Types of Variables in Statistical Modeling

Brady T. West

Review: Types of Variables when Summarizing Data

Categorical variables: take on small number of discrete values

Examples: gender, race/ethnicity, political party preference, region, binary indicators of events

Asked: Are the categories ordered in any way, or are they simply discrete values?

Review: Types of Variables when Summarizing Data

Categorical variables: take on small number of discrete values

Examples: gender, race/ethnicity, political party preference, region, binary indicators of events

Asked: Are the categories ordered in any way, or are they simply discrete values?

Continuous variables:
take on many possible values

Examples: height, age, income, blood pressure.

Asked: What does the distribution look like (shape, center, spread)? Is the variable normally distributed?

Classifying Variables for Model Fitting

Dependent Variables (DVs)

Other names: outcome, response, or endogenous variables, or variables of interest

Model distributional features of these <u>variables of interest</u> as a function of **independent variables**
(i.e., their distributions **depend on** the values of these other variables)

Classifying Variables for Model Fitting

Dependent Variables (DVs)

Other names: outcome, response, or endogenous variables, or variables of interest

Model distributional features of these variables of interest as a function of **independent variables**
(i.e., their distributions **depend on** the values of these other variables)

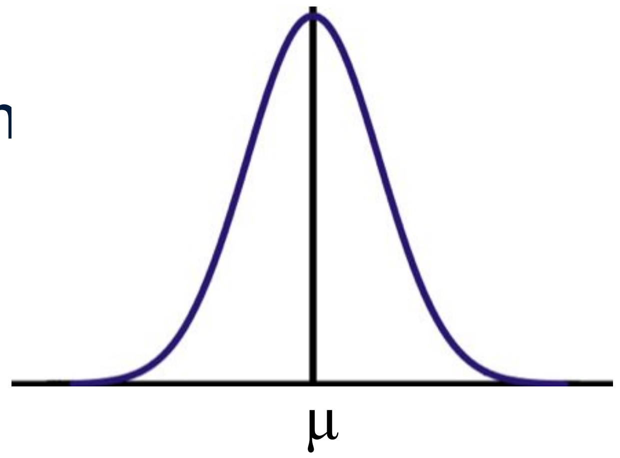
Independent Variables (IVs)

Other names: predictor variables, covariates, regressors, or exogenous variables

When fitting models, we examine the distributions of dependent variables **conditional** on the values of these independent variables

Dependent Variables (DVs)

- "Model" the DV, **which is a variable of primary interest,** as a function of other theoretically relevant IVs
  → **our research question defines all of these variables!**
- Involves selecting reasonable distribution for DV (e.g., norm
  and defining parameters of that distribution (e.g., the mean)
  as a function of (or conditional on) the IVs



- DVs could be continuous, categorical, binary, etc.

Example: Assume blood pressure is normally distributed, where mean **depends on** a person's age, BMI, and gender.

Independent Variables (IVs)

- **Theoretically relevant predictors** of dependent variables
  Interested in estimating relationships of IVs with DVs

Independent Variables (IVs)

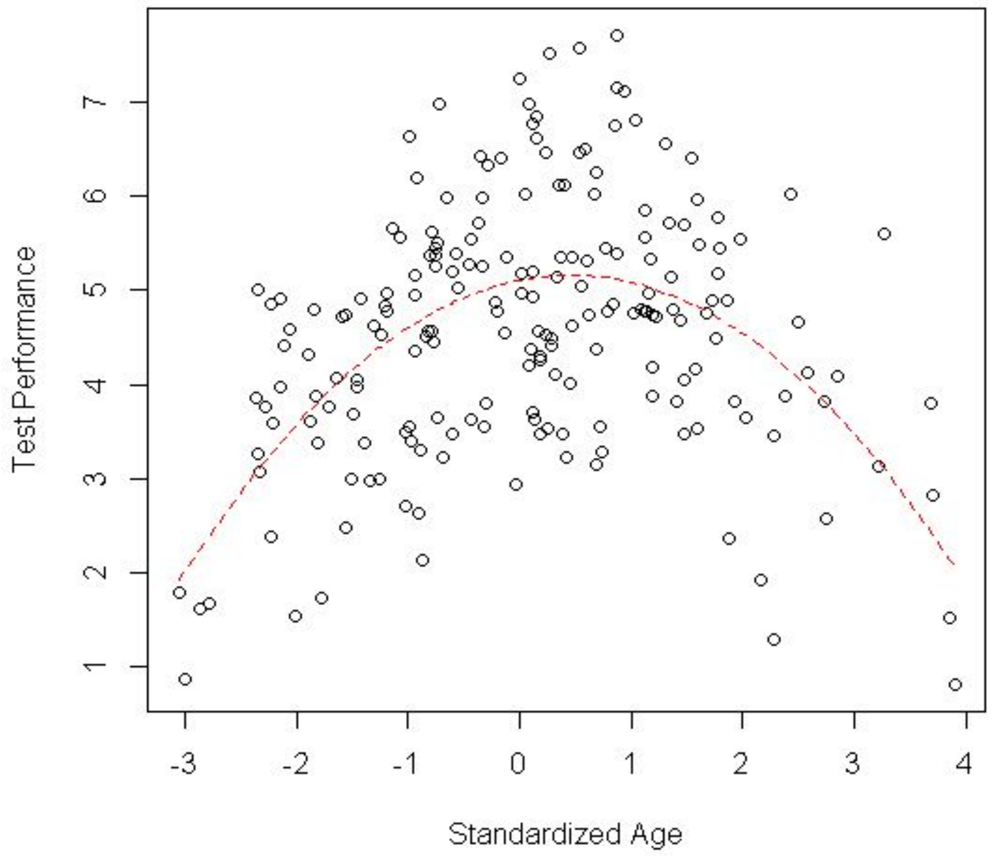
- **Theoretically relevant predictors** of dependent variables
  Interested in estimating relationships of IVs with DVs
- Might be **manipulated by an investigator**
  Randomized experiment: cases randomly assigned to an intervention or a control group ("group" is the predictor)

Independent Variables (IVs)

- **Theoretically relevant predictors** of dependent variables
  Interested in estimating relationships of IVs with DVs
- Might be **manipulated by an investigator**
  Randomized experiment: cases randomly assigned to an intervention or a control group ("group" is the predictor)
- Could simply be **observed**
  Observational studies: hard to make causal inference about relationships

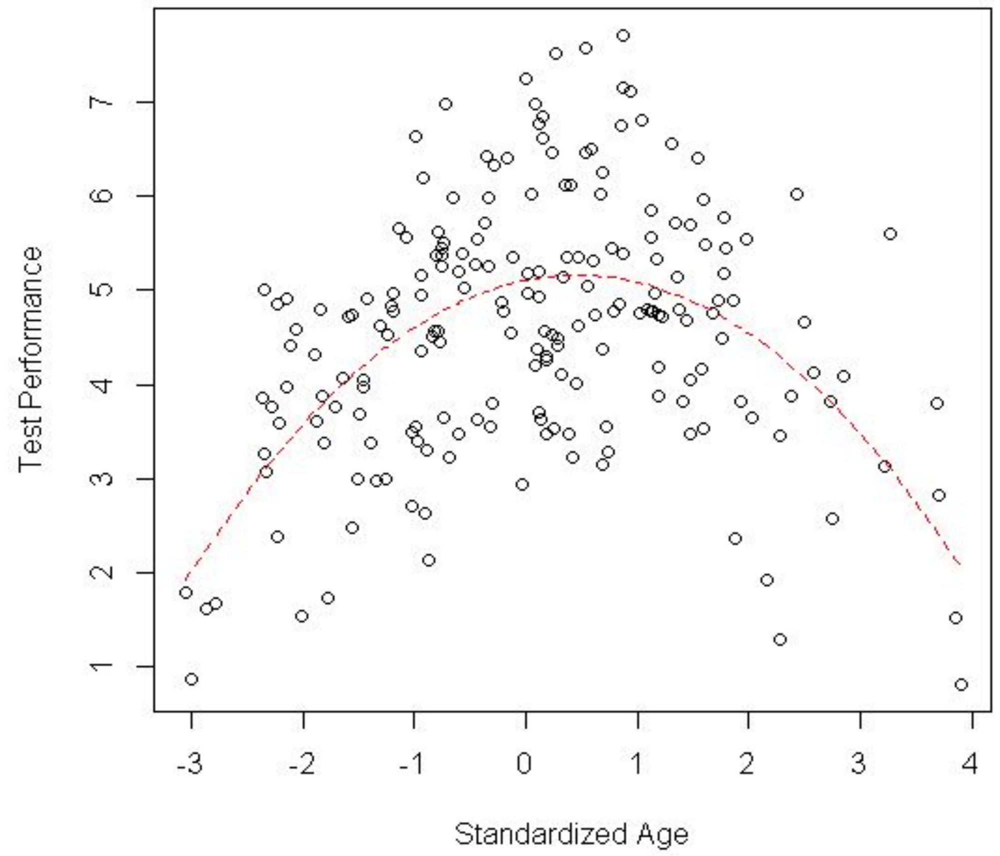
Independent Variables (IVs)

- If IVs **continuous**: estimate functional relationships of those IVs with distributional features of the DVs
- If IVs **categorical**, compare groups defined by the categories in terms of distributions on the DV



Independent Variables (IVs)

- If IVs **continuous**: estimate functional relationships of those IVs with distributional features of the DVs
- If IVs **categorical**, compare groups defined by the categories in terms of distributions on the DV



Best Practice

Avoid estimating *functional* relationships of categorical IVs (e.g., race) with DVs, since actual values of categorical IVs may not have any numerical meaning!

“Control” Variables

Goal: Estimating parameters describing **relationships** of IVs with DVs

“Control” Variables

Goal: Estimating parameters describing **relationships** of IVs with DVs

- **Randomized study designs:** attempt to ensure randomized groups are **balanced** with respect to other **“confounding” variables** that may have negative impact on estimation of relationship of group with DV

“Control” Variables

Goal: Estimating parameters describing **relationships** of IVs with DVs

- **Randomized study designs:** attempt to ensure randomized groups are **balanced** with respect to other **“confounding” variables** that may have negative impact on estimation of relationship of group with DV

- **Non-randomized (observational) designs:** groups that define IV **may not be balanced**. *Example:* males generally may weigh more than females, and an analysis looking at relationship of gender with DV related to weight may not yield clear estimate of the gender – DV relationship.

“Control” Variables, cont’d

When fitting models, include several IVs
→ effectively adjusting for this **confounding problem**

"Control" Variables, cont'd

When fitting models, include several IVs
→ effectively adjusting for this **confounding problem**

Example: Interested in comparing distribution of blood pressure (DV) between males and females, and weight is related to blood pressure.

→ Include weight as a **control variable**, and make inference about the gender – blood pressure relationship **given a value for weight**

Missing Data

Key:

Before fitting models …conduct simple descriptive and bivariate analyses of DVs and IVs

→ check for **missing data** on both DVs and IVs

Missing Data

Key:
Before fitting models …conduct simple descriptive and bivariate analyses of DVs and IVs
→ check for **missing data** on both DVs and IVs

Listwise Deletion =units of analysis with **any missing data** on **any of IVs or DVs** dropped from analysis!

If **cases dropped are systematically different** from cases analyzed
→ introduce **bias in estimated IV- DV relationships!**

Missing Data

- If units with missing data identified in descriptive analyses, can **compare units with missing data** *(would be dropped)* **to units with complete data** *(would be retained)* in terms of distributions on variables that are fully observed:

  Are there differences?

Missing Data

- If units with missing data identified in descriptive analyses, can **compare units with missing data** *(would be dropped)* **to units with complete data** *(would be retained)* in terms of distributions on variables that are fully observed:

  Are there differences?

- If evidence of differences → other techniques can be used (e.g. **imputation**) to address missing data issue

What's Next?

- **Implications of study design** (cluster sample, longitudinal study, cross-sectional convenience sample, volunteer clinical trial) on types of models to be fitted.

What's Next?

- **Implications of study design** (cluster sample, longitudinal study, cross-sectional convenience sample, volunteer clinical trial) on types of models to be fitted.
- Recognize that study designs affect properties of collected data, and **models need to reflect these properties**!

Example: Repeated measures of DV from the same person over time will be correlated!