Standard Deviation and Linear Regression

Let us begin by examining the standard deviation formulas. There are two sets of formulas that you can use to determine the standard deviation. The first set uses the squared deviations from the mean, $s=\sqrt{\frac{\sum(x_k-\bar{x})^2}{n-1}}$ and $\sigma=\sqrt{\frac{\sum(x_k-\mu)^2}{N}}$, and the second set uses the sum of the data values, $\sum x_k$, and the sum of the squares of the data values, $\sum x_k^2$, $s=\sqrt{\frac{n\sum x_k^2-\left(\sum x_k\right)^2}{n(n-1)}}$ and $\sigma=\sqrt{\frac{N\sum x_k^2-\left(\sum x_k\right)^2}{N^2}}$.

The advantage to using the first set of formulas is that you can determine if you have calculated the mean correctly by examining $\sum(x_k-\bar{x})$ and $\sum(x_k-\mu)$, respectively. Since the necessary sums are easier to determine using a table, you can construct a table that has columns for x_k, $(x_k-\bar{x})$, and $(x_k-\bar{x})^2$, for determining the sample standard deviation and x_k, $(x_k-\mu)$, and $(x_k-\mu)^2$, for determining the population standard deviation. The disadvantage to using the first set of formulas is that you CANNOT round any values used in intermediate calculations. So, that means that when you determine the deviations from the mean, $(x_k-\bar{x})$ or $(x_k-\mu)$, you CANNOT round *any* of these values and when you square these values to determine the squared deviations from the mean, $(x_k-\bar{x})^2$ or $(x_k-\mu)^2$, you CANNOT round *any* of these values either: you must keep all decimal places for determining the sum of the deviations from the mean, $\sum(x_k-\bar{x})$ or $\sum(x_k-\mu)$, and the sum of the squared deviations from the mean, $\sum(x_k-\bar{x})^2$ or $\sum(x_k-\mu)^2$. Since the mean may not be a whole number or be a terminating decimal value, this means that you will have many decimal places to record and, of course, that means that there are many opportunities to record values incorrectly by omitting or switching digits.

The second set of formulas, $s=\sqrt{\frac{n\sum x_k^2-\left(\sum x_k\right)^2}{n(n-1)}}$ and $\sigma=\sqrt{\frac{N\sum x_k^2-\left(\sum x_k\right)^2}{N^2}}$, provide a solution to the problem or recording all decimal places as well as sums that have many decimal places since using these formulas only require you to determine the sum of the data values, $\sum x_k$, and the sum of the squares of the data values, $\sum x_k^2$. Making a table to organize the necessary values and to use to determine the necessary sums is simpler as well since you need only a column for the original data values, x_k, and a column for the squares of the data values, x_k^2. Since the data values, x_k, have a fixed number of data values, their squares, x_k^2, will have a fixed number of data values as well; the number of decimal places in the squares of the data values, x_k^2, is twice as many as in the original data values. So, the required sums, the sum of the data values, $\sum x_k$, and the sum of the squares of the data values, $\sum x_k^2$, each have a fixed number of decimal places. The second set of formulas have no disadvantage since they are always easy to use. An added advantage is their similarity to the simplest of the regression formulas. We will explore these similarities shortly.

So, having discussed the formulas, it would be nice to explore an example. To that end, it is important to use data with a reasonable number of data values so that you will understand the advantages and disadvantages of these formulas. Let us consider the Nursing Home Data, http://lib.stat.cmu.edu/DASL/Datafiles/nursinghomedat.html, available on the Data and Story Library (DASL) web site http://lib.stat.cmu.edu/DASL/. In order to make sure that you understand the data, please visit http://lib.stat.cmu.edu/DASL/Datafiles/nursinghomedat.html to read the description of the data: this will help you to understand the context of the data. Let us consider the annual facility expenditures, in hundreds of dollars, that is provided in the column marked with the heading FEXP. If we want to determine the standard deviation using the first set of formulas then we need to set up and use a table to determine the sum of the data values, the mean of the data values, the deviations from the mean for the data values, the sum of the deviations from the mean for the data values, the squared deviations from the mean for the data values, and the sum of the squared deviations from the mean for the data values. For the second set of formulas, we need to set up and use a table to determine the sum of the data values, the squares of the data values, and the sum of the squares of the data values. Examining these tables, provided for your convenience on the next page, you will see that producing the first table requires more time and attention to detail than producing the second table. To put it mildly, the second set of formulas is easier to use. However, the choice of which set of formulas to use is yours.

FEXP (x)	x - mean	(x - mean)^2
5334	2486.4615384615	6182490.9822485200
493	-2354.5384615385	5543851.3668639000
6115	3267.4615384615	10676304.9053254000
6346	3498.4615384615	12239233.1360947000
6225	3377.4615384615	11407246.4437870000
449	-2398.5384615385	5752986.7514792900
4998	2150.4615384615	4624484.8284023700
966	-1881.5384615385	3540186.9822485200
1260	-1587.5384615385	2520278.3668639000
6442	3594.4615384615	12920153.7514793000
1236	-1611.5384615385	2597056.2130177500
3360	512.4615384615	262616.8284023670
4231	1383.4615384615	1913965.8284023700
1280	-1567.5384615385	2457176.8284023700
1123	-1724.5384615385	2974032.9053254400
5206	2358.4615384615	5562340.8284023700
4443	1595.4615384615	2545497.5207100600
4585	1737.4615384615	3018772.5976331400
1675	-1172.5384615385	1374846.4437869800
5686	2838.4615384615	8056863.9053254400
907	-1940.5384615385	3765689.5207100600
3351	503.4615384615	253473.5207100590
1756	-1091.5384615385	1191456.2130177500
2123	-724.5384615385	524955.9822485210
4531	1683.4615384615	2834042.7514792900
2543	-304.5384615385	92743.6745562130
4446	1598.4615384615	2555079.2899408300
1064	-1783.5384615385	3181009.4437869800
2987	139.4615384615	19449.5207100592
411	-2436.5384615385	5936719.6745562100
4197	1349.4615384615	1821046.4437869800
1198	-1649.5384615385	2720977.1360946700
1209	-1638.5384615385	2684808.2899408300
137	-2710.5384615385	7347018.7514792900
1279	-1568.5384615385	2460312.9053254400
1273	-1574.5384615385	2479171.3668639000
3524	676.4615384615	457600.2130177520
2561	-286.5384615385	82104.2899408283
3874	1026.4615384615	1053623.2899408300
6402	3554.4615384615	12634196.8284024000
1911	-936.5384615385	877104.2899408280
1122	-1725.5384615385	2977482.9822485200
3893	1045.4615384615	1092989.8284023700
2212	-635.5384615385	403909.1360946740
2959	111.4615384615	12423.6745562130
3006	158.4615384615	25110.0591715977
1344	-1503.5384615385	2260627.9053254400
1242	-1605.5384615385	2577753.7514792900
1484	-1363.5384615385	1859237.1360946700
1154	-1693.5384615385	2868072.5207100600
245	-2602.5384615385	6773206.4437869800
6274	3426.4615384615	11740638.6745562000
148072	0	193734422.9230770000

FEXP (x)	x^2
5334	28451556
493	243049
6115	37393225
6346	40271716
6225	38750625
449	201601
4998	24980004
966	933156
1260	1587600
6442	41499364
1236	1527696
3360	11289600
4231	17901361
1280	1638400
1123	1261129
5206	27102436
4443	19740249
4585	21022225
1675	2805625
5686	32330596
907	822649
3351	11229201
1756	3083536
2123	4507129
4531	20529961
2543	6466849
4446	19766916
1064	1132096
2987	8922169
411	168921
4197	17614809
1198	1435204
1209	1461681
137	18769
1279	1635841
1273	1620529
3524	12418576
2561	6558721
3874	15007876
6402	40985604
1911	3651921
1122	1258884
3893	15155449
2212	4892944
2959	8755681
3006	9036036
1344	1806336
1242	1542564
1484	2202256
1154	1331716
245	60025
6274	39363076
148072	615375138

Once we have determined the required sums, we are ready to use the formulas to determine the standard deviation. So, suppose that we are only studying these 52 licensed nursing facilities in New Mexico: if we are only studying these 52 licensed nursing facilities in New Mexico then we have data for a population and we must determine the population standard deviation. So, using the population standard deviation formula from the first set of formulas, $\sigma = \sqrt{\frac{\sum (x_k - \mu)^2}{N}}$, we find that

$$\begin{aligned} \sigma &= \sqrt{\frac{193734422.923077}{52}} \\ &= 1930.1973938667 \\ &\approx 1930.2 \text{ hundred dollars} \end{aligned}.$$

Using the population standard deviation formula from the second set of formulas, $\sigma = \sqrt{\frac{N\sum x_k^2 - \left(\sum x_k\right)^2}{N^2}}$, we find that

$$\begin{aligned} \sigma &= \sqrt{\frac{52(615375138) - (148072)^2}{(52)^2}} \\ &= 1930.1973938667 \\ &\approx 1930.2 \text{ hundred dollars} \end{aligned}.$$

As we should, we find the same value for the population standard deviation using each formula. Please remember that we record the final value our calculated population standard deviation to one more decimal place than that used in the original data; since the original data is recorded to zero decimal places, we use one decimal place when recording the value of the population standard deviation.

Now, suppose that we are only studying these the 60 licensed nursing facilities in New Mexico but that we only have data for 52 licensed nursing facilities in New Mexico: if we can only use data for 52 of the 60 licensed nursing facilities in New Mexico then we have data for a sample and we must determine the sample standard deviation. So, using the sample standard deviation formula from the first set of formulas, $s = \sqrt{\frac{\sum (x_k - \bar{x})^2}{n-1}}$, we find that

$$\begin{aligned} s &= \sqrt{\frac{193734422.923077}{52-1}} \\ &= 1949.0290338941 \\ &\approx 1949.0 \text{ hundred dollars} \end{aligned}.$$

Using the sample standard deviation formula from the second set of formulas, $s = \sqrt{\frac{n\sum x_k^2 - \left(\sum x_k\right)^2}{n(n-1)}}$, we find that

$$\begin{aligned} s &= \sqrt{\frac{52(615375138) - (148072)^2}{52 \cdot (52-1)}} \\ &= 1949.0290338941 \\ &\approx 1949.0 \text{ hundred dollars} \end{aligned}.$$

As we should, we find the same value for the sample standard deviation using each formula. Please remember, just as before, that we record the final value our calculated sample standard deviation to one more decimal place than that used in the original data; since the original data is recorded to zero decimal places, we use one decimal place when recording the value of the sample standard deviation.

You must decide for yourself which set of formulas you prefer to use. You will get the same value for the standard deviation, population or sample, but the amount of work that you do in order to determine the necessary sums will be different. Please remember that you CANNOT round any of the intermediate values used in these calculations: you must record all decimal places.

Now, let us consider the formulas for determining the correlation coefficient, r, and the slope of the least squares line, m. If you are partial to the standard deviation formulas, $s=\sqrt{\dfrac{\sum\left(x_k-\bar{x}\right)^2}{n-1}}$ and $\sigma=\sqrt{\dfrac{\sum\left(x_k-\mu\right)^2}{N}}$, then you may prefer using the correlation coefficient formula $r=\dfrac{\sum\left(\dfrac{x_k-\bar{x}}{s_x}\right)\left(\dfrac{y_k-\bar{y}}{s_y}\right)}{n-1}$ and the slope of the least squares line formula $m=\dfrac{\sum\left(x_k-\bar{x}\right)\left(y_k-\bar{y}\right)}{\sum\left(x_k-\bar{x}\right)^2}$. However, you will have to determine the quotients $\dfrac{x_k-\bar{x}}{s_x}$ and $\dfrac{y_k-\bar{y}}{s_y}$ and their *products* and its sum in order to determine the correlation coefficient as well as $x_k-\bar{x}$ and $y_k-\bar{y}$, and their product and its sum in addition to $\left(x_k-\bar{x}\right)^2$ and its sum. For each of these, you CANNOT round any intermediate values: you must record and use ALL decimal places. So, using these formulas necessitates a great deal of care in recording values and using these values to calculate the other required values. It is important to note that s_x and s_y are the sample standard deviation for the x-values and the sample standard deviation for the y-values, respectively, and that they CANNOT be rounded: you must use ALL decimal places of these values in your calculations. Once you round/truncate any values, you introduce error: error grows as you perform operations on these values (this is called propagation of error).

In order to help you to see the extent of the work necessary to using the correlation coefficient formula $r=\dfrac{\sum\left(\dfrac{x_k-\bar{x}}{s_x}\right)\left(\dfrac{y_k-\bar{y}}{s_y}\right)}{n-1}$, let us consider an example. Suppose we use the Nursing Home Data that we considered earlier. Since the annual facility expenditures, in hundreds of dollars, depends on the number of beds in the facility, we can take the number of beds in the facility as the explanatory variable and the annual facility expenditures, in hundreds of dollars, as the response variable. Using these variables, we can explore the linear relation between these variables. So, let us examine the table that we would need to create in order to determine the correlation coefficient for this linear relation using $r=\dfrac{\sum\left(\dfrac{x_k-\bar{x}}{s_x}\right)\left(\dfrac{y_k-\bar{y}}{s_y}\right)}{n-1}$. Check out the table on the next page. You will see that we must create seven columns in order to determine the one required sum, $\sum\left(\dfrac{x_k-\bar{x}}{s_x}\right)\left(\dfrac{y_k-\bar{y}}{s_y}\right)$; this sum, 23.47305042, is located at the bottom of the seventh column Please take careful note of all the decimal places that must be recorded and carefully consider if you are up to the task if you plan to use this formula for the correlation coefficient. Having determined the necessary sum, we can then determine the correlation coefficient.

$$\begin{aligned} r &= \frac{23.47305042}{52} \\ &= 0.460255891 \\ &\approx 0.5 \end{aligned}$$

It is important to remember that the linear correlation coefficient tells us the strength of the linear relation and that the correlation coefficient has no units. So, for licensed nursing facilities in New Mexico, there is a weak linear relation between the number of beds in the facility and the annual facility expenditures, in hundreds of dollars. In interpreting the correlation coefficient, you may find it helpful to associate the magnitude of the correlation coefficient with grades – 0.5 corresponding to a very weak linear relation, 0.6 corresponding to a weak linear relation, 0.7 corresponding to a relatively strong linear relation, 0.8 corresponding to a strong linear relation, 0.9 corresponding to a very strong linear relation, and 0.99 corresponding to an extremely strong linear relation; of course, we need to use more adjectives as the magnitude of r gets closer to 1. Please keep in mind that the closer to 0 the magnitude of the linear correlation coefficient is, the weaker the linear relation between the variables is. If the linear correlation coefficient is 0 then there is no *linear* relation between the variables; notice that the statement is that there is no *linear relation between the variables* not that there is no relation between the variables.

All right, let us now determine the linear correlation coefficient using the other formula.

BED (x)	x - mean	(x - mean)/s_x	FEXP (y)	y - mean	(y - mean)/s_y	[(x - mean)/s_x] * [(y - mean)/s_y]
244	150.7307692	3.689612967	5334	2486.461538	1.275743714	4.707000549
59	-34.26923077	-0.838847959	493	-2354.538462	-1.208057151	1.013376276
120	26.73076923	0.654320238	6115	3267.461538	1.676456062	1.096939129
120	26.73076923	0.654320238	6346	3498.461538	1.794976615	1.174489526
120	26.73076923	0.654320238	6225	3377.461538	1.732894421	1.133867889
65	-28.26923077	-0.691978956	449	-2398.538462	-1.230632494	0.851571789
120	26.73076923	0.654320238	4998	2150.461538	1.103350182	0.721944353
90	-3.269230769	-0.080024777	966	-1881.538462	-0.965372208	0.077253696
96	2.730769231	0.066844226	1260	-1587.538462	-0.814527867	-0.054446485
120	26.73076923	0.654320238	6442	3594.461538	1.84423191	1.206718262
62	-31.26923077	-0.765413458	1236	-1611.538462	-0.826841691	0.632875758
120	26.73076923	0.654320238	3360	512.4615385	0.262931711	0.17204154
116	22.73076923	0.556407569	4231	1383.461538	0.709820898	0.39494972
59	-34.26923077	-0.838847959	1280	-1567.538462	-0.804266347	0.674657184
80	-13.26923077	-0.324806449	1123	-1724.538462	-0.884819277	0.287395007
120	26.73076923	0.654320238	5206	2358.461538	1.210069987	0.791773282
80	-13.26923077	-0.324806449	4443	1595.461538	0.818593007	-0.265884288
100	6.730769231	0.164756894	4585	1737.461538	0.891449798	0.1468725
60	-33.26923077	-0.814369792	1675	-1172.538462	-0.601601331	0.489925951
110	16.73076923	0.409538566	5686	2838.461538	1.456346462	0.596430042
120	26.73076923	0.654320238	907	-1940.538462	-0.995643691	-0.651469817
135	41.73076923	1.021492745	3351	503.4615385	0.258314027	0.263865905
59	-34.26923077	-0.838847959	1756	-1091.538462	-0.560042176	0.469790237
60	-33.26923077	-0.814369792	2123	-724.5384615	-0.371743288	0.302736504
25	-68.26923077	-1.671105643	4531	1683.461538	0.863743695	-1.443406962
221	127.7307692	3.126615122	2543	-304.5384615	-0.156251372	-0.488537904
64	-29.26923077	-0.716457124	4446	1598.461538	0.820132235	-0.587589582
62	-31.26923077	-0.765413458	1064	-1783.538462	-0.915090761	0.700422784
108	14.73076923	0.360582232	2987	139.4615385	0.071554367	0.025801233
62	-31.26923077	-0.765413458	411	-2436.538462	-1.250129382	0.956865853
90	-3.269230769	-0.080024777	4197	1349.461538	0.692376314	-0.05540726
146	52.73076923	1.290752584	1198	-1649.538462	-0.846338578	-1.092413707
62	-31.26923077	-0.765413458	1209	-1638.538462	-0.840694742	0.64347907
30	-63.26923077	-1.548714807	137	-2710.538462	-1.390712203	2.153816582
79	-14.26923077	-0.349284616	1279	-1568.538462	-0.804779423	0.281097072
44	-49.26923077	-1.206020467	1273	-1574.538462	-0.807857879	0.974293137
120	26.73076923	0.654320238	3524	676.4615385	0.347076173	0.227098964
100	6.730769231	0.164756894	2561	-286.5384615	-0.147016005	-0.0242219
49	-44.26923077	-1.083629631	3874	1026.461538	0.526652769	-0.570696546
123	29.73076923	0.727754739	6402	3554.461538	1.823708871	1.327212774
82	-11.26923077	-0.275850115	1911	-936.5384615	-0.480515398	0.132550228
58	-35.26923077	-0.863326127	1122	-1725.538462	-0.885332353	0.764330552
110	16.73076923	0.409538566	3893	1045.461538	0.536401213	0.219676984
62	-31.26923077	-0.765413458	2212	-635.5384615	-0.326079525	0.249585657
86	-7.269230769	-0.177937446	2959	111.4615385	0.057188239	-0.010175929
102	8.730769231	0.213713229	3006	158.4615385	0.081302811	0.017375486
135	41.73076923	1.021492745	1344	-1503.538462	-0.771429484	-0.788009621
78	-15.26923077	-0.373762783	1242	-1605.538462	-0.823763235	0.307892039
83	-10.26923077	-0.251371947	1484	-1363.538462	-0.699598845	0.175859524
60	-33.26923077	-0.814369792	1154	-1693.538462	-0.868913922	0.70761725
54	-39.26923077	-0.961238795	245	-2602.538462	-1.335299996	1.28354216
120	26.73076923	0.654320238	6274	3426.461538	1.758035144	1.150317973
						23.47305042

If we want to determine the linear correlation coefficient using $r = \dfrac{n\sum x_k y_k - \left(\sum x_k\right)\left(\sum y_k\right)}{\sqrt{n\sum x_k^2 - \left(\sum x_k\right)^2}\sqrt{n\sum y_k^2 - \left(\sum y_k\right)^2}}$ then we need to determine $\sum x_k$, $\sum x_k^2$, $\sum y_k$, $\sum y_k^2$, and $\sum x_k y_k$. To do this, we can construct and use a table that has columns for x_k, x_k^2, y_k, y_k^2, and $x_k y_k$; ordering these columns as x_k^2, x_k, $x_k y_k$, y_k, and y_k^2 makes the table easy to complete

x^2	BED (x)	xy	FEXP (y)	y^2
59536	244	1301496	5334	28451556
3481	59	29087	493	243049
14400	120	733800	6115	37393225
14400	120	761520	6346	40271716
14400	120	747000	6225	38750625
4225	65	29185	449	201601
14400	120	599760	4998	24980004
8100	90	86940	966	933156
9216	96	120960	1260	1587600
14400	120	773040	6442	41499364
3844	62	76632	1236	1527696
14400	120	403200	3360	11289600
13456	116	490796	4231	17901361
3481	59	75520	1280	1638400
6400	80	89840	1123	1261129
14400	120	624720	5206	27102436
6400	80	355440	4443	19740249
10000	100	458500	4585	21022225
3600	60	100500	1675	2805625
12100	110	625460	5686	32330596
14400	120	108840	907	822649
18225	135	452385	3351	11229201
3481	59	103604	1756	3083536
3600	60	127380	2123	4507129
625	25	113275	4531	20529961
48841	221	562003	2543	6466849
4096	64	284544	4446	19766916
3844	62	65968	1064	1132096
11664	108	322596	2987	8922169
3844	62	25482	411	168921
8100	90	377730	4197	17614809
21316	146	174908	1198	1435204
3844	62	74958	1209	1461681
900	30	4110	137	18769
6241	79	101041	1279	1635841
1936	44	56012	1273	1620529
14400	120	422880	3524	12418576
10000	100	256100	2561	6558721
2401	49	189826	3874	15007876
15129	123	787446	6402	40985604
6724	82	156702	1911	3651921
3364	58	65076	1122	1258884
12100	110	428230	3893	15155449
3844	62	137144	2212	4892944
7396	86	254474	2959	8755681
10404	102	306612	3006	9036036
18225	135	181440	1344	1806336
6084	78	96876	1242	1542564
6889	83	123172	1484	2202256
3600	60	69240	1154	1331716
2916	54	13230	245	60025
14400	120	752880	6274	39363076
537472	4850	15679560	148072	615375138

and minimizes the amount of confusion when determining $x_k y_k$ since these values are on either side of the column that you are completing (please see the example table provided above). One important thing to notice in this table is that the sums will all involve a fixed number of decimal places unlike the table used with the other formula. In addition, you will use four of these five values, x_k^2, x_k, $x_k y_k$, and y_k, to determine the slope for the least squares line if you use $m=\frac{n\sum x_k y_k-\left(\sum x_k\right)\left(\sum y_k\right)}{n\sum x_k^2-\left(\sum x_k\right)^2}$ and two of these sums, x_k and y_k, if you use $b=\frac{\sum y_k-m\sum x_k}{n}$ to determine b, the y-coordinate of the y-intercept, for the least squares regression line.

Let us use $r=\frac{n\sum x_k y_k-\left(\sum x_k\right)\left(\sum y_k\right)}{\sqrt{n\sum x_k^2-\left(\sum x_k\right)^2}\sqrt{n\sum y_k^2-\left(\sum y_k\right)^2}}$ to determine the linear correlation coefficient and compare the result to that which we obtained using the other formula.

$$
\begin{aligned}
r &= \frac{52(15679560)-(4850)(148072)}{\sqrt{52(537472)-(4850)^2}\sqrt{52(615375138)-(148072)^2}} \\
&= \frac{97187920}{\sqrt{4426044}\sqrt{10074189992}} \\
&= 0.460255891 \\
&\approx 0.5
\end{aligned}
$$

So, as we should, we obtained the same value for the correlation coefficient. You might wonder why one might choose to use this formula over the other formula. Well, we would choose to use this formula due to its similarities the slope for the least squares line formula, $m=\frac{n\sum x_k y_k-\left(\sum x_k\right)\left(\sum y_k\right)}{n\sum x_k^2-\left(\sum x_k\right)^2}$: the numerator for $r=\frac{n\sum x_k y_k-\left(\sum x_k\right)\left(\sum y_k\right)}{\sqrt{n\sum x_k^2-\left(\sum x_k\right)^2}\sqrt{n\sum y_k^2-\left(\sum y_k\right)^2}}$ and $m=\frac{n\sum x_k y_k-\left(\sum x_k\right)\left(\sum y_k\right)}{n\sum x_k^2-\left(\sum x_k\right)^2}$ are the same and $n\sum x_k^2-\left(\sum x_k\right)^2$ appears in the denominator of $m=\frac{n\sum x_k y_k-\left(\sum x_k\right)\left(\sum y_k\right)}{n\sum x_k^2-\left(\sum x_k\right)^2}$ and in the first square root of the denominator of $r=\frac{n\sum x_k y_k-\left(\sum x_k\right)\left(\sum y_k\right)}{\sqrt{n\sum x_k^2-\left(\sum x_k\right)^2}\sqrt{n\sum y_k^2-\left(\sum y_k\right)^2}}$.
So, due to these similarities, we have already have determined the two values that we need to determine the slope of the least squares line, the values of the numerator and denominator of the fraction.

$$
\begin{aligned}
m &= \frac{52(15679560)-(4850)(148072)}{52(537472)-(4850)^2} \\
&= \frac{97187920}{4426044} \\
&= 21.95819111 \\
&\approx 22.0 \text{ hundred dollars per bed}
\end{aligned}
$$

Please recall that the units for the slope are the quotient of the y-units and the x-units. Since the response variable is the annual facility expenditures, in hundred dollars, and the explanatory variable is the number of beds in the facility, the y-units are hundred dollars and the x-units are beds, making the slope-units hundred dollars per bed.

Using $b=\frac{\sum y_k-m\sum x_k}{n}$, we can determine the value of b. It is important to note that we already have the necessary sums, $\sum x_k$ and $\sum y_k$.

$$b = \frac{148072 - \left(\frac{97187920}{4426044}\right)4850}{52}$$
$$= 799.5148679$$
$$\approx 799.5 \text{ hundred dollars}$$

Rather than using the value of the slope expressed as a fraction above, we could use the full decimal value,

$$b = \frac{148072 - (21.95819111)4850}{52}$$
$$= 799.5148679$$
$$\approx 799.5 \text{ hundred dollars}$$

Either way, we obtain the same value. However, we CANNOT USE AN APPROXIMATION FOR THE SLOPE in our calculation of b: we CANNOT use 22.0 in our calculation since we cannot round/truncate/approximate intermediate values used in calculations. So, the value of the y-coordinate of the y-intercept is 799.5 hundred dollars and the y-intercept is $(0,\ 799.5)$; recall that the y-intercept is a point. Combining these last two results, we find that the equation for the least squares line is $y = 22.0x + 799.5$; notice that the values of m and b have been recorded to one decimal place since the original data values have zero decimal places.

The last thing that we will do is interpret the slope and the y-intercept. To interpret the slope, it is important to remember that

$$m = \frac{\text{change in y}}{\text{change in x}}$$
$$= \frac{\text{change in response variable}}{\text{change in explanatory variable}}$$

Adding the meaning of our variables, we have

$$m = \frac{\text{change in the annual facility expenditures}}{\text{change in the number of beds in the facility}}$$

To use this, we must express our slope as a fraction. This is easy to do since all numbers can be expressed as a fraction simply by putting a 1 in the denominator.

$$m = 22.0\ \frac{\text{hundred dollars}}{\text{bed}}$$
$$= \frac{22.0}{1}\frac{\text{hundred dollars}}{\text{bed}}$$
$$= \frac{22.0 \text{ hundred dollars}}{1 \text{ bed}}$$

Combining these two ideas, we find that the numerator is a positive change, an increase by 22.0 hundred dollars or a 22.0 hundred dollar increase, and the denominator is a positive change, a 1 bed increase or an increase by 1 in the number of beds. So, adding the context as well as the variables and the dependence of the annual facility expenditures on the number of beds, we can write a sentence that interprets the slope.

Here are three interpretations of the slope:

(i) For licensed nursing facilities in New Mexico, for each additional bed in a facility, the annual facility expenditures increases by 22.0 hundred dollars.

(ii) For licensed nursing facilities in New Mexico, if the number of beds in the facility is increased by 1 then the annual facility expenditures increases by 22.0 hundred dollars.

(iii) For licensed nursing facilities in New Mexico, for each additional bed in a facility, there is a 22.0 hundred dollar increase in the annual facility expenditures.

If the slope is negative then we keep the negative in the numerator and use the word decrease in the interpretation since a negative change is a decrease.

For the interpretation of the y-intercept, $(0, 799.5)$, we must remember that the x-value is the number of beds in the facility and that the y-value is the annual facility expenditures. With this in mind, all we need to do is make sure that our interpretation includes the dependence of the annual facility expenditures on the number of beds in the facility and the context for the data.

Here are three interpretations for the y-intercept:

(i) For licensed nursing facilities in New Mexico, the annual facility expenditures for a facility having zero beds is 799.5 hundred dollars.

(ii) For licensed nursing facilities in New Mexico, a facility having zero beds has annual facility expenditures of 799.5 hundred dollars.

(iii) The annual facility expenditures for licensed nursing facilities in New Mexico with zero beds is 799.5 hundred dollars.

All interpretations for the slope m must include the dependence of the response variable on the explanatory variable as well as the idea of an increase or decrease in the value of the response variable for a unit (1) increase in the value of the explanatory variable. Interpretations of the y-intercept must include the dependence of the response variable on the explanatory variable. Each interpretation must include the context for the data.

Recommended formulas for use:

Sample Standard Deviation: $s = \sqrt{\dfrac{n\sum x_k^2 - \left(\sum x_k\right)^2}{n(n-1)}}$

Population Standard Deviation: $\sigma = \sqrt{\dfrac{N\sum x_k^2 - \left(\sum x_k\right)^2}{N^2}}$

Slope for the Least Squares Regression Line: $m = \dfrac{n\sum x_k y_k - \left(\sum x_k\right)\left(\sum y_k\right)}{n\sum x_k^2 - \left(\sum x_k\right)^2}$

Correlation Coefficient: $r = \dfrac{n\sum x_k y_k - \left(\sum x_k\right)\left(\sum y_k\right)}{\sqrt{n\sum x_k^2 - \left(\sum x_k\right)^2}\sqrt{n\sum y_k^2 - \left(\sum y_k\right)^2}}$

Y-Coordinate of the Y-Intercept for the Least Squares Regression Line: $b = \dfrac{\sum y_k - m\sum x_k}{n}$

Equation for the Least Squares Regression Line: $y = mx + b$ where you substitute in the values of m and b

Convenient table set up – order for columns: x_k^2, x_k, $x_k y_k$, y_k, and y_k^2