RegARIMA Model Diagnostics

Seasonal Adjustment With X-13ARIMA-SEATS

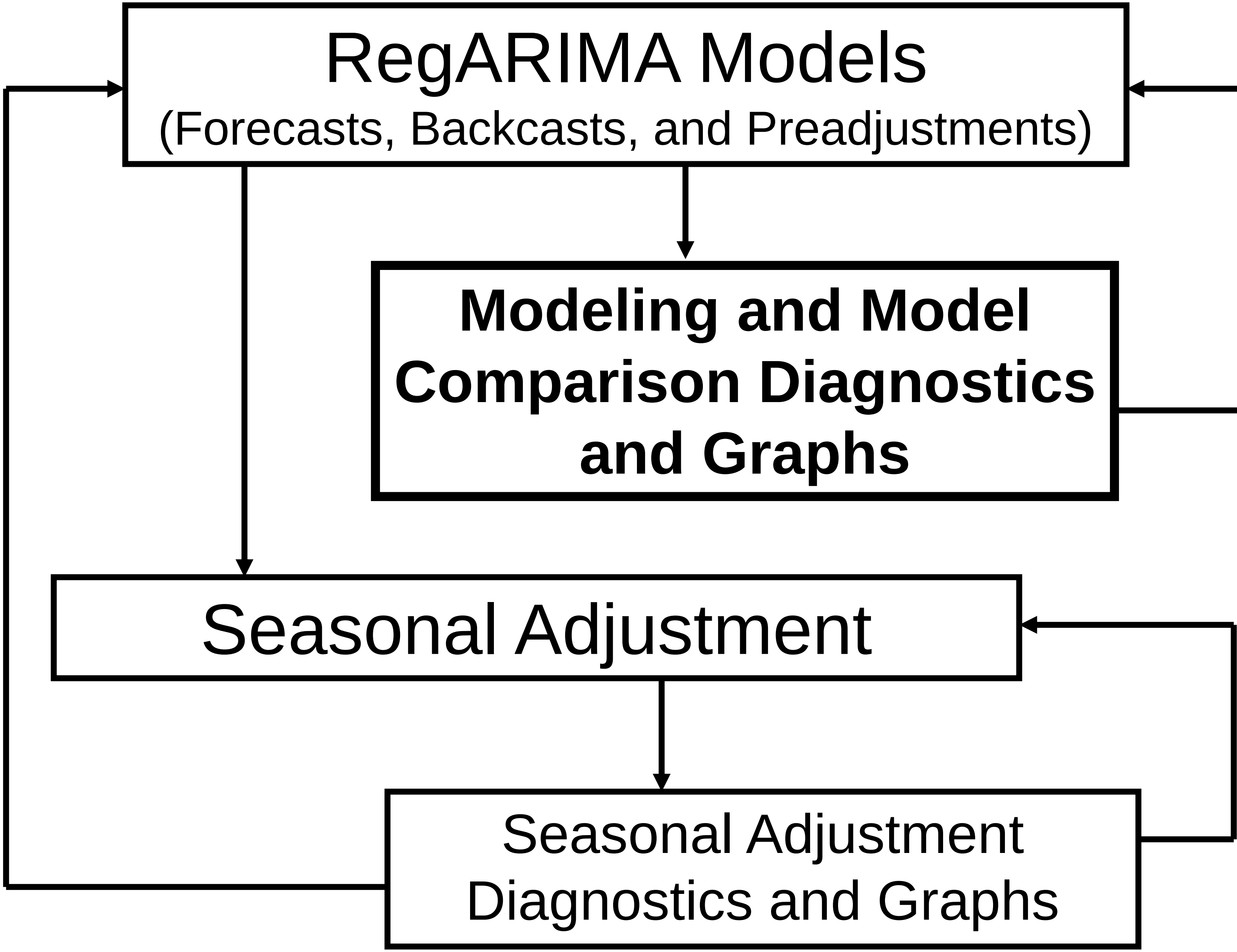
2019

Economic Statistical Methods Division

U.S. Census Bureau

Objectives

- **At the end of this unit, you should understand –**
  - How to interpret some basic regARIMA model diagnostics


RegARIMA Models
(Forecasts, Backcasts, and Preadjustments)
Modeling and Model Comparison Diagnostics and Graphs
Seasonal Adjustment
Seasonal Adjustment Diagnostics and Graphs

Modeling Diagnostics

- Standard errors, correlations, and t tests for regression and ARMA model coefficients
  - Except for outliers, |t| should be > 2 (or 1.96)
- Normality – Geary's a, kurtosis
- Sample autocorrelations and partial autocorrelations
- Ljung-Box Q
- Spectrum plot of regARIMA residuals
- QS (discussed in later class section)

Modeling Diagnostics (continued)

- Average forecast error over last three years

- AICC comparisons
  - (**`history`** spec)
- Out-of-sample forecast error plots
  - (**`history`** spec)

Other Modeling Diagnostics

- Histogram of standardized residuals
  - Do they look normal?
- Durbin-Watson Statistic for residuals
  - Tests for first order autocorrelation: generally values from 1.5-2.5 are desired
- Friedman Non-Parametric Test for seasonality of the residuals

- Available in the output, we will not cover them

`Estimate` Spec

- Default estimation happens whenever you have a regARIMA model even without the **`estimate`** spec
- Use it to gather needed diagnostics and information

`Estimate` Spec Example

```
estimate {
  print=(roots regcmatrix)
  savelog=all
#  savelog=(aicc aic bic hq afc)
}
```

`Check` Spec

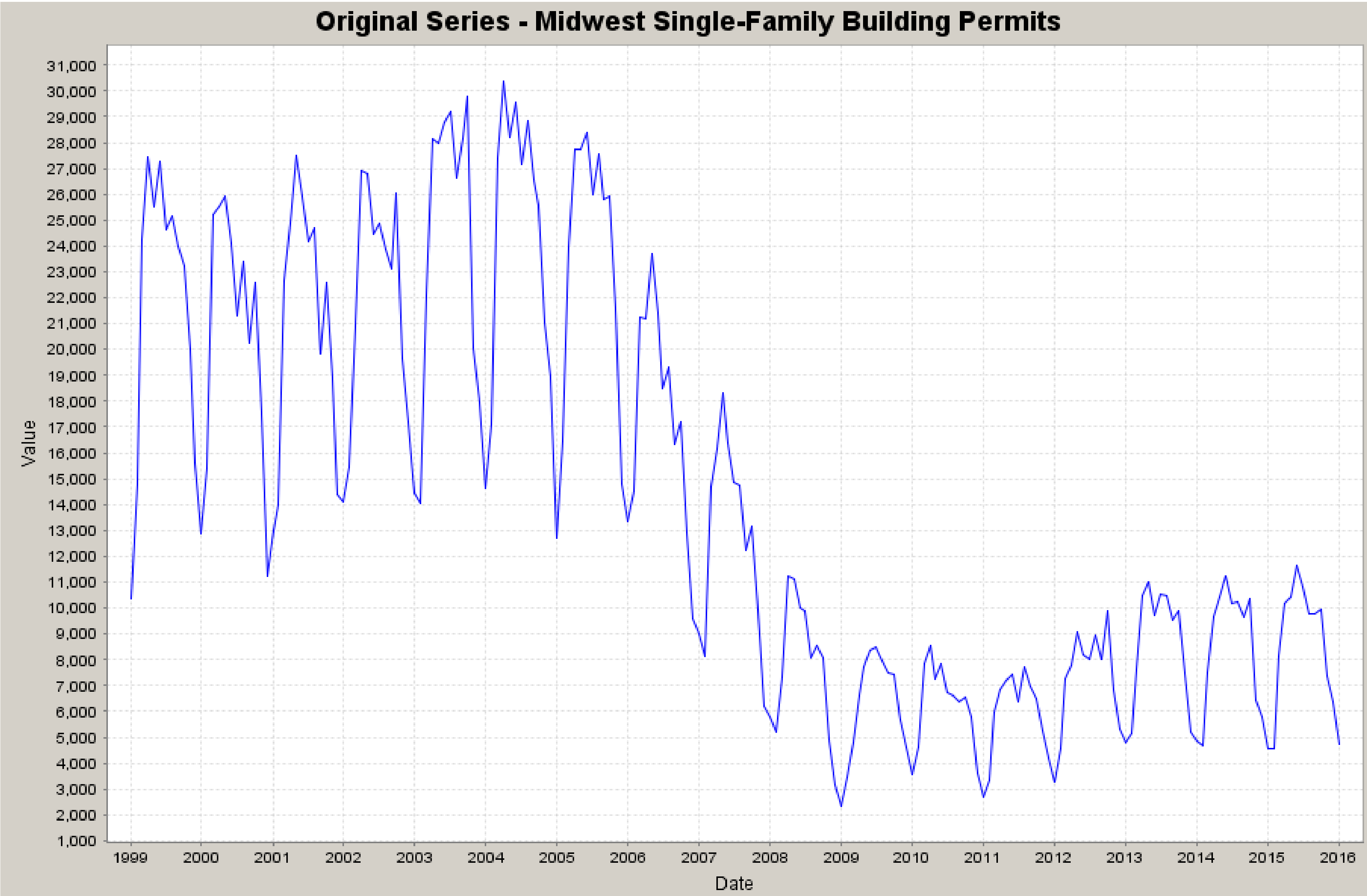
- Use the **`check`** spec to gather certain model diagnostics, control output

Check Spec Example

```
check {
  print=all
  savelog=all
#   savelog=(normalitytest
#      lbq bpq sft)
}
```

Model Information

- ARMA coefficients and standard errors
- Regression coefficients, standard errors, t-values and p-values
- Correlations between the parameters


Original Series - Midwest Single-Family Building Permits
Value
31,000
30,000
29,000
28,000
27,000
26,000
25,000
24,000
23,000
22,000
21,000
20,000
19,000
18,000
17,000
16,000
15,000
14,000
13,000
12,000
11,000
10,000
9,000
8,000
7,000
6,000
5,000
4,000
3,000
2,000
1,000
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
Date

Midwest Single-Family Building Permits

- Log transformation
- ARIMA model: (1 1 0)(0 1 1)
- Regressor: Easter[2]
- Automatically Identified Outlier: TC2008.Dec

- Run series with and without trading day regressors

Regression Model With One-Coefficient Trading Day Regression

Regression Model			
	Parameter Estimate	**Standard Error**	**t-value**
1-Coefficient Trading Day			
Weekday	0.0151	0.00118	12.77
Sat/Sun (derived)	-0.0378	0.00296	-12.77
Easter[2]	-0.0514	0.01873	-2.75
Automatically Identified Outliers			
TC2008.Dec	-0.3637	0.06537	-5.56

Regression Model With Full Six-Coefficient Trading Day Regression

Regression Model			
	Parameter Estimate	**Standard Error**	**t-value**
Trading Day			
Mon	0.0228	0.00939	2.43
Tue	0.0140	0.00919	1.52
Wed	0.0174	0.00911	1.91
Thu	0.0094	0.00935	1.01
Fri	0.0157	0.00925	1.69
Sat	-0.0341	0.00924	-3.69
Sun (derived)	-0.0451	0.00962	-4.69
Easter[2]	-0.0465	0.01938	-2.40
Automatically Identified Outliers			
TC2008.Dec	-0.3762	0.06582	-5.72

Regression Model With No Trading Day Regression

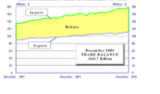
Regression Model			
	Parameter Estimate	**Standard Error**	**t-value**
Easter[2]	-0.0940	0.02731	-3.44
Automatically Identified Outliers			
TC2008.Dec	-0.3481	0.08444	-4.12

Significance Test for Variable Groups

- Most common groups
  - Trading day
  - Seasonal dummies
- F tests for regressor groups
  - Also chi-square tests, but research indicates the F tests are more appropriate

Significance Tests of the Regressor Groups

F Tests for Trading Day Regressors			
	df	**F-statistic**	***P*-Value**
Trading Day	6, 185	25.90	0.00



Normality of the residuals

- X-13ARIMA-SEATS has three tests for normality of the residuals:

1. Skewness
2. Geary's a – assumes series is not skewed
3. Sample kurtosis – only reliable for very long series, but significance may indicate need for additional outlier regressors

- If the regARIMA model fits the data well, a lack of normality of the residuals may not be a problem.
- However, these tests may indicate a need for additional or different regressors (trading day, holiday, outliers)

Normality of Midwest Building Permits, No TD

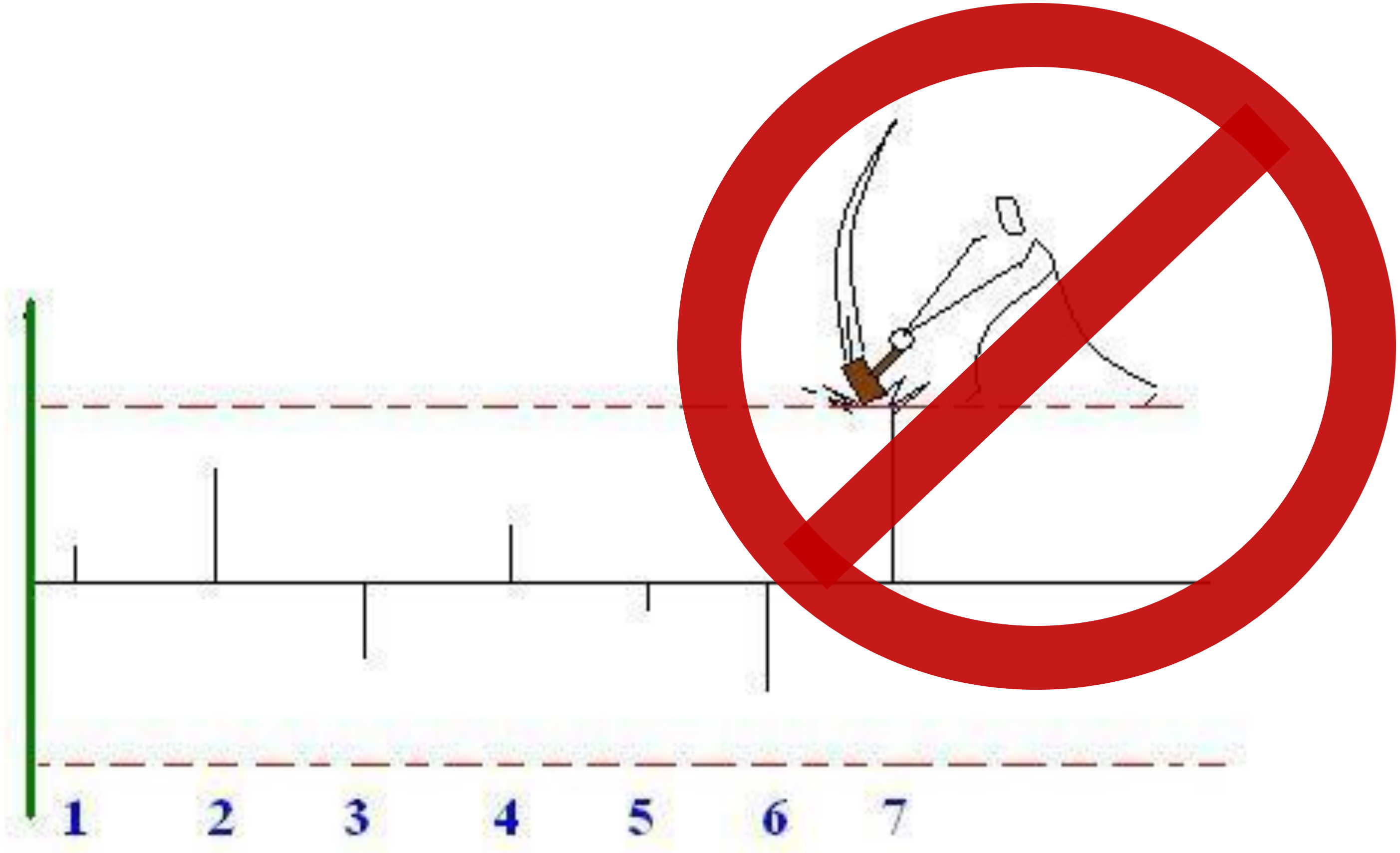
Normality Statistics for regARIMA Model Residuals

Number of residuals : 191
Skewness coeffi: -0.3917
Geary's a : 0.7977
Kurtosis : 2.8764

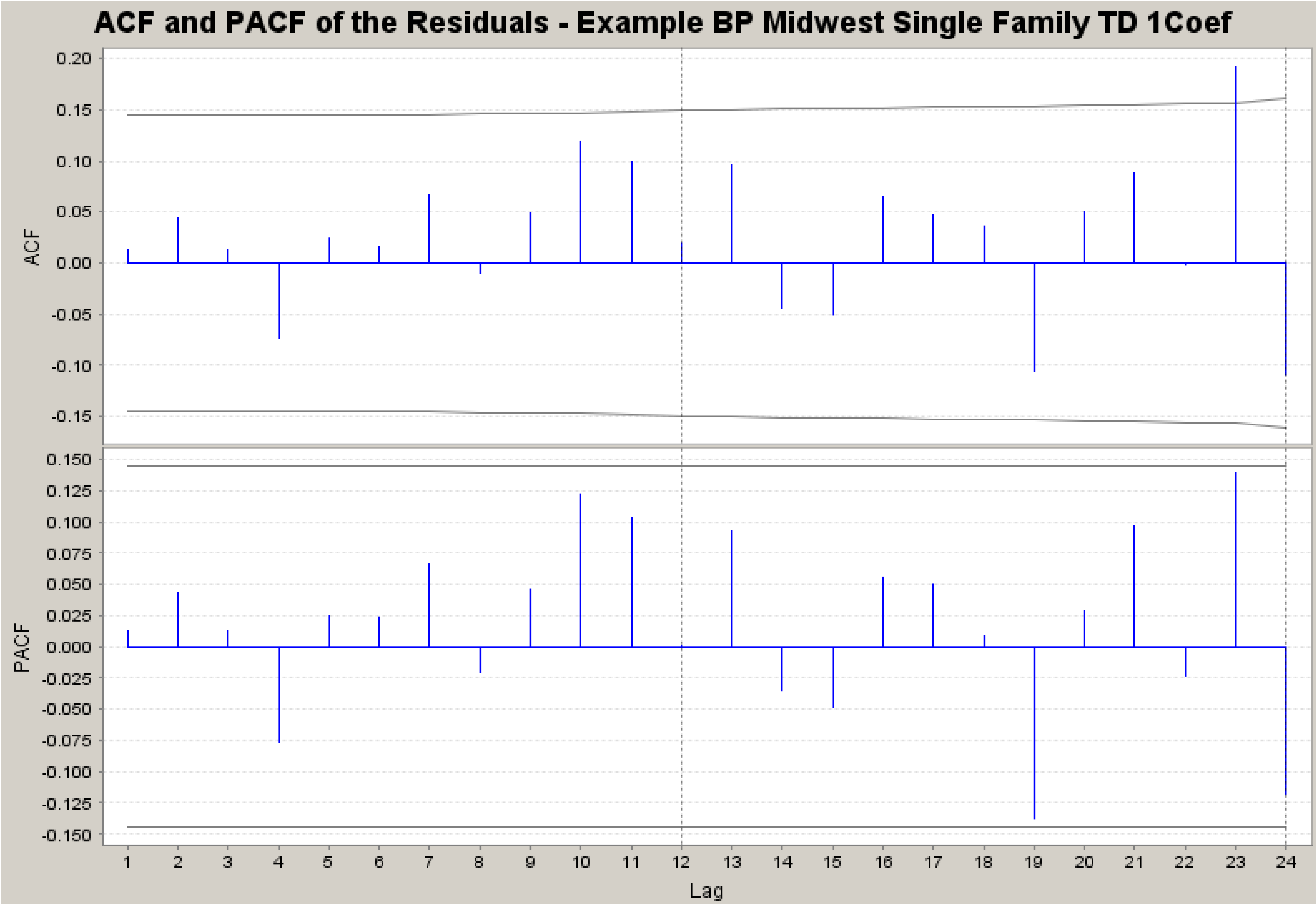
No indication of lack of normality.

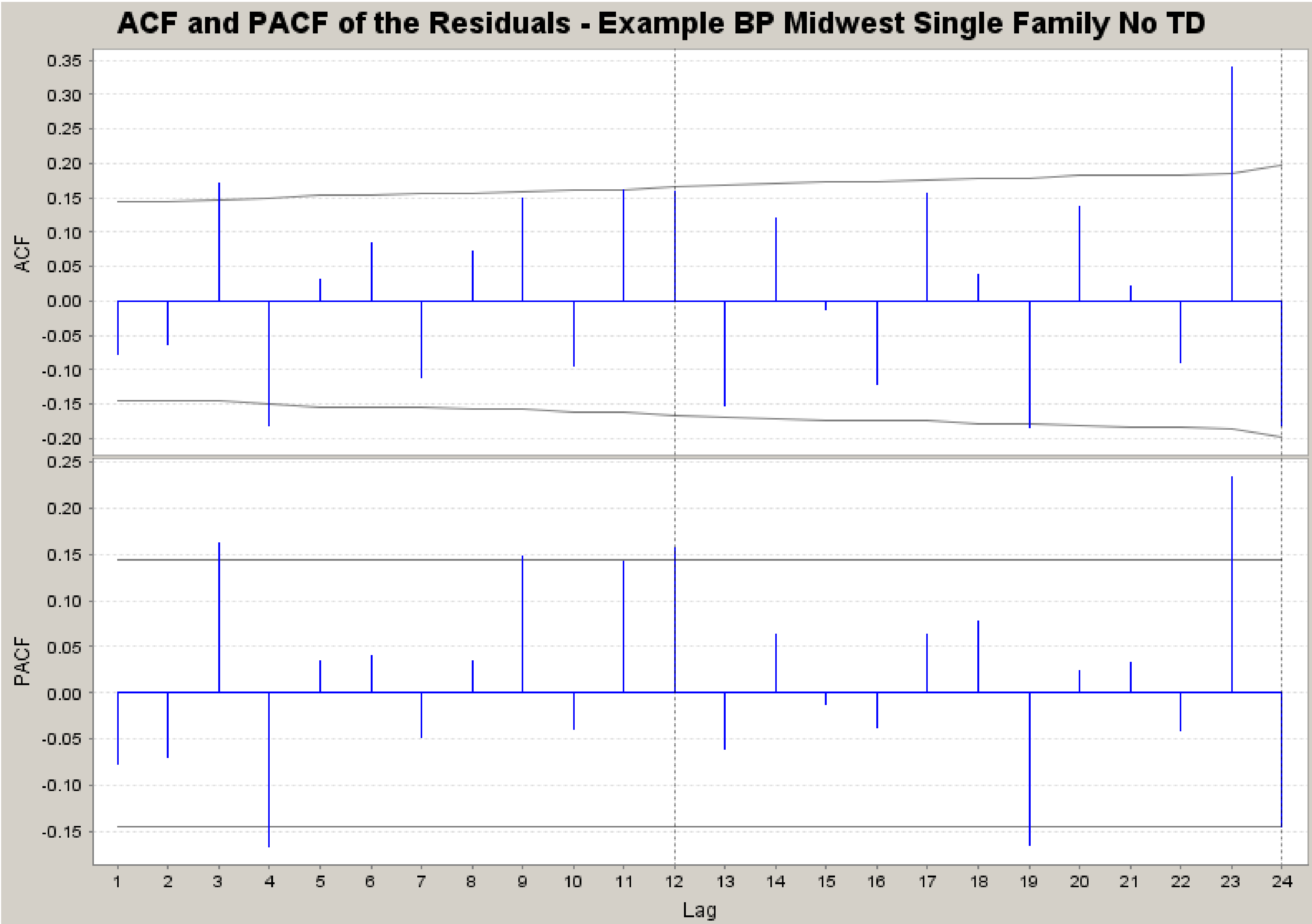
Autocorrelation Function and Partial Autocorrelation Function of Residuals

- Look for lags that spike beyond the bounds (2 standard errors)
- Consider ***important*** lags that may indicate model inadequacy
  - Low lags
  - Seasonal frequencies (multiples of 12 for monthly series, multiples of 4 for quarterly)



Kevin Tolliver (OSMREP), 2010


ACF and PACF of the Residuals - Example BP Midwest Single Family TD 1Coef
ACF
PACF
Lag
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
0.150
0.125
0.100
0.075
0.050
0.025
0.000
-0.025
-0.050
-0.075
-0.100
-0.125
-0.150
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24


ACF and PACF of the Residuals - Example BP Midwest Single Family No TD
ACF
PACF
Lag
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
-0.20
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24

Ljung-Box Q Diagnostic

- Goodness-of-fit diagnostic to indicate model inadequacy
- Function of the sample autocorrelation function

Greta Ljung and George Box (1978)

Ljung-Box Q Diagnostic (2)

- Based on the sample autocorrelation function but cumulative rather than specific to individual lags
  - One ACF peak at a low lag can result in a string of failures in LBQ
  - Some users check only the LBQ result at 24

Ljung-Box Q Formula

If r_k are the autocorrelations of the residuals and n is the series length, then the LBQ at lag K is

$$Q_K = n(n+2)\sum_{k=1}^{K}\frac{r_k^2}{n-k}$$

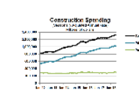
For large n, $Q_K \sim \chi^2_{K\text{-}p\text{-}q}$

Ljung-Box Qs in the Output File: Midwest Single-Family BP No TD

Sample Autocorrelation of Innovations, Part 1									
	Lag1	Lag 2	Lag 3	Lag 4	Lag 5	Lag 6	Lag 7	Lag 8	...
ACF	-0.08	-0.06	0.17	-0.18	0.03	0.08	-0.11	0.07	...
SE	0.07	0.07	0.07	0.07	0.08	0.08	0.08	0.08	...
LB Q	1.16	1.96	7.69	14.19	15.78	15.78	18.22	19.23	...
DF	0	0	1	2	3	4	5	6	...
P	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	...

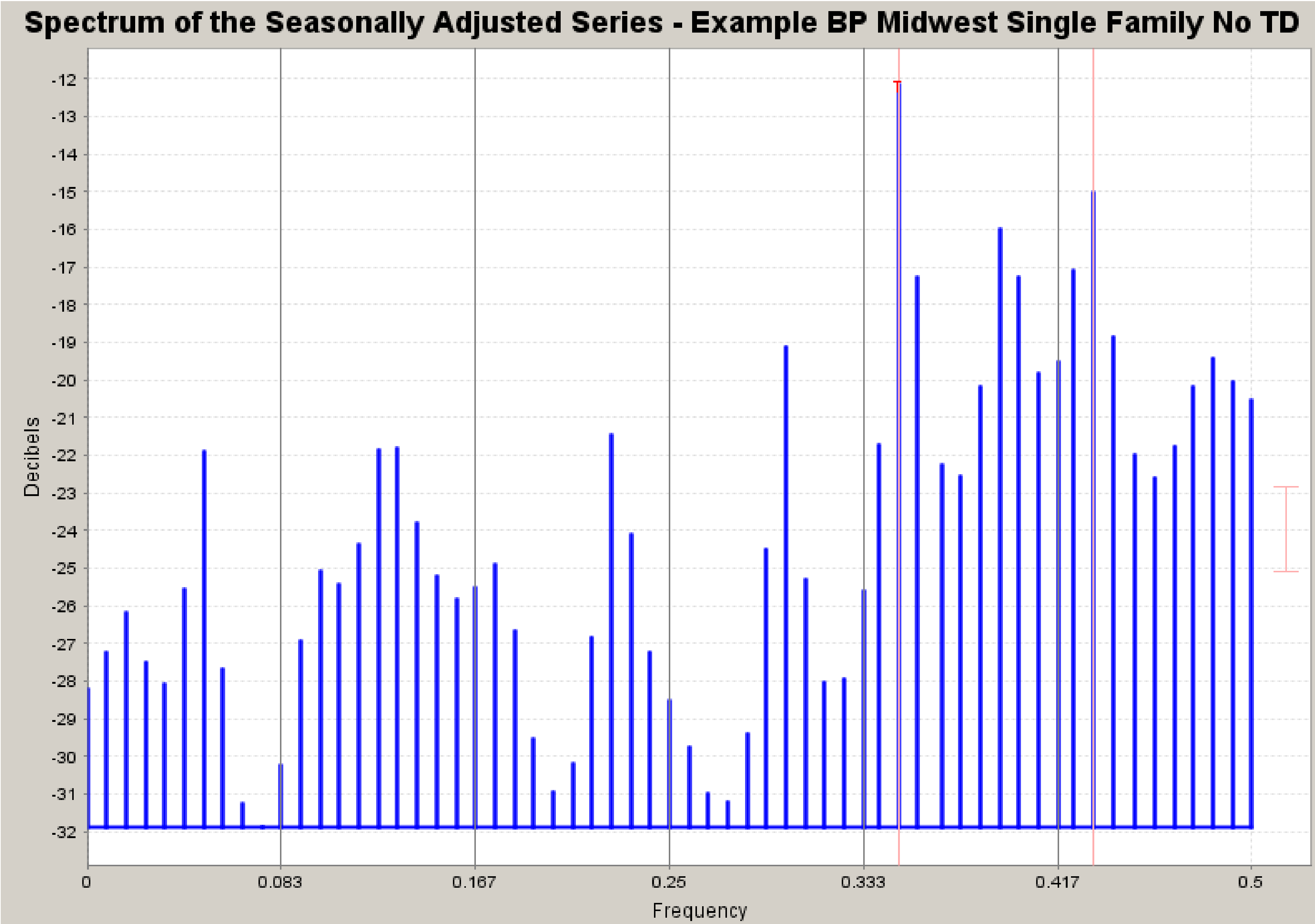
Ljung-Box Qs From Win X-13

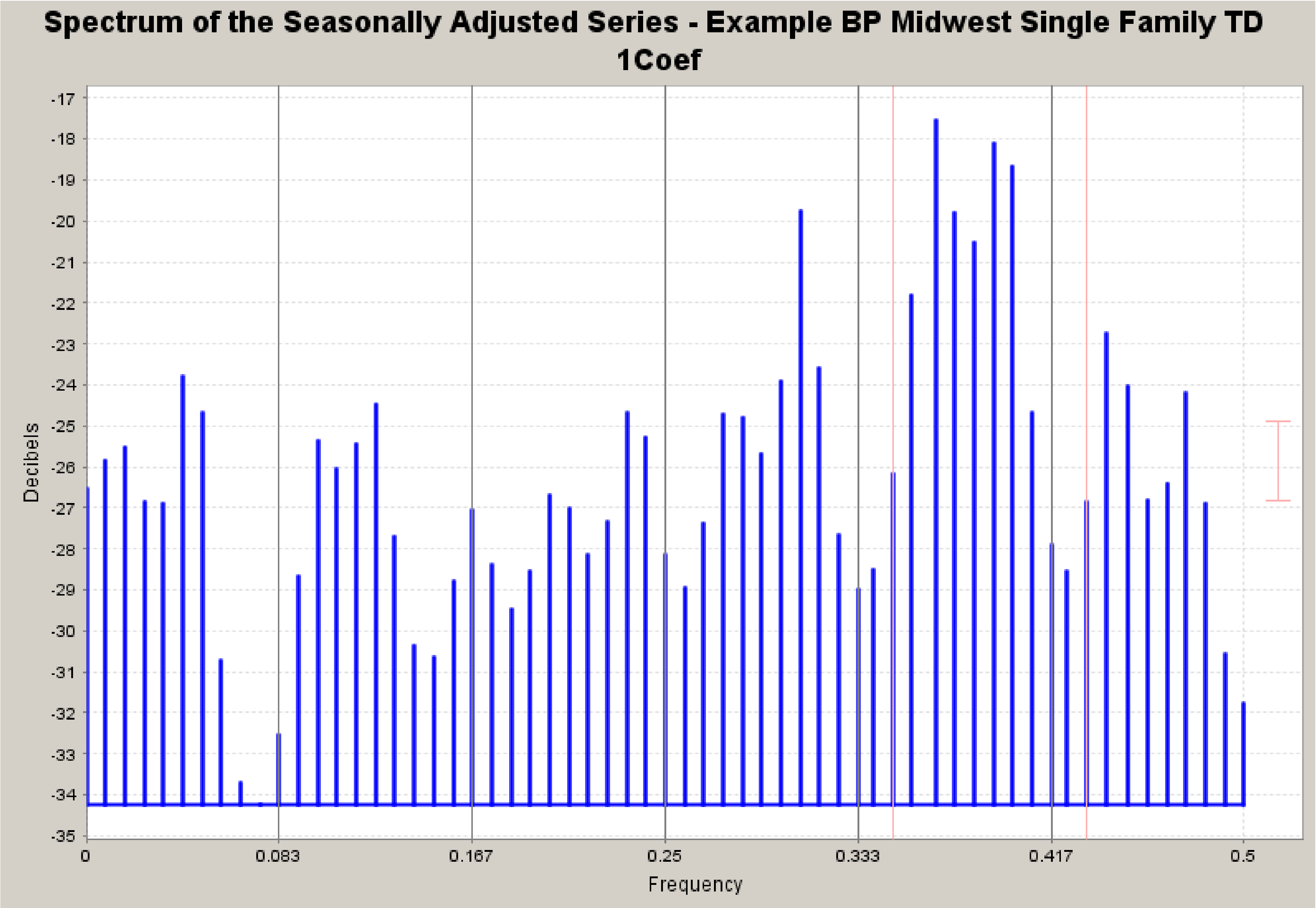
Series Name	# LBQ Fail	Sig LBQ	Sig Seas LBQ
Example BP Midwest Single Family No TD	**22**	3 – 24	*s12s24*
Example BP Midwest Single Family TD 1Coef	0		
Example BP Midwest Single Family TD (6)	0		



Spectral Diagnostic

- More information coming during discussion of seasonal adjustment diagnostics
- Monthly series only
- Look for large peaks at seasonal frequencies (1/12, 2/12, 3/12, 4/12, 5/12) and trading day frequencies (0.348, 0.432)


Spectrum of the Seasonally Adjusted Series - Example BP Midwest Single Family No TD
Decibels
-12
-13
-14
-15
-16
-17
-18
-19
-20
-21
-22
-23
-24
-25
-26
-27
-28
-29
-30
-31
-32
0
0.083
0.167
0.25
0.333
0.417
0.5
Frequency


Spectrum of the Seasonally Adjusted Series - Example BP Midwest Single Family TD
1Coef
Decibels
-17
-18
-19
-20
-21
-22
-23
-24
-25
-26
-27
-28
-29
-30
-31
-32
-33
-34
-35
0
0.083
0.167
0.25
0.333
0.417
0.5
Frequency

`Forecast` Spec

- Forecasting is so important, even without the **`forecast`** spec, the program produces a year of forecasts
  - Popular leads: (a) 2 years (b) ½ the seasonal filter length, for example 3x5=>42 monthly forecasts, 14 quarterly forecasts
- Also controls backcasts at beginning of series
  - No backcasts if model span is different from (shorter than) the adjustment span

Forecast Spec Example

```
forecast{
  maxlead = 8  # 2 years in quarters
# [ default: 12 months or 4 quarters ]
# [ range: 0 to 120 ]
  maxback = 8
# [ default: 0  max: 120 ]

  print = See Manual/Quick Reference
  save = See Manual/Quick Reference
}
```

Average Forecast Errors

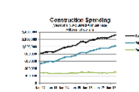
- Within sample average absolute percentage errors:
  - AAPE(Last year): Remove the last year of the series, calculate forecasts for the year, find average error
  - AAPE(Last-1 year): Remove additional year of the series, calculate forecasts for the year, find average error
  - AAPE(Last-2 years): Remove additional year of the series, calculate forecasts for the year, find average error

Average Forecast Errors, continued

- AAPE(Last 3 years): average of the other three
  - All four are in the output file; only this average is in the Win X-13 diagnostics tables
- These are within-sample errors because the default setting is not to recalculate the regARIMA parameters after removing series values
  - Full span parameters are used for forecasting
- To recalculate the parameters, use **`outofsample=yes`** in the **`estimate{}`** spec

Average Forecast Error From Win X-13

Series Name	aaFcE (3-yr)
Example BP Midwest Single Family No TD	**7.6236109489352**
Example BP Midwest Single Family TD 1Coef	6.0469965328332
Example BP Midwest Single Family TD (6)	6.0151315575393



AICC – Akaike's Information Criterion (Corrected for Sample Size)

- Alternate to usual AIC
- *Comparison* diagnostic for different models, prefer minimum
  - Not pass/fail, no “good” value
- Must have
  - Same outliers
  - Same order(s) of differencing
  - Same underlying series and model span

Want to Compare AICC for Series With Different Outliers?

- Force the same outliers then compare
  - Hard-code outliers and turn off automatic outlier identification

- Can be tricky – try…
  - Union of outlier sets
  - Multiple runs, each with the other model's outliers
  - Closer analysis to identify one good outlier set

AICC in the Output File: Midwest Single-Family BP, TD 1-Coefficient

Likelihood Statistics

Number of observations (nobs)	205
Effective number of observations (nefobs)	192
Number of parameters estimated (np)	6
Log likelihood	222.8849
Transformation Adjustment	-1797.589
Adjusted Log likelihood (L)	-1574.7041
AIC	3161.4082
AICC (F-corrected-AIC)	3161.8622
Hannan Quinn	3169.324
BIC	3180.9531

AICC Comparison

Midwest Single-Family BP, No TD
AICC: 3269.82056

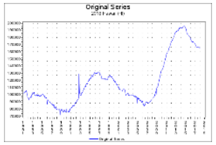
Midwest Single-Family BP, TD (6)
AICC: 3171.38135

Midwest Single-Family BP, TD 1-Coefficient
AICC: 3161.8623

- If models meet all criteria, prefer the lesser value (TD1 model)
- As of now, no clear significance levels for likelihood-statistic comparisons, but some work indicates that the AICC difference should be at least 2

Interventions, Note

- In cases of interventions comparing AICC values may be OK
  - Outliers are identified generally, not usually as known events
  - Interventions are for specific known events



Out-of-Sample Forecast Error Plot

- Prefer the model with generally smaller forecast errors
- This graph compares the forecast performance of two models over time, to see if one has consistently smaller forecast errors

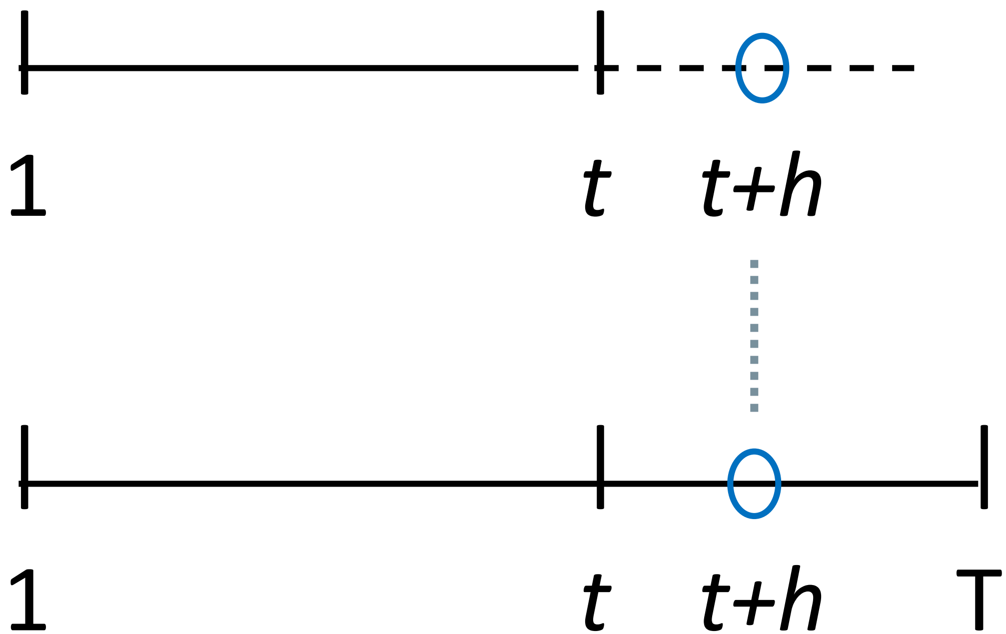
History Spec Forecasts

```
history {
   estimates = ( fcst )
   start = yyyy.mm
}
```

Forecast Performance

- Truncate series at t_0
- Fit the regARIMA model and calculate *h*-step-ahead forecast
  - One step ahead and one year ahead
- Add data point (t_0 +1)
- Repeat model fit and forecast

Out-of-Sample Forecast Error, Illustration



$\hat{Y}_{t+h|t}$ = forecast at time *t+h* given values up to time *t*

Y_{t+h} = actual value at time *t+h* (given T)

Error: $Y_{t+h} - \hat{Y}_{t+h|t}$, for $t_0 \leq t \leq T - h$

Notation for h-step Forecasting

$\hat{Y}^{(i)}_{t+h\,|\,t}$ = Model *i*'s *h*-step ahead forecast of Y_{t+h} at time *t* for

i = 1, 2

(re-estimate model parameters for each time *t*)

h: Usually compare 1-step ahead and 1-year ahead forecasts
(1 year = *s* steps, *s* is 12 for monthly, 4 for quarterly series)

Graphical Model Selection

For given h and t_0 , plot

$$SSE_{h,N}^{(1,2)} = \sum_{t=t_0}^{N-h} \left[\left(Y_{t+h} - \hat{Y}_{t+h|t}^{(1)} \right)^2 - \left(Y_{t+h} - \hat{Y}_{t+h|t}^{(2)} \right)^2 \right]$$

versus N for $(t_0+h) \leq N \leq \mathrm{T}$

N is the end date of the series (progresses over time)

T is the final end date of the full series

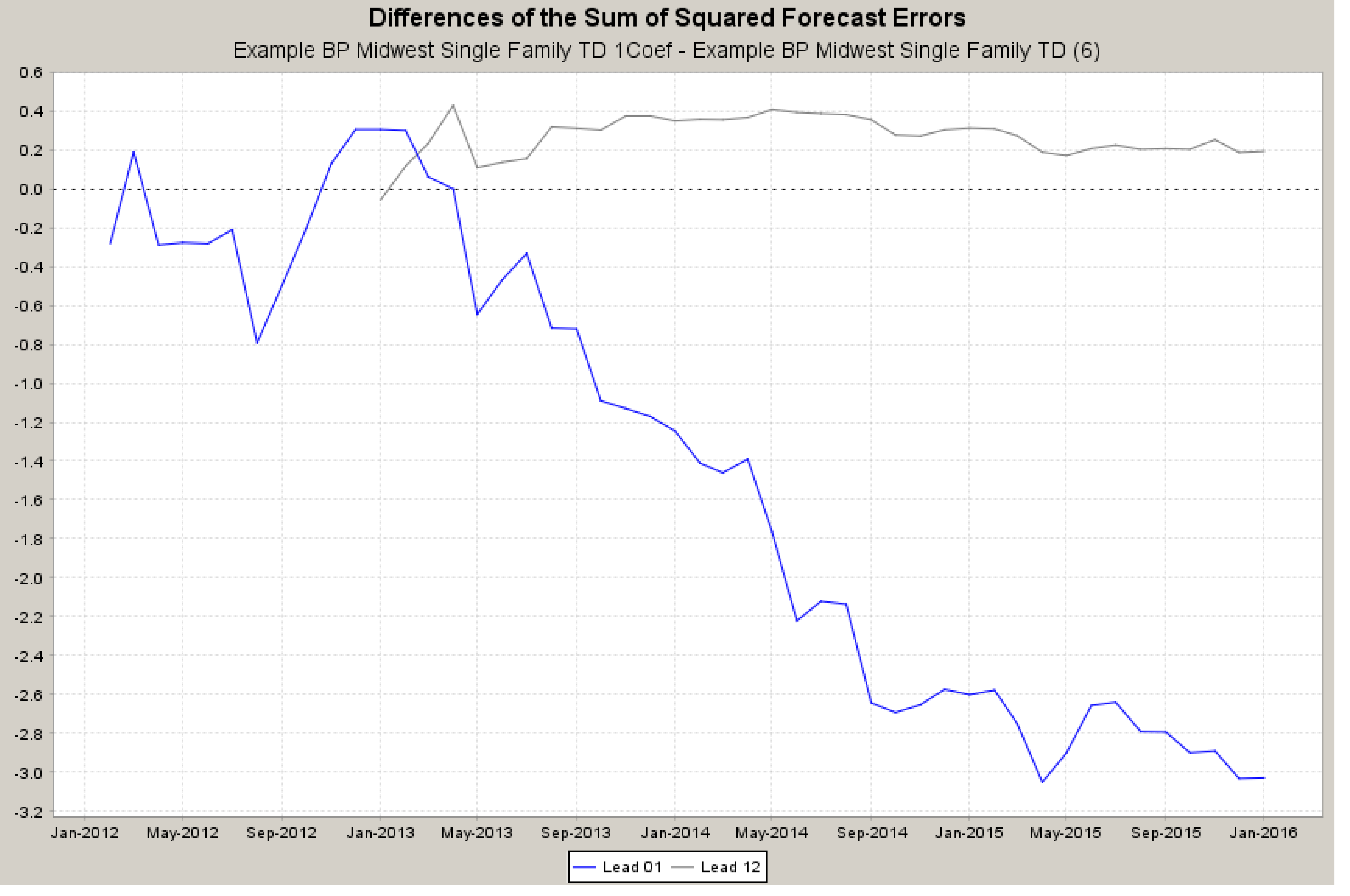
Slope Is the Key

- Persistently ***decreasing*** value (negative slope) with increasing N ➔ prefer Model 1
  - Persistently better forecasts from Model 1 (consistently summing negative values, indicating Model 1 has smaller errors)
  - Stair-stepping down, prefer Model 1
- Persistently ***increasing*** value (positive slope) with increasing N ➔ prefer Model 2
  - Stair-stepping up, prefer Model 2

Two Lines on the Graph:
1-Step Ahead, 1-Year Ahead

- Because usually the series and model are seasonal, the seasonal (year-ahead) forecast is especially important

- In this example
  - Model 1: Midwest Single-Family BP, TD 1-Coef
  - Model 2: Midwest Single-Family BP, TD (6-Coef)



Differences of the Sum of Squared Forecast Errors
Example BP Midwest Single Family TD 1Coef - Example BP Midwest Single Family TD (6)
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
-1.2
-1.4
-1.6
-1.8
-2.0
-2.2
-2.4
-2.6
-2.8
-3.0
-3.2
Jan-2012
May-2012
Sep-2012
Jan-2013
May-2013
Sep-2013
Jan-2014
May-2014
Sep-2014
Jan-2015
May-2015
Sep-2015
Jan-2016
Lead 01
Lead 12

Interpret the Graph

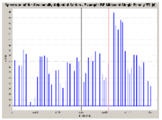
- Seasonal forecasts are inconclusive, neither model seems to forecast better one year ahead
- One-month ahead forecasts are better from the model with the one-coefficient trading day regressor
  - Negative slope

Each Line Is a Separate Comparison of Two Models

- If one line is inconclusive, the other line can indicate model preference
- Weight seasonal forecasts more
- Weight most-recent result more

Uses for Forecast Error History Diagnostic

- Compare any two regARIMA models
- Especially useful when AICC is not appropriate
  - Different spans of data
  - Different outliers
  - Different orders of differencing



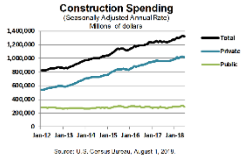
Other Forecasting Use

- Forecast next month's (quarter's) value before it is reported
- Compare the forecasted value to the actual value when it is available
- Big differences may indicate problems with the incoming value, need for editing or treating as an outlier

Programs Using Forecast Comparisons

- Census Bureau
  - International Trade forecasted year 2000 data
    - Identifying reporting problems among respondents
  - Service Sector checks forecasts for heads up on outlier results

- TERROR (TRAMO for Errors) by Gianluca Caporello and Agustín Maravall (Bank of Spain)
  - Automatic modeling and forecast checking
  - Capable of handling large number of series
  - Can point to editing needs

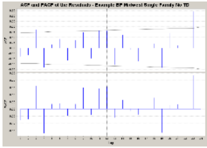


Very *Very* Generally – When Diagnostics Disagree

- AICC/AIC
  - (need to see clear separation in value)
- Out-of-Sample Forecast Error Graphs
  - (need to see clear preference)

outweigh

- ACF/PACF
- Spectrum of the Model Residuals



Diagnostics with Win X-13

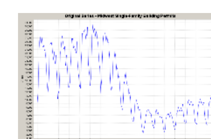
	Model Span	ARIMA Model	Trading Day	Holiday
Midwest Single Family TD 1Coef	1999.01 to 2016.01	(1 1 0)(0 1 1)	TD1[t12.8]	E2[t-2.75]
Midwest Single Family TD (6)	1999.01 to 2016.01	(1 1 0)(0 1 1)	TD[f-p0.00]	E2[t-2.40]
Midwest Single Family No TD	1999.01 to 2016.01	(1 1 0)(0 1 1)		E2[t-3.44]

	AICC	aa FcE (3-yr)	Normal?	# LBQ Fail	Sig LBQ	Sig Seas LBQ
TD 1Coef	3161.8623	6.0469965328332	ok	0		
TD (6)	3171.38135	6.01513155753926	ok	0		
No TD	3269.82056	7.62361094893522	ok	22	3--24	*s12s24*

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Win X-13 Diagnostics

- **Help -> Diagnostics** List to find a description of all the columns
- **File -> Failure/Warning Thresholds** to set the limits for what should be considered a failure (row is pink, failed diagnostic is bolded) or a warning (row is yellow, warned diagnostic is italic)



Purpose of RegARIMA Model Diagnostics

- Help users choose best model for their purposes
  - Many diagnostics are comparative rather than pass/fail
- Indicate areas of model inadequacy
  - User may need to start over, shorten the series, investigate unusual series behavior, etc.