Gaussian SLR: Hypothesis Testing

GR 5205 / GU 4205
Section 2/ Section 3

Columbia University
Xiaofei Shi

Least Square Estimator for Gaussian Model

X - predictor (random) variable **Y - response random variable**

- Build your model:

  1) relationship: $Y = \beta_0 + X\beta_1 + \epsilon, \;\; \epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$

  2) preference: choose $\widehat{\beta}_0, \widehat{\beta}_1$ to minimize $\mathbb{E}\left[\|Y - \beta_0 - X\beta_1\|^2\right]$

- Estimate your model parameters: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

  1) using observed data to express your preference: $\min_{\beta_0, \beta_1} Q := \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$

  2) get parameters estimation for your model:

- Understand your model:

  1) properties of estimations: Today!

  2) predictions: $\widehat{Y}_0 = \widehat{\beta}_0 + X_0 \widehat{\beta}_1 \qquad \hat{y}_0 = b_0 + x_0 b_1$

What is hypothesis testing?

- Null hypothesis H_0

  Alternative hypothesis H_1
- Type I error:

  rejection of a true null hypothesis;
- Type II error:

  failure to reject a false null hypothesis.
- Can we control both?

<table>
<tr><th colspan="2" rowspan="2">Table of error types</th><th colspan="2">Null hypothesis (H_0) is</th></tr>
<tr><th>True</th><th>False</th></tr>
<tr><th rowspan="2">Decision about null hypothesis (H_0)</th><th>Don't reject</th><td>Correct inference
(true negative)
(probability = $1-\alpha$)</td><td>Type II error
(false negative)
(probability = β)</td></tr>
<tr><th>Reject</th><td>Type I error
(false positive)
(probability = α)</td><td>Correct inference
(true positive)
(probability = $1-\beta$)</td></tr>
</table>

Pipeline to design a test

1) State the statistical assumptions;
2) State the relevant null hypothesis and alternative hypothesis;
3) Set a threshold α;
4) Choosing the test statistics T and test methods;
5) Under the null hypothesis, derive the distribution p of the test statistics T;
6) Insert data into T and get t_{obs};
7) Under the null hypothesis, calculate the p-value by $p\left(T \geq t_{\text{obs}}\right)$;
8) Reject the null hypothesis if and only if the p-value is less than or equal to the threshold.

In Linear Regression Models

1) State the statistical assumptions; $Y = \beta_0 + X\beta_1 + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$

2) State the relevant null hypothesis and alternative hypothesis;

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_1 : \beta_1 \neq 0.$$

3) Typical choice: $\alpha = 5\%$

Summary of Gaussian SLR:
distribution of estimator, confidence interval

	distribution	1- α confidence interval
slop β_1	$\widehat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\\|x-\bar{x}1_n\\|^2}\right)$	$\left[\widehat{\beta}_1 \pm \frac{\widehat{\sigma}_{\mathrm{LS}}}{\\|x-\bar{x}1_n\\|} t(\frac{\alpha}{2}; n-2)\right]$
intercept β_0	$\widehat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\\|x-\bar{x}1_n\\|^2}\right)\right)$	$\left[\widehat{\beta}_0 \pm \widehat{\sigma}_{\mathrm{LS}}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\\|x-\bar{x}1_n\\|^2}} t(\frac{\alpha}{2}; n-2)\right]$
noise level σ^2	$\widehat{\sigma}^2_{\mathrm{LS}} \sim \frac{\sigma^2}{n-2}\chi^2(n-2)$	$\left[\frac{\widehat{\sigma}^2_{\mathrm{LS}}}{\chi^2\left(\frac{\alpha}{2}; n-2\right)}, \frac{\widehat{\sigma}^2_{\mathrm{LS}}}{\chi^2\left(1-\frac{\alpha}{2}; n-2\right)}\right]$
mean of Y_0 at x_0 $\mathbb{E}[Y_0] = \beta_0 + x_0\beta_1$	$\widehat{\beta}_0 + x_0\widehat{\beta}_1 \sim \mathcal{N}\left(\mathbb{E}[Y_0], \sigma^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\\|x-\bar{x}1_n\\|^2}\right)\right)$	$\left[\left(\widehat{\beta}_0 + x_0\widehat{\beta}_1\right) \pm \widehat{\sigma}_{\mathrm{LS}}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\\|x-\bar{x}1_n\\|^2}} t(\frac{\alpha}{2}; n-2)\right]$
new observation at x_0 $Y_0 = \beta_0 + x_0\beta_1 + \epsilon_0$	$\widehat{\beta}_0 + x_0\widehat{\beta}_1 \sim \mathcal{N}\left(\mathbb{E}[Y_0], \sigma^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\\|x-\bar{x}1_n\\|^2}\right)\right)$	$\left[\left(\widehat{\beta}_0 + x_0\widehat{\beta}_1\right) \pm \widehat{\sigma}_{\mathrm{LS}}\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{\\|x-\bar{x}1_n\\|^2}} t(\frac{\alpha}{2}; n-2)\right]$

Wald Test:

$$T = \frac{\widehat{\beta}_1 - 0}{\widehat{\sigma}_{LS}}$$

$$\left[\widehat{\beta}_1 \pm \frac{\widehat{\sigma}_{\mathrm{LS}}}{\|x - \bar{x}1_n\|} t(\tfrac{\alpha}{2}; n-2)\right]$$

Hypothesis testing based on confidence interval

- The upper bound, lower bound and the length of the confidence interval are all random variables!
- As α shrinks, the interval widens. (High confidence comes at the price of big margins of error.)
- As sample size grows, the interval shrinks. (Large samples mean precise estimates.)
- As noise level increases, the interval widens. (The more noise there is around the regression line, the less precisely we can measure the line.)
- As grows, the interval shrinks. (Widely-spread measurements give us a precise estimate of the slope.)

ANOVA: ANalysis Of VAriance

- From HWK2 Q2: $\|Y - \bar{Y}1_n\|^2 = \|Y - \widehat{Y}\|^2 + \|\widehat{Y} - \bar{Y}1_n\|^2$
- Residual sum of squares: $RSS = \|Y - \widehat{Y}\|^2$
- Total sum of squares: $SS_{\text{total}} = \|Y - \bar{Y}1_n\|^2$
- The sum of squares due to regression: $SS_{\text{reg}} = \|\widehat{Y} - \bar{Y}1_n\|^2 = RSS - SS_{\text{total}}$
- RSS and SS_{reg} are independent (from last class!)
- F test:

ANOVA

Source	df	SS	MS	F	p-value
Regression	1	SS_{reg}	$MS_{reg} = \frac{SS_{reg}}{1}$	$F = \frac{MS_{reg}}{MS_{res}}$	
Residual	n-2	RSS	$\widehat{\sigma}^2 = \frac{RSS}{n-2}$		
Total	n-1	SS_{total}			

F test: What are we really testing?

- An F test for whether the simple linear regression model “explains” (really, predicts) a “significant” amount of the variance in the response.
- Compare two versions of the simple linear regression model.

References and further reading

- Kutner, Nachtsheim, Neter: *Applied Linear Regression Models* Chapter 2
- Agresti: *Foundations of Linear and Generalized Linear Models* Chapter 2&3
- CMU 36-401 Lecture notes