Correlation Analysis

How can you describe the relationship between two numerical variables?

- **Correlation:** Describes the degree to which two variables are related. Typically you have observed both X and Y (not manipulated either).

- **Regression:** Describes the effect of variable X on variable Y when X is manipulated in some manner.

What can you do with correlations?

- Determine strength and direction of relationship (positive or negative) with Pearson's rho
- Calculate 95% confidence interval
- Determine significance of correlation with p-value

Significant p-values do not always imply biologically meaningful correlations!

What shouldn't you do with correlations?

- Generate lines of best fit – this is only for regression.

- Infer causality. Additional experiments will be necessary for this.

- Compare results of different experimental methodologies for measuring the same variable. See mean-difference plot for visualizing these data

Temperature and Relative Humidity

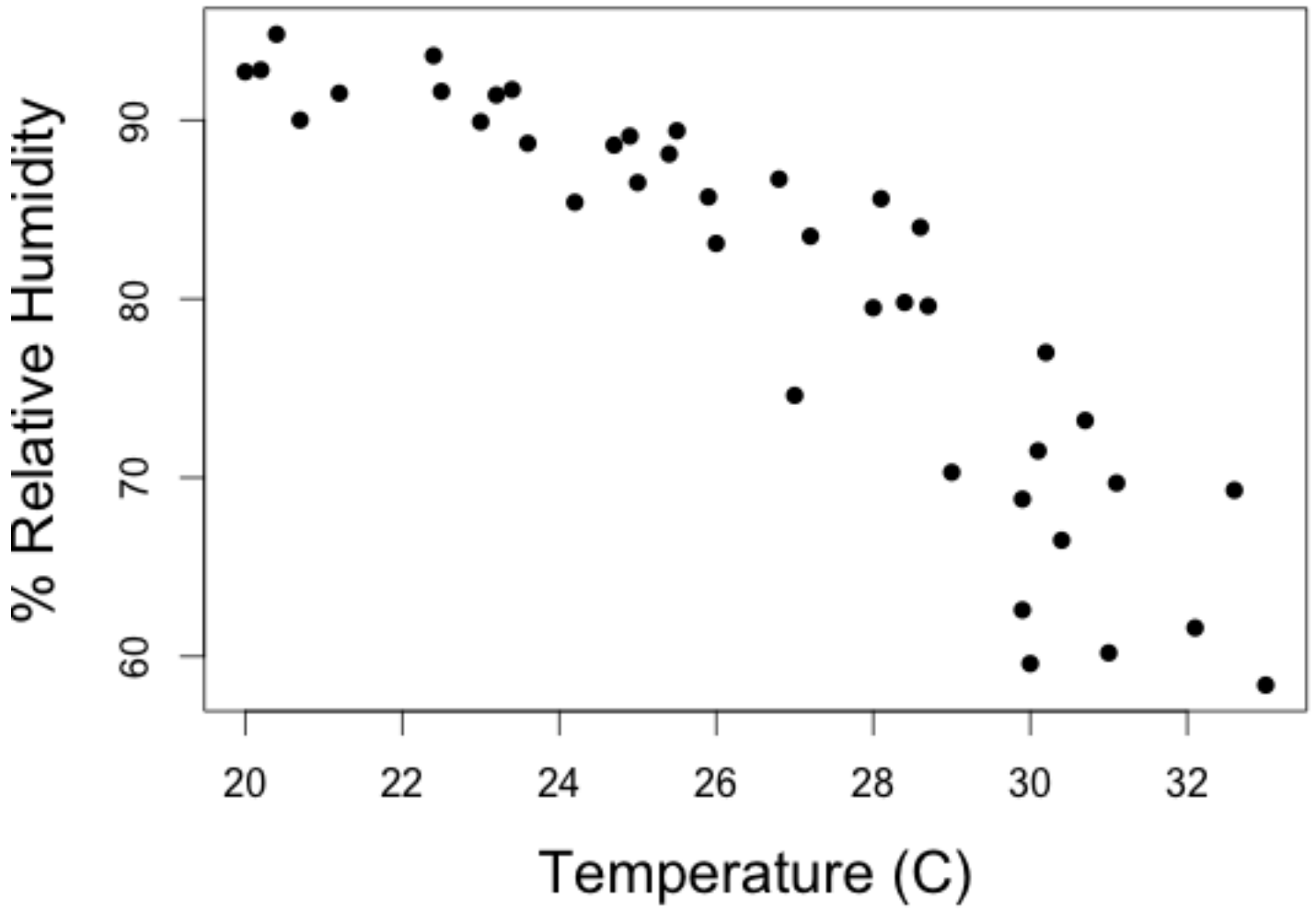
- Analyze with correlation or regression?

DATE	DAY	LOCATION	TIME	TEMP	RH
1-Jun	4	A	7:17	22.5	91.6
1-Jun	4	A	8:20	24.9	89.1
1-Jun	4	A	9:15	26.8	86.7
1-Jun	4	A	10:22	30.7	73.2
1-Jun	4	A	11:09	30.4	66.5
1-Jun	4	B	7:59	24.7	88.6
1-Jun	4	B	9:04	25.9	85.7
1-Jun	4	B	10:11	28.7	79.6
1-Jun	4	B	10:55	30.1	71.5
1-Jun	4	B	11:58	33	58.4
. . .	. . .	. . .	. . .	. . .	. . .

R Break

- Open R file “Correlation_Regression.r”
- Load data
- Using the functions provided:
  - Calculate Pearson’s rho
  - Calculate 95% Confidence Interval
  - Determine significance of correlation

Temperature and Relative Humidity



Temperature and Relative Humidity

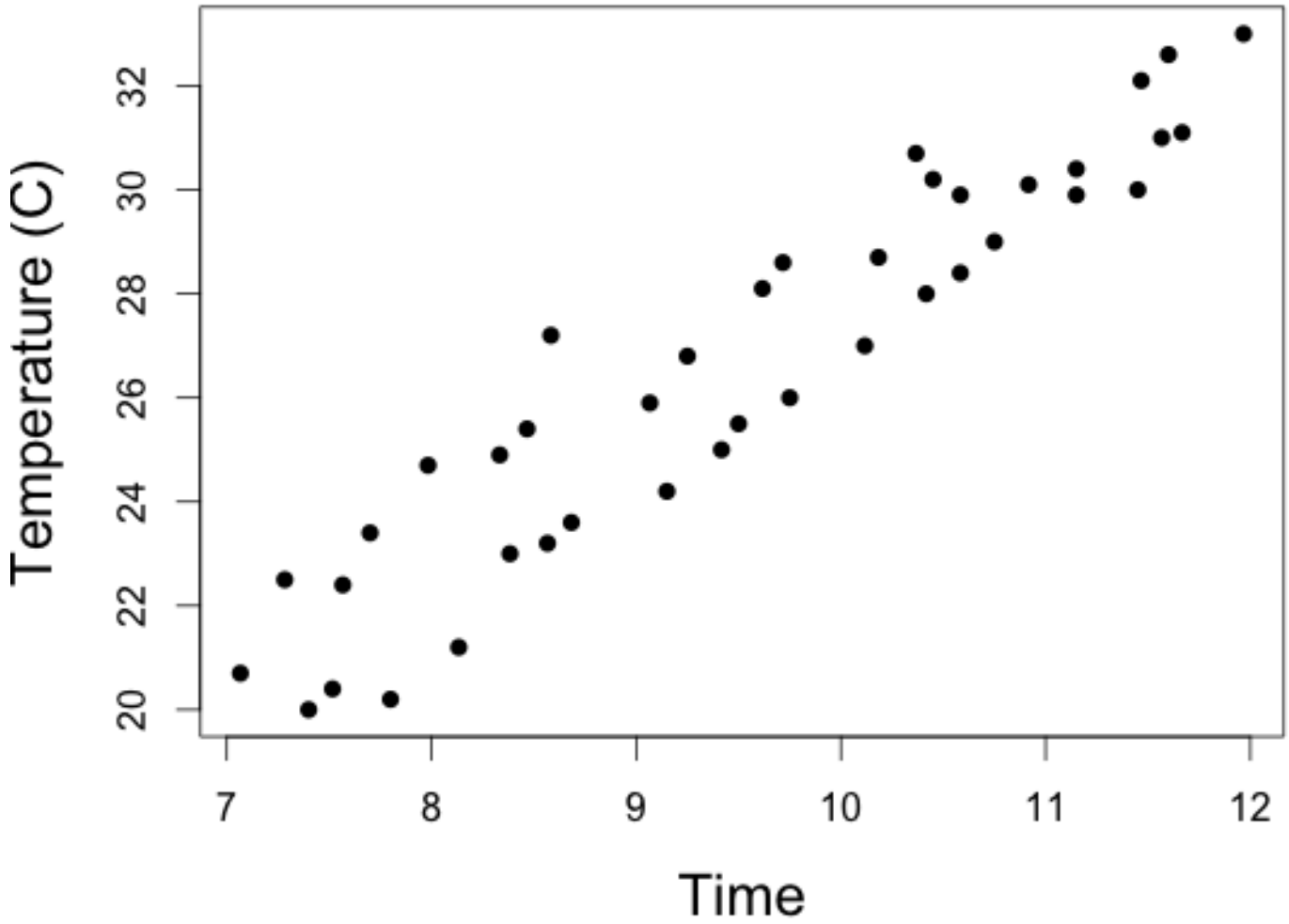
```
> ct<-cor.test(rh, temp, alternative="two.sided")
> ct

 Pearson's product-moment correlation

data:  rh and temp
t = -12.1867, df = 38, p-value = 1.067e-14
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9419978 -0.8044381
sample estimates:
       cor
-0.8923368
```

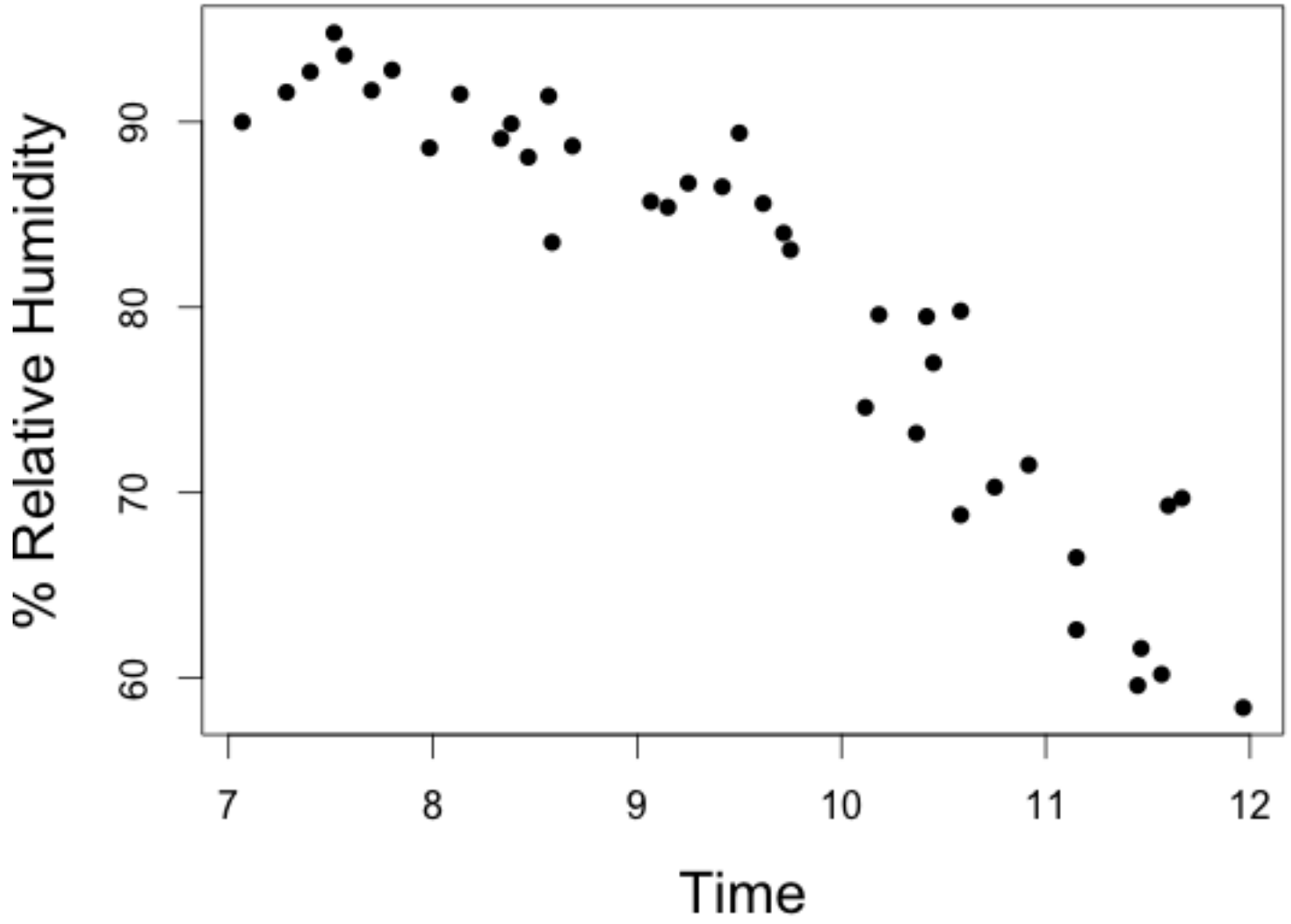
What about the other variables?

- Time



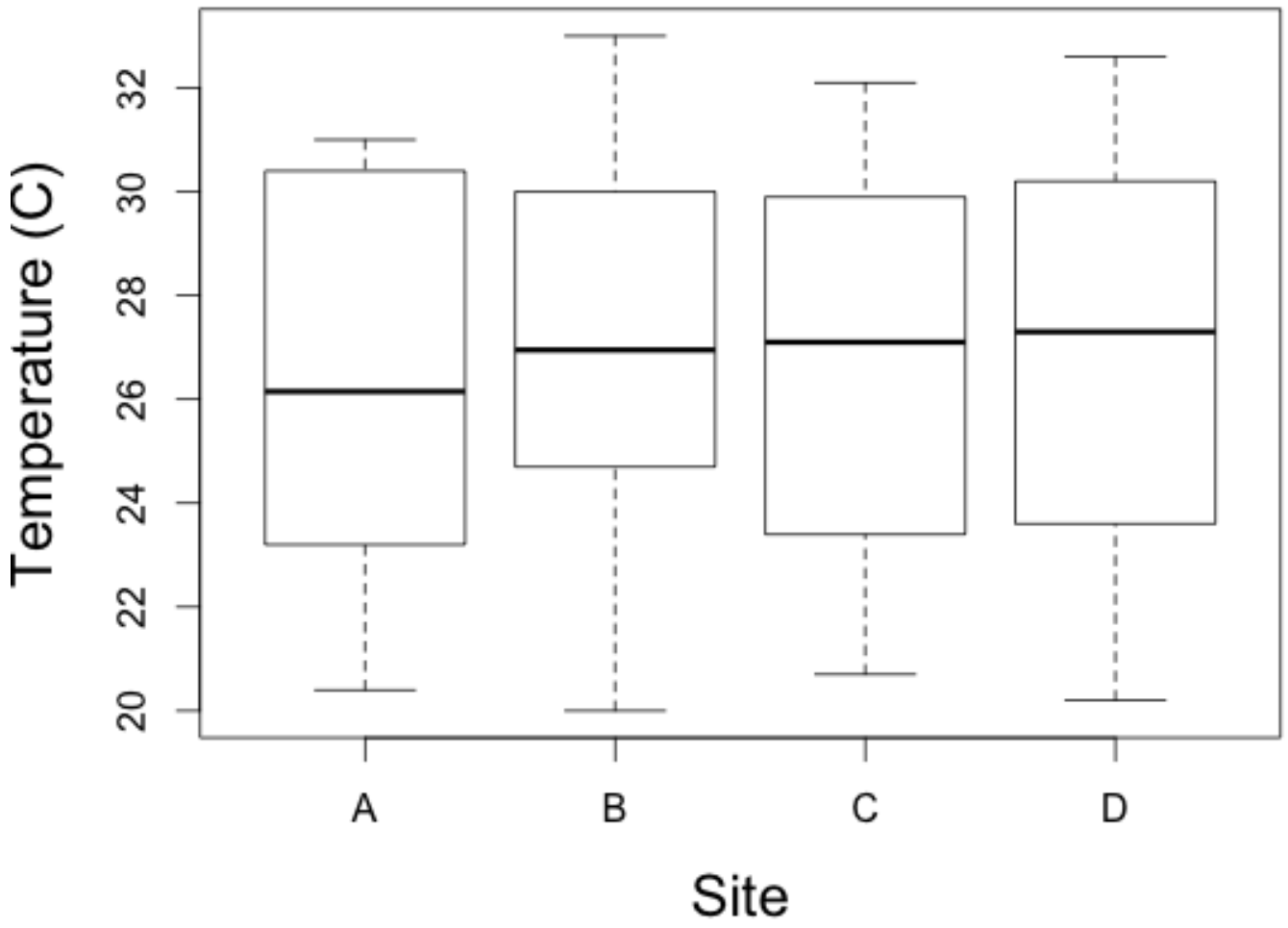
What about other variables?

- Time



What about other variables?

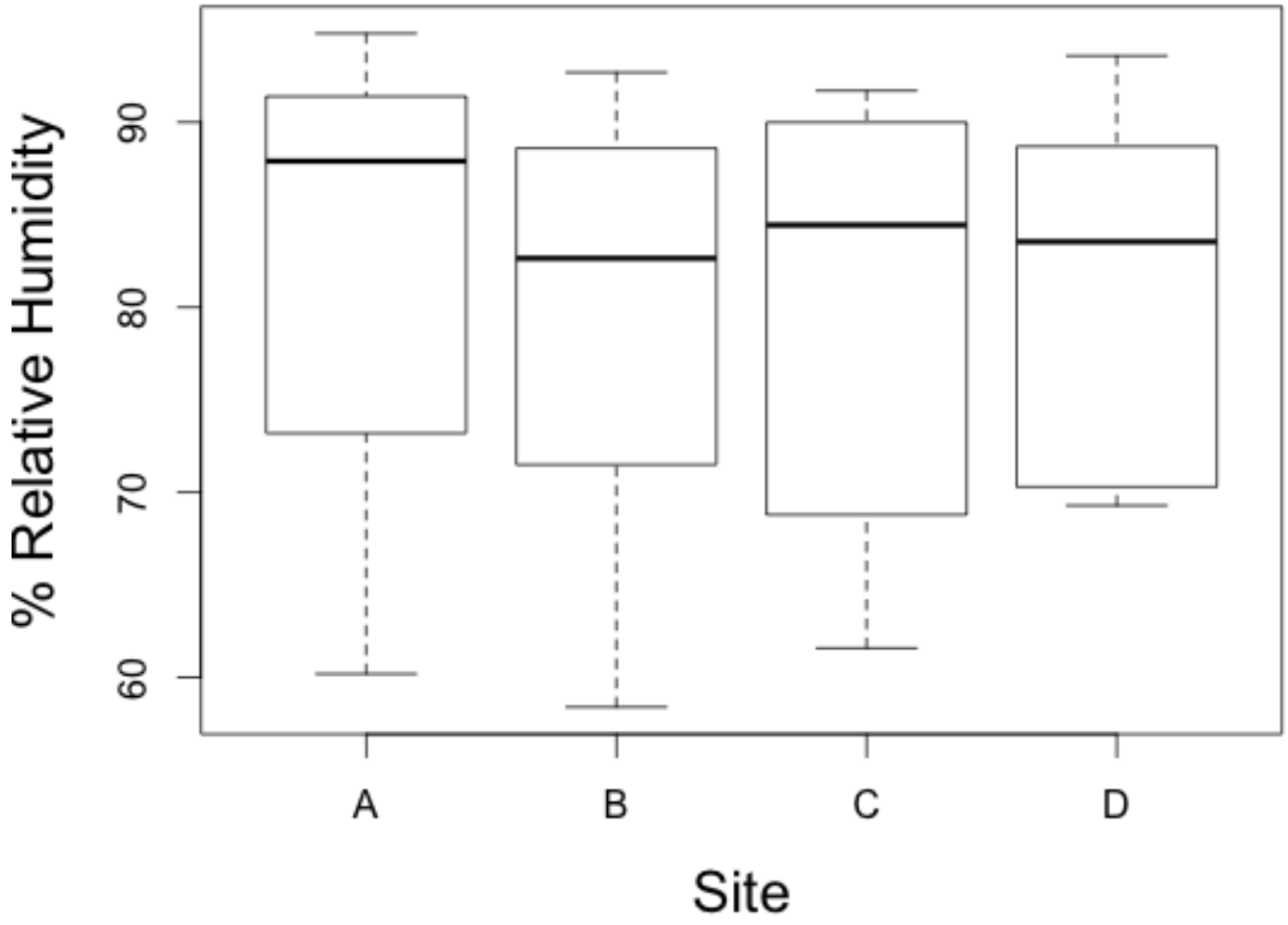
- location



Use ANOVA to quantitatively analyze these type of data.

What about other variables?

- location



Use ANOVA to quantitatively analyze these type of data.

Multiple Comparisons

- p-values tell you the probability of getting a **false positive** due to random chance
- The more times you roll a die, the more likely you are to have rolled a 6.
- The more times you look for significant relationships, the more likely you are to accidentally find false positives.

Multiple Comparisons

- Don't "fish" in your data set!
- Plan statistical analysis before gathering data.
- If multiple comparisons must happen, control for them!

See "Multiple Comparisons Concepts" in *Intuitive Biostatistics* (2010) H. Motulsky, Oxford University Press

Where do we go from here?

- Design experiments to take advantage of linear regression.
- If temperature and RH are part of a larger experiment, think about their **colinearity**.
- Determine how temporal data should be handled.

Correlation: Assumptions

- Random samples from a single population
- Paired samples (X,Y)
- Independent observations
- Y values were not computed from X values
- X values were not experimentally manipulated
- X and Y are normally distributed
- All variation is linear
- No outliers

Adapted from *Intuitive Biostatistics* (2010) H. Motulsky, Oxford University Press

Regression: Assumptions

- The model (relationship between X and Y) is correct
- Y values are normally distributed
- Variability is constant for all values of X
- Independent observations
- Y values are not computed from X values
- X values are known precisely (X is manipulated, not measured; Y is measured)

Adapted from *Intuitive Biostatistics* (2010) H. Motulsky, Oxford University Press