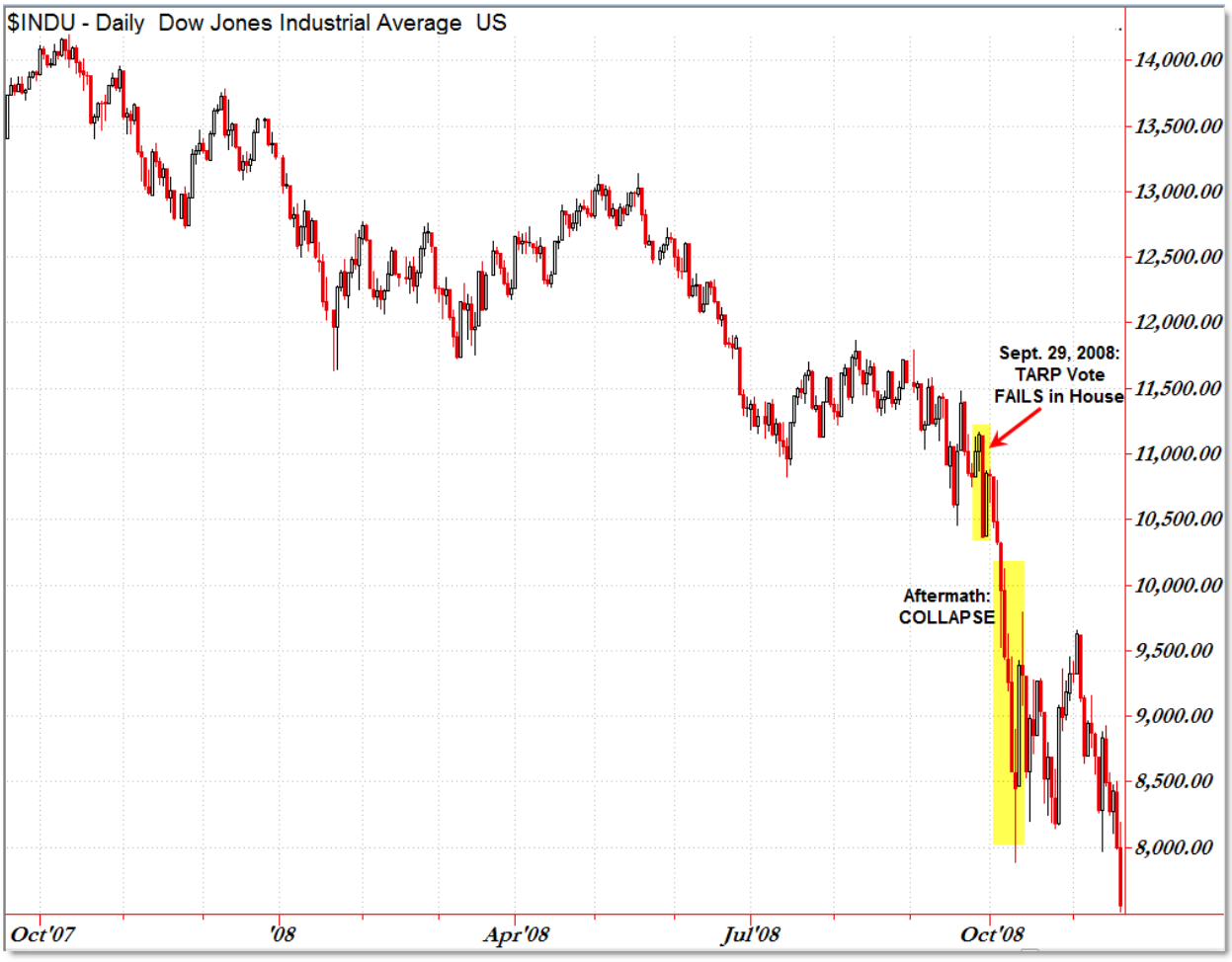
A self-contained course in the mathematical theory of statistics

for scientists & engineers with an emphasis on predictive regression modeling & financial applications.

Dr Tim Smith, Embry-Riddle Aeronautical University



Preface & Acknowledgments

This textbook is designed for a higher level undergraduate, perhaps even first year graduate, course for engineering or science students who are interested to gain knowledge of using data analysis to make predictive models. While there is no statistical perquisite knowledge required to read this book, due to the fact that the study is designed for the reader to truly understand the underlying theory rather than just learn how to read computer output, it would be best read with some familiarity of elementary statistics. The book is self-contained and the only true perquisite knowledge is a solid understanding of university level calculus, which of course it is expected that any engineering or science student will have mastery off. The intention for this textbook is for an elective type course; however, the foundations are laid here for further mathematical study and this text could well serve as a transition for an interested student with little to no prior knowledge to then go on to study in the popular fields of data scientist, big data or whatever the buzz words of the day may call it. A natural next read would be something equivalent to the popular texts “an introduction to statistical learning” or the “the elements of statistical learning,” by Hastie & Friedman et al.

The author is very grateful for the opportunity to have implemented and taught the MA 413 course at his current institution, Embry Riddle Aeronautical University in Daytona Beach, Fl. It was that course which led to me writing this textbook evolving for class notes, and I am very thankful to the many students who made corrections along the way. And, of course I am extremely grateful to my wife who supported me & encouraged me to labor on, hence I dedicate this book to her!

Table of contents

Ch 1 Introduction

- **1.1 "review of descriptive statistics" page 4**
- **1.2 "introduction to correlation & regression" page 11**

Ch 2 Overview of Probability Theory

- **2.1 "continuous random variables" page 20**
- **2.2 "introduction to density functions" page 23**
- **2.3 "expectation and variance" page 31**

Ch 3 General Linear Model

- **3.1 "foundational theory of hypothesis testing" page 38**
- **3.2 "introduction to the general linear model" page 45**
- **3.3 "one way ANOVA and the F test" page 49**
- **3.4 "single variable linear regression" page 57**
- **3.5 "examples of single variable regression" page 66**
- **3.6 "ANOVA error analysis for regression" page 70**
- **3.7 "multivariable linear regression" page 82**
- **3.9 "a brief introduction to model optimization" page 96**

Ch 4 Applications to financial modeling

- **4.1 "introduction and definition of volatility" page 103**
- **4.2 "a macro economic model & suggested further study" page 113**

1.Introduction

1.1 review of descriptive statistics

As usual when working with a data set, we use the notation x_i for the i^{th} data point in a data set. For example, if we were working with the following data set of airplane's speeds

Speed (mph)
500
700
900

Data set 1.1.1 (airplane's speeds)

we would call $x_1 = 500$ and $x_2 = 700$ etc. Of course, in the "real world" it is common to work with extremely large data sets so it becomes necessary to calculate the so called "descriptive statistics," which allow us to understand the various things a data set is telling us. These descriptive statistics are, generally speaking, divided into two categories: measures of central tendency OR measures of dispersion. The first category, measures of central tendency, attempts to simply describe the average value or middle of the data set; namely, a few examples of the measures of central tendency are the median and the mean as given in Def 1.1.1 & 1.1.2. The second category, measures of dispersion, attempts to describe how spread out the data set is; namely, a few examples of the measures of dispersion are the range and the variance as given in Def 1.1.3 & 1.1.4.

It is common that many resources will attempt to describe a data set by graphical illustration. Although these illustrations are useful, it is essential to remember that as scientists we cannot rely on graphical analyses to draw conclusions. Rather, we require formal analytical mathematical statements. For example we know that ***for a data set to be considered a normal distribution*** the data set must have ***most of the data frequency near the middle*** with a ***symmetrical pattern and the frequency should be less the further away from the middle***. Hence, if a Histogram is constructed it should look like this:

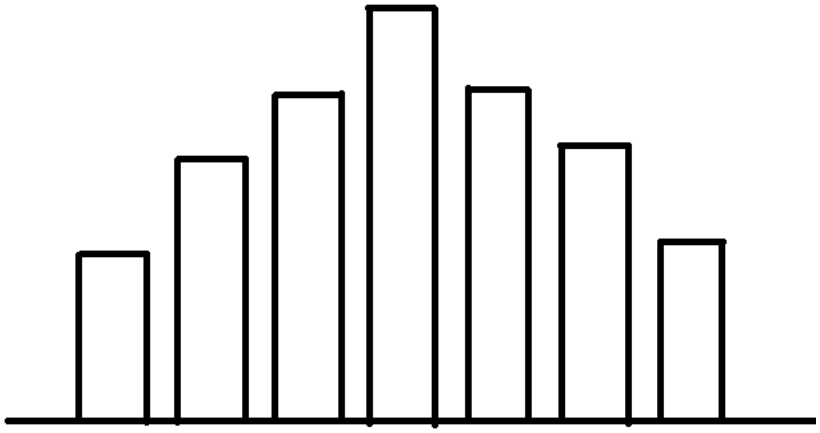


Figure 1. Histogram of a normal distribution

It is essential to understand that this graph alone does not prove nor reject the hypothesis that this distribution is normal. If one wanted to attempt to validate the hypothesis that this distribution is normal, then a formal "test of normality," including a formal computed analytical value to be compared to a formal analytical critical value, would be required. Prior to getting ahead of ourselves, let us summarize a few common descriptive statistics with proper analytical formulas.

Def 1.1.1 The ***mean*** (or arithmetic average) of a data set of n elements

$$X = \frac{1}{n}\sum x_i$$

Ex 1.1.1 Find the mean of data set 1

$$\bar{X} = \frac{1}{3}(500 + 700 + 900) = 700$$

Ex 1.1.2 Find the mean of data set 1.1.2

Speed (mph)
20
30
40

Data set 1.1.2 (Car speeds)

$$\bar{X} = \frac{1}{3}(20 + 30 + 40) = 30$$

Def 1.1.2 The ***median*** of a data set of n elements

$\tilde{X} =$ the middle value of the data set when ranked (low – high order),

NOTE: if there' a tie for middle, then the median is the average of the two.

Ex 1.1.3 Find the median of data set 1.1.1

Firstly, we must rank the data set, which in this case is already ordered, as

$X_1 = 500$ & $X_2 = 700$ & $X_3 = 900$.

Then, the median is simply found as the middle value. In this case

$\tilde{X} = 700.$

It is important to note that the measures of central tendency alone do not completely describe the data set under consideration. For example, if we compute the mean & median of both data sets 1.1.3A & 1.1.3B, then we will find the results to be the same as 50. However, it is obvious that the data sets are quite different; namely, the first data set is very clustered together while the second data set is much more spread out.

45
47
50
53
55

Data set 1.1.3A

30
35
50
65
70

Data set 1.1.3B

Thus, we will need to consider measures of dispersion in addition to finding the mean or median. Now, a small value of dispersion would imply that the data set is closely clustered together while a large value of dispersion would mean the data set is more spread out; hence we would expect data set 1.1.3A to have smaller measures of dispersion that data set 1.1.3B. This is indeed true as we will find the variance of 1.1.3A is 17, while the variance of 1.1.3B is 312.50.

Def 1.1.3 The ***range*** of a data set of n elements

= distance between the largest & smallest value of the data set when ranked

Ex 1.1.4 Find the range of data set 1.1.1

Firstly, we must rank the data set, which in this case is already ordered, as

X_1= 500 & X_2 = 700 & X_3 = 900.

Then, the range is simply the largest values minus the smallest

X3 – X1 = 900 – 500 = 400.

Def 1.1.4 The ***variance*** of a data set of n elements

$S^2 = = \frac{1}{n-1}\sum(x_i - \bar{X})^2$

Ex 1.1.5 Find the variance of data set 1.1.3B

First, we must find the mean which in this case is 50. Next, it helps to use the following table to simplify the procedure for computing our formula.

x_i	$(x_i - \bar{X})$	$(x_i - \bar{X})2$
45	-5	25
47	-3	9
50	0	0
53	3	9
55	5	25

If we sum the last column we will find the so called sum of the squares, $\sum(x_i - \bar{X})^2$ which in this example is 68. The variance is then found as this value divided by n-1. In this example we would divide by 4 to find the variance to be 68/4 = 17.

Def 1.1.5 The ***standard deviation*** of a data set of n elements

S = square root of variance

The standard deviation is essentially measuring the same thing as the variance did, however, by taking the square root we are bringing the measure back to the same dimension / units of the original data. For example if we computed the variance of data set 2 we would find it to be 100. However, this would actually be in units of mph^2 which may not be the most practical in applications, yet the standard deviation would be in units of just mph.

Ex 1.1.6 Find the standard deviation of data set 1.1.2

First, we must find the variance which as noted above is 100 mph^2. Thus, by definition, the standard deviation is the square root of variance =

$$\sqrt{(100)mph^2} = 10mph.$$

It is also very important to keep in mind “unit bias” when performing data analysis. For example if we were to compare our prior data set of car speeds

Speed (mph)
20
30
40

Data set 1.1.2 (Car speeds)

to our prior data set of plane speeds

Speed (mph)
500
700
900

Data set 1.1.1 (Aircraft speeds)

it should be apparent that a 200mph difference of speed in one context is quite different than a 200 mph difference of speed in another (have you ever been passed by another vehicle on the freeway going 200 mph faster?). This “unit bias” can be eliminated by

transforming the raw data to the “Z scores” which, as the next definition will outline, is done simply by dividing the individual data point’s deviation by the standard deviation. In fact, this standardization will show that the third car, which is going 10 mph above the mean of that data set, has the same “Z score” as the third plane, which is going 200 mph above the mean of that data set. Hence, one can infer that a 10mph deviation in car’s speed is essentially equivalent to a 200mph deviation in an airplane’s speed

Def 1.1.5 The ***Z score*** data point X_i from a data set

$$Z = \frac{X_i - mean}{st\ dev}$$

Ex 1.1.7 Compute the Z score for all data points in data sets 1.1.1 & 1.1.2.

x_i	$Z = \frac{x_i - \bar{X}}{s}$		x_i	$Z = \frac{x_i - \bar{X}}{s}$
20	-1		500	-1
30	0		700	0
40	1		900	1

1.2 introduction to correlation & regression

To conclude this introductory chapter, in this section we will briefly introduce the idea of correlation between two data sets x and y which we will assume both contain an equal number of elements, namely n elements in each data set. The main idea with correlation, or perhaps the main question to ask, is this: is there a pattern between the

two data sets? A common mistake that can be made is thinking that the only way to have a correlation between data set x and data set y is that the pattern between x and y must be linear, perhaps y = 2x or y =3x etc. However, this is not correct as there are many other correlation patterns which can occur between two data sets such as quadratic fits or exponential fits. Later on in the textbook we will study the concept of building a predictive model from an x,y data set pair known as a linear regression model, and this tool is one of the most widely applicable statistical models. Of course the linear regression model is a primary purpose of our study and for students of engineering or the sciences this linear predictive data model can be extremely useful. But, it is important to note that linear fits are not the only fit and just because data is correlated does not mean that a linear regression model will work. In mathematical terms one might say that a solid value of correlation is a necessary condition for a linear regression model to work, but it is not a sufficient condition!

The correlation between two data sets is defined in terms of the deviation between the Z scores of the x data set and the Z scores of the y data set. No deviation between the data set's Z scores is defined as perfect correlation, while an extreme amount of deviation between the data set's Z scores is defined as a low correlation or near zero correlation. For example, if we were to revisit example 1.1.5 and call the data set of car's speeds the x data set and the data set of airplane's speeds the y data set and then compute the differences of those Z scores, we would essentially be studying the correlation pattern between the cars and airplanes. It is worthy to note, prior to working out the details of this example, that in this case we expect a perfect correlation

as we have previously discovered that the car's speeds and airplane's speeds go up in a uniform pattern of 1 standard deviation each data point.

Ex 2.1.1 Compute the difference between the Z score for data points sets 1.1.1 & 1.1.2.

x_i	$Z_{xi} = \frac{x_i - \bar{X}}{s}$		y_i	$Z_{yi} = \frac{y_i - \bar{Y}}{s}$
20	-1		500	-1
30	0		700	0
40	1		900	1

To begin we recall from Ex 1.1.7 the Z score which we previously computed and then label accordingly as done above. To complete this example we then simply compute the differences

$Z_{xi} = \frac{x_i - \bar{X}}{s}$	$Z_{yi} = \frac{y_i - \bar{Y}}{s}$	Differences=$Z_{xi} - Z_{yi}$
-1	-1	0
0	0	0
1	1	0

We observed, as expected, that the total differences are zero which shows a perfect correlation between our data sets.

Now, for a formal definition of correlation we use the following definition which has the interpretation similar to a percent: r near 1 is near perfect correlation (analogous to 100% being near perfect chance) while r near 0 is low correlation (analogous to 0% being near no chance).

Def 1.2.1 The ***correlation*** between a data pair set x and y both of n elements is

$$r = 1 - \frac{1}{2(n-1)} \sum_{i=1}^{n} (Z_{xi} - Z_{yi})^2$$

Ex 2.1.2 Compute the correlation for data pairs from data sets 1.1.1 & 1.1.22.

To begin we recall from Ex 2.1.1 the differences in the Z score which we previously computed and then we must compute the squared differences

$Z_{xi} = \frac{x_i - \bar{X}}{s}$	$Z_{yi} = \frac{y_i - \bar{Y}}{s}$	Differences=$Z_{xi} - Z_{yi}$	$(Z_{xi} - Z_{yi})^2$
-1	-1	0	0^2
0	0	0	0^2
1	1	0	0^2

Now, to complete the problem we utilize the definition

$$r = 1 - \frac{1}{2(n-1)} \sum_{i=1}^{n} (Z_{xi} - Z_{yi})^2$$

with n= 3. Doing so this yields the solution

$$r = 1 - \frac{1}{2(3-1)}(0^2 + 0^2 + 0^2) = 1 - 0 = 1$$

We observed, as expected, that the correlation here is a perfect correlation = 1, which again we can informally view analogously to a percent so in an informal sense we can think this data set is 100% correlated.

Ex 2.1.3 Compute the correlation for data pairs from the data set 1.2.1 below.

X	y
1	2
2	4
3	6
4	7
5	11

Data set 1.2.1

Where it is given that the mean of x is 3 and y is 6, while the standard deviation of x is 1.58 and of y is 3.39.

To begin we must compute the Z scores in for x and y separately

$Z_{xi} = \frac{x_i - \bar{X}}{s} = \frac{x_i - 3}{1.58}$	$Z_{yi} = \frac{y_i - \bar{Y}}{s} = \frac{y_i - 6}{3.39}$
$\frac{1-3}{1.58}$	$\frac{2-6}{3.39}$

$\frac{2-3}{1.58}$	$\frac{4-6}{3.39}$
$\frac{3-3}{1.58}$	$\frac{6-6}{3.39}$
$\frac{4-3}{1.58}$	$\frac{7-6}{3.39}$
$\frac{5-3}{1.58}$	$\frac{11-6}{3.39}$

Now, the differences in the Z scores must be computed and then their squares

$Z_{xi} = \frac{x_i - 3}{1.58}$	$Z_{yi} = \frac{y_i - 6}{3.39}$	Differences=$Z_{xi} - Z_{yi}$	$(Z_{xi} - Z_{yi})^2$
-1.26582	-1.17994	-0.08588	0.007376
-0.63291	-0.58997	-0.04294	0.001844
0	0	0	0
0.632911	0.294985	0.337926	0.114194
1.265823	1.474926	-0.2091	0.043724

Finally, to complete the problem we utilize the definition

$$r = 1 - \frac{1}{2(n-1)} \sum_{i=1}^{n} \left(Z_{xi} - Z_{yi}\right)^2$$

with n= 5. Doing so this yields the solution

$$r = 1 - \frac{1}{2(5-1)} (0.007376 + 0.001844 + \cdots) = 1 - 0.02 = 0.98$$

We observed, as expected, that the correlation here is a very high correlation = 0.98 as expected since in the data set we can observe the pattern Y being

approximately 2x. Again we can informally view analogously to a percent so in an informal sense we can think this data set is 90% correlated

It is worthy to note that while the prior definition is the theoretically correct and the original definition it is not always the commonly used one. By putting in the definitions of Z scores and performing some algebraic manipulation the following alternate definition for correlation can be obtained and is useful since it is all in terms of values from the data set, hence it is not needed to first compute the Z scores. Also, for mathematical interest the correlation can be written as the covariance of X and Y divided by the products of their standard deviations. Namely, we can write $r = \frac{cov(x,y)}{s_x s_y}$

Def 1.2.2 The ***correlation*** between a data set x and set y both of n elements can

$$r = \frac{\sum(X - x_i)\left((Y - y_i)\right)}{\sqrt{\{\sum(X - x_i)^2\}\{\sum(Y - y_i)^2\}}}$$

The above definition will yield the exact same value as the correlation definition provided in the prior definition 1.2.1, and is actually derived from that prior definition, but this new formula is often preferred when coding formulas as it can be computed directly from raw data as opposed to needing the normalized "z values."

One of the most useful applications of correlation for an (x,y) pair data set is to build a predictive model to predict the y variable in terms of the x variable as input. Namely, it is desired to create an equation of the form $\hat{y} = mx + b$, where the hat notation is utilized to distinguish it as being a predicted value perhaps a future or forward data point.

Def 1.2.3 The ***linear regression line*** of a data set x and set y is

$$\hat{y} = mx + \beta,$$

where

$$m = r\left(\frac{s_y}{s_x}\right)$$

and

$$\beta = \bar{y} - m\bar{x}.$$

Ex 2.1.4 Compute the linear regression line for data pairs from the data set 1.2.1.

To begin we recall from the prior solution that

$$s_y = 1.52, s_x = 0.71, \bar{y} = 6, \bar{x} = 3 \; and \; r = 0.9$$

Hence, we can compute

$$m = r\left(\frac{s_y}{s_x}\right) = 0.9\left(\frac{1.52}{0.71}\right) = 1.93$$

and

$$\beta = \bar{y} - m\bar{x} = 6 - 1.93 * 3 = 0.21$$

Which yields our solution as the predictive model of our linear regression line

$$\hat{y} = 1.93x + 0.21.$$

The linear regression line has far reaching applications in various fields such as engineering or science and finance applications. For example, our data ended at the

value of x being 5 so one could use the linear regression line to expand beyond that, perhaps to find the predicted y value associated with a future x of 6 as

$$\hat{y}(6) = 1.93(6) + 0.21 = 11.79.$$

For another example, our data set contained only integer values of x being 1 then x being 2 etc., and one could use the linear regression line to fill in between that, perhaps to find the predicted y value associated for a half way value x of 1.5 as

$$\hat{y}(1.5) = 1.93(1.5) + 0.21 = 3.11.$$

There are many other applications, and of course restrictions to, the linear regression line and this will be a central theme of the later chapters of this textbook. However, prior to continuing with our development it is necessary to overview, or perhaps receive for the informed reader, some key principles from the mathematical theory of probability which are contained in Chapter 2. Due to the fact that this text is designed as a self-contained resource, these principals in Chapter 2 are developed "from the ground up" and the informed reader may be able to jump forward at this point. Any reader who has the knowledge equivalent to a Junior or Senior year university level course in mathematical statistics and/or introductory probability theory can most likely jump to Chapter 3. Regardless of that progression, it is important to close this Chapter with one essential principal regarding regression: while it is logical that it only makes sense to use a linear regression model for a data set which is highly correlated, that does not ensure that the linear regression model will work or be statistically valid. Namely, it is vital to understand, as previously stated, that from a mathematical point of

view one can say that a solid value of correlation is a necessary condition for a linear regression model to work, but it is not a sufficient condition!

2 Overview of Probability Theory

2.1 continuous random variables

When studying probability theory it is very important to consider the perspective we have when investigating problems. As engineers or scientists, it is expected to have solution values that predict exactly when or exactly where some event will occur, i.e. deterministic solutions. However, in probability we do not have such solutions or problems rather we define the likelihood of outcomes. This begins with how we define our variables; namely we define a **random variable**(RV) as a number whose value depends on the outcome of a random experiment. The key point here is that the outcome of the experiment are random not deterministic. A good example is the lottery: the odds say it is extremely unlikely to win but that does not mean you will not win as until the experiment of the lottery numbers being drawn is conducted we do not know the outcome.

As is well known, there are, generally speaking, two “kinds” of variables: discrete variables and continuous variables. One of the simplest illustrations to demonstrate the difference between these two kinds of variables can be illustrated from a typical classroom situation; namely, the number of students in the class is a discrete variable while the time of the class is a continuous variable. A **discrete variable** is one that is finite and countable. For example, no matter how large the class is the number of students is countable! We also notice that the number of students is a finite discrete value, identified by a positive integer, as you can think when new students enter the class there is either 1 student or 2 students, but no in between value such as a half of a

student. On the other hand a **continuous variable** is one that is infinite. For example time is an infinite continuum. A student who is studying theoretical physics will be very interested to dialog about the matter of time as a variable and observable measurements etc. However, for simplification of the idea let us just look at one interesting property of continuous real numbers. There is a famous mathematical axiom that states between any two real numbers there is always at least one more value, hence, any interval on the real number line contains an infinite number of values. Now, this idea can be illustrated by just considering two moments in time, let us say a starting time of t=1 second and an end time of t= 2 seconds, and in doing so we see that there is a halfway point

$$X = 1.5 = \tfrac{1}{2}(1 + 2)$$

Repeating this process using the original starting time of t=1 second but a new end time of t = 1.5 seconds we see that there is a new halfway point

$$X = 1.25 = \tfrac{1}{2}(1 + 1.5).$$

Repeating this process once more using again the original starting time of t=1 second but a new end time of t = 1.25 seconds we see that there is a new halfway point

$$X = 1.125 = \tfrac{1}{2}(1 + 1.25).$$

As you can see, this process could go on indefinitely, hence proving that between any two values of a continuous variable there are infinitely many points.

The prior result is very interesting from a pure mathematical number theoretic point of view alone, but it also yields one very interesting probability result for us to take

note of, namely that the probability of our RV being any one single value is exactly equal to zero. While this will be developed more formally later on, once we formally define density functions and probability integrals, we can see the idea as follows using the classical probability definition of probability

$$P(A) = \frac{size\ of\ (\ AKA\ number\ of\ elements\ in\ )\ sample\ A}{size\ of\ sample\ space\ \Omega} = \frac{1}{\infty} = 0.$$

2.2 introduction to density functions

In the following section we will define one of the most important topics in the mathematical theory of probability, the probability density function. It is from this density function that many results such as probability solutions and expected values will be derived. However, prior to doing so, it is important to note that for simplification through the remainder of this text, which is designed to for a first course in probability theory, we will only be considering examples of independent continuous random variables as outlined in the following definition (NOTE: a summary of some examples & main definitions for discrete random variables are provided in the appendix for either faculty who may desire to include such examples to their course and/or interested reader desiring to study further)

Def 2.1.1 We say that two random variables X and Y are independent if the events are independent events for every pair of intervals $A<X\leq B$ and $C<Y\leq D$.

This is equivalent to saying that

$P(A < X \leq B \ \& \ C < Y \leq D) = P(A < X \leq B) \bullet P(C < Y \leq D)$.

At this level we will not attempt to develop the derivation or underlying motivation for our probability density functions; rather, we will define a density function f(x) for a random variable X to be the function that creates the probability as

$$P(A < x \leq B) = \int_{A}^{B} f(x)dx.$$

Now, this probability density function must meet two basic properties, which are in line with the axioms of probability:

Def 2.1.2 In order for a function to be a density of a RV X it must satisfy

(i) $f(x) \geq 0$

and

(ii) $\int_{-\infty}^{+\infty} f(x)dx = 1.$

NOTE: if you have an example of a function desired to be used for a density that meets criteria (i) but not (ii) then you can create a valid density by the normalization process (similar to that of normalizing a vector) by dividing the constant $K = \int_{-\infty}^{+\infty} f(x)dx$, as one will find the function

$$\frac{1}{K} f(x)$$

will be a valid density function.

Ex 2.1.1 Find the normalized density for a density of the form e^{-Ax} defined for x > 0 and zero elsewhere.

To begin we note that we must satisfy properties (i) and (ii) of definition 2.12. Now, it is first observed that property (i) is met because the exponential function is a strictly positive function. However, property (ii) is not met because $\int_{-\infty}^{+\infty} f(x)dx = 1/A$. Thus, using the logic from above we take the constant $K = \frac{1}{A}$ to create the normalized density to be the so called ***exponential density*** $f(x) = Ae^{-Ax}$.

Ex 2.1.2 Find the value of C so that the function

$$\begin{matrix} Cx(1-X) \\ 0 \end{matrix} \quad \text{if} \quad \begin{matrix} 0 < x < 1 \\ else \end{matrix}$$

will be a valid density.

To begin we note that we must satisfy properties (i) and (ii) of definition 2.12. Now, it is first observed that property (i) is met because this parabolic function will be above the x axis, with zeros at x=1 and x=0, provided that the value of C is positive. Now, property (ii) is not met until we specify the value of C, thus we compute

$$\int_{-\infty}^{+\infty} f(x)\,dx = \int_{-\infty}^{0} 0\,dx + \int_{0}^{1} Cx(1-x)\,dx + \int_{1}^{+\infty} 0\,dx = C\left(\frac{1}{2} - \frac{1}{3}\right) = \frac{C}{6}.$$

Now, in order to make this a valid density we must choose C=6.

Ex 2.1.3 Verify that the ***uniform density***

$$f(x) = \begin{matrix} 0 \\ \dfrac{1}{R-L} \\ 0 \end{matrix} \quad \text{if} \quad \begin{matrix} x < L \\ L < x < R \\ x > R \end{matrix}$$

is a valid probability density function.

To begin we note that we must satisfy properties (i) and (ii) of definition 2.12. Now, it is observed that property (i) is met because the function is a constant positive value. In verifying property (ii), we obtain

$$\int_{-\infty}^{+\infty} f(x)\,dx = \int_{-\infty}^{L} 0\,dx + \int_{L}^{R} \frac{1}{R-L}\,dx + \int_{R}^{+\infty} 0\,dx = \frac{R-L}{R-L} = 1.$$

Ex 2.1.4 Verify that the ***standard normal density***

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$$

is a valid probability density function.

To begin we note that we must satisfy properties (i) and (ii) of definition 2.12. Now, it is first observed that property (i) is met as the exponential function is a strictly positive function. However, we must verify property (ii), and in doing so we obtain

$$\int_{-\infty}^{+\infty} f(x)\,dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}\,dx\,.$$

This is a very difficult integral to do in closed form (IF YOU DO SO, THEN PLEASE CONTACT THE PRESIDENT OF YOUR UNIVERSITY) but one can compute numerically and verify that this integral is indeed equal to 1, hence the provided density is a valid probability density function!

It is worthy to take note of a few of these common distributions as they will frequently be used in examples as we move forward and have many common real world applications! The following are the most likely examples that you will encounter are:

The ***exponential density*** is $f(x) = Ae^{-Ax}$ which is defined for x > 0,

and the ***uniform density*** is $f(x) = \frac{1}{R-L}$ which is defined for L < x < R,

where both of these densities serve useful for textbook illustrative examples due to the fact that the resulting integrals turn out to be doable without the need for complicated integration techniques.

The ***normal density*** is $f(x)=\frac{1}{o\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma^2}}$ which is defined for all x, and this density is by far one of the most applicable in real world modeling applications.

The ***standard normal density*** is $f(x)=\frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}$ which is defined for all x, and is a special case of the normal density with mean μ=0 along with variance $\boldsymbol{\sigma}=1$, serves as the backbone of many theoretical mathematical statistical results such as the famous central limit theorem.

The ***T density*** is $f(x)=\frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\Gamma\left(\frac{v}{2}\right)}\left(1+\frac{x^2}{v}\right)^{-\frac{v+1}{2}}$ which is defined for x > 0 with v being the degrees of freedom. This density is utilized in applications as an approximation for the normal density when some of the information of the population mean, μ, or variance, $\boldsymbol{\sigma}^2$, is unknown.

The ***chi squared density*** is $f(x)=\frac{1}{2^{\frac{v}{2}}\Gamma\left(\frac{v}{2}\right)}x^{\left(\frac{k}{2}\right)-1}e^{-\frac{x}{2}}$ which is defined for x > 0, with v being the degrees of freedom. This density is utilized in applications for error analysis when considering the sum of squares error and/or Goodness of fit error analysis.

The ***F density*** is $f(x)=\frac{\Gamma\left(\frac{d_1+d_2}{2}\right)}{x\Gamma\left(\frac{d_1}{2}\right)\Gamma\left(\frac{d_2}{2}\right)}\sqrt{\frac{(d_1x)^{d_1}d_2^{d_2}}{(d_1x+d_2)^{d_1+d_2}}}$ which is defined for x > 0, where d_1 and d_2 are the degrees of freedom, numerator and denominator respectively. This

density is related to a ratio of two chi squared densities, and is very useful in a great deal of applications. especially the analysis of linear regression.

The ***logistic density*** is $f(x) = \frac{e^{-(x-\mu)/s}}{s\left(1+e^{-\frac{x-\mu}{s}}\right)^2}$ which is defined for x > 0, with s representing the scale not standard deviation as one might expect. This density is very useful in the analysis of regression when applied to case when the response variable in the form of a categorical "1/0" variable (AKA logistic regression).

Some more generalized "abstract" examples are:

The ***Gamma density*** is $f(x) = \frac{a^b}{\Gamma(b)} x^{b-1} e^{-ax}$ which is defined for x > 0 and will only converge for both $a > 0, b > 0$.

The ***Beta density*** is $f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$ which is defined for x > 0.

The factor Γ is the so called "Gamma function," which normalizes the density. There is a formal definition of this function valid for any values of n, but for our purposes it will suffice to use the definition:

$$\Gamma(n) = (n-1)! \text{ for integer values}$$

$$\Gamma\left(\frac{n}{2}\right) = \sqrt{\pi}\frac{(n-2)!!}{2^{(n-1)/2}} \text{ for halves using odd n; note } (n-2)!! = (n-2)\cdot(n-4)\cdots 3\cdot 1$$

Now, we have several probability density functions let us look at some examples

Ex 2.1.5 For the ***exponential density***

$$f(x) = e^{-x}$$

Find the probability P(0<x<5).

To begin we know the density is as given above so we just need the probably integral $P(0 < x < 5) = \int_0^5 f(x)\,dx$.

Thus, we compute

$$\int_0^5 e^{-x}\,dx = \left[-e^{-x}\right]_{x=0}^{x=5} = 1 - e^{-5} \approx 0.993$$

Hence, we have computed the probability P (0<x<5) = 93%.

Ex 2.1.6 For the ***uniform density***

$$f(x) = \begin{matrix} 0 \\ \frac{1}{R-L} \\ 0 \end{matrix} \quad \text{if} \quad \begin{matrix} x < R \\ R < x < L \\ x > L \end{matrix}$$

with R = 10 and L = 0 find the probability P (0<x<2).

To begin we note that our density will be $f(x) = \frac{1}{10}$ and the probability integral will be $P(0 < x < 2) = \int_0^2 f(x)\,dx$.

Thus, we compute

$$\int_0^2 \frac{1}{10}dx = \left[\frac{x}{10}\right]_{x=0}^{x=2} = 0.2$$

Hence, we have computed the probability P(0<x<2) = 20%.

Ex 2.1.7 For the ***standard normal density***

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$$

find the probability P(0<x<2).

To begin we know the density is as given above so we just need the probably integral $P(0 < x < 2) = \int_0^2 f(x)dx$.

Thus, we compute

$$\int_0^2 \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx.$$

However, this integral, which is ultimately an integral of the form e^{u^2}, is not solvable in closed form so numerical approximations will be required (which yield the solution of approximately 47%) . In the next chapter we will further discuss how to work with normal density, as it is one of the most important densities if not the most important, and we will look at some applications of nice function in MATLAB. For now, we will move on with developing further properties of probability distribution theory, namely the expected value and variance.

2.3 expectation and variance

In the following section we will define two extremely useful properties of statistics the expectation and the variance. Generally speaking one can view these in an analogous manner as the expected value and variance are interpreted in elementary data analysis. Namely, the expectation (AKA expected value) can be viewed as the average value or “what we expect to get on average,” which is frequently just called the mean and often the symbol μ is utilized. And, the variance can be viewed as a measure of dispersion or “how spread out is the data,” which is often notated by the symbol $\boldsymbol{\sigma}^2$. For simplification we will define, for a random variable x, the expectation as E(x) and moving forward write all expressions, definitions and so forth in terms of E(x) as not only is it good practice for consistency, but it is also the proper and formal way to define things!

Def 2.2.1 The ***expectation*** of a continuous random variable X with density f(x) is

$$E(x) = \int_{\Omega} x \cdot f(x)\, dx$$

At this time we will focus on solving examples and address interpretations along with theoretical implications for later studies. However, it is good for the reader to understand the solution obtained is an expected value and not a probability, i.e. it does not have to be within the usual range of 0 to 1 rather the answer can be viewed as just a number!

Ex 2.2.1 Find the expectation for the ***Standard Normal density***

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$$

To begin we recall the above definition of the expectation is .

$$E(x) = \int_{\Omega} x \cdot f(x)\, dx$$

and we compute

$$E(x) = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}\, dx = \left[\frac{-1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \right]_{x=-\infty}^{\infty} = 0.$$

Ex 2.2.2 Find the expectation for the particular case of the ***Beta density***

$$\begin{matrix} 6x(1-X) \\ 0 \end{matrix} \quad \text{if} \quad \begin{matrix} 0 < x < 1 \\ else \end{matrix}$$

To begin we recall the above definition of the expectation is

$$E(x) = \int_{\Omega} x \cdot f(x) dx$$

and we compute

$$E(x) = \int_0^1 x(6x(1-x))dx = \frac{1}{2}.$$

Def 2.2.2 The ***variance*** of a continuous random variable X with density f(x) is

$$VAR(x) = \int_{\Omega} (x-\mu)^2 \cdot f(x) dx$$

where the symbol μ is representing the value of the expectation for the density, as often the expectation is interpreted as a mean or average value. Again, at this time we will quickly observe solving an example and leave interpretations along with theoretical implications for later studies.

Ex 2.2.3 Find the variance for the ***Standard Normal density***

$$f(x)=\frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}$$

To begin we recall the above definition of the expectation is

$$VAR(x)=\int_{\Omega}(x-\mu)^2\cdot f(x)dx$$

Thus, we compute

$$VAR(x)=\int_{-\infty}^{\infty}(x-0)^2\frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}dx=1.$$

For the purpose of this textbook study, the preceding definitions will suffice to cover all forthcoming needed mathematical theory. However, we will close this section with the following definitions and results as they can be very useful in applications to actually compute the expectation and/or variance for a density without conducting the integrals from the prior definitions.

Def 2.2.3 The ***moment generating function*** for density f(x) is

$$MGF=E(e^{tx})=\int_{\Omega}e^{tx}\cdot f(x)dx$$

where the result is actually a function of t. This MGF can be very useful, namely we can find the expectation and variance, by constructing the so called moments.

Def 2.2.4 The ***n'th moment*** of a density f(x) is

$$\mu_n = E(x^n) = \frac{d^n}{dt^n}\{MGF]_{t=0}.$$

This result can be extremely useful as one can prove that the expectation is equal to μ_1 while the variance is equal to $\mu_2 - (\mu_1)^2$.

3 General Linear Model

3.1 foundational theory of hypothesis testing

In the prior chapter the general theory of probability was summarized along with the main concepts of probability density functions, cumulative distributions and moment generating functions etc. Now, in this chapter we will embark on the study of one of the most powerful and important real world applications of mathematics: the theory of hypothesis testing! The general idea of hypothesis testing can be summarized as the process of obtaining some data from an experiment and then using probability theory to attempt to validate a claim. Moving forward, we will refer to the claim as the hypothesis and generally speaking the experiment will involve the implementation of something that wasn't utilized in the past which we will refer to as the treatment. While the idea will not be discussed in detail here many instructors effectively teach hypothesis testing through the parallel logic of a court case. Namely, in a court case the defendant is assumed innocent until proven guilty beyond a reasonable amount of doubt as decided by a jury. Likewise, in hypothesis testing we desire to validate our claim the hypothesis, but we will take the stance that it is not valid (AKA assumed innocent) until proven otherwise beyond a mathematical amount of certainty (AKA beyond reasonable doubt).

In general the hypothesis procedure will consist of 4 steps

First, the hypothesis is made as a mathematical statement.

Second, the so called "critical value" and "rejection region" are defined.

Third, calculation of the test statistic

Fourth, conclusions are stated.

In the following derivation, we will assume that the hypothesis is being studied on the simple difference on a population mean after the application of a treatment. Namely, we will consider the so called “null hypothesis” as μ = population mean value as given. The idea of this hypothesis statement is that the symbol μ is in a sense representing the population mean moving forward in time with the treatment applied consistently in the future (i.e. this statement is saying that the mean does not change when the treatment is applied). In the same manner that a defendant is assumed innocent in a court case until proven otherwise, we will assume this null hypothesis is truthful until proven otherwise.

The null hypothesis will be rejected if our soon to be defined test statistics falls outside our mathematical region which is defined from our chosen level of statistical certainty. Namely, if we define our statistical certainty to be at a level of $(1-\alpha)\%$ then the critical value z_α (AKA endpoints) of our mathematical region can be found from the probability statement: $P\left(-z_\alpha < X < z_\alpha\right) = 1-\alpha$. For example, if we take a 95% confidence level the critical value will solve the equation

$$\int_{-z_\alpha}^{z_\alpha} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1-\alpha = 0.95$$

This example will yield the solution of $z_\alpha = 1.96$ which is a very important, if not “famous,” critical value and very much worth remembering! It is worthy to note that many authors will use the notation $z_{\alpha/2}$ due to the fact that this example is an illustration of a so called “two tailed test.” A two tailed test is one where the allowable error is

allowed either above the critical value or below the negative of the critical value, hence the error is split in half. A one tailed test it where the error is not split in half, hence only outside of the critical value in “one tail.” For now, we will restrict our study to the two tailed examples for simplification.

Looking back to the original claim, we see that we have defined two regions: the range within the region is where we expect things to be and the range outside of the region, which is to be viewed as an oddity. Thus, we can define the region outside to be the region to reject the null hypothesis. Namely, if our soon to be defined test statistic is either greater than z_α (or less than -z_α) we will reject the null hypothesis. Or, in a cleaner mathematical statement we can say:

Reject the null if || test stat || > z_α.

Now, the only missing point is the so called test statistic. We will formally define and prove where this value comes from shortly but let us first accept the definition so that we can view a few examples to illustrate this process of hypothesis testing.

Def 3.1.1 The ***Test statistic*** for a single sample hypothesis test of differences of mean is given by

$$\text{TS} = \frac{X - \mu}{\sigma / \sqrt{n}}$$

Ex 3.1.1 A drug manufacturer wants to test to see if a drug has an effect on rat’s speed running through mazes. A sample of 81 rats is given this drug, and their average speed is found to be 71 miles per day. The population’s average

without this drug is 73 with a standard deviation of 21 miles per day. Perform an appropriate hypothesis test.

To begin solving, we recall that a full solution to a hypothesis test problem has four steps

First, the hypothesis is made as a mathematical statement.

Second, the so called “critical value” and “rejection region” are defined.

Third, calculation of the test statistic

Fourth, conclusions are stated.

For our present example, we will assume that the level of confidence is 95% and the test is a two tailed test (which would make sense as the researcher wanted to unbiasedly test for an effect on speed rather than specifically test for an increase). So, we already know the second step is with 1.96 being the critical value. Hence, all we really need to compute are the 1^{st} and 3^{rd} steps. To begin, we must define the desired hypothesis, which is what we really want to show and is often referred to as the alternate hypothesis. In this example, the researcher knows that the population mean is 73 and they are attempting to see if the drug has an effect on that speed. Thus, we set the alternate hypothesis to state that μ is different than 73. Then, we also must construct the null hypothesis (the logical opposite of the alternate hypothesis) which in this case will state that μ is equal to 73. Now, all that remains is to compute the test statistic from our formula and then use our results to conclude.

In doing so, we obtain the for step solution as:

First, null H: $\mu = 73$

Alt H: $\mu \neq 73$.

Second, assume the null is truthful and reject if $|| TS || > 1.96$.

Third, $TS = \dfrac{71-73}{21/\sqrt{81}} = -0.857$.

Fourth, since the TS does not fall in the rejection region we fail to reject the null.

It is very important to note in this example that the result is just simply failure to reject the null hypothesis. This wording is very important, and it is essential to understand that this conclusion does not disprove anything, nor do we accept anything, rather we have just failed to reject the null hypothesis. Perhaps, one will find it useful to think that we have attempted to do something and failed to do so. Hence, our conclusion is that we did not do anything, or perhaps a more sophisticated way it to say we have “no conclusion!” Analogously, when a jury is tasked to find a defendant guilty beyond a reasonable doubt, if they do not find the evidence, then their formal result is to say “not guilty” or “no we did not find sufficient evidence.”

Ex 3.1.2 An instructor wants to see if group activity work increase test scores. Currently the school’s average math score is 85 with a standard deviation of 4. A sample of 36 students are assigned to do group work in class. Their average is 90.

Perform an appropriate hypothesis test.

Again, we note that a full solution to a hypothesis test problem has four steps, and to begin our present example we observe that the desired hypothesis is specifically to

increase the test scores. It is known that the population mean score is 85, so the logical choice for the Alt H is : μ > 85. As in our last example, we will assume that the level of confidence is 95%, but this is a one tailed test. Therefore, the prior critical value of 1.96 would not be the correct critical value. To find the correct value we would need to go back to our probability density theory. In doing so, we obtain the desired equation to solve

$$\int_{-\infty}^{z_\alpha} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1 - \alpha = 0.95$$

which will yield the solution of $z_\alpha = 1.65.$ We are now prepared to fully develop our hypothesis testing procedure:

First, null H: μ = 85

Alt H: μ > 85.

Second, assume the null is truthful, and reject if: TS > 1.65.

Third, TS = $\frac{90 - 85}{4/\sqrt{36}} = 7.5$.

Fourth, since the TS does fall in the rejection region we reject the null.

We previously presented the definition of the test statistic formula without development nor proof. Let us now formally define and prove from where this formula comes. We assume that the population problem we are studying is modeled by a normal distribution

with mean μ and standard deviation $\boldsymbol{\sigma}$, hence $X \sim N(\mu, \boldsymbol{\sigma})$. Now, in regards to the sample we will need to utilize two Lemmas from a theorem of advanced probability theory known as the Central Limit Theorem. Namely, we will take as definition the following:

Def 3.1.2 If $X_1, X_2, \ldots, X_n$ are random variables with

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$S = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

and if $\bar{X}$ is the random variable of the sample means of all the simple random sample size n from a population with expected value E(X), and variance Var(X) then

$$E\left(\bar{X}\right) = E(X)$$
$$Var\left(\bar{X}\right) = \frac{1}{n}Var(X).$$

We now need to prove our main foundational result, which is illustrated in the following theorem definition.

Def 3.1.3 (theorem) If $X_1, X_2, \ldots, X_n$ are normally distributed random variables with mean μ and standard deviation $\boldsymbol{\sigma}$, then

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} . \sim N(0,1)$$

Proof: Let us begin by recalling a fact about random variables, namely if $X \sim N(\mu, \sigma)$ and we consider the RV =aX, where a is a fixed constant, then we can show by some

routine algebra on the cumulative distribution function that this $RV \sim N(a\mu, a\sigma)$. A similar result is well known that if we have two random variables $X \sim N(\mu, \sigma)$ and $Y \sim N(\nu, \varphi)$, and if we consider the RV = X + Y, then we will find $X + Y \sim N(\mu + \nu, \sigma + \varphi)$. Thus, we can draw the conclusion that our $X_1, X_2, \ldots, X_n$ are normally distributed random variables that $X \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

Now, we shall look at the expression

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}. = \frac{\sqrt{n}}{\sigma}\bar{X} - \frac{\sqrt{n}}{\sigma}\mu.$$

From the results above we will see that this has a mean 0 and standard deviation 1, hence we have proved that $\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}. \sim N(0,1).$, which means the use of the standard normal distribution for our critical values are validated.

3.2 Introduction to the general linear model

In the prior section the general mathematical theory of probability, which we developed in Chapter 2, was extended and applied to a problem in the so called field of inferential statistics, namely the hypothesis testing method as illustrated through examples. While it is very important to both understand the mathematical theory of probability and be able to use it to justify models, designs, & results etc., moving forward we will not necessarily prove every result at that depth. Moreover, for the

remainder of this text we will be looking specifically at applications of the so called general linear model, and while one could prove all details at the level of mathematical probability theory we will not do so here as it would require quite an amount of advanced perquisite mathematical knowledge from higher level matrix theory which is not usually studied until the graduate level. However, it is strongly emphasized that any application of the general linear model must be checked to assure that the distribution of the errors is fitting a normal distribution! In addition, while we develop our models through examples in the next few sections we take the time to note which mathematical density the model is following as this is a crucial mathematical tool to know when formally justifying hypothesis conclusions through the use of critical values.

The general linear model is a statistical model which can be written in a mathematical statement as

$$Y = XB + e$$

where Y is a matrix of response measurements, X is the so called design matrix which is commonly a matrix of input variable measurements, and B is a matrix of parameters which are to be determined. The term e is a matrix generally containing errors, and it is this term which defines our equation to be viewed as a statistical model - we expect and allow some error! Again, the assumption we make and require is that these errors follow a normal distribution.

It is very important to understand that this general linear model is the model that, depending on how the design is set up, will lead to all of our specific models moving forward in the remaining sections of this chapter. Namely, both ANOVA and multiple

variable regression are particular cases of this general linear model. For example if Y is taken as a vector of n measurements with components labeled as y_i for i=1 to n, and B is also taken to be vector of coefficients then the model would be the multiple linear regression equation

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots$$

Moreover, it is very important to understand this framework and terminology as seemingly small wording or model differences can lead to very deviating output from computer programs! A common mistake is the difference between the words multivariate and multivariable. While the terms sound very similar they are quite different! Multivariable simply means that you have multiple input variables which is fine as illustrated above with the regression equation. However, multivariate means that you have a multiple number of varying output variables which is not such a simple situation. In this chapter we will restrict our regression applications of the general linear model to the case where we have a single response variable; hence, Y in our model will be restricted to a column vector of measurements of the single response variable Y.

A very important detail to clarify in the wording is the difference between the general linear model vs generalized linear model and the difference between multivariate vs multivariable. Many people mistakenly call our above model the generalized linear model (perhaps they think of it is a generalization of regression combined with ANOVA), but this is not the correct wording! The generalized linear model, which is not something covered in this text, is a completely different thing. The generalized linear model is an extension of our model to address the case when the

error term e is not a normal distribution of continuous measurements. For example if you are dealing with categorical or ordinal data you may run into this generalized linear model. It is very important to understand this, what appears to be a minor, difference as when using the various software products the output for a general linear model (“lm” in R programing) will yield completely different output than that for a generalized linear model (“glm” in R programing). The truth is that the computer program will output the correct results for what you tell it to do, but the most common mistake is telling it to do the wrong thing! Again, for the remainder of this chapter the regression applications of the general linear model we study will be only of continuous numeric data to the case where we have a single response variable; thus, Y in our model will be restricted to a column vector of measurements of nice continuous data. Furthermore, while we may not address it in every example the underlying assumption of the general linear model we must meet is that the error terms are normally distributed.

3.3 one way ANOVA and the F test

In the prior chapter we looked at the problem of T hypothesis tests about the mean difference on two samples of data, and we will now extend this idea to the problem of three or more samples of data. The more than two sample analysis is commonly referred to as ANOVA, “analysis of variance,” and will be extremely useful later on when doing error analysis from regression models.

To begin for consistency we will utilize the notation $X_{i,j}$ to represent the i^{th} element in j^{th} treatment. For example in the data set

Treatment 1	Treatment 2	Treatment 3
75	81	79
77	83	80
79	85	81

Data set 3.3.1

we call the second data point in the 1^{st} sample $X_{2,1}=77$ and the 1^{st} data point in the 3^{rd} sample $X_{1,3} = 79$. Now, for this data set we notice that the overall mean is 80, so for notation we call X= 80. Also, if we look at the first column as a data set in isolation we notice that the mean of the first sample is 77, and likewise the mean of the second sample is 83; thus, for notation we call $\bar{x}_1 = 77$, $\bar{x}_2 = 83$, and $\bar{x}_3 = 80$. Also, we reserve the capital letter N to define the total number of data points in our overall sample, while lower case m is reserved to define how many column samples, and lower case n_j is reserved to define how many data points occur in the j^{th} sample. However, for

simplification in this text we will only address examples of equal column sample size, hence we will have $n_1=n_2=n_3$ which we will just refer to as lower case n. Therefore, we will always have N= m*n.

Now, the main concept of ANOVA “analysis of variance” is to, as one might expect from the title, analyze the deviations within the data coming from the so called sum of squares used in calculating variance. Namely, we will look to dissect the sum of squares into two components: deviation within samples & deviation between samples. Then we will conduct a formal hypothesis test to determine if there is significant difference between the samples. To begin we define the sum of squares total as

$$\text{SST}=\sum_{i=1}^{n} \quad \sum_{j=1}^{M}(\mathrm{X}_{i,j} - \bar{X})^2$$

$$\sum \sum = 72.$$

which will have a so called “degrees of freedom” = N-1. Through algebraic analysis it can be shown that the SST term can be broken down as

$$\text{SST}=\sum_{i=1}^{n} \quad \sum_{j=1}^{M}(\mathrm{X}_{i,j} - \bar{X}_J)^2 + n\sum_{j=1}^{m}(\bar{X}_J - \bar{X})^2$$

where the first term can be viewed as the sum of deviation within the column samples and the second term can be viewed as the sum of deviations between the samples. Hence, we define

$$\text{SSW}=\sum_{i=1}^{n} \quad \sum_{j=1}^{M}(\mathrm{X}_{i,j} - \bar{X}_J)^2$$

which has the so called “degrees of freedom” =m(n-1), and we define

$$\text{SSB}=n\sum_{j=1}^{m}(\bar{X}_J - \bar{X})^2$$

which has the so called “degrees of freedom” = m-1.

In order to fully understand what these sum of squares really represent it is best to work through an illustrative example fully by hand so prior to continuing with our theoretical development let us study the data set we have. Now, we observe that the grand mean is X=80 and the individual sample means are $\bar{x}_1$ =77 for the first data set, $\bar{x}_2$ = 83 for the second data set, and $\bar{x}_3$ = 80, and the third data set. Thus, a routine computation yields

$$SSB=3\sum_{j=1}^{3}(\bar{X}_j - \bar{X})^2 = 3(77-80)^2 + 3(83-80)^2 + 3(80-80)^2= 54.$$

Now, the SSW computation is a little more in depth and the best way to conduct it is to go back to the raw data set and subtract the column mean from each data point column by column. Hence, we construct the table

75-77	81-80	79-83
77-77	83-80	80-83
79-77	85-80	81-83

And, we can then compute the SSW within the three columns as

$$SSW_1 = (75-77)^2 + (77-77)^2 + (79-77)^2 = 8$$

$$SSW_2 = (81-83)^2 + (83-83)^2 + (85-83)^2 = 8$$

$$SSW_3 = (79-80)^2 + (80-80)^2 + (81-80)^2 = 2$$

Then, we compute the SSW as the sum of these, hence SSW = 8+8+2 =18.

At this point we have all of our computations needed to perform the desired hypothesis test, but a very useful trick is worthy to observe. Namely, we know that SST = SSB + SSW and usually the overall variance, hence SST, is known prior to starting any analysis (in this example SST is 72). Furthermore, we have seen that SSB is really not that lengthy of computation to conduct, but SSW is quite complicated to conduct and it would be nice to be able to avoid. If one knows SST = 72 and then computes SSB =54 they can essentially back solve

$$SST = SSB + SSW$$

To find

$$SSW = SST – SSB = 72 – 54 = 18.$$

Likewise, a similar relation exists for the degrees of freedom $df_T = N-1$, $df_B=m-1$ and $df_W=m(n-1)$ as

$$df_T = df_B + df_W.$$

Now that all of the computations have been addressed let us proceed to develop the hypothesis test. The underlying assumption of the hypothesis is to test if there is any significant difference between the samples or not. Thus, the null hypothesis will be $\mu_1 = \mu_2 = \mu_3$. We will investigate the ratio of sum of squares between to within, of course using mean squares with the division by degrees of freedom. Furthermore, since both SSB and SSW are constructed as the square of variables which are assumed to be normally distributed, then both SSB and SSW will be appropriate to relate to the chi squared density, and the ratio of two chi squared is known as the F density. Hence, we will use the F density, with numerator degrees of freedom being df_B and degrees of freedom being df_W to obtain our critical values. Doing so we construct the usual four step hypothesis test procedure as

First, null H: $\mu_1 = \mu_2 = \mu_3$

Alt H: at least one pair of μ differs.

Second, assume the null is truthful, and reject if: TS $> F_{dfB,dfW,\alpha}$

Third, TS $= \frac{\frac{SSB}{M-1}}{\frac{SSW}{m(n-1)}}$.

Fourth, conclusion.

Ex 3.3.1 Conduct the ANOVA hypothesis test for data set 3.3.1 at the significance level of α=5%.

Again, we note that a full solution to a hypothesis test problem has four steps, and a critical value will be needed. For this example the F critical value $F_{0.05,2,6} = 5.14$ can be obtained from either online tables/calculators or using the density and logic outlined in chapter 2. Thus, we can proceed to conduct the full hypothesis test utilizing our prior computations as:

First, null H: $\mu_1 = \mu_2 = \mu_3$

Alt H: at least one pair of μ differs.

Second, assume the null is truthful, and reject if: TS > 5.14

Third, TS $= \frac{\frac{SSB}{M-1}}{\frac{SSW}{m(n-1)}}$ = TS $= \frac{\frac{54}{2}}{\frac{18}{6}} = 9$

Fourth, reject the null since the Test Stat of 9 is above the critical value of 5.14.

An important observation to make is even though these results do tell us there is a significant difference between the column samples it does not tell us either which samples it is from nor if it is just one pair of samples that are significantly different or if there are multiple. In order to determine such information a follow up test, often called a post hoc test, would be needed. In our example it is noted that sample one had mean 77 while sample two had mean 83 and since those means were the furthest apart one would expect that sample 1 and sample 3 is where the difference is coming from, but in

order to exactly determine this a follow up test would be needed. Furthermore, taking the time to work through the ANOVA and analyze the following two data sets will help deepen the understanding. It is very interesting to note that while on the data set 3.3.2, due to extremely low deviation within, the hypothesis rejection is obtained even though the difference between the means does not appear to be extreme (77,80 and 83). However, the result is not obtained on the data set 3.3.3 even though one may expect to see such a rejection due to the large deviation between the sample means (76,83 and 90). This result is due to the F statistic being a ratio of different kinds of deviations, namely SSB and SSW. In addition, this result further emphasizes that there is often a lot more going on within the data and it is not sufficient to look at raw data or raw differences between the means, hence the name ANOVA "Analysis of Variance."

Treatment 1	Treatment 2	Treatment 3
76	79	81
77	80	83
78	81	85

Data set 3.3.2

Treatment 1	Treatment 2	Treatment 3
72	81	79
76	83	90
80	85	101

Data set 3.3.3

3.4 single variable linear regression

In the prior section, we looked at our first application of the general linear model, namely the application to the ANOVA analysis, to investigate problem of testing for mean difference on several samples of data. Now, we begin our “predictive modeling” study by considering another application of the general linear model to the application of the so called Linear Regression Model. The equation for this Regression Line will be defined as

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

where the coefficient B_1 will be referred to as the slope coefficient, and the coefficient B_0 will be referred to as the intercept coefficient. Furthermore, the notation “Y hat” is used to identify a predicted value while the notations x and y will be reserved for the actual data values. For example, if we had the data set of n pairs (x_i , y_i) along with the regression line

$$\hat{Y} = 3x + 1$$

we would label the original data points by usual letters and then use the “Y hat” for any computed approximated values. Let us assume the 5th data point we had was (5, 17) then we would call $x_5 = 5$ and $y_5 = 17$, but if we computed a prediction as

$$\hat{Y}(5) = 3 * 5 + 1 = 16$$

then and only then would we use the “Y hat” notation. It is important to note that while in this case $\hat{Y}(5)$ is exactly the same thing as $\hat{Y}_5$ that is not always the case. The correct

notation for $\widehat{Y_n}$ is the solution of the regression line when the variable $x=x_n$ is inputted, hence we define

$$\hat{Y}_n = \hat{Y}(x_n) = b_1 x_n + b_0.$$

In this example we had the equivalence because we had the data set consisting of $x_1 = 1$ and $x_2 = 2$ etc, thus, x_5 was indeed equal to 5, but this is not always the case and it is very important to ensure the above notation is used appropriately moving forward.

Now, one of the most important things to analyze from a regression will be the error, which we will also refer to as the residual. This is defined for the i^{th} data to be the difference between the predicted "Y hat" value and the original "true data" value, which is

$$y_i - \hat{Y}_i.$$

A common issue with error analysis, or when studying variance within data, is the issue of error being positive for some data points while it is negative for others. To avoid this when summing the error to get a total we square the individual error terms. Hence, we define for a data set of n pairs, the residual sum of squares to be

$$\sum_{i=1}^{n} \left(y_i - \hat{Y}_i\right)^2.$$

which is often noted as RSS for residual sum of squares or SSE for sum of squared error. A very interesting phenomena to observe is how we derive the formulas for the coefficients b_1 and b_0 in our regression line equation as we observe the exact same formulas through two different mathematical methods. Firstly, if we apply calculus

methods to solve for the unknown coefficients b_1 and b_0 we would do so by seeking to optimize the RSS. Namely, we seek to optimize

$$\sum_{i=1}^{n}(y_i - b_1 x_i + b_0)^2.$$

and by routine knowledge of calculus, this term can be optimized by the usual minimization technique which is conducted by solving the partial derivatives of RSS with respect to b_1 and b_0 separately set equal to zero. Doing so the algebra yields the formulas

$$b_1 = r\frac{s_y}{s_x}$$

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

where the notation s_x is defining the standard deviation of the x data, s_y is defining the standard deviation of the y data, and r is the correlation between the x and y data. On the other hand, if we approach this from a statistical point of view we will obtain the exact same formulas.

If we had a case of perfect correlation between the data set of x values and the data set of y values we can recall from the definition of correlation from section 1.2

$$r = 1 - \frac{1}{2(n-1)}\sum_{i=1}^{n}\left(z_{xi} - z_{yi}\right)^2$$

that what the case r=1 is really telling us is that these data points have no deviation between the z scores z_x and z_y. Thus, we have that the Z score for the i^{th} x data point

$$z_{xi} = \frac{x_i - \bar{x}}{s_x}$$

is equal to the Z score for the i^{th} y data point

$$z_{yi} = \frac{y_i - \bar{y}}{s_y}$$

and solving the equality

$$\frac{x_i - \bar{x}}{s_x} = \frac{y_i - \bar{y}}{s_y}$$

for the i's estimated value of y we obtain

$$y_i = \frac{s_y(x_i - \bar{x})}{s_x} + \bar{y}.$$

Again, the prior development was obtained for a case of perfect correlation, r = 1 , which of course is not practical for applications. However, it can be proven that the linear conversion rule for converting z_x to an estimate of $\hat{z}_y$ with a correlation value of r, is that

$$\hat{z}_y = rz_x.$$

Hence, we get the equation for a predicted $\hat{y}_i$ for a non-perfectly correlated data set by solving

$$\frac{\hat{y}_i - \bar{y}}{s_y} = r\frac{x_i - \bar{x}}{s_x}$$

for $\hat{y}_i$ which yields the equation for the i's estimated value of y we obtain

$$\hat{y}_i = r\frac{s_y(x_i - \bar{x})}{s_x} + \bar{y}.$$

And, by a little reorganization we see this is exactly the regression line equation

$$\hat{y}_i = \left(r\frac{s_y}{s_x}\right)x_i + \bar{y} - \left(r\frac{s_y}{s_x}\right)\bar{x}$$

as

$$\hat{Y}_i = b_1 x_i + b_0.$$

If we define

$$b_1 = r\frac{s_y}{s_x}$$

$$b_0 = \bar{y} - b_1\bar{x}.$$

Again, it is very interesting to see how this same result for our regression line coefficient formula is obtained exactly the same through two different methods.

Now, prior to studying many examples and illustrations of this extremely useful regression line “predictive model,” let us quickly address one of the most powerful questions to ask in statistical data analysis: what might cause our model not to work? An initial thought is it is not linearly correlated and that is very correct. If we had two data sets the first with r = 0.2 and the second with r = 0.9, it is obvious that since the first data has a very poor correlation value that a regression model would not be appropriate to use as if the data is not correlated then a linear fit model would not work. However, the reverse of this statement is not always true! Namely, we cannot guarantee that just because the second data set has an extremely high correlation value that a regression model will work. This is a classical illustration of the difference between a necessary condition and a suffice condition. Furthermore, one can think that a solid

value of correlation is a necessary condition for the regression model to be valid, but it is not a sufficient condition for the regression model to be valid. This can be further understood by thinking that correlation is just exactly that correlation: a solid value of correlation means that there is a pattern in the data, hence it is saying the data is correlated, but it is not necessarily guaranteeing that pattern is linear. Generally speaking, we can always find a fit for a model but it may not be linear. For example, our current model is that or a linear regression model

$$\hat{Y}_i = b_1 x_i + b_0.$$

but perhaps we could have had a quadratic model

$$\hat{Y}_i = b_2 (x_i)^2 + b_1 x_i + b_0$$

or maybe an exponential decay model

$$\hat{Y}_i = b_1 e^{-x_i} + b_0$$

and of course there could be many many more examples of possible models.

At this point we will not address nonlinear fits, this will be briefly introduced later in the text; however, we must now take a brief look into validating a set of assumptions that need to be tested to verify that our linear model is working. Now, in section 3.6 we will develop a formal error analysis known as the F test or ANOVA hypothesis test for regression which will test the hypothesis of whether our model fits or not. But prior to that, let us recall one underlying assumption of the general linear model: that the residuals must be normally distributed.

The main assumption of the general linear model, hence the main assumption of our regression model, is that the error term $\hat{Y}(x_i) - Y_i$ must be a normal distribution. A quick way to check this is if given a data set of the form

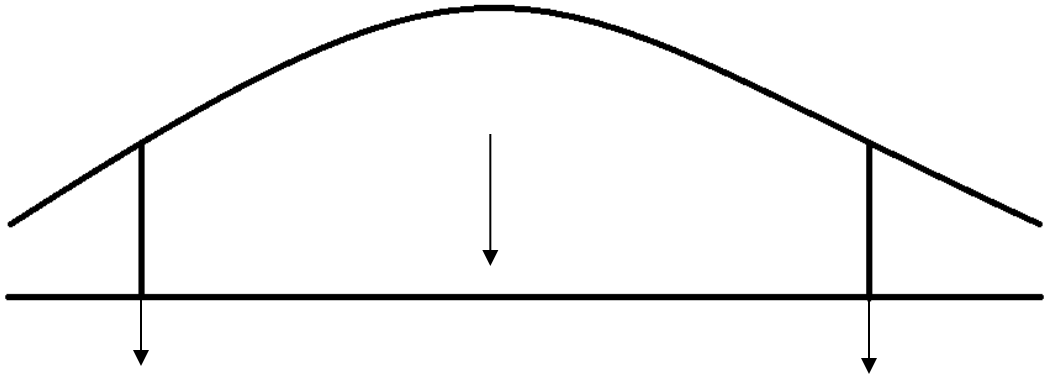
X	Y
X_1	Y_1
X_2	Y_2
.	.
.	.
.	.

conduct the usual procedure to compute regression model, $\hat{Y}_i = b_1 x_i + b_0$. Then, for each x data, compute the corresponding $\hat{Y}_i = \hat{Y}(x_i)$ which will yield a data set of the form

Y_i	$\hat{Y}(x_i)$
X_1	$\hat{Y}_1$
X_2	$\hat{Y}_2$
.	.
.	.
.	.

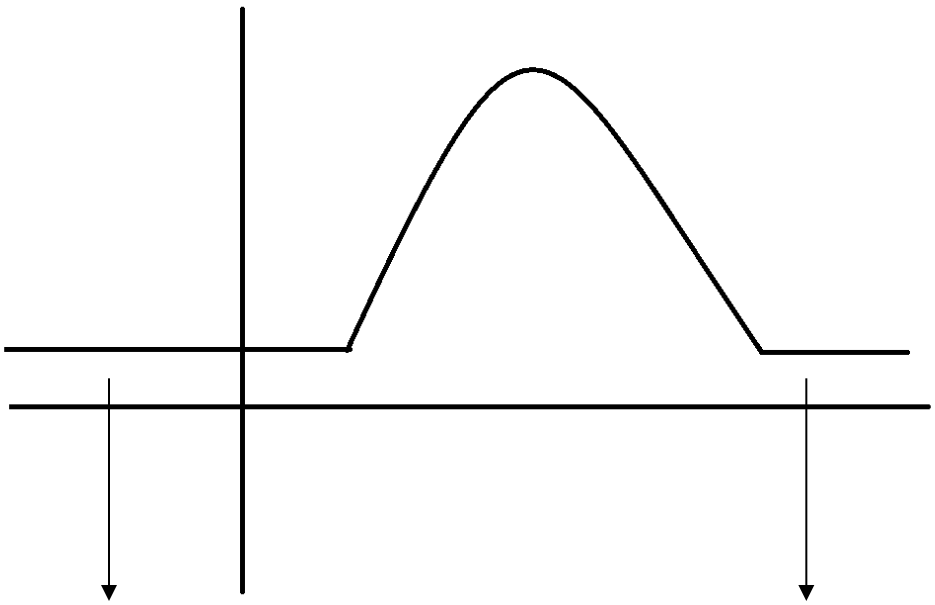
Now, from this data set the residuals $\hat{Y}_i - Y_i$ can easily be constructed. Thus, to check our assumption all that is needed to do is to verify that this newly computed data set of residuals is fitting a normal distribution. There are formal Goodness of fit tests to address this question, but often a rule of thumb can be applied to investigate; namely

that the residuals should look like a normal graph,



$\bar{X} - 2s$ $\bar{X}$ $\bar{X} + 2s$

With the empirical rule values "$\bar{X} \pm 2s$" falling near the tails. Again, this is not a formal test of normalcy, rather just a rule of thumb or a guideline. For example, we may be able to identify that a data set is not truly a normal distribution but instead a skewed distribution as the graph below illustrates:


$\bar{X}$ - 2 σ
$\bar{X}$ + 2 σ

3.5 examples of single variable regression

As developed in the prior chapter, we have defined our regression line as

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

where the coefficients b_1 and b_0 are defined as

$$b_1 = r\frac{s_y}{s_x}$$

$$b_0 = \bar{y} - b_1\bar{x}.$$

And, these computations have been proven as the unique line which fits the provided data set where the error has been minimalized, namely the sum of residual squares

$$\sum_{i=1}^{n}\left(\hat{Y}(x_i) - Y_i\right)^2 ,$$

was minimized. Let us now consider a few illustrative examples to conclude this chapter.

Ex 3.5.1 Find the regression line for data set 3.5.1

Fri	10/17	X=1	1860
Mon	10/20	X=2	1885
Tue	10/21	X=3	1910
Wed	10/22	X=4	1945
Thu	10/23	X=5	1955

data set 3.5.1 (daily S&P closing values)

Firstly, we must compute the usual descriptive statistics of the mean and standard deviation of both the x data set (note x is given the numerical

counter as x=1 is day one, x=2 is day 2 etc) and the y data values (note y is given as just the stock price) along with the correlation.

Doing so we obtain $r = 0.99, \bar{X} = 3,\ s_x = 1.58,\ \bar{Y} = 1911\ and\ s_y = 39.9$ where all values have been rounded to two decimal places. Now, we compute

$$b_1 = r\frac{s_y}{s_x} = 0.99\left(\frac{39.9}{1.58}\right) = 25$$

$$b_0 = \bar{y} - b_1\bar{x} = 1911 - 25(3) = 1836$$

Hence, we have obtained our regression line as

$$\hat{Y} = 25x + 1836.$$

.

Before continuing to more examples there a few important points to make notice of from the results of example 3.5.1. Firstly, it is very important to note that just because this example had a high value of correlation, recall r was 0.99 which is near perfect correlation, that does not ensure that the linear regression equation is will be valid. Furthermore, this example was an extremely small data set and for most any statistical model or result to be valid it is required to have at least n=30 data points (or 15-20 data points for each predictor variable as we will see later in multiple regression). In addition, it is not really appropriate to have time as an input variable in a regression model as any

data that involves time will have some sort of cyclical or seasonal effect that is not accounted for within regression. A method, which will not be discussed formally here, to deal with time data is the so called time series model. This method basically adjust the x data to remove seasonality and then creates a linear regression on that non-seasonal data. Then the predictive model will then take a form of

$$\hat{Y} = b_1 x + b_0 + c_1 \Delta \mathrm{T}$$

where the time or seasonal effect is carried within the last term. Again, time series methods or other time dependent data will not be discussed in this text and this model noted above is just one example of how this problem can be addressed and the important thing to currently understand is that time should not usually be used as a variable in a regression model. Lastly, to interpret the results of example 3.5.1 if we were a financial analyst and this model was valid, the interpretation of the coefficient b_1 being 25 is good news; this value can be interpreted directly as a slope, namely for each day forward the stock's value increases by $25.

Ex 3.5.2 Find the regression line for data set 3.5.2 and compute the residuals

x	Y
1	3
2	7
3	9
4	11
5	15

data set 3.5.2

Firstly, we must compute the usual descriptive statistics of the mean and standard deviation of both the x and y data set along with the correlation.

Doing so we obtain $r = 0.99, \bar{X} = 3,\ s_x = 1.58,\ \bar{Y} = 9\ and\ s_y = 4.47$ where all values have been rounded to two decimal places. Now, we compute

$$b_1 = r\frac{s_y}{s_x} = 0.99\left(\frac{4.47}{1.58}\right) = 2.8$$

$$b_0 = \bar{y} - b_1\bar{x} = 9 - 2.8(3) = 0.6$$

Hence, we have obtained our regression line as

$$\hat{Y} = 2.8x + 0.6.$$

Now, to compute the residuals we proceed by first computing for each x input the corresponding $\hat{y}(x_i) = 2.8x_i + 0.6$ then we can compute

y	$\hat{y}$=2.8x+0.6	$\hat{y} - y$	Squares
3	3.4	-0.4	0.16
7	6.2	0.8	0.64
9	9	0	0
11	11.8	-0.8	0.64
15	14.6	0.4	0.16

Now, analyzing the results of this example it does appear that the regression line is a good fit since the error is low. However, as previously noted, this observation alone is not enough to prove that the regression line is valid. Furthermore, even if we compute the so-called sum of squared residuals

$$\sum_{i=i}^{n}(\hat{y}(x_i) - y_i)^2 = 1.6$$

this is again just an observation that the value of error appears to be low error which is not enough to determine if the regression model is valid (e.g. what is the critical value?). In the next section we will develop a formal error analysis, which is based from an analysis of this sum of squared residuals, to formally test if the regression model is valid.

3.6 ANOVA error analysis for regression

As noted in the prior sections we have discovered two main conditions required for a regression model to be considered as a valid model. Namely, that the correlation between the input (predictor) variable x and the output (response) variable y be a very solid value of correlation, perhaps $r = 0.9$ or higher, and that the value of the error (sum of squared residuals = $\sum_{i=i}^{n}(\hat{y}(x_i) - y_i)^2$) is small. However, it is again noted that these conditions are in a sense mathematical “necessary conditions,” and in addition to these conditions being somewhat vague (e.g. what value of error is small) we have not yet developed a formal sufficient condition to say a regression model is valid. In this section we will address this issue and formalize a so called ANOVA analysis for regression model fit, but prior to doing so we must first develop some perquisite knowledge about variance within the data and the coefficient of determination.

In routine descriptive data analysis of a y data set of n elements it is well known that the variance can be computed as

$$Variance = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Now, the numerator in this term is often referred to as the sum of squares total, and after routine algebraic it can be proven that

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

Where $\hat{y}_i$ is the predicted value from the regression model, hence $\hat{y}_i = B_1 x_i + B_0$, and we refer to the latter terms as sum of squares residual and sum of squares regression respectively. Hence, we define

Def 3.6.1 The ***Sum of squares regression,*** which is often also referred to as the sum of squares model or SSM, is

$$SSM = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2.$$

And, this quantity has degrees of freedom =1.

Def 3.6.2 The ***Sum of squares residual,*** which is often also referred to as the sum of squares error or SSE, is

$$SSE = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2.$$

And, this quantity has degrees of freedom =n-2, where n is the total number of data pairs

Ex 3.6.1 Given the data set 3.5.2 has a variance of 20, find the both the sum of sum of squares regression (AKA sum of squares model “SSM”) and squares residual (AKA sum of squares error “SSE”) and for data set 3.5.2. Now, to solve this example our intuition is to compute directly the sums

$$SSM = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

and

$$SSE = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

such as previously done in the last column of Ex 3.5.2. Hence, using the last column of that example we could find SSE to be $0.16 + 0.64 + 0 + 0.64 + 0.16$ =1.6. And, while there is nothing wrong with this computational method the following trick can often be extremely time saving: Due to the fact that we know the variance is 20 and from the definition of variance we can extract the sum of squares total as

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = (n-1) \cdot variance = 4 \cdot 20 = 80 = SST.$$

As previously noted, when added together the sum of squares residual plus sum of squares regression total to SST. Hence we can solve

$$SST = SSE + SSM$$

for the desired sum of squares regression as

$$SSM = SST - SSE = 80 - 1.6 = 78.4.$$

Now, while the sum of squares computed above are very useful to get an insight of what is going on within the regression model we will now proceed to formalize some extremely useful definitions which are at the core of regression analysis.

Def 3.6.3 The ***Coefficient of determination*** which is often also referred to as the R squared or the regression model is

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSM}{SST} = \frac{(\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2)}{(\sum_{i=1}^{n}(y_i - \bar{y})^2)}.$$

This quantity can be interpreted as a percent, where it is defining what percentage of the variability with the dependent variable has been accounted from in the regression.

Ex 3.6.2 Compute the R^2 for data set 3.5.2.

As previously determined we have $SSM = 78.4$ and $SST = 80$, hence we compute the value of $R^2 = \frac{78.4}{80} = 0.98$.

In this example we have found the value of R squared is very good, essentially we are saying the 98% of the variability within the dependent variable has been accounted for. While this is an extremely positive result this alone still does not validate formally that our regression model is valid; furthermore, one of the most common mistakes in applied regression analysis is to simply conclude that a solid value of R square means that the model is valid; this is not true as there are examples where even a high value of R

squared comes from a regression model that is not valid. In fact, it can be shown that for a single variable regression model the value of R squared is exactly that, it is the correlation value squared (of course this is not the case later on in multiple variable regression models as we do not have a direct correlation value there since multiple inputs). Hence, we are still looking at a necessary condition for our regression model to be valid and prior to us developing the sufficient condition it is necessary to quickly define two side definitions that result from the coefficient of determination.

Def 3.6.4 The ***Fraction of variance unexplained*** is defined as the term subtracted from the one in the above definition, hence is given by

$$FVU = \frac{SSE}{SST}.$$

In the prior example we have R squared to be 98%, hence we can reason that the Fraction of variance unexplained was 2%. In advanced analysis it is essential to analyze how variance is effecting our data and/or models, but for our present study we shall just take this definition with the common sense interpretation.

Def 3.6.5 The ***Adjusted R squared*** is given by

$$1 - \frac{SSE/df_e}{SST/df_T}.$$

Where df_e is referring to the degrees of freedom of the residuals, which is n-2 in a single variable model, and df_T is referring to the degrees of freedom total, which is n-1.

This quantity can be interpreted as a percent and interpreted similarly to the coefficient of determination, but it addresses the sample size and how that affects natural variance within the data. Often it is written as $1 - \frac{Varince\ (residuals)}{Variance\ (total)}$.

Ex 3.6.3 Compute the adjusted R squared for data set 3.5.2.

As previously determined we have $SSE = 1.6$ and $SST = 80$, and we have a data set of $n = 5$ elements, hence we have $df_e = 3$ and $df_T = 4$ so we compute

$$1 - \frac{\frac{1.6}{3}}{\frac{80}{4}} = 1 - 0.027 = 0.973.$$

Now, from our illustration of both the coefficient of determination "R^2" and the adjusted R squared we can observe that the two values are very similar, which is generally true, but it is very important to introduce the sample size which is done through the degrees of freedom. We have now developed all of the perquisite material so let us now build a formal hypothesis test to determine if our regression model is valid, hence define our mathematical "sufficient condition." To begin, let us recall the four step process for a hypothesis test problem which has the steps

First, the hypothesis is made as a mathematical statement.

Second, the so called "critical value" and "rejection region" are defined.

Third, calculation of the test statistic

Fourth, conclusions are stated

To begin, we will define the hypothesis with the null hypothesis as the model does not fit, which mathematically will read "$\beta = 0$," and the alternate hypothesis mathematically reading "$\beta \neq 0$," hence saying there is a slope/fit. Here, we use the Greek letters to define population true values since we are under hypothesis. Now, to encourage understanding as a flow from the prior definitions we jump to the third step to define our test statistic as the F stat

$$FS = \frac{SSM/df_m}{SSE/df_e} = \frac{(\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2)}{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n-2}}.$$

Now, from a routine mathematical analysis we can observe that this test statistic is the ratio of two chi squared densities (essentially normals squared) with the numerator one having degrees of freedom $df_m = 1$ and the denominator one having degrees of freedom $df_e = n - 2$. Hence, our critical value will come from an F density with numerator degrees of freedom being 1 and denominator degrees of freedom being n-2, which at the $\alpha = 5\%$ level we will define as $F_{1,n-2,0.95}$.

We are now prepared to fully develop our hypothesis testing procedure:

First, null H: $\beta = 0$ " no fit"

Alt H: $\beta \neq 0$ " there is a fit."

Second, assume the null is truthful, and reject if: FS > $F_{1,n-2,0.95}$.

Third, $FS = \frac{SSM/df_m}{SSE/df_e} = \frac{\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}}{n-2}$.

Fourth, Conclusion

Ex 3.6.4 Compute the formal hypothesis test procedure to determine if the linear regression model, $\hat{y} = 2.8x + 0.6$, for data set 3.5.2 is valid.

We now can conduct this analysis by simply putting the appropriate values into the above four step hypothesis testing procedure and then drawing the appropriate conclusion. Prior to doing so it is known that a F density critical value will be needed, namely since we have a data set of size 5 it will be needed to find $F_{1,3,,0.95}$ and while we could revert back to our formal knowledge of probability densities and distributions from chapter 2 to formally compute this value for simplification we will just state here that $F_{1,3,,0.95} = 10.13$. Now, let us proceed to compute

First, null H: $\beta = 0$ " no fit"

Alt H: $\beta \neq 0$ " there is a fit."

Second, assume the null is truthful, and reject if: FS > 10.13.

Third, $FS = \frac{SSM/df_m}{SSE/df_e} = \frac{78.4}{1.6/3} = 147.$

Fourth, we conclude to reject the null

In example 3.6.4 the conclusion of rejecting the null hypothesis can be interpreted to mean the model is valid (technically the hypothesis test is to see if there is a slope/fit and in this case we find there is one). It is very important to understand that even though this example had a very solid value of correlation, r=0.99, along with the very solid value of the coefficient of determination, R^2 =98%, and very low sum of squared

residual error none of these facts validate the model. The model is formally found to be valid only after the results of the four step hypothesis test are completed.

Ex 3.6.5 Compute the formal hypothesis test procedure to determine if the linear regression model from data set 3.6.1 below is valid.

x	Y
1	2
2	7
3	9
4	11
5	20

data set 3.6.1

To begin we perform the routine data analysis, and doing so we obtain

$r = 0.96,\ \bar{X} = 3,\ s_x = 1.58,\ \bar{Y} = 9.8\ and\ s_y = 6.61.$

As usual all values have been rounded to two decimal places. Now, we compute

$$b_1 = r\frac{s_y}{s_x} = 0.96\left(\frac{6.61}{1.58}\right) = 4$$

$$b_0 = \bar{y} - b_1\bar{x} = 9.8 - 4(3) = -2.2$$

Hence, we have obtained our regression line as

$$\hat{Y} = 4x - 2.2.$$

Now, to compute the residuals we proceed by first computing for each x input the corresponding $\hat{y}(x_i) = 4x_i - 2.2$ then we can compute

y	$\hat{y}$=4-2.2		$\hat{y}-y$	Squares
	2	1.8	-0.2	0.04
	7	5.8	-1.2	1.44
	9	9.8	0.8	0.64
	11	13.8	2.8	7.84
	20	17.8	-2.2	4.84

We are almost prepared to fully develop our hypothesis testing procedure since we have found the sum of squares residuals by summing the last column which ` yields

$$SSE = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = 14.8$$

However, since we know that the variance (square of s_y) is 43.69 we can quickly find

$$SST = (n-1) \cdot variance = 174.77$$

Then, from that result and the known theory of sums of squares we can find

$$SSM = SST - SSE = 174.77 - 14.8 = 159.97$$

Now, since this data set is the same size as the prior example we can again use the critical value $F_{1,3,0.95} = 10.13$. Hence, we can now proceed to compute the four steps

First, null H: $\beta = 0$ " no fit"

Alt H: $\beta \neq 0$ " there is a fit."

Second, assume the null is truthful, and reject if: FS > 10.13.

Third, $FS = \frac{SSM/df_m}{SSE/df_e} = \frac{159.97}{14.8/3} = 32.43$

Fourth, we conclude to reject the null

In the same manner as in the prior example our conclusions for example 3.6.5 can be interpreted to mean the model is valid (again, technically the hypothesis test is to see if there is a slope/fit and in this case we find there is one). It is also important to note that a lot more can be obtained from our analysis than just the model works or does not work. Furthermore, the F stat

$$FS = \frac{SSM/df_m}{SSE/df_e} = \frac{(\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2)}{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n-2}}.$$

which is essentially a ratio of two kinds of errors, or more formally variances, can be viewed as a measure of how well the model is working. Namely, if one model has a higher F stat than another model we can imply that the model with the higher F stat is performing better. Often it is taught that the SSE term is “bad error” while the SSM is “natural variance,” so as this ratio gets larger the model is better. Now, for a more formal interpretation we can define the P value of a model, and the interpretation of this P value follows the usual statistical model methodology

Def 3.6.6 The ***P value*** for a regression model with $df_e = 1$ and $df_m = n - 2$ is

$$P = \int_{FS}^{\infty} F_{1,n-2}\, dx.$$

Where $F_{1,n-2}$ is referring to the F density with numerator degrees of freedom being 1 and denominator degrees of freedom of the residuals being n-2, and FS is the F stat of the model hypothesis test as previously defined.

Ex 3.6.6 Compute the P value for the regression model in data set 3.6.1.

As previously determined we have $FS = 32.43,$ hence we can now set up the P value formal to find

$$P = \int_{32.43}^{\infty} F_{1,3}\, dx.$$

where $F_{1,3}$ is referring to the F density. Computing this integral by hand can be a very complicated matter which we will not do in detail here, for the interested reader the formulas were defined in Chapter 2, but many online calculators for the F distribution and/or numerical software programs for the integration are available to compute it. Doing so we obtain

$$\int_{32.43}^{\infty} F_{1,3}\, dx = 0.01.$$

The interpretation of this P value follows the usual statistical model methodology; namely, if one model has a lower P value than another model we can imply that the model with the lower P value is performing better. Hence, we can conclude that a model with a low P value is a good model and the lower the P value the better the model!

3.7 multivariable linear regression

In the prior section we focused solely on building a regression model where we had only one input variable x. Now, we shall expand this coverage to a multivariable regression model which has several input predictor variables x1,x2… predicting a single output response variable y (we do not discuss the so called multivariate models which have multiple response variables). Prior to developing the formulas for this regression model, we must first quickly address the idea of how correlation generalizes to multivariable situations. Now, the short answer to this question is that correlation, as previously stated in the first chapter definition,

Def 1.2.1 The ***correlation*** between a data pair set x and y both of n elements is

$$r = 1 - \frac{1}{2(n-1)} \sum_{i=1}^{n} \left(Z_{xi} - Z_{yi}\right)^2$$

does not change nor generalize. Thus, when looking at a multivariable data set of the form $x_1, x_2; y$ we simply compute correlation individual as pairs of data. For example, we compute the correlation r_{y1} as the correlation between variable x_1 and y individual by using the above formula with the data columns of x_1 and y. Then, separately, we could compute the correlation r_{y2} as the correlation between variable x_2 and y individually

using the same formula with the data columns of x_1 and y. In addition, we could compute the inter-correlation r_{12} as the correlation between variable x_1 and x_2.

Ex 3.7.1 Compute the correlation r_{y1} and r_{y2} $f$or data pairs from data sets 3.7.1

x_1	x_2	y
1	2	3
2	4	6
3	7	10

data set 3.7.1

Prior to beginning the computations, we notice that this data set has near perfect correlation with the variable x_2 being two times the variable x_1 for all data points except a slight perturbation on the last value, and the response variable y is the sum $x_1 + x_2$.

Now, to compute the correlation we need to utilize the same formula two times, putting the appropriate data in each time. First, to compute the correlation between the y variable and x_1 we will need to compute

$$r_{y1} = 1 - \frac{1}{4}\sum_{i=1}^{n}\left(Z_{x1,i} - Z_{yi}\right)^2$$

Then, to compute the correlation between the y variable and x_2 we will need to compute

$$r_{y2} = 1 - \frac{1}{4}\sum_{i=1}^{n}\left(Z_{x2,i} - Z_{yi}\right)^2$$

However, prior to doing this computation, we must first put our data into the normalized Z scores. Thus, beginning with the x_1 and y pairs, we obtain

$Z_{x1,i} = \frac{x_i - \bar{X}}{s}$	$Z_{yi} = \frac{y_i - \bar{Y}}{s}$	Differences=$Z_{xi} - Z_{yi}$	$(Z_{xi} - Z_{yi})^2$
-1	-0.95	0.05	0.0025
0	-0.09	-0.09	0.0081
1	1.04	0.04	0.0016

Now, we can compute

$$r_{y1} = 1 - \frac{1}{4}(0.0025 + 0.0081 + 0.0016) = 0.9979$$

Likewise, continuing with the x_2 and y pairs, we obtain

$Z_{x2,i} = \frac{x_i - \bar{X}}{s}$	$Z_{yi} = \frac{y_i - \bar{Y}}{s}$	Differences=$Z_{xi} - Z_{yi}$	$(Z_{xi} - Z_{yi})^2$
-0.93	-0.95	0.02	0.0004
-0.13	-0.09	-0.04	0.0016
1.06	1.04	0.02	0.0004

And, from this we can compute

$$r_{y2} = 1 - \frac{1}{4}(0.0004 + 0.0016 + 0.0004) = 0.9994$$

As expected, these values are showing extremely high values of correlation, and this is due to the fact of the previously observed direct linear pattern between the input "x predictor" variables and the response variable y. In addition, it was observed that there is a near direct pattern between the input variables x_1 and x_2 which can also be observed by computing the inter-correlation

$$r_{12} = 1 - \frac{1}{4}(0.0049 + 0.0169 + 0.0036) = 0.9937.$$

A very important point to note here is that in this case, since $x_2 \approx 2x_1$ it would not make sense to consider the multiple regression line

$$\hat{y} = b_1 x_1 + b_2 x_2 + b_0.$$

Due to the fact of the high value of inter-correlation this equation is mathematically equivalent to

$$\hat{y} = c_1 x_1 + b_0.$$

where $c_1 \approx b_1 + 2b_2$. In fact, adding the two variables into the equation actually has an adverse effect known as multicollinearity. To avoid such a problem from here forward, we will use the rule of thumb that any two input predictors variables should only be utilized in the model if their cross correlation is $r_{jk} < 0.9$.

Def 3.7.1 Two variables x_j and x_k can lead to ***multicollinearity*** within a multivariable regression model if $r_{jk} > 0.9,$ and in such a case one of the variables should be removed from the model. The ***variance inflation factor*** of a variable x_j is

$$VIF = \frac{1}{R_j^2}$$

where R_j^2 is the R squared of the regression model with variable x_j put as the response variable and all other variables put as the predictors

The issue of inter-correlation and/or multicollinearity is a very serious issue which can lead to the results of a regression model being invalid. Now, there is not an exact mathematical theory of how to resolve such an issue and it is generally resolved case by case on individual data set; however, there are two "rules of thumb" worthy to be aware of. Firstly, if the variables x_j and x_k do have $r_{jk} > 0.9$ then we know that one of them should be removed, and a "common sense" approach is to look at the other correlations. For example if the correlation between y and x_j is high while the correlation between y and x_k is low, then it would make sense that x_j is probably your most deterministic variable and is the one to keep (think correlation to Y is desired). Of course, it is not always such a clear cut issue and often there are many other cross correlations. For example, if the cross correlation between the other predictor variables and x_j is very high while the cross correlation between the other predictor variables and x_k is low then it may not be advisable to keep variable x_j. Again, there is not exactly a firm mathematical rule here, honestly while the area of multicollinearity within data sets is one of the most serious issues in regression analysis it is still a bit of a "grey area," but many authors & statistician agree on this second rule of thumb. Namely, that a

variance inflation factor higher than 10 indicates a serious multicollinearity problem, and the higher the value of a variable's VIF the more severe.

Now, let us proceed to discuss our main focus of this section, the multiple linear region model

Def 3.7.2 The ***multiple linear region*** model for a data set of n elements in the form $x_1, x_2; y$ is given by

$$\hat{y} = B_{z1}Z_1 + B_{z2}Z_2$$

when the variables are normalized, or is given by

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

when the variables are in their natural coordinates. And, the coefficients of the normed line are found as

$$B_{z1} = \frac{r_{y1} - r_{y2}r_{12}}{1-(r_{12})^2},$$

$$B_{z2} = \frac{r_{y2} - r_{y1}r_{12}}{1-(r_{12})^2},$$

which are determined through the same "least squares" method as the single variable model. And, the coefficients of the regression line in the natural coordinates are found as

$$b_1 = (\frac{s_y}{s_{x1}})B_{z1},$$

$$b_2 = (\frac{s_y}{s_{x2}})B_{z2},$$

$$b_0 = \bar{y} - b_1\overline{x_1} + b_2\overline{x_2}.$$

Ex 3.7.2 Compute the linear regression line for the data set 3.7.1

To begin, we recall as previously computed that $r_{y1} = 0.9979, r_{y2} = 0.9994$ and $r_{12} = 0.9937$. Thus, we can compute

$$B_{z1} = \frac{r_{y1} - r_{y2}r_{12}}{1-(r_{12})^2} = \frac{0.9979 - 0.9994 \cdot 0.9937}{1-(0.9937)^2} = 0.38,$$

and

$$B_{z2} = \frac{r_{y2} - r_{y1}r_{12}}{1-(r_{12})^2} = \frac{0.9994 \ - 0.9979 \cdot 0.9937}{1-(0.9937)^2} = 0.62$$

which allows us to compute the normed line as

$$\hat{y} = 0.38Z_1 + 0.62Z_2.$$

Now, by computing the standard deviation of all three variables $s_{x1} = 1, s_{x2} = 2.52$ and $s_y = 3.51$, along with the means $\overline{x_1} = 2, \overline{x_2} = 4.33$ and $\bar{y} = 6.33$ we can compute

$$b_1 = (\frac{s_y}{s_{x1}})B_{z1} = \left(\frac{3.51}{1}\right)0.38 = 1.36,$$

$$b_2 = (\frac{s_y}{s_{x2}})B_{z2} = \left(\frac{3.51}{2.52}\right)0.62 = 0.86,$$

and

$$b_0 = \bar{y} - b_1\overline{x_1} - b_2\overline{x_2} = 6.33 - 1.36 \cdot 2 - 0.86 \cdot 4.33 = -0.11.$$

Hence, we can now construct our final solution of the regression line as

$$\hat{y} = b_1x_1 + b_2x_2 + b_0 = 1.36x_1 + 0.86x_2 - 0.11.$$

A very interesting result of this example is that if this example is computed on various numerical statistical software packages many of the results will null out to zero due to rounding issues and so Fourth. Thus, while we will from this point forward rely on outputs from such software programs – it is just not practical to conduct such lengthy calculations for each example – it is very important to know of the true math formals and how to use them "just in case." Furthermore, all of these formals will easily generalize to three or more variable models; for example the only change to the coefficient formula for a three variable model would be the additional term subtracted in the numerator, i.e. the numerator B_{z1} would become $r_{y1} - r_{y2}r_{12} - r_{y3}r_{13}$, likewise for B_{z2} and B_{z3} etc. Then the other formulas easily generalize as we will find that

$$b_j = (\frac{s_y}{s_{xj}})B_{zj.}$$

Now that we have developed all of the necessary formulas to build the regression line let us generalize the hypothesis test procedure from the prior section. To begin we will consider an arbitrary multivariable data set of the form $x_1, x_2, \ldots; y$ where we will have n total data points and k total predictor "x input variables." From this we will define the appropriate degrees of freedom as $df_m = k$ for the numerator "regression," and the denominator "residual" will have degrees of freedom $df_e = n - k - 1$. Hence, our critical value will come from an F density with numerator degrees of freedom being k and

denominator degrees of freedom being n-k-1, which at the $\alpha = 5\%$ level we will define as $F_{k,n-k-1,0.95}$.

We are now prepared to fully develop our hypothesis testing procedure:

First, null H: $\beta = 0$ " no fit"

Alt H: $\beta \neq 0$ " there is a fit."

Second, assume the null is truthful, and reject if: FS > $F_{1,n-2,0.95}$.

Third, $FS = \frac{SSM/df_m}{SSE/df_e} = \frac{\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{k}}{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n-k-1}}$.

Fourth, Conclusion.

Ex 3.7.3 Compute the formal hypothesis test procedure to determine if the linear regression model, $\hat{y} = 1.36x_1 + 0.86x_2 - 0.11$, for data set 3.7.1 is valid.

Analogously as to the single variable model, we now can conduct this analysis by simply putting the appropriate values into the four step hypothesis testing procedure and then drawing conclusion. However, we will first need to compute the SSM and SSE terms and due to the lengthy computations for multiple regression models this is most likely done using a statistical software program or at least a spreadsheet program such as excel to do the computations. Doing so we obtain

$$SSM = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = 24.67$$

and

$$SSE = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = 0.004$$

Now, it is known that a F density critical value will be needed, and in this example we have a data set of size n=3 and with k=2 variables. Thus, when attempting to set up the critical value $F_{k,n-k-1,0.95}$ we obtain a strange result, namely the critical value required is $F_{2,0,0.95}$ which will cause the online F distribution calculators (or manual integration formula) to report an error. Likewise, the formula

$$FS = \frac{SSM/df_m}{SSE/df_e} = \frac{\frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{k}}{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n-k-1}}$$

will yield a result of division by zero. These are due to an assumption which we have severely violated with such a small data set, so at this point we are not able to proceed with our analysis, and the most logical conclusion would be that this model is invalid due to the data set being to small

In example 3.7.3 we have obtained poor results due to the fact that we have violated a sample size requirement. While there is not a formal rule, it is generally understood for a multiple variable regression model to be valid it is needed to have at least 15 to 20 data points for each input variable. Thus, in our two variable model it would have been needed to have at least 30 to 40 data values, and of course our example was not valid

as the presented data set was extremely small. It is essential to note that while in this text book examples are provided, for nice numerical illustration, with small data set, but this is not practical for real world applications. Generally speaking, most real world data studies involve data sets with the size of n being in the hundreds if not thousands, and of course the more data collected the better. However, for formalities it is important to recall that for any regression model to be valid it is needed to have at least 15 to 20 data points for each input variable.

Ex 3.7.4 Compute the formal hypothesis test procedure to determine a 7 variable multiple regression model is valid, provided that the data set had 140 data pairs and it was found that

$$SSM = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = 200$$

and

$$SSE = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = 10$$

Now, we now can conduct this analysis by simply putting the appropriate values into the four step hypothesis testing procedure and then drawing a conclusion. However, it is known that a F density critical value will be needed, and in this example we have n=140 with k=7 variables. Thus, we use the online calculator to set up the critical value

$$F_{k,n-k-1,0.95} = F_{7,132,,0.95} = 2.08.$$

Now, let us proceed to compute

First, null H: $\beta = 0$ " no fit"

Alt H: $\beta \neq 0$ " there is a fit."

Second, assume the null is truthful, and reject if: FS > 2.08.

Third, $FS = \frac{SSM/df_m}{SSE/df_e} = \frac{200/7}{10/132} = 377.14$

Fourth, we conclude to reject the null

In example 3.7.4 the conclusion of rejecting the null hypothesis can be interpreted to mean the model is valid.

Ex 3.7.5 Compute the formal hypothesis test procedure to determine a 5 variable multiple regression model is valid, provided that the data set had 200 data pairs and it was found that

$$SSM = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = 300$$

and

$$SSE = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = 7{,}500$$

We can now conduct this analysis by simply putting the appropriate values into the four step hypothesis testing procedure and then drawing a conclusion. However, it is known that a F density critical value will be needed, and in this

example we have n=200 with k=5 variables. Thus, we use the online calculator to set up the critical value

$$F_{k,n-k-1,0.95} = F_{5,194,,0.95} = 2.26.$$

Now, let us proceed to compute

First, null H: $\beta = 0$ " no fit"

Alt H: $\beta \neq 0$ " there is a fit."

Second, assume the null is truthful, and reject if: FS > 2.26.

Third, $FS = \frac{SSM/df_m}{SSE/df_e} = \frac{300/5}{7500/194} = 1.55$

Fourth, we fail to reject the null

In example 3.7.5 the conclusion of failure to reject the null hypothesis can be interpreted to mean the model is not valid. A very important point to note when comparing example 3.7.4 and 3.7.5 is that a larger data set does not guarantee the model will be valid. Namely, in example 3.7.4 we had a nicely working model from a data set of 140, which is right on the line of "15-20 data points for each variable," but in example 3.7.5 we had a larger data set that yielded a model that was not valid. The important thing to note is that it is not sample size alone which makes the model valid, rather it is the ratio from the F statistic that concludes. Furthermore, one can think of the ratio SSM to SSE as a ratio of "OK error" to "bad error." Hence, in example 3.7.4 we had the ratio of 200 "Ok

error” to 10 “bad error,” which is in the manner it should be and therefore we find a valid model. However, in example 3.7.5 we had the ratio of 300 “Ok error” to 7,500 “bad error,” which is in not good as in this case the bad error is clearly dominant. Of course this idea of ratio is a nice way to interpret the situation informally the full and understanding what is going on, but the proper way to determine if a model is valid is to utilize the full hypothesis testing procedure along with the corresponding degrees of freedom along F distribution critical values and so Fourth.

3.8 a brief introduction to model optimization

As previously discussed one major threat to finding the best regression model from a given data set is multicollinearity from the predictor variables. While we will not do an in depth study of this matter a few comments are worthy prior to continuing. To being one crucial thing to remember is the model you design is the model you design. Yes, there are mathematical rules which yield advice and direction but some of them are to some point "grey areas." For example, a general rule of thumb is that no pair of two input variables x_j and x_k should be put into the model together if their correlation r_{jk} is higher than 0.9. However, there is not an exact formula to explain which one to keep and which one to remove, and in some cases the decision comes down to either common sense or choice of the model designer due to his or her thoughts related to what the model is actually predicting in the real world. For example if I was creating a regression model

$$\hat{y} = b_1 x_1 + b_2 x_2 + b_3 x_3 + b_0$$

where x_1 represents the market's gross domestic product while x_2 represents the producer price index, and x_3 represents the consumer price index. If a high correlation r_{23} was found, while the decision to remove x_2 or x_3 could be made from mathematical principals it may also be decided due to real world understanding. Namely, if this model was predicting a response value y that was something of interest specifically to consumer value, perhaps the price of gasoline, then it could make sense to keep the variable x_3. However, if this model was predicting a response value y that was something of interest to more of a general market response, perhaps the price of an overall stock market, then it could make sense to keep the variable x_2. In the reminder

of this section we will not discuss these details, as they are really on a case by case basis, but rather we will summarize a few mathematical results and discuss how they can be applied to optimize a regression model.

Def 3.8.1 If two variables x_j and x_k within a regression model have $r_{jk} > 0.9$ then one of the variables should be removed otherwise the model could be subject to multicollinearity which can cause numerical results and/or the model itself to be not valid. The process to decide which variable to remove should be determined by:

(i) If the model has a lot of other variables (generally k = 5 or more), the other correlations between x_j and x_k separately with the other variables should be computed. If either x_j or x_k has significant inter correlation with other variables then the variables x_j or x_k which has the highest should be removed.

(ii) The other correlations between x_j and x_k separately with the response variables should be computed. The variables x_j or x_k which has the highest correlation to the response variables should be kept and the other removed.

(iii) The single variables regression model between y and x_j then x_k separately should be run. The variables x_j or x_k which has the best result (lowest P value of the model) should be kept and the other removed.

Generally speaking the results of step (ii) and (iii) should yield the same result, and in the case where they do not it is quite possible that a linear fit is not occurring between the variables. Furthermore, some author's have suggested to first look at the full

multiple regression model with all variables included (e.g. including both x_j and x_k) then draw conclusion from the full model such as a variable test stat analysis. However, that process is not suggested here as this process is considering the case when the multicollinearity is so severe that the results from the full regression model may not be valid at all.

Def 3.8.2 An ***optimal linear regression*** model between two different models

$$\hat{y} = \sum_{j=1}^{n} b_j x_j + b_0$$

and

$$\hat{y} = \sum_{j=1}^{m} c_j x_j + c_0$$

is the model which has the lower model ANOVA analysis P value (generally speaking the model with a higher FS is a better model)

It is important to note here that these two models are not necessarily having the same number of input variables.

Def 3.8.3 An linear regression model is ***computationally improved*** if the model

$$\hat{y} = \sum_{j=1}^{n} b_j x_j + b_0$$

can be reduced to

$$\hat{y} = \sum_{j=1}^{n-1} c_j x_j + c_0$$

Without any significant change to model ANOVA analysis P value.

A notational point within definition 7.8.3 is that a variable has been removed and it is not necessarily the nt^{h} x variables. For example we may find, such as in the introductory illustrative economic example, that the 2nd of 3 variables is the one to be removed.

Now, that we have laid the definitions and ground work for our task let us solve the problem through the extremely important so called variable test stat analysis.

It is important to note here that these two models are not necessarily having the same number of input variables.

Def 3.8.4 In the linear regression model

$$\hat{y} = \sum_{j=1}^{n} b_j x_j + b_0$$

the individual predictor variables' s ***strength of variables test statistic*** is

$$TS(x_i) = \frac{b_i}{s_e}$$

where s_e is the standard error of variable x_i.

Example 3.8.1 For a large data set with the predictor variables being the value of the S&P 500 stock value during the 2000^{s} and 2010^{s} decade a regression model

$$\hat{y} = b_1 x_1 + b_2 x_2 + b_3 x_3 + b_0$$

was found. Here the 1st variables represents the United States' Gross Domestic product while the 2nd variable represents the United States' Consumer Price Index and the 3rd variables represents the total money aggregate in the US economy, all of which are readily available from government database websites. The model was found to have the test statistics of $TS(x_1) = 11.63$, $TS(x_2) = 0.27$ and $TS(x_3) = -14.13$. Use this information to find which variables is most deterministic.

Now, referring to definition 3.8.1 we observe that when considering absolute values the x_3 variable has the highest strength of coefficient test statistics, hence that variable would be by definition the most deterministic. However, it is important to note that this result does not imply that money aggregate is the most deterministic variables in predicting the stock market, rather is just say that in this model it is the strongest variable! Furthermore, at this point we don't know if either this model is valid or optimal!

Def 3.8.5 In the linear regression model which valid (hence passes the ANOVA hypothesis test)

$$\hat{y} = \sum_{j=1}^{n} b_j x_j + b_0$$

If a variables has a ***strength of variables test statistic absolute value*** greater than 1.96 it is considered to be significant and needed in the model, and

variables with test statistics less than 1.96 should be removed. This should be done one by one, first removing the variable with the lowest absolute value test stat and then rerunning the regression model with the remaining n-1 variables to obtain both a new model equation along with coefficients & strength of variable test statistics.

Example 3.8.2 For the large data set from example 3.8.1 with regression model

$$\hat{y} = b_1 x_1 + b_2\, x_2 + b_3 x_3 + b_0$$

Remove the weakest variable to create an optimal model.

Again, referring to definition 3.8.1 and results of example 3.8.1 we observe that the x_2 variable has the lowest strength of coefficient test statistics, hence that variable would be least deterministic and should be remove since it's test statistic is less than the crtical of 1.96. Then, rerunning the regression on the new data set, with the x2 variable deleted, a new regression model is created,

$$\hat{y} = b_1 x_1 + b_3 x_3 + b_0.$$

It is very important to note here that the b_1, b_3 and b_0 in this model are not the same values and are not related to the b_1, b_2, b_3 and b_0 from the prior model. Furthermore, it is interest to note that the F stat from the old model was 726, while the F stat from the new model is 1092 which clearly shows the model has improved and this process does indeed "optimize our regression model." In addition, this model has been computationally improved due to the fact that there

is one less input variable in the data set, hence less computations for the computer to compute which is essentially a savings in the "cost of computing."

Example 3.8.3 For a large data set with the predictor variables being the value of the value of a particular stock a regression model

$$\hat{y} = b_1 x_1 + b_2 x_2 + b_3 x_3 + b_0$$

was found. Here the 1st variables represents the company's profit while the 2nd variable represents the value constructed from conditional return of the stock from overall market and the 3rd variables represents a value constructed as a weighted average of values of other company's values in the same sector. The model was found to have the test statistics of $TS(x_1) = 7.1$, $TS(x_2) = 10.1$ and $TS(x_3) = -12.1$. Use this information to find which variables is most deterministic and remove any unnecessary variables to find an optimal model.

Now, referring to definition 3.8.1 we observe that when considering absolute values the x_3 variable has the highest strength of coefficient test statistics, hence that variable would be by definition the most deterministic. Furthermore, since all values are higher than the critical of 1.96 no variables need to be removed and the original model is taken as our optimal model.

Now, that we demonstrated an example of this process we have completed our study of optimizing a regression model. While it would be nice to add some further examples here to further the truth is the best way to practice with these examples is hands on with real time data sets. The reader is encourage to search online for the many readily available interesting data sets and apply the logic learned here hands on

to experiment. Furthermore, while it is quite complicated to add large data sets here into the textbook the author does have available the data sets used here along with many others. For the interested reader any, non-confidential data sets such as those used in example 3.8.1 and 3.8.2, are available by directly contacting the author through the email provided at his primary current institution.

4 Applications to financial modeling

4.1 Introduction and definition of volatility

When studying any financial model, or more generally speaking anything about the financial markets, the word volatility will come up and often will lead to more unsolved questions than answers. For example, one of the founders of the famous “Black-Sholes model,” who in 1997 won the Nobel Prize in Economic Sciences for work on that model, is infamous for a statement regarding volatility that basically said well the only problem is with volatility we are trying to describe human behavior which by definition is chaotic and random. While this statement is very true and it is quite possible a true measure of volatility may not be definable, in the following pair of definitions with examples we will outline the current two common approaches utilized to compute market volatility as this important value is needed to run any financial model.

In the later sections of this chapter we will firstly use our knowledge from this textbook to create a multiple regression model to model the S&P 500 stock market, and then in the last section we will apply that model to suggest a new method which could potentially be used in place of or alongside of existing volatility. Prior to doing so we will first give a quick summary of the main results from the famous Black-Sholes model which requires some knowledge from partial differential equations; however, it is worthy to note that for the reader who either does not have such knowledge or is not interested to study such matters they can jump now directly to definition 4.1.2 without any loss of continuity.

Now, to being our study we will start by taking the following partial differential equation as definition:

Def 4.1.1 The ***Black-Sholes partial differential equation***, which is designed to predict the fair price of an option V from a stock with price S and with volatility σ along with the current risk free interest rate value r, is given by

$$\frac{\partial V}{\partial t}+\frac{1}{2}\sigma^2 S^2\frac{\partial^2 V}{\partial S^2}+rS\frac{\partial V}{\partial S}-rV=0,$$

where as usual the variable t is used to define time.

In this text we will focus on solving examples and illustrating the importance of volatility; however, it is interesting for the reader to observe that the solution of this partial differential equation can be obtained through Fourier transforms and also how this equation can be related to the famous “heat equation” from parabolic equations.

Now, to do so we will begin by defining some values of the initial data that will be needed for the reader who wishes to complete this derivation in full detail, but since this text is not a primer for financial mathematics we will just state the results here rather than dedicate pages to fully explain and define things like answer the question “exactly what is an option and what does its value mean?” Hence, we will take as definition the initial data as

$$V(0,t)=0,\qquad \lim_{S-\infty} V(S,t)=S$$

And, in order to remove the nonlinearity in S in the coefficients of the 2nd and 3rd terms in the equation if we take $t=T-\frac{\tau}{0.5\sigma^2}$ and $S=Ke^x$ then let $V(S,T)=Kv(x,\tau)$, where T

and K can be viewed as fixed constant parameters, then the Black-Sholes partial differential equation will become

$$\frac{\partial V}{\partial \tau} = \frac{\partial^2 V}{\partial x^2} + (k-1)\frac{\partial V}{\partial x} + rV.$$

Then, after applying a routine canonical form variable transformation, this equation can be quickly identified as the famous “heat equation” parabolic partial differential equation

$$u_t = cu_{xx}.$$

And, as it is well known, after transforming this equation to Fourier space by applying the Fourier transform of $F(\omega, t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} u(x,t)\, e^{-i\omega x} dx$, the nice result of the ordinary differential equation

$$\frac{dF}{dt} = -c\omega^2 F.$$

That has a general solution of the form

$$F = Ce^{-c\omega^2 t}$$

which is interesting to see as this derivation allows one to see where the negative exponential term in our final solution comes from. Furthermore, when extracting back our solution the inverse Fourier transform

$$U(x,t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} F(x,t)\, e^{i\omega x} d\omega = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} Ce^{-c\omega^2 t}\, e^{i\omega x} d\omega \approx \int e^{-\omega^2}$$

will be required which also is interesting to see as this allows one to see where the integral form similar to our normal distribution integral comes from.

While the above derivation is logically sound, there are some details which, to avoid confusion, we will not fully expand on here such as full algebra of the Fourier transform process and the needed side data of the partial differential equation

$$V(S,T) = max(S-K,0)$$

involving the so called strike price K, as our primary focus is to just present the main idea along with solution and some examples. However, an important result from the derivation is to see where and how the certain portions of our solution came from! Now, the final solution we will be working with moving forward is presented in the following definition.

Def 4.1.2 The ***solution of the Black-Sholes*** partial differential equation, for the fair price of an "at the money" option V from a stock with price S with time to maturity T with volatility σ along with the current risk free interest rate value r, is given by

$$V = SN\left(\frac{\left(r+\frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}}\right) - ke^{rT}N\left(\frac{\left(r-\frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}}\right).$$

Here the notation N is for our standard normal distribution, hence

$$N(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}\, dx.$$

Let us now see, for illustration, a few examples of applying this formula. Prior to working these examples it is important to note that the above formal is a special case, which we are using for simplification, of a so called at the money option. This is the case where the strike price K is set to equal the initial stock price; thus, if a literature search was

done for the general solution of the Black-Sholes formula a similar formula to the above would be obtained but with one extra term and of course that extra term vanishes when considering the at the money option, which again is just done here for simplification as the focus of this textbook is on the statistical models not financial mathematics.

Ex 4.1.1 Find the fair price of an at the money, T=1 year option, with a current market interest rate of 4% which has a current price of $100 with volatility $\sigma = 0.1$.

To begin, we recall our above option pricing formula is

$$S \cdot N\left(\frac{\left(r+\frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}}\right) - ke^{rT}N\left(\frac{\left(r-\frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}}\right).$$

We compute the fair price of this option by inputting $S = 100, r = 0.04, T = 1$ and $\sigma = 0.1$, which yields the result

$$S \cdot N\left(\frac{\left(r+\frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}}\right) - ke^{rT}N\left(\frac{\left(r-\frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}}\right) = \$6.18.$$

Ex 4.1.2 For the same , T=1 year option, as in example 4.1.1. with all values the same except $\sigma = 0.25$, find the fair price of the option.

To begin, we recall our above option pricing formula is

$$S \cdot N\left(\frac{\left(r+\frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}}\right) - ke^{rT}N\left(\frac{\left(r-\frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}}\right).$$

We compute the fair price of this option by inputting $S = 100, r = 0.04, T = 1$ and $\sigma = 0.25,$ which yields the result

$$S \cdot N\left(\frac{\left(r+\frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}}\right) - ke^{rT}N\left(\frac{\left(r-\frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}}\right) = \$11.84.$$

These two examples show how important the value of volatility is when considering financial modeling, namely raising the volatility by 15% doubled the price of the option for the exact same stock under all equal market conditions.

Now, as one can see from the results of those examples the value of volatility has a major effect on the results of the financial model. We will now attempt to formally define what volatility is. The words “attempt to” are chosen appropriately here as the true definition of volatility, as truly needed within the derivation of Black-Sholes, is a value that we as human beings may not be able to accurately measure ever. In short, volatility should tell us how the market is moving and more importantly how is it going to continue that tomorrow; perhaps an acronym for volatility would be the holy grail of financial modeling! While we do not seek to resolve such a vast research problem here in the book we will conclude this section be explicitly defining two currently used measures of volatility and then in the third section, after first introducing a macroeconomic regression model in the second section, we will suggest a new method as an alternate method to measure volatility.

To begin let us consider this quick illustration of how one can interpret volatility: a value of volatility of $\sigma=0.1$ or perhaps lower really means that there is not much current market movement, perhaps just a steady up trend growing a little bit each day; and value of volatility a bit higher say $\sigma=0.2$ or 0.3 means that there is quite a bit of current market, perhaps sudden swings of one day lose a percent of value but the next day gain back half a percent; and value of volatility that is very high say $\sigma=0.5$ or even higher means that there is a lot of current market, perhaps sudden swings of one day lose two or more percent of value but the next day gain a percent or so. Now, the issue with all of these interpretations is they focus on past behavior, as we will see in the next two definitions, and we would desire a value that tells us future market behavior. Namely, we desire an interpretation such as σ = low meaning that the market is calm and is

going to stay that way for some time; while σ = medium means yes the market is currently moving a little bit, perhaps just absorbing some new news, but it is going to calm down; while σ = high means yes the market is currently moving a bit and expect to continue.

Def 4.1.3 The ***classical definition of volatility*** is that volatility is measured by computing the standard deviation of the logarithmic returns of a stock, hence

σ = standard deviation of $ln\left(\frac{S_f}{S_0}\right)$.

Ex 4.1.3 The data below is the value of the S&P 500 for 2008 showing the closing value of each month. Use S0 = which is the value of S&P 500 on January 1st 2008 to find the classic value of σ for the S&P 500 in this time period

$$SN\left(\frac{\left(r+\frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}}\right) - ke^{rT}N\left(\frac{\left(r-\frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}}\right).$$

To begin we

Def 4.1.4 The definition of ***implied volatility*** is the value of σ so that the when this value of volatility is inputted into the Black-Sholes solution, with all other parameters fixed, the solution is equal to the current market selling price of the option.

While the solution of a problem illustrating implied volatility is quite complicated, as it requires the equation to be inverted or solved by a repeating algorithm, and will not be presented here it is useful to note that both of these definitions are leading to a trailing indicator. Namely, they both rely on what happened yesterday. The first definition

directly utilizes old data while the second definition inherently utilizes old data as whoever is buying / selling will be looking to existing data, of course adding on their own speculations and opinions. Presently most market analysis utilize the second defection of implied volatility in practice.

4.2 a macro economic model & suggested further study

A widely accepted principle in the economic and financial world is the fact that various economic indicators are correlated to the market and can be used to forecast the stock market. An prior study identified that the variables of consumer price index (CPI), producer price index (PPI), gross domestic product (GDP), money supply (MS), and treasury spread (T) could be used in a multiple linear regression model to explain market (S&P 500) returns and movements. This assumes that there is a linear relationship between the S&P 500 and the aforementioned independent variables. This initial model was created using monthly data from 1974 - 2005 and yielded acceptable preliminary results; however, tests for multicollinearity were not conducted, which left much room for improvement in the model and further study.

The format of this model follows that of a normal multiple linear regression model with the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots$$

where y is the dependent variable (S&P 500), and the usual notation is utilized for the input variables and corresponding coefficients. Following the prior research [6-8] on the

subject of creating a multiple linear regression model to predict the price of the S&P 500, some important results are necessary to be taken into account for the current study. The major finding of these previous studies is the finding of substantial multicollinearity between CPI and GDP. Due to this multicollinearity that exists it becomes necessary to remove one of the two variables to improve the model and eliminate any unintended chaotic behaviour. After following routine regression coefficient analysis the variables of CPI, PPI and T were removed. At that point further data study was conducted to see if any other predictor variables could be introduced to improve the model overall; the variables were selected through a common sense approach as to what factors one would expect to predict the market, such as: the price of gasoline, or the price of boring money, or the unemployment rate etc. The final model contained variables GDP, MS, and Unemployment Rate (UI). This model was both the most statistically significant and computationally efficient model generated.

The indicators used in this new model date back to roughly 1960, however since the goal of this study was to create a new volatility model to compare with VIX, the date that the data set of our new model begins at was restricted to January 1991. This was chosen as the calculation methodology changed for the VIX in 1990, hence we started at the beginning of the closest calendar year. Taking monthly data points starting January 1st 1991 and ending December 31st 2017 yields 324 total data points for the model, with the opportunity to update the data set at any given time. The data collected was then normalized using a Z-Score transformation. The governing equation of the model over this time frame is

$$y_i = 198.8z_1 + 298.79z_2 - 210.38z_3 + 1188.91$$

where Z_1 corresponds to GDP, Z_2 corresponds to MS, and Z_3 corresponds to UI. In this model i = 1,2, … , n corresponds to the month starting with i = 1 being January 1991.

	Coefficients	*t Stat*
Intercept	1188.91	163.37
GDP	198.84	6.04
MS	298.79	9.03
UI	-210.38	-28.18

ANOVA

	df	*MS*	*F*
Regression	3	2.9E+07	1664.048
Residual	320	17159.6	
Total	323		

Regression Statistics	
Multiple R	0.96941254
R Square	0.93976068

Figure 4.2.1. The data analysis of the updated three variable model.

As one can see above, the model has a very solid F-Statistic along with a great R^2 value of 0.94. This shows that the updated multiple linear regression model is not only statistically significant, it is actually a more computationally efficient model than the original model.

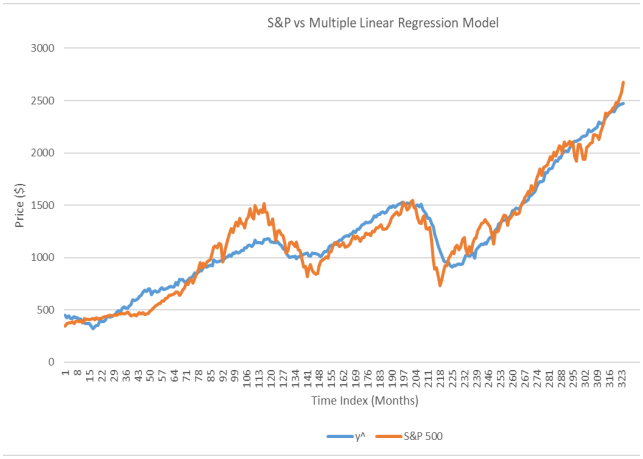


Figure 4.2.2. A graphical representation of the models

When analyzing Figure 4.2.2, the first discrepancy to look into is the large residual between the model S&P 500 and the actual S&P 500 that occurs at months 95-127. The timeframe in question is the 2000-2002 bubble and subsequent crash, known as the dot com crash. The large residuals shown in the graph are explained by intuitively understanding that the value of the market during this time period was inflated by the greed and behavior of investors, and the fundamentals for the market, which are understood to be the economic indicators used in the multiple regression model, could not justify this high price. As a result, the market eventually crashed and returned to levels similar to that of the model. With this information in hand, we will consider our multiple regression line to be the “expected value” of the market at any given time, and define volatility to be the deviation of the market from its expected value. In times of calmness, the expected value and observed value of the market would be similar, which

in turn would cause the volatility of the market to be low. Conversely, when the market starts to deviate from its expected value, say during a bubble or crash, the volatility of the market will increase at the rate that the observed value of the market deviates from its expected value. While the ongoing and future research is not detailed here in this textbook, the general idea can be thought of as suggesting a new definition of volatility perhaps similar to value

$$\sigma = \frac{(E - O)^2}{E},$$

which would fit a chi squared density model. Of course further study and experimentation is needed as one would expect a different normalization type factor in the denominator, but the idea and suggested is now completed for the interested reader to conduct their own further research!