Building the Regression Model

C. Andy Tsao

Department of Applied Math, National Dong Hwa University

Outline

Learning from Data

- 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.)

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
  Ready to explain what's going on? Ready to educt what the future will be?
- Trade-off
  Best explanatory/prediction power vs. Parsimony
  Criteria and how to use them? "Good" models?

Selection-I.1

- $R^2_{\mathbf{p}} = 1 - \frac{\mathbf{SSE}_p}{\mathbf{SSTO}}$. ID those with substantial increases. NOT the biggest one.
- $R^2_{\mathbf{a}} = 1 - (\frac{\mathbf{n}-1}{\mathbf{n}-\mathbf{p}})\frac{\mathbf{SSE}_p}{\mathbf{SSTO}} = 1 - \frac{\mathbf{MSE}_p}{\mathbf{SSTO}/(\mathbf{n}-1)}$
  $MSE_{\mathbf{p}} = SSE_{\mathbf{p}}/(n-p)$.
  ID those with smaller/smallest $MSE_{\mathbf{p}}$.
- $\Gamma_{\mathbf{p}} = \frac{E(SSE_{\mathbf{p}})}{}$

Selection-I.2

- AIC: Akaike's information criterion
  $\text{AIC}_\mathbf{p} = -2\ln likelihood + 2p \propto n\ln SSE_\mathbf{p} - n\ln n + 2p.$
  ID models with smaller AIC.
- BIC (or SBC in Text): Schwartz' Bayesian information criterion.
  $\text{BIC}_\mathbf{p} = -2\ln likelihood + p\ln n \propto$
  $n\ln SSE_\mathbf{p} - n\ln n + (\ln n)p.$
  ID models with smaller BIC.
- Does these criteria make sense?
  Increasing/Decreasing in ..

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 $<$ loer 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

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 Y vs $X_{\mathbf{q}}$ while $X_1, \cdots, X_{\mathbf{q}-1}$ in the model.
- Partial Regression Plots:
  $e(Y|X_1, \cdots, X_{\mathbf{q}-1})$ vs. $e(X_{\mathbf{q}}|X_1, \cdots, X_{\mathbf{q}-1})$.
  - $e(Y|X_1, \cdots, X_{\mathbf{q}-1})$: residual of Y regresses on $X_1, \cdots, X_{\mathbf{q}-1}$
  - $e(X_{\mathbf{q}}|X_1, \cdots, X_{\mathbf{q}-1})$: residual of $X_{\mathbf{q}}$ regresses on $X_1, \cdots, X_{\mathbf{q}-1}$
- Why bother?

Outliers-I

The model (fitted) shouldn’t be affect by just few points.

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 ! Outlying X

  $0 \leq h_{ii} \leq 1; \sum_{i=1}^{n} h_{ii} = p:$

  $h_{ii} = \frac{p}{n}$: 2p/n, extreme h_{ii} , outside (0:2; 0:5)

  $h_{new} = X'_{new}(X'X)^{-1}X_{new}$ for hidden extrapolation.

Influential obs

- $(DFFITS)_{\mathbf{i}} = \dfrac{\hat{Y}_{\mathbf{i}} - \widehat{Y_{\mathbf{i(i)}}}}{\sqrt{MSE_{\mathbf{(i)}}\mathbf{h}_{ii}}}$ Flag: If $|DFFITS| > 1$ for small/medium data set or $> 2\sqrt{p/n}$, large data set.
- Cook's Distance
  $D_{\mathbf{i}} = \dfrac{\sum_{j=1}^{n}(\hat{\mathbf{Y}}_j - \hat{\mathbf{Y}}_{j(i)})^2}{\mathbf{pMSE}} = \dfrac{\mathbf{e}_i^2}{\mathbf{pMSE}} \dfrac{\mathbf{h}_{ii}}{(1-\mathbf{h}_{ii})^2} \sim F_{\mathbf{p,n-p}}$
- $$(DFBETAS)_{\mathbf{i}} = \frac{\beta_{\mathbf{k}} - \hat{\beta}_{\mathbf{k(i)}}}{\sqrt{MSE_{\mathbf{(i)}}c_{\mathbf{kk}}}}$$
  where $c_{\mathbf{kk}}$ 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
  - Nonsignificance on important predictors
  - Wrong sign of estimated $\hat{\beta}$
  - Large coefficient in $r_{\mathbf{XX}}$, Large R^2 among X

Remedial Measure

For unequal error variances, high multicollinearity, influential obs

Model Validation