TAP TO GO BACK TO KIOSK MENU

SAS® GLOBAL FORUM 2020

MARCH 29 – APRIL 1
WASHINGTON, DC

USERS PROGRAM

BUILDING TWO REGRESSION LINES IS BETTER THAN BUILDING ONE

TAP TO GO BACK TO KIOSK MENU

G Liu

Please use the headings above to navigate through the different sections of the poster

In a time series, there can be a change in parameter coefficient or error variance at a time point *k*. Traditional simple regression modeling can limit the accuracy and predictability of this time series if this change is not accounted for. Utilizing PROC UCM in SAS/ETS®, the structural break at *k* can be identified so that two separate regression lines can be modeled to better represent the data, and to provide us with better predictions for future time periods.

Introduction	A simple walk-through of how data with structural break look like.
Methods	A detailed PROC UCM step-by-step guide to find the break point *k* and develop better models for the Nile River water level data.
Application 1	Using the structural break analysis to identify a change in the stock market trend of the SPY ETF closing prices.
Application 2	Applying this structural break idea to traditional non-time series data of PROC SQL runtimes.
Conclusion	Summary and thoughts.

BUILDING TWO REGRESSION LINES IS BETTER THAN BUILDING ONE

G Liu

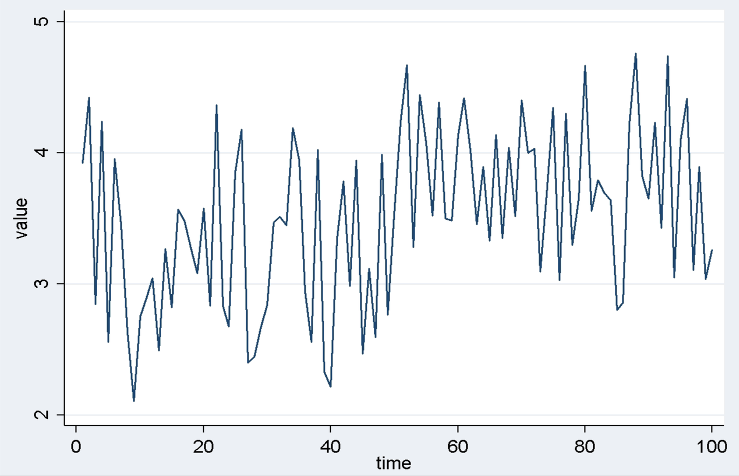
Abstract
Introduction
Methods
Application 1
Application 2
Conclusion

Please use the headings above to navigate through the different sections of the poster

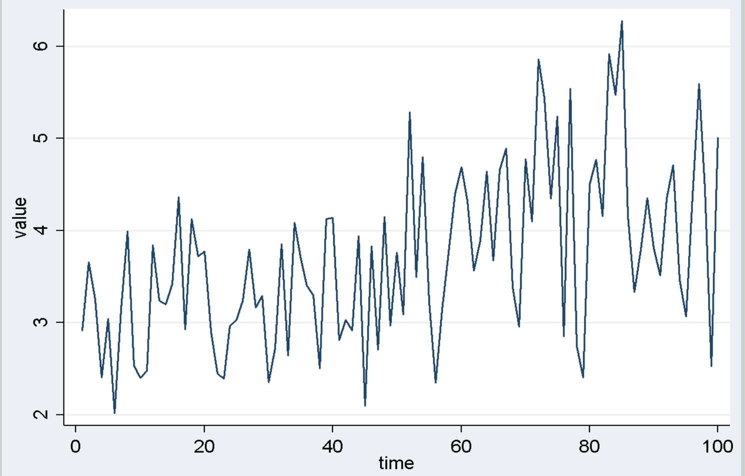
Introduction

A time series with a structural break exhibits changes in parameter coefficients and/or error variance at a time point k such that:

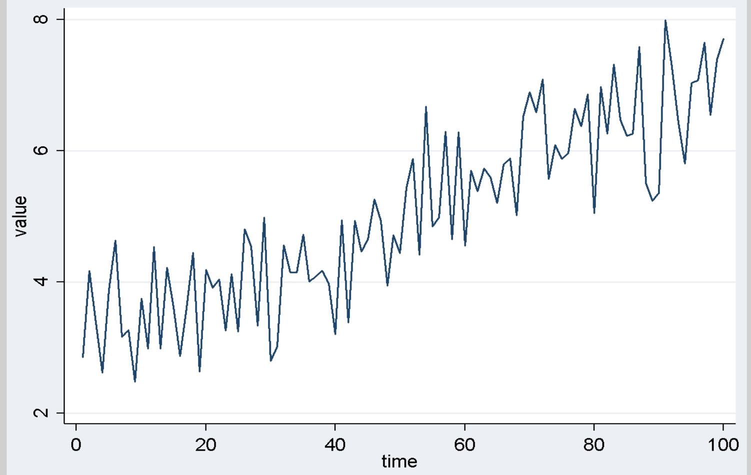
$$
\begin{aligned}
&y_t = \alpha_1 + \beta_1 x_t + \varepsilon_t\,, \quad t=1,\ \ldots,\ k \\
&y_t = \alpha_2 + \beta_2 x_t + \eta_t\,, \quad t=k+1,\ \ldots,\ T \\
&\text{where} \qquad \varepsilon_t \sim i.i.d.\ N(0,\ \sigma_1^2) \\
&\qquad\qquad\quad\ \eta_t \sim i.i.d.\ N(0,\ \sigma_2^2)
\end{aligned}
$$



$\alpha_2 > \alpha_1$



$\alpha_2 > \alpha_1$ and $\sigma_2^2 > \sigma_1^2$



$\beta_2 > \beta_1$

Examples

Nile River water levels (see Methods section)

- Aswan dam was built in 1899, lowering the mean water level, $\alpha_2 < \alpha_1$

SPY ETF closing prices (see Application 1 section)

- A break of an uptrend in closing prices, $\beta_2 < \beta_1$

SAS® Proc SQL runtimes (see Application 2 section)

- Although not a time series, this data mimic $\sigma_2^2 > \sigma_1^2$

PROC UCM

PROC UCM can be used to perform the structural break analysis.

Syntax: UCM Procedure

The UCM procedure uses the following statements:

PROC UCM <options>;
AUTOREG <options>;
BLOCKSEASON options;
BY variables;
CYCLE <options>;
DEPLAG options;
ESTIMATE <options>;
FORECAST <options>;
ID variable options;
IRREGULAR <options>;
LEVEL <options>;
MODEL dependent variable <= regressors>;
NLOPTIONS options;
PERFORMANCE options;
OUTLIER options;
RANDOMREG regressors </ options>;
SEASON options;
SLOPE <options>;
SPLINEREG regressor <options>;
SPLINESEASON options;
TF regressor <options>;

The PROC UCM and MODEL statements are required. In addition, the model must contain at least one component with nonzero disturbance variance.

BUILDING TWO REGRESSION LINES IS BETTER THAN BUILDING ONE

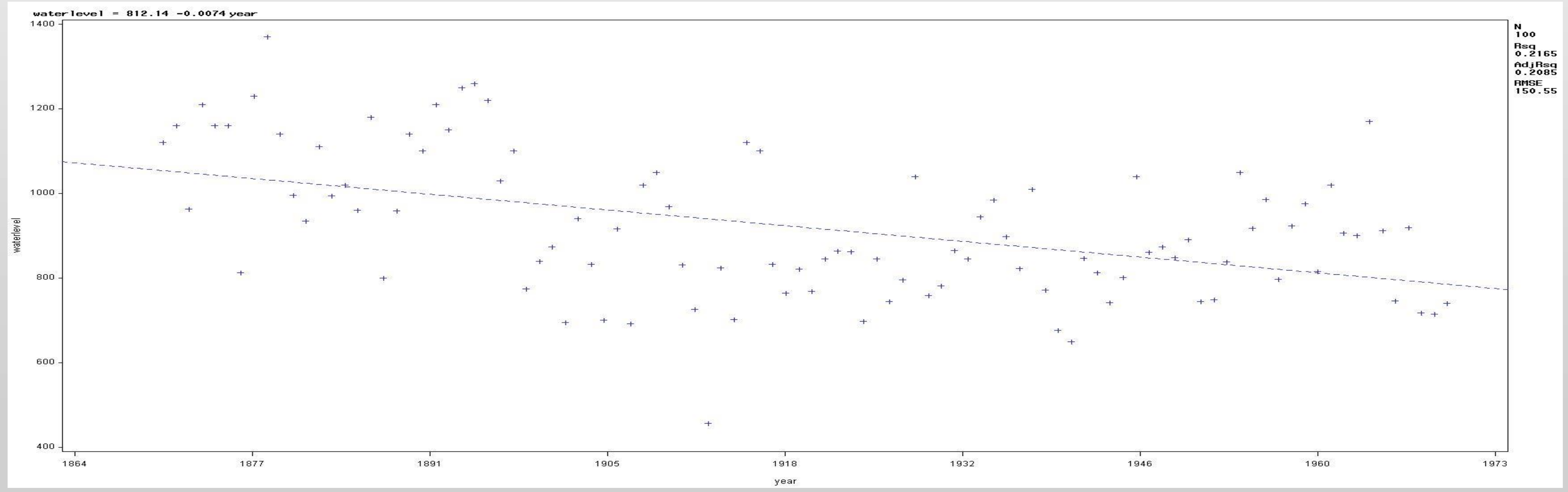
G Liu

Abstract
Introduction
Methods
Application 1
Application 2
Conclusion

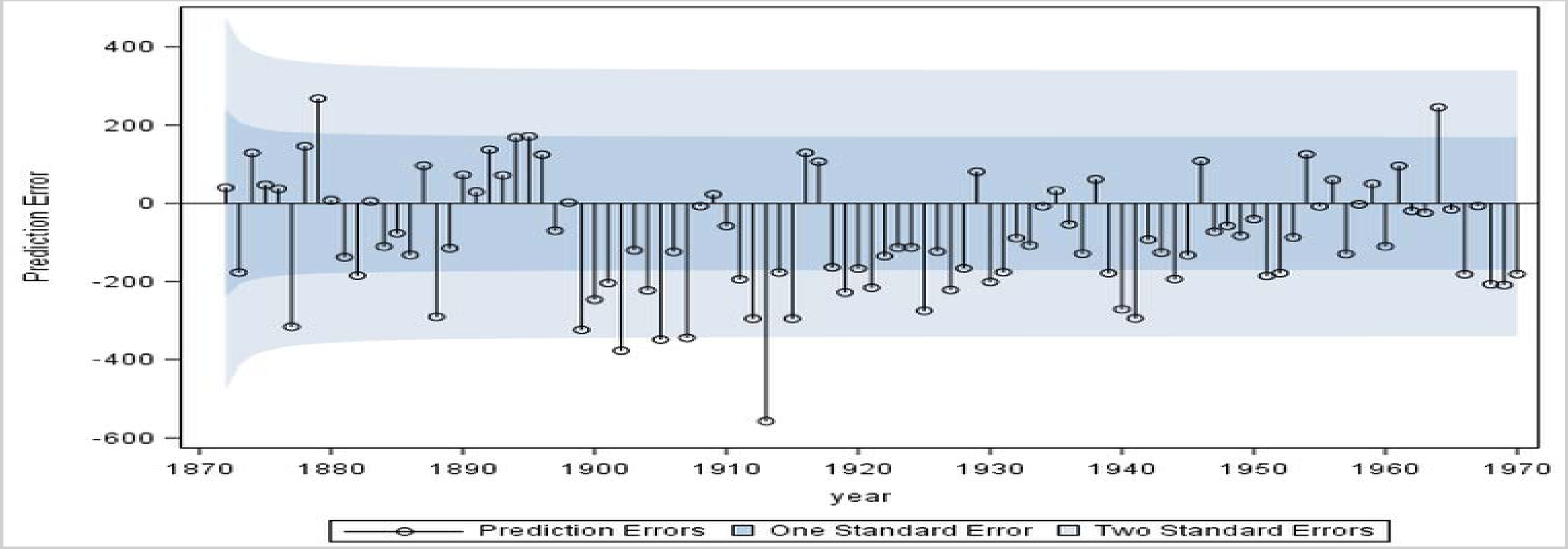
Please use the headings above to navigate through the different sections of the poster

Nile River Water Level Data

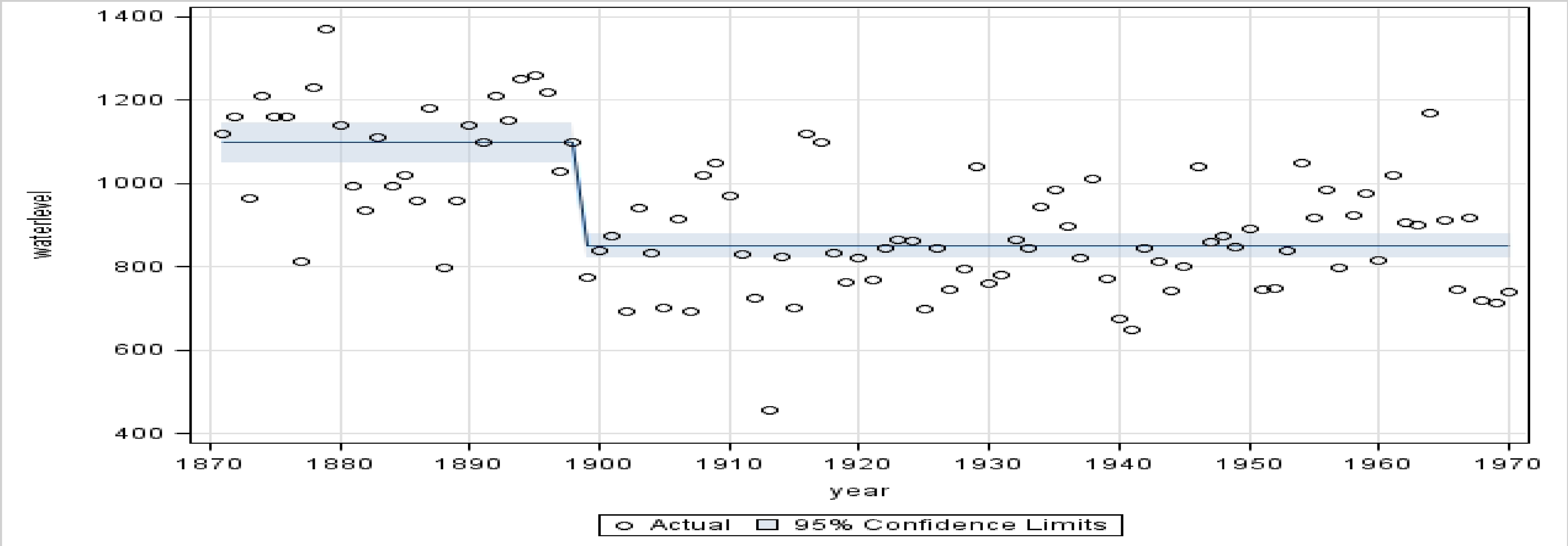
- 100-year water level data of Nile River
- Aswan dam built in 1899, lowering water level
- PROC REG with RMSE = 150.55



- Simple regression line resulted in heteroscedastic residuals



- PROC UCM with RMSE = 128.26



- More importantly, the forecasted water level using PROC REG would be significantly lower and less predictive than that of PROC UCM

Step-by-step Guide

```
/* Find break point */
/* Step 1 - Find components that are significant */
proc ucm data=nile;
    id year interval=year;
    model waterlevel;
    autoreg;
    irregular;
    level;
    slope;
run;
/* Result: autoregressive and slope components not significant */
```

Significance Analysis of Components (Based on the Final State)

Component	DF	Chi-Square	Pr > ChiSq
Irregular	1	0.19	0.6631
Level	1	187.08	<.0001
Slope	1	0.00	0.9988
AutoReg	1	1.15	0.2826

```
/* Step 2 - Drop autoregressive and slope components */
proc ucm data=nile;
    id year interval=year;
    model waterlevel;
    irregular;
    level;
run;
/* Result: level component has insignificant error variance term */
```

Final Estimates of the Free Parameters

Component	Parameter	Estimate	Approx Std Error	t Value	Approx Pr > |t|
Irregular	Error Variance	15099	3145.5	4.80	<.0001
Level	Error Variance	1469.17636	1280.4	1.15	0.2512

```
/* Step 3 - Set level variance term =0 and do not ask for estimate (noest) */
/* Print outlier diagnostics details, CUSUM plots etc */
proc ucm data=nile;
    id year interval=year;
    model waterlevel;
    irregular;
    level plot=smooth checkbreak variance=0 noest;
    outlier print=detail;
    estimate plot=(cusum cusumsq panel residual);
    forecast plot=(forecasts decomp);
run;
```

Outlier Summary

Obs	year	Break Type	Estimate	Standard Error	Chi-Square	DF	Pr > ChiSq
29	1899	Level	-247.77778	37.68996	43.22	1	<.0001

```
/* Create dummy variable representing the shift in mean in year 1899 */
data nile2;
    set nile;
    shift1899 = (year ge mdy(1,1,1899));
run;

/* Step 4 - Run final model with shift1899 as a regressor */
proc ucm data=nile2;
    id year interval=year;
    model waterlevel=shift1899;
    irregular;
    level plot=smooth checkbreak variance=0 noest;
    outlier print=detail;
    estimate plot=(cusum cusumsq panel residual);
    forecast plot=(forecasts decomp);
run;
```

Final Estimates of the Free Parameters

Component	Parameter	Estimate	Approx Std Error	t Value	Approx Pr > |t|
Irregular	Error Variance	16301	2328.7	7.00	<.0001
shift1899	Coefficient	-247.77778	28.43520	-8.71	<.0001

Fit Statistics Based on Residuals

Mean Squared Error	16452
Root Mean Squared Error	128.26485
Mean Absolute Percentage Error	11.97873
Maximum Percent Error	27.31229
R-Square	-0.06232
Adjusted R-Square	-0.06232
Random Walk R-Square	0.35099
Amemiya's Adjusted R-Square	-0.09267

Number of non-missing residuals used for computing the fit statistics = 71

Significance Analysis of Components (Based on the Final State)

Component	DF	Chi-Square	Pr > ChiSq
Irregular	1	53.42	<.0001
Level	1	2069.96	<.0001

BUILDING TWO REGRESSION LINES IS BETTER THAN BUILDING ONE

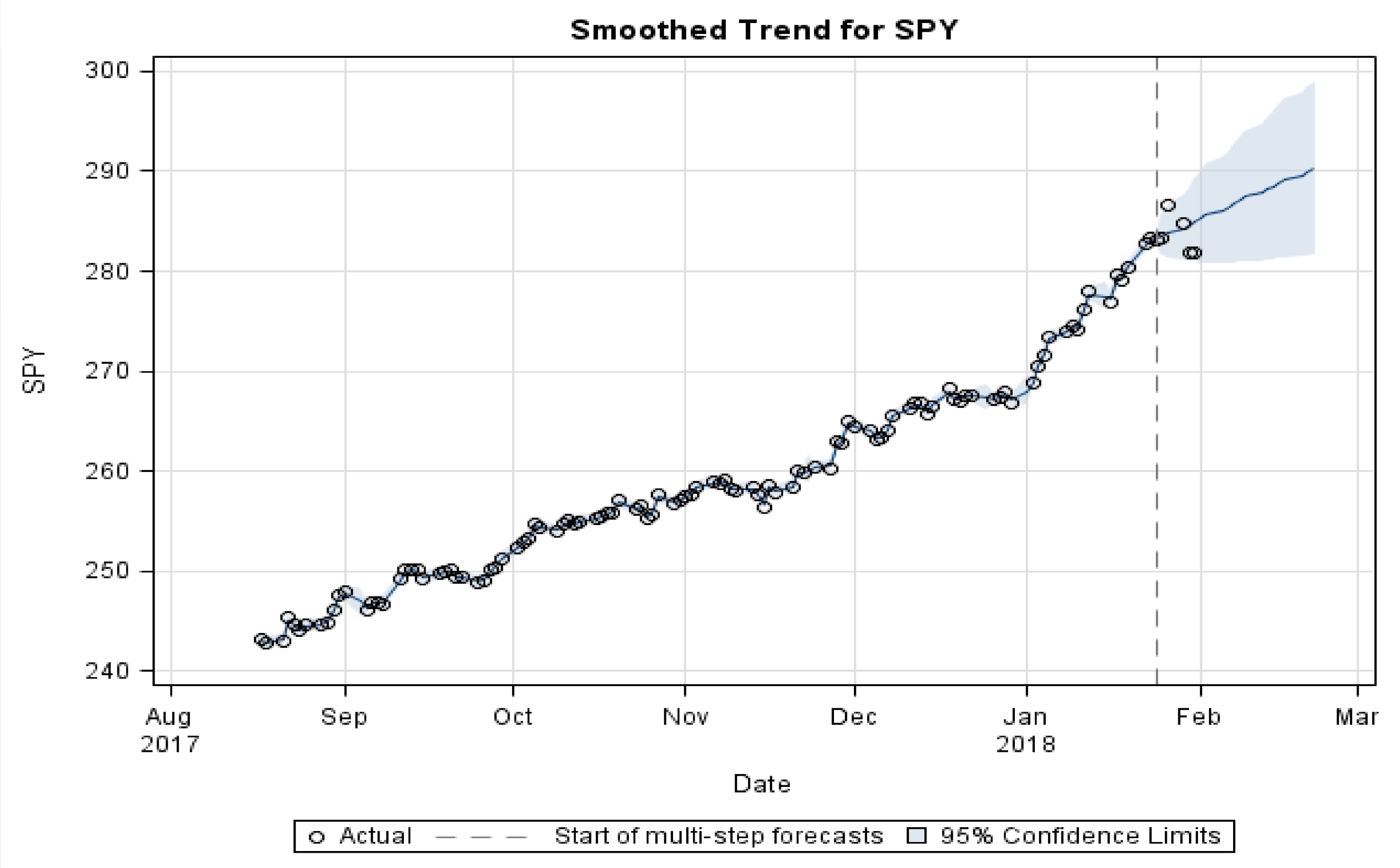
G Liu

Abstract
Introduction
Methods
Application 1
Application 2
Conclusion

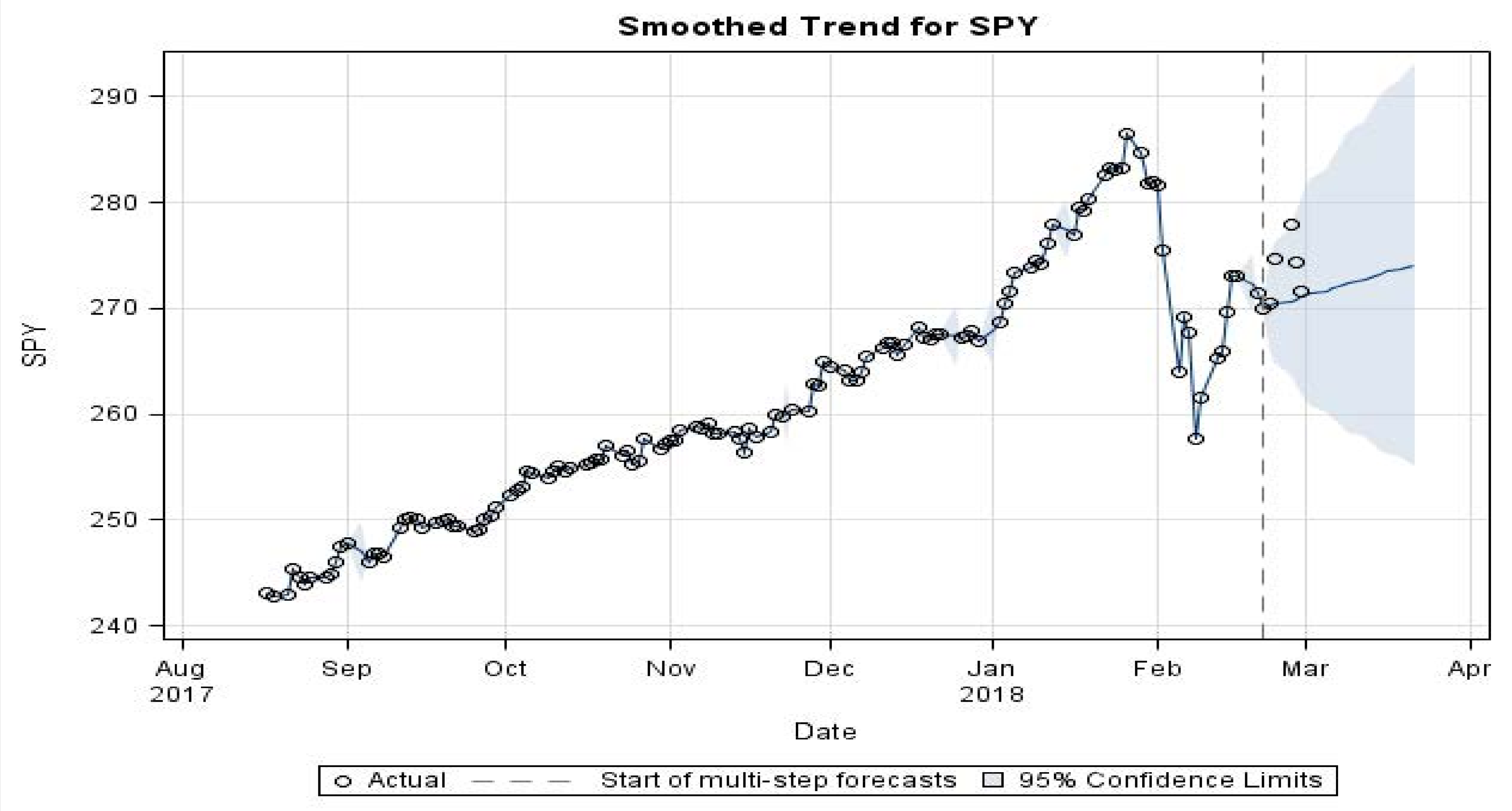
Please use the headings above to navigate through the different sections of the poster

SPY ETF Closing Price Data

- A real-time structural break analysis on Exchange Traded Funds (ETF) SPY daily closing prices starting from Aug 17th, 2017
- Daily analysis using PROC UCM performed to identify the break point indicating the end of an uptrend
- SPY closing prices on Jan 31st, 2018



- SPY closing prices on Feb 28th, 2018



PROC UCM analysis on Jan 31st, 2018

Outlier Summary

Obs	Date	Break Type	Estimate	Standard Error	Chi-Square	DF	Pr > ChiSq
117	26JAN2018	Additive Outlier	2.76431	0.7694568	12.91	1	0.0003
119	30JAN2018	Level	-3.52351	1.0288021	11.73	1	0.0006

Daily analysis results with Jan 30th closing price first identified as an outlier on Jan 30th, subsequently confirmed as a break point on Jan 31st.

Analysis On	Date	Break Type	Estimate	Std Error	ChiSq	DF	Pr > ChiSq
Jan 30, 2018	Jan 26	Additive Outlier	2.80	0.78	12.84	1	0.0003
	Jan 30	Additive Outlier	-3.54	1.04	11.66	1	0.0006
Jan 31, 2018	Jan 26	Additive Outlier	2.76	0.77	12.91	1	0.0003
	Jan 30	**Level**	-3.52	1.03	11.73	1	0.0006
Feb 1, 2018	Jan 26	Additive Outlier	2.75	0.77	12.95	1	0.0003
	Jan 30	**Level**	-3.51	1.03	11.69	1	0.0006
Feb 2, 2018	Feb 2	Additive Outlier	-6.49	1.19	29.63	1	< 0.0001
	Feb 2	Level	-6.49	1.19	29.63	1	< 0.0001

BUILDING TWO REGRESSION LINES IS BETTER THAN BUILDING ONE

G Liu

Please use the headings above to navigate through the different sections of the poster

SAS® PROC SQL Runtime Data

Anecdotally, PROC SQL join operation can have runtimes that vary greatly depending on the size of the datasets. The runtimes are generally proportional to the size but only up to a certain point. Datasets that are much larger seem to have runtimes much longer than anticipated. This sounds very similar to a structural break analysis.

The runtime data generated have the following attributes:

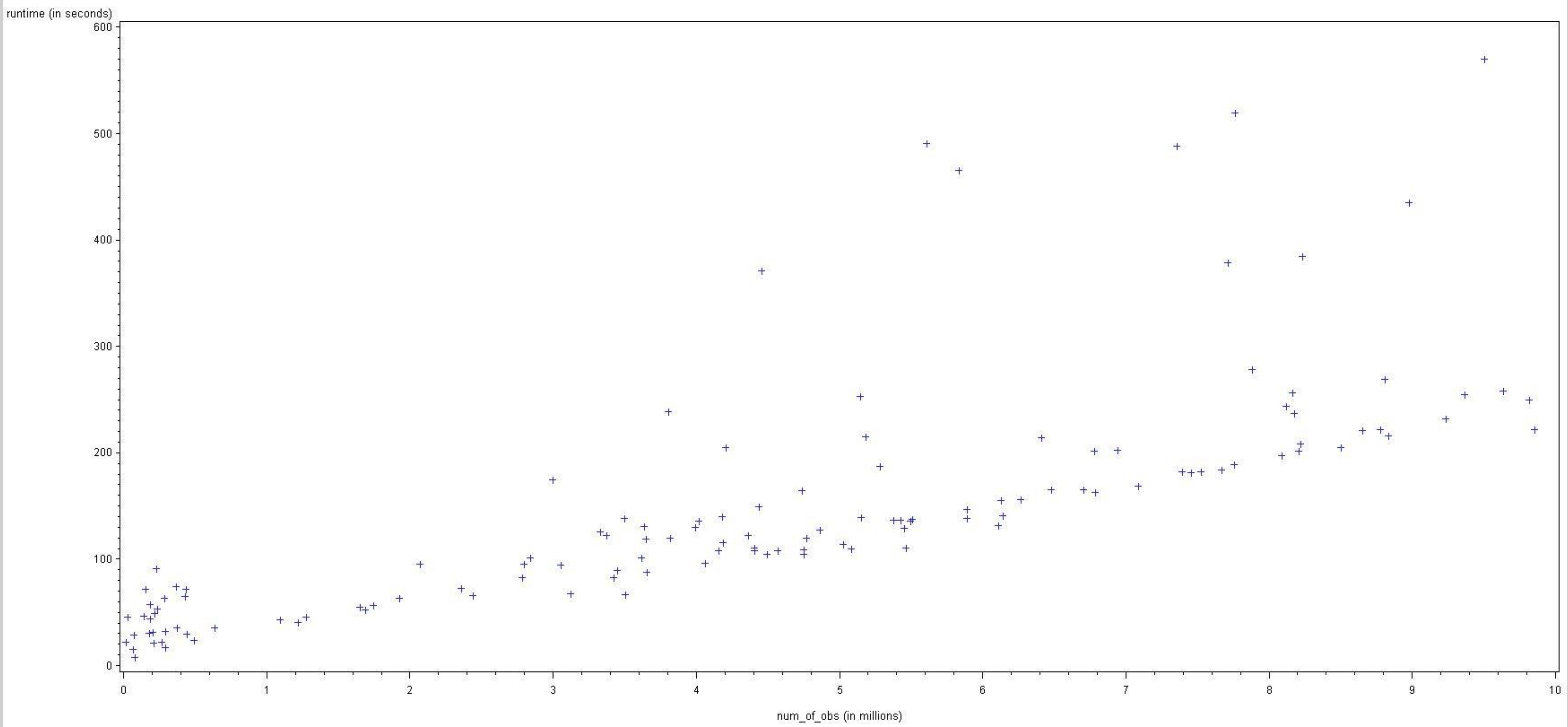
- 173 simulated runtimes
- Primary dataset ***heart*** has varying simulated number of observations from 15K to 10M
- All four reference datasets have the same 50K observations and a range of 11 to 400 variables in all simulations

PROC UCM cannot be used in this case because the independent variable (number of observations in ***heart***) does not have a regular interval like time series data. However, the modeling idea remains the same.

Visually, the break point is somewhere between 1 to 3 million observations, where the variability of runtimes is much greater after 3 million observations. A separate model can be developed using only the runtime data with 3+ million observations.

PROC SQL Code Used to Generate Data

```
%let starttime=%sysfunc(time());
proc sql;
    create table test_t as
    select  *
    from    heart h left join
            m.qtr1001 a on h.age=a.age and h.sex=a.sex left join
            m.mon1001 b on h.age=b.age and h.sex=b.sex left join
            m.mon111  c on h.age=c.age and h.sex=c.sex left join
            m.qtr111  d on h.age=d.age and h.sex=d.sex;
quit;
%let stoptime=%sysfunc(time());
```



BUILDING TWO REGRESSION LINES IS BETTER THAN BUILDING ONE

G Liu

Please use the headings above to navigate through the different sections of the poster

CONCLUSION

- Structural break can occur when parameter coefficient and/or error variance change at a particular time point
- PROC UCM can be used to find the break point, and to model the time series with a structural break
- Better models and predictions when structural break is accounted for
- Real time structural break analysis on financial data is also an application of PROC UCM
- Awareness of possible structural break in traditional non-time series data can be helpful in selecting data for modeling and predictions

References

SAS®, 2018. SAS/ETS® 15.1 User's Guide.

SAS®
GLOBAL FORUM 2020

USERS PROGRAM

MARCH 29 - APRIL 1 | WASHINGTON, DC | #SASGF