Data100 Sp22 Disc 5 Modeling/Loss

Attendance:

https://tinyurl.com/disc5michelle

Announcements

Due Dates

-Homework 5 due March 3 (start early)

-Lab 4 due Feb 22

-Weekly Check 5 due Feb 28

Other

-Review session Feb19 (tomorrow lol) 1-4pm

* I’ll be teaching visualizations :D *

Transformations

1

Motivation

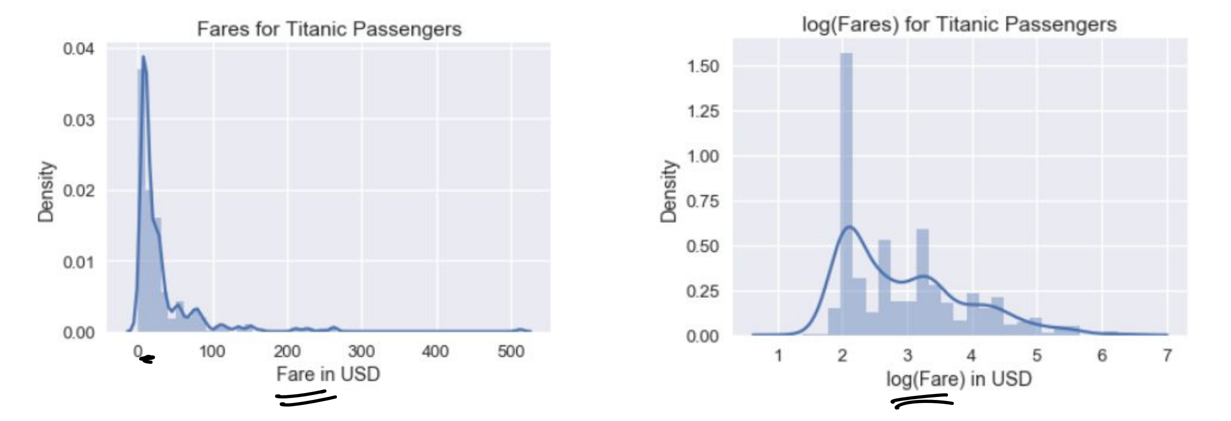
1. Transformations can help ‘normalize’ skewed data

- Normal curve has several nice properties (e.g 68-95-99.7 rule)
- Left Skew -> Square or Cube Data
- Right Skew -> Square Root or Log Data

Motivation

1. Transformations can help 'normalize' skewed data

- Normal curve has several nice properties (e.g 68-95-99.7 rule)
- Left Skew -> Square or Cube x-axis
- Right Skew -> Square Root or Log x-axis



Motivation

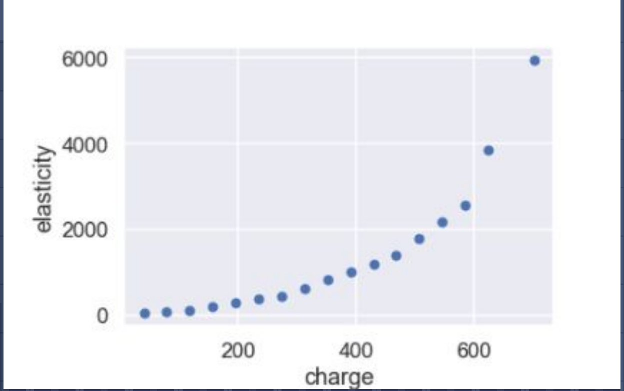
2. Extract hidden linear relationships

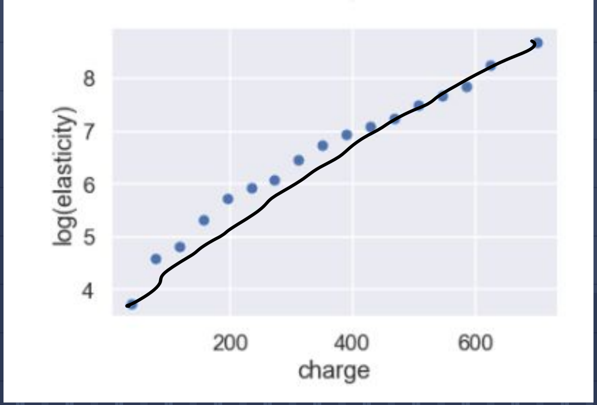
- Often difficult to visualize relationships when our data is non-linear
- Transforming data can reveal hidden linear relationships!
- We like linear relationships because they are easy to model!

Motivation

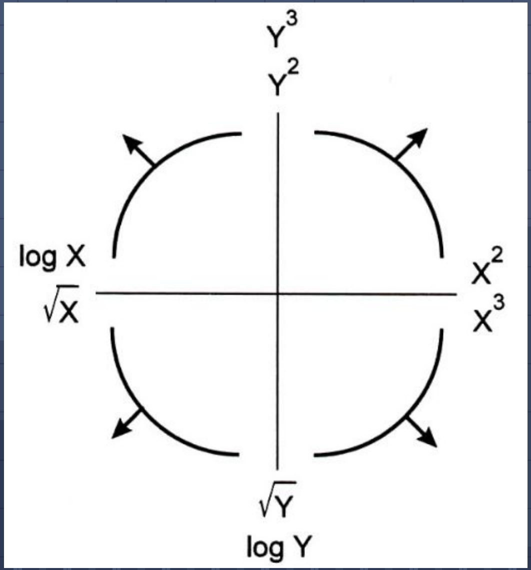
2. Extract hidden linear relationships

- Often difficult to visualize relationships when our data is non-linear
- Transforming data can reveal hidden linear relationships!
- We like linear relationships because they are easy to model!



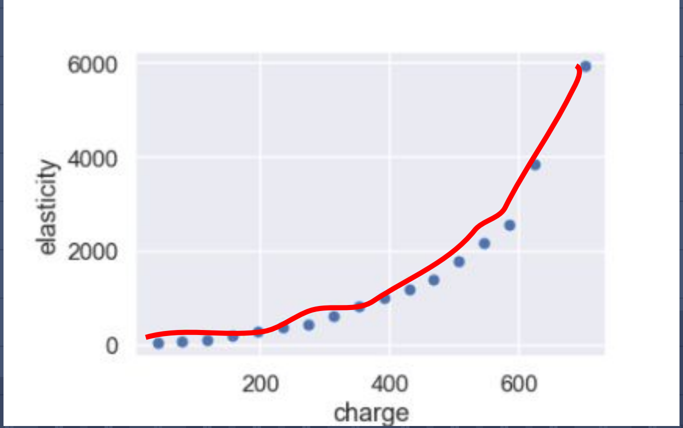


What transformation do I use?

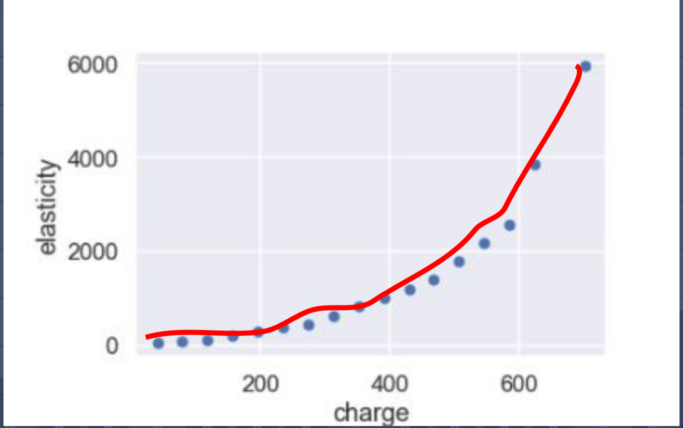


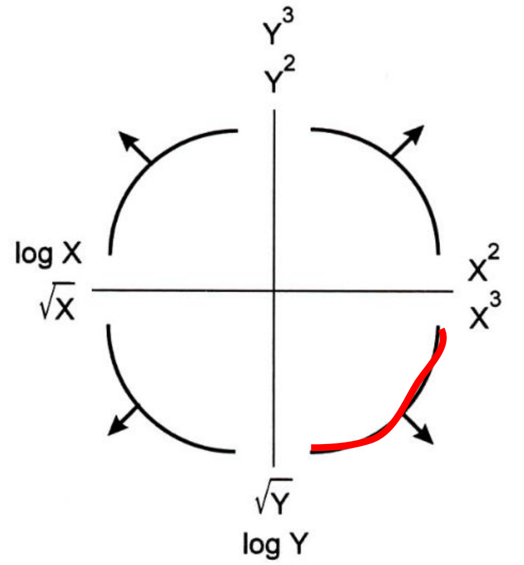
Pls save this!!

What transformation do I use?

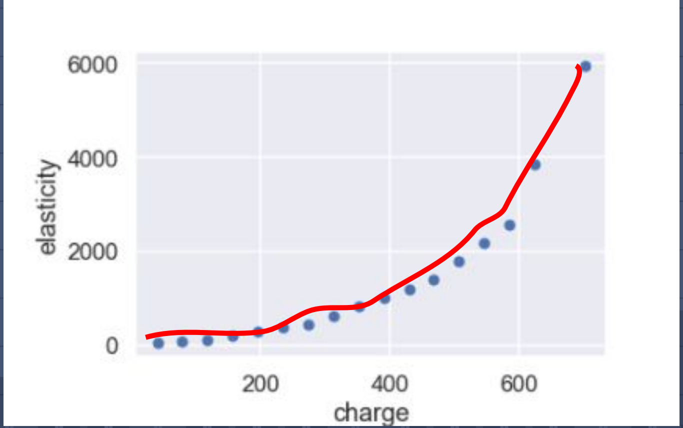


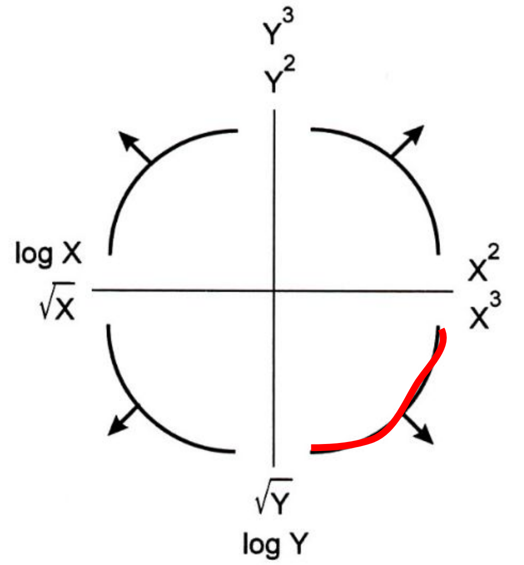
What transformation do I use?





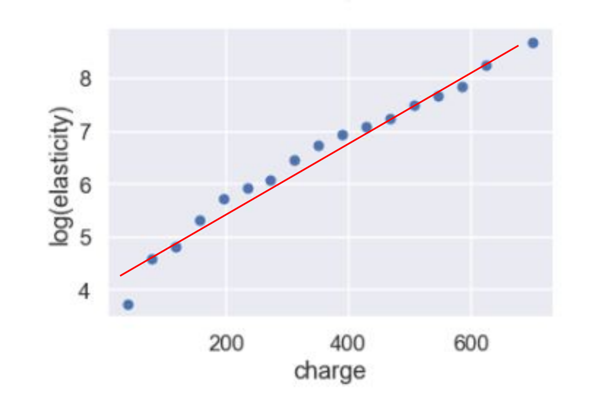
What transformation do I use?



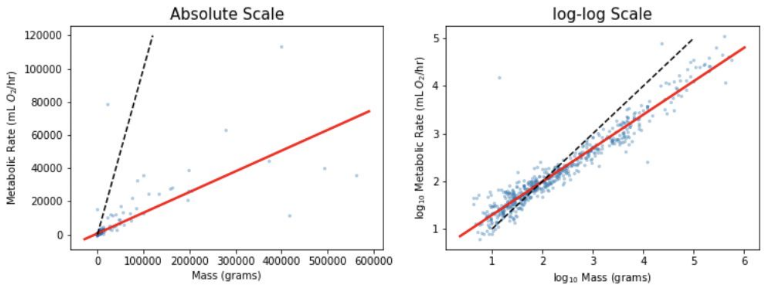


- Use log(Y) = log(elasticity) or X^2 = charge^2

What transformation do I use?



- Use log(Y) = log(elasticity) or X^2 = charge^2



(a) Let C and k be some constants and x and y represent mass and metabolic rate, respectively. Based on the plots, which of the following best describe the pattern seen in the data? Reminder: $log(ab) = log(a) + log(b)$.

○ A. $y = C + kx$ ○ B. $y = C \times 10^{kx}$ ○ C. $y = C + k\log_{10}(x)$ [✓ D. $y = Cx^k$]

(b) What parts of the plots could you use to make initial guesses on C and k?

$C = 10^a$ => a is y-int of Log-Log graph

k = slope of Log-Log graph

(c) Your friend points to the solid line on the log-log plot and says "since this line is going up and to the right, we can say that, in general, the bigger a mammal is, the greater its metabolic rate". Is this a reasonable interpretation of the plot?

Yes. slope > 0

$$\boxed{\log_{10} y = \underbrace{a}_{\text{y-inte}} + \underbrace{k}_{\text{slope}} \log_{10} x}$$

$$10^{\log_{10} y} = 10^{a + k\log_{10} x}$$

$$y = \underbrace{10^a}_{C} \cdot 10^{k\log_{10} x}$$

$$= C\left(10^{\log_{10} x}\right)^k$$

$$y = Cx^k$$

~~$C = 10$~~

$C = 10^a$

KDE

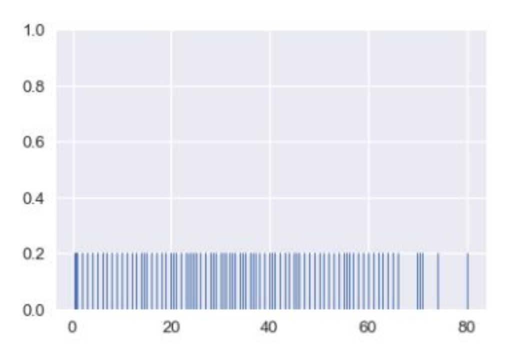
2

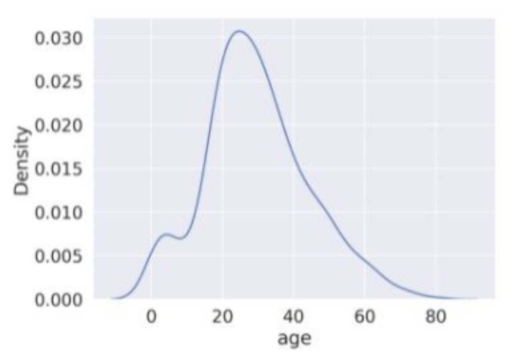
What is KDE?

- Kernel Density Estimation allows us to estimate **density curve** (probability density function)
    - Total area under curve must sum to 1

Why use KDE?

- Why do this?
  - ‘Smoothing’ 1-dimensional data





Why KDE over histogram?

- Can vary the alpha to make things more interpretable using KDE
- KDE gives us a better sense of the underlying structure (density curve) of the data -> better analysis

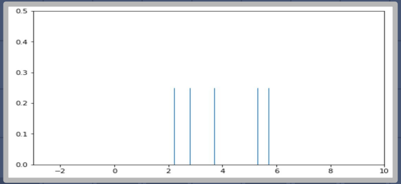
Three Steps to Create a KDE

- Place a kernel at each data point
  - E.g. Gaussian Kernel w bandwidth alpha=1 (creates tiny normals)
  - Can pick other kernels too!

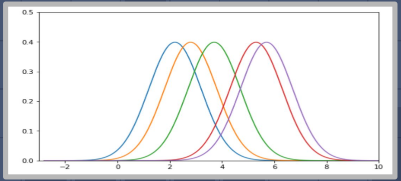
- Normalize (scale) kernels
  - Total area must be one!

- Sum kernels

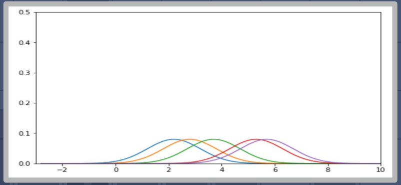
Three Steps to Create a KDE



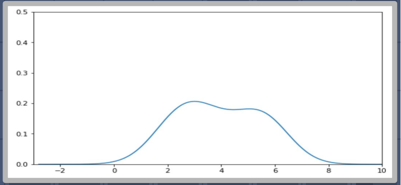
Step 1 - Placing Kernel



Step 2 - Normalizing

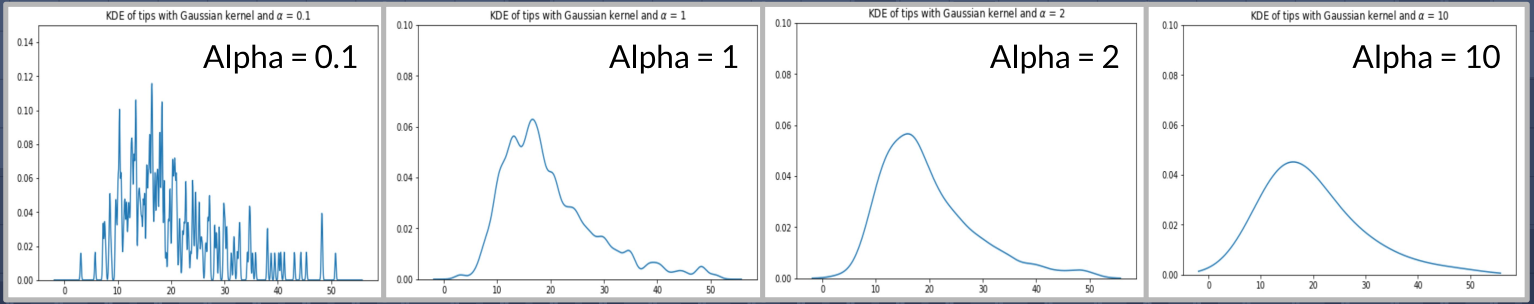


Step 3 - Summing



Alpha

- Alpha is bandwidth parameter, aka. Smoothness
  - Higher alpha = more smooth....but be careful of TOO smooth
  - Losing structure present in data



KDE of tips with Gaussian kernel and α = 0.1

Alpha = 0.1

KDE of tips with Gaussian kernel and α = 1

Alpha = 1

KDE of tips with Gaussian kernel and α = 2

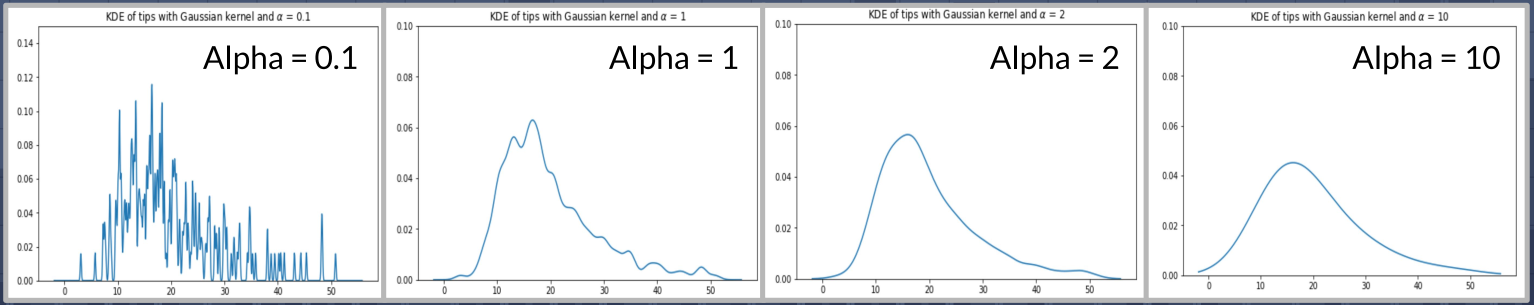
Alpha = 2

KDE of tips with Gaussian kernel and α = 10

Alpha = 10

Alpha (Bandwidth)

- Alpha is bandwidth parameter, aka. Smoothness
  - Higher alpha = more smooth....but be careful of TOO smooth
  - Losing structure present in data



KDE of tips with Gaussian kernel and α = 0.1

Alpha = 0.1

KDE of tips with Gaussian kernel and α = 1

Alpha = 1

KDE of tips with Gaussian kernel and α = 2

Alpha = 2

KDE of tips with Gaussian kernel and α = 10

Alpha = 10

Types of Kernels

Gaussian

$$K_\alpha(x, x_i) = \frac{1}{\sqrt{2\pi\alpha^2}} e^{-\frac{(x-x_i)^2}{2\alpha^2}}$$

Boxcar

$$K_\alpha(x, x_i) = \begin{cases} \frac{1}{\alpha}, & |x - x_i| \le \frac{\alpha}{2} \\ 0, & \text{else} \end{cases}$$

Loss

3

Modeling and Loss Functions

-Why do we use models?

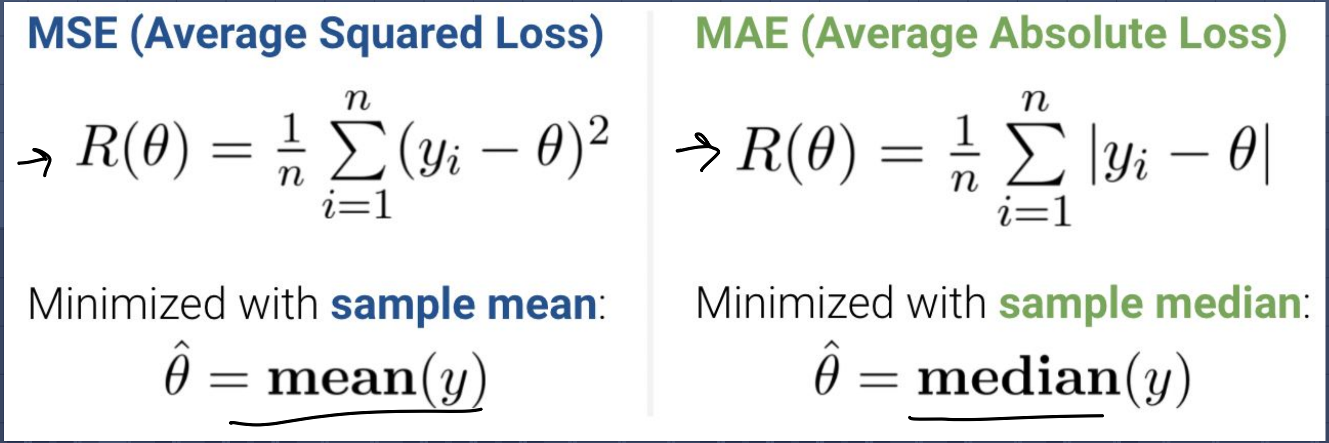
- Modeling is a way we represent the world, can help us understand data and make predictions
- Examples: Constant model, SLR (simple linear regression)

-How do we evaluate models?

- Loss functions!

L2 loss

L1 Loss



MSE (Average Squared Loss)

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

Minimized with **sample mean**:

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

MAE (Average Absolute Loss)

$$\rightarrow R(\theta) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \theta|$$

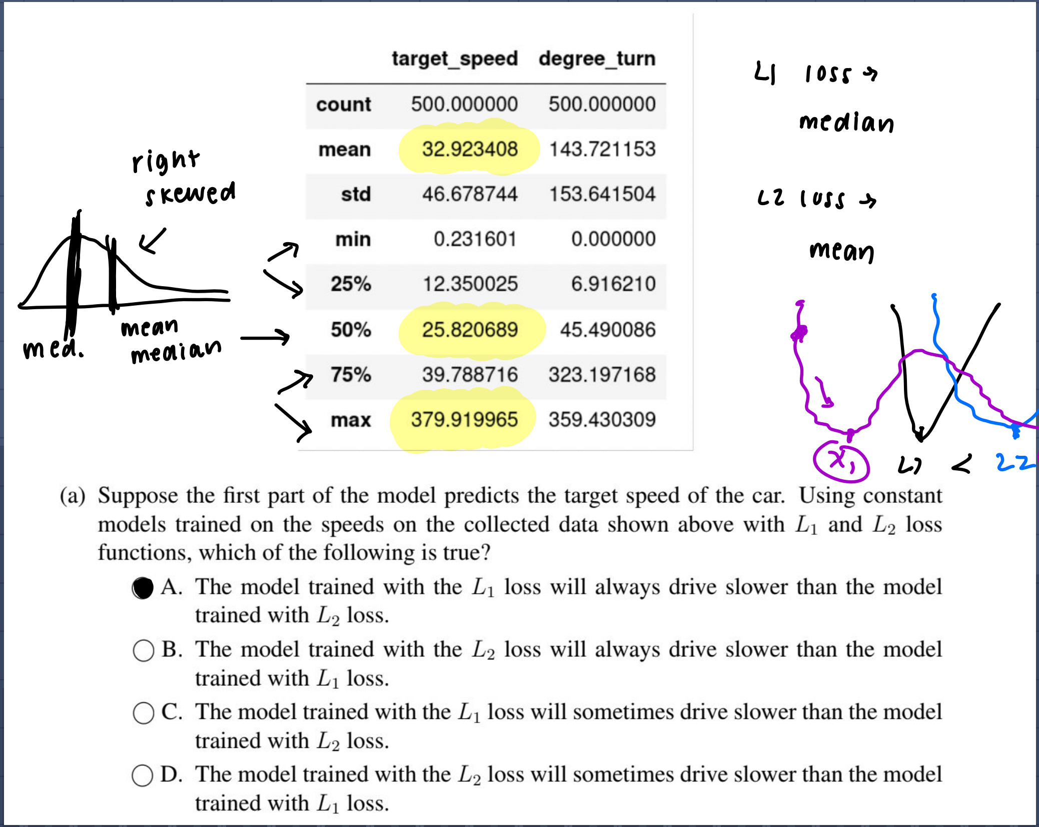
Minimized with **sample median**:

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

Choosing a loss function

-L1 loss is more robust, not affected by outliers as much

-Will compare L1 and L2 more later!


right skewed
med.
mean
median
x_1
x_2
x
$L_1 < L_2$

	target_speed	degree_turn
count	500.000000	500.000000
mean	32.923408	143.721153
std	46.678744	153.641504
min	0.231601	0.000000
25%	12.350025	6.916210
50%	25.820689	45.490086
75%	39.788716	323.197168
max	379.919965	359.430309

L1 loss → median

L2 loss → mean

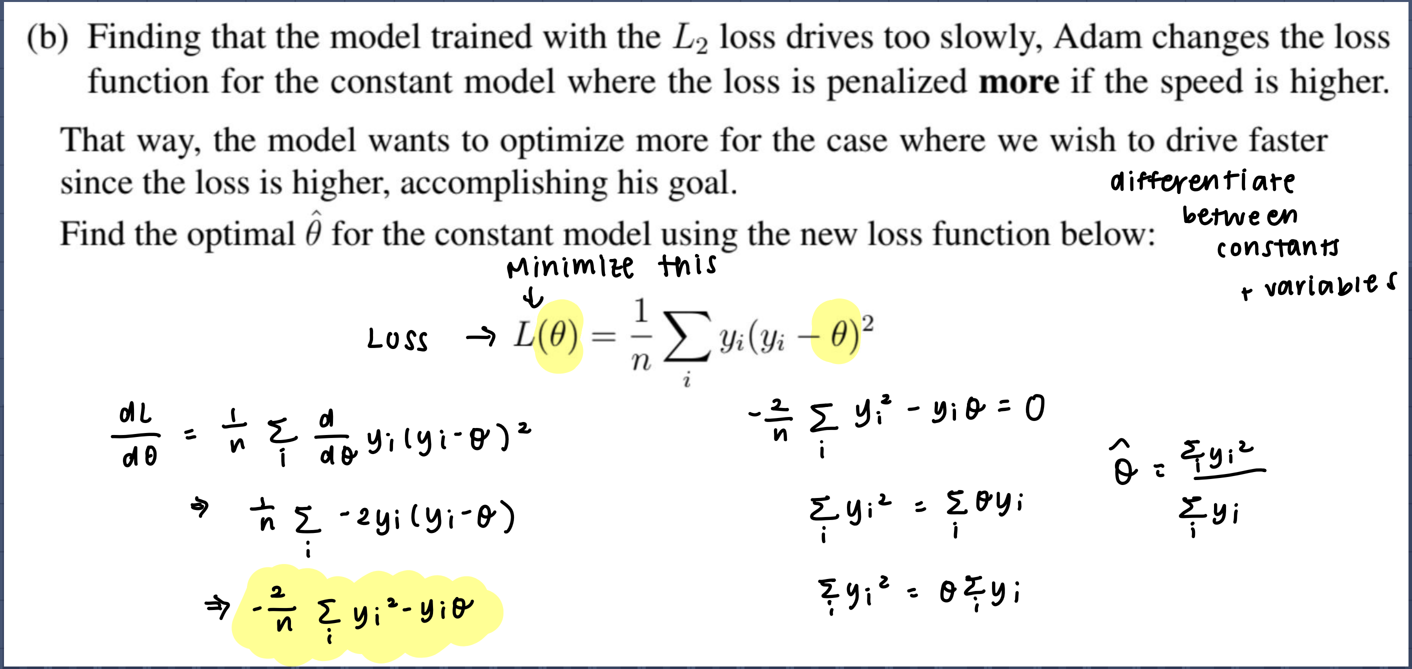
(a) Suppose the first part of the model predicts the target speed of the car. Using constant models trained on the speeds on the collected data shown above with L_1 and L_2 loss functions, which of the following is true?

● A. The model trained with the L_1 loss will always drive slower than the model trained with L_2 loss.

○ B. The model trained with the L_2 loss will always drive slower than the model trained with L_1 loss.

○ C. The model trained with the L_1 loss will sometimes drive slower than the model trained with L_2 loss.

○ D. The model trained with the L_2 loss will sometimes drive slower than the model trained with L_1 loss.



(b) Finding that the model trained with the L_2 loss drives too slowly, Adam changes the loss function for the constant model where the loss is penalized **more** if the speed is higher.

That way, the model wants to optimize more for the case where we wish to drive faster since the loss is higher, accomplishing his goal.

Find the optimal $\hat{\theta}$ for the constant model using the new loss function below:

differentiate between constants + variables

Minimize this ↓

$$\text{Loss} \rightarrow L(\theta) = \frac{1}{n}\sum_i y_i(y_i - \theta)^2$$

$$\frac{dL}{d\theta} = \frac{1}{n}\sum_i \frac{d}{d\theta} y_i(y_i-\theta)^2$$

$$\Rightarrow \frac{1}{n}\sum_i -2y_i(y_i-\theta)$$

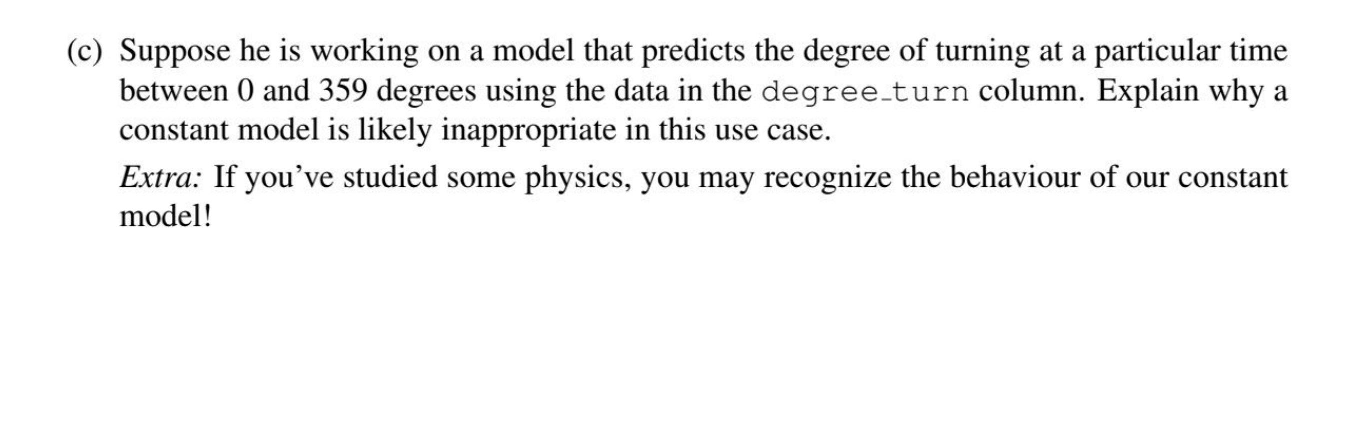
$$\Rightarrow -\frac{2}{n}\sum_i y_i^2 - y_i\theta$$

$$-\frac{2}{n}\sum_i y_i^2 - y_i\theta = 0$$

$$\sum_i y_i^2 = \sum_i \theta y_i$$

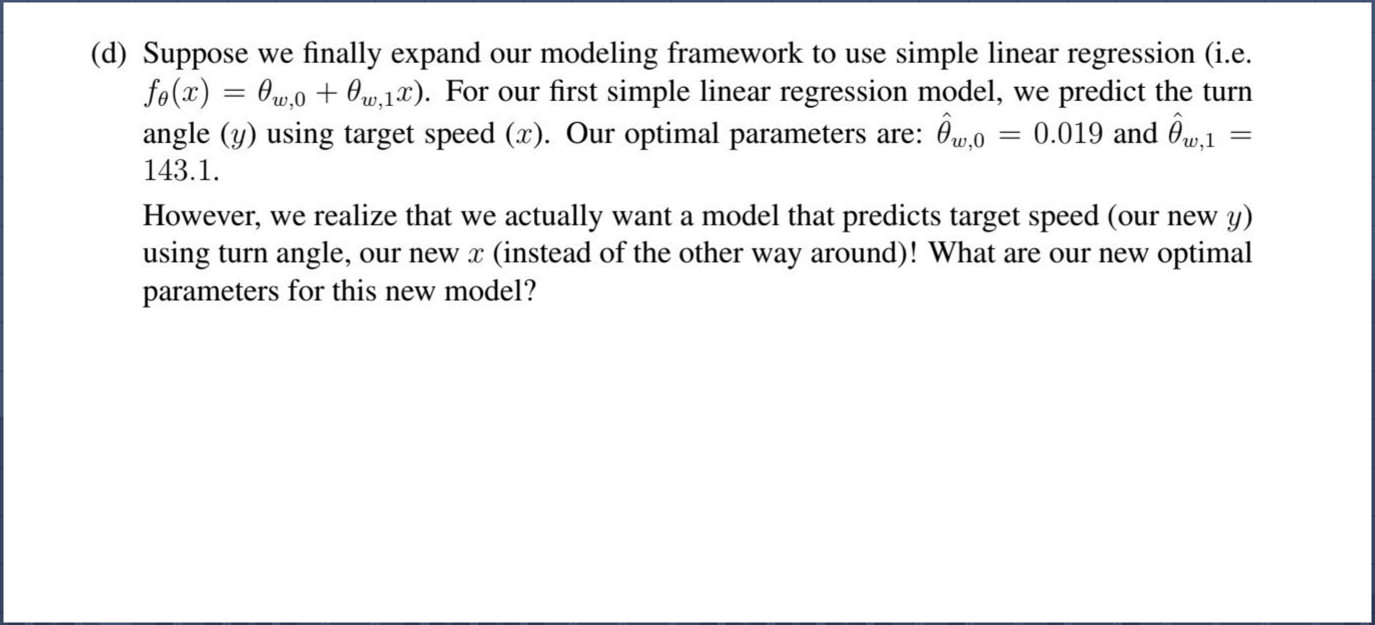
$$\sum_i y_i^2 = \theta\sum_i y_i$$

$$\hat{\theta} = \frac{\sum_i y_i^2}{\sum_i y_i}$$



(c) Suppose he is working on a model that predicts the degree of turning at a particular time between 0 and 359 degrees using the data in the `degree_turn` column. Explain why a constant model is likely inappropriate in this use case.

Extra: If you've studied some physics, you may recognize the behaviour of our constant model!



(d) Suppose we finally expand our modeling framework to use simple linear regression (i.e. $f_\theta(x) = \theta_{w,0} + \theta_{w,1}x$). For our first simple linear regression model, we predict the turn angle (y) using target speed (x). Our optimal parameters are: $\hat{\theta}_{w,0} = 0.019$ and $\hat{\theta}_{w,1} = 143.1$.

However, we realize that we actually want a model that predicts target speed (our new y) using turn angle, our new x (instead of the other way around)! What are our new optimal parameters for this new model?