Andreas Haupt, Gerd Nollmann

Technischer Anhang A zu:

Warum werden immer mehr Haushalte von Armut gefährdet?

Zur Erklärung erhöhter Armutsrisikoquoten mit unbedingten Quantilregressionen

A1. Das Gesetz iterierter Erwartungen

Wenn sich eine Verteilung der Variable Y über mehrere Subgruppen aufteilt, die durch die Variablen X_1, X_2 … X_n beschrieben werden und die Gruppen voneinander unabhängig sind, dann trägt jede gruppenspezifische Verteilung in je spezifischer Weise zur gesamten Verteilung bei. Die Art des Beitrags ist eine Funktion aus der Wahrscheinlichkeit der Werte innerhalb der Gruppe und der Wahrscheinlichkeit, dass ein Wert Teil dieser Gruppe ist.

$$Y = \sum y = \sum_{x=1}^{J} \sum_{y \in \chi} y\, P(Y|X) P(X),$$

wobei χ der Wertebereich von Y ist. Wenn die Gruppen voneinander unabhängig sind, lässt sich der Erwartungswert der gesamten Verteilung von Y als Summe der gewichteten gruppenspezifischen Erwartungswerte ausdrücken.

$$\begin{aligned} E(Y) = E(\sum y) &= \sum_{x=1}^{J} \sum_{y \in \chi} y \cdot P(Y|X) \cdot P(X)) \\ &= E[E(Y|X) \cdot P(X)] \\ &= E_x[E(Y|X)]. \end{aligned}$$

$E_x[]$ bedeutet, dass die Erwartungswertbildung über alle Subgruppen von X erfolgt. Die Aussage, dass der unbedingte Erwartungswert E(Y) als Summe bedingter Erwartungswerte ausgedrückt werden kann, wird als Gesetz iterierter Erwartungen bezeichnet.

A2. Bedingte und unbedingte Interpretation von OLS-Koeffizienten

OLS-Regressionen führen die Werte von abhängigen Variablen auf Erwartungswerte zurück. Die Regressionskoeffizienten einer OLS-Regression wird typischerweise als Steigerung eines bedingten Erwartungswertes interpretiert, wenn X um eine Einheit steigt. Formal lässt sich dies im Fall einer Dummy-Variable ausdrücken als:

$$\beta_{OLS} = E(y|X=1) - E(y|X=0).$$

Einer linearen Regression liegt die Annahme zugrunde, dass sich die Werte von Y als Kombination aller subgruppenspezifischen Erwartungswerte und eines Fehlerterms beschreiben lassen:

$$y = E(y|X) + \epsilon.$$

Unter der Annahme strikter Exogenität folgt:

$$E(y) = E_x[E(y|X)] + E(\epsilon)$$
$$= E(y|X)P(X).$$

Wird auf diesen Ausdruck das Gesetz iterierter Erwartungen angewendet, ergibt sich im Fall einer Regression mit nur einer Dummy-Variablen somit:

$$\begin{aligned}
\mathrm{E(y)} &= \sum E(y|X)P(X) \\
&= \mathrm{E(y|X = 0)P(X = 0) + E(y|X = 1)P(X = 1)} \\
&= \bar{\mathrm{y}}_{\mathrm{X=0}}\mathrm{P(X = 0)} + \bar{\mathrm{y}}_{\mathrm{X=1}}\mathrm{P(X = 1)} \\
&= \bar{\mathrm{y}}_{\mathrm{X=0}}\mathrm{P(X = 0)} + \bar{\mathrm{y}}_{\mathrm{X=1}}\big(1 - \mathrm{P(X = 0)}\big)
\end{aligned}$$

β_{OLS} drückt somit nicht nur Differenzen von *bedingten* Erwartungswerten, sondern auch Veränderungen des *unbedingten* Erwartungswertes durch die bedingten Erwartungswerte aus. Die unbedingte Interpretation des Koeffizienten lautet dann: Durch die Erhöhung von X um eine Einheit erhöht sich der Erwartungswert (Mittelwert) der gesamten Verteilung um Z. Wird der letzte Ausdruck der Gleichung nach P(X=0) abgeleitet, ergibt sich:

$$\frac{\delta E(y)}{\delta P(X = 0)} = E(y|X = 0) - E(y|X = 1) = \beta_{OLS}$$

A3. Bedingte und unbedingte Quantilregressionen

A3.1 Bedingte Quantilregressionen

Bedingte Quantilregressionen (BQR), die auf dem Verfahren von Koenker/Basset (1978) beruhen, schätzen Unterschiede der Werte subgruppenspezifischer Verteilungen für identische Quantile τ. Im Fall des 50. Quantils geben sie den Unterschied des Medianwertes an, wenn X um eine Einheit steigt. Der Wert eines Quantil τ einer Verteilung $F_Y(y)$ mit $0 < \tau < 1$ ist definiert als

$$Q(\tau) := F_Y^{-1}(\tau) = \inf\{y : F_Y(y) \geq \tau\}.$$

Im Fall des Wertes für das 20. Quantil wird also derjenige Wert geschätzt, der entweder exakt die Grenze zwischen den unteren 20% und oberen 80% der kumulierten Einkommensverteilung definiert. Ist dies nicht möglich, wird der nächsthöhere Wert als Schätzung genutzt. Das Verfahren von Koenker/Basset (1978) wird als bedingte Quantilregression bezeichnet, weil die geschätzten Koeffizienten lediglich Differenzen subgruppenspezifischer Lageparameter ausdrücken können. Wie bei einer OLS-Regression wird angenommen, dass jedes X eine Subgruppe mit einer eigenen Verteilung von Y darstellt. Anders als OLS-Koeffizienten haben Schätzungen der bedingten Quantilregression nur eine mögliche Interpretation:

$$\beta_{BQR} = Q(\tau|X = 1) - Q(\tau|X = 0).$$

Eine unbedingte Interpretation ist nicht möglich. Für eine unbedingte Interpretation müsste eine lineare Quantilregression der Form

$$Q_Y(\tau|X) = X'\beta_\tau$$

analog unter Zuhilfenahme des Gesetzes iterierter Erwartungen umformbar sein in

$$E_x[Q_Y(\tau|X)] = E_x[X'\beta_\tau] = \bar{X}'\beta_\tau \rightarrow Q_Y(\tau).$$

Wäre die letzte Umformung möglich, müsste im Fall eine Regression auf den Median mit einer Dummyvariablen gelten:

$$E_x[Q(.5)|\, X\,] = E_x[\beta_{0\tau} + \beta_{1\tau}X] = \beta_{0\tau} + \beta_{1\tau}\bar{X} \equiv Q(.5).$$

Da dies nicht der Fall ist, sind alle Verfahren, die eine unbedingte Interpretation von Koeffizienten benötigen, nicht mit BQR kombinierbar.

A3.2 Unbedingte Quantilregressionen

Für die Schätzung von Gruppeneinflüssen auf die Lage unbedingter Quantile wird die abhängige Variable (in unserem Fall die realen Haushaltsnettoäquivalenzeinkommen) durch die *rezentrierte Einflussfunktion (RIF)* des Quantils ersetzt. Für Quantile ist diese definiert als (Fortin et al. 2011: 76):

$$RIF(y; Q_\tau) = Q_\tau + \frac{\tau - \mathbb{I}(y \leq Q_\tau)}{f_Y(Q_\tau)},$$

wobei $\mathbb{I}(\cdot)$ eine Indikatorfunktion, $f_Y(Q_\tau)$ die Dichte der Haushaltsnettoäquivalenzeinkommen am Quantil τ und Q_τ den Sample-Quantilwert für τ darstellen. Die Einflussfunktion weist definitionsgemäß nur zwei verschiedene Werte auf. In der Periode 2009/11 ist der Einkommenswert für das 15. Quantil 9.257, die Dichte am 15. Quantil ist .391479. Dies ergibt für diesen Zeitpunkt zwei Werte der rezentrierten Einflussfunktion am 15. Quantil:

$$RIF(y, Q_{0.15} = 9.257) = \begin{cases} 9.257 + \dfrac{0.15 - 1}{.391479} \;\; wenn \;\; y < 9.257 = 7.085 \\ 9.257 + \dfrac{0.15 - 0}{.391479} \;\; wenn\; y \geq 9.257 = 9.640 \end{cases}$$

Die Rezentrierung der Einflussfunktion ermöglicht es, das Gesetz iterierter Erwartungen auf Quantile anzuwenden. Erst dadurch ist es möglich, eine unbedingte Interpretation von Quantilregressionen zu erhalten. Dies lässt sich wie folgt begründen: Der Erwartungswert der nicht-rezentrierten Einflussfunktion ist per Definition Null. Daher ist der Erwartungswert der rezentrierten Einflussfunktion der unbedingte Quantilwert:

$$E\big(RIF(y; Q_\tau)\big) = Q_\tau.$$

Die Anwendung des Gesetzes iterierter Erwartungen auf diesen Ausdruck ergibt:

$$E_X[E(RIF(y; Q_\tau)|X)] = Q_\tau$$

Unbedingte Quantilwerte lassen sich daher als gewichtete, bedingte Erwartungswerte der rezentrierten Einflussfunktion darstellen. Der Einfluss jeder Subgruppe auf das unbedingte Quantil ist die Erwartungswertdifferenz der subgruppenspezifischen Einflussfunktion zur Referenzgruppe. Für den Fall einer Regression auf eine Dummy-Variable ergibt sich:

$$\beta_{RIF} = E(RIF(y; Q_\tau)|X = 1) - E(RIF(y; Q_\tau)|X = 0).$$

Unbedingte Quantilregressionen geben daher Erwartungswertdifferenzen der Quantile wieder, die aus subgruppenspezifischen Wahrscheinlichkeiten resultieren, ober- oder unterhalb des Quantils zu liegen. Für die Verdeutlichung dieses Gedankens und des eigentlichen Regressionsmodells ist zunächst eine Umformung der rezentrierten Einflussfunktion notwendig:

$$\begin{aligned} RIF(y;\ Q_\tau) &= Q_\tau + \frac{\tau - \mathbb{I}(y\ \leq Q_\tau)}{f_Y(Q_\tau)} \\ &= \frac{\mathbb{I}(y > Q_\tau)}{f_Y(Q_\tau)} +\ Q_\tau + \frac{\tau - 1}{f_Y(Q_\tau)} \\ &= \frac{1}{f_Y(Q_\tau)} \cdot \mathbb{I}(y > Q_\tau) +\ Q_\tau - \frac{1}{f_Y(Q_\tau)} \cdot (1 - \tau). \end{aligned}$$

Der Ausdruck $\frac{1}{f_Y(Q_\tau)}$ ist eine Konstante ($c_{1,\tau}$). Dies trifft auch auf $Q_\tau - \frac{1}{f_Y(Q_\tau)} \cdot (1 - \tau)$ zu ($c_{2,\tau}$). Mit der letzten Umformung wird schließlich der bedingte Erwartungswert gebildet: $E(RIF(y; Q_\tau)|X) = E\big(c_{1,\tau} \cdot \mathbb{I}(y > Q_\tau) \mid X\big) + E\big(c_{2,\tau} |X\big) = c_{1,\tau} \cdot P(y > Q_\tau|X) + c_{2,\tau}$.

Wenn sich zwei Gruppen in Bezug auf die Wahrscheinlichkeit $P(y > Q_\tau| X)$ nicht unterscheiden, dann unterscheiden sich die Erwartungswerte ihrer rezentrierten Einflussfunktionen nicht und β_{RIF} ist 0. Unterschiede der Koeffizienten resultieren somit aus Differenzen der bedingten Wahrscheinlichkeit, oberhalb einer Quantilgrenze zu liegen.

Die Schätzung der durch β_{RIF} ausgedrückten Erwartungswertdifferenzen kann schließlich mit OLS erfolgen. Auch wenn die abhängige Variable nur zwei Werte annimmt, weil Beobachtungen nur ober- oder unterhalb der Quantilgrenze liegen, erfolgt die Regression auf die

Erwartungswerte der abhängigen Variablen. Somit sind die RIF-OLS-Koeffizienten unverzerrte Schätzungen der Einflussdifferenzen auf das Quantil.

A4. Dekomposition unbedingter Quantilregressionen

Ausgangspunkt für die Dekompositionsgleichung ist ein Regressionskoeffizient für eine Dummy-Variable: $\beta = E(\,\theta \mid X = \ 1) - E(\,\theta \mid X = 0)$, wobei θ für eine beliebige, metrische abhängige Variable (etwa das Einkommen) steht. Wie wir weiter oben dargelegt haben, ist eine lineare Regression gleichbedeutend mit der Summe bedingter Erwartungswerte. Aufgrund des Gesetzes iterierter Erwartungen lässt sich der unbedingte Erwartungswert in folgender Weise ausdrücken:

$$E(\theta) = E(\beta_0 + \beta_1 x_1 + \epsilon\,) = E(\beta_0) + E(\beta_1 x_1) = \ \beta_0 + \ \beta_1 E(x_1) = \ \beta_0 + \ \beta_1 \overline{x_1}.$$

Unter der Annahme, dass β_1 die Differenz von zwei voneinander unabhängigen Erwartungswerten ausdrückt, können diese mit zwei separaten linearen Regressionen geschätzt werden.

$$\begin{aligned} \beta_1 &= E(\beta_0 + \beta_1 K + \epsilon \mid X = 1) - \ E(\beta_0 + \beta_1 K + \epsilon \mid \ X = 0) \\ &= \ \left(\,\beta_{0|X=1} + \beta_{1|X=1}\,\overline{K}_{X=1}\right) - \ \left(\beta_{0|X=0} + \beta_{1|X=0}\,\overline{K}_{X=0}\right), \end{aligned}$$

wobei K einen Vektor von Kovariaten darstellt. Dieser Ausdruck wird auf der linken Seite durch $\beta_{1|X=1}\,\overline{K}_{X=0} - \ \beta_{1|X=1}\,\overline{K}_{X=0}$ erweitert. Daraufhin ergibt sich die bekannte Oaxaca-Blinder-Dekompositions-Gleichung:

$$\begin{aligned} \beta_1 &= \left(\beta_{0|X=1} + \beta_{1|X=1}\,\overline{K}_{X=1}\right) - \left(\beta_{0|X=0} + \beta_{1|X=0}\,\overline{K}_{X=0}\right) - \left(\beta_{1|X=1}\,\overline{K}_{X=0} - \ \beta_{1|X=1}\,\overline{K}_{X=0}\right) \\ &= \underbrace{\left(\beta_{0|X=0} - \ \beta_{0|X=1}\,\right) + \ \overline{K}_{X=0}\left(\beta_{1|X=0} \ - \ \ \beta_{1|X=1}\right)}_{\text{Struktureffekt}} \ + \underbrace{\beta_{1|X=1}\left(\overline{K}_{X=0} - \ \overline{K}_{X=1}\right)}_{\text{Kompositionseffekt}} \end{aligned}$$

Der erste Teil der Gleichung bezeichnet den (Einkommens-)Struktureffekt. Das ist derjenige Teil der Gleichung, der in unserer Analyse Veränderungen der ökonomischen Situation einer Gruppe über die Zeit ausdrückt. Der Kompositionseffekt drückt Unterschiede aus, die aus Veränderungen der relativen Populationshäufigkeit einer Gruppe resultieren.

Anhang B

RIF-Regressionen auf Einflussfunktionen des 15./50. Perzentils der Haushaltsnettoäquivalenzeinkommen

Periode	**2009/11**	**1992/94**	**2009/11**	**1992/94**
Perzentil	15	15	50	50
Rentner	-0.106***	-0.194***	0.007	-0.021
	(0.026)	(0.025)	(0.020)	(0.024)
Junger Haushalt	-0.340***	-0.076**	-0.076***	0.005
	(0.042)	(0.025)	(0.018)	(0.016)
Kein Verdiener	-1.116***	-0.714***	-0.150***	-0.150***
	(0.056)	(0.062)	(0.022)	(0.032)
Teilverdiener	-0.404***	-0.306***	-0.073***	-0.114***
	(0.030)	(0.030)	(0.017)	(0.020)
Zwei Vollverdiener	-0.054***	0.030*	0.068***	0.088***
	(0.013)	(0.012)	(0.015)	(0.015)
Geringe Berufserfahrung	-0.263***	-0.103***	-0.138***	-0.038*
	(0.029)	(0.025)	(0.019)	(0.018)
Hohe Berufserfahrung	0.023	0.015	0.083***	0.046***
	(0.017)	(0.012)	(0.016)	(0.014)
Geringe Bildung	-0.197***	-0.122***	-0.102***	-0.085***
	(0.031)	(0.024)	(0.015)	(0.016)
Hohe Bildung	0.119***	0.081***	0.206***	0.149***
	(0.015)	(0.018)	(0.013)	(0.016)
Geschlecht	0.003	-0.043*	-0.020	-0.070***
	(0.017)	(0.017)	(0.011)	(0.013)
Transfers	-0.004	-0.041***	0.003	-0.026***
	(0.005)	(0.007)	(0.002)	(0.004)
Steuern/Sozialabgaben	0.014***	0.015***	0.026***	0.033***
	(0.001)	(0.002)	(0.001)	(0.003)
Einpersonenhaushalt	-0.127***	-0.157***	-0.015	-0.008
	(0.020)	(0.022)	(0.013)	(0.015)
Nichtdeutsch	-0.234***	-0.128***	-0.184***	-0.096***
	(0.037)	(0.027)	(0.019)	(0.020)
Geschieden	-0.197***	-0.145***	-0.045**	-0.011
	(0.028)	(0.031)	(0.015)	(0.018)
Paar ohne Kinder	0.013	-0.041**	0.116***	0.123***
	(0.019)	(0.015)	(0.014)	(0.014)
Region	-0.141***	-0.254***	-0.179***	-0.276***
	(0.021)	(0.023)	(0.012)	(0.013)
Konstante	9.515***	9.514***	9.633***	9.641***
	(0.028)	(0.022)	(0.021)	(0.026)
R2	0.281	0.221	0.352	0.375
N	32004	19861	32004	19861

* p<0.05, ** p<0.01, *** p<0.001

Bootstrap-Standardfehler mit Replikationsgewichten, 1000 Wiederholungen