University of Applied Sciences Northwestern Switzerland
School of Business

robsurvey: Robust Survey Statistics Estimation

useR! 2022 Conference

```
> library("robsurvey")

                    88
                    88
                    88
    e8d88 .d8b.  8888b.   ___ _   _ _ _____   _____ _   _
    8P'  d8' '8b 88 '8b  / __| | | | '__\ \ / / _ \ | | |
    88   Y8. .8P 88  dP  \__ \ |_| | |   \ V /  __/ |_| |
    88    'Y8P'  88e8P'  |___/\__,_|_|    \_/ \___|\__, |
                                                     __/ |
                                          version 0.3 |___/
```

Tobias Schoch (joint work with Beat Hulliger), June 23, 2022

Outline

- Basic robust estimators
- Robust weighted regression
- Summary & outlook

Requirement

```
> library("robsurvey", quietly = TRUE)
> data("losdata")
> data("counties")
```

PART 1

Basic Robust Estimators

The `losdata` data

- **Length of stay** (LOS) in hospital (days per year)
- **Sample** of $n = 71$ patients (population size $N = 2479$)

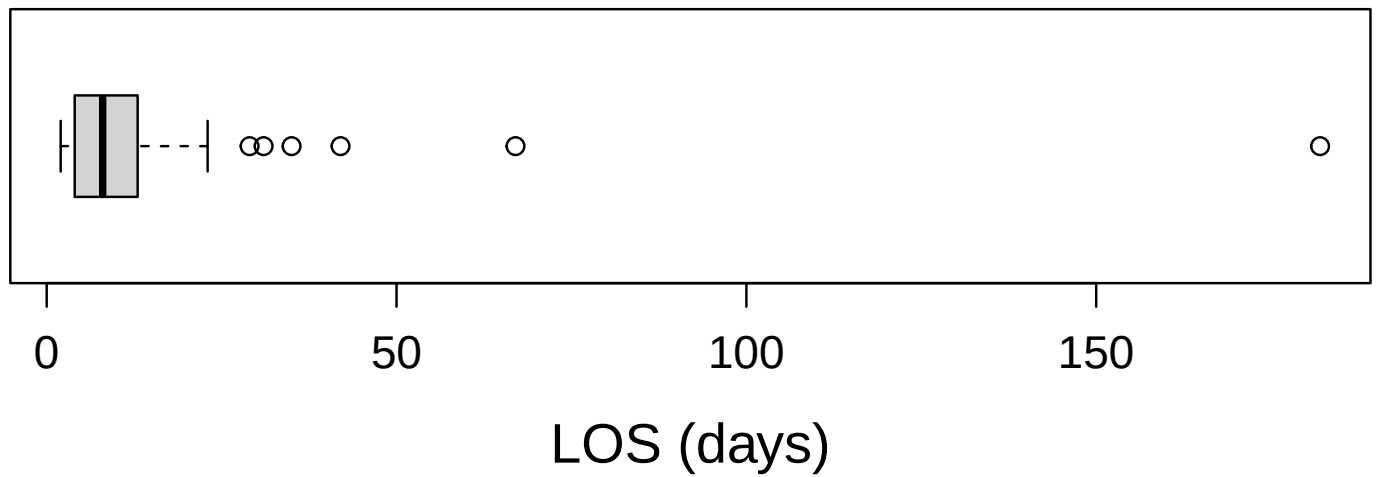
```
> head(losdata, 3)

  los    weight  fpc
1  10 34.91549 2479
2   7 34.91549 2479
3  21 34.91549 2479
```

where

`los`	length of stay in hospital
`weight`	sampling weight
`fpc`	population size (finite population correction)

The `losdata` data (cont'd)

- **Distribution** of `los`
  - Average: 13.03 (days)
  - Maximum: 182



Two "flavors" of basic robust estimators

- **bare-bone** functions: `weighted_mean` and `weighted_total`
- **survey** methods: `svymean` and `svytotal`

followed by (suffix)

`_winsorized()` and `_k_winsorized()`: winsorization

`_trimmed()`: trimming

`_huber()` and `_tukey()`: M-estimators

`_dalen()`: Dalén's estimators: censoring of value and weight

E.g., `weighted_mean_winsorized()`

Bare-bone functions

The 2% one-sided **winsorized weighted mean** is

```
> attach(losdata)
> weighted_mean_winsorized(los, weight, LB = 0,
                           UB = 0.98)
[1] 11.40845
```

- Lower bound `LB` $= 0$
- Upper bound `UB = 0.98` $\Rightarrow$ largest 2% of the observations are winsorized
- Return value: estimate (scalar)
- Light-weight, minimalistic, bare-bone
- Useful for package developers

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
```

- Computes **standard errors (SE)** using functionality of the `survey` package (Lumley, 2010, 2021)
- Return value: an instance of class `svystat_rob`

Survey methods (cont'd)

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?

- Implemented in the `C` language
- M-**estimators**
  - Huber and Tukey ψ-function
  - Interface to add other ψ-functions: see `doc_psifunction.html`
- **Vignettes**
  - Basic Robust Estimators
  - Robust Horvitz–Thompson Estimator

PART 2

Robust Weighted Regression

Regression

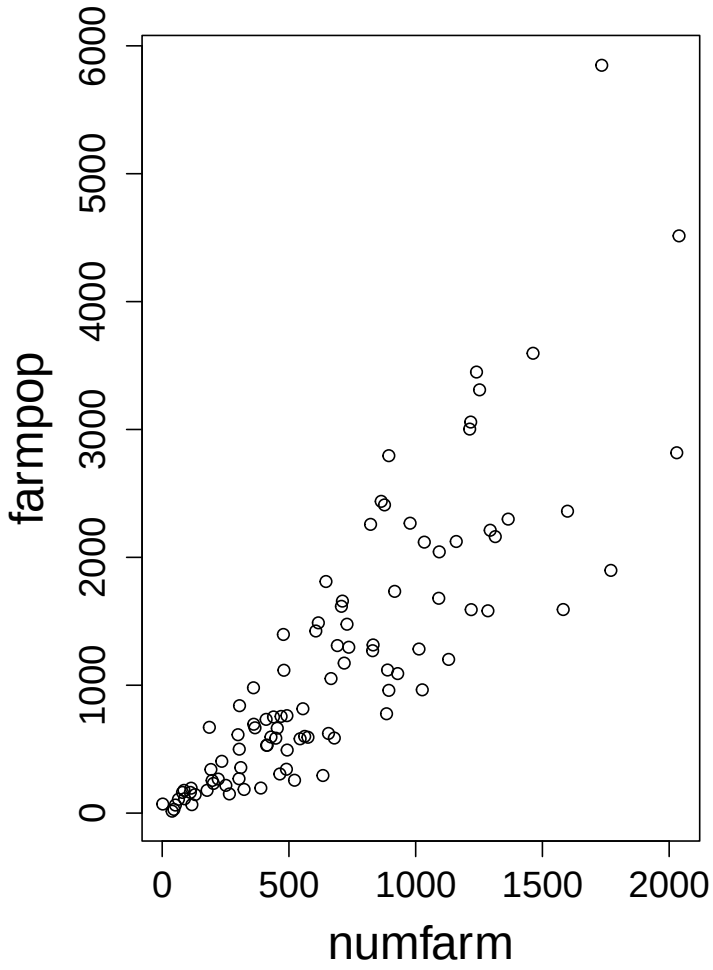
- Simple random sample of $n = 100$ counties in the U.S.
- Population: $N = 3141$ counties
- Data: Lohr (1999)

```
> 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` $\sim$ `numfarm`
- **Variance:** (heteroscedasticity)
  `var = sqrt(numfarm)`
- **Sampling design:**

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

- **Weighted least squares**

Regression: Weighted least squares (cont'd)

```
> svyreg(farmpop ~ numfarm, dn, var = ~sqrt(numfarm))

Weighted least squares

Call:
svyreg(formula = farmpop ~ numfarm, design = dn,
       var = ~vi)

Coefficients:
(Intercept)      numfarm
    -53.998        1.839

Scale estimate: 99.25
```

Regression: Weighted least squares (cont'd)

- **Methods**: `coef()`, `residuals()`, `fitted()`, `plot()`, etc.
- **Inference** under the model $y_i = \boldsymbol{x}_i^T \boldsymbol{\theta} + \sigma \sqrt{v_i} e_i,\ i \in U$
  - $\boldsymbol{\theta}$: super-population parameter
  - $\boldsymbol{\theta}_N$: census parameter, finite-population parameter
  - $\widehat{\boldsymbol{\theta}}_n$: sample-based estimator
- **Design-based**: estimate $\boldsymbol{\theta}_N$

```
summary(..., mode = "design")
vcov(..., mode = "design")
```

Regression: Weighted least squares (cont'd)

- **Model-based**: estimate $\boldsymbol{\theta}$ (ignore sampling design)

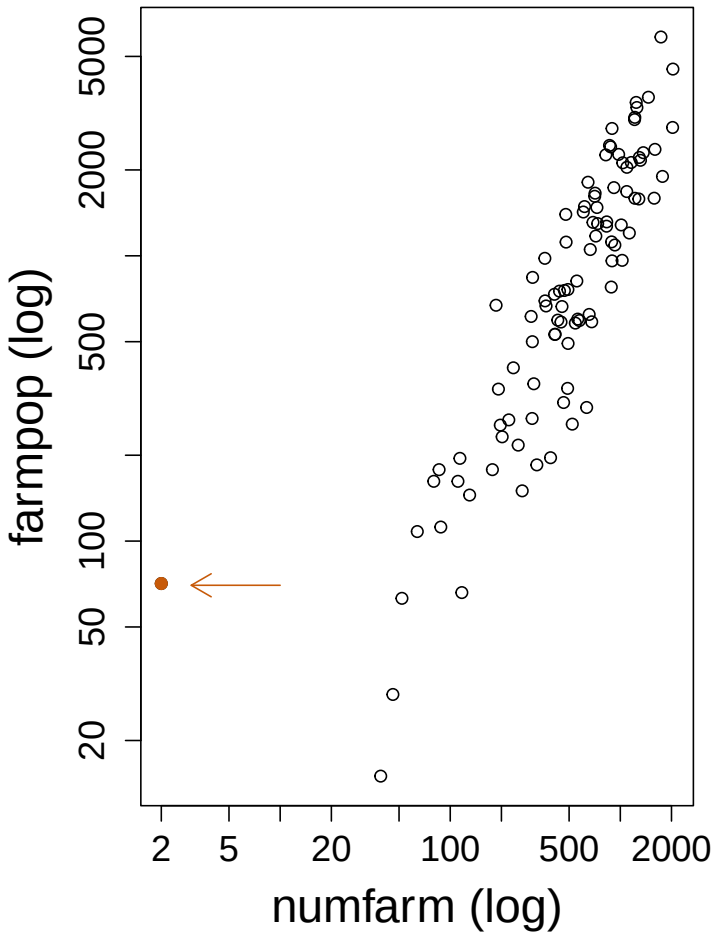
  `mode = "model"` in `summary()` and `vcov()`

- **Compound design-model**: estimate $\boldsymbol{\theta}$

  `mode = "compound"` in `summary()` and `vcov()`

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

Robust regression



- **Model:** `log(farmpop)` $\sim$ `log(numfarm)`
- **Variance:** homoscedastic
- **Outlier** $\rightarrow$ Robust regression

Robust regression: Implementation

Language		M	GM	Reference
with weighting				
`MASS`	R	√	–	Venables and Ripley (2002)
`robust`	R	√	–	Wang et al. (2022)
`robustbase`	R	√	–	Mächler et al. (2021)
`robstat`	Stata	√	–	Jann et al. (2018)
`robustreg`	SAS	√	–	SAS Institute (2022)
`robsurvey`	**R**	√	√	**this talk**
without weighting				
`robeth`	R	√	√	Marazzi (1993, 2020)
NAG	C	√	√	NAG (2022)
`GSL`	C	√	–	Galassi et al. (2019)
`[?]regress`	Stata	√	–	Verardi and Croux (2009)
`rreg`	Stata	√	–	StataCorp (2022)
`robustfit`	MATLAB	√	–	The Math Works (2022)

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

Robust regression: *M*- and *GM*-estimators (cont'd)

- **Utility methods:** `coef()`, `plot()`, etc.
- **Inference:** `summary()` and `vcov()` (3 modes of inference)
- Vignette: **Robust Weighted Regression**

PART 3

Summary and Outlook

Summary and outlook

- Package: **52 exported functions** (+ 21 S3-methods)
  - Robust generalized regression (GREG) estimator
  - Tukey's weighted line
  - ...
- Take away message: **2 "flavors"** of functions
  - Bare-bone functions
  - Survey methods (`survey` package required) $\Rightarrow$ variance
- **What is missing? What methods do you need?**

Summary and outlook (cont'd)

- Where can I find this slide deck?
  - **CRAN** webpage of `robsurvey`
  - **Link** to **GitHub** `tobiasschoch/robsurvey`
  - On GitHub: **folder:** `/inst/doc/useR_2022_conference`

I'm ready to take your questions!

Appendix: Datasets

- **counties**: Lohr SL (1999). *Sampling: Design and Analysis*. Pacific Grove (CA): Duxbury Press, Appendix C.
- **losdata**: Ruffieux C, Paccaud F, Marazzi A (2000). Comparing rules for truncating hospital length of stay. *Casemix Quarterly* 2.

Appendix: Software

- **GSL** Galassi M, Davies J, Theiler J, Gough B, Jungman G, Alken P, Booth M, Rossi F, Ulerich R (2019). GNU Scientific Library (release 2.6). 3rd edition.
- **MASS** Venables WN and Ripley BD, 2002, Modern Applied Statistics with S. 4th edition, New York: Springer-Verlag.
- **NAG** NAG (2022). The NAG Library for C. The Numerical Algorithms Group (NAG), Oxford. C library mark 28.3.
- **[?]regress** Verardi V, Croux C (2009). Robust Regression in Stata. *The Stata Journal* 9, 439–453. (Note: [?] is a wildcard for `m`, `s`, or `mm`; the methods thus read `mregress`, etc.)
- **robeth** Marazzi A, 2020, robeth: R Functions for Robust Statistics. R package version 2.7-6.
- **robstat** Jann B, Verardi V, Vermandele C, 2018. ROBSTAT: Stata module to compute robust univariate statistics. Statistical Software Components, Boston College Department of Economics.

Appendix: Software (cont'd)

- **robust** Wang J, Zamar R, Marazzi A, Yohai V, Salibián-Barrera M, Maronna R, Zivot E, Rocke D, Martin D, Mächler M, Konis K (2022). robust: A Port of the S-PLUS "Robust Library". R package version 0.7-0.
- **robustbase** Mächler M, Rousseeuw P, Croux C, Todorov V, Ruckstuhl A, Salibián-Barrera M, Verbeke T, Koller M, Conceicao ELT, Anna di Palma M (2021). robustbase: Basic Robust Statistics. R package version 0.93-9.
- **robustfit** The Math Works, Inc. (2022). MATLAB. Version R2022a.
- **robustreg** SAS Institute, Inc. (2020). SAS/STAT Software. SAS Institute Inc., Cary. Version 15.2.
- **rreg** StataCorp (2022). Stata Statistical Software. StataCorp LLC, College Station. Release 17.
- **survey** Lumley T (2021). survey: Analysis of Complex Survey Samples. R package version 4.1-1.

Appendix: Literature

- **Beaumont JF and Rivest LP (2009)**. Dealing with outliers in survey data, in *Sample Surveys: Theory, Methods and Inference*, ed. by Pfeffermann D and Rao CR, Amsterdam: Elsevier, vol. 29A of Handbook of Statistics, chap. 11, 247–280.
- **Binder DA 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.
- **Hampel FR, Ronchetti EM, Rousseeuw PJ and Stahel WA (1986)**. *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley and Sons.
- **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 BG, Binder DA, Chinnappa BN, Christianson A, Colledge MJ, Kott PS, New York: John Wiley and Sons, chap. 26, 503–526.

Appendix: Literature (cont'd)

- **Lumley T (2010)**. *Complex Surveys: A Guide to Analysis Using R.*, Hoboken (NJ): John Wiley and 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.