Homework 2

Lian Di 1801212881

Problem 1

1.（Matlab code）

```
function closed_form_1()
M=csvread('climate_change_1.csv',1,0)
Y_train = M(1:284,11)
Y_test = M(285:308,11)
X_train = [ones(284,1),M(1:284,3:10)]
X_test = [ones(24,1),M(285:308,3:10)]
%regress
theta = (X_train' * X_train) ^ (-1) * X_train' * Y_train
%get R2
e = Y_train - X_train * theta
RSS = e' * e
ESS = sum((X_train * theta-mean(Y_train)).^2)
TSS = sum((Y_train - mean(Y_train)).^2)
R2_train = ESS/TSS
%Test
e_test = Y_test - X_test * theta
RSS_test = e_test' * e_test
ESS_test = sum((X_test * theta-mean(Y_test)).^2)
TSS_test = sum((Y_test - mean(Y_test)).^2)
R2_test = ESS_test/TSS_testEnd
```

2.

Temp=
-124.594
+0.064205*MEI
+0.006457*CO2
+0.000124*CH4
-0.01653*N2O
-0.00663*CFC-11
+0.003808*CFC-12
+0.093141*TSI
-1.53761*Aerosols

Training set: R^2 = 0.7509

Testing set: R^2 = 0.2250

3.

	P-value
Intercept	1.43E-09
MEI	4.9E-20
CO2	0.005053
CH4	0.810146
N2O	0.054669
CFC-11	5.96E-05
CFC-12	0.00021
TSI	1.1E-09
Aerosols	5.41E-12

As the p-value shows, MEI, CO2, CFC-11 CFC-12 TSI Aerosols are significant in the model.

4.

The regression model is linear in parameters.
The mean of residuals is zero.
Homoscedasticity of residuals or equal variance.
No autocorrelation of residuals.
The X variables and residuals are uncorrelated.
Normality of residuals.
And X^TX should be Invertible matrix(Full rank)
When applying the closed form solution to climate_change_2.csv, as NO added in, the X^TX is very close to singular value.(not full rank) We can not get the right $(X^TX)^{-1}$, and we can not get the right answer, so the solution is unreasonable.
Matlab code:

```
M_2=csvread('climate_change_2.csv',1,0)
Y_2 = M_2(:,12)
X_2 = [ones(308,1),M_2(:,3:11)]
theta_2 = (X_2' * X_2) ^ (-1) * X_2' * Y_2
e_2 = Y_2 - X_2 * theta_2
RSS_2 = e_2' * e_2
TSS_2 = sum((Y_2 - mean(Y_2)).^2)
R2_2 = 1 - RSS_2/TSS_2
```

Problem 2

1.

L1 Regularization:(Lasso regression:)
$J(\theta)=1/2(X\theta-Y)^T(X\theta-Y)+\alpha||\theta||_1$
L2 Regularization:(Ridge regression:)
$J(\theta)=1/2*(X\theta-Y)^T(X\theta-Y)+1/2*\alpha||\theta||^2_2$

2.（Matlab code）

```
function closed_form_2(lambda)
M=csvread('climate_change_1.csv',1,0);
Y_train = M(1:284,11);
Y_test = M(285:308,11);
X_train = [ones(284,1),M(1:284,3:10)];
X_test = [ones(24,1),M(285:308,3:10)];
theta = (X_train' * X_train + lambda*eye(9)) ^ (-1) * X_train' * Y_train
e = Y_train - X_train * theta;
RSS = e' * e;
ESS = sum((X_train * theta-mean(Y_train)).^2);
TSS = sum((Y_train - mean(Y_train)).^2);
R2_train = ESS/TSS

e_test = Y_test - X_test * theta;
RSS_test = e_test' * e_test;
ESS_test = sum((X_test * theta-mean(Y_test)).^2);
TSS_test = sum((Y_test - mean(Y_test)).^2);
R2_test = ESS_test/TSS_test
end
```

3.

In OLS:

Norm of theta: 124.6038

R2_train = 0.7509

R2_test = 0.2250

In L2 Regularization(lambda = 0.1)

Norm of theta: 0.8733

R2_train = 0.6945

R2_test = 0.6733

When lambda=0.1, theta =

```
   -0.0250
    0.0507
    0.0070
    0.0001
   -0.0148
   -0.0061
    0.0037
    0.0014
   -0.8713
```

It will reduce the coefficient of unimportant prediction factors close to 0 and avoid overfitting, Temp is less sensitive to single variable, so it is more robust.

4.

```
for i = 0 : 4
    lambda = 10/10^i
    closed_form_2(lambda)
```

end

Lambda	R2_train	R2_test
0.001	0.7148	0.5625
0.01	0.7117	0.5853
0.1	0.6945	0.6733
1	0.6795	0.8468
10	0.6746	0.9409

There are some ways for cross validation: Simple Cross Validation, 2-fold Cross Validation, and K-fold Cross Validation.
I choose to use Simple Cross Validation, get the MSE to measure the model.
The original data is randomly divided into two groups, one is the training set, the other is the verification set. The training set is used to train the classifier, and then the verification set is used to verify the model. The final classification accuracy is recorded as the performance index of the classifier.
So, MSE = RSS_test/24

Lambda	MSE
0.001	0.0136
0.01	0.0139
0.1	0.0152
1	0.0177
10	0.019

Choose the lambda with the least MSE, so choose Lambda = 0.001.

Problem 3

1.Workflow:

For P features, from k = 1 to k = P:
Choose any k features from P features, establish C (P, K) models, and choose the best one (MSE minimum or R2 maximum);
Select an optimal model from the P optimal models (cross validation error).

2.

Use MEI, CO2, CFC-11 CFC-12 TSI Aerosols are significant in the model(Get rid of CH4, N2O), get the result:

Temp=
-122.253
+0.064214*MEI
+0.004061*CO2
-0.00431*CFC-11
+0.00243*CFC-12

+0.08852*TSI
-1.56651*Aerosols

Problem 4

```
M=csvread('climate_change_1.csv',1,0);
Y = M(1:284,11);
X = [ones(284,1),M(1:284,3:10)];

n = size(X,2)
m = size(X,1)
theta = 0.01*ones(n,1);
temp = zeros(n,1);
k = 0;%iteration times
alpha = 0.0001 %learning rate
while true
    for j = 1:n
        theta(j) = theta(j) - (alpha*sum((X*theta-Y).*X(:,j))/m)
    end
    k = k+1
    if 1/2*( norm(Y - X*theta) )^2 < 0.0001 || k>1000
        break;
    end
end
```