Variable Selection

version: 2018-01-26 · 11:39:52

Motivation

- **Linear Regression Model:** $y_i \mid \mathbf{x}_i \overset{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2)$
  - n observations $\mathbf{y} = (y_1, \dots, y_n)$
  - p covariates $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$
- **Maximum Likelihood:** (assume $\sigma = 1$)
  - *Likelihood:* $\ell(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) = -\frac{1}{2}\sum_{i=1}^n (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2$.
  - For $p \le n$: $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. $\mathbf{X}'\mathbf{X}$ is almost surely invertible when $p \le n$.
  - For $p > n$: $\mathbf{X}'\mathbf{X}$ is singular $\implies$ infinitely many likelihood-maximizing values:

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\alpha}} + \operatorname{span}\{\hat{\boldsymbol{\gamma}}_1, \dots, \hat{\boldsymbol{\gamma}}_{p-n}\} \qquad \implies \qquad \sup \|\hat{\boldsymbol{\beta}}\| = \infty.$$

- **Application of p > n**:
  - y_i: level of cancer-associated antigen in subject i.
  - x_{ij}: expression level for gene j.
  - Typical data: $n \sim 100$ - 1K, $p \sim$ 1M.

Penalized Likelihood

- **Model:** $y_i \mid \mathbf{x}_i \overset{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}_i'\boldsymbol{\beta}, 1) \quad \Longrightarrow \quad \ell(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) = -\frac{1}{2}\sum_{i=1}^n (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2.$
- **Unconstrained MLE**: $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}), \quad S(\boldsymbol{\beta}) = \sum_{i=1}^n (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2.$
  - For $p \le n$: $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$
  - For $p > n$: $\hat{\boldsymbol{\alpha}} + \operatorname{span}\{\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_{p-n}\}.$
- **Penalized Likelihood:**

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 \quad \text{subject to} \quad \rho(\boldsymbol{\beta}) < t,$$

  for some penalty function $\rho(\boldsymbol{\beta}) \ge 0$.
  - For $p > n$: Typically $\rho(\boldsymbol{\beta}) < t \implies \|\boldsymbol{\beta}\| < C \implies \tilde{\boldsymbol{\beta}}$ is *likely* to be unique.
  - For $p < n$: $\operatorname{Bias}(\hat{\boldsymbol{\beta}}) = E[\hat{\boldsymbol{\beta}}] - \boldsymbol{\beta}_{\text{true}} = 0$, but $\operatorname{Bias}(\tilde{\boldsymbol{\beta}}) \ne 0$. However, for $p \approx n$ $\operatorname{var}(\tilde{\boldsymbol{\beta}}) \ll \operatorname{var}(\hat{\boldsymbol{\beta}})$, such that PL can have smaller mean squared error (MSE):

$$\operatorname{MSE}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_{\text{true}}) \overset{\text{def}}{=} \sum_{i=1}^p E[(\hat{\theta}_i - \theta_{i,\text{true}})^2] = \sum_{i=1}^p \operatorname{Bias}(\hat{\theta}_i)^2 + \operatorname{var}(\hat{\theta}_i).$$

Ridge Regression

- **Model:** $y_i \mid \mathbf{x}_i \overset{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}_i'\boldsymbol{\beta}, 1) \implies \ell(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) = -\frac{1}{2}\sum_{i=1}^n (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2$.
- **Unconstrained MLE:** $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2$.
- **$\mathbf{L_2}$-Constraint:**

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 \quad \text{subject to} \quad \sum_{j=1}^p \beta_j^2 \le t.$$

Note: the constraint assumes equally weighted β_j. This is because the data are assumed to have been standardized, i.e., if $\mathbf{X}_j = (x_{1j}, \ldots, x_{nj})$ is all observations of covariate j, then

$$x_{ij} \leftarrow \frac{x_{ij} - \text{mean}(\mathbf{X}_j)}{\text{sd}(\mathbf{X}_j)},$$

i.e., $\mathbf{X}_j$ has mean 0 and variance 1. Thus, β_j is the change in $E[y \mid \mathbf{x}]$ per standard deviation of $\mathbf{X}_j$, with all other covariates being fixed. (For linear regression, it is also common to get rid of the intercept β_0 by setting $y_i \leftarrow y_i - \bar{\mathbf{y}}$.)

Ridge Regression

- **Model:** $y_i \mid \mathbf{x}_i \overset{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}_i'\boldsymbol{\beta}, 1)$.
- **Penalized Likelihood:** Let $S(\boldsymbol{\beta}) = \sum_{i=1}^n (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2$ and $\rho(\boldsymbol{\beta}) = \sum_{j=1}^p \beta_j^2$.

  Constrained minimization

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) \quad \text{subject to} \quad \rho(\boldsymbol{\beta}) \le t. \tag{1}$$

- **Unconstrained Formulation:**

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) + \lambda \cdot \rho(\boldsymbol{\beta}). \tag{2}$$

  Proof: Let $\tilde{\boldsymbol{\beta}}$ be the solution to (2) and $t = \rho(\tilde{\boldsymbol{\beta}})$.

  Then for $\boldsymbol{\beta}$ subject to $\rho(\boldsymbol{\beta}) \le t = \rho(\tilde{\boldsymbol{\beta}})$,

$$S(\tilde{\boldsymbol{\beta}}) + \lambda \cdot \rho(\tilde{\boldsymbol{\beta}}) \le S(\boldsymbol{\beta}) + \lambda \cdot \rho(\boldsymbol{\beta}) \le S(\boldsymbol{\beta}) + \lambda \cdot \rho(\tilde{\boldsymbol{\beta}})$$
$$\implies S(\tilde{\boldsymbol{\beta}}) \le S(\boldsymbol{\beta}).$$

Ridge Regression

- **Model:** $y_i \mid \mathbf{x}_i \overset{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}_i'\boldsymbol{\beta}, 1)$.
- **Parameter Estimation:**
  - Penalized likelihood with L_2 penality: $\tilde{\boldsymbol{\beta}} = \arg\max \ell(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) - \lambda \sum_{j=1}^{p} \beta_j^2$
  - *Unconstrained formulation:*

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

  - *Solution:* PL is quadratic in $\boldsymbol{\beta}$, so use complete-the-squares to obtain

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \lambda^2 \mathbf{I}_p)^{-1} \mathbf{X}'\mathbf{y}.$$

- **Questions:**

  1. How to pick λ?

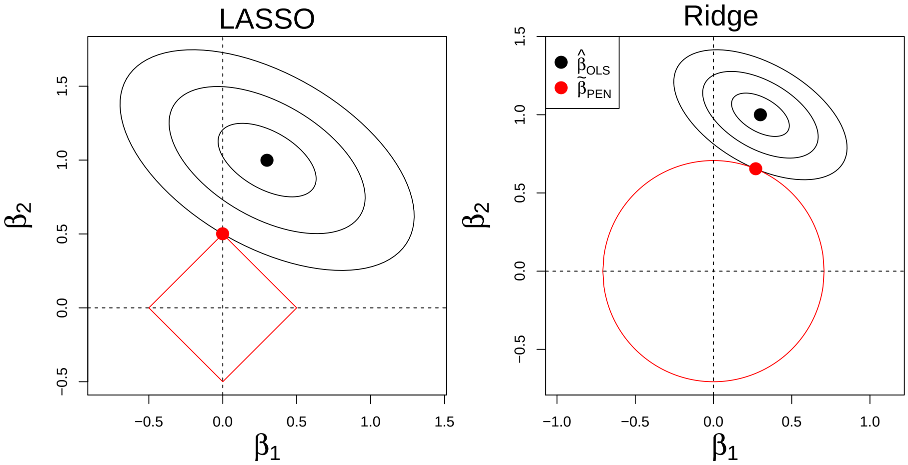
  2. How to make PL-based confidence intervals for $\boldsymbol{\beta}$?

Lasso Regression

- **Model:** $y_i \mid \mathbf{x}_i \overset{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}_i'\boldsymbol{\beta}, 1)$.
- **Unconstrained MLE**: $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2$.
- **$\mathbf{L_1}$ Constraint:**

$$
\begin{aligned}
\tilde{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 \quad \text{subject to} \quad \sum_{j=1}^p |\beta_j| \le t \\
&= \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^p |\beta_j|.
\end{aligned}
$$

Lasso Regression



Advantage of Lasso over Ridge: Variable selection, i.e., some of the $\tilde{\beta}_j$ in Lasso can equal 0. The black contours correspond to the shape of $S(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2$. The corners of the contraint region allow Lasso to perform variable selection.

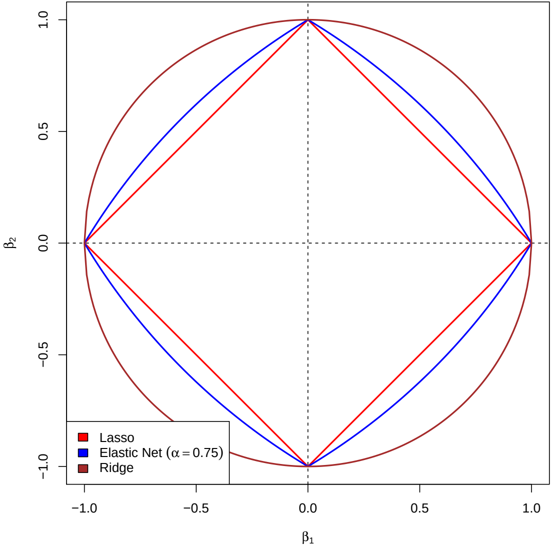
Elastic Net

- **Model:** $y_i \mid \mathbf{x}_i \overset{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}_i'\boldsymbol{\beta}, 1)$.
- **Penalized Likelihood:** $\tilde{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) - \lambda\rho(\boldsymbol{\beta})$.
  - *Ridge Regression:* $\rho(\boldsymbol{\beta}) = \sum_{j=1}^{p} \beta_j^2$.
  - *Lasso Regression:* $\rho(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|$.
- **Advantage of Lasso over Ridge:** Variable selection, i.e., some of the $\tilde{\beta}_j$ in Lasso can equal 0.
- **Advantage of Ridge over Lasso:** Better performance when covariates are highly correlated. (To see this, run demo posted online)
- **Elastic Net Regression:** Compromise between L_1 and L_2 constraints:

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda \cdot \rho_\alpha(\boldsymbol{\beta}),$$

  where $\rho_\alpha(\boldsymbol{\beta}) = \sum_{j=1}^{p} (1-\alpha)\beta_j^2 + \alpha|\beta_j|$ for $\alpha \in [0, 1]$.

Elastic Net Regression



Constraint shapes for different penalty functions.

Elastic Net Regression

- **Model:** $y_i \mid \mathbf{x}_i \overset{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}_i'\boldsymbol{\beta}, 1)$.
- **Parameter Estimation:**

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \underbrace{\sum_{i=1}^{n} (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}_{S(\boldsymbol{\beta})} + \lambda \cdot \underbrace{\sum_{j=1}^{p} (1-\alpha)\beta_j^2 + \alpha|\beta_j|}_{\rho_\alpha(\boldsymbol{\beta})}.$$

- Lots of methods of solution, but the simplest and most effective for $p \gg n$ (and fixed λ, α) is **Coordinate Descent:**
  - Minimize $\Omega(\boldsymbol{\beta}) = S(\boldsymbol{\beta}) + \lambda\rho_\alpha(\boldsymbol{\beta})$ one β_j at a time holding the others fixed.
  - Continue cycling throught β_j's until relative tolerance is reached.

Elastic Net Regression

Coordinate Descent

- Minimize $\Omega(\beta_1, \boldsymbol{\beta}^\star)$ as a function of β_1 with $\boldsymbol{\beta}^\star = (\beta_2, \ldots, \beta_p)$ fixed:

$$
\begin{aligned}
\Omega(\beta_1, \boldsymbol{\beta}^\star) &= \sum_{i=1}^{n} (y_i - x_{1i}\beta_1 - \mathbf{x}_i^{\star\prime}\boldsymbol{\beta}^\star)^2 + \lambda\alpha|\beta_1| + \lambda(1-\alpha)\beta_1^2 \\
&\quad + \lambda \sum_{j=2}^{p} (1-\alpha)\beta_j^2 + \alpha|\beta_j| \\
&= A\beta_1^2 + B\beta_1 + \lambda\alpha|\beta_1| + C
\end{aligned}
$$

$$
\Longrightarrow \qquad \hat{\beta}_1 = \frac{\mathcal{Q}(\frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^{\star\prime}\boldsymbol{\beta}^\star), 2n\lambda\alpha)}{1 + 2n\lambda(1-\alpha)}, \qquad \text{where}
$$

$$
\mathcal{Q}(z, w) = \operatorname{sgn}(z)(|z| - w)_+ = \begin{cases} z - w & z > 0, w < |z| \\ z + w & z < 0, w < |z| \\ 0 & w > z \end{cases}
$$

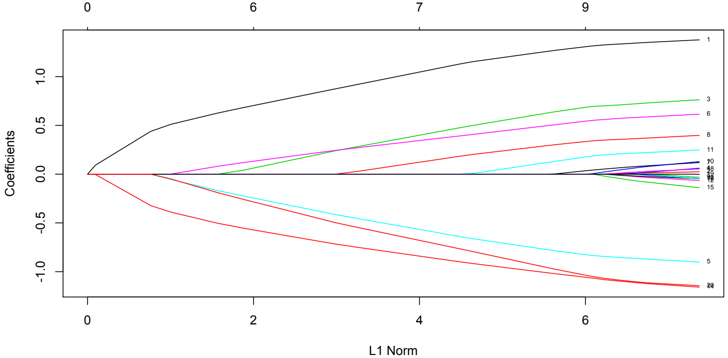
Least-Angle Regression (LARS)

- **Elastic Net:**

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda \cdot \sum_{j=1}^{p} (1-\alpha)\beta_j^2 + \alpha|\beta_j|.$$

- **Motivation:** Coordinate descent is best for fixed λ. But how to pick λ?

- **LARS:** Calculates $\tilde{\boldsymbol{\beta}}$ at *every* λ in only p steps.



Least-Angle Regression (LARS)

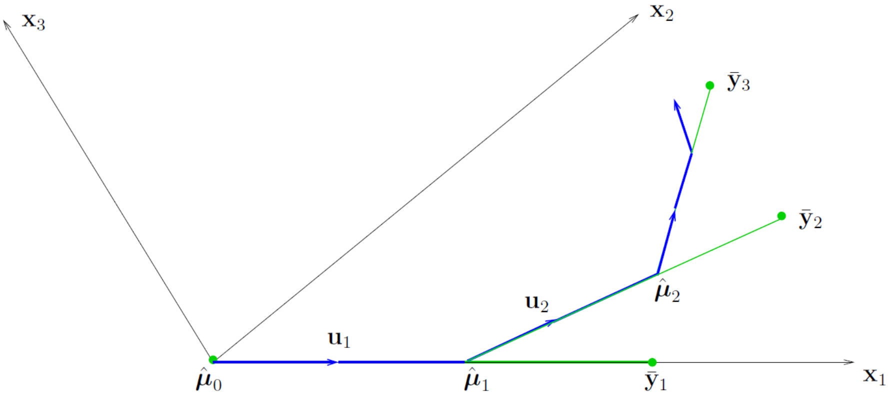
Basic Idea:

As explained by Robert Tibshirani here:

1. Start with all coefficients β_j equal to zero.
2. Find the predictor x_j most correlated with y. Increase the coefficient β_j in the direction of the sign of its correlation with y. Take residuals $r = y - \hat{y}$ along the way. Stop when some other predictor x_k has as much correlation with r as x_j has.
3. Increase (β_j, β_k) in their joint least squares direction, until some other predictor x_m has as much correlation with the residual r.
4. Continue until: all predictors are in the model.

Least-Angle Regression (LARS)

Illustration



From the original LARS paper by Efron et al (2004). The LARS estimator at step k is $\hat{\mu}_k$. Between step k and $k+1$, the estimator goes towards the OLS estimate $\bar{y}_{k+1}$.

Picking the Value of $\boldsymbol{\lambda}$

Objective 1: Minimize Prediction Error

- **Model:** $y_i \mid \mathbf{x}_i \overset{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2)$
- **Parameter Estimation:** Elastic net for fixed α as a function of λ:

$$\tilde{\boldsymbol{\beta}}_{\lambda} = \underset{\boldsymbol{\beta}}{\arg\min} \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda \cdot \sum_{j=1}^{p} (1-\alpha)\beta_j^2 + \alpha|\beta_j|.$$

- **Objective:** Suppose that $(y, \mathbf{x}) \sim f(y, \mathbf{x})$ have a joint distribution. Let $\tilde{\boldsymbol{\beta}}_{\lambda}^{\text{obs}}$ denote the elastic net estimator for the given dataset $(\mathbf{y}_{\text{obs}}, \mathbf{X}_{\text{obs}})$. Then we wish to minimize the mean square prediction error (MSPE)

$$\hat{\lambda} = \underset{\lambda}{\arg\min}\, \text{MSPE}(\lambda), \qquad \text{MSPE}(\lambda) = E[\{y - \mathbf{x}'\tilde{\boldsymbol{\beta}}_{\lambda}^{\text{obs}}\}^2].$$

Picking the Value of λ

Objective 1: Minimize Prediction Error

- **Model:** $y_i \mid \mathbf{x}_i \overset{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2)$
- **Parameter Estimation:** $\tilde{\boldsymbol{\beta}}_{\lambda} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda \cdot \sum_{j=1}^{p} (1-\alpha)\beta_j^2 + \alpha|\beta_j|.$
- **Prediction Error:** $\hat{\lambda} = \arg\min_{\lambda} \mathrm{MSPE}(\lambda), \quad \mathrm{MSPE}(\lambda) = E[\{y - \mathbf{x}'\tilde{\boldsymbol{\beta}}_{\lambda}^{\text{obs}}\}^2].$
- **MSPE Estimation:** Use cross-validation:
  - Separate data into training and test sets: $(\mathbf{y}_{\text{train}}, \mathbf{X}_{\text{train}})$ and $(\mathbf{y}_{\text{test}}, \mathbf{X}_{\text{test}})$.
  - *MSPE estimate:* $\widehat{\mathrm{MSPE}}(\lambda) = \sum_{i=1}^{n_{\text{test}}} \{y_i^{\text{test}} - \mathbf{x}_i^{\text{test}\prime}\tilde{\boldsymbol{\beta}}_{\lambda}^{\text{train}}\}^2,$

    where $\tilde{\boldsymbol{\beta}}_{\lambda}^{\text{train}}$ is calculated from $(\mathbf{y}_{\text{train}}, \mathbf{X}_{\text{train}})$.
  - *K-Fold CV:* (i) Randomly separate separate data into K sets (ii) MSPE estimate is

    $$\widehat{\mathrm{MSPE}}(\lambda) = \textstyle\sum_{k=1}^{K} \widehat{\mathrm{MSPE}}_k(\lambda),$$

    where $\widehat{\mathrm{MSPE}}_k(\lambda)$ has subset k as test set, and remaining data as training set.

Picking the value of λ

Objective 2: Don't miss any non-zero β_j's

- **Model:** $y_i \mid \mathbf{x}_i \overset{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}_i'\beta, \sigma^2)$
- **Covariance Test:** For each step k of LARS, can construct a test statistic T_{k+1} such that under

  $$H_0 : \text{All } k \text{ non-zero } \beta_j\text{'s have been identified,}$$

  as $n, p \to \infty$ (but $p < n$) we have

  $$T_{k+1} \mid H_0 \to \mathcal{F}(2, n-p).$$

- **In Practice:** Stop LARS at first value of $k-1$ such that

```
pval <- pf(q = Tk, df1 = 2, df2 = n-p, lower.tail = FALSE)
```

  is greater than 5%.

Resources

- `lars`: Package for LARS and Lasso.
- `covTest`: Implementation of covariance test. Paper by Lockhart et al (2014) can be found here.
- `glmnet`: Very efficient LARS-type elastic net calculation for many GLMs.