Math E-21b Supplement on Least Squares Approximation in Economics

You may already be familiar with the Method of Least Squares from statistics or multivariable calculus.
In Math E-21a, for example, you may have seen this as an exercise in unconstrained optimization. Given a collection of data $\{(x_i, y_i);\ 1 \le i \le n\}$, we seek the best-fitting (regression) line $y = mx + b$ for this data. By comparing the actual values y_i with those y-values that would be on the line, we seek to minimize the sum of the squares of these errors. That is we seek to minimize the function:

$$f(m,b) = \sum_{i=1}^{n} (mx_i + b - y_i)^2$$

A straightforward calculation yields that the partial derivatives of f will vanish when

$$\left(\sum_{i=1}^{n} x_i^2\right) m + \left(\sum_{i=1}^{n} x_i\right) b = \left(\sum_{i=1}^{n} x_i y_i\right)$$
$$\left(\sum_{i=1}^{n} x_i\right) m \quad + \quad n\, b \quad = \left(\sum_{i=1}^{n} y_i\right)$$

If we express these two equations in the two unknowns m and b, we get:

$$\begin{bmatrix} \left(\sum_{i=1}^{n} x_i^2\right) & \left(\sum_{i=1}^{n} x_i\right) \\ \left(\sum_{i=1}^{n} x_i\right) & n \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} \left(\sum_{i=1}^{n} x_i y_i\right) \\ \left(\sum_{i=1}^{n} y_i\right) \end{bmatrix}$$

You may verify that if we were to define the matrices

$$\mathbf{A} = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix},\ \mathbf{c} = \begin{bmatrix} m \\ b \end{bmatrix},\ \text{and}\ \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix},$$

then the previous matrix equation would simply be $[\mathbf{A}^T\mathbf{A}]\mathbf{c} = \mathbf{A}^T\mathbf{y}$, and this coincides with what we would have obtained using the Method of Least Squares in Linear Algebra.

The Cobb-Douglas Production Function

The analysis above is fine for finding best-fitting lines for a collection of data given as ordered pairs of numbers, but it is not applicable to situations where three or more variables appear to be affinely related. Such is the case with the data used a century ago to derive the Cobb-Douglas model for production that you may have seen in economics.

A production function is a function that relates various inputs such as capital and labor to the production level of an industry or a national economy. In practice, one models such a production function from available data collected over years of sampling. As with any model, a production function comes with limitations but can be quite useful in prediction and planning. As a practical matter we have to decide, out of all possible factors that affect production, which ones are significant enough to be included in our model. The Cobb-Douglas model relates the number of units of labor and capital to the production levels, all normalized to a standard base year.

The data upon which the Cobb-Douglas model is based comes from measures for the United States national economy between the years 1899 and 1922. With 1899 as the base year, and a standard level of 100 units for labor, capital, and production for that year, we have the following actual data:

Year	L	K	P	$x = \ln L$	$y = \ln K$	$z = \ln P$
1899	100	100	100	4.6052	4.6052	4.6052
1900	105	107	101	4.6540	4.6728	4.6151
1901	110	114	112	4.7005	4.7362	4.7185
1902	117	122	122	4.7622	4.8040	4.8040
1903	122	131	124	4.8040	4.8752	4.8203
1904	121	138	122	4.7958	4.9273	4.8040
1905	125	149	143	4.8283	5.0039	4.9628
1906	134	163	152	4.8978	5.0938	5.0239
1907	140	176	151	4.9416	5.1705	5.0173
1908	123	185	126	4.8122	5.2204	4.8363
1909	143	198	155	4.9628	5.2883	5.0434
1910	147	208	159	4.9904	5.3375	5.0689
1911	148	216	153	4.9972	5.3753	5.0304
1912	155	226	177	5.0434	5.4205	5.1761
1913	156	236	184	5.0499	5.4638	5.2149
1914	152	244	169	5.0239	5.4972	5.1299
1915	156	266	189	5.0499	5.5835	5.2417
1916	183	298	225	5.2095	5.6971	5.4161
1917	198	335	227	5.2883	5.8141	5.4250
1918	201	366	223	5.3033	5.9026	5.4072
1919	196	387	218	5.2781	5.9584	5.3845
1920	194	407	231	5.2679	6.0088	5.4424
1921	146	417	179	4.9836	6.0331	5.1874
1922	161	431	240	5.0814	6.0661	5.4806

Based on some reasonable economic assumptions, we surmise that production can be modeled by a production function of the form:

$$P(L,K) = bL^{\alpha}K^{\beta}$$

We must figure out which constants b, α, and β will best fit the data.

If $\alpha+\beta = 1$, we say that P has **constant returns to scale**.

If $\alpha+\beta < 1$, we say that P has **decreasing returns to scale**.

If $\alpha+\beta > 1$, we say that P has **increasing returns to scale**.

Which of these situations characterized the U.S. economy from 1899 to 1922 can only be ascertained when we try to fit the model to the data in the best way. If we wish to use the method of least squares, we must first find a linear relationship between the unknown parameters. To do this, we take the natural logarithm of both sides of the relation $P(L,K)=bL^{\alpha}K^{\beta}$ and obtain:

$$\ln(P) = \ln(b) + \alpha\ln(L) + \beta\ln(K)$$

If we let $z = \ln(P)$, $x = \ln(L)$, and $y = \ln(K)$, and call $a = \ln(b)$, then we have the following relation between among the as-yet-to-be-determined numbers a, α, and β:

$$z = a + \alpha x + \beta y$$

We can then use least squares approximation to find which choices of a, α, and β best fit the data.

We write $\mathbf{A} = [\mathbf{1}\ \ \mathbf{x}\ \ \mathbf{y}]$, where "**1**" represents a column vector consisting of all 1's, $\boldsymbol{x}$ represents a column vector consisting of the $x = \ln(L)$ data We let $\boldsymbol{y}$ represent a column vector consisting of the $y = \ln(K)$ data. We also let $\boldsymbol{z}$ represent a column vector consisting of the $z = \ln(P)$ data.

$$\mathbf{A} = \begin{bmatrix} 1 & 4.6052 & 4.6052 \\ 1 & 4.6540 & 4.6728 \\ 1 & 4.7005 & 4.7362 \\ 1 & 4.7622 & 4.8040 \\ 1 & 4.8040 & 4.8752 \\ 1 & 4.7958 & 4.9273 \\ 1 & 4.8283 & 5.0039 \\ 1 & 4.8978 & 5.0938 \\ 1 & 4.9416 & 5.1705 \\ 1 & 4.8122 & 5.2204 \\ 1 & 4.9628 & 5.2883 \\ 1 & 4.9904 & 5.3375 \\ 1 & 4.9972 & 5.3753 \\ 1 & 5.0434 & 5.4205 \\ 1 & 5.0499 & 5.4638 \\ 1 & 5.0239 & 5.4972 \\ 1 & 5.0499 & 5.5835 \\ 1 & 5.2095 & 5.6971 \\ 1 & 5.2883 & 5.8141 \\ 1 & 5.3033 & 5.9026 \\ 1 & 5.2781 & 5.9584 \\ 1 & 5.2679 & 6.0088 \\ 1 & 4.9836 & 6.0331 \\ 1 & 5.0814 & 6.0661 \end{bmatrix}$$

The inconsistent system of linear equations associated with this data is then $\mathbf{A}\begin{bmatrix} a \\ \alpha \\ \beta \end{bmatrix} = \boldsymbol{z}$, and the normal equation is:

$$\mathbf{A}^{\mathrm{T}}\mathbf{A}\begin{bmatrix} a \\ \alpha \\ \beta \end{bmatrix} = \mathbf{A}^{\mathrm{T}}\mathbf{z}.$$

Write $\mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{B}$ and $\mathbf{A}^{\mathrm{T}}\mathbf{z} = \mathbf{D}$. If we write the augmented matrix associated with the normal equation as $\mathbf{C} = [\mathbf{B} \mid \mathbf{D}]$, then the least squares solution is given by $\mathbf{S} = \mathrm{rref}(\mathbf{C})$ and is obtained by row reduction.

$$\mathbf{B} = \mathbf{A}^{\mathrm{T}}\mathbf{A} = \begin{bmatrix} 24.0000 & 119.3311 & 128.5556 \\ 119.3311 & 594.2757 & 641.1346 \\ 128.5556 & 641.1346 & 693.4555 \end{bmatrix}$$

$$\mathbf{D} = \mathbf{A}^{\mathrm{T}}\mathbf{z} = \begin{bmatrix} 121.8561 \\ 607.0907 \\ 655.4095 \end{bmatrix}$$

$$\mathbf{C} = \left[\begin{array}{ccc|c} 24.0000 & 119.3311 & 128.5556 & 121.8561 \\ 119.3311 & 594.2757 & 641.1346 & 607.0907 \\ 128.5556 & 641.1346 & 693.4555 & 655.4095 \end{array}\right]$$

$$\mathbf{S} = \left[\begin{array}{ccc|c} 1 & 0 & 0 & -0.0692 \\ 0 & 1 & 0 & 0.7689 \\ 0 & 0 & 1 & 0.2471 \end{array}\right]$$

Thus, $a = -.0692$, $\alpha = .7689$, and $\beta = .2471$. We then calculate $b = \exp(a) = .9331$

Note that $\alpha + \beta = 1.016$, hence P has increasing returns to scale, but has very nearly constant returns to scale. For this reason, it is often assumed in economics that $\alpha + \beta = 1$ in the Cobb-Douglas model.