STATISTICS TOPIC NOTES
Linear Regression with One Predictor Variable

Regression is the most important statistical tool for managing uncertainty about business questions. Regression is about relationships between numerical variables. For example, we can give a better estimate of the rent of an apartment if we know its area; we can better forecast sales of a product if we know how much will be spent on advertising. In order to use the area of an apartment to estimate the rent, we need to understand the relationship between area and rent; in order to use advertising to forecast sales, we need to understand the relationship between advertising and sales. If we have rent and area data on a representative sample of apartments, regression can provide an equation to estimate rent by plugging in the value of area; if we have data on sales at different levels of advertising, regression can provide an equation to estimate sales by plugging in the value of advertising.

The variable that we want to estimate (rent) or to forecast (sales) is variously called the **response** variable, the **dependent** variable, or the ***Y*** variable. The variable that we will plug in is called the **predictor** variable, the **independent** variable, or the ***X*** variable. The regression equation will be **linear** – that is, it will be a straight line of the form $y = \alpha + \beta x$. Regression is not limited to one predictor variable, nor to the linear form. For example, if y is apartment rent and x_1 is apartment area and x_2 is the number of bathrooms, the regression equation could be $y = \alpha + \beta_1 x_1 + \beta_2 x_2$ or $y = \alpha + \beta_1 x + \beta_2 x^2$ or even $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2^{\ 2}$. However, we will start with the simplest case first – one predictor and the linear form, called **simple linear regression**.[1] So we will be using straight line regression equations of the form $y = \alpha + \beta x$.

In order to make the presentation of the regression model concrete, I will use the example of Austin apartment rents at length. The *Y* variable will be monthly rent. The *X* variable will be the area of the apartment. I have a random sample of 60 Austin apartments for which both rent and area are known for each apartment. I will give you the complete dataset a bit later (see Tables 1 and 2). The objective is to develop a simple linear regression equation of the form $Rent = \alpha + \beta$ *Area*, so that the value of *Area* may be plugged into the right-hand side of the equation to get an estimate of *Rent*, once estimates of α and β have been obtained. I will discuss the regression model, how to do it in Excel, and interpretation of some of the computer output.

Let us begin.

To set the stage for regression, let me pose a *pre*-regression question and recall how we can answer it using previous Topic Notes. I will then show how regression can improve our answer.

Question: What is the mean rent of all Austin apartments?

[1] You may wonder why it is called "regression." The reason is historical and makes an interesting story, but would be tangential here. Just realize that all we mean by "regression" is to model the relationship between variables by equations.

The question asks for the population mean of an uncertainty distribution. Which distribution? Clearly, it is the distribution that has all Austin apartment rents as its population of outcomes (Population 1) and an unknown probability curve. Since we do not know either the outcomes or the probability curve, we cannot directly calculate the population mean. Therefore, we are uncertain. To reduce the uncertainty, I have drawn a random sample of 60 apartments from the unknown population of Austin apartments and obtained their rents:

Table 1. Rents of 60 randomly selected Austin apartments.

519	530	450	425	470	415	505	470	625	470	659	605
765	580	520	770	700	399	445	470	745	480	650	929
475	995	495	445	450	585	565	700	540	460	750	695
575	565	420	510	785	525	650	455	650	600	455	455
415	620	575	635	485	495	515	550	595	575	430	1050

Each of these 60 draws comes from the same unknown population of rents with the probability of selection governed by the same unknown probability curve. Therefore, the known distribution of these 60 rents should be representative of the unknown distribution of all Austin apartment rents. Therefore, the sample mean of the known 60 rents should be like the unknown population mean of all Austin apartments. Therefore, I calculate the sample mean of these 60 rents $\bar{x}$ = \$572.27 and use this value as an estimate of μ = unknown population mean rent of all Austin apartments.[2]

But the estimate, \$572.27, is almost certainly not the true value of μ. We should summarize our uncertainty about the estimate. One way to do that is to address the further question, How much can we expect \$572.27 to differ from the true mean μ? To answer that, we reason that our sample mean is a representative sample mean. So our sample mean probably differs from the true mean μ by about as much as a typical sample mean differs from μ. But how much does a typical sample mean deviate from μ? Answer: By the standard deviation of sample means – that is, by the standard deviation of Population 3, which is $\sigma/\sqrt{n}$. To calculate this, we need σ and *n*. The sample size *n* is known: *n* = 60. But σ, the standard deviation of Population 1, is not known. However, the sample of 60 rents taken from Population 1 are probably representative of Population 1. So the standard deviation *s* of the 60 rents is a good estimate of σ. We calculate *s* = \$140.52. Therefore, $\sigma/\sqrt{n}$ approximately = 140.52 / $\sqrt{60}$ = \$18.14.[3]

Therefore, we estimate the mean rent of the population of all Austin apartments to be \$572.27 and we can expect this estimate to deviate from the actual mean by about ±\$18.14. Furthermore, since the sample size 60 is "sufficiently large", the Population 3 of sample means is approximately normal. Thus, approximately 68% of all sample means are within ±\$18.14 of the true mean μ. Therefore, our sample mean \$572.27 has about 68% chance of being within

[2] The logic for this was explained in the Topic Note on Estimation and Sampling Distributions. If you need to review the logic, now would be a good time.

[3] The logic of using the standard deviation of the individual values in the sample as an estimate of the standard deviation of the individual values in Population 1 was also explained in the Topic Note on Estimation and Sampling Distributions.

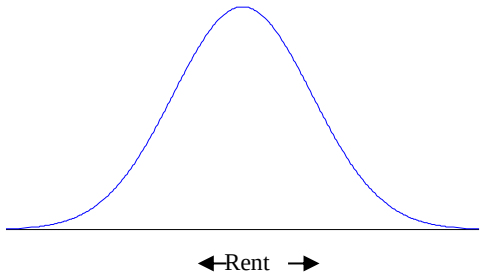
±$18.14 of μ. Wrapping up this assessment of our uncertainty, we say that our best estimate of the mean rent μ of all Austin apartments is $572.27 and we are about 68% certain that the true mean μ lies in the interval $572.27 ±$18.14.

The preceding few paragraphs summarize how we develop a good estimate of the mean rent of all Austin apartments and assess the uncertainty of the estimate. If any of this is not clear to you, please review the appropriate sections of the previous Topic Notes.

Now I will show how regression can improve the answer.

It is intuitive that large apartments should usually rent for more than smaller apartments. But the $572.27 estimate is for the mean rent of all apartments, regardless of their size. If we want an estimate for the mean rent of all apartments that have 1,000 square feet of area, then $572.27 may not be a very good estimate. It seems intuitive that the mean rent of 1000-square foot apartments should be more than the mean rent of 800-square foot apartments, which should be more than the mean rent of 600-square foot apartments. To be sure, we should not expect all 1000-square foot apartments to rent for the same amount – they may still differ in terms of number of bathrooms, age, and other factors that affect rent. Some 1000-square foot apartments may rent for less than some 800-square foot apartments. But the average 1000-square foot apartment should command higher rent than the average 800-square foot apartment. It is important to note that even if we know that a group of apartments all have the same area, we are still uncertain about the rent because there are many factors other than area that affect rent. Since we are uncertain about the rents of 1000-square foot apartments, there is a distribution of rent for 1000-square foot apartments – with a set of possible outcomes (rents) and a probability curve. The following graph shows one *possible* probability curve for the rents of 1000-square foot apartments.

Figure 1

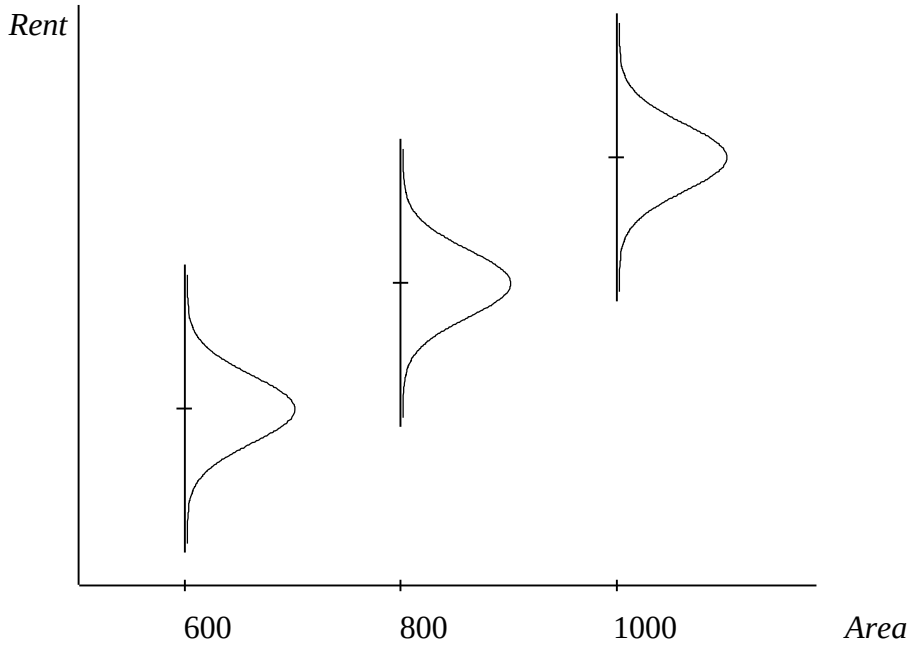


It is important to note that the uncertainty distribution shown in Figure 1 is a subset of the uncertainty distribution of rent for all Austin apartments. Figure 1 is for 1000-square foot apartments only. Every apartment in this distribution has the same area – 1000 square feet. But their rents still vary. There is a mean rent, where the deviations of higher-rent 1000-square foot apartments exactly balance the deviations of lower-rent 1000-square foot apartments. And there is a standard deviation, which measures the average magnitude of the deviations of 1000-square foot apartments from their mean rent. We do not know either that mean or that standard deviation.

Similarly, since we are uncertain about the rents of 800-square foot apartments, there is a distribution of rent for 800-square foot apartments – with a set of possible outcomes (rents) and a probability curve. Likewise for rents of 600-square foot apartments. Do you understand that for every value of area, there is a distribution of rent for apartments of that area – with a set of possible outcomes (rents) and a probability curve?

The following graph shows one *possible* set of probability curves for the rents of 600, 800, and 1000 square foot apartments. In order to show them all on the same graph along with area, I have rotated the rent axis from horizontal to vertical and put area on the horizontal axis. Although I have used the same shape for all three probability curves, the true curves may have different shapes and they may occupy different vertical positions than I have drawn. Also, apartments have other possible areas. 600, 800, and 1000 are not the only possible areas. Potentially, every possible area has a distribution of rents that could be shown on the graph. Every distribution has a mean and a standard deviation. ***The objective of regression is to estimate these means and standard deviations.***

Figure 2



Before discussing how we estimate these means and standard deviations, let me give a few reasons why these estimates are useful.

Having good estimates of the means is useful for two reasons:

- We can see how much the typical rent should be for an apartment of a given area. This gives a benchmark value for renters, developers, tax collectors, appraisers, insurance companies, and others. If an apartment of that area actually rents for more or less than the mean, the apartment should have compensating factors that justify the deviation.
- We can see how the mean rent changes as the area changes. This tells us how much we can expect apartment rent to increase if area increases by a specified amount. So how

much more rent is 100 additional square feet worth? This is useful information to renters, developers, tax collectors, appraisers, insurance companies, and others.

Having a good estimate of the standard deviation of the distribution is valuable because it summarizes the remaining uncertainty about the rent after taking the area into account. Since the apartments in one of the vertical distributions all have the same area, something other than area must be responsible for their differences in rent. Their standard deviation tells us how much the rent of apartments of a given size deviate, on average, from their expected rent. If this typical deviation is large, then factors other than area must be important in explaining why the rent is at the level it is. But if the typical deviation is small, then factors other than area have little to say about why the rent is at the level it is.

Now back to the theme:

The objective of regression is to estimate the means and standard deviations of the distributions of Y for each x.

How to do this? One approach: You might try to adapt the procedure that we followed for estimating the mean rent of all Austin apartments that was reviewed at the beginning of this Topic Note above. First get the areas of the 60 sample apartments. (I have the areas and I will show them to you in a moment.) Then divide the sample of 60 apartments into groups. Put the 600-square foot apartments into one group, the 800-square foot apartments into a second group, and the 1000-square foot apartments into a third group. Then calculate the mean rent of the sample apartments in each group and summarize the uncertainty of the estimate for each group by the standard error of each group's mean.

Unfortunately, this straightforward application does not work well. A big problem is that there is only one apartment with 600 square feet in the sample, only two with 800 square feet, and none with 1000 square feet. So you might revise the groups to include apartments that are "close" to 600, 800, and 1000 in area – like the ranges 550-650, 750-850, 950-1050. But this still leaves a small number of apartments in each group – too small for averaging to work well in balancing the large deviations in each group and too small to invoke the Central Limit Theorem in assessing the uncertainty of the estimates for each group.[4] The problem is that there are too many vertical distributions to estimate and too few data in each vertical distribution for the basic approach to work well unless we sample a lot more apartments.

The solution adopted by simple linear regression is to make some strong simplifying assumptions about the distributions and their relationships with each other. These assumptions will allow us to use *all* of the data to estimate each individual distribution and to use *all* of the data in assessing the uncertainty of the estimate for each individual distribution. However, these strong assumptions may not be correct. Fortunately, there are some simple diagnostic tests that we can use to check whether the assumptions are plausible. Often they are plausible. Later in the course, we will see some ways to correct matters if the assumptions are wrong.

[4] Nevertheless, it is a good idea – especially if you make enough groups to cover the full range of areas and "smooth" all of the group means afterward. Some advanced nonlinear regression techniques use variations on this idea.

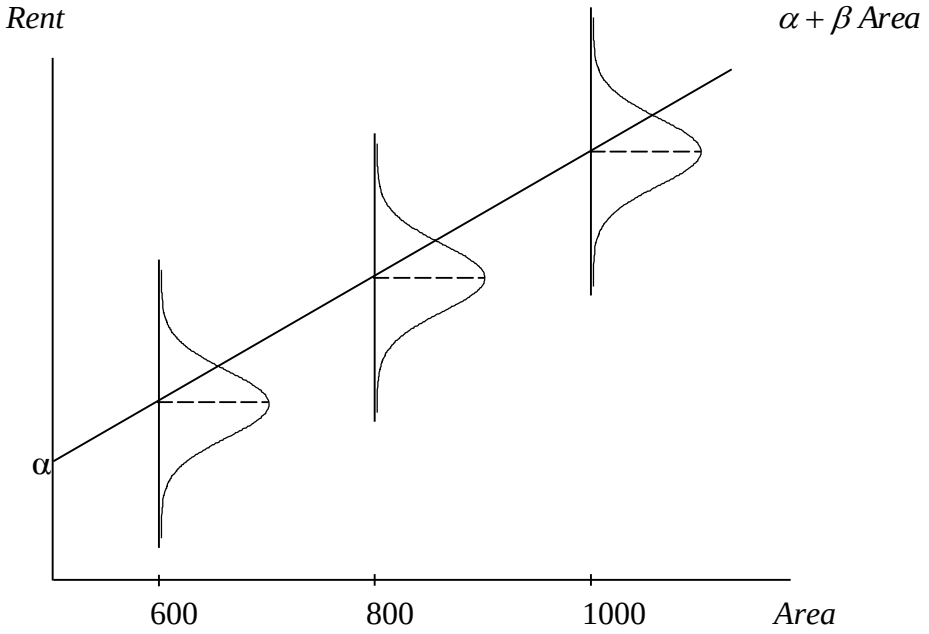
The Assumptions of the Simple Linear Regression Model

There are four assumptions. I emphasize that it is important to check whether the assumptions are plausible each time that you use regression. If one or more assumptions are violated, then the conclusions drawn from the analysis may not be correct. Later in the course, I will present some simple graphical tests that you can employ to check the validity of the regression assumptions.

Assumption 1 (L – the Linear assumption): The means of the distributions of rent (in general, *Y*) lie on a straight line $\alpha + \beta\, Area$ (in general $\alpha + \beta\, x$).

This assumption is illustrated graphically in the figure below. The line $\alpha + \beta\, Area$ that connects the means of each distribution is called the ***true* regression equation**. We do not know what the equation is because we do not know what the true mean of each distribution is. Therefore, we are uncertain about the true regression equation. Even if Assumption 1 is correct, we do not know the values of α and β. Therefore, there is a distribution of uncertainty with outcomes and probabilities for each of α and β. We will estimate α and β in order to estimate the true regression equation. But this ***estimated* regression equation** will deviate from the true regression equation. We will need to assess the uncertainty of the estimates.

Figure 3



There are four important points to understand about the L Assumption:

- It is a strong simplifying assumption. It replaces the difficult task of estimating the mean of rent (*Y*) for *every* given area (*x*) by the much simpler task of estimating only two parameters – α and β. If you knew the values of α and β, then you would know the mean of *Y* for every *x*. We can use all of the data to estimate each of the two parameters. If the L assumption is correct, it eases our task considerably. But L is an assumption and should be tested for plausibility.

- Suppose the L assumption is wrong and we proceed anyway. Then the actual means will not lie on a straight line. Nonetheless, the regression analysis will produce a straight line and it may be rather far from some of the true means.
- α is the *Y*-intercept of the true regression equation. The line crosses the vertical axis at the value α. Moreover, α is the value that you get when you plug area = 0 into the true regression equation. Theoretically, α is the true mean rent of all apartments that have zero area. Of course, that is a nonsensical idea. There are no apartments with zero area. This is a warning that regression models may not hold if extrapolated beyond the range of the available data. In other cases, plugging $x = 0$ into a regression equation may be quite valid (for example, if *Y* is stock price and *x* is earnings per share, which are sometimes zero or negative). For apartments, however, we could still plausibly interpret α as the value of fixed costs that all apartments must cover in their rents, regardless of their areas.
- β is the slope of the true regression equation. It represents the expected change in rent for an increase of one square foot in area. If β = 0.5, then additional area is worth 50 cents more rent per square foot, on average. In most regressions, β is of more interest than α. For apartments, β is a variable cost factor. It shows the effect on rent of varying the area.

Assumption 2 (H – the constant variance or Homoscedasticity assumption): The standard deviations of the distributions of rent (in general, of *Y*) are the same for every level of area (in general, of *x*).

In the illustrative figure above, I drew the distributions with the same standard deviation.

The H assumption is satisfied if the uncertainty about rent is the same for every value of area. Since rents vary *within* the vertical distributions for reasons *other than* differences in area, this says that the other-than-area factors that make rents vary operate in the same manner at every value of area. This is a fairly strong assumption. Fortunately, there is a simple graphical test for its plausibility. I will discuss this test later in the course.

Suppose the H assumption is wrong and we proceed anyway. Then the uncertainty about rent varies with area, sometimes more uncertainty in one vertical distribution, sometimes less in another. However, the regression model will produce a common estimate of uncertainty for all of them. This estimate will be too big for some and too small for others. This means that the uncertainty within vertical distributions will be overestimated for some areas and underestimated for others. So we will be underconfident about rent for some areas and overconfident about rent for other areas.

In case the H assumption is true, then all of the vertical distributions have the same standard deviation. In this case I will use the symbol σ_e to stand for this one value. (You may think of the subscript *e* as standing for "equal" as a mnemonic.) In case the H assumption is false, then the vertical distributions may have standard deviations that vary, depending upon the value of *x*. In that case I will use the symbol σ_x, where the subscript *x* indicates that the value of the standard deviation may change from one *x* to another.

Assumption 3 (I – the Independence assumption): The distributions are independent of each other.

This assumption can be interpreted graphically in terms of the above figure. Suppose we sample an apartment with 600 square feet of area, and its rent is more than expected – that is, the

rent lies above the true regression line. If we then look at sampled apartments at other levels of area (say at 800 and 1000 square feet, for example), the fact that the 600-square foot apartment has higher rent than expected should not affect whether or not the 800 and 1000-square foot apartments are above or below the regression equation. In general, the rents of two apartments are independent if the value of one rent relative to its mean (above or below the regression equation) gives no information about whether or not the rent of the other apartment is above or below its mean.

If the I assumption is not correct, there can be serious problems with regression analysis. However, if the apartments are drawn as a random sample, then the I assumption is automatically satisfied.

Assumption 4 (N – the Normal assumption): The vertical distributions are all normal.

In the above figure, I have shown the three illustrative distributions as normal. You will see later that the N assumption is not important for *estimation of parameters* if the sample size is sufficiently large. The N assumption will be important for *prediction of individual values* – but even there, we have a work-around. Finally, there are graphical and quantitative tests for the plausibility of the N assumption.

The four assumptions say that that the rent distributions are independent and have the same normal shape, but different locations along a straight line. You can get any of the distributions just by sliding any one of the distributions like a railroad car along the "track" of the regression line.

A few words about the relative importance of the four assumptions:

- The L assumption is the most important. The point of regression is to estimate the mean rent (mean Y) as area varies (as x varies). If we get that wrong, then regression has failed its major objective. We will have biased estimates of the means. Fortunately, if the L assumption is wrong, it is often easy to fix.
- The I assumption is next in importance. It is especially important for time series data because the value of a time series at one time often depends upon its value in the past. If the I assumption is wrong, it is often quite difficult to fix.
- The H assumption is of importance mostly for *prediction*, that is, for estimating the rent of an individual apartment of given area. If H is violated, then regression will give the wrong estimate of uncertainty (standard deviation) for the rent of an individual apartment of given area. However, most of the rest of regression, including estimating the regression equation, will work OK, especially if the sample size is large.
- For most parts of regression, the N assumption is not very important. That is because there are versions of the Central Limit Theorem for regression like the version we saw for means in the Topic Note on Estimation and Sampling Distributions. As long as the sample size is "sufficiently large", most of regression will work just fine without the N assumption. The N assumption is convenient in two instances: (i) if the sample size is small, or (ii) for predicting the rent of an individual apartment of a given size, whether the sample size is large or small. For (ii) we can still predict the rent by plugging area into the regression equation, and we can still compute a standard deviation to summarize our uncertainty, but without the N assumption we cannot give a percent confidence for a

margin of error. For example, we cannot say that there is a 68% probability that the actual rent will lie in a ± one standard deviation margin around the predicted value. (There are other things we can do to get a confidence percentage in case N is violated.)

All four assumptions can be tested for plausibility (later).

Now, as promised, here are the rents and areas of the 60 sample apartments:

Table 2. Rents and areas of 60 randomly selected Austin apartments.

Rent	Area	Rent	Area	Rent	Area	Rent	Area
519	725	425	620	505	672	470	751
765	995	770	1040	445	660	480	608
475	481	445	520	565	755	460	900
575	925	510	880	650	810	600	860
415	600	635	832	515	611	575	925
530	668	470	545	470	705	659	944
580	725	700	921	470	564	650	940
995	1421	450	577	700	1250	750	1048
565	672	785	1080	455	512	455	474
620	1025	485	710	550	630	430	700
450	781	415	605	625	850	605	921
520	800	399	680	745	1156	929	1229
495	870	585	730	540	932	695	896
420	700	525	687	650	755	455	630
575	800	495	703	595	1093	1050	1864

How to Run Regression in Excel

- First, set up your data with the predictor (*X*) variable(s) in contiguous columns next to each other. The response (*Y*) variable should also be in a single column. Put labels (variable names) at the tops of the columns. Be sure that you have activated the *Analysis ToolPak* and *Analysis ToolPak VBA* add-ins, if not already activated.[5]
- Click the Data tab to open the Data ribbon in Excel.
- Click the Data Analysis option in the Analysis box at the right on the ribbon. Among the options, scroll down and select Regression, then click OK. If you do not have the Data Analysis option, then you need to activate the *Analysis ToolPak* and *Analysis ToolPak VBA* add-ins.
- In the Regression dialogue box, fill in the *Y* range and the *X* range. Click the "Labels" box if you have column header names. Select the output range to store the computer output. Check the "Residuals" box and perhaps the "Residual Plots" box. Then OK.

[5] Click the File tab, then Options -> Manage: Excel Add-ins -> Go. In the Add-ins dialogue box that pops up, be sure the Analysis ToolPak and Analysis ToolPak – VBA boxes have checks. Then OK.

How to Run Regression with the StatTools Add-in

Although you can run regression in Excel itself if you have activated the Analysis ToolPak add-ins, it is more convenient to run regression in StatTools.

Notes on running regression in StatTools:[6]

- *First, define your dataset with the StatTools* Data Set Manager. This utility allows you to select a rectangular set of data in a spreadsheet and apply a name to it so that StatTools will know where the data are that you want to analyze. Put descriptive labels at the tops of the columns in your rectangle. StatTools will use those labels, by default, as the names of your variables. Click on StatTools → Data Set Manager. StatTools will guess the extent of your data, but you can type in or select the correct range if StatTools guesses incorrectly. Then fill in the dialogue box with the range of your data and give it a name. You may have more than one StatTools data set.
- To start the regression, click on StatTools → Regression & Classification → Regression… . Then fill in the dialogue box:
- For Regression Type, select "Multiple" from the pull-down menu options.
- For Data Set, select the name of the dataset containing your data from the pull-down menu options – assuming you have defined your dataset per the first bullet item above. (If you omitted the Data Set Manager step, go back and do it.)
- Select response and predictor variables by clicking their names. "D" means dependent (response or *Y*) variable. "I" means independent (predictor or *X*) variable. You can select multiple predictors for multiple regression (see later set of Topic Notes), but only one dependent variable.
- In the "Graphs" section, click the box "Residuals vs Fitted Values". This will help you verify L and H and thus help verify the assumptions of the regression model. This is the most important option in this section of the dialogue box, but the others can also be helpful on occasion. For example, a plot of the residuals in time order can be useful if the data are a time series. If time is one of the predictors, you can get this plot by checking "Residuals vs X values". If time is not a predictor, you can get this plot anyway by StatTools → Time Series & Forecasting → Time series graph and selecting the Residual column.
- Click OK. Your regression output will be put into a new Excel workbook.

Missing data: Any row that has one or more missing values will be omitted from the regression analysis – the whole row will be omitted. This is true of both Excel and StatTools.

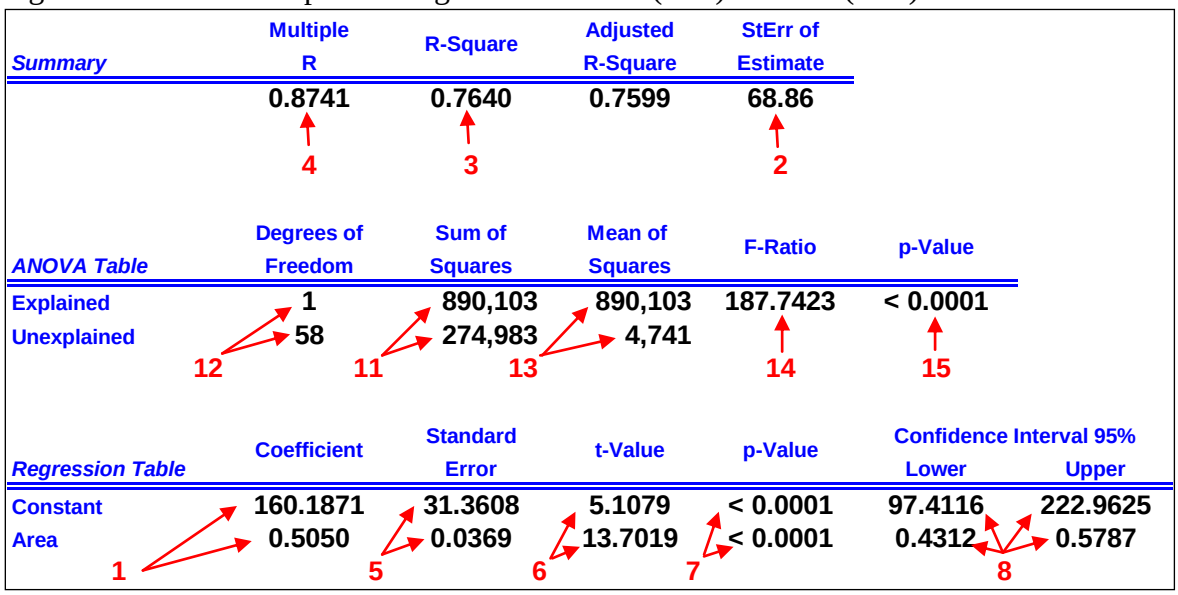
[6] StatTools is part of the Palisades DecisionTools Suite. This suite of management science add-ins for Excel is available to you for download from the McCombs site given on your syllabus.

Interpreting Regression Output

The objective of regression is to estimate the mean rent of all apartments of a given area and to assess the uncertainty of the estimate. Let us assume for now that the four regression assumptions (L,H,I N) are all satisfied. (We will return later to the issue of testing the assumptions.) Enter the rent and area data shown in Table 2 above into Excel and run the regression. The following Figure 4 shows the primary table from the StatTools output. Excel's Data Analysis add-in provides the same output, but in a little different format.

It is easy to become overwhelmed with the amount of output that regression provides. I have labeled the output with 15 sets of numbers that will be interpreted – eventually – but only a few right now. Let us begin with the basic parts, #1, #2, and #5.

Figure 4. StatTools Output for Regression of Rent (as *Y*) on Area (as *X*)



#1. The estimated regression equation.

These figures are the regression estimates of the intercept and the slope. The estimate of the intercept is ***a* = 160.1871**. The estimate of the slope is ***b* = 0.5050**. These are not the true values of the intercept and slope. That is why I use the Roman characters ***a*** and ***b*** to refer to the estimates, in order to distinguish them from the corresponding true intercept and true slope, denoted by α and β.

Thus the estimated regression equation is **160.1871 + 0.5050 *Area***. Since the intercept results from plugging in *Area* = 0, we might interpret the intercept as saying that an apartment of zero area is estimated to rent for $160.19 – a nonsensical interpretation for these data,[7] but perhaps not for other data. The slope may be realistically interpreted as saying that each additional square foot of apartment space is estimated to cost about $0.50 more per month. We could interpret the intercept, more meaningfully, as a fixed cost and the slope times area as a variable cost.

[7] There is no apartment in the dataset with area near zero. The smallest apartment has 474 square feet. Avoid extrapolating regression results beyond the range of available data. The model may not hold there.

- To estimate the mean rent of all 600-square foot apartments, we plug in 600: 160.1871 + 0.5050 * 600 = 463.17;
- To estimate the mean rent of all 800-square foot apartments, we plug in 800: 160.1871 + 0.5050 * 800 = 564.16;
- To estimate the mean rent of all 1000-square foot apartments, we plug in 1000: 160.1871+ 0.5050 * 1000 = 665.16.

See discussion in #2, next up, for assessing the uncertainty of these estimates.

#2. "Standard error of the estimate" [called just "standard error" in Excel's output]

This is the estimate of the common standard deviation σ_e referred to in the H assumption. This is not σ_e itself (which is unknown), but an estimate of σ_e. It estimates the average deviation of the individual rents from the mean rent in each of the vertical distributions depicted in Figure 2 or Figure 3. The "standard error of the estimate" is \$68.86. We can interpret this as saying that use of area in regression allows us to estimate an apartment's rent to within about \$68.86 on average. Contrast this number with the (ordinary) standard deviation of rent, which at \$140.52 is more than twice as large. Thus, knowing the area of an apartment enables us to estimate its rent much more accurately than not knowing its area:

- If we do not know the area of an apartment, the sample allows us to estimate its rent to within ±\$140.52 on average.
- If know the area of an apartment, the sample allows us to estimate its rent to within ±\$68.86 on average.

Knowing the area of the apartment reduces the uncertainty by about half.

Table 3 shows the situation. Column 1 shows the 60 actual rents. Column 2 shows the mean estimate of the 60 rents; this is the same number = mean of 60 actual rents. Column 3 shows the deviation of actual rent from the mean estimate – i.e., Rent – Mean Estimate = Rent – 572.27. Column 4 shows the regression estimate. This is the number obtained by plugging each apartment's area into 160.1871 + 0.5050**Area*. Finally, Column 5 shows the deviation of actual rent from the regression estimate – i.e., Rent – Regression Estimate.

If you look down the rows of Table 3 and compare the Mean Deviations in the blue column with the Regression Deviations in the red column, you will see that the Regression Deviations are generally smaller. A measure of the average magnitude of each of the two columns can be computed by their standard deviations, since the mean of each column is zero. The standard deviation of the Mean Deviations is \$140.52, and the standard deviation of the Regression Deviations is \$68.86.[8] On average, the regression estimates are much closer to the actual rents than the mean estimates are.

Table 3. 60 Sample Apartment Rents, Two Sets of Estimates, and Their Deviations

Rent	Mean Estimate	Mean Deviation	Regression Estimate	Regression Deviation
519	572.27	-53.27	526.29	-7.29
765	572.27	192.73	662.63	102.37
475	572.27	-97.27	403.08	71.92

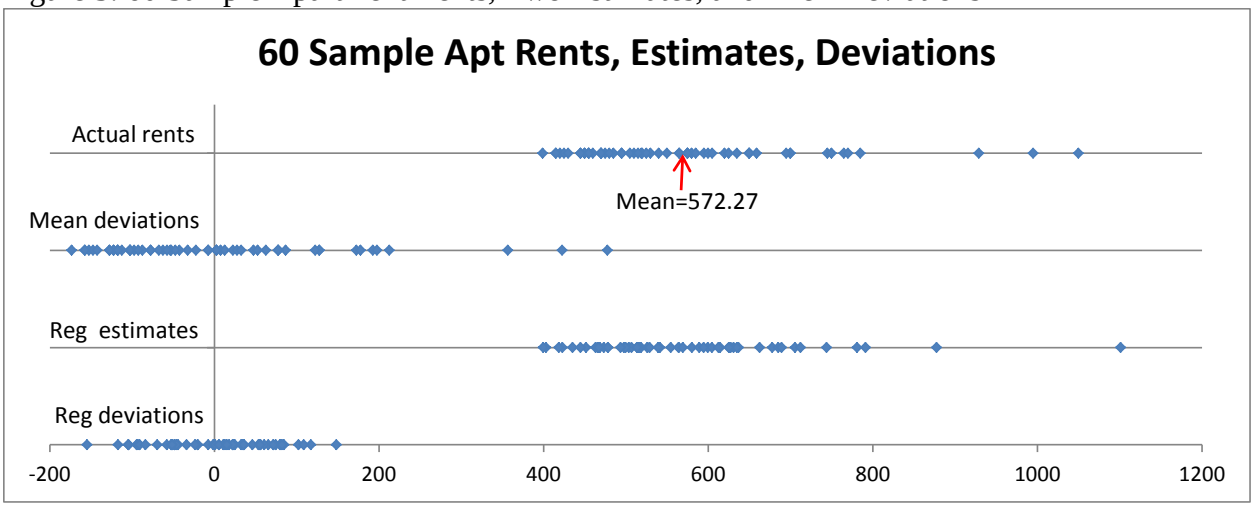
[8] Provided you compute the standard deviation for the Regression Deviations by dividing by $n - 2$, instead of $n - 1$. You are not required to understand the reason for dividing by $n - 2$, instead of $n - 1$ in simple regression.

575	572.27	2.73	627.28	-52.28
415	572.27	-157.27	463.17	-48.17
530	572.27	-42.27	497.51	32.49
580	572.27	7.73	526.29	53.71
995	572.27	422.73	877.75	117.25
565	572.27	-7.27	499.53	65.47
620	572.27	47.73	677.78	-57.78
450	572.27	-122.27	554.57	-104.57
520	572.27	-52.27	564.16	-44.16
495	572.27	-77.27	599.51	-104.51
420	572.27	-152.27	513.67	-93.67
575	572.27	2.73	564.16	10.84
425	572.27	-147.27	473.27	-48.27
770	572.27	197.73	685.35	84.65
445	572.27	-127.27	422.77	22.23
510	572.27	-62.27	604.56	-94.56
635	572.27	62.73	580.32	54.68
470	572.27	-102.27	435.39	34.61
700	572.27	127.73	625.26	74.74
450	572.27	-122.27	451.55	-1.55
785	572.27	212.73	705.55	79.45
485	572.27	-87.27	518.71	-33.71
415	572.27	-157.27	465.69	-50.69
399	572.27	-173.27	503.57	-104.57
585	572.27	12.73	528.81	56.19
525	572.27	-47.27	507.10	17.90
495	572.27	-77.27	515.18	-20.18
505	572.27	-67.27	499.53	5.47
445	572.27	-127.27	493.47	-48.47
565	572.27	-7.27	541.44	23.56
650	572.27	77.73	569.21	80.79
515	572.27	-57.27	468.72	46.28
470	572.27	-102.27	516.19	-46.19
470	572.27	-102.27	444.99	25.01
700	572.27	127.73	791.40	-91.40
455	572.27	-117.27	418.73	36.27
550	572.27	-22.27	478.32	71.68
625	572.27	52.73	589.41	35.59
745	572.27	172.73	743.93	1.07
540	572.27	-32.27	630.82	-90.82
650	572.27	77.73	541.44	108.56
595	572.27	22.73	712.12	-117.12
470	572.27	-102.27	539.42	-69.42

480	572.27	-92.27	467.21	12.79
460	572.27	-112.27	614.66	-154.66
600	572.27	27.73	594.46	5.54
575	572.27	2.73	627.28	-52.28
659	572.27	86.73	636.88	22.12
650	572.27	77.73	634.86	15.14
750	572.27	177.73	689.39	60.61
455	572.27	-117.27	399.54	55.46
430	572.27	-142.27	513.67	-83.67
605	572.27	32.73	625.26	-20.26
929	572.27	356.73	780.79	148.21
695	572.27	122.73	612.64	82.36
455	572.27	-117.27	478.32	-23.32
1050	572.27	477.73	1101.45	-51.45

Figure 5 shows the situation graphically. The top row of Figure 5 shows the 60 actual rents with the red arrow indicating the mean estimate. The second row shows the deviation of each apartment from the mean estimate of $572.27. The second row is just the top row shifted left by $572.27. The third row shows the 60 regression estimates, obtained by plugging the areas into the regression equation. Collectively, they appear to match the actual rents in the top row fairly well. Finally, the bottom row shows the regression deviations. These are the differences between the actual rents in the top row and the corresponding regression estimates in the third row. Even an eyeball comparison of the regression deviations in the bottom row with the mean deviations in the second row shows that the regression deviations collectively are much smaller in magnitude than the mean deviations. The regression deviations are much less spread out around their mean (0) than the mean deviations are around their mean (0). The average magnitude of the regression deviation from 0 is $68.86; the average magnitude of the mean deviation from 0 is $140.52.

Figure 5. 60 Sample Apartment Rents, Two Estimates, and Their Deviations



Moreover, if the N assumption is true, we can go further and say that about 68% of the rents in the whole population of all Austin apartments are within $68.86 of their regression

estimates. If the N assumption is not true, we can still say that the average Austin apartment deviates by about $68.86 from its regression estimate, but we cannot go further and say that 68% of the apartments deviate by less than $68.86 from their regression estimates.[9]

The name ("standard error of the estimate") that StatTools assigns to #2 is unfortunate. The name ("standard error") that Excel assigns to #2 is even more unfortunate. For those names suggest that this quantity is a "standard error" in the sense of a standard deviation of a sampling distribution (Population 3).[10] It is NOT. It is important to note that the so-called "standard error of the estimate" does NOT assess the uncertainty of the estimates of *mean* rents of 600, 800, and 1000-square foot apartments calculated in #1 immediately preceding. The "standard error of the estimate" assesses uncertainty about the rents of INDIVIDUAL apartments in the vertical distributions. This is the uncertainty about individual apartment rent due to factors other than area. Thus, recalling the terminology from the Topic Note on Mean and Standard Deviation, the "standard error of the estimate" corresponds to the standard deviation of the portion of Population 1 that is in any of the vertical distributions in Figure 2 or Figure 3. To assess the uncertainty of the regression estimates of the means of the vertical distributions, we need to know how much the typical regression estimate of the vertical mean deviates from the true vertical mean; this is a Population 3 property.

The development of this concept in regression (estimating the mean rent $\alpha + \beta x$ of all apartments having area = x) is analogous to the *pre*-regression development of the Three Populations. You may recall that discussion from the Topic Note on Estimation and Sampling Distributions. If the ideas from that Topic Note are fuzzy to you, I suggest that you review them before continuing. Since the ideas are so analogous in regression, I will only sketch their development. I will illustrate with the case of x = 1000-square foot apartments:

We want to estimate the mean rent of all 1000-square foot apartments and quantify the uncertainty of this estimate. Based upon my sample of 60 apartments, the estimate of mean rent of all 1000-square foot apartments is the plug-in estimate: 160.1871+ 0.5050 * 1000 = 665.16. If I had drawn a different sample of 60 apartments, I would have had a different set of rents and areas. Consequently, the estimated regression equation would have been different. So, when we plug in area = 1000, we would get a different estimate of the mean rent of 1000-square foot apartments. Imagine calculating the regression equation for every *possible* sample of 60 apartments. Each such regression equation has the form *Estimated mean rent = a + b * Area*, where *a* and *b* vary from sample to sample. Then plug *Area* = 1000 into each such equation and calculate the estimates of mean rent of 1000-square foot apartments. Put all of these estimates together in one pile. Every number in that pile is a plug-in estimate of the mean rent of all 1000-square foot apartments. Our sample estimate of $665.16 is one of the numbers in that pile. The pile (uncertainty distribution) has three important properties:

- (Property 1 – Central Limit Theorem) The distribution is (approx.) normal if the sample size 60 is sufficiently large (it is – 30 or more is the rule of thumb).
- (Property 2) The mean of the sampling distribution equals the true mean rent $\alpha + 1000\beta$ of all apartments having 1000 square feet.

[9] The apparent connection between 68 percent and 68 dollars is pure coincidence.

[10] Most other software publishers call this quantity "root mean-square error", which is both more accurate and less confusing. I am sorry that software publishers cannot settle upon one clearly descriptive name. But that is the way things are.

- (Property 3) The standard deviation of the sampling distribution is approximately equal to the "*standard error of the estimate*" ÷ $\sqrt{n}$.[11,12] Since $n = 60$, then the "*standard error of the estimate*" ÷ $\sqrt{n}$ = 68.86 ÷ $\sqrt{60}$ = 8.89.

We can now fully assess the uncertainty of the estimates calculated in #1 immediately preceding as follows:

- We estimate the mean rent of all 600-square foot apartments to be 160.1871 + 0.5050 * 600 = 463.17 ± 8.89 on average, with 68% confidence that the true mean will lie therein;
- We estimate the mean rent of all 800-square foot apartments to be 160.1871 + 0.5050 * 800 = 564.16 ± 8.89 on average, with 68% confidence that the true mean will lie therein;
- We estimate the mean rent of all 1000-square foot apartments to be 160.1871 + 0.5050 * 1000 = 665.16 ± 8.89 on average, with 68% confidence that the true mean will lie therein.

Notice that the +/- amount of expected deviation is the same regardless of the value of area. This is a direct consequence of the H assumption of regression.

#5. Standard error of the coefficients

These numbers assess the uncertainty of the estimates of the intercept and slope. The development of this concept in regression and the associated Properties is analogous to the *pre*-regression development of the Three Populations in the Topic Note on Estimation and Sampling Distributions and to the discussion just presented in #2, immediately above in this Topic Note. If the ideas from the previous Topic Note are fuzzy to you, I suggest that you review them before continuing. Since the ideas are so similar in regression, I will only sketch their development:

If I had drawn a different sample of 60 apartments, I would have had a different set of rents and areas. Consequently, the estimated regression equation would have been different. That is, the equation would have a different slope and a different intercept. Imagine calculating the regression equation for every *possible* sample of size 60. Each regression equation has the form *Estimated mean rent* = *a* + *b* * *Area*, where *a* and *b* vary from sample to sample. Put all of the slopes *b* together in one pile, one slope for every sample. That pile is a Population 3: the sampling distribution of the estimated slopes. It has three important properties:

- (Property 1 – Central Limit Theorem) The distribution is (approx.) normal if the sample size 60 is sufficiently large (it is – 30 or more is the rule of thumb).
- (Property 2) The mean of the sampling distribution equals the true slope β .
- (Property 3) The standard deviation of the sampling distribution is approximately equal to the value printed on the output at #5, namely 0.0369.[13]

---

[11] Note how this formula parallels the formula for the standard error of the mean from the pre-regression setting of the Topic Note on Estimation and Sampling Distributions. All three Properties are nearly exact parallels of the 3 Properties from that Topic Note.

[12] This formula is exact if $x = \bar{x}$. Otherwise it is an approximation that is too small. The value of the exact formula gets larger the farther *x* is from $\bar{x}$. The exact formula is approximately $(Z_x)^2 / 2$ times larger, where Z_x is the Z-score of *x*, that is, $Z_x = (x - \bar{x}) / stdev(x)$. For *x* = 1000 sq ft, the exact formula is about $(Z_x)^2 / 2$ = [(1000-816.05)/243.24]^2 / 2 = 28.6% larger. So $8.89*(1+.286) = $11.43 is a better approximation for *x*=1000 sq ft. For *x*=800 sq ft, the formula is about 0.2% larger. For *x*=600 sq ft, the formula is about 39.4% larger. If *x* is more than 2 standard deviations from the mean *x*, the correction can be substantial in percentage terms – 200% or more larger – however, the division by $\sqrt{n}$ is usually a much stronger reduction effect.

Similarly, if you put all of the intercepts *a* together in one pile, one intercept for every possible sample, that pile is a Population 3: the sampling distribution of the estimated intercepts. It has three important properties:

- (Property 1 – Central Limit Theorem) The distribution is (approx.) normal if the sample size 60 is sufficiently large (it is – 30 or more is the rule of thumb).
- (Property 2) The mean of the sampling distribution equals the true intercept α.
- (Property 3) The standard deviation of the sampling distribution is approximately equal to the value printed on the output at #5 in Figure 4, namely 31.3608. [14]

Using these Properties, we can fully assess the uncertainty of the estimates printed on the output at #2.

- We estimate the true slope to be 0.5050 ± 0.0369 on average, with 68% confidence that the true slope will lie therein;
- We estimate the true intercept to be 160.1871 ± 31.3608 on average, with 68% confidence that the true intercept will lie therein;

The relatively small standard error (0.0369) of the sample slope suggests that we can be reasonably confident that the cost of extra area is close to 50 cents per square foot, on average. As noted earlier, we do not really care about the intercept in this case (except as a possible measure of fixed costs), and we distrust the regression equation generally for areas much below the lowest area in the data. But I present the assessment of uncertainty for the intercept in the interest of completeness.

[13] There is a formula for calculating this quantity, but you are not responsible for knowing that formula.
[14] There is a formula for calculating this quantity, but you are not responsible for knowing that formula.

SUMMARY

The purpose of regression is to estimate the mean *Y* for each given value of *X* and to estimate the uncertainty (standard deviation) of *Y* when all *X* values are the same. In order to avoid having too many means and standard deviations for the data to make reasonable estimates, regression imposes four strong modeling assumptions on the data:

- **L** – As *x* varies, the means of the distributions of *Y* fall on a straight line: *mean of Y* = $\alpha + \beta x$.
- **H** - The standard deviations of the distributions of *Y* are the same for every *x*.
- **I** – The distributions of *Y* at all *x*'s are independent of each other.
- **N** – The distributions of *Y* at all *x*'s are normal.

If these four assumptions hold, then only three parameters need to be estimated: α, β, and the common standard deviation σ_e of the *Y*'s at each *x* – instead of having to estimate a different mean and standard deviation at every *x*. This is a great economy. The means can be estimated by plugging *x* into the estimated regression line *a* + *bx*, which re-uses the estimates *a* and *b* of intercept and slope repeatedly every time the mean of *Y* at another *x* is desired.

Moreover, estimates of the three parameters, α, β, and the common standard deviation σ_e of the *Y*'s at each *x*, do not need to be calculated by hand – they can be read off computer output.

I explained how to run regression with Excel and with the StatTools add-in.

Regression output has many parts. In this Topic Note, I made a start on interpreting the output. In particular:

- (#1) The coefficients in the regression table output give the estimates *a* and *b* of the intercept α and slope β. These estimates are used to calculate the plug-in estimates of the mean of *Y* at any *x*: *a* + *bx*.
- (#2) The "standard error of the estimate" provides the estimate of the common standard deviation σ_e of the *Y*'s at each *x*. This has two uses:
    - to assess the *remaining* uncertainty about the *value* of *Y* after taking your knowledge of *x* into account; and
    - (after dividing by $\sqrt{n}$) to assess the uncertainty about the *mean* of *Y* at a given *x*.
- (#5) The standard errors of the coefficient estimates assess the uncertainty of the sampling distribution of the intercept and slope estimates. These figures say how far you can expect the intercept and slope estimates *a* and *b* to be from the true intercept and slope α and β.