Measuring unexplained variation in insurance data: a non-parametric approach based on global sensitivity indices

Giovanni Rabitti

Department of Actuarial Mathematics and Statistics, Heriot-Watt University

Insurance Data Science Conference 2024

Motivation

- Measures of explained variations are useful in scientific research, as they quantify the amount of variation in an outcome variable of interest that is explained by one or more other variables.
- This information helps us understand the 'quality' of a dataset, before training any model.
- Hoessjer et al. (2009) estimate the proportion of total variation explained by a Poisson regression model for claim counts.
- However, past works rely on regression-based quantification of the explained variance.

  In this talk, we want to present a non-parametric quantification method.

Notation and setting

- We assume that the quantity of interest Y is linked to the covariates as

$$Y = f(\mathbf{X}) = f(X_1, X_2, ..., X_d).$$

- Let us denote $\mathbf{X}_u = \{X_i : i \in u\}$.
  For example, $\mathbf{X}_{\{3,4,6\}} = (X_3, X_4, X_6)$.

Relationship to variance decomposition

Variance decomposition of $Y = f(\mathbf{X})$:

$$Var(f(\mathbf{X})) = Var\left[\mathbb{E}\left(f(\mathbf{X})|\mathbf{X}\right)\right] + \mathbb{E}\left[Var\left(f(\mathbf{X})|\mathbf{X}\right)\right] \tag{1}$$

where $Var\left[\mathbb{E}\left(f(\mathbf{X})|\mathbf{X}\right)\right]$ is called the regression variance and $\mathbb{E}\left[Var\left(f(\mathbf{X})|\mathbf{X}\right)\right]$ the residual variance.

Green (1993) writes that "In analyzing a regression, we shall usually be interested in which of the two parts ot the total variance $Var(Y)$ is the larger one. [...] A natural measure is the ratio

$$\text{coefficient of determination} = \frac{\text{regression variance}}{\text{total variance}}."$$

Total sensitivity index

We consider the total importance index from global sensitivity analysis (Homma and Saltelli, 1996). Total effect of group u is

$$T_u = \frac{\mathbb{E}\left[Var\left(f(\mathbf{X})|\mathbf{X}_{-u}\right)\right]}{Var\left(f(\mathbf{X})\right)} = \frac{\mathbb{E}\left[\left(\mathbb{E}\left(f(\mathbf{X})|\mathbf{X}_{-u}\right) - f(\mathbf{X})\right)^2\right]}{Var\left(f(\mathbf{X})\right)}.$$

In words, it is the expected variance that would be left if all factors but $\mathbf{X}_u$ could be fixed.

By symmetry, the total effect of group $-u = \{1, 2, ..., d\} \setminus u$ is

$$T_{-u} = \frac{\mathbb{E}\left[Var\left(f(\mathbf{X})|\mathbf{X}_u\right)\right]}{Var\left(f(\mathbf{X})\right)}.$$

Given data setting

- However, we have available a dataset $\{(y_i, \mathbf{x}_{i,k})\}$, with $i = 1, 2, ..., n$ and $k = 1, 2, ..., d$.
- We don't want to estimate any metamodel $\hat{f}$ because it will mediate the effect of covariates, and its accuracy will affect the variance of the output explained by the covariates.
- Hence, we want a non-parametric way to estimate the unexplained variations .

From now on, denote with $\mathbf{X}_{OBS}$ the set of observed variables, and with $\mathbf{X}_{UNOBS}$ the set of unobserved variables, so that

$$\mathbf{X} = (\mathbf{X}_{OBS}, \mathbf{X}_{UNOBS}).$$

Hence,

$$T_{UNOBS} = \frac{\mathbb{E}\left[Var\left(Y|\mathbf{X}_{OBS}\right)\right]}{Var\left(Y\right)}.$$

Example

Consider the linear model

$$Y = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_d X_d + \beta_{d+1} X_{d+1}$$

where all X_i's are iid standard normal distributions. Assume that the variable X_{d+1} is not observed.

The variance explained by the model is $T_{OBS} = \sum_{j=1}^{d} \beta_j^2 / Var(Y)$ and equivalently $T_{OBS} = R^2$ where R^2 is the goodeness-of-fit index for linear models.

Consequently, $T_{UNOBS} = \beta_{d+1}^2 / Var(Y) = 1 - R^2$ is the fraction of the variance unexplained by the linear model.

Properties

As Honerkamp-Smith and Xu (2016) write, the measures for explained variations can be seen as a squared rank correlation. Kendall and Gibbons (1990) consider three properties for these rank correlations. Equivalently, measures of unexplained variations should:

1. lie between 0 and 1.
2. decrease with the strength of the association
3. have value 1 if there is no association, and value 0 if there is perfect association.

We can prove that the estimator T_{UNOBS} satisfies all three properties.

Proposition

The index T_{UNOBS} is bounded by $0 \leq T_{UNOBS} \leq 1$. Moreover, T_{UNOBS} decreases (increases, respectively) as the variance explained by $\mathbf{X}_{OBS}$ increases (decreases, respectively).

Computational algorithm: the fully observed cohorts

Consider a target observation t with $t = 1, 2, ..., n$. Define the set $C_{t,u}$ as

$$C_{t,u} = \{i = 1, 2, ..., n \quad | \quad z(x_{i,j}, x_{t,j}) = 1 \quad \text{for} \quad j \in u\} \qquad (2)$$

where $z(x_{i,j}, x_{t,j}) = 1$ if $|x_{i,j} - x_{t,j}| \leq \delta_j$, 0 otherwise.
Precisely, $C_{t,u}$ is the set of observations whose values in variables indexed by u are similar to those of the target observation t.
Mase et al. (2019) call $C_{t,u}$ the cohort of subject t for variables u.

We define the fully observed cohort for subject t as

$$C_{t,OBS} = \{i = 1, 2, ..., n \quad | \quad z(x_{i,j}, x_{t,j}) = 1 \quad \text{for every observed j}\}\,.$$

If we compute the mean of the fully observed cohort of t, we get

$$\overline{y}_{t,OBS} = \frac{1}{|C_{t,OBS}|} \sum_{i \in C_{t,OBS}} y_i .$$

This empirical mean is an estimate of

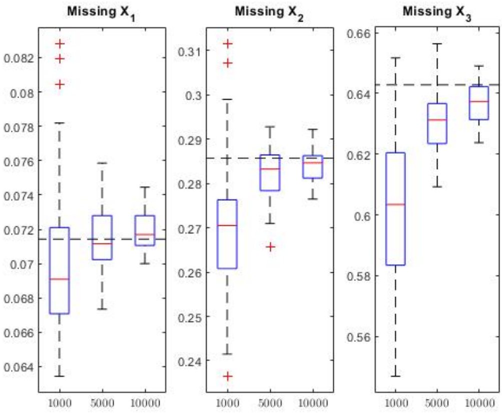
$$\mathbb{E}\left(Y | \mathbf{X}_{t,OBS} = \mathbf{x}_{t,OBS}\right) = \mathbb{E}\left(Y | \mathbf{X}_{t,-UNOBS} = \mathbf{x}_{t,-UNOBS}\right).$$

Averaging over all subjects, the total importance of unobserved variables can be estimated by

$$\hat{T}_{UNOBS} = \frac{\frac{1}{n}\sum_{t=1}^{n}\left(\overline{y}_{t,OBS} - y_t\right)^2}{\hat{\sigma}^2}.$$

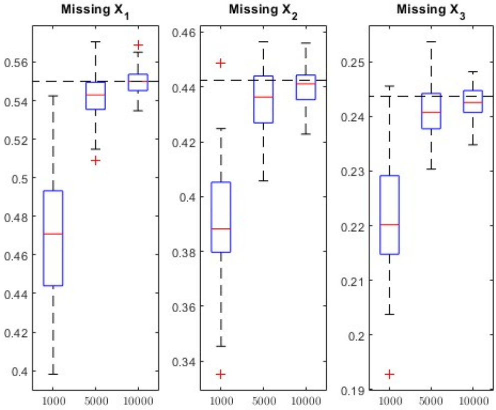
Numerical simulation: linear model

Consider the model $Y = X_1 + 2X_2 + 3X_3$ with all X_i standard normal distributed.



Numerical simulation: Ishigami function

Consider the Ishigami model $Y = 7\sin(X_2)^2 + \sin(X_1)(1 + 0.1X_3^4)$ with $X_1, X_2, X_3 \sim U(-\pi, \pi)$.



Application: Medical malpractice insurance costs dataset

- This dataset comprises 4558 insurance losses resulting from medical malpractice cases in various hospitals in Lombardy, Italy.
- Six risk factors: hospital code, medical department, type of claim, total number of hospitalizations, year of the claim, and the Case Mix Index (CMI), which represents a hospital's patient mix.
- AIM: claim amounts explained by the aforementioned risk factors.
- The code executed in 0.48 seconds, and the result is $\hat{T}_{UNOBS} = 0.5043$.

Application: Global Health Observatory life expectancy

- This dataset contains 1649 complete observations of life expectancy and various health and economic variables for 193 countries.
- twenty-one explanatory variables: Country, year, Country status (developed or developing), adult mortality rates, infant deaths, alcohol consumption, health expenditure, hepatitis B coverage immunization, number of measles, population BMI, deaths under five years, polio immunization coverage, government expenditure on health, diphtheria immunization coverage, HIV, Country GDP, Country population, prevalence of thinness for Age 10 to 19, prevalence of thinness for Age 5 to 9, human development index, and the number of years of schooling.
- The Quality of Interest (QoI) is the life expectancy in years.
- From this dataset, the estimated unexplained variation in life expectancy is $\hat{T}_{UNOBS} = 0.0013$, computed in 0.332 seconds.

Application to four insurance datasets

- Australian auto claims dataset: there are 67856 policies. The variables are related to the vehicle (age, type, value) and to the policyholder (age, area, gender).
- Singapore auto claims dataset: this dataset contains 7483 observations. The variables include vehicle variables (type of vehicle, age and if it is private or not), as well as person policyholder variables, such as age, gender and prior accident record.
- French auto claims dataset: it contains the claim frequency for 677,991 policyholders. The observed variables describe the vehicle characteristics (power, age, brand and engine type - gas/diesel/regular) plus characteristics of the policyholder (age, bonus-malus class, typology of the living area, density of inhabitants, region in France).

- Telematics auto claims dataset: it contains the observed claim frequency for 100,000 policyholders with traditional and telematics observed variables:
    - Traditional variables: policyholder's characteristics (age, gender, marital status, credit score, region type, expected annual driven miles, number of years without claims) and vehicle features (age, type of use, territorial location);
    - Telematics variables: annualized percentage of time on the road, total distance driven in miles, percent of driving day mon/tue/. . . /sun, percent vehicle driven within 2hrs/3hrs/4hrs, percent vehicle driven during wkday/wkend, percent of driving during xx rush hours: am/pm, mean number of days used per week, number of sudden acceleration 6/8/9/. . . /14 mph/s per 1000 miles, number of sudden brakes 6/8/9/. . . /14 mph/s per 1000 miles, number of left turn per 1000 miles with intensity 08/09/10/11/12, number of right turn per 1000 miles with intensity 08/09/10/11/12.

Results

Dataset	Obs	Sec	Variables	$\hat{T}_{UNOBS}$
Australian	67856	128.87	6	0.9608
Singapore	7483	0.98	6	0.9493
French	678013	19528.61	9	0.3739
Telematics (traditional)	100000	529.75	10	0.1781
Telematics (only telem.)	100000	1921.96	39	0.0241
Telematics (all variab.)	100000	2151.33	49	0.0043

- The computational time is more influence by the number of observations in the dataset rather than the number of explanatory variables.
- The number of observations seems not to affect the explained variance (see Australia vs Singapore).
- The introduction of telematics variables decreases the unexplained variance.

Conclusions

- We have considered the problem of estimating the variance explained by covariates in a non-parametric setting.
- With this insight we can quantify the "value" and "quality" of the dataset.
- Concerning the actuarial application, our results on four insurance dataset show that telematics data contain much more information than traditional insurance datasets, and we have quantified this increase.
- Future research: on this notion and on the computational scaling to the big data setting.