Decompositions

Mind the Gap

When observing a gap in mean outcomes between two groups, we often are interested in how much of the gap can be explained by differences in observable characteristics.

The simplest approach is to estimate a pooled regression of the outcome on the observable characteristics, including an indicator variable for group membership, and interpret the coefficient on the group indicator as the unexplained gap in mean differences.

The Oaxaca-Blinder (O-B) decomposition(s) represent an alternative approach that provides further insight into the relationship between the group indicator and the observable characteristics of the groups.

Oaxaca-Blinder decomposition

As an alternative to a pooled regression with a categorical (or dummy) variable d that splits our dataset into two groups, we can run regressions[1] of the form $y = X\beta + \epsilon$ to estimate the the mean difference between groups, as follows:

$$\bar{y}_0 = \bar{X}_0\beta_0; \text{ group } d = 0$$
$$\bar{y}_1 = \bar{X}_1\beta_1; \text{ group } d = 1$$

Then the mean difference in outcomes is:

$$\begin{aligned}
\bar{y}_1 - \bar{y}_0 &= \bar{X}_1\beta_1 - \bar{X}_0\beta_0 \qquad\qquad (1)\\
&= \left(\bar{X}_1 - \bar{X}_0\right)\beta_1 + \bar{X}_0\left(\beta_1 - \beta_0\right)\\
&= \left(\bar{X}_1 - \bar{X}_0\right)\beta_0 + \bar{X}_1\left(\beta_1 - \beta_0\right)
\end{aligned}$$

Where, using $\bar{X}_0$ as the baseline, $\left(\bar{X}_1 - \bar{X}_0\right)\beta_1$ is the "explained" component (since it is based on differences in characteristics), and $\bar{X}_0\left(\beta_1 - \beta_0\right))$ is the "unexplained" component (based on differences in returns to characteristics). This decomposition is not unique, as you can see from the second and third lines of equation (1).

---

[1]To keep the notation simple, regression coefficients β will contain an intercept, i.e. β is always a vector whose first element is the intercept, and the data X is a design matrix, whose first column is all ones.

We define:

$$\text{Gap}^0 = \bar{X}_0 \left(\beta_1 - \beta_0 \right); \ \bar{X}_0 \text{ is the baseline}$$

$$\text{Gap}^1 = \bar{X}_1 \left(\beta_0 - \beta_1 \right); \ \bar{X}_1 \text{ is the baseline}$$

$$\text{Gap}^{\text{OLS}} = \delta_d; \text{ where } y = \delta_0 + d\delta_d + X\delta_x + \epsilon$$

where, for Gap^{OLS}, given the strong assumptions that (i) the model is properly specified and that (ii) coefficients are equal across groups, a sensible definition of the population unexplained gap is δ_d.

Using the fact (equation 12) that

$$\frac{\text{cov}\,(d, y)}{\text{var}\,(d)} = \mathbb{E}_y \left[(y) \, | d = 1 \right] - \mathbb{E}_y \left[(y) \, | d = 0 \right]$$

then equation (1) can be written as:

$$\frac{\text{cov}\,(d, y)}{\text{var}(d)} - \frac{\text{cov}\,(x, d)}{\text{var}(d)} \times \frac{\text{cov}\,(x, y | d = 1)}{\text{var}(x | d = 1)} = \overline{X}_0 \left(\beta_1 - \frac{\text{cov}\,(x, y | d = 0)}{\text{var}(x | d = 0)} \right)$$

$$\frac{\text{cov}\,(d, y)}{\text{var}(d)} - \frac{\text{cov}\,(x, d)}{\text{var}(d)} \times \frac{\text{cov}\,(x, y | d = 0)}{\text{var}(x | d = 0)} = \overline{X}_1 \left(\frac{\text{cov}\,(x, y | d = 1)}{\text{var}(x | d = 1)} - \beta_0 \right)$$

We an define another measure, Gap^p, as

$$\bar{y}_1 - \bar{y}_0 = \left(\bar{X}_1 - \bar{X}_0 \right) \hat{\beta}_p + \text{Gap}^p$$

$$\text{Gap}^p = \bar{X}_1 \left(\hat{\beta}_1 - \hat{\beta}_p \right) + \bar{X}_0 \left(\hat{\beta}_p - \hat{\beta}_0 \right)$$

where $\hat{\beta}_p$ is the coefficient from the pooled regression of y on X (i.e. $y = X\beta_p + \epsilon$). We consider separately the cases where the coefficients are equal across groups and where the coefficients are unequal across groups.

The Case in Which Coefficients Are Equal across Groups

Here we assume that the mean outcomes between groups 0 and 1 differ only by a constant and that the outcome is influenced by only the observable characteristic(s) x.

By definition, an O-B unexplained gap can always be written as the difference in overall mean outcomes minus the difference in predicted mean outcomes, and both of these differences can be denoted by linear projections.

A general expression for an O-B unexplained gap is:

$$\text{Gap} = [\bar{y}_1 - \bar{y}_0] - \left[\hat{\beta} \left(\bar{x}_1 - \bar{x}_0 \right) \right]$$

$$= \text{b}(y | d) - \text{b}(\hat{\beta} x | d) \tag{2}$$

where $\hat{\beta}$ is some coefficient computed from sample data and $b(z|w)$ is the slope from a regression of z on w and an intercept. The choice of $\hat{\beta}$ is what identifies the different O-B decompositions.

So equation (2) resolves to Gap^1 when $\hat{\beta}$ is the slope coefficient of a regression of y on x using the data from group 1, and equation (2) gives Gap^0 when $\hat{\beta}$ is the slope coefficient using data from group 0[2].

$$
\begin{aligned}
\text{Gap} &= b(y|d) - b(\hat{\beta}x|d) \\
&= \frac{\text{cov}\,(d,y)}{\text{var}(d)} - \hat{\beta}\frac{\text{cov}\,(x,d)}{\text{var}(d)} \quad &(3) \\
&= \frac{\text{cov}\,(d, \beta_0 + \beta_d d + \beta_x x + \epsilon)}{\text{var}(d)} - \hat{\beta}\frac{\text{cov}\,(d,x)}{\text{var}(d)} \\
&= \beta_d + \beta_x \frac{\text{cov}\,(d,x)}{\text{var}(d)} - \hat{\beta}\frac{\text{cov}\,(d,x)}{\text{var}(d)} \quad &(4) \\
&\to \beta_d, \text{ when } \beta_x = \hat{\beta}
\end{aligned}
$$

Given our assumption that the mean outcomes between groups 0 and 1 differ only by a constant and that the outcome is influenced by only the observable characteristic(s) x (and so $\hat{\beta} = \beta_x$ in both group-specific regressions), then all the gap expressions are equivalent, except for Gap^p.

Gap^p is different because in the pooled regression used to compute Gap^p, d is a omitted variable. In particular:

$$
\begin{aligned}
\text{Gap}^p &= b(y|d) - b(x \times b(y|x)|d) \\
&= \frac{\text{cov}\,(d,y)}{\text{var}(d)} - \frac{\text{cov}\,(d, x \times [\text{cov}\,(x,y)\,/\text{var}\,(x)])}{\text{var}(d)} \\
&= \frac{\text{cov}\,(d,y)}{\text{var}(d)} - \frac{\text{cov}\,(d,x)}{\text{var}(d)} \times \frac{\text{cov}\,(x,y)}{\text{var}(x)} \\
&= \frac{1}{\text{var}(d)}\left(\text{cov}\,(d,y) - \frac{\text{cov}\,(d,x)\,\text{cov}\,(x,y)}{\text{var}(x)}\right) \quad &(5)
\end{aligned}
$$

Compare this to Gap^{OLS}, defining $\tilde{z}(w)$ as the component of z that is orthogonal to w in the population (so that $\tilde{z}(w) = z - w b(z|w)$, alternatively $\tilde{z}(w) = z - w\frac{\text{cov}(z,w)}{\text{var}(w)}$)[3]

---

[2]under the data generation process described by $y = \beta_0 + \beta_d d + \beta_x x + \epsilon$

[3]Note: $\tilde{d}(x) = d - P(x) = d - x\frac{\text{cov}(x,d)}{\text{var}(x)}$, the linear propensity, and $\tilde{y}(x) = y - x\frac{\text{cov}(y,x)}{\text{var}(x)}$, so the coefficient δ_d is $\frac{\text{cov}(\tilde{d}(x),\tilde{y}(x))}{\text{var}(\tilde{d}(x))}$.

$$\text{Gap}^{\text{OLS}} = \delta_d$$

$$= \frac{\text{cov}(\tilde{d}(x), \tilde{y}(x))}{\text{var}(\tilde{d}(x))}; \text{ per FWL}$$

$$= \frac{\text{cov}(d, \tilde{y}(x))}{\text{var}(\tilde{d}(x))} - \frac{\text{cov}(x, \tilde{y}(x))}{\text{var}(\tilde{d}(x))} \times \frac{\text{cov}(d, x)}{\text{var}(x)}; \text{ per definition of } \tilde{d}(x)$$

$$= \frac{\text{cov}(d, \tilde{y}(x))}{\text{var}(\tilde{d}(x))}; \text{ per definition of } \tilde{y}(x), \text{cov}(x, \tilde{y}(x)) = 0$$

$$= \frac{1}{\text{var}(\tilde{d}(x))} \left(\text{cov}(d, y) - \text{cov}(d, x) \frac{\text{cov}(x, y)}{\text{var}(x)} \right) \tag{6}$$

and so $\text{Gap}^p = \frac{\text{var}(\tilde{d}(x))}{\text{var}(d)} \text{Gap}^{\text{OLS}}$

The Case in Which Coefficients Vary across Groups

The relationship between Gap^p and Gap^{OLS} is exact and general.

Here we assume that the coefficient on x varies between the two groups, and that x is a scalar variable (i.e. a single observable characteristic, not a vector of observable characteristics):

$$y = \delta_0 + d\delta_d + \delta_x x + \delta_{dx} dx + \epsilon$$

Based on equation (4) we have

$$\text{Gap}^1 = \frac{\text{cov}(d, y)}{\text{var}(d)} - \frac{\text{cov}(x, d)}{\text{var}(d)} \times \frac{\text{cov}(x, y|d=1)}{\text{var}(x|d=1)} \tag{7}$$

and

$$\text{Gap}^0 = \frac{\text{cov}(d, y)}{\text{var}(d)} - \frac{\text{cov}(x, d)}{\text{var}(d)} \times \frac{\text{cov}(x, y|d=0)}{\text{var}(x|d=0)} \tag{8}$$

and finally,

$$\text{Gap}^{\text{OLS}} = w_1 \text{Gap}^1 + w_0 \text{Gap}^0 \tag{9}$$

where the weights are:

$$w_1 = \frac{\mathbb{P}(d=1) \text{var}(x|d=1)}{\mathbb{P}(d=1) \text{var}(x|d=1) + \mathbb{P}(d=0) \text{var}(x|d=0)} \tag{10}$$

$$w_0 = \frac{\mathbb{P}(d=0) \text{var}(x|d=0)}{\mathbb{P}(d=1) \text{var}(x|d=1) + \mathbb{P}(d=0) \text{var}(x|d=0)} \tag{11}$$

Now writing $\text{Gap}^{\text{OLS}} = w_1\text{Gap}^1 + w_0\text{Gap}^0$ with $w_1 + w_0 = 1$ (equation (9)):

$$
\begin{aligned}
w_1\text{Gap}^1 + w_0\text{Gap}^0 &= w_1\left[\frac{\text{cov}(d,y)}{\text{var}(d)} - \frac{\text{cov}(x,d)}{\text{var}(d)} \times \frac{\text{cov}(x,y|d=1)}{\text{var}(x|d=1)}\right] \\
&+ w_0\left[\frac{\text{cov}(d,y)}{\text{var}(d)} - \frac{\text{cov}(x,d)}{\text{var}(d)} \times \frac{\text{cov}(x,y|d=0)}{\text{var}(x|d=0)}\right] \\
&= \frac{\text{cov}(d,y)}{\text{var}(d)} - \frac{\text{cov}(x,d)}{\text{var}(d)}\Pi
\end{aligned}
$$

where

$$
\begin{aligned}
\Pi &= \frac{\mathbb{P}(d=1)\,\text{var}(x|d=1)}{\mathbb{P}(d=1)\,\text{var}(x|d=1) + \mathbb{P}(d=0)\,\text{var}(x|d=0)}\frac{\text{cov}(x,y|d=1)}{\text{var}(x|d=1)} \\
&+ \frac{\mathbb{P}(d=0)\,\text{var}(x|d=0)}{\mathbb{P}(d=1)\,\text{var}(x|d=1) + \mathbb{P}(d=0)\,\text{var}(x|d=0)}\frac{\text{cov}(x,y|d=0)}{\text{var}(x|d=0)}
\end{aligned}
$$

simplifying Π:

$$
\begin{aligned}
\Pi &= \frac{\pi\,\text{var}(x|d=1)}{\pi\,\text{var}(x|d=1) + (1-\pi)\,\text{var}(x|d=0)}\frac{\text{cov}(x,y|d=1)}{\text{var}(x|d=1)} \\
&+ \frac{(1-\pi)\,\text{var}(x|d=0)}{\pi\,\text{var}(x|d=1) + (1-\pi)\,\text{var}(x|d=0)}\frac{\text{cov}(x,y|d=0)}{\text{var}(x|d=0)} \\
&= \frac{\pi\,\text{cov}(x,y|d=1) + (1-\pi)\,\text{cov}(x,y|d=0)}{\pi\,\text{var}(x|d=1) + (1-\pi)\,\text{var}(x|d=0)} \\
&= \frac{\text{cov}(x,y) - \frac{\text{cov}(d,x)\text{cov}(d,y)}{\text{var}(d)}}{\text{var}(x) - \frac{\text{cov}(d,x)^2}{\text{var}(d)}}
\end{aligned}
$$

and so

$$
\begin{aligned}
w_1\text{Gap}^1 + w_0\text{Gap}^0 &= \frac{\text{cov}(d,y)}{\text{var}(d)} - \frac{\text{cov}(x,d)}{\text{var}(d)}\Pi \\
&= \frac{\text{cov}(d,y)}{\text{var}(d)} - \frac{\text{cov}(x,d)}{\text{var}(d)}\left(\frac{\text{cov}(x,y) - \frac{\text{cov}(d,x)\text{cov}(d,y)}{\text{var}(d)}}{\text{var}(x) - \frac{\text{cov}(d,x)^2}{\text{var}(d)}}\right) \\
&= \frac{\text{cov}(d,y)\,\text{var}(x) - \text{cov}(d,x)\,\text{cov}(x,y)}{\text{var}(x)\text{var}(d) - \text{cov}(d,x)^2}
\end{aligned}
$$

Now we have

$$
\text{Gap}^{\text{OLS}} = \frac{1}{\text{var}(\tilde{d}(x))}\left(\text{cov}(d,y) - \text{cov}(d,x)\frac{\text{cov}(x,y)}{\text{var}(x)}\right)
$$

5

and since $\tilde{d}(x)$ represents the residuals from a population regression of d on x,

$$\text{var}\left(\tilde{d}(x)\right) = \text{var}\left(d - x\frac{\text{cov}(d,x)}{\text{var}(x)}\right)$$

$$= \text{var}(d) - \frac{\text{cov}(d,x)^2}{\text{var}(x)}$$

and so

$$\text{Gap}^{\text{OLS}} = \frac{1}{\text{var}(\tilde{d}(x))}\left(\text{cov}(d,y) - \text{cov}(d,x)\frac{\text{cov}(x,y)}{\text{var}(x)}\right)$$

$$= \frac{\text{var}(x)}{\text{var}(x)\text{var}(d) - \text{cov}(d,x)^2}\left(\text{cov}(d,y) - \text{cov}(d,x)\frac{\text{cov}(x,y)}{\text{var}(x)}\right)$$

$$= \frac{\text{cov}(d,y)\text{var}(x) - \text{cov}(d,x)\text{cov}(x,y)}{\text{var}(x)\text{var}(d) - \text{cov}(d,x)^2}$$

$$= w_1\text{Gap}^1 + w_0\text{Gap}^0$$

Appendix

Regression on binary variable: coefficient

Recall that for a continuous variable y and a binary variable d (with the proportion of $d = 1$ in the data $= p$):

$$\mathbb{E}[d] = p$$

$$\text{var}(d) = p\left(1 - \mathbb{E}[d]\right)^2 + (1-p)\left(-\mathbb{E}[d]\right)^2$$

$$= p(1-p)^2 + (1-p)p^2$$

$$= p(1-p)$$

$$\text{covar}(d,y) = \mathbb{E}_{y,d}\left[(y-\bar{y})\left(d-\bar{d}\right)\right]$$

$$= \mathbb{E}_d\left[\mathbb{E}_y\left[(y-\bar{y})|d\right], \left(d-\bar{d}\right)\right]$$

$$= \mathbb{E}_y\left[(y-\bar{y})|d=1\right] \times p(1-p)$$

$$- \mathbb{E}_y\left[(y-\bar{y})|d=0\right] \times (1-p)p$$

Then the coefficient β_d from the regression $y \sim \beta_d d$ is

$$\beta_d = \frac{\text{cov}(d,y)}{\text{var}(d)} = \mathbb{E}_y\left[(y)|d=1\right] - \mathbb{E}_y\left[(y)|d=0\right] \tag{12}$$

6

Misc

The law of total variance, or variance decomposition formula is:

$$\text{var}(X) = \mathbb{E}[\text{var}(X|Y)] + \text{var}(\mathbb{E}[X|Y])$$

where the first term is the between-group variance and the second term is the within-group variance. In our problem, the corresponding variance decomposition is

$$\text{var}(x) = (1 - \pi)\left(\text{var}(x|d = 0) + (\mathbb{E}[x|d = 0] - \mathbb{E}[x])^2\right)$$
$$+ \pi\left(\text{var}(x|d = 1) + (\mathbb{E}[x|d = 1] - \mathbb{E}[x])^2\right)$$

but we have

$$\mathbb{E}[x] = \mathbb{E}[\mathbb{E}[x|y]] = \pi\mathbb{E}[x|d = 1] + (1 - \pi)\mathbb{E}[x|d = 0]$$

substituting the RHS for $\mathbb{E}[x]$ we have

$$\text{var}(x) = (1 - \pi)\left(\text{var}(x|d = 0) + (\mathbb{E}[x|d = 1] - \mathbb{E}[x|d = 0])^2 \pi^2\right)$$
$$+ \pi\left(\text{var}(x|d = 1) + (\mathbb{E}[x|d = 1] - \mathbb{E}[x|d = 0])^2 (1 - \pi)^2\right)$$

but we have $\mathbb{E}[x|d = 1] - \mathbb{E}[x|d = 0] = \text{cov}(x, d)/\text{var}(d)$, and $\text{var}(d) = \pi(1 - \pi)$, so

$$\text{var}(x) = (1 - \pi)\,\text{var}(x|d = 0) + \pi\text{var}(x|d = 1) + \frac{\text{cov}(d, x)^2}{\text{var}(d)^2} \times [(1 - \pi)\pi^2 + \pi(1 - \pi)^2]$$

$$= (1 - \pi)\,\text{var}(x|d = 0) + \pi\text{var}(x|d = 1) + \frac{\text{cov}(d, x)^2}{\text{var}(d)^2} \times [(1 - \pi)\text{var}(d) + \pi\text{var}(d)]$$

$$= (1 - \pi)\,\text{var}(x|d = 0) + \pi\text{var}(x|d = 1) + \frac{\text{cov}(d, x)^2}{\text{var}(d)}$$

by the same logic

$$\text{cov}(x, y) = (1 - \pi)\,\text{cov}(x, y|d = 0) + \pi\text{cov}(x, y|d = 1) + \frac{\text{cov}(d, x)\text{cov}(d, y)}{\text{var}(d)}$$

OLS Weights

With $\rho \equiv \mathbb{E}[d] = \mathbb{E}\left[\mathbb{E}[d|X]\right]$ and $\mathrm{P}(X) = \alpha_p + X\beta_p$ is the propensity score ($\beta_p = \frac{\mathrm{cov}(d,X)}{\mathrm{var}(X)}$). We derive the expression for the derivatives of w_0 and w_1 wrt the intercept of the propensity model α_p, where a shift in the intercept changes ρ by a small amount.

Define $a_0 = \pi\mathrm{var}\left(\mathrm{P}(x)|d=1\right)$ and $a_1 = (1-\pi)\mathrm{var}\left(\mathrm{P}(x)|d=0\right)$, such that $w_0 = \frac{a_0}{a_0+a_1}$ and $w_1 = \frac{a_1}{a_0+a_1}$ (NOTE: this is different from the previous definition)

Referencing the weights, per equations (11 and 10)

$$
\begin{aligned}
a_0 \;&= \pi\mathrm{var}\left(\mathrm{P}(x)|d=1\right) && \text{by definition} \\
&= \mathbb{E}\left[d\right]\mathbb{E}\left[\left(\mathrm{P}(x)-\mathbb{E}\left[\mathrm{P}(x)|d=1\right]\right)^2|d=1\right] && \text{by definition of var} \\
&= \mathbb{E}\left[d\right]\left(\frac{\mathbb{E}[\mathrm{P}(x)^2 d]}{\mathbb{E}[d]} - \left(\frac{\mathbb{E}[\mathrm{P}(x)d]}{\mathbb{E}[d]}\right)^2\right) && \text{since } \mathbb{P}\left(A.B\right) = \mathbb{P}\left(A|B\right)\mathbb{P}\left(B\right) \\
&= \mathbb{E}\left[\mathrm{P}(x)^2 d\right] - \frac{\mathbb{E}[\mathrm{P}(x)d]^2}{\mathbb{E}[d]} \\
&= \mathbb{E}\left[\mathrm{P}(x)^2\mathbb{E}\left[d|X\right]\right] - \frac{\mathbb{E}[\mathrm{P}(x)\mathbb{E}[d|X]]^2}{\mathbb{E}[\mathbb{E}[d|X]]} && \text{by definition of } \mathrm{P}(x) \\
&= \mathbb{E}\left[\mathrm{P}(x)^3\right] - \frac{\mathbb{E}[\mathrm{P}(x)^2]^2}{\mathbb{E}[\mathrm{P}(x)]}
\end{aligned}
$$

using the definition of the linear propensity ($\mathrm{P}(X) = \alpha_p + X\beta_p$), then

$$
\begin{aligned}
\frac{da_0}{d\alpha_p} &= 3\mathbb{E}\left[\mathrm{P}(x)^2\right] - 4\frac{\mathbb{E}\left[\mathrm{P}(x)^2\right]\mathbb{E}\left[\mathrm{P}(x)\right]}{\mathbb{E}\left[\mathrm{P}(x)\right]} + \frac{\left(\mathbb{E}\left[\mathrm{P}(x)^2\right]\right)^2}{\mathbb{E}\left[\mathrm{P}(x)\right]^2} \\[2mm]
&= -\mathbb{E}\left[\mathrm{P}(x)^2\right] + \frac{\left(\mathbb{E}\left[\mathrm{P}(x)^2\right]\right)^2}{\mathbb{E}\left[\mathrm{P}(x)\right]^2} \\[2mm]
&= \frac{\left(\mathbb{E}\left[\mathrm{P}(x)^2\right]\right)^2 - \mathbb{E}\left[\mathrm{P}(x)^2\right]\mathbb{E}\left[\mathrm{P}(x)\right]^2}{\mathbb{E}\left[\mathrm{P}(x)\right]^2} \\[2mm]
&= \frac{\mathbb{E}\left[\mathrm{P}(x)^2\right]\left(\mathbb{E}\left[\mathrm{P}(x)^2\right] - \mathbb{E}\left[\mathrm{P}(x)\right]^2\right)}{\mathbb{E}\left[\mathrm{P}(x)\right]^2} \\[2mm]
&= \frac{\mathbb{E}\left[\mathrm{P}(x)^2\right]\mathrm{var}\left(\mathrm{P}(x)\right)}{\mathbb{E}\left[\mathrm{P}(x)\right]^2} > 0
\end{aligned}
$$

Gap weights

Example

data:

```
# https://cran.r-project.org/web/packages/oaxaca/vignettes/oaxaca.pdf
# https://www.worldbank.org/content/dam/Worldbank/document/HDN/Health/HealthEquityCh12.pdf
# https://pmc.ncbi.nlm.nih.gov/articles/PMC8343972/
# https://www.sciencedirect.com/science/article/abs/pii/S0169721811004072
# https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2528391
# https://link.springer.com/article/10.1186/s12982-021-00100-9
# https://giacomovagni.com/blog/2023/oaxaca/
# https://journals.sagepub.com/doi/pdf/10.1177/1536867X0800800401
# https://ocw.mit.edu/courses/14-662-labor-economics-ii-spring-2015/resources/mit14_662s15_lecnotes1/
```

```r
# Load HMDA data
data('CPS1985', package = "AER") # data('PSID1982', package = "AER")
hmda <- CPS1985 |> tibble::as_tibble()

# data("nswcps", package = "hettreatreg")

# Prepare data for analysis
# Looking at income differences between racial groups
lending_data <- hmda |>
  dplyr::mutate(
    minority = factor(ethnicity != "cauc"), # minority = factor(race != "white"),
    log_income = log(wage) #log(income)
  ) |>
  dplyr::select(wage, log_income, minority, education, experience, union)
  # dplyr::select(log_income, minority, education, hrat, ccred, mcred, pubrec)
```

```r
# Fit separate regressions
model_a <-
  lm(wage ~ education + experience, data = lending_data |> dplyr::filter(union=='yes'))
model_b <-
  lm(wage ~ education + experience, data = lending_data |> dplyr::filter(union=='no'))

# Get mean characteristics (including intercept)
X_mean_a <-
  c(1,
    colMeans(
    lending_data |>
      dplyr::filter(union=='yes') |>
      dplyr::select(education, experience)
    )
  )
X_mean_b <-
  c(1,
    colMeans(
      lending_data |>
        dplyr::filter(union=='no') |>
        dplyr::select(education, experience)
    )
  )

# Get coefficients
beta_a <- coef(model_a)
beta_b <- coef(model_b)

# Calculate decomposition
tibble::tibble(
  explained = sum((X_mean_a - X_mean_b) * beta_a)
  , unexplained = sum(X_mean_b * (beta_a - beta_b))
  , total_gap <- explained + unexplained
)

# Perform Oaxaca-Blinder decomposition
decomp <-
  oaxaca::oaxaca(
    wage ~ education + experience | union
    , data =
      lending_data |>
      dplyr::mutate(union = dplyr::case_when(union=='yes'~0, TRUE ~1))
  )
```