DRP HW 4 - The Gauss-Markov Theorem

Owen Huang

July 3, 2024

Notes

This homework introduces and proves the *Gauss-Markov theorem* for linear estimators. You will see how the normal equations which you derived in the previous homework is in some sense, *optimal*, and how when choosing estimators, we are forced to choose a balance between bias and variance. It just so happens that the mean squared error (MSE) solution is the best one!

1 Introduction and Warmup

Before we can state some of the theorems in this homework precisely, there are some probabilistic prerequisites we must cover, namely some basic results about statistics in higher dimensions.

Recall that a random vector $X = (X_1, \dots X_n)$ is just a list of n random variables $X_i : \mathbb{R} \to \mathbb{R}$. So we can naturally define the expectation of a random vector to be the vector of expectations of each coordinate place.

Exercise 1.1. Show that this definition of expectation preserves the linearity property: if X, Y are random vectors, then $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.

Now, to define variance, observe that the usual approach of $\operatorname{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2]$ doesn't really make sense [why?]. This leads us to the definition of the *variance-covariance matrix*.

Definition 1.2. *Given a random vector $X = (X_1, \dots, X_n)$, the* ***variance-covariance*** *matrix is given by*

$$\operatorname{Var}(X) = \operatorname{Cov}(X, X) = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T]$$

Exercise 1.3. Show that the i-th diagonal entry in $\operatorname{Var}(X)$ is equal to $\operatorname{Var}(X_i)$.

So in some sense, this matrix generalizes the notion of normal variance in 1 dimension! Now in linear algebra, our favourite thing to do is to multiply things by a matrix, and see what happens. In other words, how does the variance change after a linear transformation? Let's start in 1 dimension to get some intuition:

Exercise 1.4. Let $X : \mathbb{R} \to \mathbb{R}$ be a random variable. Show that for any scalar $a \in \mathbb{R}$, $\operatorname{Var}(aX) = a^2 \operatorname{Var}(X)$.

And actually, the higher dimensional case follows in a similar way.

Exercise 1.5. Let X a random vector, and A a (nonrandom) matrix of appropriate size. Prove that $\operatorname{Var}(AX) = A \operatorname{Var}(X) A^T$.

And finally, a linear algebra exercise before we move onto the major theorem for this homework.

Exercise 1.6. Let x be a random vector with $\mathbb{E}[x] = \mu$ and $\operatorname{Var}(x) = \Sigma$. Prove that for a fixed nonrandom square matrix of appropriate size A, we have the identity $\mathbb{E}[x^T A x] = \mu^T A \mu + \operatorname{tr}(A\sigma)$, where $\operatorname{tr}(X)$ is the trace of the the square matrix X.

2 The Gauss-Markov Theorem

Let us quickly recall the setting in which we are working. We observe pairs of inputs, which we collect in a data set $\mathcal{D} = \{(x^{(0)}, y^{(0)}), \ldots, (x^{(d)}, y^{(d)})\} \subset \mathbb{R}^{n+1}$, with $x^{(i)} \in \mathbb{R}^n$, and we are now to pick $\beta = (\beta_1, \ldots \beta_n)^T$ to fit the d linear equations $y^{(i)} = \beta_1 x_1^{(i)} + \ldots \beta_n x_n^{(i)} + \epsilon^{(i)}$. Here, we require the $\epsilon(i)$ to be of mean zero, but curiously, they need not be independent. We will however need that the variance of all the $\epsilon^{(i)}$ to be the same, that is: $\mathrm{Var}(\epsilon^{(i)}) = \sigma^2 > 0$ for every $1 \leq i \leq d$. Those error terms should be interpreted as irreducible noise - no matter how robust our estimation system is, it can never get rid of the fundamental noise [1] We define the mean-squared-loss by

$$\mathcal{L}(\beta) = \sum_{i=1}^{d} (\hat{f}(x^{(i)}; \beta) - y^{(i)})^2 \tag{1}$$

and we try to minimize this loss over all choices of $\beta \in \mathbb{R}$. In last week's work, you proved that when $d \geq n$ (ie., more observations than regressors), we had the optimal solution $\beta^{mse} = (X^T X)^{-1} X^T Y$. Moreover, in class we showed that β^{mse} is unbiased, which we interpret [2] as the fact that when we take more and more observations, the value of β^{mse} should approximate the true β. However, such an estimator still has *variance* - we do not have enough information to conclude about the *speed* of its convergence to β. The Gauss-Markov theorem asserts that in fact, out of all unbiased linear estimators of β, β^{mse} has the lowest variance. Let us make this a little more precise.

Definition 2.1. *Given a linear regression problem, a* ***linear estimator*** *of the coefficient vector $\beta \in \mathbb{R}^n$ is of the form $\hat{\beta} = Ay$ for an appropriately sized matrix A. Such a matrix A must be dependent only on the $x^{(i)}, y^{(i)}$ as those are the only things one observes.*

Theorem 2.2 (Gauss-Markov)**.** *Suppose that α is an unbiased linear estimator of β. Then* $\mathrm{Var}(\alpha) \geq \mathrm{Var}(\beta^{mse})$

Proof. Because linear estimators are defined by their associated matrix, we can assume that $\alpha = AY$ for some matrix A. Also, here what we mean by $\mathrm{Var}(\alpha) \geq \mathrm{Var}(\beta^{mse})$ is that $\mathrm{Var}(\alpha) - \mathrm{Var}(\beta^{mse})$ is a positive semidefinite matrix. Recall that a square matrix M is *positive semidefinite* if for all x, $x^T M x \geq 0$.

Exercise 2.3. What is the size of the matrix A?

Observe that we can write $A = (X^T X)^{-1} X^T + B$ for another matrix B.

Exercise 2.4. Use the equation $Y = X\beta + \epsilon$ to compute $\mathbb{E}[\alpha] = \mathbb{E}[AY]$ in terms of only an identity matrix, β, B and X. By the unbiasedness of α, which matrix product must equal zero? This gives you a constraint which you will use later.

Exercise 2.5. Use the fact that X and β are nonrandom to show that $\mathrm{Var}(Y)$ is a diagonal matrix. What is the common value on the diagonal?

Exercise 2.6. Use the answer to the previous exercise and Exercise 1.5 to compute $\mathrm{Var}(AY)$. You will also need the constraint you found in Exercise 2.4.

Exercise 2.7. Prove that for any matrix M, MM^T is positive semidefinite. Conclude the proof for the *Gauss-Markov Theorem.*

□

[1] In real life, this might represent physical constraints, such as data lost through wires, or data compression.
[2] in conjunction with the central limit theorem