**Data Science 100**

**Lecture 13: Modeling and Estimation**

Slides by:
Joseph E. Gonzalez, jegonzal@berkeley.edu
2018 updates - Fernando Perez, fernando.perez@berkele.edu

Recap ... so far we have covered
- **Data collection**: Surveys, sampling, administrative data
- **Data cleaning and manipulation**: Pandas, text & regexes.
- **Exploratory Data Analysis**
  - Joining and grouping data
  - Structure, Granularity, Temporality, Faithfulness and Scope
  - Basic exploratory data visualization
- **Data Visualization**:
  - Kinds of visualizations and the use of size, area, and color
  - Data transformations using Tukey Mosteller bulge diagram
- **An introduction to database systems and SQL**

---

**Today – Models & Estimation**

**What is a model?**

A model is an an **idealized** representation of a system

- Atoms don't actually work like this...
- Proteins are far more complex
- We haven't really seen one of these.

“Essentially, all models are wrong, but some are useful.”

George Box
Statistician
1919-2013
Why do we build models?

- Models enable us to make **accurate predictions**

- **Provide insight** into complex phenomena

**Models: Statistical correlations (A)**

**Models: statistical correlations (B)**
Models and the World

- **Data Generation Process:** the real-world phenomena from which the data is collected
  - Example: everyday there are some number of clouds and it rains or doesn’t?
  - We don’t know or can’t compute this, could be stochastic or adversarial

- **Model:** a theory of the data generation process
  - Example: if there are more than X clouds then it will rain
  - How do we pick this model? EDA? Art?
  - May not reflect reality — “all models are wrong ...”

- **Estimated Model:** an instantiation of the model
  - Example: if there are more than 42 clouds then it will rain
  - How do we estimate it?
  - What makes the estimate “good”?

---

Example – Restaurant Tips

Follow along with the notebook ...

Step 1: Understanding the Data (EDA)

Collected by a single waiter over a month

- **Predict** which tables will tip the highest
- **Understand** relationship between tables and tips

Understanding the Tips

- **Right skewed**
- **Mode around $15**
- **Mean around $20**
- **No large bills**

Derived Variable: Percent Tip

\[
pct\_tip = \frac{\text{tip}}{\text{total\_bill}} \times 100
\]

- **Natural representation of tips**
- **Why?** Tradition in US is to tip %
- **Issues in the plot?**
  - **Outliers**
  - **What makes the estimate “good”?**
  - **Small bills... bad data?**
  - **Transformations?**
  - **Remove outliers**
Step 1: Define the Model
START SIMPLE!!

How do we estimate the parameter $\theta^*$
- Guess a number using prior knowledge: 15%
- Use the data! How?
- Estimate the value $\theta^*$ as:
  - the percent tip from a randomly selected receipt
  - the mean of the percent tips observed
  - the median of the percent tips observed
- Which is the best? How do I define best?
  - Depends on our goals ...

Defining an the Objective (Goal)
- Ideal Goal: estimate a value for $\theta^*$ such that the model makes good predictions about the future.
  - Great goal! Problem?
  - We don’t know the future. How will we know if our estimate is good?
  - There is hope! … we will return to this goal … in the future …
- Simpler Goal: estimate a value for $\theta^*$ such that the model “fits” the data
  - What does it mean to “fit” the data?
  - We can define a loss function that measures the error in our model on the data

Step 2: Define the Loss
“Take the Loss”

Loss Functions
- Loss function: a function that characterizes the cost, error, or loss resulting from a particular choice of model or model parameters.
  - Many definitions of loss functions and the choice of loss function affects the accuracy and computational cost of estimation.
  - The choice of loss function depends on the estimation task
  - quantitative (e.g., tip) or qualitative variable (e.g., political affiliation)
  - Do we care about the outliers?
  - Are all errors equally costly? (e.g., false negative on cancer test)
Squared Loss

Widely used loss!

The predicted value
The "error" in our prediction

\[ L(\theta, y) = (y - \theta)^2 \]

An observed data point

- Also known as the the \( L^2 \) loss (pronounced “el two”)
- Reasonable?
  - \( \theta = y \) → good prediction → good fit → no loss!
  - \( \theta \) far from \( y \) → bad prediction → bad fit → lots of loss!

Absolute Loss

It sounds worse than it is…

\[ L(\theta, y) = |y - \theta| \]

Absolute value

- Also known as the the \( L^1 \) loss (pronounced “el one”)
- Reasonable?
  - \( \theta = y \) → good prediction → good fit → no loss!
  - \( \theta \) far from \( y \) → bad prediction → bad fit → some loss

Huber Loss

Parameter \( \alpha \) that we need to choose.

- Reasonable?
  - \( \theta = y \) → good prediction → good fit → no loss!
  - \( \theta \) far from \( y \) → bad prediction → bad fit → some loss
- A hybrid of the \( L^2 \) and \( L^1 \) losses…

The Huber loss function, interactively

Comparing the Loss Functions

- All functions are zero when \( \theta = y \)
- Different penalties for being far from observations
- Smooth vs. not smooth
- Which is the best?
  - Let’s find out
Average Loss

A natural way to define the loss on our entire dataset is to compute the average of the loss on each record.

\[ L(\theta, D) = \frac{1}{n} \sum_{i=1}^{n} L(\theta, y_i) \]

- In some cases we might take a weighted average (when?)
- Some records might be more important or reliable
- What does the average loss look like?

Double Jeopardy
Name that Loss!

Name that loss

(a) Squared Loss
(b) Absolute Loss
(c) Huber Loss

Difference between Huber and L1

Zoomed in with only 5 data points sampled at random

Different Minimizers

Observations

Absolute and Huber Loss have nearly identical Values

Squared Loss is slightly to the right

Squared Loss / 10

Absolute Loss

Huber Loss

Corner
Sensitivity to Outliers

Small fraction of loss on outliers...

34% of loss due to a single point

Recap on Loss Functions

- **Loss functions**: a mechanism to measure how well a particular instance of a model fits a given dataset
- **Squared Loss**: sensitive to outliers but a smooth function
- **Absolute Loss**: less sensitive to outliers but not smooth
- **Huber Loss**: less sensitive to outliers and smooth but has an extra parameter to deal with
- Why is smoothness an issue → Optimization! ...

Summary of Model Estimation (so far...)

1. **Define the Model**: simplified representation of the world
   - Use domain knowledge but... keep it simple!
   - Introduce parameters for the unknown quantities
2. **Define the Loss Function**: measures how well a particular instance of the model "fits" the data
   - We introduced $L^2$, $L^1$, and Huber losses for each record
   - Take the average loss over the entire dataset
3. **Minimize the Loss Function**: find the parameter values that minimize the loss on the data
   - So far we have done this graphically
   - Now we will minimize the loss analytically

Step 3: Minimize the Loss

Minimizing a Function

- Suppose we want to minimize:
  \[ f(\theta) = (\theta - 3)^2 \]
- Solve for derivative = 0:
  \[ \frac{\partial}{\partial \theta} f(\theta) = 2(\theta - 3) = 0 \]
- Procedure:
  1. Take derivative
  2. Set equal to zero
  3. Solve for parameters

A Brief Review of Calculus
Quick Review of the Chain Rule

How do I compute the derivative of composed functions?

\[ \frac{\partial}{\partial \theta} h(\theta) = \frac{\partial}{\partial \theta} f(g(\theta)) = \left( \frac{\partial}{\partial u} f(u) \right)_{u=g(\theta)} \frac{\partial}{\partial \theta} g(\theta) \]

Derivative of \( f \) evaluated at \( g(\theta) \)

Using the Chain Rule

First application of chain rule

\[ \frac{\partial}{\partial \theta} \exp \left( \sin (\theta^2) \right) = \left( \frac{\partial}{\partial u} \exp (u) \right)_{u=\sin(\theta^2)} \frac{\partial}{\partial \theta} \sin (\theta^2) \]

Derivative of exponent

Substituting \( u = \exp (\sin (\theta^2)) \)

\[ \frac{\partial}{\partial \theta} \sin (\theta^2) = \left( \frac{\partial}{\partial u} \sin (u) \right)_{u=\exp (\sin (\theta^2))} \frac{\partial}{\partial \theta} \theta^2 \]

Second application of the chain rule

Computing the remaining derivative

\[ \frac{\partial}{\partial \theta} \sin (\theta^2) \cos (\theta^2) \frac{\partial}{\partial \theta} \theta^2 = \exp (\sin (\theta^2)) \cos (\theta^2) \frac{\partial}{\partial \theta} \theta^2 \]

\[ = \exp (\sin (\theta^2)) \cos (\theta^2) 2\theta \]

Bonus material (not covered in lecture) but useful for studying

Bonus material (not covered in lecture) but useful for studying

**Quick Review of the Chain Rule**

- How do I compute the derivative of composed functions?

\[ \frac{\partial}{\partial \theta} h(\theta) = \frac{\partial}{\partial \theta} f(g(\theta)) \]

\[ = \left( \frac{\partial}{\partial u} f(u) \right)_{u=g(\theta)} \frac{\partial}{\partial \theta} g(\theta) \]

**Using the Chain Rule**

- First application of chain rule

\[ \frac{\partial}{\partial \theta} \exp \left( \sin (\theta^2) \right) = \left( \frac{\partial}{\partial u} \exp (u) \right)_{u=\sin(\theta^2)} \frac{\partial}{\partial \theta} \sin (\theta^2) \]

- Derivative of exponent

\[ \frac{\partial}{\partial \theta} \sin (\theta^2) = \left( \frac{\partial}{\partial u} \sin (u) \right)_{u=\exp (\sin (\theta^2))} \frac{\partial}{\partial \theta} \theta^2 \]

- Second application of the chain rule

\[ \left( \frac{\partial}{\partial \theta} \sin (\theta^2) \right) \cos (\theta^2) \frac{\partial}{\partial \theta} \theta^2 = \exp (\sin (\theta^2)) \cos (\theta^2) \frac{\partial}{\partial \theta} \theta^2 \]

\[ = \exp (\sin (\theta^2)) \cos (\theta^2) 2\theta \]

**Convex sets and polygons**

- No line segment between any two points on the boundary ever leaves the polygon.
- Equivalently, all angles are ≤ 180°.
- The interior is a convex set.

**Non-Convex sets and polygons**

- There is at least one line segment between two points on the boundary that leaves the set.
Formal Definition of Convex Functions

A function $f$ is convex if and only if:
\[
 tf(a) + (1 - t)f(b) \geq f(ta + (1 - t)b)
\]
\[\forall a, \forall b, t \in [0,1]\]

Convex or Not Convex

Are our previous loss functions convex?

Yes!

Average Loss?

Yes!

(Sum of convex functions is convex)

Is a Gaussian convex?

Yes!

Formal Proof

Suppose you have two convex functions $f$ and $g$:
\[
 tf(a) + (1 - t)f(b) \geq f(ta + (1 - t)b)
\]
\[
 tg(a) + (1 - t)g(b) \geq g(ta + (1 - t)a)
\]
\[\forall a, \forall b, t \in [0,1]\]

We would like to show:
\[
 th(a) + (1 - t)h(b) \geq h(ta + (1 - t)a)
\]

Where: $h(x) = f(x) + g(x)$

Bonus material (not covered in lecture) but useful for studying.
We would like to show:
\[ th(a) + (1 - t)h(b) \geq h(ta - (1 - t)a) \]

Where: \( h(x) = f(x) + g(x) \)

Starting on the left side
Substituting definition of \( h \):
\[ th(a) + (1 - t)h(b) = t(f(a) + g(a)) + (1 - t)(f(b) + g(b)) \]
Re-arranging terms:
\[ = [tf(a) + (1 - t)f(b)] + [tg(a) + (1 - t)g(b)] \]

Convexity in \( t \) \[ \geq f(ta + (1 - t)b) + [g(a) + (1 - t)g(b)] \]

Convexity in \( g \) \[ \geq f(ta + (1 - t)b) + g(ta + (1 - t)b) \]

Definition of \( h = h(ta + (1 - t)b) \)

Bonus material (not covered in lecture) but useful for studying

Minimizing the Average Squared Loss

\[ L_D(\theta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \theta)^2 \]

Take the derivative
\[ = -2 \frac{1}{n} \sum_{i=1}^{n} (y_i - \theta) \]

Set the derivative equal to zero
\[ 0 = -2 \frac{1}{n} \sum_{i=1}^{n} (y_i - \theta) \]
\[ 0 = \sum_{i=1}^{n} (y_i - \theta) \]
\[ 0 = \left( \sum_{i=1}^{n} y_i \right) - n\bar{\theta} \]
\[ 0 = \left( \sum_{i=1}^{n} y_i \right) - n\theta \]

The estimate for percent tip that minimizes the squared loss is the mean of the percent tips
We guessed that already!

Mean (Average)!
Minimizing the Average Absolute Loss

\[ L_D(\theta) = \frac{1}{n} \sum_{i=1}^{n} |y_i - \theta| \]

\[ \frac{\partial}{\partial \theta} L_D(\theta) = \frac{1}{n} \sum_{i=1}^{n} \text{sign}(y_i - \theta) \]

- Take the derivative
- How?
- What is sign(0)?

**Convention:**

\[ \text{sign}(0) = 0 \]

Using this procedure we discovered:

\[ \hat{\theta}_{L_2} = \frac{1}{n} \sum_{i=1}^{n} y_i = \text{mean}(D) \]

\[ \hat{\theta}_{L_1} = \text{median}(D) \]

Calculus for Loss Minimization

- General Procedure:
  - Verify that function is convex (we often will assume this...)
  - Compute the derivative
  - Set derivative equal to zero and solve for the parameters
  - Using this procedure we discovered:

The **median** minimizes the absolute loss → Robust! not sensitive to outliers

- Many optimal values
- Pick one?

Minimizing the Average Absolute Loss

- Take the derivative
- How?
- Derivative at the corner?
- What is the sign of 0?

Convention:

\[ \text{sign}(0) = 0 \]

Percent Tips in sorted order

<table>
<thead>
<tr>
<th>Percent Tips</th>
<th>2</th>
<th>2</th>
<th>2</th>
<th>0</th>
<th>8</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y1</td>
<td>Y3</td>
<td>Y5</td>
<td>Y2</td>
<td>Y4</td>
<td>Y2</td>
<td>Y1</td>
</tr>
</tbody>
</table>

Minimizing the Average Absolute Loss

- Take the derivative
- Set derivative to zero and solve for parameters

\[ \left( \sum_{y_i < \theta} 1 \right) = \left( \sum_{y_i > \theta} 1 \right) \]

\[ 0 = \left( \sum_{y_i < \theta} -1 \right) + \left( \sum_{y_i > \theta} +1 \right) \]

Minimizing the Average Absolute Loss

- Take the derivative
- How?
- Derivative at the corner?
- What is the sign of 0?

Convention:

\[ \text{sign}(0) = 0 \]

Minimizing the Average Absolute Loss

- Take the derivative
- How?
- Derivative at the corner?
- What is the sign of 0?

Convention:

\[ \text{sign}(0) = 0 \]

Calculus for Loss Minimization

- General Procedure:
  - Verify that function is convex (we often will assume this...)
  - Compute the derivative
  - Set derivative equal to zero and solve for the parameters
  - Using this procedure we discovered:

The **median** minimizes the absolute loss → Robust! not sensitive to outliers

- Many optimal values
- Pick one?
Minimizing the Average Huber Loss

\[
L_{\alpha}(\theta, y) = \begin{cases} 
\frac{1}{2} (y - \theta)^2 & |y - \theta| < \alpha \\
\alpha (|y - \theta| - \frac{\alpha}{2}) & \text{otherwise}
\end{cases}
\]

➢ Take the derivative of the average Huber Loss

\[
\frac{\partial}{\partial \theta} L_D(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \begin{array}{cc}
(y_i - \theta) & |y_i - \theta| < \alpha \\
\alpha \text{sign}(y_i - \theta) & \text{otherwise}
\end{array} \right.
\]

Set derivative equal to zero:

\[
\left( \sum_{\theta \geq y_i + \alpha} \alpha \right) - \left( \sum_{\theta \leq y_i - \alpha} \alpha \right) - \left( \sum_{|y_i - \theta| < \alpha} (y_i - \theta) \right) = 0
\]

➢ Solution?

➢ No simple analytic solution ...

➢ We can still plot the derivative

Minimizing the Average Huber Loss

\[
L_{\alpha}(\theta, y) = \begin{cases} 
\frac{1}{2} (y - \theta)^2 & |y - \theta| < \alpha \\
\alpha (|y - \theta| - \frac{\alpha}{2}) & \text{otherwise}
\end{cases}
\]

➢ Take the derivative of the average Huber Loss

\[
\frac{\partial}{\partial \theta} L_D(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left\{ - (y_i - \theta) \right\} |y_i - \theta| < \alpha
\]

\[
\left( - \alpha \text{sign}(y_i - \theta) \right) \text{ otherwise}
\]

Visualizing the Derivative of the Huber Loss

\[
L_{\alpha}(\theta, y) = \begin{cases} 
\frac{1}{2} (y - \theta)^2 & |y - \theta| < \alpha \\
\alpha (|y - \theta| - \frac{\alpha}{2}) & \text{otherwise}
\end{cases}
\]

➢ Large $\alpha$ ➔ unique optimum like squared loss

\[
\alpha = 10
\]

\[
\text{Huber Loss} \quad \text{Quadratic}
\]
Minimizing the Huber Loss Numerically

Often we will use numerical optimization methods when using numerical optimization methods:

- **convex loss function**
- **smooth loss function**
- **analytic derivative**

**Summary of Model Estimation**

1. **Define the Model:** simplified representation of the world
   - Use domain knowledge but... keep it simple!
   - Introduce parameters for the unknown quantities

2. **Define the Loss Function:** measures how well a particular instance of the model “fits” the data
   - We introduced \( L_2 \), \( L_1 \), and Huber losses for each record
   - Take the average loss over the entire dataset

3. **Minimize the Loss Function:** find the parameter values that minimize the loss on the data
   - We did this graphically
   - Minimize the loss analytically using calculus
   - Minimize the loss numerically

**Going beyond the simple model**

- **percentage tip = 0**
- **How could we improve upon this model?**
- **Things to consider when improving the model**
  - Related factors to the quantity of interest
    - Examples: quality of service, table size, time of day, total bill
  - Do we have data for these factors?
  - The form of the relationship to the quantity of interest
    - Linear relationships, step functions, etc...
  - Goals for improving the model
    - Improve prediction accuracy
    - More complex models
  - Provide understanding of simpler models
  - Is my model “identifiable” (i.e. possible to estimate the parameters)?
    - percent tip = \( 0 \), \( 0 \) — many identical parameterizations

**Improving the Model**

- **Visualizing the Derivative of the Huber Loss**
  - Alpha = 1
    - Derivative is continuous
    - Small \( \alpha \) → many optima
  - Alpha = 10
    - Large \( \alpha \) → unique optimum
      - like squared loss

**Numerical Optimization**

```python
from scipy.optimize import minimize

def huber_loss_derivative(theta, y, alpha):
    def d = abs(theta - y) if alpha else alpha
    return sp.where(d < alpha, [alpha, (d - alpha) / alpha])

def huber_loss(theta, y):
    return lambda: d

def huber_loss_derivative(theta, y):
    return lambda: d

---

tip = 0.1
alpha = 1
result = minimize(huber_loss, [initial_value], method='BFGS', options={'alpha': alpha})
```

**Minimizing the Huber Loss from scratch**

```python
from scipy.optimize import minimize

def huber_loss(theta, y, alpha):
    def d = abs(theta - y) if alpha else alpha
    return sp.where(d < alpha, [alpha, (d - alpha) / alpha])

def huber_loss_derivative(theta, y, alpha):
    return lambda: d

---

tip = 0.1
alpha = 1
result = minimize(huber_loss, [initial_value], method='BFGS', options={'alpha': alpha})
```
percentage tip = $\theta_1^* + \theta_2^* \times \text{total bill}$

**Rationale:**
Larger bills result in larger tips and people tend to be more careful or stingy on big tips.

**Parameter Interpretation:**
- $\theta_1$: Base tip percentage
- $\theta_2$: Reduction/increase in tip for an increase in total bill.

Often visualization can guide in the model design process.

Estimating the model parameters:
percentage tip = $\theta_1^* + \theta_2^* \times \text{total bill}$

- Write the loss (e.g., average squared loss)

$$L_D(\theta_1, \theta_2) = \frac{1}{n} \sum_{i=1}^{n} (y_i - (\theta_1 + \theta_2 x_i))^2$$

- Take the derivative(s):

$$\frac{\partial}{\partial \theta_1} L_D(\theta_1, \theta_2) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta_1} (y_i - (\theta_1 + \theta_2 x_i))^2$$

$$= -\frac{2}{n} \sum_{i=1}^{n} (y_i - (\theta_1 + \theta_2 x_i))$$

$$\frac{\partial}{\partial \theta_2} L_D(\theta_1, \theta_2) = \frac{1}{n} \sum_{i=1}^{n} (y_i - (\theta_1 + \theta_2 x_i)) x_i$$

- Set derivatives equal to zero and solve for parameters

Solving for $\theta_1$

$$0 = -\frac{2}{n} \sum_{i=1}^{n} (y_i - (\theta_1 + \theta_2 x_i))$$

$$= -\frac{2}{n} \left( \sum_{i=1}^{n} y_i - n \theta_1 - \theta_2 \sum_{i=1}^{n} x_i \right)$$

Rearrange:
$$\sum_{i=1}^{n} y_i = n \theta_1 + \theta_2 \sum_{i=1}^{n} x_i$$
Solving for $\theta_1$

$$\sum_{i=1}^{n} y_i = n\bar{x} + \theta_2 \sum_{i=1}^{n} x_i$$

Divide by $n$

$$\frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y} = \theta_1 + \theta_2 \bar{x}$$

Define the average of $x$ and $y$:

$$\bar{x} := \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\bar{y} := \frac{1}{n} \sum_{i=1}^{n} y_i$$

$$\theta_1 = \bar{y} - \theta_2 \bar{x}$$

Solving for $\theta_2$

$$0 = \frac{2}{n} \left( \left( \sum_{i=1}^{n} y_i x_i \right) - \left( \theta_1 \sum_{i=1}^{n} x_i + \theta_2 \sum_{i=1}^{n} x_i^2 \right) \right)$$

Rearranging Terms

$$\sum_{i=1}^{n} y_i x_i = \theta_1 \sum_{i=1}^{n} x_i + \theta_2 \sum_{i=1}^{n} x_i^2$$

Divide by $n$

$$\frac{1}{n} \sum_{i=1}^{n} y_i x_i = \frac{\theta_1}{n} \sum_{i=1}^{n} x_i + \frac{\theta_2}{n} \sum_{i=1}^{n} x_i^2$$

$$\theta_2 = \frac{xy - \bar{y} \bar{x}}{\bar{x}^2 - \bar{x}^2} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}) (y_i - \bar{y})$$

Solving for $\theta_1$

$$\theta_1 = \bar{y} - \theta_2 \bar{x}$$

System of Linear Equations

Substituting $\theta_1$ and solving for $\theta_2$

$$\bar{y} = (\bar{y} - \theta_2 \bar{x}) \bar{x} + \theta_2 \bar{x}^2$$

$$\bar{x} \bar{y} = \theta_1 \bar{x} + \theta_2 \bar{x}^2$$

$$\bar{y} \bar{x} = \theta_1 \bar{x} + \theta_2 \bar{x}^2$$

$$\theta_2 = \frac{\bar{y} \bar{x} - \bar{x} \bar{y}}{\bar{x}^2 - \bar{x}^2} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}) (y_i - \bar{y})$$

$$\theta_1 = \bar{y} - \theta_2 \bar{x}$$

Completing the squares:

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

$$\sum_{i=1}^{n} (x_i^2 - 2x_i x_1 + x_1^2 - x_1^2) = \sum_{i=1}^{n} (x_i - x_1)^2 - n \bar{x}^2 + \bar{x} \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} (x_i - x_1)^2 = n \bar{x}^2 + \bar{x} \sum_{i=1}^{n} x_i$$

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

Denominator Derivation

Skipped in Class
Completing the squares:
\[ \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \bar{y} - \bar{y} + \bar{y})^2 + \bar{y} - \bar{y} + 2 \bar{y} \overline{\bar{x}} \]

**Derivation**

**Numerator**

\[ \sum_{i=1}^{n} (y_i - \bar{y}) (x_i - \bar{x}) + \bar{y} \overline{\bar{x}} \]

\[ = \sum_{i=1}^{n} (y_i - \bar{y}) (x_i - \bar{x}) + \sum_{i=1}^{n} (y_i - \bar{y} \bar{x}, \bar{y} = 2 \bar{y} \overline{\bar{x}}) \]

**Step 3:**

\[ \text{Numerically (using optimization algorithms)} \]

\[ \text{Analytically (using calculus)} \]

\[ \text{Minimize the loss} \]

\[ \text{(} D_{\theta_1, \theta_2} = \frac{1}{n} \sum_{i=1}^{n} (y_i - (\theta_1 + \theta_2 x_i))^2 \text{)} \]

**Step 3:** Minimize the loss
- **Analytically (using calculus)**
- Numerically (using optimization algorithms)

\[ \frac{\partial}{\partial \theta_1} D_{\theta_1, \theta_2} = -2 \frac{1}{n} \sum_{i=1}^{n} (y_i - (\theta_1 + \theta_2 x_i)) \]

\[ \frac{\partial}{\partial \theta_2} D_{\theta_1, \theta_2} = -2 \frac{1}{n} \sum_{i=1}^{n} (y_i - (\theta_1 + \theta_2 x_i)) x_i \]

**Summary so far ...**

**Step 1:** Define the model with unknown parameters
- Percentage tip = $\theta_1^2 + \theta_2^2 + \text{total bill}$

**Step 2:** Write the loss (we selected an average squared loss)

\[ L_{D_{\theta_1, \theta_2}} = \frac{1}{n} \sum_{i=1}^{n} (y_i - (\theta_1 + \theta_2 x_i))^2 \]

**Step 3:** Minimize the loss
- Analytically (using calculus)
- Numerically (using optimization algorithms)

**Set derivatives equal to zero and solve for parameter values**

\[ \theta_1 = \bar{y} - \bar{x} \]

\[ \theta_2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) (y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \]

**Is this a local minimum?**

\[ \frac{\partial^2}{\partial \theta_1^2} D_{\theta_1, \theta_2} = -2 \frac{1}{n} \sum_{i=1}^{n} (y_i - (\theta_1 + \theta_2 x_i)) \]

\[ \frac{\partial^2}{\partial \theta_2^2} D_{\theta_1, \theta_2} = -2 \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta_2} (y_i - (\theta_1 + \theta_2 x_i)) = -2 \frac{1}{n} \sum_{i=1}^{n} x_i^2 > 0 \]

**Visualizing the Higher Dimensional Loss**

- What does the loss look like?
- Go to notebook ...
percentage tip = $\theta_1^* + \theta_2^* \cdot \text{is Male} + \theta_3^* \cdot \text{is Smoker} + \theta_4^* \cdot \text{table size}$

**Rational:**
Each term encodes a potential factor that could affect the percentage tip.

**Possible Parameter Interpretation:**
- $\theta_1$: base tip percentage paid by female non-smokers without accounting for table size.
- $\theta_2$: tip change associated with male patrons ...
- $\theta_3$: tip change associated with male smokers ...
- $\theta_4$: table size

Maybe difficult to estimate ... what if all smokers are male?

**Define the model**
- Use python to define the function

```python
def f(theta, data):
    return [theta[0] + theta[1] * (data['sex'] == 'Male') +
            theta[2] * (data['smoker'] == 'Yes') +
            theta[3] * data['size']]
```

**Define and Minimize the Loss**

```python
def l2(theta):
    return np.mean([squared_loss(f(theta, data), data['tip']) for data in dataset.values])
```

**Define and Minimize the Loss**

```python
def l1(theta):
    return np.mean([squared_loss(f(theta, data), data['tip']) for data in dataset.values])
```

**Define and Minimize the Loss**

```python
def huber(theta):
    return np.mean([huber_loss(f(theta, data), data['tip']) for data in dataset.values])
```

**Difficult to Plot**

**Rational:**
Each term encodes a potential factor that could affect the percentage tip.

**Possible Parameter Interpretation:**
- $\theta_1$: base tip percentage paid by female non-smokers without accounting for table size.
- $\theta_2$: tip change associated with male patrons ...
- $\theta_3$: tip change associated with male smokers ...
- $\theta_4$: table size

Maybe difficult to estimate ... what if all smokers are male?

**Define the model**
- Use python to define the function

```python
def f(theta, data):
    return [theta[0] + theta[1] * (data['sex'] == 'Male') +
            theta[2] * (data['smoker'] == 'Yes') +
            theta[3] * data['size']]
```

**Define and Minimize the Loss**

```python
def l2(theta):
    return np.mean([squared_loss(f(theta, data), data['tip']) for data in dataset.values])
```

**Define and Minimize the Loss**

```python
def l1(theta):
    return np.mean([squared_loss(f(theta, data), data['tip']) for data in dataset.values])
```

**Define and Minimize the Loss**

```python
def huber(theta):
    return np.mean([huber_loss(f(theta, data), data['tip']) for data in dataset.values])
```