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 ($F = ma, PV = nRT$)
- 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 predict what the future will be?
- Trade-off
  Best explanatory/prediction power vs. Parsimony
  Criteria and how to use them? "Good" models?

Selection-I.1

- $R_p^2 = 1 - \frac{SSE_p}{SSTO}$. ID those with substantial increases. NOT the biggest one.
- $R_a^2 = 1 - (\frac{n-1}{n-p})\frac{SSE_p}{SSTO} = 1 - \frac{MSE_p}{SSTO/(n-1)}$
  $MSE_p = SSE_p/(n-p)$.
  ID those with smaller/smallest MSE_p.
- $$\Gamma_p = \frac{E(SSE_p)}{\sigma^2} - (n-2p).$$
  $$\widehat{\Gamma_p} = C_p = \frac{SSE_p}{MSE(X_1,\cdots,X_{p-1})} - (n-2p)$$
  ID those Small C_p AND $C_p \approx p$.
- $PRESS_p = \sum_{i=1}^n (Y_i - \hat{Y}_{i(i)})^2$.Prediction Sum of Squares.
  ID those with small $PRESS_p$

Selection-I.2

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

Selection-II

All-subset Selection $Y|X_1, \cdots, X_{p-1}$

Best among all 2^{p-1} combinations.

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

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

Diagnosis

Basics

Checking the adequacy of a regression model

- Improper functional form of a predictor
- Outliers
- Influential observation
- Multicollinearity

Basics

- Heuristics: When the model is correct and parameters are estimated correctly $e_i \approx \epsilon_i$
- Assumption to be checked: $\epsilon_1, \cdots, \epsilon_n \sim N(0, \sigma^2)$
- Various checks
  1. Indepedence among errors (sequence plot, time plot), common variance (original, stardardized) , normality (normal probability plots)
  2. Independence of E(Y|x) (residuals vs. fitted values), Independence of X (vs. x)

Improper functional form

- 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. 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

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

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

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
  - $0 \leq h_{ii} \leq 1, \quad \sum_{i=1}^{n} h_{ii} = p.$
  - $\bar{h_{ii}} = \frac{p}{n}$. 2p/n, extreme h_{ii}, outside $(0.2, 0.5)$
  - $h_{new} = X'_{new}(X'X)^{-1}X_{new}$ for hidden extrapolation.

Influential obs

- $(DFFITS)_i = \dfrac{\widehat{Y_i} - \widehat{Y_{i(i)}}}{\sqrt{MSE_{(i)} h_{ii}}}$ 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
  - 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 $\gg V\bar{I}F$

Remedial Measure

For unequal error variances, high multicollinearity, influential obs

- Model Assumption
- Weighted LSE, General Error
- Transformations:
  - On Y: Box-Cox Transformation: $y^* = y^r$, say $r = 0.5, 2.5$ or $y^* = log(y)$, for example.
  - On x: Standardization, polynomials, Y regresses on $g_j(x_1, x_2, \cdots, x_{p-1}), j = 1, .., J$
- Multicollinearity: Principal Component Analysis, Ridge Regression:

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

- LASSO, Bridge regression
- Robust Regression

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?