Ordinary Least Squares

Using linear algebra to generalize the simple linear regression model.

Recap: Simple Linear Regression

Simple Linear Regression

In the last lecture, we re-introduced the simple linear regression model.

$$\hat{y} = f_\theta(x) = \theta_0 + \theta_1 x$$

- Our loss function was squared loss, and so our objective function was mean squared error (MSE).
- To solve for the **optimal parameters** (also known as coefficients or weights), we minimized MSE by hand using calculus.

$$\hat{\theta}_1 = r\frac{\sigma_y}{\sigma_x} \qquad \hat{\theta}_0 = \bar{y} - \hat{\theta}_1\bar{x}$$

- These are parameter estimates.
- We also looked at r, the correlation coefficient, and its relation to the optimal coefficients.

Multiple Regression

We also extended this model to account for multiple **features**.

$$\hat{y} = f_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_p x_p = \theta_0 + \sum_{j=1}^{p} \theta_j x_j$$

Each x_j is a separate feature.

- We learned about multiple R^2, an extension of the correlation coefficient r to multiple features.
- Our loss function yet again was squared loss.
- We didn't minimize MSE by hand – we abstracted away the process of determining the theta values.
- Lastly, we introduced RMSE as a method of comparing model performance.

Today, we will learn how to find the optimal parameters ("thetas") for multiple regression, for any number of features.

Agenda

- Use vector dot products to define the multiple regression model.
- Formulate the problem statement using vector norms.
- Use a geometric derivation to solve for the optimal θ (which is now a vector).
- Explore properties of residuals.
- Understand when a unique solution exists.

Linear algebra formulation

Dot products

The dot product of two vectors $\vec{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$ $\vec{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$ and is defined as follows:

$$\vec{a} \cdot \vec{b} = a_1 b_1 + a_2 b_2 + \ldots + a_n b_n$$

$$\vec{a} \cdot \vec{b} = \vec{a}^T \vec{b}$$

- An alternate way of writing a dot product: .
  - This is the form we will use primarily moving forward.
- The dot product between two vectors is a **scalar**, not another vector.
- The dot product is only defined for two vectors of the same length.
- The dot product is a special case of the inner product.

Multiple regression as a dot product

We previously stated that the multiple regression model was of the form

$$\hat{y} = f_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_p x_p = \theta_0 + \sum_{j=1}^{p} \theta_j x_j$$

This can be restated as a dot product between two vectors.

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{bmatrix} \quad x = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

$$\hat{y} = x^T \theta$$

Even though they don't have arrows on top of them, *x* and θ are still **vectors**!

scalar

Design matrix

Our mean squared error involves all observations at once, so it may be valuable to express our model in terms of all observations, instead of just one observation. One step in that process is stacking all of our observations together into a **design matrix**. With *n* observations:

$$\mathbb{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & x_{23} & \cdots & x_{2p} \\ 1 & x_{31} & x_{32} & x_{33} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \cdots & x_{np} \end{bmatrix}$$

Rows correspond to **observations**.
e.g. all features for data point 3

Columns correspond to **features**.
e.g. feature 1, for all data points

Example design matrix

	Bias	FG	AST	PTS
0	1	1.8	0.6	5.3
1	1	0.4	0.8	1.7
2	1	1.1	1.9	3.2
3	1	6.0	1.6	13.9
4	1	3.4	2.2	8.9
...	...	...	...	...
703	1	4.0	0.8	11.5
704	1	3.1	0.9	7.8
705	1	3.6	1.1	8.9
706	1	3.4	0.8	8.5
707	1	3.8	1.5	9.4

708 rows × 4 columns

Here, n = 708, and p = 3.
- Each of the 4 columns corresponds to a different feature.
- Each of the 708 rows corresponds to a different data point.

Multiple regression as a matrix multiplication

This allows us to express our linear model on our **entire dataset** (not just one observation) as

$$\hat{\mathbb{Y}} = \mathbb{X}\boldsymbol{\theta}$$

Vector with dimensions **n x 1**.

Matrix with dimensions **n x (p + 1)**.

Vector with dimensions **(p + 1) x 1**.

Note: This means that $\mathbb{Y}$ represents an n-length vector containing all of our true y values.

Multiple regression as a matrix multiplication

$$
\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \vdots \\ \hat{y}_n \end{bmatrix}
=
\begin{bmatrix}
1 & x_{11} & x_{12} & x_{13} & \ldots & x_{1p} \\
1 & x_{21} & x_{22} & x_{23} & \ldots & x_{2p} \\
1 & x_{31} & x_{32} & x_{33} & \ldots & x_{3p} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & x_{n1} & x_{n2} & x_{n3} & \ldots & x_{np}
\end{bmatrix}
\begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \\ \vdots \\ \theta_p \end{bmatrix}
$$

Moving forward:

- $\mathbb{X}_{i,:}$ or just simply $\mathbb{X}_i$ will represent **row i** of our design matrix. **Rows are observations / data points**.
- $\mathbb{X}_{:,j}$ will represent **column j** of our design matrix. **Columns are features**.

Multiple regression as a matrix multiplication

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & x_{23} & \ldots & x_{2p} \\ 1 & x_{31} & x_{32} & x_{33} & \ldots & x_{3p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \ldots & x_{np} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \\ \vdots \\ \theta_p \end{bmatrix}$$

For instance, to compute the predicted output for the second observation:

$$\hat{y}_2 = \theta_0 + \theta_1 x_{21} + \theta_2 x_{22} + \theta_3 x_{23} + \ldots + \theta_p x_{2p}$$

Or, more compactly:

$$\hat{y}_2 = \mathbb{X}_2^T \theta$$

X_2 is a vector, not a matrix. That is why it is transposed.

Example

Consider the following design matrix and value of θ.

Bias	FG	AST	PTS
1	1.8	0.6	5.3
1	0.4	0.8	1.7
1	1.1	1.9	3.2
1	6.0	1.6	13.9
1	3.4	2.2	8.9
...	...	...	...
1	4.0	0.8	11.5

$$\theta = \begin{bmatrix} -5 \\ 2 \\ -1 \\ 3 \end{bmatrix}$$

The predicted **response** (output) for the second observation:

$$\hat{y}_2 = \begin{bmatrix} 1 & 0.4 & 0.8 & 1.7 \end{bmatrix} \begin{bmatrix} -5 \\ 2 \\ -1 \\ 3 \end{bmatrix}$$

$$= 1(-5) + 0.4(2) + 0.8(-1) + 1.7(3)$$
$$= 0.1$$

Example

Consider the following design matrix and value of θ.

The predicted **response vector**:

Bias	FG	AST	PTS
1	1.8	0.6	5.3
1	0.4	0.8	1.7
1	1.1	1.9	3.2
1	6.0	1.6	13.9
1	3.4	2.2	8.9
...	...	...	...
1	4.0	0.8	11.5

$$\theta = \begin{bmatrix} -5 \\ 2 \\ -1 \\ 3 \end{bmatrix}$$

$$\hat{\mathbb{Y}} = \begin{bmatrix} 1 & 1.8 & 0.6 & 5.3 \\ 1 & 0.4 & 0.8 & 1.7 \\ 1 & 1.1 & 1.9 & 3.2 \\ 1 & 6.0 & 1.6 & 13.9 \\ 1 & 3.4 & 2.2 & 8.9 \\ ... & ... & ... & ... \\ 1 & 4.0 & 0.8 & 11.5 \end{bmatrix} \begin{bmatrix} -5 \\ 2 \\ -1 \\ 3 \end{bmatrix} = \begin{bmatrix} 13.9 \\ 0.1 \\ 4.9 \\ 47.1 \\ 26.3 \\ ... \\ 36.7 \end{bmatrix}$$

Summary of notation

When looking at a **single observation**, our model is

$$\hat{y} = f_\theta(x) = x^T \theta$$

- x is a **vector** of size p + 1.
- $\hat{y}$ is a **scalar**.
- θ is a **vector** of size p + 1.

When looking at **multiple observations**, our model is

$$\hat{\mathbb{Y}} = \mathbb{X}\theta$$

- $\mathbb{X}$ is a **matrix** of size n x (p + 1).
- $\hat{\mathbb{Y}}$ is a **vector** of size n (i.e. $\hat{\mathbb{Y}} \in \mathbb{R}^n$).
- θ is a **vector** of size p + 1.

In many settings, we assume that we have only p (and not p + 1) columns. One of those p columns may be "1" for each observation.

Problem statement

Vector norms

- The two norms we have learned are the *L1* and *L2* norms.

Consider the vector:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}$$

L2 vector norm:

$$||x||_2 = \sqrt{x_1^2 + x_2^2 + x_3^2 + \ldots + x_n^2} = \sqrt{\sum_{i=1}^{n} x_i^2}$$

L1 vector norm:

$$||x||_1 = |x_1| + |x_2| + |x_3| + \ldots + |x_n| = \sum_{i=1}^{n} |x_i|$$

L_2 vector norm

- The L_2 vector norm can be thought of as the "length" of a vector.

$$||x||_2 = \sqrt{x_1^2 + x_2^2 + x_3^2 + \ldots + x_n^2} = \sqrt{\sum_{i=1}^{n} x_i^2}$$

- The "distance" between two vectors is the L_2 norm of their difference.
  - For instance, if *a* and *b* are two vectors of the same length, then their distance is

$$||a - b||_2$$

- Note, the square of the L_2 norm of a vector is the sum of the squares of the vector's elements:

$$||x||_2^2 = \sum_{i=1}^{n} x_i^2$$

Residuals

Residuals are defined as being the difference between an actual and predicted value, in the regression context.

- We use the letter **e** to denote residuals. The residual **i** is

$$e_i = y_i - \hat{y}_i$$

- The MSE of a model is equal to the mean of the squares of its residuals:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} e_i^2$$

The red dotted lines represent residuals.
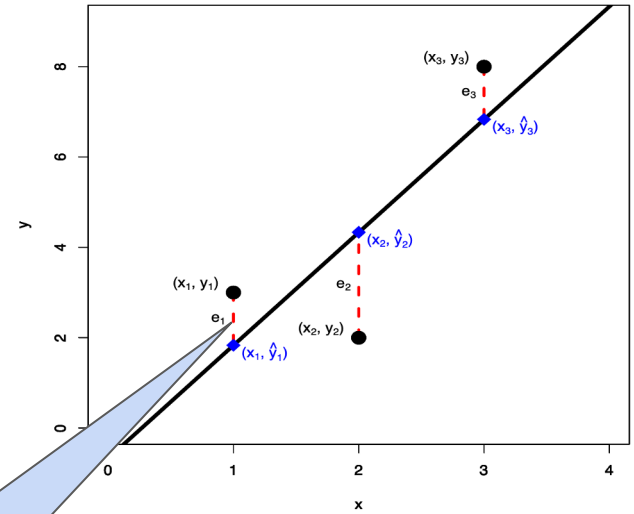
Residual vector

We can stack all *n* residuals into a vector, called the residual vector, *e*:

$$e = \mathbb{Y} - \hat{\mathbb{Y}} = \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_n - \hat{y}_n \end{bmatrix}$$

The residual vector is the "difference" between the two vectors containing our true *y* values and predicted *y* values.

Mean squared error, again

We are choosing our loss to be squared loss. This means, the average loss across our dataset is mean squared error.

$$R(\theta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbb{X}_i^T \theta)^2$$

We can write this in terms of the norm of the residual vector!

$$R(\theta) = \frac{1}{n} ||\mathbb{Y} - \hat{\mathbb{Y}}||_2^2 = \frac{1}{n} ||\mathbb{Y} - \mathbb{X}\theta||_2^2$$

This is the residual vector!

Optimization procedure

As we did in the last lecture, we note that the value of theta that minimizes $\frac{1}{n}||\mathbb{Y} - \mathbb{X}\theta||_2^2$ is the same value that minimizes

$$R(\theta) = ||\mathbb{Y} - \mathbb{X}\theta||_2^2$$

Therefore, our goal is to find the value of θ that minimizes the squared L_2 norm of the residual vector. In other words, we want the "distance" between $\mathbb{Y}$ and $\hat{\mathbb{Y}}$ to be minimized.

There are two ways we can determine the optimal $\hat{\theta}$ here.

- Using calculus, like we've done earlier.
- Using a **geometric argument**. This is what we'll do here.

Geometric derivation

A linear combination of columns

$$\hat{\mathbb{Y}} = \mathbb{X}\,\theta$$

$$\begin{bmatrix} \hat{\mathbb{Y}} \end{bmatrix} = \begin{bmatrix} \mathbb{X}_{:,1} & \mathbb{X}_{:,2} \end{bmatrix} \begin{bmatrix} \theta \end{bmatrix} = \theta_1 \mathbb{X}_{:,1} + \theta_2 \mathbb{X}_{:,2}$$

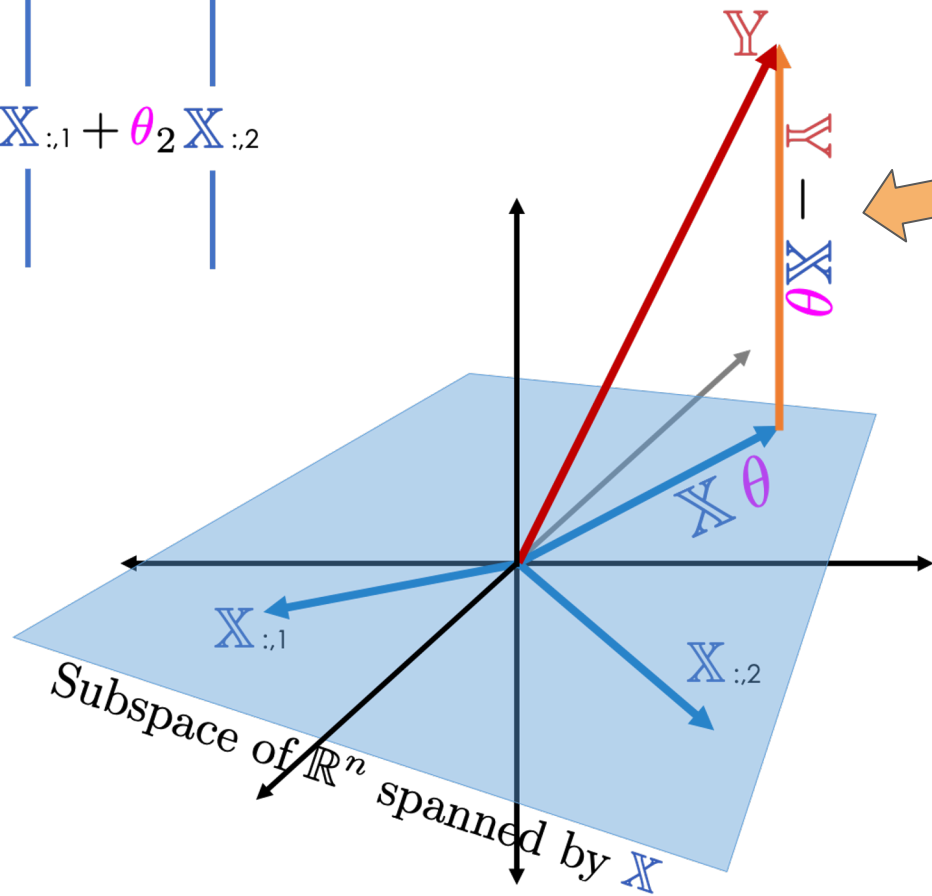The linear model represents $\hat{\mathbb{Y}}$ as a linear combination of the columns of $\mathbb{X}$.

Span

Our prediction is a linear combination of the columns of X.

- The set of all possible linear combinations of the columns of X is called the **span** of the columns of X (denoted $\text{span}(\mathbb{X})$).
  - Also called the **column space**.
- Intuitively, this is all of the vectors you can "reach" using the columns of X.
- Since each column of X has length n, $\text{span}(\mathbb{X})$ is a subspace $\mathbb{R}^n$.
- Our goal is to find the vector in $\text{span}(\mathbb{X})$ that is closest to $\mathbb{Y}$.

$$\begin{pmatrix} | \\ \hat{\mathbb{Y}} \\ | \end{pmatrix} = \theta_1 \mathbb{X}_{:,1} + \theta_2 \mathbb{X}_{:,2}$$

$\mathbb{Y}$

$\hat{\mathbb{Y}}$

$\theta \mathbb{X}$

$\mathbb{X}\theta$

$\mathbb{X}_{:,1}$

$\mathbb{X}_{:,2}$
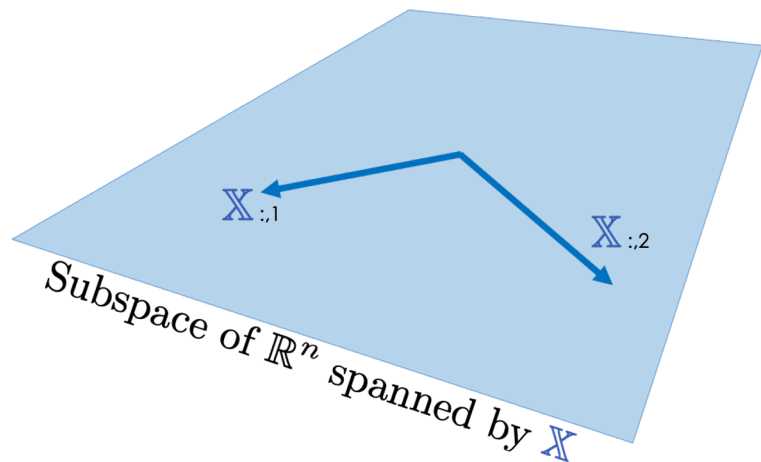
Subspace of $\mathbb{R}^n$ spanned by $\mathbb{X}$

Recall, this is the residual vector, $e = \mathbb{Y} - \hat{\mathbb{Y}}$.

Our goal is to minimize the L_2 norm of the residual vector, i.e. we want our predictions to be "as close" to our true y values as possible.
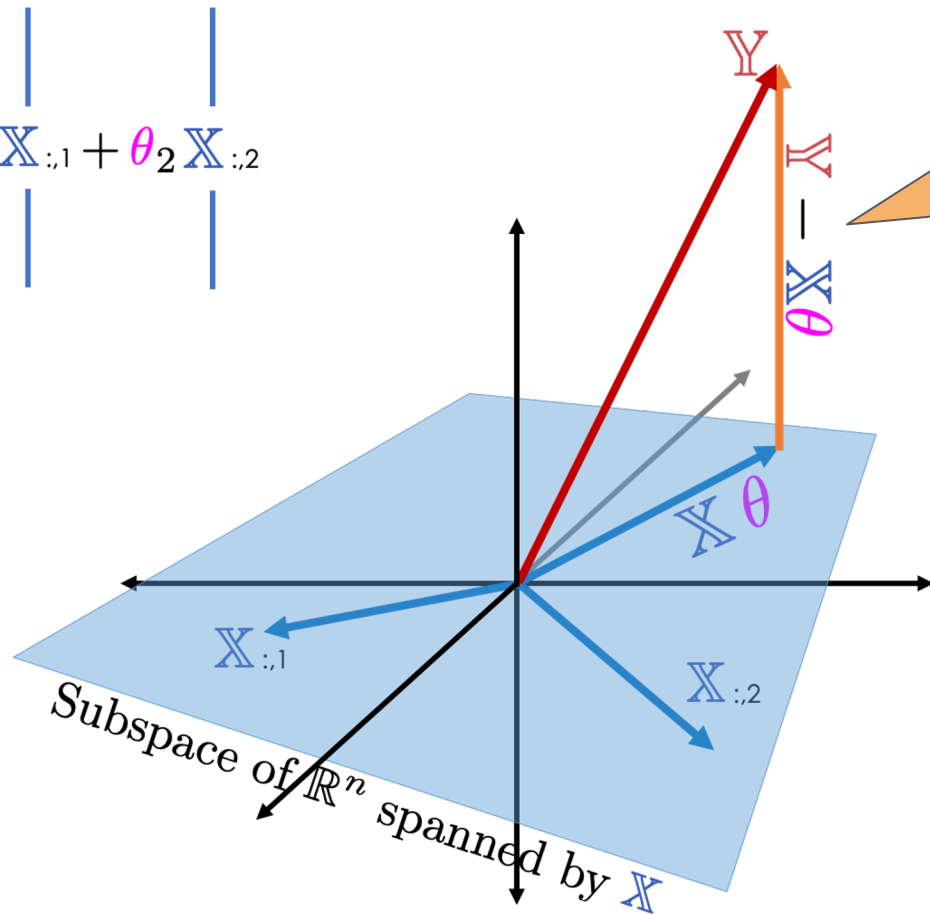
$$\begin{pmatrix} \mid \\ \hat{\mathbb{Y}} \\ \mid \end{pmatrix} = \theta_1 \mathbb{X}_{:,1} + \theta_2 \mathbb{X}_{:,2}$$

$\mathbb{Y}$

$\hat{\mathbb{Y}}$

$\mathbb{Y} - \mathbb{X}\theta$

θ

$\mathbb{X}\theta$

How do we minimize this distance – the norm of the residual vector (squared)?

The vector in span(X) that is closest to Y is the **orthogonal projection** of Y onto span(X).

$\mathbb{X}_{:,1}$

$\mathbb{X}_{:,2}$

Subspace of $\mathbb{R}^n$ spanned by $\mathbb{X}$

We will not prove this: see Khan Academy.

$$\begin{bmatrix} | \\ \hat{\mathbb{Y}} \\ | \end{bmatrix} = \theta_1 \mathbb{X}_{:,1} + \theta_2 \mathbb{X}_{:,2}$$
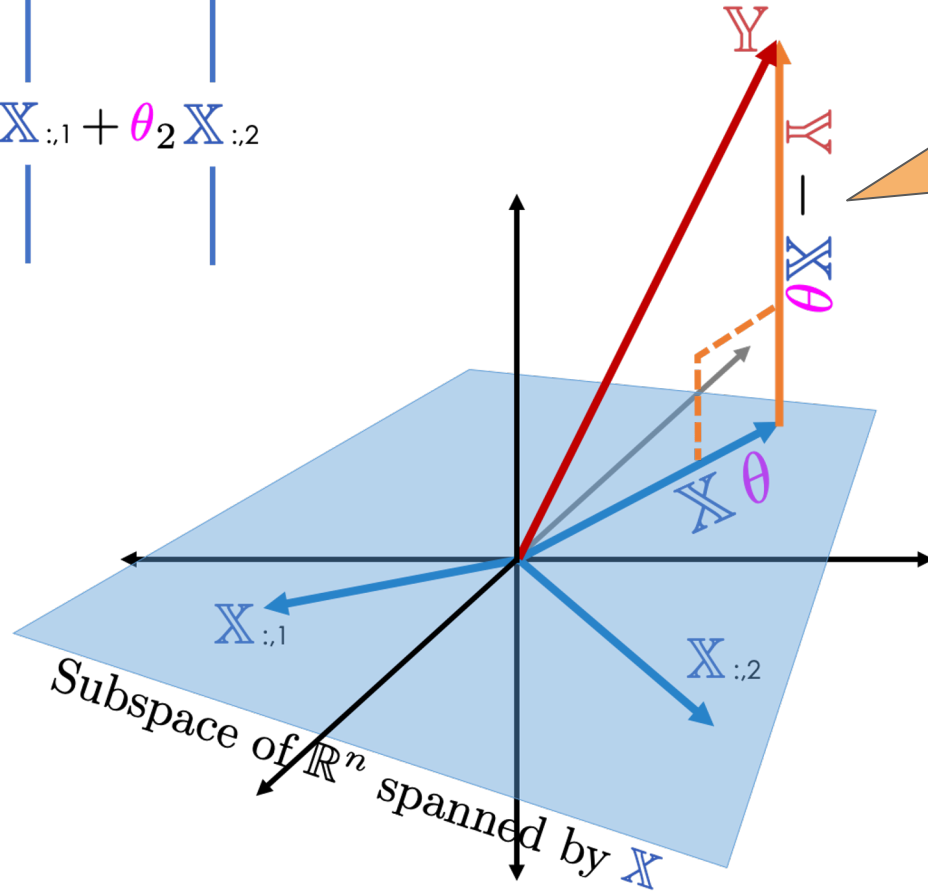
How do we minimize this distance – the norm of the residual vector (squared)?

The vector in span(X) that is closest to Y is the **orthogonal projection** of Y onto span(X).

Thus, we should choose the θ that makes the residual vector **orthogonal** to span(X).

$\mathbb{Y}$

$\mathbb{Y} - \theta \mathbb{X}$

$\mathbb{X}\theta$

Subspace of $\mathbb{R}^n$ spanned by $\mathbb{X}$

$\mathbb{X}_{:,1}$

$\mathbb{X}_{:,2}$

Orthogonality

We say two vectors are **orthogonal** if and only if their dot product is 0.

- This is a generalization of the notion of two vectors in 2D being perpendicular.

$$a^T b = 0 \iff a, b \text{ are orthogonal}$$

Suppose a vector is orthogonal to the span of the columns of a matrix.

- This is true if and only if it is orthogonal to each column individually. ([proof](#))
- Let $M \in \mathbb{R}^{n \times d}, v \in \mathbb{R}^{n \times 1}$. Suppose v is orthogonal to the span of the columns of M.
  Then:

$$M = \begin{bmatrix} | & | & & | \\ m_1 & m_2 & \cdots & m_d \\ | & | & & | \end{bmatrix}$$

$$\begin{aligned} m_1^T v &= 0 \\ m_2^T v &= 0 \\ &\vdots \\ m_d^T v &= 0 \end{aligned} \implies M^T v = \vec{0}$$

Let $M \in \mathbb{R}^{n \times d}, v \in \mathbb{R}^{n \times 1}$.

Suppose $\textbf{\textit{v}}$ is orthogonal to the span of the columns of $\textbf{\textit{M}}$.

$$M = \begin{bmatrix} | & | & & | \\ m_1 & m_2 & \cdots & m_d \\ | & | & & | \end{bmatrix}$$

$$m_1^T v = 0$$
$$m_2^T v = 0$$
$$\vdots$$
$$m_d^T v = 0$$

$\Longrightarrow$

$$M^T v = \begin{bmatrix} m_1^T v \\ m_2^T v \\ \vdots \\ m_d^T v \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \vec{0}$$

$\textbf{\textit{v}}$ is orthogonal to each column of $\textbf{\textit{M}}$ separately.
(Note, each column of $\textbf{\textit{M}}$ has length $\textbf{\textit{n}}$, and $\textbf{\textit{v}}$ also has length $\textbf{\textit{n}}$).

This product encapsulates all $\textbf{\textit{d}}$ of the equations on the left into a single equation. The quantity on the right is the **zero vector** ($\textbf{\textit{d}}$-length vector full of 0s).

Residuals are orthogonal to the span of X

We want the θ such that the residual vector is orthogonal to $\text{span}(\mathbb{X})$.

By the definition of orthogonality: $\mathbb{X}^T \left(\mathbb{Y} - \mathbb{X}\hat{\theta} \right) = 0$ ← Still the zero vector!

Rearranging: $\mathbb{X}^T\mathbb{Y} - \mathbb{X}^T\mathbb{X}\hat{\theta} = 0$

The **normal equation**: $\mathbb{X}^T\mathbb{X}\hat{\theta} = \mathbb{X}^T\mathbb{Y}$

Assuming $\mathbb{X}^T\mathbb{X}$ is full rank: $\hat{\theta} = \left(\mathbb{X}^T\mathbb{X} \right)^{-1} \mathbb{X}^T\mathbb{Y}$

$$\hat{\theta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$$

least squares estimate for θ.

Summary

1. **Choose a model**. We chose the multiple linear regression model, formulated using a matrix.

$$\hat{\mathbb{Y}} = \mathbb{X}\boldsymbol{\theta}$$

2. **Choose a loss function**. We chose squared loss, and hence our average loss was

$$R(\theta) = \frac{1}{n}||\mathbb{Y} - \hat{\mathbb{Y}}||_2^2 = \frac{1}{n}||\mathbb{Y} - \mathbb{X}\theta||_2^2$$

3. **Minimize average loss to determine optimal model parameters**. Done!

$$\hat{\theta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$$

Residuals

Residual plots

Residual plots can tell us about the quality of our model.

- In the **simple linear regression** case, with only one independent variable, we typically plot residuals vs. x.
- More generally, a residual plot is of **residuals vs. fitted values**.

Properties:

- A good residual plot has no pattern. This means that our model represents the relationship in the data well.
  - If you see a curve, it is a sign that transformations or additional variables could help.
- A good residual plot also has a similar vertical spread throughout the entire plot.
  - If this is not the case, the accuracy of the predictions is not reliable.

Residual vs. x for dataset 1

No pattern, even spread.

Residual vs. x for dataset 2
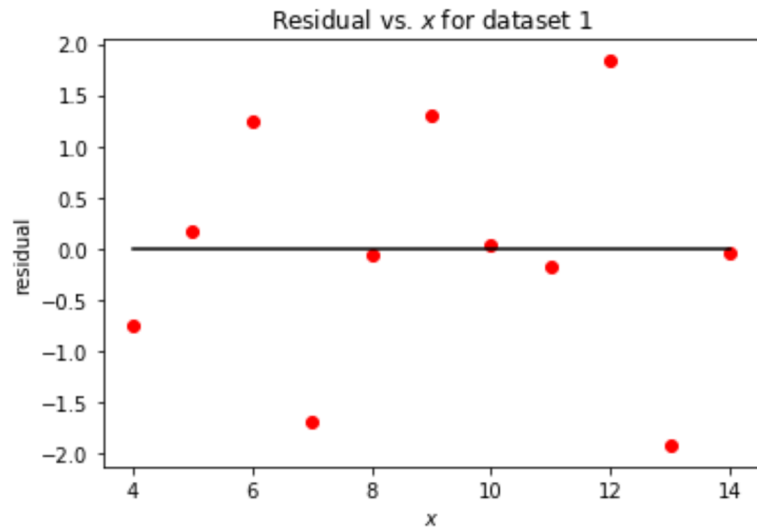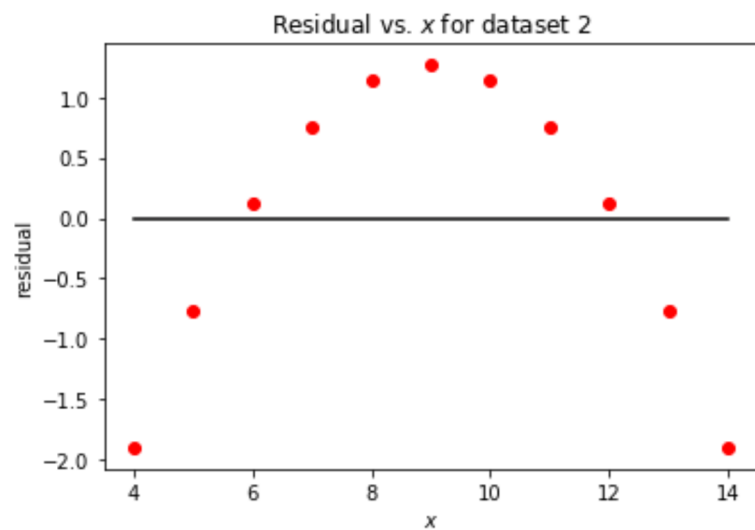
Clear quadratic relationship in the residuals.
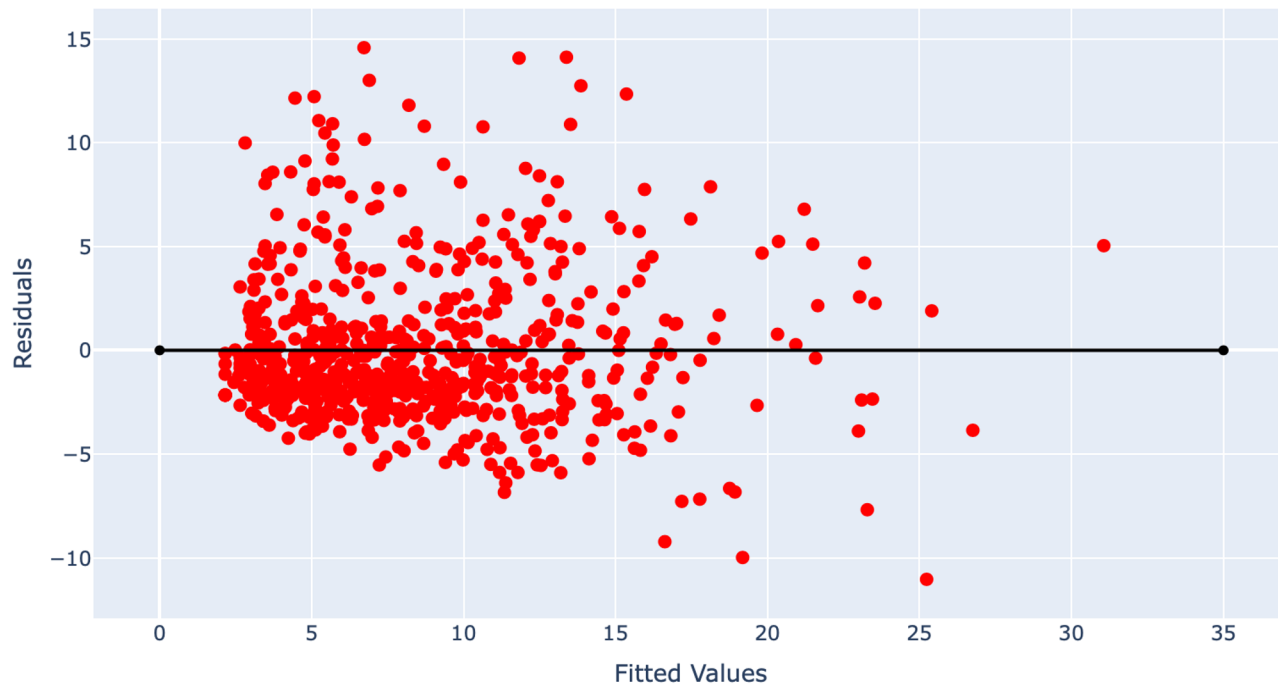
No clear relationship, but uneven spread.

Residuals are orthogonal to the span of X

When using the optimal parameter vector, our residuals are orthogonal to $\mathrm{span}(\mathbb{X})$.

$$\mathbb{X}^T e = 0$$

- Since our predicted response $\hat{\mathbb{Y}}$ is in $\mathrm{span}(\mathbb{X})$, it is orthogonal to the residuals.
    - This is true, no matter what the features of model are.
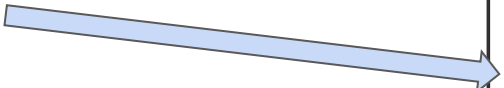
$$\hat{\mathbb{Y}}^T e = 0$$

- When our model contains an intercept term, things become slightly more interesting!

We denote this column with $\mathbb{1}$.

Since $\mathbb{X}^T e = 0$, it is also true that

$$\mathbb{1}^T e = 0$$

$$\mathbb{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & x_{13} & \cdots & x_{2p} \\ 1 & x_{31} & x_{32} & x_{33} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \cdots & x_{np} \end{bmatrix}$$

Properties when our model has an intercept term

When our linear model has an intercept term (i.e. when our design matrix has a column of all 1s), the following properties hold true:

- The sum of the residuals is 0.
  - The mean of the residuals is also 0!
  - This is why the positive and negative residuals cancel out in any residual plot where the (linear) model contains an intercept term, even if the model is terrible.
- The average true y value is equal to the average predicted y value.
  - This follows from the property above.

These properties are true when there is an intercept term, and not necessarily when there isn't.

Existence of a unique solution

Does a solution always exist?

- For all models so far, our goal has been to determine the value of θ that minimizes some average loss.
- The minimum value of both mean squared error and mean absolute error is 0.

$$MSE(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad MAE(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

- This means, there is always **at least one** model parameter that minimizes average loss.

Does a **unique** solution always exist?

- **Constant** model with **squared** loss:
  - Any set of values has a unique mean.
  - Thus, in this case, a unique solution always exists.
- **Simple** linear model with **squared** loss:
  - Any set of non-constant* values has a unique mean, SD, and correlation coefficient.
- **Constant** model with **absolute** loss:
  - This is unique when there is an odd number of *y* values.
  - But, from Lecture 11, when there is an even number of *y* values, there are infinitely many solutions!
    - Recall, in such a case, any value of *θ* between the "middle two" values minimized MAE.

$$\hat{\theta} = \mathbf{mean}(y)$$

$$\hat{\theta}_1 = r\frac{\sigma_y}{\sigma_x} \qquad \hat{\theta}_0 = \bar{y} - \hat{\theta}_1\bar{x}$$

$$\hat{\theta} = \mathbf{median}(y)$$

Understanding the solution matrices

Typically, **n is much larger than p**.

$$\hat{\theta} = \left(\mathbb{X}^T \mathbb{X}\right)^{-1} \mathbb{X}^T \mathbb{Y}$$

Understanding the solution matrices

Typically, **n is much larger than p**.

$$\hat{\theta} = \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathbb{Y}$$

Understanding the solution matrices

The **Normal Equation**:

$$\mathbb{X}^T \mathbb{X} \hat{\theta} = \mathbb{X}^T \mathbb{Y}$$

Our optimal parameter vector can be thought of as the solution to a set of p + 1 equations, with p + 1 unknowns.

Does a **unique** solution always exist?

Let's consider our optimal θ for the multiple linear regression model:

$$\hat{\theta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$$

- As mentioned previously, **at least one solution always exists**.
  - Intuitively, we can always draw a line of best fit for a given set of data. There may be multiple lines that are "equally good".
- When does a unique solution for $\hat{\theta}$ exist?
  - When $\mathbb{X}^T\mathbb{X}$ is invertible. If it is **not invertible**, a unique solution does **not exist**.
    - In such a case, there will be infinitely many values of theta that minimize average squared loss.
    - If there are infinitely many "optimal" choices of coefficients, it's unclear which to use.
    - We want a **unique solution**.

Invertibility of $\mathbb{X}^T\mathbb{X}$

When is $\mathbb{X}^T\mathbb{X}$ invertible?

- $\mathbb{X}^T\mathbb{X}$ is invertible if and only if it is full rank.
    - The shape of $\mathbb{X}^T\mathbb{X}$ is (p + 1) x (p + 1). Invertibility is only defined for square matrices.
    - The rank of a matrix is the number of linearly independent columns (or rows) it contains.
    - This is one of several conditions of the "invertible matrix theorem."
- $\mathbb{X}^T\mathbb{X}$ and $\mathbb{X}$ have the same rank.
    - The proof is beyond the scope of this class.
- The maximum possible rank of $\mathbb{X}^T\mathbb{X}$ is p + 1.
- Thus, $\mathbb{X}^T\mathbb{X}$ is invertible if and only if $\mathbb{X}$ has rank p + 1 (full column rank).
    - **That is, a unique solution for the least squares estimate exists if and only if all columns of $\mathbb{X}$ are linearly independent.**

Invertibility of $\mathbb{X}^T\mathbb{X}$

When does our design matrix $\mathbb{X}$ **not** have full column rank?

- When some features in our design matrix are linear combinations of other features.
  - If "Width", "Height", and "Perimeter" are all columns, $\mathbb{X}$ will not have full rank, since Perimeter = 2 * Width + 2 * Height (linear combination).
  - When we discuss one-hot encoding, this is something to be aware of.
- When our design matrix has more columns than rows (i.e. it is "fat").
  - In the normal setting, n > p + 1 (we typically have more observations than features).

n

p + 1
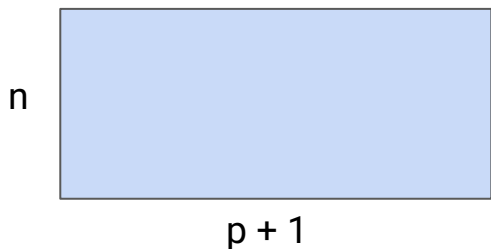
Since the row rank and column rank of a matrix are equal, rank(X) ≤ min(n, p+1).
If n < p + 1, then the column rank cannot possibly be p + 1, meaning it cannot be full column rank.

Summary

- We defined the multiple linear regression model in terms of matrices.
$$\hat{\mathbb{Y}} = \mathbb{X}\boldsymbol{\theta}$$

- We used a geometric argument to derive the optimal parameter vecto $\hat{\boldsymbol{\theta}}$, that minimizes average squared loss. This value is called the **least squares estimate**.
$$\hat{\theta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$$

- We discussed residuals and their properties.
- We explored when a unique solution for $\hat{\theta}$ exists, and when one does not.