Netflix Prize

- The Most Famous statistical learning or data mining competition in the world.
- The competition was held by Netflix. On 21 September 2009, the grand prize of US Dollar $1,000,000$ was given to the BellKor's Pragmatic Chaos team which bested Netflix's own algorithm for predicting ratings by 10.06%
- More than ten thousands high quality papers related to this project
- Netflix provided a training data set of 100,480,507 ratings that 480,189 users gave to 17,770 movies. Each training rating is a quadruplet of the form "user, movie, date of grade, grade". The user and movie fields are integer IDs, while grades are from 1 to 5 (integral) stars
  - Training set (99,072,112 ratings not including the probe set, 100,480,507 including the probe set) Probe set (1,408,395 ratings)
  - Qualifying set (2,817,131 ratings) consisting of:
    Test set (1,408,789 ratings), used to determine winners
    Quiz set (1,408,342 ratings), used to calculate leaderboard scores
- `https://en.wikipedia.org/wiki/Netflix_Prize`

Estimating Estimable Functions of β: An Example

customer	Movie		
	1	2	3
1	4	1	?
2	?	3	5
3	?	?	3
4	3	1	?

- Can we guess ratings for customer/movie combinations not in the dataset?
- Which movie is the best?
- Statistical model: y_{ij} the customer i's rating on movie j

$$y_{ij} = \mu + c_i + m_j + \epsilon_{ij}$$

$$\begin{bmatrix} 4 \\ 1 \\ 3 \\ 5 \\ 3 \\ 3 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \mu \\ c_1 \\ c_2 \\ c_3 \\ c_4 \\ m_1 \\ m_2 \\ m_3 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{33} \\ \epsilon_{41} \\ \epsilon_{42} \end{bmatrix}$$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Can We Estimate the Means that Underly the Missing Table Entries?

$$\boldsymbol{X}\boldsymbol{\beta} = \begin{bmatrix} \mu + c_1 + m_1 \\ \mu + c_1 + m_2 \\ \mu + c_2 + m_2 \\ \mu + c_2 + m_3 \\ \mu + c_3 + m_3 \\ \mu + c_4 + m_1 \\ \mu + c_4 + m_2 \end{bmatrix}$$

Can we estimate the missing values

$$\mu + c_1 + m_3, \mu + c_2 + m_1, \mu + c_3 + m_1, \mu + c_3 + m_2, \mu + c_4 + m_3?$$

Note: $m_1 - m_2$ is estimable because

$$[1, -1, 0, 0, 0, 0, 0]\boldsymbol{X}\boldsymbol{\beta} = m_1 - m_2$$

(What does $m_1 - m_2$ mean here?)

We can estimate the mean underlying the rating for any combination of customer and movie

- Likewise, $m_2 - m_3$ is estimable because

$$[0, 0, 1, -1, 0, 0, 0]\boldsymbol{X}\boldsymbol{\beta} = m_2 - m_3$$

  Thus, we can also estimate $m_1 - m_3$ (How?)

- It follows that any linear combination of the form $\mu + c_i + m_j$ can be estimated for any $i = 1, 2, 3, 4$ and $j = 1, 2, 3$ because

$$\mu + c_i + m_j = (\mu + c_i + m_j) + (m_j - m_{j'})$$

Movie lsmeans

- If our goal is to compare movies to see which is most highly rated, we can accomplish that by estimating the pairwise differences between movie effects.
- However, if we want to retain information about the mean rating rather than the difference between mean ratings, it is natural to consider estimating the average (across *all* customers) of the mean rating for each movie.
- This average for the jth movie is

$$\frac{1}{4}\sum_{i=1}^{4}(\mu + c_i + m_j) = \mu + \bar{c}_{\cdot} + m_j$$

This average is estimable for each movie in our example because it is a linear combination of estimable functions.

Suppose we consider a different model

customer	Movie		
	1	2	3
1	4	1	?
2	?	3	5
3	?	?	3
4	3	1	?

- Can we guess ratings for customer/movie combinations not in the dataset?
- Which movie is the best?
- Statistical model: y_{ij} the customer i's rating on movie j

$$y_{ij} = \mu_{ij} + \epsilon_{ij}$$

$$\begin{bmatrix} 4 \\ 1 \\ 3 \\ 5 \\ 3 \\ 3 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{13} \\ \mu_{21} \\ \mu_{22} \\ \mu_{23} \\ \mu_{31} \\ \mu_{32} \\ \mu_{33} \\ \mu_{41} \\ \mu_{42} \\ \mu_{43} \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{33} \\ \epsilon_{41} \\ \epsilon_{42} \end{bmatrix}$$

Can we estimate the means that underly the missing table entries? No.

$$\boldsymbol{X}\boldsymbol{\beta} = \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{22} \\ \mu_{23} \\ \mu_{33} \\ \mu_{41} \\ \mu_{42} \end{bmatrix}$$

- Can we estimate missing ones $\mu_{13}, \mu_{21}, \mu_{31}, \mu_{32}, \mu_{43}$?
- None of the means underlying missing table entries are estimable under this cell-means model.