Data100 Sp22 Disc 4

11 February 2022 09:06

Data100 Sp22 Disc 4

Announcements

Due Dates

-Homework 4 due Feb 17

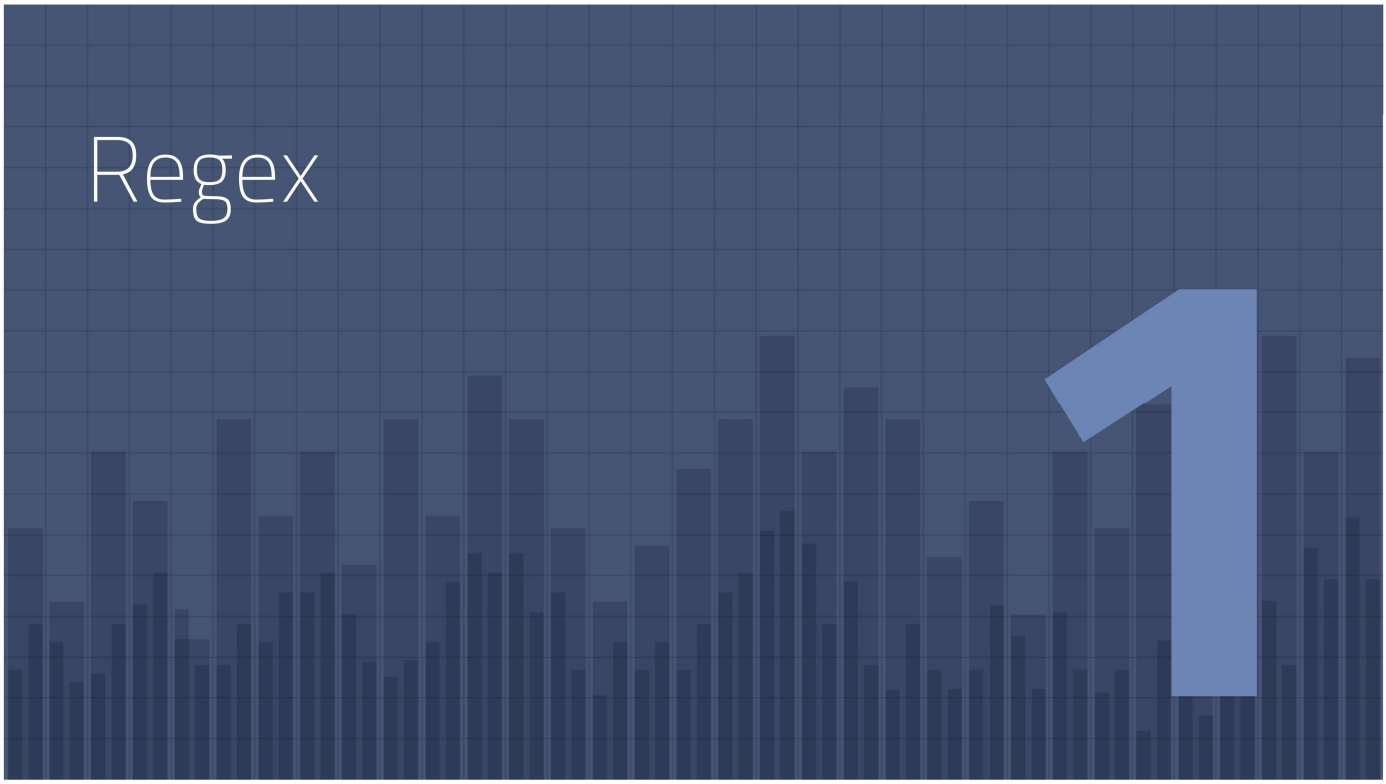
-Lab 4 due Feb 15

-Weekly Check 4 due Feb 14

Other

-Review session Feb19-20! (Weekend, will be recorded)

-Midterm Feb. 24


Regex
1

Basic Regex

operation	example	matches	does not match
or	`TEA|BOBA`	`TEA` `BOBA`	every other string
character class	`[A-Za-z]*`	`helloWorld` `BoBaTEa`	`b0b4T` `AB_AB`
beginning of line (carat)	`^word.*`	`word` `wordle`	`sword`
end of line	`.*word$`	`word` `sword`	`wordle`
character class negation (carat within [])	`[^a-z]+`	`PEPPERS3982` `17211!↑å`	`porch` `CLAmS`

Repeating Characters

operation	example	matches	does not match
Zero or more	`BOB*A`	`BOA` `BOBBBBA`	`BOOA` `BBBA`
One or more	`BOB+A`	`BOBA` `BOBBBBA`	`BOA` `BBBA`
zero or one	`cats?`	`cat` `cats`	`any other string`
repeated exactly {a} times	`c[ao]{2}t`	`coat` `caat`	`caoat` `cat`
repeated from a to b times: {a,b}	`c[ao]{1,2}t`	`coat` `cat`	`caoat` `ct`
Parenthesis (also used in catching groups)	`(CA)+T`	`CACACAT` `CAT`	`CAAT` `T`

Character Groups

1. **. (period):** any character apart from a '\n' (newline)
1. **\w:** a "word" character (letter, digit, or underscore) [a-zA-Z0-9_]
1. **\s:** matches a single whitespace character
1. **\d:** decimal digit [0-9]
1. **\:** escapes a special regex character. e.g. \(means match a opening parenthesis **(**

Regex in Python

1. re.match(pattern, string, ...): match if zero or more characters at the beginning of string match given pattern

1. re.search(pattern, string, ...): scan through string looking for the first location where the regular expression pattern produces a match

1. re.findall(pattern, string, ...): returns all non-overlapping matches of pattern in string as a list of strings

1. re.sub(pattern, repl, string, ...): return the string obtained by replacing the non-overlapping occurrences of pattern in string by the replacement repl

Other features

1. **Greedy vs Lazy**
    a. **Greedy: match as many times as possible**
    b. **Lazy: match as few times as possible**
    c. **+? (lazy): a+? only ever matches a max of 1 a**
2. **Order of operations**

	ERE Precedence (from high to low)	
1	Collation-related bracket symbols	[==] [::] [..]
2	Escaped characters	\<special character>
3	Bracket expression	[]
4	Grouping	()
5	Single-character-ERE duplication	* + ? {m,n}
6	Concatenation	
7	Anchoring	^ $
8	Alternation	|

Quickfire Regex

Which of these regex strings:

Matches: jossh, jossssh

Doesn't match: josh, joh, bjosssh

1. jos[s]+h
2. ^jos[s]+h
3. ^jos[s]*h
4. jos[s]*h

Quickfire Regex

Which of these regex strings:

Matches: jossh, jossssh

Doesn't match: josh, joh, bjosssh

1. jos[s]+h: matches b(josssh)
2. ^jos[s]+h
3. ^jos[s]*h: matches josh
4. jos[s]*h: matches b(josssh)

Quickfire Regex

Which strings would this Regex match?

```
[\w]?[\d]{2,3}[\w]+[A-Z]
```

1. j05hl15A
2. j0shli5A
3. 05lisA
4. j055shhhlisa

Quickfire Regex

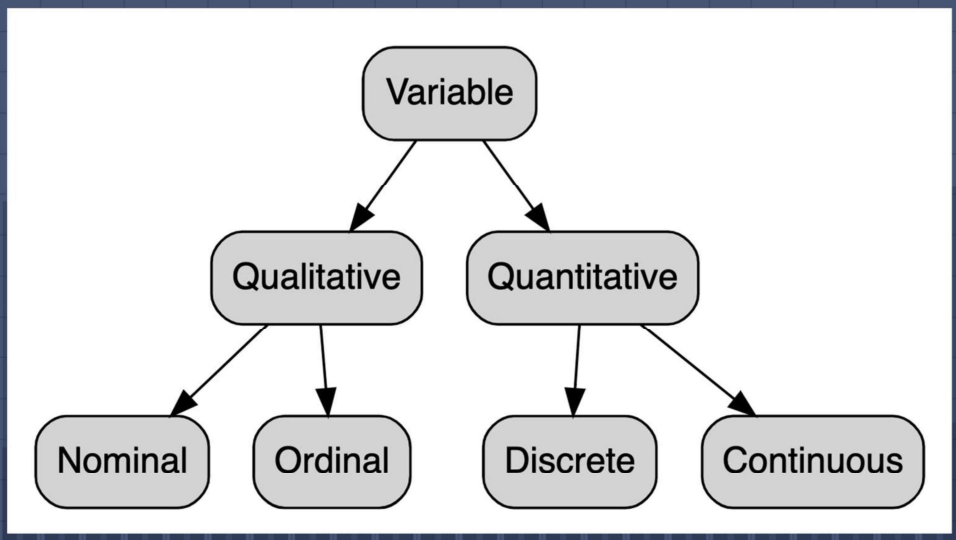
Which strings would this Regex match?

[\w]?[\d]{2,3}[\w]+[A-Z]

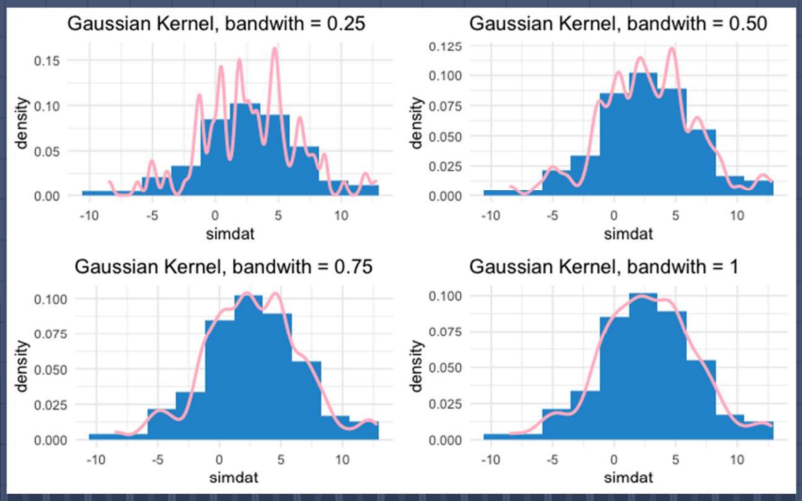
1. j05hl15A
2. j0shli5A : atleast 2 digits needed for [\d]{2,3}
3. 05lisA
4. J055shhhlisa: needs to end with an uppercase letter


Visualizations
2

Types of Variables



KDEs


Gaussian Kernel, bandwith = 0.25
Gaussian Kernel, bandwith = 0.50
Gaussian Kernel, bandwith = 0.75
Gaussian Kernel, bandwith = 1
density
simdat
0.15
0.10
0.05
0.00
0.125
0.100
0.075
0.050
0.025
0.000
-10
-5
0
5
10


United States $7,290
Dollar figures reflect all public and private spending on care, from doctor visits to hospital infrastructure. Data are from 2007 or the most recent year available.
GRAPHIC: OLIVER UBERTI, NG STAFF. SOURCE: "OECD HEALTH DATA 2009," ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT
Health care spending per person, in U.S. dollars
Average number of doctor visits a year
0
4
8
12+
Nation with universal health coverage provided by public and private insurers
Nation without universal health coverage
Average life expectancy at birth
Switzerland $4,417
Luxembourg 4,162
Canada 3,895
Austria 3,763
France 3,601
Denmark 3,512
Sweden 3,323
ABOVE AVERAGE
Australia 3,137
$2,986
U.K. 2,992
BELOW AVERAGE
Finland 2,840
Spain 2,671
Japan 2,581
New Zealand 2,510
Portugal 2,150
South Korea 1,688
Czech Republic 1,626
Slovak Republic 1,555
Hungary 1,388
Poland 1,035
ALL 30 OECD COUNTRIES NOT SHOWN
Mexico 823
80
79.2
75

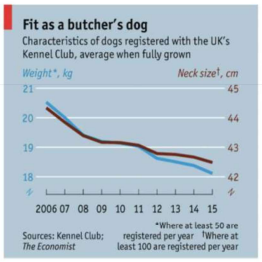
(a) Five variables are being represented visually in this graphic. What are they and what are their feature types (ie qualitative, quantitative, nominal, ordinal)?

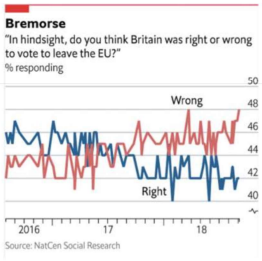
(c) How can we figure out how to interpret the visual qualities of the plot, e.g., how do we know what a color represents?

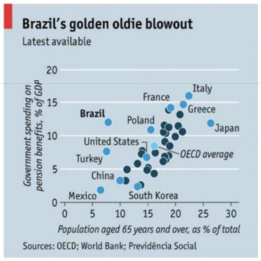
(e) Make 3 observations about the figure. Describe the feature that you are basing your observation on.

For example, South Korea's expenditure on health care is comparable to Eastern European countries (and among the lowest of all countries plotted), but the life expectancy is much higher than the Eastern European countries. In the plot we see that the left endpoint of South Korea's line segment is near the Eastern European countries, but the slope of the line segment is much steeper.

9. Creating visualizations that represent data accurately and that support the narrative we wish to create is no easy task. Even the journalists and editors at *The Economist*, a newspaper known for it's compelling, data-driven articles, have been known to make blunders. Three of their ill-thought-out plots are presented below. Consider what aspects of the visualizations are misleading, and think of ways in which you can remedy them.







Hint: The datapoints in the rightmost plot are shaded based on whether or not they are labeled.

disc04

15 February 2022 17:17

disc04

Data 100, Spring 2022 **February 11, 2022**

Discussion #4

Regular Expressions

Here's a complete list of metacharacters:

`.  ^  $  *  +  ?  {  }  [  ]  \  |  (  )`

Some reminders on what each can do (this is not exhaustive):

- **`"^"`** matches the position at the beginning of string (unless used for negation `"[^]"`)
- **`"$"`** matches the position at the end of string character.
- **`"?"`** match preceding literal or sub-expression 0 or 1 times.
- **`"+"`** match preceding literal or sub-expression *one* or more times.
- **`"*"`** match preceding literal or sub-expression *zero* or more times
- **`"."`** match any character except new line.
- **`"[ ]"`** match any one of the characters inside, accepts a range, e.g., `"[a-c]"`.
- **`"( )"`** used to create a sub-expression
- **`"\d"`** match any *digit* character. `"\D"` is the complement.
- **`"\w"`** match any *word* character (letters, digits, underscore). `"\W"` is the complement.
- **`"\s"`** match any *whitespace* character including tabs and newlines. `\S` is the complement.
- **`"*?"`** Non-greedy version of *. Not fully discussed in class.
- **`"\b"`** match boundary between words. Not discussed in class.
- **`"+?"`** Non-greedy version of +. Not discussed in class.
- **`"{m,n}"`** The preceding element or subexpression must occur between m and n times, inclusive.

Some useful `re` package functions:

- **`re.split(pattern, string)`** split the `string` at substrings that match the `pattern`. Returns a list.
- **`re.sub(pattern, replace, string)`** apply the `pattern` to `string` replacing matching substrings with `replace`. Returns a string.
- **`re.findall(pattern, string)`** Returns a list of all matches for the given `pattern` in the `string`.

Regular Expressions

1. Which strings contain a match for the following regular expression, `"1+1$"`? The character `"␣"` represents a single space.

   ◯ A. `What␣is␣1+1` ◉ B. `Make␣a␣wish␣at␣11:11` ◯ C. `111␣Ways␣to␣Succeed`

2. Write a regular expression that matches strings (including the empty string) that only contain lowercase letters and numbers.

   `^[a-z0-9]*$`

3. Given `sometext = "I've␣got␣10␣eggs,␣20␣gooses,␣and␣30␣giants."`, use `re.findall` to extract all the items and quantities from the string. The result should look like `['10 eggs', '20 gooses', '30 giants']`. You may assume that a space separates quantity and type, and that each item ends in s.

   `re.findall(r"\d\d\s\w+s", sometext)`

4. For each pattern specify the starting and ending position of the first match in the string. The index starts at zero and we are using closed intervals (both endpoints are included).

	abcdefg	abcs!	ab␣abc	abc␣123
`abc*`	[0, 2]	[0,2]	[0,1]	[0,2]
`[^\s]+`	[0,6]	[0,4]	[0,5]	[0,5]
`ab.*c`	[0,2]	[0,2]	[0,5]	[0,2]
`[a-z1,9]+`	[0,6]	[0,3]	[0,3]	[0,3]

5. *(Bonus)* Given the following text in a variable `log`:

```
169.237.46.168 - - [26/Jan/2014:10:47:58 -0800]
"GET␣/stat141/Winter04/␣HTTP/1.1" 200 2585
"http://anson.ucdavis.edu/courses/"
```

Fill in the regular expression in the variable `pattern` below so that after it executes, day is 26, month is Jan, and year is 2014.

```
pattern = ...
matches = re.findall(pattern, log)
day, month, year = matches[0]
```

6. *(Bonus)* Given that `sometext` is a string, use `re.sub` to replace all clusters of non-vowel characters with a single period. For example `"a␣big␣moon,␣between␣us..."` would be changed to `"a.i.oo.e.ee.u."`.

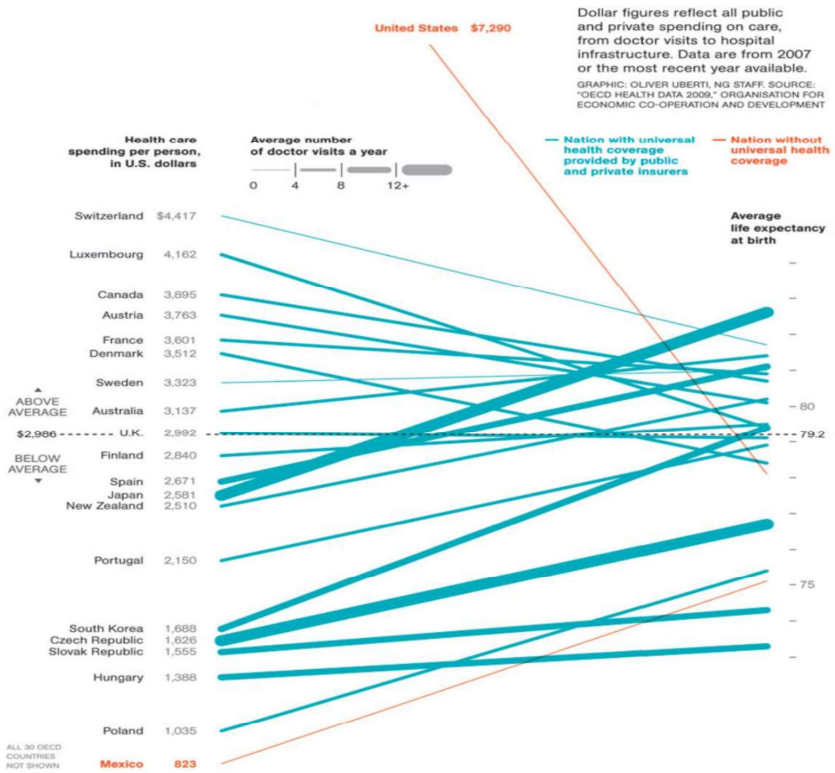
7. *(Bonus)* Given the text:

   ```
   "<record>␣Josh␣Hug␣<hug@cs.berkeley.edu>␣Faculty␣</record>"
   "<record>␣Lisa␣Yan␣<lisa.yan@berkeley.edu>␣Instructor␣</record>"
   ```

   Which of the following matches exactly to the email addresses (including angle brackets)?

   ◯ A. `<.*@.*>` ◯ B. `<[^>]*@[^>]*>` ◯ C. `<.*@\w+\..*>`

Data Visualization



8. The first part of the discussion will be centered on the above visualization.

   (a) Five variables are being represented visually in this graphic. What are they and what are their feature types (ie qualitative, quantitative, nominal, ordinal)?

   (b) How are the variables represented in the graphic, e.g., the variable `XXX` is mapped to the x-axis, the variable `WWW` is mapped to the y-axis, the variable `ZZZ` is conveyed through color, etc.?

(c) How can we figure out how to interpret the visual qualities of the plot, e.g., how do we know what a color represents?

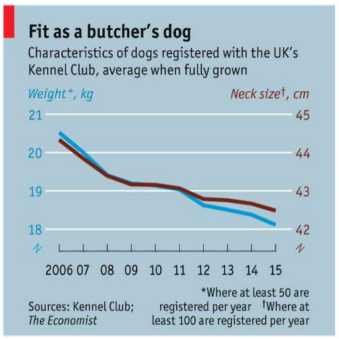
(d) What purpose does the comment at the top right of the plot serve?

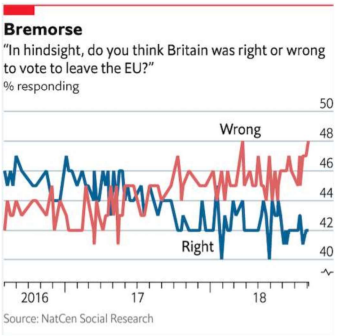
(e) Make 3 observations about the figure. Describe the feature that you are basing your observation on.

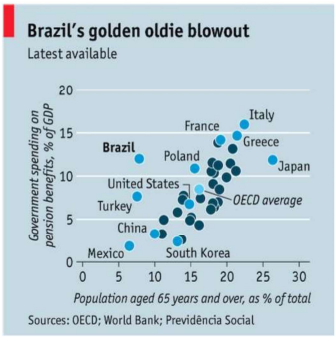
For example, South Korea's expenditure on health care is comparable to Eastern European countries (and among the lowest of all countries plotted), but the life expectancy is much higher than the Eastern European countries. In the plot we see that the left endpoint of South Korea's line segment is near the Eastern European countries, but the slope of the line segment is much steeper.

(f) Consider the steep negative slope and narrowness of the line segment that represents the data for the United States. What systemic, social, or societal issues might explain this?

9. Creating visualizations that represent data accurately and that support the narrative we wish to create is no easy task. Even the journalists and editors at *The Economist*, a newspaper known for it's compelling, data-driven articles, have been known to make blunders. Three of their ill-thought-out plots are presented below. Consider what aspects of the visualizations are misleading, and think of ways in which you can remedy them.







Hint: The datapoints in the rightmost plot are shaded based on whether or not they are labeled.

hw05_sol

Tuesday, February 22, 2022 9:25 AM

hw05_sol

Data 100, Spring 2022

Homework #5 Solutions

Total Points: 36

Submission Instructions

You must submit this assignment to Gradescope by **Friday, March 3rd at 11:59 PM Pacific**. While Gradescope accepts late submissions, you will not receive **any** credit for a late submission if you do not have prior accommodations (e.g. DSP).

You can work on this assignment in any way you like:

- One way is to download this PDF, print it out, and write directly on these pages (we've provided enough space for you to do so). Alternatively, if you have a tablet, you could save this PDF and write directly on it.
- Another way is to use some form of LaTeX. Overleaf is a great tool; visit the course website for a LaTeX template of this homework.
- You could also write your answers on a blank sheet of paper.

Regardless of what method you choose, the end result needs to end up on Gradescope, as a PDF. If you wrote something on physical paper (like options 1 and 3 above), you will need to use a scanning application (e.g. CamScanner) in order to submit your work.

When submitting on Gradescope, you **must correctly assign pages to each question** (it prompts you to do this after submitting your work). This significantly streamlines the grading process for our tutors. Failure to do this may result in a score of 0 for any questions that you didn't correctly assign pages to. If you have any questions about the submission process, please don't hesitate to ask on Piazza.

Collaborators

Data science is a collaborative activity. While you may talk with others about the homework, we ask that you write your solutions individually. If you do discuss the assignments with others please include their names at the top of your submission.

Properties of Simple Linear Regression

1. (7 points) In lecture, we spent a great deal of time talking about simple linear regression, which you also saw in Data 8. To briefly summarize, the simple linear regression model assumes that given a single observation x, our predicted response for this observation is $\hat{y} = \theta_0 + \theta_1 x$. (Note: In this problem we write (θ_0, θ_1) instead of (a, b) to more closely mirror the multiple linear regression model notation.)

   In Lecture 9 we saw that the $\theta_0 = \hat{\theta}_0$ and $\theta_1 = \hat{\theta}_1$ that minimize the average L_2 loss for the simple linear regression model are:

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

   Or, rearranging terms, our predictions $\hat{y}$ are:

   $$\hat{y} = \bar{y} + r\sigma_y \frac{x - \bar{x}}{\sigma_x}$$

   (a) (3 points) As we saw in lecture, a residual e_i is defined to be the difference between a true response y_i and predicted response $\hat{y}_i$. Specifically, $e_i = y_i - \hat{y}_i$. Note that there are n data points, and each data point is denoted by (x_i, y_i).
   Prove, using the equation for $\hat{y}$ above, that $\sum_{i=1}^{n} e_i = 0$.

   Solution:

   $$\begin{aligned}
   \sum_{i=1}^{n} e_i &= \sum_{i=1}^{n} (y_i - \hat{y}_i) \\
   &= \sum_{i=1}^{n} (y_i - (\bar{y} + r\sigma_y \frac{(x_i - \bar{x})}{\sigma_x}) \\
   &= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \bar{y} - r\frac{\sigma_y}{\sigma_x} \sum_{i=1}^{n} (x_i - \bar{x}) \\
   &= n\bar{y} - n\bar{y} - r\frac{\sigma_y}{\sigma_x}[n\bar{x} - n\bar{x}] \\
   &= 0
   \end{aligned}$$

   (b) (2 points) Using your result from part (a), prove that $\bar{y} = \bar{\hat{y}}$.

Solution:

$$\sum_{i=1}^{n} e_i = 0$$
$$\sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$$
$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{y}_i = 0$$
$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$$
$$\frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i$$
$$\bar{y} = \bar{\hat{y}}$$

(c) (2 points) Prove that $(\bar{x}, \bar{y})$ is on the simple linear regression line.

Solution: Starting with

$$y = \hat{\theta}_0 + \hat{\theta}_1 x$$

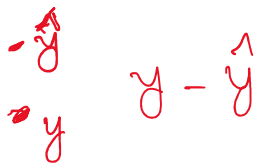
we can substitute the definition of $\hat{\theta}_0$ from above as:

$$y = \bar{y} - \hat{\theta}_1 \bar{x} + \hat{\theta}_1 x$$

When we plug $\bar{x}$ in for x, we find the right-hand side becomes

$$\bar{y} - \hat{\theta}_1 \bar{x} + \hat{\theta}_1 \bar{x},$$

which reduces to $\bar{y}$. We see that $(\bar{x}, \bar{y})$ is on the regression line.



Geometric Perspective of Least Squares

2. (7 points) In Lecture 11, we viewed both the simple linear regression model and the multiple linear regression model through the lens of linear algebra. The key geometric insight was that if we train a model on some design matrix $\mathbb{X}$ and true response vector $\mathbb{Y}$, our predicted response $\hat{\mathbb{Y}} = \mathbb{X}\hat{\theta}$ is the vector in span($\mathbb{X}$) that is closest to $\mathbb{Y}$.

   In the simple linear regression case, our optimal vector θ is $\hat{\theta} = [\hat{\theta_0}, \hat{\theta_1}]^T$, and our design matrix is

$$\mathbb{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} | & | \\ \mathbb{1} & \vec{x} \\ | & | \end{bmatrix}$$

   This means we can write our predicted response vector as $\hat{\mathbb{Y}} = \mathbb{X}\begin{bmatrix} \hat{\theta_0} \\ \hat{\theta_1} \end{bmatrix} = \hat{\theta_0}\mathbb{1} + \hat{\theta_1}\vec{x}$.

   Note, in this problem, $\vec{x}$ refers to the n-length vector $[x_1, x_2, ..., x_n]^T$. In other words, it is a feature, not an observation.

   For this problem, assume we are working with the **simple linear regression model**, though the properties we establish here hold for any linear regression model that contains an intercept term.

   (a) (3 points) Using the geometric properties from lecture, prove that $\sum_{i=1}^{n} e_i = 0$.

   Hint: Recall, we define the residual vector as $e = \mathbb{Y} - \hat{\mathbb{Y}}$, and $e = [e_1, e_2, ..., e_n]^T$.

   Solution: We can think of $\sum_{i=1}^{n} e_i$ as the dot product of the residual vector, e and the one vector, $\mathbb{1}$. That is,

$$\sum_{i=1}^{n} e_i = e \cdot \mathbb{1}$$

   The predicted $\hat{\mathbb{Y}}$ is the vector closest to $\mathbb{Y}$ in the span($\{\mathbb{1}, \vec{x}\}$). In other words, $\hat{\mathbb{Y}}$ is the projection of $\mathbb{Y}$ into the span, and the difference $\mathbb{Y} - \hat{\mathbb{Y}}$ is orthogonal to any vector in the span($\{\mathbb{1}, \vec{x}\}$). So, $(\mathbb{Y} - \hat{\mathbb{Y}})$ is orthogonal to $\mathbb{1}$. Orthogonality means that the dot product between $\mathbb{Y} - \hat{\mathbb{Y}}$ and any vector in the span is 0. In particular,

$$(\mathbb{Y} - \hat{\mathbb{Y}}) \cdot \mathbb{1} = 0$$

Since $\vec{e} = \mathbb{Y} - \hat{\mathbb{Y}}$, we have shown that

$$\vec{e} \cdot \mathbb{1} = 0$$

This same argument can be used to establish that $\vec{x} \cdot e = 0$ because $\vec{x}$ is also in the span$(\{\mathbb{1}, \vec{x}\})$.

And, again, since $\hat{\mathbb{Y}}$ is in this span (specifically, $\hat{\mathbb{Y}} = \hat{\theta}_0 \mathbb{1} + \hat{\theta}_1 \vec{x}$) it also follows that $\hat{\mathbb{Y}} \cdot e = 0$. In other words, we have just given explanations for all parts of this problem.

(b) (2 points) Explain why the vector $\vec{x}$ (as defined in the problem) and the residual vector e are orthogonal. *Hint: Two vectors are orthogonal if their dot product is 0.*

Solution: In the previous problem, we explained why $\vec{e}$ is orthogonal to any vector in the span$(\{\mathbb{1}, \vec{x}\})$. Therefore it follows that the dot product of x and e is 0.

(c) (2 points) Explain why the predicted response vector $\hat{\mathbb{Y}}$ and the residual vector e are orthogonal.

Solution: See the solution for part a. Since $\hat{\mathbb{Y}}$ is in the span of our design matrix, it is orthogonal to the residuals. Hence, $\hat{\mathbb{Y}} \cdot e = 0$.

Properties of a Linear Model With No Constant Term

Suppose that we don't include an intercept term in our model. That is, our model is now

$$\hat{y} = \gamma x,$$

where γ is the single parameter for our model that we need to optimize. (In this equation, x is a scalar, corresponding to a single observation.)

As usual, we are looking to find the value $\hat{\gamma}$ that minimizes the average L_2 loss (mean squared error) across our observed data $\{(x_i, y_i)\}, i = 1, \ldots, n$:

$$R(\gamma) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \gamma x_i)^2$$

The normal equations derived in lecture no longer hold. In this problem, we'll derive a solution to this simpler model. We'll see that the least squares estimate of the slope in this model differs from the simple linear regression model, and will also explore whether or not our properties from the previous problem still hold.

3. (4 points) Use calculus to find the minimizing $\hat{\gamma}$. That is, prove that

$$\hat{\gamma} = \frac{\sum x_i y_i}{\sum x_i^2}$$

Note: This is the slope of our regression line, analogous to $\hat{\theta_1}$ from our simple linear regression model.

Solution:

As in lecture, the value of γ that minimizes $\frac{1}{n}\sum_{i=1}^{n}(y_i - \gamma x_i)^2$ is the same value that minimizes $\sum_{i=1}^{n}(y_i - \gamma x_i)^2$.

Differentiate the sum of squares with respect to γ to find:

$$-2\sum_{i=1}^{n}(y_i - \gamma x_i)x_1$$

Set the derivative to 0 and solve for the minimizing $\hat{\gamma}$

$$\begin{aligned} 0 &= -2\sum_{i=1}^{n}(y_i - \hat{\gamma} x_i)x_1 \\ &= \sum_{i=1}^{n} y_i x_i - \hat{\gamma}\sum_{i=1}^{n} x_i^2 \end{aligned}$$

Rearrange terms

$$\sum_{i=1}^{n} x_i y_i = \hat{\gamma} \sum_{i=1}^{n} x_i^2$$

to find,

$$\hat{\gamma} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

4. (8 points) For our new simplified model, our design matrix $\mathbb{X}$ is:

$$\mathbb{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} | \\ \vec{x} \\ | \end{bmatrix}.$$

Therefore our predicted response vector $\hat{\mathbb{Y}}$ can be expressed as $\hat{\mathbb{Y}} = \hat{\gamma}\vec{x}$. ($\vec{x}$ here is defined the same way it was in Question 2.)

Earlier in this homework, we established several properties that held true for the simple linear regression model that contained an intercept term. For each of the following four properties, state whether or not they still hold true even when there isn't an intercept term. Be sure to justify your answer.

(a) (2 points) $\sum_{i=1}^{n} e_i = 0$.

Solution: This property does not hold anymore. Note that our proof for question 3 requires $\mathbb{1}$ to be one of the columns in our design matrix. Without an intercept term, the feature matrix might not have $\mathbb{1}$ as one of its columns. Hence, we do not know for sure that $\mathbb{1} \cdot e = 0$.

Intuitively speaking, without an intercept term our regression line is forced to pass through the origin, so we can't "position" it in order to guarantee that the residuals sum to 0.

For a concrete counterexample, consider the points $\{(-1, 2), (1, 3)\}$. Then, $\hat{\gamma} = \frac{-1 \cdot 2 + 1 \cdot 3}{(-1)^2 + 1^2} = \frac{1}{2}$. Then, $\hat{y}_1 = \hat{\gamma} x_1 = -\frac{1}{2}$ and $\hat{y}_2 = \hat{\gamma} x_2 = \frac{1}{2}$, making the residuals $e_1 = 2 - (-\frac{1}{2}) = \frac{5}{2}$ and $e_2 = 3 - \frac{1}{2} = \frac{5}{2}$; clearly $e_1 + e_2 = 5 \neq 0$.

(b) (2 points) The column vector $\vec{x}$ and the residual vector e are orthogonal.

Solution: This property still holds. We know that $\vec{x}$ is the only column in our design matrix. Since the residuals are orthogonal to the column space of our feature matrix, we know that $\vec{x} \cdot e = 0$.

(c) (2 points) The predicted response vector $\hat{\mathbb{Y}}$ and the residual vector e are orthogonal.

Solution: This property still holds. Note that our design matrix in this scenario is just one column, which is $\vec{x}$. Since the residuals are orthogonal to the column space of our feature matrix, we know that they are orthogonal to anything in the span of $\vec{x}$. Note that $\hat{\mathbb{Y}} = \hat{\gamma}\vec{x}$, which means that $\hat{\mathbb{Y}} \in \text{span}(\{\vec{x}\})$. Hence $\hat{\mathbb{Y}}$ is orthogonal to the residuals, which means that $\hat{\mathbb{Y}} \cdot e = 0$.

(d) (2 points) $(\bar{x}, \bar{y})$ is on the regression line.

Solution: This property does not hold anymore. Note that our derivation in question 1 relied on the value of $\hat{\theta}_0$. However, $\hat{\theta}_0$ does not exist anymore since we don't have an intercept term, which means that $\hat{\gamma}\bar{x}$ is not necessarily equal to $\bar{y}$.

For concreteness, consider the counterexample in the solutions for 4a. There, $\bar{x} = 0$, $\bar{y} = \frac{5}{2}$ and $\hat{\gamma} = \frac{1}{2}$, and so $\hat{\gamma}\bar{x} = 0 \neq \frac{5}{2}$. Thus, in this case, $(\bar{x}, \bar{y})$ is not on the regression line.

MSE "Minimizer"

5. (10 points) Recall from calculus that given some function $g(x)$, the x you get from solving $\frac{dg(x)}{dx} = 0$ is called a *critical point* of g – this means it could be a minimizer or a maximizer for g. In this question, we will explore some basic properties and build some intuition on why, for certain loss functions such as squared L_2 loss, the critical point of the empirical risk function (defined as average loss on the observed data) will always be the minimizer.

   Given some linear model $f(x) = \gamma x$ for some real scalar γ, we can write the empirical risk of the model f given the observed data $\{x_i, y_i\}, i = 1, \ldots, n$ as the average L_2 loss, also known as mean squared error (MSE):

   $$\frac{1}{n}\sum_{i=1}^{n}(y_i - \gamma x_i)^2.$$

   (a) (1 point) Let's break the function above into individual terms. Complete the following sentence by filling in the blanks using one of the options in the parenthesis following each of the blanks:

   The mean squared error can be viewed as a sum of n _____ (linear/quadratic/logarithmic/exponential) terms, each of which can be treated as a function of ____ (x_i/y_i/γ).

   Solution: The mean squared error can be treated as a sum of n quadratic terms, each of which can be treated as a function of γ.

   (b) (3 points) Let's investigate one of the n functions in the summation in the MSE. Define $g_i(\gamma) = \frac{1}{n}(y_i - \gamma x_i)^2$ for $i = 1, \ldots, n$. Recall from calculus that we can use the 2nd derivative of a function to describe its curvature about a certain point (if it is facing concave up, down, or possibly a point of inflection). You can take the following as a fact: A function is convex if and only if the function's 2nd derivative is non-negative on its domain. Based on this property, verify that g_i is a **convex function**.

   Solution: First, we have

   $$g_i(\gamma) = \frac{1}{n}(y_i - \gamma x_i)^2 = \frac{1}{n}(y_i^2 - 2y_i x_i \gamma + \gamma^2 x_i^2)$$

Then we take the 2nd derivative of $g_i(\gamma)$ with respect to γ.

$$\frac{dg_i(\gamma)}{d\gamma} = \frac{1}{n}(-2y_i x_i + 2x_i^2\gamma)$$
$$\frac{d^2g_i(\gamma)}{d\gamma^2} = \frac{2}{n}x_i^2$$

Since n is a positive number, $\frac{2}{n} > 0$. In addition, $x_i^2 \geq 0$. Thus, $\frac{2}{n}x_i^2$ is always non-negative, which means g_i is a convex function. In terms of g_i's curvature, we can see that for all γ, the function either faces concave up or is a constant, which occurs when $x_i = 0$.

(c) (2 points) Briefly explain intuitively in words why given a convex function $g(x)$, the critical point we get by solving $\frac{dg(x)}{dx} = 0$ minimizes g. You can assume that $\frac{dg(x)}{dx}$ is a function of x (and not a constant).

Solution: For a convex function $g(x)$, its 2nd derivative is always non-negative. This means that as x increases, the slope $\frac{dg(x)}{dx}$ is either increasing or staying the same. When $\frac{dg(x)}{dx} = 0$, it's at a point where the slope is turning from negative to 0. We know that when the slope is negative, $g(x)$ is decreasing and when the slope is positive, $g(x)$ increasing. This means the before the function gets to the point where the slope is 0, the function has been decreasing and as it hits this point, the function has stopped decreasing. Hence this point must be the minimum and the function can only stay the same or increase as x increases after getting to this point.

(d) (3 points) Now that we have shown that each term in the summation of the MSE is a convex function, one might wonder if the entire summation is convex given that it is a sum of convex functions.

Let's look at the formal definition of a **convex function**. Algebraically speaking, a function $g(x)$ is convex if for any two points $(x_1, g(x_1))$ and $(x_2, g(x_2))$ on the function,

$$g(cx_1 + (1-c)x_2) \leq cg(x_1) + (1-c)g(x_2)$$

for any real constant $0 \leq c \leq 1$.

Intuitively, the above definition says that, given the plot of a convex function $g(x)$, if you connect 2 randomly chosen points on the function, the line segment will always lie on or above $g(x)$ (try this with the graph of $y = x^2$).

i. (2 points) Using the definition above, show that if $g(x)$ and $h(x)$ are both convex functions, their sum $g(x) + h(x)$ will also be a convex function.

Solution: By definition, since $g(x)$ and $h(x)$ are both convex, we have that for all $0 \leq c \leq 1$:

$$g(cx_1 + (1-c)x_2) \leq cg(x_1) + (1-c)g(x_2)$$
$$h(cx_1 + (1-c)x_2) \leq ch(x_1) + (1-c)h(x_2)$$

Adding the 2 inequalities and combining the terms, we have:

$$(g+h)(cx_1 + (1-c)x_2) \leq c(g+h)(x_1) + (1-c)(g+h)(x_2)$$

Therefore, by definition, we have shown that $(g+h)(x) = g(x) + h(x)$ is also a convex function.

ii. (1 point) Based on what you have shown in the previous part, explain intuitively why the sum of n convex functions is still a convex function when $n > 2$.

Solution: We can repeatedly apply what we have shown in the previous part on the first 2 convex functions out of all the convex functions, and eventually, we will reduce the sum of any convex functions to a sum of two convex functions, which is a convex function.

(e) (1 point) Finally, explain why in our case that, when we solve for the critical point of the MSE by taking the gradient with respect to the parameter and setting the expression to 0, it is guranteed that the solution we find will minimize the MSE.

Solution: The MSE is a summation of n convex functions, which as we saw in the previous part makes the entire MSE a convex function. Since for a convex function, its critical point is a minimizer, the critical point of the MSE is the minimizer of the function.

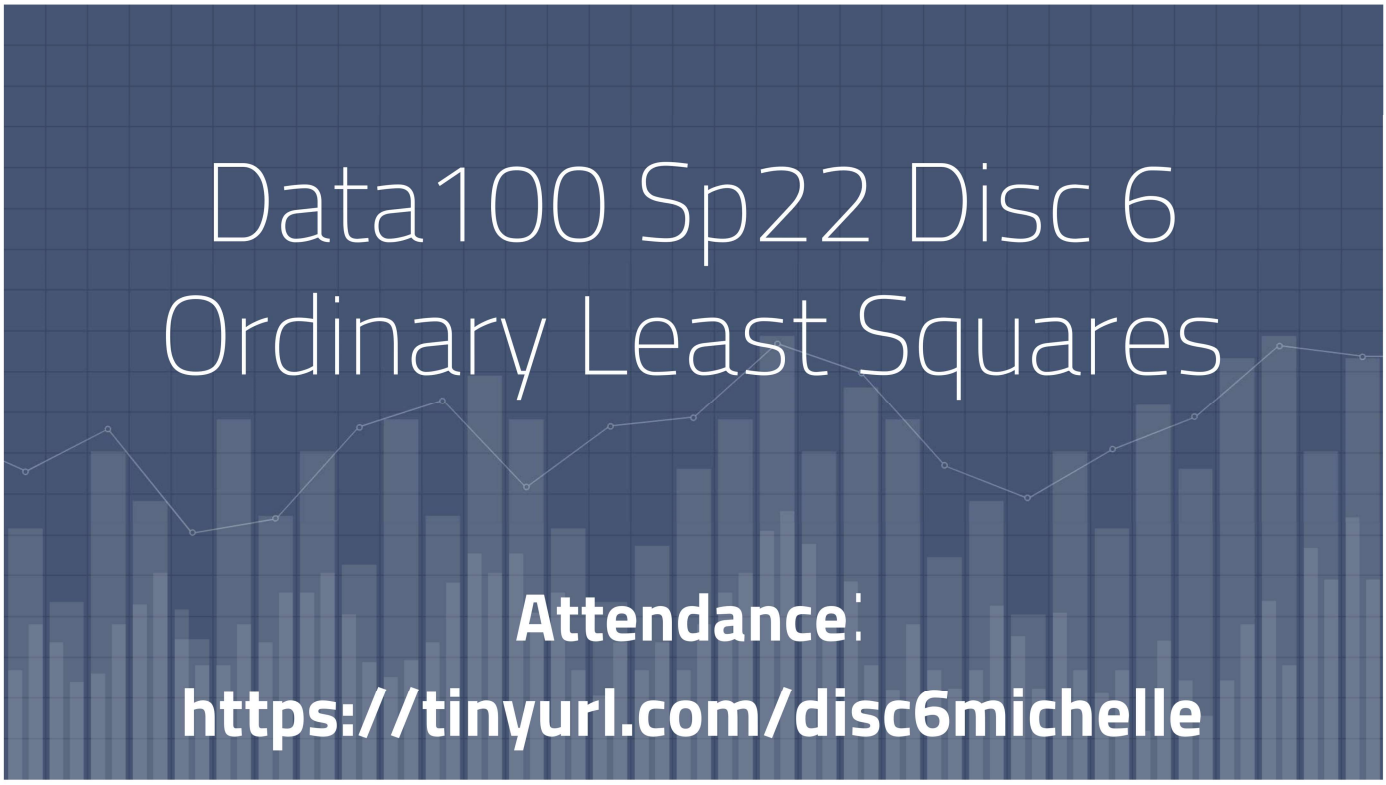
Closing note: In this question, we have discussed only the simple linear model with no constant term—a single-variable function. However, the above properties extend more generally to all multivariable linear regression models; this proof is beyond the scope of this course and is left to a future you.

Congratulations! You have finished Homework 5!

Data100 Sp22 Disc 6

25 February 2022 08:56

Data100 Sp22 Disc 6



Announcements

Due Dates

-Homework 5 due March 3 (start early)

-Lab 6 due March 1

-Weekly check 6 due Feb 28

Other

-Congrats on finishing the midterm!

-Calendly: https://forms.gle/3BJjPjPbMkNHs2ir9


Modeling Process
1

Recap: Modeling Process

1. Choose a model
2. Choose a loss function
3. Fit the model
4. Evaluate model performance

Recap: Modeling Process Simple Linear Regression

1. Choose a model

2. Choose a loss function

3. Fit the model

4. Evaluate model performance

SLR model

$$\hat{y} = \theta_0 + \theta_1 x$$

Recap: Modeling Process Simple Linear Regression

1. Choose a model	**2. Choose a loss function**	3. Fit the model	4. Evaluate model performance

SLR model

$$\hat{y} = \theta_0 + \theta_1 x$$

L1/L2 Loss, MSE

$$L(y, \hat{y}) = (y - \hat{y})^2$$

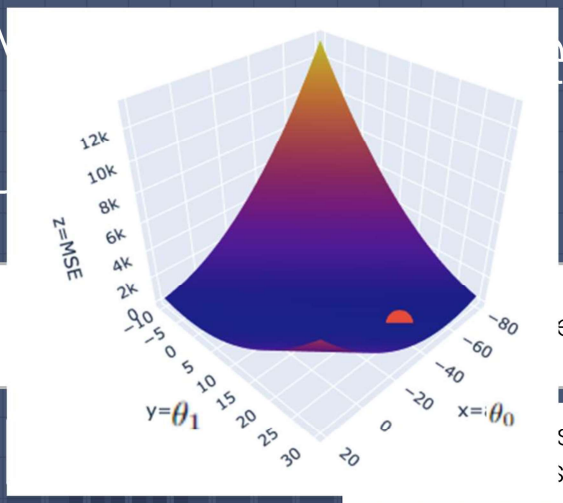
Recap: Modeling Process Simple Linear Regression

1. Choose a model	2. Choose a loss function	**3. Fit the model**	4. Evaluate model performance
SLR model	L1/L2 Loss, MSE	**Minimize Loss with Calculus**	
$\hat{y} = \theta_0 + \theta_1 x$	$L(y, \hat{y}) = (y - \hat{y})^2$	$\hat{\theta}_1 = r\frac{\sigma_y}{\sigma_x}$ $\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}$	

Recap: Modeling Process Simple Linear Regression

1. Choose a model	2. Choose a loss function	3. Fit the model	**4. Evaluate model performance**
SLR model	L1/L2 Loss, MSE	Minimize Loss with Calculus	**Visualizations, RMSE**
$\hat{y} = \theta_0 + \theta_1 x$	$L(y, \hat{y}) = (y - \hat{y})^2$	$\hat{\theta}_1 = r\frac{\sigma_y}{\sigma_x}$ $\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}$	

Recap: M ess Simple L on



1. Choose a model

SLR model

$$\hat{y} = \theta_0 + \theta_1 x$$

el

s

$$L(y, \hat{y}) = (y - \hat{y})^2$$

$$\hat{\theta}_1 = r\frac{\sigma_y}{\sigma_x}$$

$$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}$$

4. Evaluate model performance

Visualizations, RMSE

Why Multiple Linear Regression?

- **Simple Linear Regression not enough for all use cases**
  - **Often want to predict the value of the response variable based on multiple predictor variables.**

Why Multiple Linear Regression?

- **Simple Linear Regression not enough for all use cases**
  - **Often want to predict the value of the response variable based on multiple predictor variables.**
  - **E.g. predict points based on all 3 of Field Goals (FG), Assists (AST), and 3 pointers (3PA)**
  - **SLR - can only predict points based on one out of {FG, AST, 3PA}**

	FG	AST	3PA	PTS
1	1.8	0.6	4.1	5.3
2	0.4	0.8	1.5	1.7
3	1.1	1.9	2.2	3.2
4	6.0	1.6	0.0	13.9
5	3.4	2.2	0.2	8.9
6	0.6	0.3	1.2	1.7

Modeling Process
Multiple Linear Regression

1. Choose a model
y is scalar

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p$$

Modeling Process
Multiple Linear Regression

1. Choose a model
y is scalar

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p$$

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

$$x, \theta \in \mathbb{R}^{(p+1)} : x = \begin{bmatrix} 1 \\ 0.4 \\ 0.8 \\ 1.5 \end{bmatrix} \quad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$$

Modeling Process
Multiple Linear Regression

1. Choose a model (y is a vector)

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

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

Prediction vector $\mathbb{R}^n$ | Design matrix $\mathbb{R}^{n \times (p+1)}$ | Parameter vector $\mathbb{R}^{(p+1)}$

Modeling Process
Multiple Linear Regression

1. Choose a model (y is a vector)

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

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

Prediction vector $\mathbb{R}^n$ | Design matrix $\mathbb{R}^{n \times (p+1)}$ | Parameter vector $\mathbb{R}^{(p+1)}$

Special all ones feature - intercept term

Modeling Process
Multiple Linear Regression

2. Choose a loss function

L2 Loss

Mean Squared Error (MSE)

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

Modeling Process
Multiple Linear Regression

2. Choose a loss function

L2 Loss

Mean Squared Error (MSE)

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

For the n-dimensional vector $x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}$, the **L2 vector norm** is

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

Modeling Process
Multiple Linear Regression

3. Fit the Model

The value of theta that minimizes the MSE loss (R(θ)) is:

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

Normal Equation

Modeling Process
Multiple Linear Regression

3. Fit the Model

The value of theta that minimizes the MSE loss (R(θ)) is:

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

Least squares estimate

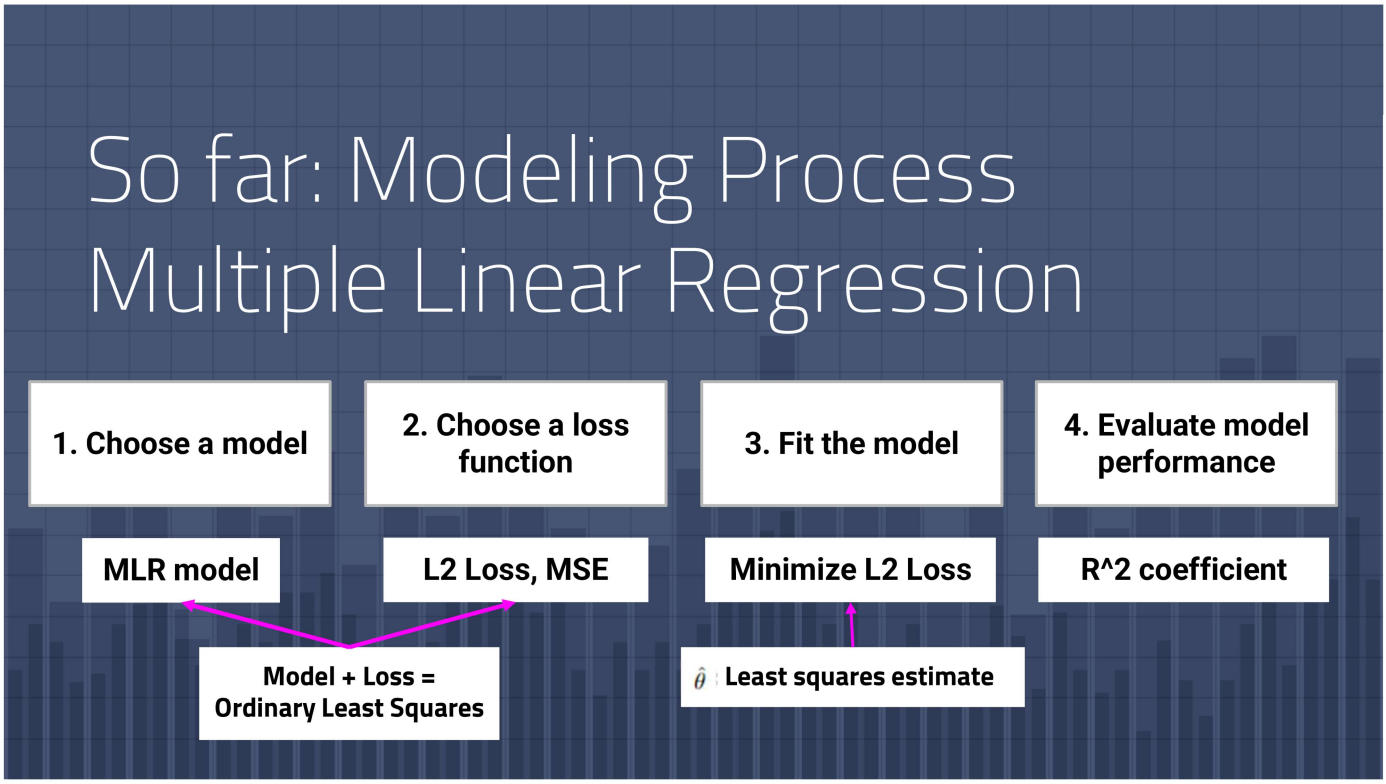
Normal Equation

Modeling Process
Multiple Linear Regression

4. Evaluate Model Performance

Multiple R^2, also called the **coefficient of determination**

$$R^2 = \frac{\text{variance of fitted values}}{\text{variance of } y} = \frac{\sigma_{\hat{y}}^2}{\sigma_y^2}$$


So far: Modeling Process
Multiple Linear Regression
1. Choose a model
2. Choose a loss function
3. Fit the model
4. Evaluate model performance
MLR model
L2 Loss, MSE
Minimize L2 Loss
R^2 coefficient
Model + Loss =
Ordinary Least Squares
$\hat{\theta}$ Least squares estimate

2. Which of the following are true about the optimal solution $\hat{\theta}$ to ordinary least squares (OLS)? Recall that the least squares estimate $\hat{\theta}$ solves the normal equation $(\mathbb{X}^T\mathbb{X})\theta = \mathbb{X}^T\mathbb{Y}$.

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

Hint: OLS optimizes the MSE

- [x] A. Using the normal equation, we can derive an optimal solution for simple linear regression with an L_2 loss.
- [ ] B. Using the normal equation, we can derive an optimal solution for simple linear regression with an L_1 loss.
- [x] C. Using the normal equation, we can derive an optimal solution for a constant model with an L_2 loss.
- [ ] D. Using the normal equation, we can derive an optimal solution for a constant model with an L_1 loss.
- [x] E. Using the normal equation, we can derive an optimal solution for the model specified option B in question 1 ($\hat{y} = \theta_1 x + \theta_2 \sin(x^2)$).

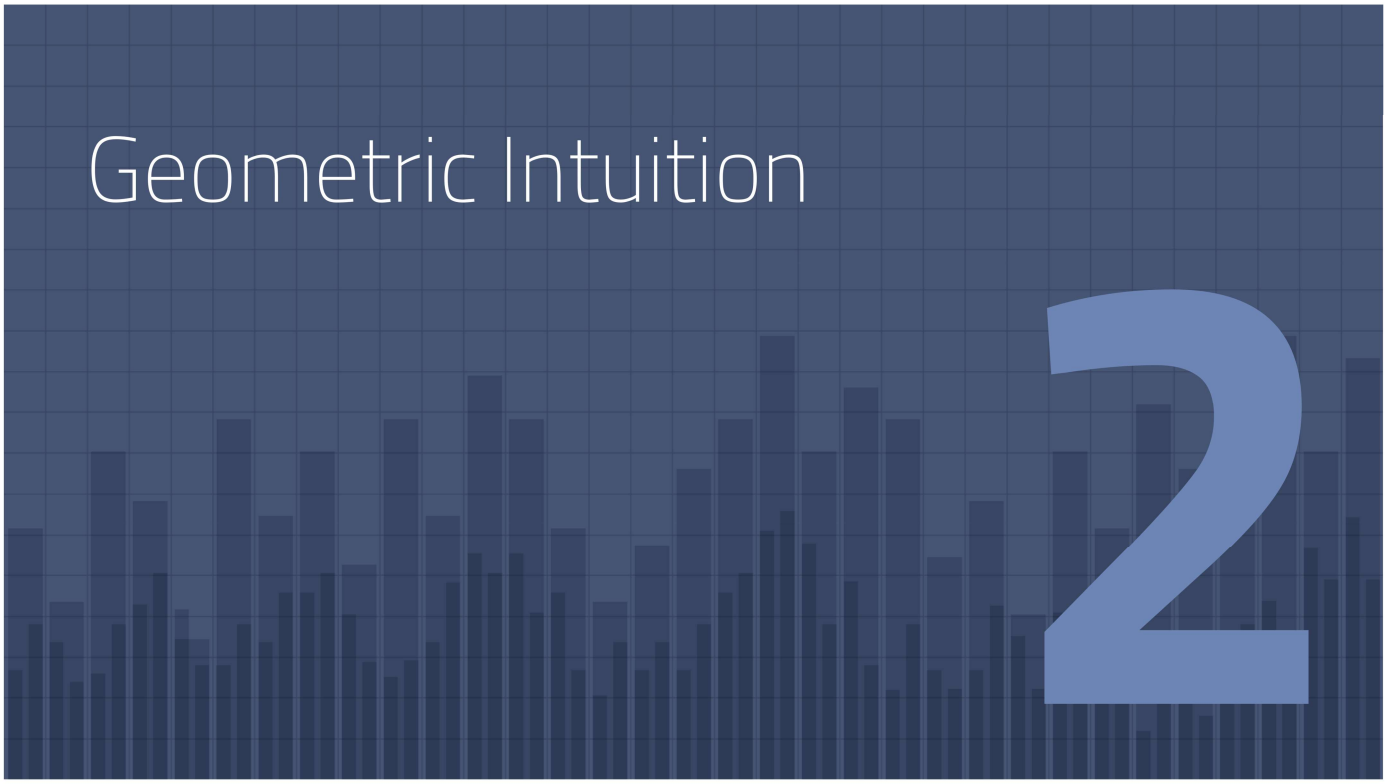
3. Which of the following conditions are required for the least squares estimate in Question 2?

- [x] A. $\mathbb{X}$ must be full column rank.
- [ ] B. $\mathbb{Y}$ must be full column rank.
- [ ] C. $\mathbb{X}$ must be invertible.
- [ ] D. $\mathbb{X}^T$ must be invertible.

2 vector

$\begin{bmatrix} a \\ a \end{bmatrix}$ X^TX to be invertible

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & \\ \vdots & & \ddots \\ 1 & & \end{bmatrix}$$


Geometric Intuition
2

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

So far, we've thought of our model as horizontally stacked predictions per datapoint:

$$\underset{n}{\begin{pmatrix} | \\ \hat{\mathbb{Y}} \\ | \end{pmatrix}} = \begin{bmatrix} \text{---}\, x_1^T \,\text{---} \\ \text{---}\, x_2^T \,\text{---} \\ \vdots \\ \text{---}\, x_n^T \,\text{---} \end{bmatrix} \underset{p+1}{\begin{pmatrix} | \\ \theta \\ | \end{pmatrix}} =$$

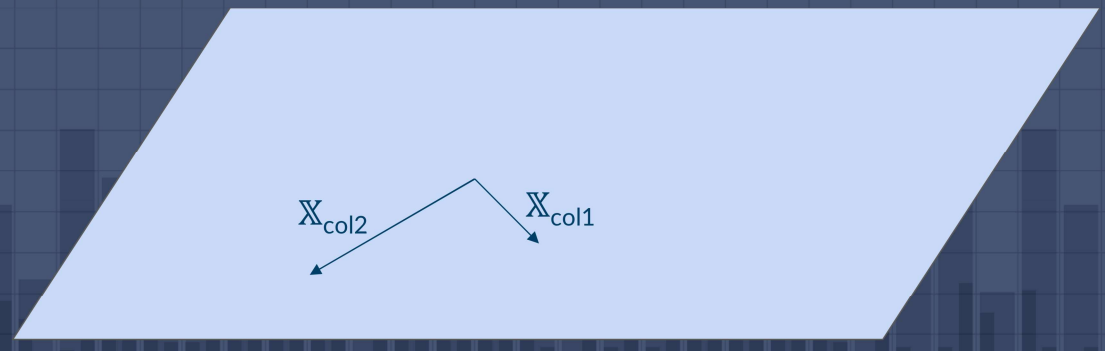
We can also think of $\hat{\mathbb{Y}}$ as a **linear combination of feature vectors**, scaled by **parameters**.

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

30

Space that can be reached any combination of columns of X
span(X)

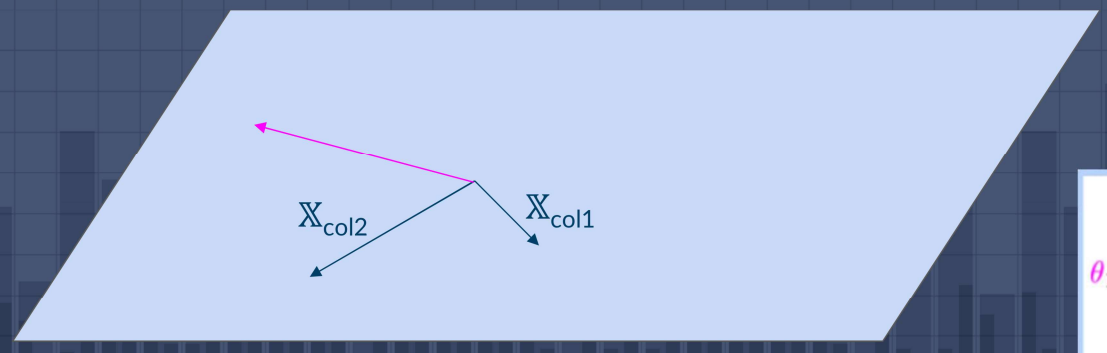
$\mathbb{X}\,\theta$



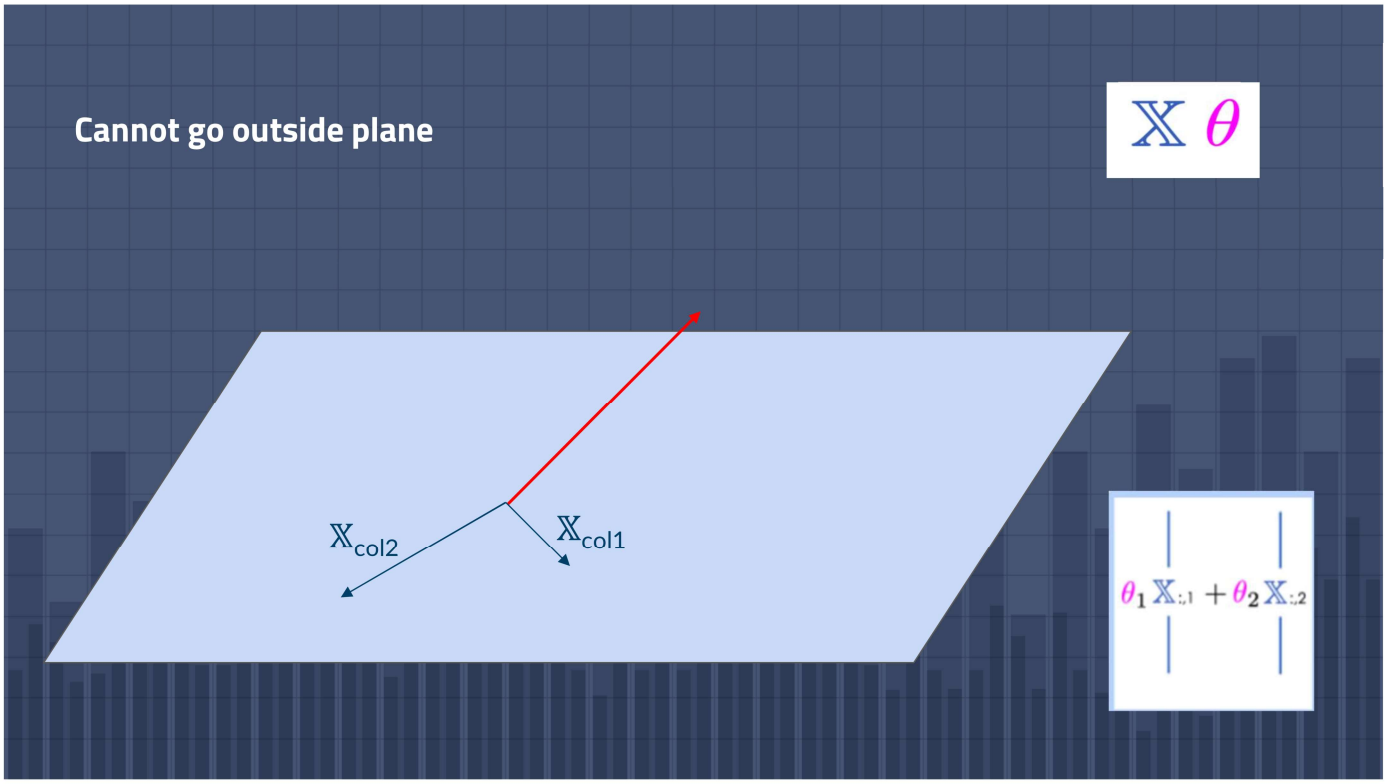
Space that can be reached by any combination of columns of X
- span(X)

Could be any linear combination (e.g. this could be -2*col1 + 0.7*col2

$$\mathbb{X}\,\theta$$

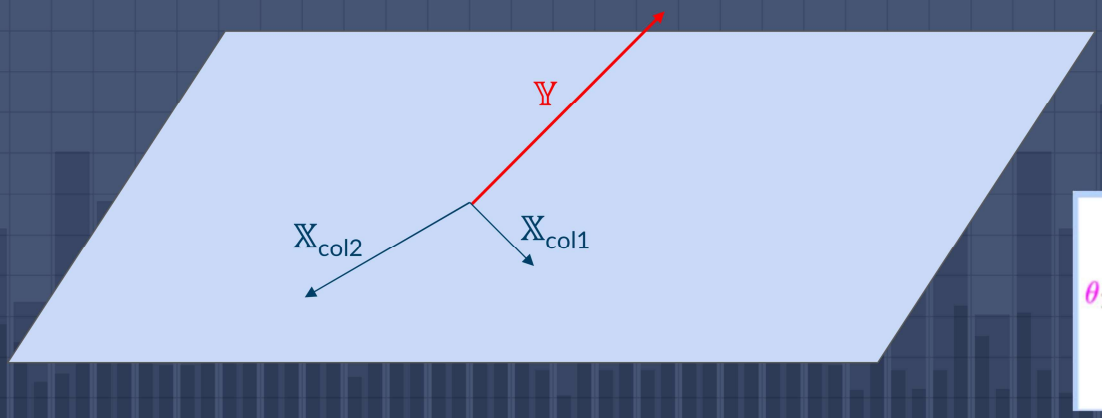


$$\theta_1 \mathbb{X}_{:1} + \theta_2 \mathbb{X}_{:2}$$


Cannot go outside plane
$\mathbb{X}\,\theta$
$\mathbb{X}_{col2}$
$\mathbb{X}_{col1}$
$\theta_1 \mathbb{X}_{:,1} + \theta_2 \mathbb{X}_{:,2}$

However, Y need not be on the plane

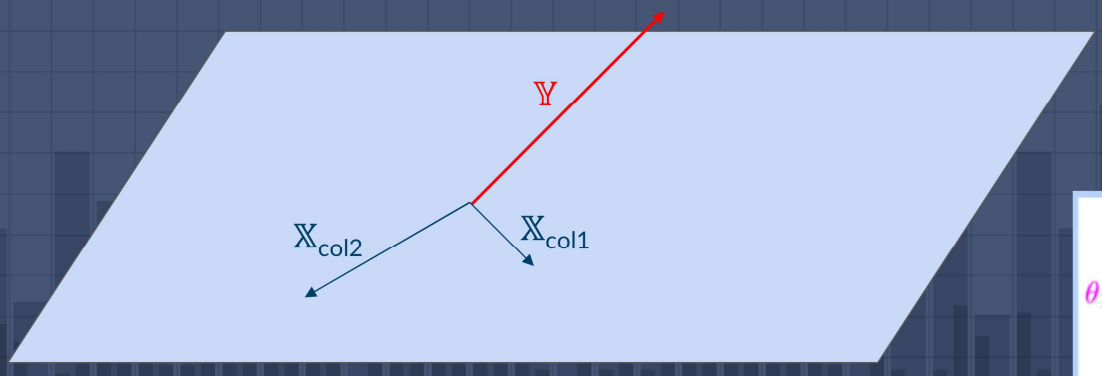
$\mathbb{X}\,\theta$



$$\theta_1 \mathbb{X}_{:1} + \theta_2 \mathbb{X}_{:2}$$

How do we predict Y? Make a guess along plane that is closest

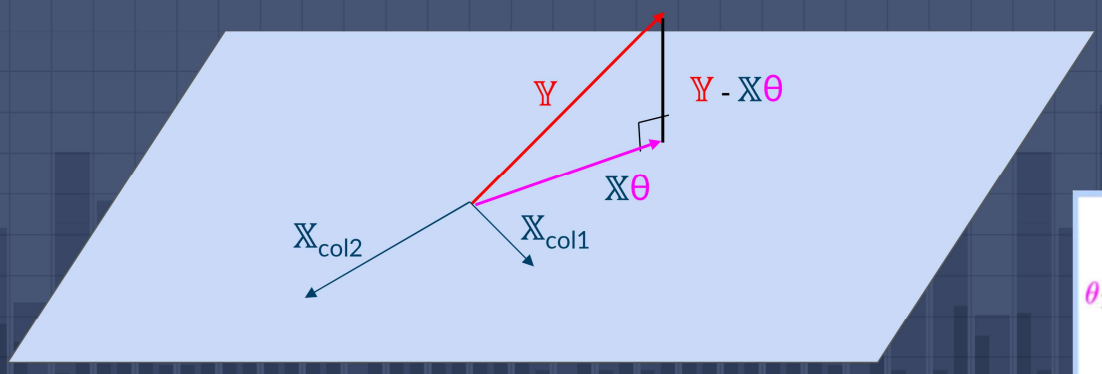
$\mathbb{X}\,\theta$



$$\theta_1 \mathbb{X}_{:,1} + \theta_2 \mathbb{X}_{:,2}$$

How do we predict Y? Make a guess along plane that is closest
How do determine closest? Drop a perpendicular
$\mathbb{X}\theta$ connects to the perpendicular

$\mathbb{X}\,\theta$

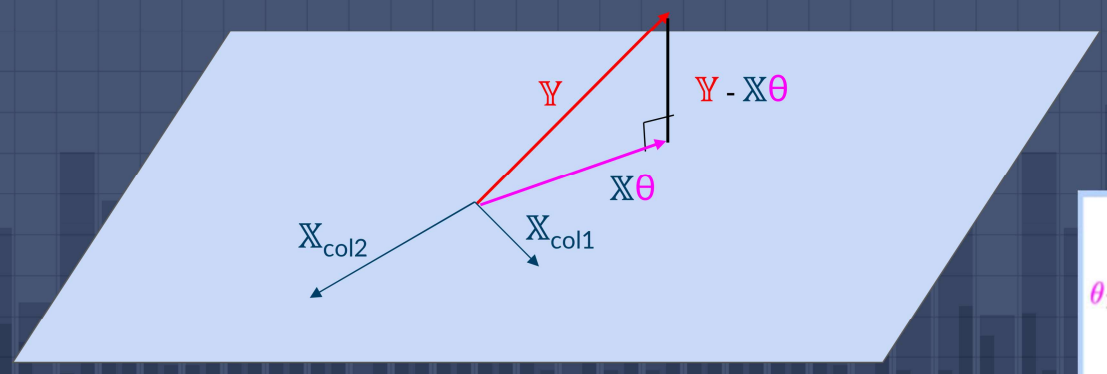


$$\theta_1 \mathbb{X}_{:1} + \theta_2 \mathbb{X}_{:2}$$

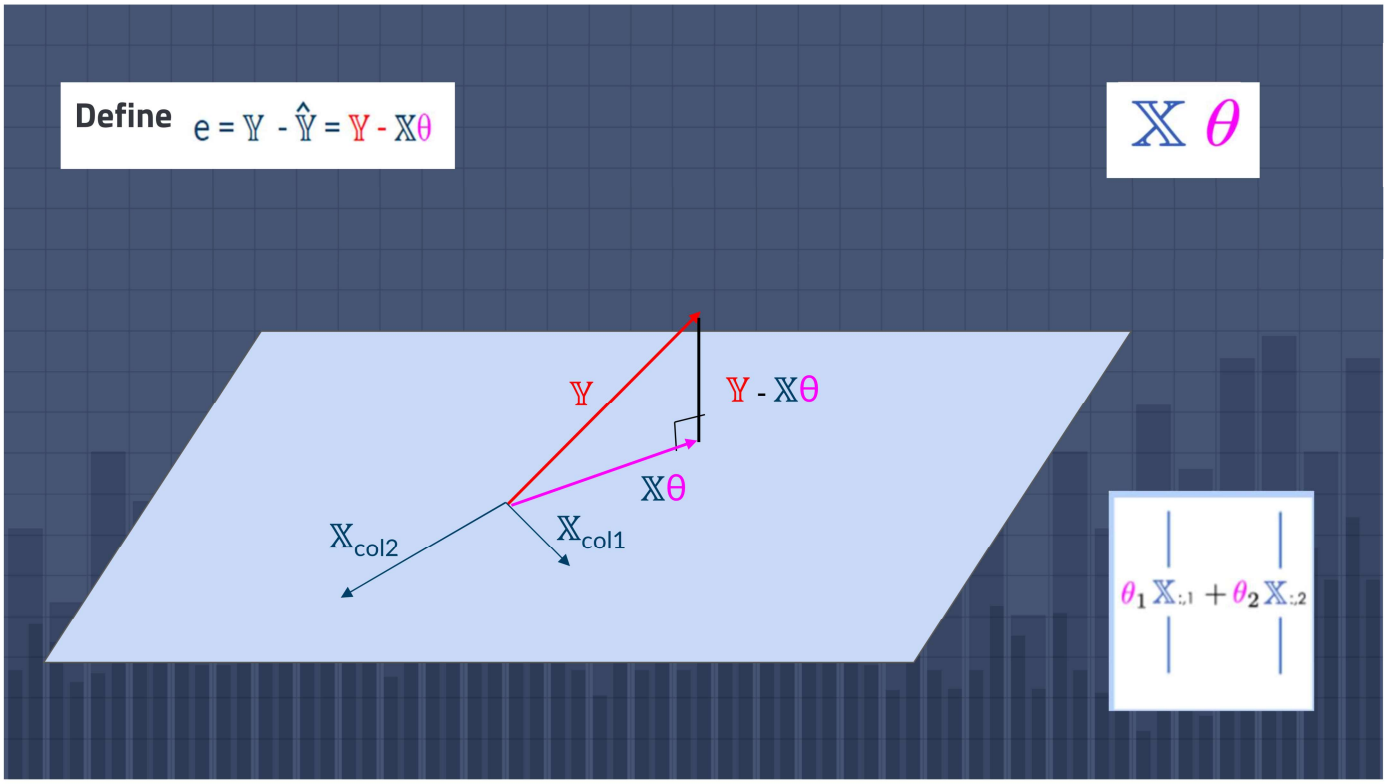
$\mathbb{X}\theta$ connects to the perpendicular

This is our best guess, $\hat{\mathbb{Y}}$ = $\mathbb{X}\theta$

$$\mathbb{X}\,\theta$$



$$\theta_1 \mathbb{X}_{:1} + \theta_2 \mathbb{X}_{:2}$$

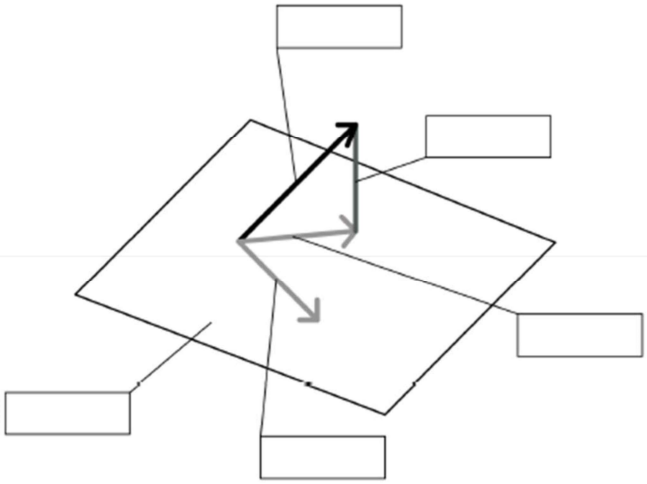

Define e = 𝕐 - $\hat{\mathbb{Y}}$ = 𝕐 - 𝕏θ
𝕏 θ
𝕐
𝕐 - 𝕏θ
𝕏θ
$\mathbb{X}_{col2}$
$\mathbb{X}_{col1}$
$\theta_1 \mathbb{X}_{:1} + \theta_2 \mathbb{X}_{:2}$

Some nice properties

- When using θ, residuals (e) are orthogonal to span($\mathbb{X}$) $\mathbb{X}^T e = 0$
- Linear models with an intercept terms WILL HAVE the sum of their residuals to be 0 $\sum_{i=1}^{n} e_i = 0$
- A least squares estimate θ is unique only if $\mathbb{X}$ is full column rank

4. Suppose we have a dataset represented with the design matrix span($\mathbb{X}$) and response vector $\mathbb{Y}$. We use linear regression to solve for this and obtain optimal weights as $\hat{\theta}$. Label the following terms on the geometric interpretation of ordinary least squares:

- $\mathbb{X}$ (i.e., span($\mathbb{X}$))
- The response vector $\mathbb{Y}$
- The residual vector $\mathbb{Y} - \mathbb{X}\hat{\theta}$
- The prediction vector $\mathbb{X}\hat{\theta}$ (using optimal parameters)
- A prediction vector $\mathbb{X}\alpha$ (using an arbitrary vector α).



(a) What is always true about the residuals in least squares regression? Select all that apply.

- ☐ A. They are orthogonal to the column space of the design matrix.
- ☐ B. They represent the errors of the predictions.
- ☐ C. Their sum is equal to the mean squared error.
- ☐ D. Their sum is equal to zero.
- ☐ E. None of the above.

(b) Which are true about the predictions made by OLS? Select all that apply.

- ☐ A. They are projections of the observations onto the column space of the design matrix.
- ☐ B. They are linear combinations of the features.
- ☐ C. They are orthogonal to the residuals.
- ☐ D. They are orthogonal to the column space of the features.

(c) We fit a simple linear regression to our data $(x_i, y_i), i = 1, 2, 3$, where x_i is the independent variable and y_i is the dependent variable. Our regression line is of the form $\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x$. Suppose we plot the relationship between the residuals of the model and the $\hat{y}s$, and find that there is a curve. What does this tell us about our model?

- ☐ A. The relationship between our dependent and independent variables is well represented by a line.
- ☐ B. The accuracy of the regression line varies with the size of the dependent variable.
- ☐ C. The variables need to be transformed, or additional independent variables are needed.

(d) Which are the following is true of the mystery quantity $\vec{v} = (I - \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T)\mathbb{Y}$?

- ☐ A. The vector $\vec{v}$ represents the residuals for any linear model.
- ☐ B. If the $\mathbb{X}$ matrix contains the $\vec{1}$ vector, then the sum of the elements in vector $\vec{v}$ is 0 (i.e. $\sum_i v_i = 0$).
- ☐ C. All the column vectors x_i of $\mathbb{X}$ are orthogonal to $\vec{v}$.
- ☐ D. If $\mathbb{X}$ is of shape n by p, there are p elements in vector $\vec{v}$.
- ☐ E. For any α, $\mathbb{X}\alpha$ is orthogonal to $\vec{v}$.