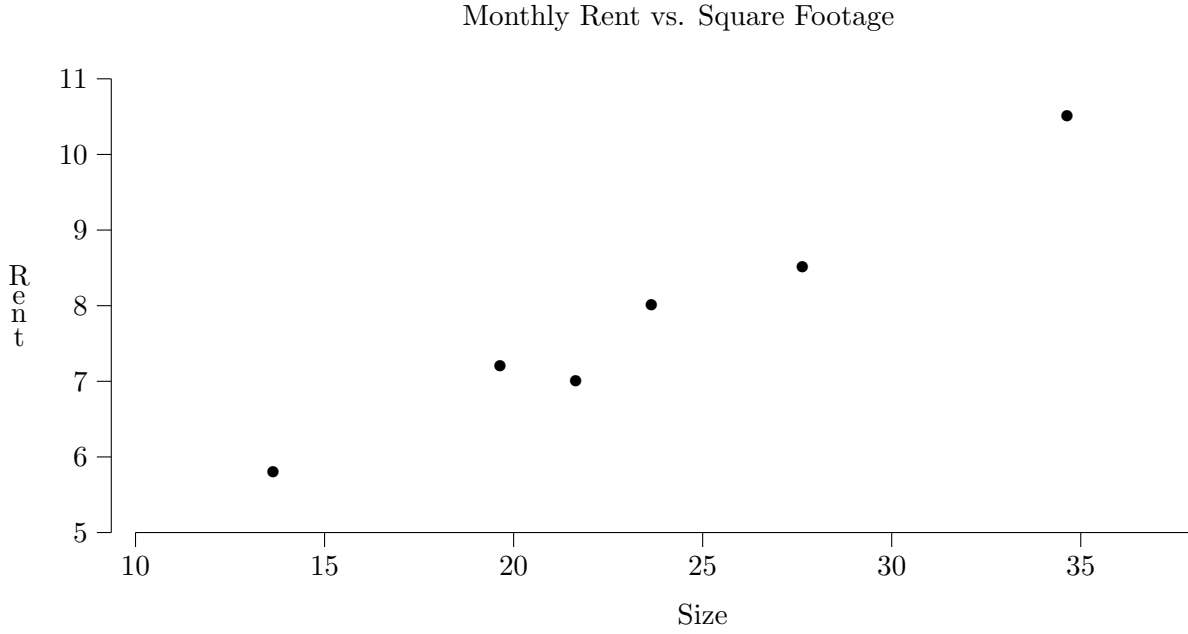
MATH 1070 Introductory Statistics
Annotated Example for Linear Regression

Problem statement

A consumer welfare agency wants to investigate the relationship between the sizes of houses and rents paid by tenants in a small city. The agency collected the following information on the sizes (in hundreds of square feet) of six houses and the monthly rents (in dollars) paid by tenants. [1]

Size of the house, x	21	13	19	27	34	23
Monthly rent, y	700	580	720	850	1050	800

Step 1: Construct a scatterplot and examine the relationship



Step 2: Compute means and standard deviations

Means

$$\text{Size } \bar{x} = \frac{21+13+19+27+34+23}{6} = 22.83$$

$$\text{Rent } \bar{y} = \frac{700+580+720+850+1050+800}{6} = 783.33$$

[1] *Introductory Statistics, 3rd ed*, Prem S. Mann, John Wiley and Sons publishers, p. 644, 1998

Standard deviations

$$\begin{aligned}\text{Size } s_x &= \sqrt{\frac{(21-22.83)^2+(13-22.83)^2+(19-22.83)^2+(27-22.83)^2+(34-22.83)^2+(23-22.83)^2}{6-1}} \\ &= 7.17\end{aligned}$$

$$\begin{aligned}\text{Rent } s_y &= \sqrt{\frac{(700-783.33)^2+(580-783.33)^2+(720-783.33)^2+(850-783.33)^2+(1050-783.33)^2+(800-783.33)^2}{6-1}} \\ &= 160.08\end{aligned}$$

Step 3: Compute correlation for the values

The next step would be to compute the correlation of all the pairs of values. This serves two purposes:

1. It explains the relationship between the two variables of interest; and
2. It can be used to compute the regression line equation.

For simplicity, we will let Excel calculate the correlation of these values for us. (For a more in depth explanation of correlation, see the annotated example.) The correlation for this data set is

$$r = 0.9854$$

Examining the value we can tell that these two variables have a strong linear relationship and hence regression is a valid technique to use.

Step 4: Plug in the values and press 'Go'

Now we just need to plug in the values for the variables and then press the 'Go' button to get our answers. First let's summarize our values to see what we've got:

mean x	$\bar{x}$	22.8333
mean y	$\bar{y}$	783.3333
standard deviation of x	s_x	7.1671
standard deviation of y	s_y	160.08
correlation	r	0.9854

Now we can plug into our formulas. Recall that we first need to compute the value for b since we need that to compute the value for a.

$$b = r \times \frac{s_y}{s_x}$$

$$a = \bar{y} - b\bar{x}$$

So plugging in our values gives us

$$b = 0.9854 \times \frac{160.08}{7.1671} = 22.0093$$

$$a = 783.3333 - (22.0093 \times 22.8333) = 280.7877$$

So now entering these values into our line equation we get

$$\hat{y} = a + bx$$

$$\hat{y} = 280.7877 + 22.0093x$$

Step 5: Testing the equation and comparing it to reality

Is this the right equation? Is this the best line? How well did we do?

We can check the "rightness" of the equation by plugging in the mean of x (22.8333). The output of the equation should be the mean of y (783.3333) if we did all the calculations correctly.

$$\begin{aligned} \hat{y} &= 280.7877 + 22.0093 \times 22.8333 \\ &= 280.7877 + 504.4561 \\ &= 783.3334 \end{aligned}$$

Since we get the same value for $\hat{y}$ as the mean of y (with allowance for rounding) we can be assured we did the math correctly.

Step 6: How'd we do?

Size of the house, x	21	13	19	27	34	23
Monthly rent, y	700	580	720	850	1050	800
Predicted rent, $\hat{y}$	742.98	566.91	698.96	875.04	1029.10	787.00
Residual $y - \hat{y}$	-42.98	13.00	21.04	-25.04	20.90	13.00

If we sum the residuals we get 0.0. This means that we have the same amount of error above the line as we do below the line. But how well did we really do? If we take the correlation coefficient and square it we can determine how much of the variation in y we are explaining with x.

$$r = 0.9855$$
$$r^2 = 0.9712$$

So from this we can see that the explanatory variable x is explaining 97.12% of the variation in y. That leaves only 2.88% of the variability to be explained by other variables.

If we plot the regression line on the same scatterplot as with the actual data points we can see how well it "fits" the data.


Monthly Rent vs. Square Footage
Rent
11
10
9
8
7
6
5
10
15
20
25
30
35
Size

/ /