ECON 0150 | Economic Data Analysis

The economist's data analysis skillset.

Part 4 | Review

Part 4: What we've learned

The bivariate GLM in four parts.

- ***4.1*** *Numerical predictors:* $y = \beta_0 + \beta_1 x + \varepsilon$
- ***4.2*** *Categorical predictors: same equation,* x *is 0 or 1*
- ***4.3*** *Model assumptions: linearity, homoskedasticity, independence, normality*
- ***4.4*** *The problem of timeseries: autocorrelation and differencing*

Today: practice problems in the style of the MiniExam.

Practice 1: Numerical predictor

A city wants to know if neighborhoods with more parks have lower crime rates.

	num_parks	crime_rate
0	6	39.7
1	14	26.9
2	11	25.6
3	9	31.1
4	2	40.5

a) How would you visualize the relationship between these two variables?

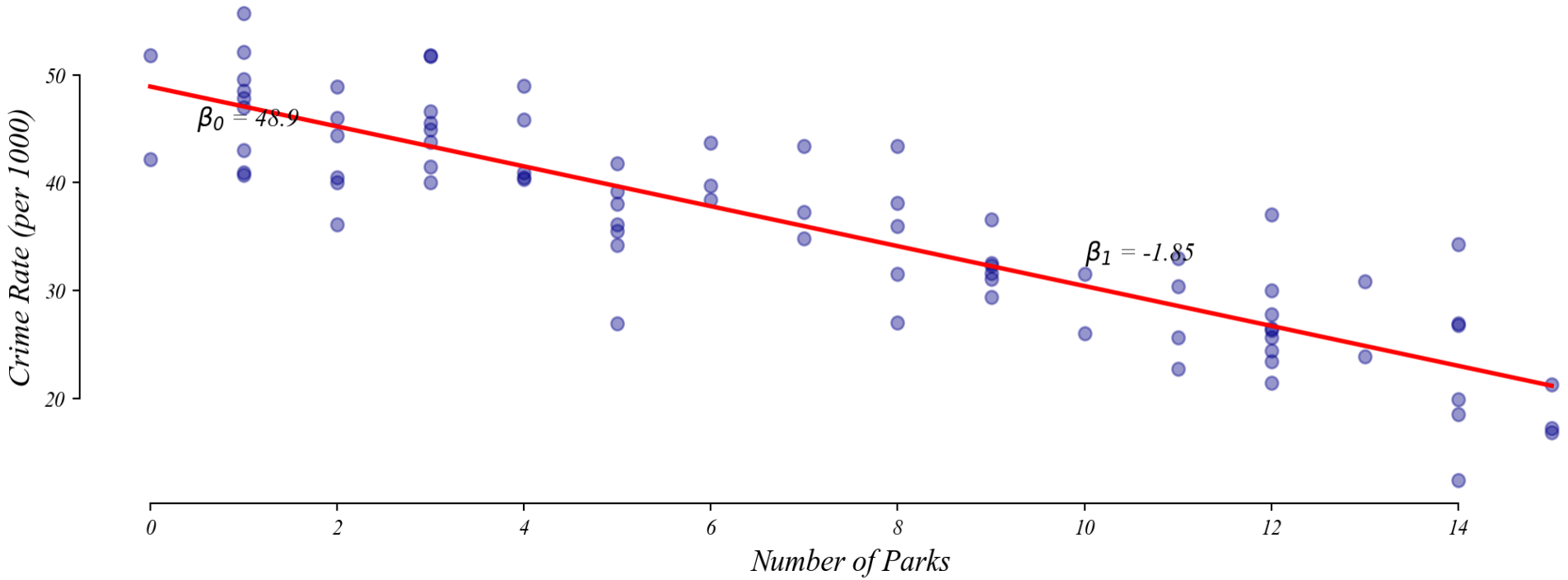
Scatterplot: num_parks on the x-axis, crime_rate on the y-axis.

b) Write down a model to test whether more parks means lower crime.

$$\text{crime\_rate} = \beta_0 + \beta_1 \cdot \text{num\_parks} + \varepsilon$$

Practice 1: Numerical predictor

A city wants to know if neighborhoods with more parks have lower crime rates.



c) What part of your model would indicate a relationship exists?

If β_1 is significantly different from zero (small p-value), there is a relationship.

Practice 2: Binary predictor

Same data. Now create a binary variable: `has_park` = 1 if `num_parks` > 0, else 0.

	num_parks	crime_rate	has_park
0	0	39.7	0
1	0	26.9	0
2	0	25.6	0
3	0	31.1	0
4	0	40.5	0

a) What is the variable type of `has_park`?

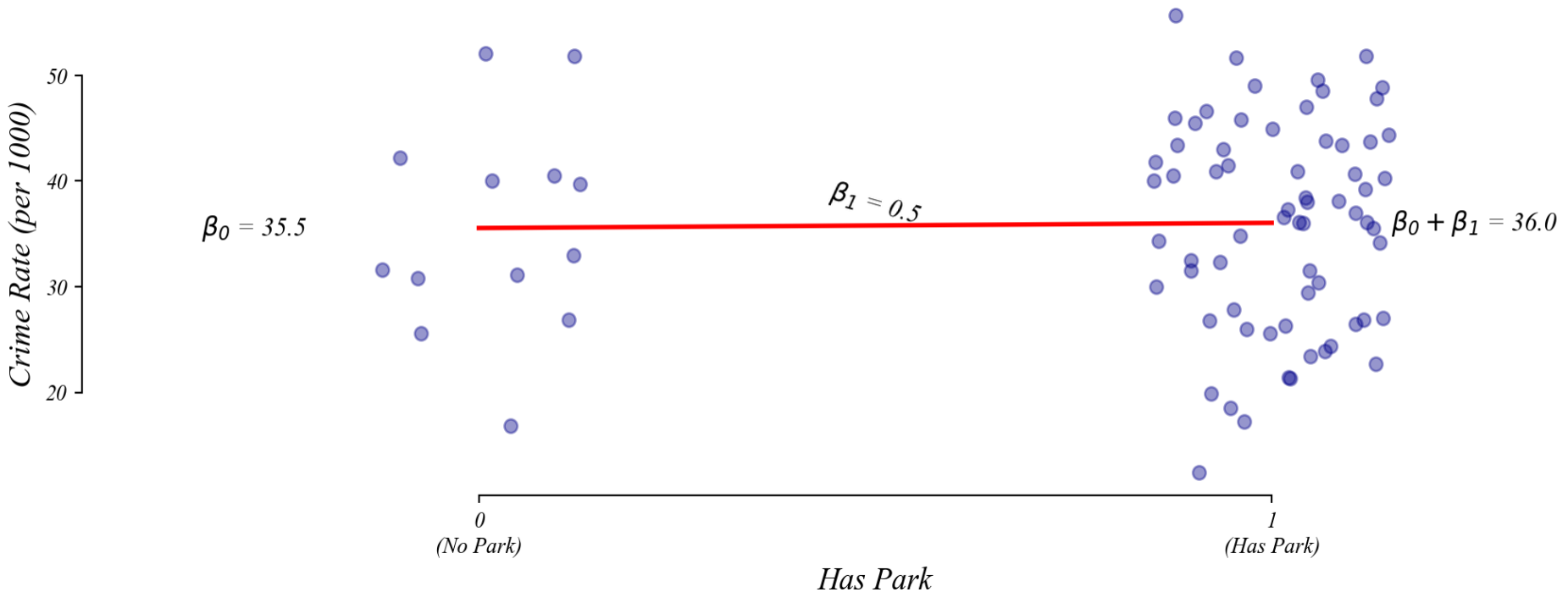
Binary categorical (0 or 1).

b) Visualize the relationship between `has_park` and `crime_rate`.

Strip plot or box plot with has_park on the x-axis and crime_rate on the y-axis.

Practice 2: Binary predictor

Same data. Now create a binary variable: `has_park` *= 1 if* `num_parks` *> 0, else 0.*



c) Write down the model. What does β_0 represent? What does β_1 represent?

$$\text{crime\_rate} = \beta_0 + \beta_1 \cdot \text{has\_park} + \varepsilon$$

β_0 *= Mean crime rate in neighborhoods without parks (x = 0).*

β_1 = *Difference in mean crime rate (has park minus no park).*

Practice 2: Binary predictor

What if we flip the coding?

d) If we instead coded `no_park` = 1 for no park, 0 for has park, what changes?

- β_0 *becomes the mean crime rate for neighborhoods* ***with*** *parks*
- β_1 *flips sign (same magnitude, opposite direction)*

β_0 *is ALWAYS the mean of the group coded as 0.*

β_1 *is ALWAYS the difference (group 1 minus group 0).*

Practice 3: Reading regression output

A researcher studies whether hours of sleep predict test scores using 150 students.

$$\text{test\_score} = \beta_0 + \beta_1 \cdot \text{hours\_sleep} + \varepsilon$$

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	42.30	5.100	8.294	0.000	32.22	52.38
hours_sleep	5.80	0.720	8.056	0.000	4.38	7.22

a) Interpret the intercept (42.30) in context.

A student who sleeps 0 hours is predicted to score 42.3 points. (Not meaningful.)

b) Interpret the slope (5.80) in context.

Each additional hour of sleep is associated with a 5.8 point increase in test score.

Practice 3: Reading regression output

A researcher studies whether hours of sleep predict test scores using 150 students.

$$\text{test\_score} = \beta_0 + \beta_1 \cdot \text{hours\_sleep} + \varepsilon$$

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	42.30	5.100	8.294	0.000	32.22	52.38
hours_sleep	5.80	0.720	8.056	0.000	4.38	7.22

c) What is the null hypothesis for the slope coefficient?

$H_0 : \beta_1 = 0$ *or hours of sleep has no effect on test scores.*

d) What test score would the model predict for a student who sleeps 8 hours?

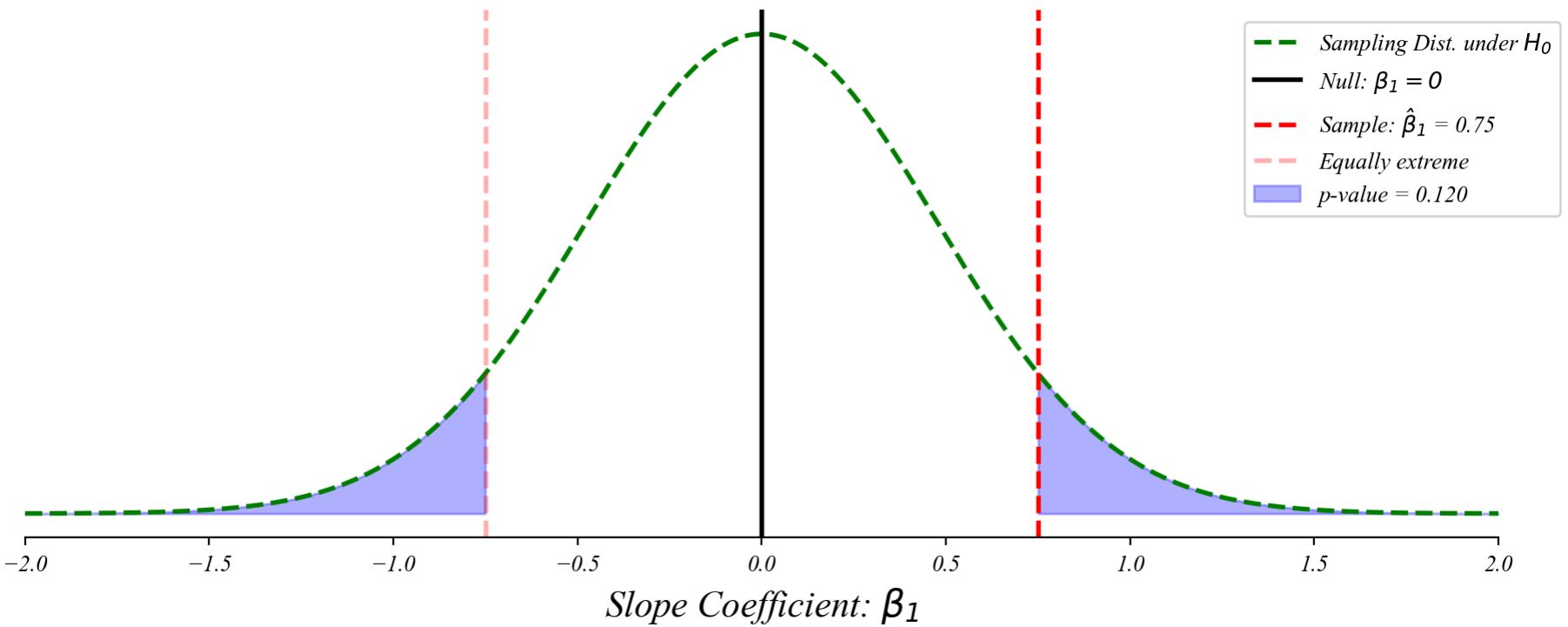
$\hat{y} = 42.3 + 5.8 \times 8 = 88.7$ *points.*

Practice 4: Drawing the sampling distribution

Now consider the experience and wages example from the MiniExam demo.

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	12.50	1.200	10.417	0.000	10.13	14.87
experience	0.75	0.478	1.570	0.120	−0.20	1.70

a) Draw the sampling distribution of β_1 under the null hypothesis.



Practice 4: Drawing the sampling distribution

Step by step: how to draw this on the exam.

*1. Draw a bell curve centered at **0** (the null hypothesis)*

2. Label the x-axis: "Slope Coefficient: β_1"

*3. Mark where **0** is (the null) with a solid line*

*4. Mark where your **observed slope** is with a dashed line*

*5. Shade the area **beyond** your observed slope (both tails) is the p-value*

Practice 5: Binary predictor

A researcher collects GPA data from 100 students.

	on_campus	gpa
0	0	3.09
1	0	2.99
2	0	2.54
3	1	2.91
4	1	2.61

a) If we code `on_campus` = 1 for yes, 0 for no, what does β_0 represent?

Mean GPA for off-campus students (the x = 0 group).

b) What does β_1 represent?

Difference in mean GPA: on-campus minus off-campus.

Practice 5: Binary predictor - the coding trick

What if we flip the coding?

c) Now code `off_campus` = 1 for off-campus, 0 for on-campus. What changes?

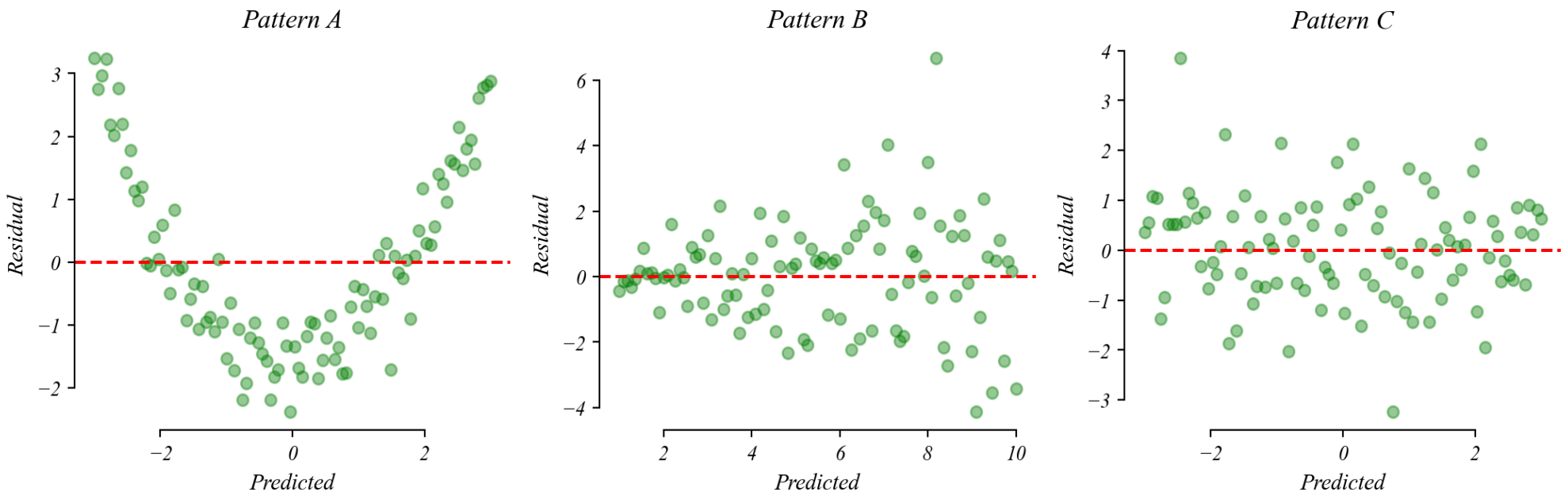
- β_0 *becomes the mean GPA for* ***on-campus*** *students*
- β_1 *flips sign (same magnitude, opposite direction)*

β_0 *is ALWAYS the mean of the group coded as 0.*

β_1 *is ALWAYS the difference (group 1 minus group 0).*

Practice 6: Reading a residual plot

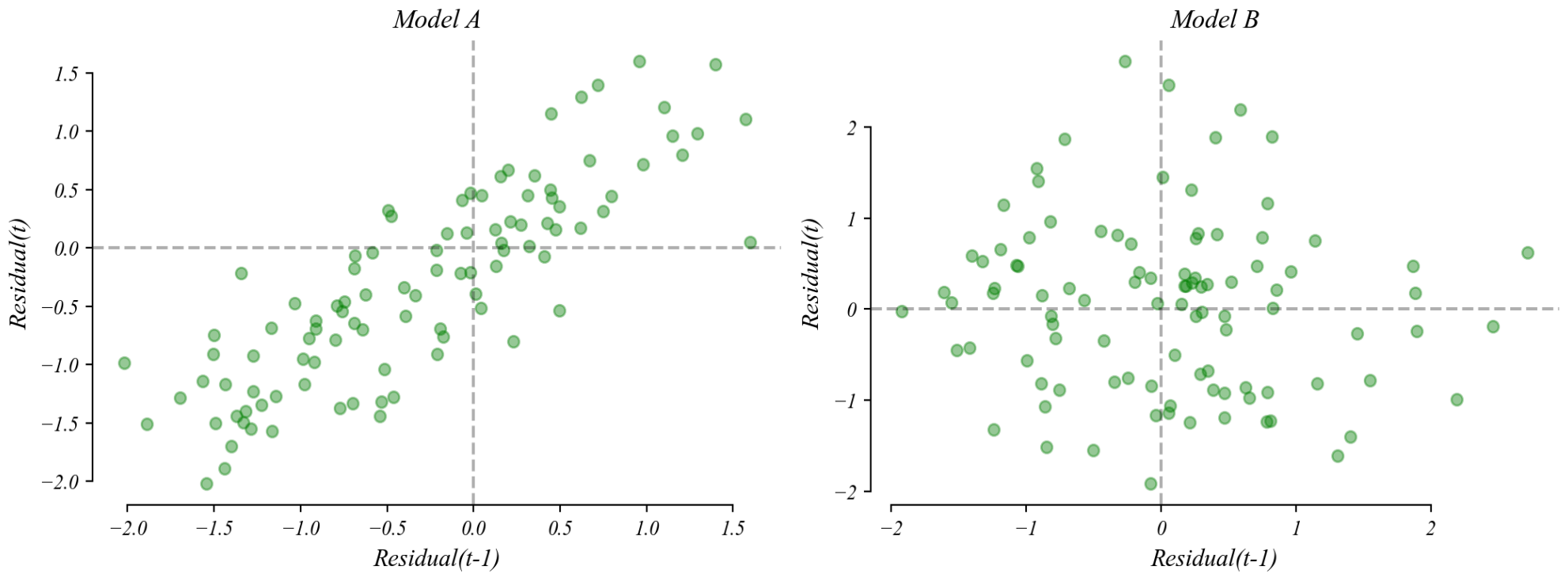
Which model assumption does each pattern suggest is violated?



- ***Pattern A**: U-shaped residuals → **linearity** violated (relationship is curved)*
- ***Pattern B**: Fan shape → **homoskedasticity** violated (variance increases)*
- ***Pattern C**: Random scatter → assumptions look fine*

Practice 7: Lagged residual plots

Which lag plot shows autocorrelation?



- ***Model A***: *Strong positive slope → residuals are* ***autocorrelated*** *(independence violated)*

- ***Model B***: *Random cloud → residuals are* ***independent*** *(assumption satisfied)*

Quick reference: what to know for the MiniExam

The core skills.

1. ***Write the model***: $y = \beta_0 + \beta_1 x + \varepsilon$ *(works for numerical AND binary* x*)*
2. ***Interpret*** β_0*: predicted* y *when* $x = 0$ *(baseline/reference group)*
3. ***Interpret*** β_1*: change in* y *for a one-unit increase in* x *(or difference between groups)*
4. ***Sampling distribution***: *bell curve centered on null, shade tails beyond observed slope*
5. ***P-value***: *probability of observing a slope as extreme as ours if* $\beta_1 = 0$
6. ***Residual plots***: *random = good; U-shape = non-linearity; fan = heteroskedasticity*
7. ***Lag plots***: *slope = autocorrelation (independence violated)*