The Journal of Open Source Software

robomit: Robustness Checks for Omitted Variable Bias

Sergei Schaub[1, 2, 3]

1 Agricultural Economics and Policy Group, ETH ZC<rich **2** Grassland Sciences Group, ETH ZC<rich **3** Chair of Ecosystem Management, ETH ZC<rich

DOI:

Software

- Review
- Repository
- Archive

Submitted:
Published:

Summary

The recently developed methodological framework by Oster (2019) (hereafter Oster framework) helps to understand if inferences based on estimation results are likely to hold despite omitted variables bias. `robomit` (Schaub, 2021) implements the Oster framework in R and offers features for sensitivity analyses of the estimates parameters of the Oster framework and visualization of those sensitivity analyses.

Statement of need

Researchers frequently encounter omitted variables bias in their estimations in nonexperimental work, which can lead to flawed inferences. Recent methodological developments help understand whether inferences of these estimations are likely to hold despite this bias (Cinelli & Hazlett, 2020; Harada, 2013; Imbens, 2003; Oster, 2019) (see Oster (2019) and Cinelli & Hazlett (2020) for a recent overview).

The Oster framework offers the option to compute i) the bias-adjusted treatment effect or correlation, β^*, and ii) the degree of selection on unobservables relative to observables (with respect to the treatment variable) that would be necessary to eliminate the result, δ^*, using standard regression output. Thus, researchers can assess the potential severity of the omitted variables bias for their inferences. The two variables, β^* and δ^*, can be estimated by only specifying two parameters.

Here, we present the R-package `robomit` (Schaub, 2021) and its features. `robomit` implements the Oster framework, i.e., the estimation of β^* and δ^*, for linear cross-sectional and panel models. Additionally, `robomit` offers features for sensitivity analyses of β^* and δ^* (concerning the sample and external parameter specification) and their visualization.[1]

The remaining sections introduce a) briefly omitted variable bias and the central intuition of the Oster framework, and b) the functions of `robomit`.

Framework

Omitted variable bias occurs when we omit one or more (unobserved) variables in our estimation correlated with our independent variable and our treatment variable, i.e., variable of interest. This omission might be because we do not observe these omitted variables, which are so-called unobservables. Let's assume the simple case and that we want to estimate (e.g., Stock & Watson, 2007; Wooldridge, 2016):

[1] The Oster framework is available in Stata under the command `psacalc`.

$$Y = \alpha + \beta T + \varphi X + \omega Z + u. \tag{1}$$

Where Y is the outcome variable, T the treatment variable, X a vector of observed control variables (i.e., observables), Z an unobserved variable (i.e., unobservables), and u the error term. If Z can be represented as a function of T, let's say $Z = \psi + \eta T + e$ (where e is the error term), and Z is not part of our estimation, we estimate:

$$Y = (\alpha + \omega\psi) + (\beta + \omega\eta)T + \varphi X + (u + \omega e). \tag{2}$$

Thus, we estimate a biased coefficient for T when we omit Z.

Oster (2019) developed a framework based on Altonji, Elder, & Taber (2005) to assess the potential severity of selection on unobservables (i.e., omitted variable bias). The central intuition of the Oster framework is that the omitted variable bias is proportional to the coefficient movements scaled by the movement of the R^2. Using this intuition, Oster (2019) presents a framework to approximate β^*, which is the bias-adjusted treatment effect (if the estimation is causal) or the bias-adjusted treatment correlation. The approximation is:

$$\beta^* \approx \widetilde{\beta} - \delta(\dot{\beta} - \widetilde{\beta})\frac{R^2_{max} - \widetilde{R}^2}{\widetilde{R}^2 - \dot{R}^2}. \tag{3}$$

Where $\widetilde{\beta}$ is the coefficient of the controlled model (i.e., the intermediate regression of Y on T and X), δ the value of relative importance of the selection of the observed variables compared to the unobserved variables, $\dot{\beta}$ the coefficient of the uncontrolled model (i.e., the auxiliary regression of Y on T), R^2_{max} the R^2 of the hypothetical model (i.e., the hypothetical regression of Y on T, X, and Z), $\widetilde{R}^2$ the R^2 of the controlled model, and $\dot{R}^2$ the R^2 of the uncontrolled model. Hence, we only need to define the values of δ and R^2_{max} to estimate β^* while all other values are automatically derived from the regression output. As a default, it is often assumed that $\delta = 1$ and $R^2_{max} = 1.3\widetilde{\beta}$ (Altonji, Elder, & Taber, 2005; Oster, 2019). $R^2_{max} = 1.3\widetilde{R}^2$ is based on the 90%-survival rate of results of randomized studies (Oster, 2019). Researchers should also examine other values for R^2_{max} to understand the sensitivity of the results to the specified value, especially when $\widetilde{R}^2$ is low (`robomit` also implements a sensitivity analysis of R^2_{max}).

Next to estimating β^* we can estimate δ^*, which is the degree of selection on unobservables relative to observables (with respect to the treatment variable) that would be necessary to produce $\beta = \hat{\beta}$. δ^* is defined as:

$$\delta^* = \frac{\begin{gathered}(\widetilde{\beta} - \hat{\beta})(\widetilde{R}^2 - \dot{R}^2)\hat{\sigma}^2_Y\hat{\tau}_X + (\widetilde{\beta} - \hat{\beta})\hat{\sigma}^2_X\hat{\tau}_X(\dot{\beta} - \widetilde{\beta})^2 + \\ 2(\widetilde{\beta} - \hat{\beta})^2(\hat{\tau}_X(\dot{\beta} - \widetilde{\beta})\hat{\sigma}^2_X) + (\widetilde{\beta} - \hat{\beta})^3(\hat{\tau}_X\hat{\sigma}^2_X - \hat{\tau}^2_X)\end{gathered}}{\begin{gathered}(R^2_{max} - \widetilde{R}^2)\hat{\sigma}^2_Y(\dot{\beta} - \widetilde{\beta})\hat{\sigma}^2_X + (\widetilde{\beta} - \hat{\beta})(R^2_{max} - \widetilde{R}^2)\hat{\sigma}^2_Y(\hat{\sigma}^2_X) - \hat{\tau}_X) + \\ (\widetilde{\beta} - \hat{\beta})^2(\hat{\tau}_X(\dot{\beta} - \widetilde{\beta})\hat{\sigma}^2_X) + (\widetilde{\beta} - \hat{\beta})^3(\hat{\tau}_X\hat{\sigma}^2_X - \hat{\tau}^2_X)\end{gathered}}. \tag{4}$$

Where σ^2_Y is the variance of Y, σ^2_X the variance of X, and $\hat{\tau}_X$ the variance of this residual in the sample. To estimate δ^* researchers need only to specify $\hat{\beta}$ and R^2_{max}. $\hat{\beta}$ is commonly defined as $\hat{\beta} = 0$ (Oster, 2019).

Demonstration of the robomit package

The demonstration uses a cross-sectional dataset of sales prices of house in the city of Windsor (Canada) in 1987, which is taken from Anglin & Gencay (1996) using the R package `Ecdat` (Croissant & Graves, 2020). The dataset contains information about house sales prices (Canadian dollars), the lot size of the property (in square feet), and other control variables. In our demonstration, we are interested in the correlation of house sales prices (dependent variable; log-transformed) and the lot size of the property (treatment variable; log-transformed) and how robust this correlation is to the potential inclusion of unobservables (i.e., omitted variable bias). This analysis aims to illustrate the functions of robomit and not to build a causal model.

Estimation of β^* and δ^*

First, we estimate β^* and δ^* using *o_beta* and *o_delta*, respectively, from `robomit`:

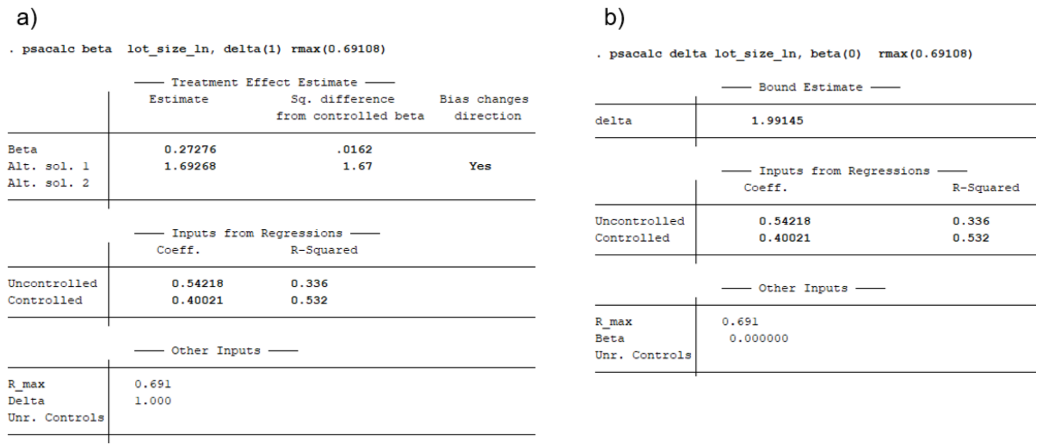
```
# estimate beta* for the lot size variable
o_beta(y = "price_ln",                        # dependent variable
       x = "lot_size_ln",                     # independent treatment variable
       con = "bedrooms + bathrooms +
              factor(driveway_dummy)",        # other control variables
       delta = 1,                             # delta (usually set to one)
       R2max = 0.5316*1.3,                    # maximum R-square (often assumed
                                              # to be the 1.3 times the R-square
                                              # of the controlled model)
       type = "lm",                           # model type
       data = Housing)                        # dataset
```

```
## # A tibble: 10 x 2
##    Name                           Value
##    <chr>                          <dbl>
##  1 beta*                         0.273
##  2 (beta*-beta controlled)^2     0.0162
##  3 Alternative Solution 1        1.69
##  4 (beta[AS1]-beta controlled)^2 1.67
##  5 Uncontrolled Coefficient      0.542
##  6 Controlled Coefficient        0.400
##  7 Uncontrolled R-square         0.336
##  8 Controlled R-square           0.532
##  9 Max R-square                  0.691
## 10 delta                         1
```

```
# estimate delta* for the lot size variable
o_delta(y = "price_ln",                        # dependent variable
        x = "lot_size_ln",                     # independent treatment variable
        con = "bedrooms + bathrooms +
               factor(driveway_dummy)",        # other control variables
        beta = 0,                              # beta (usually set to zero)
        R2max = 0.5316*1.3,                    # maximum R-square
        type = "lm",                           # model type
        data = Housing)                        # dataset
```

```
## # A tibble: 7 x 2
##   Name                     Value
##   <chr>                    <dbl>
## 1 delta*                   1.99
## 2 Uncontrolled Coefficient 0.542
## 3 Controlled Coefficient   0.400
## 4 Uncontrolled R-square    0.336
## 5 Controlled R-square      0.532
## 6 Max R-square             0.691
## 7 beta hat                 0
```

The results show that β^*, i.e., the bias-adjusted coefficient of lot size, is 0.27. Moreover, we estimated a δ^* of 1.99. Thus, the unobservables need to be 1.99 times more important than the observables (with respect to the treatment variable) to obtain a correlation of zero (as we defined: beta = 0). The results are equivalent to those of the Stata command `psacalc` (Fig. 1). All functions of robomit also offers the option to include unrelated control variables (by specifying m (Oster, 2019)) and weights (by specifying *weights*).



```
a)
. psacalc beta  lot_size_ln, delta(1) rmax(0.69108)

              ---- Treatment Effect Estimate ----
             | Estimate         Sq. difference        Bias changes
             |                  from controlled beta  direction
-------------+------------------------------------------------------
Beta         | 0.27276          .0162
Alt. sol. 1  | 1.69268          1.67                  Yes
Alt. sol. 2  |
-------------+------------------------------------------------------

              ---- Inputs from Regressions ----
             | Coeff.           R-Squared
-------------+------------------------------------------------------
Uncontrolled | 0.54218          0.336
Controlled   | 0.40021          0.532
-------------+------------------------------------------------------

              ---- Other Inputs ----
-------------+------------------------------------------------------
R_max        | 0.691
Delta        | 1.000
Unr. Controls|
-------------+------------------------------------------------------

b)
. psacalc delta lot_size_ln, beta(0)  rmax(0.69108)

              ---- Bound Estimate ----
-------------+------------------------------------------------------
delta        | 1.99145
-------------+------------------------------------------------------

              ---- Inputs from Regressions ----
             | Coeff.                         R-Squared
-------------+------------------------------------------------------
Uncontrolled | 0.54218                        0.336
Controlled   | 0.40021                        0.532
-------------+------------------------------------------------------

              ---- Other Inputs ----
-------------+------------------------------------------------------
R_max        | 0.691
Beta         | 0.000000
Unr. Controls|
-------------+------------------------------------------------------
```

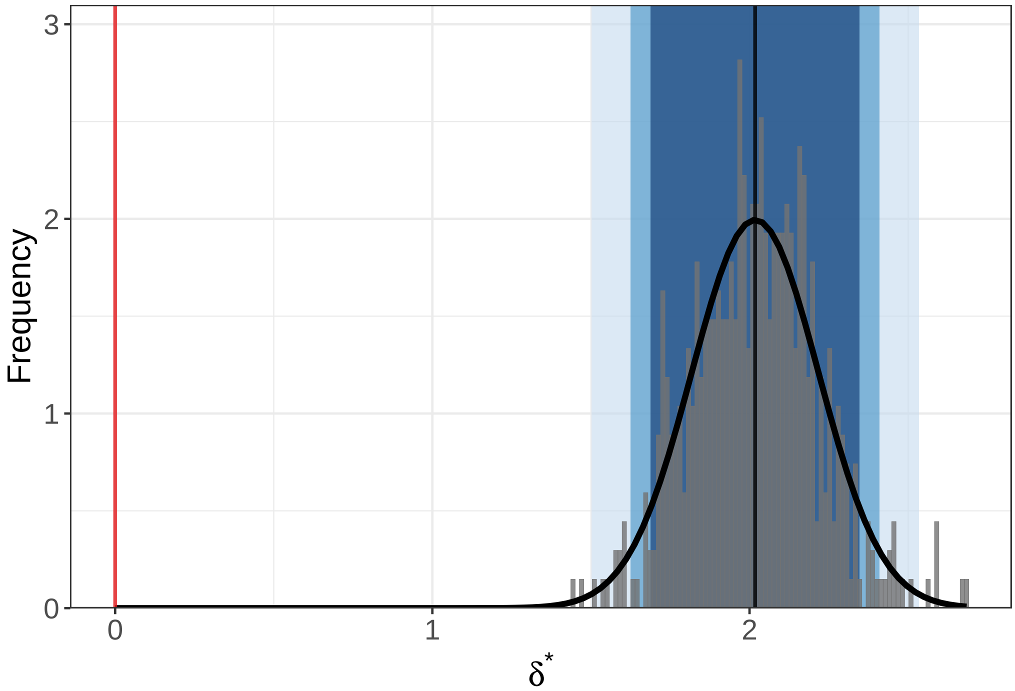
Figure 1: Stata results of β^* (panel a) and δ^* (panel b).

Features for sensitivity analyses and their visualization

`robomit` includes a set of functions for sensitivity analyses of β^* and δ^*: *o_beta_boot, o_delta_boot, o_beta_boot_inf, o_delta_boot_inf, o_beta_boot_viz, o_delta_boot_viz, o_beta_rsq, o_delta_rsq, o_beta_rsq_viz,* and *o_delta_rsq_viz*. Here, we present the visualization of bootstrapped δ^* using *o_delta_boot_viz* and δ^* over a range of R^2_{max} using *o_delta_rsq_viz* (the other functions follow the same logic). First, we use *o_delta_boot_viz*:

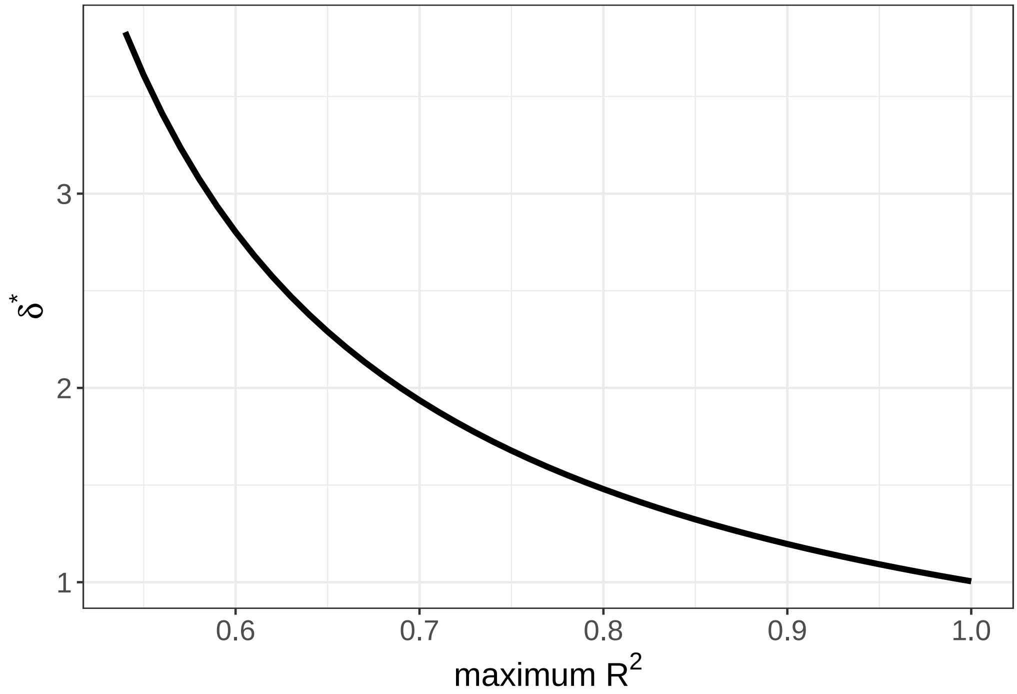
```
# visualization of bootstrapped delta*s
o_delta_boot_viz(y = "price_ln",               # dependent variable
             x = "lot_size_ln",                # independent treatment
                                               # variable
             con = "bedrooms + bathrooms +
                    factor(driveway_dummy)",   # other control variables
             beta = 0,                         # beta
```

```
             R2max = 0.5316*1.3,              # maximum R-square
             sim = 500,                       # number of simulations
             obs = 350,                       # draws per simulation
             rep = FALSE,                     # without replacement
             CI = c(90,95,99),                # confidence intervals
             type = "lm",                     # model type
             norm = TRUE,                     # normal distribution
             bin = 200,                       # number of bins
             useed = 123,                     # seed
             data = Housing)                  # dataset
```



The figure show that δ^* is not sensitive to selecting different sub-samples, thus, sample selection. Second, we use *o_delta_boot_viz*:

```
# estimate delta*s over a range of maximum R-squares
o_delta_rsq_viz(y = "price_ln",                 # dependent variable
                x = "lot_size_ln",              # independent treatment
                                                # variable
                con = "bedrooms + bathrooms +
                       factor(driveway_dummy)", # other control variables
                beta = 0,                       # beta
                type = "lm",                    # model type
                data = Housing)                 # dataset
```



δ^* decrease from 3.83 to 1.005 when the R^2_{max} increases from 0.54 to 1. δ^* remains above one when $R^2_{max} = 1$. $\delta^* \geq 1$ is suggested as a reasonable heuristic threshold for indicating robustness (Altonji, Elder, & Taber, 2005; Oster, 2019).

Acknowledgements

The R package was supported by the Mercator Foundation Switzerland within a ZC<rich-Basel Plant Science Center PhD Fellowship program.

References

Altonji, J. G., Elder, T. E., & Taber, C. R. (2005). Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of Political Economy*, *113*(1), 151–184. doi:https://doi.org/10.1086/426036

Anglin, P. M., & Gencay, R. (1996). Semiparametric estimation of a hedonic price function. *Journal of Applied Econometrics*, *11*(6), 633–648.

Cinelli, C., & Hazlett, C. (2020). Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *82*(1), 39–67. doi:https://doi.org/10.1111/rssb.12348

Croissant, Y., & Graves, S. (2020). *Ecdat: Data sets for econometrics*. Retrieved from https://CRAN.R-project.org/package=Ecdat

Harada, M. (2013). *Generalized sensitivity analysis and application to quasi-experiments*. Working Paper, New York University. Retrieved from www3.grips.ac.jp/~m-harada/docs/Harada-GSA_and_applications.pdf

Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, *93*(2), 126–132. doi:https://doi.org/10.1257/000282803321946921

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, *37*(2), 187–204. doi:https://doi.org/10.1080/07350015.2016.1227711

Schaub, S. (2021). *Robomit: Robustness checks for omitted variable bias.*

Stock, J. H., & Watson, M. W. (2007). *Introduction to econometrics.* Pearson international edition (2nd ed.). Boston: Addison-Wesley.

Wooldridge, J. M. (2016). *Introductory econometrics: A modern approach* (5th ed., [international ed.].). Melbourne: South Western Cengage Learning.