Causality

EC 421, Set 10

Edward Rubin
28 February 2019

Prologue

Schedule

Last Time

- Autocorrelation, nonstationarity, 'in-class' analysis
- **Follow up:** `EC422` (time series) is only offered in the winter. 😢
- **Follow up:** `EC410` (computational economics) in the spring! 😃
- **Follow up:** `R` is mainly written in `C`, `R`, and `Fortran`.

Today

- Return to our in-class examples
- Causality

Upcoming

Assignment due Sunday. Another one coming soon.

Problems and strategies

Step 1: Define the problem.

Q: What was the problem/goal/objective for the analysis?

Problems and strategies

Step 1: Define the problem.

Q: What was the problem/goal/objective for the analysis?

A: For y_1 and y_2 each, find the **true** model.

Problems and strategies

Step 1: Define the problem.

Q: What was the problem/goal/objective for the analysis?

A: For y_1 and y_2 each, find the **true** model.

Clarification:

Q: What does the *true model* for y_1 mean?

(**A**) The variables that best explain/predict y_1.

(**B**) The variables that are statistically significant.

(**C**) The variables that actually generated y_1.

(**D**) Something else?

Problems and strategies

Step 1: Define the problem.

Q: What was the problem/goal/objective for the analysis?

A: For y_1 and y_2 each, find the **true** model.

Clarification:

Q: What does the *true model* for y_1 mean?

(**A**) The variables that best explain/predict y_1.

(**B**) The variables that are statistically significant.

(**C**) The variables that actually generated y_1.

(**D**) Something else?

A: (C) We want to know variables and coefficients generated y_1.

Problems and strategies

Step 1: Define the problem.

Q: What was the problem/goal/objective for the analysis?

A: For y_1 and y_2 each, find the **true** model.

Clarification:

Q: What does the *true model* for y_1 mean?

(**A**) The variables that best explain/predict y_1.

(**B**) The variables that are statistically significant.

(**C**) The variables that actually generated y_1.

(**D**) Something else?

A: (C) We want to know variables and coefficients generated y_1.

The **true data-generating process** (DGP).

Problems and strategies

Step 2: Define your strategy

How did you approach this problem?

Problems and strategies

Step 2: Define your strategy

How did you approach this problem?

A few options:

1. Find the combination of variables that **maximize R^2** or **adjusted R^2**.
2. First **include all** variables. Keep statistically **significant variables**.
3. Iterate with (2.): **Drop non-significant variables** until nothing changes.
4. **Add variables one by one**. Keep statistically **significant variables**.
5. **Plot** variables' (or residuals') relationships with y.

```r
# Load the data
fun_df <- read_csv("fun_data.csv")
# Separate into two datasets
y1_df <- fun_df %>% select(-y2)
y2_df <- fun_df %>% select(-y1)
# Peak at the data
y1_df
```

```
#> # A tibble: 100 x 10
#>          y1     x1    x2     x3    x4     x5    x6       x7    x8    x9
#>       <dbl>  <dbl> <dbl>  <dbl> <dbl>  <dbl> <dbl>    <dbl> <dbl> <dbl>
#>  1   3.08   -0.777 0.405  1.23  0.762 -0.232  1.17  0.111       1 1.98
#>  2   6.04    0.473 1.59   0.584 1.53   0.349  1.52 -0.00994     2 0.511
#>  3   9.57    2.30  3.52  -0.976 3.32   0.581  1.50  0.974       3 0.936
#>  4  11.4     2.46  5.33  -1.77  4.64  -0.576  1.92  2.53        4 2.88
#>  5  -0.0319  0.313 2.09  -2.59  1.37  -0.717  3.76  2.14        5 2.20
#>  6   5.21    1.37  1.23   2.34  2.21  -1.40   3.55  1.17        6 1.83
#>  7   7.97    1.73  3.46   0.584 2.24  -1.31   3.77  1.92        7 1.75
#>  8  -5.17    2.60  4.09  -4.15  4.13  -2.57   4.60  0.886       8 1.14
#>  9   1.57    0.877 3.96   2.08  1.42  -2.89   3.68  1.32        9 2.23
#> 10   3.97   -0.197 0.875 -0.760 0.697 -1.92   1.90  1.85       10 1.90
#> # … with 90 more rows
```

R showcase

`gather`ing data

Let's plot y_1 against the nine potential explanatory variables, x_1 to x_9.

R showcase

`gather`ing data

Let's plot y_1 against the nine potential explanatory variables, x_1 to x_9.

We'll use two new functions to streamline this process.

- `gather()` (from `dplyr`): *Stacks* variables (names and values).
- `facet_wrap()`: Creates multiple plots grouped by a variable.

`gather`ing data

Example: `gather` all variables in our dataset.

```
data.frame(w = 0:1, x = 2:3, y = 4:5, z = 6:7) %>%
  gather(key = "var", value = "value")
```

```
#>   var value
#> 1   w     0
#> 2   w     1
#> 3   x     2
#> 4   x     3
#> 5   y     4
#> 6   y     5
#> 7   z     6
#> 8   z     7
```

`gather`ing data

Example: `gather` all variables in our dataset except `w`.

```
data.frame(w = 0:1, x = 2:3, y = 4:5, z = 6:7) %>%
  gather(-w, key = "var", value = "value")
```

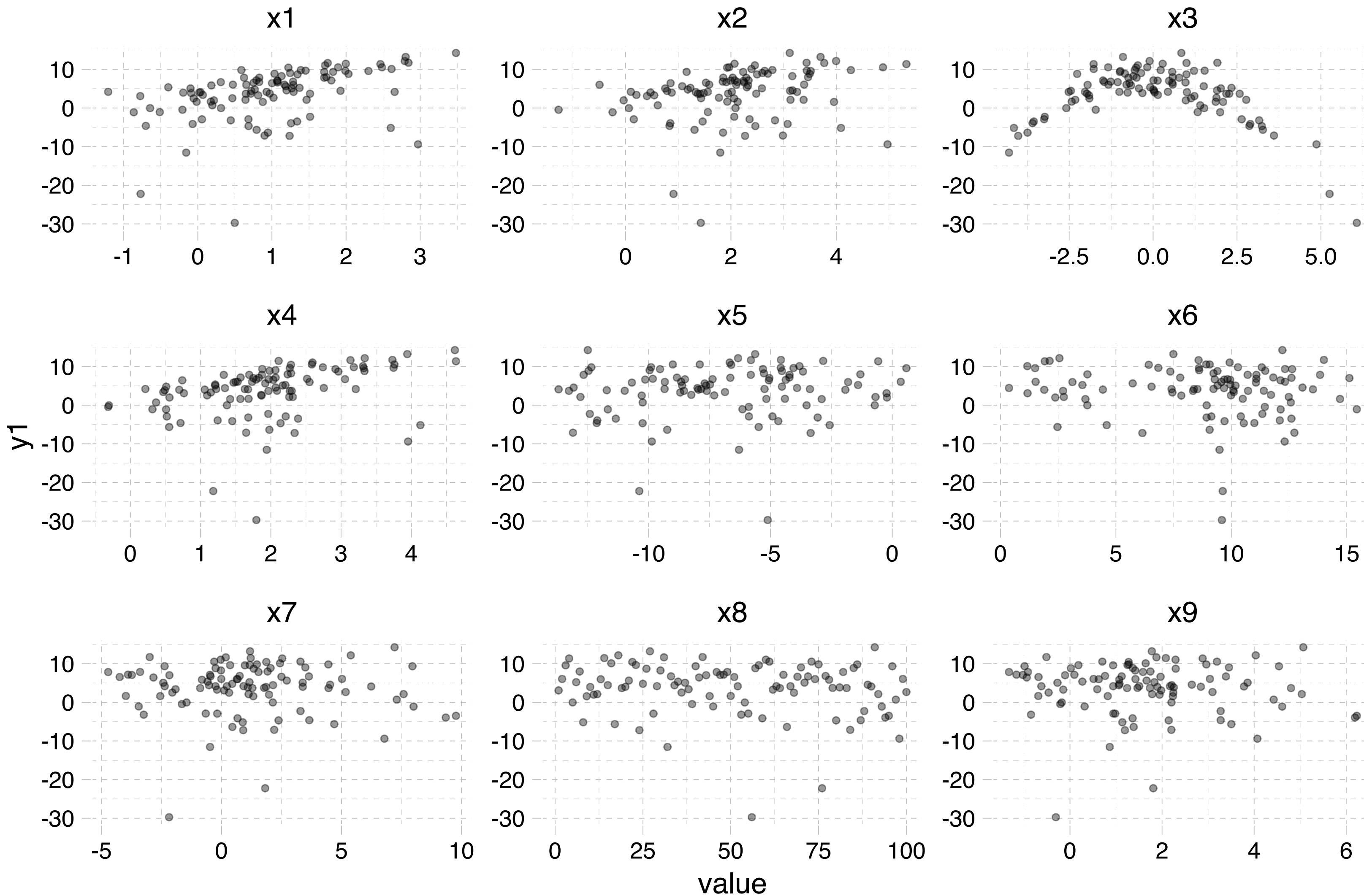
```
#>   w var value
#> 1 0   x     2
#> 2 1   x     3
#> 3 0   y     4
#> 4 1   y     5
#> 5 0   z     6
#> 6 1   z     7
```

`gather`ing data

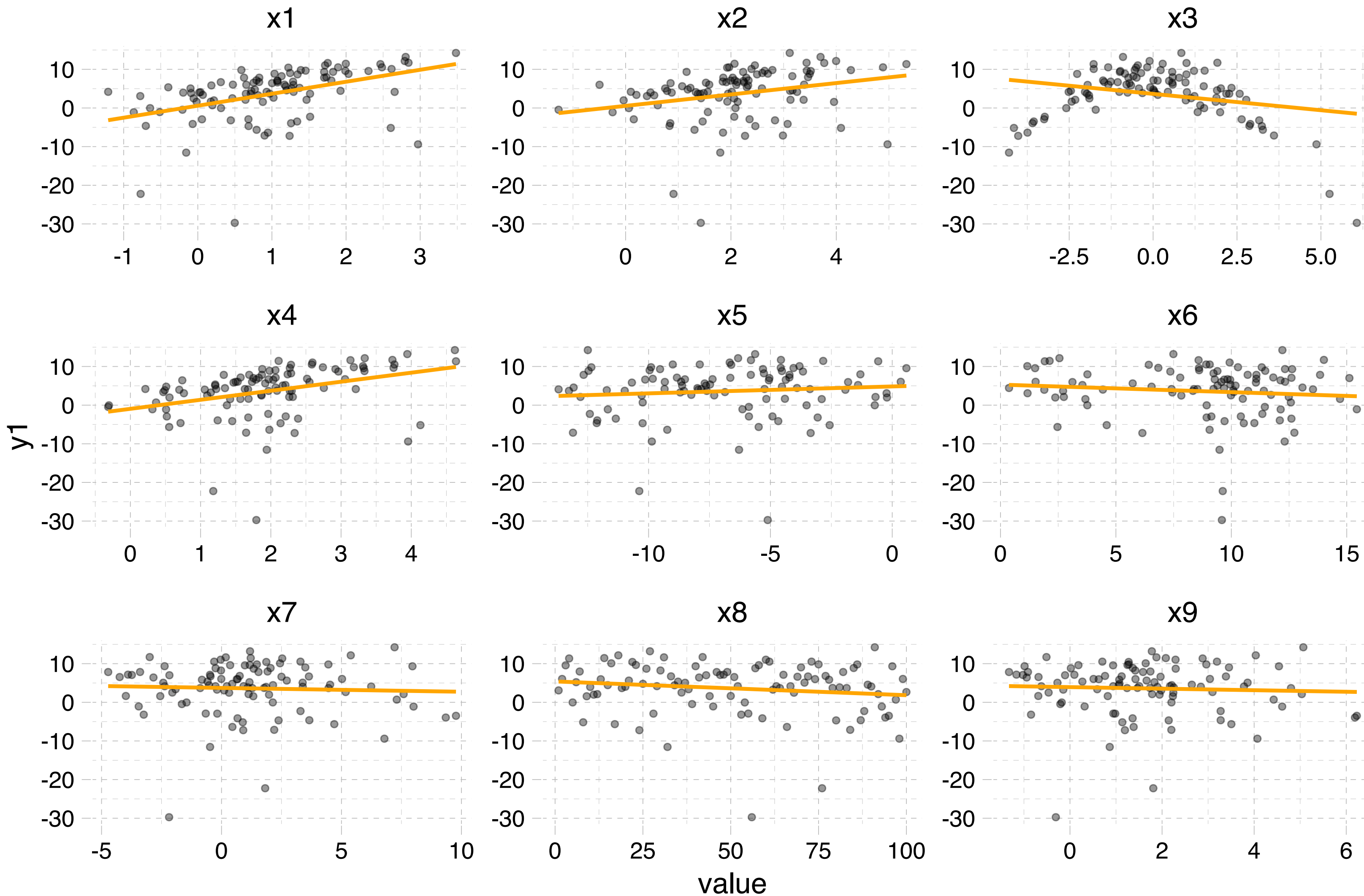
Adding these new functions to our previous `ggplot2` work...

```
y1_df %>% gather(-y1, key = "var", value = "value") %>%
  ggplot(aes(x = value, y = y1)) +
  geom_point(alpha = 0.4, size = 1.5) +
  facet_wrap(~ var, scales = "free") +
  theme_pander(base_size = 16)
```

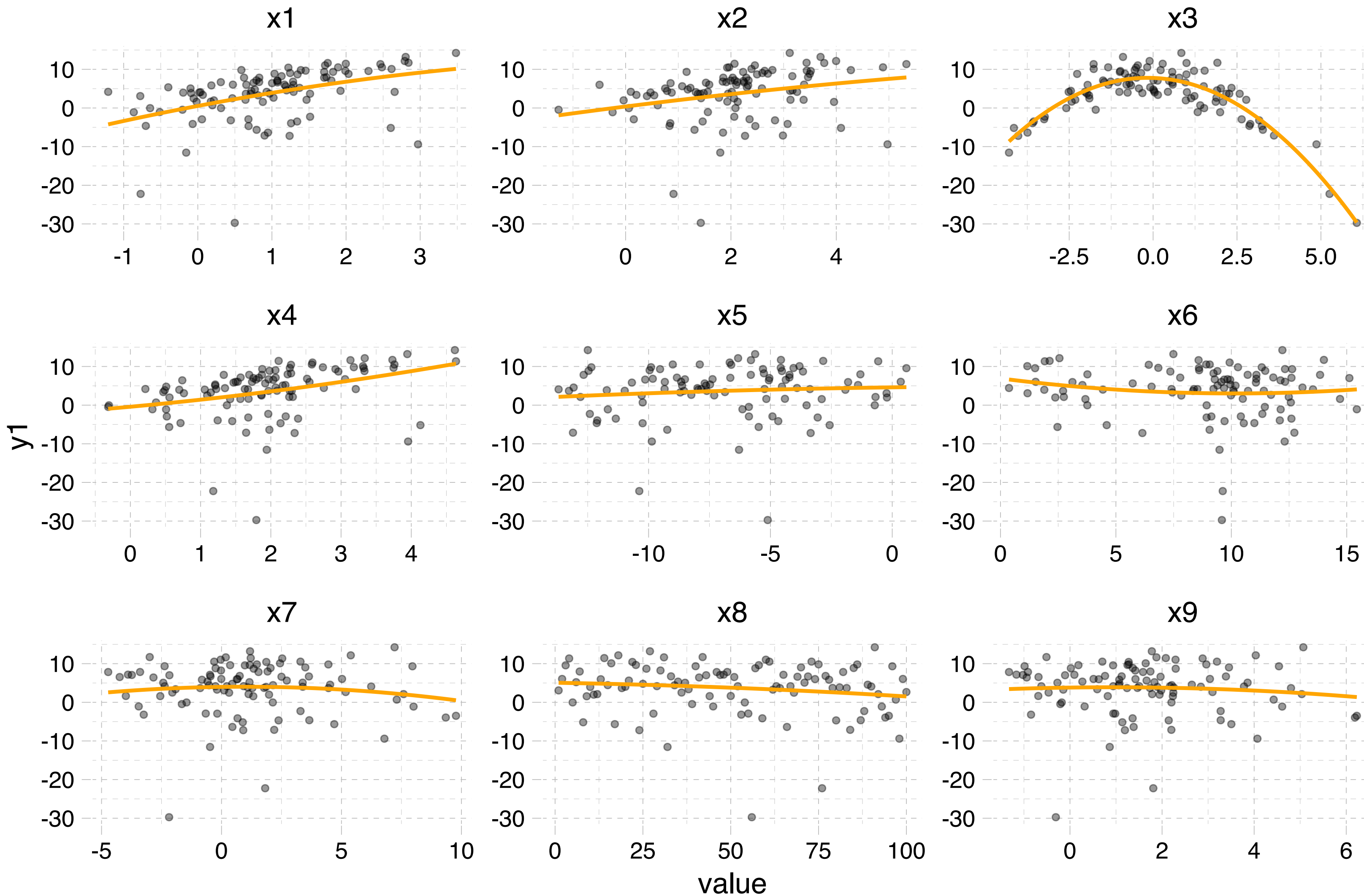
Plot: y_1 against x_1 through x_9



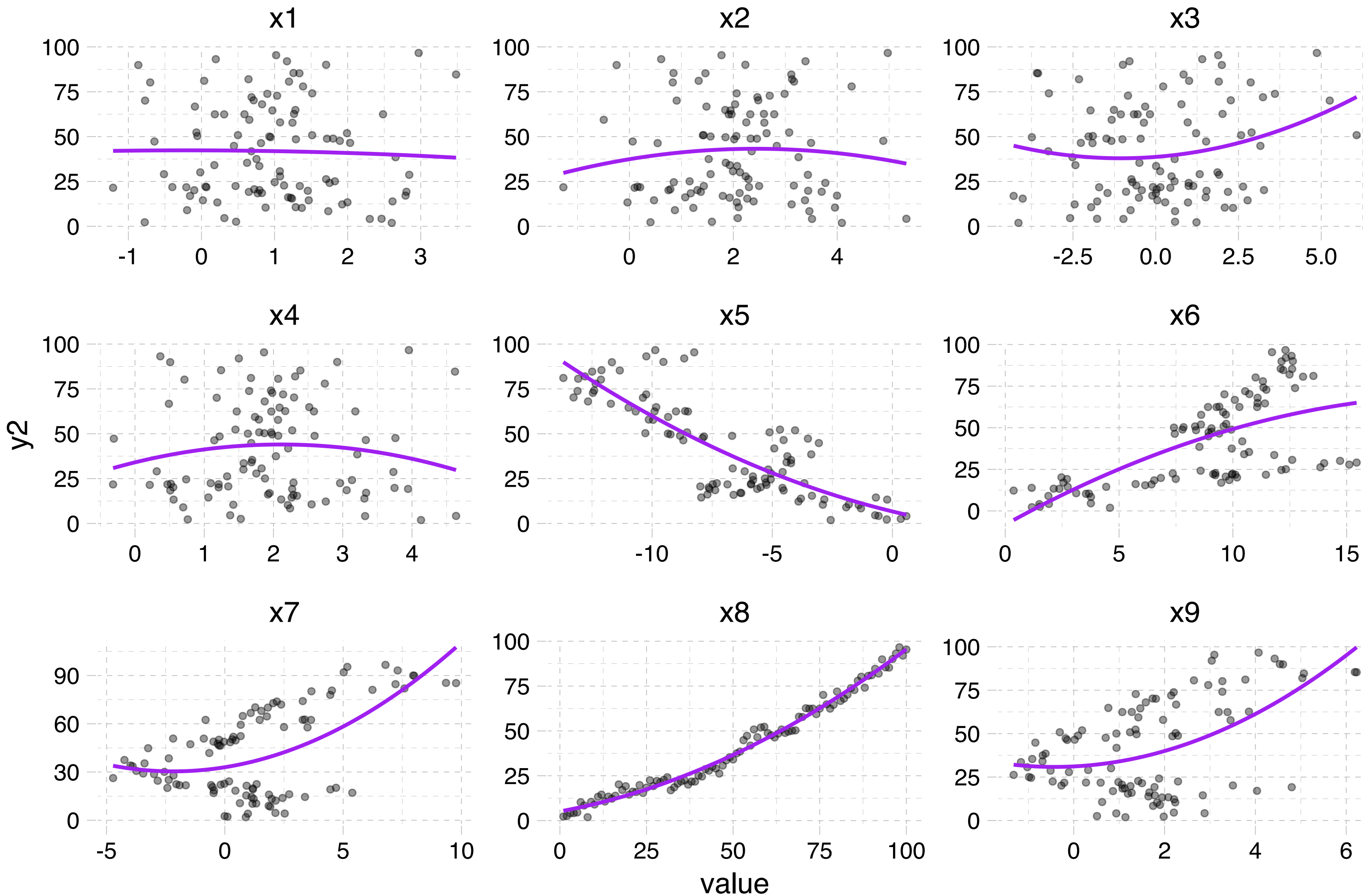
Simple linear regressions: y_1 against x_1 through x_9



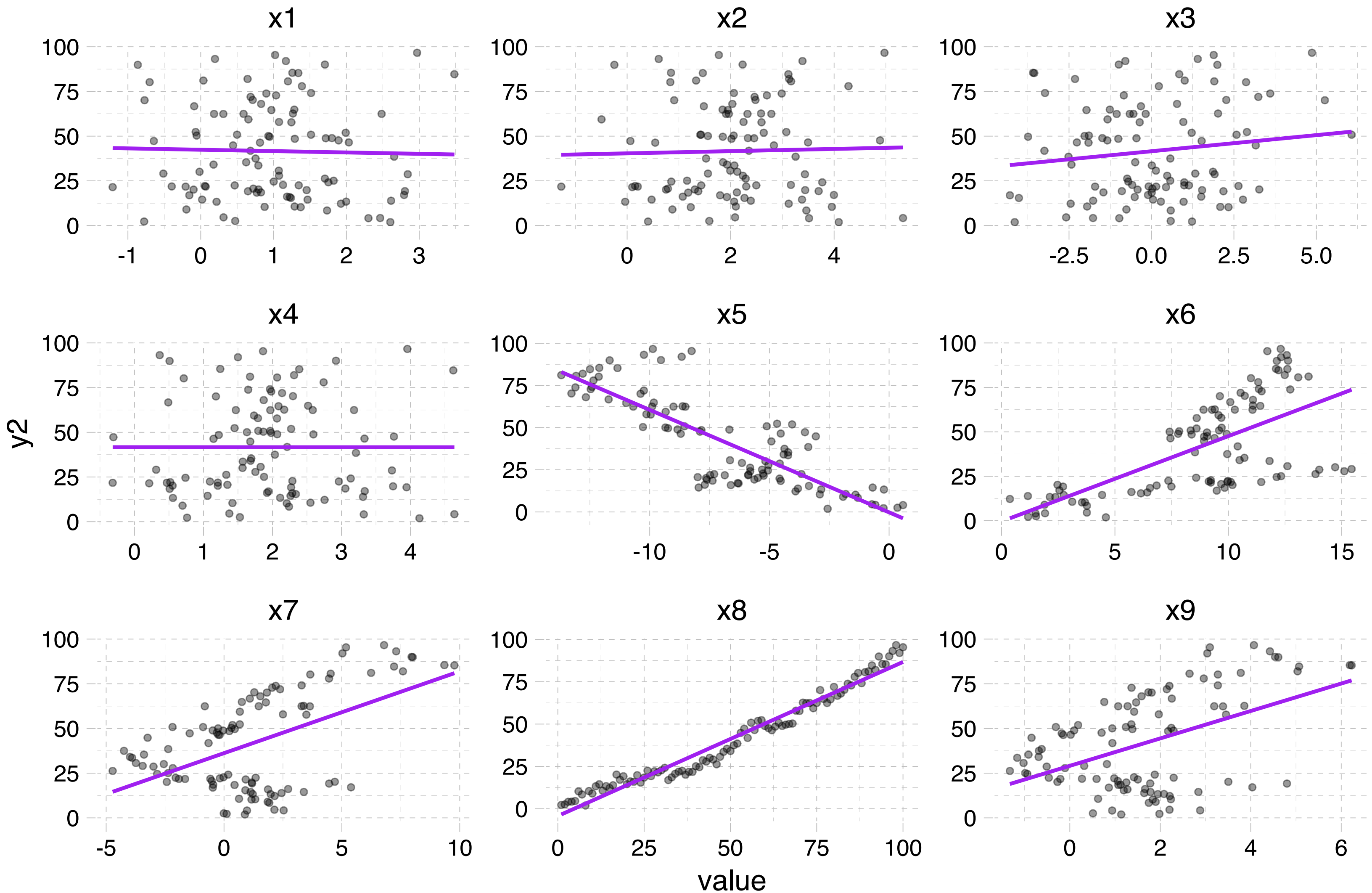
Linear regressions with quadratic RHS: y_1 against x_1 through x_9



Linear regressions with quadratic RHS: y_2 against x_1 through x_9



Simple linear regressions: y_2 against x_1 through x_9



Searching for the unknown model

Results

Your responses: Percentage who said `TRUE` (29 responses)

	X1	X2	X3	X4	X5	X6	X7	X8	X9
y1	78.6	7.1	60.7	39.3	28.6	28.6	17.9	17.9	25.0
y2	46.4	50.0	64.3	10.7	75.0	57.1	75.0	53.6	46.4

Searching for the unknown model

Results

Your responses: Percentage who said `TRUE` (29 responses)

	X1	X2	X3	X4	X5	X6	X7	X8	X9
y1	78.6	7.1	60.7	39.3	28.6	28.6	17.9	17.9	25.0
y2	46.4	50.0	64.3	10.7	75.0	57.1	75.0	53.6	46.4

Truth: The true data-generating processes

$$y_1 = 3 + x_1 - x_3^2 + 2x_4 + u$$
$$y_2 = 1 + x_3 + x_5 + x_7 + v$$

Searching for the unknown model

Results

Your responses: Percentage who said `TRUE` (29 responses)

	X1	X2	X3	X4	X5	X6	X7	X8	X9
y1	78.6	7.1	60.7	39.3	28.6	28.6	17.9	17.9	25.0
y2	46.4	50.0	64.3	10.7	75.0	57.1	75.0	53.6	46.4

Truth: The true data-generating processes

$$y_1 = 3 + x_1 - x_3^2 + 2x_4 + u$$
$$y_2 = 1 + x_3 + x_5 + x_7 + v$$

Q: Is it worse include an incorrect variable or exlcude a correct variable?

Causality

Causality

Intro

Most tasks in econometrics boil down to one of two goals:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

Causality

Intro

Most tasks in econometrics boil down to one of two goals:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

1. **Prediction:** Accurately and dependably predict/forecast y using on some set of explanatory variables—doesn't need to be x_1 through x_k. Focuses on $\hat{y}$. β_j doesn't really matter.

Causality

Intro

Most tasks in econometrics boil down to one of two goals:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

1. **Prediction:** Accurately and dependably predict/forecast y using on some set of explanatory variables—doesn't need to be x_1 through x_k. Focuses on $\hat{y}$. β_j doesn't really matter.

2. **Causal estimation:**[†] Estimate the actual data-generating process—learning about the true, population model that explains how y changes when we change x_j—focuses on β_j. Accuracy of $\hat{y}$ is not important.

† Often called *causal identification*.

Causality

Intro

Most tasks in econometrics boil down to one of two goals:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

1. **Prediction:** Accurately and dependably predict/forecast y using on some set of explanatory variables—doesn't need to be x_1 through x_k. Focuses on $\hat{y}$. β_j doesn't really matter.

2. **Causal estimation:**[†] Estimate the actual data-generating process—learning about the true, population model that explains how y changes when we change x_j—focuses on β_j. Accuracy of $\hat{y}$ is not important.

For the rest of the term, we will focus on **causally estimating** β_j.

† Often called *causal identification*.

Causality

The challenges

As you saw in the data-analysis exercise, determining and estimating the true model can be pretty difficult—both practically and econometrically.

The challenges

As you saw in the data-analysis exercise, determining and estimating the true model can be pretty difficult—both practically and econometrically.

Practical challenges

- Which variables?
- Which functional form(s)?
- Do data exist? How much?
- Is the sample representative?

The challenges

As you saw in the data-analysis exercise, determining and estimating the true model can be pretty difficult—both practically and econometrically.

Practical challenges

- Which variables?
- Which functional form(s)?
- Do data exist? How much?
- Is the sample representative?

Econometric challenges

- Omitted-variable bias
- Reverse causality
- Measurement error
- How precise can/must we be?

The challenges

As you saw in the data-analysis exercise, determining and estimating the true model can be pretty difficult—both practically and econometrically.

Practical challenges

- Which variables?
- Which functional form(s)?
- Do data exist? How much?
- Is the sample representative?

Econometric challenges

- Omitted-variable bias
- Reverse causality
- Measurement error
- How precise can/must we be?

Many of these challenges relate to **exogeneity**, *i.e.*, $\boldsymbol{E}[u_i|X] = 0$.

The challenges

As you saw in the data-analysis exercise, determining and estimating the true model can be pretty difficult—both practically and econometrically.

Practical challenges

- Which variables?
- Which functional form(s)?
- Do data exist? How much?
- Is the sample representative?

Econometric challenges

- Omitted-variable bias
- Reverse causality
- Measurement error
- How precise can/must we be?

Many of these challenges relate to **exogeneity**, *i.e.*, $\boldsymbol{E}[u_i|X] = 0$.
Causality requires us to **hold all else constant** (*ceterus paribus*).

Causality

It's complicated

Occasionally, ***causal*** relationships are simply/easily understood, *e.g.*,

Causality

It's complicated

Occasionally, ***causal*** relationships are simply/easily understood, *e.g.*,

- What caused the forest fire?
- How did this baby get here?

Causality

It's complicated

Occasionally, ***causal*** relationships are simply/easily understood, *e.g.*,

- What caused the forest fire?
- How did this baby get here?

Generally, ***causal*** relationships are complex and challenging to answer, *e.g.*,

Causality

It's complicated

Occasionally, ***causal*** relationships are simply/easily understood, *e.g.*,

- What caused the forest fire?
- How did this baby get here?

Generally, ***causal*** relationships are complex and challenging to answer, *e.g.*,

- What causes some countries to grow and others to decline?
- What caused President Trump's 2016 election?
- How does the number of police officers affect crime?
- What is the effect of better air quality on test scores?
- Do longer prison sentences decrease crime?
- How did cannabis legalization affect mental health/opioid addition?

Correlation ≠ Causation

You've likely heard the saying

> Correlation is not causation.

The saying is just pointing out that there are violations of exogeneity.

Causality

Correlation ≠ Causation

You've likely heard the saying

> Correlation is not causation.

The saying is just pointing out that there are violations of exogeneity.

Although correlation is not causation, **causation *requires* correlation**.

Causality

Correlation ≠ Causation

You've likely heard the saying

> Correlation is not causation.

The saying is just pointing out that there are violations of exogeneity.

Although correlation is not causation, **causation *requires* correlation**.

New saying:

> Correlation plus exogeneity is causation.

Let's work through a few examples.

Causation

Example: The causal effect of fertilizer[†]

Suppose we want to know the causal effect of fertilizer on corn yield.

† Many of the early statistical and econometric studies involved agricultural field trials.

Example: The causal effect of fertilizer†

Suppose we want to know the causal effect of fertilizer on corn yield.

Q: Could we simply regress yield on fertilizer?

† Many of the early statistical and econometric studies involved agricultural field trials.

Causation

Example: The causal effect of fertilizer[†]

Suppose we want to know the causal effect of fertilizer on corn yield.

Q: Could we simply regress yield on fertilizer?

A: Probably not (if we want the causal effect).

† Many of the early statistical and econometric studies involved agricultural field trials.

Example: The causal effect of fertilizer[†]

Suppose we want to know the causal effect of fertilizer on corn yield.

Q: Could we simply regress yield on fertilizer?

A: Probably not (if we want the causal effect).

Q: Why not?

† Many of the early statistical and econometric studies involved agricultural field trials.

Causation

Example: The causal effect of fertilizer†

Suppose we want to know the causal effect of fertilizer on corn yield.

Q: Could we simply regress yield on fertilizer?
A: Probably not (if we want the causal effect).

Q: Why not?
A: Omitted-variable bias: Farmers may apply less fertilizer in areas that are already worse on other dimensions that affect yield (soil, slope, water).
Violates *all else equal* (exogeneity). Biased and/or spurious results.

† Many of the early statistical and econometric studies involved agricultural field trials.

Example: The causal effect of fertilizer†

Suppose we want to know the causal effect of fertilizer on corn yield.

Q: Could we simply regress yield on fertilizer?

A: Probably not (if we want the causal effect).

Q: Why not?

A: Omitted-variable bias: Farmers may apply less fertilizer in areas that are already worse on other dimensions that affect yield (soil, slope, water).

Violates *all else equal* (exogeneity). Biased and/or spurious results.

Q: So what *should* we do?

† Many of the early statistical and econometric studies involved agricultural field trials.

Example: The causal effect of fertilizer[†]

Suppose we want to know the causal effect of fertilizer on corn yield.

Q: Could we simply regress yield on fertilizer?
A: Probably not (if we want the causal effect).

Q: Why not?
A: Omitted-variable bias: Farmers may apply less fertilizer in areas that are already worse on other dimensions that affect yield (soil, slope, water).
Violates *all else equal* (exogeneity). Biased and/or spurious results.

Q: So what *should* we do?
A: **Run an experiment!**

† Many of the early statistical and econometric studies involved agricultural field trials.

Example: The causal effect of fertilizer†

Suppose we want to know the causal effect of fertilizer on corn yield.

Q: Could we simply regress yield on fertilizer?

A: Probably not (if we want the causal effect).

Q: Why not?

A: Omitted-variable bias: Farmers may apply less fertilizer in areas that are already worse on other dimensions that affect yield (soil, slope, water).

Violates *all else equal* (exogeneity). Biased and/or spurious results.

Q: So what *should* we do?

A: Run an experiment! 💩

† Many of the early statistical and econometric studies involved agricultural field trials.

Causation

Example: The causal effect of fertilizer

Randomized experiments help us maintain *all else equal* (exogeneity).

Causation

Example: The causal effect of fertilizer

Randomized experiments help us maintain *all else equal* (exogeneity).

We often call these experiments ***randomized control trials*** (RCTs).[†]

† Econometrics (and statistics) borrows this language from biostatistics and pharmaceutical trials.

Example: The causal effect of fertilizer

Randomized experiments help us maintain *all else equal* (exogeneity).

We often call these experiments ***randomized control trials*** (RCTs).[†]

Imagine an RCT where we have two groups:

- **Treatment:** We apply fertilizer.
- **Control:** We do not apply fertilizer.

† Econometrics (and statistics) borrows this language from biostatistics and pharmaceutical trials.

Example: The causal effect of fertilizer

Randomized experiments help us maintain *all else equal* (exogeneity).

We often call these experiments ***randomized control trials*** (RCTs).[†]

Imagine an RCT where we have two groups:

- **Treatment:** We apply fertilizer.
- **Control:** We do not apply fertilizer.

By randomizing plots of land into **treatment** or **control**, we will, on average, include all kinds of land (soild, slope, water, *etc.*) in both groups.

† Econometrics (and statistics) borrows this language from biostatistics and pharmaceutical trials.

Example: The causal effect of fertilizer

Randomized experiments help us maintain *all else equal* (exogeneity).

We often call these experiments ***randomized control trials*** (RCTs).[†]

Imagine an RCT where we have two groups:

- **Treatment:** We apply fertilizer.
- **Control:** We do not apply fertilizer.

By randomizing plots of land into **treatment** or **control**, we will, on average, include all kinds of land (soild, slope, water, *etc.*) in both groups.

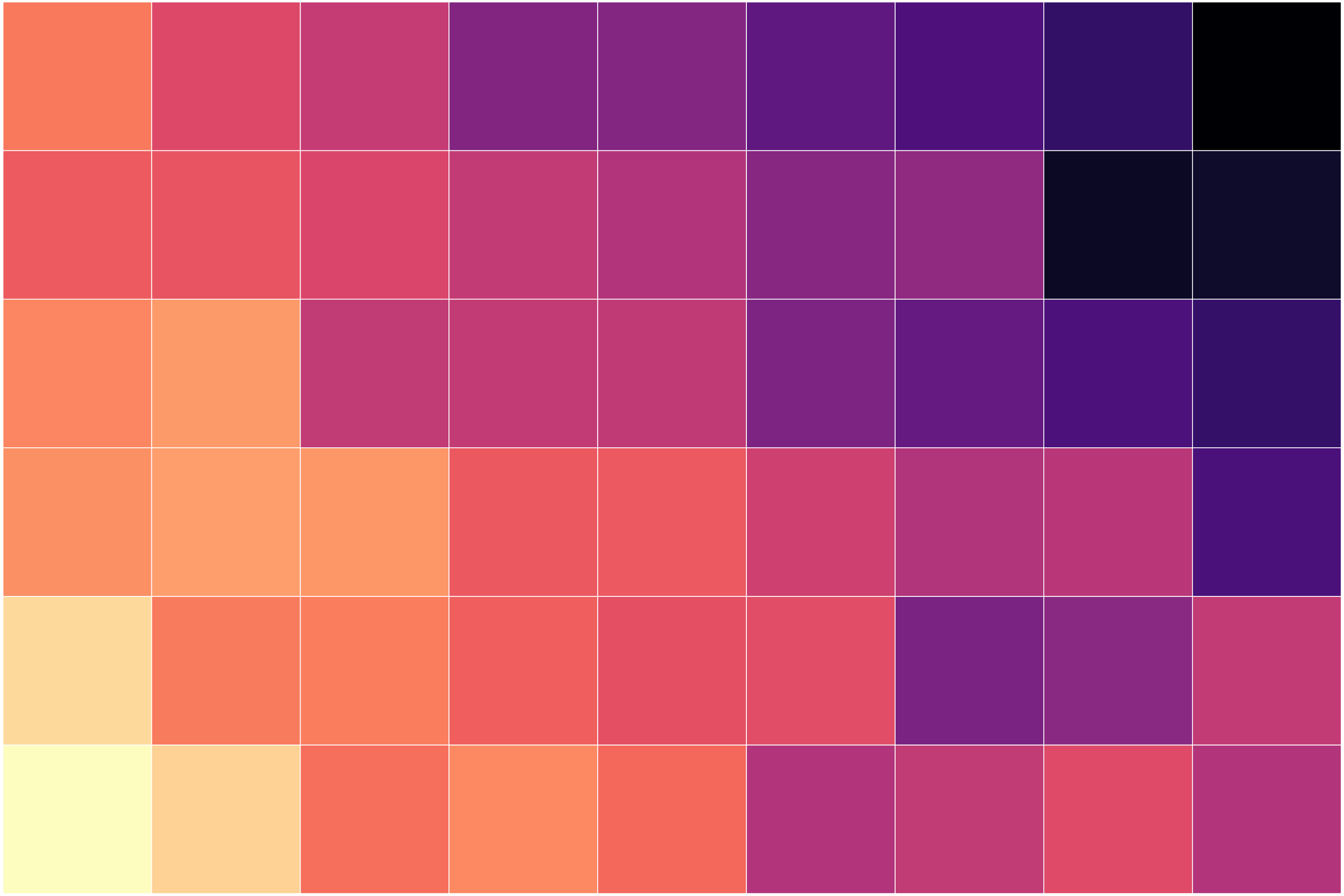
All else equal!

† Econometrics (and statistics) borrows this language from biostatistics and pharmaceutical trials.

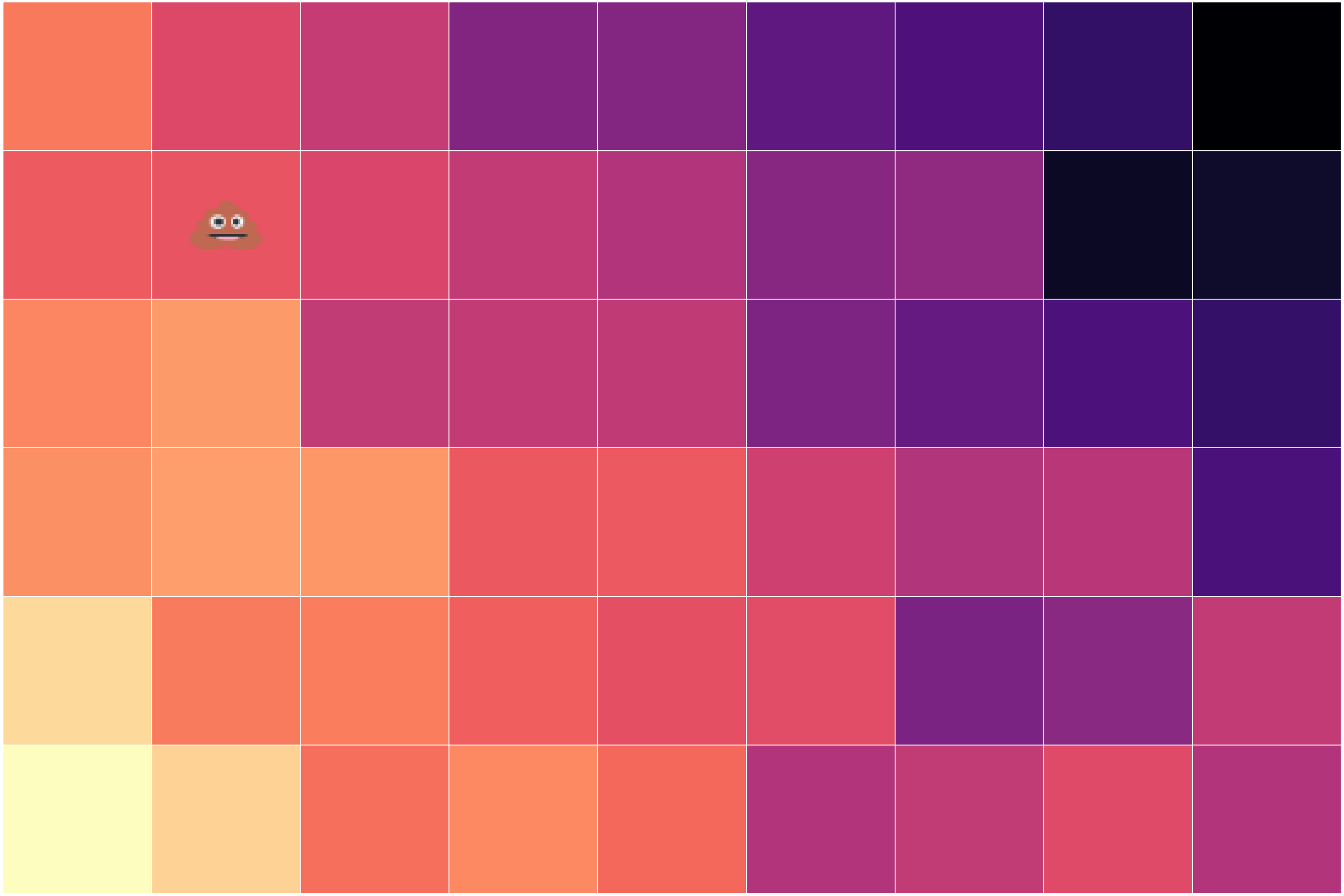
54 equal-sized plots

01	02	03	04	05	06	07	08	09
10	11	12	13	14	15	16	17	18
19	20	21	22	23	24	25	26	27
28	29	30	31	32	33	34	35	36
37	38	39	40	41	42	43	44	45
46	47	48	49	50	51	52	53	54

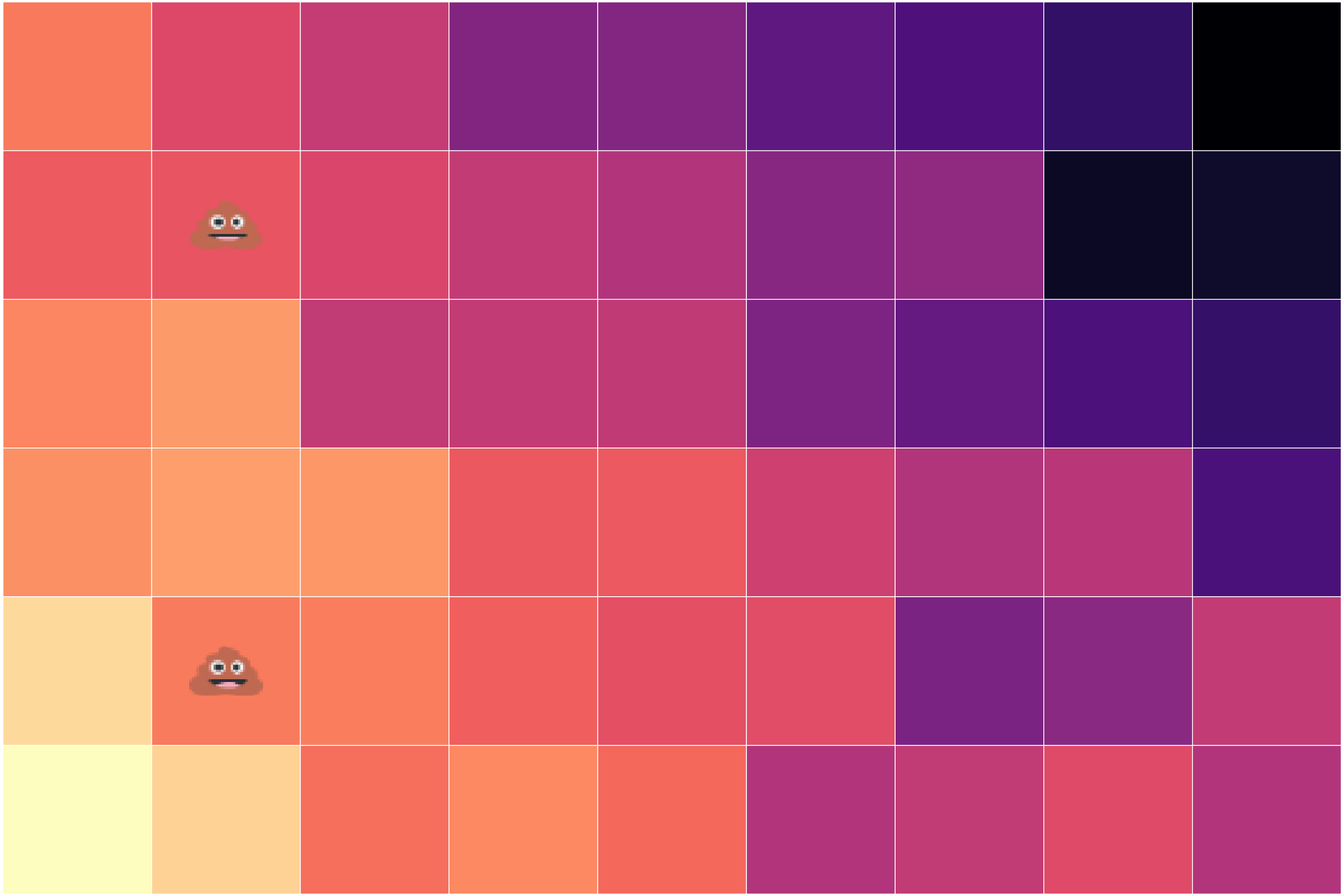
54 equal-sized plots of varying quality



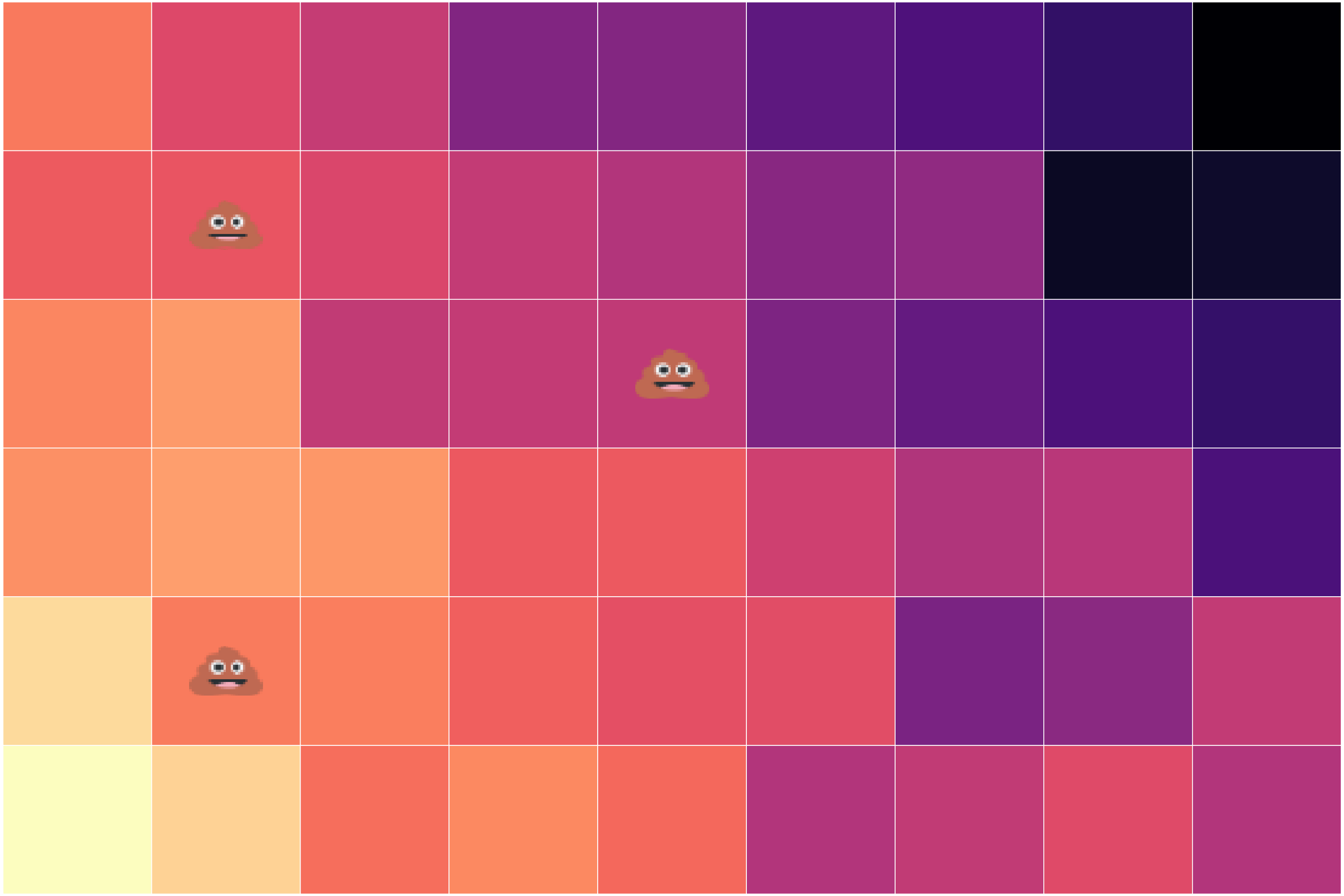
54 equal-sized plots of varying quality plus randomly assigned treatment



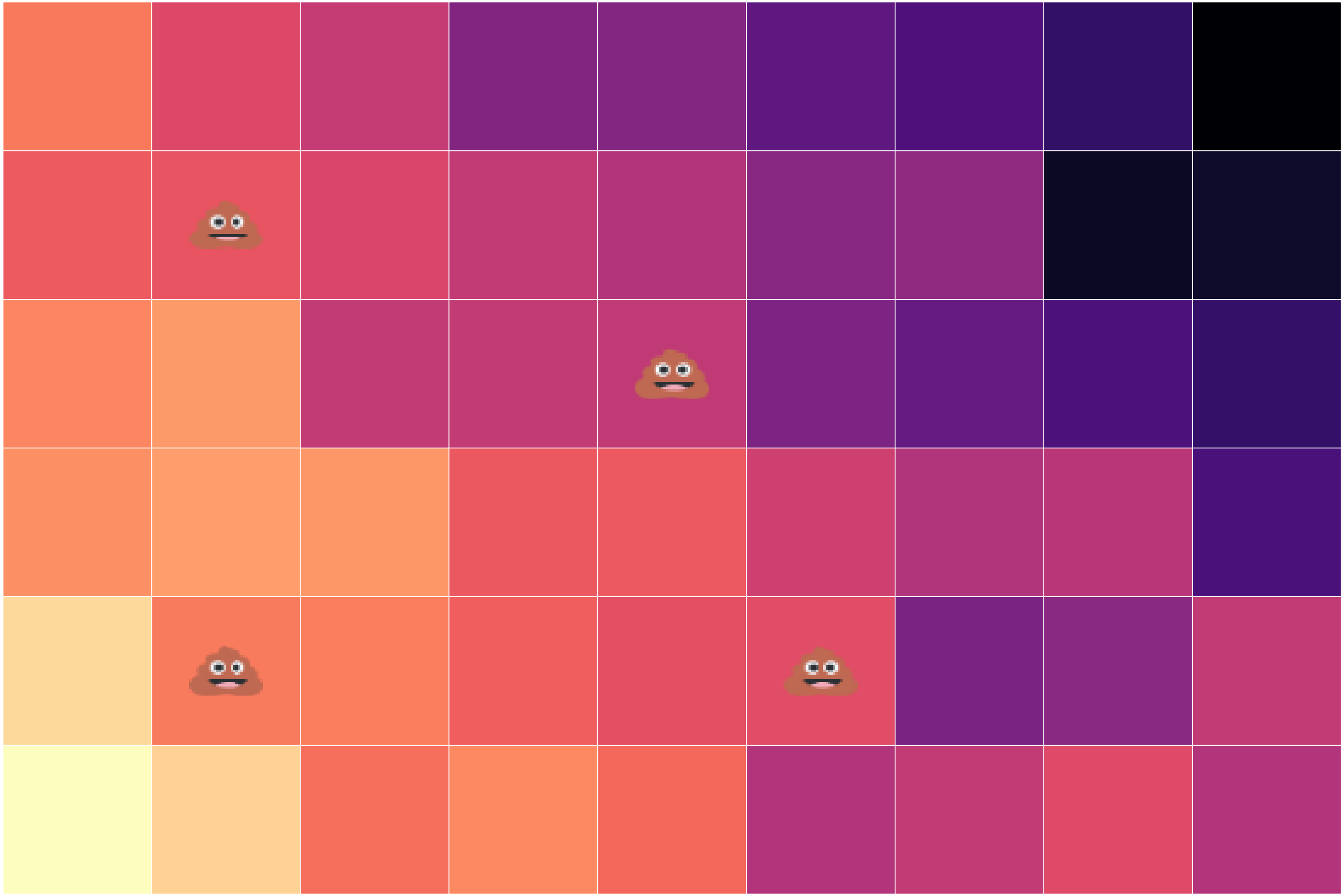
54 equal-sized plots of varying quality plus randomly assigned treatment



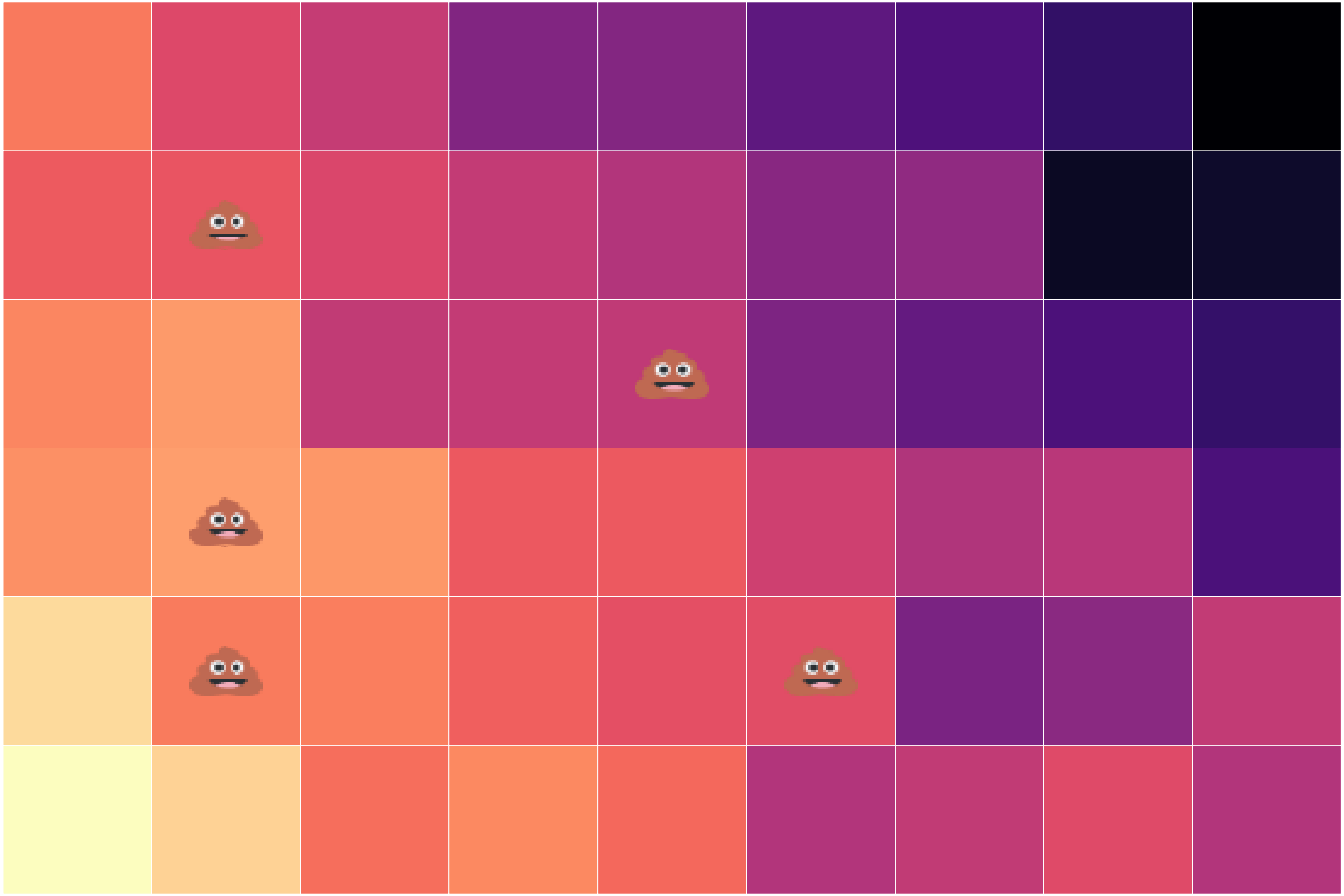
54 equal-sized plots of varying quality plus randomly assigned treatment



54 equal-sized plots of varying quality plus randomly assigned treatment

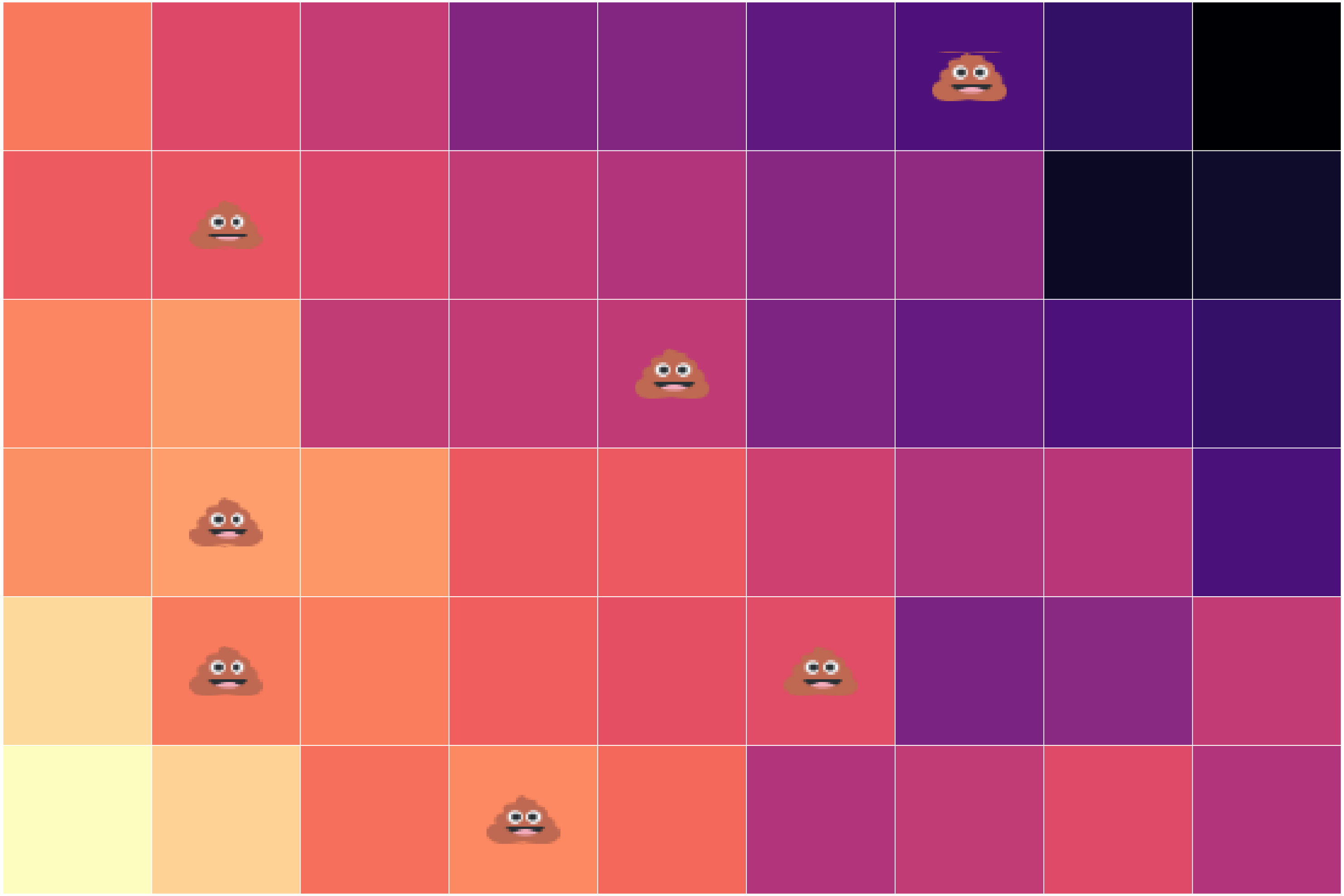


54 equal-sized plots of varying quality plus randomly assigned treatment

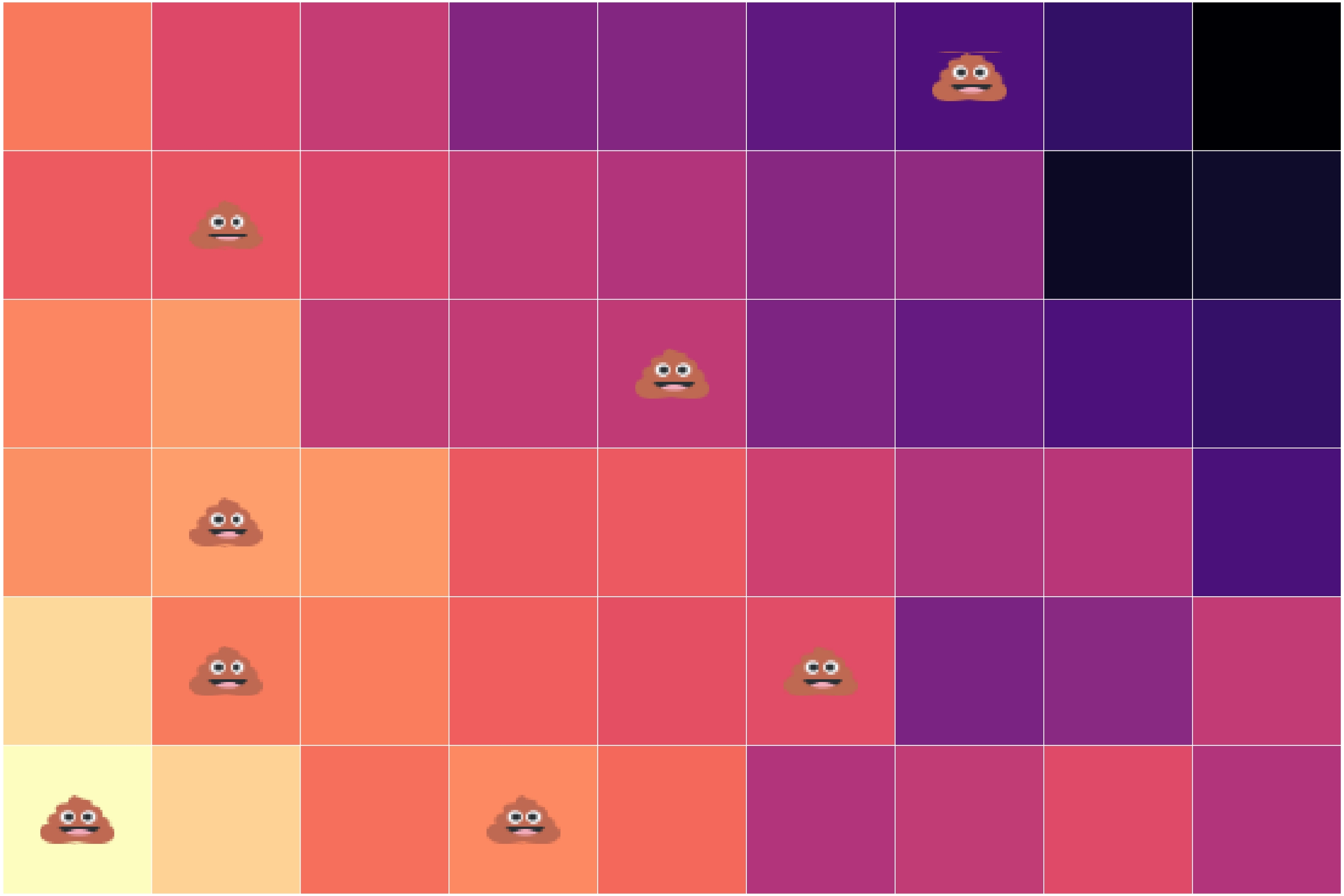


54 equal-sized plots of varying quality plus randomly assigned treatment

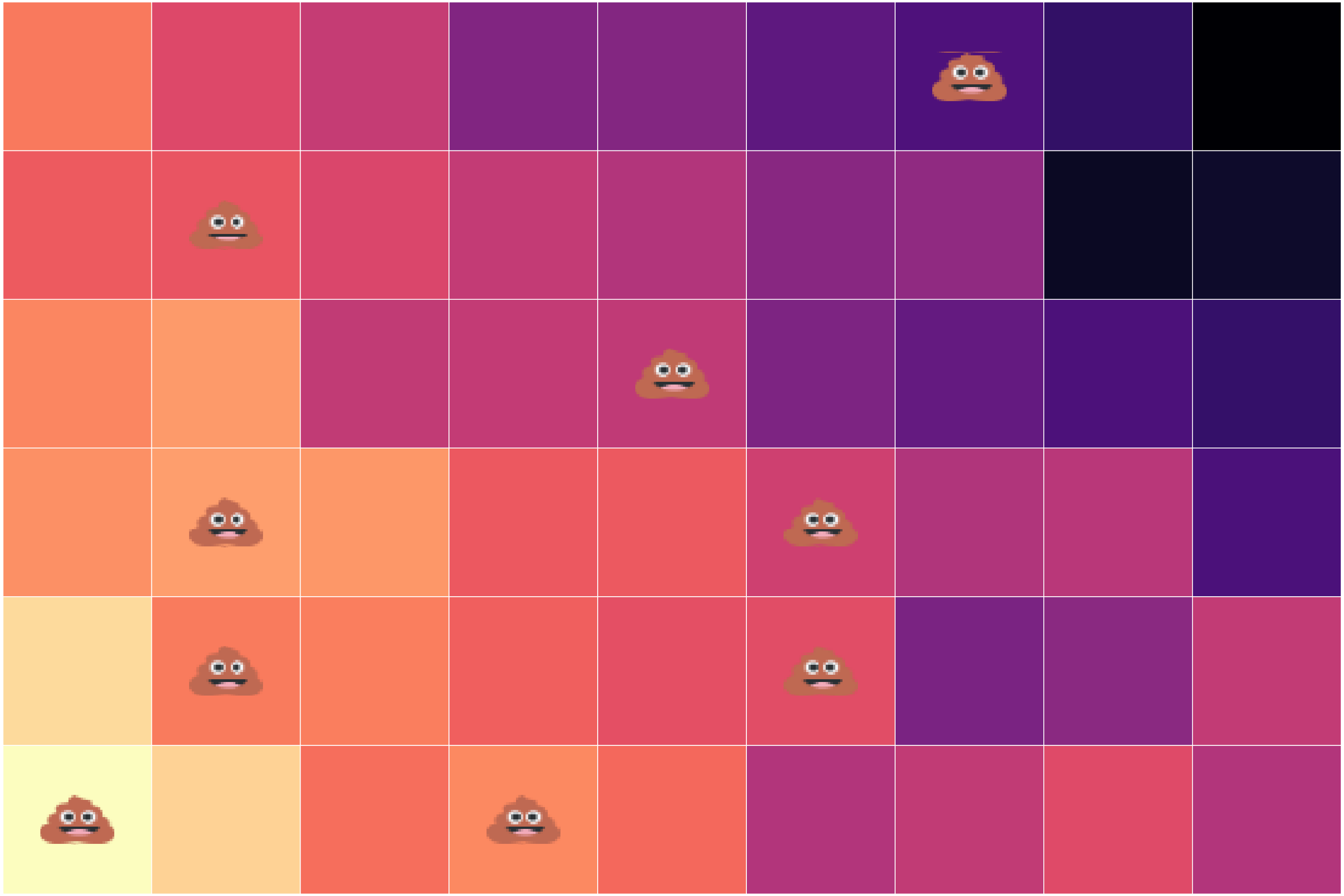
54 equal-sized plots of varying quality plus randomly assigned treatment



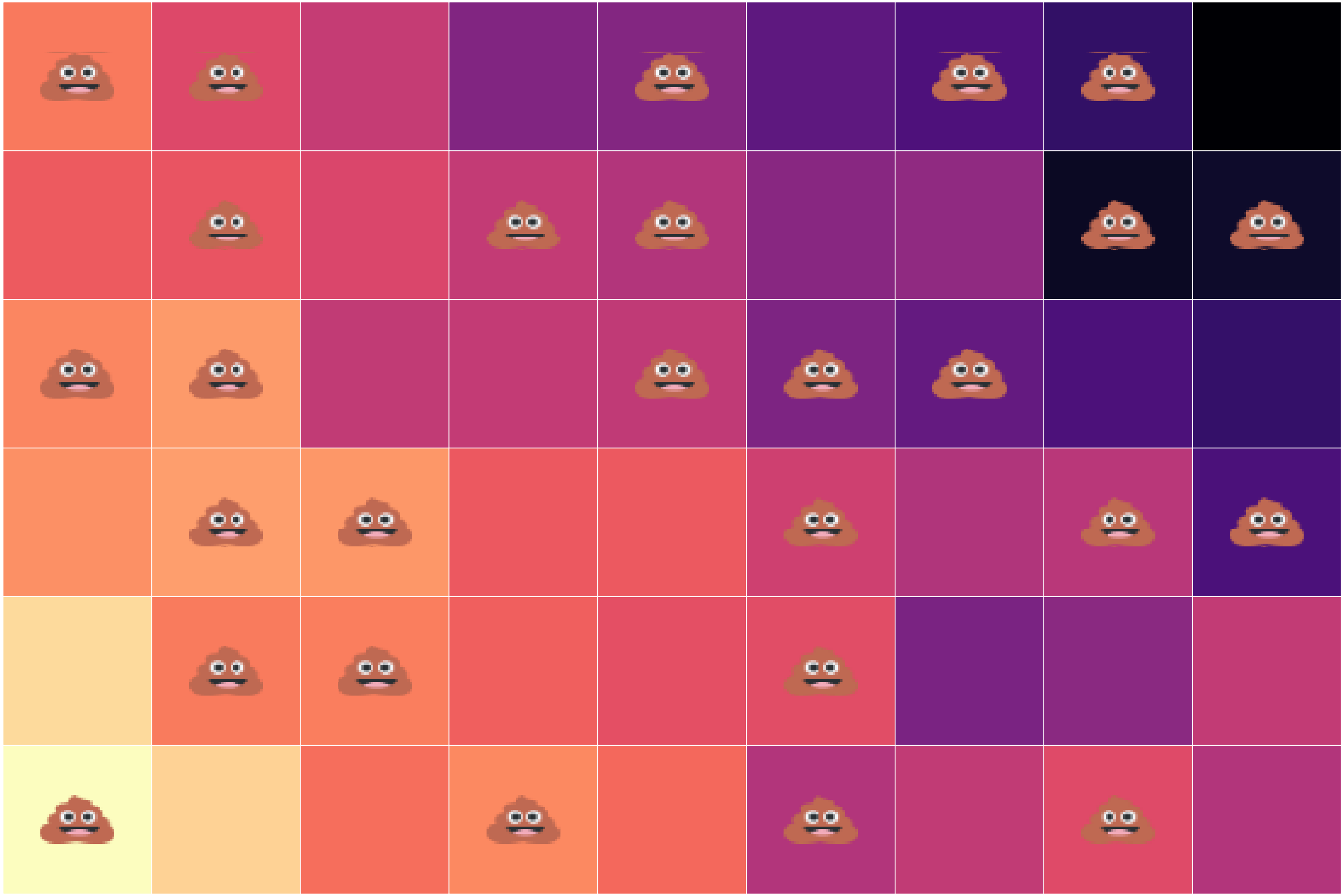
54 equal-sized plots of varying quality plus randomly assigned treatment



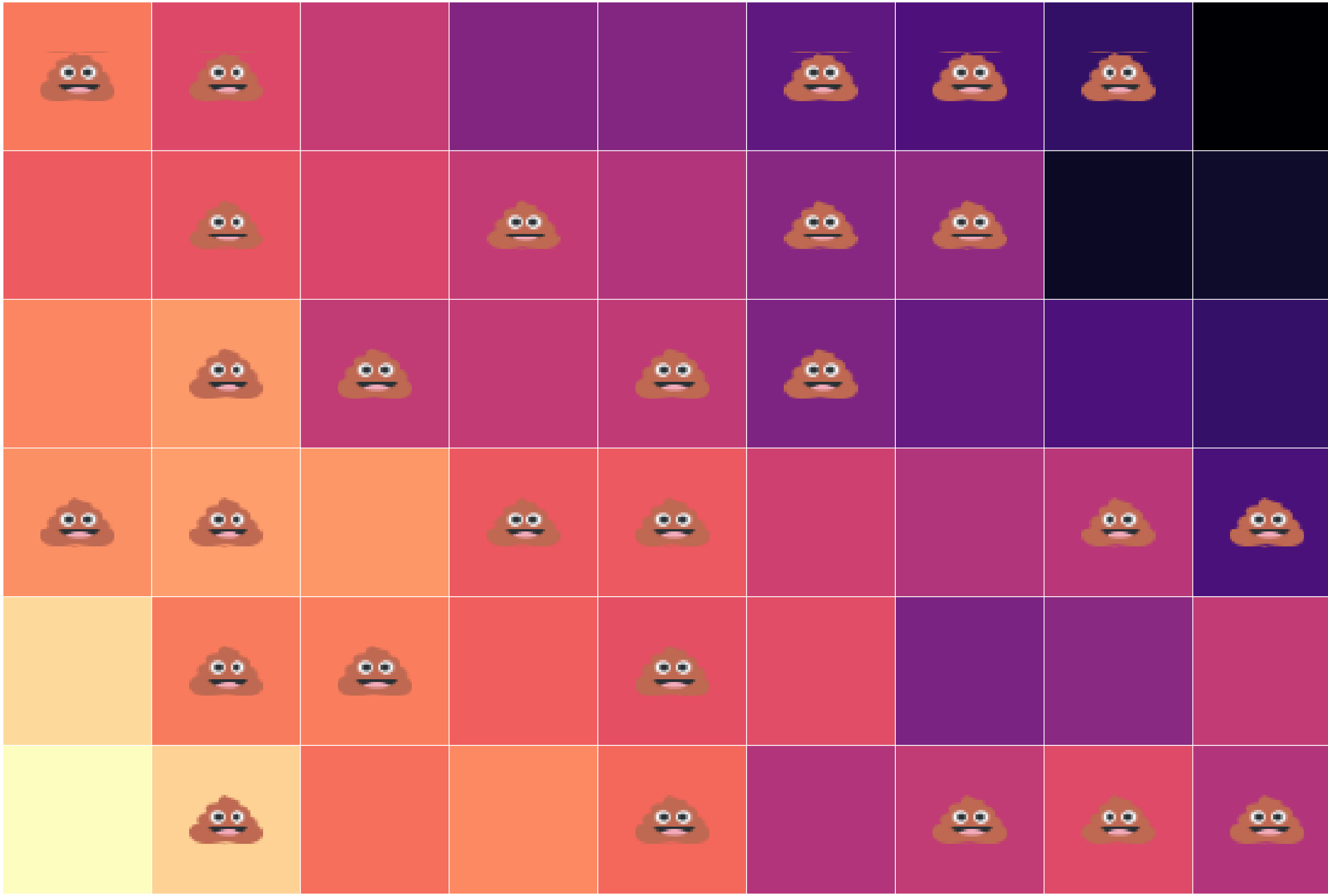
54 equal-sized plots of varying quality plus randomly assigned treatment



54 equal-sized plots of varying quality plus randomly assigned treatment



54 equal-sized plots of varying quality plus randomly assigned treatment



54 equal-sized plots of varying quality plus randomly assigned treatment

Example: The causal effect of fertilizer

We can estimate the **causal effect** of fertilizer on crop yield by comparing the average yield in the treatment group (💩) with the control group (no 💩).

$$\overline{\text{Yield}}_{\text{Treatment}} - \overline{\text{Yield}}_{\text{Control}}$$

Example: The causal effect of fertilizer

We can estimate the **causal effect** of fertilizer on crop yield by comparing the average yield in the treatment group (💩) with the control group (no 💩).

$$\overline{\text{Yield}}_{\text{Treatment}} - \overline{\text{Yield}}_{\text{Control}}$$

Alternatively, we can use the regression

Example: The causal effect of fertilizer

We can estimate the **causal effect** of fertilizer on crop yield by comparing the average yield in the treatment group (💩) with the control group (no 💩).

$$\overline{\text{Yield}}_{\text{Treatment}} - \overline{\text{Yield}}_{\text{Control}}$$

Alternatively, we can use the regression

$$\text{Yield}_i = \beta_0 + \beta_1 \text{Trt}_i + u_i \tag{1}$$

where Trt_i is a binary variable (=1 if plot i received the fertilizer treatment).

Causation

Example: The causal effect of fertilizer

We can estimate the **causal effect** of fertilizer on crop yield by comparing the average yield in the treatment group (💩) with the control group (no 💩).

$$\overline{\text{Yield}}_{\text{Treatment}} - \overline{\text{Yield}}_{\text{Control}}$$

Alternatively, we can use the regression

$$\text{Yield}_i = \beta_0 + \beta_1 \text{Trt}_i + u_i \tag{1}$$

where Trt_i is a binary variable (=1 if plot i received the fertilizer treatment).

Q: Should we expect (1) to satisfy exogeneity? Why?

Causation

Example: The causal effect of fertilizer

We can estimate the **causal effect** of fertilizer on crop yield by comparing the average yield in the treatment group (💩) with the control group (no 💩).

$$\overline{\text{Yield}}_{\text{Treatment}} - \overline{\text{Yield}}_{\text{Control}}$$

Alternatively, we can use the regression

$$\text{Yield}_i = \beta_0 + \beta_1 \text{Trt}_i + u_i \tag{1}$$

where Trt_i is a binary variable (=1 if plot i received the fertilizer treatment).

Q: Should we expect (1) to satisfy exogeneity? Why?

A: On average, **randomly assigning treatment should balance** trt. and control across the other dimensions that affect yield (soil, slope, water).

Causality

Example: Returns to education

Labor economists, policy makers, parents, and students are all interested in the (monetary) *return to education.*

Example: Returns to education

Labor economists, policy makers, parents, and students are all interested in the (monetary) *return to education.*

Thought experiment:

- Randomly select an individual.
- Give her an additional year of education.
- How much do her earnings increase?

This change in earnings gives the **causal effect** of education on earnings.

Causality

Example: Returns to education

Q: Could we simply regress earnings on education?

Causality

Example: Returns to education

Q: Could we simply regress earnings on education?

A: Again, probably not if we want the true, causal effect.

Causality

Example: Returns to education

Q: Could we simply regress earnings on education?

A: Again, probably not if we want the true, causal effect.

1. People *choose* education based upon many factors, *e.g.*, ability.
2. Education likely reduces experience (time out of the workforce).
3. Education is ***endogenous*** (violates *exogeneity*).

Causality

Example: Returns to education

Q: Could we simply regress earnings on education?

A: Again, probably not if we want the true, causal effect.

1. People *choose* education based upon many factors, *e.g.*, ability.
2. Education likely reduces experience (time out of the workforce).
3. Education is ***endogenous*** (violates *exogeneity*).

The point (2) above also illustrates the difficulty in learning about educations while *holding all else constant*.

Many important variables have the same challenge—gender, race, income.

Causality

Example: Returns to education

Q: So how can we estimate the returns to education?

Example: Returns to education

Q: So how can we estimate the returns to education?

Option 1: Run an **experiment**.

Example: Returns to education

Q: So how can we estimate the returns to education?

Option 1: Run an **experiment**.

- Randomly assign education (might be difficult).
- Randomly encourage education (might work).
- Randomly assign programs that affect education (*e.g.*, mentoring).

Example: Returns to education

Q: So how can we estimate the returns to education?

Option 1: Run an **experiment**.

- Randomly assign education (might be difficult).
- Randomly encourage education (might work).
- Randomly assign programs that affect education (*e.g.*, mentoring).

Option 2: Look for a ***natural experiment***—a policy or accident in society that arbitrarily increased education for one subset of people.

Example: Returns to education

Q: So how can we estimate the returns to education?

Option 1: Run an **experiment**.

- Randomly assign education (might be difficult).
- Randomly encourage education (might work).
- Randomly assign programs that affect education (*e.g.*, mentoring).

Option 2: Look for a ***natural experiment***—a policy or accident in society that arbitrarily increased education for one subset of people.

- Admissions cutoffs
- Lottery enrollment and/or capacity constraints

Causality

Real-world experiments

Both examples consider **real experiments** that isolate causal effects.

Characteristics

- Feasible—we can actually (potentially) run the experiment.
- Compare individuals randomized into treatment against individuals randomized into control.
- Require "good" randomization to get *all else equal* (exogeneity).

Real-world experiments

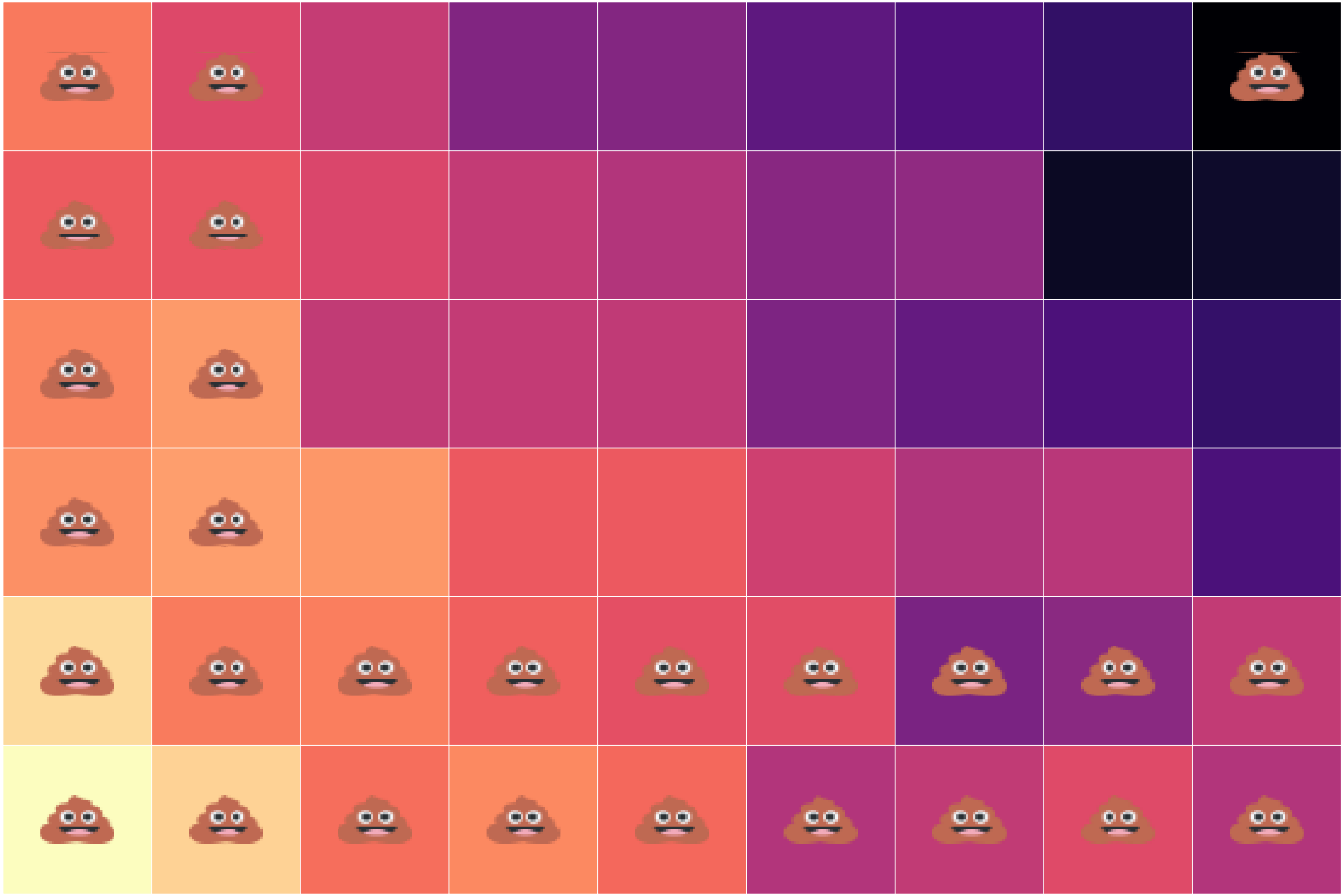
Both examples consider **real experiments** that isolate causal effects.

Characteristics

- Feasible—we can actually (potentially) run the experiment.
- Compare individuals randomized into treatment against individuals randomized into control.
- Require "good" randomization to get *all else equal* (exogeneity).

Note: Your experiment's results are only as good as your randomization.

Unfortunate randomization



Causality

The ideal experiment

The **ideal experiment** would be subtly different.

Rather than comparing units randomized as treatment vs. control, the ideal experiment would compare treatment and control **for the same, exact unit**.

The ideal experiment

The **ideal experiment** would be subtly different.

Rather than comparing units randomized as treatment vs. control, the ideal experiment would compare treatment and control **for the same, exact unit**.

$$y_{\text{Treatment},i} - y_{\text{Control},i}$$

The ideal experiment

The **ideal experiment** would be subtly different.

Rather than comparing units randomized as treatment vs. control, the ideal experiment would compare treatment and control **for the same, exact unit**.

$$y_{\text{Treatment},i} - y_{\text{Control},i}$$

which we will write (for simplicity) as

$$y_{1,i} - y_{0,i}$$

Causality

The ideal experiment

The **ideal experiment** would be subtly different.

Rather than comparing units randomized as treatment vs. control, the ideal experiment would compare treatment and control **for the same, exact unit**.

$$y_{\text{Treatment},i} - y_{\text{Control},i}$$

which we will write (for simplicity) as

$$y_{1,i} - y_{0,i}$$

This *ideal experiment* is clearly infeasible[†], but it creates nice notation for causality (the Rubin causal model/Neyman potential outcomes framework).

† Without (1) God-like abilities and multiple universes or (2) a time machine.

The ideal experiment

The *ideal* data for 10 people

```
#>      i trt  y1i  y0i
#> 1    1   1 5.01 2.56
#> 2    2   1 8.85 2.53
#> 3    3   1 6.31 2.67
#> 4    4   1 5.97 2.79
#> 5    5   1 7.61 4.34
#> 6    6   0 7.63 4.15
#> 7    7   0 4.75 0.56
#> 8    8   0 5.77 3.52
#> 9    9   0 7.47 4.49
#> 10  10   0 7.79 1.40
```

Causality

The ideal experiment

The *ideal* data for 10 people

```
#>     i trt  y1i  y0i
#> 1   1   1 5.01 2.56
#> 2   2   1 8.85 2.53
#> 3   3   1 6.31 2.67
#> 4   4   1 5.97 2.79
#> 5   5   1 7.61 4.34
#> 6   6   0 7.63 4.15
#> 7   7   0 4.75 0.56
#> 8   8   0 5.77 3.52
#> 9   9   0 7.47 4.49
#> 10 10   0 7.79 1.40
```

Calculate the causal effect of trt.

$$\tau_i = y_{1,i} - y_{0,i}$$

for each individual i.

Causality

The ideal experiment

The *ideal* data for 10 people

```
#>     i trt  y1i  y0i effect_i
#> 1   1   1 5.01 2.56     2.45
#> 2   2   1 8.85 2.53     6.32
#> 3   3   1 6.31 2.67     3.64
#> 4   4   1 5.97 2.79     3.18
#> 5   5   1 7.61 4.34     3.27
#> 6   6   0 7.63 4.15     3.48
#> 7   7   0 4.75 0.56     4.19
#> 8   8   0 5.77 3.52     2.25
#> 9   9   0 7.47 4.49     2.98
#> 10 10   0 7.79 1.40     6.39
```

Calculate the causal effect of trt.

$$\tau_i = y_{1,i} - y_{0,i}$$

for each individual i.

The ideal experiment

The *ideal* data for 10 people

```
#>     i trt  y1i  y0i effect_i
#> 1   1   1 5.01 2.56     2.45
#> 2   2   1 8.85 2.53     6.32
#> 3   3   1 6.31 2.67     3.64
#> 4   4   1 5.97 2.79     3.18
#> 5   5   1 7.61 4.34     3.27
#> 6   6   0 7.63 4.15     3.48
#> 7   7   0 4.75 0.56     4.19
#> 8   8   0 5.77 3.52     2.25
#> 9   9   0 7.47 4.49     2.98
#> 10 10   0 7.79 1.40     6.39
```

Calculate the causal effect of trt.

$$\tau_i = y_{1,i} - y_{0,i}$$

for each individual i.

The mean of τ_i is the **average treatment effect** (ATE).

Thus, $\overline{\tau} = 3.82$

The ideal experiment

This model highlights the fundamental problem of causal inference.

$$\tau_i = y_{1,i} - y_{0,i}$$

Causality

The ideal experiment

This model highlights the fundamental problem of causal inference.

$$\tau_i = y_{1,i} - y_{0,i}$$

The challenge:

If we observe $y_{1,i}$, then we cannot observe $y_{0,i}$.
If we observe $y_{0,i}$, then we cannot observe $y_{1,i}$.

The ideal experiment

So a dataset that we actually observe for 6 people will look something like

```
#>      i trt  y1i  y0i
#> 1    1   1 5.01   NA
#> 2    2   1 8.85   NA
#> 3    3   1 6.31   NA
#> 4    4   1 5.97   NA
#> 5    5   1 7.61   NA
#> 6    6   0   NA 4.15
#> 7    7   0   NA 0.56
#> 8    8   0   NA 3.52
#> 9    9   0   NA 4.49
#> 10 10   0   NA 1.40
```

Causality

The ideal experiment

So a dataset that we actually observe for 6 people will look something like

```
#>      i trt  y1i  y0i
#> 1    1   1 5.01   NA
#> 2    2   1 8.85   NA
#> 3    3   1 6.31   NA
#> 4    4   1 5.97   NA
#> 5    5   1 7.61   NA
#> 6    6   0   NA 4.15
#> 7    7   0   NA 0.56
#> 8    8   0   NA 3.52
#> 9    9   0   NA 4.49
#> 10  10   0   NA 1.40
```

We can't observe $y_{1,i}$ and $y_{0,i}$.

But, we do observe

- $y_{1,i}$ for i in 1, 2, 3, 4, 5
- $y_{0,j}$ for j in 6, 7, 8, 9, 10

Causality

The ideal experiment

So a dataset that we actually observe for 6 people will look something like

```
#>     i trt  y1i  y0i
#> 1   1   1 5.01   NA
#> 2   2   1 8.85   NA
#> 3   3   1 6.31   NA
#> 4   4   1 5.97   NA
#> 5   5   1 7.61   NA
#> 6   6   0   NA 4.15
#> 7   7   0   NA 0.56
#> 8   8   0   NA 3.52
#> 9   9   0   NA 4.49
#> 10 10   0   NA 1.40
```

We can't observe $y_{1,i}$ and $y_{0,i}$.

But, we do observe

- $y_{1,i}$ for i in 1, 2, 3, 4, 5
- $y_{0,j}$ for j in 6, 7, 8, 9, 10

Q: How do we "fill in" the `NA`s and estimate $\overline{\tau}$?

Causally estimating the treatment effect

Notation: Let D_i be a binary indicator variable such that

- $D_i = 1$ if individual i is treated.
- $D_i = 0$ if individual i is not treated (*control* group).

Causally estimating the treatment effect

Notation: Let D_i be a binary indicator variable such that

- $D_i = 1$ if individual i is treated.
- $D_i = 0$ if individual i is not treated (*control* group).

Then, rephrasing the previous slide,

- We only observe $y_{1,i}$ when $D_i = 1$.
- We only observe $y_{0,i}$ when $D_i = 0$.

Causally estimating the treatment effect

Notation: Let D_i be a binary indicator variable such that

- $D_i = 1$ if individual i is treated.
- $D_i = 0$ if individual i is not treated (*control* group).

Then, rephrasing the previous slide,

- We only observe $y_{1,i}$ when $D_i = 1$.
- We only observe $y_{0,i}$ when $D_i = 0$.

Q: How can we estimate $\bar{\tau}$ using only $(y_{1,i}|D_i = 1)$ and $(y_{0,i}|D_i = 0)$?

Causally estimating the treatment effect

Q: How can we estimate $\overline{\tau}$ using only $(y_{1,i}|D_i = 1)$ and $(y_{0,i}|D_i = 0)$?

Causally estimating the treatment effect

Q: How can we estimate $\overline{\tau}$ using only $(y_{1,i}|D_i = 1)$ and $(y_{0,i}|D_i = 0)$?

Idea: What if we compare the groups' means? *I.e.*,

$$Avg(y_i \mid D_i = 1) - Avg(y_i \mid D_i = 0)$$

Causally estimating the treatment effect

Q: How can we estimate $\overline{\tau}$ using only $(y_{1,i}|D_i = 1)$ and $(y_{0,i}|D_i = 0)$?

Idea: What if we compare the groups' means? *I.e.*,

$$Avg(y_i \mid D_i = 1) - Avg(y_i \mid D_i = 0)$$

Q: When does this simple difference in groups' means provide information on the **causal effect** of the treatment?

Causally estimating the treatment effect

Q: How can we estimate $\overline{\tau}$ using only $(y_{1,i}|D_i = 1)$ and $(y_{0,i}|D_i = 0)$?

Idea: What if we compare the groups' means? *I.e.*,

$$Avg(y_i \mid D_i = 1) - Avg(y_i \mid D_i = 0)$$

Q: When does this simple difference in groups' means provide information on the **causal effect** of the treatment?

Q$_{2.0}$: Is $Avg(y_i \mid D_i = 1) - Avg(y_i \mid D_i = 0)$ a *good* estimator for $\overline{\tau}$?

Causally estimating the treatment effect

Q: How can we estimate $\overline{\tau}$ using only $(y_{1,i} | D_i = 1)$ and $(y_{0,i} | D_i = 0)$?

Idea: What if we compare the groups' means? *I.e.*,

$$Avg(y_i \mid D_i = 1) - Avg(y_i \mid D_i = 0)$$

Q: When does this simple difference in groups' means provide information on the **causal effect** of the treatment?

Q$_{2.0}$: Is $Avg(y_i \mid D_i = 1) - Avg(y_i \mid D_i = 0)$ a *good* estimator for $\overline{\tau}$?

Time for math! 🎉

Causally estimating the treatment effect

Assumption: Let $\tau_i = \tau$ for all i.

This assumption says that the treatment effect is equal (constant) across all individuals i.

Causally estimating the treatment effect

Assumption: Let $\tau_i = \tau$ for all i.

This assumption says that the treatment effect is equal (constant) across all individuals i.

Note: We defined

$$\tau_i = \tau = y_{1,i} - y_{0,i}$$

which implies

$$y_{1,i} = y_{0,i} + \tau$$

$Q_{3.0}$: Is $Avg(y_i \mid D_i = 1) - Avg(y_i \mid D_i = 0)$ a *good* estimator for τ?

Q$_{3.0}$: Is $Avg(y_i \mid D_i = 1) - Avg(y_i \mid D_i = 0)$ a *good* estimator for τ?

Difference in groups' means

$Q_{3.0}$: Is $Avg(y_i \mid D_i = 1) - Avg(y_i \mid D_i = 0)$ a *good* estimator for τ?

Difference in groups' means

$$= Avg(y_i \mid D_i = 1) - Avg(y_i \mid D_i = 0)$$

$Q_{3.0}$: Is $Avg(y_i \mid D_i = 1) - Avg(y_i \mid D_i = 0)$ a *good* estimator for τ?

Difference in groups' means

$$= Avg(y_i \mid D_i = 1) - Avg(y_i \mid D_i = 0)$$

$$= Avg(y_{1,i} \mid D_i = 1) - Avg(y_{0,i} \mid D_i = 0)$$

$\mathbf{Q_{3.0}}$: Is $Avg(y_i \mid D_i = 1) - Avg(y_i \mid D_i = 0)$ a *good* estimator for τ?

Difference in groups' means

$$= Avg(y_i \mid D_i = 1) - Avg(y_i \mid D_i = 0)$$

$$= Avg(y_{1,i} \mid D_i = 1) - Avg(y_{0,i} \mid D_i = 0)$$

$$= Avg(\tau + y_{0,i} \mid D_i = 1) - Avg(y_{0,i} \mid D_i = 0)$$

$Q_{3.0}$: Is $Avg(y_i \mid D_i = 1) - Avg(y_i \mid D_i = 0)$ a *good* estimator for $\boldsymbol{\tau}$?

Difference in groups' means

$$= Avg(y_i \mid D_i = 1) - Avg(y_i \mid D_i = 0)$$

$$= Avg(y_{1,i} \mid D_i = 1) - Avg(y_{0,i} \mid D_i = 0)$$

$$= Avg(\boldsymbol{\tau} + y_{0,i} \mid D_i = 1) - Avg(y_{0,i} \mid D_i = 0)$$

$$= \boldsymbol{\tau} + Avg(y_{0,i} \mid D_i = 1) - Avg(y_{0,i} \mid D_i = 0)$$

$Q_{3.0}$: Is $Avg(y_i \mid D_i = 1) - Avg(y_i \mid D_i = 0)$ a *good* estimator for $\boldsymbol{\tau}$?

Difference in groups' means

$$= Avg(y_i \mid D_i = 1) - Avg(y_i \mid D_i = 0)$$

$$= Avg(y_{1,i} \mid D_i = 1) - Avg(y_{0,i} \mid D_i = 0)$$

$$= Avg(\boldsymbol{\tau} + y_{0,i} \mid D_i = 1) - Avg(y_{0,i} \mid D_i = 0)$$

$$= \boldsymbol{\tau} + Avg(y_{0,i} \mid D_i = 1) - Avg(y_{0,i} \mid D_i = 0)$$

$$= \text{Average causal effect} + \text{Selection bias}$$

Q$_{3.0}$: Is $Avg(y_i \mid D_i = 1) - Avg(y_i \mid D_i = 0)$ a *good* estimator for τ?

Difference in groups' means

$$= Avg(y_i \mid D_i = 1) - Avg(y_i \mid D_i = 0)$$

$$= Avg(y_{1,i} \mid D_i = 1) - Avg(y_{0,i} \mid D_i = 0)$$

$$= Avg(\tau + y_{0,i} \mid D_i = 1) - Avg(y_{0,i} \mid D_i = 0)$$

$$= \tau + Avg(y_{0,i} \mid D_i = 1) - Avg(y_{0,i} \mid D_i = 0)$$

$$= \text{Average causal effect} + \text{Selection bias}$$

So our proposed group-difference estimator give us the sum of

1. τ, the **causal, averate treatment effect** that we want
2. **Selection bias:** How much trt. and control groups differ (on average).

Next time: Solving selection bias.

Table of contents