Regression

Lecture #21 | GEOG 510
GIS & Spatial Analysis in Public Health

Varun Goel

Outline

- Correlation
- Regression
- Confounding

Correlation

- Correlation
  - The relationship between things that happen or change together
    - Other terms used: relationship, association
  - Using statistical techniques, we can measure and test the relationship between things

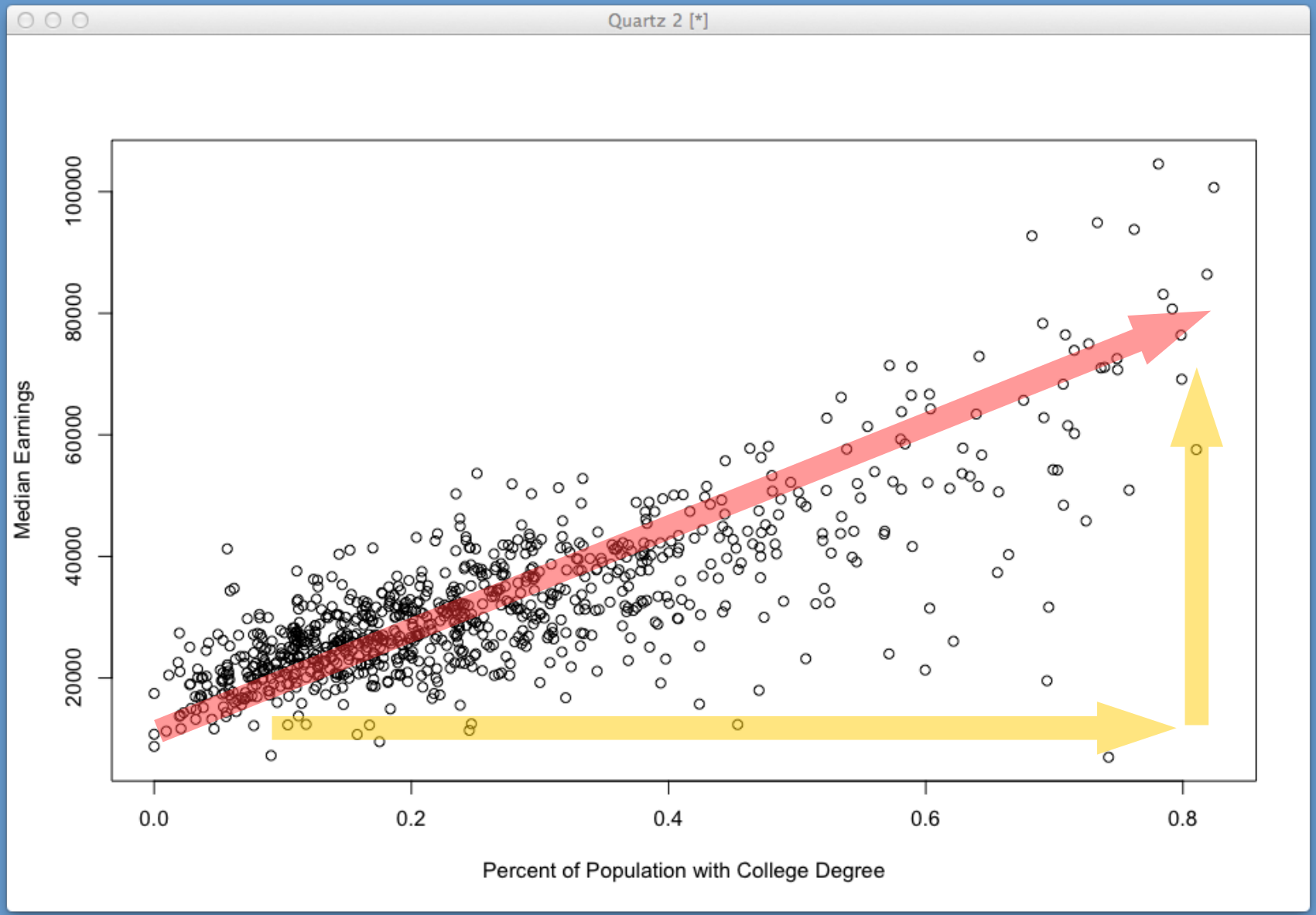
Correlation

- Correlation concepts
  - Direction
    - What is the overall trend of the relationship between the two variables?
  - Strength
    - How closely related are the two variables?
  - Form
    - What is the "shape" of the relationship?

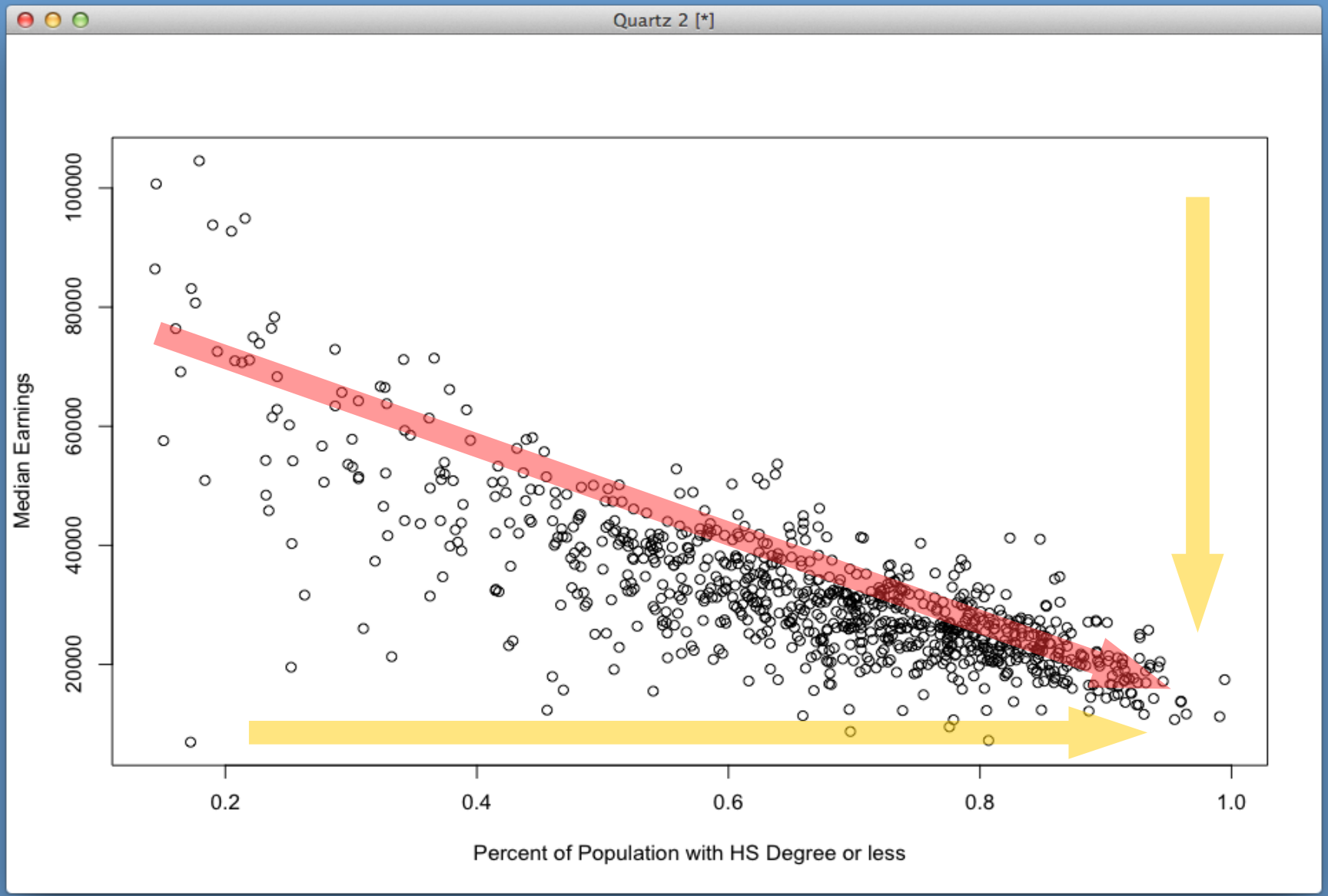
Direction

- The direction of the relationship or association between two variables, or their correlation, can be:
  - Positive
    - As variable X increases, variable Y increases
      - As variable X decreases, variable Y decreases
  - Negative
    - As variable X increases, variable Y decreases
      - As variable X decreases, variable Y increases
  - Neutral (random)
    - No systematic relationship between variables

Direction


Quartz 2 [*]
Median Earnings
20000
40000
60000
80000
100000
0.0
0.2
0.4
0.6
0.8
Percent of Population with College Degree

Direction

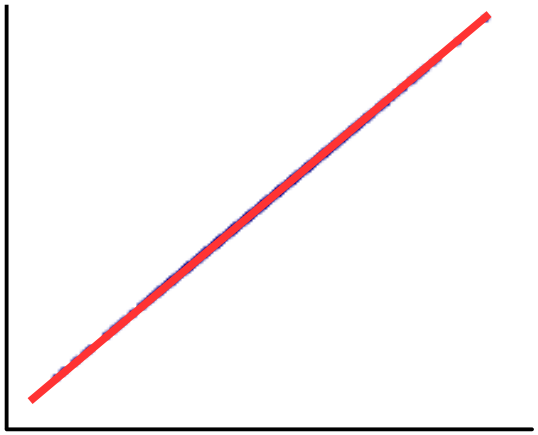

Quartz 2 [*]
Median Earnings
20000
40000
60000
80000
100000
0.2
0.4
0.6
0.8
1.0
Percent of Population with HS Degree or less

Strength of Association

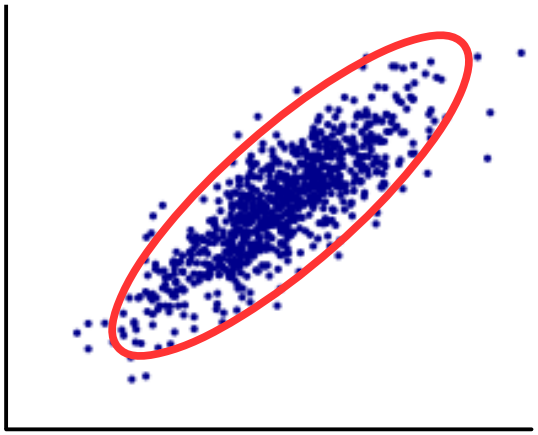
- Strength of association
  - Strength of the association is the consistency of the variation in the two variables, over the range of both
    - Can be thought of as the "precision" of the correlation
  - Can be visually estimated by viewing the scatterplot of the two variables

Strength of Association

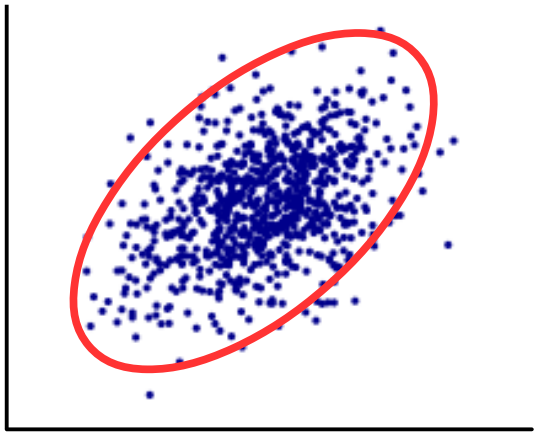
The strength is...



Perfect



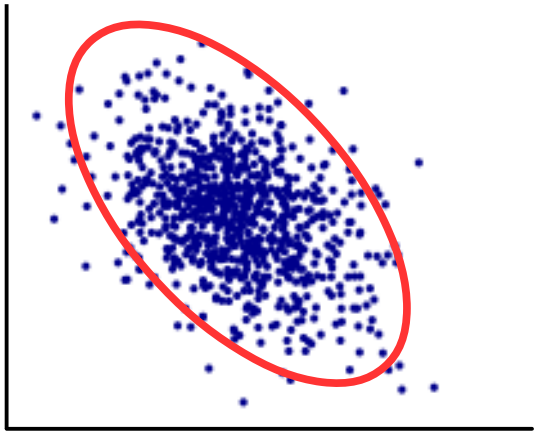
Strong



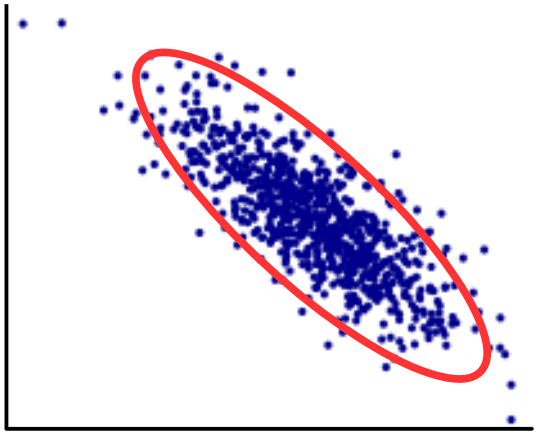
Moderate

Strength of Association

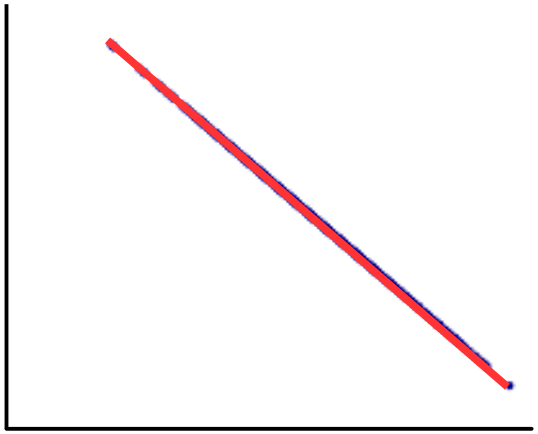
The strength is...



Moderate



Strong

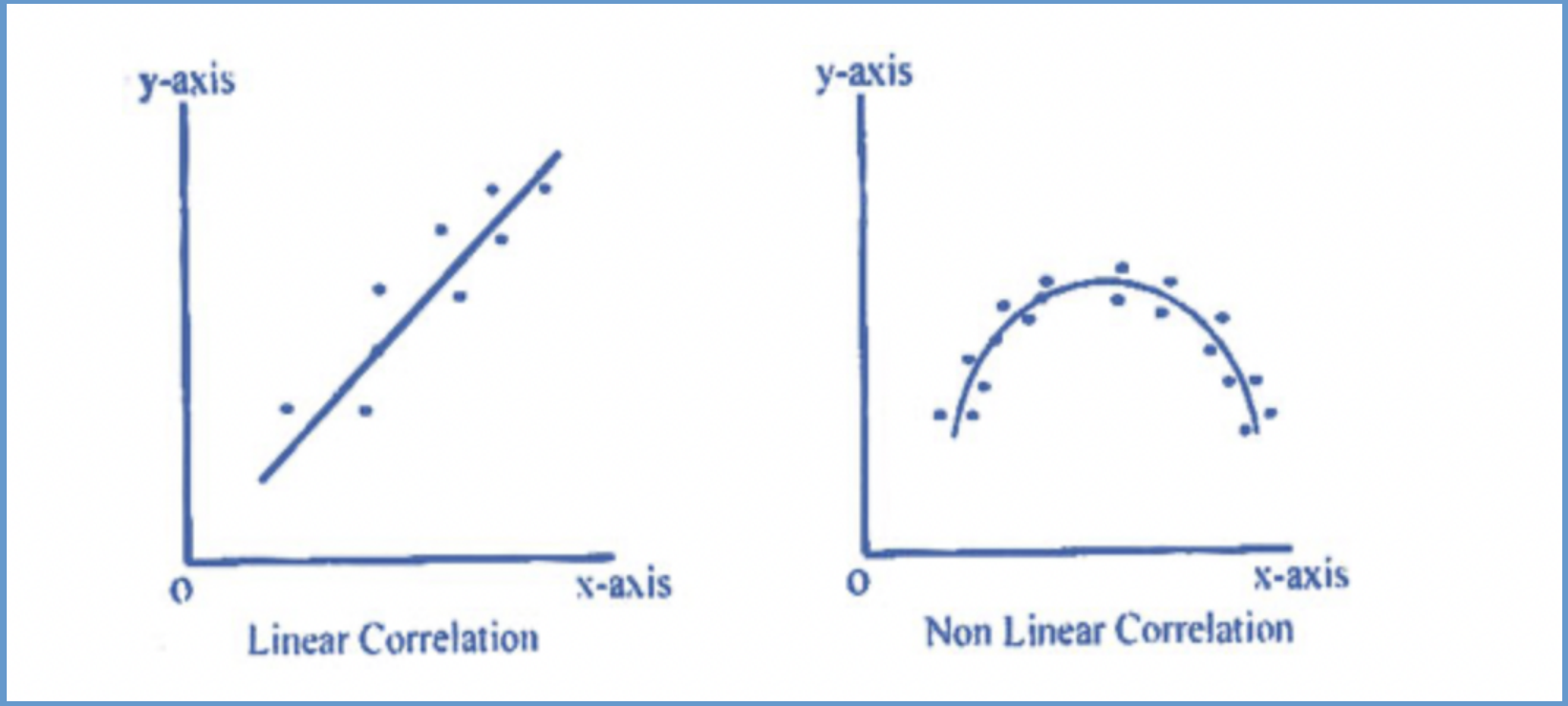


Perfect

Relationship Form

- Correlation can take many forms
  - The relationship between two variables is not bound to "straight lines"
    - We can estimate the relationship via the scatterplot of the variables
- Linear
  - The relationship can be estimated using a straight line
- Non-linear
  - The relationship cannot be estimated using a straight line

Relationship Form



Relationship Form

- What if the relationship is non-linear?
  - Transform the data
    - Common transformation
      - Logarithms (natural log and base 10 log)
  - After data is transformed, use linear techniques
    - Sometimes referred to as a "log-linear" relationship

Variable Relationships

- Any two variables may be correlated
  - However, in some cases, we expect that a functional relationship exists
  - Two variables may be correlated, but this does not prove a causal relationship
    - Causal relationships are much more difficult to prove and validate

Variable Relationships

- Any two variables may be correlated
  - However, in some cases, we expect that a functional relationship exists
  - Two variables may be correlated, but this does not prove a causal relationship
    - Causal relationships are much more difficult to prove and validate

Bivariate Correlation

- Pearson's r
  - r values range from -1 to 1
    - Provides information about *direction* and *strength* of the correlation (association) between variables
      - -1 is perfect, negative correlation
      - 0 is no correlation
      - 1 is perfect, positive correlation

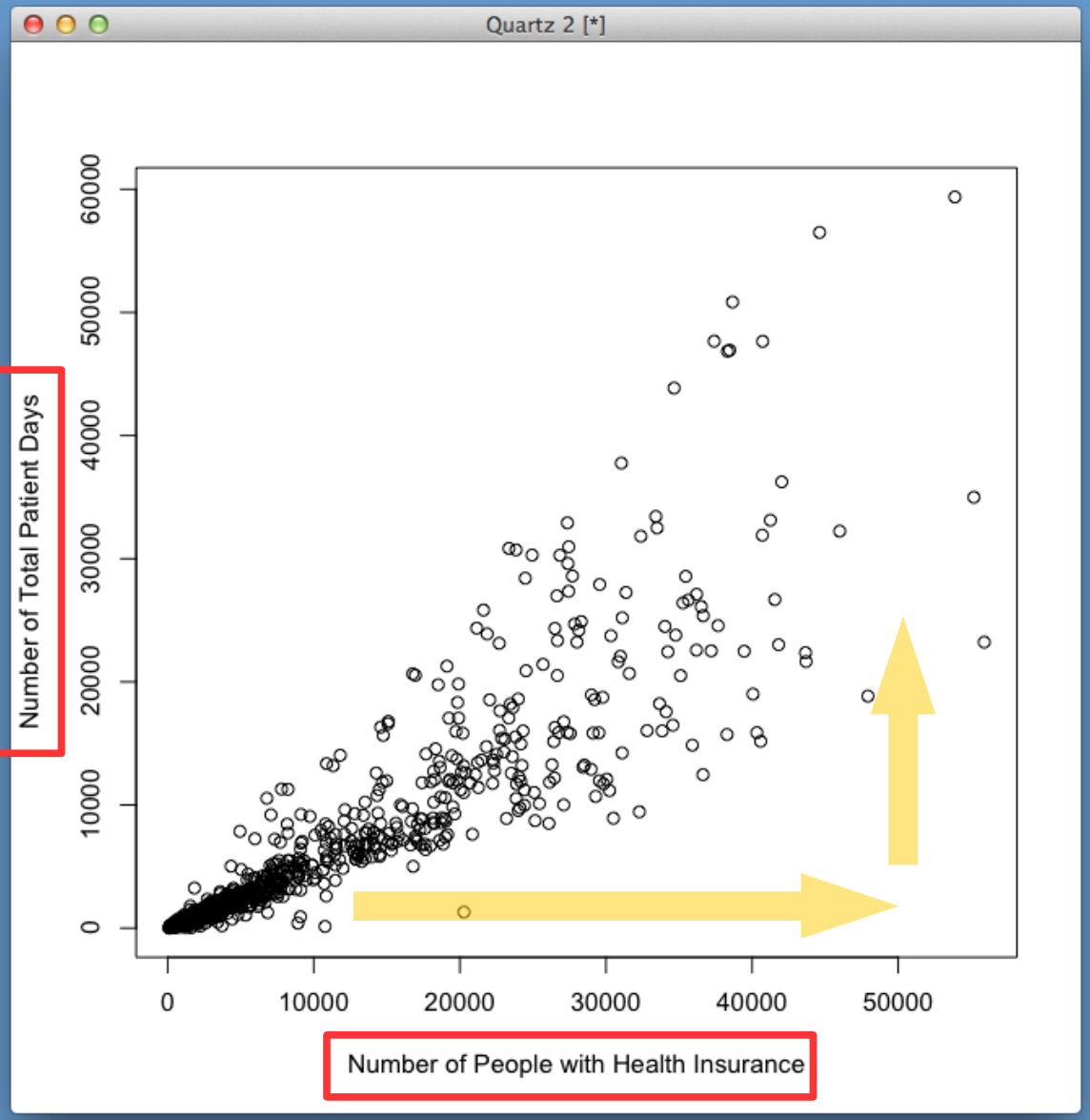
Bivariate Correlation

- Spearman's rs
  - Very similar to Pearson's r
  - Non-parametric version
    - For ordinal data (strongly ordered)
    - For interval/ratio data… if the data are severely non-normal
      - The interval/ratio data must first be converted to ordinal data
  - rs values range from -1 to 1

Aggregated Data

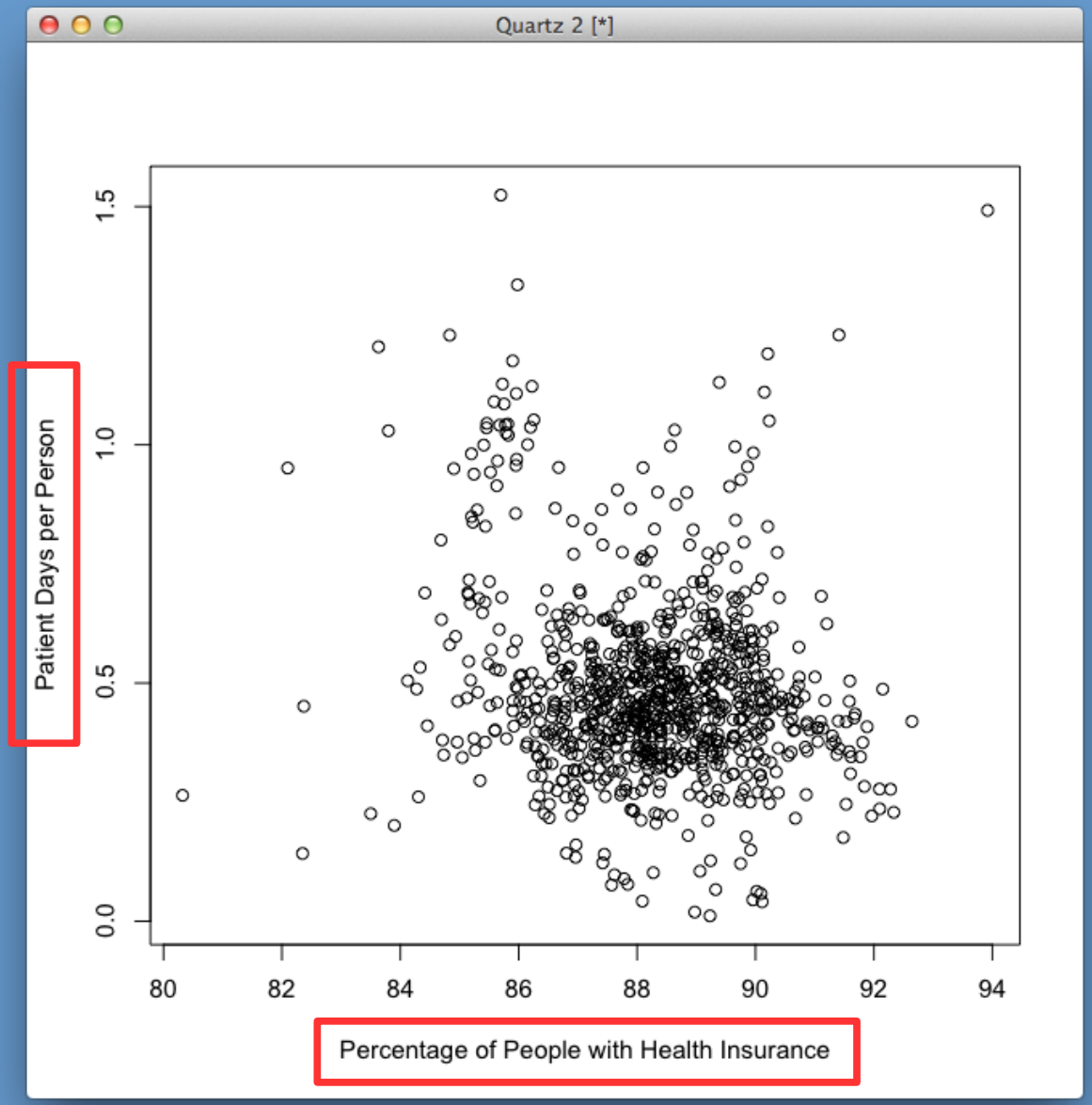
- Aggregated observation units
  - Example: Patient days and Health insurance

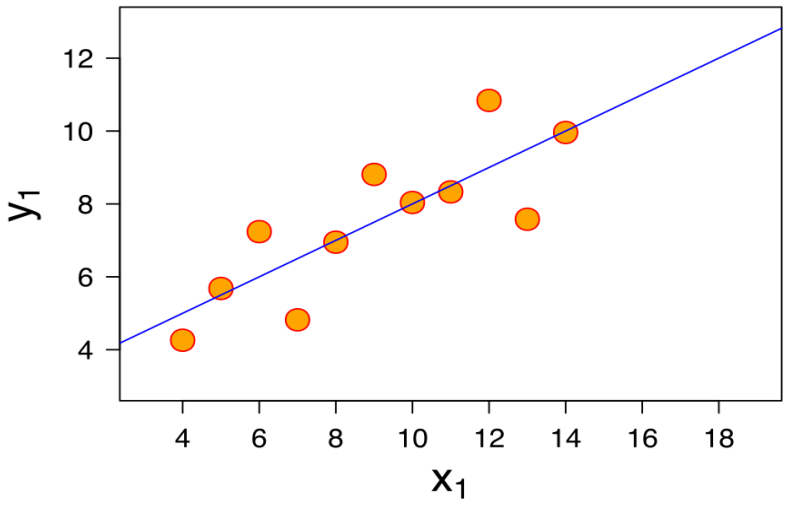
The "positive" relationship is nothing but a function of the number of people in each unit!

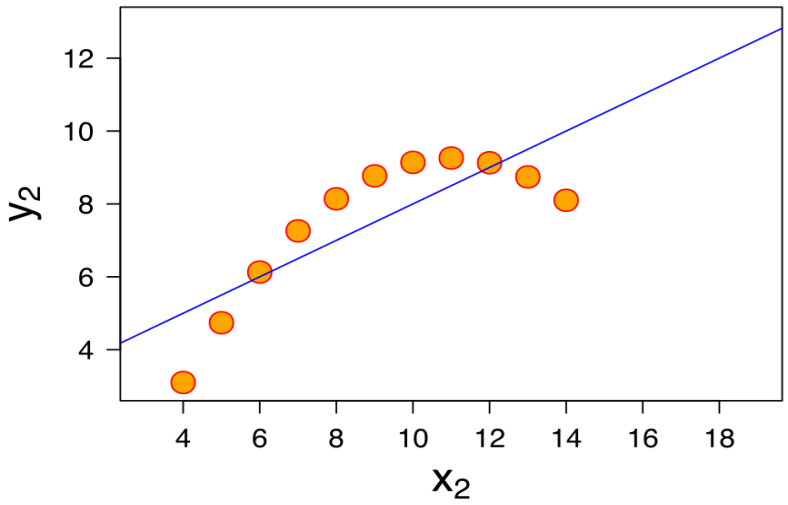


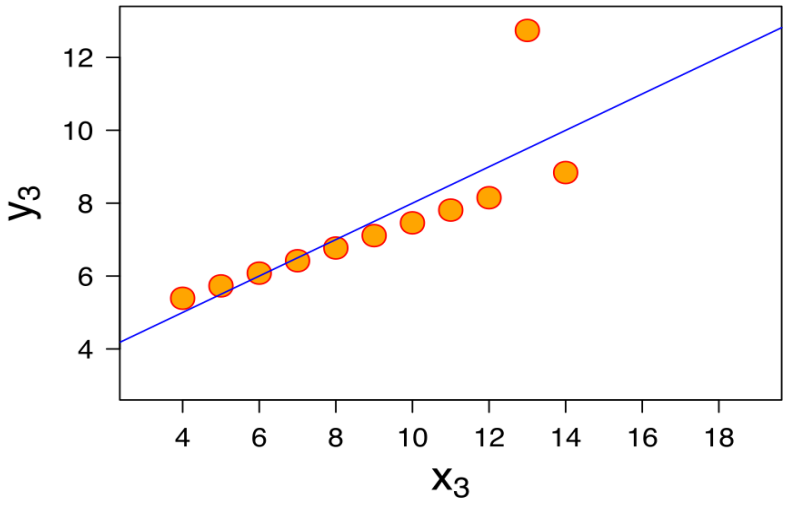
Aggregated Data

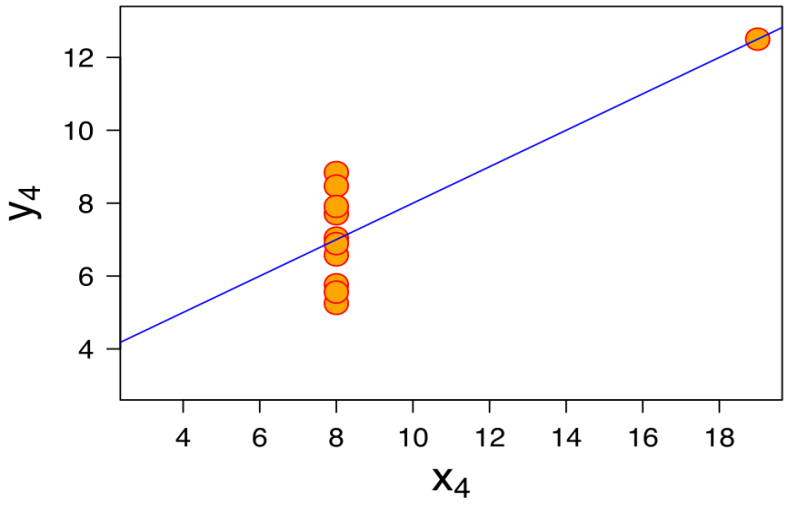
- Aggregated observation units
  - Example: Patient days and Health insurance
    - Divide by population size of each unit




12
10
8
6
4
y_1
4
6
8
10
12
14
16
18
x_1


12
10
8
6
4
y_2
4
6
8
10
12
14
16
18
x_2


12
10
8
6
4
y_3
4
6
8
10
12
14
16
18
x_3


12
10
8
6
4
y_4
4
6
8
10
12
14
16
18
x_4

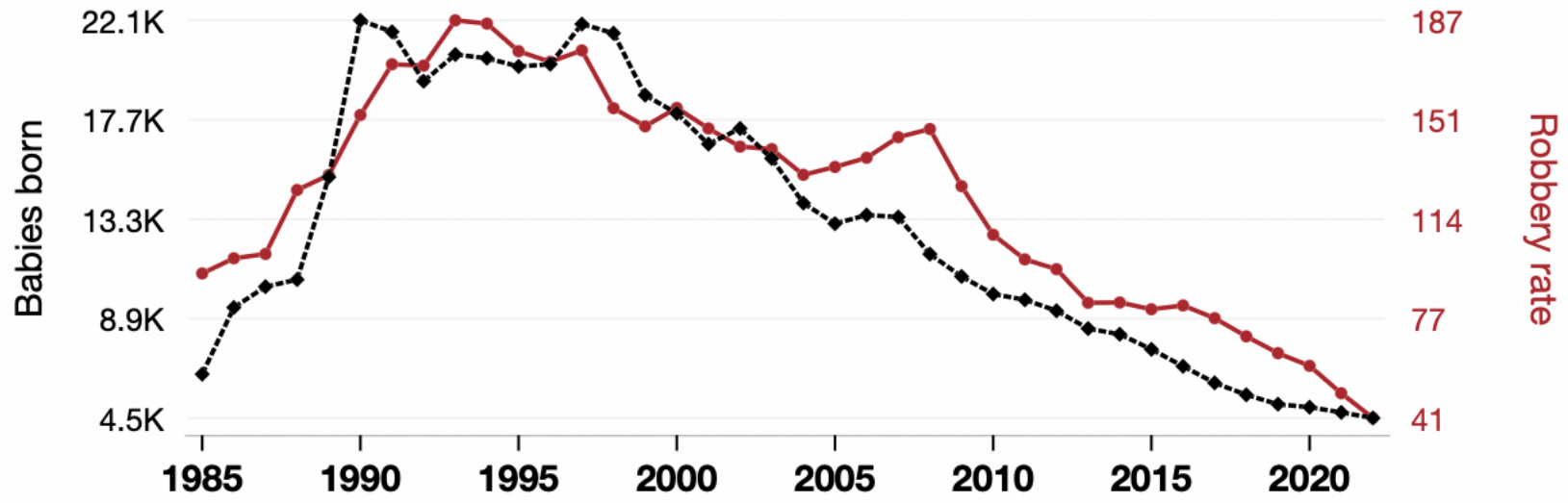
Regression

- What correlation doesn't tell us...
  - The magnitude of the influence one variable has on the other
- Regression requires that a functional relationship exists between variables
  - If there is **no clear functional relationship**, then **regression should not be used**
  - Also, the causal direction of the relationship between variables is extremely important
  - X causes Y not equal to Y causes X

Popularity of the first name Jordan

correlates with

Robberies in South Carolina



Babies of all sexes born in the US named Jordan · Source: US Social Security Administration

The robbery rate per 100,000 residents in South Carolina · Source: FBI Criminal Justice Information Services

1985-2022, r=0.947, r^2=0.897, p<0.01 · tylervigen.com/spurious/correlation/1081

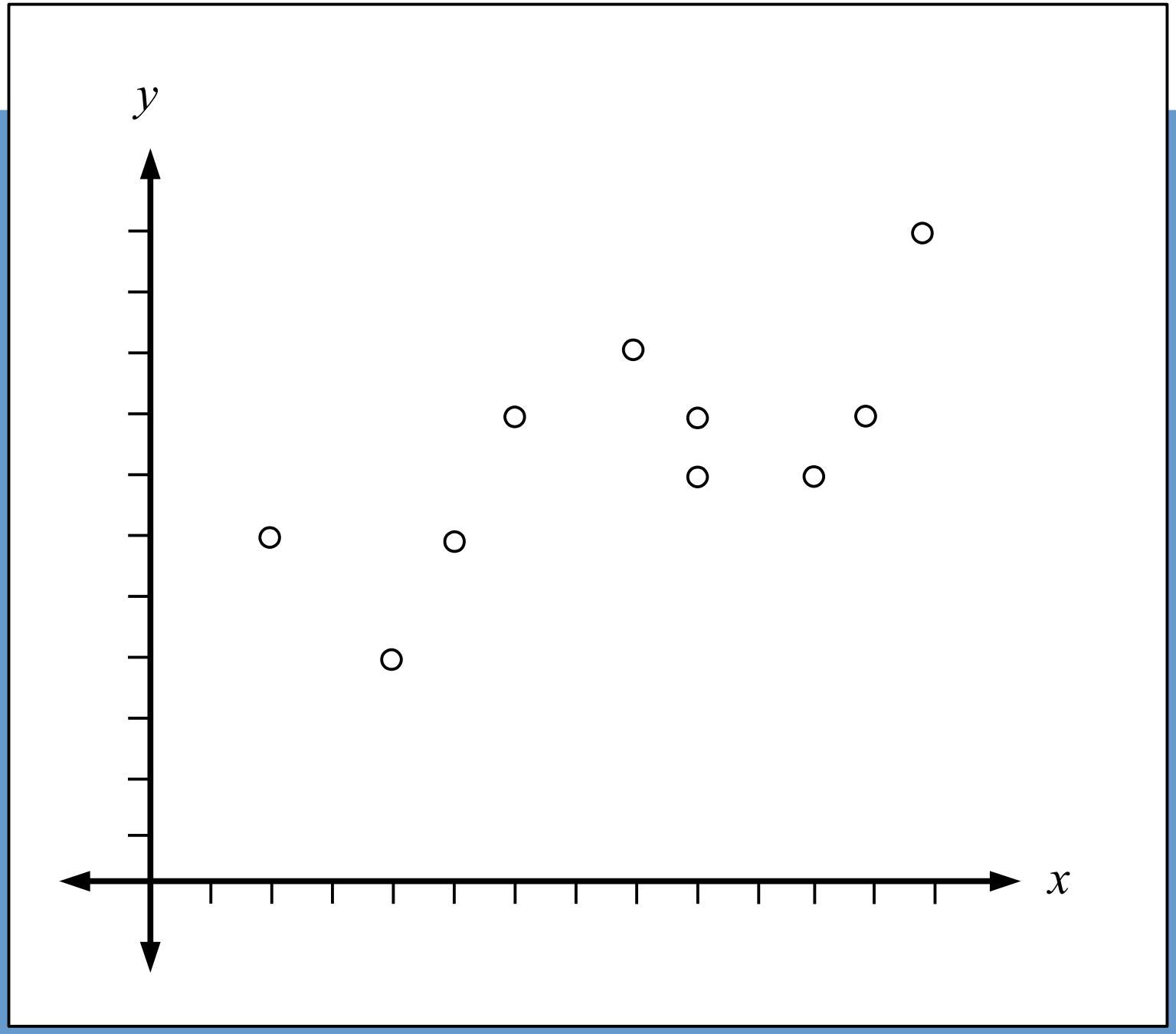
View details about correlation #1,081

Explanatory vs. Predictive

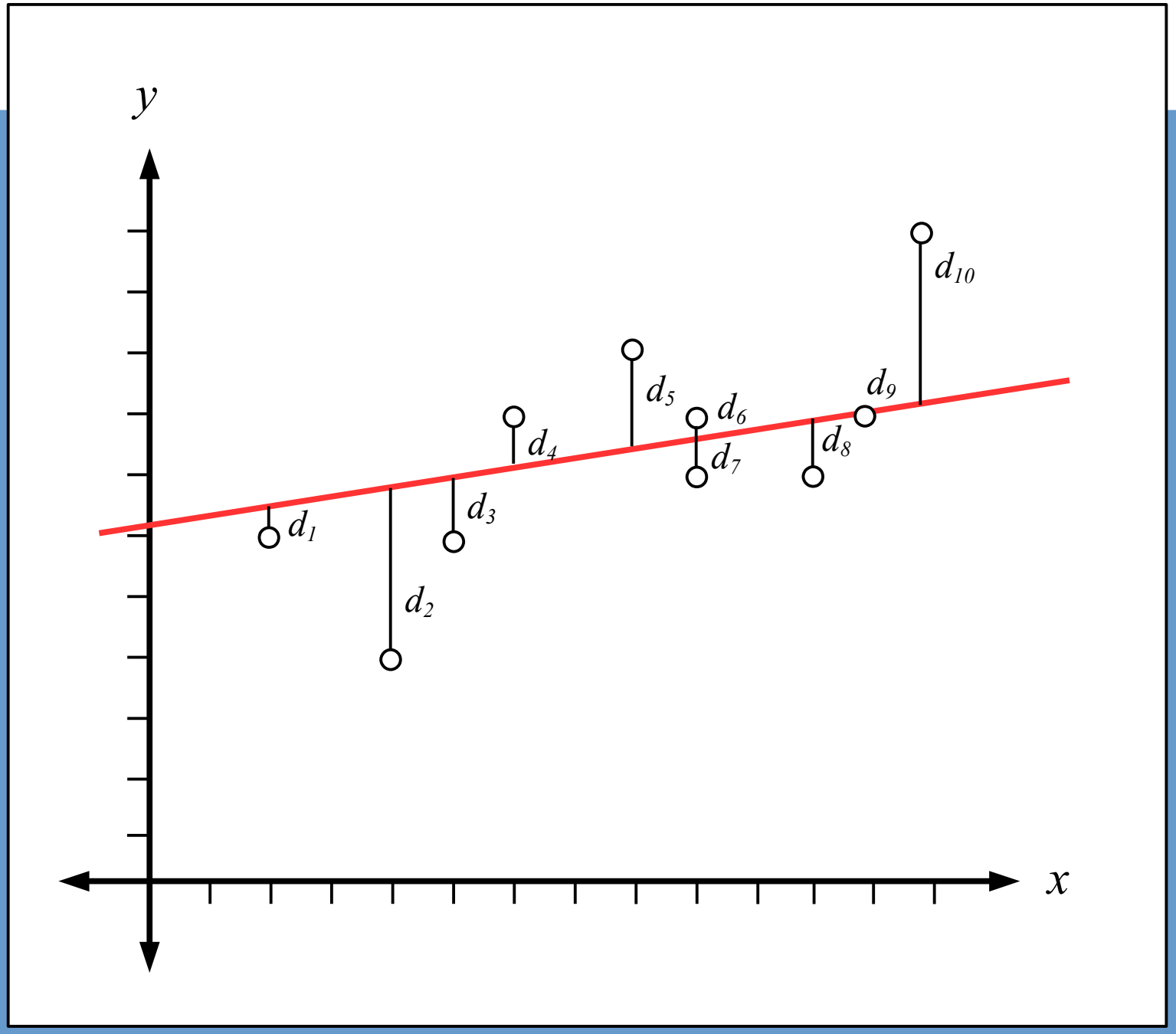
- Explanatory model
  - Used when we are trying to make inference about relationships among variables
    - This is how I generally use regression
- Predictive model
  - Use regression to "most accurately" predict unknown values of Y

Bivariate Regression

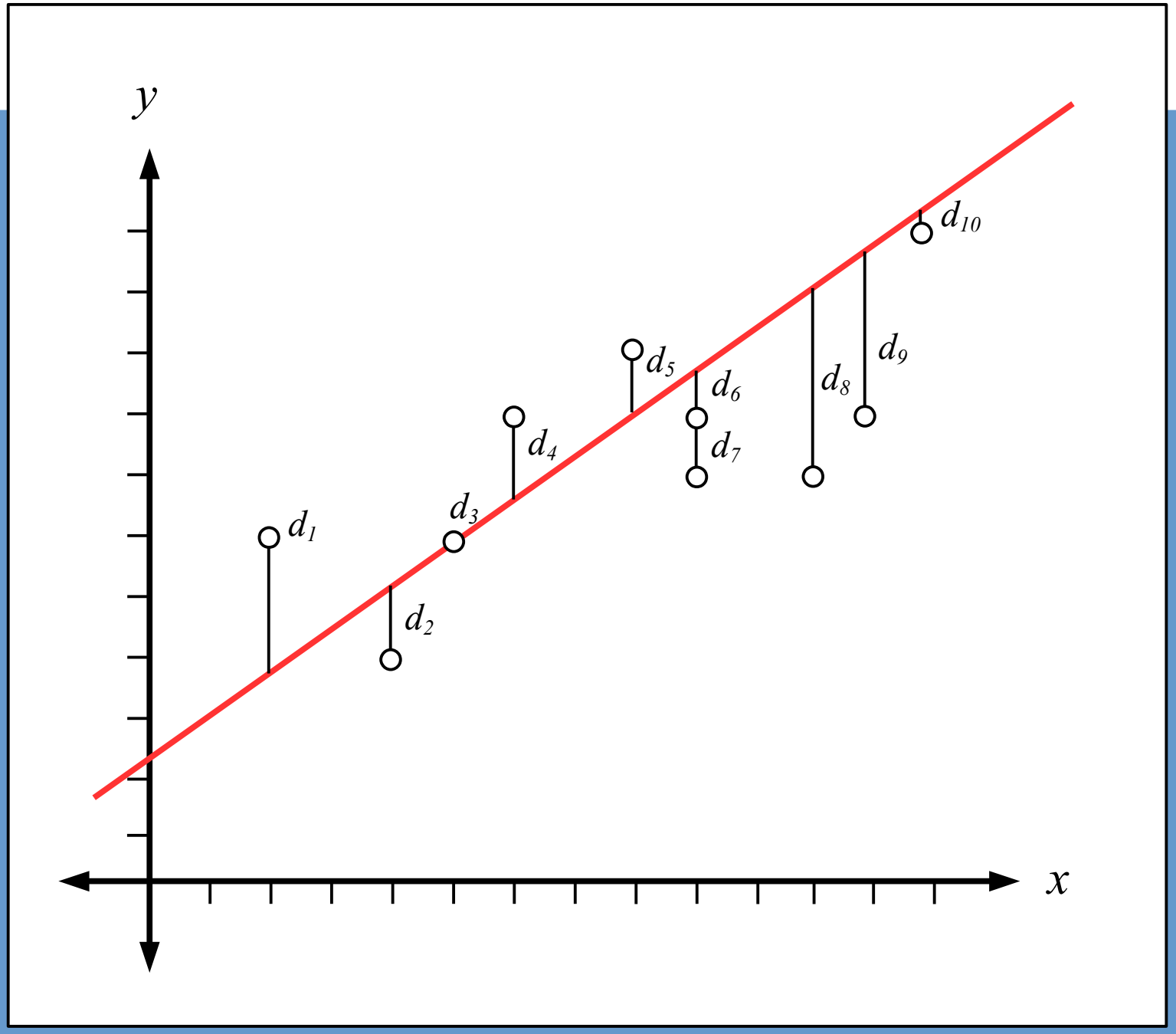
- Simple linear regression
  - Linear relationship between variables
    - Defined as $Y = a + bX$
  - Fits the regression line through the observed X, Y data
    - Ordinary Least Squares (OLS)
      - Minimizes the squared deviations from the observed Y values to the regression line



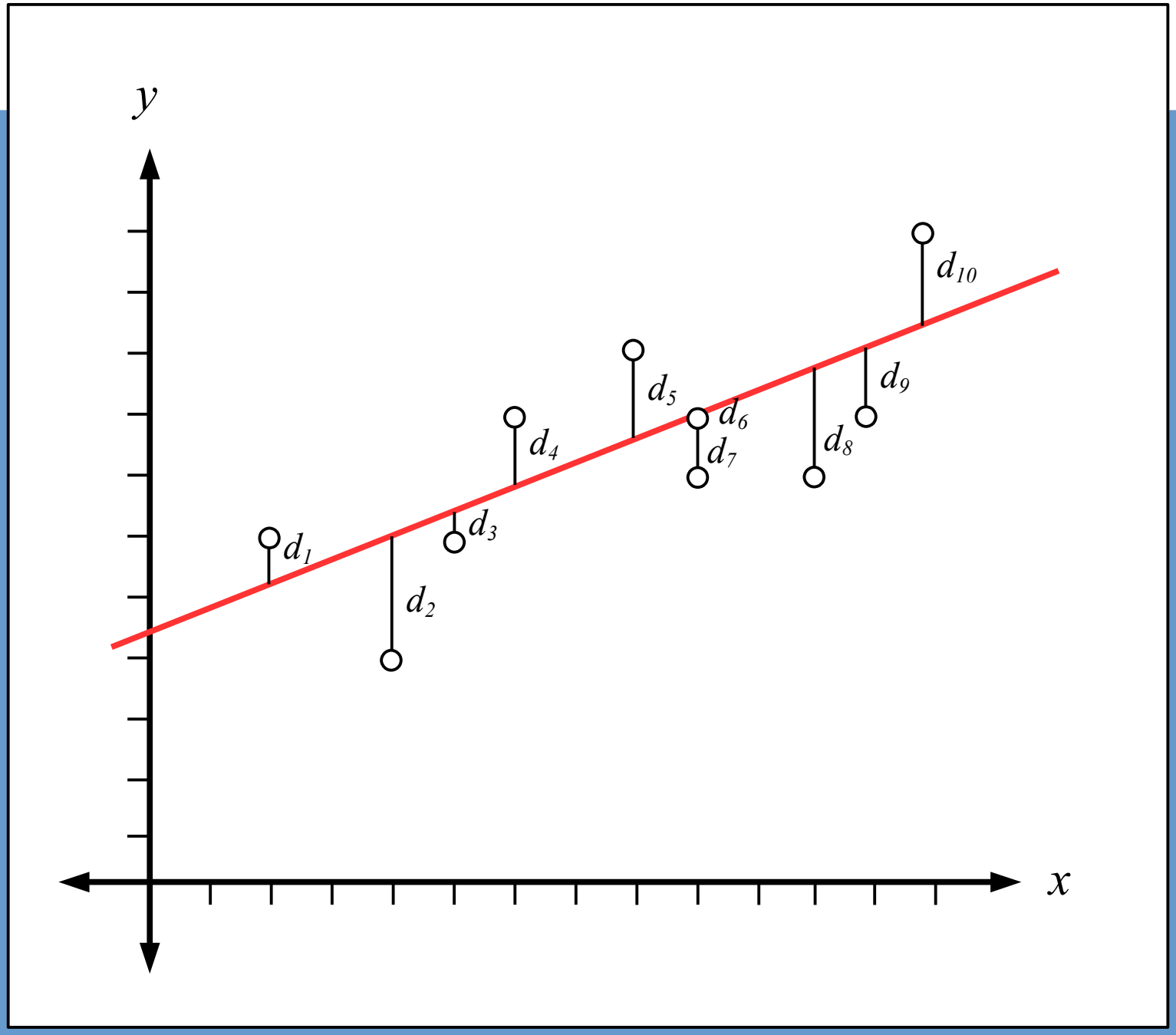
Point	*x*	*y*
1	*2*	*6*
2	*4*	*4*
3	*5*	*6*
4	*6*	*8*
5	*8*	*9*
6	*9*	*8*
7	*9*	*7*
8	*11*	*7*
9	*12*	*8*
10	*13*	*10*



Point	*x*	*y*
1	*2*	*6*
2	*4*	*4*
3	*5*	*6*
4	*6*	*8*
5	*8*	*9*
6	*9*	*8*
7	*9*	*7*
8	*11*	*7*
9	*12*	*8*
10	*13*	*10*



Point	*x*	*y*
1	*2*	*6*
2	*4*	*4*
3	*5*	*6*
4	*6*	*8*
5	*8*	*9*
6	*9*	*8*
7	*9*	*7*
8	*11*	*7*
9	*12*	*8*
10	*13*	*10*



Point	*x*	*y*
1	*2*	*6*
2	*4*	*4*
3	*5*	*6*
4	*6*	*8*
5	*8*	*9*
6	*9*	*8*
7	*9*	*7*
8	*11*	*7*
9	*12*	*8*
10	*13*	*10*

Bivariate OLS Regression

$$Y = \beta_0 + \beta X + \epsilon \quad \rightarrow \quad r^2$$

- Effects of X on Y
  - Regression parameters
    - Slope
    - $H_0 : \beta = 0$ $H_A : \beta \neq 0$

Bivariate OLS Regression

$$Y = \beta_0 + \beta X + \epsilon \quad \rightarrow \quad r^2$$

- Effects of X on Y
  - Uses a *t* test, which is based on the size of β and the number of observations (*n*)
    - IMPORTANT! We can have a scenario in which X may have a small, yet statistically significant affect on Y (e.g., small β, high *n*)
      - Like correlation and spatial autocorrelation… the analyst must interpret the difference between results that are "statistically significant" and results that are "important"

Bivariate OLS Regression

$$Y = \beta_0 + \beta X + \epsilon \quad \rightarrow \quad r^2$$

- Significance of Intercept parameter
  - Somewhat limited value
  - Can potentially provide interesting information… but generally, not evaluated

Bivariate OLS Regression

$$Y = \beta_0 + \beta X + \epsilon \quad \rightarrow \quad \boxed{r^2}$$

- Coefficient of Determination
  - The proportion of Y explained by X
  - $H_0 : r^2 = 0$ $\quad$ $H_A : r^2 \neq 0$
    - Uses an *F* test (*F* value and *df* generally reported)
    - If the *p*-value is low (e.g., $p < 0.05$), reject the null hypothesis

Multiple Regression

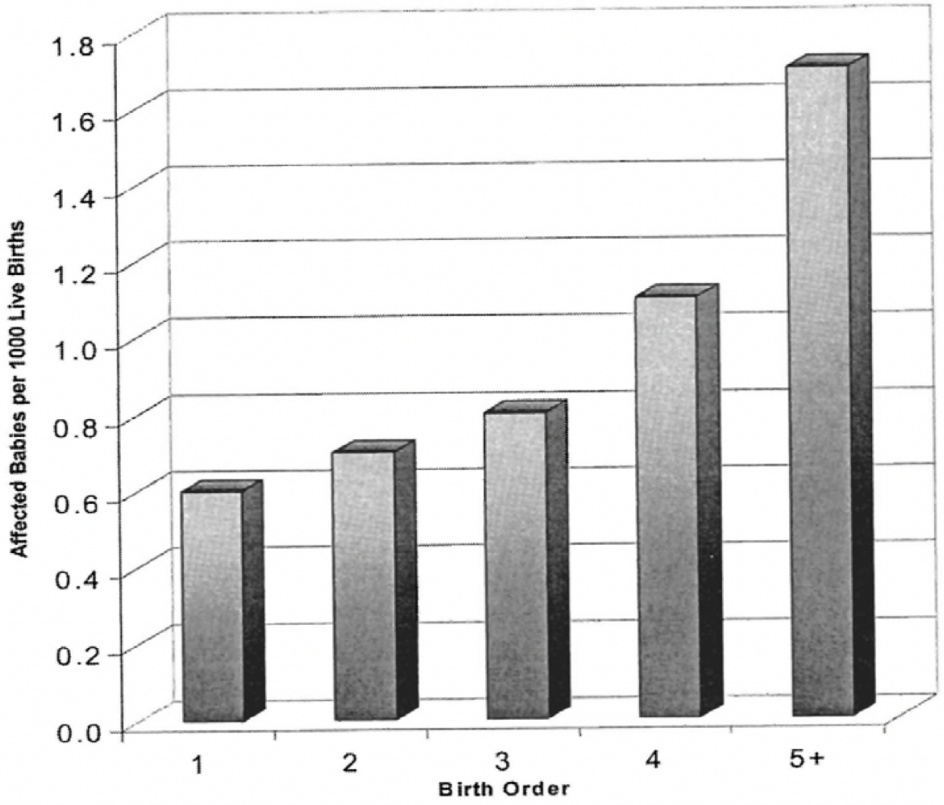
- Regression can be "extended" to include multiple independent variables
  - For many phenomena, a single explanatory variable does not provide sufficient characterization
    - Influenced by numerous factors
    - **CONFOUNDING!!!!**
  - More than one explanatory variable may be included in an OLS regression

Confounding

- BMJ Editorial: "The scandal of poor epidemiological research" [16 October 2004]
  - *"Confounding, the situation in which an apparent effect of an exposure on risk is explained by its association with other factors, is probably the most important cause of spurious associations in observational epidemiology."*

Confounding

Association between birth order and Down syndrome

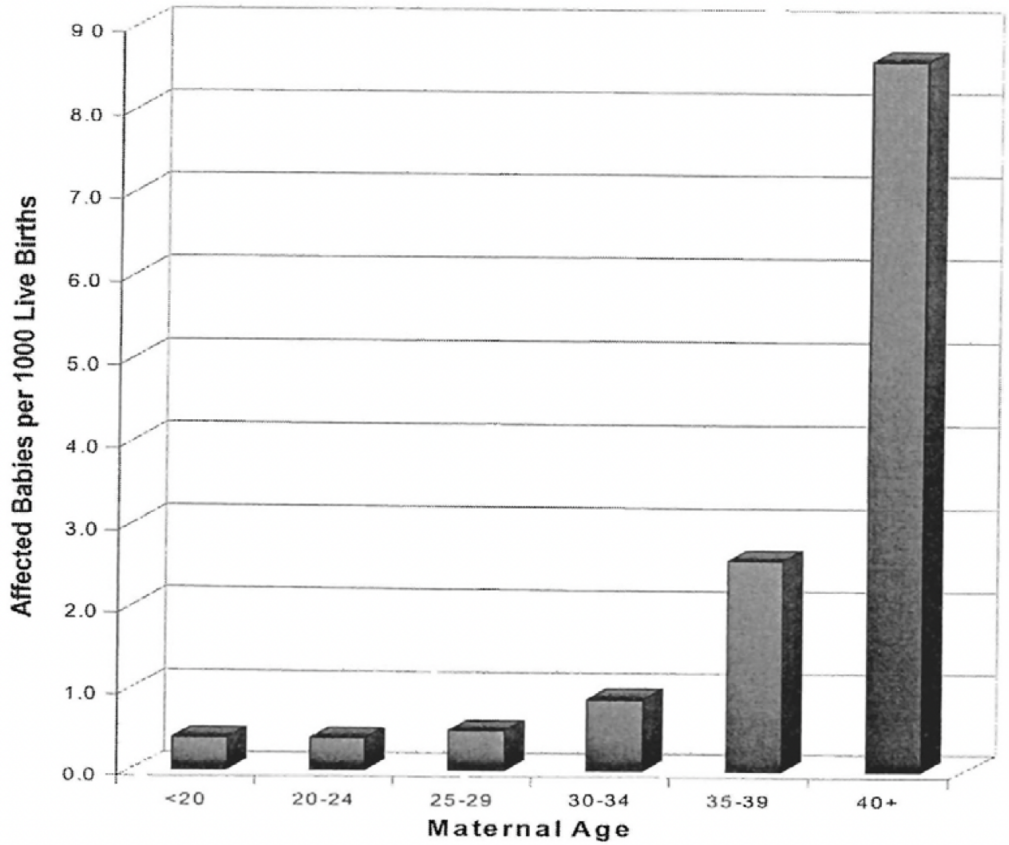


Data from Stark and Mantel (1966)

Source: Rothman 2002

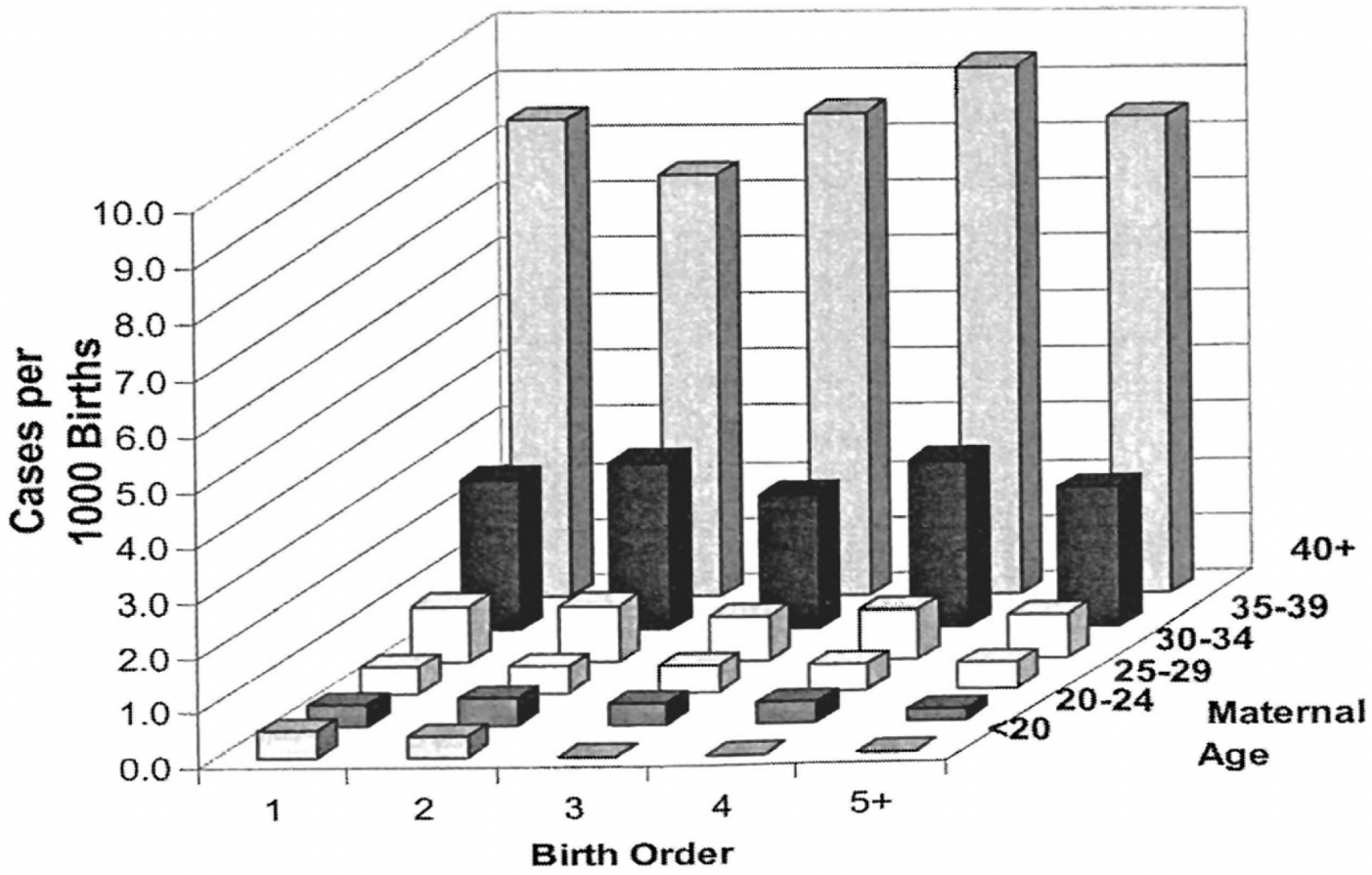
Confounding

Association between maternal age and Down syndrome

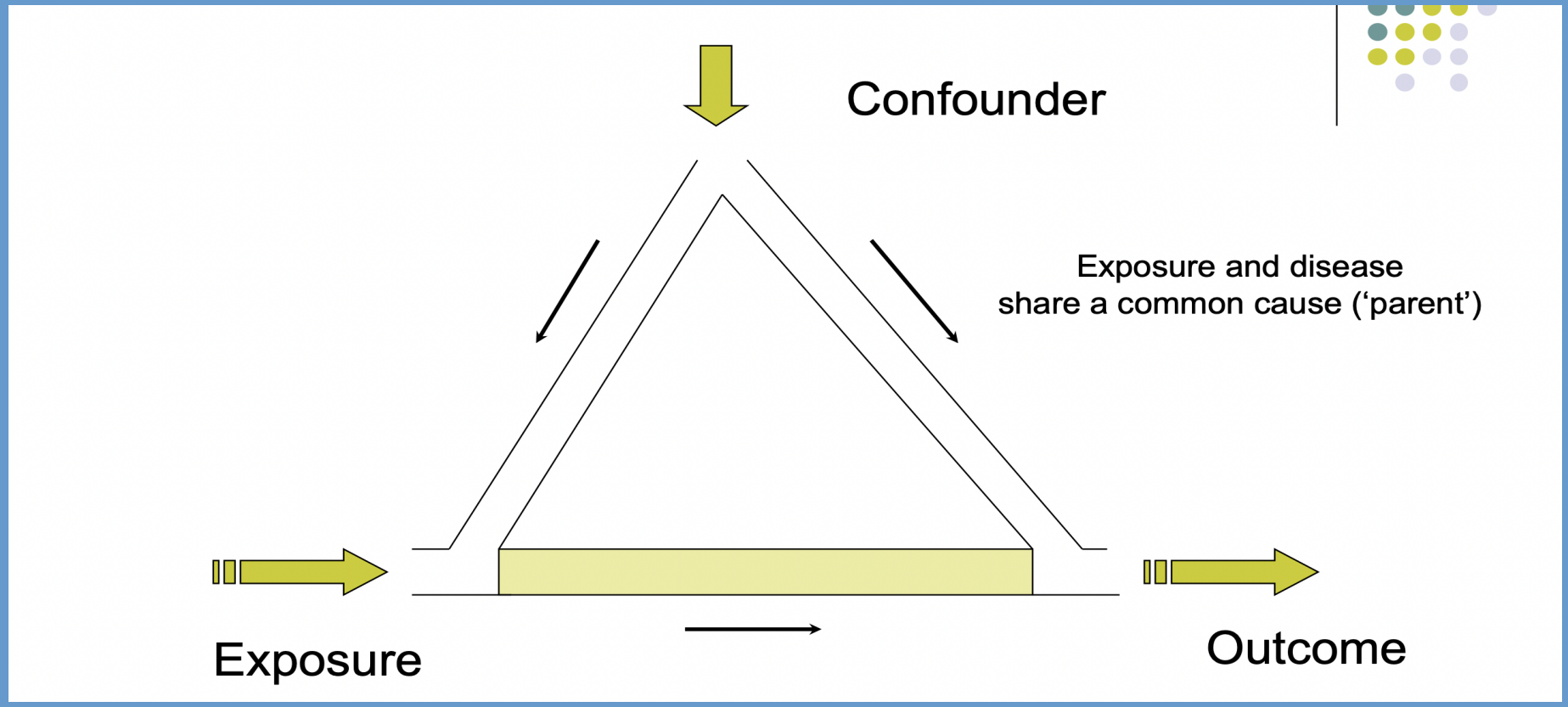


Confounding

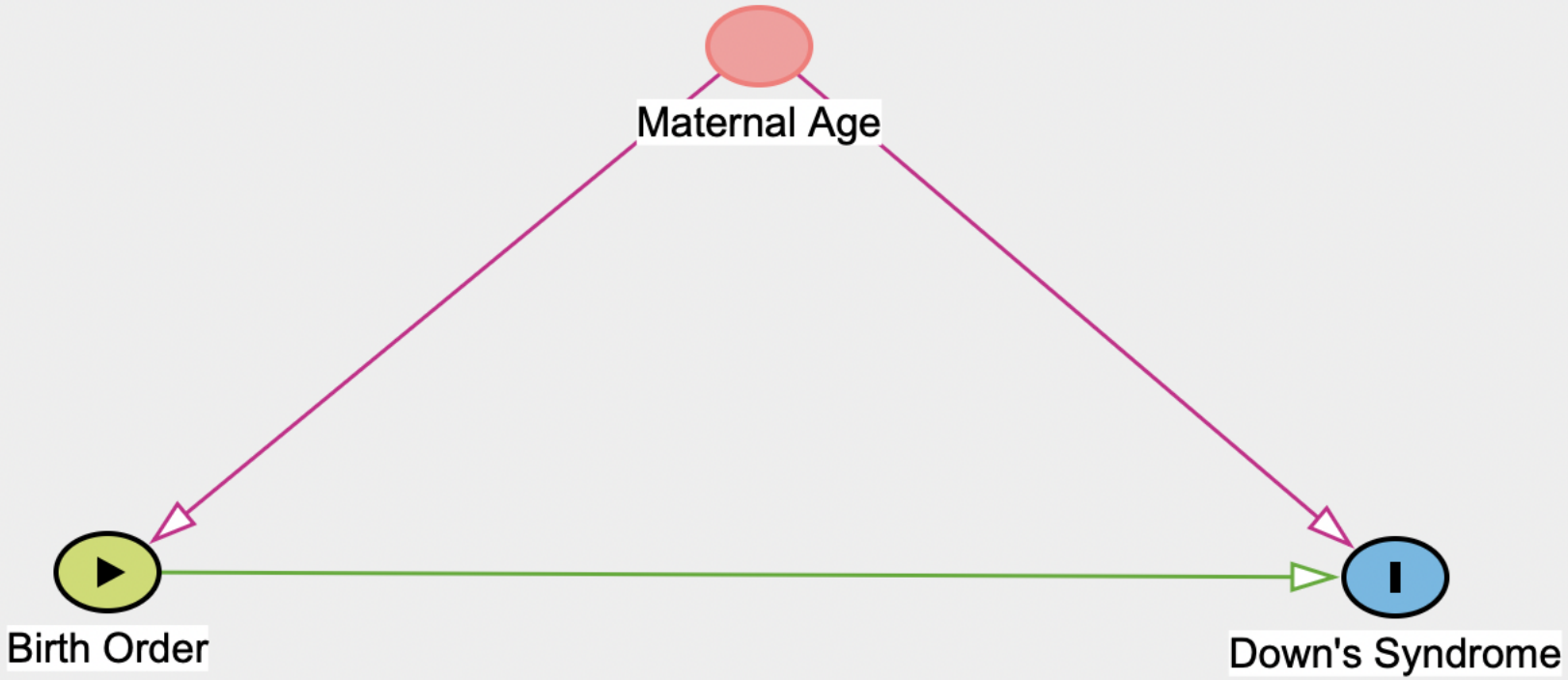
Association between maternal age and Down syndrome, stratified by birth order



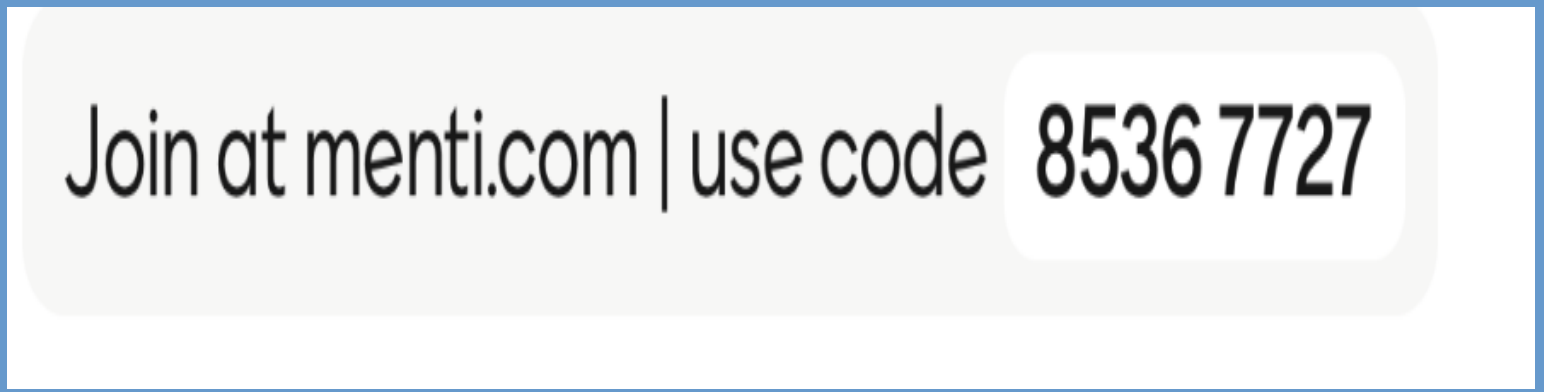
Confounding



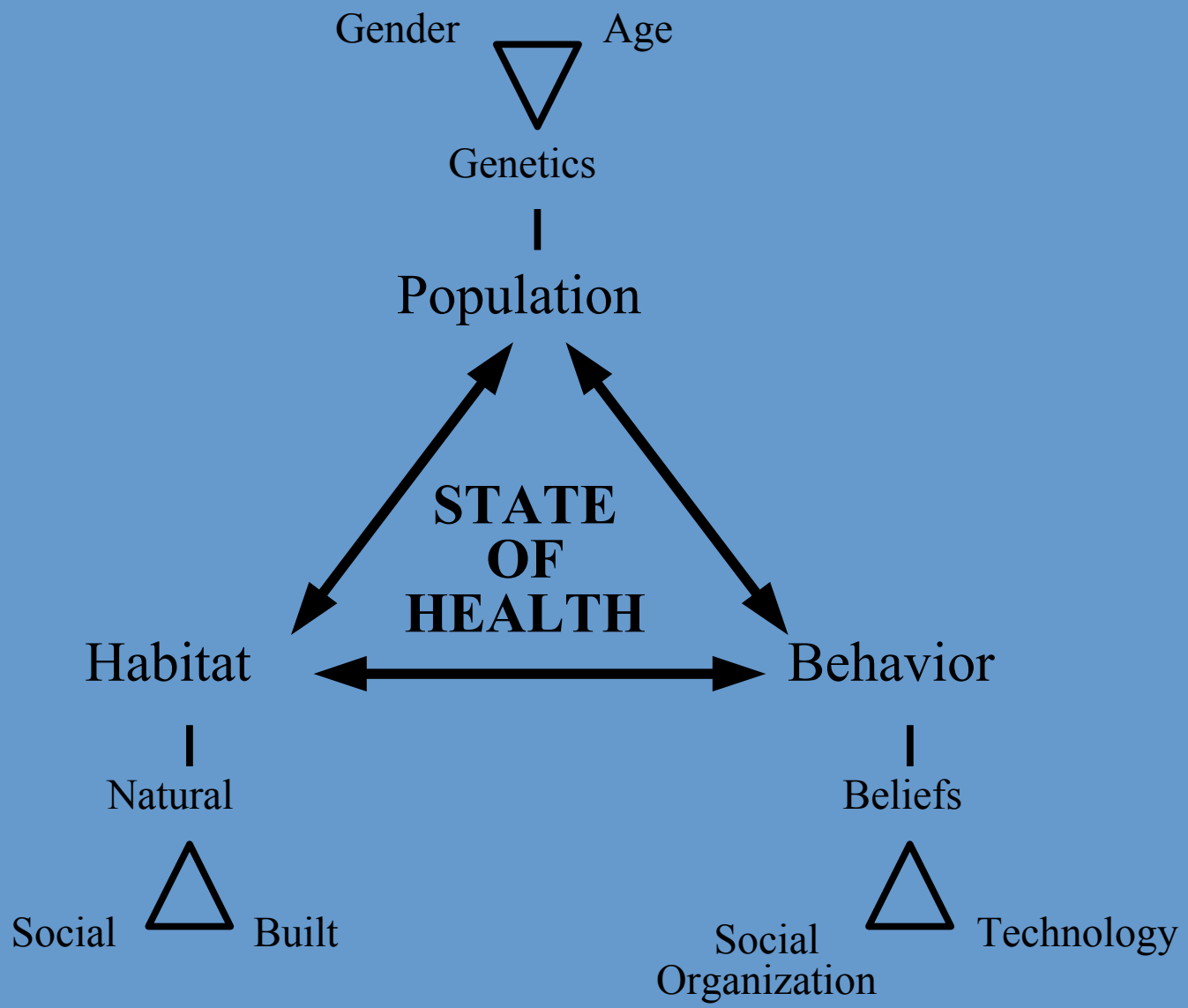
Confounding



Your Projects



Triangle of Human Ecology



Multiple Regression

- Caution, only use if…
  - There is a functional relationship with the additional predictor variable(s)/confounder
    - Do not simply "include" other variables because you can!
    - Theoretical justification for including each predictor variable is necessary

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$$

Multiple Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

- Inferential tests on the βs
  - Slope parameters on Xs
  - $H_0 : \beta_n = 0$ $H_A : \beta_n \neq 0$

Multiple Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \quad \rightarrow \quad R^2$$

- Regression coefficients
  - In multiple regression, these are sometimes referred to as "partial coefficients"
    - Because, theoretically, they will both explain a portion of the variation in the Y variable
    - These values, are generally conditional upon the other independent variables in the model
      - e.g., β_1 is the effect of X_1 on Y, when X_2 is held constant

Multiple Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \rightarrow \boxed{R^2}$$

- Coefficient of determination
  - The "fit" of the model
    - Proportion of the variation in Y that is explained by X_1 and X_2

Multiple Regression

- Via the β values, we can measure the effect of each X variable on Y
  - What if we want to compare the independent variables' effects?
    - e.g., which variable has a "stronger" effect on Y?
  - **β values are influenced by the units of the X variables**

Multiple Regression

- β values from a multiple regression **cannot** be compared directly
  - They must be "standardized"
    - Similar in concept to comparing the standard deviations from two different datasets
  - Standardize

$$\beta_k' = \beta_k \frac{s_{X_k}}{s_Y}$$

Multiple Regression

- Multiple regression assumptions
  - Sample with independent observations
  - Xs,Y are interval/ratio data
  - Linear relationship between Xs,Y
  - Independent variables are INDEPENDENT from one another
    - This is a huge issue in multiple regression

Multicollinearity

- Independent variables must be independent from one another
  - Means not correlated!
- Multicollinearity occurs when the independent variables are correlated with one another
  - Correlation can be measured using *r*
  - Avoid at all costs!
    - Will produce "junk" regression results

Multicollinearity

- A simple method to detect multicollinearity is to examine the correlation matrix
  - Prior to regression!
    - Examine all variables that you are considering including in your multiple regression
  - Not necessary to test for the significance of the correlation
    - Only the *r* value is required

Correlation Matrix

- Most software packages can produce a correlation matrix
  - The correlation coefficient (r) for multiple combinations of variables

	ENROLLMENT	PBERATE13	MedHouInc	PcEdltHS	PcEdColDeg	WhPct	AsPct	HispPct	PopDenKMsq	SchType
ENROLLMENT	1.00	-0.19	-0.08	0.19	-0.16	-0.20	-0.04	0.22	0.06	-0.54
PBERATE13	-0.19	1.00	0.08	-0.25	0.14	0.36	-0.14	-0.26	-0.16	0.10
MedHouInc	-0.08	0.08	1.00	-0.67	0.79	0.40	0.38	-0.56	-0.12	0.13
PcEdltHS	0.19	-0.25	-0.67	1.00	-0.75	-0.76	-0.22	0.90	0.24	-0.12
PcEdColDeg	-0.16	0.14	0.79	-0.75	1.00	0.46	0.43	-0.70	0.05	0.20
WhPct	-0.20	0.36	0.40	-0.76	0.46	1.00	-0.27	-0.79	-0.48	0.03
AsPct	-0.04	-0.14	0.38	-0.22	0.43	-0.27	1.00	-0.29	0.36	0.10
HispPct	0.22	-0.26	-0.56	0.90	-0.70	-0.79	-0.29	1.00	0.25	-0.09
PopDenKMsq	0.06	-0.16	-0.12	0.24	0.05	-0.48	0.36	0.25	1.00	0.10
SchType	-0.54	0.10	0.13	-0.12	0.20	0.03	0.10	-0.09	0.10	1.00

Multicollinearity

- Evaluating the correlation among independent variables
  - How much correlation is too much?
    - A somewhat difficult question!
  - Common rules of thumb for *r*
    - No more than 0.8 (highly relaxed)
      - Personally, I believe this is too much correlation
    - No more than 0.5
      - This is generally where I start to get pretty nervous about the "independence" of my independent variables

Multicollinearity

- Variance Inflation Factor (VIF)
  - Higher VIF signals more multicollinearity
  - Rules of thumb
    - VIF > 10, VIF > 7.5, VIF > 2
- Tolerance
  - VIF is reciprocal of Tolerance
    - e.g., VIF = 2 = Tolerance = 0.5
    - Lower tolerance signals more multicollinearity

Multicollinearity

- Multicollinearity Condition Number (MCN) < Geoda
  - Higher MCN signals more multicollinearity
  - Rules of thumb
    - MCN > 30, MCN > 15

Multicollinearity

- My advice…
  - Use multiples
    - Check *R* values prior
      - Help decide what to include or not include in regression
    - Check Tolerance, VIF, or MCF after
      - Help decide whether results can/should be trusted

Inference

- Required checks (post regression)
  - Residuals should be normally distributed
  - Residuals should have equal variance
  - Observations must be independent
    - For spatial data, residuals should **not** be spatially autocorrelated (should be random)

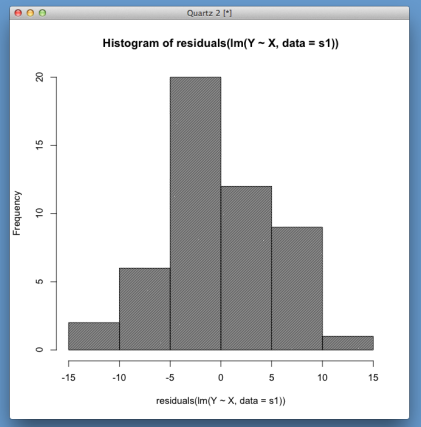
Inference

- Required checks (post regression)
  - Residuals should be normally distributed
  - Residuals should have equal variance
  - **Observations must be independent**
    - **For spatial data, residuals should <u>not</u> be spatially autocorrelated (should be random)**

$$Y = \beta_0 + \beta X + \epsilon \quad \rightarrow \quad r^2$$

Regression Residuals

- Regression residuals should be normally distributed
- How to test?
  - Look at the histogram
  - Jarque-Bera statistic
    - If $p < 0.05$, this signals non-normality
      - In practice, normality tests are extremely sensitive
      - Using "real" data, a high chance your residuals will not be normal



Regression Residuals

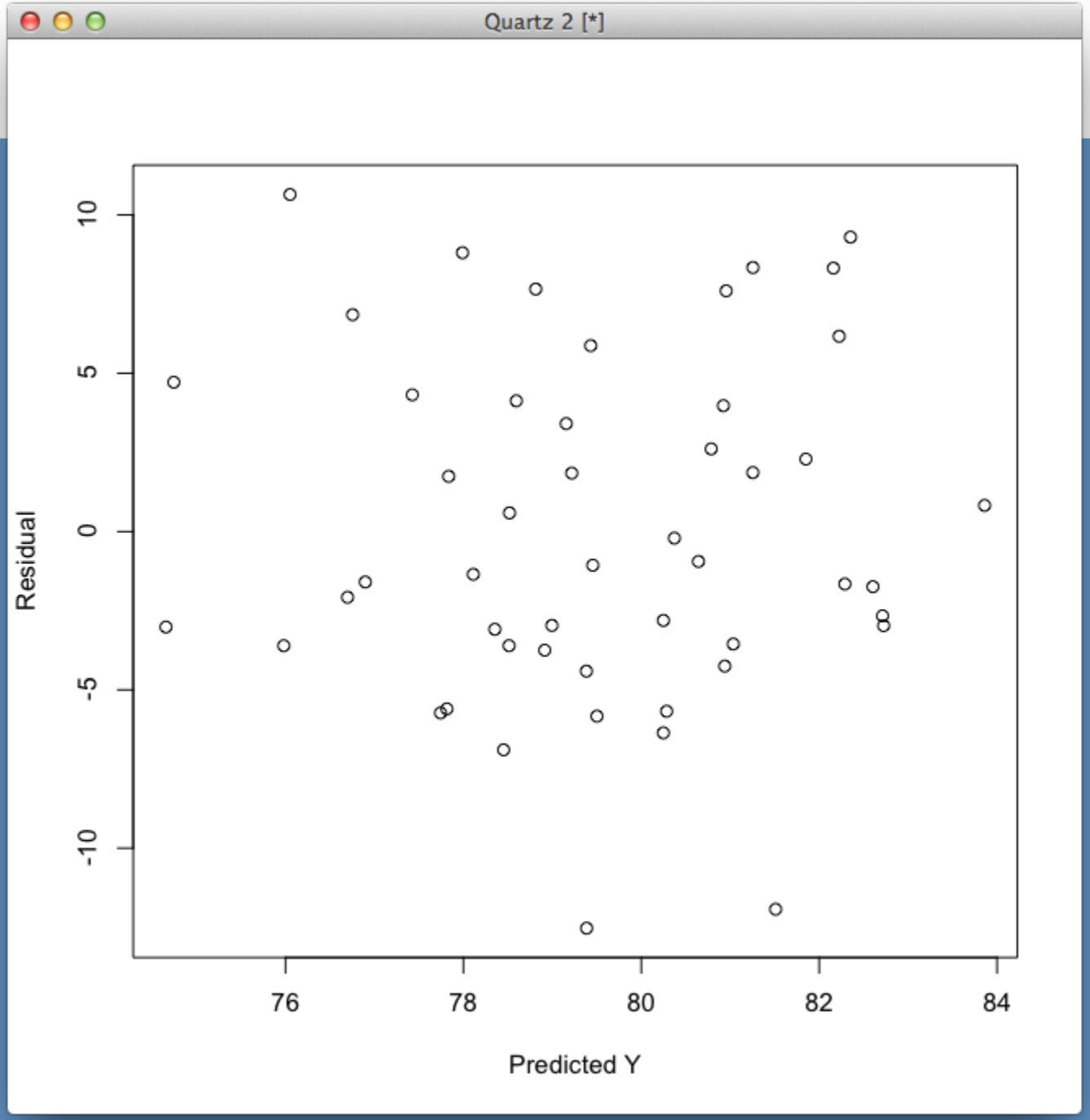
- If your residuals are extremely non-normal
  - Potential effects
    - Model is invalid (misspecified)
      - e.g., you cannot trust anything!
    - Standard errors on coefficients are unreliable (too narrow)
      - e.g., you cannot trust the *p*-values on the β coefficients

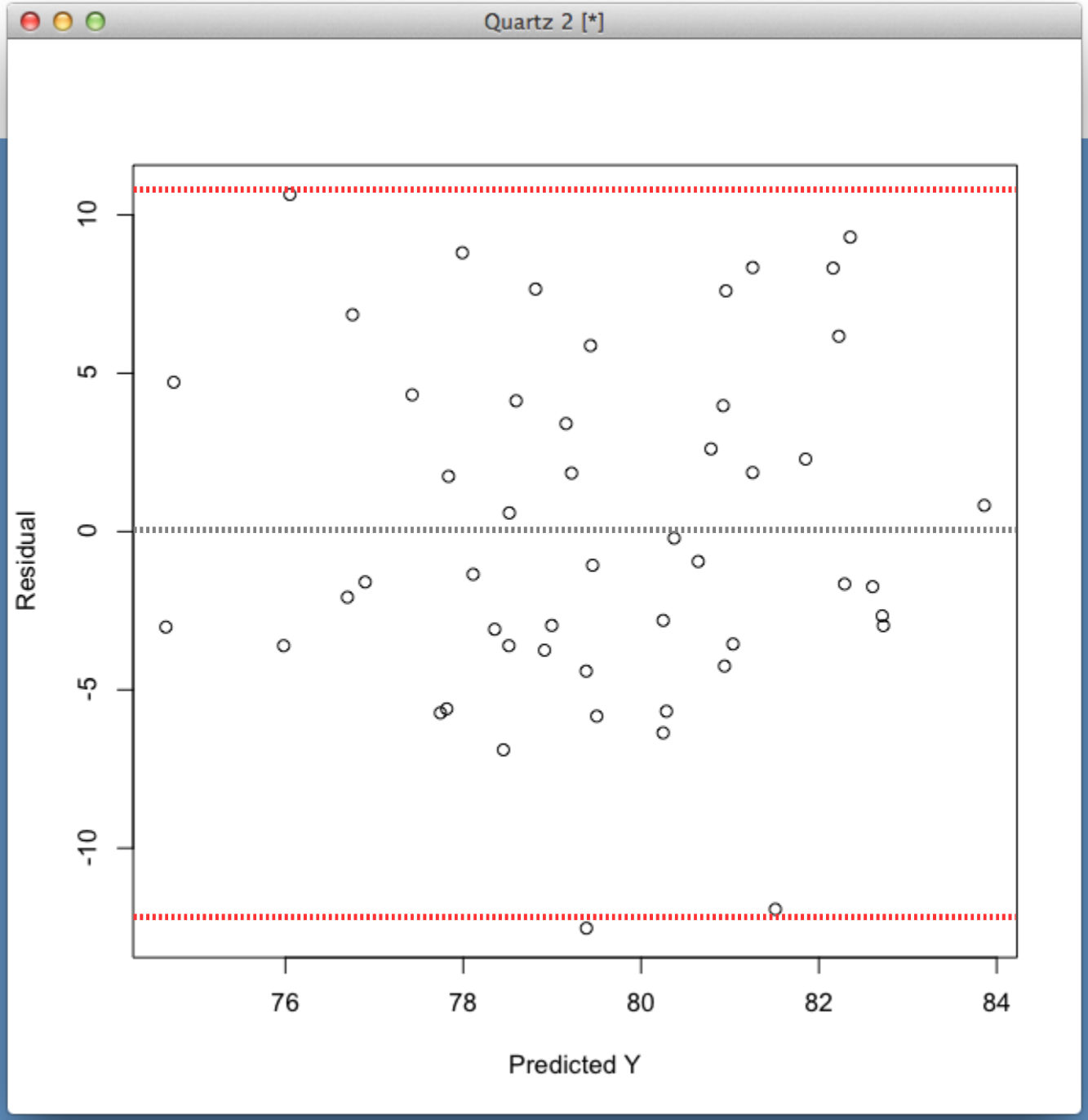
Regression Residuals

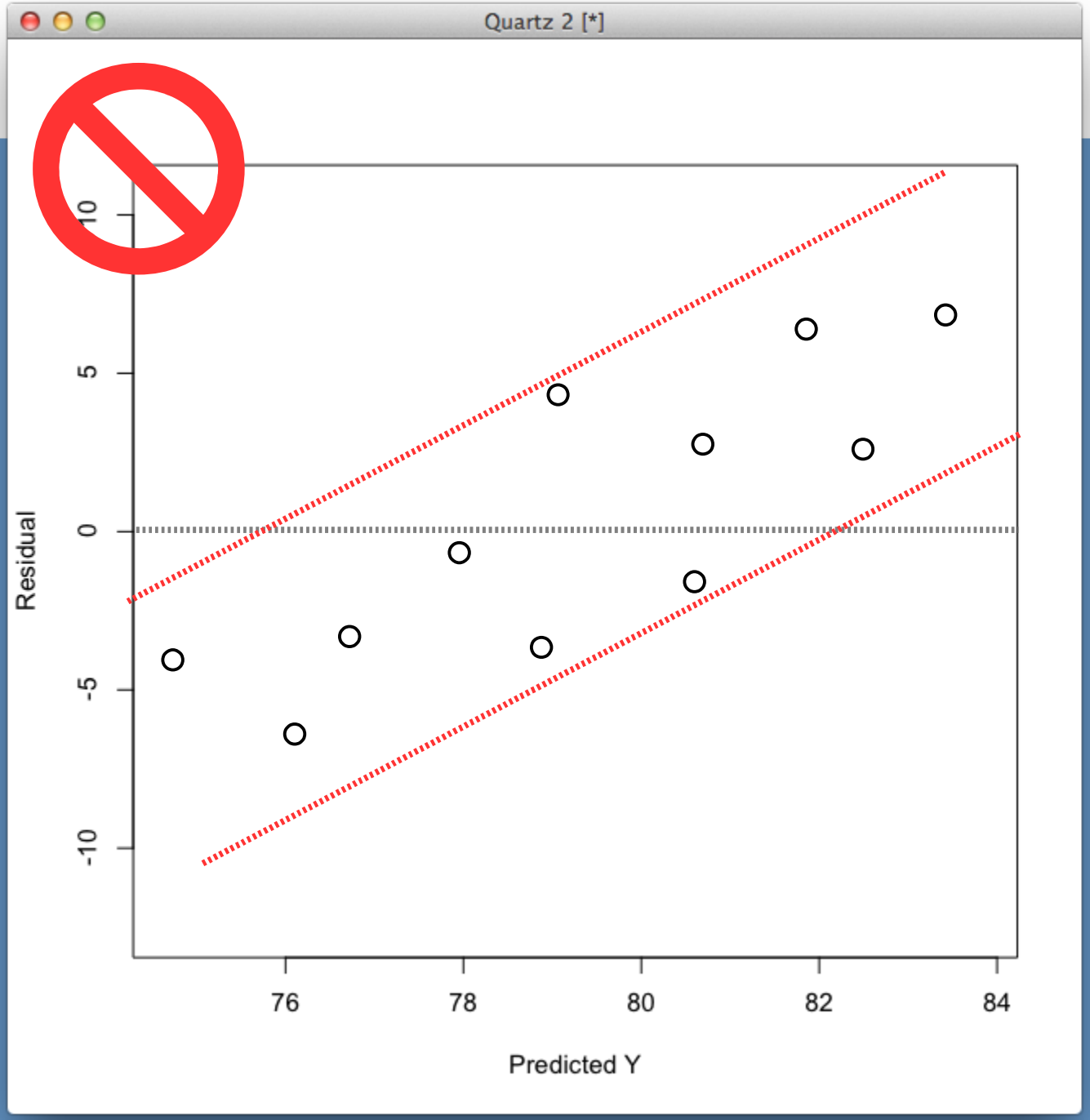
- Potential causes… and fixes
  - Non-linear relationship(s) or…
  - Extremely non-normal Y or X data
    - Mathematical transform (e.g., log)
    - Non-OLS regression model
  - Outliers
    - Removal (must be justified)
  - Robust regression approaches

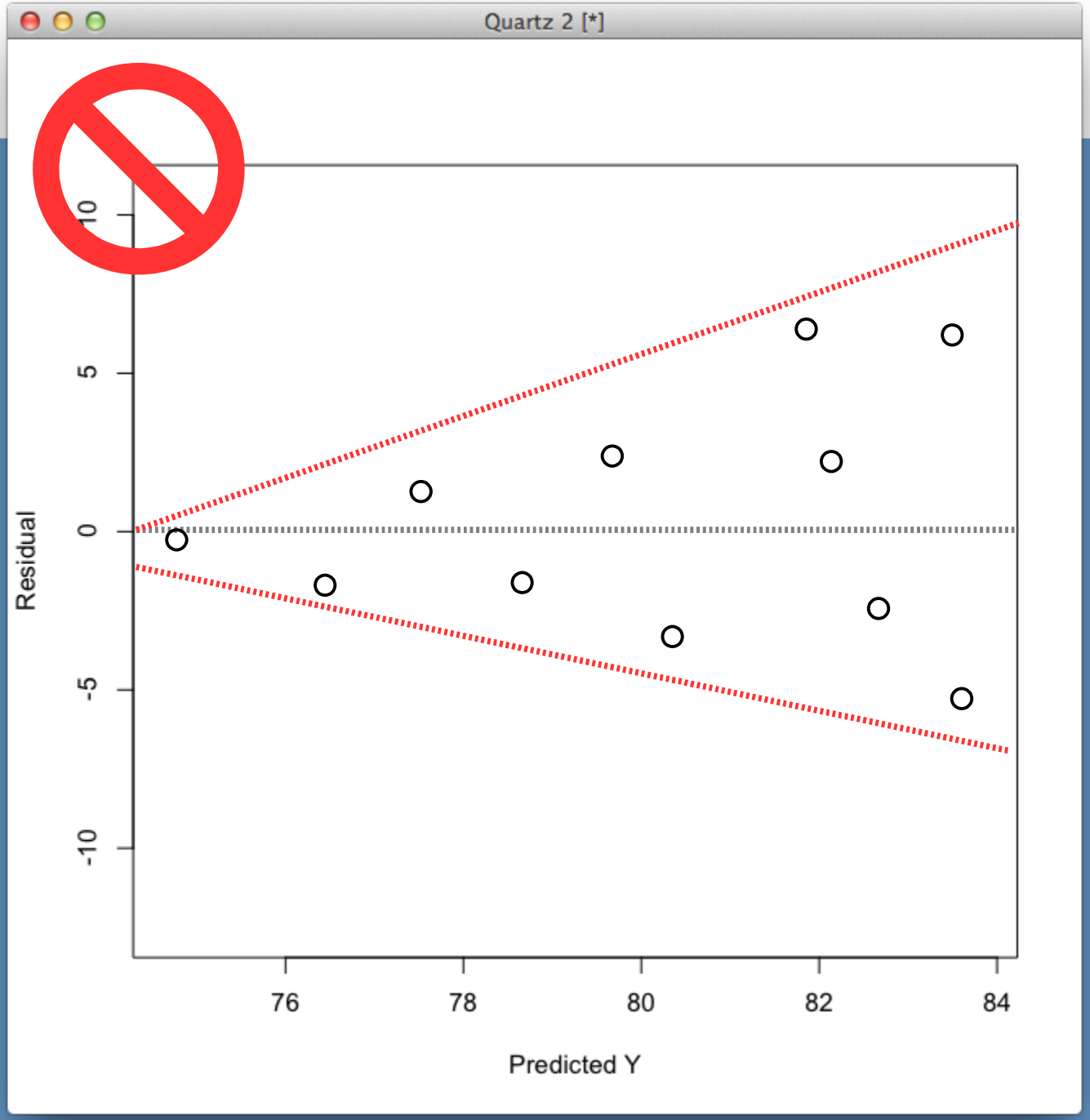
Homoscedasticity

- Residuals should have equal variance
  - Variation is similar over the range of predicted Y values
    - The opposite is heteroscedasticity
  - Plot the residuals against the modeled Y values
    - Should be box or rectangle shaped
      - Watch out for "cones" or "trends"


Quartz 2 [*]
Residual
10
5
0
-5
-10
76
78
80
82
84
Predicted Y


Quartz 2 [*]
Residual
10
5
0
-5
-10
76
78
80
82
84
Predicted Y


Quartz 2 [*]
Residual
10
5
0
-5
-10
76
78
80
82
84
Predicted Y


Quartz 2 [*]
Residual
10
5
0
-5
-10
76
78
80
82
84
Predicted Y

Homoscedasticity

- Residuals should have equal variance
  - Plot residuals against the modeled Y values
  - Breusch-Pagan test
  - Levene test
    - If $p < 0.05$, this signals heteroscedasticity (unequal variance)

Homoscedasticity

- If your residuals are heteroscedastic
  - Potential effects
    - Standard errors on coefficients are unreliable (too narrow)
      - e.g., you cannot trust the *p*-values on the β coefficients

Homoscedasticity

- Potential causes… and fixes
  - Unequal weighting among observations
    - Weighted regression approach (e.g., by population size)
  - Non-normal Y or X data
    - Mathematical transform (e.g., log of Y)
  - Another fix: White adjustment

Residual Autocorrelation

- **Residuals** should not be spatially autocorrelated – independence
  - Run a **Moran's *I*** analysis using the **residuals** as the observations
    - We hope to get "null" results (high *p* value), meaning regression residuals are randomly distributed

Please, please, please note that this DOES NOT mean that we remove variables from a regression if they have spatial autocorrelation. What this means is that regression residuals cannot be spatially autocorrelated.

Residual Autocorrelation

- Potential fixes...
  - Find missing independent variable
  - Spatial regression approaches

Keywords

- Explanatory vs predictive
- Correlation
- Regression
- Confounding
- β, R^2, p value
- Multiple regression
- Multicollinearity
- Residuals
  - Independent, normal, homoscedastic