计量经济学第五次作业

Karry

1 Oaxaca 分解法介绍（Oaxaca Decomposition Method）

男女工资存在差异，Oaxaca分解法认为可以将这一差异分解为两部分：

- 一部分代表男女个体特征差异（解释变量X的差别）；
- 一部分代表特征的回报差异（系数的差异）。

前一部分的差异是较为“公平的”，后一部分的差异则可能与劳动力市场的“歧视”有关。 假定个体的工资方程可以由下式表示：

$$Y = X\beta + u$$

其中 Y 代表工资（取对数），X是一系列影响工资的个体特征（年龄、教育程度、工作经验等）。假定 M、F 分别代表男性与女性，这样男女工资差异可以表示成：

$$\Delta_O = \overline{Y}_M - \overline{Y}_F$$

其中$\overline{Y}_M, \overline{Y}_F$分别代表男女的平均工资水平；$\Delta_O$代表男女平均工资的差异。

接下来，我们可以分别针对男女样本进行回归。基于回归可以得到：

$$\Delta_O = \overline{Y}_M - \overline{Y}_F = \overline{X}_M\hat{\beta}_M - \overline{X}_F\hat{\beta}_F$$

然后通过加减一项 $\overline{X}_M\hat{\beta}_F$ 可以得到 Oaxaca 分解的公式

$$\begin{aligned}\Delta_O &= (\overline{X}_M\hat{\beta}_M - \overline{X}_M\hat{\beta}_F) + (\overline{X}_M\hat{\beta}_F - \overline{X}_F\hat{\beta}_F) \\ &= (\overline{X}_M - \overline{X}_F)\hat{\beta}_F + \overline{X}_M(\hat{\beta}_M - \hat{\beta}_F) \quad (1)\end{aligned}$$

或者是通过加减一项 $\overline{X}_F\hat{\beta}_M$ 可以得到 Oaxaca 分解的公式

$$\begin{aligned}\Delta_O &= (\overline{X}_M\hat{\beta}_M - \overline{X}_F\hat{\beta}_M) + (\overline{X}_F\hat{\beta}_M - \overline{X}_F\hat{\beta}_F) \\ &= (\overline{X}_M - \overline{X}_F)\hat{\beta}_M + \overline{X}_F(\hat{\beta}_M - \hat{\beta}_F) \quad (2)\end{aligned}$$

定义： $\Delta_X = (\overline{X}_M - \overline{X}_F)\hat{\beta}_F$ （或者 $(\overline{X}_M - \overline{X}_F)\hat{\beta}_M$）；
$\Delta_S = \overline{X}_M(\hat{\beta}_M - \hat{\beta}_F)$ （或者 $\overline{X}_F(\hat{\beta}_M - \hat{\beta}_F)$）。 其中 Δ_X代表男女之间的特征（禀赋）差异导致的工资差别，也就是男女之间的工资差异可以被其特征差异解释的部分（explained），可以界定为相对“公平”的部分；而Δ_S代表男女生即使在基本特征相同的情况下，仍然会存在的工资差别，是无法被

可观测的二者特征差异所解释的部分（unexplained)，因此可以被界定为“不公平”的部分。

想要完成本次作业中后续结果的解释，仅仅理解上面的分解过程是不足够的，在此引出更加一般的分解方式：有理由相信的确存在非歧视性的系数向量，其剔除了所有歧视的影响来表现解释变量的偏效应，假设 β^* 就是这样一个非歧视性的系数向量，那么男女工资差异可以表示为：

$$\begin{aligned}\Delta_O &= \overline{X}_M\hat{\beta}_M - \overline{X}_F\hat{\beta}_F \\ &= (\overline{X}_M - \overline{X}_F)\beta^* + [\overline{X}_M(\beta_M - \beta^*) + \overline{X}_F(\beta^* - \beta_F)]\end{aligned}$$

现在做一个两阶段分解

$$\Delta_O = Q + U$$

定义：$Q = (\overline{X}_M - \overline{X}_F)\beta^*$ 表示可以被解释的部分；

$U = [\overline{X}_M(\beta_M - \beta^*) + \overline{X}_F(\beta^* - \beta_F)]$ 表示不可以被解释的部分；

那 β^* 该怎么确定呢？下面给出三种确定方法：

1. 假设工资歧视只针对女性，而不存在对男性的（积极）“歧视”，那么可以用 β_M 来作为 β^* 的估计值，就可以得到式（2）
2. 假设对女性来说并不存在工资歧视，而是只对男性存在（积极）“歧视”，那么可以用 β_F 作为 β^* 的估计值，就可以得到式（1）
3. 还可以用 β_M 和 β_F 的线性组合来估计 β^* 即令：

$$\beta^* = weight \times \beta_M + (1 - weight) \times \beta_F\$$

2　使用数据完成以上分析过程

（1）执行以下回归命令，并结合以上的模型推导对结果做出解释：

```
reg lropc00 age00 msa ctrlcity north_central south00 west sch_10
diploma_hs ged_hs smcol bachelor_col master_col doctor_col if
female==0 & white==1 // 在男性白人中进行回归
estimates store male // 保存回归结果

reg lropc00 age00 msa ctrlcity north_central south00 west sch_10
diploma_hs ged_hs smcol bachelor_col master_col doctor_col if
female==1 & white==1 // 在女性白人中进行回归
estimates store female // 保存回归结果

oaxaca8 male female, weight(1 0 0.651) detail notf // 对两回归结果进行
分析
```

答：上述命令执行结果可分为四部分做详细解释

Part Ⅰ收入总差异

```
. oaxaca8 male female, weight(1 0 0.651) detail notf
(high estimates: male; low estimates: female)

                                    Mean prediction 1 =   2.843339
                                    Mean prediction 2 =   2.538099
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
difference	.3052401	.0232167	13.15	0.000	.2597363	.3507439

图 1. 男性女性收入总差异

第一部分呈现了回归结果中男性女性的收入总差异，也即 $\Delta_O = \overline{Y}_M - \overline{Y}_F = 0.3052$ 这说明：男性和女性的收入水平存在差距，男性收入水平比女性收入水平高30.52%

Part Ⅱ Weight = 1 的差异情况

Linear decompositions

W=1	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
explained						
age00	-.0022252	.0020132	-1.11	0.269	-.006171	.0017206
msa	-.00265	.0026628	-1.00	0.320	-.0078689	.0025688
ctrlcity	-.0001546	.0015731	-0.10	0.922	-.0032379	.0029286
north_central	-.0028031	.0023995	-1.17	0.243	-.0075061	.0018998
south00	.001774	.0018145	0.98	0.328	-.0017824	.0053304
west	.0002001	.0008748	0.23	0.819	-.0015145	.0019147
sch_10	-.0015512	.0025023	-0.62	0.535	-.0064556	.0033532
diploma_hs	.0004179	.0030613	0.14	0.891	-.0055821	.0064179
ged_hs	.0002651	.000926	0.29	0.775	-.0015498	.0020801
smcol	-.0141515	.0059643	-2.37	0.018	-.0258414	-.0024616
bachelor_col	-.0004948	.0083319	-0.06	0.953	-.016825	.0158354
master_col	-.0126813	.0070481	-1.80	0.072	-.0264954	.0011328
doctor_col	.0150659	.0046541	3.24	0.001	.005944	.0241879
Total	-.0189887	.010135	-1.87	0.061	-.0388529	.0008756
unexplained						
age00	.5366984	.3647461	1.47	0.141	-.1781908	1.251588
msa	-.0207122	.058977	-0.35	0.725	-.1363049	.0948806
ctrlcity	-.0187396	.016153	-1.16	0.246	-.050399	.0129197
north_central	.0146831	.0206186	0.71	0.476	-.0257287	.0550948
south00	.0221033	.0217059	1.02	0.309	-.0204396	.0646461
west	-.0019994	.0128246	-0.16	0.876	-.0271351	.0231363
sch_10	.0023061	.002803	0.82	0.411	-.0031877	.0077999
diploma_hs	.0265128	.0253679	1.05	0.296	-.0232073	.0762329
ged_hs	.0070159	.0051623	1.36	0.174	-.0031019	.0171338
smcol	.0175805	.0207796	0.85	0.398	-.0231467	.0583077
bachelor_col	.010187	.0161512	0.63	0.528	-.0214688	.0418428
master_col	.0063584	.0075405	0.84	0.399	-.0084206	.0211375
doctor_col	.0002369	.0016827	0.14	0.888	-.0030611	.003535
_cons	-.2780025	.3768883	-0.74	0.461	-1.01669	.460685
Total	.3242288	.0216201	15.00	0.000	.2818542	.3666033

图 2. Weight = 1 的差异情况

这一部分的回归结果表示在假设工资歧视只针对女性，而不存在对男性的（积极）“歧视”的情况下，那么可以用 β_M 来作为 β^* 的估计值，进而得到式（2）即：$\Delta_O = (\overline{X}_M - \overline{X}_F)\hat{\beta}_M + \overline{X}_F(\hat{\beta}_M - \hat{\beta}_F)$

进而在 Part Ⅰ 中呈现的收入总差异可以被分解为两部分：一部分为“可解释部分”（$(\overline{X}_M - \overline{X}_F)\hat{\beta}_M$）。因为男性和女性在禀赋上存在差异，所以即使假设女性在劳动市场上被视为男性，还是会与真正的男性存在收入差距。值得注意的是：“可解释部分”的总差异为 -0.01899 也即男性和女性在禀赋上的差异导致男性的收入水平是略低于女性的。

另一部分为“不可解释部分”（$X_F(\hat{\beta}_M - \hat{\beta}_F)$）。该部分为被视为男性的女性和真正的女性的收入差距，无法由男性和女性的生产力条件差异解释，也即劳动力市场对女性的歧视。结果表明：由于“可解释部分”出现了负效应，所以几乎所有的收入差距与女性在劳动力市场的差别待遇或者性别歧视有关。

当然，不论是“可解释部分”还是“不可解释部分”的差异都可以分解到每一个自变量中，例如在“可解释部分中”中是否为医生（doctor_col）这一禀赋所造成的收入差异为 0.015； 在“不可解释部分”中劳动力市场对男性和女性年龄（age）歧视所带来的收入差异高达 0.537 ！

Part Ⅲ Weight = 0 的差异情况

W=0	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
explained						
age00	.0014225	.0017834	0.80	0.425	-.0020729	.0049178
msa	-.0031072	.0030158	-1.03	0.303	-.0090182	.0028037
ctrlcity	-.0003128	.0027964	-0.11	0.911	-.0057936	.005168
north_central	-.0037882	.0029931	-1.27	0.206	-.0096546	.0020782
south00	.0032558	.002562	1.27	0.204	-.0017657	.0082772
west	.0001264	.0007365	0.17	0.864	-.0013172	.00157
sch_10	-.0048192	.0032758	-1.47	0.141	-.0112397	.0016014
diploma_hs	.0002343	.0018407	0.13	0.899	-.0033734	.0038421
ged_hs	-.0002071	.0007949	-0.26	0.794	-.0017649	.0013508
smcol	-.0116016	.0049648	-2.34	0.019	-.0213325	-.0018707
bachelor_col	-.0004529	.0076271	-0.06	0.953	-.0154018	.014496
master_col	-.011292	.0062611	-1.80	0.071	-.0235635	.0009794
doctor_col	.0146444	.0048592	3.01	0.003	.0051204	.0241683
Total	-.0158977	.0107569	-1.48	0.139	-.0369808	.0051855
unexplained						
age00	.5330508	.3622671	1.47	0.141	-.1769798	1.243081
msa	-.020255	.0576755	-0.35	0.725	-.1332969	.0927869
ctrlcity	-.0185815	.0160158	-1.16	0.246	-.0499719	.012809
north_central	.0156681	.0219996	0.71	0.476	-.0274502	.0587865
south00	.0206215	.0202526	1.02	0.309	-.0190729	.0603159
west	-.0019257	.0123521	-0.16	0.876	-.0261353	.0222839
sch_10	.0055741	.0066745	0.84	0.404	-.0075077	.0186559
diploma_hs	.0266963	.0255423	1.05	0.296	-.0233657	.0767583
ged_hs	.0074881	.0055003	1.36	0.173	-.0032922	.0182685
smcol	.0150306	.0177706	0.85	0.398	-.0197991	.0498603
bachelor_col	.0101452	.0160843	0.63	0.528	-.0213794	.0416698
master_col	.0049692	.0059035	0.84	0.400	-.0066015	.0165398
doctor_col	.0006585	.004572	0.14	0.885	-.0083025	.0096195
_cons	-.2780025	.3768883	-0.74	0.461	-1.01669	.460685
Total	.3211378	.0217203	14.79	0.000	.2785667	.3637088

图 3. Weight = 0 的差异情况

这一部分的回归结果表示在假设对女性来说并不存在工资歧视，而是只对男性存在（积极）“歧视”的情况下，那么可以用 β_F 作为 β^* 的估计值，就可以得到式(1)即：$\Delta_O = (\overline{X}_M - \overline{X}_F)\hat{\beta}_F + \overline{X}_M(\hat{\beta}_M - \hat{\beta}_F))$

后续的分析和 Part Ⅲ 是基本类似的，只不过所有“不可解释的部分”带来的差异，均是劳动力市场对男性的积极“歧视”。

从结果中看出：男性和女性在禀赋上的差异导致男性的收入水平是略低于女性的，大约低 0.0157。同时劳动力市场对男性的“积极”歧视使得男性的收入水平水平较比女性高出 0.3211。

Part Ⅳ Weight = 0.651 时的差异情况

Mean prediction 1 = 2.843339
Mean prediction 2 = 2.538099

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
difference	.3052401	.0232167	13.15	0.000	.2597363	.3507439

Linear decomposition

Total	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
W=,651						
explained	-.0179099	.0098544	-1.82	0.069	-.0372242	.0014044
unexplained	.32315	.0214196	15.09	0.000	.2811683	.3651317

图 4. Weight = 0.651 时的差异情况

> 在这种条件下，用 $0.651 \times \beta_M + (1-0.651) \times \beta_F$ 作为 β^* 的估计值，也即本回归结果时在假定劳动力市场对男性和女性均存在歧视的情况下，衡量男性和女性收入水平的差异。
>
> 男性和女性在禀赋上的差异导致男性的收入水平是略低于女性的，大约低 0.0179。同时劳动力市场对性别的歧视使得男性的收入水平水平较比女性高出 0.3232。
>
> 总的来说三种权重的回归结果整体上并没有显著差别，都呈现出：
>
> - 禀赋上的差异导致男性的收入水平略低于女性的
> - 劳动力市场对女性存在一定的“歧视”

（2）在原工资方程中增加一个控制变量 afqtp89（用来衡量个人的综合素质），再次完成以上的分解过程，分析结果有何不同？

```
replace afqtp89=afqtp89/100.0 // 改变 afqtp89 的单位

reg lropc00 age00 msa ctrlcity north_central south00 west sch_10
diploma_hs ged_hs smcol bachelor_col master_col doctor_col afqtp89
if female==0 & white==1
estimates store male

reg lropc00 age00 msa ctrlcity north_central south00 west sch_10
diploma_hs ged_hs smcol bachelor_col master_col doctor_col afqtp89
if female==1 & white==1
estimates store female

oaxaca8 male female, weight(1 0 0.651) detail notf
```

> 答：和第（1）问回归结果类似，仍然可以将其分为四个部分
>
> Part Ⅰ收入总差异

```
. oaxaca8 male female, weight(1 0 0.651) detail notf
(high estimates: male; low estimates: female)

                                Mean prediction 1 =   2.843339
                                Mean prediction 2 =   2.538099
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
difference	.3052401	.0232247	13.14	0.000	.2597204	.3507597

图 5. 男性女性收入总差异

第一部分呈现了回归结果中男性女性的收入总差异，也即 $\Delta_O = \overline{Y}_M - \overline{Y}_F = 0.3052$ 和初始回归并无任何差别。

为了更好的展示和第一问的差别，我们先把第一问分析和本次分析三种权重下“可解释部分”和“不可解释部分”的整体情况列出

```
. oaxaca8 male female, weight(1 0 0.651) notf
(high estimates: male; low estimates: female)

                                    Mean prediction 1 =   2.843339
                                    Mean prediction 2 =   2.538099
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
difference	.3052401	.0232167	13.15	0.000	.2597363	.3507439

Linear decompositions

Total	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
W=1						
explained	-.0189887	.010135	-1.87	0.061	-.0388529	.0008756
unexplained	.3242288	.0216201	15.00	0.000	.2818542	.3666033
W=0						
explained	-.0158977	.0107569	-1.48	0.139	-.0369808	.0051855
unexplained	.3211378	.0217203	14.79	0.000	.2785667	.3637088
W=,651						
explained	-.0179099	.0098544	-1.82	0.069	-.0372242	.0014044
unexplained	.32315	.0214196	15.09	0.000	.2811683	.3651317

图 6. 第一问三种权重下的整体情况

```
. oaxaca8 male female, weight(1 0 0.651) notf
(high estimates: male; low estimates: female)

                                    Mean prediction 1 =   2.843339
                                    Mean prediction 2 =   2.538099
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
difference	.3052401	.0232247	13.14	0.000	.2597204	.3507597

Linear decompositions

Total	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
W=1						
explained	-.0072278	.0107371	-0.67	0.501	-.0282721	.0138165
unexplained	.3124678	.0215062	14.53	0.000	.2703164	.3546193
W=0						
explained	-.00911	.0111683	-0.82	0.415	-.0309995	.0127794
unexplained	.3143501	.0215864	14.56	0.000	.2720415	.3566587
W=,651						
explained	-.0078847	.0103505	-0.76	0.446	-.0281712	.0124019
unexplained	.3131248	.0212668	14.72	0.000	.2714427	.3548068

图 7. 第二问三种权重下的整体情况

可以发现：可解释部分的差异缩小为原来的 2 分之 1 说明增加了此变量后男性和女性的禀赋更加接近，“可解释部分”也即禀赋所造成的收入差异微乎其微。

深入到每一个具体部分来看，发现主要有以下几点区别：

Part Ⅱ Weight = 1 的差异情况

unexplained						
age00	.3914134	.3664907	1.07	0.286	-.3268952	1.109722
msa	-.034189	.058479	-0.58	0.559	-.1488057	.0804277
ctrlcity	-.0213204	.0160226	-1.33	0.183	-.0527242	.0100834
north_central	.0147435	.0204316	0.72	0.471	-.0253017	.0547886
south00	.0232481	.0215538	1.08	0.281	-.0189967	.0654928
west	.0002596	.0127114	0.02	0.984	-.0246542	.0251734
sch_10	.0028878	.0028033	1.03	0.303	-.0026065	.008382
diploma_hs	.0162527	.0254389	0.64	0.523	-.0336066	.066112
ged_hs	.0064541	.0051234	1.26	0.208	-.0035875	.0164958
smcol	.0023309	.0217106	0.11	0.915	-.040221	.0448828
bachelor_col	-.0040524	.017967	-0.23	0.822	-.0392671	.0311624
master_col	.0004805	.0081242	0.06	0.953	-.0154426	.0164037
doctor_col	-.000655	.0017448	-0.38	0.707	-.0040747	.0027647
afqtp89	.0689754	.0479943	1.44	0.151	-.0250916	.1630425
_cons	-.1543613	.3750825	-0.41	0.681	-.8895094	.5807868
Total	.3124678	.0215062	14.53	0.000	.2703164	.3546193

图 8. Weight = 1 时不可解释部分每一个自变量的影响差异

与图2.对比可以明显看到年龄(age00)所造成的差异降低。

Part Ⅲ Weight = 0 的差异情况

unexplained						
age00	.3887531	.3639999	1.07	0.286	-.3246735	1.10218
msa	-.0334343	.0571886	-0.58	0.559	-.1455219	.0786532
ctrlcity	-.0211405	.0158864	-1.33	0.183	-.0522773	.0099963
north_central	.0157326	.0218	0.72	0.470	-.0269946	.0584598
south00	.0216895	.0201108	1.08	0.281	-.0177269	.061106
west	.00025	.012243	0.02	0.984	-.0237458	.0242459
sch_10	.00698	.0066518	1.05	0.294	-.0060573	.0200174
diploma_hs	.0163652	.0256142	0.64	0.523	-.0338377	.0665681
ged_hs	.0068885	.0054597	1.26	0.207	-.0038123	.0175893
smcol	.0019928	.0185647	0.11	0.915	-.0343932	.0383789
bachelor_col	-.0040357	.0178927	-0.23	0.822	-.0391048	.0310334
master_col	.0003756	.0063558	0.06	0.953	-.0120817	.0128328
doctor_col	-.0018204	.0047276	-0.39	0.700	-.0110864	.0074456
afqtp89	.070115	.0487893	1.44	0.151	-.0255103	.1657402
_cons	-.1543613	.3750825	-0.41	0.681	-.8895094	.5807868
Total	.3143501	.0215864	14.56	0.000	.2720415	.3566587

图 9. Weight = 0 时不可解释部分每一个自变量的影响差异

与图3.对比可以得到和上述相同的结论。

综上分析发现，当引入了控制变量afqtp89（用来衡量个人的综合素质）时，oaxaca8 分解结果有以下两个显著变化：

- “可解释部分”的所带来的收入差异趋近于 0
- “不可解释部分”所带来的收入差异几乎解释了所有的收入差异，并且部分自变量所带来的收入差异有较大变化。

基于此结果我们认为：劳动力市场中存在较强的性别歧视。