



Wer "verdient" was warum? Das Oaxaca/Blinder-Dekompositions-Verfahren zur Analyse des Gender Pay Gap

Christian Pierdzioch

Georg Stadtmann

European University Viadrina Frankfurt (Oder)
Department of Business Administration and Economics

Discussion Paper No. 408

June 2019

ISSN 1860 0921

Wer "verdient" was warum? Das Oaxaca/Blinder-Dekompositions-Verfahren zur Analyse des Gender Pay Gap

Christian Pierdzioch^a

Georg STADTMANN^b

Version: Juni 2019

Zusammenfassung

In letzter Zeit stand häufig der sogenannte Gender Pay Gap in der Diskussion. Der Gender Pay Gap besagt, dass der durchschnittliche Lohnsatz von Frauen unter dem durchschnittlichen Lohnsatz von Männern liegt. Doch was sind genau die Hintergründe und Ursachen dieses Phänomens? In diesem Beitrag werden einige Faktoren herausgearbeitet. Dazu wurde ein stilisierter Datensatz generiert, der im Anschluss durch ökonometrische Verfahren ausgewertet wird. Ein Ansatz, welcher in der Literatur zum Gender Pay Gap häufig Anwendung gefunden hat, ist das Oaxaca/Blinder-Dekompositions-Verfahren (Busch 2013, S. 332). Die Idee hinter diesem Verfahren wird in diesem Aufsatz ebenfalls skizziert.

Keywords: Gender Pay Gap, Diskriminierung, Dekomposition

Kontakt:

^a Helmut Schmidt Universität, Hamburg

^b Corresponding author: Georg STADTMANN, Europa-Universität Viadrina, Department of Economics, Grosse Scharrnstrasse 59, 15230 Frankfurt (Oder), Germany, stadtman@europa-uni.de

Inhaltsverzeichnis

1	Einleitung	3
2	Deskriptive Statistik	3
3	Ökonometrische Analyse mittels multipler Regression	10
3.1	Beschreibung des Testansatzes	10
3.2	Exkurs: Hypothesenpaar und einseitiger versus zweiseitiger t - Test	12
3.3	Regressionsergebnisse	13
3.4	Exkurs: Gehaltsprognosen auf Basis der Regressionsergebnisse	14
3.5	Exkurs: Zusammenhang zwischen Gehalt und Alter	16
4	Oaxaca/Blinder-Dekomposition	17
5	Zusammenfassung und Fazit	22

Abbildungsverzeichnis

1	Höchster Bildungsabschluss: Männer versus Frauen	7
2	Gehalt in Abhängigkeit des höchsten Bildungsabschlusses	8
3	Gehalt in Abhängigkeit der Branchenzugehörigkeit	9
4	Gehalt in Abhängigkeit des Alters	17

Tabellenverzeichnis

1	Deskriptive Statistik: Gesamter Datensatz	5
2	Deskriptive Statistik: Männer versus Frauen	6
3	Regressionsergebnisse	13
4	Regressionsergebnisse, Charakteristika & Dekomposition	21

1 Einleitung

Gender Pay Gap

In letzter Zeit stand häufig der sogenannte Gender Pay Gap in der Diskussion. Der Gender Pay Gap besagt, dass der durchschnittliche Lohnsatz von Frauen unter dem durchschnittlichen Lohnsatz von Männern liegt. Doch was sind genau die Hintergründe und Ursachen dieses Phänomens? In diesem Beitrag werden einige Faktoren herausgearbeitet. Dazu wurde ein stilisierter Datensatz generiert, der im Anschluss durch ökonometrische Verfahren ausgewertet wird. Ein Ansatz, welcher in der Literatur zum Gender Pay Gap häufig Anwendung gefunden hat, ist das Oaxaca/Blinder-Dekompositions-Verfahren (Busch 2013, S. 332). Die Idee hinter diesem Verfahren wird in diesem Aufsatz ebenfalls skizziert.

Die Analysen können in Excel oder einer speziellen Ökonometrie-Software (STATA oder R) durchgeführt werden. Wir gehen auch auf die notwendigen Befehle ein, mit denen die einzelnen Analyseschritte durchgeführt werden können.

Die Excel Datei ist unter

www.wiwi.europa-uni.de/de/forschung/publikationen-projekte/dp/2019.html

verfügbar.

2 Deskriptive Statistik

Beschreibung der Daten

Die Daten sind in einem Excel-File verfügbar: In jeder Zeile ist eine einzige Person und deren Charakteristika gespeichert. Es existieren Informationen über das Geschlecht, das Alter, den höchsten Bildungsabschluss und die Branche, in der ein Beruf ausgeübt wird.

- Das Geschlecht ist über eine Dummy-Variable kodiert: Nimmt die Variable `gend` den Wert 1 an, so handelt es sich um eine weibliche Person, eine männliche Person wird durch eine 0 kodiert.
- Das Alter (`alte`) ist in Jahren angegeben.
- Bezüglich der Bildung werden drei unterschiedliche Kategorien definiert: Der höchste Bildungsabschluss könnte
 - eine abgeschlossene Schulausbildung,
 - eine abgeschlossene Berufsausbildung oder
 - ein abgeschlossenes Hochschulstudium sein.

Die zugrundeliegende Hypothese würde lauten, dass das Gehalt mit zunehmendem Bildungsgrad ansteigt.

- Um die Analyse einfach zu halten, werden bezüglich der Branchenzugehörigkeit nur zwei Branchen unterschieden: Eine Person kann in der Branche der Dachdeckenden arbeiten (`dach`) oder einem Beruf in der Pflege nachgehen (`pfle`).

Im Rahmen einer empirischen Analyse ist es immer sinnvoll, zunächst eine deskriptive Statistik der Daten vorzunehmen und z. B. Mittelwerte, Standardabweichungen, Minima und Maxima zu berechnen. Zum einen dient dies dazu, einen ersten Überblick über die Daten zu gewinnen. Zum anderen soll überprüft werden, ob die Zahlen richtig eingelesen wurden und in einem Wertebereich liegen, der sinnvoll ist. So würde es z. B. auffallen, falls das Minimum des Gehalts einen negativen Wert annehmen würde oder das maximale Alter einer erwerbstätigen Person größer als 100 Jahre ist. Gegebenenfalls müssten die Daten bereinigt werden.

Deskriptive Analyse sorgt für Überblick

Eine deskriptive Analyse kann selbstverständlich in Excel durchgeführt werden. Komfortabler geht es in einem Datenverarbeitungsprogramm: Der zugehörige Befehl in STATA heißt `summarise` oder kurz `sum`. Es erscheint

die folgende Tab. 1.

Tabelle 1: Deskriptive Statistik: Gesamter Datensatz

	Mittelwert	Standardabweichung	Minimum	Maximum
Geschlecht (gend)	0,5	0,501	0	1
Alter (alte)	40,6	12,0	20	64
Schule (schu)	0,375	0,485	0	1
Ausbildung (ausb)	0,275	0,448	0	1
Studium (stud)	0,350	0,478	0	1
Dachdeckende (dach)	0,525	0,501	0	1
Pflege (pfle)	0,475	0,501	0	1
Gehalt (lohn)	1196,04	284,58	601,2	1818,8

- Der Datensatz enthält 200 Personen, von denen 100 männlich und 100 weiblich sind.
- Das Durchschnittsalter liegt bei ca. 41 Jahren.
- 37,5 % haben als höchsten Bildungsabschluss den einfachen Schulabschluss vorzuweisen, 27,5 % eine Berufsausbildung und 35,0 % haben ein Studium erfolgreich abgeschlossen.
- 52,5 % der Personen arbeiten als Dachdeckende und 47,5 % in der Pflege.
- Das durchschnittliche Gehalt liegt bei 1.196 EUR.

Frage 1: Warum ist es sinnvoll, eine deskriptive Statistik der Daten vorzunehmen?

Die bisherigen Ausführungen gelten jedoch nur für den Durchschnitt der gesamten Stichprobe. Möchte man nun den Datensatz für Männer und Frauen

getrennt auswerten und eine separate deskriptive Statistik für beide Gruppen erstellen, so kann diese Analyse in STATA wie folgt durchgeführt werden: Zunächst ist über den Befehl `sort gend` der Datensatz bezüglich des Geschlechts zu sortieren. Anschließend kann über den Befehl `by gend: sum` eine getrennte 'Summary Statistic' für beide Geschlechter erzeugt werden.

Tabelle 2: Deskriptive Statistik: Männer versus Frauen

	Mittelwert Männer	Mittelwert Frauen
Alter (alte)	40,4	40,9
Schule (schu)	25 %	50 %
Ausbildung (ausb)	35 %	20 %
Studium (stud)	40 %	30 %
Dachdeckende (dach)	80 %	25 %
Pflege (pfle)	20 %	75 %
Gehalt (lohn)	1.376,73	1.015,35

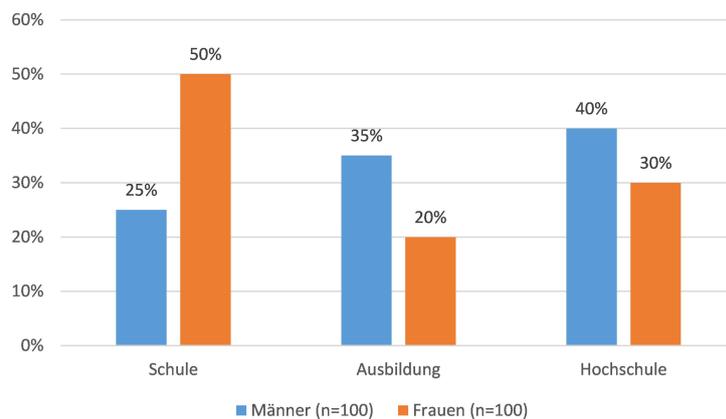
Ein Vergleich der beiden Gruppen (Tab. 2) fördert folgende Unterschiede zu Tage:

- Die Männer sind mit 40,4 Jahren etwas jünger als die Frauen (40,9 Jahre).
- Während 25 % der Männer als höchsten Bildungsabschluss eine normale Schulausbildung aufweisen, liegt dieser Wert bei den Frauen bei 50 %. Der Anteil der Männer mit Ausbildung oder Studium liegt deutlich höher als bei den Frauen. Somit weisen die Männer im Durchschnitt einen höheren Bildungsabschluss auf. Um diese Bildungsunterschiede zu verdeutlichen, werden diese Zusammenhänge in Abb. 1 noch einmal grafisch aufbereitet.
- Auch in Bezug auf die Branche gibt es deutliche Unterschiede: Frauen arbeiten zu 25 % als Dachdeckende und zu 75 % in der Pflege. Bei den

Männern ist es genau umgekehrt: 80 % arbeiten in der Dachdeckendenbranche und nur 20 % in der Pflegebranche.

Unterschiede im höchsten Bildungsabschluss

Abbildung 1: Höchster Bildungsabschluss: Männer versus Frauen



In einem nächsten Schritt könnte es interessant sein, zu überprüfen, ob mit einem höheren Schulabschluss auch ein höheres Gehalt erzielt wird. Es wird eine Variable Bildung (`bild`) erzeugt,

- die den Wert 0 annimmt, falls der höchste Bildungsabschluss eine Schulausbildung ist,
- den Wert 1 annimmt, falls der höchste Bildungsabschluss eine Ausbildung ist und
- den Wert 2 annimmt, falls der höchste Bildungsabschluss ein Hochschulstudium darstellt.

In STATA kann dies wie folgt programmiert werden:

```
gen bild=0
replace bild=1 if ausb==1
replace bild=2 if stud==1
```

Anschließend wird der Datensatz nach dieser Variablen sortiert (`sort bild`) und eine gesonderte deskriptive Statistik erzeugt (`by bild: sum`). Die Ergebnisse dieses Untersuchungsschritts sind in Abb. 2 grafisch aufbereitet.

Abbildung 2: Gehalt in Abhängigkeit des höchsten Bildungsabschlusses

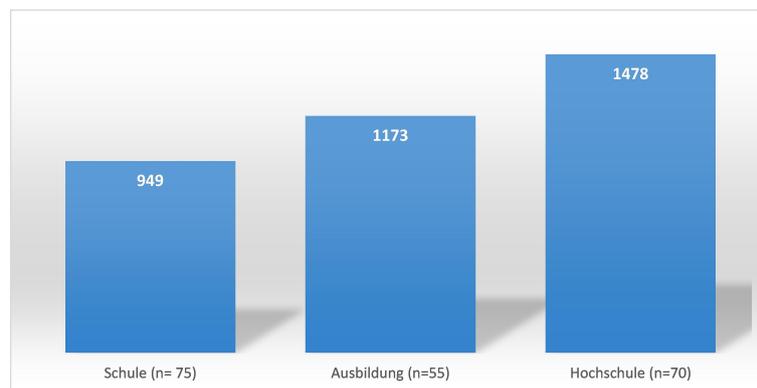


Abb. 2 zeigt deutlich, dass mit einem qualifizierteren Bildungsabschluss auch das Gehaltsniveau ansteigt. In der empirischen Untersuchung von Busch (2013, S. 321) wurde z. B. festgestellt, dass Personen, die keine Ausbildung haben im Vergleich zu Personen mit Berufsausbildung einen Gehaltsabschlag von 3,5 bis 5,5 % aufweisen. Personen mit einem Hochschulabschluss verdienen im Vergleich zu Personen, die eine Berufsausbildung haben, ca. 5 % mehr.

Stellt man nun einen Zusammenhang zwischen der Abb. 1 und Abb. 2 her, so kann man eine Ursache herausfiltern, warum Frauen ein deutlich geringeres Gehalt aufweisen: Frauen weisen ein geringeres Bildungsniveau auf als Männer. Erfolgt somit gar keine Diskriminierung am Arbeits-

markt, sondern ist die Gehaltsdifferenz auf Unterschiede in den Charakteristika zurückzuführen?

Unterscheidung in Männer- bzw. Frauenbranchen

In der Gender Pay Gap Literatur ist es üblich, Branchen in Männer- bzw. Frauenbranchen zu unterteilen. Oftmals wird ein Kriterium von 70 zu 30 eingesetzt: Arbeiten mehr als 70 % Männer in einer Branche, so wird diese Branche als Männerbranche bezeichnet, analog dazu definiert sich eine Frauenbranche. Branchen mit weniger extremen Verteilungen werden als Mischbranchen bezeichnet (Busch 2013, S. 310).

Abbildung 3: Gehalt in Abhängigkeit der Branchenzugehörigkeit

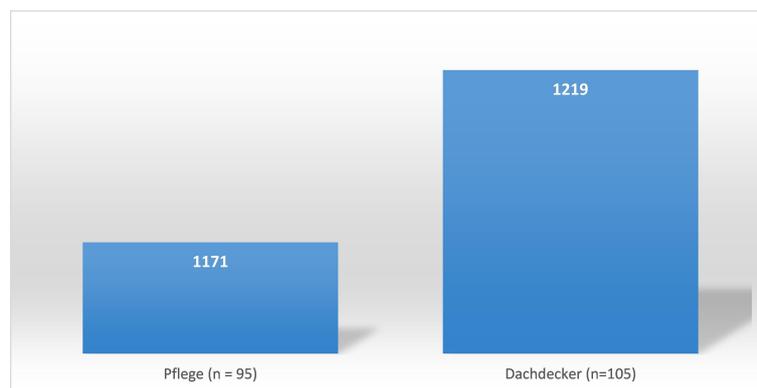


Abb. 3 zeigt deutlich, dass in der Frauenbranche (Pflege) ein geringeres Durchschnittsgehalt gezahlt wird als in der Männerbranche (Dachdeckende). Somit könnte ein Faktor für den Gender Pay Gap darin bestehen, dass Frauen in Branchen arbeiten, in denen eine geringere Entlohnung vorherrscht. Es könnte also sein, dass Frauen am Arbeitsmarkt nicht anders behandelt werden als Männer. Vielmehr könnte der Gehaltsunterschied auf die unterschiedliche Wahl der Branche zurückzuführen sein, in der gearbeitet wird.

Branchenzugehörigkeit wichtiger Erklärungsfaktor für Gehaltsunterschiede

Die deskriptive Statistik hat gezeigt, dass Frauen (1.015 EUR) in der Tat ein geringeres Durchschnittsgehalt aufweisen als Männer (1.377 EUR). Frauen verdienen somit 362 EUR weniger als Männer ($-26,3\%$). Aus der deskriptiven Analyse ist jedoch auch klar geworden, dass sich Männer und Frauen in weiteren Charakteristika unterscheiden: Frauen weisen einen geringeren Bildungsstand auf, da sie – im Durchschnitt – einen geringeren Bildungsabschluss besitzen. Ferner arbeiten Frauen oftmals in Branchen, die ein geringeres Gehaltsniveau aufweisen.

Frage 2: Was können Ursachen für Lohnunterschiede zwischen Männern und Frauen sein?

3 Ökonometrische Analyse mittels multipler Regression

3.1 Beschreibung des Testansatzes

Um die Frage zu klären, ob Frauen am Arbeitsmarkt diskriminiert werden, können verschiedene Analysen eingesetzt werden. Die einfachste Analyse besteht darin, eine Regression durchzuführen und alle vorhandenen Daten (Männer & Frauen) einzubeziehen. Als erklärende Variable werden das Alter und das Alter zum Quadrat, Dummy-Variablen für den höchsten Bildungsabschluss und eine Dummy-Variable für die Branche eingeführt. Die Variable Pflege nimmt den Wert `pfl` = 1 an, wenn in der Pflegebranche gearbeitet wird und ist ansonsten gleich Null. Die Variable Alter zum Quadrat wird eingesetzt, da in empirischen Untersuchungen häufig gezeigt wurde, dass das Gehalt nicht linear mit dem Alter ansteigt, sondern ein nicht-linearer, konkaver Zusammenhang besteht. Durch den Einbezug einer zusätzlichen Dummy-Variable für das Geschlecht (`gend`: Mann = 0, Frau = 1) kann überprüft werden, ob Frauen (signifikant) weniger verdienen als Männer. Somit wird

folgende Regressionsgleichung geschätzt:

$$\text{lohn} = \beta_0 + \beta_1 \cdot \text{alte} + \beta_2 \cdot \text{alte}^2 + \beta_3 \cdot \text{ausb} + \beta_4 \cdot \text{hoch} + \beta_5 \cdot \text{pfle} + \beta_6 \cdot \text{gend} + \epsilon \quad (1)$$

Das Modell wurde somit so spezifiziert, dass die Konstante (β_0)

- für Männer
- mit höchstem Bildungsabschluss Schulausbildung und
- einer Beschäftigung in der Dachdeckendenbranche gilt.

Spezifizierung der erwarteten Vorzeichen

In Bezug auf die Alters-Koeffizienten wird erwartet, dass das Gehalt mit zunehmenden Alter ansteigt ($\beta_1 > 0$), dieser Anstieg mit zunehmenden Alter jedoch degressiv ausfällt ($\beta_2 < 0$). Die Koeffizienten, welche mit dem höchsten Bildungsabschluss in Zusammenhang stehen, sollten beide positiv sein ($\beta_3 > 0$ und $\beta_4 > 0$). Personen, die als höchsten Bildungsabschluss eine Ausbildung oder ein Hochschulstudium vorweisen können, sollten gegenüber Personen, die einen Schulabschluss vorweisen können, ein höheres Durchschnittsgehalt aufweisen. Diese Hypothese könnte man noch weiter verfeinern und überprüfen, ob ein Hochschulabschluss auch höher entlohnt wird als eine Berufsausbildung ($\beta_4 > \beta_3 > 0$). Arbeitet eine Person in der Pflege anstatt in der Dachdeckendenbranche, so sollte diese Person ebenfalls ein geringeres Gehalt aufweisen ($\beta_5 < 0$). Für den Gender-Koeffizienten würden wir auf Grund der Voruntersuchungen erwarten, dass dieser negativ ist ($\beta_6 < 0$). Die für den Untersuchungsgegenstand jedoch entscheidende Frage besteht darin, ob dieser Koeffizient auch signifikant von Null verschieden ist. Ist dies der Fall, würden am Arbeitsmarkt Frauen – nur auf Grund ihres Geschlechts, jedoch nicht auf Grund ihrer Berufsausbildung oder Branchenzugehörigkeit – ein geringeres Gehalt aufweisen als Männer. Man hätte dann eine Form der Diskriminierung der Frauen nachgewiesen.

3.2 Exkurs: Hypothesenpaar und einseitiger versus zweiseitiger t -Test

Test besteht aus einem Hypothesenpaar

Genau genommen besteht ein Hypothesentest stets aus einem Hypothesenpaar: Der Nullhypothese und der Alternativhypothese. In der Regel wird in der wirtschaftswissenschaftlichen Forschung ein zweiseitiger Test durchgeführt. Somit werden die beiden Hypothesen für den Koeffizienten (β_1), welcher mit der Variable "Alter" in Beziehung steht, wie folgt spezifiziert:

Die Nullhypothese lautet: $H_0 : \beta_1 = 0$

Die Alternativhypothese lautet: $H_a : \beta_1 \neq 0$

Zweiseitiger Test ist Standard

Somit folgt: Obwohl wir eigentlich davon ausgehen, dass der Koeffizient β_1 ein positives Vorzeichen aufweisen sollte, wird die Alternativhypothese zweiseitig ($\beta_1 \neq 0$) und nicht einseitig ($\beta_1 > 0$) formuliert. Dies mag ein wenig inkonsistent anmuten, ist jedoch innerhalb der wirtschaftswissenschaftlichen Forschung Standard. Diese Vorgehensweise lässt sich wie folgt begründen:

- Zum einen ist ein zweiseitiger Test "konservativer". Dies sei anhand der kritischen t -Werte eines 95 % Konfidenzniveaus exemplarisch erläutert: Formuliert man die Hypothese zweiseitig, so wird die Nullhypothese erst dann verworfen, wenn der absolute Wert der geschätzten t -Statistik den kritischen Wert von $t_{krit} = 1,96$ übersteigt. Würde man "nur" einseitig testen, so müsste lediglich der Wert von $t_{krit} = 1,64$ überschritten werden.
- Zum anderen ist es oftmals so, dass man grundsätzlich auch eine alternative Theorie formulieren könnte, warum ältere Arbeitnehmer weniger verdienen als Jüngere. Man könnte z. B. argumentieren, dass mit zunehmendem Alter die körperliche und geistige Leistungsfähigkeit abnimmt, was zu einer geringeren Entlohnung führt. Da man oftmals keinen eindeutigen theoretischen Zusammenhang bestimmen kann, for-

muliert man die Alternativhypothese i. d. R. zweiseitig!

3.3 Regressionsergebnisse

In Tabellen mit empirischen Ergebnissen werden oftmals durch Sternchen angezeigt, ob ein Koeffizient auf einem 99 % (***), 95 % (**) oder 90 % (*) Konfidenzniveau von Null verschieden ist. Die kritischen t-werte lauten: 99 % (2,58), 95 % (1,96) oder 90 % (1,64). In der Tab. 3 befinden sich die geschätzten t-Werte in Klammern. Da alle absoluten Werte der geschätzten t -Werte größer sind als die kritischen t -Werte, müssen alle Nullhypothesen verworfen werden.

Nun endlich zu den Ergebnissen: Alle Koeffizienten in Tab. 3 weisen die erwarteten Vorzeichen auf und sind auf einem 99 % Konfidenzniveau von Null verschieden. Dies bedeutet z. B. für das oben erwähnte Hypothesenpaar in Bezug auf das Alter: Die Nullhypothese muss verworfen werden. Der β_1 -Koeffizient ist (mit einer Wahrscheinlichkeit größer als 99 %) von Null verschieden. Somit liegt ein signifikanter Einfluss des Alters auf das Gehalt vor.

Tabelle 3: Regressionsergebnisse

Geschätzter Koeffizient	Variable	Koeffizient (t-Wert)
$\hat{\beta}_0$	Konstante	500,8*** (10,0)
$\hat{\beta}_1$	Alter	28,3*** (11,0)
$\hat{\beta}_2$	Alter ²	-0,27*** (-8,7)
$\hat{\beta}_3$	Ausbildung	97,4*** (8,7)
$\hat{\beta}_4$	Studium	491,3*** (42,7)
$\hat{\beta}_5$	Pflege	-53,9*** (-4,7)
$\hat{\beta}_6$	Gender	-274,4*** (-24,6)

Interpretation der Regressionsergebnisse ist wichtig

Studierende haben oftmals Probleme, das Ergebnis des Hypothesentests verbal richtig zu interpretieren. Man lehnt niemals die Alternativhypothese ab. In Bezug auf die Ablehnung einer Hypothese geht es nur um die Nullhypothese. Diese muss u. U abgelehnt werden oder sie kann nicht abgelehnt werden. Eine Ablehnung der Nullhypothese hat auch NICHT zur Folge, dass die Alternativhypothese angenommen wird. Bzgl. der Annahme einer Hypothese kann man keine Aussage treffen.

In Bezug auf die Gender-Dummy Variable ist anzumerken, dass der geschätzte Koeffizient einen Wert von $\hat{\beta}_6 = -274,4$ EUR aufweist. Somit verdienen Frauen auch unter Berücksichtigung von Kontrollvariablen signifikant weniger als Männer. Zu berücksichtigen ist jedoch, dass der in der deskriptiven Analyse quantifizierte Unterschied von 362 EUR etwas "geschmolzen" ist. Zum Teil ist der in der deskriptiven Analyse ermittelte Gehaltsunterschied also darauf zurückzuführen, dass Männer einen höheren Bildungsabschluss aufweisen und in einer Branche arbeiten, in der höhere Löhne gezahlt werden. Ein Teil des Gehaltsunterschieds ist somit auf unterschiedliche Charakteristika (Bildungsniveau & Branchenzugehörigkeit) zurückzuführen.

In den folgenden beiden Abschnitten werden zwei Exkurse gemacht, um die Regressionsergebnisse noch weiter zu interpretieren.

3.4 Exkurs: Gehaltsprognosen auf Basis der Regressionsergebnisse

Im Folgenden wurden die Werte aus der Regression stark gerundet, so dass man mit glatten Werten weiterrechnen kann. Diese Ergebnisse sollen verwendet werden, um Gehaltsprognosen für Personen mit unterschiedlichen Charakteristika zu erstellen.

$$\widehat{lohn} = \widehat{500} + \widehat{30} \cdot \widehat{alte} - \widehat{0,3} \cdot \widehat{alte}^2 + \widehat{100} \cdot \widehat{ausb} + \widehat{500} \cdot \widehat{stud} - \widehat{50} \cdot \widehat{pfle} - \widehat{275} \cdot \widehat{gend} \quad (2)$$

Prognosen auf Basis der Regressionsergebnisse

Fall 1: Wie hoch ist das prognostizierte Gehalt einer 40-jährigen Frau, mit einer Schulausbildung als höchsten Bildungsabschluss, die als Dachdeckende arbeitet?

$$\begin{aligned}\widehat{lohn} &= \widehat{500} + \widehat{30} \cdot 40 - \widehat{0,3} \cdot 1600 + \widehat{100} \cdot 0 + \widehat{500} \cdot 0 - \widehat{50} \cdot 0 - \widehat{275} \cdot 1 \\ &= 500 + 1200 - 480 - 275 = 945\end{aligned}$$

Für eine Person mit oben skizzierten Charakteristika würde man ein Gehalt von 945 EUR prognostizieren.

Fall 2: Wie hoch ist das prognostizierte Gehalt einer 40-jährigen Frau, mit einer Berufsausbildung, die in der Pflege arbeitet?

$$\begin{aligned}\widehat{lohn} &= \widehat{500} + \widehat{30} \cdot 40 - \widehat{0,3} \cdot 1600 + \widehat{100} \cdot 1 + \widehat{500} \cdot 0 - \widehat{50} \cdot 1 - \widehat{275} \cdot 1 \\ &= 500 + 1200 - 480 + 100 - 50 - 275 = 995\end{aligned}$$

Für eine Person mit oben skizzierten Charakteristika würde man ein Gehalt von 995 EUR prognostizieren. Somit wirken sich die abgeschlossene Berufsausbildung (+100 EUR) positiv und die Tätigkeit in der Pflegebranche (−50 EUR) negativ auf das prognostizierte Gehalt aus.

Fall 3: Wie hoch ist das prognostizierte Gehalt eines 30-jährigen Mannes, mit einem Studium, der in der Branche der Dachdeckenden arbeitet?

$$\begin{aligned}\widehat{lohn} &= \widehat{500} + \widehat{30} \cdot 30 - \widehat{0,3} \cdot 900 + \widehat{100} \cdot 0 + \widehat{500} \cdot 1 - \widehat{50} \cdot 0 - \widehat{275} \cdot 0 \\ &= 500 + 900 - 270 + 500 = 1630\end{aligned}$$

Der Mann aus dem Fall 3 verdient also deutlich mehr als die Frau aus dem Fall 2, obwohl der Mann jünger ist. Dieses Charakteristikum wirkt sich

ja im Vergleich zum Fall 2 negativ auf das prognostizierte Gehalt aus. Das höhere Gehalt ist vielmehr damit zu begründen, dass der Mann als höchsten Bildungsabschluss ein Hochschulstudium aufweist (+500 EUR) und eben ein Mann ist und der "Abschlag" von 275 EUR nicht anfällt.

Frage 3:

Fall 4: Wie hoch ist das prognostizierte Gehalt einer 50 jährigen Frau, mit einem Hochschulabschluss und einem Job in der Pflegebranche?

3.5 Exkurs: Zusammenhang zwischen Gehalt und Alter

Zusammenhang zwischen Gehalt und Alter ist nicht-linear

In der nachfolgenden Abb. 4 ist der Zusammenhang zwischen dem Alter und dem Gehalt dargestellt. Dazu wurden für das Alter Zahlen zwischen 20 und 65 Jahren generiert und diese in die Gleichung

$$\text{Gehaltskomponente_Alter} = 30 \cdot \text{Alter} - 0,3 \cdot \text{Alter}^2 \quad (3)$$

eingesetzt. Man erkennt, wie diese Gehaltskomponente mit dem zunächst Alter ansteigt. Der Zusammenhang ist jedoch nicht linear sondern konkav. Bei einem Wert von ca. 50 Jahren erreicht die Kurve ein Maximum. Für Personen, welche älter als ca. 50 Jahre sind, steigt die entsprechende Gehaltskomponente mit jedem weiteren Lebensjahr nicht mehr an sondern reduziert sich mit jedem weiteren Lebensjahr. Das Maximum lässt sich durch Ableiten der o.g. Gleichung nach dem Alter bestimmen. Man erhält als notwendige Bedingung:

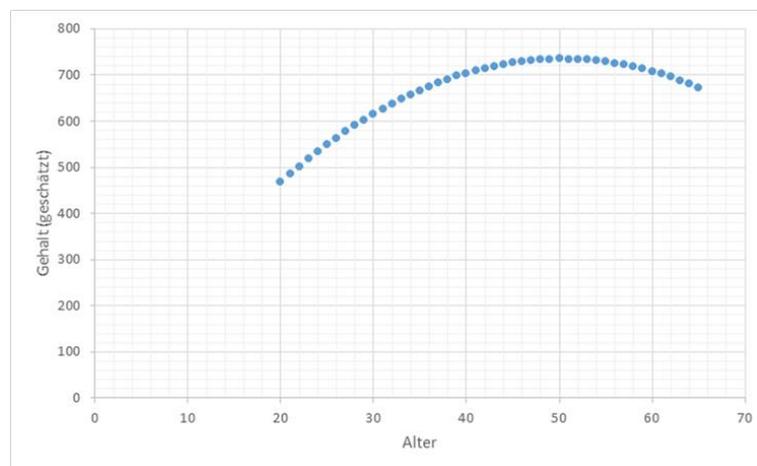
$$\frac{d\text{Gehaltskomponente_Alter}}{d\text{Alter}} = 30 + 0.6 \cdot \text{Alter} = 0 \quad (4)$$

Löst man nun nach dem Alter auf, so folgt:

$$\text{Alter} = 30/0,6 = 300/6 = 50$$

Personen mit 50 Jahren weisen das höchste Gehalt auf. Dort liegt in der Tat das Maximum bzgl. dieser Gehaltskomponente. Somit wird deutlich: Methoden wie z. B. der Maximierung aus der Mikroökonomik können mit Daten aus der empirischen Forschung gefüttert werden!

Abbildung 4: Gehalt in Abhängigkeit des Alters



4 Oaxaca/Blinder-Dekomposition

Um die Ursachen bzgl. der Gehaltsunterschiede zwischen Männern und Frauen weiter aufzuschlüsseln, müssen weitere Analysen vorgenommen werden. Es könnte sich z. B. die Frage anschließen, ob mit zunehmendem Alter Männer und Frauen einen gleich hohen Anstieg des Gehalts erwarten können.

Sind Koeffizienten für Männer bzw. Frauen gleich?

So könnte unter Umständen der geschätzte Koeffizient von $\hat{\beta}_1 \approx 30$ EUR auch darauf zurückzuführen sein, dass das Gehalt von Männern mit jedem

weiteren Geburtstag um $\beta_1^{Mann} \approx 40$ EUR ansteigt, während das Gehalt der Frauen nur um $\beta_1^{Frau} \approx 20$ EUR ansteigt. Der zuvor geschätzte Koeffizient von $\hat{\beta}_1 \approx 30$ EUR ist nämlich nur ein durchschnittlicher Wert für die gesamte Gruppe. Eine tiefergehende Analyse könnte nun darin bestehen, für die Gruppe der Männer und für die Gruppe der Frauen eine separate Schätzung der Regressionsgleichung vorzunehmen und zu prüfen, inwieweit Unterschiede zwischen den beiden Gruppen bestehen. Die Ergebnisse der beiden Regressionen sind in Tab. 4 Spalte 2 (Frauen) und Spalte 3 (Männer) zu finden.

Dekomposition bringt Klarheit

Eine Möglichkeit, die Gehaltsunterschiede zwischen Männern und Frauen zu erklären, besteht in dem Oaxaca (1973) / Blinder (1973)-Dekomposition Verfahren. Die Idee hinter dem Verfahren besteht darin, den Gehaltsunterschied in mehrere Komponenten aufzuspalten. Eine Standardprozedur besteht in der Zerlegung in drei Komponenten (threefold decomposition vgl. Jann 2008):

Eine Komponente wird als Endowment-Komponente bezeichnet. Diese Komponente wird wie folgt isoliert:

- Es wird eine Regression nur für die Gruppe der Frauen geschätzt und die geschätzten Koeffizienten gespeichert ($\hat{\beta}^F$).
- Anschließend wird der Unterschied zwischen den durchschnittlichen Charakteristika der Männer und den durchschnittlichen Charakteristika der Frauen berechnet ($\bar{X}^M - \bar{X}^F$).
- Danach wird das Produkt der durchschnittlichen Unterschiede in den Charakteristika zwischen Mann und Frau und den geschätzten Koeffizienten der Gruppe der Frauen berechnet ($(\bar{X}^M - \bar{X}^F) \cdot (\hat{\beta}^F)$).
- Zuletzt werden alle individuellen Effekte aufsummiert.

Es lässt sich folgende Frage beantworten: Was wäre, wenn Frauen die gleichen Charakteristika aufweisen würden wie Männer? Welchen Anteil des Ge-

haltsunterschieds lässt sich also dadurch erklären, dass Frauen „ungünstigere“ Charakteristika aufweisen? Diese Komponente würde man *nicht* als Diskriminierung von Frauen durch den Arbeitsmarkt bewerten.

Die entsprechenden Berechnungen für die Endowment-Komponente wurden in Tab. 4 in Spalte 8 vorgenommen. Hätte die Gruppe der Frauen die gleichen Eigenschaften wie die Gruppe der Männer, so würde das Gehalt um ca. 76 EUR höher liegen. In unserem Datensatz lässt sich also der absolute Gehaltsunterschied von 361,38 EUR (Männer 1.376,73 EUR, Frauen 1.015,35 EUR) zu 21 % ($76,31/361,38$) erklären, weil Frauen „schlechtere“ bzw. ungünstigere Charakteristika aufweisen. Somit ist der größte Teil des Unterschieds auf Grund einer aktiven Diskriminierung der Frauen am Arbeitsmarkt zurückzuführen. Diese Komponente wird nun in einem zweiten Schritt quantifiziert.

Frage 4: Welchen Wert würde die Endowment-Komponente annehmen, falls Männer und Frauen die gleichen Charakteristika aufweisen würden?

Die zweite Komponente wird als Koeffizienten-Komponente bezeichnet. Diese wird wie folgt berechnet:

- Es wird eine Regression separat für die Gruppe der Männer und eine Regression separat für die Frauen geschätzt und die Differenzen der geschätzten Regressionskoeffizienten berechnet ($\hat{\beta}^M - \hat{\beta}^F$).
- Anschließend werden die durchschnittlichen Charakteristika der Gruppe der Frauen (\bar{X}^F) mit der Differenz der geschätzten Koeffizienten multipliziert ($\bar{X}^F \cdot (\hat{\beta}^M - \hat{\beta}^F)$).
- Zuletzt werden alle individuellen Effekte aufsummiert.

Somit lässt sich die Frage beantworten, ob die Frauen mehr verdienen würden, wenn die Arbeitgeber die tatsächlichen Charakteristika der Frauen im gleichen Maße, d. h. mit den gleichen Koeffizienten vergüten würden, mit

denen auch die Charakteristika der Männer vergütet werden. Die Berechnungen erfolgen in Tab. 4 Spalte 9. Würden Frauen am Arbeitsmarkt nicht diskriminiert, so würde ihre Gehalt um ca. 263,29 EUR höher liegen. Somit lässt sich ein sehr großer Teil des Gehaltsunterschieds von 361,38 EUR durch Diskriminierung von Frauen am Arbeitsmarkt erklären.

Die dritte Komponente ist ein Interaktionseffekt und stellt ein Maß für die Kombination o.g. Effekte dar. Dieser Effekt wurde in Tab. 4 Spalte 10 berechnet und nimmt einen Wert von ca. +21,79 EUR an. Der Interaktionseffekt fällt in diesem Beispiel relativ gering aus.

Tabelle 4: Regressionsergebnisse, Charakteristika & Dekomposition

1	2	3	4	5	6	7	8	9	10
Variable	$(\hat{\beta}^F)$	$(\hat{\beta}^M)$	$(\hat{\beta}^M - \hat{\beta}^F)$	(X^F)	(\bar{X}^M)	$(\bar{X}^M - \bar{X}^F)$	$(\bar{X}^M - \bar{X}^F) \cdot (\hat{\beta}^F)$	$(\bar{X}^F) \cdot (\hat{\beta}^M - \hat{\beta}^F)$	$(\bar{X}^M - \bar{X}^F) \cdot (\hat{\beta}^M - \hat{\beta}^F)$
Alter	27,24	28,74	1,50	40,90	40,37	-0,53	-14,44	61,37	-0,80
Alter ²	-0,264	-0,276	-0,01	1809,94	1778,63	-31,31	8,26	-23,02	0,40
Ausbildung	90,38	97,36	6,97	0,20	0,35	0,15	13,56	1,39	1,05
Studium	478,88	505,17	26,29	0,30	0,40	0,10	47,89	7,89	2,69
Pflege	-38,27	-71,92	-33,65	0,75	0,20	-0,55	21,05	-25,24	18,51
Konstante	245,39	486,28	240,89	1,00	1,00	0,00	0,00	240,89	0,00
SUMME							76,31	263,29	21,79

- Spalte 1: Variablenamen
- Spalten 2 – 4: Geschätzte Koeffizienten für die Gruppe der Männer und die Gruppe der Frauen sowie deren Differenzen.
- Spalten 5 – 7: Durchschnittliche Charakteristika der Männer und der Frauen sowie deren Differenzen (aus Tab. 2 übernommen).
- Spalte 8: Endowment Effekt
- Spalte 9: Koeffizienten Effekt
- Spalte 10: Interaktionseffekt
- Befehl in STATA: `oaxaca lohn alte alte-quad ausb stud pfle, by(gend) noisily xb`

5 Zusammenfassung und Fazit

Der sogenannte Gender Pay Gap, d. h. Gehaltsunterschiede von Frauen und Männern spielen in der öffentlichen Diskussion eine große Rolle. In diesem Beitrag wurde detailliert gezeigt, wie anhand eines stilisierten Datensatzes dieser Gehaltsunterschied in unterschiedliche Komponenten aufgespalten werden kann. Ein wichtiger Effekt besteht darin, dass die Gruppe der Frauen ungünstigere Charakteristika aufweisen als die Gruppe der Männer. Frauen weisen (immer noch) einen geringeren Bildungsstand auf. Ferner arbeiten Frauen in den "falschen" Branchen, d. h. Branchen, in denen i. d. R. geringere Löhne gezahlt werden. Ferner scheint auch eine Diskriminierung der Frauen stattzufinden: Selbst wenn die Gruppe der Frauen die gleichen Charakteristika aufweisen würde wie die Gruppe der Männer, würden sie dennoch ein geringeres Gehaltsniveau aufweisen.

Reale Datensätze verfügbar

Die in diesem Beitrag vorgenommene Untersuchung wurde bewusst einfach gehalten und anhand eines stilisierten Datensatzes durchexerziert. Reale Datensätze stellt z. B. Jann (2008) für Österreich zur Verfügung. Für Deutschland könnte z. B. das sozio-ökonomische Panel (SOEP) vom DIW in Berlin herangezogen werden (Busch 2013).

In diesem Beitrag wurde das Niveau des Gehalts als abhängige Variable eingesetzt. Somit können alle Ergebnisse in Euro interpretiert werden. In Forschungsarbeiten ist es üblich, die Variablen zu logarithmieren, so dass man über relative Gehaltsunterschiede sprechen muss. In dem von Jann (2008, S. 468) vorgestellten Datensatz für Österreich wird ein Gehaltsunterschied zwischen Männern und Frauen von ca. 17,3 % berechnet. Da Frauen ungünstigere Charakteristika aufweisen als Männer sind ca. 8,5 % durch die Endowment-Komponente zu erklären. 8,3 % sind durch eine Diskriminierung der Frauen zu erklären (Koeffizienten-Komponente). Nur ein geringer Teil entfällt auf die Interaktionskomponente (0,5 %).

References

- Blinder, A. S. (1973): Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources*, Vol. 8, 436 – 455.
- Busch Anne (2013): Der Einfluss der beruflichen Geschlechtersegregation auf den "Gender Pay Gap" – Zur Bedeutung geschlechtlich konnotierter Arbeitsinhalte. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, Vol. 65, 301 – 338.
- Jann, Ben (2008): The Blinder-Oaxaca decomposition for linear regression models. *The Stata Journal*, Vol. 8(4), 453 – 479.
- Oaxaca, R. (1973): Male-female wage differentials in urban labor markets. *International Economic Review*, Vol. 14, 693 – 709.

Antworten

Antwort 1

Eine deskriptive Analyse hilft, um zu überprüfen ob die Daten richtig in das Datenverarbeitungsprogramm eingelesen wurden. So kann z.B. in einem ersten Schritt überprüft werden, ob die Anzahl an Beobachtungen im Ökonometrieprogramm mit dem Datenfile übereinstimmt. Außerdem kann dann überprüft werden, ob die Daten in einem Wertebereich liegen, die ökonomisch Sinn machen. Ferner hilft die deskriptive Statistik, einen ersten Überblick über die Daten zu gewinnen.

Antwort 2

Eine Ursache könnte darin bestehen, dass Männer andere, d.h. bessere Charakteristika aufweisen als Frauen. Weist die Gruppe der Männer im Durchschnitt z. B. einen höheren Bildungsabschluss auf als die Gruppe der Frauen, so sollte *ceteris paribus* (unter sonst gleichen Bedingungen) auch ein Gehaltsunterschied bestehen.

Eine weitere Ursache könnte auch darin bestehen, dass Frauen am Arbeitsmarkt diskriminiert werden: Hat die Gruppe der Frauen – im Durchschnitt – die gleichen Charakteristika wie die Gruppe der Männer, so könnte ein Gehaltsunterschied u. U. auf eine negative Diskriminierung der Gruppe der Frauen hindeuten.

Antwort 3

Fall 4: Wie hoch ist das prognostizierte Gehalt einer 50-jährigen Frau, mit einem Hochschulabschluss und einem Job in der Pflegebranche?

$$\begin{aligned}\widehat{lohn} &= \widehat{500} + \widehat{30} \cdot 50 - \widehat{0,3} \cdot 2500 + \widehat{100} \cdot 0 + \widehat{500} \cdot 1 - \widehat{50} \cdot 1 - \widehat{275} \cdot 1 \\ &= 500 + 1500 - 750 + 500 - 50 - 275 = 1.425\end{aligned}$$

Für eine Person mit oben skizzierten Charakteristika würde man ein Gehalt von 1.425 EUR prognostizieren.

Antwort 4

Würden die durchschnittlichen Charakteristika der Gruppe der Männer und der Gruppe der Frauen übereinstimmen, dann wäre der erste Klammerausdruck $(\bar{X}^M - \bar{X}^F) \cdot (\hat{\beta}^F)$ bei der Berechnung der Endowment-Komponente gleich Null. Somit würde die Endowment-Komponente den Wert Null annehmen.