1. Overview

Types of Studies

- Controlled Experiments ($F = ma, PV = nRT$)
- Controlled Experiments with Supplemental variables
- Confirmatory observational studies
- Exploratory observational studies
- Data Collection and preparation
  - How the data are collected? (Design, Nature of the study)
  - Data consistency. (Plots and numerical summaries, logical relations)
  - Is this data analysis-ready? (Format checking, file conversion, etc.)

– GIGO (Garbage In; Garbage Out.)

- Reduction of explanatory or predictor variables
  Find parsimonious model with good explanatory/prediction power. Trade-off.
- Model refinement and selection
  Choosing from many "good" models, checking the adequacy of the models, sensitivity of the models, fixing the weak spots.
- Model Validation
  OK to explain what's going on? OK to predict what the future will be?

3. Diagnosis

Checking the adequacy of a regression model

- Improper functional form of a predictor
- Outliers
- Influential observation
- Multicollinearity

3.1. Improper functional form of a predictor

- Goal: Detect the suitable form of Y vs X_q while $X_1, \cdots, X_{q-1}$ in the model.
- Partial Regression Plots:
  $e(Y|X_1, \cdots, X_{q-1})$ vs. $e(X_q|X_1, \cdots, X_{q-1})$.
  - $e(Y|X_1, \cdots, X_{q-1})$: residual of Y regresses on $X_1, \cdots, X_{q-1}$

– $e(X_q|X_1, \cdots, X_{q-1})$: residual of X_q regresses on $X_1, \cdots, X_{q-1}$

- Why bother?

3.2. Outliers

- Rationale: The model (fitted) shouldn't be affect by just few points.
- LSE is EXTREMELY sensitive to outliers. Example.
- Detection: Residual-based tests and plots towards outlying Y

  –

$$r_i = \frac{e_i}{\sqrt{MSE(1-h_{ii})}} \text{ since}$$

$$\sigma^2(e) = \sigma^2(I - H), \quad H = X(X X)^{-1}X .$$

– Deleted Residuals: With or Without You. Outlying Y.

$d_i = Y_i - \hat{Y}_{i(i)}$

– Studentized Deleted Residual:

$t_i = \frac{d_i}{s(d_i)}$ where $s(d_i) = \sqrt{MSE_{(i)}(1 - h_{ii})}$

- Hat matrix Leverage values → Outlying X
  - $0 \le h_{ii} \le 1, \quad \sum_{i=1}^{n} h_{ii} = p.$
  - $\bar{h}_{ii} = \frac{p}{n}$. 2p/n, extreme h_{ii}, outside (0.2, 0.5)
  - $h_{new} =$

3.3. Influential obs

- $$(DFFITS)_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}}$$

  Flag: If $|DFFITS| > 1$ for small/medium data set or $> 2\sqrt{p/n}$, large data set.

- Cook's Distance

  $$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})}{pMSE} = \frac{e_i^2}{pMSE} \frac{h_{ii}}{(1 - h_{ii})^2} \qquad F_{p,n-p}$$

- $$(DFBETAS)_i = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\sqrt{MSE_{(i)} c_{kk}}}$$

  where c_{kk} is the diagonal entries of $(X'X)^{-1}$

  Flag: DFBETAS > 1 for small/medium data; $> 2/\sqrt{n}$.

  Change of signs.

- DFINF

R_k^2: R^2 of X_k regressing on the other X s.

Flag: Larger than 10 or $V\overline{I}F$