Introduction to Simple Linear Regression

11/6/24

Housekeeping

- Homework 7 due tonight!
- Last problem set is assigned today! Atypical due date: Wednesday 11/13

Linear regression

Crash course; take STAT 211 for more depth!

Fitting a line to data

- Hopefully we are all familiar with the 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
- **Linear regression**: statistical method where the relationship between variable x and variable y is modeled as a **line + error:**

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

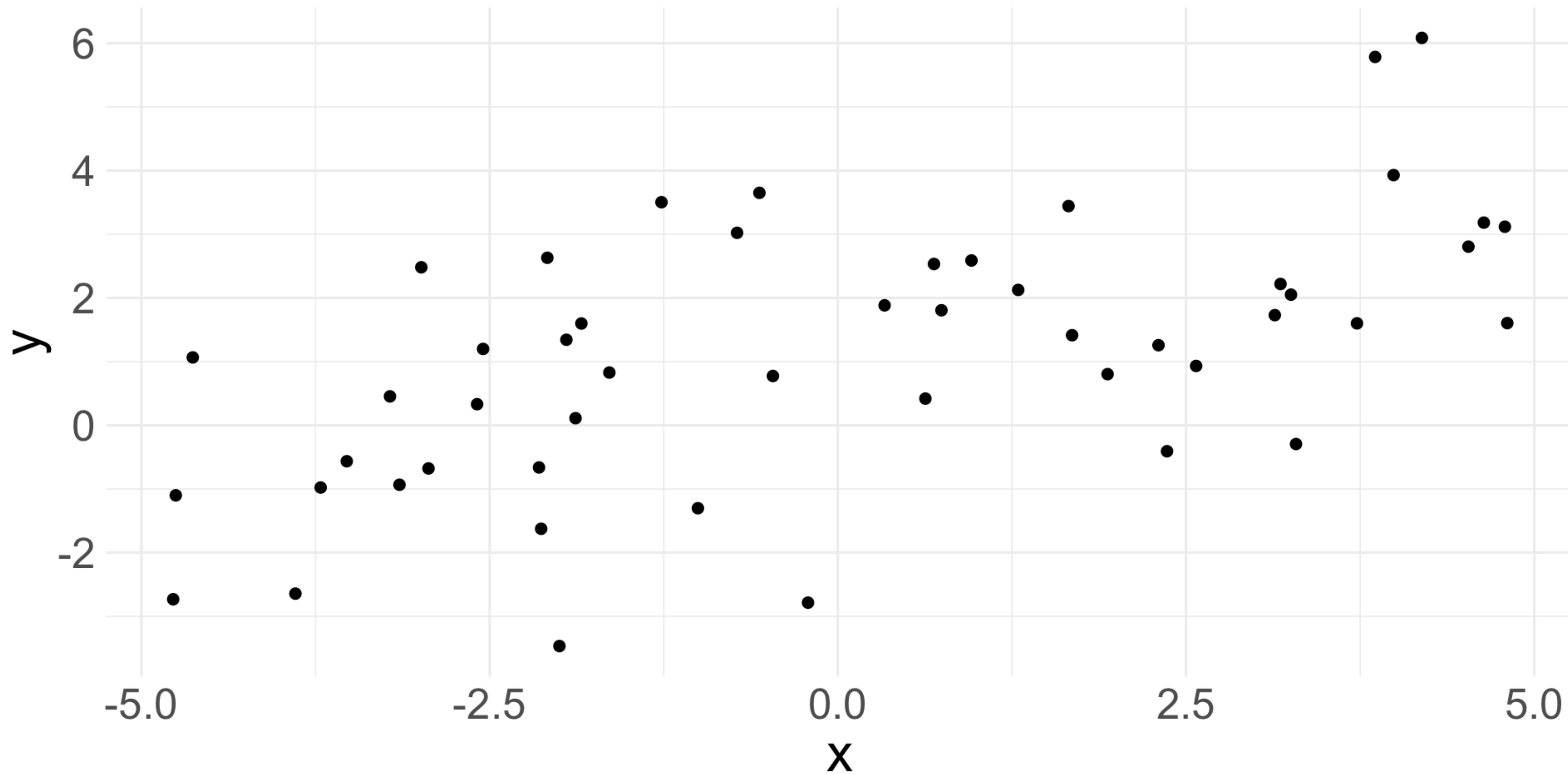
Linear regression model

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

- 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!

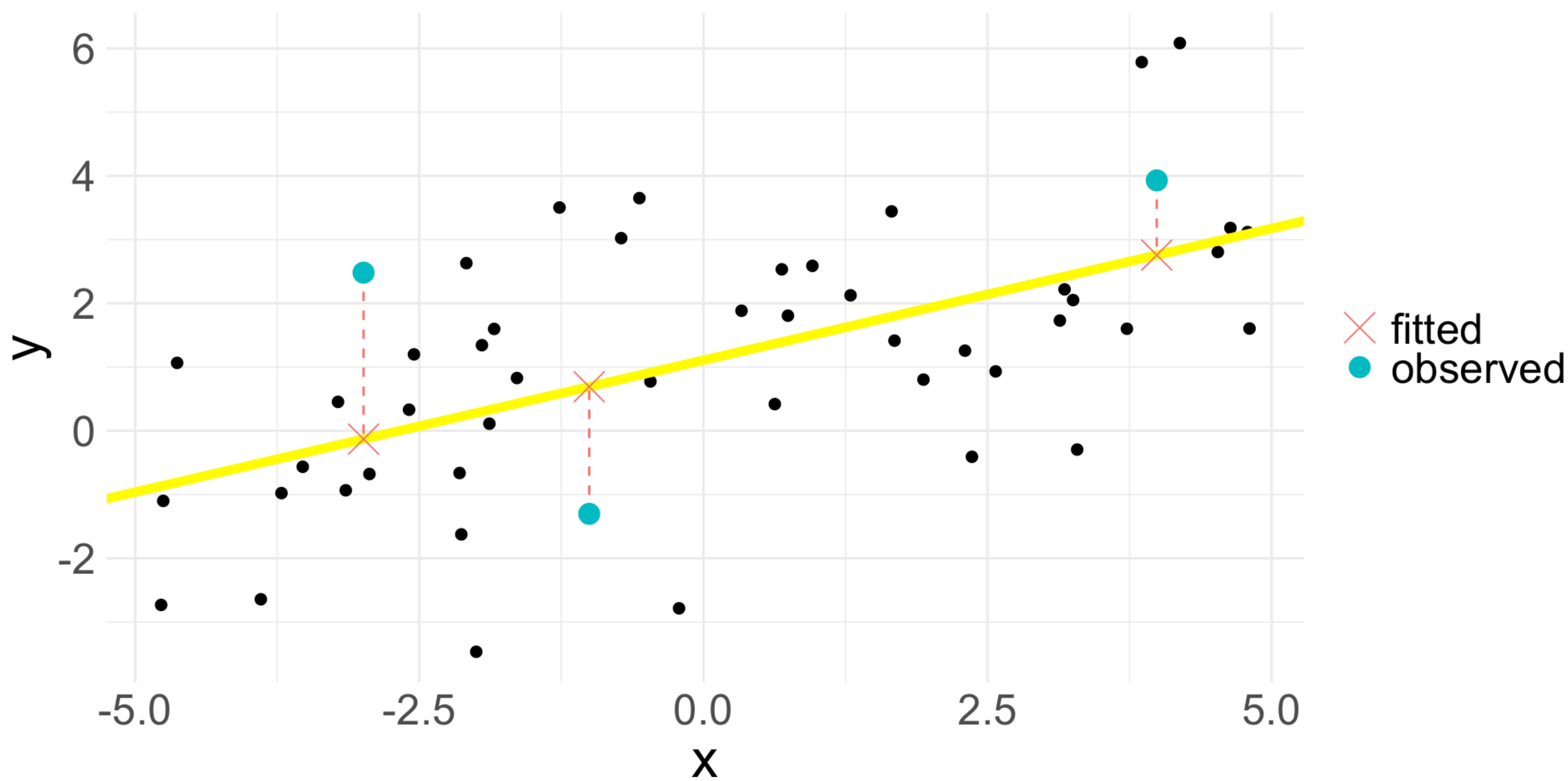
Fitted values

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

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

- The hat on y signifies that this is an estimate: the estimated/**fitted** value of y given these specific values of x, b_0 and b_1
  - We observe y, but can obtain a corresponding estimate $\hat{y}$
- Note that the fitted value is obtained *without* the error

Fitted values (cont.)



- Suppose our estimated line is the yellow one
- Every observed value y_i has a corresponding fitted value $\hat{y}_i$; the above plot just shows three specific examples

Residual

Residuals are the remaining variation in the data after fitting a model.

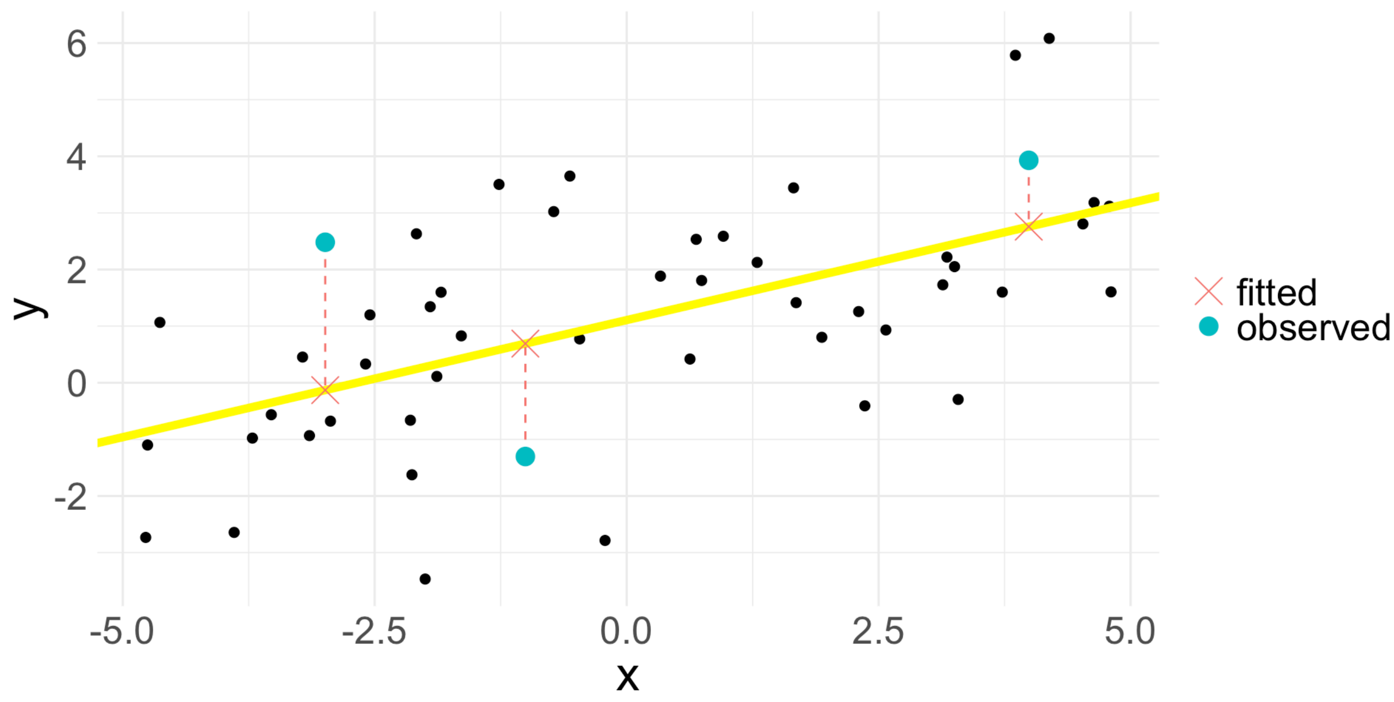
$$\text{data} = \text{fit} + \text{residual}$$

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

$$y_i = \hat{y}_i + e_i \quad \Rightarrow \quad e_i = \hat{y}_i - y_i$$

- Residual = difference between observed and expected
- Since each observation has a fitted value, each observation has a residual
  - In the linear regression case, the residual is indicated by the vertical dashed line
  - What is the ideal value for a residual?

Residual (cont.)

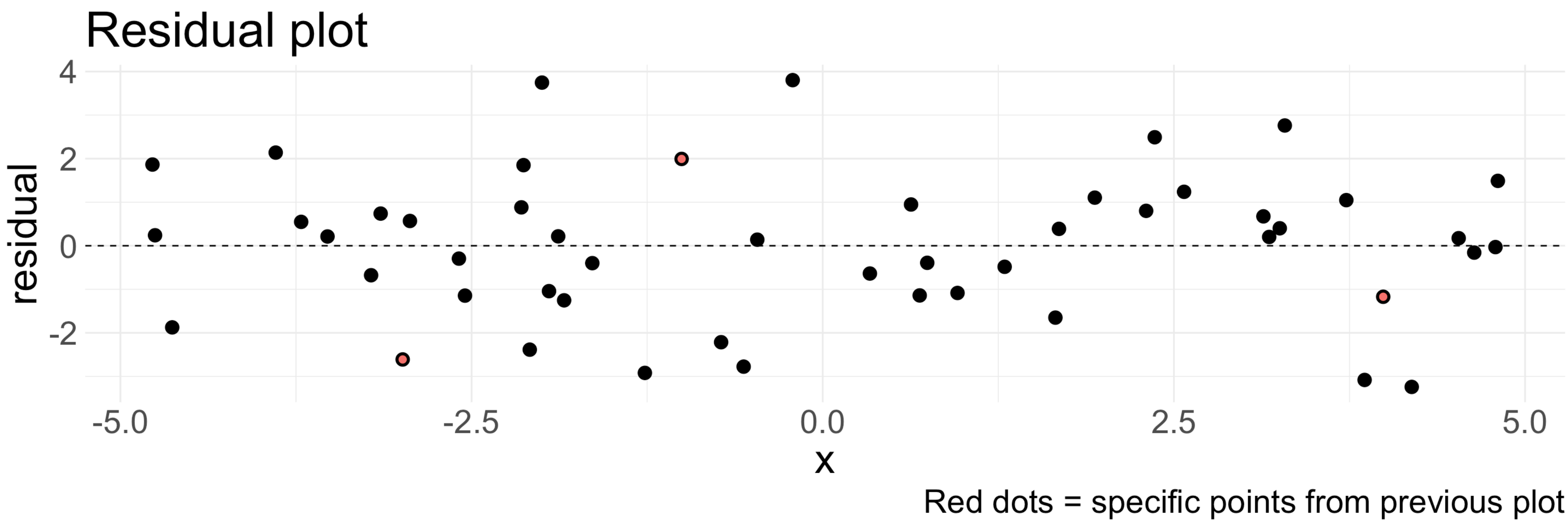


Residual values for the three highlighted observations:

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

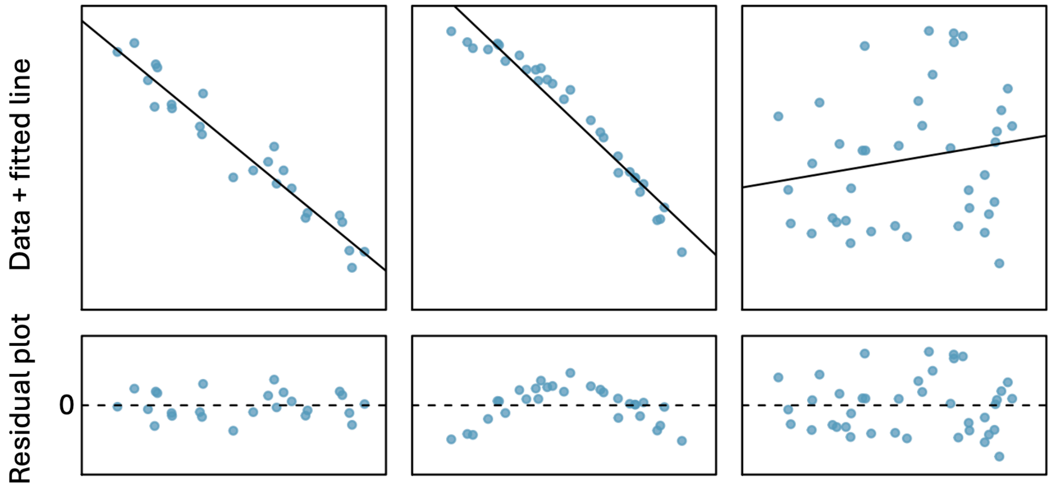
- Residuals are very helpful in evaluating how well a model fits a set of data
- **Residual plot**: original x values plotted against their corresponding residuals on y -axis



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!

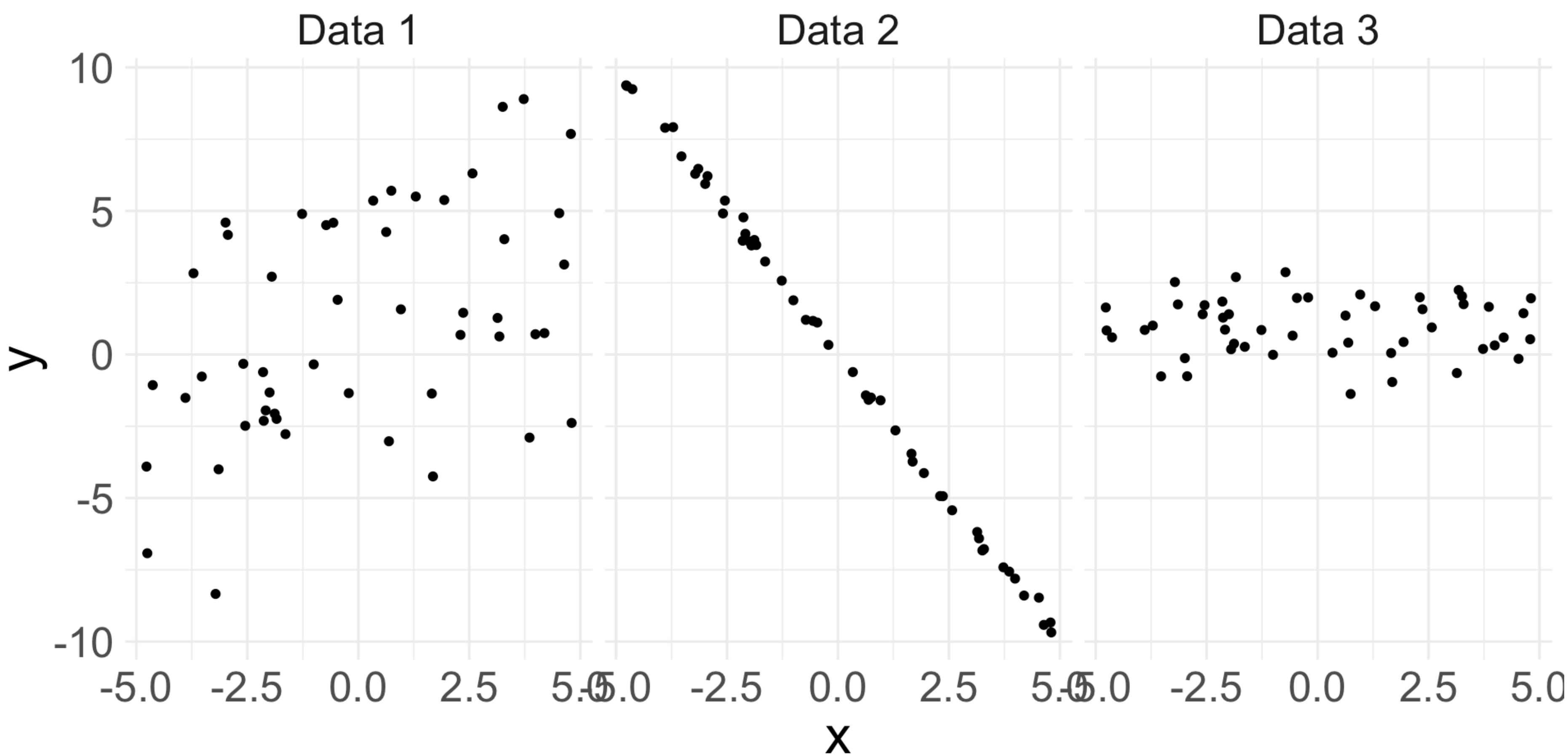


Can you identify any patterns remaining in the residuals?

- Sorry! The residuals shown here are taken as $y_i - \hat{y}_i$!

Describing linear relationships

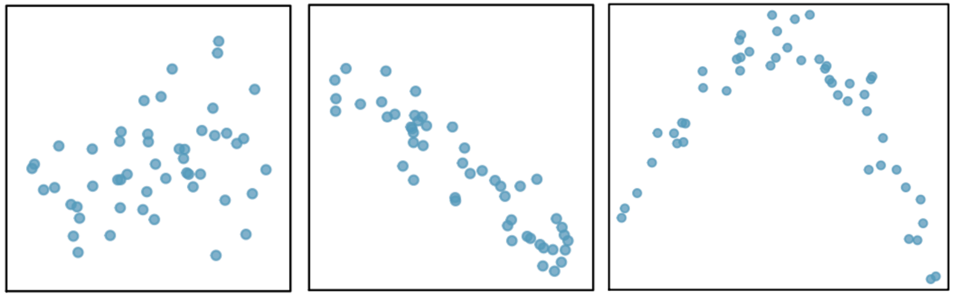
Different data may exhibit different strength of linear relationships:



- 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

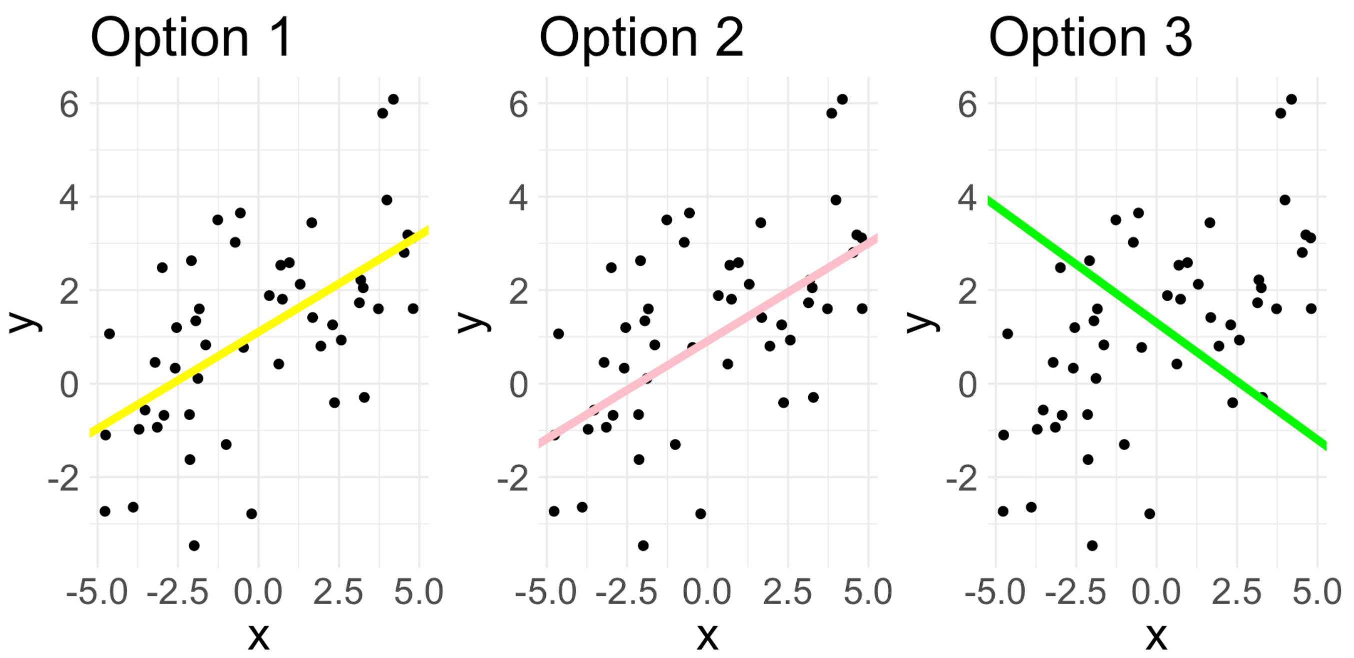
Least squares regression

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

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

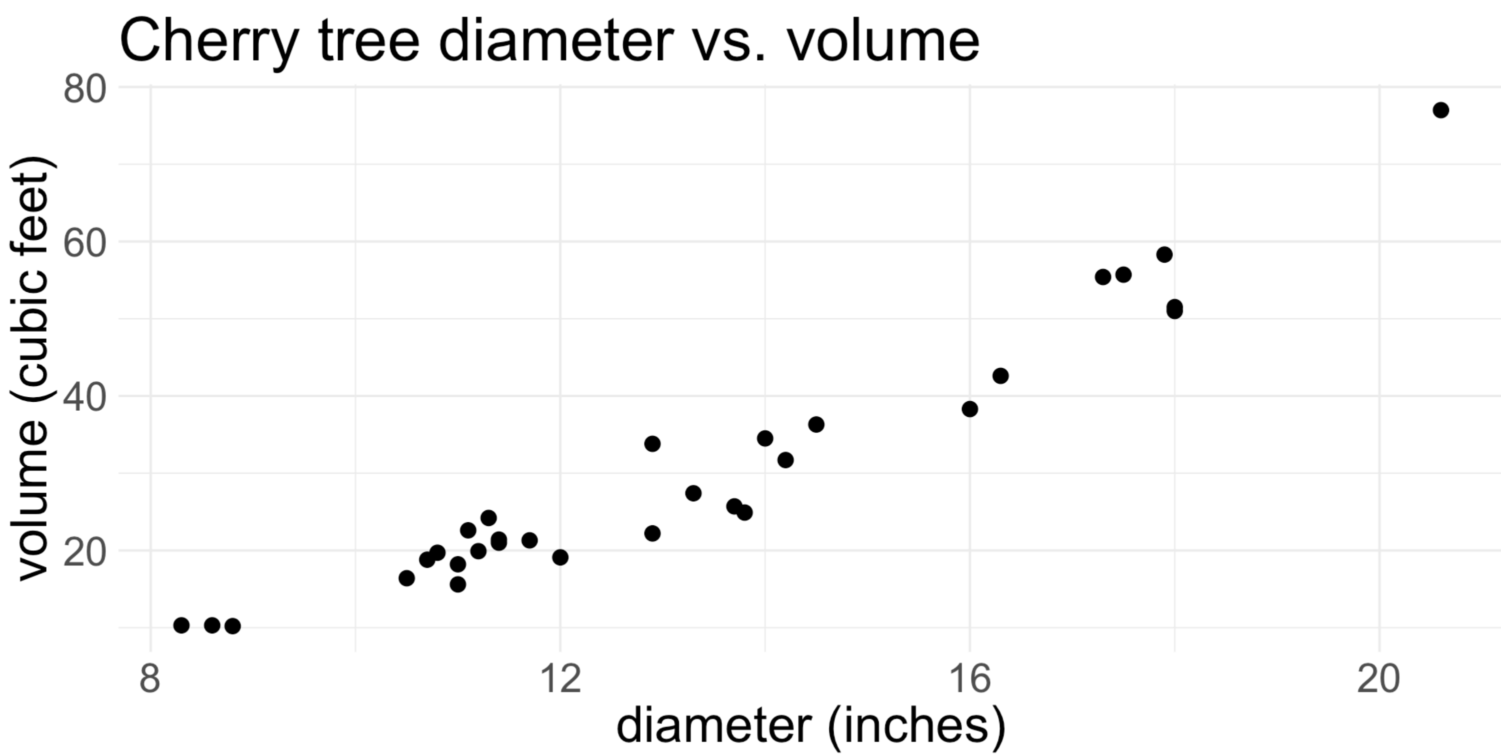
We will see how to check for these four LINE conditions using the `cherry` data from `openintro`.

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`

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?

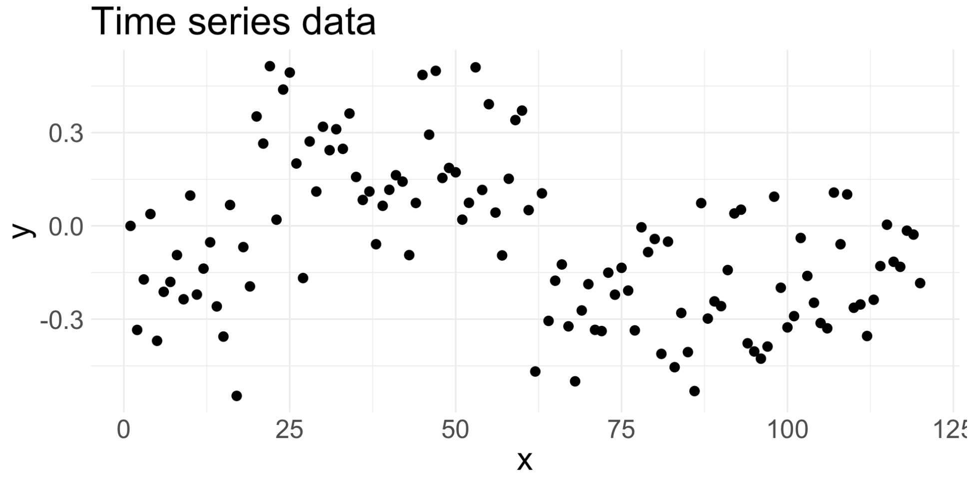
- I would say yes

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

At this point, it is time to actually fit our model

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

- 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.

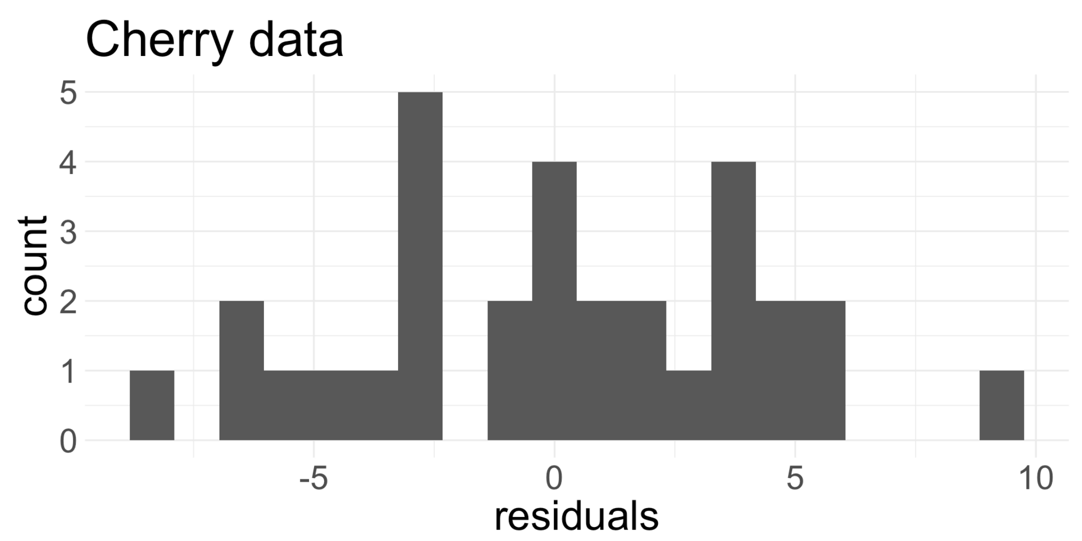
3. Nearly normal residuals

Assess *after* fitting the model by obtaining residuals and making a histogram.

- Remember, residuals are $\hat{y}_i - y_i$

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

diam	volume	volume_hat	residual
8.3	10.3	5.108	-5.192
8.6	10.3	6.628	-3.672
8.8	10.2	7.641	-2.559
10.5	16.4	16.253	-0.147
10.7	18.8	17.266	-1.534

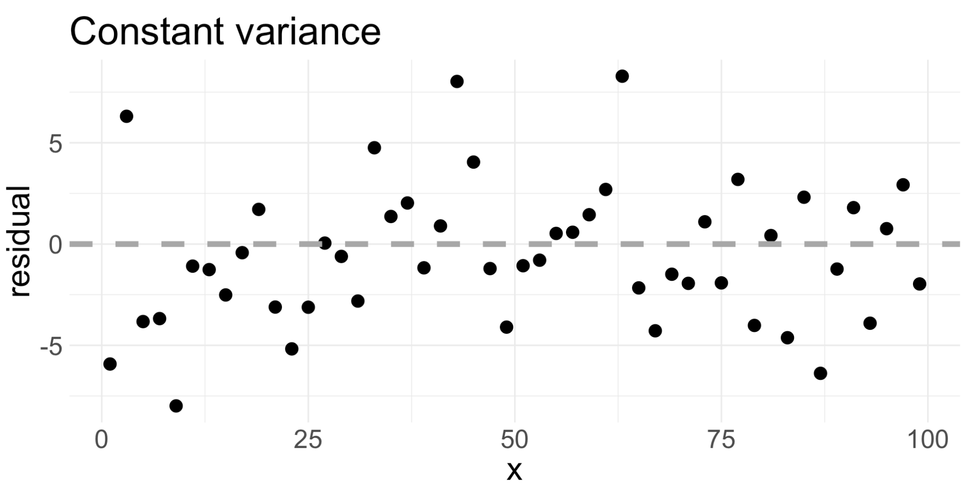


- Do the residuals appear approximately Normal?
  - I think so!

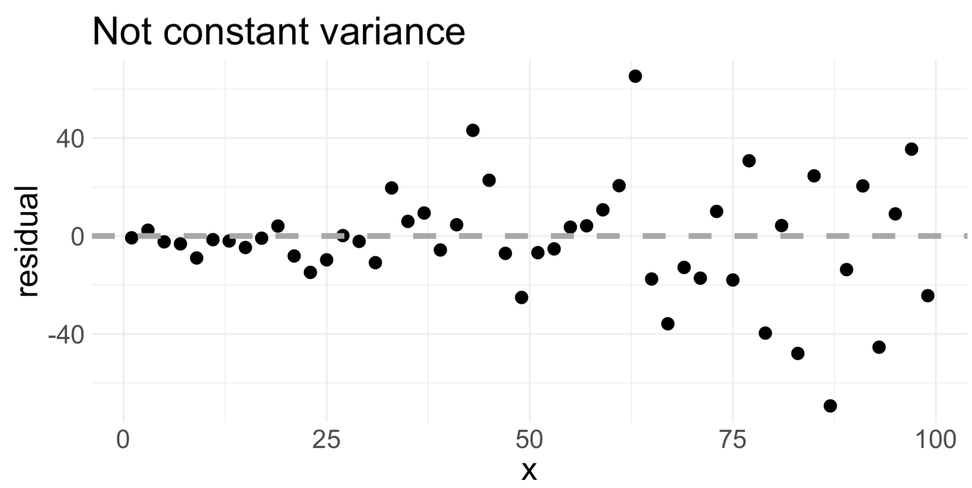
4. Equal variance

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

A good residual plot:



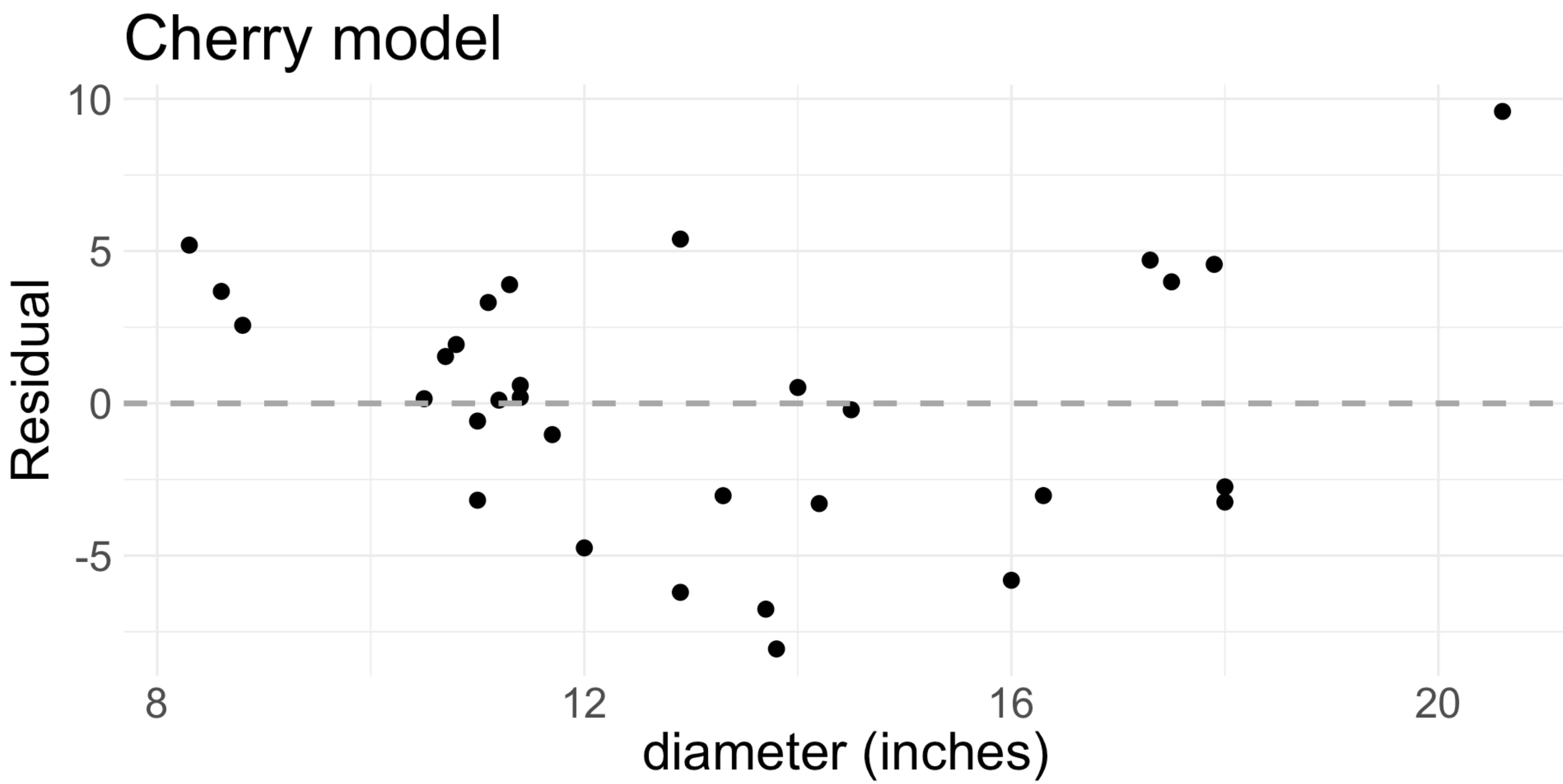
A bad residual plot:



We usually have a horizontal line at 0 to compare residuals to

4. Equal variance (cont.)

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



- Based on this plot, I would say the equal variance condition is not perfectly met.
  - Some of the variability in the errors appear related to `diameter`

Fitting the least-squares line

Parameter estimates

- Like in previous topics, we have to estimate the parameters using data
- We want to estimate β_0 and β_1 using the (x_i, y_i)
  - In practice, we let software do this for us
- However, we *can* derive the least-squares estimates using properties of the least-squares line

Estimating slope and intercept

First obtain b_1:

$$b_1 = \frac{s_y}{s_x} R$$

where:

- s_x and s_y are the sample standard deviations of the explanatory and response variables
- R is the correlation between x and y

Then obtain b_0:

$$b_0 = \bar{y} - b_1 \bar{x}$$

where

- $\bar{y}$ is the sample mean of the response variable
- x is the sample mean of the explanatory variable

- Take STAT 0211 or 0311 to see where these formulas come from!

Fitting cherry model (by hand)

Verify estimates $b_0 = -36.94$ and $b_1 = 5.07$ from our model for the cherry data:

```r
cherry |>
  pivot_longer(cols = c(diam, volume), names_to = "variable", values_to = "val") |>
  select(-height) |>
  group_by(variable) |>
  summarise(mean = mean(val), s = sd(val))
```

variable	mean	s
diam	13.248	3.138
volume	30.171	16.438

```r
R <- cor(cherry$diam, cherry$volume)
R
```

```
[1] 0.9671194
```

- Set-up the calculations:
  - $b_1 = \frac{s_y}{s_x} R$
  - $b_0 = \bar{y} - b_1 \bar{x}$

- $b_1 = \frac{16.438}{3.138} \times 0.967 = 5.07$
- $b_0 = 30.171 - 5.07 \times 13.248 = -36.94$