Regression
Part 2

Lecture #22 | GEOG 510
GIS & Spatial Analysis in Public Health

Varun Goel

Outline

- Regression
  - Multiple (Multivariate) regression
  - Motivation for Spatial Regression
  - Demo
  - In-class exercises

Regression

- The magnitude of the influence one variable has on the other

- How does **variable X** (independent/predictor variable) **influence Y** (dependent/response variable)

- Multiple (Multivariate) Regression
  - **Confounding!**
  - How does X influence Y, after controlling for other potential factors that may explain the relationship?

Your Projects

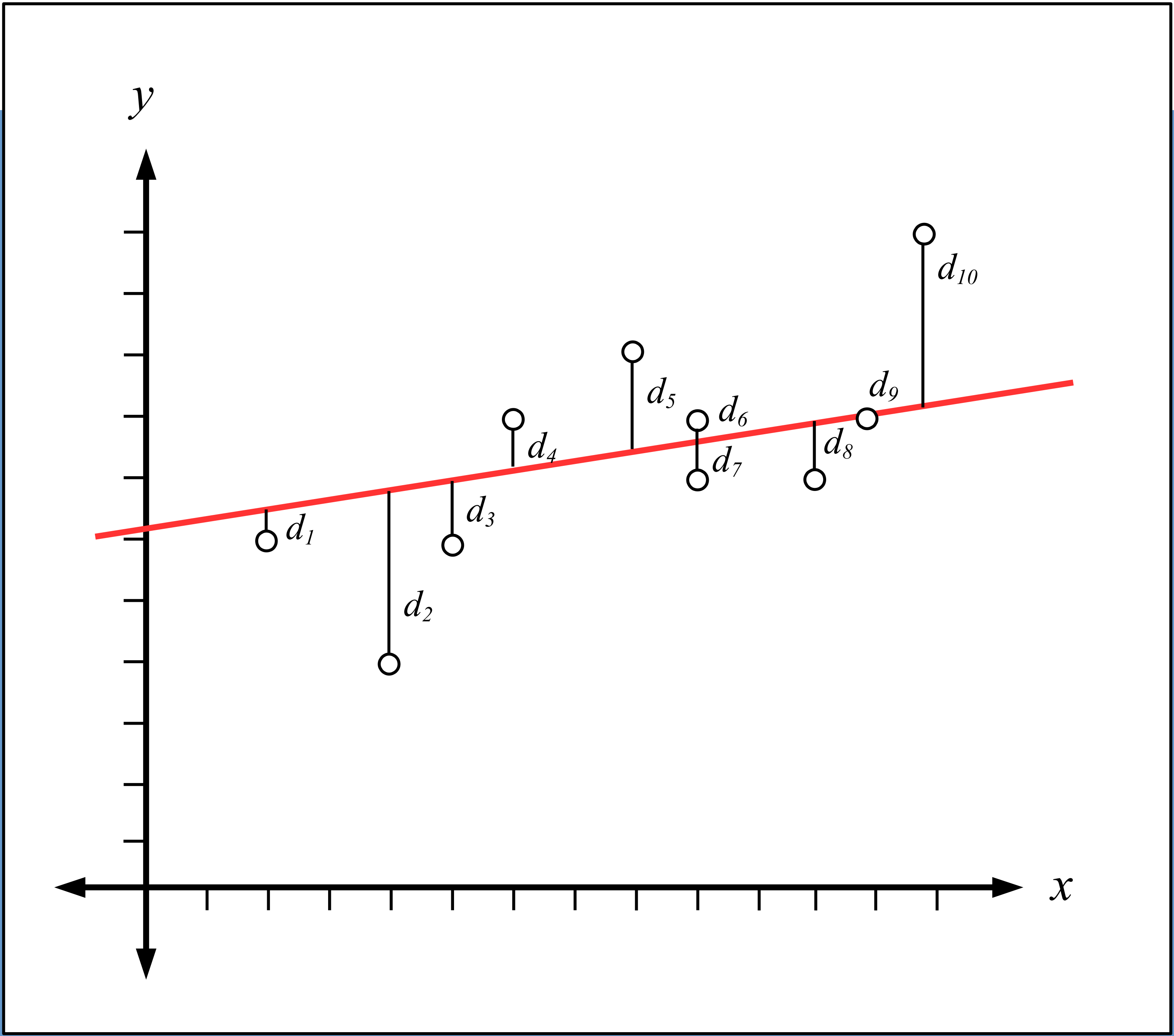
(Name): (X) on (Y) - eg. Varun: heat on mental health

15 responses

- Addy: heat on maternal health
- Grace: heat on physical inactivity
- Regan: Distance and Access to Free Clinics
- Eleni: eviction on death
- Kejsi: air pollution on childhood asthma rates
- Michael Eslick Binge drinking on Birth defects
- Haofeng: vegetation on mental distress
- Liam Baker: Water affordability on food insecurity
- Lucia: Spatial relationship between infectious diseases and temperature
- Amanda: energy burden on mental health
- Mikey: city walkability on drunk driving fatalities
- Pengyu CHEN: Road Slope & Curvature, road condition, traffic flow on traffic

Univariate Regression

- Simple linear regression
  - Linear relationship between variables
    - Defined as $\mathrm{Y} = a + b\mathrm{X}$
  - 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

Univariate OLS Regression

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

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

Univariate OLS Regression

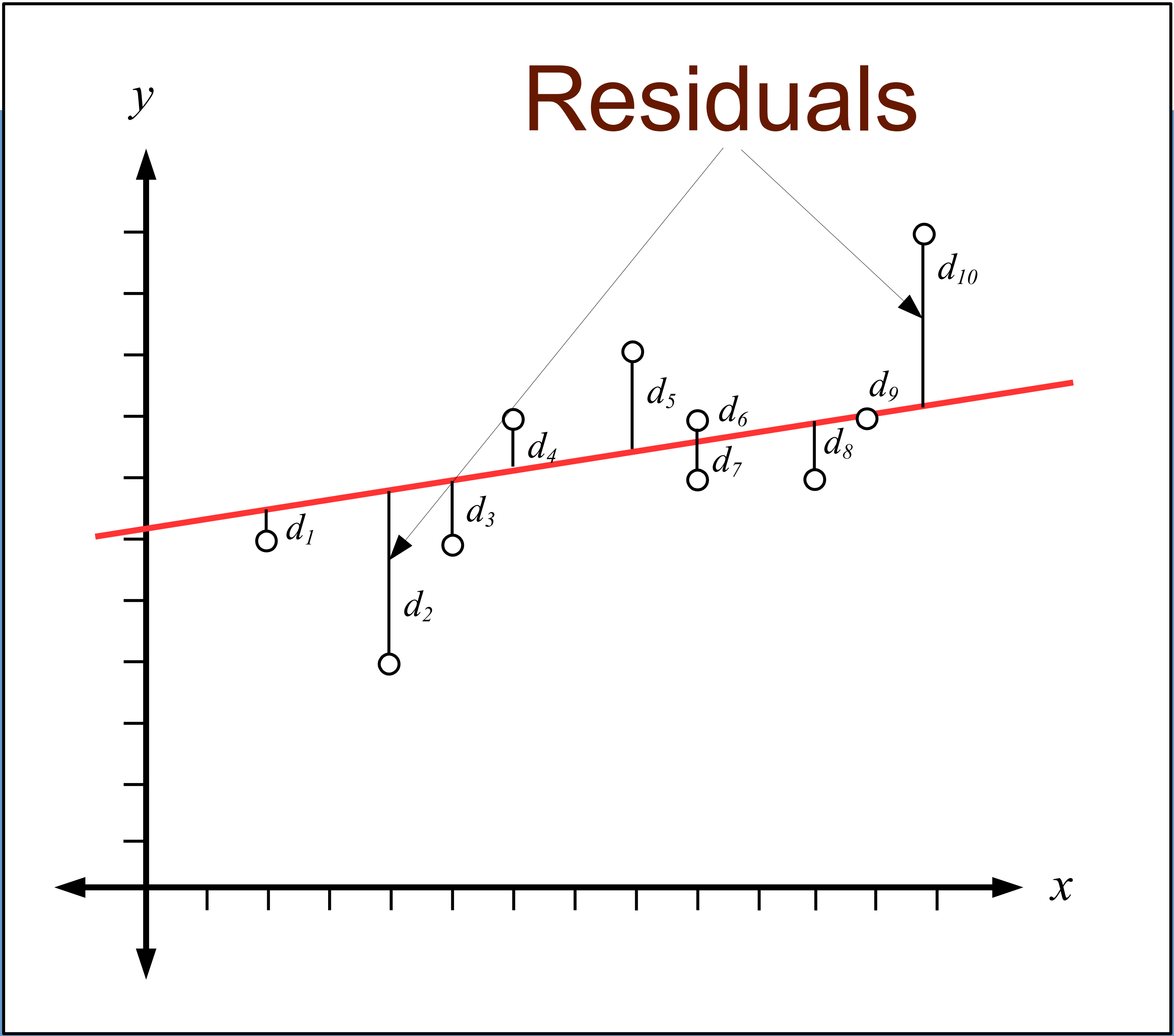
$$Y = \beta_0 + \beta X + \epsilon \quad \rightarrow \quad \boxed{r^2}$$

- Coefficient of Determination
  - The proportion of Y explained by X
  - $H_0 : r^2 = 0 \qquad H_A : r^2 \neq 0$
    - 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

Univariate OLS Regression

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

- Error (residuals)
  - Everything else not explained by the model
  - Effect of unmeasured variables
  - Random, if model is well specified
    - However, not random in practice
    - Governed by strong assumptions, that are often violated

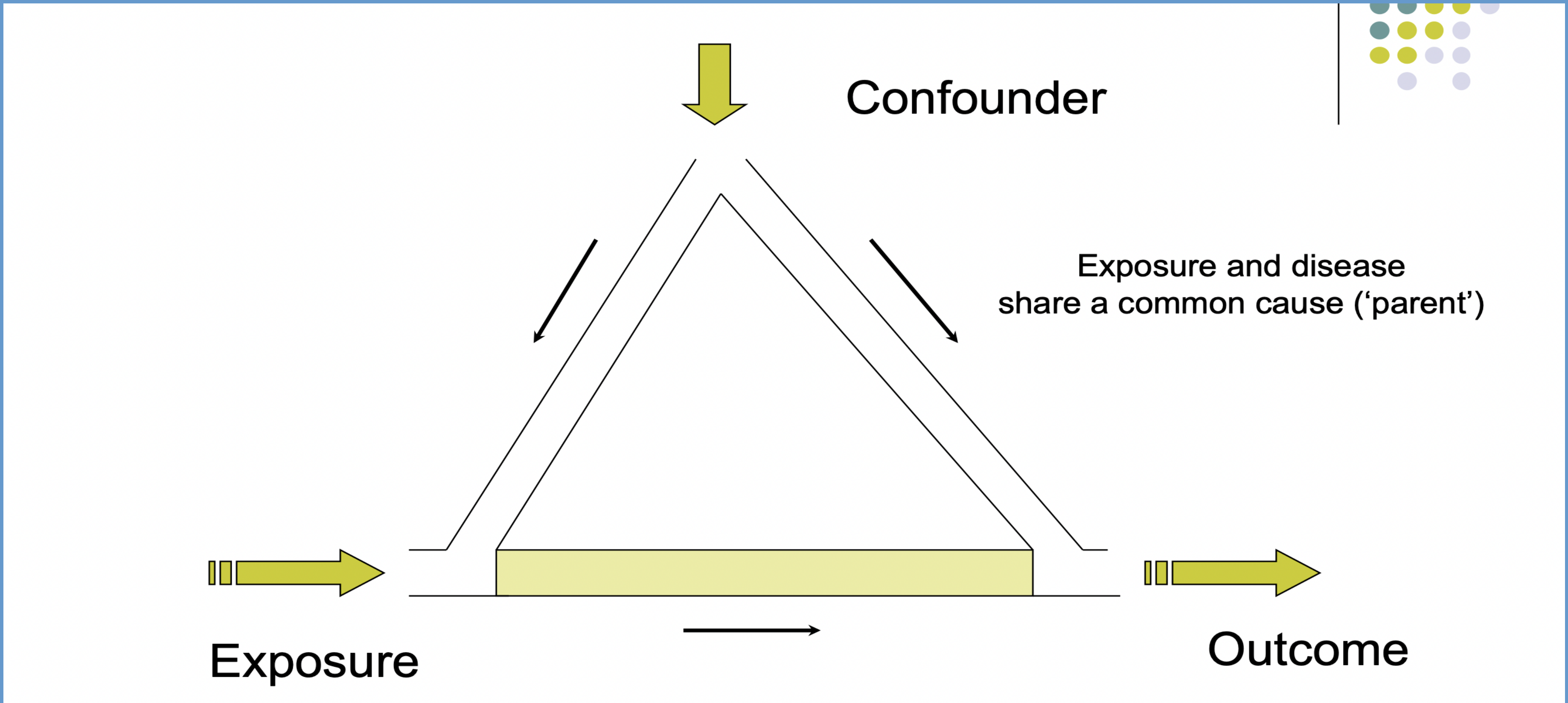


Point	*x*	*y*
1	2	6
2	4	4
3	5	6
4	6	8
5	8	9
6	9	8
7	9	7
8	11	7
9	12	8
10	13	10

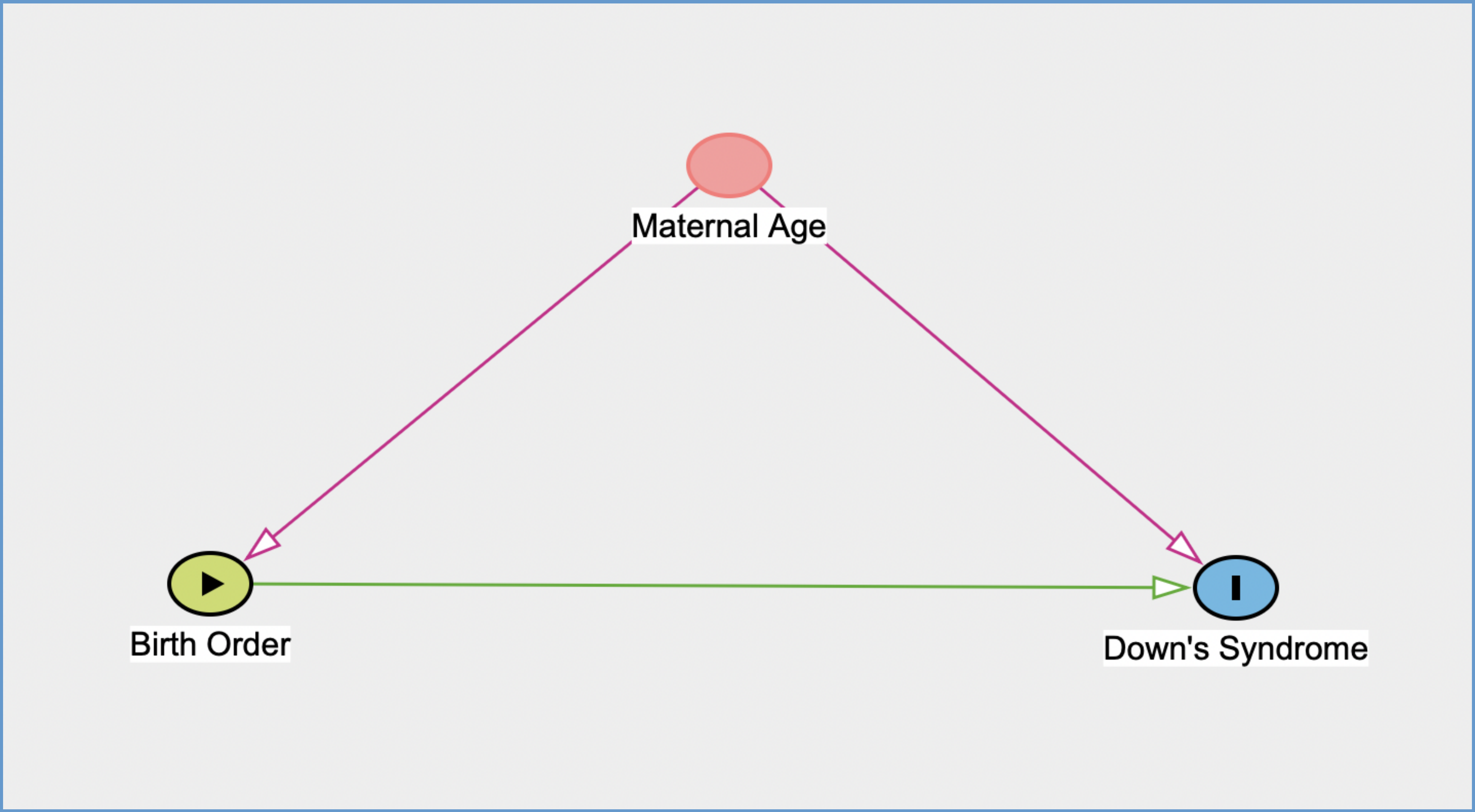
Multiple Regression

- 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

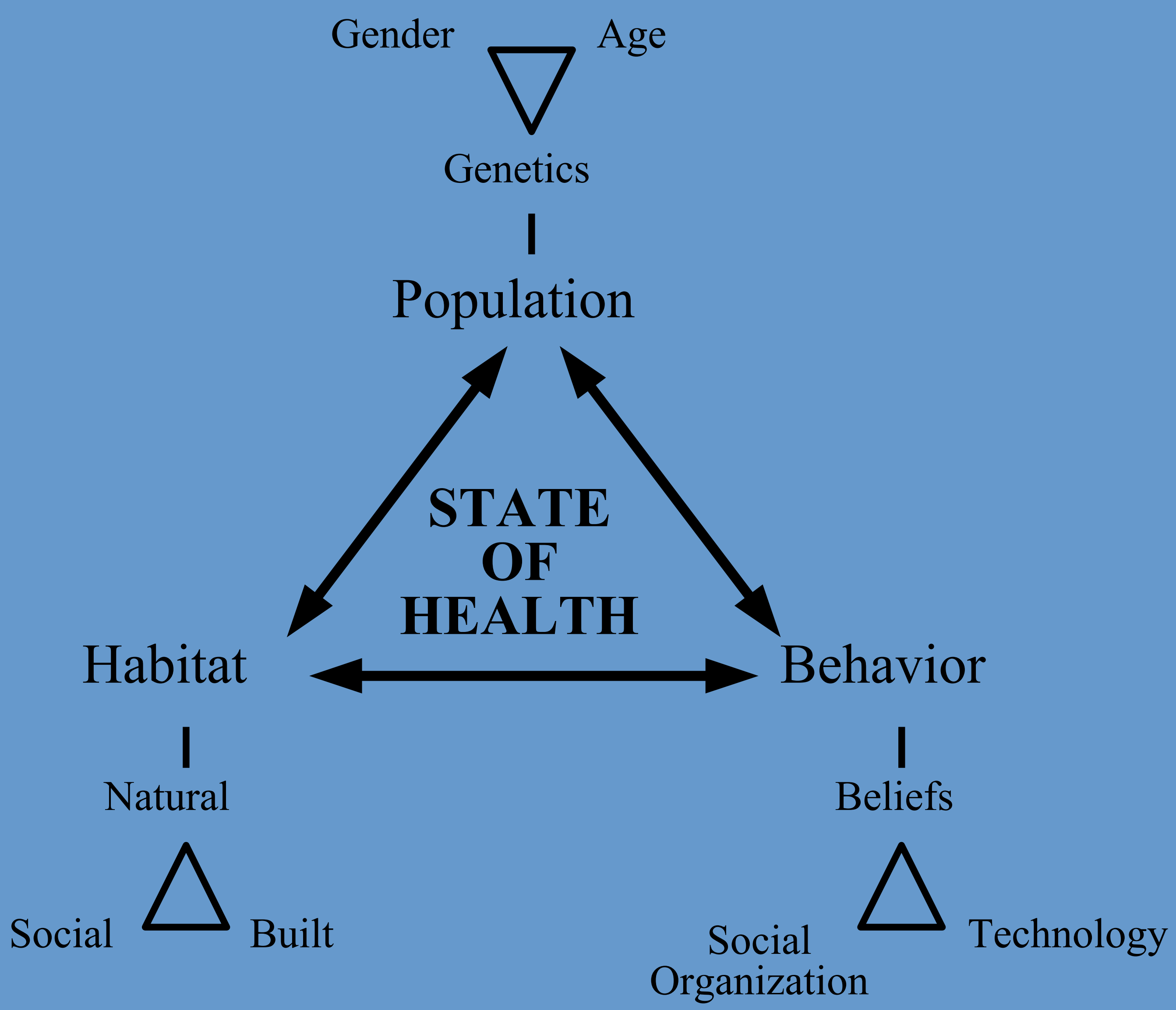
Confounding



Confounding



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

$$Y = \beta_0 + \boxed{\beta_1} X_1 + \boxed{\beta_2} X_2 + \epsilon$$

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

Multiple Regression

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

- 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 OLS Regression

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

Multicollinearity

- Independent variables must be independent from one another
  - Means not correlated!
- <u>Multicollinearity</u> 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) in 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)

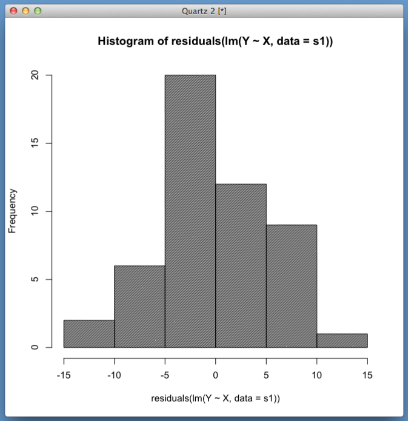
Inference

- 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)**

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

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



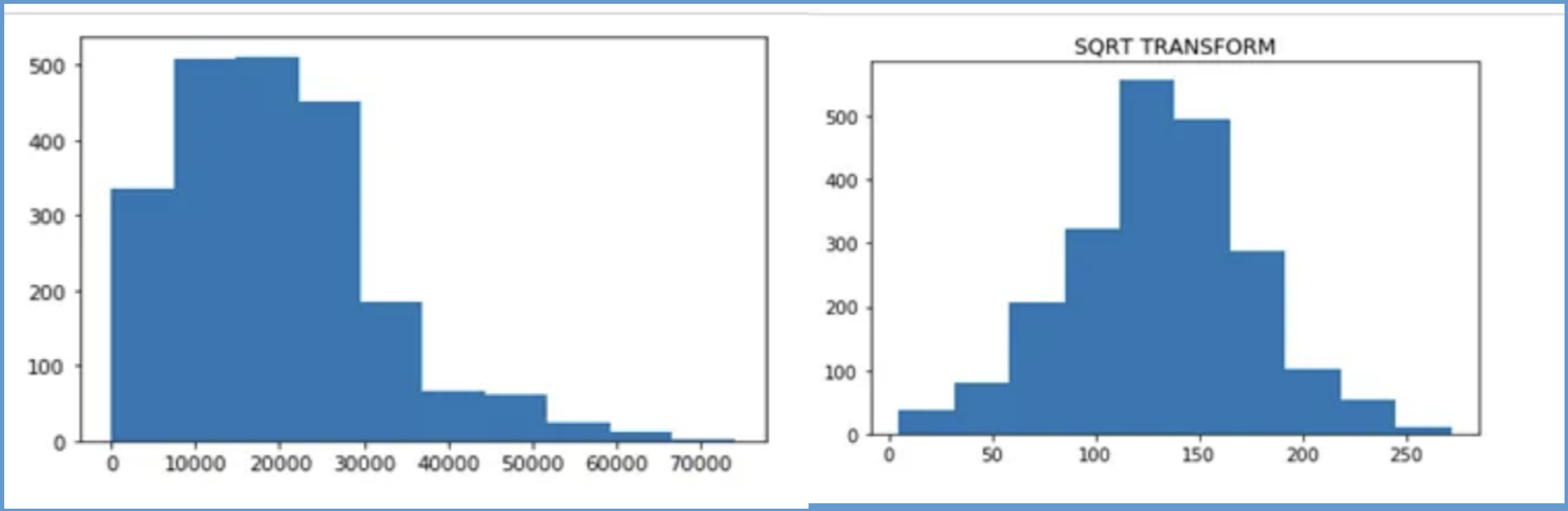
Regression Residuals

- 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

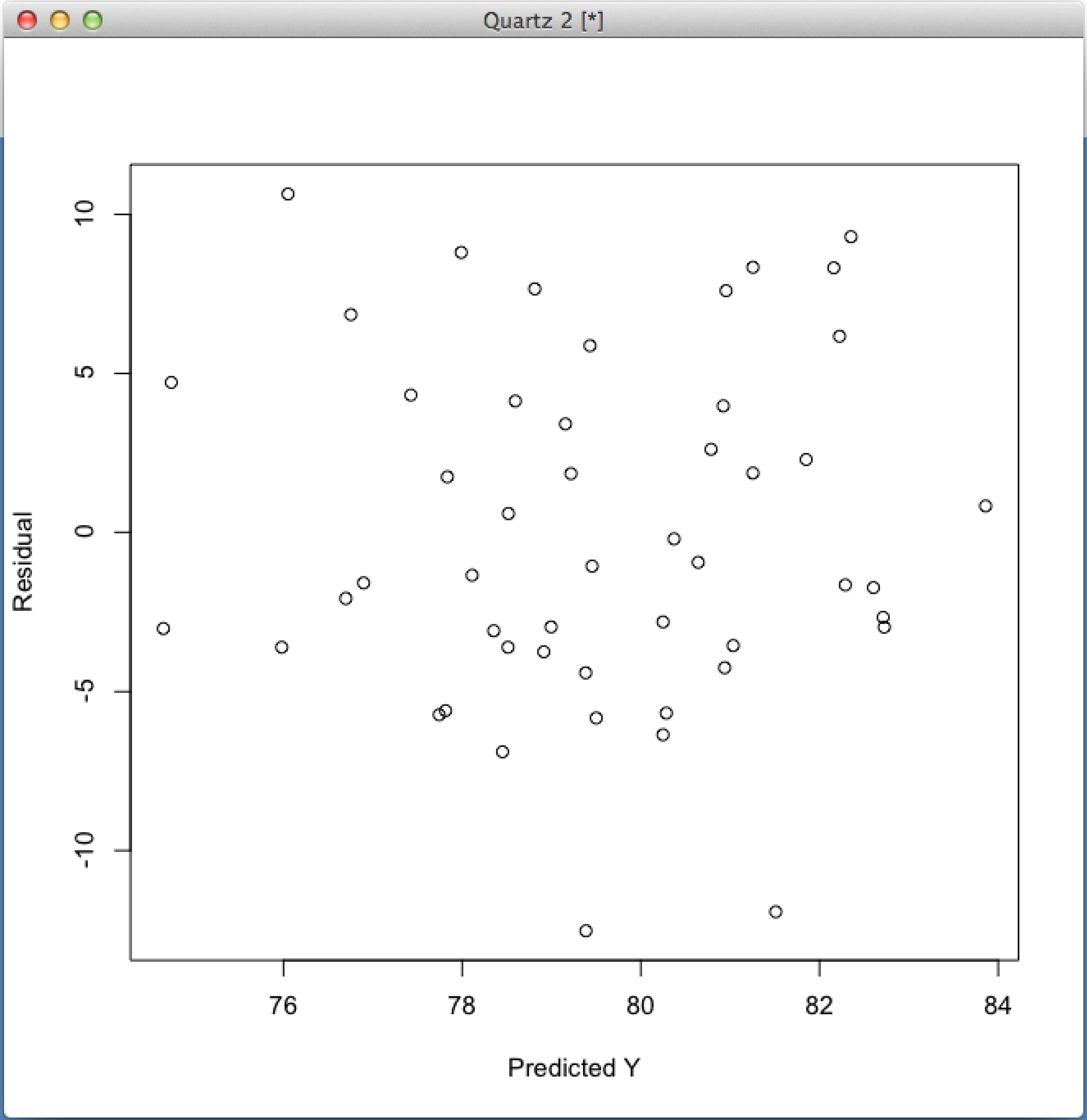
Regression Residuals

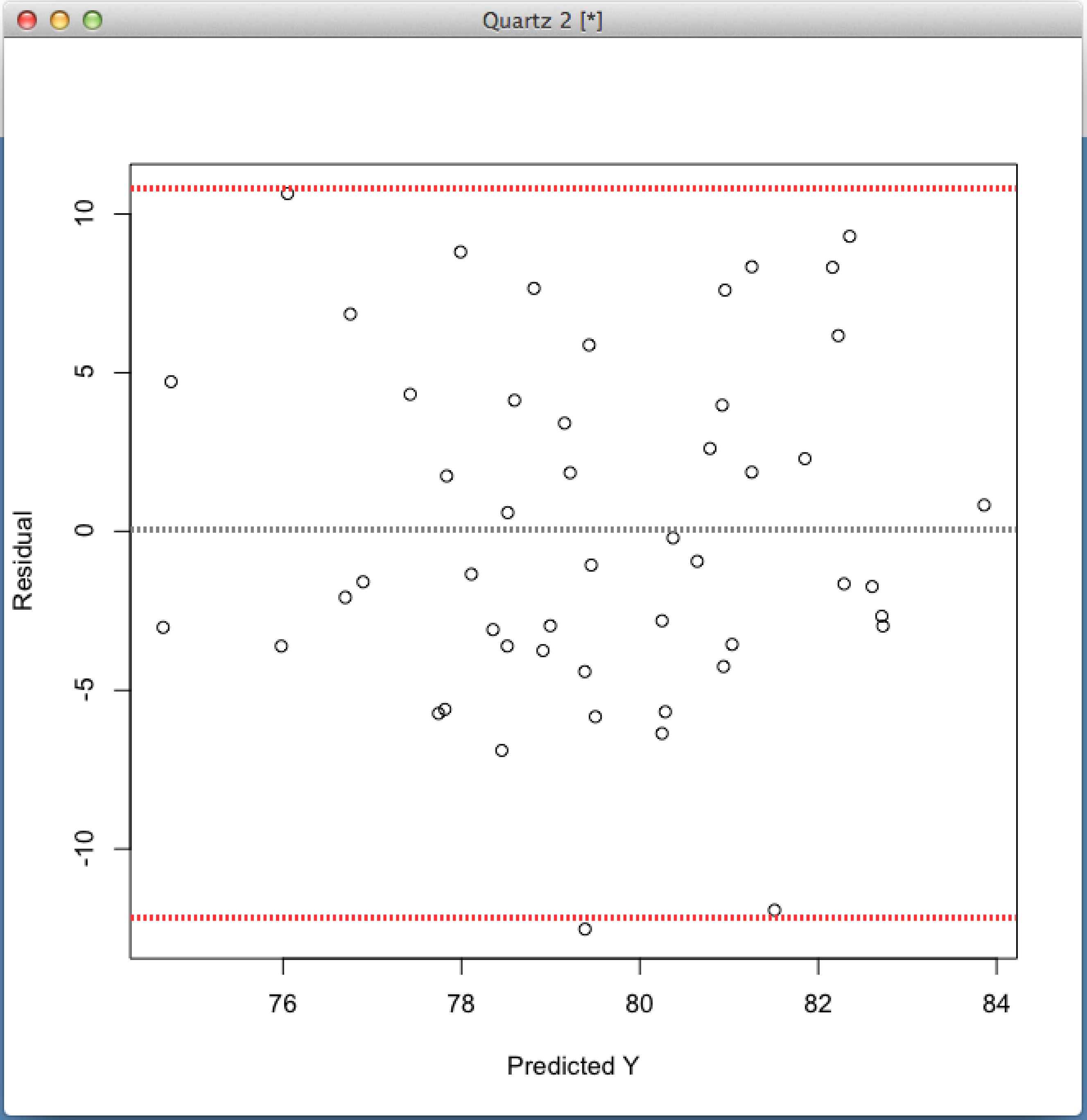


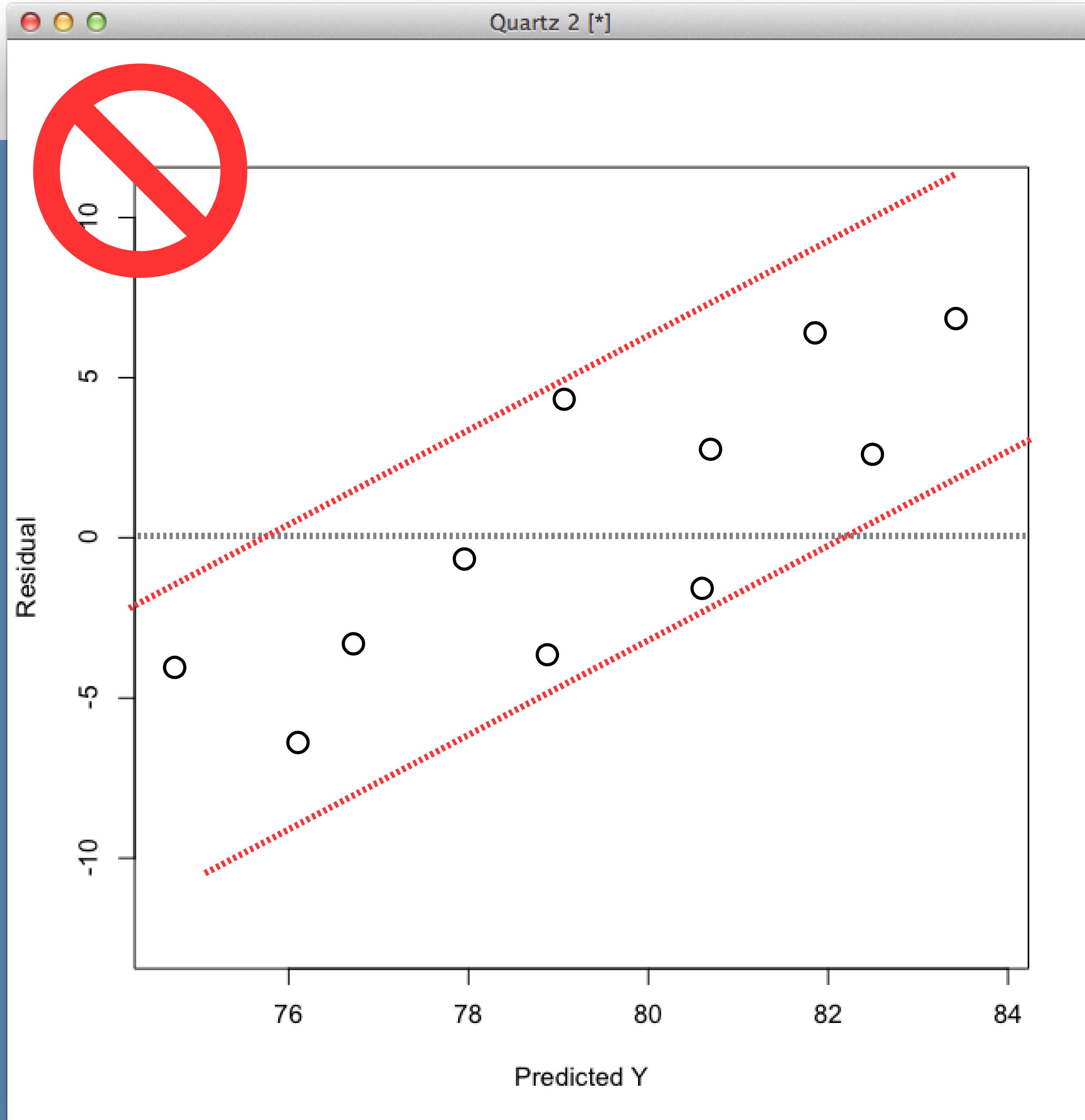
Non-normal regression residuals transformed to normal using a square root transformation

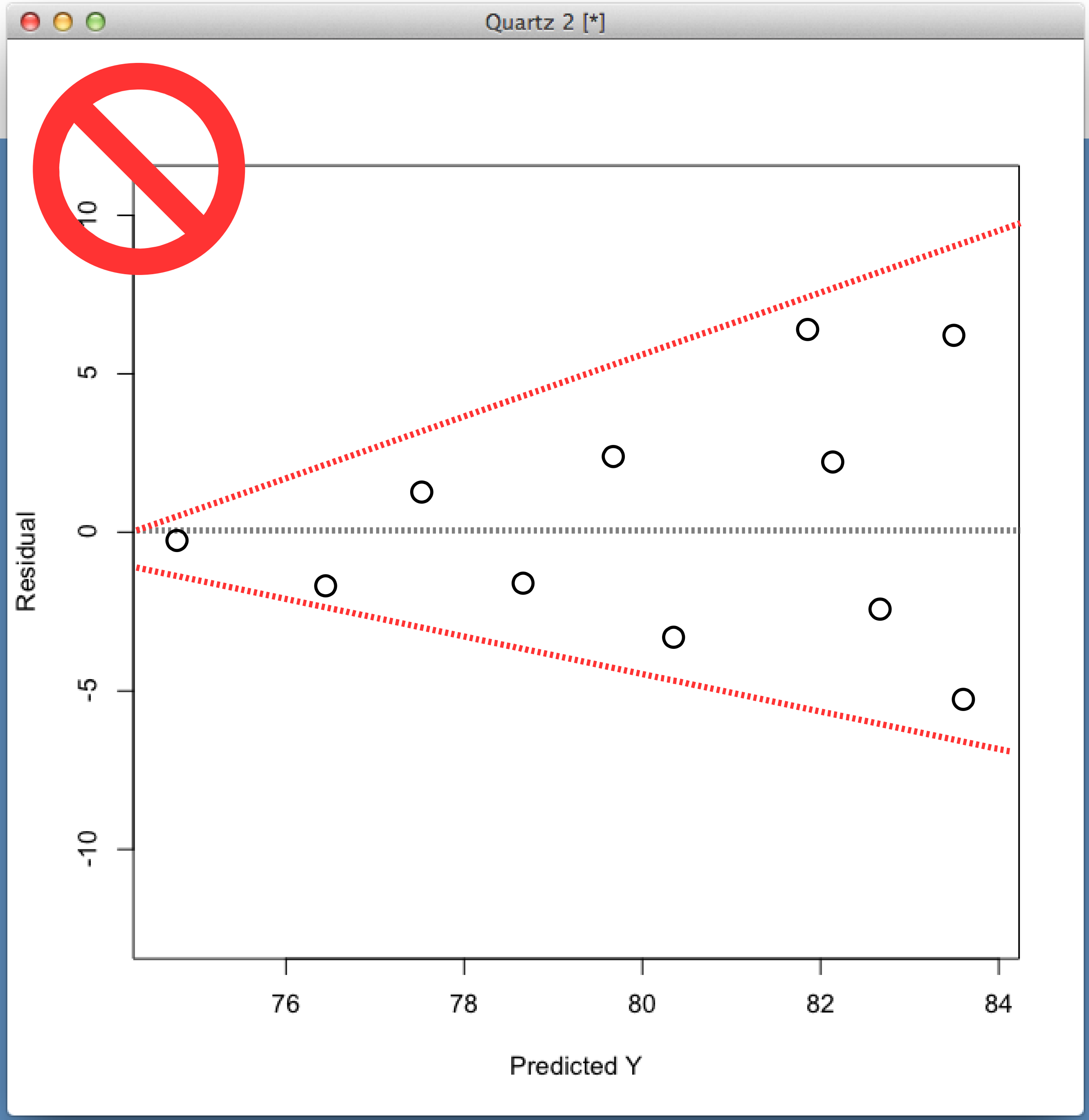
Homoscedasticity

- 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 [*]
10
5
0
-5
-10
Residual
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

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
  - Spatial autocorrelation!

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