Name: Dong Ao
ZID: z5305320
Q1:
(a): $\widetilde{Y_i} = \widetilde{\beta_0} + \widetilde{\beta_1}\widetilde{X_i} = \widetilde{\beta_0} + \widetilde{\beta_1}(cX_i + cd) = \widetilde{\beta_0} + cd\widetilde{\beta_1} + c\widetilde{\beta_1}X_i = \widehat{\beta_0} + \widehat{\beta_1}X_i$.

Thus, $\widehat{\beta_0} = \widetilde{\beta_0} + cd\widetilde{\beta_1}$ and $\widehat{\beta_1} = c\widetilde{\beta_1}$. We have $\widetilde{\beta_1} = \frac{\widehat{\beta_1}}{c}$ and $\widetilde{\beta_0} = \widehat{\beta_0} - cd\frac{\widehat{\beta_1}}{c} = \widehat{\beta_0} - d\widehat{\beta_1}$.

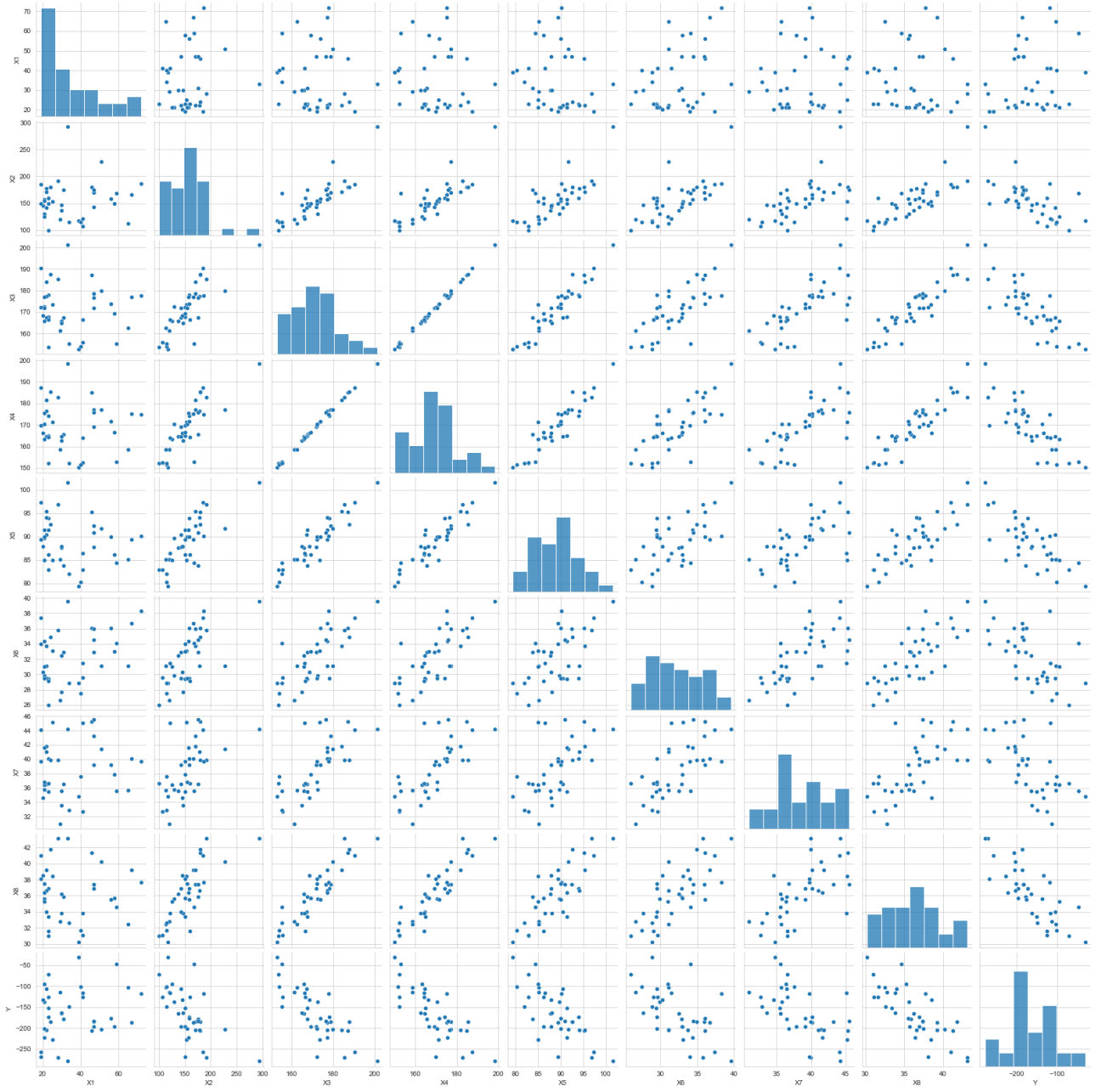
$\hat{\sigma} = \sqrt{\frac{(Y_i - \widehat{Y_i})^2}{n-2}}$. Since $\widehat{Y_i}$ is the prediction made by the model, they did not change. Thus, for $\tilde{\sigma}$, it is still the same as $\hat{\sigma}$ from the formula provided.

$\widetilde{\beta_1} = \frac{\widehat{\beta_1}}{c}$, $\widetilde{\beta_0} = \widehat{\beta_0} - d\widehat{\beta_1}$ and $\tilde{\sigma} = \hat{\sigma}$.

(b): We know that $Y_{T_i} = \widehat{\beta_0} + \widehat{\beta_1}X_i$ and for any Y_{T_i}, X_i can only be 1 since it is treatment group. Thus, $Y_{T_i} = \widehat{\beta_0} + \widehat{\beta_1}$. Similarly, $Y_{P_i} = \widehat{\beta_0} + \widehat{\beta_1} * 0 = \widehat{\beta_0}$ because it is placebo group without receiving a dose of drug. We say that n_T and n_P are respectively the number of data points of treatment group and non-treatment group. Then, $\overline{Y_T} = \frac{\sum_{i=1}^{n_T} Y_{T_i}}{n_T} = \frac{\sum_{i=1}^{n_T} \widehat{\beta_0} + \widehat{\beta_1}}{n_T} = \frac{n_T(\widehat{\beta_0} + \widehat{\beta_1})}{n_T} = \widehat{\beta_0} + \widehat{\beta_1}$ and $\overline{Y_P} = \frac{\sum_{i=1}^{n_P} Y_{P_i}}{n_P} = \frac{\sum_{i=1}^{n_P} \widehat{\beta_0}}{n_P} = \widehat{\beta_0}$.
Thus, $\widehat{\beta_0} = \overline{Y_P}$ and $\widehat{\beta_1} = \overline{Y_T} - \overline{Y_P}$.

Q2

(a):



From the pairs plot, Y has a negative correlation respectively with $X_2, X_3..X_8$. Moreover, there are no particular relationships between X_1 and Y ,and X_3 is proportional to X_4. In linear model regression, X_1 will possibly contribute nothing and feature 3 and 4 may be considered as one feature.

(b):

$\sum_{i=1}^{n} X_{ij}^2$ j = 1: X1 38.0

j = 2: 38.00000000000001

j = 3: 38.00000000000002

j = 4: 37.99999999999999

j = 5: 38.0

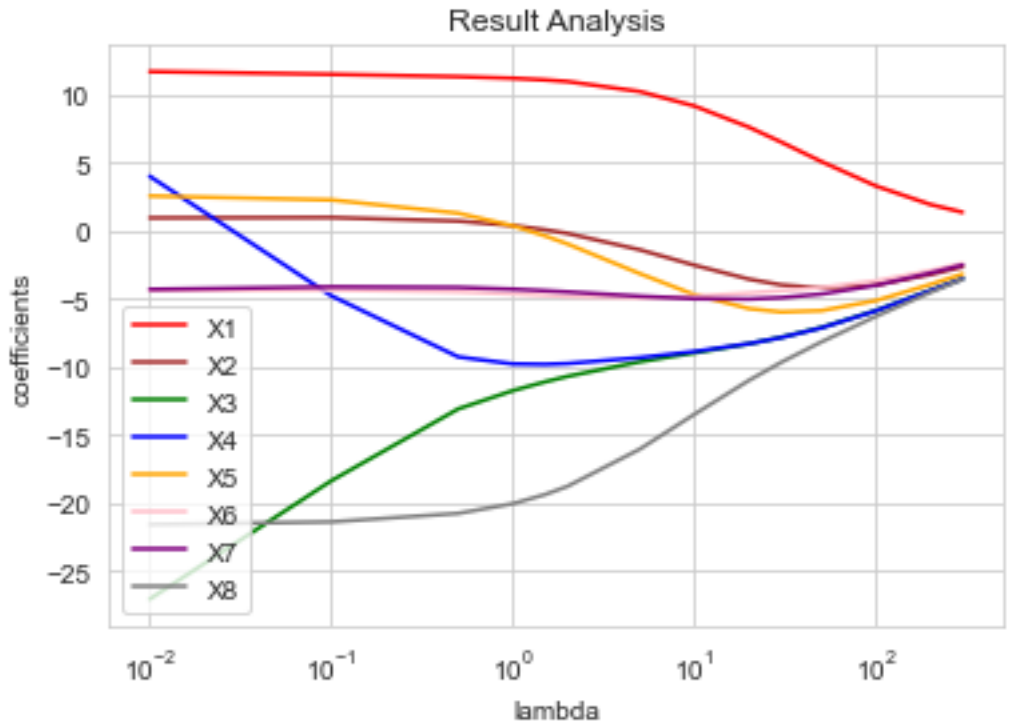
j = 6: 38.0

j = 6: 38.00000000000001

j = 1: 37.999999999999986

(c):

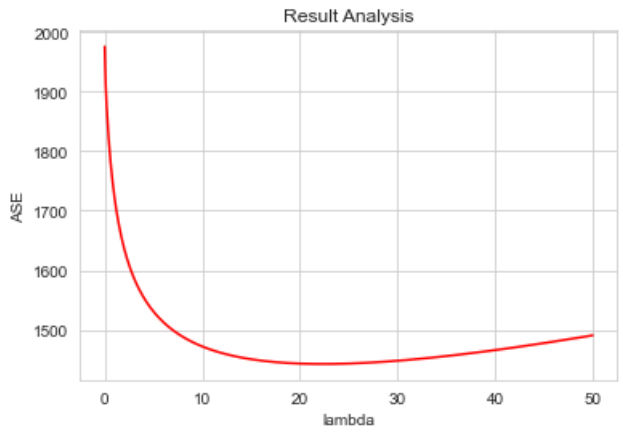
```
# load the diabetes dataset
df=pd.read_csv('data.csv')
target_name="Y"
target=df[target_name]

stand = np.std(df)
newdf = (df - df.mean())/np.std(df)
X=newdf[['X1', 'X2', 'X3', 'X4', 'X5', 'X6', 'X7', 'X8']]
Y=df[['Y']]

lamb = [0.01, 0.1, 0.5, 1, 1.5, 2, 5, 10, 20, 30, 50, 100, 200, 300]
coefs = []
for i in range(8):
    coefs.append([])
ridge = linear_model.Ridge()
for a in lamb:
    ridge.set_params(alpha=a)
    ridge.fit(X, Y)
    for i in range(8):
        coefs[i].append(ridge.coef_[0][i])

colors = ['red', 'brown', 'green','blue','orange','pink', 'purple', 'grey']
ax = plt.gca()
plt.title('Result Analysis')
for i in range(8):
    ax.plot(lamb, coefs[i], color=colors[i], label='X' + str(i + 1))
    ax.set_xscale('log')

ax.legend()
plt.xlabel('lambda')
plt.ylabel('coefficients')

plt.show()
```



I saw that when lambda get larger, the coefficients all tend to converge to 0 even it has not been reached in the plot. Feature 3 and 4 shows some strong relationship when $\log(\lambda) < 10^0$, when X_3 is increasing, X_4 decrease in almost the same speed. Feature 3, 4 and 5 changes their slope almost at the same time.

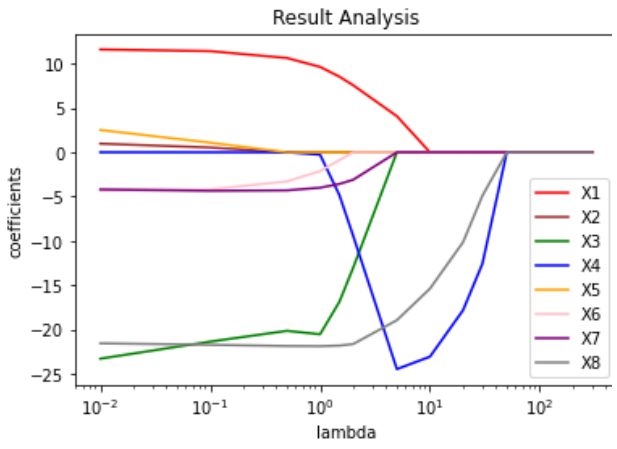
(d):

```
for lams in lmd:
    sum = 0
    Y_pre = []
    for i in range(38):
        X_test = X[i].reshape(1,-1)
        X1 = np.delete(X, i,axis=0)
        Y1 = np.delete(Y, i,axis=0)

        ridge2 = linear_model.Ridge(alpha=lams)
        ridge2.fit(X1, Y1)
        y_pre = ridge2.predict(X_test)
        Y_pre.append(y_pre[0])
        sum += math.pow(y_pre[0] - Y[i], 2)
    ASE = sum / 38
    ASEs.append(ASE)
print(f"the mimimal ASE is {min(ASEs)} with lambd ={ASEs.index(min(ASEs)) * 0.1 }")
bx.plot(lmd, ASEs, color='green')
plt.show()

sum = 0
for i in range(38):
    X_test = X[i].reshape(1,-1)
    X1 = np.delete(X, i,axis=0)
    Y1 = np.delete(Y, i,axis=0)

    ridge2 = linear_model.LinearRegression()
    ridge2.fit(X1, Y1)
    y_pre = ridge2.predict(X_test)
    sum += math.pow(y_pre[0] - Y[i], 2)
OLS = sum / 38

print(f"the OLS is {OLS}")
```



The least MSE from using ridge (1442.6982227952915 with lambda 22.3) is much le ss than the OLS one (1975.4147393421724).

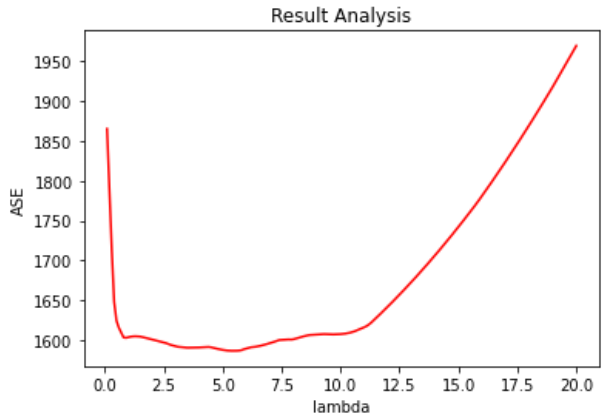
(e):




```
stand = np.std(df)
newdf = (df - df.mean())/np.std(df)

X=newdf[['X1', 'X2', 'X3', 'X4', 'X5', 'X6', 'X7', 'X8']]
Y=df[['Y']]
lamb = [0.01, 0.1, 0.5, 1, 1.5, 2, 5, 10, 20, 30, 50, 100, 200, 300]
coefs = []
for i in range(8):
    coefs.append([])
lasso = Lasso()
for a in lamb:
    lasso.set_params(alpha=a)
    lasso.fit(X, Y)
    for i in range(8):
        coefs[i].append(lasso.coef_[i])

colors = ['red', 'brown', 'green','blue','orange','pink', 'purple', 'grey']
ax = plt.gca()
plt.title('Result Analysis')
for i in range(8):
    ax.plot(lamb, coefs[i], color=colors[i], label='X' + str(i + 1))
    ax.set_xscale('log')

ax.legend()
plt.xlabel('lambda')
plt.ylabel('coefficients')

plt.show()
```

When lambda get extremely large, all the coefficients tend to 0. When lambda is around 5, the absolute value of the coefficient of X_4 becomes very large, which is increasing. Also, X_3 and X_4 have relationships from the graph. When $\log(\lambda)$ around 10^0, one of X_4 and X_3 is chosen in Lasso and the other one changes immediately, which means they are likely to be reduced to one single variable.

(f):



The least error from using lasso (1442. 1586.6715081806428
with lambda 5.4) is much less than the OLS one (1975.4147393421724).

(g):
The MSE from Ridge and Lasso are quite different. The MSE of Lasso almost remain still when lambda is in range of 1 and 10. Moreover, the MSE of Lasso gro ws extremely fast when lambda is greater than 10 but Ridge's grows relatively slow. I prefer to choose Ridge. Using the predicted model we got, the least MSE of Ridg e is less than Lasso's. From the definition of Lasso, it has an ability to select the v ariates and exclude those from our model by giving very small coefficient. But from the pairs plot we got, we found that only X_1 is irrelated to Y, which means unselect some variates in this circumstance is not that efficient. We have 7 out of 8 features in our model which influenced out model much, that is why we choose Ridge here.

Q2
(a):
We suppose that when $j = k$, $\left(\sum_{i=1}^{n} X_{ij}Y_i\right)$ is the max. Thus, $\max_j\left(\sum_{i=1}^{n} |X_{ij}Y_i|\right) = \sum_{i=1}^{n} |X_{ik}Y_i|$.

$$RHS = \max_j\left(\sum_{i=1}^{n} |X_{ij}Y_i|\right)\sum_{j=1}^{p} |\beta_j| = \sum_{i=1}^{n} |X_{ik}Y_i| \sum_{j=1}^{p} |\beta_j| = \sum_{j=1}^{p}\sum_{i=1}^{n} |X_{ik}Y_i\beta_j|.$$

$$\hat{Y} = X\beta = \left[\sum_{j=1}^{p} \beta_j X_{1j}, \sum_{j=1}^{p} \beta_j X_{2j} \ldots \sum_{j=1}^{p} \beta_j X_{nj}\right]^T$$

$$LHS = | < Y, X\beta > | = | < Y, \hat{Y} > | = \left| \sum_{j=1}^{p} \sum_{i=1}^{n} X_{ij} Y_i \beta_j \right|.$$

$$LHS = \left| \sum_{j=1}^{p} \sum_{i=1}^{n} X_{ij} Y_i \beta_j \right| \leq \sum_{j=1}^{p} \left| \sum_{i=1}^{n} X_{ij} Y_i \beta_j \right| \leq \sum_{j=1}^{p} \sum_{i=1}^{n} |X_{ij} Y_i \beta_j| \leq \sum_{j=1}^{p} \sum_{i=1}^{n} |X_{ik} Y_i \beta_j| = RHS$$

$$(The\ principle\ of\ absolute\ value\ inequality: |a+b| \leq |a| + |b|)$$

(b):

$$OE = \frac{1}{2} \left|| \mathrm{Y} - \mathrm{X}\beta \right||_2^2 + \lambda \left|| \beta \right||_1 \geq \frac{1}{2} \left|| \mathrm{Y} - \mathrm{X}\beta \right||_2^2 + \max_j \left(\sum_{i=1}^{n} \left| X_{ij} Y_i \right| \right) \left|| \beta \right||_1$$

$$\geq \frac{1}{2} \left|| \mathrm{Y} - \mathrm{X}\beta \right||_2^2 + | < Y, X\beta > | (From\ \ a)$$

$$\geq \frac{1}{2} \sum_{i=1}^{n} \left(Y_n - \hat{Y}_n \right)^2 + \sum_{i=1}^{n} Y_n \hat{Y}_n \geq \frac{1}{2} \sum_{i=1}^{n} Y_n^2 + \hat{Y}_n^2 - 2 Y_n \hat{Y}_n + \sum_{i=1}^{n} Y_n \hat{Y}_n = \frac{1}{2} \sum_{i=1}^{n} Y_n^2 + \hat{Y}_n^2$$

$$= \frac{1}{2} \sum_{i=1}^{n} \hat{Y}_n^2 + \frac{1}{2} \sum_{i=1}^{n} Y_n^2.$$

It reaches the minimum when $\sum_{i=1}^{n} \hat{Y}_n^2$ is minimal. We know $min\left(\widehat{Y_n}^2 \right) = 0$ and only when $\hat{\beta} = 0_p$, we can reach the minimum. Thus, $\hat{\beta} = 0_p$ is a solution of Lasso prob lem.

(c):

$$l(0_p) = \frac{1}{2} \left|| Y \right||_2^2 + \lambda \left|| 0_p \right||_1 = \frac{1}{2} \left|| Y \right||_2^2$$

For any $\beta \neq 0_p$, $l(\beta) = \frac{1}{2} \left|| \mathrm{Y} - \mathrm{X}\beta \right||_2^2 + \lambda \left|| \beta \right||_1 = \frac{1}{2} \left(\left|| Y \right||_2 + \left|| X\beta \right||_2 - 2 \sum_{i=1}^{n} Y_n X_n \beta \right) + \lambda$ $\left|| \beta \right||_1 = \frac{1}{2} \left(\left|| Y \right||_2 + \left|| X\beta \right||_2 \right) + \left(\lambda \left|| \beta \right||_1 - \sum_{i=1}^{n} Y_n (X_n \beta) \right)$.

We know that from (b), when λ is sufficiently large, $\lambda \left|| \beta \right||_1 > \sum_{i=1}^{n} Y_n (X_n \beta)$. Moreover, we $\left|| X\beta \right||_2 > 0$.

Thus,

$$l(\beta) > \frac{1}{2} \left(\left|| Y \right||_2 + \left|| X\beta \right||_2 \right) > \frac{1}{2} \left|| Y \right||_2 = l(0_p).$$

End of proof.