CorReg

Clément Théry, Christophe Biernacki, Gaétan Loridant

ArcelorMittal Dunkerque, Université de Lille 1,équipe MΘdal Inria

February 8, 2015

1. Steel industry databases.
2. Goal: To understand and prevent quality problems on finished product, knowing the whole process, without a priori.






Regression

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

where $\varepsilon \sim \mathcal{N}(0, \sigma_Y^2)$

OLS

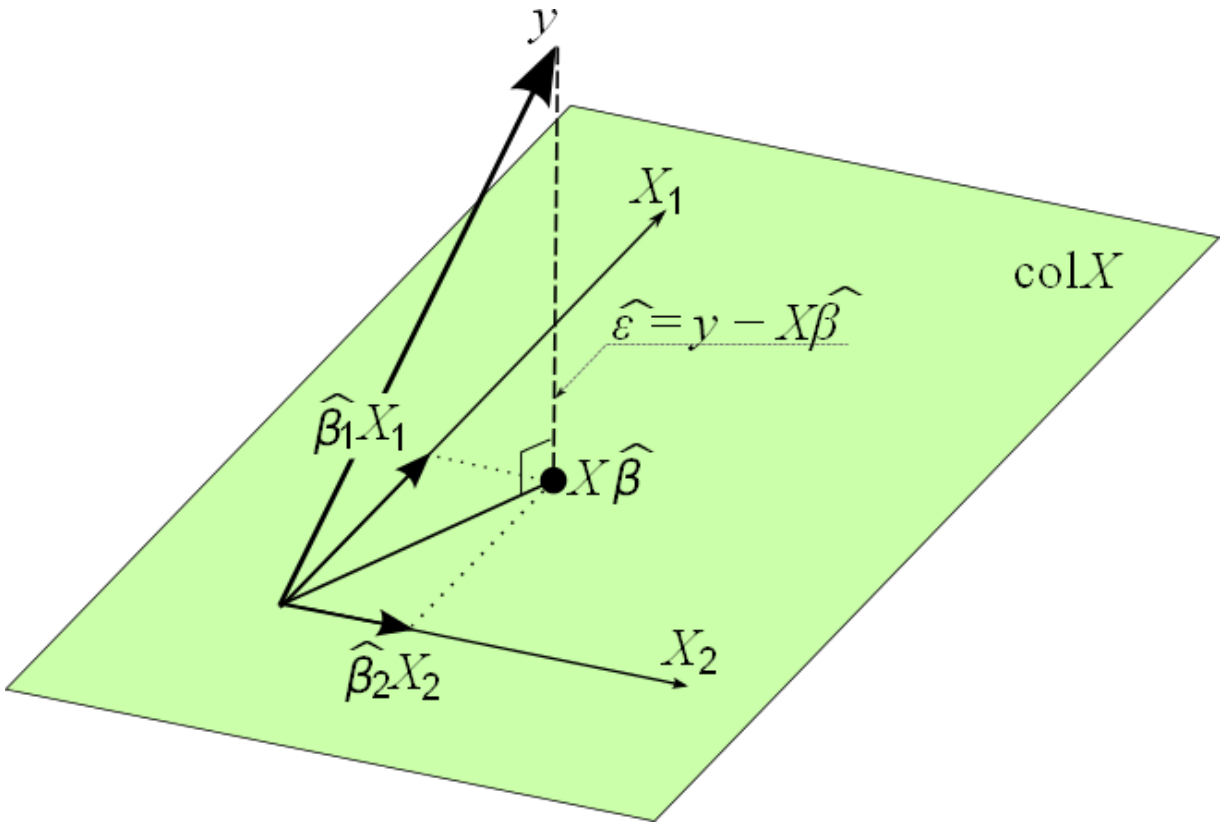

Figure: Multiple linear regression with Ordinary Least Squares seen as a projection on the $d-$dimensional hyperplane spanned by the regressors $\boldsymbol{X}$. Public domain image.

OLS

β can be estimated by $\hat{\beta}$ with Ordinary Least Squares (OLS), that is the unbiased maximum likelihood estimator [Saporta, 2006, Dodge and Rousson, 2004]:

$$\hat{\beta}_{OLS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \tag{2}$$

with variance matrix

$$\mathrm{Var}(\hat{\beta}_{OLS}) = \sigma_Y^2\,(\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{3}$$

In fact it is the Best Linear Unbiased Estimator (BLUE). The theoretical MSE is given by

$$\mathrm{MSE}(\hat{\beta}_{OLS}) = \sigma_Y^2\,\mathrm{Tr}((\boldsymbol{X}'\boldsymbol{X})^{-1}).$$

Running example

$\boldsymbol{X}^1, \boldsymbol{X}^2, \boldsymbol{X}^4, \boldsymbol{X}^5 \sim \mathcal{N}(0,1)$ and $\boldsymbol{X}^3 = \boldsymbol{X}^1 + \boldsymbol{X}^2 + \varepsilon_1$ where $\varepsilon_1 \sim \mathcal{N}(\boldsymbol{0}, \sigma_1^2 \boldsymbol{I}_n)$.
Two *scenarii* for $\boldsymbol{Y}$:
$\beta = (1,1,1,1,1)'$ and $\sigma_Y \in \{10, 20\}$.
It is clear that $\boldsymbol{X}'\boldsymbol{X}$ will become more ill-conditioned as σ_1 gets smaller. R^2 stands for the coefficient of determination which is here:

$$R^2 = 1 - \frac{\text{Var}(\varepsilon_1)}{\text{Var}(\boldsymbol{X}^3)} \tag{4}$$

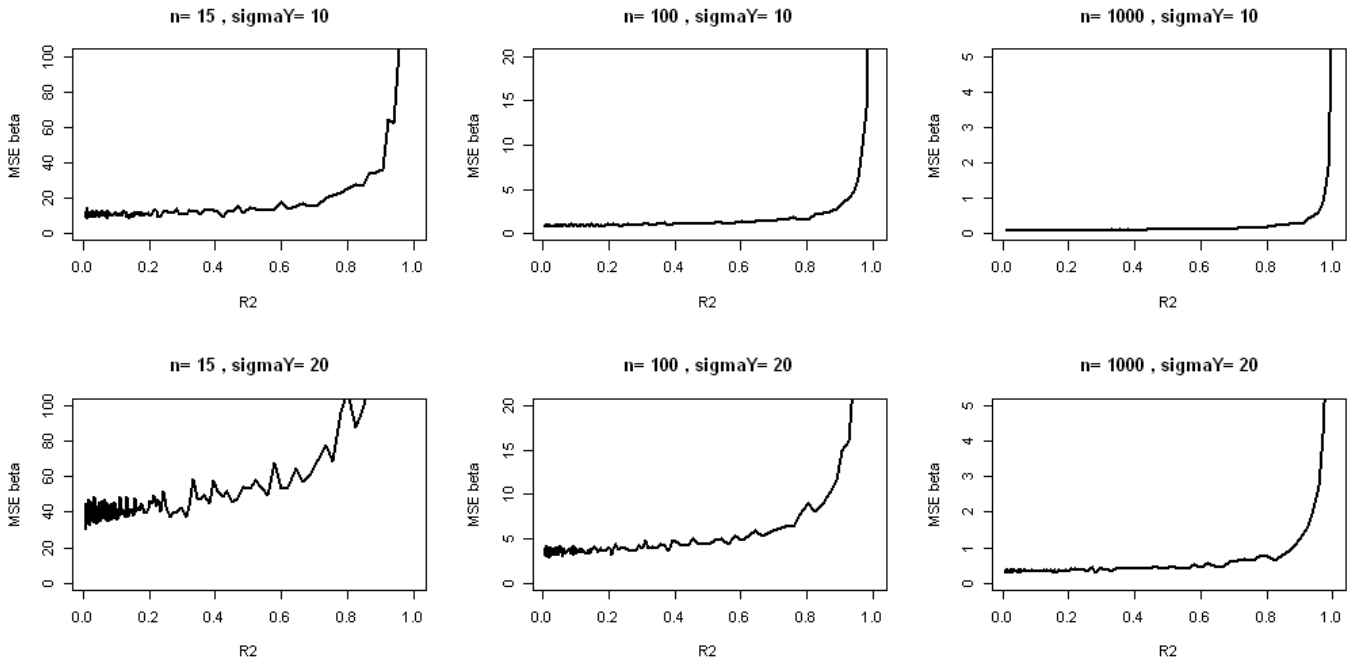

Figure: Evolution of observed Mean Squared error on $\hat{\beta}_{OLS}$ with the strength of the correlations for various sample sizes and strength of regression. $d = 5$ covariates (running example).

Ridge

Alternative ridge : bien mais pas de sélection donc inacceptable

Ridge

Alternative ridge : bien mais pas de sélection donc inacceptable

LASSO

Alternative lasso (et autres) : bien mais problèmes en cas de corrélations

LASSO

Liste des alternatives

SEM

Modélisation de la structure mais à la main et aucun impact sur l'estimation

Selvarclust

Semble très bien mais n'aboutit pas vers la régression donc on le prolonge en CorReg

Modèle génératif complet avec dépendances

Explosion des mélanges

SEM avec Gibbs

Bic pondéré

Résultats pourris

Excel, fonctions graphiques, arbres de décision


Dodge, Y. and Rousson, V. (2004).
Analyse de régression appliquée: manuel et exercices corrigés (coll. eco sup,).
Recherche, 67:02.


Saporta, G. (2006).
Probabilités, analyse des données et statistique.
Editions Technip.