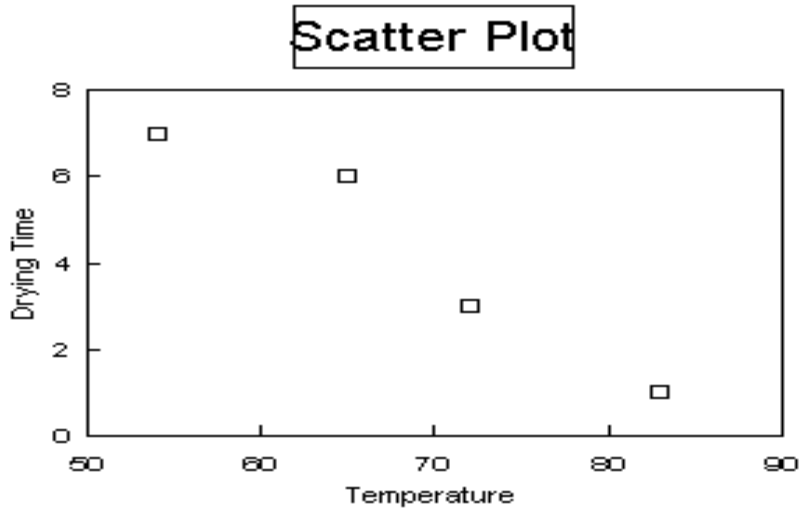
SIMPLE REGRESSION ANALYSIS

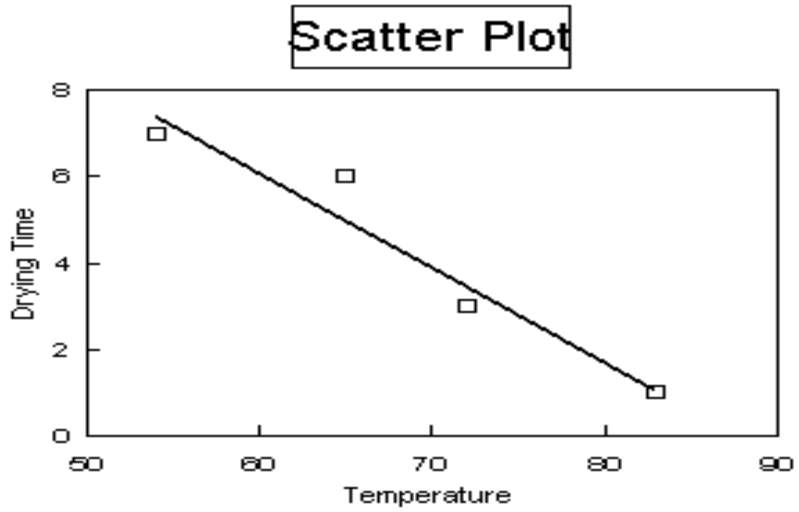
Introduction. Regression analysis is used when two or more variables are thought to be systematically connected by a linear relationship. In simple regression, we have only two – let us designate them x and y – and we suppose that they are related by an expression of the form $y = \beta_0 + \beta_1 x + \varepsilon$. We'll leave aside for a moment the nature of the variable ε and focus on the x - y relationship. $y = \beta_0 + \beta_1 x$ is the equation of a straight line; β_0 is the *intercept* (or *constant*) and β_1 is the *x coefficient*, which represents the slope of the straight line the equation describes. To be concrete, suppose we are talking about the relation between air temperature and the drying time of paint. We know from experience that as x (temperature) increases, y (drying time) decreases, and we might suppose that the relationship is linear. But suppose that we need to know the exact nature of the relationship, so that we can predict drying time at various temperatures. How could we discover the actual values of β_0 and β_1? Well, to start with, we cannot discover the actual values. Note that β_0 and β_1 are Greek letters, indicating that these are parameters, and they are somewhat in the nature of population parameters which can never be known exactly. What we can do is to get estimates of these parameters – let us call them b_0 and b_1, using Latin characters to indicate that these are statistics and only approximations of the real thing. How could we find b_0 and b_1? Let's do an experiment: let's paint some boards at different temperatures and observe how long it takes them to dry. Such an experiment might generate the following data:

Temperature	Drying Time (hours)
54	7
65	6
72	3
83	1

If we plotted these points on an axis system, we could see the relationship:



Each little box represents a temperature/drying time observation, and we can see that in a general way, as temperature goes up, drying time goes down. One way to find the exact relationship would be to take a straight edge and try to draw the line that comes closest to hitting all the dots. By measuring with a ruler, we could then determine the slope and intercept of this line. This is somewhat laborious, however, and since everyone is likely to see the scatter diagram a little differently, this technique would not give us a single, agreed-upon line. Another way to find the exact relationship would be to use regression analysis. By entering the data into the appropriate formulas (discussed below) we could find the line shown next:



This straight line has the equation DT = 19.26 – 0.22 x Temp, so that $b_0 = 19.26$ and $b_1 = -0.22$, and we may take this equation to describe the relation between our two variables. By entering a temperature, we can evaluate the equation to discover the drying time at that temperature.

We are left with a question however: why is it that all the points aren't on the straight line? Well, because other things besides temperature affect drying time, and those other things had different values for each of our sample points. This is likely always to be the case – for any y, there may be literally hundreds of things that affect its value, and we have concentrated on only one. To take explicit account of this, when we stated the regression equation above, we included a factor labeled ε. Since we cannot control or account for all the myriad things that affect drying time, we summarize them in ε and take ε to be a random variable whose value on any particular occasion is determined simply by chance; it is the operation of this ε that causes our actual observations to diverge from the straight line. The existence of these random deviations is what makes regression analysis interesting and useful. If all relations were exact and free from chance, we could simply establish two points and draw a line between them to represent the relation; as it is, we must deal with randomness, and the regression line thus represents the *average* relationship between x and y.

The Regression Model

- Assumptions of the Regression Model
  - the relation between x and y is given by $y = \beta_0 + \beta_1 x + \varepsilon$

  ε is a random variable, which may have both positive and negative values, so
  - ε is normally distributed
  - $E(\varepsilon) = 0$
  - the standard deviation of ε, σ_{yx}, is constant over the whole range of variation of x. This property is called "homoscedasticity."
    - since $E(\varepsilon) = 0$, we're supposing that $E(y) = \beta_0 + \beta 1x + E(\varepsilon) = \beta_0 + \beta_1 x$

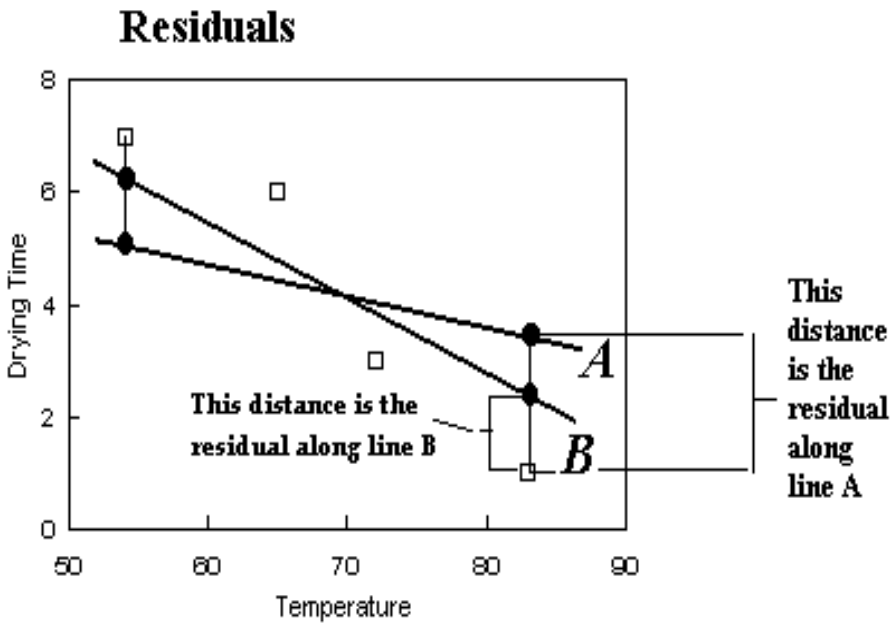
- Finding the regression line: the method of "ordinary least squares" or OLS

  begin with assumed values for b_0 and b_1 and suppose that the relation between x and y is given by $y = b_0 + b_1 x$; some b_0's and b_1's will give us better fits than others

  let $y = a + bx_i$ be the value of y estimated by the regression equation when x has the value x_i; then if y_i is actual value, $y_i - \hat{y}_i$ is called the *residual* or the *error*

  substituting, let $e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i$

  different b_0's and b_1's will cause each e_i to have a different value:



The residuals along the line marked A are larger than those along the line marked B

but the *sum* of deviations is always zero

$\Rightarrow$ square each residual and define the sum of squared errors as $\sum (y_i - b_0 - b_1 x_i)^2$

- x and y are data: the variables are b_0 and b_1, and choosing different values of these will change the size of the sum of squares
- minimizing the sum of squares with respect to b_0 and b_1, using minimization methods from differential calculus, gives unique values for the b's

Resulting formulas are rarely used explicitly anymore, but

$$b_1 = \frac{(\Sigma x_i y_i) - n \times \bar{x} \times \bar{y}}{(\Sigma x^2) - n \times \bar{x}^2}$$

$$b_0 = \bar{y} - b\bar{x}$$

Continuing the paint-drying example: (recall that temp = x and drying time = y)

Temperature	Drying Time (hours)	$x \times y$	x^2
54	7	378	2916
65	6	390	4225
72	3	216	5184
83	1	83	6889

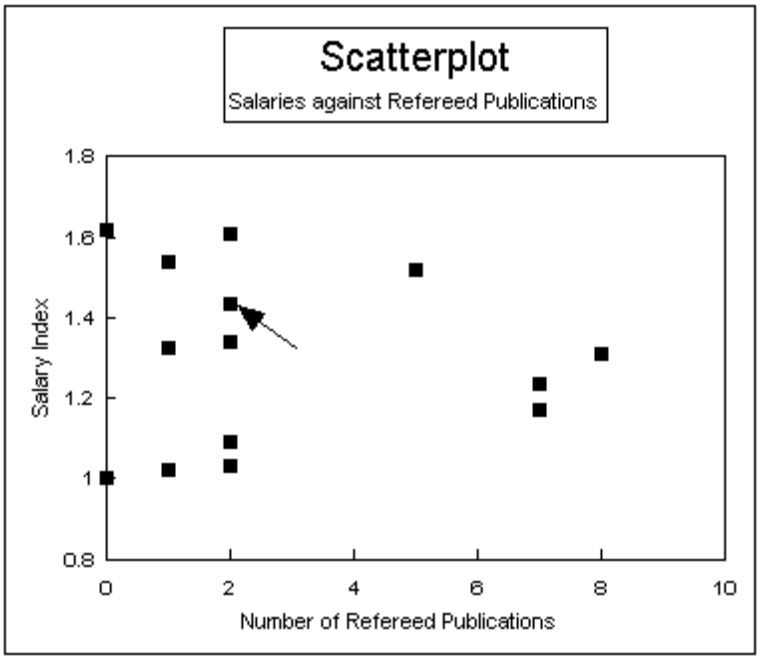
$\bar{x} = 68.5$ $\bar{y} = 4.25$ $\sum x \times y = 1067$ $\sum x^2 = 19214$

$$b_1 = \frac{1067 - 4 \times 68.5 \times 4.25}{19214 - 4 \times 68.5} = -0.21910$$

$b_0 = 4.25 - (-0.21910) \times 68.5 = 19.25842$

Using a Spreadsheet to explore Linear Relations

Looking at relations often begins with a scatter plot of the data, as in the following examples



Each point in the scatterplot represents an individual faculty member; the arrow is pointing at a point which represents a faculty member with 2 publications and salary just over 1.4 times the department's minimum at the time

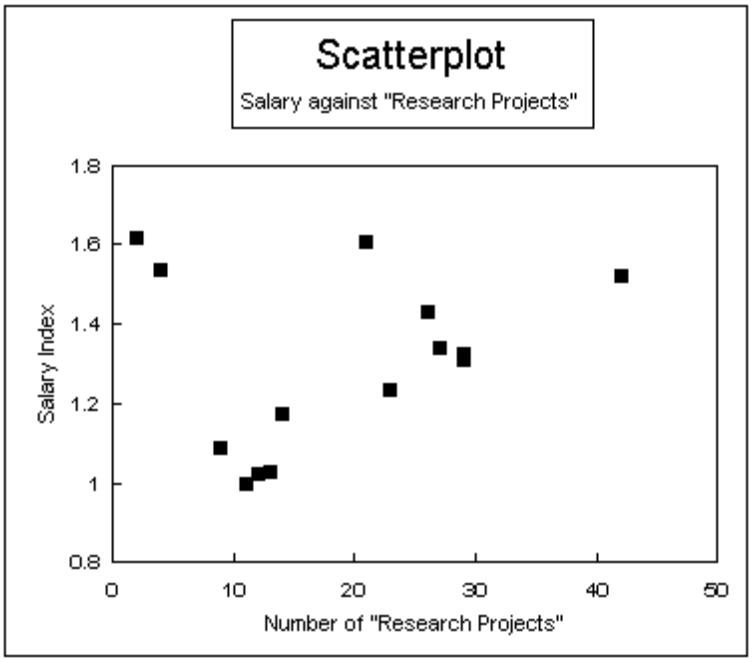
Chairman's Claim: "Salary differences are determined by differences in success in publication."

But – there is no discernible pattern in this plot

- individuals with same number of publications have very different salaries
- individuals with about the same salary have very different numbers of publications

almost any line would fit the data equally well

Same time period, the next diagram plots the salary index against "total research projects," self-declared stuff put on the Annual Report:



except for the three "outliers" in the upper left corner, all points are more or less on a straight line, and we might sensibly conclude that more research projects systematically means a larger salary

Important point: if the three outliers are left in the data set, the regression line is nearly horizontal, and the correlation is low

- ► leaving the outliers out of the data set greatly improves the fit
- ► the outliers would be difficult to spot, even in this small data set, without doing the scatter plot
- ➢ we are led to investigate the nature of these outliers

Creating a Scatter Diagram with Excel

Begin by entering the Data you want to plot in columns
Click on the **Chart Wizard** button in the task bar or click on **Insert** then click on **Chart**

- ► brings up the Chart Wizard, a series of dialog boxes
- ► choose XY (Scatter) and the variant with no lines
- ► if the X and Y variables are in adjacent columns with the X to the left, you can mark the whole data range, and Excel will assume that the left-most column is X
- ► if the X and Y variables are separated or reversed, click on the "Series" tab, then on "Add"; there are separate entry lines for X and Y variables; be careful to get them straight and note that in Microsoft's peculiar terminology, a "Series" is an X variable and a Y variable. In most of the scatter diagrams we'll be doing in Stats II, there is only one series, consisting of two columns of variables.
- ► Variable names at the head of a column will be picked up and preserved as titles on your graph
- ► Many things can be edited after the chart is prepared; others cannot
  - ◆ experience is the most reliable teacher
- ► To print, click on the chart to highlight it, then be sure that the print dialog box specifies printing the selected chart
- ► Charts may also be copied to the clipboard, then inserted into Word documents

SIMPLE REGRESSION II

Spreadsheet Calculations & Predictions

Using Excel to Calculate the Regression Line

Be sure to enter data in separate columns

Click on **TOOLS**

from TOOLS choose **DATA ANALYSIS**

then choose **REGRESSION**

you'll get a dialog box

if you marked the data before clicking on TOOLS, all you have to do is click on OK

otherwise, go ahead and use the dialog box to enter your data ranges

choose an output range: if you don't specify, Excel creates a new sheet and puts your output on it

Output example:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.96902261
R Square	0.93900482
Adjusted R Square	0.90850722
Standard Error	0.83295872
Observations	4

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	21.36235955	21.36236	30.78947	0.030977392
Residual	2	1.387640449	0.69382		
Total	3	22.75			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	19.258427	2.736669611	7.037176	0.019601	7.483479796	31.03337
X Variable 1	-0.2191011	0.03948603	-5.54883	0.030977	-0.388995917	-0.04921

Intercept Coefficient: the b_0 which estimates β_0

X Variable 1 Coefficient: the b_1 that estimates β_1 , or the slope of the regression line

Using the Regression Equation

Equation may be used to

interpolate: predict values within the limits of observation

extrapolate (or forecast): predict values outside the range of observation or experience

To make a point prediction, simply plug an x value into the equation

DT = 19.26 – 0.22 x Temp; what will drying time be at 75 degrees?

DT = 19.26 – 0.22 x 75 = 2.76 hr.

Extrapolation: drying time at 45 degrees: DT = 19.26 – 0.22 x 45 = 9.36

can lead to absurdity: drying time at 95 degrees?

Practitioners generally prefer interval estimates: we're 95% confident that drying time will be between 9 and 9.72 hours, for example

Using a Spreadsheet to Make Predictions

We must use the regression output to create the regression equation somewhere in the spreadsheet. We also need to select a set of adjacent cells as entry cells, so that we can conveniently type in the x value for which we wish to make a prediction and the confidence level for interval predictions.

For point predictions

- we must enter x value at which to make predictions
- spreadsheet output gives you
  - constant term or intercept
  - x coefficient
- thus you can create the regression equation as

  (Cell address for Constant) + (Cell address for x coefficient)*(Cell address used to enter x value)

Homework: 13.1, 13.5, 13.7

SIMPLE REGRESSION III
Evaluating the Equation

- Issue One: Have the assumptions of OLS regression been met?
  - Linear relation
  - constant *distribution* of errors
  - absence of relation between succeeding values of y variable over time
    if y_{t+1} is correlated with y_t the condition is called **autocorrelation** or **serial correlation**

  $\Rightarrow$ analysis of residuals
  if there's a pattern, there's a problem

- Issue Two: Is there a significant relationship between x and y?

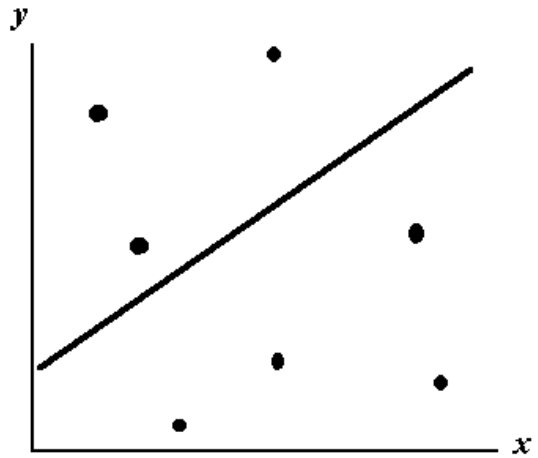
  Purposes and Intuition of the Procedures

  The least-squares technique will always give us coefficients, but
  - there may be no relation – as when a scatter plot is just a random set of dots
  - also – the data points are a sample, subject to sampling error

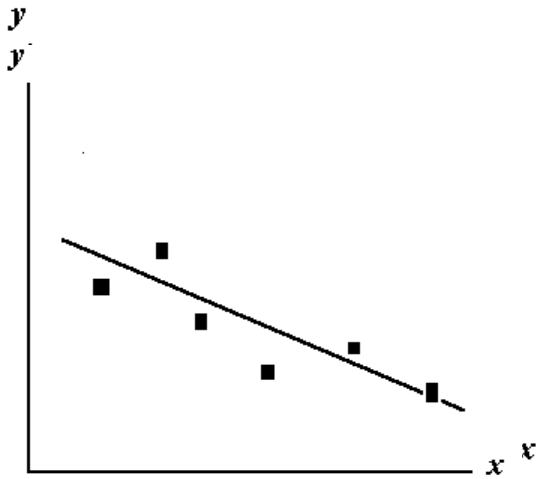
  suggests a need for techniques to evaluate whether there is a relation and, if so, how good x is at explaining y

  Ultimately, the simple techniques all come down to
  - How close the data points are to the regression line
    the coefficient of determination or r^2
  - how different the slope of the regression line is from zero: the t test on β_1
    the quantity $b_1 \div s_{b1}$ is distributed as a t distribution, so tests the hypothesis H_0: $\beta_1 = 0$

  Examples:



No Relation: The points are away from the line



Negative Relation

Positive Relation

r^2, the coefficient of determination

This number serves as an evaluation of the overall closeness of the relationship and represents

- the proportion of variation in y "explained" by variation in x
- how closely data points cluster around the regression line

necessarily, $0 \le r^2 \le 1$

values near 0 indicate a weak relation; not much of the variation in y is associated with variation in x

Cautions:

- r^2 doesn't necessarily mean anything about the relation between x and y as such

some third force, z, may be causing both x and y

the problem of "omitted variables"

- low r^2 doesn't mean there's no relation – only that x explains little of the variation in y

Definition:

$$r^2 = 1 - \frac{\Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2}$$

In this definition

- the numerator is the sum of squared residuals, or "sum of squared errors"

represents deviation of actual from predicted values, that is, variation which is not explained by the regression

- denominator is a measure of total variation in y

- somewhat the same thing as the standard deviation of the residuals divided by the standard deviation of y
  - the smaller are the residuals, that is, the closer the observations lie to the regression line, the closer this fraction is to zero
  - since the ratio (sum of squared residuals)/(sum of squared deviations) represents the portion of variation ***not*** explained, 1 – that ratio is the portion of total variation that is explained.

Evaluating the coefficients

➢ confidence intervals for the coefficients

remember that b_1 estimates the unknown parameter β_1

confidence interval for β_1:

$$b_1 \pm t^{(n-2)d.f.} \cdot s_{b1}$$

Example: we have regressed quantity sold on the price of cat food; our data were generated by setting 15 different prices in 15 different test markets. We have the estimated regression equation Sales = $4000 – $2000 × Price, with price measured in dollars. Further, $s_{b1} = 500$. Calculate a 90% confidence interval for β_1

Solution: There are 15 – 2 = 13 degrees of freedom, which gives a t value of 1.771 for a 90% confidence interval. Thus our confidence interval is

$$-2000 \pm 1.771 \times 500 \text{ or } -2000 \pm 855.5$$

we could also say that we're 90% confident that $-2855.5 \le \beta_1 \le -1144.5$

changing the price by $0.10 will reduce sales by at least 114.5 but by no more than 285.5

- confidence intervals are generated automatically by Excel

➢ Hypothesis tests for the coefficients

b_1 is based on a sample; a different set of data (different sample) will give different b_1

the mean of all these possible b_1's is β_1, the actual slope of the x-y line

if $\beta_1 = 0$ we could, by sheer chance, once in a while get a $b_1 \neq 0$

but the further b is from 0, the less likely that's due to chance

the quantity $(b_1 - \beta_1)/s_{b1}$ is distributed as a t distribution

most often we wish to test

H_0: $\beta_1 = 0$
H_1: $\beta_1 \neq 0$

that is, if $\beta_1 = 0$, there is no systematic relation between x and y, and the equation reduces to $y = \beta_0 + \varepsilon$
rejecting H_0 means concluding that there is a significant relationship between x and y
Note that since $\beta_1 = 0$ by hypothesis, the t statistic reduces to $t = b_1/s_{b1}$
From above: with 13 degrees of freedom, at 1% significance level the critical values of $t = \pm 3.012$. From our data $t = -2000/500 = -4 < -3.012$, so we reject H_0 and conclude that there is a statistically significant relationship between price and quantity

- An F test for the regression using ANOVA
  recall that F = (variation between groups)/(variation within groups), where the between group variation represented systematic variation caused by a difference of "treatments," while within group variation is random, or "error" variation

  here, "variation between groups" is equivalent to variation explained by the regression, so that

$$F = \frac{\text{Regression Mean Square}}{\text{Error Mean Square}} = \frac{\text{RMS}}{\text{EMS}}$$

  Regression Sum of Squares: $SSR = \sum (\hat{y} - \bar{y})^2$
    since the $\hat{y}$ values are predicted by the regression, this sum represents variation from the mean value of y which is due to variation in x
    then $RMS = SSR/m$ where m is the number of explanatory variables and represents the degrees of freedom in SSR
  Error Sum of Squares: $SSR = \sum (y - \hat{y})^2$
    each term represents the difference between the predicted value and actual value; or the portion of the variation in y that isn't "explained" by variation in x
    then $EMS = SSE/(n - 2)$ since n – 2 is the degrees of freedom in the residuals
  Total Sum of Squares: $SST = SSR + SSE = \sum (y - \bar{y})^2$
    which measures the total variation in y from all sources
  - We may simply give the F ratio, although it is customary to present an ANOVA table
    Excel reports the ANOVA table as part of the output, giving both the F value and the p value ("significance F") for the ANOVA test

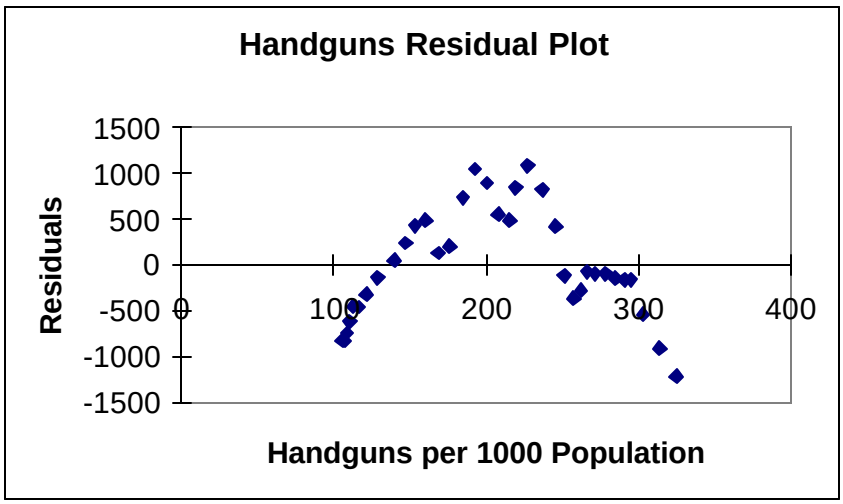
  - In simple regression, the ANOVA test merely repeats the result of the t test on the regression coefficient
  in Excel output, note that F test and t test on the x1 coefficient have the same p value
  real significance is in multiple regression

More on Residuals

- Assumptions of OLS regression:
  - relation is linear, that is, $y = \beta_0 + \beta_1 X + \varepsilon$
  - ε is normally distributed with
    $E(\varepsilon) = 0$
    σ_{YX} constant
  - the residuals may be taken as a sample of the ε term and used to check whether these assumptions are met

- Some possibilities:
  - relation is not linear: positive residuals, then negative, then positive (or the reverse)
  - σ_{YX} is not constant (heteroscedasticity): the residuals will fan out, or get larger, as x increases

- Autocorrelation (serial correlation): a problem with ***time-series*** data
  - the residuals themselves trace a pattern, typically rising then falling then rising
  - ***Example:*** A classic picture of autocorrelation



Effect of autocorrelation: biases r^2 upward
biases s_b downward ⇒ calculated t statistic is biased upward and p-values downward

β_1 appears to be significantly different from 0 when it isn't
Amount of variation explained appears to be greater than it actually is

Testing for autocorrelation: the Durbin-Watson statistic

Making Predictions with a Regression Equation

- Equation may be used to
  interpolate: predict values within the limits of observation
  extrapolate (or forecast): predict values outside the range of observation or experience
- To make a ***point prediction***, simply plug an x value into the equation
  DT = 19.26 – 0.22 x Temp; what will drying time be at 75 degrees?
  DT = 19.26 – 0.22 x 75 = 2.76 hr.
  Extrapolation: drying time at 45 degrees: DT = 19.26 – 0.22 x 45 = 9.36
  can lead to absurdity: drying time at 95 degrees?

Statisticians generally prefer interval estimates
- interval for the average (or mean) value of y, given a value of x
  if we paint twenty boards at 70 degrees, find an interval for the average drying time of the twenty boards
- interval for a specific value of y, given a value of x
  if we paint a board, find a confidence interval for the time it'll take to dry

the first of these is somewhat narrower since it's an average of repeated cases. Like all averages, positive deviations tend to be offset by negative ones

Interval for the average y value, using a t statistic

$$\hat{y} \pm t_C \times s_{yx} \times \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\left(\frac{s_{y.x}}{s_b}\right)^2}}$$

here, t has n - 2 degrees of freedom

$$s_{yx} = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{n-2}}$$

s_{yx} appears as Standard Error in Excel

$$s_b = \frac{s_{y.x}}{\sqrt{\sum x^2 - n\bar{x}^2}}$$

this term appears in Excel as Standard Error for X Variable 1 (or whatever name is used)

NOTE: The formula for the confidence interval seems different from that in your text because it is adapted to use spreadsheet output; interested students may prove to themselves, by appropriate substitutions, that the two formulae are in fact equivalent

Example: From above,

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.96902261
R Square	0.93900482
Adjusted R Square	0.90850722
Standard Error	0.83295872
Observations	4

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	21.36235955	21.36236	30.78947	0.030977392
Residual	2	1.387640449	0.69382		
Total	3	22.75			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	19.258427	2.736669611	7.037176	0.019601	7.483479796	31.03337
X Variable 1	-0.2191011	0.03948603	-5.54883	0.030977	-0.388995917	-0.04921

from spreadsheet output, $s_{yx} = 0.83295872$ and $s_b = 0.03948603$; there are two degrees of freedom, so the appropriate t value for a 90% confidence interval is 2.92. Let us calculate a confidence interval for the average value of drying time at a temperature of 50 degrees.

$\hat{y}$ = 19.26 – 0.22 x 50 = 8.30

plugging numbers into the above formula, the confidence interval is

$$\hat{y} \pm t_C \times s_{yx} \times \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\left(\frac{s_{yx}}{s_b}\right)^2}} = 8.3 \pm 2.92 \times 0.833 \times \sqrt{\frac{1}{4} + \frac{(50-68.5)^2}{\left(\frac{0.833}{0.0395}\right)^2}} = 8.30 \pm 2.46$$

Note: although I've rounded off at intermediate steps in this calculation, you should **NEVER** do so in practice; carry your calculations out to as many decimal places as possible, rounding off only with the final answer.

Notice that the quantity (x - $\bar{x}$) gets larger the further we get from the mean; hence the confidence interval widens

Interval for a specific value of y, given a value of x, using t

$$\hat{y} \pm t_C \times s_{yx} \times \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\left(\frac{s_{yx}}{s_b}\right)^2}}$$

where all terms are as previously defined

This expression is very similar to that for an average value: it differs only in the inclusion of a "1" under the radical

Example: from above, the confidence interval is

$$\hat{y} \pm t_C \times s_{yx} \times \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\left(\frac{s_{yx}}{s_b}\right)^2}} = 8.3 \pm 2.92 \times 0.833 \times \sqrt{1 + \frac{1}{4} + \frac{(50 - 68.5)^2}{\left(\frac{0.833}{0.0395}\right)^2}} = 8.30 \pm 3.46$$

Using a Spreadsheet to Make Predictions

Use the regression output to create the regression equation somewhere in the spreadsheet. We also need to select a set of adjacent cells as entry cells, so that we can conveniently type in the x value for which we wish to make a prediction and the confidence level for interval predictions. Use the TINV function to find the t value for use calculating the confidence interval: =TINV(0.05, df) gives the t value for a 95% confidence interval; df can be found by entering the cell address for number of observations – 2.

Correlation

- Simple Correlation

  the coefficient of correlation (or Pearson's coefficient of correlation) is a measure of the closeness of the relation between x and y

  Definition: $r = \frac{\Sigma(x - \bar{x}) \times (y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2 \times \Sigma(y - \bar{y})^2}}$

  here the x's and y's are paired observations

  - Intuition of the correlation coefficient

    denominator serves to scale the expression; note that numerator ≤ denominator, so that $-1 \le r \le +1$

    look at terms in numerator

    when x is greater than its mean and y is greater than its mean, both terms are positive and product is positive

    if $x < \bar{x}$ while $y < \bar{y}$, the product is again positive

    but if when $x < \bar{x}$, $y > \bar{y}$, or vice versa, product is negative

    so, if x and y deviate from their respective means in the same direction, the sum of the products is positive and $r > 0$

    but if deviations are opposite, sum of products is negative, and $r < 0$

    Magnitude: if x and y are not associated, then large $(x - \bar{x})$ is as likely as not to be associated with small $(y - \bar{y})$ – the size of the products is simply random and the sum will be close to 0

    but if x and y are associated, then large x differences will be associated with large y differences, producing large products and a large sum

Example: Prices and unit sales of a particular model of rat trap were recorded on five occasions with these results:

Price	Units Sold
$10	3
8	5
12	2
6	8
4	12

Calculate r, the correlation coefficient

Solution: First, note that $\bar{x} = 8$ and $\bar{y} = 6$. Then arrange the data in a table as follows

x	$(x - \bar{x})$	$(x - \bar{x})^2$	y	$(y - \bar{y})$	$(y - \bar{y})^2$	$(x - \bar{x}) \times (y - \bar{y})$
10	2	4	3	-3	9	2 x - 3 = 6
8	0	0	5	-1	1	0
12	4	16	2	-4	16	-16
6	-2	4	8	2	4	-4
4	-4	16	12	6	36	-24
Σ =		40			66	-50

then $r = \dfrac{-50}{\sqrt{40 \times 66}} = -0.97$

In these calculations, look at the pattern of positive and negative deviations

if all pairs $(x - \bar{x}) \times (y - \bar{y})$ are positive or negative, then r will be positive or negative respectively

if some are positive and some are negative, the will tend to offset one another

the more this occurs, the closer will r be to zero

Now, actually, r is a statistic, that is, it's based on a sample. It is an estimate of the unknown parameter ρ (Greek letter *rho*) and is subject to sampling error.

ρ is what we would get if we computed the above expression for every possible pair of x and y – it represents the "true" degree of correlation between x and y

we can test for whether ρ actually is different from zero; for samples drawn independently from the same population, the following quantity is distributed as a t distribution with n - 2 degrees of freedom

$$t = \frac{r}{\sqrt{\dfrac{1 - r^2}{n - 2}}}$$

Thus we can test $H_0: \rho = 0$ vs. $H_1: \rho \neq 0$

from above

$$t = \frac{-0.97}{\sqrt{\dfrac{1 - (-0.97)^2}{5 - 2}}} = \frac{-0.97}{0.14} = -6.91$$

for a two-tailed test, with 3 degrees of freedom, at $\alpha = 0.01$, $t_C = \pm 5.841$, so we can reject H_0 and conclude $r \neq 0$

that is, we reject the premise that the observed correlation is due to sampling error and conclude that there is a linear relation between price and sales

- Using Excel to find r
  - Appears as "Multiple R" in regression output
    - attach the sign from regression coefficient
    - DO NOT DO THIS if there is more than one X variable
  - Use the Tools -Data Analysis -Correlation Procedure

- Data must be entered into adjacent columns: designate the whole block as Input Range

Output:

	Column 1	*Column 2*
Column 1	1	
Column 2	-0.93044	1

- the correlation between the variables in column 1 and column 2 is the correlation between X and Y

▶ Use the CORREL spreadsheet formula

CORREL(X range, Y range) returns the correlation coefficient

CORRELATION DOES NOT PROVE CAUSATION!