14 Regression diagnostics (14.5-14.6)

Applied regression analysis and other multivariable methods

Yi Zhou

May 16, 2016

14.5 Colinearity

Issues: unreliable and unstable parameter estimates and standard errors

Example on pp.358. Collinearity exists due to the association between AGE and AGE^2

- Regression coefficients are inconsistent
- Standard error are increased

Mathematical concepts in collinearity

Considering fitting the model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + E_i$$

- $\hat{\beta}_1 = (\frac{r_{y,x1} - r_{y,x2} r_{x1,x2}}{1 - r^2_{x1,x2}})(\frac{SD_y}{SD_{x1}})$
- $\hat{\beta}_2 = (\frac{r_{y,x2} - r_{y,x1} r_{x1,x2}}{1 - r^2_{x1,x2}})(\frac{SD_y}{SD_{x2}})$
- $\hat{\beta}_1, \hat{\beta}_2, \bar{Y} - \hat{\beta}_0$ are all proportional to $\frac{1}{1 - r^2_{X1,X2}}$ (inflation factor)
- If $r^2_{x1,x2} = 1$, then the estimates of the coefficients are indetrminate
- As $r^2_{x1,x2}$ decreases, the collinearity problem becomes less severe
- If $r^2_{x1,x2}$ is near 1, the regression coefficients are highly unstable

Collinearity concept

Examine collinearity

- If each predictor variable is treated as the response variable with the the independent variables are the remaining predictors
- If any of the associated R^2-values equals to 0, then collinearity exists
- `Collinearity` indicates that one of the predictors is nearly an exact combination of the others
- `Perfect collinearity` means that the parameters in the model cannot be estimated uniquely
- A model containing a perfect collinearity is overparameterized
- `Near collinearity` exists when R^2-values is nearly 1

`The variance inflation factor` (VIF)

$$VIF_j = \frac{1}{1 - R_j^2}$$

- The larger the value of VIF_j, the more troublesome the variable X_j is
- Larger than 10
- Equivalent to $R_j^2 > 0.9$ or $R_j > 0.95$

`Tolerance`

$$Tolerance_j = \frac{1}{VIF_j} = 1 - R_j^2$$

Intercept requires special treatment

- If the means of all X_j's are 0 (centered data), $\bar{Y}$ is the estimated intercept.
- VIF_0

The treatment of intercept in regression diagnostics

- it is another predictor
- it should be eliminated from discussion

Solutions to the presence of collinearity

- Computational algorithm to detect collinearity
- Scale the data properly: scaling, cetering, and computing `z scores`
- `Principle component analysis`
- Centering may help decrease collinearity

Principle component ananlysis

- `Principle components`: a set of new variables that are linear combinations of the original predoctors
    - Components are not correlated with each other
    - Each in turn has maximum variance (`eigenvalue`)
    - The larger eigenvalue, the more important the associated principle component
    - Eigenvalue approaching zero indicates the presence of a near collinearity
    - Eigenvalue equal to zero indicates an exact collinearity

- The number of zero eigenvalues is the number of collinearities
- Using eigenvalues to determine the presense of near collinearity
  - `condition index` (CI): $CI_j = \sqrt{\frac{\lambda_1}{\lambda_j}}$
  - `condition number` (CN): $CN = \sqrt{\frac{\lambda_1}{\lambda_k}}$
  - `variance proportions`: two or more loadings less than 0.5 doesn't indicates a major problem

Collinearity diagnostics

Three steps

- Simple descriptive analyses
- VIF values
  - in model with two predictors, the two VIF values are identical
- Condition index and varaince proportion should be examined

Treating collinearity problems

- Eliminating one or more of the predictors in the cllinear set
- Scientific expertise and previous experience to indentify the variables that would be accaptable to drop
- Orthogonal polynomials
- Attention to dummy variables and interaction terms
- Study design
- Using centered data
- Regression on principle components, ridge regression