Out of controls

Hal Varian

Incorporating work by Steve Scott,
Kay Brodersen and Art Owen

European Central Bank
June 2018

Opinions expressed in these slides are solely those of the author.

Outline

- Randomized experiments are gold standard for causality
- But controls can be costly
  - Opportunity cost
  - Actual costs
- Control groups are only one way of estimating counterfactual
- This talk is about alternatives:
  - Synthetic controls
  - Regression discontinuity

Motivating problem

An advertiser contemplates changing its (bid, budget, creative) and wants to know what will happen to some performance measure (clicks, revenue, conversions). Solution: run an experiment.

Some possible designs

- Apply treatment for some users, compare to non-treated users
- Apply treatment for some geos, compare to non-treated geos
- Apply treatment for some advertisers compare to similar advertisers that did not get treatment
- Apply treatment and compare actual to *prediction* of would have happened without the treatment (interrupted regression, synthetic control)
- Last method is nice since don't have to explicitly manage controls
- We want an "automatic" way to build predictive models for counterfactual
    - Time series methods to model seasonality and trend
    - Regression methods incorporate contemporaneous predictors

Time series + regression component

Steve Scott

Kalman filter for time series component

- Handles trend and seasonality
- Bayesian-friendly
- Andrew Harvey [1989], Durbin and Koopman [2012]

Spike-and-slab for regression component

- George-McCulloch [1997]); Madigan-Raftery [1994]
- Probability variable is included in regression (spike)
- Probability distribution over coefficient value (slab)
- Sample from simulated posterior, average to get point prediction
- See Scott and Varian (2013, 2014) for details

Steve Scott's "Bayesian Structural Time Series"

- Download R package from CRAN (`BoomSpikeSlab, bsts`)

Modeling with Bayesian structural time series

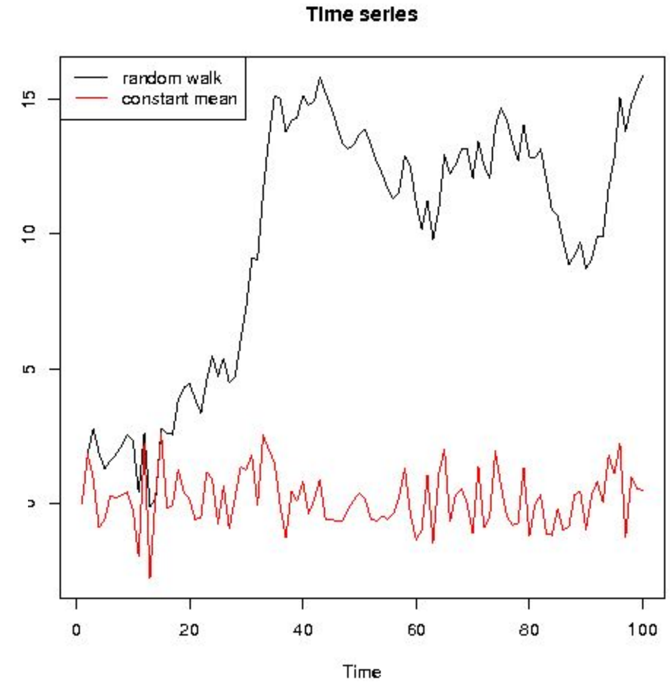
Model = Trend + Seasonal + Regression

- Trend and Seasonal components are adaptive and nonparametric.
- Can add other model components as necessary (e.g. holidays, weather).
- The regression component uses Google Trends as predictors.
  - Could use other predictors as well
- Accounting for the time series structure of the problem is important.

One slide tutorial on Kalman forecasting

Two important time series models:

$$y_t = y_{t-1} + e_t \qquad \text{random walk}$$
$$y_t = \mu_0 + e_t \qquad \text{constant mean}$$



Both have obvious predictors. Kalman models the in-between cases...

$$y_t = \mu_t + v_t \qquad v_t \sim N(0, V) \quad \text{observation}$$
$$\mu_t = \mu_{t-1} + w_t \qquad w_t \sim N(0, W) \quad \text{state}$$

- constant mean: $W = 0$ and $V > 0$
- random walk: $W > 0$ and $V = 0$

Best predictor for $m_t = E\mu_t$ for in-between model turns out to be:

$$m_t = k_t y_t + (1 - k_t) m_{t-1}$$

$$m_t = m_{t-1} + k_t (y_t - m_{t-1})$$

$$k_t = \frac{k_{t-1} V + W}{k_{t-1} V + W + V}$$

Basic structural model

$$
\begin{aligned}
y_t &= \mu_t + \gamma_t + v_t & v_t &\sim N(0, V) & &\text{observation} \\
\mu_t &= \mu_{t-1} + b_{t-1} + w_{1t} & w_{1t} &\sim N(0, W_1) & &\text{level} \\
b_t &= b_{t-1} + w_{2t} & w_{2t} &\sim N(0, W_2) & &\text{slope} \\
\gamma_t &= -\sum_{s=1}^{S} \gamma_{t-s} + w_{3t} & w_{3t} &\sim N(0, W_3) & &\text{seasonal}
\end{aligned}
$$

Stochastic generalization of constant trend model: $y_t = \mu + bt + \gamma_t + v_t.$

We add a regression component: $y_t = \mu_t + \gamma_t + X_t\beta + v_t$

But not just any old regression….

Spike and slab regression for variable selection

George and McCulloch (1997)

- We think most elements of β are zero.
- Let $\gamma_j = 1$ if $\beta_j \neq 0$ and $\gamma_j = 0$ if $\beta_j = 0$.

$$\gamma = (1, 0, 0, 1, \cdots, 1, 0, 0)$$

- Now factor the prior distribution

$$p(\beta, \gamma, \sigma^{-2}) = p(\beta_\gamma | \gamma, \sigma^2) p(\sigma^2 | \gamma) p(\gamma)$$

$$\gamma \sim \prod_j \pi_j^{\gamma_j} (1 - \pi_j)^{1-\gamma_j} \qquad \text{“Spike”}$$

$$\beta_\gamma | \gamma, \sigma^2 \sim \mathcal{N}\left(b_\gamma, \sigma^2 \left(\Omega_\gamma^{-1}\right)^{-1}\right) \qquad \text{“Slab”}$$

$$\frac{1}{\sigma^2} \sim \Gamma\left(\frac{df}{2}, \frac{ss}{2}\right)$$

Estimation by MCMC sampler

1. Draw Kalman state variables (level, slope, etc) and regression parameters γ and β from posterior
2. Draw variances of the state components
3. Repeat a few thousand times
4. Results
    a. Posterior distn of variances in state components
    b. Posterior distn of inclusion for each predictor γ
    c. Posterior distn of coefficient values β
    d. Posterior distn of forecast y_t
5. Point estimates are averages over these distributions
6. Model uncertainty is natural due to huge number of possible models

Estimating a model using bsts

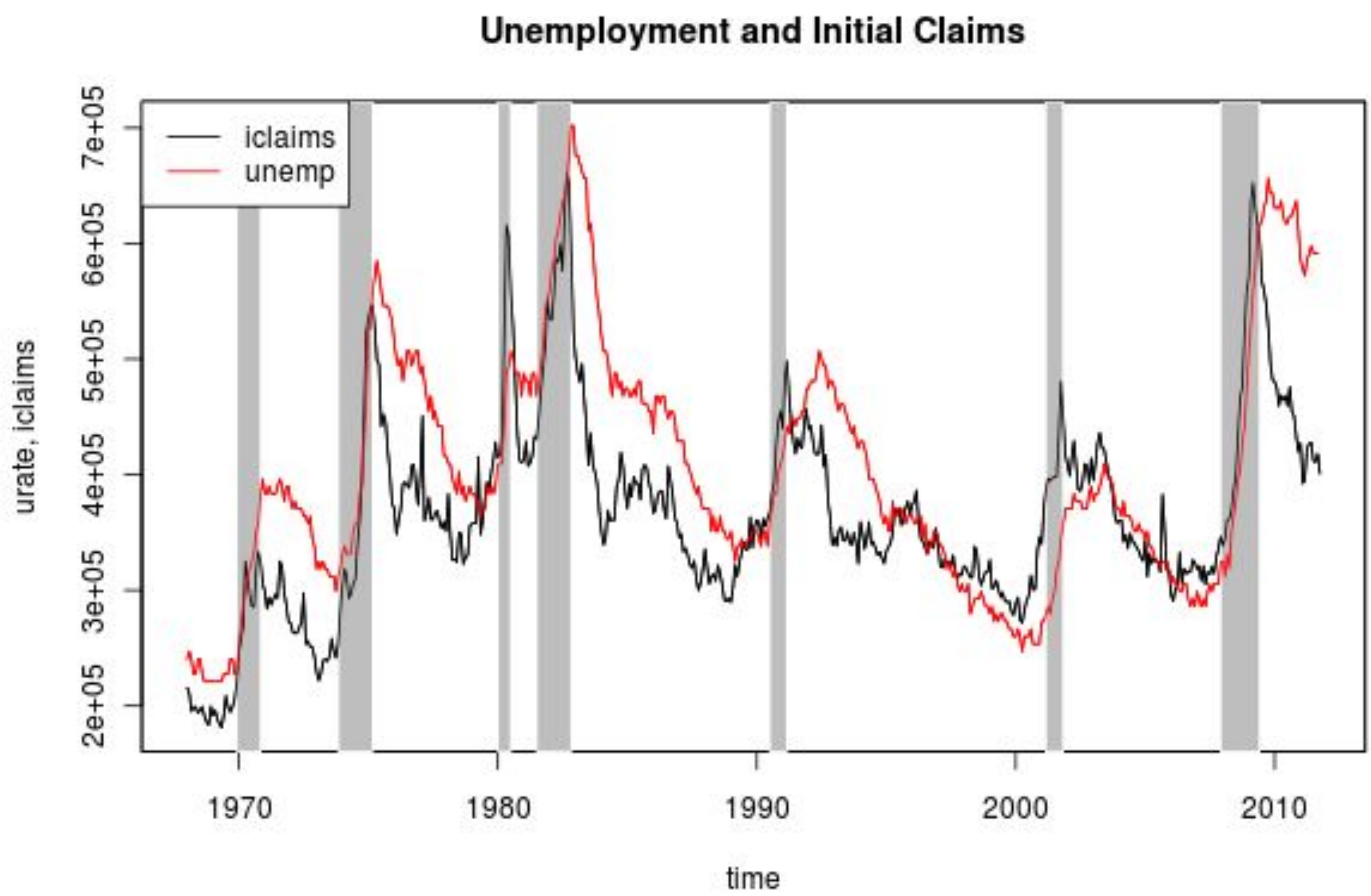
```
y <- my.data$ResponseVariable

ss <- AddLocalLinearTrend(
    list(),     ## No previous state specification.
    y)          ## Peek at the data for scaling.

ss <- AddSeasonal(
    ss,             ## Adding state to ss.
    y,              ## Peek at the data for scaling.
    nseasons = 7)   ## 7 "seasons" for day of week effect

model <- bsts(y ~ .,                   ## regression formula like 'lm'
              state.specification = ss,  ## time series spec
              niter = 1000,              ## MCMC iterations
              data = my.data,
              expected.model.size = 1)   ## spike-slab
```

Example: initial claims for unemployment benefits



Grey bars indicate recessions

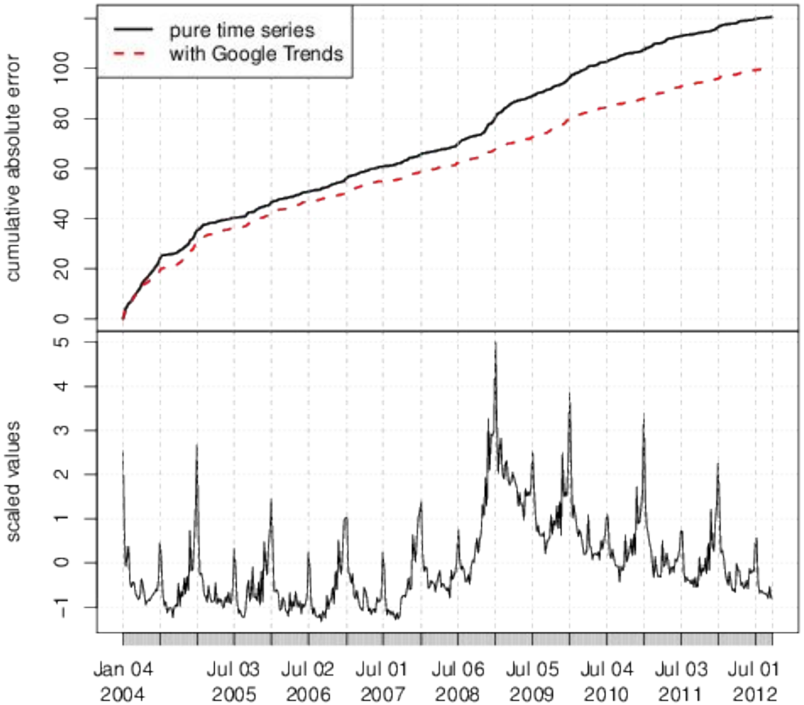
Does Google data help explain initial claims?

Cumulative absolute 1-step-ahead prediction errors

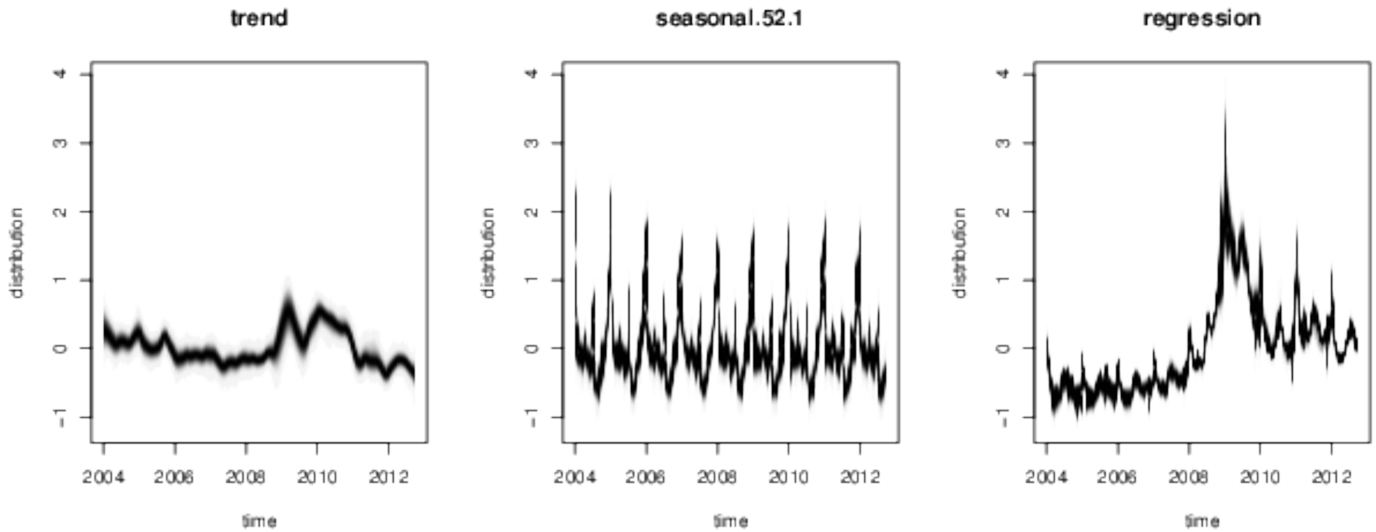
Plain time series model vs. model that includes Google Trends (correlate) data.

Models perform about the same when things are "boring".

Google Trends helps the model react to sudden changes (e.g. the recession).



What parts of the model did the explaining?



- The output is a "dynamic distribution" that shows the uncertainty around each component’s contribution.
- The regression help prediction

Which variables were important in the regression?

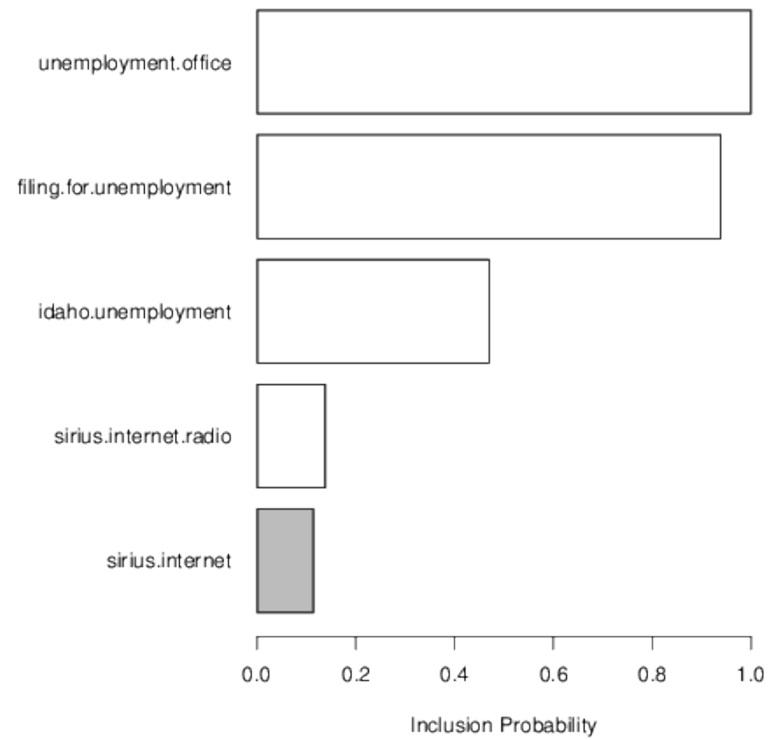
The important search terms:

- unemployment office
- filing for unemployment
- Idaho unemployment
  (Almost any state would do, but Idaho is slightly more predictive than the others in these data).

Other 95 search terms have inclusion probability less than 0.1.

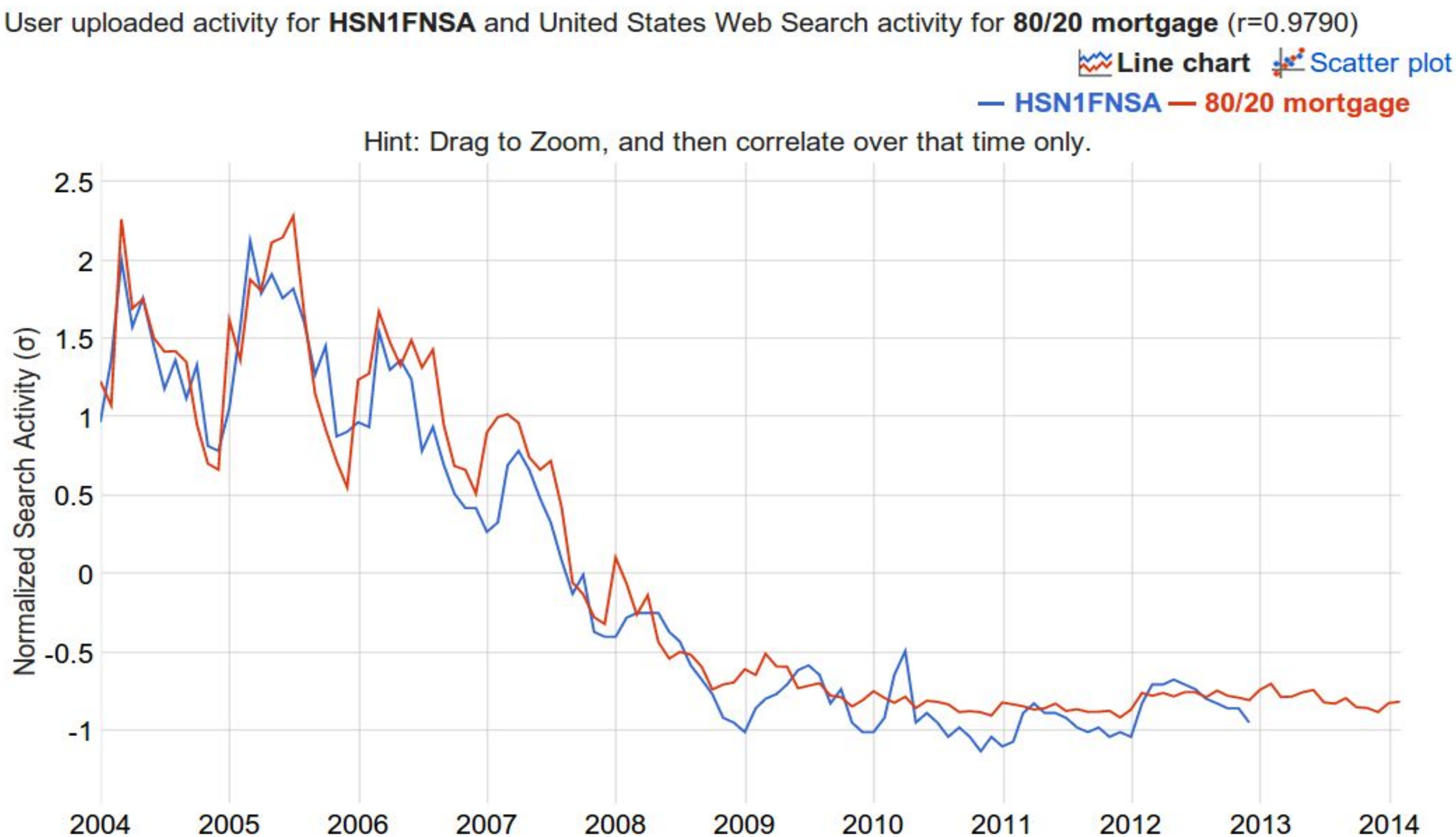
The predictions from this model average over which variables are in/out.

Sirius internet radio is spurious.

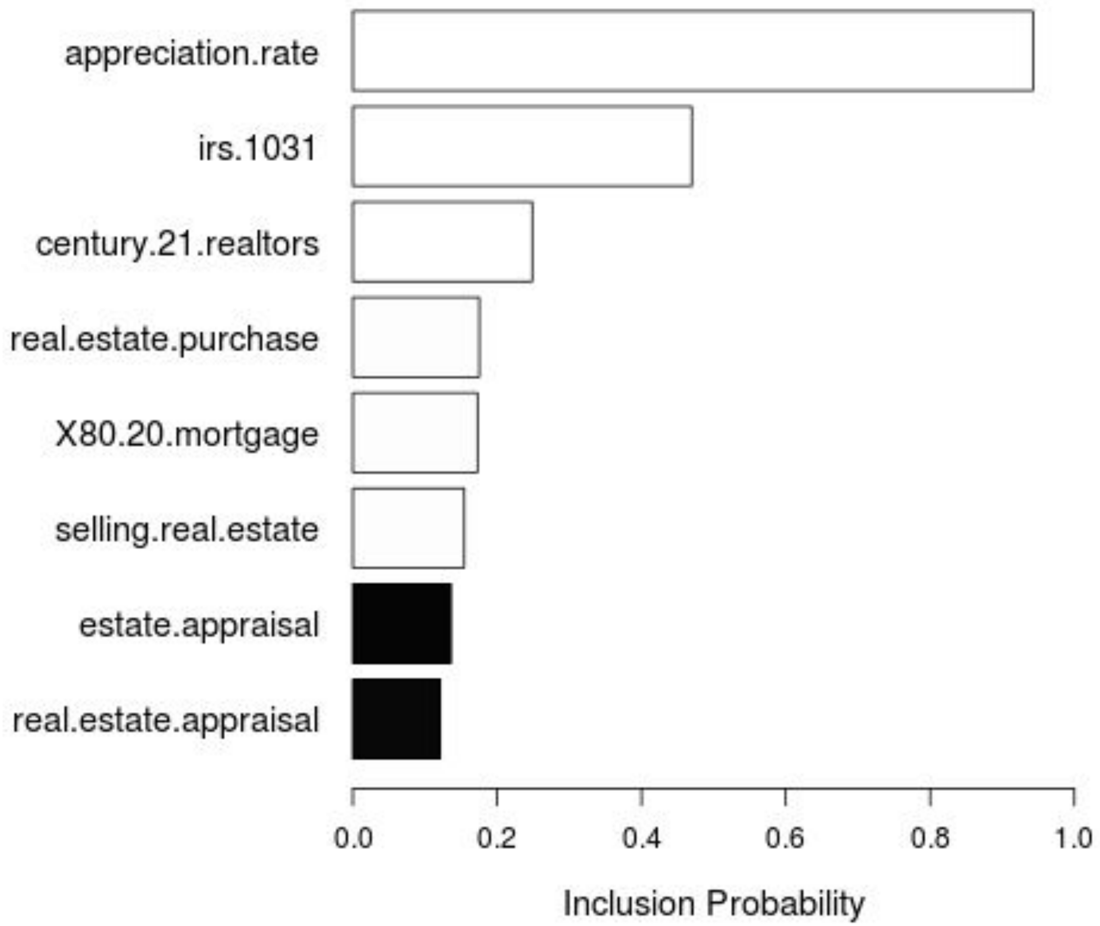


New Home Sales in US

Raw correlation



Predictors chosen by model



Incremental fit plots

Visualize how much each predictor contributes to model fit

model: $y_t = trend_t + seasonal_t + b_1\ x_{1t} + b_2\ x_{2t}$

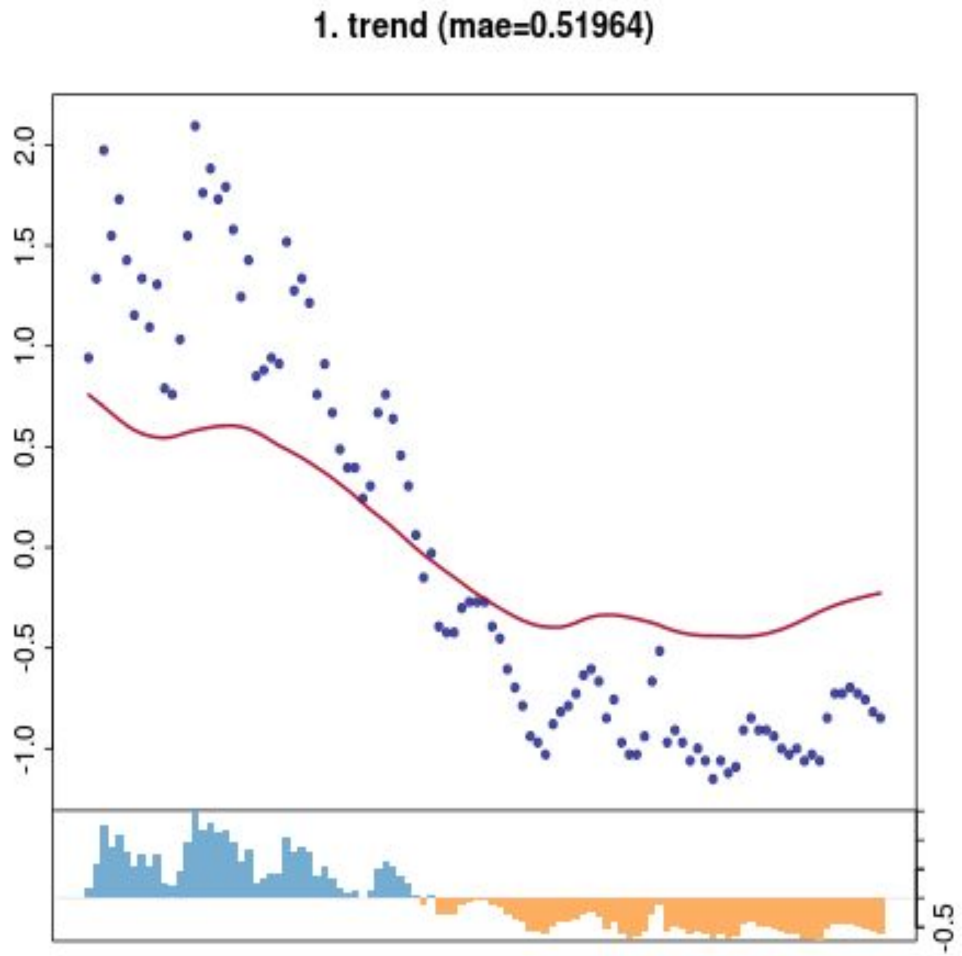
plot1: $y_t = trend_t$

plot2: $y_t = trend_t + seasonal_t$

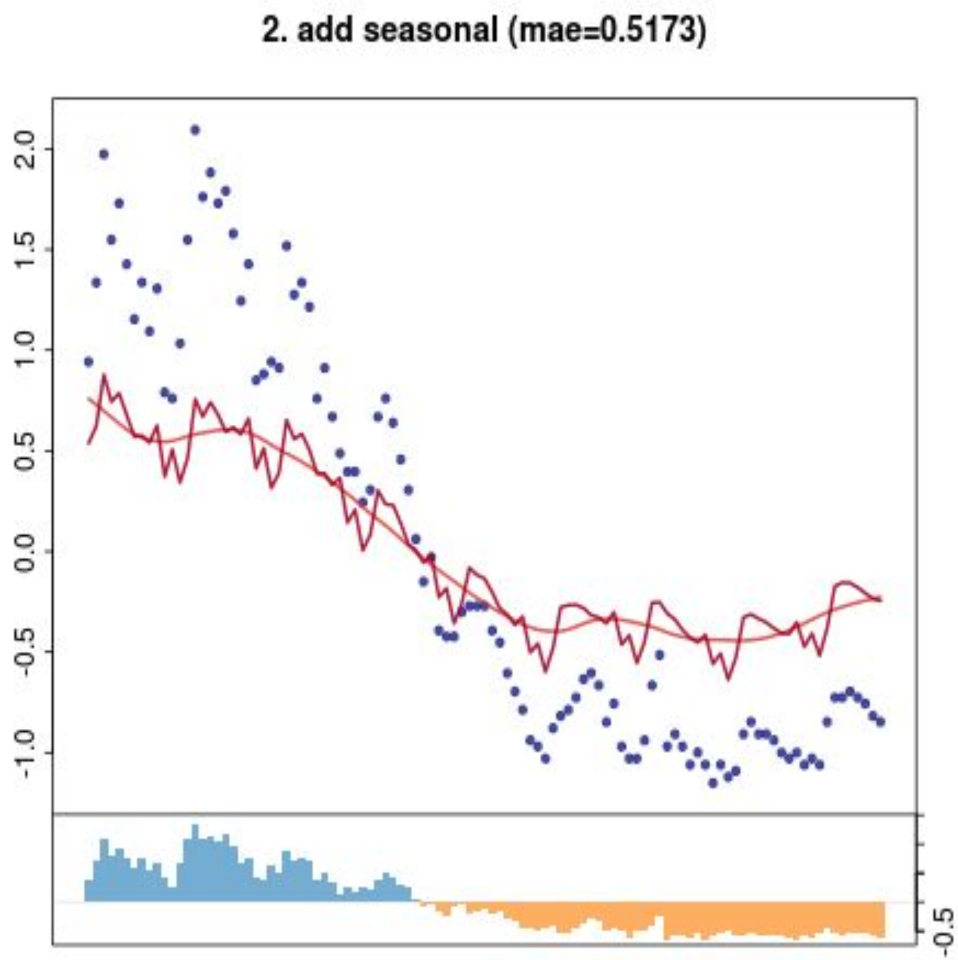
plot3: $y_t = trend_t + seasonal_t + b_1\ x_{1t}$

plot4: $y_t = trend_t + seasonal_t + b_1\ x_{1t} + b_2 x_{2t}$

Trend

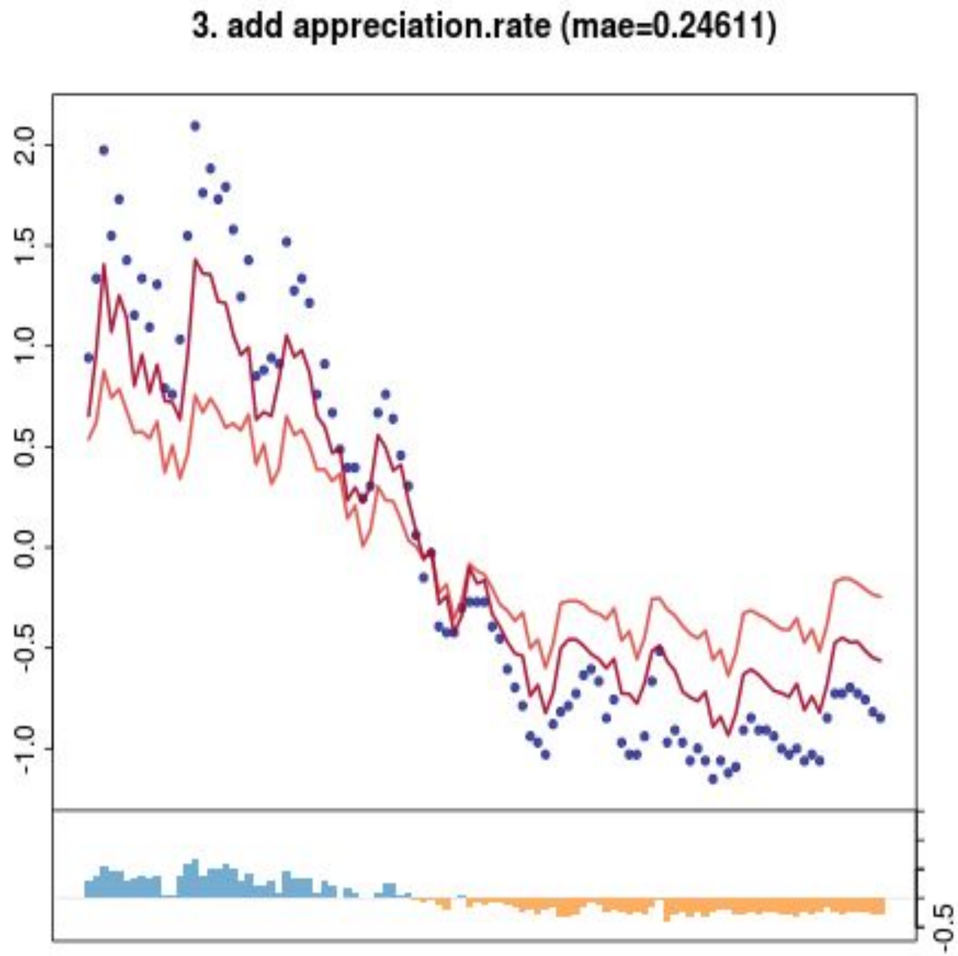


Seasonal

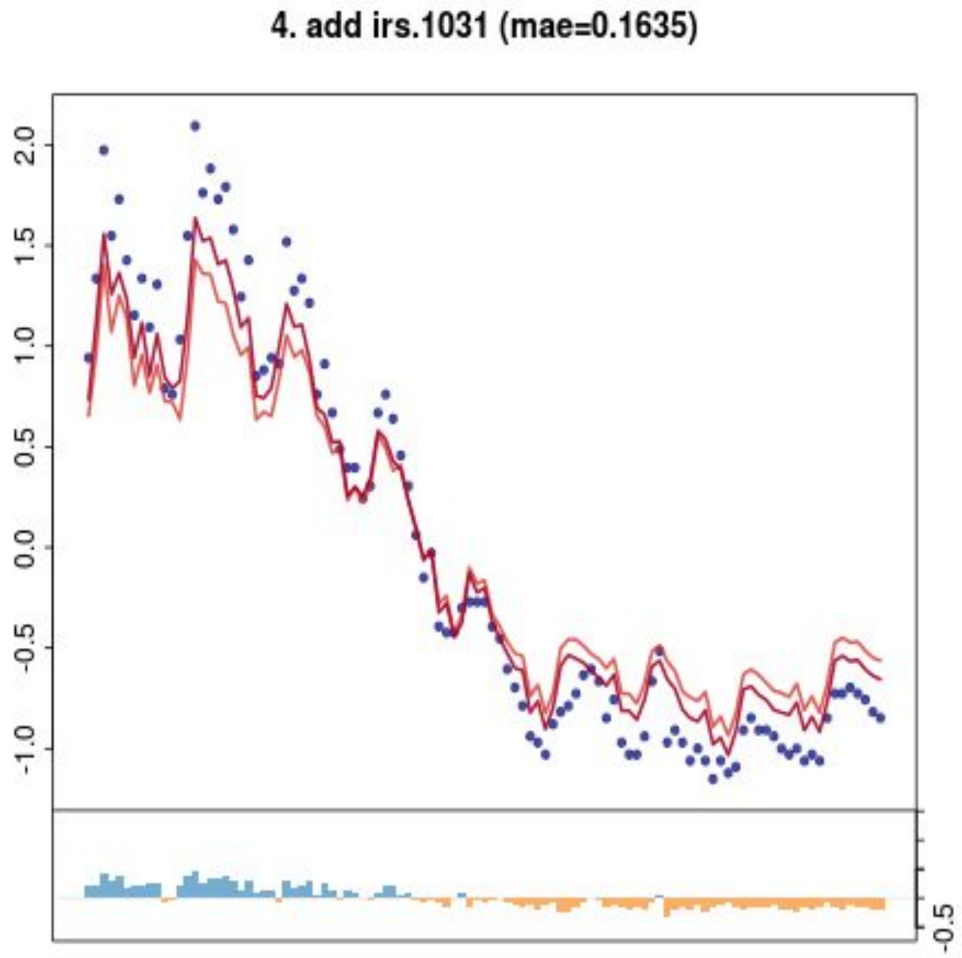


Google

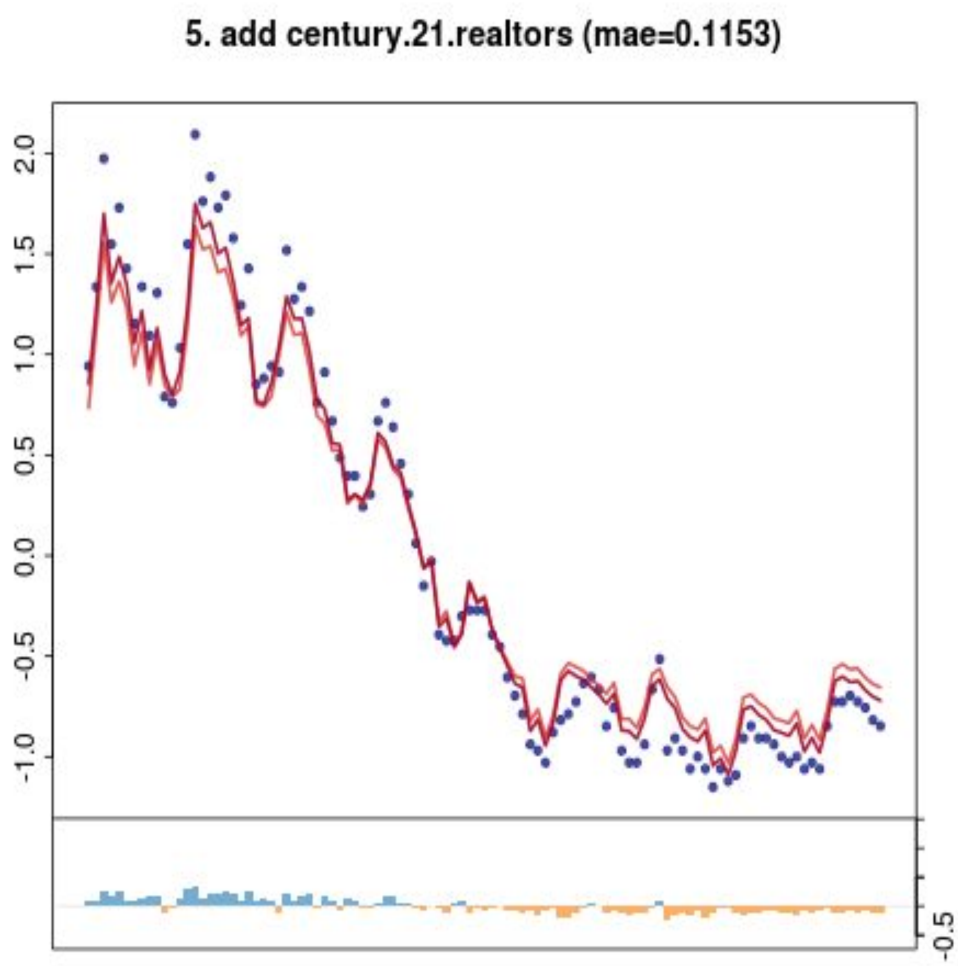
[appreciation rate]



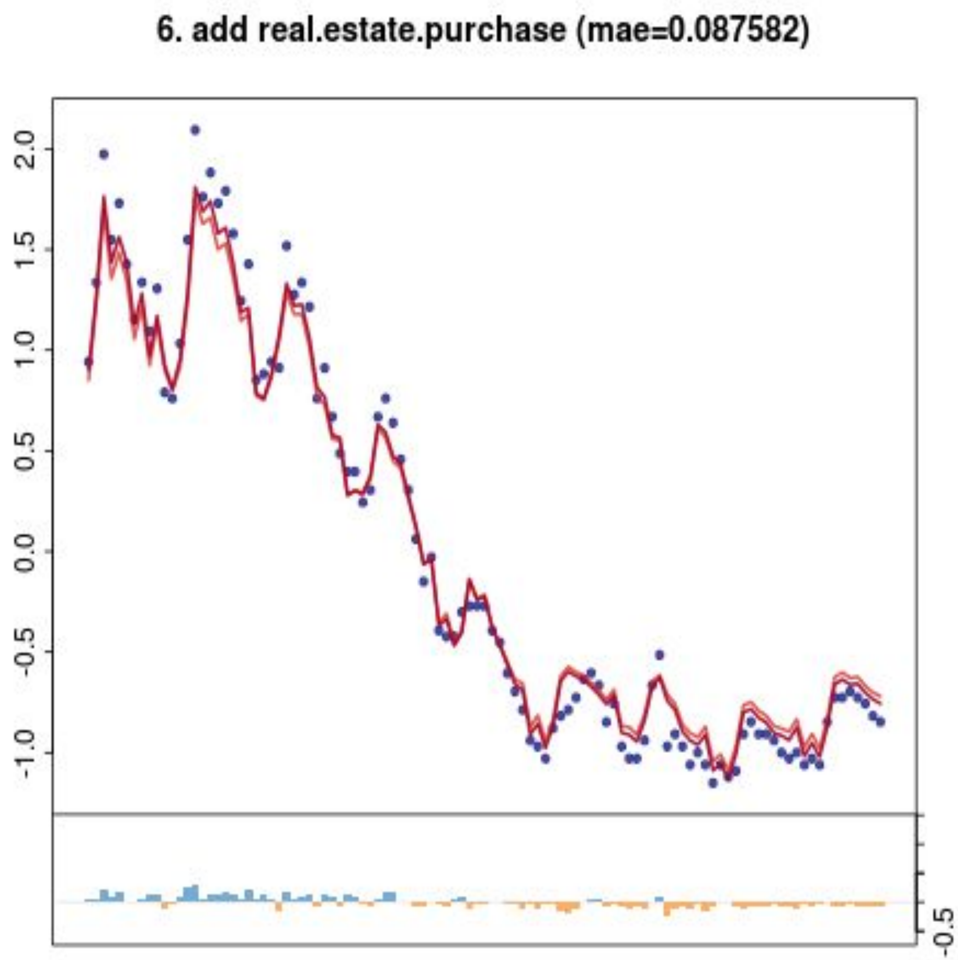
[irs 1031]



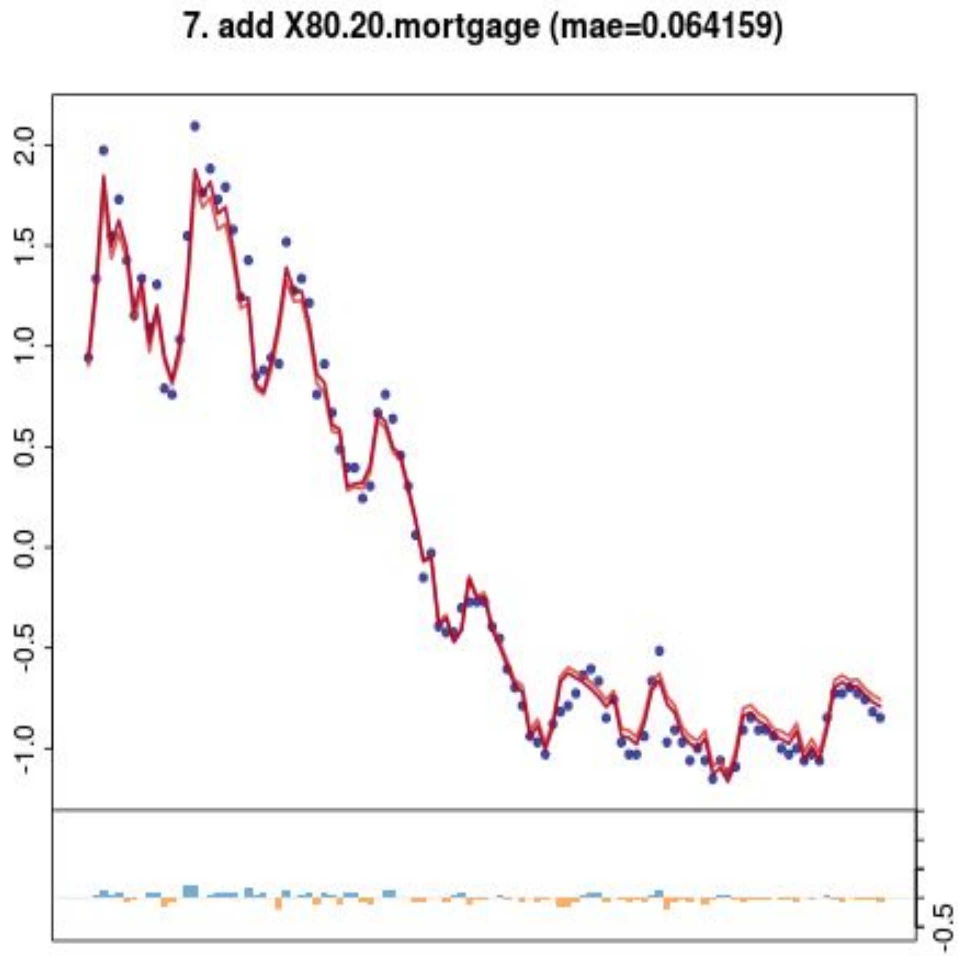
[century 21 realtors]



[real estate purchase]



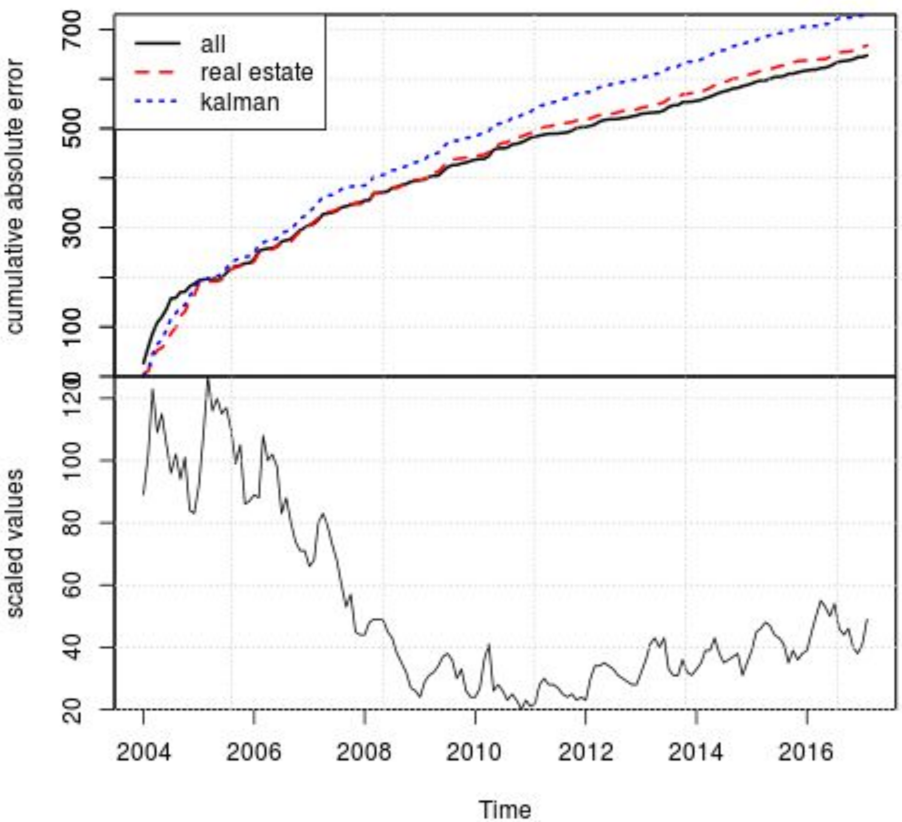
[80-20 mortgage]



Another approach: use query categories

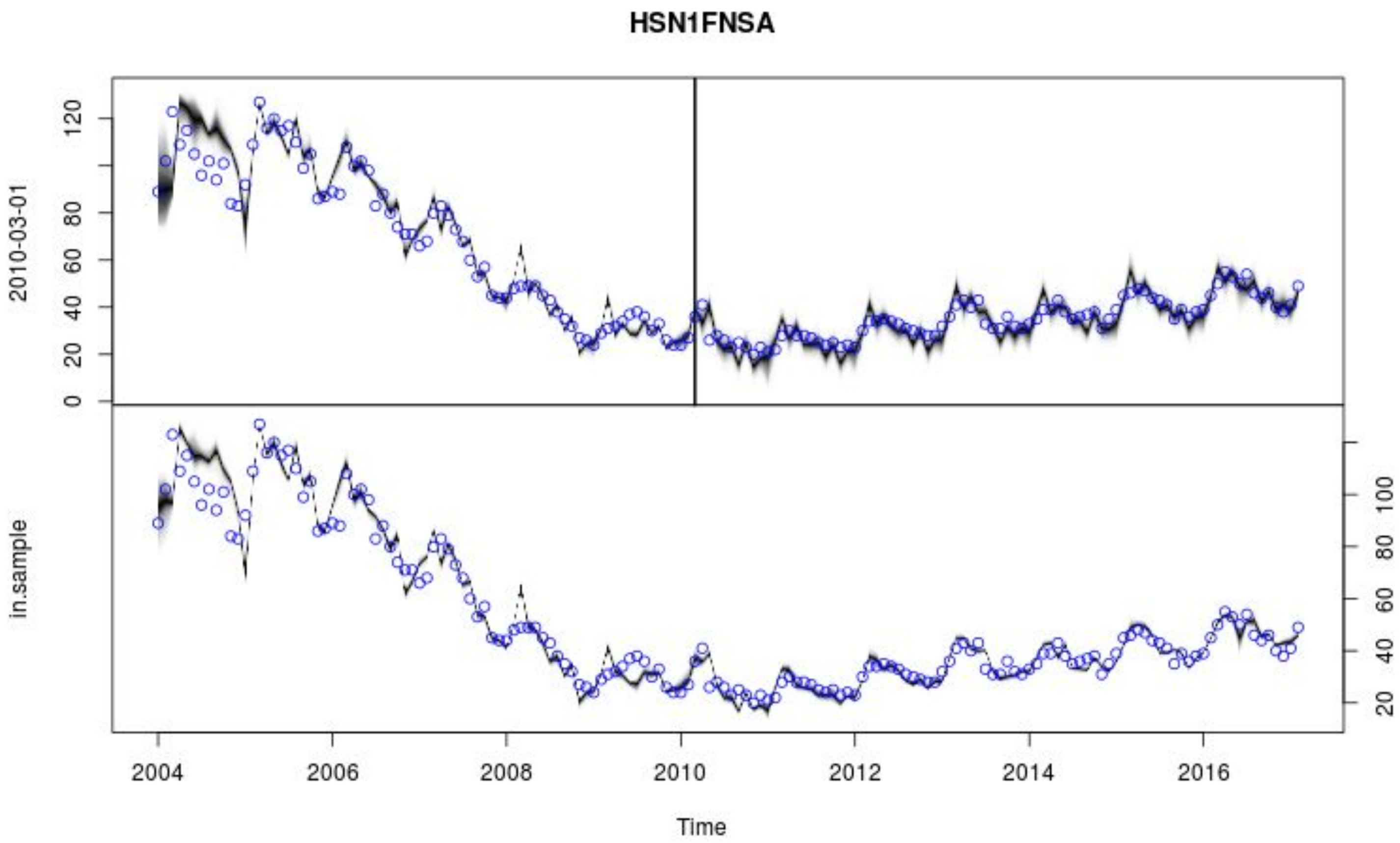
Advantage: exact queries may not persist in future but categories will (probably).

1. all 148 commercial categories;
2. only the 8 real estate categories;
3. no regression predictors

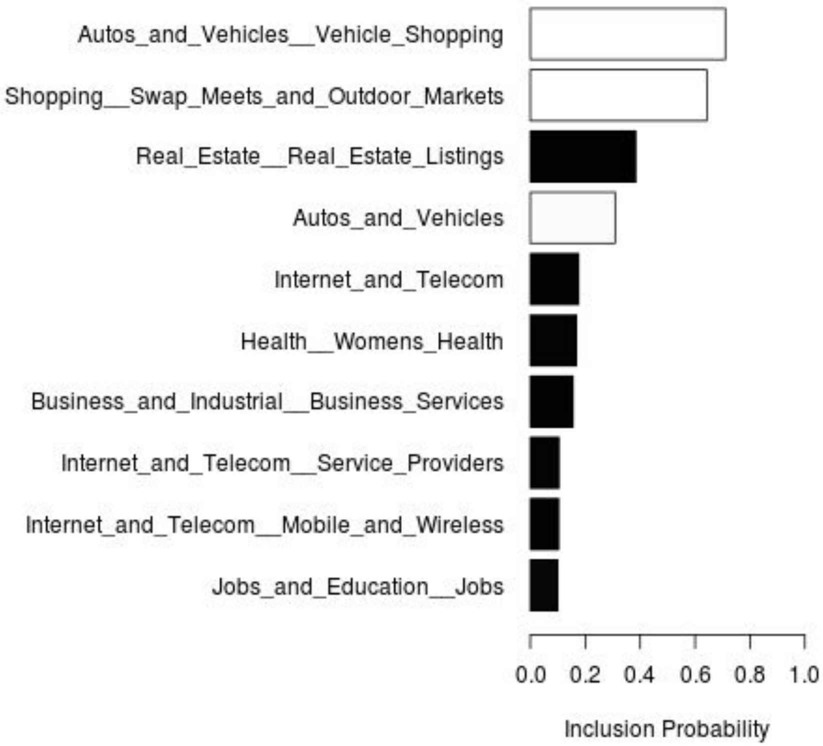


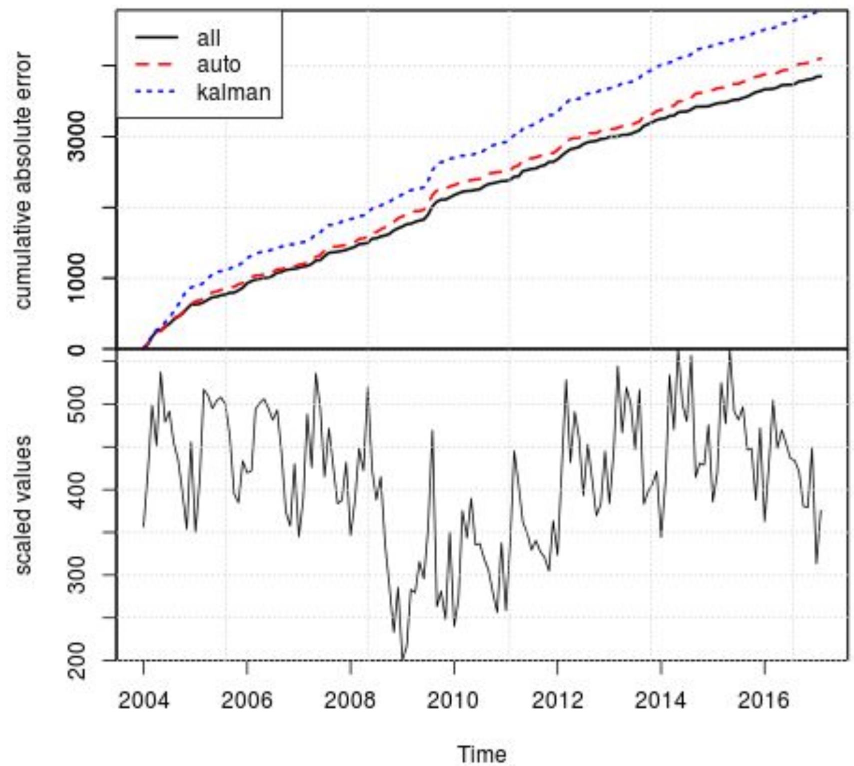
Out of sample prediction

Estimate up until time t and then freeze all the posterior distributions. This freezes regression and the variance posteriors, but allows for Kalman updating and regression predictions.

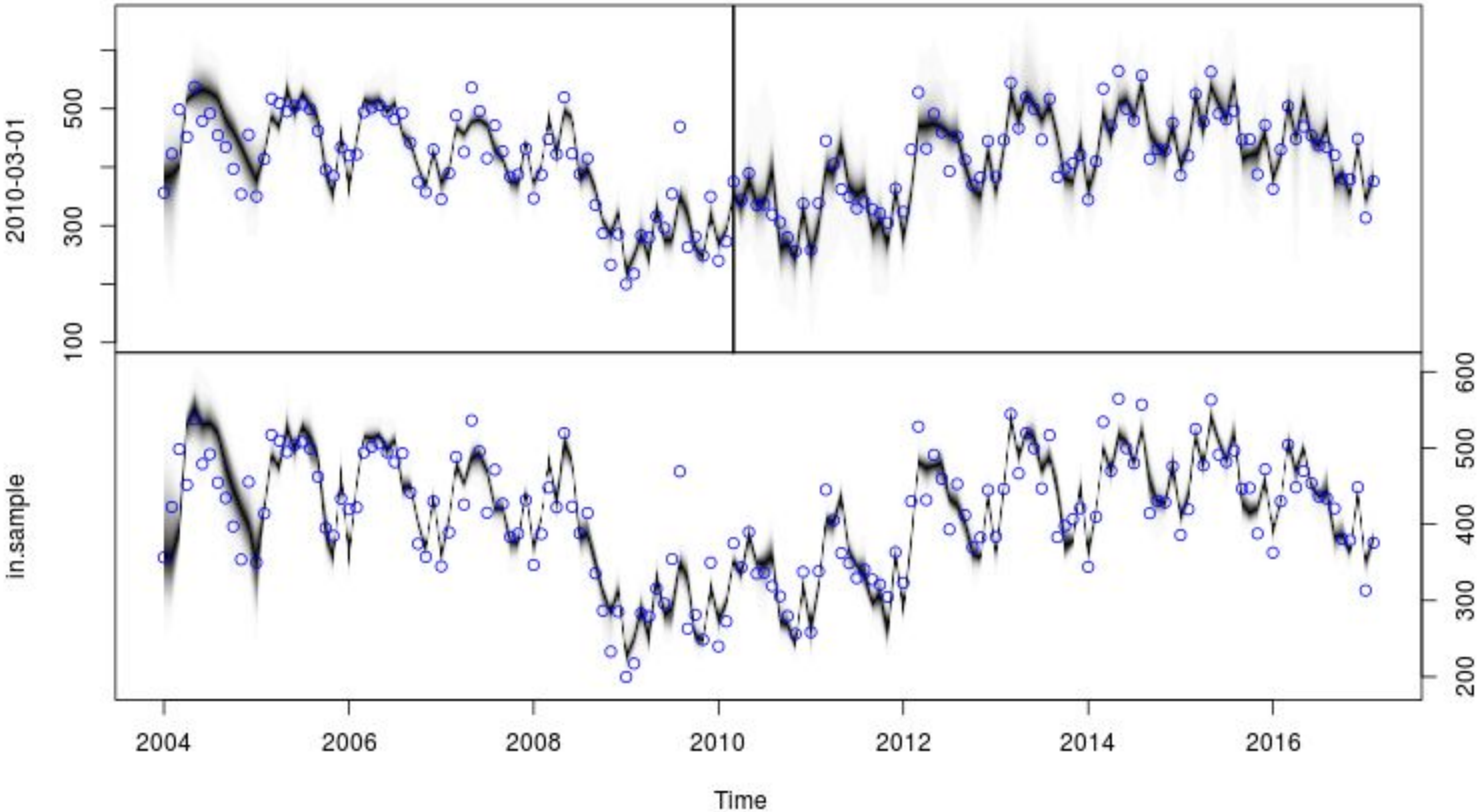


Same thing for motor vehicle sales





Motor vehicles out of sample forecast

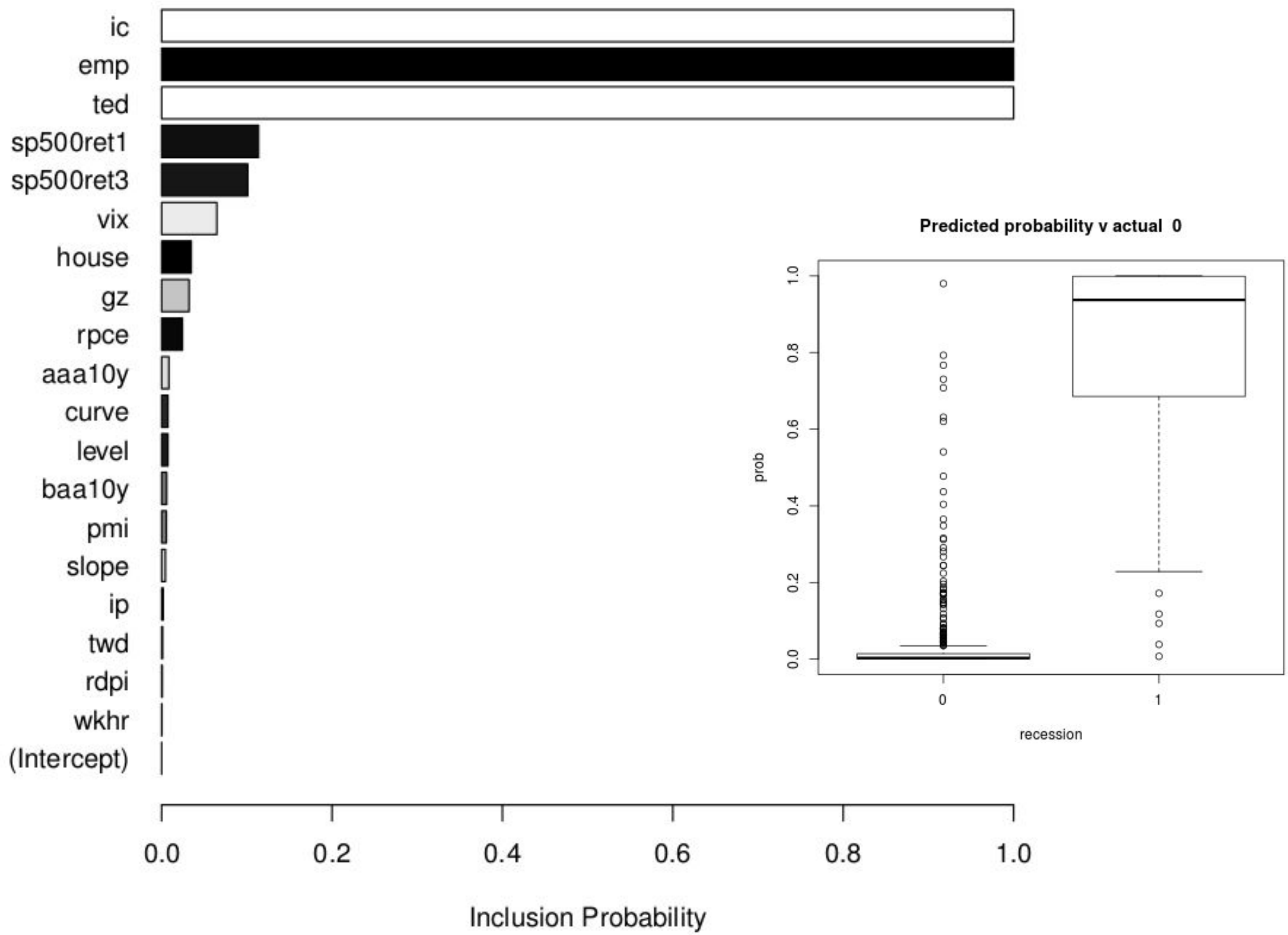


Predict recessions at different forecast horizons

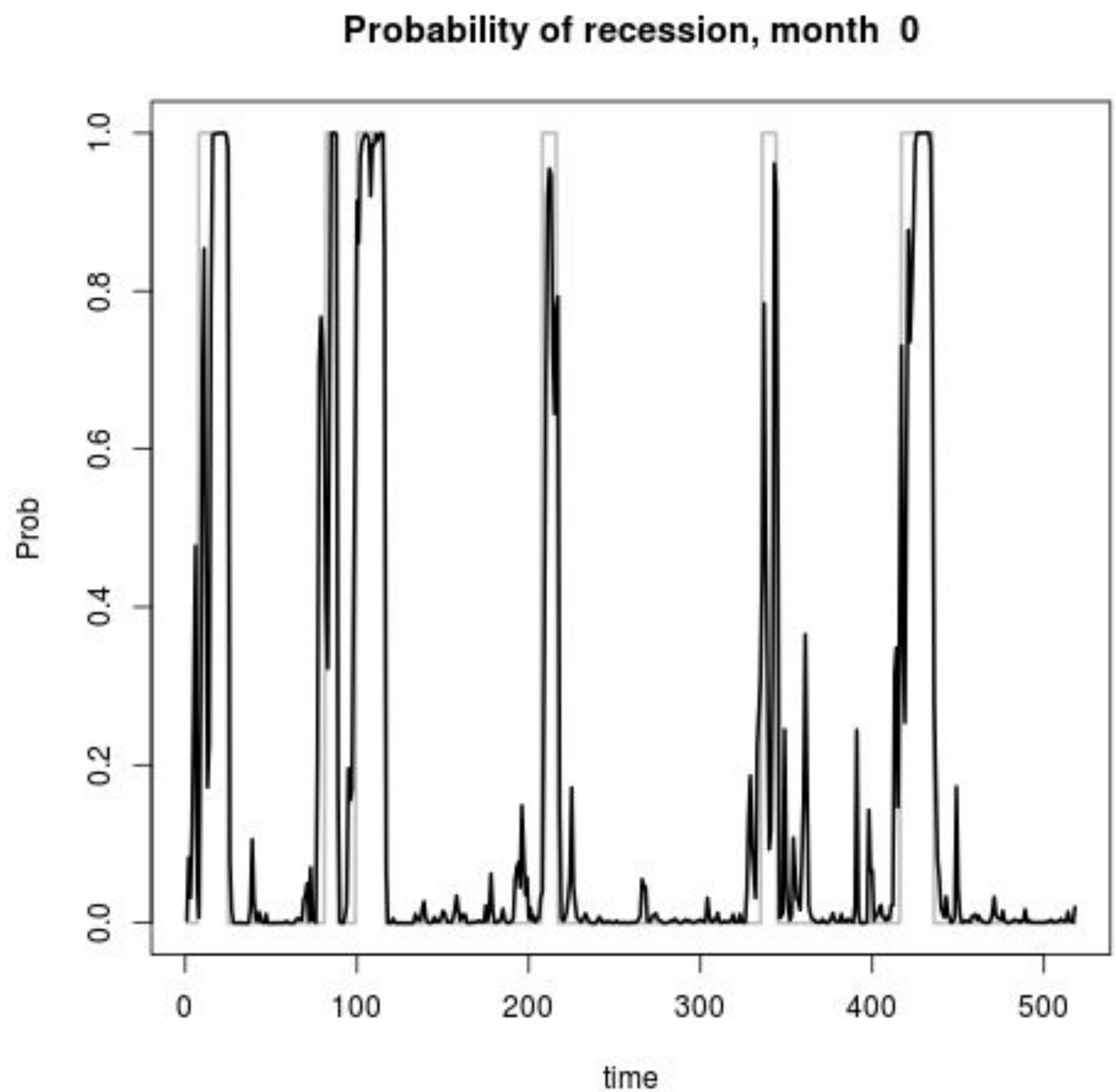
Based on Berge, Sinha, Smolyansky (2016), "[Which market indicators best forecast recessions?]" FEDS Notes. Logistic model, BMA.

Variable	Definition/notes	Transformation
Curvature of yield curve	2 x 2-year minus 3-month and 10-year	
GZ index	Gilchrist and Zakrajsek (AER, 2012)	
TED spread	3-month ED less 3-month Treasury yield	
BBB corporate spread	BBB less 10-year Treasury yield	
S 500, 1-month return		1-month log diff.
S 500, 3-month return		3-month log diff.
Trade-weighted dollar		3-month log diff.
VIX	CBOE and extended following Bloom	
Macroeconomic Indicators		
Real personal consumption expend.		3-month log diff.
Real disposable personal income		3-month log diff.
Industrial production		3-month log diff.
Housing permits		3-month log diff.
Nonfarm payroll employment		3-month log diff.
Initial claims	4-week moving average	3-month log diff.
Weekly hours, manufacturing		3-month log diff.
Purchasing managers index		3-month log diff.

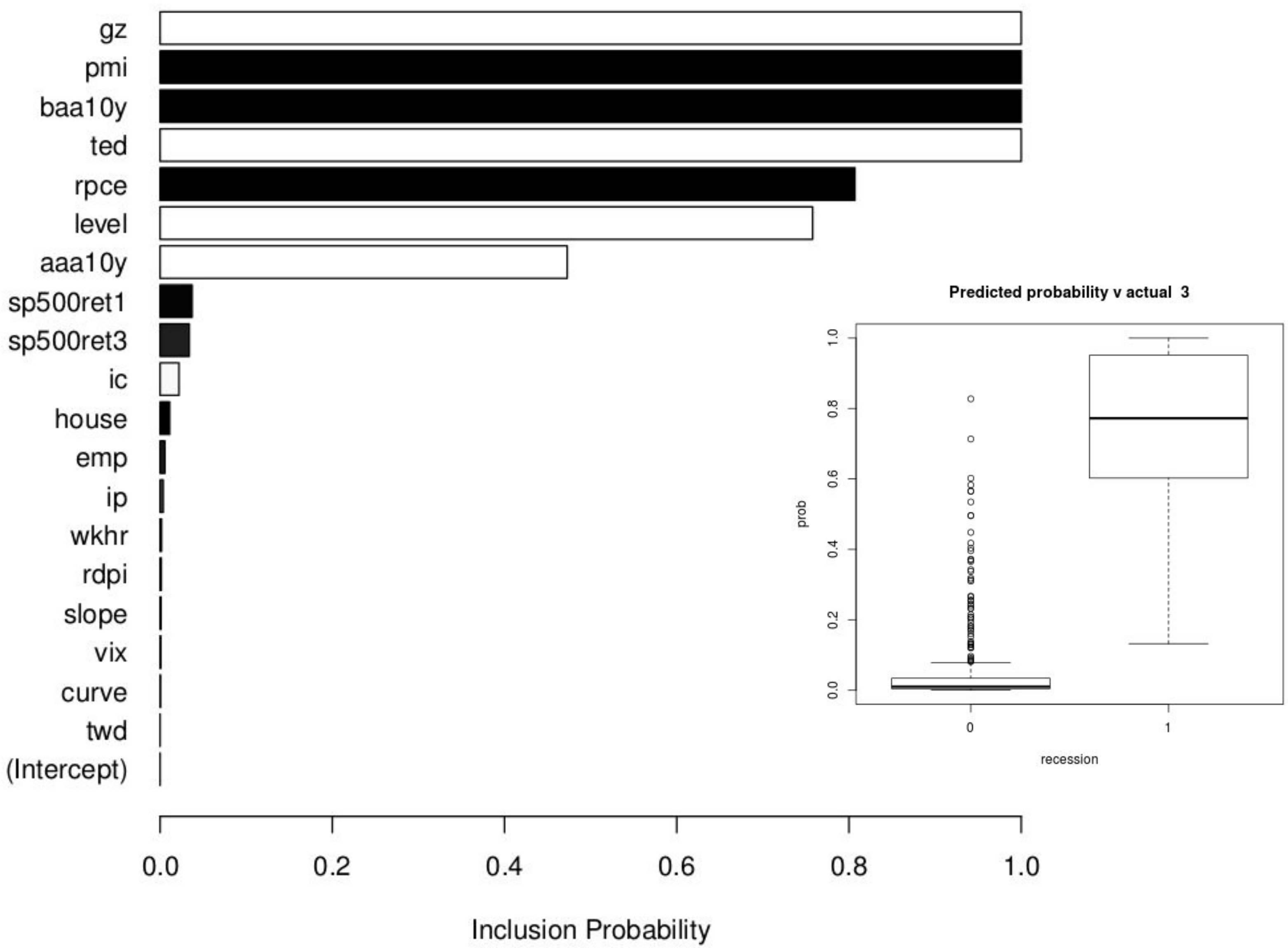
Predicting recessions 0 months out



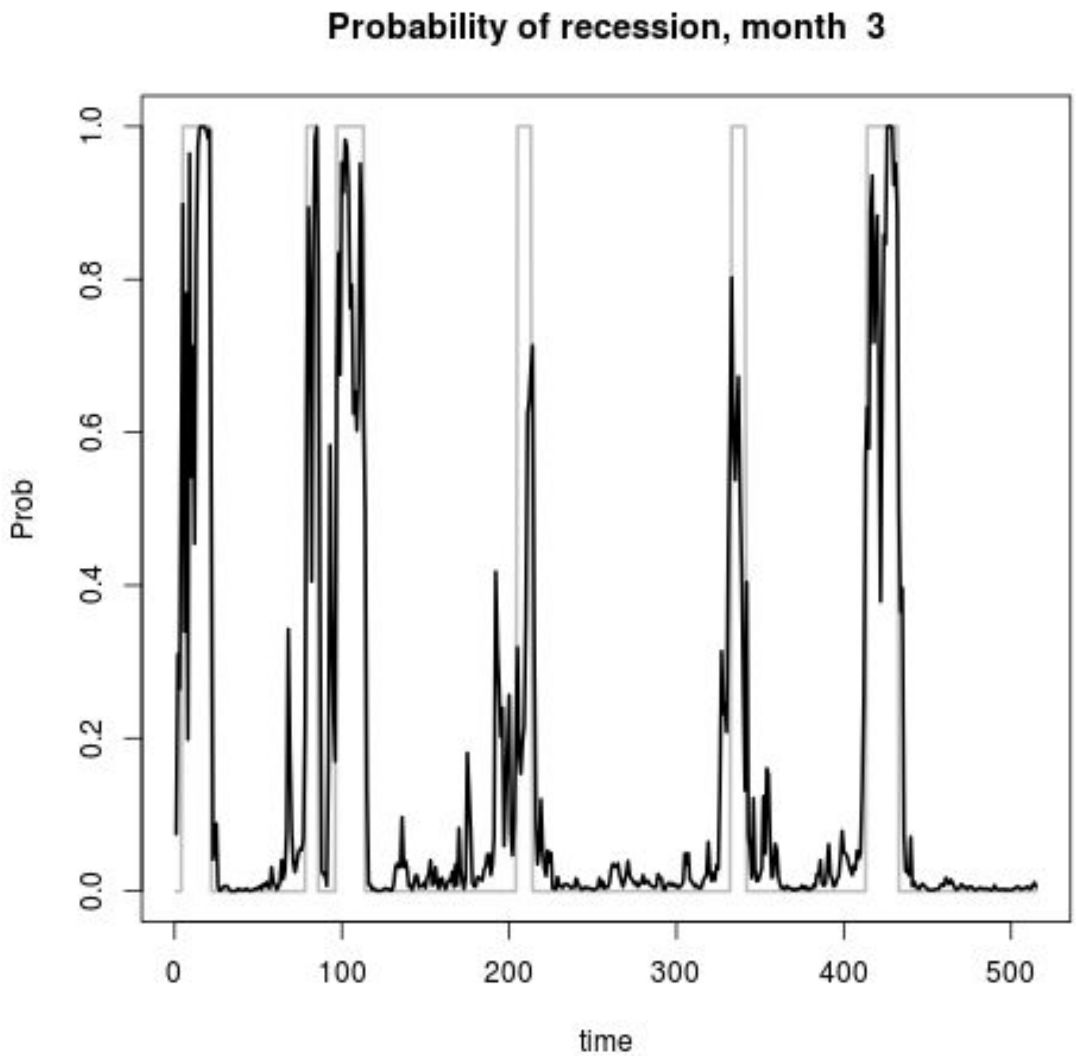
Predicting recessions 0 months out



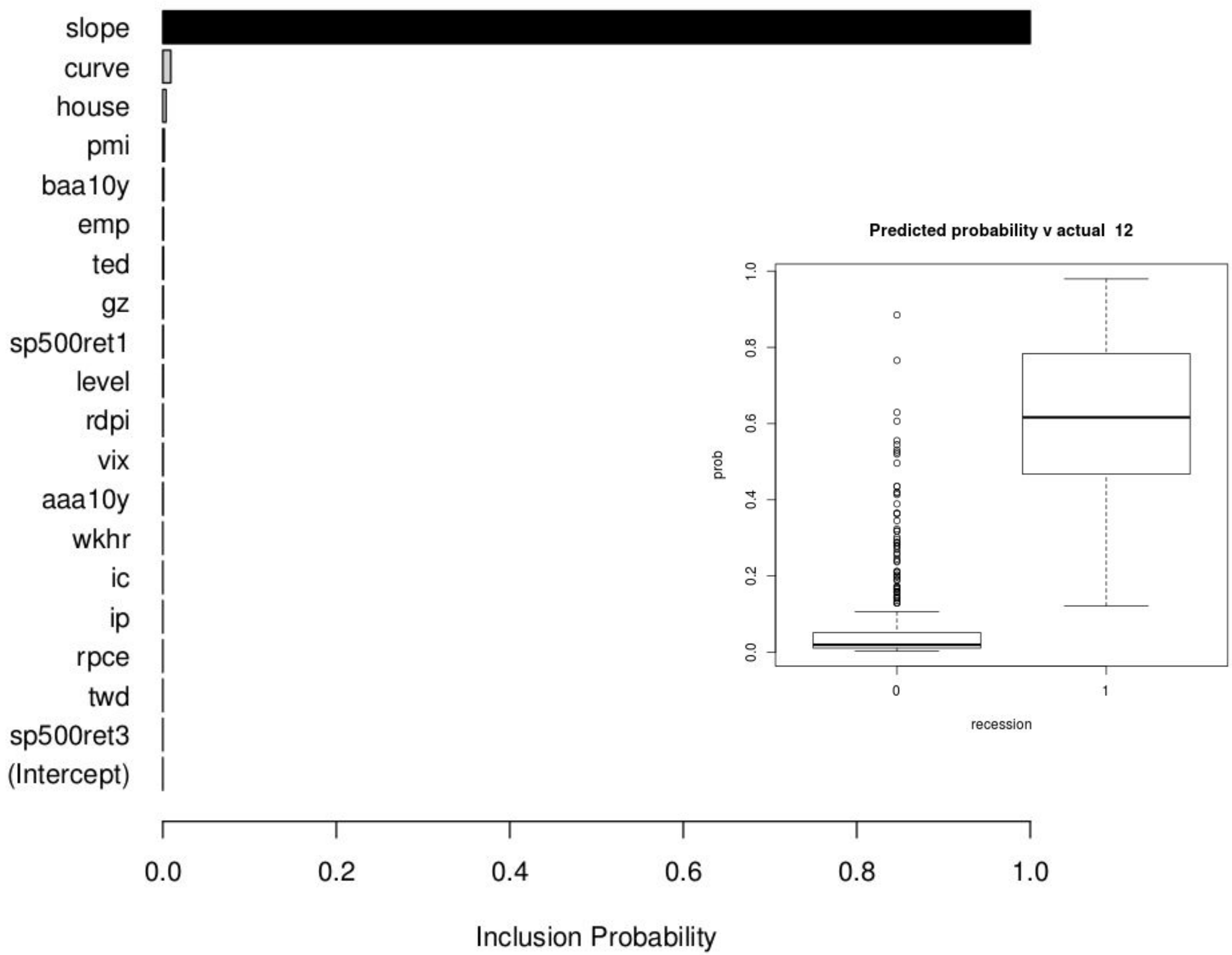
Predicting recessions 3 months out



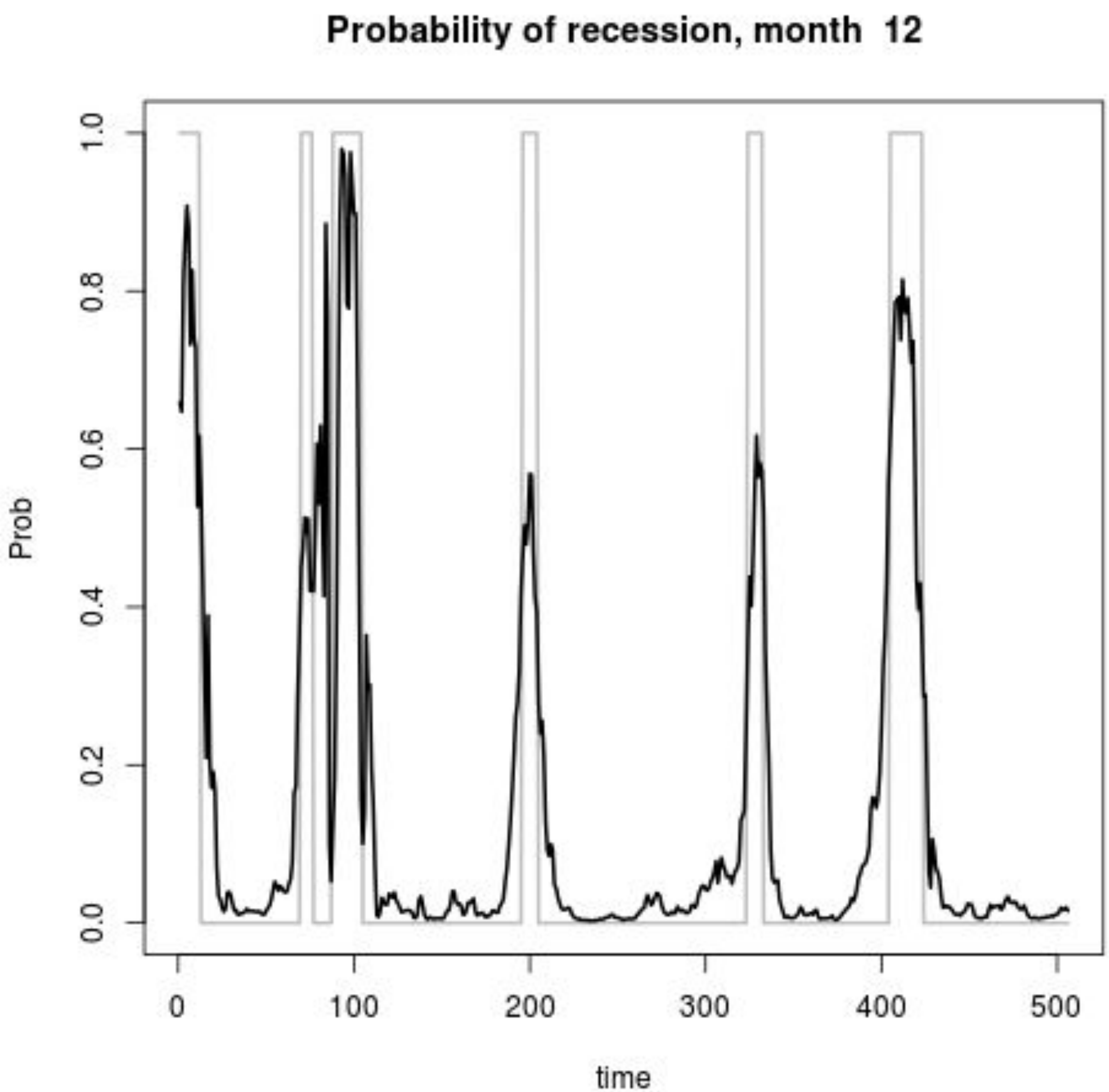
Predicting recessions 3 months out



Predicting recessions 12 months out



Predicting recessions 12 months out



Why time series estimation is important: persistence

T+1

T

	Not in recession	In recession
Not in recession	0.84	0.01
In recession	0.01	0.14

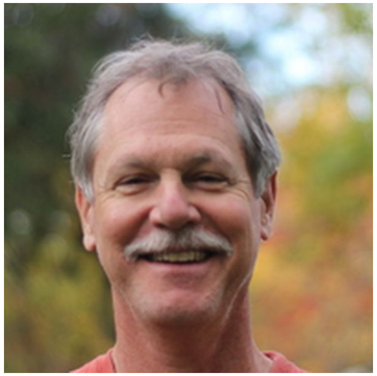
Predictors are essentially the same as without a time series model, but you assign higher probability to recessions.

Motivating example for causal inference

An advertiser contemplates changing its (bid, budget, creative) and wants to know what will happen to its (clicks, revenue, conversions). Apply treatment and compare actual to prediction of would have happened without the treatment (interrupted regression, synthetic control).

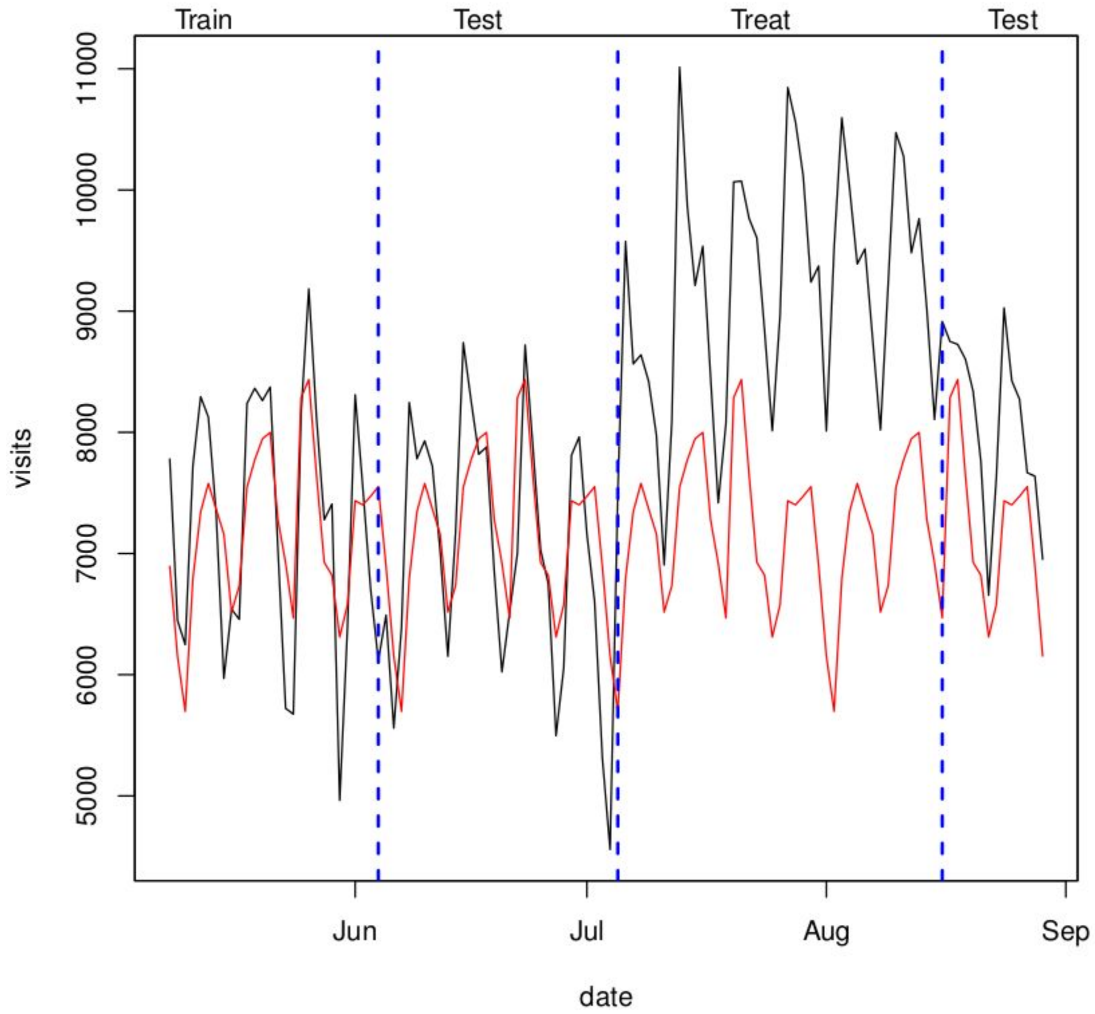
Brodersen, Gallusser, Koehler, Remy and Scott *Annals of Applied Statistics*, vol. 9 (2015), pp. 247-274

Brodersen and Varian (2016) Estimating online ad effectiveness: a practical guide

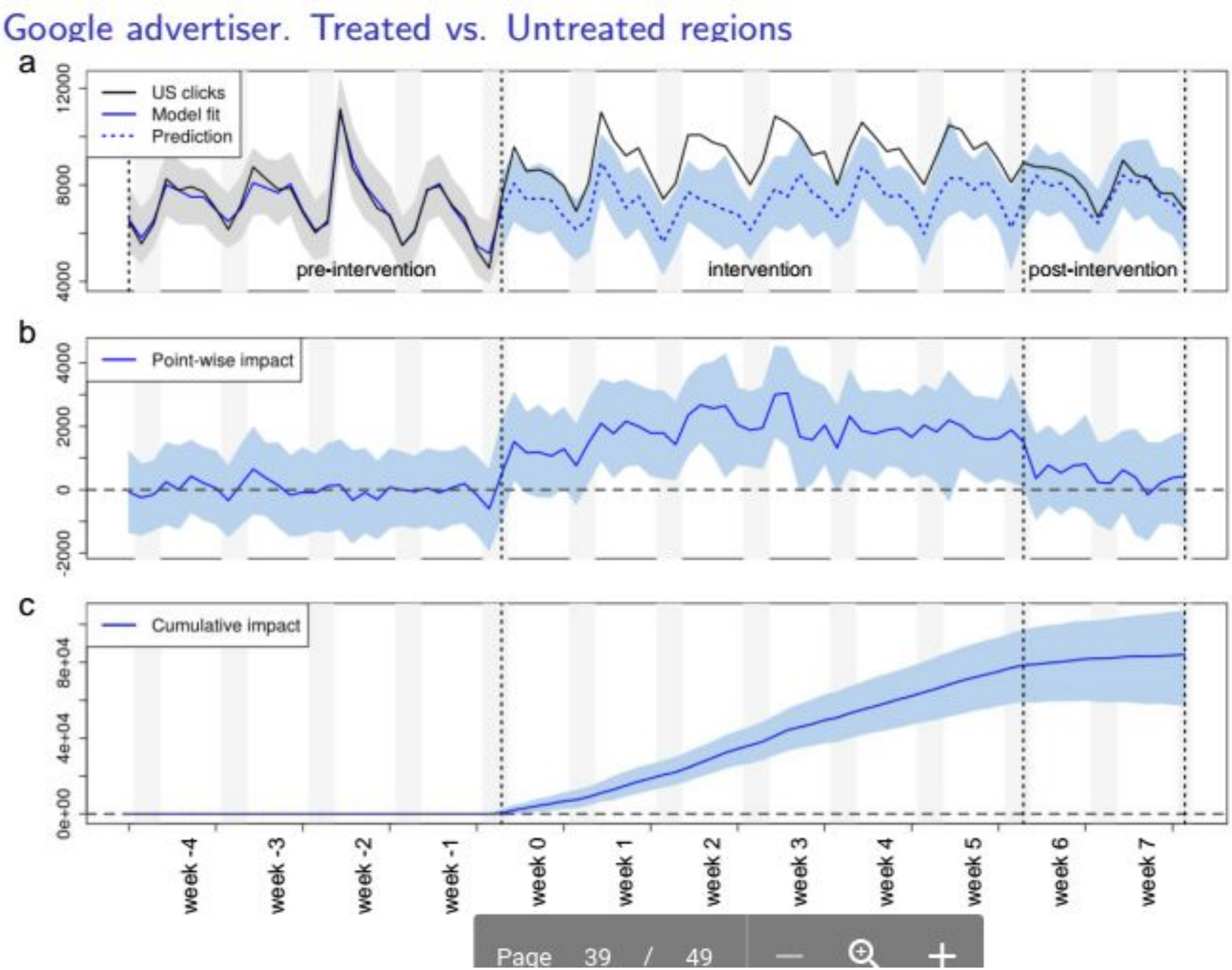


Causal inference

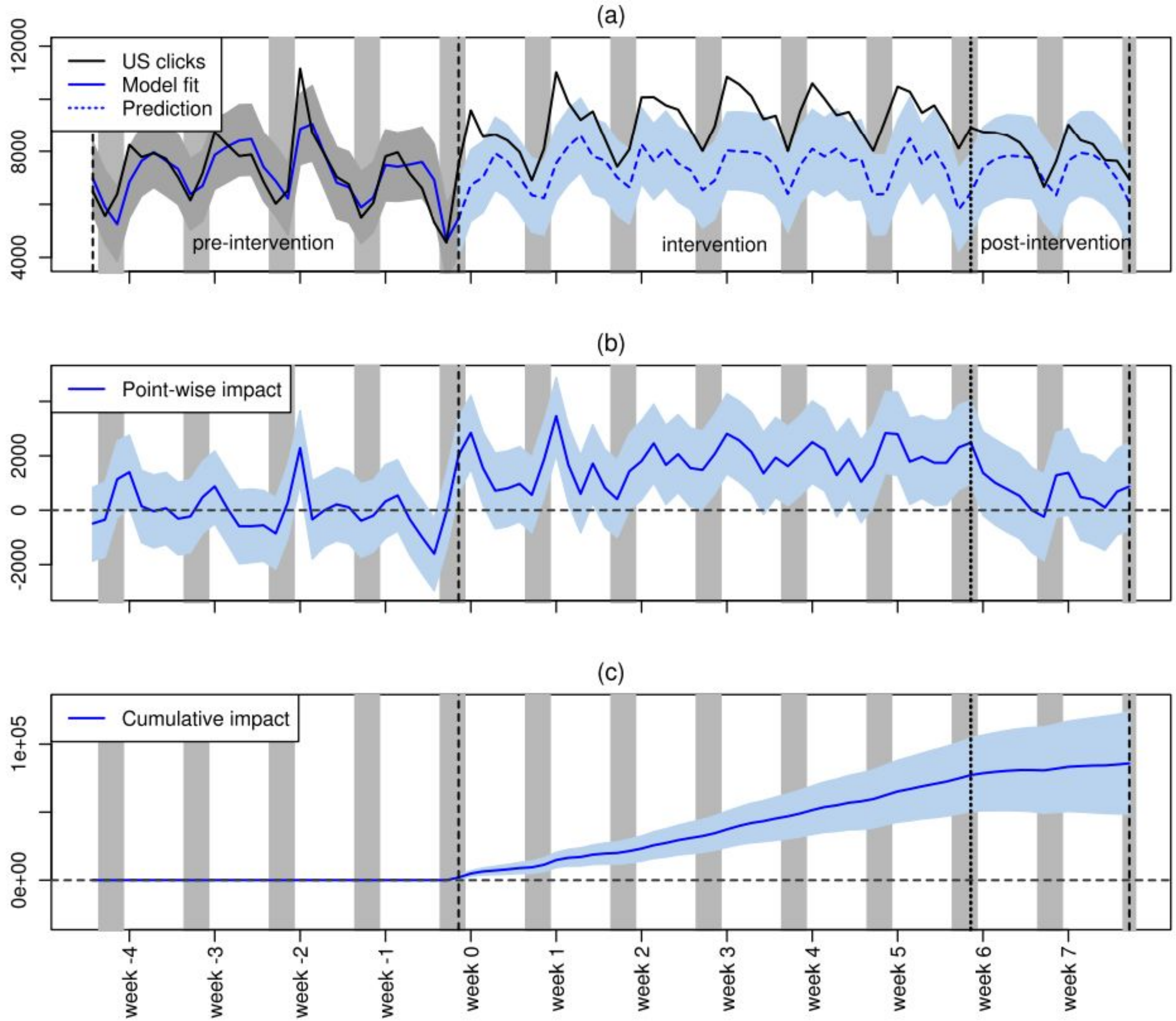
Can build model of counterfactual: train, test, treat, compare. Related to "synthetic control" in Abadie (2003, 2010).



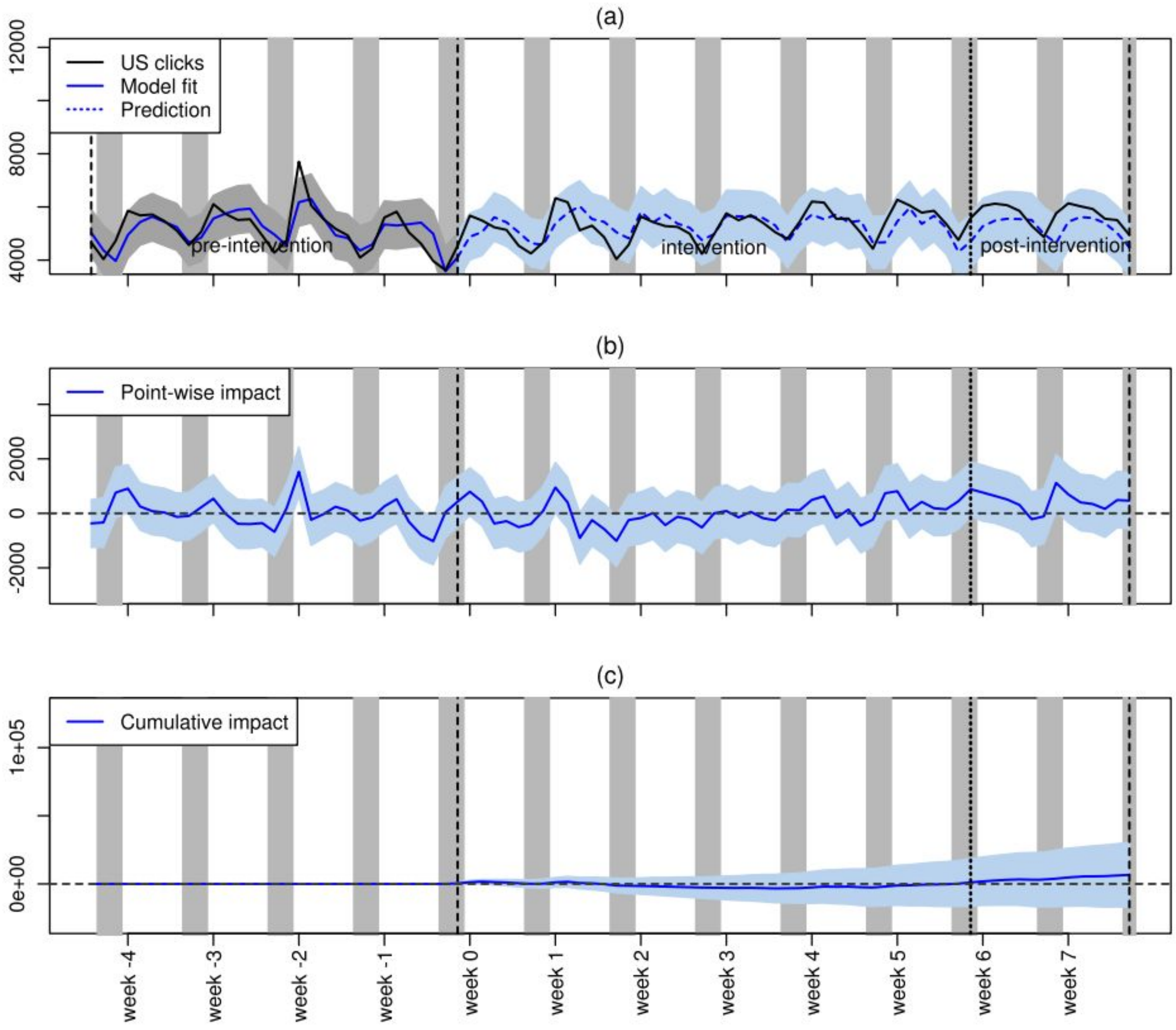
Cross section model: treated v untreated regions



Time series model: treated v untreated times (Trends)

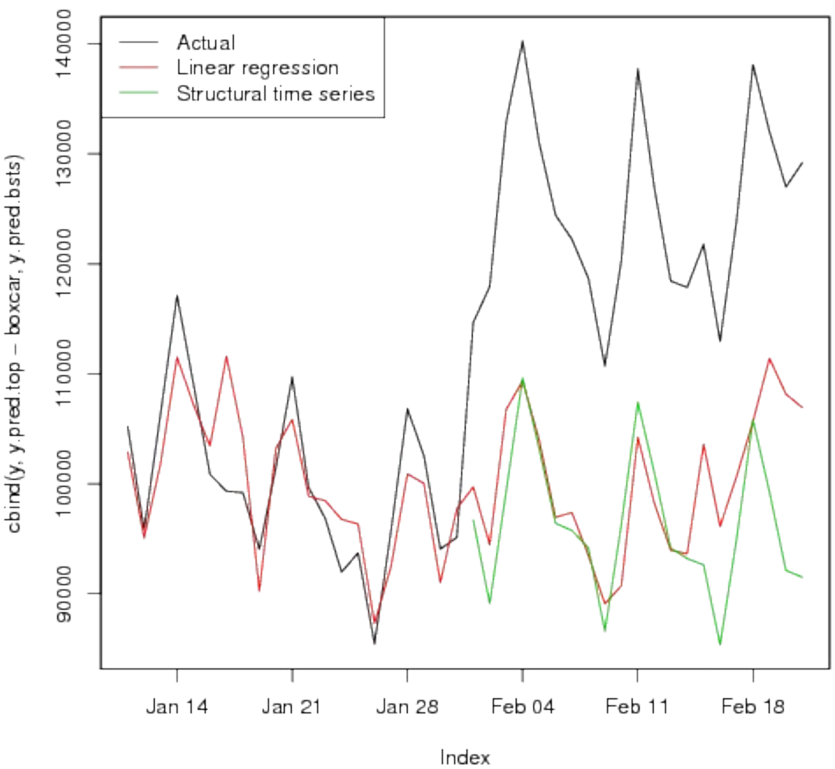


Predicted clicks in untreated regions (Trends)

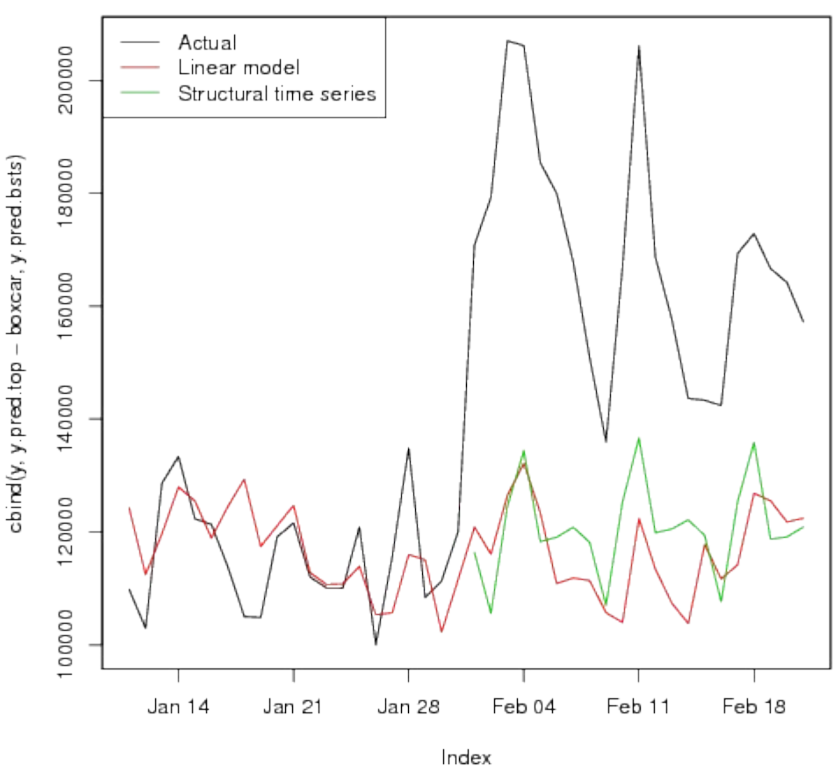


Another ad experiment: clicks and sales

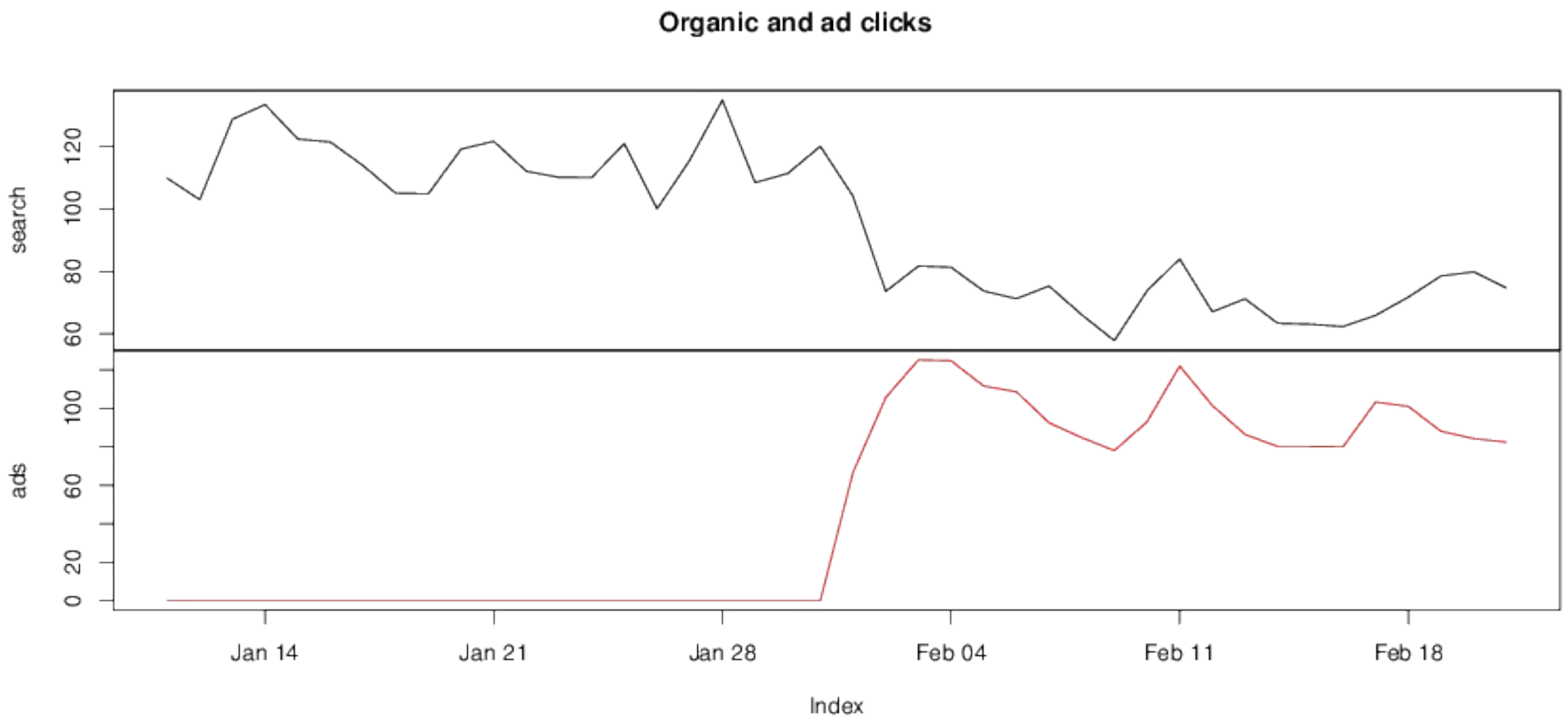
Ad clicks



Sales



Cannibalization of organic clicks

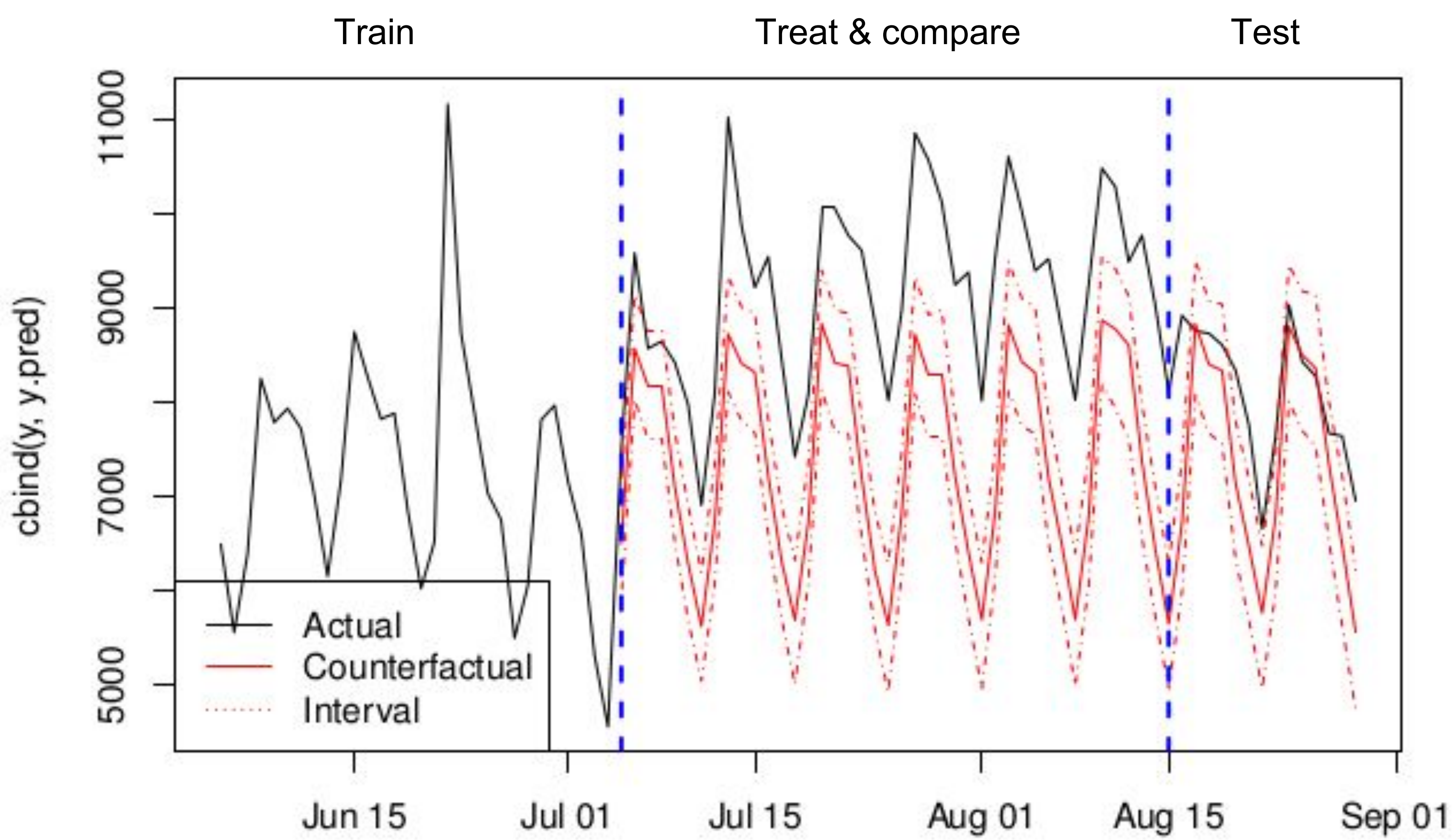


Many ways to predict counterfactual

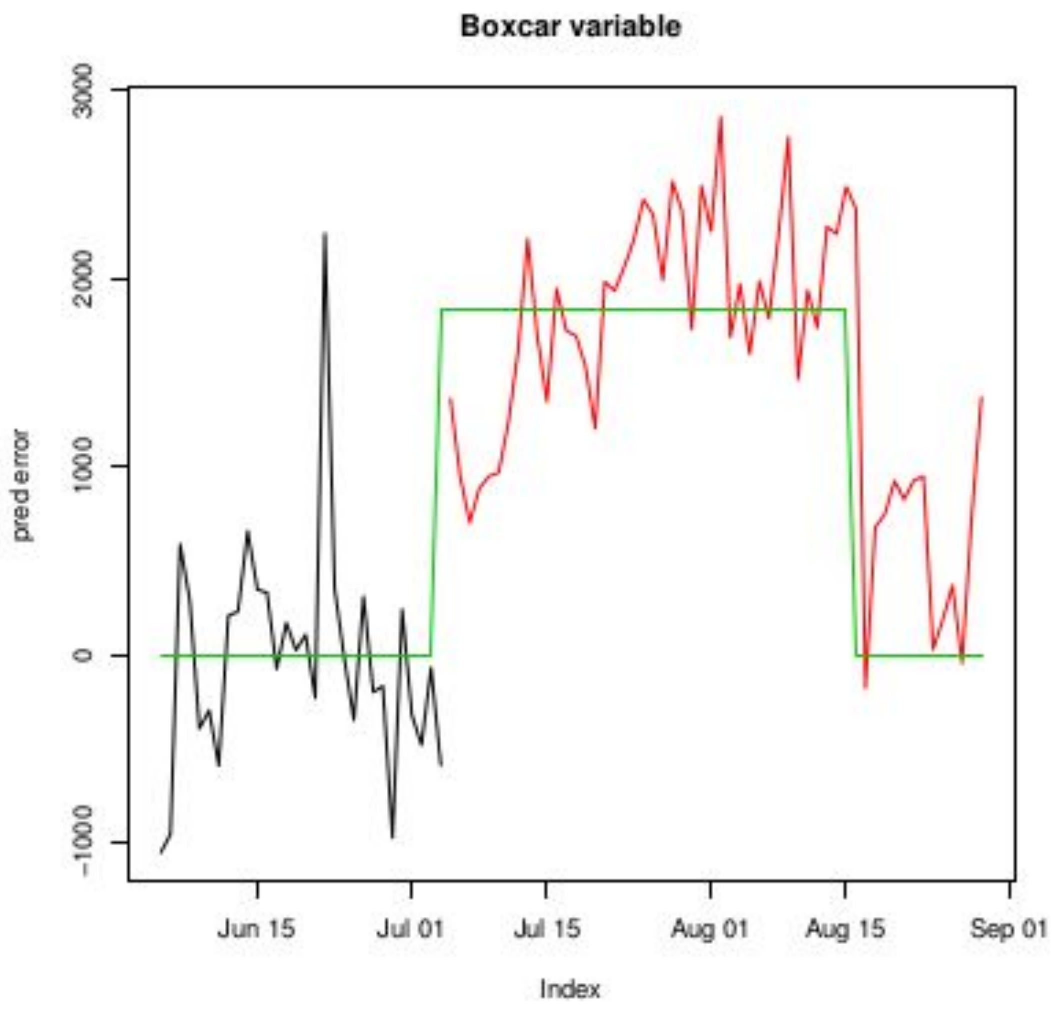
(joint work with Kay Brodersen)

- BSTS
  - Extrapolation
  - Boxcar variable
    - All predictors
    - Top predictors
- Linear model
  - Simple linear model with extrapolation
  - Simple linear model with boxcar variable
- Deseasonalize data or not
  - Deseasonalized with extrapolation
  - Deseasonalized with boxcar
  - Use weekly data

BSTS extrapoliation

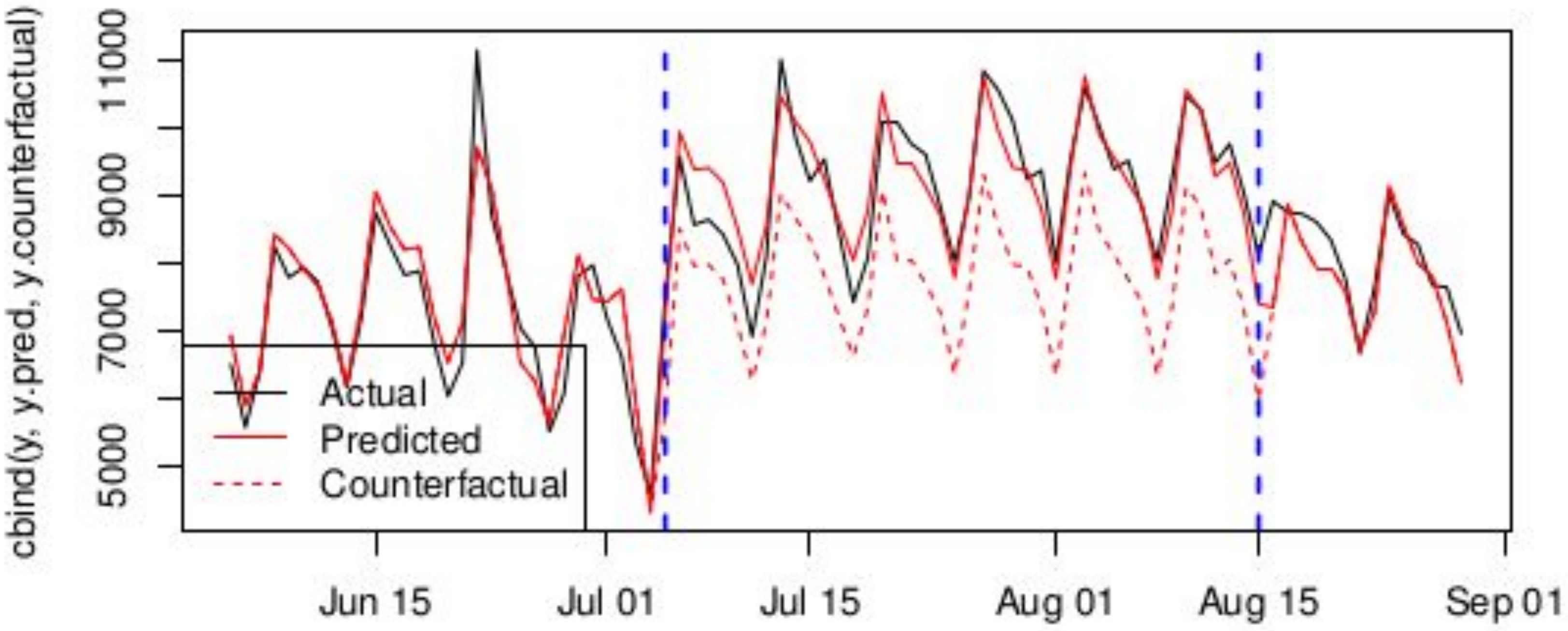


Boxcar variable



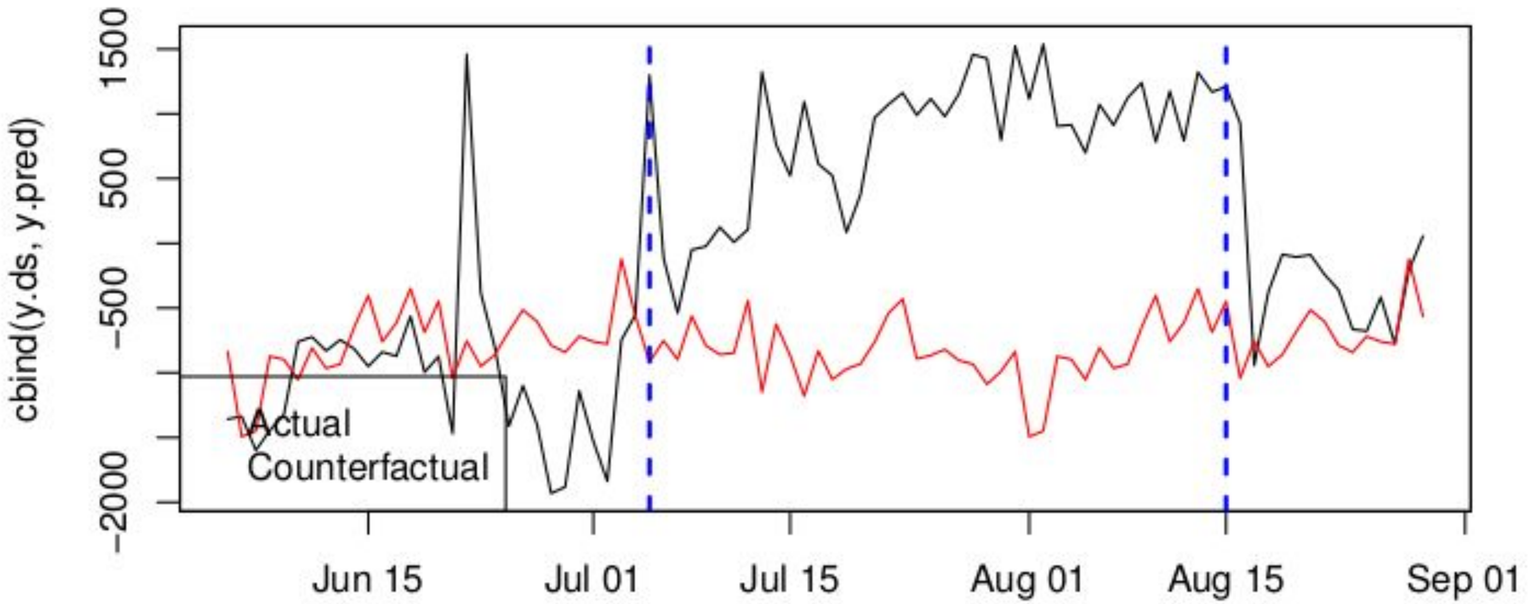
Simple linear model

- Two predictors from Trends
- July 4 dummy variable

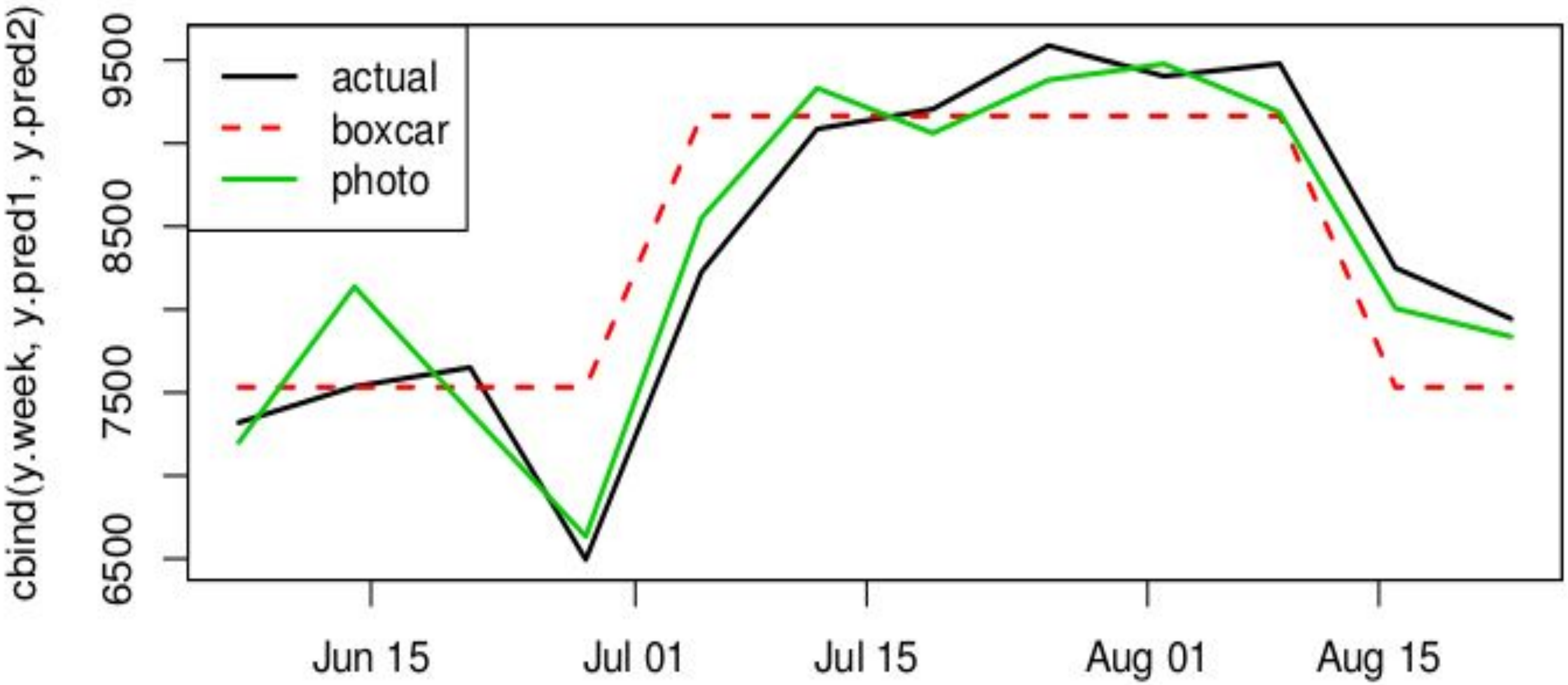


Deseasonalized data

- Day of week dummies
- July 4 dummy
- Explain other spike



Weekly data



actual
boxcar
photo
cbind(y.week, y.pred1, y.pred2)
6500
7500
8500
9500
Jun 15
Jul 01
Jul 15
Aug 01
Aug 15

Comparison

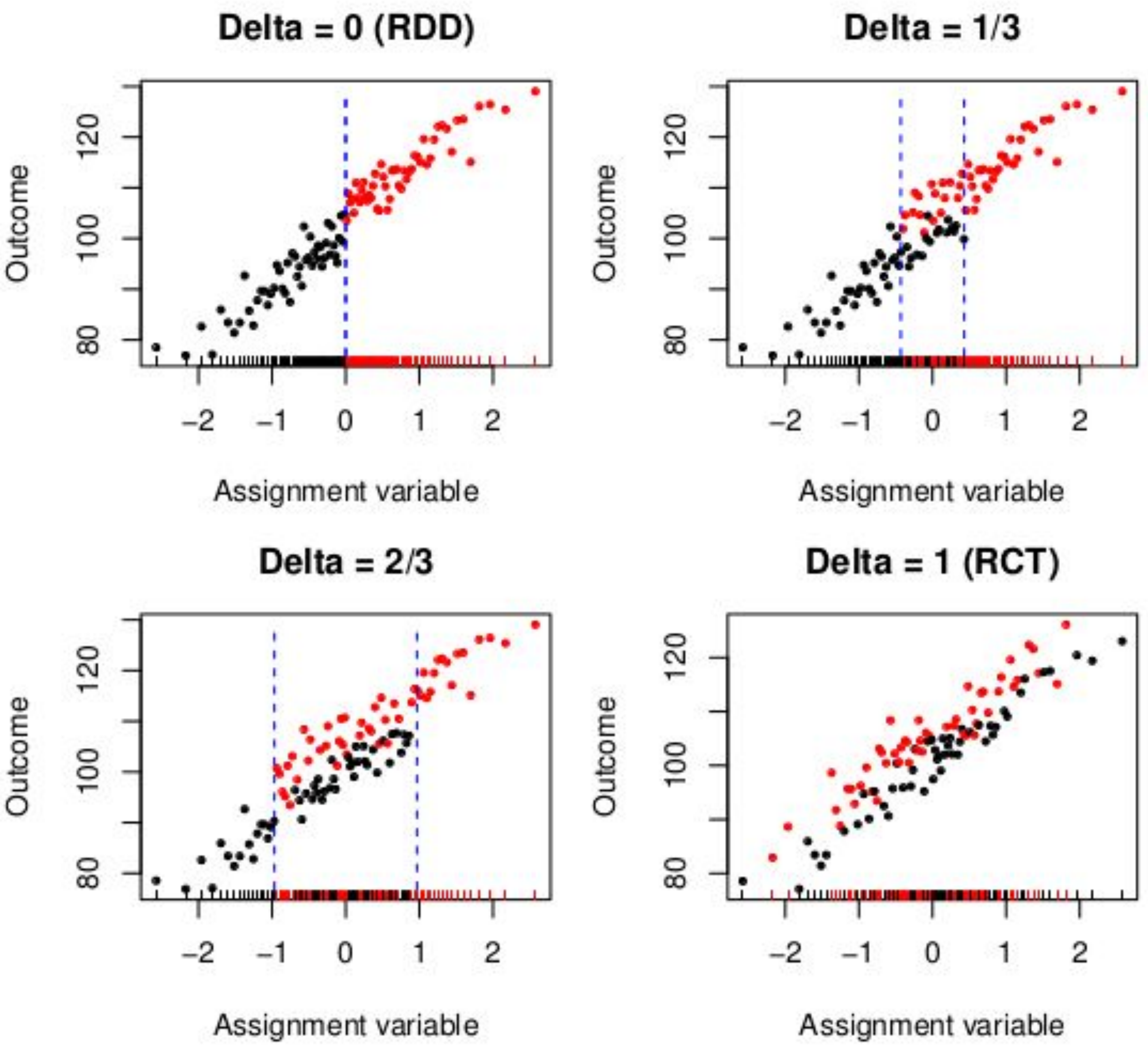
	method	estimate
1	bsts-extrap	1830.43
2	bsts-boxcar	1362.88
3	bsts-boxcar-all-predictors	1279.05
4	bsts-boxcar-top-predictors	1327.06
5	lm-boxcar	1434.57
6	lm-extrap	1289.19
7	not deseasonalized	1393.41
8	deseasonalized-boxcar	1300.67
9	deseasonalized-extrap	1298.37
10	week-boxcar	1248.61

Regression discontinuity

- Motivating examples
  - Causal impact of incentive program
  - Causal impact of merit scholarships
- Estimation methods
  - Randomized controlled trial (RCT): good statistical properties
  - Regression discontinuity (RDD): no disruption of ordering
- HYBRID = RCT + RDD
  - Related literature: tie-breaker design: if two subjects have same score, use lottery to break ties

Art Owen

Local randomization



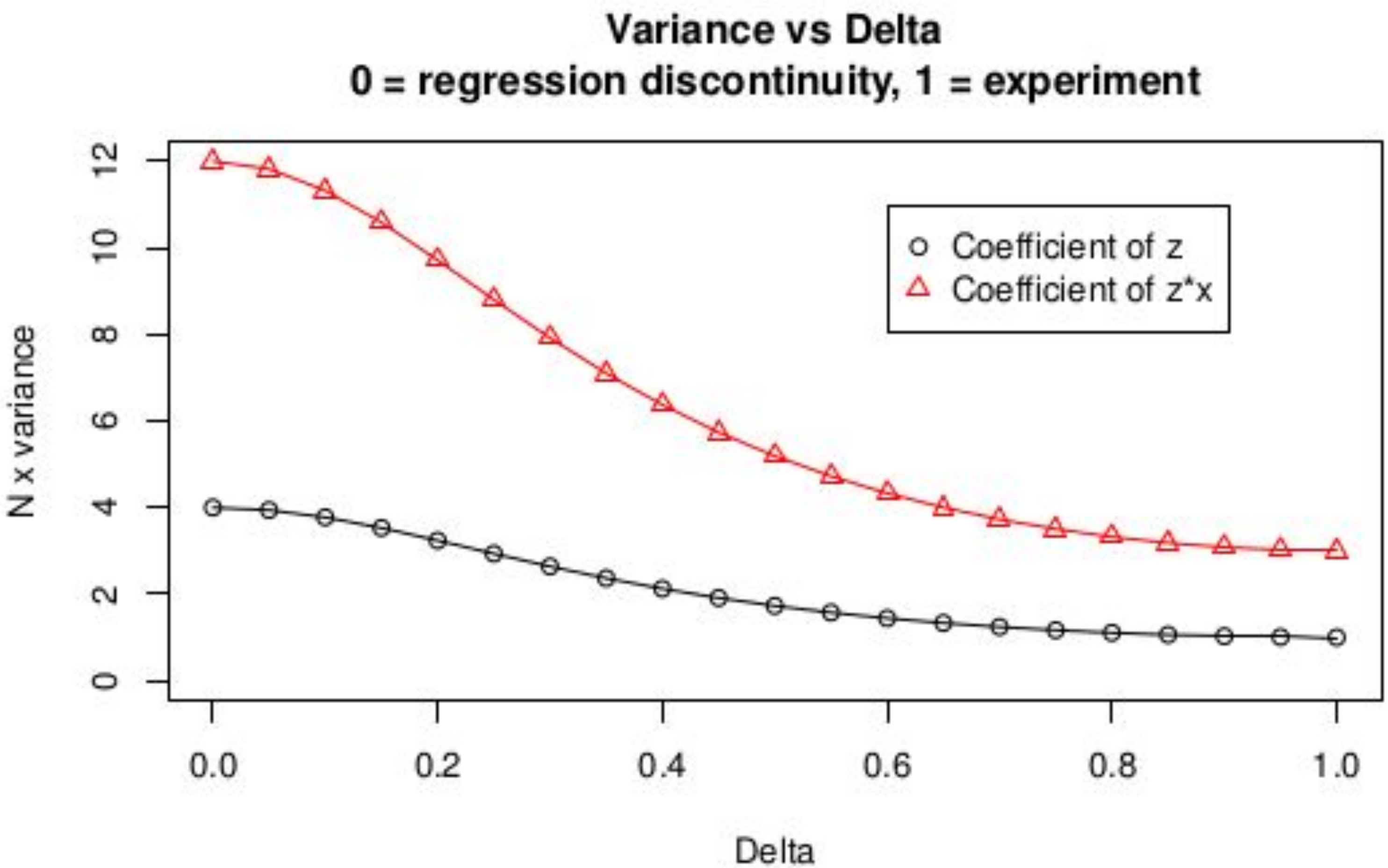
Set up the model

- $Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i + \varepsilon_i$
  - Y_i = impact of treatment;
  - x_i = assignment variable (aka "running variable")
  - z_i = treatment $(-1, +1)$
  - Assume $\beta_3 > 0$. (If not, reverse x.)
- Treatment assignment:

$$z_i = \begin{cases} +1, & x_i \geq \Delta \\ \text{random choice of } +1 \text{ or -1}, & |x_i| < \Delta \\ -1, & x_i \leq -\Delta \end{cases} \tag{1}$$

- How does variance of (β_2, β_3) change as Δ changes?

Benefit of randomization: tighter coefficient estimates



Cost of randomization: ranking is perturbed

- Alas, randomization disturbs the ranking
- Tradeoff: tighter estimate vs noisier ordering
- Example of classic “explore v exploit” tradeoff
  - More experimentation now leads to better optimization in future
  - Randomize now, reap rewards come later!

Objective function

- $v(\Delta) = g(\Delta) + \lambda p(\Delta)$
  - $g(\Delta) =$ expected gain per subject
  - $p(\Delta) =$ precision of estimate
  - $\lambda =$ weight on precision
- Some calculation yields:

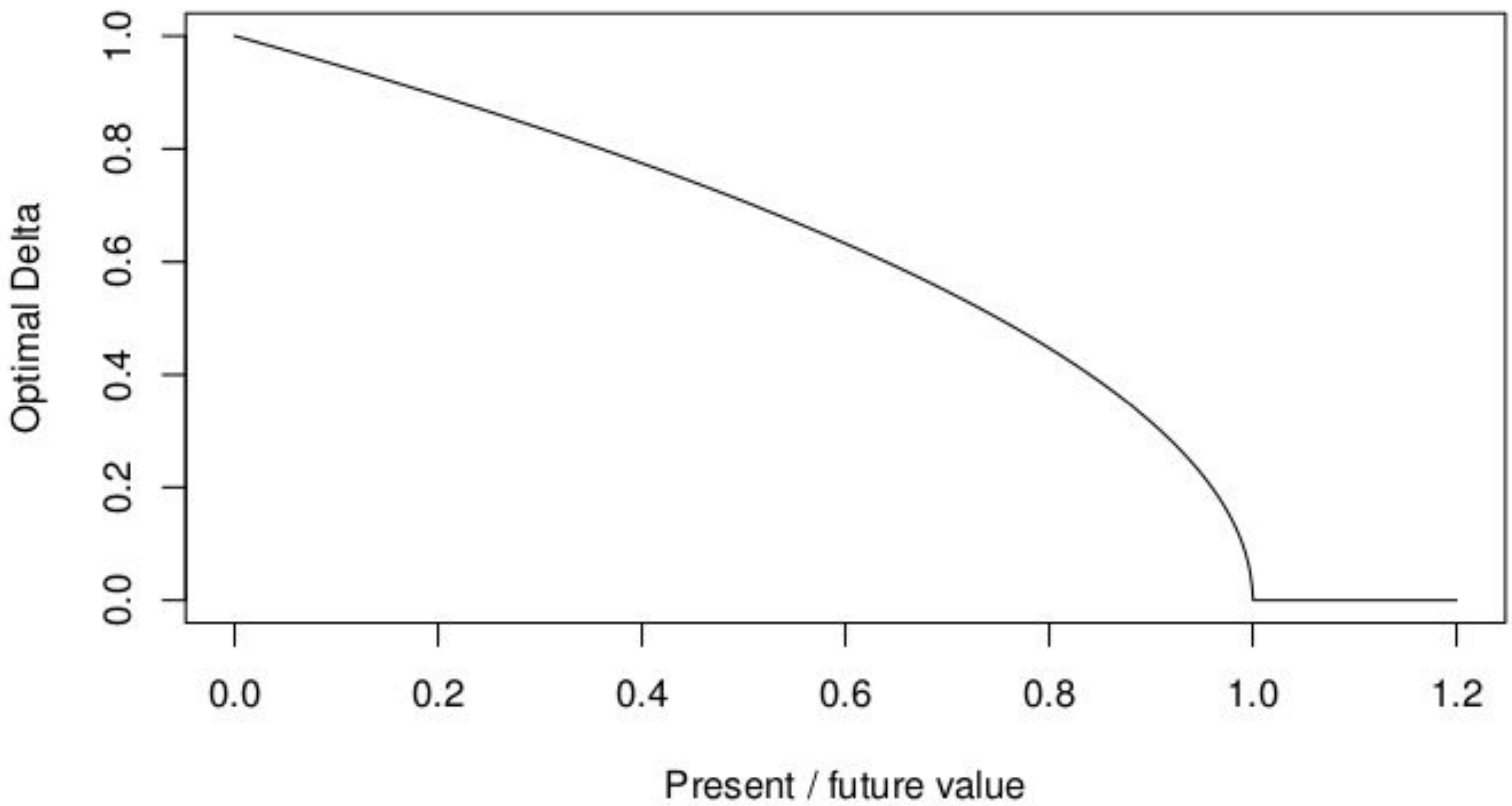
$$v(\Delta) = \beta_0 + \beta_3 \frac{1 - \Delta^2}{2} + \lambda \Big(\frac{1}{3} - \frac{(1 - \Delta^2)^2}{4} \Big)$$

Optimal Δ

$$\Delta^* = \begin{cases} 1, & \beta_3/\lambda \leq 0 \\ \sqrt{1 - \beta_3/\lambda}, & 0 \leq \beta_3/\lambda \leq 1 \\ 0, & 1 \leq \beta_3/\lambda. \end{cases}$$

- $\beta_3 =$ efficiency of ordering
- $\lambda =$ value of higher precision
- Full RCT ($\Delta = 1$) is optimal when $\beta_3 = 0$
- Full RDD ($\Delta = 0$) is optimal when $\lambda \leq \beta_3$
- Otherwise a hybrid approach is preferred

Optimal Δ plot



Summary

- RDD is a very cheap experiment since you do what you would do anyway
- The only extra cost is gathering data on those not treated
- But RCT is 4 times more efficient
- A hybrid experiment gives you the best of both worlds
  - Only disturbs the ordering a little bit around cuttoff
  - But can improve statistical efficiency significantly
- Amount of randomization (size of Δ) depends on exploration/exploitation tradeoff