Quantitative Methods

Lecture 14 Introduction to Time Series

Outline

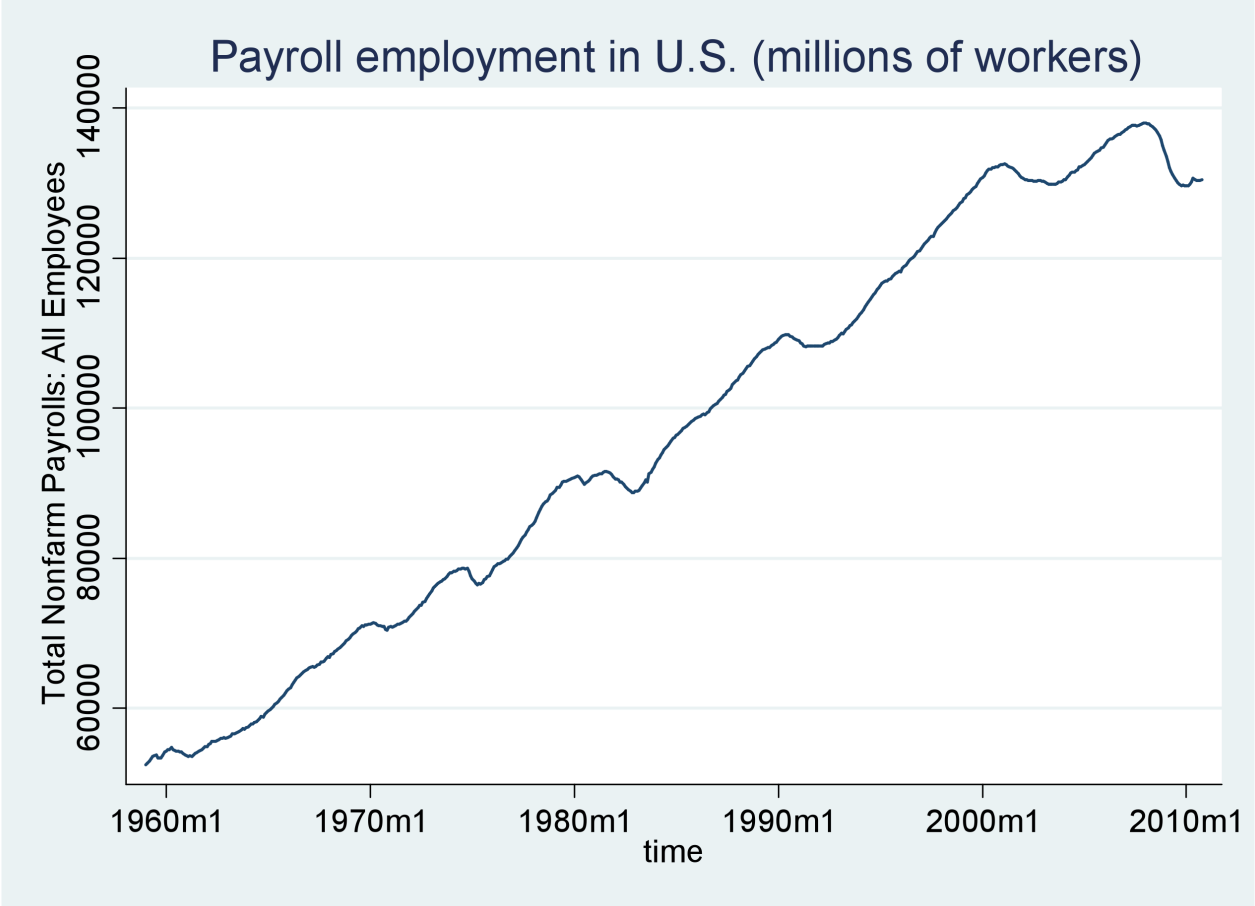
1. Time Series Data: What's Different?
2. Using Regression Models for Forecasting
3. Lags, Differences, Autocorrelation, & Stationarity
4. Autoregressions
5. The Autoregressive – Distributed Lag (ADL) Model
6. Forecast Uncertainty and Forecast Intervals
7. Lag Length Selection: Information Criteria
8. Nonstationarity I: Trends
9. Nonstationarity II: Breaks
10. Summary

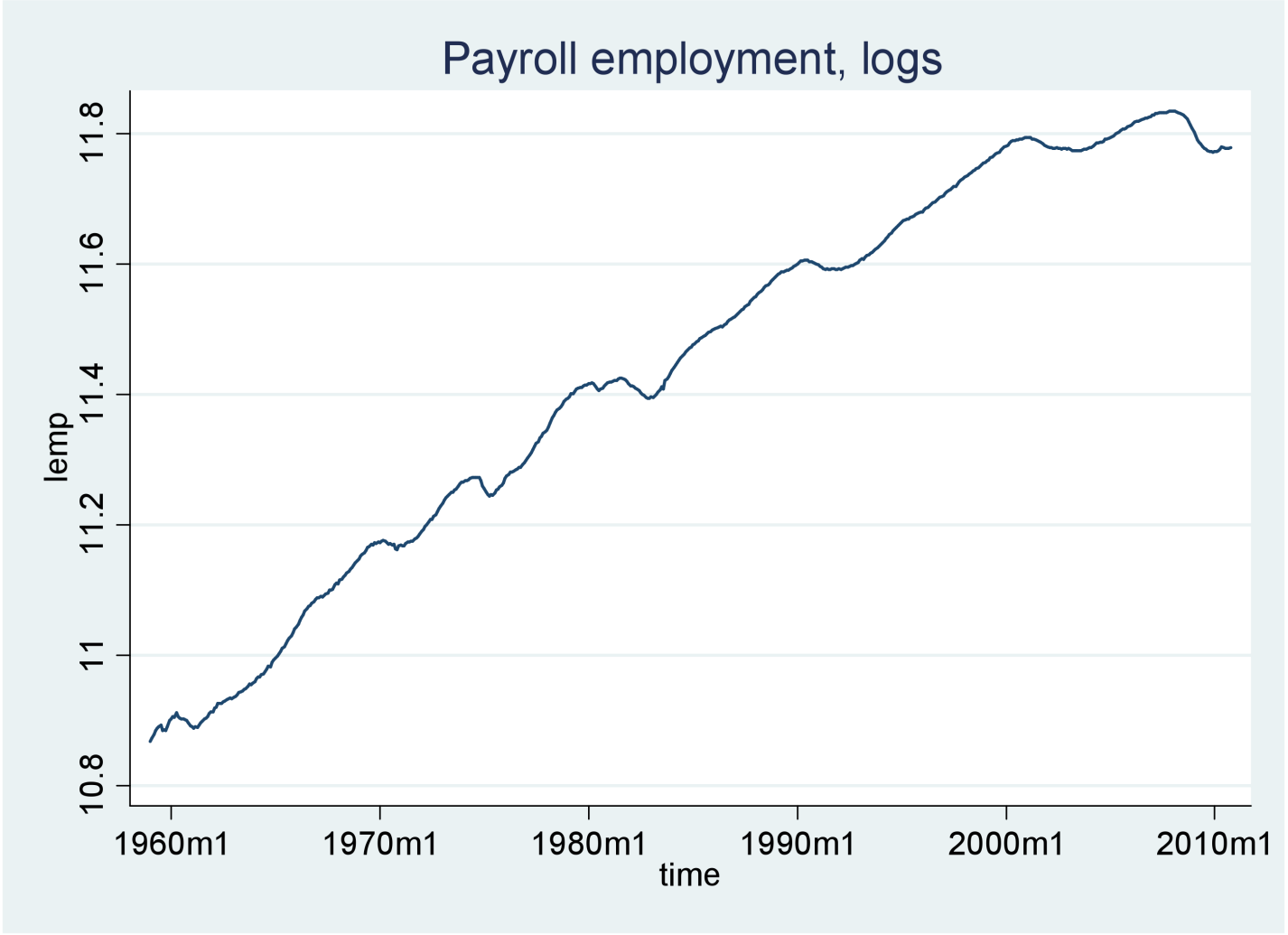
1. Time Series Data: What's Different?

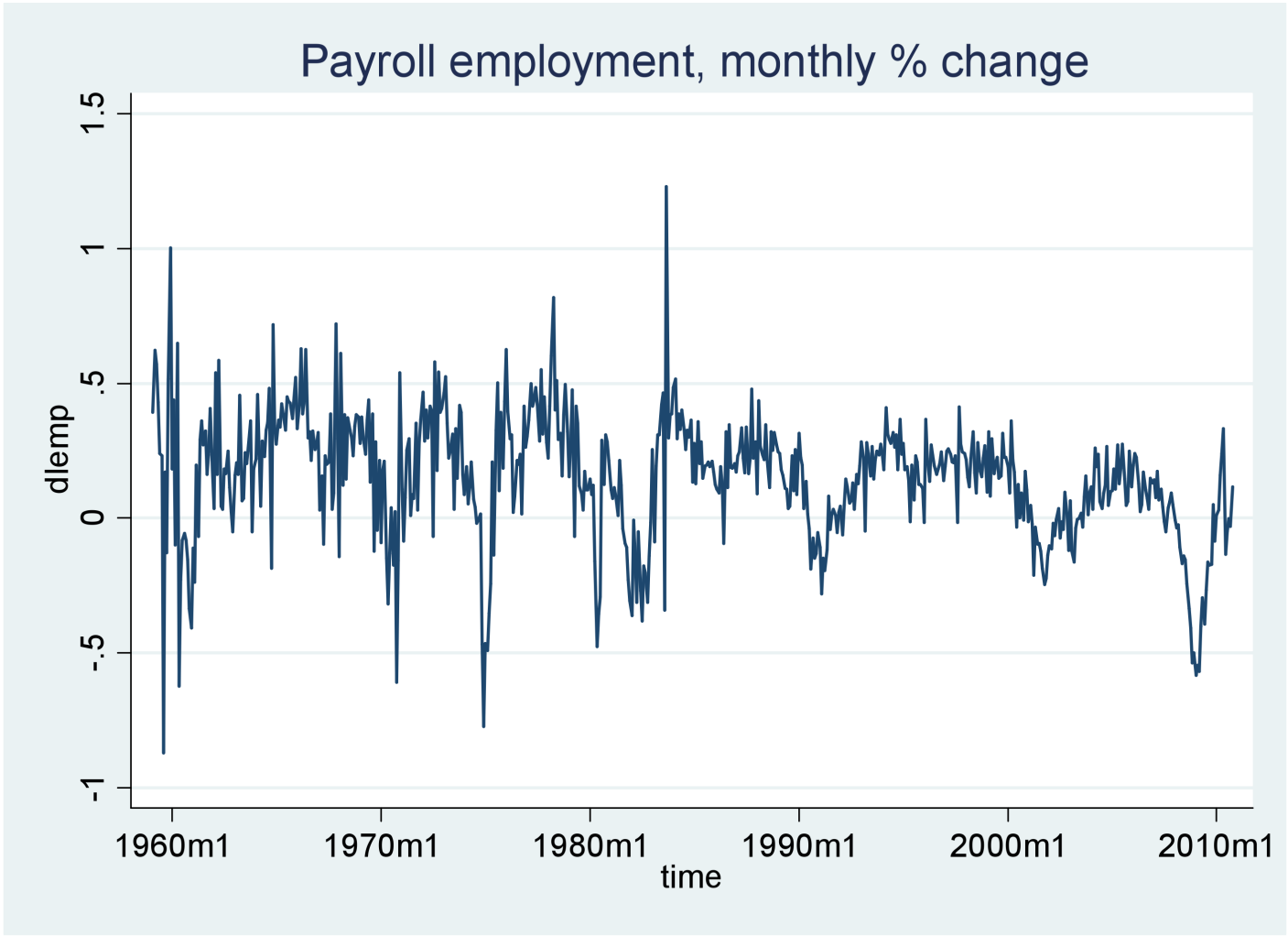
Time series data are data collected on the same observational unit at multiple time periods

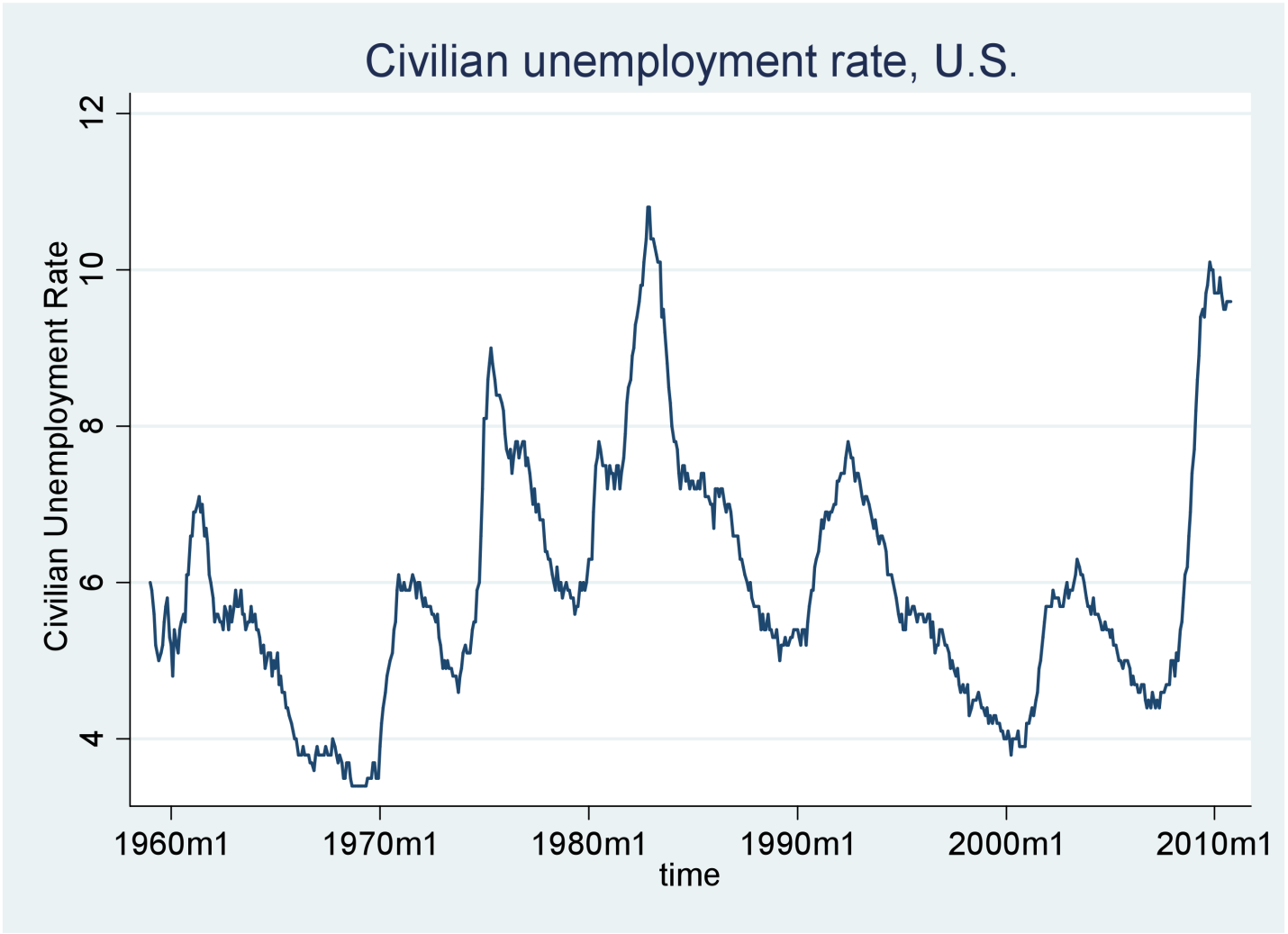
- Aggregate consumption and GDP for a country (for example, 20 years of quarterly observations = 80 observations)
- Yen/$, pound/$ and Euro/$ exchange rates (daily data for 1 year = 365 observations)
- Cigarette consumption per capita in California, by year (annual data)

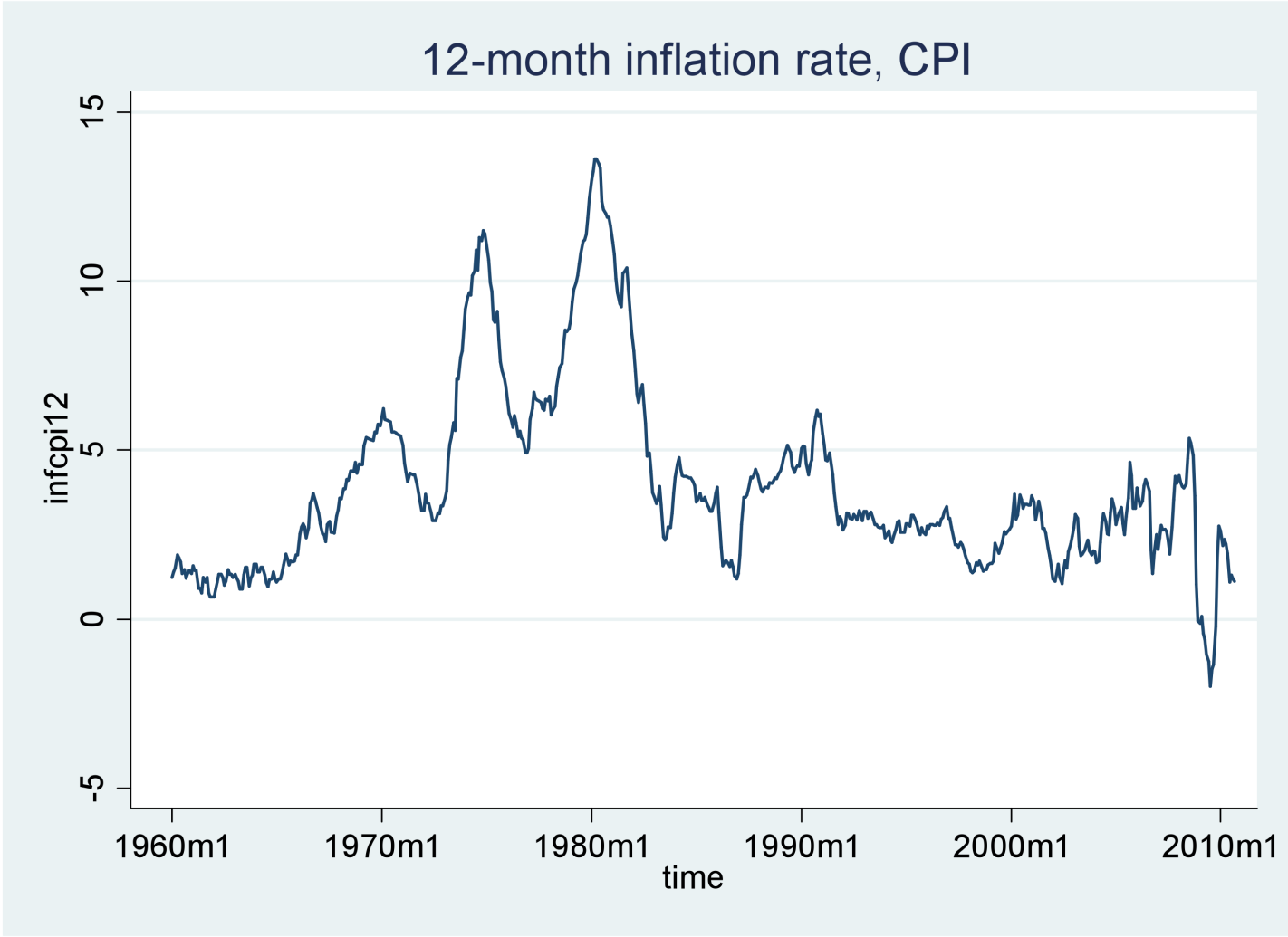
Some monthly U.S. macro and financial time series




Payroll employment, logs
lemp
10.8
11
11.2
11.4
11.6
11.8
1960m1
1970m1
1980m1
1990m1
2000m1
2010m1
time


Payroll employment, monthly % change
dlemp
1.5
1
.5
0
-.5
-1
1960m1
1970m1
1980m1
1990m1
2000m1
2010m1
time


Civilian unemployment rate, U.S.
Civilian Unemployment Rate
4
6
8
10
12
1960m1
1970m1
1980m1
1990m1
2000m1
2010m1
time


12-month inflation rate, CPI
infcpi12
15
10
5
0
-5
1960m1
1970m1
1980m1
1990m1
2000m1
2010m1
time

Some uses of time series data

- Forecasting (SW Ch. 14)
- Estimation of dynamic causal effects (SW Ch. 15)
  - If the Fed increases the Federal Funds rate now, what will be the effect on the rates of inflation and unemployment in 3 months? In 12 months?
  - What is the effect over time on cigarette consumption of a hike in the cigarette tax?
- Modeling risks, which is used in financial markets (one aspect of this, modeling changing variances and "volatility clustering," is discussed in SW Ch. 16)
- Applications outside of economics include environmental and climate modeling, engineering (system dynamics), computer science (network dynamics),…

Time series data raises new technical issues

- Time lags
- Correlation over time (*serial correlation*, a.k.a. *autocorrelation* – which we encountered in panel data)
- Calculation of standard errors when the errors are serially correlated

2. Using Regression Models for Forecasting

- Forecasting and estimation of causal effects are quite different objectives.
- For forecasting,
  - $\bar{R}^2$matters (a lot!)
  - Omitted variable bias isn't a problem!
  - We won't worry about interpreting coefficients in forecasting models – no need to estimate causal effects if all you want to do is forecast!
  - External validity is paramount: the model estimated using historical data must hold into the (near) future

3. Introduction to Time Series Data and Serial Correlation

Time series basics:

A. Notation
B. Lags, first differences, and growth rates
C. Autocorrelation (serial correlation)

A. Notation

- Y_t = value of Y in period t.
- Data set: $\{Y_1,\ldots,Y_T\}$ are T observations on the time series variable Y
- We consider only consecutive, evenly-spaced observations (for example, monthly, 1960 to 1999, no missing months) (missing and unevenly spaced data introduce technical complications)

B. Lags, first differences, and growth rates

Lags, First Differences, Logarithms, and Growth Rates

- The first lag of a time series Y_t is Y_{t-1}; its j^{th} lag is Y_{t-j}.
- The first difference of a series, ΔY_t, is its change between periods $t-1$ and t; that is, $\Delta Y_t = Y_t - Y_{t-1}$.
- The first difference of the logarithm of Y_t is $\Delta \ln(Y_t) = \ln(Y_t) - \ln(Y_{t-1})$.
- The percentage change of a time series Y_t between periods $t-1$ and t is approximately $100\Delta \ln(Y_t)$, where the approximation is most accurate when the percentage change is small.

Example: Quarterly rate of inflation at an annual rate (U.S.)
CPI = Consumer Price Index (Bureau of Labor Statistics)

- CPI in the first quarter of 2004 (2004:I) = 186.57
- CPI in the second quarter of 2004 (2004:II) = 188.60
- Percentage change in CPI, 2004:I to 2004:II

$$= 100 \times \left(\frac{188.60 - 186.57}{186.57} \right) = 100 \times \left(\frac{2.03}{186.57} \right) = 1.088\%$$

- Percentage change in CPI, 2004:I to 2004:II, *at an annual rate* = 4×1.088 = **4.359**% ≈ **4.4**% (percent per year)
- Like interest rates, inflation rates are (as a matter of convention) reported at an annual rate.
- Using the logarithmic approximation to percent changes yields 4×100×[log(188.60) – log(186.57)] = **4.329**%

Example: US CPI inflation – its first lag and its change

TABLE 14.1 Inflation in the United States in 2004 and the First Quarter of 2005

Quarter	U.S. CPI	Rate of Inflation at an Annual Rate (Inf_t)	First Lag (Inf_{t-1})	Change in Inflation (ΔInf_t)
2004:I	186.57	3.8	0.9	2.9
2004:II	188.60	4.4	3.8	0.6
2004:III	189.37	1.6	4.4	−2.8
2004:IV	191.03	3.5	1.6	1.9
2005:I	192.17	2.4	3.5	−1.1

The annualized rate of inflation is the percentage change in the CPI from the previous quarter to the current quarter, multiplied by four. The first lag of inflation is its value in the previous quarter, and the change in inflation is the current inflation rate minus its first lag. All entries are rounded to the nearest decimal.

C. Autocorrelation (serial correlation)

The correlation of a series with its own lagged values is called ***autocorrelation*** or ***serial correlation***.

- The first ***autocovariance*** of Y_t is cov(Y_t,Y_{t-1})
- The first ***autocorrelation*** of Y_t is corr(Y_t,Y_{t-1})
- Thus

$$\text{corr}(Y_t,Y_{t-1}) = \frac{\text{cov}(Y_t, Y_{t-1})}{\sqrt{\text{var}(Y_t)\,\text{var}(Y_{t-1})}} = \rho_1$$

- These are population correlations – they describe the population joint distribution of (Y_t, Y_{t-1})

Autocorrelation (Serial Correlation) and Autocovariance

The j^{th} autocovariance of a series Y_t is the covariance between Y_t and its j^{th} lag, Y_{t-j}, and the j^{th} autocorrelation coefficient is the correlation between Y_t and Y_{t-j}. That is,

$$j^{\text{th}} \text{ autocovariance} = \text{cov}(Y_t, Y_{t-j}) \tag{14.3}$$

$$j^{\text{th}} \text{ autocorrelation} = \rho_j = \text{corr}(Y_t, Y_{t-j}) = \frac{\text{cov}(Y_t, Y_{t-j})}{\sqrt{\text{var}(Y_t)\text{var}(Y_{t-j})}}. \tag{14.4}$$

The j^{th} autocorrelation coefficient is sometimes called the j^{th} serial correlation coefficient.

Sample autocorrelations

The ***j*th *sample autocorrelation*** is an estimate of the *j*th population autocorrelation:

$$\hat{\rho}_j = \frac{\text{cov}(Y_t, Y_{t-j})}{\text{var}(Y_t)}$$

where

$$\widehat{\text{cov}(Y_t, Y_{t-j})} = \frac{1}{T}\sum_{t=j+1}^{T}(Y_t - \bar{Y}_{j+1,T})(Y_{t-j} - \bar{Y}_{1,T-j})$$

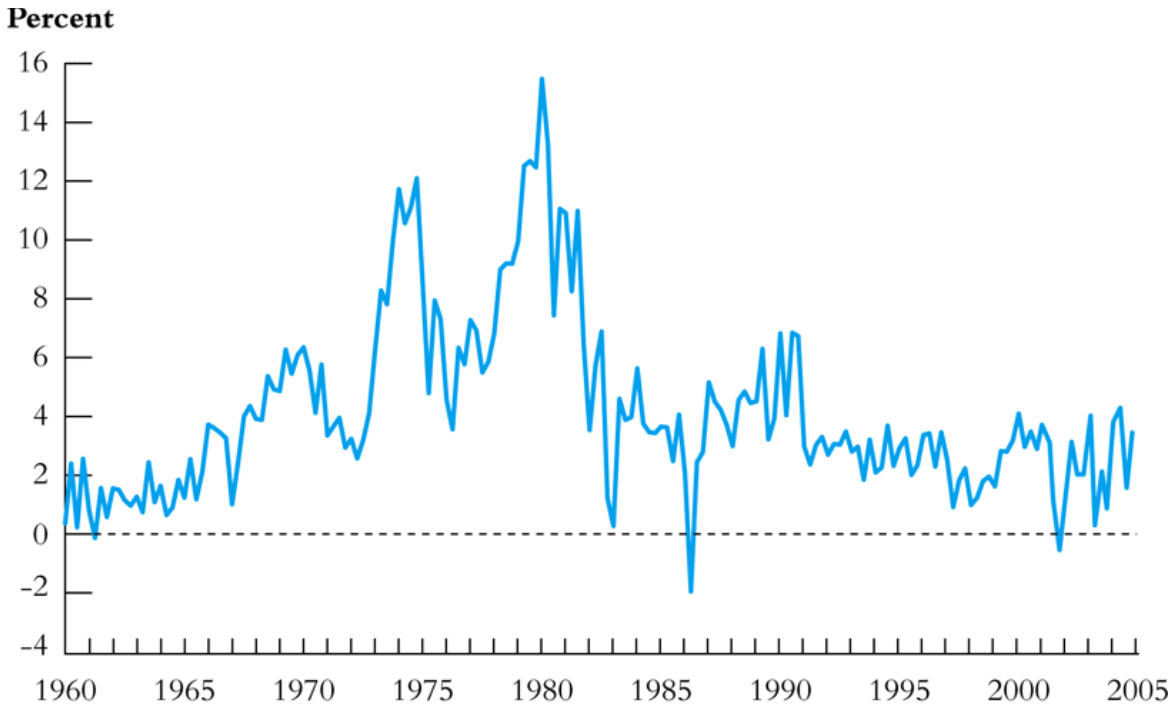
Where $\bar{Y}_{j+1,T}$ is the sample average of Y_t computed over observations $t = j+1,\ldots,T$. *NOTE*:

- The summation is over $t=j+1$ to T (*why?)*
- The divisor is T, not $T - j$ (this is the conventional definition used for time series data)

Example: Autocorrelations of:
(1) the quarterly rate of U.S. inflation
(2) the quarter-to-quarter change in the quarterly rate of inflation

TABLE 14.2 First Four Sample Autocorrelations of the U.S. Inflation Rate and Its Change, 1960:I–2004:IV

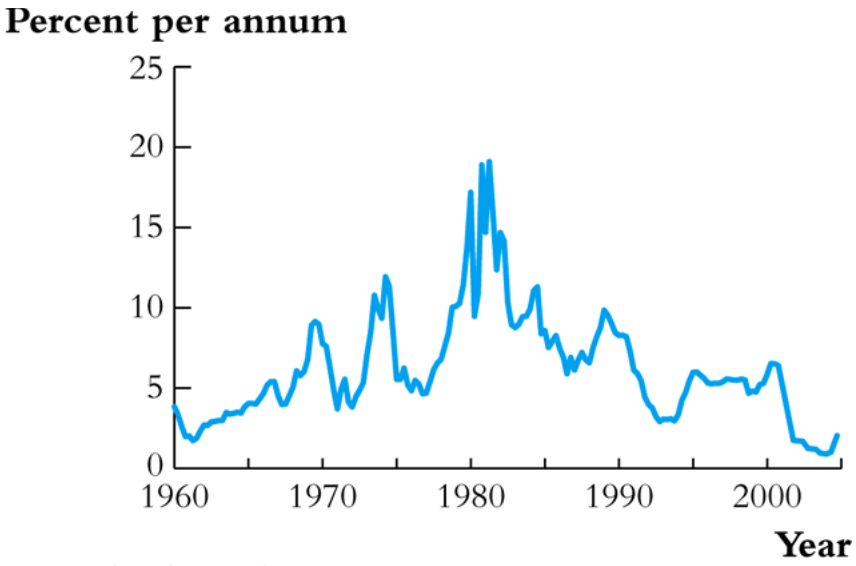
	Autocorrelation of:	
Lag	**Inflation Rate (Inf_t)**	**Change of Inflation Rate (ΔInf_t)**
1	0.84	−0.26
2	0.76	−0.25
3	0.76	0.29
4	0.67	−0.06



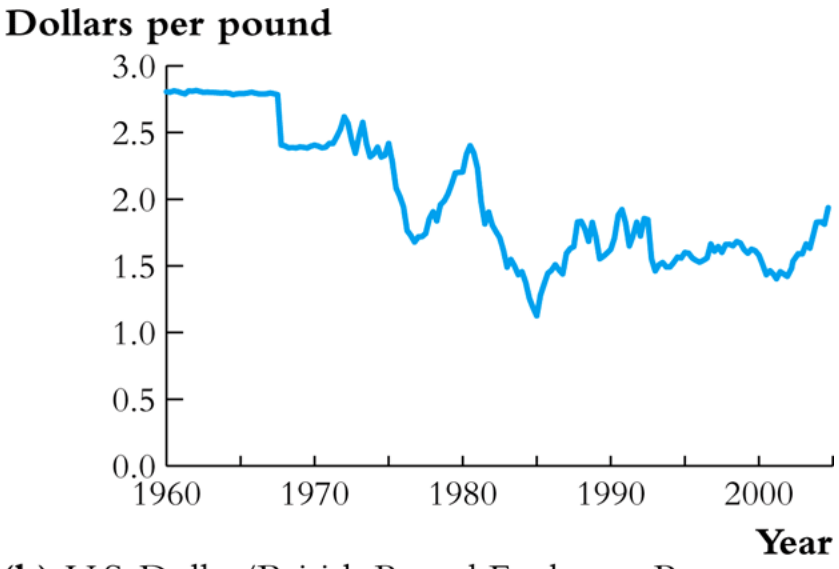
(a) U.S. CPI Inflation Rate

- The inflation rate is highly serially correlated (ρ_1 = .84)
- Last quarter's inflation rate contains much information about this quarter's inflation rate
- The plot is dominated by multiyear swings
- But there are still surprise movements!

Other economic time series: Do these series look serially correlated (is Y_t strongly correlated with Y_{t+1}?)

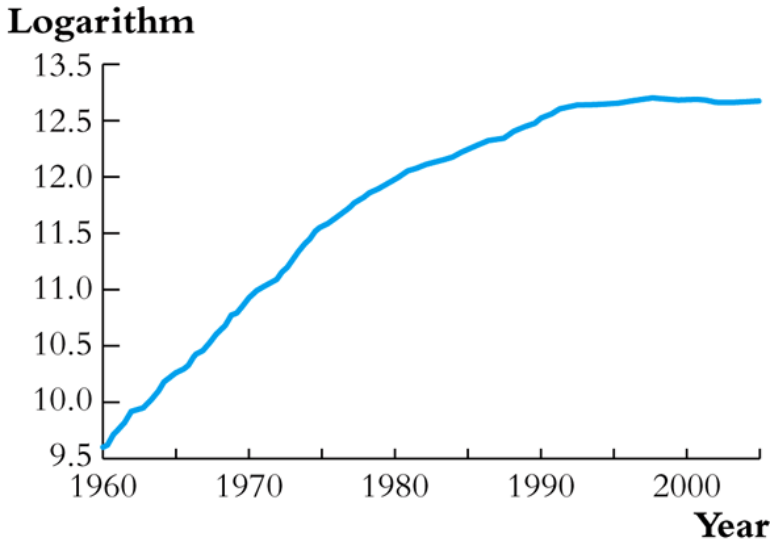


(a) Federal Funds Interest Rate

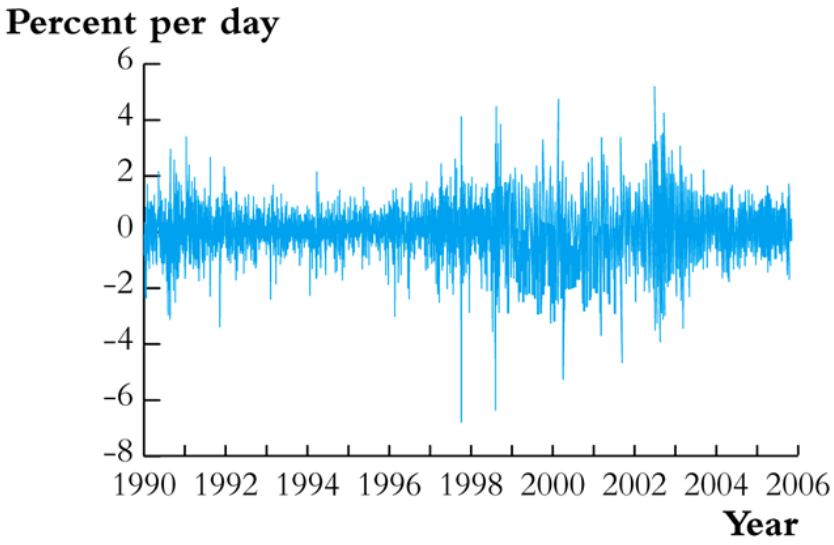


(b) U.S. Dollar/British Pound Exchange Rate

Other economic time series, ctd:



(c) Logarithm of GDP in Japan



(d) Percentage Changes in Daily Values of the NYSE Composite Stock Index

4. Autoregressions (SW Section 14.3)

- A natural starting point for a forecasting model is to use past values of Y (that is, Y_{t-1}, Y_{t-2},…) to forecast Y_t.
- An ***autoregression*** is a regression model in which Y_t is regressed against its own lagged values.
- The number of lags used as regressors is called the ***order*** of the autoregression.
  - In a ***first order autoregression***, Y_t is regressed against Y_{t-1}
  - In a ***p^{th} order autoregression***, Y_t is regressed against $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}$.

The First Order Autoregressive (AR(1)) Model

The population AR(1) model is

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t$$

- β_0 and β_1 *do not* have causal interpretations
- if $\beta_1 = 0$, Y_{t-1} is not useful for forecasting Y_t
- The AR(1) model can be estimated by an OLS regression of Y_t against Y_{t-1} (mechanically, how would you run this regression??)
- Testing $\beta_1 = 0$ v. $\beta_1 \neq 0$ provides a test of the hypothesis that Y_{t-1} is not useful for forecasting Y_t

Example: AR(1) model of the change in inflation

Estimated using data from 1962:I – 2004:IV:

$$\Delta Inf_t = 0.017 - 0.238\Delta Inf_{t-1} \quad \overline{R}^2 = 0.05$$
$$\qquad\qquad (0.126) \quad (0.096)$$

Is the lagged change in inflation a useful predictor of the current change in inflation?

- $t = -.238/.096 = -2.47 > 1.96$ (in absolute value)
- Reject H_0: $\beta_1 = 0$ at the 5% significance level
- Yes, the lagged change in inflation is a useful predictor of current change in inflation–but the $\overline{R}^2$ is pretty low!

Example: AR(1) model of inflation – STATA

First, let STATA know you are using time series data

```
generate time=q(1959q1)+_n-1;
```
_n is the observation no.
So this command creates a new variable time that has a special quarterly date format

```
format time %tq;
```
Specify the quarterly date format

```
sort time;
```
Sort by time

```
tsset time;
```
Let STATA know that the variable time is the variable you want to indicate the time scale

Example: AR(1) model of inflation – STATA, ctd.

```
. gen lcpi = log(cpi);                    variable cpi is already in memory

. gen inf = 400*(lcpi[_n]-lcpi[_n-1]);    quarterly rate of inflation at an
                                          annual rate
```

This creates a new variable, inf, the "nth" observation of which is 400 times the difference between the nth observation on lcpi and the "n-1"th observation on lcpi, that is, the first difference of lcpi

```
                                          compute first 8 sample
   autocorrelations
. corrgram inf if tin(1960q1,2004q4), noplot lags(8);
```

LAG	AC	PAC	Q	Prob>Q
1	0.8359	0.8362	127.89	0.0000
2	0.7575	0.1937	233.5	0.0000
3	0.7598	0.3206	340.34	0.0000
4	0.6699	-0.1881	423.87	0.0000
5	0.5964	-0.0013	490.45	0.0000
6	0.5592	-0.0234	549.32	0.0000
7	0.4889	-0.0480	594.59	0.0000
8	0.3898	-0.1686	623.53	0.0000

`if tin(1962q1,2004q4)` *is STATA time series syntax for using only observations between 1962q1 and 1999q4 (inclusive). The "tin(.,.)" option requires defining the time scale first, as we did above*

Example: AR(1) model of inflation – STATA, ctd

```
. gen dinf = inf[_n]-inf[_n-1];
. reg dinf L.dinf if tin(1962q1,2004q4), r;    L.dinf is the first lag of dinf


Linear regression                                      Number of obs =     172
                                                       F(  1,   170) =    6.08
                                                       Prob > F      =  0.0146
                                                       R-squared     =  0.0564
                                                       Root MSE      =  1.6639

------------------------------------------------------------------------------
             |               Robust
        dinf |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dinf |
         L1. |  -.2380348   .0965034    -2.47   0.015    -.4285342   -.0475354
       _cons |   .0171013   .1268831     0.13   0.893    -.2333681    .2675707
------------------------------------------------------------------------------

.  dis "Adjusted Rsquared = " _result(8);
Adjusted Rsquared = .05082278
```

Forecasts: terminology and notation

- *Predicted values* are "in-sample" (the usual definition)
- *Forecasts* are "out-of-sample" – in the future
- *Notation*:
  - $Y_{T+1|T}$ = forecast of Y_{T+1} based on $Y_T, Y_{T-1}, \ldots$, using the population (true unknown) coefficients
  - $\hat{Y}_{T+1|T}$ = forecast of Y_{T+1} based on $Y_T, Y_{T-1}, \ldots$, using the estimated coefficients, which are estimated using data through period T.
  - For an AR(1):
    - $Y_{T+1|T} = \beta_0 + \beta_1 Y_T$
    - $\hat{Y}_{T+1|T} = \hat{\beta}_0 + \hat{\beta}_1 Y_T$, where $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimated using data through period T.

Forecast errors

The one-period ahead forecast error is,

forecast error = $Y_{T+1} - \hat{Y}_{T+1|T}$

The distinction between a forecast error and a residual is the same as between a forecast and a predicted value:

- a *residual* is "in-sample"
- a *forecast error* is "out-of-sample" – the value of Y_{T+1} isn't used in the estimation of the regression coefficients

Example: forecasting inflation using an AR(1)

AR(1) estimated using data from 1962:I – 2004:IV:

$$\Delta Inf_t = 0.017 - 0.238\Delta Inf_{t-1}$$

$Inf_{2004:III} = 1.6$ (units are percent, at an annual rate)

$Inf_{2004:IV} = 3.5$

$\Delta Inf_{2004:IV} = 3.5 - 1.6 = 1.9$

The forecast of $\Delta Inf_{2005:I}$ is:

$\Delta Inf_{2005:I|2004:IV} = 0.017 - 0.238 \times 1.9 = -0.44 \approx -0.4$

so

$$Inf_{2005:I|2004:IV} = Inf_{2004:IV} + \Delta Inf_{2005:I|2004:IV} = 3.5 - 0.4 = 3.1\%$$

The AR(p) model: using multiple lags for forecasting

The p^{th} order autoregressive model (AR(p)) is

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p} + u_t$$

- The AR(p) model uses p lags of Y as regressors
- The AR(1) model is a special case
- The coefficients do not have a causal interpretation
- To test the hypothesis that $Y_{t-2},\ldots,Y_{t-p}$ do not further help forecast Y_t, beyond Y_{t-1}, use an F-test
- Use t- or F-tests to determine the lag order p
- Or, better, determine p using an "information criterion" (*more on this later…*)

Example: AR(4) model of inflation

$$\Delta Inf_t = .02 - .26\Delta Inf_{t-1} - .32\Delta Inf_{t-2} + .16\Delta Inf_{t-3} - .03\Delta Inf_{t-4},$$
$$(.12) \quad (.09) \quad (.08) \quad (.08) \quad (.09)$$

$\overline{R}^2 = 0.18$

- *F*-statistic testing lags 2, 3, 4 is 6.91 (*p*-value < .001)
- $\overline{R}^2$ increased from .05 to .18 by adding lags 2, 3, 4
- So, lags 2, 3, 4 (jointly) help to predict the change in inflation, above and beyond the first lag – both in a statistical sense (are statistically significant) and in a substantive sense (substantial increase in the $\overline{R}^2$)

Example: AR(4) model of inflation – STATA

```
. reg dinf L(1/4).dinf if tin(1962q1,2004q4), r;

Linear regression                               Number of obs =     172
                                                F(  4,   167) =    7.93
                                                Prob > F      =  0.0000
                                                R-squared     =  0.2038
                                                Root MSE      =  1.5421
------------------------------------------------------------------------------
             |               Robust
        dinf |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dinf |
         L1. |  -.2579205   .0925955    -2.79   0.006    -.4407291   -.0751119
         L2. |  -.3220302   .0805456    -4.00   0.000     -.481049   -.1630113
         L3. |   .1576116   .0841023     1.87   0.063    -.0084292    .3236523
         L4. |  -.0302685   .0930452    -0.33   0.745    -.2139649    .1534278
       _cons |   .0224294   .1176329     0.19   0.849    -.2098098    .2546685
------------------------------------------------------------------------------
```

NOTES

- **L(1/4).dinf** ***is A convenient way to say "use lags 1-4 of dinf as regressors"***
- ***L1,…,L4 refer to the first, second,… 4th lags of dinf***

Example: AR(4) model of inflation – STATA, ctd.

```
.  dis "Adjusted Rsquared = " _result(8);
Adjusted Rsquared = .18474733
```

result(8) is the rbar-squared of the most recently run regression

```
.  test L2.dinf L3.dinf L4.dinf;

 ( 1)  L2.dinf = 0.0
 ( 2)  L3.dinf = 0.0
 ( 3)  L4.dinf = 0.0

       F(  3,   147) =    6.71
            Prob > F =    0.0003
```

L2.dinf is the second lag of dinf, etc.

Digression: we used Δ*Inf, not Inf, in the AR's. Why?*

The AR(1) model of Inf_{t-1} *is an AR(2) model of* Inf_t*:*

$$\Delta Inf_t = \beta_0 + \beta_1 \Delta Inf_{t-1} + u_t$$

or

$$Inf_t - Inf_{t-1} = \beta_0 + \beta_1 (Inf_{t-1} - Inf_{t-2}) + u_t$$

or

$$Inf_t = Inf_{t-1} + \beta_0 + \beta_1 Inf_{t-1} - \beta_1 Inf_{t-2} + u_t$$

$$= \beta_0 + (1 + \beta_1) Inf_{t-1} - \beta_1 Inf_{t-2} + u_t$$

So why use ΔInf_t, not Inf_t?

AR(1) model of ΔInf: $\Delta Inf_t = \beta_0 + \beta_1 \Delta Inf_{t-1} + u_t$

AR(2) model of Inf: $Inf_t = \gamma_0 + \gamma_1 Inf_t + \gamma_2 Inf_{t-1} + v_t$

- When Y_t is strongly serially correlated, the OLS estimator of the AR coefficient is biased towards zero.
- In the extreme case that the AR coefficient = 1, Y_t isn't stationary: the u_t's accumulate and Y_t blows up.
- If Y_t isn't stationary, the regression output can be un reliable (this is complicated – regressions with trending variables can be misleading, *t*-stats don't have normal distributions, etc. – more on this some other time)
- Here, Inf_t is strongly serially correlated – so to keep ourselves in a framework we understand, the regressions are specified using ΔInf

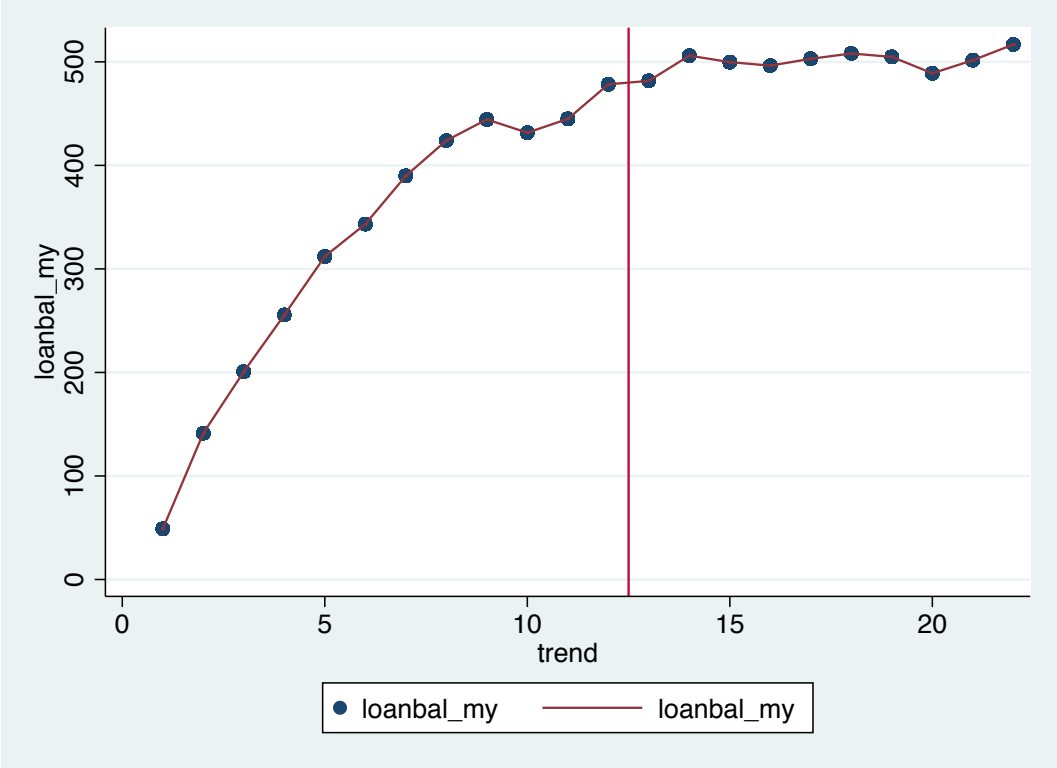
5. Time Series Regression with Additional Predictors and the Autoregressive Distributed Lag (ADL)

- So far we have considered forecasting models that use only past values of *Y*
- It makes sense to add other variables (*X*) that might be useful predictors of *Y*, above and beyond the predictive value of lagged values of *Y*:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \ldots + \beta_p Y_{t-p} + \delta_1 X_{t-1} + \ldots + \delta_r X_{t-r} + u_t$$

- This is an ***autoregressive distributed lag model*** with *p* lags of *Y* and *r* lags of *X* … ***ADL(p,r).***

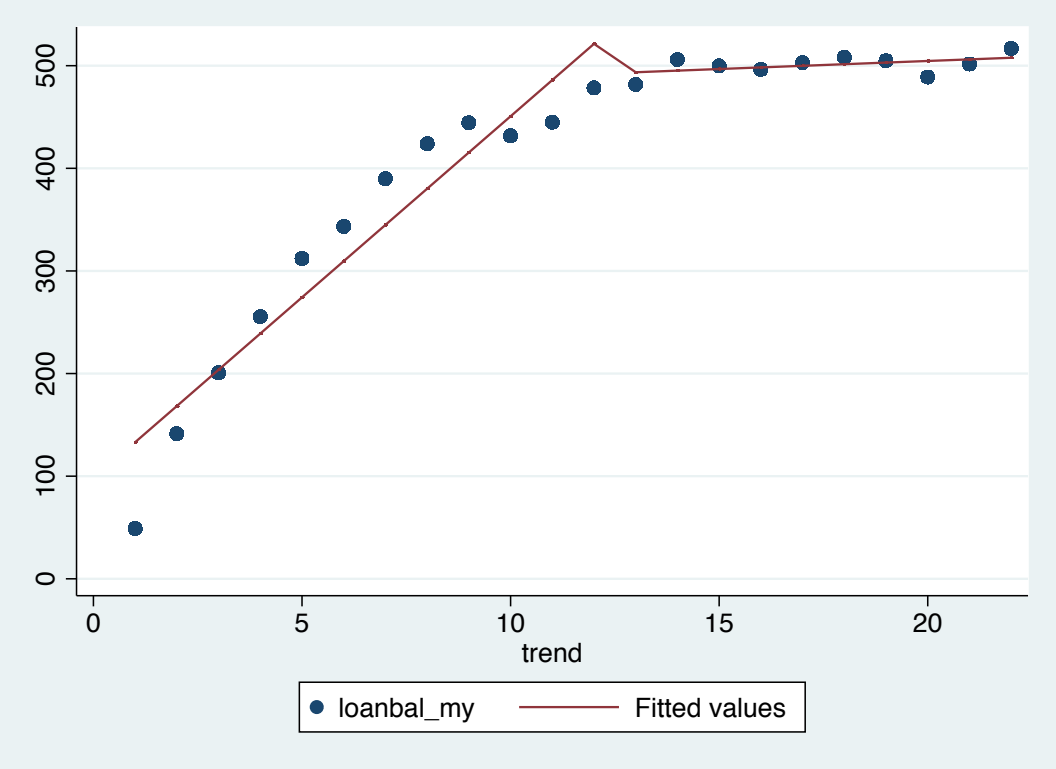
Application: Interrupted Time Series Design



- You want to estimate the effect of a policy on an outcome, but you have only a single time series data set.
- Is this change due to the policy?

The simplest ITSA

```
reg loanbal trend trend_post post if TIKA==1,
cluster(monthyear)
predict hat
graph twoway (scatter loanbal_my trend if TIKA==1) (line hat
trend if TIKA==1)
```



The simplest ITSA

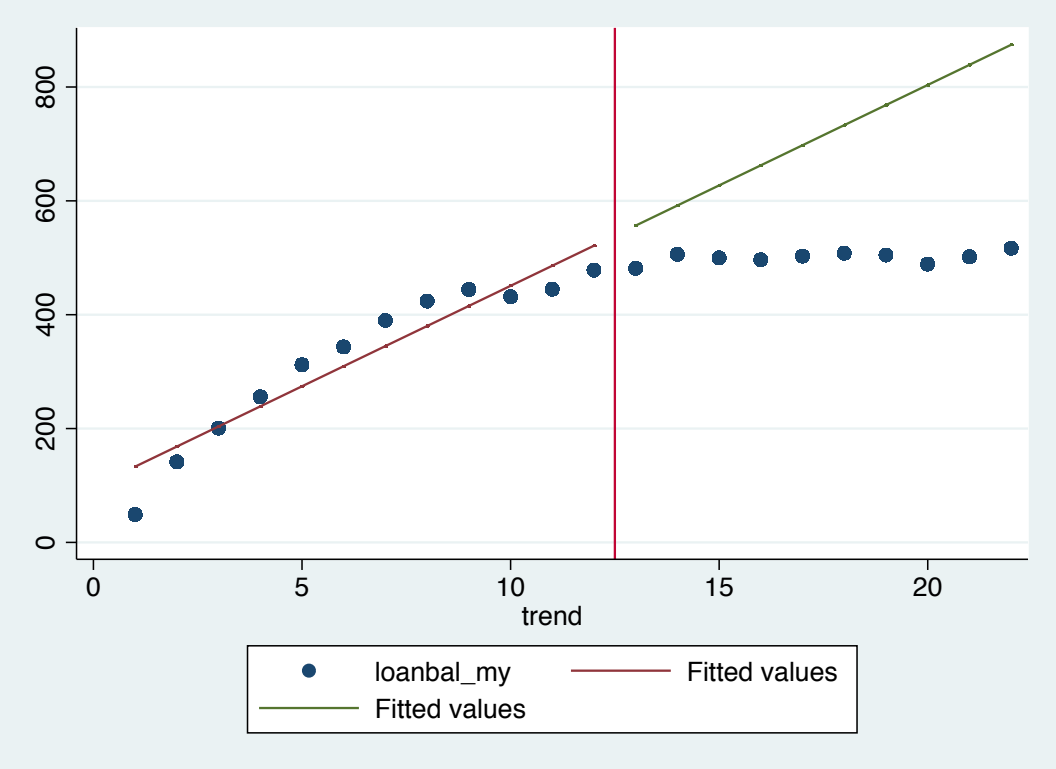
```
. reg loanbal trend trend_post post if TIKA==1, cluster(monthyear)

Linear regression                               Number of obs     =     49,551
                                                F(3, 21)          =      87.65
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0337
                                                Root MSE          =     635.48

                                  (Std. Err. adjusted for 22 clusters in monthyear)
------------------------------------------------------------------------------
             |               Robust
     loanbal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       trend |   35.29128   3.905853     9.04   0.000     27.16861    43.41394
  trend_post |  -33.72453   4.055094    -8.32   0.000    -42.15756    -25.2915
        post |   375.4563   35.30229    10.64   0.000     302.0411    448.8714
       _cons |   97.79587   29.61329     3.30   0.003     36.21166    159.3801
------------------------------------------------------------------------------
```

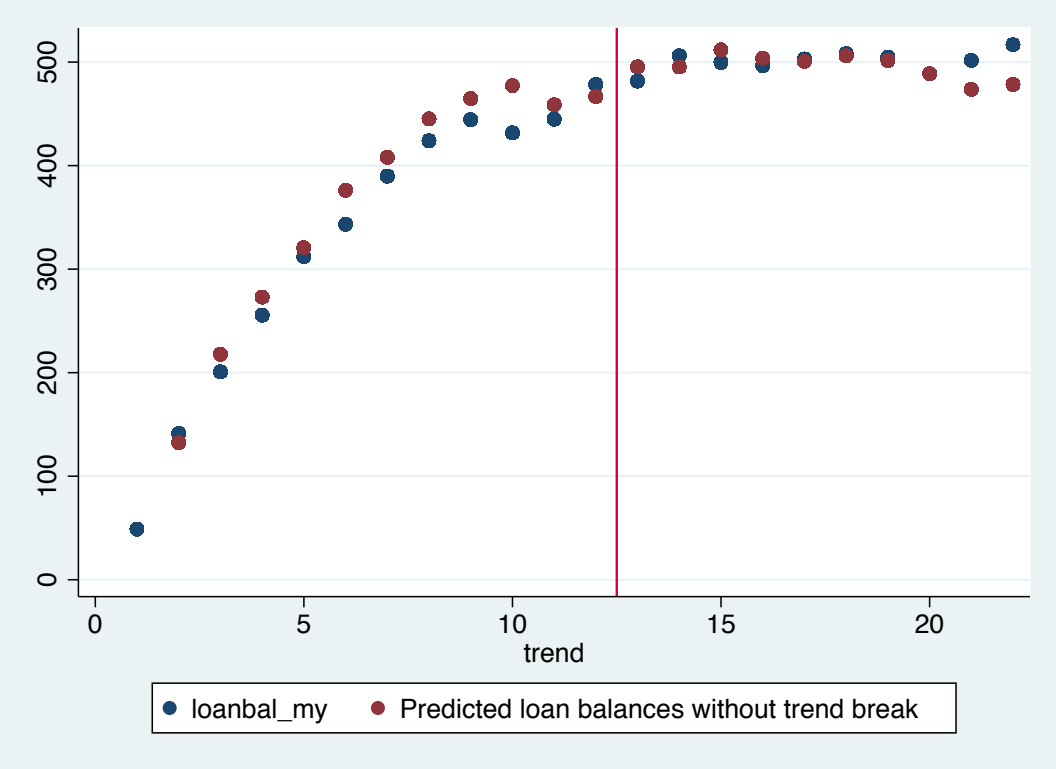
The simplest ITSA

The counterfactual



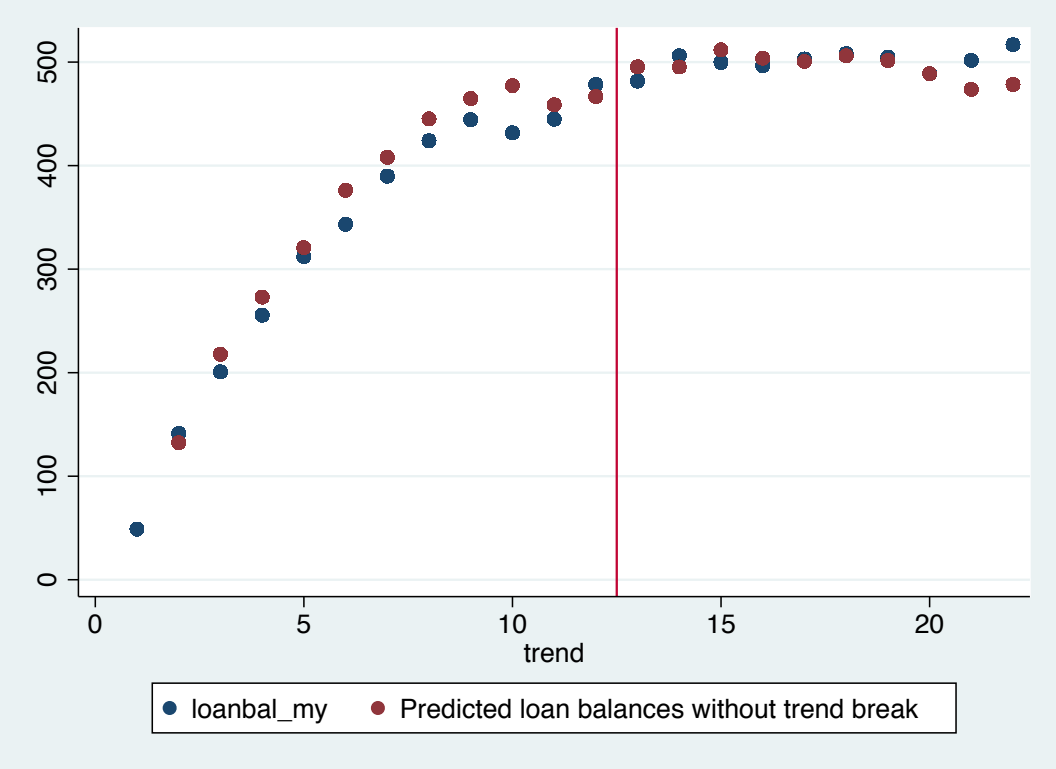
The simplest ITSA

But this is not very strong evidence – if we model the curvature in the pre period then…



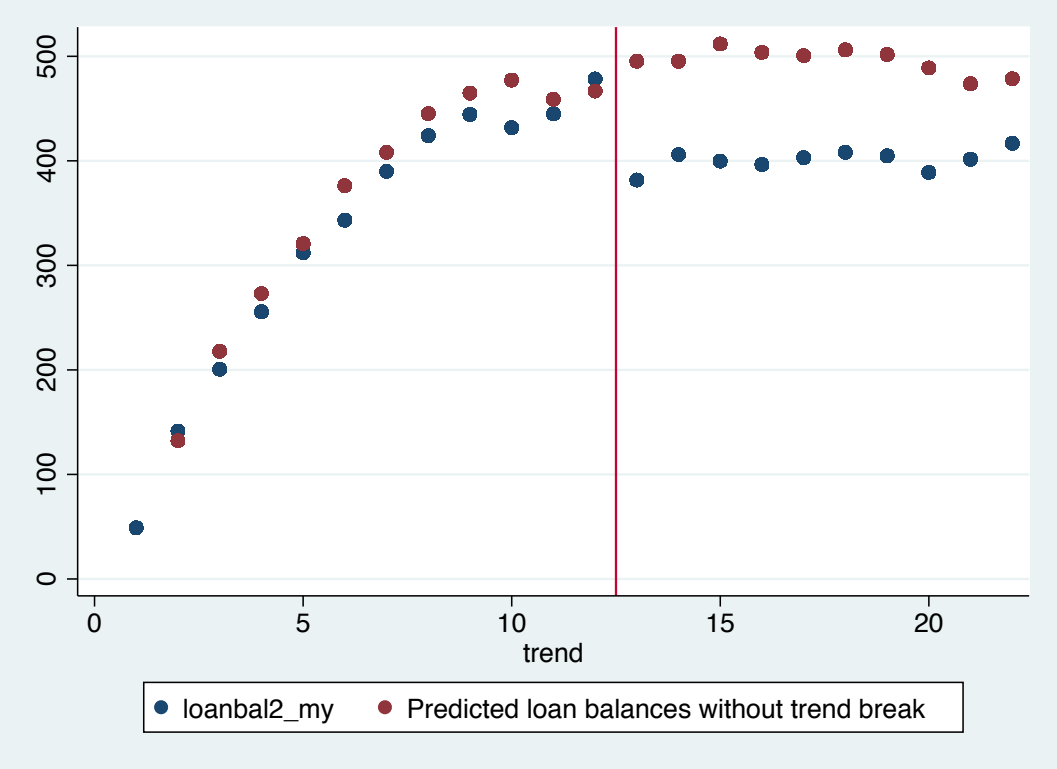
The simplest ITSA

But this is not very strong evidence – if we model the curvature in the pre period then…



The simplest ITSA

But not always – another data sets looks like this



7. Conclusion: Time Series Forecasting Models

- For forecasting purposes, it isn't important to have coefficients with a causal interpretation!

- However, with an interrupted time series design you can come close to policy claims

- Caution is needed though: lack of comparison group heavily exposes this approach to invalid claims.