Choosing the Correct Regression Model from Types of Variables

AI-05 Yujin Kim

Types of Dependent Variable (DV)

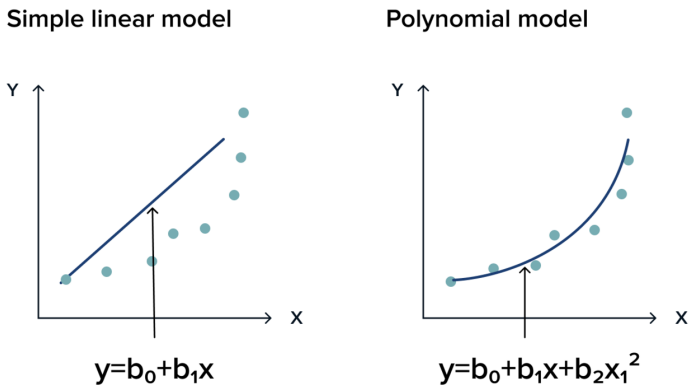
Continuous DV

- continuous scale (e.g., price)

model:

- Linear regression (simple, multivariate, polynomial)
- Ridge, Lasso regression

- most common, straightforward to use
- understand the mean change in a DV given a one-unit change in each IV
- estimate parameters by minimizing the sum of the squared error (SSE)



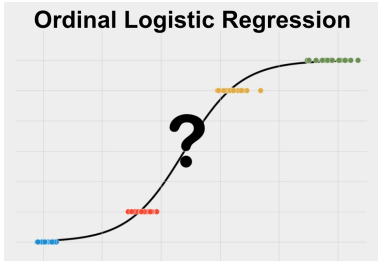
Categorical DV

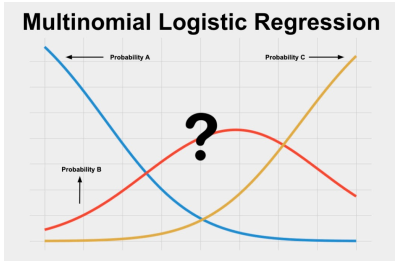
- categorical scale

model:

- Binary logistic regression (a Binomial probability distribution, e.g. pass/fail)
- Ordinal logistic regression (an order of outcomes, e.g. movie review)
- (multi)nominal logistic regression (no order of preference or ranking, e.g. race)

- apply maximum likelihood estimation which assumes probability distribution given the observed data

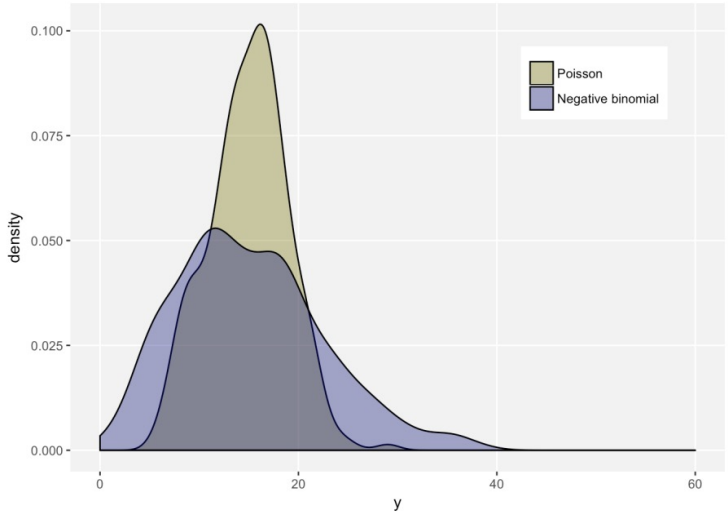




Count DV

- discrete, non-negative values
e.g., # of occurrences of an event

- do not meet the LM's assumptions (e.g., the errors follow a normal dist.)
- skewed distribution



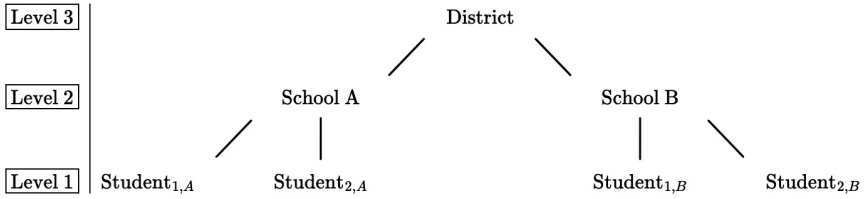
model:

- Poisson regression (count)
- Negative binomial regression (count + overdispersion)
- Zero-inflated models (zero-part, regression part)

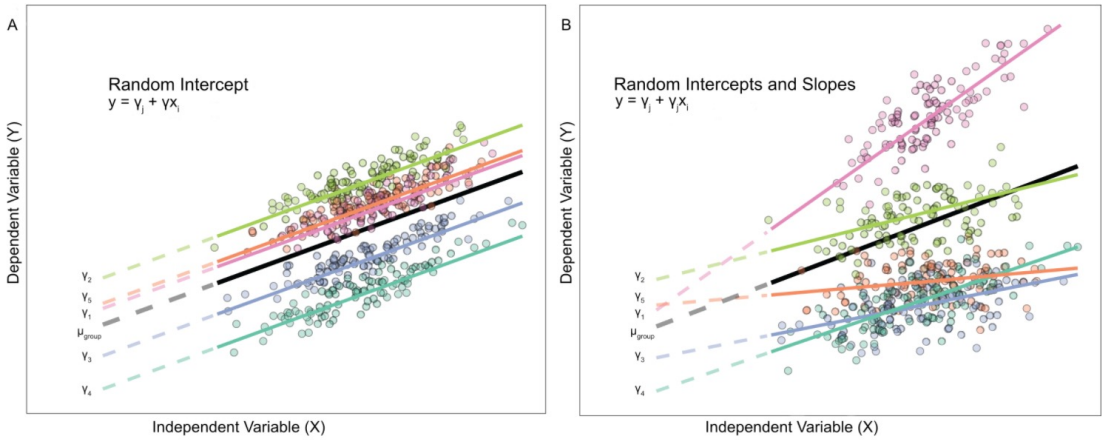
Types of Data Structure

Hierarchical or Multilevel Data (also linear-mixed model)

- assumes there is non independence (nested or grouped)
- parameters vary at more than one level

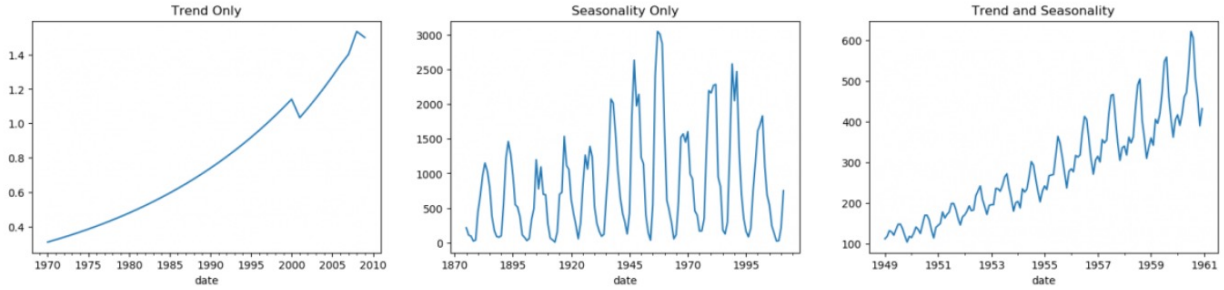


- estimate generalizations of linear models
- must decide fixed and random effects (for intercept or slope)
- DV must be the lowest level of analysis



Time Series Regression

- assumes evenly spaced time series data (if not, should be in this format using interpolation, aggregation etc.)
- need to decompose patterns of time series
  - trend: there is a long-term increase or decrease in the data, "changing direction"
  - seasonality: affected by seasonal factors (of the year, the day of the week, etc), with regularity. e.g., ice cream sales in summer
  - cycle: patterns of rises and falls that are ***not*** a fixed frequency. e.g., socio-economic factors



- also need to remove other components for time series model (e.g.)
  - stationarity : the values of the series is NOT a function of time
    -> should make non-stationary stationary
  - autocorrelation : the correlation of the series with its previous values
  - moving average (MA) : removing seasonality and/or trend
  - ARIMA