STATS 205: Hierarchical Linear Models **Fall 2018**

Lecture 2: t-Test, The Wald Test and Likelihood Ratio Test

Lecturer: Prof. Jingyi Jessica Li *Subscriber: Kexin Li*

2.1 Multivariate Linear Model

The vector form of the model is:

$$Y_{n\times 1} = \mathbf{X}_{n\times p}\ \beta_{n\times 1} + \epsilon_{n\times 1},$$

where the *design matrix*

$$\mathbf{X} = \begin{bmatrix} 1_{p\times 1} & \mathbf{x}_1 & \dots & \mathbf{x}_{p-1} \end{bmatrix}, \mathbf{x}_j \in \mathbb{R}^n \text{ are predictors/features/covariates.}$$

Fixed Design vs. Random Design

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_n^{\mathrm{T}} \end{bmatrix}, \mathbf{x}_i \in \mathbb{R}^p \text{ are observations.}$$

For random design, we assume that $\mathbf{x}_i$'s are random and thus $(\mathbf{x}_i, Y_i)$ are jointly random. Then we have to check whether $\{\mathbf{x}_i, Y_i\}_{i=1}^n$ can be reasonably assumed to be independent and identically distributed (*i.i.d*). Random design is widely used in econometrics. A related concept is the measurement error model, which assumes that the marginal distribution of $\mathbf{x}_i$ takes a specific form.

For fixed design, the only randomness comes from ϵ_i's. The assumptions that all the observations follow the same linear model and the condition that ϵ_i's are *i.i.d* is easier to satisfy. Fixed design is widely used statistics.

If the linear model is reasonably assumed, it is sufficient to consider fixed design in most cases.

For a detailed discussion about random design and fixed design, please refer to

Buja, A., et al. "A conspiracy of random predictors and model violations against classical inference in regression." Arxiv preprint. Available at `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.753.718&rep=rep1&type=pdf` (2014).

2.2 Inference of β and σ^2

- β indicates the effect of predictors on the response

- $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1})$
  - exact if $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n)$
  - the distribution of $\hat{\beta}$ totally depends on our assumptions
  - asymptotically holds when ϵ_i's are *i.i.d* but not Gaussian (Central Limit Theorem)
    For the asymptotic result, please see `http://cameron.econ.ucdavis.edu/e240a/asymptotic.pdf`.
- the distribution of $\hat{\beta}$ can be viewed as from infinitely repeated sampling (Frequentist View)

2.2.1 *t*-test

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \qquad \beta_1 : \text{ the intercept}$$

Hypothesis testing:

$$H_0 : \beta_j = 0; \quad H_1 : \beta_j \neq 0$$

The t-statistic:

$$T = \frac{\hat{\beta}_j}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_j)}} = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 \times j^{th} \text{ diagonal entry of } (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}}}$$

For the denominator, we use the *plug-in estimator* (replacing the true value by the estimator).

In order to do a valid statistical inference, we need to know the distribution of the t-statistic.

2.2.2 Hat Matrix

The fitted value:

$$\hat{Y} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}Y$$

We define the *Hat Matrix* as:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$$

The residuals can be expressed as:

$$e = Y - \hat{Y} = Y - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}Y = (\mathbf{I}_n - \mathbf{H})Y$$

Properties of the Hat Matrix

- The Hat Matrix is symmetric, i.e.

$$\mathbf{H}^{\mathrm{T}} = \mathbf{H}$$

- The Hat Matrix is idempotent, i.e.

$$\mathbf{H}^2 = \mathbf{H}$$

- $\mathrm{trace}(\mathbf{H}) = \mathrm{trace}(\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}) = \mathrm{trace}(\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}) = \mathrm{trace}(\mathbf{I}_p) = p$, $\mathrm{trace}(\mathbf{I}_n - \mathbf{H}) = n - p$.

2.2.3 Estimators for σ^2

Given the Gaussian assumption, the MLE estimator:

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

We can show that

$$\frac{n\hat{\sigma}^2_{MLE}}{\sigma^2} = \frac{\sum_{i=1}^{n} e_i^2}{\sigma^2} \sim \chi^2_{n-p} \qquad \text{given that } \epsilon_i \sim \mathcal{N}(0,\sigma^2)$$

and thus

$$\mathbb{E}[\frac{n\hat{\sigma}^2_{MLE}}{\sigma^2}] = n-p, \qquad \mathbb{E}[\hat{\sigma}^2_{MLE}] = \frac{n-p}{n}\sigma^2 \text{ (biased)}$$

An easy modification to generate an unbiased estimator is:

$$\hat{\sigma}^2_{unbiased} = \frac{1}{n-p}\sum_{i=1}^{n} e_i^2 = \frac{RSS}{n-p} = \frac{SSE}{n-p}$$

where *RSS* (Residual Sum of Squares) and *SSE* (Sum of Squared Errors) and different names for the same item.

2.2.4 t-statistic and t-distribution

t-distribution:

(i) The numerator follows the standard Gaussian distribution.

(ii) The denominator is the square root of a variable following the χ^2 distribution divided by the degree of freedom.

(iii) The numerator and the denominator are independent.

We are to show that the t-statistic T follows the $t-$distribution under the null hypothesis only conditional on normality assumption on the error term: $\epsilon \sim \mathcal{N}(0,\sigma^2\mathbf{I}_n)$.

Proof. For (iii), an important property about $\mathcal{N}(\cdot,\cdot)$ is:

If $\mathbb{R}^p \ni U \sim \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, $\mathbf{A} \in \mathbb{R}^{n\times p}$, $\mathbf{B} \in \mathbb{R}^{m\times p}$, then

$$\mathbf{A}U \text{ and } \mathbf{B}U \text{ are independent } \iff \mathbf{A}\boldsymbol{\Sigma}\mathbf{B}^{\mathrm{T}} = \mathrm{Cov}(\mathbf{A}U,\mathbf{B}U) = \mathbf{0}_{n\times m}$$

Here in our case,

$$\left.\begin{array}{rr} \hat{\beta} = \mathbf{A}Y, & \mathbf{A} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}} \\ e = \mathbf{B}Y, & \mathbf{B} = \mathbf{I}_n - \mathbf{H} \\ Y \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta},\boldsymbol{\Sigma}), & \boldsymbol{\Sigma} = \sigma^2\mathbf{I}_n \end{array}\right\} \Rightarrow \mathbf{A}\boldsymbol{\Sigma}\mathbf{B}^{\mathrm{T}} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{H}) = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}(\mathbf{X}^{\mathrm{T}} - \mathbf{X}^{\mathrm{T}}\mathbf{H}) = \mathbf{0}$$

$$\Rightarrow \hat{\beta} \perp e$$

$$\Rightarrow \hat{\beta}_j \perp \hat{\sigma}^2_{unbiased} = \frac{1}{n-p}\sum_{i=1}^{n} e_i^2$$

$$\Rightarrow \hat{\beta}_j \perp \sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_j)}$$

Together with

$$\hat{\beta}_j \sim \mathcal{N}(\beta_j,\ \sigma^2 \cdot j^{th} \text{ diagonal entry of } (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1})$$

$$\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_j)} = \sqrt{\hat{\sigma}^2_{unbiased} \cdot j^{th} \text{ diagonal entry of } (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}}$$

and

$$\frac{\hat{\sigma}^2_{unbiased}}{\sigma^2} \sim \frac{\chi^2_{n-p}}{n-p}$$

Under the null hypothesis (H_0), we can say that the t-statistic follows the t-distribution with $(n-p)$ degree of freedom:

$$T = \frac{\hat{\beta}_j}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_j)}} \overset{H_0}{\sim} t_{n-p}$$

Asymptotically, the t-distribution converges to the standard normal distribution if viewing $\hat{\sigma}^2$ as the true value of σ^2 as $n \to \infty$:

$$t_{n-p} \overset{n\to\infty}{\longrightarrow} \mathcal{N}(0,1)$$

2.3 The Wald Test

$$\beta = \begin{pmatrix} \beta_{1\,(p_1\times 1)} \\ \beta_{2\,(p_2\times 1)} \end{pmatrix}, \qquad \widehat{\mathrm{Var}}(\hat{\beta}) = \begin{pmatrix} \widehat{\mathrm{Var}}(\hat{\beta}_1) & \widehat{\mathrm{Cov}}(\hat{\beta}_1, \hat{\beta}_2) \\ \widehat{\mathrm{Cov}}(\hat{\beta}_1, \hat{\beta}_2) & \widehat{\mathrm{Var}}(\hat{\beta}_2) \end{pmatrix}$$

Hypothesis testing:

$$H_0 : \beta_2 = \mathbf{0}_{p_2\times 1}; \quad H_1 : \beta_2 \neq \mathbf{0}_{p_2\times 1}$$

The Wald statistic:

$$W = \hat{\beta}_2^{\mathrm{T}} \widehat{\mathrm{Var}}(\hat{\beta}_2)^{-1} \hat{\beta}_2$$

Note that if $p_2 = 1$, $W = T^2$.

Given the normality assumptions that $\hat{\beta}_2 \sim \mathcal{N}(\cdot,\cdot), e \sim \mathcal{N}(\cdot,\cdot)$, the Wald statistic follows the exact F-distribution:

$$W/p_2 \overset{H_0}{\sim} F(p_2, n-p)$$

Asymptotically, the normality assumptions can be relaxed with large sample size (Central Limit Theorem):

$$W \underset{n\to\infty}{\overset{H_0}{\sim}} \chi^2_{p_2}$$

2.4 Likelihood Ratio Test

The test-statistic Λ:

$$\begin{aligned}
\Lambda &= \frac{\max\limits_{\beta \in H_0} L(\beta)}{\max\limits_{\beta} L(\beta)} \\
\log \Lambda &= \max_{\beta \in H_0} \log L(\beta) - \max_{\beta} \log L(\beta) \\
&= -\frac{1}{2\sigma^2}[RSS_{H_0} - RSS] \qquad \text{(in linear model case with Gaussian assumptions)} \\
-2\log \Lambda &= \frac{RSS_{H_0} - RSS}{\sigma^2} \\
-2\log \hat{\Lambda} &= \frac{RSS_{H_0} - RSS}{RSS/(n-p)}
\end{aligned}$$

Question: why do we use $RSS/(n-p)$ instead of $RSS_{H0}/(n-p+p_2)$ as the estimator of σ^2?

Answer: the reason is that we assume the full model with a p-dim β as the true model.

Asymptotically,

$$-2\log \hat{\Lambda} \underset{n\to\infty}{\overset{H_0}{\sim}} \chi^2_{p_2}$$

Notes:

(i) LRT: the most powerful test

(ii) Wald $\Longleftrightarrow$ LRT under linear model assumptions.

(iii) **Practical tips:**

- do t-test when only 1 coefficient interested
- do LRT(calculate RSS) or the Wald test when more coefficients interested

(iv)

Tests	Decisions to make	Description
$H_0^{(j)}: \beta_j = 0,\ H_1^{(j)}: \beta_j \neq 0$ $j = 1, 2, \cdots, p$	p	Multiple testing Type I error is hard to control
$H_0: \beta = 0,\ H_1: \beta \neq 0$	1	One test for p parameters jointly

2.5 ANOVA (Analysis of Variance)

$$\begin{aligned}
\sum_{i=1}^{n}(Y_i - \bar{Y})^2 &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \\
SST &= RSS(\text{or } SSE) + SSR
\end{aligned}$$

- $\bar{\hat{Y}}_i = \bar{Y}_i$