# Supplement to "A Multi-Step Kernel-Based Regression Estimator that Adapts to Error Distributions of Unknown Form"

***Communications in Statistics: Theory-Methods*, 2021, 50(24), 6211-6230**
https://doi.org/10.1080/03610926.2020.1741625

By Jan G. De Gooijer[a] and Hugo Reichardt[b]

[a] *University of Amsterdam and* [b] *London School of Economics*

The following R files are available for replication of the empirical and simulation results.

**Simulation:**

1. simulation.r: Computes mean, variance and RMSE for all estimates (and percentage of errors for YDG) for the slope and intercept. The results are summarized in Table3.csv, Table4.csv, Table5.csv, and Table6.csv (see below).
2. bandwidth-simulation.r: For the choice of the bandwidth under 8 error distributions. The results for the mean are summarized in Table1.csv (see below) and briefly discussed in Section 4.1, last paragraph, of the paper.
3. std-error-simulation.r: Computes the results in Tables 3–6 of the paper.

**Output files:**

1. Table1.csv: mean and standard deviations of rule-of-thumb bandwidth under different distributions based on 500 replications. Column headers (a)–(h) refer to the distributions described in Section 4.1 of the paper. Bandwidth selection is based on LSE residuals of the samples considered in Table2.csv.
2. Table2.csv: RMSE of kernel-based estimators for different bandwidth values for sample size $n = 100$ and the number of variables (including intercept) $p = 2$. Results are based on 500 replications. For scenarios marked with $*$, the YDG estimator failed for some replications. In that case, the results are averages over the replications where it did not fail. The results in Table2.csv are *not* summarized in the paper.
3. Table3.csv: RMSE of the slope coefficient for all investigated scenarios. Results are based on 500 replications. The results are summarized in Table 1, and discussed in Section 4.2 (2nd paragraph) of the paper.
4. Table4.csv: RMSE of the intercept coefficient for all investigated scenarios. Results are based on 500 replications. The results are summarized in Table 1, and discussed in Section 4.2 (2nd paragraph) of the paper.
5. Table5.csv: Comparison of bias of the slope coefficient. Results are based on 500 replications. For $p = 5$ and $p = 10$, the bias of the slope coefficients is defined as the mean of the absolute bias of the $p - 1$ slope coefficients. The results are summarized in Table 2, and discussed in Section 4.2 of the paper.
6. Table6.csv: Comparison of bias of the intercept coefficient. Results are based on 500 replications. The results are summarized in Table 2, and discussed in Section 4.2 of the paper.

**Data:**

The empirical illustration in Section 5 of the paper is based on the educational data set of Andrabi (2017); see the folder Data. The data is also available at http://www.aeaweb.org/articles?id=10.1257/aer.20140774.

File residuals-models-1-3.csv contains the LSE residuals of models 1–3 as given by Eq. (17) in the paper.

**Reference:**

Andrabi, T., J. Das and A. I. Khwaja (2017). Report cards: The impact of providing school and child test scores on educational markets. *American Economic Review 107*, 1535–1563.