IE 555 – Programming for Analytics

Module #6 – OR Applications
Stock Forecasting

---

1 Stock Market Data

In this section we'll import data from an external source (e.g., a Website) and plot it in Python. Our focus will be on stock market data, although there are myriad other types of data that might be of interest to you.

See `stock_prices_quandl.py` for some code to get you started.

2 A (very) Brief Review of Forecasting

We'll apply three common/basic forecasting techniques to our stock market data.

2.1 Moving Average

For an n-period moving average forecast, simply calculate the arithmetic average of the n most recent observations.

$$F_t = \frac{D_{t-1} + D_{t-2} + \ldots + D_{t-n}}{n},$$

where F_t is the forecast for period t (e.g., tomorrow), D_{t-1} is the observed value from today, D_{t-2} is the observed value from yesterday, ..., D_{t-n} is the observed value from n periods ago.

Your choice of n will affect the forecasted value. Increasing n will "smooth" the forecast; decreasing n will make it more responsive to trends.

Note that, for moving averages, $F_t = F_{t+1} = F_{t+2} = \ldots$. In other words, all future forecasts are the same (given a set of n observations). This "feature" is not present in the other two methods described below.

2.2 Linear Regression

- Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be n paired data points for 2 variables X and Y.
- y_i is the observed value of Y when x_i is the observed value of X.
- Y is the dependent variable, X is the independent variable.

- If there's a linear relationship between X and Y:

$$\hat{Y} = a + bX,$$

  where $\hat{Y}$ is the predicted value of Y.

- For forecasting demand, X typically corresponds to time; Y typically corresponds to demand.

Our regression model is:

$$\hat{Y} = a + bX$$

We need to find values for a and b that minimize the sum of squared errors.

$$\begin{aligned} S_{xy} &= n\sum_{i=1}^{n} iD_i - \frac{n(n+1)}{2}\sum_{i=1}^{n} D_i, \\ S_{xx} &= \frac{n^2(n+1)(2n+1)}{6} - \frac{n^2(n+1)^2}{4}, \text{and} \\ \bar{D} &= \frac{1}{n}\sum_{i=1}^{n} D_i. \end{aligned}$$

The optimal values of a and b are:

$$\begin{aligned} b &= \frac{S_{xy}}{S_{xx}} \\ a &= \bar{D} - \frac{b(n+1)}{2} \end{aligned}$$

What is our forecast for some time t (in the future)?

$$\hat{D}_t = a + bt$$

2.3 Holt's Method (double exponential smoothing)

Intuition: We are trying to estimate a trend (i.e., a line). At each step we use the newest observation to improve our estimate of the actual slope and "intercept."

Interpret G_t as the slope at time t, and S_t as "what I would have forecast for period t in period $t-1$, had I known what I know now."

$$
\begin{aligned}
S_t &= \alpha D_t + (1-\alpha)(S_{t-1} + G_{t-1}) \\
G_t &= \beta(S_t - S_{t-1}) + (1-\beta)G_{t-1},
\end{aligned}
$$

for some smoothing constants α and β.

How do we get initial values for S_t and G_t (e.g., S_0 and G_0)? Typically:

- Use a subset of the data as a baseline.
- Use regression analysis to find estimates of the slope and intercept values using the baseline data.

The τ-step-ahead forecast using Holt's method is:

$$F_{t,t+\tau} = S_t + \tau G_t$$