Statistical Data Analysis

Jana de Wiljes

wiljes@uni-potsdam.de

www.dewiljes-lab.com

7. November 2022

Universität Potsdam

Coefficient of determination

Lemma: The coefficient of determination R^2 can be transformed into

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{\hat{\beta}^\top \mathbf{X}^\top \mathbf{y} - n\bar{y}^2}{\mathbf{y}^\top \mathbf{y} - n\bar{y}^2} \tag{1}$$

Def: The corrected coefficient of determination $\bar{R}^2$ is defined by

$$\bar{R}^2 = 1 - \left(\frac{n-1}{n-p-1}\right)(1 - R^2) \tag{2}$$

Asymptotic Properties of the LS-Estimator

Proposition: Consider the setting

$$\mathbf{y}_n = \mathbf{X}_n\beta + \epsilon_n \quad \text{with } \mathbb{E}[\epsilon_n] = \mathbf{0} \quad \text{and } Cov(\epsilon_n) = \sigma^2\mathbf{I}_n \tag{3}$$

with the following assumption being fulfilled:

$$\lim_{n\to\infty} \frac{1}{n}\mathbf{X}_n^\top \mathbf{X}_n = \mathbf{V} \tag{4}$$

where $\mathbf{V}$ is positive definite. Then

- The LS-estimator $\hat{\beta}_n$ for β as well as the ML- and REML-estimators $\hat{\sigma}_n^2$ for σ^2 are consistent. ($\text{MSE}_\theta(\hat{\theta}) \to 0 \ n \to \infty$)
- The LS-estimator $\hat{\beta}_n$ for β is asymptotically normally distributed:

$$\sqrt{n}(\hat{\beta}_n - \beta) \to \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{V}^{-1}) \text{ (in distribution)} \tag{5}$$

Asymptotic Properties of the LS-Estimator

Proposition: Hence, for sufficiently large n it follows that $\hat{\beta}_n$ is approximately normally distributed with

$$\hat{\beta}_n \to \mathcal{N}(\beta, \sigma^2 \mathbf{V}^{-1}/n) \text{ (almost surely)} \tag{6}$$

Proposition:

- Similar to the error terms, also the residuals have expectation zero.
- In contrast to the error terms, the residuals are not uncorrelated.

Asymptotic Properties of the LS-Estimator

Proposition: Beside the usual assumptions, additionally assume that the error terms are normally distributed. Then the following properties hold:

- The distribution of the squared sum of residuals is given by:

$$\frac{\hat{\epsilon}^\top \hat{\epsilon}}{\sigma^2} = (n - p - 1)\frac{\hat{\sigma}^2}{\sigma^2} \tag{7}$$

- The squared sum of residuals $\hat{\epsilon}^\top \hat{\epsilon}$ and the LS-estimator $\hat{\beta}$ are independent.

Prediction

Proposition:

1. The expected prediction error is zero i.e., $\mathbb{E}[\hat{\mathbf{y}}_0 - \mathbf{y}_0] = 0$, i.e., $\mathbb{E}[\hat{\mathbf{y}}_0 - \mathbf{y}_0] = 0$
2. Prediction error covariance matrix is given by:

$$\mathbb{E}[(\hat{\mathbf{y}}_0 - \mathbf{y}_0)(\hat{\mathbf{y}}_0 - \mathbf{y}_0)^\top] = \sigma^2(\mathbf{X}_0(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_0^\top + \mathbf{I}_{T_0}) \tag{8}$$

Proof

Proof of (i): The true value is given by $y_0 = X_0\beta + \epsilon_0$. For the prediction error $\hat{y}_0 - y$ one obtains

$$\begin{aligned}
\mathbb{E}[\hat{y}_0 - y_0] &= \mathbb{E}[X_0\hat{\beta} - X_0\beta - \epsilon_0] && (9)\\
&= \mathbb{E}[X_0(\hat{\beta} - \beta) - \epsilon] && (10)\\
&= X_0 \underbrace{\mathbb{E}[\hat{\beta} - \beta]}_{\mathbb{E}[\hat{\beta}] - \beta} - \underbrace{\mathbb{E}[\epsilon_0]}_{=0} = 0 && (11)
\end{aligned}$$

Proof of (ii): For the prediction error variance are obtains

$$\begin{aligned}
\mathbb{E}[(\hat{y}_0 - y_0)^\top(\hat{y}_0 - y_0)] &= \mathbb{E}[(X_0(\hat{\beta} - \beta - \epsilon))(X_0(\hat{\beta} - \beta - \epsilon)^\top]\\
&= X_0\mathbb{E}[(\hat{\beta} - \beta - \epsilon)(\hat{\beta} - \beta - \epsilon)^\top]X_0^\top + \mathbb{E}[\epsilon_0\epsilon_0^\top]\\
&- X_0\mathbb{E}[(\hat{\beta} - \beta)\epsilon_0^\top] - \underbrace{\mathbb{E}[\epsilon_0(\hat{\beta} - \beta)^\top]}_{\epsilon_0 \text{ and } (\hat{\beta} - \beta) \text{ are independent}} X_0^\top\\
&= \sigma^2(X_0(X^\top X)^{-1}X_0^\top + I)
\end{aligned}$$

□

Iterative Solvers for Least-Squares Regression

So far: Given $\mathbf{y} \in \mathbb{R}^n$, solve

$$\min_{\beta} \frac{1}{2} \|\mathbf{X}\beta - \mathbf{y}\|$$

directly using $\beta^* = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}^\top$. Here

$$\mathbf{X} \in \mathbb{R}^{n \times (p+1)} \quad \text{and} \quad \beta \in \mathbb{R}^{(p+1)}.$$

Problems:

1. Generating $\mathbf{X}^\top \mathbf{X}$ and solving normal equations is too costly for large-scale problems.
2. Exact solution not useful when problem is ill-posed $\rightsquigarrow$ add explicit regularization or do so implicitly by early stopping.

Iterative methods that avoid working with $\mathbf{X}^\top \mathbf{X}$

- Steepest descent
- Conjugate gradient for least-squares (CGLS)

Iterative Methods

Idea: obtain a sequence $\beta_1, \ldots, \beta_j, \ldots$ that converges to least-squares solution β^*, i.e., $\beta_j \longrightarrow \beta^*$ for $j \to \infty$.

Question: How fast does the sequence converge?

Definition: Assume

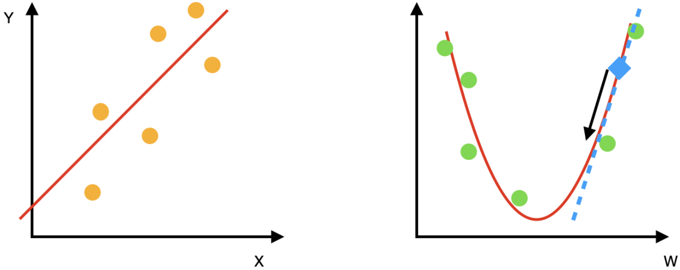
$$\|\beta_{j+1} - \beta^*\| < \gamma_j \|\beta_j - \beta^*\|$$

where all $\gamma_j < 1$. Then

- If γ_j is bounded away from 0 and 1 the convergence is referred to linear
- If $\gamma_j \to 0$ the convergence is referred to superlinear
- If $\gamma_j \to 1$ the convergence is referred to sublinear
- The sequence converges is called quadratically if γ_j is bounded away from 0 and 1 and

$$\|\beta_{j+1} - \beta^*\| < \gamma_j \|\beta_j - \beta^*\|^2$$

Steepest Descent for Least-Squares [Cauchy 1847]



Approach: Consider now

$$R_N(\beta) = \frac{1}{2}\|\mathbf{X}\beta - \mathbf{y}\|^2 \quad \text{with} \quad \nabla_\beta R_N(\beta) = \mathbf{X}^\top(\mathbf{X}\beta - \mathbf{y}).$$

Steepest descent direction is $\mathbf{d}_j = \mathbf{X}^\top(\mathbf{y} - \mathbf{X}\beta_j)$ and

$$\beta_{j+1} = \beta_j + \alpha_j \mathbf{d}_j$$

Steepest Descent for Least-Squares

How to choose α_j?

Idea: Minimize R_N along direction $\mathbf{d}_j$

$$\alpha_j = \arg\min_{\alpha} R_N(\beta_j + \alpha \mathbf{d}_j) = \arg\min_{\alpha} \frac{1}{2} \|\alpha \mathbf{X} \mathbf{d}_j - \mathbf{r}_j\|^2$$

with residual $\mathbf{r}_j = \mathbf{y} - \mathbf{X}\beta_j$.

This leads to simple quadratic equation in 1D whose solution is

$$\alpha_j = \frac{\mathbf{r}_j^\top \mathbf{X} \mathbf{d}_j}{\|\mathbf{X} \mathbf{d}_j\|^2}$$

Algorithm: Steepest Descent for Least-Squares

Algorithm 1 Steepest Descent for Least-Squares

for $j = 1, \ldots$ **do**
  Compute residual $\mathbf{r}_j = \mathbf{y} - \mathbf{X}\beta_j$
  Determine the SD direction $\mathbf{d}_j = \mathbf{X}^\top \mathbf{r}_j$
  Compute step size $\alpha_j = \frac{\mathbf{r}_j^\top \mathbf{X}\mathbf{d}_j}{\|\mathbf{X}\mathbf{d}_j\|^2}$
  Take the step $\beta_{j+1} = \beta_j + \alpha_j \mathbf{d}_j$
end for

Remark: The algorithm converges linearly, i.e.,

$$\|\beta_{j+1} - \beta^*\| < \gamma \|\beta_j - \beta^*\| \quad \text{with} \quad \gamma \approx \left| \frac{\kappa - 1}{\kappa + 1} \right|$$

Here, κ depends on condition number of $\mathbf{X}$, i.e.,

$$\kappa \approx \frac{\sigma^2_{\max}}{\sigma^2_{\min}}$$

Accelerating Steepest Descent: Post-Conditioning

Idea: Improve convergence by transforming the problem

$$\phi(\beta) = \frac{1}{2}\|\mathbf{XSS}^{-1}\beta - \mathbf{y}\|^2$$

Here: $\mathbf{S}$ is invertible

Solve in two steps:

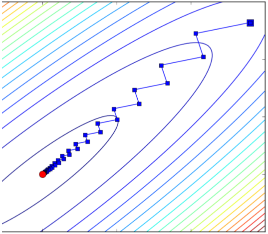
1. Set $\mathbf{z} = \mathbf{S}^{-1}\beta$ and compute

$$\mathbf{z}^* = \arg\min_{\mathbf{z}} \frac{1}{2}\|\mathbf{XSz} - \mathbf{y}\|^2$$

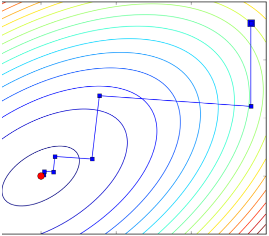
2. Then $\beta = \mathbf{Sz}$.

Pick $\mathbf{S}$ such that $\mathbf{XS}$ is better conditioned.

original problem:



post-conditioned:



Conjugate Gradient Method for Least-Squares

CG is designed to solve quadratic optimization problems

$$\min_{\beta} \frac{1}{2}\beta^\top \mathbf{H}\beta - \mathbf{b}^\top \beta$$

with $\mathbf{H}$ symmetric positive definite. In our case

$$\arg\min_{\beta} \frac{1}{2}\|\mathbf{X}\beta - \mathbf{y}\|^2 = \arg\min_{\beta} \frac{1}{2}\beta^\top \underbrace{\mathbf{X}^\top \mathbf{X}}_{=\mathbf{H}}\beta - \underbrace{\mathbf{y}^\top \mathbf{X}}_{=\mathbf{b}^\top}\beta$$

CG improves over SD by using previous step (not a memory-less method) and constructing a basis for the solution.

Facts:

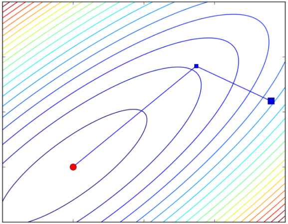
- terminates after at most n steps (in exact arithmetic)
- good solutions for $j \ll n$
- convergence $\gamma_j \approx \left|\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right|^j$

Conjugate Gradient Least-Squares

- Uses the structure of the problem to obtain stable implementation
- Typically converges much faster than SD
- Accelerate using post conditioning

$$\min_{\beta} \frac{1}{2}\|\mathbf{XSS}^{-1}\beta - \mathbf{y}\|^2$$

- Faster convergence when eigenvalues of $\mathbf{S}^\top\mathbf{X}^\top\mathbf{XS}$ are clustered.



Iterative Regularization

Consider

$$\min_{\beta} \|\mathbf{X}\beta - \mathbf{b}\|^2$$

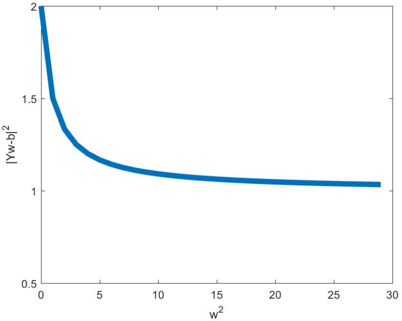
- Assume that $\mathbf{X}$ has non-trivial null space
- The matrix $\mathbf{X}^\top \mathbf{X}$ is not invertible
- Can we still use iterative methods (CG, CGLS, ...)?

What are the properties of the iterates?

Iterative Regularization: L-Curve

The CGLS algorithm has the following properties

- For each iteration $\|\mathbf{X}\beta_k - \mathbf{y}\|^2 \leq \|\mathbf{X}\beta_{k-1} - \mathbf{y}\|^2$
- If starting from $\beta = 0$ then $\|\beta_k\|^2 \geq \|\beta_{k-1}\|^2$
- $\beta_1, \beta_2, \ldots$ converges to the minimum norm solution of the problem
- Plotting $\|\beta_k\|^2$ vs $\|\mathbf{X}\beta_k - \mathbf{y}\|^2$ typically has the shape of an L-curve



Ill-posedness and Regularization

Proposition: If the least squares problem is ill-posed, i.e., solution does not exist or is unstable.

Small perturbations in $\mathbf{y}$ or $\mathbf{X}$ yield large perturbations in β

Solve regularized problem: For some $\lambda > 0$ and matrix $\mathbf{G}$

$$\min_{\beta} \frac{1}{2}\|\mathbf{X}\beta - \mathbf{y}^\top\|^2 + \frac{\lambda}{2}\|\mathbf{G}\beta\|^2$$

Ridge Regularization (L_2)

Definition: The solution to the so called ridge regression is given by

$$\begin{aligned}\hat{\beta}^{Ridge} &= \arg\min_{\beta\mathbb{R}} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_2^2 \\ &= (\mathbf{X}^\top \mathbf{X} + \lambda I_p)^{-1}\mathbf{X}^\top \mathbf{y}\end{aligned}$$

Properties

- decreases variance but increases bias (for increasing λ)
- Can improve predictive performance
- special case of Tikhonov regularization

Lasso Regularization (L_1)

Definition:

$$\hat{\beta}^{Lasso} = \arg\min_{\beta\mathbb{R}} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1 \tag{12}$$

Properties

- LASSO=Least Absolute Shrinkage and Selection Operator
- This penalty allows coefficients to shrink towards exactly zero
- LASSO usually leads to sparse models, that are easier to interpret

Cross Validation

Finding good least-squares solution requires good parameter selection.

- λ when using Tikhonov regularization (weight decay)
- number of iteration (for SD and CGLS)

Suppose that we have two different "solutions"

$$\begin{aligned} \beta_1 \quad &\rightarrow \quad \|\beta_1\|^2 = \eta_1 \quad \|\mathbf{X}\beta_1 - \mathbf{y}\|^2 = \rho_1. \\ \beta_2 \quad &\rightarrow \quad \|\beta_2\|^2 = \eta_2 \quad \|\mathbf{X}\beta_2 - \mathbf{y}\|^2 = \rho_2. \end{aligned}$$

How to decide which one is better?

Cross Validation

Goal: Gauge how well the model can predict new examples.

Let $\{\mathbf{X}_{\mathrm{CV}}, \mathbf{y}_{\mathrm{CV}}\}$ be data that is **not used** for the training

Idea: If $\|\mathbf{X}_{\mathrm{CV}}\beta_1 - \mathbf{y}_{\mathrm{CV}}\|^2 \leq \|\mathbf{X}_{\mathrm{CV}}\beta_2 - \mathbf{y}_{\mathrm{CV}}\|^2$, then β_1 is a better solution that β_2.

When the solution depends on some hyper-parameter(s) λ, we can phrase this as bi-level optimization problem

$$\lambda^* = \arg\min_{\lambda} \|\mathbf{X}_{\mathrm{CV}}\beta(\lambda) - \mathbf{y}_{\mathrm{CV}}\|^2,$$

where $\beta(\lambda) = \arg\min_{\beta} \frac{1}{2}\|\mathbf{X}\beta - \beta\|^2 + \frac{\lambda}{2}\|\beta\|^2$.

Cross Validation

To assess the final quality of the solution cross validation is not sufficient (why?).

Need a final testing set.

Procedure:

- Divide the data into 3 groups $\{\mathbf{X}_{\text{train}}, \mathbf{X}_{\text{CV}}, \mathbf{X}_{\text{test}}\}$.
- Use $\mathbf{X}_{\text{train}}$ to estimate $\beta(\lambda)$
- Use $\mathbf{X}_{\text{CV}}$ to estimate λ
- Use $\mathbf{X}_{\text{test}}$ to assess the quality of the solution

Important - we are not allowed to use $\mathbf{X}_{\text{test}}$ to tune parameters!