Survey Research Methods (2012)
Vol.6, No.2, pp. 95-103
ISSN 1864-3361
http://www.surveymethods.org

Imputation of Housing Rents for Owners Using Models With Heckman Correction

Beat Hulliger and Gordon Wiegand
University of Applied Sciences and Arts Northwestern Switzerland FHNW

The direct income of owners and tenants of dwellings is not comparable since the owners have a hidden income from the investment in their dwelling. This hidden income is considered a part of the disposable income of owners. It may be predicted with the help of a linear model of the rent. Since such a model must be developed and estimated for tenants with observed market rents a selection bias may occur. The selection bias can be minimised through a Heckman correction. The paper applies the Heckman correction to data from the Swiss Statistics on Income and Living Conditions. The Heckman method is adapted to the survey context, the modeling process including the choice of covariates is explained and the effect of the prediction using the model is discussed.

Keywords: Selection bias, Statistics on Income and Living Conditions, survey data, imputed rents

1 Introduction

Income is the most important factor for many statistics on the social situation of a society (United Nations 2009). Many apparent and hidden sources are part of the income. Precise definitions of income sources are needed and the national and international scientific community and official statistics have elaborated on the underlying concepts and definitions of income, on its components, its reference unit (household, family, person) (International Labour Organisation 2003). The main problem is comparability across time and space (Aaberge et al. 2006).

One special issue is hidden income in the form of living in an owned dwelling. Obviously the dwelling is an asset which would normally provide income. Since an owner consumes the service of its dwelling immediately the benefit of owning a dwelling does not show up in the market. This may happen with many income components for example, when somebody owns a fruit tree or a water well. However consuming the service from an owned dwelling is an important part of income and such ownership is widespread. Therefore in any comparison of income the hidden income from using the services of an owned dwelling should be taken into account (Sauli and Törmälehto 2010). Otherwise important statistics like the indicators on poverty and inequality of the European Union, the so called Laeken indicators, would be flawed (European Commission 2003; Atkinson et al. 2002).

Comparing the income of tenants and homeowners is difficult since owners pay the service of their dwelling in a very different way than tenants. Tenants do not have this hidden income except if they pay a below-market rent. This may be the case when a rent subsidy is part of their income like for janitors.

Contact information: Beat Hulliger, School of Business FHNW, Olten, Switzerland, e-mail: beat.hulliger@fhnw.ch

In order to allow a cross-European comparison of income, Eurostat suggests to add an imputed rent to the income of owners (Eurostat 2006). The imputed rent is defined as the predicted rent minus home mortgage interest rate for our purposes.[1] Consideration of the imputed rent is necessary to allow international comparisons since the percentage of home owners differs considerably among the European countries. In Switzerland 57% of the households are tenants, in Spain 11% and in Belgium 29%.

The imputed rent is the amount of money a household would pay if it would rent its property, minus the mortgage interest or the below-market rent respectively. To capture the imputed rent three different approaches are possible and recommended by Eurostat (Frick et al. 2007).

Self-assessment Home-owners assess the imputed rent themselves. This approach is highly error-prone.

User-cost method The decision of a household to move into home ownership may be seen as a decision against the opportunity to invest in financial assets. The imputed rent is calculated using a nominal interest rate of the difference of the (owner estimated) market value of the house minus outstanding mortgages.

Rental equivalence method Imputed rents are calculated following a two step procedure: First, a regression model is fitted to the rent with appropriate explanatory variables, which may include characteristics of the dwelling and tenants. Next, the resulting coefficients and the model are applied to owners to predict a rent. A simplified variant of this approach is to use a stratification of the tenants with auxiliary variables, classify owners into the strata and impute mean rent of the stratum to an owner.

[1] A more precise definition of imputed rent is given by the Commission Regulation (EC) No 1980/2003 of 21 October 2003 implementing Regulation (EC) No 1177/2003 of the European Parliament and of the Council concerning Community Statistics on Income and Living Conditions (EU-SILC).

The rental equivalence approach is the most promising because it exploits available auxiliary information best. However, this approach requires many observations from which the model for rents can be estimated. Since the sample size is limited and usually no oversampling of tenants is possible, the proportion of tenants in the population must be large in order to achieve a sufficient number of observations in the sample. That is the case for Switzerland, where 65.4% of the dwellings are rented. A source of error may be that the model is developed for tenants and then applied to owners. Consequently a statistical technique like the Heckman method, which corrects for the selection bias using an additional model for the selection, has to be applied. While the use of regression models to impute rent is well established, the use of the Heckman correction is less. In an overview on the comparability of imputed rents Juntto and Reijo (2010, Table 6) find that in 14 European countries the regression method is used but only in 7 of these countries a Heckman correction is used.

However, the data from which rents have to be estimated usually stem from a sample survey as the Statistics on Income and Living Conditions (SILC). The modeling of the rent therefore has to take into account the sample design and non-response in order to be valid for the population. This is not common use, as for example the code proposed in (Eurostat 2006) does not consider sampling weights. In addition to unit-nonresponse in surveys, often also income components or covariates for a model are missing. For example in the case of the Swiss SILC (CH-SILC) the floor area of the dwelling is often missing though it is an important predictor of the rent. These practical issues have to be taken into account.

The Heckman method is introduced in detail in the next section. In Section 3.1 the CH-SILC data is presented. Section 3.2 describes the models to estimate the probability to own a dwelling and to predict the imputed rents. The focus lies on the building of the models. The article closes in Section 4 with a short description of the resulting imputed rents.

2 Predicting Imputed Rents With Heckman Correction

The Heckman method is a two step regression model. First, the probability to be a tenant, the selection probability, is estimated using a probit regression. Second, the housing rents of the tenants are modeled using a linear regression which incorporates the result of the first model. The model for the rent of tenants is then used to predict imputed rents of owners.

The Heckman method (Heckman 1979) is also explained in Eurostat (2006). Neither source takes into account, that the models should hold for the population and hence the population likelihood functions must be estimated from the sample. We describe the method considering the sampling design and following Heckman (1979) though using a simpler notation.

We consider two sets of explanatory variables x_i and z_i and the observed rent y_i for $i \in U$, U denoting the population of dwellings considered. While the dwelling is the unit of interest, properties of the household and the inhabitants living in the dwelling and of the location and neighbourhood are also important. These regressors x_i and z_i are usually multivariate. The underlying model on the level of the population is

$$y_i^* = x_i'\beta + \epsilon_i \tag{1}$$

$$d_i^* = x_i'\gamma_x + z_i'\gamma_z + \delta_i \tag{2}$$

The apostrophe "′" indicates a transposed vector. The dependent variable d_i^* in (2) denotes a latent variable which relates to the indicator of being a tenant (with the rent observed). The observed indicator of a dwelling having a tenant is then $d_i = 1_{\{d_i^*>0\}}$, where 1_A is an indicator function taking on the value 1 on the set A and 0 otherwise. Assuming that the error δ_i is normally distributed, equation (1) is a linear regression model for the rent y_i^*. The rent may be observed if i is a tenant, or not, if i is an owner. Thus y_i^* denotes a semi-latent variable here. The observable rent of a tenant is denoted $y_i = y_i^*$. The error terms δ_i and ϵ_i have a bivariate normal distribution with an expected value of 0, a variance σ_δ^2 and σ_ϵ^2 respectively and a covariance of $\rho\sigma_\delta\sigma_\epsilon$. The separation of the explanatory variables in x_i and z_i is necessary to distinguish the variables x, which are in both equations (1) and (2), from the variables z, which are included only in (2). For later use we introduce the short-hand notation $\mu_i = x_i'\gamma_x + z_i'\gamma_z$ for the linear predictor in (2).

To estimate the parameters the log-likelihood-function of the population has to be estimated and maximized according to the parameters. The well known two-stage procedure from Heckman can be used in a straightforward way when data on x and z is available for the whole population or for a simple random sample of the population. The estimates of the probit model are computed by applying an iteratively reweighted least squares method. For data stemming from a complex survey the population means involved in the algorithm can be estimated by Horvitz-Thompson estimators. (That is the so called pseudo-likelihood procedure, see Chambers and Skinner (2003, Chapter 2)). To estimate the variance with respect to the survey design a linear approximation may be applied. Algorithms for the probit model which take into account the sample design are implemented, for example, in SAS (PROC SURVEYLOGISTIC), R (package `survey`) and SPSS (CS procedures). These algorithms yield consistent estimates of γ_x and γ_z, which is the first step of the Heckman-method: a probit model of d_i on x_i and z_i. The dependent variable y_i^* in equation (1) may be observed for $\left\{i \in U | d_i^* > 0\right\}$. The condition $d_i^* > 0$ is equivalent to $\delta_i > -(x_i'\gamma_x + z_i'\gamma_z) = -\mu_i$. The conditional expectation of y_i^* equals

$$E = \left[y_i^* | d_i^* > 0, x_i, z_i\right] = x_i'\beta + E\left[\epsilon_i | \delta_i > -\mu_i\right]. \tag{3}$$

If β is estimated with the observed values y_i only, a bias is likely, because the expected value of the second summand in equation (3) is not necessarily zero. It may be calculated as:

$$E\left[\epsilon_i|\delta_i > -\mu_i\right] = \rho\sigma_\epsilon\lambda_i, \tag{4}$$

where λ_i is the inverse of the so called Mills ratio (IMR):

$$\lambda_i = \frac{\varphi(\mu_i/\sigma_\delta)}{\phi(\mu_i/\sigma_\delta)}, \tag{5}$$

where φ denotes the density of the standardized normal distribution and ϕ its cumulative distribution function. An estimate $\hat{\lambda}_i$ of the IMR may be derived by plugging in the estimate $\hat{\mu}_i = x_i'\hat{\gamma}_x + z_i'\hat{\gamma}_z$ from the probit model. If there is no correlation between the errors of the equations (1) and (2), i.e. if $\rho = 0$, then the IMR $\lambda_i = 0$. Since in our case only the binary variable d_i is observed but not the latent variable d_i^* we cannot estimate σ_δ. Actually in the probit model used to estimate γ_x and γ_z the estimation of σ_δ is not necessary (apart from a possible overdispersion) and we may assume that $\sigma_\delta = 1$.

The second step of the Heckman method is to estimate a linear model with response y_i and explanatory variables x_i and the estimate $\hat{\lambda}_i$ of the IMR:

$$y_i = x_i'\beta + \hat{\lambda}_i\alpha + \eta_i, \tag{6}$$

where $\eta_i = \epsilon_i - E\left[\epsilon_i|\delta_i > -\mu_i\right]$ now has expected value 0. Hence, the correction consists in introducing $\hat{\lambda}$ as an explanatory variable in the model. When model (6) is estimated from survey data the survey design must be taken into account. Again this is possible with the special programs adapted to survey sampling in SAS, SPSS and R and other statistical software packages.

A test of the hypothesis $\alpha = 0$ for the coefficient α of the IMR ($\hat{\lambda}$) indicates whether the Heckman correction is necessary. Failure to reject the null hypothesis does not confirm that $\lambda = 0$, and so $\hat{\lambda}$ may still be included in the linear model for observed rents. This does not introduce any bias but may be less efficient. Note that the usual tests for the other coefficients β underestimate the p-values since the uncertainty of estimating $\hat{\lambda}_i$ is not taken into account and therefore the (uncorrected) tests reject too often. Consistent variance estimators and correct tests exist for the case where the data covers the population or stems from a simple random sample with no nonresponse. However, neither in R nor in SAS there is currently an implementation of the method available which takes into account the sample design as well. We do take into account the sample design but have to admit that the tests for β may overstate the significance.

The interpretation of the estimates of the coefficients is relatively clear when we analyse the rents of tenants. However in the context of imputation of the predicted values for owners it is somewhat unclear whether the IMR should be included in the model to predict the rent or not, i. e. shall rent of owners be predicted as $\hat{y}_i^* = x_i'\hat{\beta}$ or $\hat{y}_i = x_i'\hat{\beta} + \hat{\lambda}_i\hat{\alpha}$? The decision depends on the interpretation of y_i^* and $\hat{y}_i$ respectively.

In our case $\hat{y}_i^*$ and $\hat{y}_i$ differ significantly, since the IMR is higher for house owners than for tenants. There is no recommendation in the literature. Three possible perspectives are possible:

1. The IMR is a proxy for different variables not represented by x. Especially it may contain information about the status of qualities of the dwelling, like "popularity of the neighbourhood", "view" or "special architecture" which are not reflected in the explanatory variables x. Under this interpretation the IMR should be used in the prediction.
2. The IMR is a premium for the owner to put the dwelling on the market. This premium would have to be included in the imputed rent when it is assumed that the rented dwellings do not have such a premium included in their rents. Since owned and rented dwellings are often specifically built and marketed for being sold or rented this premium may be justified.
3. The IMR is not used in the prediction since the imputed rent should reflect the theoretical income without consideration of market costs or other hidden variables.

We decided to not include the IMR in the prediction model. The main reason is that the $\hat{\lambda}_i$ differ widely between tenants and owners. It seems questionable whether the size of the difference may be explained by hidden variables or a premium. The influence of the IMR on the imputed rents would be very strong. In addition, Eurostat also excludes the IMR in the prediction model.

The Heckman method allows a consistent estimation of β. However, the least squares estimation of the variances is biased because the variability of the estimation of λ_i is not taken into account. An exception is the test for $\alpha = 0$. If the IMR has no significant influence in the model for rents we may assume that there is no selection bias. Assuming no selection bias and assuming the model (1) is correct, the variance estimators in the linear model are consistent. In the construction and selection of x_i and z_i we have to consider the underestimation of the variance of the errors by the usual tests.

Unfortunately the software at our disposal has no joint implementation of the Heckman method and survey modelling. The analysis was carried out with the software R (R-Development-Core- Team 2007). We used the package `survey` to take into account sampling[2] and paid more attention to relevance than to significance for the selection of the variables. To phrase it differently, the test statistics t, χ^2 and F and the p-values of the tests in the linear model are used only as relative indicators.

We selected the variables by the following criteria:

- The Schwartz Criterion or Bayes Information (BIC) and the Akaike Information Criterion (AIC)
- Nagelkerke R^2 for the probit model and complementary McFadden and Cox-Snell R^2
- Significant tests for continuous variables and particular levels of categoric variables and Wald tests for the whole variables (the Wald test is the alternative to the F-test if the dependent variable is binary)
- F-values

[2] A description of the model building and prediction process is given in the appendix.

- Kappa (κ) and variance inflation factors as measures for collinearities.
- Diagnostic plots like binned plot, Tukey-Anscombe plot and normal quantile-quantile plot

Further criteria result from the analysis of the predicted rents. They can be compared to the real rents paid by tenants (via plots and indicators of the distribution like minima, maxima and median). Furthermore the predicted rents of different models can be compared. Particularly the comparison of the predicted rents of a reduced model with those of a full model including all available explanatory variables may be instructive.

3 Data and Models

3.1 Data

SILC is a survey aiming at collecting timely and comparable cross-sectional and longitudinal multidimensional data on income, poverty, social exclusion and living conditions. This instrument is anchored in the European Statistical System. After two pre-test waves Switzerland participated fully since 2007. A preliminary version of CH-SILC of 2007 is used in this analysis. Figures in this article are therefore not official statistics.

For SILC the survey units are households and the survey population are all permanent residents of Switzerland. The sample design is stratified according to regions. The household data is weighted according to the inclusion probability derived from the sample design. Additionally, a calibration method based on the known distribution of the demographic characteristics of the population is applied to reduce possible bias from non-response. The results are therefore based on a household structure which accurately represents the permanent resident population in Switzerland (Graf 2008).

6 612 households have responded to CH-SILC 2007 and are included in the data set. From all available variables we selected 103 which might be relevant to predict rents. The variables cover topics like socio-demographic questions, dwelling characteristics including information on the neighborhood (like distance to the next playground in minutes to walk). The information on the type of the dwelling itself includes e.g. questions about the satisfaction with certain aspects.[3]

We have taken into account the stratification of the sample, including the size of the strata and the finite population correction. Calibration weights were computed by the Swiss Federal Statistical Office. These weights account for both the sample design and non-response (Graf 2008).

3.2 Models

The full model included the two steps: a probit model to estimate the probability that a household rents the dwelling and a linear model of the rent. We started by fitting a full model including all reasonable variables. We used 38 variables for the probit model and 29 of them plus the IMR for the linear model. All monetary variables like rent and income were subjected to the log transformation. The resulting Nagelkerke R^2 equals 0.62 and $AIC = 4\ 675$ in the probit model. Not surprisingly the multicolinearities are high, indicated by $\kappa > 10\ 000$. The variance inflation factors proved, that this is not attributable to any single variable, nor a small set of them. The information criteria for the linear model are lower than for the probit model: McFadden $R^2 = 0.46, AIC = 180$ and $\kappa = 75$.

Then we gradually reduced the number of variables and tried out different ways to recode them in order to find parsimonious models, that satisfy the information criteria best. We had to reduce the high multi-collinearity of the initial probit model by recoding and selecting variables. A backward selection strategy was used. The model should be as small as possible keeping in mind that the same equations should be applied for forthcoming waves of SILC. To make sure that a feasible set of variables can be identified which will be surveyed in coming waves without considerable additional effort, the set has to be as small as possible.

A disadvantage of a small set of variables in this context is that the influence of a single variable is rather high. That becomes a problem if the meaning of this variable changes over time. For instance the amount of money spent on life insurance may differ significantly if changes on the capital market take place.

The variables were split into two groups. One group contains all dwelling related variables like area of the dwelling or number of dwellings in the building. A second group of variables are related to the household or persons like income of the household or the nationality of the household member with highest income. We used both groups of variables for the probit model and only the variables directly related to the dwelling for the linear model. In other words the dwelling variables were candidates for the x-variables while personal variables were candidates for z-variables. Hence the IMR represents personal variables. Allowing also personal variables in the linear model for the rent did not enhance the fit.

The final models are shown in Table 1. The final probit model consisted of fifteen variables. Nine of them were adopted for the linear model. Note that there were variables retained in the probit model, even though they were not significant, because they became significant in the linear model. Leaving them out of the probit model may lead to inconsistencies.

One problem when developing a model was to detect and treat outliers. They were a minor problem in the linear model. For instance the distribution of the residuals did not show very large outliers but rather somewhat heavier tails than a normal distribution (cf. Figure 2). This did not hold for the probit model. A part of the sample seemed not to be fitted well by the probit model. The influence of these households on the probit model was judged too high. Actually the estimated probability of these households to own the dwelling was very high, but they were tenants nevertheless.

[3] The costs of the utilities are not itemized in CH-SILC, such that gross rent cannot be broken down to net rent and utilities like heating costs.

Table 1: Final probit and linear model

	Probit model		Linear model	
Variable	Estimate	Std. Error	Estimate	Std. Error
(Intercept)	2.572	0.181	2.837	0.016
Rooms	-0.216	0.020	0.083	0.003
Tenure(6-10)	-0.186	0.060	-0.017	0.007
Tenure(11-20)	-0.405	0.061	-0.036	0.008
Tenure(21+)	-0.310	0.067	-0.080	0.009
Heating-bad	-0.412	0.095	0.049	0.009
Pollution	-0.111	0.059	0.022	0.007
Vandalism	-0.090	0.058	-0.019	0.007
Noisy	0.061	0.062	-0.017	0.007
Washing-private	0.557	0.046	-0.045	0.008
Region2	0.172	0.060	0.087	0.007
Region3	0.074	0.050	0.027	0.006
Region4	-0.183	0.079	-0.074	0.011
Municipality-rich	-0.011	0.090	0.063	0.013
Municipality-developed	0.019	0.071	-0.060	0.010
Municipality-agrarian	0.020	0.083	-0.073	0.018
Occupation-part-time	-0.084	0.065	-0.020	0.007
Occupation-retired	0.058	0.072	-0.004	0.008
Occupation-other	-0.551	0.109	-0.010	0.013
Housing-Type-2	-1.080	0.052		
Housing-Type-3	-0.632	0.101		
Rent-not-burden	-0.303	0.055		
Private-car	0.248	0.062		
Foreigner	0.339	0.064		
Education-low	0.006	0.058		
Education-high	0.120	0.046		
Age	-0.015	0.002		
IMR			-0.061	0.013

After revision of the data with the help of the data owner most of these outliers could be corrected and only 3 observations had to be discarded as outliers in the final data set. The criteria to identify the outliers were:

1. All households with a housing rent over 6 000 CHF (max. 11 000 CHF)
2. All households with standardized residuals lower than −50 in the probit model.
3. All households which appear to be outliers in the space of explanatory variables according to the diagonal of the hat matrix (leverage points).

After identifying households as outliers and dropping them from the model, we re-fitted both models and checked again for outliers. For the final model 110 additional observations had to be excluded due to missing values. Thus the final reviewed data set contained 6499 observations.

The resulting Nagelkerke $R^2 = 0.53$ of the final probit model was lower than the $R^2 = 0.63$ of the full model. The AIC dropped to 5 728, which means that the relative goodness of fit improved. The measure of collinearity $\kappa = 400$ indicates that there are still some collinearities but much less than in the full model. The coefficient of determination of the linear model $R^2 = 0.37$ also indicates a slightly worse fit of the final model. On the other hand AIC dropped to 91 indicating a better fit, and the collinearity measure for the linear model was $\kappa = 24$.

A binned plot shows the averages of residual over bins versus the expected values. The binned plot for the final probit model is shown in Figure 1.[4] There is no large deviation from the expected values. The higher variability at the extremes is to be expected. Figure 2 shows the diagnostic plots for the final linear model. The residual distribution has heavier tails than expected under the normal distribution. No enhancement was found when using different transformations. There is still a possible leverage point in the data. However, the leverage of this point is much smaller than the values which occurred with the full model without outlier and leverage elimination.

The IMR has a significant influence in the linear model. The Heckman method was therefore appropriate and crucial. That is not necessarily the case. We developed the method for two more surveys. In one case the IMR was not significant despite intense modeling attempts. A more parsimonious model for the prediction of imputed rents might drop the selection correction in this case. However, it may be safer to leave the IMR in the linear model even if the final model is more complex and possibly less efficient.

The standard software for statistical production at the Swiss Federal Statistical Office is SAS. Though SAS has the neces-

[4] The binned plot was realized with the R-package "arm", Gelman et al. (2011).

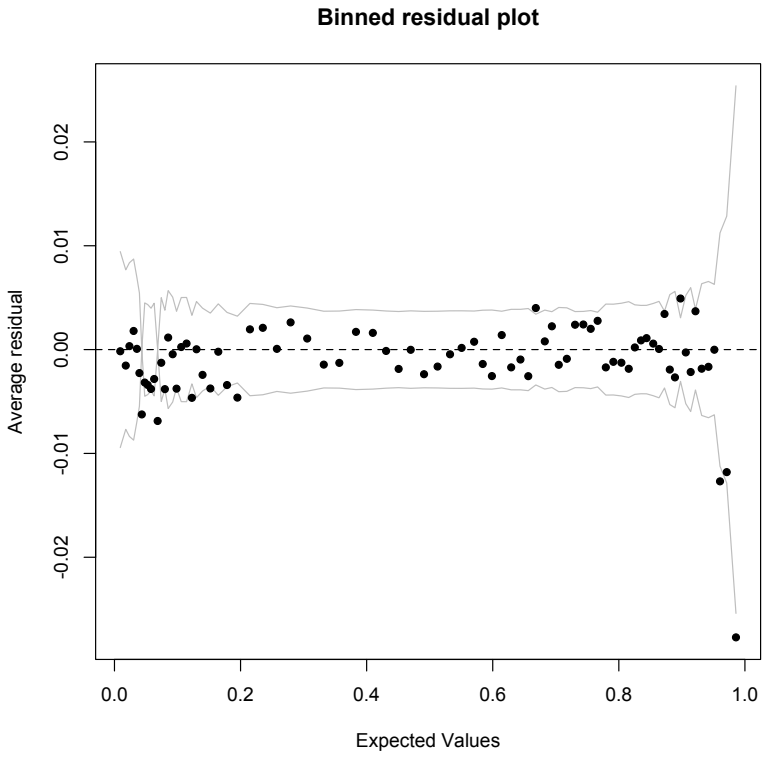

Figure 1. Binned plot of the residuals of the final probit model

Table 2: Imputed rents of owners and real rents of tenants

Rent	Model	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Real rents	none	200	960	1236	1340	1600	6000
Imputed rents	full	566	1260	1570	1690	1970	7140
Imputed rents	final	613	1332	1676	1796	2116	7390
Imputed rents	final, no IMR	582	1179	1467	1557	1825	5895

Real rents refers to the real rents of tenants. Imputed rents refers to the imputed rents of owners. The final model without IMR has no Heckman correction, i.e. uses just the linear model. The distribution refers to the sample only and is not an estimate of the population distribution.

sary building blocks the overall method is not available. We used the open-source programming language R (Ihaka and Gentleman 1996) with the package `survey` (Lumley 2004) to develop the models and impute rents. After finishing the modeling process we double-checked the models and results with SAS. A comparison of the models with and without consideration of the sample design showed considerable differences. Hence, taking the sampling design and the sampling weights into account was necessary.

4 Resulting Imputed Rents

For several explanatory variables the value range for the owners exceeds the range for the tenants. For instance the maximum number of rooms is 8 for tenants and 13 for owners. This is a potential danger because the model may fit poorly at the extremes of the range of the explanatory variables of the tenants and beyond.

One of the most important variables was the floor area of the dwelling, which is highly predictive for the rent. Unfortunately for more than 500 observation the area was missing. Therefore an imputation for area was developed where the variables used for the Heckman models were avoided as far as possible. An exception was the number of rooms of the dwelling which was highly significant in all models. In addition, the area of the dwelling may contain considerable measurement error. The mere number of missing values for dwelling-area is a good indicator for the size of measurement error to be expected.

The overall distribution of the imputed rents was similar under the full and the final model, as Table 2 shows. However, the imputed rent of a particular dwelling may differ considerably under the full and the final model.

The estimate of the population mean of the imputed rents of owners is 27% higher than the estimated population mean

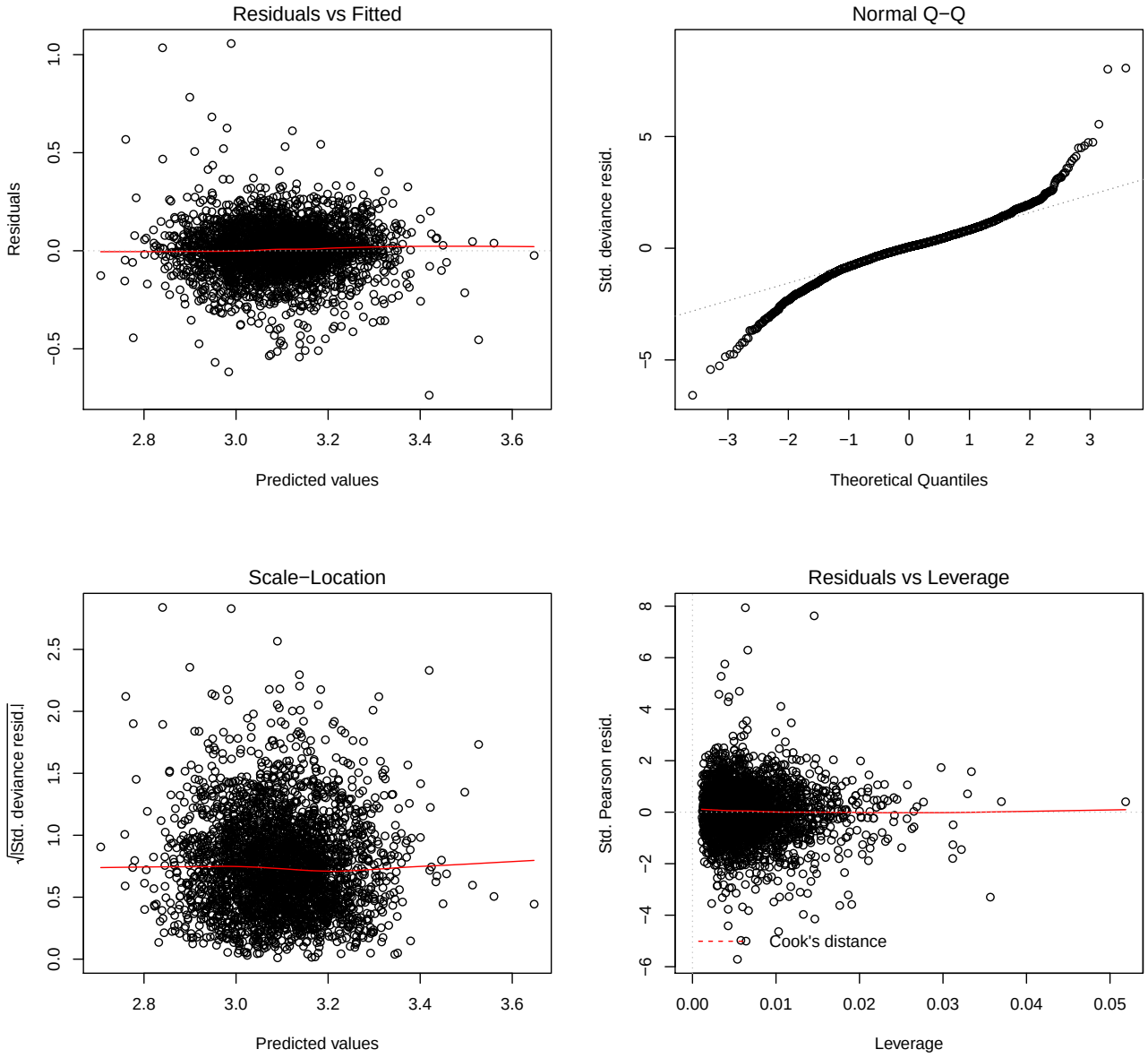

Figure 2. Diagnostic plots of the final linear model

rent of the tenants. That is to be expected since the dwellings for sale are typically of a higher quality in terms of size and furnishing.

Additional income due to the imputed rents occurs more frequently for high income households. The overall effect of the imputed rents on indicators of poverty and social exclusion is moderate but clearly visible. The estimate of the mean annual income of the population increases by 10.8% when adding the imputed rents. The estimate of the population at-risk-of-poverty rate (ARPR) decreases 1.1 percentage points with the additional income.

The Heckman correction has a significant impact on the prediction of the imputed rents. The last row in Table 2 shows the imputed rents computed without Heckman correction. In other words the last row shows the distribution of the imputed rents of owners in the sample when the final linear model does not include the IMR as an explanatory variable. The imputed rents using the Heckman correction have a considerably higher sample mean than without correction. Hence the Heckman correction is highly recommended to reduce the selection bias.

5 Conclusion

Regression models using a set of good predictor variables can predict the rent of tenants well. Apart from the sample design also a selection bias due to the decision of renting versus buying a dwelling should be taken into account. Correction of a selection bias with the Heckman method is feasible and leads to useful results.

The sample design and the sample weights may have a high impact and should be taken into account when fitting the models. Otherwise the predictions may be valid for the sample at hand but the inference to the population may not be sound.

Using the model developed for tenants to predict rents of owners is a feasible way to impute rents. However, since the range of the explanatory variables for owners and tenants is not the same such that the prediction is actually an extrapolation beyond the range of the explanatory variables used to estimate the model, there is an additional risk of misspecification of the model for the owners.

The imputed rents for the owners have a moderate effect on the mean income and on important poverty indicators.

Since non-response and variables differ from survey to survey there is no guarantee that the Heckman correction is always necessary and that a particular model fits well across surveys and time. However, the Heckman correction may have a considerable impact and should be tested in any case.

Careful model building and checking is necessary to ensure that a well founded model is used for the prediction of rents. In particular when a periodic survey is used the models should be parsimonious to diminish the risk of changing the model for every survey edition.

Acknowledgements

A large part of the work for this article has been carried out under a contract with the Swiss Federal Statistical Office. We are grateful for being allowed to use the data and results for this publication. In particular we would like to thank Anne Cornali Schweingruber and her team for their support and fruitful comments. We thank the two anonymous referees for their valuable comments.

References

Aaberge, R., Fjaerli, E., Langorgen, A., & Mogstad, M. (2006). Comparability of income data across households/individuals and over time. In Eurostat (Ed.), *Comparative EU statistics on Income and Living Conditions: Issues and Challenges Proceedings of the EU-SILC conference* (p. 57-75).

Atkinson, T., Cantillon, B., & Marlier, E. (2002). *Social Indicators: The Eu and Social Inclusion*. Oxford: Oxford University Press.

Chambers, R. L., & Skinner, C. J. (2003). *Analysis of Survey Data*. Chichester: John Wiley & Sons.

European Commission. (2003). *Laeken indicators. detailed calcultion methodology.* Technical report, EUROSTAT Working Group statistics on income, poverty and social exclusion.

Eurostat. (2006). Hbs and eu-silc imputed rent. In D. E.-S. . H. DOC HBS/161/2006/EN and E.-S. I. rent. In Eurostat-Luxembourg (Ed.), *Meeting of the working group on living conditions (HBS, EU-SILC and IPSE), Number Doc. HBS/161/2006/EN, Doc. EU-SILC/162/06/EN.* Luxembourg: Eurostat.

Frick, J. R., Goebel, J., & Grabka, M. M. (2007, November). Comparative EU statistics on Income and Living Conditions: Issues and Challenges. In *Chapter Assessing the Distributional Impact of Imputed Rent and Non-Cash Employee Income in Micro-Data* (p. 117-142). Helsinki: Proceedings of the EU-SILC conference.

Gelman, A., Su, Y.-S., Yajima, M., Hill, J., Pittau, M. G., Kerman, J., et al. (2011). *arm: Data Analysis Using Regression and Multilevel/Hierarchical Models.* R package version 1.4-07.

Graf, E. (2008). *Pondérations du SILC pilote, SILC_I vague 2, SILC_II vague 1, SILC_I et SILC_II combinés*. Neuchâtel: Bundesamt für Statistik (BFS), Statistik der Schweiz.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*(1), 153-161.

Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, *5*(3), 299-314.

International Labour Organisation. (2003). *Household income and expenditure statistics (report ii).* Technical report, Seventeenth International Conference of Labour Statisticians, Geneva.

Juntto, A., & Reijo, M. (2010). *The comparability of imputed rent.* Eurostat Methodologies and Working Papers. Eurostat.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, *9*(1), 1-19.

R-Development-Core-Team. (2007). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Sauli, H., & Törmälehto, V. (2010). *Income and living conditions in Europe.* Chapter The distributional impact of fictive rent, pp. 155-178. eurostat statistical books.

United Nations. (2009). *Rethinking Poverty: Report on the World Social Situation 2010.* Department of Economic and Social Affairs.

Appendix: Procedure for Model Building

The data set of CH-SILC is not publicly available. Hence the source code of our R routines is available from the authors but would not be helpful for researchers with other data. Nevertheless it may be helpful to outline the general procedure we followed. The data set must contain the following elements:

- all explanatory variables for both model steps
- the rents for the tenants

Additionally three variables are needed to incorporate the survey design:

- the calibration weights
- a variable with the assignment to a stratum
- a variable with the size of the strata in the population

It may be helpful to recode and/or transform some variables, e.g. to use log transforms of the rent and other monetary variables. Furthermore it is advisable to impute missing values for crucial variables like floor area.
In a first step the probit model is fitted using the survey design for the complete sample. Subsequently possible outliers with respect to this model are identified and eliminated from the data set and the survey design is restricted to the domain of non-outliers. The probit model is refitted with the reduced data set. This may be an iterative process. The Inverse Mills ratio (IMR) is calculated by (5).
The IMR has to be added to the set of variables designated to predict the rent and the survey design has to be restricted to the domain of the tenants. The rents for the tenants is modeled by a linear regression. Again outliers have to be identified and eliminated and a parsimonious model is built in an iterative process. After obtaining a definitive model with a consistent data set the model can be applied to the owners. The IMR is removed from the linear model to impute rents for home-owners.