Math 141, 511, 516, 517, Spring 2010

Lecture 23

04/27/2010

The last day when you can make up quizzes or tests is Tuesday, May 4.

The last Math 142 help session of this semester will be held on Monday, May 3.

Section 8-5. **Method of least squares.**

Regression analysis is the process of fitting an elementary function to set of data points by the **method of least squares.** We begin with **linear regression**, the process of finding the equation of the line that is the "best" approximation to set of data points.
We want to approximate the cost function by a linear function

$$y = ax + b,$$

where a and b are reals to be determined.
For a set of n points (x_1, y_1), (x_2, y_2),...,(x_n, y_n), we constract a table

x	y	$ax + b$	**Residual**
x_1	y_1	$ax_1 + b$	$y_1 - (ax_1 + b)$
x_2	y_2	$ax_2 + b$	$y_2 - (ax_2 + b)$
...	...	...	...
x_n	y_n	$ax_n + b$	$y_n - (ax_n + b)$

The difference $y_k - (ax_k + b)$ for $k = 1, ..., n$ is called the **residual**.

Our criteria of the "best" approximation is the following: Determine the values of a and b that *minimize the sum of the squares* of the residuals. The resulting line is called the **least squares line**, or the **regression line.** We minimize

$$F(a,b) = [y_1 - (ax_1 + b)]^2 + [y_2 - (ax_2 + b)]^2 + ... + [y_n - (ax_n + b)]^2.$$

Step 1. Find critical points ($F_a(a,b) = 0, F_b(a,b) = 0$).
Step 2. Compute $A = F_{aa}(a,b)$, $B = F_{ab}(a,b)$, and $C = F_{bb}(a,b)$.
Step 3. Evaluate $AC - B^2$ and try to classify the critical point (a,b) by using the second derivative test.

Example 2. Find the least squares line and use it to estimate y for the indicated value of x. Round answers to two decimal places.

x	1	2	2	3
y	3	1	2	0

Estimate y when $x = 2.5$.

Least squares approximation. For a set of points (x_1, y_1), (x_2, y_2),...,(x_n, y_n), the coefficients of the least squares line $y = ax + b$ are the solutions of the system of **normal equations**

$$\left(\sum_{k=1}^{n} x_k^2\right) a + \left(\sum_{k=1}^{n} x_k\right) b = \sum_{k=1}^{n} x_k y_k$$
$$\left(\sum_{k=1}^{n} x_k\right) a + nb = \sum_{k=1}^{n} y_k$$

and are given by the formulas

$$a = \frac{n\left(\sum_{k=1}^{n} x_k y_k\right) - \left(\sum_{k=1}^{n} x_k\right)\left(\sum_{k=1}^{n} y_k\right)}{n\left(\sum_{k=1}^{n} x_k^2\right) - \left(\sum_{k=1}^{n} x_k\right)^2}$$
$$b = \frac{\sum_{k=1}^{n} y_k - a\left(\sum_{k=1}^{n} x_k\right)}{n}$$

The method of least squares can also be applied to find the quadratic equation

$$y = ax^2 + bx + c$$

that best fits a set of a data points.