Building the Regression Model

C. Andy Tsao

Department of Applied Math, National Dong Hwa University

Outline

- Overview
- Selection of Predictors
- Diagnosis
- Remedial Measure
- Validation

Learning from Data

Objectives

- Reduction of explanatory or predictor variables
  Find parsioniousodel with good
  explanatoryprediction power. Trade-off.
- Model refinement and slection
  Choosing from many "good" odls, checkig the
  adequacy of the odels, snsiiviy of the odls, fixng
  the weak sots.
- Model Val13.5594(i)dation
  Ready to explain what’s go13.5594(i)ng on? Read to predict
  what the uture will be?
- Trade-off
  Best explanatoryprediction power v. Parsiony
  Criteria and how to us them? "Good" odls

Selection-I

- $\mathbf{R}_p^2 = 1 - \frac{SSE_p}{SSTO}$. ID those with substantial increases. NOT the biggest one.
- $\mathbf{R}_a^2 = 1 - ($

Comments

- No easy, clear-cut way to ID the best model
- Usually, many "good" models rather than one best model
- Respect the hierarchy of models
  - Higher order terms **<** lower order terms ($\mathbf{X}^4$ **<** $\mathbf{X}^1$)
  - Interaction terms **<** main effect terms ($\mathbf{X}_1\mathbf{X}_2$ **<** $\mathbf{X}_1$ or $\mathbf{X}_2$)
- Chapter 10 Variable Selection of Faraway, J. (2002). Also his Chapter 11 is highly recommended

Diagnosis

Checking the adequacy of a regression model

- Improper functional form of a predictor
- Outliers

  Influential observation
- Multicollinearity

Improper functional form of a predicto

- Goal: Detect the suitable form of

Outliers-II

- Studentized Deleted Residual:
  $\mathbf{t}_i = \frac{d_i}{s(d_i)}$ where $\mathbf{s}(\mathbf{d}_i) = \sqrt{\mathbf{MSE}_{(i)}(1 - \mathbf{h}_{ii})}$
- Hat matrix Leverage values $\rightarrow$ Outlying $\mathbf{X}$
  - $0 \leq \mathbf{h}_{ii} \leq 1, \quad \sum_{i=1}^{n} \mathbf{h}_{ii} = \mathbf{p}.$
  - $\bar{\mathbf{h}}_{ii} = \frac{p}{n}$. 2p/n, extreme $\mathbf{h}_{ii}$, outside $(0.2, 0.5)$
  - $\mathbf{h}_{new} = \mathbf{X}'_{new}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_{new}$ for hidden extrapolation.

Influential obs

- $(\mathbf{DFFITS})_i = \dfrac{\hat{Y}_i - \widehat{Y_{i(i)}}}{\sqrt{\mathbf{MSE}_{(i)} h_{ii}}}$ Flag: If $|\mathbf{DFFITS}| > 1$ for small/medium data set or $> 2\sqrt{p/n}$, large data set.
- Cook's Distance

  $$\mathbf{D}_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{pMSE} = \frac{e_i^2}{pMSE} \frac{h_{ii}}{(1-h_{ii})^2} \sim \mathbf{F}_{p,n-p}$$

- $$(\mathbf{DFBETAS})_i = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\sqrt{\mathbf{MSE}_{(i)} \mathbf{c}_{kk}}}$$

  where $\mathbf{c}_{kk}$ is the diagonal entries of $(\mathbf{X}'\mathbf{X})^{-1}$
  Flag: DFBETAS > 1 for small/medium data; $> 2/\sqrt{n}$.
  Change of signs.
- DFINF
- One vs many trouble makers.

Multicollinearity: VIF

- Problems of MLCL: X, Extra SSR, **s**$(\hat{\boldsymbol{\beta}})$, nonsignificance
- Informal Diagnosis
  - Sensitive incl/exclud of **X** or data
    Nonsignificance on important predictors
  - Wrong sign of estimated **β**
  - Large coefficient in $\mathbf{r}_{XX}$, Large $\mathbf{R}^2$ among **X**
  - Wide confidence ervals of **β**
- Variat68 -156.36on I6 -0 1647.6flation Factor (T6 -0 1647.6try $^{-1}$) **VIF**$_k$ diagonal L
  $\mathbf{r}_{XX}$. R

Model Validation

- Estimation/Fit the past; Predict the future
- Consistency with New Data
- Comparison with theoretical expectation, earlier empirical and simulation results
- Cross-Validation: Use of a holdout sample to check the model and predictive ability.

What's next?