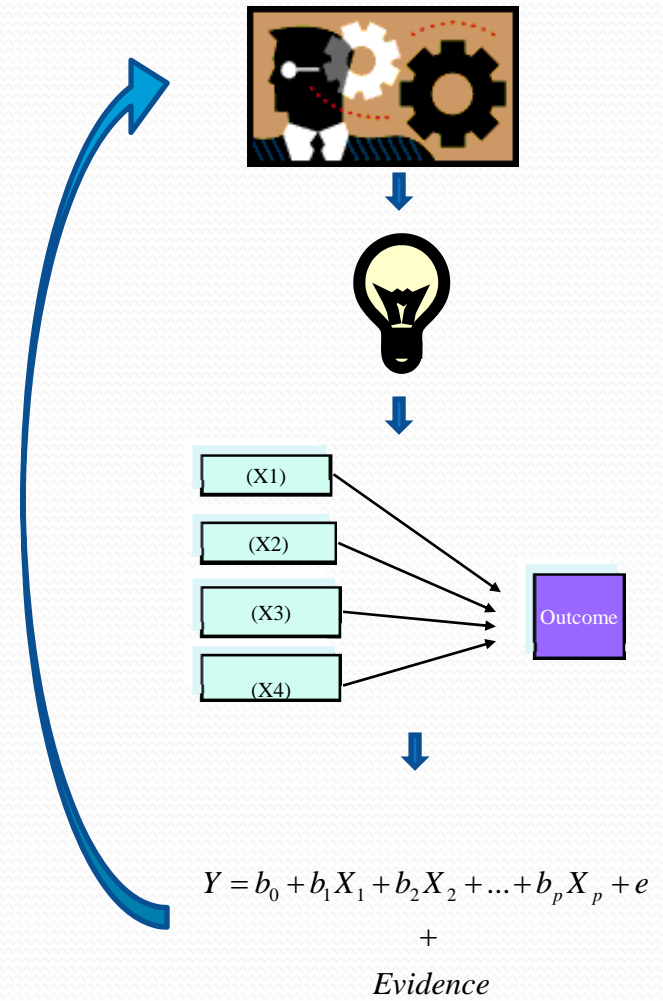


# Linear Least Squares Regression

Summary

# Goal

- The goal of linear regression is to understand the uncertainty in some aspect of human nature based on its association with other facets of experience
- Distinction: simple correlation vs. scientific understanding
  - Sample vs. Inference to a population
  - Anecdote vs. Data
- The *model* is a reflection of your theory about what potentially explains that aspect of human nature
- Evidence comes from data i.e. information from the real world
- Regression offers a possibility to answer the question of whether the evidence matches your theory that relies on more than 'well it seems that way to me'



# Products of the Analysis

- How can one tell whether the model fits?
  - Variability in the outcome/response/behavior explained by the model
    - Statistic:  $R^2$
  - Errors in prediction
    - Statistic: Residual standard error
  - Other fit indices for model comparison (e.g. Bayesian Information Criterion)
- So it fits, what about the details?
  - Which predictors are important?
    - Statistic: raw coefficient
  - Which predictors are *most* important?
    - Statistic: Partial, semi-partial, average semi-partial correlation, Pratt's, Dominance analysis
- How good are our guesses?
  - Range of possibilities
    - Statistic: confidence intervals for  $R^2$ , coefficients, etc.

# Validity of the Procedure

- OLS Regression isn't applicable for inference beyond the sample if the assumptions are not met
- Primary assumptions
  - Normality: of the distribution residuals
  - Homoscedasticity: Residual variance is constant about the regression surface
  - Independence of observations: One case's score is not influenced by another's
  - Reliable measures: Capture the true variance in the construct measured
  - Linearity: Predictors are linearly related to the DV
  - Correctly specified model: All relevant predictors, no unnecessary ones, interactions if needed, no indirect effects, etc.
- Additional concerns
  - Lack of outliers: extreme in terms of model fit, not the individual predictors
  - Overfitting: inherent bias
  - Collinearity among predictors: predictor variance is not already accounted for by other predictors

# Consequences of Using an Invalid Procedure

- The consequences all have tremendous impact on the theory that formed the basis for investigating this aspect of human nature and how one would assess the theory's potential to explain cognition, behavior, personality etc.
- Lack of linear association between predictors and outcome, model misspecification etc. means that you have the wrong *theory*. Yep, your whole idea is off.
- Biased, inefficient coefficients due to poor reliability, collinearity etc. lead to an incorrect interpretation regarding your *theory*.
- Outliers may suggest you aren't able to apply your *theory* to the entire population you drew from
- Overfitting implies you are overconfident in your *theory*

# Solutions to Analytical Problems

- There are many, many approaches to deal with problems, and the excuse that they aren't easy to pull off went away about 20 years ago.
- Some examples
  - Model Misspecification: add more relevant predictors, path analysis
  - Curvilinearity: Polynomial regression, Interactions
  - Outliers: Quantile regression, robust/resistant regression
  - Lack of normality: use bootstrap interval estimates, variable transformation

# Validity of the Model

- Science entails prediction and this requires new data.  
We can:
- Replicate
  - Fit the model to whole new data set (potentially huge cost in time and money)
- Simulate
  - Fit the model to simulated data that looks like the original
- Split
  - Randomly split the original data into parts, form the model with one, test it on the others.



# MR: Basic Guidelines/Considerations

- Is it appropriate for your theoretical model?
  - Example: a single MR cannot get at indirect effects
- Exploratory or Confirmatory?
  - A wide variety of approaches are available for the former, however they are *useless* if models are selected solely by p-values and still useless without subsequent validation<sup>1</sup>
  - Example of a more confirmatory approach would be sequential/hierarchical regression, but it offers little more than presentation organization in the results section, and even though subsets are tested until the full model is realized, the approach does not tell you which model is best among subsets or full model
- Prediction vs. Explanation
  - A balanced approach is best for psych considerations

# MR: Basic Guidelines/Considerations

- Reliability
  - Very straightforward, strong and positive correlation between reliability and quality of analysis
  - MR assumption is technically having *perfectly* reliable measures. You might want to get in the ballpark.
- Initial data examination
  - If this is not done adequately then there is no point in doing regression. Know your data well or you'll look silly with the analysis.
- Model fit
  - $R^2$ : what's large? How do you normally gauge things in percents? What do you typically see in that area of research?
  - Residual standard error: can only make sense if you know the scale of the DV
  - Statistical sig: marginal utility. Is  $R^2$  different from zero?
- Generalization
  - Variable importance
    - If you want to say in your conclusion "Results suggest variable X is more important in predicting Y than variable Z" additional steps must be taken to show that X is statistically different from Z in its contribution
    - If you do not take those steps, all you can say is "*In this sample*, X had a larger stat than Z" with a "I have no idea how it'd turn out in another sample" qualification in your discussion of limitations (of course in the time you write that sentence you potentially could have done the analysis).
  - Assumptions
    - If they are not tested, you essentially have to qualify every statement in the discussion/conclusion with "In this sample", as you have done nothing to justify any generalization beyond it. Implying that you can generalize is academically dishonest.
  - Validation
    - You're not really doing science if there is no predictive aspect to the study. Only with extremely small samples is this not viable.

# MR: Basic Guidelines/Considerations

- Data considerations
  - MR is flexible enough to handle any type of predictor
  - Extendible to different outcomes, data situations.  
Examples include:
    - Categorical outcome
      - Logistic regression
    - Count outcome
      - Poisson regression
    - Censored data (e.g. ceiling/floor effects)
      - Tobit regression

# Modern approaches

- Modern approaches are useful for:
  - Dealing with data problems- violating assumptions, outliers
  - Answering different, perhaps more interesting questions
- Robust or nonparametric techniques used to deal with data problems rarely require new understanding in interpretation, but if you don't know OLS well they aren't going to make sense either
  - If a robust estimate of a coefficient is confusing, it's because you don't understand a coefficient in general
  - Example: if a bootstrapped CI for a coefficient is confusing to you, it's because the CI for the coefficient is confusing to you.
- However, there are also methods now computationally feasible that allow us to get more out of our data
  - Examples: Model exploration techniques, Quantile regression

# Alternative Approaches

- Practically innumerable, but a few examples:
- Within OLS
  - More confirmatory: Sequential regression
  - More exploratory: a whole host of procedures are available—stepwise, all subsets, model averaging
  - Interactions
- Extension to a *Generalized* Linear Model
  - Logistic regression, Survival regression, Time Series
- Multivariate
  - Path analysis, Partial Least Squares
- Contextual
  - Multilevel modeling
  - Bayesian

# Summary

- Linear regression is a very flexible tool, and the basic model can be applied in a variety of ways
- Though the theoretical model being tested might be relatively simple, a well done linear regression is not a simple endeavor
- Modern approaches allow us to deal with data problems, as well as use the analysis in new ways to capture more of what the data has to offer