Lecture 7: Regression
Modeling Social Data, Spring 2019
Columbia University

Vatsala Swaroop (vs2671)

March 13, 2019

1 Continuation of Reproducibility, Replication Lecture

The Lecture begins with a continued overview into what we should do -

- Read the literature - It is important to know what work has already been done. This might also help you think and understand about the problem more.
- Formulate your study
- Run a simple pilot - a basic version is run to get initial results. This can result in revisions in the study
- Analyze the results
- Revise your study (null != nil)
- Do a power calculation (think about effect size here - reference to homework 2)
- Pre-register your plans
- Run your study
- Create a reproducible report - important for others to check the work
- Think critically about results
- Disclose everything you did - this also includes details around data collection

Next, Professor Jake gives an example of his own research, findings and interpretation. We see an experiment revolving around the probability to win if an item is picked. Multiple tests are run for significance of difference in willingness to pay between the two items.

In order to collect the data, they used Mechanical Turk - a platform that typically allows people to earn small amounts of money for tasks/surveys completed. It was found by him that the average responses by a random set of people on Turk were not typically different from how someone who knew about the subject would respond.

We also see a brief overview of the data file they get back. Typically, a lot of data is collected and can be used for different purposes -
eg. We revisit the boulder experiment covered two lectures before. Data like when they saw start boulder, timestamps so that they can check for bots and filter them out

The steps he has followed closely match the description of what we should do stated above. They can be briefly summarized as -

Data collection $\rightarrow$ Run pilot $\rightarrow$ See Rmd and finalize result $\rightarrow$ Study to check $\rightarrow$ Analysis and summary of findings

1.1 Brief revisit to standard error and deviation

For a sample size of 10 people, we compute the average value for a quantity

Standard error - variation in the average decreases with more samples

Standard deviation - Measure of uncertainity in the population

$$\sigma = \frac{\sigma_x}{N^{1/2}}$$

The standard deviation graph is relevant for effect size. If means are different but the variation overlaps a lot, then that is not very significant.

2 Regression

Regression analysis basically refers to using the available data to infer relationships from variables and predict future outcomes.

2.1 Goals

The primary goals of regression analysis are to -

- Predict future outcomes
- Explain associations between predictors and outcomes
- Describe or summarize outcomes under different conditions

An example of regression analysis would be SAT scores for Asian vs Hispanic students.

2.2 Prediction problem

We basically want to find a function f whose output matches the data well such that

$$y_i = f(x_i)$$

Here, y is the output/outcome vector represented by

$$\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

And x is the input vector such that there are n training samples of d dimensions/features each

$$\begin{bmatrix} x_{11} & \dots & x_{1d} \\ \vdots & & \\ x_{n1} & & x_{nd} \end{bmatrix}$$

2.3 Loss Function

To arrive at our desired function mapping f(x), we define a loss function. Our goal is to **minimize this loss function**

A reasonable choice of loss function is Least squares error

$$L_i[f] = \frac{1}{N}\sum_{1}^{N}(y_i - f(x_i))^2$$

Another choice is the absolute loss function

$$L_i[f] = \frac{1}{N}\sum_{1}^{N}|y_i - f(x_i)|$$

Here y_i is actual outcome and $f(x_i)$ is the predicted outcome.

2.4 Maximum Likelihood Interpretation

Assumption: We imagine that the data is generated by

$$y_i = f(x_i) + \epsilon_i$$

such that $\epsilon_i \sim N(0, \sigma^2)$
We can then calculate the likelihood of seeing the observed data D as -

$$p(D|f) = \prod_{i=1}^{N} p(\epsilon_i|f) = (\frac{1}{\sqrt{2\pi\sigma^2}})^N \prod_{i=1}^{N} e^{-\frac{1}{2\sigma^2}(y_i - f(x_i))^2}$$

Observe : We have assumed independence here.
What we basically want to find is under what function f is the probability of the observed data maximized.

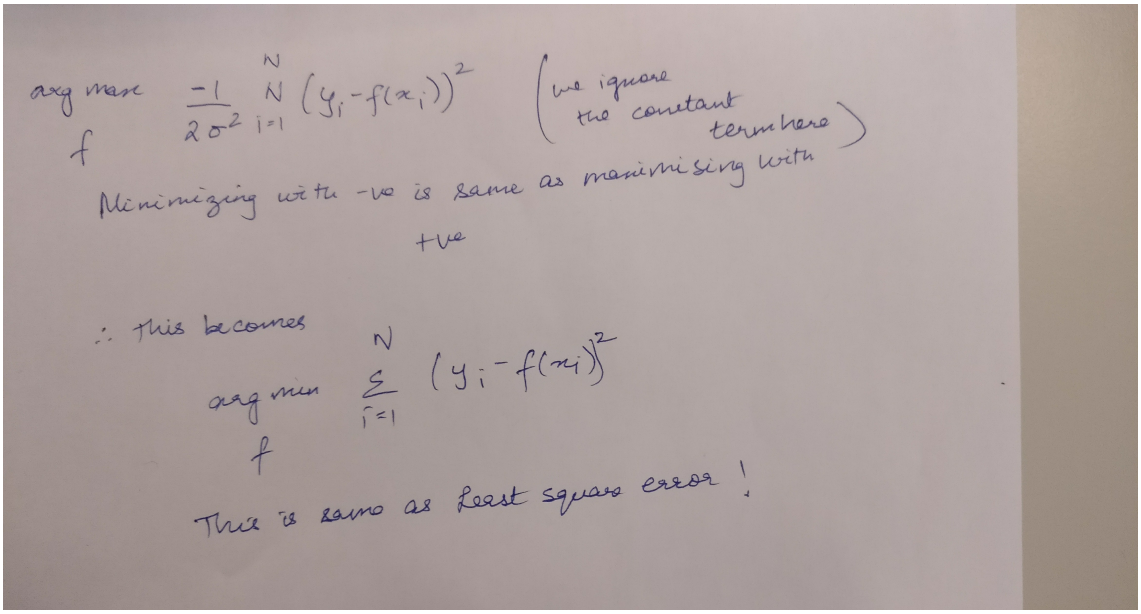
This is given by

$$f^* = \underset{f}{\operatorname{argmax}}\, p(D|f)$$

where f^* Maximum Likelihood Solution.

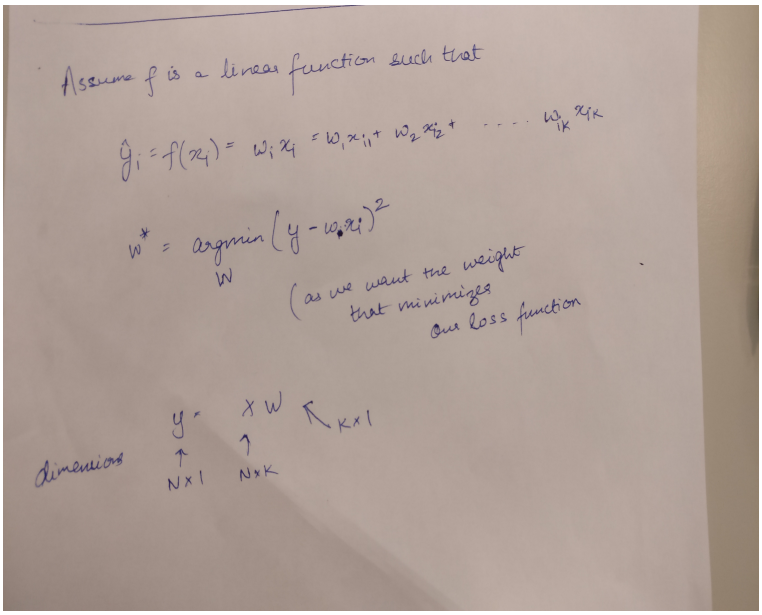
2.5 Maximum Log Likelihood and relation to Least square error

Since it is difficult to calculate in the product form stated above, we take a log to deal with the products! It is okay to do this because of the monotonically increasing nature of log function.
Then log of maximum likelihood becomes -

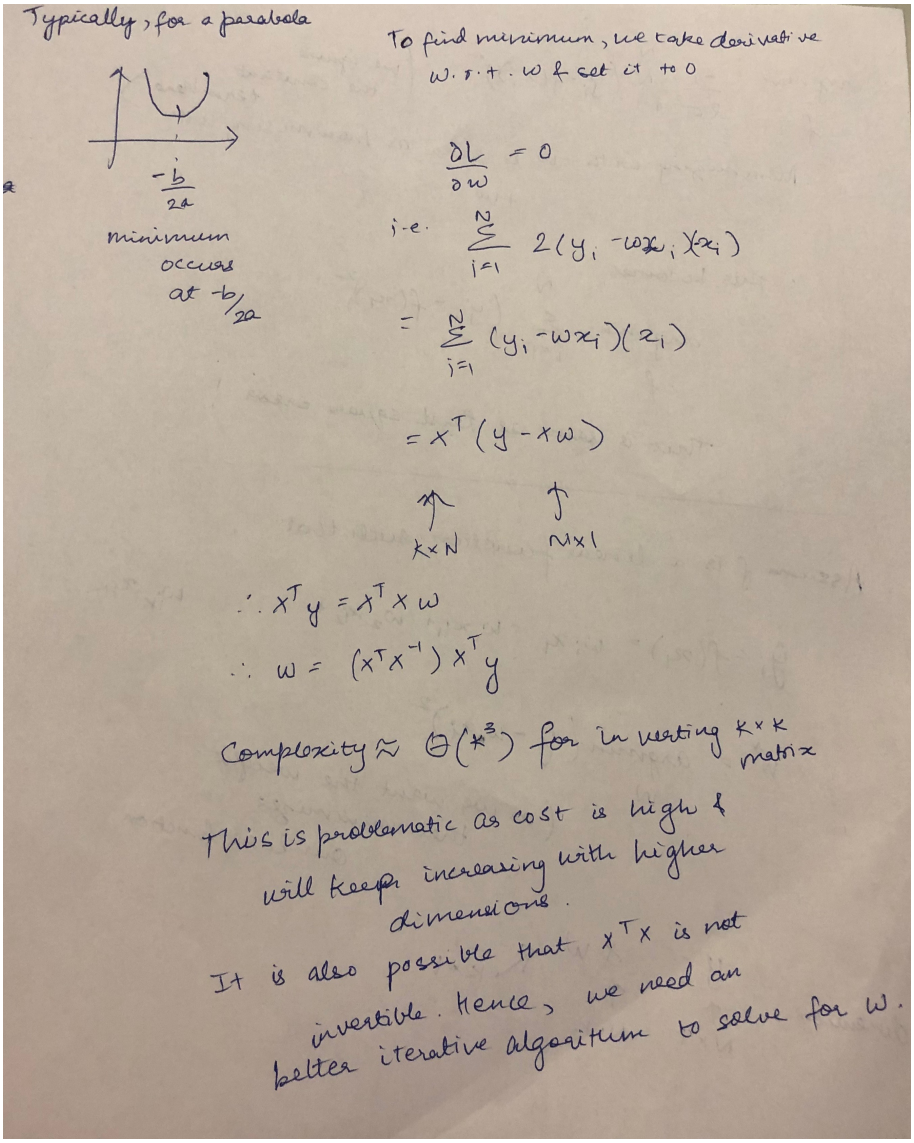


Hence, we can achieve the Maximum Likelihood Solution by solving Least Squared Error.

2.6 Computing Desired weight vector



Continued computation given on next page



2.7 Gradient Descent

We update weights as

$$w := w - \alpha \frac{\partial L}{\partial w}$$

where L is the loss function
which is

$$w := w - \alpha \frac{2}{N} X^T (y - Xw)$$

as the update rule at each step.

where α is called the learning rate, which is the step size we take at each iteration in the direction of negative gradient. This method take $O(kN)$ to compute the update for w at each step.

There are variations to this vanilla form of gradient descent- One such variation is the stochastic gradient descent where we update w using a smaller batch size (or 1 data point)

$$w := w - \alpha(y_i - w \cdot x_i)x_i$$

This method only requires $O(mK)$ where m is the batch size