OPSR: A Package for Estimating Ordered Probit Switching Regression Models in R

Daniel Heimgartner
ETH Zürich

Xinyi Wang
MIT Boston

Abstract

Selection bias may arise if unobserved factors simultaneously influence the selection process for who gets treated (or not), and the outcome of (not) receiving the treatment. Different methods exist to correct for this bias depending on whether longitudinal or cross-sectional data is available. A possible cure in the latter case (where the counterfactual treatment outcome is never observed) is to explicitly account for the arising error correlation and estimate the covariance matrix of the selection and outcome processes. This is known as endogenous switching regression. The R package **OPSR** introduced in this article provides an easy-to-use, fast and memory efficient interface to ordered probit switching regression, accounting for self-selection into an ordinal treatment. It handles log-transformed outcomes which need special consideration when computing conditional expectations and thus treatment effects.

Keywords: ordered probit switching regression, endogenous switching regression, Heckman selection, selection bias, treatment effect, R.

1. Introduction

The goal of the program evaluation literature is to estimate the effect of a treatment program (e.g., a new policy, technology, medical treatment, or agricultural practice) on an outcome. To evaluate such a program, the "treated" are compared to the "untreated". In an experimental setting, the treatment can be (randomly) assigned by the researcher. However, in an observational setting, the treatment is not always exogenously prescribed but rather self-selected. This gives rise to a selection bias when unobserved factors influencing the treatment adoption also influence the outcome (also known as *selection on unobservables*). Simple group comparison no longer yield an unbiased estimate of the treatment effect. In more technical terms, the counterfactual outcome of the treated ("if they had not been treated") does not necessarily correspond to the factual outcome of the untreated. For example, cyclists riding without a helmet (the "untreated") might have a risk-seeking tendency. We therefore potentially overestimate the benefit of wearing a helmet if we compare the accident (severity) rate of the two groups. Risk-seeking is not readily measured and it is easy to imagine that it becomes part of the error in applied research and thus leading cause of a selection bias.

To properly account for the selection bias, various techniques exist, both for longitudinal and cross-sectional data. In the first case, difference in differences is a widely adopted measure. In the latter case, instrumental variables, matching propensity scores, regression-discontinuity design, and the endogenous switching regression model have been applied (Wang and Mokhtarian 2024). The latter method is particularly well-suited to correct for both selection on observables and unobservables (unlike other methods which only address and correct for selection on observables).

The seminal work by Heckman (1979) proposed a two-part model to address the selection bias that often occurs when modelling a continuous outcome which is only observable for a subpopulation. A very nice exposition of this model is given in Cameron and Trivedi (2005, Chapter 16). The classical Heckman model consists of a probit equation and continuous outcome equation. A natural extension is then switching regression, where the population is partitioned into different groups (regimes) and separate parameters are estimated for the continuous outcome process of each group. This model is originally known as the Roy model (Cameron and Trivedi 2005) or Tobit 5 model (Amemiya 1985). These classical models (the Tobit models for truncated, censored or interval data and their extensions) are implemented in various environments for statistical computing and in R's (R Core Team 2017) **sampleSelection** package (Toomet and Henningsen 2008).

Many different variants can then be derived by either placing different distributional assumptions on the errors and/or how the latent process manifests into observed outcomes (i.e., the dependent variables can be of various types, such as binary, ordinal, censored, or continuous) more generally known as conditional mixed-process (CMP) models. CMP models comprise a broad family involving two or more equations featuring a joint error distribution assumed to be multivariate normal. The Stata (StataCorp 2023) command `cmp` (Roodman 2011) can fit such models. The variant at the heart of this paper is an ordered probit switching regression (OPSR) model, with ordered treatments and continuous outcome. Throughout the text we use the convention that OPSR refers to the general methodology, while **OPSR** refers specifically to the package.

OPSR is available as a Stata command, `oheckman` (Chiburis and Lokshin 2007), which however, does not allow distinct specifications for the continuous outcome processes (i.e., the same explanatory variables must be used for all treatment groups). The relatively new R package **switchSelection** (Potanin 2024) allows to estimate multivariate and multinomial sample selection and endogenous switching models with multiple outcomes. These models are systems of ordinal, continuous and multinomial equations and thus nest OPSR as a special case.

OPSR is tailored to one particular method, easy to use (understand, extend and maintain), fast and memory efficient. Unlike the mentioned implementations, it handles log-transformed continuous outcomes which need special consideration for the computation of conditional expectations. It obeys to R's implicit modeling conventions (by extending the established generics such as `summary()`, `predict()`, `update()`, `anova()` among others) and produces production-grade output tables. This work generalizes the learnings from Wang and Mokhtarian (2024) and makes the OPSR methodology readily available. The mathematical notation presented here translates to code almost verbatim which hopefully serves a pedagogical purpose for the curious reader.

The remainder of this paper is organized as follows: Section 2 outlines the ordered probit switching regression model, lists all the key formulas underlying the software implementation and details **OPSR**'s architecture. In Section 3 the key functionality is demonstrated both on simulated data and the data from Wang and Mokhtarian (2024) which we use to reproduce their core model. The case study in Section 4 leverages tracking data from the TimeUse+ study (Winkler, Meister, and Axhausen 2024) investigating telework treatment effects on weekly distance traveled. There, we also compare the OPSR models to the ones not accounting for error correlation and discuss the implications for treatment effects. The summary in Section 5 concludes.

2. Model and software

In the following, we outline the ordered probit switching regression model as well as list all the key formulas underlying the software implementation. **OPSR** follows the R-typical formula interface to a workhorse fitter function. Its architecture is detailed after the mathematical part.

As alluded, OPSR is a two-step model: One process governs the ordinal outcome and separate processes (for each ordinal outcome) govern the continuous outcomes. The ordinal outcome can also be thought of as a regime or treatment. In the subsequent exposition, we will refer to the two processes as *selection* and *outcome* process.

We borrow the notation from Wang and Mokhtarian (2024) where also all the derivations are detailed. For a similar exhibition, Chiburis and Lokshin (2007) can be consulted. Individual subscripts are suppressed throughout, for simplicity.

Let $\mathcal{Z}$ be a latent propensity governing the selection outcome

$$\mathcal{Z} = \boldsymbol{W}\boldsymbol{\gamma} + \epsilon, \tag{1}$$

where $\boldsymbol{W}$ represents the vector of attributes of an individual, $\boldsymbol{\gamma}$ is the corresponding vector of parameters and $\epsilon \sim \mathcal{N}(0,1)$ a normally distributed error term.

As $\mathcal{Z}$ increases and passes some unknown but estimable thresholds, we move up from one ordinal treatment to the next higher level

$$Z = j \quad \text{if } \kappa_{j-1} < \mathcal{Z} \leq \kappa_j, \tag{2}$$

where Z is the observed ordinal selection variable, $j = 1, \dots, J$ indexes the ordinal levels of Z, and κ_j are the thresholds (with $\kappa_0 = -\infty$ and $\kappa_J = \infty$). Hence, there are $J-1$ thresholds to be estimated. The probability that an individual self-selects into treatment group j is

$$\begin{aligned}
\mathsf{P}[Z = j] &= \mathsf{P}[\kappa_{j-1} < \mathcal{Z} \leq \kappa_j] \\
&= \mathsf{P}[\kappa_{j-1} - \boldsymbol{W}\boldsymbol{\gamma} < \epsilon \leq \kappa_j - \boldsymbol{W}\boldsymbol{\gamma}] \\
&= \Phi(\kappa_j - \boldsymbol{W}\boldsymbol{\gamma}) - \Phi(\kappa_{j-1} - \boldsymbol{W}\boldsymbol{\gamma}).
\end{aligned} \tag{3}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

The outcome model for the j^{th} treatment group is expressed as

$$y_j = \boldsymbol{X_j}\boldsymbol{\beta_j} + \eta_j, \tag{4}$$

where y_j is the observed continuous outcome, $\boldsymbol{X_j}$ the vector of observed explanatory variables associated with the j^{th} outcome model, $\boldsymbol{\beta_j}$ is the vector of associated parameters, and $\eta_j \sim \mathcal{N}(0, \sigma_j^2)$ is a normally distributed error term. At this point it should be noted that $\boldsymbol{X_j}$ and $\boldsymbol{W}$ may share some explanatory variables but not all, due to identification problems otherwise (Chiburis and Lokshin 2007).

The key assumption of OPSR is now that the errors of the selection and outcome models are jointly multivariate normally distributed

$$\begin{pmatrix} \epsilon \\ \eta_1 \\ \vdots \\ \eta_j \\ \vdots \\ \eta_J \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_1\sigma_1 & \cdots & \rho_j\sigma_j & \cdots & \rho_J\sigma_J \\ \rho_1\sigma_1 & \sigma_2^2 & & & & \\ \vdots & & \ddots & & & \\ \rho_j\sigma_j & & & \sigma_j^2 & & \\ \vdots & & & & \ddots & \\ \rho_J\sigma_J & & & & & \sigma_J^2 \end{pmatrix}\right), \tag{5}$$

where ρ_j represents the correlation between the errors of the selection model (ϵ) and the j^{th} outcome model (η_j). If the covariance matrix should be diagonal (i.e., no error correlation), no selection-bias exists and the selection and outcome models can be estimated separately.

As shown in Wang and Mokhtarian (2024), the log-likelihood of observing all individuals self-selecting into treatment j and choosing continuous outcome y_j can be expressed as

$$\ell(\theta \mid \boldsymbol{W}, \boldsymbol{X_j}) = \sum_{j=1}^{J} \sum_{\{j\}} \Bigg\{ \ln\left[\frac{1}{\sigma_j}\phi\left(\frac{y_j - \boldsymbol{X_j\beta_j}}{\sigma_j}\right)\right] \quad + \\ \ln\left[\Phi\left(\frac{\sigma_j(\kappa_j - \boldsymbol{W\gamma}) - \rho_j(y_j - \boldsymbol{X_j\beta_j})}{\sigma_j\sqrt{1-\rho_j^2}}\right) - \Phi\left(\frac{\sigma_j(\kappa_{j-1} - \boldsymbol{W\gamma}) - \rho_j(y_j - \boldsymbol{X_j\beta_j})}{\sigma_j\sqrt{1-\rho_j^2}}\right)\right]\Bigg\} \tag{6}$$

where $\sum_{\{j\}}$ means the summation of all the cases belonging to the j^{th} selection outcome, $\phi(\cdot)$ and $\Phi(\cdot)$ are the density and cumulative distribution function of the standard normal distribution.

The conditional expectation can be expressed as

$$\begin{aligned} \mathsf{E}[y_j \mid Z = j] &= \boldsymbol{X_j\beta_j} + \mathsf{E}[\eta_j \mid \kappa_{j-1} - \boldsymbol{W\gamma} < \epsilon \le \kappa_j - \boldsymbol{W\gamma}] \\ &= \boldsymbol{X_j\beta_j} - \rho_j\sigma_j\frac{\phi(\kappa_j - \boldsymbol{W\gamma}) - \phi(\kappa_{j-1} - \boldsymbol{W\gamma})}{\Phi(\kappa_j - \boldsymbol{W\gamma}) - \Phi(\kappa_{j-1} - \boldsymbol{W\gamma})}, \end{aligned} \tag{7}$$

where the fraction is the ordered probit switching regression model counterpart to the inverse Mills ratio (IMR) term of a binary switching regression model. We immediately see, that regressing $\boldsymbol{X_j}$ on y_j leads to an omitted variable bias if $\rho_j \neq 0$ which is the root cause of the selection bias. However, the IMR can be pre-computed based on an ordered probit model and then included in the second stage regression, which describes the Heckman correction (Heckman 1979). It should be warned, that since the Heckman two-step procedure includes an estimate in the second step regression, the resulting OLS standard errors and heteroskedasticity-robust standard errors are incorrect (Greene 2002).

To obtain unbiased treatment effects, we must further evaluate the "counterfactual outcome", which reflects the expected outcome under a counterfactual treatment (i.e., for $j' \neq j$)

$$\begin{aligned} \mathsf{E}[y_{j'} \mid Z = j] &= \boldsymbol{X_{j'}\beta_{j'}} + \mathsf{E}[\eta_{j'} \mid \kappa_{j-1} - \boldsymbol{W\gamma} < \epsilon \le \kappa_j - \boldsymbol{W\gamma}] \\ &= \boldsymbol{X_{j'}\beta_{j'}} - \rho_{j'}\sigma_{j'}\frac{\phi(\kappa_j - \boldsymbol{W\gamma}) - \phi(\kappa_{j-1} - \boldsymbol{W\gamma})}{\Phi(\kappa_j - \boldsymbol{W\gamma}) - \Phi(\kappa_{j-1} - \boldsymbol{W\gamma})}. \end{aligned} \tag{8}$$

Let's assume that $y_j = \ln(Y_j + 1)$ in the previous equations. I.e., the continuous outcome was log-transformed as is usual in regression analysis. We have to note, that in such cases the Equations 7-8 provide the conditional expectation of the log-transformed outcome. Therefore we need to back-transform $Y_j = \exp(y_j) - 1$ which yields

$$\mathsf{E}[Y_j \mid Z = j] = \exp\left(\boldsymbol{X_j}\boldsymbol{\beta_j} + \frac{\sigma_j^2}{2}\right)\left[\frac{\Phi(\kappa_j - \boldsymbol{W\gamma} - \rho_j\sigma_j) - \Phi(\kappa_{j-1} - \boldsymbol{W\gamma} - \rho_j\sigma_j)}{\Phi(\kappa_j - \boldsymbol{W\gamma}) - \Phi(\kappa_{j-1} - \boldsymbol{W\gamma})}\right] - 1 \quad (9)$$

for the factual case, and

$$\mathsf{E}[Y_{j'} \mid Z = j] = \exp\left(\boldsymbol{X_{j'}}\boldsymbol{\beta_{j'}} + \frac{\sigma_{j'}^2}{2}\right)\left[\frac{\Phi(\kappa_j - \boldsymbol{W\gamma} - \rho_{j'}\sigma_{j'}) - \Phi(\kappa_{j-1} - \boldsymbol{W\gamma} - \rho_{j'}\sigma_{j'})}{\Phi(\kappa_j - \boldsymbol{W\gamma}) - \Phi(\kappa_{j-1} - \boldsymbol{W\gamma})}\right] - 1 \quad (10)$$

for the counterfactual case (Wang and Mokhtarian 2024).

This concludes the mathematical treatment and we briefly outline **OPSR**'s architecture which can be conceptualized as follows:

- We provide the usual formula interface to specify a model. To allow for multiple parts and multiple responses, we rely on the **Formula** package (Zeileis and Croissant 2010).
- After parsing the formula object, checking the user inputs and computing the model matrices, the Heckman two-step estimator is called in `opsr_2step()` to generate reasonable starting values.
- These are then passed together with the data to the basic computation engine `opsr.fit()`. The main estimates are retrieved using maximum likelihood estimation by passing the log-likelihood function `loglik_cpp()` (Equation 6) to `maxLik()` from the **maxLik** package (Henningsen and Toomet 2011).
- All the above calls are nested in the main interface `opsr()` which returns an object of class '`opsr`'. Several methods then exist to post-process this object as illustrated below.

The likelihood function `loglik_cpp()` is implemented in C++ using **Rcpp** (Eddelbuettel and Balamuta 2018) and relying on the data types provided by **RcppArmadillo** (Eddelbuettel and Sanderson 2014). Parallelization is available using OpenMP. This makes **OPSR** both fast and memory efficient (as data matrices are passed by reference).

3. Illustrations

We first illustrate how to specify a model using **Formula**'s extended syntax and simulated data. Then the main functionality of the package is demonstrated. We conclude this section by demonstrating some nuances, reproducing the core model of Wang and Mokhtarian (2024).

Let us simulate date from an OPSR process with three ordinal outcomes and distinct design matrices $\boldsymbol{W}$ and $\boldsymbol{X}$ (where $\boldsymbol{X} = \boldsymbol{X}_j \ \forall j$) by

```
R> sim_dat <- opsr_simulate()
R> dat <- sim_dat$data
R> head(dat)
```

```
  ys      yo     xs1    xs2    xo1    xo2
1  3 -2.1007 -0.7227  1.026 -0.530  0.618
2  3  0.6958  0.7511  1.806 -0.452 -0.309
3  3 -0.0204  0.5161 -0.111  0.575  0.555
4  1  2.4690  0.0773 -1.421  0.595  0.377
5  1 -0.4258 -1.4454  0.425 -1.085 -0.733
6  2  1.5742 -0.8670  0.278 -0.724 -0.750
```

where `ys` is the selection dependent variable (or treatment group), `yo` the outcome dependent variable and `xs` respectively `xo` the corresponding explanatory variables.

Models are specified symbolically. A typical model has the form `ys | yo ~ terms_s | terms_o1 | terms_o2 | ...` where the `|` separates the two responses and process specifications. If the user wants to specify the same process for all continuous outcomes, two processes are enough (`ys | yo ~ terms_s | terms_o`). Hence, the minimal `opsr()` interface call reads

```
R> fit <- opsr(ys | yo ~ xs1 + xs2 | xo1 + xo2, data = dat,
+    printLevel = 0)
```

where `printLevel = 0` omits working information during maximum likelihood iterations.

As usual, the fitter function does the bare minimum model estimation while inference is performed in a separate call to

```
R> summary(fit)

Call:
opsr(formula = ys | yo ~ xs1 + xs2 | xo1 + xo2, data = dat, printLevel = 0)

BFGS maximization, 117 iterations
Return code 0: successful convergence
Runtime: 0.503 secs
Number of regimes: 3
Number of observations: 1000 (153, 510, 337)
Estimated parameters: 19

Log-Likelihood: -2054
AIC: 4147
BIC: 4240
Pseudo R-squared (EL): 0.5
Pseudo R-squared (MS): 0.449
Multiple R-squared: 0.802 (0.802, 0.752, 0.844)

Estimates:
              Estimate Std. error t value Pr(> t)
kappa1         -1.9487     0.0915  -21.30 < 2e-16 ***
kappa2          0.8899     0.0637   13.98 < 2e-16 ***
s_xs1           0.9374     0.0588   15.94 < 2e-16 ***
```

```
s_xs2            1.4375     0.0722   19.90 < 2e-16 ***
o1_(Intercept)   1.0148     0.1900    5.34 9.3e-08 ***
o1_xo1           2.1570     0.1031   20.93 < 2e-16 ***
o1_xo2           0.9863     0.0923   10.68 < 2e-16 ***
o2_(Intercept)   0.8937     0.0475   18.80 < 2e-16 ***
o2_xo1          -1.0523     0.0481  -21.87 < 2e-16 ***
o2_xo2           1.5105     0.0501   30.17 < 2e-16 ***
o3_(Intercept)   1.0812     0.0838   12.90 < 2e-16 ***
o3_xo1           1.5836     0.0625   25.35 < 2e-16 ***
o3_xo2          -2.0598     0.0605  -34.04 < 2e-16 ***
sigma1           1.1495     0.0590   19.48 < 2e-16 ***
sigma2           1.0679     0.0345   30.92 < 2e-16 ***
sigma3           1.1350     0.0399   28.41 < 2e-16 ***
rho1             0.2328     0.1683    1.38  0.1666
rho2             0.2131     0.0708    3.01  0.0026 **
rho3             0.1734     0.0929    1.87  0.0621 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Wald chi2 (null): 4155 on 8 DF, p-value: < 0
Wald chi2 (rho): 14.5 on 3 DF, p-value: < 0.002
```

The presentation of the model results is fairly standard and should not warrant further explanation with the following exceptions

1. The number of regimes along absolute counts are reported.

2. Pseudo R-squared (EL) is based on the log-likelihood of the "equally likely" model, while Pseudo R-squared (MS) is based on the log-likelihood of the "market share" model. These indicators reflect the goodness of fit for the selection process. The multiple R-squared is reported for all continuous outcomes collectively and for the regimes separately in brackets. These indicators reflect the goodness of fit for the outcome processes.

3. Coefficient names are based on the variable names as passed to the formula specification, except that `"s_"` is prepended to the selection coefficients, `"o[0-9]_"` to the outcome coefficients and the structural components `"kappa"`, `"sigma"`, `"rho"` (aligning with the letters used in Equation 6) are hard-coded (but can be over-written).

4. The coefficients table reports robust standard errors based on the sandwich covariance matrix as computed with help of the **sandwich** package (Zeileis 2006). `rob = FALSE` reports conventional standard errors.

5. Two Wald-tests are conducted. One, testing the null that all coefficients of explanatory variables are zero and two, testing the null that all error correlation coefficients (`rho`) are zero. The latter being rejected indicates that selection bias is an issue.

A useful benchmark is always the null model with structural parameters only. The null model can be derived from an '`opsr`' model fit as follows

```
R> fit_null <- opsr_null_model(fit, printLevel = 0)
```

A model can be updated as usual

```
R> fit_intercept <- update(fit, . ~ . | 1)
```

where we have removed all the explanatory variables from the outcome processes.
Several models can be compared with a likelihood-ratio test using

```
R> anova(fit_null, fit_intercept, fit)

Likelihood Ratio Test

Model 1: ~Nullmodel
Model 2: ys | yo ~ xs1 + xs2 | 1
Model 3: ys | yo ~ xs1 + xs2 | xo1 + xo2
  logLik    Df  Test Restrictions Pr(>Chi)
1  -3292     8
2  -2842    13   900            5   <2e-16 ***
3  -2054    19  1575            6   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

If only a single object is passed, then the model is compared to the null model. If more than one object is specified a likelihood ratio test is conducted for each pair of neighboring models. As expected, both tests reject the null.

Models can be compared side-by-side using the **texreg** package (Leifeld 2013), which also allows the user to build production-grade tables as illustrated later.

```
R> texreg::screenreg(list(fit_null, fit_intercept, fit),
+    include.pseudoR2 = TRUE, include.R2 = TRUE, single.row = TRUE)

==============================================================================
                  Model 1              Model 2              Model 3
------------------------------------------------------------------------------
kappa1              -1.02 (0.05) ***     -1.95 (0.09) ***     -1.95 (0.09) ***
kappa2               0.42 (0.04) ***      0.89 (0.06) ***      0.89 (0.06) ***
sigma1               2.52 (0.14) ***      2.52 (0.14) ***      1.15 (0.06) ***
sigma2               2.12 (0.07) ***      2.12 (0.07) ***      1.07 (0.03) ***
sigma3               2.84 (0.12) ***      2.84 (0.12) ***      1.13 (0.04) ***
rho1                                     -0.07 (0.16)          0.23 (0.17)
rho2                                      0.12 (0.07)          0.21 (0.07) **
rho3                                      0.12 (0.09)          0.17 (0.09)
s_xs1                                     0.94 (0.06) ***      0.94 (0.06) ***
s_xs2                                     1.43 (0.07) ***      1.44 (0.07) ***
o1_(Intercept)       0.59 (0.20) **       0.45 (0.38)          1.01 (0.19) ***
o1_xo1                                                         2.16 (0.10) ***
```

```
o1_xo2                                                        0.99 (0.09) ***
o2_(Intercept)       0.89 (0.09) ***      0.92 (0.10) ***      0.89 (0.05) ***
o2_xo1                                                       -1.05 (0.05) ***
o2_xo2                                                        1.51 (0.05) ***
o3_(Intercept)       1.17 (0.15) ***      0.98 (0.20) ***      1.08 (0.08) ***
o3_xo1                                                        1.58 (0.06) ***
o3_xo2                                                       -2.06 (0.06) ***
------------------------------------------------------------------------------
AIC               6599.54              5709.50              4146.59
BIC               6638.80              5773.30              4239.84
Log Likelihood   -3291.77             -2841.75             -2054.30
Pseudo R^2 (EL)      0.09                 0.50                 0.50
Pseudo R^2 (MS)     -0.00                 0.45                 0.45
R^2 (total)          0.01                 0.01                 0.80
R^2 (1)              0.02                 0.02                 0.80
R^2 (2)              0.00                 0.01                 0.75
R^2 (3)              0.01                 0.01                 0.84
Num. obs.         1000                 1000                 1000
==============================================================================
*** p < 0.001; ** p < 0.01; * p < 0.05
```

Finally, the key interest of an OPSR study almost certainly is the estimation of treatment effects which relies on (counterfactual) conditional expectations as already noted in the mathematical exposition.

```
R> p1 <- predict(fit, group = 1, type = "response")
R> p2 <- predict(fit, group = 1, counterfact = 2, type = "response")
```

where `p1` is the result of applying Equation 7 and `p2` is the counterfactual outcome resulting from Equation 8. The following `type` arguments are available

- `type = "response"`: Predicts the continuous outcome according to the Equations referenced above.
- `type = "unlog-response"`: Predicts the back-transformed response if the continuous outcome was log-transformed according to Equations 9-10.
- `type = "prob"`: Returns the probability vector of belonging to `group`.
- `type = "mills"`: Returns the inverse Mills ratio.

Elements are `NA_real_` if the `group` does not correspond to the observed regime (selection outcome). This ensures consistent output length.

Now that the user understands the basic workflow, we illustrate some nuances by reproducing a key output of Wang and Mokhtarian (2024) where they investigate the treatment effect of telework (TW) on weekly vehicle miles driven. The data is attached, documented (`?telework_data`) and can be loaded by

```
R> data("telework_data", package = "OPSR")
```

The final model specification reads

```
R> f <-
+    twing_status | vmd_ln ~
+    edu_2 + edu_3 + hhincome_2 + hhincome_3 + flex_work + work_fulltime +
+    twing_feasibility + att_proactivemode + att_procarowning + att_wif +
+    att_proteamwork + att_tw_effective_teamwork + att_tw_enthusiasm +
+    att_tw_location_flex |
+    female + age_mean + age_mean_sq + race_black + race_other + vehicle +
+    suburban + smalltown + rural + work_fulltime + att_prolargehouse +
+    att_procarowning + region_waa |
+    edu_2 + edu_3 + suburban + smalltown + rural + work_fulltime +
+    att_prolargehouse + att_proactivemode + att_procarowning |
+    female + hhincome_2 + hhincome_3 + child + suburban + smalltown +
+    rural + att_procarowning + region_waa
```

and the model can be estimated by

```
R> start_default <- opsr(f, telework_data, .get2step = TRUE)
R> fit <- opsr(f, telework_data, start = start, method = "NM", iterlim = 50e3,
+    printLevel = 0)
```

where we demonstrate that

1. Default starting values as computed by the Heckman two-step procedure can be retrieved.
2. `start` values can be overridden (we have hidden the `start` vector here for brevity). If the user wishes to pass start values manually, some minimal conventions have to be followed as documented in `?opsr_check_start`.
3. Alternative maximization methods (here "Nelder-Mead") can be used (as in the original paper).

With help of the **texreg** package, production-grade tables (in various output formats) can be generated with ease.

```
R> texreg::texreg(
+    fit, beside = TRUE, include.structural = FALSE, include.R2 = TRUE,
+    include.pseudoR2 = TRUE, custom.model.names = custom.model.names,
+    custom.coef.names = custom.coef.names, reorder.coef = reorder.coef,
+    groups = groups, scalebox = 0.83, booktabs = TRUE, dcolumn = TRUE,
+    use.packages = FALSE, float.pos = "t!", single.row = TRUE,
+    caption = "Replica of \\cite{Wang+Mokhtarian:2024}, Table 3.",
+    label = "tab:wang-replica"
+  )
```

Dot arguments (`...`) passed to `texreg()` (or similar functions) are forwarded to a S4 method `extract()` which extracts the variables of interest from a model fit (see also `?extract.opsr`). We demonstrate here that

	Selection	NTWer (535)	NUTWer (322)	UTWer (727)
Education (ref: high school or less)				
Some college	0.32 (0.14)*		0.15 (0.33)	
Bachelor's degree or higher	0.47 (0.13)***		0.62 (0.32)*	
Household income (ref: less than \$50,000)				
\$50,000 to \$99,999	0.06 (0.12)			0.47 (0.23)*
\$100,000 or more	0.25 (0.11)*			0.31 (0.23)
Flexible work schedule	0.31 (0.10)**			
Full time worker	0.33 (0.10)**	0.45 (0.13)***	0.69 (0.17)***	
Teleworking feasibility	0.13 (0.01)***			
Attitudes				
Pro-active-mode	0.08 (0.04)*		−0.18 (0.08)*	
Pro-car-owning	−0.08 (0.04)*	0.14 (0.07)*	0.16 (0.09)	0.25 (0.06)***
Work interferes with family	0.11 (0.04)**			
Pro-teamwork	0.09 (0.04)*			
TW effective teamwork	0.32 (0.04)***			
TW enthusiasm	0.09 (0.04)*			
TW location flexibility	0.08 (0.04)*			
Pro-large-house		0.18 (0.05)***	0.18 (0.08)*	
Intercept		3.64 (0.27)***	2.49 (0.37)***	2.38 (0.26)***
Female		−0.21 (0.10)*		−0.36 (0.11)***
Age		0.01 (0.00)*		
Age squared		−0.00 (0.00)		
Race (ref: white)				
Black		−0.40 (0.24)		
Other races		−0.06 (0.18)		
Number of vehicles		0.12 (0.05)*		
Residential location (ref: urban)				
Suburban		0.07 (0.15)	0.45 (0.17)**	0.28 (0.14)*
Small town		0.47 (0.18)**	0.19 (0.29)	0.29 (0.28)
Rural		0.60 (0.23)**	0.81 (0.31)**	0.88 (0.34)**
Region indicator (WAA)		−0.25 (0.11)*		−0.27 (0.11)*
Number of children				0.18 (0.06)**
AIC	7191.35	7191.35	7191.35	7191.35
BIC	7491.94	7491.94	7491.94	7491.94
Log Likelihood	−3539.67	−3539.67	−3539.67	−3539.67
Pseudo R^2 (EL)	0.49	0.49	0.49	0.49
Pseudo R^2 (MS)	0.46	0.46	0.46	0.46
R^2 (total)	0.24	0.24	0.24	0.24
R^2 (1)	0.28	0.28	0.28	0.28
R^2 (2)	0.23	0.23	0.23	0.23
R^2 (3)	0.21	0.21	0.21	0.21
Num. obs.	1584	1584	1584	1584

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 1: Replica of Wang and Mokhtarian (2024), Table 3.

1. Model components can be omitted. Here, the structural coefficients (`kappa`, `sigma`, `rho`) are disgarded (`include.structural = FALSE`).

2. The model components can be printed side-by-side (`beside = TRUE`).

3. Additional goodness-of-fit indicators can be included (`include.R2 = TRUE` and `include.pseudoR2`

= TRUE). Note that the indicators are repeated for all the model components.

4. The output formatting can be controlled flexibly, by reordering, renaming and grouping coefficients (the fiddly but trivial details are hidden here for brevity).

4. Case study

Now, that the reader is familiar with the main functionality of **OPSR**, this section demonstrates how to employ it in a real-world example. The emphasis, therefore, lies not on what each function does but on guiding the reader through the modeling and post-estimation steps. We investigate once again, telework treatment effects on weekly distance traveled (aggregated over all modes of transport).

We first discuss the model building strategy to arrive at an appropriately specified OPSR model. We then demonstrate, why error correlation occurs, having omitted a variable simultaneously influencing the selection and outcome process. The OPSR models are compared to models not accounting for this error correlation and implications for treatment effects are shown. The case study concludes with a discussion on unit treatment effects investigating to what degree foregone commutes (when teleworking) are compensated with leisure travel.

We use the TimeUse+ dataset (Winkler *et al.* 2024), a smartphone-based diary, recording travel, time use, and expenditure data. Our analytical sample comprises employed individuals and is based on what Winkler and Axhausen (2024) identified as valid days. A valid day has at least 20 hours of information where 70% of the events were validated by the user. Users who did not have at least 14 valid days were excluded. For the remaining 824 participants mobility indicators for a typical week were constructed. The telework status is based on tracked (and labelled) work activities and three regimes are differentiated: Non-teleworkers (NTW), Non-usual teleworkers (NUTW; <3 days/week) and Usual teleworkers (UTW; 3+ days/week).

The data, underlying this analysis, is attached, documented (`?timeuse_data`) and can be loaded by

```
R> data("timeuse_data", package = "OPSR")
```

A basic boxplot of the response variable against the three telework statuses is displayed in Figure 1. By simply looking at the data descriptively, we might prematurely conclude that telework does not impact weekly distance traveled. However, the whole value proposition of OPSR (and of models in general) is that we really are interested in a counterfactual. If the teleworkers self-select, the counterfactual is not simply the group average. More prosaically, if the usual telewokers (UTW) would choose to be non-teleworkers (NTW), they might travel more or less than the actual NTWers.

Meanwhile, commute distance increases across the three teleworker groups, suggesting that, one, longer commutes increase the propensity to telework and two, teleworkers have a higher share of leisure travel (given the similar overall distance traveled).

Before blindly trying to specify a full model using **OPSR** the analyst is advised to first, think of an identification restriction as mentioned in Section 2 and second, estimate the models separately, e.g., using `polr()` from the **MASS** package (Venables and Ripley 2002), and `lm()`

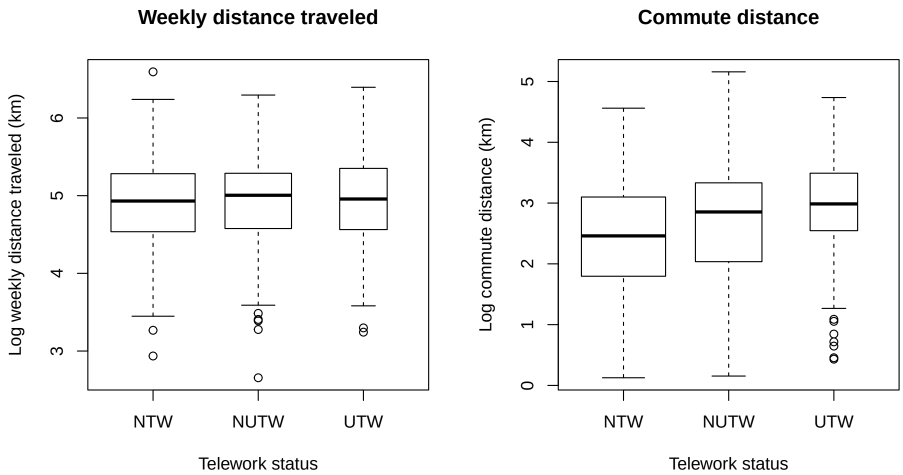

Figure 1: Log weekly distance traveled and log commute distance for different telework statuses.

for the treatment groups separately. We reserve the international standard classification of occupations (ISCO-08) variables for the selection process.

```
R> drop <- c("id", "weekly_km", "log_weekly_km", "commute_km", "log_commute_km",
+    "wfh_days")
R> dat_polr <- subset(timeuse_data, select = !(names(timeuse_data) %in% drop))
R> dat_polr$wfh <- factor(dat_polr$wfh)
R> fit_polr <- MASS::polr(wfh ~ ., dat_polr, method = "probit")
```

The `stepAIC()` function chooses a selection model specification by AIC in a stepwise algorithm.

```
R> fit_step <- MASS::stepAIC(fit_polr, trace = FALSE)
R> fit_step$anova

Stepwise Model Path
Analysis of Deviance Table

Initial Model:
wfh ~ start_tracking + age + car_access + dogs + driverlicense +
    educ_higher + fixed_workplace + grocery_shopper + hh_income +
    hh_size + isco_clerical + isco_craft + isco_elementary +
    isco_managers + isco_plant + isco_professionals + isco_service +
    isco_agri + isco_tech + married + n_children + freq_onl_order +
    parking_home + parking_work + permanent_employed + rents_home +
    res_loc + sex_male + shift_work + swiss + vacation + workload +
    young_kids

Final Model:
```

```
wfh ~ age + car_access + educ_higher + fixed_workplace + grocery_shopper +
    hh_income + isco_clerical + isco_craft + isco_elementary +
    isco_tech + freq_onl_order + parking_home + permanent_employed +
    shift_work + workload + young_kids


                   Step Df Deviance Resid. Df Resid. Dev  AIC
1                                         778       1429 1521
2      - start_tracking  6   2.0260       784       1431 1511
3             - res_loc  3   2.8697       787       1434 1508
4       - isco_managers  1   0.0133       788       1434 1506
5           - isco_agri  1   0.0415       789       1434 1504
6            - vacation  1   0.1212       790       1434 1502
7       - driverlicense  1   0.2959       791       1434 1500
8            - sex_male  1   0.5118       792       1435 1499
9          - n_children  1   0.4769       793       1435 1497
10            - hh_size  1   0.4137       794       1436 1496
11            - married  1   0.3909       795       1436 1494
12       - isco_service  1   0.4773       796       1437 1493
13         - isco_plant  1   0.6176       797       1437 1491
14         - rents_home  1   1.2327       798       1438 1490
15       - parking_work  1   1.1424       799       1440 1490
16              - swiss  1   1.5091       800       1441 1489
17               - dogs  1   1.8158       801       1443 1489
18 - isco_professionals  1   1.8728       802       1445 1489
```

Fitting the linear models separately, benefits an understanding of potential identification problems (e.g., colinear variables or missing factor levels in one of the groups). While the resulting estimates are potentially biased and their standard errors not reliable, it can still help to have a closer look at resulting estimates and goodness of fit indicators.

```
R> fit_lm <- function(data, group) {
+    f <- paste0("log_weekly_km ~ . - wfh")
+    dat <- subset(data, subset = wfh == group)
+    fit <- lm(f, dat)
+    fit
+ }
R> drop <- c("id", "weekly_km", "commute_km", "log_commute_km", "wfh_days")
R> dat_lm <- subset(timeuse_data,
+    select = !(names(timeuse_data) %in% drop) & !grepl("^isco_", names(timeuse_data)))
R> fit_ntw <- fit_lm(dat_lm, group = 1)
R> fit_nutw <- fit_lm(dat_lm, group = 2)
R> fit_utw <- fit_lm(dat_lm, group = 3)
R> summary(fit_utw)

Call:
lm(formula = f, data = dat)
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-1.1358 -0.3212  0.0632  0.3096  0.9904 

Coefficients: (2 not defined because of singularities)
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)           4.45333    0.73297    6.08  2.1e-08 ***
start_tracking8      -0.03287    0.23776   -0.14   0.8903    
start_tracking9       0.07150    0.23190    0.31   0.7585    
start_tracking10      0.09169    0.21147    0.43   0.6655    
start_tracking11     -0.18267    0.21163   -0.86   0.3901    
start_tracking12      0.02403    0.22293    0.11   0.9144    
age                  -0.00358    0.00664   -0.54   0.5907    
car_access           -0.12417    0.14169   -0.88   0.3829    
dogs                 -0.01616    0.15978   -0.10   0.9196    
driverlicense        -0.09628    0.29254   -0.33   0.7428    
educ_higher          -0.01777    0.11027   -0.16   0.8723    
fixed_workplace      -0.36434    0.15000   -2.43   0.0169 *  
grocery_shopper       0.02009    0.12016    0.17   0.8675    
hh_income4001_8000    0.25702    0.25458    1.01   0.3151    
hh_income8001_12000   0.17664    0.25069    0.70   0.4827    
hh_income12001_16000  0.27060    0.26953    1.00   0.3178    
hh_income16001+       0.26340    0.29683    0.89   0.3770    
hh_incomeNA           0.54669    0.37849    1.44   0.1517    
hh_size              -0.05528    0.07244   -0.76   0.4472    
married              -0.02904    0.12535   -0.23   0.8173    
n_children            0.01198    0.09314    0.13   0.8979    
freq_onl_order       -0.00987    0.10955   -0.09   0.9284    
parking_home               NA         NA      NA       NA    
parking_work          0.28751    0.12155    2.37   0.0199 *  
permanent_employed    0.46645    0.33901    1.38   0.1719    
rents_home            0.00874    0.11645    0.08   0.9403    
res_locrural          0.08757    0.27202    0.32   0.7482    
res_locsuburban      -0.03755    0.25928   -0.14   0.8851    
res_locurban         -0.24778    0.27434   -0.90   0.3685    
sex_male              0.24664    0.12610    1.96   0.0532 .  
shift_work                 NA         NA      NA       NA    
swiss                 0.39131    0.14266    2.74   0.0072 ** 
vacation             -0.02974    0.12819   -0.23   0.8170    
workload              0.01438    0.04129    0.35   0.7283    
young_kids           -0.05015    0.15969   -0.31   0.7541    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.531 on 102 degrees of freedom
Multiple R-squared:  0.381,	Adjusted R-squared:  0.187 
```

```
F-statistic: 1.96 on 32 and 102 DF,  p-value: 0.00593
```

Here, we have two singularity issues for the UTWers: First, `shift_work` is a constant and second, `parking_home` is colinear with `car_access`.

We then follow the conventional (somewhat heuristic) model building strategy to specify the full identified model and then exclude all variables that do not produce significant estimates (at the 10% level). The formula specification of the full model is hidden here for brevity.

```
R> fit_full <- opsr(f_full, timeuse_data, printLevel = 0)
R> f_red <- wfh | log_weekly_km ~
+    age + educ_higher + hh_income + young_kids + workload + fixed_workplace +
+    shift_work + permanent_employed + isco_craft + isco_tech + isco_clerical +
+    isco_elementary + car_access + parking_home + freq_onl_order +
+    grocery_shopper |
+    sex_male + res_loc + workload + permanent_employed + parking_work |
+    swiss + res_loc + young_kids + workload + parking_work |
+    sex_male + swiss + fixed_workplace + permanent_employed + parking_work
R> fit_red <- opsr(f_red, timeuse_data, printLevel = 0)
R> print(anova(fit_red, fit_full), print.formula = FALSE)

Likelihood Ratio Test

   logLik      Df    Test Restrictions Pr(>Chi)
1 -1337.0    50.0
2 -1316.8    99.0    40.4           49      0.8

R> summary(fit_red)

Call:
opsr(formula = f_red, data = timeuse_data, printLevel = 0)

BFGS maximization, 234 iterations
Return code 0: successful convergence
Runtime: 3.25 secs
Number of regimes: 3
Number of observations: 824 (424, 265, 135)
Estimated parameters: 50

Log-Likelihood: -1337
AIC: 2774
BIC: 3010
Pseudo R-squared (EL): 0.202
Pseudo R-squared (MS): 0.126
Multiple R-squared: 0.214 (0.203, 0.19, 0.291)

Estimates:
```

	Estimate	Std. error	t value	Pr(> t)	
kappa1	0.13345	0.40919	0.33	0.74433	
kappa2	1.25047	0.40781	3.07	0.00217	**
s_age	0.00725	0.00403	1.80	0.07219	.
s_educ_higher	0.44929	0.09295	4.83	1.3e-06	***
s_hh_income4001_8000	-1.06428	0.25627	-4.15	3.3e-05	***
s_hh_income8001_12000	-0.89366	0.25137	-3.56	0.00038	***
s_hh_income12001_16000	-0.72192	0.26184	-2.76	0.00583	**
s_hh_income16001+	-0.69387	0.28776	-2.41	0.01590	*
s_hh_incomeNA	-0.63145	0.34501	-1.83	0.06722	.
s_young_kids	0.29617	0.10095	2.93	0.00335	**
s_workload	0.05353	0.02404	2.23	0.02598	*
s_fixed_workplace	-0.55419	0.14298	-3.88	0.00011	***
s_shift_work	-0.82518	0.16677	-4.95	7.5e-07	***
s_permanent_employed	0.33270	0.18560	1.79	0.07305	.
s_isco_craft	-0.67913	0.22364	-3.04	0.00239	**
s_isco_tech	0.21921	0.13246	1.65	0.09794	.
s_isco_clerical	0.55330	0.09817	5.64	1.7e-08	***
s_isco_elementary	-4.46545	1.29525	-3.45	0.00057	***
s_car_access	-0.71446	0.26447	-2.70	0.00690	**
s_parking_home	0.64134	0.25170	2.55	0.01083	*
s_freq_onl_order	0.20944	0.08812	2.38	0.01747	*
s_grocery_shopper	-0.13267	0.08788	-1.51	0.13116	
o1_(Intercept)	3.90114	0.17240	22.63	< 2e-16	***
o1_sex_male	0.09334	0.05623	1.66	0.09691	.
o1_res_locrural	0.21702	0.09467	2.29	0.02188	*
o1_res_locsuburban	0.10923	0.09818	1.11	0.26593	
o1_res_locurban	-0.01088	0.10899	-0.10	0.92049	
o1_workload	0.06058	0.01314	4.61	4.0e-06	***
o1_permanent_employed	0.29905	0.11690	2.56	0.01052	*
o1_parking_work	0.23222	0.05204	4.46	8.1e-06	***
o2_(Intercept)	3.88702	0.21570	18.02	< 2e-16	***
o2_swiss	0.18517	0.10900	1.70	0.08935	.
o2_res_locrural	0.43405	0.14799	2.93	0.00336	**
o2_res_locsuburban	0.22649	0.14363	1.58	0.11482	
o2_res_locurban	0.17387	0.16409	1.06	0.28933	
o2_young_kids	-0.15630	0.06958	-2.25	0.02469	*
o2_workload	0.07455	0.01515	4.92	8.7e-07	***
o2_parking_work	0.16130	0.07033	2.29	0.02182	*
o3_(Intercept)	3.85223	0.33710	11.43	< 2e-16	***
o3_sex_male	0.23879	0.08908	2.68	0.00735	**
o3_swiss	0.39109	0.12054	3.24	0.00118	**
o3_fixed_workplace	-0.36832	0.12432	-2.96	0.00305	**
o3_permanent_employed	0.54238	0.25706	2.11	0.03487	*
o3_parking_work	0.28905	0.09202	3.14	0.00168	**
sigma1	0.51246	0.02225	23.04	< 2e-16	***
sigma2	0.54030	0.02954	18.29	< 2e-16	***

```
sigma3                  0.54263    0.05799    9.36 < 2e-16 ***
rho1                    0.28712    0.19141    1.50 0.13360
rho2                   -0.18889    0.14349   -1.32 0.18804
rho3                    0.46193    0.22531    2.05 0.04034 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Wald chi2 (null): 2584 on 39 DF, p-value: < 0
Wald chi2 (rho): 8.93 on 3 DF, p-value: < 0.03
```

The reduced model specification (`fit_red`) is not rejected in the likelihood ratio test. Further, there is significant error correlation between the selection process and the outcome process for the UTWers (`rho3`). The Wald-test suggests that the null hypothesis (`rho1` = `rho2` = `rho3` = 0) can be rejected at the 5% level, suggesting that OPSR is beneficial given our model assumptions.

However, so far we have neglected the commute distance which most likely impacts the propensity to telework (see Figure 1) and naturally influences the weekly distance traveled. To illustrate this, the reduced model specification can be updated to include `log_weekly_km`

```
R> fit_commute <- update(fit_red, ~ . + log_commute_km | . + log_commute_km | . +
+    log_commute_km | . + log_commute_km)
R> print(summary(fit_commute), print.call = FALSE)

BFGS maximization, 252 iterations
Return code 0: successful convergence
Runtime: 3.78 secs
Number of regimes: 3
Number of observations: 824 (424, 265, 135)
Estimated parameters: 54

Log-Likelihood: -1195
AIC: 2499
BIC: 2754
Pseudo R-squared (EL): 0.22
Pseudo R-squared (MS): 0.145
Multiple R-squared: 0.42 (0.42, 0.423, 0.411)

Estimates:
                       Estimate Std. error t value Pr(> t)
kappa1                  0.62599    0.33120    1.89 0.05875 .
kappa2                  1.77835    0.33544    5.30 1.1e-07 ***
s_age                   0.00720    0.00415    1.73 0.08301 .
s_educ_higher           0.44123    0.09697    4.55 5.4e-06 ***
s_hh_income4001_8000   -1.16003    0.15062   -7.70 1.3e-14 ***
s_hh_income8001_12000  -1.01456    0.16563   -6.13 9.1e-10 ***
s_hh_income12001_16000 -0.83329    0.16961   -4.91 9.0e-07 ***
s_hh_income16001+      -0.78611    0.20550   -3.83 0.00013 ***
```

```
s_hh_incomeNA           -0.72432    0.22230   -3.26 0.00112 **
s_young_kids             0.31155    0.10171    3.06 0.00219 **
s_workload               0.04390    0.02398    1.83 0.06710 .
s_fixed_workplace       -0.47995    0.14363   -3.34 0.00083 ***
s_shift_work            -0.84467    0.17413   -4.85 1.2e-06 ***
s_permanent_employed     0.29698    0.18939    1.57 0.11686
s_isco_craft            -0.63875    0.23164   -2.76 0.00582 **
s_isco_tech              0.19158    0.13026    1.47 0.14134
s_isco_clerical          0.58015    0.10092    5.75 9.0e-09 ***
s_isco_elementary       -4.01667    4.38821   -0.92 0.36002
s_car_access            -0.80368    0.24670   -3.26 0.00112 **
s_parking_home           0.65692    0.23484    2.80 0.00515 **
s_freq_onl_order         0.22414    0.08937    2.51 0.01214 *
s_grocery_shopper       -0.09852    0.08733   -1.13 0.25929
s_log_commute_km         0.26436    0.04610    5.73 9.8e-09 ***
o1_(Intercept)           3.46983    0.15893   21.83 < 2e-16 ***
o1_sex_male              0.09656    0.04880    1.98 0.04784 *
o1_res_locrural          0.09827    0.09533    1.03 0.30261
o1_res_locsuburban      -0.00984    0.09638   -0.10 0.91867
o1_res_locurban         -0.00992    0.10848   -0.09 0.92712
o1_workload              0.04449    0.01286    3.46 0.00054 ***
o1_permanent_employed    0.19371    0.10561    1.83 0.06661 .
o1_parking_work          0.11790    0.04459    2.64 0.00820 **
o1_log_commute_km        0.32497    0.02930   11.09 < 2e-16 ***
o2_(Intercept)           3.66032    0.19123   19.14 < 2e-16 ***
o2_swiss                 0.13600    0.08849    1.54 0.12430
o2_res_locrural          0.13492    0.12265    1.10 0.27129
o2_res_locsuburban      -0.03779    0.11644   -0.32 0.74550
o2_res_locurban         -0.08499    0.13607   -0.62 0.53222
o2_young_kids           -0.16561    0.06068   -2.73 0.00635 **
o2_workload              0.05065    0.01321    3.83 0.00013 ***
o2_parking_work          0.04425    0.05650    0.78 0.43351
o2_log_commute_km        0.29014    0.03989    7.27 3.5e-13 ***
o3_(Intercept)           3.33474    0.29376   11.35 < 2e-16 ***
o3_sex_male              0.10902    0.08520    1.28 0.20071
o3_swiss                 0.39746    0.10344    3.84 0.00012 ***
o3_fixed_workplace      -0.24254    0.09842   -2.46 0.01373 *
o3_permanent_employed    0.33954    0.18947    1.79 0.07313 .
o3_parking_work          0.23440    0.08658    2.71 0.00678 **
o3_log_commute_km        0.27830    0.05043    5.52 3.4e-08 ***
sigma1                   0.43374    0.01941   22.34 < 2e-16 ***
sigma2                   0.45466    0.02510   18.12 < 2e-16 ***
sigma3                   0.47517    0.04689   10.13 < 2e-16 ***
rho1                     0.24339    0.21835    1.11 0.26499
rho2                    -0.17064    0.15079   -1.13 0.25780
rho3                     0.35299    0.24432    1.44 0.14851
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Wald chi2 (null): 1142 on 43 DF, p-value: < 0
Wald chi2 (rho): 5.37 on 3 DF, p-value: < 0.147
```

where now all goodness of fit indicators improved (in particular R^2 for the continuous outcomes) and the rho coefficients are no longer significant at conventional levels. Similarly, the Wald-test (rho) can no longer reject the null at the 10% level. Meanwhile, some of the coefficients slightly changed in magnitude and rendered insignificant or vice versa. For example, the effect of residential location (`o1_res_loc_rural` and `o2_res_loc_rural`) moderated the effect of commute distance in `fit_red`, suggesting that individuals living in more rural locations tend to have longer commutes.

While this discussion (of omitted variable bias and/or endogeneity) is common for all regression analysis, it highlights here, why error correlation can occur. Both model specifications (`fit_red` and `fit_commute`) produce similar insights in the post-estimation that follows. However, we will demonstrate later, that not accounting for error correlation can lead to reverse (and most likely false) conclusions.

We first define some helper functions to compute treatment effects.

```
R> tw_status <- c("NTW", "NUTW", "UTW")
R> estimated_weekly_km <- function(object, type = "unlog-response") {
+    nReg <- object$nReg
+    out <- vector("list", nReg)
+    counterfacts <- vector("list", nReg)
+    for (g in 1:nReg) {
+      for (c in 1:nReg) {
+        counterfacts[[c]] <- predict(object, group = g, counterfact = c, type = type)
+      }
+      df <- as.data.frame(counterfacts)
+      names(df) <- tw_status
+      out[[g]] <- df
+    }
+    names(out) <- tw_status
+    out
+  }
R> average <- function(object) {
+    ae <- lapply(object, function(x) {
+      apply(x, 2, function(x) mean(x, na.rm = TRUE))
+    })
+    as.data.frame(ae)
+  }
R> pairwise_diff <- function(mat) {
+    n <- nrow(mat)
+    m <- ncol(mat)
+    result <- matrix(NA, nrow = n, ncol = m)
+    for (j in 1:m) {
+      result[, j] <- c(
```

```
+         mat[2, j] - mat[1, j],
+         mat[3, j] - mat[1, j],
+         mat[3, j] - mat[2, j]
+       )
+     }
+     rownames(result) <- c("NTW -> NUTW", "NTW -> UTW", "NUTW -> UTW")
+     colnames(result) <- c("NTW", "NUTW", "UTW")
+     result
+   }
R> ate <- function(object) {
+     awk <- average(estimated_weekly_km(object))
+     ate <- pairwise_diff(awk)
+     ate
+   }
R> ate(fit_commute)
```

```
              NTW  NUTW    UTW
NTW -> NUTW  14.0 -20.5 -59.87
NTW -> UTW  -50.2 -53.1 -63.14
NUTW -> UTW -64.2 -32.7  -3.27
```

Telework reduces weekly kilometers traveled across all groups with the exception of NTWers who would slightly be more mobile when switching from NTW to NUTW (`NTW -> NUTW`). The treatment effects when switching from NTW to NUTW are strongest for UTWers, who have generally longer commutes. Treatment effects for NTW to UTW are similar across all three groups, again slightly stronger for UTWers. Interestingly, NTWers show a non-linear pattern, first increasing weekly kilometers when adopting some telework (NTW to NUTW) but then substantially decreasing weekly kilometers with more telework (NUTW to UTW). An explanation could be, that these individuals (living closer to their workplace) do initially not adjust activity chains and location choices when only occasionally teleworking. For example, an individual might stay subscribed to the gym close to the workplace and visit that facility even on a home office day. On the other hand, UTWers show somewhat an inverse pattern, first (NTW to NUTW) strongly reducing weekly kilometers but upon further telework adoption (NUTW to UTW) only minimally adjusting weekly kilometers. A similar argument could be made, that these individuals (living further from their workplace) already from the start adjust activity chains and location choices. One can therefore conclude, that the treatment effect over the full range (NTW to UTW) is similar across all groups but the main travel reduction happens at different treatment intensities. Figure 2 (panel d) visualizes these average treatment effects and shows the linear pattern for NUTW and the (mirrored) hockey stick pattern for NTW and UTW.

While the discussion above was based on average treatment effects, Figure 2 shows the distributions of predicted weekly distance traveled by teleworker group. Each panel presents a pair of (un)treated telework statuses as the margins and the dashed lines are the empirical sample means. The solid black reference line marks the instances where weekly distance traveled is equal for both of the paired (un)treated telework statuses. I.e., points below the reference line indicate more travel under the regime depicted on the x-axis.

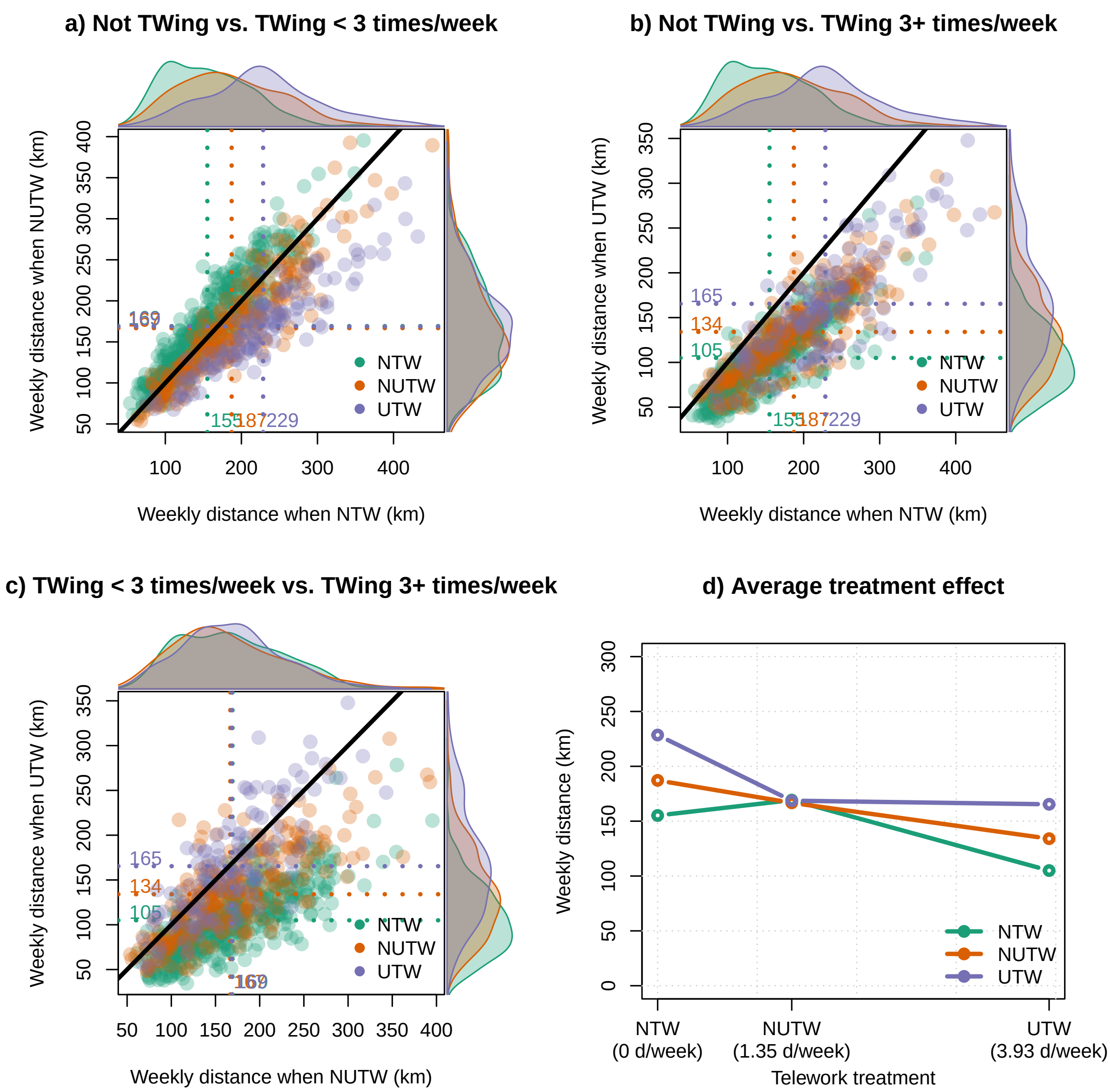

Figure 2: Treatment effects.

As already alluded, not controlling for commute distance implies that selection on unobservables exists, leading to error correlation and selection bias if not accounted for. As we will illustrate now, this also compromises treatment effects. Recalling that `fit_red` is our final model without commute distance (but significant error correlation, as we have seen), we derive a model (`fit_nocor`) without error correlation by setting the `rho` coefficients to 0. I.e., this is the same as separately estimating an ordered probit model and three linear regression models.

```
R> start <- coef(fit_red)
R> fixed <- c("rho1", "rho2", "rho3")
R> start[fixed] <- 0
```

```
R> fit_nocor <- opsr(f_red, timeuse_data, start = start, fixed = fixed,
+    printLevel = 0)
```

The average treatment effects are

```
R> ate(fit_red)

              NTW  NUTW   UTW
NTW -> NUTW  37.7 -10.6 -54.2
NTW -> UTW  -54.7 -47.0 -44.3
NUTW -> UTW -92.4 -36.4   9.9

R> ate(fit_nocor)

              NTW  NUTW   UTW
NTW -> NUTW 17.72 14.14 14.55
NTW -> UTW   8.80 12.31  7.89
NUTW -> UTW -8.93 -1.83 -6.66
```

While resulting treatment effects based on `fit_red` are comparable to the ones based on `fit_commute`, `fit_nocor` yields completely different insights, in particular, that telework generally increases weekly distance traveled.

Recall that for `fit_commute`, the Wald-test (rho) could not reject the null (p value 0.15). Therefore, adding `log_commute_km` to the model without error correlation (`fit_nocor`) might yield less biased treatment effects

```
R> start <- coef(fit_commute)
R> start[fixed] <- 0
R> fit_nocor2 <- opsr(fit_commute$formula, timeuse_data, start = start,
+    fixed = fixed, printLevel = 0)
R> ate(fit_nocor2)

               NTW   NUTW    UTW
NTW -> NUTW  0.421  -2.15  -3.61
NTW -> UTW  -9.054 -11.11 -20.93
NUTW -> UTW -9.475  -8.96 -17.32
```

where now the direction of the treatment effects aligns with the OPSR models but the values are still considerably different. Since `fit_nocor2` is a nested model of `fit_commute` we can conduct a likelihood ratio test

```
R> print(anova(fit_nocor2, fit_commute), print.formula = FALSE)

Likelihood Ratio Test

    logLik      Df      Test Restrictions Pr(>Chi)
1 -1198.04   51.00
2 -1195.47   54.00      5.14            3     0.16
```

which does not reject the null (at the 10% level) that the simpler model is sufficient. As a conclusion should be noted that the modeled covariance matrix (in particular the magnitude of `rho`) potentially strongly influences the treatment effects.

Lastly (using `fit_commute`), we would like to investigate to what degree foregone commute distance (when teleworking) is compensated with leisure travel. Therefore, we compute unit treatment effects and compare them to the average two-way commute distance for each group. The unit treatment effect is calculated by dividing the total treatment effect by the corresponding average teleworking frequency difference (`twdiff1` to `twdiff3` below). I.e., the treatment effect is standardized and therefore also comparable for different regime switching (e.g., NTW to NUTW vs. NUTW to UTW).

```
R> dat_ute <- subset(timeuse_data, select = c(commute_km, wfh, wfh_days))
R> dat_ute <- aggregate(cbind(wfh_days, 2 * commute_km) ~ wfh, data = dat_ute,
+    FUN = mean)
R> top <- t(dat_ute[2:3])
R> colnames(top) <- c("NTW", "NUTW", "UTW")
R> rownames(top) <- c("WFH (days)", "2-way commute (km)")
R> i <- "WFH (days)"
R> twdiff1 <- top[i, "NUTW"] - top[i, "NTW"]
R> twdiff2 <- top[i, "UTW"] - top[i, "NTW"]
R> twdiff3 <- top[i, "UTW"] - top[i, "NUTW"]
R> twdiff <- matrix(c(rep(twdiff1, 3), rep(twdiff2, 3), rep(twdiff3, 3)), nrow = 3)
R> bottom <- ate(fit_commute) / twdiff
R> ute <- rbind(top, bottom)
R> ute
```

```
                    NTW   NUTW    UTW
WFH (days)          0.0   1.35   3.93
2-way commute (km)  30.1  43.33  51.07
NTW -> NUTW         10.4  -5.20 -23.15
NTW -> UTW         -37.3 -13.51 -24.42
NUTW -> UTW        -47.7  -8.31  -1.26
```

Generally, telework reduces weekly distance traveled by less than the foregone commute distance, which indicates, that a rebound effect (compensating leisure travel) exists. For example, the NUTWers could save 43.33 km in commute travel but only reduce 5.2 km per marginal teleworking day when switching from NTW to NUTW. This compensating travel exists for all TW groups except the NTWers (NTW to UTW and NUTW to UTW), where we observe diminished travel activity beyond foregone commutes. The insights from the discussion on average treatment effects caries over: Adjustments in weekly distance traveled are very different both across the three teleworker groups but also across the regime switching.

5. Summary and discussion

In a real-world setting, the treatment is usually not exogenously prescribed but self-selected. Various methods in various statistical environments exist to account for selection-bias which arises if unobserved factors simultaneously influence both the selection and outcome process. OPSR is introduced as a special case of endogenous switching regression. The model frame for such Heckman-type models as well as their implementation in the R system for statistical computing is reviewed. The here presented R implementation in package **OPSR** re-uses design and functionality of the corresponding R software. Hence, the new function `opsr()` is straightforward to apply for model fitting and diagnostics. Further, it is fast and memory efficient thanks to the C++ implementation of the log-likelihood function which can also be parallelized. **OPSR** handles log-transformed outcomes which need special consideration when computing conditional expectations and thus treatment effects. In the case study, the OPSR method is applied to a tracking and activity diary dataset, investigating the telework treatment effects on weekly distance traveled. We demonstrate, first, why error correlation occurs and second, in how far computed treatment effects differ if the error correlation is not accounted for. We find that non-teleworkers tend to have shorter commutes and adjust mobility patterns mainly when switching from non-usual telework to usual telework. On the other hand, weekly distance traveled slightly increases when initially adopting some telework. Contrary, usual teleworkers (had they not been teleworking) adjust mobility patterns strongly when adopting some telework but then only marginally reduce distance traveled when further adopting telework. Comparing the unit treatment effects to the two-way commute distance indicates that telework generally reduces weekly distance traveled and it does so by less than the foregone commute. Therefore, some compensating travel (rebound effects) exists for most of the teleworker groups.

Computational details

The results in this paper were obtained using R 4.4.0 with the packages **OPSR** 0.1.2.9001, **MASS** 7.3.60, **texreg** 1.39.4 and **gridExtra** 2.3. R itself and all packages used are available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/`.

Acknowledgments

References

Amemiya T (1985). *Advanced Econometrics*. Harvard University Press, Cambridge, Massachusetts.

Cameron AC, Trivedi PK (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press, Cambridge.

Chiburis R, Lokshin M (2007). "Maximum Likelihood and Two-Step Estimation of Ordered-Probit Selection Model." *The Stata Journal*, **7**(2), 167–182. `doi:10.1177/1536867X0700700202`.

Eddelbuettel D, Balamuta JJ (2018). "Extending R with C++: A Brief Introduction to **Rcpp**." *The American Statistician*, **72**(1), 28–36. `doi:10.1080/00031305.2017.1375990`.

Eddelbuettel D, Sanderson C (2014). "**RcppArmadillo**: Accelerating R with high-performance C++ linear algebra." *Computational Statistics and Data Analysis*, **71**, 1054–1063. doi:10.1016/j.csda.2013.02.005.

Greene WH (2002). *LIMDEP Version 8.0 Econometric Modeling Guide, vol. 2.* Econometric Software, Plainview, New York.

Heckman J (1979). "Sample Selection Bias as a Specification Error." *Econometrica*, **47**(1), 153–161. doi:10.2307/1912352.

Henningsen A, Toomet O (2011). "**maxLik**: A package for maximum likelihood estimation in R." *Computational Statistics*, **26**(3), 443–458. doi:10.1007/s00180-010-0217-1.

Leifeld P (2013). "**texreg**: Conversion of Statistical Model Output in R to LaTeX and HTML Tables." *Journal of Statistical Software*, **55**(8), 1–24. doi:10.18637/jss.v055.i08.

Potanin B (2024). ***switchSelection**: Endogenous Switching and Sample Selection Regression Models.* R package version 2.0.0, URL https://CRAN.R-project.org/package=switchSelection.

R Core Team (2017). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Roodman D (2011). "Fitting Fully Observed Recursive Mixed-Process Models with **cmp**." *The Stata Journal*, **11**(2), 159–206. doi:10.1177/1536867X110110020.

StataCorp (2023). *Stata Statistical Software.* StataCorp LLC, College Station, Texas. URL https://www.stata.com/.

Toomet O, Henningsen A (2008). "Sample Selection Models in R: Package **sampleSelection**." *Journal of Statistical Software*, **27**(7), 1–23. doi:10.18637/jss.v027.i07.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S.* 4th edition. Springer-Verlag, New York. doi:10.1007/978-0-387-21706-2.

Wang X, Mokhtarian PL (2024). "Examining the Treatment Effect of Teleworking on Vehicle-Miles Driven: Applying an Ordered Probit Selection Model and Incorporating the Role of Travel Stress." *Transportation Research Part A*, **186**, 104072. doi:10.1016/j.tra.2024.104072.

Winkler C, Axhausen KW (2024). "How Do the Swiss Spend Their Time?" *Findings*. doi:10.32866/001c.108600.

Winkler C, Meister A, Axhausen KW (2024). "The TimeUse+ Data Set: 4 Weeks of Time Use and Expenditure Data Based on GPS Tracks." *Transportation*, pp. 1–27. doi:10.1007/s11116-024-10517-1.

Zeileis A (2006). "Object-Oriented Computation of Sandwich Estimators." *Journal of Statistical Software*, **16**(9), 1–16. doi:10.18637/jss.v016.i09.

Zeileis A, Croissant Y (2010). "Extended Model Formulas in R: Multiple Parts and Multiple Responses." *Journal of Statistical Software*, **34**(1), 1–13. doi:10.18637/jss.v034.i01.

Affiliation:

Daniel Heimgartner
Institute for Transport Planning and Systems
Eidgenössische Technische Hochschule Zürich
IFW C 46.1
Haldeneggsteig 4
8092 Zürich, Switzerland
E-mail: daniel.heimgartner@ivt.baug.ethz.ch

Xinyi Wang
Department of Urban Studies and Planning
Massachusetts Institute of Technology
105 Massachusetts Avenue
Cambridge, MA 02139
E-mail: xinyi174@mit.edu