Sebastián Pastor Ferrari

Data Mining – Assignment 02

02.21.2020

Question 01:

i. **Which predictors are likely to be measuring the same thing among the entire set of predictors? Discuss the relationships among INDUS, NOX, and TAX.**
RAD-TAX which have a .91 correlation coefficient. INDUS-NOX have a high coefficient (0.763), as well as INUDS-TAX (0.721), as well as NOX-TAX (0.668).

ii. **Compute the correlation table for the numerical predictors and search for highly correlated pairs. These have potential redundancy and can cause multicollinearity. Choose which ones to remove based on this table.**

Multivariate

Correlations

	CRIM	ZN	INDUS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	LSTAT	MEDV
CRIM	1.0000	-0.2005	0.4066	0.4210	-0.2192	0.3527	-0.3797	0.6255	0.5828	0.2899	0.4556	-0.3883
ZN	-0.2005	1.0000	-0.5338	-0.5166	0.3120	-0.5695	0.6644	-0.3119	-0.3146	-0.3917	-0.4130	0.3604
INDUS	0.4066	-0.5338	1.0000	0.7637	-0.3917	0.6448	-0.7080	0.5951	0.7208	0.3832	0.6038	-0.4837
NOX	0.4210	-0.5166	0.7637	1.0000	-0.3022	0.7315	-0.7692	0.6114	0.6680	0.1889	0.5909	-0.4273
RM	-0.2192	0.3120	-0.3917	-0.3022	1.0000	-0.2403	0.2052	-0.2098	-0.2920	-0.3555	-0.6138	0.6954
AGE	0.3527	-0.5695	0.6448	0.7315	-0.2403	1.0000	-0.7479	0.4560	0.5065	0.2615	0.6023	-0.3770
DIS	-0.3797	0.6644	-0.7080	-0.7692	0.2052	-0.7479	1.0000	-0.4946	-0.5344	-0.2325	-0.4970	0.2499
RAD	0.6255	-0.3119	0.5951	0.6114	-0.2098	0.4560	-0.4946	1.0000	0.9102	0.4647	0.4887	-0.3816
TAX	0.5828	-0.3146	0.7208	0.6680	-0.2920	0.5065	-0.5344	0.9102	1.0000	0.4609	0.5440	-0.4685
PTRATIO	0.2899	-0.3917	0.3832	0.1889	-0.3555	0.2615	-0.2325	0.4647	0.4609	1.0000	0.3740	-0.5078
LSTAT	0.4556	-0.4130	0.6038	0.5909	-0.6138	0.6023	-0.4970	0.4887	0.5440	0.3740	1.0000	-0.7377
MEDV	-0.3883	0.3604	-0.4837	-0.4273	0.6954	-0.3770	0.2499	-0.3816	-0.4685	-0.5078	-0.7377	1.0000

The correlations are estimated by Row-wise method.

I will remove TAX and NOX, given they are highly correlated with each other and with other predictors.

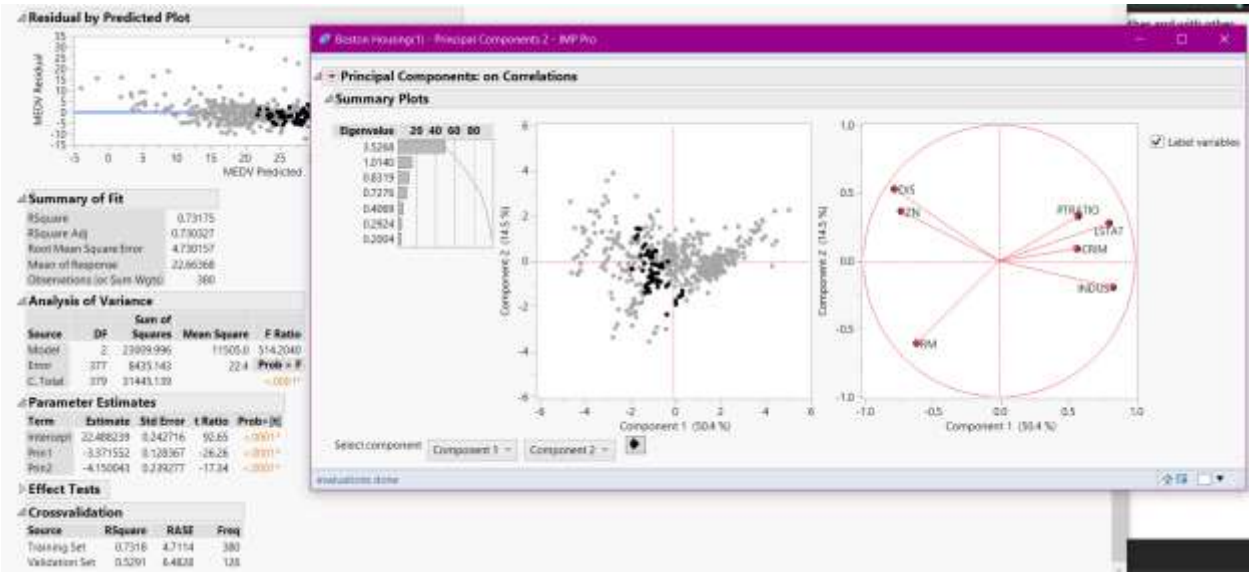
iii. **Use an exhaustive search (All Possible Models) to reduce the remaining predictors as follows: First, choose the top three models. Then run each of these models and compare their predictive accuracy for the validation set. Compare RMSE, Cp, $AICc$, and Validation RSquare. Finally, describe the best model.**

All Possible Models

Model	Number	RSquare	RMSE	AICc	BIC
PTRATIO,LSTAT	2	0.6322	5.5384	2384.38	2400.04
RM,PTRATIO,LSTAT	3	0.7189	4.8482	2284.28	2303.82
RM,RAD,LSTAT	3	0.6805	5.1694	2333.03	2352.57
CRIM,RM,LSTAT	3	0.6789	5.1820	2334.88	2354.42
CRIM,RM,PTRATIO,LSTAT	4	0.7254	4.7984	2277.49	2300.90
RM,DIS,PTRATIO,LSTAT	4	0.7247	4.8045	2278.45	2301.86
RM,RAD,PTRATIO,LSTAT	4	0.7204	4.8418	2284.32	2307.74
CRIM,RM,DIS,PTRATIO,LSTAT	5	0.7339	4.7303	2267.68	2294.96
INDUS,RM,DIS,PTRATIO,LSTAT	5	0.7332	4.7358	2268.57	2295.85
RM,DIS,RAD,PTRATIO,LSTAT	5	0.7288	4.7756	2274.91	2302.19
CRIM,INDUS,RM,DIS,PTRATIO,LSTAT	6	0.7413	4.6698	2258.97	2290.10
CRIM,ZN,RM,DIS,PTRATIO,LSTAT	6	0.7370	4.7087	2265.26	2296.40
CRIM,RM,AGE,DIS,PTRATIO,LSTAT	6	0.7361	4.7169	2266.59	2297.72
CRIM,ZN,INDUS,RM,DIS,PTRATIO,LSTAT	7	0.7445	4.6471	2256.34	2291.32
CRIM,INDUS,RM,AGE,DIS,PTRATIO,LSTAT	7	0.7429	4.6620	2258.77	2293.74
CRIM,INDUS,RM,DIS,RAD,PTRATIO,LSTAT	7	0.7413	4.6759	2261.04	2296.01
CRIM,ZN,INDUS,RM,AGE,DIS,PTRATIO,LSTAT	8	0.7456	4.6432	2256.79	2295.60
CRIM,ZN,INDUS,RM,DIS,RAD,PTRATIO,LSTAT	8	0.7445	4.6533	2258.44	2297.25
CRIM,INDUS,RM,AGE,DIS,RAD,PTRATIO,LSTAT	8	0.7429	4.6681	2260.85	2299.66
CRIM,ZN,INDUS,RM,AGE,DIS,RAD,PTRATIO,LSTAT	9	0.7456	4.6494	2258.90	2301.53

The best model I found was: MEDV = CRIM,ZN,INDUS,RM,DIS,PTRATIO,LSTAT, it has a predictive value (Adj. R2) = 0.7445.

iv. Run a PCA and use only a few principal components in a regression model to predict MEDV. Compare this model with the one you obtained in part iii.



After running a PCA with 7 predictors (CRIM,ZN,INDUS,RM,DIS,PTRATIO,LSTAT), I saved 2 components as columns and then ran a Principal Components Regression Model (PCR) to predict MEDV.

My new $R^2 = 0.73$, slightly less than the 0.7445 that I obtained from the original Stepwise Regression.

Question 02:

a. Make sure the variables are coded correctly in JMP (Nominal, Ordinal, or Continuous), then use the Columns Viewer to summarize the data. Are there any missing values? How many Nominal columns are there?

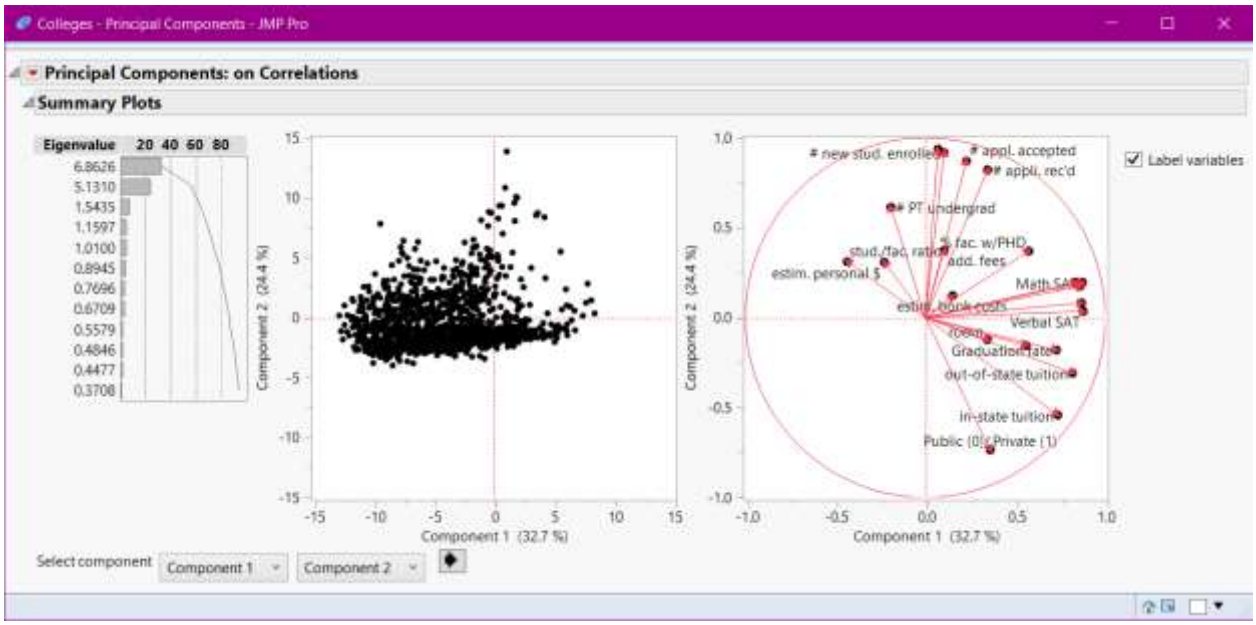
Columns	N	N Missing
Math SAT	777	525
Verbal SAT	777	525
ACT	714	588
# appli. rec'd	1292	10
# appl. accepted	1291	11
# new stud. enrolled	1297	5
% new stud. from top 10%	1067	235
% new stud. from top 25%	1100	202
# FT undergrad	1299	3
# PT undergrad	1270	32
in-state tuition	1272	30
out-of-state tuition	1282	20
room	981	321
board	804	498
add. fees	1028	274
estim. book costs	1254	48
estim. personal $	1121	181
% fac. w/PHD	1270	32
stud./fac. ratio	1300	2
Graduation rate	1204	98

-Public/Private must be coded to ordinal, I also recoded it into {Public:0,Private:1}.

- There is a ton of missing data, notably in standardized testing. This is curious because it alludes to the trend of making standardized testing optional by American colleges.

- There are 20 nominal columns.

b. Conduct a principal components analysis on the data and comment on the results. Recall that, by default, JMP will conduct the analysis on correlations rather than covariances. Is this necessary? Do the data need to be normalized in this case? Discuss key considerations in this decision.



Things such as standardized testing scores max out at certain values, and they are being weighted in comparison with things like costs, out/in state tuition rates, personal $, and % features (scaled 0-100). This is not an accurate PCA because all the features have entirely different scales, which are not accounted in their weights by JMP. The issue with different numerical scales, and different measures means that it is probably best to normalize the data to a certain scale and run the PCA based on their covariances.