ISyE 7406: Data Mining and Statistical Learning **(Due: 01/30/2020)**

Homework Assignment #2

Instructor: Yajun, Mei *Name:* Chen-Yang(Jim), Liu *GTID:* 90345****

Problem 1: R exercise for linear Regression (50 points)

(a) Introduction

The dataset in this problem is from faraway library in R. This dataset contains 252 observations with 18 features. We will use this dataset to perform linear regression. The methods include linear regression with all predictors and 5 best variables, linear regression based on AIC, Lasso regression, Ridge Regression, Principal component regression and partial least squares. Before analysis is performed, the columns of this datasets are:

- brozek: Percent body fat using Brozek's equation
- siri: Percent body fat using Siri's equation
- age: Age
- weight: Weight
- height: Height
- adipos: Adiposity index = $Weight/Height^2$
- free: Fat Free Weight
- chest: Chest circumference
- abdom: Abdomen circumference
- hip: Hip circumference
- thigh: Thigh circumference
- knee: Knee circumference
- ankle: Ankle circumference
- biceps: Extended biceps circumference
- forearm: Forearm circumference

- wrist: Wrist circumference

Now we would like to train our linear regression models based on above features. In this dataset, brozek is assigned as a response variable and other columns are seen as exploratory variables. The dataset will be split into testing examples and training examples. We will keep 30% data as our testing examples and 70% data as our training examples which we will use to build our linear regression models.

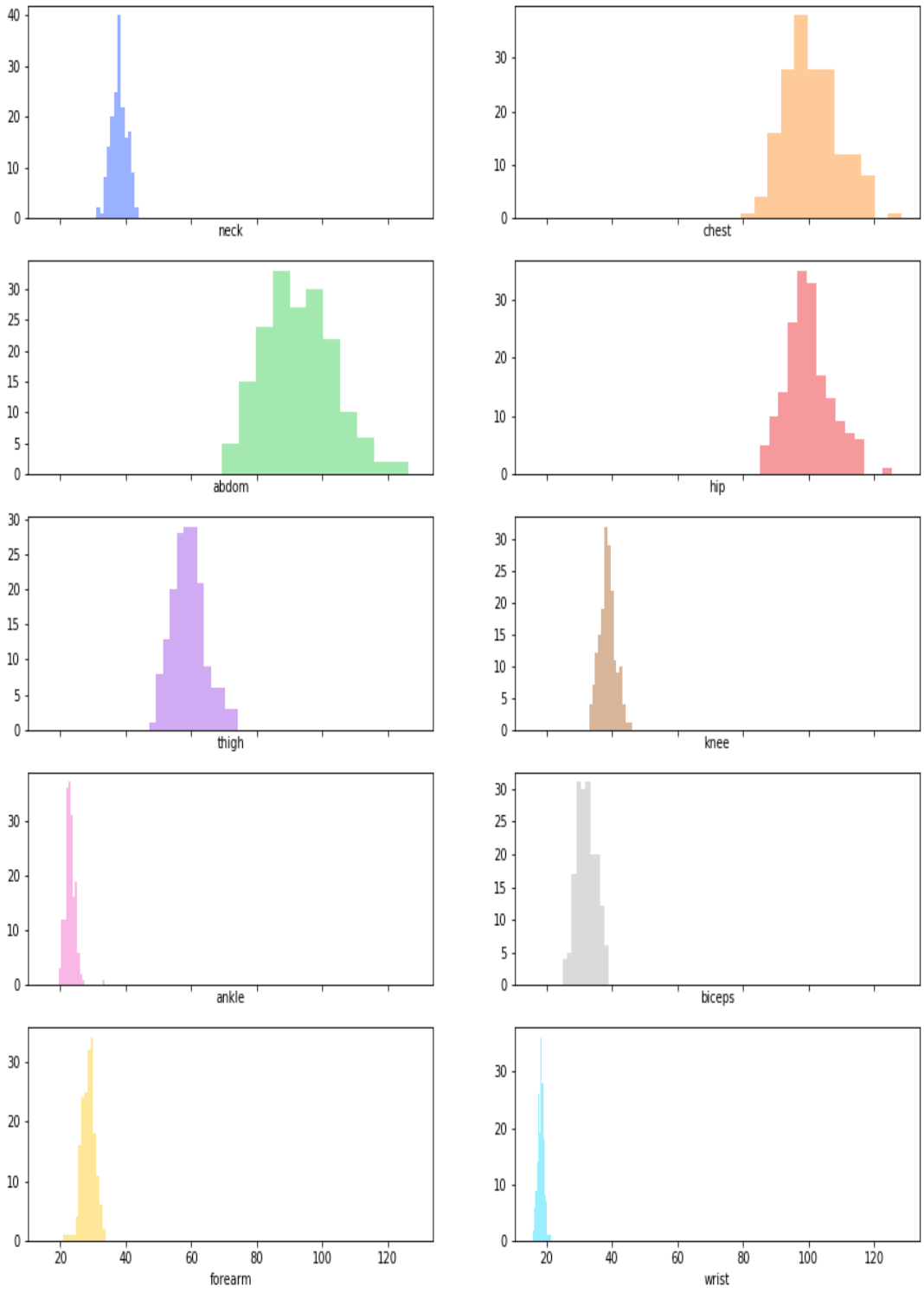
(b) Exploratory Data Analysis

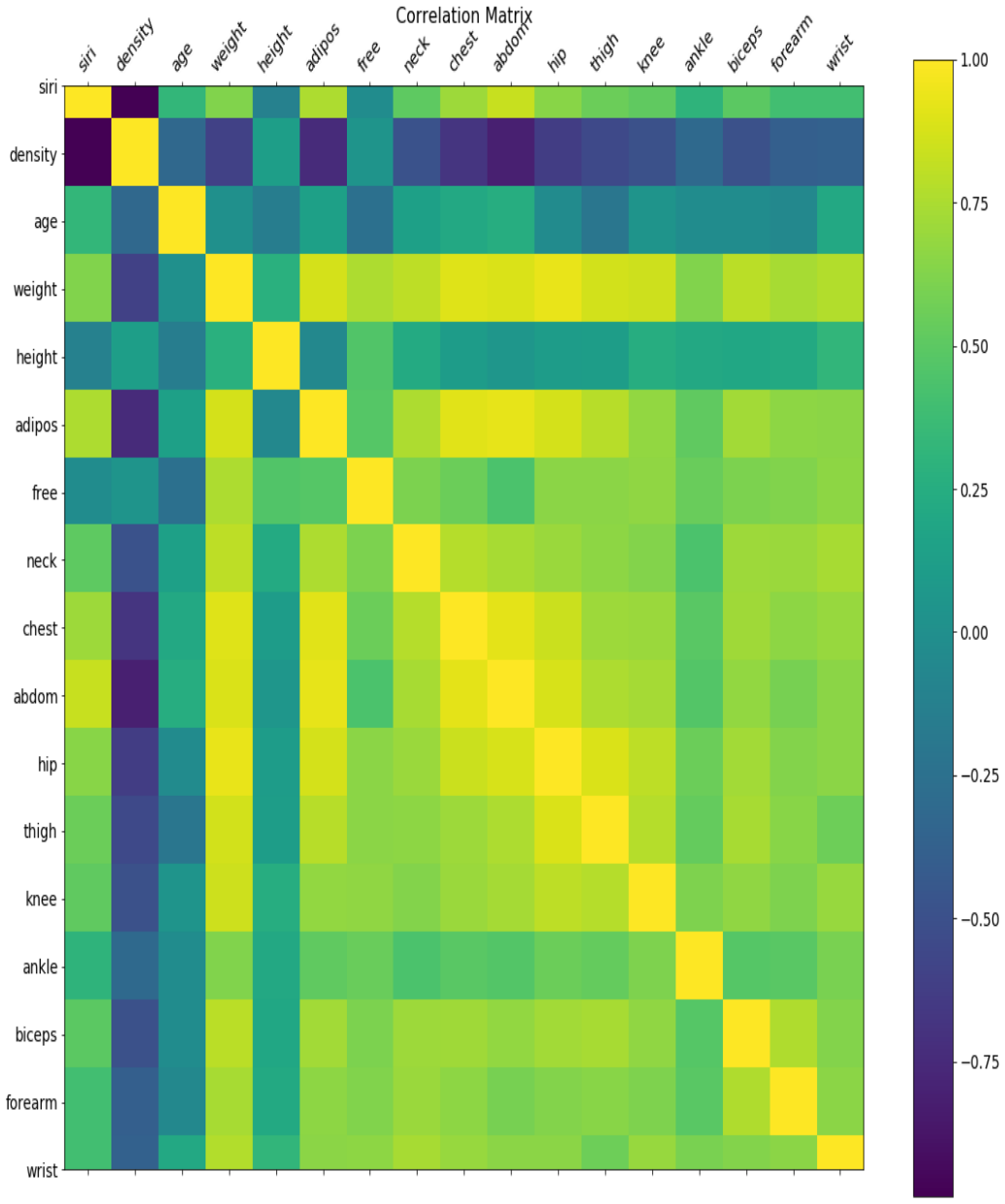
Exploratory data analysis is an important step to understand the data set. We will select interesting features for data exploration. Since brozek is related to body density which is further related to weight, naturally weight is related to brozek values. Therefore, we are interested in humans physical attributes. From below descriptive statistics, we can realize data distribution of each feature. Interestingly, abdom feature has more widespread distribution compared to other features. And, as for wrist, there is little variation which is reasonable. After descriptive statistics are derived, we try to visualize these features to clearly observe possible relationship.

	neck	chest	abdom	hip	thigh	knee	ankle	biceps	forearm	wrist
count	176.000000	176.000000	176.000000	176.000000	176.000000	176.000000	176.000000	176.000000	176.000000	176.000000
mean	37.990909	100.917045	92.928977	99.975000	59.354545	38.622159	23.041477	32.169318	28.645455	18.231250
std	2.393713	8.541084	10.704072	6.984288	5.164835	2.438950	1.585339	2.994504	2.036547	0.939234
min	31.100000	79.300000	69.400000	85.000000	47.200000	33.000000	19.700000	24.800000	21.000000	15.800000
25%	36.400000	94.000000	85.025000	95.475000	56.000000	37.275000	22.000000	30.100000	27.300000	17.600000
50%	38.000000	99.900000	91.550000	99.300000	59.100000	38.500000	22.800000	32.000000	28.800000	18.300000
75%	39.650000	106.100000	100.025000	103.950000	62.150000	40.000000	24.000000	34.325000	29.925000	18.800000
max	43.900000	128.300000	126.200000	125.600000	74.400000	46.000000	33.700000	39.100000	33.800000	21.400000

From below histogram picture, we could clearly see the distribution of data from each feature. Clearly, abdom is more widespread, but wrist is more central. As for other features, there might be outliers in hip and chest features. Based on the below figure, clearly one tick appears outside distribution of chest and hip. If this person were super fat, other columns might have same situation happening which he might have a larger abdom as we expect. However, the distribution of abdom does not follow this trend.

Then, we switch our focus to correlation among features. The below correlation matrix is provided and we can understand whether each feature has strong correlation or not which is really important because when we utilize linear regression, we often delete some of columns that are related. In this way, we could avoid collinearity problems. From the correlation matrix, height, age, ankle and density have weak correlation among other features. Note that correlation does not mean causation. In other words, we could not make conclusions that weight causes bigger hips, chest or knee, for example. To sum up, after data exploration, we can clearly understand our dataset and then we could clean our dataset for further analysis. After cleaned dataset is present, models could be built accordingly.


neck
chest
abdom
hip
thigh
knee
ankle
biceps
forearm
wrist


Correlation Matrix
siri
density
age
weight
height
adipos
free
neck
chest
abdom
hip
thigh
knee
ankle
biceps
forearm
wrist
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

(c) Methods

1. **Linear Regression with All Predictors**
   Common linear regression method in this problem is that we do not remove any predictors.

   - Assumption: the regression function $E(Y|X)$ is linear in the inputs $X_1, \ldots, X_{17}$. The linear regression model has the form $f(X) = \beta_0 + \sum_{j=1}^{17} X_j\beta_j + \epsilon$ where ϵ is random error(normal distributed).
   - α-level: for each feature, we assign α as 0.05
   - VIF analysis: in vif output, we could clearly see there is colinearity among predictor variables within a multiple regression
   - Variable selection: in this problem, we do not select any variables. (All value of predictors is greater than 5)
   - P-value: some of features, like age, height, adipos, neck, chest, abdom, hip, ankle, forearm and wrist, are greater than α.
   - Goodness of fit: R^2 is extremely high in this multiple regression.
   - Validation: Monte Carlo Cross-Validation algorithm is performed to validate our model.
   - Statistical Package: scikit-learn and statsmodels package are used for this problem.

Dep. Variable:	brozek	R-squared:	1.000
Model:	OLS	Adj. R-squared:	0.999
Method:	Least Squares	F-statistic:	1.878e+04
Date:	Wed, 29 Jan 2020	Prob (F-statistic):	1.29e-251
Time:	18:14:37	Log-Likelihood:	51.348
No. Observations:	176	AIC:	-66.70
Df Residuals:	158	BIC:	-9.628
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	11.9563	4.876	2.452	0.015	2.326	21.587
siri	0.8823	0.014	61.928	0.000	0.854	0.910
density	-10.0529	4.334	-2.320	0.022	-18.613	-1.493
age	-0.0009	0.002	-0.522	0.602	-0.004	0.002
weight	0.0106	0.006	1.911	0.058	-0.000	0.022
height	-0.0010	0.005	-0.187	0.852	-0.011	0.009
adipos	-0.0201	0.017	-1.185	0.238	-0.054	0.013
free	-0.0141	0.006	-2.241	0.026	-0.026	-0.002
neck	0.0009	0.012	0.072	0.943	-0.023	0.025
chest	0.0072	0.006	1.206	0.230	-0.005	0.019
abdom	-0.0004	0.006	-0.058	0.954	-0.013	0.012
hip	-0.0098	0.008	-1.270	0.206	-0.025	0.005
thigh	0.0265	0.008	3.302	0.001	0.011	0.042
knee	-0.0247	0.013	-1.908	0.058	-0.050	0.001
ankle	0.0066	0.013	0.498	0.619	-0.019	0.033
biceps	-0.0211	0.009	-2.308	0.022	-0.039	-0.003
forearm	0.0211	0.013	1.641	0.103	-0.004	0.046
wrist	0.0552	0.031	1.792	0.075	-0.006	0.116

Omnibus:	134.752	Durbin-Watson:	1.954
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6140.722
Skew:	2.172	Prob(JB):	0.00
Kurtosis:	31.609	Cond. No.	1.42e+05

2. **Linear regression with the best subset of k = 5 predictors variables**

   - Assumption: the regression function $E(Y|X)$ is linear in the inputs $X_1, \ldots, X_{17}$. The linear regression model has the form $f(X) = \beta_0 + \sum_{j=1}^{5} X_j\beta_j + \epsilon$ where ϵ is random error(normal distributed).
   - α-level: for each feature, we assign α as 0.05
   - VIF analysis: in vif output, we could clearly see there is colinearity among 5 predictor variables within the multiple regression
   - Variable selection: in this problem, we would like to choose 5 best predictors. We perform regsubsets in R first to select features since Python does not have preogram or package to support Best subset selection. We get Siri, density, thigh, knee and wrist as best 5 features.
   - P-value: wrist feature is greater than α.
   - Goodness of fit: R^2 is extremely high in this multiple regression (0.99).
   - Validation: Monte Carlo Cross-Validation algorithm is performed to validate our model.
   - Statistical Package: leaps package in R, scikit-learn and statsmodels package are used for this problem.

OLS Regression Results

Dep. Variable:	brozek	R-squared:	0.999
Model:	OLS	Adj. R-squared:	0.999
Method:	Least Squares	F-statistic:	6.220e+04
Date:	Wed, 29 Jan 2020	Prob (F-statistic):	2.90e-275
Time:	20:02:20	Log-Likelihood:	42.628
No. Observations:	176	AIC:	-73.26
Df Residuals:	170	BIC:	-54.23
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	11.5415	4.650	2.482	0.014	2.362	20.721
siri	0.9024	0.010	91.872	0.000	0.883	0.922
density	-9.6991	4.259	-2.277	0.024	-18.106	-1.292
thigh	0.0135	0.005	2.845	0.005	0.004	0.023
knee	-0.0286	0.011	-2.590	0.010	-0.050	-0.007
wrist	0.0366	0.022	1.679	0.095	-0.006	0.080

Omnibus:	179.458	Durbin-Watson:	2.074
Prob(Omnibus):	0.000	Jarque-Bera (JB):	10288.723
Skew:	3.422	Prob(JB):	0.00
Kurtosis:	39.826	Cond. No.	3.29e+04

3. **Linear regression with variables (stepwise) selected using AIC**

- Assumption: the regression function $E(Y|X)$ is linear in the inputs $X_1, \ldots, X_{17}$. The linear regression model has the form $f(X) = \beta_0 + \sum_{j=1}^{7} X_j\beta_j + \epsilon$ where ϵ is random error(normal distributed).
- α-level: for each feature, we assign α as 0.05
- VIF analysis: in vif output, we could clearly see there is colinearity among 7 predictor variables within the multiple regression(in appendix)
- Variable selection: in this problem, we would like to choose predictors based on AIC. We wrote a program to select predictors based on AIC. And then, we get 'siri', 'density', 'thigh', 'knee', 'wrist', 'biceps', 'forearm' and 'hip' in this model.
- P-value: wrist feature is greater than α.
- Goodness of fit: R^2 is extremely high in this multiple regression (0.99).
- Validation: Monte Carlo Cross-Validation algorithm is performed to validate our model.
- Statistical Package: leaps package in R, scikit-learn and statsmodels package are used for this problem.

OLS Regression Results

Dep. Variable:	brozek	R-squared:	0.999
Model:	OLS	Adj. R-squared:	0.999
Method:	Least Squares	F-statistic:	4.038e+04
Date:	Wed, 29 Jan 2020	Prob (F-statistic):	3.70e-270
Time:	20:17:13	Log-Likelihood:	47.541
No. Observations:	176	AIC:	-77.08
Df Residuals:	167	BIC:	-48.55
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	11.2750	4.582	2.461	0.015	2.229	20.321
siri	0.9050	0.010	92.305	0.000	0.886	0.924
density	-9.4675	4.205	-2.252	0.026	-17.769	-1.166
thigh	0.0238	0.007	3.497	0.001	0.010	0.037
knee	-0.0243	0.011	-2.193	0.030	-0.046	-0.002
wrist	0.0484	0.024	1.994	0.048	0.000	0.096
biceps	-0.0218	0.009	-2.484	0.014	-0.039	-0.004
forearm	0.0199	0.012	1.694	0.092	-0.003	0.043
hip	-0.0089	0.006	-1.592	0.113	-0.020	0.002

Omnibus:	166.573	Durbin-Watson:	1.992
Prob(Omnibus):	0.000	Jarque-Bera (JB):	8090.463
Skew:	3.078	Prob(JB):	0.00
Kurtosis:	35.640	Cond. No.	5.78e+04

4. Ridge regression

   - Assumption: The ridge regression model has the form $f(X) = \beta_0 + \sum_{j=1}^{17} X_j\beta_j + \epsilon$ where ϵ is random error(normal distributed). But, in Ridge regression, we want to find $\hat{\beta}^{ridge}$ and add a penalty function. $\hat{\beta}^{ridge} = argmin \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{i=1}^{17} x_{ij}\beta_j)^2 + \lambda(\sum_{i=1}^{17} \beta_j^2)$
   - λ: choosing right λ is important, so we use cross validation to get $\lambda = 1e-06$ gives great result.
   - Validation: Monte Carlo Cross-Validation algorithm is performed to validate our model.
   - Statistical Package: scikit-learn is used for this problem.

5. Lasso Regression

   - Assumption: The Lasso regression model has the form $f(X) = \beta_0 + \sum_{j=1}^{17} X_j\beta_j + \epsilon$ where ϵ is random error(normal distributed). But, in Lasso regression, we want to find $\hat{\beta}^{lasso}$ and add a penalty function. $\hat{\beta}^{lasso} = argmin \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{i=1}^{17} x_{ij}\beta_j)^2 + \lambda(\sum_{i=1}^{17} |\beta_j|)$
   - λ: choosing right λ is important, so we use cross validation to get $\lambda = 0.001$ gives great result.
   - Validation: Monte Carlo Cross-Validation algorithm is performed to validate our model.
   - Statistical Package: scikit-learn is used for this problem.

6. Principal Component Regression

   - Assumption: The Principal Component Regression forms the derived input columns $z_m = Xv_m$ and then regresses y on $z_1, \ldots, z_M$ for some $M \leq p$. Since $z_i \perp z_j, \forall i \neq j$, a new dimension space is derived. That is, $\hat{y}^{pcr} = \bar{y}1 + \sum_{m=1}^{M} \hat{\theta}_m z_m$ where $\hat{\theta}_m = \langle z_m, y\rangle / \langle z_m, z_m\rangle$
   - Number of components:in this problem, explained variance ratio provides component selection which means the number of component could explain how much variance in this data set. If two or three components could explain most variance, these two or three components will be chosen.
   - Validation: Monte Carlo Cross-Validation algorithm is performed to validate our model.
   - Statistical Package: scikit-learn is used for this problem.

7. Partial Least Squares

   - Assumption: The Partial Least Squares has the form $max_\alpha Corr^2(y, X\alpha)Var(X\alpha)$ subject to $||\alpha|| = 1$, $\alpha^T S\hat{l}_i = 0, i = 1, \ldots, 16$ where S is smaple variance matrix of the x_j
   - Number of components: Cross validation is utilized to get the number of component that could show strong relationship with the response variable. For this training data, we select three components to build our model.
   - Validation: Monte Carlo Cross-Validation algorithm is performed to validate our model.
   - Statistical Package: scikit-learn is used for this problem.

(d) Results

The results are derived and there are some interesting numbers in the below table. Our training data contains 176 observations and the rest is test data. For this training data, the MSE of linear regression methods is lower than the mean of MSEs from Monte Carlo. However, using principal components of partial least squares gives higher MSE than MSEs of Monte Carlo.

Methods	MSE	MC MSE Mean	MC MSE Variance
Linear Regression with All	0.017	0.057	0.0024
Linear Regression with best 5	0.013	0.0395	0.001
Linear Regression on AIC	0.016	0.0399	0.001
Ridge Regression	0.017	0.057	0.002
Lasso Regression	0.017	0.043	0.0008
Principal Component Regression	2.83	2.19	0.24
Partial Least Squares	3.477	0.76	0.12

(e) Findings

From the Results section, we found different MSE based on different methods. In linear regression methods, we found if we use variable selection based on AIC or Best Subset selection, we could see our MSE going down which means that our model is improved. But as for Ridge regression or Lasso regression, those are different approaches, but in this problem, we did not observe improvement of Linear Regression Model with all predictors based on MC MSE mean. What's more, in Principal Component Regression and Partial Least squares, first we select possible components and then create a new dimension reduction dataset. The reason that MSE from those two methods are higher than other MSEs might be that since we just choose two or three components, this dimension-deduction dataset contains less comprehensive dataset that have higher variance. But if we selected 17 components which means we do not do dimension deduction, we would derive around 0.017 MSE. After we investigate MSE from our training data, we perform Monte Carlo Algorithm to check whether our model is suitable or not. Variance of each method from Monte Carlo is small, but MSEs are different from MSEs of training data. Therefore, we could make sure whether we should use models built based on training data or not. If MC MSE mean is greater than MSE of training data, we could consider using trained models. The last important thing to mention is that MSE of PLS is greater than MC MSE Mean. At this time, we would think PLS based on our training data is not persuasive. We might create new training data to train our model again.