The two-variable gaussian distribution model

Susan Thomas
IGIDR, Bombay

18 September, 2008

Recap

- MLE for single variable data, one-parameter distribution (birth data, Bernoulli)
- MLE for single variable data, two-parameter distribution (wage data, log-normal/normal)
- MLE for two-variable data, one-parameter conditional distribution, explained by multiple parameters (workforce participation-education, Bernoulli with logit transformation)

The two-variable regression model

- Variable to model: wages. Data: weekly wages, in USD.
- Additional information: education, in number of years.
- What we want to estimate: the expected wage – conditional on education.

Data description

- As with workforce participation, we check whether the conditional behaviour of log(wages) y is different from the unconditional behaviour.
- Unlike workforce participation, the comparison is done on the conditional density distribution of y.
- Observed: the median of y conditional on education increases with length of schooling.
- Conditional variance varies less with length of schooling: the range of values of y conditional on schooling does not change.
- *Suggested model*: y with varying conditional expectation but with unconditional variance.

The model

- Transform wages W to log(wages) w – lognormal –> normal
- Model for w:
  1. (X, y) pairs are independent
  2. Variable X is exogenous
  3. Conditional normality -
     1. E$(y|X = X_i) = (\beta_0 + beta_1 X_i)$,
     2. Variance is unconditional, σ^2, not $\sigma_i^2 = f(X_i)$.
- Model: $E(y|X = X_i) \sim N((\beta_0 + \beta_1 X_i), \sigma^2)$
- Model parameter space: $\beta_0, \beta_1, \sigma^2$
- $y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

Model parameter interpretation

- β_0 is $E(y)$ with no schooling.
- β_1 is the marginal increase in $E(y)$ with one more year of schooing.
  Unlike the logit model, where β_1 was the log(odds) ration – marginal increase in the log(odds) of workforce participation.
- What is the unconditional $E(y)$?

$$\begin{aligned} E(y|X_j) &= \beta_0 + \beta_1 X_j \\ E(y) &= \sum_{j=0}^{J} (\beta_0 + \beta_1 X_j) f(X_j) \\ &= \beta_0 \times 1 + \beta_1 \sum_{j=0}^{J} X_j f(X_j) \\ &= \beta_0 + \beta_1 E(X) \end{aligned}$$

- Syntax: β_0, β_1 are called “regression coefficients”.

Setting up the log(likelihood) for estimation

- Since y is conditionally gaussian distributed:

$$
\begin{aligned}
f_{\beta_0,\beta_1,\sigma^2}(y_1,..,y_N|X_1,..,X_N) &= \frac{1}{\sqrt{(2\pi\sigma^2)}} e^{-\frac{1}{2}\frac{(y_i - E(y_i))^2}{\sigma^2}} \\
&= \frac{1}{\sqrt{(2\pi\sigma^2)}} e^{-\frac{1}{2}\frac{(y_i - \beta_0 - \beta_1 X_i)^2}{\sigma^2}}
\end{aligned}
$$

- Using this, we set up the L and the log(L), l as:

$$
\begin{aligned}
L &= \prod_i \frac{1}{\sqrt{(2\pi\sigma^2)}} e^{-\frac{1}{2}\frac{(y_i - \beta_0 - \beta_1 X_i)^2}{\sigma^2}} \\
&= \frac{1}{\sqrt{(2\pi\sigma^2)}} e^{-\frac{1}{2}\sum_{i=1}^{N}\frac{(y_i - \beta_0 - \beta_1 X_i)^2}{\sigma^2}} \\
l &= -\frac{N}{\sqrt{(2\pi\sigma^2)}} - \frac{1}{2}\sum_{i=1}^{N}\frac{(y_i - \beta_0 - \beta_1 X_i)^2}{\sigma^2}
\end{aligned}
$$

Maximising L, minimising l

- The optimisation now involves *three* parameters: $\beta_0, \beta_1, \sigma^2$. This gives three equations:

$$
\begin{aligned}
\frac{\partial}{\partial \beta_0} l &= -2 \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 X_i) \\
\frac{\partial}{\partial \beta_1} l &= -2 \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 X_i) X_i \\
\frac{\partial}{\partial \sigma^2} l &= -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 X_i)^2
\end{aligned}
$$

- Set these to zero, and we get three equations to solve for $(\beta_0, \beta_1, \sigma^2)$.

Solution for β_0, β_1

- We see that the equation for σ^2 involves both β_0, β_1.
- Equations for β_0, β_1 do not have σ^2, so we solve for them first.
- Solution for β_0:

$$
\begin{aligned}
\frac{\partial}{\partial \beta_0} l &= -2 \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 X_i) = 0 \\
\sum_{i=1}^{N} (y_i) - N\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{N} (X_i) &= 0 \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{X}
\end{aligned}
$$

Solution for β_1

- Solve for the β_1 using $\frac{\partial}{\partial \beta_1} l = -2\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 X_i)X_i = 0$

$$\hat{\beta}_1 \sum_{i=1}^{N} X_i^2 = \sum_{i=1}^{N}(y_i X_i - \bar{X}(\bar{y} - \hat{\beta}_1 \bar{X}))$$

$$\hat{\beta}_1 (\sum_{i=1}^{N} X_i^2 - \bar{X}^2) = \sum_{i=1}^{N}(y_i X_i - \bar{X}\bar{y})$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} y_i X_i - N\bar{X}\bar{y}}{\sum_{i=1}^{N} X_i^2 - \bar{X}^2}$$

$$= \frac{cov\hat{(}Xy)}{\hat{s}_X^2}$$

$$\text{And, } \hat{\beta}_0 = \bar{y} - \bar{X}\frac{cov\hat{(}Xy)}{\hat{s}_X^2}$$

Sample correlation between y, X, $r_{(X,Y)}$

- We see MLE for the regression coefficients is a function of the sample correlation between the dependent variable y and the independent variable X.
- Sample correlation, $r_{(x,y)} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2 \sum_{i=1}^{N}(Y_i - \bar{Y})^2}}$
- This has properties:
  1. It is free of the unit of X, Y.
     Linear transformations of X, Y does not affect the value of $r_{(X,Y)}$. Eg., $r_{(X,Y)} = r_{(aX, b+cY)}$
     Non-linear transformations of X, Y do affect the value of $r_{(X,Y)}$ Eg., $r_{(X,Y)} \neq r_{(\log X, Y)}$
  2. $r_{(X,Y)} = 0$ if the sample covariance is zero.
  3. $r_{(X,Y)}$ takes values between -1 and +1.
     If $Y_i = a + bX_i$, then $r_{(X,Y)} = 1$

Population correlation between y, X, $\rho_{(X,Y)}$

- $\rho_{(X,Y)} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sqrt{E(X-\mu_X)^2 E(Y-\mu_Y)^2}}$
- It has the same properties as $r_{(X,Y)}$:
  - invariance across linear transformations,
  - $\rho_{(X,Y)} = 0$ if X, Y are uncorrelated,
  - $\rho^2_{(X,Y)} \leq 1$ unless $Y = a + bX$ in which case it $\rho_{(X,Y)} = 1$.

$r_{(X,Y)}$ and MLE $\hat{\beta}_1$

- The model clearly specifies the flow of the relationship between X, Y.
- However, consider the following pair of equations:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i \\ X_i &= \gamma_0 + \gamma_1 Y_i + \eta_i \end{aligned}$$

- We know $r_{X,Y} = r_{Y,X}$
  Then, $r^2_{(X,Y)} = \hat{\beta}_1 \hat{\gamma}_1$.
- When $\sigma_x = \sigma_y$, then $\hat{\beta}_1 = r_{X,Y} = r_{Y,X} = \hat{\gamma}_1$.

Example with data

- $\sum_{i=1}^{N} X_i = 48943$
- $\sum_{i=1}^{N} y_i = 19460.1$
- $\sum_{i=1}^{N} N = 3877$
- $\sum_{i=1}^{N} X_i^2 = 645663$
- $\sum_{i=1}^{N} y_i^2 = 99876$
- $\sum_{i=1}^{N} y_i X_i = 247775$
- What is $\beta_0, \beta_1, \sigma^2$?

Example with data

- $\hat{\beta}_0 = 4.06$
- $\hat{\beta}_1 = 0.076$
- $\hat{\sigma}^2 = 0.526$

$r_{(X,Y)}$ and MLE $\hat{\sigma}^2$

- To understand the relationship between $r_{(X,Y)}$ and MLE $\hat{\sigma}^2$, we need to understand the relationship between X and Y better.
- Consider a new form of the X, Y relationship:

$$Y_i = \beta_0 X_{0,i} + \beta_1 X_{1,i} + \epsilon_i$$

where $X_{0,i} = 1$ and $X_{1,i}$ is the following regression equation

$$X_i = \gamma_1 1 + \omega_i = \gamma_1 X_{0,i} + \omega_i$$

Interpreting equation: $X_i = \gamma_1 1 + \epsilon_i$

- X_i is a continuous rv that can take any value.
- ϵ_i is gaussian distributed.
- The MLE $\hat{\gamma}_1$ will be:

$$\hat{\gamma}_1 = \frac{1}{N} \sum_{i=1}^{N} X_i = \bar{X}$$

- Thus, this regression equationt gives you the value of the sample mean of the dependent variable as the regression coefficient, γ_1.

Reparameterised model $Y_i|X_i$

- We can then rewrite $Y_i = \beta_0 X_{0,i} + \beta_1 X_{1,i} + \epsilon_i$ as

$$
\begin{aligned}
Y_i &= \beta_0 X_{0,i} + \beta_1(\bar{X}X_{0,i} + (X_i - \bar{X})) + \epsilon_i \\
&= (\beta_0 + \beta_1\bar{X})X_{0,i} + \beta_1(X_i - \bar{X}) + \epsilon_i \\
&= \delta_0 X_{0,i} + \delta_1(X_i - \bar{X}) + \epsilon_i
\end{aligned}
$$

- This is a useful reparameterisation because $X_{0,i}, (X_i - \bar{X})$ are orthogonal to each other (by definition).
- Two outcomes:

$$
\begin{aligned}
\sum_{i=0}^{N}(X_i - \bar{X}) &= 0 \\
\sum_{i=0}^{N} X_{0,i}(X_i - \bar{X}) &= 0
\end{aligned}
$$

MLE coefficients for the reparameterised model $Y_i | X_i$

- Differentiating l wrt δ_0 gives:

$$-2 \sum_{i=1}^{N} (Y_i - \delta_0 - \delta_1 (X_i - \bar{X})) = 0$$

  Solution: $\hat{\delta}_0 = \bar{Y}$

- Differentiating l wrt δ_1 gives:

$$-2 \sum_{i=1}^{N} (Y_i - \delta_0 - \delta_1 (X_i - \bar{X}))(X_i - \bar{X}) = 0$$

  Solution: $\hat{\delta}_1 = \frac{\sum_{i=1}^{N} Y_i (X_i - \bar{X})}{\sum_{i=1}^{N} (X_i - \bar{X})^2}$

$$\hat{\delta}_1 = \hat{\beta}_1!$$

  MLE estimates are the same across rescaling of data.

Link 1 between $r_{(X,Y)}$ and MLE $\hat{\sigma}^2$

- We found out that:

$$\hat{Y}_i = \beta_0 + \beta_1 X_i = \delta_0 + \delta_1 (X_i - \bar{X})$$

- Or $\beta_1 = \frac{\sum_{i=1}^{N} Y_i (X_i - \bar{X})}{\sum_{i=1}^{N} (X_i - \bar{X})^2}$
- Using the above equality of model and reparameterised model, we calculate that:

$$\hat{\sigma}^2 == (1 - r^2) \sum_{i=1}^{N} (Y_i - \bar{Y})^2$$

Link 2 between $r_{(X,Y)}$ and MLE $\hat{\sigma}^2$

- The one-variable model is a restriction on the model $Y_i = \beta_0 + \beta_1 X_i$ where the restriction is $\beta_1 = 0$.
- Variance of the one-variable model is $\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \hat{\sigma}_R^2$.
- Variance of the error from the unrestricted model is $\sum_{i=1}^{N}(Y_i - \beta_0 - \beta_1 X_i)^2 = \hat{\sigma}^2$
- The ratio is:

$$\frac{\hat{\sigma}^2}{\hat{\sigma}_R^2} = (1 - r^2)$$

Link 3 regression R^2

- The variance of the restricted model can be rewritten as:

$$n\sigma_R^2 = \sum_{i=1}^{N}((Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}))^2$$

where $\hat{Y}_i = \hat{\delta}_0 + \hat{\delta}_1(X_i - \bar{X})$

- This collapses to:

$$n\sigma_R^2 = \sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2$$

- The first term is the *residual sum of squared* or RSS $= n\hat{s}igma^2 = \sum_{i=1}^{N} \hat{\epsilon}_i^2$
- The second term is the *explained sum of squares* or ESS
- σ_R^2 is called the *total sum of squares* or TSS.
- Sample correlation, $r^2_{(X,Y)}$ is:

$$r^2_{(X,Y)} = 1 - \frac{\text{RSS}}{\text{TSS}} = \frac{\text{ESS}}{\text{TSS}}$$