Statistics

Luc Veldhuis

November 2016

Specific tests

Test designed for specific situation/distribution

- t-test
- Gauss-test

The likelihood is known, then we may use this test.

Definition: Likelihood-ratio test

Suppose X distributed in accordance with density p_θ.
Then the likelihood ratio statistic for testing $H_0 : \theta_0 \in \Theta_0$ versus $H_a : \Theta\Theta_0$ is:

$$\lambda(X) = \frac{sup_{\theta \in \Theta} p_\theta(X)}{sup_{\theta \in \Theta_0} p_\theta(X)} = \frac{p_{\hat{ml}}(x)}{p_{\hat{\theta}_0}}$$

Here $p_{\hat{ml}}(x)$ is the Maximum likelihood estimator, $p_{\hat{\theta}_0}$ is the constrained Maximum likelihood estimator.

Likelihood ratio test

In general $\lambda(X) \geq 1$
$\lambda(X) = 1$ if $\hat{\theta_{ml}} \in \Theta_0$
If $\lambda(X)$ is large, this points to H_a

Example

Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$
$H_0 : \mu : \mu_0$ vs $H_a : \mu \neq \mu_0$
Determine $\lambda_n(X)$ with $n =$ sample size.

$$p_\theta(X) = \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} exp(-\frac{1}{2}\sigma^2(X_i - \mu)^2)$$

Likelihood ratio test

Example continued

$\hat{\mu_{ml}} = \bar{x}, \hat{\mu_0} = \mu_0$ then:

$$\lambda_\theta(X) = \frac{\prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} exp(-\frac{1}{2}\sigma^2(X_i - \bar{x})^2)}{\prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} exp(-\frac{1}{2}\sigma^2(X_i - \mu_0)^2)}$$

But take *log* + multiply by 2:

$2log(\lambda_\theta(X)) = \frac{n}{\sigma^2}(\bar{X} - \mu_0)^2 = (\sqrt{n}\frac{(\bar{X}-\mu_0)}{\sigma})^2 \sim N(0,1)$

Which follows a χ_1^2 distribution.

Hence, $K_T = \{X : 2log(\lambda_n(X) \geq \chi^2_{1;1-\alpha})\}$

Likelihood ratio test

Notation

- $\Theta - \theta_0 = \{\theta - \theta_0 : \theta \in \Theta\}$ for all $\theta_0 \in \Theta_0$
- $\sqrt{n}(\Theta - \theta_0) = \{\sqrt{n}(\theta - \theta_0) : \theta \in \Theta\}$ for all $\theta_0 \in \Theta_0$

Example

$X_1, \ldots X_n \sim exp(1), \theta_{ml} = \bar{x}, \theta_0 = 1$
$H_0 : \theta = 1$ vs $H_a : \theta \neq 1$
$\Theta = (0, \infty), \theta_0 = 1 : \sqrt{n}(\Theta - \theta_0) = \sqrt{n}(-1, \infty) \rightarrow \mathbb{R}$ if $n \rightarrow \infty$
$\Theta_0 = \{1\} : \sqrt{n}(\Theta_0 - \theta_0) = \sqrt{n}\{0\} = \{0\}$ if $n \rightarrow \infty$ generates a 0 dimentional set.

Likelihood ratio test

Theorem

Suppose $\theta \to p_\theta$ differentiable for all x and the sets $\sqrt{n}(\Theta - \theta_0)$ and $\sqrt{n}(\Theta_0 - \theta_0)$ converge for a given θ_0 to a k dimentional set and a k_0 dimentional set with $k > k_0$ for $n \to \infty$. Then, under certain conditions, and assuming $\theta = \theta_0$ (H_0 is true):

$$2log(\lambda_n(X)) \text{ follows } \chi^2_{k=k_0}$$

Then they converge in distribution

$$lim_{n\to\infty} F_n(X) = F(X)$$

Likelihood ratio test

Theorem continued

Often warranted by Central Limit Theorem:

$X_i \sim U(-1,1)$ study $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$

Then **drawings**

$$lim_{n\to\infty} F_n(\bar{X}_n) = N(0, \frac{1}{3})$$

Example

$X_1, \ldots, X_n \sim exp(\theta), \theta > 0$
$H_0 : \theta = 1$ vs $H_a : \theta \neq 1$

$$\begin{aligned} p_\theta(x) &= \prod_{i=1}^{2} \theta exp(-\theta x_i) \\ &= \theta^n exp(-n\theta\bar{x}) \\ \lambda_n(x) &= \frac{\theta_{ml}^{\hat{n}} exp(-n\bar{\theta}_{ml}\bar{x})}{exp(-n\bar{x})} \end{aligned}$$

Apply theorem:

$$\begin{aligned} 2log(\lambda_n(X)) &\sim \chi_1^2 \text{ if } n \to \infty \\ & 2nlog(\bar{X}) - 2n(\bar{X}^2 - \bar{X}) \end{aligned}$$

The critical region (approximately): $\{X : 2log(\lambda_n(X)) \geq \chi^2_{1,1-\alpha}\}$
So $K_T = (\chi^2_{1,1-\alpha}, \infty)$. This is always one sided, irrespective of H_0

Likelihood ratio test

Remark

If $H_0 : \theta \leq \theta_0$, then convergence of $\sqrt{n}(\Theta_0 - \theta_0)$ does not go well for all $\theta_0 \in \Theta_0$, but not for a boundary point. Then, $2log(\lambda_n(X))$ follows a mixture of χ^2

Regression models

Definition

Statistical methods to estimate the relationship among variables. E.g.:

$$Y = f(x_1, \dots, x_n) + error$$

Y is the response, x's are covariates/explanatory variables and f is a function deemed appropriate. Commonly, f is linear in the parameters.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon$$

Linearity: eases estimation.
May seem a limitation: broad class. E.g. 'hockey-stick' curve.

Regression models

Example

Y : income of former VU student
X : age of former VU student
In the simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i \qquad \forall i f(x_{1i}) = E(Y_i)$$

$$\begin{aligned} E(Y_i) &= E(\beta_0 + \beta_1 x_{1i} + \epsilon_i) \\ &= E(\beta_0) + E(\beta_1 x_{1i}) + E(\epsilon_i) \\ &= \beta_0 + \beta_1 x_{1i} \end{aligned}$$

When $\epsilon_i \sim N(0, \sigma^2)$. Parameters $(\beta_0, \beta_1, \sigma^2)$
The close the points are to the line, the better the model. Is reflected in σ^2 (relative to scale of line).

$$Y_i \sim N(\beta_0 + \beta_1 x_{1i}, \sigma^2)$$

General model

Definition

$$Y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} \epsilon_i$$

with $\epsilon_i \sim N(0, \sigma^2)$

ML estimation

Definition

Likelihood of simple linear regression model:

$$L(\theta; Y_1, \dots Y_n) = \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} exp(-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 x_{1i})^2)$$

Take logarithm:

$$-\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_{1i})^2$$

First maximize with respect to β_0 and β_1 using lease squares method:

$$min_{\beta_0,\beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_{1i})^2$$

ML estimation

Definition continued

Maximize the quadratic distance between line and observants. This gives: $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ and $\hat{\beta}_1 = \frac{S_y}{S_x} r_{x,y}$ where

$$r_{x,y} = \frac{\frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X})}{(S_x S_y)}$$

which is the sample correlation coefficient (measure of linear relatedness).

The estimated line is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i}$

Second, maximize log-likelihood with respect to σ^2 (after substitution of $\hat{\beta}_0$ and $\hat{\beta}_1$ for β_0 and β_1). Then

$$\hat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i})^2$$

Likelihood ratio statistic

Testing

Test $H_0 : \beta - 1 = 0$ and $H_a L\beta_0 \neq 0$ in simple linear regression model. Using the likelihood ration statistic.

$$\lambda_n(Y_1, \ldots, Y_n) = \frac{L(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2; y_1, \ldots, y_n)}{L(\tilde{\beta}_0, 0, \tilde{\sigma}^2; y_1, \ldots, y_n)}$$

Theorem

$$2log(\lambda(Y_1, \ldots, Y_n)) \text{ follows } \chi^2_1$$
$$\Theta_0 - \theta_0 = \{\beta_0^{(0)} - \beta_0, \beta_1^{(0)} - 0, \sigma^{2(0)} - \sigma_0^2 :$$
$$(\beta_0^{(0)}, \beta_1^{(0)}, \sigma^{2(0)}) \in \Theta_0\} \subset \mathbb{R} \; k_0 = 2$$
$$\Theta - \theta_0 \subset \mathbb{R}^3$$

$H_0 : \beta_0 = 0, \beta - 1 = 0$
Usually intercept is not in H_0