Statistical Inference and Multivariate Analysis (MA324)

Lecture Slides
Lecture 32

Regression: Influential observations, Variable selection, Multicolinearity

Indian Institute of Technology Guwahati

Jan-May 2023

The PRESS Statistic

- PRESS: **PR**ediction **E**rror **S**um of **S**quares
- Delete i^{th} observation. Fit the model on remaining $(n-1)$ observations. Now, predict y_i. Let the corresponding prediction error be $e_{(i)} = y_i - \hat{y}_{(i)}$ (PRESS Residual).
- $e_{(i)} = \frac{e_i}{1-h_{ii}}$. (Can be shown)
- Large values of $e_{(i)}$ implies potential influential observations.
- Large difference between e_i and $e_{(i)}$ indicates an observation where model fit is quite well but a model built without that predicts poorly.
- $Var(e_{(i)}) = \frac{\sigma^2}{1-h_{ii}}$.

- Standardized PRESS residual is $\frac{e_{(i)}}{\sqrt{Var(e_{(i)})}} = \frac{e_i}{\sqrt{\sigma^2(1-h_{ii})}}$, which is same as the Studentized residuals. It can be estimated by replacing σ^2 by MS_{Res}.

- $PRESS = \sum_{i=1}^{n} e_{(i)}^2 = \sum_{i=1}^{n} \left(\frac{e_i}{1-h_{ii}}\right)^2$.

- PRESS is a measure of how well a regression model perform in predicting new observations.

- R^2 for prediction :

  $R^2_{prediction} = 1 - \frac{PRESS}{SS_T}$: gives indication of the prediction capability of the regression model.

- Using PRESS, we may compare model.

Variable Selection

- Criteria for Evaluating Subset Regression Models:
  - R^2
  - R^2_{Adj}
  - Residual Mean Square : $MS_{Res}(p) \equiv R^2_{Adj(p)}$.
  - PRESS Statistic
- Techniques:
  - All possible Regression.
  - Step-wise Type Procedures :
    - Forward Selection : $F = \frac{SS_R(x_2|x_1)}{MS_{Res}(x_1,x_2)}$
    - Backward elimination
    - Step-wise Regression

Multicollinearity

- Near linear relationship among regressor $(\sum_{j=1}^{p} t_j \underset{\sim}{x_j} \simeq 0)$
- Effect of Multicollinearity:
  - Consider scaled response and regressor (length unit).
  - Consider $y = \beta_1 x_1 + \beta_2 x_2 + \epsilon$.
  - $\hat{\beta}_1 = \frac{r_{1y} - r_{12} r_{2y}}{1 - r_{12}^2}$, and $\hat{\beta}_2 = \frac{r_{2y} - r_{12} r_{1y}}{1 - r_{12}^2}$; where r_{12} is the simple correlation between x_1 and x_2, and r_{jy} is the simple correlation between x_j and y, $j = 1, 2$.
  - $Var(\hat{\beta}_j) = \frac{\sigma^2}{1 - r_{12}^2}$, $Cov(\hat{\beta}_1, \hat{\beta}_2) = \frac{-r_{12}\sigma^2}{1 - r_{12}^2}$.
  - Strong multicollinearity between x_1 and x_2 indicates the r_{12} will be large.
  - If $|r_{12}| \to 1$, $Var(\hat{\beta}_j) \to \infty$, and $|Cov(\hat{\beta}_1, \hat{\beta}_2)| \to \infty$.
  - The above large variances and covariances means different sample taken at the same x level could lead to widely different estimates of the model parameters.

- Effect of Multicollinearity (contd.):
  - $L_1^2 = (\hat{\beta} - \beta)^T(\hat{\beta} - \beta)$.
  - $E(L_1^2) = \sum_{j=1}^{p} Var(\hat{\beta}_j) = \sigma^2 Tr(X^{'}X)^{-1} = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}$, where λ_j's are eigenvalues of $(X^{'}X)$.
  - If $(X^{'}X)$ is ill-conditioned then at least one λ_j will be small $\Rightarrow E(L_1^2)$ is big.
  - Therefore, we have $E(\underset{\sim}{\hat{\beta}}^T \underset{\sim}{\hat{\beta}}) = \underset{\sim}{\beta}^{'} \underset{\sim}{\beta} + \sigma^2 Tr(X^{'}X)^{-1}$, implies magnitude of $\underset{\sim}{\hat{\beta}}$ are large.

Multicollinearity Diagnostics

- Examination of correlation matrix $(X^{'}X)$:
  - If x_i and x_j are nearly linearly dependent, then $|r_{ij}|$ should be close to 1.
  - However, this procedure is helpful to detect near linear dependence between a pair of regressors only.
- Variance Inflation Factors (VIFs):
  - $Var(\hat{\beta}_j) = \sigma^2 c_{jj}, C = (X'X)^{-1}$. It can be shown that $c_{jj} = (1 - R_j^2)^{-1} = \frac{1}{1-R_j^2}$, where R_j^2 is the coefficient of determination obtained when x_j is regressed on remaining $(k-1)$ regressors.
  - $VIF_j = \frac{1}{1-R_j^2}$: This measures the factor by which the variance of $\hat{\beta}_j$ inflated due to the near linear dependence.
  - Rule of thumb : If any of $VIF > 5$, the associated coefficient is estimated poorly due to multicollinearity.

- Eigen system Analysis of $(X^{'}X)$:
  - The eigen values, $\lambda_1, \lambda_2, ..., \lambda_p$, can be used to see the extent of multicollinearity.
  - Small eigen values (one or more) $\Rightarrow$ multicollinearity.
  - Condition number, $k = \frac{\lambda_{max}}{\lambda_{min}}$.
  - Rule of thumb :

    $k < 100 \rightarrow$ No serious problem with multicollinearity.

    $100 \leq k < 1000 \rightarrow$ moderate to strong multicollinearity.

    $k \geq 1000 \rightarrow$ severe multicollinearity.
  - Condition indices : $k_j = \frac{\lambda_{max}}{\lambda_j},\ j = 1, 2, ..., p$

    The number of j's such that, $k_j \geq 1000 \rightarrow$ provide useful information on the number of near linear dependence.

Method for dealing with multicollinearity

- Source of multicollinearity:
  - Data collection method (ex: biased sample) $\rightarrow$ collecting more data.
  - Constraints in model or population (ex: family income (x_1) and household size (x_2)) $\rightarrow$ Model respecification
  - Model specification (ex: range of x is small, then adding x^2 in the model) $\rightarrow$ Model respecification
  - An overdefined model (ex: adding more regressors) $\rightarrow$ Model respecification, and other method of estimate like Ridge regression.