Predictive Analytics Lecture 2

Adam Kapelner

Stat 422/722
at The Wharton School of the University of Pennsylvania

January 17 & 18, 2017

Inference

We haven't spoken about t tests. Why is that?

Inference

We haven't spoken about t tests. Why is that?

In order to have inference, we need to make explicit random variable model assumptions

$$Y \sim g(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p, \sigma^2, \ldots)$$

must be assumed to be something like

$$Y \sim \mathcal{N}\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p,\, \sigma^2\right)$$

(we will explore next time)

R^2 vs. F test

In this case R^2 will be related to F, the omnibus test statistic for whether the model has any signal whatsoever.

$$R^2 = \frac{SSE_0 - SSE}{SSE_0} = \ldots = 1 - \left(1 + F\frac{p-1}{n-p}\right)^{-1}$$

$$\begin{aligned} F &= \frac{\frac{SSE_0 - SSE}{p-1}}{\frac{SSE}{n-p}} = \frac{SSE_0 - SSE}{SSE}\frac{n-p}{p-1} = \ldots \\ &= \underbrace{\frac{R^2}{1-R^2}}_{\substack{\text{ratio of variance}\\ \text{explained to}\\ \text{unexplained}}} \quad \underbrace{\frac{n-p}{p-1}}_{\substack{\text{penalty for}\\ \text{too many features}}} \end{aligned}$$