University of Applied Sciences and Arts Northwestern Switzerland
School of Business

Robust Estimation with Survey Data

The `robsurvey` package

Tobias Schoch

uRos2025 Conference

November 26, 2025, Bucharest

Outline

- Robustness in finite population sampling
- Basic robust estimators
- Robust regression
- Robust model-assisted estimation (GREG)

Notes

- The `robsurvey` package complements/extends the functions in the `survey` (Lumley, 2010, 2024) package.
- The package is joint work with Beat Hulliger.

PART 1

Robustness in the Context of Finite Population Sampling

Definition (Representative vs. nonrepresentative outliers; Chambers, 1986)

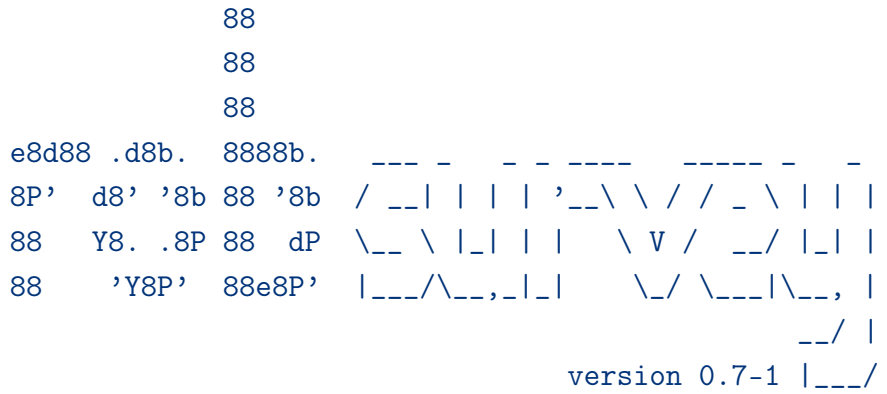
- **Representative outlier:** An extreme
  - but correct value that
  - is thought to represent other population units similar in value.
- **Nonrepresentative outlier:** An extreme observation
  - whose value is either deemed erroneous or
  - unique in the sense that there is no other unit like it in the population.

Clearly, an **erroneous value** is (usually) treated at the **editing stage**.

Length-Of-Stay (LOS) Data

- Patients in hospital inpatient care (population size: $N = 2\,479$)
- **Simple random sample** without replacement of size $n = 71$
- Variable of interest: **length of stay** (LOS, number of days)

```
> library("robsurvey", quietly = TRUE)
> head(losdata)
  los   weight  fpc
1  10 34.91549 2479
2   7 34.91549 2479
3  21 34.91549 2479
```




```
type: package?robsurvey to learn more
use:  library(robsurvey, quietly = TRUE) to suppress the
      start-up message
```

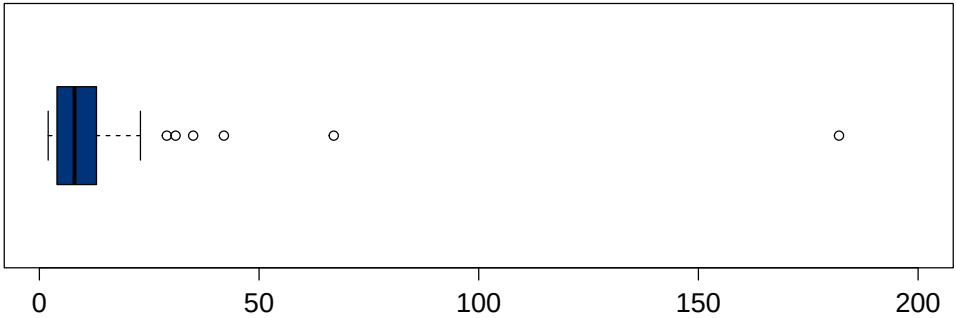


Figure: Length of stay in inpatient care (days, mean: 13.03, max. 182)

```
> attach(losdata)
> weighted_mean(los, weight)
1] 13.02817
```

Definition (Influential observation; Lee, 1995)

The **influence** of an observation varies depending on the type of estimator.

- For the Horvitz–Thompson (HT) estimator of the population y-total, $\widehat{t}_y$, an **influential observation** is an extreme value based on its expanded value.

$$\widehat{t}_y = \sum_{i \in s} \underbrace{w_i y_i}_{\text{expanded value}}, \qquad w_i = \frac{1}{\pi_i},$$

  where π_i denotes the sample inclusion probability.

- An influential observation for the HT-estimator **does not necessarily** have to be influential for **other types of estimators** (e.g., ratio estimator) and vice versa.

- A general way to **robustify** the HT-estimator is to
  - introduce a **robustness weight**, u_i, taking values in the interval $[0, 1]$,
  - define

$$\widehat{t}_y^{\text{rob}} = \sum_{i \in s} u_i \, w_i \, y_i.$$

- The u_i's are allowed to depend on y_i and/or w_i.
- For **example**,
  - u_i is a function of the expanded value, $w_i y_i$, (e.g., Dalén's estimators)
  - u_i depends only the value y_i.

Important takeaways

- The robustness weights u_i
  - **differ** from **variable to variable** (of interest),
  - are defined for a particular **type** of estimator (here, HT-estimator).

  $\Rightarrow$ Thus, it is not meaningful to "fix" the robustness weights at the design or editing stage.

- **Also**, (many) robust estimators require the specification of a robustness **tuning constant** (degree of robustness).

Why robustness?

- Let's consider estimating the mean LOS in hospital (the data contain representative outliers).
- The HT-estimator is **design-unbiased** for the population y-total.
- The estimator $\widehat{t}_y^{\text{rob}}$ is biased, right? **Yes, it is biased.**
- However, in terms of **mean square error**,

$$\text{MSE}_p\left(\widehat{t}_y^{\text{rob}}\right) = \text{Var}_p\left(\widehat{t}_y^{\text{rob}}\right) + \text{Bias}_p\left(\widehat{t}_y^{\text{rob}}\right)^2,$$

  estimator $\widehat{t}_y^{\text{rob}}$ **can be** superior to the HT-estimator.

- **Bias-variance trade-off**
  - The unbiasedness of the HT estimator comes at the cost of a large variance.
  - Robust estimators attempt to **balance** bias and variance.
- An **out-of-the-box** robust estimator is usually not superior.

  $\Rightarrow$ Statisticians must pay attention to **appropriate tuning**.

Frequently Raised Objections to Robustness

"At our office, we don't need robust estimators because…

1. we use **stratified sampling designs** (or pps designs),
2. we do **careful editing**,
3. we **restrict the weights** in calibration (prevent excessively large weights, see e.g., `survey::trimWeights`)."

⇒ While I agree that these **precautions are important**, they are **not enough**.

⇒ Rather than replacing any of the precautions, robust estimators **complement** them.

PART 2

Basic Robust Estimators

Two “Flavors” of Basic Robust Estimators

- **Bare-bone** functions: `weighted_mean` and `weighted_total`
- **Survey** methods: `svymean` and `svytotal`

followed by (suffix):

`_winsorized()` or `_k_winsorized()`	winsorization
`_trimmed()`	trimming
`_huber()` or `_tukey()`	M-estimators
`_dalen()`	Dalén’s estimators (weight reduction)

For example, `weighted_mean_winsorized()`

Bare-Bone Functions

For example, the **one-sided 2% winsorized** weighted mean of LOS

```
> weighted_mean_winsorized(los, weight, LB = 0, UB = 0.98)
```

```
[1] 11.40845
```

- Lower bound: `LB = 0`
- Upper bound: `UB = 0.98` $\Rightarrow$ only 2% largest observations are winsorized
- Return value: scalar
- Minimalistic function

Survey Methods

```
> library("survey")
> dn <- svydesign(ids = ~1, fpc = ~fpc, weights = ~weight,
                  data = losdata)
> svymean_winsorized(~los, dn, LB = 0, UB = 0.98)
      mean   SE
los  11.41  1.5
```

- Requirement: `survey` package (Lumley, 2010, 2024)
- Variance estimation (standard error: SE) based on `survey` package
- Return value: Instance of class `svystat_rob`

```
> m <- svymean(~los, dn)
confint(m)
      2.5%   97.5%
los    7.8    18.3
```

Winsorized estimator

```
> m_rob <- svymean_winsorized(~los, dn, LB = 0, UB = 0.98)
confint(m_rob)
      2.5%   97.5%
los    8.5    14.3
```

Comparison (gains in relative efficiency)

```
> 1 - mse(m_rob) / mse(m)
[1] 0.318095
```

Utility Methods

`coef()`	extracts estimates	$\surd$
`vcov()`	variance-covariance matrix	$\surd$
`SE()`	standard error	$\surd$
`summary()`	shows summary of fitted model	
`mse()`	computes mean square error	
`residuals()`	extracts residuals	
`fitted()`	computes fitted values	
`robweights()`	robustness weights (M-estimators)	
`scale()`	estimate of scale (M-estimators)	

Note: $\surd$ indicates methods that are also available in the `survey` package.

What more?

- **M-estimators**
  - Huber and Tukey ψ-function
  - Interface to add other ψ-functions: see `doc_psifunction.html`
- Other functions: `weighted_mad()`, `weighted_IQR()`, etc.
- **Vignettes**
  - Basic Robust Estimators
  - Robust Horvitz–Thompson Estimator

PART 3

Robust Weighted Regression

Regression

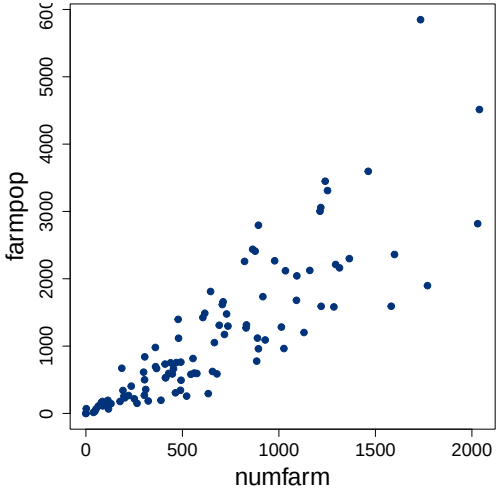
- **Simple random sample** of $n = 100$ counties in the U.S.
- Population: $N = 3\,141$, data: Lohr (1999; U.S. Bureau of the Census)

```
> head(counties[, c(2, 6, 7, 9, 10)], 2)
    county  farmpop  numfarm  weights   fpc
1 Escambia      531      414    31.41  3141
2 Marshall     1592    15824    31.41  3141
```

where

`farmpop`	farm population,	`weights`	weights,
`numfarm`	number of farms,	`fpc`	population size

Regression: Weighted Least Squares



- **Model:** `farmpop ~ -1 + numfarm`
- **Variance:** heteroscedasticity
- **Sampling design:**

```
dn <- svydesign(ids = ~1,
    fpc = ~fpc,
    weights = ~weights,
    data = subset(counties,
    numfarm > 0))
```

- **Weighted least squares**

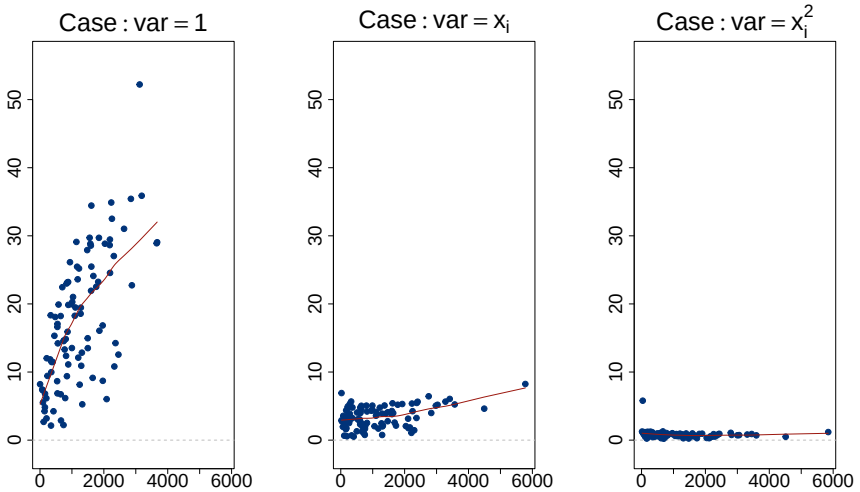


Figure: Regression diagnostic plot: $\sqrt{\text{abs. residuals}}$ vs. fitted values

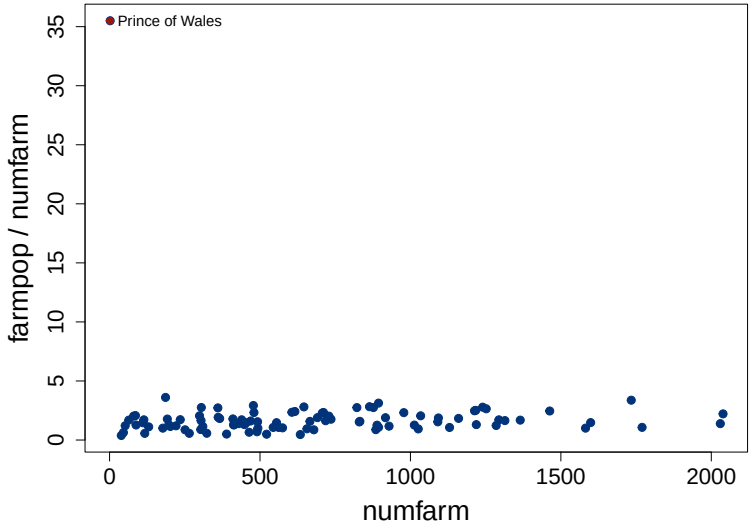


Figure: Heteroscedasticity diagnostic plot

```
> svyreg(farmpop ~ -1 + numfarm, dn, var = ~I(numfarm^2))
Weighted least squares

Call:
svyreg(formula = farmpop ~ -1 + numfarm, design = dn,
    var = ~I(numfarm^2))

Coefficients:
numfarm
   1.97

Scale estimate: 3.477
```

Robust Regression: *M*- and *GM*-Estimators

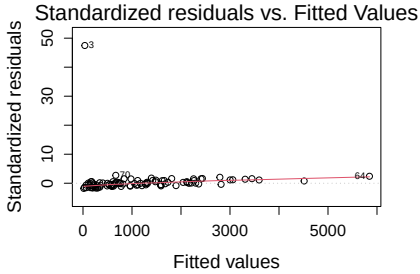
Function `svyreg` followed by

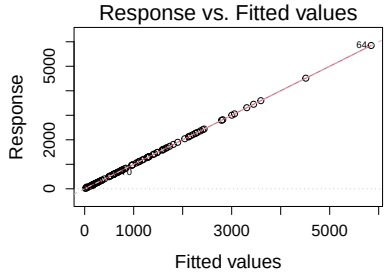
```
 _huberM(formula, design, k, var = NULL, ...)
_huberGM(formula, design, k,
         type = c("Mallows", "Schweppe"),
         xwgt, var = NULL, ...)
```

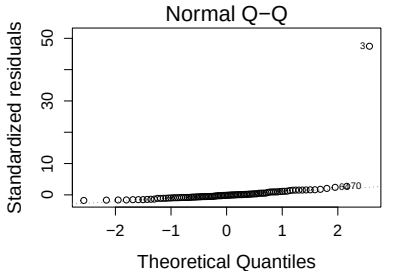
- `k`: robustness tuning constant of Huber ψ-function
- `type`: Mallows or Schweppe *GM*-estimator
- `xwgt`: downweight high-leverage observations
- **Also** `_tukeyM()` and `_tukeyGM()` with Tukey ψ-function

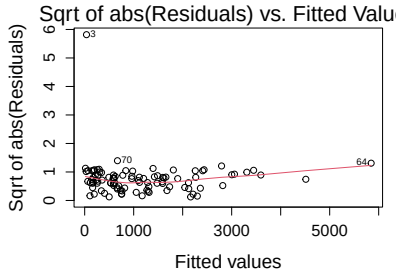
```
> svyreg_huberM(farmpop ~ -1 + numfarm, dn, k = 5,
                var = ~I(numfarm^2))
Survey regression M-estimator (Huber psi, k = 5)
Call:

svyreg_huberM(formula = farmpop ~ -1 + numfarm, k = 5,
              design = dn_positive, var = ~I(numfarm^2))

IRWLS converged in 3 iterations

Coefficients:
numfarm
  1.661
```

```
Scale estimate: 0.7128 (weighted MAD)
```


Standardized residuals vs. Fitted Values
Standardized residuals
Fitted values
0
10
30
50
0
1000
3000
5000
3
70
64


Response vs. Fitted values
Response
Fitted values
0
2000
5000
0
1000
3000
5000
64


Normal Q−Q
Standardized residuals
Theoretical Quantiles
0
10
30
50
−2
−1
0
1
2
3
70


Sqrt of abs(Residuals) vs. Fitted Values
Sqrt of abs(Residuals)
Fitted values
0
1
2
3
4
5
6
0
1000
3000
5000
3
70
64

Three Modes of Inference

- **Inference** under the model

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\theta} + \sqrt{v_i} e_i, \qquad i \in U$$

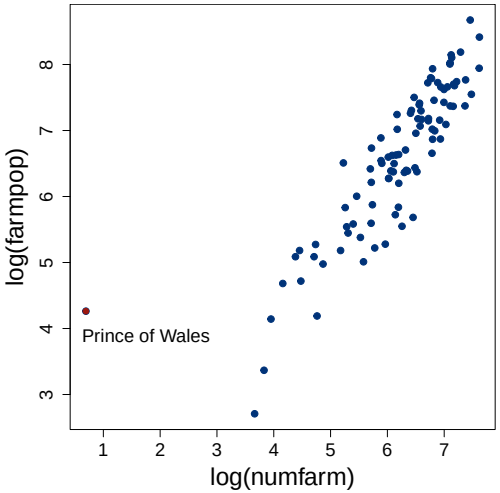
  (under standard assumptions on the random variables e_i)

- **Parameter** of interest
  - θ super-population parameter
  - θ_N census parameter, finite-population parameter
  - $\widehat{\theta}_n$ sample-based estimator (ignores sampling design)

Literature: Rubin-Bleuer and Schriopu-Kratina (2005), Binder and Roberts (2009)

- The `summary()` method has a **mode** argument:
  - `mode = "design"` $\Rightarrow$ **design-based**: estimate θ_N
  - `mode = "model"` $\Rightarrow$ **model-based**: estimate θ (ignores sampling design)
  - `mode = "compound"` $\Rightarrow$ **compound design-model**: estimate θ
- Additional **methods**: `coef()`, `vcov()`, `residuals()`, `fitted()`, `plot()`, etc.

Alternative Model: Log-Log Specification



- **Model:**
  $\texttt{log(farmpop)} \sim \texttt{log(numfarm)}$
- **Variance:** homoscedastic
- **Outlier** $\Rightarrow$ Robust regression

PART 4

Model-Assisted Estimation (Generalized Regression Estimation)

- **Goal**: estimate **total farm population**, t_y, (counties data)
- **Auxiliary information**: total number of farms, t_x, is known (register)
- **Model**

$$\underbrace{\text{farmpop}_i}_{y_i} = \theta \cdot \underbrace{\text{numfarm}_i}_{x_i} + \sqrt{v_i} e_i, \qquad i \in U$$

(under standard assumptions on the random variables e_i; $v_i = \text{numfarm}_i^2$)

$\Rightarrow$ (Robust) estimate $\widehat{\theta}$ (see above)

- Two approaches
  - **Model-based** (BLUP) prediction (Valliant et al., 2000)
  - **Model-assisted** (GREG) estimation (Särndal et al., 1992)

- The approaches differ in terms of the **emphasis** placed on the model.

  $\Rightarrow$ High belief in the validity of the model $\Rightarrow$ BLUP

- **Model-assisted** estimator (GREG, Särndal et al., 1992)

$$\hat{t}_y^{\text{greg}} = t_x \hat{\theta} + \frac{N}{n} \sum_{i \in s} (y_i - x_i \hat{\theta})$$

  - Model-unbiased
  - **Approximately design-unbiased** (ADU), i.e., bias $\to 0$ as $n, N \to \infty$
  - Has some robustness w.r.t. model misspecification

Step 1: Estimate the regression parameter (θ)

```
> theta <- svyreg_huberM(farmpop ~ -1 + numfarm, dn, k = 5,
                         var = ~I(numfarm^2))
```

Step 2: GREG-estimate of the total (given θ)

```
> svytotal_reg(theta, totals = 2087759, type = "ADU")
```

```
            total      SE
farmpop  3633302  178352
```

Results of BLUP not shown (affected by outlier county “Prince of Wales”)

Arguments of `svytotal_reg()` and `svymean_reg()`

- **Method selection** (argument `type`):
  - `type = "projective"` (not robust)
  - `type = "ADU"` (**default**, “standard” GREG, not robust)
  - `type = "huber"`
  - `type = "tukey"`
  - `type = "lee"`
  - `type = "BR"` (Beaumont–Rivest; Ronchetti–Welsh and Chambers)
  - `type = "duchesne"`
- **Tuning constant** *k*

PART 5

Summary and Outlook

Summary

- The package **offers more** (was not covered in the talk)
  - Huber proposal-2 estimator
  - Adaptive estimation (minimum estimated risk estimator)
  - Tukey's weighted line
  - ...
- Learn more
  - **CRAN** webpage of `robsurvey`
  - **GitHub** `tobiasschoch/robsurvey`
  - **4 vignettes**

Takeaway

- **2 "flavors"** of functions
  - Bare-bone functions
  - Survey methods (`survey` package required) $\Rightarrow$ variance

    $\Rightarrow$ **What is missing? What methods do you need?**

I’m ready to take your questions!

Literature

- Beaumont, J. F. and Rivest, L. P. (2009). Dealing with outliers in survey data, in: *Sample Surveys: Theory, Methods and Inference*, ed. by Pfeffermann, D. and Rao, C. R., Amsterdam: Elsevier, vol. 29A of *Handbook of Statistics*, Chap. 11, 247–280.
- Binder, D. A. and Roberts, G. (2009). Design- and Model-Based Inference for Model Parameters, in: *Sample Surveys: Inference and Analysis*, ed. by Pfeffermann, D. and Rao, C. R. Volume 29B of *Handbook of Statistics*, Amsterdam: Elsevier, Chap. 24, 33–54.
- Chambers, R. (1986). Outlier Robust Finite Population Estimation. *Journal of the American Statistical Association*, 81, 1063–1069
- Hulliger, B. (1995). Outlier Robust Horvitz–Thompson Estimators. *Survey Methodology* 21, 79–87.
- Lee, H (1995). Outliers in Business Surveys, in: *Business Survey Methods*, ed. by Cox, B. G., Binder, D. A., Chinnappa, B. N., Christianson, A., Colledge, M. J., and Kott, P. S., New York: John Wiley & Sons, Chap. 26, 503–526.

- Lohr, S. L. (2019). *Sampling: Design and Analysis*, Boca Raton (FL): Chapman and Hall/CRC, 2nd ed.
- Lumley, T. (2024). survey: Analysis of Complex Survey Samples. R package version 4.1-1. URL `https://CRAN.R-project.org/package=survey`
- Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R*, Hoboken (NJ): John Wiley & Sons.
- Rubin-Bleuer, S. and Schiopu-Kratina, I. (2005). On the Two-phase framework for joint model and design-based inference. *The Annals of Statistics* 33, 2789–2810.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, New York: Springer.
- Valliant, R., Dorfman, A. H. and Royall, R. M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*, Hoboken (NJ): John Wiley & Sons.