Business Statistics II: QM 122

Lecture Notes
by
Stefan Waner

(Ninth printing: 2009)

Department of Mathematics, Hofstra University

Table of Contents

1. Simple Linear Regression 2
2. The Coefficient of Determination and the Variability of the Random Term 10
3. Inferences about the Slope and Correlation Coefficient 16
4. Using Regression to Predict 27
5. Assumptions for Simple Linear Regression 33
6. Multiple Regression: The Model & Inferences about the β Parameters 38
7. Using F-statistics: Evaluating the Whole Model and Portions of the Model 44
8. Quadratic and Interactive Terms 50
9. Qualitative Variables 56
10. Single Factor Analysis of Variance 64
11. Multi Factor Analysis of Variance 73
12. Quality Improvement: Types of Variation 78
13. Using the Chi-Square ($\approx$2) Distribution 84
14**. Cyclic Fluctuations & Trigonometric Models** 94

Tables 93

Some Useful Pages:
Calculating Everything by Hand for Simple Regression: 17
The Excel Output Explained: 17

Topic 1
Simple Linear Regression (Based on §§ 14.1-14.2 in book)

A **linear function of one variable** is a function of the form

$$y = f(x) = \beta_0 + \beta_1 x,$$

where β_0 and β_1 are the **parameters** of the model. Its graph is a straight line with y-intercept β_0 and slope β_1. We also call a linear model a **first order** model.

In general, models specified by a mathematical equation are called **deterministic** models, since they hypothesize an exact relationship between x and y.

Examples 1
(a) $y = 4 + 3x$; $y = -22.345 - 4.01x$
(b) A **linear demand equation** has the form $y = \beta_0 + \beta_1 x$ where x is the unit price of an item, and y is a measure of the demand (e.g. monthly orders or sales).

Two Data Points
If we are given only two data points (x_1, y_1) and (x_2, y_2), then the equation of the line through them is given by

$$y = \beta_0 + \beta_1 x$$

where

$$\beta_1 = \frac{y_2 - y_1}{x_2 - x_1},$$

and $\beta_0 = y_1 - \beta_1 x_1$.

Worksheet 1
Find the equation of the line through (1, 3) and (3.2, 5).

Solution

$$\beta_1 = \frac{y_2 - y_1}{x_2 - x_1} = \frac{\qquad\qquad}{\qquad\qquad} =$$

$$\beta_0 = y_1 - \beta_1 x_1. =$$

Therefore, the equation of the line is

$$y = \beta_0 + \beta_1 x$$

$$\boxed{y = \qquad\qquad}$$

Probabilistic Models
In real life, we cannot expect an exact mathematical relationship between, say, price and demand, but we might hypothesize instead that the actual demand is given by, say

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

where ε is a **random error** component. Such a model is called a **probabilistic model**.

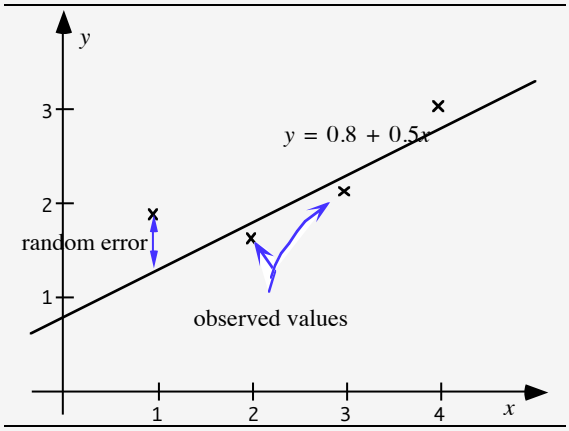
Q Exactly what is the random error?
A Actually, ε is a random variable, specified for each value of x, as follows. For a fixed value of x, the experiment consists of measuring y, and then subtracting the theoretical prediction $\beta_0 + \beta_1 x$ from the result.

Q What do you mean by a "random variable specified for each value of x?"
A This means, we actually have lots of random variables ε_x, one for each value of x. However, we shall be making the assumption that all of the ε_xs have the same normal distribution, so can drop the subscript and write ε instead.

Probabilistic First Order Model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

y = dependent variable (also called the response variable) — that is what is being modeled.
x = independent variable (also called the predictor variable)
β_0 = y-intercept
β_1 = slope; the increase of y per one unit increase in x.
ε = random error: a random normal variable with mean 0 that does not depend on x.



The deterministic part of the function,

$$E(y) = \beta_0 + \beta_1 x$$

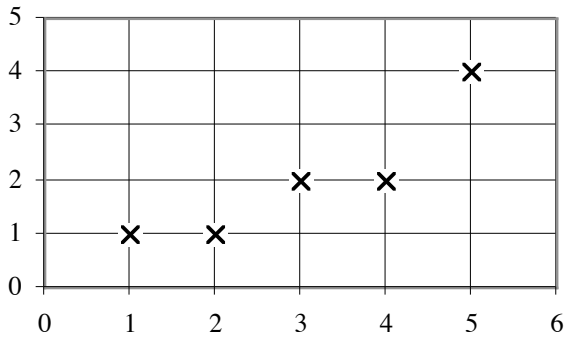
is referred to as the **line of means**, since the mean of $\bar{y} = \beta_0 + \beta_1 x + \bar{\varepsilon} = \beta_0 + \beta_1 x$.

Best Fit Line: Least Squares
In the simplest case we have two data points and we only need to find the equation of the line passing through them. However, it often happens that we have many data points that don't quite all lie on one line. The problem then is to find the line coming closest to passing through all of the points.

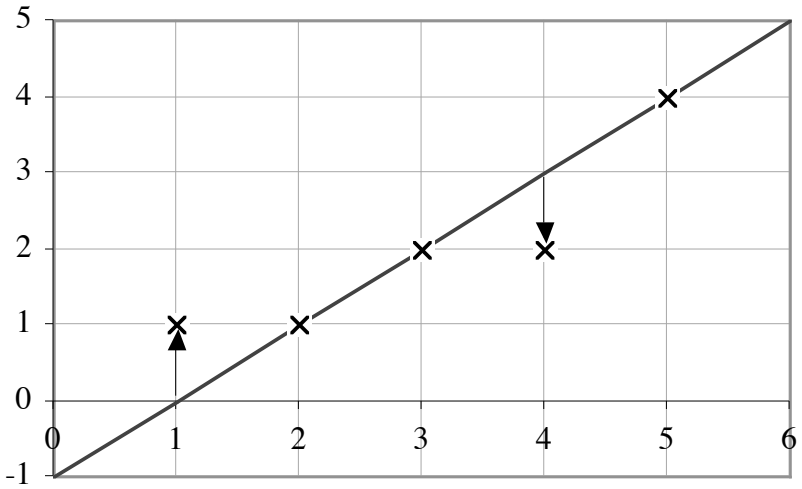
Example 2
The following table list measured values of sales revenues for various advertising expenditures.

Advertising Expenditure (x) ($100)	1	2	3	4	5
Sales Revenue (y) ($1000)	1	1	2	2	4

First, we plot the given data in a **scattergram.**



The regression line will be the one that minimizes the **sum of the squares of the errors (SSE)** (also known as the **sum of the squares of the residuals**), as shown. (The errors are given by **observed value − predicted value**):



From the chart, we see that SSE = 2 and SE = 0 for the given line (although it is not the regression one.) To obtain the actual regression line, we would have to adjust the line above until we obtained the lowest SSE.

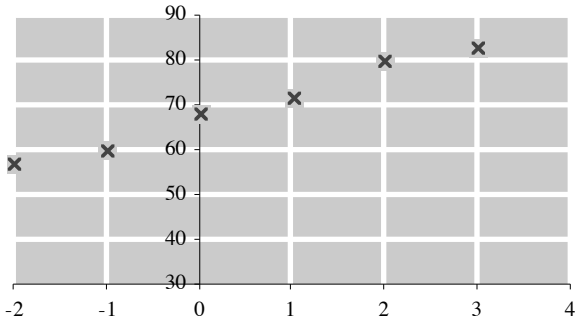
Worksheet 2

You are conducting research for a cable TV company interested in expanding into China and we come across the following figures showing the growth of the cable market there.

Year (x) (x= 0 represents 2000)	−2	−1	0	1	2	3

Households with Cable (y) (Millions)	57	60	68	72	80	83

Data are approximate, and the 2001–2003 figures are estimates. Sources: HSBC Securities, Bear Sterns/*New York Times*, March 23, 2001, p. C1.



Use Excel to compute SSE for the linear models $y = 72 + 8x$ and $y = 68 + 5x$. Which model is the better fit?

Model 1: $y = 72 + 8x$

Year x	Observed y	Predicted $\hat{y} = 72 + 8x$	**Residual** $y - \hat{y}$	**Residual^2** $(y - \hat{y})^2$
−2	57	56	$57-56=1$	$1^2=1$
−1	60			
0	68			
1	72			
2	80			
3	83			
			SSE =	

Model 2: $y = 68 + 5x$

Year x	**Observed** y	**Predicted** $\hat{y} = 68 + 5x$	**Residual** $y - \hat{y}$	**Residual^2** $(y - \hat{y})^2$
−2	57			
−1	60			
0	68			
1	72			
2	80			
3	83			
			SSE =	

Better model = Model with smaller SSE = _____

Least Squares (Regression) Line

The least squares line associated with the points (x_i, y_1) is the line the *minimizes the sum-of-squares error, SSE*, and has the form

$$\hat{y} = b_0 + b_1 x$$

with

$$\text{Slope } b_1 = \frac{SS_{xy}}{SS_{xx}}$$

$$\text{Intercept } b_0 = \bar{y} - b_1\bar{x}$$

where

$\bar{x}$ and $\bar{y}$ are te sample means of x and y,

$$SS_{xy} = \Sigma x_i y_i - \frac{(\Sigma x_i)(\Sigma y_i)}{n} \qquad = \Sigma(x_i - \bar{x})(y_i - \bar{y})$$

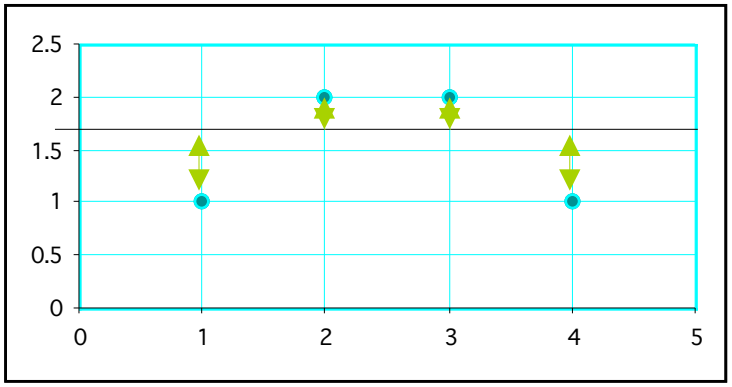
$$SS_{xx} = \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n} \qquad = \Sigma(x_i - \bar{x})^2$$

n = sample size

Also,

$$\text{SSE} = \text{sum of squares of errors} = \Sigma(y_i - \hat{y}_i)^2 = SS_{yy} - b_1 SS_{xy}$$

Question Why minimize SSE and not, say, the absoute values of the errors?
Answer There are two important reasons:
(1) Mathematically, it generalizes the notion of the sample mean. For instance, look at the following points:



Any horizontal line will minimize the sum of the distances shown (the absolute values of the errors). However, only the line $y = 1.5$ (which happens to be the mean of the given y-values) will minimize SSE. In general, the sample mean $\bar{x}$ of a collection of values of X is the uniqu number that minimizes SSE, and not the absolute values of the errors.
(2) Statistically, the regression line gives us unbiased estimators of the population parameters β_0 and β_1. (Also see the discussion in Topic 2.)

Worksheet 3 A Tabular Approach for By-Hand Calculation or Excel Calculation

Compute the regression line using the given data, and supply the missing information:

x = Advertising expenditure in hundreds of dollars

y = Sales revenue in thousands of dollars

	x	y	x^2	xy	y^2
	1	1			
	2	1			
	3	2			
	4	2			
	5	4			
Σ (Sum)					
Means					

Note that the values in the bottom rows are the sums of the entries in that column. Substituting these values in the formula gives ($n = 5$)

$$SS_{xy} = \Sigma x_i y_i - \frac{(\Sigma x_i)(\Sigma y_i)}{n} = \square - \frac{\square}{\square} = \square$$

$$SS_{xx} = \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n} = \square - \frac{\square}{\square} = \square$$

$$SS_{yy} = \Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n} = \square - \frac{\square}{\square} = \square$$

Regression Coefficients:

$$b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\square}{\square} = \square$$

$$b_0 = \bar{y} - b_1\bar{x} = \square - \square \quad \square = \square$$

Thus, the least squares line is

$$\hat{y} = b_0 + b_1 x : \boxed{\hat{y} =}$$

$$SSE = SS_{yy} - b_1 SS_{xy} = \square - \square \quad \square = \square$$

Interpreting the coefficients:

For every 1-unit increase in x, y increases by b_1 units

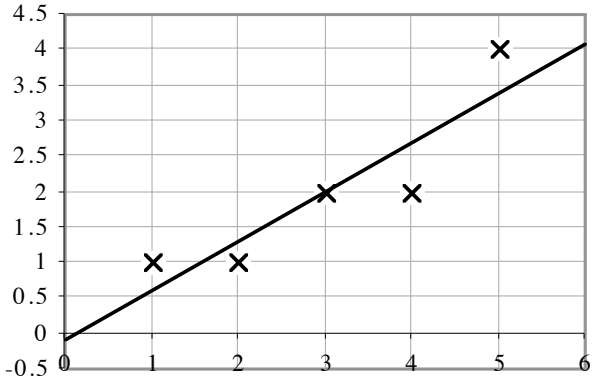
For every _____ increase in _____ , _____ increases by _____ .

Estimating y:

If I pay \$3500 ($x = 3.5$) for advetrtising, I can expect sales revenues of

$\hat{y}$ = _____

Here is a graph of the regression line with the associated data:



Using Excel Regression Output to obtain the regression line:
Following is some of the Excel regression output for this data

Regression Statistics	
Multiple R	0.90369611
R Square	0.81666667
Adjusted R Square	0.75555556
Standard Error	0.60553007
Observations	5 ← n

ANOVA

	df	*SS*	*MS*	*F*
Regression	1	SST → 4.9	4.9	13.3636364
Residual	3	SSE → 1.1	0.36666667	
Total	4	SS_{yy} → 6		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	b_0 → -0.1	0.6350853	-0.1574592	0.88488398
X Variable 1	b_1 → 0.7	0.19148542	3.65563078	0.03535285

The values of the regression coefficients are shown in the indicated cells, and the sum of the squares error (SSE) is listed in the ANOVA (Analysis of Variance) part under SS.

Execicses for this topic:

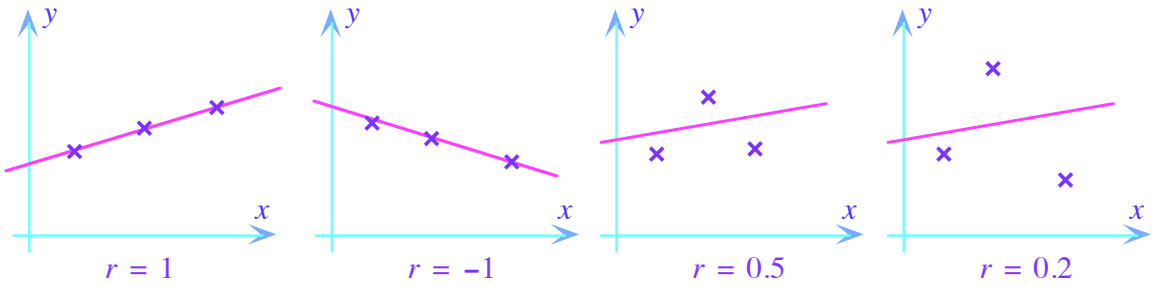
p. 553: #4 (compute the regression equation two ways: (1) by hand (2) Using the Excel regression analysis.)
p. 557 #12 (As above)

Topic 2
The Coefficient of Determination and the Variability of the Random Term.
(§14.3 and part of §14.5 in the text)
Coefficient of Determination

Question If my data points do not all lie on one straight line, how can I measure how closely they can be approximated by a straight line?

Answer Think of SSE for a moment. It measures the sum of the squares of the deviations from the regression line, and therefore itself constitutes a measurement of goodness of fit. (For instance, if SSE = 0, then all the points lie on a straight line.) However, SSE depends on the units we use to measure y, and also on the number of data points (the more data points we use, the larger SSE tends to be). Thus, while we can (and do) use SSE to compare the goodness of fit of two lines to the same data, we cannot use it to compare the goodness of fit of one line to one set of data with that of another to a different set of data.

To remove this dependency, statisticians have found a related quantity that can be used to compare the goodness of fit of lines to different sets data. This quantity, called the **coefficient of determination, coefficient of correlation** or **correlation coefficient**, and usually denoted r, is between -1 and 1. The closer r is to -1 or 1, the better the fit. For an *exact* fit, we would have $r = -1$ (for a line with negative slope) or $r = 1$ (for a line with positive slope). For a bad fit, we would have r close to 0.the figure shows several collections of data points with least squares lines and the corresponding values of r.



In the Excel printout, r is found here:

SUMMARY OUTPUT

Regression Statistics		
Multiple R	0.90369611	$\leftarrow r$
R Square	0.81666667	$\leftarrow r^2$
Adjusted R Square	0.75555556	
Standard Error	0.60553007	
Observations	5	

Question How do we compute and interpret r?

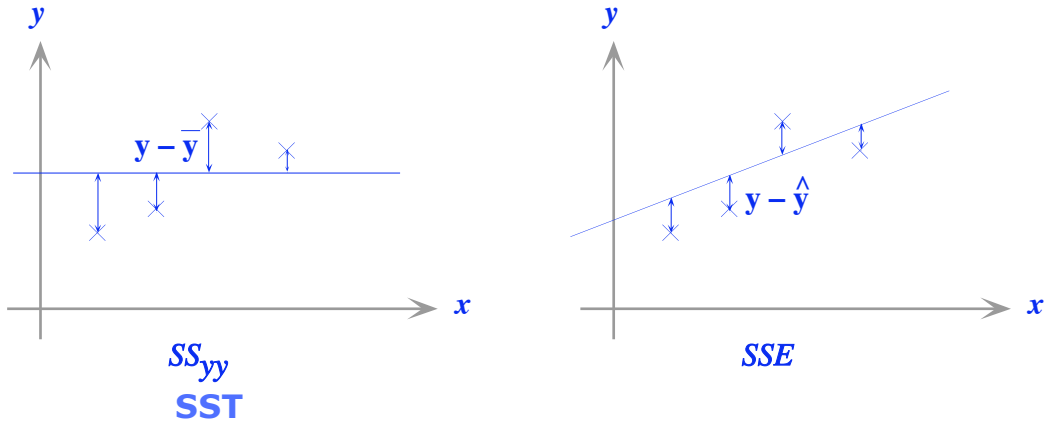
Answer Actually, it's easier to compute and interpret r^2. First, let's see what the quantities SS_{yy} and SSE measure:

SS_{yy} is the sample variation in y (in fact, its expected value is the variance of y and measures the deviation of y from the mean $\bar{y}$ of all values of y.

Note SS_{yy} is also called SST, the **Total Sum of Squares** and can also be computed by

$$\text{SST} = \Sigma(y_i - \bar{y})^2 = SS_{yy}$$

SSE measures the deviation of y from the linear predicted $\hat{y}$



SST − SSE measures the part of the deviation of y from the mean that can be attributed to x. (For a perfect linear set of data SSE = 0, so all the deviation of y from the mean can be attributed to the value of x. On the other hand, if β_1 was 0, then SST = SSE, so none of the deviation of y from the mean is attributable to the value of x.)

This quantity SST − SSE is also called the **sum of squares due to regression,** and is denoted by SSR.

Thus, the *proportion* of the total sample variation that can be attributed to x is given by

$$r^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\text{SST} - \text{SSE}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

and this is just the square of the coefficient of determination.

Coefficient of Determination r

$$r^2 = \frac{SS_{xy}{}^2}{SS_{xx}\,SS_{yy}} \qquad \text{or} \qquad r^2 = \frac{\text{SSR}}{\text{SST}}$$

By Hand **Excel**

It appears under "Multiple R" in the Excel regression analysis.

Interpretation r^2 is the proportion of the sample variation in y attributable to the value of x in a linear relationship.

Quick Example If $r^2 = 0.85$, then approximately 85% of the sample variation of the value of y is due to the value of x (assuming a linear relationship).

Worksheet 1 — Obtaining and Interpreting r^2 from Excel Output

Following is a *partial* Excel Output for a regression analysis of profit ($ million) as a function of time (years since 1995). Use the given data to compute r, and interpret the result.

SUMMARY OUTPUT

ANOVA

	df	*SS*	*MS*	*F*
Regression	1	2.9575	2.9575	2.87874574
Residual	5	5.13678571	1.02735714	
Total	6	8.09428571		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-2.6714286	0.85663676	-3.118508	0.02629654
X	-0.325	0.1915498	-1.6966867	0.15051834

Solution We use

$$r^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\square}{\square} = \square$$

so $r = \sqrt{r^2} \approx \square$

Thus, approximately _____% of the variation of profits is due to the value of time, if we assume a linear model. (The rest of the variation is due to "statistical noise.")

Worksheet 2 — Calculating Correlation by Hand

Use the "By Hand" table above to compute r^2 for the following data:

x	−2	0	2	4	6
y	−1	−2	−4	−3	−5

x	y	x^2	xy	y^2
−2	−1			
0	−2			
2	−4			
4	−3			
6	−5			
Σ (Sum)				

$$r^2 = \frac{SS_{xy}^{\ 2}}{SS_{xx}\,SS_{yy}} = \frac{\square}{\square} \approx \square$$

so $r = \sqrt{r^2} \approx$ ☐

Variability of the Random Term

First we record some consequences of the thre basic assumptions for lienar regression. Revall that our original probabilitic model is

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where ε is normal with mean 0 an standard deviation σ. This quantity ε meqasures the variability of the random term, so the question is, how do we estimate it?

Variability of the Random Term

An unbiased estimator for σ^2 is given by

$$s^2 = \frac{\text{SSE}}{n-2} = \frac{\text{SSE}}{\text{Degrees of freedom}}$$ [1]

where SSE can be calculated from the following formula:

$$\text{SSE} = \text{sum of squares of errors} = \Sigma(y_i - \hat{y}_i)^2 = SS_{vv} - b_1 SS_{xv}$$

s is called the **estimated standard error of the regression model**. (s^2 is also called the **mean square error.**)

---

[1] There are $n-2$ degrees of freedom because we are estimating two parameters β_0 and β_2.

In the Excel output, s appears in the "Regression Statistics" part of the table. See if you can also find the Mean square error in the table.

SUMMARY OUTPUT		
Regression Statistics		
Multiple R	0.90369611	
R Square	0.81666667	
Adjusted R Square	0.75555556	
Standard Error	0.60553007	← s
Observations	5	

Interpretation of s

We can use the Empirical rule here. It says that approximately 67% of all the observed values of y are within $\pm\sigma$ of $\bar{y}$, Using this for sample statitics, we deduce that around 67% of all the observed values of y are within $\pm s$ of $\hat{y}$. Similarly, approximately 95% of all the observed values of y should lie within $\pm 2s$ of $\hat{y}$.

Quick Example Look at the above Excel regression analysis.

SSE = sum of squares of residuals = 1.1

Thus, $s^2 = \dfrac{1.1}{n-2} = \dfrac{1.1}{5-2} \approx 0.36666667,$

so $s \approx 0.605530.$

Worksheet 3 — Calculating s by Hand

Let us go back to the original data we were using before:

x = Advertising expenditure in hundreds of dollars

y = Sales revenue in thousands of dollars

Compute S_{YX}, and confirm this by looking up the value in the Excel output.

	x	y	x^2	xy	y^2
	1	1			
	2	1			
	3	2			
	4	2			
	5	4			
Σ (Sum)					
Means					

$$SS_{xy} = \Sigma x_i y_i - \frac{(\Sigma x_i)(\Sigma y_i)}{n} = \square - \frac{\square}{\square} = \square$$

$$SS_{xx} = \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n} = \square - \frac{\square}{\square} = \square$$

$$SS_{yy} = \Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n} = \square - \frac{\square}{\square} = \square$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\square}{\square} = \square$$

$$SSE = SS_{yy} - b_1 SS_{xy} = \square - \square \; \square = \square$$

$$s^2 = \frac{\text{SSE}}{n-2} = \frac{\square}{\square} \approx \square$$

$$s = \sqrt{s^2} \approx \square$$

Excel Regression Analysis: $s \approx$ ☐

Interpretation: In the *residual plot,* we can expect to find approximatley 95% (that is, almost all) the observed values between _____ and _____ .

Exercises for this topic:
p.563 # 15: Use the By-Hand formulas above to compute r^2 and not the ones in the book. Then use Excel to compuate it as we did.
p. 564 # 20: Excel only when computing r^2.
Also, compute s for both above exercises.

Topic 3
Inferences about the Slope and Correlation Coefficient
(Based on §14.5 in book)

In this topic, we will be able to answer several questions. Here is the first one.

Question 1
Does *y* really depend on *x* based on the given data?

For example, if you did a regression of a person's blood pressure as a function of his or her age, you would expect the answer to be yes, but if you did a regression of a soccer player's scoring average per game as a function of the number of kibbles and bits my pet chia ate on that day, you would expect the answer to be "no." How can we analyze this in less obvious situations?

First Recall that we have made an assumption that $y = \beta_0 + \beta_1 x + \varepsilon$. The quantities β_0 and β_1 are thus, in effect, hypothesized **parameters** of the population from which the data is sampled. When we computed, we were *making estimates of these parameters*.

Q Why can't we just take the mean of a sample to estimate β_0 and β_1? It worked in QM1...
A We have no way of sampling β_0 and β_1 separately: all we can sample is y.

Q OK, so we used those strange formulas to compute b_0 and b_1. Are they at least unbiased estimators of β_0 and β_1?
A Yes: given our assumptions about the model (above), the sampling distribution of b_1 is always normal with mean β_1 (since it is an unbiased estimator) and standard deviation

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{SS_{xx}}} \approx \frac{s}{\sqrt{SS_{xx}}} .$$

We call this latter quantity s_{b_1}. So,

$$s_{b_1} = \frac{s}{\sqrt{SS_{xx}}}$$

Before going on, we really need to (1) know where all this stuff is in the Excel table, and (2) know how to get these numbers by hand. The following formulas include things we have'nt yet gotten to, but have patience...

The Excel Table Explained (So Far):

Regression Statistics	
Multiple R	r
R Square	r^2
Adjusted R Square	
Standard Error	s
Observations	n

ANOVA

	df	*SS*	*MS*	*F*
Regression	k	*SSR*	*MSR*	*MSR/MSE*
Residual	$n-k-1$	SSE	*MSE*	
Total	$n-1$	SS_{yy}	*SSyy*/$(n-1)$	

	Coefficients	*Standard Error*	*t Stat*	*P-value*	
Intercept	b_0	s_{b_0}	*test statistic* for H_0: $\beta_0 = 0$:	*p*-value for two-tailed test	0ne-tailed: divide this by 2
X	b_1	s_{b_1}	*test statistic* for H_0: $\beta_1 = 0$:	*p*-value for two-tailed test	

Calculating Everything By Hand

All we need are the quantities SS_{xx}, SS_{xy}, SS_{yy}, and $\bar{x}$, $\bar{y}$

$$SS_{xx} = \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n} \qquad SS_{xy} = \Sigma x_i y_i - \frac{(\Sigma x_i)(\Sigma y_i)}{n} \qquad SS_{yy} = \Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n}$$

$$\bar{x} = \frac{\Sigma x_i}{n} \qquad \bar{y} = \frac{\Sigma y_i}{n}$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}} \qquad b_0 = \bar{y} - b_1\bar{x}$$

$$\text{SSE} = SS_{yy} - b_1 SS_{xy}$$

$$r^2 = 1 - \frac{\text{SSE}}{SS_{yy}} \qquad s^2 = \frac{\text{SSE}}{n-2}$$

$$s_{b_1} = \frac{s}{\sqrt{SS_{xx}}} \qquad s_{b_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}}}$$

$$t = \frac{b_1}{s_{b_1}} \quad (\text{for } H_0: \beta_1 = 0) \qquad t = \frac{b_0}{s_{b_0}} \quad (\text{for } H_0: \beta_0 = 0)$$

Confidence Interval for β_1: $b_1 \pm t_{\alpha/2}\, s_{b_1}$

Worksheet 1 — Computing the Terms in the Excel Output by Hand

	x	y	x^2	xy	y^2
	1	1	1	1	1
	2	1	4	2	1
	3	2	9	6	4
	4	2	16	8	4
	5	4	25	20	16
Σ (Sum)	15	10	55	37	26
Means	3	2			

$$SS_{xy} = \Sigma x_i y_i - \frac{(\Sigma x_i)(\Sigma y_i)}{n} = 37 - \frac{(15)(10)}{5} = 7$$

$$SS_{xy} = \Sigma x_i y_i - \frac{(\Sigma x_i)(\Sigma y_i)}{n} = 37 - \frac{(15)(10)}{5} = 7$$

$$SS_{xx} = \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n} = 55 - \frac{15^2}{5} = 10$$

$$SS_{yy} = \Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n} = 26 - \frac{10^2}{5} = 6$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{7}{10} = 0.7$$

$$b_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{10}{5} - 0.7\frac{15}{5} = -0.1$$

$$SSE = SS_{yy} - b_1 SS_{xy} = \square - \square \; \square = \square$$

$$s^2 = \frac{\text{SSE}}{n-2} = \frac{\square}{\square} \approx \square$$

$$s = \sqrt{s^2} \approx \square$$

$$r^2 = 1 - \frac{\text{SSE}}{SS_{yy}} = 1 - \frac{\square}{\square} = \square$$

$$r = \sqrt{r^2} \approx \square$$

$$S_{b_1} = \frac{s}{\sqrt{SS_{xx}}} = \frac{\square}{\square} \approx \square$$

$$S_{b_0} = s\sqrt{SSE} = \square \sqrt{\square} = \square$$

$$t_{b_1} = \frac{b_1}{s_{b_1}} = \frac{\square}{\square} \approx \square$$

$$t_{b_0} = \frac{b_0}{s_{b_0}} = \frac{\square}{\square} \approx \square$$

Regression Statistics	
Multiple R	
R Square	
Adjusted R Square	
Standard Error	
Observations	

ANOVA

	df	*SS*	*MS*	*F*
Regression				
Residual				
Total				

	Coefficients	*Standard Error*	*t Stat*
Intercept			
X			

OK back to the task at hand.

Now, recall that when we measured a sample mean, we used the sample information to test a hypothesis about the population mean. Here, we will test a hypothesis about the parameter β_1 to answer the following question:

Question 1: Does *y* depend on *x* at all? Note that, if y did not depend on x, then β_1 would be zero. Thus, let us test the following hypothesis:

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$ or $H_1: \beta_1 > 0$ or $H_1: \beta_1 < 0$.

To test these at a given significance level, we use the above information about the sampling distribution of b_1 to obtain a test statistic, and use the t-distribution based on $(n-2)$ degrees of freedom.

Testing Model Utility

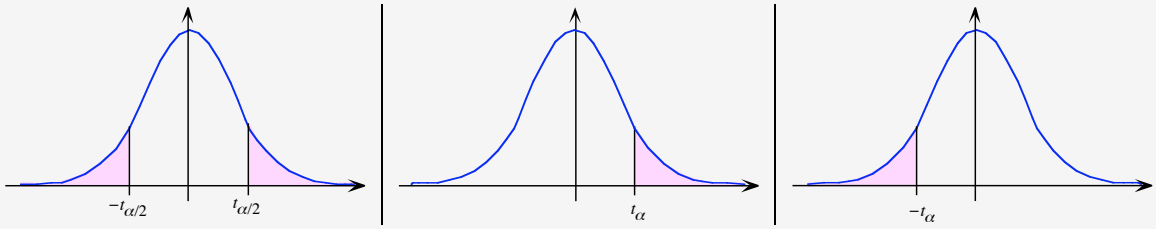
Test Statistic

$$t = \frac{b_1 - \text{Hypothesized value of } \beta_1}{s_{b_1}} = \frac{b_1}{s_{b_1}} = t\text{-stat on Excel}$$

where $s_{b_1} = \dfrac{s}{\sqrt{SS_{xx}}}$ = Standard error on Excel Table

Two-Tailed	**One-Tailed; Upper**	**One-Tailed; Lower**
H_0: $\beta_1 = 0$	H_0: $\beta_1 = 0$	H_0: $\beta_1 = 0$
H_1: $\beta_1 \neq 0$	H_1: $\beta_1 > 0$	H_1: $\beta_1 < 0$

Rejection Regions (t_α and $t_{\alpha/2}$ are based on $(n-2)$ degrees of freedom)



Using the *p*-Statistic:

Two-Tailed Tests	**One-Tailed Tests:**
Use it as is	Divide it by 2

Worksheet 2 - Testing the Regression Coedfficient (*t* Test)

The following data suggests a relationship between famy income and SAT scores.

Parents' Income ($1000)	5	15	25	35	45	55	65
Verbal SAT	350	377	402	416	429	437	446

Source: The College Board/*The New York Times*, March 5, 1995, p. E16.

Test, at the 95% level of significance, whether SAT scores go up as family income increases. Interpret the coefficient b_1 and also the result of the hypothesis test.

Regression Statistics	
Multiple R	0.97214532
R Square	0.94506653
Adjusted R Square	0.93407983
Standard Error	8.86364968
Observations	7

ANOVA

	df	*SS*	*MS*	*F*
Regression	1	6758.03571	6758.03571	86.0191836
Residual	5	392.821429	78.5642857	
Total	6	7150.85714		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	353.767857	6.75243248	52.3911728	4.7898E-08
x	1.55357143	0.16750723	9.27465275	0.00024501

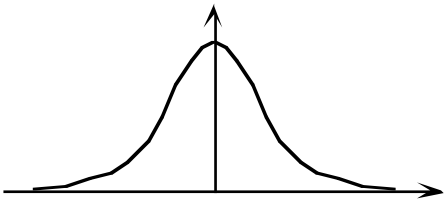
H_0: _____

H_a: _____

$df = n - 2$ = _____

Rejection region:

critical t = _____



$t-$statistic: _____ In rejection region? _____

Conclusion: _____

Interpret ting result of hypothesis test:

Interpreting coefficient b_1:

Note: If t did not fall in the rejection region, that would not have meant that we must accept H_0: $\beta_1 = 0$. All it means that we cannot conclude that β_1 is positive.

Guess what: The corresponding p-value (which works as is for the two-tailed test) is right next to it! Thus, if we are testing at the 95% confidence level, $\alpha = 0.05$ and the p-value is 0.03, we can safely reject H_0 since the p-value is smaller than α.
For the one-tailed test, we use half the given p-value to estimate α, and so

$$p \approx .000\,24501/2 \approx 0.000\,1225$$

so we can certainly reject H_0 with a signficance level of

$$1 - 0.000\,1255 = 0.999\,8745,$$

or 99.99%.

Excel Note: We can compute the 1-tailed or 2-tailed p-value from any t-statistic using the formula

=TDIST(t-stat. ,df , Number of tails (1 or 2))

For instance, the p-value for the above test is

```
=TDIST(52.3911728,5,1)
```

We are ready for the next question.

Question 2 I got a slope of b_1. What is the confidence interval for that answer?

To answer this question, we use knowledge we already have: if a random variable X is normally distributed with population mean μ and standard deviation σ, then if we take a sample of X, the $(1-\alpha)$ confidence interval for μ is $\bar{x} \pm z_{\alpha/2}\sigma/\sqrt{n}$. The reason this works is that the sampling distribution of $\bar{x}$ is normal with mean μ and standard deviation $\sigma_{\bar{X}} = \sigma/\sqrt{n}$. But here, we saw above that the sampling distribution of b_1 is always normal with mean β_1 and standard deviation S_{b_1}. Thus, we get the following confidence interval test:

How to find the $(1-\alpha)$ Confidence Interval for the Slope β_1
We can be $100(1-\alpha)\%$ certain that β_1 is in the interval

$$b_1 \pm t_{\alpha/2}\, S_{b_1}$$

where $t_{\alpha/2}$ is based on $(n-k-1)$ degrees of freedom. (For simple regression, $k = 1$, so use $n-2$ degrees of freedom.)

Excel

- The quantity S_{b_1} is the standard error in the slope, and appears in the X Variable 1 row under "Standard Error."
- To obtain $t_{.025}$ without using a table, enter `=TINV(.05, DF)`

Worksheet 3 – Confidence Interval for the Slope

Use the Excel printout of the preceding worksheet to compute a 95% confidence interval for the slope.

b_1 = _____ df = _____

$t_{\alpha/2} = t_{\_\_\_}$ = _____

S_{b_1} = _____

CI = [ ] ± [ ]

= [ ] ± [ ]

= [,]

Interpretation:

Question 3: What is the F-statistic and what does it tell us?

The F-statistic is defined to by

$$F = \frac{MSR}{MSE} = \frac{SSR/k}{SSE/(n-k-1)}$$

If $\beta_1 = 0$, then

$SS_{yy} \approx SSE$ Since the regression line should be close to horizontal

Therefore

$SSR = SS_{yy} - SSE$

should be close to 0, resulting in a small value of F. The probability distribution for F assuming that $\beta_1 = 0$ is known, and depends on two degrees of freeedom: $k = 1$ in the numerator and $n-k-1 = n-2$ in the denominator. Some of its critical values are given by the following table (a more complete table appears at the end of this booklet):

Critical Values of F ($\alpha = 0.05$)

Excel: =FINV(0.05,df_n,df_d)

df Denominator ↓	Numerator → 1	2	3	4	5	6
1	161.446	199.499	215.707	224.583	230.160	233.988
2	18.513	19.000	19.164	19.247	19.296	19.329
3	10.128	9.552	9.277	9.117	9.013	8.941
4	7.709	6.944	6.591	6.388	6.256	6.163
5	6.608	5.786	5.409	5.192	5.050	4.950
6	5.987	5.143	4.757	4.534	4.387	4.284
7	5.591	4.737	4.347	4.120	3.972	3.866
8	5.318	4.459	4.066	3.838	3.688	3.581
9	5.117	4.256	3.863	3.633	3.482	3.374
10	4.965	4.103	3.708	3.478	3.326	3.217
11	4.844	3.982	3.587	3.357	3.204	3.095
12	4.747	3.885	3.490	3.259	3.106	2.996
13	4.667	3.806	3.411	3.179	3.025	2.915
14	4.600	3.739	3.344	3.112	2.958	2.848
15	4.543	3.682	3.287	3.056	2.901	2.790
16	4.494	3.634	3.239	3.007	2.852	2.741
17	4.451	3.592	3.197	2.965	2.810	2.699
18	4.414	3.555	3.160	2.928	2.773	2.661
19	4.381	3.522	3.127	2.895	2.740	2.628
20	4.351	3.493	3.098	2.866	2.711	2.599

Note: A slight disadvantage of the *F*-statistic is that it does not differentiate between positive and negative slopes. Therefore, we only use it with an alternate hypothesis of the form

H_a: $\beta_1 \neq 0$

Worksheet 4 - Testing the Regression Coedfficient (*F* Test)

Life expectancies at birth in the United States for people born in various years is given in the following table. You can download this data at

http://www.zweigmedia.com/qm203/

under "Life Expectancy Data".

Year Since 1920	0	10	20	30	40	50	60	70	78
Life Expectancy	54.1	59.7	62.9	68.2	69.7	70.8	73.7	75.4	76.7

Source: Centers for Disease Control and Prevention, National Center for Health Statistics, *National Vital Statistics Report*, Feb. 7, 2001. http://www.cdc.gov/nchs/fastats/pdf/nvsr48_18tb12.pdf

Use an F-test to determine, at the 95% level of significance, whether the life-expectancy has been changing with time. Interpret the coefficient b_1

Regression Statistics

Multiple R	0.97369201
R Square	0.94807614
Adjusted R Square	0.94065844
Standard Error	1.85070591
Observations	9

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	437.773102	437.773102	127.812772	9.4773E-06
Residual	7	23.9757864	3.42511235		
Total	8	461.748889			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	57.0236538	1.14367512	49.8600107	3.4175E-10	54.3192938
x	0.27370703	0.02421022	11.3054311	9.4773E-06	0.21645898

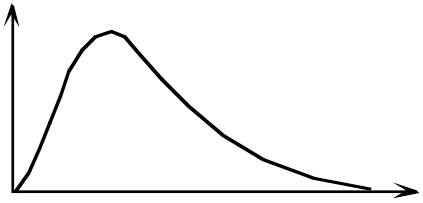
H_0: _____

H_a: _____

df (numerator) = k = _____ *df*(denominator) = $n - k - 1$ = _____

F_{critical} = _____

Rejection region:



F−statistic: _____ In rejection region? _____

Conclusion: _____

Interpret ting result of hypothesis test:

Interpreting coefficient b_1:

Note: The "*p*-value" for *F* is listed above as Significance *F*, and gives the signficance with which we can reject Note that it is the same as the *p*-value for the *t*-statirci associated with β_1. (Why?)

Exercises for this topic:
p. 575 #26. First, compute all the terms in the Excel Sheet by hand. Then answer questions a-c. Finally, check your calculations by doing an Excel regression analysis

p. 576 #30 Excel only

Topic 4
Using Regression to Predict *y* from *x*
(Based on §14.6 in the book)

Let us go back to the original scenario: x = monthly advertising expenditure in \$100, y = monthly sales revenue in \$1000. We can already predict y by using the formula for $\hat{y}$. What we don't have is a confidence interval. Here are two questions we can ask:

Question 1 What is a confidence interval for my *average* sales revenue if I pay \$3500 (x = 3.5) for advertising in a month? That is, find a conrfidence interval for $\bar{y}$, the population mean of y, given a specific value of x. This confidence interval is called a **confidence interval (CI) for the mean of *y*.**

Question 2 What is a confidence interval for my sales revenue *in a particular month* if I pay \$3500 per month (x = 3.5) for advertising? That is, find a conrfidence interval for y, a particular value of y, given a specific value of x. This confidence interval is called a **prediction interval (PI) for an individual response of *y*.**

Question What is the difference between these two confidence intervals?
Answer We are less certain about a particular month's revenues than about the mean revenue. Therefore, the confidence interval for a particular value of y will be larger.

All we need is a confidence interval for this prediction.

Question Wait a minute! Since $y = \beta_0 + \beta_1 x + \varepsilon$, and the standard deviation ε can be estimated by s, why not just use that s to give us a confidence interval for y?
Answer If we knew the values of β_0 and β_1 exactly, that would be fine. But we don't know those values; we only have *estimates* b_0 and b_1 of those values. Thus, we can't use s to form our confidence interval.

Note The tiniest error in b_0 or b_1 can have disastrous consequences for long-term prediction (illustration in class). Thus, the standard deviation and resulting confidence interval for our prediction of y from a given value of x should depend on how far that value of x is from the mean $\bar{x}$.

Here are the underlying facts:

- If we use the regression value $\hat{y}$ to predict $\bar{y}$ at a specified value x_p *of* x, then the standard deviation of the error is given by
$$\sigma_{(y-\bar{y})} = \sigma\sqrt{\frac{1}{n} + \frac{(x_p-\bar{x})^2}{SS_{xx}}}$$
where σ is the standard deviation of ε.

- If we use the regression value $\hat{y}$ to predict y at a specified value x_p *of* x, then the standard deviation of the error is given by

$$\sigma_{(y-\hat{y})} = \sigma\sqrt{1 + \frac{1}{n} + \frac{(x_p-\bar{x})^2}{SS_{xx}}}$$

If we now recall that $\sigma \approx S_{YX}$, where $S_{YX}^2 = \frac{\text{SSE}}{n-2}$, we see that our $(1-\alpha)$ confidence intervals for $\bar{y}$ and y_p are given as follows:

Predicting $\bar{y}$ and y from a given x

A $(1-\alpha)$ confidence interval for $\bar{y}$ is given by

$$\hat{y}_p \pm t_{\alpha/2} s\sqrt{\frac{1}{n} + \frac{(x_p-\bar{x})^2}{SS_{xx}}} \qquad \textbf{CI for the Mean of } y$$

A $(1-\alpha)$ confidence interval for y_p is given by

$$\hat{y}_p \pm t_{\alpha/2} s\sqrt{1 + \frac{1}{n} + \frac{(x_p-\bar{x})^2}{SS_{xx}}} \qquad \textbf{PI for an Individual Response of } y$$

In both cases,
x_p is the given value of x
$\hat{y}_p$ is the resulting value of y (given by the regression equation)
$t_{\alpha/2}$ is based on $(n-k-1) = (n-2)$ degrees of freedom.

Excel:

- To obtain $t_{.025}$ without using a table, enter `=TINV(.05, DF)`
- To obtain SS_{xx} from the Excel output we can use

$$SS_{xx} = \left[\frac{s}{s_{b_1}}\right]^2$$

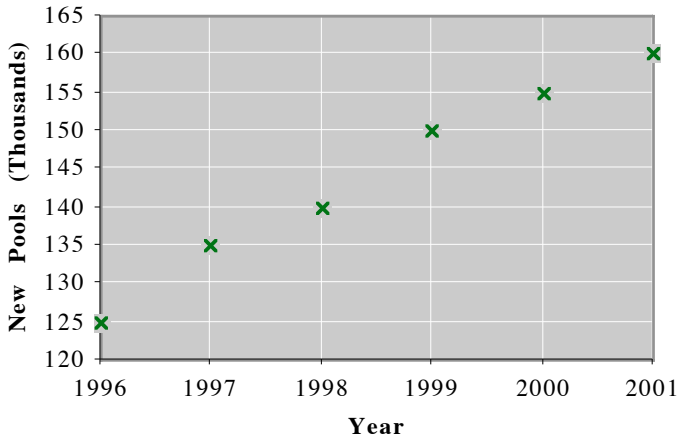
Note: To get the above formula for SS_{xx}, we use the formula $s_{b_1} = \frac{s^2}{\sqrt{SS_{xx}}}$, giving

$$SS_{xx} = \left[\frac{s}{s_{b_1}}\right]^2.$$

Worksheet 1 — Mean and Individual Predicted Values

The following graph shows approximate annual sales of new in-ground swimming pools in the U.S.[2]

[2] 2001 figure is an estimate. Source: PK Data/*New York Times,* July 5, 2001, p. C1.



2001 figure is an estimate. Source: PK Data/*New York Times*, July 5, 2001, p. C1.

Here is the underlying data:

Year Since 1995	1	2	3	4	5	6
Number of Pools	125	135	140	150	155	160

Use regression to compute confidence intervals for both the predicted value of y and the mean value of y in 2003. Are both results meaningful in the contect of this problem?

Step 1 Do the regression:
Taking x = year since 1995 and y = number of pools, we obtain the following output from Excel:

Regression Statistics	
Multiple R	0.99231497
R Square	0.984689
Adjusted R Square	0.98086124
Standard Error	1.82574186
Observations	6

ANOVA

	df	*SS*	*MS*	*F*
Regression	1	857.5	857.5	257.25
Residual	4	13.3333333	3.33333333	
Total	5	870.833333		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	119.666667	1.69967317	70.4056925	2.4386E-07
X Variable 1	7	0.43643578	16.0390149	8.8363E-05

Step 2: Do the calculations for the CI and PI:

Regression equation: $\hat{y} =$ ☐

$x_p =$ _____ $\hat{y}_p =$ _____

$n =$ _____ $\bar{x} =$ _____

$$SS_{xx} = \left[\frac{s}{s_{b_1}}\right]^2 = \left[\frac{\square}{\square}\right]^2 = \_\_\_\_\_$$

$t_{\alpha/2} =$ _____ $S_{YX} =$ _____

$$\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}} = \frac{1}{\square} + \frac{(\_\_\_\_\_ - \_\_\_\_\_)^2}{\square} = \_\_\_\_\_$$

$$\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}} = \_\_\_\_\_$$

$$\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}} = \_\_\_\_\_$$

CI: $\hat{y}_p \pm t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$

$= \square \pm \square$

$= \square \pm \square$

$= [\square , \square]$

PI: $\hat{y}_p \pm t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$

$= \square \pm \square$

$= \square \pm \square$

$= [\square , \square]$

Setting this up on the Excel Sheet Here is the PI calculuation:

Xp:	8
Yp:	175.666667
X-bar:	3.5
SSxx:	17.5
t:	2.77645086
Delta PI:	7.72733298
PI Lower:	167.939334
PI Upper:	183.394

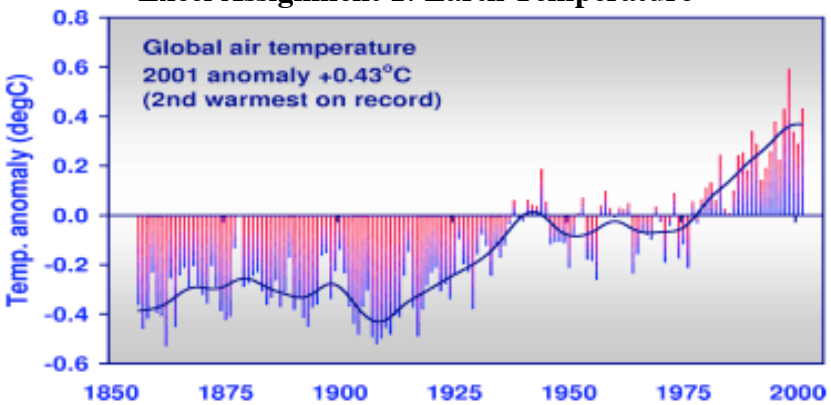
If you use formulas for everything, then you can change the value of x_p and automatically see the effect on the confidence interval.

Interpretation:

Comments on CI for $\bar{y}$:

Exercises for this topic:
p. 581 #32

Excel Assignment 1: Earth Temperature



Source: Climatic Research Unit, University of East Anglia
http://www.cru.uea.ac.uk/

Go to

http://www.zweigmedia.com/qm203/

and download the Excel spreadsheet called **Surface Temperatures**.

A. Obtain residual plots and use then to judge the extent to which the regression assumptions are met. Comment on your conclusions.

B. Test for evidence of first order autocorrelation.

C. Perform a linear regression on the data and give the regression model. **Important:** First rescale the year data so that $x = 0$ corresponds to 1950. (The regression computations are more accurate for small values of x.)

D. Interpret the slope of the regression equation.

E. Perform a hypothesis test at the 95% significance level to test whether temperature is increasing with time.

F. Obtain a 95% PI for $x = 70$. Interpret the result.

G. Repeat Steps A–F using the 1980–2006 data only . (Take $x = 0$ to represent 1980 here.)

H. Compare slopes of the regression equations for 1880-2006 and 1980–2006. What does this comparison suggest about global warming?

Topic 5
Assumptions for Simple Linear Regression
(§14.8 in the text)
We list the assumptions we are making when we perform regression analysis:

Assumptions for Simple Linear Regression
(§11.4–11.6 in text)

1. Normalcy:
We assume a relationship of the form $y = \beta_0 + \beta_1 x + \varepsilon$, where ε is a **normal** random variable with mean 0 and variance σ^2. In practice, we require that the residuals are more-or-less normally distributed.

2. Homoscendacity:
We assume that ε has the same standard deviation σ at every value of x.

3. Independence of errors
The values of each measurement of y (and hence ε) are independent of each other: getting a certain value for one measurement does not effect the probability of the others. (Think, for example, of the DOW.)

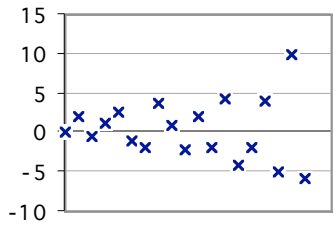
Question Why are these assumptions necessary?
Answer: Although we can always construct a regresion line without these assumptions (just obtain the line that minimizes SSE), we cannot say that b_0 and b_1 are unbiased estimators of β_0 and β_1 without them. Nor can we make statitical inferences (see later) about β_0 and β_1 without these assumptions.

Illustration of Violations of the Assumptions

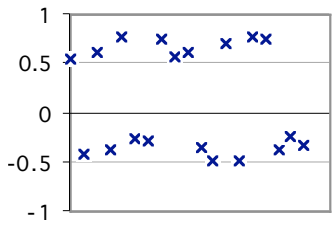
We look at scatter plots of the residuals $y - \hat{y}$ versus x.
Note We can have Excel plot them for us as an option when we do regression.

Worksheet 1
Identify what, if any, violations are present in the given residual plots:

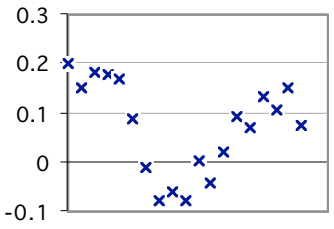
A

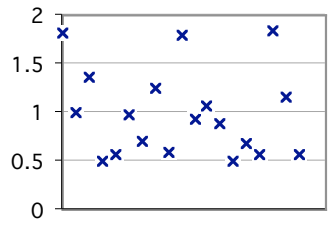
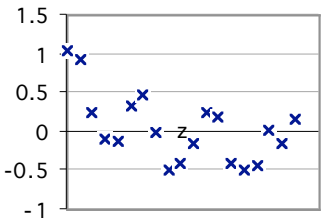
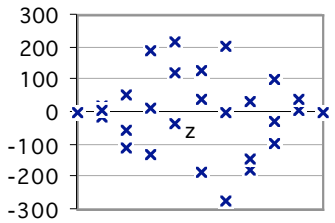


B



C



D	E	F
		

There is a very precise way of detercting a certin kind of violation of Assumption 3 (Independence) called **first order autocorrelation.** This only makes sense in **time-series** data (that is, data where the *x*-axis measures time) such as the DOW or the price of gold, etc. If the scores in a time-seris plot are autocorrelated, it means that each residual depends positively (positive autocorreplation) or negatiely (negative autcorrelation) on the preceding score. We measure this phenomenon using a statstic called the **Durbin-Watson statistic:**

Durbin-Watson Statistic

$$d = \frac{\Sigma(e_i - e_{i-1})^2}{\Sigma e_i^2} = \frac{\Sigma(e_i - e_{i-1})^2}{SSE}$$

where e_i is the residual at time i.
Only makes sense for time-series data.

Properties:
(1) $0 \le d \le 4$[†]
(2) $d \approx 2$ is reiduals are uncorrelated.
(3) For positive correlation, $d < 2$, and approaches 0 for strong positive correlation.
(4) For negative correlation, $d > 2$ and approaches 4 for strong negative correlation.

Minitab: Check the Durbin-Watson box under regression options.

We do a hypothesis test:

H_0: There is no positive autocorrelation

[†] Why? Look at the numerator: $(e_i - e_{i-1})^2 = e_i^2 - 2e_ie_{i-1} + e_{i-1}^2 \le e_i^2 + 2|e_i||e_{i-1}| + e_{i-1}^2$. However, in general $2ab \le a^2 + b^2$ (which comes from the inequality $(a-b)^2 \ge 0$). This gives

$$(e_i - e_{i-1})^2 \le e_i^2 + e_i^2 + e_{i-1}^2 + e_{i-1}^2$$

Therefore, summing from 2 to n:

$$\Sigma(e_i - e_{i-1})^2 \le \Sigma e_i^2 + \Sigma e_i^2 + \Sigma e_{i-1}^2 + \Sigma e_{i-1}^2 \le 4\Sigma_{1 \le i \le n} e_i^2 = 4\text{SSE}$$

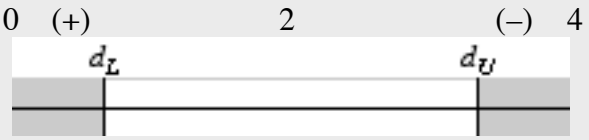
(Note that the sums on the left-hand side go from 2 to n, whereas the sum on the right-hand side groes from 1 to n and is therefore larger in general). This gives the result

H_a: There is positive autocorrelation (That is, $e_i = me_{i-1} + k$ for some $m > 0$ and k)

Reject H_0 if d is in the left-hand rejection region shown shaded (yes, positive autocorrelation)

Accept H_0 if d is in the right-hand portion of the rejection region shown (not only is there no positive autocorrelation evidenced, but there is evidence of *negative* autocorrelation, meaning we iaccept *the null-hypothesis*)

If d is in the "in-between" region, the test is inconclusive (fail to reject or accept)



Testing for Negative autocorrelation:

Replace d_L by $4-d_U$ and d_U by $4 - d_L$ and proceed as above, but do the opposite: the right-hand region gives negative autocorrelation, and the left-hand region gives none.

We can look up the lower and upper limits d_U and d_L in table E.10 in the textbook, using $k = 1$ (k is the number of independent variables we are using in regression).

Question Where do the numbers in the table come from?

Answer What Durbin & Watson did was to compute the sampling distribution of D. The rejection regions correspond to the tail-areas under the curve with an area of 0.05. It is not built into Excel, so we need the table at hand. Here is a partial table for $\alpha = .05$ (the whole one is in the back of the book):

Critical Values for Durbin-Watson (α = 0.05)

	$k = 1$		$k = 2$		$k = 3$		$k = 4$		$k = 5$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1.08	1.36	0.95	1.54	0.81	1.75	0.69	1.98	0.56	2.22
16	1.11	1.37	0.98	1.54	0.86	1.73	0.73	1.94	0.62	2.16
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.66	2.10
18	1.16	1.39	1.05	1.54	0.93	1.70	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.54	0.97	1.69	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.89	1.83	0.79	1.99

Worksheet 2 Computing the Durbin-Watson Statsitic

Here is some more data on the households with cable (y = millions of China households with cable, and x is the year since 2000. You can download this data at

http://www.zweigmedia.com/qm203/

under "China Cable Data".

Test for first order autocorrelation:

Year x	**Observed** y	**Residual** e	**(Delta-R)^2** $(e_i - e_{i-1})^2$
−4	50		
−3	55		
−2	57		
−1	60		
0	68		
1	72		
2	80		
3	83		
4	85		
5	87		
6	95		
7	101		
8	103		
9	111		
10	114		
11	118		
		Sum:	

Step 1: Compute the regression line (Use an excel data analysis for a quick answer)

Step 2: Have Excel show the residualsfor you..

Step 3: Compute $D = \frac{\text{Sum}}{\text{SSE}} = \frac{\square}{\square} = \square$

Step 4: Perform the hypothesis tests:

H_0: _____

H_a: _____

d_L = _____ d_U = _____

Conclusion: _____

H_0: _____

H_a: _____

d_L = _____ d_U = _____
Conclusion: _____

Execicses for this topic
Go to
http://www.zweigmedia.com/qm203/
and download the Surface Temperatures spreadsheet.
(1) Use the data for 1980–2006 only: Plot the residuals versus *x* and determine if there are any violations of model assumtions.
(2) Compute the Durin-Watson Statistic to determine whether there is evidende of autocorrelation.

Topic 6
Multiple Regression: The Model & Inferences about the β Parameters
(Based on §15.1–15.3 and part of 15.5 in book)

A **linear function of k variables** is a function of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k.$$

a **probabilistic linear function of k variables** has the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon,$$

where ε is normally distributed and independent of the values of the x_i.

Worksheet 1 - Interpreting the coefficients:
Here is an example of a **linear function of two variables**: Chrysler's percentage share of the US mini-van market in the period 1993–1994 could be approximated by the linear function

$$c(x_1, x_2, x_3) = 72.3 - 0.8x_1 - 0.2x_2 - 0.7x_3,$$

where x_1 is the percentage of the market held by foreign manufacturers, x_2 is *General Motors*' percentage share, and x_3 is *Ford*'s percentage share.[3]

Interpretation of β_1:

For every 1-unit increase in x_1, y _____ by _____ units.
In other words, Chrysler's percentage share of the US mini-van market in the period 1993–1994 _____

_____.

Interpretation of β_2:

For every 1-unit increase in x_2, y _____ by _____ units.
In other words, Chrysler's percentage share of the US mini-van market in the period 1993–1994 _____

_____.

Interpretation of β_3:

For every 1-unit increase in x_3, y _____ by _____ units.
In other words, Chrysler's percentage share of the US mini-van market in the period 1993–1994 _____

_____.

[3] The model is your instructor's. Source for raw data: Ford Motor Company/*The New York Times*, November 9, 1994, p. D5.

Obtaining the best-fit coefficients for this kind of model is called **multiple linear regression.** The procedure to compute the regression coefficients b_0, b_1, ... by hand involves the use of matrix algebra, and is beyond the scope of this course. Therefore, we will use Excel output exclusively (sometimes aided by PHStat).

Worksheet 2 — Basics of Multiple Regression

Go to

http://www.zweigmedia.com/qm203/

and download the Excel spreadsheet for the Multiple Regression Example on Radio and TV (from Exercise 12.25 in the textbook, but with different data). Since the textbook does not bother to say exactly what the variables are, we are forced to invent them: Take

y = Number of Psychic Crystal pendants sold per day

x_1 = Minutes per day in late-nite TV and radio advertising

x_2 = Number of half-page ads per day in the local newspapers

(a) Write down the probabilistic model, the regression equation, and interpret the slopes.

(b) Find a 95% CI for the population slope β_1.

(c) Determine at the 95% level of significance whether sales go up as newspaper advertising goes up.

(d) Interpret the p-values for the coefficients β_1 and β_2. On this basis, which explanatory variables should be retained?

y Sales	x_1 Radio/TV	x_2 Newspaper	y Sales	x_1 Radio/TV	x_2 Newspaper
563	0	40	1295	45	45
308	0	40	1045	50	0
384	25	25	1076	50	0
436	25	25	1257	55	25
362	30	30	1359	55	25
677	30	30	1199	60	30
764	35	35	1520	60	30
996	35	35	1648	65	35
852	40	25	991	65	35
680	40	25	1439	70	40
656	45	45	1214	70	40

Here is the resulting regression output:

Regression Statistics	
Multiple R	0.8311521
R Square	0.6908139
Adjusted R Square	0.6582679
Standard Error	233.86664
Observations	22

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	2321834.06	1160917	21.22583	1.44E-05
Residual	19	1039178.53	54693.61		
Total	21	3361012.59			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	155.422	186.555473	0.833114	0.415131	-235.043
RadioTV	16.866047	2.58935264	6.513615	3.07E-06	11.44647
Newspaper	1.9378349	4.36134205	0.444321	0.66183	-7.19056

(a) Probabilistic model:

$y =$ _____

Regression equation:

$\hat{y} = b_0 + b_1x_1 + b_2x_2$

$\hat{y} =$ _____

Interpretation of b_1:

Interpretation of b_2:

(b) Preparation for doing hypothesis CI for β_1 at the 95% significance level:

$n =$ _____ $k =$ _____

$df = n-k-1 =$ _____ $t_{\alpha/2} =$ _____

$S_{b_1} =$ _____

$CI = b_1 \pm t_{\alpha/2}\, S_{b_1}$

$= \square \pm \square$

$= \square \pm \square$

$= [\square , \square]$

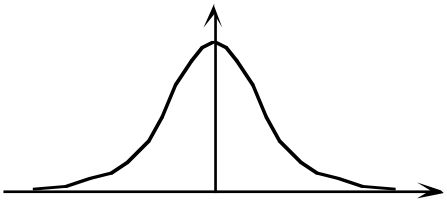
(c) H_0: _____

H_a: _____

$df = n - k - 1 =$ _____

Rejection region:

critical $t =$ _____



$t-$statistic: _____ In rejection region? _____

Conclusion: _____

Interpret ting result of hypothesis test:

(d) p-value for $\beta_1 =$ _____

Interpretation:

p-value for $\beta_2 =$ _____

Interpretation:

Conclusion:

Variability of the Random Term

The standard deviation σ of ε has an unbiased estimator for σ^2 given as follows.

First, the **number of degrees of freedom** is given by

$$d_v = n - \#\beta\text{-terms} = n - (k+1) = n-k-1$$

Then the unbiased estimator is

$$s^2 = \frac{\text{SSE}}{n-k-1} = \text{MSE} \qquad \text{(Mean Square Error)}$$

where

$$\text{SSE} = \text{sum of squares of errors} = \Sigma(y_i - \hat{y}_i)^2$$

s is called the **estimated standard error of the regression model** and appears under "Regression Statistics" whereas its square, MSE appears in the ANOVA section.

Interpretation of *s*

Just as with simple regression, we can make the following inference: Around 95% of the observations will lie within 2*s* of the predicted value $\hat{y}$.

Coefficients of Multiple Determination

As we saw for simple regression, the proportion of the total sample variation that can be attributed to the independent variables is given by

$$r^2 = \frac{SS_{yy} - \text{SSE}}{SS_{yy}} = \frac{\text{SSR}}{\text{SST}}$$

The textbook calls this quantity. As for simple regression, r^2 gives the proportion of the sample variation in *y* attributable to the values of the independent variables in a linear relationship. A disadvantage of r^2 is that it cannot be used to compare models with different numbers of explanatory variables. The larger the number of variables, the smaller SSE tends to become. In fact, it is possible to construct models with $n - 1$ variables that result in an exact fit of the regression models, and hence SSE = 0, so that $r^2 = 1$. The **adjusted $\mathbf{r^2}$** is defined by the following formula [which scales the quantity $1-r^2$ by the ratio $(n-1)/)n-k-1)$]:

$$r_{adj}^2 = 1 - \left[(1-r^2)\frac{n-1}{n-k-1}\right]$$

Question: How do we interpret the adjusted r^2?

Answer: Suppose, for instance, that $r_{adj}^2 = .89$. We can say that "if we take model size into account, 89% of the variation in *y* is explained by the values of the independent variables."

Answer: Suppose, for instance, that $r_{adj}^2 = .89$. We can say that "if we take model size into account, 89% of the variation in y is explained by the values of the independent variables."

Exercises for this topic:
p. 632 #5, 15
For both of these, also perform a *t*-test for individual significance for the coefficients of x_1 and x_2.

Topic 7
Using *F*-statistics: Evaluating the Whole model and Portions of the Model
(§15.5 and16.2 in the book)

In Topic 3 we saw that a F statistic could also be used to evaluate a simple linear regression. If we look at its definition,

$$F = \frac{(SS_{yy} - \text{SSE})/k}{\text{SSE}/[n-(k+1)]} = \frac{SSR/k}{SSE/(n-k-1)}$$

we see that it related to the fit of the *entire* model. To understand its meaning, recall that

$$SS_{yy} = \Sigma\,(y - \bar{y})^2 \qquad \text{and} \qquad \text{SSE} = \text{Sum of } \Sigma(y - \hat{y})^2$$

Therefore, if $\beta_1 = \beta_2 = \ldots = \beta_k = 0$, then

$$SS_{yy} \approx SSE$$

(Since the regression equation should be close to a constant, so that $\bar{y} \approx \hat{y} \approx$ that constant).
Therefore

$$SSR = SS_{yy} - SSE \approx 0,$$

and so F is close to zero. In fact, its sampling distribution (assuming $\beta_1 = \ldots = \beta_k = 0$) is the known distribution tabulated at the end of this booklet .

The *F*-Statistic for Multiple Regression: Testing the Usefulness of the Overall Model
The F-statistic is used to test the following hypothesis:

H_0: $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$

H_a: At least one of these coefficients is non-zero.

Test statistic:

$$F = \frac{(SS_{yy} - \text{SSE})/k}{\text{SSE}/(n-k-1)} = \frac{SSR/k}{SSE/(n-k-1)} = \frac{R^2/k}{(1-R^2)/(n-k-1)}$$

$$= \frac{\text{Sum of Squares(Regression)/df(Regression)}}{\text{Sum of Sqaures(Error)/df(Error)}} = \frac{\text{Mean Square(Model)}}{\text{Mean Sqare(Regression)}}$$

Using *F*:
Compare the F-statistic to the one in the table with k df in the numerator & $[n-(k+1)]$ df in the denominator. If $F > F_\alpha$, then we reject H_0.

Note: Rejecting H_0 does not mean that the model is the *best* one; another model might give an even better confidence level.

Some terminology for the terms in the Excel table:

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	df (Regression)	SSR (Regression)	MSR	MSR/MSE	
Residual	df (Error)	SSE (Error)	MSE		
Total	dfM + dfE	SST = SSM + SSE			

Workbook 1 _ Using the F-Test to Evaluate the Entire Model

Fill in the missing values of the following Excel sheet, and compute the overall usefulness of the given regression model at the 95% confidence level.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.78881784
R Square	0.62223358
Adjusted R Square	0.54128363
Standard Error	8.55161132
Observations	18

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	1686.37453			
Residual	14	1023.82079			
Total	17	2710.19531			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	-72.848775	43.6506747	-1.6689037	0.1173366	-166.47024
X1	-0.0662404	0.18831924	-0.351745	0.73026749	-0.4701453
X2	85.7340025	22.6308238	3.78837303	0.00199612	37.1956697
X3	-0.0222909	0.02994571	-0.7443784	0.46895881	-0.0865182

Missing values:

$$\text{MSR} = \frac{\text{SSR}}{\text{df(R)}} = \frac{\square}{\square} \approx \square$$

$$\text{MSE} = \frac{\text{SSE}}{\text{df(E)}} = \frac{\square}{\square} \approx \square$$

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{\square}{\square} \approx \square$$

Significance F (p-value) =FDIST(F-stat, df_1, df_2) = $\square$

Evaluating the Model

Linear Model: y = _____

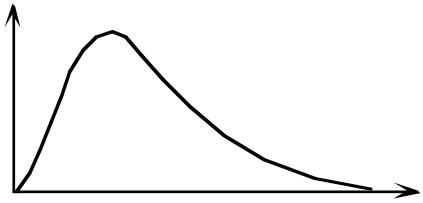
H_0: _____

H_a: _____

df (numerator) = k = _____ *df*(denominator) = $n - k - 1$ = _____

$F_{critical}$ = _____

Rejection region:



F−statistic: F = _____

In rejection region? _____

Conclusion: _____

Interpret ting result of hypothesis test:

Interpreting the F significance value:

Testing a Portion of a Model

Suppose we want to test a bunch of terms at once (this was more reliable than testing them one-at-a-time, due to type 1 error accumulation). To do this, we compute an F-statistic showing the percentage of new errors caused by reducing the model as follows.

Testing a Portion of the Model

Do *two* regression analyses:

Reduced Model: $\bar{y} = \beta_0 + \beta_1 x_1 + \ldots + \beta_g x_g$

Complete (larger) Model: $\bar{y} = \beta_0 + \beta_1 x_1 + \ldots + \beta_g x_g + \ldots + \beta_k x_k$

H_0: $\beta_{g+1} = \ldots = \beta_k = 0$

H_a: at least one of them is not zero

$$F = \frac{(SSE_{partial} - SSE_{complete})/[k-g]}{SSE_{complete}/[n-(k+1)]} .$$

Rejection Region: $F > F_\alpha$, based on $(k-g)$ numerator and $n-(k+1)$ denominator df.

n = # data points

$k-g$ = number of β's tested

Interpretation: The null hypothesis asserts that the additional explanatory variables $x_{g+1}, x_{g+2}, \ldots, x_k$ do not contribute significantly to the usefulness of the model. In other words, the reduced model is preferable. Rejecting it implies that the additional variables do contribute significantly to the usefulness of the model.

Note: If the two models differ by a single term, then the *F*-test can be replaced by a *t*-test: Just do the regression analysis for the larger model, and test for the extra coefficient.

Worksheet 2 — Testing a portion of a model

Here is a model for the selling price of a home (y) as a function of the list price (x_1), the number of bedrooms (x_2), and the time on the market in weeks (x_3). Use a comparison of models to determine whether x_2 and x_3 contribute significantly more than x_1 alone.

Reduced Model

Regression Statistics	
Multiple R	0.99474985
R Square	0.98952725
Adjusted R Square	0.98917816
Standard Error	36747.5119
Observations	32

ANOVA

	df	*SS*
Regression	1	3.8278E+12
Residual	30	4.0511E+10
Total	31	3.8683E+12

	Coefficients	*Standard Error*

Complete Model

Regression Statistics	
Multiple R	0.99515812
R Square	0.99033968
Adjusted R Square	0.98930464
Standard Error	36532.137
Observations	32

ANOVA

	df	*SS*
Regression	3	3.8309E+12
Residual	28	3.7369E+10
Total	31	3.8683E+12

	Coefficients	*Standard Error*

Intercept	-26306.993	10738.6428
X1	0.99674558	0.01872148

Intercept	-7477.8395	20471.9887
X1	0.99787257	0.0189959
X2	-2663.1189	5370.9791
X3	-534.62597	421.174382

Complete Model: $y =$ _____

Reduced Model: $y =$ _____

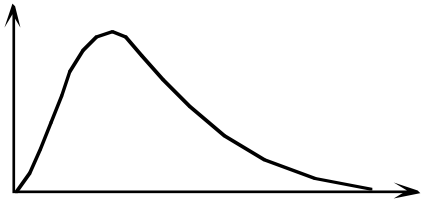
H_0: _____

H_a: _____

df (numerator) $= k - g =$ _____ df(denominator) $= n - k - 1 =$ _____

$F_{critical} =$ _____

Rejection region:



$F-$statistic: $F = \dfrac{(SSE_{partial} - SSE_{complete})/[k-g]}{SSE_{complete}/[n-(k+1)]}$

$= \dfrac{[\square - \square]/\square}{\square/\square}$

$\approx \dfrac{\square}{\square} \approx \square$

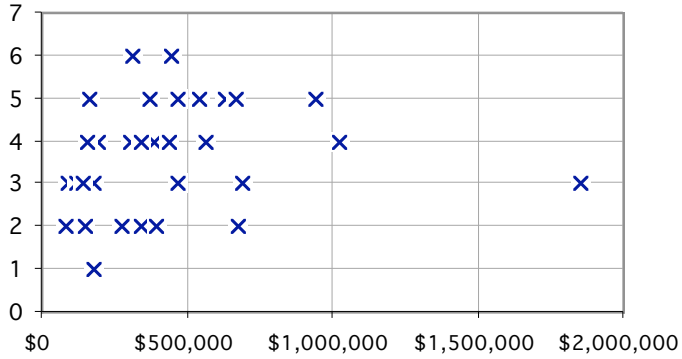
In rejection region? _____

Conclusion: _____

Interpret ting result of hypothesis test:

What does the result suggest about housing prices?

FYI: Here is a Excel plot of price vs. number of bedrooms for the data used above.



Exercises for This Topic:
Go the download place at

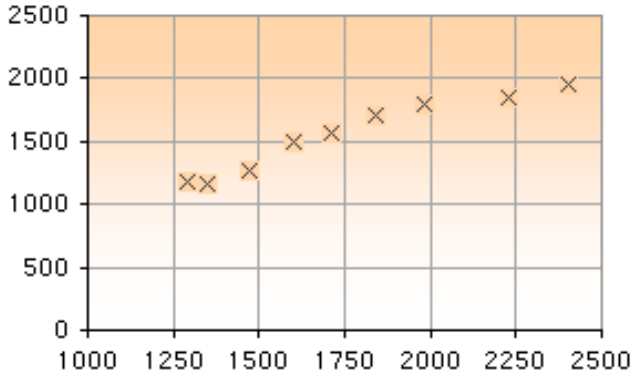
http://www.zweigmedia.com/qm203/

and download the data under "Homework Assignment on Ford Stock Pricel". Determine whether the Yen and Mark rates contribute significantly more information to the Ford stock price than the S&P index alone.

Topic 8
Quadratic and Interactive Terms

Quadratic Terms
(Based on 16.1 in the text)
Here are some data and a graph showing monthly electricity use vs. size of home in square ft.

Size (sq. ft) *x*	1290	1350	1470	1600	1710	1840	1980	2230	2400	2930
Usage (kw-hrs) *y*	1180	1170	1260	1490	1570	1710	1800	1840	1960	1950



The Excel scatter chart suggests a quadratic relation. $y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$, so we take x_2 to be x_1^2, by adding an extra column on the spreadsheet, and then do a multiple linear regression. The curvature is accounted for by the x_1^2 term, so:

$$\text{There is no curvature} \Leftrightarrow \text{Coefficient of } x^2 = 0 \Leftrightarrow \beta_2 = 0$$

Therefore, to test whether there is evidence of curvature, all we need to do is a *t*-test to look at the coefficient of x^2.

Worksheet 1 — Testing for Curvature
Use the above data in Excel, complete the given table, find the regression quadratic model, use it top predict electric usage for a 2500 square ft home, and test for curvature at the 95% significance level.

First, download the data from

http://www.zweigmedia.com/qm203/

by using the **Curvature Data** link. Then create an extra column in the Excel workbook using the squares of the values of x (Notice that we put the y-column first):

	A	B	C
1	y	x	x^2
2	1180	1290	=B2^2
3	1170	1350	
4	1260	1470	
5	1490	1600	
6	1570	1710	
7	1710	1840	
8	1800	1980	
9	1840	2230	
10	1960	2400	
11	1950	2930	

Now do a regression using both x and x^2 as the explanatory variables. Now do the regression and check that the given values match the part of the sheet shown below, and fill in the remaining ones.

Regression Statistics	
Multiple R	0.99091051
R Square	0.98190365
Adjusted R Square	0.97673326
Standard Error	46.9099376
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	835806.204	417903.102	189.909147	7.972E-07
Residual	7	15403.7957	2200.54225		
Total	9	851210			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	-1234.7	243.36981	-5.0733492	0.00144169	-1810.1778
x	2.41543711	0.24640607	9.80266875	2.4394E-05	1.83277975
x^2	-0.0004538	5.9214E-05	-7.6636835	0.00011978	-0.0005938

The regression equation is

$\hat{y} =$ _____

Predicted value of y for a 2500 sq ft home =

_____

Now test for curvature:

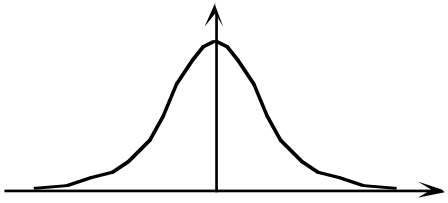
H_0: _____

H_a: _____

$df = n - k - 1$ = _____

Rejection region:

critical t = _____



t−statistic: _____ In rejection region? _____

Conclusion: _____

Interpret ting result of hypothesis test:

p-value for β_2 = _____

Interpretation:

Does the cost accelerate or decelerate as the size of the home increases? Explain

Interactive Models

(Based on nothing in this book, but a nice section in "Statistics for Business and Economics;" by McClave, Benson, Sincich; 8th Ed., Prentice Hall.)

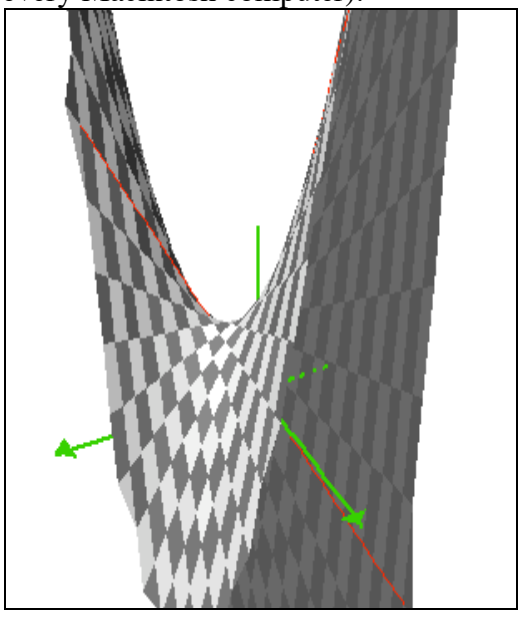
Here is some data showing the auction prices of 32 grandfather clocks together with the number of bidders and the age of the clock.

Age	# Bidders	Auction Price	Age	# Bidders	Auction Price
X1	X2	Y	X1	X2	Y
127	13	1200	170	14	2100
115	12	1100	182	8	1600
127	7	850	162	11	1900
150	9	1500	184	10	2000
156	6	1000	143	6	800
182	11	2000	159	9	1500
156	12	1800	108	14	1100
132	10	1300	175	8	1500
137	9	1300	108	6	700
113	9	1000	179	9	1800
137	15	1700	111	15	1200
117	11	1000	187	8	1600
137	8	1100	111	7	800
153	6	1100	115	7	700
117	13	1200	194	5	1400
126	10	1300	168	7	1300

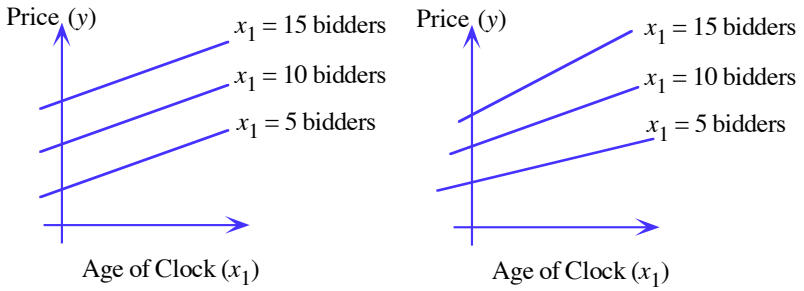
If we suspect that the age of the clock (x_1) and the number of bidders (x_2) will *interact* (that is, different numbers of bidders may cause the price to vary differently as a function of the age), we try a model of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon.$$

The effect on the graph is to twist the surface (see the following picture generated by the grapher that comes with every Macintosh computer):



If $\beta_3 = 0$, then each value of x_2 will result in a line of the same slope for y as a function of x_1. Otherwise, if $\beta_3 \neq 0$, the slope will vary.



No Interaction ($\beta_3 = 0$) **Interaction ($\beta_3 \neq 0$)**

We suspect that the number of bidders will have a positive impact on the variation of price with age, so we test the alternate hypothesis H_a $\beta_3 > 0$. To do this on Excel, we introduce a third column for x_1x_2, and do a regression at the 99% significance level:

Regression Statistics	
Multiple R	0.97309468
R Square	0.94691325
Adjusted R Square	0.94122538
Standard Error	95.4955358
Observations	32

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	4554578.75	1518192.92	166.479523	5.9656E-18
Residual	28	255343.126	9119.39735		
Total	31	4809921.88			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	61.6599935	316.98622	0.19451948	0.84717364	-587.65757
X1	2.62232358	2.18256632	1.20148632	0.23962706	-1.8484659
X2	-65.831044	32.1040495	-2.0505527	0.04977643	-131.59328
X1X2	1.11008423	0.22804843	4.86775643	3.9815E-05	0.64294766

The t-statistic for β_3 is 4.86. We look up $t_\alpha = t_{0.001}$ for $n-(k+1) = 28$ degrees of freedom as usual, and get 3.408. Since $t > t_\alpha$, we reject H_0 and conclude that there is a strong interaction here. Also note the large value of R^2 (95%) and the tiny value for Significance F, showing that the model is a good one.

Exercises for this Section

Quadratic: p. 709 #8 (test for curvature the way we do, and also read up about using the log transformation and try it).

Interaction: The following problem comes from "Statistics for Business and Economics;" by McClave, Benson, Sincich; 8th Ed., Prentice Hall.

Download the CEO Data sheet at

http://www.zweigmedia.com/qm203/

(a) What does it mean for the salary and percentage stock price to interact?

(b) Find the interaction regression equation, and test the overall model at the 95% level of significance. (You will fail to reject H_0.)

(c) Is there evidence at the 95% level of significance that CEO income and stock percentage interact?

(c) Looking at the *p*-values, try to eliminate one of the variables in order to obtain a model with a satisfactory *F*-value. What is the resulting regression model?

(d) Using the better regression model, predict the change in profit for every $1000 increase in a CEO's income when the CEO owns 2% of the company stock.

(e) If you were on a board of directors and wished to use the regression analysis above, would you be in favor giving a new CEO more stock and a lower salary? Explain.

Topic 9
Qualitative Variables
(Based on §15.7 in the book, but we go further)

If, for example, we are interested in the market activity in real estate, the time a home has to wait on the market may depend on whether it is a house or a condominium, and also on the asking price. Here, the asking price is a **quantitative variable** since it is a number, while the kind of home (house vs. condo) is a **qualitative variable**. In this topic, we see how to include qualitative variables (or "dummy variables" as they are called in the textbook.).

Suppose we are interested in the sale price of a home as a function of whether it is (A) a house, (B) a condominium, or (C) a co-op. We can do so by defining *two* new variables (not three)

$$x_1 = \begin{cases} 1 & \text{if the property is a condominium (Catgegory B);} \\ 0 & \text{if not} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{if the property is a co-op (Category C);} \\ 0 & \text{if not} \end{cases}$$

Then, in the model, we can plug in $x_1 = x_2 = 0$ for a house, $x_1 = 1$ & $x_2 = 0$ for a condo, etc. We say that x_1 and x_2 constitute a **qualitative variable with 3 levels.**

Q How do we interpret the coefficients β_i for the model?
A If we use the example of housing sales, then, writing

$$\bar{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

we find

μ_A = mean sale price of a house (put $x_1 = x_2 = 0$) = $f(0,0) = \beta_0$
μ_B = mean sale price of a condominium (put $x_1 = 1$ & $x_2 = 0$) = $f(1,0) = \beta_0 + \beta_1$
μ_C = mean sale price of a co-op (put $x_1 = 0$ & $x_2 = 1$) = $f(0,1) = \beta_0 + \beta_2$

Thus,

$\beta_0 = \mu_A$
$\beta_1 = \mu_B - \mu_A$
$\beta_2 = \mu_C - \mu_A$.

In other words, the coefficients measure the difference between the three categories. For instance, β_1 measures the effect on the price of a home of switching from a house (the **base level)** to a condominium.

Qualitative Variable With *k* Levels: Comparing *k* Means

Model: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{k-1} x_{k-1}$

$$x_i = \begin{cases} 1 & \text{if } y \text{ is observed at level } i; \\ 0 & \text{if not} \end{cases}$$

$\mu_A = \beta_0$	$\beta_0 = \mu_A$
$\mu_B = \beta_0 + \beta_1$	$\beta_1 = \mu_B - \mu_A$
$\mu_C = \beta_0 + \beta_2$	$\beta_2 = \mu_C - \mu_A$
....	...

Worksheet 1 — Comparing 3 Means

Go to

http://www.zweigmedia.com/qm203/

and download the Housing Prices Excel file. There you will find selling prices of various homes in 1995 (y) sold in (A) Manhattan, (B) Connecticut, and (C) Long Island.

(a) Obtain the regression model and interpret the coefficients in the model.

(b) Test at the 95% level of significance whether Connecticut homes are cheaper tan Manhattan homes, and whether Long Island homes are cheaper than Manhattan homes.

(c) Test the overall model, and interpret the result.

(d) According to the regression model, how much more expensive is a home in Connecticut than in Long Island?

(a) The explanatory variables are:

$$x_1 = \begin{cases} 1 & \text{if } \_\_\_\_\_ \\ 0 & \text{if } \_\_\_\_\_ \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{if } \_\_\_\_\_ \\ 0 & \text{if } \_\_\_\_\_ \end{cases}$$

The model is $\bar{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, where

β_0 = _____

β_1 = _____

β_2 = _____

Interpretation of coefficients:

β_0:

β_1:

β_2:

(b) Comparing Connecticut and Manhattan:

H_0: _____

H_a: _____

$p-$value for 2-tail test: _____

p-value for 1-tail test: $=\frac{1}{2}$ p-value for 1-tail test = _____

Conclusion & Interpretation:

Comparing Long Island and Manhattan:

H_0: _____

H_a: _____

$p-$value for 2-tail test: _____

p-value for 1-tail test: $=\frac{1}{2}$ p-value for 1-tail test = _____

Conclusion & Interpretation:

(c) Testing overall model:

H_0: _____

H_a: _____

$F-$significance: _____

Conclusion & Interpretation:

(d) $\mu_C - \mu_B = (\mu_C - \mu_A) - (\mu_B - \mu_A)$

$=$ _____ $-$ _____ $=$ _____

Interpretation:

NOTE H_0 only tests whether $\mu_B = \mu_A$ and $\mu_C = \mu_A$. It does *not* test whether $\mu_C = \mu_B$. However—and this is why we use the term "level"—. If we were comparing the effectiveness of 5 different brands of detergents on a ketchup stain, we do not have

levels, and would like to know whether there is any difference at all among the 5 brands. Since there is no sense of "levels" here, we will use ANOVA (analysis of variance) later in this course address this.

Using Qualitative Variables to Compare Two Slopes

The effectiveness of an advertising medium can be measured by the number of items sold per $1,000 spent on advertising. That is, the *slope* of the Sales vs. Expenditures line. Suppose we want to compare (A) newspaper, (B) television, and (C) radio advertising. Let y be the monthly sales , and let x be the advertising expenditure (all three categories). Note that x is a *quantitative variable*. Let

$$x_1 = \begin{cases} 1 & \text{if we were advertising on the radio;} \\ 0 & \text{if not} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{if we were advertising on television;} \\ 0 & \text{if not} \end{cases}$$

Then consider first the *linear* model

$$\bar{y} = \beta_0 + \beta_1 x + \beta_2 x_1 + \beta_3 x_2.$$

Claim: This model gives the same sales vs. expenditure slope for all three categories. Indeed, the slope for newspaper sales is obtained by setting $\beta_2 = \beta_3 = 0$, and we get

$$\bar{y} = \beta_0 + \beta_1 x ,$$

yielding a slope of β_1. Similarly, if we look at radio advertising, we get

$$\bar{y} = \beta_0 + \beta_1 x + \beta_2, \qquad (1)$$

again a slope of β_1.

To obtain different slopes, we need the following *interactive* model:

$$\bar{y} = \beta_0 + \beta_1 x + \beta_2 x_1 + \beta_3 x_2 + \beta_4 x x_1 + \beta_5 x x_2. \qquad (2)$$

This gives a slope of β_1 for newspaper, $\beta_1+\beta_4$ for radio, and $\beta_1+\beta_5$ for television. Thus, if we test the hypothesis

$$H_0: \beta_4 = \beta_5 = 0$$

we are in fact testing whether or not the slopes for the three media are the same by testing the model for interaction. To do this, we use the reduced model test, comparing the interactive model and linear model..

Worksheet 2 — Comparing 2 Slopes

We want to compare the response to monetary bonuses for two types of worker: union and non-union. Here is the data, available under "Productivity" at

http://www.zweigmedia.com/qm203/

productivity	bonus	union?	productivity	bonus	union?
y	x	x1	y	x	x1
1435	0.2	1	1635	0.3	0
1512	0.2	1	1589	0.3	0
1491	0.2	1	1661	0.3	0
1575	0.2	0	1610	0.4	1
1512	0.2	0	1574	0.4	1
1488	0.2	0	1636	0.4	1
1583	0.3	1	1654	0.4	0
1529	0.3	1	1616	0.4	0
1610	0.3	1	1689	0.4	0

(a) Using a regression model, determine the coefficients that give the slope of (1) productivity vs. bonus for non-union workers and (2) productivity vs. bonus for union workers Do the data provide evidence that non-union workers are more responsive to bonuses than union workers?

(b) If we disregard bonuses, do the data provide evidence that non-union workers produce less than union workers?

Solution

(a) We use two levels: (A) non-union and (B) union. The model that shows

We compare the following models:

$\bar{y}$ = _____ Linear

$\bar{y}$ = _____ Interactive

Normally, we would do a regression for each of these models. But, since they differ only by a single term, we need only do the interaction model and look at the (single) interaction term using a *t*-test. Set up the interactive model to obtain the following output:

Interactive Model

Regression Statistics	
Multiple R	0.85002282
R Square	0.72253879
Adjusted R Square	0.66308282
Standard Error	40.433563
Observations	18

ANOVA

	df	*SS*	*MS*	*F*

Regression	3	59603.3889	19867.7963	12.1525012
Residual	14	22888.2222	1634.87302	
Total	17	82491.6111		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	1410.11111	51.3221403	27.475688	1.4008E-13
x	640	165.06933	3.87715877	0.00167544
x1	-47.777778	72.5804668	-0.6582732	0.52104129
x*x1	-3.3333333	233.443285	-0.014279	0.9888089

Model for non-union workers: x_1 = ____ $\bar{y}$ = _____

Slope for non-union workers: ☐

Model for union workers: x_1 = _____ $\bar{y}$ = _____

Slope for union workers: ☐

Hypothesis test to check whether union members are less responsive:

H_0: _____

H_a: _____

$p-$value for 2-tail test: _____

p-value for 1-tail test: _____

Conclusion & Interpretation:

(b) We are asked to compare two means for (A) non-union (B) union, regardless of bonuses, so we ignore x. The model we use is therefore a simple regression:

$$\bar{y} = \beta_0 + \beta_1 x_1$$

Hypothesis test to check whether union members produce less:

H_0: _____

H_a: _____

$p-$value for 2-tail test: _____

p-value for 1-tail test: _____

Conclusion & Interpretation:

Exercises for this topic:

Comparing 4 means: Download the Sales Data (Seasonal) worksheet at

http://www.zweigmedia.com/qm203/

and use it to compare average sales for (A) first quarter (B) second quarter, (C) third quarter and (D) fourth quarter.

(a) Construct the model and define each of the variables.
(b) Perform the regression and interpret each of the slopes.
(c) Is there a significant difference between sales in the different quarters?
(d) According to the model, what is the difference between third and fourth quarter sales?
(e) Are there any variables that do not contribute significantly to the model? If so, reduce the model appropriately and repeat parts (b) and (c).

Comparing two slopes: p. 622, #12.40 For part (a), they mean an ordinary linear model. Omit (i), (j), (k), (l) since they do not tell us anything interesting. Part (m) refers to the linear model.

Also, answer the following question: If we ignore shelf space (notice the nice distribution of shelf sizes anyway) do items placed in front sell better than object placed at the back?

Excel Assignment 2

Go the web site at

http://www.zweigmedia.com/qm203/

and download the Assignment 2 Excel worksheet.

Important: Use the 90% level of significance throughout.

Part 1:

Extract the Long Island home sales data only. Let

y = Sales price in \$1000

x_1 = Number of bedrooms

x_2 = Time on market in weeks

x_3 = Taxes & Maintenance

Perform three regressions as follows. In each case, write down the regression model with coefficients rounded to 4 significant digits, and use the model to predict the selling price of a 3-bedroom home whose with \$25,000 taxes after 10 weeks on the market.

(a) Multiple linear model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

(b) Interactive Model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3$

(c) Quadratic model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 {x_1}^2 + \beta_5 {x_2}^2 + \beta_6 {x_3}^2$

(d) Full second order model: $y =$ Quadratic model + interactive terms

(e) Compare the model in (d) with those in (b) and (c). Based on the outcomes, decide which of the three models (b), (c), (d) is best for predicting the cost of a home. [Hint: The comparison of (d) and (b) tells you whether the quadratic terms contribute significantly, and the comparison of (d) and (c) tells you whether any of the interactive terms contribute significantly.]

Part 2:
Using the same data sheet, compare housing prices in (A) Manhattan, (B) Westchester, (C) Connecticut, and (D) New Jersey:
(a) Is there any significant difference between housing prices in the four areas?
(b) What does your regression model predict for the difference between the cost of a home in New Jersey and Westchester?

Topic 10
Analysis of Variance (ANOVA): Single Factor Analysis
(Based on 13.2 in the book)

In the language of ANOVA, we are interested in the **response** (dependent variable, which we called y in regression) to one or more **factors** (independent variables which we called x_1, x_2, ... in regression). These factors may be qualitative or quantitative, and their values are called **levels**. This is where they differ from the qualitative variables as we used them in regression. For instance, a qualitative factor may have non-numerical levels, such as Soccer, Football, etc., while quantitative ones have numerical levels. Finally, the **treatments** in an experiment are the levels (in a single factor experiment) or pairs of levels (in a multiple factor experiment), and the **units** are the elements of the sample space n the experiment (e.g. students for SAT measurements).

Design of Experiments To design an experiment for factor analysis, one needs to first select a random sample of experimental units (e.g. soccer players) and then assign them (possibly randomly) to individual treatments for a given factor (e.g. have them practice in different brands of cleats and measure the resulting wear and tear). In an **observational** experiment, you would not decide who wears what cleats, but simply observe the wear and tear on the brands of cleats they already use. In a **completely randomized design**, one assigns experimental units (soccer players) to treatments completely randomly and independently.

The objective is usually to compare the sample means for the different levels: μ_1, μ_2, ..., μ_c and we will test the null hypothesis

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_c$$

against H_a: at least two of the treatment means are different.

Q How do we test this?
A For *two treatments*, we need to compare two means, μ_1 and μ_2.

Method 1: Comparison of Two Means Statistics
For this, we have the following, based on the sampling distribution of $\bar{x}_2 - \bar{x}_1$.

Comparing two Means
Large Samples:

$$\sigma_{(\bar{x}_2-\bar{x}_1)} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Confidence Interval for $(\mu_2-\mu_1)$

$$(\bar{x}_2-\bar{x}_1) \pm z_{\alpha/2}\ \sigma_{(\bar{x}_2-\bar{x}_1)}$$

Hypothesis Test

H_0: $(\mu_2-\mu_1) = D_0$ ($D_0 = 0$ for our purposes here)
H_a: $(\mu_2-\mu_1) \neq D_0$ or $(\mu_2-\mu_1) > D_0$ or $(\mu_2-\mu_1) < D_0$
where D_0 is some hypothesized difference between the two parameters.

Test statistic: $z = \dfrac{(\bar{x}_2-\bar{x}_1) - D_0}{\sigma_{(\bar{x}_2-\bar{x}_1)}}$

Assumptions
The two samples are randomly and independently selected from the two samples, and the sample sizes are sufficiently large so that the sampling distributions are approximately normal.

Small Samples
When one or both of the sample sizes is small, we cannot use the above approximation of $\sigma_{(\bar{x}_2-\bar{x}_1)}$, since it is *not* an unbiased estimator. An unbiased estimator is given by:

$$s_{(\bar{x}_2-\bar{x}_1)} = \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \quad . \qquad \text{("pooled sample variance")}$$

This estimate allows to proceed as usual for small samples, using the *t*-distribution instead of the normal one.

Assumptions
The distributions of x_1 and x_2 are normal with *the same population variance*. (The latter assumption is needed in order to guarantee that we can still use the *t*-distribution (otherwise we would need to use a new distribution).

Method 2: Regression

Comparing Two Means with Regression

Construct the following model, but call the treatments *A* and *B* (rather than 1 and 2)

Let

$$x_1 = \begin{cases} 1 & \text{if the measurement is made with treatement } B; \\ 0 & \text{if not} \end{cases}$$

and use

$$E(y) = \beta_0 + \beta_1 x_1.$$

Then,

$$\mu_A = \beta_0$$

$$\mu_B = \beta_0 + \beta_1$$

or $\beta_1 = \mu_B - \mu_A$,

Confidence Interval for $(\mu_B - \mu_A)$

This is the confidence interval for β_1

Hypothesis Testing

H_0: $\beta_1 = 0$

H_a: $\beta_1 \neq 0, \beta_1 > 0$, or $\beta_1 < 0$

(If we want to use $D_0 \neq 0$, then first subtract D_0 from all the data for treatment *B*.)

Assumptions

Same as for Method 1. Note that the regression assumption about the "noise" amounts to saying once again that the population variances are the same: Why? Because: If treatment B is not applied, then $y = \beta_0 + \varepsilon$, so the st. deviation of the noise ε is the st. deviation for treatment A. If B is applied, then $y = \beta_0 + \beta_1 + \varepsilon$, so that same st. deviation is the st. deviation for treatment B.

Q What about more than two treatments?

A We use various statistics:

(1) SSA = Sum of Squares Among different treatments or groups, measuring the variability of the treatment means (weighted with the number of samples in each treatment)

$$\text{SSA} = n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + \ldots + n_c(\bar{x}_c - \bar{x})^2$$

Related to that is

MSA = the Mean Square Among different treatments, obtained by dividing SSA by $\nu_1 = c - 1$, which is the number of degrees of freedom for the *c* treatments.

$$\text{MSA} = \frac{\text{SSA}}{c-1}$$

Note that, if $n_1 = .. = n_c = n$, then MSA is *n* times the usual sample variance of the means.

(2) SSW = Sum of Squares Within, measuring the variability within each treatment

$$\text{SSW} = \Sigma_j (x_{1j} - \bar{x}_1)^2 + \Sigma_j (x_{2j} - \bar{x}_2)^2 + \ldots + \Sigma_j (x_{pj} - \bar{x}_{pc})^2$$

where the sums are taken over all measurements within the corresponding treatment.

MSW, obtained by dividing SSW by its degrees of freedom: $\nu_2 = n-c$.

$$\text{MSW} = \frac{\text{SSW}}{n-c}$$

We take the ration of the above statistics to obtain an F-statistic:

$$\text{F} = \frac{\text{MSA}}{\text{MSW}} = \frac{\text{Variation of Means}}{\text{Variation Within}}$$

If F is close to 1, then the variation among sample means is completely explained by variation within treatments, and we will tend to not reject H_0. If it is much larger than 1, we will reject H_0.

Comparison of More than Two Means: Single Factor ANOVA

H_0: $\mu_1 = \mu_2 = ... = \mu_c$

H_a: at least two of the treatment means are different.

Test statistic: $F = \dfrac{\text{MSA}}{\text{MSW}} = \dfrac{\text{Variation of Means}}{\text{Variation Within}}$

Rejection region: $F > F_\alpha$, where F_α is based on $\nu_1 = (c-1)$ numerator and $\nu_2 = (n-c)$ denominator degrees of freedom.

Assumptions

1. Samples are selected independently and randomly
2. All c population distributions are normal with the *same* variance.

We usually summarize all the ANOVA statistics in an "ANOVE table":

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	SSA	$c-1$	$\text{MSA} = \dfrac{\text{SSA}}{c-1}$	$\text{F} = \dfrac{\text{MSA}}{\text{MSW}}$	Observed Significance Level	F_α
Within Groups	SST	$n-c$	$\text{MSW} = \dfrac{\text{SSW}}{n-c}$			

Worksheet 1—Single Factor Analysis: Doing it all By Hand

Your employment agency tracks 15 people after placing them in permanent jobs, obtaining the following results, after 1 year.

Blue Collar Job (A)	White Collar Job (B)	Unemployed (C)
9	11	13
12	11	15
10	11	11

8	13	12
11	9	9

Use a Single Factor ANOVA to determine whether there is any significant difference among the three outcome means at the 95% significance level. (Give the ANOVA table, state the hypotheses, and obtain the conclusion.)

Solution:

H_0: _____

H_a: _____

n = _____ c = _____

$\bar{x}_1$ = _____ $\bar{x}_2$ = _____ $\bar{x}_3$ = _____ $\bar{x}$ = _____

SSA $= n_1(\bar{x}_1-\bar{x})^2 + n_2(\bar{x}_2-\bar{x})^2 + \ldots + n_c(\bar{x}_c-\bar{x})^2$

$= \square\left(\square - \square\right)^2 + \square\left(\square - \square\right)^2$

$+ \square\left(\square - \square\right)^2$

$= \square$

$$\text{MSA} = \frac{\text{SSA}}{c-1} = \frac{\square}{\square} = \square$$

SSW $= \Sigma_j(x_{1j}- \bar{x}_1)^2 + \Sigma_j(x_{2j}- \bar{x}_2)^2 + \ldots + \Sigma_j(x_{pj}- \bar{x}_{pc})^2$

$= \left(\square - \square\right)^2 + \left(\square - \square\right)^2 + \left(\square - \square\right)^2$

$+ \left(\square - \square\right)^2 + \left(\square - \square\right)^2 + \left(\square - \square\right)^2$

$+ \left(\square - \square\right)^2 + \left(\square - \square\right)^2 + \left(\square - \square\right)^2$

$= \square$

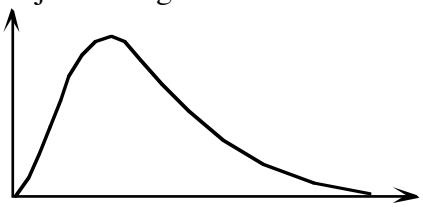
$$\text{MSW} = \frac{\text{SSW}}{n-c} = \frac{\square}{\square} = \square$$

$F = \frac{\square}{\square} = \square$

$F_\alpha = \square$

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups						
Within Groups						

Rejection region:



F in rejection region? _____

Conclusion and Interpretation:

Worksheet 2—Single Factor Analysis with Excel

We want to compare the distance 4 different brands of golf balls will travel when hit with a driver, using a robotic driver. Go to

http://www.zweigmedia.com/qm203/

and download the Golf ball file to see the data.

Brand A	Brand B	Brand C	Brand D
251	263	270	252
245	263	263	249
248	265	278	249
251	255	267	242
261	264	271	247
250	257	266	251
254	263	271	262
245	264	273	249

255	261	276	247
249	256	267	246

Here is the resulting Excel ANOVA analysis (for $\alpha = 0.05$)

Groups	*Count*	*Sum*	*Average*	*Variance*
Column 1	10	2509	250.9	23.4333333
Column 2	10	2611	261.1	13.6555556
Column 3	10	2702	270.2	21.5111111
Column 4	10	2494	249.4	27.3777778

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	2827.8	3	942.6	43.853192	4.149E-12	2.86626545
Within Groups	773.8	36	21.4944444			
Total	3601.6	39				

Solution:

H_0: _____

H_a: _____

Conclusion:

Q How do we decide which specific golf ball goes further than the others?
A We can compare the four brands pairwise using any of the above procedures. To do this, we use a **Tukey-Kramer** procedure:

Tukey-Kramer Procedure for Pairwise Comparison
This is used when we reject the null hypothesis in the ANOVA test (so that there is a significant difference among the means)

Procedure:

(1) Compute the magnitudes of all the pairwise differences $|\bar{x}_i - \bar{x}_j|$

(2) Compute the Critical Range Q_{ij} for this pair (if the numbers in each group are the same, then so are the Q_{ij}):

$$Q_{ij} = Q\sqrt{\frac{\text{MSW}}{2}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

where

Q is the upper-tail critical value from a **studentized range distribution** with c df in the numerator and $n-c$ in the denominator. (Table at the rear of the booklet.)

However, if we have, say, a 95% confidence level for each hypothesis test, we cannot be 95% confident in the result of *all* of them.

Conclusion: If $|\bar{x}_i - \bar{x}_j|$ exceeds Q_{ij}, then there is a statistically significant difference between μ_i and μ_j. Otherwise, there is not.

Worksheet 3—Using Tukey-Kramer

Let us continue our analysis of the 4 brands of golf balls above:

$\bar{x}_1$ = _____ $\bar{x}_2$ = _____ $\bar{x}_3$ = _____

Enter these values in a spreadsheet as shown:

	A	B	C	D	E
1		$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	$\bar{x}_4$
2	$\bar{x}_1$	`=ABS(B1-A2)`			
3	$\bar{x}_2$				
4	$\bar{x}_3$				
5	$\bar{x}_4$				

Then copy across the rows and columns to *instantaneously* compute all the absolute values of the differences.

Next, compute Q_{critical} (there is only one of them .. why?)

MSW = ☐ c = ☐ $n-c$ = ☐

Q = ☐ (from table)

$$Q_{\text{critical}} = Q\sqrt{\frac{\text{MSW}}{2}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

$$= \square\sqrt{\frac{\square}{2}\left[\frac{1}{\square}+\frac{1}{\square}\right]}$$

$$\approx \square \times \sqrt{\square} \approx \square$$

Conclusion:

Exercises for this topic:
p. 505 #1 Do this "by hand"
10 (Use Excel)

Topic 11
ANOVA—Two-Factor
(Based on §13.5 in book)

If we are looking at two factors (e.g. A: Type of club used to hit a ball, and B (2 of them, say): the brand of golf ball (4 of them, say)) then we might want to look at all possible combinations, or **treatments**: 8 of them. An experiment which includes all the possible treatments is called a **complete factorial experiment.**

S'pose that Factor 1 has a levels and Factor 2 has b levels, so that there are ab treatments altogether. We are interested in two kinds of results:

Main Effect of Factor *A*:
H_0: No difference among the a levels in Factor *A*:
(i.e., the brand of golf ball does not effect the distance traveled
H_a: at least two of the factor *A* means differ.)
Test statistic: $F = \frac{\text{MS(A)}}{\text{MSW}}$ Rejection region $F>F_\alpha$ based on $(a-1)$ numerator & $(n-ab)$ denominator.

Main Effect of Factor *B*:
H_0: No difference among the b levels in Factor *B*:
(i.e., the type of golf club does not effect the distance traveled
H_a: at least two of the factor *B* means differ.)
Test statistic: $F = \frac{\text{MS(B)}}{\text{MSW}}$ Rejection region $F>F_\alpha$ based on $(b-1)$ numerator & $(n-ab)$ denominator.

Interaction:
H_0: Factors *A* and *B* do not interact to effect the response mean
(i.e., changing the golf ball has no effect on the ratios of the mean distances for the type of club used.)
H_a: *A* and *B* do interact to effect the response mean.)
Test statistic: $F = \frac{\text{MS(AB)}}{\text{MSW}}$ Rejection region $F>F_\alpha$ based on $(a-1)(b-1)$ numerator & $(n-ab)$ denominator.

Q Exactly what are all these things MS(A), MS(B) and MS(AB), etc?
A They are obtained as follows.

$$\text{SS}(A) = n_1(\bar{x}_1-\bar{x})^2 + n_2(\bar{x}_2-\bar{x})^2 + \ldots + n_a(\bar{x}_a-\bar{x})^2,$$

where the $\bar{x}_1$ is the mean for all data from level 1 of Factor A, $\bar{x}_2$ is the mean for level 2, etc. (just ignore which level of Factor *B* they belong to). $\bar{x}$ is the overall mean. Then, to get MS(A), divide by the (number of treatments for Factor A) -1:

$$\text{MS(A)} = \frac{\text{SS}(A)}{a-1}$$

SST(B) and MS(B) are defined similarly.

For SS(AB), use the sum of all terms $n_{ij}\,(\bar{x}_{ij} - \bar{x}_j - \bar{x}_j + \bar{x})^2$

where $\bar{x}_{ij}$ is the mean of all data from treatment (i, j); that is, level i of factor A and level j of factor B, and n_{ij} is the number of these data points.

$$\text{So, MS(AB)} = \frac{\text{SS(AB)}}{(a-1)(b-1)}$$

Question: What do we do when there is interaction?
Answer: The interaction test should be done *first* because, it there is interaction, then the results for main effects are not informative (certain levels of Factor A might respond favorably with certain levels of Factor B) and the Main Effects statistics combine all the levels of one of the factors. When there is interaction, the only meaningful pairwise comparisons are among all the *ab* treatments. That is, regard the entire experiment as a single factor one with *ab* different levels, and do a pairwise comparison using Tukey-Kramer.

Question: If there is no interaction?
Answer Then do the entire analysis. If Factor A has an effect, then do a pairwise comparison among the *a* levels in Factor A using the following critical value for *Q* (due to Tukey —by himself, this time):

$$Q_{\text{critical}} = Q\sqrt{\frac{\text{MSW}}{bn'}}$$

where n' = number of data scores within each treatment, and where *Q* has the following degrees of freedom: numerator: *a*, denominator: $ab(n'-1)$

Worksheet 1—Two Factor ANOVA with Excel
We consider more golf club data where, this time the factors are:

Factor A: type of club; $a = 2$
Factor B: brand of club; $b = 4$

The following data is at

http://www.zweigmedia.com/qm203/

under *Golfball Two Factor*.

		B: BRAND			
		1	**2**	**3**	**4**
A: TYPE	**Driver**	227	238	241	220
		232	232	247	229
		234	227	240	233
		221	237	245	238
	5 Iron	164	184	186	170
		180	181	193	179
		167	180	190	184

173	186	192	187

Here is the Excel Two Factor (with replication) output for this data (we replicate this in class). Note that the input data must include the headings: it is the block outlined above.

SUMMARY	Brand 1	Brand 2	Brand 3	Brand 4	Total
Driver					
Count	4	4	4	4	16
Sum	914	934	973	920	3741
Average	228.5	233.5	243.25	230	233.8125
Variance	33.6666667	25.6666667	10.9166667	58	60.8291667
Count	4	4	4	4	16
Sum	684	731	761	720	2896
Average	171	182.75	190.25	180	181
Variance	50	7.58333333	9.58333333	55.3333333	75.0666667
Total					
Count	8	8	8	8	
Sum	1598	1665	1734	1640	
Average	199.75	208.125	216.75	205	
Variance	980.5	750.125	811.357143	762.857143	

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
	SS(A)		**MS(A)**			
Sample	22313.2813	1	22313.2813	711.889332	2.3699E-19	4.25967528
	SS(B)		**MS(B)**			
Columns	1217.84375	3	405.947917	12.9514789	3.1138E-05	3.00878611
	SS(AB)		**MS(AB)**			
Interaction	68.34375	3	22.78125	0.72681954	0.54597736	3.00878611
	SSW		**MSW**			
Within	752.25	24	31.34375			
Total	24351.7188	31				

Let us now test the following hypotheses, as shown in the box before the example:

Factor A = type of club; $a = 2$

Factor B = brand of club; $b = 4$

n = number of data points = 32

Interaction: H_0: _____

H_a: _____

$P-$value: _____

Conclusion:

Main Effect of Factor *A*: H_0: _____

H_a: _____

$P-$value: _____

Conclusion:

Main Effect of Factor *B*: H_0: _____

H_a: _____

$P-$value: _____

Conclusion:

Tukey for Factor A:

$\bar{x}_1$ = Average for all drivers = 233.8125 (*Get this from the above output*)

$\bar{x}_2$ = Average for all 5-irons = ☐

$|\bar{x}_2 - \bar{x}_1|$ = |☐ − ☐| = ☐

n' = Number of data scores within each treatment = ☐

Degrees of freedom for Q: Numerator: a = ☐

Denominator: $ab(n'-1)$ = ☐

$$Q_{\text{crit}} = Q\sqrt{\frac{\text{MSW}}{bn'}} = \square\sqrt{\frac{\square}{\square\ \square}}$$

Conclusion:

Tukey for Factor B:

$\bar{x}_1$ = Average for all brand 1 golf balls = 199.75 (Get this from the above output)

$\bar{x}_2$ = Average for all brand 2 golf balls = ☐

$\bar{x}_3$ = Average for all brand 3 golf balls = ☐

$\bar{x}_4$ = Average for all brand 4 golf balls = ☐

Differences Table: $|\bar{x}_i - \bar{x}_j|$

	A	B	C	D	E
1		$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	$\bar{x}_4$
2	$\bar{x}_1$				
3	$\bar{x}_2$				
4	$\bar{x}_3$				
5	$\bar{x}_4$				

n' = Number of data scores within each treatment = ☐

Degrees of freedom for Q: Numerator: b = ☐

Denominator: $ab(n'-1)$ = ☐

$$Q_{\text{crit}} = Q\sqrt{\frac{\text{MSW}}{an'}} = \square\sqrt{\frac{\square}{\square\ \square}}$$

Conclusion:

Exercises for this topic:
p. 527 #30

Topic 12
Quality Improvement: Types of Variation
(Based on §20.2 in book)

Basically, we are interested in monitoring the output of some (industrial) process to check for patterns that might indicate a production problem. We use **time series plot** together with a **centerline** drawn at the intended mean.

Examples of Variation:
Oscillation up & down relative to mean
Uptrend/downtrend
increasing variance
cyclical behavior
meandering (autocorrelation)
outlier/shock
level shift (jump to new level)

The output distribution is characterized by its *mean* and *variance*. If neither of these quantities changes with time, then the process is in a state of **statistical control**. Otherwise, it's **out of statistical control**. Note that random behavior is not a sign of being out of control. However, variation of its mean and variance is.

Testing for Statistical Control

The First Test: The 3σ-Control Limit

Let H_0: Process is under control (normally distributed with st. deviation σ)
H_a: Process is out of control

If the process is under control, then 0.0027 (or 0.27%) of the data point should lie outside 3 standard deviations from the mean. So, every time the machine produces a widget, we can measure its size, and then test H_o. Let our rule for rejection be this:

Rule for rejecting H_0: If the reading is outside 3σ, then reject H_0.

The probability of a Type I error is then

P(we reject H_0 | H_0 is valid)
= P(reading is outside 3σ | data is normally distributed with st. deviation σ)
= 0.0027.

Question What do we use as the mean and standard deviation for the process?
Answer The sample mean and st. deviation, if that's all we can find.

Types of Charts:
First, we look at charts to monitor the ranges of a process (R charts) and the sample means ($\bar{x}$ charts). The first chart that should be examined for a specific process is the R chart, because if it is out of statistical control, then the information given on the $\bar{x}$ chart may not be meaningful.

1. *R* chart

R stands for Range, and this gives a useful estimate for monitoring the standard deviation of smallish samples. To measure *R*, we use small samples of output readings, compute the range of each, and then plot the data. In the graph, we use a **centerline** and **upper and lower control limits** (UCL, LCL). These are estimates of the following quantities:

> Centerline = estimate of μ (population mean of the *ranges)*
> UCL = estimate of $\mu + 3\sigma/\sqrt{n}$ (n = sample size, σ = standard deviation of the ranges)

Unbiased estimators are given as follows:

Estimator of *R*: $\overline{R}$	← Centerline
Estimator for $\mu - 3\sigma/\sqrt{n} = D_3\overline{R}$	← LCL
Estimator for $\mu + 3\sigma/\sqrt{n} = D_4\overline{R}$	← UCL

where D_3 and D_4 are obtained from the control chart table at the back.. (It is computed from intermediate statistics called d_2 and d_3, also shown in that table).

Given these limits, we graph the time series in question (*R* in this case) and use the following pattern recognition rules to determine whether the process is out of statistical control:

Pattern Analysis Rules

These are rules to spot **rare events**; that is, events that indicate a likelihood that a process is out of control.

H_0: Process is under control (normally distributed with a fixed standard deviation)
H_a: Process is out of control

We reject H_0 if any one of the following conditions are found:

Rule 1: One or more points beyond the UCL or LCL (Outlier or increasing variance)
Rule 2: 8 points in a row on the same side of the centerline (Meandering, Uptrend/downtrend, or level shift)
Rule 3: Six points in a row monotonically increasing or decreasing (Uptrend/downtrend)
Rule 4: 14 points in a row oscillating up & down (Oscillation)

Note: Detecting cyclical behavior is more tricky, and may require more sophisticated methods (such as Fourier transform methods).

When a process is suspected of being out of control, the process should be analyzed to determine what, if any, changes should be made. Even when it is in statistical control, the process might not be satisfactory — for instance, the mean value of *R* might be too large, reflecting an unacceptably large variation in the product being manufactured.

Worksheet 1—*R* Chart

The following data is at

http://www.zweigmedia.com/qm203/

under *Soda Bottle Fills:*

Sample #	Soda Bottle Fills (liters)				
1	1.006	1.013	1.015	0.987	1.006
2	0.997	1.002	1	1.006	1.013
3	0.996	0.997	0.986	0.99	1.012
4	0.999	0.993	0.986	1.003	0.986
5	0.995	0.992	0.994	1.012	1.003
6	1.009	0.992	0.999	1.013	1.002
7	0.989	1.007	0.997	0.987	1.001
8	0.995	0.998	0.989	0.993	0.996
9	1	0.997	1.008	1.008	1.006
10	1.006	0.987	0.998	0.994	1.002
11	1.009	0.988	0.999	1.003	0.992
12	0.996	0.994	0.99	0.988	0.992
13	0.992	0.995	1.012	1.013	0.997
14	0.986	1.001	0.99	1.008	1.004
15	1.005	1	0.999	1.002	1.008
16	0.988	1.009	0.993	0.99	1.01
17	0.997	1.008	1.011	1.007	1.006
18	0.993	1.001	0.986	0.987	1.002
19	1.014	1.006	1.001	0.986	1.009
20	0.991	1.001	0.993	1.01	1.007
21	1.008	1.008	1.012	0.989	0.999
22	0.988	0.988	0.99	0.987	0.996
23	1.013	1.009	1.014	0.989	0.993
24	0.988	0.988	1	0.99	1.007
25	0.997	0.995	1.008	1.013	1.011
26	0.996	1.013	1.01	0.996	0.989
27	1.008	1.014	0.996	0.991	0.989
28	1.002	1.005	1.006	0.991	0.993
29	0.988	1.013	0.993	0.996	1.014
30	1.013	1.01	1.006	1.005	1.013

Control limits for R-chart:

Centerline = $\bar{R}$ = ☐

D_3 = ☐ D_4 = ☐

LCL = $D_3\bar{R}$ = ☐

UCL = $D_4\bar{R}$ = ☐

Now we graph the ranges with the control limits, and look at each of the pattern recognition rules:

Rule 1: One or more points beyond the UCL or LCL (Outlier or increasing variance)

☐ Yes ☐ No

Rule 2: 8 points in a row on the same side of the centerline (Meandering, Uptrend/downtrend, or level shift)

☐ Yes ☐ No

Rule 3: Six points in a row monotonically increasing or decreasing (Uptrend/downtrend)

☐ Yes ☐ No

Rule 4: 14 points in a row oscillating up & down (Oscillation)

☐ Yes ☐ No

Conclusion:

2. Means Chart ($\bar{\bar{x}}$ Chart)

Used to detect changes in the sample mean: Here we plot successive sample means vs. time.

Centerline = estimate of μ (population mean of the x-values)

UCL = estimate of $\mu + 3\sigma/\sqrt{n}$ (n = sample size, σ = standard deviation of the samples)

Unbiased estimators are given as follows:

Estimator of R: $\bar{\bar{x}}$	← Centerline
Estimator for $\mu - 3\sigma/\sqrt{n} = \bar{\bar{x}} - A_2\bar{R}$	← LCL
Estimator for $\mu + 3\sigma/\sqrt{n} = \bar{\bar{x}} + A_2\bar{R}$	← UCL

where A_2 is obtained from the control chart table at the back.. (It is computed using the estimate

$$\hat{\sigma} = \frac{\bar{R}}{d_2}, \qquad (d_2 \text{ is an estimator of } R/\sigma \text{ and depends on } n)$$

where $\bar{R}$ is the mean of the sample ranges

Worksheet 2—$\bar{x}$ Chart

We again use the soda data at

http://www.zweigmedia.com/qm203/

under *Soda Bottle Fills*.

Control limits for *R*-chart:

Centerline = $\bar{\bar{x}}$ = ☐

A_2 = ☐ $\bar{R}$ = ☐

LCL = $\bar{\bar{x}} - A_2\bar{R}$ = ☐

UCL = $\bar{\bar{x}} + A_2\bar{R}$ = ☐

Now we graph the means with the control limits, and look at each of the pattern recognition rules:

Rule 1: One or more points beyond the UCL or LCL (Outlier or increasing variance)

☐ Yes ☐ No

Rule 2: 8 points in a row on the same side of the centerline (Meandering, Uptrend/downtrend, or level shift)

☐ Yes ☐ No

Rule 3: Six points in a row monotonically increasing or decreasing (Uptrend/downtrend)

☐ Yes ☐ No

Rule 4: 14 points in a row oscillating up & down (Oscillation)

☐ Yes ☐ No

Conclusion:

3. Proportions Chart (*p*-Chart)

This is used for monitoring *qualitative* processes: e.g., is the product defective or not? The statistic we monitor here is

$$\hat{p} = \frac{\text{\# defective items in the sample}}{\text{\# items in the sample}} \qquad \text{(we are monitoring this statistic)}$$

$$\bar{p} = \frac{\text{total \# defective items}}{\text{total \# items monitored}} \qquad \text{(an estimator of } p\text{)} \leftarrow \text{Centerline}$$

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \qquad (n = \text{size of each sample})$$

Q Where does the standard deviation formula come from?
A When $n = 1$, we are dealing with samples of size 1, and so

$$\hat{p} = \begin{cases} 0 & \text{if the item is not defective} \\ 1 & \text{if it is} \end{cases} .$$

But this is precisely the binomial random variable x which we know from QM I has standard deviation $\sqrt{p(p-1)}$. For larger samples, $\hat{p}$ is just the mean $\bar{x}$ of this binomial random variable, so we are in the sampling distribution of the mean of x. The Central Limit Theorem tells us that the sampling distribution of $\bar{p}$ has standard deviation $\sqrt{p(p-1)} \ / \sqrt{n}$.

Exercises for this topic:
p. 864 #7

Topic 13

Using the Chi-Square (χ^2) Distribution: Multinomial Distributions and Testing for Independence

(12.1, 12.,2 in book)

Multinomial Distributions

Recall that a Bernoulli trial has two outcomes: success/failure. What about three or more outcomes? For this, we talk of the **Multinomial Probability Distribution**. This is really a "vector-valued" binomial distribution: Suppose for example, there are n possible outcomes at every trial. Then let

$$x_1 = \begin{cases} 1 & \text{if the outcome is \#1} \\ 0 & \text{if not} \end{cases}, \quad x_2 = \begin{cases} 1 & \text{if the outcome is \#2} \\ 0 & \text{if not} \end{cases}, \ \ldots,$$

$$x_k = \begin{cases} 1 & \text{if the outcome is \#}n \\ 0 & \text{if not} \end{cases}.$$

Then, if $p_i = P(x_i = 1)$, one has $p_1 + p_2 + \ldots + p_k = 1$. For example, p_1 is the probability of consumers who prefer Brand i.

Here, we test the following hypothesis:

H_0: $p_1, p_2, \ldots, p_k$ have specified values $P_0, \ldots, P_k$

(for instance, in a no-preference situation, $P_i = 1/k$)

H_a: At least one of the probabilities differs from the prescribed value.

The experiment to test the hypothesis: In a sample of size n, let

n_1 = # of responses in which the outcome is outcome #1

n_2 = # of responses in which the outcome is outcome #2

...

n_k = # of responses in which the outcome is outcome #k

and let $n = n_1 + \ldots + n_k$ (total sample size).

If H_0 is valid, then each outcome has expected value $E(x_i) = nP_i$ where n is the total sample size. We use the following test statistic:

$$\chi^2 = \frac{[n_1 - E(x_1)]^2}{E(x_1)} + \ldots + \frac{[n_k - E(x_k)]^2}{E(x_k)}$$

(Note: n_i = observed value, $E(x_i)$ = predicted value)

For sufficiently large[4] sample size n, the (sampling) distribution of χ^2 (in a situation where H_0 is true) is approximately the Chi-Square distribution with $(k-1)$ degrees of freedom.

Q What do we mean by that?

[4] large enough, that is, so that $E(n_i)$ exceeds 5.

A Take many samples of size n, and measure χ^2 for all of these. The resulting probability distribution is approximately Chi-Square with $(k-1)$ degrees of freedom.

Q Why $k-1$ degrees of freedom?

A The loss of one is due to the equation $p_1 + p_2 + \ldots + p_k = 1$.

Q Why is the binomial formula for $E(x_i)$ still valid in a multinomial experiment?

A To compute $E(x_i)$, lump all the other outcomes together as "failure" and you are in the binomial distribution.

Multinomial Probability Distribution

This is a sequence of n independent (identical) trials, where there are k possible outcomes in each trial. With

$$x_i = \begin{cases} 1 & \text{if the outcome is \#}i \\ 0 & \text{if not} \end{cases}$$

and $p_i = P(x_i = 1)$, one has $p_1 + p_2 + \ldots + p_k = 1$.

Hypotheses

H_0: $p_1 = P_1, p_2 = P_2, \ \ldots, p_k = P_k$

H_a: At least one of the $p_i \neq P_i$.

Test Statistic

With n_i = # of responses in which the outcome is outcome #i

Take

$$\chi^2 = \frac{[n_1 - E(x_1)]^2}{E(x_1)} + \ldots + \frac{[n_k - E(x_k)]^2}{E(x_k)}, \text{ where } E(x_i) = nP_i$$

Critical value for $\chi^2 = \chi_\alpha^2$ `=CHIINV(ALPHA, DF)`

Uses df = $k-1$

Notes

1. This is a two-sided test; the test statistic does not differ between positive and negative values of $n_i - E(x_i)$.

2. The test statistic can be rewritten as follows if we divide top & bottom by n:

$$\chi^2 = \frac{n[\hat{p}_1 - P_1]^2}{P_1} + \ldots + \frac{n[\hat{p}_k - P_k]^2}{P_k}$$

where $\hat{p}_i$ is the observed proportion corresponding to outcome i:

$$\hat{p}_i = \frac{n_i}{n}$$

Assumptions

Same as for the binomial distribution: The probability of each outcome is fixed and independent of the history.

Worksheet 1—Multinomial Distribution

Pay increases at Company X depend on evaluation scores, as follows:

Scores > 80 → Merit pay increase

Scores in [50,80] → Standard pay increase

Scores < 50 → No pay increase

The company designed the system with the expectation that 25% would get merit increases, 65% would get standard increases and 10% no increases. The actual results in a survey of the 600 instances (after several years of doing this) was:

$n_1 = 193$; $n_2 = 365$; $n_3 = 42$.

(a) Test these data at the 95% level as to whether the actual pay increases differed significantly from the desired outcomes (indicating to management that they had better change the test...)

(b) Construct a 95% confidence interval for the merit pay outcome.

Solution

(a) $n =$ ☐ $k =$ Number of outcomes = ☐

$P_1 =$ ☐ $P_2 =$ ☐ $P_3 =$ ☐

H_0: ☐

H_a: ☐

$E(x_1) = nP_1 =$ ☐ ☐ = ☐

$E(x_2) = nP_2 =$ ☐ ☐ = ☐

$E(x_3) = nP_3 =$ ☐ ☐ = ☐

$$\chi^2 = \frac{[n_1 - E(x_1)]^2}{E(x_1)} + \frac{[n_2 - E(x_2)]^2}{E(x_2)} + \frac{[n_3 - E(x_3)]^2}{E(x_3)}$$

$$= \frac{[\square - \square]^2}{\square} + \frac{[\square - \square]^2}{\square} + \frac{[\square - \square]^2}{\square}$$

$=$ ☐ $+$ ☐ $+$ ☐

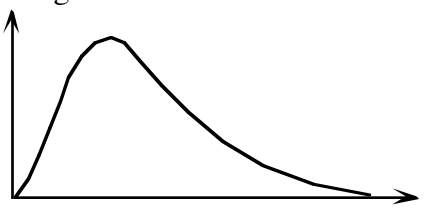
$=$ ☐

Critical value:

df = $k-1$ = ☐

$\chi_\alpha^2 =$ ☐

Rejection region:



Conclusion:

(b) Think of this as a test with two outcomes: Success = Merit pay outcome (outcome #1); Failure = any of the other two. Then we have a binomial distribution, which we approximate with a normal distribution. The confidence interval for x_1 can now be computed using the binomial distribution confidence interval:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

where

$\hat{p}$ = Probability of success (observed) = [ ]

n = [ ] $z_{\alpha/2} = z_{\_\_\_\_} =$ [ ]

$$\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$= \sqrt{\frac{[\ ] \times [\ ]}{[\ ]}} \approx [\ ]$$

This gives the CI as

CI = [ ] ± [ ]

≈ [ ] ± [ ]

≈ [[ ] , [ ]]

Note We can use the above procedure to compare two probability distributions as in the textbook: Think of one of them as the observed set of probabilities and the other as the hypothesized set of probabilities. As far as rejecting the null hypothesis is concerned, it does not mater which set is which.

Testing for Independence

In QM I we all saw tables like following:

Example 1 A survey of 100 stocks shows the following performance after one year

	Increased ≥ 20%	Stayed Within ±20%	Decreased ≥ 20%	Totals
Pharmaceutical Companies	$n_{11} = 10$	$n_{12} = 30$	$n_{13} = 10$	50
Electronic Companies	$n_{21} = 5$	$n_{22} = 0$	$n_{23} = 5$	10
Banking Company	$n_{31} = 15$	$n_{32} = 10$	$n_{33} = 15$	40
Totals	30	40	30	$n = 100$

Associated the cells are probabilities given by

$$p_{ij} = \frac{n_{ij}}{n} \qquad (n = \text{total sample size} = 100 \text{ here})$$

These are called **marginal probabilities**. Let us designate these probabilities as follows:

	Increased ≥ 20%	Stayed Within ±20%	Decreased ≥ 20%	Totals
Pharmaceutical Companies	p_{11}	p_{12}	p_{13}	P_{R1}
Electronic Companies	p_{21}	p_{22}	p_{23}	P_{R2}
Banking Company	p_{31}	p_{32}	p_{33}	P_{R3}
Totals	P_{C1}	P_{C2}	P_{C3}	1

In QM I we were asked such questions as "are "Increased>20%" and "Banking Company" independent? In real life, they were practically *always* dependent, since the requirement for independence is that $p_{11} = P_{R1}P_{C1}$ etc., and getting exact equality would be next-to impossible, given random errors. Instead, we ask the question in the form of a hypothesis:

H_0: The row events are independent of the column events.
H_a: At least one of the row events is not independent of a column event.

Mathematically, H_0 means that

$p_{ij} = P_{Ri}P_{Cj}$ (recalling QM I) for every pair (i, j)

For this example, it means:

H_0: The performance of a stock does not depend on the type of stock
H_a: The performance of a stock does depend on the type of stock

We use, as estimates of the marginal probabilities, the relative frequencies found in the table above, and compute a test statistic.

$$\chi^2 = \sum_{i,j} \frac{[n_{ij} - E(x_{ij})]^2}{E(x_{ij})} \qquad \text{or} \qquad \chi^2 = \sum_{i,j} \frac{n[p_{ij} - P_{Ri}P_{Cj}]^2}{P_{Ri}P_{Cj}}$$

where $E(x_{ij}) = nP_{Ri}P_{Cj}$ = Expected frequency. (Note that this is not the same as the observed frequency p_{ij}.)

Q How many degrees of freedom are there?

A If we are thinking of the P_{ri} and P_{cj} as fixed, the last probability in each row and column is determined by the others, and so df = $(r-1)(c-1)$. In this example, $df = (3-1)(3-1) = 4$.

Worksheet 2 — Testing for Independence

Test the above data for statistical independence.

We use Excel as follows (careful with the totals!)

Observed Freq

	Incr	Same	Decr	Totals
Pharmaceutical	10	30	10	50
Electronic	5	0	5	15
Banking	15	10	15	40
Totals	30	40	30	100

Expected Freq

	Incr	Same	Decr	Totals
Pharmaceutical				
Electronic				
Banking				
Totals				

$$\frac{[n_{ij} - E(x_{ij})]^2}{E(x_{ij})}$$

	Incr	Same	Decr	Totals
Pharmaceutical				
Electronic				
Banking				
Totals				

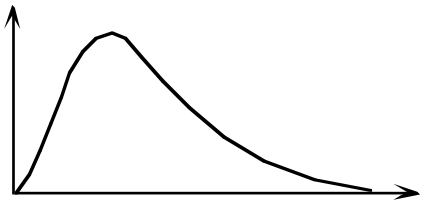
$\chi^2 =$ ☐

Critical value:

df = $(r-1)(c-1)$ = ☐

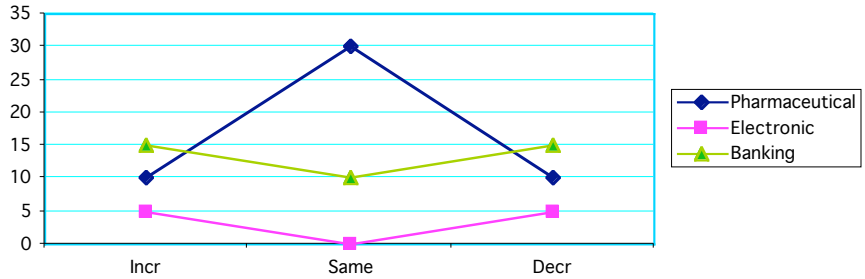
$\chi_\alpha^2 =$ ☐

Rejection region:



Conclusion:

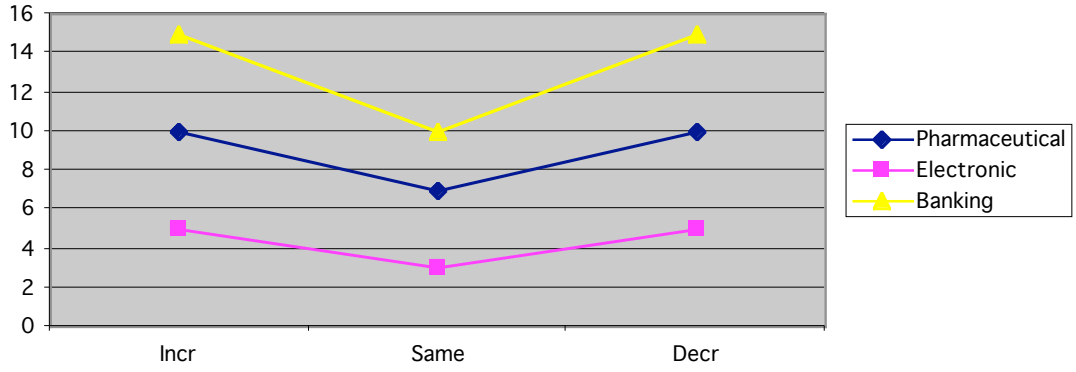
Here is a plot of the observed data:



Had they been statistically independent, the graphs would have been close to parallel

Example 2 (Independent Events)
Here is some adjusted data that gives no rejection of H_0 (try it on your spreadsheet —if it is set up properly, you need do nothing except enter the given data in the unshaded cells:

	Incr	Same	Decr	Totals
Pharmaceutical	10	7	10	27
Electronic	5	3	5	13
Banking	15	10	15	40
Totals	30	20	30	80



Exercises for this topic
Multinomial Distributions: p. 452, #3, #6

Independence: p. 470 #14 Include a plot of the observed data.

Excel Assignment 3
Uncovering Tax Fraud using Benford's Law and Chi Square

You are a tax fraud specialist working for the Internal Revenue Service (IRS), and have just been handed a portion of the tax return from Colossal Conglomerate. The agency suspects that the portion you were handed may be fraudulent, and would like your opinion. Is there any mathematical test, you wonder, that can point to a suspicious tax return based on nothing more than the numbers entered?

You decide, on an impulse, to make a list of the first digits of all the numbers entered in the portion of the Colossal Conglomerate tax return (there are 625 of them). You first reason that, if the tax return is an honest one, the first digits of the numbers should be uniformly distributed. More precisely, if the experiment consists of selecting a number at random from the tax return, and the random variable X is defined to be the first digit of the selected number, then X should have the following probability distribution:

y	1	2	3	4	5	6	7	8	9
$P(X=x)$	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9

You then do a quick calculation based on this probability distribution, and find an expected value of $E(X) = 5$.

Next, you turn to the Colossal Conglomerate return data and calculate the relative frequency (experimental probability) of the actual numbers in the tax return. You find the following results.

Colossal Conglomerate Return

y	1	2	3	4	5	6	7	8	9
$P(Y=y)$	0.31	0.16	0.13	0.11	0.07	0.07	0.05	0.06	0.04

It certainly does look suspicious! For one thing, the smaller digits (especially 1) seem to occur a lot more often than any of the larger digits. Moreover, when you compute the expected value, you obtain $E(Y) = 3.42$, considerably lower than the value of 5 you predicted. Gotcha!

You are about to file a report recommending a detailed audit of Colossal Conglomerate when you recall an article you once read about first digits in lists of numbers. The article dealt with a remarkable discovery in 1938 by Dr. Frank Benford, a physicist at the General Electric company. What Dr. Benford noticed was that the pages of logarithm tables that listed numbers starting with the digits 1 and 2 tended to be more soiled and dog-eared than the pages whose listed numbers started with larger digits—say, 8. For some reason, numbers starting with low digits seemed more prevalent than

numbers starting with high digits. He subsequently analyzed more than 20,000 sets of numbers, such as tables of baseball statistics, listings of widths of rivers, half-lives of radioactive elements, street addresses, and numbers in magazine articles. The result was always the same: inexplicably, numbers starting with low digits tended to appear more frequently than the high ones, with numbers beginning with the digit 1 most prevalent of all.[5] Moreover, the expected value of the first digit was not the expected 5, but 3.44.

Since the first digits in Colossal Conglomerate's return have an expected value of 3.42; very close to Benford's value, it might appear that your suspicion was groundless after all. (Back to the drawing board...)

Out of curiosity, you decide to investigate Benford's discovery more carefully. What you find is that Benford did more than simply observe a strange phenomenon in lists of numbers. He went further and derived the following formula for the probability distribution of first digits in lists of numbers.

$$P(X=x) = \log(1 + 1/x) \qquad (x = 1, 2, ..., 9)$$

You compute these probabilities, and find the following distribution.

x	1	2	3	4	5	6	7	8	9
$P(X=x)$	0.30	0.18	0.12	0.10	0.08	0.07	0.06	0.05	0.05

(a) Give a bar graph which compares the probabilities of the first digits predicted by Benford's law with those observed in the tax return. It should look something like the following (although the data are different).

(b) Apply a Chi-Square test to the hypothesis that $p(X=1)$, $p(X=2)$, ... ,$p(X=9)$ have the values specified by Benford's law at the 95% significance level. Use the Excel setup shown below. (Note that $n = 625$ here and is entered into the cell that calculates χ^2.) What do you conclude about Colossal Conglomerate's tax return?

[5] The does not apply to all lists of numbers. For instance, a list of randomly chosen numbers between 1 and 999 will have first digits uniformly distributed between 1 and 9.

(c) Repeat parts (a) and (b) for the Honest Growth Funds Stockholder Report ($n = 400$) where the distribution of first digits is shown below.

Honest Growth Funds Return

y	1	2	3	4	5	6	7	8	9
$P(Y=y)$	0.28	0.16	0.1	0.11	0.07	0.09	0.05	0.07	0.07

14. Cyclic Fluctuations & Trigonometric Models

Seasonal linear models do not work well to predict smooth cyclic fluctuations because, their graphs are generally sawtooth shapes (see the graph in the preceding section) whereas within each cycle, we would like to model a gradual increase and then decrease (see any of the graphs that illustrated C_t above).

$$y = T_t + C_t$$

To model C_t we use a **trigonometric model**. Before we can do this we need to know the **period *P*** of the cyclical fluctuations. If $f(t)$ is a function of t, then its **period *P*** is the smallest positive number P such that $f(t+P) = f(t)$ for every t (assuming such a P exists). If such a P exists, we refer to $f(t)$ as **periodic**. For instance, if t is time in months, then many business-related time series are periodic with period $P = 12$.

Modeling Periodic Fluctuations
If the period is known to be P, then we use

$$C_t = \beta_0 + \beta_1 \cos\left(\frac{2\pi t}{P}\right) + \beta_2 \sin\left(\frac{2\pi t}{P}\right)$$

The **amplitude** (heights of peaks) and **phase shift** (location of first peak) are determined by β_1 and β_2, while β_0 determines the **baseline** (level about which fluctuations occur).

Example 1
Go back to the utilities data in

www.zweigmedia.com/qm203 → Utilities

but this time we do trigonometric modeling for a model of the form

$$y = T_t + C_t = \beta_0 + \beta_1 t + \beta_2 \cos\left(\frac{2\pi t}{P}\right) + \beta_3 \sin\left(\frac{2\pi t}{P}\right)$$

where $P = 2$ (note that t is in *quarters)* and the period is 6 months (2 quarters):

Setup

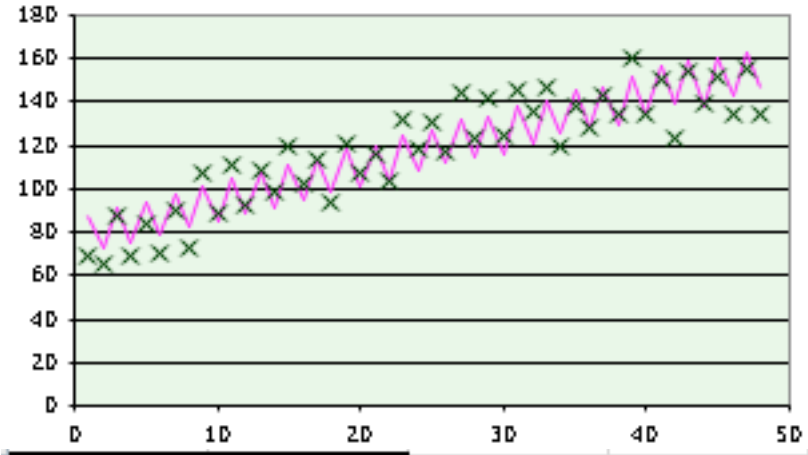
Quarter	Power Load	Month	Cos	Sin
1	69	1	=COS(2*PI()*D3/2)	
2	65	2	1	-2.449E-16

(Use a similar formula for Sin)

After the regression, which we pasted in cell A52 we use the above formula to get he predicted *y*-values:

Month	Cos	Sin	Fitted y
1	-1	1.2246E-16	=$B$68+$B$69*D3+$B$70*E3 + $B$71*F3

Resulting fit:



Regression Statistics	
Multiple R	0.95335072
R Square	0.90887759
Adjusted R S	0.9026647
Standard Erro	8.014483
Observations	48

ANOVA

	df	*SS*	*MS*	*F*
Regression	3	28189.2739	9396.42463	146.28898
Residual	44	2826.20526	64.2319378	
Total	47	31015.4792		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	77.2751382	2.35614279	32.7973069	1.4525E-32
X Variable 1	1.65470579	0.08418351	19.6559376	2.0504E-23
X Variable 2	-8.3111381	1.419494	-5.8550005	5.5101E-07
X Variable 3	2E+14	2.3497E+14	0.85118993	0.39927268

Q Wait a minute! This looks the same as the seasonal regression model. How come?
A The trouble is that we only have data by the quarter, following the patters high/low/high/low... which, if you model by a sine wave, gives the same zigzag shape as linear regression above.

There is another *serious* issue with this data: Look at the coefficient of β_3!!! Its p-value is also ridiculously high, so it really looks like that term should not be there. (Reason: for all the data points we used, the value of $\sin\left(\frac{2\pi t}{P}\right)$ is theoretically zero, which Excel renders as something very close to zero.

But using zero as an independent variable in regression is always asking for trouble!!

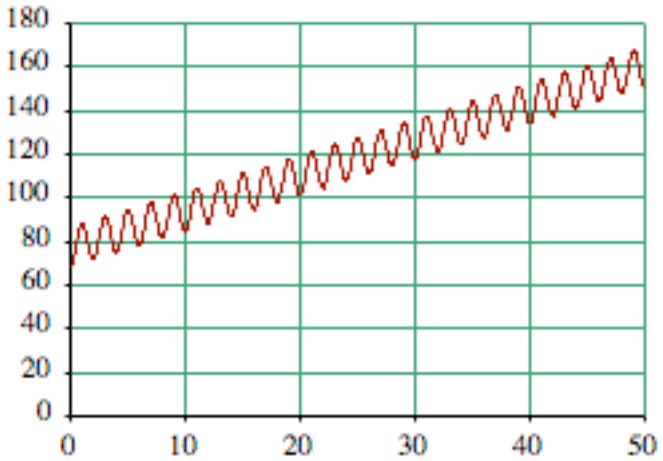
So we immediately jettison the sine term and repeat the regression to get a more acceptable result:

Regression Statistics				
Multiple R	0.95256345			
R Square	0.90737712			
Adjusted R S	0.90326055			
Standard Erro	7.98991428			
Observations	48			
ANOVA				
	df	*SS*	*MS*	*F*
Regression	2	28142.7363	14071.3682	220.420552
Residual	45	2872.74286	63.8387303	
Total	47	31015.4792		
	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	77.4026993	2.34416376	33.0193225	3.458E-33
X Variable 1	1.64597826	0.08330061	19.7594972	8.2603E-24
X Variable 2	-9.0104891	1.15399666	-7.8080721	6.5677E-10

Notice that the *adjusted* r^2 has increased slightly (although as expected the actual r^2 has decreased..)

If we had had some in-between data we would have seen the curved fluctuations more clearly). In fact, here is the actual graph of the regression model

$$y = 77.4 + 1.65t - 9.01\cos\left(\frac{2\pi t}{2}\right)$$
$$= 77.4 + 1.65t - 9.01\cos(\pi t)$$



$y = 77.4 + 1.65t - 9.01\cos(\pi t)$

Exercise for this topic:
Get the pollen count data from

www.zweigmedia.com/qm203 → Airline Empty Seat Volume

The goal here is to model the empty seat volume using a model of the form

$$y = \beta_0 + \beta_1 t + \beta_2 \cos\left(\frac{2\pi t}{P}\right) + \beta_3 \sin\left(\frac{2\pi t}{P}\right)$$

for a suitable P.

(a) Obtain the regression.

(b) Graph the regression function versus time.

(c) Identify the secular trend. What does it say about the trend in empty seats?

Normal Distribution: $P(Z \leq z)$ **Excel: =NORMSDIST(z)**

Negative z

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-0.0	0.50000	0.49601	0.49202	0.48803	0.48405	0.48006	0.47608	0.47210	0.46812	0.46414
-0.1	0.46017	0.45620	0.45224	0.44828	0.44433	0.44038	0.43644	0.43251	0.42858	0.42465
-0.2	0.42074	0.41683	0.41294	0.40905	0.40517	0.40129	0.39743	0.39358	0.38974	0.38591
-0.3	0.38209	0.37828	0.37448	0.37070	0.36693	0.36317	0.35942	0.35569	0.35197	0.34827
-0.4	0.34458	0.34090	0.33724	0.33360	0.32997	0.32636	0.32276	0.31918	0.31561	0.31207
-0.5	0.30854	0.30503	0.30153	0.29806	0.29460	0.29116	0.28774	0.28434	0.28096	0.27760
-0.6	0.27425	0.27093	0.26763	0.26435	0.26109	0.25785	0.25463	0.25143	0.24825	0.24510
-0.7	0.24196	0.23885	0.23576	0.23270	0.22965	0.22663	0.22363	0.22065	0.21770	0.21476
-0.8	0.21186	0.20897	0.20611	0.20327	0.20045	0.19766	0.19489	0.19215	0.18943	0.18673
-0.9	0.18406	0.18141	0.17879	0.17619	0.17361	0.17106	0.16853	0.16602	0.16354	0.16109
-1	0.15866	0.15625	0.15386	0.15151	0.14917	0.14686	0.14457	0.14231	0.14007	0.13786
-1.1	0.13567	0.13350	0.13136	0.12924	0.12714	0.12507	0.12302	0.12100	0.11900	0.11702
-1.2	0.11507	0.11314	0.11123	0.10935	0.10749	0.10565	0.10383	0.10204	0.10027	0.09853
-1.3	0.09680	0.09510	0.09342	0.09176	0.09012	0.08851	0.08692	0.08534	0.08379	0.08226
-1.4	0.08076	0.07927	0.07780	0.07636	0.07493	0.07353	0.07215	0.07078	0.06944	0.06811
-1.5	0.06681	0.06552	0.06426	0.06301	0.06178	0.06057	0.05938	0.05821	0.05705	0.05592
-1.6	0.05480	0.05370	0.05262	0.05155	0.05050	0.04947	0.04846	0.04746	0.04648	0.04551
-1.7	0.04457	0.04363	0.04272	0.04182	0.04093	0.04006	0.03920	0.03836	0.03754	0.03673
-1.8	0.03593	0.03515	0.03438	0.03362	0.03288	0.03216	0.03144	0.03074	0.03005	0.02938
-1.9	0.02872	0.02807	0.02743	0.02680	0.02619	0.02559	0.02500	0.02442	0.02385	0.02330
-2	0.02275	0.02222	0.02169	0.02118	0.02068	0.02018	0.01970	0.01923	0.01876	0.01831
-2.1	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578	0.01539	0.01500	0.01463	0.01426
-2.2	0.01390	0.01355	0.01321	0.01287	0.01255	0.01222	0.01191	0.01160	0.01130	0.01101
-2.3	0.01072	0.01044	0.01017	0.00990	0.00964	0.00939	0.00914	0.00889	0.00866	0.00842
-2.4	0.00820	0.00798	0.00776	0.00755	0.00734	0.00714	0.00695	0.00676	0.00657	0.00639
-2.5	0.00621	0.00604	0.00587	0.00570	0.00554	0.00539	0.00523	0.00508	0.00494	0.00480
-2.6	0.00466	0.00453	0.00440	0.00427	0.00415	0.00402	0.00391	0.00379	0.00368	0.00357
-2.7	0.00347	0.00336	0.00326	0.00317	0.00307	0.00298	0.00289	0.00280	0.00272	0.00264
-2.8	0.00256	0.00248	0.00240	0.00233	0.00226	0.00219	0.00212	0.00205	0.00199	0.00193
-2.9	0.00187	0.00181	0.00175	0.00169	0.00164	0.00159	0.00154	0.00149	0.00144	0.00139
-3	0.00135	0.00131	0.00126	0.00122	0.00118	0.00114	0.00111	0.00107	0.00104	0.00100
-3.1	0.00097	0.00094	0.00090	0.00087	0.00084	0.00082	0.00079	0.00076	0.00074	0.00071
-3.2	0.00069	0.00066	0.00064	0.00062	0.00060	0.00058	0.00056	0.00054	0.00052	0.00050
-3.3	0.00048	0.00047	0.00045	0.00043	0.00042	0.00040	0.00039	0.00038	0.00036	0.00035
-3.4	0.00034	0.00032	0.00031	0.00030	0.00029	0.00028	0.00027	0.00026	0.00025	0.00024
-3.5	0.00023	0.00022	0.00022	0.00021	0.00020	0.00019	0.00019	0.00018	0.00017	0.00017
-3.6	0.00016	0.00015	0.00015	0.00014	0.00014	0.00013	0.00013	0.00012	0.00012	0.00011
-3.7	0.00011	0.00010	0.00010	0.00010	0.00009	0.00009	0.00008	0.00008	0.00008	0.00008
-3.8	0.00007	0.00007	0.00007	0.00006	0.00006	0.00006	0.00006	0.00005	0.00005	0.00005
-3.9	0.00005	0.00005	0.00004	0.00004	0.00004	0.00004	0.00004	0.00004	0.00003	0.00003

Positive *z*

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91308	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983
3.6	0.99984	0.99985	0.99985	0.99986	0.99986	0.99987	0.99987	0.99988	0.99988	0.99989
3.7	0.99989	0.99990	0.99990	0.99990	0.99991	0.99991	0.99992	0.99992	0.99992	0.99992
3.8	0.99993	0.99993	0.99993	0.99994	0.99994	0.99994	0.99994	0.99995	0.99995	0.99995
3.9	0.99995	0.99995	0.99996	0.99996	0.99996	0.99996	0.99996	0.99996	0.99997	0.99997

t-Statistic **Excel:** `=TINV(2*α,df)`

df	0.1	0.05	0.01	0.025	0.005
1	3.078	6.314	31.821	12.706	63.656
2	1.886	2.920	6.965	4.303	9.925
3	1.638	2.353	4.541	3.182	5.841
4	1.533	2.132	3.747	2.776	4.604
5	1.476	2.015	3.365	2.571	4.032
6	1.440	1.943	3.143	2.447	3.707
7	1.415	1.895	2.998	2.365	3.499
8	1.397	1.860	2.896	2.306	3.355
9	1.383	1.833	2.821	2.262	3.250
10	1.372	1.812	2.764	2.228	3.169
11	1.363	1.796	2.718	2.201	3.106
12	1.356	1.782	2.681	2.179	3.055
13	1.350	1.771	2.650	2.160	3.012
14	1.345	1.761	2.624	2.145	2.977
15	1.341	1.753	2.602	2.131	2.947
16	1.337	1.746	2.583	2.120	2.921
17	1.333	1.740	2.567	2.110	2.898
18	1.330	1.734	2.552	2.101	2.878
19	1.328	1.729	2.539	2.093	2.861
20	1.325	1.725	2.528	2.086	2.845
21	1.323	1.721	2.518	2.080	2.831
22	1.321	1.717	2.508	2.074	2.819
23	1.319	1.714	2.500	2.069	2.807
24	1.318	1.711	2.492	2.064	2.797
25	1.316	1.708	2.485	2.060	2.787
26	1.315	1.706	2.479	2.056	2.779
27	1.314	1.703	2.473	2.052	2.771
28	1.313	1.701	2.467	2.048	2.763
29	1.311	1.699	2.462	2.045	2.756
30	1.310	1.697	2.457	2.042	2.750
31	1.309	1.696	2.453	2.040	2.744
32	1.309	1.694	2.449	2.037	2.738
33	1.308	1.692	2.445	2.035	2.733
34	1.307	1.691	2.441	2.032	2.728
35	1.306	1.690	2.438	2.030	2.724
40	1.303	1.684	2.423	2.021	2.704
45	1.301	1.679	2.412	2.014	2.690
50	1.299	1.676	2.403	2.009	2.678
75	1.293	1.665	2.377	1.992	2.643
100	1.290	1.660	2.364	1.984	2.626
200	1.286	1.653	2.345	1.972	2.601
1000	1.282	1.646	2.330	1.962	2.581

Critical Values for Durbin-Watson (α = 0.05)

	k = 1		k = 2		k = 3		k = 4		k = 5	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1.08	1.36	0.95	1.54	0.81	1.75	0.69	1.98	0.56	2.22
16	1.11	1.37	0.98	1.54	0.86	1.73	0.73	1.94	0.62	2.16
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.66	2.10
18	1.16	1.39	1.05	1.54	0.93	1.70	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.54	0.97	1.69	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.89	1.83	0.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.65	1.04	1.77	0.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	0.98	1.87
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.75	1.00	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.14	1.81
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80
36	1.41	1.52	1.35	1.59	1.30	1.65	1.24	1.73	1.18	1.80
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.20	1.79
39	1.44	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.33	1.77
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.37	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
75	1.60	1.65	1.57	1.68	1.54	1.71	1.52	1.74	1.49	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
85	1.62	1.67	1.60	1.70	1.58	1.72	1.55	1.75	1.53	1.77
90	1.64	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
95	1.65	1.69	1.62	1.71	1.60	1.73	1.58	1.76	1.56	1.78
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

Critical values of F ($\alpha = 0.05$)

Excel: =FINV(0.05,df_n,df_d)

df	Numerator →						
Denominator ↓		**1**	**2**	**3**	**4**	**5**	**6**
	1	161.446	199.499	215.707	224.583	230.160	233.988
	2	18.513	19.000	19.164	19.247	19.296	19.329
	3	10.128	9.552	9.277	9.117	9.013	8.941
	4	7.709	6.944	6.591	6.388	6.256	6.163
	5	6.608	5.786	5.409	5.192	5.050	4.950
	6	5.987	5.143	4.757	4.534	4.387	4.284
	7	5.591	4.737	4.347	4.120	3.972	3.866
	8	5.318	4.459	4.066	3.838	3.688	3.581
	9	5.117	4.256	3.863	3.633	3.482	3.374
	10	4.965	4.103	3.708	3.478	3.326	3.217
	11	4.844	3.982	3.587	3.357	3.204	3.095
	12	4.747	3.885	3.490	3.259	3.106	2.996
	13	4.667	3.806	3.411	3.179	3.025	2.915
	14	4.600	3.739	3.344	3.112	2.958	2.848
	15	4.543	3.682	3.287	3.056	2.901	2.790
	16	4.494	3.634	3.239	3.007	2.852	2.741
	17	4.451	3.592	3.197	2.965	2.810	2.699
	18	4.414	3.555	3.160	2.928	2.773	2.661
	19	4.381	3.522	3.127	2.895	2.740	2.628
	20	4.351	3.493	3.098	2.866	2.711	2.599
	21	4.325	3.467	3.072	2.840	2.685	2.573
	22	4.301	3.443	3.049	2.817	2.661	2.549
	23	4.279	3.422	3.028	2.796	2.640	2.528
	24	4.260	3.403	3.009	2.776	2.621	2.508
	25	4.242	3.385	2.991	2.759	2.603	2.490
	26	4.225	3.369	2.975	2.743	2.587	2.474
	27	4.210	3.354	2.960	2.728	2.572	2.459
	28	4.196	3.340	2.947	2.714	2.558	2.445
	29	4.183	3.328	2.934	2.701	2.545	2.432
	30	4.171	3.316	2.922	2.690	2.534	2.421

Studentized Q Distribution ($\alpha = 0.05$)

df	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	18.0	27.0	32.8	37.1	40.4	43.1	45.4	47.4	49.1	50.6	52.0	53.2	54.3	55.4
2	6.09	8.33	9.80	10.9	11.7	12.4	13.0	13.5	14.0	14.4	14.8	15.1	15.4	15.7
3	4.50	5.91	6.83	7.50	8.04	8.48	8.85	9.18	9.46	9.72	9.95	10.2	10.4	10.5
4	3.93	5.04	5.76	6.29	6.71	7.05	7.35	7.60	7.83	8.03	8.21	8.37	8.53	8.66
5	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	7.00	7.17	7.32	7.47	7.60	7.72
6	3.46	4.34	4.90	5.31	5.63	5.90	6.12	6.32	6.49	6.65	6.79	6.92	7.03	7.14
7	3.34	4.17	4.68	5.06	5.36	5.61	5.82	6.00	6.16	6.30	6.43	6.55	6.66	6.76
8	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92	6.05	6.18	6.29	6.39	6.48
9	3.20	3.95	4.42	4.76	5.02	5.24	5.43	5.60	5.74	5.87	5.98	6.09	6.19	6.28
10	3.15	3.88	4.33	4.65	4.91	5.12	5.31	5.46	5.60	5.72	5.83	5.94	6.03	6.11
11	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49	5.61	5.71	5.81	5.90	5.98
12	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.40	5.51	5.62	5.71	5.80	5.88
13	3.06	3.74	4.15	4.45	4.69	4.89	5.05	5.19	5.32	5.43	5.53	5.63	5.71	5.79
14	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25	5.36	5.46	5.55	5.64	5.71
15	3.01	3.67	4.08	4.37	4.60	4.78	4.94	5.08	5.20	5.31	5.40	5.49	5.57	5.65
16	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15	5.26	5.35	5.44	5.52	5.59
17	2.98	3.63	4.02	4.30	4.52	4.71	4.86	4.99	5.11	5.21	5.31	5.39	5.47	5.54
18	2.97	3.61	4.00	4.28	4.50	4.67	4.82	4.96	5.07	5.17	5.27	5.35	5.43	5.50
19	2.96	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04	5.14	5.23	5.32	5.39	5.46
20	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01	5.11	5.20	5.28	5.36	5.43
24	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92	5.01	5.10	5.18	5.25	5.32
30	2.89	3.49	3.85	4.10	4.30	4.46	4.60	4.72	4.82	4.92	5.00	5.08	5.15	5.21
40	2.56	3.44	3.79	4.04	4.23	4.39	4.52	4.64	4.74	4.82	4.90	4.98	5.04	5.11
60	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65	4.73	4.81	4.88	4.94	5.00
120	2.80	3.36	3.69	3.92	4.10	4.24	4.36	4.47	4.56	4.64	4.71	4.78	4.84	4.90
∞	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47	4.55	4.62	4.69	4.74	4.80

Statistical Control Chart Factors

n	A	A_2	A_3	d_2	d_3	D_1	D_2	D_3	D_4
2	2.121	1.880	2.659	1.128	0.853	0.000	3.686	0.000	3.267
3	1.732	1.023	1.954	1.693	0.888	0.000	4.358	0.000	2.575
4	1.500	0.729	1.628	2.059	0.880	0.000	4.698	0.000	2.282
5	1.342	0.577	1.427	2.326	0.864	0.000	4.918	0.000	2.114
6	1.225	0.483	1.287	2.534	0.848	0.000	5.079	0.000	2.004
7	1.134	0.419	1.182	2.704	0.833	0.205	5.204	0.076	1.924
8	1.061	0.373	1.099	2.847	0.820	0.388	5.307	0.136	1.864
9	1.000	0.337	1.032	2.970	0.808	0.547	5.394	0.184	1.816
10	0.949	0.308	0.975	3.078	0.797	0.686	5.469	0.223	1.777
11	0.905	0.285	0.927	3.173	0.787	0.811	5.535	0.256	1.744
12	0.866	0.266	0.886	3.258	0.778	0.923	5.594	0.283	1.717
13	0.832	0.249	0.850	3.336	0.770	1.025	5.647	0.307	1.693
14	0.802	0.235	0.817	3.407	0.763	1.118	5.696	0.328	1.672
15	0.775	0.223	0.789	3.472	0.756	1.203	5.740	0.347	1.653
16	0.750	0.212	0.763	3.532	0.750	1.282	5.782	0.363	1.637
17	0.728	0.203	0.739	3.588	0.744	1.356	5.820	0.378	1.622
18	0.707	0.194	0.718	3.640	0.739	1.424	5.856	0.391	1.609
19	0.688	0.187	0.698	3.689	0.733	1.489	5.889	0.404	1.596
20	0.671	0.180	0.680	3.735	0.729	1.549	5.921	0.415	1.585
21	0.655	0.173	0.663	3.778	0.724	1.606	5.951	0.425	1.575
22	0.640	0.167	0.647	3.819	0.720	1.660	5.979	0.435	1.565
23	0.626	0.162	0.633	3.858	0.716	1.711	6.006	0.443	1.557
24	0.612	0.157	0.619	3.895	0.712	1.759	6.032	0.452	1.548
25	0.600	0.153	0.606	3.931	0.708	1.805	6.056	0.459	1.541

Chi-Square

DF	$\chi^2_{0.1}$	$\chi^2_{0.05}$	$\chi^2_{0.01}$
1	2.7055	3.8415	6.6349
2	4.6052	5.9915	9.2104
3	6.2514	7.8147	11.3449
4	7.7794	9.4877	13.2767
5	9.2363	11.0705	15.0863
6	10.6446	12.5916	16.8119
7	12.0170	14.0671	18.4753
8	13.3616	15.5073	20.0902
9	14.6837	16.9190	21.6660
10	15.9872	18.3070	23.2093
11	17.2750	19.6752	24.7250
12	18.5493	21.0261	26.2170
13	19.8119	22.3620	27.6882
14	21.0641	23.6848	29.1412
15	22.3071	24.9958	30.5780
16	23.5418	26.2962	31.9999
17	24.7690	27.5871	33.4087
18	25.9894	28.8693	34.8052
19	27.2036	30.1435	36.1908
20	28.4120	31.4104	37.5663
21	29.6151	32.6706	38.9322
22	30.8133	33.9245	40.2894
23	32.0069	35.1725	41.6383
24	33.1962	36.4150	42.9798
25	34.3816	37.6525	44.3140
26	35.5632	38.8851	45.6416
27	36.7412	40.1133	46.9628
28	37.9159	41.3372	48.2782
29	39.0875	42.5569	49.5878
30	40.2560	43.7730	50.8922
60	74.3970	79.0820	88.3794
120	140.2326	146.5673	158.9500
1000	1057.7240	1074.6794	1106.9690