Einführung Mikroökonometrie

310208 (VO)

Dietmar Bauer

Wiederholung der Einführung in die Ökonometrie

- Im MLR-Modell können wir die Koeffizienten mittels Minimierung der Fehlerquadratsumme schätzen.
- Die so erhaltenen KQ-Schätzer (OLS) sind unter den Annahmen KLR unverzerrt, normalverteilt von minimaler Varianz (=BLUS).
- Die Varianz können wir unverzerrt schätzen.
- Durch den Vergleich der Koeffizienten mit deren geschätzter Standardabweichung können wir t-Tests durchführen auf die Nullhypothese $H_0 : \beta_j = 0$.
- Das R^2 gibt uns an, wie viel von der Variation der abhängigen Variablen durch das Modell erklärt wird.

Wiederholung der Einführung in die Ökonometrie

- Haben wir eine Variable zu viel im Modell, ist dies kein großes Problem. Die Varianz der geschätzten Koeffizienten steigt, die Tests sind aber alle gültig.
- Haben wir eine Variable zu wenig im Modell, erzeugt das möglicherweise eine Verzerrung, die Tests sind dann nicht mehr gültig (= wir haben keine Kontrolle über den Fehler erster Art α).
- Ist die Annahme der Homoskedastie (Varianz der Fehler hängt nicht von den Regressoren ab) verletzt, dann stimmen die Formeln für die Varianz nicht und der KQ-Schätzer ist nicht mehr von minimaler Varianz.

Gepoolte Querschnittsdaten

3. Gepoolte Querschnittsdaten

- Querschnittsdaten gehen von der Beobachtung der gleichen Variablen von verschiedenen Individuen zu einem Zeitpunkt (oder kurzem Zeitraum, etwa eine Erhebungswoche) aus.
- Manchmal werden die gleichen Erhebungen parallel von verschiedenen Gruppen durchgeführt (etwa in verschiedenen Ländern wie bei OECD Untersuchungen) und ergeben dann eine Sammlung an strukturell gleichen Datensätzen.
- Manchmal werden die gleichen Erhebungen mehrmals durchgeführt (etwa in unterschiedlichen Jahren) und geben so eine Sammlung an strukturell gleichen Datensätzen.

Wir betrachten hier also eine Zahl an Datensätzen
$(y_i^{(j])}, X_i^{(j)}; i = 1, ..., n(j)), j = 1, ..., J.$

3. Gepoolte Querschnittsdaten

Für jeden der Datensätze wird das Modell:

$$\mathbf{y}_i^{(j)} = (\beta^{(j)})'\mathbf{X}_i^{(j)} + \mathbf{u}_i^{(j)}$$

angenommen, wobei für jedes *j* die KLR Annahmen gelten sollen.

Frage: Wie können wir/sollen wir die Koeffizienten $\beta^{(j)}$ schätzen?

Oder spezieller: Untersuchen wir jeden Datensatz getrennt oder zahlt es sich aus, alle gemeinsam zu betrachten?

3. Gepoolte Querschnittsdaten

Für jeden der Datensätze wird das Modell:

$$\mathbf{y}_i^{(j)} = (\beta^{(j)})'\mathbf{X}_i^{(j)} + \mathbf{u}_i^{(j)}$$

angenommen, wobei für jedes j die KLR Annahmen gelten sollen.

Frage: Wie können wir/sollen wir die Koeffizienten $\beta^{(j)}$ schätzen?

Oder spezieller: Untersuchen wir jeden Datensatz getrennt oder zahlt es sich aus, alle gemeinsam zu betrachten?

Antwort: das kommt drauf an.

Zwei Extremfälle (für den Fall $J = 2$, um die Notation einfach zu halten)

1. $\beta^{(1)} = \beta^{(2)}$: Datensätze *poolen*.
2. $\beta^{(1)} \neq \beta^{(2)}$: Daten getrennt untersuchen.

3. Gepoolte Querschnittsdaten

Getrennte Analyse: dann gilt für jeden der beiden Datensätze die Annahme KLR und somit die Eigenschaften des KQ-Schätzers (normalverteilt, unverzerrt, BLUS).

Schätzung also mit KQ-Schätzung einzeln für jeden Datensatz in der Gleichung:

für $j = 1$:

$$\mathbf{y}_i^{(1)} = (\beta^{(1)})'\mathbf{X}_i^{(1)} + \mathbf{u}_i^{(1)}$$

Für $j = 2$:

$$\mathbf{y}_i^{(2)} = (\beta^{(2)})'\mathbf{X}_i^{(2)} + \mathbf{u}_i^{(2)}$$

Die Gleichungen kann man auch alle untereinander schreiben und gemeinsam schätzen, man bekommt dennoch die KQ-Schätzer des jeweiligen Datensatzes.

3. Gepoolte Querschnittsdaten

Stellt man die Beobachtungen des zweiten Datensatzes unter den ersten, kann man beide Gleichungen so gemeinsam schätzen.

In Matrixschreibweise:

$$
Y = \left(\begin{array}{c} \mathbf{y}_1^{(1)} \\ \mathbf{y}_2^{(1)} \\ \vdots \\ \mathbf{y}_{n(1)}^{(1)} \\ \hline \mathbf{y}_1^{(2)} \\ \vdots \\ \mathbf{y}_{n(2)}^{(2)} \end{array} \right), \quad
X = \left(\begin{array}{cccc|cccc}
1 & \mathbf{x}_{1,1}^{(1)} & \dots & \mathbf{x}_{1,K}^{(1)} & 0 & 0 & \dots & 0 \\
1 & \mathbf{x}_{2,1}^{(1)} & \dots & \mathbf{x}_{2,K}^{(1)} & 0 & 0 & \dots & 0 \\
\vdots & \vdots & & \vdots & \vdots & & & \vdots \\
1 & \mathbf{x}_{n(1),1}^{(1)} & \dots & \mathbf{x}_{n(1),K}^{(1)} & 0 & 0 & \dots & 0 \\
\hline
0 & 0 & \dots & 0 & 1 & \mathbf{x}_{1,1}^{(2)} & \dots & \mathbf{x}_{1,K}^{(2)} \\
\vdots & & & \vdots & \vdots & \vdots & & \vdots \\
0 & 0 & \dots & 0 & 1 & \mathbf{x}_{n(2),1}^{(2)} & \dots & \mathbf{x}_{n(2),K}^{(2)}
\end{array} \right)
$$

$$
Y = \begin{pmatrix} Y_{(1)} \\ Y_{(2)} \end{pmatrix}, \qquad X = \begin{pmatrix} X_{(1)} & 0 \\ 0 & X_{(2)} \end{pmatrix}, \qquad \alpha = \left(\frac{\beta_{(1)}}{\beta_{(2)}} \right)
$$

3. Gepoolte Querschnittsdaten

Das ergibt den KQ-Schätzer:

$$\begin{aligned}
\hat{\alpha} &= (X'X)^{-1}X'Y = \left[\begin{pmatrix} X'_{(1)} & 0 \\ 0 & X'_{(2)} \end{pmatrix}\begin{pmatrix} X_{(1)} & 0 \\ 0 & X_{(2)} \end{pmatrix}\right]^{-1}\begin{pmatrix} X'_{(1)}Y_{(1)} \\ X'_{(2)}Y_{(2)} \end{pmatrix} \\
&= \begin{bmatrix} (X'_{(1)}X_{(1)})^{-1} & 0 \\ 0 & (X'_{(2)}X_{(2)})^{-1} \end{bmatrix}\begin{bmatrix} X'_{(1)}Y_{(1)} \\ X'_{(2)}Y_{(2)} \end{bmatrix} \\
&= \begin{pmatrix} \hat{\beta}_{(1)} \\ \hat{\beta}_{(2)} \end{pmatrix}
\end{aligned}$$

Die gemeinsam geschätzten Koeffizienten sind also ident mit den getrennt geschätzten!

Andere Parametrisierung:

Betrachte die folgende Gleichung:

$$\mathbf{y}_i^{(j)} = (\beta^{(1)})'\mathbf{X}_i^{(j)} + \gamma'(\delta(j=2)\mathbf{X}_i) + \mathbf{u}_i^{(j)}, \qquad (*)$$

für $i = 1, ..., n(j), j = 1, 2$. Damit für $j = 1$:

$$\mathbf{y}_i^{(1)} = (\beta^{(1)})'\mathbf{X}_i + \mathbf{u}_i$$

Für $j = 2$:

$$\mathbf{y}_i^{(2)} = (\beta^{(1)})'\mathbf{X}_i + \gamma'\mathbf{X}_i + \mathbf{u}_i$$

und damit

$$\beta^{(1)} + \gamma = \beta^{(2)}$$

Die Koeffizienten gleich, dann und nur dann, wenn $\gamma = 0$.

3. Gepoolte Querschnittsdaten

Stellt man die Beobachtungen des zweiten Datensatzes unter den ersten, kann man beide Gleichungen so gemeinsam schätzen.

In Matrixschreibweise:

$$Y = \left(\begin{array}{c} \mathbf{y}_1^{(1)} \\ \mathbf{y}_2^{(1)} \\ \vdots \\ \mathbf{y}_{n(1)}^{(1)} \\ \hline \mathbf{y}_1^{(2)} \\ \vdots \\ \mathbf{y}_{n(2)}^{(2)} \end{array}\right), \quad X = \left(\begin{array}{cccc|cccc} 1 & \mathbf{x}_{1,1}^{(1)} & \dots & \mathbf{x}_{1,K}^{(1)} & 0 & 0 & \dots & 0 \\ 1 & \mathbf{x}_{2,1}^{(1)} & \dots & \mathbf{x}_{2,K}^{(1)} & 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & & \vdots \\ 1 & \mathbf{x}_{n(1),1}^{(1)} & \dots & \mathbf{x}_{n(1),K}^{(1)} & 0 & 0 & \dots & 0 \\ \hline 1 & \mathbf{x}_{1,1}^{(2)} & \dots & \mathbf{x}_{1,K}^{(2)} & 1 & \mathbf{x}_{1,1}^{(2)} & \dots & \mathbf{x}_{1,K}^{(2)} \\ \vdots & & & \vdots & \vdots & \vdots & & \vdots \\ 1 & \mathbf{x}_{n(2),1}^{(2)} & \dots & \mathbf{x}_{n(2),K}^{(2)} & 1 & \mathbf{x}_{n(2),1}^{(2)} & \dots & \mathbf{x}_{n(2),K}^{(2)} \end{array}\right)$$

$$Y = \left(\begin{array}{c} Y_{(1)} \\ Y_{(2)} \end{array}\right), \qquad X = \left(\begin{array}{c|c} X_{(1)} & 0 \\ \hline X_{(2)} & X_{(2)} \end{array}\right), \qquad \alpha = \left(\begin{array}{c} \beta^{(1)} \\ \hline \gamma \end{array}\right)$$

3. Gepoolte Querschnittsdaten

Das ergibt den KQ-Schätzer:

$$
\begin{aligned}
\hat{\alpha} &= (X'X)^{-1}X'Y \\
&= \begin{pmatrix} (X'_{(1)}X_{(1)}) + (X'_{(2)}X_{(2)}) & (X'_{(2)}X_{(2)}) \\ (X'_{(2)}X_{(2)}) & (X'_{(2)}X_{(2)}) \end{pmatrix}^{-1} \begin{pmatrix} (X'_{(1)}Y_{(1)} + (X'_{(2)}Y_{(2)})) \\ (X'_{(2)}Y_{(2)})) \end{pmatrix} \\
&= \left(\begin{bmatrix} 0 & 0 \\ 0 & (X'_{(2)}X_{(2)})^{-1} \end{bmatrix} + \begin{bmatrix} I_{K+1} \\ -I_{K+1} \end{bmatrix} (X'_{(1)}X_{(1)})^{-1} \begin{bmatrix} I_{K+1} & -I_{K+1} \end{bmatrix} \right) \left[\begin{array}{c} (X'_{(1)}Y_{(1)} + \\ (X'_{(2)}Y \end{array} \right. \\
&= \begin{pmatrix} 0 \\ \hat{\beta}_{(2)} \end{pmatrix} + \begin{bmatrix} I_{K+1} \\ -I_{K+1} \end{bmatrix} \hat{\beta}_{(1)} = \begin{pmatrix} \hat{\beta}_{(1)} \\ \hat{\beta}_{(2)} - \hat{\beta}_{(1)} \end{pmatrix}
\end{aligned}
$$

Daher ist der KQ-Schätzer für γ gleich der Differenz der einzeln geschätzten Gleichungen.

Nachdem wir beide Datensätze in einer Gleichung haben, können wir die KQ-Theorie verwenden, um folgendes Theorem zu zeigen:

Theorem (Chow-Test)

Wenn das gemeinsame Modell (*) die Annahmen KLR erfüllt, dann ist die Test-Statistik

$$F = (\hat{\beta}_{(2)} - \hat{\beta}_{(1)})'(M_1^{-1} + M_2^{-1})^{-1}(\hat{\beta}_{(2)} - \hat{\beta}_{(1)})/((K+1)\hat{\sigma}^2)$$

mit $M_j = X'_{(j)}X_{(j)}$ verteilt nach einer F-Verteilung mit $K+1$ und $n(1) + n(2) - 2(K+1)$ Freiheitsgraden.

Da die beiden Datensätze unabhängig voneinander sind, sind auch die beiden Schätzer $\hat{\beta}_{(1)}$ und $\hat{\beta}_{(2)}$ voneinander unabhängig. Daher gilt:

$$\mathrm{Var}(\hat{\beta}_{(2)} - \hat{\beta}_{(1)}) = \mathrm{Var}(\hat{\beta}_{(1)}) + \mathrm{Var}(\hat{\beta}_{(2)}) = \sigma^2 M_2^{-1} + \sigma^2 M_1^{-1} = \sigma^2(M_1^{-1} + M_2^{-1})$$

Chow-Test:

- Nullhypothese: $H_0 : \gamma = 0$.
- Gegenhypothese: $H_1 : \gamma \neq 0$.
- Test-Statistik: F verteilt nach F mit $K + 1$ und $n(1) + n(2) - 2(K + 1)$ Freiheitsgraden.
- Test mit Konfidenz $1 - \alpha$ verwirft, wenn p-Wert kleiner als α ist.

3. Gepoolte Querschnittsdaten

Man kann zeigen, dass

$$F = \frac{(SSR_p - SSR_1 - SSR_2)}{SSR_1 + SSR_2} \frac{n(1) + n(2) - 2(K+1)}{K+1}$$

wobei SSR_1 und SSR_2 die Residuenquadratsumme der gemeinsamen Schätzung sind und SSR_p jene, die unter der Nullhypothese, also in einem Modell mit $\gamma = 0$ erhalten würde.

Interpretation:

- Darstellung von F mit Differenz der Schätzer zeigt, dass der Test verwirft, wenn die geschätzten Koeffizienten zu unterschiedlich sind.
- Darstellung mit Residuenquadratsumme zeigt, dass der Test dann verwirft, wenn die Residuenquadratsumme stark verringert wird, wenn man in den beiden Datensätzen unterschiedliche Koeffizienten zulässt.

Beweis der Darstellung: Der Beweis benutzt die Blockmatrizeninversion:

$$(X'X)^{-1} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -A^{-1}B \\ I \end{bmatrix} (\underbrace{D - CA^{-1}B}_{D_\perp})^{-1} \begin{bmatrix} -CA^{-1} & I \end{bmatrix}$$

wobei $A = M_1 + M_2$ und $B = C = D = M_2$. Dann gilt $\hat{\gamma} = [0, I](X'X)^{-1}X'Y$ und somit

$$\begin{aligned}
\hat{\gamma}'(M_1^{-1} + M_2^{-1})^{-1}\hat{\gamma} &= Y'X(X'X)^{-1} \begin{bmatrix} 0 \\ I \end{bmatrix} (D_\perp^{-1})^{-1}[0, I](X'X)^{-1}X'Y \\
&= Y'X \left(\begin{bmatrix} -A^{-1}B \\ I \end{bmatrix} D_\perp^{-1} D_\perp D_\perp^{-1} \begin{bmatrix} -CA^{-1} & I \end{bmatrix} \right) X'Y \\
&= Y'X \left((X'X)^{-1} - \begin{bmatrix} I \\ 0 \end{bmatrix} (M_1 + M_2)^{-1}[I, 0] \right) X'Y \\
&= Y' \left(X(X'X)^{-1}X' - I + I - X_1(M_1 + M_2)^{-1}X_1' \right) Y \\
&= Y'(I - X_1(X_1'X_1)^{-1}X_1)Y - Y'(I - X(X'X)^{-1}X')Y
\end{aligned}$$

Diese Ausdrücke kennen wir als Residuenquadratsumme der Regression von Y auf X_1, die ersten Spalten der Datenmatrix X (das ergibt SSR_p) beziehungsweise auf X (das ist $SSR_1 + SSR_2$).

3. Gepoolte Querschnittsdaten

Der Chow-Test verwendet auch die Annahme, dass in beiden Datensätzen die Varianzen der Fehler identisch sind.

- FGLS Schätzer ist leicht zu konstruieren:
  - Schätze in jedem der beiden Datensätze mittels KQ-Schätzung die Varianz $\hat{\sigma}^2_{(j)}$.
  - Berechne WLS mit Gewichten $1/\hat{\sigma}_{(j)}$ (Gewicht $1/\hat{\sigma}_{(1)}$ in allen Beobachtungen in Datensatz 1 und Gewicht $1/\hat{\sigma}_{(2)}$ in allen Beobachtungen in Datensatz 2).
- Der Chow-Test kann dann mit den FGLS Schätzern berechnet werden. Das gibt die gleichen Schätzer, aber andere Varianzen.

3. Gepoolte Querschnittsdaten

Datensatz *CPS*78_85.*RAW* in Wooldridge: Durchschnittseinkommen als Funktion der Ausbildung.

- Wir haben schon gesehen, dass die Ausbildung und die Erfahrung und einige andere Variable einen Einfluss haben.
- Wir haben auch gesehen, dass Logarithmen zu Homoskedastizität führen.
- Jetzt werden zwei Wellen der Erhebung in 78 und 85 gemeinsam verwendet.
- Die Inflation ist nicht berücksichtigt, sodass schon allein deswegen damit gerechnet werden muss, dass die beiden Datensätze nicht mit dem gleichen Modell beschrieben werden können: Die Inflation kann als multiplikativer Faktor gesehen werden.

3. Gepoolte Querschnittsdaten

```
Call:
lm(formula = lwage ~ educ + south + female + exper + expersq +
    union)

Residuals:
     Min       1Q   Median       3Q      Max
-2.69617 -0.27930 -0.01868  0.29430  2.31078

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.48834432  0.08518813   5.733 0.000  ***
educ         0.08692861  0.00560766  15.502 < 0.00 ***
south       -0.06856410  0.03089558  -2.219 0.0267 *
female      -0.25227665  0.02837287  -8.891 < 0.00 ***
exper        0.03210662  0.00392788   8.174 0.000  ***
expersq     -0.00045035  0.00008542  -5.272 0.000  ***
union        0.13216848  0.03341396   3.955 0.000  ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4551 on 1077 degrees of freedom
Multiple R-squared: 0.3009,Adjusted R-squared: 0.297
F-statistic: 77.24 on 6 and 1077 DF,  p-value: < 2.2e-16
```

3. Gepoolte Querschnittsdaten

```
Call:
lm(formula = lwage ~ educ + south + female + exper + expersq +
    union, subset = (y85 == 0))

Residuals:
     Min       1Q   Median       3Q      Max
-2.55578 -0.23808  0.01535  0.21889  1.31010

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.4677126  0.0983131   4.757 2.52e-06 ***
educ         0.0765125  0.0065206  11.734  < 2e-16 ***
south       -0.0157631  0.0370184  -0.426   0.6704
female      -0.3156381  0.0344931  -9.151  < 2e-16 ***
exper        0.0249556  0.0047169   5.291 1.77e-07 ***
expersq     -0.0002857  0.0001016  -2.812   0.0051 **
union        0.2008916  0.0375991   5.343 1.35e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.387 on 543 degrees of freedom
Multiple R-squared: 0.3835,Adjusted R-squared: 0.3766
F-statistic: 56.29 on 6 and 543 DF,  p-value: < 2.2e-16
```

3. Gepoolte Querschnittsdaten

```
Call:
lm(formula = lwage ~ educ + south + female + exper + expersq +
    union, subset = (y85 == 1))

Residuals:
     Min       1Q   Median       3Q      Max
-2.22617 -0.27678 -0.00632  0.28205  2.10718

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.6560242  0.1214769   5.400 1.01e-07 ***
educ         0.0880661  0.0079352  11.098  < 2e-16 ***
south       -0.1024121  0.0422090  -2.426 0.015588 *
female      -0.2337426  0.0385604  -6.062 2.57e-09 ***
exper        0.0336233  0.0053604   6.273 7.41e-10 ***
expersq     -0.0005097  0.0001177  -4.329 1.79e-05 ***
union        0.1916060  0.0504258   3.800 0.000162 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.436 on 527 degrees of freedom
Multiple R-squared: 0.3251,Adjusted R-squared: 0.3174
F-statistic:  42.3 on 6 and 527 DF,  p-value: < 2.2e-16
```

3. Gepoolte Querschnittsdaten

Die Zahlen ändern sich, aber ist das genug, um auf andere Koeffizienten schließen zu können?

Chow-Test:

SSR_p	SSR_1	SSR_2	n	K	F	F_c	p
223.09	81.32	100.19	1084	6	35.02	1.95	0

p-Wert kleiner als 0.05 $\Rightarrow$ Test verwirft Gleichheit der Koeffizienten ganz klar.

Das sagt uns aber noch nichts, warum der Test verwirft.

3. Gepoolte Querschnittsdaten

```
Call: lm(formula = lwage ~ y85 + educ + I(south * y85) + I(y85 * educ) +
    exper + expersq + union + female + I(y85 * female))

Residuals:
     Min       1Q   Median       3Q      Max
-2.55656 -0.25804  0.00597  0.26069  2.08614

Coefficients:
                  Estimate  Std. Error t value             Pr(>|t|)
(Intercept)      0.46595135  0.09323821   4.997 0.000000677924112228 ***
y85              0.18037940  0.12577523   1.434              0.15182
educ             0.07458564  0.00665882  11.201 < 0.0000000000000002 ***
I(south * y85)  -0.10318974  0.03968689  -2.600              0.00945 **
I(y85 * educ)    0.01599902  0.00937712   1.706              0.08826 .
exper            0.02917736  0.00356123   8.193 0.000000000000000719 ***
expersq         -0.00039218  0.00007738  -5.068 0.000000473075546964 ***
union            0.19796150  0.03025619   6.543 0.000000000093246480 ***
female          -0.31740547  0.03652471  -8.690 < 0.0000000000000002 ***
I(y85 * female)  0.08345961  0.05117571   1.631              0.10322
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4116 on 1074 degrees of freedom
Multiple R-squared: 0.4298,Adjusted R-squared: 0.425
F-statistic: 89.94 on 9 and 1074 DF,  p-value: < 0.00000000000000022
```