STAT 3008 Applied Regression Analysis
Tutorial 10.

XU Yongze & DONG Fangyuan

Nov 25 & 26 & 28, 2013

Contents

1 Overview of Course Content

- Simple linear regression & Multiple linear regression
  - scatter plot
  - model(definition and notation)
  - residual plots
  - parameters' estimates
  - hypothesis tests
  - confidence intervals
  - draw conclusion(association/causal)
- Further discussion
  - Look at the linear assumption:
    - polynomial regression
    - transformation
  - Look at the linear predictors:
    - aliased
    - misfitted model(overfitted/lurking variable)
    - model selection
    - qualitative $\rightarrow$ factors
  - Look at errors:
    - normality assumption $\rightarrow$ QQ-plot
    - unequal variance $\rightarrow$ weighted least square
  - Look at particular cases of observations:
    - outlier tests
    - leverage
    - Cook's distance
- Some techniques
  - added-variable plot(another way to find $\hat{\beta}_1$)
  - Delta method(find out that $g(\hat{\theta}) \sim N(g(\theta), \sigma^2 g'(\theta)^T D g'(\theta))$, so as to find C.I. for $g(\theta)$)

2 Diagnostics with Residuals

2.1 What is the diagnostics?

♣ Regression diagnostics are used after fitting so as to check whether assumptions(mean/var/error) are consistent with observed data. The basic tools are residuals or scaled residuals. The basic idea is to check if the residuals look reasonable(null plot: mean zero, constant variance, no seperated points).

2.2 About H

$$H = (h_{ij})_{n \times n}$$

$$h_{ii} \text{ is called leverage.}$$

$H = X(X'X)^{-1}X'$
$H^t = H$
$H^2 = H$
$tr(H) = p + 1$
$\sum_{i=1}^n h_{ji} = \sum_{i=1}^n h_{ij} = 1$
$HJ = J$
$H\mathbf{1} = \mathbf{1}$
$JJ = nJ$
$HX = X$
$X'H = X'$
$(I - H)X = 0$
$H(I - H) = 0$
$\text{Cov}(\hat{e}, \hat{Y}) = 0$
$\text{Cov}(\hat{e}, Y) = \sigma^2(I - H)$
$\text{Cov}(e, \hat{Y}) = \sigma^2 H$
$\text{Cov}(e, Y) = \sigma^2 I$
RSS: $\sum(Y_i - \hat{Y}_i)^2 = Y'(I - H)Y$
TSS(SYY): $\sum(Y_i - \bar{Y})^2 = Y'(I - \frac{1}{n}J)Y$
SSreg: $\sum(\hat{Y}_i - \bar{Y})^2 = Y'(H - \frac{1}{n}J)Y$

2.3 Residuals

$$\hat{\mathbf{e}} = Y - X\hat{\beta} = (I - H)Y = (I - H)\mathbf{e}$$

♣ **Assumptions for** e:

$\mathrm{E}(e)$, $\mathrm{Var}(e) = \sigma^2 I$, e is normally distributed.

♣ **Properties of** $\hat{e}$:

$\mathrm{E}(\hat{e}) = 0$, $\mathrm{Var}\hat{e} = \sigma^2(I - H)$, dependent & non-identically distributed.

Then,

$$\mathrm{Var}(\hat{e}_i) = \sigma^2(1 - h_{ii})$$

$$\mathrm{Cov}(\hat{e}_i, \hat{e}_j) = -\sigma^2 h_{ij}$$

The higher the leverage, the smaller the variance of $\hat{e}_i$.

♣**Checkthe residual**:

The residual plot should look like the null plot.

2.4 Leverage

$\hat{Y} = HY$
$\hat{Y}_i = \sum_{k=1}^{n} h_{ik} Y_k = h_{ii} Y_i + \sum_{k \neq i}^{n} h_{ik} Y_k$

As h_{ii} approaches to 1, $h_{ik}(k \neq i)$ approachesto 0, so $\hat{Y}$ gets closer to Y_i. h_{ii} is pulling $\hat{Y}_i$ towards Y_i, giving the name levrage. For cases with large values of h_{ii}, no matter what value of y_i is observed, we are nearly certain to get a residual near 0. But, be careful of leverage point if $h_{ii} \sim 1$.

3 Solutions of Homework3