FÖRDJUPNING

Dekomponering av löneskillnader

Den här fördjupningen ger en detaljerad beskrivning av dekomponeringen av skillnader i genomsnittlig lön. Först beskrivs metoden i generella termer. Därefter beskrivs och jämförs dekomponering med log-linjära respektive dekomponering med exponentiella löneekvationer.

Ett sätt att analysera skillnader i löner mellan män och kvinnor är att göra en så kallad dekomponering av differensen av den genomsnittliga lönen för respektive grupp.[172] Denna metod utgår från att skillnaden i lön mellan män och kvinnor beror på sammansättningsskillnader eller på prisskillnader. Sammansättningsskillnader följer av att män och kvinnor som grupper har olika fördelning av lönepåverkande faktorer. Om exempelvis lön stiger med ålder och män har en högre medelålder än kvinnor kommer detta att bidra till att män har en högre medellön än kvinnor. Prisskillnader följer av att de lönepåverkande faktorerna är prissatta olika för män och kvinnor. Om exempelvis mäns löner ökar med ålder i snabbare takt än kvinnors så kommer män att ha en högre medellön vid en given medelålder.

Dekomponering bygger på att en så kallad kontrafaktisk medellön beräknas. Givet denna beräkning kan differensen av medellön (μ) mellan män (M) och kvinnor (K) skrivas som:

$$\mu_M - \mu_K = \underbrace{\mu_M - \mu_K^C}_{förklarad} + \underbrace{\mu_K^C - \mu_K}_{oförklarad}$$

där μ_K^C är den kontrafaktiska medellön som hade uppmätts för kvinnor om deras lönepåverkande faktorer prissatts som för män.[173] Differensen av mäns medellön och den kontrafaktiska medellönen för kvinnor utgör den förklarade delen av löneskillnaden. Den förklarade delen anger således vad löneskillnaden hade varit om den enda skillnaden mellan män och kvinnor varit fördelningen av lönepåverkande faktorer. Den oförklarade delen, som utgörs av differensen mellan kvinnors kontrafaktiska medellön och kvinnors faktiska medellön, anger vad löneskillnaden hade varit om den enda skillnaden mellan män och kvinnor

[172] Det går även att dekomponera differensen av andra statistikor (se Fortin m.fl., 2011). I avsnitt 4.4 dekomponeras till exempel differensen av två precentilkvoter. En annan metod är så kallad standardvägning (se till exempel Medlingsinstitutet, 2012).

[173] Alternativt kan det kontrafaktiska utfallet där män erätts som kvinnor användas. Valet av kontrafaktiskt utfall är ett indexproblem som handlar om vilka priser, mäns eller kvinnors, som ska vara referensen för dekomponeringen. När det gäller löneskillnader mellan män och kvinnor är det standard i litteraturen att välja mäns priser som referens.

Aritmetiskt och geometriskt medelvärde

Aritmetiskt medelvärde: För n observationer av en variabel (y) är det aritmetiska medelvärdet summan av de observerade värdena dividerat med antalet observationer, det vill säga

$$AM = \frac{\sum_{i=1}^{n} y_i}{n}$$

Geometriskt medelvärde: För n observationer av en variabel (y) är det geometriska medelvärdet den n:te roten ur produkten av de observerade värdena, det vill säga

$$GM = \left(\prod_{i=1}^{n} y_i\right)^{1/n}$$

hade varit prissättningen av de lönepåverkande faktorerna. De lönepåverkande faktorernas observerbarhet och bristande tillgänglighet till statistik gör emellertid att det i praktiken är omöjligt att inkludera alla lönepåverkande faktorer. Den oförklarade delen kommer därför att bestå av prisskillnader för de inkluderade faktorerna och effekten av utelämnade faktorer. Detta gör att resultaten bör tolkas med försiktighet.

Det finns olika sätt att beräkna den kontrafaktiska medellönen, men för löneskillnader genomförs ofta en så kallad Blinder-Oaxaca-dekomponering där den kontrafaktiska medellönen beräknas med hjälp av löneekvationer. Vanligtvis används log-linjära löneekvationer eftersom det är den i litteraturen dominerande funktionsformen för att beskriva sambandet mellan löner och olika variabler. Dels är den teoretiskt förankrad, dels fungerar den väl empiriskt.[174] En dekomponering baserad på log-linjära löneekvationer innebär emellertid att det är differensen av medelvärdet av *logaritmerad* lön och inte medelvärdet av lön som dekomponeras. Denna distinktion kan förefalla obetydlig, men den innebär att löneskillnaden mellan grupperna kvantifieras utifrån geometriska medelvärden och inte utifrån aritmetiska medelvärden (se förklaring i marginalen).[175]

Dekomponering med log-linjära löneekvationer

Dekomponeringen utgår från att en löneekvation, det vill säga ett statistiskt samband (regressionsekvation) som beskriver sambandet mellan lön och ett antal förklarande variabler, skattas separat för män och kvinnor. Med log-linjära löneekvationer har detta samband följande utseende:

$$\ln y_i = \alpha + x_i\beta + u_i,$$

där y_i är lön för person i, α är en konstant, β är en vektor av koefficienter (priser), x_i är en vektor av förklarande variabler, och u_i är en slumpterm med medelvärdet noll. Med hjälp av de skattade löneekvationerna kan medelutfallet för män respektive kvinnor beräknas som $\overline{\ln y}_M = \hat{\alpha}_M + \bar{x}_M\hat{\beta}_M$ och $\overline{\ln y}_K = \hat{\alpha}_K + \bar{x}_K\hat{\beta}_K$ där M och K står för män respektive kvinnor, strecken indikerar att det är frågan om (aritmetiska) medelvärden och $\hat{\alpha}$ samt $\hat{\beta}$ är de skattade regressionskoefficienterna. Det går även att beräkna det kontrafaktiska utfallet $\overline{\ln y}_C = \hat{\alpha}_M + \bar{x}_K\hat{\beta}_M$ som är det medelutfall som kvinnor haft om de haft mäns regressionskoefficienter (priser).

---

[174] Se Kaiser (2012).

[175] Att det är frågan om geometriska medelvärden poängterades av Oaxaca (1973).

Med hjälp av de beräknade utfallen kan differensen av mäns och kvinnors medelutfall (G) skrivas som

$$
\begin{aligned}
G &= \overline{\ln y_M} - \overline{\ln y_K} \\
&= \overline{\ln y_M} - \overline{\ln y_C} + \overline{\ln y_C} - \overline{\ln y_K} \\
&= \underbrace{(\bar{x}_M - \bar{x}_K)\hat{\beta}_M}_{förkarad} + \underbrace{(\hat{\alpha}_M - \hat{\alpha}_K) + \bar{x}_K(\hat{\beta}_M - \hat{\beta}_K)}_{oförklarad}.
\end{aligned}
$$

I det sista uttrycket framgår tydligt att den förklarade delen beror på skillnader i förklarande variabler och den oförklarade delen beror på skillnader i koefficienter. Här kan noteras att utelämnade faktorer kommer påverka konstanterna och effekten av dessa kommer således att inkluderas i den oförklarade delen. Det framgår också att det som dekomponeras är differensen av gruppernas aritmetiska medelvärde av *logaritmerad* lön. Detta är det samma som differensen av gruppernas logaritmerade geometriska medellön.[176] Detta är i sin tur en approximation av den relativa differensen av geometrisk medellön (G_{GM}), det vill säga

$$
\begin{aligned}
G &= \ln(GM_M) - \ln(GM_K) \\
&= \ln\left(\frac{GM_M}{GM_K}\right) \approx \frac{GM_M - GM_K}{GM_K} = G_{GM},
\end{aligned}
$$

där GM_j är respektive grupps geometriska medellön. Approximationen blir mer precis ju mindre G_{GM} är.

Att mäta löneskillnader i termer av geometriska medelvärden kan vara problematiskt. Då det geometriska medelvärdet avviker från den gängse uppfattningen av vad ett medelvärde är kan resultaten vara svåra att förmedla. Ett annat problem, som exemplifierades i inledningen av kapitlet, är att jämförelser av geometriska medelvärden fångar upp skillnader i spridning mellan lönefördelningar. Hur stort detta problem är i praktiken beror på hur lönefördelningarna ser ut. Det går att visa att G är en kombination av differensen av logaritmerad aritmetisk medellön och differensen av de logaritmerade lönefördelningarnas högre moment.[177] För två lönefördelningar som har likartad spridning kommer G att vara nära en approximation av den relativa differensen av aritmetiska medellöner. I andra fall riskerar G att ge en missvisande bild av löneskillnaden. För att jämföra genomsnittliga löner mellan män och kvinnor är ett mått baserat på aritmetiska medelvärden mer intuitivt.

[176] Logaritmeras båda sidor av formeln för det geometriska medelvärdet fås att det aritmetiska medelvärdet av logaritmerade värden är lika med det logaritmerade geometriska medelvärdet, det vill säga: $\ln GM = \sum_{i=1}^{n} \ln y_i / n$.

[177] Leslie och Murphy (1997).

Dekomponering med exponentiella löneekvationer

För att dekomponera differensen av aritmetiska medellöner kan dekomponeringen baseras på exponentiella löneekvationer.[178] Sambandet mellan lön och de förklarande variablerna specificeras då som

$$y_i = e^{\gamma + x_i\delta + u_i}.$$

Notationen är den samma som ovan. Logaritmeras denna ekvation erhålls den log-linjära ekvationen. De båda ekvationerna beskriver således sambandet mellan lön och de förklarande variablerna på samma sätt. Det finns olika sätt att skatta den exponentiella löneekvationen. I det här kapitlet används så kallad *Poisson quasi maximum likelihood*.[179] Med de skattade löneekvationerna kan medellönen i respektive grupp beräknas som $\bar{y}_M = \overline{e^{\hat{\gamma}_M + x_{i,M}\hat{\delta}_M}}$ respektive $\bar{y}_K = \overline{e^{\hat{\gamma}_K + x_{i,K}\hat{\delta}_K}}$, och den kontrafaktiska medellönen som $\bar{y}_C = \overline{e^{\hat{\gamma}_M + x_{i,K}\hat{\delta}_M}}$. Med dessa beräkningar kan nu differensen av gruppernas medellön skrivas som

$$\bar{y}_M - \bar{y}_K = \underbrace{\left(\overline{e^{\hat{\gamma}_M + x_{i,M}\hat{\delta}_M}} - \overline{e^{\hat{\gamma}_M + x_{i,K}\hat{\delta}_M}}\right)}_{\textit{förklarad}} + \underbrace{\left(\overline{e^{\hat{\gamma}_M + x_{i,K}\hat{\delta}_M}} - \overline{e^{\hat{\gamma}_K + x_{i,K}\hat{\delta}_K}}\right)}_{\textit{oförklarad}}.$$

Medellönen är nu den aritmetiska medellönen. I övrigt är tolkningen den samma som för dekomponeringen med log-linjära löneekvationer.

Detaljerad dekomponering

Hittills har bara dekomponering på en aggregerad nivå beskrivits. För att analysera löneskillnaden ytterligare kan en detaljerad dekomponering göras. En detaljerad dekomponering visar hur mycket var och en av de inkluderade variablerna bidrar till den förklarade delen och den oförklarade delen. I det här kapitlet görs emellertid bara en detaljerad dekomponering av den förklarande delen. Detta på grund av det invariansproblem som präglar detaljerad dekomponering av den oförklarade delen.[180] Invariansproblemet innebär att den detaljerade dekomponeringen varierar beroende på vilken variabel av en grupp indikatorvariab-

[178] Kaiser (2012).

[179] Se till exempel Wooldridge (2009).

[180] Se till exempel Andrén (2012b).

ler som utelämnas vid skattningarna av löneekvationerna och är därmed mer eller mindre godtycklig.

För dekomponering med log-linjära löneekvationer är den detaljerade dekomponeringen rättfram då den förklarade delen är summan av varje variabels bidrag. Det vill säga,

$$(\bar{x}_M - \bar{x}_K)\hat{\beta}_M = \sum_l (\bar{x}_M^l - \bar{x}_K^l)\hat{\beta}_M^l.$$

Variabel l bidrar med $(\bar{x}_M^l - \bar{x}_K^l)\hat{\beta}_M^l$ till den förklarade delen.

För dekomponering med exponentiella löneekvationer är den detaljerade dekomponeringen inte lika rättfram då den är icke-linjär. För att genomföra den detaljerade dekomponeringen används i det här kapitlet vikter som erhålls genom en linjärisering runt $\bar{x}_M\hat{\delta}_M$.[181] Vikten för variabel l beräknas som

$$w^l = \frac{(\bar{x}_M^l - \bar{x}_K^l)\hat{\delta}_M^l}{\sum_l (\bar{x}_M^l - \bar{x}_K^l)\hat{\delta}_M^l}.$$

Summan i nämnaren omfattar alla variabler. Med vikterna kan den förklarade delen skrivas som

$$E = \overline{e^{\hat{\gamma}_M + x_{i,M}\hat{\delta}_M}} - \overline{e^{\hat{\gamma}_M + x_{i,K}\hat{\delta}_M}} = \sum_l w^l E$$

Variabel l bidrar med $w^l E$ till den förklarade delen.

En jämförelse av dekomponering med log-linjära respektive exponentiella löneekvationer

I det här avsnittet jämförs dekomponering med log-linjära respektive exponentiella löneekvationer med hjälp av löneskillnader bland privatanställda tjänstemän respektive arbetare. Tabell 27 visar de olika löneskillnadsmått som är inblandade i jämförelsen.

[181] Yun (2004).

Tabell 27 Olika löneskillnadsmått för tjänstemän och arbetare i privat sektor, 2012

Procent

Löneskillnadsmått	Tjänstemän	Arbetare
ln GM_M – ln GM_K (log-linj. löneekv.)	19,5	9,5
$(GM_M - GM_K)/GM_K$	21,5	10,0
$(AM_M - AM_K)/AM_K$ (exp. löneekv.)	24,9	10,0

Anm. GM=geometrisk medellön, AM=artimetisk medellön. Index M och K för män respektive kvinnor. Vid dekomponering med exponentiella löneekvationer dekomponeras $AM_M - AM_K$; här redovisas den relativa differensen.

Källor: SCB och Konjunkturinstitutet.

För tjänstemän uppgår differensen av logaritmerad geometrisk medellön till 19,5 procent. Detta är en approximation av den relativa differensen av geometrisk medellön som uppgår till 21,5 procent. Approximationsfelet är således 2 procentenheter. Den relativa differensen av aritmetisk medellön uppgår för tjänstemän till 24,9 procent. Valet av löneekvation och därmed löneskillnadsmått påverkar alltså storleken på den löneskillnad som rapporteras. Detta är i sig inget problem, men om resultaten från en dekomponering med log-linjära löneekvationer uppfattas som en differens av aritmetiska medellöner kommer löneskillnaden att underskattas. I detta fall med 5,4 procentenheter eller 22 procent. Skillnaden mellan de två löneskillnadsmåtten beror på att bland tjänstemän har män en större lönespridning än kvinnor (se avsnitt 4.4).

För arbetare uppgår differensen av logaritmerad geometrisk medellön till 9,5 procent vilket är 0,5 procentenheter lägre än både den relativa differensen av geometrisk medellön och den relativa differensen av aritmetisk medellön. Valet av löneekvation för arbetare spelar således en mindre roll för den uppmätta löneskillnaden. Detta beror på att lönefördelningarna för män och kvinnor i detta fall har likartad spridning (se avsnitt 4.4).

Tabell 28 redovisar resultaten av en dekomponering inklusive en detaljerad dekomponering av den förklarade löneskillnaden med respektive löneekvation för tjänstemän och arbetare. Överlag är det liten skillnad i resultaten. För tjänstemän förklaras en något större del av löneskillnaden med log-linjära löneekvationer, 54 procent jämfört med 51 procent med exponentiella löneekvationer. För arbetare är den förklarade löneskillnaden 87 procent av den totala löneskillnaden oavsett ekvationsform. Andelarna som varje faktor förklarar vid detaljerad dekomponering av den förklarade löneskillnaden varierar något mellan de olika funktionsformerna, men det föreligger inga anmärkningsvärda skillnader i resultat.

Tabell 28 Dekomponering med log-linjära respektive exponentiella löneekvationer för tjänstemän och arbetare i privat sektor, 2012

Procent

	Tjänstemän		Arbetare	
	Log-linj.	Exp.	Log-linj.	Exp.
Oförklarad löneskillnad	*9,0*	*12,3*	*1,2*	*1,3*
Förklarad löneskillnad	*10,5*	*12,6*	*8,4*	*8,6*
Varav:				
Yrkesgrupp	6,8	8,5	3,2	3,6
Utbildningsnivå	–0,1	–0,3	–0,0	–0,1
Utbildningsinriktning	0,7	0,9	0,5	0,6
Födelseland	0,1	0,1	0,1	0,1
Åldersgrupp	0,9	1,2	1,2	1,2
Region	–0,5	–0,7	–0,1	–0,2
Tjänsteomfattning	0,2	0,3	–0,0	–0,2
Timlön	0,3	0,3	0,2	0,2
Företagsstorlek	–0,3	–0,4	–0,2	–0,2
Näringsgren	2,3	2,6	3,6	3,6

Anm. Samtliga förklarande variabler är indikatorvariabler (dummyvariabler), Variablerna indikerar vilken *yrkesgrupp* (3-ställig SSYK) och *åldersgrupp* (18–24, 25–29, 30–34,…, 55–59 respektive 60–64 år) individen tillhör; individens *utbildningsnivå* och *utbildningsinriktning* (SUN 2000), *födelseland* (Sverige, Norden exkl., Sverige, Europa exkl. Norden, utanför Europa) och *tjänsteomfattning* (1–25, 26–50, 51–75, 76–90 respektive 91–100 procent); vilken *region* (riksområde NUTS-2) individen arbetar i; om individen har *timlön;* samt *företagsstorlek* (1–4, 5–9, 10–19, 20–49, 50–99, 100–199, 200–499 respektive 500- anställda) och *näringsgren* (SNI2002 bokstavsnivå med A+B) för det företag där individen arbetar.

Källor: SCB och Konjunkturinstitutet.