9 Analyzing Relationships Among Variables

9.1 Introduction to Analyzing Relations

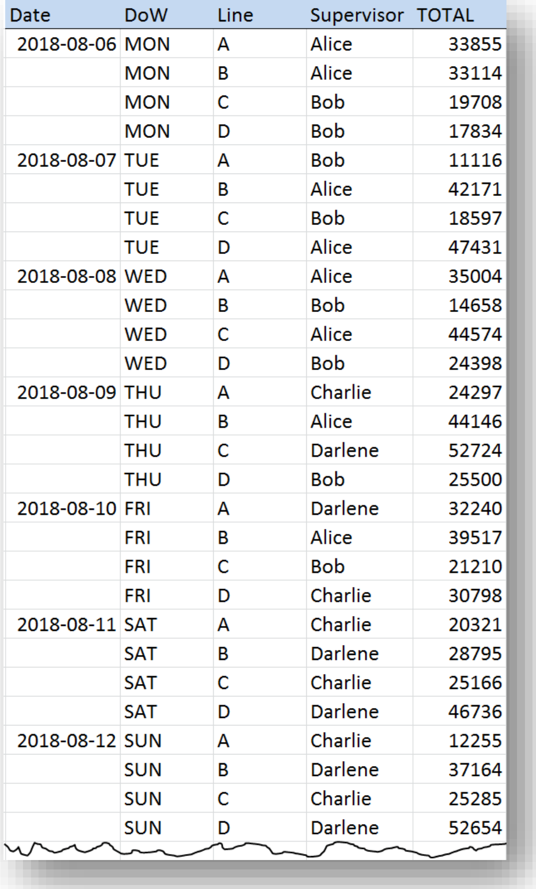
There are many cases where we need to discuss more than one aspect of the entities we are interested in understanding. For example, in analyzing productivity figures in a factory as part of an operations research study, we may collect information about several aspects of each product line beyond just the daily production such as

- **Day of the week** – maybe production changes systematically during the work week; keeping track of exact dates and therefore allowing us to track Mondays, Tuesdays and so on could be useful.
- **Shift supervisor** – perhaps particular supervisors differ in their effects on productivity; a nasty supervisor, for example, might cause resentment among workers and get poorer productivity than a good supervisor. Alternatively, perhaps lower production in "Bob's" shift is due to theft orchestrated by Bob!
- **Type of equipment on each production line** – maybe specific lines are affected by differences in the machinery.
- **Ambient temperature in factory** – perhaps differences in productivity can be traced to working conditions such as excessive heat during afternoon shifts compared with nighttime or morning shifts.

Figure 9-1 shows some sample data with a few of these observations. There is no theoretical limit on the level of detail that we can collect in scientific and professional studies. What limits our ability to study all aspects of reality are such factors as

- **The difficulty or impossibility of quantifying specific attributes of reality;** e.g., there is no easy, simple measure of such human attributes as *honesty* or *originality*; and there is no immediate, simple measure of a product's *utility* or *marketing appeal*.
- **Ability to define metrics (ways of measuring something);** e.g., marketing appeal might be measured through studies of purchasing habits for that product.
- **The complexity of acquiring the data;** e.g., a measure that we think might be indicative of *honesty* might require extensive testing of every subject, possibly in different situations and environments. Similarly, measuring a product's *utility* might involve extensive studies of how consumers actually use the product over a period of years.
- **Ability to identify independent factors possibly affecting the variable(s) of interest;** e.g., in a wide range of different populations grouped by such factors as age, gender, socio-economic status, and so on.
- **The controllability of factors;** e.g., it might be possible to impose experimental conditions on subjects in a study of honesty by giving them tasks in a laboratory or via computer; it might be much more difficult to perform such studies in the real world.
- **Increased costs resulting from increased complexity of data gathering:** It's cheaper to weigh bolts to see if they vary in weight than it is to study the marketing benefits of offering five different shades of those bolts.

Figure 9-1.Multiple variables in observations about production line (first week of data only).



Date	DoW	Line	Supervisor	TOTAL
2018-08-06	MON	A	Alice	33855
	MON	B	Alice	33114
	MON	C	Bob	19708
	MON	D	Bob	17834
2018-08-07	TUE	A	Bob	11116
	TUE	B	Alice	42171
	TUE	C	Bob	18597
	TUE	D	Alice	47431
2018-08-08	WED	A	Alice	35004
	WED	B	Bob	14658
	WED	C	Alice	44574
	WED	D	Bob	24398
2018-08-09	THU	A	Charlie	24297
	THU	B	Alice	44146
	THU	C	Darlene	52724
	THU	D	Bob	25500
2018-08-10	FRI	A	Darlene	32240
	FRI	B	Alice	39517
	FRI	C	Bob	21210
	FRI	D	Charlie	30798
2018-08-11	SAT	A	Charlie	20321
	SAT	B	Darlene	28795
	SAT	C	Charlie	25166
	SAT	D	Darlene	46736
2018-08-12	SUN	A	Charlie	12255
	SUN	B	Darlene	37164
	SUN	C	Charlie	25285
	SUN	D	Darlene	52654

Such considerations lead to *multifactorial analysis*. Simple approaches to handling such multifactorial data include *cross-tabulations* (also known as *contingency tables*), *scatterplots* for showing two variables at a time, *correlation coefficients* to express the intensity of a relationship between two variables, and *regression equations* to predict a variable's value as a function of one or more other variables.

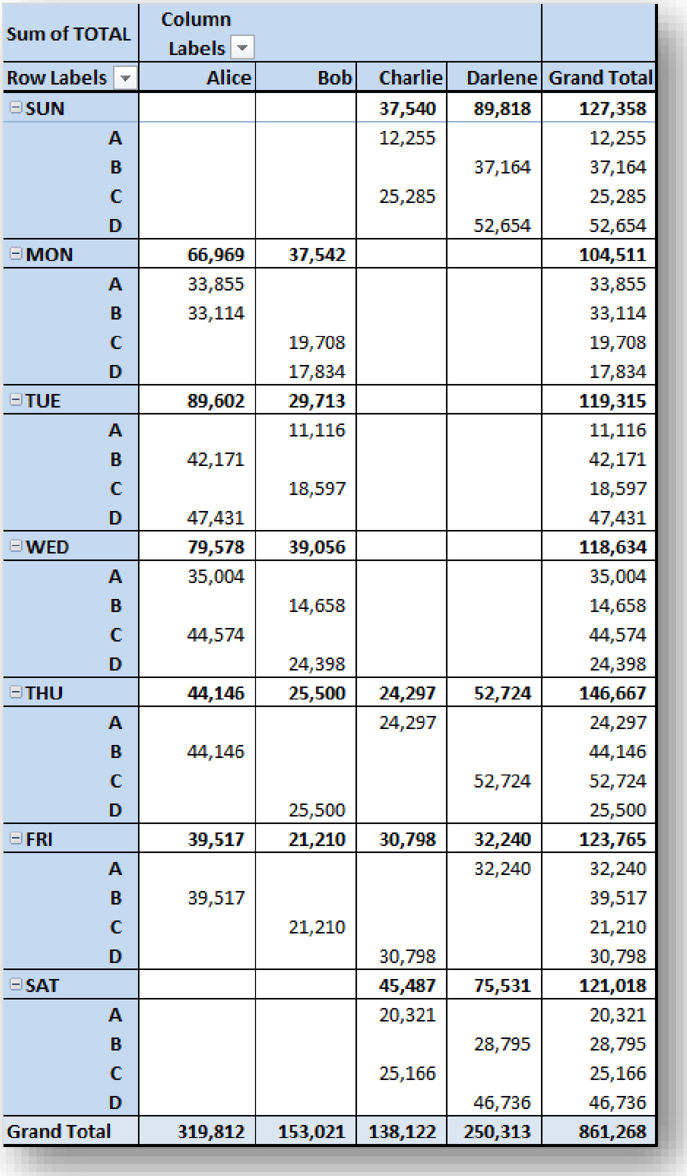
9.2 Cross-Tabulations (Contingency Tables)

Let's consider the production-line example introduced above. Figure 9-2 shows the first week of collected data, with information on the day of week (Monday through Sunday), production line (A, B, C, or D) and supervisors (Alice, Bob, Charlie and Darlene). As you can see, it's difficult to make sense of these unaggregated data.

One summary representation, shown in Figure 9-2 is three-way cross-tabulation (or *three-way contingency* table) that shows production classified by three of the variables: it collapses the data by combining the different dates and adding the production values by *day of the week*, *production line* (A, B, C, D) and *supervisor*. The table also provides subtotals.

The table in Figure 9-2 was generated using the EXCEL Insert | PivotTable function accessible (Figure 9-3).

Figure 9-2. Cross-tabulation of production data showing breakdown by Day of Week and Supervisor.



Sum of TOTAL	Column Labels ▾				
Row Labels ▾	Alice	Bob	Charlie	Darlene	Grand Total
⊟SUN			37,540	89,818	127,358
A			12,255		12,255
B				37,164	37,164
C			25,285		25,285
D				52,654	52,654
⊟MON	66,969	37,542			104,511
A	33,855				33,855
B	33,114				33,114
C		19,708			19,708
D		17,834			17,834
⊟TUE	89,602	29,713			119,315
A		11,116			11,116
B	42,171				42,171
C		18,597			18,597
D	47,431				47,431
⊟WED	79,578	39,056			118,634
A	35,004				35,004
B		14,658			14,658
C	44,574				44,574
D		24,398			24,398
⊟THU	44,146	25,500	24,297	52,724	146,667
A			24,297		24,297
B	44,146				44,146
C				52,724	52,724
D		25,500			25,500
⊟FRI	39,517	21,210	30,798	32,240	123,765
A				32,240	32,240
B	39,517				39,517
C		21,210			21,210
D			30,798		30,798
⊟SAT			45,487	75,531	121,018
A			20,321		20,321
B				28,795	28,795
C			25,166		25,166
D				46,736	46,736
Grand Total	319,812	153,021	138,122	250,313	861,268

Figure 9-3. Insert | PivotTable symbol.

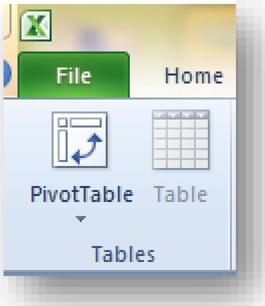
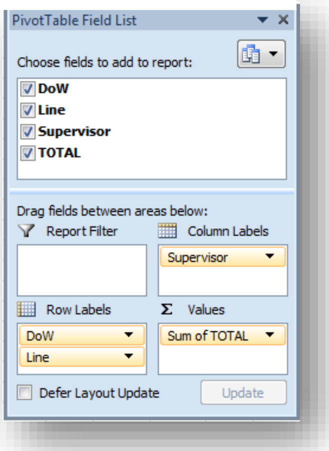


Figure 9-2 was generated using the options shown in the PivotTable Field List (Figure 9-4).

Figure 9-4. PivotTable settings for three-way cross-tabulation.



PivotTable Field List

Choose fields to add to report:

☑ DoW
☑ Line
☑ Supervisor
☑ TOTAL

Drag fields between areas below:

Report Filter

Column Labels: Supervisor

Row Labels: DoW, Line

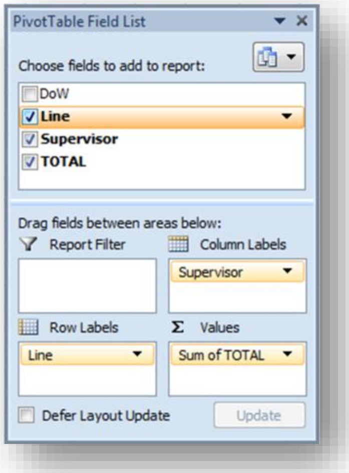
Values: Sum of TOTAL

□ Defer Layout Update

Update

Another representation summarizes production by production line and by supervisor. Figure 9-5 shows the PivotTable settings for this arrangement and Figure 9-6 shows the resulting table.

Figure 9-5.PivotTable settings for two-way contingency table.



PivotTable Field List

Choose fields to add to report:

□ DoW
☑ Line
☑ Supervisor
☑ TOTAL

Drag fields between areas below:

Report Filter

Column Labels: Supervisor

Row Labels: Line

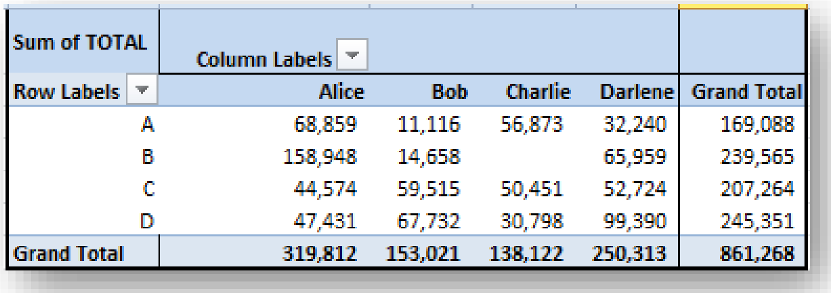
Values: Sum of TOTAL

□ Defer Layout Update

Update

This table (Figure 9-6Figure 9-5) is a *two-way* contingency table because it shows counts classified by two variables: *production line* and *supervisor*. The table includes subtotals by those variables.

Figure 9-6. Two-way contingency table for production line and supervisor.



Sum of TOTAL	Column Labels				
Row Labels	Alice	Bob	Charlie	Darlene	Grand Total
A	68,859	11,116	56,873	32,240	169,088
B	158,948	14,658		65,959	239,565
C	44,574	59,515	50,451	52,724	207,264
D	47,431	67,732	30,798	99,390	245,351
Grand Total	319,812	153,021	138,122	250,313	861,268

9.3 Filtering Data for Temporary Views

Another useful EXCEL tool for working with complex tables is the Filter function accessible through the Sort & Filter drop-down menu (Figure 9-7). Clicking on the Filter option inserts pull-down menus in the columns of the table that one has highlighted.

Figure 9-7. Sort & Filter menu.

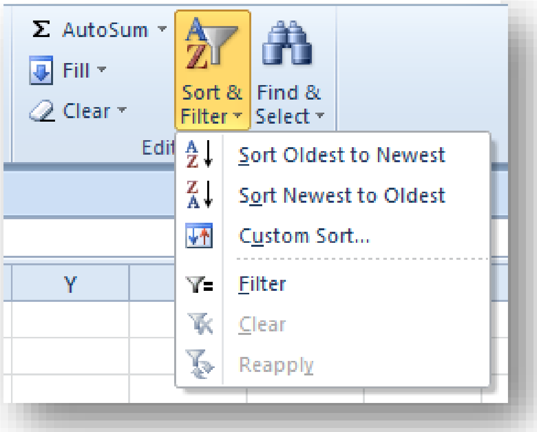
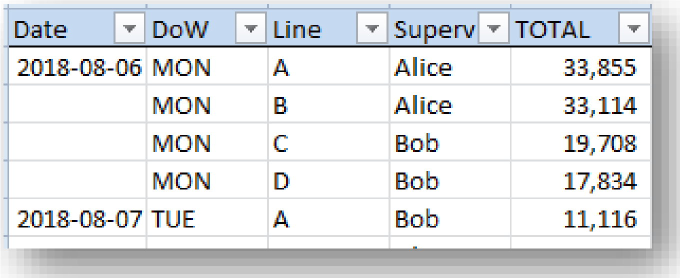


Figure 9-8 shows the headings and a few rows resulting from highlighting the entire table shown in Figure 9-1 at the start of §9.2 and selecting the Filter function.

Figure 9-8. Heading row showing pull-down menu tabs.



Date ▾	DoW ▾	Line ▾	Superv ▾	TOTAL ▾
2018-08-06	MON	A	Alice	33,855
	MON	B	Alice	33,114
	MON	C	Bob	19,708
	MON	D	Bob	17,834
2018-08-07	TUE	A	Bob	11,116

Figure 9-9 shows the pull-down menu:

Figure 9-9. Pull-down menu for DoW column.

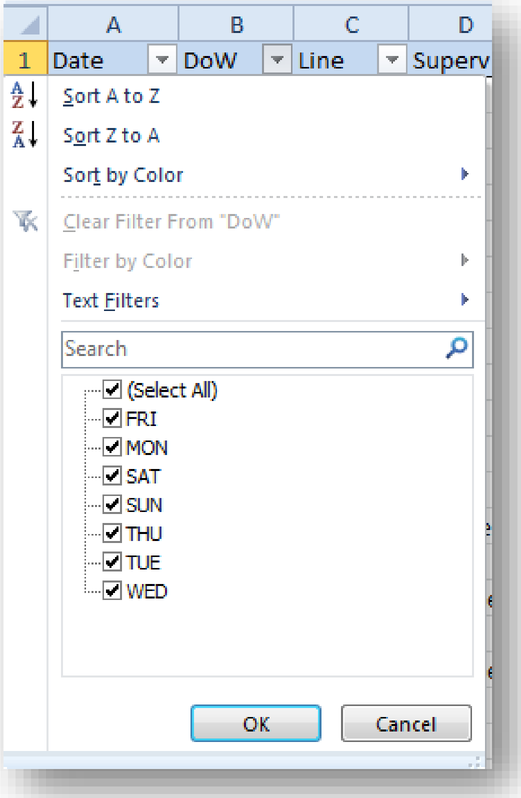
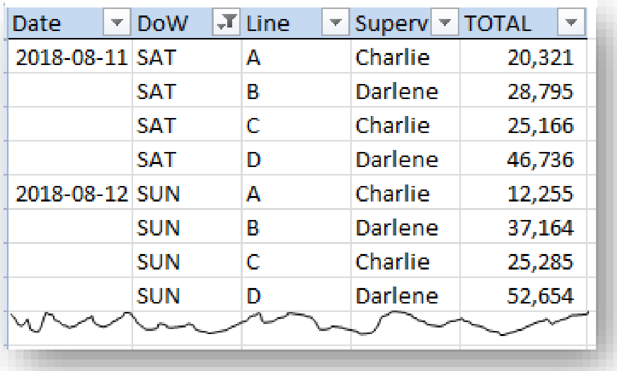


Figure 9-10 shows the results of clicking on (Select All) to remove all the check marks and then clicking only on the SAT and SUN checkboxes.

Figure 9-10. Filtered subset of table showing only SAT and SUN data.



Date ▾	DoW	Line ▾	Superv ▾	TOTAL ▾
2018-08-11	SAT	A	Charlie	20,321
	SAT	B	Darlene	28,795
	SAT	C	Charlie	25,166
	SAT	D	Darlene	46,736
2018-08-12	SUN	A	Charlie	12,255
	SUN	B	Darlene	37,164
	SUN	C	Charlie	25,285
	SUN	D	Darlene	52,654

9.4 Charts for Contingency Tables

Two-way contingency tables can be represented graphically using a variety of charts. A popular graph type is the vertical bar chart with one variable represented by position along the abscissa and the other by the position (and usually color) of a bar in a series clustered over each value of the abscissa variable. In Figure 9-11, the abscissa shows the days of the week and the clustered bars represent the production lines. Such charts are easily created in EXCEL.

Figure 9-11. Clustered bar chart showing production totals for day of week and production line

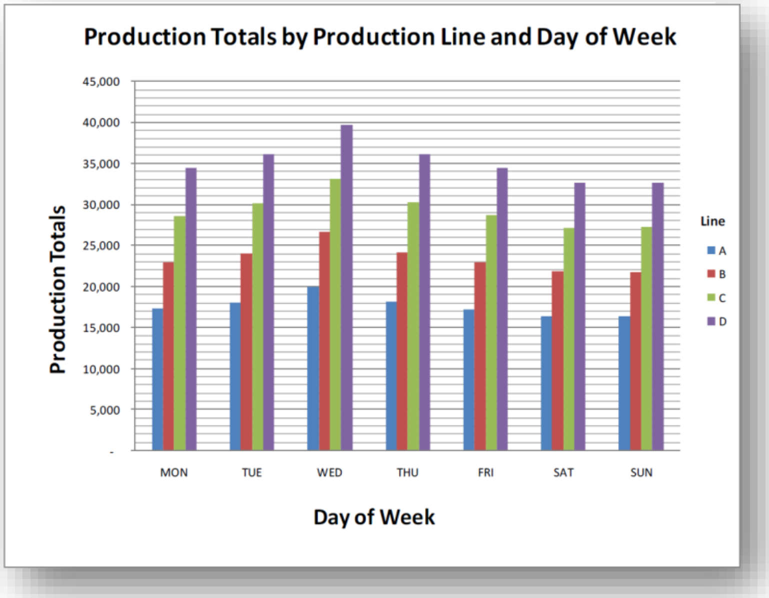
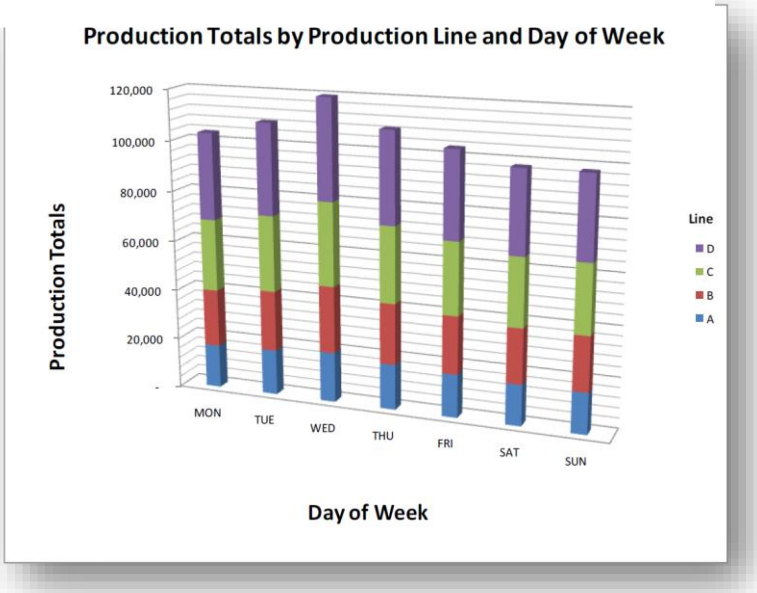


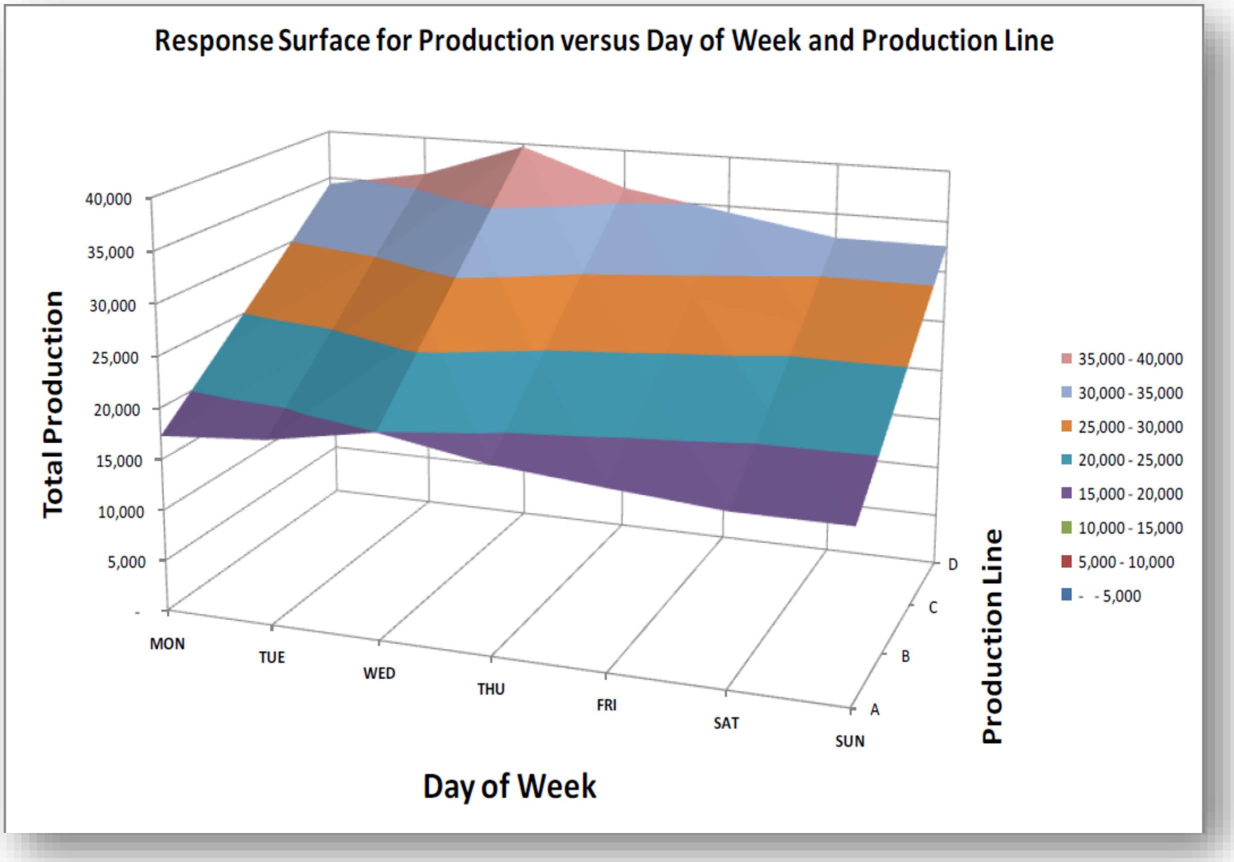
Figure 9-12. Stacked bar chart for production totals by day of week.



Another version of these data is the stacked bar chart (Figure 9-12), in which the second variable has portions of a single bar representing the total for all of its values. This particular example also uses features of EXCEL such as the ability to tilt and rotate three-dimensional graphs.

Another useful representation is the response surface, which shows a two-dimensional surface combining information from two variables versus a third. Figure 9-13 is a response surface for production versus both day of week and production line. Such charts can easily be prepared in EXCEL and there are many options for enhancing their appearance. In particular, one can rotate the image on any axis to clarify relationships that are of interest.

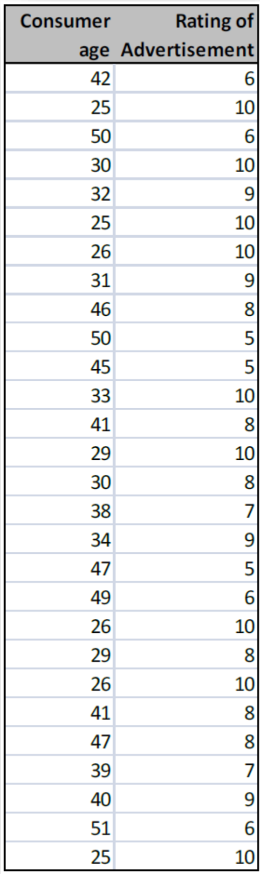
Figure 9-13. Three-variable response-surface chart.



9.5 Scatterplots and the Intuitive Grasp of Relationships

Often we measure two quantitative variables for the same entities and want to see how the data fall out in general terms. For example, Figure 9-14 shows data about the rating of a new advertising campaign by people of different ages as part of a marketing study to see if the advertisements appeal differently to viewers as a function of their age.

Figure 9-14. Data on responses to ads and age of respondent.



Consumer age	Rating of Advertisement
42	6
25	10
50	6
30	10
32	9
25	10
26	10
31	9
46	8
50	5
45	5
33	10
41	8
29	10
30	8
38	7
34	9
47	5
49	6
26	10
29	8
26	10
41	8
47	8
39	7
40	9
51	6
25	10

Figure 9-15. Scatterplot of rating vs age of respondent.

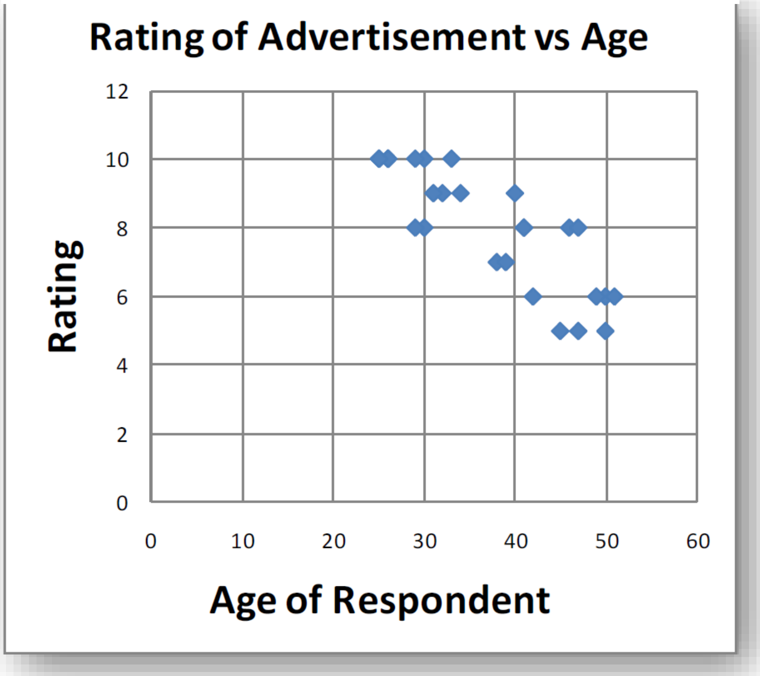


Figure 9-15 shows the scatterplot derived from these marketing-research data. Each point represents the particular combination of consumer age and rating of the advertisements for a specific respondent. These figures are easy to create in EXCEL.

At an intuitive level, a scatterplot in which the dots form a *slanted, tight* cloud naturally suggests to even a naïve viewer that perhaps there is some sort of relationship between the two variables shown. In Figure 9-15, for example, the immediate impression is that perhaps the ratings fall as the age of the respondent rises.

Contrariwise, a formless, wide cloud of dots in a scattergram suggests that perhaps there is no relation between the variables. However, without rigorous analysis, such impressions remain just that – impressions rather than actionable conclusions. For real-world decisions, we need to have quantitative estimates of the strength of the relationships and the probability that we are seeing the results of raw chance in the random sampling.

Proper statistical analyses of these impressions can involve *correlation* and *regression*, depending on the assumptions of the analysis. Correlation gives us an estimate of the *strength* of the relationship, if any; regression gives us estimates of the precise quantitative nature of the relationship, if any.

9.6 Pearson Product-Moment Correlation Coefficient, *r*

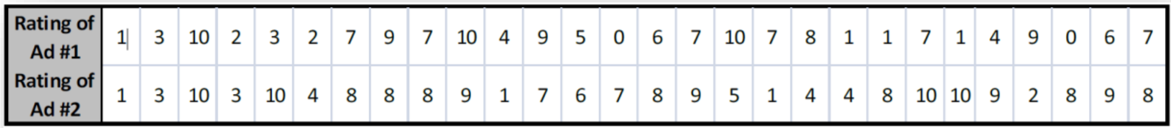
The intensity of the relationship between two variables that are both measured independently, and for which neither is viewed as a predictor of the other, is measured by an important statistic called the *Pearson product-moment correlation coefficient*, *r*.[97] This statistic applies to Normally distributed interval scales; there are other forms of correlation coefficient suitable for ordinal scales.

The notion of *independence* is important because it can determine whether to represent the relation between two variables in a data set in terms of the *correlation coefficient*, which implies no predictive rule and treats both variables as equally free to vary, in contrast with the *regression coefficient* (discussed in the next section) which usually assumes that one of the variables (the *independent variable*) can be used as a predictor of the other (the *dependent variable*).

In the example shown in Figure 9-14 and Figure 9-15, on the previous page, we collected data about the *age of respondents* and about their *rating of an advertisement*. Intuitively, we wouldn't expect anyone to be interested in predicting the age of a respondent by asking them how they feel about an ad; however, it could mean a lot to be able to measure the strength of relationship between how people feel about an ad based on their age if one wants to reach a particular demographic slice of the population. Judging from the scattergram in Figure 9-15, it looks roughly as if the older people liked the ad being tested less than the younger people. The natural, spontaneous tendency is to put the age on the abscissa and the response to the ad on the ordinate; that's a typical arrangement: the independent variable goes on the X-axis and the dependent variable goes on the Y-axis. Reversing the two would look funny in this case, although not necessarily in other cases.

Consider now a study of the responses to two different advertisements shown in Figure 9-16. Each vertical pair (e.g., 1, 1) represents the rating by an individual of their response to each of two ads.

Figure 9-16. Ratings of two different ads by 28 people.



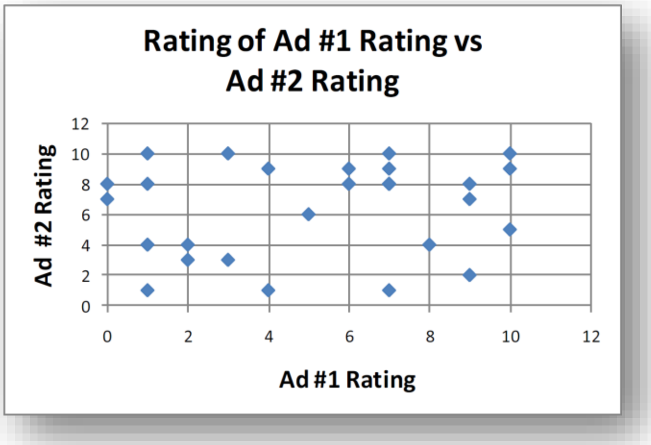
Rating of Ad #1	1	3	10	2	3	2	7	9	7	10	4	9	5	0	6	7	10	7	8	1	1	7	1	4	9	0	6	7
Rating of Ad #2	1	3	10	3	10	4	8	8	8	9	1	7	6	7	8	9	5	1	4	4	8	10	10	9	2	8	9	8

In this case, there's no particular reason why we would assign priority to one ad or the other; both are what we call *independent variables*. We can still be interested in the intensity of the relationship – the association – between these responses, but we don't normally think in terms of predicting one from the other.

Figure 9-17 shows a scatterplot for the responses to the two ads. The artificial data in this example were constructed to be randomly related – and the scatterplot shows a typical appearance for such data, with no obvious pattern.

The computation of the product-moment correlation coefficient, *r*, is best left to the statistical package.

Figure 9-17. Scatterplot for paired responses to two ads.



[97] Karl Pearson (1857-1936) was one of the founders of modern statistics. (O'Connor and Robertson 2003)

EXCEL has a function for computing r: =CORREL(array1, array2) which instantly provides the coefficient. In our current example, r = 0.0.156246 ≈ 0.156.

Notice that in this case, it doesn't matter which variable is selected for which axis; we say that these are *two independent variables*.

The correlation coefficient r has properties that are easy to understand:

- Two variables with *no relationship* between them at all have a correlation coefficient *r = 0*.
- Two variables in which a larger value of one variable is *perfectly* associated with a correspondingly larger value of the other have an r = +1. E.g., if we calculate *r* for the height of individuals measured in inches and then measured in centimeters, the data should have r = 1 because knowing one measurement should allow computation of the other measurement without error.
- If a larger value of one variable is *perfectly* associated with a systematically smaller value of the other, the r = -1. For example, imagine giving children bags of candies containing exactly 20 candies to start with; in this silly situation, calculating the *r* for the number of candies eaten and the number of candies left should produce r = -1 because the more candies are eaten, the fewer are left – and there should be zero error in the data.
- In both of these extremes, knowing the value of one of the variables allows perfect computation of the value of the other variable. We say that there is no unexplained error in the relationship. However, if we introduce errors, r will decline from +1 or increase from -1. For example, if we measure the height of a subject with a laser system correct to 0.01cm but measure the height in inches using a piece of string with only whole numbers of inches marked on the string, it's likely that the correlation coefficient will be less than perfect. Similarly, if a three-year-old is counting the candies that are left (and perhaps eating some surreptitiously), the data for the candy correlation may very well produce an *r* > -1.

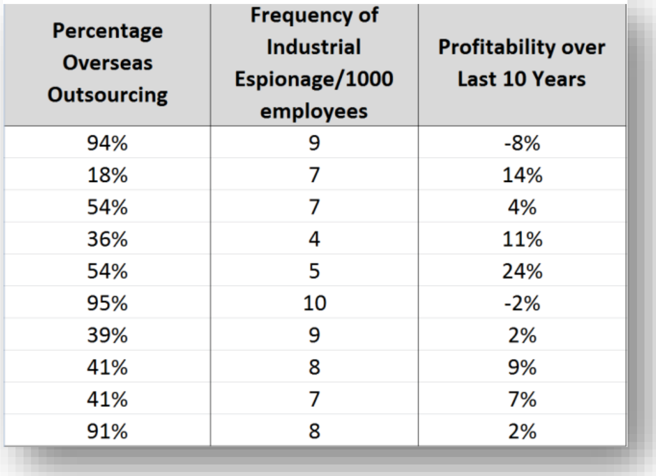
9.7 Computing the Correlation Coefficient Using EXCEL

As mentioned in the previous section, the function =CORREL(array1, array2) mentioned in the previous section computes *r* with no additional information and no options.

However, the Data Analysis function Correlation offers more options than the simple function when we need to compute r for more than one pair of variables. It's especially useful because it calculates correlations for all possible pairs of the data.

For example, imagine that a study (Figure 9-18) of the percentage of foreign outsourcing for ten companies is tested for a possible correlation with the frequency of industrial espionage expressed in relation to the number of employees in the company. In addition, the researchers also record the percentage profitability over the past decade for each company.[98]

Figure 9-18. Outsourcing, espionage and profitability.

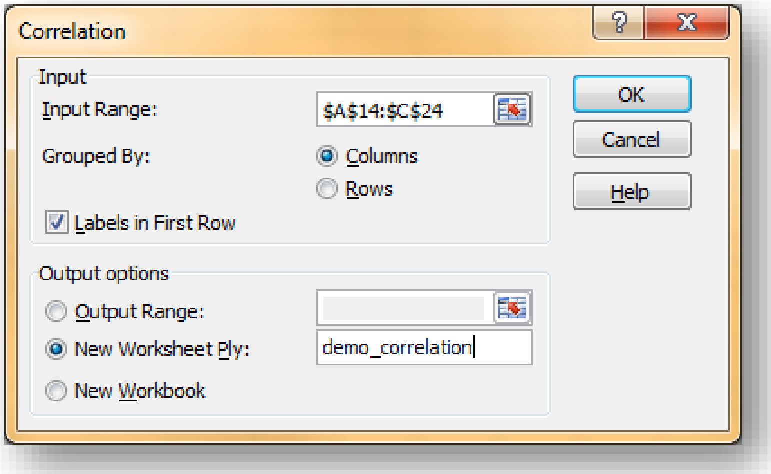


Percentage Overseas Outsourcing	Frequency of Industrial Espionage/1000 employees	Profitability over Last 10 Years
94%	9	-8%
18%	7	14%
54%	7	4%
36%	4	11%
54%	5	24%
95%	10	-2%
39%	9	2%
41%	8	9%
41%	7	7%
91%	8	2%

[98] These are *made-up figures*: do not interpret them as indicating anything real – they are created only for the purpose of illustration in this text.

Using the Correlation function from the Data Analysis menu (Figure 9-19),

Figure 9-19. Entering ranges and options into Correlation menu.



Correlation

Input

Input Range: $A$14:$C$24

Grouped By: ● Columns ○ Rows

☑ Labels in First Row

Output options

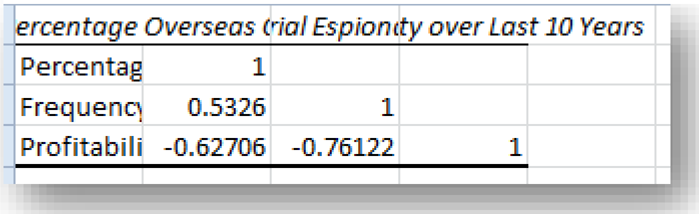
○ Output Range:

● New Worksheet Ply: demo_correlation

○ New Workbook

we generate a simple table of results showing the correlation coefficients *r* for each pair of variables in this made-up example. The default output lacks effective formatting, as you can see in Figure 9-20, where the columns are too narrow to display the column headings.

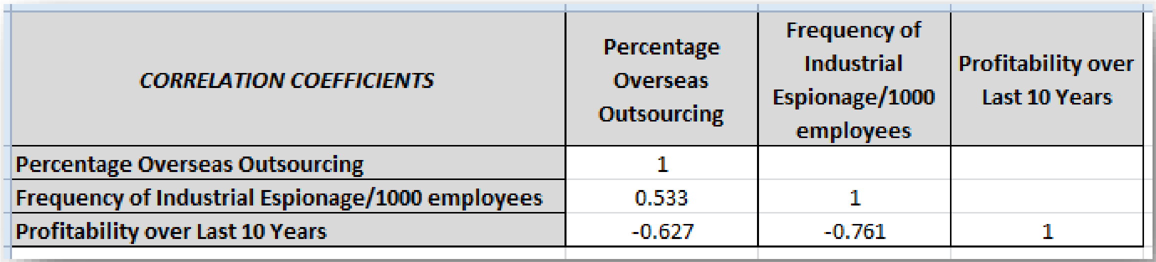
Figure 9-20. Unformatted output from Data Analysis | Correlation tool.



	ercentage Overseas (	rial Espiona	ity over Last 10 Years
Percentag	1		
Frequency	0.5326	1	
Profitabili	-0.62706	-0.76122	1

However, some simple formatting (e.g., highlighting the columns and double-clicking any of the column boundaries – and the same for the heading row) produces a more usable report (Figure 9-21).

Figure 9-21. More readable formatted output from Data Analysis | Correlation tool.



CORRELATION COEFFICIENTS	Percentage Overseas Outsourcing	Frequency of Industrial Espionage/1000 employees	Profitability over Last 10 Years
Percentage Overseas Outsourcing	1		
Frequency of Industrial Espionage/1000 employees	0.533	1	
Profitability over Last 10 Years	-0.627	-0.761	1

For example, the correlation coefficient *r* for *Percentages Overseas Outsourcing* and *Frequency of Industrial Espionage/1000 Employees* is 0.533 and *r* for the *Outsourcing* percentage and *10-year profitability* is about -0.63.

9.8 Testing the Significance of the Correlation Coefficient

How do we know if a sample's correlation coefficient (*r*) is consistent with the hypothesis that the parametric correlation (*ϱ*)[99] has a particular value? As usual, it's possible to compute a test statistic based on a correlation coefficient (*r*) based on a sample of size *n* that is distributed as a Student's-t statistic:

$$t_{[\nu]} = (r - \varrho)/s_r$$

where

$t_{[\nu]}$ = the test statistic with $\nu = n - 2$ degrees of freedom

s_r = standard error of the correlation coefficient:

$$s_r = \sqrt{(1 - r^2)/(n - 2)}$$

Thus

$$t_{(n-2)} = (r - \rho)/\sqrt{(1 - r^2)/(n - 2)}$$

If our hypotheses are H0: $\rho = 0$ and H1: $\rho \neq 0$, then

$$t_{(n-2)} = r/\sqrt{(1 - r^2)/(n - 2)}$$

In the example discussed in §9.7, n = 10 and the correlation coefficient for overseas outsourcing (*o*) and industrial espionage (*e*) was $r_{oe} = 0.533$.

The test for correlation between outsourcing and espionage is thus

$$t_{oe[8]} = 0.533 /[\ \sqrt{}(1 - 0.533^2)/8] = 5.034$$

and the calculation of the two-tailed probability that the null hypothesis is true is

=T.DIST.2T(t_{oe},8) = 0.001***

which is extremely significant. We can reasonable reject the null hypothesis; the positive correlation between overseas outsourcing and industrial espionage appears to be real.[100]

The negative correlations (outsourcing and profitability; espionage and profitability) need to be converted using the absolute value function (ABS). The t-tests are

$$t_{op[8]} = -0.627 /[\ \sqrt{}(1 - 0.627^2)/8] = -5.927$$

and the calculation of the two-tailed probability that the null hypothesis is true is

=T.DIST.2T(ABS(t_{op}),8) = 0.0004***

so the negative correlation between outsourcing and profitability is extremely significant too.

Finally, the correlation between industrial espionage and profitability, -0.761, has a t_{ep} = -0.761 with p(H0) = 0.0001***. So that correlation is extremely significant, too.

[99] The Greek symbol is *rho* which corresponds to our letter *r*. See §7.3 on page 7-3.

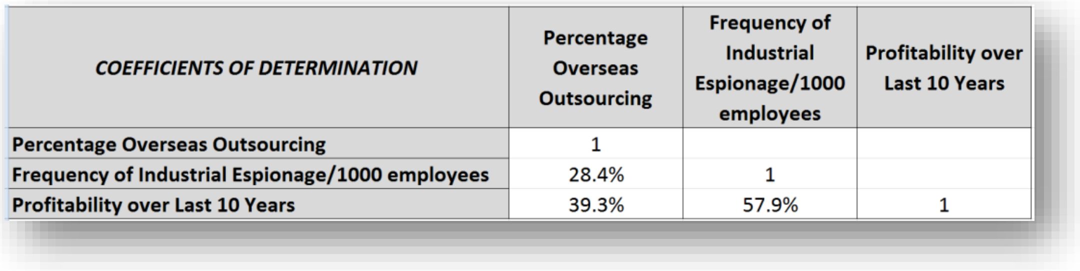
[100] Remember this is a completely made-up example! The example does *not* speak to the issue of outsourcing and espionage or anything else in the real world.

9.9 Coefficient of Determination, r^2

A valuable aspect of the correlation coefficient r is that its square, r^2, known as the *coefficient of determination*, tells us *what proportion of the variation* in one of the variables can be *explained* by the other variable. For example,

- If we learn that the correlation coefficient between the incidence of a type of hacker attack on a network and the occurrence of disk errors on the network disk drives is r = 0.9, then $r^2 = 0.81$ and we can assert that in this study, 81% of the variation in one of the variables may be *explained* or *accounted for* by variations in the other. More frequent hacker attacks are positively associated with damaged disk drives; damaged disk drives are associated with a higher frequency of hacker attacks. The 81% figure based on r^2 implies that if we were to define one of the variables as an *independent variable* and the other as a *dependent variable*, we could predict the dependent variable with about 81% of the total variance explained by knowing the value of the dependent variable and 19% left unexplained.
- In our made-up example about outourcing, espionage and profitability (§9.7), the values of the correlation coefficients can easily be squared in EXCEL to show the coefficients of determination:

Figure 9-22. Coefficients of determination.



COEFFICIENTS OF DETERMINATION	Percentage Overseas Outsourcing	Frequency of Industrial Espionage/1000 employees	Profitability over Last 10 Years
Percentage Overseas Outsourcing	1		
Frequency of Industrial Espionage/1000 employees	28.4%	1	
Profitability over Last 10 Years	39.3%	57.9%	1

Too often, you will hear a journalist or some other statistically naïve person asserting that "the correlation between A and B was 60%, which implies a strong relationship between A and B." Well, not really: $r = 0.6$ means $r^2 = 0.36$ or in other words, that only 36% of the variation in A can be explained by knowing the value of B or vice versa. All the rest of the variation is unexplained variance. Always mentally square correlation coefficients to estimate the coefficient of determination when you are told about correlations.

Two factors may be positively or negatively correlated because they are *both* determined to some extent by a third, unmeasured factor. Keep in mind is that *correlation does not imply or prove causation* in either direction. For example,

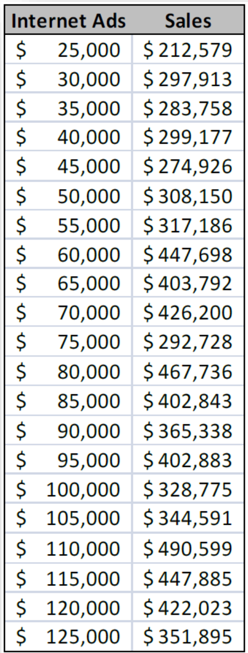
- Just because weight is correlated with age does not mean that weight determines age – or, for that matter, that age determines weight.
- In the outsourcing/espionage/profitability model, there is no implication of causality one way or another.
- And although studies may find a positive correlation between playing violent video games and having aggressive thoughts, the correlation does not mean that playing violent games necessarily causes the increase in aggressivity or that increased aggressivity causes an increase in playing violent video games.[101]

[101] (Anderson and Dill 2000)

9.10 Linear Regression in EXCEL

Sometimes one of the variables in a two-variable data set has been deliberately chosen (the *independent variable*) and the other varies without imposed controls (the *dependent variable*). For example, Figure 9-23 shows the results of an study of the amount of money spent in a month on Internet advertising at Urgonian Corporation and the corresponding monthly sales of the product advertised in the year 2125.

Figure 9-23. Internet ads and sales at Urgonian Corporation in 2125.

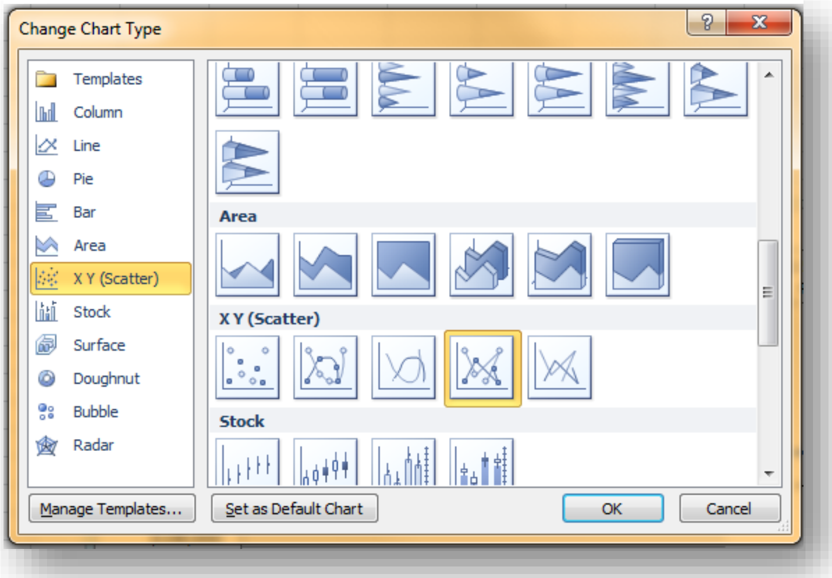


Internet Ads	Sales
$ 25,000	$ 212,579
$ 30,000	$ 297,913
$ 35,000	$ 283,758
$ 40,000	$ 299,177
$ 45,000	$ 274,926
$ 50,000	$ 308,150
$ 55,000	$ 317,186
$ 60,000	$ 447,698
$ 65,000	$ 403,792
$ 70,000	$ 426,200
$ 75,000	$ 292,728
$ 80,000	$ 467,736
$ 85,000	$ 402,843
$ 90,000	$ 365,338
$ 95,000	$ 402,883
$ 100,000	$ 328,775
$ 105,000	$ 344,591
$ 110,000	$ 490,599
$ 115,000	$ 447,885
$ 120,000	$ 422,023
$ 125,000	$ 351,895

The sales figures are not usually directly under our control (assuming that we aren't selling out our entire production every week) but the amount we spend on advertising is under our control (assuming our marketing manager is not using a Ouija board to determine the spending). This situation is a classic *Model I regression* case in which the X variable – the independent variable – will be the advertising budget and the Y value – the dependent variable – will be the sales figures.

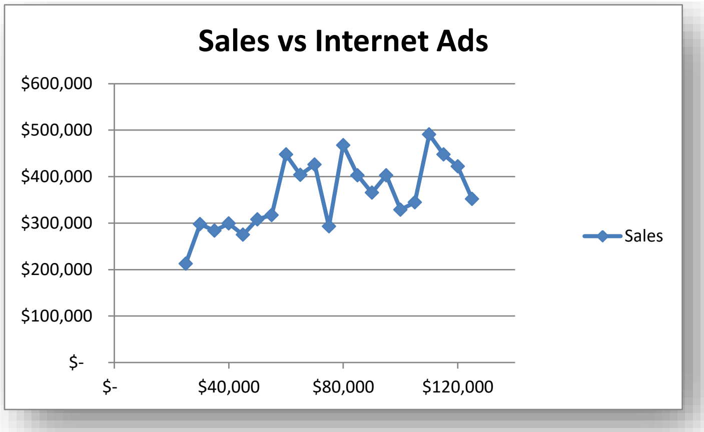
In graphing these data, one can choose the Insert | Scatter option:

Figure 9-24. Creating a scatterplot.



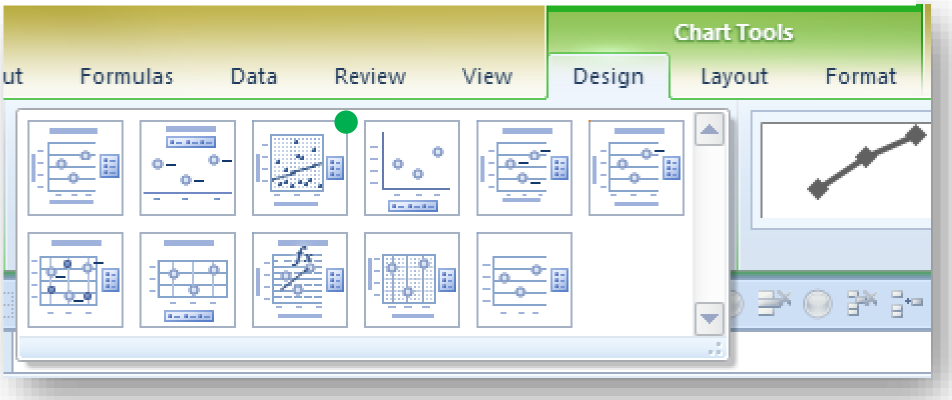
This selection generates a simple XY chart (Figure 9-26) which we can then modify to include a regression line and regression equation:

Figure 9-25. Simple regression without line..



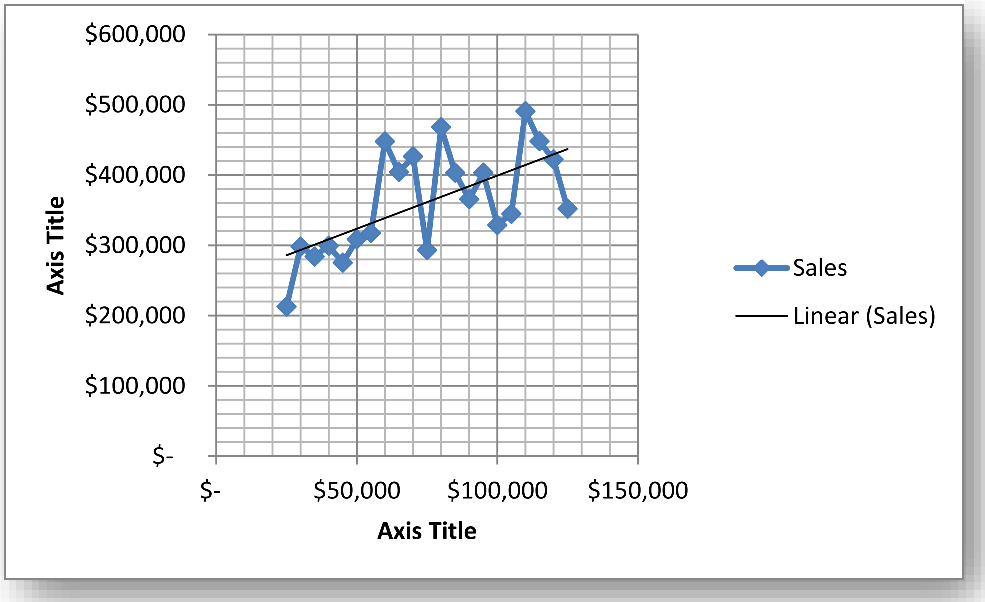
Click on the chart and then use Chart Tools | Design to select a version of the chart that shows a linear regression (top row, highlighted with green dot in Figure 9-27):[102]

Figure 9-27. Choosing Layout 3 to show regression line in existing graph.



Instantly, the graph is converted to the form shown in Figure 9-28:

Figure 9-28. Conversion to *Type 3* XY plot showing regression line added to raw data.



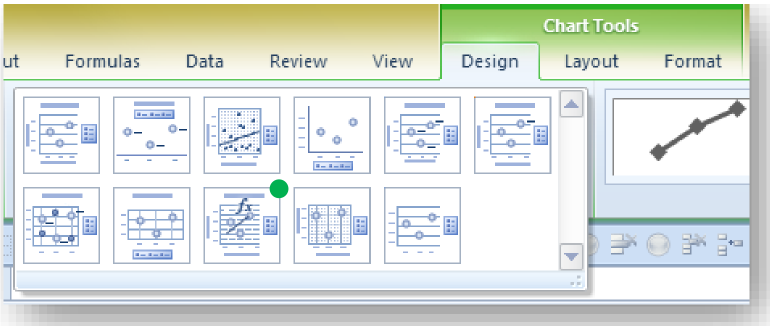
The chart needs additional work, such as labeling the axes and adding a title, but it's a quick and easy way to show the linear regression line without additional computation.

But what if we want to see the regression equation? We have another option.

[102] The green dot does not appear in Excel; it was added in creating the figure.

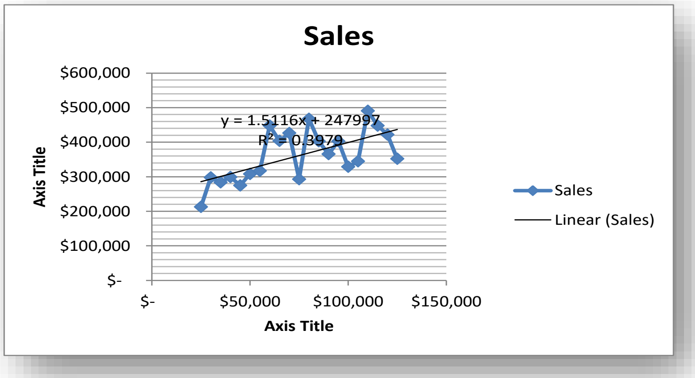
Using Layout 9, shown with the added green dot in Figure 9-29, we can instantly generate a graph that includes the regression equation and the coefficient of determination for the data:

Figure 9-29. Choosing regression line with equation for graph.



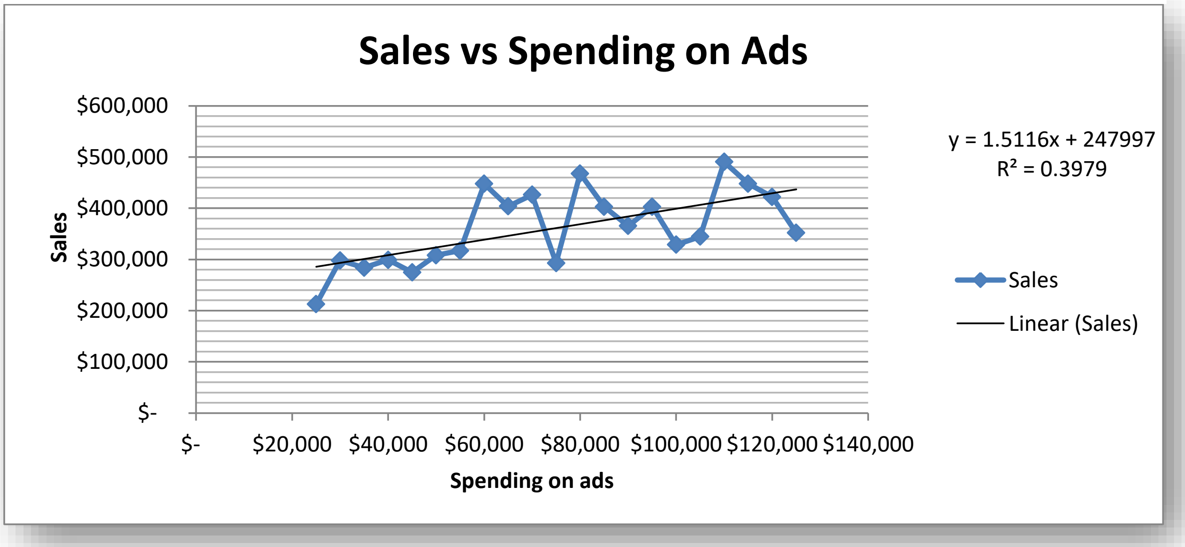
The results (Figure 9-31) are a start:

Figure 9-31. Initial chart produced using Layout 9.



Moving the equation to a more readable area of the chart, stretching the chart sideways, and adding or improving titles, we end up with a presentable representation (Figure 9-30) that includes the regression equation and the coefficient of determination ("R^2"):

Figure 9-30. Chart modified for better readability and with axis labels and better title.



9.11 ANOVA with Linear Regression

Typically, we represent the *best-fit linear regression model* as

$$\hat{Y} = a + bX + \varepsilon$$

where

$\hat{Y}$ is the estimated value of the dependent variable for a given independent value X;

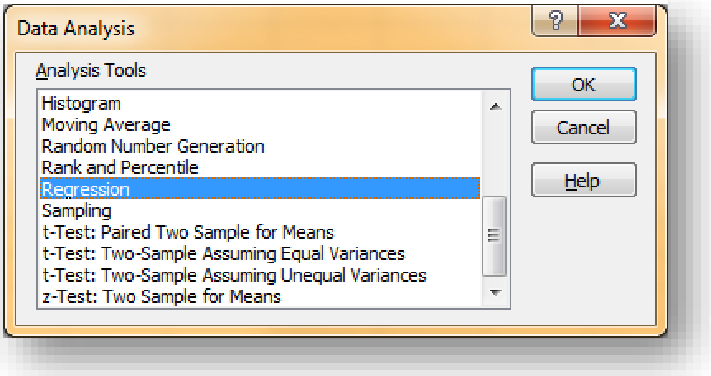
a is the Y-intercept, or the value of Y for X = 0;

b is the regression coefficient, or the amount Y rises for a unit increment in X;

ε is the residual error, also called the unexplained error – a measure of the average (squared) difference between the predicted values and the observed values.

Figure 9-32 shows the pop-up panel for Regression in the Data Analysis section of EXCEL 2010.

Figure 9-32. Selecting the Regression tool in Data Analysis



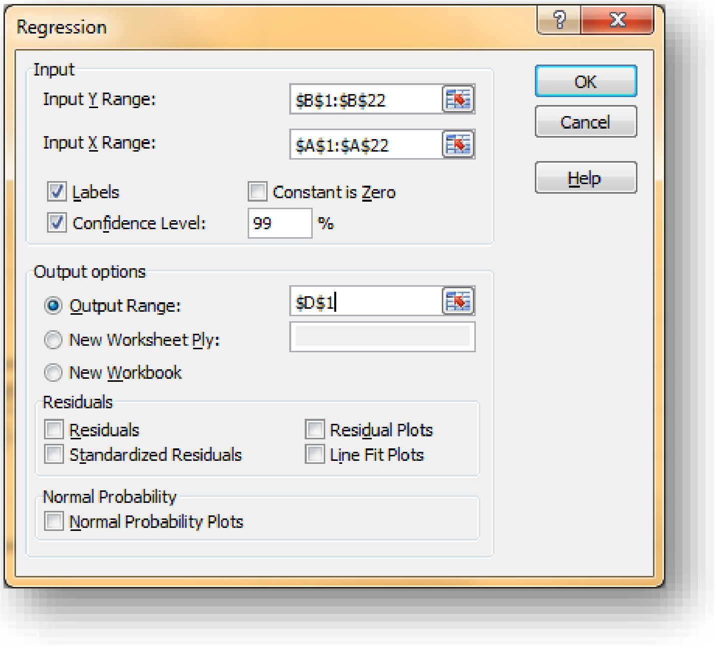
Data Analysis

Analysis Tools

- Histogram
- Moving Average
- Random Number Generation
- Rank and Percentile
- Regression
- Sampling
- t-Test: Paired Two Sample for Means
- t-Test: Two-Sample Assuming Equal Variances
- t-Test: Two-Sample Assuming Unequal Variances
- z-Test: Two Sample for Means

For this demonstration, the Regression menu includes the following settings:

Figure 9-33. Regression menu settings for demonstration.



Regression

Input

Input Y Range: $B$1:$B$22

Input X Range: $A$1:$A$22

☑ Labels □ Constant is Zero

☑ Confidence Level: 99 %

Output options

● Output Range: $D$1

○ New Worksheet Ply:

○ New Workbook

Residuals

□ Residuals □ Residual Plots

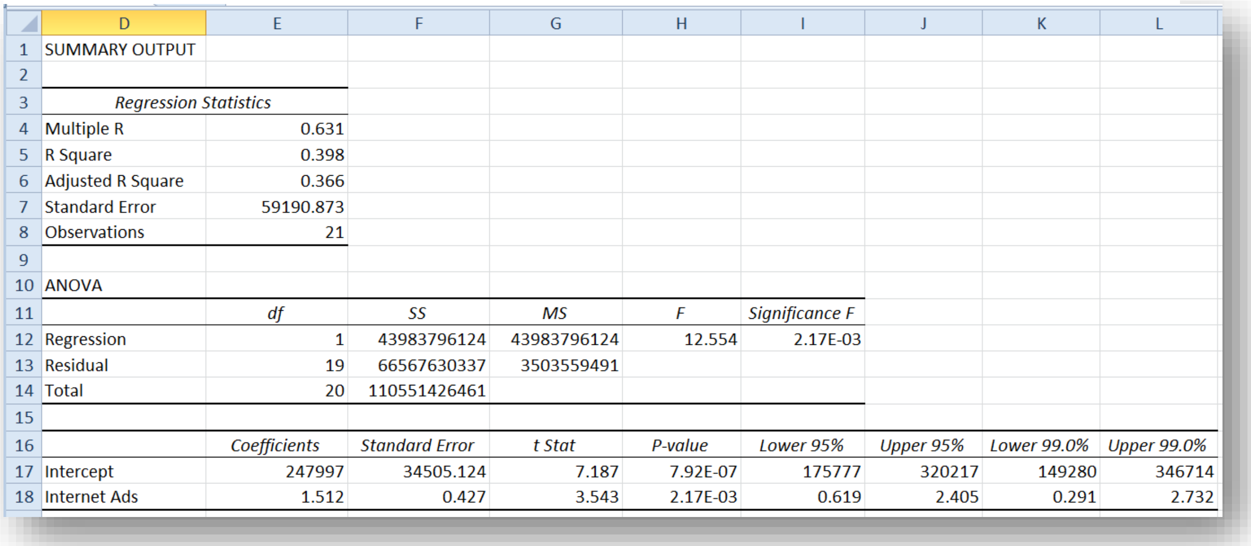
□ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

Figure 9-34 shows the results of the Regression function, including the *ANOVA with linear regression* table (labeled simply ANOVA in the figure).

Figure 9-34. ANOVA with linear regression for Sales vs Advertising data.



	D	E	F	G	H	I	J	K	L
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.631							
5	R Square	0.398							
6	Adjusted R Square	0.366							
7	Standard Error	59190.873							
8	Observations	21							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	43983796124	43983796124	12.554	2.17E-03			
13	Residual	19	66567630337	3503559491					
14	Total	20	110551426461						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
17	Intercept	247997	34505.124	7.187	7.92E-07	175777	320217	149280	346714
18	Internet Ads	1.512	0.427	3.543	2.17E-03	0.619	2.405	0.291	2.732

The regression equation coefficients are

a = 247,997 (the Intercept shown in the bottom table in cell E17) and

b = 1.51 (the coefficient called “Internet Ads” positioned in cell E22) so the predictive equation (without the error term) is

$\hat{Y} = 247{,}997 + 1.51X$

In the Regression Statistics section of the output in Figure 9-34, the key statistic of interest in this course is *R Square* – the coefficient of determination, r^2 (§0) – which is the variability explained by the regression (Regress SS) as a proportion of the total variability (Total SS): 0.398 or 39.8%.

The ANOVA section shows us that $F_{[1,19]} = MS_{regression} / MS_{residual} = 12.554$**. The regression is highly significant (p(H0: ρ=0) = 0.00217) at the 0.01 level of significance.

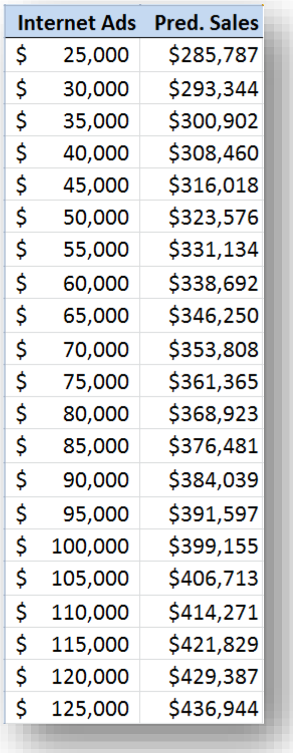
The last section of the output includes Coefficients: the Y-intercept (a) = 247,997. That represents the predicted sales with zero Internet ads. The 95% confidence limits for a are 175,777 and 320,217. Because the menu in Figure 9-33 includes specification of 99% confidence limits, those are also provided (149,280 and 346,714). These values represent the estimate of how much in sales would occur with zero Internet ads.

The slope (b) is 1.512 (the Coefficient in the second row, listing the statistics for Internet Ads). The P-value is exactly what is shown in cell I12 of the ANOVA table (2.17E-03) and the confidence limits for b are also shown: 0.619 and 2.405 for the 95% confidence limits; 0.291 and 2.732 for the 99% confidence limits. These values represent the change in expected sales as a proportion of expenditures in Internet ads. Thus at the point estimate $b = 1.512$, the model predicts that every expenditure of a dollar in Internet ads will generate sales of $1.512 or 151.2% return on investment. On the other hand, the confidence limits also warn that the uncertainty left in the regression (r^2 = 39.8%) means that it is also possible that the return on investment (slope) could be as low as 0.291 or 29.1%. This result indicates that there is a 99% probability that we are correct in asserting that the return on investment in Internet ads will meet or exceed 29.1%.

9.12 Predicted Values in Linear Regression & Confidence Limits

It's easy to generate the predicted values of Y for given values of X by plugging the X-values into the best-fit linear regression equation. Figure 9-35 shows the predicted values of sales for the original list of Internet Ad expenditures (Figure 9-23) using Y-intercept *a* ($247,997) and slope *b* (1.512) calculated by the Data Analysis | Regression tool.

Figure 9-35. Predicted sales as function of Internet ad expenditures.



Internet Ads	Pred. Sales
$ 25,000	$285,787
$ 30,000	$293,344
$ 35,000	$300,902
$ 40,000	$308,460
$ 45,000	$316,018
$ 50,000	$323,576
$ 55,000	$331,134
$ 60,000	$338,692
$ 65,000	$346,250
$ 70,000	$353,808
$ 75,000	$361,365
$ 80,000	$368,923
$ 85,000	$376,481
$ 90,000	$384,039
$ 95,000	$391,597
$ 100,000	$399,155
$ 105,000	$406,713
$ 110,000	$414,271
$ 115,000	$421,829
$ 120,000	$429,387
$ 125,000	$436,944

Suppose we want to estimate the sales predicted for expenditures of $150,000 on Internet ads. We calculate

$$\hat{Y} = \$247{,}997 + 1.512 * \$150{,}000 = \$474{,}734$$

A more involved calculation is to compute the upper and lower (1 – α) confidence limits for a predicted Y ($\hat{Y}$) for a given X (symbolized by X_i). This measure of uncertainty is smallest at the center of the regression, where the selected X_i is the mean, $\bar{X}$. As X_i moves further away from the mean, the uncertainty of the prediction increases.

The standard error of $\hat{Y}$ is symbolized $s_{\hat{Y}}$ and is a function of the given value of X_i. it is defined as follows:

$$s_{\hat{Y}} = \sqrt{MS_{residual}\left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{ns_x^2}\right]}$$

- $MS_{residual}$ is the error mean square from the ANOVA;
- n is the sample size;
- X_i is the specific value of the independent variable, X for which we want to computer the predicted value $\hat{Y}$ and its confidence limits;
- $\bar{X}$ is the mean of X;
- s^2_x is the variance of the values of X in the dataset; can be calculated using =VAR.P(range); unusually, we are using VAR.P instead of VAR.S because it provides the value needed in the computation.

Our example has the following values for these components in the ANOVA with regression for our example:

- $MS_{residual}$ = 3,503,559,491
- n = 21
- X_i = 150,000
- $\bar{X}$ = 75,000
- s^2_x = 916,666,667

The calculation yields $s_{\hat{Y}}$ = 34,505.

The distribution of $\hat{Y}$ follows Student's-t with ν = n – 2 degrees of freedom.

Computation of the confidence limits for $\hat{Y}$ can use the =CONFIDENCE.T(alpha, standard_dev, size) function from EXCEL 2010. The only wrinkle is that =CONFIDENCE.T is defined for computing confidence limits of the *mean*, and it therefore assumes that the degrees of freedom are size – 1. Because the degrees of freedom of $s_{\hat{Y}}$ are $\nu = n - 2$, we have to trick the function by entering the size parameter as n – 1 to force the function to use $\nu = n - 2$ for its calculation.

The =CONFIDENCE.T value is one-half the confidence interval. Thus lower (L_1) and upper (L_2) $(1 - \alpha)$ confidence limits are

$$L_1 = \hat{Y} - \text{CONFIDENCE.T}(\alpha, s_{\hat{Y}}, n\text{-}1)$$

$$L_2 = \hat{Y} + \text{CONFIDENCE.T}(\alpha, s_{\hat{Y}}, n\text{-}1)$$

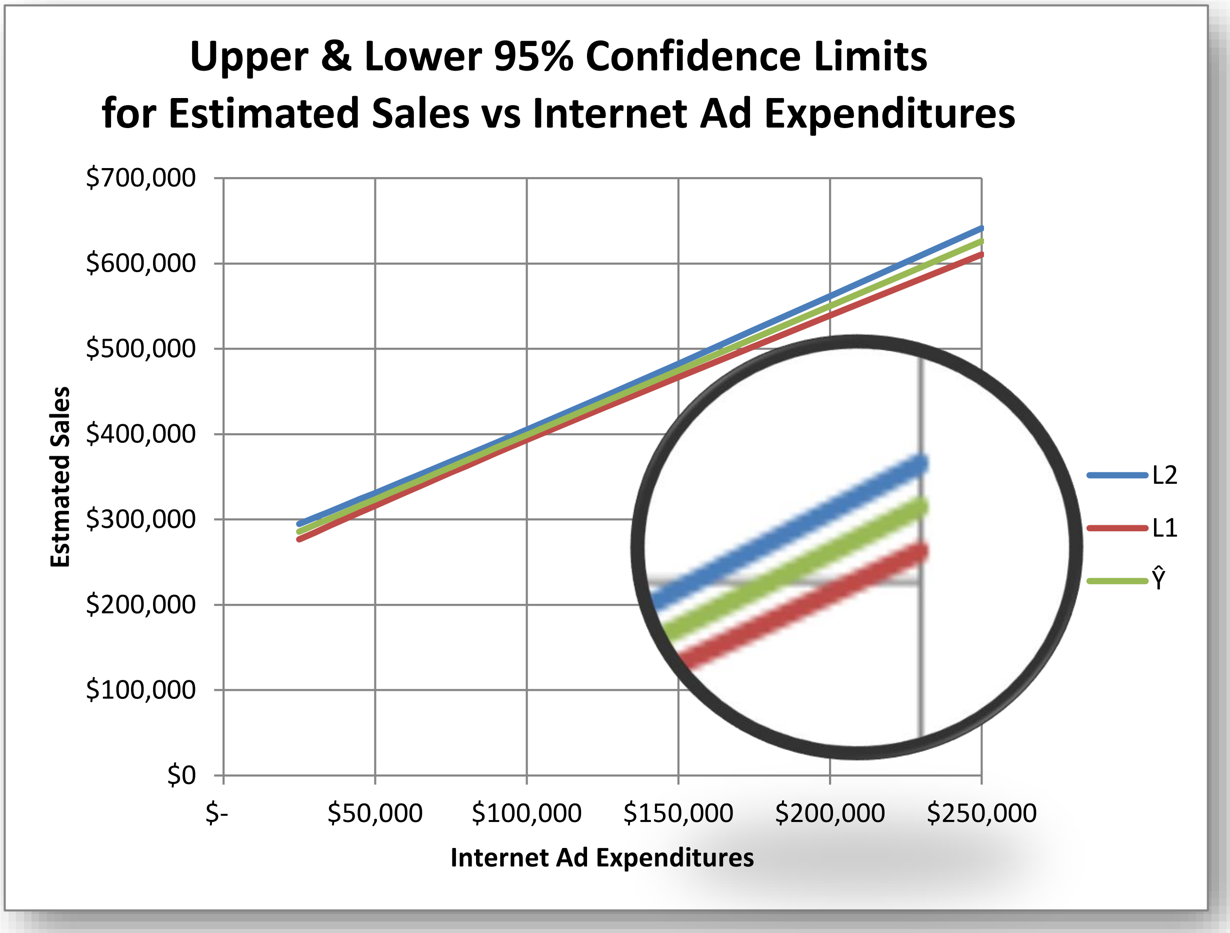
In our example, the 95% confidence limits for the estimated value $\hat{Y}$ = \$474,734 for X_i = \$150,000 are

L_1 = \$458,585

L_2 = \$490,883

After calculating the upper and lower 95% confidence limits for all the values in the original data chart, it's easy to create a chart illustrating the bowed (parabolic) nature of the upper and lower confidence limits. Figure 9-36shows the modest widening of the confidence limits around the estimated sales. The values include Internet Ad expenditures reaching up to \$250,000 to demonstrate the curve of the confidence limits. The circular inset expands the right-hand side of the predictions to illustrate the divergence between upper bound (blue) expected value (green) and lower bound (red).

Figure 9-36. Confidence limits get wider away from the mean.


Upper & Lower 95% Confidence Limits
for Estimated Sales vs Internet Ad Expenditures
Estmated Sales
$700,000
$600,000
$500,000
$400,000
$300,000
$200,000
$100,000
$0
$-
$50,000
$100,000
$150,000
$200,000
$250,000
Internet Ad Expenditures
L2
L1
Ŷ