c) I_{[X_i>c]} + \epsilon_i$
Polynomial (Tensor) Products	$Y_i = \beta_1 + \beta_2 X_{1,i} + \beta_3 X_{2,i} + \beta_4 X_{1,i}^2 + \beta_5 X_{2,i}^2 + \beta_6 X_{1,i} X_{2,i} + \epsilon_i$

Interaction Variables

- Market Negative: $I_{[VWM^e<0]}$
- Negative Return: $VWM^e \times I_{[VWM^e<0]}$
- Squared Return: $(VWM^e)^2$

Date	Market Negative	Negative Return	Squared Returns
2019-11	0	0	14.98
2019-12	0	0	7.67
2020-01	1	-0.11	0.01
2020-02	1	-8.13	66.10
2020-03	1	-13.38	179.02
2020-04	0	0	186.32
2020-05	0	0	31.14
2020-06	0	0	6.05
2020-07	0	0	33.29
2020-08	0	0	58.06

Monthly Dummy Variables

Date	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1963-07	0	0	0	0	0	0	1	0	0	0	0	0
1963-08	0	0	0	0	0	0	0	1	0	0	0	0
1963-09	0	0	0	0	0	0	0	0	1	0	0	0
1963-10	0	0	0	0	0	0	0	0	0	1	0	0
1963-11	0	0	0	0	0	0	0	0	0	0	1	0
1963-12	0	0	0	0	0	0	0	0	0	0	0	1
1964-01	1	0	0	0	0	0	0	0	0	0	0	0
1964-02	0	1	0	0	0	0	0	0	0	0	0	0

A Caveat for Using Dummy Variables

The Dummy Variable Trap

- Cannot include an intercept and all dummies
  - $I_{1,i} = 1$ if Monday, $I_{2,i} = 1$ if Tuesday, etc.
  - Problematic specification:

$$Y_i = \beta_1 + \beta_2 I_{1,i} + \beta_3 I_{2,i} + \beta_4 I_{3,i} + \beta_5 I_{4,i} + \beta_6 I_{5,i} + \epsilon_i$$

  - $\sum_{j=1}^{5} I_{j,i} = 1$ always
  - Perfect Collinearity: Cannot estimate model
- Solution 1: Remove the constant

$$Y_i = \beta_1 I_{1,i} + \beta_2 I_{2,i} + \beta_3 I_{3,i} + \beta_4 I_{4,i} + \beta_5 I_{5,i} + \epsilon_i$$

- Solution 2: Remove one dummy

$$Y_i = \beta_1 + \beta_2 I_{2,i} + \beta_3 I_{3,i} + \beta_4 I_{4,i} + \beta_5 I_{5,i} + \epsilon_i$$

- Interpretation changes, models identical
- Most software will produce an error or warning

Parameter Estimation and Model Fit

Estimating the unknown parameters

- Many possible ways to estimate β
  - Take k data points and solve (Gaussian Elimination)
    - Exact and simple solution
    - Doesn't work if $n > k$
  - Minimize the maximum error
    - Maximum Score
    - Computationally challenging
  - Minimize the average error
    - Many solutions
  - Minimize some non-negative function of the errors
    - Least squares

$$\underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \mathbf{x}_i \boldsymbol{\beta})^2$$

    - Least absolute deviations

$$\underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} |Y_i - \mathbf{x}_i \boldsymbol{\beta}|$$

Calculus of Least Squares

- Formal problem

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \mathbf{x}_i \boldsymbol{\beta})^2 = \sum_{i=1}^{n} \epsilon_i^2$$

- Matrix equivalent

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\epsilon}'\boldsymbol{\epsilon}$$

- k First Order Conditions (F.O.C)

$$-2\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = -2\mathbf{X}'\hat{\boldsymbol{\epsilon}} = \mathbf{0}$$

$$\Rightarrow -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0}$$

- Solve for $\boldsymbol{\beta}$ to get LS estimator, denoted $\hat{\boldsymbol{\beta}}$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

- Second derivative is always positive definite as long as $\operatorname{rank}(\mathbf{X}) = k$.

$$2\mathbf{X}'\mathbf{X}$$

Standard Regression Summary Information

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right)' \left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - \mathbf{x}_i\boldsymbol{\beta}\right)^2$$

Benchmark Model

$$R_i^e = \beta_c + \beta_{VWM} VWM_i^e + \beta_{SMB} SMB_i + \beta_{HML} HML_i + \beta_{MOM} MOM_i + \epsilon_i$$

	$\hat{\beta}$	s.e.$(\hat{\beta})$	Z	$\Pr(Z)$	Conf. Int.
Const.	-0.0859	0.043	-1.991	0.046	[-0.170, -0.001]
VWM^e	1.0798	0.012	93.514	0.000	[1.057, 1.102]
SMB	0.0019	0.017	0.110	0.912	[-0.032, 0.036]
HML	0.7643	0.021	36.380	0.000	[0.723, 0.805]
MOM	-0.0354	0.013	-2.631	0.009	[-0.062, -0.009]

Least Absolute Deviations (LAD)

$$\hat{\boldsymbol{\beta}}_{LAD} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} |Y_i - \mathbf{x}_i \boldsymbol{\beta}|$$

	$\hat{\beta}$	s.e.($\hat{\beta}$)	T	$\Pr(T)$	Conf. Int.
Const.	-0.0306	0.044	-0.696	0.487	[-0.117, 0.056]
VWM^e	1.0716	0.010	103.257	0.000	[1.051, 1.092]
SMB	0.0161	0.015	1.090	0.276	[-0.013, 0.045]
HML	0.7503	0.016	47.702	0.000	[0.719, 0.781]
MOM	-0.0272	0.011	-2.581	0.010	[-0.048, -0.007]

Other estimators

- Fit values

$$\hat{Y}_i = \mathbf{x}_i \hat{\boldsymbol{\beta}}$$

- Estimated errors

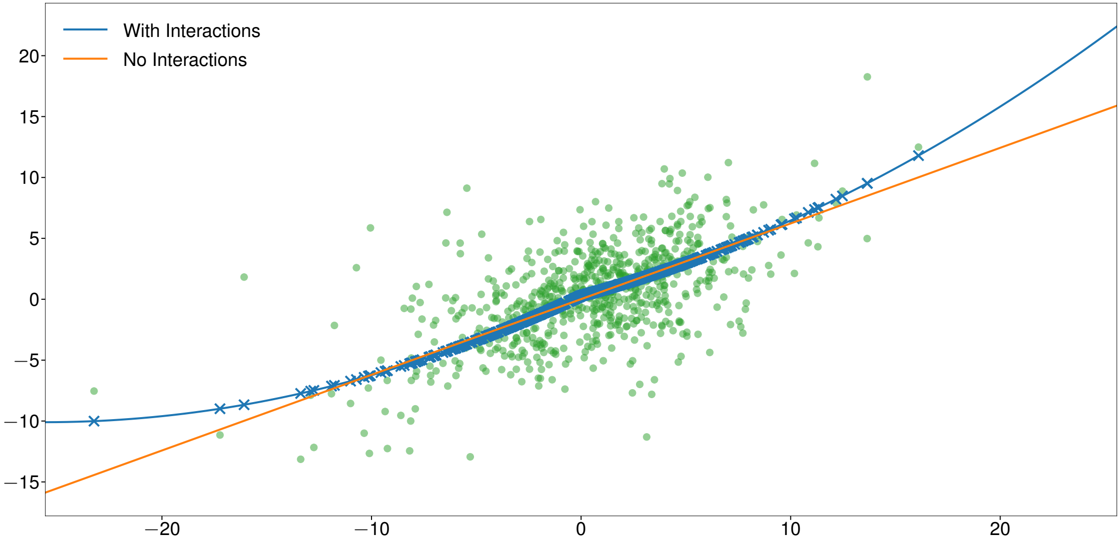
$$\hat{\epsilon}_i = Y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}} = Y_i - \hat{Y}_i$$

- Error variance estimator

$$s^2 = \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{n-k} \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{n}$$

  - $n-k$ is a degree of freedom correction
  - $\hat{\epsilon}_i$ are too close to zero.

Linear vs. Nonlinear Fit



Features of the OLS estimator

- Only assumption needed for estimation

$$\text{rank}(\mathbf{X}) = k \Rightarrow \mathbf{X}'\mathbf{X} \text{ is invertible}$$

- Estimated errors are orthogonal to $\mathbf{X}$

$$\mathbf{X}'\hat{\boldsymbol{\epsilon}} = \mathbf{0} \text{ or for each variables, } \sum_{i=1}^{n} X_{ij}\hat{\epsilon}_i = 0,\ j = 1, 2, \ldots, k$$

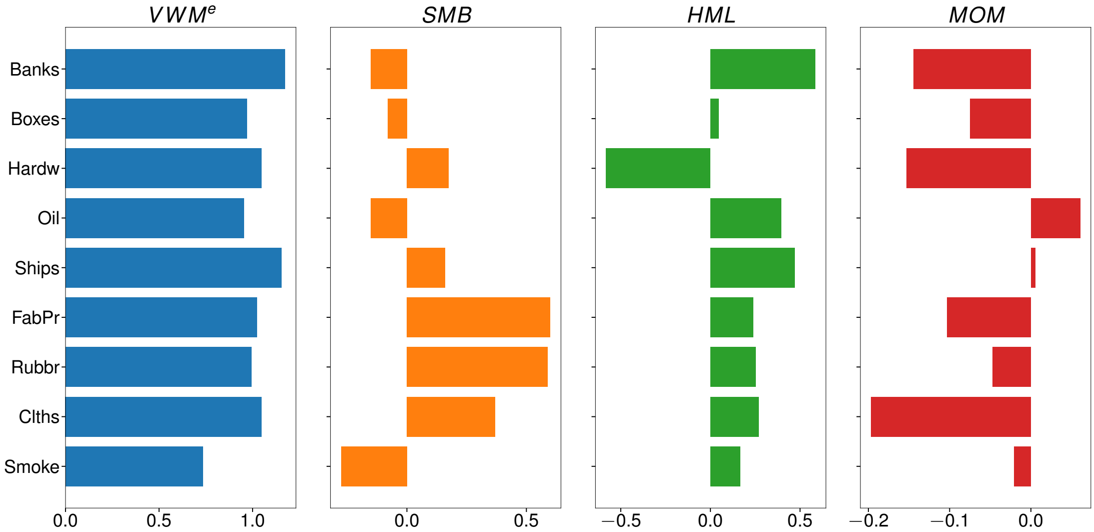
- If model includes a constant, estimated errors have mean 0

$$\sum_{i=1}^{n} \hat{\epsilon}_i = 0$$

- Closed under linear transformations to either $\mathbf{X}$ or $\mathbf{y}$
  Linear: $a\mathbf{z}$, a nonzero
- Closed under affine transformation to $\mathbf{X}$ or $\mathbf{y}$ if model has constant
  Affine: $a\mathbf{z} + c$, a nonzero

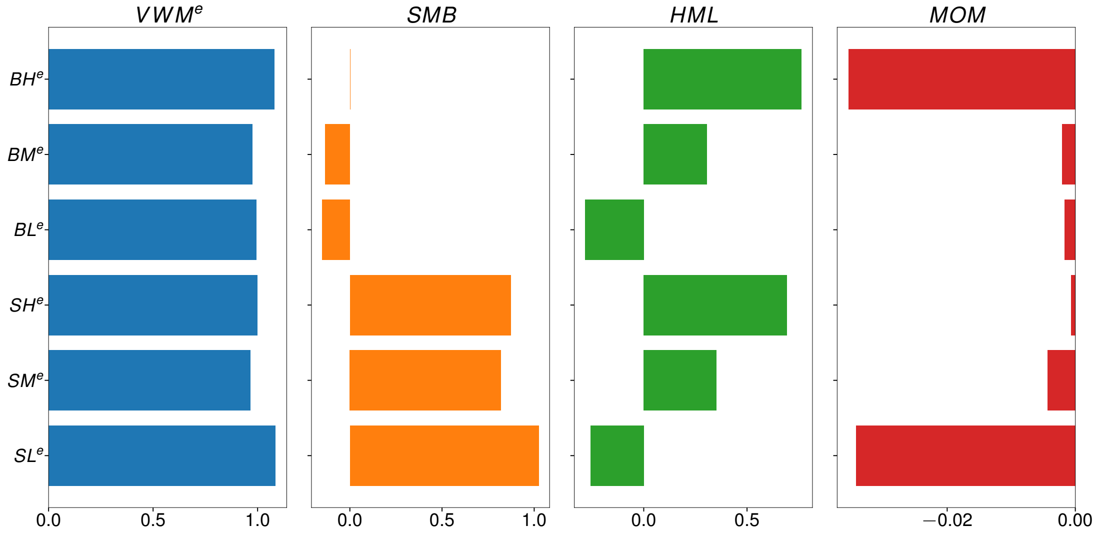
Factor Loadings

Industry Portfolios

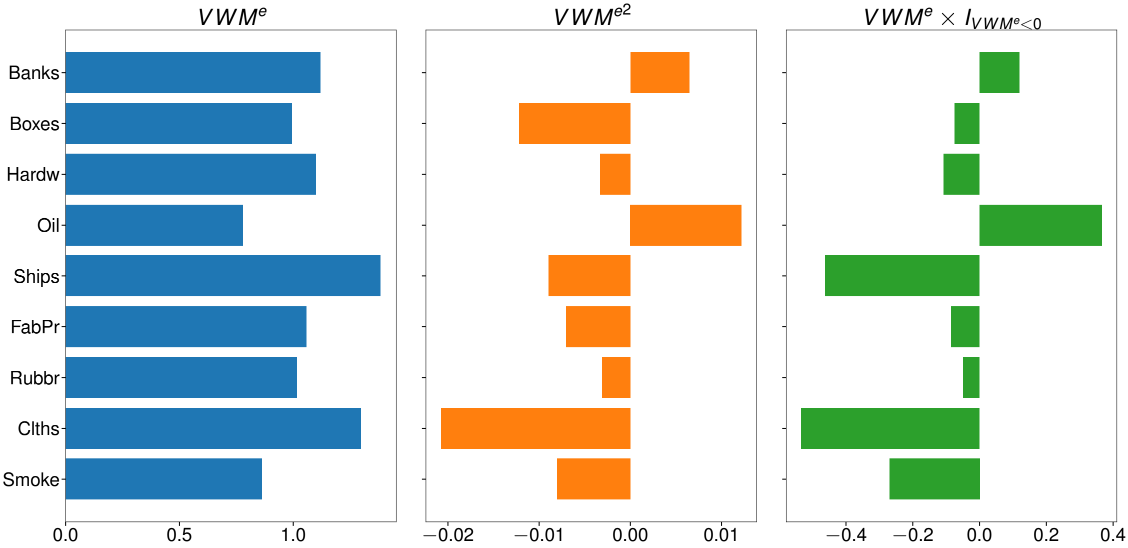


Factor Loadings

Component Portfolios



Evidence of Nonlinearity



Measuring Model Fit

Assessing fit

Next step: Does my model fit?

- A few preliminaries

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \text{Total Sum of Squares (\textbf{TSS})}$$

$$\sum_{i=1}^{n}(\mathbf{x}_i\hat{\boldsymbol{\beta}} - \bar{\mathbf{x}}\hat{\boldsymbol{\beta}})^2 \text{Regression Sum of Squares (\textbf{RSS})}$$

$$\sum_{i=1}^{n}(Y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}})^2 \text{ Sum of Squared Errors (\textbf{SSE})}$$

  - ι is a $k \times 1$ vector of 1s.

- Note: $\bar{y} = \bar{\mathbf{x}}\hat{\boldsymbol{\beta}}$ if the model contains a constant

$$TSS = RSS + SSE$$

- Can form ratios of explained and unexplained

$$R^2 = \frac{RSS}{TSS} = 1 - \frac{SSE}{TSS}$$

Uncentered R^2: R_u^2

- Usual R^2 is formally known as *centered* R^2 (R_c^2)
  - Only appropriate if model contains a constant
- Alternative definition for models without constant

$$\sum_{i=1}^{n} Y_i^2 \text{Uncentered Total Sum of Squares } (\mathrm{TSS_U})$$

$$\sum_{i=1}^{n} (\mathbf{x}_i\hat{\boldsymbol{\beta}})^2 \text{Uncentered Regression Sum of Squares } (\mathrm{RSS_U})$$

$$\sum_{i=1}^{n} (Y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}})^2 \text{Uncentered Sum of Squares Errors } (\mathrm{SSE_U})$$

- Uncentered R^2: R_u^2
- Warning: Most software packages return R_c^2 for any model
  - Inference based on R_c^2 when the model does not contain a constant will be wrong!
- Warning: Using the wrong definition can produce nonsensical and/or misleading numbers

The limitation of R^2

- R^2 has one crucial shortcoming:
  - Adding variables cannot decrease the R^2
  - Limits usefulness for selecting models : Bigger model always preferred
- Enter $\overline{R}^2$

$$\overline{R}^2 = 1 - \frac{\frac{SSE}{n-k}}{\frac{TSS}{n-1}} = 1 - \frac{s^2}{s_y^2} = 1 - \frac{SSE}{TSS}\frac{n-1}{n-k} = 1 - (1 - R^2)\frac{n-1}{n-k}$$

- $\overline{R}^2$ is read as "Adjusted R^2"
- $\overline{R}^2$ increases if and only if the estimated error variance decreases
- Adding noise variables should generally decrease $\bar{R}^2$
- Caveat: For large n, penalty is essentially nonexistent
- Much better way to do model selection coming later...

Model Fit

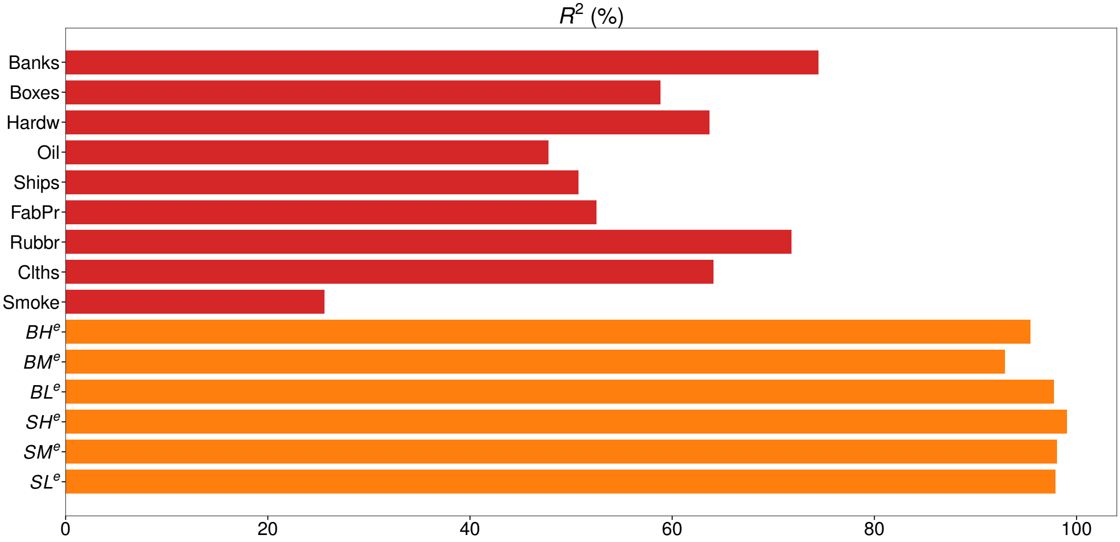
- Reference

$$R_i^e = \alpha + \beta_{VWM} VWM^e + \beta_{SMB} SMB + \beta_{HML} HML + \beta_{MOM} MOM + \epsilon_i$$

- Add Shift: $R_i^e \to R_i^e + 99$
- Rescaled: $R_i^e \to \pi R_i^e$
- Different LHS: $R_i^e \to R_i^e - VWM^e - HML$
- No Constant: $R_i^e \to R_i^e + 99$, α excluded from model

	Reference	Add. Shift	Rescaled	Diff. LHS	No. Const.
R^2	95.4%	95.4%	95.4%	38.2%	8.8%
$\bar{R}^2$	95.4%	95.4%	95.4%	37.8%	8.3%

Model Fit



R^2 (%)
Banks
Boxes
Hardw
Oil
Ships
FabPr
Rubbr
Clths
Smoke
BH^e
BM^e
BL^e
SH^e
SM^e
SL^e
0
20
40
60
80
100

Properties of the OLS Estimator

Making sense of estimators

- Only one assumption in 30 slides
  - $\mathbf{X}'\mathbf{X}$ is nonsingular (Identification)
  - More needed to make any statements about unknown parameters
- Two standard setups:
  - Classical (also Small Sample, Finite Sample, Exact)
    - Make strong assumptions $\Rightarrow$ get clear results
    - Easier to work with
    - Implausible for most finance data
  - Asymptotic (also Large Sample)
    - Make weak assumptions $\Rightarrow$ hope distribution close
    - Requires limits and convergence notions
    - Plausible for many financial problems
    - Extensions to make applicable to most finance problem
- We'll cover only the Asymptotic framework since the Classical framework is not appropriate for most financial data.

Assumptions

The assumptions

Assumption (Linearity)

$Y_i = \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i$

- Model is correct and conformable to requirements of linear regression
- Strong (kind of)

Assumption (Stationary Ergodicity)

$\{(\mathbf{x}_i, \epsilon_i)\}$ *is a strictly stationary and ergodic sequence.*

- Distribution of $(\mathbf{x}_i, \epsilon_i)$ does not change across observations
- Allows for applications to time-series data
- Allows for i.i.d. data as a special case

The assumptions

Assumption (Rank)

$\mathrm{E}[\mathbf{x}_i'\mathbf{x}_i] = \mathbf{\Sigma}_{\mathbf{XX}}$ *is nonsingular and finite.*

- Needed to ensure estimator is well defined in large samples
- Rules out some types of regressors
  - Functions of time
  - Unit roots (random walks)

Assumption (Moment Existence)

$\mathrm{E}[X_{j,i}^4] < \infty$, $i = 1, 2, \ldots$, $j = 1, 2, \ldots, k$ *and* $\mathrm{E}[\epsilon_i^2] = \sigma^2 < \infty$, $i = 1, 2, \ldots$.

- Needed to estimate parameter covariances
- Rules out very heavy-tailed data

The assumptions

Assumption (Martingale Difference)

$\{\mathbf{x}_i'\epsilon_i, \mathcal{F}_i\}$ *is a martingale difference sequence,* $\mathrm{E}\left[(X_{j,i}\epsilon_i)^2\right] < \infty$ $j = 1, 2, \ldots, k$, $i = 1, 2\ldots$ *and* $\mathbf{S} = \mathrm{V}[n^{-\frac{1}{2}}\mathbf{X}'\boldsymbol{\epsilon}]$ *is finite and nonsingular.*

- Provides conditions for a central limit theorem to hold

Definition (Martingale Difference Sequence)

Let $\{\mathbf{z}_i\}$ be a vector stochastic process and $\mathcal{F}_i$ be the information set corresponding to observation i containing all information available when observation i was collected except $\mathbf{z}_i$. $\{\mathbf{z}_i, \mathcal{F}_i\}$ is a martingale difference sequence if

$$\mathrm{E}[\mathbf{z}_i|\mathcal{F}_i] = \mathbf{0}$$

Large Sample Properties of the OLS Estimator

Large Sample Properties

$$\hat{\boldsymbol{\beta}}_n = \left(n^{-1}\sum_{i=1}^{n}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\left(n^{-1}\sum_{i=1}^{n}\mathbf{x}_i'Y_i\right)$$

Theorem (Consistency of $\hat{\beta}$)

Under these assumptions

$$\hat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}$$

- Consistency means that the estimate will be close – eventually – to the population value
- Without further results it is a very weak condition

Large Sample Properties

Theorem (Asymptotic Distribution of $\hat{\beta}$)

Under these assumptions

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} N(0, \boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\mathbf{S}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}) \tag{1}$$

where $\boldsymbol{\Sigma}_{\mathbf{XX}} = \mathrm{E}[\mathbf{x}_i'\mathbf{x}_i]$ *and* $\mathbf{S} = \mathrm{V}[n^{-1/2}\mathbf{X}'\boldsymbol{\epsilon}]$.

- CLT is a strong result that will form the basis of the inference we can make on β
- What good is a CLT?

Estimating the parameter covariance

- Before making inference, the covariance of $\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ must be estimated

Theorem (Asymptotic Covariance Consistency)

Under the large sample assumptions,

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_{\mathbf{XX}} &= n^{-1}\mathbf{X}'\mathbf{X} \xrightarrow{p} \boldsymbol{\Sigma}_{\mathbf{XX}} \\ \hat{\mathbf{S}} &= n^{-1}\sum_{i=1}^{n}\hat{\epsilon}_i^2\mathbf{x}_i'\mathbf{x}_i \xrightarrow{p} \mathbf{S} \\ &= n^{-1}\left(\mathbf{X}'\hat{\mathbf{E}}\mathbf{X}\right)\end{aligned}$$

and

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}^{-1}\hat{\mathbf{S}}\hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}^{-1} \xrightarrow{p} \boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\mathbf{S}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}$$

where $\hat{\mathbf{E}} = \mathrm{diag}(\hat{\epsilon}_1^2, \ldots, \hat{\epsilon}_n^2)$.

Parameter Estimation Error

Covariance

	Const.	VWM^e	SMB	HML	MOM
Const.	0.001860	-0.000171	0.000079	-0.000157	-0.000154
VWM^e	-0.000171	0.000133	-0.000060	0.000039	0.000019
SMB	0.000079	-0.000060	0.000297	0.000042	0.000019
HML	-0.000157	0.000039	0.000042	0.000441	0.000122
MOM	-0.000154	0.000019	0.000019	0.000122	0.000181

Correlation

	Const.	VWM^e	SMB	HML	MOM
Const.	–	-34.3%	10.6%	-17.3%	-26.5%
VWM^e	-34.3%	–	-29.9%	16.1%	12.3%
SMB	10.6%	-29.9%	–	11.5%	8.2%
HML	-17.3%	16.1%	11.5%	–	43.2%
MOM	-26.5%	12.3%	8.2%	43.2%	–

Bootstrap Estimation of Parameter Covariance

Alternative estimators of parameter covariance

1. Residual Bootstrap
   - Appropriate when data are conditionally homoskedastic
   - Separate selection of $\mathbf{x}_i$ and $\hat{\epsilon}_i$ when constructing bootstrap $\tilde{Y}_i$
2. Non-parametric Bootstrap
   - Works under more general conditions
   - Resamples $\{Y_i, \mathbf{x}_i\}$ as a pair

- Both are for data where the errors are not cross-sectionally correlated

Bootstraping Heteroskedastic Data

Algorithm (Nonparametric Bootstrap Regression Covariance)

1. *Generate a sets of* n *uniform integers* $\{U_i\}_{i=1}^{n}$ *on* $[1, 2, \ldots, n]$.
2. *Construct a simulated sample* $\{Y_{u_i}, \mathbf{x}_{u_i}\}$.
3. *Estimate the parameters of interest using* $Y_{u_i} = \mathbf{x}_{u_i}\boldsymbol{\beta} + \epsilon_{u_i}$, *and denote the estimate* $\tilde{\boldsymbol{\beta}}_b$.
4. *Repeat steps 1 through 3 a total of* B *times.*
5. *Estimate the variance of* $\hat{\boldsymbol{\beta}}$ *using*

$$\widehat{\mathrm{V}}\left[\hat{\boldsymbol{\beta}}\right] = \mathrm{B}^{-1}\sum_{b=1}^{B}\left(\tilde{\boldsymbol{\beta}}_j - \hat{\boldsymbol{\beta}}\right)\left(\tilde{\boldsymbol{\beta}}_j - \hat{\boldsymbol{\beta}}\right)' \text{ or } \mathrm{B}^{-1}\sum_{b=1}^{B}\left(\tilde{\boldsymbol{\beta}}_j - \overline{\tilde{\boldsymbol{\beta}}}\right)\left(\tilde{\boldsymbol{\beta}}_j - \overline{\tilde{\boldsymbol{\beta}}}\right)'$$

Bootstrap Iterations

b	Const.	VWM^e	SMB	HML	MOM
1	-0.032329	1.073407	-0.002263	0.747751	-0.059899
2	-0.098095	1.082409	0.047483	0.750626	-0.023940
3	-0.037763	1.085002	0.036935	0.776792	-0.025469
4	-0.019000	1.083847	-0.002470	0.689036	-0.053007
5	-0.052940	1.067186	0.032972	0.783204	-0.034621

Estimator Correlation

Bootstrap

	Const.	VWM^e	SMB	HML	MOM
Const.	–	-36.2%	16.2%	-11.5%	-23.0%
VWM^e	-36.2%	–	-30.3%	14.8%	9.7%
SMB	16.2%	-30.3%	–	9.4%	7.5%
HML	-11.5%	14.8%	9.4%	–	38.7%
MOM	-23.0%	9.7%	7.5%	38.7%	–

White's Estimator

	VWM^e	SMB	HML	MOM
VWM^e	–	30.1%	-22.6%	-15.0%
SMB	30.1%	–	-17.5%	-2.4%
HML	-22.6%	-17.5%	–	-19.5%
MOM	-15.0%	-2.4%	-19.5%	–

Hypothesis Testing

Elements of a hypothesis test

Definition (Null Hypothesis)

The null hypothesis, denoted H_0, is a statement about the population values of some parameters to be tested. The null hypothesis is also known as the maintained hypothesis.

- Null is important because it determines the conditions under which the distribution of $\hat{\beta}$ must be known

Definition (Alternative Hypothesis)

The alternative hypothesis, denoted H_1, is a complementary hypothesis to the null and determines the range of values of the population parameter that should lead to rejection of the null.

- Alternative is important because it determines the conditions where the null should be rejected

$$H_0 : \lambda_{\text{Market}} = 0, \;\; H_1 : \lambda_{\text{Market}} > 0 \quad \text{or} \quad H_1 : \lambda_{\text{Market}} \neq 0$$

Elements of a hypothesis test

Definition (Hypothesis Test)

A hypothesis test is a rule that specifies the values where H_0 should be rejected in favor of H_1.

- The test embeds a test statistic and a rule which determines if H_0 can be rejected
- Note: Failing to reject the null does not mean the null is accepted.

Definition (Critical Value)

The critical value for an α-sized test, denoted C_α, is the value where a test statistic, T, indicates rejection of the null hypothesis when the null is true.

- CV is the value where the null is just rejected
- CV is usually a point although can be a set

Elements of a hypothesis test

Definition (Rejection Region)

The rejection region is the region where $T > C_\alpha$.

Definition (Type I Error)

A Type I error is the event that the null is rejected when the null is *actually* valid.

- Controlling the Type I is the basis of frequentist testing
- Note: Occurs only when null is true

Definition (Size)

The size or level of a test, denoted α, is the probability of rejecting the null when the null is true. The size is also the probability of a Type I error.

- Size represents the preference for being wrong and rejecting true null

Elements of a hypothesis test

Definition (Type II Error)

A Type II error is the event that the null is not rejected when the alternative is true.

- A Type II occurs when the null is not rejected when it should be

Definition (Power)

The power of the test is the probability of rejecting the null when the alternative is true. The power is equivalently defined as 1 minus the probability of a Type II error.

- High power tests can discriminate between the null and the alternative with a relatively small amount of data

Type I & II Errors, Size and Power

- Size and power can be related to correct and incorrect decisions

		Decision	
		Do not reject H_0	Reject H_0
Truth	H_0	Correct	Type I Error (Size)
	H_1	Type II Error	Correct (Power)

Hypothesis Testing in Regression Models

Hypothesis testing in regressions

- Distribution theory allows for inference
- Hypothesis

$$H_0 : \mathbf{R}(\boldsymbol{\beta}) = 0$$

  - $\mathbf{R}(\cdot)$ is a function from $\mathbb{R}^k \to \mathbb{R}^m$, $m \leq k$
  - All equality hypotheses can be written this way

$$H_0 : (\beta_1 - 1)(\beta_2 - 1) = 0 \qquad H_0 : \frac{\beta_1 \beta_2}{\beta_1 + \beta_2} - 1 = 0$$

- Linear Equality Hypotheses (LEH)

$$H_0 : \mathbf{R}\boldsymbol{\beta} - \mathbf{r} = 0 \text{ or in long hand, } \sum_{j=1}^{k} R_{i,j}\beta_j = r_i, \; i = 1, 2, \ldots, m$$

  - $\mathbf{R}$ is an m by k matrix
  - $\mathbf{r}$ is an m by 1 vector
- Attention limited to linear hypotheses in this chapter
- Nonlinear hypotheses examined in GMM notes

What is a linear hypothesis

3-Factor FF Model: $BH_i^e = \beta_1 + \beta_2 VWM_i^e + \beta_3 SMB_i + \beta_4 HML_i + \epsilon_i$

- $H_0 : \beta_2 = 0$ [Market Neutral]
  - $\mathbf{R} = [\ 0\ 1\ 0\ 0]$
  - $\mathbf{r} = 0$
- $H_0 : \beta_2 + \beta_3 = 1$
  - $\mathbf{R} = [\ 0\ 1\ 1\ 0]$
  - $\mathbf{r} = 1$
- $H_0 : \beta_3 = \beta_4 = 0$ [CAPM with nonzero intercept]
  - $\mathbf{R} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$
  - $\mathbf{r} = [0 \quad 0]'$
- $H_0 : \beta_1 = 0, \beta_2 = 1, \beta_2 + \beta_3 + \beta_4 = 1$
  - $\mathbf{R} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix}$
  - $\mathbf{r} = [0 \quad 1 \quad 1]'$

Estimating linear regressions subject to LER

- Linear regressions subject to linear equality constraints can *always* be directly estimated using a transformed regression

$$BH_i^e = \beta_1 + \beta_2 VWM_i^e + \beta_3 SMB_i + \beta_4 HML_i + \epsilon_i$$

$$\begin{aligned} H_0 &: \beta_1 = 0, \beta_2 = 1, \beta_2 + \beta_3 + \beta_4 = 1 \\ &\Rightarrow \beta_2 = 1 - \beta_3 - \beta_4 \\ &\Rightarrow 1 = 1 - \beta_3 - \beta_4 \\ &\Rightarrow \beta_3 = -\beta_4 BH_i^e \end{aligned}$$

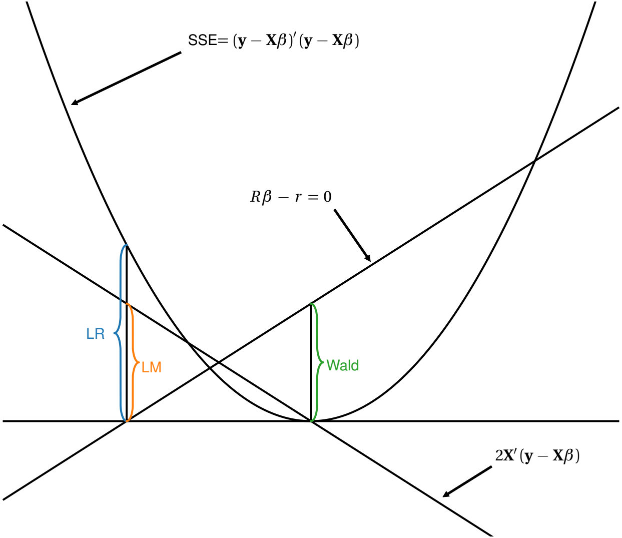
- Combine to produce restricted model

$$\begin{aligned} BH_i^e &= \mathbf{0} + \mathbf{1} VWM_i^e + \beta_3 SMB_i - \beta_3 HML_i + \epsilon_i \\ BH_i^e - \mathbf{VWM_i^e} &= \beta_3 (\mathbf{SMB_i} - \mathbf{HML_i}) + \epsilon_i \\ \tilde{R}_i &= \beta_3 \tilde{R}_i^P + \epsilon_i \end{aligned}$$

3 Major Categories of Tests

- Wald
  - Directly tests magnitude of $\mathbf{R}\boldsymbol{\beta} - \mathbf{r}$
  - t-test is a special case
  - Estimation only under alternative (unrestricted model)
- Lagrange Multiplier (LM)
  - Also Score test or Rao test
  - Tests how close to a *minimum* the sum of squared errors is if the null is true
  - Estimation only under null (restricted model)
- Likelihood Ratio (LR)
  - Tests magnitude of log-likelihood difference between the null and alternative
  - Invariant to reparameterization
    - Good thing!
  - Estimation under both null and alternative
  - Close to LM in asymptotic framework

Visualizing the three tests



Wald and t-Tests

Refresher: Normal Random Variables

- A univariate normal RV can be transformed to have any mean and variance

$$Y \sim N\left(\mu, \sigma^2\right) \Rightarrow \frac{Y - \mu}{\sigma} \sim N(0, 1)$$

- Same logic extends to m-dimensional multivariate normal random variables

$$\begin{aligned} \mathbf{y} &\sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \mathbf{y} - \boldsymbol{\mu} &\sim N(\mathbf{0}, \boldsymbol{\Sigma}) \\ \boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu}) &\sim N(\mathbf{0}, \mathbf{I}) \end{aligned}$$

- Uses property that positive definite matrix has a square root: $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{1/2}\left(\boldsymbol{\Sigma}^{1/2}\right)'$

$$\text{Cov}\left[\boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})\right] = \boldsymbol{\Sigma}^{-1/2}\text{Cov}\left[(\mathbf{y} - \boldsymbol{\mu})\right]\left(\boldsymbol{\Sigma}^{-1/2}\right)' = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}^{-1/2}\right)' = \mathbf{I}$$

- If $\mathbf{z} \equiv \boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu}) \sim N(\mathbf{0}, \mathbf{I})$ is multivariate standard normally distributed, then

$$\mathbf{z}'\mathbf{z} = \sum_{i=1}^{m} z_i^2 \sim \chi^2_m$$

t Tests

t-tests

- Single linear hypothesis: $H_0 : \mathbf{R}\boldsymbol{\beta} = r$

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\mathbf{S}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}) \Rightarrow \sqrt{n}\left(\mathbf{R}\hat{\boldsymbol{\beta}} - r\right) \xrightarrow{d} N(\mathbf{0}, \mathbf{R}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\mathbf{S}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\mathbf{R}')$$

  - Note: Under the null $H_0 : \mathbf{R}\boldsymbol{\beta} = r$
- Transform to standard normal random variable

$$z = \sqrt{n}\frac{\mathbf{R}\hat{\boldsymbol{\beta}} - r}{\sqrt{\mathbf{R}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\mathbf{S}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\mathbf{R}'}}$$

- Infeasible: Depends on unknown covariance
- Construct a feasible version using the estimate

$$t = \sqrt{n}\frac{\mathbf{R}\hat{\boldsymbol{\beta}} - r}{\sqrt{\mathbf{R}\hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}^{-1}\hat{\mathbf{S}}\hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}^{-1}\mathbf{R}'}}$$

  - Estimated variance of $\mathbf{R}\hat{\boldsymbol{\beta}}$
  - Note: Asymptotic distribution is unaffected since covariance estimator is consistent

t-test and t-stat

Unique property of t-tests

- Easily test *one-sided* alternatives

$$H_0 : \beta_1 = 0 \text{ vs. } H_1 : \beta_1 > 0$$

  - More powerful if you know the sign (e.g. risk premia)

t-stat

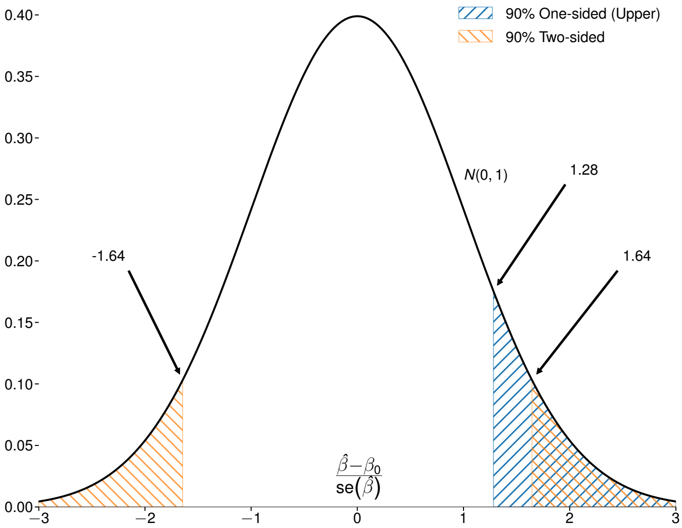
Definition (t-stat)

The t-stat of a coefficient $\hat{\beta}_k$ is test of $H_0 : \beta_k = 0$ against $H_0 : \beta_k \neq 0$, and is computed

$$\sqrt{n}\frac{\hat{\beta}_k}{\sqrt{\left(\hat{\mathbf{\Sigma}}_{\mathbf{XX}}^{-1}\hat{\mathbf{S}}\hat{\mathbf{\Sigma}}_{\mathbf{XX}}^{-1}\right)_{[kk]}}}$$

- Single most common statistic
- Reported for nearly every coefficient

Distribution and rejection region



Implementing a t Test

Algorithm (t-test)

1. *Estimate the unrestricted model* $y_i = \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i$
2. *Estimate the parameter covariance using* $\hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}^{-1} \hat{\mathbf{S}} \hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}^{-1}$
3. *Construct the restriction matrix,* $\mathbf{R}$*, and the value of the restriction,* r*, from null*
4. *Compute*

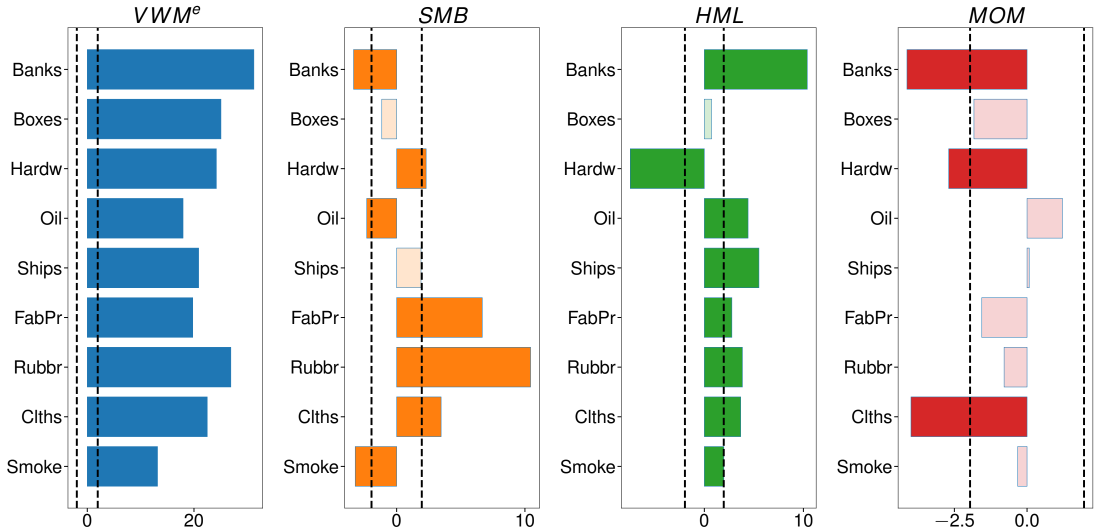
$$t = \sqrt{n} \frac{\mathbf{R}\hat{\boldsymbol{\beta}}_n - r}{\sqrt{v}}, \quad v = \mathbf{R} \hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}^{-1} \hat{\mathbf{S}} \hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}^{-1} \mathbf{R}'$$

5. *Make decision (*C_α *is the upper tail* α*-CV from* $N(0,1)$*):*
   a. *1-sided Upper: Reject the null if* $t > C_\alpha$
   b. *1-sided Lower: Reject the null if* $t < -C_\alpha$
   c. *2-sided: Reject the null if* $|t| > C_{\alpha/2}$

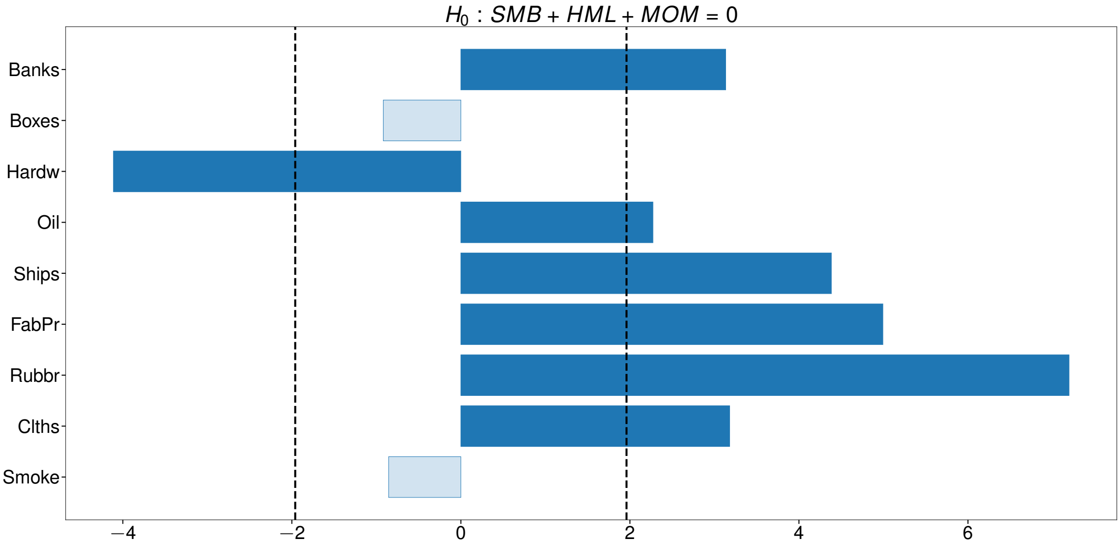
Note: Software automatically adjusts for sample size and returns $\hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}^{-1} \hat{\mathbf{S}} \hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}^{-1} / n$

t-statistics

$H_0: \beta_j = 0, H_1: \beta_j \neq 0$



t-test statistics



Wald Tests

Wald tests

- Wald tests examine validity of one or more equality restriction by measuring magnitude of $\mathbf{R}\boldsymbol{\beta} - \mathbf{r}$
  - For same reasons as t-test, under the null

$$\sqrt{n}\left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}\right) \xrightarrow{d} N(\mathbf{0}, \mathbf{R}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\mathbf{S}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\mathbf{R}')$$

  - Standardized and squared

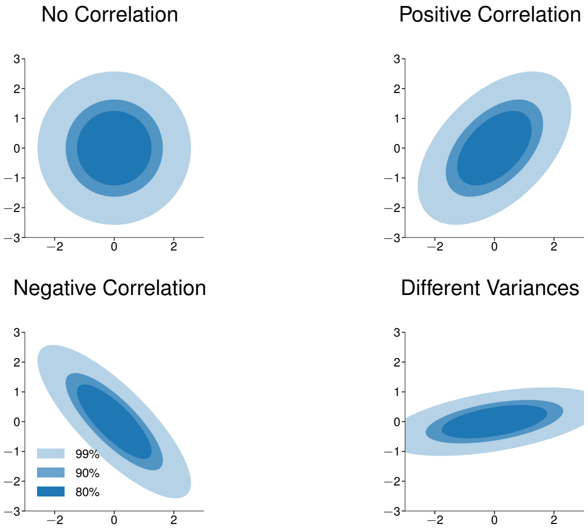
$$W = n(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})' \left[\mathbf{R}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\mathbf{S}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\mathbf{R}'\right]^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}) \xrightarrow{d} \chi^2_m$$

  - Again, this is infeasible, so use the feasible version

$$W = n(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})' \left[\mathbf{R}\hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}^{-1}\hat{\mathbf{S}}\hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}^{-1}\mathbf{R}'\right]^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}) \xrightarrow{d} \chi^2_m$$

Bivariate confidence sets

Correlation between $\hat{\beta}_1$ and $\hat{\beta}_1$



Implementing a Wald Test

Algorithm (Large Sample Wald Test)

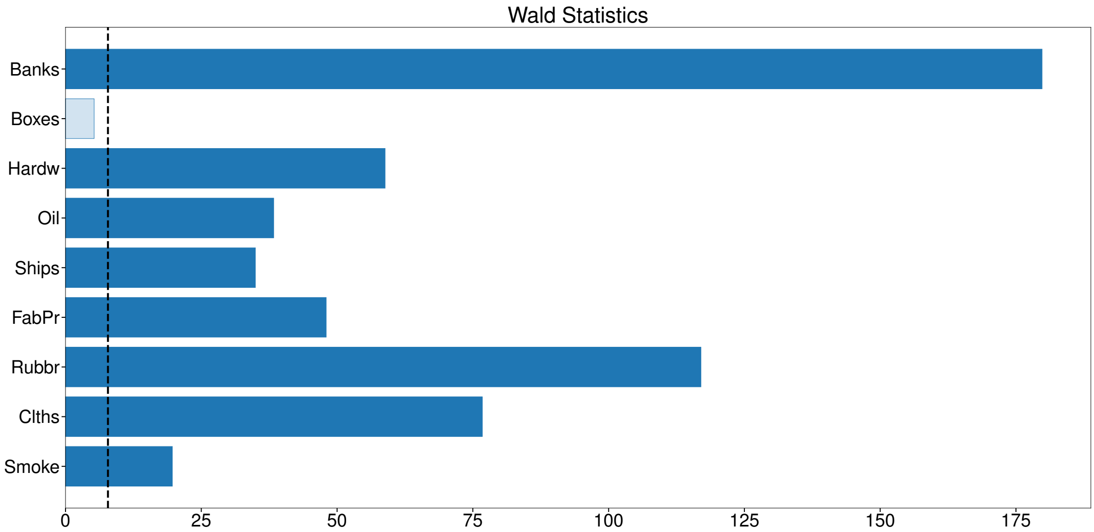
1. *Estimate the unrestricted model* $y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i$.
2. *Estimate the parameter covariance using* $\hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}^{-1}\hat{\mathbf{S}}\hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}^{-1}$ *where*

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{XX}} = n^{-1}\sum_{i=1}^{n}\mathbf{x}_i'\mathbf{x}_i, \quad \hat{\mathbf{S}} = n^{-1}\sum_{i=1}^{n}\hat{\epsilon}_i^2\mathbf{x}_i'\mathbf{x}_i$$

3. *Construct the restriction matrix,* $\mathbf{R}$*, and the value of the restriction,* $\mathbf{r}$*, from the null hypothesis.*
4. *Compute* $W = n(\mathbf{R}\hat{\boldsymbol{\beta}}_n - \mathbf{r})'\left[\mathbf{R}\hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}^{-1}\hat{\mathbf{S}}\hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}^{-1}\mathbf{R}'\right]^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}}_n - \mathbf{r})$.
5. *Reject the null if* $W > C_\alpha$ *where* C_α *is the critical value from a* χ^2_m *using a size of* α.

Wald Test Statistics

$H_0 : \beta_{SMB} = \beta_{HML} = \beta_{MOM} = 0$



Lagrange Multiplier Tests

Lagrange Multiplier (LM) tests

- LM tests examine *shadow price* of the constraint (null)

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\,(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \text{ subject to } \mathbf{R}\boldsymbol{\beta}-\mathbf{r}=0.$$

- Lagrangian

$$\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\lambda}) = (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + (\mathbf{R}\boldsymbol{\beta}-\mathbf{r})'\boldsymbol{\lambda}$$

- If null true, then $\boldsymbol{\lambda} \approx \mathbf{0}$
- FOC:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}}) + \mathbf{R}'\tilde{\boldsymbol{\lambda}} = \mathbf{0}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\lambda}} = \mathbf{R}\tilde{\boldsymbol{\beta}} - \mathbf{r} = \mathbf{0}$$

- A few minutes of matrix algebra later

$$\tilde{\boldsymbol{\lambda}} = 2\left[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}}-\mathbf{r})$$

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\left[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}}-\mathbf{r})$$

  - $\hat{\boldsymbol{\beta}}$ is the OLS estimator, $\tilde{\boldsymbol{\beta}}$ is the estimator computed under the null

Why LM tests are also known as score tests...

$$\tilde{\boldsymbol{\lambda}} = 2\left[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})$$

- $\tilde{\boldsymbol{\lambda}}$ is just a function of normal random variables (via $\hat{\boldsymbol{\beta}}$, the OLS estimator)
- Alternatively,

$$\mathbf{R}'\tilde{\boldsymbol{\lambda}} = -2\mathbf{X}'\tilde{\boldsymbol{\epsilon}}$$

  - $\mathbf{R}$ has rank m, so $\mathbf{R}'\boldsymbol{\lambda} \approx \mathbf{0} \Leftrightarrow \mathbf{X}'\tilde{\boldsymbol{\epsilon}} \approx \mathbf{0}$
  - $\tilde{\boldsymbol{\epsilon}}$ are the estimated residuals *under the null*
- Under the assumptions,

$$\sqrt{n}\tilde{\mathbf{s}} = \sqrt{n}\left(n^{-1}\mathbf{X}'\tilde{\boldsymbol{\epsilon}}\right) \xrightarrow{d} N(\mathbf{0}, \mathbf{S})$$

- We know how to test multivariate normal random variables for equality to 0

$$LM = n\tilde{\mathbf{s}}'\mathbf{S}^{-1}\tilde{\mathbf{s}} \xrightarrow{d} \chi^2_m$$

- But we always have to use the feasible version,

$$LM = n\tilde{\mathbf{s}}'\hat{\tilde{\mathbf{S}}}^{-1}\tilde{\mathbf{s}} = n\tilde{\mathbf{s}}'\left(n^{-1}\mathbf{X}'\tilde{\mathbf{E}}\mathbf{X}\right)^{-1}\tilde{\mathbf{s}} \xrightarrow{d} \chi^2_m$$

Note: $\hat{\tilde{\mathbf{S}}}$ (and $\tilde{\mathbf{E}}$) is estimated using the errors from the *restricted* regression.

Implementing a LM test

Algorithm (Large Sample Lagrange Multiplier Test)

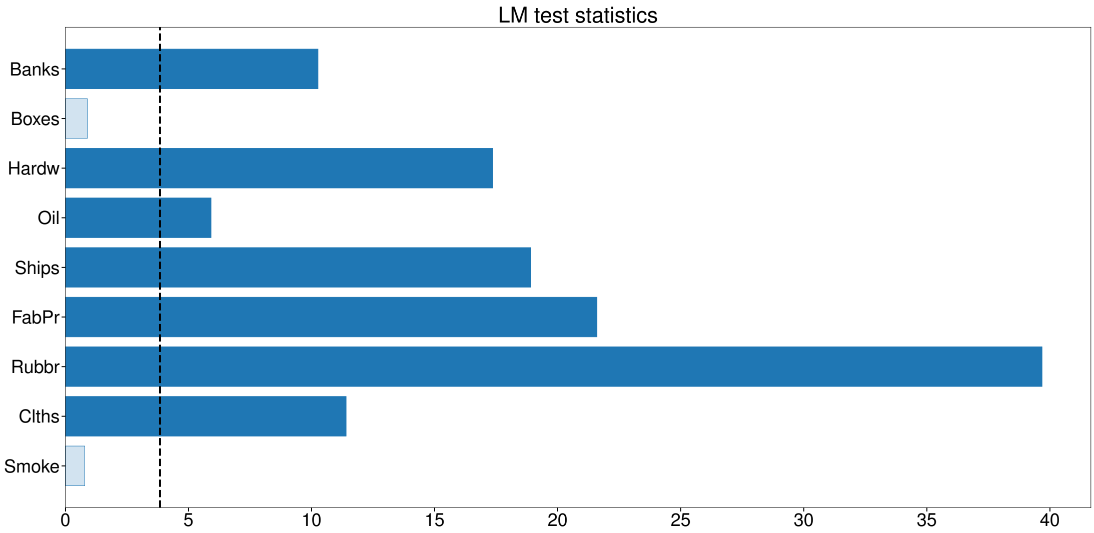
1. *Form the unrestricted model,* $Y_i = \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i$.
2. *Impose the null on the unrestricted model and estimate the restricted model,* $Y_i = \tilde{\mathbf{x}}_i \boldsymbol{\beta} + \epsilon_i$.
3. *Compute the residuals from the restricted regression,* $\tilde{\epsilon}_i = Y_i - \tilde{\mathbf{x}}_i \tilde{\boldsymbol{\beta}}$.
4. *Construct the score using the residuals from the restricted regression from both models,* $\tilde{\mathbf{s}}_i = \mathbf{x}_i' \tilde{\epsilon}_i$ *where* $\mathbf{x}_i$ *are the regressors from the unrestricted model.*
5. *Estimate the average score and the covariance of the score,*

$$\tilde{\mathbf{s}} = n^{-1} \sum_{i=1}^{n} \tilde{\mathbf{s}}_i, \qquad \hat{\tilde{\mathbf{S}}} = n^{-1} \sum_{i=1}^{n} \tilde{\mathbf{s}}_i \tilde{\mathbf{s}}_i' \tag{2}$$

6. *Compute the LM test statistic as* $LM = n \tilde{\mathbf{s}}' \hat{\tilde{\mathbf{S}}}^{-1} \tilde{\mathbf{s}}$ *and compare to the critical value from a* χ^2_m *using a size of* α.

Lagrange Multiplier (LM) Test Statistics

$H_0 : \beta_{SMB} = \beta_{HML} = \beta_{MOM} = 0$



Likelihood Ratio Tests

Likelihood ratio (LR) tests

- A "large" sample LR test can be constructed using a test statistic that looks like the LM test
- Formally the large-sample LR is based on testing whether the difference of the scores, evaluated at the restricted and unrestricted parameters, is large – in a statistically meaningful sense
- Suppose $\mathbf{S}$ is known, then

$$n\left(\tilde{\mathbf{s}}-\hat{\mathbf{s}}\right)'\mathbf{S}^{-1}\left(\tilde{\mathbf{s}}-\hat{\mathbf{s}}\right)=n\left(\tilde{\mathbf{s}}-\mathbf{0}\right)'\mathbf{S}^{-1}\left(\tilde{\mathbf{s}}-\mathbf{0}\right) \quad (\text{ Why?})$$

$$n\tilde{\mathbf{s}}'\mathbf{S}^{-1}\tilde{\mathbf{s}} \xrightarrow{d} \chi^2_m$$

- Leads to definition of large sample LR – identical to LM but uses a difference variance estimator

$$LR = n\tilde{\mathbf{s}}'\hat{\mathbf{S}}^{-1}\tilde{\mathbf{s}} \xrightarrow{d} \chi^2_m$$

Note: $\hat{\mathbf{S}}$ (and $\hat{\mathbf{E}}$) is estimated using the errors from the *unrestricted* regression.

- $\hat{\mathbf{S}}$ is estimated *under the alternative* and $\tilde{\mathbf{S}}$ is estimated *under the null*
- $\hat{\mathbf{S}}$ is usually "smaller" than $\tilde{\mathbf{S}} \Rightarrow LR$ is usually larger than LM

Implementing a LR test

Algorithm (Large Sample Likelihood Ratio Test)

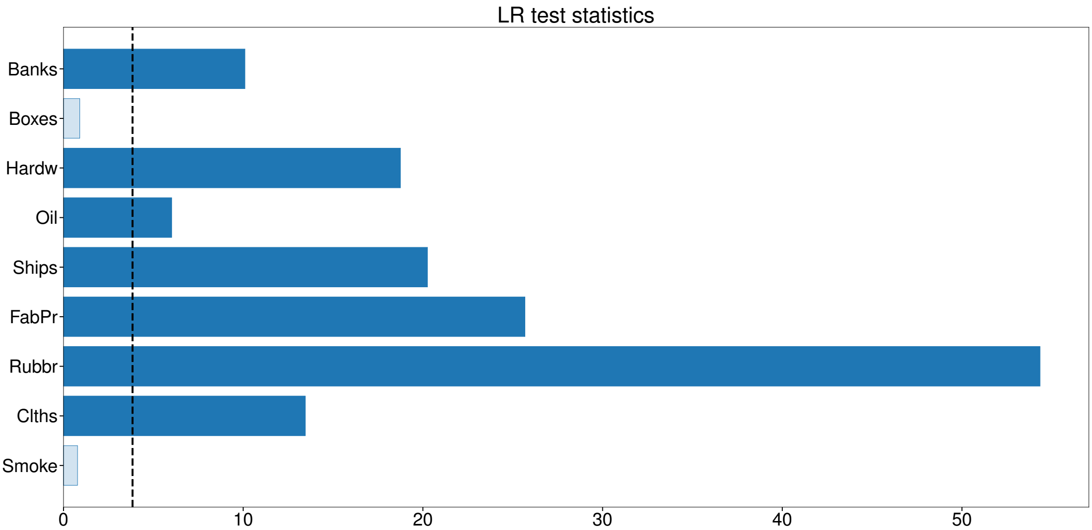
1. *Estimate the unrestricted model* $Y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i$.
2. *Impose the null on the unrestricted model and estimate the restricted model,* $Y_i = \tilde{\mathbf{x}}_i\boldsymbol{\beta} + \epsilon_i$.
3. *Compute the residuals from the restricted regression,* $\tilde{\epsilon}_i = y_i - \tilde{\mathbf{x}}_i\tilde{\boldsymbol{\beta}}$*, and from the unrestricted regression,* $\hat{\epsilon}_i = y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}$.
4. *Construct the score from both models,* $\tilde{\mathbf{s}}_i = \mathbf{x}_i'\tilde{\epsilon}_i$ *and* $\hat{\mathbf{s}}_i = \mathbf{x}_i'\hat{\epsilon}_i$*, where in both cases* $\mathbf{x}_i$ *are the regressors from the unrestricted model.*
5. *Estimate the average score and the covariance of the score,*

$$\tilde{\mathbf{s}} = n^{-1}\sum_{i=1}^{n}\tilde{\mathbf{s}}_i, \qquad \hat{\mathbf{S}} = n^{-1}\sum_{i=1}^{n}\hat{\mathbf{s}}_i\hat{\mathbf{s}}_i' \tag{3}$$

6. *Compute the LR test statistic as* $LR = n\tilde{\mathbf{s}}'\hat{\mathbf{S}}^{-1}\tilde{\mathbf{s}}$ *and compare to the critical value from a* χ^2_m *using a size of* α.

Asymptotic Likelihood Ratio (LR) Test Statistics

$H_0 : \beta_{SMB} = \beta_{HML} = \beta_{MOM} = 0$



Likelihood ratio (LR) tests (Classic Assumptions)

- If null is *close* to alternative, log-likelihood should be similar under both

$$LR = -2\ln\left(\frac{\max_{\boldsymbol{\beta},\sigma^2} f(\mathbf{y}|\mathbf{X};\boldsymbol{\beta},\sigma^2) \quad \text{subject to} \quad \mathbf{R}\boldsymbol{\beta}=\mathbf{r}}{\max_{\boldsymbol{\beta},\sigma^2} f(\mathbf{y}|\mathbf{X};\boldsymbol{\beta},\sigma^2)}\right)$$

- A little simple algebra later...

$$LR = n\ln\left(\frac{SSE_R}{SSE_U}\right) = n\ln\left(\frac{s_R^2}{s_U^2}\right)$$

- In classical setup, distribution LR is

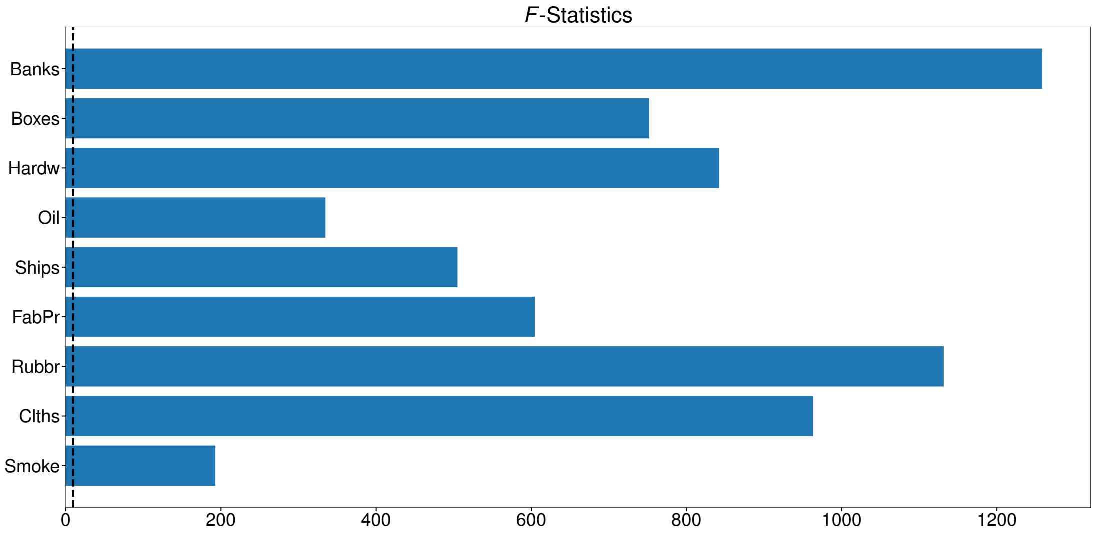
$$\frac{n-k}{m}\left[\exp\left(\frac{LR}{n}\right)-1\right] \sim F_{m,n-k}$$

- Although $m \times LR \to \chi^2_m$ as $n \to \infty$

Warning: The distribution of the LR *critically* relies on homoskedasticity and normality

F-statistics

H_0: All $\beta_j = 0$ except constant



Choosing a Test

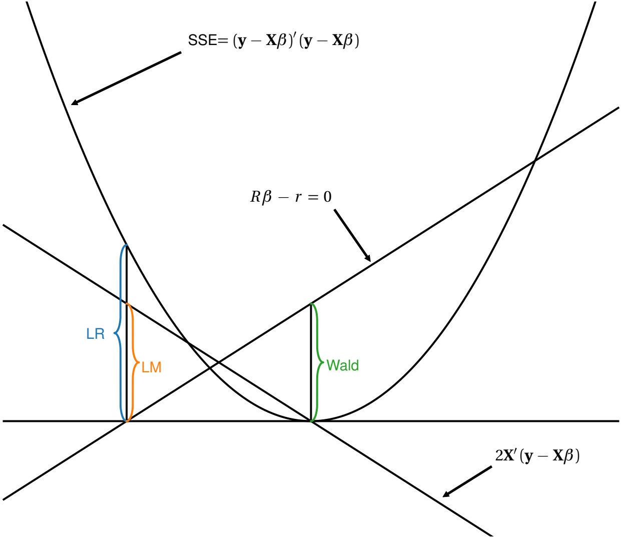
Comparing the three tests

- Asymptotically all are equivalent
- Rule of thumb: $W \approx LR > LM$ since W and LR use errors estimated under the alternative
  - Larger test statistics are good since all have same distribution $\Rightarrow$ more power
- If derived from MLE (Classical Assumptions: normality, homoskedasticity), an exact relationship:

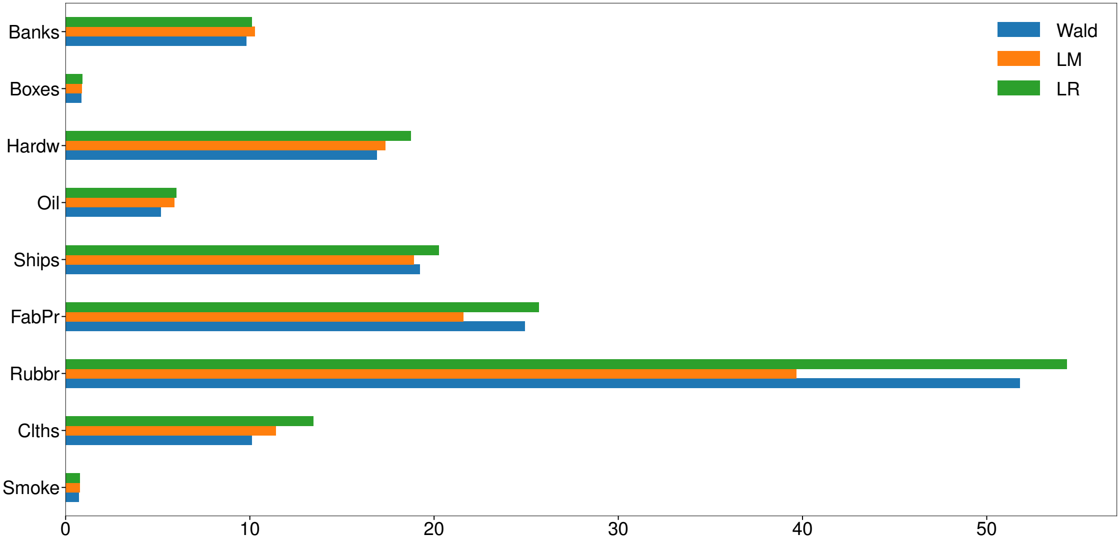
$$W = LR > LM$$

- In some contexts (not linear regression) ease of estimation is a useful criteria to prefer one test over the others
  - Easy estimation of null: LM
  - Easy estimation of alternative: Wald
  - Easy to estimate both: LR or Wald

Comparing the three



Comparing the Three Test Statistics



Heteroskedasticity

Heteroskedasticity

- Heteroskedasticity:
  - *hetero*: Different
  - *skedannumi*: To scatter
- Heteroskedasticity is pervasive in financial data
- Usual covariance estimator (previously given) allows for Heteroskedasticity of unknown form
- Tempting to always use “Heteroskedasticity Robust Covariance” estimator
  - Also known as White’s Covariance (Eicker/Huber) estimator
- Finite sample properties are generally worse if data are homoskedastic
- If data are homoskedastic can use a simpler estimator
- Required condition for simpler estimator:

$$\mathrm{E}\left[\epsilon_i^2 X_{j,i} X_{l,i} | X_{j,i}, X_{l,i}\right] = \mathrm{E}\left[\epsilon_i^2\right] X_{j,i} X_{l,i}$$

for $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, k$, and $l = 1, 2, \ldots, k$ to justify simpler estimator.

White's Estimator	Classic Estimator
Heteroskedasticity Robust	Requires Homoskedasticity
$\hat{\mathbf{\Sigma}}_{\mathbf{XX}}^{-1}\hat{\mathbf{S}}\hat{\mathbf{\Sigma}}_{\mathbf{XX}}^{-1}$	$\hat{\sigma}^2\hat{\mathbf{\Sigma}}_{\mathbf{XX}}^{-1}$

- White's Covariance estimator has worse finite sample properties
- Should be avoided if homoskedasticity plausible

White's test

- Implemented using an auxiliary regression

$$\hat{\epsilon}_i^2 = \mathbf{z}_i\boldsymbol{\gamma} + \eta_i$$

- $\mathbf{z}_i$ consist of all cross products of $X_{i,p}X_{i,q}$ for $p, q \in \{1, 2, \ldots, k\}$, $p \neq q$
- LM test that all coefficients on parameters (except the constant) are zero

$$H_0 : \gamma_2 = \gamma_3 = \ldots = \gamma_{k\cdot(k+1)/2} = \mathbf{0}$$

- $Z_{1,i} = 1$ is always a constant – never tested

Implementing White's Test for Heteroskedasticity

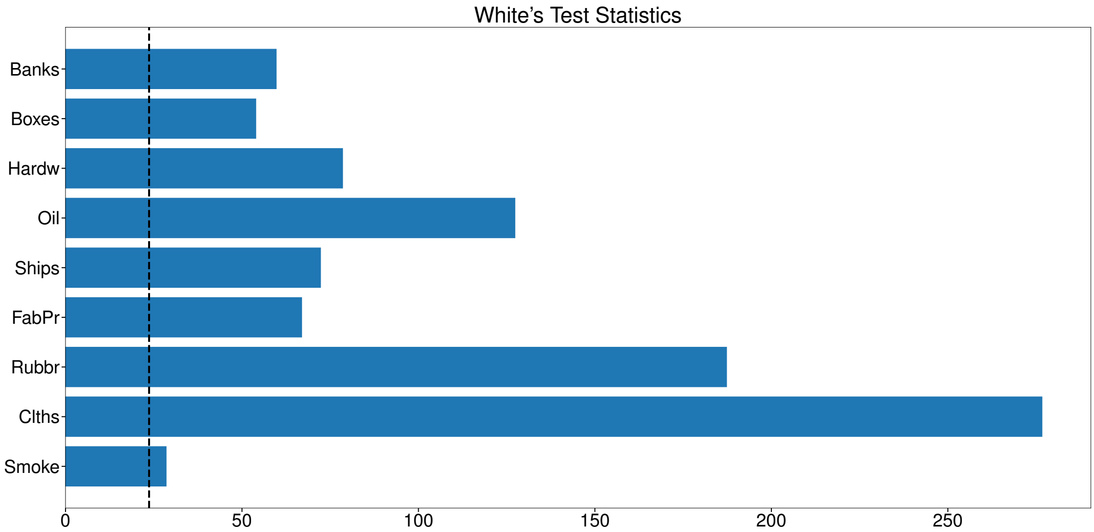
Algorithm (White's Test for Heteroskedasticity)

1. *Fit the model* $Y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i$
2. *Construct the fit residuals* $\hat{\epsilon}_i = Y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}$
3. *Construct the auxiliary regressors* $\mathbf{z}_i$ *where the* $k(k+1)/2$ *elements of* $\mathbf{z}_i$ *are computed from* $X_{i,o}X_{i,p}$ *for* $o = 1, 2, \ldots, k$, $p = o, o+1, \ldots, k$.
4. *Estimate the auxiliary regression* $\hat{\epsilon}_i^2 = \mathbf{z}_i\boldsymbol{\gamma} + \eta_i$
5. *Compute White's Test statistic as* nR^2 *where the* R^2 *is from the auxiliary regression and compare to the critical value at size* α *from a* $\chi^2_{k(k+1)/2-1}$.

Note: This algorithm assumes the model contains a constant. If the original model *does not* contain a constant, then $\mathbf{z}_i$ should be augmented with a constant, and the asymptotic distribution is a $\chi^2_{k(k+1)/2}$.

White's Test for Heteroskedasticity

Base model is CAPM



Estimating the parameter covariance (Homoskedasticity)

Theorem (Homoskedastic CLT)

Under the large sample assumptions, and if the errors are homoskedastic,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} N(0, \sigma^2 \boldsymbol{\Sigma}_{\mathbf{XX}}^{-1})$$

where $\boldsymbol{\Sigma}_{\mathbf{XX}} = \mathrm{E}[\mathbf{x}_i'\mathbf{x}_i]$ *and* $\sigma^2 = \mathrm{V}[\epsilon_i]$

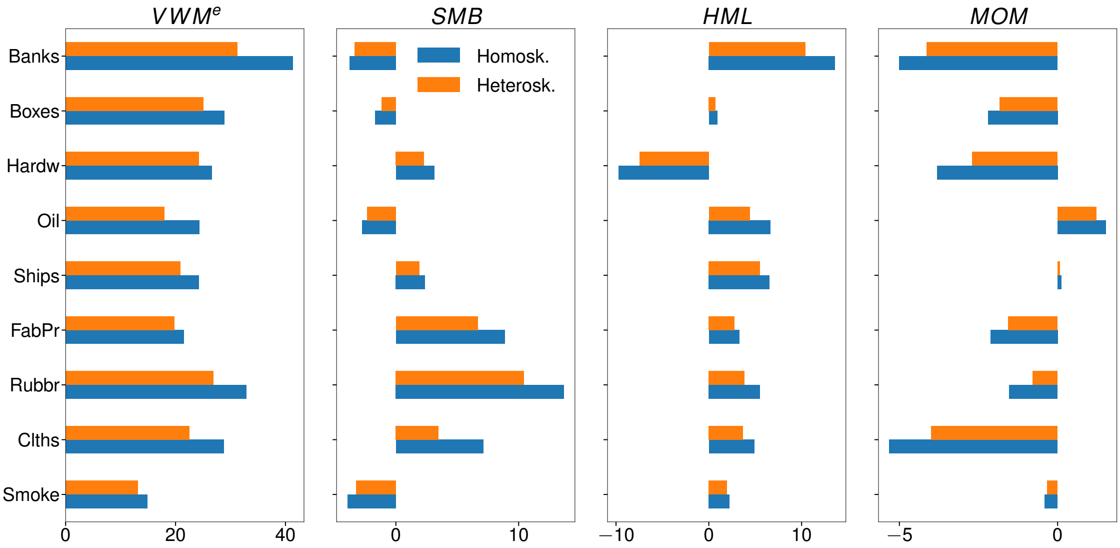
Theorem (Homoskedastic Covariance Estimator)

Under the large sample assumptions, and if the errors are homoskedastic,

$$\hat{\sigma}^2 \hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}^{-1} \xrightarrow{p} \sigma^2 \boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}$$

- Homoskedasticity justifies the "usual" estimator $\hat{\sigma}^2 \left(n^{-1}\mathbf{X}'\mathbf{X}\right)^{-1}$
  - When using financial data this is the "unusual" estimator

Comparing t-statistics



Bootstraping Homoskedastic Data

Algorithm (Residual Bootstrap Regression Covariance)

1. *Generate 2 sets of* n *uniform integers* $\{U_{1,i}\}_{i=1}^{n}$ *and* $\{U_{2,i}\}_{i=1}^{n}$ *on* $[1, 2, \ldots, n]$.
2. *Construct a simulated sample* $\left\{\tilde{Y}_{u_{1,i}} = \mathbf{x}_{u_{1,i}}\hat{\boldsymbol{\beta}} + \hat{\epsilon}_{u_{2,i}}\right\}$.
3. *Estimate the parameters of interest using* $\tilde{Y}_{u_{1,i}} = \mathbf{x}_{u_{1,i}}\boldsymbol{\beta} + \tilde{\epsilon}_{u_{1,i}}$, *and denote the estimate* $\tilde{\boldsymbol{\beta}}_b$.
4. *Repeat steps 1 through 3 a total of* B *times.*
5. *Estimate the variance of* $\hat{\boldsymbol{\beta}}$ *using*

$$\widehat{\mathrm{V}}\left[\hat{\boldsymbol{\beta}}\right] = \mathrm{B}^{-1}\sum_{b=1}^{B}\left(\tilde{\boldsymbol{\beta}}_j - \hat{\boldsymbol{\beta}}\right)\left(\tilde{\boldsymbol{\beta}}_j - \hat{\boldsymbol{\beta}}\right)' \text{ or } \mathrm{B}^{-1}\sum_{b=1}^{B}\left(\tilde{\boldsymbol{\beta}}_j - \overline{\tilde{\boldsymbol{\beta}}}\right)\left(\tilde{\boldsymbol{\beta}}_j - \overline{\tilde{\boldsymbol{\beta}}}\right)'$$

Specification Failures

Problems with models

What happens when the assumptions are violated?

- Model misspecified
  - Omitted variables
  - Extraneous Variables
  - Functional Form
- Heteroskedasticity
- Too few moments
- Errors correlated with regressors
  - Rare in Asset Pricing and Risk Management
  - Common on Corporate Finance

Not enough moments

- Too few moments causes problems for both $\hat{\boldsymbol{\beta}}$ and t-stats
  - Consistency requires 2 moments for $\mathbf{x}_i$, 1 for ϵ_i
  - Consistent estimation of variance requires 4 moments of $\mathbf{x}_i$ and 2 of ϵ_i
- Fewer than 2 moments of $\mathbf{x}_i$
  - Slopes can still be consistent
  - Intercepts cannot
- Fewer than 1 for ϵ_i
  - $\hat{\boldsymbol{\beta}}$ is inconsistent
    - Too much noise!
- Between 2 and 4 moments of $\mathbf{x}_i$ or 1 and 2 of ϵ_i
  - Tests are inconsistent

Omitted Variables

What if the linearity assumption is violated?

- Omitted variables

Correct Model	$y_i = \mathbf{x}_{1,i}\boldsymbol{\beta}_1 + \mathbf{x}_{2,i}\boldsymbol{\beta}_2 + \epsilon_i$
Model Estimated	$y_i = \mathbf{x}_{1,i}\boldsymbol{\beta}_1 + \epsilon_i$

- Can show

$$\hat{\boldsymbol{\beta}}_1 \xrightarrow{p} \boldsymbol{\beta}_1 + \boldsymbol{\delta}'\boldsymbol{\beta}_2$$

$$\mathbf{x}_{2,i} = \mathbf{x}_{1,i}\boldsymbol{\delta} + \boldsymbol{\nu}_i$$

- $\hat{\boldsymbol{\beta}}_1$ captures any portion of Y_i explainable by $\mathbf{x}_{1,i}$
  - $\boldsymbol{\beta}_1$ from model
  - $\boldsymbol{\beta}_2$ through correlation between $\mathbf{x}_{1,i}$ and $\mathbf{x}_{2,i}$
- Two cases where omitted variables do not produce bias
  - $\mathbf{x}_{1,i}$ and $\mathbf{x}_{2,i}$ uncorrelated, .e.g, some dummy variable models
    - Estimated variance remains inconsistent
  - $\boldsymbol{\beta}_2 = \mathbf{0}$: Model correct

Extraneous Variables

Correct model	$Y_i = \mathbf{x}_{1,i}\boldsymbol{\beta}_1 + \epsilon_i$
Model Estimated	$Y_i = \mathbf{x}_{1,i}\boldsymbol{\beta}_1 + \mathbf{x}_{2,i}\boldsymbol{\beta}_2 + \epsilon_i$

- Can show:

$$\hat{\boldsymbol{\beta}}_1 \xrightarrow{p} \boldsymbol{\beta}_1$$

- No problem, right?
  - Including extraneous regressors increase parameter uncertainty
  - Excluding marginally relevant regressors reduces parameter uncertainty but increases chance model is misspecified
- Bias-Variance Trade off
  - Smaller models reduce variance, even if introducing bias
  - Large models have less bias
  - Related to model selection...

Heteroskedasticity

- Common problem across most financial data sets
  - Asset returns
  - Firm characteristics
  - Executive compensation
- Solution 1: Heteroskedasticity robust covariance estimator

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}^{-1}\hat{\mathbf{S}}\hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}^{-1}$$

- Partial Solution 2 : Use data transformations
  - Ratios:
    - Volume vs. Turnover (Volume/Shares Outstanding)
  - Logs: Volume vs. $\ln$ Volume
    - Volume = Size · Shock
    - $\ln$ Volume = $\ln$ Size + $\ln$ Shock

GLS and FGLS

Solution 3: Generalized Least Squares (GLS)

$$\hat{\boldsymbol{\beta}}_n^{\text{GLS}} = (\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{-1}\mathbf{y}, \quad \mathbf{W} \text{ is } n \times n \text{ positive definite}$$

$$\hat{\boldsymbol{\beta}}_n^{\text{GLS}} \xrightarrow{p} \boldsymbol{\beta}$$

- Can choose $\mathbf{W}$ cleverly so that $\mathbf{W}^{-\frac{1}{2}}\boldsymbol{\epsilon}$ is homoskedastic and uncorrelated
- $\hat{\boldsymbol{\beta}}^{\text{GLS}}$ is asymptotically efficient
- In practice $\mathbf{W}$ is unknown, but can be estimated

$$\hat{\epsilon}_i^2 = \mathbf{z}_i\boldsymbol{\gamma} + \eta_i$$

$$\hat{\mathbf{W}} = \text{diag}(\mathbf{z}_i\hat{\boldsymbol{\gamma}})$$

- Resulting estimator is Feasible GLS (FGLS)
  - Still asymptotically efficient
  - Small sample properties are not assured – may be quite bad
- Compromise implementation: Use pre-specified but potentially sub-optimal $\mathbf{W}$
  - Example: Diagonal which ignores any potential correlation
  - Requires alternative estimator of parameter covariance, similar to White (notes)

Model Selection

Model Building

- The Black Art of econometric analysis
- Many rules and procedures
  - Most contradictory
- Always a trade-off between bias and variance in finite sample
- Better models usually have a finance or economic theory behind them
- Three distinct steps
  - Model Selection
  - Specification Checking
  - Model Evaluation using pseudo out-of-sample (OOS) evaluation
    - Common to use actual out-of-sample data in trading models

Strategies

- General to Specific
  - Fit largest specification
  - Drop largest p-val
  - Refit
  - Stop if all p-values indicate significance at size α
    - α is the econometrician's choice
- Specific to General
  - Fit all specifications that include a single explanatory variable
  - Include variable with the smallest p-val
  - Starting from this model, test all other variables by adding in one-at-a-time
  - Stop if no p-val of an excluded variable indicates significance at size α

Information Criteria

- Information Criteria
  - Akaike Information Criterion (AIC)

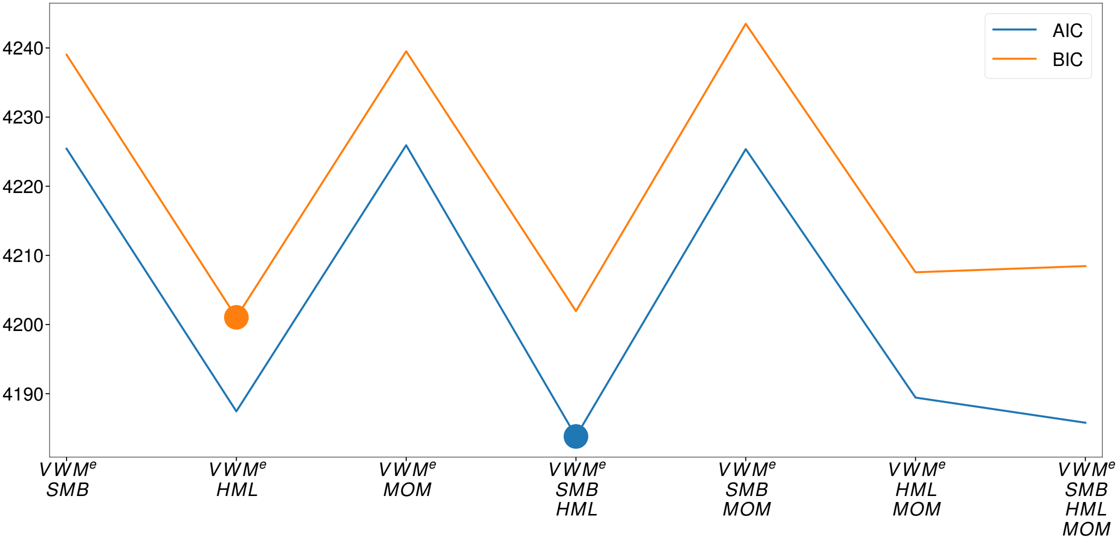
$$AIC = \ln \hat{\sigma}^2 + 2\frac{k}{n}$$

  - Schwartz (Bayesian) Information Criterion (SIC/BIC)

$$BIC = \ln \hat{\sigma}^2 + k\frac{\ln n}{n}$$

- Both have versions suitable for likelihood based estimation
- Reward for better fit: Reduce $\ln \hat{\sigma}^2$
- Penalty for more parameters: $2\frac{k}{n}$ or $k\frac{\ln n}{n}$
- Choose model with smallest IC
  - AIC has fixed penalty $\Rightarrow$ inclusion of extraneous variables
  - BIC has larger penalty if $\ln n > 2$ ($n > 7$)

Information Criteria Model Selection



Cross-Validation

- Use $100 - m\%$ to estimate parameters, evaluate using remaining $m\%$
- $m = 100 \times k^{-1}$ in k-fold cross-validation

Algorithm (k-fold cross-validation)

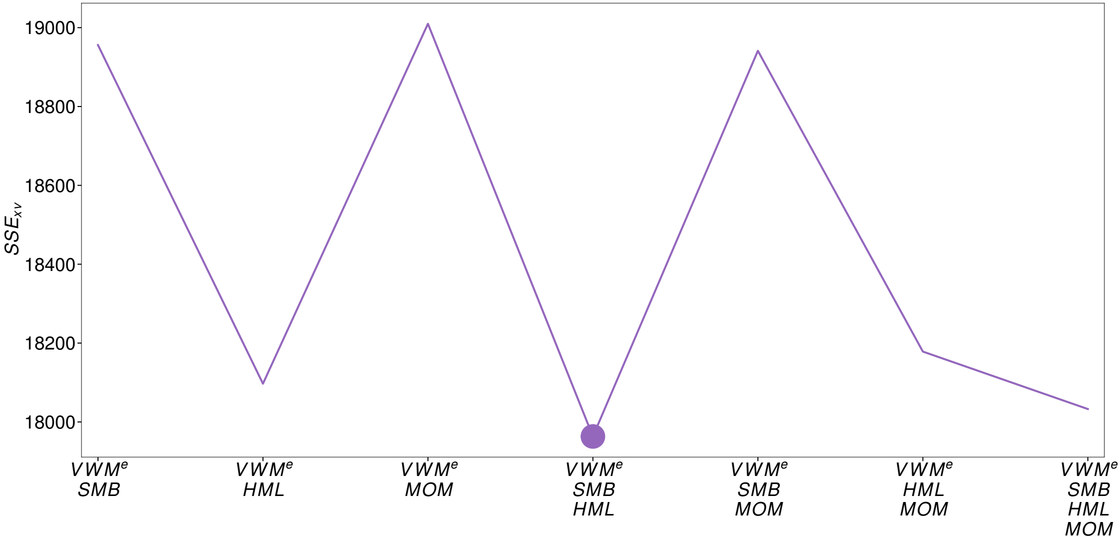
1. *For each model:*
   a. *Randomly divide observations into k-equally sized blocks, $\mathcal{S}_j$, $j = 1, \ldots, k$*
   b. *For $j = 1, \ldots, k$ estimate $\hat{\boldsymbol{\beta}}_j$ by excluding the observations in block j*
   c. *Compute cross-validated SSE using observations in block j and $\hat{\boldsymbol{\beta}}_j$*

$$\mathrm{SSE}_{xv} = \sum_{j=1}^{k} \sum_{i \in \mathcal{S}_j} \left(y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}_j \right)^2$$

2. *Select model with lowest cross-validated SSE*

- Typical values for k are 5 or 10

Cross-validation Model Selection



Checking for Specification Errors

Specification Analysis

- Is a selected model any good?

$$Y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i$$

Common Specification Tests

- Stability Test: Chow

$$Y_i = \mathbf{x}_i\boldsymbol{\beta} + I_{[i>C]}\mathbf{x}_i\boldsymbol{\gamma} + \epsilon_i$$

  - $H_0 : \boldsymbol{\gamma} = \mathbf{0}$
- Nonlinearity Test: Ramsey's RESET

$$Y_i = \mathbf{x}_i\boldsymbol{\beta} + \gamma_1\hat{Y}_i^2 + \gamma_2\hat{Y}_i^3 + \ldots + \gamma_{L-1}\hat{Y}_i^L + \epsilon_i$$

  - $H_0 : \boldsymbol{\gamma} = \mathbf{0}$
- Recursive and/or Rolling Estimation
- Influence Function
  - Influence: $\mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i' \Leftarrow$ Normalized length of $\mathbf{x}_i$
- Normality Tests: Jarque-Bera

$$JB = n\left(\frac{sk^2}{6} + \frac{(\kappa - 3)^2}{24}\right) \sim \chi^2_2$$

Implementing a Chow & RESET Tests

Algorithm (Chow Test)

1. *Estimate the model* $Y_i = \mathbf{x}_i\boldsymbol{\beta} + I_{[i>C]}\mathbf{x}_i\boldsymbol{\gamma} + \epsilon_i$*.*
2. *Test the null* $H_0 : \boldsymbol{\gamma} = 0$ *against the alternative* $H_1 : \gamma_i \neq 0$*, for some* i*, using a Wald, LM or LR test using a* χ^2_k *test.*

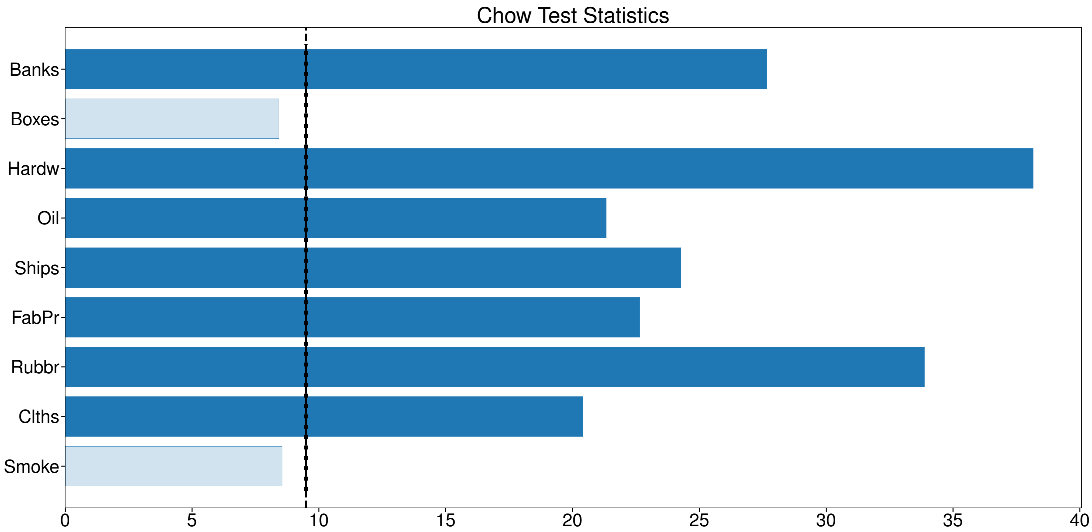
Note: Chow tests can only be used when the break date is known. Taking the maximum Chow test statistic over multiple possible break dates changes the distribution of the test statistic under the null of no break.

Algorithm (RESET Test)

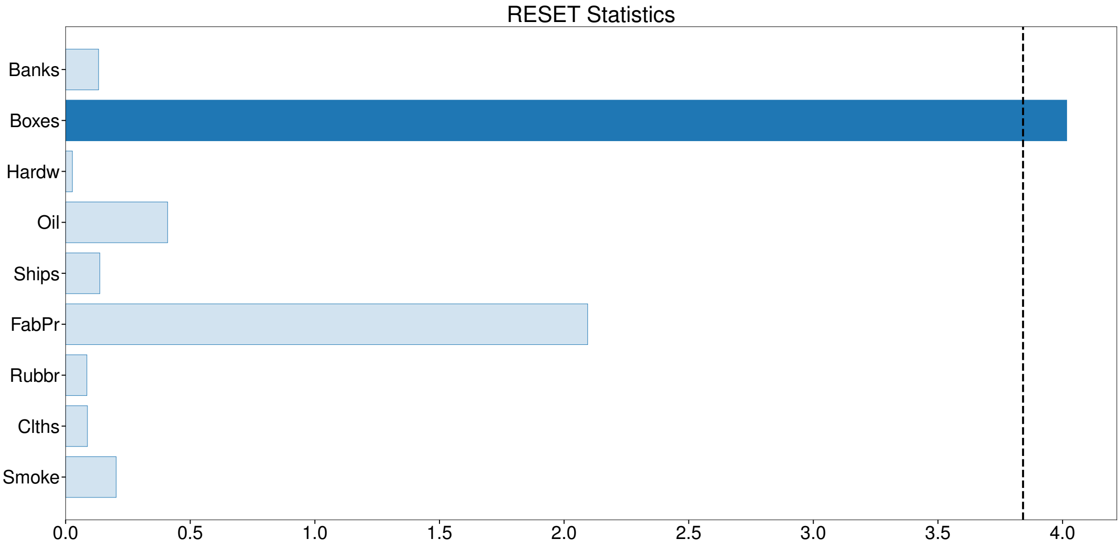
1. *Estimate the model* $Y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i$ *and construct the fit values ,* $\hat{Y}_i = \mathbf{x}_i\hat{\boldsymbol{\beta}}$*.*
2. *Re-estimate the model* $Y_i = \mathbf{x}_i\boldsymbol{\beta} + \gamma_1\hat{Y}_i^2 + \gamma_2\hat{Y}_i^3 + \ldots + \epsilon_i$*.*
3. *Test the null* $H_0 : \gamma_1 = \gamma_2 = \ldots = \gamma_m = 0$ *against the alternative* $H_1 : \gamma_i \neq 0$*, for some* i*, using a Wald, LM or LR test, all of which have a* χ^2_m *distribution.*

Chow Test for Structural Change

October 1987 Break



Ramsey's RESET Test for Neglected Nonlinearity



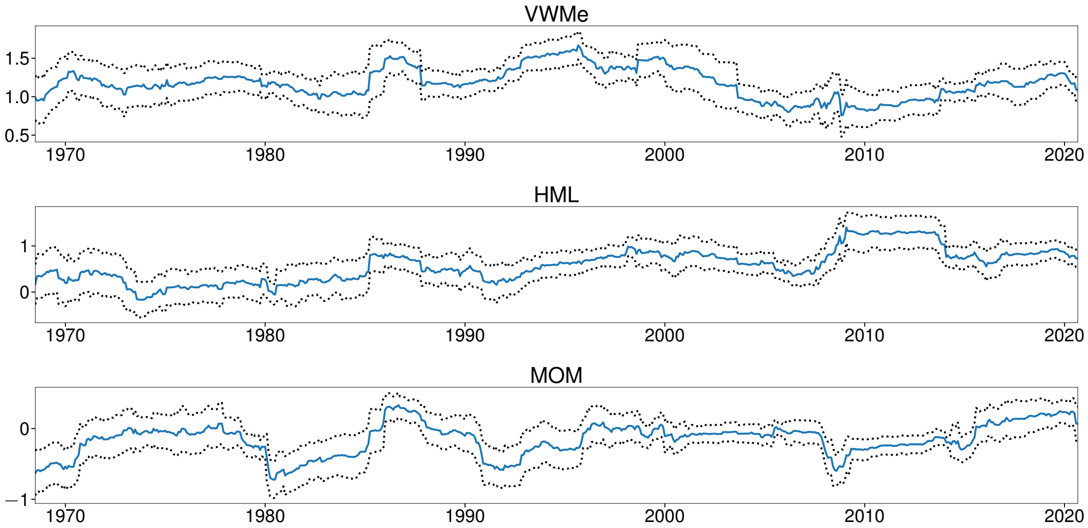
Evidence of Nonlinear Relationship

$$R_i^e = \alpha + \beta_{VWM} VWM^e + \beta_{VWM^2} \left(VWM^e\right)^2 + \beta_{asymVWM^e} VWM^e \times I_{[VWM^e<0]}$$
$$+ \beta_{SMB} SMB + \beta_{HML} HML + \beta_{MOM} MOM + \epsilon_i$$

	$\hat{\beta}$	s.e.$(\hat{\beta})$	Z	$\Pr(Z)$	Conf. Int.
Const.	0.2857	0.225	1.268	0.205	[-0.156, 0.727]
VWM^e	0.4594	0.089	5.154	0.000	[0.285, 0.634]
$(VWM^e)^2$	0.0159	0.007	2.240	0.025	[0.002, 0.030]
$VWM^e \times I_{[VWM^e<0]}$	0.3524	0.188	1.870	0.061	[-0.017, 0.722]
SMB	-0.1972	0.048	-4.087	0.000	[-0.292, -0.103]
HML	0.3470	0.060	5.810	0.000	[0.230, 0.464]
MOM	0.0611	0.039	1.578	0.114	[-0.015, 0.137]

Rolling Window Estimation

60-month window



Outliers

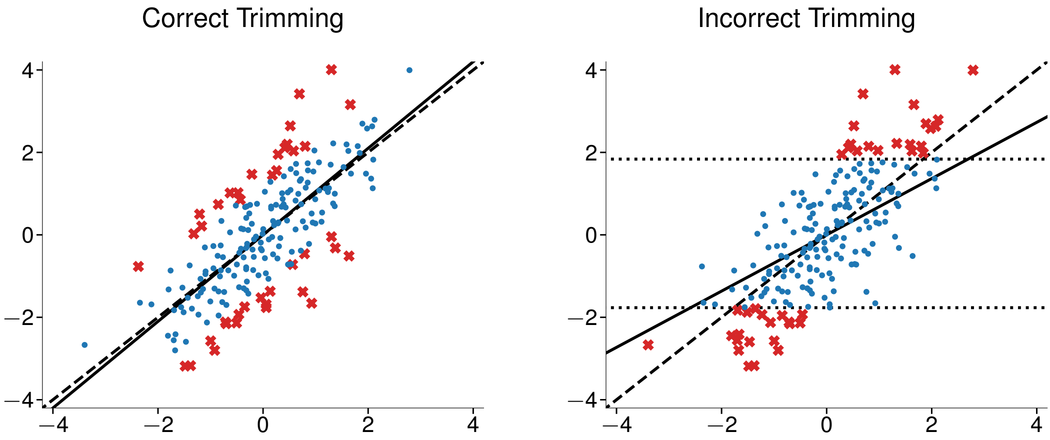
- Outliers happen for a number of reasons
  - Data entry errors
  - Funds “blowing-up”
  - Hyper-inflation
- Often interested in results which are “robust” to some outliers
- Three common options
  - Trimming
  - Windsorization
  - (Iteratively) Reweighted Least Squares (IRWLS)
    - Similar to GLS, only uses functions based on “outlyingness” of error

Trimming

- Trimming involves removing observations
- Removal must be based on values of ϵ_i not Y_i
  - Removal based on Y_i can lead to bias
- Requires initial estimate of $\hat{\boldsymbol{\beta}}$, denoted $\tilde{\boldsymbol{\beta}}$
  - Could include full sample, but sensitive to outliers, especially if extreme
  - Use a subsample that you believe is “good”
  - Choose subsamples at random and use a “typical” value
- Construct residuals $\tilde{\epsilon}_i = Y_i - \mathbf{x}_i\tilde{\boldsymbol{\beta}}$ and delete observations if $\tilde{\epsilon}_i < \hat{q}_\alpha$ or $\tilde{\epsilon}_i > \hat{q}_{1-\alpha}$ for some small α (typically 2.5% or 5%)
  - $\hat{q}_\alpha$ is the α-quantile of the empirical distribution of $\tilde{\epsilon}_i$
- Estimate final $\hat{\boldsymbol{\beta}}$ using OLS on remaining (non-trimmed) data

Correct and Incorrect Trimming

- Removal based on Y_i leads to bias



Windsorization

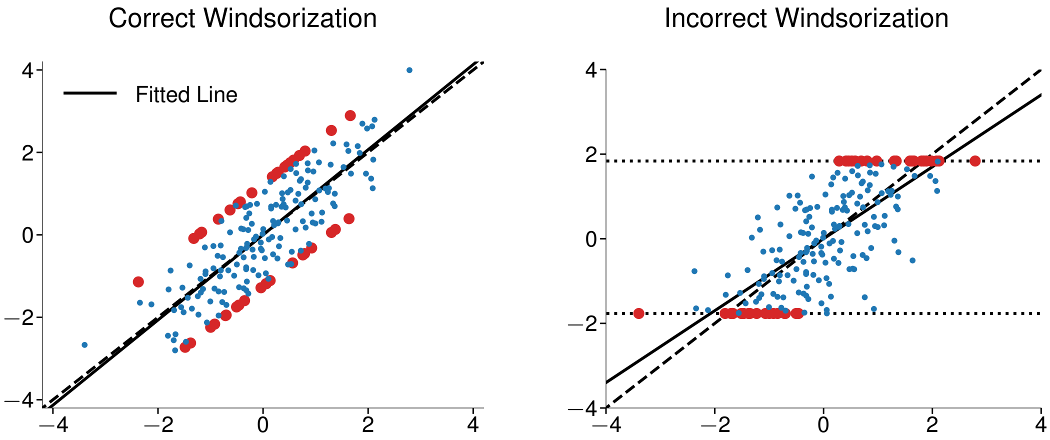
- Windsorization involves replacing outliers with less outlying observations
- Like trimming, removal must be based on values of ϵ_i not Y_i
- Requires initial estimate of $\hat{\boldsymbol{\beta}}$, denoted $\tilde{\boldsymbol{\beta}}$
- Construct residuals $\tilde{\epsilon}_i = Y_i - \mathbf{x}_i\tilde{\boldsymbol{\beta}}$
- Reconstruct Y_i as

$$Y_i = \begin{cases} \mathbf{x}_i\tilde{\boldsymbol{\beta}} + \hat{q}_\alpha & \tilde{\epsilon}_i < \hat{q}_\alpha \\ Y_i & \hat{q}_\alpha \leq \tilde{\epsilon}_i \leq \hat{q}_{1-\alpha} \\ \mathbf{x}_i\tilde{\boldsymbol{\beta}} + \hat{q}_{1-\alpha} & \tilde{\epsilon}_i \geq \hat{q}_{1-\alpha} \end{cases}$$

- Estimate final $\hat{\boldsymbol{\beta}}$ using OLS the reconstructed data

Correct and Incorrect Windsorization

- Removal based on Y_i leads to bias



Rolling and Recursive Regressions

- Parameter stability is often an important concern
- Rolling regressions are an easy method to examine parameter stability

$$\hat{\boldsymbol{\beta}}_i = \left(\sum_{i=j}^{j+m} \mathbf{x}_i'\mathbf{x}_i\right)^{-1}\left(\sum_{i=j}^{j+m} \mathbf{x}_i' Y_i\right), \quad j = 1, 2, \ldots, n-m$$

  - Constructing confidence intervals formally is difficult
  - Approximate method computes full sample covariance matrix, and then scales by n/m to reflect the smaller sample used
  - Similar to building a confidence interval under a null that the parameters are constant
- Recursive regression is defined similarly only using an expanding window

$$\hat{\boldsymbol{\beta}}_i = \left(\sum_{i=1}^{j} \mathbf{x}_i'\mathbf{x}_i\right)^{-1}\left(\sum_{i=1}^{j} \mathbf{x}_i' Y_i\right), \quad j = m, m+1, \ldots, n$$

  - Similar issues with confidence intervals
  - Often hard to observe variation in $\boldsymbol{\beta}$ near the end of the sample if n is large

Machine Learning Approaches:
Best Subset and Stepwise Regression

Regression and Machine Learning

- Many machine learning methods are modifications of regression analysis
  - Best Subset Regression
  - Stepwise Regression
  - Ridge Regression and LASSO
  - Regression Trees and Random Forests
  - Principal Component Regression (PCR) and Partial Least Squares (PLS)
- Key design concerns for ML algorithms:
  - Work well in scenarios where the number of variables available p is large relative to the sample size n
    - $k \leq p$ is the number of variables in a specific model
  - Explicitly make bias-variance trade-off to optimize out-of-sample performance
  - Perform model selection using methods that have been rigorously statistically analyzed

Empirical Application

- ML is well suited to prediction problems
- Focus on tracking portfolio construction
  - Target: VWM^e
  - Assets: 49 Industry portfolios
- Full, unrestricted model

$$VWM_i^e = \sum_{j=1}^{49} \beta_j R_{i,j}^e + \epsilon_i$$

- ML methods differ:
  - Which components of $\boldsymbol{\beta}$ are set to exactly 0
  - How the remaining components are estimated

Best Subset Regression

Selecting the best model from all distinct models

- Consider all 2^p models

Algorithm (Best Subset Regression)

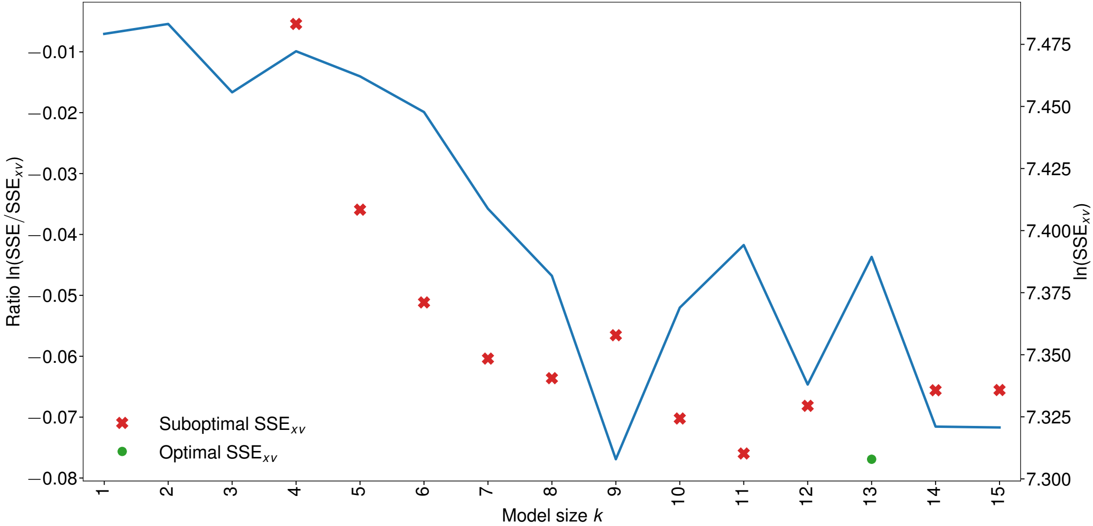
Select the preferred model using:

1. *For each* $k = 0, 1, \ldots, p$ *find the model containing* k *variables that minimizes the* SSE
2. *Select the best model from the* $p + 1$ *models selected in the first step by minimizing a criterion*
   - *Common choices include cross-validated* SSE*, AIC or BIC*
3. *Estimate model parameters of preferred model using OLS*

- In practice only feasible when the number of available variables $p \lessapprox 25$
- Preferred model parameters are still estimated using OLS and so may over fit the in-sample data
- Note: Combinations of reasonable models likely perform the best single model

Best Subset Regression

Randomly selected set of 15 industry portfolios



Forward Stepwise Regression

Approximating Best Subset

- When p is large, Best Subset Regression is infeasible
- Forward Stepwise adds 1 variable at a time to build a sequence of $p+1$ models

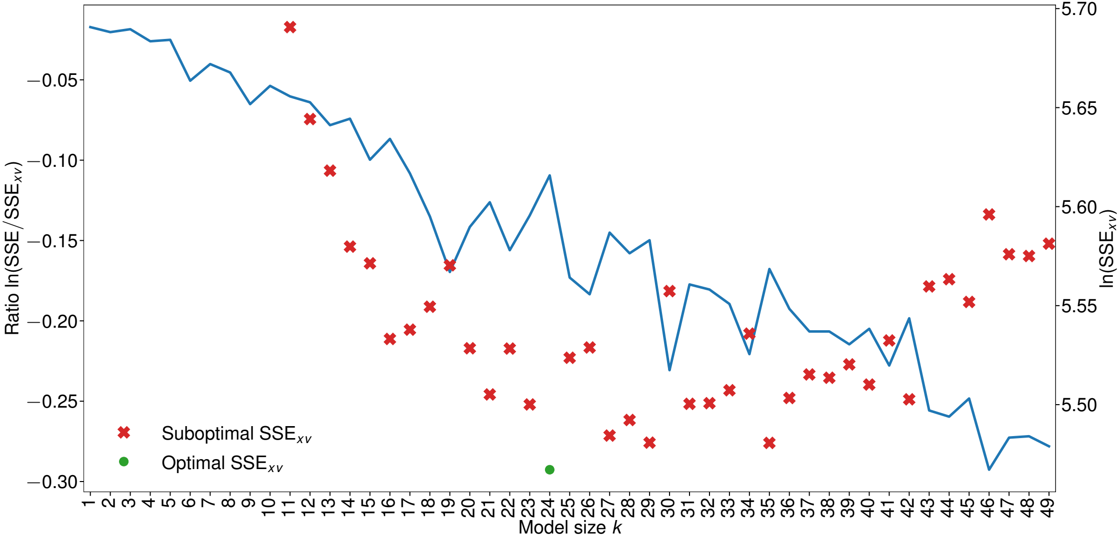
Algorithm (Forward Stepwise Regression)

Select the preferred model using:

1. *Initialize $\mathcal{M}_0$ with only a constant*
2. *For $i = 1, \ldots, p$ estimate all $p - i + 1$ models that add a single variable to model $\mathcal{M}_{i-1}$ and select the model the minimizes the SSE as $\mathcal{M}_i$*
3. *Select the best model from the $p+1$ models selected in the first step by minimizing a criterion*
4. *Estimate model parameters of preferred model using OLS*

- Only requires fitting $O\left(p^2\right)$ models rather than 2^p models
- Path dependence means that it may not find the model as Best Subset Regression

Forward Stepwise Regression



Backward Stepwise Regression

- Backward Stepwise removes 1 variable at a time to build a sequence of $p+1$ models

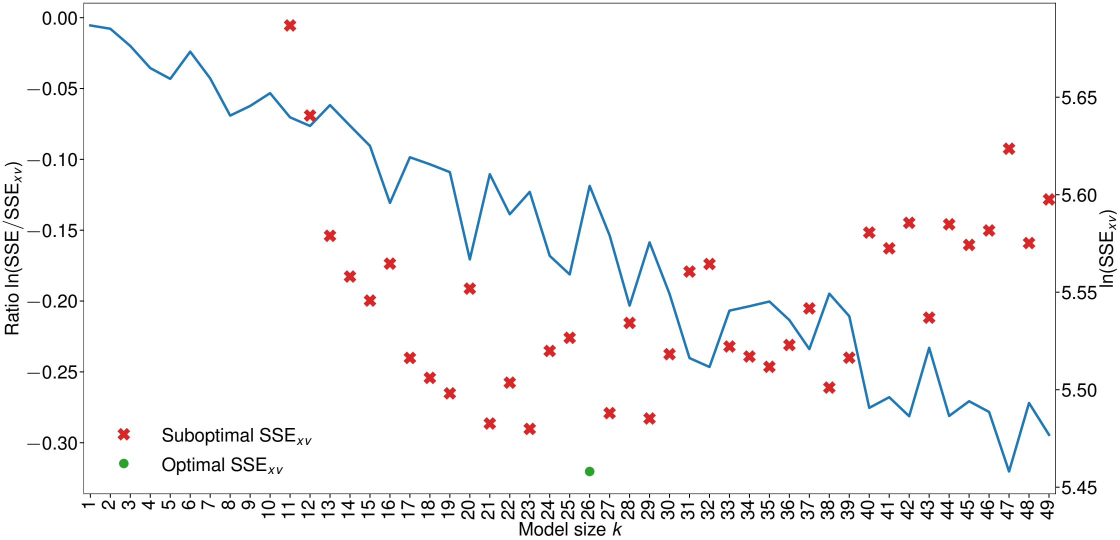
Algorithm (Backward Stepwise Regression)

Select the preferred model using:

1. *Initialize* $\mathcal{M}_p$ *with all variables including a constant*
2. *For* $i = p-1, \ldots, 0$ *estimate all* i *models that remove a single variable from model* $\mathcal{M}_{i+1}$ *and select the model the minimizes the* SSE *as* $\mathcal{M}_i$
3. *Select the best model from the* $p+1$ *models selected in the first step by minimizing a criterion*
4. *Estimate model parameters of preferred model using OLS*

- Same complexity as forward stepwise: $O\left(p^2\right)$
- Generally selects a different model than forward stepwise regression

Backward Stepwise Regression



Hybrid Approaches

Combining Forward and Backward Stepwise Regression

- Forward and backward can be combined to produce alternative collections of candidate models
- Multiple passes may better approximate Best Subset Regression

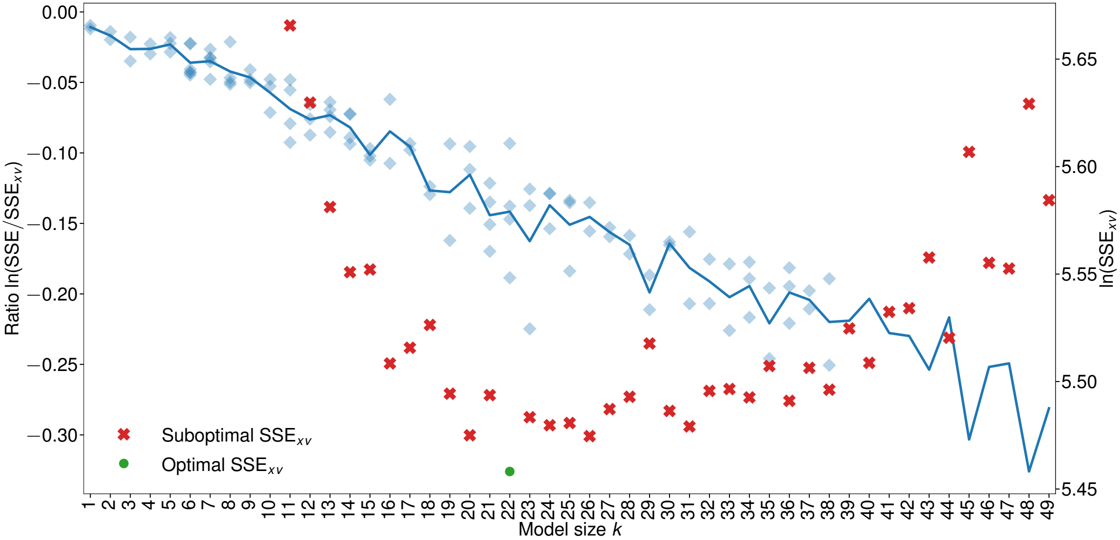
Algorithm (Hybrid Stepwise Selection (2-Level))

Select the preferred model using:

1. *For $k = 3, \ldots, p-2$, use forward select a model with k variables*
2. *Use backward to select $k-1$ candidate models from the k-variable model*
3. *Select the preferred model from all candidate models by minimizing a criterion*
4. *Estimate model parameters of preferred model using OLS*

- Two passes produces a set of $O\left(p^2\right)$ candidate models
- In general m-passes produces a set of $O\left(p^m\right)$ candidate models

Hybrid Stepwise Regression



Machine Learning Approaches: Shrinkage Estimators

Ridge Regression

- Fit a modified least squares problem

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\ (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \text{ subject to } \sum_{j=1}^{k}\beta_j^2 \le \omega.$$

- Equivalent formulation

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\ (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \lambda\sum_{j=1}^{k}\beta_j^2$$

- Analytical solution

$$\hat{\boldsymbol{\beta}}^{\text{Ridge}} = \left(\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}_k\right)^{-1}\mathbf{X}'\mathbf{y}$$

  - Solution is well-defined even if $p > n$
  - In practice complementary to model selection
- *Shrinks* parameters toward 0 when compared to OLS

$$\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}_k > \mathbf{X}'\mathbf{X} \Rightarrow \left(\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}_k\right)^{-1} < \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

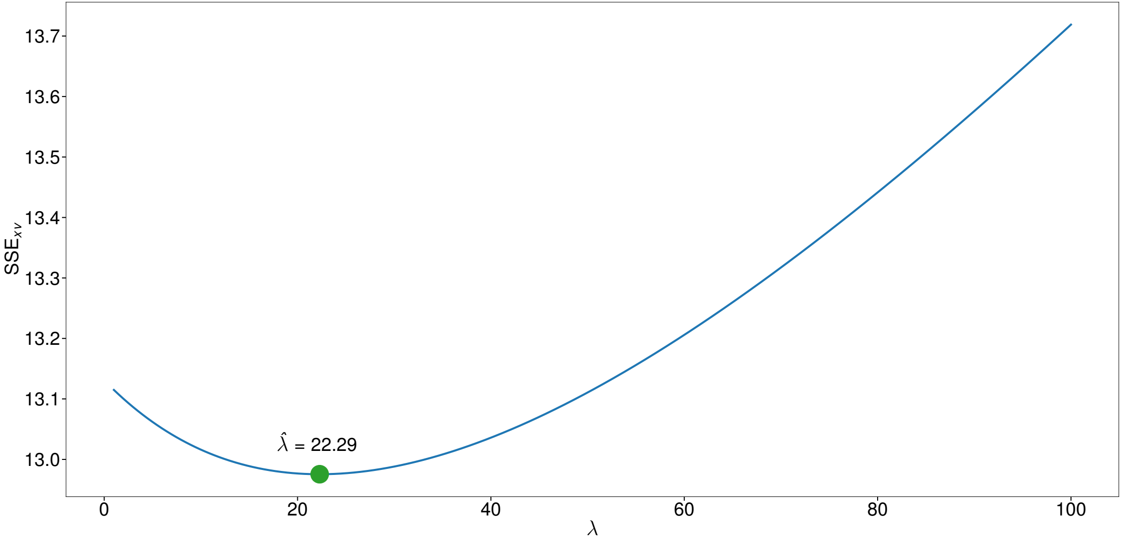
Choosing λ

- λ is a tuning parameter that controls the bias-variance trade-off
- Small λ produces estimates that are similar to OLS and so have only small bias
- Large λ produces estimates with a stronger shrinkage towards 0
  - For any fixed value of λ, as $n \to \infty$ the information in $\mathbf{X}'\mathbf{X}$ dominates the shrinkage $\lambda \mathbf{I}_k$ so that the estimator converges to OLS
- λ is selected by minimizing the cross-validated SSE across a reasonable grid of values $\lambda_1, \ldots, \lambda_m$

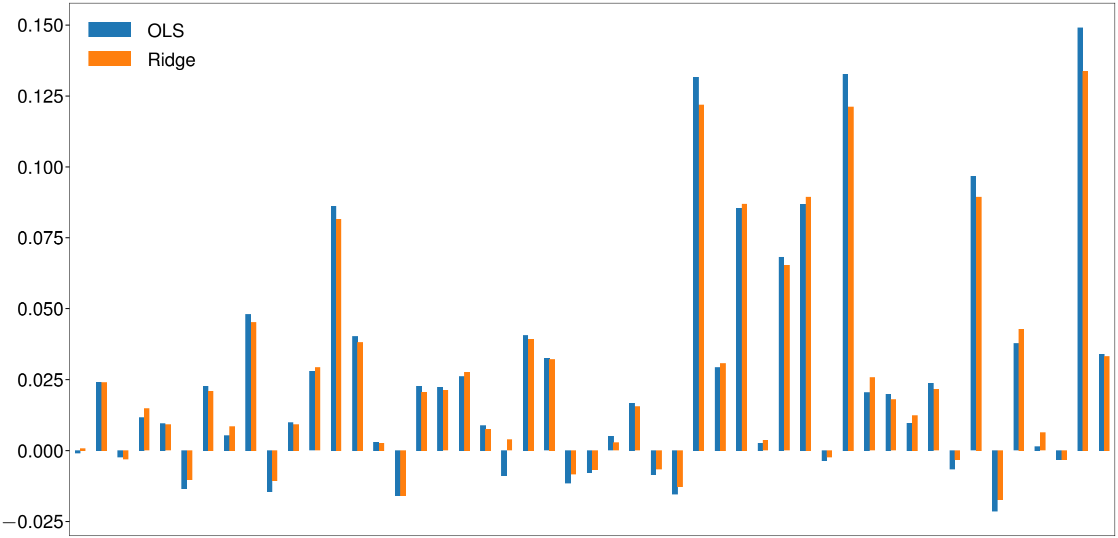
Important: Regressors should be standardized before selecting an optimal λ

$$\tilde{X}_{i,j} = \frac{X_{i,j} - \bar{X}_j}{\hat{\sigma}_j}$$

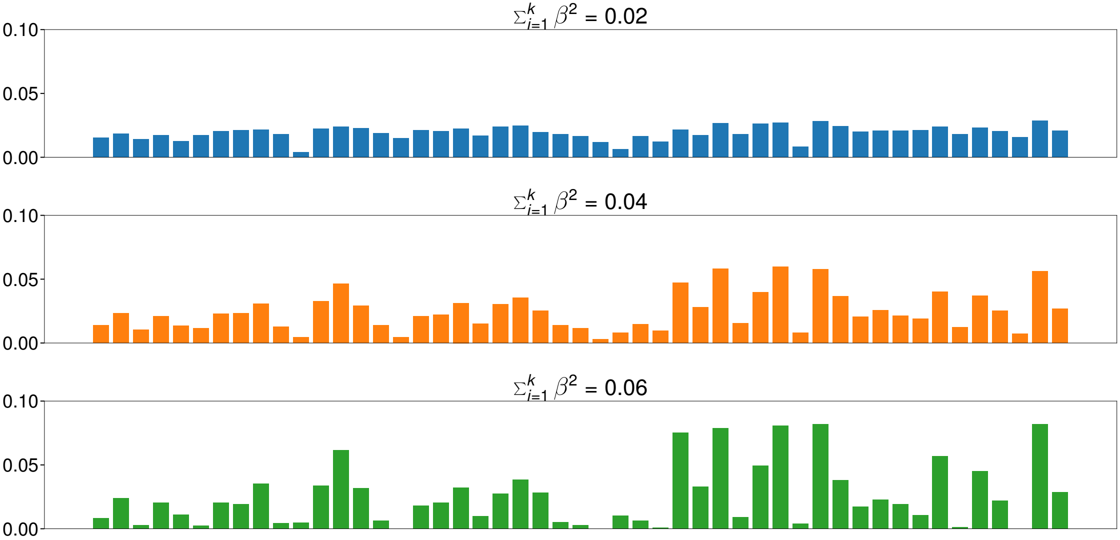
Ridge Regression Penalty Optimization



Ridge Regression



The Effect of Penalization in Ridge Regression



LASSO

Least Absolute Shrinkage and Selection Operator

- LASSO is also defined as a constrained least squares problem

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \ (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \ \text{ subject to } \sum_{j=1}^{k} |\beta_j| < \omega$$

- Equivalent formulation

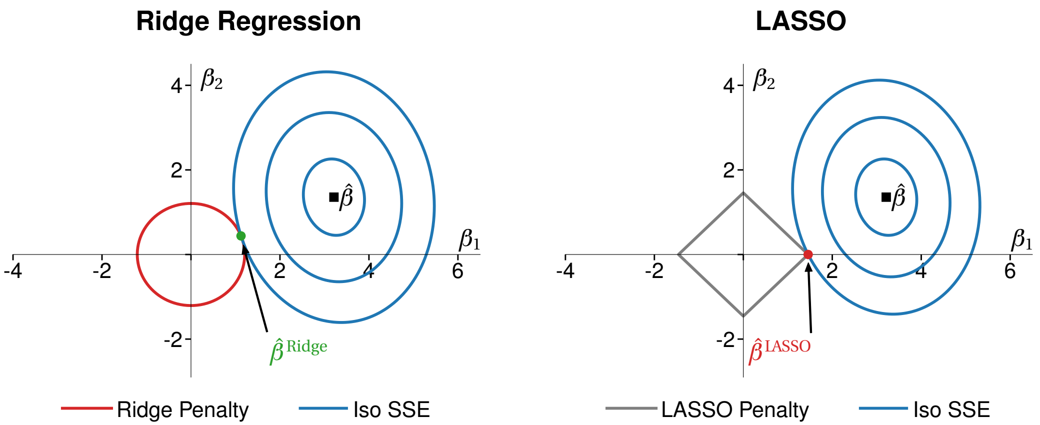
$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \ (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{k} |\beta_j|$$

- Key difference is swap from L_2 (quadratic) penalty to L_1 (absolute value) penalty
- Shape of penalty near $\beta_j \approx 0$ make a large difference
- LASSO tends to estimate coefficients that are exactly 0
  - This is the selection component of LASSO
  - Also shrinks non-zero coefficient
- Ridge does not estimate coefficients to be exactly zero (in general)

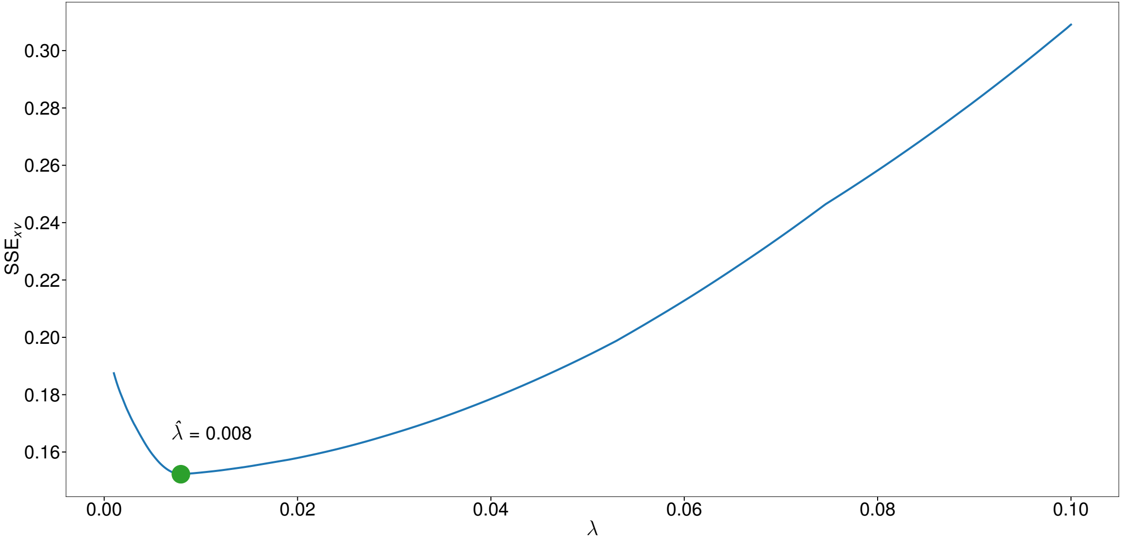
LASSO

- Calibration of λ is identical to calibration in Ridge Regression
- Common to use Post-LASSO parameter estimation
  1. Optimize λ and select variables with non-zero coefficient using LASSO
  2. Exclude variables with 0 coefficient and re-estimate model using OLS
- OLS parameter inference and hypothesis testing is valid in Post-LASSO
- Many variants of LASSO
  - Elastic net: Combine L_1 and L_2 penalties
  - Adaptive LASSO: Consistent Model Selection and Parameter Estimation
  - Group LASSO: Selection across groups of variables rather than individual variables
  - Graphical LASSO: Network estimation
  - Prior LASSO: Selection and shrinkage around a non-zero target

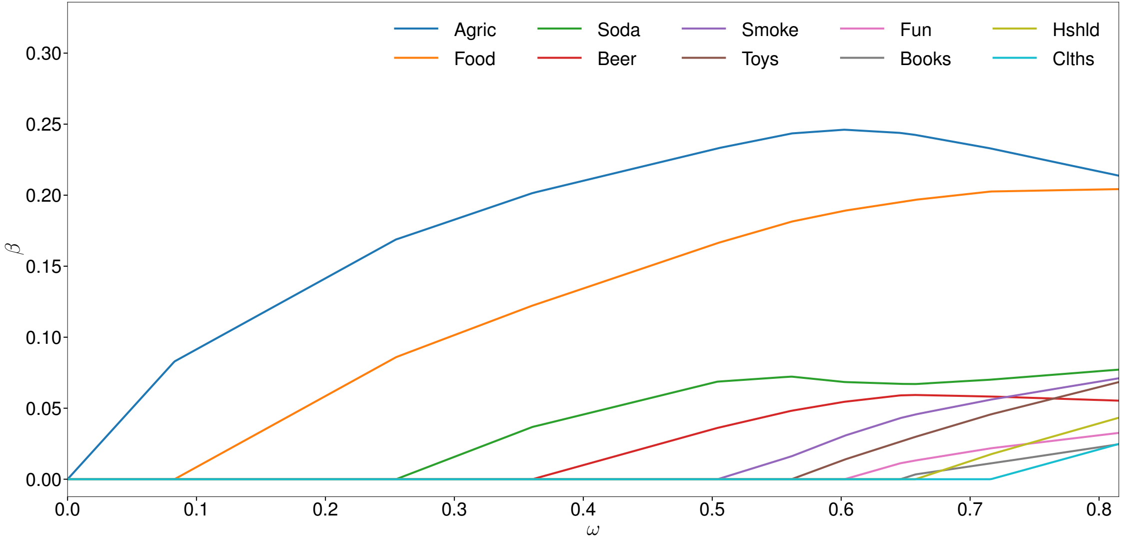
Ridge Regression and LASSO



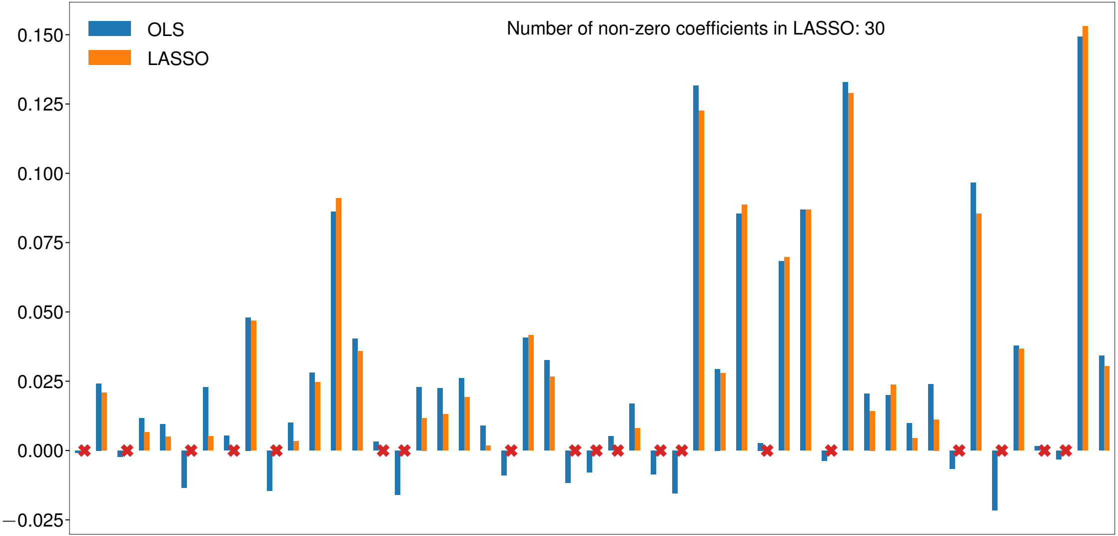
LASSO Regression Penalty Optimization



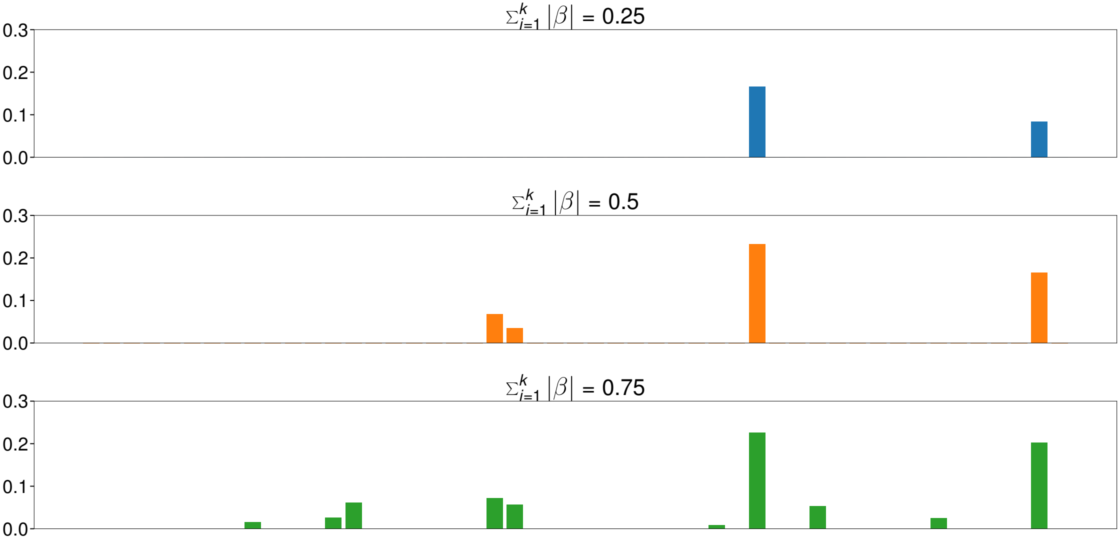
The LASSO Penalty and Path



LASSO Parameter Estimation



The Effect of Penalization in LASSO Regression



Machine Learning Approaches:
Tree Estimators and Random Forests

Regression Trees

- Regression trees built models that rely exclusively on indicator functions.
- A tree is built starting from a root node and splitting the data into two buckets considering all possible splits based on the values of regressors

Algorithm (Regression Tree)

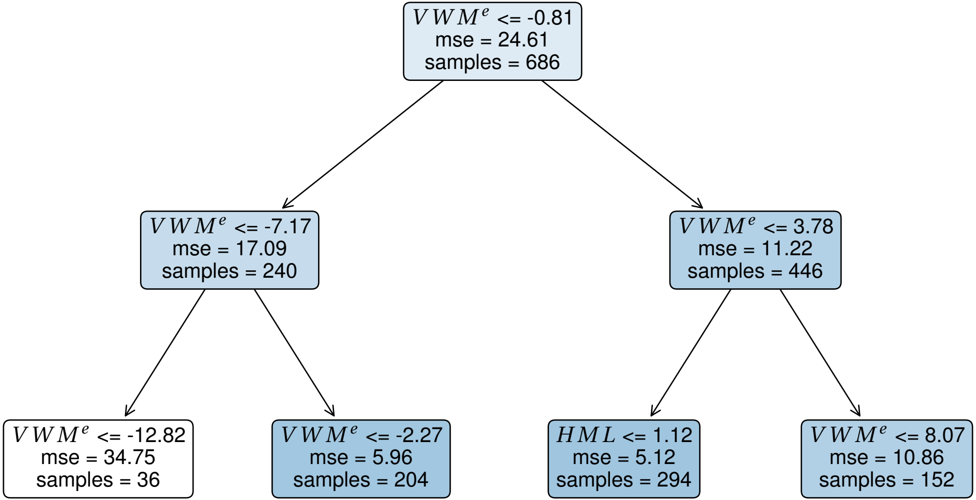
Initialize the tree with a single node (root) that contains all data points. Repeat until the stopping criterion is met:

1. *For each non-terminal node in the tree, compute the split that minimizes the* SSE *by splitting the data by each regressor*
2. *Split the node that shows the largest reduction in* SSE *into two child nodes*

- This process of splitting a node into two leaves continues until a stopping criterion is met:
  - A maximum depth is reached
  - The number of nodes d is reached
  - The number of observations in all terminal nodes falls below some threshold
  - The reduction in SSE for further splits in all terminal nodes falls below some threshold
- The latter two conditions may also stop individual nodes from being further split

Basic Regression Tree Application

- Tree estimated on BH^e using four factors
- Only first three levels visualized

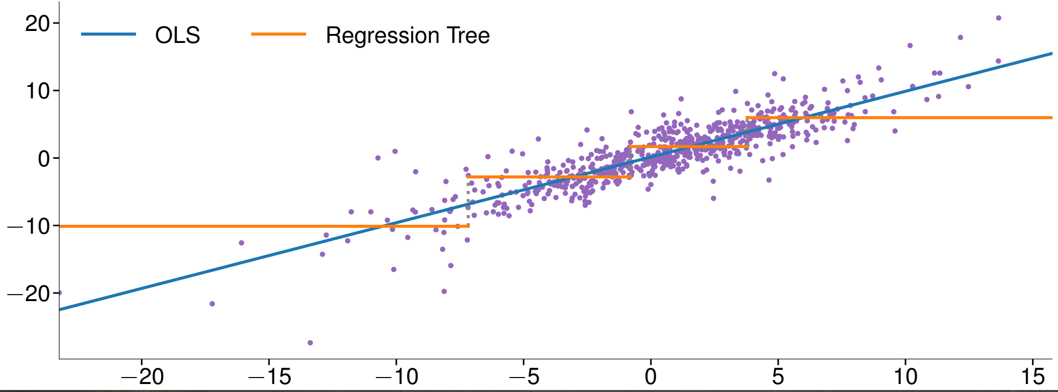


Regression Tree as a Regression

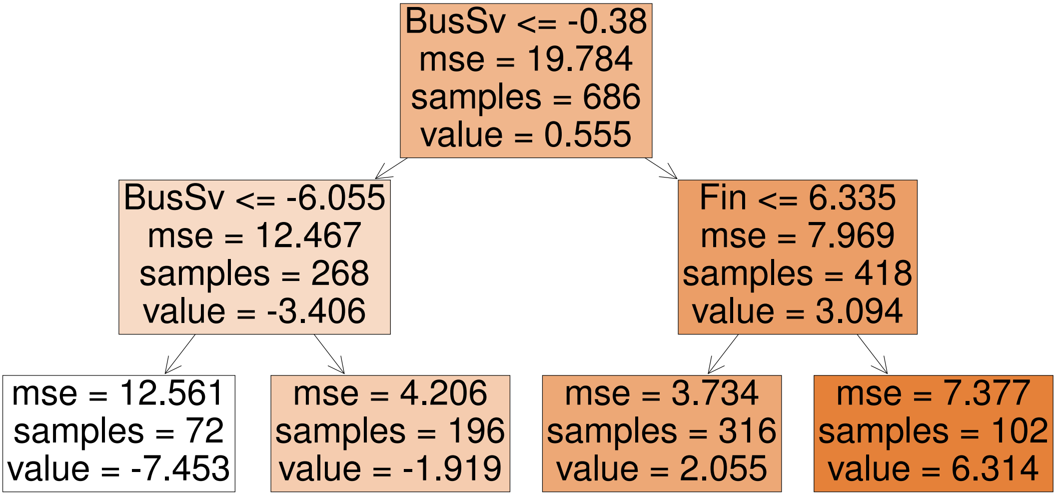
First two levels of BH^e regression tree

- Regression trees build dummy-variable regressions

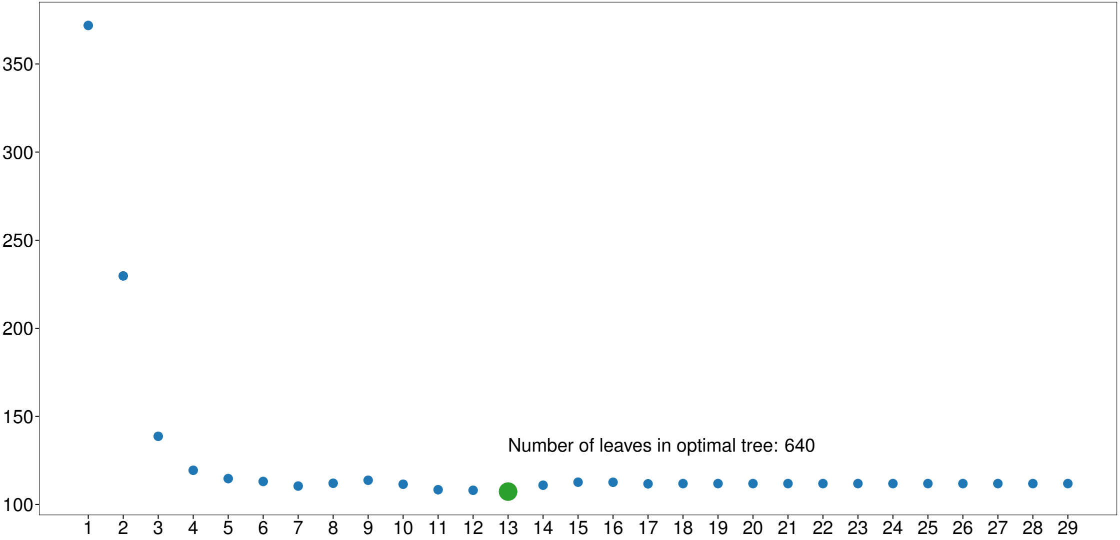
$$BH^e = \beta_1 I_{[VWM_i^e \leq -7.17]} + \beta_2 I_{[-7.17 < VWM_i^e \leq -0.81]} + \beta_3 I_{[-0.81 < VWM_i^e \leq 3.78]} + \beta_4 I_{[VWM^e > 3.78]} + \epsilon_i$$



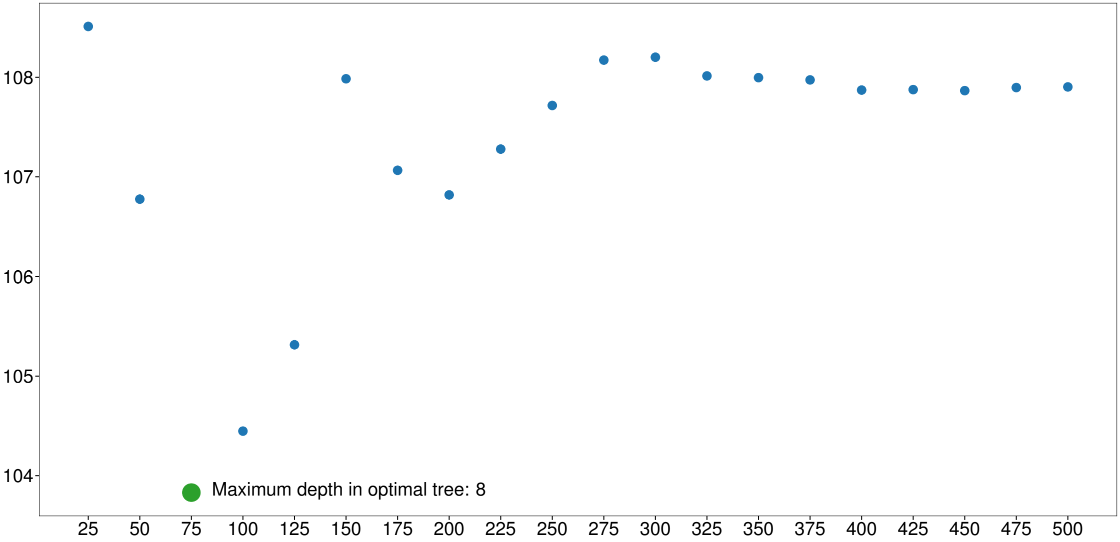
Regression tree visualization



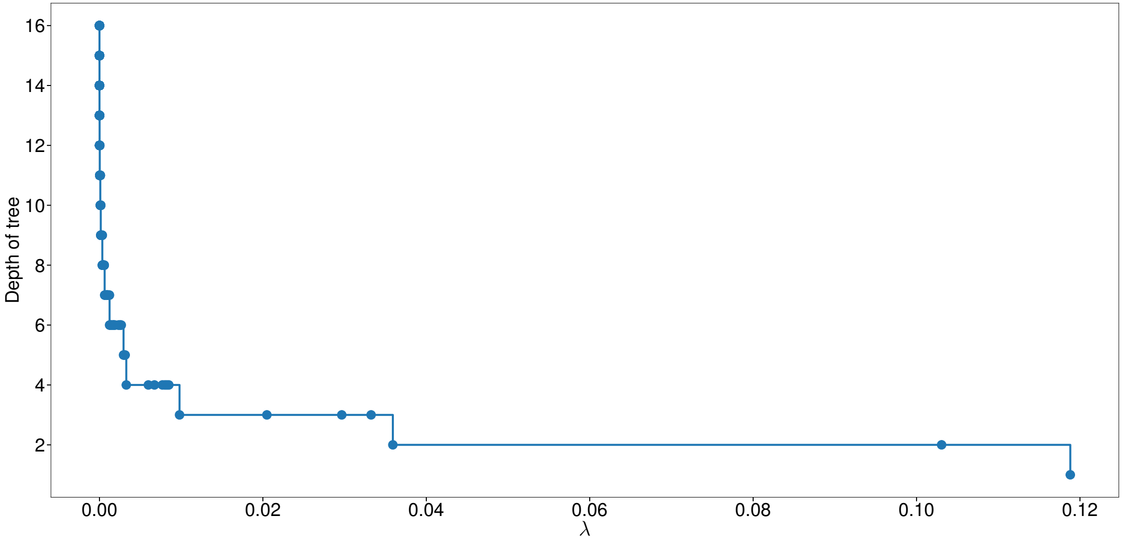
Cross-validating tree depth



Cross-validating tree leaves



Depth as a function of the pruning penalty



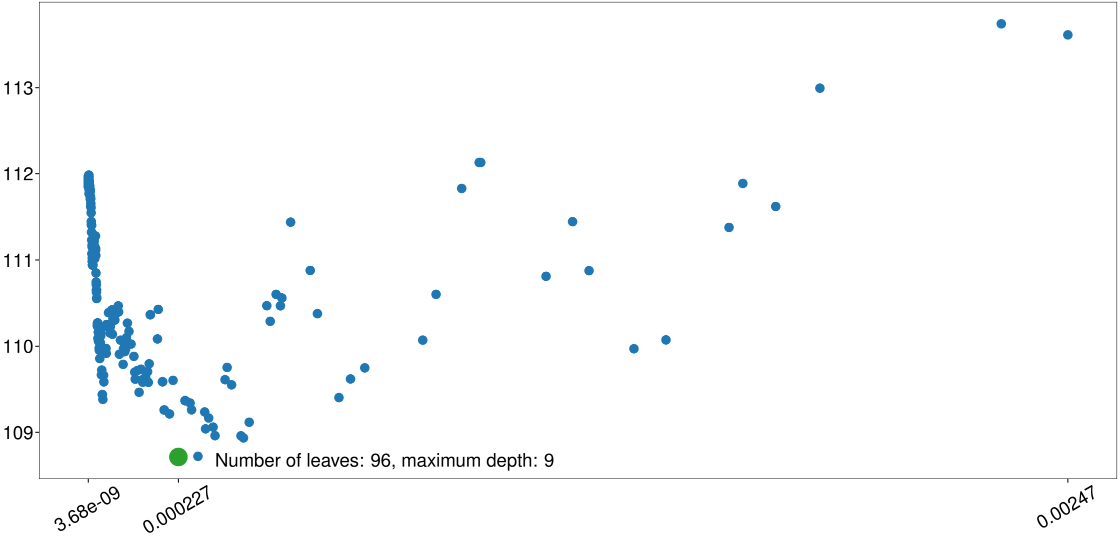
Improvements: Pruning

- Common to prune a tree by recursively removing leaves using a modified objective function

$$\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 + \alpha \left|T\right|$$

  - $\hat{Y}_i$ is the predicted value for a given tree
  - $|T|$ is the number of terminal nodes in the tree
- Pruning starts with a large tree with T_0 nodes that is only terminated when one of the two stopping criteria are satisfied
- For values of α on a grid of plausible values $\{\alpha_1 < \alpha_2 < \ldots < \alpha_q\}$ select the corresponding tree that minimizes the modified objective function
- $\hat{\alpha}$ is selected by computing the best cross-validated fit from the set of q trees
- Using $\hat{\alpha}$ and the original data, estimate the regression tree
- Note: While not required, standardizing Y simplifies the interpretation of α

Selecting the pruning penalty



Improvements: Bagging

Bootstrap Aggregation

- Bagging (**B**ootstrap **AGG**regation) fits trees to B bootstrapped samples
- Each bootstrap sample is used to generate a tree $\hat{f}^{(b)}(\mathbf{x})$
- The bagged predicted value for $\mathbf{x}_i$ is

$$\hat{f}^{\text{bagged}}(\mathbf{x}_i) = B^{-1} \sum_{b=1}^{B} \hat{f}^{(b)}(\mathbf{x}_i)$$

Improvements: Random Forests

Extending Bagging to Reduce Prediction Correlation

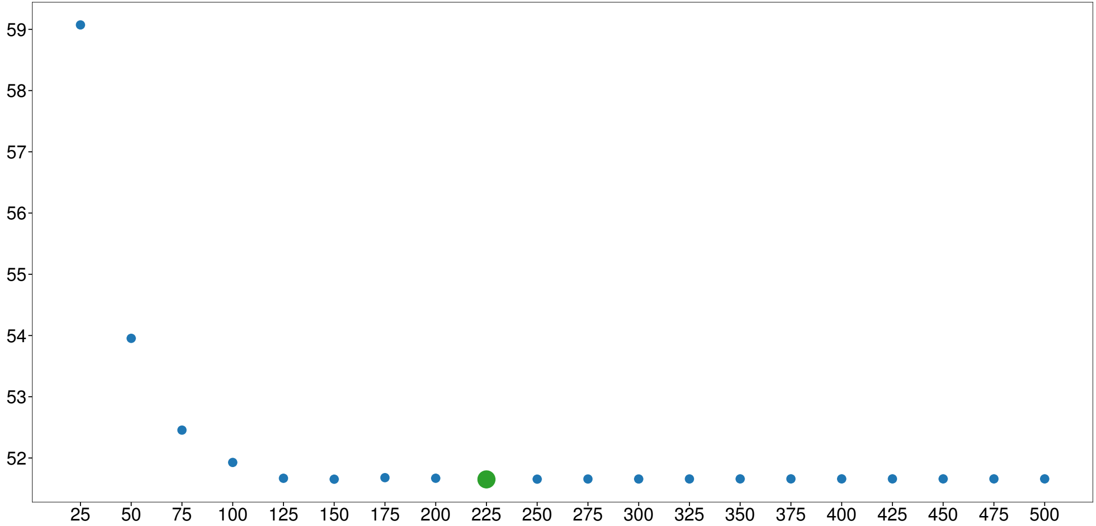
- Random Forests builds B trees using B bootstrapped samples
- Each tree is built using only $k \approx \sqrt{p}$ of the variables
- Produces a set of trees that are weekly correlated because most regressors are excluded from each tree
- Used when two criteria are met
  - p is large
  - A small number of strong predictors
- Predictions are produced using the same method as the bagged forecast

$$\hat{f}^{\text{RF}}\left(\mathbf{x}_i\right) = B^{-1}\sum_{b=1}^{B}\hat{f}^{(b)}\left(\mathbf{x}_i\right)$$

- Bagging is a special case of a Random Forest when $k = p$

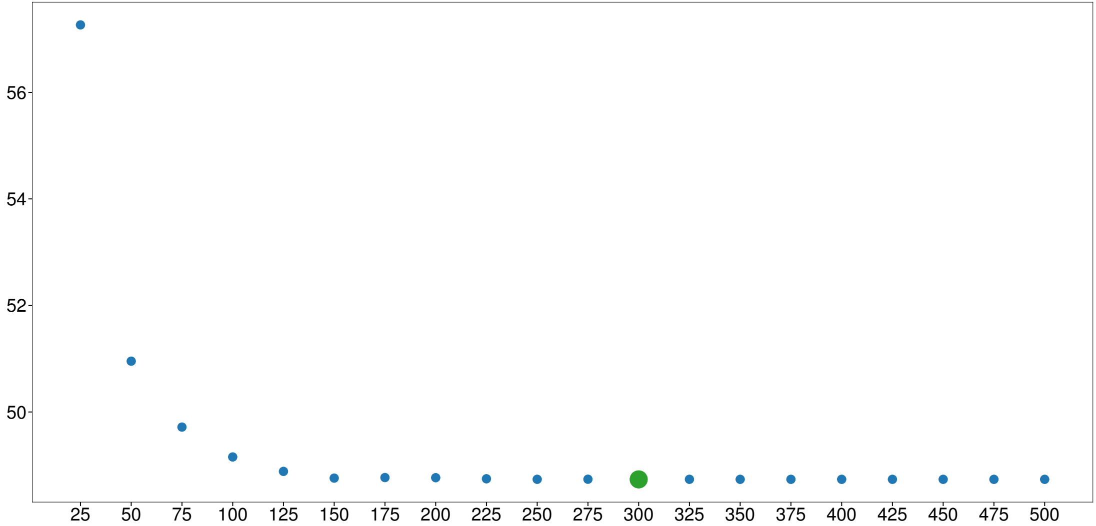
Random Forest Cross Validation

Maximum $\sqrt{m}$ features



Random Forest Cross Validation

Maximum m features



Improvements: Boosting

Focusing Learning on Hard-to-Fit Observations

- Boosting fits a sequence of trees each with d terminal nodes
- Each tree is fit to the *residuals* of the previous tree
  - Child trees focus on fitting observations that were hard to fit by previous trees
  - Nodes are not added for observations that have small prediction errors
- Building a fresh tree collects all observations in to a single leaf
- Allows for models with many low-interaction terms to be built

Algorithm (Boosted Regression Tree)

Compute a boosted regression tree by:

1. *Initialize* $\hat{f}(\mathbf{x}) = 0$ *and* $\epsilon_i^{(0)} = Y_i$
2. *For* $b = 1, \ldots, B$*:*
   a. *Fit a tree with* d *splits and* $d+1$ *terminal nodes to* $\left(\epsilon_i^{(b-1)}, \mathbf{x}_i\right)$
   b. *Update the forecast as* $\hat{f}(\mathbf{x}) = \hat{f}(\mathbf{x}) + \lambda \hat{f}^{(b)}(\mathbf{x})$ *and compute* $\hat{\epsilon}_i^{(b)} = \hat{\epsilon}_i^{(b-1)} - \lambda \hat{f}^{(b)}(\mathbf{x}_i)$

Improvements: Boosting

- Predictions are produced from

$$\hat{f}(\mathbf{x}) = \sum_{b=1}^{B} \lambda \hat{f}^{(b)}(\mathbf{x})$$

- Three tuning parameters
  - $\lambda \in (0, 1]$ is a tuning parameter that shrinks forecasts towards 0
    - In practice $\lambda \in (0.001, 0.2)$
    - Small λ slows learning, and requires large B to fit well
  - d controls the individual tree depth
    - d is the maximum number of interaction terms in the regression model representation
    - Often set to 1 (no interactions)
  - B controls the depth of the tree
- All three parameter interact and serve as substitutes
  - Increase one, decrease the others to maintain approximately constant fit
- Note: Data should be standardized when using boosting

Cross Validation optimization of Boosting