Georgian Implementation Science Fogarty Training (GIFT)

Ilia State University & Yale University

Summer Bootcamp in Implementation Science & Biostatistics

July 21–24, 2026
Tbilisi, Georgia

Yale Center for Analytical Sciences

Correlation and Regression

Different data types

X Independent	Y Dependent	Statistical Test
Categorical	Categorical	chi square
Categorical	Continuous	t-test, ANOVA
Continuous	Continuous	correlation and regression

Association Between 2 Continuous Variables

Correlation

- relationship between two variables where the magnitude of one variable changes as the magnitude of the second variable changes

- Y does not have to depend on X (i.e. do not distinguish between dependent and independent variable)

- A child's arm length and leg length are correlated, but neither "depends on" the other

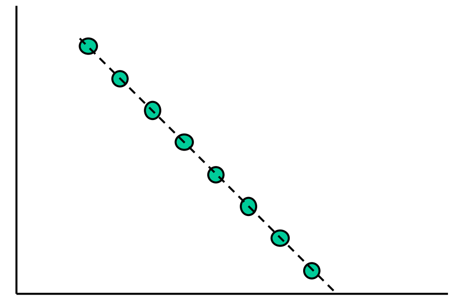
Pearson Correlation Coefficient

- Measures the linear relation between 2 variables

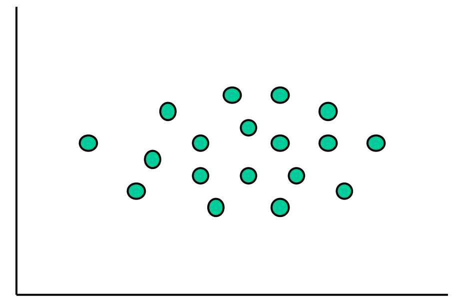
$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$$

Correlation Coefficient

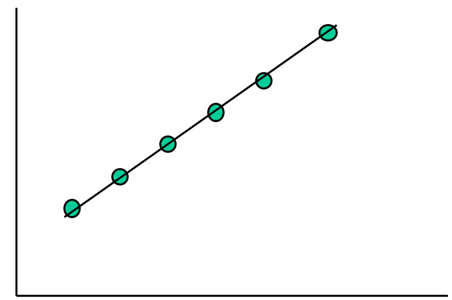
- Values between –1 and 1



r=-1



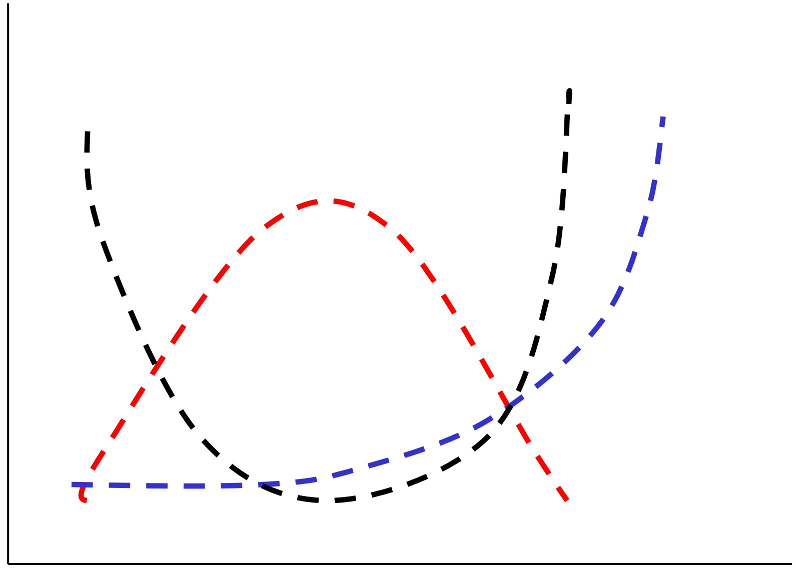
r=0



r=1

Correlation Coefficient

- Can only describe linear associations
  - Variables not always related in a linear fashion



Coefficient of Determination

- The amount (proportion) of variation in the dependent variable that is explained by the independent variable

$$r^2 = \frac{TSS - SSE}{TSS}$$

Correlation Coefficient: Hypothesis Testing

- Ho: $\rho=0$ (i.e. the correlation in the population is zero)
- Ha: $\rho\neq0$

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

- Under Ho, this statistic is distributed as the t-distribution with n-2 d.f.

Assumptions of Correlation

- Bivariate normality
- Independence of pairs of observations
- Linear relationship
- Random sampling

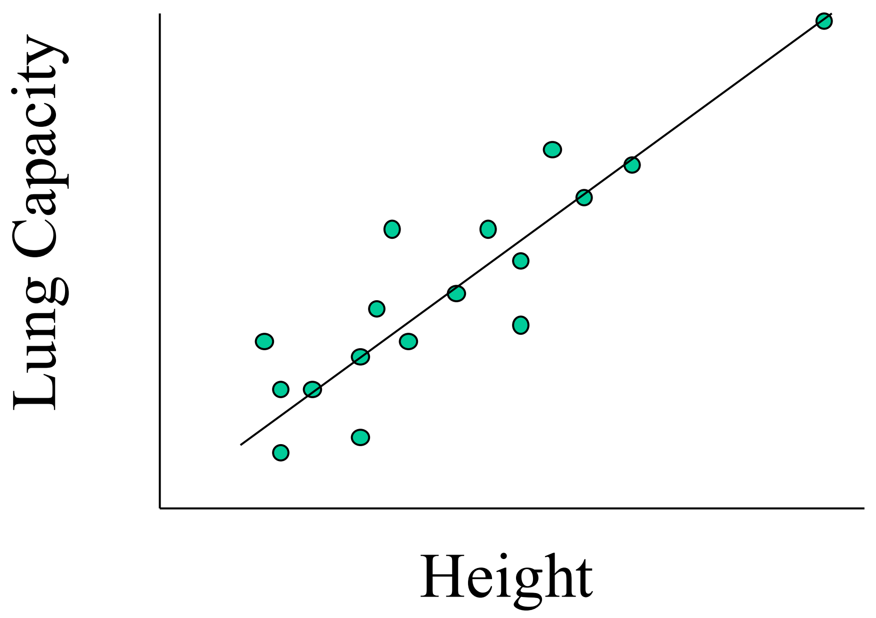
R EXAMPLE OF CORRELATION

Misuses of Correlation

- Correlation and repeated measurements
- Spurious correlations with time
- Restricted sampling
- Associating change with initial value or part to a whole
- Assessing agreement
- Cause and effect

Linear Regression

- Used to predict the value of a dependent variable from one or more independent variables



Simple Linear Regression

- Equation for a straight line

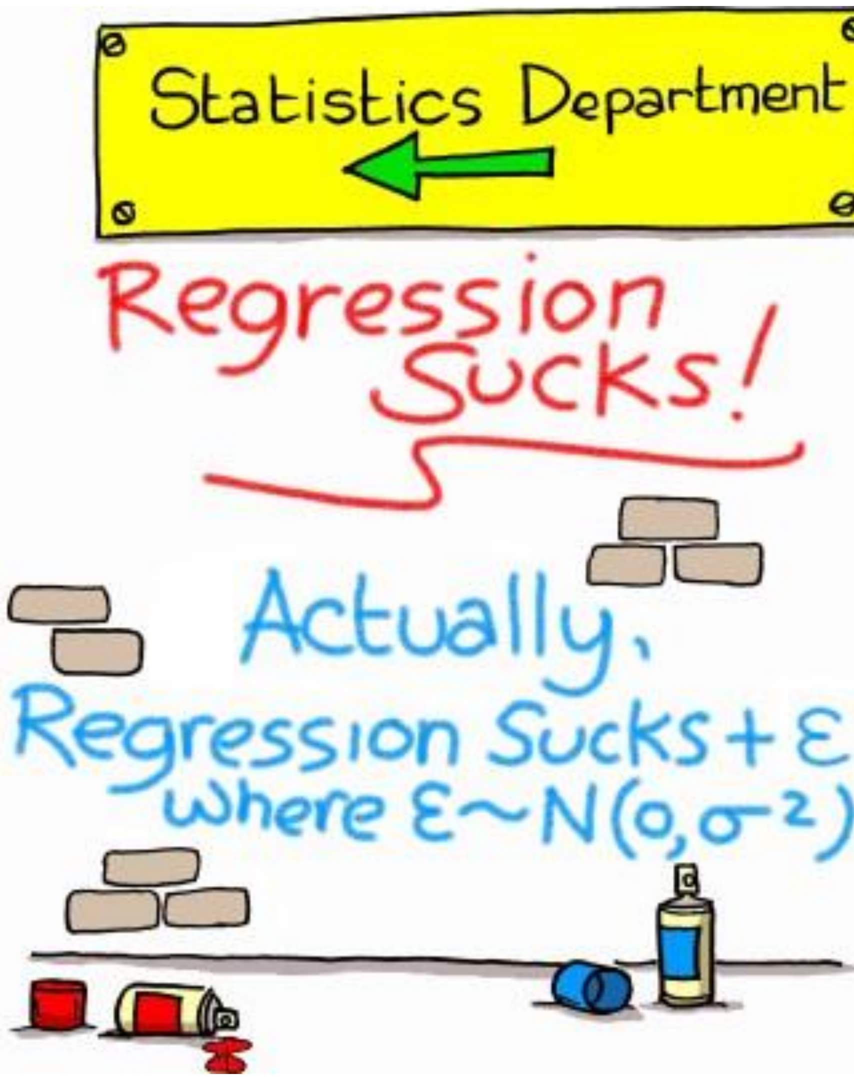
$$y = mx + b$$

slope intercept

- Probabilistic model

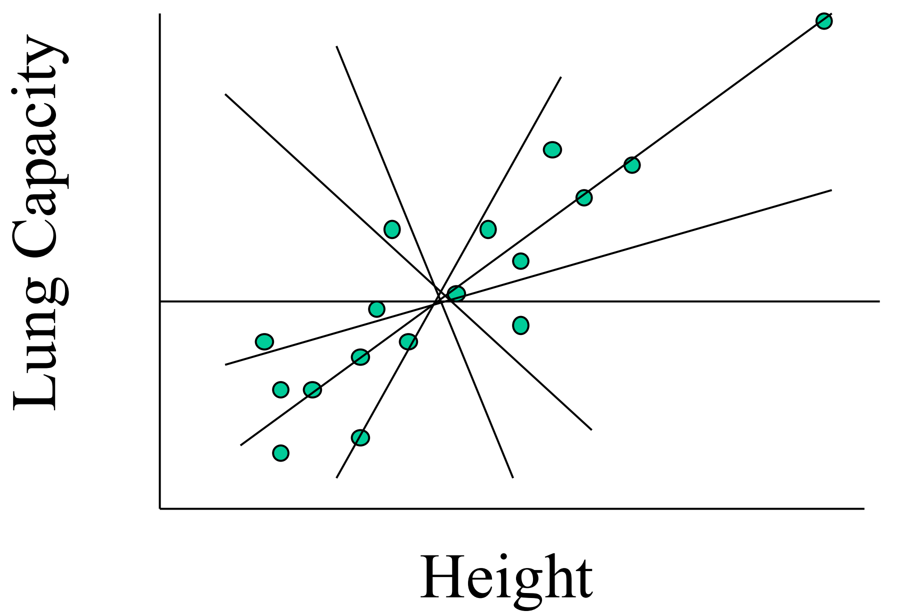
$$y = \beta_0 + \beta_1 x + \varepsilon$$

Random error (residual)


Statistics Department
Regression Sucks!
Actually,
Regression Sucks + ε
where ε ~ N(0, σ^2)

Simple Linear Regression

- Assume that random error on average will be 0 (i.e. sum of the residuals is zero)



Method of Least Squares

- There's only one line that minimizes the squared residuals

$$SSE = \sum \left(\underset{\text{obs}}{y} - \underset{\text{pred}}{\hat{y}} \right)^2$$

Assumptions for Linear Regression

- Normality – values of the outcome variable (y) should be normally distributed at each value of the predictor variable (x)
- Homoscedasticity – variability of y should be same at each value of x
- Linearity – relation between x and y should be linear
- Independence

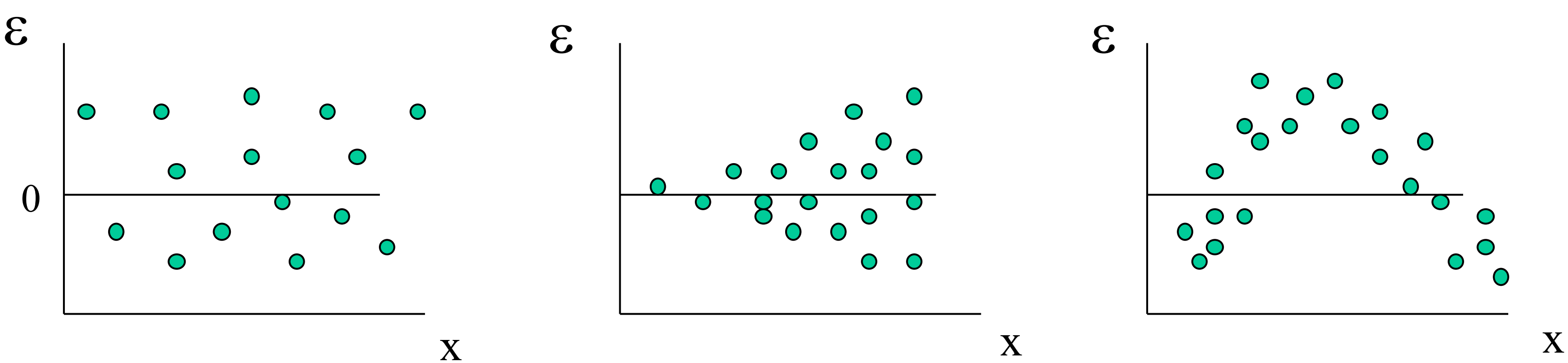
Inference About the Slope

- If x contributed no information to the value of y, what would the slope be?

- To test the significance of the slope:
  - Ho: $\beta_1=0$
  - Ha: $\beta_1\neq0$

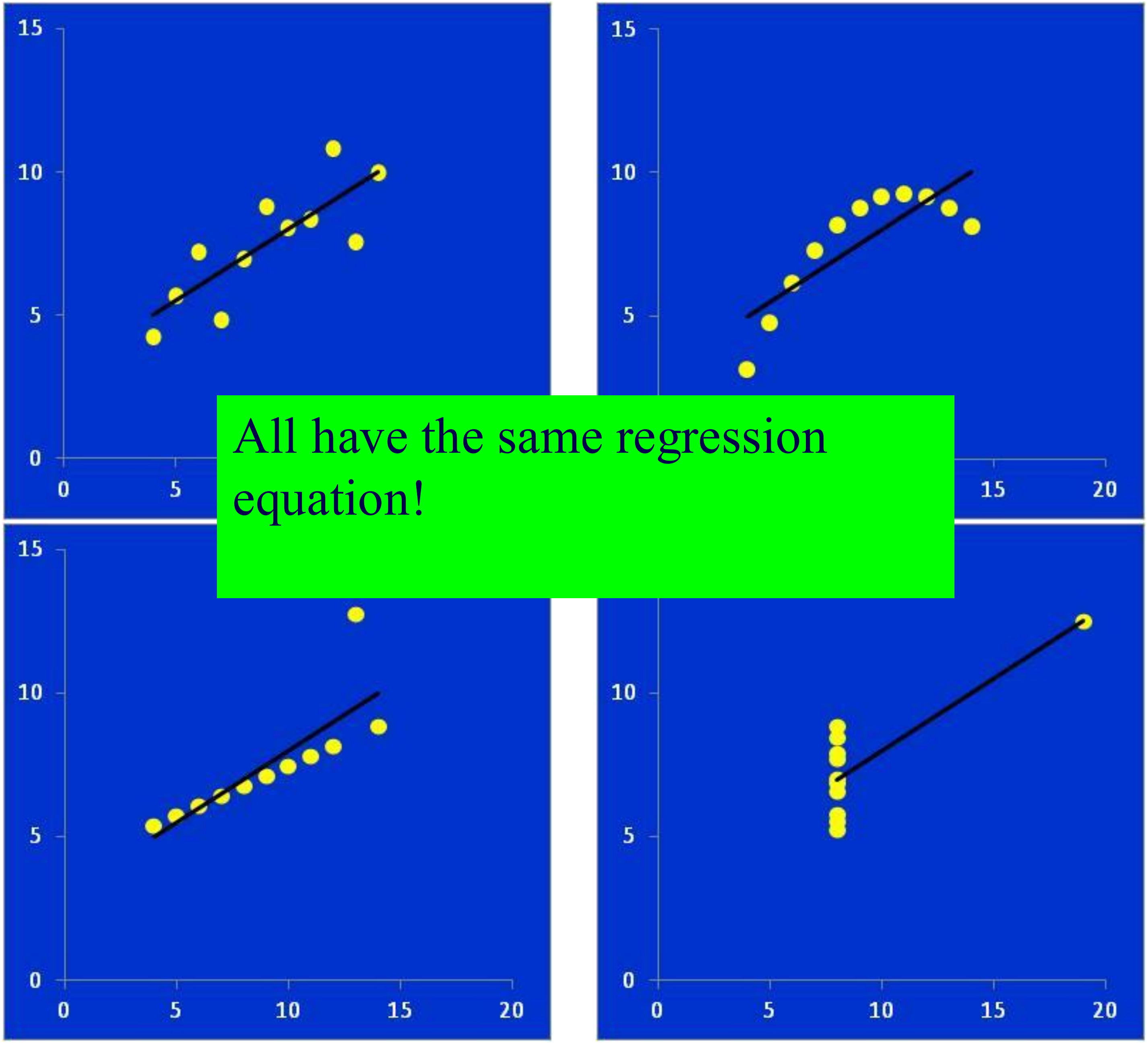
Evaluating the Assumptions

- Normality – plot stem and leaf or histogram of residuals
- Homoscedasticity and Linearity – plot residuals by predicted values or by x and look for patterns



Outliers and Influence

- Concern: omission of one or a few influential points could change a coefficient estimate and qualitatively effect conclusions
  - Outliers – in linear regression – values with a much different value for the dependent variable compared to what is predicted by the regression
  - High Leverage observations - outliers in the independent variable (or combinations of the independent variables) with the potential to exert undue influence on the regression coefficient estimate
  - Influential observations – exert undue influence on the regression coefficients

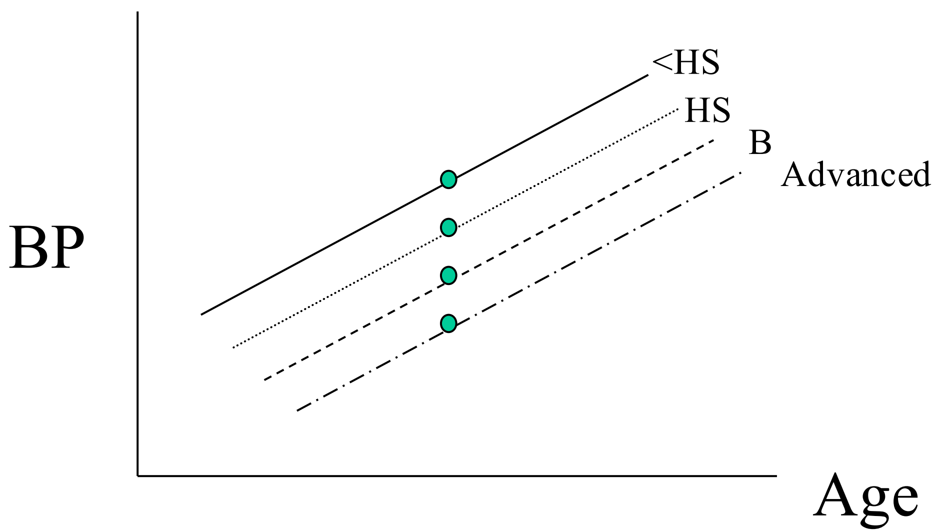


All have the same regression equation!

R EXAMPLE OF LINEAR REGRESSION

Multiple Linear Regression

- To determine the effect of one or more variables on an outcome variable while adjusting for one or more variables
- Example: Analysis of Covariance



Multiple Linear Regression

- General form of the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + \varepsilon$$

- Each one of the individual slopes represents the slope of a line as a function of an independent variable while keeping all other variables constant

Dummy Variables

- Create dummy (indicator) variables to include categorical variables in the regression equation
- For k levels, need k-1 dummy variables

	x1	**x2**	**x3**
<HS	0	0	0
HS	1	0	0
B	0	1	0
Advanced	0	0	1

Dummy Variables

$$BP = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

$$\mu_{<HS} = \beta_0$$

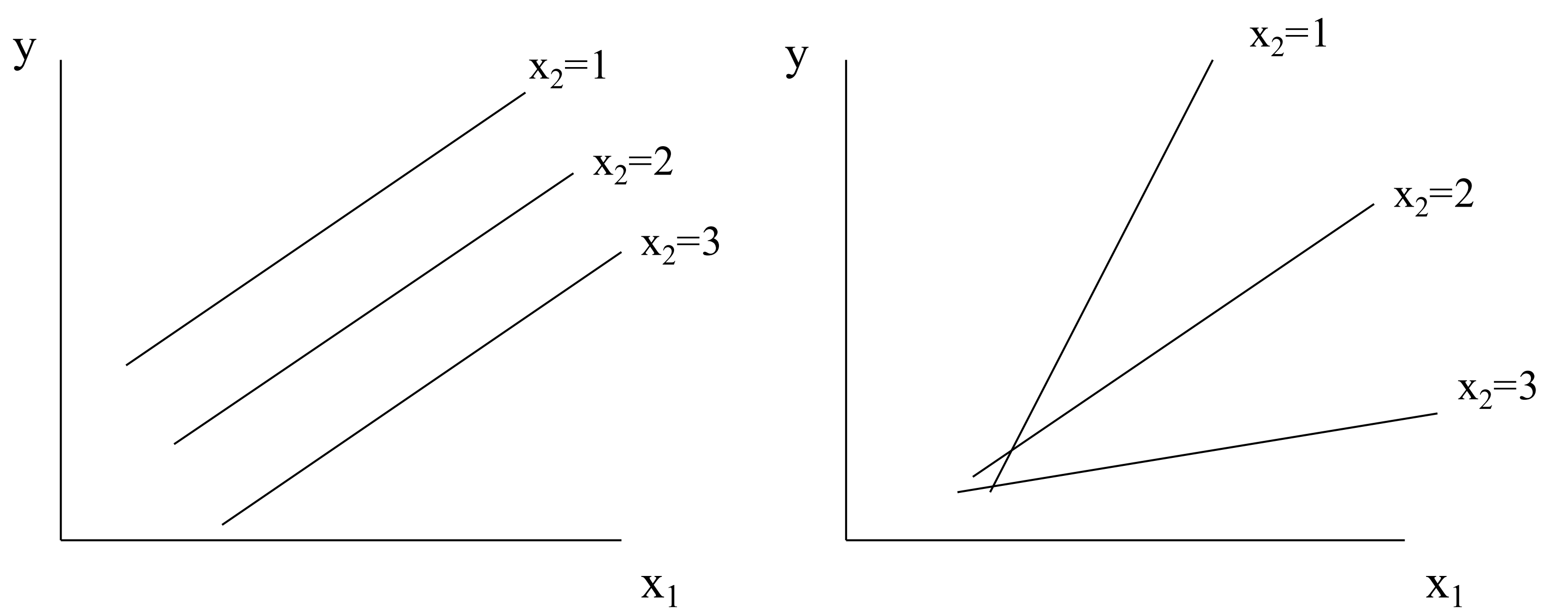
$$\mu_{HS} = \beta_0 + \beta_1$$

$$\mu_B = \beta_0 + \beta_2$$

$$\mu_{Ad} = \beta_0 + \beta_3$$

Interaction

- Is the effect of one independent variable on the outcome variable dependent on the level of another independent variable.



Interaction

- Interaction terms in multiple regression are specified by the inclusion of crossproducts

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Model Selection

- How do you choose which variables should be in the model?

  - Depends on the Purpose of the Model
    - Prediction
    - Evaluating an independent variable of primary interest
    - Identifying important independent variables associated with an outcome

Model Selection

- *Prediction* – goal is to minimize prediction error (how well the model predicts an outcome for a new observation)
  - bias-variance tradeoff – larger model vs smaller model
    - Estimate prediction error
      - Cross-validation or other measures
    - Screen candidate models to find one that minimizes

Model Selection

- Evaluating an *independent variable* of primary interest
  - Rule out confounding
    - Do not include independent variables that are:
      - essentially alternative measures of either the outcome or the predictor
      - those hypothesized to mediate the effect
    - Include independent variables that are very well established causal antecedents of the outcome regardless of statistical significance
    - If cannot define potential confounders *a priori* may eliminate on statistical grounds
      - *Automated selection procedures*

Model Selection

- Identifying *important independent variables* associated with an outcome - exploratory
  - very difficult because now making inferences about most or all of the variables
    - Overfitting and false positives
    - Evaluation of effect modification
    - Collinearity
    - Mediation

Model Selection

- Identifying important independent predictors of an outcome- cont'd
  - Recommendations
    - *Rule out confounding*
      - *face validity*
      - *Do not exclude variables based on a strict (eg. 0.05) observed significance*
    - *Eliminate poorly motivated or implausible interactions from consideration*
    - *Do not correct for multiple comparisons*
    - *Cautious interpretation*

R EXAMPLE OF MULTIPLE REGRESSION

"The Least-Squares Line didn't fit my data, so I decided to try the Hollywood Squares line."