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 ($F = ma, PV = n \quad T$)
- Controlled Experiments with Supplemental variables
- Confirmatory observational studies
- Exploratory observational studies

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

- ${}^{2}_{p} = 1 - \frac{SSE_p}{SSTO}$. ID those with substantial increases. NOT the biggest one.
- ${}^{2}_{a} = 1 - (\frac{n-1}{n-p})\frac{SSE_p}{SSTO} = 1 - \frac{MSE_p}{SSTO/(n-1)}$
  $MSE_p = SSE_p/(n-p)$.
  ID those with smaller/smallest MSE_p.
- $\Gamma_p = \frac{E(SSE_p)}{\sigma^2} - (n - 2p)$.
  □$\Gamma_p = C_p = \frac{SSE_p}{MSE(X_1,\cdots,X_{p-1})} - (n - 2p)$
  ID those Small C C$_p$

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

Diagnosis

Improper functional form of a predi

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

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).

    $$e_i^* = \frac{e_i}{\sqrt{MSE}}$$

Outliers-II

- 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 $\rightarrow$ Outlying X
  - $0 \leq h_{ii}$

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

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

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

  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
- One vs many trouble makers.

Multicollinearity: VIF

- Problems of MLCL: X, Extra SSR, $s(\hat{\beta})$, nonsignificance
- Informal Diagnosis
  - Sensitive incl/exclud of X or data
  - Nonsig on important predictors
  - Wrong sign of estimated β
  - Large coefficient in r_{XX}, Large ^{2} among X
  - Wide confidence intervals of β
- Variation Inflation Factor (TL^{-1}) VIF_k diagonal entry of r_{XX}.

$$(VIF)_k = (1 - \;{}_k^2),$$

$_k^2$: ^{2} of $X_k$regressing on the other $X's$.

Flag: Larger than 10 or □ $V\bar{I}F$

Remedial Measure

For UEQ error variances, high MTCL, INFLU obs

- Model Assumption
- Box-Cox Transformation: $\ ^* = \ ^r$, say $r = 0.5, 2.5$

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?