You do not need to submit this assignment.

Curve fitting

In class we discussed fitting a specific curve to a set of n data points $(x_1, y_1), \ldots, (x_n, y_n)$. To do this we used the least error squared method in which we defined the predicted value of y for a given x value as $y' = f(x; a, b)$. (Note that in this notation, y' is not the derivative of y, it is the predicted value of Y.) The error associated with using y' instead of the true y value is $e_i = y'_i - y_i$. If we square each error value and add them over all x values, we have a representative measure for how well the predicting curve fits the data. The goal is to find a and b that minimize the value of the sum of the errors squared. Thus we set $\frac{\partial}{\partial a}\sum_{i=1}^{n}(y'_i - y_i)^2 = 0$ and $\frac{\partial}{\partial b}\sum_{i=1}^{n}(y'_i - y_i)^2 = 0$ and solve the resulting set of linear simultaneous equations.

Straight Line If one selects a straight line $y' = a + bx$, the resulting set of simultaneous equations

$$\begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum y_i x_i \end{pmatrix} \tag{1}$$

leads to the desired values of a and b. To simplify the notation, the limits on the summations in the above matrices are not written, but it is understood that each summation is taken over the number of data points, n. For example, the term $\sum x_i^2$ really means $\sum_{i=1}^{n}(x_i)^2$.

Quadratic Curve Finally, one might also try fitting a quadratic equation $y' = a + bx + cx^2$ to the data points. This form can be directly used in a linear regression, however one will need to solve a 3×3 system of linear equations to find the coefficients. The result is

$$\begin{pmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum y_i x_i \\ \sum y_i x_i^2 \end{pmatrix}$$

where the summation is done over all the n data pairs.

Power Curve Rather than fitting a straight line to some data, one might try a power curve $y' = ax^b$. While this form cannot directly be used in a linear regression, it may be transformed to a form that is amenable to linear regression by taking the natural log of each side. This leads to $\ln y' = \ln a + b \ln x$. If one defines the new variables $Y' = \ln y$, $A = \ln a$, and $X = \ln x$, the transformed version of the predicting equation, $Y' = A + bX$, is perfect for use in (1). Just remember that the coefficients that result from solving (1) yield the values of A and b. Hence you will need to calculate $a = e^A$ to get the final value of a.

Exponential Curve One might also try an exponential curve $y' = ab^x$. Again, while this form cannot directly be used in a linear regression, it may be transformed by taking the natural log of each side. This leads to $\ln y' = \ln(a) + \ln(b)x$. If one defines the new variables $Y' = \ln y$, $A = \ln a$, and $B = \ln b$, the transformed version of the predicting equation, $Y' = A + Bx$, can be used in (1). In this case, remember that the coefficients that result from solving (1) yield the values of A and B, so you will need to calculate $a = e^A$ and $b = e^B$to get the final values of a and b.

1. Write out the resulting system of equations if you used $y' = ax + bx^3 + cx^4$ to predict y for a given x value.
2. Develop two predicting equations $y' = f(x; a, b)$ that would require a transformation, as well as a newly defined set of parameters and/or variables (such as A, B, or perhaps X, Y). Clearly show the necessary transformations, the newly defined parameters and/or variables, and most importantly, the resulting systems of linear equations.