ØAMET2200
Business Decision Making Using Data
Lecture 10

Instructor: Fenella Carpena

November 8, 2019

Announcements

- Problem Set 4 has been posted. Due Nov. 21 at 11:59 PM.
- Lecture and Office Hours (OH) next week are cancelled
  - No lecture on Nov. 15
  - No OH on Nov. 11
- Last meeting for this course is Nov. 22
  - Course Wrap Up
  - Practice Mini-Final Exam
- Extra OH schedule
  - Nov. 18 (Monday) 4-5PM
  - Nov. 25 (Monday) 4-5PM
  - Nov. 29 (Friday) 10-11AM
  - Dec. 2 (Monday) 4-5PM
  - Dec. 3 (Tuesday) 4-5PM

Part 4 of this Course: Building regression models

- You want to build a model to predict outcomes from data you have
- What's the **best model** you can build?
- Today: Modeling time series for forecasting into the future

Agenda for Today

- Smoothing Time Series Data
- Regression Models for Time Series
  - Polynomial Trends
  - Autoregressions
- Checking Regression Conditions
- Chapter 27

Time Series

- A time series is a sequence of data points ordered in time
  - Example: daily price of Apple stock, annual inflation rate
- A time series consists of three components:

$$y_t = \underbrace{\text{Trend}_t + \text{Seasonal}_t}_{\text{pattern}} + \underbrace{\text{Irregular}_t}_{\text{random}}$$

- Trend: smooth, slowly meandering pattern
- Seasonal: cyclical oscillations related to seasons
  - Example: Sales of skis are lower after winter
- When we forecast a time series, we are extrapolating the trend or seasonal component
- The irregular variation affects how accurate our forecasts are

Time Series

- Sometimes the time series data we have are "smoothed out"
  - **Smoothing** means removing irregular and seasonal components to enhance the visibility of a trend
  - In other words, removing short-term fluctuations and highlighting longer-term trends
- We'll briefly discuss three ways of smoothing
  1. Simple Moving Average
  2. Exponentially-Weighted Moving Average
  3. Seasonal Adjustment
- Often seen in finance (stock trading) and macroeconomic applications

1. Simple Moving Average

- The 5-period Simple Moving Average (SMA) is the unweighted average of the 5 most recent data points prior to and including the current period

$$\text{SMA}_{t,5} = \frac{Y_t + Y_{t-1} + Y_{t-2} + Y_{t-3} + Y_{t-4}}{5}$$

- How does it work in practice? Example:

	date	price
1	02jan2001	128
2	03jan2001	134
3	04jan2001	133
4	05jan2001	129
5	08jan2001	129
6	09jan2001	130
7	10jan2001	131

1. Simple Moving Average

- Note that we can also calculate the SMA for any number of periods (other than 5)
- The more terms that are averaged, the smoother the estimate of the trend

2. Exponentially-Weighted Moving Average (EWMA)

- EWMA is a weighted average of past observations with geometrically declining weights
  - More recent values of Y receive more weight than those further in the past
- Can be written $S_t = (1 - w)Y_t + wS_{t-1}$, where w is a weight
  - The current smoothed value is the weighted average of the current data point and the prior smoothed value
- The choice of w affects the level of smoothing
  - $\uparrow w \implies$ smoother S_t
  - $\uparrow w \implies S_t$ trailing more behind the observations
- How to pick w? Depends on the application
  - In finance, it is common to pick a number of days N and set $w = (N - 1)/(N + 1)$

2. Exponentially-Weighted Moving Average (EWMA)

- Larger N (higher w): time series is smoother
- Larger N (higher w): trailing more behind the observations

Comparison of SMA and EWMA

- EWMA reacts quicker to price changes than SMA
- EWMA is more "sensitive" than SMA

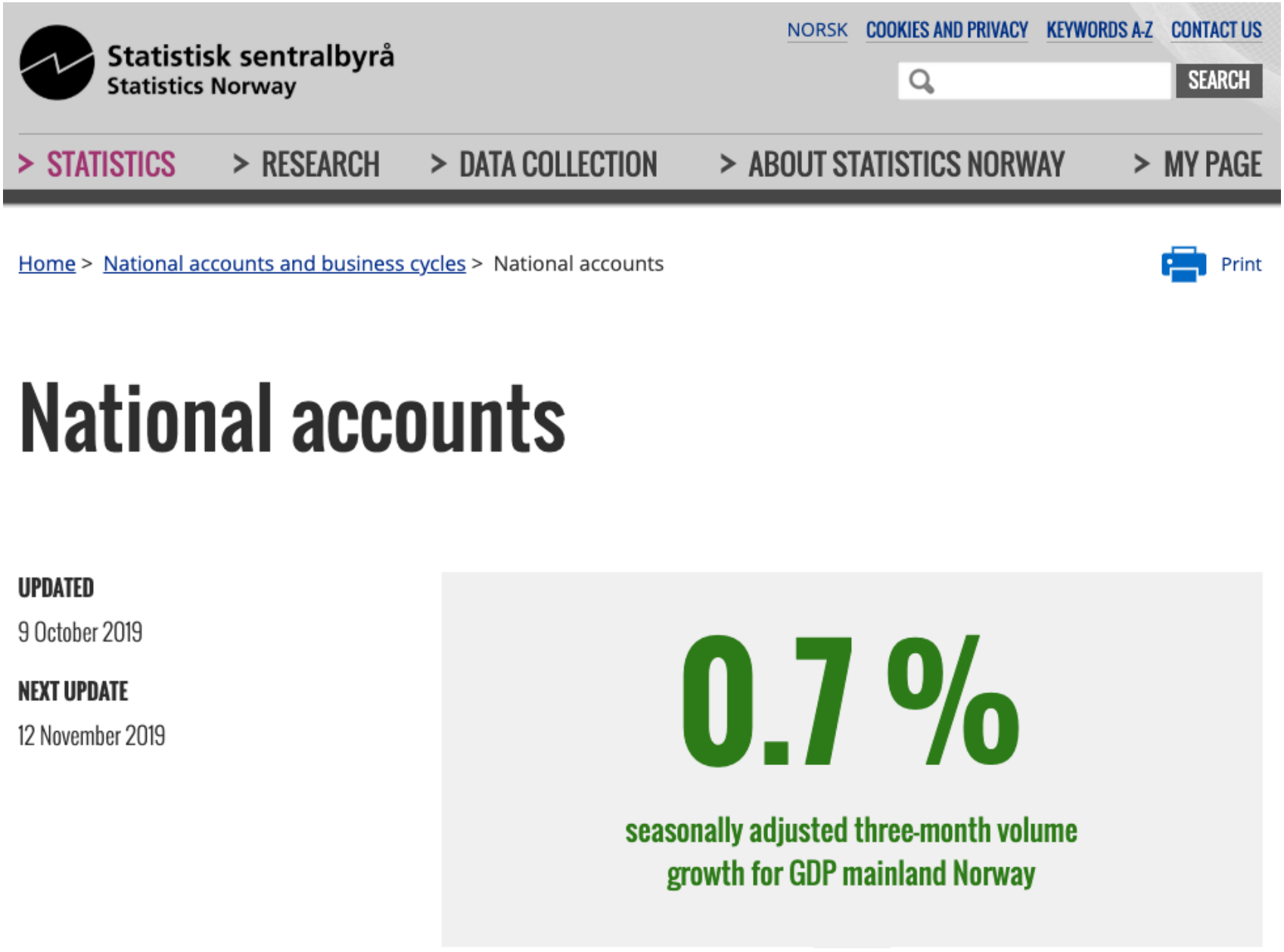
3. Seasonal Adjustment

- Seasonal adjustment means removing the seasonal component of a time series
- The difference between the observed time series and the seasonally adjusted time series is the seasonal component
- Many government-reported time series are seasonally adjusted
- For example:
  - *bruttonasjonalprodukt* or Gross Domestic Product (GDP)
  - See `https://www.ssb.no/en/nasjonalregnskap-og-konjunkturer/statistikker/knr`

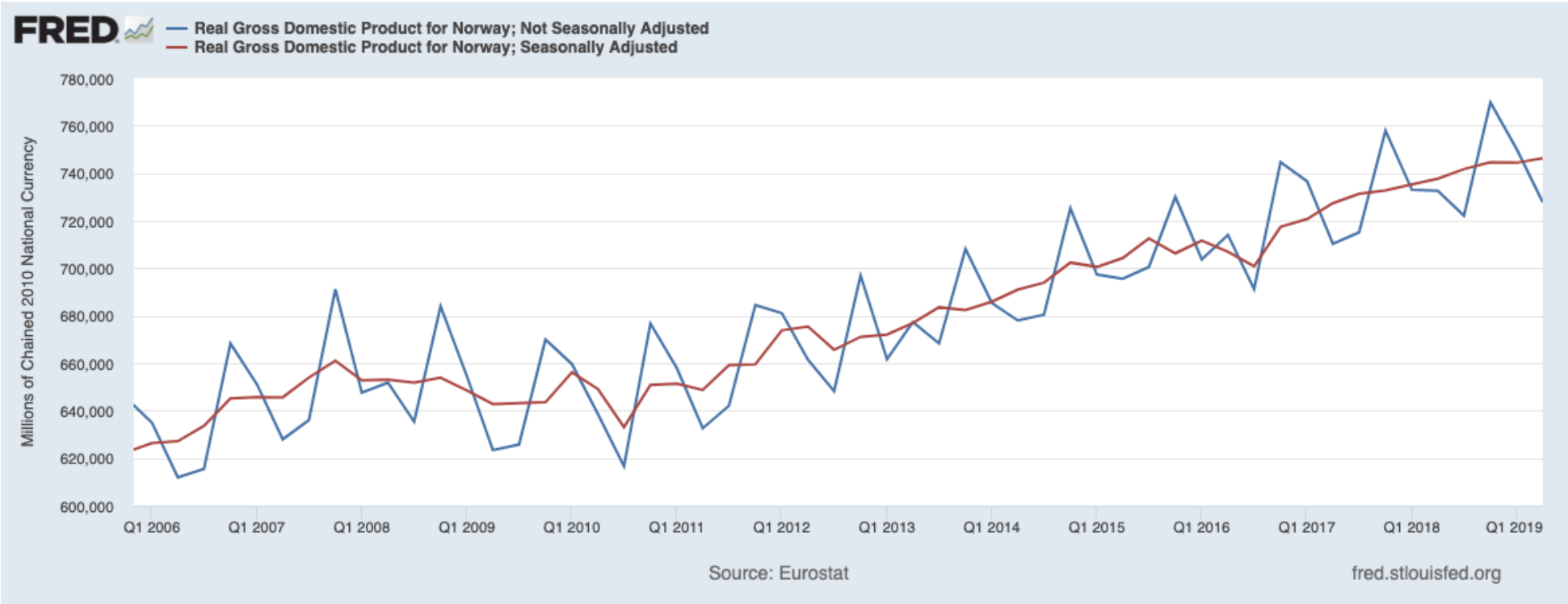
Why seasonally adjust these statistics?

Because of climatic conditions, public holidays and holidays in July and December, the intensity of the production varies throughout the year. The same applies to household consumption and other parts of the economy.

3. Seasonal Adjustment



3. Seasonal Adjustment



Case Study for Today: Forecasting Unemployment Rate

- The **unemployment rate** (the percentage of the labor force that is unemployed) is one of the most watched macroeconomic variables
- Many government agencies track the unemployment rate
  - If the unemployment rate is high, there will be greater demand for government services (e.g., at NAV)
- Businesses also watch the unemployment rate
  - If the unemployment rate is decreasing, managers worry that there will be pressure to raise wages, cutting into profits
- Suppose that you are an economist at Norges Bank, and your job is to develop statistical models for the unemployment rate

Case Study for Today: Forecasting Unemployment Rate

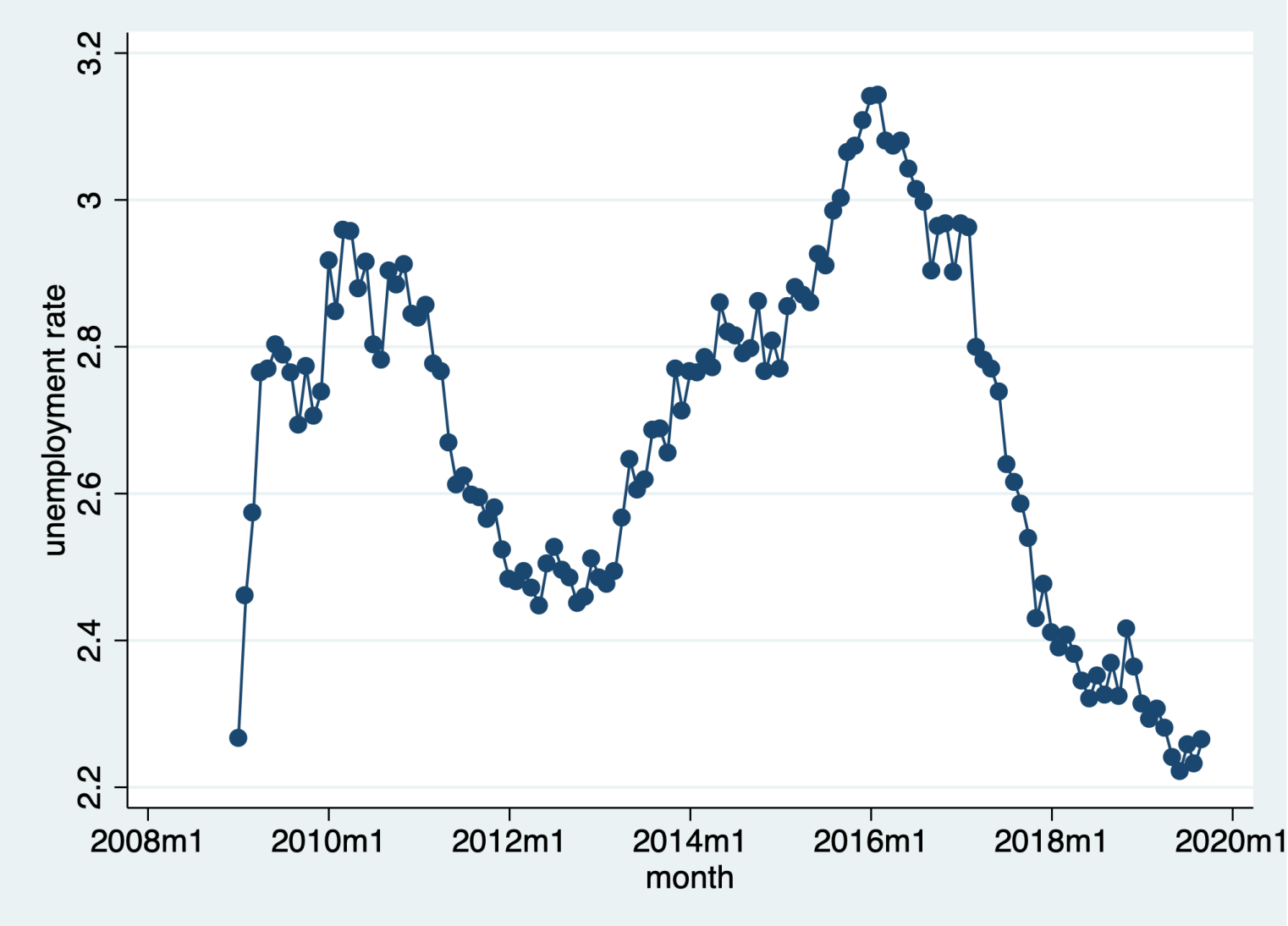
- You have monthly data on the seasonally-adjusted unemployment rate from 2009m1-2019m9 (time series)
  - We have $n = 129$ months in our data
- The data are compiled by the OECD and was downloaded from `fred.stlouisfed.org/series/LMUNRRTTNOM156S`
- You want to build a model to make a forecast and prediction interval for the next month's unemployment rate
  - To get a forecast, we need a regression model for time series
  - To get a valid prediction interval, we need to check the regression conditions

Case Study for Today: Forecasting Unemployment Rate

	month	t	unemprate
1	2009m1	1	2.266534
2	2009m2	2	2.460776
3	2009m3	3	2.573925
4	2009m4	4	2.764229
5	2009m5	5	2.769858
6	2009m6	6	2.802247
7	2009m7	7	2.78899

Case Study for Today: Forecasting Unemployment Rate

When analyzing a time series, always begin with a time plot!



Regression Models for Time Series

- We can use regression models to generate forecasts for time series data
- We'll discuss two types of explanatory variables that we can put in our regression model for forecasting
  1. The time index t: **polynomial trend**
  2. Prior values of the response variable y: **autoregression**
- Regression with time series data require the same assumptions as in the MRM
  - The most important one is usually the independence of errors

Polynomial Trends

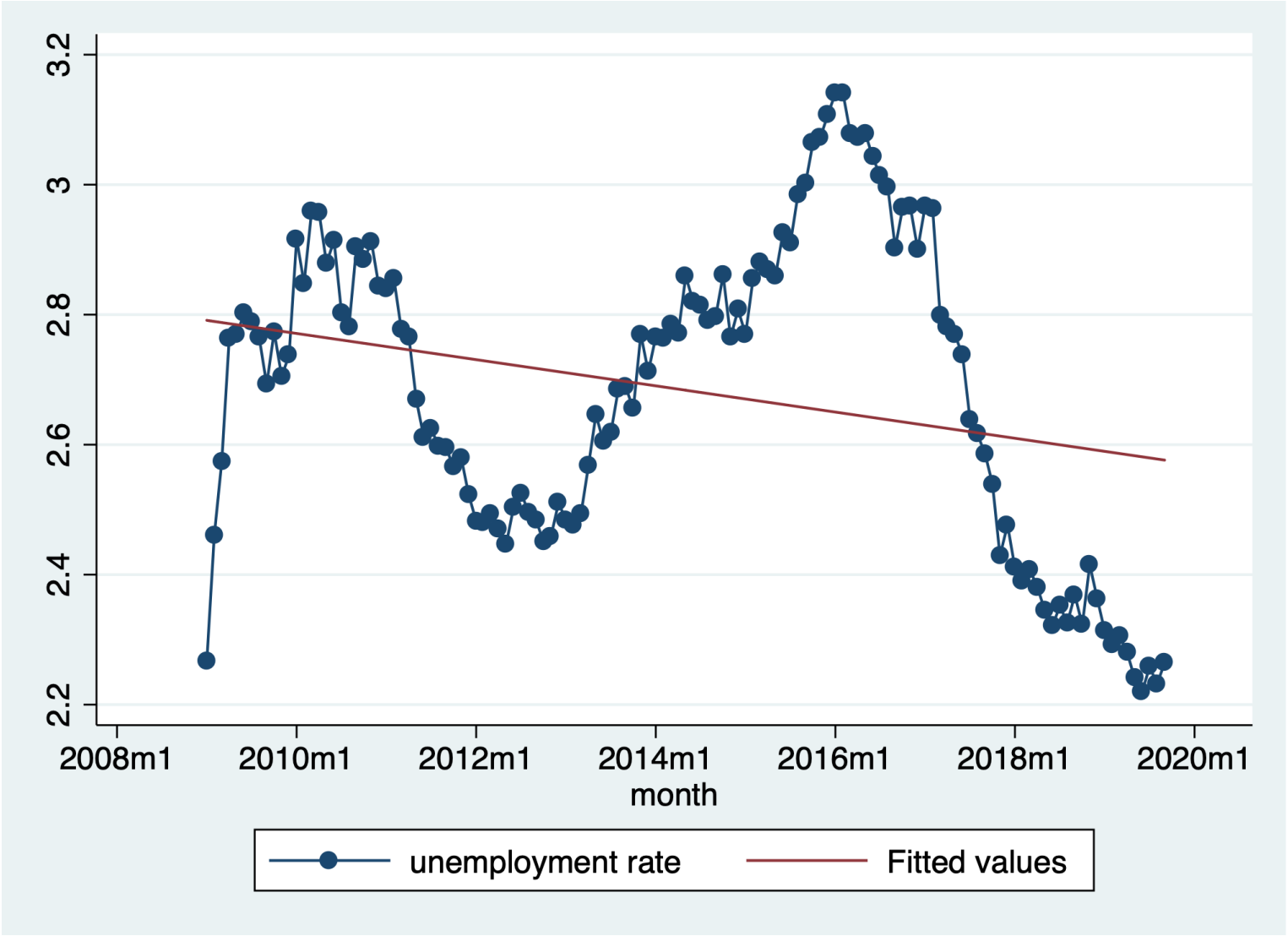
- A regression model for a time series that uses **powers** of t as explanatory variables
- Example: a third-degree or cubic polynomial

$$unemprate_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \epsilon_t$$

- Note that we can also have powers other than three (e.g., sixth-degree polynomial)
- Why do we care? Polynomials helps to better capture curvature in the data

Case Study: Linear Fit

$$unemprate_t = \beta_0 + \beta_1 t + \epsilon$$



Case Study: Polynomial Trends

- The simple regression of *unemprate* on t (without polynomial terms) clearly does not give us a good fit

- We can try a fifth-degree polynomial

$$unemprate_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 t^4 + \beta_5 t^5 + \epsilon_t$$

Case Study: Polynomial Trends

```
. forvalues i = 2/5 { ;
  2.         gen t`i' = t^`i' ;
  3. } ;

. regress unemprate t t2 t3 t4 t5 ;
```

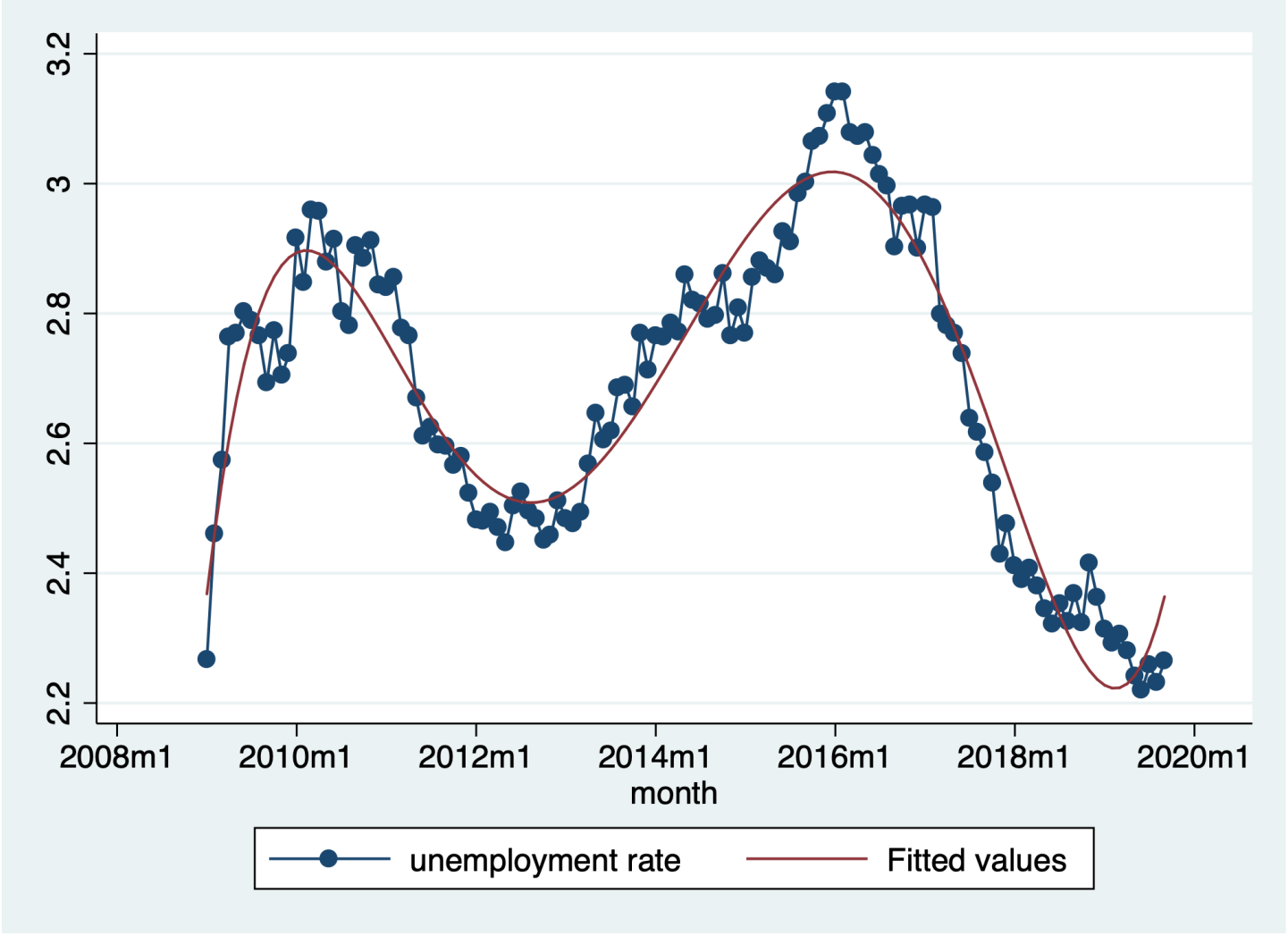
Source	SS	df	MS
Model	6.62389355	5	1.32477871
Residual	.656163388	123	.005334662
Total	7.28005693	128	.056875445

Number of obs =	129
F(5, 123) =	248.33
Prob > F =	0.0000
R-squared =	0.9099
Adj R-squared =	0.9062
Root MSE =	.07304

unemprate	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
t	.1084633	.0063412	17.10	0.000	.0959114	.1210153
t2	-.006105	.0002994	-20.39	0.000	-.0066976	-.0055125
t3	.0001275	5.81e-06	21.93	0.000	.000116	.000139
t4	-1.10e-06	4.92e-08	-22.32	0.000	-1.20e-06	-1.00e-06
t5	3.31e-09	1.51e-10	21.96	0.000	3.01e-09	3.61e-09
_cons	2.265558	.041356	54.78	0.000	2.183696	2.347419

Case Study: Polynomial Trends

$$unemprate_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 t^4 + \beta_5 t^5 + \epsilon_t$$



Case Study: How to use polynomial trends?

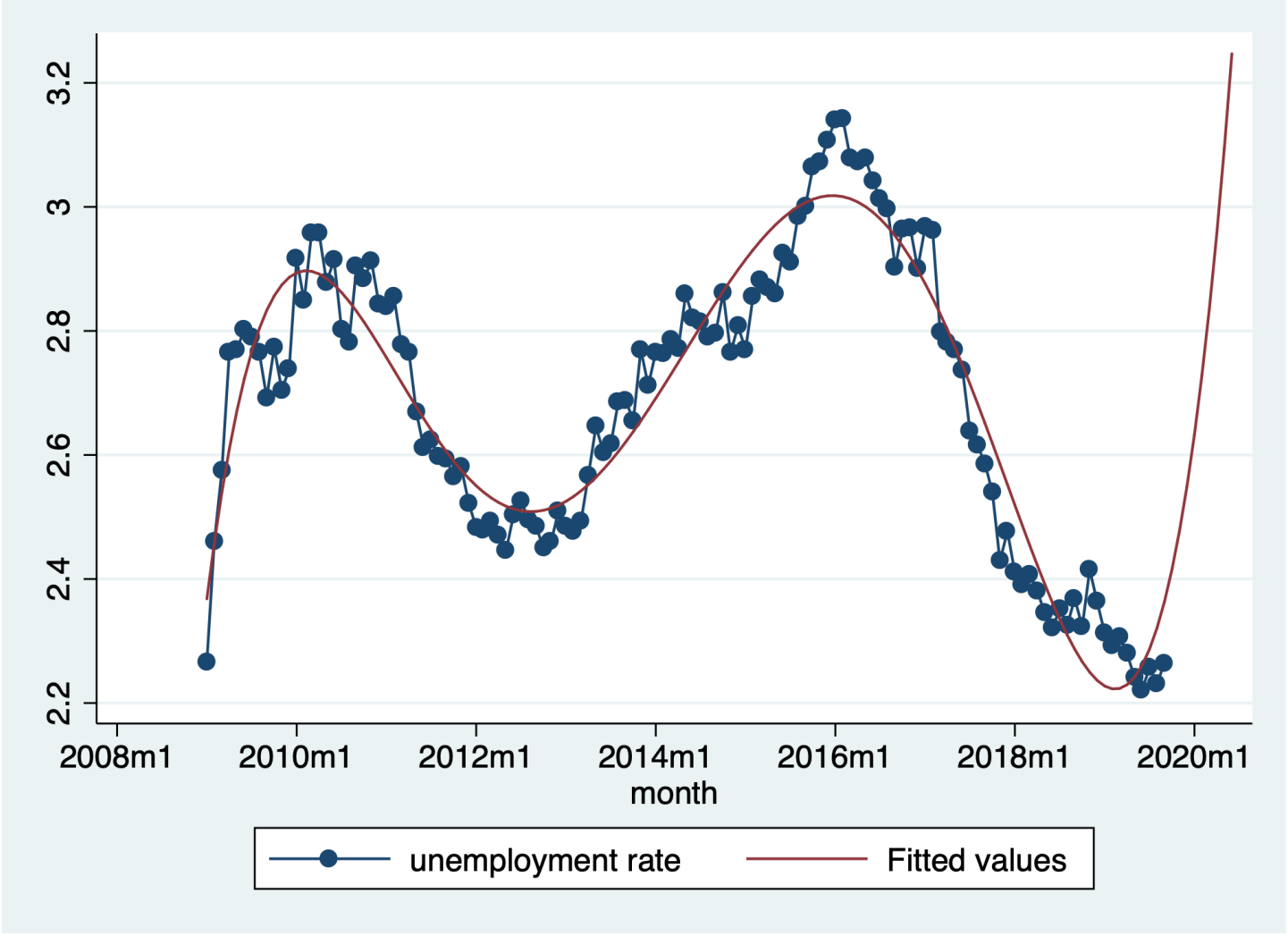
- How do we make forecasts using the regression with polynomial trends?
  - Plug in the time period in the variable t
- Example: Using the fifth-degree polynomial trend model, what is the forecast and prediction interval of the unemployment rate in November 2019?

Problems with Polynomial Trends

- Be **cautious** when using polynomial trend to forecast time series
- Polynomials extrapolate past trends
  and lead to **poor forecasts** when trends change direction
- Forecasting with polynomial is useful only if the patterns persist into the future
  - You cannot rely on a polynomial that fits well during the stable period if you suspect that the stability may be short-lived
- Polynomials with high powers of time can be very **unreliable**
  - They will fit historical data,
    but not necessarily generate accurate forecasts
  - Relates to "overfitting"

Case Study: Forecasting with Polynomial Trend

$$unemprate_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 t^4 + \beta_5 t^5 + \epsilon_t$$



Autoregressions

- An **autoregression** is a regression that uses prior values of the response as predictors
  - These prior values are called **lagged** variables
- The simplest autoregression has one lag called a **first-order** autoregression, AR(1)

$$y_t = \beta_0 + \beta_1 y_{t-1} + \epsilon_t$$

- We can also have **more lags** than one
- Example: third-order autoregression, AR(3)

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_3 y_{t-3} + \epsilon_t$$

Case Study: Lagged Unemployment Rate

	month	t	unemprate
1	2009m1	1	2.266534
2	2009m2	2	2.460776
3	2009m3	3	2.573925
4	2009m4	4	2.764229
5	2009m5	5	2.769858
6	2009m6	6	2.802247
7	2009m7	7	2.78899

- Declare dataset as time series in Stata to take advantage of "lagged" operator `L.`
- AR(1): $unemprate_t = \beta_0 + \beta_1 unemprate_{t-1} + \epsilon_t$
- AR(3): $unemprate_t = \beta_0 + \beta_1 unemprate_{t-1} + \beta_2 unemprate_{t-2} + \beta_3 unemprate_{t-3} + \epsilon_t$

Case Study: AR(1) for Unemployment Rate

```
. regress unemprate L1.unemprate if t >= 4 ;
```

Source	SS	df	MS
Model	6.65136274	1	6.65136274
Residual	.388360312	124	.003131938
Total	7.03972306	125	.056317784

Number of obs	= 126
F(1, 124)	= 2123.72
Prob > F	= 0.0000
R-squared	= 0.9448
Adj R-squared	= 0.9444
Root MSE	= .05596

unemprate	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
unemprate L1.	.9838605	.0213494	46.08	0.000	.9416041	1.026117
_cons	.0409929	.057692	0.71	0.479	-.0731957	.1551816

Even with just 1 explanatory variable, this regression explains more variation in unemployment rate than the fifth-order polynomial

Case Study: AR(3) for Unemployment Rate

```
. regress unemprate L1.unemprate L2.unemprate L3.unemprate if t >= 4;
```

Source	SS	df	MS
Model	6.66924396	3	2.22308132
Residual	.370479091	122	.003036714
Total	7.03972306	125	.056317784

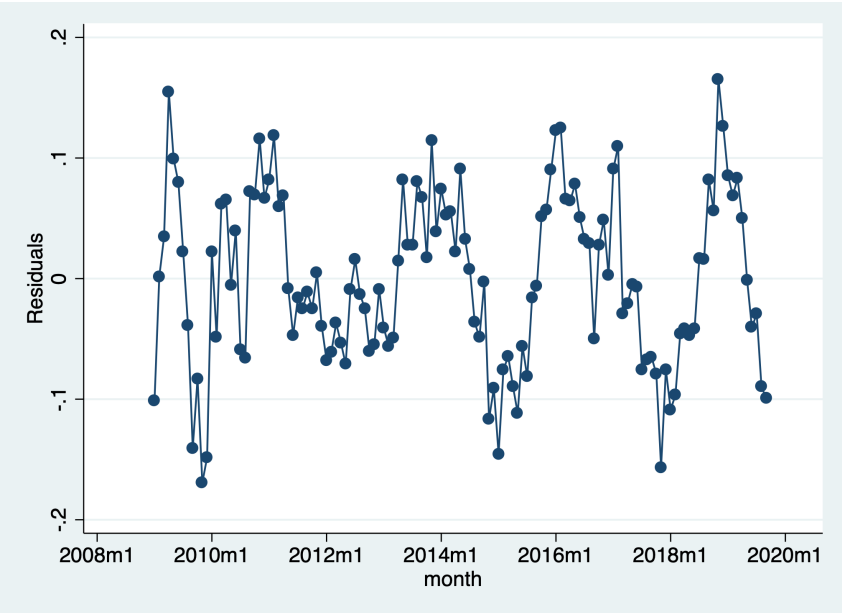
Number of obs	126
F(3, 122)	732.07
Prob > F	0.0000
R-squared	0.9474
Adj R-squared	0.9461
Root MSE	.05511

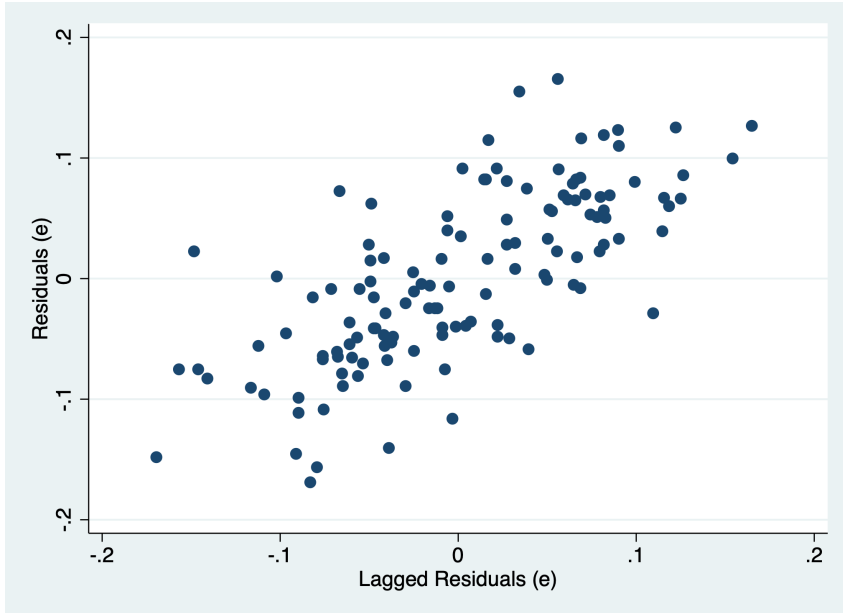
unemprate	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
unemprate						
L1.	.8959153	.0871901	10.28	0.000	.7233138	1.068517
L2.	.2664708	.1175873	2.27	0.025	.0336951	.4992466
L3.	-.180357	.083877	-2.15	0.034	-.3463999	-.0143141
_cons	.0457792	.0581256	0.79	0.432	-.0692862	.1608445

Understanding and Building Autoregressions

- The slopes in autoregressions are **not** easily interpretable
  - Autoregressions are generally used for **short-term forecasting**
  - **Not** for analyzing how explanatory variables impact y
- How many lags to include? Usually done by **trial and error**
  - Try different lags, select the model that offers the best combination of goodness of fit ($\overline{R}^2$ and s_e) and statistically significant coefficients
- Case study example: AR(3) has almost the same $\overline{R}^2$ as AR(1), so we pick AR(1) model for simplicity.
- Why do we care about autoregressions?
  - Lets us capture large amounts of **dependence** in the y variable
  - Helps to ensure that independence of residuals holds

Residuals e in fifth-order polynomial regression

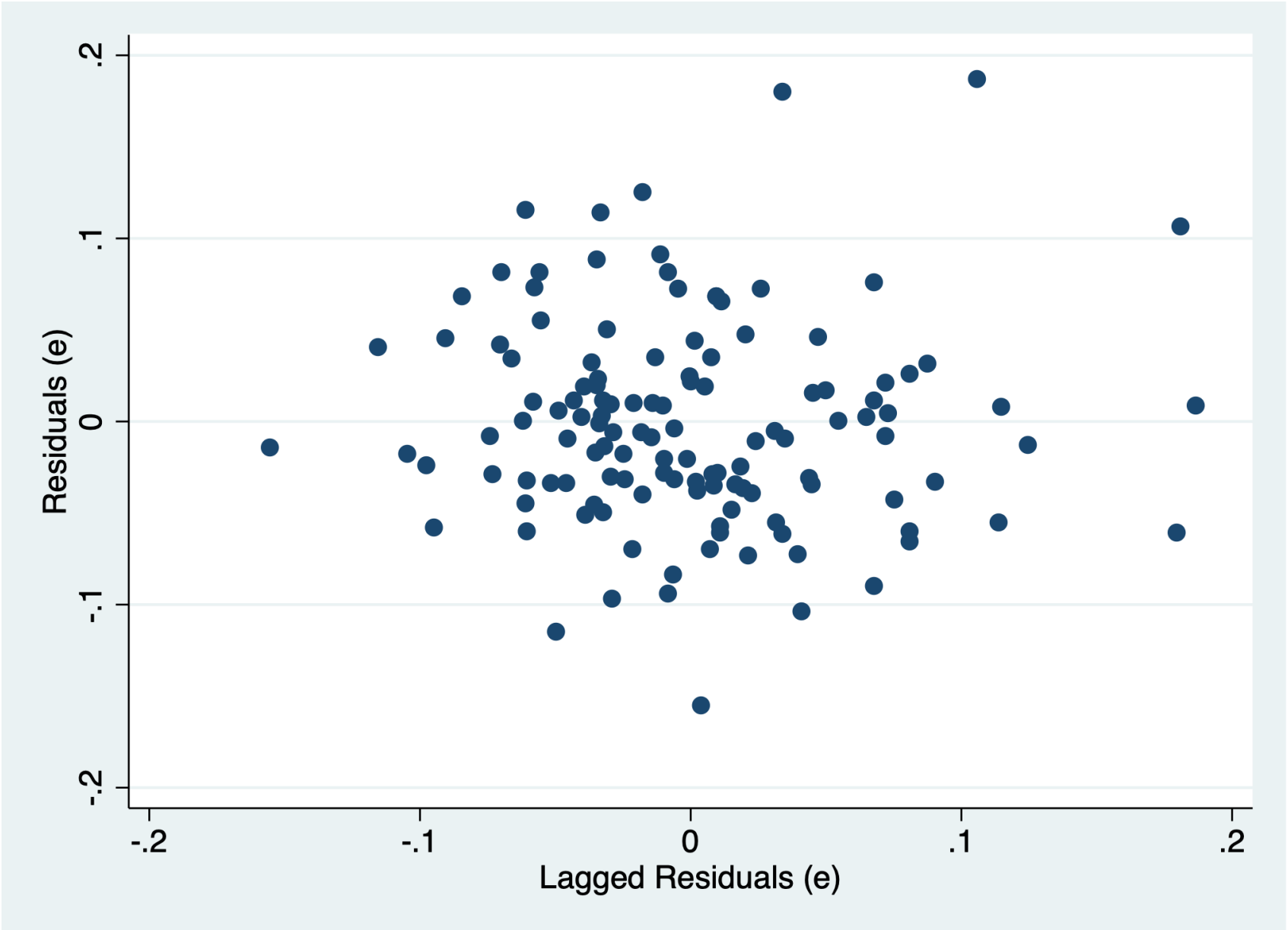




$$corr(e, e_{t-1}) = 0.7209$$

Durbin Watson (DW) statistic $= 0.5496$

Residuals e in AR(1) regression



Case Study: Forecasting unemployment with AR(1) Model

- What is the forecast for Oct. 2019 (1 period beyond data)?

- What is the 95% prediction interval for this forecast?

Case Study: Forecasting unemployment with AR(1) Model

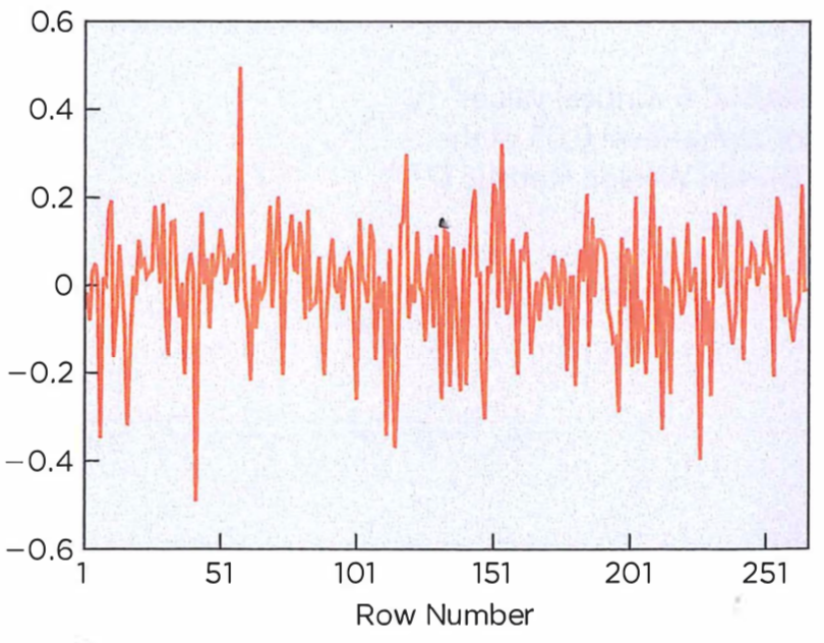
- What is the forecast for Nov. 2019 (2 periods beyond data)?

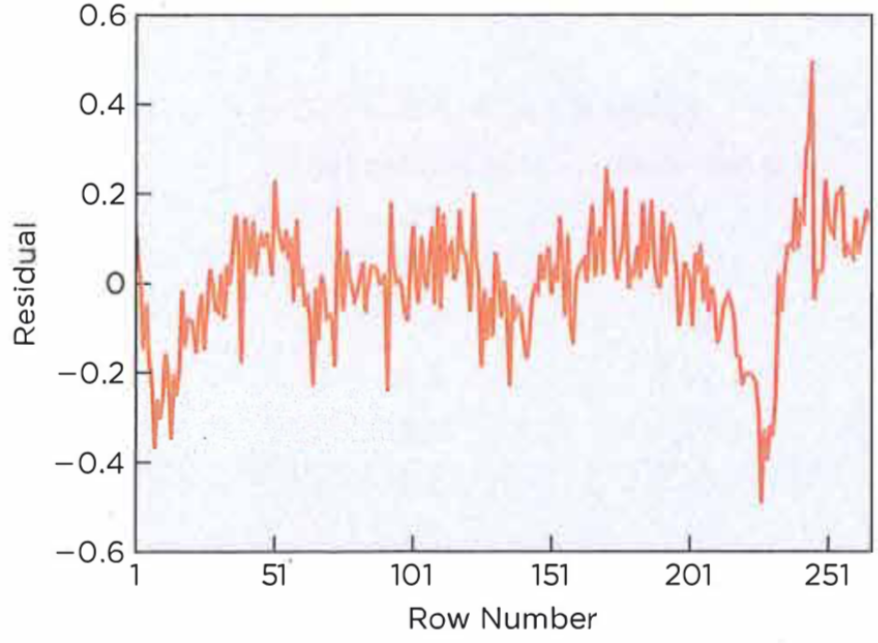
- Prediction interval for Nov. 2019 is hard to compute by hand (not covered in this course)
  - Forecasts are used in place of observations
  - Uncertainty is compounding

Checking the Model

- As before, we need to check the multiple regression conditions
  1. Linear
  2. No obvious lurking variables
  3. Errors are independent
  4. Errors have similar variances
  5. Errors are nearly normal
- The most important of these in time series context is the independence of the model errors
  - Usually the residual e_t is correlated with e_{t-1}
  - If the errors are not independent, s_e and standard error of the slopes are too small
- Use Durbin-Watson statistic to check dependence
  - Can't apply DW test when using an autoregression

Independent vs. Dependent Residuals





Summary and Best Practices

- Examine these plots of residuals when fitting a time series regression
  - Time plot of residuals
  - Scatter plot of residuals vs. fitted values
  - Scatter plot of residuals versus lags of the residuals
- The above plots can show dependence in the residuals
- Using autoregressions can help deal with dependence
- Avoid polynomials with high powers