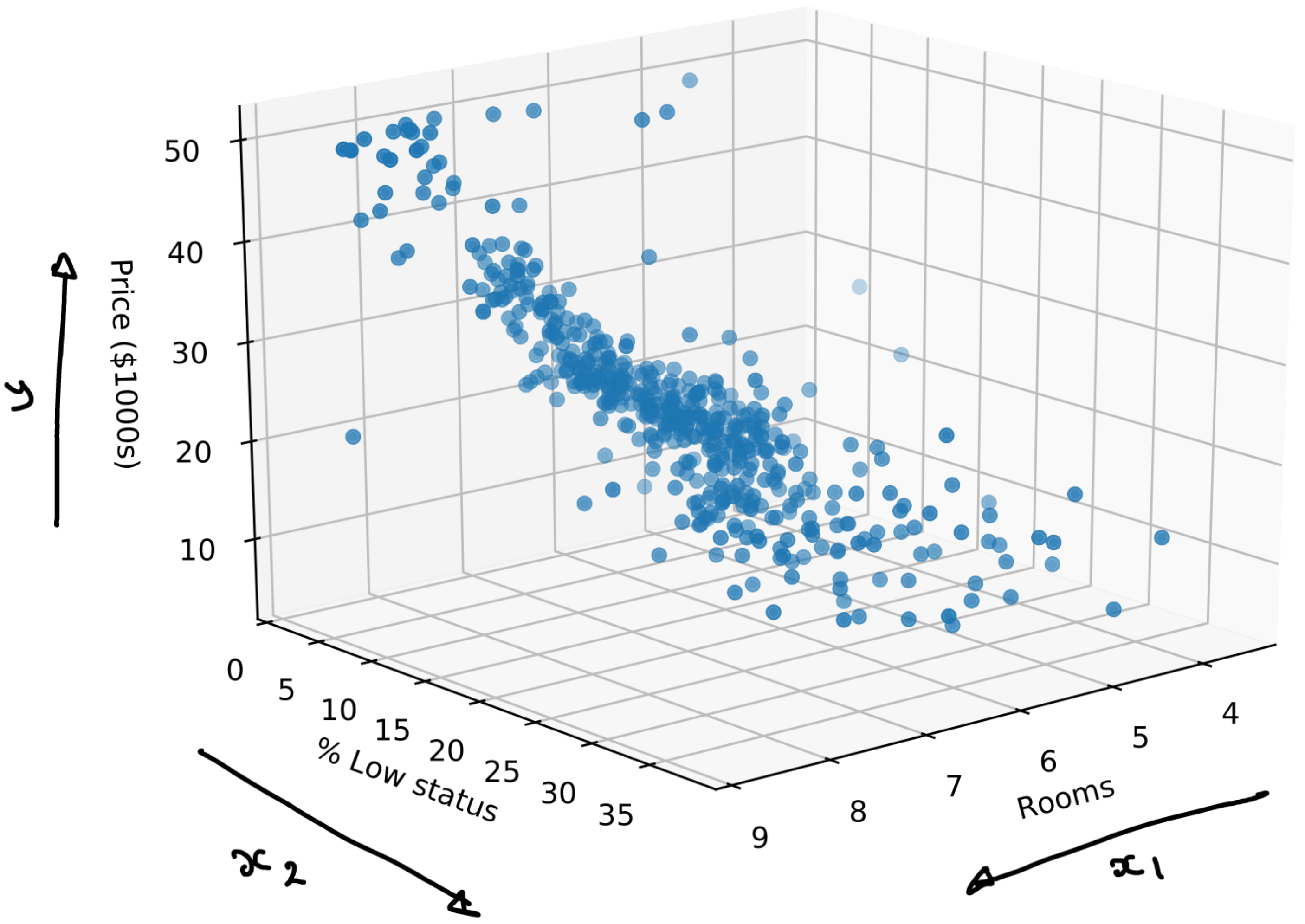
Multiple linear regression

Model and loss

Herman Kamper

http://www.kamperh.com/

Boston house prices



Multiple linear regression

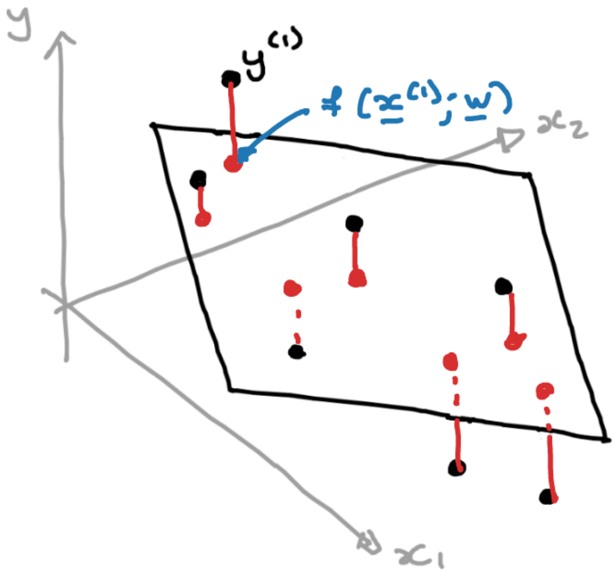
The model:

$$f(x_1, x_2, \ldots, x_D; w_0, w_1, w_2, \ldots, w_D)$$
$$= \underbrace{w_0}_{\text{"bias"}} + w_1 x_1 + w_2 x_2 + \ldots + w_D x_D$$

$$f(\underline{x}; \underline{w}) = \underline{w}^T \underline{x}$$ [Pretending $x_0 = 1$]

with

$$\underline{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{bmatrix} \quad \text{and} \quad \underline{x} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix}$$



Loss function:

Squared loss:

$$J(\underline{w}) = \sum_{n=1}^{N} \left(y^{(n)} - f(\underline{x}^{(n)}; \underline{w})\right)^2$$
$$= \sum_{n=1}^{N} \left(y^{(n)} - \left(w_0 + w_1 x_1^{(n)} + w_2 x_2^{(n)} + \ldots + w_D x_D^{(n)}\right)\right)^2$$

Optimization:

Derive $\frac{\partial J}{\partial w_0}, \frac{\partial J}{\partial w_1}, \ldots, \frac{\partial J}{\partial w_D}$ and set equal to zero.

Painful to do term-by-term.

Idea: Rather write in vector form and find $\frac{\partial J}{\partial \underline{w}}$.

Interlude: Watch vector and matrix derivatives.

Writing the loss in matrix form

We want to minimize:

$$J(\underline{w}) = \sum_{n=1}^{N} \left(y^{(n)} - f(\underline{x}^{(n)}; \underline{w})\right)^2$$

$$= \sum_{n=1}^{N} \left(y^{(n)} - \underline{w}^T \underline{x}^{(n)}\right)^2 \quad \ldots \text{1}$$

"design matrix"

Define:

$$\underline{X} = \begin{bmatrix} — (\underline{x}^{(1)})^T — \\ — (\underline{x}^{(2)})^T — \\ \vdots \\ — (\underline{x}^{(N)})^T — \end{bmatrix} ; \quad \underline{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{bmatrix}$$

Now we can write 1 as:

$$J(\underline{w}) = (\underline{y} - \underline{X}\underline{w})^T (\underline{y} - \underline{X}\underline{w}) \quad \ldots \text{2}$$

To see this, you can define an error vector:

$$\underline{e} = \begin{bmatrix} y^{(1)} - \underline{w}^T \underline{x}^{(1)} \\ y^{(2)} - \underline{w}^T \underline{x}^{(2)} \\ \vdots \\ y^{(N)} - \underline{w}^T \underline{x}^{(N)} \end{bmatrix}$$

Then note that 1 can be written as $J(\underline{w}) = \underline{e}^T \cdot \underline{e}$, and $\underline{e} = \underline{y} - \underline{X}\underline{w}$, which leads to 2. We now use the form in 2 to determine $\frac{\partial J}{\partial \underline{w}}$.

$$J(\underline{w}) = \underline{y}^T \underline{y} - \underbrace{\overset{1\times N}{\underline{y}^T} \overset{N\times D}{\underline{X}} \overset{D\times 1}{\underline{w}}}_{1\times 1} - \underline{w}^T \underline{X}^T \underline{y} + \underline{w}^T \underline{X}^T \underline{X}\underline{w}$$

$$= \underline{y}^T \underline{y} - \underline{w}^T \underline{X}^T \underline{y} - \underline{w}^T \underline{X}^T \underline{y} + \underline{w}^T \underline{X}^T \underline{X}\underline{w}$$

$$= \underline{y}^T \underline{y} - 2\underline{w}^T \underline{X}^T \underline{y} + \underline{w}^T \underline{X}^T \underline{X}\underline{w}$$

Multiple linear regression

Optimization

Herman Kamper

http://www.kamperh.com/

The normal equations

Now we set $\frac{\partial J}{\partial \underline{w}} = \underline{0}$, i.e. $\frac{\partial J}{\partial w_0} = 0, \ldots, \frac{\partial J}{\partial w_D} = 0$

$$\frac{\partial J}{\partial \underline{w}} = \frac{\partial}{\partial \underline{w}} \left[\underline{y}^T \underline{y} - 2 \underline{w}^T \underline{X}^T \underline{y} + \underline{w}^T \underline{X}^T \underline{X} \underline{w} \right]$$

$$= -2 \underline{X}^T \underline{y} + \left(\underline{X}^T \underline{X} + (\underline{X}^T \underline{X})^T \right) \underline{w} \quad \ldots (3)$$

$$= -2 \underline{X}^T \underline{y} + 2 \underline{X}^T \underline{X} \underline{w}$$

Set $\frac{\partial J}{\partial \underline{w}} = 0$, then:

$$\underline{X}^T \underline{X} \underline{w} = \underline{X}^T \underline{y}$$

$$\hat{\underline{w}} = (\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{y}$$

This is called the normal equations. With one line of Python, can get the estimates of D+1 parameters.

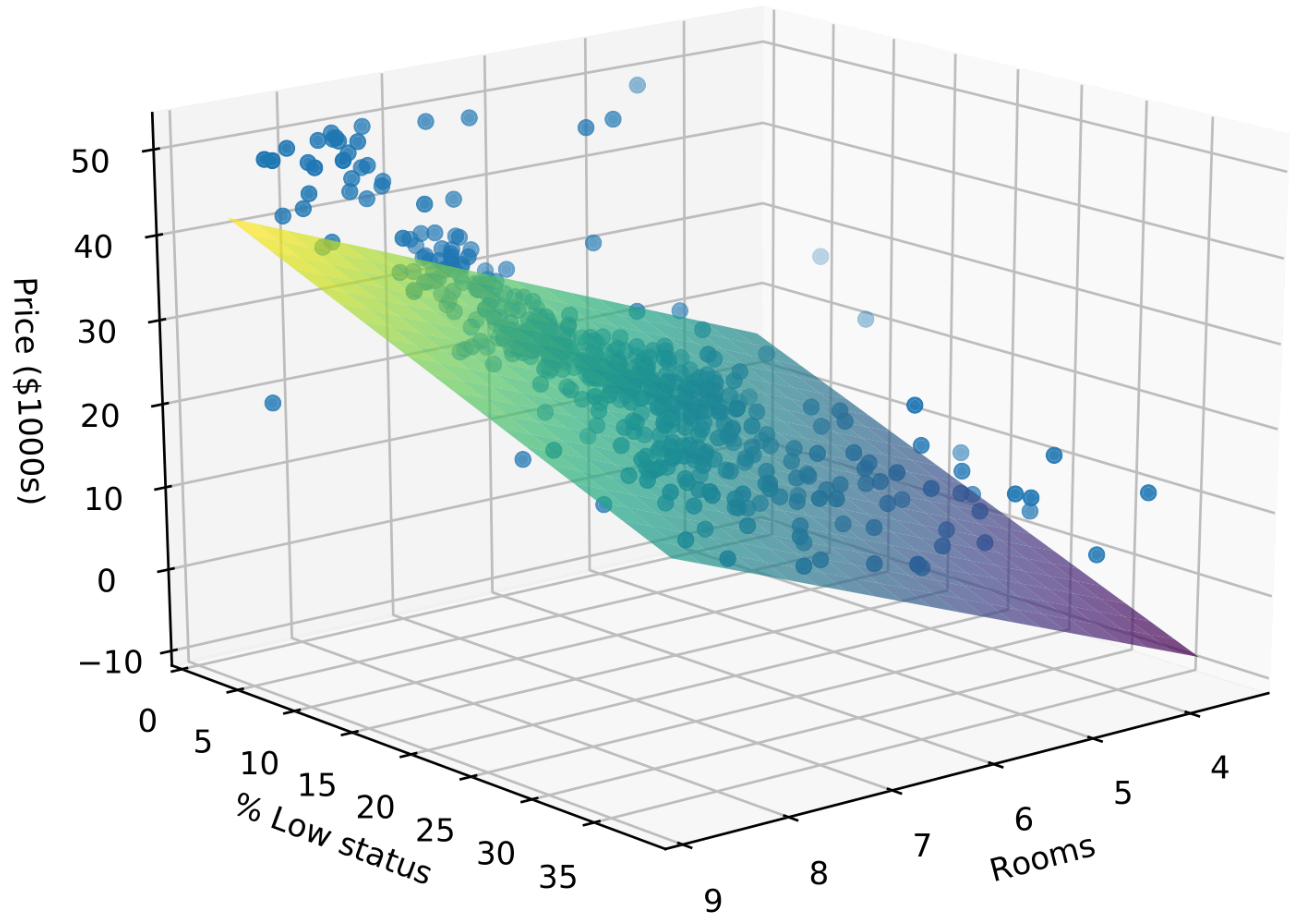
To get to (3), we used the following identities:

$$\frac{\partial \underline{x}^T \underline{A} \underline{x}}{\partial \underline{x}} = (\underline{A} + \underline{A}^T) \underline{x}$$

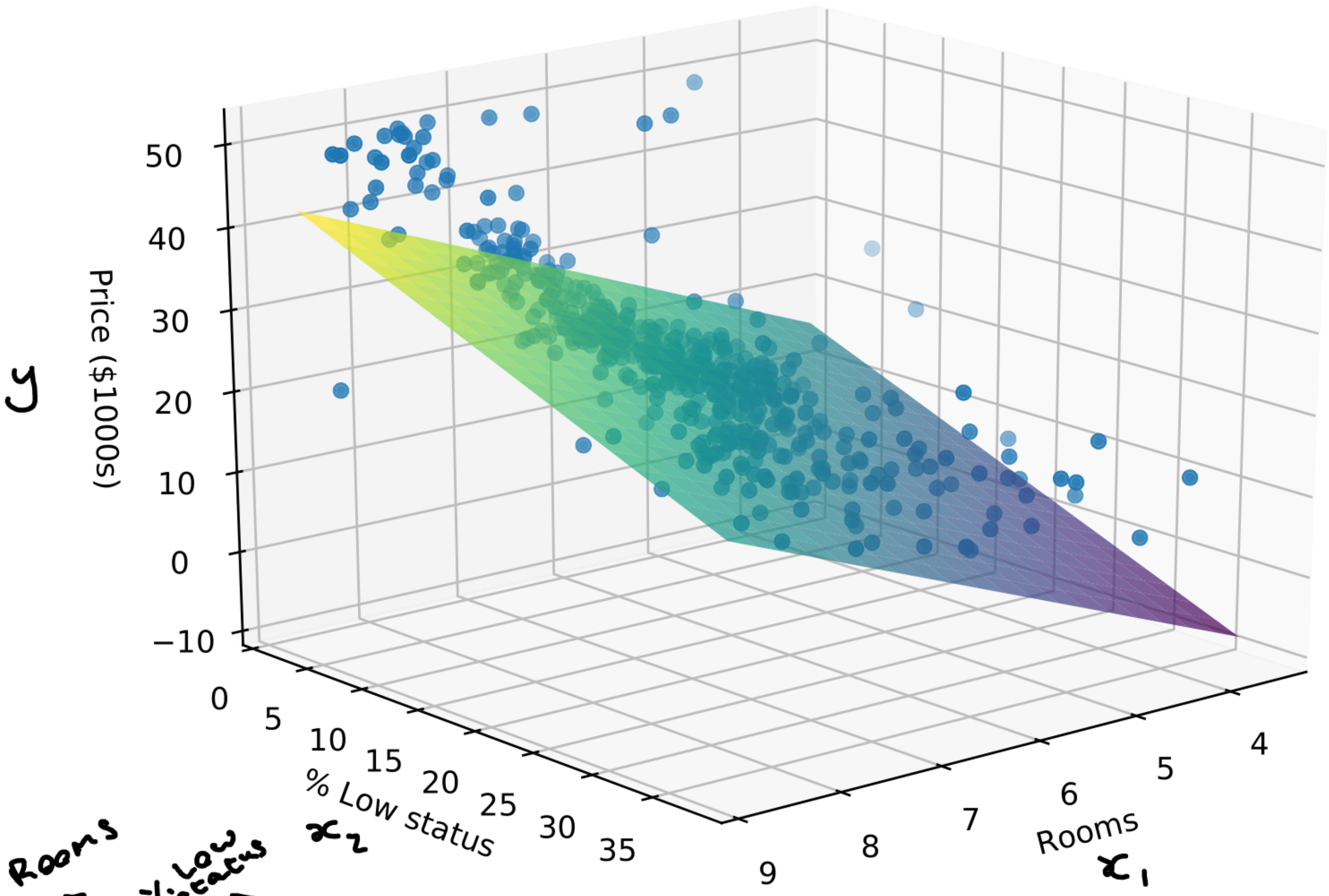
$$\frac{\partial \underline{x}^T \underline{a}}{\partial \underline{x}} = \underline{a}$$

You can find these in the Matrix calculus Wikipedia article.

Boston house prices fit



Boston house prices fit



Rooms
% Low status

$$f(\mathbf{x}; \hat{\mathbf{w}}) = \hat{w}_0 + \hat{w}_1 x_1 + \hat{w}_2 x_2 = -1.358 + 5.095 x_1 - 0.642 x_2$$