Regressão Linear: Probabilidade vs. Otimização

Uma dedução dos pressupostos sobre os erros

Luca WB

2026-08-09

Table of contents

1 Breve introdução

Esse post surgiu por conta da minha primeira aula da cadeira de Aprendizado Profundo na universidade, nessa aula, o professor fez a demonstração da formula normal da regressão linear. Não era exatamente uma novidade essa formula, eu já sabia demonstrar essa formula de cabeça usando valor esperado condicional, o que me intrigou foi um detalhe da demonstração. Meu professor usou a abordagem de otimização para dedução, calculando o ponto minimo global da função perda/avaliação. Isso me intrigou pelo fato de que, a primeira vista, não tinhamos levantado nenhum pressuposto sobre a natureza do erro do nosso modelo, coisa que temos que fazer para realiar a dedução via valor esperado.

Assim, gastei meu fim de semana seguinte pensando sobre o assunto, pesquisei um pouco na internet sobre a correlção entre as duas abordagens e os pressupostos dela, e não achei nada muito satisfatório, então comecei a rabiscar no papel para ver se esclarecia alguma coisa, e acredito ter achado uma perspectiva interessante para ver a relação entre as duas abordagens no que tange os pressupostos sobre o erro

2 Dedução via Otimização

Primeiro, antes de mais nada, é preciso estar familiarizado sobre as deduções, pare que então possamos falar sobre elas.

A ideia central na regressão linear é encontrar os parâmetros (ou pesos) que minimizam o erro das nossas previsões. Assim como no exemplo do gradiente onde olhamos para a derivada para encontrar a direção que reduz o valor da função, aqui o nosso objetivo principal é minimizar a função de custo conhecida como Soma dos Erros Quadráticos (SSE - *Sum of Squared Errors*).

A fórmula clássica do SSE é a soma das diferenças ao quadrado entre o valor real (y_i) e o valor previsto ($\hat{y}_i$):

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

2.1 O Truque de Notação (Notation Trick)

Antes de derivarmos, precisamos simplificar nossa notação. Um modelo linear comum tem um viés (intercepto ou *bias*) b e os pesos das características $w_1, \dots, w_m$. A previsão para uma única amostra é escrita como:

$$\hat{y}_i = b + w_1 x_{i1} + \cdots + w_m x_{im}$$

Para evitar o trabalho de calcular as derivadas do viés b separadamente dos outros pesos, nós usamos um **truque de notação** (*notation trick*). Nós incorporamos o b no vetor de pesos w (sendo o primeiro elemento, w_0) e adicionamos uma coluna inteira de 1s na nossa matriz de características X.

Dessa forma, a previsão de todas as amostras ao mesmo tempo se transforma em uma elegante multiplicação de matrizes:

$$\hat{y} = Xw$$

Com esse truque, a nossa função de erro SSE pode ser reescrita na sua forma vetorializada:

$$SSE(w) = (y - Xw)^T (y - Xw)$$

2.2 Expandindo a Equação Algébrica

Para conseguir derivar a função e encontrar o ponto de mínimo, vamos primeiro expandir a multiplicação dessas matrizes:

$$SSE(w) = (y^T - (Xw)^T)(y - Xw)$$
$$SSE(w) = y^T y - y^T Xw - (Xw)^T y + (Xw)^T (Xw)$$

Aplicando a propriedade da matriz transposta, onde $(AB)^T = B^T A^T$:

$$SSE(w) = y^T y - y^T Xw - w^T X^T y + w^T X^T Xw$$

Aqui entra um detalhe geométrico e algébrico importante, se você observar as dimensões, o termo $y^T Xw$ resulta em um único número, ou seja, é um **escalar**. Como a transposta de um escalar é ele mesmo, podemos dizer que $(y^T Xw)^T = w^T X^T y$. Portanto, os dois termos do meio são idênticos e podemos agrupá-los:

$$SSE(w) = y^T y - 2w^T X^T y + w^T X^T Xw$$

2.3 A Derivada Matricial (Calculando o Gradiente)

A ideia básica é que, se queremos minimizar o valor da função, precisamos olhar para a derivada dela. Como SSE é uma função convexa, o seu ponto de mínimo global ocorre exatamente onde a inclinação (o gradiente) é zero. Vamos derivar a função em relação ao vetor de pesos w:

$$\frac{\partial}{\partial w} SSE(w) = \frac{\partial}{\partial w}(y^T y - 2w^T X^T y + w^T X^T Xw)$$

Analisando termo a termo:

1. A derivada de $y^T y$ em relação a w é 0 (pois é apenas uma constante, não depende de w).
2. Usando a regra de derivada matricial $\frac{\partial}{\partial w}(w^T a) = a$, a derivada de $-2w^T X^T y$ é $-2X^T y$.
3. Para formas quadráticas, a regra é $\frac{\partial}{\partial w}(w^T Aw) = 2Aw$. Logo, a derivada de $w^T X^T Xw$ é $2X^T Xw$.

Juntando todas as peças, o nosso gradiente é:

$$\frac{\partial}{\partial w} SSE(w) = -2X^T y + 2X^T Xw$$

2.4 Minimização e a Equação Normal

Agora, basta aplicar a ideia fundamental da otimização: igualar o gradiente a zero para encontrar o fundo do "vale" da função de custo.

$$0 = -2X^T y + 2X^T Xw$$
$$2X^T Xw = 2X^T y$$

Dividindo os dois lados por 2, temos:

$$X^T Xw = X^T y$$

Finalmente, para isolar o nosso vetor de pesos ótimos w, multiplicamos ambos os lados da equação pela inversa da matriz $(X^T X)$:

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

E esta é a **Equação Normal** da regressão linear. Ela nos dá, de forma direta e analítica, os pesos exatos que minimizam o erro.

3 Dedução via Probabilidade/Esperança

Na seção anterior obtivemos a Equação Normal resolvendo um problema de otimização. Agora vamos chegar ao mesmo resultado partindo de uma interpretação probabilística do modelo linear.

3.1 O pressuposto fundamental

Considere o modelo

$$y = Xw + u,$$

onde u é nosso erro, isto é, representa tudo aquilo que não está sendo explicado pelas variáveis observadas em X (Já estamos usando o notation trick e notação matricial). Por definição do nosso modelo $u = y - Xw$

A hipótese mais importante é que iremos assumir, é:

$$\boxed{E[u \mid X] = 0}.$$

Em palavras:

> Depois de observarmos os valores de X, o erro continua tendo média zero.

Essa hipótese é mais forte do que simplesmente assumir

$$E[u] = 0.$$

De fato,

$$E[u \mid X] = 0 \implies E[u] = 0,$$

mas a recíproca não é verdadeira.

3.2 Por que assumir $E[u \mid X] = 0$?

A ideia é que toda a estrutura sistemática explicável por X já foi incorporada ao termo Xw.

Se a média do erro ainda dependesse de X, então existiria informação em X que o modelo não está utilizando.

Por exemplo, suponha que

$$E[u \mid X = x] = x^2.$$

Nesse caso, conhecer x permite prever o erro em média, indicando que o modelo ainda está omitindo uma relação importante.

A condição

$$E[u \mid X] = 0$$

formaliza exatamente a ideia de que o erro contém apenas a parte não explicada do fenômeno.

3.3 Ortogonalidade entre erro e variáveis explicativas

Da propriedade da esperança condicional segue que

$$E[Xu] = 0.$$

Isso é verdade, pois como ambos são independentes na média, podemos escrever como $E[X]E[u]$, e como $E[u] = 0$, temos a formula acima

Substituindo

$$u = y - Xw,$$

obtemos

$$E[X(y - Xw)] = 0.$$

Expandindo:

$$E[Xy] - E[XX^T]w = 0.$$

Logo,

$$\boxed{E[XX^T]w = E[Xy]}$$

Esta é a versão **populacional** da Equação Normal.

3.4 Da população para a amostra

Até aqui trabalhamos com quantidades populacionais, isto é, esperanças teóricas que normalmente não conhecemos.

Na prática observamos apenas uma amostra de tamanho n. Assim, substituímos os momentos populacionais pelos seus análogos amostrais:

$$E[Xy] \approx \frac{1}{n} X^T y,$$

e

$$E[XX^T] \approx \frac{1}{n} X^T X.$$

Substituindo na equação anterior:

$$\frac{1}{n} X^T X \hat{w} = \frac{1}{n} X^T y.$$

Multiplicando ambos os lados por n:

$$X^T X \hat{w} = X^T y.$$

Finalmente,

$$\boxed{\hat{w} = (X^T X)^{-1} X^T y.}$$

---

4 Comparando as duas deduções

Bom, como viram, na dedução via esperança assumimos explicitamente algo sobre o erro, que ele deve ser independente das variaveis preditoras. Ao passo de que a principio não assumimos nada sobre o erro quando deduzimos via otimização. Entretanto isso não é exatamente verdade

A ideia fundamental aqui é que Otimizar o Erro Quadrático e Garantir a Premissa de Ortogonalidade são, na verdade, duas faces da mesma moeda. Quando pedimos para um otimizador (como o Gradient Descent) minimizar a distância entre a previsão e o valor real, ele faz isso calculando derivadas. Relembrando, a formula do gradiente:

$$\frac{\partial}{\partial w} SSE(w) = -X^T y + X^T X w$$

E idendica a formula do erro pelo lado da esperança

$$E[Xu] = E[Xy] - E[XX^T]w.$$

E nesse momento, o que fizemos com o gradiente na perspectiva da otimização? Igualamos a zero Na perspectiva da otimização, isso apenas nos diz que queremos que o erro seja o menor possivel, e a principio, parece que estamos assumino apenas a condição fraca que a $E[u] = 0$. Porem agora que sabemos que o gradiente é igual ao valor esperado do erro, podemos ver alguma relações interessantes:

$$\frac{\partial}{\partial w} SSE(w) = E[-Xu]$$
$$\frac{\partial}{\partial w} SSE(w) = -X^T \hat{u}$$

Onde $\hat{u}$ são os residuos (pense nele como os erros concretos, que realmente aconteceram, e não o objeto teorico do erro) Considerando que igualamos $-X^T \hat{u}$ a 0, a unica forma disso ser possivel, é se X e $\hat{u}$ forem ortogonais, isso é, possuirem correlação igual a 0, e caso não saiba, a formula $X^T \hat{u}$ é exatamente a formula da covariância multivarida.

Abaixo há um grafico interativo, onde ploto uma regressão linear com duas variaveis preditoras X_1, X_2 e uma terceira variavel alvo y. Esse espaço vetorial das nossas variaveis nos permite visualizar como existe um espaço de erro pelo simples fato de as variaveis preditoras não conseguirem apontar para a varivel alvo atraves de combinações lineares, dessa forma, podemos ver claramente que o erro é ortogonal as variaveis preditoras

```
Unable to display output for mime type(s): text/html
```

(a) Projeção Ortogonal

```
Unable to display output for mime type(s): text/html
```

(b)

Figure 4.1

Com isso podemos ver que na realidade, em ambos os casos estamos assumindo exatamente os mesmos pressupostos, apesar de no início não parecer o caso.

Um último detalhe interessante, como pudermos ver, não assumimos que o erro é gaussiano para chegar nesse resultado, acredito que muitas pessoas se confundam achando que a premissa de otimizar um modelo lienar com SSE, ou melhor, via a formula normal pressupoem estritamente isso, o que não é verdade. Entretanto, acredito que essa confusão seja devido ao Teorema de Gauss-Markov que diz que caso o erro possua variância constante, nossa regressão linear é um BLUE (Best Linear Unbiased Estimator), e se assumirmos que todos os erros vem de uma mesma gaussiana, o erro possui variância constante.

Nota: Existe um outro teorema, chamado Teorema de Rao-Blackwell, que segundo ele, caso o erro siga uma gaussiana, a formula normal não apenas nos garante um BLUE, como um BUE (Best Unbiased Estimator)