Building the Regression Model

C. Andy Tsao

Department of Applied Math, National Dong Hwa University

Outline

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

Selection-I.1

- $\mathrm{R}_p^2 = 1 - \frac{SSE_p}{SSTO}$.

Selection-I.2

- AIC: Akaike's information criterion
  $\text{AIC}_p = -2\ln \text{likelihood} + 2\text{p} \propto \text{n}\ln \text{SSE}_p - \text{n}\ln \text{n} + 2\text{p}.$
  ID models with smaller AIC.
-

Selection-III

General Linear Test Approach Given $Y|X_1, \cdots, X_2$
Should $X_3, \cdots, X_5$ be added?.
Testing H_0 : Reduced vs H_1 : Full

- Fit the *full* model ($Y|X_1, \cdots, X_5$) and get $SSE(F), df_F$
- Fit the *reduced* model ($Y|X_1, \cdots, X_2$) and get $SSE(R), df_R$
- Calculate
  $$F = \frac{SSE(R) \;\; SSE(F)}{df_R \;\; df_F} / \frac{SSE(F)}{df_F} \sim F_{df_R \;\; df_F, df_F}$$
  Then perform a α level 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
  ($X^4 < X^1$)

  Interaction terms $<$ main effect terms
  ($X_1X_2 < X_1$ or X_2)
- Chapter 10 Variable Selection of Faraway, J. (2002). Also his Chapter 11 is highly recommended

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 $X_1, \cdots, X_{q-1}$
  - $e(X_q|X_1, \cdots, X_{q-1})$: 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. ExampX0.
- Detection: Residual-based tests and pXots towards outlying Y. Why? What to expect?

  Semistudentized residual: Same scale (Naive).

  $\mathrm{e}_i^* = \frac{e_i}{\sqrt{MSE}}, \mathrm{e}_i = \mathrm{Y}_i - \hat{\mathrm{Y}}_i$

  Studentized residual: In the same scale (Refined).

  $\mathrm{r}_i = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$ since

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

  Deleted Residuals: With or Without You. Outlying Y.

  $\mathrm{d}_i = \mathrm{Y}_i - \hat{\mathrm{Y}}_{i(i)}$

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

  $$0 \leq h_{ii} \leq 1, \qquad {}_{i=1}^{n} h_{ii} = p.$$

Influential obs

- $(\mathrm{DFFITS})_i = \dfrac{\widehat{Y}_i - \widehat{Y_{i(i)}}}{\sqrt{\mathrm{MSE}_{(i)} h_{ii}}}$ Flag: If $|\mathrm{DFFITS}| > 1$ for small/medium 13433] > 1

Multicollinearity: VIF

- Problems of MLCL: X, Extra SSR, $s(\hat{\beta})$, nonsignificance
- Informal Diagnosis

  Sensitive incl/exclud of X or data

  Nonsignificance on important predictors

  Wrong sign of estimated β

  Large coefficient in r_{XX}, Large R^2 among X

  Wide confidence intervals of β
- Variation Inflation Factor (TL^{-1}) VIF_k diagonal entry of r_{XX}.

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

R_k^2: R^2 of $X_k$regressing on the other $X's$.
Flag: Larger than 10 or $\gg \bar{VIF}$

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?