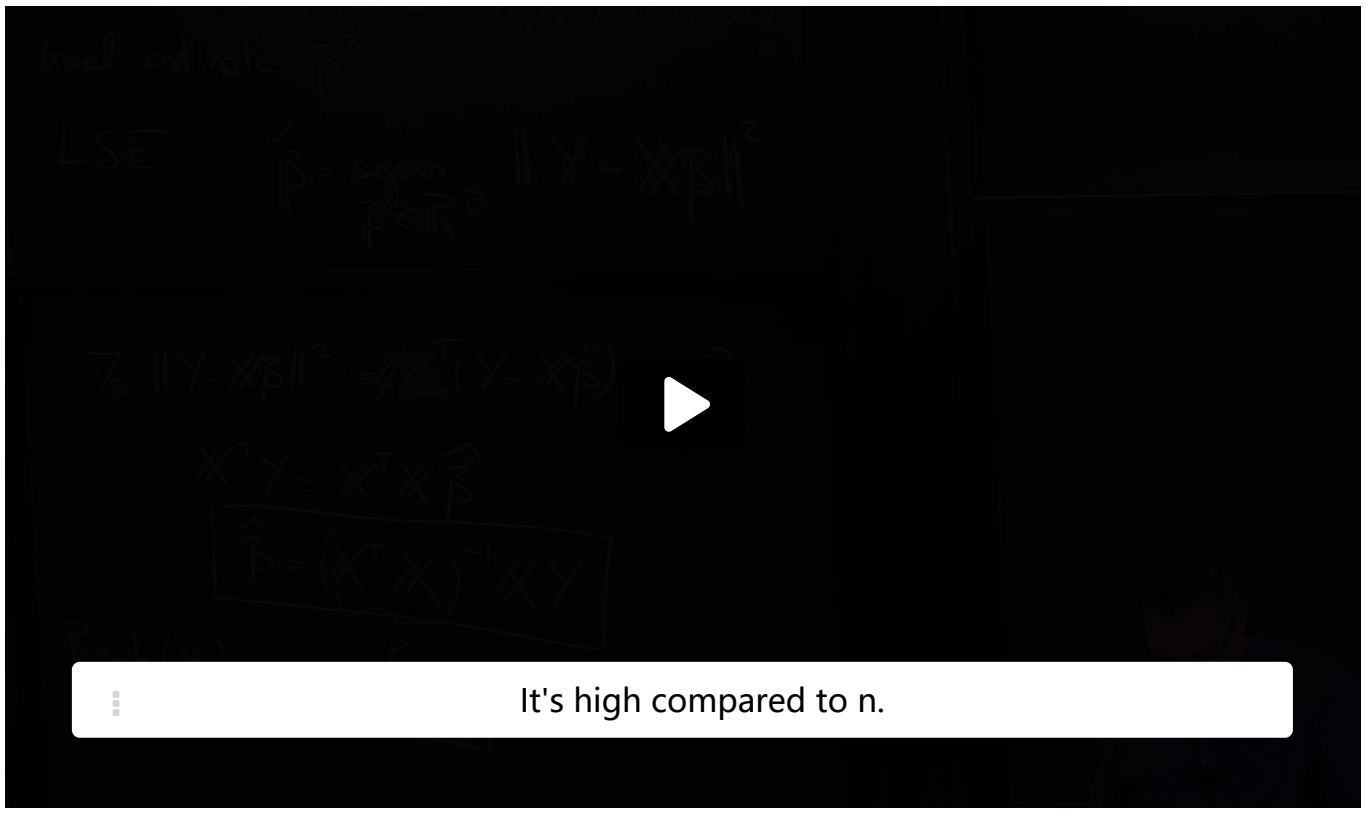
3. Multivariate Linear Regression
Review: Setup of Multivariate Linear Regression



8:14 / 8:14 1.0x

But when p is lower than n, this is
what's called the realm of high dimensional statistics,
and you need to do something different.
OK.
Now high dimensional statistics happens when n is equal to 2
and p is equal to 3.
OK?
It doesn't have to be.
It's high compared to n.

End of transcript. Skip to the start.

Video
Download video file

Transcripts
Download SubRip (.srt) file
Download Text (.txt) file

Review of Matrix Notation

6/6 points (graded)
Assume that we have collected data from one thousand patients from a hospital: $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_{1000}, Y_{1000})$, where each $\mathbf{X}_i = \begin{pmatrix} 1 \\ X^{(1)} \\ \vdots \\ X^{(4)} \end{pmatrix}$ is a column vector containing four different measurements: $X^{(1)}$ is age, $X^{(2)}$ is height, $X^{(3)}$ is heart rate,$X^{(4)}$ is blood pressure. We also include a constant covariate, $X^{(0)} = 1$ by convention to allow for an intercept β_0. On the other hand, Y_i represents the individual's cholesterol level. In order to understand how these four measurements correlate with cholesterol level, we perform a linear regression (regardless of whether this is truly a linear relationship).

Setting up the equation for linear regression, we get

$$\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{Y}$ is the column vector of dependent variables,

$$\mathbb{X} = \begin{pmatrix} \mathbf{X}_1^T \\ \hline \mathbf{X}_2^T \\ \hline \vdots \\ \hline \mathbf{X}_{1000}^T \end{pmatrix}$$

is the **design** matrix of covariates

and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots \varepsilon_{1000})^T$ is the (random) vector representing the error $\mathbf{Y} - \mathbb{X}\boldsymbol{\beta}$.

In the given setup, what are the dimensions of $\mathbf{Y}$ and $\boldsymbol{\varepsilon}$?

1000 ✔ **Answer:** 1000 rows × 1 ✔ **Answer:** 1 columns

What are the dimensions of $\mathbb{X}$?

1000 ✔ **Answer:** 1000 rows × 5 ✔ **Answer:** 5 columns

What are the dimensions of $\boldsymbol{\beta}$?

5 ✔ **Answer:** 5 rows × 1 ✔ **Answer:** 1 columns

Solution:

Pictorially, here is what the matrix equation $\mathbf{Y} = \mathbb{X}\beta + \varepsilon$ looks like:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{999} \\ Y_{1000} \end{pmatrix} = \begin{pmatrix} -\mathbf{X}_1^T - \\ -\mathbf{X}_2^T - \\ \vdots \\ -\mathbf{X}_{999}^T - \\ -\mathbf{X}_{1000}^T - \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_4 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{999} \\ \varepsilon_{1000} \end{pmatrix}$$

As stated in the problem, $\mathbf{Y}$ and ε are column vectors with 1000 entries. Such a vector has 1000 rows and 1 column. To make the dimensions match up, $\mathbb{X}$ must have 1000 rows and 5 columns, since each $\mathbf{X}$ is a column vector of dimension 5. The vector β must have as many rows as $\mathbb{X}$ has columns, so it is a column vector with $\mathbf{5}$ rows.

Submit You have used 2 of 3 attempts

Answers are displayed within the problem

The Least-Squares Estimator

0/1 point (graded)

When computing the least-squares estimator, we are computing some $\boldsymbol{\beta}$ which minimizes the error

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{Y} - \mathbb{X}\boldsymbol{\beta}\|_2^2$$

where $\|v\|_2$ is the Euclidean norm.

Let n be the number of samples and let each $\mathbf{X}_i$ be p-dimensional. (For example, n might be the number of patients, and $p-1$ is the number of covariates that we are trying to study - e.g. height, weight, age and blood pressure as in the previous problem.)

Recall that by employing the same technique of computing the gradient (with respect to the components of $\boldsymbol{\beta}$) and setting it equal to zero, we can show that $\hat{\boldsymbol{\beta}}$ must satisfy the **score equation**

$$\mathbb{X}^T\mathbb{X}\hat{\boldsymbol{\beta}} = \mathbb{X}^T\mathbf{Y}.$$

We would like to isolate $\hat{\boldsymbol{\beta}}$ by multiplying by $(\mathbb{X}^T\mathbb{X})^{-1}$ from the left. Which of the following conditions, each on its own, **guarantees** that $\mathbb{X}^T\mathbb{X}$ is invertible? Choose all that apply.

Hint: Use the fact from linear algebra that $\mathbf{rank}\,(A^TA) = \mathbf{rank}\,(A)$. *What are the dimensions of* $\mathbb{X}^T\mathbb{X}$*?*

- [x] There are at least as many observations as covariates (i.e. $n \geq p$).
- [ ] There are at least as many covariates as observations (i.e. $n \leq p$).
- [ ] $\mathbb{X}$ has rank n, where n is the number of samples.
- [x] $\mathbb{X}$ has rank p, where p is the number of covariates. ✔
- [ ] There are at least p distinct samples amongst the X's, so that $\mathbb{X}$ has at least p distinct rows.
- [ ] There are at least p distinct values amongst the Y.

✖

Solution:

The only correct choice is "$\mathbb{X}$ **has rank** p". We examine each choice in order. A key observation here is to notice that $\mathbb{X}^T\mathbb{X}$ is a $p \times p$ matrix.

- "**There are at least as many observations as covariates ($n \geq p$).**" The condition $n \geq p$ is not enough to guarantee invertibility. Consider the scenario where $\mathbf{X}_1 = \cdots = \mathbf{X}_n$.
- "**There are at least as many covariates as observations ($n \leq p$).**" This is also incorrect, via the same exact argument as in the first choice.
- "$\mathbb{X}$ **has rank** n." This is not sufficient. If there are fewer observations than covariates ($n < p$), then $\mathbf{rank}\left(\mathbb{X}^T\mathbb{X}\right) = \mathbf{rank}\left(\mathbb{X}\right) = n < p$, which means that the $p \times p$ matrix $\mathbb{X}^T\mathbb{X}$ is singular.
- "$\mathbb{X}$ **has rank** p." This gives $\mathbf{rank}\left(\mathbb{X}^T\mathbb{X}\right) = p$, so it has full rank. This is equivalent to $\mathbb{X}^T\mathbb{X}$ being invertible.
- "**There are at least p distinct observations of X.**" Consider the scenario where $\mathbf{X}_1, \ldots, \mathbf{X}_n$ all lie on the same line (span a subspace of dimension 1), and $p \geq 2$. Then $\mathbf{rank}\left(\mathbb{X}\right) = 1$, which certainly does not work.
- "**There are at least p distinct observations of Y.**" The number of Y's has no bearing on the invertibility of $\mathbb{X}^T\mathbb{X}$.

Submit You have used 3 of 3 attempts

Answers are displayed within the problem

Uniqueness of the LSE

1/1 point (graded) 满秩的条件

An $n \times n$ matrix $\mathbf{M}$ has "full rank"i.e. $\mathbf{rank}\left(\mathbf{M}\right) = n$ if and only if its determinant is non-zero.

Each choice below specifies a collection of $\mathbf{x}$'s in $\mathbb{R}^2$ and y's in $\mathbb{R}$, where each $\mathbf{x}$ can be written as $\mathbf{x} = (x^{(1)}, x^{(2)})$. Which one admits a **unique** least-squares estimator $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ for the linear model $y = \beta_0 + \beta_1 x^{(1)} + \beta_2 x^{(2)}$?

- $(x_1^{(1)}, x_1^{(2)}, y_1) = (0,1,10), (x_2^{(1)}, x_2^{(2)}, y_2) = (0,1,9), (x_3^{(1)}, x_3^{(2)}, y_3) = (0,1,-3)$
- $(x_1^{(1)}, x_1^{(2)}, y_1) = (0,1,10), (x_2^{(1)}, x_2^{(2)}, y_2) = (0,3,9)$
- $(x_1^{(1)}, x_1^{(2)}, y_1) = (0,1,10), (x_2^{(1)}, x_2^{(2)}, y_2) = (0,1,9), (x_3^{(1)}, x_3^{(2)}, y_3) = (0,0,7), (x_4^{(1)}, x_4^{(2)}, y_4) = (3,0,-1)$ ✔
- $(x_1^{(1)}, x_1^{(2)}, y_1) = (0,1,10), (x_2^{(1)}, x_2^{(2)}, y_2) = (0,3,9), (x_3^{(1)}, x_3^{(2)}, y_3) = (0,-1,14)$

Solution:

The third option is the correct choice. Recall that the design matrix (with intercept β_0 accounted for) is

$$\mathbf{X} = \begin{pmatrix} 1 & x_1^{(1)} & x_2^{(2)} \\ & \vdots & \\ 1 & x_n^{(1)} & x_n^{(2)} \end{pmatrix}.$$

There exists a unique least-squares estimator $\hat{\beta}$ if $\boldsymbol{rank}\,(\mathbf{X}) = 3$, since β has three components. The first and fourth choices each yield square 3×3 matrices $\mathbf{X}$ with zero determinant, so their respective ranks are strictly smaller than 3. In the second choice, the resulting matrix has size 2×3, so its rank can be at most 2. In contrast, the third choice yields a 4×3 matrix with full rank (rank 3).

Submit You have used 1 of 3 attempts

Answers are displayed within the problem

Uniqueness of the LSE, continued

0/1 point (graded)

所以要降维

Let $p = 5, n = 1000$.

有p个观测值，观测1000次

但是rank3，也就是有只有3个变量的信息是完全不同的

Assume that the design matrix $\mathbb{X}$ has rank 3. Which of the following choices correctly characterizes the space of all estimators $\boldsymbol{\beta}$ that satisfy the score equation, $\mathbb{X}^T\mathbb{X}\boldsymbol{\beta} = \mathbb{X}^T\mathbf{Y}$?

- A unique least-squares estimator $\hat{\boldsymbol{\beta}}$.
- A finite (larger than one) collection of estimators $\hat{\boldsymbol{\beta}}$. ✖
- An infinite collection of estimators $\hat{\boldsymbol{\beta}}$. ✔

Solution:

The correct answer is "**An infinite collection of estimators $\hat{\boldsymbol{\beta}}$.**"
In general, if $\mathbf{rank}\,(\mathbb{X}^T\mathbb{X}) = \mathbf{rank}\,(\mathbb{X}) = \boldsymbol{k}$ such that $\boldsymbol{k < p}$, then $\mathbb{X}^T\mathbb{X}$ has a linear subspace of dimension $(\boldsymbol{p - k})$ such that $\mathbb{X}^T\mathbb{X}\boldsymbol{v} = 0$, which is called the **kernel** or the **nullspace** of the matrix. Therefore, if $\hat{\boldsymbol{\beta}}$ is any vector such that $\mathbb{X}^T\mathbb{X}\hat{\boldsymbol{\beta}} = \mathbb{X}^T\mathbf{Y}$, then so is $\hat{\boldsymbol{\beta}} + \boldsymbol{v}$:

$$\mathbb{X}^T\mathbb{X}\,(\hat{\boldsymbol{\beta}} + \boldsymbol{v}) = \mathbb{X}^T\mathbb{X}\hat{\boldsymbol{\beta}} + \mathbb{X}^T\mathbb{X}\boldsymbol{v} = \mathbb{X}^T\mathbf{Y} + 0.$$

Note: An illustrative example is the single-variable model, $\boldsymbol{Y = a + bX + \varepsilon}$, which required the assumption that $\mathbf{Var}\,(X) \neq 0$. As it turns out, this is a special case of the condition $\mathbf{rank}\,(\mathbb{X}^T\mathbb{X}) = 2$.

需要满秩

Observe that, for this case ($\boldsymbol{p = 2}$ with intercept $\boldsymbol{a}$), we get

$$\mathbb{X}^T\mathbb{X} = \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}.$$

The determinant of this product is equal to $n\sum x_i^2 - (\sum x_i) = n^2\hat{\sigma}_x^2$, where $\hat{\sigma}^2$ is the empirical variance of the $\boldsymbol{x}$'s. The condition for matrices being invertible is equivalent to the determinant being nonzero. From this point of view, it should come as no surprise that we obtained an infinite famliy of best-fitting lines whenever we only had one value of $\boldsymbol{x}$ to work with. As a reminder, any line that crossed $(\boldsymbol{x}, \bar{\boldsymbol{y}})$ yielded the same error - this is the one-dimensional space parametrized by $(\boldsymbol{a}, \boldsymbol{b})$.

Submit You have used 2 of 2 attempts

Answers are displayed within the problem