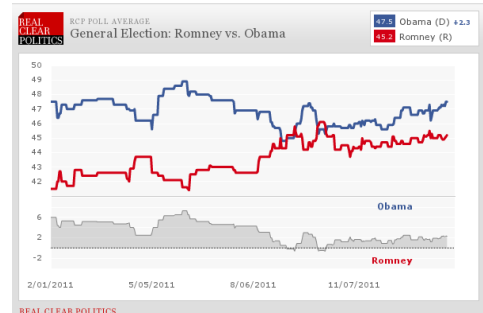
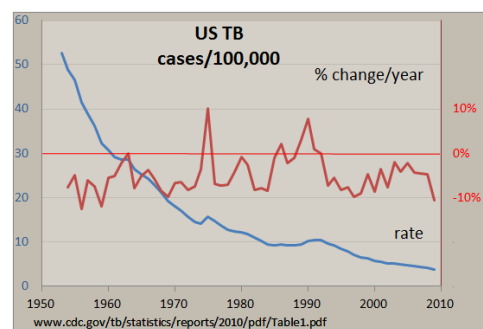


# Zeitreihen

## Statistik II

Wiederholung  
Zeitreihen  
Zusammenfassung

- 1 Wiederholung  
Literatur
- 2 Zeitreihen  
Zeitreihen-Daten  
Modelle  
Probleme  
Trends und Saisonalität  
Fehlerstruktur
- 3 Zusammenfassung



## Zum Nachlesen

- Wooldridge ch 10.1 & 10.2

## Für nächste Woche

- *Einfache* Modelle für Paneldaten
- Wooldridge ch. 13.1-13.4 (im Reader)

## Was sind Zeitreihen?

- Bisher: „Fälle“
- Länder, Parteien, Städte, Personen . . .
- Unabhängig voneinander erhoben, in der Regel durch Zufallsverfahren aus Population (Grundgesamtheit) ausgewählt
- Jeder Fall hat Ordnungsnummer (Index)  $n$ :  $y_1, y_5, y_{27}, y_n \dots$
- *Reihenfolge im Datensatz beliebig*
- Datenmatrix kann umsortiert werden (Spalten und Zeilen)

## Inwiefern sind Zeitreihen anders?

- Zeitreihen beziehen sich auf **ein** Objekt (z. B. ein politisches System)
- Beobachtungen sind **nicht** voneinander unabhängig
- Reihenfolge der Beobachtungen im Datensatz ist nicht beliebig, sondern durch Kalenderzeit vorgegeben
- Spalten der Datenmatrix können umsortiert werden, Zeilen nicht
- Beobachtungen haben Ordnungsnummer (Index)  $t$ , beginnend mit  $t = 1$  (erste Beobachtung)
- Abstände zwischen Beobachtungen sind (normalerweise) gleich groß

- Typischerweise an Kalenderzeit orientiert
- In der Politikwissenschaft gängige Abstände:
  - Tage (tracking polls im Wahlkampf)
  - Wochen (Anzahl terroristischer Anschläge im Irak)
  - Monate (Anteil der Parteiidentifizierer in Deutschland)
  - Quartale (Wirtschaftsleistung und Zufriedenheit mit US-Präsident)
  - Jahre (Zustimmung zur EU und politische Variablen)
  - (Legislaturperioden) (Entwicklung von Parteiprogrammen über die Zeit)
- Beispiel: US-Präsident 1949-85

## Presidential Approval & Unemployment

### Presidential Approval

„Do you approve or disapprove of the way . . . . . is handling his job as President?“

→ Wieviel Prozent Zustimmung zur Amtsführung?

% Approval	% Unemployed	Kalenderzeit	$t$
⋮	⋮	⋮	⋮
69.7	2.8	1953q1	17
73.0	2.7	1953q2	18
74.0	2.6	1953q3	19
⋮	⋮	⋮	⋮

## Lags und Leads

- Lag: Wert aus Vergangenheit (früherer Zeitpunkt)
- (Lead: Wert aus Zukunft)
- Bsp.: Zum Zeitpunkt  $t = 19$ 
  - Ist das erste Lag der Arbeitslosenquote (Vorquartal,  $t - 1$ ) = 2.7
  - Das zweite Lag der Arbeitslosenquote ( $t - 2$ ) = 2.8

% Unemployed	Kalenderzeit	$t$
⋮	⋮	⋮
2.8	1953q1	17
2.7	1953q2	18
2.6	1953q3	19
⋮	⋮	⋮

## Wie kommen Zeitreihen zustande?

- „Normale“ Regression (über Stichproben)
  - ① Stichprobe wird zufällig aus Grundgesamtheit ausgewählt
  - ② Eine Stichprobe, die auch anders aussehen könnte
  - ③ Mathematisches Modell der Stichprobenziehung → Inferenzstatistik
  - ④ Rückschluß auf wahre Parameter in Grundgesamtheit
- Zeitreihenanalyse
  - ① „Datengenerierender Prozeß“ (DGP) erzeugt historischen Ablauf
  - ② Eine Zeitreihe, die auch anders hätte aussehen können
  - ③ Mathematisches Modell des Prozesses (Inferenzstatistik)
  - ④ Rückschluß auf wahre Parameter des Prozesses
- Stärkere Annahmen, u. a. über Stabilität des DGP

## Statisches Modell

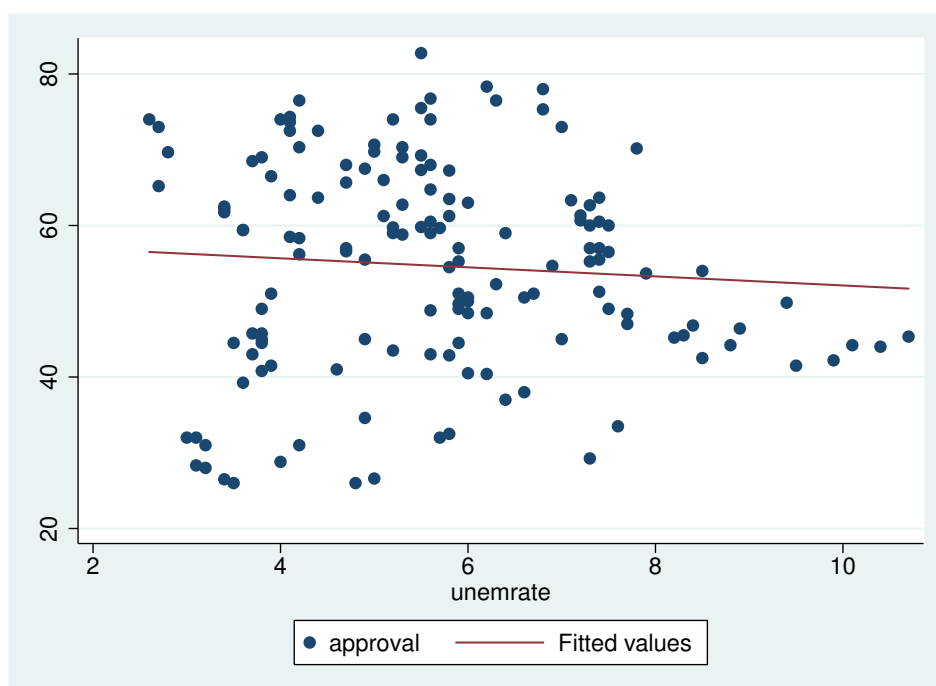
- Lineare Beziehung zwischen  $y$  und  $x$
- Schätzung des Modells
  - Nicht auf Grundlage von unabhängigen Fällen
  - Sondern auf Grundlage von zeitlich gestaffelten Beobachtungen am selben Objekt (z. B. USA)
  - Veränderungen in  $x$  haben **unmittelbaren** Effekt auf  $y$
  - Unemployment  $\rightarrow$  approval (economic voting)

### Statisches Modell

$$\text{approval}_t = \beta_0 + \beta_1 \text{unemployment}_t + \epsilon_t$$

$$y_t = \beta_0 + \beta_1 x_{1,t} + \epsilon_t$$

## Lineare Beziehung $x$ und $y$



## Implikation statisches Modell

```
. reg approval unemrate
```

Source	SS	df	MS			
Model	159.716707	1	159.716707	Number of obs =	148	
Residual	27291.3802	146	186.927262	F( 1, 146) =	0.85	
				Prob > F =	0.3568	
				R-squared =	0.0058	
				Adj R-squared =	-0.0010	
				Root MSE =	13.672	
Total	27451.0969	147	186.742156			

approval	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
unemrate	-.5973806	.6462674	-0.92	0.357	-1.874628	.6798672
_cons	58.05901	3.814606	15.22	0.000	50.52003	65.59799

- Zu jedem Zeitpunkt  $t \dots$
- Ist Zustimmung zur Amtsführung eine lineare Funktion der ALQ ( $58 - 0.6 \times \text{unemrate}$ )
- Z. B. Zunahme ALQ um zwei Prozentpunkte  $\rightarrow$  *sofortige* Abnahme Zustimmung um 1.2 Punkte
- Plausibel?

## (Finite) Distributed Lag Model

- Veränderungen in  $x$  brauchen Zeit, um auf  $y$  zu wirken
- $y$  eine Funktion von  $x$  heute plus vergangene 1, 2,  $\dots$  Werte von  $x$  (Lags)
- Anzahl der vergangenen Zeitpunkte, die wir berücksichtigen können, endlich (finit)
- Im Text: andere Bezeichnung für Koeffizienten der Lags, um Sache klarer zu machen

### FDL mit zwei Lags

$$y_t = \beta_0 + \delta_0 x_t + \delta_1 x_{t-1} + \delta_2 x_{t-2} + \epsilon_t$$

$$\text{approval}_t = \beta_0 + \delta_0 \text{unemployment}_t + \delta_1 \text{unemployment}_{t-1}$$

# Approval & unemployment: lags

```
. list date approval unemrate L1.unemrate L2.unemrate in 1/8, clean
```

	date	approval	unemrate	L. unemrate	L2. unemrate
1.	1949q1	69	3.8	.	.
2.	1949q2	57	4.7	3.8	.
3.	1949q3	57	5.9	4.7	3.8
4.	1949q4	51	6.7	5.9	4.7
5.	1950q1	45	7	6.7	5.9
6.	1950q2	37	6.4	7	6.7
7.	1950q3	43	5.6	6.4	7
8.	1950q4	41	4.6	5.6	6.4

# Verzögerter Effekt der Arbeitslosigkeit

- Annahme: Modell ist korrekt, d. h.  

$$\text{approval}_t = 57.6 - 2.6 \times \text{alq}_t + 3.2 \times \text{alq}_{t-1} - 1 \times \text{alq}_{t-2}$$
- ALQ einmalig 4 Prozent, sonst 2

	date	approval	alq	L. alq	L2. alq
7.	7	56.8	2	2	2
8.	8	56.8	2	2	2
9.	9	56.8	2	2	2
10.	10	51.6	4	2	2
11.	11	63.2	2	4	2
12.	12	54.8	2	2	4
13.	13	56.8	2	2	2
14.	14	56.8	2	2	2

- ALQ 2 Prozent, dann dauerhafter Anstieg auf 4 Prozent

	date	approval	alq	L. alq	L2. alq
7.	7	56.8	2	2	2
8.	8	56.8	2	2	2
9.	9	56.8	2	2	2
10.	10	51.6	4	2	2
11.	11	56.8	2	2	2
12.	12	56.8	2	2	2
13.	13	56.8	2	2	2

13.	13	56	4	4	4
-----	----	----	---	---	---

- Annahme: die beobachteten  $x$  und  $y$  Werte werden zufällig vom DGP hervorgebracht
- Deshalb können wir sie ähnlich wie Stichprobe behandeln
- Tatsächlich: Trends
- Beispiel
  - $x$  und  $y$  nehmen beide über Zeit zu
  - Regression von  $y$  auf  $x$  zeigt positiven Effekt, selbst wenn Korrelation 0 oder schwach negativ
- Drittvariablenproblem → Drittvariablenkontrolle durch Einschluß Kalenderzeit
- Beispiel: Presidential Approval und Verbraucherpreise (CPI)

## Presidential Approval und CPI

```
. reg approval cpi
```

Source	SS	df	MS			
Model	1719.69082	1	1719.69082	Number of obs = 148		
Residual	25731.4061	146	176.242507	F( 1, 146) = 9.76		
Total	27451.0969	147	186.742156	Prob > F = 0.0022		
				R-squared = 0.0626		
				Adj R-squared = 0.0562		
				Root MSE = 13.276		

approval	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
cpi	-.1348399	.0431667	-3.12	0.002	-.2201522	-.0495277
_cons	60.95396	2.283144	26.70	0.000	56.44168	65.46624

```
. reg approval date
```

Source	SS	df	MS			
Model	1632.71801	1	1632.71801	Number of obs = 148		
Residual	25818.3789	146	176.838212	F( 1, 146) = 9.23		
Total	27451.0969	147	186.742156	Prob > F = 0.0028		
				R-squared = 0.0595		
				Adj R-squared = 0.0530		
				Root MSE = 13.298		

approval	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
date	-.0777434	.0255856	-3.04	0.003	-.1283094	-.0271774
_cons	56.98289	1.32836	42.90	0.000	54.35759	59.60819

```
. reg cpi date
```

Source	SS	df	MS			
Model	73781.978	1	73781.978	Number of obs = 148		
Total	94583.0401	147	643.422041	F( 1, 146) = 517.87		
				Prob > F = 0.0000		
				Root MSE = 11.936		

- In vielen ökonomischen Zeitreihen Saisonalität (Zyklen)
  - Arbeitslosigkeit im Winter höher
  - Mehr Autos, Fahrräder etc. verkauft im Sommer
  - Anstieg Ölpreis im Herbst
- Saisonalität in politikwissenschaftlichen Zeitreihen weniger klar, aber
  - Midterm-Effekt in amerikanischen Wahlen
  - Beliebtheit der Regierungsparteien in Deutschland über BTW-Wahlzyklus
  - Häufung von Terrorakten zu Jahrestagen/Festen
- Ggf. z. B. durch Dummies modellieren, um Verzerrungen zu vermeiden

## Saisonalität: CPI und Approval

```
. reg cpi _I* date
```

Source	SS	df	MS			
Model	73783.4552	4	18445.8638	Number of obs =	148	
Residual	20799.5849	143	145.451643	F( 4, 143) =	126.82	
Total	94583.0401	147	643.422041	Prob > F =	0.0000	
				R-squared =	0.7801	
				Adj R-squared =	0.7739	
				Root MSE =	12.06	

cpi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
_Iquarter_2	.1125771	2.804067	0.04	0.968	-5.4302	5.655354
_Iquarter_3	.2359647	2.804355	0.08	0.933	-5.307382	5.779311
_Iquarter_4	.2431365	2.804835	0.09	0.931	-5.301159	5.787432
date	.5225581	.0232122	22.51	0.000	.4766747	.5684414
_cons	30.8954	2.086516	14.81	0.000	26.771	35.0198

```
. reg approval honeymoon date
```

Source	SS	df	MS			
Model	2688.40949	2	1344.20475	Number of obs =	148	
Residual	24762.6874	145	170.777155	F( 2, 145) =	7.87	
Total	27451.0969	147	186.742156	Prob > F =	0.0006	
				R-squared =	0.0979	
				Adj R-squared =	0.0855	
				Root MSE =	13.068	

approval	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
honeymoon	12.58499	5.061734	2.49	0.014	2.580679	22.5893
date	-.0791643	.0251498	-3.15	0.002	-.1288719	-.0294567
_cons	56.42957	1.324231	42.61	0.000	53.81228	59.04686

## Recap: Schätzverfahren

Welche Eigenschaften sollen Schätzverfahren haben?

- ① Möglichst unverzerrt (wenig bias)
  - ② Möglichst wenig Varianz der Schätzungen (relative Effizienz)
  - ③ Konsistenz (durch größere Fallzahlen beliebig kleine Abweichungen von wahren Werten)
- Wiederholte Stichprobenziehung (Umfang  $n$ ) → zufällige Verteilung von Schätzungen
  - Mittelwert der Schätzungen stimmt mit wahren Parameter in Grundgesamtheit überein
  - Systematische Abweichung: bias (Mittelwert  $\neq$  wahrer Parameter)
  - Wiederholte Stichprobenziehung (Umfang  $n$ ) → zufällige Verteilung von Schätzungen
  - Unter verschiedenen Schätzverfahren das mit der geringsten

- (Relativ) effizientes Verfahren

## Recap: Voraussetzungen für OLS

- Wann ist OLS unverzerrt, effizient und konsistent?
- (Gauß-Markov-Bedingungen):
  - ①  $y$  intervallskaliert und unbeschränkt, Varianz aller  $x$ , keine perfekte Multikollinearität
  - ② Keine Autokorrelation der zufälligen Einflüsse  $\epsilon$ , d. h. für zwei Beobachtungen  $i$  und  $h$  sind  $\epsilon_i$  und  $\epsilon_h$  bei wiederholter Stichprobenziehung unkorreliert
  - ③ Kein Zusammenhang zwischen  $\epsilon$  und  $x$ -Variablen, d. h.
    - ① Konditionaler Mittelwert von  $\epsilon = 0$  für alle Werte von  $x$
    - ② Konstante konditionale Varianz von  $\epsilon$  (Homoskedastizität) für alle Werte von  $x$
  - ④ (Normalverteilung von  $\epsilon$  – wird nur für Berechnung Standardfehler gebraucht, kein Problem bei großen Stichproben)
- OLS = BLUE: Best Linear Unbiased Estimator

## Probleme mit Zeitreihendaten

- Grundproblem: Fälle sind nicht zufällig ausgewählt, sondern zeitlich strukturiert
- Fiktives Beispiel: nationale Bildungsausgaben ( $x$ ) und nationaler Schulleistungstests ( $y$ ), jährliche Messung
- Typische Probleme:
  - ① (Zusammenhang zwischen Variablen bzw. Verteilung einer Variablen ändert sich über die Zeit ((Non-)Stationarität))
  - ② (Extreme Autokorrelation einer  $x$ -Variablen)
  - ③ Heteroskedastizität (Varianz von  $\epsilon$  nicht konstant)
  - ④ Abhängigkeit zwischen  $x$  und  $\epsilon$
  - ⑤ Autokorrelation von  $\epsilon$

## Heteroskedastizität

- Höhere Bildungsausgaben
  - Bessere Testergebnisse
  - Tests werden sorgfältiger durchgeführt → weniger zufällige Varianz
- Abhängigkeit der *Varianz* von  $\epsilon$  vom Niveau von  $x$
- Standardfehler zu optimistisch
- Mögliche Lösungen
  - Qualität der Testdurchführung direkt messen (möglicherweise aber hoch mit  $x$  korreliert)
  - „Robuste“ Standardfehler berechnen lassen

## Endogenität

- Idealerweise: Alle Werte von  $x$  komplett von aktuellen, vergangenen, zukünftigen Werten von  $\epsilon$  unabhängig (strikt exogen)
- D. h. Gedankenexperiment:
  - DGP würde sehr häufig wiederholt
  - Werte für z. B.  $x_t$  und  $\epsilon_{t-1}$  würden notiert
  - Korrelation zwischen beiden?
- Endogenität im Beispiel leicht möglich
  - Zufälliger negativer  $\epsilon_{t-1}$  (z. B. Fußball-WM 2014) beeinflusst Leistungen im Test 2014 ( $y_{t-1}$ )
  - Panische Erhöhung der Bildungsausgaben in 2015 ( $x_t$ )
  - Zusammenhang zwischen  $x$  und  $\epsilon$ ;  $x$  ist nicht exogen
- Bias, aber möglicherweise noch konsistent  $\rightarrow$  besondere Modelle

Autokorrelation von  $\epsilon$ 

- In Zeitreihen sind die Werte von  $\epsilon$  häufig (positiv) mit sich selbst korreliert
  - Neue Show mit G. Jauch führt zu besonderen Anstrengungen zum Zeitpunkt  $t \rightarrow \epsilon_t > 0, y_t$  ungewöhnlich hoch
  - Ein Jahr später: positiver Effekt der Show abgeschwächt, aber noch vorhanden  $\rightarrow \epsilon_t > \epsilon_{t+1} > 0$
  - Allgemein:  $\epsilon_t = \rho\epsilon_{t-1} + u_t$  mit  $|\rho| < 1$
- OLS immer noch konsistent und unverzerrt, aber nicht mehr effizient
- $R^2$  zu hoch, Signifikanztests nicht mehr gültig
- Statistischer Test (Durbin-Watson-Test und Alternativen)
- Ggf. besonderes Modell schätzen



# Presidential Approval

```
. reg approval honeymoon date
```

Source	SS	df	MS			
Model	2688.40949	2	1344.20475	Number of obs =	148	
Residual	24762.6874	145	170.777155	F( 2, 145) =	7.87	
				Prob > F	= 0.0006	
				R-squared	= 0.0979	
				Adj R-squared	= 0.0855	
Total	27451.0969	147	186.742156	Root MSE	= 13.068	

approval	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
honeymoon	12.58499	5.061734	2.49	0.014	2.580679	22.5893
date	-.0791643	.0251498	-3.15	0.002	-.1288719	-.0294567
_cons	56.42957	1.324231	42.61	0.000	53.81228	59.04686

```
. estat durбина
```

Durbin's alternative test for autocorrelation

lags(p)	chi2	df	Prob > chi2
1	639.669	1	0.0000

H0: no serial correlation

```
. prais approval honeymoon date, robust
```

Iteration 0: rho = 0.0000  
 Iteration 1: rho = 0.8928  
 Iteration 2: rho = 0.8933  
 Iteration 3: rho = 0.8933  
 Iteration 4: rho = 0.8933

Prais-Winsten AR(1) regression -- iterated estimates

```
Linear regression
```

Statistik II		Zeitreihen (26/31)		
		Number of obs =	148	
		prob > r	= 0.0000	
		R-squared	= 0.3154	
		Root MSE	= 5.9474	

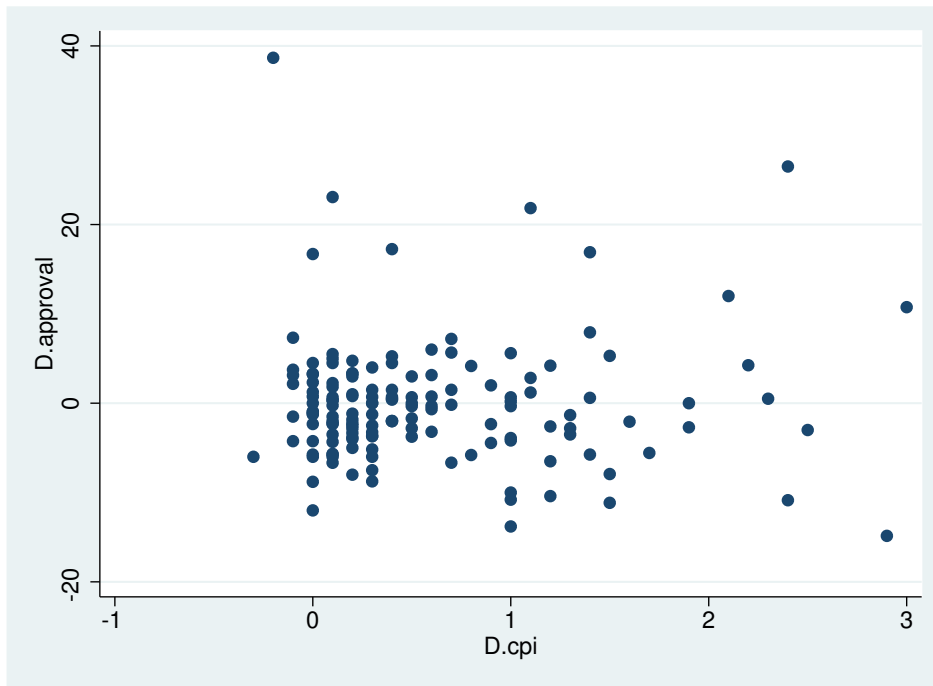
## (1.) Differenzen

- Zeitreihen mit starken Trends machen Probleme bei Schätzung
- Oft ist nicht Niveau, sondern Veränderung interessant
- Differenzen (gegenüber Vorperiode) für abhängige und/oder unabhängige Variablen
- In Stata: D.-Operator

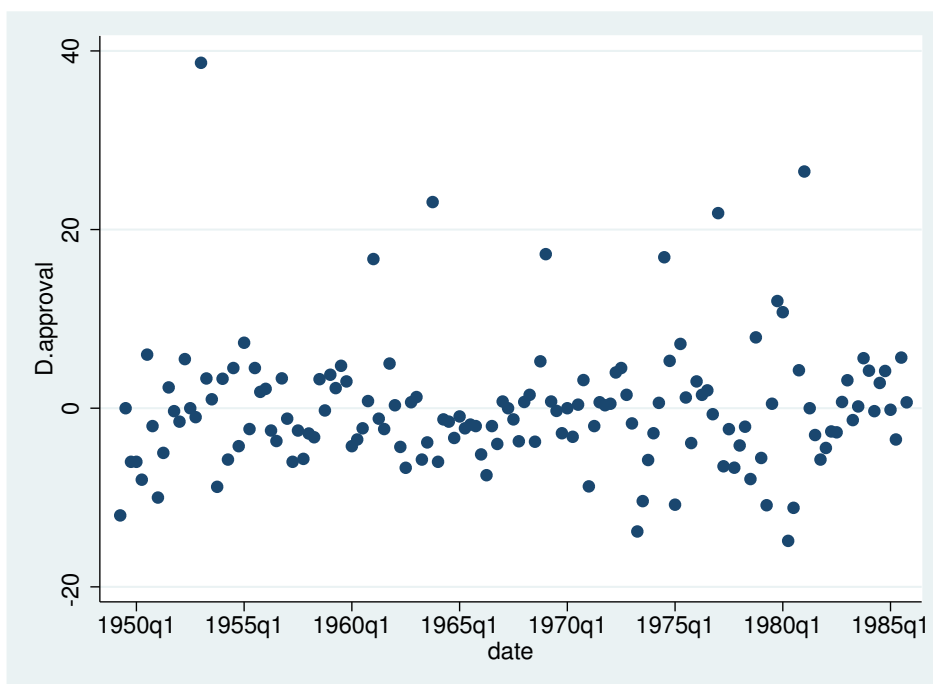
```
. list date approval D.approval cpi D.cpi in 1/6, clean
```

	date	approval	D. approval	cpi	D. cpi
1.	1949q1	69	.	23.8	.
2.	1949q2	57	-12	23.8	0
3.	1949q3	57	0	23.8	0
4.	1949q4	51	-6	23.8	0
5.	1950q1	45	-6	23.5	-.2999992
6.	1950q2	37	-8	23.7	.2000008

## Veränderung approval / Veränderung CPI



## Veränderung approval / Zeit



## Finales (?) Modell

```
. prais D.approval honeymoon D.cpi date

Iteration 0: rho = 0.0000
Iteration 1: rho = 0.0628
Iteration 2: rho = 0.0641
Iteration 3: rho = 0.0642
Iteration 4: rho = 0.0642

Prais-Winsten AR(1) regression -- iterated estimates

-----+-----+-----+-----+-----+-----+-----+-----+
Source |         SS      df      MS              Number of obs =   147
-----+-----+-----+-----+-----+-----+-----+-----+
Model | 3936.77694      3 1312.25898              F( 3, 143) = 55.48
Residual | 3382.07009    143 23.6508398              Prob > F      = 0.0000
-----+-----+-----+-----+-----+-----+-----+-----+
Total | 7318.84703    146 50.1290893              R-squared     = 0.5379
                                           Adj R-squared = 0.5282
                                           Root MSE    = 4.8632

-----+-----+-----+-----+-----+-----+-----+-----+
D.approval |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----+-----+-----+-----+-----+-----+-----+
honeymoon |      24.13866   1.876509     12.86   0.000     20.42938     27.84794
      cpi |
      D1. |     -1.264644   .8988159     -1.41   0.162     -3.041326     .5120389
      date |      .0182705   .0142053      1.29   0.200     -.0098091     .04635
      _cons |     -1.00744   .5734237     -1.76   0.081     -2.140922     .1260424
-----+-----+-----+-----+-----+-----+-----+-----+
rho |      .0641721

Durbin-Watson statistic (original)    1.843234
Durbin-Watson statistic (transformed) 1.963893
```

## Zusammenfassung

- Zeitreihen sind anders:
  - Kein Stichprobenziehung
  - Nur ein Objekt wird beobachtet
  - Viele Informationen über dieses Objekt
- Zeitreihen sind schwach
  - Beobachtungen sind voneinander abhängig
  - Effektive Fallzahl kleiner als  $n$
- Gleichzeitig informativer als Querschnittsdaten: Dynamik
- *Sehr* sorgfältige Analyse notwendig