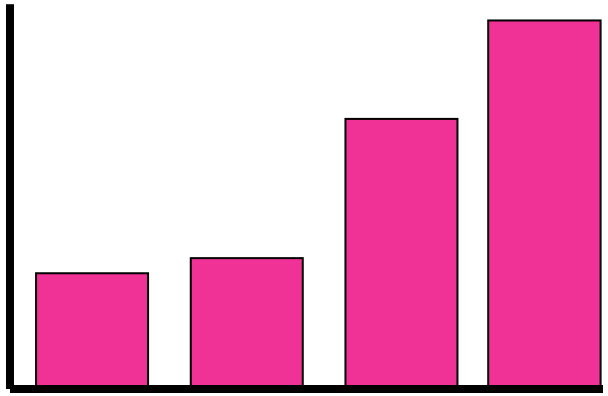
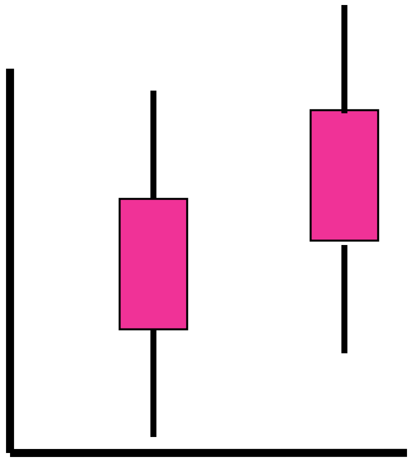
Linear Regression

Questions and Announcements?

T-TESTS & ANOVA

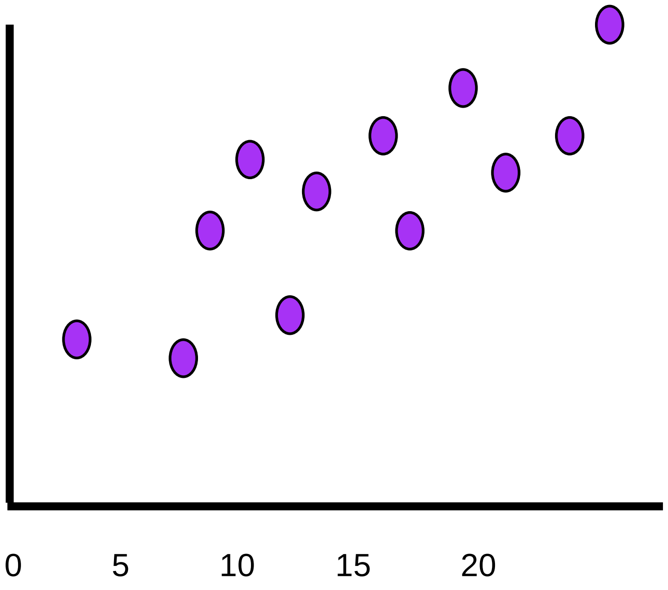
- Differences between groups
    - Predictor is a Factor = Categorical
- Response variable = Continuous
- **Graphs:** Boxplots, bar charts, etc.





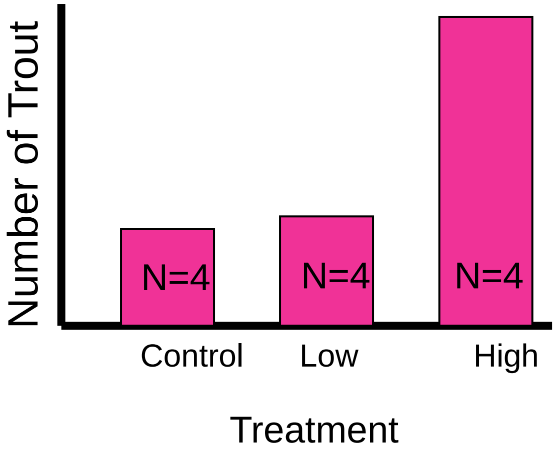
Linear Regression

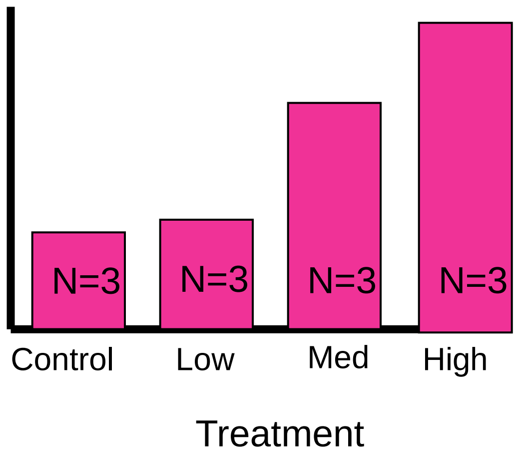
- Relationship between continuous variables
- Develop a predictive relationship
- **Graphs:** Scatterplots

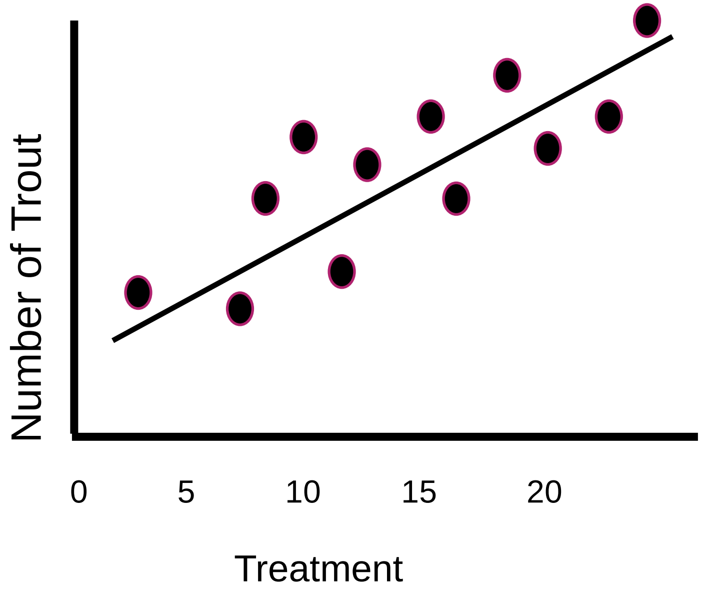


CONTRAST ANOVA VERSUS REGRESSION

12 Experimental Units (e.g., streams):
How allocate these among treatments?







Interested in a specific treatment → ANOVA

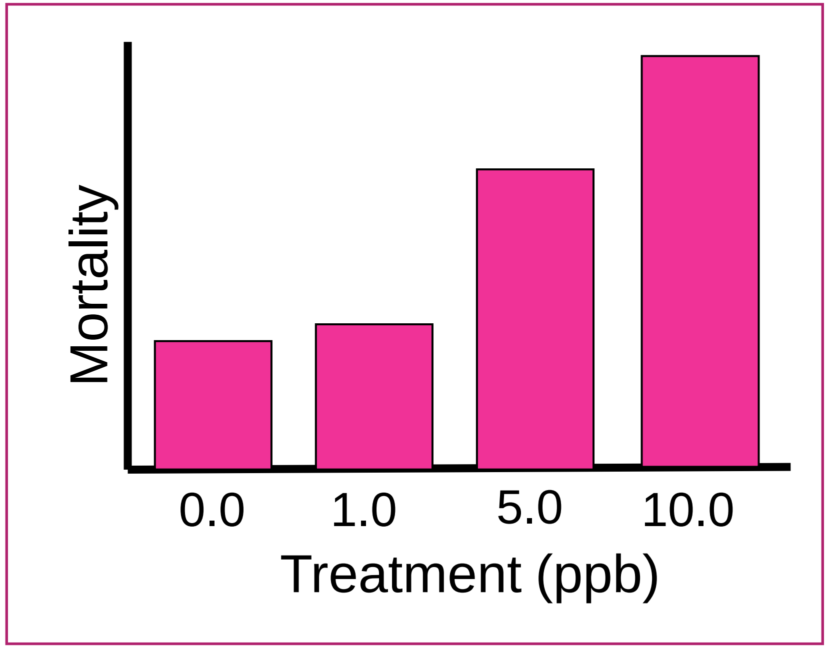
Interested in a predictable relationship → regression

ANOVA VERSUS REGRESSION

Must decide before start of experiment

Trout are sensitive to Cu

Research Q:
Is 5 ppb safe?



ANOVA:
N=6 for each treatment

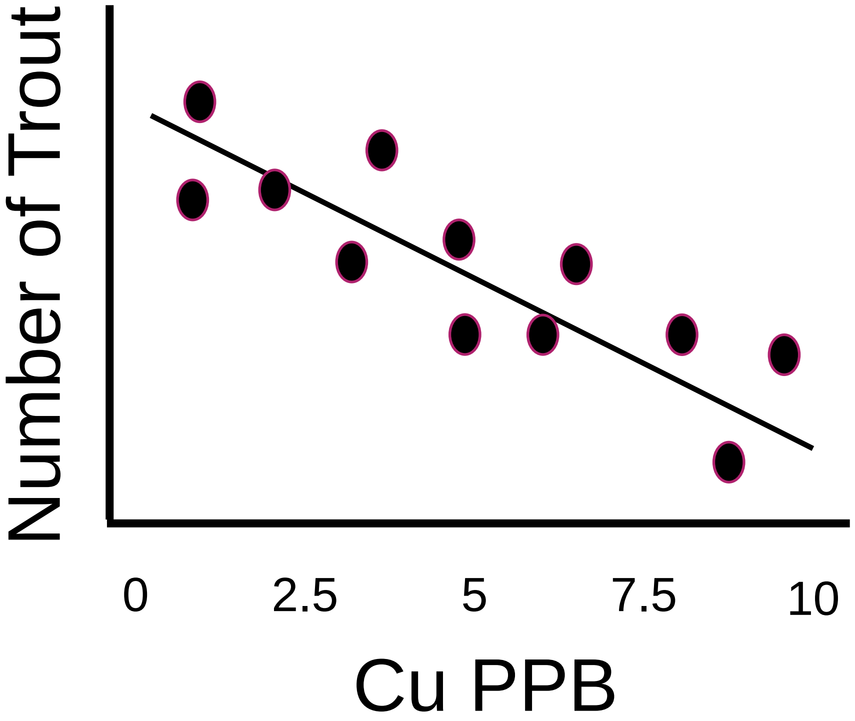
ANOVA VERSUS REGRESSION

Must decide before start of experiment

Trout are sensitive to Cu

Research Question:

What conc. Kills 50% of trout?



Regression:

Each tank gets unique Cu concentration

Does not require strict “replication”

Ensure cover range of x-values

-don’t “clump” them

Linear Regression

- A <u>linear</u> relationship between a dependent variable (Y) and an independent variable(s) (X)
  → Implicit: X <u>causes</u> Y
  (not simply correlation)

- Dependent, independent <u>should</u> be obvious:
  -Island area and number of species
  -Food quality and # of offspring
  -Predator abundance and prey abundance (??)

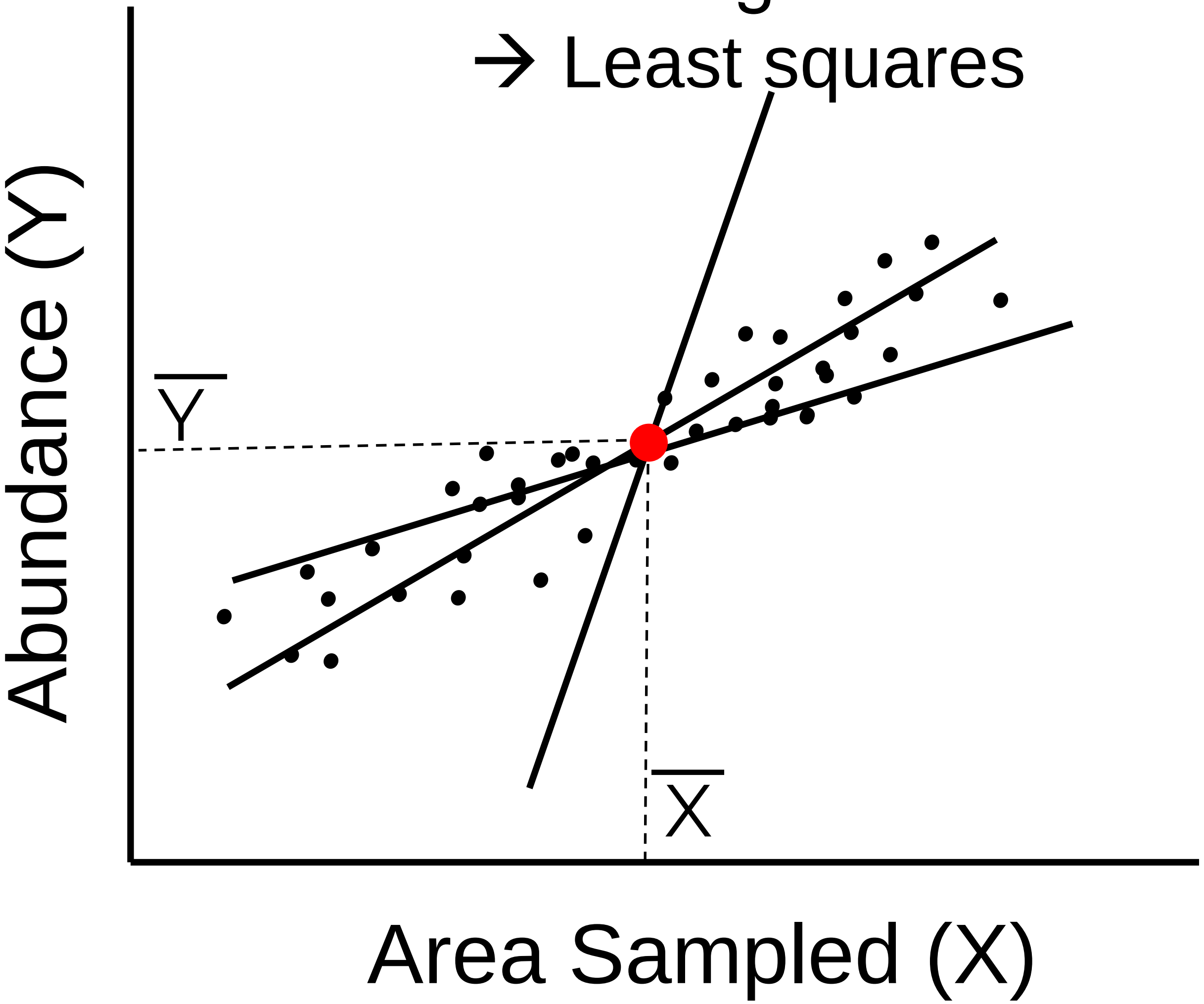
- Make predictions:
  Given X → predict Y
  Given Y → predict X (inverse prediction)

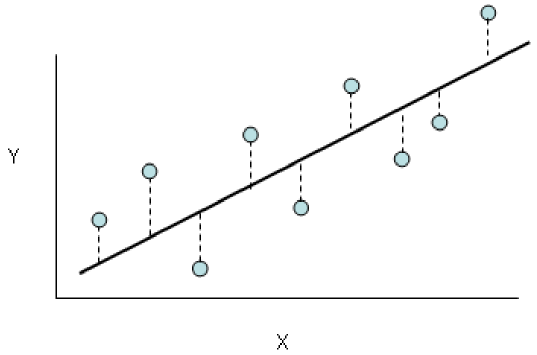
Goal: Find the best straight line through a set of points

2 Conditions: Should pass through $\bar{X}$, $\bar{Y}$

"Best fit" regression

→ Least squares



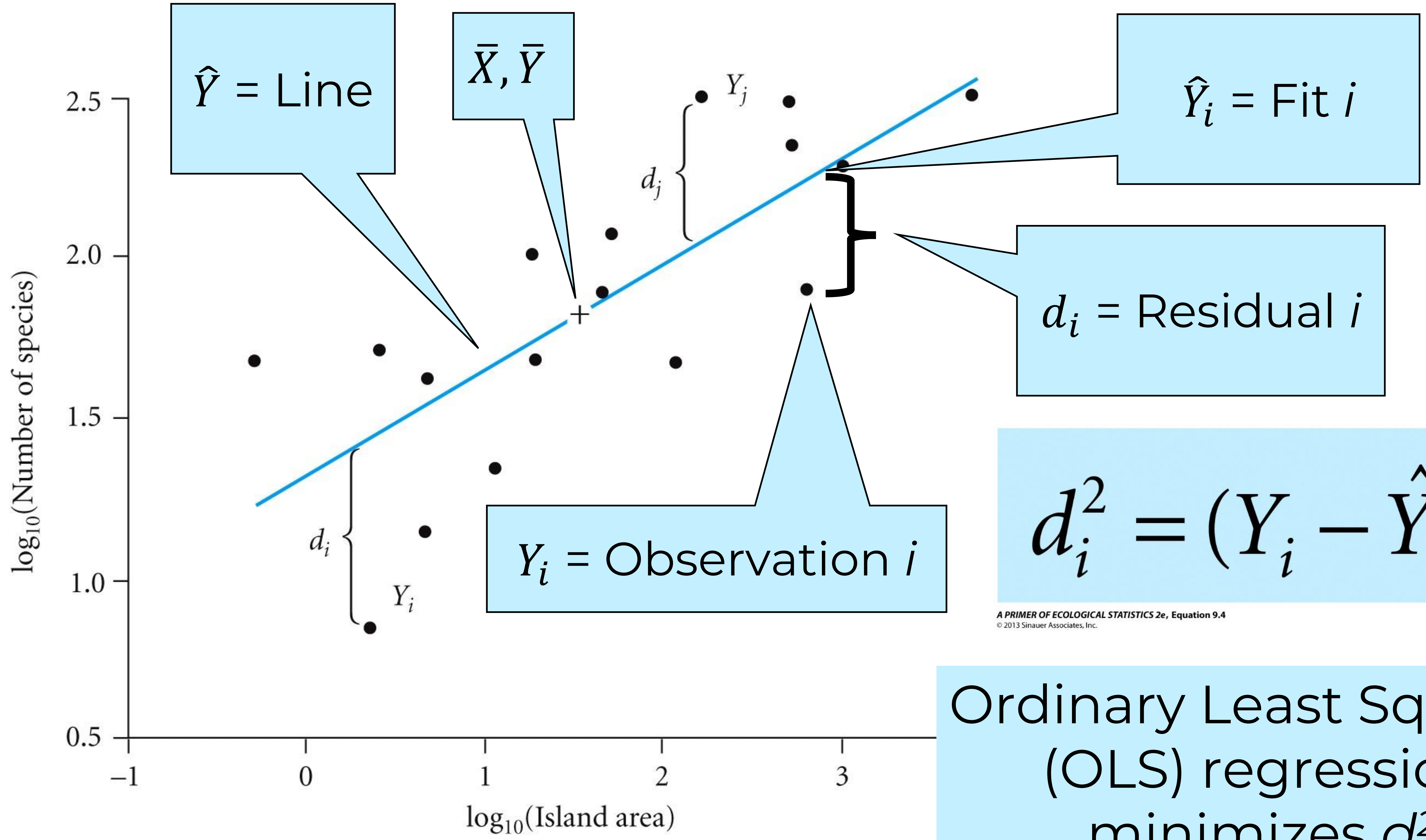


Minimize <u>Residuals</u>

$$(Y_i - \hat{Y}_i)$$

Y_i = observation

$\hat{Y}_i$ = predicted



***A PRIMER OF ECOLOGICAL STATISTICS 2e*, Figure 9.3**

$$d_i^2 = (Y_i - \hat{Y}_i)^2$$

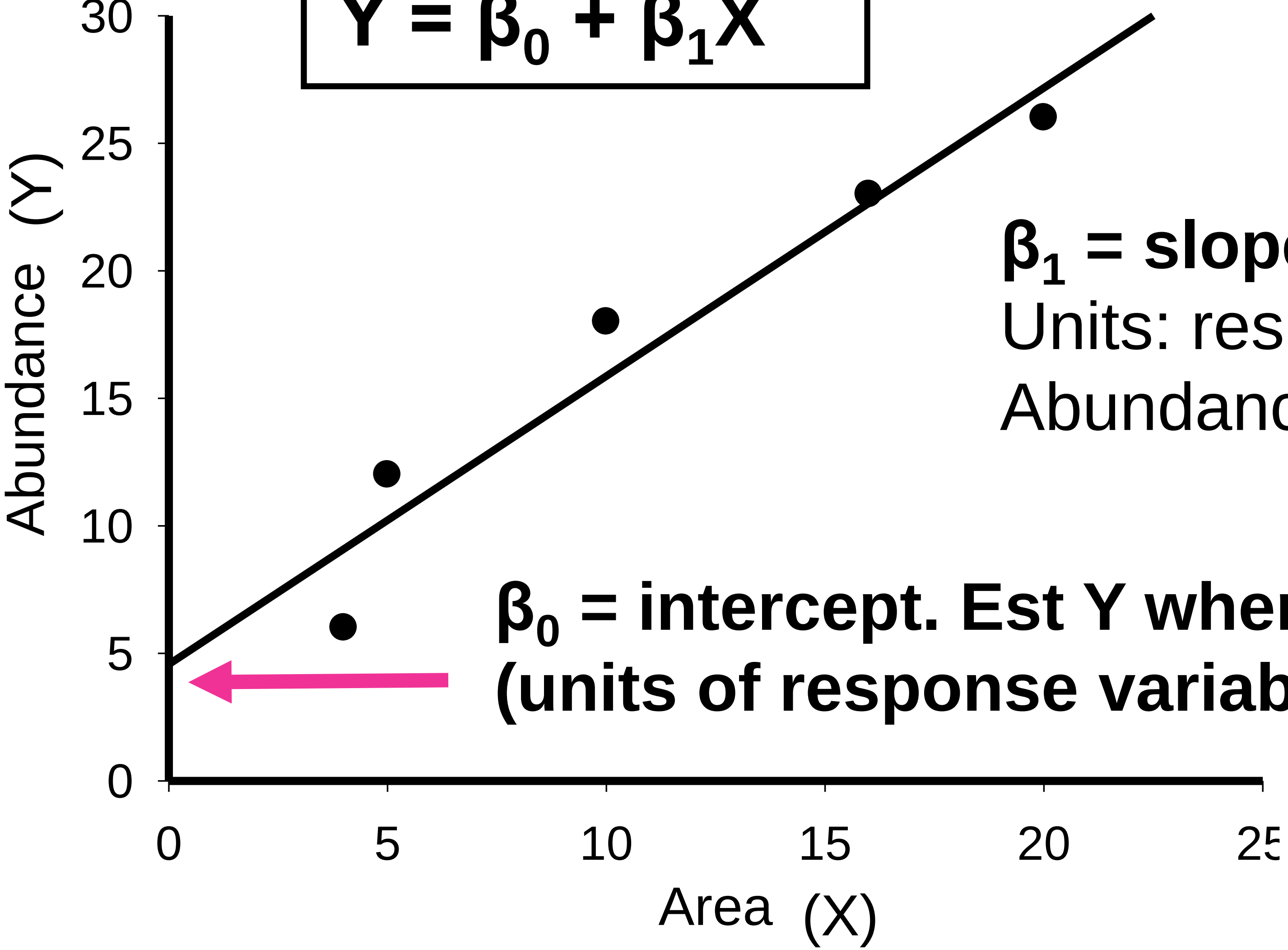
***A PRIMER OF ECOLOGICAL STATISTICS 2e*, Equation 9.4**

Ordinary Least Squares (OLS) regression minimizes d^2

Every line can be described by 2 Parameters

β_0 = Y intercept
β_1 = slope

$$\boxed{Y = \beta_0 + \beta_1 X}$$



β_1 = slope, rate of change
Units: response / predictor
Abundance / Area

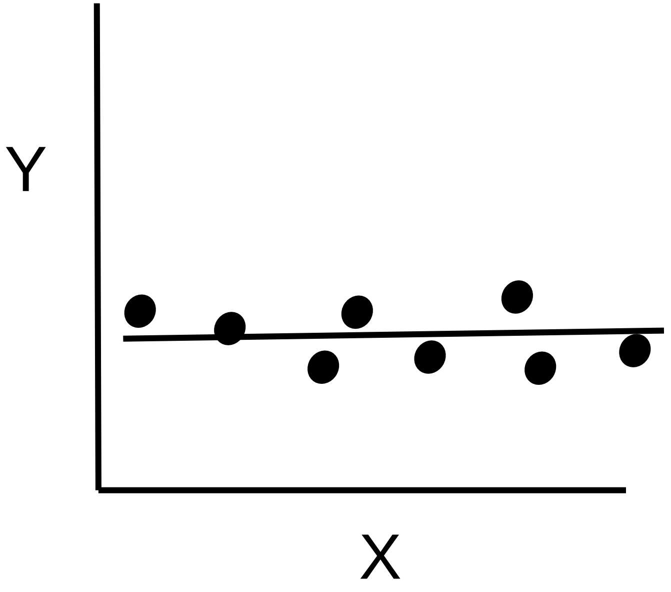
β_0 = intercept. Est Y when X = 0
(units of response variable, Abundance)

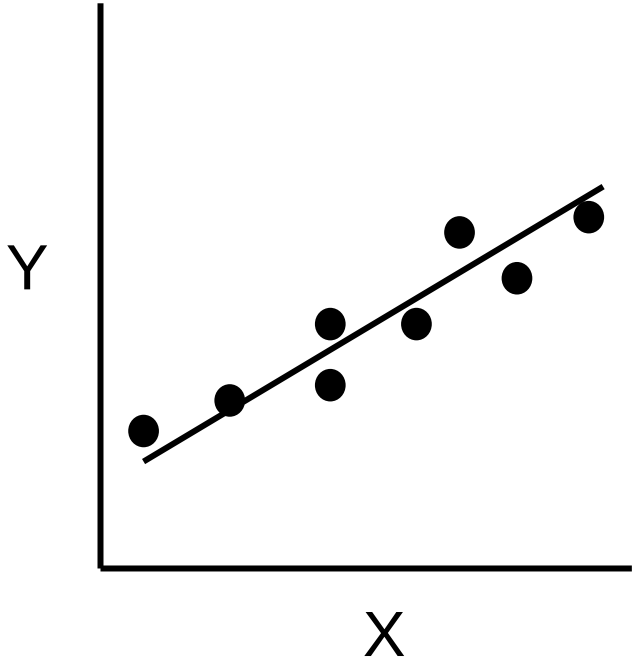
Hypothesis Tests with Regression

→ Is the relationship significant?

β_1 = Regression Coefficient

- Slope Null hypothesis:
  $\beta_1 = 0$, no relationship
- Alternative hypothesis
  $\beta_1 \neq 0$, IS a relationship

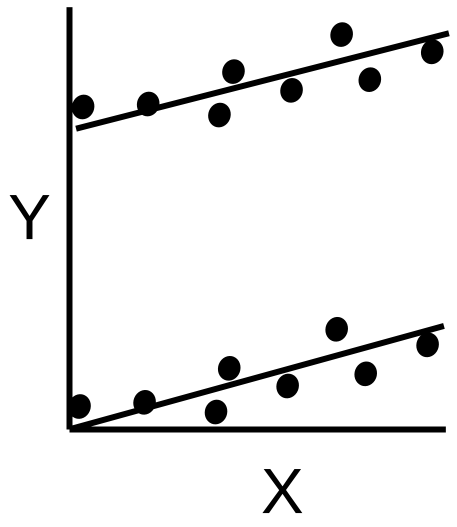




Hypothesis Tests with Regression

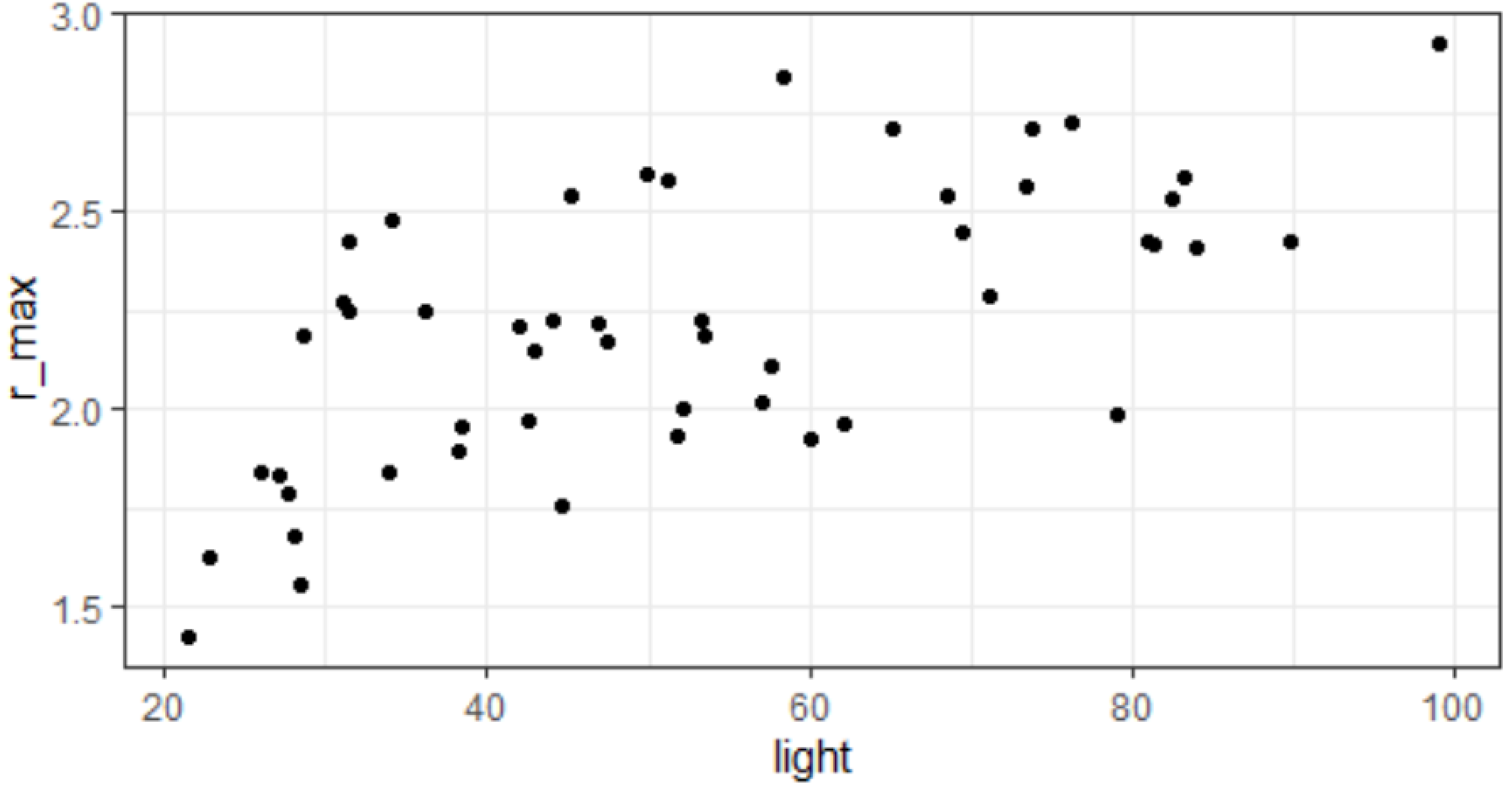
→ is the Y-intercept = 0?

- Intercept = expected value of Y when X = 0
- Intercept Null hypothesis:
  $\beta_0 = 0$
- Alternative hypothesis
  $\beta_0 \neq 0$



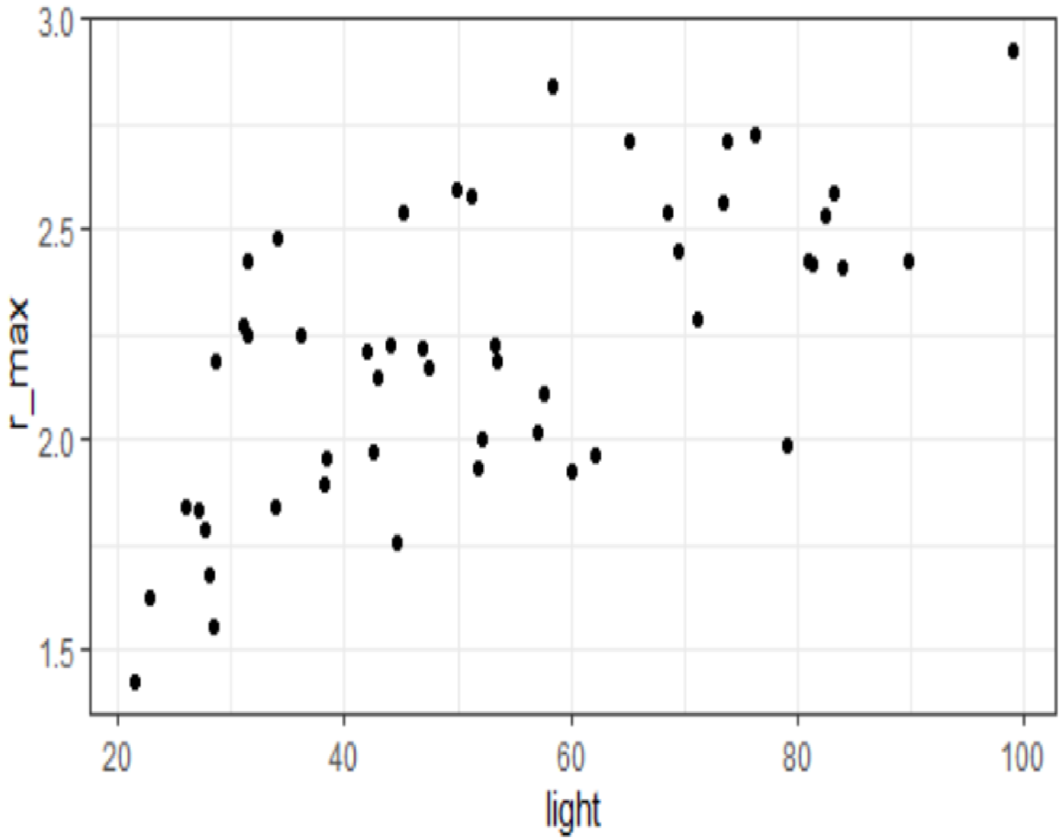
- Often, not really interested in Intercept
- Have to pick a value, and 0 is as good as any other

Example: Does Light intensity influence plant growth?



Example: Does increasing light intensity affect plant growth?

1. Plot data
2. Fit `lm()`
3. ANOVA Table (β_1 Hypothesis)
4. `summary()` of model fit
5. Linear Equation
    1. Predictions



Fit `lm()` and ANOVA table

- `Fit_lm < - lm(r_max ~ light, data = plant_light)`
- Check Assumptions (later)
- `anova(fit_lm)`

```
Analysis of Variance Table

Response: r_max
          Df Sum Sq Mean Sq F value    Pr(>F)
light      1 2.5338 2.53380  35.527 2.881e-07 ***
Residuals 48 3.4234 0.07132
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA interpretation: β_1 only

- We reject the null hypothesis and conclude that the slope (beta_1) is not equal to 0 ($F_{1, 48} = 35.5$, $p < 0.001$)

```
Analysis of Variance Table

Response: r_max
          Df Sum Sq Mean Sq F value    Pr(>F)    
light      1 2.5338 2.53380  35.527 2.881e-07 ***
Residuals 48 3.4234 0.07132                      
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
summary(fit_lm)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.625210   0.105377   15.42  < 2e-16 ***
light       0.011167   0.001874    5.96 2.88e-07 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2671 on 48 degrees of freedom
Multiple R-squared:  0.4253,    Adjusted R-squared:  0.4134

F-statistic: 35.53 on 1 and 48 DF,  p-value: 2.881e-07
```

summary(fit_lm)

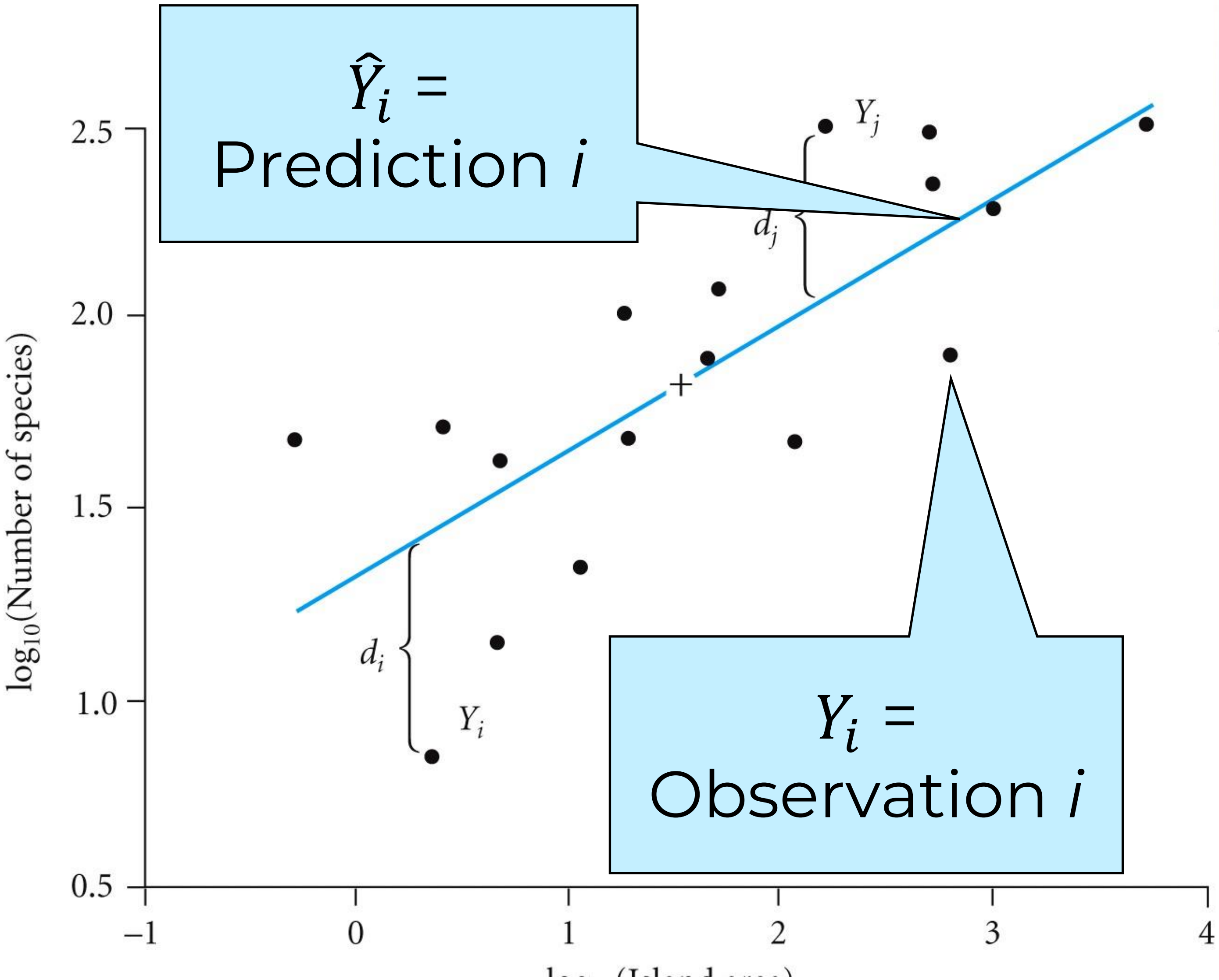
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.625210   0.105377   15.42  < 2e-16 ***
light       0.011167   0.001874    5.96 2.88e-07 ***
---
```

We reject both the null hypotheses and conclude that the intercept is not equal to zero ($t = 15.42$, $p < 0.001$) nor is the regression coefficient (beta_1) equal to 0 ($t = 5.96$, $p < 0.001$).

```
Residual standard error: 0.2671 on 48 degrees of freedom
Multiple R-squared:  0.4253,    Adjusted R-squared:  0.4134
```

Adjusted $R^2 = 0.4134$; model explains 41% of the variation in the data

Predictive ability increases with *decreasing* Residual Sum of Squares (RSS)



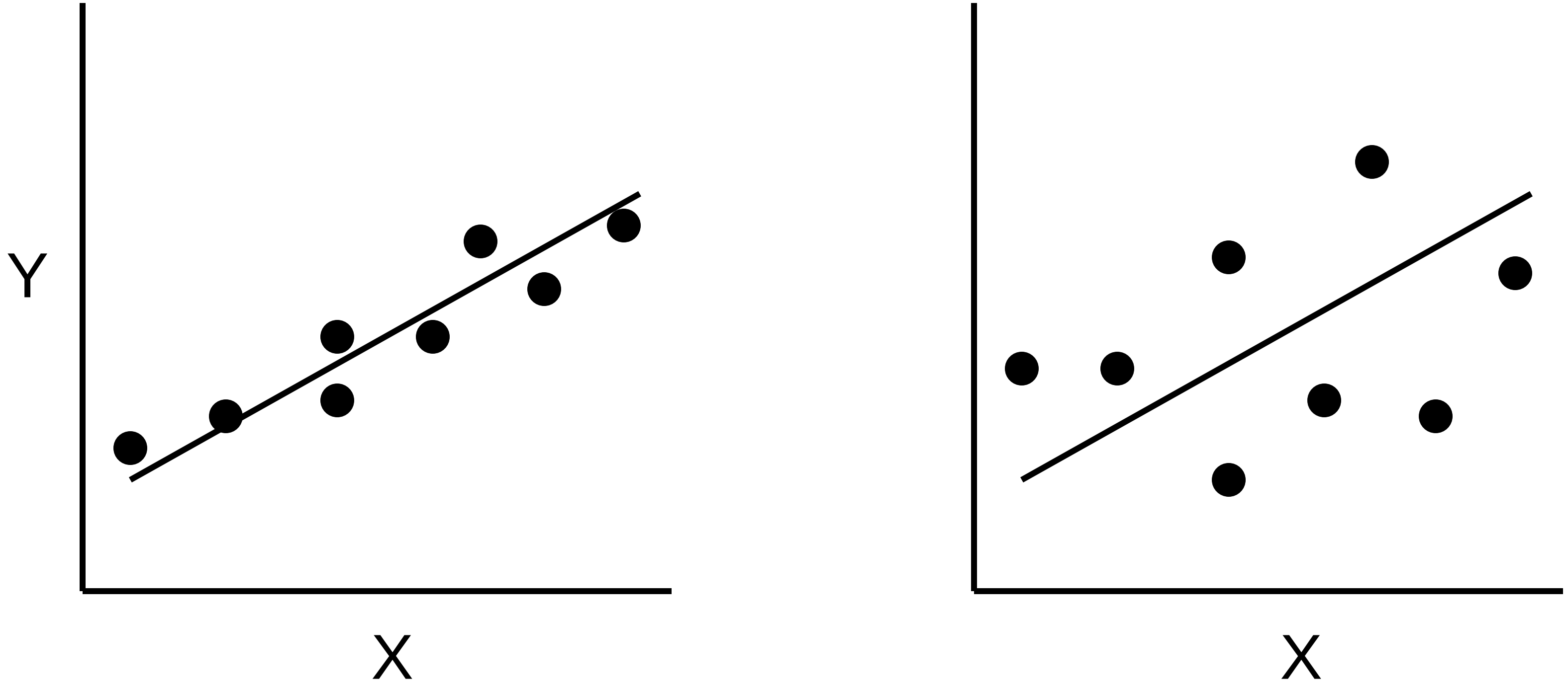
$$RSS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

***A PRIMER OF ECOLOGICAL STATISTICS 2e*, Equation 9.5**

OLS tries to minimize RSS
Small RSS = Good predictive ability → high R^2

Closer the points are to the line, higher R^2 value

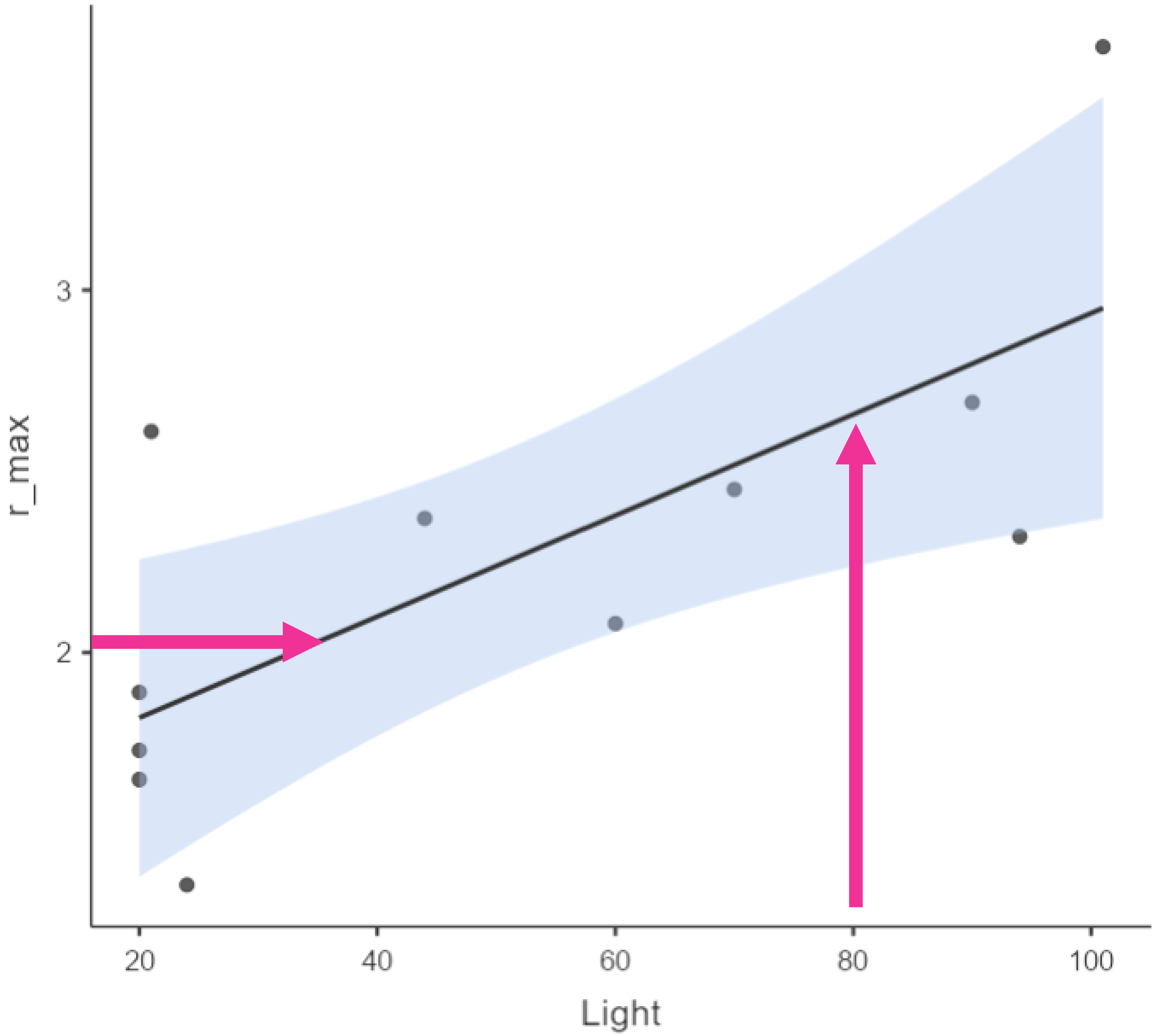
Visualization tool:
https://demonstrations.wolfram.com/VisualizingRSquaredInStatistics/
Change ρ value to increase/decrease R^2



$R^2 \sim 0.8$ $R^2 \sim 0.4$

Estimate values

- Estimate response ($\hat{Y}$) for unmeasured levels of predictor (80W)
- Estimate predictor for future response measurements (2.0)



Equation of the line

```
Coefficients:
            Estimate Std. Error
(Intercept) 1.625210   0.105377
light       0.011167   0.001874
```

- Estimate = the coefficient value
- Write out the equation of the regression line
- $Y = \beta_0 + \beta_1 X$
- $\beta_0 = 1.63$; $\beta_1 = 0.01$; Y = r_max; X = Light

- `R_max = 1.63 + 0.011 * Light`

Using equation, Estimate value

- `R_max = 1.63 + 0.011 * Light`
- How much growth would you expect with an 80W bulb?

Using equation, Estimate value

- `R_max = 1.63 + 0.011 * Light`
- How much growth would you expect with an 80W bulb?

- `R_max = 1.63 + 0.011 * 80`
- = 2.51

Using equation, estimate value

- `R_max = 1.63 + 0.011 * Light`
- Measured a plant that grew 1.9, what bulb was used?

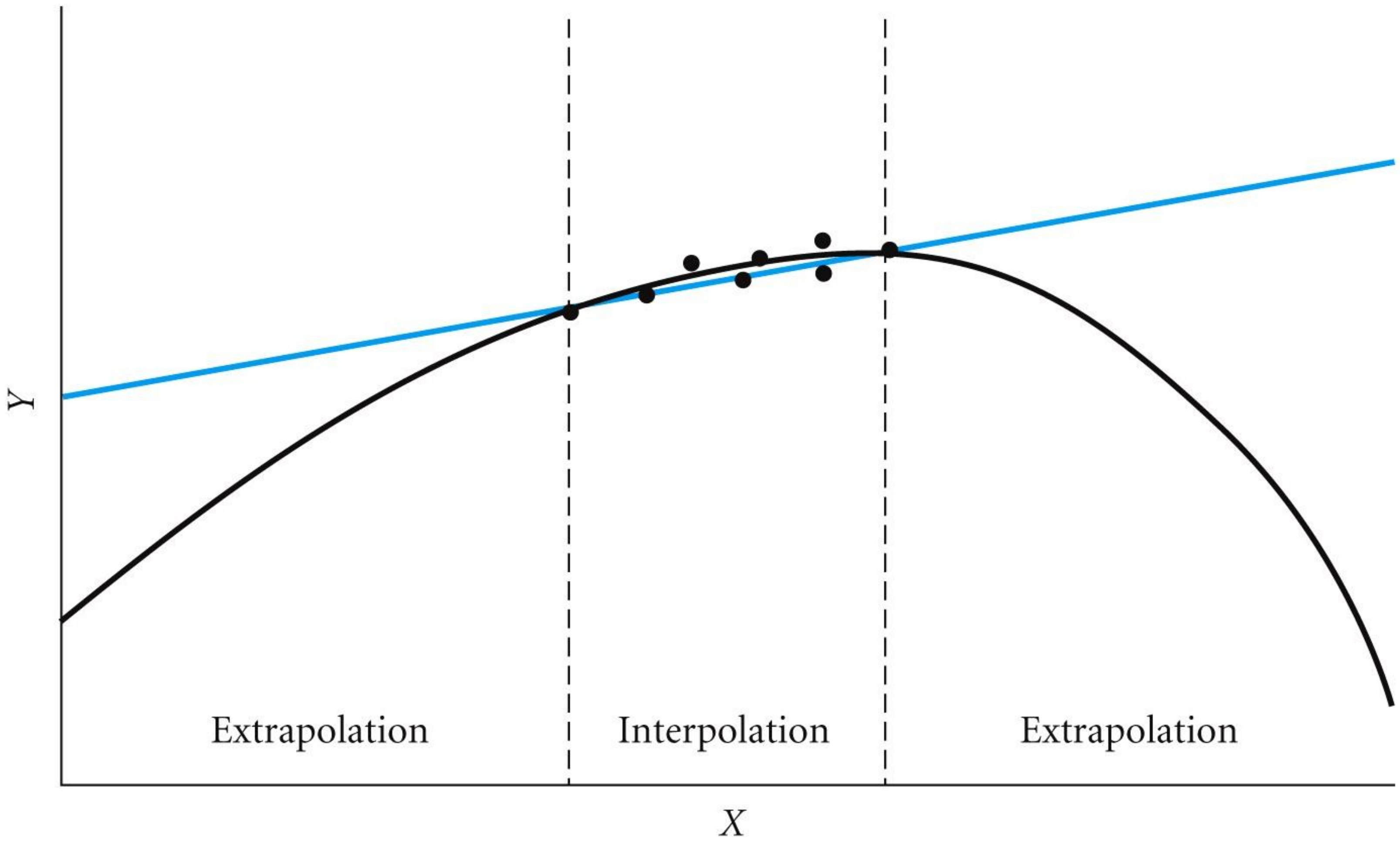
Using equation, Estimate value

- `R_max = 1.63 + 0.011 * Light`
- Measured a plant that grew 1.9, what bulb was used?

- Re-arrange equation
- `(R_max - 1.63) / 0.011 = Light`

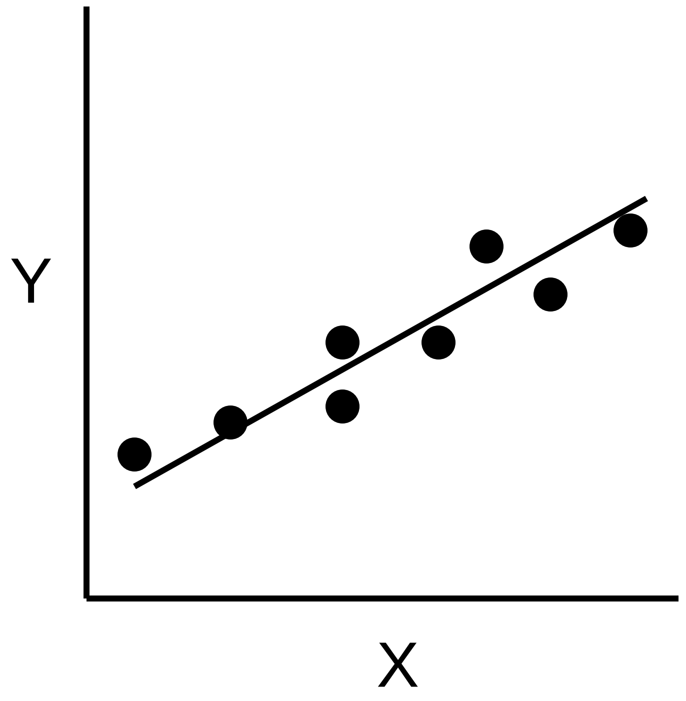
- `(1.9 - 1.63) / 0.011 = 24.5`

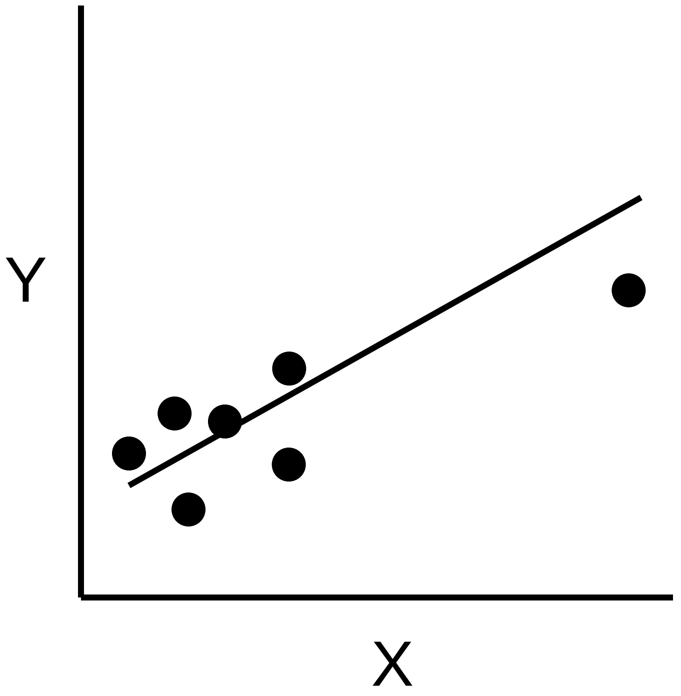
Interpolation Vs. Extrapolation



***A PRIMER OF ECOLOGICAL STATISTICS 2e*, Figure 9.2**

Range of X-values should be ~ uniform





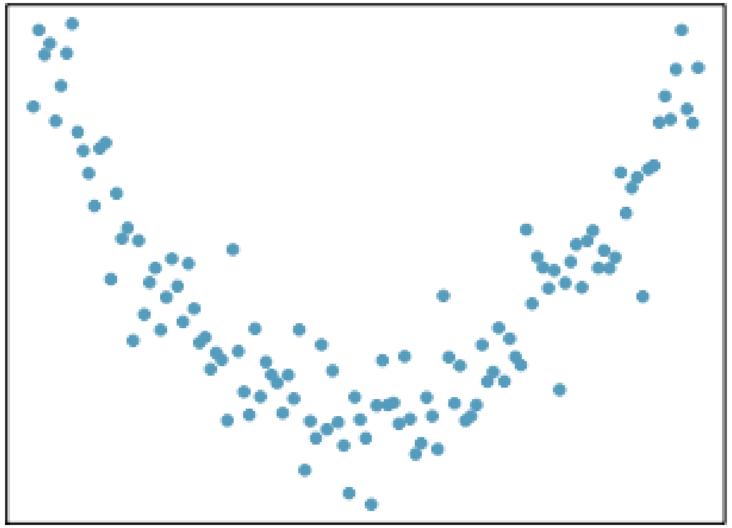
Assumptions

1. Linear Model is correct
   - Check for non-linear pattern in scatter plot
2. X variable accurately measured
3. For X value, Y’s are independent with normally distributed errors
   - Residual vs. Fitted Plots
4. Variances are constant along regression line
   - Points are approx. Uniform distance from regression line

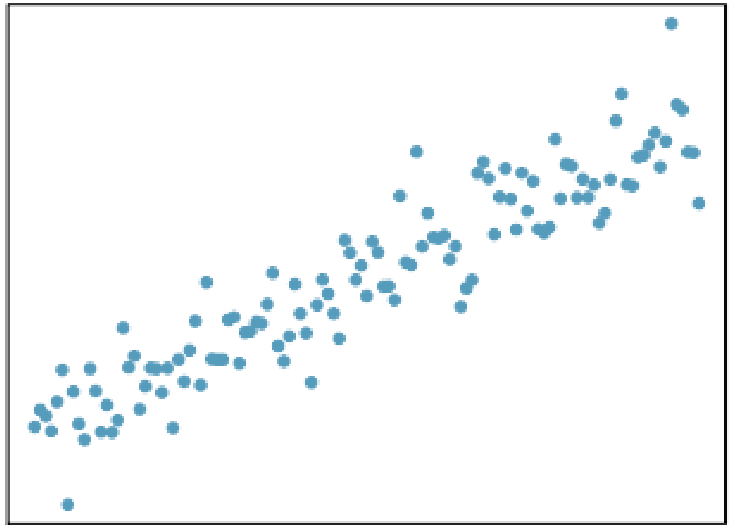
Linear Model (Assumption 1)

Make a scatter plot

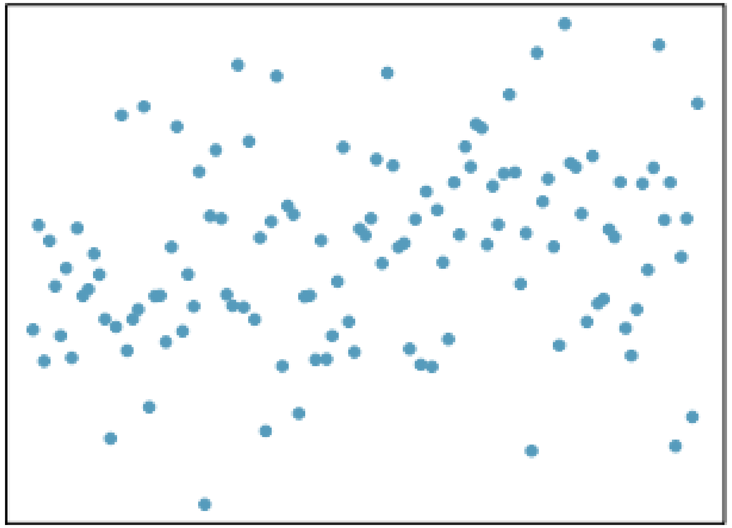
Which ones are linear?



(a)



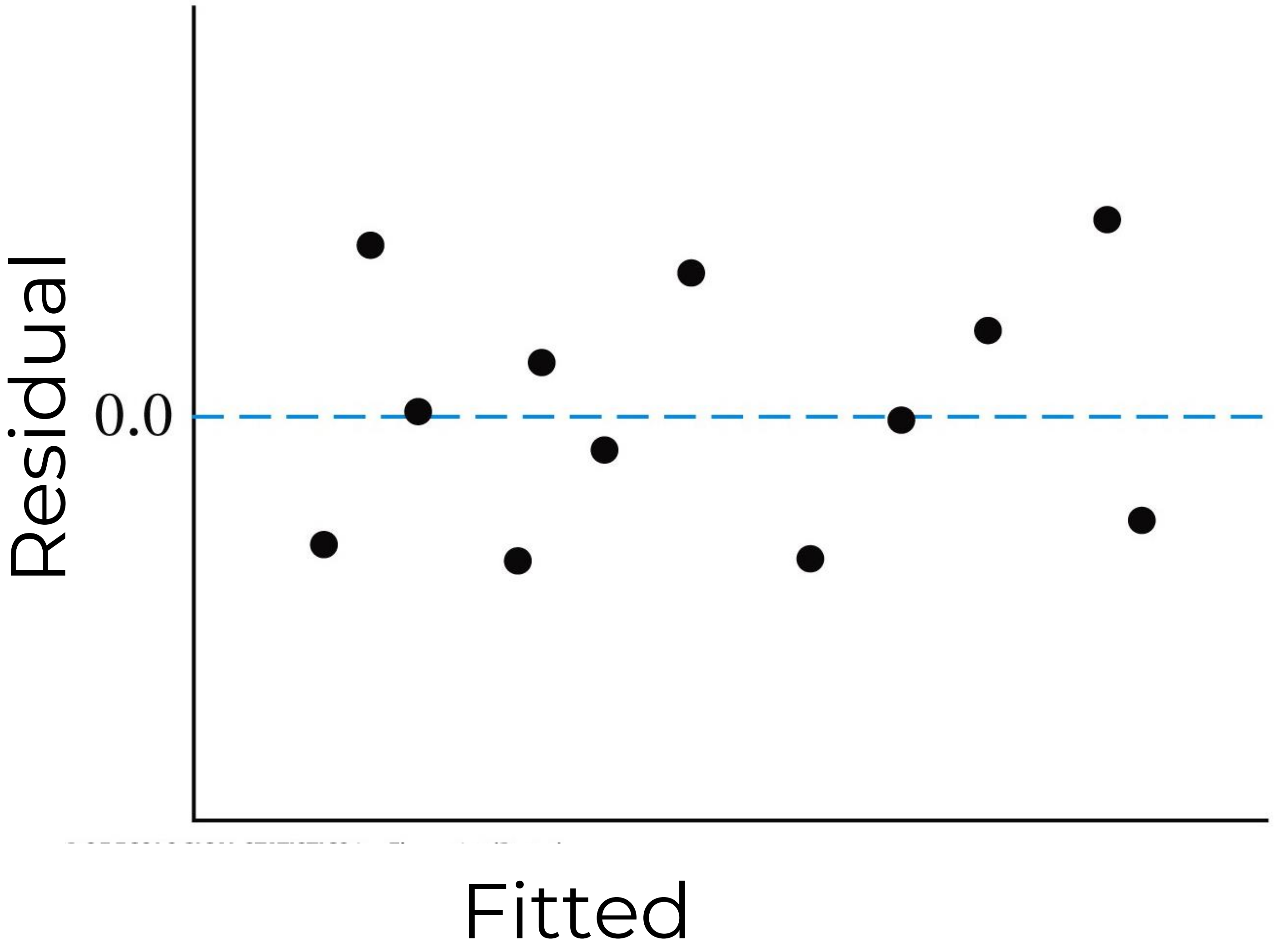
(b)



(c)

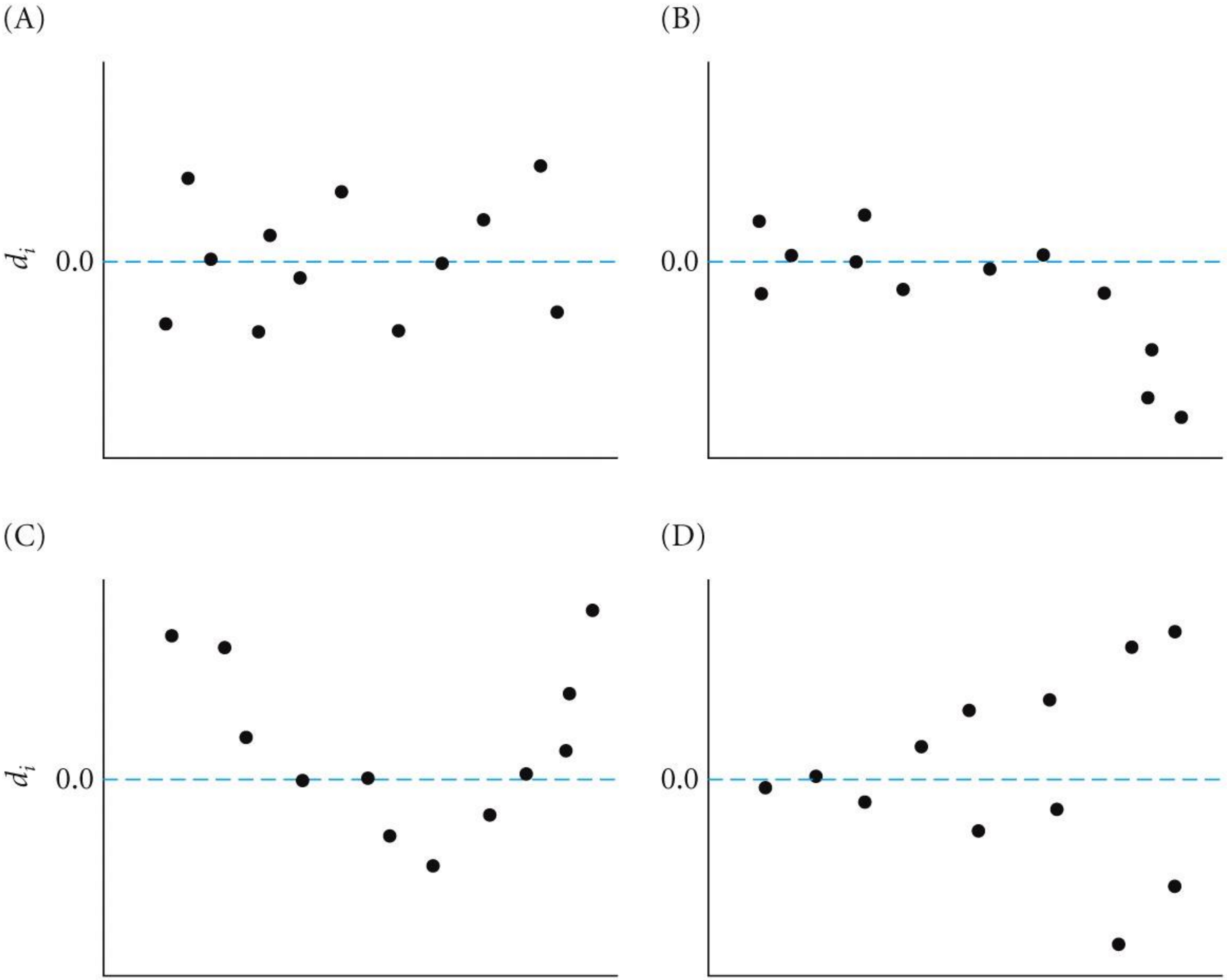
Residual vs. fitted (assumption 3 and 4)

- Residuals are normal
  - Approx. same height above/below 0-line
- Residuals have equal variance
  - Equal distribution across fitted values
  - No apparent pattern



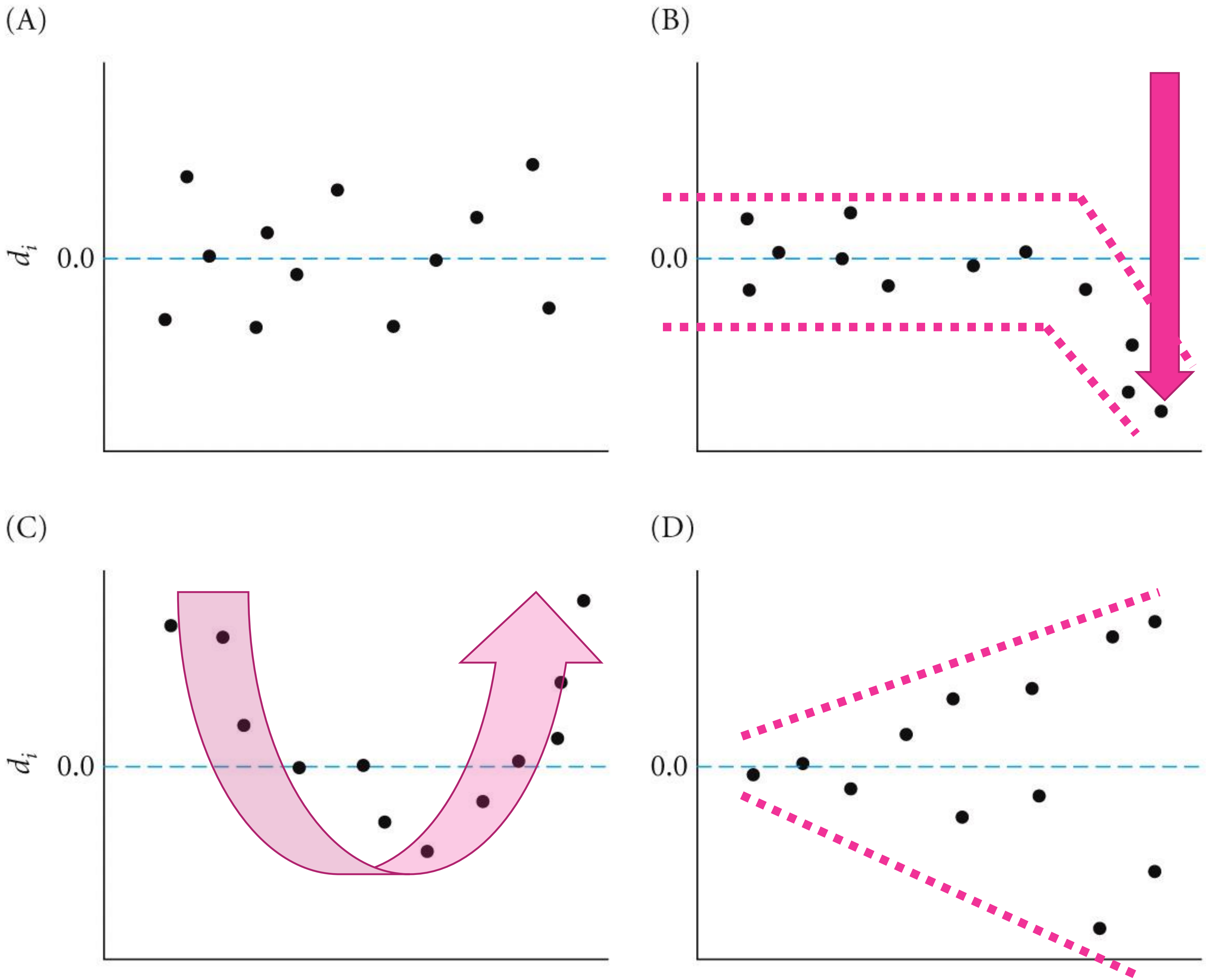
Residual vs. Fitted plots:

What's wrong?



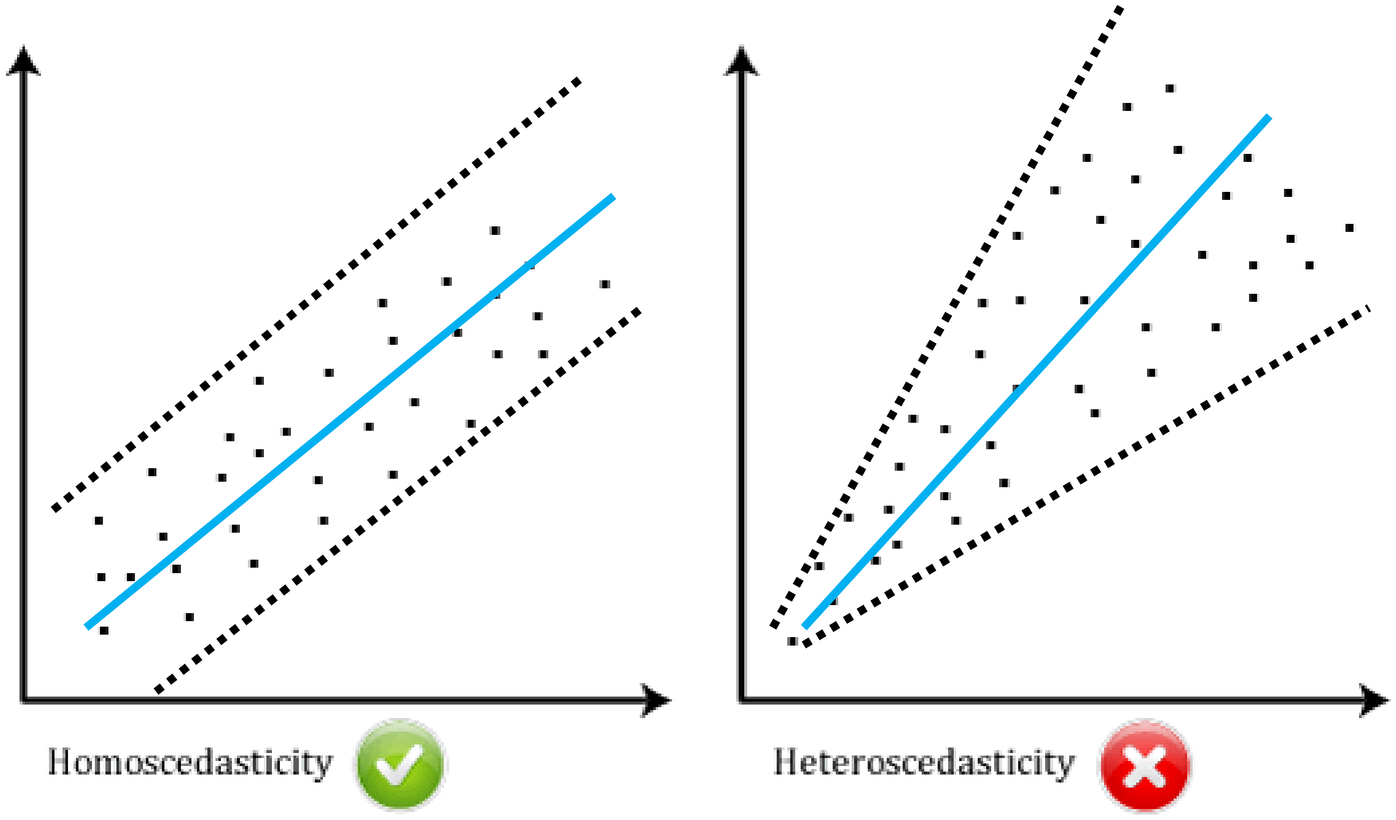
What's wrong?

- B: "drop" at high fitted values
    - Unequal distance above/below 0-line
- C: "U" pattern
- D: "Fan" pattern



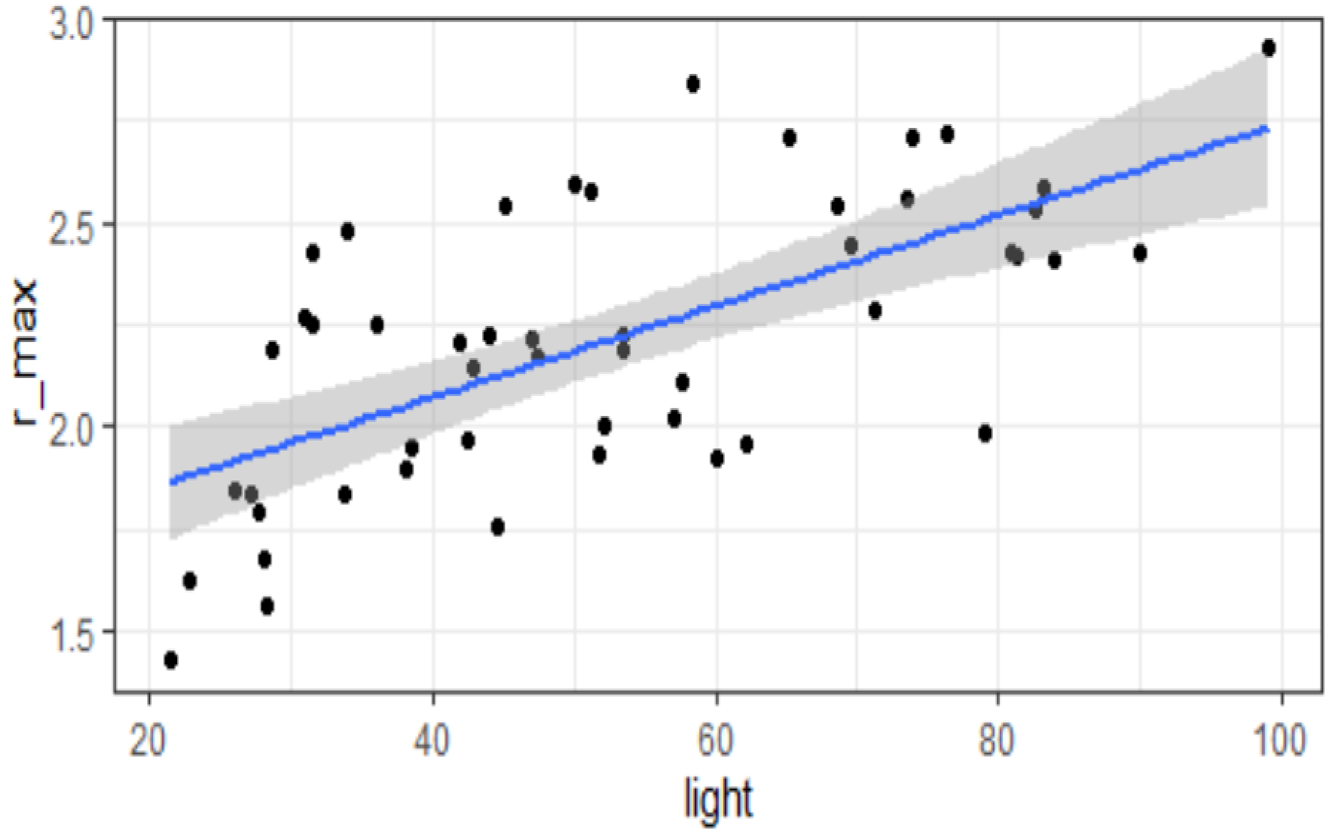
Constant Variance Along Regression Line (Assumption 4)

- Plot regression line + raw data points
- Points should be ~ same distance from regression line across x-axis



Line of best fit to ggplot()

```
ggplot(df,
  aes(x = light,
     y = r_max)) +
 geom_point() +
 geom_smooth(
     method = "lm") +
 theme_bw()
```



Looking forward

- Linear Regression lab on Wednesday
- Friday is open for Homework questions