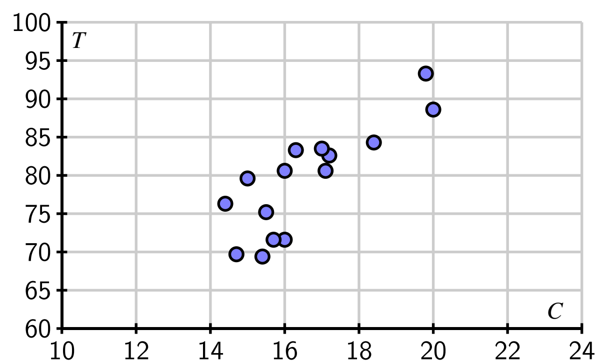
Mathematics 327
Least squares problems

Open the Jupyter notebook on `CoCalc` called `Linear_Regression` before beginning this activity.

1. The rate at which crickets chirp C is related to the outside temperature T as shown in the data below. The rate C is expressed in chirps per second while temperature T is in degrees Fahrenheit. The data is displayed graphically below and is also given in the Jupyter notebook. For now, call the data points (C_i, T_i).

C	20.0	16.0	19.8	18.4	17.1	15.5	14.7	15.7	15.4	16.3	15.0	17.2	16.0	17.0	14.4
T	88.6	71.6	93.3	84.3	80.6	75.2	69.7	71.6	69.4	83.3	79.6	82.6	80.6	83.5	76.3



We would like to model this with a linear function

$$T = \beta_0 + \beta_1 C.$$

Write a linear equation that relates C_1 and T_1.

Describe how you will set up a linear system $A\mathbf{x} = \mathbf{b}$ for the y-intercept β_0 and the slope β_1; that is, what is the form of the matrix A and vector $\mathbf{b}$?

The first cell of the Jupyter notebook defines some common functions we have been using recently along with a new function `ones(m)`, which will give you a vector consisting of `m` 1's. This cell will also give you two vectors, `chirps` and `temperature`. Use this to find the least squares approximate solution giving β_0 and β_1. What line is given by the least squares approximate solution?

If a cricket is chirping at 22 chirps per second, what is your prediction for the temperature?

2. Suppose we want to find the least squares approximate solution to the equation $A\mathbf{x} = \mathbf{b}$ and that we have a QR factorization $A = QR$. Explain why the least squares approximate solution is given by

$$A\widehat{\mathbf{x}} = QQ^T\mathbf{b}.$$

Since $A = QR$, we have $A\widehat{\mathbf{x}} = QR\widehat{\mathbf{x}} = QQ^T\mathbf{b}$. Multiply both sides of this equation by Q^T and explain why

$$R\widehat{\mathbf{x}} = Q^T\mathbf{b}.$$

Remember that R is upper triangular, which means that solving this equation is simple.

3. Brozaks formula, which is used to calculate a person's body fat index (BFI), is

$$BFI = 100\left(\frac{4.57}{\rho} - 4.142\right)$$

where ρ denotes a person's body density in grams per cubic centimeter. Obtaining an accurate measure of ρ is difficult, however, because it requires submerging the person in water and measuring the volume of water displaced. Instead, we can take several other body measurements and use it to predict BFI. For instance, we take 10 patients and measure their weight w in pounds, height h in inches, abdomen a in centimeters, wrist circumference r in centimeters, neck circumference n in centimeters, and BFI. We find that:

w	h	a	r	n	BFI
154	68	85	17	36	13
173	72	83	18	39	7
154	66	88	17	34	25
185	72	86	18	37	11
184	71	100	18	34	28
210	75	94	19	39	21
181	70	91	18	36	19
176	73	89	19	38	13
191	74	83	18	38	5
199	74	89	19	42	12

We would like to model the relationship between BFI and the five measurements using a linear function:

$$BFI = \beta_0 + \beta_1 w + \beta_2 h + \beta_3 a + \beta_4 r + \beta_5 n.$$

Using the first patient, write an equation for the unknowns $\beta_0, \beta_1, \ldots, \beta_5$.

Describe how you would set up a linear system $A\mathbf{x} = \mathbf{b}$ for these constants; that is, how will you form A and $\mathbf{b}$?

Use the Jupyter notebook to find the least squares approximate solution and hence the best constants $\beta_0, \beta_1, \ldots, \beta_5$.

Suppose a patient's measurements are $w = 190$, $h = 70$, $a = 90$, $r = 18$ and $n = 35$. Estimate this patient's BFI.