Lecture 9:

Gentle introduction to regression with time series

Jan Zouhar

Introductory Econometrics

From random variables to random processes

- in cross-sectional regression, we were making inferences about the whole population based on a small sample
- a crucial assumption: random sampling
  - the bridge between population characteristics (distribution of wages in a country) and the probabilistic machinery of random variables (distribution of a wage of a randomly drawn person)
- unfortunately, with time series, random sampling makes no sense:

year	GDP	inflation	unemp
2004	1,957.6	2.6	5.4
2005	2,035.4	2.8	4.5
…	…	…	…

- what would the underlying population be?

From random variables to random processes (cont'd)

- random sampling makes the characteristics of different individuals *independent*

- it is difficult to imagine that GDP in 2004 is independent of that in 2005

- therefore, we will have to switch to a more advanced theoretical vehicle: *random processes*

- those who had taken courses in random processes would tell you it was difficult

→ we will omit many mathematical details, and focus on the intuition

New issues with time series

- good news: everything we learnt with cross-sectional data will be used in time-series analysis, too
- bad news: many new pitfalls that can spoil the analysis
  1. **trends and seasonality**: can result in *spurious regression* (see next slide)
  2. **lags in economic behaviour**: government's expenditure cuts will slowly percolate through the economy → lagged effect (the effect of today's cuts will spread over quarters or even years)
  3. **persistence in time series**: governments expenditure itself cannot change too dramatically from one year to another
     - most real-life time series persistent, but the degree differs
     - strong persistence of time series can again produce spurious regression (*stationarity*, *unit-root* issues)
     - weak persistence problematic only if applies to u (*serial correlation*, or *autocorrelation* of u)

Spurious regression problem

- both y and x both exhibit a monotonous trend, we will find a relationship even though they have nothing in common

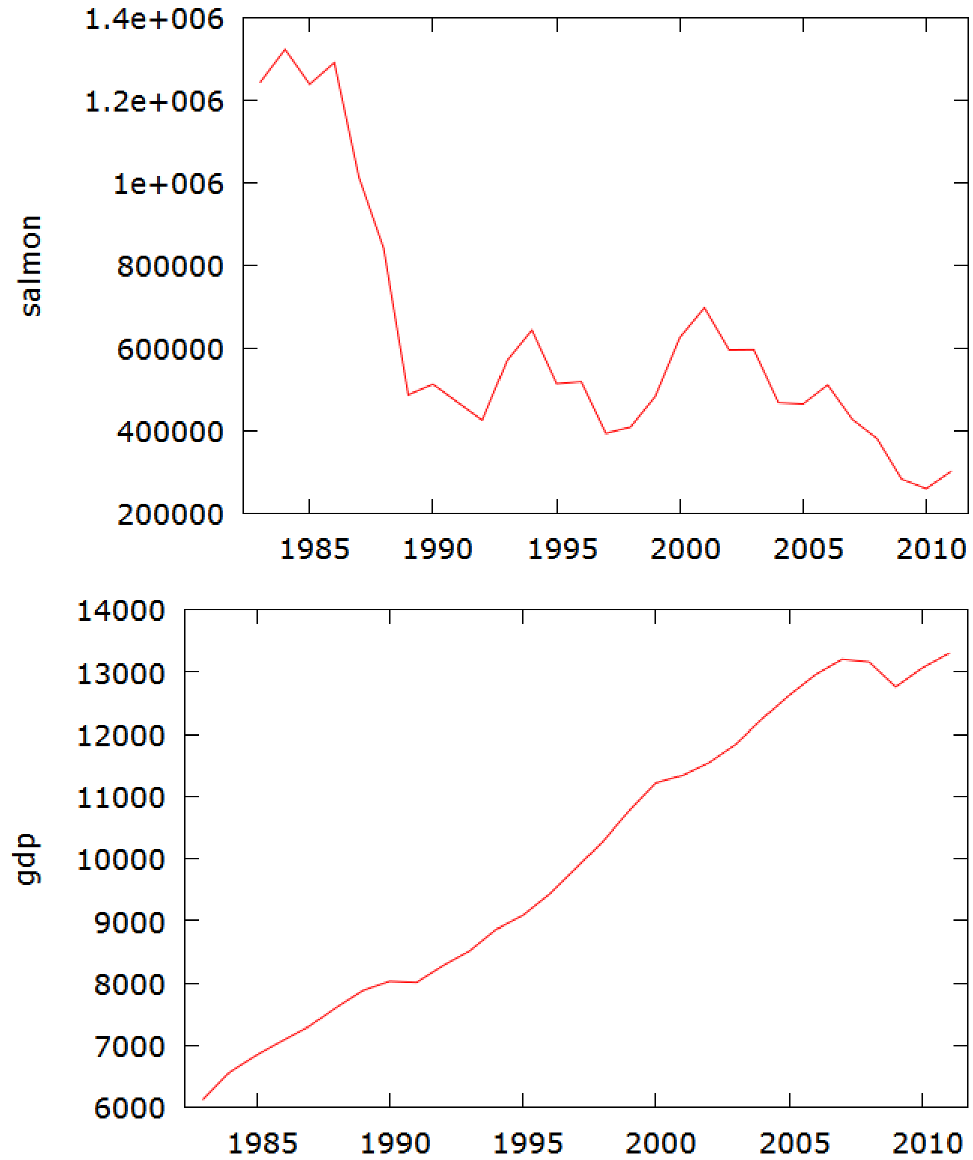
Example: Norwegian salmon production vs GDP in the U.S.

- the data in `salmon.gdt` contain two annual time series (1983–2011)
  - annual *salmon* production in Norway
  - *GDP* in the U.S. (bln. of 2005 dollars)
- do you think that there is a strong causal relationship?
- estimated equation in Gretl:

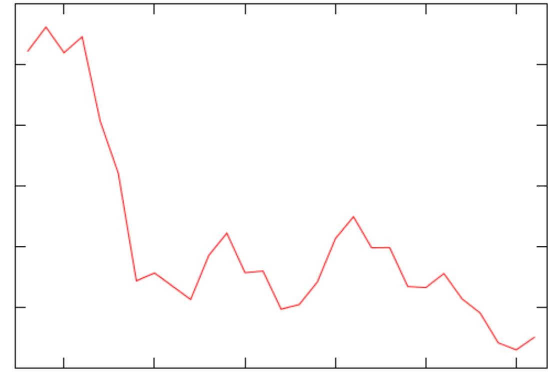
```
^gdp = 1.34e+04 - 0.00551*salmon
      (713)       (0.00103)

T = 29, R-squared = 0.514
(standard errors in parentheses)
```

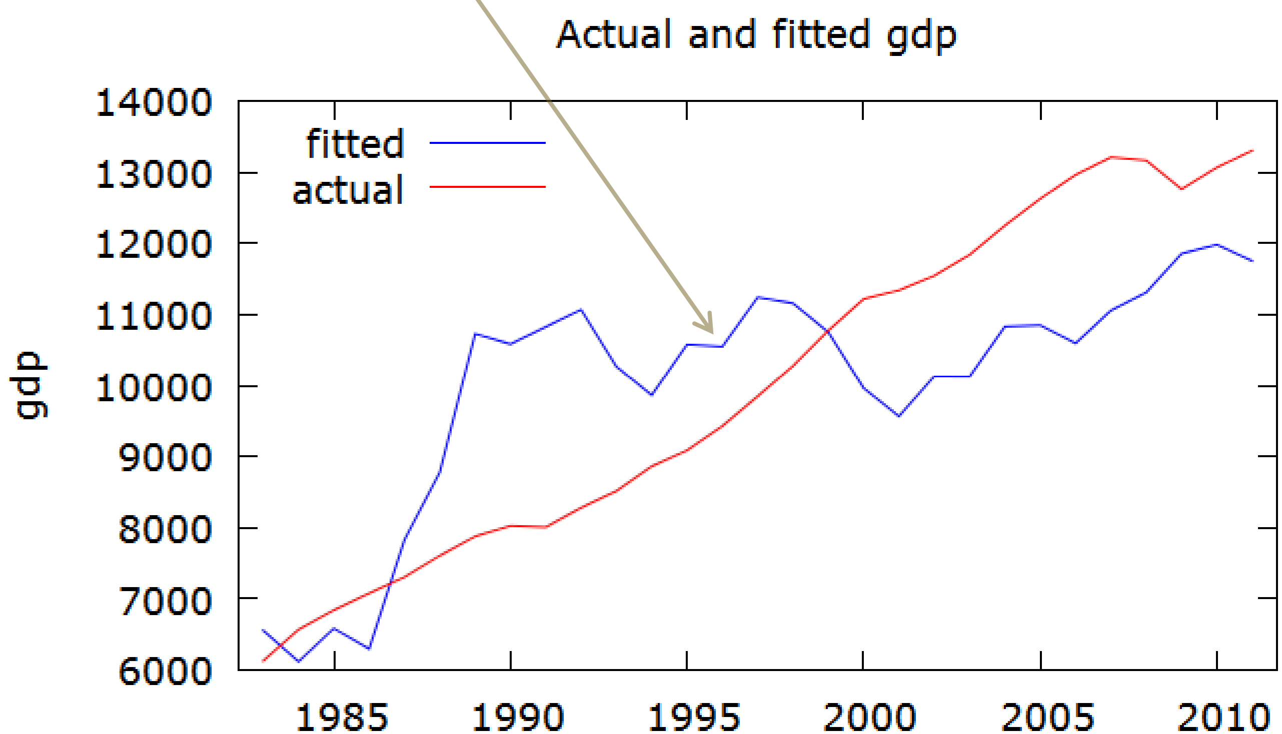
- Quizz: is *salmon* significant at the 5 % level? And how about 1 %?

Let's look at the time series first:



Fitted values are calculated as: 13,414 – 0.00551 ×





Accounting for trends in the regressions

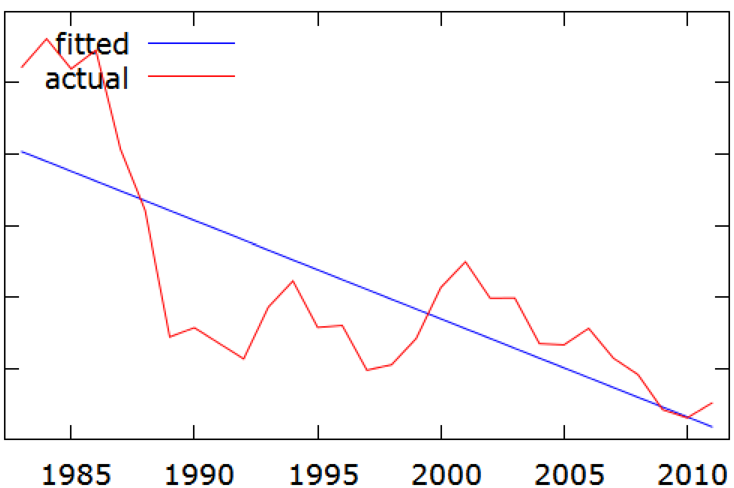
Option 1

- *detrend* all time series, i.e. create new variables where the linear trends have been subtracted
- two steps involved:

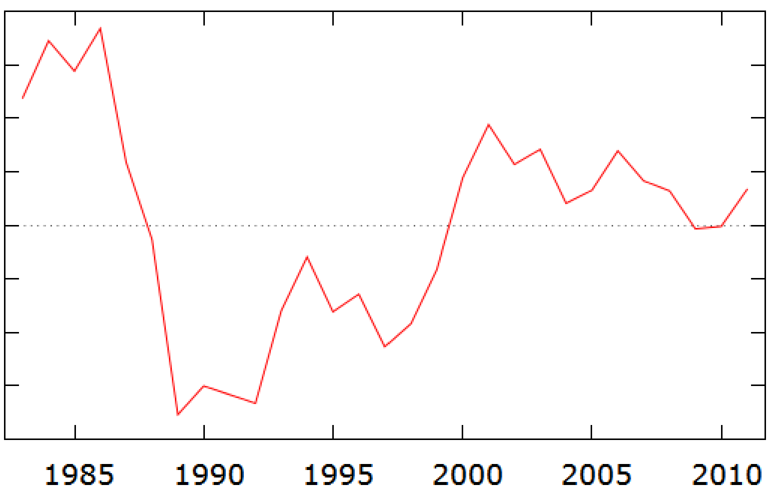
  (1) regress x_t on t (time, values 1,2,…,n)

  (2) save residuals (this is the detrended x_t)

salmon series and its linear trend



detrended *salmon* series



Accounting for trends in the regressions (cont'd)

Option 2: add variable *t* (time) to the estimated equation

Option 1: results

```
^gdp_detrended = 1.46e-013 + 0.000701*salmon_detrended
                 (52.7)      (0.000269)

T = 29, R-squared = 0.201
(standard errors in parentheses)
```

Option 2: results

```
^gdp = 5.14e+03 + 0.000701*salmon + 295*time
       (304)      (0.000274)         (9.90)

T = 29, R-squared = 0.986
(standard errors in parentheses)
```

- *salmon* coefficient identical, std. errors nearly identical → can use both
- R-squareds different, but most variation explained by *time* in Option 2 → use detrended dependent variable for the R-squared!

Frisch-Waugh-Lovell theorem

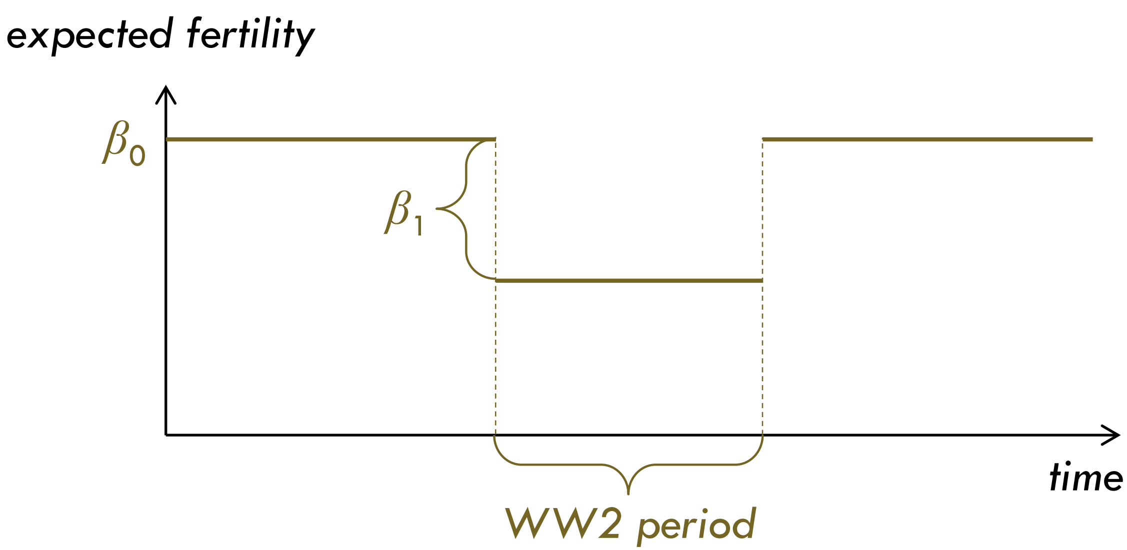
	Dependent variable:			
	gdp	gdp_detrended		gdp
	(1)	(2)	(3)	(4)
year	294.72010***		19.29931*	
	(9.90449)		(9.90449)	
salmon	0.00070**		0.00070**	
	(0.00027)		(0.00027)	
salmon_detrended		0.00070**		0.00070
		(0.00026)		(0.00990)
Constant	-578,999.30000***		-38,976.06000*	
	(19,909.28000)		(19,909.28000)	
Observations	29	29	29	29
R2	0.98613	0.20106	0.20106	0.00018

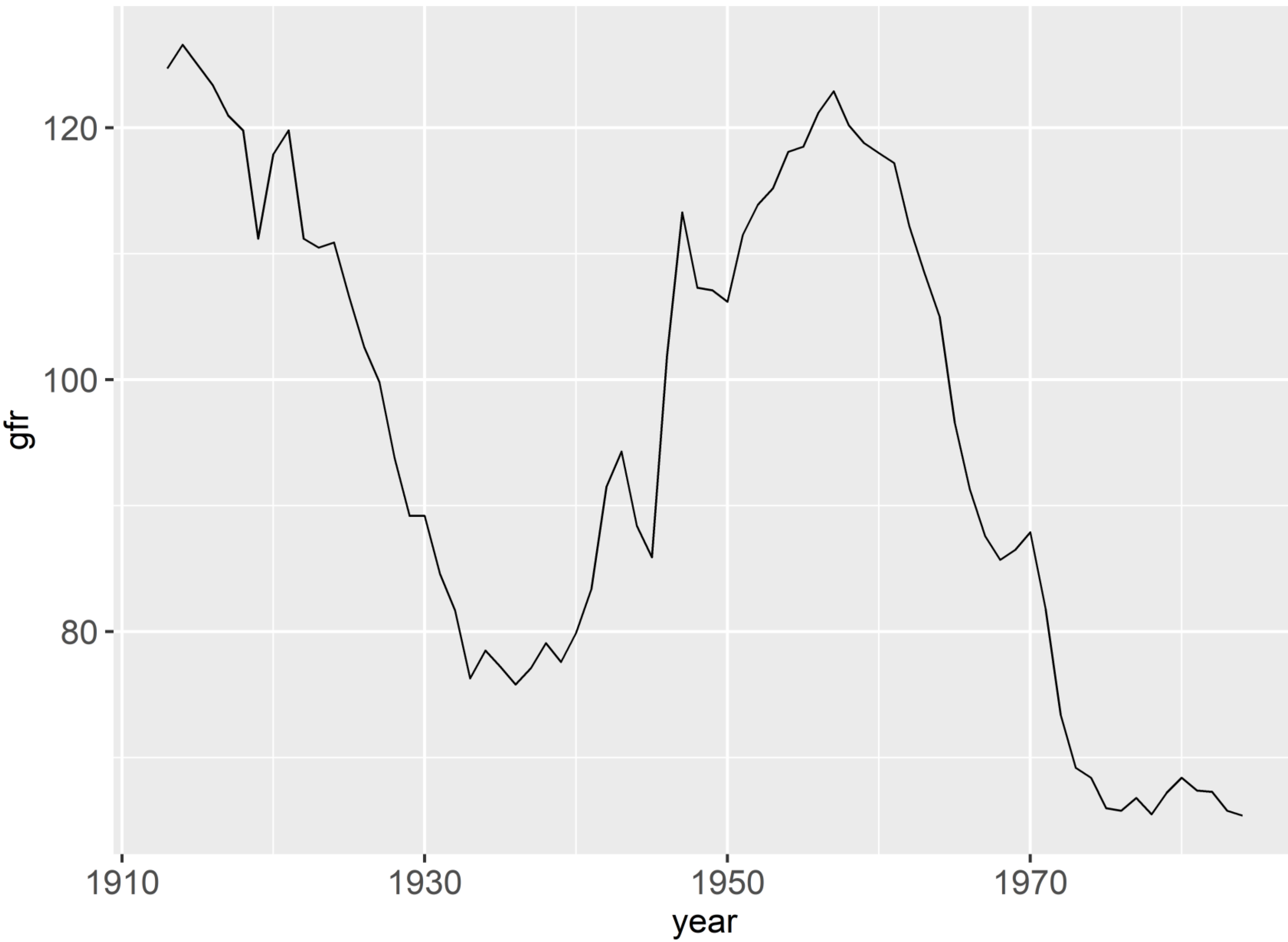
Using dummies to account for specific events

Example: fertility equation

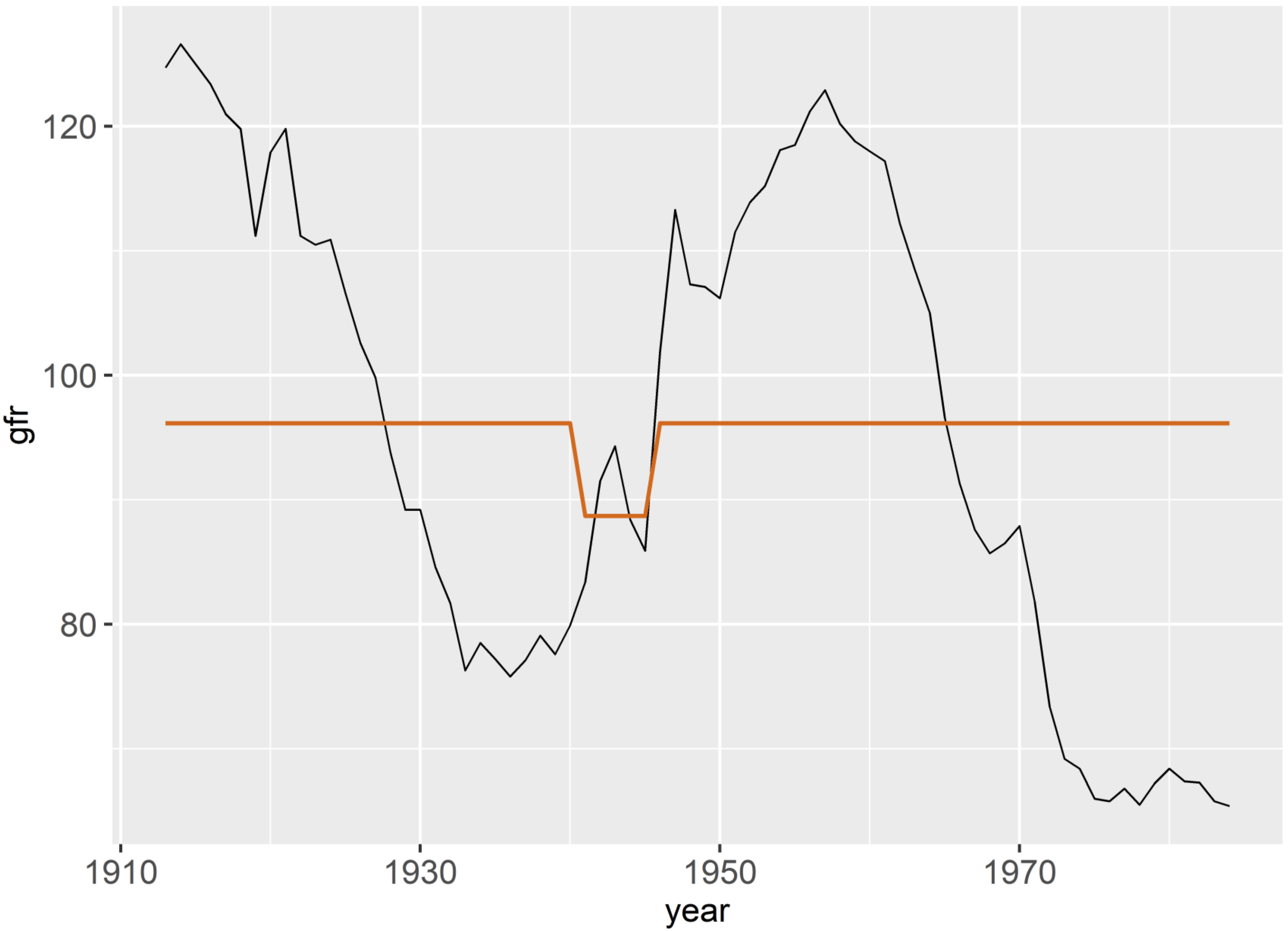
- frequency: annual data, 1913–1984
- *gfr* = the number of births per 1000 women aged 15–44
- *ww2* = 1 for years 1941–1945, = 0 otherwise

$$gfr_t = \beta_0 + \beta_1 ww2_t + u_t$$





120
100
80
gfr
1910
1930
1950
1970
year

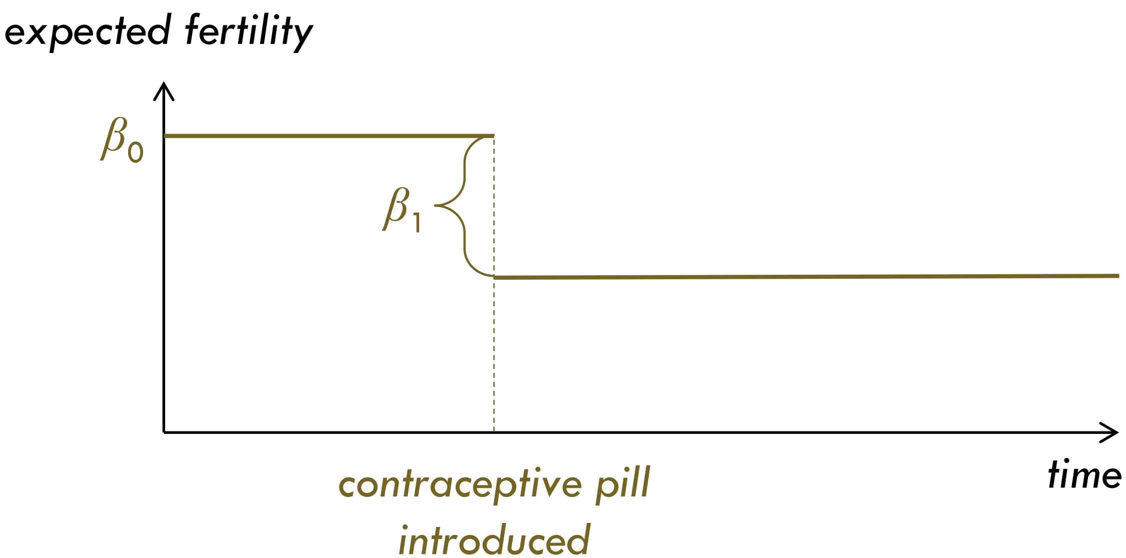


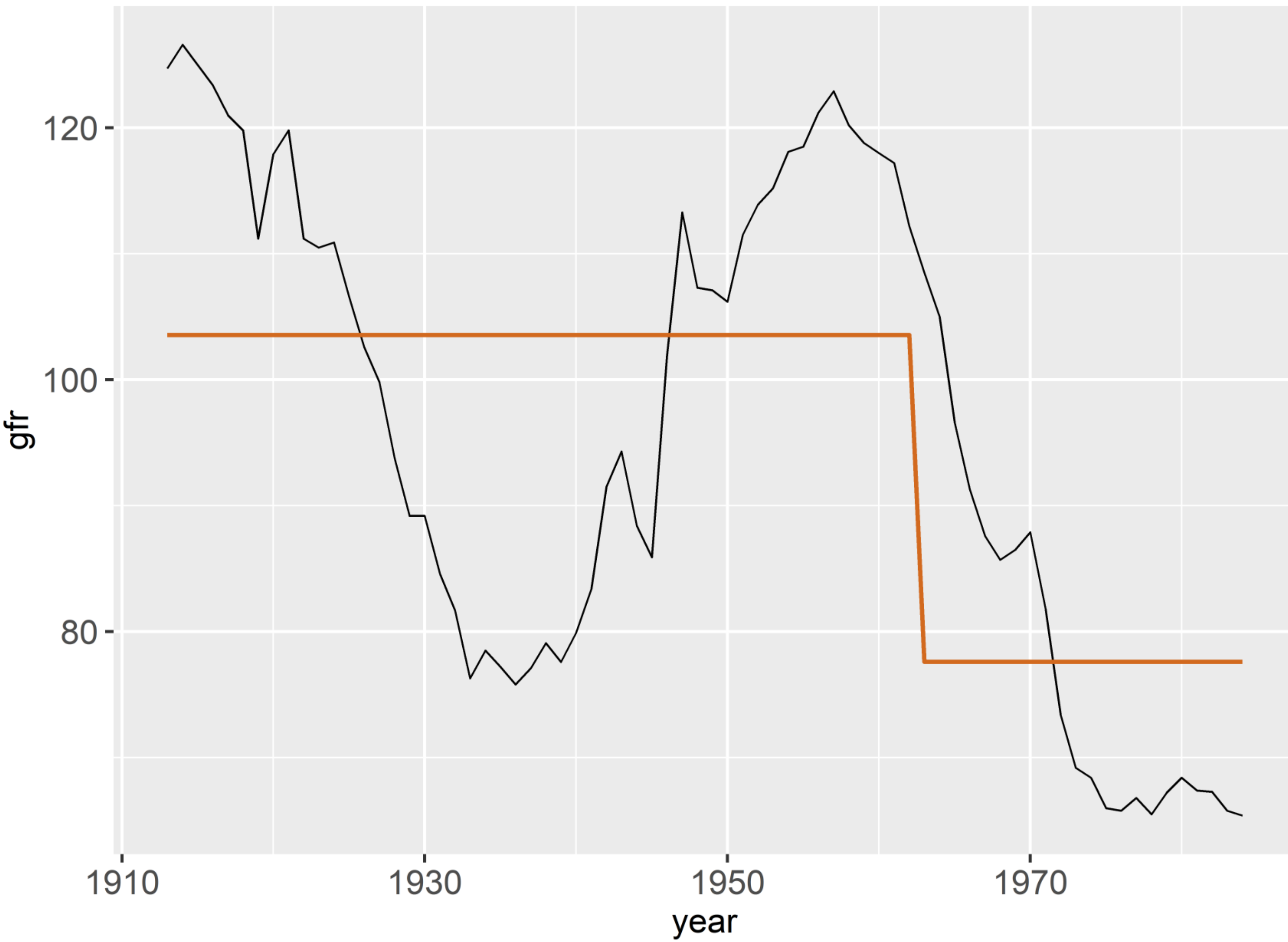
120
100
80
gfr
1910
1930
1950
1970
year

Example: fertility equation 2

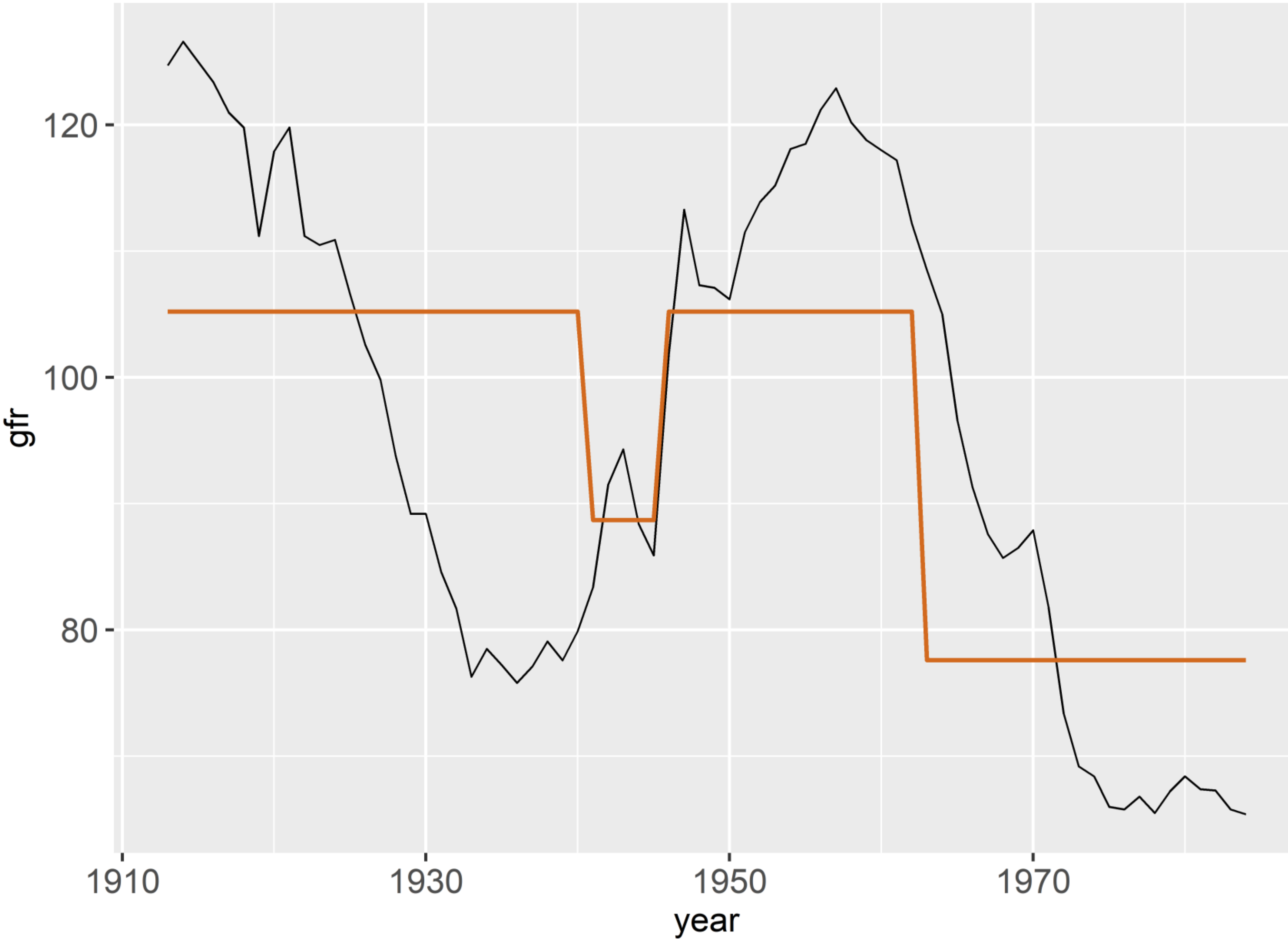
- $pill = 0$ before 1963, $= 1$ afterwards

$$gfr_t = \beta_0 + \beta_1 pill_t + u_t$$




120
100
80
gfr
1910
1930
1950
1970
year

$$gfr_t = \beta_0 + \beta_1 ww2_t + \beta_2\, pill_t + u_t$$

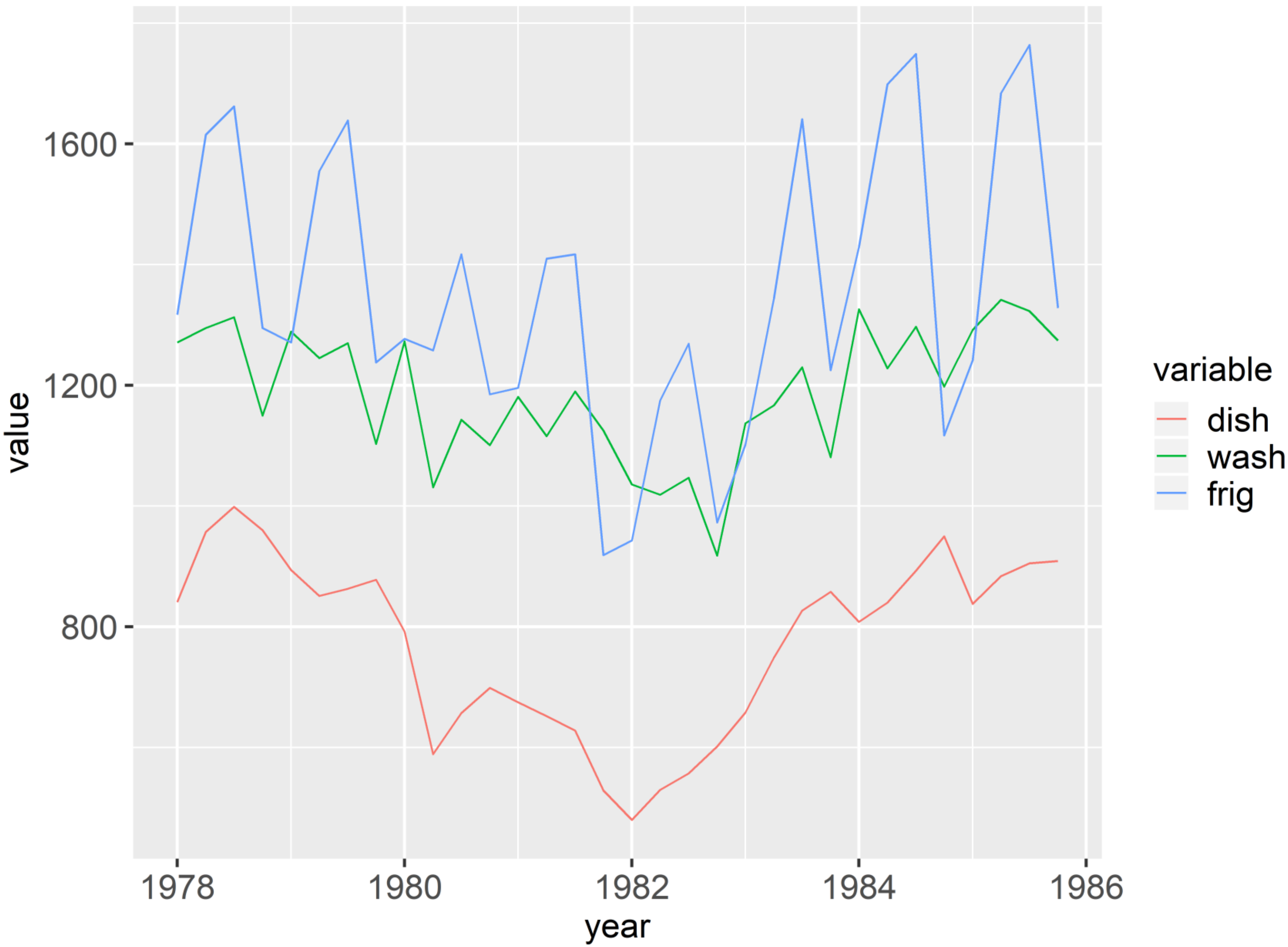


Seasonality

- seasonal patterns are noticeable with quarterly, monthly, or daily data
- note that many time series in the online databases are "seasonally adjusted", meaning that specialized algorithms have been used to even out the differences between seasons → these series can be used without further ado
- when using a seasonally unadjusted series, we can still use a simple fix that accounts for the seasonal variation: periodic dummies, i.e. dummy variables that identify individual periods

Example: durable goods

- open `durgoods.gdt` in Gretl
- change dataset structure to a quarterly time series (Data → Dataset structure)
- add periodic dummies (Add → Periodic dummies)
- this creates variables *dq1*,…, *dq4*
  (*dq1* stands for "dummy for quarter 1")


1600
1200
800
value
1978
1980
1982
1984
1986
year
variable
dish
wash
frig

values of the periodic dummies

gretl: edit data

dish 1978:1

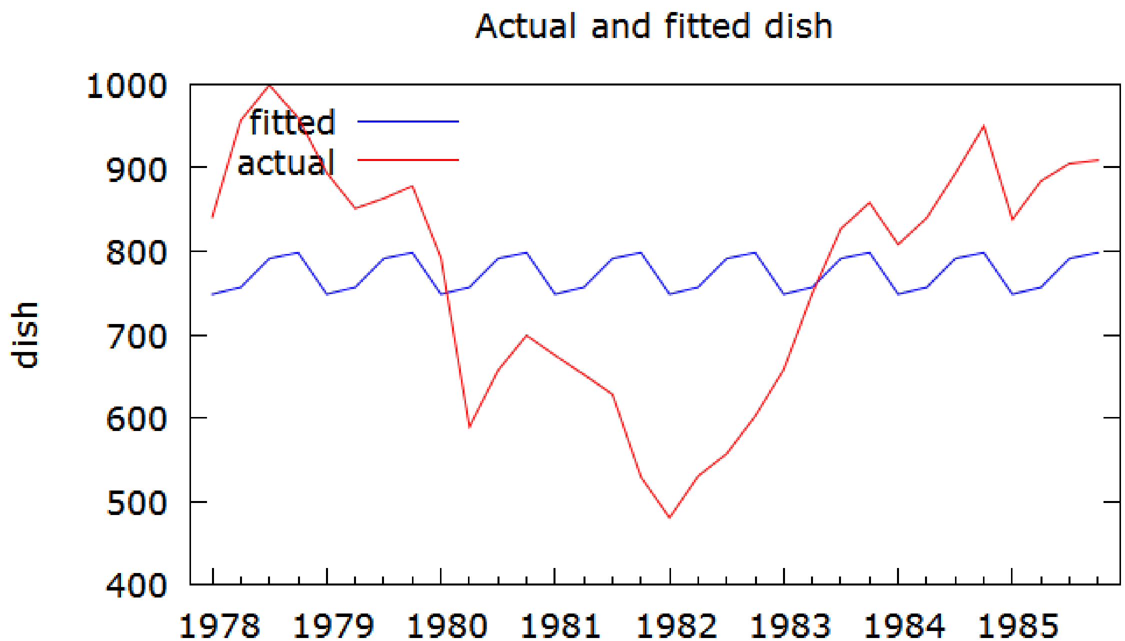
	dish	frig	wash	dur	dq1	dq2	dq3	dq4
1978:1	841	1317	1271	252.6	1	0	0	0
1978:2	957	1615	1295	272.4	0	1	0	0
1978:3	999	1662	1313	270.9	0	0	1	0
1978:4	960	1295	1150	273.9	0	0	0	1
1979:1	894	1271	1289	268.9	1	0	0	0
1979:2	851	1555	1245	262.9	0	1	0	0
1979:3	863	1639	1270	270.9	0	0	1	0
1979:4	878	1238	1103	263.4	0	0	0	1
1980:1	792	1277	1273	260.6	1	0	0	0
1980:2	589	1258	1031	231.9	0	1	0	0
1980:3	657	1417	1143	242.7	0	0	1	0
1980:4	699	1185	1101	248.6	0	0	0	1
1981:1	675	1196	1181	258.7	1	0	0	0
1981:2	652	1410	1116	248.4	0	1	0	0
1981:3	628	1417	1190	255.5	0	0	1	0
1981:4	529	919	1125	240.4	0	0	0	1

Seasonality (cont'd)

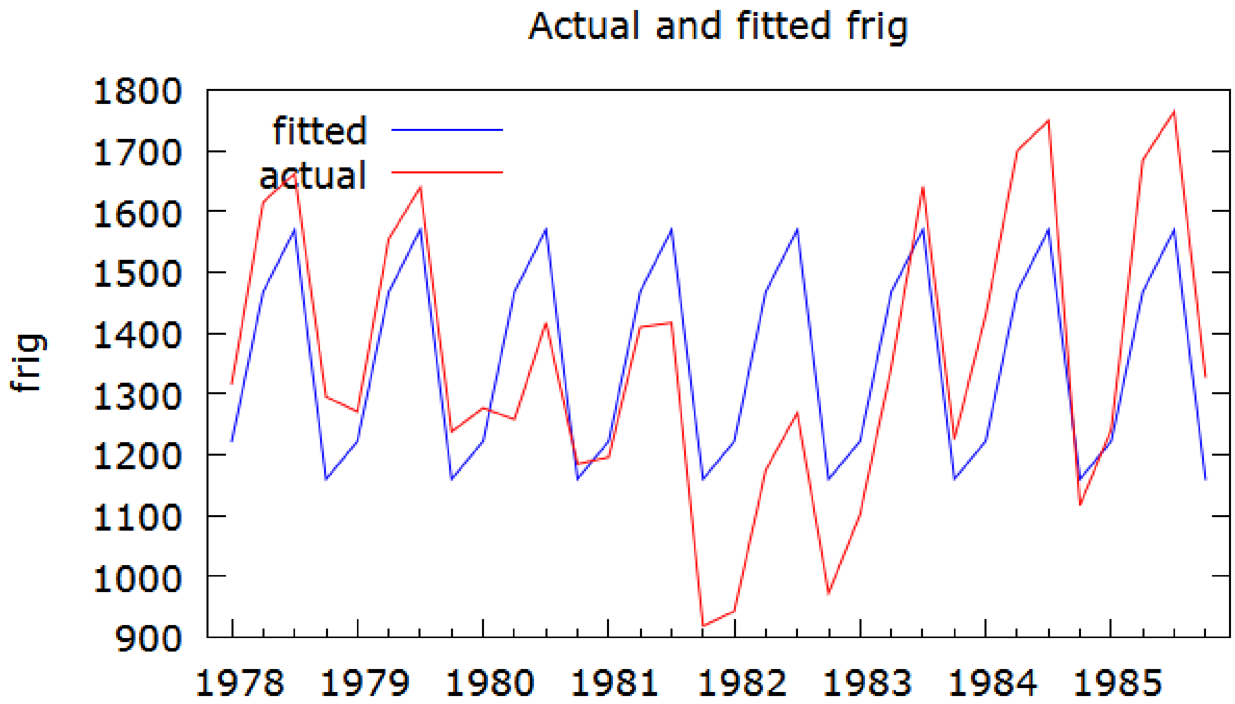
- to describe the seasonal pattern in dishwasher sales, run the regression

$$dish_t = \beta_0 + \beta_1 dq1_t + \beta_2 dq2_t + \beta_3 dq3_t + u_t$$

- the dishwasher time series and the fitted values are shown below, F-test for joint significance: p-value = 0.89 → no statistical evidence of seasonality



- with refrigerator series, that's a different story:



- joint significance: p-value = 0.000 079, strong evidence of seasonality (i.e. we reject the null of no seasonal pattern)

Seasonality (cont'd)

- interpretation: just as with other category dummies
- we omitted *dq4* → quarter 4 is the base period
- e.g., the coefficient on *dq1* tells us that in quarter 1, sales are higher by 62,125 than in quarter 4 (on average)

```
Model 2: OLS, using observations 1978:1-1985:4 (T = 32)
Dependent variable: frig

             coefficient   std. error   t-ratio    p-value
  --------------------------------------------------------
  const        1160.00       59.9904     19.34     9.81e-018 ***
  dq1            62.1250     84.8393      0.7323   0.4701
  dq2           307.500      84.8393      3.625    0.0011    ***
  dq3           409.750      84.8393      4.830    4.42e-05  ***

Mean dependent var   1354.844   S.D. dependent var   235.6719
Sum squared resid    806142.4   S.E. of regression   169.6785
R-squared            0.531797   Adjusted R-squared   0.481632
F(3, 28)             10.60102   P-value(F)           0.000079
...
```

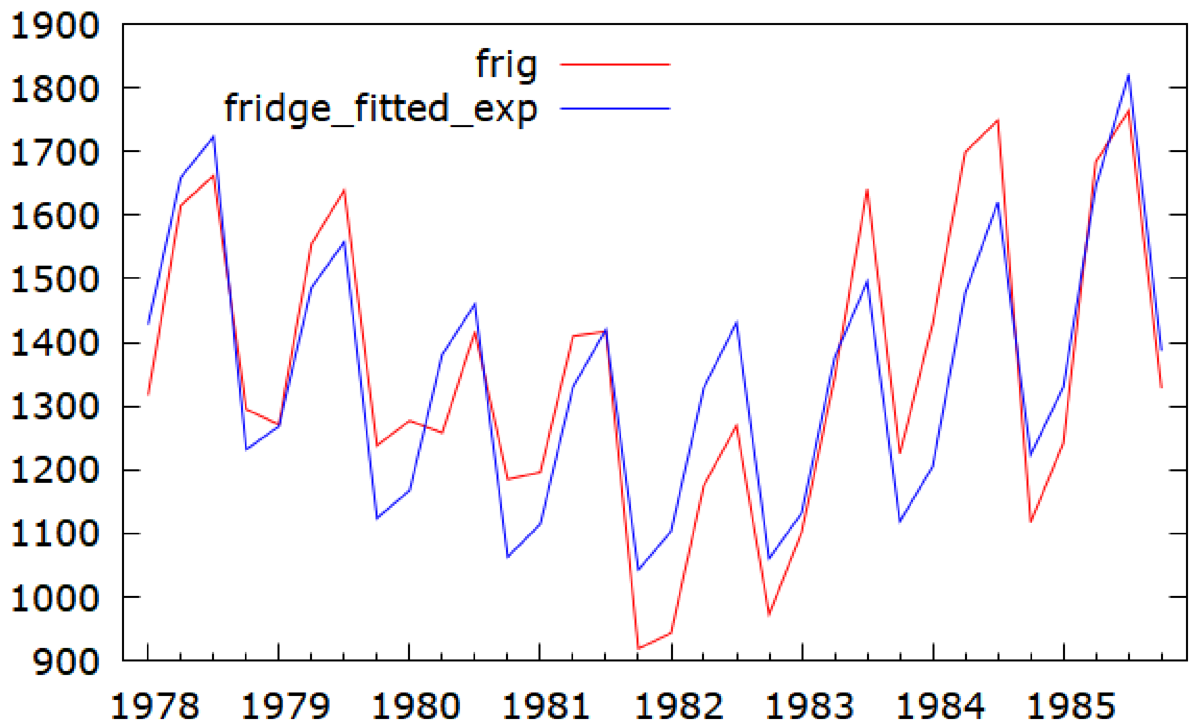
Seasonality (cont'd)

- **conclusion**: with seasonally unadjusted data, it makes sense to add both a time trend and periodic dummies in addition to your independent variables of interest
- note that this can be done also in case the dependent variable is logged, only the interpretation changes:

```
^l_frig = 7.05 +   0.0530*dq1 + 0.234*dq2 + 0.304*dq3
          (0.0458)(0.0647)     (0.0647)    (0.0647)

T = 32, R-squared = 0.517
(standard errors in parentheses)
```

- results imply that in quarter 1, sales increase by **5.3 %** compared with the baseline level of quarter 4

```
^l_frig = 7.30 + 0.183*dq2 + 0.252*dq3 - 0.0555*dq4 - 0.0365*time + 0.00113*sq_time
         (0.0590)(0.0470)   (0.0471)    (0.0473)     (0.00742)     (0.000218)

T = 32, R-squared = 0.764
```




```
 Null hypothesis: the regression parameters are zero for the variables
    dq2, dq3, dq4
  Test statistic: F(3, 26) = 19.323, p-value 8.49373e-007
```

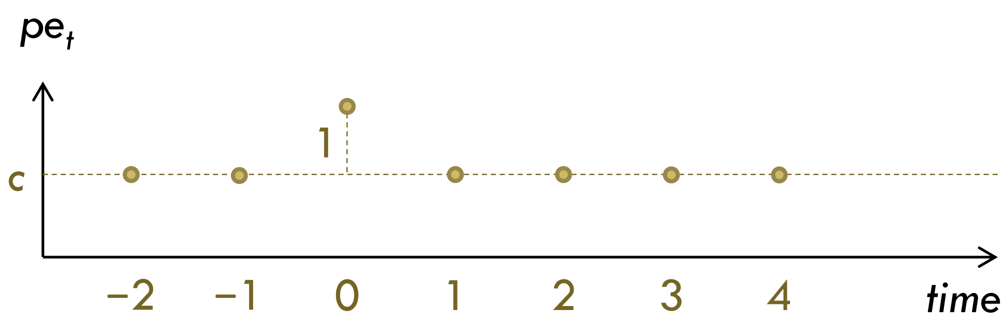
Finite distributed lag (FDL) model

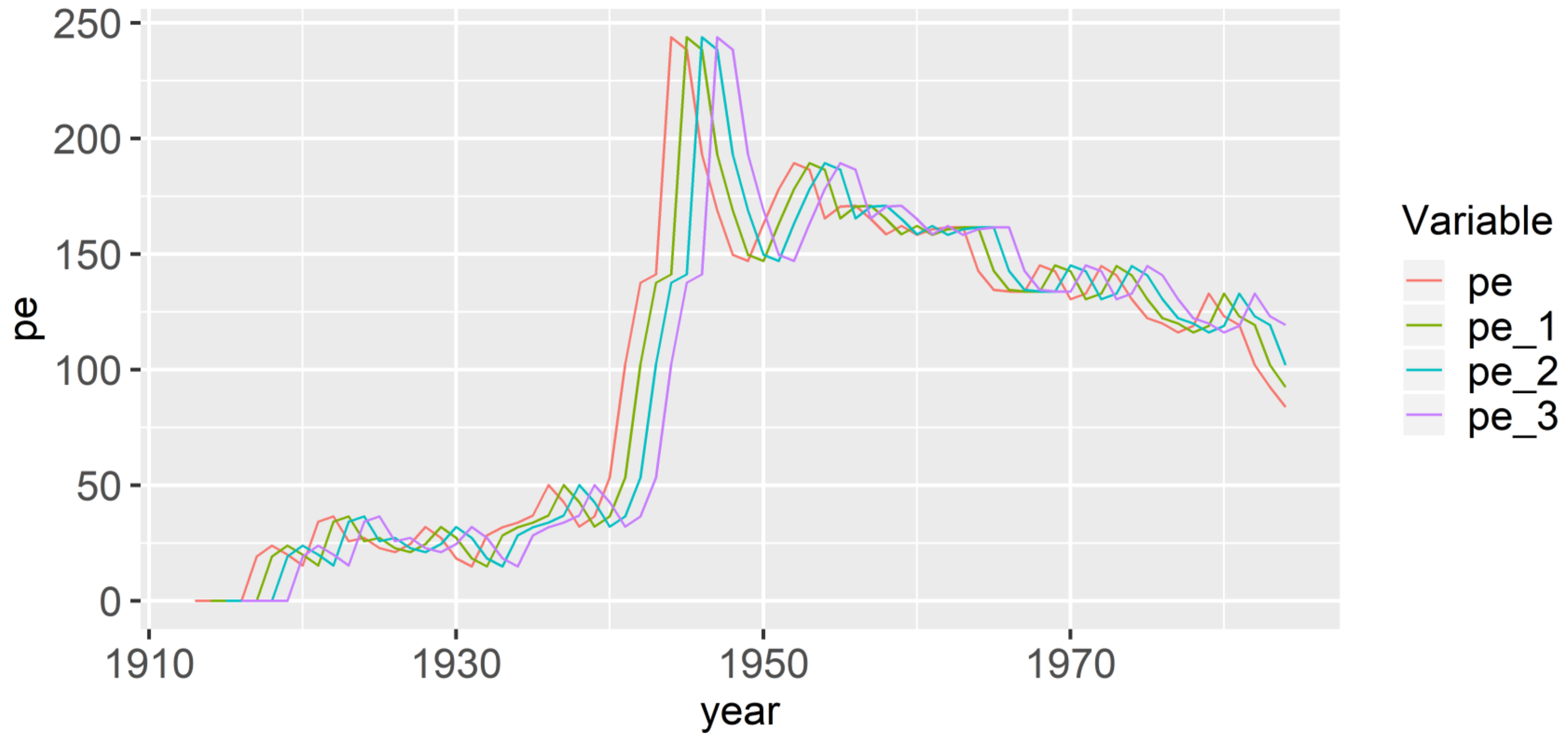
Example: fertility equation 3

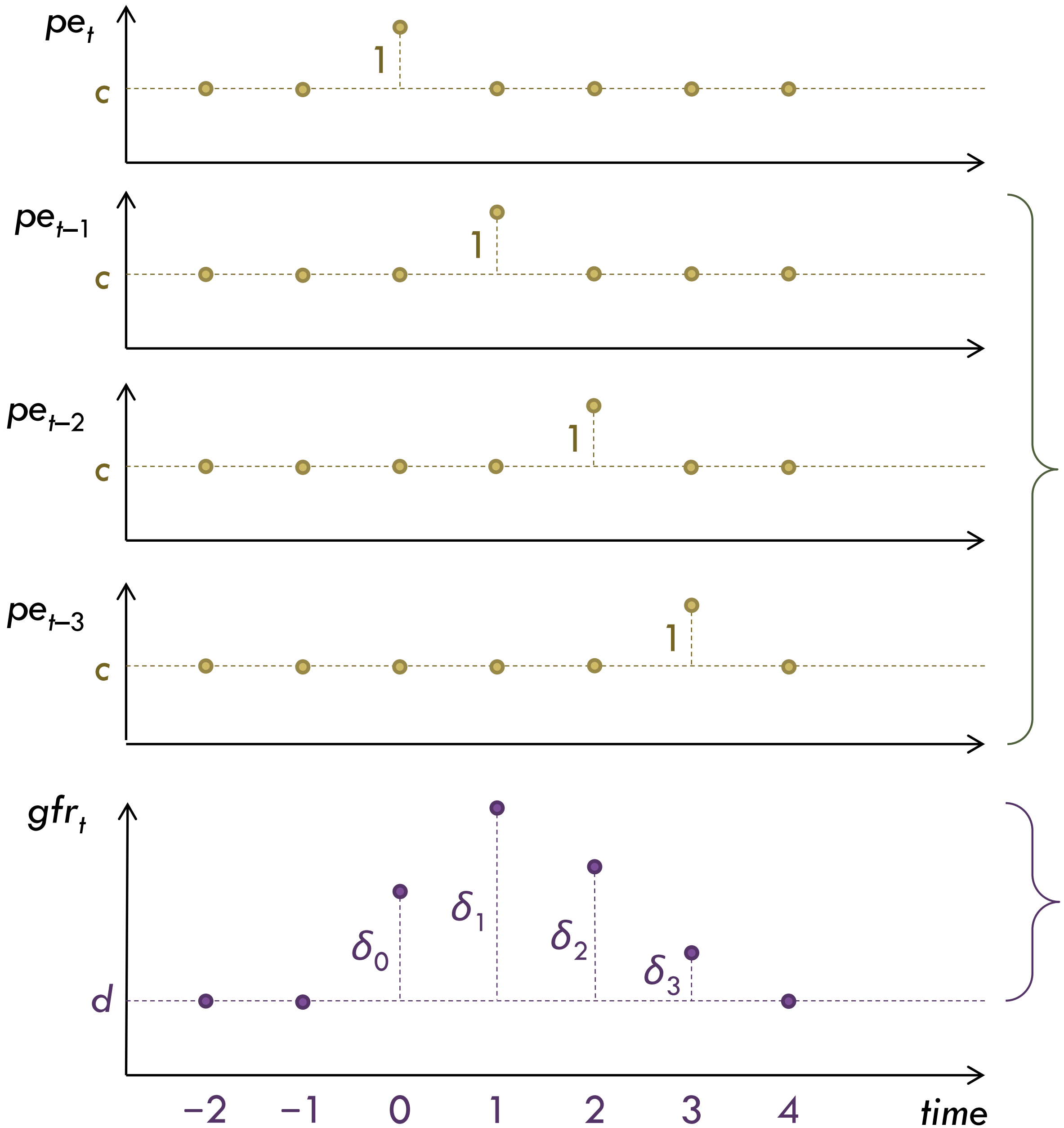
- pe = real dollar value of personal tax exemption

$$gfr_t = \beta_0 + \delta_0 pe_t + \delta_1 pe_{t-1} + \delta_2 pe_{t-2} + \delta_3 pe_{t-3} + u_t$$

- here, δ_0 is the **impact propensity** (= immediate effect) of a unit increase in pe
- the δ parameters capture the effect of a **temporary increase** in pe:
  - assume that pe equals c except for period 0, where it increases to $c + 1$:

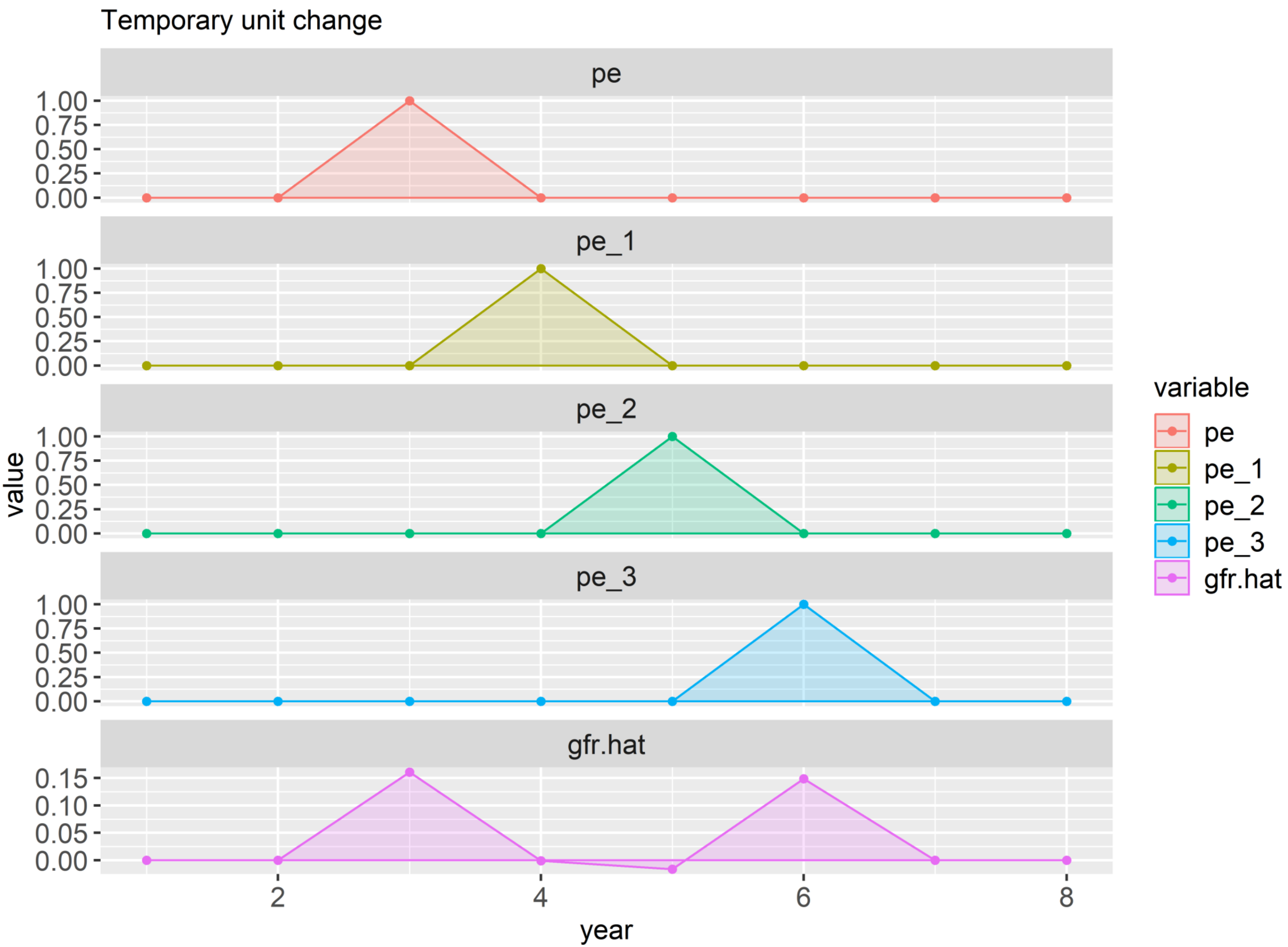



250
200
150
100
50
0
pe
1910
1930
1950
1970
year
Variable
pe
pe_1
pe_2
pe_3



a temporary change in *pe* occurs in subsequent periods in lagged versions of *pe*

the effect of a temporary change in *pe*, or the **lag distribution**

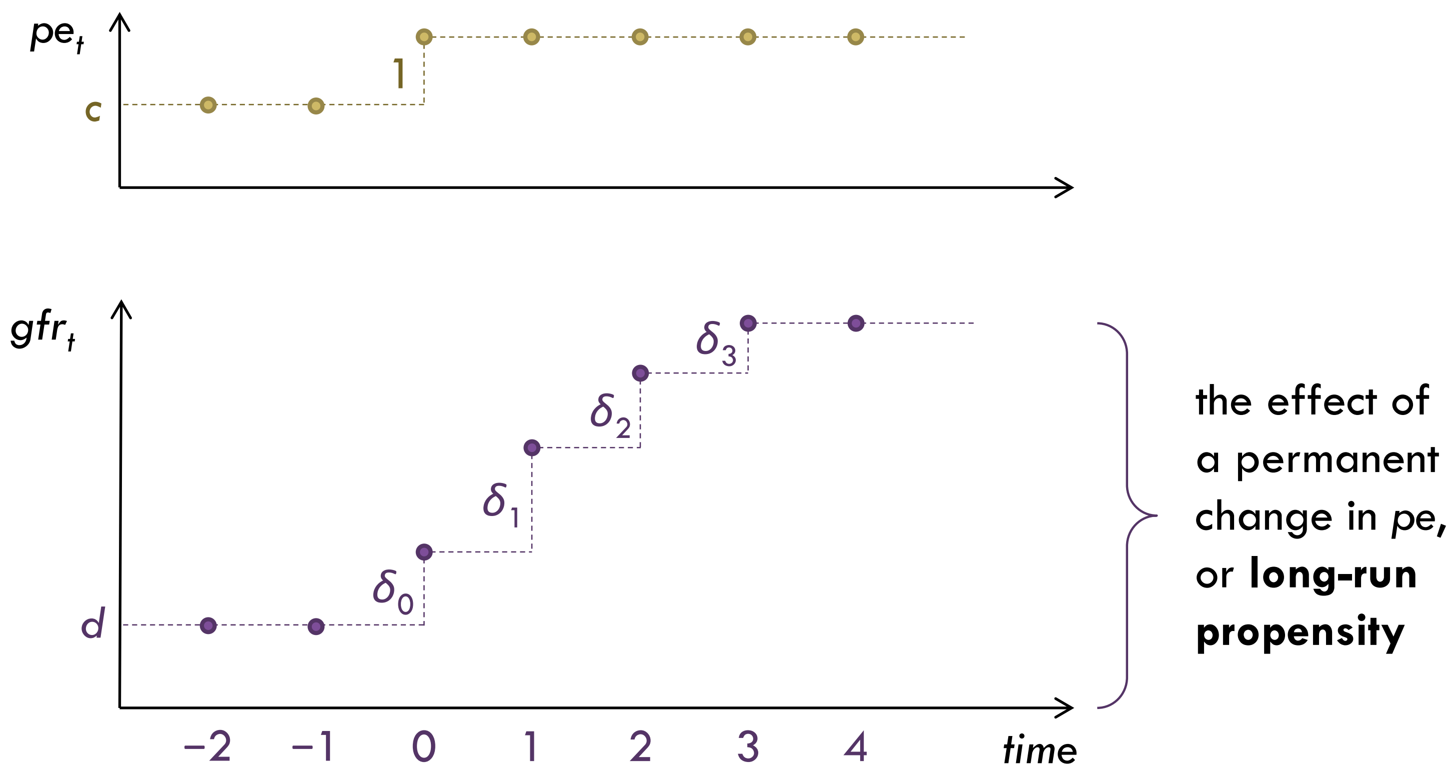


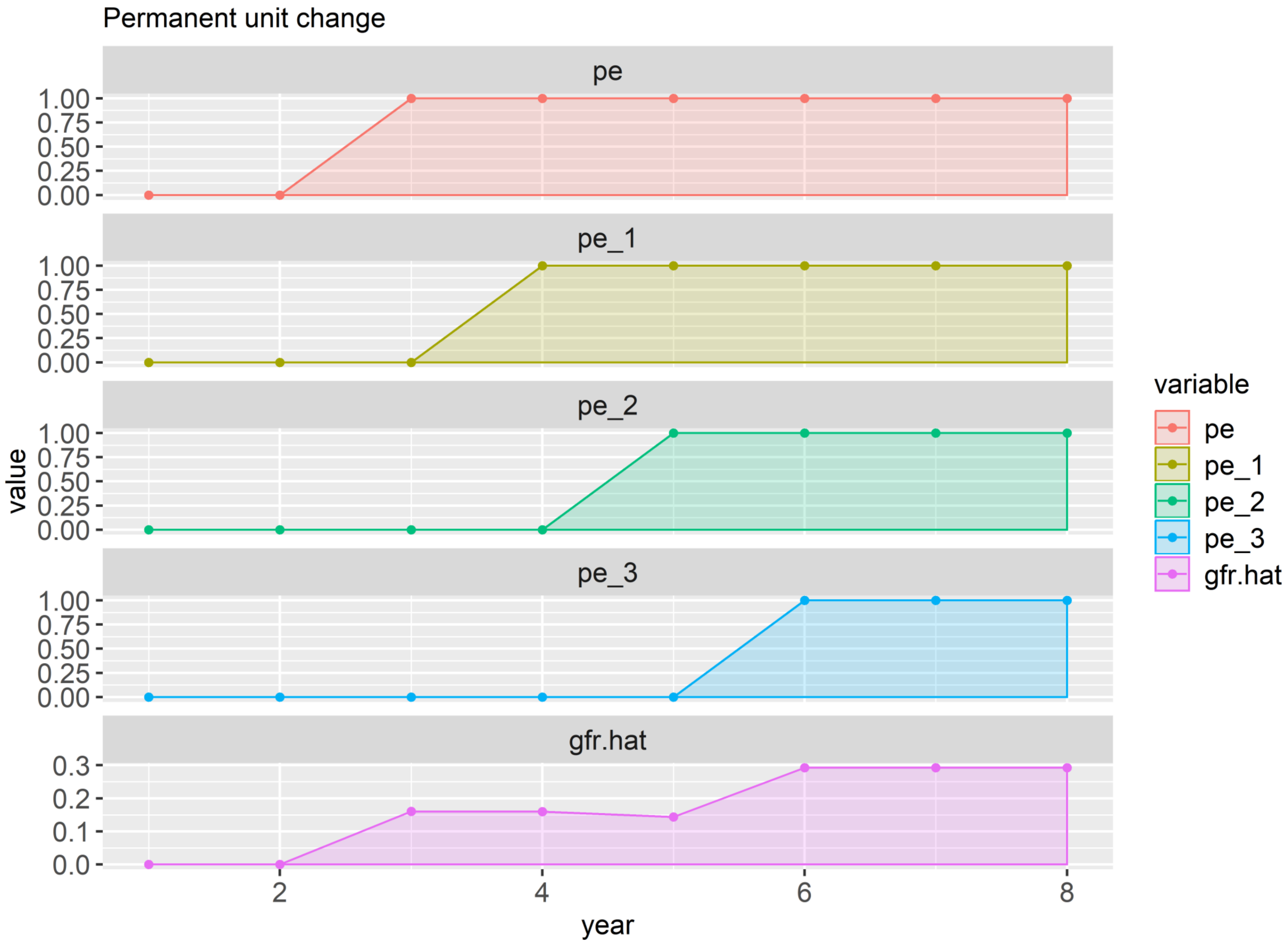
Temporary unit change
pe
pe_1
pe_2
pe_3
gfr.hat
1.00
0.75
0.50
0.25
0.00
0.15
0.10
0.05
value
2
4
6
8
year
variable
pe
pe_1
pe_2
pe_3
gfr.hat

Finite distributed lag (FDL) model (cont'd)

- **long-run propensity** (LRP): the effect of a **permanent** unit increase in *pe*

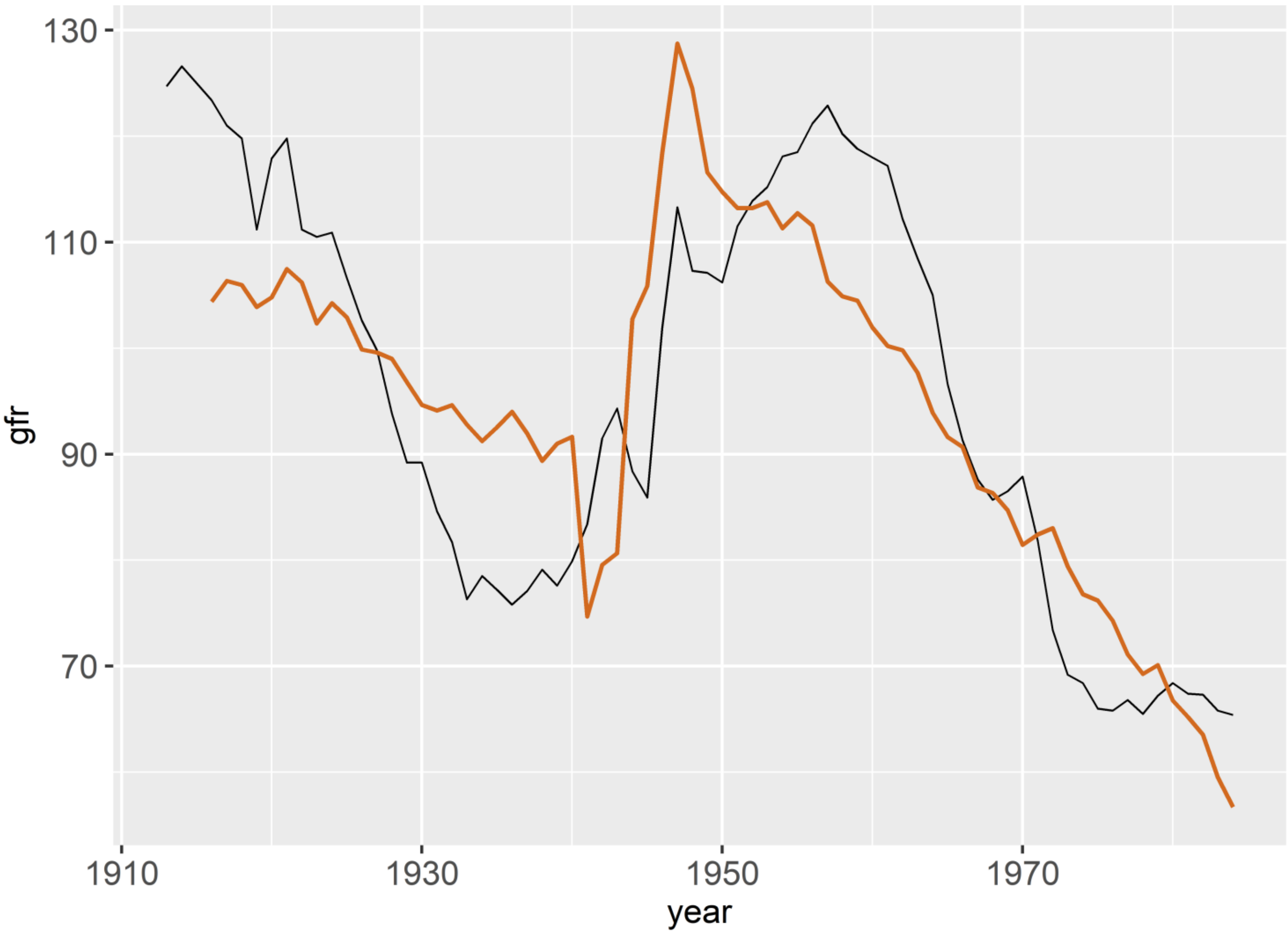
$$\text{LRP} = \delta_0 + \delta_1 + \delta_2 + \delta_3$$





Permanent unit change
pe
pe_1
pe_2
pe_3
gfr.hat
1.00
0.75
0.50
0.25
0.00
0.3
0.2
0.1
0.0
value
2
4
6
8
year
variable
pe
pe_1
pe_2
pe_3
gfr.hat

$$gfr_t = \beta_0 + \beta_1 ww2_t + \beta_2 pill_t + \beta_3 t + \delta_0 pe_t + \delta_1 pe_{t-1} + \delta_2 pe_{t-2} + \delta_3 pe_{t-3} + u_t$$



Estimating LRP

- a natural estimator of LRP is $\text{LRP} = \hat{\delta}_0 + \hat{\delta}_1 + \hat{\delta}_2 + \hat{\delta}_3$
- so we just add up the coefficients on *pe* and its lags
- more work is required in case we need std. errors or 95% CI for LRP
- we'll use a simple trick: the equation can be rewritten as follows

$$\begin{aligned} gfr_t &= \beta_0 + \delta_0 pe_t + \delta_1 pe_{t-1} + \delta_2 pe_{t-2} + \delta_3 pe_{t-3} + u_t \\ &= \beta_0 + \text{LRP}\, pe_t + \delta_1 \underbrace{(pe_{t-1} - pe_t)}_{A} + \delta_2 \underbrace{(pe_{t-2} - pe_t)}_{B} + \delta_3 \underbrace{(pe_{t-3} - pe_t)}_{C} + u_t \end{aligned}$$

- this gives us the following procedure:
  1. Create variables *A*, *B*, and *C*.
     - in Gretl: Add → Define new variable... → A = pe(−1) − pe etc.
  2. Regress *gfr* on *pe*, *A*, *B* and *C*; now, LRP is the coefficient on *pe*, and we can read off its std. error and calculate the 95% CI if needed.

Lecture 9:

Gentle introduction to regression with time series

Jan Zouhar

Introductory Econometrics