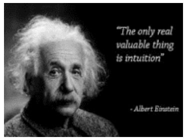
Multi-Collinearity

XI(✗) – X’s are not independent (are correlated)



Multi-Collinearity: Intuition

- Dimensionality issues are important, but multi-collinearity is **critical**
- Multi-collinearity is **tolerable** when **moderate**
- But it is a **problem** when it is **severe**
- If you include two predictor variables that are **perfectly correlated** the regression **cannot be estimated** – the solution is **indefinite**
- If **nearly perfectly correlated**, there is a solution but it is **unstable** and the **standard errors** will be very **large**
- With **multi-collinearity** there are linear dependencies among the predictors, beyond pair-wise correlation → the problem is **aggravated**
- The model has a solution but the **standard errors** will be **very large**
- Which makes the **model unstable** → if you drop a few observations (i.e., **"shake the tree")** the results may **change substantially**
- Which is why models with **severe** multi-collinearity may do **deceptively** well with the **training** set, but may perform **poorly** with the **test set**

Testing for Multi-Collinearity

- First, you need to analyze the **correlation matrix** and inspect for **desirable** correlations → **high** between the **dependent** and any **independent** variable; and **low** among **independent** variables.
- Run your regression model and report **multi-collinearity statistics** in the results. Two are most widely used:

  ➢ **Condition Index (CI):** a composite score of the linear association of all independent variables for the **model** as a **whole**

  ✓ **Rule of thumb: CI < 30** no problem, **30 < CI < 50** some concern, **CI > 50 severe**, no good

  ➢ **Variance Inflation Factors (VIF):** a statistic measuring the contribution of **each variable** to the model's multicollinearity → helps figure out which variables are problematic

  ✓ **Rule of thumb: VIF < 10** no problem, **VIF >= 10** too high,

R Tips

`colldiag(){perturb}` → Function to compute Condition Index (CI) statistics for multicollinearity analysis for the entire model

`collin.diag = colldiag(mod = lm.fit, scale = FALSE, center = FALSE, add.intercept = TRUE)` → The scale, center and add.intercept attributes can be used to evaluate collinearity with or without standardizing variab les, centering variables or the intercept → Concern: `CI>30` → Severe: `CI>50`

`Vif(){car}` → Function to compute Variance Inflation Factors (VIF's) to evaluate the contribution of each variable to the model's multicollinearity → Concern: `VIF>5`; Severe: `VIF>10`

`vif(lm.fit)` → Print VIF's for linear model `lm.fit`

KOGOD SCHOOL

of

BUSINESS