In most settings, the number of observations (n) is much greater than the number of features (p). Note that at least one solution always exists because intuitively, we can always draw a line of best fit for a given set of data, but there may be multiple lines that are “equally good”. (Formal proof is beyond this course.) Let’s now revisit the interpretation for uniqueness of a solution at the end of the last lecture, but with the new notation of p instead of $p+1$ features.

The Least Squares estimate $\hat{\theta}$ is **unique** if and only if $\mathbb{X}$ is **full column rank**.

> Proof:
>
> - We know the solution to the normal equation $\mathbb{X}^T\mathbb{X}\hat{\theta} = \mathbb{X}^T\mathbb{Y}$ is the least square estimate that minimizes the squared loss.
> - $\hat{\theta}$ has a **unique** solution $\iff$ the square matrix $\mathbb{X}^T\mathbb{X}$ is **invertible** $\iff$ $\mathbb{X}^T\mathbb{X}$ is full rank.
>   - The **column rank** of a square matrix is the max number of linearly independent columns it contains.
>   - An n x n square matrix is deemed full column rank when all of its columns are linearly independent. That is, its rank would be equal to n.
>   - $\mathbb{X}^T\mathbb{X}$ has shape $p \times p$, and therefore has max rank p.
> - $rank(\mathbb{X}^T\mathbb{X})$ = $rank(\mathbb{X})$ (proof out of scope).
> - Therefore, $\mathbb{X}^T\mathbb{X}$ has rank p $\iff$ $\mathbb{X}$ has rank p $\iff$ $\mathbb{X}$ is full column rank.

Therefore, if $\mathbb{X}$ is not full column rank, we will not have unique estimates. This can happen for two major reasons.

1. If our design matrix $\mathbb{X}$ is “**wide**”:
   - If n < p, then we have way more features (columns) than observations (rows).
   - Then $rank(\mathbb{X})$ = min(n, p) < p, so $\hat{\theta}$ is not unique.
   - Typically we have n >> p so this is less of an issue.
2. If our design matrix $\mathbb{X}$ has features that are **linear combinations** of other features:
   - By definition, rank of $\mathbb{X}$ is number of linearly independent columns in $\mathbb{X}$.
   - Example: If “Width”, “Height”, and “Perimeter” are all columns,
     - Perimeter = 2 * Width + 2 * Height $\rightarrow$ $\mathbb{X}$ is not full rank.
   - Important with one-hot encoding (to discuss later).

Let’s now explore how to use the normal equations with a real-world dataset in the next section.

13.2 `sklearn`

13.2.1 Implementing Derived Formulas in Code

Throughout this lecture, we’ll refer to the `penguins` dataset.

▶ Code

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	Male
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	Female
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	Female
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	Female
5	Adelie	Torgersen	39.3	20.6	190.0	3650.0	Male

Our goal will be to predict the value of the `"bill_depth_mm"` for a particular penguin given its `"flipper_length_mm"` and `"body_mass_g"`. We’ll also add a bias column of all ones to represent the intercept term of our models.