Model Selection

Hui Chen

MIT Sloan

15.450, Spring 2023

Motivation: The Factor Zoo

- In finance, we have spent a long time looking for a model to explain stock returns with a small number of factors.
- The problem is that new "anomalies" keep popping up, adding to the factor zoo:

$$\text{CAPM: } R^e_{it} = \alpha_i + \beta_{i,m}\, R^e_{mt} + \varepsilon_{it}$$

$$\text{FF3: } R^e_{it} = \alpha_i + \beta_{i,m}\, R^e_{mt} + \beta_{i,hml}\, HML_t + \beta_{i,smb}\, SMB_t + \varepsilon_{it}$$

$$\text{4-factor: } R^e_{it} = \alpha_i + \beta_{i,m}\, R^e_{mt} + \beta_{i,hml}\, HML_t + \beta_{i,smb}\, SMB_t + \beta_{i,wml}\, WML_t + \varepsilon_{it}$$

$$\text{5-factor: } R^e_{it} = \alpha_i + \beta_{i,m}\, R^e_{mt} + \beta_{i,hml}\, HML_t + \beta_{i,smb}\, SMB_t + \beta_{i,rmw}\, RMW_t + \beta_{i,cma}\, CMA_t + \varepsilon_{it}$$

- Which is the "right" model?
- Is there any hope of finding a "simple" model?

Replicating Anomalies

Kewei Hou
The Ohio State University and China Academy of Financial Research

Chen Xue
University of Cincinnati

Lu Zhang
The Ohio State University and National Bureau of Economic Research

Most anomalies fail to hold up to currently acceptable standards for empirical finance. With microcaps mitigated via NYSE breakpoints and value-weighted returns, 65% of the 452 anomalies in our extensive data library, including 96% of the trading frictions category, cannot clear the *single* test hurdle of the absolute t-value of 1.96. Imposing the higher multiple test hurdle of 2.78 at the 5% significance level raises the failure rate to 82%. Even for replicated anomalies, their economic magnitudes are much smaller than originally reported. In all, capital markets are more efficient than previously recognized. (*JEL* C58, G12, G14, G17, M41)

Outline

Model Selection

- True model:

$$y_i = f(\mathbf{x}_i) + \varepsilon_i$$

  where $\mathbf{x}_i$ is a $p \times 1$ vector, $\mathrm{E}[\varepsilon_i] = 0$, $\mathrm{Var}[\varepsilon_i] = \sigma^2$, and $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$.

- We try to find a "good" fitted model $\widehat{f}(\mathbf{x})$ to predict y.
- However, not knowing the true model, model selection requires:
  1. Feature selection: Which variables to include in $\mathbf{x}$?
  2. Functional form selection: Which class of $\widehat{f}$ to use?
- Feature selection will be the primary concern if we only consider linear models, or if we use a class of $\widehat{f}$ that is flexible.
  - ↪ Q: With p potential features, how many possible linear models?
- How to choose among the candidate models?

Model Evaluation

- We can measure the accuracy of the fitted model over the sample $(y_1, \mathbf{x}_1), \cdots, (y_n, \mathbf{x}_n)$ ("training set") using:

$$MSE_{Tr} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \widehat{f}(\mathbf{x}_i)\right)^2$$

- However, by now we know it is a bad idea to evaluate/select the fitted model based on MSE_{Tr}, which would lead to over-fitting.
- We should instead compute the MSE over a new set of data $(\tilde{y}_1, \tilde{\mathbf{x}}_1), \cdots, (\tilde{y}_m, \tilde{\mathbf{x}}_m)$ ("test set").

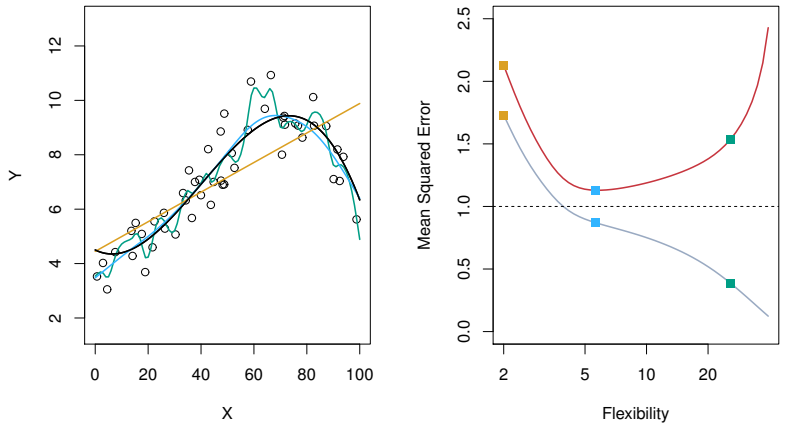
$$MSE_{Te} = \frac{1}{m} \sum_{i=1}^{m} \left(\tilde{y}_i - \widehat{f}(\tilde{\mathbf{x}}_i)\right)^2$$

- In the case of classification, use classification error instead of MSE:

$$CE = \frac{1}{n} \sum_{i=1}^{n} I(y_i \neq \widehat{y}_i)$$

Example: Training Error vs. Test Error

Source: ISL Figure 2.9



Left: Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves). Right: Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line).

Bias and Variance

- What do we expect the MSE to be over new data $(y_0, \mathbf{x}_0)$?

$$MSE(\widehat{f}(\mathbf{x}_0)) = \mathrm{E}\left[\left(y_0 - \widehat{f}(\mathbf{x}_0)\right)^2 | \mathbf{x}_0\right]$$

$$= \underbrace{\sigma^2}_{\text{Irreducible}} + \underbrace{\left(f(\mathbf{x}_0) - \mathrm{E}[\widehat{f}(\mathbf{x}_0)]\right)^2}_{\textbf{Bias}} + \underbrace{\mathrm{E}\left[(\mathrm{E}[\widehat{f}(\mathbf{x}_0)] - \widehat{f}(\mathbf{x}_0))^2 | \mathbf{x}_0\right]}_{\textbf{Variance}}$$

- Q: Conditional on $\mathbf{x}_0$, what is random in the above equation?

Sources of prediction error

- Sources of prediction error:
  1. σ^2: irreducible error
  2. Bias: error due to using the wrong model (f vs. $\widehat{f}$)
  3. Variance: impact of the randomness of training sample on the fitted model $\widehat{f}$

- Bias-variance trade-off:
  - ↪ Bias is not necessarily a bad thing. Having a bigger bias in exchange for smaller variance might be good for overall prediction accuracy.
  - ↪ Example: OLS = BLUE. But there exists other (biased) linear estimators that produces smaller MSE.

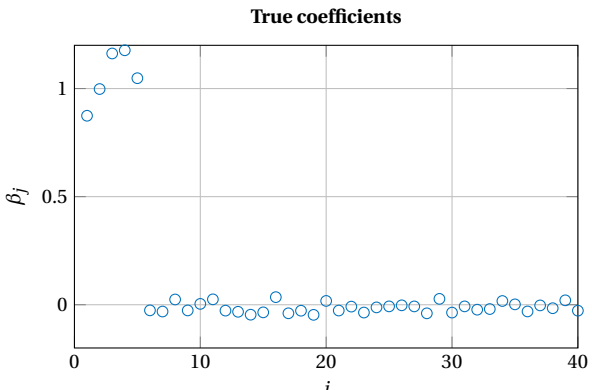
Q: When would unbiasedness be important?

Example: The Factor Zoo

- Suppose standardized returns are generated from a linear model (true model):

$$r_i = \mathbf{x}_i'\beta + \varepsilon_i$$

- Factors $\mathbf{x}_i$ are standardized: $E[\mathbf{x}_i] = 0, \text{Var}[x_i] = 1$.
- $\text{E}[\varepsilon_i] = 0, \text{Var}[\varepsilon_i] = \sigma^2 = 1$.
- Size of training sample is small: $n = 50, p = 40$



Example: The Factor Zoo

1. Use the true β to simulate a sample of $(\mathbf{x}_i, r_i)$ with size $n = 50$ (training set).
2. Estimate $\widehat{\beta}$ using this training set.
3. Simulate a new sample of size $m = 1$ (test set).
4. Compute the prediction error based on $\widehat{\beta}$ over the test set.
5. Repeat Steps 1-4 for $R = 10,000$ times. Compute the bias, variance, and MSE_{Te} across the simulations.

	σ^2	$Bias(\widehat{f}(\mathbf{x}_0))^2$	$\text{Var}(\widehat{f}(\mathbf{x}_0))$	MSE_{Te}
OLS	1	0.0001	0.8030	1.8031
Ridge	1	0.1074	0.5539	1.6613

How to trade off bias against variance?

- Indirect methods: Make (theoretical) adjustment to the training error by penalizing model flexibility.
  - ↪ Adjusted R^2 (for linear models):
  $$\overline{R}^2 = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)}$$
  - ↪ AIC:
  $$\frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$$
  - ↪ BIC:
  $$\frac{1}{n\hat{\sigma}^2}(RSS + \log(n)d\hat{\sigma}^2)$$

- Direct methods: Estimate MSE_{Te} directly via validation or cross-validation.

Outline

Validation

- What is a good model? As discussed earlier, a reasonable criterion is the ability to predict accurately with new data.
- For now, assume our data are IID.

Validation

1. Randomly split the data sample into two, a training set and a validation (or hold-out) set.
2. Fit the model $\widehat{f}(\cdot)$ using the training set.
3. Use the fitted model to make predictions on the validation set and compute the prediction errors (e.g., using MSE).

- This is inefficient: We are not fully utilizing all the information available to determine $\widehat{f}(\cdot)$.
- When the sample is small, both fitting and testing become unreliable.

$K-$fold cross-validation

1. Partition the sample of size n randomly into K separate sets (typically with equal size).
   - ↪ Let F_k denote the set of observations in the kth subset.
2. For each $k = 1, 2, \cdots, K$, fit the model $\widehat{f}_{-k}(\cdot)$ to the full training set excluding the kth subset.
3. Compute the total cross-validation error for the kth subset:

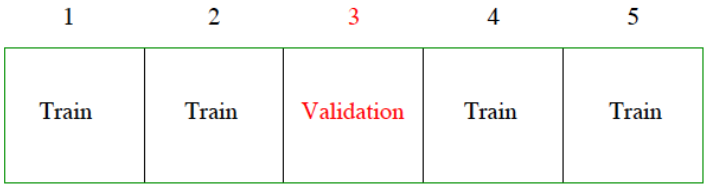
$$TE_k = \sum_{i \in F_k} (y_i - \widehat{f}_{-k}(\mathbf{x}_i))^2$$

4. Compute the average cross-validation error over all subsets:

$$CVE = \frac{1}{n} \sum_{k=1}^{K} TE_k = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in F_k} (y_i - \widehat{f}_{-k}(\mathbf{x}_i))^2$$

5. Repeat Steps 2-4 for each candidate model. We then select the model that minimizes CVE.

$K-$fold cross validation: Illustration



- Typical choices for K: 5 or 10
- If $K = n$: "Leave-one-out" cross validation (LOOCV).

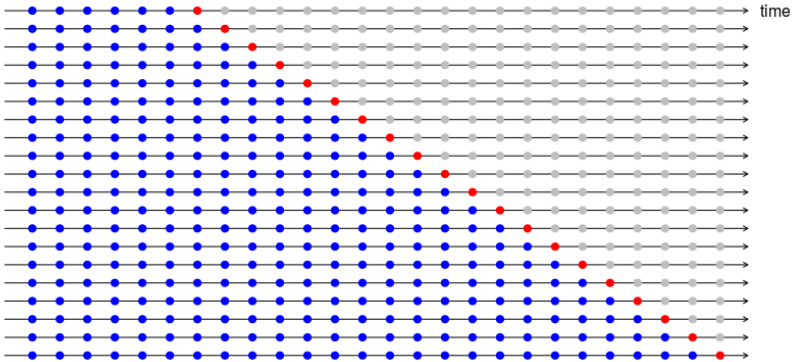
Cross validation for time series data

- In time series, need to avoid using future information to construct the forecast.
- For validation, we can use the early portion of the sample as training sample, the remainder as the test sample.
- For cross validation, one possibility is to gradually expand the window for the training set, with the following observation as the test set.

Cross validation for time series data

1. Choose the minimum size of training set k.
2. For $i = 1, 2, \cdots, T-k$, select the observation at $t = k+i$ as the test set. Use the observations from $t = 1$ to $t = k+i-1$ as the training sample.
3. Fit the model on the training sample and compute the prediction error for observation at $t = k+i$.
4. Repeat Steps 2-3, and compute the overall average cross-validation error.

Cross validation for time series data



Outline

Best subset selection

- A naive way to conduct feature selection for linear models is to simply try all the possible models.
- The algorithm is straightforward to implement. But the number of potential models can be huge (2^p), and it could lead to severe over-fitting.

Best subset selection

1. Start with the null model, $\mathcal{M}_0$, which contains no predictors.
2. For $k = 1, 2, \cdots, p$, fit all $\binom{p}{k}$ models that contain k predictors. Pick the best model among them, $\mathcal{M}_k$, with the smallest RSS or largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \cdots, \mathcal{M}_p$ using cross-validation, AIC, BIC, adjusted R^2 ...

Forward and backward stepwise selection

Forward stepwise selection

1. Start with the null model, $\mathcal{M}_0$, which contains no predictors.
2. For $k = 0, \cdots, p-1$, consider all $p-k$ models, each augments the predictors in $\mathcal{M}_k$ with one additional predictor. Choose the best among the $p-k$ models, based on RSS or R^2.
3. Select a single best model from among $\mathcal{M}_0, \cdots, \mathcal{M}_p$ using cross-validation, AIC, BIC, adjusted R^2 ...

- The benefit: Computationally more manageable than best subset selection. The number of models is on the order of p^2.
- The cost: It will miss some candidate models and so does not guarantee to yield the best model.
- There is also a backward step-wise selection.

Example: Stepwise selection for the credit model

Source: ISL, Chapter 6

# Variables	Best subset	Forward stepwise
One	`rating`	`rating`
Two	`rating`, `income`	`rating`, `income`
Three	`rating`, `income`, `student`	`rating`, `income`, `student`
Four	`cards`, `income`, `student`, `limit`	`rating`, `income`, `student`, `limit`

TABLE 6.1. *The first four selected models for best subset selection and forward stepwise selection on the* `Credit` *data set. The first three models are identical but the fourth models differ.*

Example: Stepwise selection for the credit model

Source: ISL, Chapter 6

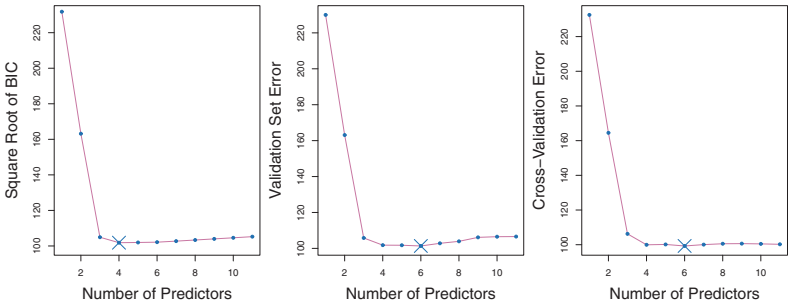


FIGURE 6.3. *For the* `Credit` *data set, three quantities are displayed for the best model containing d predictors, for d ranging from* 1 *to* 11. *The overall* best *model, based on each of these quantities, is shown as a blue cross.* Left: *Square root of BIC.* Center: *Validation set errors.* Right: *Cross-validation errors.*

Outline

Ridge Regression

- Control variance $\text{Var}(\mathbf{x}_0'\widehat{\beta})$ by regularizing the coefficients.
- Let y_i be centered (mean 0) and $\mathbf{x}_i$ be standardized (mean 0, unit variance).

$$\underbrace{\min_{\beta} \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\beta)^2}_{OLS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

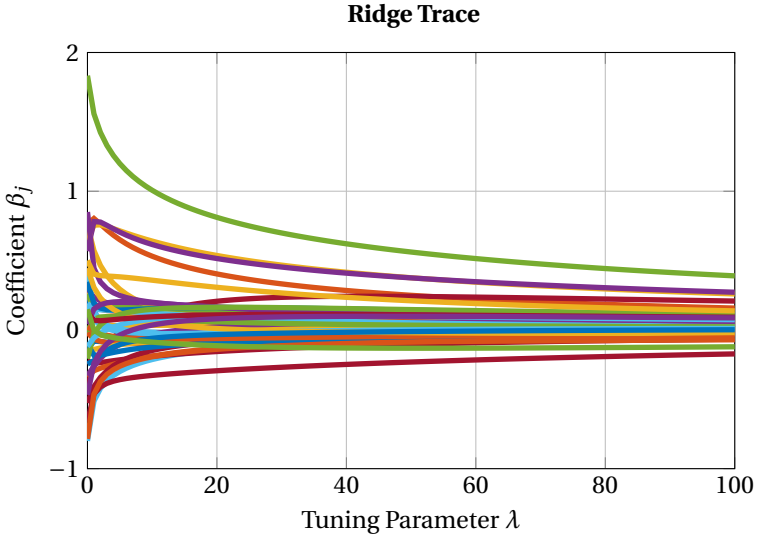
- The new part is called an ℓ_2 penalty.
- Solution:

$$\widehat{\beta}_{\lambda}^{ridge} = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{y}$$

- λ – tuning parameter
  - ↪ Generally, bias increases while variance decreases as λ increases.
  - ↪ $\lambda \to 0$: $\widehat{\beta}_{\lambda}^{ridge} \to \widehat{\beta}^{OLS}$
  - ↪ $\lambda \to \infty$: $\widehat{\beta}_{\lambda}^{ridge} \to 0$
  - ↪ How to choose λ? Methods like AIC or BIC won't work. (Q: Why?) Cross-validation (more later).

Q: Is $\widehat{\beta}_{\lambda}^{ridge}$ biased?

Example: Predicting returns



The LASSO

- The ℓ_2-penalty in Ridge regression reduces the size of $\widehat{\beta}$, but does not drive its elements to zero. Thus, it does not help with variable (model) selection.
  - ↪ Particularly useful if $p \gg n$.
- LASSO (least absolute shrinkage and selection operator): ℓ_1 penalty.
- With centered y_i (mean 0) and standardized $\mathbf{x}_i$ (mean 0, unit variance),

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

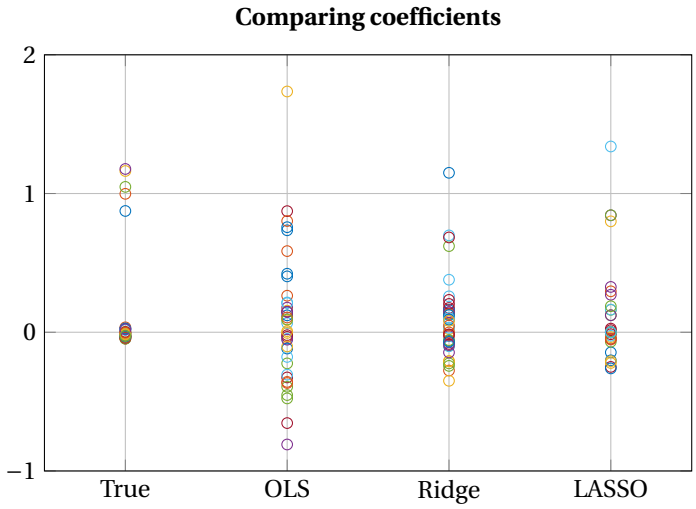
- Compare ℓ_1 and ℓ_2 penalty.
- No closed-form solution for $\widehat{\beta}_{\lambda}^{lasso}$.
  - ↪ Use the *lars* or *glmnet* package in R.
  - ↪ How to choose λ? Cross-validation.

LASSO vs. Ridge

- LASSO and ridge regression generally perform comparably in terms of prediction error.
- LASSO has advantage in doing variable selection by setting some of the coefficients to zero.

Example: Fitted coefficients from OLS, Ridge, LASSO

From the factor zoo example



Choosing λ using $K-$fold cross-validation

1. Partition the training data n into K separate sets of equal size.
2. For each $k = 1, 2, \cdots, K$, fit the model $\widehat{f}^{\lambda}_{-k}(\cdot)$ to the full training set excluding the kth subset, under a particular value of λ.
3. Compute the fitted values and the total cross-validation error for the kth subset:

$$e_k(\lambda) = \sum_{i \in F_k} (y_i - \widehat{f}^{\lambda}_{-k}(\mathbf{x}_i))^2$$

4. Compute the average cross-validation error over all subsets:

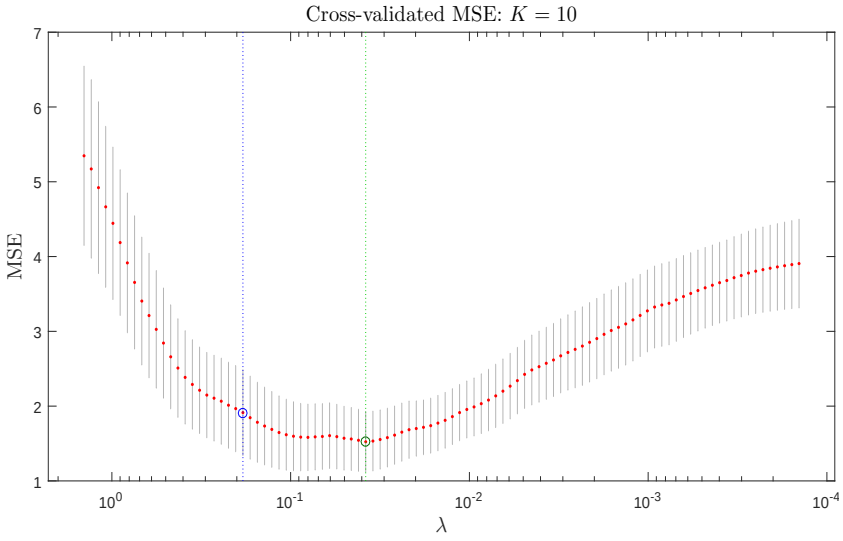
$$CVE(\lambda) = \frac{1}{n} \sum_{k=1}^{K} e_k(\lambda) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in F_k} (y_i - \widehat{f}^{\lambda}_{-k}(\mathbf{x}_i))^2$$

5. Optimal tuning parameter minimizes the average cross-validation error

$$\lambda^* = \arg\min_{\lambda} CVE(\lambda)$$

6. Refit the model with λ^* on the entire training set.

Example: $K-$fold cross validation



Summary and Readings

- The Bias-Variance tradeoff.
- Cross-validation for IID and time series.
- Subset selection: Feature selection for linear models.
- Shrinkage methods aim to reduce variance at the expense of a bigger bias.
  - ↪ LASSO has the additional advantage in doing feature selection.
  - ↪ The bias also hurts the interpretability of the model.
- Reading: ISL Chapter 5.1, 6.1-6.2