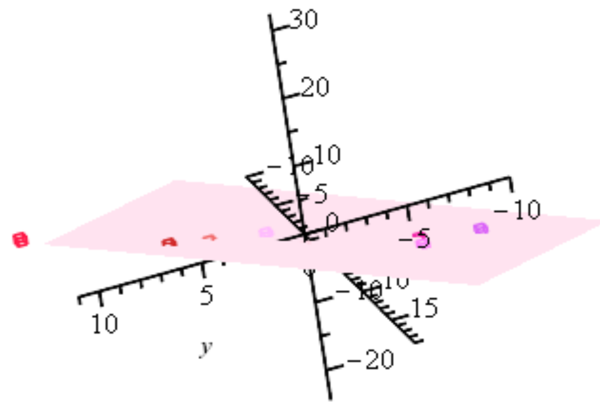

Figure 55.6: The graph of the plane $f(x,y) = 1.1706x + 1.1401y - 1.2298$ as an approximate fit to the data (X, y_1) of Example 55.1.

See Figure 55.6. We can see how the choice of K affects the quality of the solution (w, b) by computing the norm $\|\xi\|_2$ of the error vector $\xi = y - Xw - b\mathbf{1}_m$. As in Example 55.1 we notice that the smaller K is, the smaller is this norm. We also observe that for a given value of K, Program (**RR6′**) gives a slightly smaller value of $\|\xi\|_2$ than (**RR3**b) does.

As pointed out by Hastie, Tibshirani, and Friedman [88] (Section 3.4), a defect of the approach where b is also penalized is that the solution for b is not invariant under adding a constant c to each value y_i. This is not the case for the approach using Program (**RR6′**).

55.3 Kernel Ridge Regression

One interesting aspect of the dual (of either (**RR2**) or (**RR3**)) is that it shows that the solution w being of the form $X^\top\alpha$, is a linear combination

$$w = \sum_{i=1}^{m} \alpha_i x_i$$

of the data points x_i, with the coefficients α_i corresponding to the dual variable $\lambda = 2K\alpha$ of the dual function, and with

$$\alpha = (XX^\top + KI_m)^{-1}y.$$

If m is smaller than n, then it is more advantageous to solve for α. But what really makes the dual interesting is that with our definition of X as

$$X = \begin{pmatrix} x_1^\top \\ \vdots \\ x_m^\top \end{pmatrix},$$