

Solving Time Series Problems

John Mount, Ph.D.
Win Vector LLC
jmount@win-vector.com
<https://www.win-vector.com>

<https://odsc.com/california/odsc-west-schedule-2024/>
Tutorials | Workshops, 30 Oct Day 2, 3:30 PM
(Hyatt Regency San Francisco Airport, 1533 Old Bayshore Highway, Burlingame, CA 94010)
All slides, code and data are shared here: <https://github.com/WinVector/Examples/tree/main/TimeSeries#readme>



1 Hold on this slide until ready to start.

Motivation

"One of the biggest mistakes in my data science career was not paying enough attention to time series models.

Time series is everywhere—finance, sales, marketing—and it can drastically improve your decision-making process."

— Mark Eltsefon, Staff Scientist at Meta
https://www.linkedin.com/posts/mark-eltsefon_datascience-timeseries-activity-7253363004694491137-94DQ?utm_source=share&utm_medium=member_desktop



2 Forecasting problems tend to be high value.

Outline

- Motivation
- Who I am
- What is time series forecasting?
- Families of methodologies
- Our example problem
- The liar's graph and out of sample evaluation
- Solving the problem using Stan
- What is Stan?
- Results/Observations



Luca Cambiaso FightingFigures



3 We will fill in some details as we go.

Who I am

- John Mount, General Partner at Win Vector LLC.
- Co-author of *Practical Data Science with R*, 2nd edition, Manning, 2020.
- Co-author of the `vttreat` R package for re-encoding high cardinality explanatory variables.
- Win Vector LLC is a statistics, machine learning, and data science consultancy and training organization.
- Specialize in solution design, technology evaluation, and prototyping.
- We help greatly speed up solution and production deployment.
- Looking for some more engagements!
- Please contact: jmount@win-vector.com.
- Please follow us on the Win Vector blog: <https://win-vector.com/blog-2/>



4

What is Time Series Forecasting?

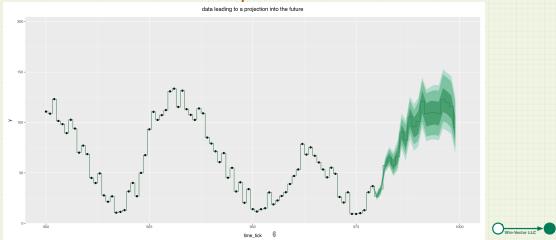
(a slightly heterodox view)

5

Time Series Forecasting

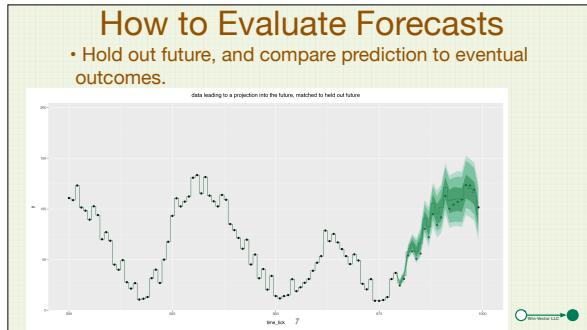
- Produce an estimate of the future using trends and relations observed in the past.

data leading to a projection into the future



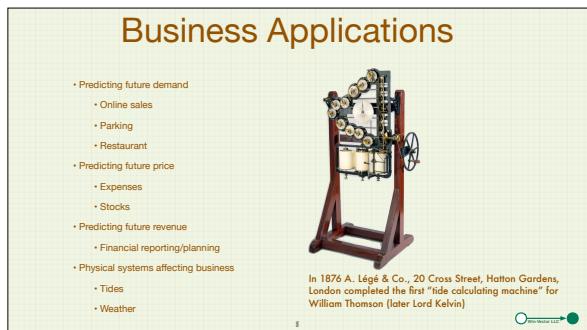
6

The dots are actual counts as a function of time, and the green line is the central prediction with uncertainty bands.

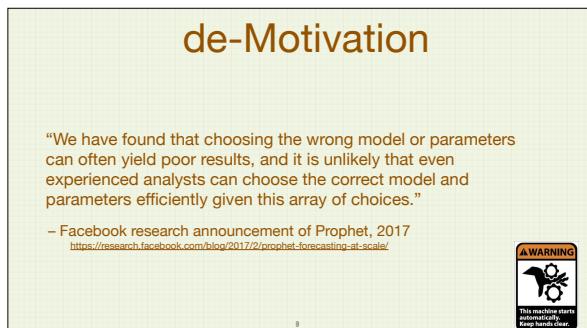


7

A prediction is good if it matches future data. Don't get bullied out of this.

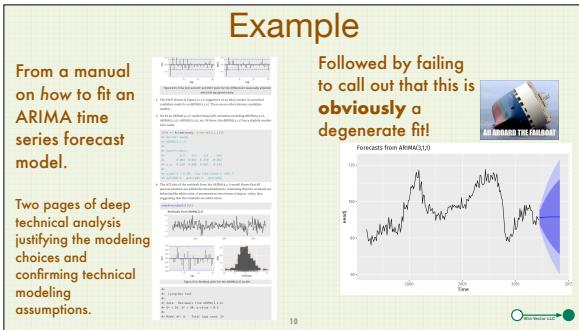


8

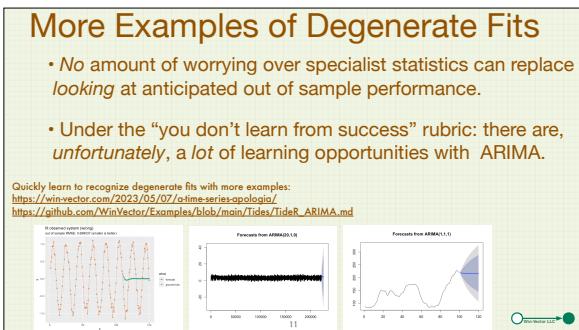


9

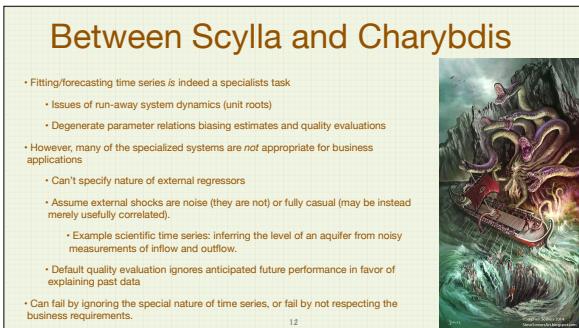
It is easy to get your hand caught in the gears.



At no point in the training period was the curve flat. But that is precisely the predicted expected value. Yes, models of expected tend to have less variance than actuals <https://win-vector.com/2024/06/15/good-models-do-not-match-variance/>. However, this is a degenerate fit. Most references avoid making the obvious by the simple expedient of never showing the final graph. Can't over emphasize the importance of insisting on method-agnostic hold out tests.



Each of these is a degenerate fit. Some are from “auto ARIMA” which claims to pick the right parameters. One of these is even from a package help page.



We have to sail between two dangers: under modeling and over-modeling.

Families of Methodologies

13



Methodology Families

- ARIMA derived methods
- Brute force methods
- Decomposition or attribution methods
- Hidden state inference

14



ARIMA derived methods

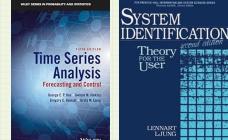
- "Auto regression integrated moving average"
- Fit for a conserved relation between observations
 - AR example $\hat{y}_t = \hat{E}[y_t] = 1.975y_{t-1} - y_{t-2}$ for many t
 - Push that forward as our prediction

$$\begin{aligned}\hat{y}_1 &= 1.975y_0 - y_{-2} \\ \hat{y}_{-1} &= 1.975y_1 - y_{-3} \\ \hat{y}_{-2} &= 1.975y_{-1} - \hat{y}_1 \\ \dots \\ y_{t+1} &= 1.975y_{t+k-1} - \hat{y}_{t+k-2}\end{aligned}$$
 - Can run away!
- Regression with ARIMA residuals:

$$(1 - \phi_1 B - \phi_2 B^2)(y_t - \beta_0 - \beta_1 x_t - \beta_2 z_t) = c + (1 - \theta_1 B - \theta_2 B^2)y_t$$
- Very powerful
 - Can dominate other solution method when only measuring loss or fit quality.
 - Highly technical (avoiding issues such as "unit roots").
 - Wrongly thought of as "the only way to solve time series problems."
 - https://github.com/WinVector/Examples/blob/main/TimeSeries/ts_example.md
 - https://github.com/WinVector/Examples/blob/main/TimeSeries/sm_example.ipynb

15

Among the most powerful methods. The “get k steps out by running a 1 step out process k times” is very powerful, but also very sensitive.



Brute force

- Treat as a standard machine learning problem
- Model $\hat{y}_{t+k} = \hat{E}[y_{t+k}] = f_k(y_t, \dots, y_{t-n})$ for $k = 1 \dots u$
- Fit for all the $f_k()$
 - Linear methods
 - <https://win-vector.com/2023/05/07/a-time-series-apologia/>
- Neural net methods such as LSTMs

https://github.com/WinVector/Examples/blob/main/TimeSeries/nested_model_example.ipynb

16

Not actually a bad idea. Misses a lot of the traps in forecasting by ignoring them. Kind of treated as “not allowed” until deep learning had great success with it. Low explainability even with linear methods, as we are a bundle of different models.

Decomposition or Attribution

- Fit past to things we know future values of:
 - Month of year ...
- Use the combination of the future values of these things as our future prediction.



https://github.com/WinVector/Examples/blob/main/TimeSeries/Prophet_example.ipynb

17

Fit the past to things we know the future of: such as seasonal averages. Use this fit to project into the future. Very useful in attributing or explaining past effects. Tends to under-perform ARIMA on model quality evaluations—but returns more explainable results and useful data decompositions.

Hidden State Inference

- Includes methods such as EM and particle filters
- Guess hidden state and parameters β, ζ

$$\text{Estimate } \hat{\beta} = \hat{E}[\beta] = \frac{\int_{\beta, \zeta} d\beta, \zeta P[\text{training data} | \beta, \zeta] \beta}{\int_{\beta, \zeta} d\beta, \zeta P[\text{training data} | \beta, \zeta]}$$

https://github.com/WinVector/Examples/blob/main/TimeSeries/Stan_soln.ipynb

18

Computationally expensive, but very versatile. This is the method we will pursue in this session.

Opinion

- Good time series and system identification solutions combine a few ideas from a few of these methodological themes.
 - Many of the methodologies simplify if you delegate the solution to a systematic solver.
- Most of the solving methods are excessively technical and restrictive.
 - There are in fact issues with time series forecasting require expert packages.
 - If the package wasn't built for your application, you may not be able to adapt it to your application. You are instead forced to adapt your goals to the package.
- A case where there are more packages in Python, but more good packages in R.



19

Notice that none of the methods are “just a linear regression”

Our Example Problem

20

External Regressors

- For many business applications we need to model future values as a function of past values and additional facts called “external regressors.”
- This is where it all goes wrong.
 - ARMAX, ARIMAX, SARIMAX, transfer functions, regression with ARIMA residuals all claim to solve this.
 - No two of them seem to be solving the same problem (if you can even find a description of what problem they claim to solve).
 - <https://win-vector.com/2024/07/15/arimax-offerings-remain-a-muddle/>
 - The ARIMAX model muddle <https://robjhyndman.com/hyndtsight/arimax/>
- To the extent time series modeling is magic, it obscures out the effects of external regressors.

21

There is a fiction that ARMAX is how time series researches work. That appears to be wishful thinking. Time series research moved on to transfer functions and other terms of art.

Our Notional Example Problem

22

- We are modeling a restaurant
- We imagine number of diners on a given day is related to number of guests in the last two days.
 - These are called "lags" or the auto-regressive portion of the model.
 - In practice we would also add most recent same day of week as a lag.
- Improve prediction and utility of prediction by bringing in information
 - We have transient external regressors such as events being near our restaurant.
 - We have durable external regressors such as price changes and restaurant reviews.
 - Usually also add a few more easy external regressors to represent seasonality (day of week, month of year, and so on).
- Can also model this as two sub-populations of guests: loyal and transient
 - The issue is: these subpopulations are not labeled in our training data!



22

As Equations

23

This immediately requires bi-temporal notation: what was known about date A on date B. We will use "date" and "time" as synonyms in this talk.

$$\begin{aligned} \text{loyal_guests}_{\text{date}} &= \beta_{\text{loyal}} + \beta_1 \text{loyal_guests}_{\text{date}-1} + \beta_2 \text{loyal_guests}_{\text{date}-2} + \beta_{\text{review}} x_{\text{review}, \text{date}-1} \\ \text{transient_guests}_{\text{date}} &= \beta_{\text{imp}} + \beta_{\text{events}} x_{\text{events}, \text{date}} \\ \text{total_guests}_{\text{date}} &= \text{loyal_guests}_{\text{date}} + \text{transient_guests}_{\text{date}} \end{aligned}$$

- These equations are our specification of the problem. Under different assumptions we would write different equations.
- Subtle point: we are modeling realized values, not the traditional expected values.
- We only observe **total_guests** (not **loyal_guests** or **transient_guests**).
- In practice we benefit from a *lot* more external regressors- such as reservations by date.
- Without controllable external regressors forecasting devolves into Cassandra's curse.



23

The Equations in Detail

24

- $$\begin{aligned} \text{loyal_guests}_{\text{date}} &= \beta_{\text{loyal}} + \beta_1 \text{loyal_guests}_{\text{date}-1} + \beta_2 \text{loyal_guests}_{\text{date}-2} + \beta_{\text{review}} x_{\text{review}, \text{date}-1} \\ \text{total_guests}_{\text{date}} &= \text{loyal_guests}_{\text{date}} + \beta_{\text{imp}} + \beta_{\text{events}} x_{\text{events}, \text{date}} \end{aligned}$$
- $x_{\text{review}, \cdot}, x_{\text{events}, \cdot}$ features engineered by the data scientist.
 - Notice we only know about reviews in the past, but claim to know about events in the future.
 - $\text{total_guests}_{\text{date}}$ observed
 - $\text{loyal_guests}_{\text{date}}$ unobserved (to be inferred)
 - Classic Bayesian set up. Our population is a mixture of unlabeled sub-populations with different behaviors. Could divide into even more sub-populations if we cared.
 - $\beta_{\text{loyal}}, \beta_{\text{imp}}, \beta_1, \beta_2, \beta_{\text{review}}, \beta_{\text{events}}$ the linear model coefficients to be inferred
 - We call $\beta_{\text{loyal}}, \beta_1, \beta_2$ the "durable" auto-regressive portion of the model.
 - We call β_{review} a "durable" effect, associated with a durable external regressor.
 - We call β_{events} a "transient" or "impermanent" effect, associated with a transient external regressor.

24

How can we know events in the future?

- Buy the information from somebody that curates records about future events.



(Not an endorsement, I just used them for a couple of clients.
Some notes of mine on problems in using past predictions of the future:
<https://win-vector.com/2024/09/09/please-version-data/>)

25

25

Can Standard Packages Solve This?

$$\text{loyal_guests}_{\text{date}} = \beta_{\text{loyal}} + \beta_1 \text{loyal_guests}_{\text{date}-1} + \beta_2 \text{loyal_guests}_{\text{date}-2} + \beta_{\text{review}} x_{\text{review}, \text{date}-1}$$

$$\text{total_guests}_{\text{date}} = \text{loyal_guests}_{\text{date}} + \beta_{\text{imp}} + \beta_{\text{events}} x_{\text{events}, \text{date}}$$

- If we force $\beta_{\text{review}} = 0$, then this is “regression with ARIMA residuals” variation of ARIMAX.
- If we force $\beta_{\text{events}} = 0$, then this is the sort of system social scientists want to study policy changes (such as tariffs or seatbelt laws). Also called an ARMAX variation (may or may not in fact be).
- None of the ARIMA style packages we surveyed offered structural choice or control. The equations the package solves either match yours, or do not.

26

26

An Artificial Example



27

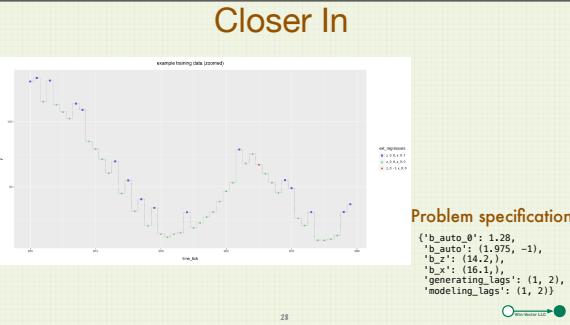
27

Time series methods tend to be really good at modeling $\sin(x)$. So if they are not good and noised up copies of $\sin(x)$, what good are they going to be on brutal real world data?

No real world data is ever this easy, so using the methodology does take some engineering (adding more lags, feature engineering and so on).

Closer In

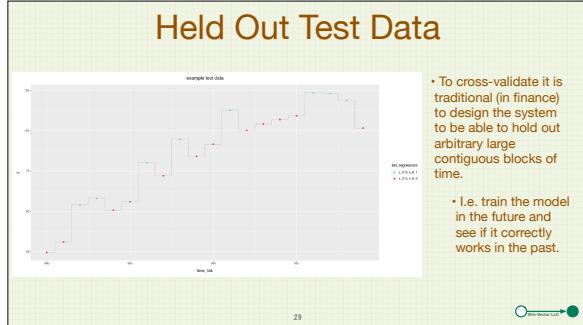
28



b_z is our data generating durable effect coefficient, and b_x is our transient effect coefficient. We generate the data and then see if the system can recover these coefficients from unlabeled data.

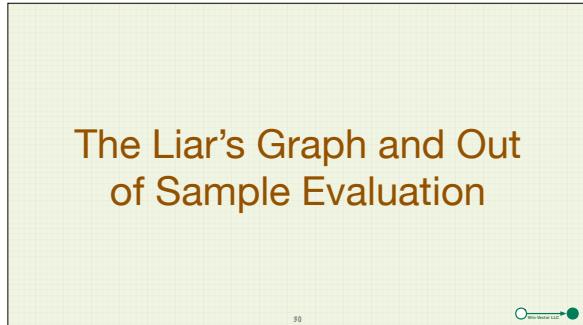
Held Out Test Data

29



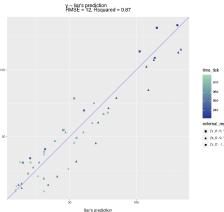
The Liar's Graph and Out of Sample Evaluation

30



The Liar's Graph

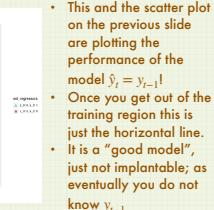
- If you take away one thing today, please take away this.
- Most common scatter plots of “time series performance” plot how well $\hat{y}_t \sim y_t$
- This is often “one tick out performance” even if your application requires predicting many time periods out!
- Do NOT get conned by the complicated diagnostics and the scatter plot. Insist on seeing the model applied.



31

Look at Out of Sample Predictions!

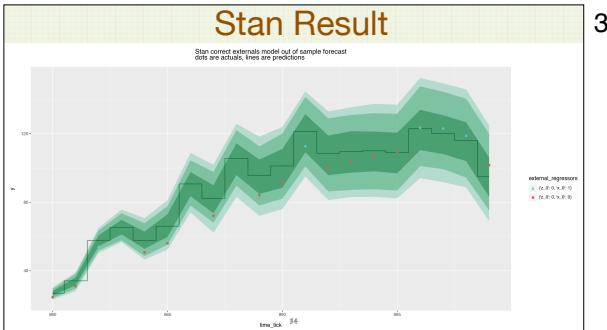
- This and the scatter plot on the previous slide are plotting the performance of the model $\hat{y}_t = y_{t-1}$.
- Once you get out of the training region this is just the horizontal line.
- It is a “good model”, just not implantable; as eventually you do not know y_{t-1} .



32

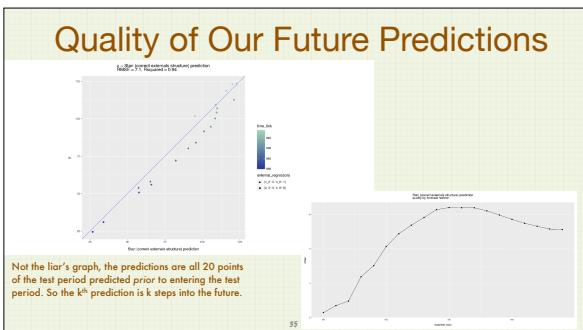
Solving the Problem

33

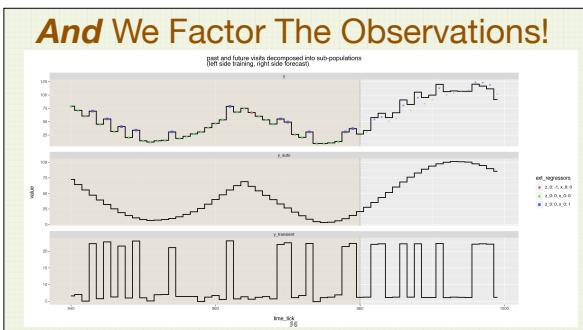


34

We get predictions, and they are indeed close the future realized actuals (not seen during training or forecasting). In this, all predictions are made with only information available at time = 979.



35

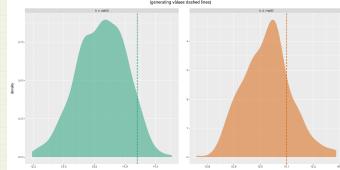


36

Even the historic factorization can be valuable to the business. y_{auto} is the loyal or durable customers, $y_{\text{transient}}$ are the transient or impulse customers, and y is the sum of the y_{loyal} and $y_{\text{transient}}$. Only y is observed, and only in the training region.

And We Get Coefficient Estimates!

Coefficient estimates:



Generating actuals:

```
{'b_auto_0': 1.2804125781056719,  
'b_imp_0': 10.4,  
'b_auto': (1.9753766811982755, -1),  
'b_z': (14.2,),  
'b_x': (16.1,),  
'generating_tags': (1, 2),  
'modeling_tags': (1, 2)}
```

37

Again even estimating b_z (the durable external regressor effects) and b_x (the transient external regressor effects) can be very valuable.

How We Solve

- Copy our equations into "Stan"

into "Stan"

- And "turn the crank."



```
y_auto[3:(N_y_observed + N_y_future)] ~ normal(  
b_auto_0  
+ b_auto[1] * y_auto[2:(N_y_observed + N_y_future - 1)]  
+ b_auto[2] * y_auto[1:(N_y_observed + N_y_future - 2)]  
+ b_x_dur[1] * x_dur_1[3:(N_y_observed + N_y_future)],  
b_var_y_auto);  
y ~ normal(  
y_auto  
+ b_imp_0 + b_x_imp[1] * x_imp_1,  
b_var_y);
```

Complete workbook: https://github.com/WinVector/Examples/blob/main/TimeSeries/Stan_sol.ipynb

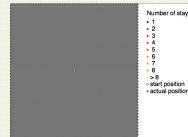
38

What is Stan?

39

Stan

- A Markov Chain Monte Carlo sampler.
- Can be used for complicated Bayesian inference
 - Guesses likely values of parameters, and nuisance variables, and unobserved intermediate state conditioned on observed data.
 - Returns distributional answers.
- Subsumes the implementation portion of many other systems/ideas (such as “particle filters” and “importance sampling”).
- Fairly high dependency (requires C++ compiler and linker)
- May eventually see competition from Torch based alternatives



40

40

What We Did

- Asked Stan to use hidden state methods to solve an ARIMAX style formulation of the problem.
- We directly controlled how different regressors last over time.
- We directly specified two sub-populations of customers (with different behavior) that we only observe the sum of.
- We asked Stan to “un-stir” the mixture.

41

41

Why it all Works

- We want to infer parameters β and hidden state ζ given data. That is pick β, ζ such that $P[\beta, \zeta | data]$ is maximal.
- By Bayes' Law $P[\beta, \zeta | data] = P[\beta, \zeta]P[data | \beta, \zeta]/P[data]$.
 - Can ignore $P[data]$ as it is free of our parameter estimates.
 - Therefore: picking β, ζ such that $P[\beta, \zeta]P[data | \beta, \zeta]$ is large is the same as picking such that $P[\beta, \zeta | data]$ is large.
- Structural mis-specification much more risky than “wrong priors” when we have a lot of training data.

42

(splash slide, move from quickly)



42

The Stan Solution in Detail

https://github.com/WinVector/Examples/blob/main/TimeSeries/Stan_soln.ipynb

43

(may or may not go into the workbook)

43



What we get from Stan

b_0	accept_stat	stepsize	trenddepth	n_leapfrog	divergent	energy	b_warmup_0	b_warmup_0	lp_warmup_0	y[0]	y[0]	y[0]	y			
0	2722.44	0.951920	0.000706	0.0	0	-2239.00	1.70910	6.79484	1.94931	...	103.270	121.744	109.826	112		
1	2748.47	0.871088	0.000706	10.0	0	-1023.0	0.0	2255.93	1.74729	6.52748	1.93935	...	108.3290	125.114	106.9760	104
2	2732.40	0.823397	0.000706	10.0	0	-2239.06	1.64489	5.99184	1.93849	...	108.0620	127.683	113.549	116		
3	2726.63	0.786470	0.000706	10.0	0	-2232.23	1.73269	5.97778	1.93883	...	119.6260	135.204	117.8630	118		
4	2699.50	0.863398	0.000706	10.0	0	-2225.57	1.65860	5.87227	1.94142	...	119.229	145.286	125.279	121		
...			
395	2746.19	0.870762	0.020688	8.0	0	-2150.90	1.65880	5.26812	1.93254	...	78.4539	86.6416	82.7909	84		
396	2725.33	0.844742	0.020688	8.0	0	-2054.02	0.0	1850.00	1.75150	5.64120	1.94312	...	84.8620	94.520	86.520	86
397	2703.35	0.846784	0.020688	8.0	0	-2158.27	1.75156	5.64338	1.94312	...	73.0254	85.1592	81.5279	80		
398	2693.27	0.765531	0.020688	8.0	0	-2190.17	1.60223	5.64163	1.93846	...	85.9863	107.9670	95.5999	101		
399	2703.47	0.871172	0.020688	8.0	0	-2050.56	1.78176	5.75108	1.93982	...	108.5280	131.6860	120.1980	122		

400 rows x 2035 columns

44



Each row is a sample. The columns we are interested in for a given row are the parameters (“b_ *”) and hidden state (“y[**]”, “y_auto[**]”, “y_future[**]”). One thing to keep in mind: the estimated state and parameters in a given row are mutually consistent as they were picked with the plausibility $(\exp(lp_{\text{row}}) / Z, \text{ for some } Z)$ high. It is a domain question if and when we can average these across rows.

Our Thesis

In business forecasting: ability to specify actionable structure beats technique in forecasting.

45

By actionable structure I mean properly structured external regressors that are under our control.

45



Follow up resources

46

- This talk
 - <https://github.com/WinVector/Examples/tree/main/TimeSeries#readme>
- Some of the Perils of Time Series Forecasting
 - <https://win-vector.com/2023/05/25/some-of-the-perils-of-time-series-forecasting/>
- A Time Series Apologia
 - <https://win-vector.com/2023/05/07/a-time-series-apologia/>
- Please Version Data
 - <https://win-vector.com/2024/09/09/please-version-data/>
- The `vtreat` data preparation system
 - <https://github.com/WinVector/pyvtreat>
 - <https://github.com/WinVector/vtreat>

46



Thank you

47

All materials: <https://github.com/WinVector/Examples/tree/main/TimeSeries#readme>

47



Part of Our Using Stan to Solve Problems Training Offering

48

(stop on this slide)

- The series currently includes:
 - [Dealing with range censored data, or tobit style regression.](#)
 - [Learning rank preferences from observed actions.](#)
 - [Time series with external explanatory variables.](#)
- Please contact jmount@win-vector.com for custom data science research, training and consulting.

All materials: <https://github.com/WinVector/Examples/tree/main/TimeSeries#readme>

48



Appendices

ARIMAX Solutions

- https://github.com/WinVector/Examples/blob/main/TimeSeries/ts_example.md
- https://github.com/WinVector/Examples/blob/main/TimeSeries/sm_example.ipynb

Just going to show the R versions. Or skip.

(possibly skip)

A Few Things From Experience

- At first it feels like forecasting software binding fitting and prediction together is a mistake.
 - Violates the very useful `sklearn` separate `.fit()` and `.predict()` API by forcing a `.fit_predict()` pattern.
 - However, part of forecast inference is to also estimate un-observed details of past state. So fitting is in fact not separate from forecasting.
- And you may want to re-run with different values of future external regressors under your control.
 - Such as offering a discount on a day that is predicted to have low attendance.
- Our own `vtaest` high cardinality variable re-encoding package uses the same restriction to prevent over fitting.
- Stan great for prototyping
 - Not forced to use full Stan methodology in production.
- Can, in many cases, export inferences from Stan for use by simpler implementations.

<https://win-vector.com/2019/07/03/replicating-a-linear-model/>

52

(skip)