ST4242
Lecture 10

Jialiang Li

Purpose of residual analysis

- Checking model assumptions.
- Identify outliers.
- Identify model influential observations.

Checking assumptions

- Model assumptions:
    1. Linearity assumption
    2. Homogeneity assumption
    3. Normality assumption
    4. Independence assumption
- Will check assumptions 2 and 3 using residuals.

Homogeneity assumption

- For “repeated” models, assume that each subject observed at common time points to have a common covariance matrix.
- For “random” models, within-subject variation is assumed to be the same.

Normality assumption

- For “repeated” models, assume marginal distribution is normal.
- For “random” models, assume conditional error and random effects to follow normal distribution.

Residuals

- Can be used to check any systematic departures from the model for the mean response.
- Raw residuals *r*
- Standardized residuals
- Relation between standardized residuals and raw residuals.

Residuals for longitudinal data

- The raw residuals are correlated and do not necessarily have constant variance.
- The scatter plot of residuals against the predicted values will not necessarily have a constant range.
- May be correlated with the covariates and show an apparent systematic trend in residual plots.

Transformed residuals

- Standardization itself may not be useful.
- Normalization or de-correlate transformation is needed.
- Technical step: Cholesky decomposition of a positive definite covariance matrix.
- Will be used to produce diagnostic plots such as residual plots and QQ plots.

Influence analysis for linear regression

- Traditional influence analysis for linear regression model is concentrated around 2 major concepts:
- 1. Leverage, as the diagonal element of the hat matrix, and standardized residuals.
- 2. Case deletion diagnostics and Cook’s distance.

Leverage

- Turn to Wikipedia for the definition of hat matrix.
- The diagonal element of the hat matrix H is called leverage.
- The leverage *h* is positive because (X’X) is positive definite. The sum of the leverages equal to the rank of X.
- Cases with high leverage are called influential.

Cook’s distance

- Comes from the idea of case deletion.
- A case is called influential if the least squares estimate changes significantly after deleting this case from the data.
- It is not necessary to recalculate the regression after case deletion.