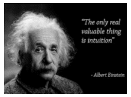
Cross-Validation



Cross-Validation: Intuition

- The methods discussed above (i.e., Cp, AIC, BIC and Adjusted R2) are useful because they make **adjustments** for **model size**, thus providing an **estimate** of the test error using the adjusted training error
- This is important because the **training MSE** in complex models often **underestimates** the test **MSE**.
- **Cross-Validation** is an alternative approach involving **fitting** a model with **training data** and **testing** it directly with the held-out or **test data**, which is an alternative approach
- Cross validation is particularly **important** with **over-fitted** or over-identified models in which the model will perform really well with the training data, but not so well with held-out data.
- The most common **cross-validation** technique is to: (1) **fit** the model with the **training** set; (2) use this model to **predict** values in the **test** set; and (3) then compute the **MSE** of the test set predictions.
- When comparing models, the one with the **lowest** cross-validation error or **MSE** is preferred.

Partitioning and Re-Sampling

- **Re-sampling** involve drawing **samples** from the data many times and re-fitting (i.e., **re-training**) the model each time to test the model more thoroughly.
- There are **many ways** to **partition** the data when sampling and re-sampling data.
- Most popular partitioning **methods** include:
  - **Hold-out** random splitting (**pre-set** percentage).
  - **K-Fold**
  - **Leave-One (or P) Out**
  - **Bootstrap**

KOGOD SCHOOL
of
BUSINESS