STAT 302: Linear Regression with Factors

Sheridan Grant

University of Washington Statistics Department

slgstats@uw.edu

May 4, 2020

Factors

In the April 29 class, we discussed what might happen if the `sex` variable in the SAT data had more than two possible responses. Suppose `0 == male`, `1 == female`, and `2 == other/no response`. We make this assumption in class for pedagogical reasons, but real-life justifications for this might be:

- Breaking response 2 down further might make model coefficients impossible to estimate, if there are too few responses in each group.
- If the reasons we expect GPA to vary by sex—probably societal factors, including discrimination—affect members of the 2 response group similarly (i.e. trans discrimination), then a single coefficient for this group is reasonable.

The Problem with Factors

Suppose for the "sex" variable S `0 == male`, `1 == female`, and `2 == other/no response`. The response, `FYGPA`, is Y and the other covariate, `HSGPA`, is X. We fit the model

$$Y = \beta_0 + \beta_X X + \beta_S S + \epsilon.$$

Then the expected difference in Y between a student with $S = 0$ and a student with $S = 2$, holding X fixed, is $2\beta_S$—twice the expected difference between $S = 1$ and $S = 2$. Why should we assume this? If we switched the coding for male/female, then the modeling assumption would be different—isn't that dumb? (Yes.)

Dummy Variables and One-Hot Encoding

1. Pick a level of the factor variable to be the "baseline" (irrelevant mathematically, but helpful for interpretation).
2. If there are k levels, define binary variables $S_1, \ldots, S_{k-1}$ where S_j is 1 if S is at the jth level, and the kth level is baseline. ($S_1 = \mathbf{1}[\text{female}]$, $S_2 = \mathbf{1}[\text{other/no response}]$.)
3. Model $Y = \beta_0 + \beta_X X + \beta_1 S_1 + \beta_2 S_2 + \epsilon$ (for our hypothetical SAT data).

In R, just do `sat$sex <- as.factor(sat$sex)`, specify the order of the levels with the `levels` argument.

Dummy Variables and One-Hot Encoding

$$Y = \beta_0 + \beta_X X + \beta_1 S_1 + \beta_2 S_2 + \epsilon$$

- Expected college GPA for male student with HS GPA x: $\beta_0 + \beta_X x$
- Expected college GPA for female student with HS GPA x: $\beta_0 + \beta_X x + \beta_1$
- Expected difference between male and female students (`female-male`), holding HS GPA fixed: subtract the previous two bullets to get β_1
- Expected difference between `female` and `other/no response` students (`female - other/no response`)?

Dummy Variables and One-Hot Encoding

$$Y = \beta_0 + \beta_X X + \beta_1 S_1 + \beta_2 S_2 + \epsilon$$

- Expected college GPA for male student with HS GPA x: $\beta_0 + \beta_X x$
- Expected college GPA for female student with HS GPA x: $\beta_0 + \beta_X x + \beta_1$
- Expected difference between male and female students (`female-male`), holding HS GPA fixed: subtract the previous two bullets to get β_1
- Expected difference between `female` and `other/no response` students (`female - other/no response`)? Just do `(female - male) - (other/no response - male)` to get $\beta_1 - \beta_2$.