SPE International

Gulf Coast Section

Data Analytics Study Group

# Data Science Convention

April 04, 2019 Norris Confrence Center, CityCenter Houston

# Interrupted Time Series Regression for the Estimation of Intervention Effects

Justin Clark

Data Scientist

California Resources Corporation

CALIFORNIA RESOURCES CORPORATION

# Agenda

- Introduction to the Interrupted Time Series model
  - Value of ITS
  - When should ITS be used?
  - Standard ITS specification
  - *Centered* ITS
  - Modeling nonlinear decline behavior
  - Alternative Regression Methods
  - Challenges to Validity
  - Examples from Low Volume Acid Maintenance Program in Long Beach, CA

# Project Statement

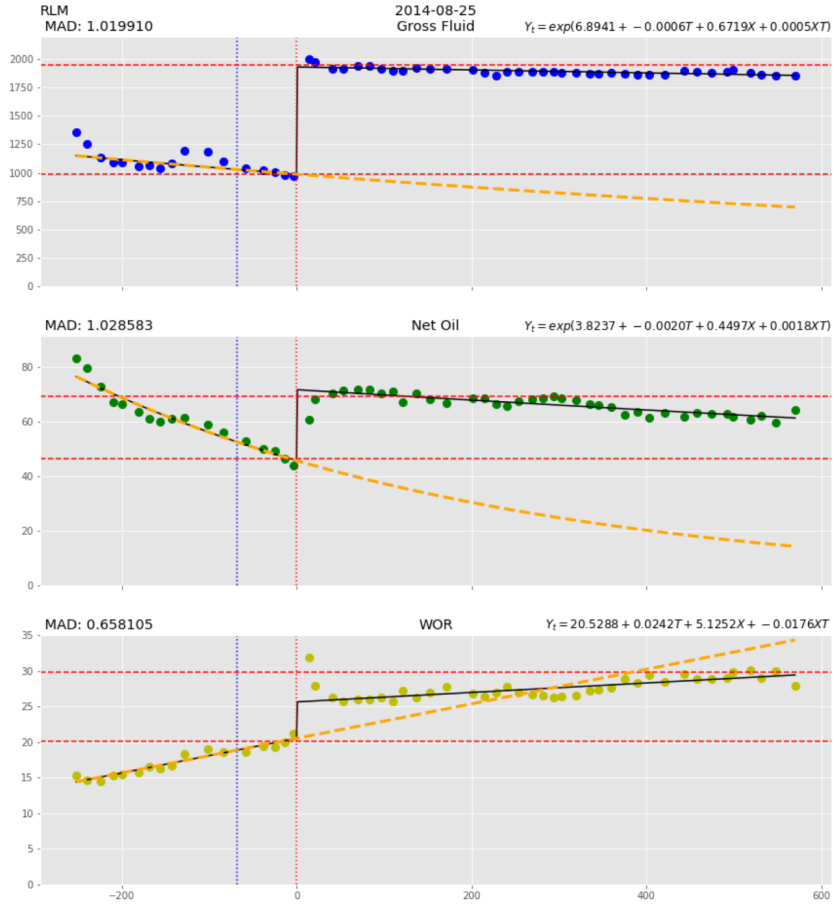
“Can you improve our producer acid maintenance program?”

-Manager

“I don’t know” - Me

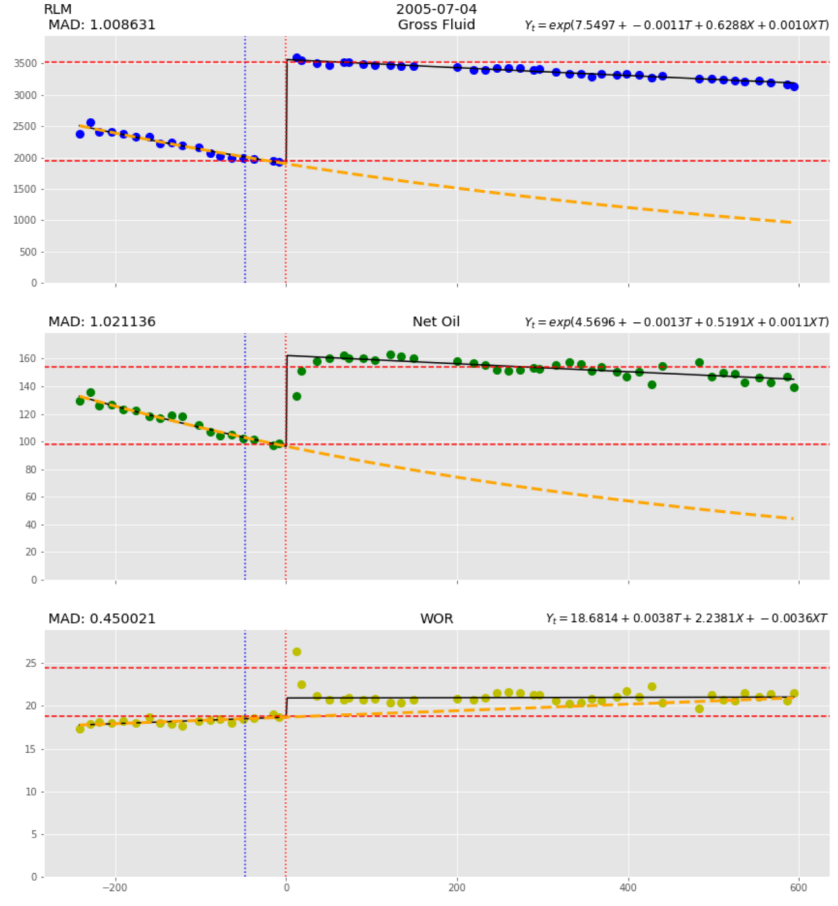
# Project Statement

- What do you mean by “**improve**?”
- How is/should performance be estimated?
- What does historical performance look like?
- Outcomes of this step will be used as target variables in future analysis
- 630 Low Volume Producer Acid Maintenance Jobs in data set (small data)



# What is Interrupted Time Series Regression?

- **Interrupted Time Series Regression**: A linear regression technique used to estimate the causal effect of policy interventions



# Value of Interrupted Time Series

## Pros

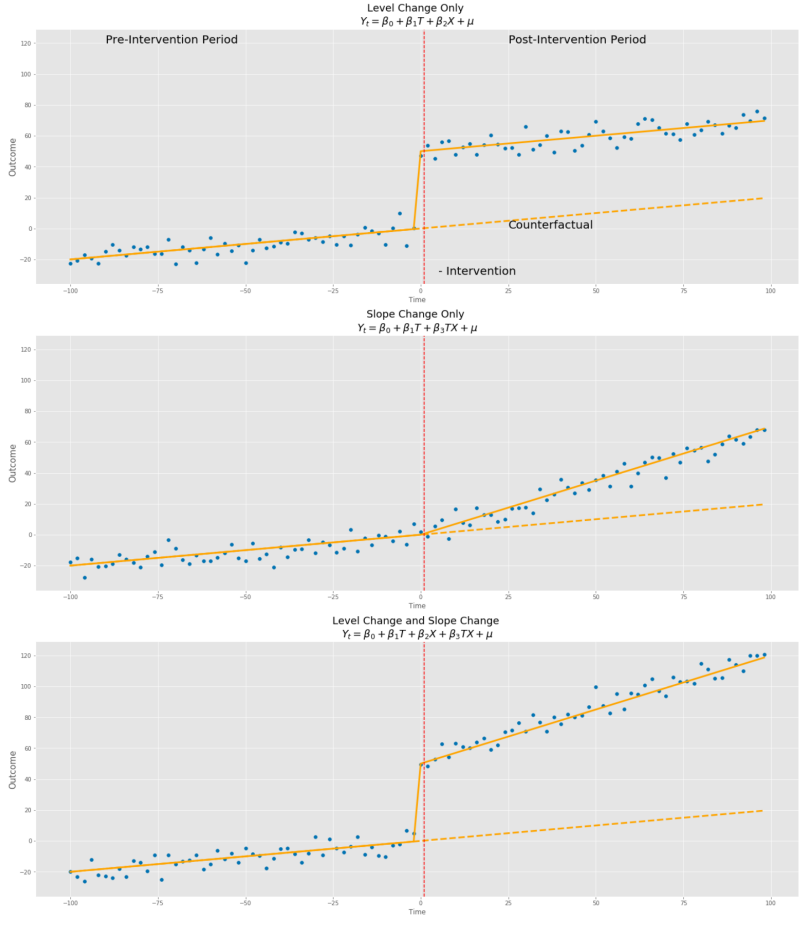
- Allows the estimation of the effects of policy changes
- ITS is simple: Simple to use, simple to interpret
- ITS is the strongest quasi-experimental design
- Automatic counterfactual forecast
- Randomized Control Trials (RCT) are expensive and often infeasible

## Cons

- Not as good as RCTs
- Time Series signals in the oilfield are easily compromised by non-intervention effects that may be difficult to model
- Hypothesis space limited to class of linear models.

# When can ITS be used?

- We have time series data
- We are interested in modeling the effect of a policy intervention
- The data can appropriately be modeled using a linear form
- We have a sufficient number of pre and post intervention observations
- The intervention began at a specific and known time
- The period over which the intervention is introduced is well defined
- We have an a priori belief regarding the expected intervention response



# Standard ITS Model Specification

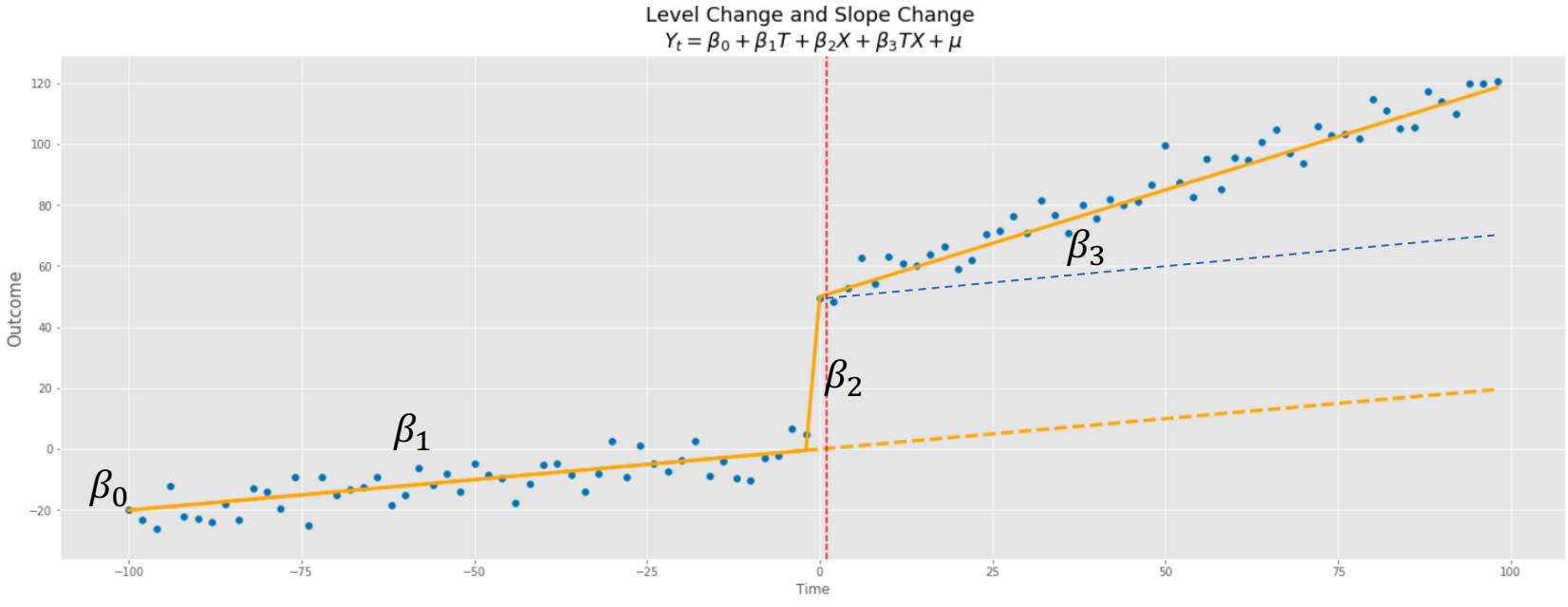
$$Y = \beta_0 + \beta_1 T + \beta_2 X + \beta_3 XT + \epsilon$$

**Y**: The outcome of interest

**T**: Time from 1,…,N, in common units (days, weeks, months)

**X**: A dummy variable that takes the value of 0 for T ≤ Ti, and the value 1 for T > Ti. Where Ti is the time of the intervention.

# Standard ITS Model




Pre-Intervention trend

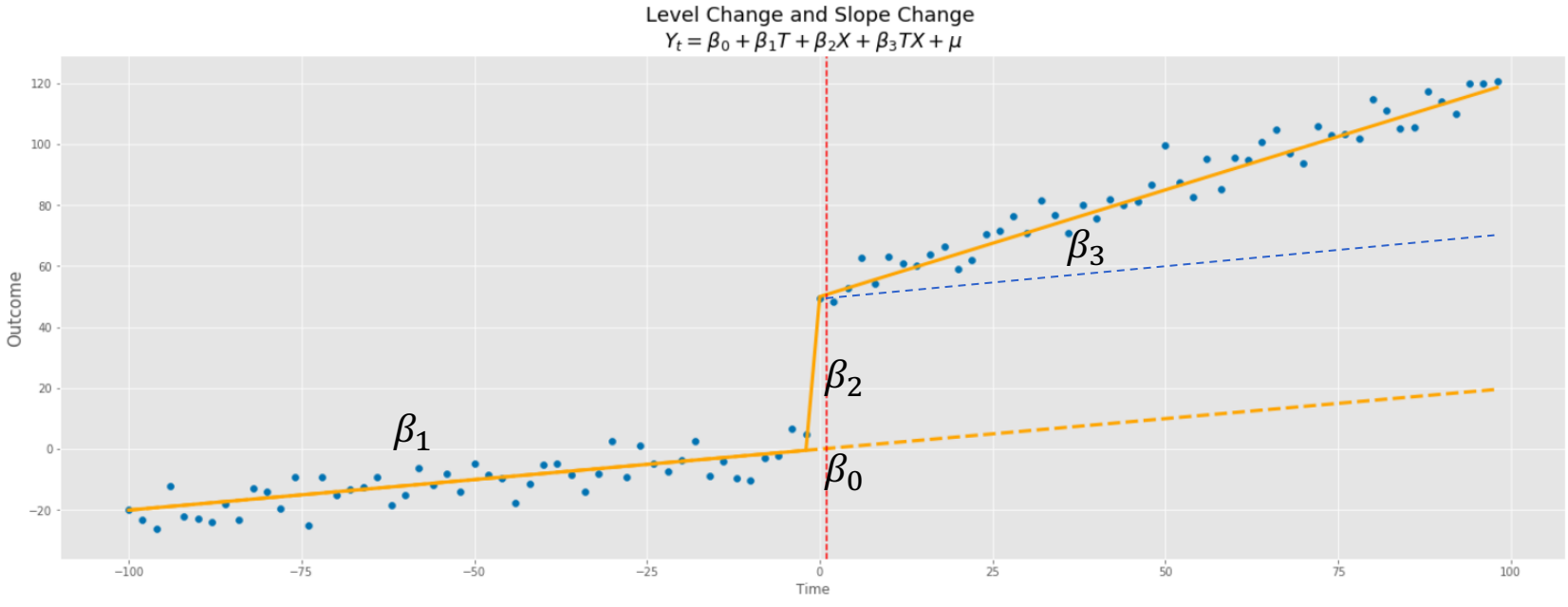
Trend change

$$Y = \beta_0 + \beta_1 T + \beta_2 X + \beta_3 XT + \epsilon$$

Initial Level

Level change

# Centered ITS Model




Pre-Intervention trend  Trend change

$$Y = \beta_0 + \beta_1 T + \beta_2 X + \beta_3 XT + \epsilon$$

Terminal Pre rate  Level change

# Exponential Decline ITS

$$\ln(Y) = \beta_0 + \beta_1 T + \beta_2 X + \beta_3 XT + \epsilon$$

$$Y = \exp\{\beta_0 + \beta_1 T + \beta_2 X + \beta_3 XT + \epsilon\}$$

We model gross production and net production using this model

In this form additive effects become multiplicative effects

# Data Setup

- Event_Date = 9/16/2012
- T = Datediff('d',Date,Event_Date)
- X_t = (T > 0)
- We take log transform of Gross and Net Oil

| Date | Gross | Net Oil | WOR | $X_t$ | T | $T*X_t$ |
|---|---|---|---|---|---|---|
| … | … | … | … | … | … | … |
| 7/9/2012 | 1331.5 | 32 | 39.8 | 0 | -69 | 0 |
| 7/25/2012 | 1234.8 | 31.5 | 38.2 | 0 | -53 | 0 |
| 8/7/2012 | 1229 | 31 | 38.2 | 0 | -40 | 0 |
| 8/26/2012 | 1216.9 | 32 | 36.7 | 0 | -21 | 0 |
| 9/14/2012 | 1213.7 | 32 | 36.2 | 0 | -2 | 0 |
| 9/17/2012 | 1767 | 39 | 44.3 | 1 | 1 | 1 |
| 9/23/2012 | 1766.8 | 39 | 43.9 | 1 | 7 | 7 |
| 9/24/2012 | 1753.6 | 38 | 45.1 | 1 | 8 | 8 |
| 9/25/2012 | 1735.8 | 35 | 47.5 | 1 | 9 | 9 |
| … | … | … | … | … | … | … |

Outcomes of interest

# Data

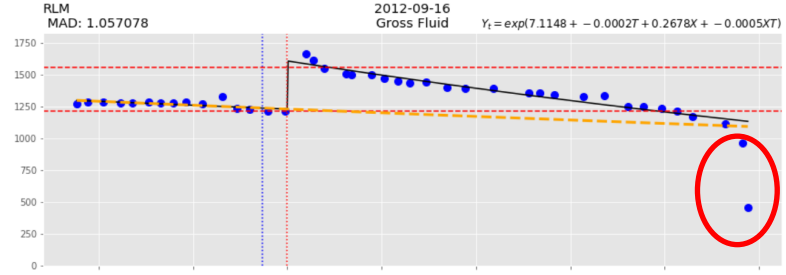
## Issue

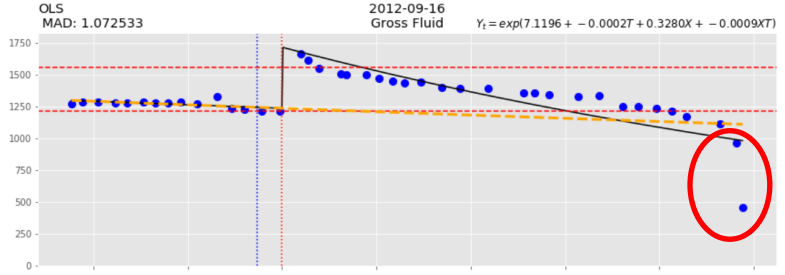
- Irregularly spaced data: Higher frequency during early production
- Outliers
- Insufficient observations
- Inclining pre-intervention production
- Non-exponential / hyperbolic decline
- Scale to uneconomic
- Coincidental well work, e.g Acid Maintenance + Pump Replacement

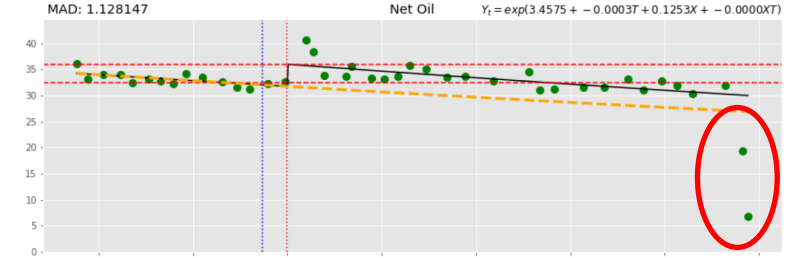
## Solution

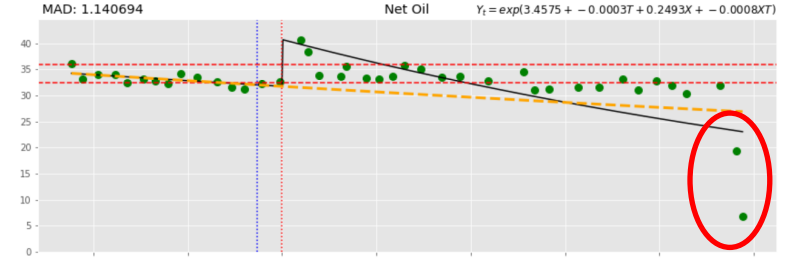
- Resample data to mode sampling frequency, biweekly.
- Use Robust Regression
- Discarded before fit
- Discarded after fit
- Pre-Intervention Data Truncation
  - 200 Days prior to day job scheduled in system.
- No Solution Implemented
- Discarded

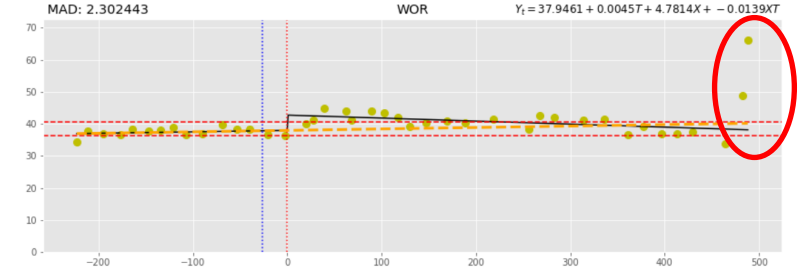
# Outliers - Robust Regression vs OLS



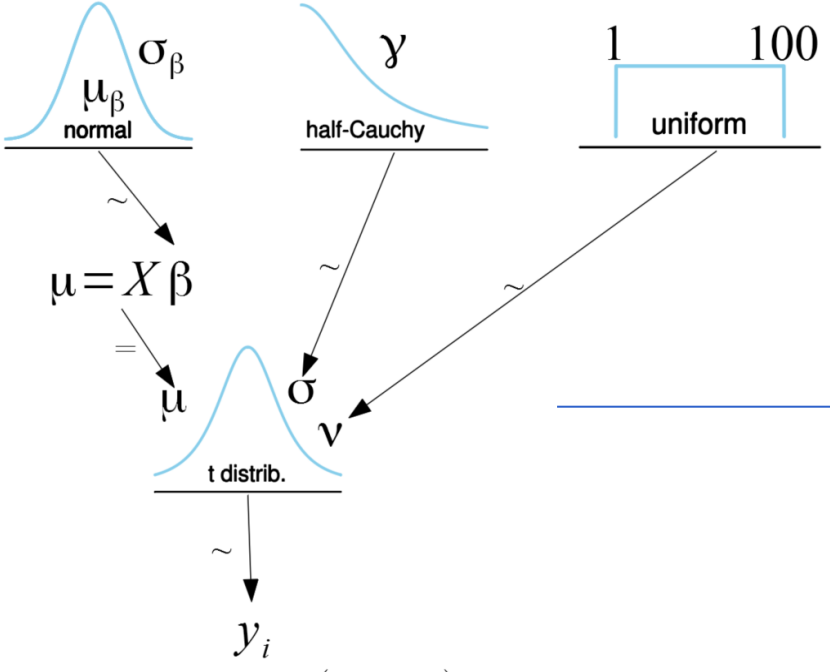



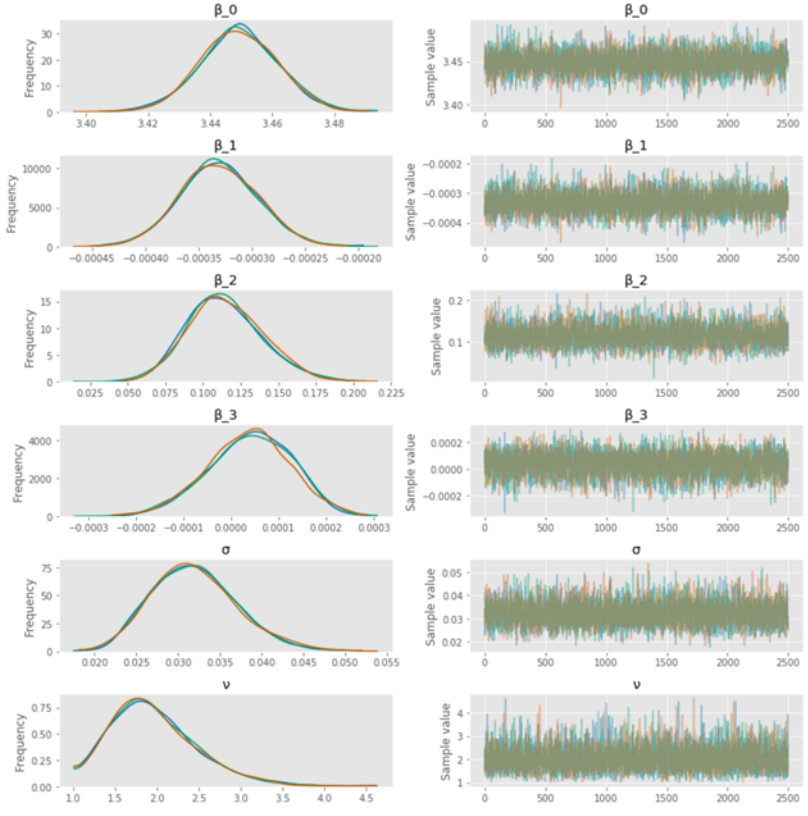

# Outliers – Bayesian Regression – “Studentized”

Prior

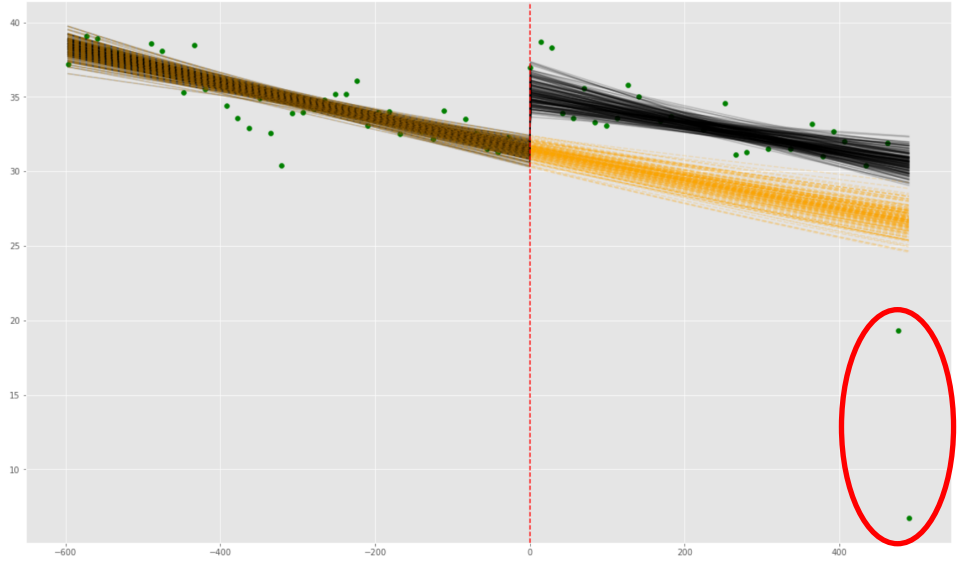



$$y_i \sim \text{StudentT}(\mu, \sigma, \nu)$$
$$\mu \sim \text{Normal}(\mu_\beta, \sigma_\beta)$$
$$\sigma \sim \text{Half-Cauchy}(\gamma)$$

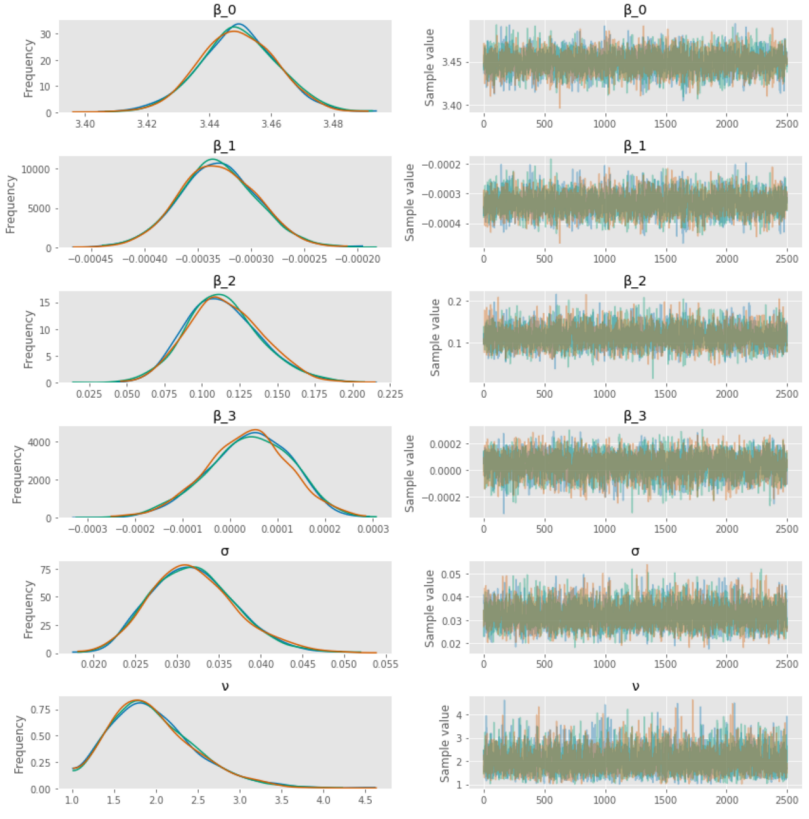
Posterior



# Outliers - Bayesian Regression – “Studentized”



Posterior




:

| | mean | sd | mc_error | hpd_2.5 | hpd_97.5 | n_eff | Rhat |
|---|---|---|---|---|---|---|---|
| **β_0** | 3.449807 | 0.012747 | 2.347507e-04 | 3.424910 | 3.474907 | 3405.407859 | 1.000728 |
| **β_1** | -0.000330 | 0.000038 | 6.984639e-07 | -0.000401 | -0.000253 | 3462.058504 | 1.000844 |
| **β_2** | 0.114960 | 0.025650 | 4.178945e-04 | 0.066767 | 0.164904 | 3746.702074 | 1.000155 |
| **β_3** | 0.000033 | 0.000090 | 1.699922e-06 | -0.000139 | 0.000210 | 3619.713255 | 1.001198 |
| **σ** | 0.032126 | 0.005061 | 7.945818e-05 | 0.022733 | 0.042328 | 5131.284969 | 1.000058 |
| **ν** | 1.986191 | 0.538352 | 8.037547e-03 | 1.081039 | 3.026092 | 5193.961618 | 0.999908 |

# Additional Challenges to Validity

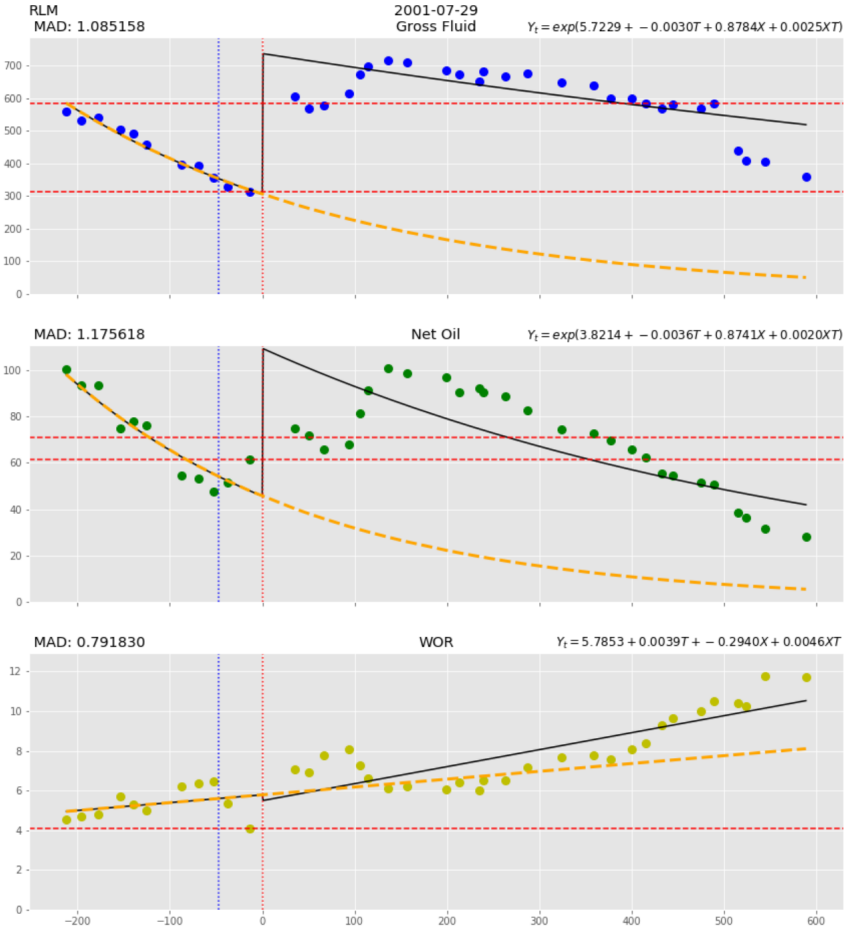
## Challenges

- Unmodeled effects
- Autocorrelation (i.i.d violation)
- Simultaneous intervention
- Gradual Effect
- Delayed Effect
- Temporary Effect
- Erroneous Data
  - Missing Intervention labels

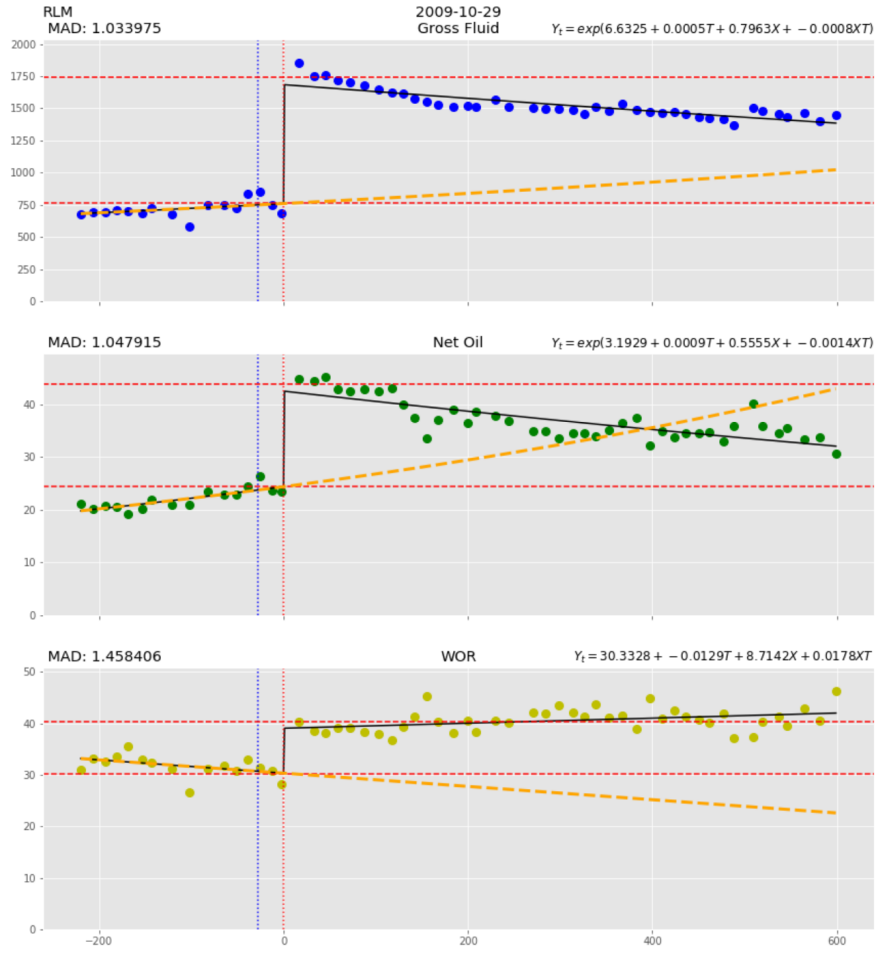
## Solutions

- Consider including control variables
- Use heteroskedasticity and autocorrelation robust standard errors (Newey-West), or use model that accounts for autocorrelation (ARIMA)

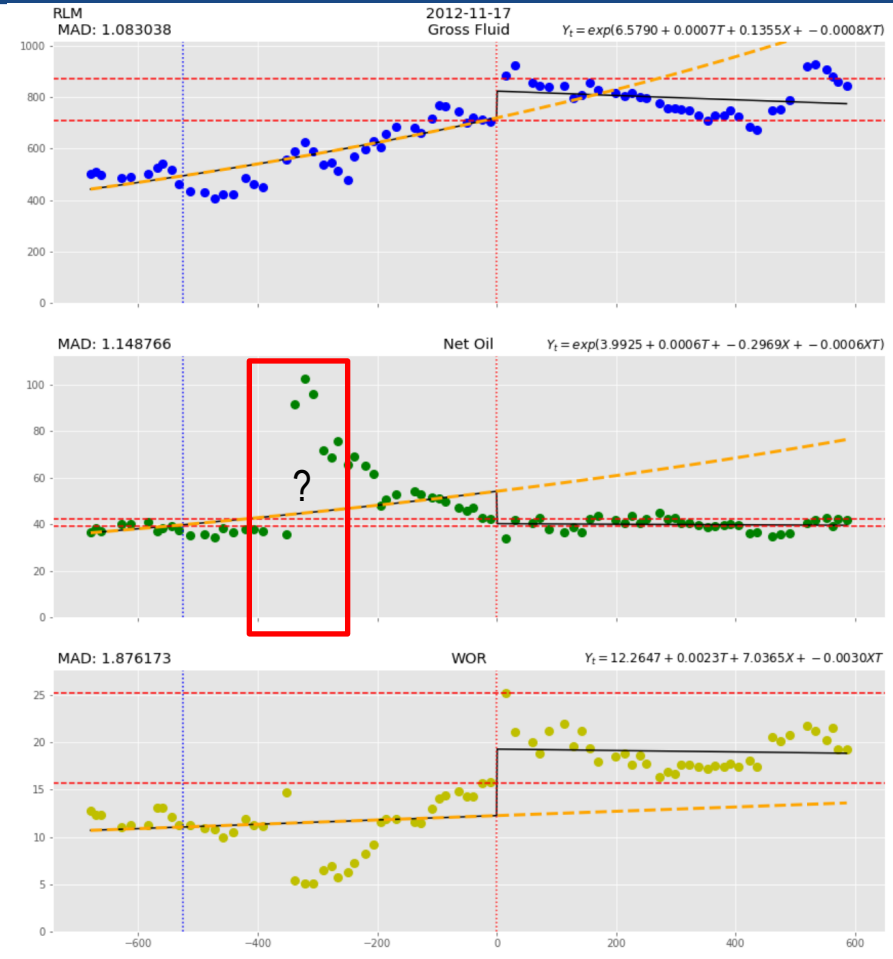
# Gradual Effect Example



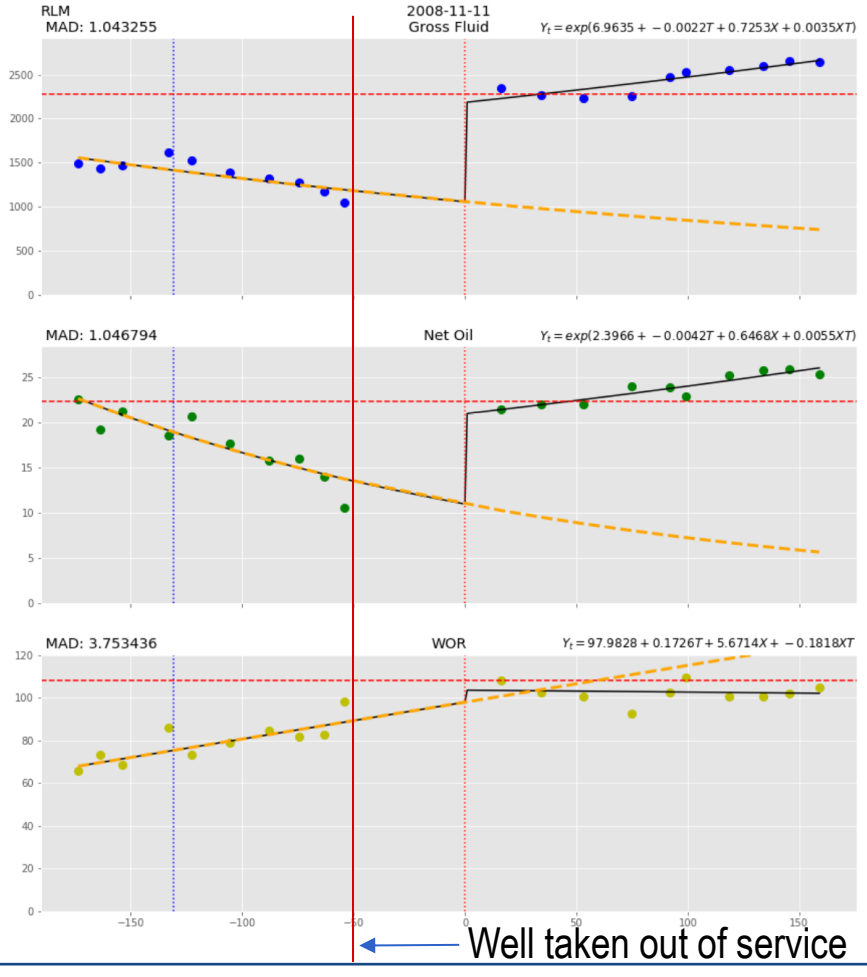
# Inclining Production Example



# Erroneous Data – Missing Events

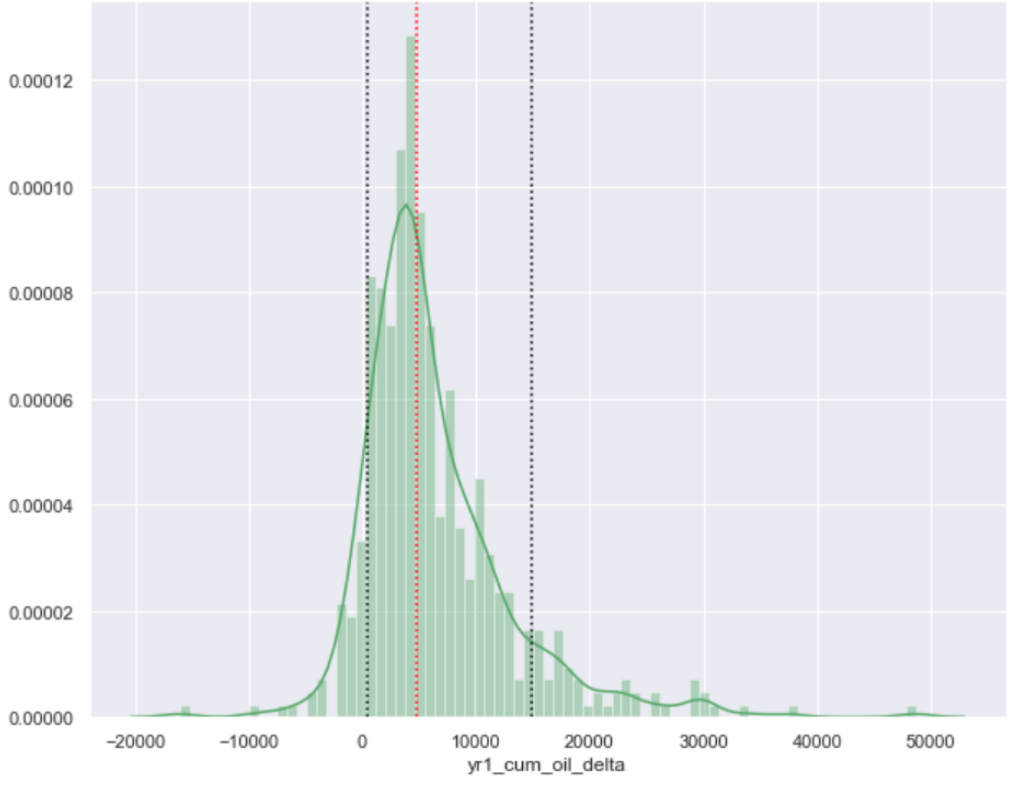


# Scale to Uneconomic



# 1 Year Incremental Cumulative

```
Mean: 6429.440170972158
10th Pctile: 497.4827429460315
50th Pctile: 4751.906340676091
90th Pctile: 14813.906427023663
```



# Conclusion

- The Interrupted Time Series model is a simple and effective tool for estimating the effects of policy intervention
- Start Simple. Disregard the additional complexity (at first) of the robust and Bayesian methods

- Thank you to SPE for inviting me to speak
- Special thanks to CRC for supporting the Data Analytics team and for allowing me to present

Q&A

CALIFORNIA
RESOURCES CORPORATION

# Where to Learn More