MthStat 465, Spring 2005, Lecture Number 5
General Least Squares

1. Least Squares in General

In the method of least squares, a numerical measure of how well a curve fits the data is given. A curve is then chosen to minimize this measure. Specifically, if $y = f(x)$ is proposed as a fit to the data, then

$$E(f) := \sum_{k=1}^{N} (f(x_k) - y_k)^2$$

measures how well $y = f(x)$ fits the data: the bigger $E(f)$ is, the worse the fit. $E(f)$ is called the sum of squared errors. The goal of the method of least squares is to choose f so that $E(f)$, the **sum of squared errors**, is as small as possible. This minimum value is called the **residual sum of squares**. Usually there is a restriction on the choices for f. The most common is that the graph of $y = f(x)$ must be a line, that is, $f(x) = mx + b$. In this case

$$E(f) = \sum_{k=1}^{N} (y_k - (mx_k + b))^2$$

may be regarded as a function of two variables, m and b, with $-\infty < m < \infty$ and $-\infty < b < \infty$. The function

$$Q(m, b) := \sum_{k=1}^{N} (y_k - (mx_k + b))^2$$

is a quadratic function of two variables and may be minimized by a variety of means:

1. Repeated completion of squares;
2. Calculus: compute the gradient of Q and find where it is equal to the zero vector;
3. Vector projection.

The same can be said for solving the least squares problem with other types of functions. For example, we can try to fit quadratic functions to the data. This means we should consider functions of the form $f(x) = A + Bx + Cx^2$, where A, B and C are allowed to vary over all real numbers. In this case we want to minimize

$$Q(A, B, C) := \sum_{k=1}^{N} (A + Bx_k + Cx_k^2 - y_k)^2$$

as a function of A, B and C. The same methods apply here as in the linear case.

More generally, if we are given M functions, $f_1(x)$, $f_2(x)$, ..., $f_M(x)$ we can try to fit a curve of the form $f(x) = A_1 f_1(x) + \cdots + A_M f_M(x)$ by minimizing

$$Q(A_1, A_2, \ldots, A_M) := \sum_{k=1}^{N} (A_1 f_1(x_k) + \cdots A_M f_M(x_k) - y_k)^2$$

as a function of the A_j's.

In the case of linear regression, $M = 2$, $f_1(x) = 1$, and $f_2(x) = x$. In the case of quadratic regression, $M = 3$, $f_1(x) = 1$, $f_2(x) = x$ and $f_3(x)x^2$.

We are not limited to polynomials. We could have $f_1(x) = 1$, $f_2(x) = \cos(x)$ and $f_3(x) = \sin(x)$.

2. Correlation coefficient

Crudely, the sign of the quantity

$$\sum_{k=1}^{N}(x_k - \overline{x})(y_k - \overline{y})$$

meaures if there is a positive or negative trend in the data. Indeed, the sign of this quantity is the same as the sign of the slope of the least squares regression line. To standardize this measurement, we use

$$r := \frac{\sum_{k=1}^{N}(x_k - \overline{x})(y_k - \overline{y})}{\sqrt{\sum_{k=1}^{N}(x_k - \overline{x})^2 \sum_{k=1}^{N}(y_k - \overline{y})^2}},$$

the **correlation coefficient**. By considering vectors and dot product, it is easy to see that $-1 \leq r \leq 1$. If $|r| \approx 1$ then the least squares line predicts y from x very well. If $|r| \approx 0$ then the predictions will not be very good at all. We will quantify this more later when we discuss the general linear model.