Homework 2

Xiaonan Hu
Liqi Zhu

Part 2: Data Exploration and Sampling Bias

a.

Correlations	rocky1	rocky2	rocky3	rocky4	rocky5
rocky1	1.0000000	0.6007483	0.5163356	0.4755442	0.3855623
rocky2	0.6007483	1.0000000	0.7009254	0.6441270	0.5315010
rocky3	0.5163356	**0.7009254**	1.0000000	0.6981126	0.6058606
rocky4	0.4755442	0.6441270	0.6981126	1.0000000	0.6195678
rocky5	**0.3855623**	0.5315010	0.6058606	0.6195678	1.0000000

Based on the correlations of ratings, we can see that **Rocky2 and Rocky3** are most similar, and **Rocky5** is the most different one from the others.

b.

Througout the dataset, mean rating of **Rocky1** is the best, and mean rating of **Rocky5** is the worst.

c.

Movie	rocky1	rocky2	rocky3	rocky4	rocky5
all_mean	**3.771838**	3.660645	3.59801	3.395633	**3.033764**
rocky4_subset_mean	**4.134366**	3.877048	3.85813	3.395633	**3.136088**

The mean ratings of all movies are **higher** from consumers who rated Rocky4, than those from all ratings. That means consumers who rated Rocky4 tend to rate higher for the others, because this sample method may imply consumers preference.

d.

Missing ratings may because of consumers' non-response or absence of watching.
Omitting incomplete data may cause several problems:
1. It will reduce the size of the dataset and make further models more variance and less precise;
2. As missing data may caused by absence of watching, that implies consumers' preference, so that simply omitting will cause bias.

Part 3: Explanatory Models

c.

Model	AIC	BIC
lm(rocky5 ~ firstInteractions)	11006.86	11044.57
lm(rocky5 ~ secondInteractions)	10946.57	**11021.98**
lm(rocky5 ~ thirdInteractions)	10934.50	11035.06
lm(rocky5 ~ fourthInteractions)	**10929.51**	11036.34

Discussion:

h.

According to AIC, linear regression on **fourthinteractions** is the best model. Accoeding to BIC, lineat regression on **secondinteractions** is the best model. The BIC model is smaller than the AIC model because BIC panelizes models more strictly on the size of a model, the multipier of the number of parameters in BIC is log(nRows) greater than that in AIC which is 2.

i.

The model with the penalty parameter s=.5 is easier to interpret. Because larger panelty parameter will reduce the coefficients in the lasso regression, which is prettier with fewer independent variables, so that fewer relations need to be interpreted.

j.

According to the lasso regression where the penalty parameter is set to .5, **Rocky3 and Rocky4** are most important in predicting Rocky5. And the **interactions** are more important.

k.

Coefficients	OLS	Lasso	Ridge
(Intercept)	1.90938502	0.5645371164	1.2136780089
rocky1	-0.52188369	0.0429589563	-0.1067536758
rocky2	-0.67410327	0.3765732778	0.0719900275
rocky3	-0.05092965	0.1650684449	0.1199391446
rocky4	-0.77936057	0.1249643465	0.1294720613
rocky1:rocky2	0.22031033	-0.0825150711	-0.0121432811
rocky1:rocky3	0.16909344	.	0.0029894188
rocky1:rocky4	0.26899986	0.0102817830	0.0082412850
rocky2:rocky3	0.27582020	-0.0002782899	0.0096437888
rocky2:rocky4	0.53780504	0.0197389913	0.0161255775
rocky3:rocky4	0.14059109	.	0.0172902763
rocky1:rocky2:rocky3	-0.09139339	.	-0.0001226338
rocky1:rocky2:rocky4	-0.12849751	.	0.0010730461
rocky1:rocky3:rocky4	-0.05444169	0.0038956091	0.0022077215
rocky2:rocky3:rocky4	-0.11746679	.	0.0027222861
rocky1:rocky2:rocky3:rocky4	0.03396713	0.0016494383	0.0003536575

Compared to OLS regression on fourthinteractions, **Lasso regression has less coefficiects, and Ridge regression is smaller in the magnitude of the estimated coefficients**. Because Lasso regression includes a penalty term based on the absolute value of the coefficients, many relatively insignificant coefficients are forced to be 0. Although there is also a penalty term in Ridge regression, it is based on the sum of squares of the coefficients, so that coefficients small in magnitude will be kept because their squares will be very small.

l.

If the lasso regression is estimated with a penalty parameter of 0, then the estimated coefficients will be the same as OLS regression. Lasso regression finds coefficients minimize $(y-XB)^2 + \lambda sum|B|$, where $|B|$ is the sum of the absolute value of the coefficients, while OLS minimizes $(y-XB)^2$, so, when $\lambda = 0$, they do exactly the same work.

Part 4: Predictive Modelling

C.

MSE	LM	MARS	Nnet	knn
1	0.9850555	0.9748798	0.9726182	1.114994
2	0.9229779	0.9230449	0.9187437	1.073585
3	0.9430657	0.9377928	0.9345890	1.092402
4	0.9392933	0.9289758	0.9357366	1.044224
5	1.0064112	0.9946266	0.9945080	1.132360
Mean	0.9593607	0.9518640	**0.9512391**	**1.091513**

Neural networks model is the strongest, and k-nearest neighbour is the weakest.