Chapter 6

Model Diagnostics

Small area estimation models used in Chapters 3 and 4 are particular cases of linear mixed models.[1] Under model-based SAE, using either unit-level (Chapter 4) or area-level (Chapter 3) models, a series of useful model diagnostics can help verify model assumptions and assess model fit. These checks also include residual analysis to detect deviations from the assumed model and detection of influential observations (Rao and Molina 2015). Performing thorough and rigorous model diagnostics as part of the SAE exercise is crucial to ensure the validity of small area estimates.

This chapter describes the underlying model considered in Section 6.1. Then it provides brief recommendations for model and variable selection in Section 6.2 and concludes with regression diagnosis in Section 6.3.

6.1 The nested-error model

The model used for small area estimation of poverty and welfare, such as the ones proposed by Elbers, Lanjouw, and Lanjouw (2003) and Molina and Rao (2010), assume that the transformed welfare y_{ch}, for each household h within each location c in the population is linearly related to a $1 \times K$ vector of characteristics (or correlates) x_{ch} for that household, according to the nested-error model:

$$y_{ch} = x_{ch}\beta + \eta_c + e_{ch},\ h = 1, \ldots, N_c,\ c = 1, \ldots, C, \tag{6.1}$$

where η_c and e_{ch} are respectively location and household-specific idiosyncratic errors, assumed to be independent from each other, following:

$$\eta_c \overset{iid}{\sim} N\left(0, \sigma_\eta^2\right),\ e_{ch} \overset{iid}{\sim} N\left(0, \sigma_e^2\right),$$

where the variances σ_η^2 and σ_e^2 are unknown. Here, C is the number of locations in which the population is divided and N_c is the number of households in location c, for $c = 1, \ldots, C$. Finally, β is the $K \times 1$ vector of coefficients.

As illustrated in Chapter 4, the assumption of normality plays a considerable in EB methods. Deviations from this assumption may lead to biased and noisier estimates, as shown in Corral et al. (2021). Isolated deviations from the model (outliers) and influential observations or even outlying locations may also

[1] Except for the gradient boosting application shown.

exist. Thus, the following sections provide some insights towards selecting a suitable model, as well as checks that may be done to ensure that the chosen model for SAE is adequate.

6.2 Model and variable selection

The objective of model selection is to determine the relevant covariates out of a pool of candidate model variables such that the resulting SAE model generates the most precise estimates possible, noting that it must be possible to measure this precision accurately. Classic approaches to model selection include lasso or stepwise regression. Other literature on the subject includes the fence method, described in Pfeffermann (2013), which involves selecting a model out of a subgroup of correct models that fulfill specific criteria. Rao and Molina (2015) also elaborate on other methods such as Akaike Information Criteria (AIC) type methods, which rely on the marginal restricted log likelihood based on normality of the random effects and the errors. For variable selection under area-level models, particularly Fay-Herriot models, Lahiri and Suntornchost (2015) propose a method where the approximation error converges to zero in probability for large sample sizes.[2]

Since the aim is to arrive at the "true" model, removing all non-significant covariates from the model is recommended as these may introduce noise. It is important not to confuse the estimated noise and the true noise of a given small area estimate. The most common uncertainty measure for an area-specific prediction is the MSE (Tzavidis et al. 2018). In applications of SAE, such as those illustrated for unit-level models, an estimate of the MSE for the small area estimator is obtained through a parametric bootstrap procedure (see section 4.4.2.2). This method differs considerably from the method used in the ELL method, where a single computational algorithm is used to obtain predictors and assess uncertainty. Corral, Molina, and Nguyen (2021), through model-based simulations, present evidence of the fact that the single computational algorithm to estimate noise used in ELL could underestimate the actual MSE of the method.[3]

The script example provided in section 4.5.1 of Chapter 4 uses a lasso approach for model selection that initially includes all suitable covariates and uses 10 fold cross-validation with a shrinkage parameter λ that is within one standard error of the one that minimizes the cross-validated average prediction squared error. The lasso approach employed here does not consider the nested error structure of the model; that is, it is done with the corresponding linear model without the random area effects. Practitioners who are comfortable with R may rely on the `glmmlasso` R package, which may be used for model selection in the model with the assumed nested error structure (see Groll and Tutz (2014) and Groll (2017)).

The lasso selection process yields a selected set of covariates, although some of the included covariates may be non-significant. Consequently, the next step is to remove all the non-significant covariates sequentially. The process starts by removing the most non-significant covariates, one by one. When a covariate is removed, the significance of other covariates may change; thus, after removing each covariate, the model is fit again to determine which covariate to remove next until the remaining ones are all significant. Note that the process used here ignores the magnitude of the coefficients and thus could be further improved.

Finally, it is recommended to remove highly collinear covariates. Once these are removed, the following steps are to identify outliers and influential observations, which may lead to considerably different estimated model parameters, see the next section. For the Fay-Herriot model of Chapter 3 a different

[2]The authors define the approximation error as the difference between standard variable selection criterion and the ideal variable selection criterion, where it is assumed that the direct estimates are not measured with noise.

[3]See Corral et al. (2021) for a detailed discussion on the previous ELL bootstrap. Also see Elbers, Lanjouw, and Lanjouw (2002) for the sources of noise in their algorithm.

approach than the one detailed here was taken. In the implemented approach for area-level models, the model started with all possible covariates and began removing covariates, starting with the least significant ones. The removal was done considering the random effects (see section 3.2).

6.3 Regression diagnostics

After the fitting process, checking whether the assumptions from the underlying model are satisfied is recommended. Regression diagnostics for Linear Mixed Models (LMM)[4] are more difficult to interpret than those of the standard linear models, since these models include random effects, which lead to a different covariance structure. Moreover, tools for diagnostics of linear mixed models are less common in software packages, including Stata. Thus, practitioners either must code their own diagnostics or rely on diagnostics often used for linear regression. The model assumptions to keep in mind are: [5]

1. Linearity: the dependent variable y_{ch} is a linear function of the selected vector of covariates x_{ch}.
2. Normality: random effects η_c and errors e_{ch} are normally distributed.
3. Homoskedasticity: errors' variance σ_e^2 is constant, although this assumption may be relaxed by modeling heteroskedasticity using a model such as the alpha model specification provided by ELL (2003). This can be done when choosing Henderson's Method III fitting in Stata's `sae` package.
4. Independence: errors e_{ch} are independently distributed. Under the nested-error model, the assumption is extended so that e_{ch} in location c is unrelated to the corresponding location effect η_c, as well as to all other location effects η_l, $l \neq c$.

Other issues to consider are the detection of influential observations and outliers and, as discussed above, multicollinearity. Although these are not part of the assumptions, their presence may affect the precision of estimates of model parameters and, in turn, model predictions.

The following subsections include some formal and informal ways to check if the assumptions of the underlying model are satisfied. The process starts by eliminating covariates with high multicollinearity, followed by influence analysis and, finally, residual analysis which encompasses most assumptions.[6] Model diagnostics based on residuals and influence measures for special cases of the general linear mixed model can be found in Rao and Molina (2015).

6.3.1 Multicollinearity

Collinearity reduces the accuracy of estimates of regression coefficients, leading to larger standard errors of the coefficients. Under the multiple imputation (MI) inspired bootstrap methods, such as the one often used in the ELL method (see Ch. 4), larger standard errors in the coefficients typically led to larger estimates of noise for the estimators of the indicators of interest. Thus, care was taken to avoid collinearity and multicollinearity. A simple way to detect collinearity is via a correlation matrix, where large absolute correlation coefficients may suggest collinearity problems (James et al. 2013). However,

[4] LMM are an extension of general linear models which include both fixed and random effects.

[5] These assumptions are similar to the ones of a classic linear regression, but adding those for the random effects inclusion.

[6] Many of the commands provided in the following subsections can be easily reviewed by typing `help regress postestimation` in Stata's command window.

collinearity may occur between more than a pair of variables leading to multicollinearity, which cannot be detected under the pairwise correlation matrix (*ibid*).

Under the presence of multicollinearity, the inspection of the correlation matrix is no longer sufficient, and one must instead compute the variance inflation factors (VIFs) (James et al. 2013). The smallest possible value for a VIF is 1. There are multiple rules of thumb as to what is an acceptable VIF. According to James et al. (2013), values exceeding 5 or 10 may require action. After the model is fit, the command `estat vif` may be used to check the variance inflation factor for the covariates included in the model.

Example 1:

Variance inflation factor values above 10 might require special attention since the variable in question could be a linear combination of other independent variables. In the example below, a special Mata function,[7] `_f_stepvif()`, is used to remove covariates with a VIF above a specified threshold; in this case, the chosen value is 3.[8] The function expects the covariate list as its first argument, followed by the sample weights, and finally, the threshold. After the removal of high VIF covariates, the resulting covariate list is returned in a Stata local macro called `vifvar`.

```
*================================================================================
// Collinearity
*================================================================================

reg y $postsign [aw=Whh],r

    //Check for multicollinearity, and remove highly collinear (VIF>3)
        cap drop touse              //remove vector if it is present to avoid error in next step
        gen touse = e(sample)           //Indicates the observations used
        estat vif                       //Variance inflation factor
        local hhvars $postsign

    //Remove covariates with VIF greater than 3
        mata: ds = _f_stepvif("`hhvars´","Whh",3,"touse")
        global postvif `vifvar´

        //VIF check
        reg y $postvif [aw=Whh], r
        vif

    // For ilustration
    // Henderson III GLS - model post removal of non-significant
    sae model h3 y $postsign [aw=Whh], area(HID_mun)

    // Henderson III GLS - model post removal of non-significant
        sae model h3 y $postvif [aw=Whh], area(HID_mun)
```

6.3.2 Influence analysis

Influence analysis is used to detect observations that may have considerable impact on the estimates of the parameters and, consequently, model predictions. These observations include outliers (observations

[7] The Mata function is included in Stata's `sae` package.
[8] The specific threshold value is up to the practitioner, although it is not recommended to use thresholds above 10.

with a large residual, i.e., an observation poorly predicted by the model) and influential observations (the omission of which considerably changes the point estimates of β). These observations can be identified by measuring how far the observation's value for a predictor variable is from the mean or by the size of their studentized residual. Influence analysis is recommended prior to calculating any small area estimate, as these observations may impact the bias and noise of the final small area estimates.

Cook's distance (Cook 1977), also known as Cook's D, measures the effect on the estimated coefficients when an observation is left out or deleted (Rao and Molina 2015). Under ordinary least squares regression, Cook's distance can be obtained in Stata with the `cooksd` option after the `predict` command. Nevertheless, under regular OLS, Cook's distance focuses solely on isolated observations, whereas, under the nested-error model used for SAE shown in equation 6.1, the analysis of the influence of particular locations may be more relevant. The Stata package `mlt` by Möhring and Schmidt-Catran (2013) may be used to assess the influence on the estimated parameters of particular locations or groups.[9] The `mlt` command estimates Cook's D empirically, making it computationally intensive. The rule of thumb for classifying influential locations is absolute Cook's D values greater than $4/C$, where C is the number of locations into which the population is divided. The `mlt` command will also calculate DFBETAs, which measure the influence of a single location on the coefficient of each covariate. It represents the standardized difference between the coefficient with and without the given location (Möhring and Schmidt-Catran 2013). The rule of thumb for classifying locations as influential is a DFBETA absolute value above $2/\sqrt{C}$, although this should be applied with caution.

Leverage measures the influence on the fitted values of a given observation. Unfortunately, data packages that can obtain leverage under the assumed model are not available. Cameron and Trivedi (2005) present an alternative toward handling influential observations under OLS by using the post estimation command `dfits`, which shows the difference in fits (predictions) with and without the unusual observation. The command combines outliers and leverage into a single statistic. A rule of thumb to identify these observations is if $|dfits| > 2\sqrt{k/n}$, where k is the number of covariates and n is the number of observations. Nevertheless, the removal of observations always entails loss of information (which may be fair) and thus much care should be taken before deciding on removal, and should be done only when, after inspecting the offender, it is determined to be mistaken.

Example 2:

After fitting the model and calculating residuals, problematic observations are identified using several rules of thumb: $Cooks'd > 4/n$, $leverage > (2k+2)/n$ and $abs(rstu) > 2$.

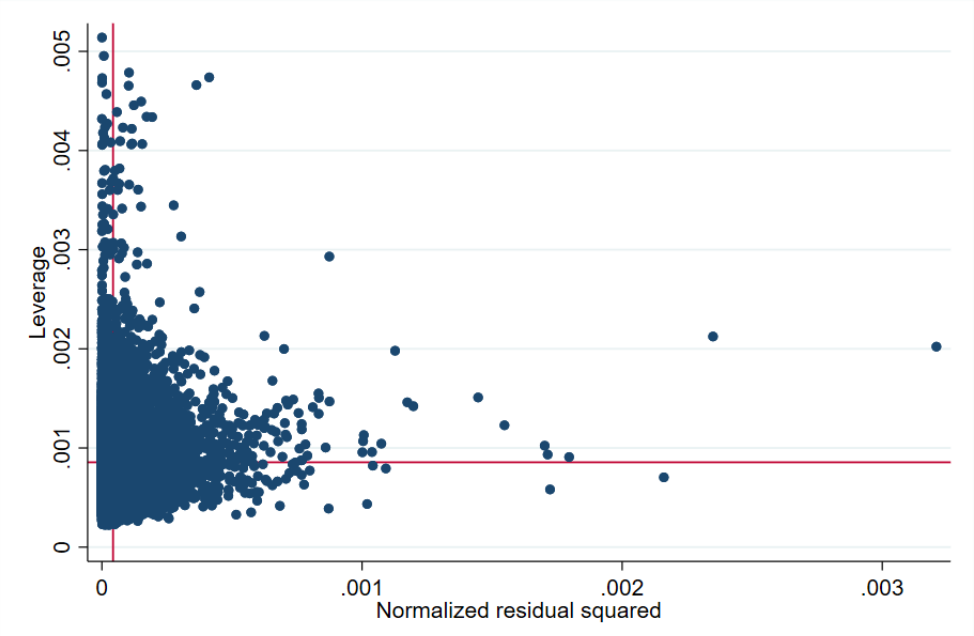
```
// Step 1
    reg y $postvif

    // After regression without weights...

    // Calculate measures to identify influential observations
        predict cdist, cooksd      // calculates the Cook´s D influence statistic
        predict rstud, rstudent    // calculates the Studentized (jackknifed) residuals
// Step 2
    reg y $postvif [aw=Whh]
```

[9] https://www.stata.com/meeting/germany12/abstracts/desug12_moehring.pdf

```
// Predict leverage and residuals
    predict lev, leverage  // calculates the diagonal elements of the
                        // projection ("hat") matrix
    predict r, resid       // calculates the residuals

// Save useful locals
local myN=e(N)              // # observations
    local myK=e(rank)           // rank or k
local KK =e(df_m)           // degrees of freedom (k-1)

sum cdist, d
* return list
    local max = r(max)          // max value
    local p99 = r(p99)                  // percentile 99

// Step 3

// For ilustration...
// We have influential data points...
reg lny $postvif if cdist<4/`myN´ [aw=Whh]
reg lny $postvif if cdist<`p99´   [aw=Whh]
    reg lny $postvif if cdist<`max´   [aw=Whh]

// Identified influential / outliers observations
gen nogo = abs(rstud)>2 & cdist>4/`myN´ & lev>(2*`myK´+2)/`myN´

count if nogo==1        // these are the obs that we want to eliminate
```

A graphical representation of the squared normalized residual versus leverage (`lvr2plot`), before and after the elimination of influential observations, is an easy way to identify potentially influential observations and outliers (Fig. 6.1). The two reference lines are the means of the leverage and the squared normalized residual. Many points are outside these two reference points and should be scrutinized before deciding to remove them.

Figure 6.1: Squared normalized residual vs leverage



Source: own elaboration from code in Appendix 6.4.2. The red lines are references to the mean leverage on the Y axis and the mean squared normalized residual on the X axis. Observations that are far away from these reference points should be inspected more closely.

6.3.3 Residual analysis

Most techniques for residual analysis rely on visual inspection of the graphed residuals. Residuals should be checked for linearity, normality, and constant variance in case homoskedasticity is assumed.

6.3.3.1 Linearity

The nested-error model used in SAE (Eq. 6.1) assumes that the outcome variable y is a linear function of the covariates.[10] If a single covariate is used, a scatter plot of the residual versus the covariate is enough to see if a linear relationship exists. When several covariates are used, checking for linearity is somewhat more complex.

To assess linearity in a simple regression, use `scatter` to produce a plot of y versus x, `lfit` to fit a regression line, and `lowess` to show a smoothed fit.

```
. reg depvar indepvar
. twoway (scatter depvar indepvar)(lfit depvar indepvar)(lowess depvar indepvar)
```

For multiple regression:

```
. reg depvar indepvars
. predict r, residual
. scatter r indepvar1
. scatter r indepvar2
```

Other checks for non-linearity include `acprlot` and `cprplot`, which produce graphs of an augmented component-plus-residual plot and a residual plot, respectively.

```
. reg depvar indepvars
. acprplot indepvar1, lowess lsopts(bwidth(1))
. acprplot indepvar2, lowess lsopts(bwidth(1))
```

When a clear non-linear pattern is observed, transformations of the independent variable might help.

```
. graph matrix depvar indepvars, half
. kdensity indepvar, normal
. gen logvar=log(indepvar)
. kdensity logvar, normal
```

```
// Linearity
    reg y $postsign, r

    // Augmented component-plus-residual plot; works better than the
    // component-plus-residual plot for identifying nonlinearities in the data.
    acprplot age_hh, lowess lsopts(bwidth(1)) graphregion(color(white)) msize(small)

    graph export "$figs\acprplot_age_hh.png", as(png) replace
```

[10] Much of the information in this section is borrowed from UCLA: Statistical Consulting Group (2022); https://stats.idre.ucla.edu/stata/webbooks/reg/chapter2/stata-webbooksregressionwith-statachapter-2-regression-diagnostics/

```
// Kernel density plot  for log_age_hh with a normal density overlaid
kdensity age_hh, normal graphregion(color(white)) msize(small)

graph export "$figs\kdensity_age_hh.png", as(png) replace

// log transformation
gen log_age_hh =log(age_hh)

// Kernel density plot  for log_age_hh with a normal density overlaid
kdensity log_age_hh, normal graphregion(color(white)) msize(small)

graph export "$figs\kdensity_log_age_hh.png", as(png) replace
```

6.3.3.2 Tests for normality of residuals and random effects

Model errors and random effects are assumed to be normally distributed under the nested-error model used for SAE. Deviations from normality may lead to considerable bias in the final small area estimates. Scatter plots of residuals against fitted values and normal Q-Q plots of residuals provide a natural way to identify outliers or influential observations that might affect the precision of the estimates. West, Welch, and Galecki (2014) mention that the random effects vector is assumed to follow a multivariate normal distribution. Thus, the information from the observations sharing the same random effect is used to predict (instead of estimating) the values of that random effect in the model.

The usual predictors of the random effects under linear mixed models are known as Empirical Best Linear Unbiased Predictors (EBLUPs), since they are the most precise linear unbiased predictors. They use the weighted least squares estimates of β, and the variance parameters are replaced by suitable estimates (Robinson 1991). Rao and Molina (2015) provide a comprehensive formal derivation of the EBLUPs, and applications beyond small area estimation can be found in West, Welch, and Galecki (2014).

The `xtmixed` or `mixed` command may be used to check the validity of a linear mixed model in Stata. Deviations from normality can be observed in a normal Q-Q plot of residuals (Figure 6.2), which displays sample quantiles of unit-level residuals against the theoretical quantiles of a normal distribution plot:

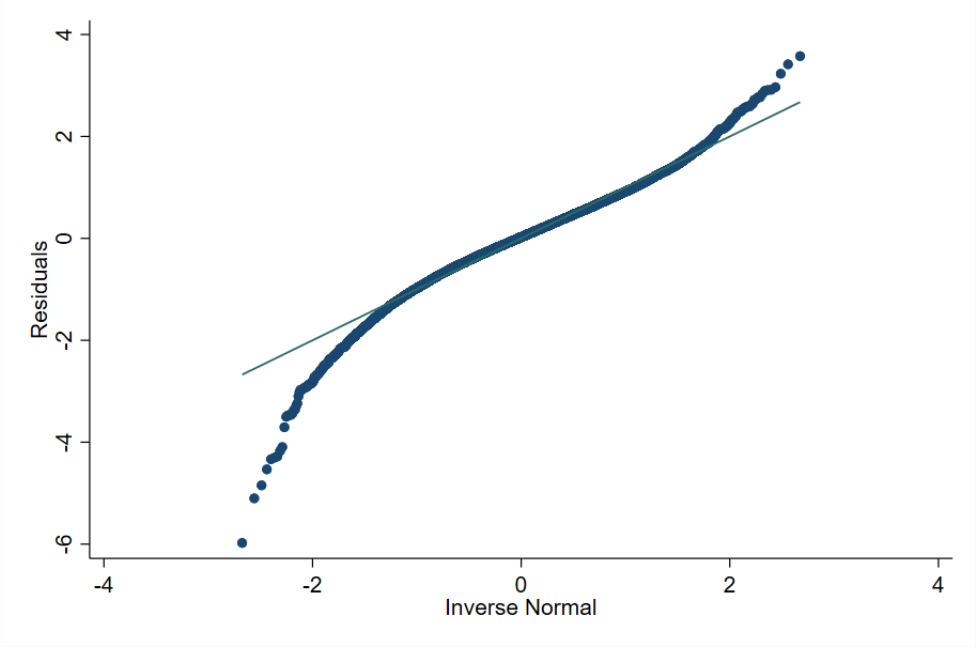
```
mixed depvar indepvars || area:, reml

predict res, residual

qnorm res
```

It is also important to check the assumptions regarding the distribution of the random effects. This may be done after fitting the linear mixed model and obtaining the predicted random effects (Figure 6.3):

```
predict eta, reffects

bysort area:  gen first = _n

qnorm eta if first==1
```

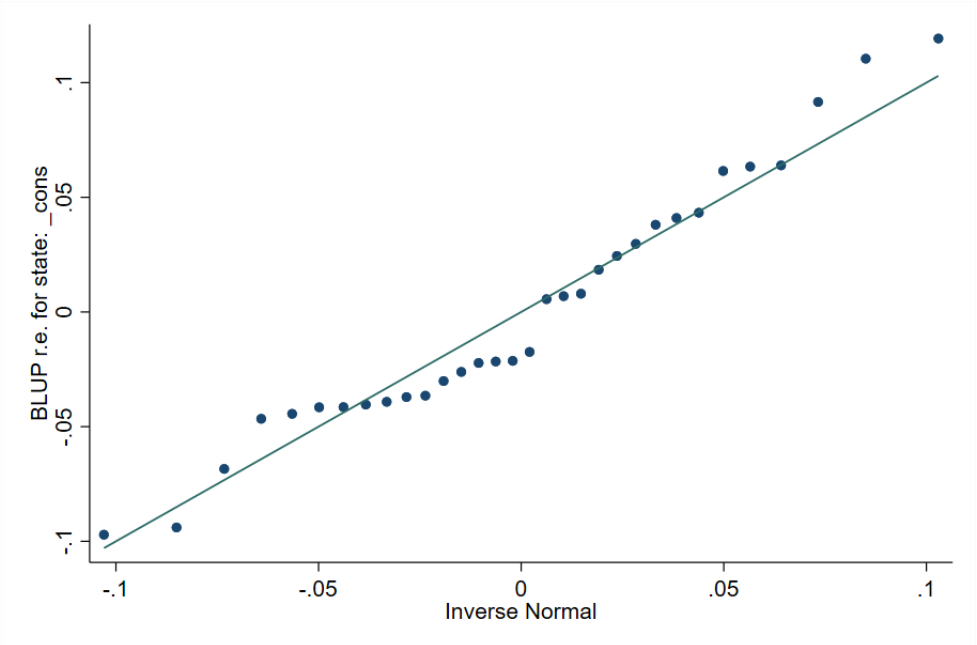
The plots of empirical quantiles compared to theoretical quantiles of the normal distribution (normal Q-Q plots) are helpful to detect deviations from normality. Transformations of the dependent variable are often taken to ensure that empirical quantiles are better aligned to the theoretical ones. As Marhuenda et al. (2017) note, in real life applications, the exact fit to a distribution is hardly met.

Figure 6.2: Sample quantiles of residuals against theoretical quantiles of a Normal distribution



Source: own elaboration from code in Appendix 6.4.2. Deviations from normality can be observed in the figure particularly at the bottom. Marhuenda et al. (2017) notes that, in applications using real data, the exact fit to a distribution is hardly met; it is recommended that practitioners apply several transformations to the dependent variable and select the one that provides the best approximation to the nested-error model's assumptions.

Figure 6.3: Sample quantiles of predicted random effects against theoretical quantiles of a Normal distribution



Source: own elaboration.

6.4 Appendix

6.4.1 Useful definitions and concepts

Definitions are based on Molina and Marhuenda (2015), Ghosh and Rao (1994), Rao and Molina (2015), and Cochran (2007).

- **Small area:** a domain (area) is regarded as small if the domain-specific sample size is not large enough to support direct estimates of adequate precision. That is to say, a small area is any domain of interest for which direct estimates of adequate precision cannot be produced.
- **Large area:** an area or domain where a direct estimate of the parameter of interest for the area has a sufficiently large sample to yield estimates with the desired precision.
- **Domain:** domain may be defined by geographic delimitation/territories (e.g., state, county, school district, health service area), socio-demographic groups, or both (e.g., specific age-sex-race group within a large geographic area) or even other types of sub-populations (e.g., set of firms belonging to a census division by industry group).
- **Direct (domain) estimators:** estimators based only on the domain-specific sample area. These estimators are typically design-unbiased but tend to have low precision in small areas.
- **Indirect (domain) estimators:** estimator that uses information from other areas, under the assumption that there exists some homogeneity relationship between them.
- **Target parameter:** indicator to be estimated. Some examples are population mean, proportion, and rate.
- **Efficiency/precision:** 1/variance when an estimator is unbiased; 1/MSE otherwise.
- **Sampling error:** error from using a sample from the population rather than the whole population.

For a better understanding of statistical inference, the following concepts/definitions are necessary. Definitions and observations are from Molina and García-Portugues (2021) and Rao and Molina (2015). Note that, when dealing with SAE, the bias for each area $c = 1, \ldots, C$ is of interest, not the average bias over all the areas.

- **Unbiased estimator:** if an estimator's bias is equal to zero for all the parameter values. If the expected value of the estimator is equal to the parameter. The estimator $\hat{\vartheta}_c$ of parameter ϑ_c is unbiased if and only if $E(\hat{\vartheta}_c - \vartheta_c) = 0$
- **Estimation error:** the estimation error $\hat{\vartheta}_c - \vartheta_c$ is typically different from zero even if the estimator is unbiased. The bias is the mean estimation error: $Bias(\hat{\vartheta}_c) = E(\hat{\vartheta}_c - \vartheta_c)$.
- **Mean squared (estimation) error (MSE):** is also called mean squared prediction error (MSPE) or prediction mean squared error (PMSE). $MSE(\hat{\vartheta}_c) = E[(\hat{\vartheta}_c - \vartheta_c)^2]$.
- **Standard error (SE) of $\hat{\vartheta}_c$:** is the standard deviation of the sampling distribution of $\hat{\vartheta}_c$.
- **Coefficient of variation (CV) of ϑ:** $cv(\hat{\vartheta}) = s(\hat{\vartheta})/\vartheta$ is the associated standard error of the estimate over the true value of ϑ. It is also known as the relative standard deviation (RSD). The estimated CV is then $\hat{cv}(\hat{\vartheta}) = s(\hat{\vartheta})/\hat{\vartheta}$.
- **Consistency:** when increasing the sample size n, the probability that the estimator differs from the true value ϑ_c by more than ε vanishes for every $\varepsilon > 0$.

6.4.2 Regression diagnostics do-file

The following Stata do-file provides an example of how regression diagnostics are often incorporated in the model selection process for small area estimation. Note that the checks are not exhaustive and only serve as a guide for the practitioner.

```
clear all
set more off
/*==============================================================================
Do-file prepared for SAE Guidelines
- Regression Diagnostics
- authors Paul Corral, Minh Nguyen & Sandra Segovia
*==============================================================================*/

global main       "C:\Users\\`c(username)´\OneDrive\SAE Guidelines 2021"
global section    "$main\4_Model_selection"

global data      "$section\1_data"
global dofile    "$section\2_dofiles"
global figs      "$section\3_figures"

local survey       "$data\survey_public.dta"
*local census      "$main\3_Unit_level\1_data\census_public.dta"

//global with candidate variables.
global myvar rural  lnhhsize age_hh male_hh  piped_water no_piped_water ///
no_sewage sewage_pub sewage_priv electricity telephone cellphone internet ///
computer washmachine fridge television share_under15 share_elderly ///
share_adult max_tertiary max_secondary HID_* mun_* state_*

version 15
set seed 648743

local graphs graphregion(color(white)) xsize(9) ysize(6) msize(small)
*==============================================================================
// End of preamble
*==============================================================================

// First part is just as the model selection dofile

// Load in survey data
use "`survey´", clear

        //Remove small incomes affecting model
        drop if e_y<1

    // Kernel density plot  for e_y with a normal density overlaid
    kdensity e_y, normal `graphs´

    graph export "$figs\kdensity_e_y.png", as(png) replace

        //Log shift transformation to approximate normality
        lnskew0 double bcy = exp(lny)

    // Kernel density plot  for lnywith a normal density overlaid
    kdensity lny, normal `graphs´

    graph export "$figs\kdensity_lny.png", as(png) replace

    // Kernel density plot  for bcy with a normal density overlaid
    kdensity bcy, normal `graphs´

    graph export "$figs\kdensity_bcy.png", as(png) replace
```

```
// removes skeweness from distribution
sum e_y, d
    sum lny, d
    sum bcy, d

    // Data has already been cleaned and prepared. Data preparation and the creation
    // of eligible covariates is of extreme importance.
    // In this instance, we skip these comparison steps because the sample is
    // literally a subsample of the census.
    codebook HID //10 digits, every single one
    codebook HID_mun //7 digits every single one

    //We rename HID_mun
    rename HID_mun MUN
    //Drop automobile, it is missing
    drop *automobile* //all these are missing

    //Check to see if lassoregress is installed, if not install
    cap which lassoregress
    if (_rc) ssc install elasticregress

    //Model selection - with Lasso
    gen lnhhsize = ln(hhsize)
    lassoregress bcy  $myvar [aw=Whh], lambda1se epsilon(1e-10) numfolds(10)
    local hhvars = e(varlist_nonzero)
    global postlasso  `hhvars´

    //Try Henderson III GLS
    sae model h3 bcy $postlasso [aw=Whh], area(MUN)

    //Rename HID_mun
    rename MUN HID_mun

    //Loop designed to remove non-significant covariates sequentially
    forval z= 0.5(-0.05)0.05{
            qui:sae model h3 bcy `hhvars´ [aw=Whh], area(HID_mun)
            mata: bb=st_matrix("e(b_gls)")
            mata: se=sqrt(diagonal(st_matrix("e(V_gls)")))
            mata: zvals = bb´:/se
            mata: st_matrix("min",min(abs(zvals)))
            local zv = (-min[1,1])
            if (2*normal(`zv´)<`z´) exit

            foreach x of varlist `hhvars´{
                    local hhvars1
                    qui: sae model h3 bcy `hhvars´ [aw=Whh], area(HID_mun)
                    qui: test `x´
                    if (r(p)>`z´){
                            local hhvars1
                            foreach yy of local hhvars{
                                    if ("`yy´"=="`x´") dis ""
                                    else local hhvars1 `hhvars1´ `yy´
                            }
                    }
                    else local hhvars1 `hhvars´
                    local hhvars `hhvars1´
            }
    }

    global postsign `hhvars´
```

```
*================================================================================
// Regression diagnostics
*================================================================================

/*This is not a complete diagnostic; it is just a preview, steps & repetitions
depend on the underlying model. Check all vars, check different transformations,
do not forget a model for heteroskedasticity (alpha model) if needed */

rename bcy y

*================================================================================
// Collinearity
*================================================================================

reg y $postsign [aw=Whh],r

    //Check for multicollinearity, and remove highly collinear (VIF>3)
        cap drop touse              //remove vector if it is present to avoid error in next step
        gen touse = e(sample)           //Indicates the observations used
        estat vif                                   //Variance inflation factor
        local hhvars $postsign

    //Remove covariates with VIF greater than 3
        mata: ds = _f_stepvif("`hhvars´","Whh",3,"touse")
        global postvif `vifvar´

        //VIF check
        reg y $postvif [aw=Whh], r
        vif

    // For ilustration
    // Henderson III GLS - model post removal of non-significant
    sae model h3 y $postsign [aw=Whh], area(HID_mun)

    // Henderson III GLS - model post removal of non-significant
        sae model h3 y $postvif [aw=Whh], area(HID_mun)

*================================================================================
// Residual Analysis
*================================================================================

// Linearity
    reg y $postsign, r

    // Augmented component-plus-residual plot; works better than the
    // component-plus-residual plot for identifying nonlinearities in the data.
    acprplot age_hh, lowess lsopts(bwidth(1)) `graphs´

    graph export "$figs\acprplot_age_hh.png", as(png) replace

    // Kernel density plot  for log_age_hh with a normal density overlaid
    kdensity age_hh, normal `graphs´

    graph export "$figs\kdensity_age_hh.png", as(png) replace

    // log transformation
    gen log_age_hh =log(age_hh)

    // Kernel density plot  for log_age_hh with a normal density overlaid
    kdensity log_age_hh, normal `graphs´

    graph export "$figs\kdensity_log_age_hh.png", as(png) replace
```

```
// Normality

    reg y $postvif [aw=Whh],r
    predict resid, resid

    // Kernel density plot  for residuals with a normal density overlaid
    kdensity resid, normal `graphs´

    graph export "$figs\kdensity_resid.png", as(png) replace

    // Standardized normal probability
    pnorm resid , `graphs´

    graph export "$figs\pnorm.png", as(png) replace

    // Quantiles of a variable against the quantiles of a normal distribution
    qnorm resid , `graphs´

    graph export "$figs\qnorm.png", as(png) replace

    // Numerical Test:  Shapiro-Wilk W test for normal data
    swilk resid


// Heteroscedasticity

    reg y $postvif

    // Residuals vs fitted values with a reference line at y=0
    rvfplot , yline(0)  `graphs´

    graph export "$figs\rvfplot_1.png", as(png) replace

    // Cameron & Trivedi´s decomposition of IM-test / White test
    estat imtest

    // Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
    estat hettest


*================================================================================
// Influence Analysis
*================================================================================

// Graphic method < before >

    reg y $postvif [aw=Whh]

    // residuals vs fitted vals
    rvfplot , yline(0)  `graphs´

    graph export "$figs\rvfplot_2.png", as(png) replace


    // normalized residual squared vs leverage
    lvr2plot ,  `graphs´

    graph export "$figs\lvr2plot.png", as(png) replace

// Numerical method

// Step 1

    reg y $postvif
```

```
    // After regression without weights...

    // Calculate measures to identify influential observations
        predict cdist, cooksd      // calculates the Cook´s D influence statistic
        predict rstud, rstudent    // calculates the Studentized (jackknifed) residuals

// Step 2

    reg y $postvif [aw=Whh]

    // Predict leverage and residuals
        predict lev, leverage  // calculates the diagonal elements of the
                           // projection ("hat") matrix
        predict r, resid       // calculates the residuals

    // Save useful locals
    local myN=e(N)              // # observations
        local myK=e(rank)           // rank or k
    local KK =e(df_m)           // degrees of freedom (k-1)

    sum cdist, d
    * return list
        local max = r(max)          // max value
        local p99 = r(p99)                  // percentile 99

 // Step 3

    // For ilustration...
    // We have influential data points...
    reg lny $postvif if cdist<4/`myN´ [aw=Whh]
    reg lny $postvif if cdist<`p99´   [aw=Whh]
        reg lny $postvif if cdist<`max´   [aw=Whh]

    // Identified influential / outliers observations
    gen nogo = abs(rstud)>2 & cdist>4/`myN´ & lev>(2*`myK´+2)/`myN´

    count if nogo==1        // these are the obs that we want to eliminate


// Graphic method < after >

    reg y $postvif [aw=Whh] if nogo==0

    // residuals vs fitted vals
    rvfplot , yline(0)  `graphs´

    graph export "$figs\rvfplot_2_after.png", as(png) replace


    // normalized residual squared vs leverage
    lvr2plot ,        `graphs´

    graph export "$figs\lvr2plot_after.png", as(png) replace


*================================================================================
// Model Specification tests
*================================================================================

   reg y $postvif

// Wald test for ommited vars < will compare with previous regression>
   boxcox y $postvif, nolog // Box - Cox model
```

```
// Functional form of the conditional mean
    reg y $postvif

    estat ovtest // performs regression specification error test (RESET) for omitted variables

    linktest //performs a link test for model specification

// Omnibus tests + Heteroscedasticity tests
    reg y $postvif

    estat imtest   // Cameron & Trivedi´s decomposition of IM-test / White test

    estat hettest // Breusch-Pagan / Cook-Weisberg test for Heteroscedasticity

*================================================================================
// Diagnostics for random effects
*================================================================================

// Multilevel mixed-effects linear regression
    mixed y $postvif || state:, reml

    predict res, residual

    predict eta, reffects

    bysort state: gen first=_n

    // For state == 1
    qnorm eta if first==1 , `graphs´

    graph export "$figs\qnorm_mixed_1.png", as(png) replace

// Quantiles of a variable against the quantiles of a normal distribution
    qnorm res , `graphs´

    graph export "$figs\qnorm_mixed.png", as(png) replace
```

References

Cameron, A Colin and Pravin K Trivedi (2005). *Microeconometrics: Methods and Applications*. Cambridge university press. ISBN: 978-0-521-84805-3.

Cochran, William G (2007). *Sampling Techniques*. John Wiley & Sons.

Cook, R Dennis (1977). "Detection of Influential Observation in Linear Regression". In: *Technometrics* 19.1, pp. 15–18.

Corral, Paul, Kristen Himelein, Kevin McGee, and Isabel Molina (2021). "A Map of the Poor or a Poor Map?" In: *Mathematics* 9.21. ISSN: 2227-7390. DOI: 10.3390/math9212780. URL: https://www.mdpi.com/2227-7390/9/21/2780.

Corral, Paul, Isabel Molina, and Minh Cong Nguyen (2021). "Pull Your Small Area Estimates up by the Bootstraps". In: *Journal of Statistical Computation and Simulation* 91.16, pp. 3304–3357. DOI: 10.1080/00949655.2021.1926460. URL: https://www.tandfonline.com/doi/abs/10.1080/00949655.2021.1926460.

Elbers, Chris, Jean O Lanjouw, and Peter Lanjouw (2003). "Micro–level Estimation of Poverty and Inequality". In: *Econometrica* 71.1, pp. 355–364.

Elbers, Chris, Jean Olson Lanjouw, and Peter Lanjouw (2002). "Micro-level Estimation of Welfare". In: *World Bank Policy Research Working Paper* 2911.

Ghosh, Malay and JNK Rao (1994). "Small Area Estimation: An Appraisal". In: *Statistical science* 9.1, pp. 55–76.

Groll, Andreas (2017). "glmmLasso: Variable Selection for Generalized Linear Mixed Models by L1-penalized Estimation". In: *R package version* 1.1, p. 25.

Groll, Andreas and Gerhard Tutz (2014). "Variable Selection for Generalized Linear Mixed Models by L 1-penalized Estimation". In: *Statistics and Computing* 24.2, pp. 137–154.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2013). *An Introduction to Statistical Learning*. Vol. 112. Springer. ISBN: 978-1-0716-1418-1. URL: https://link.springer.com/book/10.1007/978-1-0716-1418-1?noAccess=true.

Lahiri, Partha and Jiraphan Suntornchost (2015). "Variable Selection for Linear Mixed Models with Applications in Small Area Estimation". In: *Sankhya B* 77.2, pp. 312–320.

Marhuenda, Yolanda, Isabel Molina, Domingo Morales, and JNK Rao (2017). "Poverty Mapping in Small Areas under a Twofold Nested Error Regression Model". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180.4, pp. 1111–1136. DOI: 10.1111/rssa.12306.

Möhring, Katja and Alexander Schmidt-Catran (2013). *MLT: Stata Module to Provide Multilevel Tools*. URL: https://econpapers.repec.org/software/bocbocode/s457577.htm.

Molina, Isabel and Eduardo García-Portugues (2021). *A First Course on Statistical Inference*. Accessed: 2010-06-22. Bookdown.org. URL: https://bookdown.org/egarpor/inference/.

Molina, Isabel and Yolanda Marhuenda (2015). "Sae: An R Package for Small Area Estimation". In: *The R Journal* 7.1, pp. 81–98.

Molina, Isabel and JNK Rao (2010). "Small Area Estimation of Poverty Indicators". In: *Canadian Journal of Statistics* 38.3, pp. 369–385.

Pfeffermann, Danny (2013). "New Important Developments in Small Area Estimation". In: *Statistical Science* 28.1, pp. 40–68.

Rao, JNK and Isabel Molina (2015). *Small Area Estimation*. 2nd. John Wiley & Sons.

Robinson, George K (1991). "That BLUP Is a Good Thing: The Estimation of Random Effects". In: *Statistical science* 6.1, pp. 15–32.

Tzavidis, Nikos, Li-Chun Zhang, Angela Luna, Timo Schmid, and Natalia Rojas-Perilla (2018). "From Start to Finish: A Framework for the Production of Small Area Official Statistics". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 181.4, pp. 927–979.

UCLA: Statistical Consulting Group (2022). *Regression with Stata Chapter 2 – Regression Diagnostics*. URL: https://stats.oarc.ucla.edu/stata/dae/robust-regression/.

West, Brady T, Kathleen B Welch, and Andrzej T Galecki (2014). *Linear Mixed Models: A Practical Guide Using Statistical Software*. 2nd. Chapman and Hall/CRC. DOI: 10.1201/b17198. URL: https://www.taylorfrancis.com/books/mono/10.1201/b17198/linear-mixed-models-brady-west-kathleen-welch-andrzej-galecki.