14 Regression diagnostics (14.1-14.4)

Applied regression analysis and other multivariable methods

Yi Zhou

April 27, 2016

Preview

Regression diagnostics to detect inaccuracy and invalid regression results:

- detecting outliers
- checking regression assumptions
- detecting the presence of collineairy

Outliers

- recording errors
- (not recording errors) may influence the fit of regression models
- affect the understanding of the relationship between dependent and independent variables

Assumption cheking

- a thorough check of assumptions
- potential limitations of the analysis

Collinearity

- regression results are unstable
- small and practically unimportant changes in data can result in meaningful large change in the estimated model

Simple approaches to diagnosing problems in data

Simple descriptive analyses to be performed:

- examine five largest and five smallest values for every numeric variable
- examine descriptive statistics
- scatterplot if possible
- **partial regression plot for mutiple regression**
- calculate pairwise correlations

Partial regression plot

- $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{k-1} X_{k-1} + E$
- $X_k = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_{k-1} X_{k-1} + E$
- The residuals $(Y - \hat{Y})$ vs $(X_k - \hat{X}_k)$ are plotted

Residual analysis

Regression residuals

- $\hat{E}_i = Y_i - \hat{Y}_i$
- independent and following $N(0, \sigma^2)$

The standardized residuals: $z_i = \dfrac{\hat{E}_i}{S}$

- S is MSE
- unit variance

The studentized residuals: $r_i = \dfrac{\hat{E}_i}{S\sqrt{1 - h_i}}$

- h_i is the leverage
- if the assumptions are satisfied, $r_i \sim t_{n-k-1}$

The jackknife residuals

Detecting outliers

Outlier : rare or unsual value at the extreme range
Regression diagnostic statistics: leverage, jackknife residuals, Cook's distance

Leverages

- the geometric distance of the ith predictor point from the center point of the predictor space
- the larger value indicates farther distance the outlier is from the center
- the leverage score for the ith data unit is: $h_i = [\mathbf{H}]_i$, the ith `diagonal element` from the projection matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}$
- $0 \leqslant h_i \leqslant 1$
- the average leverage value: $\dfrac{(k+1)}{n}$
- scrutinizing those $h_i > \dfrac{2(k+1)}{n}$

Jackknife residuals

- For the ith observation, the jackknife residual $r_{(-i)} = \frac{\hat{E}_i}{S_{(-i)}\sqrt{1-h_i}}$
- $S^2_{(-i)}$ is the MSE with ith observation deleted
- S^2 will be larger than $S^2_{(-i)}$ if the outlier is masking its effect
- $r_{(-i)} \sim t_{n-k-2}$
- If the absolute value of $r_{(-i)}$ is greater than 95th percentile t distribution, i observation may be an outlier

Cook's distance

- the change of the regression coefficients when an observation is deleted
- $d_i = (\frac{1}{k+1})(\frac{h_i}{1-h_i})r_i^2$
- $d_i > 1$ may deserve closer scrutiny

Assessing assumptions

Plots of the residuals (ordinary, studentized, jackknife) vs predicted values (Figure 14.6) - four senarios
Plots of the residuals vs each predictor
Assumption violation or small sample size or sparse data?

Assessing the normality assumption

- Kolmogorov-Smirnov test or Shapiro-Wilks test
- P-P (probability-probability) plot or Q-Q (quantile-quantile) plot

Strategies for addressing violations of regression assumptions

Transformation

- to stabilize the variance of the dependent variable if the homoscedasticity is violated
- to normalize the dependent variable
- to linearize the regression model

Commonly used transformations

- Log transformation
- Square root transformation
- Square transformation
- Arcsin transformation

Weighted least-square analysis

- the asssumptions of homoscedasiticity and/or independence do not hold
- when the variance of Y varies for different values of the independent variables