Bias and variance for linear regression

Fraida Fund

In this set of notes, we derive the bias and variance for linear regression models, including linear basis function models.

Linear basis function model

For data $(x_i, y_i), i = 1, \ldots, n$, consider the linear basis function model:

$$\hat{y} = f(x, w) = \phi(x)^T w = w_1 \phi_1(x) + \ldots + w_p \phi_p(x)$$

The least squares fit is

$$w = (\Phi^T \Phi)^{-1} \Phi^T y$$

where

$$\Phi = \begin{bmatrix} \phi_1(x_1) & \cdots & \phi_p(x_1) \\ \vdots & \ddots & \vdots \\ \phi_1(x_n) & \cdots & \phi_p(x_n) \end{bmatrix}$$

Assume the true function is

$$y = t(x) + \epsilon, \quad \epsilon \sim N(0, \sigma_\epsilon^2)$$

When there is no under-modeling,

$$t(x) = f(x, w_t) = \phi(x)^T w_t$$

where w_t is the true parameter vector.

Unique solution to ordinary least squares estimate

For $\Phi \in R^{n \times p}$, there is a unique solution to the ordinary least squares estimate

$$w = (\Phi^T \Phi)^{-1} \Phi^T y$$

only if Rank$(\Phi) = n$. This will be the case if the columns of Φ are linearly independent, and $n \geq p$.

In other words, the unique solution exists only if the number of data samples for training (n) is greater than or equal to the number of parameters p.

This limits the model complexity you can use (greater $p \implies$ greater model complexity).

For the rest of these notes, we will assume a unique least squares solution ($n \geq p$).

Bias of linear model

Let us give a definition of *bias* on a test point, (x_t, y_t) for a function f with parameter estimate $\widehat{w}$:

$$\text{Bias}(x_t) := t(x_t) - E[f(x_t, \widehat{w})]$$

We will try to derive the bias for a linear regression when the true functino is in the assumed model class, i.e. there is no under-modeling.

Suppose that there is no under-modeling, so there is a parameter vector w_t such that

$$t(x) = f(x, w_t) = \phi(x)^T w_t$$

Then for each training sample $i = 1, \dots, n$,

$$y_i = \phi(x_i)^T w_t + \epsilon_i$$

and for the entire training set, $y = \Phi w_t + \epsilon$.

For a fixed training set, the least squares parameter estimate will be

$$\begin{aligned}\widehat{w} &= (\Phi^T\Phi)^{-1}\Phi^T y \\ &= (\Phi^T\Phi)^{-1}\Phi^T(\Phi w_t + \epsilon) \\ &= w_t + (\Phi^T\Phi)^{-1}\Phi^T\epsilon\end{aligned}$$

Now we can find $E[\widehat{w}]$ over the samples of noisy training data: since $E[\epsilon] = 0$, we have $E[\widehat{w}] = w_t$.

Informally, we can say that on average, the parameter estimate matches the "true" parameter.

Then $E[f(x_t, \widehat{w})] = E[f(x_t, w_t)] = t(x_t)$.

Conclusion: We can see that when the model is linear and there is no under-modeling, there is no bias:

$$\text{Bias}(x_t) = 0$$

Random vectors

Before we look at the variance, we will review some terminology of random vectors:

- A **random vector** $x = (x_1, \dots, x_d)^T$ is a vector where each x_j is a random variable.
- The **vector of means** of x is $\mu = (E[x_1], \dots, E[x_d])^T = (u_1, \dots, u_d)^T$.
- The **covariance** of x_i, x_j is $Cov(x_i, x_j) = E[(x_i - \mu_i)(x_j - \mu_j)]$
- The **variance matrix** (which is a $d \times d$ matrix) is:

$$Var(x) := E[(x-u)(x-u)^T] = \begin{bmatrix} Cov(x_1, x_1) & \cdots & Cov(x_1, x_d) \\ \vdots & \ddots & \vdots \\ Cov(x_d, x_1) & \cdots & Cov(x_d, x_d) \end{bmatrix}$$

- In a **linear transform** $y = Ax + b$, the input $x \in R^N$ is mapped to $Ax \in R^M$ by $A \in R^{M \times N}$
- The mean and variance matrix under this linear transform are given by $E(y) = AE(x) + b$ and $Var(y) = AVar(x)A^T$, respectively.

Variance of linear model

Variance of parameter estimate

Recall that ϵ_i are independent for different samples, with $E[\epsilon_i] = 0$ and $Var(\epsilon) = \sigma_\epsilon^2$.
Then,

$$Cov(\epsilon_i, \epsilon_j) = \begin{cases} 0, & i \neq j \\ \sigma_\epsilon^2, & i = j \end{cases}$$

so the variance matrix for the ϵ noise is

$$Var(\epsilon) = \sigma_\epsilon^2 I$$

(I is the identity matrix). Also recall from our discussion of bias that with no under-modeling,

$$\hat{w} = w_t + (\Phi^T \Phi)^{-1} \Phi^T \epsilon$$

Let us think of this as a linear transform of $\hat{w}$, $y = Ax + b$ where:

- $y = \hat{w}$
- $A = (\Phi^T \Phi)^{-1} \Phi^T$
- $x = \epsilon$
- $b = w_t$

and recall that for a linear transform $y = Ax + b$, $Var(y) = AVar(x)A^T$.
Then we can compute the variance matrix of the *parameter estimate* for the linear model as

$$\begin{aligned} Var(\hat{w}) &= [(\Phi^T \Phi)^{-1} \Phi^T][Var(\epsilon)][(\Phi^T \Phi)^{-1} \Phi^T]^T \\ &= [(\Phi^T \Phi)^{-1} \Phi^T][\sigma_\epsilon^2 I][(\Phi^T \Phi)^{-1} \Phi^T]^T \\ &= [(\Phi^T \Phi)^{-1} \Phi^T][\sigma_\epsilon^2 I][\Phi(\Phi^T \Phi)^{-1}] \\ &= \sigma_\epsilon^2 (\Phi^T \Phi)^{-1} \end{aligned}$$

Variance of model output

Now, we will use $Var(\hat{w})$ to compute $Var(x_t)$ for the linear model.
First, recall from our discussion of bias that when there is no under-modeling

$$E[f(x_t, \hat{w})] = \phi(x_t)^T \hat{w} = \phi(x_t)^T w_t$$

Then the variance of the linear model output for a test point is

$$\begin{aligned} Var(x_t) &= E[f(x_t, \hat{w}) - E[f(x_t, \hat{w})]]^2 \\ &= E[\phi(x_t)^T \hat{w} - \phi(x_t)^T w_t]^2 \\ &= E[\phi(x_t)^T (\hat{w} - w_t)]^2 \end{aligned}$$

Also note the following trick: if a is a non-random vector and z is a random vector, then

$$E[a^T z]^2 = E[a^T z z^T a] = a^T E[z z^T] a$$

Therefore,

$$\begin{aligned} Var(x_t) &= E[\phi(x_t)^T(\hat{w} - w_t)]^2 \\ &= \phi(x_t)^T E[(\hat{w} - w_t)(\hat{w} - w_t)^T]\phi(x_t) \end{aligned}$$

Finally, recall that

$$Var(\hat{w}) = E[(\hat{w} - w_t)(\hat{w} - w_t)^T] = \sigma_\epsilon^2(\Phi^T\Phi)^{-1}$$

so

$$\begin{aligned} Var(x_t) &= \phi(x_t)^T E[(\hat{w} - w_t)(\hat{w} - w_t)^T]\phi(x_t) \\ &= \sigma_\epsilon^2 \phi(x_t)^T(\Phi^T\Phi)^{-1}\phi(x_t) \end{aligned}$$

This derivation assumed there is no under-modeling. However, in the case of under-modeling, the variance expression is similar.

For the next part, we will compute the variance term from the *in-sample* prediction error, i.e. the error if the test point is randomly drawn from the training data:

- Training data is $(x_i, y_i), i = 1, \dots, n$
- $x_t = x_i$ with probability $\frac{1}{n}$

Since the rows of Φ are $\phi(x_i)^T$, then

$$\Phi^T\Phi = \sum_{i=1}^{n} \phi(x_i)\phi(x_i)^T$$

We will use a trick: for random vectors u, v, $E[u^T v] = Tr(E[vu^T])$, where $Tr(X)$ is the sum of diagonal of X.

Then the expectation (over the test points) of the variance of the model output is:

$$\begin{aligned} E[Var(x_t)] &= \sigma_\epsilon^2 E[\phi(x_t)^T(\Phi^T\Phi)^{-1}\phi(x_t)] \\ &= \sigma_\epsilon^2 Tr\left(E[\phi(x_t)\phi(x_t)^T](\Phi^T\Phi)^{-1}\right) \\ &= \frac{\sigma_\epsilon^2}{n} Tr\left(\sum_{i=1}^{n}[\phi(x_i)\phi(x_i)^T](\Phi^T\Phi)^{-1}\right) \\ &= \frac{\sigma_\epsilon^2}{n} Tr\left((\Phi^T\Phi)(\Phi^T\Phi)^{-1}\right) \\ &= \frac{\sigma_\epsilon^2}{n} Tr\left(I_p\right) \\ &= \frac{\sigma_\epsilon^2 p}{n} \end{aligned}$$

The average variance increases with the number of parameters p, and decreases with the number of samples used for training n, as long as the test point is distributed like the training data.

Summary of results for linear regression

Suppose the model class is linear with n samples and p parameters.

Result 1: Uniqueness of coefficient estimate

When $n < p$, the least squares estimate of the coefficients is not unique.

Result 2: Bias of estimate of target variable

When $n \geq p$ and the least squares estimate of the coefficients is unique, *and* there is no under-modeling, then the estimate of the target variable is unbiased.

Result 3: Variance of estimate of target variable

When $n \geq p$, the least squares estimate of the coefficients is unique, there is no under-modeling, *and* the test point is drawn from the same distribution as the trainng data, then the variance of the estimate of the target variable increases linearly with the number of parameters and inversely with the number of samples used for training:

$$Var = \frac{p}{N}\sigma_\epsilon^2$$

Result 4: Overall prediction error

The overall expected in-sample prediction error for the ordinary least squares linear regression is

$$0 + \frac{p}{N}\sigma_\epsilon^2 + \sigma_\epsilon^2$$

where the three terms represent the squared bias, the variance, and the irreducible error.