

# PLSC 503 – Spring 2021

## Regression, Conceptually and Bivariate

January 20, 2021

# “Multivariate Analysis for Political Research”

- “Regression” course
- Texts: Weisberg (2013) plus some other readings
- Course materials: <https://github.com/PrisonRodeo/PLSC503-2021-git>
- Preceptor: [Brandon Bolte](#)
- Software:  $R > Stata > \text{others}$
- Grading: Ten homework assignments (@ 50 points), plus a final project (500 points)

# Things We Will And Won't Do

Will: "Regression":

$$Y = f(\mathbf{X})$$

Won't: Multivariate regression:

$$\mathbf{Y} = f(\mathbf{X})$$

Won't: Measurement (e.g. PCA, factor analysis, etc.):

$$\mathbf{Y} = \mathbf{W}^T \mathbf{X}$$

Won't: Classification:

- Cluster Analysis
- Classification and Regression Trees  $\rightarrow$  Random Forests, etc.
- Pattern Recognition
- Machine Learning, Support Vector Machines, etc.

“Regression,” conceptually:

$$\Pr(Y|\mathbf{X}) = f(\mathbf{X})$$

Two important things:

- The distribution of  $Y$  is *conditional on all variables in  $\mathbf{X}$* , and
- The conditional distribution of  $Y$  is conditional on the *joint distribution* of the elements of  $\mathbf{X}$ .

→ Regression is hard...

Figure: Infant Mortality and Life Expectancy (data from 2000)

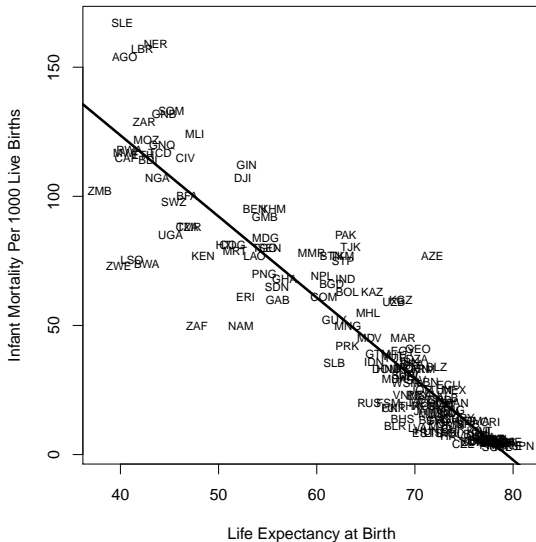
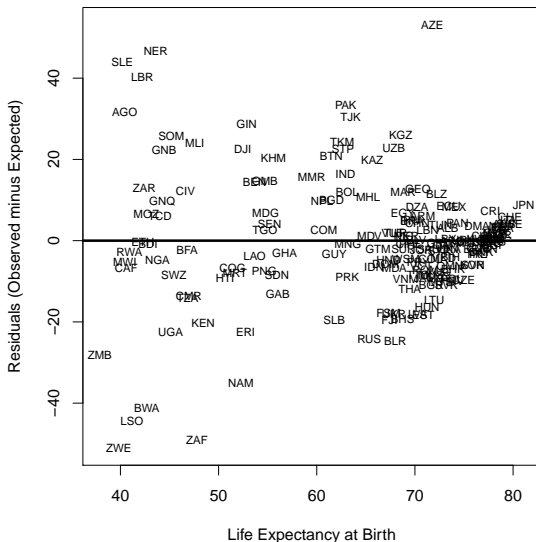
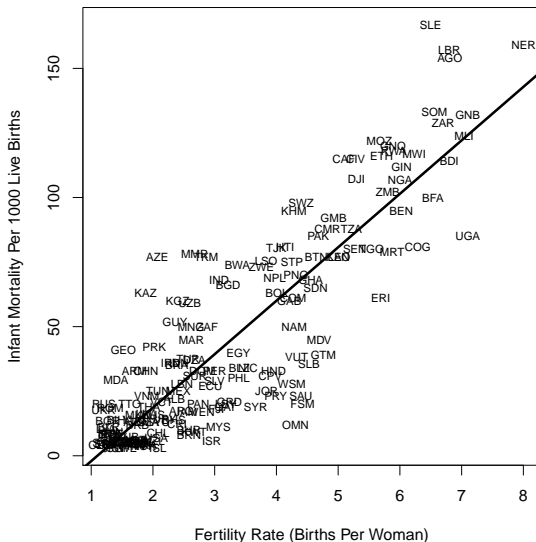


Figure: Infant Mortality and Life Expectancy: “Residuals”



### Figure: Infant Mortality and Fertility



### Figure: Infant Mortality and Wealth

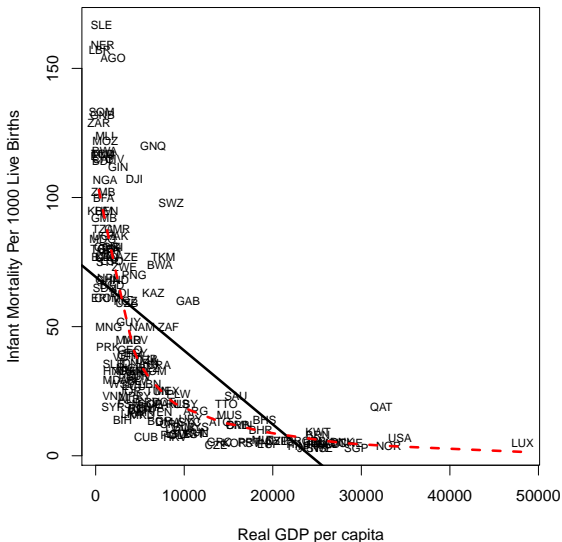




Figure: (Logged) Infant Mortality and (Logged) Wealth

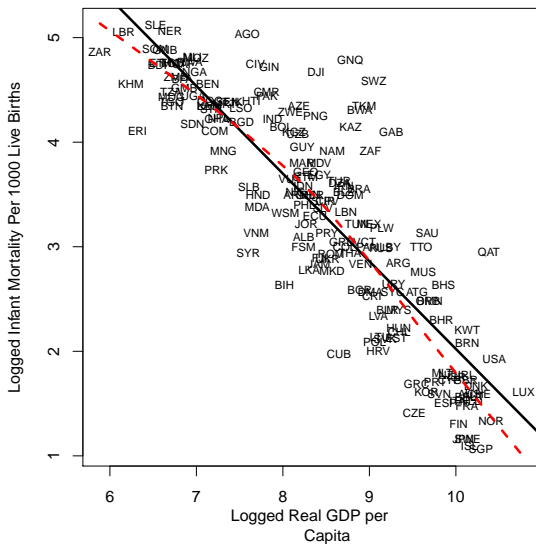




Figure: Infant Mortality, (Dichotomized) Wealth, and Democracy

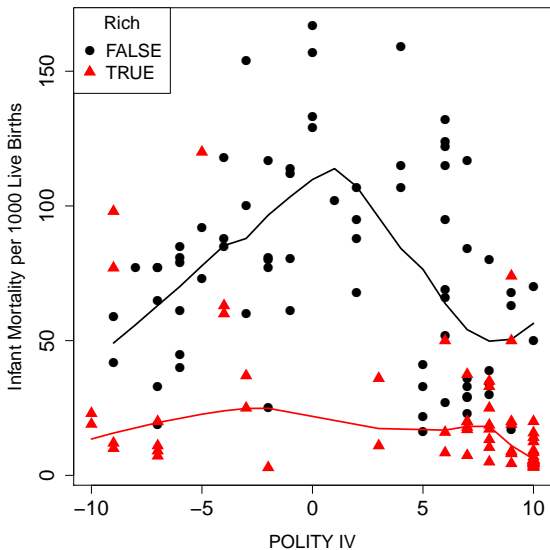


Figure: Measurement: National Health Indicators

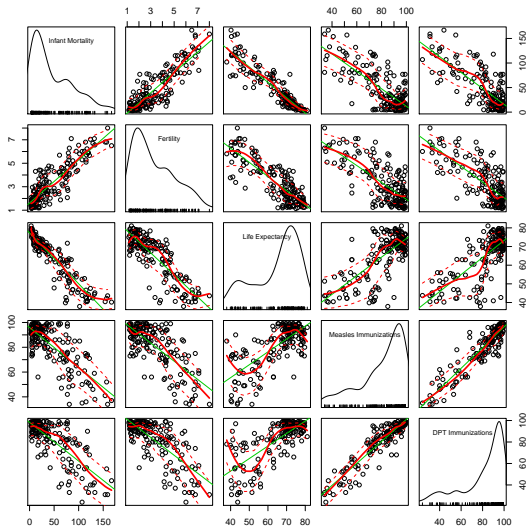
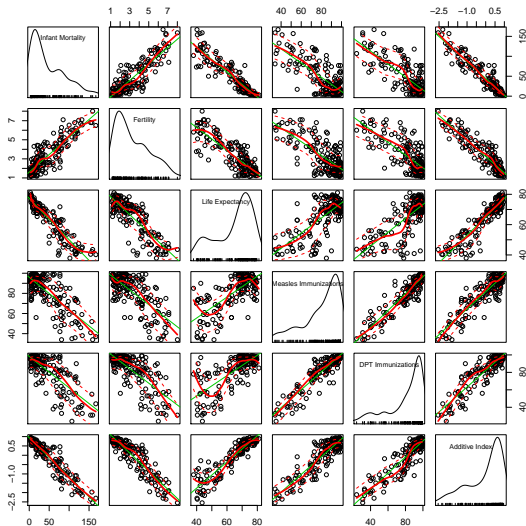


Figure: Measurement: National Health Indicators, Plus Additive Index



# Why regression?

	Description	Explanation	Prediction
<b>Task</b>	Summarize data	Correlation/causation	Forecast OOS / future data
<b>Emphasis</b>	Data	Theory / Hypotheses	Outcomes
<b>Focus</b>	Univariate	Multivariate	Multivariate
<b>Typical Application</b>	Summarize / "reduce" data	Discuss marginal associations between predictors and an outcome of interest	Optimize out-of-sample predictive power / minimize prediction error