WED NOV 29

Quantifying uncertainty

Final due Fri next week (Available by this Fri 11:59 PM)

MODEL: Suppose have categorical variable (neighborhood)

DATA housing costs model as function of size neighborhood

$$y_i = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_k x_{i,k} + \varepsilon_i$$

CATEGORICAL VARIABLE

r categories (r neighborhoods)

For each category (neighborhood) $j = 1, \dots, r$

$$\delta_{i,j} = \begin{cases} 1 & \text{if } i^{\text{th}} \text{ unit is in neighborhood } j \\ 0 & \text{o/w} \end{cases}$$

neighborhoods river, bethel, churchill, campus, whitaker

house bethel

river	bethel	churchill	campus	whitaker
0	1	0	0	0

$$\boxed{\begin{aligned} y_i = \beta_{\text{river}} \underline{\delta_{i,\text{river}}} + \beta_{\text{BETHEL}} \underline{\delta_{i,\text{BETHEL}}} + \dots + \beta_{\text{WHIT}} \underline{\delta_{i,\text{WHIT}}} \\ + \underline{\beta_{\text{area}}}\, \underline{\text{area}_i} + \beta_{\text{year}}\, \text{year}_i + \varepsilon_i \end{aligned}}$$

MODEL: parameters are β_{river}, $\beta_{\text{BETHEL}}, \dots$ β_{area} β_{year}

estimate unknown parameters $\beta_{\text{river}}, \dots$ via least-squares

$\mathbb{E}\, y_i = \beta_{river}\, \underline{\delta_{i,river}} + \beta_{BETHEL}\, \underline{\delta_{i,BETHEL}} + \cdots + \beta_{WAIT}\, \delta_{i,WAIT}$
$+ \beta_{area}\, \underline{area_i} + \beta_{year}\, year_i$

If i^{th} data point belongs to river road neighborhood what is expected rent?

$$\mathbb{E}\, y_i = \beta_{river} + \beta_{area}\, area_i + \beta_{year}\, year_i$$

Center the numerical values at their average's

area $\longrightarrow$ $\underbrace{area - \overline{area}}_{area'}$ year $\longrightarrow$ $\underbrace{year - \overline{year}}_{year'}$

$area' = 0 \iff area = \overline{area}$

Then coefficient

$\beta_{river} = \mathbb{E}$ (rent in river road) when area average year is average

β_{area} interpretation? rate per year of observed increase, holding everything else constant.

Estimate all these coefficients using LS,
$\hat{\beta}_{area}$, $\hat{\beta}_{year}$, $\hat{\beta}_{RIVER}$, etc

TRUE VALUES, ESTIMATES

$\hat{\beta} \pm z * SE$

$\hookrightarrow$ S.E. = ? ESTIMATE OF THE S.D. OF $\hat{\beta}$

β is is in the interval $\hat{\beta} \pm 2SE$ about 95% of the time

MATHEMATICAL FORMULAS FOR SE'S

ESTIMATE S.E. OF $\hat{\beta}$ VIA SIMULATION

Algorithm for determining (estimating) S.E. of $\hat{\beta}$:

MODEL FOR DATA: see above

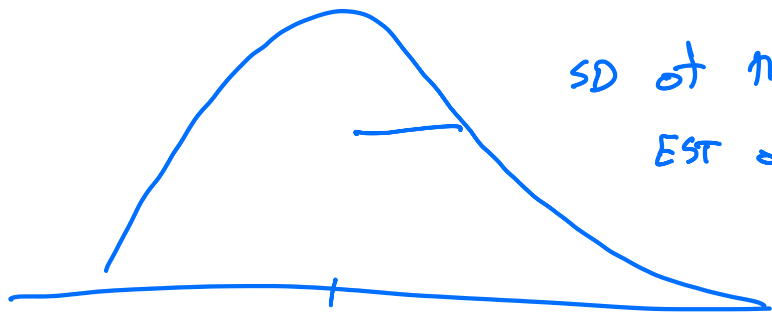
PARAMETRIC BOOTSTRAP: USE MODEL ABOVE

PUT IN ESTIMATED PARAMETER VALUES (FROM ORIGINAL DATA!)

SIMULATE FROM MODEL (HAVE EVERYTHING NEED)

FOR EACH SIM: GET AN EST OF $\hat{\beta}_{BETHEL}$

DIST'N OF $\hat{\beta}_{BETHELS}$



SD of this distribution

EST OF $SD(\hat{\beta}_{BETHEL})$

BOOTSTRAP

LIKE NEW DATASET DON'T TRUST MODEL WE SPECIFIED

TAKE NEW DATA POINTS BY DRAWING FROM EXISTING DATASET (W REPLACEMENT)

SUMMARIZE! CALCULATE COEF ESTIMATES USINGS LS $\hat{\beta}$

UNSATISFACTORY! ANY ESTIMATE w/o A (UN)CERTAINTY QUANTIFICATION IS USELESS

USE CI TO GIVE UNCERTAINTY QUANTIFICATION

$SE(\hat{\beta})$ = EST. SD OF $\hat{\beta}$ TO FORM A CI

SIMULATION IS UNIVERSAL METHOD

PARAMETRIC ← PUT IN ESTIMATED PARAMETERS IN MODEL, SIMULATE

NON-PARAMETRIC ← RESAMPLE THE SAMPLE

USE SIMULATED DIST'N OF $\hat{\beta}$ TO GET 95% CI (EMPIRICAL QUANTILES)