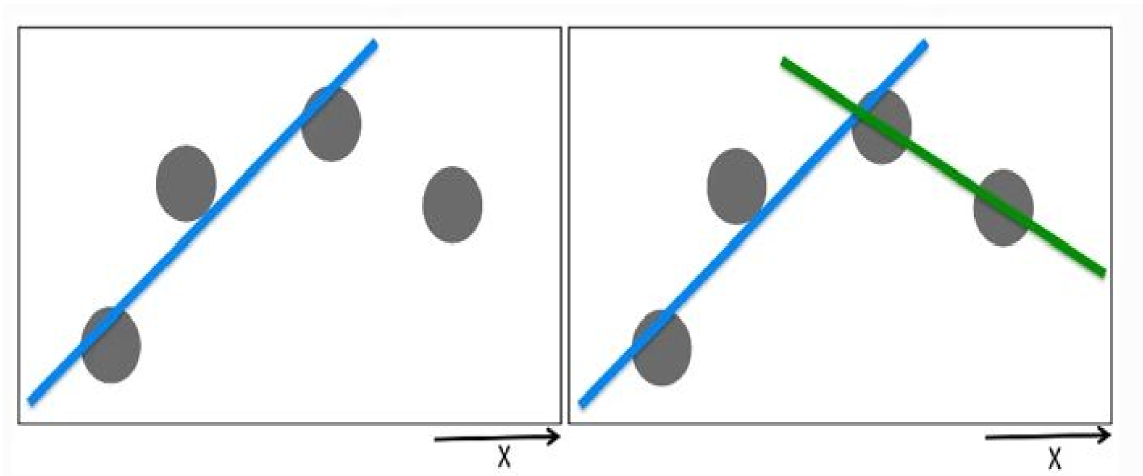
Assignment

In this assignment, we will apply F-test to detect whether there is a (statistically) significant change in the pricing behavior of your stock within some time interval. F-tests are often used to test equivalence of models that have been fitted to data usingthe least squares (such as linear regression). Some xamples include

1. testing whether regression fits the data well
2. testing for equality of means of normally distributed populations
3. testing whether two regression lines fits data better than onecolumn

We will focus on the last item. For each time period T (e.g. month), we want to check if there is a (statistically) significant change in pricing pattern for your stock.

We proceed as follows. Assume that the time period contains n days and let $P_1, \ldots, P_n$ denote the (adjusted closing) prices for days $i = 1, \ldots, n$. We construct a simple linear regression model for the price $P_i = \alpha \cdot i + \beta$. This model has two unknown parameters: slope α and intercept β. Therefore, for



this model, the number of degrees of freedom d=2. In general, if we have a linear regression on m variables, we would need to compute m slope coefficients and intercept - in this case $d = m + 1$. Let S$SE(T)$ denote the sum of the squared residuals ("loss" function) for the regression line that "fits" prices $P_1, \ldots, P_n$.

Next, we look for a day $1 < k < n$ where we suspect there is a change in linear trend. To find such a day, we divide our period T into two time periods: T_1 containing days $1, \ldots, k$ and T_2 containing days $k+1, \ldots, n$. Within each period, we construct two regressions and compute the corresponding loss functions S$SE(T_1)$ and S$SE(T_2)$. We look for k that minimizes the total loss from using two regressions S$SE(T_1)$+S$SE(T_2)$. Note that for each regression, the number of degrees of freedom is $d_1 = 2$

and $d_2 = 2$.

Once we computed our "break" day candidate k, we construct the following F statistics. To simplify the notation, let us define $L = \mathrm{S}SE(T)$, $L_1 = \mathrm{S}SE(T_1)$ and $L_2 = \mathrm{S}SE(T_2)$. For a single line, we have parameters to estimate, namely slope and intercept. For a single model, we need $d = 2$ parameters and for the 2-segment model we need the $d_1 + d_2 = 4$ parameters where $d_1 = d(L_1) = 2$ and $d_2 = d(L_2) = 2$ parameters to estimate. If there are n data points, we compute the following F statistics:

$$
\begin{aligned}
F &= \left[\frac{L - [L_1 + L_2)}{d - d_1}\right] \cdot \left[\frac{L_1 + L_2}{n - (d_1 + d_2)}\right]^{-1} \\
&= \left[\frac{L - [L_1 + L_2)}{2}\right] \cdot \left[\frac{L_1 + L_2}{n - 4}\right]^{-1}
\end{aligned}
$$

Under the null hypothesis that two regression lines do not provide a significantly better fit than one regression line, F will have an (Fisher) F-distribution with $(2, n-4)$ degrees of freedom. The null hypothesis is rejected if the F is greater than some critical value (e.g. 0.05)

In Python you can compute the F-distribution as follows

```
from scipy.stats import f as fisher_f
```

```
p_value = fisher_f.cdf(f_statistics,
                       2, n-4)
```

Questions:

1. take years 1 and 2. For each month, compute the "candidate" days and decide whether there is a significant change of pricing trend in each month. Use 0.1 as critical value.

2. how many months exhibit significant price changes for your sotck ticker.

3. are there more "changes" in year 1 or in year 2?