LECTURE 01

Exercise Hints & Worked Solutions

Local Projections: From First Principles to Empirical Research

Contents of this companion

No.	Topic	Page
	Exercises and Solutions	2
	Exercise 1: Two paths, one difference	2
	Exercise 2: Iterate forward	4
	Exercise 3: Rows by hand	7
	Exercise 4: The same rows in Stata	9
	Exercise 5: One episode	13
	Exercise 6: A response in a loop	15
	Exercise 7: Many draws	19
	Exercise 8: What the residual contains	23
	Exercise 9: First contact with military news	26
	Exercise 10: Korea is not the average	31

This companion is meant to be used in stages. Try the exercise first. Open the hint if you need a direction, then compare your work with the detailed solution. The solutions emphasize the reasoning behind each calculation, not only the final number.

On the website, the hints and solutions are collapsed until selected. In the PDF, every hint and solution is printed in full so the document can be read offline or used as a conventional solutions manual.

Exercises and Solutions

Exercise 1: Two paths, one difference

Problem. [core, pencil] The twelve-period economy of the notes follows $y_t = 0.5\,y_{t-1} + s_t + v_t$ with $y_0 = 0$, and its interventions fall at $t = 1, 5, 10, 12$. Around the second intervention the realized path is $y_4 = 0.0625$, $y_5 = -0.66875$, $y_6 = -0.134375$, $y_7 = 2.0328125$, and $y_8 = -0.28359375$, and the disturbances are $v_5 = -1.7$, $v_6 = 0.2$, $v_7 = 2.1$, and $v_8 = -1.3$.

(a) Build the counterfactual path y_t^c for $t = 5, \dots, 8$ by switching off the intervention at $t = 5$ and nothing else. Tabulate y_t, y_t^c, the gap $y_t - y_t^c$, and the change since the date before, $y_t - y_4$.

(b) Split each change $y_t - y_4$ into the causal response and the change the economy would have made without the intervention. At which dates does reading the change as the effect get the sign wrong, and by what factor does it overstate the effect at $t = 7$?

(c) Without repeating the recursion, give the four gaps that would result if every v_t were replaced by zero, and explain why.

Hint

Only one input differs between the two recursions: s_5 equals 1 on the realized path and 0 on the counterfactual one. Start both paths from y_4, which an intervention dated $t = 5$ cannot have moved, and feed the same v_t into both. For (c), subtract one recursion from the other before substituting any number.

Detailed solution

The two paths obey the same rule and receive the same disturbances:

$$y_t = 0.5\,y_{t-1} + s_t + v_t, \qquad y_t^c = 0.5\,y_{t-1}^c + s_t^c + v_t,$$

with $s_5^c = 0$ and $s_t^c = s_t$ at every other date. Before $t = 5$ the rule never uses s_5, so the paths coincide

up to that date and $y_4^c = y_4 = 0.0625$. The plan is to run the second recursion forward from there and compare.

(a) Four applications of the rule give the counterfactual path:

$$y_5^c = 0.5(0.0625) + 0 - 1.7 = -1.66875,$$
$$y_6^c = 0.5(-1.66875) + 0 + 0.2 = -0.634375,$$
$$y_7^c = 0.5(-0.634375) + 0 + 2.1 = 1.7828125,$$
$$y_8^c = 0.5(1.7828125) + 0 - 1.3 = -0.40859375.$$

Every input has a short decimal expansion, so the table is exact; nothing is rounded.

t	y_t	y_t^c	Gap $y_t - y_t^c$	Change $y_t - y_4$	$y_t^c - y_4$
5	−0.66875	−1.66875	1	−0.73125	−1.73125
6	−0.134375	−0.634375	0.5	−0.196875	−0.696875
7	2.0328125	1.7828125	0.25	1.9703125	1.7203125
8	−0.28359375	−0.40859375	0.125	−0.34609375	−0.47109375

The gap is measured in units of y for a one-unit intervention. Its halving can be read off the two rules. Subtracting the counterfactual rule from the realized one,

$$y_t - y_t^c = 0.5\,(y_{t-1} - y_{t-1}^c) + (s_t - s_t^c),$$

and the disturbance has cancelled. At $t = 5$ the gap is $0.5 \times 0 + 1 = 1$; after that $s_t - s_t^c = 0$, so each period keeps half of the previous gap.

(b) Add and subtract y_t^c:

$$y_t - y_4 = \underbrace{(y_t - y_t^c)}_{\text{causal response}} + \underbrace{(y_t^c - y_4)}_{\text{change without the intervention}}.$$

The last three columns of the table carry out the split. At $t = 5$, $-0.73125 = 1 + (-1.73125)$; at $t = 7$, $1.9703125 = 0.25 + 1.7203125$. The second bracket contains no intervention at all. Iterating the counterfactual rule from $t = 4$ gives $y_7^c = 0.5^3 y_4 + 0.5^2 v_5 + 0.5\,v_6 + v_7 = 0.0078125 - 0.425 + 0.1 + 2.1$, so the counterfactual change 1.7203125 is the fading of y_4 plus three disturbances, and $v_7 = 2.1$ alone is

larger than all of it.

Reading the change as the effect gets the sign wrong at $t = 5$ (-0.73125 against 1), at $t = 6$ (-0.196875 against 0.5), and at $t = 8$ (-0.34609375 against 0.125). At $t = 7$ it reports 1.9703125 against 0.25, which is 7.88125 times the effect. The intervention did the same thing at every date; the disturbances did not.

(c) The gap recursion in (a) contains no v_t. Replacing every disturbance by zero moves both paths but leaves their difference alone, so the gaps are still 1, 0.5, 0.25, and 0.125. They are the formula $\theta_h = \theta_0 \rho^h$ from the notes' section on the smallest model, with $\theta_0 = 1$ and $\rho = 0.5$, read at $h = 0, \dots, 3$. This is the check on the whole calculation: a gap that changed when the disturbances changed would signal an arithmetic slip.

Two mistakes are common. The first is the one the lecture opens with: reporting $y_5 - y_4 = -0.73125$ as the effect of the intervention. The second is to build the counterfactual by setting $s_t = 0$ at every date, as if "no intervention" meant no interventions at all. The intervention at $t = 1$ then disappears as well, its fading contribution 0.5^{t-1} joins each gap, and the gaps become 1.0625, 0.53125, 0.265625, and 0.1328125. The causal response to one intervention removes that intervention and nothing else.

Exercise 2: Iterate forward

Problem. [core, pencil] Let $y_t = \rho\, y_{t-1} + \theta_0 s_t + v_t$, where the interventions s_t and the disturbances v_t each have mean zero, are independent over time, and are independent of each other, and $|\rho| < 1$ (Assumptions A1–A4 of the notes).

(a) Substitute the rule into itself to write y_{t+1} and then y_{t+2} in terms of y_{t-1} and the interventions and disturbances dated t onward. State the general formula for y_{t+h} and prove it by induction.

(b) Change s_t by one unit while holding y_{t-1}, every disturbance, and every other intervention fixed. Derive the causal response θ_h.

(c) For the population regression of y_{t+h} on $(1, s_t, y_{t-1})$, read off μ_h, β_h, γ_h, and the residual $u_{t,h}$ from your formula. Show that $\mathbb{E}[u_{t,h}] = \mathbb{E}[s_t u_{t,h}] = \mathbb{E}[y_{t-1} u_{t,h}] = 0$, and name the assumption each step uses.

(d) With $\theta_0 = 1$ and $\rho = 0.9$, compute θ_4 and θ_7, and find the first horizon at which the response is below half of its impact value.

Hint

Do two substitutions by hand before writing any sum, and watch what the second substitution does to

every term the first one produced. That operation is the induction step. For (c), write down the date of every term in $u_{t,h}$ before taking a single expectation.

Detailed solution

The object we want is y_{t+h} written as "the pre-intervention outcome, plus everything that arrives from t onward." The rule gives each outcome in terms of the previous one, so we replace the previous outcome repeatedly until only y_{t-1} is left.

(a) At $h = 0$ the rule itself is already in that form. For $h = 1$, write the rule one period later and substitute for y_t:

$$\begin{aligned} y_{t+1} &= \rho\, y_t + \theta_0 s_{t+1} + v_{t+1} \\ &= \rho(\rho\, y_{t-1} + \theta_0 s_t + v_t) + \theta_0 s_{t+1} + v_{t+1} \\ &= \rho^2 y_{t-1} + \rho\,\theta_0 s_t + \rho\, v_t + \theta_0 s_{t+1} + v_{t+1}. \end{aligned}$$

For $h = 2$, write $y_{t+2} = \rho\, y_{t+1} + \theta_0 s_{t+2} + v_{t+2}$ and substitute the last line:

$$y_{t+2} = \rho^3 y_{t-1} + \rho^2\theta_0 s_t + \rho^2 v_t + \rho\,\theta_0 s_{t+1} + \rho\, v_{t+1} + \theta_0 s_{t+2} + v_{t+2}.$$

The second substitution multiplied every existing term by ρ and appended the new period's intervention and disturbance with coefficient one. Repeating that operation suggests

$$y_{t+h} = \rho^{h+1} y_{t-1} + \sum_{j=0}^{h} \rho^{h-j}(\theta_0 s_{t+j} + v_{t+j}).$$

Induction. At $h = 0$ the formula reads $\rho\, y_{t-1} + \theta_0 s_t + v_t$, which is the rule. Suppose it holds at h. Then

$$\begin{aligned} y_{t+h+1} &= \rho\, y_{t+h} + \theta_0 s_{t+h+1} + v_{t+h+1} \\ &= \rho^{h+2} y_{t-1} + \sum_{j=0}^{h} \rho^{h+1-j}(\theta_0 s_{t+j} + v_{t+j}) + \rho^0(\theta_0 s_{t+h+1} + v_{t+h+1}), \end{aligned}$$

and the last term is the $j = h + 1$ term of the sum at horizon $h + 1$. The formula therefore holds at $h + 1$, and so at every horizon.

(b) The intervention s_t appears only in the $j = 0$ term of the sum, with coefficient $\rho^h\theta_0$. Raising s_t by one unit with everything else fixed raises y_{t+h} by that coefficient, so

$$\theta_h = \theta_0 \rho^h,$$

in units of y per unit of s. The rule is linear and its coefficients do not depend on the date (A1), so the response is the same for every t and for any starting value of s_t.

(c) Match the formula with $y_{t+h} = \mu_h + \beta_h s_t + \gamma_h y_{t-1} + u_{t,h}$. The proposal is $\mu_h = 0$, $\beta_h = \theta_0 \rho^h$, $\gamma_h = \rho^{h+1}$, and everything left over is the residual:

$$u_{t,h} = \sum_{j=0}^{h} \rho^{h-j} v_{t+j} + \theta_0 \sum_{j=1}^{h} \rho^{h-j} s_{t+j}.$$

Its terms are dated as follows: the disturbances $v_t, \dots, v_{t+h}$ and the interventions $s_{t+1}, \dots, s_{t+h}$. The second sum starts at $j = 1$ because s_t has moved into the β_h term. A proposal is the population projection if its residual satisfies the three orthogonality conditions, because only one set of coefficients does when s_t and y_{t-1} both vary and neither is an exact linear function of the other. Here $\sigma_s^2 > 0$ (A2), and y_{t-1}, built from terms dated $t-1$ or earlier, varies and is uncorrelated with s_t (A2, A3).

1. *Mean zero.* Every v_{t+j} and every s_{t+j} has mean zero (A2, A3), so $\mathbb{E}[u_{t,h}] = 0$.
2. *Orthogonal to s_t.* The product $s_t u_{t,h}$ is a sum of terms $s_t v_{t+j}$ with $j \geq 0$ and $s_t s_{t+j}$ with $j \geq 1$. The first kind has expectation $\mathbb{E}[s_t]\,\mathbb{E}[v_{t+j}] = 0$ because the two sequences are independent (A3). The second kind has expectation $\mathbb{E}[s_t]\,\mathbb{E}[s_{t+j}] = 0$ because interventions are independent over time with mean zero (A2).
3. *Orthogonal to y_{t-1}.* Iterating the rule backward, $y_{t-1} = \sum_{i \geq 0} \rho^i (\theta_0 s_{t-1-i} + v_{t-1-i})$, a sum that settles down because $|\rho| < 1$ (A4). Every term of y_{t-1} is dated $t-1$ or earlier, every term of $u_{t,h}$ is dated t or later, and by A2 and A3 the two groups are independent. Each term of $u_{t,h}$ has mean zero, so $\mathbb{E}[y_{t-1} u_{t,h}] = 0$.

All three conditions hold, so the projection coefficients are $\mu_h = 0$, $\beta_h = \theta_h = \theta_0 \rho^h$, and $\gamma_h = \rho^{h+1}$. As a special-case check, at $h = 0$ the residual is $u_{t,0} = v_t$, $\beta_0 = \theta_0$, and $\gamma_0 = \rho$: the projection is the rule itself.

Two of the three steps used the assumptions about how interventions and disturbances arise, and none used anything about least squares. That is where the causal content of $\beta_h = \theta_h$ sits.

(d) With $\theta_0 = 1$ and $\rho = 0.9$,

$$\theta_4 = 0.9^4 = 0.6561, \qquad \theta_7 = 0.9^7 = 0.4782969.$$

The response falls below half of its impact value when $0.9^h < 0.5$, that is, when $h > \ln 0.5 / \ln 0.9 = 6.58$ (rounded to two decimals). The first such horizon is $h = 7$; one period earlier, $\theta_6 = 0.531441$ is still above one half. The control's coefficient over the same horizons is $\gamma_h = 0.9^{h+1}$, for example $\gamma_4 = 0.59049$, but it measures the fading of the pre-intervention outcome, not a response to anything.

Three slips recur. Starting the intervention sum in $u_{t,h}$ at $j = 0$ puts s_t into the residual, which makes $\mathbb{E}[s_t u_{t,h}] = \theta_0 \rho^h \sigma_s^2$ and suggests a bias that does not exist. Writing $\gamma_h = \rho^h$ forgets that y_{t-1} has had $h+1$ periods to fade. And attributing $\mathbb{E}[y_{t-1} u_{t,h}] = 0$ to A2 alone overlooks that y_{t-1} contains past disturbances, so A3 is needed as well.

Exercise 3: Rows by hand

Problem. [core, pencil] In the twelve-period economy the interventions are $s_t = 1$ at $t = 1, 5, 10, 12$ and $s_t = 0$ otherwise, and the realized outcomes are

t	y_t	t	y_t	t	y_t
1	1.9	5	−0.66875	9	0.658203125
2	1.45	6	−0.134375	10	1.9291015625
3	1.325	7	2.0328125	11	0.06455078125
4	0.0625	8	−0.28359375	12	2.532275390625

For $h = 0, 1, 2$, consider the regression of y_{t+h} on a constant and s_t, using only rows with $t \geq 2$ so that y_{t-1} would exist in every row.

(a) List the rows of each regression and give T_h. Check your count against $T_h = T - h - p$.

(b) Identify the intervention rows and compute the mean of y_{t+h} over the intervention rows, $\bar{y}_1$, and over the other rows, $\bar{y}_0$.

(c) Compute $\hat{\beta}_h$ and compare it with $\theta_h = 0.5^h$.

(d) Row 4 of the $h = 1$ regression sits in the comparison group. What is its outcome, and why is that outcome not a clean comparison?

Hint

Draw the staircase before computing anything. For each h, write t in one column and $t + h$ in the next, then strike out every row whose outcome date passes 12. Only after that sort the surviving rows by s_t.

Detailed solution

A regression row is dated by its intervention: row t pairs s_t with the outcome y_{t+h}. Row t needs the lag y_{t-1}, which exists only for $t \geq 2$ because y_0 is not part of the data, and the lead y_{t+h}, which exists only for $t \leq 12 - h$. The plan is to fix those rows first, then use the fact from the notes that with a 0–1 regressor the least-squares slope is the difference in means, $\hat{\beta}_h = \bar{y}_1 - \bar{y}_0$.

(a) The rows are $t = 2, \ldots, 12$ at $h = 0$, $t = 2, \ldots, 11$ at $h = 1$, and $t = 2, \ldots, 10$ at $h = 2$, so $T_0 = 11$, $T_1 = 10$, and $T_2 = 9$. With $T = 12$ and one lag, $p = 1$, the formula gives $T_h = 12 - h - 1$: 11, 10, and 9. Each horizon costs the last row, and the lag costs row 1 at every horizon.

(b) At $h = 0$ the intervention rows are 5, 10, and 12. At $h = 1$ row 12 has left the sample, because its outcome y_{13} does not exist, and the intervention rows are 5 and 10; they remain 5 and 10 at $h = 2$. Each row contributes its outcome h periods later:

- $h = 0$. Intervention outcomes y_5, y_{10}, y_{12} sum to 3.792626953125, so $\bar{y}_1 = 1.264208984375$. The eight other outcomes $y_2, y_3, y_4, y_6, y_7, y_8, y_9, y_{11}$ sum to 5.17509765625, so $\bar{y}_0 = 0.64688720703125$.
- $h = 1$. Intervention outcomes y_6 and y_{11} sum to -0.06982421875, so $\bar{y}_1 = -0.034912109375$. The eight other rows $t = 2, 3, 4, 6, 7, 8, 9, 11$ have outcomes $y_3, y_4, y_5, y_7, y_8, y_9, y_{10}, y_{12}$, which sum to 7.587548828125, so $\bar{y}_0 = 0.948443603515625$.
- $h = 2$. Intervention outcomes y_7 and y_{12} sum to 4.565087890625, so $\bar{y}_1 = 2.2825439453125$. The seven other rows $t = 2, 3, 4, 6, 7, 8, 9$ have outcomes $y_4, y_5, y_6, y_8, y_9, y_{10}, y_{11}$, which sum to 1.62763671875, so $\bar{y}_0 = 0.23251953125$.

(c) Subtracting,

$$\hat{\beta}_0 = 0.61732177734375, \qquad \hat{\beta}_1 = -0.983355712890625, \qquad \hat{\beta}_2 = 2.0500244140625.$$

			Intervention				
h	T_h	Rows	rows	$\bar{y}_1$	$\bar{y}_0$	$\hat{\beta}_h$	θ_h
0	11	2–12	5, 10, 12	1.264209	0.646887	0.617322	1
1	10	2–11	5, 10	−0.034912	0.948444	−0.983356	0.5
2	9	2–10	5, 10	2.282544	0.232520	2.050024	0.25

The table rounds to six decimals; the display above it is exact. The slopes are in units of y per unit of s. None is close to its target: the impact estimate is 0.62 against 1, the $h = 1$ estimate has the wrong

sign, and the $h = 2$ estimate is eight times too large. Two or three intervention rows cannot average away disturbances such as $v_5 = -1.7$ and $v_7 = 2.1$.

Check. The covariance form of the slope must agree. At $h = 0$, $\bar{s} = 3/11$, so $\sum_t (s_t - \bar{s})^2 = n_1 n_0 / n = 3 \times 8/11 = 24/11$, and $\sum_t (s_t - \bar{s})\, y_t = 1.346883877841$ (rounded to twelve decimals). Their ratio is $0.617321777\ldots$, the same slope. The intercept of each regression equals $\bar{y}_0$, which is a second check.

(d) Row 4 has $s_4 = 0$, so it belongs to the comparison group, but its outcome at $h = 1$ is $y_5 = 0.5\, y_4 + s_5 + v_5 = -0.66875$, which contains the intervention at $t = 5$. In the notation of the notes this is the term $\theta_0 s_{t+1}$ inside the residual $u_{t,1}$. Rows 9 and 11 have the same feature: their outcomes y_{10} and y_{12} contain the interventions at $t = 10$ and $t = 12$, while neither intervention row has an intervention one period later. Through that term alone, the comparison mean $\bar{y}_0$ carries three units of intervention spread over eight rows, which adds $3/8 = 0.375$ to $\bar{y}_0$ and lowers $\hat{\beta}_1$ by the same amount. In population this does no harm, because s_{t+1} is independent of s_t and averages out across many rows. In ten rows it does not average out.

The most common slip is to compute $\bar{y}_1$ minus the mean of *all* rows instead of the mean of the other rows. That gives $(n_0/n)(\bar{y}_1 - \bar{y}_0) = (8/11) \times 0.61732177734375 = 0.448961\ldots$ at $h = 0$. Another is to leave row 12 in the $h = 1$ regression with a blank outcome treated as zero, which adds a fictitious intervention row.

Exercise 4: The same rows in Stata

Problem. [core, computational] Reproduce the regressions of Exercise 3 in Stata, then add the lagged outcome as a control on the same rows. Write a do-file that

(a) enters the twelve rows of the hand table (the values of s_t and v_t are $s_t = 1$ at $t = 1, 5, 10, 12$ and $v_t = 0.9, 0.5, 0.6, -0.6, -1.7, 0.2, 2.1, -1.3, 0.8, 0.6, -0.9, 1.5$), rebuilds y_t from $y_t = 0.5\, y_{t-1} + s_t + v_t$ with $y_0 = 0$, and declares the data with `tsset t`;

(b) loops over $h = 0, 1, 2$ with `forvalues`, computes $\bar{y}_1$ and $\bar{y}_0$ with `summarize`, and runs `regress F`h'.y s if t >= 2` with and without `L.y`, storing h, T_h, and both slopes;

(c) asserts that $T_h = 11 - h$ in both regressions, that the slope without the control equals $\bar{y}_1 - \bar{y}_0$ to within 10^{-6}, that both regressions use exactly the same rows, and that the slopes without the control equal those of Exercise 3.

Report the three slopes with the control and explain why they are no longer differences in means.

Hint

Check `e(N)` before you look at any coefficient. Stata's `F.` operator returns a missing value past the last period and `regress` drops such rows without comment, so the row count is the first thing that can go wrong. Ask, too, which rows the regression with `L.y` loses that the one without it keeps.

Detailed solution

The do-file below types the rows rather than drawing them, so it depends on neither a data file nor a random-number stream. It stores v_t and y_t as `double`, for a reason given at the end. The assertions sit where the claims of Exercise 3 are made: the row count right after each regression, the difference in means before the slope is stored, and the comparison of samples before the control's slope is posted.

```
* ex04_hand_rows.do  Lecture 01, Exercise 4: the hand-table rows in Stata
* Run from the lab project's top folder with StataNow/SE 19.5.
version 19
clear all
set more off
capture log close
log using output/ex04_hand_rows.txt, text replace

* The twelve rows of the hand table, typed from the notes. Storing v and y
* as double keeps Stata's arithmetic equal to the exact decimals.
input byte t byte s double v
 1 1  0.9
 2 0  0.5
 3 0  0.6
 4 0 -0.6
 5 1 -1.7
 6 0  0.2
 7 0  2.1
 8 0 -1.3
 9 0  0.8
10 1  0.6
11 0 -0.9
```

```
12 1  1.5
end
tsset t
generate double y = s + v in 1                  // y_0 = 0
replace y = 0.5*L.y + s + v in 2/12
assert abs(y[5]  - (-0.66875))     < 1e-12   // check the typed rows
assert abs(y[12] - 2.532275390625) < 1e-12

tempname P
postfile `P' h T_h n1 n0 double(ybar1 ybar0 b_noctrl b_ctrl) ///
    using output/ex04_results, replace
forvalues h = 0/2 {
    * group means of the outcome h periods ahead, rows t >= 2
    quietly summarize F`h'.y if t >= 2 & s == 1
    local ybar1 = r(mean)
    local n1 = r(N)
    quietly summarize F`h'.y if t >= 2 & s == 0
    local ybar0 = r(mean)
    local n0 = r(N)

    * the regression without the control
    quietly regress F`h'.y s if t >= 2
    assert e(N) == 11 - `h'                     // T_h = T - h - p
    assert `n1' + `n0' == e(N)
    assert abs(_b[s] - (`ybar1' - `ybar0')) < 1e-6
    assert abs(_b[_cons] - `ybar0') < 1e-6
    local b_noctrl = _b[s]
    tempvar rows
    generate byte `rows' = e(sample)

    * the regression with y_(t-1), which must use the same rows
    quietly regress F`h'.y s L.y if t >= 2
```

```
    assert e(N) == 11 - `h'
    quietly count if e(sample) != `rows'
    assert r(N) == 0
    post `P' (`h') (e(N)) (`n1') (`n0') (`ybar1') (`ybar0') ///
        (`b_noctrl') (_b[s])
    drop `rows'
}
postclose `P'

use output/ex04_results, clear
format ybar1 ybar0 b_noctrl b_ctrl %12.8f
list, clean noobs

* exact slopes from the notes; with-control slopes to six decimals
assert n1 == cond(h == 0, 3, 2)
assert abs(b_noctrl - 0.61732177734375)     < 1e-9 if h == 0
assert abs(b_noctrl - (-0.983355712890625)) < 1e-9 if h == 1
assert abs(b_noctrl - 2.0500244140625)      < 1e-9 if h == 2
assert abs(b_ctrl - 0.575667)    < 5e-7 if h == 0
assert abs(b_ctrl - (-1.050405)) < 5e-7 if h == 1
assert abs(b_ctrl - 2.153022)    < 5e-7 if h == 2
display "Exercise 4: all assertions passed"
log close
```

Run in StataNow/SE 19.5, every assertion passes. The listing in the log is:

```
    h   T_h   n1   n0         ybar1        ybar0      b_noctrl        b_ctrl
    0    11    3    8    1.26420898   0.64688721    0.61732178    0.57566739
    1    10    2    8   -0.03491211   0.94844360   -0.98335571   -1.05040519
    2     9    2    7    2.28254395   0.23251953    2.05002441    2.15302246
```

Without the control, each slope is the difference in means of Exercise 3, printed to eight decimals: 0.61732178 is 0.61732177734375 rounded. The row counts confirm $T_h = T - h - p$, and the group sizes confirm that row 12's intervention leaves at $h = 1$.

With y_{t-1} included, the slopes are 0.575667, −1.050405, and 2.153022 (rounded to six decimals), on exactly the same rows. They are no longer differences in means. By the Frisch–Waugh–Lovell result that the notes preview in a footnote, the slope now uses s_t after removing its linear relation to y_{t-1} in these rows. In population that relation is zero, because the intervention is independent of the previous outcome (A2 and A3). In nine to eleven rows it is not exactly zero, so the control moves the three estimates by 0.04, 0.07, and 0.10 in absolute value (rounded to two decimals). The control's own coefficients tell the same story: exact least squares on these rows gives $\hat{\gamma}_0 = -0.061$, $\hat{\gamma}_1 = -0.115$, and $\hat{\gamma}_2 = 0.233$ (rounded to three decimals), against $\rho^{h+1} = 0.5$, 0.25, and 0.125. The population coefficient the two regressions target is the same, $\beta_h = \theta_h$; neither estimate comes near it in twelve periods.

Why the tolerance. With `double` storage, Stata's slopes agree with exact decimal arithmetic to about 10^{-15}, so even the 10^{-9} assertions pass. Stored as `float`, which is what `generate` and `input` use by default, v_t becomes a binary approximation such as 0.89999998; delete `double` from the `input` and `generate` lines to see it. The slopes then print as 0.61732179, −0.98335569, and 2.05002437, the values the notes quote, because the shipped `hand_table.csv` was generated from float draws. They still satisfy every 10^{-6} assertion, while the 10^{-12} checks on the typed rows and the 10^{-9} checks on the slopes fail, as they should. A test written as `assert _b[s] == ybar1 - ybar0` would be testing storage, not the regression.

What the first assertion catches. Leaving out `if t >= 2` changes one regression but not the other. At $h = 0$ the regression without the control then uses all twelve rows, including row 1 with its intervention, and returns 0.7762695 (rounded to seven decimals). The regression with `L.y` still uses eleven rows, because y_0 is missing. The two slopes would differ in their rows as well as in their control, and the first assertion, `e(N) == 11 - h`, stops the do-file before anyone compares them.

Exercise 5: One episode

Problem. [core, pencil] Replace the intervention column of the hand table by an indicator that equals 1 at $t = 5$ and 0 at every other date, and keep the rows $t \geq 2$ and the outcomes listed in Exercise 3.

(a) Show that when exactly one of n rows, row τ, has the indicator equal to 1, the least-squares slope of y_{t+h} on a constant and the indicator is $y_{\tau+h} - \bar{y}_{\text{others}}$, where $\bar{y}_{\text{others}}$ is the mean outcome over the other $n - 1$ rows.

(b) Compute the slope at $h = 0, 1, 2$ and compare each with $\theta_h = 0.5^h$.

(c) Use the forward equation $y_{\tau+h} = \theta_h + \rho^{h+1} y_{\tau-1} + u_{\tau,h}$ to split the $h = 0$ slope into θ_0 and everything else. Which term dominates?

(d) Return to the original intervention column ($s_t = 1$ at $t = 1, 5, 10, 12$). As h grows, the regression of y_{t+h} on s_t over rows $t \geq 2$ becomes a single-episode regression without any change to the data. At which horizon does that first happen, and what is the slope there?

Hint

With a single intervention row, the "mean of the intervention rows" is one outcome. Start from the difference-in-means result of Exercise 3 with $n_1 = 1$, or compute $\bar{s}$, $\sum_t (s_t - \bar{s})^2$, and the weight on each row directly.

Detailed solution

(a) The slope on a 0–1 regressor is the difference in means, so with one intervention row $\bar{y}_1 = y_{\tau+h}$ and $\bar{y}_0 = \bar{y}_{\text{others}}$, and the result follows at once. The weights make the same point more concretely. With one intervention among n rows, $\bar{s} = 1/n$ and

$$\sum_t (s_t - \bar{s})^2 = \Big(1 - \frac{1}{n}\Big)^2 + (n-1)\frac{1}{n^2} = \frac{(n-1)^2 + (n-1)}{n^2} = \frac{n-1}{n}.$$

The slope is $\sum_t \omega_t y_{t+h}$ with $\omega_t = (s_t - \bar{s}) / \sum_\tau (s_\tau - \bar{s})^2$. Row τ gets $\omega_\tau = (1 - 1/n) \big/ ((n-1)/n) = 1$, and every other row gets $-(1/n) \big/ ((n-1)/n) = -1/(n-1)$. Hence

$$\hat{\beta}_h = y_{\tau+h} - \frac{1}{n-1} \sum_{t \neq \tau} y_{t+h} = y_{\tau+h} - \bar{y}_{\text{others}}.$$

The episode's outcome gets full weight, and the other rows share a weight of -1 equally.

(b) At $h = 0$ the rows are $t = 2, \ldots, 12$, so $n = 11$. The episode's outcome is $y_5 = -0.66875$, and the ten other outcomes sum to 9.636474609375, a mean of 0.9636474609375. At $h = 1$ ($n = 10$) the episode's outcome is $y_6 = -0.134375$, and the nine others, $y_3, y_4, y_5, y_7, \ldots, y_{12}$, sum to 7.652099609375, a mean of 0.850233289.... At $h = 2$ ($n = 9$) the episode's outcome is $y_7 = 2.0328125$, and the eight others, $y_4, y_5, y_6, y_8, \ldots, y_{12}$, sum to 4.159912109375, a mean of 0.519989013671875. So

$$\hat{\beta}_0 = -1.6323974609375, \qquad \hat{\beta}_1 = -0.984608\ldots, \qquad \hat{\beta}_2 = 1.512823486328125,$$

or -1.632397, -0.984608, and 1.512823 rounded to six decimals, against $\theta_h = 1$, 0.5, and 0.25. Two estimates have the wrong sign and the third is six times too large.

(c) At $\tau = 5$ and $h = 0$ the forward equation is the rule itself: $y_5 = \theta_0 + \rho\, y_4 + u_{5,0}$ with $u_{5,0} = v_5$.

Substituting $\theta_0 = 1$, $\rho\, y_4 = 0.5 \times 0.0625 = 0.03125$, and $v_5 = -1.7$,

$$\hat{\beta}_0 = \underbrace{1}_{\theta_0} + \underbrace{(0.03125 - 1.7 - 0.9636474609375)}_{\text{everything else}} = 1 + (-2.6323974609375) = -1.6323974609375.$$

Everything else is more than twice the response, and the episode's own disturbance, -1.7, is by itself larger than θ_0. The same split works at $h = 1$: $y_6 = \theta_1 + \rho^2 y_4 + u_{5,1}$ with $u_{5,1} = \rho\, v_5 + v_6 + \theta_0 s_6 = -0.85 + 0.2 + 0 = -0.65$, so $\hat{\beta}_1 = 0.5 + (0.015625 - 0.65 - 0.850233\ldots) = 0.5 + (-1.484608\ldots)$. The data contain one number at each horizon. Nothing in them says how much of it is θ_h and how much is the bracket.

(d) At $h = 3$ the rows are $t = 2, \ldots, 9$. Row 10 and row 12 have no outcome three periods later, and row 1 is excluded, so row 5 is the only intervention row left. The original regression is then exactly the single-episode regression, and part (a) applies to it unchanged:

$$\hat{\beta}_3 = y_8 - \frac{y_5 + y_6 + y_7 + y_9 + y_{10} + y_{11} + y_{12}}{7} = -0.28359375 - \frac{6.413818359375}{7} = -1.199853515625,$$

against $\theta_3 = 0.125$. This is also a check on (a): the slope from the difference-in-means formula of Exercise 3 and the slope from the one-episode formula must coincide here, and they do. Row 5 stays the only intervention row through $h = 7$, so the regressions at $h = 4, \ldots, 7$, on 7, 6, 5, and 4 rows, are single-episode regressions as well. At $h = 8$ the rows are $t = 2, 3, 4$, none of them has an intervention, s_t does not vary, and no slope can be computed; Stata's `regress` reports `s` as omitted because of collinearity.

The Korean episode in the notes has the same structure, with taxes, the Treasury–Federal Reserve Accord, and price controls inside the bracket.

It is easy to subtract the mean of all n rows, including the episode itself. That shrinks the answer by the factor $(n-1)/n$: at $h = 0$ it gives $(10/11) \times (-1.6323974609375) = -1.483998$ (rounded to six decimals). The error is small here, but it hides the logic of the formula, which compares the episode with everything else.

Exercise 6: A response in a loop

Problem. [core, computational] The Practicum 01 lab project ships `data/raw/onedraw_rho09_seed2.csv`, one simulated sample of $T = 200$ periods from $y_t = 0.9\, y_{t-1} + s_t + v_t$ with $\sigma_s = \sigma_v = 1$ (seed 2; variables

`t`, `s`, `v`, `y`). Lab 3 of the Shock-to-Response Explorer displays the same observations.

(a) Write a do-file that loops over $h = 0, \dots, 12$, regresses y_{t+h} on a constant, s_t, and y_{t-1}, and stores h, $\hat{\beta}_h$, and T_h with `postfile`.

(b) Assert that $T_h = 199 - h$, that $\hat{\beta}_0 = 0.878940$ and $\hat{\beta}_{12} = -0.293297$ to within 10^{-6}, and state how many of the thirteen estimates lie below $\theta_h = 0.9^h$.

(c) Draw the response graph with θ_h overlaid and write a caption that names the intervention and its units, the units of the outcome, the horizon unit, and the counterfactual.

(d) In two or three sentences, say what the graph shows about this sample and what it cannot show.

Hint

Put `e(N)` in the same `post` command as `_b[s]`, so each coefficient travels with the number of rows behind it. For the caption, fill four slots before writing a sentence: what one unit of the intervention is, what the vertical axis measures, what the horizontal axis counts, and what a value of zero is being compared with.

Detailed solution

One regression per horizon produces one point on the curve, so the loop has thirteen regressions and a results file with thirteen rows. The design decision that matters most is to store each coefficient together with its sample size and to test the sample size before trusting the coefficient.

```
* ex06_loop.do  Lecture 01, Exercise 6: a response in a loop
* Run from the lab project's top folder with StataNow/SE 19.5.
version 19
clear all
set more off
capture log close
log using output/ex06_loop.txt, text replace

import delimited using data/raw/onedraw_rho09_seed2.csv, clear   // t s v y
tsset t
assert _N == 200

tempname P
postfile `P' h double beta_h T_h using output/results_onedraw, replace
```

```
forvalues h = 0/12 {
    quietly regress F`h'.y s L.y
    post `P' (`h') (_b[s]) (e(N))          // coefficient with its rows
}
postclose `P'

use output/results_onedraw, clear
generate double theta_h = 0.9^h
assert T_h == 199 - h                       // T - h - p, T = 200, p = 1
assert abs(beta_h[1]  - 0.878940)    < 1e-6  // h = 0
assert abs(beta_h[13] - (-0.293297)) < 1e-6  // h = 12
quietly count if beta_h < theta_h
assert r(N) == 13                           // below the truth at every h
format beta_h theta_h %9.6f
list h T_h beta_h theta_h, clean noobs

twoway (line theta_h h, lcolor(gs8) lpattern(dash))                  ///
       (connected beta_h h, lcolor(navy) mcolor(navy) msize(small)), ///
       yline(0, lcolor(gs13))                                        ///
       xlabel(0(2)12) ylabel(-0.4(0.2)1, format(%3.1f) angle(0))     ///
       xtitle("Horizon h (periods after the intervention)")          ///
       ytitle("Gap from the no-intervention path (units of y)")      ///
       legend(order(2 "Estimate, this sample"                        ///
                    1 "True response 0.9{sup:h}") rows(1) position(6)) ///
       graphregion(color(white)) plotregion(color(white))
graph export output/ex06_response.pdf, replace
display "Exercise 6: all assertions passed"
log close
```

The assertions pass, and the log lists:

```
     h   T_h      beta_h    theta_h
     0   199    0.878940   1.000000
```

```
 1   198    0.853260   0.900000
 2   197    0.612737   0.810000
 3   196    0.611046   0.729000
 4   195    0.548477   0.656100
 5   194    0.583099   0.590490
 6   193    0.509120   0.531441
 7   192    0.417788   0.478297
 8   191    0.240595   0.430467
 9   190    0.130293   0.387420
10   189   -0.028821   0.348678
11   188   -0.117200   0.313811
12   187   -0.293297   0.282430
```

An independent least-squares calculation in Python on the same CSV reproduces all thirteen coefficients to the six decimals shown. All thirteen estimates lie below θ_h. In the exported graph the connected estimate starts near 0.88 and runs below the dashed truth by between 0.01 and 0.20 through $h = 8$ (rounded to two decimals). After that the gap widens at every step, and the estimate is negative from $h = 10$ on while 0.9^h stays positive.

A caption that does its job. "Response of y to a one-unit intervention s_t (one unit is one standard deviation, since $\sigma_s = 1$), estimated by local projections of y_{t+h} on a constant, s_t, and y_{t-1} in one simulated sample of $T = 200$ periods with $\rho = 0.9$ (seed 2). Horizons are periods after the intervention; $h = 0$ is the period in which it occurs. The vertical axis is the gap between the path with the intervention and the path without it, in units of y. The dashed line is the true response $\theta_h = 0.9^h$. No sampling uncertainty is shown." Each of the four slots is filled: the intervention and its unit, the outcome's units, the horizon unit, and the counterfactual that zero refers to.

What the graph shows and what it does not. In this sample the estimated response is below the truth at every horizon. The shortfall grows from 0.19 at $h = 8$ to 0.58 at $h = 12$ (rounded to two decimals), and a reader shown only this curve would say the effect turns negative at $h = 10$. The graph cannot say whether the sample is typical. Neighbouring estimates share almost all their rows and most of their residual terms, so a curve that errs in one direction at every horizon may be one error seen thirteen times. Only the distribution across many samples, which Exercise 7 builds, or a band computed from the sample itself, which Lectures 5 and 6 build, can separate the two.

The mistake to expect here is in the caption rather than the code. Writing "response of y" on the vertical axis without saying that zero means "no different from the path without the intervention" invites the reading that zero means "y is back where it started". The graph measures gaps between paths, not changes along one.

Exercise 7: Many draws

Problem. [computational] Wrap the loop of Exercise 6 in a Monte Carlo experiment. Each sample comes from $y_t = \rho\, y_{t-1} + s_t + v_t$ with $y_0 = 0$ and s_t and v_t independent standard normal draws: 300 periods, all 300 values of s drawn before any v (`generate s = rnormal()`, then `generate v = rnormal()`), the first 100 periods discarded, so $T = 200$. Set the seed to 3 once for each $\rho \in \{0.5, 0.9\}$, outside the replication loop, and expose the number of replications as `global R` with a default of 200. In each sample, estimate $\hat{\beta}_h$ for $h = 0, \ldots, 12$ with y_{t-1} ($199 - h$ rows) and without it, on every row that has the lead ($200 - h$ rows).

(a) For each ρ and h, report the mean, standard deviation, and 5th and 95th percentiles of $\hat{\beta}_h$ with the control, and the mean and standard deviation without it. With $R = 200$, assert that at $\rho = 0.9$ and $h = 8$ the mean and standard deviation with the control are 0.370337 and 0.211226 to within 10^{-6}.

(b) Using the Monte Carlo standard error, the standard deviation divided by $\sqrt{R}$, decide at which horizons the mean differs from $\theta_h = \rho^h$ by more than simulation noise.

(c) Compare the spread with and without the control at each ρ. Explain the pattern with the variance of the term $\rho^{h+1} y_{t-1}$ that the regression without the control leaves in its residual.

(d) Where does the sample of Exercise 6 lie relative to the 5–95 percent band at $h = 8$ and at $h = 12$?

Hint

Post every estimate with its replication number, ρ, h, and `e(N)`, and compute summaries after the loop from the stored estimates rather than inside it. For (c), work out $\rho^{2(h+1)} \operatorname{Var}(y)$ at $h = 0$ and $h = 4$ for both values of ρ before you look at the simulated standard deviations.

Detailed solution

A Monte Carlo experiment answers a question one sample cannot: how an estimator behaves when the true response is known. The design fixes everything except the draws, so differences across replications are sampling variation and differences between the mean and θ_h are properties of the estimator. Seeding once per design makes the $R = 200$ run the first 200 replications of the same stream as the notes' $R = 500$ table.

```
* ex07_many_draws.do  Lecture 01, Exercise 7: many draws
* Run from the lab project's top folder with StataNow/SE 19.5.
version 19
clear all
set more off
capture log close
log using output/ex07_many_draws.txt, text replace

if "$R" == "" global R 200          // replications (D2 student default)
timer clear 1
timer on 1

tempname P
postfile `P' rep double rho h double(b_ctrl b_noctrl) n_ctrl n_noctrl ///
    using output/mc_draws, replace
foreach rho in 0.5 0.9 {
    set seed 3                      // once per design, before the loop
    quietly forvalues r = 1/$R {
        clear
        set obs 300
        generate t = _n
        tsset t
        generate s = rnormal()                  // s first, then v
        generate v = rnormal()
        generate y = s + v in 1
        replace y = `rho'*L.y + s + v in 2/300
        drop if t <= 100                        // burn-in of 100 periods
        replace t = _n
        tsset t
        forvalues h = 0/12 {
            regress F`h'.y s L.y
            local b_c = _b[s]
```

```
            local n_c = e(N)
            regress F`h'.y s
            post `P' (`r') (`rho') (`h') (`b_c') (_b[s]) (`n_c') (e(N))
        }
    }
}
postclose `P'
timer off 1
quietly timer list 1
display "Monte Carlo runtime: " %6.1f r(t1) " seconds for R = $R"

use output/mc_draws, clear
assert n_ctrl == 199 - h & n_noctrl == 200 - h
generate double theta_h = rho^h
collapse (mean) theta_h mean_ctrl=b_ctrl mean_noctrl=b_noctrl ///
         (sd) sd_ctrl=b_ctrl sd_noctrl=b_noctrl               ///
         (p5) p5=b_ctrl (p95) p95=b_ctrl, by(rho h)
generate double bias = mean_ctrl - theta_h
generate double mcse = sd_ctrl/sqrt($R)
format theta_h mean_* sd_* p5 p95 bias mcse %9.6f
list rho h theta_h mean_ctrl bias mcse sd_ctrl p5 p95 ///
    mean_noctrl sd_noctrl, clean noobs
save output/mc_R${R}_summary, replace

* checks
assert sd_noctrl > sd_ctrl if rho == 0.9   // the control buys precision
assert abs(bias) < 3*mcse if rho == 0.5 & h <= 4
if $R == 200 {
    assert abs(mean_ctrl - 0.370337) < 1e-6 if rho == 0.9 & h == 8
    assert abs(sd_ctrl   - 0.211226) < 1e-6 if rho == 0.9 & h == 8
}
display "Exercise 7: all assertions passed"
```

```
log close
```

Two runs took 27.3 and 27.8 seconds in StataNow/SE 19.5, and every assertion passed in both. The no-control regressions use $200 - h$ rows, one more than the controlled ones, because they do not need y_0; this matches the design behind the notes' table. Selected rows of the summary, rounded to four decimals:

ρ	h	θ_h	Mean	MCSE	SD	5th pct.	95th pct.	Mean, no control	SD, no control
0.5	0	1.0000	1.0028	0.0051	0.0717	0.8800	1.1171	0.9994	0.0970
0.5	4	0.0625	0.0507	0.0085	0.1208	−0.1459	0.2439	0.0548	0.1217
0.5	8	0.0039	−0.0169	0.0083	0.1173	−0.2065	0.1901	−0.0172	0.1178
0.5	12	0.0002	−0.0105	0.0085	0.1197	−0.2034	0.1764	−0.0120	0.1206
0.9	0	1.0000	1.0031	0.0051	0.0720	0.8783	1.1135	0.9690	0.2209
0.9	4	0.6561	0.6279	0.0134	0.1899	0.3424	0.9418	0.6219	0.2651
0.9	8	0.4305	0.3703	0.0149	0.2112	0.0608	0.6944	0.3705	0.2568
0.9	12	0.2824	0.2303	0.0155	0.2185	−0.1138	0.6167	0.2320	0.2449

(a) The log prints the asserted cell at full precision: at $\rho = 0.9$ and $h = 8$ the mean is 0.370337 and the standard deviation 0.211226. Every estimate is in units of y per unit of s, and one unit of s is one standard deviation.

(b) The Monte Carlo standard error measures how far a simulated mean could plausibly lie from the mean that unlimited replications would give. Divide each gap between the mean estimate and θ_h by its MCSE, and read a ratio beyond about two as more than simulation noise. At $\rho = 0.9$ every mean from $h = 1$ to $h = 12$ is below θ_h, and the ratio passes two at $h = 2$ and at every horizon from $h = 4$ on. The gap at $h = 2$ is -0.0233, 2.2 standard errors; at $h = 4$ it is -0.0282, 2.1; from $h = 5$ on, every gap is between 2.9 and 4.0 standard errors, the largest being -0.0601 at $h = 8$. Only $h = 0$, 1, and 3 stay inside two. That is more than simulation noise: the estimator with $T = 200$ falls systematically short at this persistence, by more at long horizons than at short ones, which is the small-sample bias Lecture 2 takes up.

At $\rho = 0.5$ the ratio also passes two at four horizons: $h = 2$, 5, 7, and 8, at 2.2, 2.8, 2.0, and 2.5 standard errors, with the largest gap -0.0241 at $h = 5$. These gaps stay within 0.025 of θ_h, about a fifth of the estimator's standard deviation, and they do not grow with h. Thirteen correlated comparisons easily produce a few ratios of that size, and the notes' $R = 500$ run puts every gap at $\rho = 0.5$ within 0.014. A

small shortfall at $\rho = 0.5$ cannot be ruled out, but at its largest it is less than half the one at $\rho = 0.9$. One caution applies to both readings: the thirteen means come from the same samples, so their errors are correlated and the ratios are not thirteen independent tests.

(c) With $\rho = 0.9$ the control cuts the spread at impact by a factor of three, from 0.2209 to 0.0720. Those are close to the large-sample values in the notes, $\sqrt{9.53/200} = 0.218$ and $1/\sqrt{199} = 0.071$. The ratio of the two standard deviations falls with the horizon, to 1.40 at $h = 4$, 1.22 at $h = 8$, and 1.12 at $h = 12$ (rounded to two decimals). The means with and without the control are close at $h = 4$, 8, and 12; the control changes the spread much more than the centre.

The omitted term explains both patterns. Without the control, the residual carries $\rho^{h+1} y_{t-1}$, whose variance is $\rho^{2(h+1)} \operatorname{Var}(y)$ with $\operatorname{Var}(y) = (\theta_0^2 \sigma_s^2 + \sigma_v^2)/(1 - \rho^2)$. At $\rho = 0.9$, $\operatorname{Var}(y) = 10.53$, and the added variance is still 3.67 at $h = 4$. At $\rho = 0.5$, $\operatorname{Var}(y) = 2.67$ and the added variance at $h = 4$ is 0.0026. So at $\rho = 0.5$ the control matters most at impact, 0.0717 against 0.0970, a little at $h = 1$ and $h = 2$, where the ratios are 1.13 and 1.07, and hardly at all after that. From $h = 4$ on the two standard deviations are within one percent of each other, and at $h = 5$ and $h = 6$ the one without the control is even slightly smaller, which is simulation noise.

(d) In Exercise 6, $\hat{\beta}_8 = 0.240595$ lies inside the band from 0.0608 to 0.6944, below the centre but not unusually so. At $h = 12$, $\hat{\beta}_{12} = -0.293297$ lies below the 5th percentile, -0.1138. The same sample is ordinary at one horizon and unusual four horizons later. That fits what the notes say about the residual: neighbouring estimates share rows and residual terms, so a sample that starts low tends to stay low and can drift out of the band.

Seeding inside the loop. Putting `set seed 3` inside the replication loop resets the generator in every replication. Each replication then draws the same sample, so the experiment is a single seed-3 sample estimated two hundred times. A standard deviation of exactly zero at every horizon is the symptom.

Exercise 8: What the residual contains

Problem. [core, pencil] In the AR(1) economy with $\theta_0 = \sigma_s = \sigma_v = 1$, the horizon-$h$ residual of the regression on $(1, s_t, y_{t-1})$ is $u_{t,h} = \sum_{j=0}^{h} \rho^{h-j} v_{t+j} + \sum_{j=1}^{h} \rho^{h-j} s_{t+j}$.

(a) Write $u_{t,1}$ and $u_{t+1,1}$ term by term, list the terms they share, and derive $\operatorname{Corr}(u_{t,1}, u_{t+1,1})$ as a function of ρ.

(b) Write $u_{t,2}$ and $u_{t+k,2}$ for $k = 1, 2, 3$. Count the shared terms and compute $\operatorname{Corr}(u_{t,2}, u_{t+k,2})$ at $\rho = 0.5$ for each k.

(c) Explain in one or two sentences why the correlation is zero beyond lag h.

(d) Show that $\mathbb{E}[s_t u_{t,2}\, s_{t+1} u_{t+1,2}] = 0$, even though $u_{t,2}$ and $u_{t+1,2}$ are correlated.

Hint

Write the two residuals one above the other, aligned by date, with each term's weight. A covariance collects only the terms that appear in both rows, and each shared term contributes the product of its two weights.

Detailed solution

Every term of a residual is either a disturbance v or an intervention s, each with variance one, and all of them are mutually independent (A2 and A3). So the variance of a residual is the sum of its squared weights, and the covariance of two residuals is the sum, over the terms they share, of the products of the two weights. Nothing else enters.

(a) At $h = 1$,

$$u_{t,1} = \rho\, v_t + v_{t+1} + s_{t+1}, \qquad u_{t+1,1} = \rho\, v_{t+1} + v_{t+2} + s_{t+2}.$$

The rows share one term, v_{t+1}, with weight 1 in row t and weight ρ in row $t+1$. Each variance is $\rho^2 + 1 + 1 = 2 + \rho^2$, and the covariance is $1 \times \rho = \rho$. Hence

$$\operatorname{Corr}(u_{t,1}, u_{t+1,1}) = \frac{\rho}{2 + \rho^2},$$

which at $\rho = 0.5$ is $0.5/2.25 = 2/9 = 0.2222$ (rounded to four decimals).

(b) At $h = 2$ the terms and weights are:

Residual	v_t	v_{t+1}	v_{t+2}	v_{t+3}	v_{t+4}	s_{t+1}	s_{t+2}	s_{t+3}	s_{t+4}
$u_{t,2}$	ρ^2	ρ	1			ρ	1		
$u_{t+1,2}$		ρ^2	ρ	1			ρ	1	
$u_{t+2,2}$			ρ^2	ρ	1			ρ	1

and $u_{t+3,2}$ contains $v_{t+3}, v_{t+4}, v_{t+5}$ and s_{t+4}, s_{t+5}. The variance of each residual is

$$\operatorname{Var}(u_{t,2}) = \rho^4 + \rho^2 + 1 + \rho^2 + 1 = 2 + 2\rho^2 + \rho^4.$$

- *Lag 1.* Rows t and $t+1$ share three terms: v_{t+1} (weights ρ and ρ^2), v_{t+2} (1 and ρ), and s_{t+2} (1 and ρ). The covariance is $\rho^3 + \rho + \rho = 2\rho + \rho^3$.
- *Lag 2.* Rows t and $t+2$ share one term, v_{t+2} (1 and ρ^2), so the covariance is ρ^2.
- *Lag 3.* Row $t+3$ begins at date $t+3$, after row t's last term. Nothing is shared, and the covariance is zero.

At $\rho = 0.5$ the variance is $2 + 0.5 + 0.0625 = 2.5625 = 41/16$, and

$$\begin{aligned}
\operatorname{Corr}(u_{t,2}, u_{t+1,2}) &= \frac{1.125}{2.5625} = \frac{18}{41} = 0.4390, \\
\operatorname{Corr}(u_{t,2}, u_{t+2,2}) &= \frac{0.25}{2.5625} = \frac{4}{41} = 0.0976, \\
\operatorname{Corr}(u_{t,2}, u_{t+3,2}) &= 0,
\end{aligned}$$

with the decimals rounded to four places. These are the $\rho = 0.5$ entries of the notes' table of residual autocorrelations, where the long simulated sample gives 0.2213, 0.4375, 0.0980, and 0.0011. The same counting at $\rho = 0.9$ gives 0.3203, 0.5914, and 0.1894, so the correlations grow with persistence.

(c) Row t's residual contains terms dated t through $t+h$, and row $t+k$'s residual contains terms dated $t+k$ through $t+k+h$. For $k > h$ the two windows do not overlap, and independent terms with no overlap have zero covariance. The residual is therefore at most a moving average of order h.

(d) Look for s_t in the product $s_t\, u_{t,2}\, s_{t+1}\, u_{t+1,2}$. It appears once, as the first factor. It is not in $u_{t,2}$, whose interventions start at s_{t+1}; it is not s_{t+1}; and it is not in $u_{t+1,2}$. By A2 and A3, s_t is independent of the other three factors, so

$$\mathbb{E}[s_t u_{t,2}\, s_{t+1} u_{t+1,2}] = \mathbb{E}[s_t]\ \mathbb{E}[u_{t,2}\, s_{t+1} u_{t+1,2}] = 0.$$

Because each product also has mean zero, the covariance of the two products is zero while the residuals themselves have correlation 0.4390. The estimation error of $\hat{\beta}_h$ is a weighted sum of residuals with weights built from the interventions, so it is these products, not the bare residuals, whose dependence matters for its variance. Lecture 5 builds standard errors on that fact.

A frequent error is to leave the interventions $s_{t+1}, \ldots, s_{t+h}$ out of the residual. At $h = 2$ and lag 1 that drops one shared term and two variance terms, and gives $(\rho^3 + \rho)/(\rho^4 + \rho^2 + 1) = 0.625/1.3125 = 0.4762$ at $\rho = 0.5$ instead of 0.4390. Another is to conclude from correlated residuals that $\hat{\beta}_h$ is biased; part (d) and the orthogonality conditions of Exercise 2 show why it is not.

Exercise 9: First contact with military news

Problem. [data, computational] Ramey and Zubairy's quarterly U.S. data are in the workbook `RZDAT.xlsx`, sheet `rzdat`, from their February 2018 replication package. The course does not redistribute the file; the Practicum 01 lab project ships `data/raw/get_rz.do`, which downloads it from the authors' archive, checks its SHA-256, and prints manual download steps if the download fails. The variables you need are `quarter` (fractional years), `rgdp` (real GDP) and `rgdp_pott6` (trend GDP, the series the authors call potential GDP and estimate as a sixth-degree polynomial trend in real GDP), both in billions of chained 2009 dollars at an annual rate, `pgdp` (the GDP deflator), and `news` (military news, billions of nominal dollars).

(a) Run `get_rz.do`, import the sheet, keep quarters from 1889Q1 on, and declare quarterly dates.

(b) Construct $y_t = \texttt{rgdp}_t/\texttt{rgdp_pott6}_t$ and $s_t = \texttt{news}_t/(\texttt{rgdp_pott6}_{t-1} \times \texttt{pgdp}_{t-1})$. Assert that s_t has 504 nonmissing values.

(c) Estimate the local projection of y_{t+h} on a constant, s_t, and y_{t-1} for $h = 0, \ldots, 20$, storing $\hat{\beta}_h$ and T_h. Assert $T_h = 504 - h$ and explain why the formula $T_h = T - h - p$, applied with $T = 504$ quarters of news and $p = 1$, seems to predict one row fewer. Report $\hat{\beta}_8$, $\hat{\beta}_{10}$, and the peak horizon.

(d) On the $h = 8$ estimation sample, compute each quarter's share of $\sum_t (s_t - \bar{s})^2$ and list the six largest.

(e) Re-estimate the $h = 8$ regression without 1950Q3. Does removing Korea's news quarter raise or lower the estimate?

(f) Draw the response graph and write its caption. Then write a paragraph of five sentences headed "What has not yet been argued": one sentence each on units, controls, identification, uncertainty, and dependence on a few episodes.

Hint

Date the news by the quarter in which it arrived, and the control one quarter earlier. Before running any loop, list 1950Q2 and 1950Q3 and check by hand that the 1950Q3 value of s_t divides that quarter's news by 1950Q2's nominal trend GDP.

Detailed solution

The do-file repeats the notes' first contact with data, starting from the raw workbook. It adds a check wherever a quarterly data set can go wrong: the date sequence, the timing of the scaling, the row count, and one coefficient that an independent source can confirm.

```
* ex09_first_contact.do  Lecture 01, Exercise 9: first contact with data
* Run from the lab project's top folder with StataNow/SE 19.5.
version 19
clear all
set more off
capture log close
log using output/ex09_first_contact.txt, text replace

* 1. Data: fetch the workbook or verify the copy already there (D5, D35)
do data/raw/get_rz.do "data/raw"
import excel using data/raw/RZDAT.xlsx, sheet("rzdat") firstrow clear
drop if quarter < 1889
generate qdate = yq(1889, 1) + _n - 1
format qdate %tq
tsset qdate, quarterly
assert _N == 508
assert abs(quarter - (1889 + (_n - 1)/4)) < 1e-6   // consecutive quarters

* 2. Outcome, intervention, and (for the check in step 6) spending:
*    y = real GDP / trend GDP, newsy = news / lagged nominal trend GDP,
*    g = real government spending / trend GDP
generate double y     = rgdp/rgdp_pott6
generate double newsy = news/(L.rgdp_pott6*L.pgdp)
generate double g     = (ngov/pgdp)/rgdp_pott6
quietly count if newsy < .
assert r(N) == 504
quietly count if newsy != 0 & newsy < .
display "quarters with nonzero news: " r(N)
summarize y newsy

* 3. One regression per horizon, h = 0,...,20
tempname P
```

```
postfile `P' h double beta_h T_h using output/first_contact, replace
forvalues h = 0/20 {
    quietly regress F`h'.y newsy L.y
    post `P' (`h') (_b[newsy]) (e(N))
}
postclose `P'

* 4. Where the variation comes from: shares in the h = 8 sample
quietly regress F8.y newsy L.y
generate byte in8 = e(sample)
quietly summarize newsy if in8
generate double dev2 = (newsy - r(mean))^2 if in8
quietly summarize dev2
generate double share = dev2/r(sum)
gsort -share
generate double cumshare = sum(share)
format newsy share cumshare %9.6f
list qdate newsy share cumshare in 1/6, clean noobs
assert qdate[1] == tq(1941q4) & abs(share[1] - 0.261211) < 1e-6
assert qdate[2] == tq(1950q3) & abs(share[2] - 0.195838) < 1e-6
sort qdate

* 5. Drop one episode at h = 8
quietly regress F8.y newsy L.y if qdate != tq(1950q3)
display "h = 8 without 1950Q3: " %9.6f _b[newsy] "  (T_8 = " e(N) ")"
assert e(N) == 495 & abs(_b[newsy] - 0.433969) < 1e-6

* 6. Check against a source record: Ramey and Zubairy's controls at h = 10
*    equal irf_gdp_linear in the REP04 benchmark (their jordagk.do)
quietly regress F10.y newsy L(1/4).newsy L(1/4).y L(1/4).g
display "h = 10 with RZ controls: " %10.8f _b[newsy] "  (N = " e(N) ")"
assert e(N) == 490 & abs(_b[newsy] - 0.29376385) < 1e-6
```

```
* 7. The response, its checks, and the graph
use output/first_contact, clear
assert T_h == 504 - h
assert abs(beta_h[9]  - 0.361906) < 1e-6        // h = 8
assert abs(beta_h[11] - 0.404210) < 1e-6        // h = 10
quietly summarize beta_h
assert beta_h[11] == r(max)                     // the peak is at h = 10
format beta_h %9.6f
list h T_h beta_h, clean noobs

twoway connected beta_h h, lcolor(navy) mcolor(navy) msize(small)      ///
       yline(0, lcolor(gs13)) xlabel(0(4)20)                           ///
       ylabel(0(0.1)0.5, format(%3.1f) angle(0))                       ///
       xtitle("Quarters after the news arrives")                       ///
       ytitle("Percent of trend GDP, per news worth 1 percent")        ///
       graphregion(color(white)) plotregion(color(white))
graph export output/ex09_first_contact.pdf, replace
display "Exercise 9: all assertions passed"
log close
```

In the verification run the workbook was already in `data/raw`, and `get_rz.do` reported "Ramey-Zubairy data present in data/raw and verified (SHA-256)." Every later assertion passed.

(a)–(b) The sheet holds 564 rows; dropping those before 1889 leaves 508 consecutive quarters, 1889Q1–2015Q4, and the date assertion confirms that no quarter is missing or repeated. The log's summary gives y_t a mean of 0.9868665 over 508 quarters and s_t a mean of 0.0072219 and standard deviation 0.0597282 over 504 quarters, 108 of them nonzero. The four missing values of s_t are the 1889 quarters, where `news` is missing. A unit of s_t is news worth a full year of nominal trend GDP, so the natural reading unit is 0.01, news worth one percent of trend GDP.

(c) The assertion $T_h = 504 - h$ passes. The formula $T_h = T - h - p$ with $T = 504$ and $p = 1$ would give $503 - h$, but it counts a lag as a lost row only when the lagged variable begins where the sample does. Here the first row with news is 1890Q1, and y_{t-1} exists there because GDP starts four quarters earlier. The lag costs nothing, and each horizon still costs one row at the end. The estimates rise from $\hat{\beta}_0 = 0.080313$ to

$\hat{\beta}_4 = 0.259141$ and $\hat{\beta}_8 = 0.361906$, peak at $\hat{\beta}_{10} = 0.404210$, and fall to $\hat{\beta}_{20} = 0.079552$. News worth one percent of trend GDP moves s_t by 0.01 and the prediction for y_{t+h} by $0.01\hat{\beta}_h$; since y is a ratio to trend GDP, that is $\hat{\beta}_h$ percent of trend GDP. Ten quarters after such news, GDP is about 0.40 percent of trend GDP above what the regression otherwise predicts from the previous quarter's output.

(d) The six largest shares of $\sum_t (s_t - \bar{s})^2$ in the $h = 8$ sample (496 quarters, 1890Q1–2013Q4) are, as the log lists them:

qdate	newsy	share	cumshare
1941q4	0.691893	0.261211	0.261211
1950q3	0.600073	0.195838	0.457049
1942q3	0.432791	0.100897	0.557946
1950q4	0.402163	0.086893	0.644839
1917q2	0.372915	0.074496	0.719335
1941q2	0.340130	0.061734	0.781069

Two quarters supply 45.7 percent of the variation and six supply 78.1 percent. The shares use $s_t - \bar{s}$, as in the notes' slope formula; with y_{t-1} in the regression the exact weights use s_t purged of y_{t-1}, which is close but not identical.

(e) Without 1950Q3 the $h = 8$ estimate is 0.433969 on 495 rows, against 0.361906 on the full sample. Removing Korea's news quarter *raises* the estimate by 0.072. Measured against the regression's average relation, what followed the Korean news eight quarters later was smaller than what followed the other large episodes, so that quarter was pulling the slope down.

Check against a source record. With Ramey and Zubairy's own controls (four lags each of the news, y, and government spending relative to trend), the $h = 10$ coefficient is 0.29376385 on 490 rows. That equals, at 10^{-6}, the linear GDP response `irf_gdp_linear` at $h = 10$ in the course's REP04 benchmark. The benchmark comes from an unmodified run of the authors' `jordagk.do`, lines 253–386 and 398–417 of the frozen package. The data construction therefore matches the authors' to the digits shown, and the difference from 0.404 comes from the controls, not from the data.

(f) The exported graph is a hump: 0.08 at impact, a peak of 0.40 at ten quarters, and 0.08 again at twenty. A caption that states what the reader needs:

> Response of real GDP, in percent of trend GDP, to military-spending news worth one percent of the previous quarter's nominal trend GDP. Local projections of y_{t+h} on a constant, the news, and y_{t-1}, 1890Q1–2015Q4, with $T_h = 504 - h$ rows at horizon h. Horizons are quarters after

the news arrives. Read causally, the vertical axis is the gap between GDP and the path it would have followed without the news, a reading not yet argued. No bands are shown.

A paragraph that lists, without defending, what the graph has not earned:

> *What has not yet been argued.* The coefficients are responses of a level relative to a fitted trend per unit of scaled news, and whether that level, its change, or its accumulation answers the policymaker's question is not yet settled. One lag of GDP is a choice, and Ramey and Zubairy's controls lower the ten-quarter coefficient from 0.404 to 0.294. Reading the curve as the effect of the news requires the news to be unpredictable and unrelated to the taxes, monetary policy, price controls, and wars that arrived with it, which nothing here has shown. No standard error or band has been computed, and the overlapping horizons make the residuals serially correlated. Two quarters, 1941Q4 and 1950Q3, supply 45.7 percent of the variation in the news, and dropping 1950Q3 alone moves the eight-quarter estimate from 0.362 to 0.434.

Each sentence points to a later lecture: units to Lecture 2, controls and identification to Lectures 3 and 4, uncertainty to Lectures 5 and 6, and episode dependence to Lecture 13.

Two mistakes are predictable. The first is a timing slip in the scaling: dividing the news by the *same* quarter's nominal trend GDP, or lagging the news itself instead of its denominator. Either changes the variable the notes define, the second also dates every episode a quarter late, and the hand check on 1950Q3 in the hint catches both. The second mistake is to assert $T_h = 503 - h$ from the formula and then "fix" the data until it passes, which throws away a quarter of real data. The formula counts rows that are missing; it does not decide which rows should be.

Exercise 10: Korea is not the average

Problem. [extra, pencil] In the Ramey–Zubairy data, GDP relative to trend was $y = 0.972500$ in 1950Q2, and the news in 1950Q3 was $s = 0.600073$ of lagged nominal trend GDP. Later values of y were 1.000572 in 1950Q3, 1.033931 in 1951Q3, and 1.051212 in 1953Q1. The first-contact local projection of Exercise 9 gives $\hat{\beta}_0 = 0.080313$, $\hat{\beta}_4 = 0.259141$, and $\hat{\beta}_{10} = 0.404210$.

(a) For $h = 0, 4, 10$, compute the realized change $y_{1950Q3+h} - y_{1950Q2}$ and the scaled coefficient $0.600073\,\hat{\beta}_h$, both in units of trend GDP.

(b) Which of the two numbers, if either, measures a gap between two paths?

(c) Give three distinct reasons why the two numbers differ, and say what follows if both are taken at face value.

Hint

Write each number as a difference between two outcomes, and label each outcome with its date and with the path it lies on, realized or counterfactual. For (c), look for reasons on both sides: in what the realized change contains, and in what the coefficient averages and assumes.

Detailed solution

(a) Subtracting and multiplying,

h	Quarter	y	Realized change	$\hat{\beta}_h$	$0.600073\,\hat{\beta}_h$	Ratio
0	1950Q3	1.000572	0.0281	0.080313	0.0482	1.72
4	1951Q3	1.033931	0.0614	0.259141	0.1555	2.53
10	1953Q1	1.051212	0.0787	0.404210	0.2426	3.08

The changes and products are rounded to four decimals, which gives the same digits whether they are computed from the six-decimal inputs or from the unrounded data. From the unrounded data, the realized changes are 0.028072, 0.061430, and 0.078712, and the scaled coefficients 0.048194, 0.155503, and 0.242556 (six decimals); the ratio column uses those values. Both columns are in units of trend GDP, so the last realized change is the 7.87 points of the notes' opening, and the scaled coefficient is about three times larger.

(b) The realized change is not a gap between paths. It compares one path with itself at two dates, and by the decomposition in the notes' first section it equals the causal response plus the change GDP would have made without the news:

$$y_{t+h} - y_{t-1} = \theta_h\, s_t + (y^c_{t+h} - y_{t-1}).$$

The scaled coefficient estimates $0.600073\,\beta_h$, and that equals a gap, $\theta_h\, s_t$, only under conditions this lecture has not established for the news. The news would have to be an observed shock, its effects would have to be linear and the same in every episode, and one lag of GDP would have to be an adequate control.

(c) Three distinct reasons:

1. *Everything else in 1950–1953.* The realized change contains the counterfactual change: tax increases, the Treasury–Federal Reserve Accord, wage and price controls, and the war itself apart from its

spending news. A path comparison removes those; a change along one path keeps them.

2. *An average applied to one episode.* $\hat{\beta}_{10}$ averages over every quarter with news, weighted by $(s_t - \bar{s})^2$; in the eight-quarter sample of Exercise 9, 1941Q4 alone carries 26 percent of that weight. Exercise 9 also found that dropping 1950Q3 raises the eight-quarter estimate, so eight quarters after its news Korea's own outcome sat below the average relation. Multiplying the average by Korea's news also assumes that news worth 60 percent of trend GDP has sixty times the effect of news worth one percent, whatever the state of the economy (A1).
3. *Specification and identification.* The coefficient depends on the controls: with Ramey and Zubairy's, $0.600073 \times 0.29376385 = 0.1763$ (rounded), already a quarter lower. And the reading of any coefficient as θ_h rests on an identification argument that Lecture 3 has yet to make. Neither number carries a measure of uncertainty.

Taken at face value together, the decomposition says that without the news GDP relative to trend would have *fallen* by $0.242556 - 0.078712 = 0.163844$ between 1950Q2 and 1953Q1, about sixteen points of trend GDP (unrounded values). Nobody has argued that, and the tension is a reason to doubt at least one of the maintained assumptions rather than to pick the more convenient number.

Watch for a units slip: scaling by 60.0073 because the news is "60 percent" of trend GDP. That gives 24.26 at $h = 10$, a number one hundred times too large, because s_t is measured as a fraction, not in percent. A second slip is to conclude that the regression overstates the effect threefold. Neither column is the effect of Korea's news.