Using Statistical Learning for Model Specification Selection

Steven Sawyer
Economist
Producer Price Index
FedCASIC
4/16/2019

Background

- Statistical learning
- Hedonic models
- Measures of model performance
- Validation
- Cross validation

Statistical Learning

- Statistical analysis with algorithms
- Understand and use data
- Inference and prediction
- Important to understand algorithms

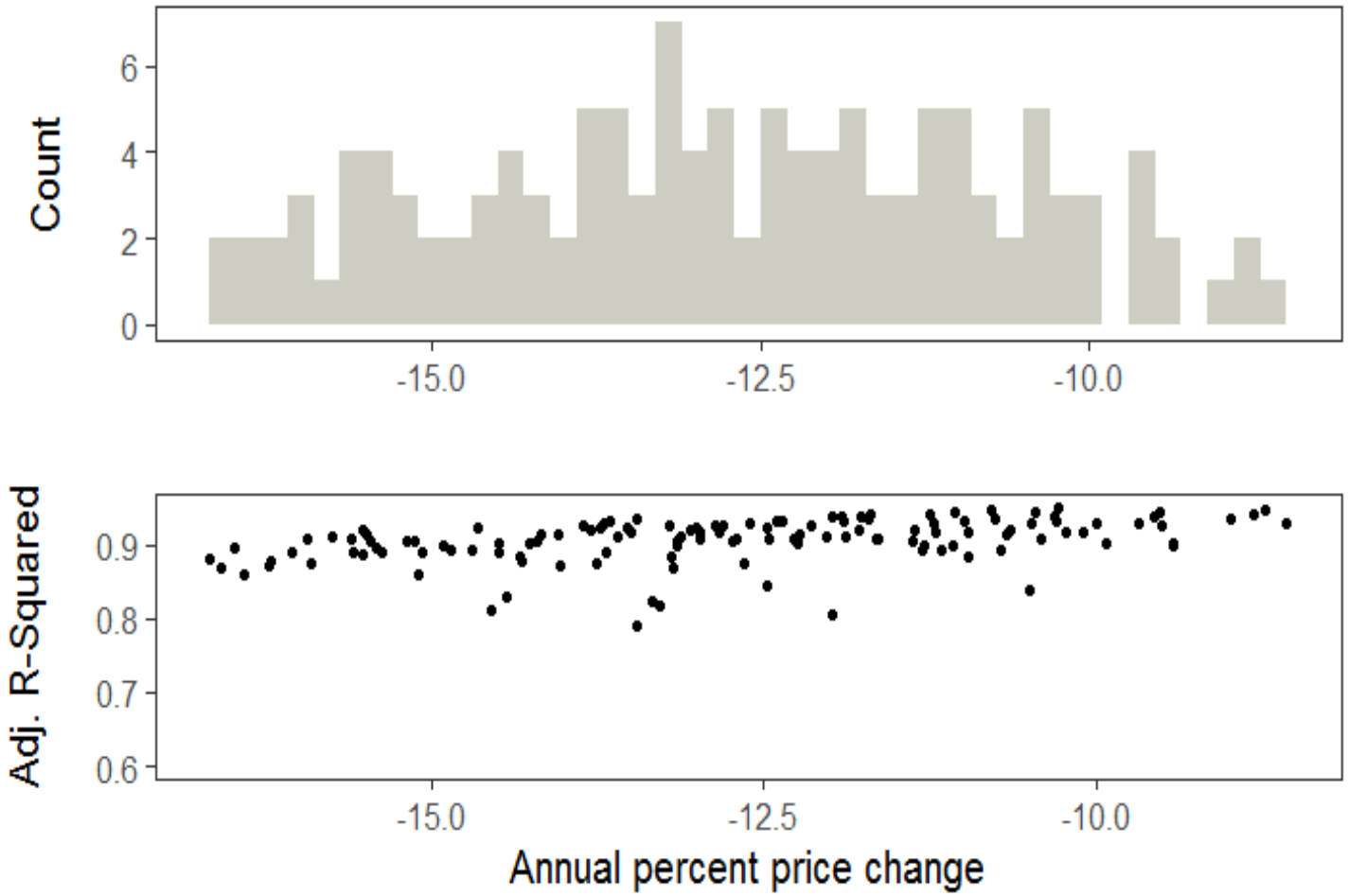
Hedonic Models

- Regression that quantifies relationship between price and characteristics of a product
- Time dummy variable gives estimate of quality adjusted price change
- Selection of other independent variables helps determine magnitude of time dummy
- $\log(price) = a_0 + \Delta d_{t+1} + b_2 \log(x_2) \ldots b_k \log(x_k)$

Microprocessors Characteristics

- Log (base frequency)
- Log (turbo frequency)
- Log (threads)
- Log (cores)
- Log (cache/cores)
- Log (TDP)
- Log (graphics execution units)
- Log (PassMark benchmark)

Range of Price Changes



Fitting the Model

- Underfitting – variables do not fully capture relationships in data
- Overfitting – variables capture random variation in a particular data set

Prediction

- Error can be quantified by mean squared error (MSE)
- Training error – in sample
- Test error – out of sample

Model evaluation measures

- Adjusted R2
  - Well-known, but has limitations
- Information criteria – AIC, BIC, etc.
  - Allow us to adjust **training error** to estimate **test error**

Validation

Data set is split

- Estimate model on **training set**
- Use model to calculate MSE on **test set**

K-Fold Cross Validation

- Randomly split data into *k* parts
- Estimate model on *k*-1 parts
- Use model to calculate MSE for held out part
- Repeat for all *k* parts

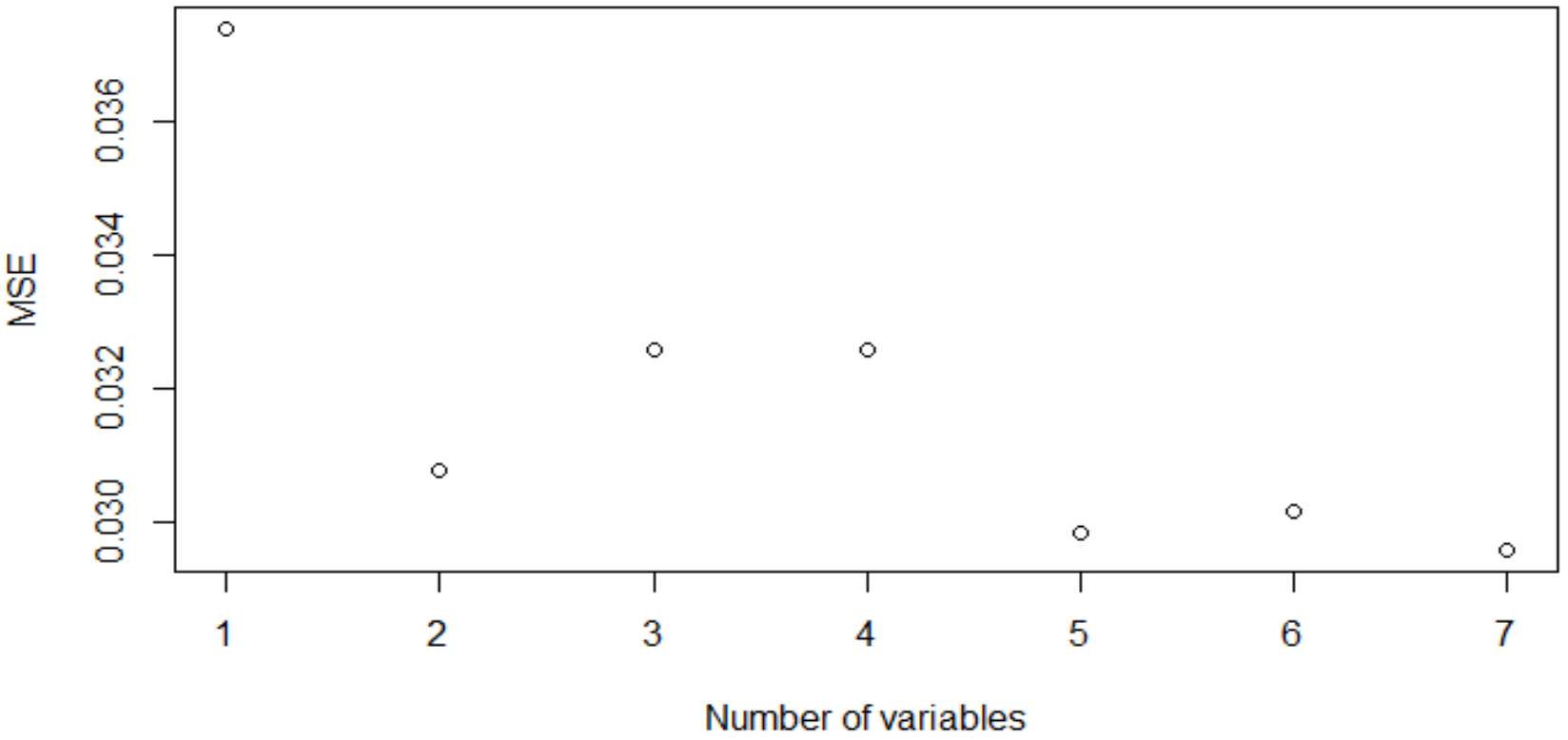
K-Fold Cross Validation

- Typically 5 to 10 folds
- Process can be repeated
- Important variables can be constrained to be in all candidate models
- Lowest MSE model estimated on entire dataset

Pre-screening

- Use best subsets to select lowest Residual Sum of Squares (RSS) 1 to *p* models
- Reduces computational load

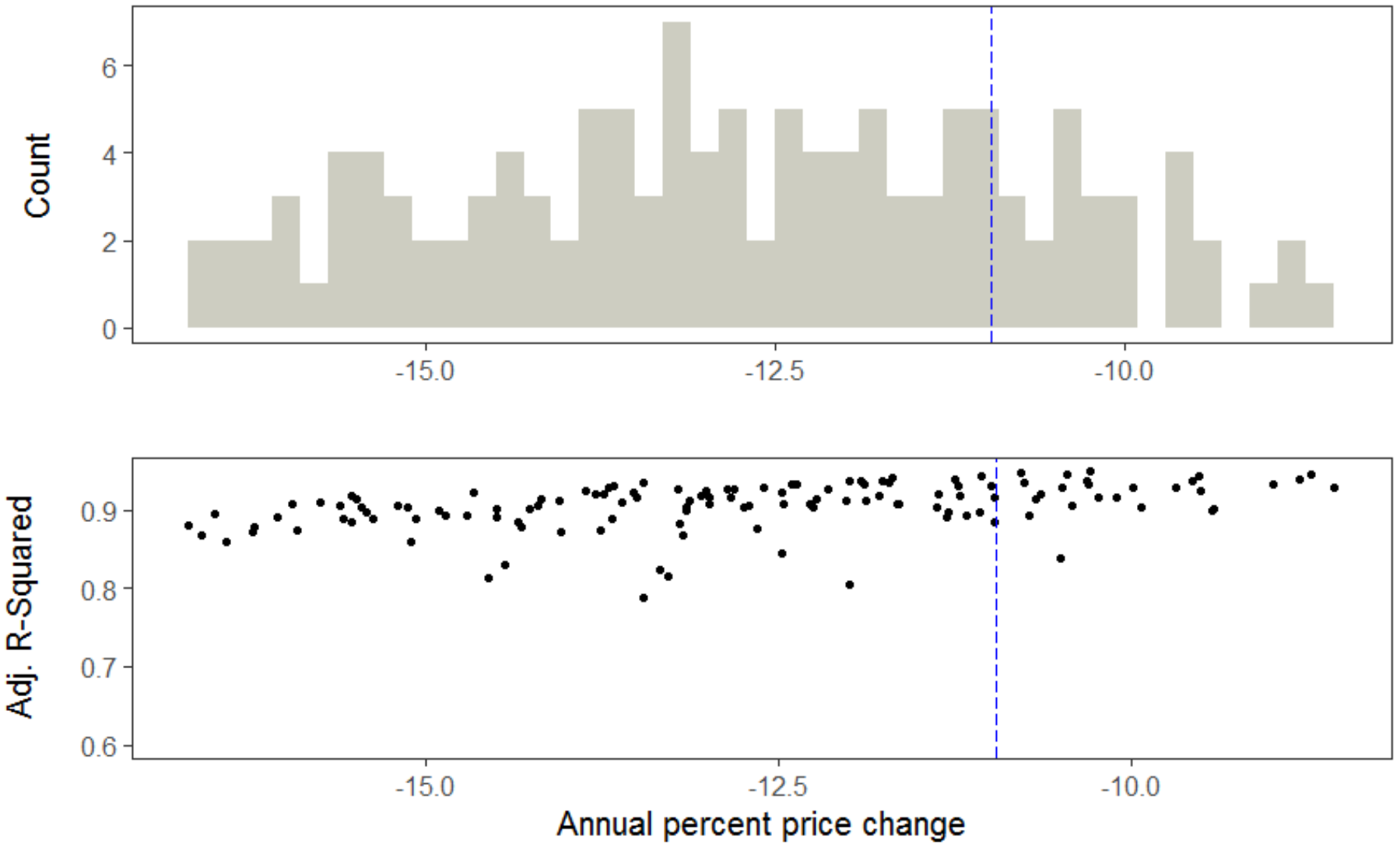
Example Results



K-Fold Cross Validation

- Advantages
  - Directly calculates test error
  - Uses entire dataset
- Disadvantage
  - Computationally intensive

Range of Price Changes



Benefits to PPI

- Transparent method for model selection
- Efficient to implement
- Specification can change over time
- Possibilities for future development

References

- *An Introduction to Statistical Learning*
- *A New Approach to Quality Adjusting PPI Microprocessors*

Contact Information

Steven Sawyer
Economist
PPI
www.bls.gov/ppi
202-691-7845
sawyer.steven@bls.gov