Lecture 4

- Linear Algebra of Linear Regression
  - Residual Sum of Squares
  - Normal Equation
  - Ordinary Least Squares

In the next few lectures we will discuss linear methods for solving regression problems.

Let $X = (X_1, \dots X_p)^T$ be an input vector.

We want to predict output $Y \in \mathbb{R}$, $\hat{Y} = f(X) \approx Y$.

Let's revisit the linear regresion method.

In linear regression: $f(X) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j = X^T \beta$, where $X = \begin{bmatrix} 1 \\ X_1 \\ \vdots \\ X_p \end{bmatrix}$ $\underbrace{\beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix}}_{\substack{\text{unknown} \\ \text{parameters} \\ \text{or} \\ \text{coefficients}}}$

Remark: Note that inputs $X_1, \dots, X_p$ can come from different sources:

- Quantitative inputs, $X_j \in \mathbb{R}$
- Transformation of quantitative inputs, $X_j = \log \tilde{X}_j$, $X_j = \tilde{X}_j^2$, etc.
- Basis expansions, such as $X_2 = X_1^2$, $X_3 = X_1^3, \dots, X_p = X_1^p$. $\Rightarrow$ fitting a polynomial.
- Interactions between inputs, such as $X_1, X_2, X_3 = X_1 \cdot X_2$
- Dummy coding of qualitative inputs: if $\tilde{X}_j$ has three levels (say, red, blue, pink) then we can create X_1, X_2, X_3 to represent $\tilde{X}_j$

$$X_1 = \begin{cases} 1 & \text{if } \tilde{X}_j = \text{red} \\ 0 & \text{otherwise} \end{cases} \quad X_2 = \begin{cases} 1 & \text{if } \tilde{X}_j = \text{blue} \\ 0 & \text{otherwise} \end{cases} \quad X_3 = \begin{cases} 1 & \text{if } \tilde{X}_j = \text{pink} \\ 0 & \text{otherwise} \end{cases}$$

Remark: The linear regresion model is called "linear" because it is linear in parameters (not in inputs).

In Lecture 2, we found an estimate $\hat{\beta}$ of the vector of regression parameters β by first finding its optimal value with respect to the mean squared error and then estimating this optimal value using the training data.

$$MSE(\beta) = \mathbb{E}\left[(Y - \hat{Y})^2\right] = \mathbb{E}\left[(Y - X^T\beta)^2\right] \to \min \xRightarrow[\text{page 9}]{\text{Lecture 2}} \beta = \mathbb{E}[XX^T]^{-1}\mathbb{E}[XY]$$

$$\Downarrow$$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$$

where

$$\underset{N \times (p+1)}{\mathbf{X}} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & & & \\ 1 & x_{i1} & \dots & x_{ip} \\ \vdots & & & \\ 1 & x_{N1} & \dots & x_{Np} \end{bmatrix} \qquad y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

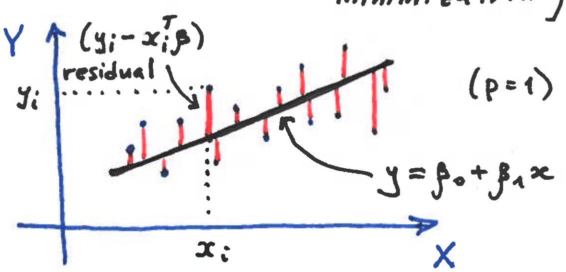
training data

It is instructive to see how β can be estimated directly from the data using the residual sum of squares:

$$\boxed{RSS(\beta) = \sum_{i=1}^{N} (y_i - x_i^T\beta)^2} \quad x_i^T = (1, x_{i1}, \dots, x_{ip})$$

Remark: RSS is the finite sample version of MSE.
(up to a multiplicative factor $\frac{1}{N}$ which is irrelevant for minimization)

Geometrically, $RSS(\beta)$ is an intuitive measure of the quality of the linear fit to the data:



$RSS(\beta_1) < RSS(\beta_2)$

$\Rightarrow$ linear fit corresponding to β_1 is better.

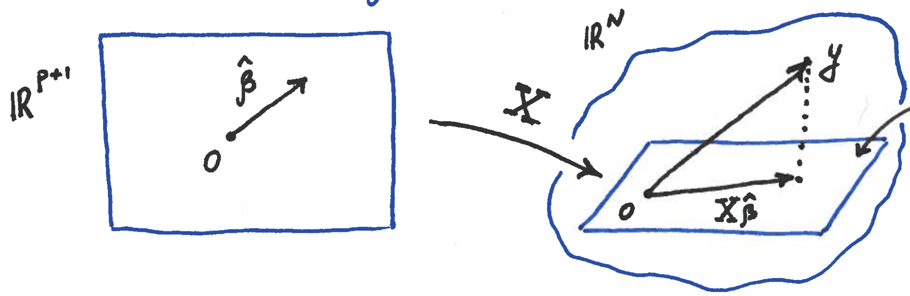
The value $\hat{\beta}$ of β that minimizes the RSS, $\hat{\beta} = \arg\min\limits_{\beta \in \mathbb{R}^{p+1}} RSS(\beta)$ has an important linear algebraic interpretation.

Let's rewrite RSS in the matrix form: $RSS(\beta) = \|y - X\beta\|^2$.

Therefore, $\hat{\beta} = \arg\min\limits_{\beta \in \mathbb{R}^{p+1}} RSS(\beta) = \arg\min\limits_{\beta \in \mathbb{R}^{p+1}} \|X\beta - y\|$ ← Euclidean norm

$\Rightarrow$ $\boxed{\hat{\beta} \text{ is a solution of linear system } X\beta = y \text{ in the least squares sense.}}$

This interpretation of $\hat{\beta}$ allows to conclude that a global minimizer of RSS exists.



$\operatorname{Im} X = \{X\beta, \beta \in \mathbb{R}^{p+1}\}$

$\operatorname{Im} X \subset \mathbb{R}^N$ subspace

→ hyperplane that goes through zero.

since linear systems always have least squares solutions

discussed in detail in ACM 104

To find $\hat{\beta}$, let's solve $\nabla_\beta RSS(\beta) = 0$.

Remark: In ACM 104, we discuss how to find least squares solutions of linear systems without Calculus

$$RSS(\beta) = (y - X\beta)^T(y - X\beta) = y^Ty - \beta^TX^Ty - y^TX\beta + \beta^TX^TX\beta$$

$$\nabla_\beta RSS(\beta) = -\nabla_\beta \beta^TX^Ty - \nabla_\beta y^TX\beta + \nabla_\beta \beta^TX^TX\beta$$

- $\beta^TX^Ty = \sum\limits_{j=0}^{p} \beta_j (X^Ty)_j \Rightarrow \nabla_\beta \beta^TX^Ty = X^Ty$

  since ∇f is a column vector.

- $y^TX\beta = \sum\limits_{j=0}^{p} (y^TX)_j \beta_j \Rightarrow \nabla_\beta y^TX\beta = (y^TX)^T = X^Ty$

- Let $X^TX = A$ ← $(p+1)\times(p+1)$ matrix $\Rightarrow \left(\nabla_\beta \beta^TX^TX\beta\right)_k = \dfrac{\partial}{\partial \beta_k}\left(\sum\limits_{i,j=0} a_{ij}\beta_i\beta_j\right)$

$$= \sum_{i,j=0}^{p} a_{ij} \underbrace{\frac{\partial \beta_i}{\partial \beta_k}}_{=\delta_{ik} = \begin{cases}1 & i=k\\ 0 & i\neq k\end{cases}} \cdot \beta_j + \sum_{i,j=0}^{p} a_{ij}\beta_i \underbrace{\frac{\partial \beta_j}{\partial \beta_k}}_{=\delta_{jk} = \begin{cases}1 & j=k\\ 0 & j\neq k\end{cases}} = \sum_{j=0}^{p} a_{kj}\beta_j + \sum_{i=0}^{p} \underbrace{a_{ik}}_{= a_{ki} \text{ since } A \text{ is symmetric}} \beta_i$$

$$= 2\sum_{j=0}^{p} a_{kj}\beta_j = 2(A\beta)_k = 2(X^TX\beta)_k \Rightarrow \nabla_\beta \beta^TX^TX\beta = 2X^TX\beta$$

So, $\nabla_\beta RSS(\beta) = -2X^Ty + 2X^TX\beta$.

Therefore, $\hat{\beta}$ is a solution of $\boxed{X^TX\beta = X^Ty}$ ← normal equation (linear system of equations on $\beta_0, \dots, \beta_p$)

The normal equation always has a solution.

The $(p+1)\times(p+1)$ matrix X^TX is the <u>Gram matrix</u> associated with $\underbrace{x^{(0)},\dots,x^{(p)}}_{\text{columns of } X}$.

It is always <u>positive semidefinite</u>, $X^TX \geq 0$.

$x^{(0)} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$

$x^{(j)}$ consists of N observations of input X_j.

① Suppose X^TX is <u>nonsingular</u> $\iff$ $X^TX > 0$ is positive definite

$\iff$ $x^{(0)},\dots,x^{(p)}$ are linearly indepen.

$\iff$ $\ker X = \{\beta : X\beta = 0\} = \{0\}$

$\iff$ $\operatorname{rank} X = p+1$

typically happens in applications.

Then the normal equation has the unique solution

$$\boxed{\hat\beta = (X^TX)^{-1}X^Ty}$$

To show that $\hat\beta$ is the global minimum, we need to check that the <u>Hessian</u> of RSS at $\hat\beta$ is positive definite, $H(RSS) > 0$.

$$H(RSS) = \left(\frac{\partial^2 RSS}{\partial\beta_s\,\partial\beta_k}\right) \leftarrow \text{matrix of second order partial derivatives.}$$

- $\dfrac{\partial RSS}{\partial\beta_k} = (\nabla_\beta RSS)_k = (2X^TX\beta - 2X^Ty)_k = 2\sum_{j=0}^{p}(X^TX)_{kj}\beta_j - (2X^Ty)_k$
- $\dfrac{\partial^2 RSS}{\partial\beta_s\,\partial\beta_k} = \dfrac{\partial}{\partial\beta_s}\left(2\sum_{j=0}^{p}(X^TX)_{kj}\beta_j\right) = 2\sum_{j=0}^{p}(X^TX)_{kj}\,\delta_{js} = 2(X^TX)_{ks}$

Therefore $H(RSS) = 2X^TX > 0 \Rightarrow \boxed{\hat\beta = (X^TX)^{-1}X^Ty}$ is indeed the global minimum of RSS.

It is called the <u>ordinary least squares</u> (<u>OLS</u>) estimate of the regression parameter $\beta = (\beta_0 \dots \beta_p)^T$

deserves to be boxed twice on the same page

② Suppose X^TX is <u>singular</u>.

does not happen often in applications

This can happen if

- $N < p+1 \Rightarrow \operatorname{rank} X < p+1$

$X = \boxed{}$ is "flat"

<u># observations $\leq$ # predictors</u>

- There are <u>redundant</u> <u>predictors</u>: one X_j is a linear combination of other predictors.

In this case, the normal equation has infinitely many solutions, all delivering the smallest possible value of RSS.

<u>Example</u>: Let $p=3$, $X_3 = X_1 + X_2 \Rightarrow y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 = \beta_0 + (\beta_1+a)X_1 + (\beta_2+a)X_2 + (\beta_3-a)X_3$.

Then if $(\hat\beta_0,\hat\beta_1,\hat\beta_2,\hat\beta_3)$ is a global minimizer of RSS, then so is $(\hat\beta_0,\hat\beta_1+a,\hat\beta_2+a,\hat\beta_3-a)$ for any $a\in\mathbb{R}$.

From linear algebra we know that the general solution of a linear system $Ax=b$ is $x^* + \ker A$, where x^* is a particular solution and $\ker A = \{x : Ax = 0\}$.

The general solution of the normal equation is then $\hat{\beta} = \hat{\beta}^* + \ker(X^TX)$

It can be shown that $\ker(X^TX) = \ker X$:

1) $\beta \in \ker X \Rightarrow X\beta = 0 \Rightarrow X^TX\beta = 0$
$\Rightarrow \beta \in \ker(X^TX)$

2) $\beta \in \ker(X^TX) \Rightarrow X^TX\beta = 0$
$\Rightarrow \beta^TX^TX\beta = 0 \Rightarrow (X\beta)^T \cdot X\beta = 0$
$\Rightarrow \|X\beta\|^2 = 0 \Rightarrow X\beta = 0$
$\Rightarrow \beta \in \ker X$

A particular solution $\hat{\beta}^*$ can be found using the notion of pseudoinverse matrix, which is based on the king of matrix decompositions: the singular value decomposition (SVD)

Proved in ACM 104

Thm Let A be an $m\times n$ matrix of rank $r>0$. Then it can be decomposed as follows:

$$\boxed{A = P\cdot\Sigma\cdot Q^T} \longleftarrow \text{SVD of } A$$

- P is $m\times r$ with orthonormal colums, $P^TP = I_r$
- Σ is $r\times r$ diagonal, $\Sigma = \text{diag}\underbrace{(\sigma_1,\dots,\sigma_r)}$ ← singular values of A
- Q is $n\times r$ with orthonormal colums, $Q^TQ = I_r$

Remark Since the 1st column of X is $\begin{bmatrix}1\\\vdots\\1\end{bmatrix}$, its rank ≥ 1.
$\Rightarrow$ SVD of X exists.

The SVD allows to generalize the notion of inverse matrix.

Def Let A be an $m\times n$ matrix and $A = P\Sigma Q^T$ be its SVD.
The pseudoinverse (aka Moore-Penrose inverse) is $\boxed{A^+ = Q\Sigma^{-1}P^T}$

It can be shown that

- If A is nonsingular $\Rightarrow A^+ = A^{-1}$
- If $\ker A = \{0\} \Rightarrow A^+ = (A^TA)^{-1}A^T$

} Proved in ACM 104 (direct check)

Let $X = P\Sigma Q^T$ be the SVD of the training inputs matrix.

Consider $\hat{\beta}^* = X^+y$. Then $\hat{\beta}^*$ is a solution of the normal equation $X^TX\beta = X^Ty$

Indeed:
- $X^TX\hat{\beta}^* = Q\Sigma \underbrace{P^TP}\Sigma \underbrace{Q^TQ}\Sigma^{-1}P^Ty = Q\Sigma P^Ty$
- $X^Ty = Q\Sigma P^Ty$

So, if X^TX is singular, then the general solution of the normal equation is

$$\boxed{\hat{\beta} = X^+y + \ker X}$$

Combining the two cases together: The general solution of the normal equation is

only if $\ker X = \{0\}$ $\swarrow$

$\ker X = \{0\}$ $\Updownarrow$

$$\hat{\beta} = \begin{cases} X^+y = (X^TX)^{-1}X^Ty & \text{if } X^TX \text{ is nonsing} \\ X^+y + \ker X & \text{if } X^TX \text{ is sing} \Leftrightarrow \ker X \neq \{0\} \end{cases}$$