Penalized GLM Guide

Tomas Nykodym
Patrick Hall

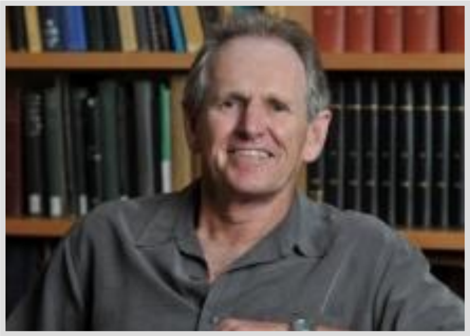
H2O.ai

Linear Modeling Methods

Ordinary Least Squares

Carl Friedrich Gauss
(1777–1855)

Elastic Net



Hui Zou and Trevor Hastie
Regularization and variable selection via the elastic net,
Journal of the Royal Statistical Society, 2005

Ordinary Least Squares Requirements

Requirements	If broken …
Linear relationship between inputs and targets; normal y, normal errors	Inappropriate application/unreliable results; use a machine learning technique; use GLM
N > **p**	Underspecified/unreliable results; use LASSO or elastic net penalized regression
No strong multicollinearity	Ill-conditioned/unstable/unreliable results; Use ridge(L2/Tikhonov)/elastic net penalized regression
No influential outliers	Biased predictions, parameters, and statistical tests; use robust methods, i.e. IRLS, Huber loss, investigate/remove outliers
Constant variance/no heteroskedasticity	Lessened predictive accuracy, invalidates statistical tests; use GLM in some cases
Limited correlation between input rows (no autocorrelation)	Invalidates statistical tests; use time-series methods or machine learning technique

Anatomy of GLM

Family/distribution defines mean and variance of Y

Nonlinear link function between linear component and E(Y)

$$\boxed{E(\boldsymbol{Y})} = \mu = \boxed{g^{-1}}(\boxed{\boldsymbol{X}\beta})$$

Linear component

$$\boxed{Var(\boldsymbol{Y}) = V(\mu) = V(g^{-1}(\boldsymbol{X}\beta))}$$

Family/distribution allows for non-constant variance

Distributions / Loss Functions

For **regression** problems, there's a large choice of different distributions and related loss functions:

- **Gaussian** distribution, squared error loss, sensitive to outliers
- **Laplace** distribution, absolute error loss, more robust to outliers
- **Huber** loss, hybrid of squared error & absolute error, robust to outliers
- **Poisson** distribution (e.g., number of claims in a time period)
- **Gamma** distribution (e.g, size of insurance claims)
- **Tweedie** distribution (compound Poisson-Gamma)

- **Binomial** distribution, log-loss for binary classification

Also, H2O supports:
- **Offsets**
- **Observation weights**

Iteratively Reweighed Least Squares

Iteratively reweighted least squares (IRLS) complements model fitting methods in the presence of outliers by:

- Initially setting all observations to an equal weight
  - Fitting GLM parameters (β's)

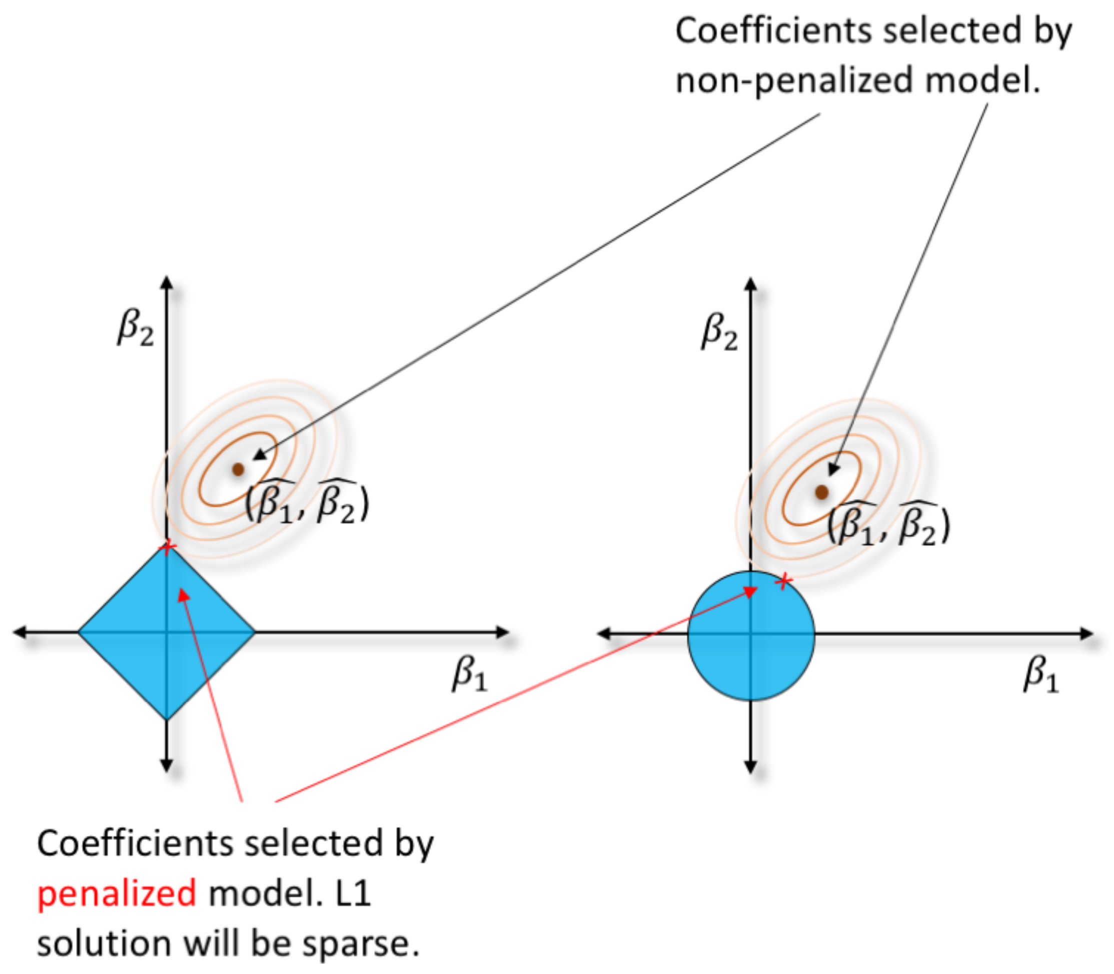
"Outer loop"

$$\tilde{\beta} = \min_{\beta}\left\{\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} * \beta_j\right)^2\right\}$$

"Inner loop"
(ADMM optimization in H2O)

  - Calculating the residuals of the fitted GLM
  - Assigning observations with high residuals a lower weight
  - Repeating until GLM parameters (β's) converge

Regularization (e.g. Penalties)



Combining GLM, IRLS and Regularization

λ controls magnitude of penalties. Variable selection conducted by refitting model many times while varying λ. Decreasing λ allows more variables in the model.

"Outermost loop"

L1/LASSO helps with variable selection.

L2/Ridge/Tinkhonov penalty helps address multicollinearity.

$$\tilde{\beta} = \min_{\beta} \left\{ \boxed{\sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} * \beta_j \right)^2} + \lambda \sum_{j=1}^{p} \left(\alpha * |\beta_j| + (1-\alpha) * \beta_j^2 \right) \right\}$$

Error function for a GLM.

α tunes balance between L1 and L2 penalties, i.e. elastic net.

- Inner loop: Fitting GLM parameters for a given λ and α
- Outer loop: IRLS until β's converge
- Outermost loop: λ varies from λ_{max} to 0

Elastic net advantages over L1 or L2:

- Does not saturate at min(p, N)
- Allows groups of correlated variables

H2O.ai

Hyperparameter Selection

Outer most loop(s):

- λ search from λ_{max} (where all coefficients = 0) to $\lambda = 0$
- Grid search on alpha usually not necessary
  - Just try a few: 0, 0.5, 0.95
  - Always keep some L2
  - Set max_predictors, large models take longer
- Models can also be validated:
  - Validation and test partitioning available
  - Cross-validated (k-fold, CV predictions available)

Practical Pointers

- P-values available for non-penalized models
- β constraints available, i.e. for all positive β's
- Use IRLS optimization for tall, skinny data sets
- L1 **OR** LBFGS for wide data (> 500 predictors)
  - (L1 **AND** LBFGS possible, but can be slower)

H2O.ai