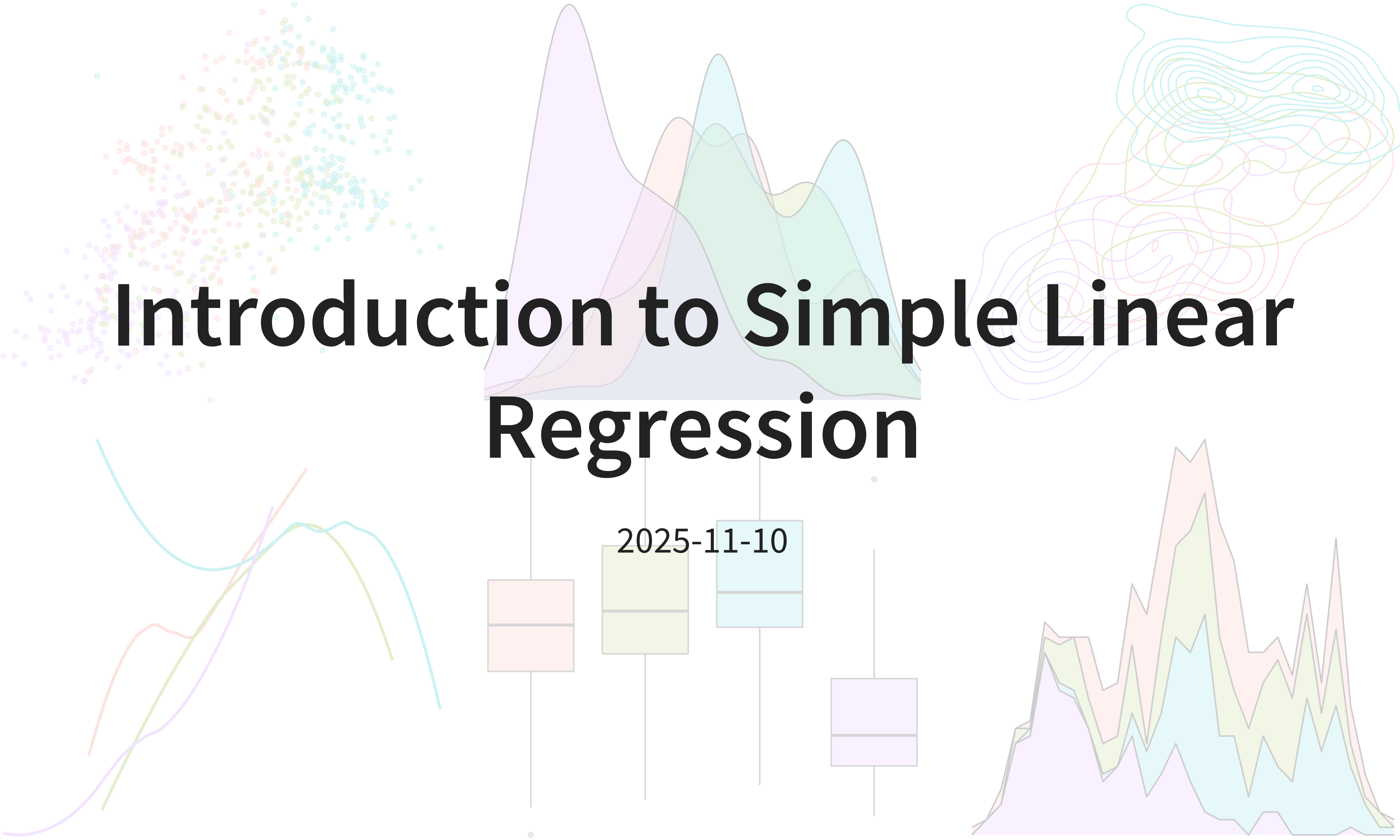
Introduction to Simple Linear Regression

2025-11-10



Housekeeping

- Homework 8 due tonight!
- Project proposal feedback (revisions due tonight midnight)

Linear regression

Crash course; take STAT 211 for more depth!

Fitting a line to data

- Recall equation of a line: $y = mx + b$
  - Intercept b and slope m determine specific line
  - This function is *deterministic*: as long as we know x, we know value of y exactly
- **Simple linear regression**: statistical method where the relationship between variables x and y is modeled as a **line + error:**

$$y = \underbrace{\beta_0 + \beta_1 x}_{\text{line}} + \underbrace{\epsilon}_{\text{error}}$$

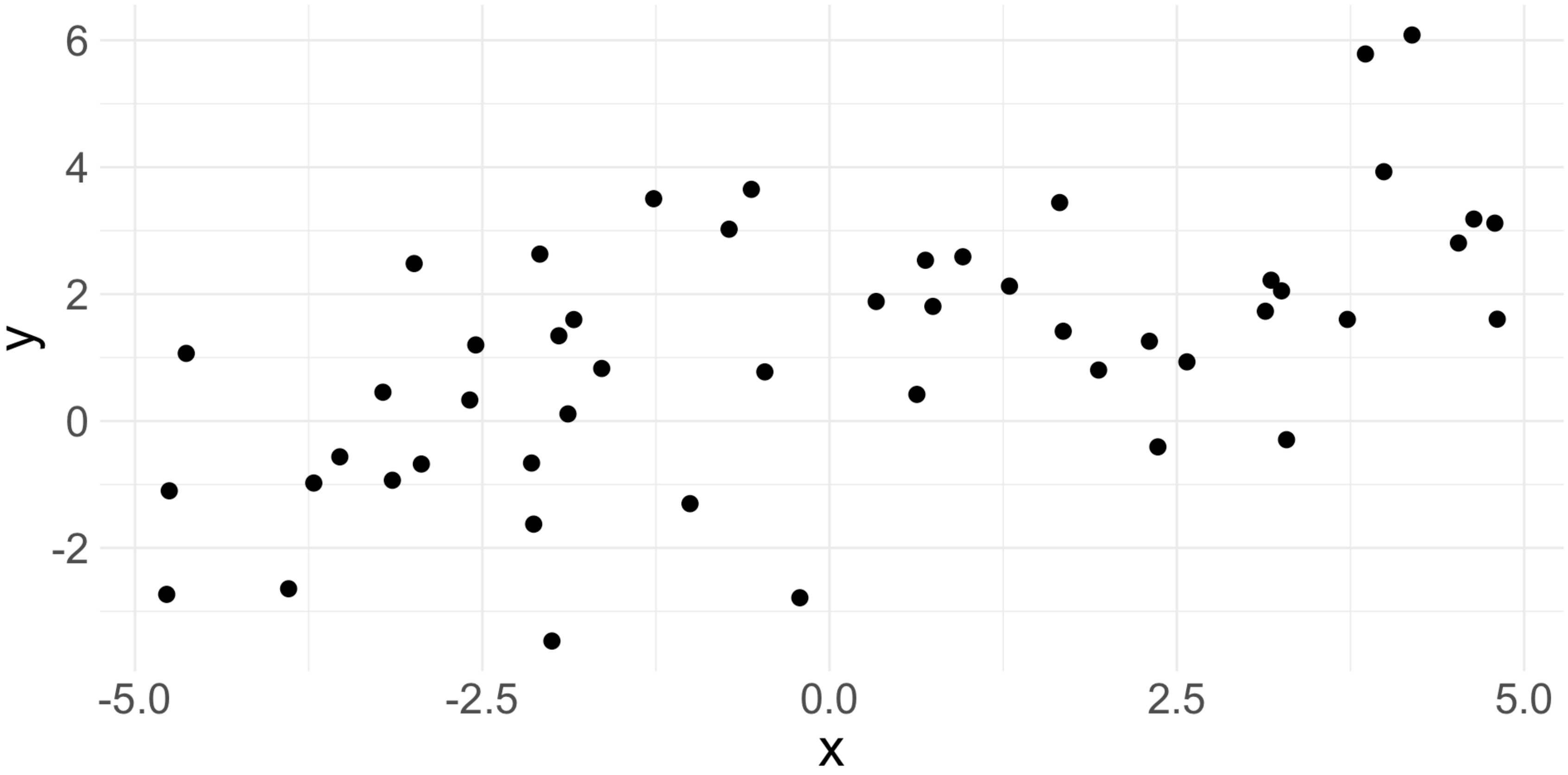
Simple linear regression model

$$y = \beta_0 + \beta_1 x + \epsilon$$

- We have two variables:
  1. y is response variable. **Must be (continuous) numerical.**
  2. x is explanatory variable, also called the **predictor** variable
     - Can be numerical or categorical
- β_0 and β_1 are the model **parameters** (intercept and slope)
  - Estimated using the data, with point estimates b_0 and b_1
- ϵ (epsilon) represents the **error**
  - Accounts for variability: we do not expect all data to fall perfectly on the line!
  - Sometimes we drop the ϵ term for convenience

Linear relationship

Suppose we have the following data:



- Observations won’t fall exactly on a line, but do fall around a straight line, so maybe a linear relationship makes sense!

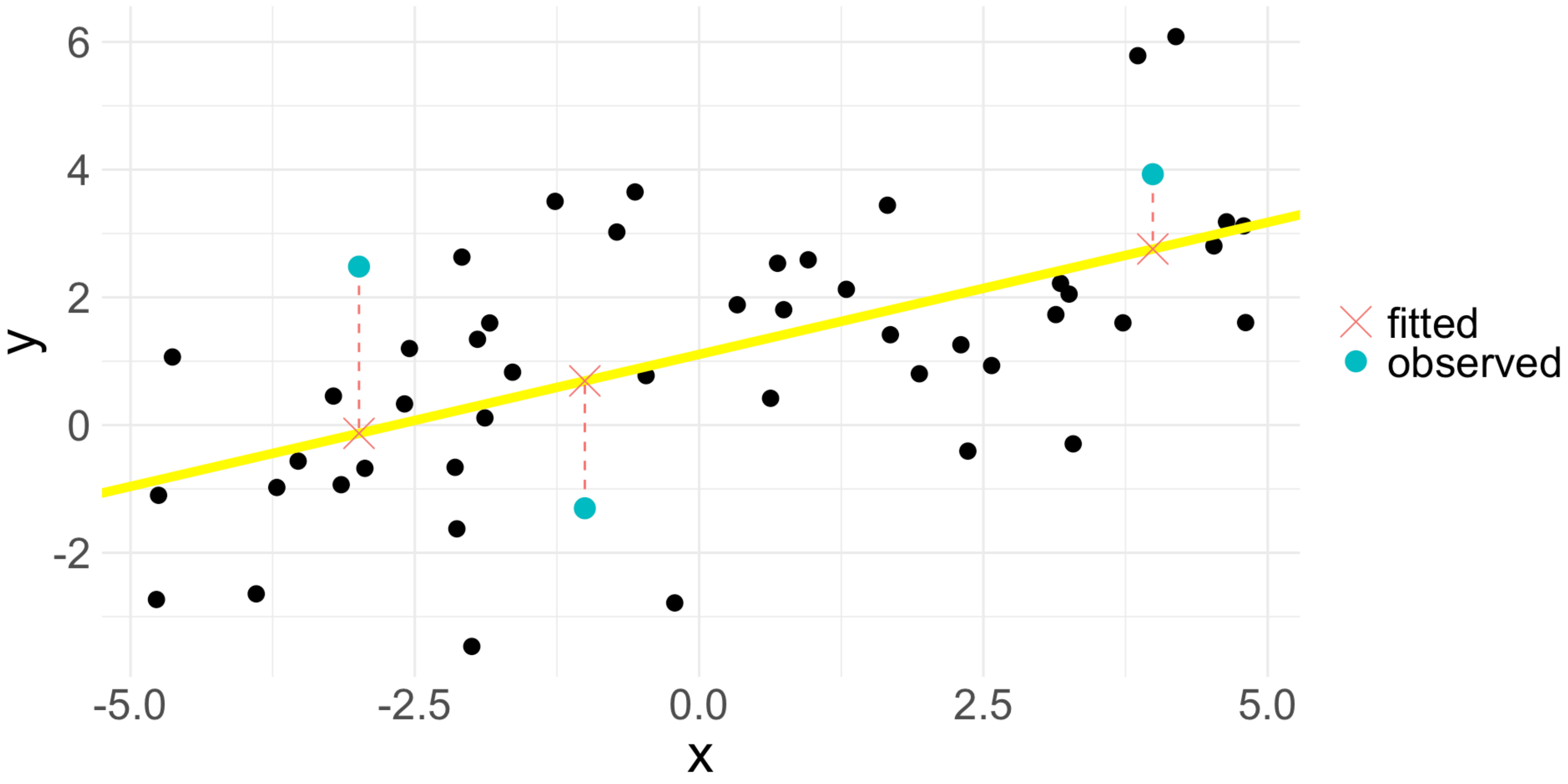
Fitted values

Suppose we have some specific estimates b_0 and b_1. We could **approximate** the linear relationship using these values as:

$$\hat{y} = b_0 + b_1 x$$

- The hat on y signifies an estimate: $\hat{y}$ is the **estimated/fitted** value of y given these specific values of x, b_0 and b_1
  - Can obtain a estimate $\hat{y}$ for every observed response y
- Note that the fitted value is obtained *without* the error

Fitted values (cont.)



- Suppose our estimated line is the yellow one: $\hat{y} = 1.11 + 0.41x$
- The fitted value $\hat{y_i}$ for y_i **lies on the line**; the above plot shows three specific examples

Residual

Residuals (denoted as e) are the remaining variation in the data after fitting a model.

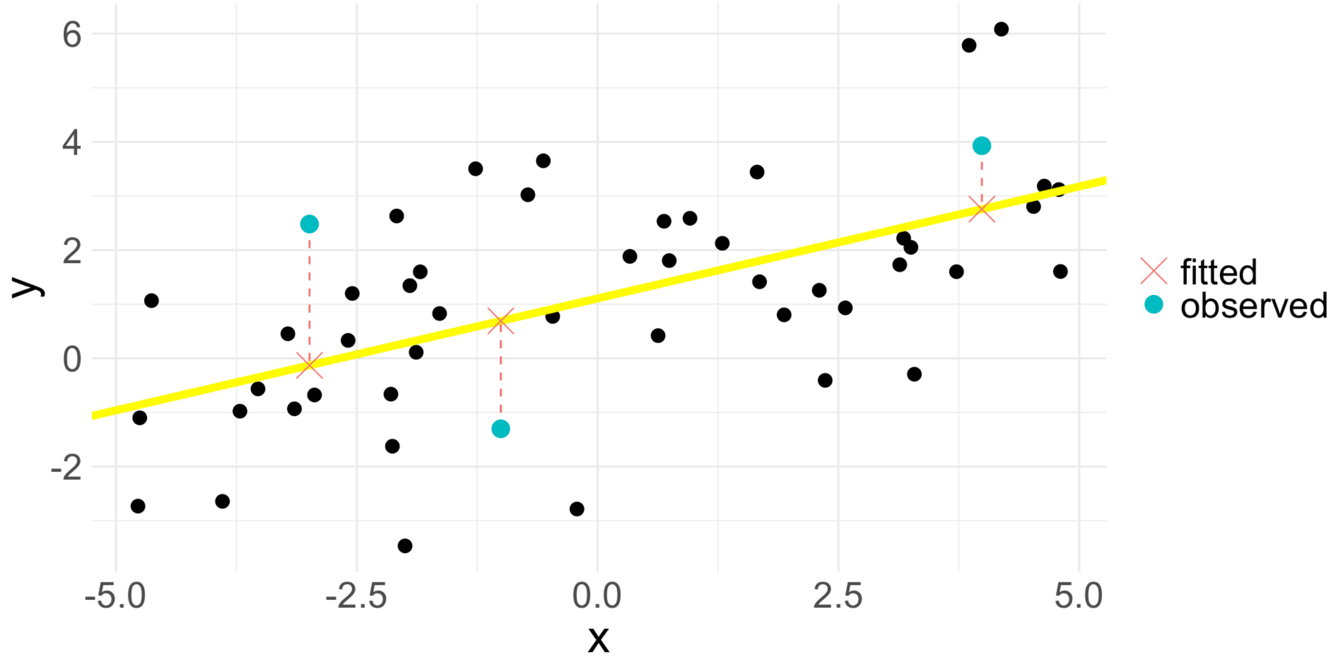
$$\text{observed response} = \text{fit} + \text{residual}$$

- For **each** observation i, we obtain the residual e_i via:

$$y_i = \hat{y_i} + e_i \Rightarrow e_i = y_i - \hat{y_i}$$

- Residual = difference between observed and expected
- In the plot, the residual is indicated by the vertical dashed line
  - What is the ideal value for a residual? What does a positive/negative residual indicate?

Residual (cont.)

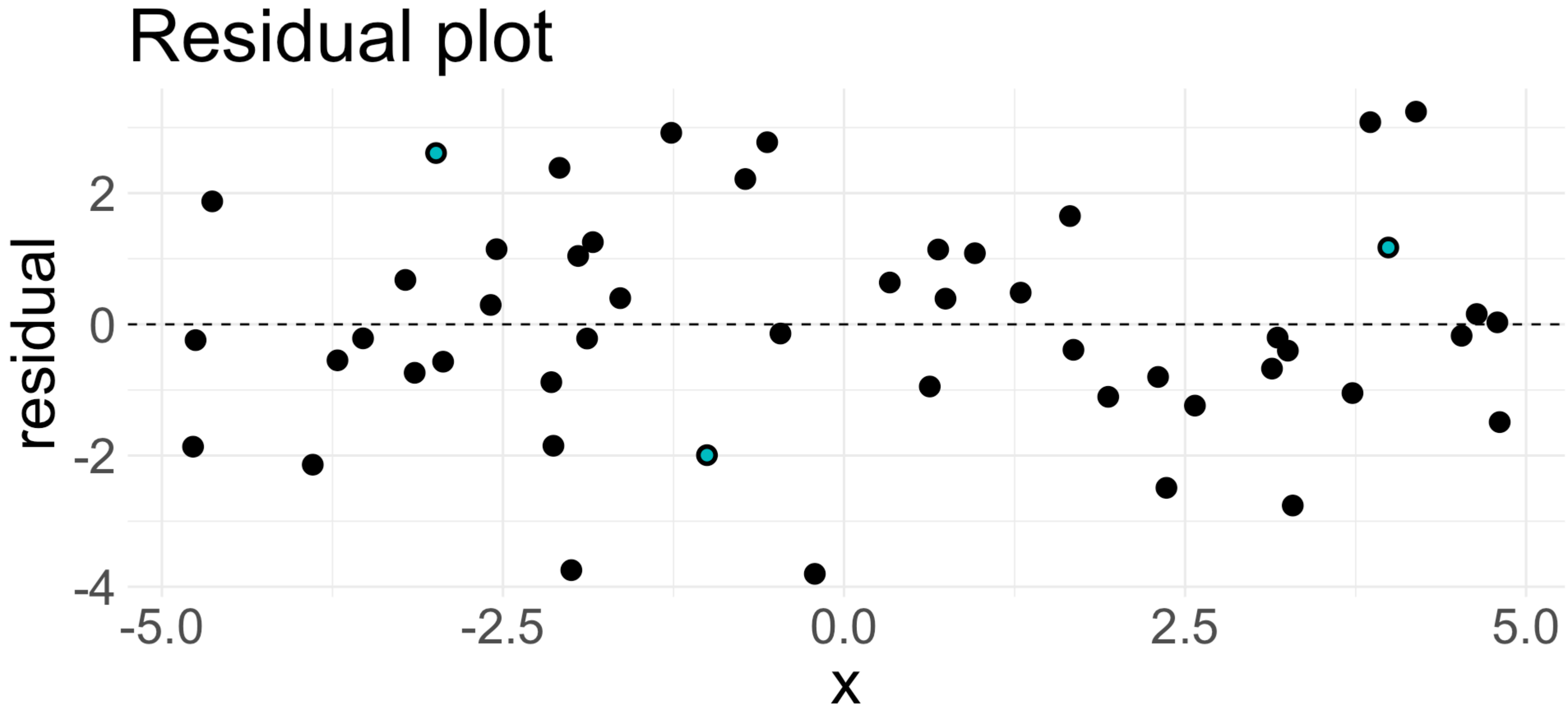


Residual values for the three highlighted observations:

x	y	y_hat	residual
-2.991	2.481	-0.130	2.611
-1.005	-1.302	0.691	-1.994
3.990	3.929	2.757	1.172

Residual plot

- Residuals are very helpful in evaluating how well a model fits a set of data
- **Residual plot**: original x values plotted against corresponding residuals on y-axis

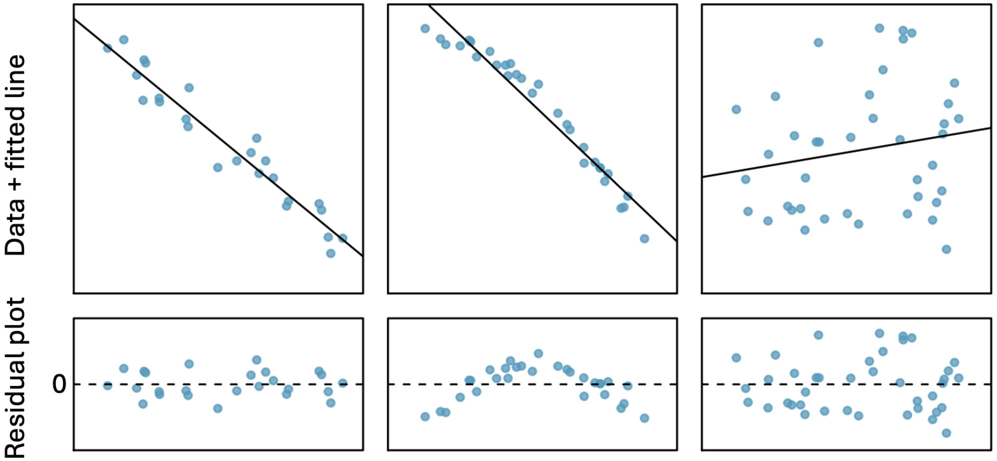


Blue dots = residuals for specific points from previous plot

Residual plot (cont.)

Residual plots can be useful for identifying characteristics/patterns that remain in the data even after fitting a model.

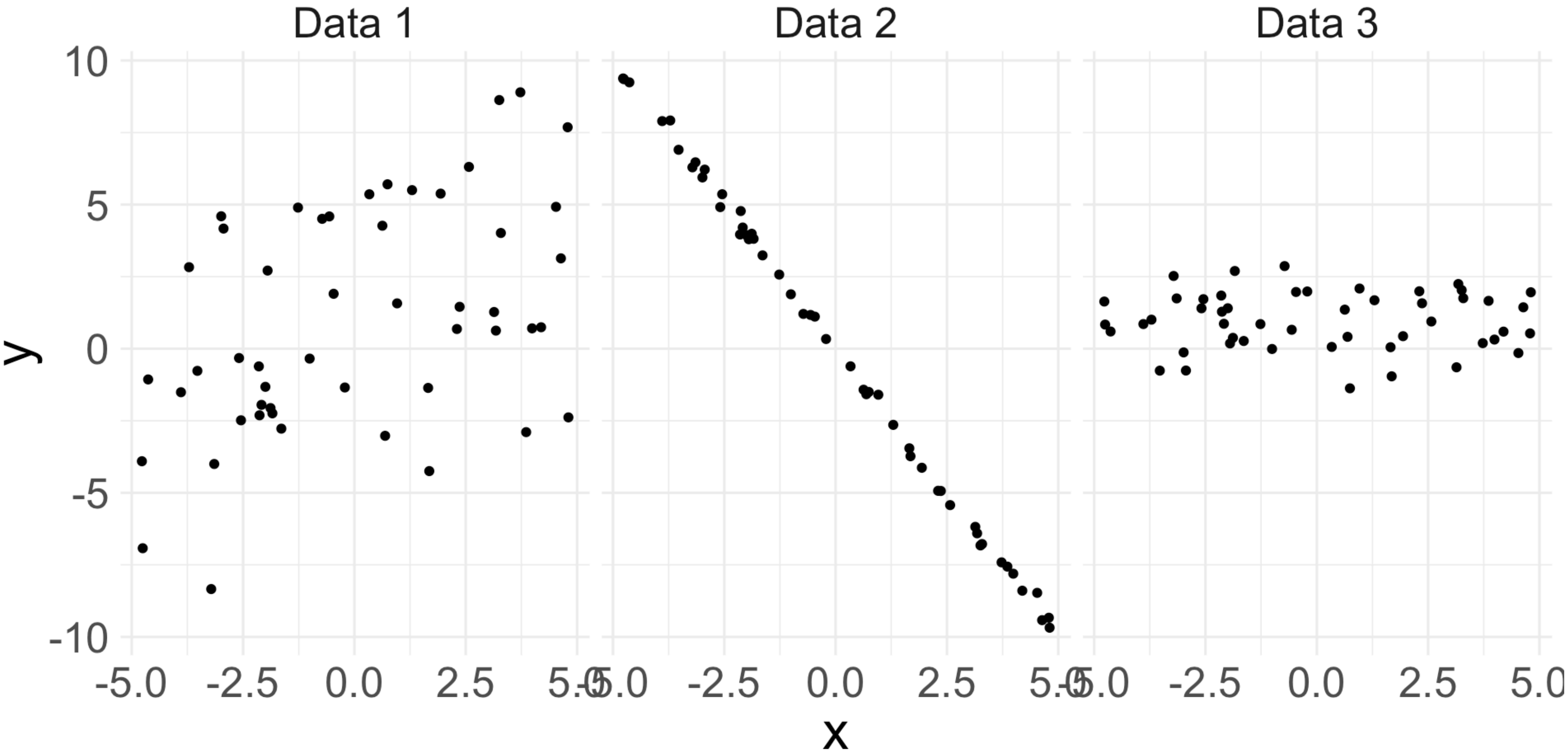
- Just because you fit a model to data, does not mean the model is a good fit!



Can you identify any patterns remaining in the residuals?

Describing linear relationships

Different data may exhibit different strength of linear relationships:

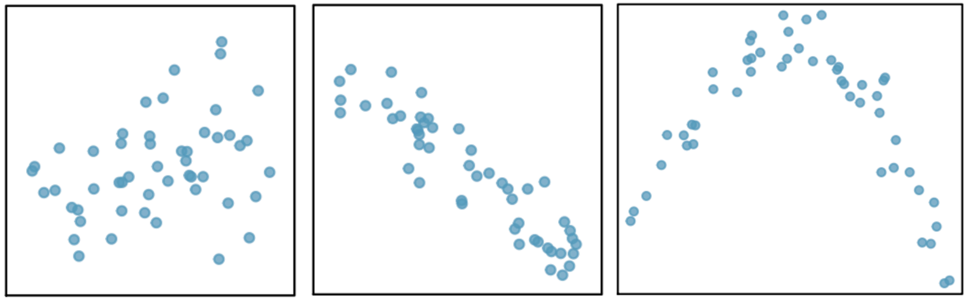


- Can we quantify the strength of the linear relationship?

Correlation

- **Correlation** is describes the strength of a *linear* relationship between two variables
    - The observed sample correlation is denoted by R
    - Formula (not important): $R = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$

- Always takes a value between -1 and 1
    - -1 = perfectly linear and negative
    - 1 = perfectly linear and positive
    - 0 = no linear relationship
- Nonlinear trends, even when strong, sometimes produce correlations that do not reflect the strength of the relationship



R = 0.33 R = –0.92 R = –0.23

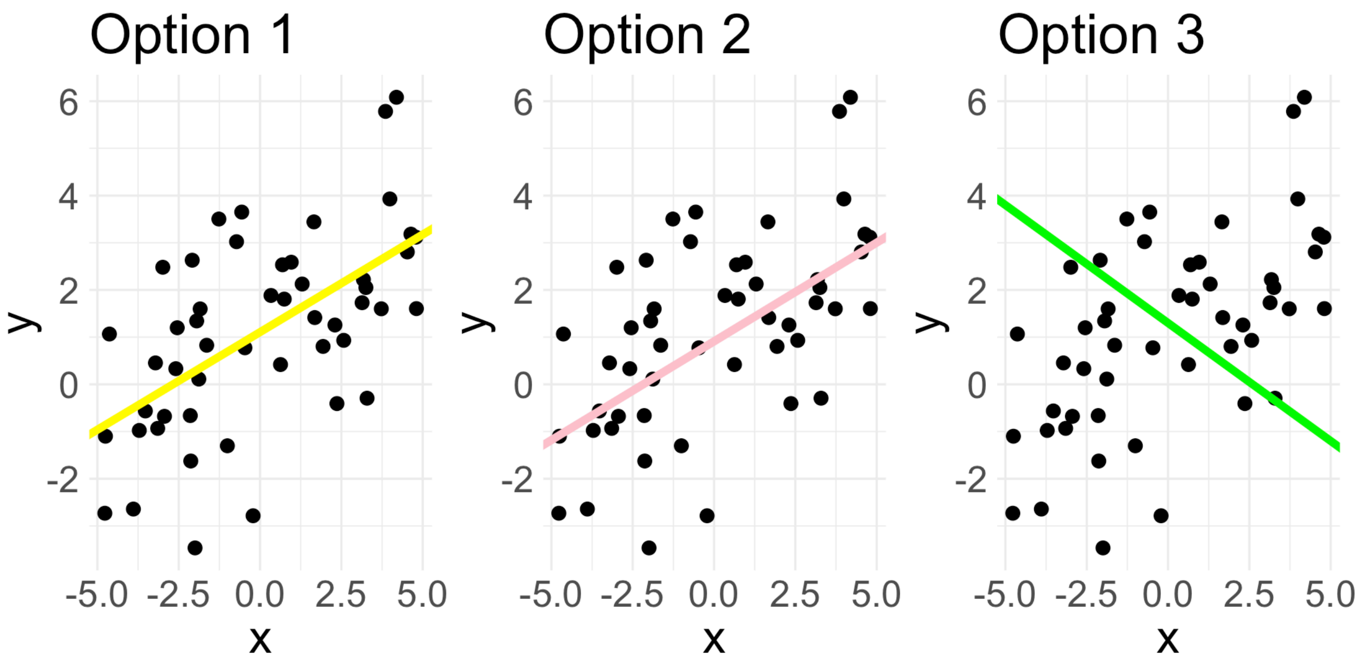
Least squares regression

In Algebra class, there exists a single (intercept, slope) pair because the (x, y) points had no error; all points landed on the line.

- Now, we assume there is error
- How do we choose a single “best” (b_0, b_1) pair?

Different lines

The following display the same set of 50 observations.



Which line would you say fits the data the best?

- There are infinitely many choices of (b_0, b_1) that could be used to create a line
- We want the BEST choice (i.e. the one that gives us the “line of best fit”)
  - How to define “best”?

Line of best fit

One way to define a “best” is to choose the specific values of (b_0, b_1) that minimize the total residuals across all n data points. Results in following possible criterion:

1. **Least absolute criterion**: minimize sum of residual magnitudes:

$$|e_1| + |e_2| + \ldots + |e_n|$$

2. **Least squares criterion**: minimize sum of squared residuals:

$$e_1^2 + e_2^2 + \ldots + e_n^2$$

- The choice of (b_0, b_1) that satisfy least squares criterion yields the **least squares line**, and will be our criterion for “best”
- On previous slide, yellow line is the least squares line, whereas pink line is the least absolute line

Linear regression model

Remember, our linear regression model is:

$$y = \beta_0 + \beta_1 x + \epsilon$$

While not wrong, it can be good practice to be specific about an observation i:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \ldots, n$$

- Here, we are stating that each observation i has a specific:
  - explanatory variable value x_i
  - response variable value y_i
  - error/randomness ϵ_i
- In SLR, we further assume that the errors ϵ_i are independent and Normally distributed

Conditions for the least squares line (LINE)

Like when using CLT, we should check some conditions before saying a linear regression model is appropriate!

Assume for now that x is continuous numerical.

1. **Linearity**: data should show a linear trend between x and y
2. **Independence**: the observations i are independent of each other
    - e.g. random sample
    - Non-example: time-series data
3. **Normality/nearly normal residuals**: the residuals should appear approximately Normal
    - Possible violations: outliers, influential points (more on this later)
4. **Equal variability**: variability of points around the least squares line remains roughly constant

Running example

We will see how to check for these four LINE conditions using the `cherry` data from `openintro`.

diam	volume
8.3	10.3
8.6	10.3
8.8	10.2
10.5	16.4
10.7	18.8

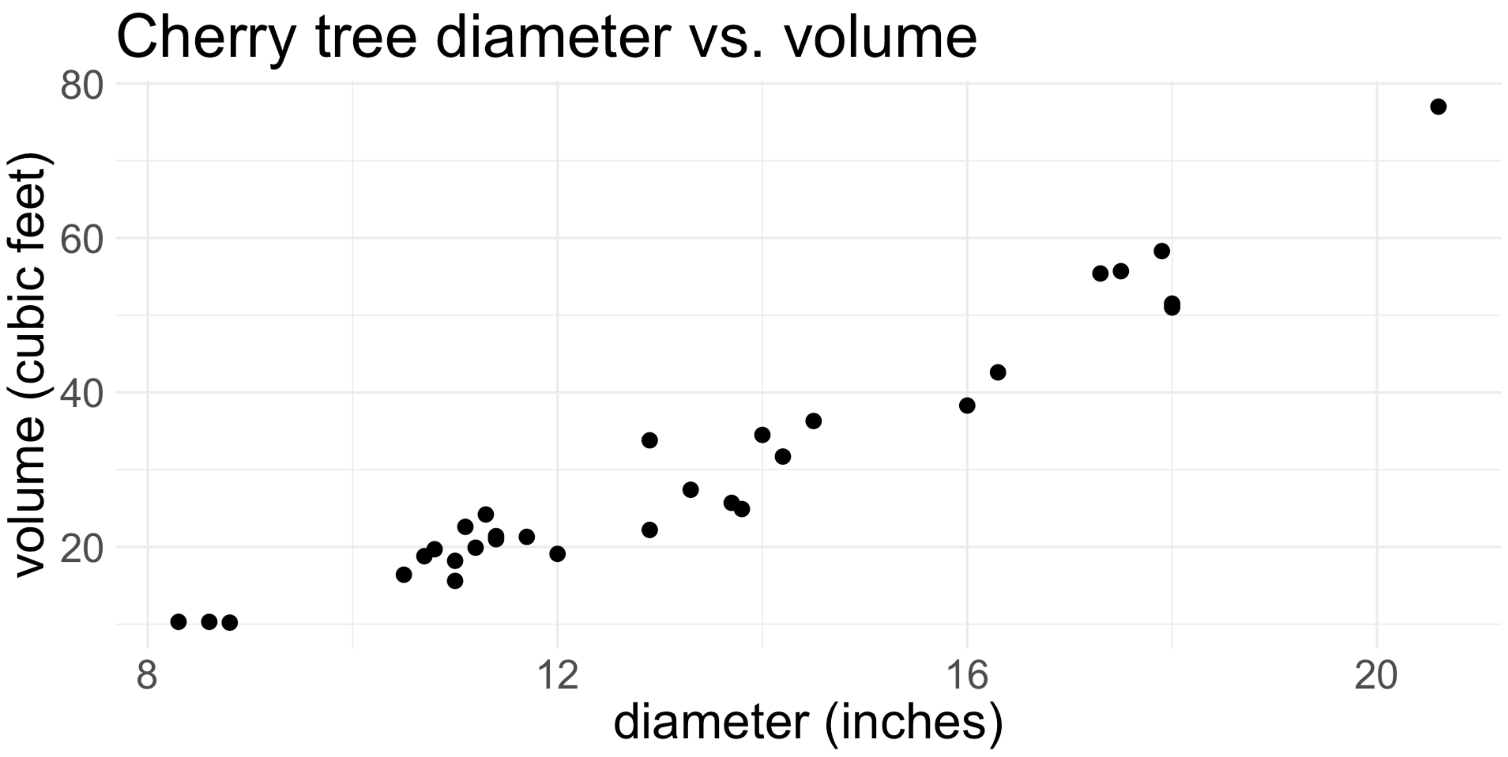
- Explanatory variable x: `diam`
- Response variable y: `volume`

Our candidate linear regression model is as follows

$$\text{volume} = \beta_0 + \beta_1 \text{diameter} + \epsilon$$

1. Linearity

Assess *before* fitting the linear regression model by making a scatterplot of x vs. y:



Does there appear to be a linear relationship between diameter and volume?

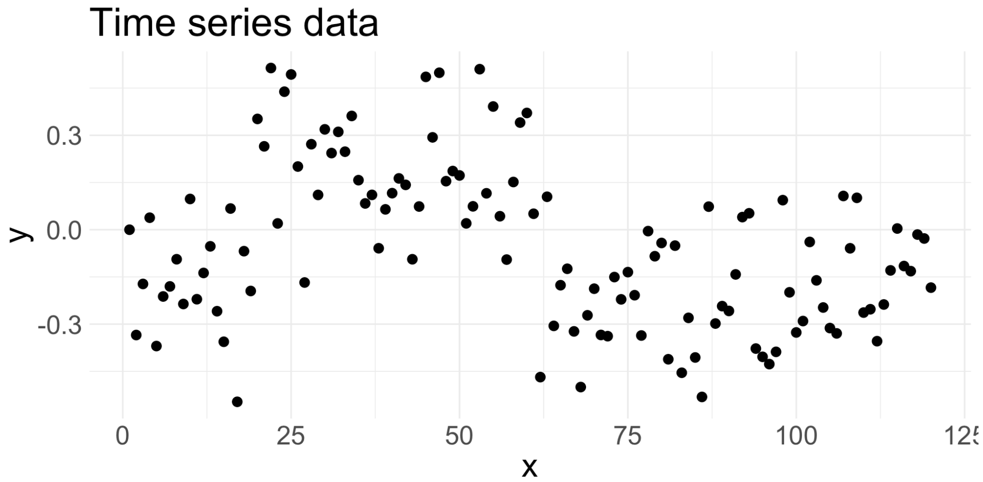
- I would say yes

2. Independence

Assess *before* fitting the linear regression model by understanding how your data were sampled.

- The `cherry` data do not explicitly say that the trees were randomly sampled, but it might be a reasonable assumption

An example where independence is violated:



Here, the data are a time series, where observation at time point i depends on the observation at time $i - 1$.

- Successive/consecutive observations are highly correlated

Fitting the model

Because the first two conditions are met, we can go ahead and fit the linear regression model (i.e. estimate the values of the coefficients)

- After fitting the model, we get the following estimates: $b_0 = -36.94$ and $b_1 = 5.07$. So our **fitted model** is:

$$\widehat{\text{volume}} = -36.94 + 5.07 \times \text{diameter}$$

Remember: the “hat” denotes an estimated/fitted value!

- We will soon see how b_0 and b_1 are calculated and how to interpret them
- The next two checks can only occur *after* fitting the model.

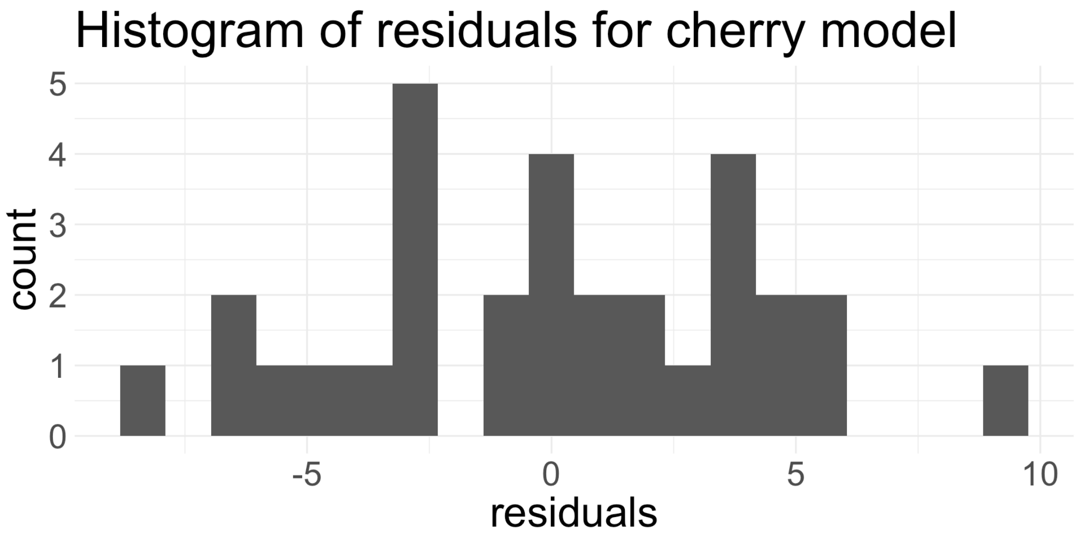
3. Nearly normal residuals

Assess *after* fitting the model by making histogram of residuals and checking for approximate Normality.

- Remember, residuals are $e_i = y_i - \hat{y_i}$

```r
cherry |>
  mutate(volume_hat = -36.94 + 5.07*diam)
  mutate(residual = volume - volume_hat)
```

diam	volume	volume_hat	residual
8.3	10.3	5.141	-5.159
8.6	10.3	6.662	-3.638
8.8	10.2	7.676	-2.524
10.5	16.4	16.295	-0.105
10.7	18.8	17.309	-1.491



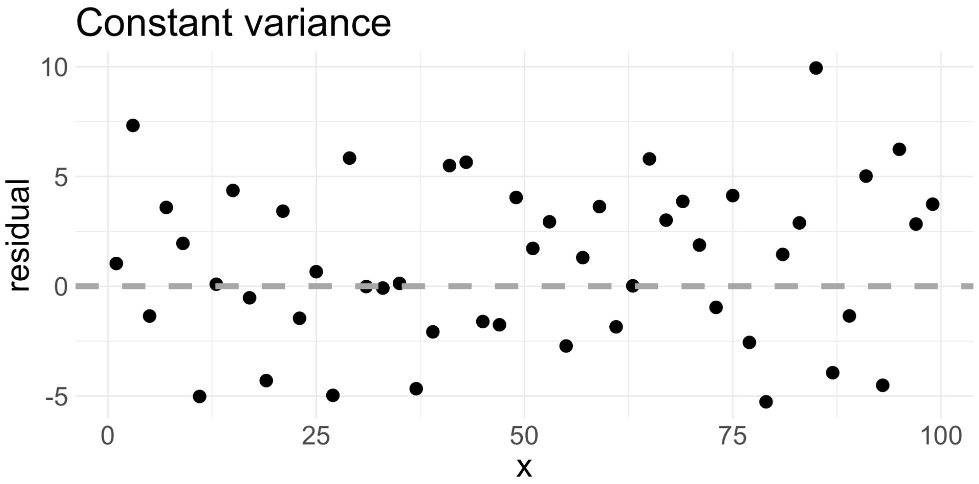
Do the residuals appear approximately Normal?

- I think so!

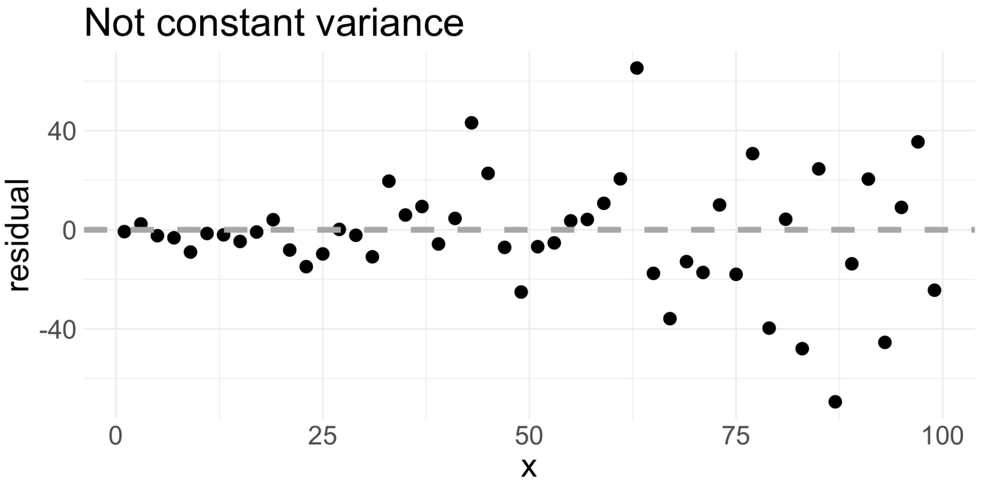
4. Equal variance

Assess *after* fitting the model by examining a residual plot and looking for patterns.

A good residual plot:



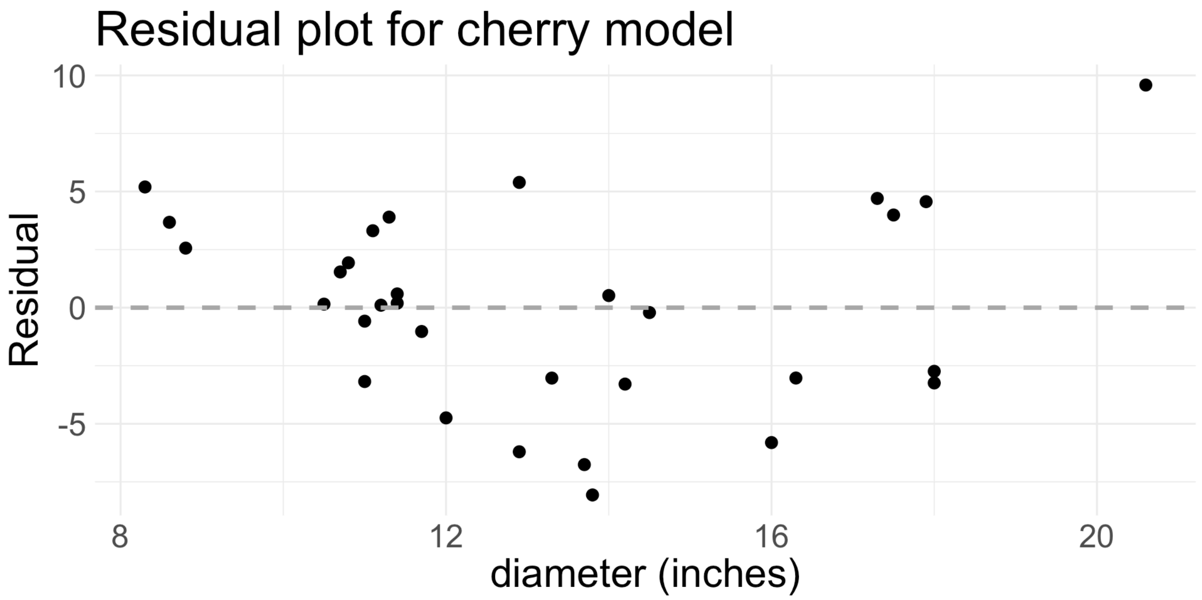
A bad residual plot:



We usually add a horizontal line at 0.

4. Equal variance (cont.)

Let’s examine the residual plot of our fitted model for the `cherry` data:



- Do we think equal variance is met?
  - I would say there is a definite pattern in the residuals, so equal variance condition is not met.
  - Some of the variability in the errors appear related to `diameter`