Methodische Aspekte der Analyse von Einkommensdiskriminierung

Ben Jann

ETH Zürich, jannb@ethz.ch

Institut für Arbeitsmarkt- und Berufsforschung
Nürnberg, 22. Juni 2009

Gliederung

- Einleitung
- Die Blinder-Oaxaca-Dekomposition
- Zwei methodische Aspekte
  - Das Index-Nummer-Problem
  - Statistische Inferenz
- Zusammenfassung

Einleitung

- Dass in den meisten westlichen Gesellschaften weiterhin bedeutende Lohnunterschiede zwischen Frauen und Männern bestehen, ist ein bekanntes Phänomen.
- Diese Unterschiede werden üblicherweise zumindest zum Teil auf Diskriminierung zurückgeführt.
- Ein beliebter statistischer Ansatz, um Lohnunterschiede zwischen Frauen und Männern (oder anderen Gruppen) zu analysieren, ist die kontrafaktische Dekomposition nach Blinder (1973) und Oaxaca (1973). Zu einer Meta-Analyse siehe Weichselbaumer und Winter-Ebmer (2005).

Einleitung

- Ausgehend von aus der Humankapitaltheorie abgeleiteten Lohngleichungen (Mincer 1974) versucht die Blinder-Oaxaca-Zerlegung zwei zentrale Fragen zu beantworten:
  - Welcher Teil der Lohndiskrepanz kann durch Gruppenunterschiede bezüglich lohnwirksamer Merkmale erklärt werden?
  - Welcher Teil des Unterschieds ist auf andere Ursachen wie z.B. unterschiedliche Bildungsrenditen oder Lohndiskriminierung zurückzuführen?
- Es handelt sich allerdings um einen indirekten Ansatz, bei dem Diskriminierung lediglich als nicht erklärter Rest übrig bleibt. Das Verfahren kann also höchstens gewisse Indizien liefern.
- Auf allgemeiner Ebene kann der Ansatz zudem aufgrund einiger weiterer Punkte kritisiert werden (z.B. statische Betrachtungsweise, Beschränkung auf Mittelwerte).

Einleitung

- Aber auch wenn wir die grundsätzliche Nützlichkeit des Verfahrens nicht anzweifeln, gilt es verschiedene methodische Probleme zu überwinden.
- Einige dieser Probleme können unter den folgenden Punkten zusammengefasst werden:
  - Das Index-Nummer-Problem
  - Statistische Inferenz
  - Identifikation der Effekte kategorialer Merkmale
  - Korrektur für Selektionseffekte
- Ich werde mich mit den ersten beiden Aspekten befassen (für die anderen beiden Punkte sei auf Oaxaca und Ransom 1999 und Yun 2005 sowie Neuman und Oaxaca 2004 verwiesen).

Die Blinder-Oaxaca-Dekomposition

- Drei-Komponenten-Zerlegung (Winsborough und Dickinson 1971): Ausgehend vom linearen Modell

$$Y_j = X_j'\beta_j + \epsilon_j, \quad E(\epsilon_j) = 0, \quad j \in \{1, 2\}$$

kann der mittlere Gruppenunterschied $R = \bar{Y}_1 - \bar{Y}_2 = \bar{X}_1'\hat{\beta}_1 - \bar{X}_2'\hat{\beta}_2$ zerlegt werden zu

$$R = (\bar{X}_1 - \bar{X}_2)'\hat{\beta}_2 + \bar{X}_2'(\hat{\beta}_1 - \hat{\beta}_2) + (\bar{X}_1 - \bar{X}_2)'(\hat{\beta}_1 - \hat{\beta}_2)$$

Effekt der Ausstattung	Effekt der Koeffizienten	Interaktion

$\bar{Y}$: Mittelwert der Ergebnisvariable (z.B. logarithmierte Löhne)
$\bar{X}$: Mittelwertsvektor der Regressoren (z.B. Bildung, Berufserfahrung, etc.)

Die Blinder-Oaxaca-Dekomposition

- Zwei-Komponenten-Zerlegung:

$$R = \underbrace{(\bar{X}_1 - \bar{X}_2)'\beta^*}_{\text{„erklärter“ Teil } (Q)} + \underbrace{\left[\bar{X}_1'(\hat{\beta}_1 - \beta^*) + \bar{X}_2'(\beta^* - \hat{\beta}_2)\right]}_{\text{„unerklärter“ Teil } (U)}$$

  wobei β^* ein Vektor von Referenzkoeffizienten ist (i.e. der Vektor der „nicht-diskriminierenden“ Koeffizienten).

- Alternative Formulierung:

$$\begin{aligned} R = {} & (\bar{X}_1 - \bar{X}_2)'[W\hat{\beta}_1 + (I - W)\hat{\beta}_2] \quad \text{(erklärter Teil)} \\ & + [\bar{X}_1'(I - W) + \bar{X}_2'W](\hat{\beta}_1 - \hat{\beta}_2) \quad \text{(unerklärter Teil)} \end{aligned}$$

  wobei W eine Matrize mit relativen Gewichten ist (I = Identitätsmatrize).

Aspekt 1: Das Index-Nummer-Problem

- Wie ist β^* bzw. W in der Zwei-Komponenten-Zerlegung zu wählen?
- Einige Vorschläge:
  - $\beta^* = \hat{\beta}_1$ (bzw. $W = I$) oder $\beta^* = \hat{\beta}_2$ (bzw. $W = 0$) (Oaxaca 1973; Blinder 1973)
  - $\beta^* = 0.5\hat{\beta}_1 + 0.5\hat{\beta}_2$ (bzw. $W = 0.5I$) (Reimers 1983)
  - Relative Gruppengrössen als Gewichte (Cotton 1988)
- Weiterer populärer Vorschlag:
  - Verwendung der Koeffizienten eines über beide Gruppen zusammengefassten Modells als Schätzer für β^* (Neumark 1988)
  - äquivalent: $W = (X_1'X_1 + X_2'X_2)^{-1}X_1'X_1$ (Oaxaca und Ransom 1994)
- Der letzte Vorschlag erscheint bestechend, ist aber fragwürdig, da ein Teil der Lohndifferenz in unangemessener Weise dem erklärten Teil zugeschlagen wird.

Aspekt 1: Das Index-Nummer-Problem

- Gegeben sei ein einfaches Modell (z.B. $Y =$ logarithmierter Lohn, $Z =$ Bildung)

$$Y = \alpha + \gamma Z + \delta G + \epsilon$$

  wobei δ ein Diskriminierungsparameter ist ($\delta < 0$) und G ein Indikator für das Geschlecht (1 falls weiblich).

- Wird nun γ^* aus einem „gepoolten“ Modell

$$Y = \alpha^* + \gamma^* Z + \epsilon^*$$

  in die Dekomposition eingesetzt, erhalten wir für den erklärten Teil

$$Q = (\bar{Z}_M - \bar{Z}_F)\gamma^* = (\bar{Z}_M - \bar{Z}_F)\left(\gamma + \delta \frac{\mathrm{Cov}(Z, G)}{\mathrm{Var}(Z)}\right)$$

  (Standardresultat aus der Theorie der weggelassenen Variablen).

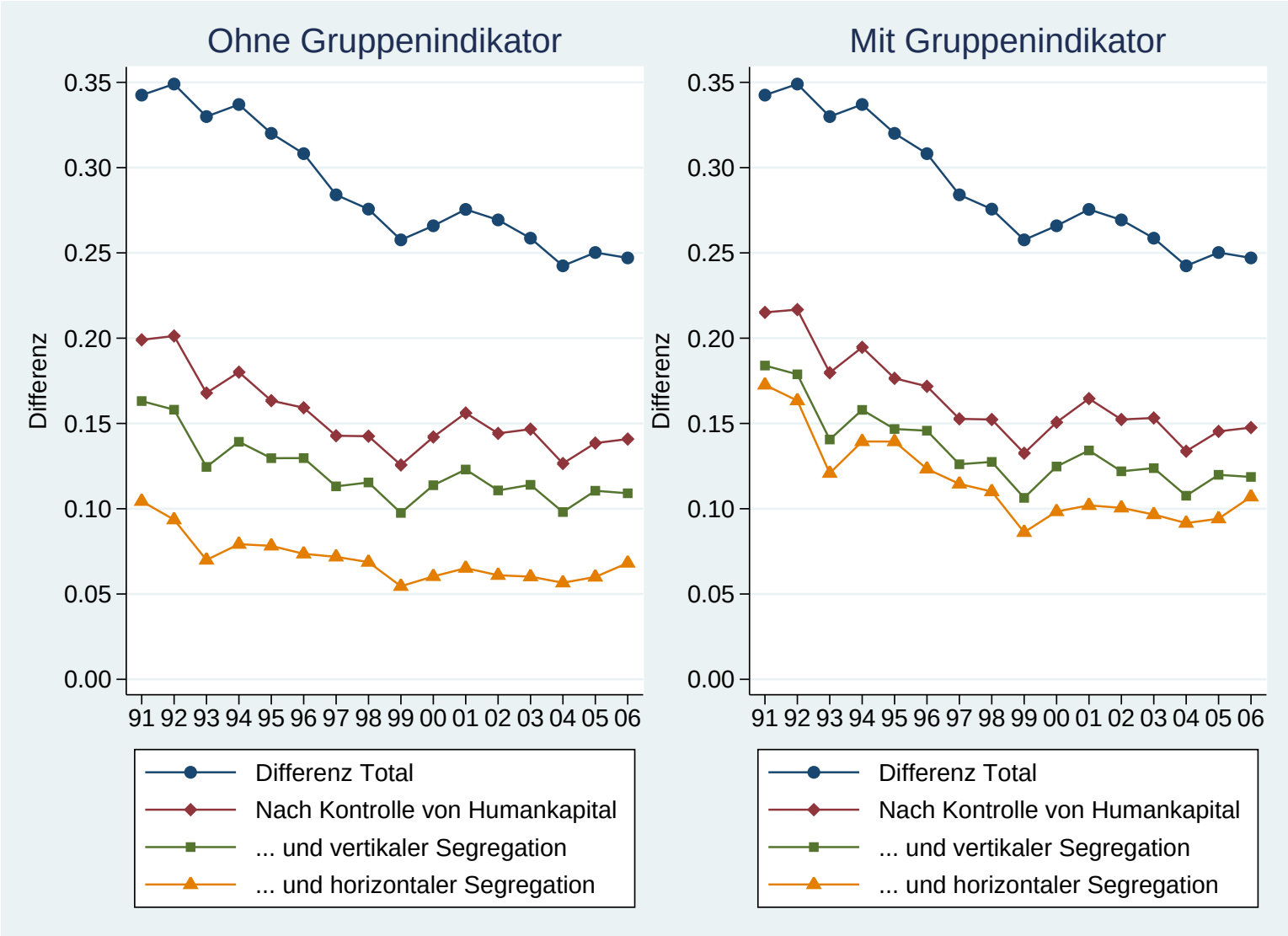
Aspekt 1: Das Index-Nummer-Problem

$$Q = (\bar{Z}_M - \bar{Z}_F)\gamma^* = (\bar{Z}_M - \bar{Z}_F)\left(\gamma + \delta\frac{\mathrm{Cov}(Z, G)}{\mathrm{Var}(Z)}\right)$$

- Sind nun Männer im Schnitte besser ausgebildet als Frauen, ist $\mathrm{Cov}(Z, G)$ negativ und Q wird überschätzt (gegeben $\gamma > 0$ und $\delta < 0$).
- Faktisch heisst das, dass ein Teil des Lohnunterschiedes zwischen Männern und Frauen durch das Geschlecht erklärt wird.
- Um das angesprochene Problem zu vermeiden, sollte im gepoolten Modell also immer auch ein Indikator für die Gruppenzugehörigkeit enthalten sein, was aber in der bisherigen Literatur meistens übersehen wurde.
- Und es macht wirklich einen Unterschied ...

Das Index-Nummer-Problem: Beispiel

- Lohnunterschiede zwischen Frauen und Männern in der Schweiz
  (Schweizerische Arbeitskräfteerhebung 1991-2006, Bundesamt für Statistik)



Aspekt 2: Statistische Inferenz

- Die Berechnung der Komponenten der Blinder-Oaxaca-Dekomposition ist trivial: Koeffizienten gruppenspezifischer OLS-Modelle und Mittelwerte der Regressoren schätzen und in die Formeln einsetzen.
- Die Bezifferung der statistischen Unsicherheit scheint etwas mehr Probleme zu bereiten. Zumindest werden in den meisten Anwendungen keine Standardfehler oder Konfidenzintervalle berichtet.
- Eine adäquate Interpretation der Resultate ist aber leider ohne ungefähre Angaben zur statistischen Präzision nur schlecht möglich.

Aspekt 2: Statistische Inferenz

- Ein erster Vorschlag zur Berechnung der Standardfehler wurde von Oaxaca und Ransom (1998) gemacht (vgl. auch Greene 2003:53–54).
- Oaxaca und Ransom gehen (implizit) von fixen Regressoren aus und vernachlässigen somit eine wichtige Quelle statistischer Unsicherheit.
- Dass die stochastische Natur der Regressoren für die Varianzen in Regressionsmodellen vernachlässigt werden kann, ist ein zentrales Resultat der Ökonometrie. Dies lässt sich jedoch nicht auf die Blinder-Oaxaca-Dekomposition übertragen.

Aspekt 2: Statistische Inferenz

- Wie lässt sich die Varianz eines Ausdrucks $\bar{X}'\hat{\beta}$ schätzen?
  - Wenn die Regressoren fix sind, dann ist $\bar{X}$ konstant. Somit:

$$\widehat{V}(\bar{X}'\hat{\beta}) = \bar{X}'\widehat{V}(\hat{\beta})\bar{X}$$

  - Wenn die Regressoren jedoch stochastisch sind, erhalten wir

$$\widehat{V}(\bar{X}'\hat{\beta}) = \bar{X}'\widehat{V}(\hat{\beta})\bar{X} + \hat{\beta}'\widehat{V}(\bar{X})\hat{\beta} + \mathrm{tr}\Big(\widehat{V}(\bar{X})\widehat{V}(\hat{\beta})\Big)$$

    (Beweis im Anhang; aus den Standardannahmen der Regression folgt, dass $\bar{X}$ and $\hat{\beta}$ unkorreliert sind).
  - Der letzte Term, tr(...), ist asymptotisch vernachlässigbar.

Aspekt 2: Statistische Inferenz

- Dieses Ergebnis lässt sich nun unmittelbar übertragen auf die Komponenten der Dekomposition. Beispielsweise (angenommen die Gruppen sind unabhängig):

$$\widehat{V}([\bar{X}_1 - \bar{X}_2]'\hat{\beta}_2) \approx (\bar{X}_1 - \bar{X}_2)'\widehat{V}(\hat{\beta}_2)(\bar{X}_1 - \bar{X}_2) + \hat{\beta}_2' \left[\widehat{V}(\bar{X}_1) + \widehat{V}(\bar{X}_2)\right] \hat{\beta}_2$$

$$\widehat{V}(\bar{X}_2'[\hat{\beta}_1 - \hat{\beta}_2]) \approx \bar{X}_2' \left[\widehat{V}(\hat{\beta}_1) + \widehat{V}(\hat{\beta}_2)\right] \bar{X}_2 + (\hat{\beta}_2 - \hat{\beta}_2)'\widehat{V}(\bar{X}_2)(\hat{\beta}_2 - \hat{\beta}_2)$$

- Ähnliche Formeln lassen sich für die anderen Varianten der Dekomposition herleiten (am einfachsten mit Hilfe der Delta-Methode). Auch die Verallgemeinerung auf komplexe Surveydaten ist einfach möglich.

Aspekt 2: Statistische Inferenz

- Monte-Carlo-Simulation: Relative Abweichungen der Varianzschätzer

	$n = 1000$		$n = 5000$	
	fix	stochastisch	fix	stochastisch
Mittelwert Gruppe A	−0.505	−0.007	−0.505	−0.009
Mittelwert Gruppe B	−0.392	0.015	−0.413	−0.023
Differenz	−0.421	−0.005	−0.430	−0.025
$W = 0$:				
– erklärt	−0.790	0.016	−0.802	−0.023
– unerklärt	−0.057	0.001	−0.048	0.008
$W = I$:				
– erklärt	−0.888	0.007	−0.892	−0.010
– unerklärt	−0.140	−0.009	−0.132	−0.007

Aspekt 2: Statistische Inferenz

- Anwendungsbeispiel: Lohnunterschiede zwischen Frauen und Männern in der Schweiz (Schweizerische Arbeitskräfteerhebung 2000)

	$\hat{\theta}$	Standardfehler			
		fix	stochastisch	Bootstrap	Jackknife
Mittelwert Männer	3.808	0.00605	0.00732	0.00729	0.00731
Mittelwert Frauen	3.568	0.0100	0.0115	0.0116	0.0115
Differenz	0.241	0.0117	0.0136	0.0137	0.0136
$W = 0$:					
– erklärt	0.0976	0.00705	0.00987	0.0101	0.0101
– unerklärt	0.143	0.0136	0.0136	0.0135	0.0137
$W = I$:					
– erklärt	0.113	0.00418	0.00820	0.00821	0.00815
– unerklärt	0.128	0.0125	0.0126	0.0125	0.0126

Zusammenfassung

- Es wurde gezeigt, dass ...
  - ... bei der Schätzung der Referenzkoeffizienten der Dekomposition anhand eines „gepoolten“ Modells für die Gruppenzugehörigkeit kontrolliert werden sollte,
  - ... Formeln für die Standardfehler einfach herzuleiten sind
  - ... und der stochastischen Natur der Regressoren bei der Schätzung der Standardfehler Rechnung getragen werden sollte.
- Benutzerfreundliche Software für den Einsatz des Blinder-Oaxaca-Verfahrens in der angewandte Forschung wurde verfügbar gemacht.
  - Jann, Ben (2008). The Blinder-Oaxaca decomposition for linear regression models. The Stata Journal 8: 453-479.

Vielen Dank für Ihre Aufmerksamkeit!

Beweis I

LEMMA: Die Varianz des Produkts von zwei unkorrelierten Zufallsvektoren ist

$$V(u_1'u_2) = \mu_1'\Sigma_2\mu_1 + \mu_2'\Sigma_1\mu_2 + \text{tr}(\Sigma_1\Sigma_2)$$

wobei $u_j \sim (\mu_j, \Sigma_j)$, $j = 1, 2$

BEWEIS: Es gilt

$$E(x+y) = E(x) + E(y), \quad E(xy) = E(x)E(y) + Cov(x, y)$$

Wenn u_1 und u_2 unkorreliert sind, folgt

$$E(u_1'u_2) = \mu_1'\mu_2, \quad E(u_j u_j') = \mu_j\mu_j' + \Sigma_j$$

Beweis II

und

$$\begin{aligned} E([u_1' u_2]^2) &= E(u_1' u_2 u_2' u_1) = \mathrm{tr}(E(u_1 u_1' u_2 u_2')) \\ &= \mathrm{tr}(E(u_1 u_1') E(u_2 u_2')) \\ &= \mathrm{tr}((\mu_1 \mu_1' + \Sigma_1)(\mu_2 \mu_2' + \Sigma_2)) \\ &= \mathrm{tr}(\mu_1 \mu_1' \mu_2 \mu_2') + \mathrm{tr}(\mu_1 \mu_1' \Sigma_2) \\ &\quad + \mathrm{tr}(\Sigma_1 \mu_2 \mu_2') + \mathrm{tr}(\Sigma_1 \Sigma_2) \\ &= (\mu_1' \mu_2)^2 + \mu_1' \Sigma_2 \mu_1 + \mu_2' \Sigma_1 \mu_2 + \mathrm{tr}(\Sigma_1 \Sigma_2) \end{aligned}$$

Schliesslich:

$$\begin{aligned} V(u_1' u_2) &= E([u_1' u_2]^2) - [E(u_1' u_2)]^2 \\ &= \mu_1' \Sigma_2 \mu_1 + \mu_2' \Sigma_1 \mu_2 + \mathrm{tr}(\Sigma_1 \Sigma_2) \end{aligned}$$

Populationswerte der Simulation

		Gruppe A	Gruppe B
X_1	Mittelwert	1.0	0.5
	Varianz	1.0	2.0
X_2	Mittelwert	1.0	0.7
	Varianz	1.0	1.5
Kovarianz (X_1, X_2)		-0.5	-0.3
Populationsanteil		0.6	0.4

	Gruppe A	Gruppe B
β_0	1.0	0.5
β_1	1.0	0.5
β_2	1.0	1.3
σ	1.0	2.0

Regressionsmodelle SAKE 2000

	Männer		Frauen	
	Koeffizient	Mittelwert	Koeffizient	Mittelwert
Bildungsjahre	0.0770	12.17	0.0768	11.73
	(0.0028)	(0.042)	(0.0051)	(0.056)
Berufserfahrung (in Jahren)	0.0216	19.59	0.0289	14.35
	(0.0018)	(0.23)	(0.0034)	(0.29)
Berufserfahrung^2/100	−0.0305	5.279	−0.0540	3.154
	(0.0040)	(0.10)	(0.0084)	(0.12)
Firmentreue (in Jahren)	0.00260	10.62	0.00650	7.651
	(0.00077)	(0.18)	(0.0015)	(0.22)
Vorgesetztenfunktion	0.139	0.558	0.0482	0.386
	(0.012)	(0.0093)	(0.021)	(0.014)
Konstante	2.504		2.353	
	(0.039)		(0.067)	
R-Quadrat	0.316		0.238	
Fallzahl	2825		1287	

Quelle: Schweizerische Arbeitskräfteerhebung (SAKE) 2000, ungewichtet.
Anmerkungen: Abhängige Variable ist der logarithmierte Bruttostundenlohn; Standardfehler in Klammern; Auswahl: Vollzeitarbeitnehmer (mit nur einer Stelle) im Alter von 20 bis 62 Jahren, ohne Ausländer.

(Jann 2008a:161)

Literaturhinweise I

- Blinder, Alan S. (1973). Wage Discrimination: Reduced Form and Structural Estimates. The Journal of Human Resources 8: 436-455.
- Cotton, Jeremiah (1988). On the Decomposition of Wage Differentials. The Review of Economics and Statistics 70: 236-243.
- Greene, William H. (2003). Econometric Analysis. 5. Upper Saddle River, NJ: Pearson Education.
- Jann, Ben (2008a). Erwerbsarbeit, Einkommen und Geschlecht. Studien zum Schweizer Arbeitsmarkt. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Jann, Ben (2008b). The Blinder-Oaxaca decomposition for linear regression models. The Stata Journal 8: 453-479.
- Mincer, Jacob (1974). Schooling, Experience and Earnings. New York and London: Columbia University Press.

Literaturhinweise II

- Neuman, Shoshana, Ronald L. Oaxaca (2004). Wage decompositions with selectivity-corrected wage equations: A methodological note. Journal of Economic Inequality 2: 3-10.
- Neumark, David (1988). Employers' Discriminatory Behavior and the Estimation of Wage Discrimination. The Journal of Human Resources 23: 279-295.
- Oaxaca, Ronald (1973). Male-Female Wage Differentials in Urban Labor Markets. International Economic Review 14: 693-709.
- Oaxaca, Ronald L., Michael R. Ransom (1994). On discrimination and the decomposition of wage differentials. Journal of Econometrics 61: 5-21.
- Oaxaca, Ronald L., Michael Ransom (1998). Calculation of approximate variances for wage decomposition differentials. Journal of Economic and Social Measurement 24: 55-61.
- Oaxaca, Ronald L., Michael R. Ransom (1999). Identification in Detailed Wage Decompositions. The Review of Economics and Statistics 81: 154-157.

Literaturhinweise III

- Reimers, Cordelia W. (1983). Labor Market Discrimination Against Hispanic and Black Men. The Review of Economics and Statistics 65: 570-579.
- Yun, Myeong-Su (2005). A Simple Solution to the Identification Problem in Detailed Wage Decompositions. Economic Inquiry 43: 766-772.
- Weichselbaumer, Doris, Rudolf Winter-Ebmer (2005). A Meta-Analysis of the International Gender Wage Gap. Journal of Economic Surveys 19: 479-511.