Homework Week 3: **Ridge and Lasso regression (22pts)**

3.1 Simulation of the data (2pts)

Re-use the data from the homework week1

```
import numpy as np
import pandas as pd
import statsmodels.api as sm
import matplotlib.pyplot as plt
# from sklearn import linear_model
# from sklearn.metrics import mean_squared_error, r2_score
np.random.seed(6996)

# error term
epsilon_vec = np.random.normal(0,1,500).reshape(500,1)
# X_matrix or regressors or predictiors
X_mat = np.random.normal(0,2,size = (500,500))
# Slope
slope_vec = np.random.uniform(1,5,500)
# Simulate Ys
Y_mat = 1 + np.cumsum(X_mat * slope_vec,axis=1)[:,1:] + epsilon_vec
# each col of Y_mat representing one simulation vector: starting with 2 re
gressors, end with 500
print(Y_mat.shape)
```

3.2 Fitting linear models (5pts)

1) Fit linear model with the first 10 predictors. Store the result in the variable m10.

2) Fit linear model with 491 predictors. Store the result in the variable v490.

3.3 Ridge regression (5pts)

1) Apply ridge regression to the data with 10 predictors.

2) Separate the sample into `train` and `test`.

3) Select the best parameter λ using cross validation on train set.

4) Calculate mean squared prediction error for the best selected λ

5) Compare the mean squared prediction error of linear model

What you should observe:

Ridge regression did not select predictors. It is expected because we simulated all predictors to be significant.
Ridge regression made a small improvement to mean squared prediction error. This is consistent with expectation because it has one additional parameter.

Regularization is expected to reduce number of predictors when there are collinear (highly correlated) predictors.
Predictors in this example are not collinear.

3.4 Lasso regression (5pts)

1) Fit lasso regression to the first 10 predictors.

2) Fit the model to the entire data.

Lasso regression marginally improved the mean squared error relative to the linear model, but did worse than ridge regression.
It kept all 10 predictors and produced similar estimates of parameters.

3.5 Large number of significant predictors (5pts)

1) Apply lasso regression analysis to data with 490 predictors.

2) Note that there are no actual slopes close to zero, but lasso regression still pushes them to zero when $\lambda=0$.

3) Calculate mean squares prediction error for the best lambda.

4) Fit lasso regression model to the entire data.
Plot the set of true slopes used in simulation and mark slopes removed by lasso.

Lasso removed predictors seemingly randomly regardless of the value of slope.
Given the way the sample was simulated (independent predictors with slopes between 1 and 3) it would be more reasonable removing none or removing the predictors with smallest slopes.