From least squares to general linear models with variables selection

May 2, 2012

Background

Suppose you have outputs of a system y_i, $i = 1, \ldots, n$ under some particular inputs represented by a set of features $\mathbf{x}_i = [x_{i,1}, \ldots, x_{i,p}]^T$, $i = 1, \ldots, n$.

An essential goal in this setting is prediction (or extrapolation): for a given new generic set of features $x_1, \ldots, x_p$, what is the expected output y ?

A seemingly more simple question than the prediction problem is the variables selection problem: which features do influence the response y ?

An industrial example

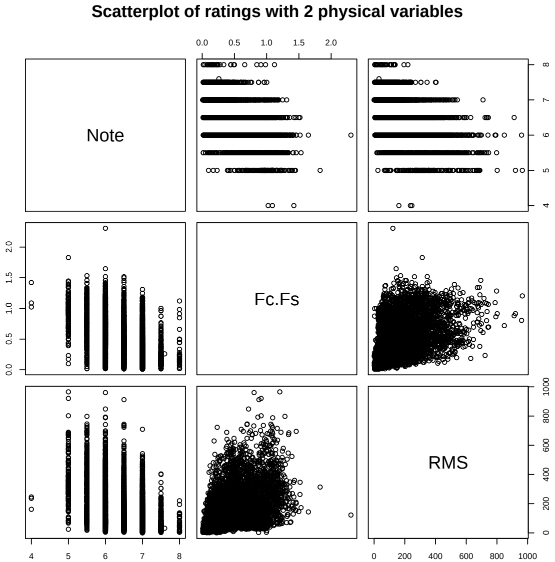
Renault Technocentre data (confidential)

The variable y are ratings that evaluate the quality of a car manual transmission using a gear shifting lever with half integer values from 0 to 10. The features are either

1. quantitative : particular physical production settings (RMS, Fc/Fs ...).
2. categorical : “Temp” (cold/warm), “Boite” (3 categories), “Rapports” (6 categories of shifts), “Vehicule” (3 categories), “Juge” (2 categories), “Type” (trial conditions: on the road/ in the lab).

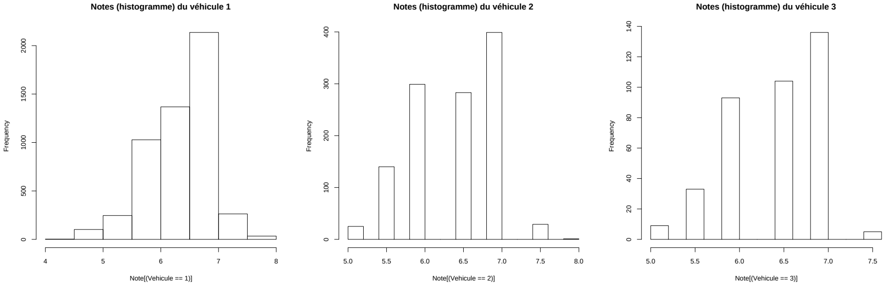
An industrial example (cont.)

Examples of quantitative variables: RMS, Fc/Fs.



An industrial example (cont.)

Example of qualitative variables (factor): "Vehicule".



Linear regression model

Assuming your data is collected in an i.i.d. context, you want to evaluate the common conditional density $\mathrm{p}(y|x_1, \ldots, x_p)$, that is a regression model.
The most simple one is to assume that y is a linear function of $\mathbf{x}$, up to an additive noise.

linear model

Assume that

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad i = 1, 2, \ldots, n \; ,$$

where $\boldsymbol{\beta}$ is an unknown parameter (same dimension as $\mathbf{x}$) and ϵ_i are i.i.d. with zero mean and (possibly unknown) variance σ^2.

Linear regression model: some remarks

1. If a feature is qualitative taking values say in $1, 2, \ldots, q$, one defines quantitative variables $x_i = \mathbb{1}(x = i)$, for each $i = 1, 2, \ldots, q-1$.
2. $\boldsymbol{\beta}$ is set as a vector column $[\beta_1, \ldots, \beta_p]^T$, where β_k is the regression coefficient of the k-th feature.
3. In the linear model, one has $\mathbb{E}[y|\mathbf{x}] = \mathbf{x}^T\boldsymbol{\beta}$.
4. One often adds up an artificial feature $x_{i,1} = 1$ (the so called intercept), that is, y is an affine function of the original $\mathbf{x}$, up to an additive noise.
5. It is often assumed that the additive noise ϵ_i is Gaussian.
6. Some transformation of y and/or $\mathbf{x}$ prior to applying a linear model can be useful.
7. To obtain a more general model (polynomial), one can add features built from the available ones, e.g. $x_{i,k}^2$, $x_{i,j}x_{i,k}$, ...
8. In the linear model, the variables selection problem amounts to find the feature indices k for which $\beta_k \neq 0$.

Least square estimation

Since $\mathbb{E}[y|\mathbf{x}] = \mathbf{x}^T\boldsymbol{\beta}$, it is sufficient to estimate $\boldsymbol{\beta}$ to solve the prediction problem. Mimicking the relationship

$$\boldsymbol{\beta} = \operatorname*{Argmin}_{\phi} \mathbb{E}[(y - \mathbf{x}^T\phi)^2] \ ,$$

one gets the least squares estimator, based on observations $(y_i, \mathbf{x}_i)_i$, $i = 1, \ldots, n$,

$$\widehat{\boldsymbol{\beta}}_n = \operatorname*{Argmin}_{\phi} \frac{1}{n}\sum_{i=1}^{n}((y_i - \mathbf{x}_i^T\phi)^2 \ .$$

Set $\mathbf{y} = [y_1, \ldots, y_n]^T$ and $\mathbf{X}_n = [\mathbf{x}_1, ..., \mathbf{x}_n]^T$, and suppose that $\mathbf{X}_n$ has full rank. One obtains

$$\widehat{\boldsymbol{\beta}}_n = (\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{X}_n^T\mathbf{y}_n \ .$$

Linear optimality

Gauss-Markov Theorem

One easily shows that $\widehat{\boldsymbol{\beta}}_n$ is unbiased,

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}_n] = \boldsymbol{\beta} \ .$$

Moreover, for all $\lambda \in \mathbb{R}^p$, $\lambda^T \widehat{\boldsymbol{\beta}}_n$ is the best unbiased linear estimator of $\lambda^T \boldsymbol{\beta}$: for all linear function $S : \mathbb{R}^n \to \mathbb{R}$ such that $\mathbb{E}[S(\mathbf{y}_n)] = \lambda^T \boldsymbol{\beta}$,

$$\mathrm{var}(S(\mathbf{y}_n)) \geq \mathrm{var}(\lambda^T \widehat{\boldsymbol{\beta}}_n) \ .$$

Linear optimality

Gauss-Markov Theorem

One easily shows that $\widehat{\boldsymbol{\beta}}_n$ is unbiased,

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}_n] = \boldsymbol{\beta} \ .$$

Moreover, for all $\lambda \in \mathbb{R}^p$, $\lambda^T \widehat{\boldsymbol{\beta}}_n$ is the best unbiased linear estimator of $\lambda^T \boldsymbol{\beta}$: for all linear function $S : \mathbb{R}^n \to \mathbb{R}$ such that $\mathbb{E}[S(\mathbf{y}_n)] = \lambda^T \boldsymbol{\beta}$,

$$\operatorname{var}(S(\mathbf{y}_n)) \geq \operatorname{var}(\lambda^T \widehat{\boldsymbol{\beta}}_n) \ .$$

However the unbiased assumption is purely artificial! Biased estimator may enjoy much better performances, especially when p/n is not "small".

Gaussian Likelihood

Gaussian assumption

Suppose that $(\epsilon_i)_{i\geq 1}$ are i.i.d. centered Gaussian r.v.'s with variance σ.

Then The Log-likelihood writes

$$(\phi, s) \mapsto -\frac{p}{2}\log(2\pi s^2) - \frac{1}{2s^2}\|\mathbf{y}_n - \mathbf{X}_n\phi\|^2 \ ,$$

where $\|\cdot\|$ here denotes the Euclidean norm in $\mathbb{R}^n$.
Hence the least square estimator is the maximum likelihood estimator.

Non-asymptotic distributions

Moreover many statistics of interest have well known distribution: χ^2, Student (σ is unknown). This allows for

1. Confidence intervals for $\boldsymbol{\beta}$ coefficients.
2. Confidence intervals for outliers detection.
3. Confidence intervals for predictors.
4. Statistical hypotheses testing, e.g. $H_0 = \{\boldsymbol{\beta} = 0\}$.

The ANOVA provides this kind of features.
Moreover hypotheses testing are sometimes used for variable selection.

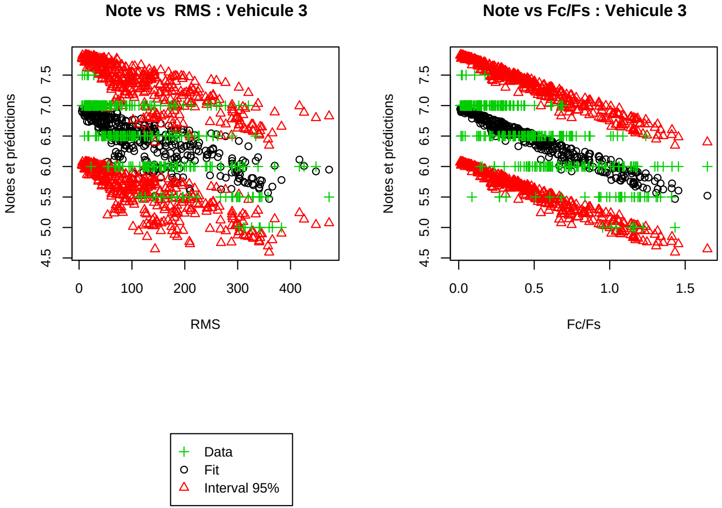
An industrial example (cont.)

Using R software, a linear regression is performed with y =ratings explained by 1 factor (the car : 3 categories) and 2 quantitative variables (physical measures of Fc/Fs, RMS).

```
Call:
lm(formula = Note ~ factor(Vehicule) + Fc.Fs + RMS)

Residuals:
    Min       1Q   Median       3Q      Max
-2.07527 -0.27582  0.05428  0.25067  1.95240

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        7.055e+00  9.825e-03 718.069   <2e-16 ***
factor(Vehicule)2  4.286e-02  1.675e-02   2.559   0.0105 *
factor(Vehicule)3 -7.641e-02  2.375e-02  -3.218   0.0013 **
Fc.Fs             -7.942e-01  2.013e-02 -39.464   <2e-16 ***
RMS               -1.037e-03  5.552e-05 -18.670   <2e-16 ***
---
Signif. codes:  0 **0.001 *0.01 0.05 0.1 1

Residual standard error: 0.4464 on 6731 degrees of freedom
Multiple R-squared: 0.3684, Adjusted R-squared: 0.3681
F-statistic: 981.7 on 4 and 6731 DF,  p-value: < 2.2e-16
```

An industrial example (cont.)

Plots of fitted values for “Vehicule 3”, with confidence intervals.



An industrial example (cont.)

An Anova of this regression gives

```
Analysis of Variance Table

Response: Note
                   Df  Sum Sq Mean Sq  F value    Pr(>F)
factor(Vehicule)    2   32.52   16.26   81.605 < 2.2e-16 ***
Fc.Fs               1  680.46  680.46 3414.978 < 2.2e-16 ***
RMS                 1   69.45   69.45  348.553 < 2.2e-16 ***
Residuals        6731 1341.20    0.20
---
Signif. codes:  0 **0.001 *0.01 0.05 0.1 1
```

Robust regression

Least square estimators are know to be sensitive to outliers.
To avoid an outliers detection step, one can use less sensitive cost functions such as minima absolute deviation (MAD) :

$$\widehat{\boldsymbol{\beta}}_n = \underset{\phi}{\mathrm{Argmin}} \frac{1}{n} \sum_{i=1}^{n} |y_i - \mathbf{x}_i^T \phi| \ .$$

The contrast is still convex and can thus be minimized using convex minimization techniques although uniqueness is no longer assured. Indeed, recall that

$$\underset{m}{\mathrm{Argmin}} \frac{1}{n} \sum_{i=1}^{n} |y_i - m| = \mathrm{median}(y_1, \ldots, y_n) \ .$$

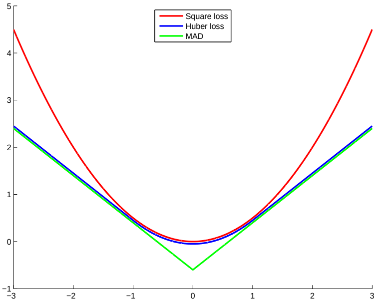
Robust regression (cont.)

A mixture of MAD and mean squared has been proposed by Huber (1981)

$$\widehat{\beta}_n = \underset{\phi}{\mathrm{Argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i - \mathbf{x}_i^T \phi) ,$$

where ℓ is quadratic inside a centered interval and linear outside,

$$\ell(z) = \frac{1}{2} z^2 \mathbb{1}_{|z|\leq 1} + (|z| - 1/2)\mathbb{1}_{|z|>1} .$$



General linear model

The linear model can be extended as follows: $y_1, \ldots, y_n$ are i.i.d. with exponential distribution $f_{\eta,\tau}$ where (η, τ) are two parameters such that η determines the mean of the distribution through a link function. Moreover , in a GLM, one sets

$$\eta = \mathbf{x}^T \boldsymbol{\beta} \ .$$

Examples

1. (Gaussian) Linear model: $f_{\eta,\tau}$ is the Gaussian distribution with mean η and variance τ.
2. Logit regression: the y's take values 0 and 1 with mean

$$\mathbb{E}[y] = \frac{\exp(\mathbf{x}^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}^T \boldsymbol{\beta})} \ .$$

(no τ parameter here!)

General linear model : estimation

Define a profile log-likelihhod

$$\ell(z, y) = -\sup_{t} \log f_{z,t}(y)$$

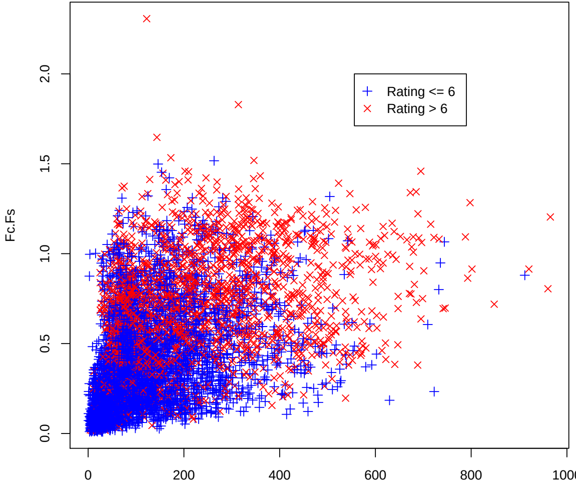
The least square estimator is replaced by

$$\widehat{\boldsymbol{\beta}}_n = \mathop{\mathrm{Argmin}}_{\phi} \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{x}_i^T \phi, y_i) \ .$$

The function ℓ is always convex and $\widehat{\boldsymbol{\beta}}_n$ can thus be computed using numerical convex optimization procedures.

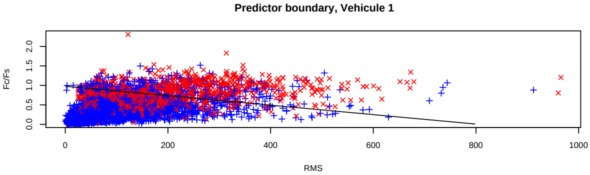
An industrial example (cont.)

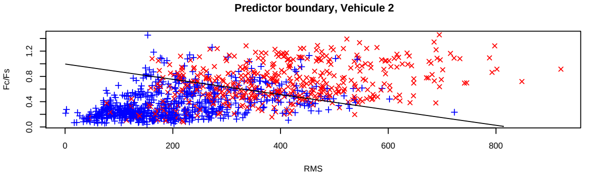
Map of the ratings to two categories : low (≤ 6) and high (> 6).
This gives the following distribution in the Fc/Fs vs RMS plane:

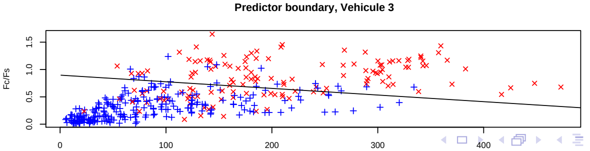


An industrial example (cont.)

Based on a logit regression, a partition of the Fc/Fs vs RMS plane for the 3 categories of the “Vehicule” factor is deduced.







Penalized regression

The performance of th regression highly depends on

1. how small p/n is (curse of dimensionality),
2. the distribution of the explanatory variables $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in $\mathbb{R}^p$ (potentially ill-posed inverse problems).

One needs a safeguard against unstable estimates!
Replace our basic contrast estimator

$$\widehat{\beta}_n = \operatorname*{Argmin}_{\phi} \quad \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{x}_i^T \phi, y_i)$$

Penalized regression

The performance of th regression highly depends on

1. how small p/n is (curse of dimensionality),
2. the distribution of the explanatory variables $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in $\mathbb{R}^p$ (potentially ill-posed inverse problems).

One needs a safeguard against unstable estimates!

by a penalized contrast estimator

$$\widehat{\boldsymbol{\beta}}_n = \underset{\phi}{\mathrm{Argmin}} \left[\frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{x}_i^T \phi, y_i) + \mathrm{pen}(\phi) \right]$$

Here $\mathrm{pen}(\phi)$ increases with the dimension of ϕ. Standard choice are

- Ridge regression: $\mathrm{pen}(\phi) = \lambda_n \sum_{i=1}^{p} \phi_i^2$
- Bayesian prior: $\mathrm{pen}(\phi) = -\log g_n(\phi)$
- Model selection via parameter dimension: Mallow's C_p, AIC, BIC ...

Curse of dimensionality

In the linear regression problem, the dimension of the unknown parameter β is p, the dimension of $\mathbf{x} = [x_1, \ldots, x_p]^T$. For a fixed dispersion parameter (say the variance),

the larger p/n, the more difficult the estimation

2 possible approaches

1. biased estimation: estimate "small coefficients" of β by 0.
2. model selection: find a correct submodel.

Goals are different but methods are similar:

$$\boxed{\text{decrease the number of coefficients to estimate.}}$$

Variable selection using penalized regression

Initiated with the AIC proposed by Akaïke (1971) based on information theory arguments, several variables selection methods based on penalties: Mallows Cp Mallows (1973), BIC Schwarz (1978).
In these approaches, $\mathrm{pen}(\phi)$ is proportional to the dimension of $\boldsymbol{\beta}$,

$$\mathrm{pen}(\phi) \propto \sum_{k=1}^{p} \mathbb{1}(\phi \neq 0) .$$

These methods have been revisited in the late 1990's using sophisticated probabilistic tools introduced by Birg and Massart (1998) such as concentration inequalities.

LASSO

For penalties based on parameter dimension, the computation of

$$\widehat{\beta}_n = \underset{\phi}{\mathrm{Argmin}} \left[\frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{x}_i^T \phi, y_i) + \mathrm{pen}(\phi) \right],$$

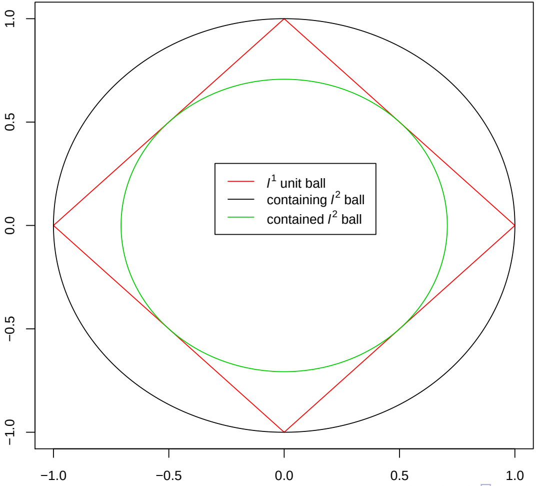
requires performing 2^p regressions, which makes its use limited to $p \leq 14, 15$. Much higher values of p are required in genomics or nowadays data mining applications.
To circumvent this, convex penalties have been proposed, the "closest" one to parameter dimension being

$$\mathrm{pen}(\phi) = \lambda_n \sum_{i=1}^{p} |\phi_i| \,.$$

For $\ell(z, y) = (z - y)^2$, one gets the LASSO.

Variable selection

Under ℓ^1 constraints, the points that are the ℓ^2-furthest away from the origin are on the axes (zero coefficient):



Regularization path

A quick algorithm has been proposed by Efron *et al* (2002) to compute

$$\widehat{\boldsymbol{\beta}}_n(\lambda) = \underset{\phi}{\mathrm{Argmin}} \left[\frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \phi)^2 + \lambda \sum_{i=1}^{p} |\phi_i| \right],$$

for all $\lambda > 0$.

As λ decreases one obtains more and more active (i.e. non zero) coefficients. The regularization path

$$\lambda \mapsto \widehat{\boldsymbol{\beta}}_n(\lambda)$$

thus defines as $\lambda \downarrow 0$ a sequence of models with increasing dimension.

Industrial example (cont.)

Consider the Renault dataset slightly changed into:

- Ratings reduced to $q = 4$ possible values.
- 15 quantitative variables (physical production settings)

It is important for the manufacturer to evaluate which variables have a significant impact on the rating.

Model

We use a general linear model for qualitative output $y \in \{1, 2, 3, 4 = q\}$: the multinomial model with logit link function,

$$\mathbb{P}(y = k) = \frac{\exp([\mathbf{x}^T \boldsymbol{\beta}]_k)}{1 + \sum_{j=1}^{q-1} \exp([\mathbf{x}^T \boldsymbol{\beta}]_k)}\ , \quad k = 1, 2, \ldots, q-1\ ,$$

where the parameter $\boldsymbol{\beta}$ is a $p \times q$ matrix, with $p = 16$ (15 variables + intercept).

Industrial example (cont.)

Following Park and Hastie (2007), the LASSO can be extended to general linear models

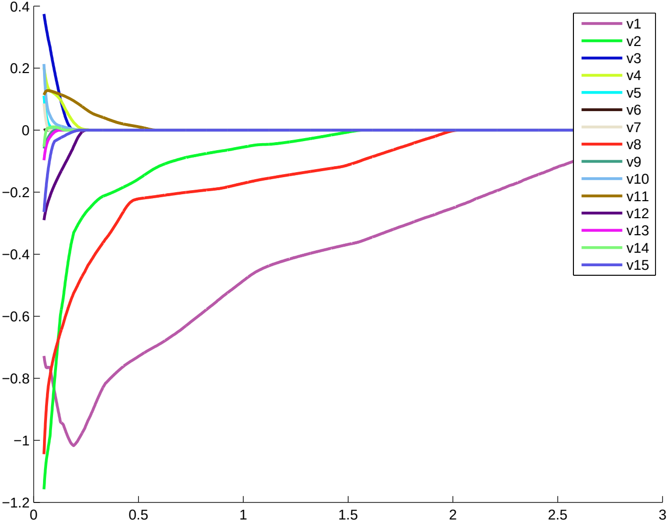
$$\widehat{\boldsymbol{\beta}}_n = \underset{\phi}{\mathrm{Argmin}} \left[\frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{x}_i^T \phi, y_i) + \lambda \sum_{i=2}^{p} \sum_{j=1}^{q-1} |\phi_{ij}| \right],$$

where ϕ is $p \times (q-1)$ matrix and for all $y \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^p$,

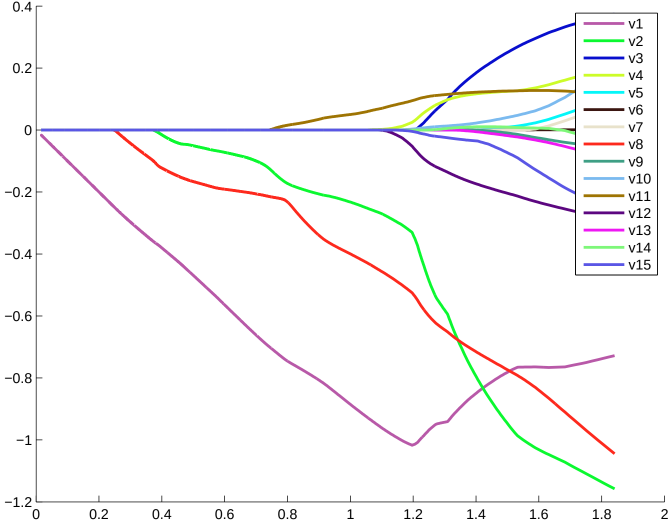
$$\ell(\mathbf{x}^T \phi, y) = \sum_{k=1}^{q-1} \frac{\mathbb{1}(y=k) \exp([\mathbf{x}^T \boldsymbol{\beta}]_k)}{1 + \sum_{j=1}^{q-1} \exp([\mathbf{x}^T \boldsymbol{\beta}]_k)} + \frac{\mathbb{1}(y=q)}{1 + \sum_{j=1}^{q-1} \exp([\mathbf{x}^T \boldsymbol{\beta}]_k)}.$$

(log-likelihood of the multinomial logit model)

Industrial example (cont.)
Regularization path: $\boldsymbol{\beta}_{1,2}, \ldots, \boldsymbol{\beta}_{1,p}$ VS λ



Industrial example (cont.)
Regularization path: $\boldsymbol{\beta}_{1,2}, \dots, \boldsymbol{\beta}_{1,p}$ VS its norm



Industrial example (cont.)

From this example, variables v1, v2 and v8 appear to be the most important in the rating.

An additional BIC penalty selection step confirms this result.

Variables v1, v2 (RMS and Fc/Fs already mentioned) were already well known to be key parameters by the engineers, v8 were discovered as a new potential one.

Concluding remarks

- The linear model is the basic regression model.
- It allows a quick analysis and simple to interpret in a statistical framework.
- Model selection and high dimensional data analysis are nowadays open issues in statistics.
- Computationally light methods are available, based on convex optimization.
- Statistical learning/regression can be a useful support to engineers.