Autocorrelation

EC 421, Set 8

Edward Rubin

Winter 2022

Prologue

Schedule

Last Time

Introduction to time series

Today

Autocorrelation

Upcoming

- Midterm project due on Sunday (before midnight).
- Assignment 3 soon.

Functions

Writing your own functions.

Writing functions

Functions are everywhere

Everything you do in `R` involves some sort of function, *e.g.*,

- `mean()`
- `lm()`
- `summary()`
- `read_csv()`
- `ggplot()`
- `+`

The basic idea in `R` is doing things to objects with functions.

Writing functions

Functions can help

We write functions to make life easier. Instead of copying and pasting the same line of code a million times, you can write one function.

In R, you use the `function()` function to write functions.[†]

```
# Our first function
the_name <- function(arg1, arg2) {
  # Insert code that involves arg1 and arg2 (this is where the magic happens)
}
```

- `the_name`: The name we are giving to our new function.
- `arg1`: The first argument of our function.
- `arg2`: The second argument of our function.

† Meta.

Writing functions

Our first real function

Let's write a function that multiplies two numbers. (It needs two arguments.)

```
# Create our function
the_product ← function(x, y) {
  x * y
}
```

Our first real function

Let's write a function that multiplies two numbers. (It needs two arguments.)

```r
# Create our function
the_product <- function(x, y) {
  x * y
}
```

Did it work?

Writing functions

Our first real function

Let's write a function that multiplies two numbers. (It needs two arguments.)

```r
# Create our function
the_product <- function(x, y) {
  x * y
}
```

Did it work?

```r
the_product(7, 15)
```

```
#> [1] 105
```

💪

Writing functions

Functions can do anything

… that you tell them.

If you are going to repeat a task (*e.g.*, a simulation), then you have a good situation for writing your own function.

R offers many functions (via its many packages), but you will sometimes find a scenario for which no one has written a function.

Now you know how to write your own.

```
# An ad lib function
ad_lib <- function(noun1, verb1, noun2) {
  paste("The next", noun1, "of our lecture", verb1, noun2)
}
```

Writing functions

```r
ad_lib(noun1 = "part", verb1 = "reviews", noun2 = "time series.")
```

```
#> [1] "The next part of our lecture reviews time series."
```

Time series

Review

Time series

Review

Changes to our model/framework.

- Our model now has t subscripts for **time periods**.
- **Dynamic models** allow **lags** of explanatory and/or outcome variables.
- We changed our **exogeneity** assumption to **contemporaneous exogeneity**, *i.e.*, $\boldsymbol{E}[u_t|X_t] = 0$
- Including **lags of outcome variables** can lead to **biased coefficient estimates** from OLS.
- **Lagged explanatory variables** make **OLS inefficient**.

Autocorrelation

Autocorrelation

What is it?

Autocorrelation occurs when our disturbances are correlated over time, *i.e.*, $\text{Cov}(u_t,\, u_s) \neq 0$ for $t \neq s$.

Autocorrelation

What is it?

Autocorrelation occurs when our disturbances are correlated over time, *i.e.*, $\mathrm{Cov}(u_t,\, u_s) \neq 0$ for $t \neq s$.

Another way to think about: If the *shock* from disturbance t correlates with "nearby" shocks in $t-1$ and $t+1$.

Autocorrelation

What is it?

Autocorrelation occurs when our disturbances are correlated over time, *i.e.*, $\mathrm{Cov}(u_t,\, u_s) \neq 0$ for $t \neq s$.

Another way to think about: If the *shock* from disturbance t correlates with "nearby" shocks in $t-1$ and $t+1$.

Note: **Serial correlation** and **autocorrelation** are the same thing.

Autocorrelation

What is it?

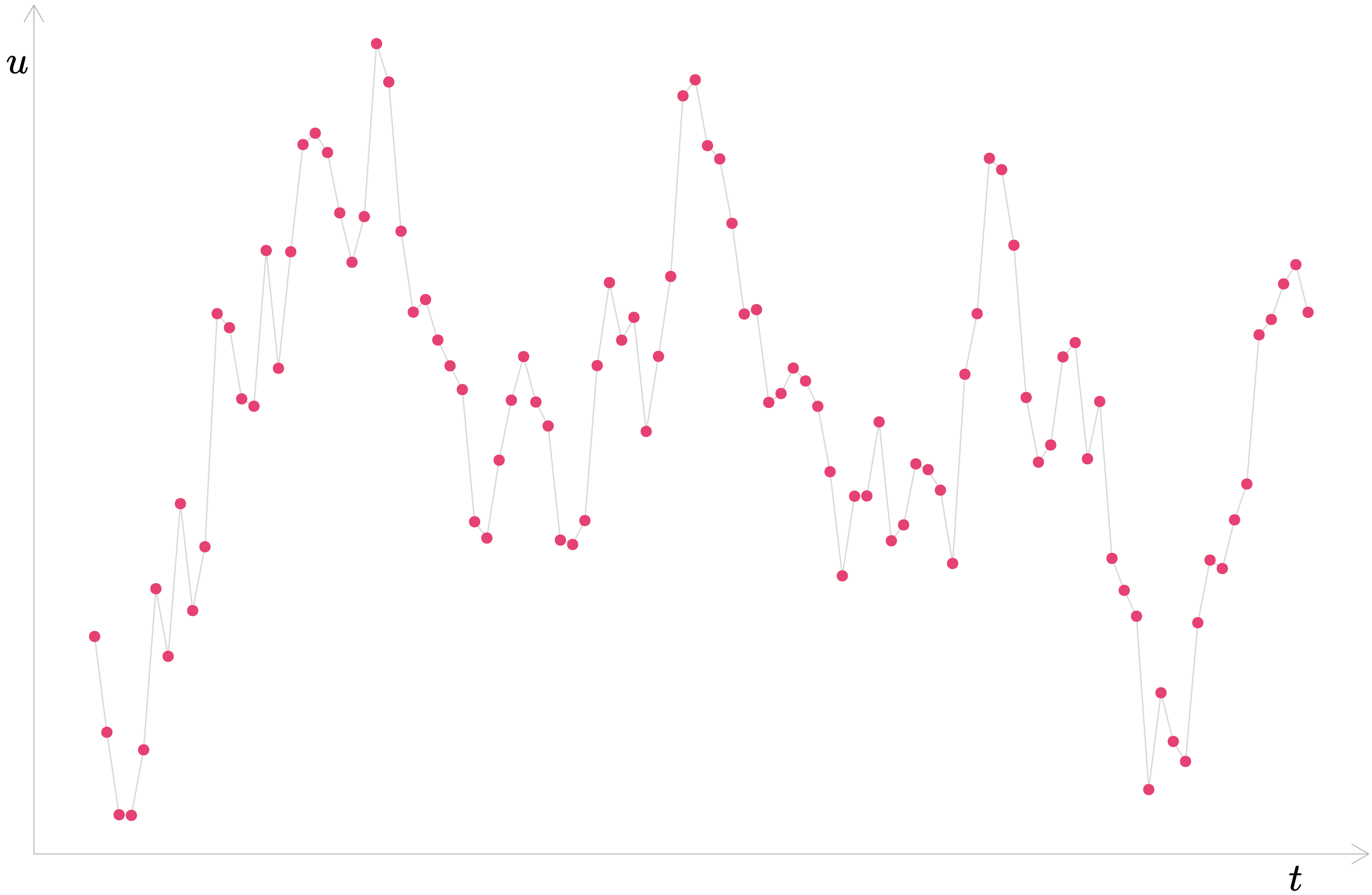
Autocorrelation occurs when our disturbances are correlated over time, *i.e.*, $\text{Cov}(u_t,\, u_s) \neq 0$ for $t \neq s$.

Another way to think about: If the *shock* from disturbance t correlates with "nearby" shocks in $t-1$ and $t+1$.

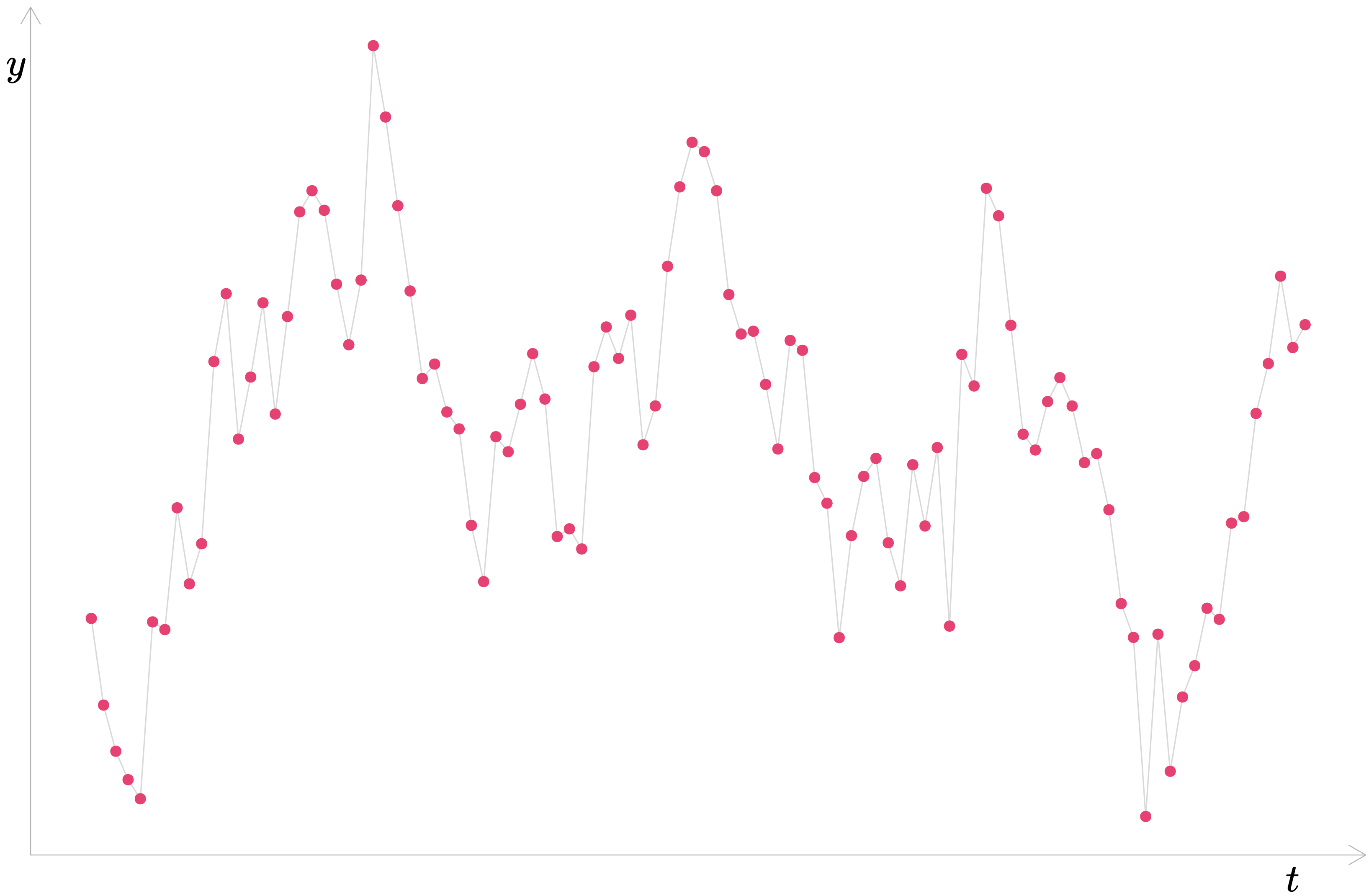
Note: **Serial correlation** and **autocorrelation** are the same thing.

Why is autocorrelation prevalent in time-series analyses?

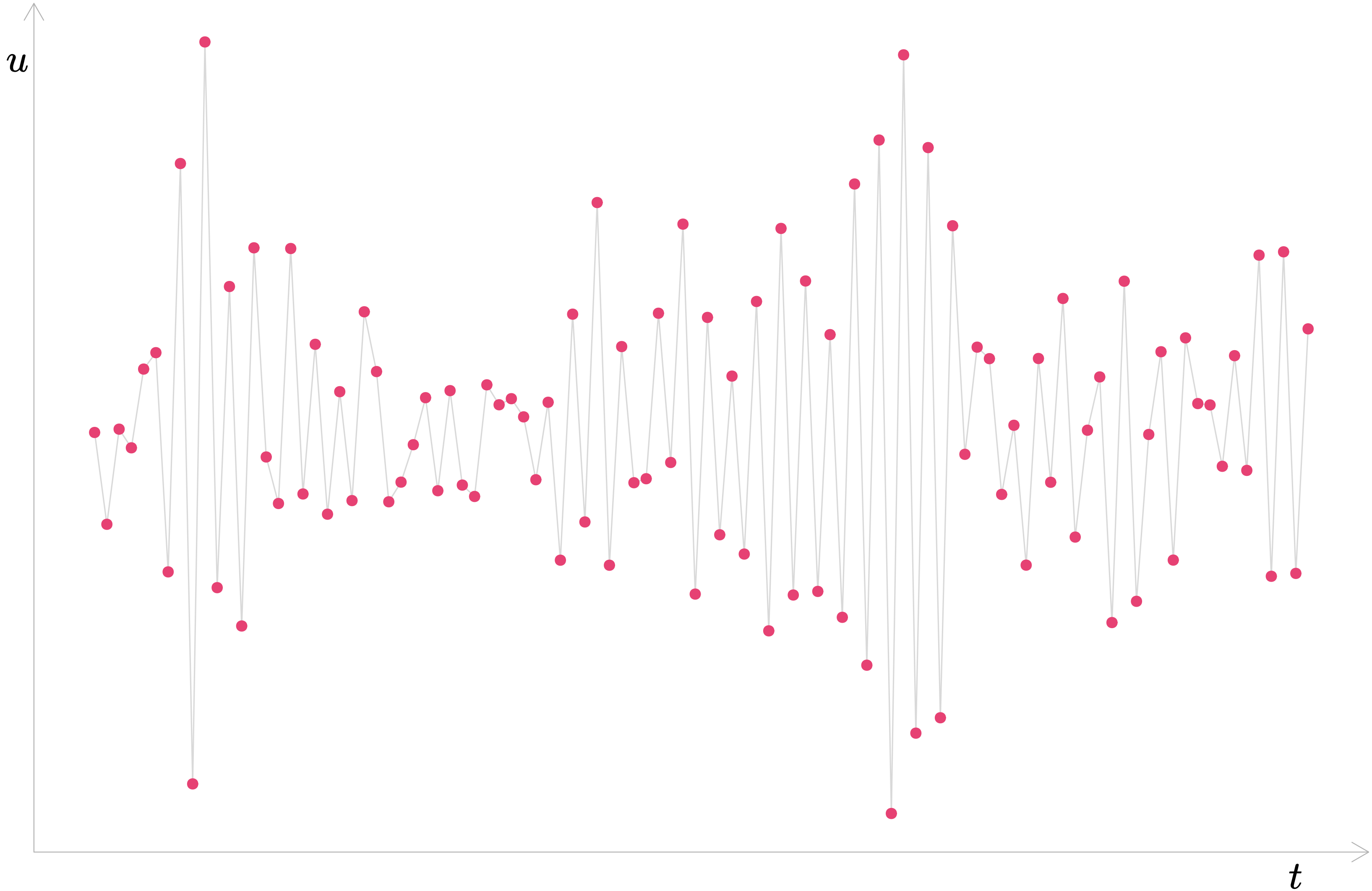
Positive autocorrelation: Disturbances (u_t) over time



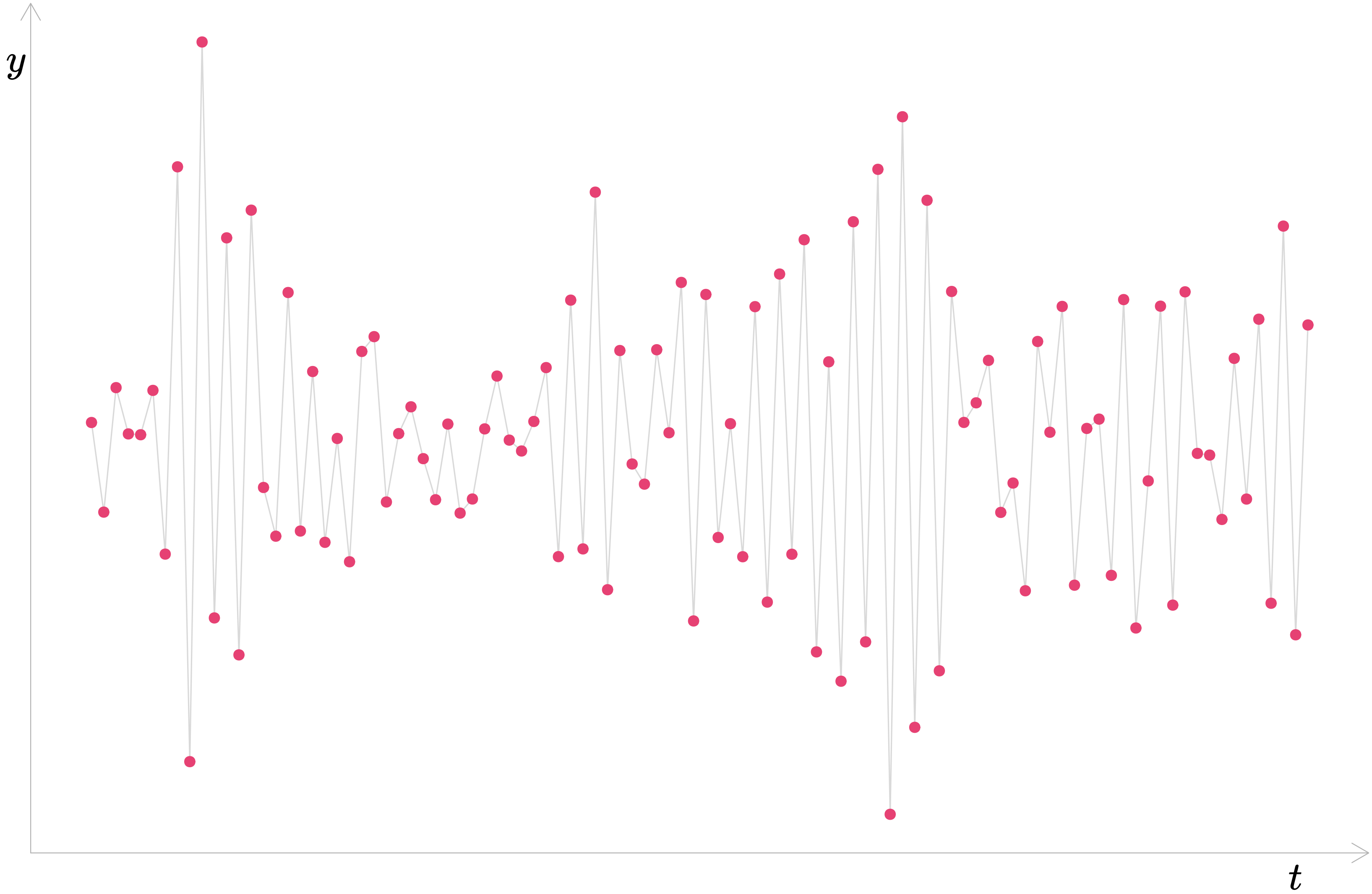
Positive autocorrelation: Outcomes (y_t) over time



Negative autocorrelation: Disturbances (u_t) over time



Negative autocorrelation: Outcomes (y_t) over time



Autocorrelation

In static time-series models

Let's start with a very common model: a static time-series model whose disturbances exhibit **first-order autocorrelation**, *a.k.a.* AR(1):

$$\text{Births}_t = \beta_0 + \beta_1 \text{Income}_t + u_t$$

where

$$u_t = \rho\, u_{t-1} + \varepsilon_t$$

and the ε_t are independently and identically distributed (*i.i.d.*).

Autocorrelation

In static time-series models

Let's start with a very common model: a static time-series model whose disturbances exhibit **first-order autocorrelation**, *a.k.a.* AR(1):

$$\text{Births}_t = \beta_0 + \beta_1 \text{Income}_t + u_t$$

where

$$u_t = \rho\, u_{t-1} + \varepsilon_t$$

and the ε_t are independently and identically distributed (*i.i.d.*).

Second-order autocorrelation, or AR(2), would be

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \varepsilon_t$$

Autocorrelation

In static time-series models

An AR(p) model/process has a disturbance structure of

$$u_t = \sum_{j=1}^{p} \rho_j u_{t-j} + \varepsilon_t$$

allowing the current disturbance to correlated with up to p of its lags.

Autocorrelation

OLS

For **static models** or **dynamic models with lagged explanatory variables**, in the presence of autocorrelation

1. OLS provides **unbiased** estimates for the coefficients.
2. OLS creates **biased** estimates for the standard errors.
3. OLS is **inefficient**.

Recall: Same implications as heteroskedasticity.

Autocorrelation get trickier with lagged outcome variables.

OLS and lagged outcome variables

Consider a model with one lag of the outcome variable—ADL(1, 0)—model with AR(1) disturbances

$$\text{Births}_t = \beta_0 + \beta_1 \text{Income}_t + \beta_2 \text{Births}_{t-1} + u_t$$

where

$$u_t = \rho u_{t-1} + \varepsilon_t$$

OLS and lagged outcome variables

Consider a model with one lag of the outcome variable—ADL(1, 0)—model with AR(1) disturbances

$$\text{Births}_t = \beta_0 + \beta_1 \text{Income}_t + \beta_2 \text{Births}_{t-1} + u_t$$

where

$$u_t = \rho u_{t-1} + \varepsilon_t$$

Problem:

OLS and lagged outcome variables

Consider a model with one lag of the outcome variable—ADL(1, 0)—model with AR(1) disturbances

$$\text{Births}_t = \beta_0 + \beta_1 \text{Income}_t + \beta_2 \text{Births}_{t-1} + u_t$$

where

$$u_t = \rho u_{t-1} + \varepsilon_t$$

Problem: Both Births_{t-1} (a regressor in the model for time t) and u_t (the disturbance for time t) depend upon u_{t-1}. *I.e.*, a regressor is correlated with its contemporaneous disturbance.

OLS and lagged outcome variables

Consider a model with one lag of the outcome variable—ADL(1, 0)—model with AR(1) disturbances

$$\text{Births}_t = \beta_0 + \beta_1 \text{Income}_t + \beta_2 \text{Births}_{t-1} + u_t$$

where

$$u_t = \rho u_{t-1} + \varepsilon_t$$

Problem: Both Births_{t-1} (a regressor in the model for time t) and u_t (the disturbance for time t) depend upon u_{t-1}. *I.e.*, a regressor is correlated with its contemporaneous disturbance.

Q: Why is this a problem?

OLS and lagged outcome variables

Consider a model with one lag of the outcome variable—ADL(1, 0)—model with AR(1) disturbances

$$\text{Births}_t = \beta_0 + \beta_1 \text{Income}_t + \beta_2 \text{Births}_{t-1} + u_t$$

where

$$u_t = \rho u_{t-1} + \varepsilon_t$$

Problem: Both Births_{t-1} (a regressor in the model for time t) and u_t (the disturbance for time t) depend upon u_{t-1}. *I.e.*, a regressor is correlated with its contemporaneous disturbance.

Q: Why is this a problem?
A: It violates contemporaneous exogeneity

OLS and lagged outcome variables

Consider a model with one lag of the outcome variable—ADL(1, 0)—model with AR(1) disturbances

$$\text{Births}_t = \beta_0 + \beta_1 \text{Income}_t + \beta_2 \text{Births}_{t-1} + u_t$$

where

$$u_t = \rho u_{t-1} + \varepsilon_t$$

Problem: Both Births_{t-1} (a regressor in the model for time t) and u_t (the disturbance for time t) depend upon u_{t-1}. *I.e.*, a regressor is correlated with its contemporaneous disturbance.

Q: Why is this a problem?

A: It violates contemporaneous exogeneity, *i.e.*, $\text{Cov}(x_t,\, u_t) \neq 0$.

OLS and lagged outcome variables

To see this problem, first write out the model for t and $t-1$:

$$\begin{aligned} \text{Births}_t &= \beta_0 + \beta_1 \text{Income}_t + \beta_2 \text{Births}_{t-1} + u_t \\ \text{Births}_{t-1} &= \beta_0 + \beta_1 \text{Income}_{t-1} + \beta_2 \text{Births}_{t-2} + u_{t-1} \end{aligned}$$

and now note that $u_t = \rho u_{t-1} + \varepsilon_t$. Substituting...

$$\text{Births}_t = \beta_0 + \beta_1 \text{Income}_t + \beta_2 \text{Births}_{t-1} + \overbrace{\left(\rho u_{t-1} + \varepsilon_t\right)}^{u_t} \tag{1}$$

$$\text{Births}_{t-1} = \beta_0 + \beta_1 \text{Income}_{t-1} + \beta_2 \text{Births}_{t-2} + u_{t-1} \tag{2}$$

In (1), we can see that u_t depends upon (covaries with) u_{t-1}.

In (2), we can see that Births_{t-1}, a regressor in (1), also covaries with u_{t-1}.

OLS and lagged outcome variables

To see this problem, first write out the model for t and $t-1$:

$$\begin{aligned} \text{Births}_t &= \beta_0 + \beta_1 \text{Income}_t + \beta_2 \text{Births}_{t-1} + u_t \\ \text{Births}_{t-1} &= \beta_0 + \beta_1 \text{Income}_{t-1} + \beta_2 \text{Births}_{t-2} + u_{t-1} \end{aligned}$$

and now note that $u_t = \rho u_{t-1} + \varepsilon_t$. Substituting...

$$\begin{aligned} \text{Births}_t &= \beta_0 + \beta_1 \text{Income}_t + \beta_2 \text{Births}_{t-1} + \overbrace{\left(\rho u_{t-1} + \varepsilon_t\right)}^{u_t} && (1) \\ \text{Births}_{t-1} &= \beta_0 + \beta_1 \text{Income}_{t-1} + \beta_2 \text{Births}_{t-2} + u_{t-1} && (2) \end{aligned}$$

In (1), we can see that u_t depends upon (covaries with) u_{t-1}.

In (2), we can see that Births_{t-1}, a regressor in (1), also covaries with u_{t-1}.

$\therefore$ This model violates our contemporaneous exogeneity requirement.

OLS and lagged outcome variables

Implications: For models with **lagged outcome variables** and **autocorrelated disturbances**

1. The models **violate contemporaneous exogeneity**.

2. OLS is **biased and inconsistent** for the coefficients.

OLS and lagged outcome variables

Intuition? Why is OLS inconsistent and biased when we violate exogeneity?

Think back to omitted-variable bias...

$$y_t = \beta_0 + \beta_1 x_t + u_t$$

When $\mathrm{Cov}(x_t,\, u_t) \neq 0$, we cannot separate the effect of u_t on y_t from the effect of x_t on y_t. Thus, we get inconsistent estimates for β_1.

OLS and lagged outcome variables

Intuition? Why is OLS inconsistent and biased when we violate exogeneity?

Think back to omitted-variable bias...

$$y_t = \beta_0 + \beta_1 x_t + u_t$$

When $\mathrm{Cov}(x_t,\, u_t) \neq 0$, we cannot separate the effect of u_t on y_t from the effect of x_t on y_t. Thus, we get inconsistent estimates for β_1. Similarly,

$$\text{Births}_t = \beta_0 + \beta_1 \text{Income}_t + \beta_2 \text{Births}_{t-1} + \overbrace{(\rho u_{t-1} + \varepsilon_t)}^{u_t} \tag{1}$$

we cannot separate the effects of u_t on Births_t from Births_{t-1} on Births_t, because both u_t and Births_{t-1} depend upon u_{t-1}.

OLS and lagged outcome variables

Intuition? Why is OLS inconsistent and biased when we violate exogeneity?

Think back to omitted-variable bias...

$$y_t = \beta_0 + \beta_1 x_t + u_t$$

When $\mathrm{Cov}(x_t,\, u_t) \neq 0$, we cannot separate the effect of u_t on y_t from the effect of x_t on y_t. Thus, we get inconsistent estimates for β_1. Similarly,

$$\text{Births}_t = \beta_0 + \beta_1 \text{Income}_t + \beta_2 \text{Births}_{t-1} + \overbrace{(\rho u_{t-1} + \varepsilon_t)}^{u_t} \tag{1}$$

we cannot separate the effects of u_t on Births_t from Births_{t-1} on Births_t, because both u_t and Births_{t-1} depend upon u_{t-1}. $\hat{\beta}_2$ is **biased** (w/ OLS).

Autocorrelation and bias

Simulation

To see how this bias can look, let's run a simulation.

$$y_t = 1 + 2x_t + 0.5y_{t-1} + u_t$$
$$u_t = 0.9u_{t-1} + \varepsilon_t$$

One (easy) way generate 100 disturbances from AR(1), with $\rho = 0.9$:

```
arima.sim(model = list(ar = c(0.9)), n = 100)
```

Autocorrelation and bias

Simulation

To see how this bias can look, let's run a simulation.

$$
\begin{aligned}
y_t &= 1 + 2x_t + 0.5y_{t-1} + u_t \\
u_t &= 0.9u_{t-1} + \varepsilon_t
\end{aligned}
$$

One (easy) way generate 100 disturbances from AR(1), with $\rho = 0.9$:

```
arima.sim(model = list(ar = c(0.9)), n = 100)
```

We are going to run 10,000 iterations with $T = 100$.

Autocorrelation and bias

Simulation

To see how this bias can look, let's run a simulation.

$$
\begin{aligned}
y_t &= 1 + 2x_t + 0.5y_{t-1} + u_t \\
u_t &= 0.9u_{t-1} + \varepsilon_t
\end{aligned}
$$

One (easy) way generate 100 disturbances from AR(1), with $\rho = 0.9$:

```
arima.sim(model = list(ar = c(0.9)), n = 100)
```

We are going to run 10,000 iterations with $T = 100$.

Q: Will this simulation tell us about *bias* or *consistency*?

Autocorrelation and bias

Simulation

To see how this bias can look, let's run a simulation.

$$y_t = 1 + 2x_t + 0.5y_{t-1} + u_t$$
$$u_t = 0.9u_{t-1} + \varepsilon_t$$

One (easy) way generate 100 disturbances from AR(1), with $\rho = 0.9$:

```
arima.sim(model = list(ar = c(0.9)), n = 100)
```

We are going to run 10,000 iterations with $T = 100$.

Q: Will this simulation tell us about *bias* or *consistency*?

A: Bias. We would need to let $T \to \infty$ to consider consistency.

Simulation

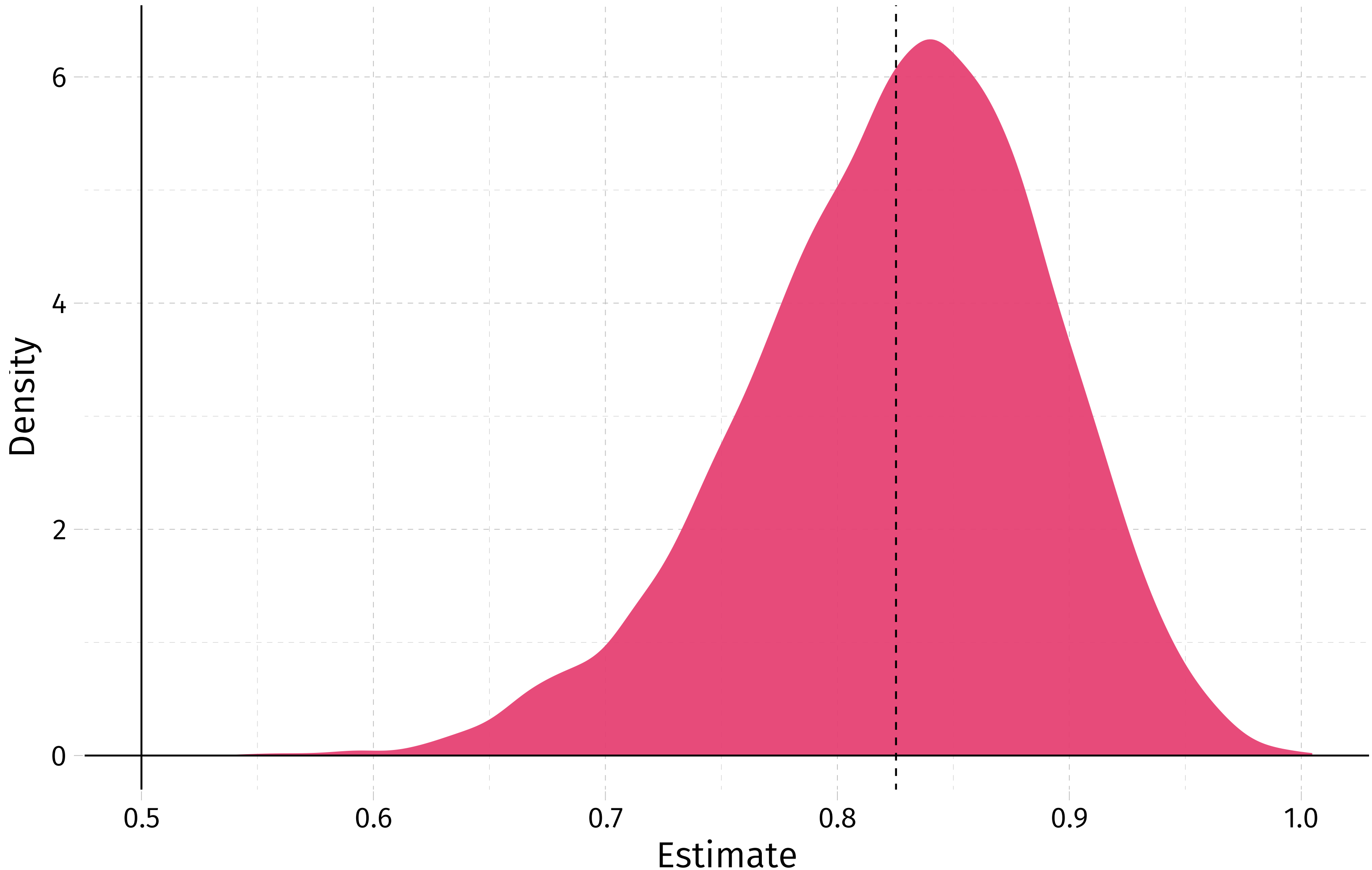
Outline of our simulation:

1. Generate T=100 values of x
2. Generate T=100 values of u
   - Generate T=100 values of ε
   - Use ε and ρ=0.9 to calculate $u_t = \rho\, u_{t-1} + \varepsilon_t$
3. Calculate $y_t = \beta_0 + \beta_1\, x_t + \beta_2\, y_{t-1} + u_t$
4. Regress y on x; record estimates

Repeat 1–4 10,000 times

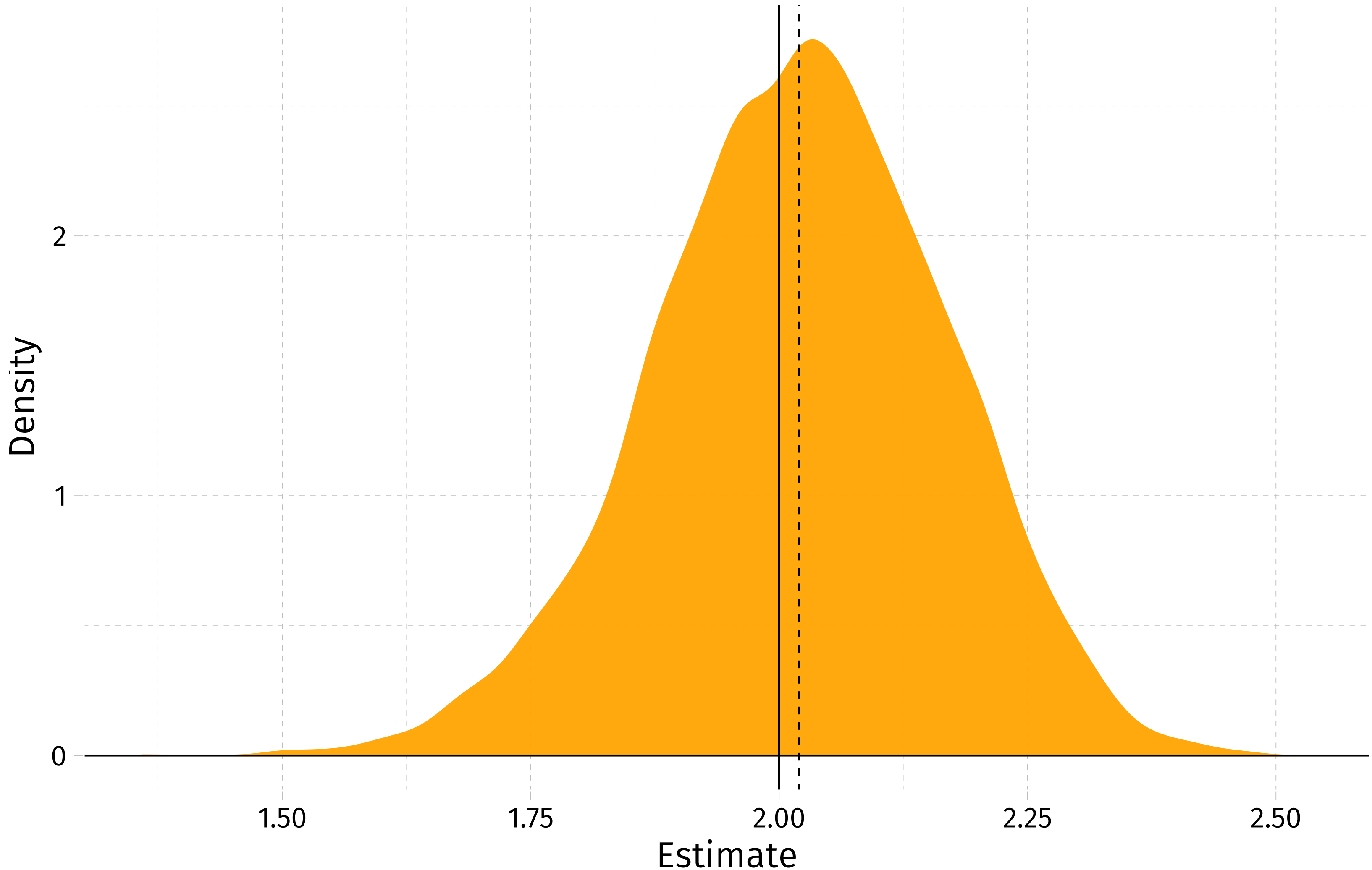
Distribution of OLS estimates, $\hat{\beta}_2$

$$y_t = 1 + 2x_t + 0.5y_{t-1} + u_t$$



Distribution of OLS estimates, $\hat{\beta}_1$

$$y_t = 1 + 2x_t + 0.5y_{t-1} + u_t$$



Testing for autocorrelation

Testing for autocorrelation

Static models

Suppose we have the **static model**,

$$\text{Births}_t = \beta_0 + \beta_1 \text{Income}_t + u_t \tag{A}$$

and we want to test for an AR(1) process in our disturbances u_t.

Testing for autocorrelation

Static models

Suppose we have the **static model**,

$$\text{Births}_t = \beta_0 + \beta_1 \text{Income}_t + u_t \tag{A}$$

and we want to test for an AR(1) process in our disturbances u_t.

Test for autocorrelation: Test for correlation in the lags of our residuals:

Testing for autocorrelation

Static models

Suppose we have the **static model**,

$$\text{Births}_t = \beta_0 + \beta_1 \text{Income}_t + u_t \tag{A}$$

and we want to test for an AR(1) process in our disturbances u_t.

Test for autocorrelation: Test for correlation in the lags of our residuals:

$$e_t = \rho e_{t-1} + v_t$$

Testing for autocorrelation

Static models

Suppose we have the **static model**,

$$\text{Births}_t = \beta_0 + \beta_1 \text{Income}_t + u_t \tag{A}$$

and we want to test for an AR(1) process in our disturbances u_t.

Test for autocorrelation: Test for correlation in the lags of our residuals:

$$e_t = \rho e_{t-1} + v_t$$

Does $\hat{\rho}$ differ significantly from zero?

Testing for autocorrelation

Static models

Suppose we have the **static model**,

$$\text{Births}_t = \beta_0 + \beta_1 \text{Income}_t + u_t \tag{A}$$

and we want to test for an AR(1) process in our disturbances u_t.

Test for autocorrelation: Test for correlation in the lags of our residuals:

$$e_t = \rho e_{t-1} + v_t$$

Does $\hat{\rho}$ differ significantly from zero?

Familiar idea: Use residuals to learn about disturbances.

Testing for autocorrelation

Static models

Specifically, to test for AR(1) disturbances in the static model

$$\text{Births}_t = \beta_0 + \beta_1 \text{Income}_t + u_t \tag{A}$$

1. Estimate $(\mathbf{A})$ via OLS.

2. Calculate residuals from the OLS regression in step 1.

3. Regress the residuals on their lags (without an intercept).

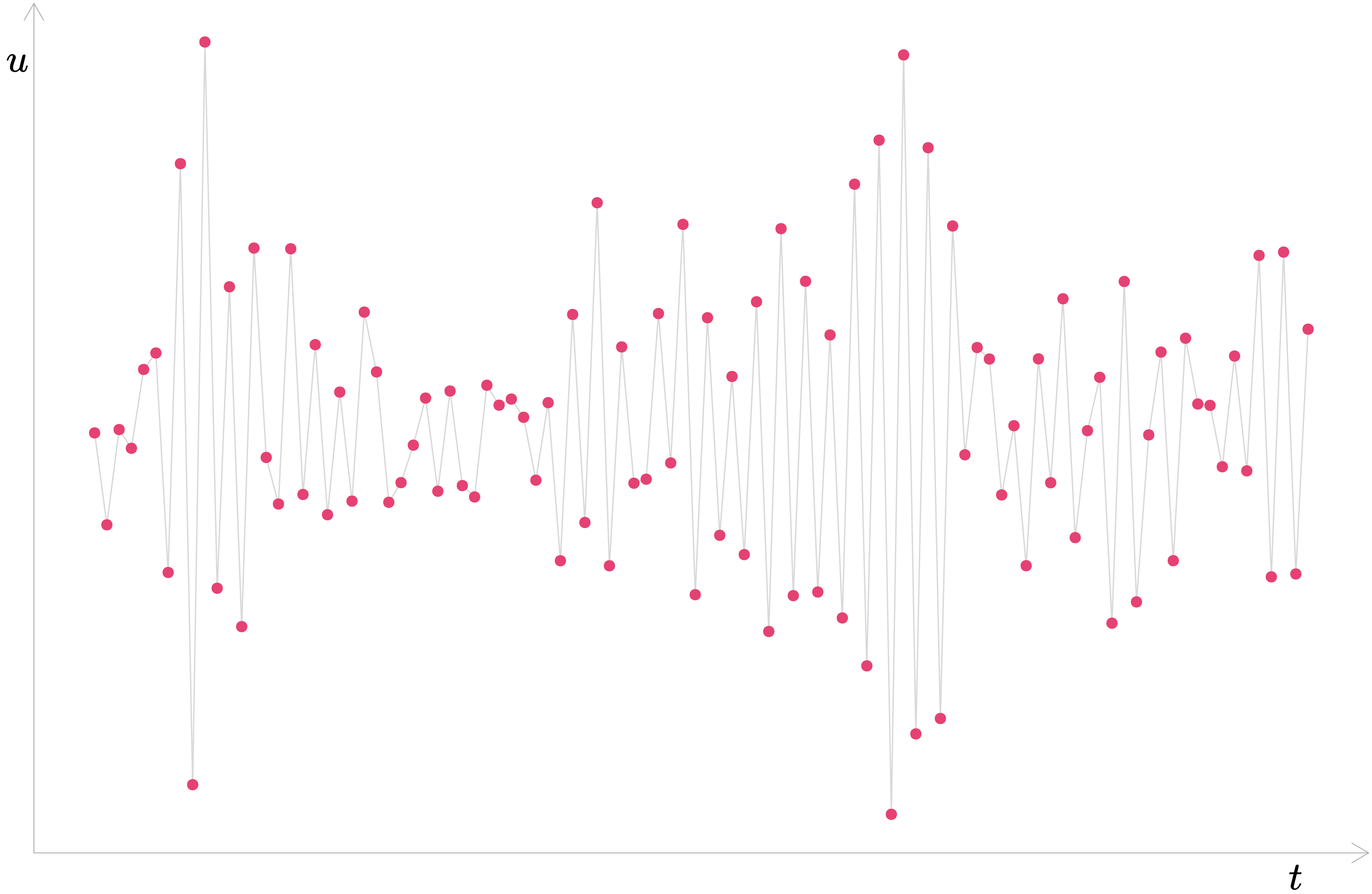
$$e_t = \rho \; e_{t-1} + v_t$$

4. Use a t test to determine whether there is statistically significant evidence that ρ differs from zero.

5. Rejecting H_0 implies significant evidence of autocorrelation.

For an example, let's return to our plot of negative autocorrelation.

Negative autocorrelation: Disturbances (u_t) over time



Testing for autocorrelation

Example: Static model and AR(1)

Step 1: Estimate the static model ($y_t = \beta_0 + \beta_1 x_t + u_t$) with OLS

```
reg_est ← lm(y ~ x, data = ar_df)
```

Testing for autocorrelation

Example: Static model and AR(1)

Step 1: Estimate the static model ($y_t = \beta_0 + \beta_1 x_t + u_t$) with OLS

```
reg_est <- lm(y ~ x, data = ar_df)
```

Step 2: Add the residuals to our dataset

```
ar_df$e <- residuals(reg_est)
```

Testing for autocorrelation

Example: Static model and AR(1)

Step 1: Estimate the static model ($y_t = \beta_0 + \beta_1 x_t + u_t$) with OLS

```
reg_est ← lm(y ~ x, data = ar_df)
```

Step 2: Add the residuals to our dataset

```
ar_df$e ← residuals(reg_est)
```

Step 3: Regress the residual on its lag (**no intercept**)

```
reg_resid ← lm(e ~ -1 + lag(e), data = ar_df)
```

Testing for autocorrelation

Example: Static model and AR(1)

Step 4: t test for the estimated ($\hat{\rho}$) coefficient in step 3.

```
tidy(reg_resid)
```

```
#> # A tibble: 1 × 5
#>   term    estimate std.error statistic  p.value
#>   <chr>      <dbl>     <dbl>     <dbl>    <dbl>
#> 1 lag(e)    -0.851    0.0535     -15.9 6.88e-29
```

Example: Static model and AR(1)

Step 4: t test for the estimated $(\hat{\rho})$ coefficient in step 3.

```
tidy(reg_resid)
```

```
#> # A tibble: 1 × 5
#>   term   estimate std.error statistic  p.value
#>   <chr>     <dbl>     <dbl>     <dbl>    <dbl>
#> 1 lag(e)   -0.851    0.0535     -15.9 6.88e-29
```

That's a very small p-value—much smaller than 0.05.

Example: Static model and AR(1)

Step 4: t test for the estimated $(\hat{\rho})$ coefficient in step 3.

```
tidy(reg_resid)
```

```
#> # A tibble: 1 × 5
#>   term     estimate std.error statistic  p.value
#>   <chr>       <dbl>     <dbl>     <dbl>    <dbl>
#> 1 lag(e)     -0.851    0.0535     -15.9 6.88e-29
```

That's a very small *p*-value—much smaller than 0.05.

Reject H_0 (H_0 was $\rho = 0$, *i.e.*, no autocorrelation).

Testing for autocorrelation

Example: Static model and AR(1)

Step 4: t test for the estimated $(\hat{\rho})$ coefficient in step 3.

```
tidy(reg_resid)
```

```
#> # A tibble: 1 × 5
#>   term   estimate std.error statistic  p.value
#>   <chr>     <dbl>     <dbl>     <dbl>    <dbl>
#> 1 lag(e)   -0.851    0.0535     -15.9 6.88e-29
```

That's a very small *p*-value—much smaller than 0.05.

Reject H_0 (H_0 was $\rho = 0$, *i.e.*, no autocorrelation).

Step 5: Conclude. Statistically significant evidence of autocorrelation.

Example: Static model and AR(3)

What if we wanted to test for AR(3)?

- We add more lags of residuals to the regression in *Step 3*.

- We **jointly** test the significance of the coefficients (*i.e.*, LM or F).

Let's do it.

Testing for autocorrelation

Example: Static model and AR(3)

Step 1: Estimate the static model ($y_t = \beta_0 + \beta_1 x_t + u_t$) with OLS

```
reg_est ← lm(y ~ x, data = ar_df)
```

Testing for autocorrelation

Example: Static model and AR(3)

Step 1: Estimate the static model ($y_t = \beta_0 + \beta_1 x_t + u_t$) with OLS

```
reg_est ← lm(y ~ x, data = ar_df)
```

Step 2: Add the residuals to our dataset

```
ar_df$e ← residuals(reg_est)
```

Example: Static model and AR(3)

Step 1: Estimate the static model ($y_t = \beta_0 + \beta_1 x_t + u_t$) with OLS

```
reg_est ← lm(y ~ x, data = ar_df)
```

Step 2: Add the residuals to our dataset

```
ar_df$e ← residuals(reg_est)
```

Step 3: Regress the residual on its lag (**no intercept**)

```
reg_ar3 ← lm(e ~ -1 + lag(e) + lag(e, 2) + lag(e, 3), data = ar_df)
```

Testing for autocorrelation

Example: Static model and AR(3)

Step 1: Estimate the static model ($y_t = \beta_0 + \beta_1 x_t + u_t$) with OLS

```
reg_est ← lm(y ~ x, data = ar_df)
```

Step 2: Add the residuals to our dataset

```
ar_df$e ← residuals(reg_est)
```

Step 3: Regress the residual on its lag (**no intercept**)

```
reg_ar3 ← lm(e ~ -1 + lag(e) + lag(e, 2) + lag(e, 3), data = ar_df)
```

Note: `lag(v, n)` from `dplyr` takes the $\texttt{n}^{\text{th}}$ lag of the variable `v`.

Testing for autocorrelation

Example: Static model and AR(3)

Step 4: Calculate the $\text{LM} = n \times R_e^2$ test statistic—distributed χ_k^2.
k is the number of regressors in the regression in *Step 3* (here, $k = 3$).

```
# Grab R squared
r2_e <- summary(reg_ar3)$r.squared
# Calculate the LM test statistic: n times r2_e
(lm_stat <- 100 * r2_e)
```

```
#> [1] 72.38204
```

```
# Calculate the p-value
(pchisq(q = lm_stat, df = 3, lower.tail = F))
```

```
#> [1] 1.318485e-15
```

Testing for autocorrelation

Example: Static model and AR(3)

Step 5: Conclude.

Example: Static model and AR(3)

Step 5: Conclude.

Recall: Our hypotheses consider the model

$$e_t = \rho_1 e_{t-1} + \rho_2 e_{t-2} + \rho_3 e_{t-3}$$

Example: Static model and AR(3)

Step 5: Conclude.

Recall: Our hypotheses consider the model

$$e_t = \rho_1 e_{t-1} + \rho_2 e_{t-2} + \rho_3 e_{t-3}$$

which we are actually using to learn about the model

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \rho_3 u_{t-3}$$

Testing for autocorrelation

Example: Static model and AR(3)

Step 5: Conclude.

Recall: Our hypotheses consider the model

$$e_t = \rho_1 e_{t-1} + \rho_2 e_{t-2} + \rho_3 e_{t-3}$$

which we are actually using to learn about the model

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \rho_3 u_{t-3}$$

H_0: $\rho_1 = \rho_2 = \rho_3 = 0$ *vs.* H_A: $\rho_j \neq 0$ for at least one j in $\{1,\, 2,\, 3\}$

Our p-value is less than 0.05.

Testing for autocorrelation

Example: Static model and AR(3)

Step 5: Conclude.

Recall: Our hypotheses consider the model

$$e_t = \rho_1 e_{t-1} + \rho_2 e_{t-2} + \rho_3 e_{t-3}$$

which we are actually using to learn about the model

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \rho_3 u_{t-3}$$

H_0: $\rho_1 = \rho_2 = \rho_3 = 0$ *vs.* H_A: $\rho_j \neq 0$ for at least one j in $\{1,\ 2,\ 3\}$

Our p-value is less than 0.05. **Reject H_0.**

Testing for autocorrelation

Example: Static model and AR(3)

Step 5: Conclude.

Recall: Our hypotheses consider the model

$$e_t = \rho_1 e_{t-1} + \rho_2 e_{t-2} + \rho_3 e_{t-3}$$

which we are actually using to learn about the model

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \rho_3 u_{t-3}$$

H_0: $\rho_1 = \rho_2 = \rho_3 = 0$ *vs.* H_A: $\rho_j \neq 0$ for at least one j in $\{1,\, 2,\, 3\}$

Our p-value is less than 0.05. **Reject H_0.**

Conclude there is statistically significant evidence of autocorrelation.

Testing for autocorrelation

Dynamic models with lagged outcome variables

Recall: OLS is biased and inconsistent when our model has *both*

1. a lagged dependent variable
2. autocorrelated disturbances

Dynamic models with lagged outcome variables

Recall: OLS is biased and inconsistent when our model has *both*

1. a lagged dependent variable
2. autocorrelated disturbances

Problem: If OLS is biased for β, then it is also biased for u_t.

Dynamic models with lagged outcome variables

Recall: OLS is biased and inconsistent when our model has *both*

1. a lagged dependent variable
2. autocorrelated disturbances

Problem: If OLS is biased for β, then it is also biased for u_t.

$\therefore$ We can't apply our nice trick of *just* using e_t to learn about u_t.

Testing for autocorrelation

Dynamic models with lagged outcome variables

Recall: OLS is biased and inconsistent when our model has *both*

1. a lagged dependent variable

2. autocorrelated disturbances

Problem: If OLS is biased for β, then it is also biased for u_t.

$\therefore$ We can't apply our nice trick of *just* using e_t to learn about u_t.

Solution: **Breusch-Godfrey** test includes the other explanatory variables,

$$e_t = \underbrace{\gamma_0 + \gamma_1 x_{1t} + \gamma_2 x_{2t} + \cdots}_{\text{Explanatory variables (RHS)}} + \underbrace{\rho_1 e_{t-1} + \rho_2 e_{t-2} + \cdots}_{\text{Lagged residuals}} + \varepsilon_t$$

Testing for autocorrelation

Dynamic models with lagged outcome variables

Specifically, to test for AR(2) disturbances in the ADL(1, 0) model

$$\text{Births}_t = \beta_0 + \beta_1 \text{Income}_t + \beta_2 \text{Births}_{t-1} + u_t \tag{B}$$

1. Estimate (B) via OLS.

2. Calculate residuals (e_t) from the OLS regression in step 1.

3. Regress residuals on an intercept, explanatory variables, and lagged residuals.

$$e_t = \gamma_0 + \gamma_1 \text{ Income}_t + \gamma_3 \text{ Births}_{t-1} + \rho_1 e_{t-1} + \rho_2 e_{t-2} + v_t$$

4. Conduct LM or F test for $\rho_1 = \rho_2 = 0$.

5. Rejecting H_0 implies significant evidence of AR(2).

Testing for autocorrelation

Dynamic models with lagged outcome variables

For an example, let's consider the relationship between monthly presidential approval ratings and oil prices during President George W. Bush's[†] presidency.

We will specify the process as ADL(1, 0) and test for an AR(2) process in our disturbances.

$$\text{Approval}_t = \beta_0 + \beta_1 \text{Approval}_{t-1} + \beta_2 \text{Price}_t + u_t$$

† Fun with approval ratings.

Testing for autocorrelation

Dynamic models with lagged outcome variables

For an example, let's consider the relationship between monthly presidential approval ratings and oil prices during President George W. Bush's[†] presidency.

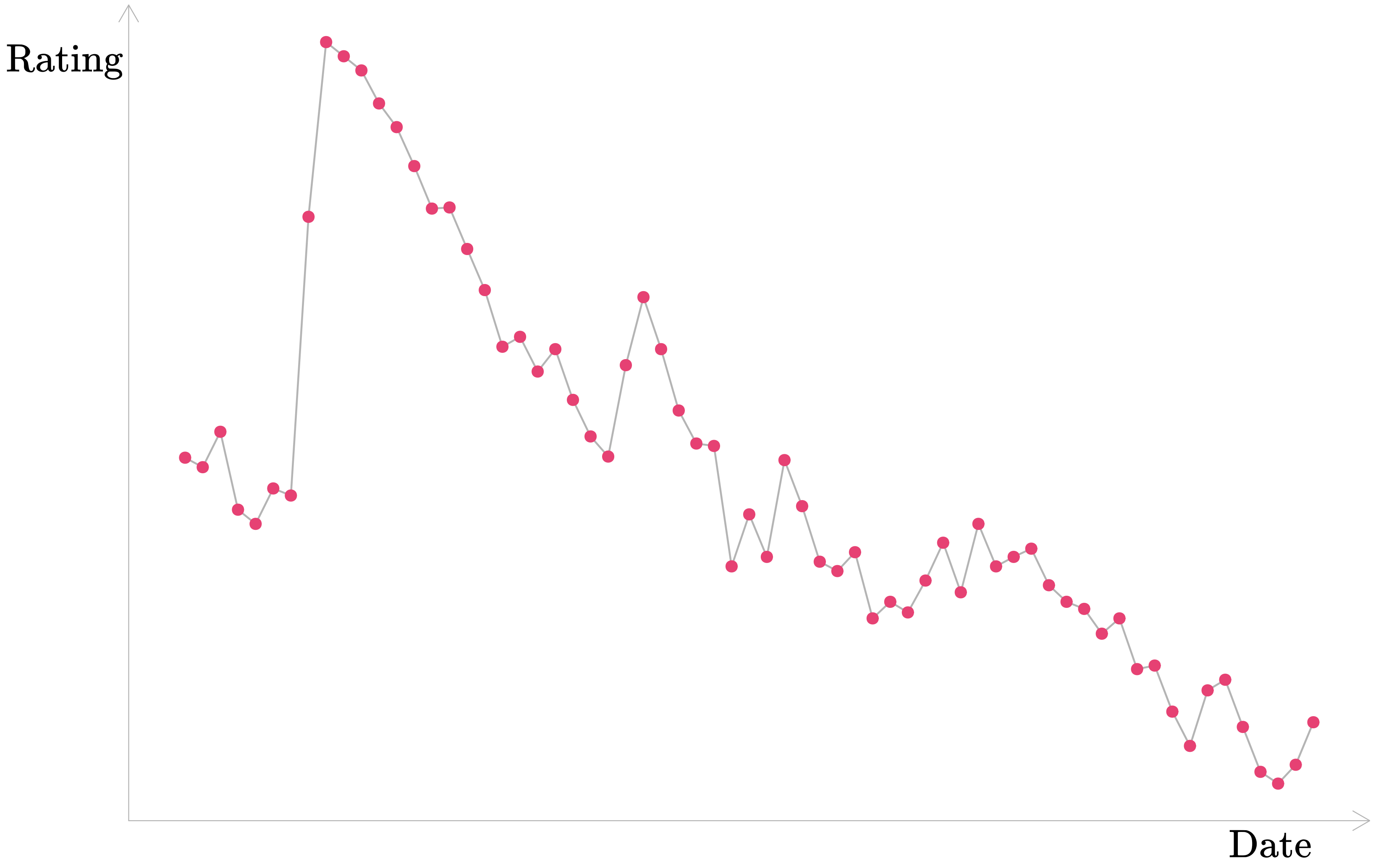
We will specify the process as ADL(1, 0) and test for an AR(2) process in our disturbances.

$$\text{Approval}_t = \beta_0 + \beta_1 \text{Approval}_{t-1} + \beta_2 \text{Price}_t + u_t$$

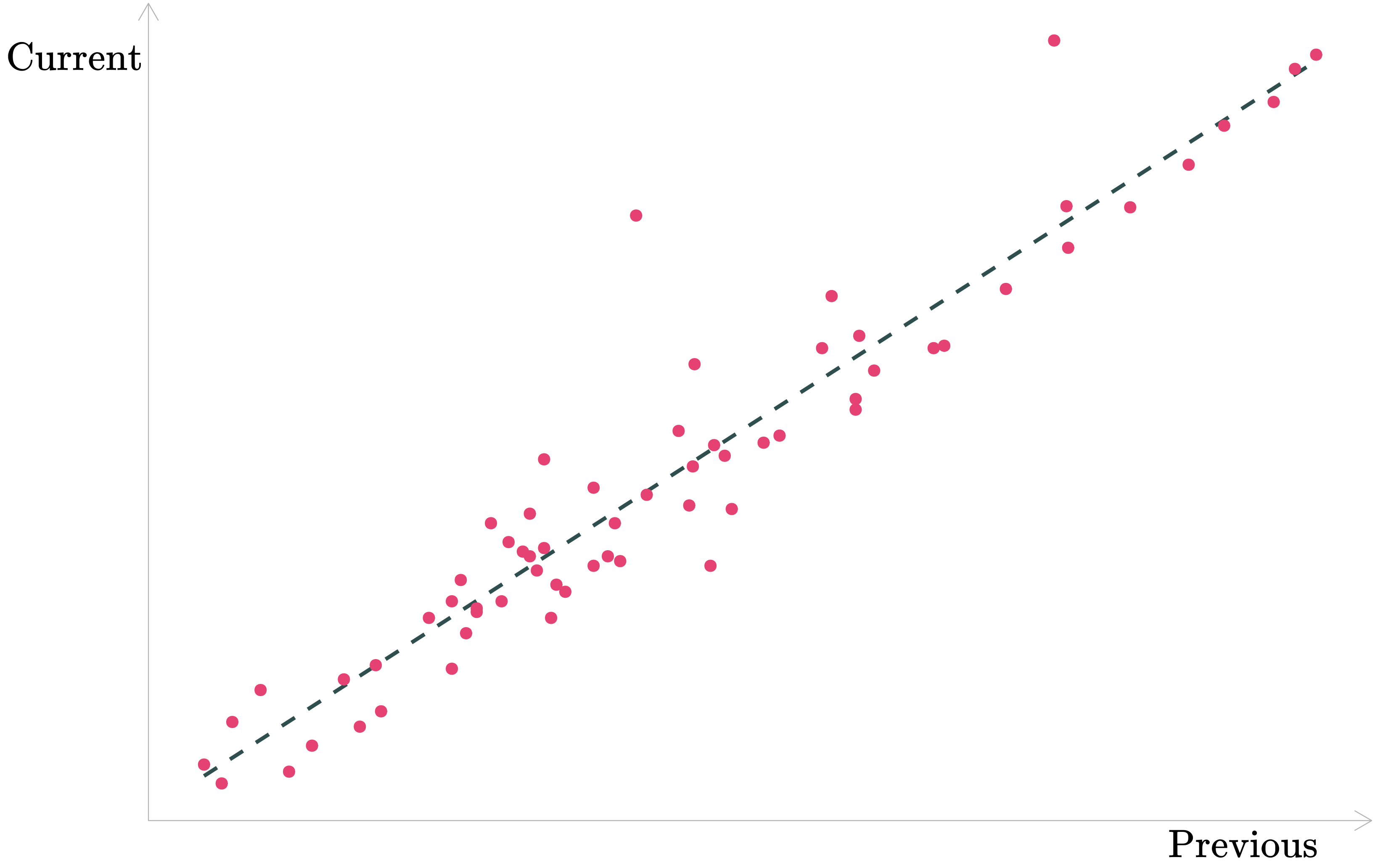
Note: We're ignoring any other violations of exogeneity for the moment.

† Fun with approval ratings.

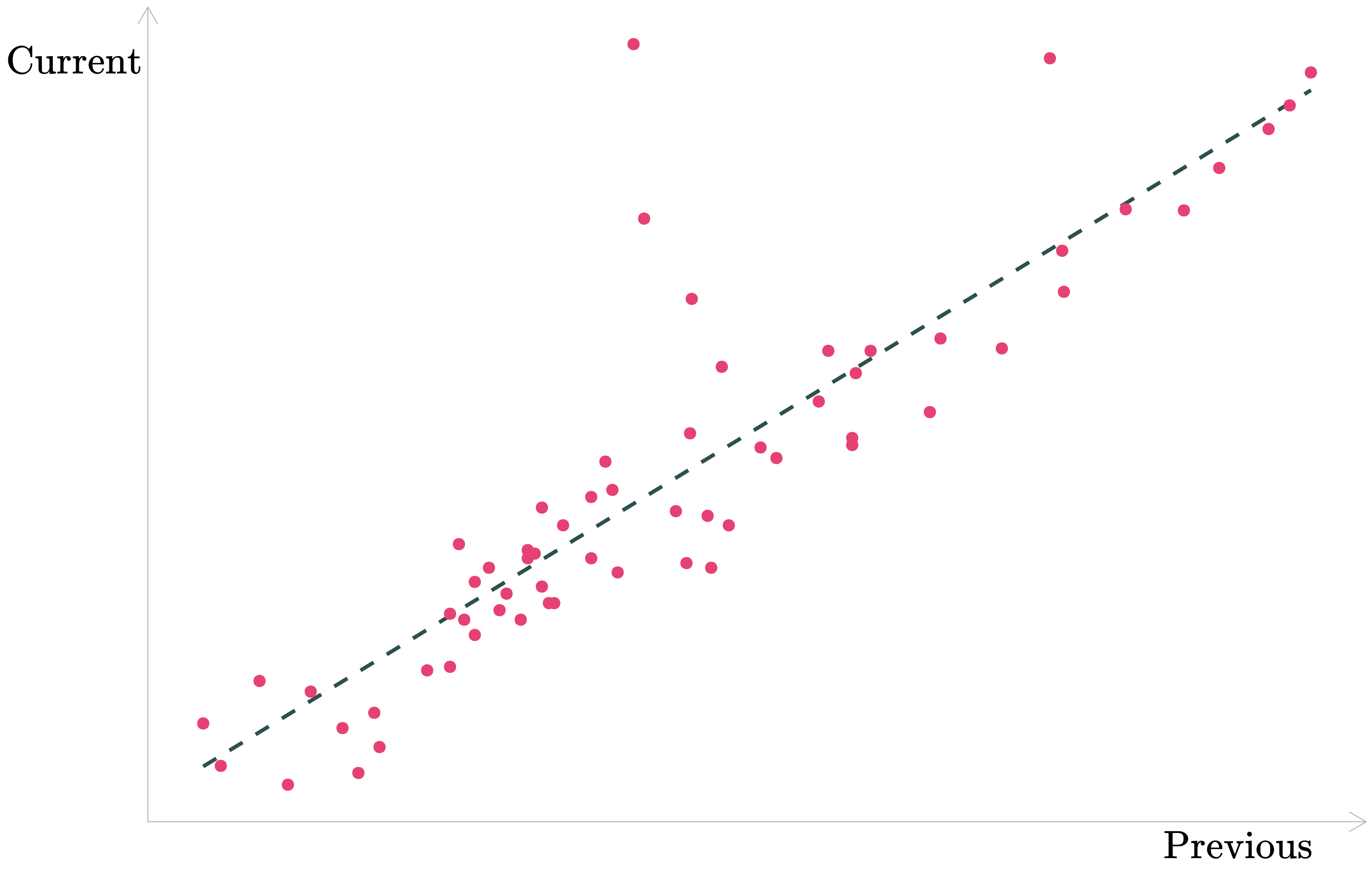
Monthly presidential approval ratings, 2001–2006



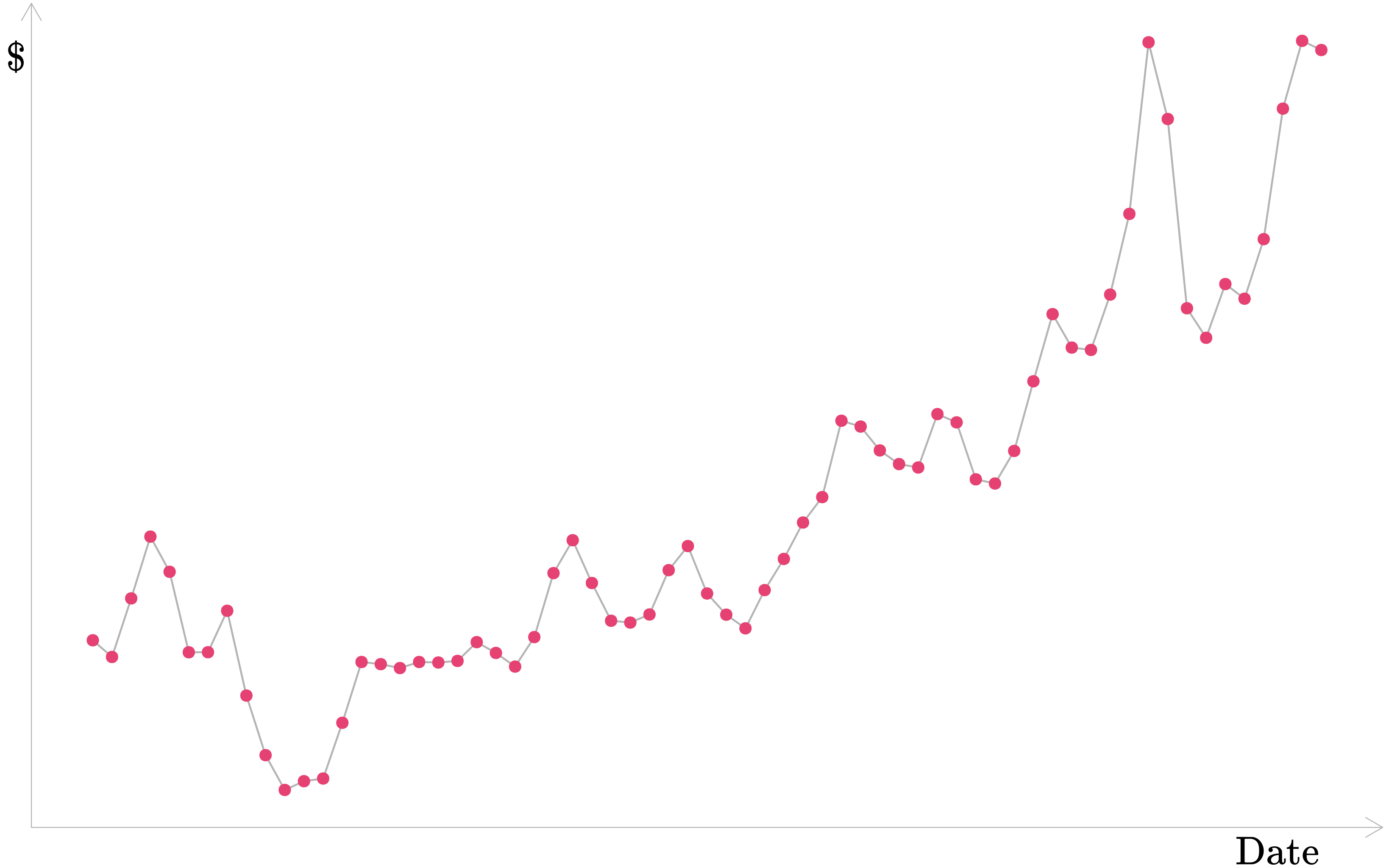
Approval rating vs. its one-month lag, 2001–2006



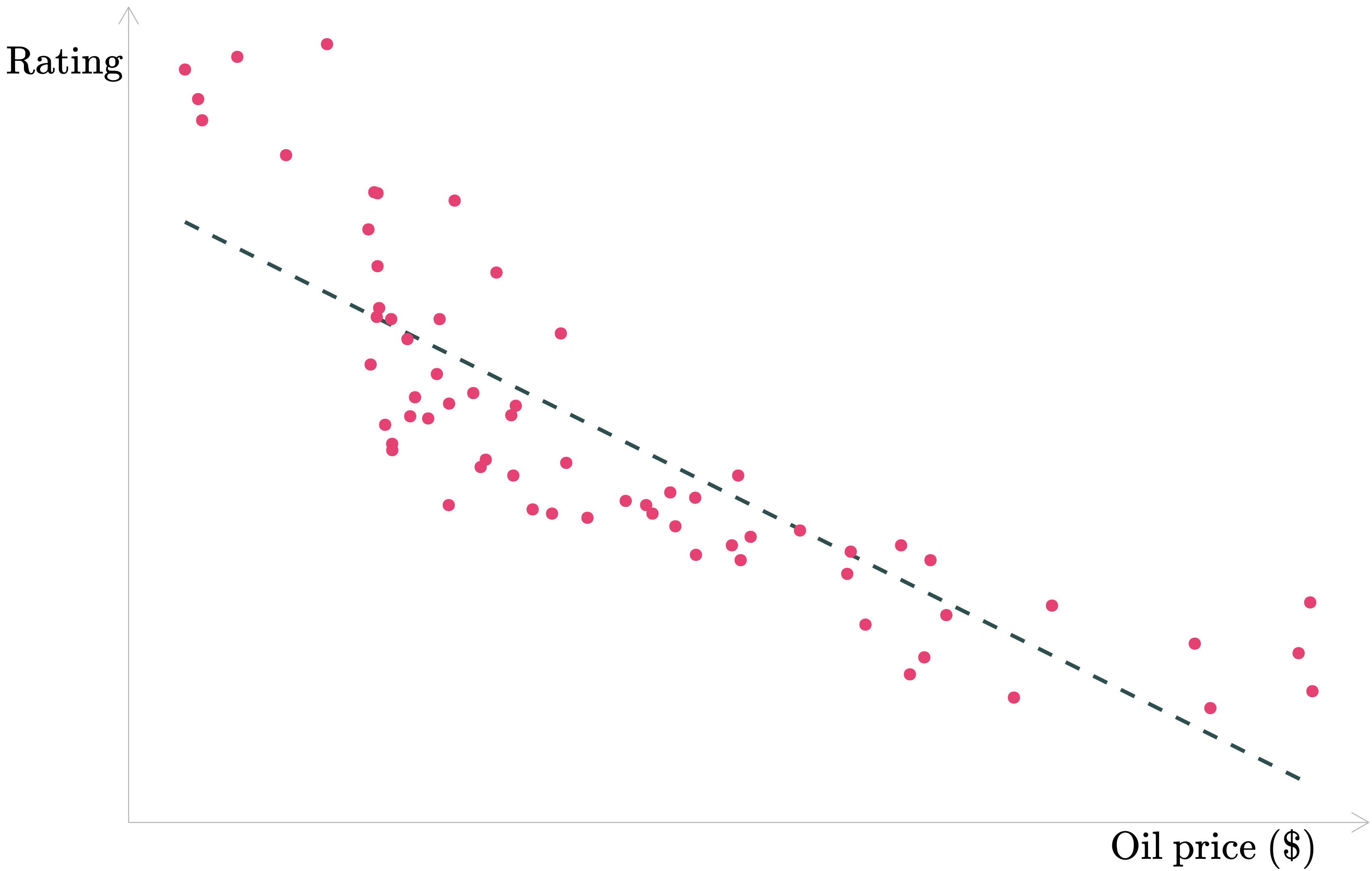
Approval rating vs. its two-month lag, 2001–2006



Oil prices, 2001–2006



Approval rating vs. oil prices, 2001–2006



Testing for autocorrelation

Example: Approval ratings and oil prices

Step 1: Estimate our ADL(1, 0) model with OLS.

```
# Estimate the model
ols_est <- lm(
  approve ~ lag(approve) + price_oil,
  data = approval_df
)
# Summary
tidy(ols_est)
```

```
#> # A tibble: 3 × 5
#>   term         estimate std.error statistic  p.value
#>   <chr>           <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)   16.2       7.86        2.06 4.40e- 2
#> 2 lag(approve)   0.841     0.0752     11.2  2.17e-16
#> 3 price_oil     -0.0410    0.0215     -1.90 6.15e- 2
```

Testing for autocorrelation

Example: Approval ratings and oil prices

Step 2: Record residuals from the OLS regression.

```
# Grab residuals
approval_df$e <- c(NA, residuals(ols_est))
```

Testing for autocorrelation

Example: Approval ratings and oil prices

Step 2: Record residuals from the OLS regression.

```
# Grab residuals
approval_df$e <- c(NA, residuals(ols_est))
```

Note: We add an `NA` because we use a lag—the first element is missing.

Testing for autocorrelation

Example: Approval ratings and oil prices

Step 2: Record residuals from the OLS regression.

```
# Grab residuals
approval_df$e ← c(NA, residuals(ols_est))
```

Note: We add an `NA` because we use a lag—the first element is missing.

E.g.,

```
{1, 2, 3, 4, 5, 6, 7, 8, 9} = x
```

Example: Approval ratings and oil prices

Step 2: Record residuals from the OLS regression.

```
# Grab residuals
approval_df$e ← c(NA, residuals(ols_est))
```

Note: We add an `NA` because we use a lag—the first element is missing.

E.g.,

```
{1, 2, 3, 4, 5, 6, 7, 8, 9} = x
{?, 1, 2, 3, 4, 5, 6, 7, 8} = lag(x)
```

Testing for autocorrelation

Example: Approval ratings and oil prices

Step 2: Record residuals from the OLS regression.

```
# Grab residuals
approval_df$e ← c(NA, residuals(ols_est))
```

Note: We add an `NA` because we use a lag—the first element is missing.

E.g.,

```
{1, 2, 3, 4, 5, 6, 7, 8, 9} = x
{?, 1, 2, 3, 4, 5, 6, 7, 8} = lag(x)
{?, ?, 1, 2, 3, 4, 5, 6, 7} = lag(x, 2)
```

Testing for autocorrelation

Example: Approval ratings and oil prices

Step 2: Record residuals from the OLS regression.

```
# Grab residuals
approval_df$e <- c(NA, residuals(ols_est))
```

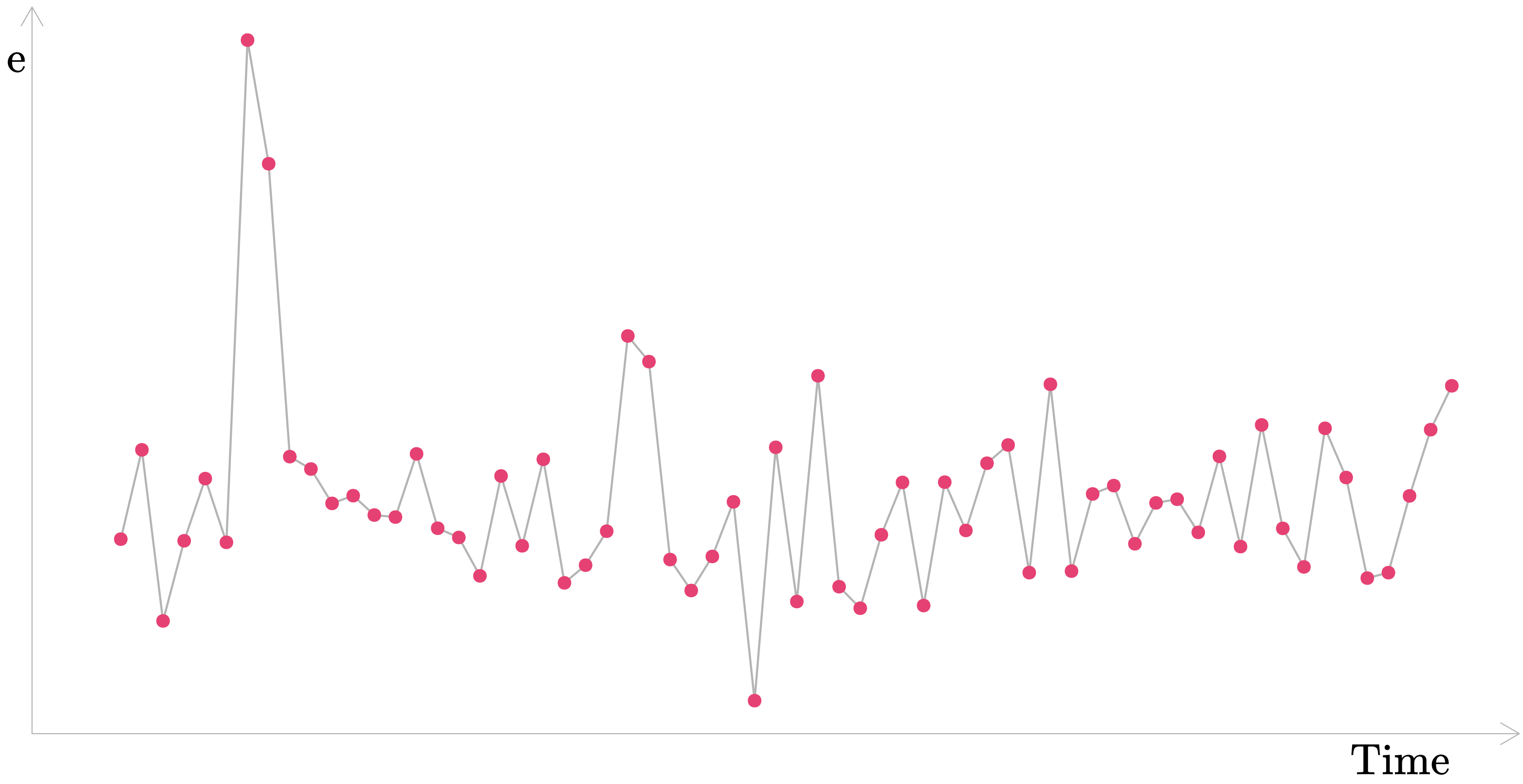
Note: We add an `NA` because we use a lag—the first element is missing.

E.g.,

```
{1, 2, 3, 4, 5, 6, 7, 8, 9} = x
{?, 1, 2, 3, 4, 5, 6, 7, 8} = lag(x)
{?, ?, 1, 2, 3, 4, 5, 6, 7} = lag(x, 2)
{?, ?, ?, 1, 2, 3, 4, 5, 6} = lag(x, 3)
```

Testing for autocorrelation

Example: Approval ratings and oil prices



Testing for autocorrelation

Example: Approval ratings and oil prices

Step 3: Regress residuals on an intercept, the explanatory variables, and lagged residuals.

```r
# BG regression
bg_reg <- lm(
  e ~ lag(approve) + price_oil + lag(e) + lag(e, 2),
  data = approval_df
)
```

```
#>               Estimate Std. Error t value Pr(>|t|)
#> (Intercept)   7.92474    9.30455   0.852   0.3979
#> lag(approve) -0.08503    0.09192  -0.925   0.3589
#> price_oil    -0.01690    0.02407  -0.702   0.4854
#> lag(e)        0.25236    0.14648   1.723   0.0903 .
#> lag(e, 2)     0.07865    0.14471   0.544   0.5889
```

Example: Approval ratings and oil prices

Step 4: F (or LM) test for $\rho_1 = \rho_2 = 0$.

Recall: We can test joint significance using an F test that compares the restricted (here: $\rho_1 = \rho_2 = 0$) and unrestricted models.

$$F_{q,\,n-p} = \dfrac{\left(\text{SSE}_r - \text{SSE}_u\right) \big/ q}{\text{SSE}_u \big/ \left(n - p\right)}$$

where q is the number of restrictions and p is the number of parameters in our unrestricted model (include the intercept).

We can use the `waldtest()` function from the `lmtest` package for this test.

Testing for autocorrelation

Example: Approval ratings and oil prices

Step 4: F (or LM) test for $\rho_1 = \rho_2 = 0$.

```r
# BG regression
bg_reg <- lm(
  e ~ lag(approve) + price_oil + lag(e) + lag(e, 2),
  data = approval_df
)
# Test significance of the lags using 'waldtest' from 'lmtest' package
p_load(lmtest)
waldtest(bg_reg, c("lag(e)", "lag(e, 2)"))
```

Here, we're telling `waldtest` to test

- the model we specified in `bg_reg` (our **unrestricted model**)
- against a model without `lag(e)` and `lag(e, 2)` (our **restricted model**)

Example: Approval ratings and oil prices

Step 4: F (or LM) test for $\rho_1 = \rho_2 = 0$.

```
# BG regression
bg_reg <- lm(
  e ~ lag(approve) + price_oil + lag(e) + lag(e, 2),
  data = approval_df
)
# Test significance of the lags using 'waldtest' from 'lmtest' package
p_load(lmtest)
waldtest(bg_reg, c("lag(e)", "lag(e, 2)"))
```

```
#> Wald test
#>
#> Model 1: e ~ lag(approve) + price_oil + lag(e) + lag(e, 2)
#> Model 2: e ~ lag(approve) + price_oil
#>   Res.Df Df      F Pr(>F)
#> 1     57
#> 2     59 -2 1.6153 0.2078
```

Testing for autocorrelation

Example: Approval ratings and oil prices

Step 5: Conclusion of hypothesis test

With a *p*-value of ~0.208, **we fail to reject the null hypothesis.**

- We cannot reject $\rho_1 = \rho_2 = 0$.

- We cannot reject "no autocorrelation".

Testing for autocorrelation

Example: Approval ratings and oil prices

Step 5: Conclusion of hypothesis test

With a p-value of ~0.208, **we fail to reject the null hypothesis.**

- We cannot reject $\rho_1 = \rho_2 = 0$.
- We cannot reject "no autocorrelation".

However, **we tested for a specific type of autocorrelation**: AR(2).

We might get different answers with different tests.

The p-value for AR(1) is 0.0896—suggestive of first-order autocorrelation.

Living with autocorrelation

Autocorrelation

Working with it

Suppose we believe autocorrelation is present. What do we do?

Working with it

Suppose we believe autocorrelation is present. What do we do?

I'll give you three options.†

1. **Misspecification**

2. **Serial-correlation robust standard errors** (a.k.a. *Newey-West*)

3. **FGLS**

† You should take `EC 422` to go much deeper into time-series analysis/forecasting.

Autocorrelation

Option 1: Misspecification

Misspecification with autocorrelation is very similar to our discussion with heteroskedasticity.

By incorrectly specifying your model, you can create autocorrelation.

Omitting variables that are correlated through time will cause your disturbances to be correlated through time.

Autocorrelation

Option 1: Misspecification

Example: Suppose births depend upon income and previous births

$$\text{Births}_t = \beta_0 + \beta_1 \text{Births}_{t-1} + \beta_2 \text{Income}_t + u_t$$

Autocorrelation

Option 1: Misspecification

Example: Suppose births depend upon income and previous births

$$\text{Births}_t = \beta_0 + \beta_1 \text{Births}_{t-1} + \beta_2 \text{Income}_t + u_t$$

but we write down the model as only depending upon previous births, *i.e.*,

$$\text{Births}_t = \beta_0 + \beta_1 \text{Births}_{t-1} + v_t$$

Option 1: Misspecification

Example: Suppose births depend upon income and previous births

$$\text{Births}_t = \beta_0 + \beta_1 \text{Births}_{t-1} + \beta_2 \text{Income}_t + u_t$$

but we write down the model as only depending upon previous births, *i.e.*,

$$\text{Births}_t = \beta_0 + \beta_1 \text{Births}_{t-1} + v_t$$

Then our disturbance v_t is

$$v_t = \beta_2 \text{Income}_t + u_t$$

Autocorrelation

Option 1: Misspecification

Example: Suppose births depend upon income and previous births

$$\text{Births}_t = \beta_0 + \beta_1 \text{Births}_{t-1} + \beta_2 \text{Income}_t + u_t$$

but we write down the model as only depending upon previous births, *i.e.*,

$$\text{Births}_t = \beta_0 + \beta_1 \text{Births}_{t-1} + v_t$$

Then our disturbance v_t is

$$v_t = \beta_2 \text{Income}_t + u_t$$

which is likely autocorrelated, since income is correlated in time.

Autocorrelation

Option 1: Misspecification

Example: Suppose births depend upon income and previous births

$$\text{Births}_t = \beta_0 + \beta_1 \text{Births}_{t-1} + \beta_2 \text{Income}_t + u_t$$

but we write down the model as only depending upon previous births, *i.e.*,

$$\text{Births}_t = \beta_0 + \beta_1 \text{Births}_{t-1} + v_t$$

Then our disturbance v_t is

$$v_t = \beta_2 \text{Income}_t + u_t$$

which is likely autocorrelated, since income is correlated in time.

Note: This autocorrelation has nothing to do with u_t.

Option 1: Misspecification

"Proof"

$$
\begin{aligned}
v_t &= \beta_2 \text{Income}_t + u_t \\
v_{t-1} &= \beta_2 \text{Income}_{t-1} + u_{t-1}
\end{aligned}
$$

$\text{Cov}(v_t,\, v_{t-1})$

Option 1: Misspecification

"Proof"

$$
\begin{aligned}
v_t &= \beta_2 \text{Income}_t + u_t \\
v_{t-1} &= \beta_2 \text{Income}_{t-1} + u_{t-1}
\end{aligned}
$$

$\text{Cov}(v_t,\, v_{t-1})$

$= \text{Cov}(\beta_2 \text{Income}_t + u_t,\, \beta_2 \text{Income}_{t-1} + u_{t-1})$

Autocorrelation

Option 1: Misspecification

"Proof"

$$v_t = \beta_2 \text{Income}_t + u_t$$
$$v_{t-1} = \beta_2 \text{Income}_{t-1} + u_{t-1}$$

$\text{Cov}(v_t,\, v_{t-1})$

$= \text{Cov}(\beta_2 \text{Income}_t + u_t,\, \beta_2 \text{Income}_{t-1} + u_{t-1})$

$= \text{Cov}(\beta_2 \text{Income}_t,\, \beta_2 \text{Income}_{t-1}) + \text{Cov}(\beta_2 \text{Income}_t,\, u_t)$
$\quad + \text{Cov}(u_t,\, \beta_2 \text{Income}_{t-1}) + \text{Cov}(u_t,\, u_{t-1})$

Option 1: Misspecification

"Proof"

$$v_t = \beta_2 \text{Income}_t + u_t$$
$$v_{t-1} = \beta_2 \text{Income}_{t-1} + u_{t-1}$$

$\text{Cov}(v_t,\, v_{t-1})$

$= \text{Cov}(\beta_2 \text{Income}_t + u_t,\, \beta_2 \text{Income}_{t-1} + u_{t-1})$

$= \text{Cov}(\beta_2 \text{Income}_t,\, \beta_2 \text{Income}_{t-1}) + \text{Cov}(\beta_2 \text{Income}_t,\, u_t)$
$\quad + \text{Cov}(u_t,\, \beta_2 \text{Income}_{t-1}) + \text{Cov}(u_t,\, u_{t-1})$

$\neq 0$ (in general) even if u_t is exogenous and without autocorrelation.

Option 2: Newey-West standard errors

As was also the case with heteroskedasticity, you can still estimate consistent standard errors (and inference) in the presence of autocorrelation.

These standard errors are called *serial-correlation robust standard errors* (or *Newey-West standard errors*).

We are not going to derive the estimator for these standard errors.

Option 3: FGLS

If we do not have a lagged outcome variable, then feasible generalized least squares (**FGLS**) can give us efficient and consistent standard errors.

Option 3: FGLS

If we do not have a lagged outcome variable, then feasible generalized least squares (**FGLS**) can give us efficient and consistent standard errors.

Let's start with a simple static model that includes an AR(1) disturbance u_t.

$$\text{Births}_t = \beta_0 + \beta_1 \text{Income}_t + u_t \tag{1}$$
$$u_t = \rho u_{t-1} + \varepsilon_t \tag{2}$$

Option 3: FGLS

If we do not have a lagged outcome variable, then feasible generalized least squares (**FGLS**) can give us efficient and consistent standard errors.

Let's start with a simple static model that includes an AR(1) disturbance u_t.

$$\begin{align} \text{Births}_t &= \beta_0 + \beta_1 \text{Income}_t + u_t \tag{1} \\ u_t &= \rho u_{t-1} + \varepsilon_t \tag{2} \end{align}$$

Now our old trick: Write out (1) for period $t-1$ (and then multiple by ρ)

$$\begin{align} \text{Births}_{t-1} &= \beta_0 + \beta_1 \text{Income}_{t-1} + u_{t-1} \tag{3} \\ \rho \text{Births}_{t-1} &= \rho\beta_0 + \rho\beta_1 \text{Income}_{t-1} + \rho u_{t-1} \tag{4} \end{align}$$

Option 3: FGLS

If we do not have a lagged outcome variable, then feasible generalized least squares (**FGLS**) can give us efficient and consistent standard errors.

Let's start with a simple static model that includes an AR(1) disturbance u_t.

$$\text{Births}_t = \beta_0 + \beta_1 \text{Income}_t + u_t \quad (1)$$
$$u_t = \rho u_{t-1} + \varepsilon_t \quad (2)$$

Now our old trick: Write out (1) for period $t-1$ (and then multiple by ρ)

$$\text{Births}_{t-1} = \beta_0 + \beta_1 \text{Income}_{t-1} + u_{t-1} \quad (3)$$
$$\rho \text{Births}_{t-1} = \rho\beta_0 + \rho\beta_1 \text{Income}_{t-1} + \rho u_{t-1} \quad (4)$$

And now subtract (4) from (1)...

Autocorrelation

Option 3: FGLS

$$\begin{aligned}\text{Births}_t - \rho\text{Births}_{t-1} = &\beta_0\left(1-\rho\right) + \\ &\beta_1\text{Income}_t - \rho\beta_1\text{Income}_{t-1} + \\ &u_t - \rho u_{t-1}\end{aligned}$$

Autocorrelation

Option 3: FGLS

$$\begin{aligned}\text{Births}_t - \rho\text{Births}_{t-1} = & \beta_0 (1-\rho) + \\ & \beta_1 \text{Income}_t - \rho\beta_1\text{Income}_{t-1} + \\ & u_t - \rho u_{t-1}\end{aligned}$$

which gives us a very specific dynamic model

$$\begin{aligned}\text{Births}_t = & \beta_0 (1-\rho) + \rho\text{Births}_{t-1} + \\ & \beta_1 \text{Income}_t - \rho\beta_1\text{Income}_{t-1} + \\ & \underbrace{u_t - \rho u_{t-1}}_{=\varepsilon_t}\end{aligned}$$

Autocorrelation

Option 3: FGLS

$$\begin{aligned}\text{Births}_t - \rho\text{Births}_{t-1} =& \beta_0 \left(1 - \rho\right) + \\ & \beta_1 \text{Income}_t - \rho\beta_1\text{Income}_{t-1} + \\ & u_t - \rho u_{t-1}\end{aligned}$$

which gives us a very specific dynamic model

$$\begin{aligned}\text{Births}_t =& \beta_0 \left(1 - \rho\right) + \rho\text{Births}_{t-1} + \\ & \beta_1 \text{Income}_t - \rho\beta_1\text{Income}_{t-1} + \\ & \underbrace{u_t - \rho u_{t-1}}_{=\varepsilon_t} \\ =& \beta_0 \left(1 - \rho\right) + \rho\text{Births}_{t-1} + \\ & \beta_1 \text{Income}_t - \rho\beta_1\text{Income}_{t-1} + \varepsilon_t\end{aligned}$$

that happens to be **free of autocorrelation**.

Option 3: FGLS

This **transformed model** is free of autocorrelation.

$$\begin{aligned}\text{Births}_t = &\beta_0\left(1 - \rho\right) + \rho\text{Births}_{t-1} + \\ &\beta_1\text{Income}_t - \rho\beta_1\text{Income}_{t-1} + \varepsilon_t\end{aligned}$$

Q: How do we actually estimate this model? (We don't know ρ.)

Autocorrelation

Option 3: FGLS

This **transformed model** is free of autocorrelation.

$$
\begin{aligned}
\text{Births}_t =& \beta_0 \left(1 - \rho\right) + \rho \text{Births}_{t-1} + \\
& \beta_1 \text{Income}_t - \rho \beta_1 \text{Income}_{t-1} + \varepsilon_t
\end{aligned}
$$

Q: How do we actually estimate this model? (We don't know ρ.)

A: FGLS (of course)...

Option 3: FGLS

This **transformed model** is free of autocorrelation.

$$
\begin{aligned}
\text{Births}_t = &\beta_0 (1 - \rho) + \rho\text{Births}_{t-1} + \\
&\beta_1\text{Income}_t - \rho\beta_1\text{Income}_{t-1} + \varepsilon_t
\end{aligned}
$$

Q: How do we actually estimate this model? (We don't know ρ.)

A: FGLS (of course)...

1. Estimate the original (untransformed) model; save residuals.
2. Estimate ρ: Regress residuals on their lags (no intercept).
3. Estimate the **transformed model**, plugging in $\hat{\rho}$ for ρ.

Table of contents