

Example 44. (again) Determine the line that best fits the data points $(2,1), (5,2), (7,3), (8,3)$.

Solution. Let's repeat the computation we did last class. This time, we let Sage do the actual work for us:

```
Sage] X = matrix([[1,2],[1,5],[1,7],[1,8]]); y = vector([1,2,3,3])
```

```
Sage] (X.transpose()*X).solve_right(X.transpose()*y)
```

$$\left(\frac{2}{7}, \frac{5}{14}\right)$$

Here are some intermediate steps to help see what's going on (and that it matches our earlier work):

```
Sage] X.transpose()*X
```

$$\begin{bmatrix} 4 & 22 \\ 22 & 142 \end{bmatrix}$$

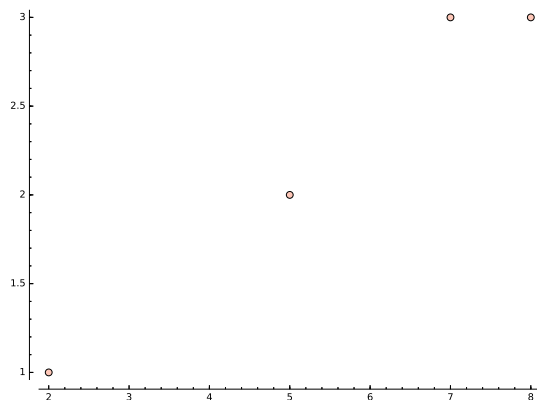
```
Sage] X.transpose()*y
```

$$(9, 57)$$

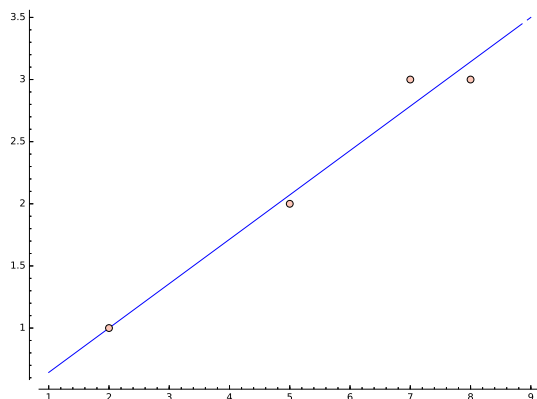
Let's plot the least squares line $y = \frac{2}{7} + \frac{5}{14}x$ in Sage to marvel at the good fit!

```
Sage] points = [[2,1],[5,2],[7,3],[8,3]]
```

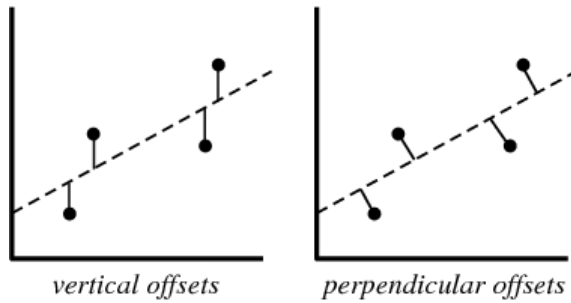
```
Sage] scatter_plot(points)
```



```
Sage] scatter_plot(points) + plot(2/7+5/14*x,1,9)
```



Comment. As usual in practice, we are minimizing the (sum of squares of the) vertical offsets:



<http://mathworld.wolfram.com/LeastSquaresFitting.html>

Comment. We get a (slightly) different “best fit” line if we change the role of x and y ! Can you explain that?

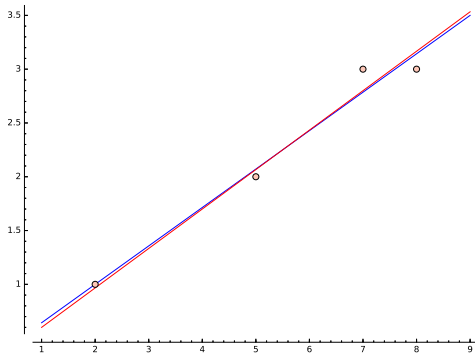
```
Sage] X = matrix([[1,1],[1,2],[1,3],[1,3]]); y = vector([2,5,7,8])
```

```
Sage] (X.transpose()*X).solve_right(X.transpose()*y)
```

$$\left(-\frac{7}{11}, \frac{30}{11}\right)$$

Note that $x = -\frac{7}{11} + \frac{30}{11}y$ is equivalent to $y = \frac{7}{30} + \frac{11}{30}x$.

```
Sage] scatter_plot([[2,1],[5,2],[7,3],[8,3]]) + plot(2/7+5/14*x,1,9) + plot(7/30+11/30*x,1,9,color='red')
```



The explanation is that (see pictures at the beginning of this example) we are minimizing vertical offsets in one case and horizontal offsets in the other case.

In linear regression, the relationship between a dependent variable and one or more explanatory variables is modeled. If y is the dependent variable, with x the explanatory variable, then it is natural to minimize the error we make in “predicting y through x ” (vertical offsets).

Example 45. A car rental company wants to predict the annual maintenance cost y (in 100USD/year) of a car using the age x (in years) of that car (as an explanatory variable). Based on the observations $(x, y) = (2, 1), (5, 2), (7, 3), (8, 3)$, predict the cost for a 4.5 year old car (using linear regression).

Solution. Once we compute the regression line $y = a + bx$ (we already did that: $y = \frac{2}{7} + \frac{5}{14}x$), our prediction is $\frac{2}{7} + \frac{5}{14} \cdot 4.5 = \frac{53}{28} \approx 1.89$, that is, 189 USD/year.

Application: Fitting data to other curves

We can also fit the experimental data (x_i, y_i) using other curves.

Example 46. Set up a linear system to find values for the parameters a, b, c that result in the quadratic curve $y = a + bx + cx^2$ that best fits some given points $(x_1, y_1), (x_2, y_2), \dots$

Solution. $y_i \approx a + bx_i + cx_i^2$ with parameters a, b, c .

The equations $y_i = a + bx_i + cx_i^2$ in matrix form:

$$\underbrace{\begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ \vdots & \vdots & \vdots \end{bmatrix}}_{\text{design matrix } A} \underbrace{\begin{bmatrix} a \\ b \\ c \end{bmatrix}} = \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \end{bmatrix}}_{\text{observation vector } \mathbf{y}}$$

Again, we determine values for a, b, c by computing a least squares solution to that system.

That is, we need to solve the system $A^T A \begin{bmatrix} a \\ b \\ c \end{bmatrix} = A^T \mathbf{y}$.

Example 47. (extra) Use Sage to find values for a, b, c that result in the quadratic curve $y = a + bx + cx^2$ that best fits the points $(0, 1), (1, 2), (2, 3), (3, -4), (4, -7), (5, -12)$.

Solution. We first input the points:

```
Sage] points = [[0,1], [1,2], [2,3], [3,-4], [4,-7], [5,-12]]
```

We set up the system described in the previous example, then determine a least-squares solution.

```
Sage] X = matrix([[1,0,0], [1,1,1], [1,2,4], [1,3,9], [1,4,16], [1,5,25]])
```

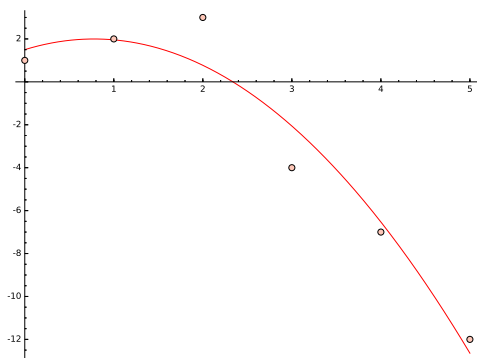
```
Sage] y = vector([1,2,3,-4,-7,-12])
```

```
Sage] (X.transpose()*X).solve_right(X.transpose()*y)
```

$$\left(\frac{3}{2}, \frac{179}{140}, -\frac{23}{28} \right)$$

Hence, the best fitting quadratic curve is $y = \frac{3}{2} + \frac{179}{140}x - \frac{23}{28}x^2$. Here's a plot:

```
Sage] scatter_plot(points) + plot(3/2+179/140*x-23/28*x^2,0,5,color='red')
```



Advanced comment. If you are comfortable with Python, you can avoid typing out X and \mathbf{y} :
[The plot command above now won't work anymore because we are overwriting x with numbers.]

```
Sage] X = matrix([[1,x,x^2] for x,y in points])
```

```
Sage] y = vector([y for x,y in points])
```