Scientific Annals of Computer Science vol. 33 (1), 2023, pp. 35–52
doi: 10.7561/SACS.2023.1.35

Shrinkage Estimators for the Intercept in Linear and Uplift Regression

Szymon Jaroszewicz[1,2] and Krzysztof Rudaś[1,2]

Abstract

Shrinkage estimators modify classical statistical estimators by scaling them towards zero in order to decrease their prediction error. We propose shrinkage estimators for linear regression models which explicitly take into account the presence of the intercept term, shrinking it independently from other coefficients. This is different from current shrinkage estimators, which treat the intercept just as an ordinary regression coefficient. We demonstrate that the proposed approach brings systematic improvements in prediction accuracy if the true intercept term differs in magnitude from other coefficients, which is often the case in practice. We then generalize the approach to uplift regression which aims to predict the causal effect of a specific action on an individual with given characteristics. In this case the proposed estimators improve prediction accuracy over previously proposed shrinkage estimators and achieve impressive performance gains over original models.

1 Introduction

Linear regression is arguably the most important type of statistical model. The most frequently used estimator for the model is the so called Ordinary Least Squares (OLS) estimator based on minimizing the squared error [2]. The method is very well understood theoretically and offers good predictive accuracy in many practical cases.

[1]Institute of Computer Science, Polish Academy of Sciences
[2]Faculty of Mathematics and Information Science, Warsaw University of Technology

However, the OLS estimator is rarely optimal, and several methods of lowering its predictive error have been developed, primarily by reducing the estimator's variance at the expense of introducing bias [2]. The most popular class of such methods is based on regularization. Another, somewhat less frequently used, class of such estimators are *shrinkage estimators* which scale the Ordinary Least Squares estimate by a factor $\alpha < 1$. While less popular, shrinkage estimators offer several advantages over regularization based methods; for example, there is no need to select regularization parameters through cross-validation. The best known such estimator is the James-Stein estimator [5] which has a provably lower expected mean squared error than the classical OLS estimator, even though the shrinkage factor α is calculated based on the training data. Another choice is a class of shrinkage estimators based on minimizing predictive MSE directly and substituting sample estimates for unknown parameters [7].

In this work we analyze ways to shrink OLS estimators for models which involve the intercept term. The intercept coefficient is estimated somewhat differently from other coefficients and often takes values very different from them, so it seems reasonable to apply a different shrinkage factor to it. Current shrinkage estimators typically ignore the different nature of the intercept and use the same shrinkage factor as for the remaining coefficients. Another strategy (also used in regularization based estimators) is to keep the original OLS estimate for the intercept term and only shrink the remaining coefficients. Here, we propose to use of a separate shrinkage factor for the intercept term of the OLS estimator and demonstrate the benefits of such an approach.

1.1 Uplift Modeling

Another estimation problem discussed in this paper is the development of shrinkage estimators for *uplift regression*. The aim of uplift modeling is to estimate the *causal* effect of an action, such as a medical treatment or a marketing campaign on a given individual [8, 11].

To clarify the nature of the problem, let us give an example. Consider an online shop which, in order to increase sales, offers discounts to selected customers. Some of the customers were not going to buy but the discount changed their mind. Clearly, this results in increased revenue for the shop. Another kind of customers were going to buy from the shop anyway, and used the discount simply to spend less money. Putting issues of customer loyalty aside, the discount resulted in a loss of income for the store. Of course, only

the first group is of interest to shop owners, but classical regression analysis is not able to distinguish such customers.

In fact, the proper way to select targets for an action is to consider the difference between the response in case an individual is subjected to the action (treated) and the response when the individual is not subjected to the action (control). Unfortunately these two pieces of information are never known to us simultaneously. Once we send the discount we cannot make the customer forget about it. This is known in literature as the Fundamental Problem of Causal Inference [3].

Uplift modeling is a solution to this problem based on dividing the available training sample into two parts: the *treatment* group, subjected to the action, and the *control* group on which the action is not taken. This second group is used as a background against which the true benefit of taking the action can be assessed. In uplift regression our aim will be to estimate the *difference* between treatment and control responses for an object described by a feature vector x [9]. More details on the problem are given in Section 4.

The second contribution of this work is developing new shrinkage estimators for uplift regression. The first such estimators have been proposed in [10] showing clear practical benefits. However, those methods shrunk the intercept term using the same factor as the remaining coefficients. Here we develop estimators which use separate shrinkage factors for the intercept term and demonstrate experimentally that they give improvements in uplift regression's predictions accuracy.

1.2 Notation

Let us now introduce the notation used throughout the paper. Lowercase Greek and Latin letters, e.g. α, β, x, y will denote vectors, and uppercase letters, e.g. X will denote matrices. Matrix transpose will be denoted with the prime symbol $'$, matrix trace with Tr, and $0_{n\times m}$ and $1_{n\times m}$ will be used to denote matrices of, respectively, all zeros and all ones with n rows and m columns. I_n will denote the $n \times n$ identity matrix.

Vector and matrix random variables will be denoted, respectively, with lower- and uppercase boldface letters $\boldsymbol{X}$, $\boldsymbol{y}$. Scalar random variables will also be denoted with bold face lowercase letters. Statistical estimators will be denoted with the usual symbol $\hat{}$ above a variable name. Even though the estimators are random variables, boldface will not be used to avoid notational clutter. E will denote the expectation of a random variable and

Var the covariance matrix of a random vector. Quantities related to test data will be denoted with subscript t.

Notation specific to uplift modeling will be introduced in Section 4.

2 Shrinkage Estimator for the Intercept Term in Ordinary Least Squares

We begin by describing the classical Ordinary Least Squares (OLS) regression methodology. Only facts needed to understand the remaining part of the paper are given, full exposition can be found e.g. in [2]. We assume that we have a random vector $\boldsymbol{x} \in \mathbb{R}^p$ of predictor variables, a real-valued response variable $\boldsymbol{y}$, and that there exists a fixed joint distribution P over $\boldsymbol{x}$ and $\boldsymbol{y}$.

Further we assume that we have a *training set* which consist of n samples from the joint distribution arranged in an $n \times (p+1)$ matrix $\boldsymbol{X}$ of predictors and an n-dimensional vector $\boldsymbol{y}$ of responses. We assume that the first column of $\boldsymbol{X}$ is a vector of ones and the remaining columns correspond to the p predictor variables. The column of ones will allow for easy treatment of the intercept term.

Further, we assume that $\boldsymbol{y}$ is related to $\boldsymbol{X}$ through a linear equation

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}, \tag{1}$$

where β is an unknown coefficient vector and $\boldsymbol{\varepsilon}$ is a random noise vector satisfying the usual assumptions that $\mathrm{E}\,\boldsymbol{\varepsilon} = 0_{n\times 1}$, $\mathrm{Var}\,\boldsymbol{\varepsilon} = \sigma^2 I_n$, and the components of $\boldsymbol{\varepsilon}$ are independent of each other and independent from $\boldsymbol{X}$ [2].

The first component of β, i.e. β_0 is called the *intercept term* and is responsible for the constant offset of $\boldsymbol{y}$.

Notice that we consider the training set $\boldsymbol{X}$, $\boldsymbol{y}$ to be random. This point of view will be used in our analyses. In practice we only have one realization of $\boldsymbol{X}$, $\boldsymbol{y}$ available, which we will denote with letters X, y.

Our goal is to find an estimator $\hat{\beta}$ of β which, on a new test sample $\boldsymbol{x}_t$ (for notational simplicity we assume that the test sample is augmented with a constant 1 in the first coordinate), $\boldsymbol{y}_t$ drawn from the distribution P achieves the lowest possible *predictive mean squared error*

$$\mathrm{MSE}(\hat{\beta}) = \mathrm{E}_{\boldsymbol{X}}\,\mathrm{E}_{\boldsymbol{y}}\,\mathrm{E}_{\boldsymbol{x}_t}\,\mathrm{E}_{\boldsymbol{y}_t}(\boldsymbol{y}_t - \boldsymbol{x}_t\hat{\beta})^2, \tag{2}$$

where the expectation is taken over the test sample as well as over the training set used to obtain $\hat{\beta}$. The most popular estimator is the Ordinary

Least Squares (OLS) estimator obtained by minimizing the training set squared error $\|\boldsymbol{y} - \boldsymbol{X}\hat{\beta}\|^2$. The estimator is given by the equation

$$\hat{\beta}_{\text{OLS}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{3}$$

It is well known that $\hat{\beta}_{\text{OLS}}$ is unbiased, i.e. $\mathrm{E}_{\boldsymbol{X}}\,\mathrm{E}_{\boldsymbol{y}}\,\hat{\beta}_{\text{OLS}} = \beta$ [2].

We now move on to derive the new proposed shrinkage estimators. We will begin by using a very general form of a shrinkage estimator based on Ordinary Least Squares:

$$\hat{\beta}_\alpha = \alpha \odot \hat{\beta}_{\text{OLS}} = \alpha \odot (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}, \tag{4}$$

where $\alpha \in \mathbb{R}^{p+1}$ is a vector of nonnegative shrinkage coefficients and $\odot$ is the Hadamard (elementwise) product of vectors and matrices. Notice that in this setting every regression coefficient β_i may be shrunk separately with a different shrinkage factor α_i. We begin by analyzing this general shrinkage estimator and later present several restrictions to it. In the most important one, all coefficients will have equal shrinkage factors, except for the intercept which will be shrunk using a separate coefficient.

Let us now calculate the predictive MSE (Equation 2) of the general shrinkage estimator given in Equation 4. Let us first present a simple relationship between the Hadamard product of vectors and diagonal matrices. For any two vectors α and β we have

$$\alpha \odot \beta = \operatorname{diag}(\alpha)\beta = \operatorname{diag}(\beta)\alpha, \tag{5}$$

where $\operatorname{diag}(\alpha)$ is a diagonal matrix with main diagonal equal to α. The expectation of the shrinkage estimator $\hat{\beta}_\alpha$ can thus be written as

$$\mathrm{E}_{\boldsymbol{X}}\,\mathrm{E}_{\boldsymbol{y}}\,\hat{\beta}_\alpha = \operatorname{diag}(\alpha)\,\mathrm{E}_{\boldsymbol{X}}\,\mathrm{E}_{\boldsymbol{y}}\,\hat{\beta}_{\text{OLS}} = \operatorname{diag}(\alpha)\beta = \alpha \odot \beta, \tag{6}$$

where β is the true coefficient vector (see Equation 1), and the second equality comes from the fact that $\hat{\beta}_{\text{OLS}}$ is unbiased. Define $S_t = \mathrm{E}_{\boldsymbol{x}_t}\,\boldsymbol{x}_t\boldsymbol{x}_t'$ and let $\boldsymbol{\varepsilon}_t$ denote the random noise term on the test sample. We now have

$$\begin{aligned}
\operatorname{MSE}(\hat{\beta}_\alpha) &= \mathrm{E}_{\boldsymbol{x}_t}\,\mathrm{E}_{\boldsymbol{y}_t}\,\mathrm{E}_{\boldsymbol{X}}\,\mathrm{E}_{\boldsymbol{y}}(\boldsymbol{y}_t - \boldsymbol{x}_t'\hat{\beta}_\alpha)'(\boldsymbol{y}_t - \boldsymbol{x}_t'\hat{\beta}_\alpha) && (7)\\
&= \mathrm{E}_{\boldsymbol{x}_t}\,\mathrm{E}_{\boldsymbol{\varepsilon}_t}\,\mathrm{E}_{\boldsymbol{X}}\,\mathrm{E}_{\boldsymbol{y}}(\boldsymbol{x}_t'\beta + \boldsymbol{\varepsilon}_t - \boldsymbol{x}_t'\hat{\beta}_\alpha)'(\boldsymbol{x}_t'\beta + \boldsymbol{\varepsilon}_t - \boldsymbol{x}_t'\hat{\beta}_\alpha) && (8)\\
&= \mathrm{E}_{\boldsymbol{x}_t}\,\mathrm{E}_{\boldsymbol{X}}\,\mathrm{E}_{\boldsymbol{y}}(\beta - \hat{\beta}_\alpha)'\boldsymbol{x}_t\boldsymbol{x}_t'(\beta - \hat{\beta}_\alpha) + \mathrm{E}_{\boldsymbol{\varepsilon}_t}\,\boldsymbol{\varepsilon}_t^2 && (9)\\
&= \mathrm{E}_{\boldsymbol{X}}\,\mathrm{E}_{\boldsymbol{y}}(\beta - \hat{\beta}_\alpha)'S_t(\beta - \hat{\beta}_\alpha) + \sigma^2 && (10)\\
&= (\beta - \mathrm{E}_{\boldsymbol{X}}\,\mathrm{E}_{\boldsymbol{y}}\,\hat{\beta}_\alpha)'S_t(\beta - \mathrm{E}_{\boldsymbol{X}}\,\mathrm{E}_{\boldsymbol{y}}\,\hat{\beta}_\alpha) + \operatorname{Tr}(S_t\operatorname{Var}(\hat{\beta}_\alpha)) + \sigma^2 && (11)
\end{aligned}$$

$$= (\beta - \alpha \odot \beta)' S_t (\beta - \alpha \odot \beta) + \mathrm{Tr}(S_t \,\mathrm{diag}'(\alpha)(\mathrm{Var}\,\hat{\beta}_{\mathrm{OLS}})\,\mathrm{diag}(\alpha)) + \sigma^2 \quad (12)$$

$$= (\beta - \alpha \odot \beta)' S_t (\beta - \alpha \odot \beta) + \alpha'(S_t \odot \mathrm{Var}\,\hat{\beta}_{\mathrm{OLS}})\alpha + \sigma^2, \quad (13)$$

where Equation 8 follows from the linear model assumptions given in Equation 1, Equation 9 from the independence of $\boldsymbol{\varepsilon}_t$ from all other variables, and Equation 10 follows from the independence of $\boldsymbol{x}_t$ from $\boldsymbol{X}$ and $\boldsymbol{y}$. Equation 11 follows from the bias variance decomposition [2, 10] and the properties of the trace of a matrix, Equation 12 by substituting the expectation of $\hat{\beta}_\alpha$ given in Equation 6, and Equation 13 from the properties of the Hadamard product, specifically [4, Lemma 5.1.5].

The shrinkage coefficient vector α will be chosen by minimizing the predictive mean squared error given in Equation 13. First, however, we need to provide a way to obtain concrete estimators from the very general Equation 4.

Equation 4 gives a form of shrinkage estimators with every regression coefficient having a separate shrinkage factor. In practice we want several coefficients to share a single shrinkage factor. To make this possible, we will assume that the shrinkage coefficient vector has the form

$$\alpha = B\gamma, \quad (14)$$

where γ is a vector of $q \leq p+1$ *unique* shrinkage coefficients and B is an $(p+1) \times q$ matrix. For example, to obtain an estimator in which a single shrinkage factor is shared by all coefficients except the intercept term, which has its own shrinkage factor we can use

$$\alpha = B\gamma = \begin{bmatrix} 1_{1\times 1} & 0_{1\times 1} \\ 0_{p\times 1} & 1_{p\times 1} \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \gamma_1 \end{bmatrix},$$

where $0_{n\times m}$ and $1_{n\times m}$ denote respectively the matrices of all zeros and all ones. Now, γ_0 is the shrinkage factor for the intercept term and γ_1 the common shrinkage factor for the remaining coefficients.

Formula 4 now becomes

$$\hat{\beta}_\alpha = \alpha \odot \hat{\beta}_{\mathrm{OLS}} = (B\gamma) \odot \hat{\beta}_{\mathrm{OLS}} = (B\gamma) \odot (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}, \quad (15)$$

which is the final form of the proposed shrinkage estimator.

We now need to find the optimal value of γ. To this end we will take the vector derivative of Equation 13 over γ:

$$\begin{aligned}
&\frac{\partial}{\partial\gamma}\left[(\beta-\alpha\odot\beta)'S_t(\beta-\alpha\odot\beta)+\alpha'(S_t\odot\operatorname{Var}\hat{\beta}_{\mathrm{OLS}})\alpha+\sigma^2\right]\\
&\quad=\frac{\partial}{\partial\gamma}(\beta-(B\gamma)\odot\beta)'S_t(\beta-(B\gamma)\odot\beta)+\frac{\partial}{\partial\gamma}(B\gamma)'(S_t\odot\operatorname{Var}\hat{\beta}_{\mathrm{OLS}})(B\gamma)\\
&\quad=2(\beta-(B\gamma)\odot\beta)'S_t\frac{\partial}{\partial\gamma}(\beta-(B\gamma)\odot\beta)+2(B\gamma)'(S_t\odot\operatorname{Var}\hat{\beta}_{\mathrm{OLS}})\frac{\partial B\gamma}{\partial\gamma}\\
&\quad=2(\operatorname{diag}(\beta)B\gamma-\beta)'S_t\operatorname{diag}(\beta)B+2(B\gamma)'(S_t\odot\operatorname{Var}\hat{\beta}_{\mathrm{OLS}})B,
\end{aligned}$$

where the first equality follows by substituting Equation 14, and the remaining ones from basic rules of matrix calculus. Transposing and equating to zero we get

$$B'\operatorname{diag}(\beta)S_t(\operatorname{diag}(\beta)B\gamma-\beta)+B'(S_t\odot\operatorname{Var}\hat{\beta}_{\mathrm{OLS}})B\gamma=0_{q\times1}\qquad(16)$$

and finally

$$B'\left[\operatorname{diag}(\beta)S_t\operatorname{diag}(\beta)+(S_t\odot\operatorname{Var}\hat{\beta}_{\mathrm{OLS}})\right]B\gamma=B'\operatorname{diag}(\beta)S_t\beta.\qquad(17)$$

By solving the last system of equations we obtain the optimal shrinkage vector α or, more specifically, the vector of its unique components γ.

Unfortunately, the estimate is not operational, since we do not know the true regression coefficients β or the matrix S_t. In order to obtain an operational estimate $\hat{\gamma}$, we will take the approach used e.g. in [7, 10] by replacing the unknown parameters with estimates obtained from the training sample. Such estimators are known in statistics as *plugin estimators*.

The final proposed estimator $\hat{\gamma}$ of γ is obtained by solving the system of equations

$$B'\left[\operatorname{diag}(\hat{\beta}_{\mathrm{OLS}})\hat{S}\operatorname{diag}(\hat{\beta}_{\mathrm{OLS}})+(\hat{S}\odot\operatorname{Var}\hat{\beta}_{\mathrm{OLS}})\right]B\hat{\gamma}=B'\operatorname{diag}(\hat{\beta}_{\mathrm{OLS}})\hat{S}\hat{\beta}_{\mathrm{OLS}},\qquad(18)$$

where all estimates will be based on a specific realization X, y of random training data $\boldsymbol{X}$, $\boldsymbol{y}$ available to us. We will use the following estimators

$$\hat{\beta}_{\mathrm{OLS}}=(X'X)^{-1}X'y\qquad\hat{S}=\frac{1}{n}X'X\qquad\operatorname{Var}\hat{\beta}_{\mathrm{OLS}}=\frac{r'r}{n-p}(X'X)^{-1},\qquad(19)$$

where r is the residual vector $r=y-X\hat{\beta}_{\mathrm{OLS}}$. Note that the variance of the OLS estimator is computed assuming a fixed, nonrandom X. This simplification is necessary since there is no general estimator of variance of the OLS estimator under random predictors.

2.1 Concrete Shrinkage Estimators for the OLS

In this section we present several shrinkage estimators for linear regression obtained by choosing a different matrix B in Equation 15.

Single This case corresponds to the classical shrinkage estimators with a single common shrinkage factor for all coefficients including the intercept. Here we have

$$B_{\text{single}} = 1_{(p+1)\times 1},$$

i.e. B is just a column of ones, and γ is a one element vector.

Intercept This model is one of the main contributions of this paper. It uses two separate shrinkage coefficients: one for the intercept term and another one for all remaining coefficients. Here

$$B_{\text{int}} = \begin{bmatrix} 1_{1\times 1} & 0_{1\times 1} \\ 0_{p\times 1} & 1_{p\times 1} \end{bmatrix}, \tag{20}$$

and γ is a two element vector.

Full For the sake of completeness we also investigated the possibility of each model coefficient (including the intercept) having a separate shrinkage factor. This corresponds to the B matrix equal to the identity matrix

$$B_{\text{full}} = I_{(p+1)\times(p+1)},$$

and the γ vector having $p+1$ components.

No Intercept We also tested a strategy where the intercept is not shrunk at all: its scaling factor is kept constant at 1. The remaining coefficients share a common shrinkage factor. Unfortunately, this type of estimator cannot be achieved by simply choosing an appropriate matrix B.

To achieve the desired result, we used the B matrix of the Intercept estimator (Equation 20) and replaced the first equation in the system given in Equation 18 with $\gamma_0 = 0$.

3 Shrinked OLS Regression. Experimental Evaluation

In this section we present an experimental evaluation of the proposed shrinkage estimators. In order to be able to fully control the experiment and achieve credible results we resorted to the use of simulated data.

The simulation procedure was conducted as follows. First we chose the true coefficient vector β based on which the data were simulated. We set $\beta_i = 1$ for odd $i > 0$ and $\beta_i = 0.5$ for even $i > 0$. The experiments were conducted for five different values of the intercept term β_0: 0.01, 0.1, 1, 10, and 100. This way, values of intercept of magnitude smaller and larger than the remaining coefficients were tested.

We generated a random matrix X with $n = 30$ rows and $p = 20$ columns from a standard normal distribution with all variables uncorrelated. Small value of n was chosen to illustrate the estimators behavior in small sample scenarios where shrinkage estimators are most useful. Then, the response vector y was generated by adding standard normal random noise ($\sigma = 1$) to the vector $X\beta$.

Model coefficients were then estimated using the standard OLS estimator and the shrinkage estimators described above. The predicted mean squared error was computed on a test set with 10 000 records generated using the same procedure.

The experiment was repeated 100 000 times for each value of the intercept term, and the results have been averaged. Table 1 presents the outcomes. The **OLS** estimator is the standard least squares estimator without shrinkage. Notice that the **OLS** estimator and the **No Intercept** estimators do not depend on the true intercept term used.

It can be seen the proposed **Intercept** shrinkage estimator, which uses a separate shrinkage factor for the intercept, achieves the lowest error, except for the case when the true intercept is of the same order of magnitude as the remaining coefficients. The differences are small but consistent. The proposed estimator always outperforms the **No intercept** estimator which does not shrink the intercept. Moreover all shrinkage estimators except **Full** outperform the original **OLS** estimator in all cases. The **Full** estimator, which shrinks each coefficient independently is a clear loser.

The third column provides the standard deviation of the estimated MSE and the fourth its standard error (that is standard deviation devided by the square root of the number of experiments). Values in the last column

Estimator	test set MSE	std. deviation	std. error
$\beta_0 = 0.01$			
OLS	3.2163	1.4245	0.0045
Intercept	**3.0254**	1.2903	0.0041
No intercept	3.0597	1.3083	0.0041
Single	3.0417	1.2938	0.0041
Full	3.2171	1.2316	0.0039
$\beta_0 = 0.1$			
OLS	3.2163	1.4244	0.0045
Intercept	**3.0271**	1.2902	0.0041
No intercept	3.0597	1.3083	0.0041
Single	3.0421	1.2939	0.0041
Full	3.2171	1.2316	0.0039
$\beta_0 = 1$			
OLS	3.2163	1.4244	0.0045
Intercept	3.0573	1.3003	0.0041
No intercept	3.0597	1.3083	0.0041
Single	**3.0536**	1.3022	0.0041
Full	3.2863	1.2597	0.0040
$\beta_0 = 10$			
OLS	3.2163	1.4244	0.0045
Intercept	**3.0496**	1.2979	0.0041
No intercept	3.0597	1.3083	0.0041
Single	3.1956	1.4086	0.0045
Full	3.2687	1.2566	0.0040
$\beta_0 = 100$			
OLS	3.2163	1.4244	0.0045
Intercept	**3.0495**	1.2977	0.0041
No intercept	3.0597	1.3083	0.0041
Single	3.2161	1.4243	0.0045
Full	3.2683	1.2561	0.0040

Table 1: Mean Squared Error of various shrinkage estimators for linear regression for different values of the true intercept term.

provide the precision of the mean MSE averaged over all simulations. It can be seen that, thanks to a relatively large number of simulations, the differences between estimators are significant.

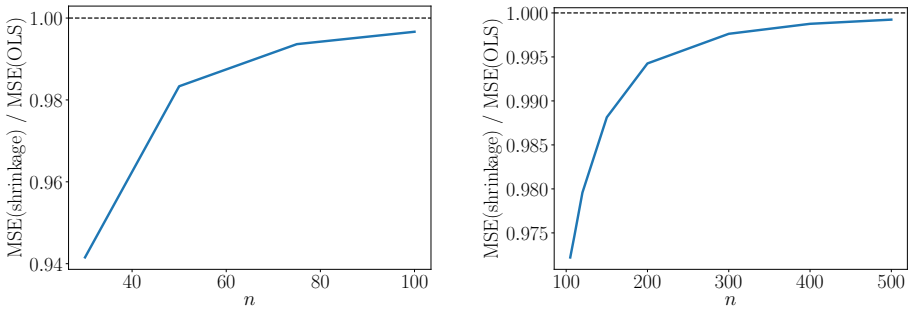

Figure 1: Ratio of mean squared errors of the **Intercept** shrinkage estimator and ordinary least squares estimator for $p = 20$ variables (left) and $p = 100$ variables (right) for growing number of training records (n)

To evaluate the gains form using the estimators for different data sizes we compared the **Intercept** shrinkage estimator which performed best in Table 1 with the standard **OLS** estimator. The value of $\beta_0 = 1$ was used for all charts. The results are shown in Figure 1 for the cases of $p = 20$ and $p = 100$ variables in the model and growing numbers of data records.

It can be seen that the gains in expected mean squared error are relatively small but consistent. Shrinkage estimators are most useful for small datasets. In the following sections we will demonstrate that larger gains are possible for uplift models.

4 Shrinkage Estimators for Uplift Regression

Let us now proceed to the case of uplift modeling. In this problem we have two training sets: treatment and control. The quantities related to the treatment group will be denoted with superscript T and to the control group with superscript C. Quantities related to the uplift (i.e. the conditional treatment effect) will be denoted with superscript U. For example, β^C will denote the true coefficient vector of the linear response in control cases and β^U the true coefficient vector of the linear strength of effect of the action. Let us now state model assumptions, analogously to Equation 1:

$$\begin{aligned}\boldsymbol{y}^C &= \boldsymbol{X}^C\beta^C + \boldsymbol{\varepsilon}^C,\\ \boldsymbol{y}^T &= \boldsymbol{X}^T\beta^C + \boldsymbol{X}^T\beta^U + \boldsymbol{\varepsilon}^T = \boldsymbol{X}^T\beta^T + \boldsymbol{\varepsilon}^T.\end{aligned}$$

Notice that we assume linear response in the control group (with true coefficient vector β^C) and a linear conditional effect of the action (with true coefficient vector β^U). As a result, the response in the treatment group is also linear with coefficients $\beta^T = \beta^C + \beta^U$. The parameter of interest, which we want to estimate is β^U.

We will assume randomized assignment of cases to the treatment group to guarantee causal nature of discovered relationships [8, 9].

As the base uplift estimator to which shrinkage will be applied, we use the *double model* approach [9] (also known as T-learner), which is also the base model used to obtain shrinkage uplift estimators in [10]. The model is simply the difference of two OLS regression estimators built independently on the treatment and control datasets:

$$\hat{\beta}^U_d = \hat{\beta}^T_{\mathrm{OLS}} - \hat{\beta}^C_{\mathrm{OLS}} = ((\boldsymbol{X}^T)'\boldsymbol{X}^T)^{-1}\boldsymbol{X}^T\boldsymbol{y}^T - ((\boldsymbol{X}^C)'\boldsymbol{X}^C)^{-1}\boldsymbol{X}^C\boldsymbol{y}^C, \quad (21)$$

where the subscript d stands for 'double' and $\hat{\beta}^T_{\mathrm{OLS}}$, $\hat{\beta}^C_{\mathrm{OLS}}$ are OLS estimators of, respectively, treatment and control response coefficient vectors. Notice that since we assume that the treatment assignment is random, this simple model is able to obtain causal predictions.

In order to repeat the derivation given in Section 2 we will now need two shrinkage coefficient vectors α^C and α^T, and two vectors of unique shrinkage coefficients γ^C and γ^T satisfying

$$\alpha^C = B\gamma^C, \qquad \alpha^T = B\gamma^T,$$

where the matrix B is assumed to be identical for both groups. The final form of the proposed uplift shrinkage estimator is

$$\hat{\beta}^U_{\alpha^C,\alpha^T} = \alpha^T \odot \hat{\beta}^T_{\mathrm{OLS}} - \alpha^C \odot \hat{\beta}^C_{\mathrm{OLS}} = (B\gamma^T) \odot \hat{\beta}^T_{\mathrm{OLS}} - (B\gamma^C) \odot \hat{\beta}^C_{\mathrm{OLS}}, \quad (22)$$

which has a similar form to the estimator used in [10], except that we will now allow different shrinkage factors for different coefficients through the use of the matrix B. It is easy to see that the expectation of the estimator is equal to

$$\mathrm{E}\,\hat{\beta}^U_{\alpha^T,\alpha^C} = \mathrm{E}_{\boldsymbol{X}^C}\,\mathrm{E}_{\boldsymbol{y}^C}\,\mathrm{E}_{\boldsymbol{X}^T}\,\mathrm{E}_{\boldsymbol{y}^T}\,\hat{\beta}^U_{\alpha^T,\alpha^C} = \alpha^T \odot \beta^T - \alpha^C \odot \beta^C.$$

Let us now derive the expression for the MSE of the estimator, analogously to Equations 1–13. We have

$$\mathrm{MSE}(\hat{\beta}^U_{\alpha^C,\alpha^T}) = \mathrm{E}_{\boldsymbol{x}_t}\,\mathrm{E}_{\boldsymbol{X}^C}\,\mathrm{E}_{\boldsymbol{y}^C}\,\mathrm{E}_{\boldsymbol{X}^T}\,\mathrm{E}_{\boldsymbol{y}^T}(\boldsymbol{x}_t'\beta^U - \boldsymbol{x}_t\hat{\beta}^U_{\alpha^T,\alpha^C})^2 \quad (23)$$

$$= \mathrm{E}_{\boldsymbol{X}^C}\,\mathrm{E}_{\boldsymbol{y}^C}\,\mathrm{E}_{\boldsymbol{X}^T}\,\mathrm{E}_{\boldsymbol{y}^T}(\beta^U - \hat{\beta}^U_{\alpha^T,\alpha^C})'S_t(\beta^U - \hat{\beta}^U_{\alpha^T,\alpha^C}) \quad (24)$$

$$
\begin{aligned}
&= (\beta^U - \mathrm{E}\,\hat{\beta}^U_{\alpha^T,\alpha^C})' S_t (\beta^U - \mathrm{E}\,\hat{\beta}^U_{\alpha^T,\alpha^C}) + \mathrm{Tr}(S_t \,\mathrm{Var}(\hat{\beta}^U_{\alpha^T,\alpha^C})) && (25)\\
&= (\beta^U - \alpha^T \odot \beta^T + \alpha^C \odot \beta^C)' S_t (\beta^U - \alpha^T \odot \beta^T + \alpha^C \odot \beta^C)\\
&\quad + \mathrm{Tr}(S_t \,\mathrm{Var}(\alpha^T \odot \hat{\beta}^T_{\mathrm{OLS}})) + \mathrm{Tr}(S_t \,\mathrm{Var}(\alpha^C \odot \hat{\beta}^C_{\mathrm{OLS}})) && (26)\\
&= (\beta^U - \alpha^T \odot \beta^T + \alpha^C \odot \beta^C)' S_t (\beta^U - \alpha^T \odot \beta^T + \alpha^C \odot \beta^C)\\
&\quad + \mathrm{Tr}(S_t \,\mathrm{diag}(\alpha^T) \,\mathrm{Var}(\hat{\beta}^T_{\mathrm{OLS}}) \,\mathrm{diag}(\alpha^T))\\
&\quad + \mathrm{Tr}(S_t \,\mathrm{diag}(\alpha^C) \,\mathrm{Var}(\hat{\beta}^C_{\mathrm{OLS}}) \,\mathrm{diag}(\alpha^C)) && (27)\\
&= (\beta^U - \alpha^T \odot \beta^T + \alpha^C \odot \beta^C)' S_t (\beta^U - \alpha^T \odot \beta^T + \alpha^C \odot \beta^C)\\
&\quad + (\alpha^T)'(S_t \odot \mathrm{Var}(\hat{\beta}^T_{\mathrm{OLS}}))\alpha^T + (\alpha^C)'(S_t \odot \mathrm{Var}(\hat{\beta}^C_{\mathrm{OLS}}))\alpha^C,
\end{aligned}
$$

where Equation 26 follows from the independence of treatment and control OLS estimators. We note that the MSE of the treatment effect is often called PEHE in literature [1].

Let us now take the derivative of the above expression with respect to γ^T (the derivation for γ^C is analogous). We have

$$
\begin{aligned}
&\frac{\partial}{\partial \gamma^T}\Big[(\beta^U - \alpha^T \odot \beta^T + \alpha^C \odot \beta^C)' S_t (\beta^U - \alpha^T \odot \beta^T + \alpha^C \odot \beta^C)\\
&\qquad + (\alpha^T)'(S_t \odot \mathrm{Var}(\hat{\beta}^T_{\mathrm{OLS}}))\alpha^T + (\alpha^C)'(S_t \odot \mathrm{Var}(\hat{\beta}^C_{\mathrm{OLS}}))\alpha^C\Big] && (28)\\
&= \frac{\partial}{\partial \gamma^T}(\beta^U - \alpha^T \odot \beta^T + \alpha^C \odot \beta^C)' S_t (\beta^U - \alpha^T \odot \beta^T + \alpha^C \odot \beta^C)\\
&\qquad + \frac{\partial}{\partial \gamma^T}(\alpha^T)'(S_t \odot \mathrm{Var}(\hat{\beta}^T_{\mathrm{OLS}}))\alpha^T && (29)\\
&= 2(\beta^U - (B\gamma^T) \odot \beta^T + (B\gamma^C) \odot \beta^C)' S_t \frac{\partial}{\partial \gamma^T}(\beta^U - \alpha^T \odot \beta^T + \alpha^C \odot \beta^C)\\
&\qquad + 2(B\gamma^T)'(S_t \odot \mathrm{Var}(\hat{\beta}^T_{\mathrm{OLS}}))\frac{\partial}{\partial \gamma^T} B\gamma^T && (30)\\
&= 2(\mathrm{diag}(\beta^T)B\gamma^T - \mathrm{diag}(\beta^C)B\gamma^C - \beta^U)' S_t \,\mathrm{diag}(\beta^T)B\\
&\qquad + 2(B\gamma^T)'(S_t \odot \mathrm{Var}(\hat{\beta}^T_{\mathrm{OLS}}))B. && (31)
\end{aligned}
$$

Transposing and equating to zero we get

$$
\begin{aligned}
&B'\left[\mathrm{diag}(\beta^T) S_t \,\mathrm{diag}(\beta^T) + S_t \odot \mathrm{Var}(\hat{\beta}^T_{\mathrm{OLS}})\right] B\gamma^T\\
&\qquad - B' \,\mathrm{diag}(\beta^T) S_t \,\mathrm{diag}(\beta^C) B\gamma^C = B' \,\mathrm{diag}(\beta^T) S_t \beta^U. && (32)
\end{aligned}
$$

By an analogous argument, taking the derivative over γ^C yields

$$-B' \operatorname{diag}(\beta^C) S_t \operatorname{diag}(\beta^T) B\gamma^T + B' \Big[\operatorname{diag}(\beta^C) S_t \operatorname{diag}(\beta^C) + S_t \odot \operatorname{Var}(\hat{\beta}^C_{\mathrm{OLS}}) \Big] B\gamma^C = -B' \operatorname{diag}(\beta^C) S_t \beta^U. \tag{33}$$

Together, Equations 32 and 33 make up a system of equations through which we can find the shrinkage coefficient vectors γ^C and γ^T.

Unfortunately, as was the case with shrinking the Ordinary Least Squares estimator in Section 2, the system depends on several quantities which are unknown to us during estimation. We take the same strategy as we did in Section 2, i.e. substituting the following training set estimators of those quantities:

$$\hat{\beta}^T_{\mathrm{OLS}} = ((X^T)'X^T)^{-1}(X^T)'y^T, \quad \hat{\beta}^C_{\mathrm{OLS}} = ((X^C)'X^C)^{-1}(X^C)'y^C, \quad \hat{S} = \frac{1}{n} X'X, \tag{34}$$

where X^T, y^T is the treatment training sample, X^C, y^C is the control training sample, X is a concatenation of X^T and X^C, and n is the total number of records in both training sets. The formulas for treatment and control OLS variances are analogous to Equation 19 and the unknown β^U is replaced with $\hat{\beta}^U_d$ given in Equation 21.

Definition 1 *The estimator defined jointly by Equations 22, 32, 33 and 34 will be called the* separately shrinked uplift estimator.

When all coefficients are shrunk identically the estimator becomes the uplift MSE-minimizing estimator from [10].

4.1 An Alternative Definition of Uplift Shrinkage Estimators

Notice that the double model given in Equation 21 yields a single parameter vector $\hat{\beta}^U_d$. If we were able to estimate the variance of this vector, we could directly apply to it the shrinkage estimators developed for OLS regression in Section 2. Since the least squares estimates on treatment and control training sets are independent, the variance of $\hat{\beta}^U_d$ can be estimated simply as

$$\operatorname{Var} \hat{\beta}^U_d = \frac{(r^T)'r^T}{n^T - p}((X^T)'X^T)^{-1} + \frac{(r^C)'r^C}{n^C - p}((X^C)'X^C)^{-1}, \tag{35}$$

where n^T and n^C are the numbers of, respectively, treatment and control training samples, and r^T, r^C treatment and control residual vectors. Notice that the expression above is the sum of variances of treatment and control least squares estimators given in Equation 19. We will substitute this equation in place of the variance of OLS estimator into Equation 19 to obtain the following shrinked uplift estimator:

Definition 2 *An estimator*

$$\alpha \odot \hat{\beta}_d^U = (B\hat{\gamma}) \odot \hat{\beta}_d^U, \tag{36}$$

where $\hat{\gamma}$ is obtained by solving the system of equations

$$B' \left[\operatorname{diag}(\hat{\beta}_d^U) \hat{S} \operatorname{diag}(\hat{\beta}_d^U) + (\hat{S} \odot \operatorname{Var} \hat{\beta}_d^U) \right] B\hat{\gamma} = B' \operatorname{diag}(\hat{\beta}_d^U) \hat{S} \hat{\beta}_d^U, \tag{37}$$

with $\hat{S}$ given in Equation 34 and $\operatorname{Var} \hat{\beta}_d^U$ in Equation 35 will be called the jointly shrinked uplift estimator*.*

5 Shrinked Uplift Regression. Experimental Evaluation

In this section we present an experimental evaluation of uplift shrinkage estimators on simulated data. The simulation protocol is similar to that used in Section 3, except that we now have two training sets: treatment and control. They both have 30 data records and 20 variables (excluding intercept). The control response coefficients β_C are the same as the regression coefficients in Section 3 with the control group intercept $\beta_0^C = 0.1$. The coefficient vector β^U of the linear conditional effect (the quantity of interest) has all coefficients equal to 0.1, except for the intercept, β_0^U, for which four different values of 0.01, 0.1, 1, and 10 were used.

For each value of the intercept, the simulation has been repeated 100 000 times. The results are shown in Table 2. **Double** denotes the double estimator given in Equation 21 which does not use shrinkage.

First, it should be noted that several shrinkage estimators allowed for achieving dramatic reduction in MSE over the original double regression model. MSE was in some cases reduced more than ten times.

It can also be seen that for small values of the uplift intercept term β_0^U, using a single shrinkage factor for all coefficients yielded the best model. The proposed separately shrinked **Intercept** method was slightly worse.

	Separately shrinked estimator (Definition 1)			Jointly shrinked estimator (Definition 2)		
Estimator	test MSE	s. dev.	s. err.	test MSE	s. dev.	s. err.
$\beta_0^U = 0.01$						
Double	4.4384	2.2681	0.0072			
Intercept	0.4460	0.5087	0.0016	1.709	1.5438	0.0049
Single	**0.3647**	0.4712	0.0015	1.6875	1.5412	0.0049
Full	3.4116	1.6822	0.0053	2.1944	1.6603	0.0053
$\beta_0^U = 0.1$						
Double	4.4393	2.2478	0.0071			
Intercept	0.4479	0.5016	0.0016	1.7102	1.5644	0.0049
Single	**0.3694**	0.4642	0.0015	1.6904	1.5632	0.0049
Full	3.4161	1.6633	0.0053	2.1962	1.6808	0.0053
$\beta_0^U = 1$						
Double	4.4326	2.2539	0.0071			
Intercept	**0.499**	0.5175	0.0016	1.8654	1.5783	0.005
Single	1.1209	0.5666	0.0018	2.1796	1.5641	0.0049
Full	3.5075	1.6981	0.0054	2.3714	1.701	0.0054
$\beta_0^U = 10$						
Double	4.4292	2.2382	0.0071			
Intercept	**0.4946**	0.5127	0.0016	1.825	1.5843	0.005
Single	7.4138	2.9874	0.0094	5.0334	2.5958	0.0082
Full	3.4993	1.6825	0.0053	2.3296	1.7155	0.0054

Table 2: Mean Squared Error of various shrinkage estimators for uplift regression for different values of the true intercept term β_0^U.

The picture changed for larger values of the intercept, where the proposed method was a clear winner. The **Single** strategy actually achieved the worst result for $\beta_0^U = 10$.

Also, the jointly shrinked uplift estimators performed significantly worse than separately shrinked estimators. An exception was the **Full** case, which, however, was still not competitive against the **Intercept** method.

Figure 2 compares the ratio of the MSE's of the proposed separately shrinked **Intercept** method and the simple double model for $p = 20$ and $p = 100$ variables and growing number of data records. The value of $\beta_0^U = 10$ was used. Overall, it can be seen that for uplift regression potential gains

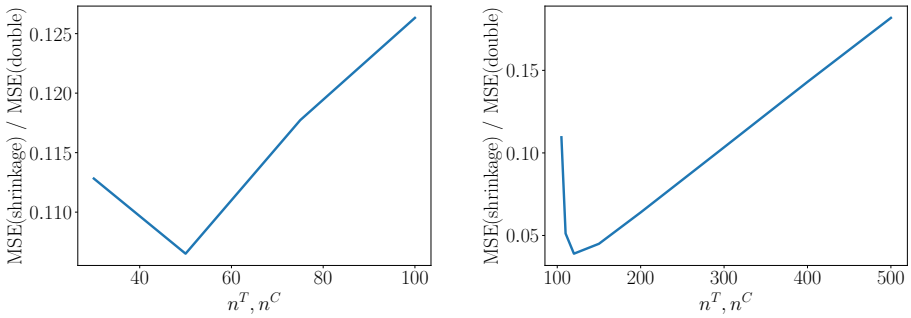

Figure 2: Ratio of mean squared errors of the separately shrinked **Intercept** estimator and the double uplift estimator for $p = 20$ variables (left) and $p = 100$ variables (right) for growing number of training records (n^T, n^C)

from using shrinkage estimators can be much larger than for classical linear regression. Moreover, the gains remain very high over a broad range of parameter values.

6 Conclusions

The paper investigated various shrinkage estimators for ordinary linear regression and for uplift regression whose aim is the estimation of causal effect of some action. The novelty of the proposed estimators lies in how shrinkage was applied to the intercept term: a topic ignored in the current literature. We have demonstrated the benefits of using a separate shrinkage factor for the intercept term. Our proposed shrinkage estimators achieved consistent improvements in predictive mean squared error for both ordinary and uplift regression, with significant gains in the latter case. A possible topic of future research is extending the results to other types of models, such as those applicable to survival data used frequently in medicine [6].

References

[1] Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016. doi:10.1073/pnas.1510489113.

[2] Christian Heumann, Thomas Nittner, Sandro Scheid, C.Radhakrishna Rao, and Helge Toutenburg. *Linear Models: Least Squares and Alternatives*. Springer New York, 2013. doi:10.1007/978-1-4899-0024-1.

[3] Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986. doi:10.2307/2289064.

[4] Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1994. doi:10.1017/CBO9780511840371.

[5] Willard James and Charles Stein. Estimation with quadratic loss. In Jerzy Neyman, editor, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 361–379, 1961.

[6] Szymon Jaroszewicz and Piotr Rzepakowski. Uplift modeling with survival data. In *ACM SIGKDD Workshop on Health Informatics (HI-KDD'14)*, 2014.

[7] Akio Namba and Kazuhiro Ohtani. MSE performance of the weighted average estimators consisting of shrinkage estimators. *Communications in Statistics - Theory and Methods*, 47(5):1204–1214, 2018. doi:10.1080/03610926.2017.1316860.

[8] Nicholas J. Radcliffe and Patrick D. Surry. Real-world uplift modelling with significance-based uplift trees. Portrait Technical Report TR-2011-1, Stochastic Solutions, 2011.

[9] Krzysztof Rudaś and Szymon Jaroszewicz. Linear regression for uplift modeling. *Data Mining and Knowledge Discovery*, 32(5):1275–1305, Sep 2018.

[10] Krzysztof Rudaś and Szymon Jaroszewicz. Shrinkage estimators for uplift regression. In Ulf Brefeld, Elisa Fromont, Andreas Hotho, Arno Knobbe, Marloes Maathuis, and Céline Robardet, editors, *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD'19)*, pages 607–623. Springer-Verlag, 2019. doi:10.1007/978-3-030-46150-8_36.

[11] Piotr Rzepakowski and Szymon Jaroszewicz. Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems*, 32(2):303–327, 2012. doi:10.1007/s10115-011-0434-0.