15.095: Machine Learning under a Modern Optimization Lens

Lectures 4: Holistic Regression

Outline

The Art of Building Regression Models: Current Practice

- Transform variables.
- Pairwise scatterplots, correlation matrix.
- Delete highly correlated variables.
- Delete variables with insignificant t-tests. Examine residuals.
- See if additional variables can be dropped/new variables brought in.
- Validate the final model.

Aspirations: From Art to Science

- Propose an algorithmic (automated) process to build regression models.
- Approach: Express all desirable statistical properties as MIO constraints.
- Find a model for which are properties are satisfied or obtain a proof that it is not possible.

What is a good regression model?

- Sparsity.
- Selective sparsity: Group Sparsity and Limited Pairwise Multi-Collinearity.
- Detecting Appropriate Nonlinear Transformations.
- Robustness.
- Statistical Significance.
- Low Global Multicollinearity.

Preprocessing

- Split data 50%/25%/25% into a training, validation, and test set.
- Normalize Training set: columns with zero mean and unit ℓ_2-norm.
- Set robustification parameter Γ.
- Set maximum pairwise correlation ρ.
- $\mathcal{HC}$ set of pairs of variables that are highly correlated.
- Modeler can specify any additional group-sparsity structure. We denote the m^{th} set of group-sparse variables as $\mathcal{GS}_m$.

Preprocessing, continued

- Modeler can specify a set of variables for nonlinear transformation. Default transformations: $x^2, x^{1/2}$, and $\log x$.
- $\mathcal{T}_m$ is the m^{th} set of transformed variables.
- k_{max}, the maximum possible subset size such that the selective sparsity constraints are still feasible

$$
\begin{aligned}
k_{max} = \max_{\mathbf{z}} \quad & \sum_{i=1}^{p} z_i \\
s.t. \quad & z_i + z_j \leq 1 \quad \forall (i,j) \in \mathcal{HC} \\
& z_i \in \{0,1\}, i = 1, \ldots, p.
\end{aligned}
$$

The initial MIO model: For $k := 1, \ldots, k_{max}$

$$\min_{\beta, z} \quad \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \Gamma\|\beta\|_1 \qquad \textit{Robustness}$$

$$s.t. \quad z_l \in \{0,1\}, \quad l = 1, \ldots, p$$

$$-\mathcal{M} z_l \leq \beta_l \leq \mathcal{M} z_l, \quad l = 1, \ldots, p$$

$$\sum_{l=1}^{p} z_l \leq k \qquad \textit{Sparsity}$$

$$z_1 = \ldots = z_l \quad \forall i = 1, \ldots, l \in \mathcal{GS}_m, \quad \forall m \qquad \textit{Group Sparsity}$$

$$z_i + z_j \leq 1 \quad \forall (i,j) \in \mathcal{HC} \qquad \textit{Pairwise Collinearity}$$

$$\sum_{i \in \mathcal{T}_m} z_i \leq 1 \quad \forall m \qquad \textit{Nonlinear Transform}$$

Controlling for Statistical Significance

- $$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

  with $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I)$

- $$\frac{\hat{\beta}_j - \beta_j}{\tilde{\sigma}\sqrt{(X^T X)^{-1}_{jj}}} \sim t_{n-p}$$

  with $\tilde{\sigma}^2 = \frac{\mathbf{Y}^T(I_n - X(X^T X)^{-1}X^T)\mathbf{Y}}{n-p}$.

Controlling for Statistical Significance

- So, if $z_j = 1$, then

$$\frac{|\hat{\beta}_j|}{\tilde{\sigma}\sqrt{(X^T X)^{-1}_{jj}}} \geq t_{sign}$$

$$t_{sign} = T^{-1}\left(\frac{1-\alpha}{2}\right).$$

- Add the constraint for $b_j\{0,1\}$:

$$\frac{\beta_j}{\tilde{\sigma}\sqrt{(X^T X)^{-1}_{jj}}} + Mb_j \geq t_{sign} z_j$$

$$-\frac{\beta_j}{\tilde{\sigma}\sqrt{(X^T X)^{-1}_{jj}}} + M(1-b_j) \geq t_{sign} z_j$$

Controlling Global Multicollinearity

- A set of variables $\mathbf{X}_1, \cdots \mathbf{X}_p$ has an ϵ-*multicollinear relationship* if for some $\mathbf{a} \in \Re^p$, $\|\mathbf{a}\| = 1$, we have that:

$$\left\| \sum_{j=1}^{n} a_i \mathbf{X}_j \right\| < \epsilon.$$

- We show that existence of an ϵ-multicollinear relationship implies existence of an eigenvector $\mathbf{v}$ for the matrix $\mathbf{X}^T\mathbf{X}$ that has a small ($O(\sqrt{\epsilon})$) eigenvalue.

Key Result

- $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ be the set of orthonormal eigenvectors of $\mathbf{M} = \mathbf{X}^T\mathbf{X} \in \Re^{p \times p}$ such that the eigenvalues associated with V are less than ϵ. Then for $\mathbf{a} \in \Re^p$, $\|\mathbf{a}\| = 1$:
- If $\left\|\sum_{j=1}^{p} a_j \mathbf{X}_j\right\| < \epsilon$, then there exists a vector $\mathbf{b} \in \Re^p$, $\|\mathbf{b}\| < (p-m)\sqrt{\epsilon}$ such that $\mathbf{a} - \mathbf{b} \in Span(V)$.
- If there exists a vector $\mathbf{b} \in \Re^p$, $\|\mathbf{b}\| < \sqrt{\epsilon}$ such that $\mathbf{a} - \mathbf{b} \in Span(V)$, then we have:

$$\left\|\sum_{j=1}^{p} a_j \mathbf{X}_j\right\| < \sqrt{(1 + \lambda_{m+1} + \ldots + \lambda_p)\epsilon},$$

where $\lambda_{m+1}, \ldots, \lambda_p$ are the eigenvalues associated with the set of orthonormal eigenvectors of M that have value greater or equal to ϵ.

Some Observations

- For $\dim(V) = r$, we have $r - 1$ linearly independent multicollinear relationships.
- There are infinite number of ways the basis of the $r - 1$ multicollinear relationships could be constructed,
- Suppose we know that $x_1 + x_2 = x_3$ and $x_4 + x_5 = x_6$.
- Letting $\mathbf{a}_1 = (1, 1, -1, 0, 0, 0)^T$ and $\mathbf{a}_2 = (0, 0, 0, 1, 1, -1)^T$, we have $V = Span(\mathbf{a}_1, \mathbf{a}_2)$. Using previous theorem and ignoring $\mathbf{b}$ as $\|\mathbf{b}\| = O(\sqrt{\epsilon})$, we identify the two multicollinear relationships as $\mathbf{a}_1$ and $\mathbf{a}_2$.
- Add $z_1 + z_2 + z_3 \leq 2, \qquad z_4 + z_5 + z_6 \leq 2.$
- Letting $\bar{\mathbf{a}}_1 = (1, 1, -1, 1, 1, -1)^T$ and $\bar{\mathbf{a}}_2 = (1, 1, -1, -1, -1, 1)^T$, then V is also $V = Span(\bar{\mathbf{a}}_1, \bar{\mathbf{a}}_2)$.
- Add $z_1 + z_2 + z_3 + z_4 + z_5 + z_6 \leq 4$.

Minimum Support

Find a vector $\mathbf{a} \in Span(V)$ of minimum support.

$$\begin{aligned}
\min \quad & \sum_{j=1}^{m} z_j \\
\text{subject to} \quad & \mathbf{a} = \sum_{i=1}^{m} \theta_i \mathbf{v}_i \\
& |a_j| \leq M \cdot z_j, \quad j = 1, \ldots, m \\
& \left| \sum_{i=1}^{m} \theta_i \right| \geq \delta \\
& z_j \in \{0, 1\}, \quad j = 1, \ldots, m,
\end{aligned}$$

$\delta > 0$. Once $\mathbf{a}$ is identified, we add

$$\sum_{i \in Supp(\mathbf{a})} z_i \leq |Supp(\mathbf{a})| - 1.$$

Holistic Regression

$$\min_{\beta,\mathbf{z}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|^2 + \Gamma\|\beta\|_1$$

subject to:

$$z_i, b_i \in \{0,1\} \qquad i = 1, \cdots, p$$

$$-Mz_i \leq \beta_i \leq Mz_i, \qquad \text{Big-M constraint } (i = 1, \cdots, p)$$

$$\sum_{i=1}^{p} z_i \leq k \qquad \text{Sparsity}$$

$$z_i = z_l \qquad \text{Group Sparsity } (\forall i, j \in GS_m \; \forall m)$$

$$z_i + z_j \leq 1 \qquad \text{Pairwise Collinearity } (\forall i, j \in HC)$$

$$\sum_{i \in \mathcal{T}_j} z_i \leq 1, \quad j = 1, \ldots, p \text{Nonlinear transformations}$$

HolisticRegression

$$\frac{\beta_j}{\tilde{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}} + Mb_j \geq t_\alpha z_j \qquad \text{Significance } (j = 1, \cdots, p)$$

$$-\frac{\beta_j}{\tilde{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}} + M(1 - b_j) \geq t_\alpha z_j \qquad \text{Significance } (j = 1, \cdots, p)$$

$$\sum_{i \in Supp(\mathbf{a})} z_i \leq |Supp(\mathbf{a})| - 1 \qquad \forall \text{ multicollinear relations } \mathbf{a}$$

identified by MC-detection Algorithm.

Contrast with existing practice

- All desired properties are simultaneously enforced.
- MIO does not have to choose which model properties to favor by performing the steps in a certain order.
- MIO is capable of handling datasets with more variables than a modeler can address manually.

Example 1: Croq'Pain Stores

Variable	Description
EARN	Operating earnings in $1000s
SIZE	Total area inside store
EMPL	# of employees as of Dec. 31, 1994
P15	# of 15-24 year olds in a 3 km radius
P25	# of 25-34 year olds in a 3 km radius
P35	# of 35-44 year olds in a 3 km radius
P45	# of 45-54 year olds in a 3 km radius
P55	# of people age 55+ in a 3 km radius

Example 1: Croq'Pain Stores

Variable	Description
TOTAL	Total population in a 3 km radius
INC	Average income in town/neighborhood surrounding site
COMP	# competitors within 1 km
NCOMP	# restaurants that do not compete with Croq'Pain within 1 km
NREST	# non-restaurant businesses in a 1 km radius
PRICE	Monthly rent per square meter of retail properties in the same locale
CLI	Cost of Living Index
K	Invested capital

The manual approach

- The model with all 14 independent variables has an $R^2 = 0.867$.
- Five of the fourteen variables are significant at the 0.05 level.
- Coefficients for number of employees and for total surrounding population are both negative.
- P35 and TOTAL have a correlation coefficient of 0.96.
- Remove variables one at a time until all variables left are significant.
- Remaining variables: (SIZE, P15, INC, NREST, and PRICE) and training set $R^2 = 0.856$. Correlation are less than 0.18.

Holistic Regression

- MIO using the default settings: 0.8 as the maximum pairwise correlation and 10 potential values of Γ.
- 1 minute to run, and returns a model with the same 5 independent variables: SIZE, P15, INC, NREST, and PRICE.
- The first 4 variables are significant at the 0.001 level, the last at the 0.01 level.
- The model has an out-of-sample validation set $R^2 = 0.80$.

Example 2: Ames Housing-Manual Approach

- Ames Housing Dataset on property sales in Ames, Iowa.
- The variables describe the quality and quantity of physical attributes of each property sold.
- 2271 observations and 315 variables.
- Using only the categorical variable for neighborhood and property's total square footage $R^2 = 0.8$.
- After a day of trying, a 36-variable model (using new variables through transformation and interactions): $R^2 = 0.92$.

Example 2: Ames Housing-MIO approach

- MIO with default settings: $\rho = 0.8$, 10 potential values of Γ.
- A 20 variables (significant at the 0.05 level), and had a test set $R^2 = 0.920$.
- MIO variables:
  - the overall quality and condition of the property,
  - whether the property is identified as being in a particular neighborhood,
  - the number of half bathrooms,
  - whether the foundation of the home constructed from stone or not,
  - the year the garage was built,
  - various measurements of square footage,
  - and the type of electrical system.

Real World Examples

Dataset	**n**	**p**	**HR k***	**R^2**	**Cor**	**Lasso k***	**R^2**	**Cor**
CPU	105	6	5	0.869	0.716	6	0.861	0.716
Yacht	154	6	1	0.600	NA*	1	0.602	NA*
White Quality	2499	11	10	0.270	0.619	9	0.280	0.828
Red Quality	800	11	6	0.384	0.40	7	0.386	0.69
Compact	4096	21	15	0.717	0.733	21	0.725	0.942
Elevator	8280	18	10	0.808	0.678	15	0.809	0.999
Pyrimidines	37	26	15	0.175	0.781	20	0.367	0.928
LPGA 2008	78	6	2	0.877	0.02	3	0.873	0.234
LPGA 2009	73	11	7	0.814	0.784	10	0.807	0.943
Airline Costs	15	9	2	0.672	0.501	9	0.390	0.973
Diabetes	221	64	4	0.334	0.423	14	0.381	0.672
HIV	528	98	11	0.945	0.662	39	0.944	0.760

Sparsity on Synthetic Data, $n = 500$, $p = 100$, $\rho = 0$

SNR	HR k^*	TP	R^2	Cor	Cond	Time
6.32	10.6	10	0.716	0.119	1.61	0.448
	0.358	0	0.007	0.007	0.02	0.011
3.16	10.6	10	0.909	0.119	1.27	0.439
	0.358	0	0.003	0.007	0.29	0.011
1.58	10	10	0.975	0.117	1.58	0.304
	0	0	0.001	0.007	0.04	0.011
SNR	**Lasso k^***	**TP**	**R^2**	**Cor**	**Cond**	
6.32	34.8	10	0.701	0.148	2.782	
	3.51	0	0.007	0.010	0.127	
3.16	34.4	10	0.904	0.148	2.805	
	3.72	0	0.002	0.010	0.146	
1.58	34.6	10	0.974	0.160	2.797	
	4.40	0	0.001	0.016	0.194	

Pairwise MC , $n = 500$, $p = 100$, True $k = 10$, $\rho = 0.9$

SNR	HR k^*	TP	R^2	Cor	Cond	Time
8.73	10.00	10.00	0.99	0.40	4.15	0.30
	0.00	0.00	0.00	0.01	0.17	0.02
4.37	10.40	10.00	0.95	0.47	5.65	0.34
	0.36	0.00	0.00	0.07	1.25	0.04
2.18	11.40	9.60	0.81	0.63	7.92	0.63
	0.54	0.22	0.01	0.08	2.25	0.15
SNR	**Lasso k^***	**TP**	**R^2**	**Cor**	**Cond**	
8.73	34.40	10.00	0.99	0.91	126.28	
	2.65	0.00	0.00	0.00	13.15	
4.37	37.20	10.00	0.94	0.91	146.36	
	3.66	0.00	0.00	0.00	20.83	
2.18	36.60	10.00	0.81	0.91	142.17	
	3.37	0.00	0.01	0.00	18.89	

Robust: $n = 500$, $p = 100$, $k = 10$, $\rho = 0$, $\mathbf{\Delta X} \sim$ U(0,2)

SNR	HR k*	TP	R^2	Cor	Cond	Time
6.32	10 0.000	10 0	0.975 0.001	0.117 0.007	1.58 0.04	0.448 0.011
3.16	10.6 0.358	10 0	0.909 0.003	0.119 0.007	1.27 0.29	0.439 0.011
1.58	10.6 0.358	10 0	0.716 0.007	0.119 0.007	1.61 0.02	0.304 0.011
SNR	**Lasso k***	**TP**	**R^2**	**Cor**	**Cond**	
6.32	34.6 4.40	10 0	0.974 0.001	0.160 0.016	2.797 0.194	
3.16	34.4 3.72	10 0	0.904 0.002	0.148 0.010	2.805 0.146	
1.58	34.8 3.51	10 0	0.701 0.007	0.148 0.010	2.782 0.127	

Significance

- **Accuracy** - the percentage of non-zero β's in the ground truth we identify correctly:

$$ACC = \frac{Supp(\beta) \cap Supp(\tilde{\beta})}{Supp(\beta)}.$$

- **False Positive Rate** - the percentage of non-zero β's recovered that are zero in the ground truth. It is defined as:

$$FPR = \frac{Supp(\tilde{\beta}) \backslash Supp(\beta)}{Supp(\tilde{\beta})}.$$

- **Time** - the time the algorithm takes to converge to a final solution.

Significance for Synthetic Data for $\alpha = 0.05$

n	Error	p	k	ACC	FPR	Time
100	$N(0, 0.1)$	10	3	100%	7.4%	0.65s
200	$N(0, 0.1)$	10	3	100%	4.5%	2.9s
500	$N(0, 0.1)$	10	3	100%	0%	6.67 s
1000	$N(0, 0.1)$	10	3	100%	0%	36.4s
500	$N(0, 1)$	10	3	100%	26%	7.93s
1000	$N(0, 1)$	10	3	100%	0%	38.9s
500	t_5	10	3	95%	35%	6.75s
1000	t_5	10	3	100%	21%	35.2s

Significance for Real-World Data for $\alpha = 0.05$

Dataset	**p**	k^*	**Loss**	**Time**
NCAA	11	6	39.9	17.3s
Pollution	6	4	249.8	2.74s
Diabetes	6	2	1553.2	9.4s
Baseball	6	4	6397.3	27.7s
Pyrimidine	11	7	0.5970	2.3s

Effects of Nonlinear Transformations and Group Sparsity

- $n = 1030,\ p = 8.$

Variable	**Units**
Concrete Compressive Strength	MPa
Cement	kg/m^3
Blast Furnace Slag	kg/m^3
Fly Ash	kg/m^3
Water	kg/m^3
Superplasticizer	kg/m^3
Coarse Aggregate	kg/m^3
Fine Aggregate	kg/m^3
Age	day

- Design a concrete mixture which will have high compressive strength. Known to be a highly nonlinear function of its age and ingredients.

Concrete dataset: Nonlinear Transformations

- MIO and Lasso chose all covariates, $R^2 = 0.609$.
- Used x^2, $\sqrt{x}$, and $\log(x)$ but at most one of them is used.
- MIO chose 6 covariates: blast furnace slag, water, log(fly ash), log(super plasticizer), log(day), and cement$^{1/2}$.
- Significant at the 0.001 level and test set $R^2 = 0.823$.
- Lasso selected a model with twelve covariates which resulted in a test set $R^2 = 0.834$.
- Variables went from 8 to 32. MIO took 1 minute.

Energy Efficiency dataset : Group Sparsity

- $n = 768$, six continuous independent variables and two categorical independent variables.

Variable	**Type**
Relative Compactness	Continuous
Surface Area	Continuous
Wall Area	Continuous
Roof Area	Continuous
Overall Height	Continuous
Orientation	Categorical; 3 levels
Glazing Area	Continuous
Glazing Area Distribution	Categorical; 5 levels

- Two dependent variables available: heating load and cooling load. 14 independent variables.

Energy Efficiency dataset : Results

- Manual approach : all variables selected.
- MIO: wall area, overall height, and glazing area both for heating and cooling load.
- The heating load model had test set $R^2 = 0.88$ and the cooling load model had test set $R^2 = 0.85$.
- Group Lasso selected all variables except surface area.
- The heating load model had a test set $R^2 = 0.91$ and the cooling load model had a test set $R^2 = 0.86$.

Multicollinearity detection

n	p	3	4	4+	Noise	ACC	FPR	Time
1000	100	3	1	1	$N(0, 0.01)$	100%	0%	0.27s
1000	500	3	1	1	$N(0, 0.01)$	100%	0%	2.37s
1000	1000	3	1	1	$N(0, 0.01)$	100%	5%	520.23s
1000	500	5	3	2	$N(0, 0.01)$	100%	0%	33.40s
1000	1000	5	3	2	$N(0, 0.01)$	100%	24%	5940.56s
1000	500	3	1	1	$N(0, 0.03)$	100%	0%	2.29s
1000	1000	3	1	1	$N(0, 0.03)$	100%	11%	32.17s

Combined Example

- Synthetic example, $n = 100$ and $p = 500$, $\rho = 0.8$.
- For each x, we also included $\sqrt{x}$, $p = 1000$.
- $\beta_i = 1$ for 7 variables in the original 500 and 3 in the 500 transformed ones.
- $\mathbf{y}$ was generated as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.
- $\Delta\mathbf{X} \sim \text{Unif}(0, f)$ and considered $\mathbf{X} + \Delta\mathbf{X}$ for various values of f.
- Modeler knows one of the values of i such that β_i is truly nonzero.
- Modeler knows $\beta_a, \beta_b, \beta_c$, and β_d are all either all zero or all nonzero and that $\beta_e, \beta_f, \beta_g$, and β_h are either all zero or all nonzero, where $\{a, b, c, d\} \in \{i | \beta_i = 1\}$ and $\{e, f, g, h\} \in \{i | \beta_i = 0\}$.

Results

ϵ	ΔX	MIO Γ^*	K^*	TP	R^2	Max Cor	Cond	T	Las K^*	TP	R^2	Max Cor	Cond
0.5	0	0.026	10.4	10	0.981	0.437	4.654	1.14	46.6	10	0.969	0.836	118.1
		0.020	0.22	0	0.001	0.020	0.382	0.17	4.15	0	0.004	0.007	17.4
0.5	1	0.000	11.2	10	0.913	0.556	6.995	1.34	65.8	10	0.854	0.798	424.0
		0.000	0.52	0	0.013	0.073	1.422	0.22	6.97	0	0.017	0.006	177.5
0.5	2	0.030	11.0	9	0.742	0.501	5.291	1.88	69	9.2	0.598	0.708	8147.1
		0.027	0.49	0.3	0.030	0.061	0.508	0.42	8.54	0.2	0.045	0.006	6993.5
1	0	0.026	11.2	10	0.931	0.468	5.322	1.08	45.6	10	0.878	0.836	113.9
		0.022	0.52	0	0.007	0.018	0.531	0.04	3.99	0	0.016	0.007	18.2
1	1	0.041	10.4	10	0.878	0.478	4.998	1.61	69.2	10	0.759	0.796	362.8
		0.036	0.22	0	0.016	0.059	0.696	0.43	4.92	0	0.033	0.006	96.4
1	2	0.099	9.8	7.6	0.573	0.436	4.224	1.64	72.4	8.6	0.503	0.702	573.8
		0.041	0.87	0.2	0.042	0.061	0.576	0.42	5.89	0.4	0.064	0.006	228.0
2	0	0.090	10	8.8	0.720	0.451	4.687	1.35	39.4	8.6	0.599	0.836	74.0
		0.045	0.28	0.3	0.046	0.025	0.289	0.23	4.01	0.5	0.055	0.007	9.76
2	1	0.116	9.4	8.2	0.614	0.426	4.262	2.05	53.4	7.8	0.509	0.782	113.5
		0.037	0.36	0.6	0.078	0.025	0.137	0.39	4.41	0.9	0.067	0.010	20.7
2	2	0.032	8.2	4	0.245	0.403	3.506	1.46	55.8	5.8	0.368	0.680	8141.4
		0.017	0.66	1	0.129	0.074	0.720	0.22	10.1	0.5	0.061	0.011	7236.0

Conclusions

- Holistic Regression : MIO imposes all statistical properties simultaneously and systematically.
- Method generally applicable. Saves substantial modeler's time.
- Scalable, fast and reliable.