

How to measure benefits of social policies using R?

An example: Social housing programs ¹

Resul Akay
Vienna < – R
Meeting

May 14, 2019

¹This presentation is based on [1]

Agenda

- 1 What do I try to answer and why?
- 2 Microeconomics of housing
- 3 The data
- 4 Estimation methods and results
- 5 Conclusion
- 6 Questions

What do I try to answer and why?

- Background story.
- What are the benefits of social housing for tenants?
- Who benefits more low-income or high-income tenants?
- How are the benefits distributed?
- According to housing authority high-income tenants benefit more. Do they?

What do I try to answer and why?

- Background story.
- What are the benefits of social housing for tenants?
- Who benefits more low-income or high-income tenants?
- How are the benefits distributed?
- According to housing authority high-income tenants benefit more. Do they?

"In God we trust. All others must bring data." It has been said some time ago by W. Edwards Deming

The microeconomics of housing [1]

- What is a Homogeneous good?
- What is a housing unit? A housing unit is a bundle of a certain quantity of capital assets that generates housing services.
- What is housing service? Housing service is an unobservable quantity of service which is provided by a housing unit per unit of time.
- The housing market.

Characteristics of housing units

Housing is not one complete consumption good; rather, it is a bundle of complex attributes such as the neighborhood of the housing units and distance to desired destinations. In the literature, the characteristics of a housing unit are listed as follows:

- **Immobility:** A regular housing units provide a stream of services for accommodation during a period in a specific location.
- **Durability:** Housing units are modifiable and are suitable for long-term use.
- **Heterogeneity:** Housing units differ in many characteristics, in particular, neighborhood, accessibility of public services and most importantly the distance to the desired destinations, and other housing attributes.

Characteristics of housing units: How to overcome this complexity:

- Due to these characteristics of housing units, it is difficult to find a classification of a homogeneous housing good which is the only thing that consumers or households attach value to.
- To overcome this complexity, Muth(1958)[7] developed a competitive theory of the housing market. He introduced the concept of "housing service".

Demand for housing service

I assume that households have preferences over consumption bundles with some amount of a non-housing composite good N and housing services H .

Households choose the optimal consumption of N and H to maximise their utility $u = u(N, H)$ subject to the budget constraint

$$Y = P_n N + P_h H, \quad (1)$$

where Y is household's income, $P_n = 1$ and P_h are the unit price of non-housing composite good and housing services respectively.

- In the following I will employ the Cobb-Douglas (C-D) utility function.

Who will participate a social housing program?

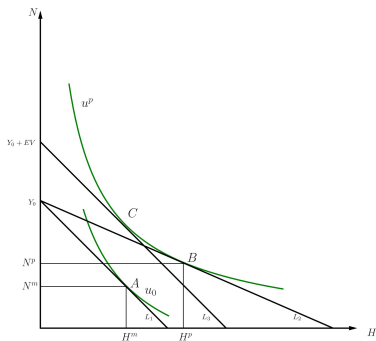
- Household will participate a housing project if she has higher utility than not participating.
- Consider following indirect utility function:

$$v(Y^*, P_h) = \underset{H, N > 0}{\text{Max}} u(H, N) \quad (2)$$

which allows household to reach the highest utility level with a given income Y^* , and the unit price of housing services P_h . So a utility maximizer household will join the program if she attains higher utility when she attends the program than when she does not i.e. she attends if $v(Y_0, P_h^p) > v(Y_0, P_h^m)$.

Equivalent variation

Household will need income $Y = Y_0 + EV$ to attain utility level u^P when she asked to pay the market rent.



Benefits of social housing and Deadweigth loss

- Households expenditure function:

$$e(P_h, u) = \underset{H, N > 0}{\text{minimize}} P_h H + N \quad \text{given} \quad u(H, N). \quad (3)$$

- Benefits of social housing

$$EV = B = e(P_h^m, u^p) - e(P_h^p, u^p) \quad (4)$$

- $EV = B = Y - Y_0$

- $B = e(P_h^m, v(P_h^m, Y)) - e(P_h^p, v(P_h^p, Y_0))$

$$B = \left(\frac{E_h^m}{\alpha} \right)^\alpha \left(\frac{Y_0 - E_h^p}{1 - \alpha} \right)^{1-\alpha} - Y_0. \quad (5)$$

Benefits of social housing and Deadweight loss

- DWL due to social housing programs is the difference between the benefits that occurs to the household and the effective subsidy or cost of the program to the housing authorities:

$$DWL = EV - (P^m - P^s)H^P = B - (E_h^m - E_h^P), \quad (6)$$

The Data (EU-SILC)

- I used European Union statistics on income and living conditions data, which is also known as *EU-SILC*, to estimate the benefits of social housing tenant.
- The data was conducted in 2008 with 5707 households. The data provides information on four tenure types: 4207 of households are homeowners, 1019 households live in rented dwellings, 282 households are living in social dwellings and 199 households have access to free accommodation.
- The household data contains two files (Household register file (*D-File*) and Household Data file (*H-file*)). *H-file* reports 111 variables and *D-file* reports 6 variables (in publicly available version).

Required variables for measuring the benefits of living in social houses

To derive the benefits of social housing dwellings we need the following variables:

- Initial income Y_0 (which is known to the social housing authority it is also provided in *EU-SILC* data set),
- Housing expenditure $E_h^P = HP_h^P$ (provided as current rent in *EU-SILC* data set),
- The market price of the social dwelling.

However, the data provides many other variables. The market value of social housing service is not documented, hence it needs to be estimated.

Missing data points

- An important problem of the data: the existence of missing observations.
- In our data one of the most important variables, *rent* which is our dependent variable, has missing observations.
- To overcome the problem of having missing observations in variable *rent*, I use the multiple imputation (MI) method to impute the missing data.

The variables that are used in imputation model

■ The data

Table 1: The list of the variables

Variable Description	Label in our Subset	Variable Label in Dataset
Household ID	household_id	HB030
Dwelling type*	dwel_type	HH010
Tenure status*	tenure_status	HH021
Number of rooms available to the household*	rooms	HH030
Leaking roof, damp walls,floors,foundations etc.*	leaking_roof	HH040
Current rent related to occupied dwelling*	rent	HH060
Total housing cost*	total_housing_cost	HH070
Bath or Shower in Dwelling*	bath_room	HH081
Problems with dwelling: too dark*	dark_home	HS160
Noise from neighbours or from the street*	noisy_home	HS170
Pollution,grime or other environmental problems*	bad_surrounding	HS180
Crime Violence or vandalism in the area*	crime_surrounding	HS190
Total household gross income*	total_income	HY010
Housing Allowances*	housing_allowency	HY070G
Region* (from D file in [9])	region	DB040
Family/Children related Allowances*	child_alw	HY050G
Capacity to afford paying for holiday*	holiday	HS040
Capacity to afford a meal with meat (or veg equivalent)*	good_meal	HS050
Financial burden of total housing cost*	hous_bur	HS140

The data and imputation method

- Multiple imputation modeling follows three steps, imputation, analysis and pooling. Figure 3 illustrates the MI procedure::

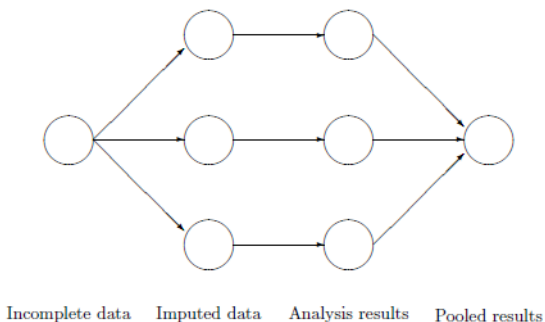


Figure: 3

- MICE algorithm is a Gibbs sampling: so it *is a Monte Carlo Markov chain algorithm for producing samples from the joint probability distribution of multiple random variables. The basic idea is to construct a Markov chain by updating each variable based on its conditional distribution given the state of the others.* (See Restricted Boltzmann Machines)
- In general further analysis is conducted with the pooled results of each imputed data set, I merge multiple imputed data sets into a single data frame *computing the mean or selecting the most likely imputed values* (For details see [2] and [6]).

Number missing data points in each variable	
region	0
dwel_type	36
tenure_status	0
leaking_roof	0
bath_room	0
indoor_wc	0
dark_home	0
noisy_home	0
bad_surrounding	0
crime_surrounding	0
rooms	0
rent	4427
total_income	0
total_housing_cost	141

MICE Algorithm: Notation

Let Y_j with $(j = 1, \dots, p)$ be one of the p incomplete variables. I denote set of observed and missing observations of Y_j by $Y^{obs} = (Y_1^{obs}, \dots, Y_p^{obs})$ and $Y^{mis} = (Y_1^{mis}, \dots, Y_p^{mis})$, respectively. The idea behind MI is to impute more than one value where missing observations occurs hence the number of imputed values are to be able to capture uncertainty in the data. Hence imputed data sets are more than one i.e. $m \geq 1$ and I denote the h th imputed dataset as Y^h where $h = 1, \dots, m$. Now let $Y_{-j} = Y_1, \dots, Y_{j-1}, Y_{j+1}, \dots, Y_p$ denote the collection of the $p - 1$ variable in Y except Y_j .

MICE Algorithm

$$\theta_1^{*(t)} \sim P(\theta_1 | Y_1^{obs}, Y_2^{(t-1)}, \dots, Y_p^{(t-1)}) \quad (7)$$

$$Y_1^{*(t)} \sim P(Y_1 | Y_1^{obs}, Y_2^{(t-1)}, \dots, Y_p^{(t-1)}, \theta_p^{*(t)}) \quad (8)$$

$$\vdots$$

$$\theta_p^{*(t)} \sim P(\theta_p | Y_p^{obs}, Y_1^{(t)}, \dots, Y_{p-1}^{(t)}) \quad (9)$$

$$Y_p^{*(t)} \sim P(Y_p | Y_p^{obs}, Y_1^{(t)}, \dots, Y_{p-1}^{(t)}, \theta_p^{*(t)}) \quad (10)$$

where $Y_j^t = (Y_j^{obs}, Y_j^{*(t)})$ is the j th imputed variable at iteration t

Diagnostic check after multiple imputation

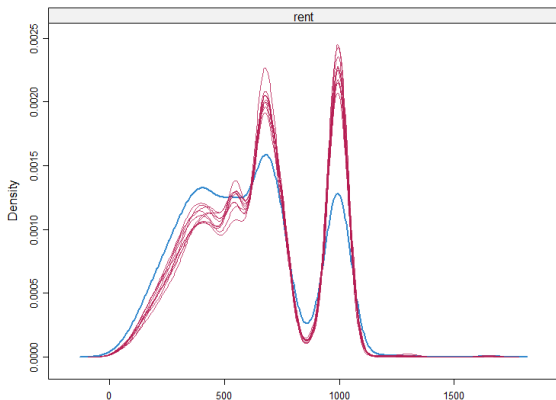


Figure: 4

Diagnostic check after multiple imputation

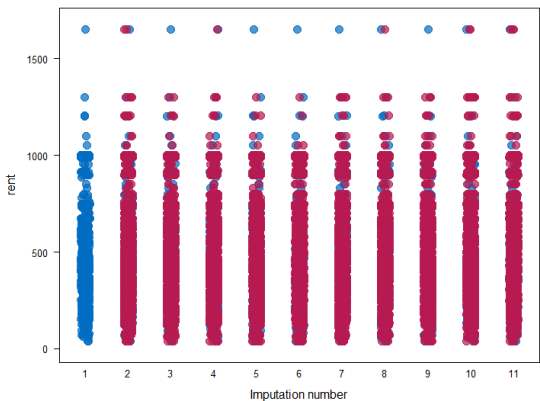


Figure: 5

However, there is no exact test which can measure the quality of imputed data. I am convinced that the imputed data set is a representative of the original sample because the MICE algorithm has been subject of many simulation experiments showing the efficiency of MICE algorithm (see [3] for details)

Rosen's (1976) hedonic regression

By following Rosen (1976)[9] arguments, we can predict the market value of social housing service with the following regression equation,

$$\log(\text{rent}) = m(X) + \epsilon, \quad (11)$$

where $m(\cdot)$ is a function with an unknown form, and X is a vector of the attributes *region*, *dwel_type*, *leaking_roof*, *bath_room*, *indoor_wc*, *noisy_home*, *bad_surrounding*, *crime_surrounding*, *rooms*.

Hedonic regression

- Rosen (1974) does not make any assumption on the functional form of the regression equation (11)
- Avoiding strict assumptions this.
- I trained a linear model, KNN regression[4], a non-parametric[5] regression and deep learnign model (Keras) to estimate equation (11).
- I take Rosen's warning into consideration,(as he stated "it is inappropriate to place too many restrictions on it (the equation (11)) at the outset")
- KNN reg had a large RMSE. However DL model had the lowest RMSE and OLS and Non-parametric regression predicted almost at the same accuracy level in training set, the difference between models was not too large.

Hedonic regression: OLS vs Machine Learning

- I did the further analysis with OLS, because ...
- It is computationally easy to perform (for this example), and also the test results show that there is a linear relationship between response and explanatory variables. So there is indeed a linear relationship between Y and X variables.
- Note: Important variables are also missing such as lot-size, district, age of the building ...

$$\log(\text{rent}) = m(X) + \epsilon = \beta_0 + \beta_1 * \text{region} + \beta_2 * \text{dwel_type} + \beta_3 * \text{leaking_roof} + \beta_4 * \text{bath_room} + \beta_6 * \text{noisy_home} + \beta_7 * \text{crime_surrounding} + \beta_8 * \text{rooms} + \epsilon$$

Residual Diagnostics

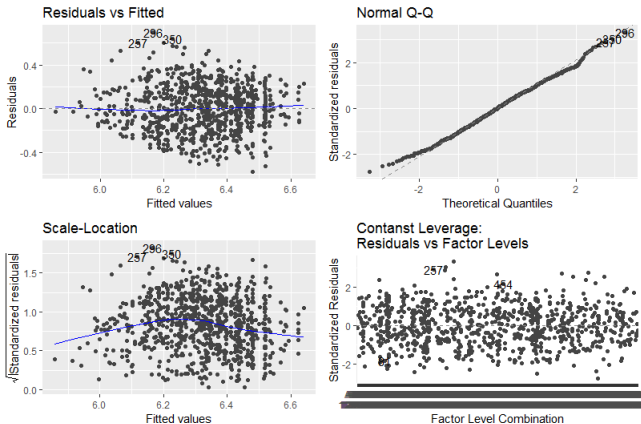


Figure:

Global Validation of Linear Models Assumptions

	Value	p-value	Decision
Global Stat	4.9224	0.2954	Assumptions acceptable.
Skewness	0.3941	0.530	Assumptions acceptable.
Kurtosis	1.9386	0.1638	Assumptions acceptable.
Link Function	1.2454	0.2644	Assumptions acceptable.
Heteroscedasticity	1.3443	0.2463	Assumptions acceptable.

Table: Test for Global validation of linear model assumptions [8]

The benefits of living in a social dwelling

The mean benefits for these high, moderate, and low income households equal to -63.1 , -15.6 , and 45.26 Euros respectively, so in average low income households benefit more.

The benefits of living in a social dwelling and distributional effect

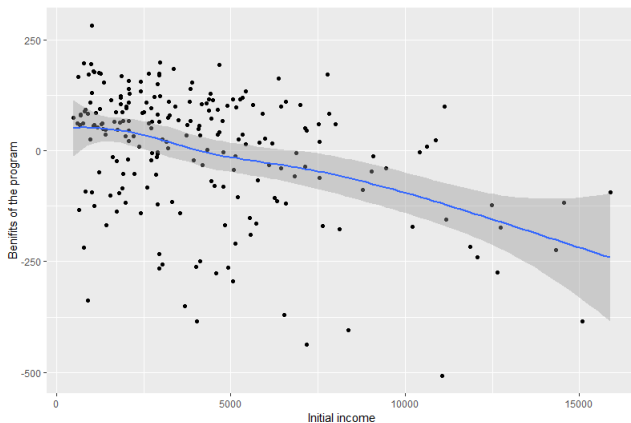


Figure:



Deadweight loss

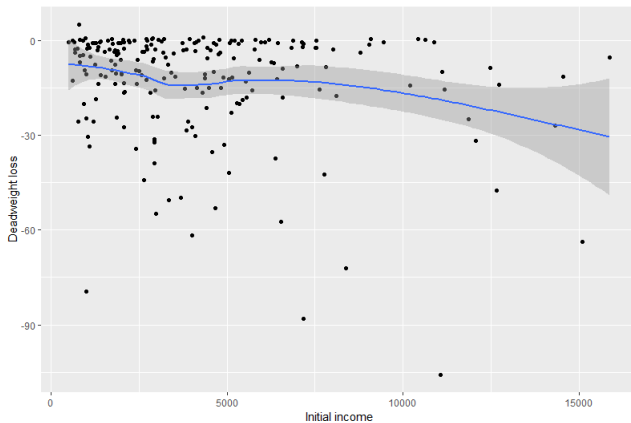


Figure:



Results

Household ID	E_h^P	E_h^m	Y_0	α	B	DWL
61	616.07	676.75	1280.00	0.48	60.00	-0.68
109	540.00	607.75	2738.92	0.20	65.00	-2.75
144	384.64	607.18	786.25	0.49	197.00	-25.54
192	519.63	580.51	606.83	0.86	61.00	0.12
459	475.45	604.81	1847.92	0.26	119.00	-10.36
604	559.86	607.18	1415.83	0.40	47.00	-0.32
1038	379.50	489.03	1997.50	0.19	99.00	-10.53
1784	540.51	583.81	4681.17	0.12	42.00	-1.30

Conclusion

- Equivalent variation measure helped us to estimate the benefits of living in social houses
- I identified that low income tenants are on average achieving higher benefits
- The benefits are thus distributed in favour of low-income families.

Thank you for your attention

Questions

References I



R. Akay.

The distribution of tenant benefits in Austrian social housing programs.

2019.






R. A. Burns, P. Butterworth, K. M. Kiely, A. A. Bielak, M. A. Luszcz, P. Mitchell, H. Christensen, C. V. Sanden, and K. J. Anstey.

Multiple imputation was an efficient method for harmonizing the mini-mental state examination with missing item-level data.

Journal of Clinical Epidemiology, 64(7):787 – 793, 2011.

References II

-  J. S. Granberg-Rademacker.
A comparison of three approaches to handling incomplete state-level data.
State Politics & Policy Quarterly, 7(3):325–338, 2007.
-  T. Hastie, R. Tibshirani, and J. Friedman.
The Elements of Statistical Learning.
Springer, 2013.
-  Q. Li and J. Racine.
Cross-validated local linear nonparametric regression.
Statistica Sinica, 14(2):485–512, 2004.

References III



D. Ldecke.

sjmisc: Data and variable transformation functions.
Journal of Open Source Software, 3(26):754, 2018.



R. Muth.

The Demand for Non-farm Housing.
University of Chicago, Department of Economics, 1958.



E. A. Pea and E. H. Slate.

Global validation of linear model assumptions.
Journal of the American Statistical Association,
101(473):341–354, 2006.
PMID: 20157621.

References IV



S. Rosen.

Hedonic prices and implicit markets: Product differentiation in pure competition.

Journal of Political Economy, 82(1):34–55, 1974.