Multiple variable models

Susan Thomas
IGIDR, Bombay

21 October, 2008

Recap

- For a simple one-exogenous variable model,

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

- β_0 is the intercept on the "regression line" and β_1 is the slope.
- The above equation is called the "population regression line".
- After estimation, we have:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + u_i$$

  which is called the "estimated/sample regression line"
- $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$, ie, the line passes through the mean of the dataset.
- $\hat{\beta}_1 = S_{xy}/S_{xx}$ where S_{xy} is the sample covariance and S_{xx} is the sample variance of the exogenous data X.
- $\hat{\sigma}^2 = (N-1)/N\hat{\sigma}^2$

Moving to multiple-variable models

Model $Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \epsilon_i$

- Extend the two variable, gaussian distribution model with intercept to include one more exogenous variable, X_2.
- Economic example: log wages (Y_i) as a function of education ($X_{1,i}$) and age ($X_{2,i}$). The model for log wages becomes:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + u_i$$

- The model is:
  - Independence: $Y_i, X_{1,i}, X_{2,i}$ are independent across i
  - Normality of Y_i conditional on $X_{1,i}, X_{2,i}$: $Y_i \sim N[\beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i}, \sigma^2]$
  - $X_{1,i}, X_{2,i}$ is exogenous
- Parameters: $\beta_0, \beta_1, \beta_2, \sigma^2$

Log Likelihood and MLE solutions

- $l = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(Y_i - \beta_0 - \beta_1 X_{1,i} - \beta_2 X_{2,i})^2$
- MLE solution involves differentiating log L wrt three parameters and setting each to zero: three equations, three unknowns.
- Solution space looks like:

$$\beta_1 = \frac{\sum_i Y_i X_{1.0.2,i}}{\sum_i X_{1.0.2,i}^2}$$

where

$$\begin{aligned} X_{1.0,i} &= X_i - \bar{X} \\ X_{1.0.2,i} &= X_{1,i} - \hat{X}_{1,i} = X_{1,i} - \bar{X}_{1,i} - \frac{cov X_1 X_2}{\mathrm{var} X_2} X_{2,i} \end{aligned}$$

Log Likelihood and OLS solutions

- Maximise the log L is the same as minimising the SSD:

$$\mathrm{SSD}(\beta_0, \beta_1, \beta_2) = \sum_{i=1}^{N} (Y_i - \beta_0 - \beta_1 X_{1,i} - \beta_2 X_{2,i})^2$$

- Here, the solution to minimising the SSE is the OLS solution:

$$\beta = (X'X)^{-1}(X'Y)$$

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \dots \\ Y_N \end{bmatrix}, \quad X = \begin{bmatrix} 1 & X_{1,1} & X_{2,1} \\ 1 & X_{1,2} & X_{2,2} \\ 1 & X_{1,3} & X_{2,3} \\ \dots & \dots & \dots \\ 1 & X_{1,N} & X_{2,N} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

- The OLS solution will be of the form:

$$\beta_1 = \frac{\sum_i Y_i X_{1.0.2,i}}{\sum_j X^2_{1.0.2,j}}$$

where the solution contains a “new” form of X_1 which is conditional on it’s partial correlation with X_2: (useful for interpreting the model).

Recap on reparameterisation in the two-variable model

- We started with:

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

  and reparameterised it as:

$$Y_i = \gamma_0 X_{0,i} + \gamma_1 X_{1.0,i} + \omega_i$$

  where $\hat{\gamma}_1 = \hat{\beta}_1$ and $X_{1.0,i} = (X_{1,i} - \bar{X})$

- This was convenient for interpretation: β_1 is the effect on Y_i of an additional unit increase in X_1.

Reparameterisation in the three-variable model

- Start with:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \epsilon_i \\ \epsilon_i &\sim N(0, \sigma^2) \end{aligned}$$

and reparameterised it as:

$$\begin{aligned} Y_i &= \beta_0 X_{0,i} + \beta_1 (X_{1,i} - \bar{X}_1) + \beta_2 (X_{2,i} - \bar{X}_2 - \alpha X_1) + \omega_i \\ &= \delta_0 X_{0,i} + \delta_1 X_{1.0,i} + \delta_2 X_{2.0,1,i} + \omega_i \end{aligned}$$

Where:

$$\begin{aligned} X_{1.0,i} &= (X_{1,i} - \bar{X}_1) \\ X_{2.0,1,i} &= (X_{2,i} - \bar{X}_2) - \alpha X_{1,i} \\ \delta_0 &= \beta_0 + \beta_1 \bar{X}_1 + \beta_2 \bar{X}_2 \\ \delta_1 &= \beta_1 + \alpha \beta_2, \quad \alpha = \frac{\sum_i (X_{2,i} - \bar{X}_2)(X_{1,i} - \bar{X}_1)}{\sum_i X_{1.0,i}^2} \\ \delta_2 &= \beta_2 \end{aligned}$$

Solution minimising SSE

- Take the first derivative of SSE wrt $\delta_0, \delta_1, \delta_2$:
  $\sum_i (Y_i - \delta_0 X_0 - \delta_1 X_{1.0,i} - \delta_2 X_{2.0,1,i})$
- We first solve for δ_2.

$$\frac{\partial SSE}{\partial \delta_2} = -2 \sum_i (Y_i - \delta_0 X_0 - \delta_1 X_{1.0,i} - \delta_2 X_{2.0,1,i}) X_{2.0,1,i}$$

- By construction:
  - $\sum_i X_0 . X_{1.0,i} = 0$
  - $\sum_i X_0 . X_{2.0,1,i} = 0$
  - $\sum_i X_{1.0,i} . X_{2.0,1,i} = 0$
- $\hat{\delta}_2$ solves for:

$$\begin{aligned}
0 &= \sum_i (Y_i X_{2.0,1,i} - \hat{\delta}_0 (X_0 . X_{2.0,1,i}) - \hat{\delta}_1 (X_{1.0,i} X_{2.0,1,i}) - \hat{\delta}_2 X^2_{2.0,1,i}) \\
0 &= \sum_i (Y_i - \hat{\delta}_2 X_{2.0,1,i}) X_{2.0,1,i} \\
\hat{\delta}_2 &= \sum_i Y_i X_{2.0,1,i} / \sum_i X^2_{2.0,1,i}
\end{aligned}$$

Solution minimising SSE

- $\hat{\delta}_2 = \sum_i Y_i X_{2.0,1,i} / \sum_i X^2_{2.0,1,i}$
- This is the partial correlation between $Y_i, X_{2,i}$ given $X_{1,i}$.

What is **partial correlation**?

- Given Y_i, X_i, standard correlation is $\rho_{y.x.z} = \frac{\sum_i (Y_i - \bar{Y})(X_i - \bar{X})}{\mathrm{var}(Y)\mathrm{var}(X)}$
- Partial correlations are correlations between Y, X, given a third variable Z.
  - First two models:

$$\begin{aligned} Y_i &= \alpha_0 X_0 + \alpha_1 Z_i + e_i \\ X_i &= \gamma_0 X_0 + \gamma_1 Z_i + u_i \\ \text{where } X_0 &= 1 \end{aligned}$$

  - Then

$$\begin{aligned} \hat{y}_{y.0,z,i} &= e_i \\ \hat{x}_{x.0,z,i} &= u_i \\ r_{y.x,z} &= \frac{\sum_i \hat{y}_{(y.0,z,i)} \hat{x}_{(x.0,z,i)}}{\sqrt{\sum_i \hat{y}^2_{(y.0,z,i)} \sum_i \hat{x}^2_{(x.0,z,i)}}} \end{aligned}$$

$r_{y.x,z}$ is the partial correlation between (Y, X) given Z.

Partial correlations and standard correlations

FYI: Partial correlations can be rewritten as functions of standard (pair-wise) correlations as:

$$r_{y.x,z} = \frac{r_{(y,z)} - r_{(y,x)} * r_{(x,z)}}{\sqrt{(1 - r^2_{(y,x)})(1 - r^2_{(x,z)})}}$$

Numerical example of $r_{y.x.z}$ and $r_{y.x}, r_{x.z}, r_{y.z}$

- Given **w** = log wages, **A** is age and **S** is years of schooling.
- Given: $r_{w,S} = 0.270, r_{w,A} = 0.115, r_{S,A} = -0.139.$
- What is the partial correlation between log wages and age, given schooling?

$$\begin{aligned} r_{w.A,S} &= \frac{r_{(w,A)} - r_{(w,S)} * r_{(S,A)}}{\sqrt{(1 - r^2_{(W,S)})(1 - r^2_{(S,A)})}} \\ &= \frac{0.115 - (0.270 * -0.139)}{\sqrt{(1 - 0.270^2)(1 - (-0.139)^2)}} \\ &= 0.1599 \sim 0.160 \end{aligned}$$

- Interpretation: For people with the same schooling, age explains around $r^2_{w,A.S} = 3\%$ of the variation in log wages.
- Calculate the partial correlation between log wages and schooling, given age?

$$r_{w.A,S} = \frac{r_{(w,A)} - r_{(w,S)} * r_{(S,A)}}{\sqrt{(1 - r^2_{(W,S)})(1 - r^2_{(S,A)})}} = \frac{0.270 - (0.115 * -0.139)}{\sqrt{(1 - 0.115^2)(1 - (-0.139)^2)}} \sim 0.291$$

Back to the reparameterised model

$$\begin{aligned} Y_i &= \beta_0 X_{0,i} + \beta_1 (X_{1,i} - \bar{X}_1) + \beta_2 (X_{2,i} - \bar{X}_2 - \alpha X_1) + \omega_i \\ &= \delta_0 X_{0,i} + \delta_1 X_{1.0,i} + \delta_2 X_{2.0,1,i} + \omega_i \end{aligned}$$

- $\hat{\delta}_2 = \sum_i Y_i X_{2.0,1,i} / \sum_i X^2_{2.0,1,i}$
- $\hat{\delta}_1 = \frac{\sum_i Y_i X_{1.0,i}}{\sum_i X^2_{1.0,i}}$
- $\hat{\delta}_0 = \bar{Y}$.
- Giving: $\hat{\beta}_2 = \hat{\delta}_2$,
  $\hat{\beta}_1 = \hat{\delta}_1 + \hat{c}ov(X_2, X_1) * s^2_{X_2} \hat{\beta}_2$
  $\hat{\beta}_0 = \hat{\delta}_0 - \hat{\delta}_1 \bar{X}_1 - \hat{\delta}_2 \bar{X}_2$

$$
\begin{aligned}
\hat{Y}_i &= \hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \hat{\beta}_2 X_{2,i} \\
&= \hat{\delta}_0 + \hat{\delta}_1 X_{1.0,i} + \hat{\delta}_2 X_{2.0,1,i} \\
\hat{u}_i &= Y_i - \hat{Y}_i \\
\mathrm{RSS} &= \sum_i \hat{u}_i^2 = N\hat{\sigma}^2
\end{aligned}
$$

An unbiased estimator for $\sigma^2 = s^2 = \frac{1}{N-3}\mathrm{RSS}$.

Intrepreting the parameters

- β_0 is the conditional expectation of Y_i when $X_{1,i} = X_{2,i} = 0$.

$$E(Y_i | X_{1,i} = 0, X_{2,i} = 0) = \beta_0$$

- β_1 is the marginal increase in Y_i for an additional increase in X_1 – *conditional on X_2 remaining the same*.
- Similarly, β_2 is the marginal increase in Y_i for an additional increase in X_1 – *conditional on X_2 remaining the same*.

What is new?

- A new twist in the estimation tale: correlation between $X_{1,i}$ and $X_{2,i}$.
- If there exists ρ_{X_1,X_2} that would affect the conditional mean of Y_i.
- What if $\rho_{X_1,X_2} = 0$?
  The two exogenous variables are orthogonal and contribute different information to Y_i. Two separate regressions can be run:
  $Y_i = \beta_0 + \beta_1 X_{1,i} + u_i$; $Y_i = \beta_0' + \beta_2 X_{2,i} + u_i'$
  and $\hat{\beta}_1, \hat{\beta}_2$ would be the same as in the joint estimation.
- What if if $\rho_{X_1,X_2} = 1$?
  Trouble in estimation!
  This is called "perfect collinearity" – using the same information through two different sources *and* trying to estimate two different parameters.
- More typically, $\rho_{X_1,X_2} \sim 1 \Rightarrow$ "near collinearity".