Support Vector Regression

Stuart McMenamin, David Simons

©2021 ITRON CONFIDENTIAL PROPRIETARY

Oct 2021

Agenda

» Data and Linear Model Specification

» Cross Validation (CV) Testing

» Bootstrap Estimation of Confidence Bands

» Overview of Alternative Linear Methods

» Intro to Support Vector Regression (SVR)

» Selecting SVR Hyper Parameters

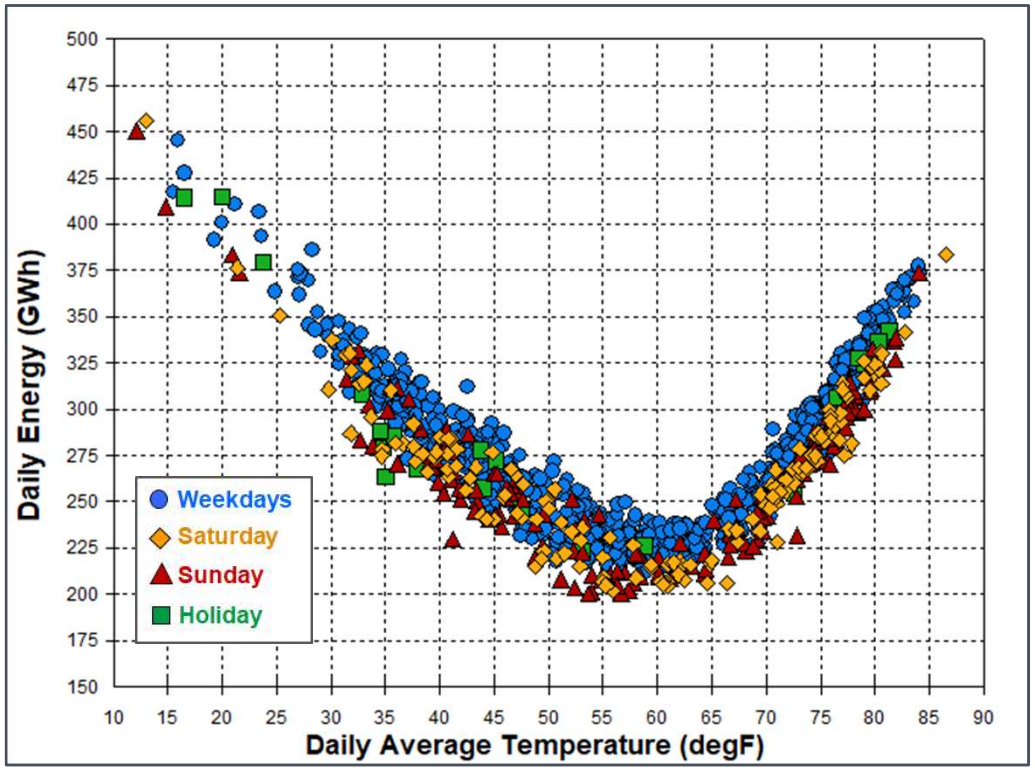
» Comparison of OLS and SVR Results

» Conclusions

Itron ©2021 ITRON CONFIDENTIAL PROPRIETARY

Overview of Data

- » Daily energy in GWh
  - PJM Data for Dominion
  - Daily data for 2017 to 2019
  - N=1,095 observations
  - All pre COVID
- » Hourly weather data
  - Dry bulb temperature (deg F)
  - Wind speed (mph)
  - Cloud cover index (0 to 8)
- » Calendar variables
  - Month
  - Day type
  - Holidays



©2021 ITRON CONFIDENTIAL PROPRIETARY

Model Specification

» Y Variable = Daily Energy in GWh

» Calendar Variables

- Monthly binaries (Jan to Nov)
- Day of week (Mon to Sat) (+)
- Bank Holidays, Retail Holidays (-)
- XMas week before and after (-)
- Time trend (+1/365 each day) (+)

» Solar variables

- Clouds = avg clouds 8 am to 8 pm (+)
- CloudsHot (Clouds × CD70) (-)
- CloudsCold (Clouds × HD55) (+)

» Wind variables

- WindHot (WindSpeed × CD70) (-)
- WindCold (WindSpeed × HD55) (+)

» Warm-side variables

- CD60_Cap5, CD65_Cap5, CD70_Cap5, Cd75_Cap5, CD80 (+)
- LagCD=.8*Lag(65,1)+.2*Lag(65,2) (+)
- LagCDCold=LagCD*(HD55>0) (-)
- SpringCD60, FallCD60 (-,?)
- Sat×CD65, Sun×CD65 (-)

» Cold-side variables

- HD60_Cap5, HD55_Cap5,
- HD50_Cap5, HD45_Cap10, HD35 (+)
- LagHD (.8*Lag(55,1)+.2*Lag(55,2))
- LagHDHot=LagHD*(CD65>0) (-)
- Fall×HD55, Spring×HD55 (-,?)
- Sat×HD55, Sun×HD55 (?)

©2021 ITRON CONFIDENTIAL PROPRIETARY

OLS Regression Model – Coefficients and Statistics

Variable	Coef	StdErr	T-Stat
CONST	179.608	2.218	80.979
Jan	1.318	0.855	1.542
Feb	-1.737	0.856	-2.028
Mar	-8.656	1.016	-8.522
Apr	-11.432	1.113	-10.269
May	-10.562	1.254	-8.422
Jun	-7.519	1.576	-4.772
Jul	-6.387	1.646	-3.882
Aug	-4.385	1.604	-2.733
Sep	-5.674	1.328	-4.271
Oct	-8.736	1.128	-7.746
Nov	-5.637	1.015	-5.553
Monday	14.931	0.823	18.148
Tuesday	16.422	0.826	19.874
Wednesday	16.429	0.830	19.792
Thursday	16.224	0.832	19.492
Friday	13.804	0.829	16.657
Saturday	0.485	0.949	0.512
BankHols	-9.700	1.588	-6.108
RetailHols	-15.809	1.710	-9.246
WkB4XMas	-4.805	2.535	-1.896
WkAftXMas	-9.715	1.617	-6.010
Trend2010	4.228	0.190	22.225

Variable	Coef	StdErr	T-Stat
CD60Cap5	2.512	0.289	8.679
CD65Cap5	4.091	0.250	16.384
CD70Cap5	7.218	0.269	26.847
CD75Cap5	8.095	0.228	35.477
CD80	9.044	0.450	20.093
LagCD	0.847	0.082	10.271
LagCDCold	-0.784	0.224	-3.502
SpringCD	-1.782	0.221	-8.051
FallCD	-0.482	0.102	-4.704
SatCD	-0.059	0.092	-0.639
SunCD	-0.123	0.090	-1.374
WindHot	-0.117	0.023	-5.167
CloudsHot	-0.089	0.045	-1.982
AvgClouds	0.433	0.114	3.790

Variable	Coef	StdErr	T-Stat
HD60Cap5	0.194	0.244	0.793
HD55Cap5	2.115	0.254	8.326
HD50Cap5	2.457	0.209	11.780
HD45Cap10	3.132	0.102	30.671
HD35	4.937	0.118	41.736
LagHD	1.123	0.040	28.054
LagHDHot	-0.626	0.228	-2.743
SpringHD	-0.434	0.112	-3.889
FallHD	-0.817	0.139	-5.899
SatHD	0.158	0.058	2.738
SunHD	-0.053	0.058	-0.912
WindCold	0.048	0.006	8.311
CloudsCold	0.025	0.009	2.602
AvgClouds	0.433	0.114	3.790

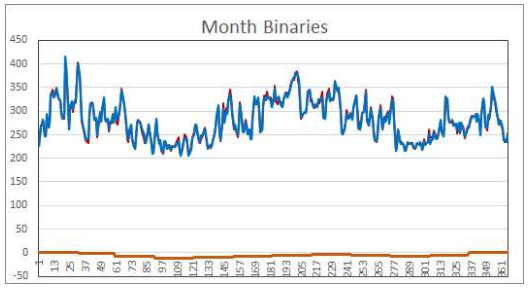
Statistic	OLS	OLS+AR1	Statistic	OLS	OLS+AR1
Observations	1095	1094	MAD	3.932	3.450
Coefficients	52	53	RMSE	4.921	4.390
R Square	0.987	0.989	MAPE	1.43%	1.26%
BIC	3.508	3.289	DW	1.162	2.029

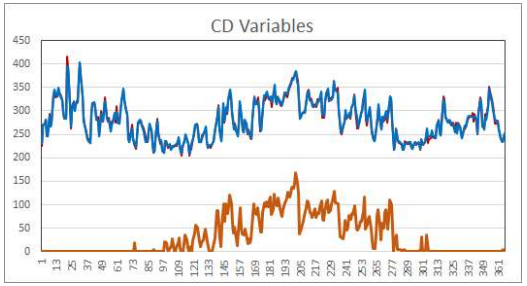
T-Stat<2

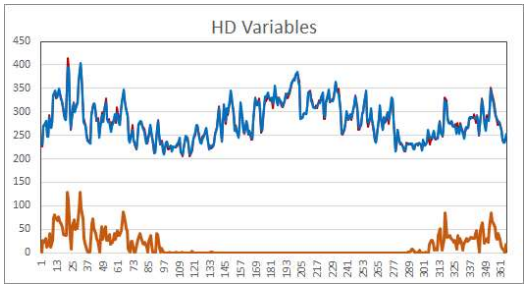
T-Stat<1

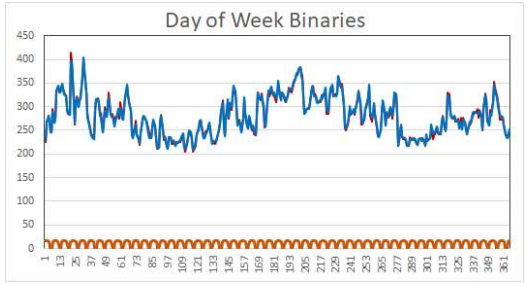
©2021 ITRON CONFIDENTIAL PROPRIETARY

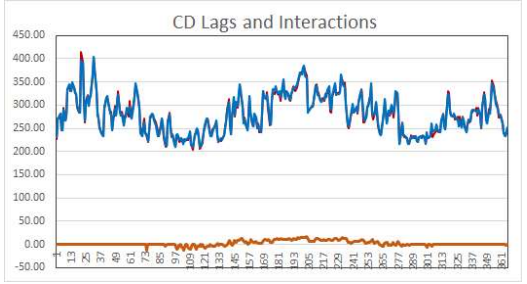
Decomposition of Predicted Values (BX tab) -- 2017

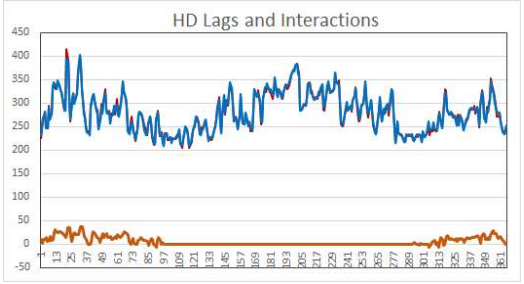


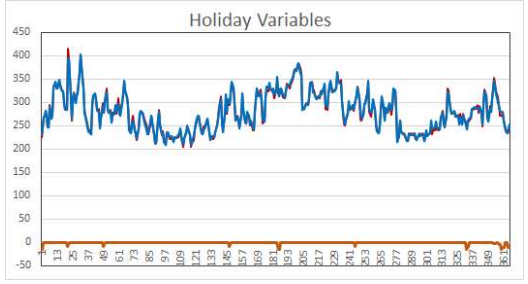


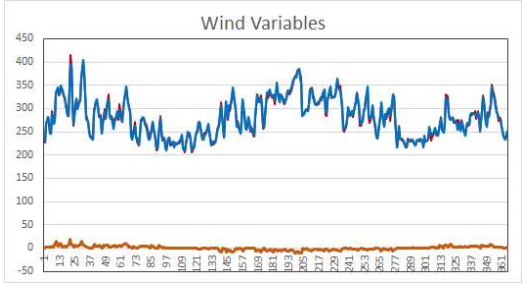


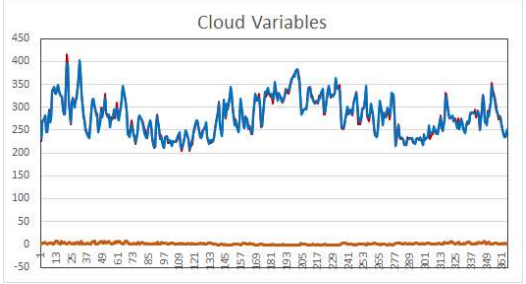












©2021 ITRON CONFIDENTIAL PROPRIETARY

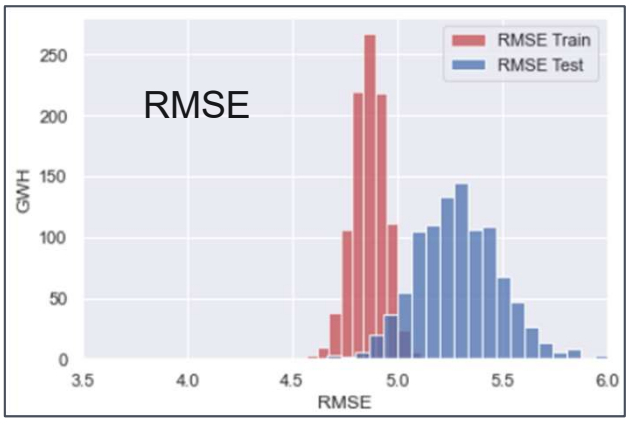
Cross Validation (CV) Testing

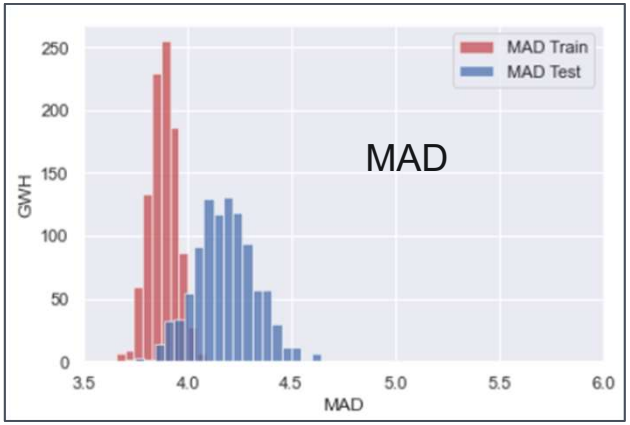
©2021 ITRON CONFIDENTIAL PROPRIETARY

Cross Validation (CV) – Out of Sample Testing

- Split data into Train/Test or Train/Tune/Test subsets
  - K Fold (various versions), Leave P Out, Shuffle & Split
  - Used here – Shuffle&Split with sklearn.train_test_split
  - Test Fraction = 30%. Train on 766, Test on 329.
  - 1000 splits using random_start = 1 to 1000
- Quantifies loss of model performance for out of sample observations.

	Average		Median		Best		Worst	
	Train	Test	Train	Test	Train	Test	Train	Test
RMSE	4.863	5.284	4.867	5.278	4.566	4.661	5.112	6.005
MAD	3.881	4.178	3.882	4.177	3.655	3.699	4.088	4.641
MAPE	1.416	1.523	1.417	1.522	1.338	1.351	1.488	1.686



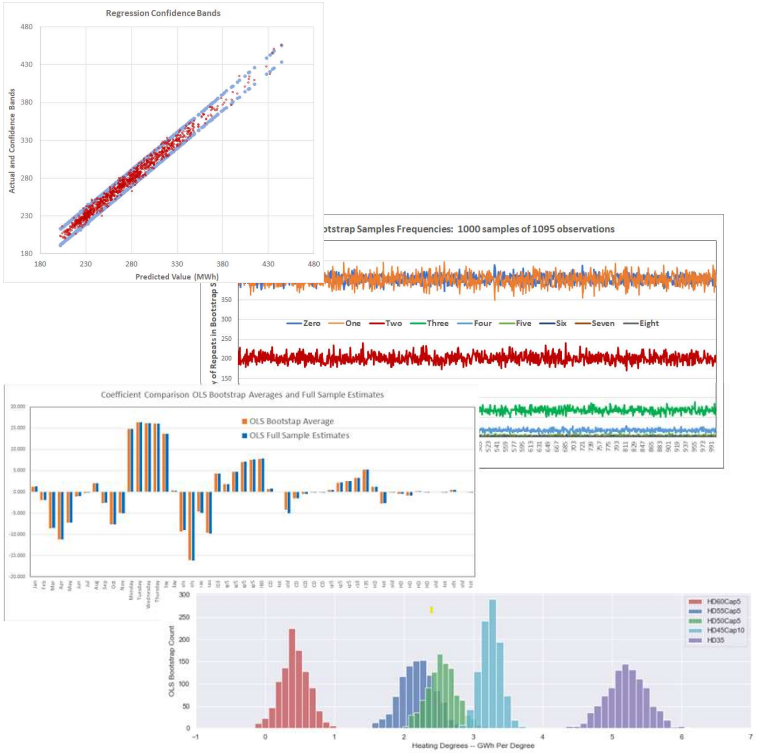


Bootstrap Estimation of Confidence Bands

©2021 ITRON CONFIDENTIAL PROPRIETARY

Bootstrap Methods

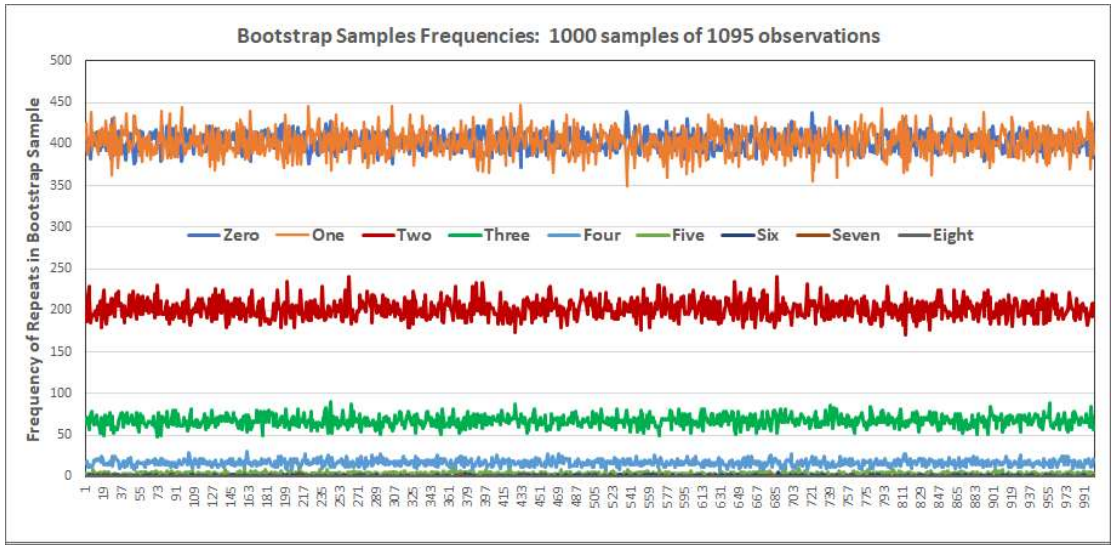
- » Bootstrap samples:
  - Sampling with replacement
  - Sample size is the same as the history (1,095)
  - Some points are selected multiple times
  - Some points are not selected at all
- » Post processing of bootstrap estimation results
  - Analyze coefficient estimates (standard deviations)
  - Analyze predicted values (confidence intervals)
- » Why do we need this?
  - We don't with OLS, because:
    - OLS provides coefficient standard errors
    - OLS provides prediction confidence intervals
  - But SVR does not provide either
- » So, let's look at how bootstrap works in a regression context



What is a Bootstrap Sample

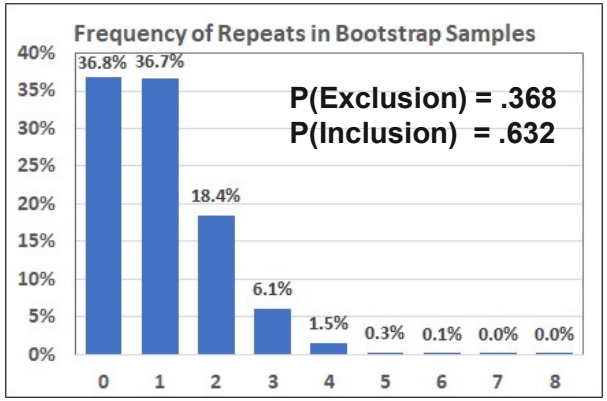
» Bootstrap samples use sampling with replacement

- Pick an observation (data row) at random. Leave it in the list. Repeat N times.
- We started with 1,095 data rows. Each bootstrap sample will also have 1,095 selected rows
- In a sample, some data rows are selected multiple times and some are not selected at all
- In what follows, we repeat the bootstrap sampling and model estimation process 1,000 times

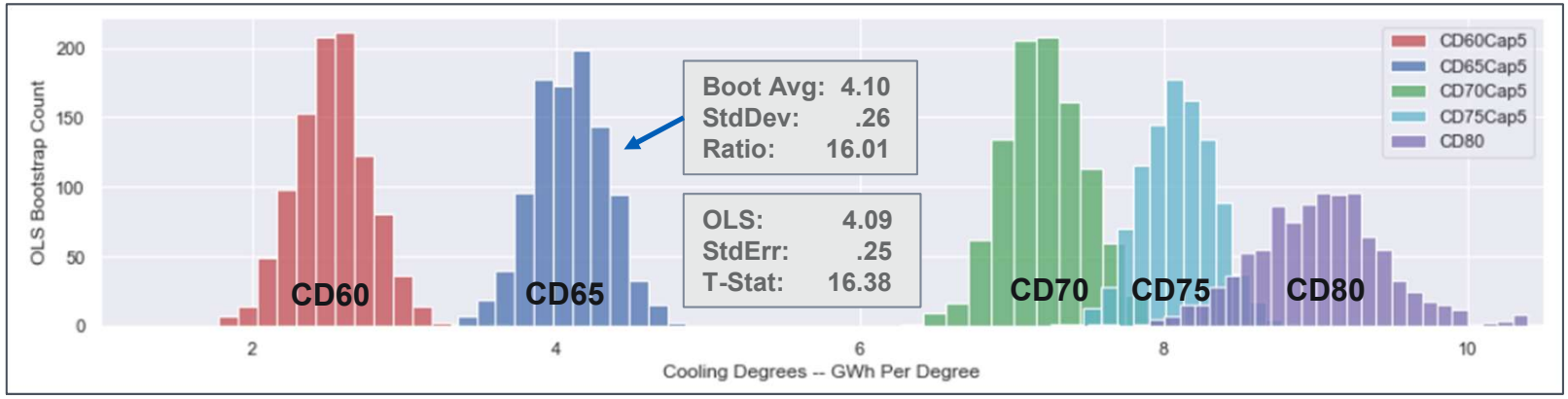


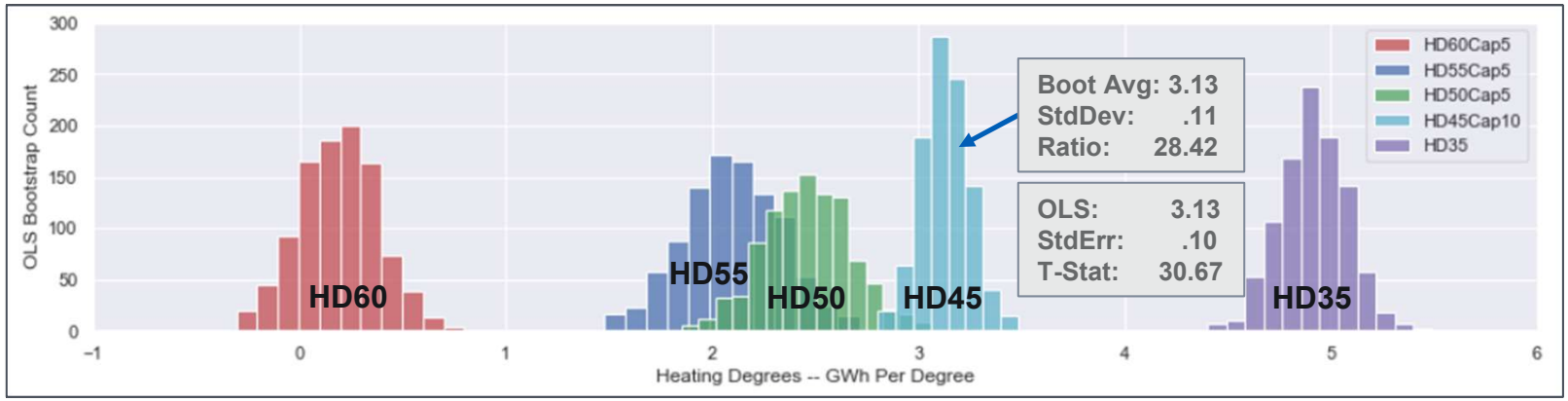
$$\text{Probability of Exclusion} = \left(1 - \frac{1}{N}\right)^{N}$$

.368 for N = 1,095



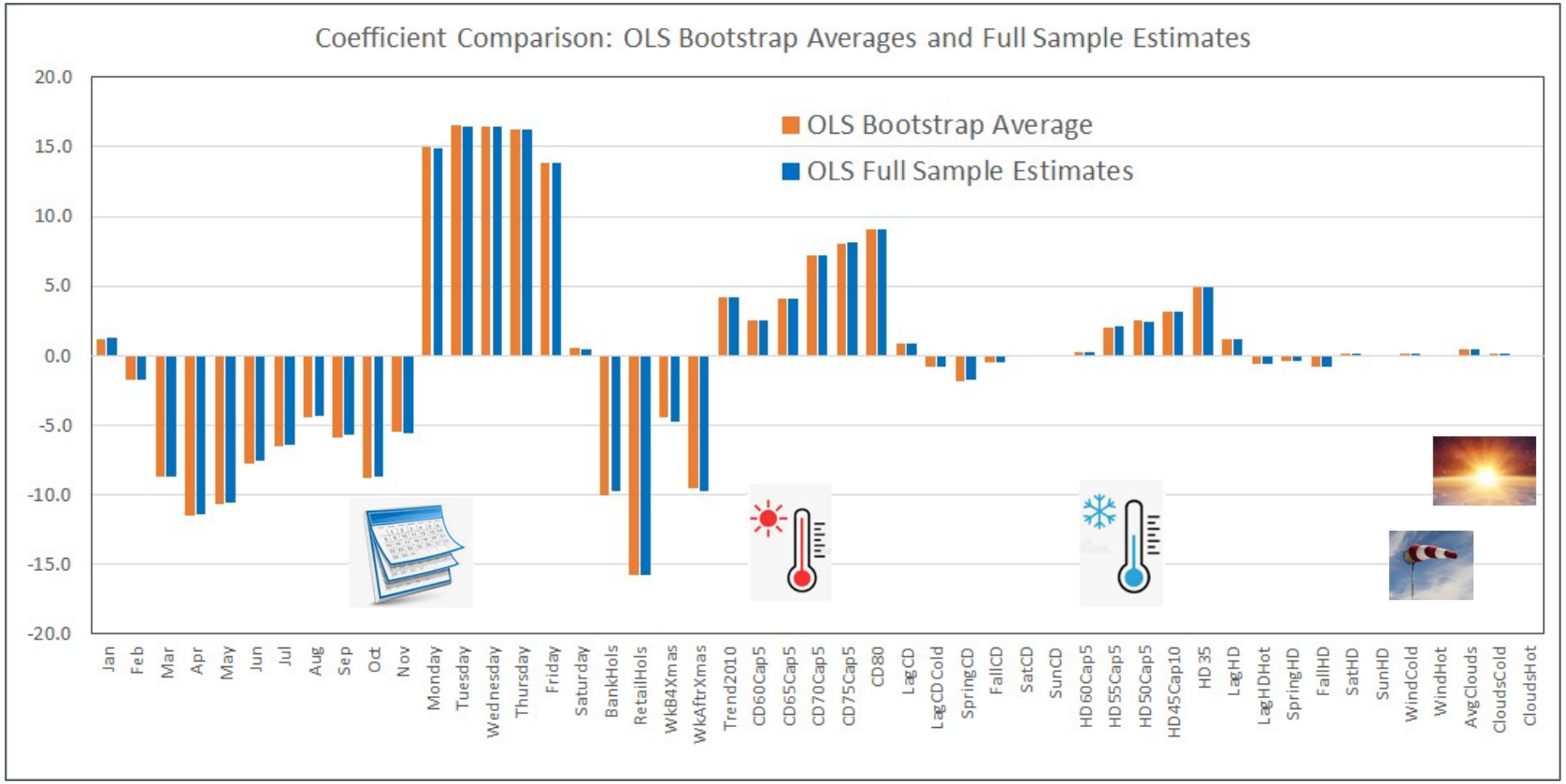
Using Bootstrap to Estimate Coefficient Statistics



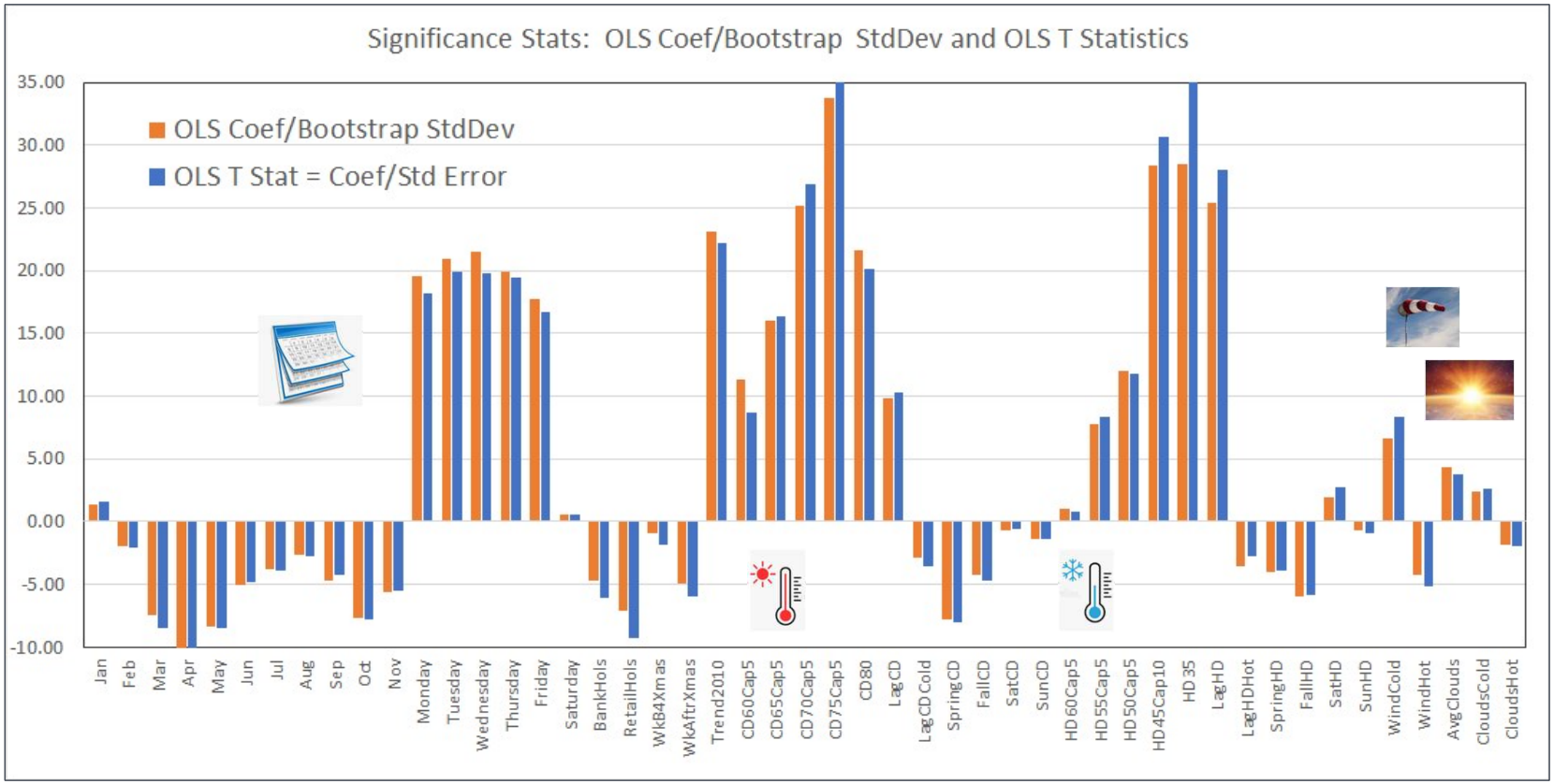


©2021 ITRON CONFIDENTIAL PROPRIETARY

OLS Bootstrap Average vs. Full Sample Coefficients

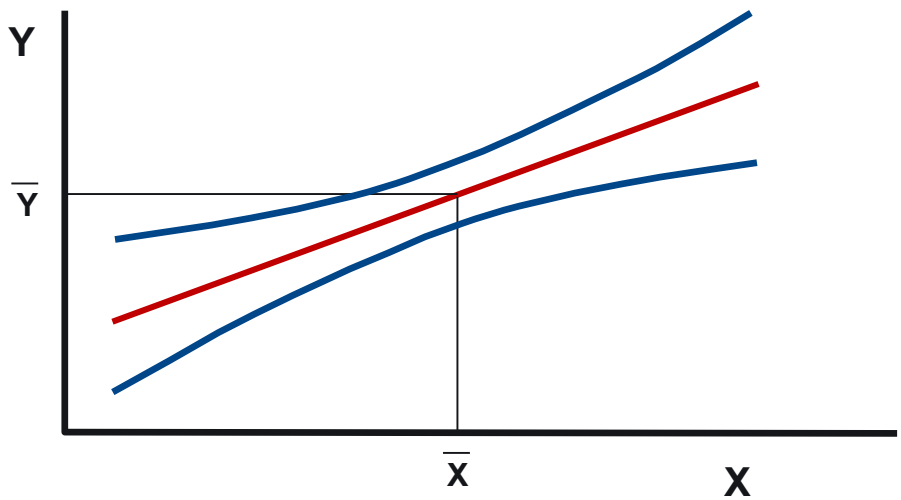


OLS Bootstrap Average vs. Full Sample “T Ratios”



Reminder on OLS Confidence Bands

» Sigma (σ) is the OLS model standard error
- Confidence bands are a bit wider than 1.96σ at the mean
- Bands get wider as X moves away from its mean
- Additional uncertainty if X is not known (as in forecasting)



Formula with multiple X Variables

$$\hat{\sigma}_f^2 = \hat{\sigma}^2 \times (1 + X_f(X'X)^{-1}X_f')$$

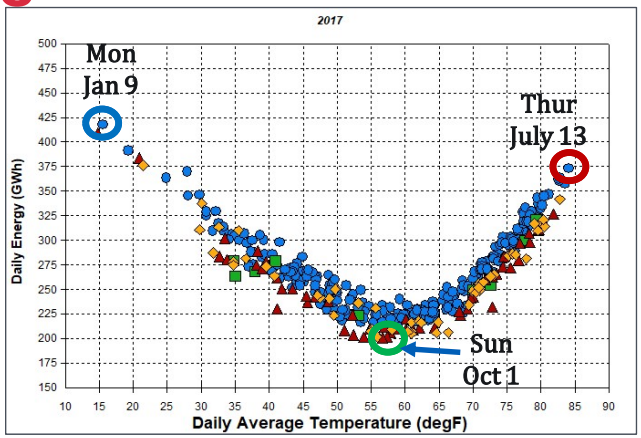
Formula with one X Variable

$$\hat{\sigma}_f^2 = \hat{\sigma}^2 \times \left(1 + \frac{1}{N} + \frac{(X_f - \bar{X})^2}{\sum_{t=1}^{N}(X_t - \bar{X})^2}\right)$$

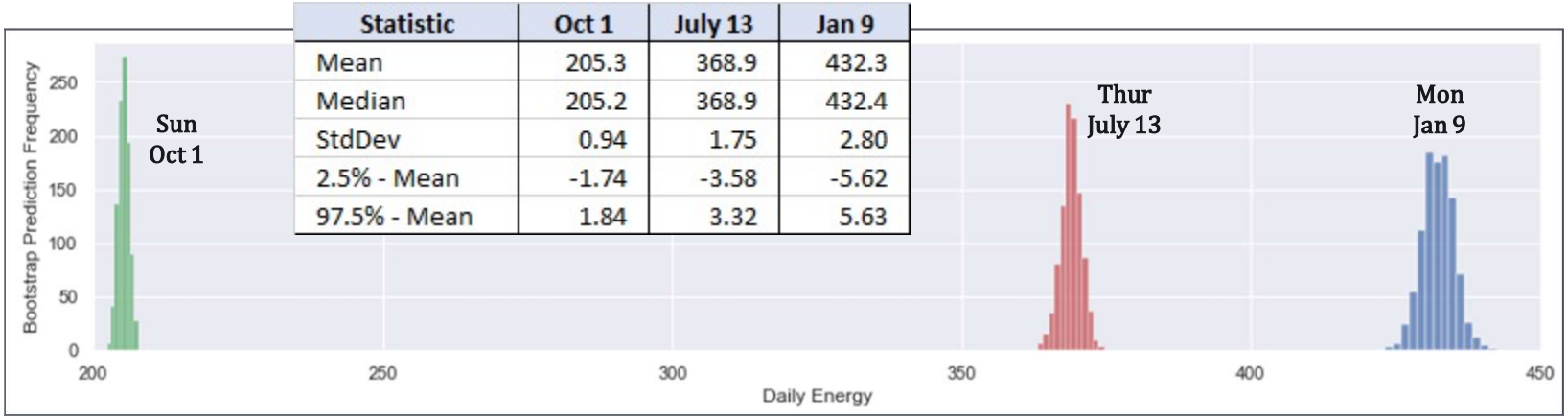
©2021 ITRON CONFIDENTIAL PROPRIETARY

Bootstrap Prediction Distributions

- Each bootstrap estimation generates parameters
- Parameters generate predicted values
- So, we have 1,000 predictions for each day
  - Compute statistics (mean, std dev, percentiles)
- Quantifies prediction variance from parameter uncertainty



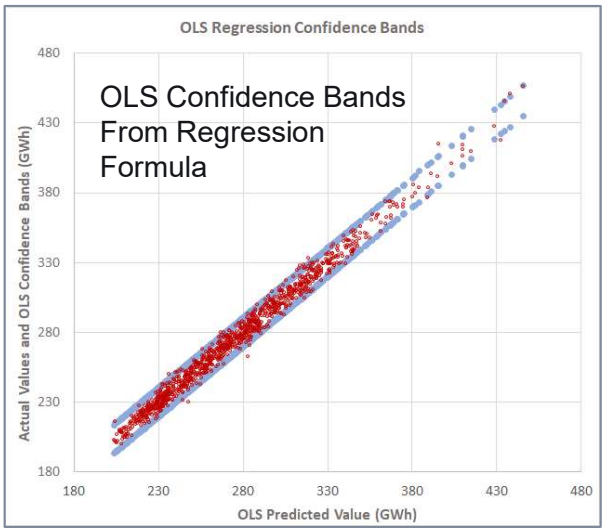
Statistic	Oct 1	July 13	Jan 9
Mean	205.3	368.9	432.3
Median	205.2	368.9	432.4
StdDev	0.94	1.75	2.80
2.5% - Mean	-1.74	-3.58	-5.62
97.5% - Mean	1.84	3.32	5.63

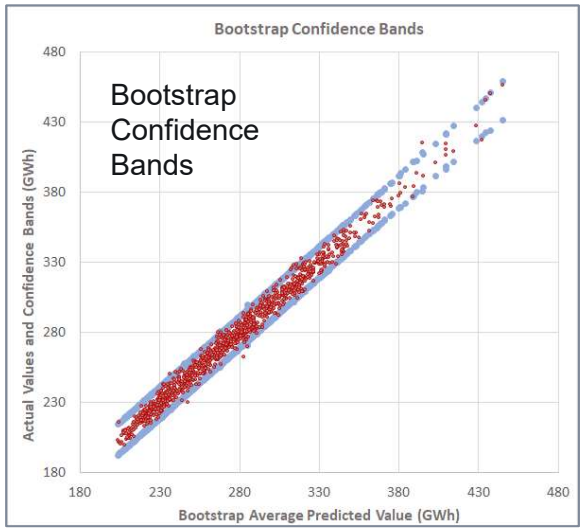


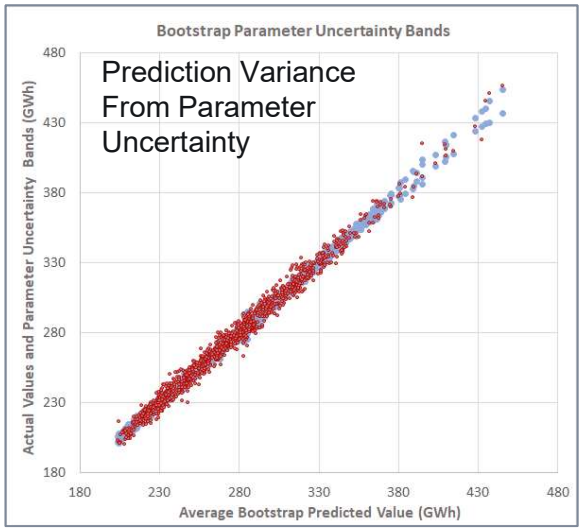
Confidence Bands from Bootstrap Estimations

» Use Bootstrap results to estimate OLS confidence bands

- Results are a bit wider (about 10%) than the results from regression confidence band formula
  - This is consistent with the accuracy loss we saw in OLS cross validation
  - This suggests that the OLS confidence bands are optimistic for out-of-sample predictions
- Quantify "parameter variance" component from distribution of bootstrap predicted values







©2021 ITRON CONFIDENTIAL PROPRIETARY

Overview of Alternative Linear Methods

©2021 ITRON CONFIDENTIAL PROPRIETARY

Why Look at Alternatives to OLS

» OLS is the best linear unbiased estimator under the standard assumptions
- Errors are uncorrelated with zero mean
- Errors have equal variance (homoscedastic)
- Note, the term “best” relates to the precision of the estimated parameter values

» Criticisms of OLS
- The above conditions often are not met
- Even if they are met:
  - Squaring residuals makes regression sensitive to outliers
  - Multicollinearity causes individual parameters to be poorly identified
  - Alternative estimation approaches can have better accuracy for out-of-sample predictions

» Alternative linear regression methods
- Penalization/Regularization methods: Ridge, Lasso, Elastic Net, SVR
- Minimize absolute errors: Quantile (q=.5), SVR

©2021 ITRON CONFIDENTIAL PROPRIETARY

Summary of Alternative Approaches

Model: $Y_t = a + \sum_k b_k X_{k,t} + e_t$

No penalization
No hyper parameters
Scaling unnecessary

Note: With penalization, you need to scale the X variables to put the b's on an equal footing.
-- Standardize (X-Avg)/StDev
-- Min/Max normalize (0,1)

OLS: $\underset{\hat{b}}{\text{Min}} \sum_t \hat{e}_t^2$

Ridge: $\underset{\hat{b}|\lambda}{\text{Min}} \sum_t \hat{e}_t^2 + \lambda \sum_k \hat{b}_k^2$

Penalization: L2 Regularization
Lambda (λ) is a hyper parameter

Lasso: $\underset{\hat{b}|\lambda}{\text{Min}} \frac{1}{2N} \sum_t \hat{e}_t^2 + \lambda \sum_k |\hat{b}_k|$

Penalization: L1 Regularization (Least Absolute Shrinkage)
May cause coefficients to go to zero – variable selection
Lambda (λ) is a hyper parameter

Elastic Net: $\underset{\hat{b}|\lambda_1,\lambda_2}{\text{Min}} \frac{1}{2N} \sum_t \hat{e}_t^2 + \lambda_1 \sum_k |\hat{b}_k| + \lambda_2 \sum_k \hat{b}_k^2$

Penalization: L1 and/or L2 Regularization
May cause coefficients to go to zero – variable selection
Lambdas are hyper parameters

Quantile: $\underset{\hat{b}|q=.5}{\text{Min}} \sum_t |\hat{e}_t|$

No penalization. q is a hyper parameter
With q=.5, quantile minimizes the sum of absolute errors

SVR: $\underset{\hat{b}|c,\varepsilon}{\text{Min}} \; C \sum_{t \in \varepsilon+} (|\hat{e}_t| - \varepsilon) + \sum_k \hat{b}_k^2$

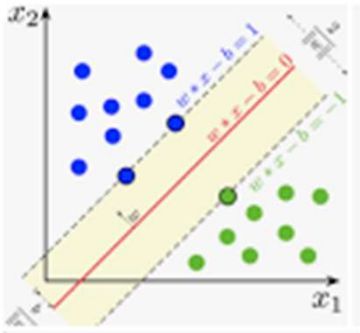
Penalization: L2 Regularization
The error sum includes only the errors larger than ε
C and epsilon (ε) are hyper parameters

©2021 ITRON CONFIDENTIAL PROPRIETARY

Introduction to Support Vector Regression (SVR)

©2021 ITRON CONFIDENTIAL PROPRIETARY

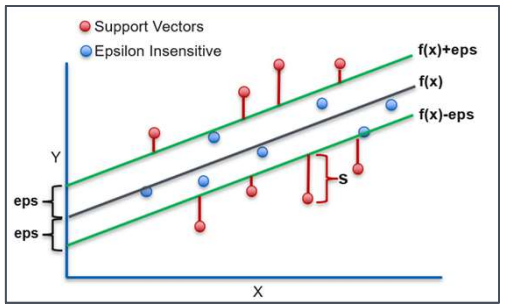
Moving on to SVR



» Started with SVM (support vector machines)
- Classification problems with labeled data (early 1990's)
- Success at character rec, image classification, and bioinformatics
- Initially, find separating plane with the widest margin
- Extended to minimize the distance of misclassified points from the margin boundary
- In these applications, only points that are on the boundary or misclassified count

» Extension to Regression in mid 1990's
- Various kernels: Linear, Radial Basis, Polynomial, …
- With Linear kernel, the model is $y_t = a + bX_t + e_t$
- The estimation objective can be expressed as:

$$\underset{\hat{b}|c,\varepsilon}{\text{Min}} \sum_k \hat{b}_k^2 + C\sum_t s_t \qquad \text{where } s_t = |\hat{e}_t| - \varepsilon \ \text{ if } |\hat{e}_t| > \varepsilon$$
$$= 0 \qquad \text{if } |\hat{e}_t| \le \varepsilon$$

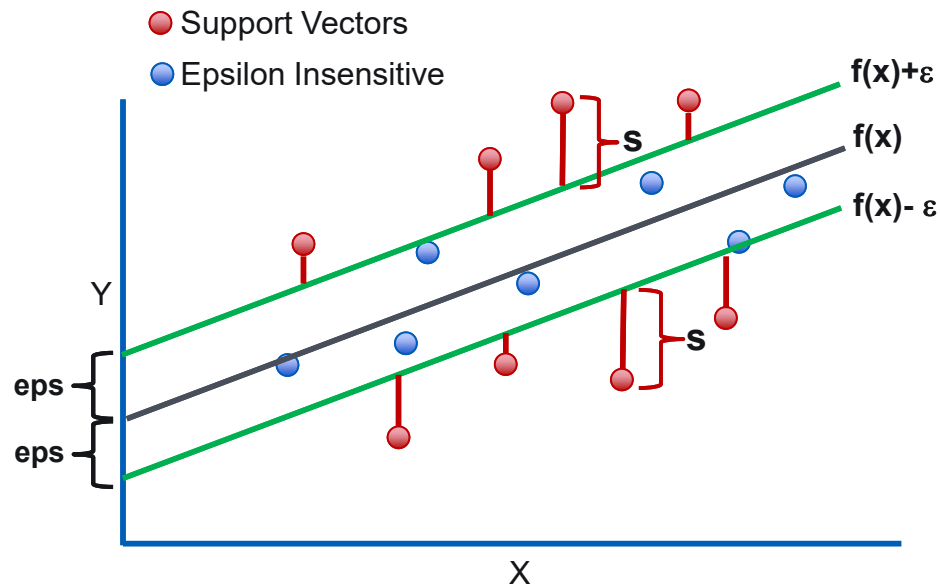


SVR Estimation Given C, eps

$$\underset{\hat{b}|c,\varepsilon}{\text{Min}} \sum_k \hat{b}_k^2 + C\sum_t s_t$$

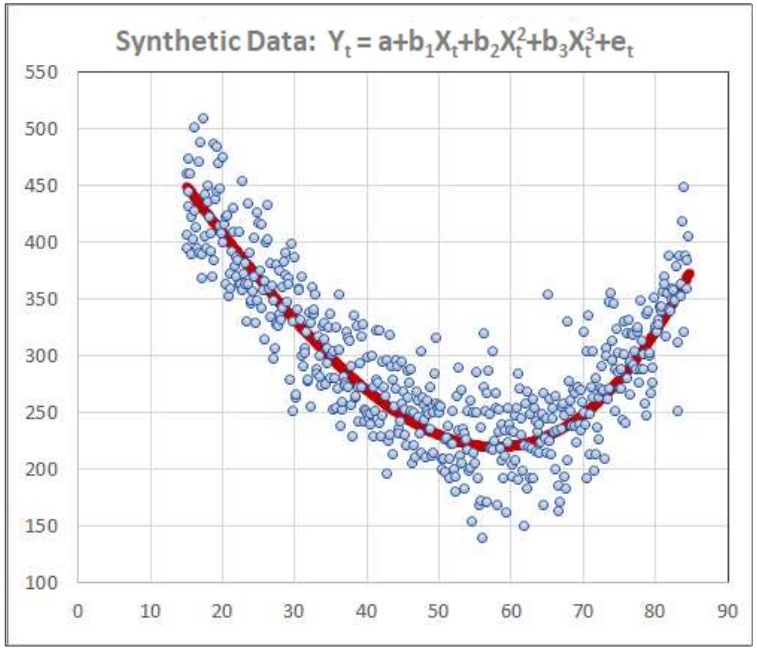
$$\text{where } s_t = |\hat{e}_t| - \varepsilon \quad \text{if } |\hat{e}_t| > \varepsilon$$
$$= 0 \quad \text{if } |\hat{e}_t| \le \varepsilon$$

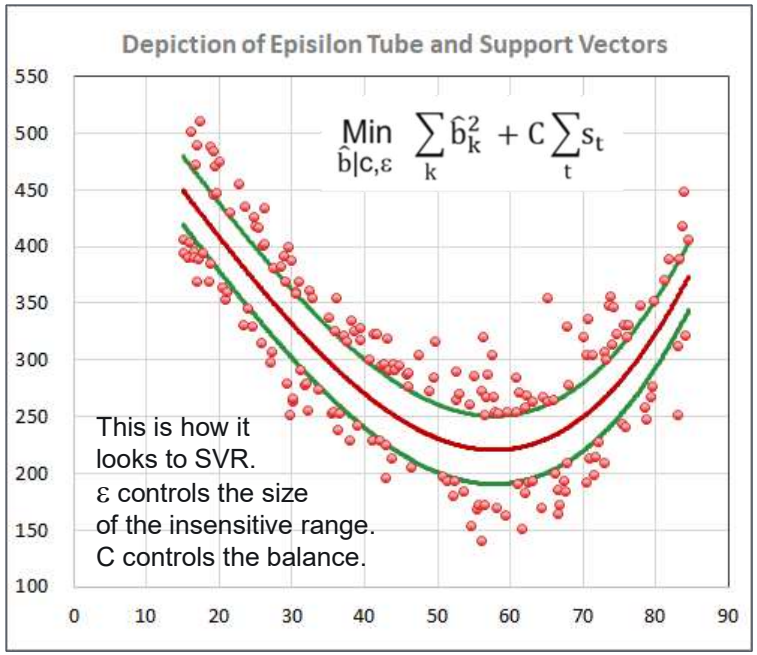
- » Epsilon insensitive method
  - Points inside epsilon tube (blue) don't count
  - Points outside (red) do count
  - S is the vertical distance to the margin
  - Red points are the "support vectors"
- » C and ε are hyper parameters
  - Can't minimize over C or ε in training
  - If you minimized over C, ε, b
    - C would go to 0
    - eps would get big (sending s to 0)
    - b would go to 0
    - Model would fail out of sample tests
  - Select C, ε using cross validation



What SVR looks like for a Daily Energy Model

» It is a different idea: Make the sum of squared parameters small while also making the sum of the vertical distances from the epsilon boundaries small.





Selecting SVR Hyper Parameters

©2021 ITRON CONFIDENTIAL PROPRIETARY

Selecting C and eps

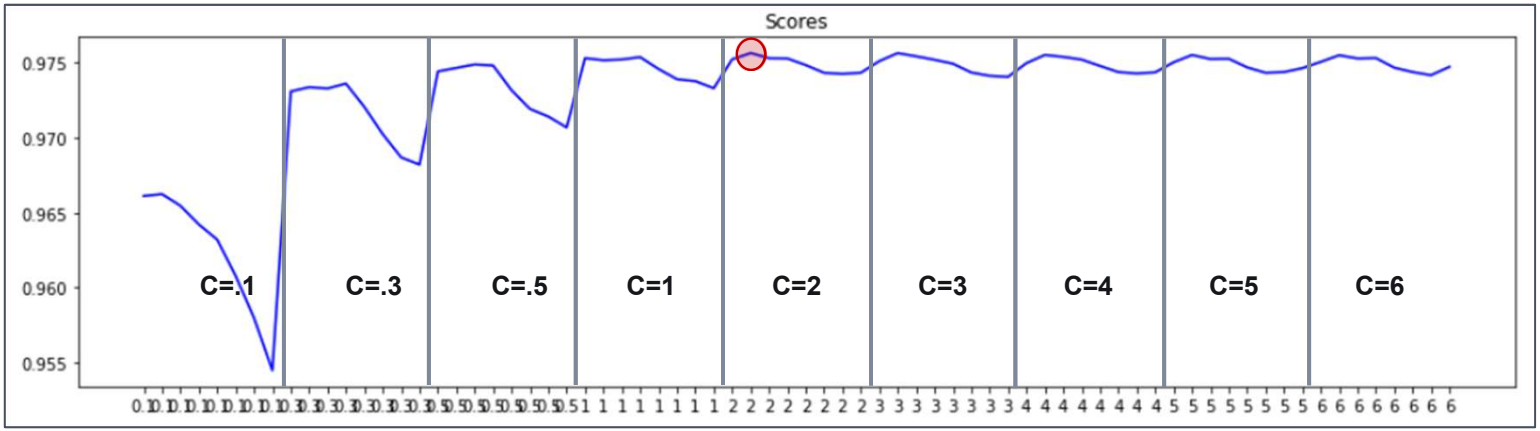
» So, how do we select a value for C and ε to support estimation

» One approach is a grid search using Train/Test cross validation (CV)

- Loop over values of C and eps
  - Loop over cross validation folds (10-fold used)
    - Estimate model using CV training set
    - Calculate score (R-square) using CV test set
  - Average test scores across the folds

C=.1, .3, .5, 1, 2, 3, 4, 5, 6

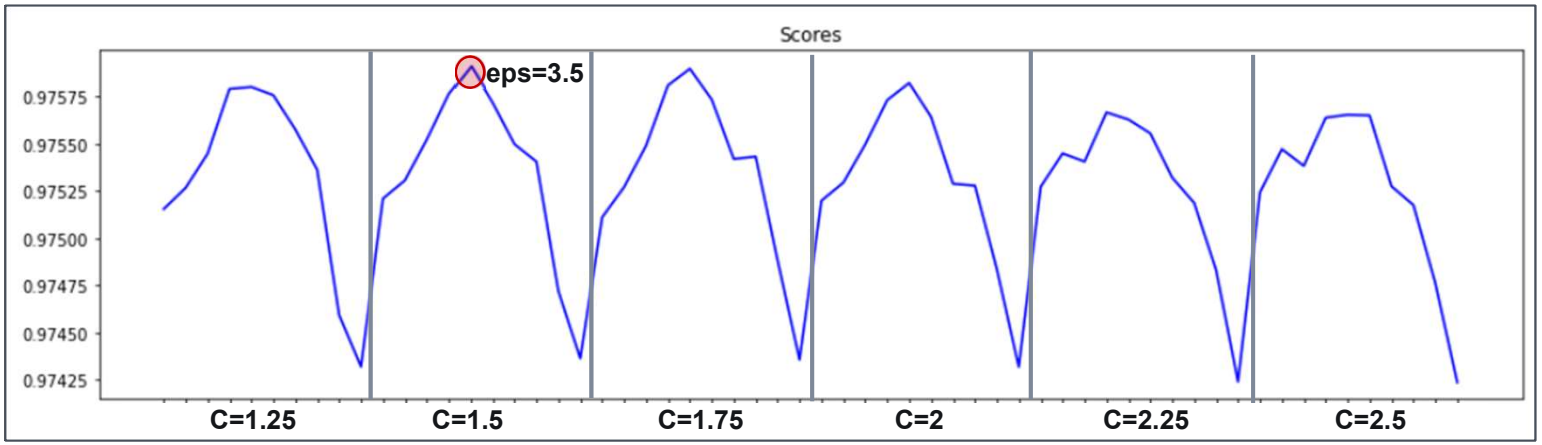
ε =1, 2, 3, 4, 5, 6, 7, 8



Selecting C and eps (cont.)

- Second round, focused on C between 1.25 and 2.5
- Explored values of eps in more detail and focused in on ε in the 3 to 4 range
- Best score is C=1.5, eps = 3.5
  - ε value is slightly below the OLS regression MAD (about 4.0)
  - 51% of observations are epsilon insensitive
  - 49% of observations are support vectors

C=1.25, 1.5, 1.75, 2, 2.25, 2.5
ε =2.75, 3, 3.25, 3.4, 3.5, 3.6, 3.75, 4, 5, 6



Comparison of OLS and SVR Results

©2021 ITRON CONFIDENTIAL PROPRIETARY

How Does SVR Stack Up for This Problem

- Training Statistics (Full Sample) are slightly worse than regression
  - RMSE: OLS = 4.93, SVR = 4.98
  - MAD: OLS = 3.93, SVR = 3.96
- How about out of sample performance?
  - Shuffle Train/Test splits, 30% Test, 1000 splits
  - OLS is slightly better out of sample for most statistics
  - But SVR is a solid contender
  - It could be different for problems with fewer unbalanced variables

OLS Train/Test Results from 1000 Train/Test Splits

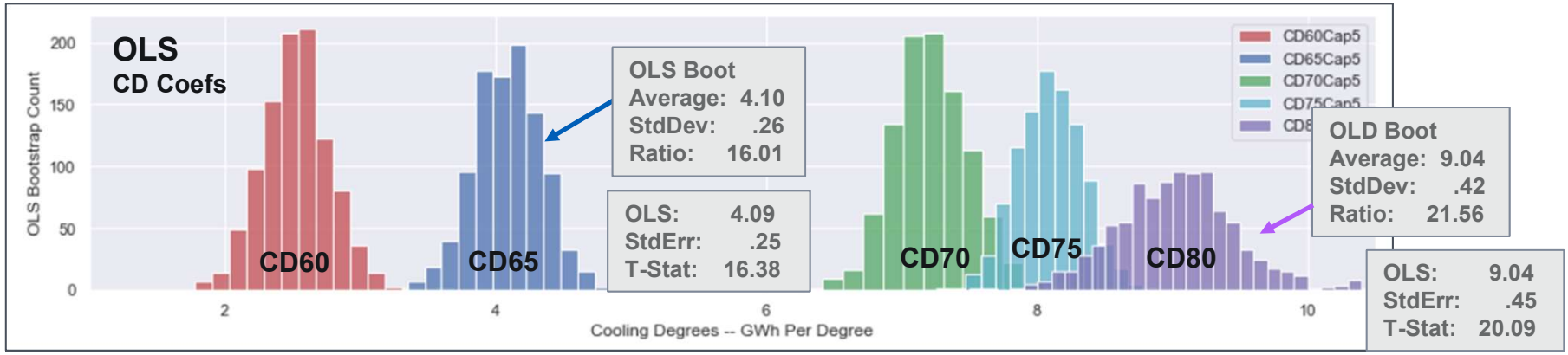
	Average		Median		Best		Worst	
	Train	Test	Train	Test	Train	Test	Train	Test
RMSE	4.863	5.284	4.867	5.278	4.566	4.661	5.112	6.005
MAD	3.881	4.178	3.882	4.177	3.655	3.699	4.088	4.641
MAPE	1.416	1.523	1.417	1.522	1.338	1.351	1.488	1.686

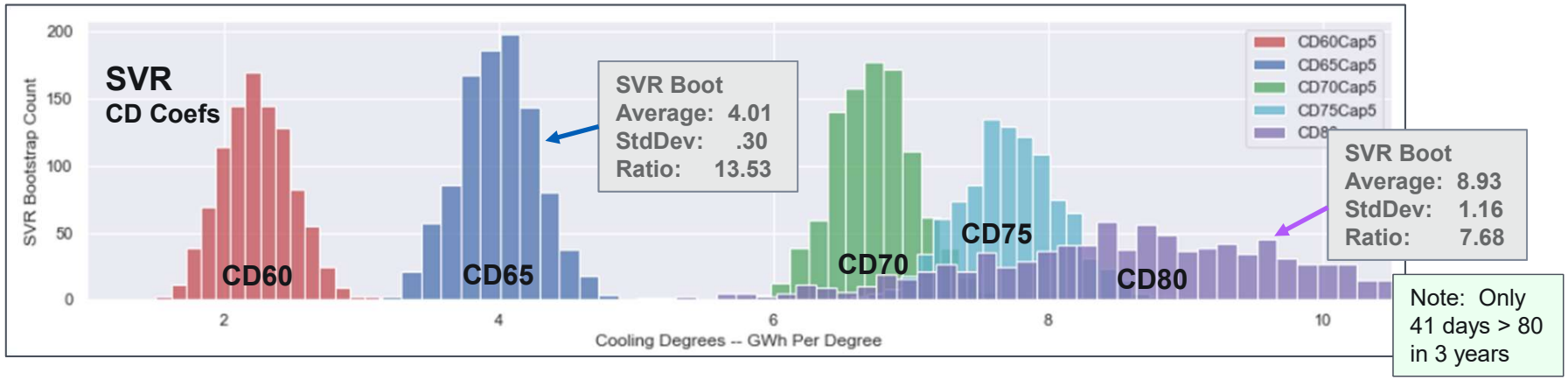
SVR Train/Test Results from 1000 Train/Test Splits

	Average		Median		Best		Worst	
	Train	Test	Train	Test	Train	Test	Train	Test
RMSE	4.955	5.344	4.957	5.340	4.669	4.513	5.197	6.163
MAD	3.934	4.200	3.936	4.196	4.150	3.618	4.150	4.770
MAPE	1.437	1.533	1.439	1.531	1.363	1.325	1.514	1.717

©2021 ITRON CONFIDENTIAL PROPRIETARY

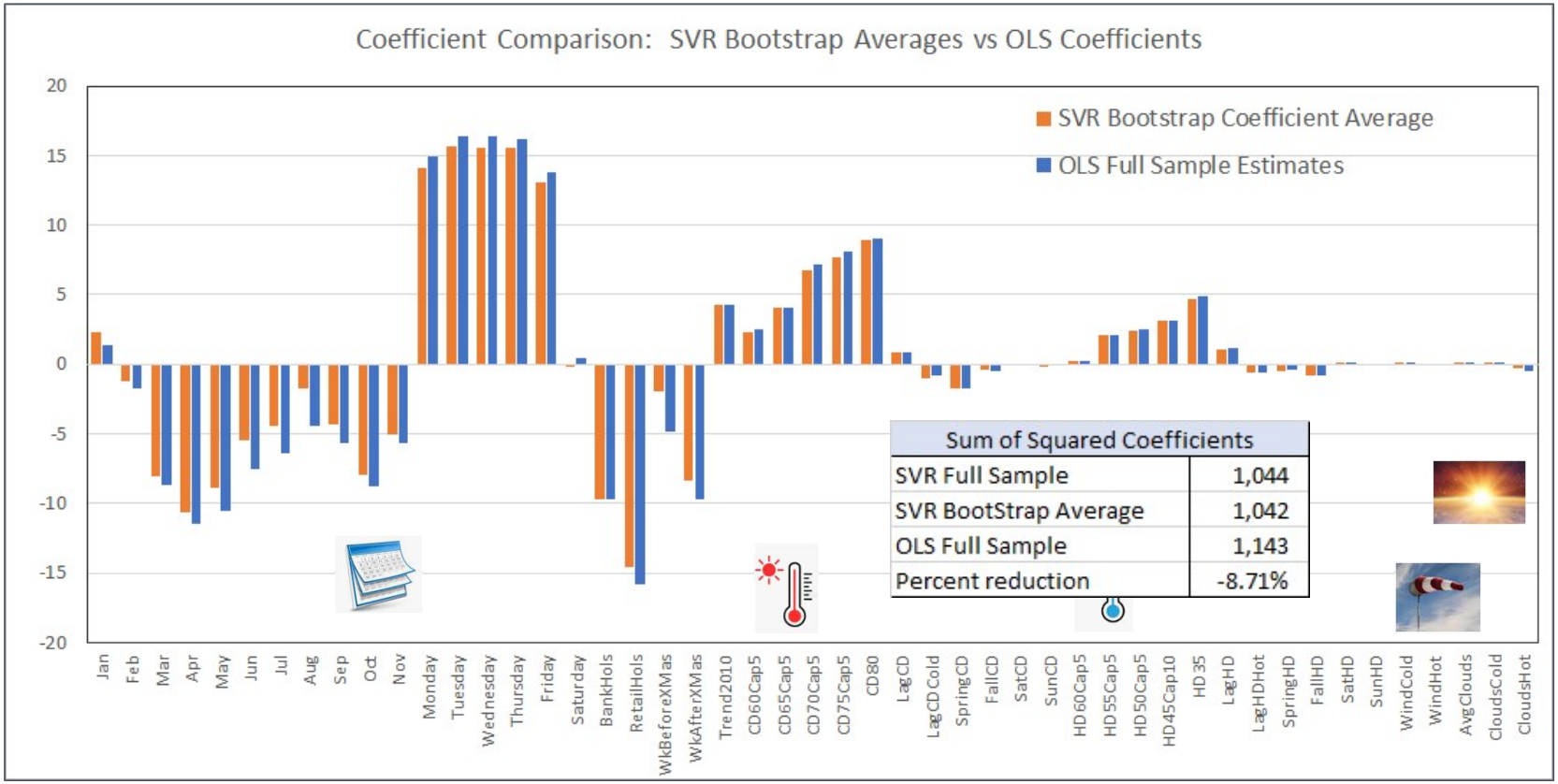
Coefficient Comparison – 1,000 Bootstrap Samples





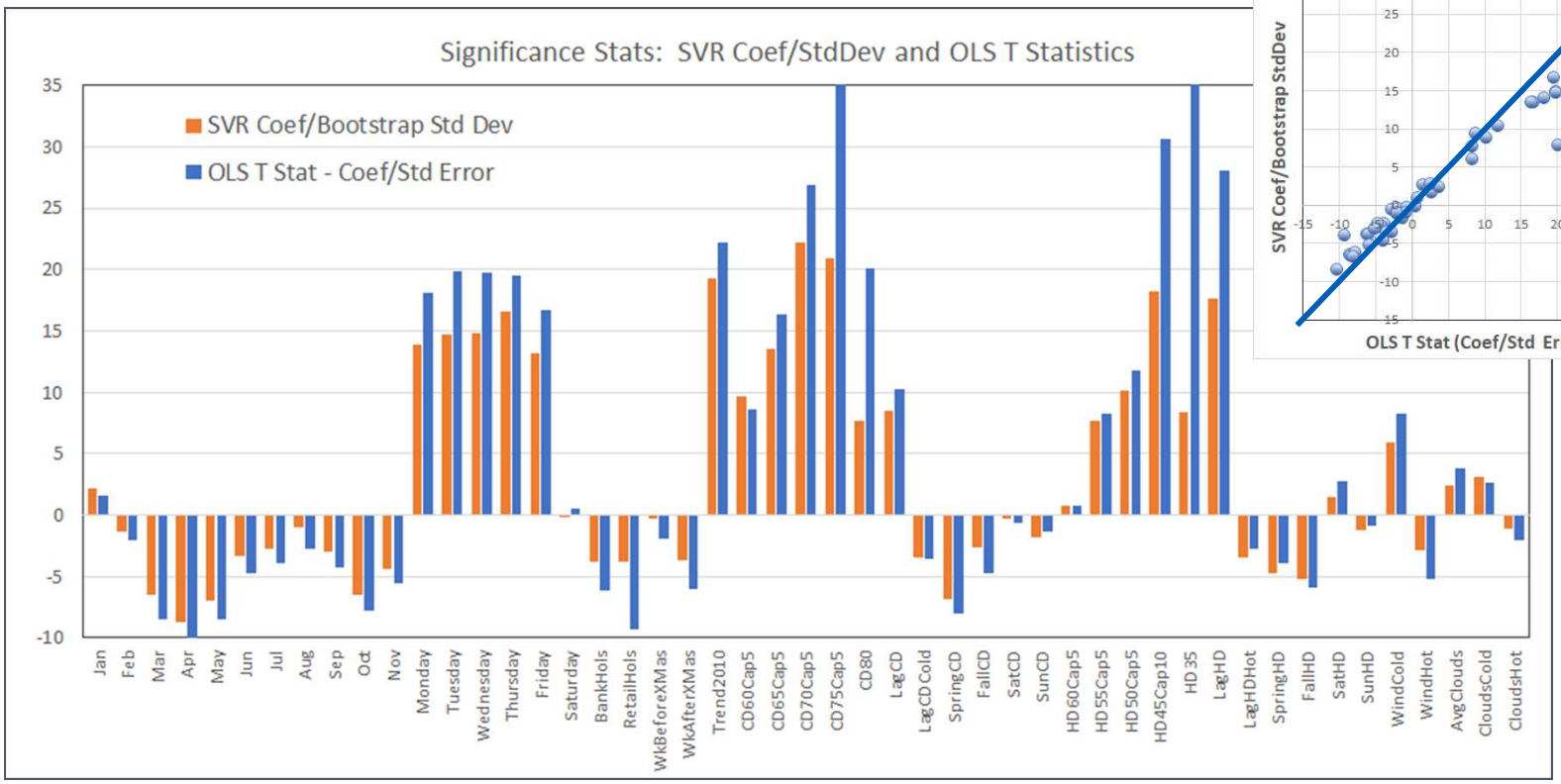
©2021 ITRON CONFIDENTIAL PROPRIETARY

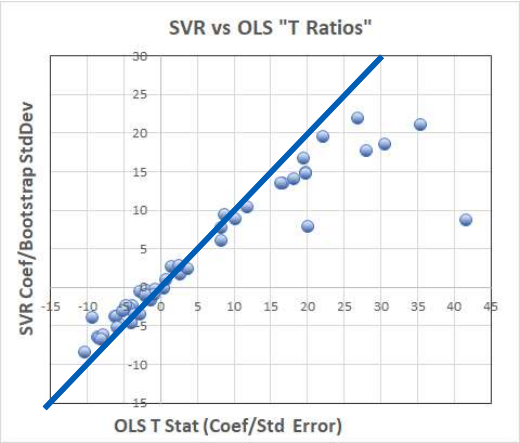
Compare SVR and OLS Coefficients



Sum of Squared Coefficients	
SVR Full Sample	1,044
SVR BootStrap Average	1,042
OLS Full Sample	1,143
Percent reduction	-8.71%

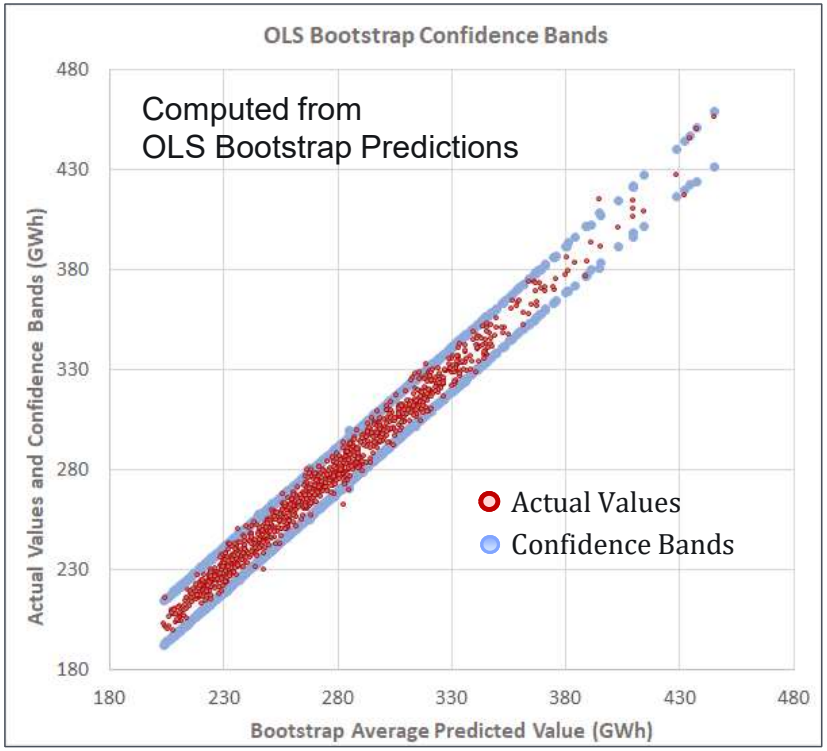
Compare SVR and OLS Coef Significance

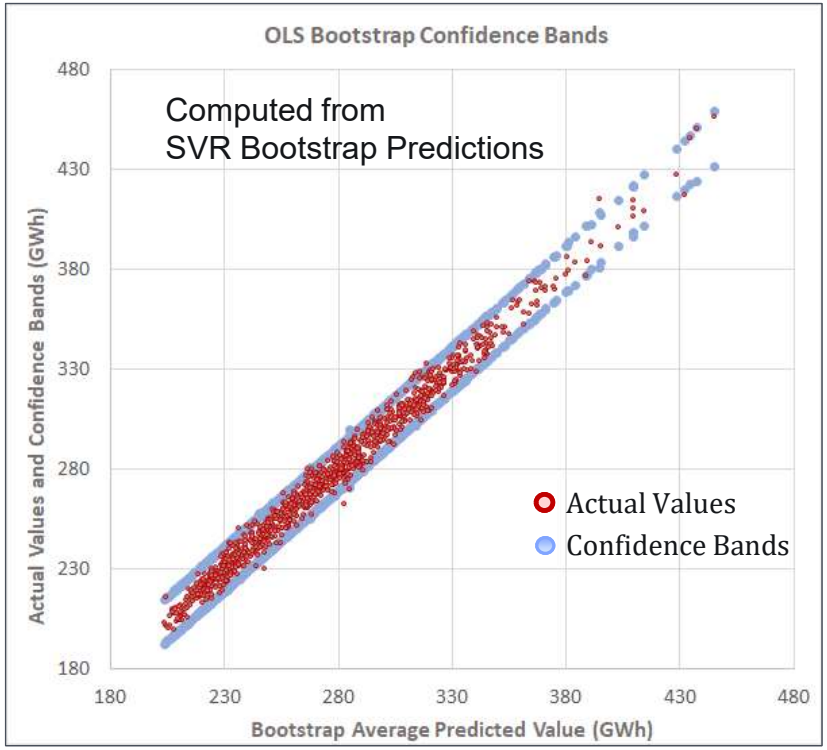




Compare SVR and OLS Bootstrap Confidence Bands

» SVR and OLS Bootstrap Confidence Bands are basically the same





©2021 ITRON CONFIDENTIAL PROPRIETARY

Conclusions

©2021 ITRON CONFIDENTIAL PROPRIETARY

Conclusions

» Support Vector Regression is a relatively new method

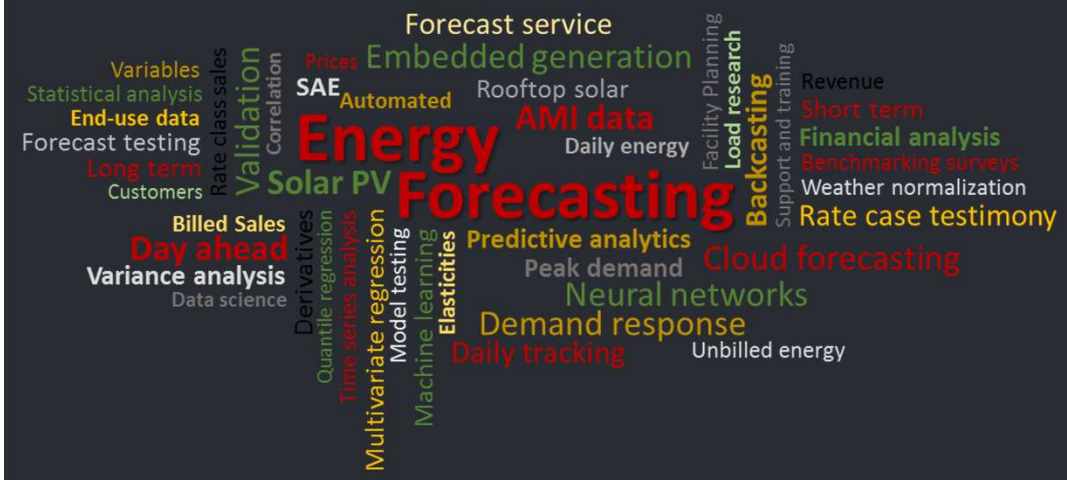
» It is different:

• Errors don't count for points close to the line (inside the epsilon tube)	OLS does not do this
• Errors do count for points outside the epsilon margins	OLS counts all errors
- Uses the sum of absolute distances from the margin	OLS uses squared errors
• Coefficient size is penalized (sum of squared coefficients)	OLS does not penalize
• You need to select hyper parameter values (C, ε) before estimation	OLS has no HPs
• You need to bootstrap to get standard errors and confidence bands	OLS has formulas

» For the data analyzed here:

- Overall performance is similar to OLS
- OLS out-of-sample fit is slightly better
- OLS significance estimates are a bit stronger
- OLS prediction intervals are a bit tighter
- For both methods, the important thing is to get the model specification right to capture nonlinearities and important interactions. **Feature engineering is everything**.

©2021 ITRON CONFIDENTIAL PROPRIETARY



Thank You

SAN DIEGO

Stuart McMenamin
stuart.mcmenamin@itron.com
858.724.2641
https://www.linkedin.com/in/stuart-mcmenamin

David Simons
david.simons@itron.com
858.724.2685
https://www.linkedin.com/in/davidsimons1

Itron

http://blogs.itron.com/forecasting/

www.itron.com