Testing random assignment to peer groups

Koen Jochmans

June 7, 2021

Motivation

Inference on peer effects has received considerable attention.

One difficulty is that of self-selection of peer groups.

(Quasi) randomization of peer assignment has proven a fruitful way forward.

- Education: Sacerdote (2001) and Zimmerman (2003).
- Workplace: Bandiera et al. (2009) and Mas and Moretti (2009).

Even given network exogeneity, (accurate) inference on peer effects is known to be challenging.

Develop a test that can be used to verify

- The (conditional) random assignment to peer groups.
- The presence of peer effects in linear-in-means model.

Connections to the literature

The test is a bias-corrected version of an idea introduced in Sacerdote (2001).

Related literature:

Guryan, Kroft and Notowidigdo (2009):

Augmented-regression test and randomization test.

Stevenson (2015, 2017):

Sample-splitting approach.

Caeyers and Fafchamps (2020):

Calculation of probability limit in a simple case.

Test developed here is more general.

Underlying calculations allow to derive theory for Guryan et al. (2009).

Setting

Stratified data on r independent urns of size $n_1, \ldots, n_r$.

Peer assignment in urn g is recorded in the $n_g \times n_g$ adjacency matrix

$$(\boldsymbol{A}_g)_{i,j} := \begin{cases} 1 & \text{if } i \text{ and } j \text{ are peers} \\ 0 & \text{if they are not} \end{cases} .$$

Individuals cannot be their own peer.

Peer groups can be of different sizes and are allowed to overlap; $m_g(i)$ is the number of peers, $m_g(i \cap j)$ is the number of common peers.

In Sacerdote (2001): Freshmen are put into urns based on their response to a set of survey questions (gender, smoker, etc). Then randomly assigned to rooms within each urn.

In Guryan, Kroft and Notowidigdo (2009): PGA golf players are randomly assigned playing partners from the set of participants within the same player category.

Default test

Random assignment implies that observables $x_{g,i}$ and $x_{g,j}$ are uncorrelated within urns for all $j \in [i]$, with

$$[i] := \{j : (\boldsymbol{A}_g)_{i,j} = 1\}$$

the set of i's peers.

A standard test (Sacerdote 2001) is based on within-urn regression of $x_{g,i}$ on

$$\bar{x}_{g,[i]} := m_g(i)^{-1} \sum_{j=1}^{n_g} (\boldsymbol{A}_g)_{i,j}\, x_{g,j},$$

the average characteristic of i's peers.

Test whether the slope coefficient is zero via a (two-sided) t-test.

Under the null this test tends to find negative assortative matching (Guryan et al. 2009).

Bias

The within-urn estimator, $\hat{\rho}$, is defined as

$$\sum_{g=1}^{r} \sum_{i=1}^{n_g} \bar{x}_{g,[i]} \left(\tilde{x}_{g,i} - \hat{\rho}\, \tilde{\bar{x}}_{g,[i]} \right) = 0,$$

where $\tilde{x}_{g,i}$ and $\tilde{\bar{x}}_{g,[i]}$ are deviations from within-urn means.

Impose the urn-level homoskedasticity assumption $\mathbb{E}_0((x_{g,i} - \mathbb{E}_0(x_{g,i}))^2) =: \sigma_g^2$ (for now).

The normal equation has bias

$$\mathbb{E}_0 \left(\sum_{g=1}^{r} \sum_{i=1}^{n_g} \bar{x}_{g,[i]}\, \tilde{x}_{g,i} \right) = -\sum_{g=1}^{r} \sigma_g^2,$$

so that $\hat{\rho}$ is inconsistent under many-urn asymptotics.

Calculation

Without loss of generality, set urn effects to zero. Then

$$\mathbb{E}_0(x_{g,i}\, x_{g,j} | \boldsymbol{A}_g) = \begin{cases} \sigma_g^2 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} .$$

The bias is

$$\mathbb{E}_0\left(\sum_{g=1}^{r}\sum_{i=1}^{n_g} \bar{x}_{g,[i]}\, \tilde{x}_{g,i}\right) = \sum_{g=1}^{r} \mathbb{E}_0\left(\sum_{i=1}^{n_g} \bar{x}_{g,[i]}\, x_{g,i}\right) - \sum_{g=1}^{r} \mathbb{E}_0\left(\sum_{i=1}^{n_g} \bar{x}_{g,[i]}\, \bar{x}_g\right).$$

Here,

$$\begin{aligned} \mathbb{E}_0\left(\sum_{i=1}^{n_g} \bar{x}_{g,[i]}\, x_{g,i}\right) &= \mathbb{E}_0\left(\sum_{i=1}^{n_g}\sum_{j=1}^{n_g} \frac{(\boldsymbol{A}_g)_{i,j}\, x_{g,j}\, x_{g,i}}{m_g(i)}\right) \\ &= \mathbb{E}_0\left(\sum_{i=1}^{n_g}\sum_{j\neq i} \frac{(\boldsymbol{A}_g)_{i,j}\, \mathbb{E}_0(x_{g,j}\, x_{g,i} | \boldsymbol{A}_g)}{m_g(i)}\right) \\ &= 0. \end{aligned}$$

Also,

$$
\begin{aligned}
\mathbb{E}_0\left(\sum_{i=1}^{n_g} \bar{x}_{g,[i]}\,\overline{x}_g\right) &= \mathbb{E}_0\left(\frac{1}{n_g}\sum_{i=1}^{n_g}\sum_{j=1}^{n_g}\sum_{j'=1}^{n_g}\frac{(\boldsymbol{A}_g)_{i,j}\,x_{g,j}\,x_{g,j'}}{m_g(i)}\right) \\
&= \mathbb{E}_0\left(\frac{1}{n_g}\sum_{i=1}^{n_g}\sum_{j=1}^{n_g}\sum_{j'=1}^{n_g}\frac{(\boldsymbol{A}_g)_{i,j}\,\mathbb{E}_0(x_{g,j}\,x_{g,j'}|\boldsymbol{A}_g)}{m_g(i)}\right) \\
&= \mathbb{E}_0\left(\frac{1}{n_g}\sum_{i=1}^{n_g}\sum_{j=1}^{n_g}\frac{(\boldsymbol{A}_g)_{i,j}\,\mathbb{E}_0(x_{g,j}^2|\boldsymbol{A}_g)}{m_g(i)}\right) \\
&= \mathbb{E}_0\left(\frac{1}{n_g}\sum_{i=1}^{n_g}\frac{\sum_{j=1}^{n_g}(\boldsymbol{A}_g)_{i,j}}{m_g(i)}\right)\sigma_g^2 \\
&= \sigma_g^2,
\end{aligned}
$$

from which the result follows.

Probability limit

Can show that, under the null, $\text{plim}_{r\to\infty}\, \hat{\rho}$ equals

$$-\frac{\lim_{r\to\infty} \frac{1}{r}\sum_{g=1}^{r} \sigma_g^2}{\lim_{r\to\infty} \frac{1}{r}\sum_{g=1}^{r} \sigma_g^2\, \mathbb{E}_0 \left(\sum_{i=1}^{n_g} \frac{1}{m_g(i)} - \frac{1}{n_g}\sum_{i=1}^{n_g}\sum_{j=1}^{n_g} \frac{m_g(i\cap j)}{m_g(i)\, m_g(j)}\right)}.$$

When $n_1 = \cdots = n_r =: n$, and $m_g(i) =: m$ and $m_g(i \cap j) = 0$ for all individuals and urns,

$$\text{plim}_{r\to\infty}\, \hat{\rho} = -\frac{m}{n-m},$$

which agrees with a result of Caeyers and Fafchamps (2020) but is obtained under weaker conditions.

Bias adjustment

An unbiased estimator of σ_g^2 (under the null) is

$$\frac{1}{n_g - 1} \sum_{i=1}^{n_g} x_{g,i}\, \tilde{x}_{g,i}.$$

The re-centered covariance

$$q_r^{\mathrm{HO}} := \sum_{g=1}^{r} \sum_{i=1}^{n_g} \tilde{x}_{g,i} \left(\bar{x}_{g,[i]} + \frac{x_{g,i}}{n_g - 1} \right)$$

will be exactly unbiased under random assignment.

Its standard deviation can be estimated by

$$s_r^{\mathrm{HO}} := \sqrt{\sum_{g=1}^{r} \left(\sum_{i=1}^{n_g} \tilde{x}_{g,i} \left(\bar{x}_{g,[i]} + \frac{x_{g,i}}{n_g - 1} \right) \right)^2}.$$

Corrected test

An adjusted test statistic follows as

$$t_r^{\mathrm{HO}} := q_r^{\mathrm{HO}} / s_r^{\mathrm{HO}}.$$

Let $b_r := \mathbb{E}(q_r^{\mathrm{HO}}) = O(\sqrt{r})$ be a non-stochastic sequence of constants.

Let $\mathbb{P}(n_g > 2) = 1$. If $\max_{g,i} \mathbb{E}(x_{g,i}^8) = O(1)$ and $\max_{g,i}(\mathrm{var}(x_{g,i}^2))^{-1} = O(1)$, then

$$t_r^{\mathrm{HO}} - \frac{b_r}{s_r^{\mathrm{HO}}} \xrightarrow{d} N(0,1),$$

as $r \to \infty$.

Implication is that, for any $\alpha \in (0,1)$,

$$\lim_{r\to\infty} \mathbb{P}_0 \left(t_r^{\mathrm{HO}} > z_{1-\alpha} \right) = \alpha,$$

where z_α is the α-quantile of the standard-normal distribution.

Power

The test is consistent against endogenous-, contextual-, and correlated-effect alternatives (Manski 1993).

Endogenous-effect alternatives:

$$x_{g,i} = \rho\, \bar{x}_{g,[i]} + \varepsilon_{g,i}, \qquad \varepsilon_{g,i} \sim \text{independent}\,(\alpha_g, \sigma_g^2),$$

where $-1 < \rho < 1$ and $\varepsilon_{g,i}$ independent of $\boldsymbol{A}_g$.

Local alternative: $\rho = \varrho/\sqrt{r}$.

With $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_r$ i.i.d. , homoskedasticity, and no overlap in peer groups, $t_r^{\mathrm{HO}} \xrightarrow{d} N(\mu, 1)$ where

$$\mu := \varrho \sqrt{2\,\mathbb{E}\left(\sum_{i=1}^{n_g} \frac{1}{m_g(i)} - \frac{n_g}{n_g - 1}\right)} > 0.$$

Locally asymptotically equivalent to contextual-effect alternatives:

$$x_{g,i} = \varepsilon_{g,i} + \frac{\theta}{m_g(i)} \sum_{j=1}^{n_g} (\boldsymbol{A}_g)_{i,j}\, \varepsilon_{g,j}$$

for $\theta = \vartheta/\sqrt{r}$.

That is, non-centrality parameter is the same:

$$\mu := \vartheta \sqrt{2\,\mathbb{E}\left(\sum_{i=1}^{n_g} \frac{1}{m_g(i)} - \frac{n_g}{n_g - 1}\right)} > 0.$$

Not a surprising finding in light of the time-series literature on testing against autoregressive alternatives and moving-average alternatives (e.g., Godfrey 1981).

Correlated-effect alternatives:

$$\mathbb{E}(x_{g,i}\, x_{g,i'} | (\boldsymbol{A}_g)_{i,i'} = 1) = \sigma_\eta^2$$

and

$$\mathbb{E}(x_{g,i}\, x_{g,i'} | (\boldsymbol{A}_g)_{i,i'} = 0) = \begin{cases} \sigma_\eta^2 + \sigma_g^2 & \text{if } i = i' \\ 0 & \text{if } i \neq i' \end{cases}$$

for $\sigma_\eta^2 > 0$.

Here, local alternatives have $\sigma_\eta^2 = \varsigma^2 / \sqrt{r}$, and

$$\mu = \frac{\varsigma^2}{\sigma^2} \frac{\mathbb{E}\left((n_g - 1) - \frac{1}{n_g} \sum_{i=1}^{n_g} \frac{m_g(i)}{n_g - 1} \right)}{\sqrt{2\, \mathbb{E}\left(\sum_{i=1}^{n_g} \frac{1}{m_g(i)} - \frac{n_g}{n_g - 1} \right)}} > 0$$

depends on the ratio ς^2/σ^2.

An alternative test

Guryan et al. (2009): 'default test fails because i cannot be his own peer.'

Informal adjustment to the default test is to include the leave-own-out urn average

$$\frac{1}{n_g - 1} \sum_{j \neq i} x_{g,j} = \frac{n_g}{n_g - 1} \left(\overline{x}_g - \frac{x_{g,i}}{n_g} \right)$$

as a control variable in the within-urn regression.

Because of the presence of fixed effects, this is equivalent to including $x_{g,i}/(n_g - 1)$ as additional control variable.

Requires variation in urn sizes to prevent a perfect fit (that satisfies the null).

By now a commonly-used test.

No theory available for this approach.

Can show that this approach tests whether

$$\sum_{g=1}^{r}\sum_{i=1}^{n_g}\tilde{x}_{g,i}\left(\bar{x}_{g,[i]}+\frac{x_{g,i}}{n_g-1}\right)\left(1-\frac{\delta}{n_g-1}\right)+o_p(\sqrt{r}),$$

is statistically different from zero. Here,

$$\delta := \frac{\lim_{r\to\infty}\frac{1}{r}\sum_{g=1}^{r}\sigma_g^2}{\lim_{r\to\infty}\frac{1}{r}\sum_{g=1}^{r}\sigma_g^2\ \mathbb{E}_0\left(\frac{1}{n_g-1}\right)},$$

is the probability limit of the slope coefficient of a within-group regression of $x_{g,i}$ on $x_{g,i}/(n_g-1)$, under the null.

This finding can be used to confirm that the test is size correct but also to show that it will often have low power.

This formalizes discussions in Stevenson (2015, 2017) and Caeyers and Fafchamps (2020).

Illustrations

Urns of two different sizes $n_g \in \{\overline{n}_1, \overline{n}_2\}$, with $\overline{n}_1 < \overline{n}_2$ and $p_n := \mathbb{P}(n_g = \overline{n}_2)$.

Non-centrality parameter of the Guryan et al.(2009) statistic is

$$\mu^* := \sqrt{p_n(1-p_n)} \frac{b(\bar{n}_2) - b(\bar{n}_1)}{\sqrt{v(\bar{n}_1)\, p_n + v(\bar{n}_2)\,(1-p_n)}},$$

where $b(n)$ and $v(n)$ are the bias and variance of

$$\sum_{i=1}^{n_g} \tilde{x}_{g,i} \left(\bar{x}_{g,[i]} + x_{g,i}/(n_g - 1)\right)$$

conditional on $n_g = n$.

Clearly, $\mu^* \to 0$ as $p_n(1-p_n) \to 0$.

Also low power when $b(\bar{n}_2) - b(\bar{n}_1)$ is small.

Here, bias from different urn sizes cancel out. Such situations can easily be constructed.

Illustrate this graphically for settings where $r = 25$ and

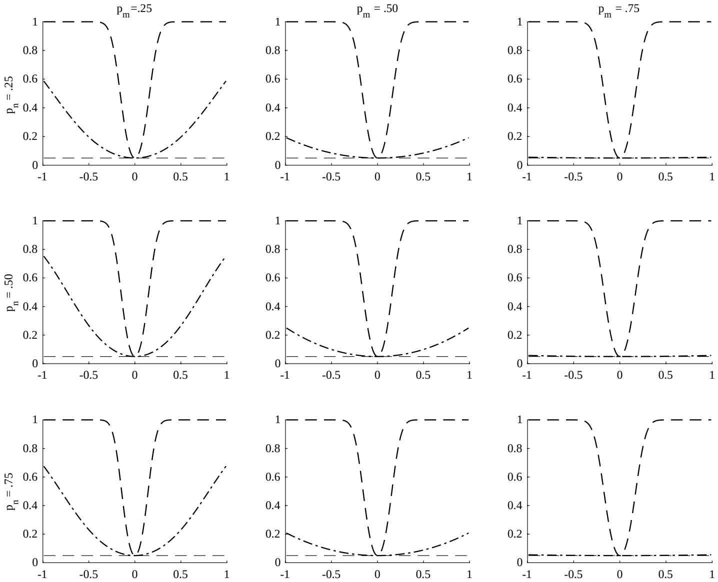
$\overline{n}_1 = 4$ and $\overline{n}_2 = 6$.

Non-overlapping peer groups, with

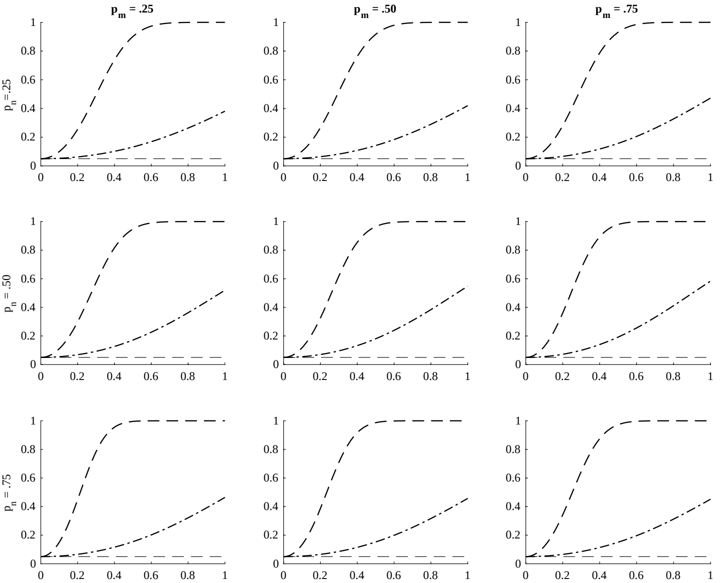
$m_g(i) = 1$ when $n_g = 4$

$m_g(i) = 2$ with $p_m := \mathbb{P}(m_g(i) = 2)$, $m_g(i) = 1$ with $(1 - p_m)$ when $n_g = 6$.

Endogenous-effect alternatives



Correlated-effect alternatives

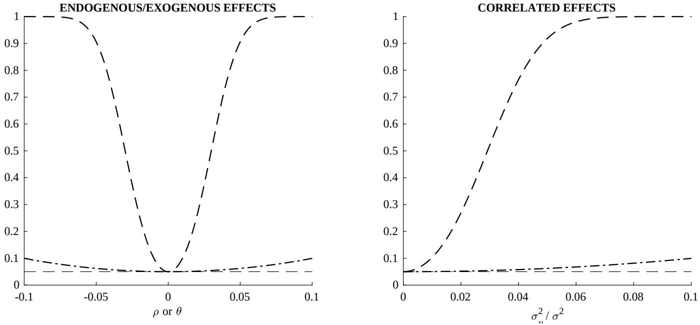


Alternative design (Guryan et al. 2009; Stevenson 2015)

$r = 100.$

$n_g \in \{39, 42, 45, 48, 51\}.$

$m_g(i) = 2$, no overlap.



Heteroskedasticity

Now let $\sigma^2_{g,i} := \mathbb{E}_0((x_{g,i} - \mathbb{E}_0(x_{g,i}))^2)$.

We have

$$\mathbb{E}_0\left(\sum_{g=1}^{r}\sum_{i=1}^{n_g} \bar{x}_{g,[i]}\,\tilde{x}_{g,i}\right) = -\sum_{g=1}^{r}\mathbb{E}_0\left(\frac{1}{n_g}\sum_{i=1}^{n_g}\frac{1}{m_g(i)}\sum_{j=1}^{n_g}(\boldsymbol{A}_g)_{i,j}\,\sigma^2_{g,j}\right).$$

An unbiased estimator of this bias is

$$-\sum_{g=1}^{r}\sum_{i=1}^{n_g}\omega_{g,i}\,x_{g,i}\,\tilde{x}_{g,i}, \qquad \omega_{g,i} := \frac{1}{n_g-2}\left(\sum_{i'\in[i]}\frac{1}{m_g(i')} - \frac{1}{n_g-1}\right).$$

This builds on Hartley, Rao and Kiefer (1969).

When peer groups do not overlap $m_g(i') = m_g(i)$ for all $i' \in [i]$, and so

$$\omega_{g,i} = \frac{1}{n_g-1}.$$

Consequently, t_r^{HO} is robust to heteroskedasticity in this case.

More generally,

$$q_r^{\mathrm{HC}} := \sum_{g=1}^{r} \sum_{i=1}^{n_g} \tilde{x}_{g,i} \; \left(\bar{x}_{g,[i]} + \omega_{g,i}\, x_{g,i}\right)$$

is exactly unbiased.

A heteroskedasticity-robust test statistic is

$$t_r^{\mathrm{HC}} := q_r^{\mathrm{HC}} / s_r^{\mathrm{HC}},$$

where

$$s_r^{\mathrm{HC}} := \sqrt{\sum_{g=1}^{r} \left(\sum_{i=1}^{n_g} \tilde{x}_{g,i} \; \left(\bar{x}_{g,[i]} + \omega_{g,i}\, x_{g,i}\right) \right)^2}.$$

Controlling for covariates

The approach can be modified to a setting where assignment to peer groups is random only conditional on a further set of covariates, $\boldsymbol{w}_{g,i}$, say, by appealing to the Frisch-Waugh-Lovell theorem.

Let $\dot{x}_{g,i}$ be the residual from a within-urn regression of $x_{g,i}$ on $\boldsymbol{w}_{g,i}$.

Then the test statistic is

$$\hat{t}_r^{\mathrm{HO}} := \hat{q}_r^{\mathrm{HO}} / \hat{s}_r^{\mathrm{HO}},$$

where

$$\hat{q}_r^{\mathrm{HO}} := \sum_{g=1}^{r} \sum_{i=1}^{n_g} \dot{x}_{g,i} \left(\bar{x}_{g,[i]} + \frac{x_{g,i}}{n_g - 1} \right)$$

and

$$\hat{s}_r^{\mathrm{HO}} := \sqrt{\sum_{g=1}^{r} \left(\sum_{i=1}^{n_g} \dot{x}_{g,i} \left(\bar{x}_{g,[i]} + \frac{x_{g,i}}{n_g - 1} \right) \right)^2}.$$

The heteroskedasticity-robust statistic t_r^{HC} can be modified in the same way.

Empirical example: PGA tournaments

Participants to PGA tournaments get randomly assigned playing partners from the same 'player category' (1, 1a, 2 or 3).

Marginal on 'player category' assignment is not random.

Data from Guryan et al. (2009), spanning 3 seasons (2002, 2005, 2006) and covering 81 tournaments.

Ability measure used is golfer's 'handicap' (centered around 72).

PGA Tour data

	n obs	mean	std	min	max
			ability ($x_{g,i}$)		
cat 1	3,205	-3.138	0.769	-5.159	1.440
cat 1a	3,436	-2.808	0.740	-4.326	6.732
cat 2	1,503	-2.857	0.894	-4.776	3.275
cat 3	657	-1.662	1.470	-4.776	6.315
			peer ability ($\bar{x}_{g,[i]}$)		
cat 1	3,205	-3.132	0.599	-5.081	0.672
cat 1a	3,436	-2.811	0.591	-4.530	3.275
cat 2	1,503	-2.850	0.744	-4.776	3.275
cat 3	657	-1.690	1.270	-4.776	6.315
			urn size (n_g)		
tourn by cat	8,791	39.292	16.869	3	83
			weight ($(n_g - 1)^{-1}$)		
tourn by cat	8,791	0.037	0.040	0.012	0.500

Stata: rassign.do

```
. rassign handicap hand_i if round==1, group(tourncat)

-----------------------------------------------------------------------------------------------
> ----
Test for (conditional) random assignment to peer groups (or absence of conditional correlation
> ):
-----------------------------------------------------------------------------------------------
> ----

T-statistic: -.85202348      (reference distribution is standard normal)

P-values left-sided:  0.1971     two-sided:  0.3942     right-sided:  0.8029
-----------------------------------------------------------------------------------------------
> ----
The null is absence of correlation.

The grouping variable is tourncat. There are 300 groups.
The smallest group is of size 3 while the largest is of size 83
```

```
. areg handicap hand_i mean_handicap if round==1, vce(cluster grouping_id) absorb(tourncat)

Linear regression, absorbing indicators         Number of obs     =      8,801
Absorbed variable: tourncat                     No. of categories =        305
                                                F(   2,    3227)  =      24.79
                                                Prob > F          =     0.0000
                                                R-squared         =     0.5415
                                                Adj R-squared     =     0.5250
                                                Root MSE          =     0.6403

                          (Std. Err. adjusted for 3,228 clusters in grouping_id)
-------------------------------------------------------------------------------
              |               Robust
     handicap |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+----------------------------------------------------------------
       hand_i |  -.0175858   .0145419    -1.21   0.227    -.0460981    .0109264
mean_handicap |  -10.80304     1.6287    -6.63   0.000    -13.99643   -7.609652
        _cons |  -28.68232   3.879736    -7.39   0.000    -36.28931   -21.07532
-------------------------------------------------------------------------------
```