Simple and multiple linear regression

Amm Quamruzzaman
ISF, UC Berkeley

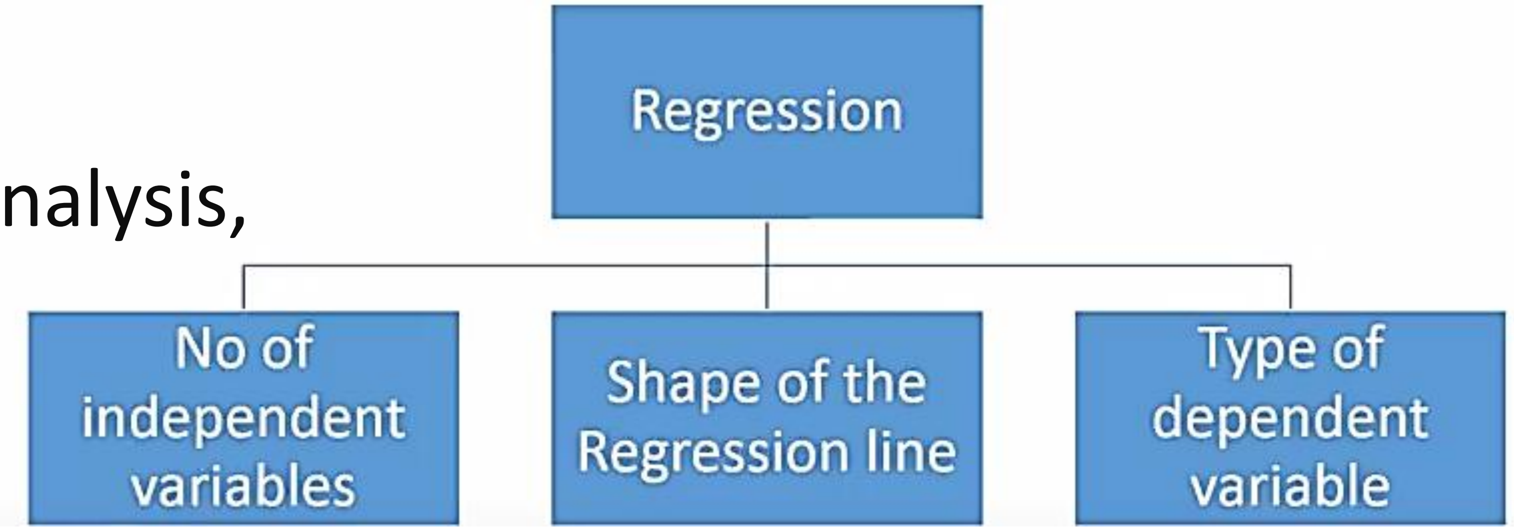
March 15, 2022

Outline

I. Introduction
II. Regression analysis
   A. Linear regression
   B. Multiple regression
   C. Curvilinear/polynomial regression
   D. Logistic regression
   E. Multilevel regression
III. Remarks

I. Introduction

- Regression analysis is a way of mathematically sorting out which factors do have an impact on an outcome (DV).
- It answers the questions: Which factors matter most? Which can we ignore? How do those factors interact with each other? And, perhaps most importantly, how certain are we about all these factors?
- There are various types of regression analysis, depending on:

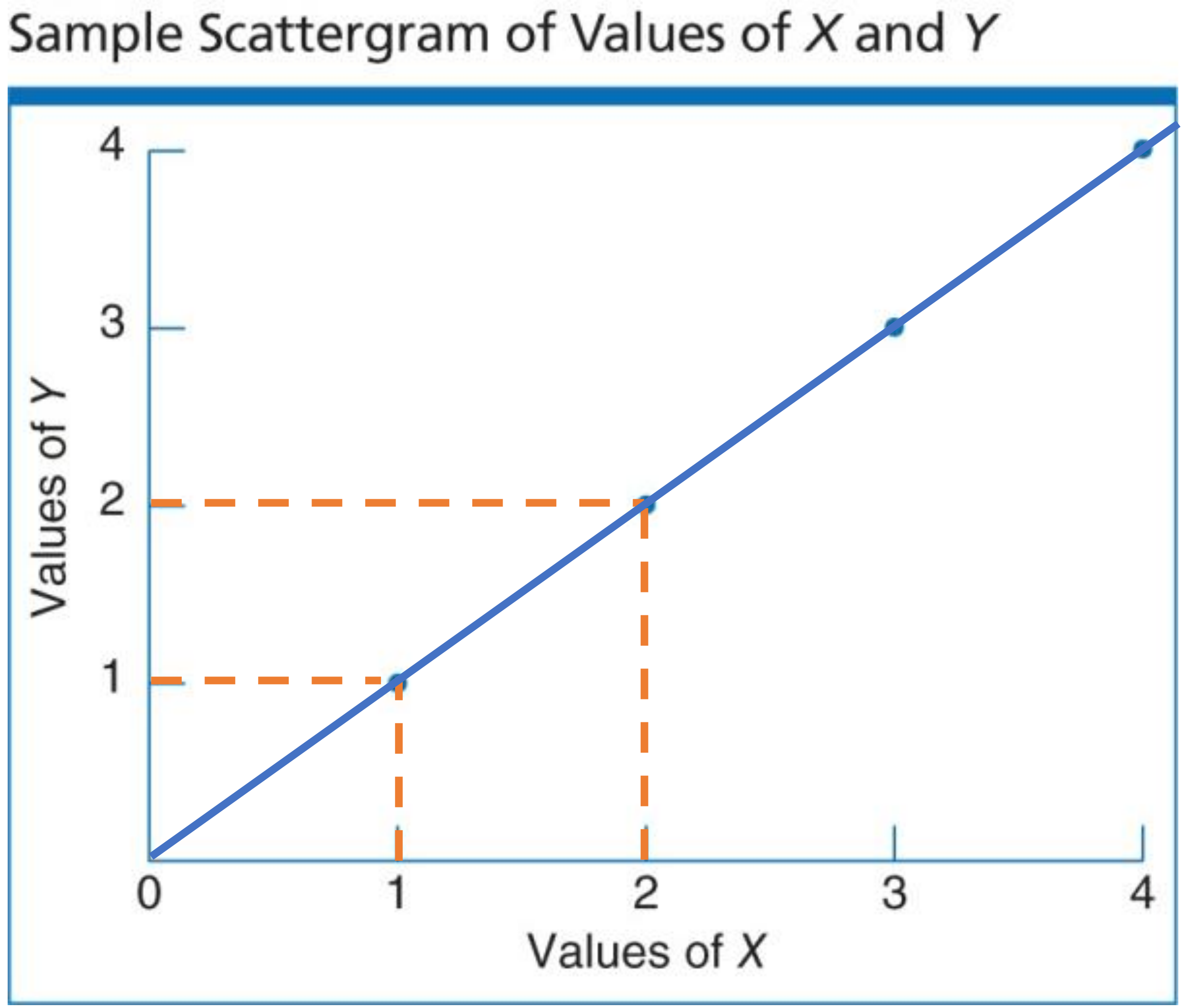


II. Regression analysis

- A method of data analysis in which the relationships among variables are represented in the form of an equation.
- A simple regression equation looks like: Y= *f*(X).
- The starting point is the linear regression, where there is a perfect linear association between two variables (X and Y).
- If values of X increase, the values of Y either increase (positive relationship) or decrease (inverse or negative relationship).
- More complex are multiple regression, logistic regression, curvilinear regression, and multilevel regression.

II.A. Linear regression

- Linear regression: seeks the explanation for the straight line that best describes the relationship between two continuous variables.
- Assumptions: Samples are drawn randomly; samples are large enough (normal distribution); and there is no non-sampling error.



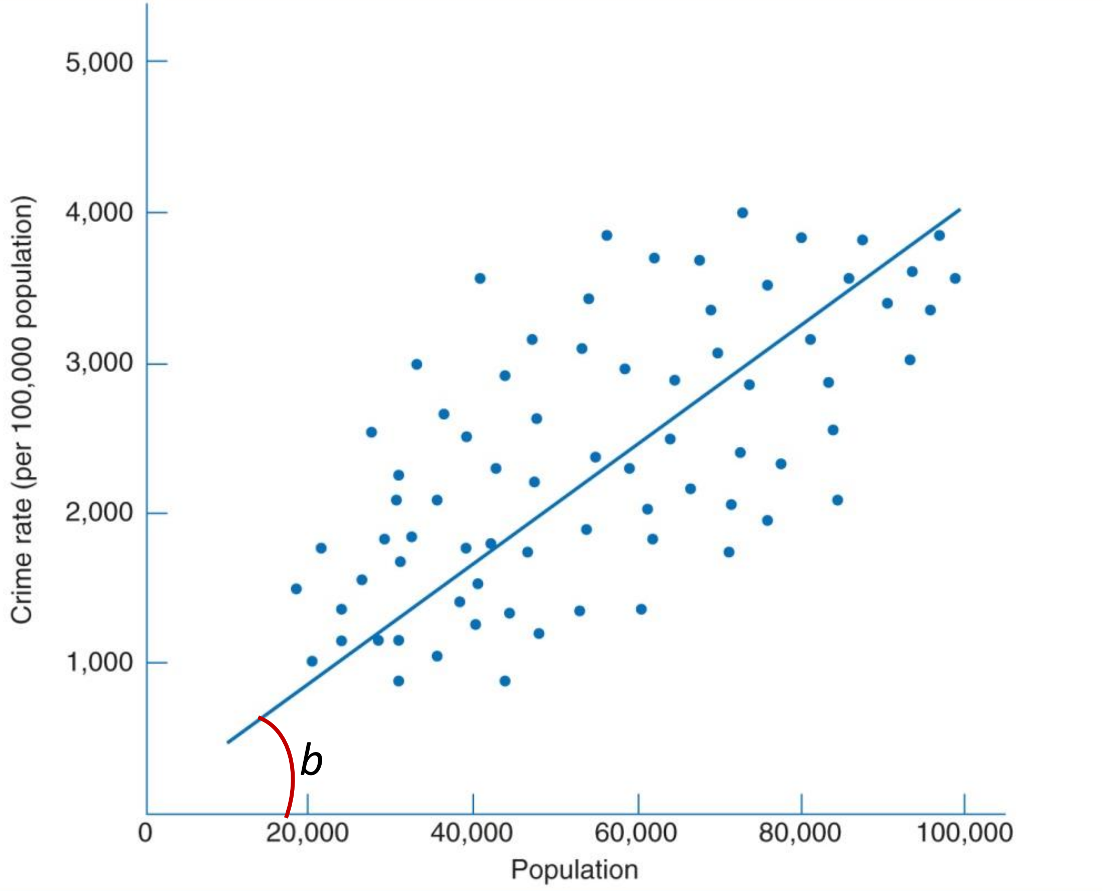
II.A. Linear regression...

- For multiple observations or values of X and Y, the equation changes to:

  $y = a + b(x)$ where

- x is any given value of X
- y is a predicted value of Y
- a is a constant (called Y-intercept)
- b is the slope of the regression line (regression coefficient).

A Scattergram of the Values of Two Variables with Regression Line Added (Hypothetical)



II.B. Multiple regression

- Y = B0 + B1*X1 + B2*X2 + … + BnXn + e
- The variables in the model are:
- Y = the dependent variable (outcome);
- X1 = the first predictor variable;
- X2 = the second predictor variable;
- X3 … Xn = control variables; and
- e = the residual error (unmeasured variables).

The parameters in the model are:

B0 = the Y-intercept;

B1 = the first regression coefficient;

B2 = the second regression coefficient

B3 … Bn = coefficients for other variables.

II.B. Multiple regression…

Regression Statistics	
Multiple R	0.665
R Square	0.442
Adjusted R	0.436
Standard E	5.899
Observatic	182

Model fit statistic

Number of observations

ANOVA

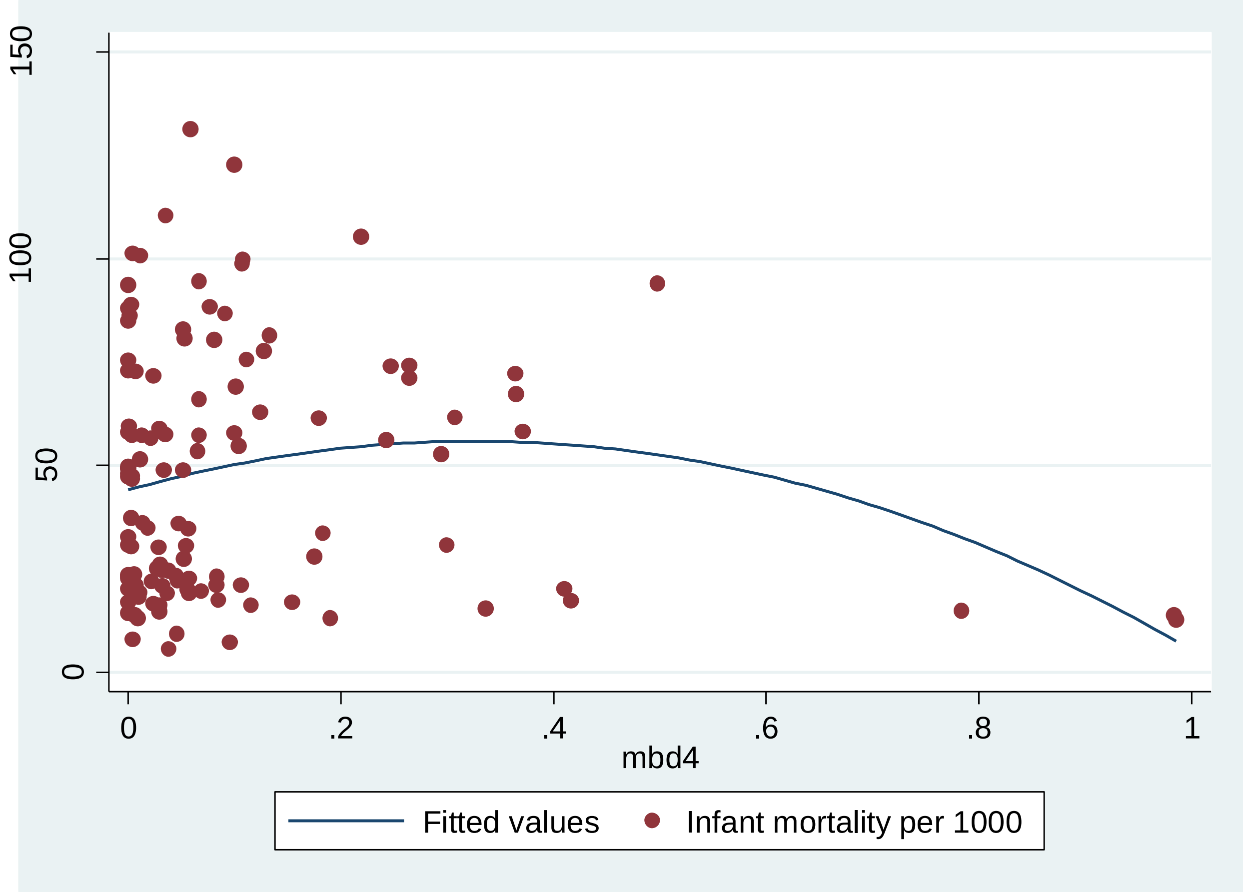
	df	SS	MS	F	Significance F
Regressior	2	4942.612	2471.306	71.019	0.000
Residual	179	6228.852	34.798		
Total	181	11171.464			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	63.532	1.222	52.007	0.000	61.122	65.943
HE	0.513	0.168	3.057	0.003	0.182	0.845
GDP	0.000	0.000	10.955	0.000	0.000	0.000

II.C. Curvilinear/polynomial regression

- Curvilinear regression allows relationships among variables to be expressed with curved geometric lines.
- Here, we fit a curved line within a linear model by using powers of our IV (e.g., a squared term).

Bivariate scatterplot showing the quadratic relationship between infant mortality and medical brain drain in 121 LMICs in the year 2004



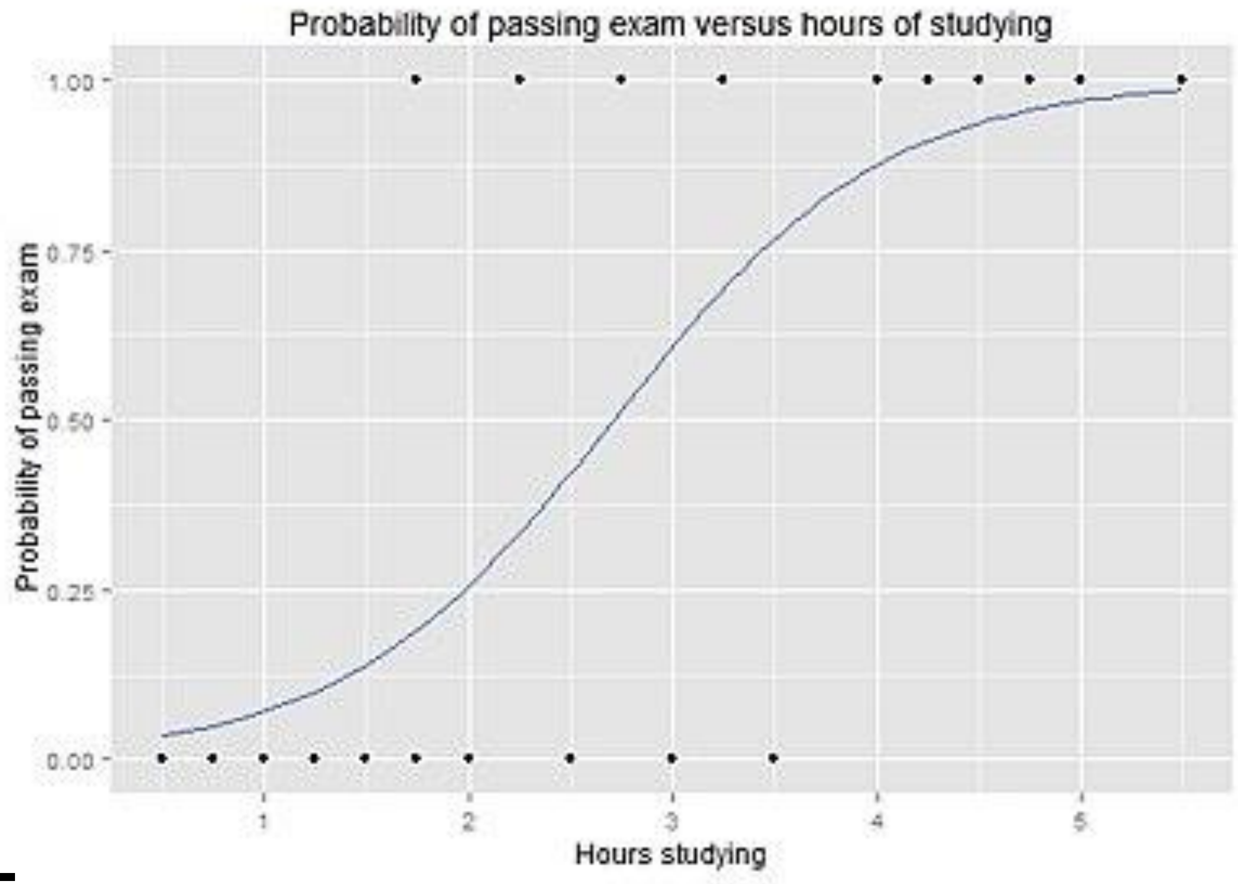
II.C. Curvilinear/polynomial regression...

Table 1. Fixed-effects regression of medical brain drain (4-year lag) on infant mortality in 121 LMICs over 1995-2008

	Infant mortality			
Variables	(1)	(2)	(3)	(4)
MBD (4-year lag)	-9.411	-9.896**	-184.012***	-24.840***
	(6.460)	(3.782)	(36.069)	(5.007)
Log of GDP per capita		-17.859***	-17.446***	-19.153***
		(1.184)	(1.175)	(1.208)
Remittances		-0.109**	-0.097**	-0.099**
		(0.034)	(0.034)	(0.034)
Log of primary gross enrollment		-27.120***	-26.775***	-26.180***
		(1.554)	(1.539)	(1.554)
Health expenditure		-1.588***	-1.428***	-1.489***
		(0.248)	(0.247)	(0.246)
MBD (4-year lag) squared			63.123***	
			(13.005)	
MBD and GDP interaction				0.003***
				(0.001)
Number of country/observations	121/1216	121/1216	121/1216	121/1216

II.D. Logistic regression

- If the DV is binary, we can fit the model using a logistic regression model.
- The odds of success are defined as the ratio of the probability of success over the probability of failure.

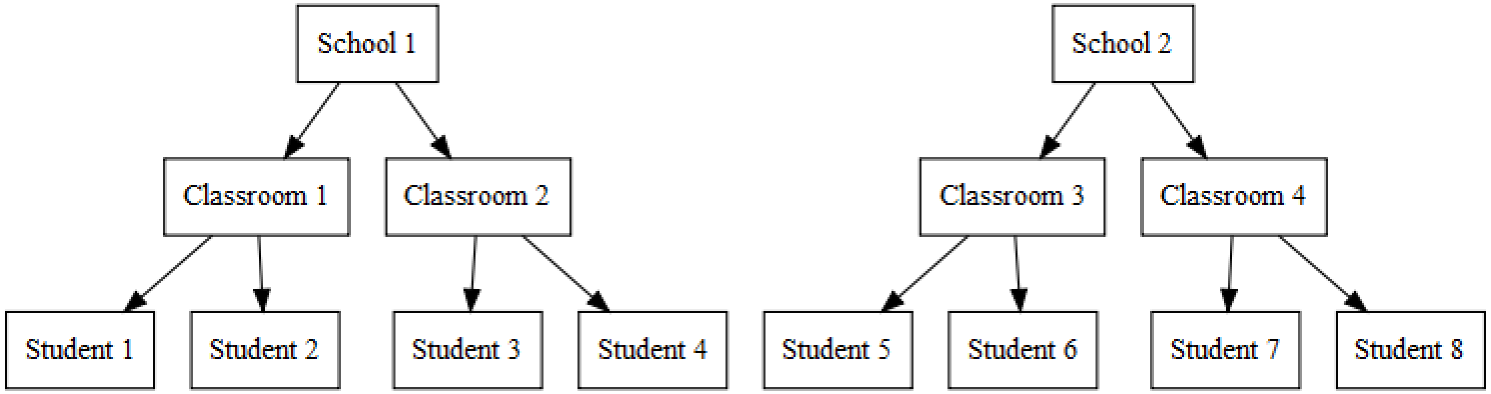


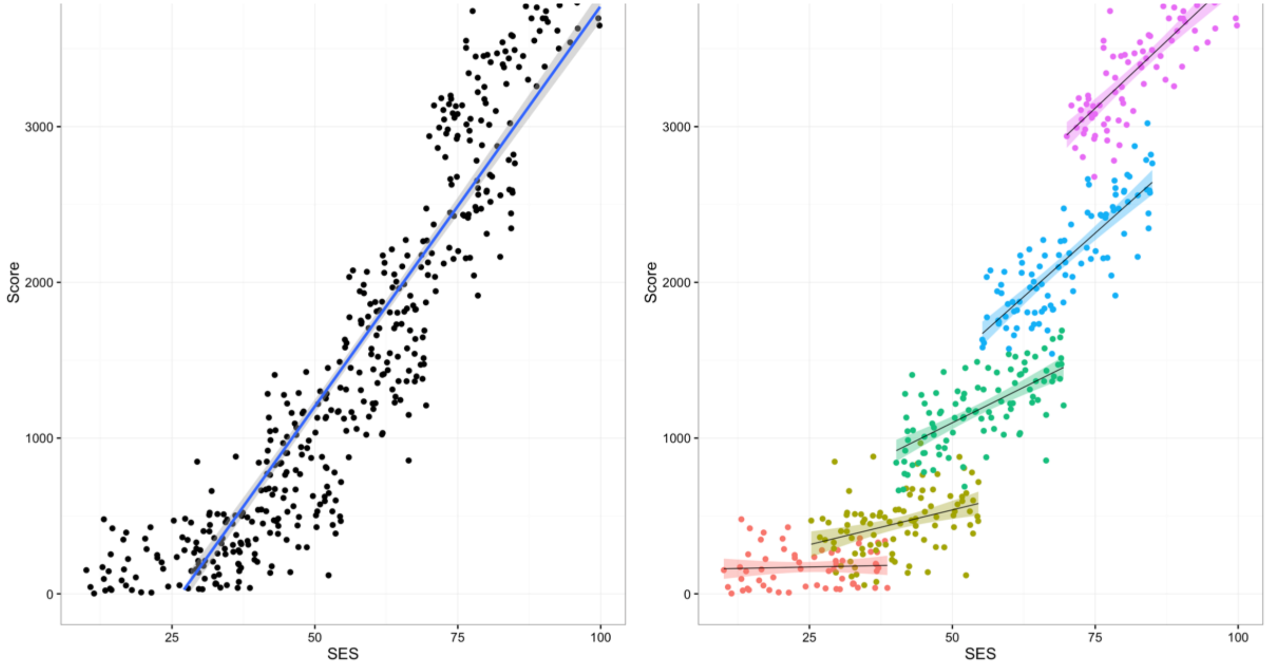
$$odds = \frac{p}{1-p}$$

$$logit(p) = \ln\left(\frac{p}{1-p}\right) \quad p = \frac{1}{1+e^{-logit(p)}}$$

II.E. Multilevel regression

- When our data is hierarchical, we cannot use one-stage linear regression.
- Stata calculates regression coefficients at each level, for example:
  - *mixed Score SES || school: || class:*





Berkeley
UNIVERSITY OF CALIFORNIA

III. Remarks

- Inferential statistics can be problematic when the sample is not drawn randomly, sample is small, there are missing values and non-responses.
  - Solution: Must meet the criteria/assumptions.
- Most of the time, a regression model only explains the correlation between two or more variables. A causal analysis requires suitable data and regression models.
  - Solution: Use longitudinal data and IV regression; but finding a suitable IV is hard.
- Not all statistical relationships derived from the sample can be generalized to the entire population.
  - Solution: Limit the interpretation only to the sample.