MovieLens Project

bart-g

4/8/2020

Contents

1. Overview

The goal of the project is to the predict movie rating given features of the data set provided.

Each record of the data set corresponds to a historical rating of a movie made by a user.

Projection goal of movie rating:

$$y_{u,i}$$

depicts rating projection for user (u) on a movie (i).

Two sets of these records, as generated within prerequisite code section, are made available:

- edx for training purpose (9M rows)
- validation for RMSE score testing (1M rows)

Features available:

- userId (integer identifier)
- movieId (numerical)
- rating (numerical)
- timestamp (integer)
- title (character string)
- genres (character string)

Snapshot of the data:

	userId	movieId	rating	timestamp	title	genres
1	1	122	5	838985046	Boomerang (1992)	Comedy|Romance
2	1	185	5	838983525	Net, The (1995)	Action|Crime|Thriller
4	1	292	5	838983421	Outbreak (1995)	Action|Drama|Sci-Fi|Thriller
5	1	316	5	838983392	Stargate (1994)	Action|Adventure|Sci-Fi
6	1	329	5	838983392	Star Trek: Generations (1994)	Action|Adventure|Drama|Sci-Fi
7	1	355	5	838984474	Flintstones, The (1994)	Children|Comedy|Fantasy

Unique users & movies in edx training set:

Unique users	Movies being rated
69878	10677

2. Analysis

Original training set(edx) was partitioned into training set, and a test set constituting 10% of the entire edx set. Additionally due to required test set consistency with users and movies of training set, necessary records were relocated to the test set. Given training set we approach the problem by handpicking selected features, an regressing ratings onto them. Due to practical considerations direct linear regression cannot be used:

- while model is augmented residuals must be stripped of the impact made (with previously identified features)

Worth noting that if the vast amount of data had been applied in least squares regression, one could argue the slope of each variable used in the modelling could be more accurately optimized.

On the downside: strick linear fit to training data could lead to overtraining when out-of-sample/test data validated the model. Such overfitting would be arguably greater, than our "modest" effect-selective model, due to growing number of degrees of freedom).

Error function

RMSE is defined as a mean square error of actual and predicted vectors.

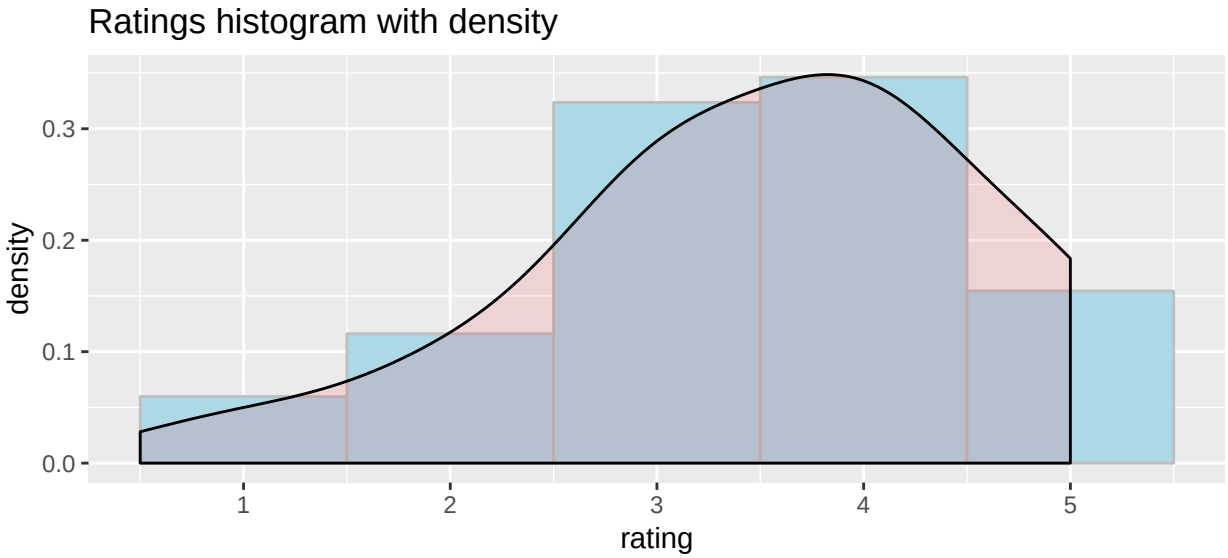
$$RMSE = \sqrt{\frac{1}{n}\sum_{u,i}(y_{u,i} - \hat{y}_{u,i})^2}$$

Edx training & test sets preparation

We start by partitioning edx data into training set and test set. Learning of the model is conducted entirely with edx data.

1st model: intercept

Histogram of the ratings directs the attention towards the mean value. Mean rating, the intercept, offers an introductory explanatory power. Had the dependent variable been centered around zero, we could argue how useful it is for building the model. Profoundly intercept mean is greater than zero, so is the mass of the distribution.

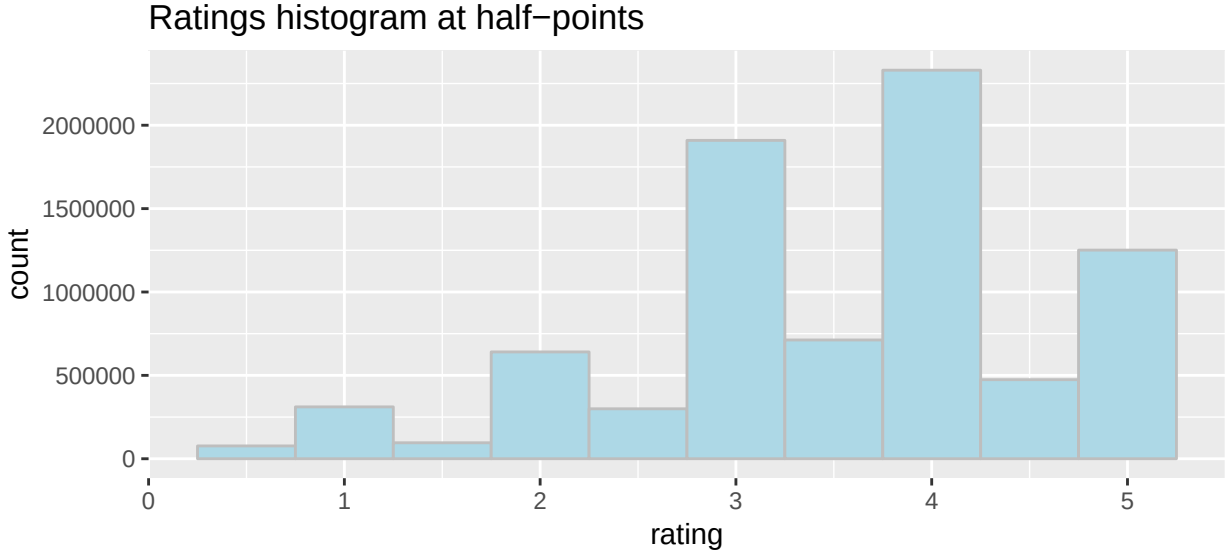


There are two things worth noting:

- distribution exposes negative skewness,
- mass of higher than average ratings is concentrated in the right tail

Indication of abnormality in terms of excess kurtosis is not significant.
Both moments however, are duly noted:

skewnes	kurtosis
-0.5958878	0.008005



Another finding from half-points histogram:

- half-point ratings aren't as popular as whole point ratings

Proposed model:

$$y_{u,i} = \mu + \epsilon_{u,i}$$

The intercept estimated from the arithmetic mean is:

Intercept
3.512457

RMSE misses the edx actual ratings by more than entire rating point:

Model RMSE
1.060054

2nd model: movie effect

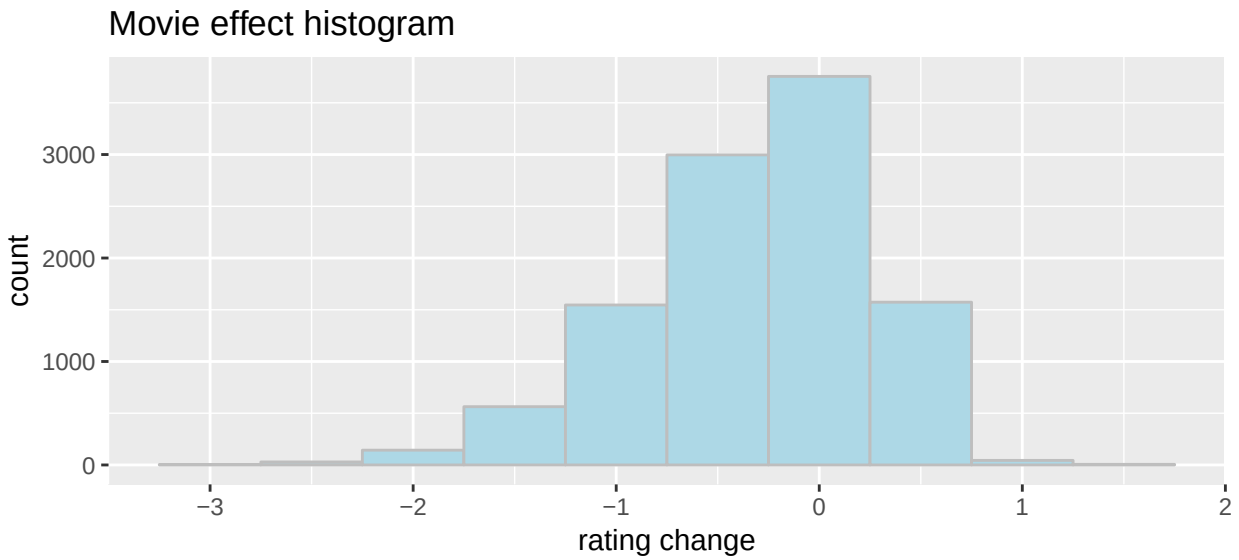
Movies may have inherent bias acquiring ratings particular way.

Hence the attempt to group ratings per movie, called movie effect b_i

Proposed model:

$$y_{u,i} = \mu + b_i + \epsilon_{u,i}$$

It's interesting to see that more movies are ranked below the average, with small percentile of top 5-stars scorers:



RMSE of ratings, once movie effect is incorporated, improves the prediction:

Model RMSE
0.9429615

3rd model: movie + user effect

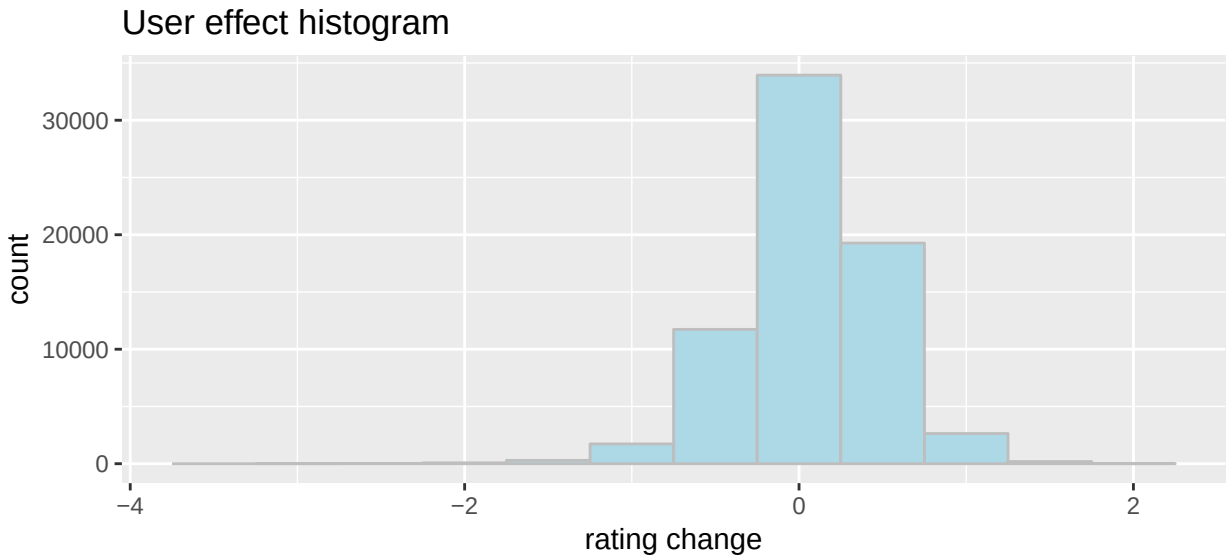
Individuals may differ in their preferences and characteristics of rating process.

Hence the attempt to group their ratings per user, called user effect b_u

Proposed model:

$$y_{u,i} = \mu + b_u + b_i + \epsilon_{u,i}$$

Usually there is more positive critisism among users:



RMSE of ratings, when user and movie effects are incorporated:

Model RMSE
0.8646844

4th model: user + movie + genres effect

Genres can be subjected to preference in rating, some enjoying greater biases,

than others. Such as comedies which typically attract more critisism than dramas.

	Action	Drama	Comedy	Crime	Sci-Fi	Thriller	Romance
total	2560545	3910127	3540930	1327715	1341183	2325899	1712100
avg	3.421405	3.673131	3.436908	3.665925	3.395743	3.507676	3.553813

The genres effect is depicted as b_g which is the effect of user's rating of a movie classified by genres variable: $g_{u,i}$.

Technically, b_g is the captured intercept within each group of genres, calculated (as for other effects) using arithmetic mean in each group.

In order to simplify group filtering an indicator function is introduced:

$$1_A(\omega) = 1$$

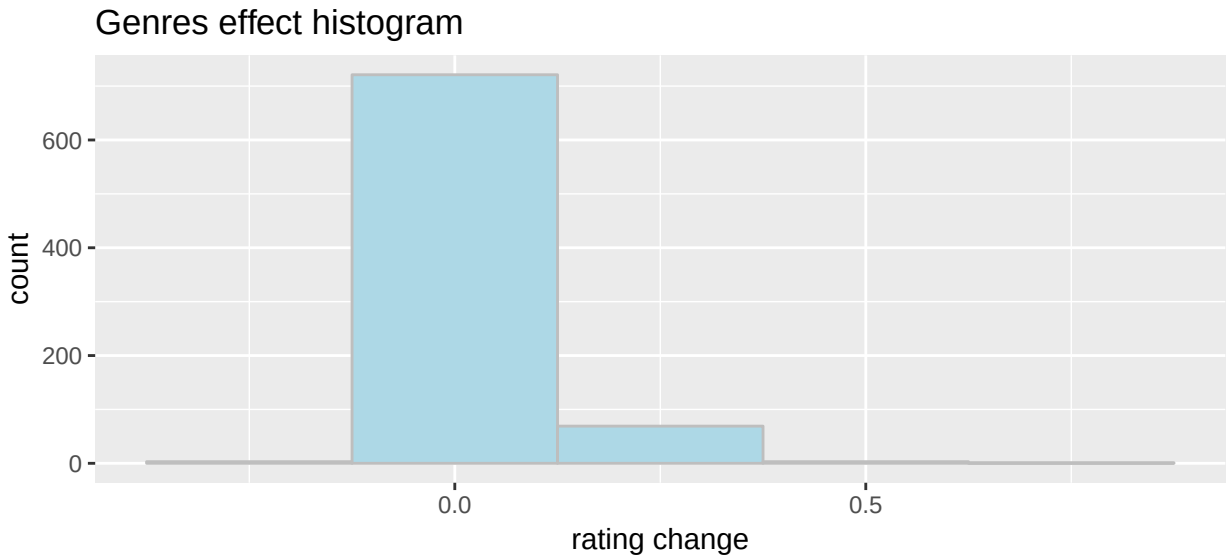
whenever $\omega \in A$, $1_A = 0$ otherwise.

Consequently, matching $g_{u,i}$ genres with selected $b_g(k)$ effect, simplifies model formulation.

Proposed model becomes:

$$y_{u,i} = \mu + b_u + b_i + 1_{g_{u,i}=m} b_g(m) + \epsilon_{u,i}$$

Notably genres effect is hard to discern on distribution plot, despite bin's width being reduced to 1/4th of the rating point.



RMSE of ratings, once genres effect is incorporated, improves the prediction by a tiny fraction (4th decimal place vs previous model):

Model RMSE
0.8643242

Regularisation of the best model

Here, we consider regularisation formula being applied, using our candidate model:

- intercept + movie + user + genres effect (4th model)

Unlike linear regression as a function of lambda, we minimize the formula below (based on model 4), containing the key penalty term:

$$F(\lambda) = \frac{1}{N} \sum_{u,i} (y_{u,i} - \mu - b_u - b_i - 1_{g_{u,i}=m} b_g(m))^2 + \lambda (\sum_u b_u^2 + \sum_i b_i^2 + \sum_{u,i} (1_{g_{u,i}=m} b_g(m))^2)$$

Initially for $\lambda = 0$ we expect no regularisation impact on RMSE. The shrinkage penalty grows with $\lambda -> \text{inf}$, so finding optimised λ estimate strikes the balance between over-trained model using training data (first part of) and reducing the impact of various effects.

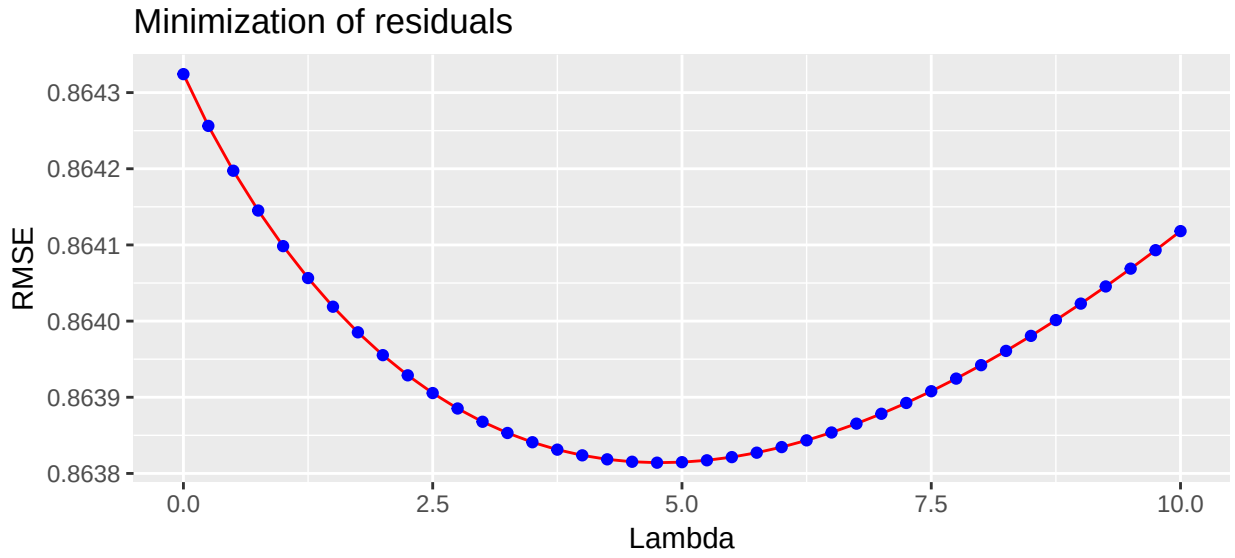
Once $F(\lambda)$ is differentiated with respect to each of the effect, regularisation routine can be put together. Here, the user effect penalized by shrinkage lamba:

$$\frac{\delta F}{b_i} = b_i(\lambda) = \frac{1}{\lambda + n_i} \sum_{u}^{n_i} (y_{u,i} - \mu)$$

Differentiation vs all effects was implemented in the same fashion as $\frac{\delta F}{b_i}$ by extending residuals term on the right hand side of the formula.

Regularisation process was conducted on the $\lambda \in [0, 10]$ at 0.25 increments.

More granular values of λ produced no improvement of residuals and were discarded.



From the curve we obtain the optimal lambda of the regularized model:

Regularisation Lambda
4.75

It yields an improved scored of RMSE, be aware that we're looking at the model built on training data:

RMSE
0.8638141

3. Results

At this point we aggregated user, movie, and generes effect, trained using edx data set.

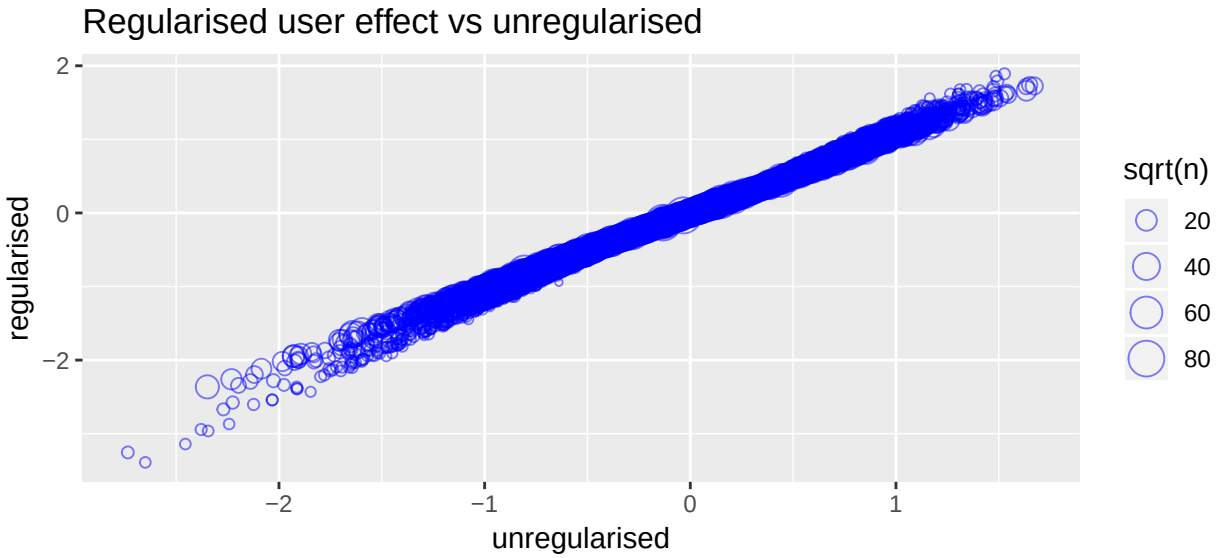
Optimal shrinkage penalty λ can be included in our final computation:

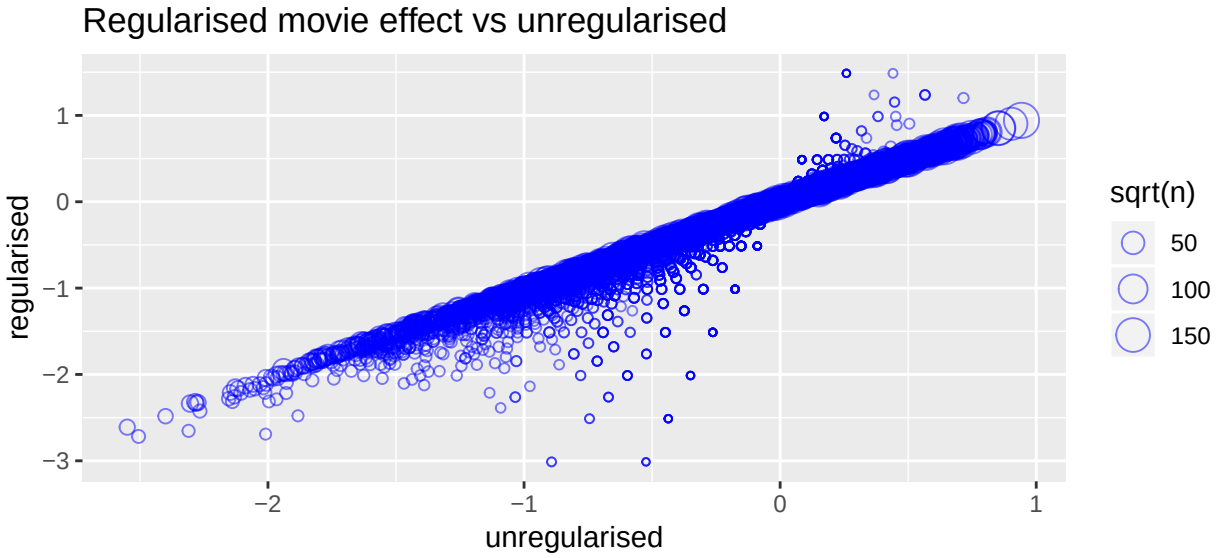
Lambda
4.75

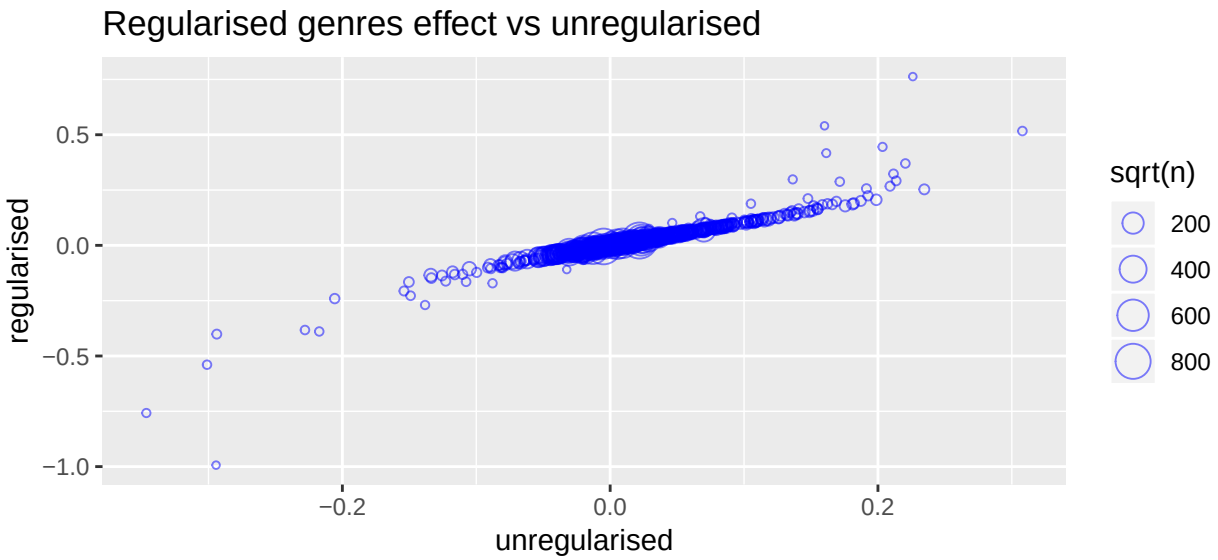
Validation set calculation result:

RMSE
0.8644514

Let's look at the comparison of regularised effects collected.







One can notice that shrinkage affects the explanatory variables:

- the least rated movies,
- users with least ratings,
- genres with fewer movie ratings (but more evident concentration of popular genres)

Models comparison

Given past results from training set (edx data) and the final model performance (validation data) it's possible to compare the impact of selected effect variables.

Model Type	RMSE	Improvement
intercept only	1.0600537	1.0600537
movie effect	0.9429615	0.1170922
movie + user effect	0.8646844	0.0782771
movie + user + genres effect	0.8643242	0.0003602
regularized 'movie + user + genres effect'	0.8638141	0.0005101
regularized 'movie + user + genres effect' (validation set)	0.8644514	-0.0006373

Looking at rating improvement one notices that adding movie effect in general provided the best predicting power once the mean intercept effect is discounted.

If one predicts rating of an out-of-sample movie, bidding solely on average rating (the intercept model) is the strongest feature of all. Yet, it's an error hardly acceptable, residing within an entire rating point.

Regularisation of the full model (intercept + movie + user + genres effect), had greater fractional impact than genres effect for instance. The latter, genres effect, was the least important factor.

Finally once out-of-sample, the validation set, data was employed, the drop in predicting accuracy was less than 1/1000-th only.

4. Conclusion

Modicum of intuition produced a number of 'effect' variables with minimum statistical apparatus. The objective of modeling movie ratings, with ultimate minimisation of the RMSE proved a viable statistical exercise.

Limitation of available CPU power available, played a major role during the study. However, thanks to multi-threaded math libraries provided by Microsoft R Open the research took half of time, or even less compared with initial effort using single-threaded framework.

Model-wise, several other features, or additional instrumental variables could have been investigated. Such as:

- time&calendar impact,
- cross-effects of existing features, or
- tail-anomalies for users and movies

Cross-validation technique wasn't employed for numerical simplicity reasons, but would be a welcome continuation of the study. Same applies to residuals analysis of the unregularised model with emphasis on heteroscedasticity and serial correlation, with latent effects in mind.

Going beyond linear modelling, we could achieve greater accuracy with an ensemble of different techniques, still being adversely limited due to computation costs.

Summarizing, a very basic approach led to a promising entry model for movie rating.