Inference for the two-variable gaussian model

Susan Thomas
IGIDR, Bombay

23 September 2008

Recap

- Two variable, gaussian distribution model:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i \\ \epsilon_i &\sim N(0, \sigma^2) \end{aligned}$$

- MLE $\hat{\beta}_1 = \frac{\sum_{i=1}^{N} Y_i (X_i - \bar{X})}{\sum_{i=1}^{N} (X_i - \bar{X})^2}$
  $\hat{\beta}_1$ is a function of r_{XY}, sample correlation between X, Y.
- MLE $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
- MLE $\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} \hat{\epsilon}_i^2 = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$

Links between r_{XY} and the MLE estimates

- $\hat{\beta}_1 = r_{XY} \frac{s_Y}{s_X}$
- $\hat{\sigma}^2_{mle} = (1 - r^2_{XY})\sigma^2_Y$
- $\hat{\sigma}^2_{mle} = (1 - r^2_{XY})\sigma^2_Y = (1 - r^2_{XY})\sigma^2_R$
- $r^2_{XY} = 1 - \frac{\sigma^2_{mle}}{\sigma^2_Y}$

Inference for the two-variable gaussian distribution model

Two paths of inferences

- Confidence intervals for the parameters
- Confidence intervals for $E(y)$
- LR test and it's asymptotic distribution.
- Variants of the LR test.

Confidence intervals for MLE $\hat{\beta}_1$

- The reparameterised version of the model is simpler to work with:

$$Y_i = \gamma_0 X_{0,i} + \gamma_1 (X_{1,i} - \bar{X}) + \omega_i$$

- This gives:

$$\begin{aligned}
\hat{\gamma}_1 = \hat{\beta}_1 &= \frac{\sum_{i=1}^{N} (X_{1,i} - \bar{X}) Y_i}{\sum_{i=1}^{N} (X_{1,i} - \bar{X}_1)^2} \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^{N} (X_{1,i} - \bar{X})(\beta_0 X_{0,i} + \beta_1 X_{1,i} + \epsilon_i)}{\sum_{i=1}^{N} (X_{1,i} - \bar{X}_1)^2} \\
\text{Fact:} E((X_1 - \bar{X}_1) X_0) &= 0 \\
\text{Fact:} E((X_1 - \bar{X}_1) X_1) &= E(X_1^2) - (\bar{X}_1)^2 \\
&= E(X_1 - \bar{X}_1)(X_1 - \bar{X}_1) = E(X_1 - \bar{X}_1)^2 \\
\text{Then, } \hat{\beta}_1 &= \beta_1 + \frac{\sum_{i=1}^{N} (X_{1,i} - \bar{X}_1) \epsilon_i}{\sum_{i=1}^{N} (X_{1,i} - \bar{X}_1)^2}
\end{aligned}$$

What is $E(\hat{\beta}_1)$?

$$
\begin{aligned}
E(\hat{\beta}_1) &= \beta_1 + E(\frac{\sum_{i=1}^{N}(X_{1,i} - \bar{X}_1)\epsilon_i}{\sigma^2_{X_1}}) \\
&= \beta_1 + \frac{1}{\sum_{i=1}^{N}(X_{1,i} - \bar{X}_1)} \sum_{i=1}^{N}(X_{1,i} - \bar{X}_1)E(\epsilon_i) \\
&= \beta_1
\end{aligned}
$$

$\hat{\beta}_1$ is an unbiased estimator of β_1.

What is $\mathrm{var}(\hat{\beta}_1^2)$?

$$
\begin{aligned}
\mathrm{var}(\hat{\beta}_1) &= E(\hat{\beta}_1 - \beta_1)^2 = E(\beta_1 + \frac{\sum_{i=1}^N (X_{1,i} - \bar{X}_1)\epsilon_i}{\sum_{i=1}^N (X_{1,i} - \bar{X}_1)^2} - \beta_1)^2 \\
&= E\frac{\sum_{i=1}^N (X_{1,i} - \bar{X}_1)\epsilon_i}{(\sum_{i=1}^N (X_{1,i} - \bar{X}_1)^2)^2} \\
&= \frac{1}{(\sum_{i=1}^N (X_{1,i} - \bar{X}_1)^2)^2} E(\sum_{i=1}^N (X_{1,i} - \bar{X}_1)\epsilon_i)^2 \\
&= \frac{1}{(\sum_{i=1}^N (X_{1,i} - \bar{X}_1)^2)^2} \sum_{i=1}^N ((X_{1,i} - \bar{X}_1)^2 E(\epsilon_i^2) \\
&\quad + \sum_{j=1, j \neq i}^N (X_{1,i} - \bar{X}_1)(X_{1,j} - \bar{X}_1) E(\epsilon_i \epsilon_j)) \\
&= \frac{\sum_{i=1}^N (X_{1,i} - \bar{X}_1)^2 \sigma_\epsilon^2}{(\sum_{i=1}^N (X_{1,i} - \bar{X}_1)^2)^2} = \frac{\sigma_\epsilon^2}{\sum_{i=1}^N (X_{1,i} - \bar{X}_1)^2}
\end{aligned}
$$

Distribution of ($\hat{\beta}_1$) and confidence intervals

- $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma_\epsilon^2}{\sum_{i=1}^N (X_{1,i} - \bar{X}_1)^2})$
- The 95% confidence interval for β_1 is:

$$\hat{\beta}_1 - 2\frac{\sigma^2}{\sum_{i=1}^N (X_{1,i} - \bar{X}_1)^2} \leq \beta_1 \leq \hat{\beta}_1 + 2\frac{\sigma^2}{\sum_{i=1}^N (X_{1,i} - \bar{X}_1)^2}$$

- Similarly, $\hat{\beta}_0 \sim N(\beta_0, \frac{\sigma^2}{\sum_{i=1}^N (X_{1,i} - \bar{X}_1)^2})$
  Using this, we can create a similar confidence interval for β_0

95% confidence intervals for the given data

- $\sum_{i=1}^{N} X_i = 48943$
- $\sum_{i=1}^{N} y_i = 19460.1$
- $\sum_{i=1}^{N} N = 3877$
- $\sum_{i=1}^{N} X_i^2 = 645663$
- $\sum_{i=1}^{N} y_i^2 = 99876$
- $\sum_{i=1}^{N} y_i X_i = 247775$
- $\hat{\beta}_0 = 4.06$
- $\hat{\beta}_1 = 0.076$
- $\hat{\sigma}^2 = 0.526$
- What is the 95% confidence interval for β_0, β_1?
- $\text{se}(\beta_0) = 0.056$, $\text{se}(\beta_1) = 0.0043$
- 95% CI: $3.95 \leq \beta_1 \leq 4.17$, $0.067 \leq \beta_1 \leq 0.085$

LR test, hypothesis, form

- $H_0 : \beta_1 = 0$
- LR statistic: $-2 log(L_R/L)$
- What is the form of L?

Likelihood evaluated at MLE

- L at the mle $(\hat{\beta}_0, \hat{;}\beta_1, \hat{\sigma}^2)$ works out to be:

$$L = (2\pi\hat{\sigma}^2 e)^{-\frac{n}{2}}$$

- The L can be expressed as only a function of $\hat{\sigma}^2$.
- Thus, we can rewrite the LR statistic as:

$$-2\left(\frac{\hat{\sigma}_R^2}{\hat{\sigma}_U^2}\right)^{-\frac{N}{2}}$$

Likelihood evaluated at MLE, and restriction

- Restricted model is: $\beta_1 = 0$ or

$$Y_i = \beta_0 + \epsilon_i$$

- β_0 works out to be $\bar{Y}$, and $\epsilon_i = (Y_i - \bar{Y})$.
- Therefoore: $\hat{\sigma}^2_{mle} = \hat{\sigma}^2_{\epsilon} = \hat{\sigma}^2_Y$
- Therefore, $\hat{\sigma}^2_R = \hat{\sigma}^2_Y$
- LR statistic is:

$$-2\log(\frac{\hat{\sigma}^2_Y}{\hat{\sigma}^2_{mle}})^{-\frac{N}{2}} = -N\log(\frac{\hat{\sigma}^2_{mle}}{\hat{\sigma}^2_Y})$$

- But we know that $\hat{\sigma}^2_{mle} = (1 - r^2_{XY})\sigma^2_Y$
- Therefore:

$$\frac{\hat{\sigma}^2_Y}{\hat{\sigma}^2_{mle}} = (1 - r^2_{XY})^{-N/2}$$

$$\mathrm{LR} = -N\log(1 - r^2_{XY})$$

LR test distribution

- Two tailed test distribution: $\chi^2(1)$
- One tailed test statistic: $\omega = (\text{sign H}_\text{A})\sqrt{(\text{LR})} \sim N(0, 1)$