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

- Controlled Experiments $(\mathbf{F} = \mathbf{ma}, \mathbf{PV} = \mathbf{nRT})$
- 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.)

Objectives

- 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?
- Trade-off
  Best explanatory/prediction power vs. Parsimoniousness
  Criteria and how to use them? "Good" models?

Selection-I

- $R_p^2 = 1 - \frac{SSE_p}{SS\quad O}$. ID those with substantial increases. NOT the biggest one.

- $R^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE_p}{SS\quad O} = 1 - \frac{MSE_p}{SS\quad O\ (n-1)}$

  $\mathbf{MSE}_p = \mathbf{SSE}_p / (\mathbf{n} - \mathbf{p})$.

  ID those with smaller/smallest $\mathbf{MSE}_p$.

- $\Gamma_p = \frac{\mathbf{E}(\mathbf{SSE}_p)}{\sigma^2} - (\mathbf{n} - 2\mathbf{p})$.

  $\hat{\Gamma}_p = \mathbf{C}_p = \frac{SSE_p}{MSE(X_1 \cdots X_{p-1})} - (\mathbf{n} - 2\mathbf{p})$

  ID those Small **C** C

Selection-II

All-subset Selection

Best among all 2^{p-1} combinations. Guidelines. **Forward stepwise Selection and other search procedures**

- Forward/Backward Stepwise Selection: One at a time, marginal effect, partial **F**-test
- Forward/Backward Selection: Marginal effect, partial (group) **F**-test

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 predi

- Goal: Detect the suitable form of $\mathbf{Y}$ vs $\mathbf{X}_q$ while $\mathbf{X}_1, \cdots, \mathbf{X}_{q-1}$ in the model.
- Partial Regression Plots: $\mathbf{e}(\mathbf{Y} | \mathbf{X}_1, \cdots, \mathbf{X}_{q-1})$ vs. $\mathbf{e}(\mathbf{X}_q | \mathbf{X}_1, \cdots, \mathbf{X}_{q-1})$.

Outliers-I

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. Why? What to expect?
  - Semistudentized residual: Same scale (Naive).
    $e_i^* \,/\, \frac{e_i}{\sqrt{MSE}}; \; e_i \,/\, Y_i - \hat{Y}_i$
  - Studentized residual: In the same scale (Refined).
    $r_i \,/\, \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$ since

$$^2(\mathbf{e}) / \quad ^2(I - H); \quad H / X(X'X)^{-1}X':$$

  - Deleted Residuals: With or Without You. Outlying Y.
    $d_i \,/\, Y_i - \hat{Y}_{i(i)}$

Outli2.32 l S 9923(e)-1.280.4(r)13.765

Influential obs

- $(\mathbf{DFFITS})_i = \frac{\hat{\mathbf{Y}}_i - \hat{\mathbf{Y}}_{i(i)}}{\sqrt{\mathbf{MSE}_{(i)} h_{ii}}}$ Flag: If $|\mathbf{DFFITS}| > 1$ for small/medium data set or $> 2\sqrt{\mathbf{p/n}}$, large data set.
- Cook's Distance
  $\mathbf{D}_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})}{pMSE} = \frac{e_i^2}{pMSE} \frac{h_{ii}}{(1-h_{ii})^2} \sim \mathbf{F}_{p\ n-p}$
-
  $$(\mathbf{DFBETAS})_i = \frac{\hat{\boldsymbol{\beta}}_k - \hat{\boldsymbol{\beta}}_{k(i)}}{\sqrt{\mathbf{MSE}_{(i)} \mathbf{c}_{kk}}}$$
  where $\mathbf{c}_{kk}$ is the diagonal entries of $(\mathbf{X}'\mathbf{X})^{-1}$
  Flag: DFBETAS > 1 for small/medium data; $> 2/\sqrt{\mathbf{n}}$.
  Change of signs.
- DFINF
- One vs many trouble makers.

Multicollinearity: VIF

- Problems of MLCL: X, Extra SSR, $\mathbf{s}(\hat{\boldsymbol{\beta}})$, nonsignificance
- Informal Diagnosis
  - Sensitive incl/exclud of **X** or data
  - Nonsig on important predictors
  - Wrong sign of estimated **β**
  - Large coefficient in $\mathbf{r}_{XX}$, Large $\mathbf{R}^2$

oi /R8 25.6

Remedial Measure

For UEQ error variances, high MTCL, INFLU obs

- Model Assumption
- Box-Cox Transformation: $\mathbf{y}^* = \mathbf{y}^r$, say $\mathbf{r} = 0.5, 2.5$ or $\mathbf{y}^* = \mathbf{log}(\mathbf{y})$, for example.
- Weighted LSE, General Error
- MTCL: Ridge Regression:

$$(\mathbf{X}'\mathbf{X} + \mathbf{cI})^{-1}$$

- Robust Regression
- Nonparametric regression, Boostraping.

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?