STAT 151A Lecture 24

Henry Liev

22 October 2025

Remark 0.1 (More on Inference for Linear Models)

NLM: $Y_{|\mathbf{X}} = \mathbf{X}\beta + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_n)$
Big issue: What if $\vec{\varepsilon}$ is non-normal
What if $\text{Var}(\varepsilon_i|x_i) \neq \text{Var}(\varepsilon_j|x_j)$: Heteroskedasticity? If this happens, NLM inference can be very bad
Solutions:

(1) Model/Adjust for Heteroskedasticity $\rightarrow$ log transform (weighted least squares)

(2) Bootstrap cases $\rightarrow$ no assumptions on heteroskedasticity

(3) Heteroskedastic Linear Model (HLM)

Method 0.2 (Heteroskedastic Linear Model)

$$Y_i = x_i^\intercal \beta + \varepsilon_i, \mathbb{E}(\varepsilon_i) = 0, \varepsilon_i \perp\!\!\!\perp \varepsilon_j, \text{Var}(\varepsilon_i) = \sigma_i^2$$

How can we get hypothesis tests and confidence intervals under the HLM? Key $\rightarrow$ understand distribution of $\hat{\beta}$. We will need to let $m \rightarrow \infty$ and use CLT to do this
$\hat{\beta} = (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal Y$
$\mathbb{E}(\hat{\beta}) \overset{HLM}{=} \beta$
Steps:

(1) $\text{Var}(\hat{\beta}) \rightarrow$ Define some quantities
$B_n = \frac{1}{n} \sum x_i x_i^\intercal \in \mathbb{R}^{(p+1)\times(p+1)} = \frac{1}{n}\mathbf{X}^\intercal \mathbf{X}$
$M_n = \frac{1}{n} \sum \sigma_i^2 x_i x_i^\intercal$
$\text{Var}(\hat{\beta}) = \frac{1}{n} B_n^{-1} M_n B_n^{-1}$
M_n depends on unknowns σ_i^2. In practice estimate σ_i^2 by e_i^2 the residual of $y_i - x_i^\intercal \hat{\beta}$
$\widehat{\text{Var}}(\hat{\beta}) = \frac{1}{n} B_n^{-1} \hat{M}_n B_n^{-1}$
$\hat{M}_n = \frac{1}{n} \sum e_i^2 x_i x_i^\intercal \in \mathbb{R}^{(p+1)\times(p+1)} = \frac{1}{n}\mathbf{X}^\intercal \hat{\Omega} \mathbf{X}, \hat{\Omega} = \text{diag}(e_1^2, e_2^2, \dots, e_n^2)$

(2) Use CLT (generalized form) $\hat{\beta} \overset{n\rightarrow\infty}{\sim} \mathcal{N}(\beta, (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\hat{\Omega}\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1})$

Remark 0.3 (Confidence Intervals Using HLM)

In Practice:

HLM: $\hat{\beta}_j + z_{1-\alpha/2}\sqrt{\left[(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\hat{\Omega}\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\right]_{jj}}$

NLM: $\hat{\beta}_j \pm t_{n-p-1,(1-\alpha/2)}\sqrt{\hat{\sigma}\left[(\mathbf{X}^\intercal\mathbf{X})^{-1}\right]_{jj}}$

Hypothesis tests are similar, instead of F-tests, you can get χ^2-tests

R: sandwich package $\rightarrow$ vcov H(c), calculates $\widehat{\mathrm{Var}}(\hat{\beta})$

lmtest $\rightarrow$ coeftest(), coefci()

Remark 0.4 (Regression Diagnostics)

(1) look for individual points with outsize impact

   (a) Measurement error → get rid of/fix
   (b) Technically correct but unduly influences the model
   (c) Indicate the presence of a ? but important group or subtle weakness in your model

   How to detect/define "unsual" data points numerically?
   Univariate data: another point is far away from all other points
   Bivariate/Multivariate Data: