Robust VIF Regression with Application to Variable Selection in Large Datasets *

Revised - May 30, 2012

Abstract

The sophisticated and automated means of data collection used by an increasing number of institutions and companies leads to extremely large datasets. Subset selection in regression is essential when a huge number of covariates can potentially explain a response variable of interest. The recent statistical literature has seen an emergence of new selection methods that provide some type of compromise between implementation (computational speed) and statistical optimality (e.g. prediction error minimization). Global methods such as Mallows' C_p have been supplanted by sequential methods such as stepwise regression. More recently, streamwise regression, faster than the former, has emerged. A recently proposed streamwise regression approach based on the variance inflation factor (VIF) is promising but its least-squares based implementation makes it susceptible to the outliers inevitable in such large datasets. This lack of robustness can lead to poor and suboptimal feature selection. In our case, we seek to predict an individual's educational attainment using economic and demographic variables. We show how classical VIF performs this task poorly and a robust procedure is necessary for policy makers. This article proposes a robust VIF regression, based on fast robust estimators, that inherits all the good properties of classical VIF in the absence of outliers, but also continues to perform well in their presence where the classical approach fails.

Keywords: Variable selection; Linear regression; Multicollinearity, M-estimator.

*The first author acknowledges the support of the Natural Sciences The authors thank the Editor and three referees for comments that improved the presentation.

1 Introduction

Datasets with millions of observations and a huge number of variables are now quite common, especially in business- and finance-related fields, as well as computer sciences, health sciences, etc. An important challenge is to provide statistical tools and algorithms that can be used with such datasets. In particular, for regression models, a first data analysis requires that the number of potential explanatory variables be reduced to a reasonable and tractable amount. Consider p potential explanatory variables $[1\ x_1 \dots x_p]^T = \boldsymbol{x}$ and a response variable y observed on n subjects. The classical normal linear model supposes $y|\boldsymbol{x} \sim N(\boldsymbol{x}^T\boldsymbol{\beta}; \sigma^2)$ with slope parameters $\boldsymbol{\beta} = [\beta_0, \beta_1, \dots, \beta_p]^T$. The aim is to find a subset of explanatory variables that satisfies a given criterion and such that the regression model holds.

The selection criteria are numerous and can be based on prediction, fit, etc. The available selection procedures can be broadly classified into three classes according to their general strategy and, as a result, their computational speed. A first class considers all the possible combinations of covariates as potential models, evaluates each according to a fixed criterion, and chooses the model which best suits the selected criterion. A second class is formed of sequential selection procedures in which a covariate at a time is entered in (or removed from) the model, based on a criterion that can change from one step to the next and that is computed for all potential variables to enter (or to exit) until another criterion is reached. Finally, the third class of selection procedures is also sequential in nature, but each covariate is only considered once as a potential covariate. For the first class, we find criteria such as the AIC (Akaike 1973), BIC (Schwarz 1978), Mallows' C_p (Mallows 1973), cross-validation, etc (see also Efron 2004). These methods are not adapted to large datasets since the number of potential models becomes too large and the computations are no longer feasible. For the second class, we find for example the classical stepwise regression which can

be considered as a simple algorithm to compute the estimator of regression coefficients $\boldsymbol{\beta}$ that minimizes an l_q penalized sum of squared errors $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda_q \|\boldsymbol{\beta}\|_{l_q}$, with $q = 0$ and $\boldsymbol{X} = [\mathbf{1}\ \boldsymbol{x}_j]_{j=1,\ldots,p}$ and $\mathbf{1}$ a vector of ones, i.e. $\|\boldsymbol{\beta}\|_{l_0} = \sum_{j=1}^{p} \eta(\beta_j \neq 0)$ (see Lin, Foster, and Ungar 2011), with $\eta(\beta_j \neq 0) = 1$ if $\beta_j \neq 0$ and 0 otherwise. Fast algorithms for stepwise regressions are available, e.g. Foster and Stine (2004). Procedures for the l_1 problem are also available, e.g. Lasso/LARS (Efron, Hastie, Johnstone, and Tibshirani 2004), the Dantzig Selector (Candes and Tao 2007), or coordinate descent (Friedman, Hastie, and Tibshirani 2010). But these algorithms may also become very slow for large datasets, not only because all remaining variables are evaluated at each stage, but also because the penalty λ_q needs to be computed, and often via cross-validation. The last class is a variation of stepwise regression in which covariates are tested sequentially but only *once* for addition to the model. An example is the *streamwise regression* of Zhou, Foster, Stine, and Ungar (2006) which uses the α-investing rule (Foster and Stine 2008), is very fast, and guards against overfitting. An improved streamwise regression approach was recently proposed in Lin, Foster, and Ungar (2011) where a very fast to compute test statistic based on the variance inflation factor (VIF) of the candidate variable given the currently selected model is proposed. The approach takes into account possible multicollinearity, seeking to find the best predictive model, even if it is not the most parsimonious. Comparisons in Lin, Foster, and Ungar (2011) establish that the method performs well and is the fastest available.

Our concern in this paper is to provide model selection tools for the regression model that are robust to small model deviations. As argued in Dupuis and Victoria-Feser (2011) (see also Ronchetti and Staudte 1994), spurious model deviations such as outliers can lead to a completely different, and suboptimal, selected model when a non robust criterion, like Mallows' C_p or the VIF, is used. This happens because under slight data contamination, the estimated model parameters, using for example

the least squares estimator (LS), and consequently the model choice criterion can be seriously biased. The consequence is that when the estimated criteria are compared to an absolute level (like a quantile of the χ^2 distribution), the decisions are taken at the wrong level. For the first class of selection procedures, robust criteria have been proposed such as the robust AIC of Ronchetti (1982), the robust BIC of Machado (1993), the robust Mallows' C_p of Ronchetti and Staudte (1994) and a robust criterion based on cross-validation (CV) in Ronchetti, Field, and Blanchard (1997). Since standard robust estimators are impossible to compute when the number of covariates is too large, Dupuis and Victoria-Feser (2011) proposed the use of a forward search procedure together with adjusted robust estimators when there is a large number of potential covariates. Their selection procedure, called Fast Robust Forward Selection (FRFS), falls in the second class of selection procedures. FRFS outperforms classical approaches such as Lasso/LARS when data contamination is present and outperforms, in all studied instances, a robust version of the LARS algorithm proposed by Khan, Van Aelst, and Zamar (2007).

However, although FRFS is indeed very fast and robust, it too can become quite slow when the number of potential covariates is very large as *all* covariates are reconsidered after one is selected for entry in the model. It is therefore important to have a robust selection procedure in the streamwise regression class so that very large datasets can be analyzed in a robust fashion. In this paper, we develop a robust VIF approach that is fast, very efficient, and clearly outperforms non-robust VIF in the presence of outliers.

The remainder of the paper is organized as follows. In Section 2, we review the classical VIF approach and present our robust VIF approach. A simulation study in Section 3 shows the good performance of the new approach. In Section 4, we analyze educational attainment data and show how policy makers are better served by robust VIF regression than by classical VIF or Lasso. In Section 5, we present a shorter

analysis of a large crime dataset that highlights more problems with classical VIF for real data. Section 6 contains a few closing remarks.

2 Robust VIF Regression

2.1 The classical approach

Lin, Foster, and Ungar (2011) propose a procedure that allows one to sweep through all available covariates and to enter those that can reduce a statistically sufficient part of the variance in the predictive model. Let $\boldsymbol{X}_S$ be the design matrix that includes the selected variables at a given stage, and $\tilde{\boldsymbol{X}}_S = [\boldsymbol{X}_S \;\; \boldsymbol{z}_j]$ with $\boldsymbol{z}_j$ the new potential covariate to be considered for inclusion. Without loss of generality, we suppose all variables have been standardized. Consider the following two models

$$\boldsymbol{y} = \boldsymbol{X}_S\boldsymbol{\beta}_S + \boldsymbol{z}_j\beta_j + \boldsymbol{\varepsilon}_{step}, \qquad \boldsymbol{\varepsilon}_{step} \sim N(\boldsymbol{0}, \sigma^2_{step}\boldsymbol{I}) \tag{1}$$

$$\boldsymbol{r}_S = \boldsymbol{z}_j\gamma_j + \boldsymbol{\varepsilon}_{stage}, \qquad \boldsymbol{\varepsilon}_{stage} \sim N(\boldsymbol{0}, \sigma^2_{stage}\boldsymbol{I}) \tag{2}$$

where $\boldsymbol{r}_S = (\boldsymbol{I} - \boldsymbol{X}_S(\boldsymbol{X}_S^T\boldsymbol{X}_S)^{-1}\boldsymbol{X}_S^T)\boldsymbol{y}$ are the residuals of the projection of $\boldsymbol{y}$ on $\boldsymbol{X}_S$. All known estimators of the parameters β_j, σ^2_{step} and $\gamma_j, \sigma^2_{stage}$ will provide different estimates when the covariates present some degree of multicollinearity, and consequently, significance tests based on estimates of β_j or γ_j do not necessarily lead to the same conclusions. While in stepwise regression the significance of β_j in model (1) is at the core of the selection procedure, in streamwise regression, one estimates more conveniently γ_j. Lin, Foster, and Ungar (2011) show that, when LS are used to estimate, $\hat{\gamma}_j = \rho\hat{\beta}_j$ where $\rho = \boldsymbol{z}_j^T(\boldsymbol{I} - \boldsymbol{X}_S(\boldsymbol{X}_S^T\boldsymbol{X}_S)^{-1}\boldsymbol{X}_S^T)\boldsymbol{z}_j$. They then compare $T_\gamma = \hat{\gamma}_j/(\rho^{1/2}\sigma)$, with suitable estimates for ρ and σ, to the standard normal distribution to decide whether or not $\boldsymbol{z}_j$ should be added to the current model. The

procedure is called VIF regression since Marquardt (1970) called $1/\rho$ the VIF for $\boldsymbol{z}_j$.

2.2 A robust weighted slope estimator

Since the test statistic T_γ is based on : (1) the LS estimator $\hat{\gamma}_j$, (2) ρ, in-turn based on the design matrix $\boldsymbol{X}_S$ and $\boldsymbol{z}_j$, and (3) the classical estimator of σ, it is obviously very sensitive to outliers, a form of model deviation. An extreme response or a very badly placed design point can have a drastic effect on T_γ. The latter is then compared to the null distribution : the correct asymptotic distribution under the hypothesis that the regression model holds. With model deviations, the null distribution is not valid, and hence selection decisions (to add the covariate or not) are taken rather arbitrarily. We propose here to limit the influence of extreme observations by considering weighted LS estimators of the form

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^{wT}\boldsymbol{X}^{w})^{-1}\boldsymbol{X}^{wT}\boldsymbol{y}^{w} \tag{3}$$

with $\boldsymbol{X}^w = diag(\sqrt{w_i^0})\boldsymbol{X}$ and $\boldsymbol{y}^w = diag(\sqrt{w_i^0})\boldsymbol{y}$. The weights w_i^0 depend on the data and are such that extreme observations in the response and/or in the design have a nil or limited effect on the value of $\widehat{\boldsymbol{\beta}}$. Dupuis and Victoria-Feser (2011) propose Tukey's redescending biweight weights

$$w_i(r_i; c) = \begin{cases} \left(\left(\frac{r_i}{c}\right)^2 - 1\right)^2 & \text{if } |r_i| \leq c, \\ 0 & \text{if } |r_i| > c, \end{cases} \tag{4}$$

where $r_i = (y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})/\sigma$ are standardized residuals that are computed in practice for chosen estimators of $\boldsymbol{\beta}$ and σ (see below). The constant c controls the efficiency and the robustness of the estimator. Indeed, the most efficient estimator is the LS estimator, i.e. (3) with all weights equal to one (i.e. $c \to \infty$), but it is very sensitive to

(small) model deviations, while a less efficient but more robust estimator is obtained by downweighting observations that have a large influence on the estimator, i.e. by setting $c < \infty$ in (4). The value $c = 4.685$ corresponds to an efficiency level of 95% for the robust estimator compared to the LS estimator at the normal model and is the value used throughout the paper.

We follow Dupuis and Victoria-Feser (2011) and use for the weights $w_i^0 = w_i(r_i^0; c)$ in (3), where the residuals $r_i^0 = (y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}^0)/\hat{\sigma}^0$ and $\hat{\sigma}^0 = 1.483\text{med}|\tilde{r}_i^0 - \text{med}(\tilde{r}_i^0)|$, the median absolute deviation (MAD) of the residuals $\tilde{r}_i^0 = y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}^0$. The slope estimates are $\widehat{\boldsymbol{\beta}}^0 = \left[(\boldsymbol{X}_0^w)^T \boldsymbol{X}_0^w\right]^{-1} (\boldsymbol{X}_0^{w2})^T \boldsymbol{y}$, with $\boldsymbol{X}_0^w = [\ 1 \quad \sqrt{w_{i1}} x_{i1} \quad \ldots \quad \sqrt{w_{ip}} x_{ip}\ ]$ and $\boldsymbol{X}_0^{w2} = [\ 1 \quad w_{i1} x_{i1} \quad \ldots \quad w_{ip} x_{ip}\ ]$, $i = 1, \ldots, n$, with weights w_{ij}, for all $j = 1, \ldots, p$, computed using (4) at the residuals $r_{ij} = (y_i - \widehat{\beta}_{0j} - x_{ij}\widehat{\beta}_j)/\widehat{\sigma}_j$, with $\widehat{\sigma}_j = \text{MAD}(y_i - \widehat{\beta}_{0j} - x_{ij}\widehat{\beta}_j)$. The slope estimators $\widehat{\beta}_1, \ldots, \widehat{\beta}_p$ and the intercept estimators $\widehat{\beta}_{01}, \ldots, \widehat{\beta}_{0p}$ are computed on the p marginal models $y = \beta_{01} + x_1\beta_1 + \varepsilon_1, \ldots, y = \beta_{0p} + x_p\beta_p + \varepsilon_p$ using a robust weighted estimator defined implicitly through

$$\sum_{i=1}^{n} w_i(r_i; c) r_i \boldsymbol{x}_i = 0\ . \tag{5}$$

Here we consider Huber's weights given for the regression model by

$$w_i(r_i; c) = min\left\{1; \frac{c}{|r_i|}\right\} \tag{6}$$

with $c = 1.345$. Estimators in (5) belong to the class of M-estimators (Huber 1964, 1967). With (6) in (5), the marginal intercepts and slope estimators are simpler (and faster) to compute than the ones based on Tukey's biweight weights as originally proposed in Dupuis and Victoria-Feser (2011). For the scale in the weights in (5), we propose to use the MAD of the residuals.

The estimator in (3) is a one-step estimator that is actually biased when there is

multicollinearity in the covariates. Dupuis and Victoria-Feser (2011) show that the bias can be made smaller and even nil if $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^1$ is iterated further to get say $\widehat{\boldsymbol{\beta}}^k$ computed at the updated weights $w_i^1, \ldots, w_i^{k-1}$ based on the residuals $r_i^{(1)} = (y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}^{(1)})/\widehat{\sigma}^{(1)}, \ldots, r_i^{(k-1)} = (y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}^{(k-1)})/\widehat{\sigma}^{(k-1)}$. In the simulation study in Section 3 however, we find that the bias is very small even with relatively large multicollinearity, so that in practice there is often no need to proceed with this iterative correction.

Finally, $\widehat{\boldsymbol{\beta}}^0$ is a coordinate-wise robust estimator and Alqallaf, Van Aelst, Yohai, and Zamar (2009) show, through the computation of a generalized version of the influence function (Hampel 1968, Hampel 1974) and different contamination schemes in the multivariate normal (MVN) setting, that coordinate-wise robust estimators can be less sensitive to extreme observations when they occur independently at the univariate level.

2.3 Robust VIF selection criterion

Let $\boldsymbol{X}_S^w = diag(\sqrt{w_{iS}^0})\boldsymbol{X}_S$ be the weighted design matrix at stage S with say q columns (hence $q-1$ covariates) and $\boldsymbol{z}_j^w = diag(\sqrt{w_{ij}})\boldsymbol{z}_j$ the new candidate covariate that is evaluated at the current stage $S+1$. One could use the weights w_{iS}^0 for $\boldsymbol{z}_j^w$ instead of the weights w_{ij} computed at the marginal models with only $\boldsymbol{z}_j$ as covariate, but this would require more computational time. The simulation results in Section 3 show that one gets very satisfactory results with w_{ij}. Let also $\tilde{\boldsymbol{X}}_S^w = [\boldsymbol{X}_S^w | \boldsymbol{z}_j^w]$ and define $\widehat{\beta}_j^w$ as the last element of the vector $[\tilde{\boldsymbol{X}}_S^{wT} \tilde{\boldsymbol{X}}_S^w]^{-1} \tilde{\boldsymbol{X}}_S^{wT} \boldsymbol{y}^w$ with $\boldsymbol{y}^w = diag(\sqrt{w_{iS}^0})\boldsymbol{y}$. $\widehat{\beta}_j^w$ is actually a robust estimator of β_j in (1). Let $\boldsymbol{H}_S^w =$

$\boldsymbol{X}_S^w(\boldsymbol{X}_S^{wT}\boldsymbol{X}_S^w)^{-1}\boldsymbol{X}_S^{wT}$ and $\widehat{\boldsymbol{\beta}}_S = (\boldsymbol{X}_S^{wT}\boldsymbol{X}_S^w)^{-1}\boldsymbol{X}_S^{wT}\boldsymbol{y}^w$, then

$$\begin{aligned}
\widehat{\beta}_j^w &= -(\boldsymbol{z}_j^{wT}\boldsymbol{z}_j^w - \boldsymbol{z}_j^{wT}\boldsymbol{H}_S^w\boldsymbol{z}_j^w)^{-1}\boldsymbol{z}_j^{wT}\boldsymbol{X}_S^w(\boldsymbol{X}_S^{wT}\boldsymbol{X}_S^w)^{-1}\boldsymbol{X}_S^{wT}\boldsymbol{y}^w \\
&\quad +(\boldsymbol{z}_j^{wT}\boldsymbol{z}_j^w - \boldsymbol{z}_j^{wT}\boldsymbol{H}_S^w\boldsymbol{z}_j^w)^{-1}\boldsymbol{z}_j^{wT}\boldsymbol{y}^w \\
&= (\boldsymbol{z}_j^{wT}\boldsymbol{z}_j^w - \boldsymbol{z}_j^{wT}\boldsymbol{H}_S^w\boldsymbol{z}_j^w)^{-1}\boldsymbol{z}_j^{wT}(\boldsymbol{y}^w - \boldsymbol{X}_S^w\widehat{\boldsymbol{\beta}}_S) \\
&= (\boldsymbol{z}_j^{wT}\boldsymbol{z}_j^w - \boldsymbol{z}_j^{wT}\boldsymbol{H}_S^w\boldsymbol{z}_j^w)^{-1}\boldsymbol{z}_j^{wT}\boldsymbol{r}_S^w \\
&= (\boldsymbol{z}_j^{wT}\boldsymbol{z}_j^w - \boldsymbol{z}_j^{wT}\boldsymbol{H}_S^w\boldsymbol{z}_j^w)^{-1}(\boldsymbol{z}_j^{wT}\boldsymbol{z}_j^w)(\boldsymbol{z}_j^{wT}\boldsymbol{z}_j^w)^{-1}\boldsymbol{z}_j^{wT}\boldsymbol{r}_S^w
\end{aligned}$$

where $\boldsymbol{r}_S^w$ are the residuals of the weighted fit of $\boldsymbol{y}^w$ on $\boldsymbol{X}_S^w$. Let

$$\rho^w = (\boldsymbol{z}_j^{wT}\boldsymbol{z}_j^w)^{-1}(\boldsymbol{z}_j^{wT}\boldsymbol{z}_j^w - \boldsymbol{z}_j^{wT}\boldsymbol{H}_S^w\boldsymbol{z}_j^w)$$

then

$$\widehat{\beta}_j^w = (\rho^w)^{-1}\widehat{\gamma}_j^w$$

with $\widehat{\gamma}_j^w = (\boldsymbol{z}_j^{wT}\boldsymbol{z}_j^w)^{-1}\boldsymbol{z}_j^{wT}\boldsymbol{r}_S^w$, i.e. the weighted estimator of the fit of $\boldsymbol{z}_j^w$ on the weighted residuals $\boldsymbol{r}_S^w$, i.e. model (2). Note however that $\widehat{\beta}_j^w$ is not equal to the last element of $\widehat{\boldsymbol{\beta}}_{S+1}^1$ unless the weights w_{iS}^0 are used for $\boldsymbol{z}_j^w$. Note also that we can write

$$\rho^w = 1 - R_{jS}^{w2}$$

with

$$R_{jS}^{w2} = \boldsymbol{z}_j^{wT}\boldsymbol{H}_S^w\boldsymbol{z}_j^w(\boldsymbol{z}_j^{wT}\boldsymbol{z}_j^w)^{-1} \tag{7}$$

a robust estimate of the coefficient of determination R^2. Renaud and Victoria-Feser (2010) propose a robust R^2 based on weighted responses and covariates and (7) is equivalent to their proposal (with $a = 1$, see their Theorem 1) but with other sets of weights. Moreover, ρ^w is the partial variance of $\boldsymbol{z}_j^w$ given $\boldsymbol{X}_S^w$ (see Dupuis and

Victoria-Feser 2011).

Lin, Foster, and Ungar (2011) note that using all the data to compute ρ (in the classical setting) is quite computationally expensive and they propose a subsampling approach. For the same reason, we also propose to actually estimate ρ^w by computing (7) on a randomly chosen subset of size $m = 200$.

To derive the t-statistic based on $\widehat{\gamma}_j^w$, we follow Lin, Foster, and Ungar (2011) who base their comparison on the expected value of the estimated variance of respectively $\widehat{\beta}_j^w$ and $\widehat{\gamma}_j^w$. Let $\widehat{\sigma}^2_{step}$ and $\widehat{\sigma}^2_{stage}$ be respectively robust residual variance estimates for models (1) and (2). Let also $\boldsymbol{A}_{(i)(j)}$ denote the element (i,j) of matrix $\boldsymbol{A}$. For $\widehat{\beta}_j^w$, supposing that $w_{ij}/w_i^0 \approx 1$, we can use

$$\begin{aligned}
\widehat{\mathrm{Var}(\hat{\beta}_j^w)} &\approx \hat{\sigma}^2_{step}\left[\tilde{\boldsymbol{X}}_S^{wT}\tilde{\boldsymbol{X}}_S^w\right]^{-1}_{(q+1)(q+1)} e_c^{-1} \\
&= \hat{\sigma}^2_{step}(\boldsymbol{z}_j^{wT}\boldsymbol{z}_j^w - \boldsymbol{z}_j^{wT}\boldsymbol{H}_S^w\boldsymbol{z}_j^w)^{-1}e_c^{-1} \\
&= \hat{\sigma}^2_{step}(\rho^w)^{-1}(\boldsymbol{z}_j^{wT}\boldsymbol{z}_j^w)^{-1}e_c^{-1} \\
&= \frac{\hat{\sigma}^2_{step}}{n}(\rho^w)^{-1}\left(\frac{1}{n}\sum_i (z_{ij}^w)^2\right)^{-1} e_c^{-1}
\end{aligned}$$

with

$$e_c = \left[\int_{-c}^{c}\left(5\left(\frac{r}{c}\right)^4 - 6\left(\frac{r}{c}\right)^2 + 1\right)d\Phi(r)\right]^2 \Bigg/ \int_{-c}^{c} r^2\left(\left(\frac{r}{c}\right)^2 - 1\right)^4 d\Phi(r)$$

and Φ the standard normal cumulative distribution (see Heritier et al. 2009, equation (3.20)). For $\hat{\gamma}_j^w$, based on the model with $\boldsymbol{r}_S^w$ as the response and $\boldsymbol{z}_j^w$ as the explanatory variable (without intercept) we have

$$\begin{aligned}
\widehat{\mathrm{Var}(\hat{\gamma}_j^w)} &\approx \hat{\sigma}^2_{stage}(\boldsymbol{z}_j^{wT}\boldsymbol{z}_j^w)^{-1}\tilde{e}_c^{-1} \\
&= \frac{\hat{\sigma}^2_{stage}}{n}\left(\frac{1}{n}\sum_i (z_{ij}^w)^2\right)^{-1}\tilde{e}_c^{-1}
\end{aligned}$$

with $\tilde{e}_c^{-1}$ the efficiency of a robust slope estimator computed using Huber's weights relative to the LS, which is not equal to e_c^{-1}, the efficiency of a robust slope estimator computed using Tukey's weights relative to the LS. We will see below that the computation of the former is not needed. Hence, approximating $\hat{\sigma}^2_{step} \approx \hat{\sigma}^2_{stage} = \hat{\sigma}^2$, we have

$$\widehat{\mathrm{Var}(\hat{\beta}_j^w)} \approx (\rho^w)^{-1}\widehat{\mathrm{Var}(\hat{\gamma}_j^w)}(e_c/\tilde{e}_c)^{-1}.$$

An honest approximate robust test statistic T_w is then given by

$$\frac{\hat{\beta}_j^w}{\sqrt{\mathrm{Var}(\hat{\beta}_j^w)}} \approx \frac{(\rho^w)^{-1}\hat{\gamma}_j^w}{\sqrt{(\rho^w)^{-1}\widehat{\mathrm{Var}(\hat{\gamma}_j^w)}(e_c/\tilde{e}_c)^{-1}}}$$

i.e.

$$T_w = (\rho^w)^{-1/2}\frac{\hat{\gamma}_j^w}{\sqrt{\frac{\hat{\sigma}^2}{n}\left(\frac{1}{n}\sum_i z_{ij}^{w2}\right)^{-1} e_c^{-1}}} \tag{8}$$

with $\hat{\sigma}^2$ a robust mean squared error for the model with $\boldsymbol{r}_S^w$ as response and $\boldsymbol{z}_j^w$ as explanatory variable (i.e. model (2)). We use $\widehat{\sigma} = \mathrm{MAD}(\boldsymbol{r}_S^w - \boldsymbol{z}_j^w(\boldsymbol{z}_j^{wT}\boldsymbol{z}_j^w)^{-1}\boldsymbol{z}_j^{wT}\boldsymbol{r}_S^w)$.

Our *fast robust evaluation procedure* is summarized by the following five steps. Suppose that we are at stage S and a set of $q-1$ covariates has been chosen in the model. We are considering covariate $\boldsymbol{z}_j$ for possible entry. We are working with $c = 4.685$ and have computed e_c and the weights w_{ij} and w_{iS}^0.

1. Obtain the residuals $\boldsymbol{r}_S^w = \boldsymbol{y}^w - \boldsymbol{X}_S^w(\boldsymbol{X}_S^{wT}\boldsymbol{X}_S^w)^{-1}\boldsymbol{X}_S^{wT}\boldsymbol{y}^w$.

2. Set $\boldsymbol{z}_j^w = diag(\sqrt{w_{ij}})\boldsymbol{z}_j$. Compute $\widehat{\gamma}_j^w = (\boldsymbol{z}_j^{wT}\boldsymbol{z}_j^w)^{-1}\boldsymbol{z}_j^{wT}\boldsymbol{r}_S^w$ and $\widehat{\sigma} = \mathrm{MAD}(\boldsymbol{r}_S^w - \boldsymbol{z}_j^w(\boldsymbol{z}_j^{wT}\boldsymbol{z}_j^w)^{-1}\boldsymbol{z}_j^{wT}\boldsymbol{r}_S^w)$.

3. Sample a small subset $\mathcal{I} = \{i_1, \ldots, i_m\} \in \{1, \ldots, n\}$ of the observations and let

${}_{\mathcal{I}}\boldsymbol{x}$ denote the corresponding subsample from the regressor $\boldsymbol{x}$.

4. Let ${}_{\mathcal{I}}\boldsymbol{H}_S^w = {}_{\mathcal{I}}\boldsymbol{X}_S^w({}_{\mathcal{I}}\boldsymbol{X}_S^{wT}{}_{\mathcal{I}}\boldsymbol{X}_S^w)^{-1}{}_{\mathcal{I}}\boldsymbol{X}_S^{wT}$, compute $R_{jS}^{w2} = {}_{\mathcal{I}}\boldsymbol{z}_j^{wT}{}_{\mathcal{I}}\boldsymbol{H}_S^w{}_{\mathcal{I}}\boldsymbol{z}_j^w({}_{\mathcal{I}}\boldsymbol{z}_j^{wT}{}_{\mathcal{I}}\boldsymbol{z}_j^w)^{-1}$, and find $\rho^w = 1 - R_{jS}^{w2}$.

5. Compute the approximate t-ratio $T_w = (\rho^w)^{-1/2}\hat{\gamma}_j^w / \sqrt{\hat{\sigma}^2 \left(\sum_i z_{ij}^{w2}\right)^{-1} e_c^{-1}}$ and compare it to an adapted quantile to decide to add or not $\boldsymbol{z}_j$ to the current set.

A more detailed algorithm in which the decision rule (add or not the new variable) is also specified is given in the Appendix. Note that in Step 5 above, the rejection quantile, or corresponding probability α_j, is adapted at each step j so that α_j increases/decreases if a rejection is made/not made. As explained in Lin, Foster, and Ungar (2011), one can think of α_j as a gambler's wealth and the game is over when $\alpha_j \leq 0$.

2.4 Comparison with the robust t-statistic of FRFS

The t-statistic proposed by Dupuis and Victoria-Feser (2011) (equation (5)) and used to test whether a candidate covariate is entered in the current model, can be written as

$$T^2 = \frac{1}{\sigma^2\rho^w}\frac{n}{\sum w_{ij}} e_c \boldsymbol{y}_j^{wT}\boldsymbol{z}_j^w(\boldsymbol{z}_j^{wT}\boldsymbol{z}_j^w)^{-1}\boldsymbol{z}_j^{wT}(\boldsymbol{I} - \boldsymbol{H}_S^w)\boldsymbol{y}_j^w$$

with $\boldsymbol{y}_j^w = diag(\sqrt{w_{ij}})\boldsymbol{y}$. Supposing that $\boldsymbol{y}_j^w \approx \boldsymbol{y}^w$ and $n/\sum w_{ij} \approx 1$, then

$$\begin{aligned} T^2 &\approx \frac{1}{\sigma^2\rho^w} e_c \boldsymbol{y}^{wT}\boldsymbol{z}_j^w\hat{\gamma}_j^w \\ &= \frac{(\hat{\gamma}_j^w)^2}{\sigma^2\rho^w(\boldsymbol{z}_j^{wT}\boldsymbol{z}_j^w)^{-1}} e_c \frac{1}{\hat{\gamma}_j^w}\boldsymbol{y}^{wT}\boldsymbol{z}_j^w(\boldsymbol{z}_j^{wT}\boldsymbol{z}_j^w)^{-1} \\ &= \frac{(\hat{\gamma}_j^w)^2}{\sigma^2\rho^w(\boldsymbol{z}_j^{wT}\boldsymbol{z}_j^w)^{-1}} e_c \frac{\boldsymbol{y}^{wT}\boldsymbol{z}_j^w}{\boldsymbol{z}_j^{wT}(\boldsymbol{I} - \boldsymbol{H}_S^w)\boldsymbol{y}^w} . \end{aligned} \tag{9}$$

Hence, T_w^2 in (8) and T^2 in (9) differ by a multiplicative factor of

$$\kappa = \frac{\boldsymbol{y}_j^{wT}\boldsymbol{z}_j^w}{\boldsymbol{z}_j^{wT}(\boldsymbol{I} - \boldsymbol{H}_S^w)\boldsymbol{y}_j^w}$$

which is the ratio of the robustly estimated covariance between $\boldsymbol{z}_j$ and $\boldsymbol{y}$, and the robustly estimated partial covariance between $\boldsymbol{z}_j$ and $\boldsymbol{y}$ given $\boldsymbol{X}_S$. One can notice that in the orthogonal case (and standardized covariates), we have $\boldsymbol{z}_j^{wT}\boldsymbol{H}_S^w \approx 0$ so that $\kappa \approx 1$. The value of κ was computed in some of the simulations outlined in the following section. While κ maintained a median value of 1 when aggregating over the 200 simulated datasets at a given setting, its variability changed with the theoretical R^2 and the absence or presence of outliers. For example, the interquartile range went from a value near 0 for $R^2 = 0.20$ and no outliers, to 5 for $R^2 = 0.80$ and 5% outlying responses with high leverage in the $p = 100$ case. There can thus be a considerable difference in the two test statistics.

3 Simulation Study

We carry out a simulation study to assess the effectiveness of the model selection approaches outlined above. First, we create a linear model

$$y = X_1 + X_2 + \ldots + X_k + \sigma\varepsilon \tag{10}$$

where $X_1, X_2, \ldots, X_k$ are multivariate normal (MVN) with $\mathrm{E}(X_i) = 0$, $\mathrm{Var}(X_i) = 1$, and $\mathrm{corr}(X_i, X_j) = \theta$, $i \neq j, i, j = 1, \ldots, k$, and ε an independent standard normal variable. We choose θ to produce a range of theoretical $R^2 = (\mathrm{Var}(y) - \sigma^2)/\mathrm{Var}(y)$ values for (10) and σ to give t values for our target regressors of about 6 under normality as in Ronchetti et al. (1997). The covariates $X_1, \ldots, X_k$ are our k target covariates. Let $e_{k+1}, \ldots, e_p$ be independent standard normal variables and use the

first $2k$ to give the $2k$ covariates

$$\begin{aligned} X_{k+1} &= X_1 + \lambda e_{k+1}, \qquad X_{k+2} = X_1 + \lambda e_{k+2}, \\ X_{k+3} &= X_2 + \lambda e_{k+3}, \qquad X_{k+4} = X_2 + \lambda e_{k+4}, \\ &\vdots \\ X_{3k-1} &= X_k + \lambda e_{3k-1}, \qquad X_{3k} = X_k + \lambda e_{3k}; \end{aligned}$$

and the final $p - 3k$ to give the $p - 3k$ covariates

$$X_i = e_i, \quad i = 3k + 1, \ldots, p.$$

Variables $X_{k+1}, \ldots, X_{3k}$ are noise covariates that are correlated with our target covariates, and variables $X_{3k+1}, \ldots, X_p$ are independent noise covariates. Note that the covariates $X_1, \ldots, X_p$ are then relabeled with a random permutation of $1 : p$ so that the target covariates do not appear in position $1 : k$, but rather in arbitrary positions. This is necessary to test the effectiveness of the streamwise variable selection as covariates considered early on are favored for entry when many covariates are correlated.

We consider samples without and with contamination. Samples with no contamination are generated using $\varepsilon \sim \mathrm{N}(0, 1)$. To allow for 5% outliers, we generate using $\varepsilon \sim 95\%\mathrm{N}(0, 1) + 5\%\mathrm{N}(30, 1)$. These contaminated cases also have high leverage X-values: $X_1, \ldots, X_k \sim$ MVN as before, except $\mathrm{Var}(X_i) = 5$, $i = 1, \ldots, k$. This represents the most difficult contamination scheme: large residuals at high leverage points. We also investigate the less challenging cases of 5% outlying in response only, and 5% high leverage only. We choose $\lambda = 3.18$ so that $\mathrm{corr}(X_1, X_{k+1}) = \mathrm{corr}(X_1, X_{k+2}) = \mathrm{corr}(X_2, X_{k+3}) = \ldots = \mathrm{corr}(X_k, X_{3k}) = 0.3$.

In all simulations, we simulated n independent samples, with or without contamination, to use for variable selection. Then, another n independent samples without contamination were simulated for out-of-sample performance testing. The out-of-

sample performance was evaluated using the mean sum of squared errors (MSE), $\sum_{i=n+1}^{2n}(y_i - \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})^2/n$ where $\widehat{\boldsymbol{\beta}}$ is the estimated coefficient determined by the classical and robust VIF regression selection procedures or FRFS applied to the training set. Because the true predictors are known, we also compute the out-of-sample performance measure using the true $\boldsymbol{\beta}$. Classical VIF selection was carried out using the `VIF` package for R and default argument settings. Robust VIF was also implemented in R. FRFS is also implemented in R as outlined in Dupuis and Victoria-Feser (2011).

It should be noted that when evaluating the performance of a given criterion (here a selection procedure), the evaluation measure should be chosen in accordance with the performance measure (see Gneiting 2011). In our case, although the data are generated from contaminated conditional Gaussian models, the core model is still Gaussian and we wish to find the model that best predicts the conditional mean response. Consequently, a suitable performance measure is the expected squared error. However, when estimating the expected squared error from data, one can resort to the mean (i.e. the MSE) only if the data are purely issued from the postulated (core) Gaussian model. If this is not the case, or if there is no guarantee that this is the case, like for example with real data, then a robust location estimator of the expected squared error such as the median of the absolute errors (i.e. the median absolute prediction error, MAPE) should be chosen. Hence, in the simulations we use the MSE, while with real datasets we use the MAPE to estimate the evaluation measure for the comparison of the variable selection methods.

Simulations results for $n = 1000$, $k = 5$, and $p = 100$ and $p = 1000$, are presented in Table 1 and Figures 1 and 2, respectively. Entries in the top panel of the table give the percentage of runs falling into each category. The category "Correct" means that the correct model was chosen. "Extra" means that a model was chosen for which the true model is a proper subset. "Missing 1" means that the model chosen differed from the true model only in that it was missing one of the target covariates; "Missing

2" and "Missing 3" are defined analogously. The Monte Carlo standard deviation of entries is bounded by 3.5%. We also report the empirical marginal false discovery rate (mFDR) $\widehat{\text{mFDR}} = \widehat{E(V)}/(\widehat{E(V)} + \widehat{E(S)} + \eta)$ where $\widehat{E(S)}$ is the average number of true discoveries, $\widehat{E(V)}$ is the average number of false discoveries and $\eta = 10$ is selected following Lin, Foster, and Ungar (2011). We also report the required computation time. Note the particularly frugal robust approach to VIF regression: the cost of robustness is no more than a doubling of the computation time.

Both VIF algorithms do not perform well in terms of the proportion of correctly selected models and the FRFS-Marginal procedure clearly outperforms in this respect. The execution time of the FRFS-Marginal procedure, the faster of the two FRFS approaches presented in Dupuis and Victoria-Feser (2011), is roughly 25 times longer than that of the robust VIF procedure for these sizes of datasets. Both VIF algorithms do however choose a model for which the true model is a subset when there are no outliers. The classical VIF approach fails miserably in the presence of outliers (outlying response/high leverage), while the robust VIF approach is only slightly affected by the presence of outliers. The classical VIF approach is less affected by the presence of high leverage points only, but the effect is increased under more highly correlated regressors or a higher number of potential regressors. Results (not shown) for response variable only outliers are very similar to outlying response/high leverage outliers. Finally, other simulations (not shown) reveal that for less outlying contamination, the robust approaches always maintain good performance, while the negative impact on classical VIF is proportional to the level of outlyingness.

As the simulated datasets have noise covariates that are correlated with target covariates, the poor performance in terms of %Correct is expected given the streamwise approach of VIF regressions. But as pointed out by Lin, Foster, and Ungar (2011), the goal here is different: good fast out-of-sample prediction, i.e. one sacrifices parsimony for speed. The streamwise approach is fast and the main purpose

Table 1: Model selection results. Simulated data, as described in §3, have $n = 1000$ observations with $p = 100$ and $p = 1000$ potential regressors, including $k = 5$ target regressors. Correlation among target regressors is $\theta = 0.1$ ($R^2 = 0.20$) and $\theta = 0.85$ ($R^2 = 0.80$). Correlation among each target regressor and two other regressors is 0.3 in all cases. Remaining regressors are uncorrelated. Methods are classical (C) and robust (R) VIF regression, and FRFS-Marginal (F). Table entries are % of cases in categories listed in first column. Empirical mFDR appears in the second to last row. Mean execution times (in seconds) appear in the last row. Data were either not contaminated, had 5% high leverage only (hl only) or 5% outliers (outlying response and high leverage). Results are based on 200 simulations for each configuration.

$p = 100$

	$R^2 = 0.20$									$R^2 = 0.80$								
	No contam.			5% hl only			5% outliers			No contam.			5% hl only			5% outliers		
	C	R	F	C	R	F	C	R	F	C	R	F	C	R	F	C	R	F
%Correct	13.5	33	68.5	17.5	24.5	61	0	20	76.5	11.5	18.5	86	6.5	15.5	83.5	0	15	88.5
%Extra	83.5	58	29.5	59.5	40	24.5	0	65.0	20	86.5	76.5	12.5	27.5	62.5	12.5	0	73.5	7.5
%Missing 1	1.5	3.5	1	4.5	10	10	0	6.5	3.5	0.5	1	1.5	7.5	5.5	4	0	3	4
%Missing 2	0	0.5	0	0.5	3	0	1	0.5	0	0	0	0	2.5	0	0	1.5	0	0
%Missing 3	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	11	0	0
%Other	1.5	5	1	18	22.5	4.5	97	8	0	1.5	4	0	56	16.5	0	87.5	8.5	0
%mFDR	11.0	6.3	2.2	9.9	6.4	2.2	6.1	9.3	1.4	16.1	13.2	0.9	16	13.1	1.0	10.7	13.8	0.5
Time	0.63	1.11	25	0.45	0.87	25	0.54	1.09	25	0.69	1.20	25	0.48	0.93	25	0.59	1.20	25

$p = 1000$

	$R^2 = 0.20$									$R^2 = 0.80$								
	No contam.			5% hl only			5% outliers			No contam.			5% hl only			5% outliers		
	C	R	F	C	R	F	C	R	F	C	R	F	C	R	F	C	R	F
%Correct	30	32	66.5	14.5	8	44	0	25	66.5	14.5	16	81	1.5	8.5	69	0	10	72
%Extra	53	27	26.5	15.5	5	15.5	0	26	17.5	77	54	12.5	13.5	29	12	0	43	9.5
%Missing 1	5.5	17	6	25.5	24	30.5	0	20	12	3	7.5	6.5	5	9	18	0	6	18.5
%Missing 2	1	5.5	0	8.5	24	3	0	6	2	0	0	0	7	2	1	1.5	1.5	0
%Missing 3	0	0.5	0	1.5	6.5	0	0	1	0	0	0	0	0.5	0	0	13	0	0
%Other	10.5	18	1	34.5	32.5	7	100	22	2	5.5	22.5	0	72.5	51.5	0	85.5	39.5	0
%mFDR	7.0	4.4	2.4	5.4	3.5	2.0	4.6	6.0	1.4	15.5	13.6	1.0	15.6	12.9	0.9	6.2	13.5	0.7
Time	5.8	10.8	253	4.4	8.9	250	6.1	11.7	238	5.86	10.9	251	4.6	9.3	254	5.47	11.3	243

of an α-investing control is to avoid model overfitting. We assess the latter through out-of-sample performance. Figure 1 shows out-of-sample MSE for the case $p = 100$. Robust VIF is as efficient as classical VIF when there are no outliers (top panel) and clearly outperforms classical VIF when there is 5% contaminated observations (bottom panels). Robust VIF also loses very little with respect to FRFS-Marginal. Note that classical VIF seems to offer some resistance to contamination by high-leverage points only (as was also seen in Table 1), but completely falls apart in the presence of outlying response values, and this whether the outlying responses appear at high-leverage points or not. Much of the same can be seen in Figure 2 where results for the case $p = 1000$ are shown.

4 College Data

Understanding the factors impacting an individual's educational attainment is a preoccupation for many governmental and nongovernmental organizations. For example, a nation's government that recognizes the potential economic benefits of higher education will seek to write public policies to promote it. Private industry that benefits from a well educated labor market will let it affect decision-making, e.g. a company may choose to establish itself where lifelong education is easily accessible to its personnel. Finally, an individual's family who associates personal achievement with higher levels of education may also act accordingly.

Since the first work by Wetterlind (1976) on projecting community college enrollments in Arizona, many researchers have sought to identify the factors impacting educational attainment, see e.g. Pennington, McGinty and Williams (2002), Petrongolo and San Segundo (2002), Kienzl, Alfonso and Melguizo (2007), and Clark (2011) (and references therein) for a list of various studies.

The data analyzed here are in the R package `AER` and are a subset of the data

Figure 1: Out-of-sample mean square errors of the models chosen by classical and robust VIF regression, and FRFS-Marginal. Simulated data, as described in §3, have $n = 1000$ observations with $p = 100$ potential regressors, including $k = 5$ target regressors. Correlation among target regressors is $\theta = 0.1$ ($R^2 = 0.20$) and $\theta = 0.85$ ($R^2 = 0.80$). Correlation among each target regressor and two other regressors is 0.3 in all cases. Remaining regressors are uncorrelated. Results are based on 200 simulations for each configuration.

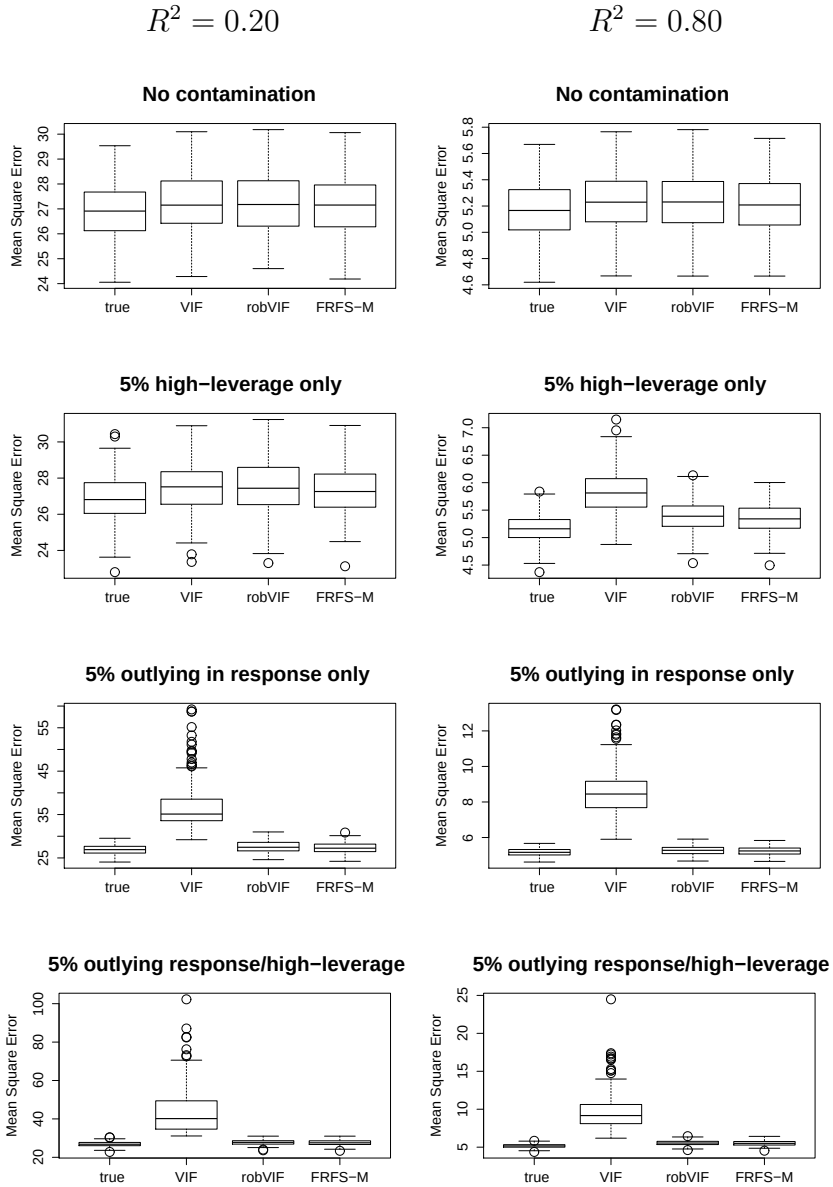
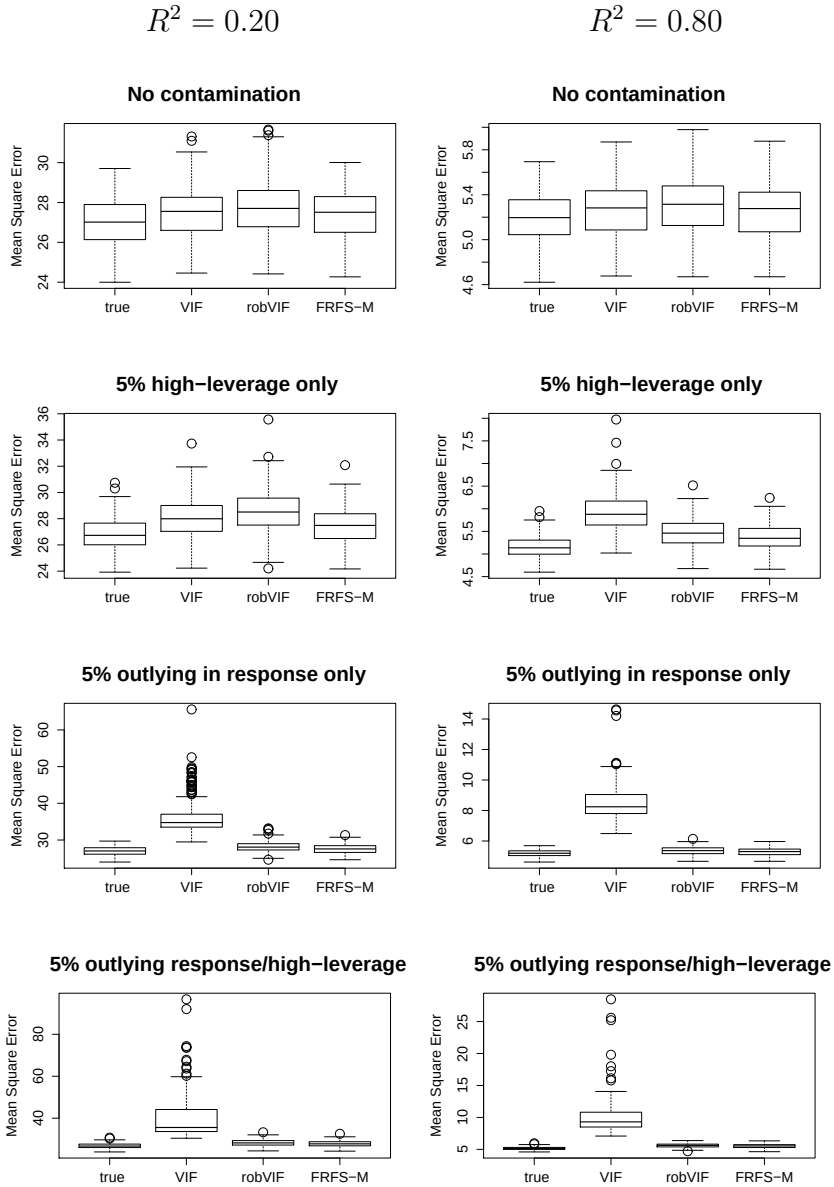


Figure 2: Out-of-sample mean square errors of the models chosen by classical and robust VIF regression, and FRFS-Marginal. Simulated data, as described in §3, have $n = 1000$ observations with $p = 1000$ potential regressors, including $k = 5$ target regressors. Correlation among target regressors is $\theta = 0.1$ ($R^2 = 0.20$) and $\theta = 0.85$ ($R^2 = 0.80$). Correlation among each target regressor and two other regressors is 0.3 in all cases. Remaining regressors are uncorrelated. Results are based on 200 simulations for each configuration.



previously analyzed in Rouse (1995). There are 4739 observations on 14 variables. The variables are listed in Table 2. We seek to predict the number of years of education using 13 economic and demographic variables. There are continuous and binary variables along with one categorical variable with three categories which is converted to two dummy variables. When considering only first-order variables we thus have $n = 4739$ and $p = 14$, when we include second-order interaction terms p rises to 104 (some interaction terms are constant and are removed). We have standardized the variables. Our analysis will show how classical, i.e. non robust, VIF regression can be inadequate for the policy maker by failing to keep important features.

The selected models are compared using the median absolute prediction error (MAPE), as measured by 10-fold CV. That is, we split the data into 10 roughly equal-sized parts. For the kth part, we carry out model selection using the other nine parts of the data and calculate the MAPE of the chosen model when predicting the kth part of the data. We do this for $k = 1, \ldots, 10$ and show boxplots of the 10 estimates of the MAPE. For all methods, the data were split in the same way. For the college data, we randomly generated the folds. Note here that we look at MAPE instead of mean squared prediction error as these real data can contain outliers (as opposed to the simulated data which were clean) and the MAPE is a better choice.

For completeness, we compare the models selected by classical and robust VIF approaches with those of FRFS-Marginal and FRFS-Full where feasible, as well as that of the popular least angle regression (LARS) of Efron, Hastie, Johnstone, and Tibshirani (2004), an extremely efficient algorithm for computing the entire Lasso (Tibshirani 1996) path. We use the R package `lars` to do the computations.

Tables 3 and 4 list the VIF and robust VIF regression selected features, along with estimated slopes, for the $p = 14$ and $p = 104$ scenarios, respectively. For both scenarios, the robust VIF regression approach selects slightly more, and/or slightly different, features. When considering only first-order terms, we see that the classical

Table 2: Original 14 Variables in College Data.

Variable	Description
`gender`	Factor indicating gender.
`ethnicity`	Factor indicating ethnicity (African-American, Hispanic or other).
`score`	Base year composite test score. These are achievement tests given to high school seniors in the sample.
`fcollege`	Factor. Is the father a college graduate?
`mcollege`	Factor. Is the mother a college graduate?
`home`	Factor. Does the family own their home?
`urban`	Factor. Is the school in an urban area?
`unemp`	County unemployment rate in 1980.
`wage`	State hourly wage in manufacturing in 1980.
`distance`	Distance from 4-year college (in 10 miles).
`tuition`	Average state 4-year college tuition (in 1000 USD).
`income`	Factor. Is the family income above USD 25,000 per year?
`region`	Factor indicating region (West or other).
`education`	Number of years of education.

and robust estimates of commonly selected features are almost the same. This serves as a good form of validation for the relative importance of these features. However, the presence of outliers in the data has lead classical VIF regression to completely miss two important features which are identified by robust VIF regression : `unemp` and `wage`. Even LS estimates (not shown) of the robust VIF regression selected model find these two features important with t-values of 3.15 and -2.70, but the classical VIF regression selection procedure could not detect this importance for the reasons outlined in Section 1. FRFS-Marginal and FRFS-Full selected features are identical. The latter features, along with estimated slopes, are also shown in Table 3.

VIF regression also misses the two important features in the $p = 104$ scenario, see Table 4. As both the county unemployment rate and the state hourly wage in manufacturing are directly impacted by economic policy, policy makers must be equipped with the best feature selection tools to have an effective strategy to reach sought after goals: in this case, increasing the level of education among its constituents. These

Table 3: VIF and robust VIF selected variables and estimated slope parameters (t-values) when only considering first-order terms. FRFS-Marginal and FRFS-Full selected variables and estimated slope parameters (t-values) are also shown. Significance : $^{*}0.05$, $^{**}0.01$, $^{***}0.001$.

Variable	VIF $\widehat{\beta}_{LS}$	robVIF $\widehat{\beta}_{rob}$	FRFS-Marg/Full $\widehat{\beta}_{rob}$
`ethnicityafam`	0.130 (5.28)***	0.129 (4.90)***	0.133 (5.16)***
`ethnicityhispanic`	0.142 (5.97)***	0.124 (4.92)***	0.130 (5.19)***
`score`	0.772 (31.3)***	0.820 (31.8)***	0.824 (31.9)***
`fcollegeyes`	0.219 (8.40)***	0.233 (8.51)***	0.232 (8.52)***
`mcollegeyes`	0.131 (5.25)***	0.146 (5.60)***	0.145 (5.55)***
`homeyes`	0.054 (2.39)*	0.057 (2.38)*	0.057 (2.41)*
`urbanyes`	-	0.024 (0.96)	-
`unemp`	-	0.077 (3.00)**	0.075 (2.94)**
`wage`	-	-0.064 (-2.56)*	-0.062 (-2.50)*
`distance`	-0.064 (-2.81)**	-0.083 (-3.22)***	-0.088 (-3.57)***
`incomehigh`	0.163 (6.70)***	0.180 (7.07)***	0.183 (7.20)***
`genderfemale`	-	-	0.066 (2.81)**

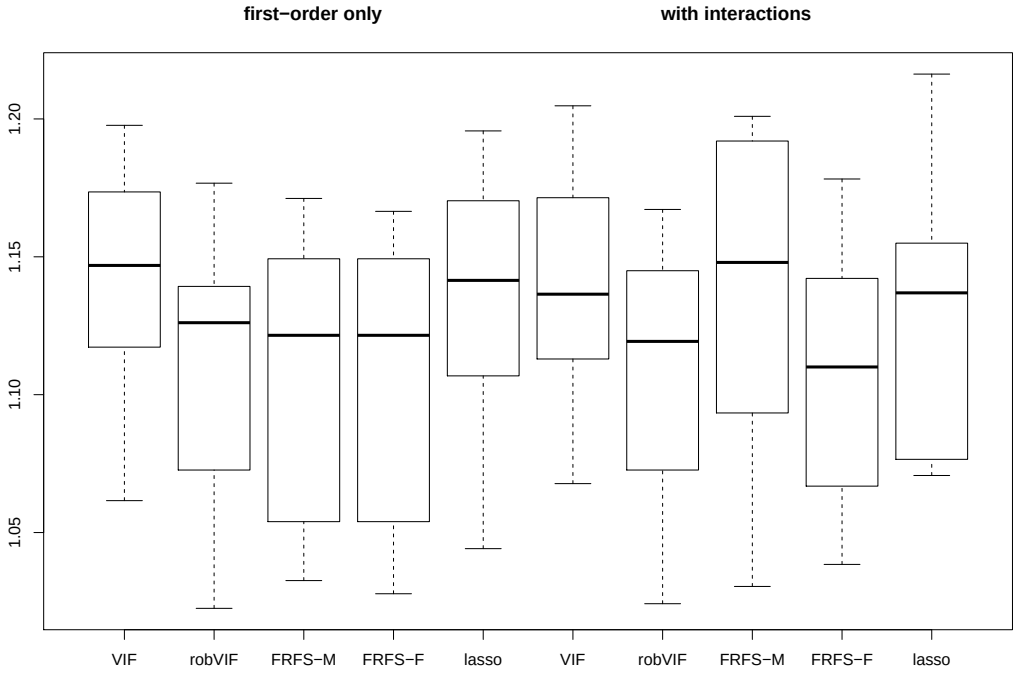
tools, we argue, must include a robust selection procedure as shown effectively by this example. Further evidence is given in Figure 3 where MAPE for VIF, robust VIF, FRFS-Marginal and FRFS-Full and Lasso are shown for both scenarios. Robust VIF outperforms both of its non-robust competitors, and even does better than FRFS-Marginal in the highly collinear case including interactions. It was shown in Dupuis and Victoria-Feser (2011) that FRFS-Marginal could select too few features in the highly collinear case and this motivated the development of FRFS-Full therein.

As the solution for VIF and robust VIF regression can depend on the order of the covariates, we ran each procedure several times with the covariates presented in random order to investigate the stability of the selected models in terms of model size and prediction performance. Table 5 shows the distribution of the size of the selected model over 100 analyses and Table 6 shows how often each variable was selected over these 100 analyses. As expected, there is considerable variability in the size of the model, and this both in the classical and robust approaches. We see however that the

Table 4: VIF and robust VIF selected variables and estimated slope parameters (t-values) when including second-order interactions. Significance : *0.05, **0.01, ***0.001.

Variable	$\widehat{\beta}_{LS}$	$\widehat{\beta}_{rob}$
`ethnicityafam`	0.132 (5.39)***	0.127 (4.83)***
`ethnicityhispanic`	-0.143 (6.02)***	0.122 (4.83)***
`score`	0.772 (31.1)***	0.809 (27.3)***
`fcollegeyes`	0.222 (8.52)***	-0.033 (-0.17)
`mcollegeyes`	0.056 (1.62)	0.045 (0.25)
`homeyes`	0.056 (2.46)*	0.041 (1.61)
`urbanyes`	-	0.028 (1.12)
`unemp`	-	0.059 (2.10)*
`wage`	-	-0.067 (−2.36)*
`distance`	-0.062 (−2.75)**	-0.078 (−3.00)**
`incomehigh`	0.167 (6.87)***	0.040 (0.27)
`genderfemale:score`	0.030 (1.24)	-
`genderfemale:fcollegeyes`	-	0.002 (0.06)
`genderfemale:mcollegeyes`	0.104 (3.07)**	0.132 (3.43)***
`score:incomehigh`	-	0.150 (0.98)
`fcollegeyes:homeyes`	-	0.115 (1.74)
`fcollegeyes:unemp`	-	0.087 (1.24)
`fcollegeyes:wage`	-	0.001 (0.01)
`fcollegeyes:tuition`	-	0.085 (1.44)
`mcollegeyes:wage`	-	0.002 (0.01)

Figure 3: College data: Out-of-sample median absolute prediction errors of the models chosen by classical and robust VIF regression, FRFS-Marginal, FRFS-Full and the Lasso, in 10-fold cross-validation.



dominating features are nearly always present. Note also that `unemp` and `wage` are selected twice as often in the robust approach compared to the classical approach. In terms of prediction performance, we see in Figure 4 that the variability in the latter is considerably less, each of the 10 random analyses shown yielding more or less the same prediction performance despite the differences in terms of selected model size and features.

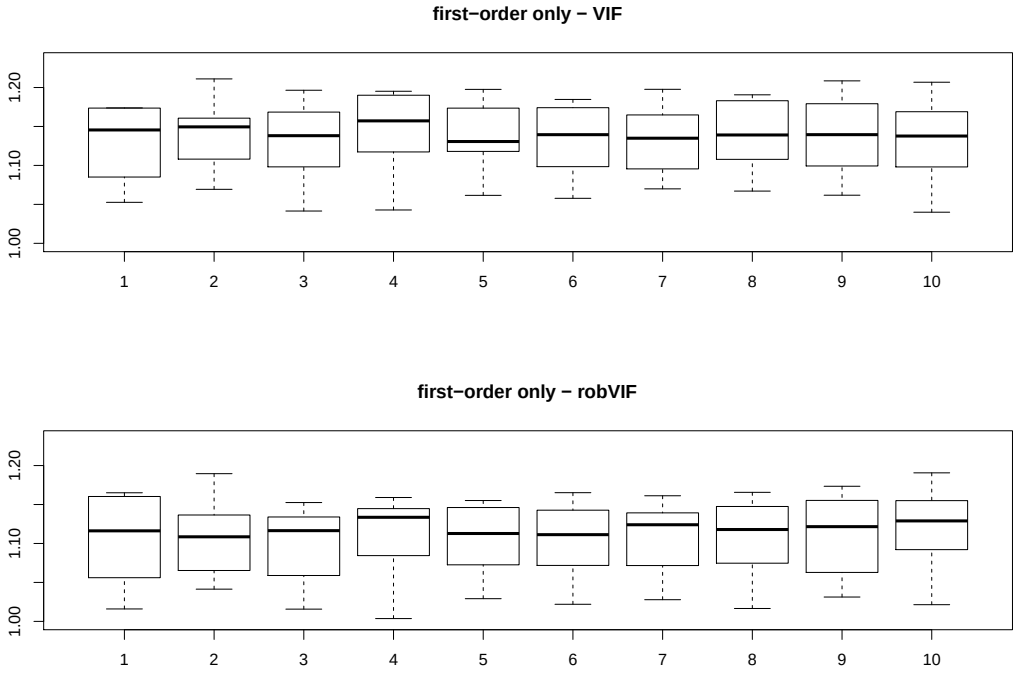
Table 5: Number of variables selected by VIF and robust VIF in 100 analyses of college data, each analysis with covariates presented in a random order.

# selected	7	8	9	10	11	12	13	14
VIF	11	29	24	22	10	3	1	-
robVIF	4	13	8	23	17	12	10	13

Table 6: Number of analyses where variable was selected by VIF and robust VIF in 100 analyses of college data, each analysis with covariates presented in a random order.

Variable	VIF	robVIF
`genderfemale`	43	47
`ethnicityafam`	100	100
`ethnicityhispanic`	67	73
`score`	100	100
`fcollegeyes`	100	99
`mcollegeyes`	100	100
`homeyes`	79	94
`urbanyes`	3	38
`unemp`	24	54
`wage`	31	63
`distance`	100	98
`tuition`	26	56
`incomehigh`	100	98
`regionwest`	31	57

Figure 4: College data: Out-of-sample MAPE of 10 random chosen analyses among 100 analyses reported in Tables 5 and 6 for classical and robust VIF regression. MAPE calculated based on 10-fold cross-validation.

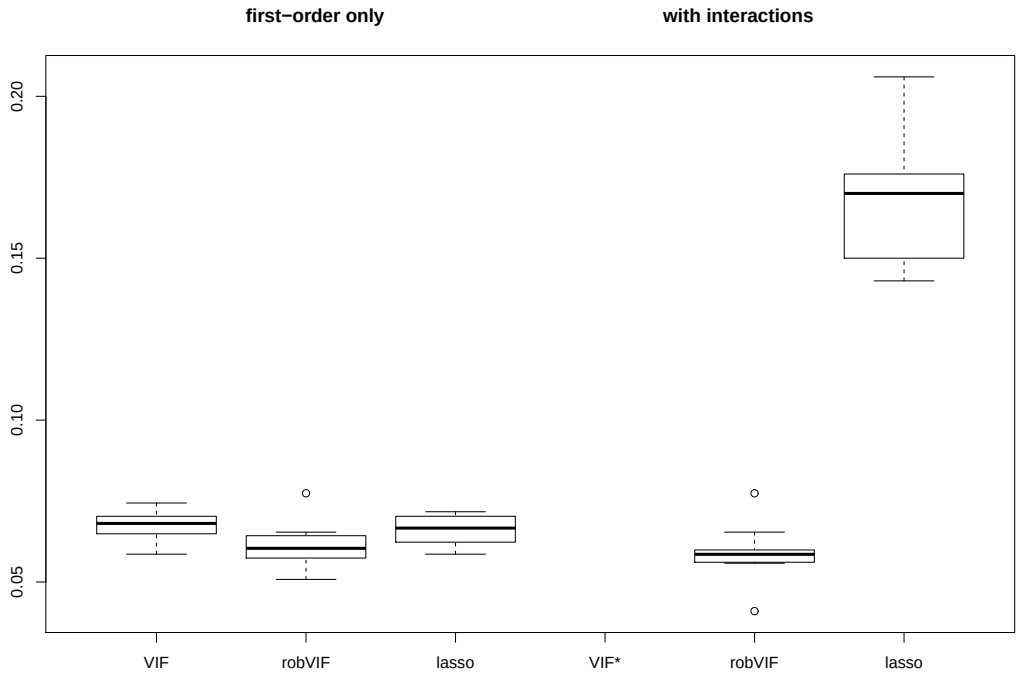


5 Crime Data

In this section, we present a shorter analysis of another dataset to show how the classical approach can even fail to give a usable result. Also, by looking at a considerably larger dataset we can show how robust VIF provides robust prediction where no other robust method is feasible.

We analyze recently made available crime data. These data are from the UCI Machine Learning Repository (Frank and Asuncion 2010) and are available at `http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime`. We seek to predict the per capita violent crimes rate using economic, demographic, community and law enforcement related variables. After removing variables with missing data, we are left with $n = 1994$ observations on $p = 97$ first-order covariates. If we include second-order interactions (removing those that are constant) we have $p = 4753$. In both cases, we standardized the variables. VIF regression selects 33 and 1437 variables, in the respective scenarios, while robust VIF regression selects 20 variables is both cases. Classical VIF experiences problems with the larger dataset, which contains outliers in a highly multicollinear setting, and chooses too many covariates. This shows how the guarantee of no overfitting only holds at the model, i.e. without any outliers in the data. For these data, robust VIF regression provides the only viable option for policy makers as the 1437 features returned by classical VIF regression do not provide useful information. As can be seen in Figure 5, robust VIF is clearly the best performer for both scenarios. VIF regression chooses too many features for many of the folds and this leads to catastrophic results out-of-sample.

Figure 5: Crime and communities data: Out-of-sample median absolute prediction errors of the models chosen by classical and robust VIF regression, and the Lasso, in 10-fold cross-validation. * Results are not shown as VIF collapses in 4 folds, yielding MAPE of 5.62, 6.55, 6.82, 9.41, and 15.1, respectively. Results for other folds were good, 0.0652, 0.0676 0.0686, 0.0694, 0.0744, but are excluded from the boxplot to allow for a better comparisons of all methods.



6 Concluding remarks

In Lin, Foster, and Ungar (2011), it was also shown that classical VIF regression equates or outperforms *stepwise regression*, *Lasso*, *FoBa*, an adaptive forward-backward greedy algorithm focusing on linear models (Zhang 2009), and *GPS*, the generalized path-seeking algorithm of Friedman (2008). In this paper, we present a very efficient robust VIF approach that clearly outperforms classical VIF in the case of contaminated datasets. This robust implementation comes with a very small cost in speed, computation time is less than doubled, and provides much needed robust model selection for large datasets.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (Eds.), *Proceedings of the Second International Symposium on Information Theory*, Budapest, pp. 267–281. Akademiai Kiado.

Alqallaf, F., S. Van Aelst, V. J. Yohai, and R. H. Zamar (2009). Propagation of outliers in multivariate data. *The Annals of Statistics*, 37, 311–331.

Candes, E. J. and T. Tao (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35, 2313–2351.

Clark, D. (2011). Do recessions keep students in school? The impact of youth unemployment on enrollment in post-compulsory education in England. *Economica*, 78, 523–545.

Dupuis, D. J. and M.-P. Victoria-Feser (2011). Fast robust model selection in large datasets. *Journal of the American Statistical Association*, 106, 203–212.

Efron, B. (2004). The estimation of prediction error: Covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99, 619–632.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression (with discussion). *Annals of Statistics*, 32, 407–499.

Foster, D. P. and R. Stine (2004). Variable selection in data mining: Building a predictive model for bankruptcy. *Journal of the American Statistical Association*, 99, 303–313.

Foster, D. P. and R. A. Stine (2008). Alpha-investing: A procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society, Ser. B*, 70, 429–444.

Frank, A. and A. Asuncion (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Friedman, J.H. (2008). Fast sparse regression and classification. Technical report, Stanford University.

Friedman, J. H., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33, 1–22.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106, 746–762.

Hampel, F. R. (1968). *Contribution to the Theory of Robust Estimation*. Ph. D. thesis, University of California, Berkeley.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69, 383–393.

Heritier, S., E. Cantoni, S. Copt, and M. Victoria-Feser (2009). *Robust Methods in Biostatistics*. Wiley.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35, 73–101.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under non-standard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, pp. 221–233.

Khan, J. A., S. Van Aelst, and R. H. Zamar (2007). Robust linear model selection based on least angle regression. *Journal of the American Statistical Association*, 102, 1289–1299.

Kienzl, G. S., M. Alfonso, and T. Melguizo (2007). The effect of local labor market conditions in the 1990s on the likelihood of community college students' persistence and attainment. *Research in Higher Education*, 48, 751–774.

Lin, D., D. P. Foster, and L. H. Ungar (2011). VIF regression: A fast regression algorithm for large data. *Journal of the American Statistical Association*, 106, 232–247.

Machado, J. A. F. (1993). Robust model selection and M-estimation. *Econometric Theory*, 9, 478–493.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, 15, 661–675.

Marquardt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*, 12, 591–612.

Pennington, K.L., D. McGinty and M. R. Williams (2002). Community college enrollment as a function of economic indicators. *Community College Journal of Research and Practice*, 26, 431–437.

Petrongolo, B. and M. J. San Segundo (2002). Staying-on at school at 16: the impact of labor market conditions in Spain. *Economics of Education Review*, 21, 353–365.

Renaud, O. and M.-P. Victoria-Feser (2010). A robust coefficient of determination for regression. *Journal of Statistical Planning and Inference*, 140, 1852–1862.

Ronchetti, E. (1982). *Robust Testing in Linear Models: The Infinitesimal Approach.* Ph. D. thesis, ETH, Zurich, Switzerland.

Ronchetti, E., C. Field, and W. Blanchard (1997). Robust linear model selection by cross-validation. *Journal of the American Statistical Association*, 92, 1017–1023.

Ronchetti, E. and R. G. Staudte (1994). A robust version of Mallows's C_p. *Journal of the American Statistical Association*, 89, 550–559.

Rouse, C.E. (1995). Democratization or diversion? The effect of community colleges on educational attainment. *Journal of Business and Economic Statistics*, 12, 217–224.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461–464.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

Wetterlind, P. J. (1976). A multi-variable input model for the projection of higher education enrollments in Arizona. Unpublished doctoral dissertation, University of Arizona, Tucson.

Zhang, T. (2009). Adaptive forward-backward greedy algorithm for sparse learning with linear models. *Advances in Neural Information Processing Systems*, 21, 1921–1928.

Zhou, J., D. P. Foster, R. A. Stine, and L. H. Ungar (2006). Streamwise feature selection. *Journal of Machine Learning Research*, 7, 1861–1885.

Appendix - **Algorithm** Robust VIF regression

The robust VIF regression procedure, based on a streamwise regression approach and α-investing, can be summarized by the following algorithm :

Input data $\boldsymbol{y}, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ (standardized)

Set initial wealth $a_0 = 0.50$, pay-out $\Delta a = 0.05$, subsample size m, and robustness constant c.

Compute efficiency e_c^{-1} where

$$e_c = \left[\int_{-c}^{c}(5r^4/c^4 - 6r^2/c^2 + 1)d\Phi(r)\right]^2 \Big/ \int_{-c}^{c} r^2(r^2/c^2 - 1)^4 d\Phi(r).$$

Get All marginal weights w_{ij} by fitting p marginal models $y = \beta_{01} + x_1\beta_1 + \varepsilon_1, \ldots, y = \beta_{0p} + x_p\beta_p + \varepsilon_p$ using (5) and (6).

Initialize $j = 1$, $S = \{0\}$, $\boldsymbol{X}_S = \mathbf{1}$, $\boldsymbol{X}_S^w = diag(\sqrt{w_{iS}^0})\boldsymbol{X}_S$ and $\boldsymbol{y}^w = diag(\sqrt{w_{iS}^0})\boldsymbol{y}$ where w_{iS}^0 is computed using (4) where $\boldsymbol{r}^0 = (\boldsymbol{y} - \mathbf{1}\hat{\boldsymbol{\beta}}^0)/\hat{\sigma}^0$ using $\boldsymbol{X}_0^w = \boldsymbol{X}_0^{w2} = \mathbf{1}$, $\widehat{\boldsymbol{\beta}}^0 = \left[(\boldsymbol{X}_0^w)^T\boldsymbol{X}_0^w\right]^{-1}(\boldsymbol{X}_0^{w2})^T\boldsymbol{y}$, where $\hat{\sigma}^0 = 1.483\text{med}|\tilde{\boldsymbol{r}}^0 - \text{med}(\tilde{\boldsymbol{r}}^0)|$ and $\tilde{\boldsymbol{r}}^0 = \boldsymbol{y} - \mathbf{1}\widehat{\boldsymbol{\beta}}^0$.

repeat

set $\alpha_j = a_j/(1 + j - f)$

get T_w from the five-step *Fast Robust Evaluation Procedure* in §2.3.

if $2(1 - \Phi(|T_w|)) < \alpha_j$ **then**

$S = S \cup \{j\}$, $\boldsymbol{X}_S = [\mathbf{1} \;\; \boldsymbol{x}_j]$, $\boldsymbol{X}_S^w = diag(\sqrt{w_{iS}^0})\boldsymbol{X}_S$, $\boldsymbol{y}^w = diag(\sqrt{w_{iS}^0})\boldsymbol{y}$,

where w_{iS}^0 is computed using (4) where $\boldsymbol{r}^0 = (\boldsymbol{y} - \boldsymbol{X}_S\hat{\boldsymbol{\beta}}^0)/\hat{\sigma}^0$ using $\boldsymbol{X}_0^w = [1 \;\; \sqrt{w_{ij}}x_{ij}]$, $\boldsymbol{X}_0^{w2} = [1 \;\; w_{ij}x_{ij}]$, $i = 1, \ldots, n$, $\widehat{\boldsymbol{\beta}}^0 = \left[(\boldsymbol{X}_0^w)^T\boldsymbol{X}_0^w\right]^{-1}(\boldsymbol{X}_0^{w2})^T\boldsymbol{y}$, where $\hat{\sigma}^0 = 1.483\text{med}|\tilde{\boldsymbol{r}}^0 - \text{med}(\tilde{\boldsymbol{r}}^0)|$ and $\tilde{\boldsymbol{r}}^0 = \boldsymbol{y} - \boldsymbol{X}_S\widehat{\boldsymbol{\beta}}^0$.

$a_{j+1} = a_j + \Delta a$

$f = j$

else $a_{j+1} = a_j - \alpha_j/(1 - \alpha_j)$

end if

$j = j + 1$

until all p covariates have been considered