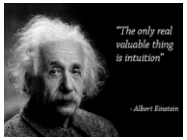
Dimensionality Issues

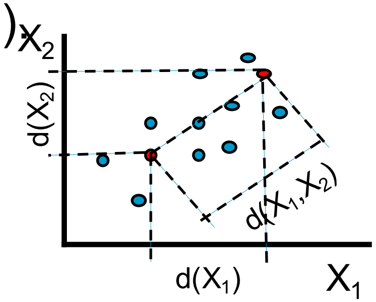


Dimensionality: Intuition

- **Is it good to add more variables to a model?** → **Up to a point!!**
- **Fewer** variables yield models with **lower R^2s** and **"omitted variable bias"**
- **Omitted variable bias** → the **included** variables **pick up** some of the **effects** of the **omitted** variables (e.g. is engine size a predictor of car mileage? If we omit a control variable for vehicle weight the effect of engine size may be biased because some of its effect will be due to vehicle weight)
- Models with more variables have **higher R^2s** and are **less biased**, but adding too many variables create a number of problems referred to as **"dimensionality"** issues (often called the **"curse of dimensionality"**).

Dimensionality Problems

- **Multi-Collinearity:** high correlation between independent variables cause the model to be unstable (e.g., dropping a few data points may yield substantially different results).
- **Over-Identification:** more variables force the model to fit the data tighter, but this is no guarantee that the model will make accurate predictions for new data.
- **Less Degrees of Freedom:** every added variable reduces the degrees of freedom of a model (n-p-1).
- **Less Parsimony:** complex models difficult to interpret. Some variables will be highly significant, others not so much (keep them or not?)
- **High Variance:** while adding more variables to a model reduces bias, the additional dimensions increase the variance of the model because the distance between points becomes larger



- **Nuisance (or Noise) Variables:** adding variables that are not very relevant for the model distorts its predictive accuracy and increases variance

In a nutshell, how many variables to include in the model is a **tradeoff!!**

Dimensionality Illustration

- Consider the following 3 regression models:
  (1) $mpg = \beta_0 + \beta_{Horsepower}(Horsepower) + \varepsilon$
  (2) $mpg = \beta_0 + \beta_{Size}(Size) + \varepsilon$
  (3) $mpg = \beta_0 + \beta_{Horsepower}(Horsepower) + \beta_{Size}(Size) + \varepsilon$
- If *Horsepower* and *Size* are perfectly **uncorrelated** (i.e., truly independent), it can be shown mathematically that:
  - ✓ $\beta_{Horsepower}$ and β_{Size} in (1, 2 and 3) are **unbiased**
  - ✓ $\beta_{Horsepower}$ and β_{Size} in (3) are **identical** to those of (1) and (2)
  - ✓ The R^2 of (3) = R^2 of (1) + R^2 of (2)
  - ✓ That is, it makes **no difference** modeling *Horsepower* and *Size* as a **multivariate** model or as two **simple** models
- However if *Horsepower* and *Size* are **correlated**, it can be shown that:
  - ✓ $\beta_{Horsepower}$ and β_{Size} in (1 and 2 reduced models) are **biased** → the **included** variable **picks up** the effect of the **omitted** variable
  - ✓ $\beta_{Horsepower}$ and β_{Size} in (3) **unbiased** but the model has **more variance**

Addressing Dimensionality Issues

- There are a number of modeling **techniques** to deal with high dimensionality. The most popular types are:
  - ✓ **Variable Selection** – if there are too many variables in the model, the most obvious solution is to carefully **select** which ones to include or not and testing the resulting models
  - ✓ **Shrinkage or Regularization** – when business rationale suggests that all or many available variables should be included in the model, dimensionality problems can be minimized by assigning **low weight** to unimportant variables by **shrinking** their **coefficients**, rather than removing them all together.
  - ✓ **Dimension Reduction Methods** – variables can be **grouped** and **combined** into fewer (i.e., reduced) **components**
  - ✓ **Structural Equations** – estimation is done with two or more **related models**, rather than a single model – i.e., a dependent variable in one model can be an independent variable in another model (covered later in the semester)

Kogod School
of
Business