STAT021 Statistical Methods II

Lecture 21 MLR Review

Lu Chen
Swarthmore College
11/27/2018

Four-step process of MLR

CHOOSE	Exploratory data analysis	
(Re-CHOOSE)	Summary statistics and visualization for individual variables	**Model building:**
	Summary statistics and visualization for the relationship btw Y and each X	Exhaustive
	Multicollinearity between X's (scatterplot & correlation matrix, VIF)	Forward
	Potential predictors (polynomial, interaction, categorical ←→ quantitative)	Backward
	State the MLR statistical model (and assumptions)	Stepwise
FIT	Estimate the parameters using maximum likelihood method	
(Re-FIT)	Write down the estimated regression line	**Model comparisons:**
ASSESS	Model	Nested F test
(Re-ASSESS)	t tests for individual slopes (different from t tests for correlations)	Adjusted R^2
	(*Categorical predictors, polynomial and interaction terms)	Mallow's Cp
	F test and R^2 for the whole model	*AIC*
	Error	
	Check assumptions (linearity, 0 mean, constant var, Normality, independence)	
	Search for unusual points (leverage, standardized residuals, Cook's D)	
USE	Understanding relationships: interpreting the values of intercept and slopes	
	(*Categorical predictors, polynomial and interaction terms)	
	Prediction	
	Mean response and confidence interval	
	Individual response and prediction interval	

Four-step process: CHOOSE

Exploratory data analysis

- Individual variables
  - Summary statistics: sample size, mean, SD (quantitative); table of counts and proportions (categorical); missing data
  - Visualization: histogram (quantitative) and bar plot (categorical)
- The relationship between the response variable Y and each predictor X_k
  - Summary statistics: correlation coefficient and/or SLR
  - Visualization: scatterplot (quantitative vs. quantitative) and boxplot (quantitative vs. categorical)

Four-step process: CHOOSE

Exploratory data analysis

- Multicollinearity between the predictors (usually including all the predictors in their original scale without any transformation)
  - Scatterplot matrix
  - Correlation matrix
  - Variance inflation factor (VIF)
- When any categorical variable is involved in the MLR model
  - `ggpairs()` function provides a barplot on the diagonal, histograms by categories in the scatterplot matrix and a boxplot in the correlation matrix.
  - `vif()` function provides a slightly different output. Read VIF values from the `GVIF^(1/(2*Df))` column and compare these values to $\sqrt{5}$ instead of 5.

Four-step process: CHOOSE

Consider potential predictors

- Quantitative predictors
  - Linear terms (the variables in the original scale)
  - Polynomial terms (NO polynomial terms for categorical predictors)
  - Transformation by a certain function (eg., $\log()$)
  - Cateogorization
- Categorical predictors
  - A categorical predictor with m categories $\Longrightarrow m - 1$ dummy variables
  - As quantitative if the categories are ordinal and the relationship is linear
- Interaction terms (two-way, three-way, and more)
  - Quantitative $\times$ quantitative
  - Quantitative $\times$ categorical
  - Categorical $\times$ categorical

Four-step process: CHOOSE

Model building and comparisons

- Model building strategies
  - Exhaustive
  - Foward selection
  - Backward elimination
  - Stepwise procedure
- Model comparison criteria
  - Nested F test (not applicable if two models have the same number of predictors OR a simpler model has samller SSE)
  - R^2_{adj}
  - Mallow's C_p
  - AIC

Four-step process: CHOOSE

State the MLR statistical model (and assumptions)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \epsilon, \text{ where } \epsilon \overset{iid}{\sim} N(0, \sigma)$$

- K predictors
- $K + 2$ parameters
- 5 assumptions
  - Linearity
  - Zero mean
  - Constant variance
  - Normality
  - Independence

Four-step process: FIT

Estimate the parameters using maximum likelihood method

- "Write down the estimated regression line"

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_K x_K$$

  - One intercept and K slopes
- Residual standard error

$$\hat{\sigma} = \sqrt{MSE} = \sqrt{\frac{\sum (y - \hat{y})^2}{n - K - 1}}$$

  - Degree of freedom: $n - K - 1$

Four-step process: ASSESS Model

- Three tests
  - Individual t tests for the slopes
    - Caution: the t tests for the slopes of categorical predictors, polynomial terms and interaction terms have their specific interpretations.
  - ANOVA F test for the significance of the model

$$F = \frac{MSModel}{MSE} = \frac{SSModel/K}{SSE/(n-K-1)} \sim F(K, n-K-1)$$

  - t tests for correlation between the response variable and each predictor (this is usually done in the exploratory data analysis)
- R^2 for the stength of the model

$$R^2 = \frac{SSModel}{SSTotal} = 1 - \frac{SSE}{SSTotal}$$

Four-step process: ASSESS Error

Check model assumptions

1. **Linearity**: linear relationship between Y and Xs.
   - Scatterplot of residuals e on fitted values $\hat{y}$ (pattern?)
2. **Zero mean**: mean of the errors is 0 - always true
3. **Constant variance**: variability of the errors is the same for all X values.
   - Scatterplot of residuals e on fitted values $\hat{y}$ (spread?)
   - Breusch-Pagan test for H_0: constant variance
4. **Normality**: distribution of the errors is Normal.
   - Normal Q-Q plot (sometimes histogram of residuals is helpful)
5. **Independence** - check data collecting process

Four-step process: ASSESS Error

Search for unusual points

Statistic	Moderately unusual	Very unusual
Leverage, h_i	$> 2(K+1)/n$	$> 3(K+1)/n$
Standardized residual, stdres$_i$	beyond ± 2	beyond ± 3
Studentized residual, studres$_i$	beyond ± 2	beyond ± 3
Cook's distance, D_i	> 0.5	> 1

$$D_i = \frac{(\text{stdres}_i)^2}{K+1}\left(\frac{h_i}{1-h_i}\right)$$

- K is the number of predictors

Re-CHOOSE, re-FIT, re-ASSESS

If any model assumption is violated and thus the response variable needs to be transformed OR model fitting could be improved by transforming the predictor(s), re-choose, re-fit and re-assess the model, and select the final best model.

Four-step process: USE

- Understanding relationships
  - Interpret the values of intercept and slopes
  - Caution: the values of the slopes of categorical predictors, polynomial terms and interaction terms have their specific interpretations.
- Prediction
  - Mean response and confidence interval
  - Individual response and prediction interval
  - If the response variable is transformed, it needs to be transformed back to the original scale in prediction.

Example: Happy Planet Index

See *Lecture21_Examples.Rmd*.