LINE: Assumptions for Linear Regression

- L: $f^*(x) = \mathbb{E}(Y \mid X = x)$ is "assumed" to be a linear function of x. This is not really an assumption, but a restriction. If the truth f^* is not a linear function, then regression just returns us the best linear approximation of f^*.
- INE: error terms at all x_i's are iid $\mathcal{N}(0, \sigma^2)$ (can be relaxed to be uncorrelated with mean zero and constant variance). This assumption is related to the objective function, an unweighted sum of the squared errors at all x_i's. If the errors have unequal variances (heteroscedasticity) or correlated, then we should use a different objective function.
- No assumptions on X's. But to achieve a good performance, we would like $\mathbf{x}_i$'s to be uniformly sampled.

Outliers

- Outlier test based on leave-one-out prediction error. Let $\hat{\boldsymbol{\beta}}_{(-i)}$ be the LS estimate of $\boldsymbol{\beta}$ based on $(n-1)$ samples excluding the i-th sample $(\mathbf{x}_i, y_i)$, then

$$\frac{y_i - \mathbf{x}_i^t \hat{\boldsymbol{\beta}}_{(-i)}}{\text{some normalizing term}} \sim \mathcal{N}(0,1), \text{ if } i\text{th sample is NOT an outlier.}$$

- Datasets from real applications are usually large (in terms of both n and p). Do not recommend to test outliers. Why?
  - Need to adjust for multiple comparison; cannot detect a cluster of outliers.
- But do recommend to do some of the following:
  - Run the `summary` command in `R` to know the range of each variable;
  - Apply log, square-root or other transformations on right-skewed predictors and Y.
  - Apply winsorization to remove the effect of extreme values.