



THE PETER J. TOBIN  
COLLEGE  
OF BUSINESS

# Review of

# BUSINESS

Volume 26, Number 3  
Fall 2005

**Special Issue:**  
**Applications of Computer Information Systems  
and Decision Sciences**

Integrating Inter- and Extra-Enterprise Applications  
Using Web Services

Wireless Networks and Security Issues

Using Probability Analysis to Define a Defendant's  
Intent Level in a Criminal Case

Analyzing the Pharmaceutical Industry Through  
Quantitative Models

The Effect of the Internet on Stock Market Volume  
and Volatility

Transforming a University from a Teaching  
Organization to a Learning Organization

Regression Analysis Revisited

# review of business

## The Peter J. Tobin College of Business

Richard A. Highfield  
Dean

Charles M. Clark  
Associate Dean

## Review of Business

Larry W. Boone  
Editor

Maxine Brady  
Contributing Editor

Gloria Brana  
Secretary

## Board of Advisors

Nina Aversano  
Aversano Consulting, LLC

Richard Carbone  
Prudential Financial

Robert J. Chrenc  
A.C. Neilsen, Retired

Jill M. Considine  
The Depository Trust  
Company

Joseph F. D'Angelo  
Hearst Corporation

John Donachie  
Fleet Specialists, Inc.

James F. Dowd  
Fairfax Inc.

Jack Foley  
Aer Lingus

Frank Fusaro  
Forum Personnel Inc.

Joseph Garcia  
Spanish Broadcasting  
System

Ken Gorman  
Apollo Partners, LLC

Robert Kalenka  
Automatic Data Processing, Inc.

Jerome Karter  
SCOR U.S. Corporation

William K. Lavin  
Consultant

George Maggiore  
Keyspan

Charles P. Menges, Jr.  
Bernstein Investment Research  
and Management

Richard E. Meyer  
Global Excess Partners LLC

Kathryn Morrissey  
AT&T

William L. Munson  
The Toa Reinsurance  
Company of America

Joseph O'Connor  
Gartmore Separate  
Accounts LLC

Thea Graves Pellman  
Professional Design Systems

John F. Robinson  
National Minority Business  
Council, Inc.

Lawrence J. Ruisi  
Loews Cineplex  
Entertainment

Joseph M. Saggese  
Borden, Inc., Retired

Joseph Scharfenberger  
J.P. Morgan Chase Corp.

Ronald T. Schroeder  
J. & W. Seligman & Co., Inc.

Edward Smith  
KPMG

Peter J. Tobin  
The Peter J. Tobin  
College of Business  
St. John's University

John P. Tutunjian  
Gourmet Events, Inc.

Charles Walsh  
Chase Manhattan Bank,  
Retired



# table of contents

Fall 2005  
Volume 26, Number 3

From the Editors . . . . .	2
FangLieh Victor Lu and Farok Vakil	
Integrating Inter- and Extra-Enterprise Applications Using Web Services . . . . .	3
Hui-Mei Yang and FangLieh Victor Lu	
Wireless Networks and Security Issues. . . . .	10
Farok Vakil	
Using Probability Analysis to Define a Defendant's Intent Level in a Criminal Case. . . . .	13
Mark I. Marpet and Christopher M. Farella	
Analyzing the Pharmaceutical Industry Through Quantitative Models . . . . .	22
Andrew Russakoff and James Gould	
The Effect of the Internet on Stock Market Volume and Volatility . . . . .	26
Farok Vakil and FangLieh Victor Lu	
Transforming a University from a Teaching Organization to a Learning Organization . . . . .	31
Hershey H. Friedman, Linda W. Friedman and Simcha Pollack	
Regression Analysis Revisited. . . . .	36
Athanasios Vasilopoulos	

# from the editors

Special Issue: Applications of Computer Information Systems and Decision Sciences

2

FangLieh Victor Lu, The Peter J. Tobin College of Business, St. John's University  
Farok Vakil, The Peter J. Tobin College of Business, St. John's University

In recent years, the demonstrated effectiveness of operations research, statistical analysis and computer information systems as aids to decision-making has made the related fields of Computer Information Systems and Decision Sciences (CIS/DS) areas of significant interest to practitioners in businesses as well as other types of organizations. CIS/DS supports decision-makers by formulating logical models which depict causal relationships among a few or many factors; measuring the magnitudes of factors involved in decisions; and establishing orderly procedures for collecting, processing and analyzing data. This issue contains examples of the wide variety of useful CIS/DS applications.

In the first article, Hui-Mei Yang and FangLieh Victor Lu discuss the ways in which Web Services have improved information-sharing within and among organizations. In general, Web Services employ XML-related technologies such as SOAP as the communication protocol, WSDL as the interface description language and UDDI for registering and searching services. The authors examine the benefits and current limitations of Web Services, then propose an integrated framework under which heterogeneous systems and applications across different networks can exchange and share data with minimal need to add or change existing hardware and software.

Second, Farok Vakil addresses business and personal/residential applications of wireless networks and their related security issues. Various wireless standards currently in use are presented, and the different characteristics of those standards are discussed. The author offers practical steps that businesses and personal computer users can take to ensure the security of their wireless networks.

Through the example of a criminal case involving driving under the influence of alcohol, Mark I. Marpet and Christopher M. Farella demonstrate how probabilistic analysis can be used in a legal situation. Their analysis helps determine whether a "reasonably-prudent individual" with full information concerning the consequences of driving while intoxicated at a specific blood alcohol content level could conclude such driving would probably cause injury. The analysis presented offers an intriguing look at the intersection of the law and the field of decision science.

Andrew Russakoff and James Gould summarize the major events in the pharmaceutical industry since the 1930s, tracing the effects of different marketing approaches, government regulations and economic incentives for innovation, among other factors. Then they explore possible approaches for examining and analyzing the industry using quantitative modeling.

Acknowledging the widespread effects of the internet with its capability to disseminate information globally in a matter of seconds, Farok Vakil and FangLieh Victor Lu investigate the effects of the online trading boom on the volume and volatility of trading in the stock market. The authors discuss the advances in business information availability, how investment firms have adapted to the democratization of business information via the internet, and the growth of online trading.

In the long run, the only sustainable source of competitive advantage is an organization's ability to learn faster than its competition. Peter Drucker has warned that 30 years from now the big university campuses will be relics. Universities as we know them won't survive. The rising costs of education, decreasing government support, the rise of for-profit institutions, globalization demands in education, technological advances, the growing number of working adults needing continuing education to avoid obsolescence and distance education are major forces causing universities to transform themselves. Hershey H. Friedman, Linda W. Friedman and Simcha Pollack discuss why it is necessary for universities to become learning organizations. The authors provide some stimulating ideas that can help facilitate the challenging transformation.

In this issue's final article, Athanasios Vasilopoulos leads the reader through a discussion of regression analysis. Since regression is at the center of almost every forecasting technique, it's important to develop a comfort level with this important statistical methodology. The theory behind regression analysis is discussed, then a step-by-step procedure is outlined to permit almost everyone to construct a regression forecasting function for both linear and multivariate cases.

# integrating inter- and extra-enterprise applications using web services

3

Integrating Inter- and Extra-Enterprise Applications Using Web Services

Hui-Mei Yang, Tatung Institute of Commerce and Technology  
FangLieh Victor Lu, The Peter J. Tobin College of Business, St. John's University

## Abstract

The fast-paced advancement of information technology is rapidly changing the business world. Fulfilling on-demand service requests is vital to today's business survival. Enterprises must incessantly upgrade and integrate their e-infrastructures in order to maintain their competitive advantage. One major roadblock to integration is that there is usually an inherent heterogeneity of system platforms and software applications existing within the same enterprise and among collaborative partners. Thus, it is common that enterprises must spend an astronomical amount of time and money to link the new and legacy systems in order to exchange data, information and knowledge.

The recent development of Web Services has greatly improved information-sharing and -exchange. Simply stated, Web Services use XML-related technologies such as SOAP as the communication protocol; WSDL as the interface description language; and UDDI for registering and searching services. In this paper, we establish an integration framework to develop web-services-enabled enterprise applications utilizing the UML tools. To achieve the desired integration, a uniformly integrated user interface will be designed and developed. Under the proposed framework, heterogeneous systems and applications across different networks can exchange and share data with a minimal need to add or change the existing hardware or software. Lastly, a completed case study [12] of the presented integration framework is briefly illustrated.

## Introduction

In facing increasingly competitive markets due to the globalization of business, business enterprises are required to unceasingly adjust to new challenges and transform themselves. Since the inception of computers, the computerization of business organizations has itself gone through a sequence of transformations due to the rapid development in information technology and increasingly complex business needs. The continuous e-revolution of business processes has encountered a number of challenging technological issues, but none more critical than the incompatibility and importability which currently exist between information systems and applications.

Historically, the inherent heterogeneity of system platforms that exist within the same enterprise and among business partners (e.g., suppliers, distributors and customers), has been a major obstacle to the smooth and timely flow of data and information required to successfully support business operations or processes. More importantly, there exists another serious incompatibility issue among software applications. "The reason is because historically most applications have not been designed for interoperability with other applications. Even when running on the same hardware platform and using the same database, different formats and implementation of business logic makes exchanging information difficult without custom built Application Program Interfaces (APIs). What's more, these applications were not designed for the web" [11]. Clearly, there is an urgent need to create an integration backbone to support all or most of the enterprise applications and data across different platforms and the World Wide Web.

## Enterprise Application Integration (EAI)

Usually, an enterprise has existing legacy systems, applications and databases and wants to continue to use them. One solution to address the incompatibility issue is the emergence of the Enterprise Application Integration (EAI) concept and its related products. The essential concept of EAI is to integrate enterprise applications and resources so that they can easily access or share business processes and exchange data. EAI's goal is to integrate the business processes, applications and data sources. This is expected to be achieved without incurring heavy-toll changes to the existing applications and the data.

Linthicum [4] states that "... Enterprise Application Integration (EAI) offers a solution to this increasingly urgent business need. It encompasses technologies that enable business processes and data to speak to one another across applications, integrating many individual systems into a seamless whole." "EAI solutions are software products that completely or partially automate the process of enabling custom-built and/or packaged business applications to exchange business-level information in formats and contexts that each understands" [5]. The key to achieve the EAI goal is to create a "middleware" platform that links to the internal and external legacy systems through adding or migrating to

*The recent development of Web Services has greatly improved the hope of making information-sharing and -exchange easier and less expensive.*

a new set of applications that use the Internet, World Wide Web, intranet, extranet and other related new technologies.

#### **Certain benefits of implementing EAI:**

- EAI dramatically reduces effort, time and cost of system maintenance.
- EAI leverages all existing legacy applications and systems by exploiting the infrastructure of the World Wide Web and Internet to reduce the efforts of application recoding.
- EAI provides a single, open architecture that is scalable, non-invasive and operates across environments, connecting databases, warehouses, packaged software and custom legacy systems.
- EAI increases efficiency – key and routine processes are automated.
- EAI adds flexibility to the reconfiguration of enterprise organizations and processes when reflecting changes in the business environment and strategies.

The growth of EAI spending is increasing sharply; \$500 million and \$900 million were spent on EAI tools in 1999 and 2000, respectively. IT analysts predict that by 2005, companies will spend \$7.3 billion on EAI.

The adoption and implementation of EAI can streamline data flows. The entire enterprise can enjoy quick and easy access to consistent, valuable information. Less programming effort is needed to create interface among systems due to EAI's process- and business-driven approach to linking various applications. Data and information channels are integrated to provide a unified view of business processes among customers, suppliers and vendors along the entire supply chain.

Unfortunately, EAI tools have many shortcomings and potential pitfalls. Pender [6] points out that "Before implementing EAI, CIOs must first be prepared to deal with the technology's potential problems: cost overruns, poor interface design and the inability to provide full system integration, among others." In addition, there are other severe EAI technology shortcomings such as the lack of open standards, difficulty in supporting many-to-many connections and limited ability for business process integration. Component-based programming plays the key role in the installation of EAI. Most software vendors often communicate distributed components within a system via a proprietary protocol, and the protocol is usually connection-

oriented. Well-known distributed technologies using proprietary protocol are DCOM (Microsoft's Distributed Component Object Model), CORBA and Java RMI. However, because of the limitations of existing technologies in facilitating communication between heterogeneous computer systems, software vendors have often resorted to building their own infrastructure. This means resources and efforts that could have been applied to add improved functionality of components or services have instead been devoted to writing proprietary network protocols or creating new interfaces.

Clearly, EAI technologies are not the final answer to the challenges of integration issues. To remediate the shortcomings encountered implementing EAI, we created an integration platform using the new and promising Web Services technology. The project has been carried out with the aid of popular UML tools, where the UML (Unified Modeling Language) is a standard language for specifying, visualizing, constructing and documenting the artifacts of software systems, as well as for business modeling and other non-software systems [3]. The integration platform is designed by taking full advantage of the cross-platform feature of Web Services technology that standardizes communication, description and discovery mechanisms. The platform allows us to dynamically integrate and distribute service components of the enterprise across the Internet through the simple creation of a user interface.

#### **A Brief Discussion on Core Technologies Involved**

##### **Client/Server Computing and Distributed Object Technology**

Developments in client/server computing since the 1980s have been helping to improve business information processing in the last 20 years. A client is a computer that requests services, and a server is defined as the computer that provides services. The client/server architecture has been created as a versatile, message-based and modular infrastructure that is intended to improve usability, flexibility, interoperability and scalability of computing as contrasted to centralized, mainframe, time sharing computing. However, as the volume of data and information flowing among enterprises increases explosively, the traditional client/server structure becomes hard pressed to keep up with complicated requests, and it is also more difficult to enlarge and maintain.

Distributed Object Technology has been introduced to alleviate some of the aforementioned client/server difficulties. The distributed object technology places a middleware between clients and servers. The middleware serves as a mediation point between applications, and provides generic interfaces with which all integrated applications pass messages to each other. Distributed Object Technology hence provides the communications channel that enables plug-and-play interoperability of distributed components across diverse networks and operating systems, allowing them to be easily assembled, reused and managed.

However, distributed object technology usually encounters a number of serious obstacles when applied on the Internet.

##### **a) It is difficult to penetrate the firewalls.**

Under TCP/IP, every frequently used protocol is given a port number, and every request packet using the specific protocol is attached with the same port number. Firewalls will usually block most other ports from access due to security reasons. This is the main cause that renders the distributed object technology impotent. Usually, the protocol of the distributed object technology is dynamically assigned a port number when there is a request for it. If no firewall exists between the server and the client, then the distributed object technology can be effectively utilized on the network. However, with the enactment of a firewall, the connection of the two communicating endpoints using the protocol will be disrupted, because the firewall disallows the arbitrary assignment of port numbers.

##### **b) The communication connection must be tightly coupled.**

Coupling implies the dependency between interactive systems or application components. Tightly coupled components tend to make maintenance and reuse much more difficult because a change in one component automatically means changes in others. For example, it will be difficult to keep the unified structure of the computing platform using the tight coupling scheme when a merger or acquisition between enterprises occurs.

##### **c) It is difficult to integrate application programs.**

The major obstacle to the smooth flow of data and information is due to the heterogeneity of IT products inherently existing within the same enterprise and among supply chain partners: suppliers, distributors and customers. The enormous variety of designs and technical specifications in products has made the integration of application components, especially over a network, a formidable task. Some of the challenges to applications integration include the methods of transport, specifications of call procedures and formats of messages.

#### **Web Services Technology**

What are Web Services? As it happens in many emerging technologies, it is not easy to provide an exact, concise definition for Web Services. A simple Google web search on the definition of Web Services yields a large number of them, and each places different emphasis on certain distinct characteristics. We list three of them here that roughly complement one another.

- Web Services are a modular collection of web-protocol based applications that can be mixed and matched to provide business functionality through an internet connection. Web Services use standard Internet protocols such as HTTP, XML and SOAP to provide connectivity and interoperability between companies [1].
- A Web Service is a component that runs on a Web server and allows client programs to call its methods over HTTP. Each method on the component appears as a URL and may return data (perhaps an XML document) and accept parameters. This technology is based on the open SOAP specification, so now our server-side components can be available to virtually any client, regardless of language or platform [2].
- A Web Service is an application or business logic that is accessible using standard Internet protocols. Web Services combine the best aspects of component-based development and the World Wide Web. Like components, Web Services represent black-box functionality that can be used and reused without regard to how the service is implemented [13].

In a nutshell, Web Services use the World Wide Web as the gigantic platform for computing, so that all the service (or components) requests and fulfillments will be carried out on the platform by simply shifting the messages exchanging activities from a low layer of communications infrastructure to a higher layer that performs program logic.

As described by Short [7], Web Services encompass: Discovery, Description, Message Format, Encoding and Transport. For background information, the roles of the five building blocks are briefly discussed below.

Whenever the client application wants to access certain functionality exposed by a Web Service, the Discovery process finds the location of the remote Web Service. Discovery can be facilitated via a centralized directory such as the increasingly popular UDDI (Universal Description, Discovery and Integration Service). The infrastructure that supports UDDI is composed of a set of registries and registrars, where a registry contains a full copy of the UDDI



The Services Side includes all of an enterprise's service components and application objects already existing internally and externally. These components and objects are described by WSDL, published on UDDI servers and are made accessible to users via SOAP technology. The Client Side includes server managers, normal (client) users and other computing systems. The Server Side collects and records all the provided services/applications on servers and dynamically generates an integrated interface for users when a service is requested. The requested service can then be called into action according to the mapping between the selected interface item and the actual record of the stored service item.

For our purpose, as shown in Exhibit 3, the application server of the integrated platform is organized into a hierarchy of four layers: the interface layer, the enterprise objects layer, the physical layer and the transport transactions layer.

#### The Processing Mechanisms of the Proposed Platform Architecture

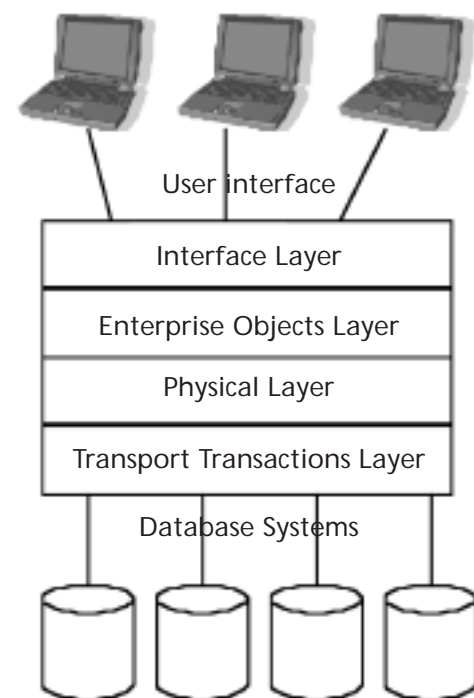


EXHIBIT 3. A LAYERS HIERARCHY OF AN APPLICATIONS SERVER

The processing of the proposed system is sequenced into six mechanisms: *the exploring and Web Services record, action record, process record, procedure record, modeling record and interface record* mechanisms. Each one of the six mechanisms includes five possible operations: Defining, Saving, Querying, Modifying and Deleting. For example, the "action record" mechanism includes action record defining, action record saving, action record querying, action record modifying and action record deleting operations.

The basic principle is to build sequentially the six mechanisms, starting from recording the characteristics and behavior of the provided Web Services to manipulating the records of action, process, procedure and modeling, and eventually to handling records of interface via a mapping scheme. This will lead to the generation of a user-friendly interface for the integrated platform so that users can easily obtain the desired services via the Web.

#### Design and Analysis of the System Components

To validate the proposed framework for applications and platform integration discussed in this paper, a case study [12] has been implemented and completed successfully. The study first illustrated the analysis and design of the needed system components: services integration components, message delivery components, procedure agent components, interface mapping components and others. The activities of the identified components were then classified using a three-layered approach: General Classification, Services Integration Classification and Interface Mapping Feedback Classification. In the Uniform Classification, interface templates were created separately for administration interface, agents and information entities. Under both the Services Integration Classification and Interface Mapping Feedback Classification, components were classified into the classes of administration interface, message delivery, agents, information entity and services integration, respectively, and the only difference between the two classifications resides in different calling modes.

*Web Services use industry standard protocols with universal support to provide a simplified mechanism to connect business applications and to exchange data regardless of the locations, technology or platforms.*

#### Conclusion

The emergence of Web Services technology has made adoption of the Internet as the new computing platform a reality. Web Services use industry standard protocols with universal support to provide a simplified mechanism to connect business applications and to exchange data regardless of locations, technology or platforms. Employing Web Services can lead to many business and technology benefits including delivering platform and technology independence, improving business process efficiency, simplifying the complexity of integration design, reducing the cost of integration and enjoying a wide choice of middleware vendors. In addition, Web Services can remediate the pitfalls of the current implementation of Enterprise Applications Integration (EAI) by adopting the popularly accepted open protocols to standardize processes of communication, description and discovery.

Our proposed integration framework based on Web Services allows IT administrators to employ an interface template to easily integrate intra- and extra-enterprise applications and services without additional coding development effort, and then authorize the developed interface to field or customer users to execute effortlessly their desired requests. Furthermore, the developed framework reduces the impact of change, and response to changes can be automated. The design flexibility of the proposed integration framework leads to broad applicability not restricted to a single information management model; it can help revise or improve business models.

Web Services technology will continue to evolve. We expect that the inclusion of the related growing technologies such as WSEL and BPEL4WS (jointly developed by BEA, IBM and Microsoft) will be easily incorporated into our proposed integration framework.

#### References

1. AZTEC glossary, 2004. <http://www.aztec.soft.net/glossary.htm> Accessed online on (date: December 26, 2004)
2. DevX.Com, Jupitermedia Corp. [archive.devx.com/free/press/2000/030100.asp](http://archive.devx.com/free/press/2000/030100.asp) Accessed online on (date: December 26, 2004)
3. Eriksson, H.-E., M. Penker, B. Lyons, and D. Fado, *UML 2 Toolkit*, Indianapolis, Indiana: OMG Press, Wiley Publishing, Inc., 2004.
4. Linthicum, D.S. *Enterprise Application Integration*, Boston, Massachusetts: Addison Wesley, 1999.
5. *Ovum Evaluates: What is Enterprise Application Integration? 2001*. <http://www.ovum.com/go/content/005004.htm> . Accessed online on (date: June 14, 2003)
6. Pender, L. "Damned if You Do ... - Will integration tools patch the holes left by an unsatisfactory ERP implementation?" *CIO Magazine*, September 15, 2000. <http://www.cio.com/archive/091500/erp.html> Accessed online on (date: May 15, 2005)
7. Short S. *Build Web Services for the Microsoft.NET Platform*. Redmond, Washington, Microsoft Press, 2002.
8. UDDI, *UDDI Specification*, ver. 3, 2005. <http://www.uddi.org> Accessed online on (date: May 15, 2005)
9. W3C, *SOAP Specification, 2003*. <http://www.w3c.org/tr/soap> Accessed online on (date: May 15, 2005)
10. W3C, *WSDL Specification, 2001*. <http://www.w3c.org/tr/wsdl> Accessed online on (date: May 15, 2005)
11. WRQ Verastream Whitepaper, *Integrating your Legacy Hosts: A Critical Step for B2B'S Success, 2001*. <http://www.wrq.com> Accessed online on (date: Dec 14, 2002)
12. Yang, H.M., F.V. Lu, S.Q. Huang, and C.B. Wang. *Development of Web-Services Based Enterprise Application: An Integration Framework*. A Completed Working Paper, November 2004.
13. ZEROONESOFTWARE Glossary 2004. [www.zeroonesoftware.com/glossary.html](http://www.zeroonesoftware.com/glossary.html) Accessed online on (date: December 26, 2004)

# wireless networks and security issues

## Wireless Networks and Security Issues

Farok Vakil, The Peter J. Tobin College of Business, St. John's University

### Abstract

The evolution of different wireless standards (known as 802.11 standards) in the late 1990s produced an astonishing volume of global demand for wireless networks. Many organizations and residential computer users have embraced the convenience and mobility of the wireless networks. In this article, the various wireless standards currently in use will be examined. Furthermore, different characteristics of these standards will be investigated. Finally, the security issues of wireless networks affecting corporations and residential setups will be addressed.

### Introduction

Recent developments in wireless technology have created an enormous opportunity for consumers and corporations alike. Consumers in general have moved away from wired networks and embraced the new technology. Wired networking—which is based on sharing data, hardware and software among connected computers by using cable or wire—is also known as Local Area Networks (LANs). On the other hand, in wireless networks, computers are connected by transmitting radio waves or, less commonly, infrared light. Generally, wireless networks can be constructed in two ways: 1) by using a peer-to-peer setup or 2) by using an access point.

In the peer-to-peer setup, several computers—each equipped with a wireless network interface card—can communicate directly with other wireless enabled computers. Although each connected computer can use file sharing and printing resources, it may not be able to access wired LAN resources, unless one of the computers with special networking software (Software Access Points) is used as a bridge to the wired LAN. In the second type of setup, a wireless network uses an access point or a router. Here the access point acts like a hub, providing connectivity for the wireless computers, which are equipped with network interface cards. It also can connect a wireless LAN to a wired LAN and create what is known as a Hybrid Network, allowing wireless computer access to the wired LAN resources, such as existing Internet connectivity or print and file sharing. Since the price of a router or access point has declined substantially during the last few years, these days the typical home or office wireless network is based on the

router. In the wireless LAN (WLAN) several mobile computers are connected to the access point or router. The access point has two functions: 1) to control the operation of a wireless station (through transmission power, for instance) and 2) to link mobile users to wired LANs [4].

Using a home or office wireless network, one can connect several computers to share hardware, software and resources—such as stored files, photos, printers and an Internet connection. Furthermore, from each computer, one can print stored files, photos or documents by sharing a single printer attached to just one computer (i.e., a printer server)—all without using cables running throughout the physical space. By using a wireless network a home or office can have the ability to share a single high-speed broadband cable or DSL connection among several computers without significant reduction in the connection throughput (speed). Indeed, wireless networks can be expanded easily to serve a dozen users or more. However, one must make certain that the equipment included within a wireless network is Wi-Fi (Wireless Fidelity) certified. In order to check if a particular device is Wi-Fi certified, one can visit the Wi-Fi Alliance web site [2]. Also, through this web site, one can make sure that various equipment one purchases is compatible with equipment already installed in a system.

In workplaces and offices, the built-in flexibility of wireless networks has created a valuable connectivity between the mobile salespeople and behind-the-scenes workforces. In today's dynamic business environment, a wireless network provides a superb and affordable means of instant communication. Important characteristics of various wireless standards are summarized in Exhibit 1.

### Security

Security has been one of the major problems with wireless networks. By default, a wireless network is designed to provide easy access. Generally, wireless networks need to announce their existence so that potential users can link up and use the services provided by the network. However, radio signals traveling through the open atmosphere can be intercepted by individuals in the vicinity who have access to wireless devices with the right software for interception. As a result, if not properly configured, the signals can be located and monitored quite readily.

*“Using a home or office wireless network, one can connect several computers to share hardware, software, and resources - such as stored files, photos, printers, and an Internet connection.”*

EXHIBIT 1. WIRELESS STANDARDS

Standard	Data Throughput (Theoretical Speed)	Data Throughput (Practical Speed)	Frequency Band	High Interference
802.11	Up to 2 Mbps	1 Mbps	2.4-2.4835 Ghz	Yes*
802.11b	11 Mbps	6 Mbps	2.4-2.4835 Ghz	Yes
802.11g**	54 Mbps	27 Mbps (g-only networks)  9-13 (b/g combination networks)	2.4-2.4835 Ghz	Yes
802.11a	54 Mbps	27 Mbps	5.12-5.25 Ghz  5.47-5.725 Ghz  5.725-5.825 Ghz	No
802.11 super a + g	108 Mbps	54 Mbps	2.4 GHz 802.11 b/g or dual band 2.4/ 5 GHz 802.11 a/b/g	No
802.11h	54 Mbps	27 Mbps	Same as 802.11a	No

\*Since 2.4 GHz has only three non-overlapping channels, users may share this frequency with their neighbors' wireless networks or other wireless devices, such as cordless phones and microwave ovens. These wireless devices are competing for space on the 2.4 GHz frequency. Consequently, a high interference rate might be observed. On the other hand, 5GHz wireless has 16 non-overlapping channels (802.11 super a + g) and 24 non-overlapping channels (802.11a and 802.11 h), so the possibility of interference is very low.

\*\* (802.11g) devices work with (802.11 b) devices albeit at lower speed.

Another security problem of the wireless network involves rogue access points. Any local computer user can purchase an access point, and potentially connect it to the corporate network or a home network in the vicinity without authorization. Rogue access points deployed by end users pose great security risks.

In addition, traffic analysis and eavesdropping present a third common problem. Not all the wireless standards available in the market provide protection against attacks

that passively observe traffic. The main risk is that these standards do not provide a way to secure data in transit against eavesdropping. A computer user equipped with a wireless network analyzer could easily capture unprotected Internet protocol radio transmission. A very generic solution available is a Wired Equivalent Privacy (WEP) system. However, it will protect only the initial contact with the network. When WEP is employed, data is not encrypted or authenticated, leaving an attacker with the opportunity to

*“In workplaces and offices, the built-in flexibility of wireless networks has created a valuable connectivity between the mobile salespeople and behind-the-scenes workforces.”*

disrupt transmissions.

Generally, home computer users are not security experts and may not be aware of the risks posed by wireless networks. Furthermore, wireless security is a work-in-progress, with evolving standards. However, home wireless network users can take several steps to reduce the security risk posed by the wireless transmission of data and sensitive information by changing their router default setups [3]. The following steps can be taken by using the router manual.

1. Change the default password of the router. Routers use a preset password initially, and it's easy for an unauthorized user to figure out.
2. Disable remote router access. This will keep anyone from accessing your router from a remote location through the Internet. However, it does not prevent local wireless users from accessing your wireless networks.
3. Change the Service Set Identifier (SSID) in order to disable broadcasting. SSID is the ID of your own local wireless network and it reveals the network to anyone in the vicinity who is using a wireless-equipped computer. All wireless routers come with a default SSID that you should change. You'll need to remember it in order to set up other wireless clients on your network.
4. Turn on your router firewall. Routers usually have their firewall turned on by default, but make sure that's the case. Also, enable any additional firewall features, such as the ability to block anonymous Internet requests. To increase your security, run a software firewall on every PC on your network.
5. Enable data encryption. Data transmitted by a wireless network can be read by anyone who picks it up, unless it's encrypted. Wireless routers have encryption capabilities. Wi-Fi Protected Access (WPA) is the standard that offers the most protection for data. Some routers are equipped with 'WPA Pre-Shared Key' (WPAPSK). This value will provide higher security for home or small-business networks.
6. Enable MAC filtering. The Media Access Control (MAC) address is a unique identifying number assigned to each network device. Enabling MAC filtering in your router improves your network's security by accepting transmissions only from PCs with specific MAC addresses. You can also prevent certain MAC addresses from accessing the network.

When sensitive information is transmitted over the Internet, a few more precautionary steps must be taken in addition to the above-mentioned measures. Wireless network users, just like other Internet users, must make sure that they are using a secure connection. Using Internet browser windows, there are two ways to recognize if the web site one is accessing is secure:

1. You should see that the "http" in the address line is replaced with "https" and
2. You should see a small padlock (resembling a small lock) in the status bar at the bottom of the browser window.

#### Conclusion

Wireless technology has been evolving rapidly during the past few years. Faster devices and more secure equipment are being introduced into the market every day. "Once the stuff of science fiction, wireless networks have rapidly become an integral part of most organizations' network structure. In fact, for small offices and many homes, wireless networks have evolved to be the only network structure" [1]. Recent developments in wireless networking have made the mobility and convenience of the wireless network even more appealing to the average computer user.

#### References

1. Goldman, J. and P. Rawles. *Applied Data Communications: A Business-Oriented Approach, 4th edition*. Danvers, MA: Johns Wiley & Sons, Inc., 2004.
2. [http://www.wifi.org/OpenSection/certified\\_products.asp?TID=2](http://www.wifi.org/OpenSection/certified_products.asp?TID=2) (accessed May 1, 2005).
3. Miastkowski, S. "How to Build a Safe, Secure Network," *PC World*, May 2004. [www.pcworld.com/news/article/0,aid,115066,pg4,00.asp](http://www.pcworld.com/news/article/0,aid,115066,pg4,00.asp)
4. Panko, R. *Business Data Networks and Telecommunications, 5th edition*. Upper Saddle River, NY: Prentice Hall, 2005, p. 218.

# using probability analysis to define a defendant's intent level in a criminal case

## Using Probability Analysis to Define a Defendant's Intent Level in a Criminal Case

Mark I. Marpet, The Peter J. Tobin College of Business, St. John's University  
Christopher M. Farella, The Law Offices of Robert G. Stahl, LLC

#### Abstract

Using a real-world criminal case, this paper demonstrates how probabilistic analysis can be used to determine whether a hypothetical, Reasonably-Prudent Individual, with full information concerning the consequences of driving while intoxicated at a specific Blood Alcohol Content (BAC) level, could reasonably conclude such driving would probably cause injury.

#### Introduction

This analysis lies at the intersection of the law and decision science. There are interesting similarities between these two disciplines, starting with the fact that their shared raison d'être is to make the best possible decisions in the face of uncertainty and often incomplete or conflicting information. McClave, et al. [10], to cite one example, make a compelling analogy between the structure of the judicial trial and of statistical hypothesis testing, emphasizing the similarities between, on one hand, the relationship between Type I and Type II errors in statistical hypothesis testing and, on the other, the relationship of the errors in convicting innocent parties versus setting guilty individuals free.<sup>1</sup> Another significant similarity is that both disciplines require the decision maker/trier of fact to assume a specific, similar persona: for decision-making, it is the Rational Decision-Maker; for the trier of fact, it is the Reasonably-Prudent Individual.

In spite of the common elements between the disciplines, the law—at least in the eyes of the decision science community, and at the risk of overgeneralization—does not make full use of quantitative decision-making tools. Finkelstein [8] introduces *Quantitative Methods in Law* with the following:

Legal thought is threaded with assessments of probability and descriptions of a statistical nature, but the data underlying these ideas seldom are collected and rarely are used for other than general impressions. Even far-reaching and debatable questions of fact are decided on a thoroughly subjective basis and in panoramic terms that are congenial to moral or political judgment.

This paper, which stems from a real-world criminal case, applies probabilistic analysis to determine whether a hypothetical, Reasonably-Prudent Individual, with full information concerning the consequences of driving while intoxicated at a specific Blood Alcohol Content (BAC) level, could reasonably conclude such driving would *probably* cause injury.

#### The Accident

A five-car accident occurred in New Jersey on an interstate highway at a point where the roadway narrows to two lanes at the beginning of an over-water bridge. Traffic had been stopped because of ongoing construction. It was night, but the area was lit by construction-zone illumination. Driver A, proceeding towards the bridge, did not stop, colliding with the car ahead of him, which proceeded, in billiard-ball-like fashion, to involve the other vehicles. The driver of the vehicle directly in front of driver A's vehicle sustained the most serious injury: a broken ankle. Driver A, who was returning from dinner, reportedly gave indication of intoxication, and was tested for alcohol by means of a blood test. The State Police reported the driver's Blood Alcohol Concentration to be 0.191%, almost twice the legal limit (at that time) in New Jersey.

#### The Law

The matter was presented to a Grand Jury, which returned, among other charges: Second Degree Aggravated Assault and Fourth Degree Assault by Automobile. The former charge is far more serious, carrying with it a significant and mandatory jail sentence. Both charges relate to defendant driver A's causing injury to the driver who had been in the vehicle that driver A's automobile had rear-ended. The difference between the two charges relates to driver A's intent.

For a trier of fact to convict a defendant on the fourth-degree charge, it must find (beyond a reasonable doubt) that the defendant driver drove recklessly, and that that reckless conduct caused the victim's injury. The fact that Driver A plowed into a stopped car does not *per se* prove recklessness. Driver A might have been distracted, for example, by an insect in the car at the instant of the

accident. In the matter under study, an element of recklessness could be inferred by the trier of fact from the fact that Driver A had been intoxicated.

For a trier of fact to convict a defendant on the second-degree charge, it must find everything that was found in the fourth-degree charge and, in addition, must find (again, beyond reasonable doubt) that the defendant driver manifested extreme indifference to the value of human life.

New Jersey case law casts the difference between Reckless Assault and Aggravated Assault as follows. In *State v. Curtis* [14], the court wrote:

Recklessly under circumstances manifesting extreme indifference to the value of human life has been construed as being distinct from mere recklessness in that under the former there is a probability of injury while under the latter there is a possibility.

**Who's Point of View?**

Who should judge whether or not an injury was caused "recklessly under circumstances manifesting extreme indifference to the value of human life?" The *who* is clear: not any given person but, rather, the hypothetical Reasonably-Prudent Individual. The trier of fact, a jury constructed to aggregate the thoughts and views of its members (or, in some cases, a judge), forms a surrogate for that Reasonably-Prudent Individual. How to make the judgment is also clear: prospectively. In other words, our hypothetical, Reasonably-Prudent Individual should look at the totality of circumstances (here, that Driver A had been driving an automobile and had a Blood Alcohol Concentration of almost twice the legal limit) and ask, what is the chance that Driver A will cause injury to another? Is it not-possible, possible or probable?

Two things are clear: first, the decision concerning whether or not "extreme indifference to human life" is manifested should never be evaluated retrospectively on the basis of the fact that injury resulted from Driver A's actions. Second, the question of whether Driver A manifested extreme indifference should never be evaluated on the extent or nature of any injury. For example, Driver A's actions would not manifest more (or less) indifference to human life had the accident broken, say, the victim's pelvis, rather than his ankle.

**"Possible" vs. "Probable" (or "a Probability") vs. "Probability"**

The meanings of the terms *possible*, *probable* and a *probability* are the same in both the law and in common usage. Probability has a specific mathematical definition that relates to the other words. Redmayne [13] discusses the link between probability and "wider theories" of epistemic justification.

Something is *possible* if it is not infeasible or impossible. Even very rare occurrences are possible. It is *possible* (but extremely unlikely) for a person who buys or otherwise obtains a lottery ticket to win the lottery. Conversely, it is not possible—impossible—for a person who does not own a lottery ticket to win the lottery.

Something is *probable* (*i.e.*, a *probability*) if it is as-or-more likely than not to occur: an injury is probable if you leap out of a second-story window. That is the sense of the phrase "a probability of injury" in *State v. Curtis*.

The uses of the words *possible* and *probable* in mathematics are similar to their usage in the common and legal lexicon. Probability has an explicitly mathematical definition: the quantification of chance, a numeric expression of the strength of a possibility. Probability is expressed as a number between zero and unity: the probability of a head (or a tail) for a fair coin fairly tossed is, for example, one-half: P(Head) = P(Tail) = 1/2 = 50%.

It is important not to confuse the common/legal-usage of something being a probability—an event that will more likely than not occur—with the mathematical concept being discussed here: the probability of something (expressed as a number).

**Probability**

Probability is the branch of mathematics that deals with the study of random phenomena. It originated hundreds of years ago when mathematicians [6] systematically studied games of chance to increase their tax-free earnings. It is key to understanding probability that, while the pattern of outcomes, including the frequency of the various outcomes, can be known with great precision, no specific outcome can be predicted in advance. In other words, if we were to toss a "fair coin," say, 10,000 times, we would realize close to 5,000 heads and close to 5,000 tails. It would be impossible, however, to predict in advance the outcome of the first or third or 386th or any other specific toss.

*A driver with a BAC [Blood Alcohol Content] of 0.04% or more will, cet. par., have a higher risk of causing an accident and, therefore, a higher value for the λ parameter.*

Year	Fatalities
1990	24,092
1991	22,385
1992	21,387
1993	21,566
1994	21,997
1995	22,423
1996	22,505
1997	21,989

The occurrence of any specific automobile accident cannot be predicted in advance. The pattern of automobile accidents is, however, predictable to a high degree of certainty. For example, in the last few years data has become available which tells us that the number of people killed each year in automobile accidents in the United States is about 22,000 [15]. The consistency is rather remarkable when you consider that the underlying accidents that generate those fatalities are generally unrelated. This probabilistic consistency is one important basis that insurance companies use to set premiums that allow them to simultaneously remain in business and generate a return.

**The Probability Distribution of Automobile Accidents**

Disparate phenomena often have the same mathematical underpinning. A fairly well-known example is that many types of measurements tend to distribute themselves in a bell-shaped pattern around the group average, called the Normal Probability Distribution. This includes the life span of people, IQ-test results and the gas mileage within a fleet of similar cars. The distribution of the number of automobile accidents in a given number of miles follows the Poisson Probability Distribution [12].

**The Poisson Distribution**

The Poisson Probability Distribution that underpins the occurrence of automobile accidents is commonly seen in situations that are characterized by the number of random events that can occur in a predefined "interval," e.g., the probability distribution of the number of flat tires in 100,000 miles of driving, or the number of automobile accidents in a region in a month.

λ, the single parameter that characterizes the Poisson, is the mean number of occurrences of the event of interest in the period of interest.<sup>2</sup> It represents the background risk of an event's occurrence. In this matter, it represents the background risk of an accident. It will vary over time and circumstance. For example, the background risk and, thus, the value of λ, will vary with the age and gender of the driver. The background risk will also vary with a driver's Blood Alcohol Concentration. A driver with a BAC of 0.04% or more will, *cet. par.*, have a higher risk of causing an accident and, therefore, a higher value for the λ parameter.

The increase in risk relative to the background risk of a sober driver is called the Intoxication-Hazard Multiplier or, more simply, the Hazard Multiplier.

**The Poisson Process**

Underpinning the Poisson Distribution is the Poisson Process, a chance-generating mechanism that has an extremely simple axiomatic structure:

- For a small trip, the probability of an automobile accident is proportional to the length of the trip. (The meaning of "small" will be explored below.)
- Accidents are independent.

The Poisson Process is useful here for two reasons. First, the Poisson Process assumptions are congruent with the way automobile accidents occur. Secondly, because the trip distance will be shown (below) to be "small" with respect to the average-mileage-between-accidents constant (1/λ), the probability of an accident can be estimated directly from the proportionality axiom.

**Is the Poisson Process appropriate to model automobile injuries?**

Mathematical models are an abstraction—a simplified version—of reality. The test of whether or not a mathematical model is appropriate for use in a given situation is whether or not the mathematical model produces results realistic enough to be of use. It is completely inappropriate to suggest that mathematical models, unless they are perfect mirrors of reality, are worthless for the purpose of casting light on a courtroom issue. That would be like asserting that eyewitness testimony, unless it was a *perfect* mirror of what had been witnessed, would be worthless for the purpose of determining what had in fact occurred.

Here, it is clear that the probability distribution of automobile accidents follows Poisson exactly. Injuries—because more than one injury can occur per accident—follow Poisson only approximately. Compared with the Poisson Process and Distribution, one can conceive of normatively realistic (and more complex) methods of modeling automobile-accident-generated injuries. For example, one can explicitly account for the fact that any given accident can cause no, one, or multiple injuries. Such a situation can be modeled by a compound probability distribution. That is, the occurrence of accidents can be modeled by a Poisson distribution, and the number of injuries in each accident can be modeled by, say, an

... [T]he Poisson Probability Distribution underpins the occurrence of automobile accidents.

Using Probability Analysis to Define a Defendant's Intent Level in a Criminal Case

16

empirically derived probability distribution. If one were to create a simulation-based model, one would use exactly that approach.

In spite of this, the Poisson distribution here provides a reasonable predictor of injury rates because the question that is here being addressed is "Was the injury to the driver of the vehicle in front of Driver A's car probable?" For that question, because we are exploring the circumstances surrounding a single injury, and assuming the average accident rate ( $\lambda$ ) is appropriately selected, the Poisson Process will be an appropriate fit to this situation.

#### **What is "small" in the context of the Poisson Process?**

One of the key axioms of the Poisson Process is that, for small intervals, the probability of an occurrence is proportional to the length of the interval. Here, that means that the probability of causing an injury with a motor vehicle is proportional, for small trips, to the length of the trip. (The proportionality constant,  $\lambda$ , is a function of, among other things, whether the driver is intoxicated and, if so, the extent of intoxication.) The overall passenger-car-caused injury rate in the United States in 1997, the year of interest for Driver A's accident, was 156 injuries per 100,000,000 miles. That is, the probability of an injury by an "average" driver is 0.000,001,580 per mile. The reciprocal of that value is the mean number of miles traveled between injuries: slightly over 640,000 miles. For a driver, typical except for the fact of being intoxicated at the level of driver A, we can expect an injury every 16,000 miles. Thus, any automobile trip that can be accomplished in a day (or two or three) is clearly small in the context of the Poisson Process.

#### **Admissibility-of-Expert-Testimony Issues**

In the courtroom, only experts are permitted to testify as to their opinion. Such opinion testimony is permitted only if the specialized knowledge of the expert is of use in helping the trier of fact "understand the evidence or determine a fact in issue" [7]. In order to be allowed to testify, the opinion must not be a "net opinion," i.e., an opinion not grounded in the science and art of the expert's field. Opinions given must be based upon sound and accepted methodology [5]. Among the evidence for such soundness is the acceptance of the opinion in the field of the expert. This can be shown by its appearance in peer-reviewed journal articles or research monographs, authoritative textbooks or other gold-standard publications.

In the matter under discussion, we believe that the use of probabilistic analysis to determine an actor's intent is novel. On the other hand, we know that the methodology and data that underpin the analysis come from a number of disparate but high-level and reliable sources. Here are the separate elements that, together, constitute the inputs to the analysis; we will take each of these in turn.

- The Poisson Process and Poisson Distribution
- The applicability of the Poisson Process to automobile accidents and automobile-accident-caused injury
- The data that is used to determine the number of automobile accidents and/or injuries. This value forms the numerator of any accident- or injury-rate fractions used in  $\lambda$
- The data on the number of miles driven per year. Elements of these data form the denominator of any accident- or injury-rate fractions used in  $\lambda$
- The hazard multiplier, the multiplier in  $\lambda$  to account for the fact that the driver is intoxicated

**The Poisson Process and Poisson Distribution.** Starting with the very simple Poisson Process assumptions, anyone with reasonable knowledge of probability and differential equations can derive the Poisson Probability distribution. Thus, the Poisson process and distribution is transparent; it should need nothing further to justify itself to the court. It is worth noting that the Poisson process and distribution are over 100 years old and are presently taught in most stochastic-processes and probability-and-statistics courses; they are bedrocks in the foundation of probability theory.

**The applicability of the Poisson Process to automobile accidents and automobile-accident-caused injury.** It does not automatically follow that, because the Poisson Process and Distribution are beyond question, their application in modeling automobile accidents is also beyond question. The utility of Poisson in this context is an issue that must be addressed independently of the issue of the validity of Poisson itself. It is not difficult to see that individual automobile accidents follow the two Poisson-process axioms; as a consequence, it is appropriate to use the Poisson distribution. The first axiom states that the chance of an accident is proportional to the trip length. If that was incorrect, then it would be meaningless to use the everyday statistics that are used to describe vehicular hazard: accidents per million miles driven, accidents per day, and so forth. The second axiom, that accidents are independent, can be guaranteed simply by counting any chain of accidents, e.g., a 50-car pile-up, as one accident.

The increase in risk relative to the background risk of a sober driver is called the Intoxication-Hazard Multiplier or, more simply, the Hazard Multiplier.

Using Probability Analysis to Define a Defendant's Intent Level in a Criminal Case

17

The application of the Poisson distribution to injury data is not as straightforward. As to the Poisson process' first axiom, it is certainly true that the chance of an injury is proportional to the exposure (but except in the case of single-injury accidents, injury accident). The second axiom—that injuries are independent—is not generally correct, as a single accident can generate multiple injuries.<sup>3</sup> Since the accident under study generated only a single injury, this did not turn out to be problematic. (That might not always be the case.)

**The data that is used to determine the number of automobile accidents and injuries.** The data comes primarily from two sources: a census of all fatal accidents (Fatal Accident Reporting System (FARS)) and a sampling of non-fatal accidents in the United States (General Estimates System (GES)). Both datasets are collected by the National Highway Traffic Safety Administration. (Details are described in the endnotes.) The acceptability aspects of these data are rather obvious; they are explicitly used in the setting of Federal Regulatory Policy. There exists a large body of information on how the data was generated, and much of the raw data is available to the public. Furthermore, the data is revised over time if any problems are discovered or as refined estimates are developed. The short of it is that this data is used for decision making on an everyday basis by the Federal Government; on that basis, its use as an input in an analysis directed towards litigation should simply not be problematic.

**The data on the number of miles driven per year.** In a similar manner, the miles-driven data is collected using a sampling system by the Federal Highway Administration. It is seriously scrutinized by Congress, as this dataset forms an input to the apportionment of highway funds. Importantly, there are robust reality checks on this data, e.g., the quantity of motor fuels sold in a geographic area has to correlate with miles driven. Since use taxes are collected on motor fuels, the fuel consumption data is similarly closely tracked.

**The intoxication-hazard multiplier.** The basis for hazard-multiplier adjustments to  $\lambda$  comes primarily from Borkenstein's seminal Grand Rapids study [3]. Prosecution experts frequently use this study as the basis for their determination of how hazardous a given level of alcohol in the blood makes a driver, e.g., Driver A is 40 times more likely to be involved in an accident as an unimpaired driver. As such, the use and legitimacy of the Grand Rapids study is typically hotly contested by defendants. (As a practical matter, it is simply not contested by prosecutors.) In order to keep things simple, attorneys for the defendant did not ask

of their alcohol expert either criticism of the prosecutor's hazard-multiplier analysis or for an independent assessment of the hazard ratio. Defense simply accepted, i.e., offered to stipulate to, the prosecution expert's hazard. Below, we will see that, if Driver A were to have driven 16,000 miles per year with the BAC that he had at the time of his accident, he would generate, on average, one injury every year. That is clearly unacceptable, and it is the reason that driving while under the influence of alcohol is strictly proscribed by law.

#### **What Injury-Rate Data is Appropriate?**

There are any number of accident rates that can ostensibly be used to reflect the risk of a driver in Driver A's risk cohort injuring someone. Briefly, the injury rate is determined by multiplying a baseline injury rate for an unimpaired driver—determined by statistical data gathered by the Federal government—by a hazard multiplier, characterizing the effect of alcohol on driving ability. The question of whether to use New Jersey or National data, male-driver or mixed-gender data, age-specific or age-aggregated data has no "correct" answer. The finer the demographic category, the more "statistically representative" the baseline is of the defendant. The coarser the demographic category, the closer one gets to our hypothetical Reasonably-Prudent Driver. In the case under study, the differences between relevant categories were slight indeed, and the probability of an accident was not particularly affected by which baseline was utilized. (It is not clear that this is always the case.) We have chosen to utilize National accident statistics on normative grounds: we do not think that the standard of conduct for the defendant driver should be different from the National norm.

#### **NHTSA Accident Statistics**

Accident rate data comes from the Department of Transportation's National Automotive Sampling System's General Estimates System [11].<sup>4</sup> GES estimates that 2,378,000 passenger-car occupants were injured in 1997. Accident data from police motor-vehicle accident reports is the starting point of Department of Transportation's GES data-collection efforts. Much of the data is summarized in the annual publication *Traffic Safety Facts*.

#### **Federal Highway Administration (FHWA) Mileage Statistics**

Vehicle Miles Traveled is determined by aggregating data collected by State Highway Departments, and summarized in FHWA summary Table VM-1. The use of the aggregate data is described in the literature [2]. Passenger cars traveled

1,502 billion miles [1].<sup>5</sup>

### The Alcohol Multiplier

In order to estimate the increased hazard that inheres to consuming alcohol and driving, one needs to be able to estimate the distribution of blood-alcohol concentration (BAC) in drivers on the road. By comparing the BAC distributions of accident-involved drivers with the population of non-involved drivers, the relative risk of an accident at a given BAC can be determined. The Grand Rapids Study was and is a (if not *the*) primary data source in this endeavor.

In the Grand Rapids Study, two populations were compared: accident-involved drivers and non-accident-involved drivers. Non-accident-involved drivers were selected according to a sample design. Essentially, the landscape of Grand Rapids, Michigan, was subdivided into discrete “blocks” and the control group was sampled at times and places designed to replicate the place-and-time distribution of accidents that occurred in the years 1959-1962. Because both the questionnaire answers and breath samples were held in the strictest confidence (neither could be used as evidence against the driver), the response rate was reported to be approximately 97%.

In the matter at hand, the alcohol expert for the prosecutor, a biological psychologist, estimated that the 0.191% BAC that was exhibited by the defendant at the time of the BAC test was equivalent to a 0.20-0.22% BAC at the time of the accident. The expert writes: “the relative risk for a crash is conservatively estimated at approximately 40 times that compared to being in a sober state.” As noted above, the defendant had accepted (and this paper will accept) at face value the hazard multiplier value of 40 that was proffered by the State.

### Analysis

As described above, the Poisson-process determined probability is given by the rather simple formula:  $P(\text{injury in } t) = \lambda \cdot t$ . The actual analysis is rather simple, and consists of two steps. First,  $\lambda$  must be calculated. Second, to determine the probability, it must be multiplied by  $t$ . (In this analysis,  $t$  is the trip length: about 20 miles.)  $\lambda$  is built-up from two factors: the baseline injury risk multiplied by the hazard multiplier. The baseline risk, the mean chance of a passenger-car-based injury, is the quotient of the number of passenger-car injuries divided by the number of passenger-car miles.

### The Probability of an Accident, Based Upon Most-Likely Estimates

As we noted above, 2,378,000 passenger-car occupants (including drivers) were injured in 1997, and passenger cars drove a total of 1,502 billion miles. The quotient, the baseline risk, is calculated to be 158 injuries per hundred-million vehicle miles (0.000,001,580  $\frac{\text{accidents}}{\text{mile}}$ ). Multiplying that by the hazard multiplier (40, dimensionless) and the trip length (20 miles) gives the odds of an accident of just over one in a thousand (0.0013)

Thus, if the defendant driver had been in the possession of complete information and had, before turning the ignition key, been able to put himself in the Rational-Decision-Maker/Reasonably-Prudent-Person persona that we discussed above, and had asked, “What is the chance that I will become involved in an accident?,” and further assuming that he was able to accurately assess his level of inebriation (in terms of a hazard multiplier), and in general going through the same logic that is seen in this paper, he would assess his chances of an accident at just above one in a thousand. At the risk of writing what should be plain, Driver A's chance of an accident is far, far less than probable, far less than a probability. To get the hazard to rise to the level of a probability, the hazard would have to be about 400 times larger.

### Worst-Case Probabilities Based upon Confidence-Interval Estimates of Accident-Generated Injuries.

Because the GES data that we used to estimate the baseline probabilities is based upon a sample of accidents, it is subject to sampling error. By using confidence limits to encapsulate the sampling variability in our measure of the baseline risk, we can effectively eliminate that as a consideration in our analysis.<sup>6</sup> The magnitude of the involved numbers clearly implies that there are simply no not-outlandish perturbations of the figures that could transport these very low probabilities into the land of the probable.

### The probability of injuring someone as given above dramatically overstates the true likelihood of injury.

There are a number of reasons for this. First, the baseline  $\lambda$  is an aggregate of all accidents, including those caused by intoxicated drivers. Secondly, the injury count includes injuries to drivers responsible for their own injury. Let us look at each in turn.

*[T]he relative risk for a crash is conservatively estimated at approximately 40 times that compared to being in a sober state.*

### The baseline $\lambda$ is an aggregate of all accidents, including those caused by intoxicated drivers.

Our analysis has been simple: we took the count of those injured in passenger cars and divided it by the number of miles driven by passenger cars to get what we called our baseline  $\lambda$ . We multiplied that baseline  $\lambda$  by the hazard multiplier. We have been going along as if the numerator of our baseline  $\lambda$ , the injury totals, are generated by not-intoxicated drivers. That is not in fact correct.  $\lambda$  has as its numerator the count of all those injured in passenger cars in 1997. A disproportionate number of those injuries relative to miles driven were generated by those who had a positive BAC. Borkenstein, *et al.* [3], estimate that drivers with BACs at or above 0.05% constituted 3% of the driving population in The Grand Rapids Study but generated 15% of the accidents. They estimate that eliminating all drivers with positive BACs from the road will eliminate about one fifth of the accidents. We made no attempt to correct our baseline  $\lambda$ , as it simply did not make any real difference in the end-result of the analysis, which concerns one defendant's intent level in a criminal case. Utilizing this correction would have lowered the probability of an accident from about 0.0013 to about 0.001, further strengthening our argument.

### The injury count included injuries to drivers responsible for their own injury.

Self-inflicted injury is not relevant in the context of an assault charge. Again, while it is possible to correct for this injury “overcount” and concomitant overestimate of  $\lambda$ , this has not been accomplished; the injury probability is so low relative to it being a *probability*, as the law would require to prevail on the aggravated-assault charge, and given that this correction would actually lower the accident probability, correction for self-inflicted injury seemed not worth the effort.

### Discussion

It is crystal clear that the prosecutor's aggravated assault charge, which required that the accident was probable under the circumstances, was completely untenable. Similarly, the analysis above clearly shows that the hypothesis—that Driver A (whose felonious conduct revolved around the act of drinking too much at dinner) manifested “extreme indifference to the value of human life”—is obviously unsupportable.

It should also be crystal clear that Driver A should never have attempted to drive in his condition. Quite the opposite. This can easily be seen by noting that the mean mileage between accidents for a person intoxicated at the level of

Driver A is the reciprocal of  $\lambda$ : just under 16,000 miles. That is, Driver A's intoxication will cause an injury for every 16,000 miles he drives. In short, driving while intoxicated at the level of Driver A is simply not acceptable behavior. Reckless? Certainly. But manifestly indifferent to the value of life? Certainly not.<sup>7</sup>

### A general rule is implied by the analysis.

That rule is: *Absent exogenous exacerbating factors, it is simply inappropriate to charge aggravation.*

What would the hazard-multiplier have to be to get the chance of injury to the level of a probability, i.e., to around 0.5? It is easy to show that the hazard multiplier would have to be around 16,000.<sup>8</sup>

This orders-of-magnitude difference between the real hazard level caused by drinking too much and the hypothetical hazard level needed to change the nature of the act from reckless to aggravated (as expressed by the hazard multiplier of 15,800) strongly confirms that the acts implied by the two hazard levels are different in kind, and not just in degree. That suggests that a person could not, simply by “ordinary” drinking, create the kind of hazard that would change the very nature of this reckless act to an aggravated act. To cross the threshold between reckless and aggravated conduct, one would have to do something beyond ordinary drinking. (That's why it's called aggravated assault.)

That aggravated acts are different in kind—and not just in degree—from reckless acts was surely what the legislature contemplated when it set significantly stiffer penalties for the aggravated crime *vis à vis* the reckless. That aggravated and reckless acts differ in kind and not just degree is similarly reflected in the New Jersey case-law criterion—the difference between a possibility and a probability—that is the underpinning for our analysis in this paper.

### Generalizing this Analysis

While it is clear that probabilistic analysis can, in the matter under study, give insight into the intent of a Reasonably-Prudent Actor with full information,<sup>9</sup> there are limits to the extent this analysis can be generalized. First, while the analysis was developed in a mathematical context that hinged upon a possibility-versus-probability case-law dichotomy, we do not perceive the analysis to ultimately hinge upon that dichotomy. The showing of a two-and-a-half order of magnitude differential between Driver A's conduct and the degree of recklessness to get to an aggravated act could easily be recast in other than strict

*While...probabilistic analysis can...give insight into the intent of a Reasonably-Prudent Actor with full information, there are limits to the extent this analysis can be generalized.*

possibility-versus-probability terms. Far more limiting, we think, is the fact that this analysis used both a readily available probability model and extensive, readily available data to use within that model. The former—the probability distribution—does not seem to be problematic: first, the Poisson model has significant extensions available that are explored in the field of stochastic processes [6] and, ultimately, mathematical analysis is far more general than the Poisson model. The latter limitation—the availability of relevant data—appears to be generally problematic. It is unclear that manna-from-Washington in the form of relevant and rigorously generated data will in general be available. Parameter estimates not underpinned by hypothetical data are unacceptable;<sup>10</sup> the dataset that was used in the instant analysis would cost millions and take years to duplicate.

References

1. *Annual Vehicle Distance Traveled in Miles and Related Data-1997 by Highway Category and Vehicle Type (Table VM-1)*. Federal Highway Administration (FHWA), October, 1998.
2. *Annual Vehicle Miles of Travel and Related Data: Procedures Used to Derive the Data Elements of the 1994 Table VM-1*. FHWA-PL-96-024, June, 1996, 18.
3. Borkenstein, R.F., R.F. Crowther, W.B. Shumate, W.B. Ziel, and R. Zylman. "The Role of the Drinking Driver in Traffic Accidents (THE GRAND RAPIDS STUDY)," 2nd ed. *Blutalkohol: Alcohol, Drugs and Behavior*, Vol.11, Supplement 1, 1974.
4. Cox, D.R. and H.D. Miller. *The Theory of Stochastic Processes*. New York: Wiley, 1968. 155-156, 199-200.
5. *Daubert v. Merrill Dow Pharmaceuticals* 509 US 579, 1993, 592-93.
6. De Moivre, A. *A Doctrine of Chances*, 3rd ed., 1756. Reprinted by New York: Chelsea Publishing, 1967.
7. *Federal Rules of Evidence* 702.
8. Finkelstein, M.O. *Quantitative Methods in Law*. New York: Free Press, 1978.
9. Levitt, S. D., and J. Porter. "How Dangerous are Drinking Drivers?" *Journal of Political Economy*, 109:6, 2001, 1198-1237.
10. McClave, J.T., P.G. Benson and T. Sincich. *Statistics for Business and Economics*, 7th ed. New Jersey: Prentice Hall, 1998, 322.

11. *National Automotive Sampling System General Estimates System (GES)*. [http://www.nhtsa.dot.gov/people/ncsa/nass\\_ges.html](http://www.nhtsa.dot.gov/people/ncsa/nass_ges.html)
12. Parzan, E. *Modern Probability Theory and its Applications*. New York: John Wiley & Sons, 1960, 252.
13. Redmayne, M. "Objective Probability and the Assessment of Evidence," *Law Probability and Risk*, 2, 2003, 275–294.
14. *State v. Curtis* 195 NJ Super. 354, App. Div., 984, 366-367.
15. *Traffic Safety Facts 1997: A Compilation of Motor Vehicle Crash Data from the Fatality Analysis Reporting System and the General Estimates System*. DOT HS 808 806. November, 1998.

Endnotes

1. In hypothesis testing, a candidate hypothesis, connecting what we observe with "the evidence" ( $H_1$  in hypothesis testing or *guilty* in the criminal-justice system), is tested against the hypothesis that our observations and the evidence are connected only by chance ( $H_0$  -in hypothesis testing or *innocent* in the criminal-justice analogy). The result of the test/trial is that we either find that there is a causal connection between the evidence and the hypothesis (*reject  $H_0$  or guilty*) or, alternatively, we find that the nexus between the evidence and the hypothesis does not sustain causal connection (*don't reject  $H_0$  or not guilty*). There are two types of errors that can occur: A *Type I Error* occurs when we incorrectly find a causal connection between the data and the candidate hypothesis (*we reject  $H_0$  when it is correct*, or find an *innocent person guilty*); a *Type II Error* occurs when we incorrectly find a causal connection between the evidence and the causal hypotheses, finding causality in results that were actually generated by chance (*we don't reject  $H_0$  when  $H_1$  is correct*, or we find *not guilty a guilty individual*).
2. The Poisson distribution has the following formula:  

$$P(x = x_0) = \frac{e^{-\lambda} \lambda^{x_0}}{x_0!}$$
 where  $x$  is an integer random variable, here, the number of automobile accidents (or injuries due to auto accidents);  
 $x_0$  is any specific value of  $x$ ;  
 $\lambda$  (pronounced "lambda," it is a Greek lower-case "L") and here represents the mean number of car accidents (or injuries) in a given area in a given distance traveled;

$e$  is the base of natural logarithms: 2.718...

$!$  is the factorial (repeated product) operator:

$$x_0! = 1 \cdot 2 \cdot \dots \cdot x_0; 0! = 1.$$

3. To model multiple injuries in an accident, one can use a Compound Poisson Distribution, the convolution of the Poisson with the distribution of injuries per accident.
4. "Data for the General Estimates System (GES) come from a nationally representative sample of police reported motor vehicle crashes of all types, from minor to fatal. The system began operation in 1988 and was created to identify traffic safety problem areas, provide a basis for regulatory and consumer initiatives and form the basis for cost and benefit analyses of traffic safety initiatives. The information is used to estimate how many motor vehicle crashes of different kinds take place, and what happens when they occur. ... GES data are used in traffic safety analyses by NHTSA as well as other DOT agencies. GES data are also used to answer motor vehicle safety questions from Congress, lawyers, doctors, students, researchers, and the general public."
5. VM-1 is described as a widely referenced source of information: "The Federal Highway Administration, State Highway Agencies and Metropolitan Planning Agencies use VM-1 for planning, budgeting, and legislative purposes. Academia uses VM-1 for course work or as a source for research. Private organizations such as insurance companies rely upon VM-1 for travel and registration data that affect the insurance industry. In addition, transportation-related trade associations use the data for legislative efforts. These are only some of the wide variety of uses of Table VM-1. ... Table VM-1 is a robust national transportation data source. The fact that Table VM-1 is referenced in many transportation and research documents is a testament to its impact."
6. Appendix C of *Traffic Safety Facts: 1997* contains estimates of the Standard Error of its estimated parameters. The number of injured (3.4 million) has associated with it a Standard Error of 152,100. A 95% two-sided confidence band ( $Z_{0.025} = \pm 1.96$ ) can be shown to be  $\pm 298,000$ . That is, we are 95% confident that the true number of injuries is  $2,378,000 \pm 298,000$ . In short, allowing for the sampling variation in the persons-injured statistic could potentially raise the probability of an accident from 0.0013 to 0.0014, an insignificant amount.

7. Levitt and Porter [9] estimate the externality stemming from drunk driving at  $30\text{¢/mile}$ , and calculate the fine per occurrence to compensate for that externality at \$8,000.
8. Divide 0.5 by the  $\lambda$  -value for an unimpaired driver. Given the same 20-mile trip:  

$$\lambda_{\text{impaired}} = \lambda_{\text{unimpaired}} \cdot (\text{hazard multiplier}) \Rightarrow (\text{hazard multiplier})_{\text{threshold}} = \frac{\lambda_{\text{impaired}}}{\lambda_{\text{unimpaired}}} = \frac{0.5}{0.00003165} = 15,800$$
 Comparing this figure with the hazard multiplier of 40, we can easily see that the 15,800 value is almost 400 times larger: two and a half orders of magnitude.
9. The utilization of the Reasonably-Prudent-Individual persona as a surrogate for the actor in this matter has allowed us to dodge the *individualization* bullet that can be fatal to the use of probability data. This principle was put forth in *Smith v. Rapid Transit, Inc.* [317 Mass. 469, 58 N.E.2d 754 (1945)]. In that matter, Smith was hit by a bus, and Rapid Transit, Inc., was the only bus company chartered to operate on that street. The court, in a directed verdict for the bus company, indicated that in spite of the fact that it was "mathematically" probable that a Rapid Transit, Inc., bus was involved in the accident – most of the buses on that street were Rapid Transit, Inc., buses – that alone was not enough. Here, when we explore the question of intent, not from the point of view of what the actor actually intended, but rather, what prospective risk of accident would a Reasonably-Prudent Individual in full command of the relevant accident data and mathematics determine, individualization issues do not intrude.
10. One important case, *People v. Collins* [68 Cal. 2d 319, 438 P2d 33 66 Cal Rptr. 497] concerning the use of probabilistic information in a criminal trial, stands for (among other things) the propositions that (a) merely hypothesized probabilities of event-of-interest occurrences, and (b) the unproven hypothesis that the probabilities are independent, are both simply unacceptable.

# analyzing the pharmaceutical industry through quantitative models

## Analyzing the Pharmaceutical Industry Through Quantitative Models

Andrew Russakoff, The Peter J. Tobin College of Business,  
St. John's University  
James Gould, Pace University

### Abstract

This paper undertakes two tasks. First, we shall give a summary of the last 50 years of the pharmaceutical industry trying to establish in which ways it is typical of American industry and in which ways it is unique. Second, we shall discuss the applicability of quantitative models for analyzing the industry.

### Introduction

The importance of the historical context for examination of the pharmaceutical industry is easily established. The pharmaceutical industry (or drug industry, for this paper) was simply another aspect of the American economy before 1930. Much of the industry's effort was expended in marketing various patent drugs directly to consumers. These drugs were more what we now would call elixirs and potions. They were rarely tested. They often had little or no effect. Probably the creams and lotions did more good because of the fatty base than for the specially added herbs.

We must understand this commerce as occurring in an unregulated setting. The only control was *caveat emptor*. In such a setting it is easy to see one of the odd peculiarities of the drug industry. There is some effect of supply exerting pressure on price. Some customers might prefer a less expensive product. Some customers prefer the more expensive product in the hope that the greater price reflected greater efficacy. This was an industry with little innovation in the modern sense.

This situation began to change after 1930. A series of natural products with health benefits far more scientifically based came onto the market. Foremost among these were vitamins and hormones (insulin, in particular). It was no longer a matter of advertising claims, but the results of laboratory testing. Still these were often products sold directly to consumers. The other players in the modern pharmaceutical industry had yet to become so important.

The next big change was the discovery of anti-infectives, sulfanilamide in 1935 and penicillin in 1940. They were such a big change in the world of the drug industry that the implications of the discoveries were not clear. The scientists

who established the anti-bacterial properties of penicillin did not imagine selling, licensing and marketing. Here, American commerce built on the science of the British.

A significant step forward in this direction was achieved with the American patent awarded in 1948 for streptomycin. This 17-year legal monopoly changed the nature of the drug industry. The American drug industry already had a dominant position after the second World War. The industry was poised to exploit the discovery of penicillin. It was strong due to the war effort. And its (European) competition was still recovering from the devastation of war (except for Switzerland, which still is a global competitor). In this respect, the drug industry was like the automobile industry in so far as it was dominant internationally because of the stimulus of the war effort and the devastation of the competition.

With penicillin as a model, drug companies looked for new wonder drugs for the enormous profits that came with a popular drug and patent protection. This defining rush for the industry was brought to a slower pace by the amended Food and Drug Act of 1962. This was the result of the hasty introduction of thalidomide to treat nausea in pregnant women. Although it was in use in Europe, its wider introduction into the U.S. came with the discovery that it produced deformations in children (phocomilia). The shock of this led to stricter safety standards [1].

It is apparent already that the American and international markets were closely entwined. Some of the American companies had already established branches overseas. Some European companies had branches in the U.S. The other options for U.S. companies were export (subject to lots of country-specific taxes and regulations) and licensing.

Pharmaceutical products usually fit into one of the following categories:

1. *proprietary drugs* (over-the-counter): lots of advertising, little research. Example: Preparation H.
2. *generic drugs*: lots of price competition. Example: aspirin
3. *patented drugs*: distributed by prescription. Example: Lipitor.

*“With penicillin as a model, drug companies looked for new wonder drugs for the enormous profits that came with a popular drug and patent protection.”*

It was the third group that most interested the larger drug companies. The profits from group three far exceeded those of groups one and two. There were great rewards for the discovery and production of new drugs, even if these drugs were just close imitations of previous drugs (so-called “me-too” drugs). The research and development costs were lower, and the rewards higher.

In some respects this created two distinct departments of the industry, like banks. Banks, which were content to make their profit from retail banking, could produce a steady if unremarkable profit like drug companies selling aspirin or toothpaste. Banks which emphasized the investment side were often much more profitable, embodying different attitudes, different leaders and a different level of profit—like drug companies with special patented drugs such as Viagra, or (until recently) Vioxx.

At this point it is necessary to consider that the drug industry is part of a complex, changing health care system. In addition to the drug companies, there are doctors, hospitals, pharmacies, insurance companies, health maintenance organizations, Medicare, Medicaid, etc.

We have seen that until 1930, drug advertising was pointed directly at the consumer-user. With the rise of doctor-prescribed wonder drugs, it was necessary to change the focus. Now the object of marketing efforts was the doctor. With the patients in a passive role, doctors now had a very unusual economic position. They were appealed to by the companies but they neither consumed the product, nor did they pay for it. Pharmacists had a reduced role in this transaction. One did not seek advice from the pharmacist, but rather asked him or her to fill the prescription. In this setting, prices for drugs rarely fit into normal economic models.

The supposed quality of the drug largely determined the launch price of the drug. In this sense, the price was a demand variable. Likewise for generics with the more crowded market, prices would be forced lower, again an example of a demand variable. Finally, another peculiarity of the drug market was that there was a high variation in the prices paid for the same product in different parts of the country. This was dependent upon many factors: insurance, local conditions, state laws and so forth. Again—all demand factors.

However, demand was limited to the countries with the economic power to make it profitable for big companies. The less-developed countries had severe health needs, presumably addressable by drug intervention, but with little

economic incentives, industry R&D was pointed more toward impotence, say, than malaria [4].

### Economics of the Pharmaceutical Industry

Now it is important to look more closely at the particular economics of the drug industry. The standard explanation of pharmaceutical attention is that research is so expensive and so unlikely to produce marketable results, that the industry needs lengthy patent protection to have any chance for reasonable profit. The expression one often hears is that the companies need to have a number of candidates “in the pipeline” hoping that enough will become successes. This explanation also justifies the very large profits for the successful drugs.

In practice, this is a somewhat threadbare analysis. Much of the R&D budget is devoted to marketing activities rather than research. Many of the actual research results have been produced at government laboratories, or at universities with government grants (see the role of the Bayh-Dole Act of 1980). Others have been produced by small companies, which the larger multinationals have then acquired (similar in many ways to the behavior of computer software companies, such as Microsoft).

Furthermore, it is in the interests of the drug companies to try to tweak the patented drugs of other companies (or their own!) to win separate new patent protection. In this way they try to make a product which is similar to a previous one, but can now be marketed as new, improved and different. This type of “me-too” drug often needs far less research or development, but yields as much profit (see Gladwell [2] for a good discussion of this phenomenon with Prilosec and Nexium).

At this point, generally, discussion suddenly shifts. The drug industry is held to a very high standard of scrutiny. Why are they getting such large profits for such minimally different products? Why are they putting their money (and in a sense, ours, since patent protection means higher prices for us) into “me-too” drugs instead of grappling with other different concerns? Why another cholesterol reducer rather than a cure for ALS?

In this respect, society makes rather inconsistent demands of the industry. While this is often enough the challenge in economics, it is nonetheless a difficult question. Because health concerns are common across the entire population, and health insurance is not universal, it is easy to see why the questions are raised. Gladwell identifies this problem by saying that “it is only by the most spectacular feat of

*“There were great rewards for the discovery and production of new drugs, even if these drugs were just close imitations of previous drugs (so-called “me-too” drugs).”*

Analyzing the  
Pharmaceutical Industry  
Through  
Quantitative Models

24

cynicism that our political system's moral negligence has become the fault of the pharmaceutical industry” [2:90].

This same critique ultimately may be applied to Gladwell's critique of Angell [3]. Gladwell feels that Angell wants the industry to be responsible for public health policy when that is not their business. Angell might well reply that he holds her responsible (in her former capacity as editor of the *New England Journal of Medicine*), when it is not her business, either.

These decisions are properly part of national political debate. It is not for the *New England Journal* to establish public policy, nor Pfizer. Rather it is for the people of this country to decide how much we are willing to pay in taxes for health care, and how that money should be allocated—for hospitals, for health insurance, for doctors, for pharmaceuticals.

Again, in practical terms, this is a difficult level for such a debate. Health care is not the only priority in national life. It is difficult to imagine a sensible debate about the balance between the needs of health care and the needs of national defense, let alone the needs for environmental protection, job creation and everything else.

We believe that having established the political field upon which this discussion must be held, we can now turn away to a special examination of the marketing aspects of the industry.

We have seen how drugs were marketed directly to consumers, then marketed to the prescribers, namely doctors, and now back to consumers. It does seem to be unhelpful to sell drugs to consumers when, in fact, what is needed is a doctor's prescription. However, the drug companies have discovered that this works. Patients come to doctors with their needs and interests, and prescriptions increase for widely advertised drugs. Other good examples of this, scarcely mentioned in any of the sources below, are the medicines for depression and anxiety.

To create the proper economic atmosphere for this process, there needs to be some price incentive. If there is no incentive to use a less expensive drug, there will be no less expensive drug. If there is the political will to re-shape the

industry to focus its efforts on drugs for orphan diseases (this expression is often used for diseases which affect many fewer people, and thus less investment, say for example ALS), then there must be economic incentive for the industry.

#### Implications for Examining the Pharmaceutical Industry Using Quantitative Models

For our purposes, we will consider the general quantitative model to be:

$$y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 \dots$$

This is a sufficiently flexible model to incorporate non-linear factors (e.g., we can take one variable to be the square of another). Likewise there are relatively minor technical adjustments, which can include exponential or logarithmic patterns. The cyclical aspects of many time series data sets can similarly be included with trigonometric variables (e.g., the time variable may be subjected to a sine transformation before being included in the model). The seductiveness of this model is that it is so flexible and such a powerful weapon for research.

There are many drawbacks to the application of this for pharmaceuticals. We have seen that costs are at this time not a major determining factor in understanding drug company sales and profits. That does not tell us which variable we ought to be trying to predict. There is a spurious excitement to predicting stock price. Spurious because it is so well known to depend upon company strength and rumors about drugs and legislation. Sales might be a better item if it were not so specifically tied to each separate product. Once sales are aggregated for the entire company, the sales figures are not so interesting. Sales of band-aids will mask the volatility in sales of cholesterol medicine.

We thus reach our first conclusion, namely that as powerful as the quantitative model is, it must predict a single dependent variable. It may be valuable to try to predict sales of a single prescription drug, but this is so dependent upon anomalous factors, such as a research study linking the drug (or a competing drug) with some nasty side-effect that the use of the “science” is strictly limited.

*“If there is no incentive to use a less expensive drug, there will be no less expensive drug.”*

Analyzing the  
Pharmaceutical Industry  
Through  
Quantitative Models

25

The next conclusion is that there might be a wonderful use for the model to predict the stock price. This application, however, is subject to all the usual caveats. Because the stock price is an amalgam of so many impressions, feelings, predictions, rumor and external factors like research studies, competing products, etc., the price resembles the proverbial random walk as much as it does any other predictable pattern. Generally the effort to predict stock prices is attendant upon the use in quick profits buying and selling on the exchange.

For this use, much of the power of the model (including the polynomial, exponential and other factors) is lost. For the purposes of quick turnaround, all these might as well be replaced by linear factors, since the short-term effect is all that is needed in the prediction. Such a use would suit the legions of day-traders. Not only are the powerful functions negated in this use, but the effects of one-time events—rumors, laws, side-effects—completely overwhelm the longer term trends.

This leaves us with a different perspective on the modeling process. We know that the greater our understanding of a single pharmaceutical firm, the less help a quantitative model is in predicting stock price. We have seen that apart from stock price, it is not clear that a predictive model is all that helpful.

This allows for a re-examination of the use of the model for short-term stock price prediction. In this use, the stock price must be regarded as an entirely separate *sui generis* number that follows its own “logic.” We do not then worry much about the competition between generic and prescription drugs. We do not try to project out the value of drugs in the pipeline and the likelihood that they will remain in one form or another under patent protection. Instead we look at the pattern of stock prices and claim that this is a phenomenon that has its own patterns and predictabilities. The model predicts after such and such a rise there will be such and such a dip, therefore we buy and sell accordingly. This “technicians” point of view removes us from discussion of the mathematics of model construction as well as the historical patterns of the pharmaceutical industry and takes us to a kind of theology of stock price behavior.

In conclusion, it seems that the more we know about the pharmaceutical industry the less confidence we may have in the usefulness of quantitative predictive models. Conversely, the more confidence we have in the predictability of the stock price of pharmaceuticals, the more the quantitative model seems to promise as an adjunct to intuition.

#### References

1. Edwards, T. *The Competitive Status of the U.S. Pharmaceutical Industry: The Influences of Technology in Determining International Industrial Competitive Advantage*. Washington D.C.: National Research Council, 1983.
2. Gladwell, M. “High Prices,” *The New Yorker*, October 25, 2004. An essay inspired by Marcia Angell, *The Truth About the Drug Companies*, Random House, 2004, and John Abramson, *Overdosed America*, HarperCollins, 2004.
3. Hall, S.S., “The Drug Lords,” *New York Times*, November 14, 2004, reviewing Marcia Angell, *The Truth About the Drug Companies*, Random House, 2004, and Jerry Avorn, *Powerful Medicines*, Alfred A. Knopf, 2004.
4. Schweitzer, S.O., *Pharmaceutical Economics and Policy*. Oxford University Press, 1997.

#### Bibliographic Note:

The background for this paper has been derived from the above four sources. Edwards is a thoughtful quick summary. Schweitzer is a magnificent and thorough examination of the industry. Gladwell is an always interesting writer for the *New Yorker Magazine*. Hall did the review for the *New York Times Book Review*.

*“In conclusion, it seems that the more we know about the pharmaceutical industry the less confidence we may have in the usefulness of quantitative predictive models.”*

# the effect of the internet on stock market volume and volatility

## The Effect of the Internet on Stock Market Volume and Volatility

Farok Vakil, The Peter J. Tobin College of Business, St. John's University  
 FangLieh Victor Lu, The Peter J. Tobin College of Business, St. John's University

### Abstract:

Technological advances have revolutionized data communication, voice communication and the delivery of information. With the help of the Internet, useful information can be disseminated around the world in a matter of a few seconds. This new method of delivering information has had an enormous effect on business activities in general and on the financial sector and stock investors in particular. These days, the Internet is making stock information of high quality available almost instantly to individual investors everywhere. Armed with such information, these investors can make more informed financial decisions regarding buying and selling stocks. Furthermore, there are lower commissions charged for executing orders through the Internet compared to the traditional broker's commissions. Due to these two factors, a new group of investors has emerged. And, as a direct result of this phenomenon, the volume of trading and the volatility of stock prices are expected to increase substantially. Although there are other issues, such as general economic conditions, market conditions and interest rates, which can contribute to change in the volume and volatility of the market, in this study we investigate the effect of the Internet on the stock market in terms of volume and volatility. Various data analysis methods will be utilized in order to examine the effect of this new phenomenon on the Dow Jones Industrial Average (DJIA), its volatility and its volume.

### Introduction

As technology advances and the computer becomes increasingly popular, useful information can be obtained very rapidly. Availability of information can have a profound effect on the decision-making process. Consequently, it is important to examine the effect of new technology and of the instant availability of information on the decision-making process employed by investors in general, and by stock investors in particular.

The Internet has become one of the most powerful vehicles for disseminating information [6]. Along with its related components, such as e-mail and messenger services, the Internet has given people throughout the world immediate

access to new financial information. As a direct result of this phenomenon, stock investors have immediate access to an enormous amount of data. These new means of disseminating information have created a tremendous opportunity in the economy in general and in the financial sector in particular. Consequently, over the last 10 years the Internet has become a key channel for distributing and servicing financial products. More and more companies are marketing their investment products over the "Net," because distribution through this channel will give a company a competitive advantage in the increasingly challenging financial industry.

The instant availability of financial news has a profound effect on the decision-making process of group and individual investors. In the past, people obtained financial news and advice solely through their traditional brokers, who charged considerable fees and commissions. In many instances, the high cost of traditional brokerages discouraged many low- to moderate-income earners (potential investors) from trading in the equity markets. By contrast, the Internet is leveling the playing field for everybody. Instant availability of information, in addition to lower fees and commissions offered by online accounts, has created a new opportunity for every potential investor, regardless of his or her financial means.

### Background Information

When the online brokerage account was introduced in the 1990s, it was expected that this new form of investment would revolutionize the way people handled their stock transactions. Optimistic industry analysts predicted that online brokerage accounts would soon replace traditional brokerage accounts. They envisioned that lower cost, convenience and 24/7 accessibility would make the online brokerage account extremely attractive. In addition, online stock investors have total control over the selection of their portfolio, their level of risk, initiation of sell and buy transactions and the research function. In short, online investors are responsible for all financial aspects of their investments.

Industry analysts predicted that during the late 1990s and early 2000s, the Web would grow and online financial tools

**EXHIBIT 1. FEES AND FEATURES OF FIVE LEADING BROKERAGES IN THE U.S.**

Firm Name	Cost to Customers	Maximum number of shares	Minimum Initial Balance
Schwab	\$30	1,000	\$5,000
E*Trade	\$15 listed stocks, \$20 unlisted stocks	5,000	\$1,000
TD Waterhouse	\$18	5,000	\$1,000
Fidelity	\$19.95	1,000	\$2,500
Datek	\$10.95 NASDAQ and NYSE Trades	Unlimited	\$2,000

would improve. They foresaw that these factors would cause the number of online accounts to increase and the value of these accounts to rise very rapidly [2]. Indeed, as the financial markets rose and continued rising through the year 2000, so did the number of online brokerage accounts. Some firms reported 100+% annual growth rates.

However, the principal downside of online brokerage accounts is the absence of the personal financial advisor. And so, when the financial markets started going haywire in 2001, online investors by and large found themselves very vulnerable [5]. A personal financial advisor could have assisted these investors in assessing their tolerance level for market risk, and then helped further by making reasoned recommendations. Instead, online investors had to deal with all aspects of stock investments by themselves. Consequently, the high rate of growth of online brokerage accounts—as had been optimistically predicted by industry analysts—did not materialize in 2001.

The unexpected change in the stock market during the year 2001, which resulted in substantial trading losses, left many investors questioning the prudence behind do-it-yourself investing. As a result, while many online investors kept their online trading accounts, they also turned to some firm or person for financial advice. Having both an advisor and the ability to trade online is an appealing arrangement for some people, and so, ever since the year 2001, there has been both a higher demand for online brokerage accounts, as well as a significant increase in the number of online accounts [4]. A number of brokers also realized this, and began adopting strategies that use both online trading and traditional brokerage services, to attract new accounts.

Generally, there are three requirements for opening an online account. First, a computer with Internet access is a necessity. The second requirement is a new Web browser, such as Netscape 8.0, Internet Explorer 6.0, Opera 8.0 or Firefox 1.0. When you log on to your account, encryption technology encodes all the data transmitted through your browser. Your browser and your broker's trading system will agree on a "key" or combination, which allows the broker to decode and read the messages you send, while preventing others from seeing them. This enables the communication between your browser and your online broker to remain secure. Finally, the last requirement is money (fees and account balances). Exhibit 1 summarizes the online account fees and features for five leading online brokerages in the U.S. as of March 2005.

*"...over the last 10 years the Internet has become a key channel for distributing and servicing financial products."*

### Data Sources and Data Collection

Given the nature of this study, different data sources are cited. Electronic data available on the Internet, such as historical data related to the DJIA, was downloaded from the financial sites of www.msn.com, and www.yahoo.com. This data was divided into two groups: 1) data previous to

*“The instant availability of financial news has a profound effect on the decision-making process of group and individual investors.”*

**EXHIBIT 2. SUMMARY STATISTICS FOR DJIA**

	Monthly DJIA Closing*		Average Monthly Volume of DJIA*	
	1990-1995	1996-2004	1990-1995	1996-2004
Mean	3,464.61	9,072.83	2,506,222.93	20,714,625,648.15
Median	3,382.30	9,332.71	2,528,390.50	20,875,150,000.00
Standard Deviation	635.49	1,637.85	719,857.52	8,237,472,130.18
Range	2,674.79	6,101.82	2,362,760	34,848,390,000.00
Number of Observations	72	108	72	108
Statistical Test (T-test)	T-Observed = 32.14**		T-Observed = 19.33**	
Statistical Test (F-test)	F-Observed = 6.64**		F= Observed =32.10**	

\* Monthly DJIA Closing and Monthly Volume were obtained from the financial site of MSN.com.

\*\* significant at  $\alpha=.01$

and including December 1995, which preceded the popularity of online accounts, and 2) data after December 1995, when online accounts became ever more popular. Other related data, such as the number of online accounts, was obtained by using print sources.

**Methodology**

Historical information about the Dow Jones Industrial Average was downloaded in spreadsheet format. Excel 2003 was then used for preliminary data analysis. For advanced statistical tests, the add-in command of Excel 2003 was utilized. As mentioned above, the data first was divided into two parts: January 1990 to December 1995, and January 1996 to December 2004. By analyzing these two sets of data, we investigated the change in volatility and volume of the DJIA Index.

Although there are several statistical methods, which can be used to measure the volatility of stocks, the *standard deviation* is one of the most widely used statistics for measuring stock volatility [1]. In carrying out this study, we performed a statistical test to compare the standard deviations for two different time periods: 1990-1995 and 1996-2004. It can be argued that if the *volatility* of the

market has not changed, then our statistical test must confirm that fact by concluding that the standard deviations for the two time periods are equal. Otherwise, if there is a significant increase in the standard deviation in the 1996-2004 period compared to the 1990-1995 period, then we can conclude that the volatility has increased. We also test for increase in the means of the two periods. The summary statistics for the change in the Monthly DJIA Closing and Average Monthly Volume of the DJIA are given in Exhibit 2. We performed the T-test for equality of two means and the F-test to investigate the equality of two standard deviations (1990-1995 period and 1996-2004 period). Our observed T-test and F-statistics support the hypotheses that there is a significant increase in the mean and standard deviation during the 1996-2004 period compared to the 1990-1995 period. Consequently, we can conclude that the average of our index and its volatility has significantly increased.

Another measure of variation and volatility is *range*, which is very easy to compute and interpret [1]. Based on our data, we calculated the yearly range for the DJIA for 1995 to 2003. Furthermore, we collected the data related to the number of online brokerage accounts during the same period [3]. The result is summarized in Exhibit 3.

**EXHIBIT 3. YEARLY RANGE OF DJIA AND NUMBER OF ONLINE ACCOUNTS**

Year	Range of DJIA	Number of online accounts (in millions)
1995	1472.68	0.30
1996	1623.91	0.90
1997	2024.30	3.10
1998	2078.25	7.30
1999	2664.62	12.9
2000	2337.10	16.6
2001	3509.49	22.1
2002	3551.21	29.2
2003	3097.09	36.4

**EXHIBIT 4. REGRESSION RESULTS**

	Coefficients	Standard Error	t Stat	P-value
<b>Intercept</b>	1724.4848130	196.0742881	8.795058392	4.954E-05
<b>Number of Online Accounts (in millions)</b>	53.09081276	10.4445654	5.083104059	0.0014261

An examination of the data indicates that there might be some correlation between the number of online accounts and the volatility of the DJIA. To investigate this notion, we used Excel 2003 to calculate the correlation coefficient (R). The result indicates a high correlation of .887. A high positive correlation supports the notion that the two variables are changing in the same direction. To further analyze the relationship between the volatility of the DJIA and number of online accounts, we ran a regression model. The result is provided in Exhibit 4.

Based on Exhibit 4, we can conclude that the explanatory variable (number of online accounts) contributed significantly to predicting the volatility of the DJIA index (since the T-Observed is larger than the T-Table at  $\alpha=.01$ ). This conclusion supports the notion that the number of online accounts has statistically contributed significantly to the increase in volatility of the DJIA.

$$Y = 1724.48413 + 53.09081276 * (\text{number of online accounts})$$

P-Value for Slope = .001, significant at  $\alpha=.01$

Based on these statistical tests, we can conclude that the volatility of the market has increased during the 1996-2004 period compared to the 1990-1995 period. Furthermore, our statistical tests support the hypothesis that the number of online accounts contributed significantly to the volatility of the stock market during the last 10 years.

**Future Research**

Future research in this area might focus on the relationship between the number of online accounts and other indexes such as the NASDAQ and S&P 500. Indeed, one possible extension of the present study might be an investigation of the degree of sensitivity of each market to the change in the

“...the number of online accounts contributed significantly to the volatility of the stock market during the last 10 years.”

number of online accounts. A multivariate model approach might be necessary to determine which index is more sensitive to a change in this number. Another possible extension of this study would be to focus on larger data sets and include newer data (2004 and 2005). By including the newer data one can determine whether the correlation between the index under study (DJIA) and the number of online accounts remains stable and strong or at some point becomes weak.

#### Conclusion

The effect of the Internet on business activities has been enormous during the last several years. The Internet is making high quality stock information available to individual investors almost instantly. Online trading of stocks has become very popular, both to those trading online and to those using traditional brokers; many investors are also moving from doing their own research to trading online. Consequently, the number of online brokerage accounts has increased significantly during the last 10 years. Based on our results, this increase has resulted in a larger volatility of the DJIA. Although there are other factors at play, such as general economic conditions, market conditions and interest rates, each of which can clearly contribute to change in the volume and volatility of the market, in this study we have limited ourselves to concentrating on the effect of the Internet on the stock market in terms of the volume and volatility of the market.

#### References

1. McClave, J., G. Benson and T. Sincich. *Statistics for Business and Economics*, 9th Ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2005.
2. “Online Brokerage Growth Surge.” *Web Finance*, Vol. 4, Issue 5, Feb. 14, 2000, 12.
3. “Online Brokerage.” *European and U.S. Market Trend Report*, MR-0557, March 1, 2001, Global Industry Analysts, Inc., www.globind.com.
4. “Online Trading Market-U.S. Report.” Mintel International Group LTD, www.marketresearch.com, Nov. 1, 2004.
5. Perkins, A. and M. Perkins. *The Internet Bubble*, Revised Ed., New York: HarperBusiness, 2001.
6. Shelly, G., T. Cashman and M. Vermaat. *Discovering Computers*, Boston: Thompson Course Technology, 2005.

# transforming a university from a teaching organization to a learning organization

## Transforming a University from a Teaching Organization to a Learning Organization

Hershey H. Friedman, Brooklyn College of the City University of New York  
Linda W. Friedman, Baruch College of the City University of New York  
Simcha Pollack, The Peter J. Tobin College of Business, St. John's University

#### Abstract

Successful 21-century universities will have to be lean, flexible and nimble. In fact, Peter Drucker claims that 30 years from now the “big universities will be relics” and will not survive. In the corporate world, businesses are becoming learning organizations in order to survive and prosper. This paper explains why it is necessary for universities to become learning organizations and provides ideas as how to make the transformation.

#### Introduction

Peter Drucker noted in an interview that: “Thirty years from now the big university campuses will be relics. Universities won't survive. It's as large a change as when we first got the printed book” [18]. This may be an exaggeration, but there is no question that universities that refuse to change may not survive. The rise of for-profit universities (e.g., the University of Phoenix), decreased government support for universities, the rising costs of education, the globalization of education, technological change, the growing number of working adults who need continuing education to avoid obsolescence and distance education are forcing universities to transform themselves. In fact, Andrews et al. [1] urge academia to respond to the “wake-up call” and recognize that inflexibility and the failure to respond quickly and decisively to environmental change can be dangerous.

For colleges to change, they not only have to learn to run their organizations in a more business-like fashion, they have to be willing, when necessary, to add and shrink programs quickly. This is not easy when the organizational structure of today's university has more to do with the convenience of establishing accounting budgets than with the demands of intellectual growth and education [12,14]. Edwards [9] notes that “the actual elimination of departments is extremely rare and usually generates a wave of unflattering national news, so the substitution strategy is driven toward less visible, more surreptitious methods.”

It is becoming quite apparent that being inflexible and resistant to change in an extremely fast-moving environment is a prescription for disaster, whether we are dealing with a business or academic institution. Several

visionaries believe that the university of the future will be very different from the university of today: more interdisciplinary programs, and the substantial modification of the current prevalent academic organizational structure.

Duderstadt [7] suggests that the university of the future will be divisionless, i.e., there will be many more interdisciplinary programs. There will also be “a far more intimate relationship between basic academic disciplines and the professions.” He asks “whether the concept of the disciplinary specialist is relevant to a future in which the most interesting and significant problems will require ‘big think’ rather than ‘small think’” [8]. Kolodny [16:40-41] asserts that the antiquated way of organizing colleges—by departments—will have to “evolve into collaborative and flexible units.” Students with narrowly defined majors will have great difficulty comprehending a world in which the knowledge required of them is complex, interconnected and, by its very nature, draws from many areas. Edwards [9] maintains that “in so many cases, the most provocative and interesting work is done at the intersections where disciplines meet, or by collaborators blending several seemingly disparate disciplines to attack real problems afresh.”

#### The Learning Organization

Clearly, there are great changes ahead for higher education, but changing the culture of an organization is a daunting task. Forward-thinking institutions have to consider what can be done to make their organizations more responsive to change. In the corporate world, many firms are recognizing that the ability of an organization to learn is the key to survival and growth, and “organizational learning” has become the mantra of many companies [3,21].

What is organizational learning? Organizational learning has been defined in many ways: Stata [24] asserts that: “organizational learning occurs through shared insights, knowledge and mental models ... [and] builds on past knowledge and experience.” Senge [21] writes: “learning organizations are not only adaptive, which is to cope, but generative, which is to create.” Pedler et al. [20] state: “A learning company is an organization that facilitates the learning of all its members and continually transforms itself.” Garvin [11] believes that a learning organization is

“an organization skilled at creating, acquiring, and transferring knowledge, and at modifying its behavior to reflect new knowledge and insights.”

What should we find in a learning organization? The following briefly summarizes what one would expect:

- Awareness of the external environment. Knowing what the competition is doing.
- Belief that individuals can change their environment. A learning culture.
- Shared vision. One that encourages individuals to take risks.
- Learning from past experience and mistakes—experience is the best teacher.
- Learning from the experiences of others in the organization. Organizational memory in order to know what worked in the past and what did not.
- Willingness to experiment and take chances. Tolerance for failure.
- Double-loop or generative learning. With double-loop, as opposed to single-loop, learning, assumptions are questioned. “Double loop learning asks questions not only about objective facts but also about the reasons and motives behind those facts” [2].
- Concern for people. Respect for employees. Diversity is seen as a plus since it allows for new ideas. Empowerment of employees.
- Infrastructure allowing the free flow of knowledge, ideas and information. Open lines of communication. Sharing of knowledge, not just information. Team learning where colleagues respect and trust each other. An organization where one employee will compensate for another’s weaknesses, as in a successful sports team.
- Utilization of shared knowledge. Emphasis on cooperation, not turf.
- Commitment to lifelong learning. Constant learning and growth.
- Ability to adapt to changing conditions. Ability to renew, regenerate and revitalize an organization.

Knowledge sharing is a necessary condition for having a learning organization. To foster the sharing of knowledge, computer software has been developed to make it easy for coworkers to share their expertise. For instance, the AskMe

Corporation (<http://www.askmecorp.com/>) claims that it is “the leading provider of software solutions that enable global 2000 companies to create and manage Employee Knowledge Networks (EKNs).” AskMe notes on its website that creating EKNs helps ensure that employees do not have to solve problems that others have already solved, i.e., “reinventing the wheel.” It also enables employees in a firm to quickly find the individual with the appropriate expertise to solve a problem.

One thing the AskMe company discovered is that knowledge sharing is difficult in pyramid-shaped organizations with tall organizational structures, i.e., characterized by numerous layers of management. Knowledge sharing works much better where there is a flat organizational structure with a relatively short chain of command. However, managers have to be willing to accept suggestions, ideas and answers from their employees. When information flows in all directions—even from the bottom of the organizational pyramid to the top—some managers might feel that they are losing some of the status and authority of their position. After all, it is quite conceivable that someone in the mailroom might be able to answer a question that stumps top management. Knowledge can be found anywhere and everywhere.

The power of knowledge sharing should not be underestimated. Linux, the extremely successful computer operating system, was developed by the collaboration of programmers all over the globe.

#### Are Universities Learning Organizations?

Before discussing universities, it might be instructive to examine whether schools—especially primary and secondary ones—are learning organizations. The evidence, albeit limited, indicates that they are not. Shields and Newton [22] examined schools that participated in the Saskatchewan School Improvement Program and found that they were not learning organizations. Isaacson and Bamburg [15] also came to the same conclusion. Schools rarely have visions, teachers rarely share knowledge with colleagues, and schools are managed with a top-down approach. Many others agree that schools have not functioned as learning organizations [5,10]. When Senge was asked by O’Neill [19] whether or not schools were learning organizations, he replied: “definitely not.”

Universities are not run like high schools or elementary schools and stress research/learning as much as (or more than) teaching. Despite this, it seems that very few universities would qualify as learning organizations. It is quite ironic that teaching organizations do not know how to learn. Most universities have little knowledge sharing and

*It is quite ironic that teaching organizations do not know how to learn. Most universities have little knowledge sharing and are notorious for turf battles.*

are notorious for turf battles. Smith [23] asserts that: “Academic departments serve as organizations that exhibit all the segmentary politics described by anthropologists: segmentation for largely demographic reasons, balanced opposition among themselves, and unitary resistance to a superordinate entity, usually the college or university as a whole.” Harrington [13] believes that departments encourage loyalty to the discipline rather than to the university. Apparently, most universities are not learning organizations.

#### Transforming the University into a Learning Organization

The following are some suggestions that can be used to help transform the university into a learning organization.

1. Establish a message board to function as a research matchmaking service. As noted above, the most exciting research is often at the interface of two disciplines. Furthermore, researchers with expertise in one area (e.g., biology) might need to collaborate with a faculty member with expertise in another area (e.g., computer simulation or geology) in order to write a paper. Universities could provide a central message board where faculty members could state the area(s) in which they are doing research and the kind of co-author, if any, they seek. This site could also be used to find ideas for research. Senior faculty members might be willing to provide ideas for research in return for a byline on any resulting article. If successful, this service can be extended to include faculty in other colleges. Universities have to understand that discouraging professors from writing co-authored papers is counterproductive. It is certainly not consistent with a key philosophy of a learning organization: sharing knowledge. Moreover, working with scholars from other disciplines creates a synergy that can result in truly innovative research. It is not uncommon in academe to find professors who continue to write essentially the same paper over and over with very little new information. There is nothing wrong with collaboration if it produces exciting research. One wonders whether James Watson and Francis Crick would have been as successful if they had worked alone. The Human Genome Project took 13 years and involved researchers from at least 18 countries.
2. Establish an online archive where faculty can post papers for review by colleagues before submitting them to journals. If the faculty at a university work together as a team and want their institution to flourish, they are more likely to provide helpful criticism. The late OpenTextProject ([www.opentextproject.org](http://www.opentextproject.org)) was an

international site that allowed individuals to post their papers for pre-submission review.

3. There could be a Web site for every course, especially multiple-section courses taught by a number of different faculty. Faculty could submit their best ideas on how to teach the course and their best lectures. This site would then be a resource for students who have difficulties with the course and would also be a resource for faculty teaching the course. Most professors teaching a course have gotten useful ideas from other faculty teaching the same course. For instance, suppose we have a site for elementary statistics. This might be a course taught by 10 different instructors. Faculty teaching the course would be encouraged to post material dealing with statistics. This might take the form of syllabi, lectures, interesting examples, humorous ways to illustrate difficult concepts, computer programs to solve statistics problems, solved exercises, etc. One of the authors has a site for his corporate finance class and has heard that students taking the course with other instructors go to the site since it contains dozens of problems with solutions in the area of mathematics of finance. Professor N. Duru Ahanotu created a Web site (<http://www.drdu.com/knowledge.html>) for anyone interested in learning organizations and knowledge management.

The corporate world is learning the value of the Web for e-training. The type of Web site described above can be especially useful to faculty teaching a course for the first time. Rather than learning the best way to teach a course through trial and error, they can go to the Web site for a particular course and see how colleagues have been teaching it. Many professors do indeed go to the Web to examine syllabi and course material from the same courses taught at schools all over the country. The problem is that the caliber of student may not be exactly the same. While it is still a good idea to see how a particular course is taught at other colleges, it will often be more useful to examine the materials used by colleagues in the same school. Interestingly, Dill [6] notes that a major weakness of universities has been in the “internal transfer of new knowledge.” This is why it is not uncommon to find that “within most universities there are better performing units that have knowledge about improving teaching and learning from which other units could learn.”

*Knowledge-sharing software could be used by administrators to get fresh ideas from the entire faculty.*

Transforming a University from a Teaching Organization to a Learning Organization

34

4. Knowledge sharing should not be limited to a university. Knowledge should be shared with the public. A Web site could be created providing helpful information for the general public. For instance, this site could have links to subjects such as small business management, marketing, personal finance, ESL, etc. Outsiders could learn these subjects online for free. Brew and Doud [4] assert that work-based learning is important for students. This means that there has to be a partnership between educators and workplace supervisors, especially with professional education. This, of course, requires knowledge sharing between academics and practitioners.
5. University administrators have to realize that the pyramid-shaped organizational structure makes little sense for an academic institution. Information should not only flow from the top to the bottom, i.e., president to provost to dean to chairs to faculty. The biggest impediments to the creation of learning organizations are the twin fears of change and of things that are new. Senior faculty often resist change. Indeed, Kuhn [17] found a similar phenomenon in the sciences. Kuhn described "normal science," as where scientists who adhere to the old dominant paradigm resist the adoption of a new paradigm. Kuhn [17:52] notes that "normal science does not aim at novelties of fact or theory and, when successful, finds none." Some of the best ideas might originate from junior faculty who often have a new perspective. Universities that want to be innovative have to allow information to flow from the bottom to the top, otherwise they will stagnate. Knowledge-sharing software could be used by administrators to get fresh ideas from the entire faculty.
6. Students have to be part of the knowledge sharing for a university to become a true learning organization. Many faculty members resist providing students with e-mail addresses and brick-and-mortar office hours of three hours per week are ludicrous in the age of asynchronous communication. How many faculty members today would deal with a bank that was only open from 9 a.m. to 3 p.m., had no ATM machines and no online banking? Information about majors can be automated. There could be a Web site where students can find out about any major, including requirements for the major and opportunities in the field. Sites consisting of FAQs (frequently asked questions) could be provided for students. Expert systems could be used to advise students as to whether they have the necessary prerequisites for a course. When you purchase a book at Amazon.com, the next time you come back you are greeted by name and

other books are recommend to you based on your purchase history. Students could also automatically receive recommendations for courses based on their major and their registration history.

7. As noted above, many futurists believe that interdisciplinary majors will be vital to the future of universities. Many of the newer programs being developed at colleges all over the country are interdisciplinary. It is often very difficult to get academic departments to create interdisciplinary majors when each department is interested in protecting its own turf. Learning organizations stress cooperation, not protection of turf, and this might require a new organizational structure not based on departments. Alternatively, department chairs could report to a "super" chair or dean with the responsibility for an entire school. The job of the "super" chair or dean would be to ensure that departments work together to create interdisciplinary programs and focus on what is best for the university as a whole, not just their own department. A discussion group in which faculty members could provide ideas for new programs could be established. Administrators could reward faculty and departments that create successful programs.
8. A learning organization cannot last long if members of the organization have no interest in learning. Unfortunately, a significant number of faculty (one number often quoted is 60%) never publish an article after they receive tenure and become associate professors. Incentives must be put in place to ensure that faculty continue to learn even after being promoted to full professor. Lifelong learning is now necessary in many professions, including medicine and law. It should also be encouraged in academe.

#### Conclusion

Establishing a paradigm of knowledge sharing and continuous growth through lifelong learning is not easy even, or perhaps especially, in academe. Interestingly, in these very turbulent times, many academicians are complacent and feel that there is no compelling need to make any serious changes. This is definitely a myopic way of thinking. Transforming colleges into learning organizations will not solve all problems, but it is certainly an important first step.

*Lifelong learning is now necessary in many professions including medicine and law. It should also be encouraged in academe.*

Transforming a University from a Teaching Organization to a Learning Organization

35

#### References

1. Andrews, R. L., M. Flanigan, and D. S. Woundy. "Are Business Schools Sleeping Through a Wake-Up Call?" *Decision Sciences Institute 2000 Proceedings*, 1, 2000, 194-196.
2. Argyris, C. "Good Communication that Blocks Learning." *Harvard Business Review*, Vol. 72, No.4, 1994, 77-85.
3. Argyris, C. and D. Schoen. *Organizational Learning II: Theory, Method, and Practice*. Reading, MA: Addison-Wesley, 1996.
4. Brew, A. and D. Boud "Preparing for New Academic Roles: A Holistic Approach to Development." *The International Journal for Academic Development*, Vol. 1 No. 2, 17-25.
5. Conzemius, A. and W. Conzemius. "Transforming Schools into Learning Organizations." *Adult Learning*, Vol. 7 No. 4, 1996, 23-25.
6. Dill, D. D. "Academic Accountability and University Adaptation: The Architecture of the Academic Learning Organization." *Higher Education*, 38, 1999, 127-154.
7. Duderstadt, J. J. "A Choice of Transformations for the 21st-Century University." *The Chronicle of Higher Education*, 46, Feb. 4, 2000, B6-B7.
8. Duderstadt, J. J. "The Future of the University in an Age of Knowledge." *The Journal of Asynchronous Learning Networks*, 1, 1997, 78-88.
9. Edwards, R. "The Academic Department: How Does it Fit into the University Reform Agenda?" *Change* 31, 1999, 17-27
10. Fullan, M. The School as Learning Organization: Distant Dreams. *Theory into Practice*, Vol. 34, 1995, No. 4, 230-235.
11. Garvin, D. A. "Building a Learning Organization." *Harvard Business Review*, Vol. 71, No. 4, 1993, 78-91.
12. Gazzaniga, M. "How to Change the University." *Science*, 1998, 237.
13. Harrington, F. H. "Shortcomings of Conventional Departments." In D. E. McHenry (Ed.), *Academic Departments: Problems, Variations, and Alternatives*. San Francisco: Jossey-Bass, 1977, 53-62.

14. Hollander, S. "Second Class Subjects? Interdisciplinary Studies at Princeton." *The Daily Princetonian*, April 24, 2000, 3.
15. Isaacson, N. and J. Bamberg "Can Schools Become Learning Organizations?" *Educational Leadership*, Vol. 50, No. 3, 1992, 42-44.
16. Kolodny, A. *Failing the Future: A Dean Looks at Higher Education in the Twenty-First Century*. Durham, NC: Duke University Press, 1998.
17. Kuhn, T. *The Structure of Scientific Revolutions*, 2nd ed. Chicago: University of Chicago Press, 1970.
18. Lenzner, R. and S. S. Johnson. "Seeing Things as They Really Are." *Forbes*, March 10, 1997, 122-131.
19. O'Neil, J. "On Schools as Learning Organizations." *Educational Leadership*, Vol. 52, No. 7, April 1995, 20-23.
20. Pedler, M., J. Burgoyne and T. Boydell. *The Learning Company: A Strategy for Sustainable Development*. New York: McGraw-Hill, 1991.
21. Senge, P.M. *The Fifth Discipline*. New York: Doubleday, 1990.
22. Shields, C. and E. E. Newton. "Empowered Leadership: Realizing the Good News." *Journal of School Leadership*, Vol. 4, No. 2, 1994, 171-196.
23. Smith, J. Z. "To Double Business Bound." In C. G. Schneider and W. S. Green (Eds.), *Strengthening the College Major*. San Francisco, CA: Jossey-Bass Inc. Publishers, 1993, 13-23.
24. Stata, R. "Organizational Learning—The Key to Management Innovation." *Sloan Management Review*, Spring 1989, 63-74.

# regression analysis revisited

## Regression Analysis Revisited

Athanasios Vasilopoulos, The Peter J. Tobin College of Business, St. John's University

### Abstract

Regression analysis is at the center of almost every forecasting technique, yet few people are comfortable with the regression methodology. We hope to improve the level of comfort with this article. Briefly we will discuss the theory behind the methodology and then outline a step-by-step procedure, which will allow almost everyone to construct a regression forecasting function for both the linear and multivariate cases. Also discussed, in addition to the model estimation mentioned above, is model testing (to establish significance) and the procedure by which the final regression equation is derived and retained to be used as the forecasting equation. Hand solutions are derived for two small-sample problems (one each for the linear and multivariate cases) and their solutions are compared to the MINITAB-derived solutions to establish confidence in the statistical tool, which is to be used exclusively for larger problems.

### Introduction and Model Estimation

Regression analysis, in which an equation is derived that connects the value of one dependent variable (Y) to the values of one independent variable X (linear model) or p independent variables  $X_1, X_2, \dots, X_p$  (multivariate case), starts with a given bivariate data set (or multivariate data set) and uses the Least Squares Method to assign the best possible values to the unknown multipliers found in the models we wish to estimate. The bivariate data, used to estimate the linear model, consists of n ordered pairs of values:  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ , while the multivariate data set used to estimate the multivariate model consists of n p-tuples of values:

$$(x_{11}, x_{21}, \dots, x_{p1}, y_1), (x_{12}, x_{22}, \dots, x_{p2}, y_2), \dots, (x_{1n}, x_{2n}, \dots, x_{pn}, y_n)$$

The linear model we wish to estimate, using the given data, is:

$$Y = a + bX, \tag{1}$$

while the multivariate model is given by:

$$Y = a + b_2X_2 + b_3X_3 + \dots + b_pX_p \tag{2}$$

$$\text{or } Y = b_1X_1 (=1) + b_2X_2 + b_3X_3 + \dots + b_pX_p \tag{3}$$

Note that the first two terms of the multivariate model are identical to the linear model, and in equation (3) we introduced a variable  $X_1$ , whose value is always equal to 1, to make the handling of the multivariate model easier, using matrix operations. Note also that the  $a$  in equation (2) is set equal to  $b_1$  in equation (3). To estimate both models (1) and (3) we use the Least Squares Methodology, which calls for the formation of the two Quadratic functions:

For the Linear Model:

$$\begin{aligned} Q(a,b) &= \sum_{i=1}^n [y_{\text{actual}} - y_{\text{linear equation}}]^2 = \sum_{i=1}^n [y_i - a - bX_i]^2 \\ &= [y_1 - a - bX_1]^2 + [y_2 - a - bX_2]^2 + \dots + [y_n - a - bX_n]^2 \\ &= [y_1^2 + y_2^2 + \dots + y_n^2] - 2a[y_1 + y_2 + \dots + y_n] - 2b[X_1y_1 + X_2y_2 + \dots + X_ny_n] \\ &\quad + 2ab[X_1 + X_2 + \dots + X_n] + b^2[X_1^2 + X_2^2 + \dots + X_n^2] \end{aligned}$$

$$Q(a,b) = \sum_{i=1}^n y_i^2 - 2a \sum_{i=1}^n y_i - 2b \sum_{i=1}^n X_i y_i + 2ab \sum_{i=1}^n X_i + b^2 \sum_{i=1}^n X_i^2 \tag{4}$$

For the Multivariate Model:

$$\begin{aligned} Q(b_1, b_2, \dots, b_p) &= \sum_{i=1}^n [y_{\text{actual } i} - y_{\text{Multivariable function } i}]^2 \\ &= \sum_{i=1}^n [y_i - b_1 - b_2 X_{2i} - b_3 X_{3i} - \dots - b_p X_{pi}]^2 \end{aligned} \tag{5}$$

To derive the "Normal" equations for the linear model from which the values of  $a$  and  $b$  of the linear model are obtained, we take the partial derivative of  $Q(a,b)$  of equation (4) with respect to  $a$  and  $b$ , set each equal to zero, and then simplify:

The result is:

$$\frac{\partial Q(a,b)}{\partial a} = -2 \sum_{i=1}^n y_i + 2b \sum_{i=1}^n X_i + 2a n \tag{6}$$

$$\text{and } \frac{\partial Q(a,b)}{\partial b} = -2 \sum_{i=1}^n X_i y_i + 2a \sum_{i=1}^n X_i + 2b \sum_{i=1}^n X_i^2 \tag{7}$$

When (6) and (7) are set equal to zero and simplified, we obtain the "Normal" equations for the linear model [2]:

$$n a + b \sum_{i=1}^n X_i = \sum_{i=1}^n Y_i \tag{8}$$

$$a \sum_{i=1}^n X_i + b \sum_{i=1}^n X_i^2 = \sum_{i=1}^n X_i Y_i \tag{9}$$

The only unknowns in equation (8) and (9) are  $a$  and  $b$ , and they should be solved for them simultaneously, thus deriving (or estimating) the linear model. This is so because all the other values of equations (8) and (9) come from the given data, where:

- $n$  = number of ordered pairs  $(X_i, Y_i)$
- $\sum_{i=1}^n X_i = X_1 + X_2 + \dots + X_n$  = sum of the X values
- $\sum_{i=1}^n Y_i = Y_1 + Y_2 + \dots + Y_n$  = sum of the y values
- $\sum_{i=1}^n X_i^2 = X_1^2 + X_2^2 + \dots + X_n^2$  = sum of the given X values, which are first squared
- $\sum_{i=1}^n X_i Y_i = X_1 Y_1 + X_2 Y_2 + \dots + X_n Y_n$  = sum of the products of the  $X_i$  and  $Y_i$  values in each ordered pair.

Note: The values of  $(a)$  and  $(b)$  obtained from the Normal equations correspond to a minimum value for the quadratic function  $Q(a,b)$  given by equation (4), as can be easily demonstrated by using the optimization methodology of differential calculus for functions of two independent variables.

To derive the Normal equations for the multivariate model, we take partial derivatives of  $Q(b_1, b_2, \dots, b_p)$  with respect to  $b_1, b_2, b_3, \dots,$  and  $b_p$  respectively, and set each equal to 0. But, because the algebraic results are too complicated we express them in the following matrix form [5]:

$$(X'X) \mathbf{b} = X'Y \tag{10}$$

where:  $X$  = Matrix formed from the values of the p independent variables ( $X$  is  $n \times p$ ), and:

$$X = \begin{pmatrix} X_1 & X_2 & X_3 & \dots & X_p \\ 1 & X_{21} & X_{31} & \dots & X_{p1} \\ 1 & X_{22} & X_{32} & \dots & X_{p2} \\ 1 & X_{23} & X_{33} & \dots & X_{p3} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & X_{2n} & X_{3n} & \dots & X_{pn} \end{pmatrix}$$

$X'$  = Transposed of  $X$  matrix ( $X'$  is  $p \times n$ ), and:

$$X' = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ X_{21} & X_{22} & X_{23} & \dots & X_{2n} \\ X_{31} & X_{32} & X_{33} & \dots & X_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ X_{p1} & X_{p2} & X_{p3} & \dots & X_{pn} \end{pmatrix}$$

and the product  $(X'X)$  is  $p \times p$ .  $Y$  is the  $n \times 1$  vector consisting of the given  $Y$  values while  $(X'Y)$  is a  $p \times 1$  vector and is equal to:

$$X'Y = \begin{pmatrix} y_1 + y_2 + y_3 + \dots + y_n \\ y_1 X_{21} + y_2 X_{22} + y_3 X_{23} + \dots + y_n X_{2n} \\ y_1 X_{31} + y_2 X_{32} + y_3 X_{33} + \dots + y_n X_{3n} \\ \dots \\ y_1 X_{p1} + y_2 X_{p2} + y_3 X_{p3} + \dots + y_n X_{pn} \end{pmatrix}$$

The matrix solution to equation (10) is given by:

$$\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \dots \\ b_p \end{pmatrix} = (X'X)^{-1} (X'Y) \tag{11}$$

where  $(X'X)^{-1}$  is the inverse of matrix  $(X'X)$  and can be found using either the Gauss-Elimination method or the Adjoint Matrix method.

Note: If the  $X'X$  matrix is diagonal, finding the inverse is trivial.

For example, if  $X^T X = \begin{pmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{pmatrix}$ , the inverse is

$$\text{given by: } (X^T X)^{-1} = \begin{pmatrix} \frac{1}{d_1} & 0 & 0 \\ 0 & \frac{1}{d_2} & 0 \\ 0 & 0 & \frac{1}{d_3} \end{pmatrix}$$

To complete the estimation of the linear model we need to find the standard deviation for a,  $\sigma(a)$ , and b,  $\sigma(b)$ , which are needed for the testing of the significance of the model. The standard deviations,  $\sigma(a)$ , and  $\sigma(b)$ , are given by:

$$\sigma(a) = \frac{\hat{\sigma}}{\sqrt{n}} \left[ \frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]^{1/2} = \frac{\hat{\sigma}}{\sqrt{n}} \left[ \frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}} \right]^{1/2} \quad (12)$$

$$\text{and } \sigma(b) = \frac{\hat{\sigma}}{\left[ \sum_{i=1}^n (x_i - \bar{x})^2 \right]^{1/2}}, \quad (13)$$

$$\text{where: } \hat{\sigma} = \left[ \frac{\sum_{i=1}^n y_i^2 - a \sum_{i=1}^n y_i - b \sum_{i=1}^n x_i y_i}{n-2} \right]^{1/2} \quad (14)$$

The a and b in equation (14) come from the solution of equations (8) and (9) while  $\sum_{i=1}^n y_i^2$ ,  $\sum_{i=1}^n y_i$ , and  $\sum_{i=1}^n x_i y_i$  come directly from the given bivariate data.

To complete the estimation of the multivariate model we need to first find the variances  $V(b_1)$ ,  $V(b_2)$ ,  $V(b_3)$ , ...,  $V(b_p)$  from which then we can obtain:  $\sigma(b_1) = \sqrt{V(b_1)}$ , ...,  $\sigma(b_p) = \sqrt{V(b_p)}$ .

The variance of the

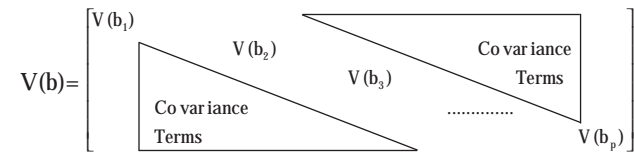
b vector  $(b = \begin{pmatrix} b_1 \\ b_2 \\ \dots \\ b_p \end{pmatrix})$  is given by [4]:

$$V(b) = (X^T X)^{-1} \hat{\sigma}^2 \quad (15)$$

$$\text{where } \hat{\sigma}^2 = \frac{Y^T Y - b^T X^T Y}{n-p} = \frac{Q^*}{n-p} \quad (16)$$

and  $Y^T Y = \sum_{i=1}^n y_i^2 = y_1^2 + y_2^2 + \dots + y_n^2$ ,  $X^T Y$  was derived in equation (10) and  $b^T$  is the transposed of vector b, or  $b^T = (b_1 \ b_2 \ \dots \ b_p)$ .

After equation (16) is substituted into equation (15) and the multiplication of the matrix  $(X^T X)^{-1}$  by  $\hat{\sigma}^2$  takes place,  $V(b)$  assumes the form:



Therefore, the variances  $V(b_1), V(b_2), \dots, V(b_p)$  are the values along the main diagonal of the  $V(b)$  matrix, while the off-the-main-diagonal terms are covariance terms.

Note: At this point, if we are estimating the linear model we have, for the given data set: a, b and  $\sigma(a)$  and  $\sigma(b)$ . If we are estimating the multivariate model we have, for the given data set:

$b_1, b_2, b_3, \dots, b_p$  and  $\sigma(b_1), \sigma(b_2), \sigma(b_3), \dots, \sigma(b_p)$ .

**Model Testing**

Now that our model of interest has been estimated, we need to test for the significance of the terms found in the estimated model. This is very important because the results of this testing will determine the final equation which will be retained and used for forecasting purposes.

**Testing the Linear Model**

The testing consists of the following steps:

a.) Testing for the significance of each factor separately [6]

Here we test the hypotheses:

1.)  $H_0: \beta = 0$  vs  $H_1: \beta \neq 0$  and 2)  $H_0: \alpha = 0$  vs.  $H_1: \alpha \neq 0$  based on our knowledge of b,  $\sigma(b)$ , a, and  $\sigma(a)$ .

If  $n \geq 30$ , we calculate  $Z_b^* = \frac{b}{\sigma(b)}$  and  $Z_a^* = \frac{a}{\sigma(a)}$

and compare each to  $Z_{\alpha/2}$  (where  $Z_{\alpha/2}$  is a value obtained from the standard Normal Table when  $\alpha$  (or  $1 - \alpha$ ) is specified).

For example if  $\alpha = 0.05$ ,  $Z_{\alpha/2} = Z_{0.025} = 1.96$ ; if  $\alpha = 0.10$ ,  $Z_{\alpha/2} = Z_{0.05} = 1.645$ ; if  $\alpha = 0.02$ ,  $Z_{\alpha/2} = Z_{0.01} = 2.33$  and if  $\alpha = 0.01$ ,  $Z_{\alpha/2} = Z_{0.005} = 2.58$ . If  $Z_b^* > Z_{\alpha/2}$  (or  $Z_b^* < -Z_{\alpha/2}$ ), the hypothesis  $H_0: \beta = 0$  is rejected and we conclude that  $\beta \neq 0$  and the term bX (in the estimated model  $\hat{y} = a + bX$ ) is important for the calculation of the value of y. Similarly, if  $Z_a^* > Z_{\alpha/2}$  (or  $Z_a^* < -Z_{\alpha/2}$ ),  $H_0: \alpha = 0$  is rejected, and we conclude that the linear equation  $\hat{y} = a + bX$  does not go through the origin. If  $n < 30$ , we calculate  $t_b^* = \frac{b}{\sigma(b)}$  and  $t_a^* = \frac{a}{\sigma(a)}$  and compare each to  $t_{n-2(\alpha/2)}$ , for a given  $\alpha$  value, where  $t_{n-2(\alpha/2)}$  is obtained from the t-distribution table, with the same interpretation for  $H_0: \beta = 0$  and  $H_0: \alpha = 0$  as above.

But, instead of testing, we can construct confidence intervals for  $\beta$  and  $\alpha$  using the equations:

$$P[b - Z_{\alpha/2} \sigma(b) \leq \beta \leq b + Z_{\alpha/2} \sigma(b)] = 1 - \alpha \quad (17)$$

$$\text{and } P[a - Z_{\alpha/2} \sigma(a) \leq \alpha \leq a + Z_{\alpha/2} \sigma(a)] = 1 - \alpha, \text{ if } n \geq 30, \quad (18)$$

$$\text{or: } P[b - t_{n-2(\alpha/2)} \sigma(b) \leq \beta \leq b + t_{n-2(\alpha/2)} \sigma(b)] = 1 - \alpha \quad (19)$$

$$\text{and } P[a - t_{n-2(\alpha/2)} \sigma(a) \leq \alpha \leq a + t_{n-2(\alpha/2)} \sigma(a)] = 1 - \alpha, \text{ if } n < 30 \quad (20)$$

If the hypothesized values:  $\beta = 0$  falls inside the confidence intervals given by equations (17) or (19), or  $\alpha = 0$  falls inside the confidence intervals given by equations (18) or (20), the corresponding hypotheses  $H_0: \beta = 0$  and  $H_0: \alpha = 0$  are not rejected and we conclude that  $\beta = 0$  (and  $b = 0$  and the term

bX is not important for the calculation of y) and  $\alpha = 0$  (i.e.  $a = 0$  and the line goes through zero). If for a given data set, we performed the above discussed tests, we will obtain one of 4 possible conclusions:

- A)  $H_0: \beta = 0$  and  $H_0: \alpha = 0$  are both rejected. Therefore  $\beta \neq 0$ , and  $\alpha \neq 0$ , and both the terms a and bX are important to the calculation of y. In this case the final equation is  $\hat{y} = a + bX$ , with both terms staying in the equation.
- B)  $H_0: \beta = 0$  is rejected, but  $H_0: \alpha = 0$  is not rejected. Therefore  $\beta \neq 0$  but  $\alpha = 0$  and the term a is not important to the calculation of y. In this case the final equation is  $\hat{y} = bX$ , with the term a dropping out of the equation.
- C)  $H_0: \beta = 0$  is not rejected but  $H_0: \alpha = 0$  is rejected. Therefore  $\beta = 0$  and the term bX is not important for the calculation of y, while  $a \neq 0$  and is important to the calculation of y. In this case the final equation is  $\hat{y} = a$ , with the term bX dropping out of the equation.
- D)  $H_0: \beta = 0$  and  $H_0: \alpha = 0$  are both not rejected. Therefore  $\beta = 0$ , and  $\alpha = 0$ , and both terms a and bX are not important to the calculation of y. In this case the final equation will be  $\hat{y} = 0$ , with both terms a and bX dropping out of the equation.

b.) Testing for the Significance of the Entire Equation [1]

To perform this test, which consists of testing the hypothesis:

$$H_0: \alpha = \beta = 0 \text{ vs } H_1: \alpha \text{ and } \beta \text{ are not both equal to } 0$$

or

$$H_0: \text{The entire regression equation is not significant}$$

vs

$$H_1: \text{The entire regression equation is significant.}$$

For a given bivariate data set and a given  $\alpha$  value, we need to first calculate:

$$\text{TOTAL SUM OF SQUARES: } TSS = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n} \quad (21)$$

$$\text{REGRESSION SUM OF SQUARES} = RSS_b = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = b^2 \sum_{i=1}^n (x_i - \bar{x})^2$$

$$b^2 \left[ \sum_{i=1}^n x_i^2 - \frac{\left( \sum_{i=1}^n x_i \right)^2}{n} \right] \quad (22)$$

ERROR SUM OF SQUARES = ESS =

$$\sum_{i=1}^n (y_i - \hat{y})^2 = Q^* = \sum_{i=1}^n y_i^2 - a \sum_{i=1}^n y_i - b \sum_{i=1}^n x_i y_i \quad (23)$$

$$\text{SUM OF SQUARES DUE TO THE CONSTANT} = SS_a = \frac{\left( \sum_{i=1}^n y_i \right)^2}{n} \quad (24)$$

Then we calculate:  $F_{\text{Total}}^* = \frac{(RSS_b + SS_a) / 2}{ESS / n - 2} \quad (25)$

and compare to  $F_{n-2}^2(\alpha)$ , which is a tabulated value, for a specified  $\alpha$  value. If  $F_{\text{Total}}^* > F_{n-2}^2(\alpha)$ , we reject  $H_0$  and conclude that the entire regression equation (i.e.  $\hat{y} = a + bX$ ) or that both the constant term  $a$ , and the factor  $X$  (and term  $bX$ ) are significant to the calculation of the  $y$  value.

Note 1: When TSS,  $RSS_b$ , and ESS are known, we can also define the coefficient of determination  $R^2$ , where:

$$R^2 = \frac{RSS_b}{TSS} = 1 - \frac{ESS}{TSS} \quad (\text{where } 0 \leq R^2 \leq 1) \quad (26)$$

which tells us how well the regression equation  $\hat{y} = a + bX$  fits the given bivariate data [6]. A value of  $R^2$  close to 1 implies a good fit.

Note 2:  $r = \text{correlation coefficient} = \sqrt{R^2} \quad (27)$

c.) A Bivariate Example

A sample of 5 adult men for whom heights and weights are measured gives the following results:

X=H	y=W	x <sup>2</sup> =H <sup>2</sup>	y <sup>2</sup> =W <sup>2</sup>	xy=HW
64	130	64 <sup>2</sup>	130 <sup>2</sup>	64x130
65	145	65 <sup>2</sup>	145 <sup>2</sup>	65x145
66	150	66 <sup>2</sup>	150 <sup>2</sup>	66x150
67	165	67 <sup>2</sup>	165 <sup>2</sup>	67x165
68	170	68 <sup>2</sup>	170 <sup>2</sup>	68x170

Given bivariate data set

For this bivariate data set we have:  $n = 5$

$$\begin{aligned} \sum_{i=1}^5 x_i &= 64 + 65 + 66 + 67 + 68 = 330 \\ \sum_{i=1}^5 x_i^2 &= 64^2 + 65^2 + 66^2 + 67^2 + 68^2 = 21,790 \\ \sum_{i=1}^5 y_i &= 130 + 145 + 150 + 165 + 170 = 760 \\ \sum_{i=1}^5 y_i^2 &= 130^2 + 145^2 + 150^2 + 165^2 + 170^2 = 116,550 \\ \sum_{i=1}^5 x_i y_i &= (64 \times 130) + (65 \times 145) + (66 \times 150) + (67 \times 165) \\ &\quad + (68 \times 170) = 50,260 \end{aligned}$$

To obtain the linear equation  $\hat{y} = a + bX$ , we substitute the values of:

$$\begin{aligned} n, \sum_{i=1}^5 x_i, \sum_{i=1}^5 x_i^2, \sum_{i=1}^5 x_i y_i \text{ to equations (8) and (9) and obtain:} \\ 5a + 330b = 760 \\ 330a + 21,790b = 50,260 \end{aligned}$$

When these equations are solved simultaneously we obtain:

$a = -508$  and  $b = 10$ , and the regression equation is:  $\hat{y} = a + bX = -508 + 10X$ . Then, using the values of  $a = -508$ ,  $b=10$ , and  $\sum_{i=1}^5 y_i, \sum_{i=1}^5 y_i^2, \sum_{i=1}^5 x_i y_i$  we obtain from equation (14):

$$\hat{\sigma} = \left[ \frac{116,550 - (-508)(760) - (10)(50,260)}{5 - 2} \right]^{1/2} = \left[ \frac{30}{3} \right]^{1/2} = \sqrt{10}$$

and from equations (12) and (13):

$$\sigma(a) = \frac{\sqrt{10}}{\sqrt{5}} \left[ \frac{21,790}{21,790 - \frac{(330)^2}{5}} \right]^{1/2} = \sqrt{2} \left[ \frac{21,790}{10} \right]^{1/2} = \left[ \frac{2 \times 21,790}{10} \right]^{1/2} = \sqrt{4358} = 66.015$$

$$\text{and } \sigma(b) = \frac{\sqrt{10}}{\left[ 21,790 - \frac{(330)^2}{5} \right]^{1/2}} = \frac{\sqrt{10}}{[10]^{1/2}} = \frac{\sqrt{10}}{\sqrt{10}} = 1$$

Since  $n = 5 < 30$ ,  $a$  and  $b$  are distributed as  $t_{n-2} = t_3$  variables and when  $\alpha = 0.05, t_3(\alpha/2) = t_3(0.025) = \pm 3.1824$ .

Then the hypotheses:  $H_0: \beta = 0$  vs.  $H_1: \beta \neq 0$ , and  $H_0: \alpha = 0$

vs.  $H_1: \alpha \neq 0$  are both rejected because :

$$t_a^* = \frac{a}{\sigma(a)} = \frac{-508}{66,015} = -7.695 < -3.1824 \text{ and:}$$

$$t_b^* = \frac{b}{\sigma(b)} = \frac{10}{1} = 10 > 3.1824.$$

Therefore, the final equation is  $\hat{y} = a + bX = -508 + 10X$ . To test for the significance of the entire equation, and to calculate the coefficient of determination, we first evaluate TSS,  $RSS_b$ , ESS, and  $SS_a$ , using equations (21) - (24) and obtain:

$$\begin{aligned} TSS &= 116,550 - \frac{(760)^2}{5} = 1030 \\ RSS_b &= 10^2 \left[ 21,790 - \frac{(330)^2}{5} \right] = 10^2(10) = 1000 \\ ESS &= 116,550 - (-508)(760) - 10(50,260) = 30 \\ SS_a &= \frac{(760)^2}{5} = 115,520. \end{aligned}$$

From equation (26) we obtain  $R^2 = \frac{1000}{1030} \approx 0.971$ , which tells

us that 97% of the variation in the values of  $Y$  can be explained (or accounted for) by the variable  $X$  included in the regression equation and only 3% is due to other factors. Since  $R^2$  is close to 1, the fit of the equation to the data is very good.

Note: The correlation coefficient  $r$ , which measures the strength of the linear relationship between  $Y$  and  $X$  is related to the coefficient of determination by:  $r = \sqrt{R^2} = \sqrt{0.97} = 0.985$  for this example. Clearly  $X$  and  $Y$  are very strongly linearly related.

Using equation (25) we obtain:

$$F_{\text{Total}}^* = \frac{(RSS_b + SS_a) / 2}{ESS / n - 2} = \frac{(1000 + 115,520) / 2}{30 / 3} = \frac{58,260}{10} = 5.826.$$

When  $F_{\text{Total}}^*$  is compared to  $F_{n-2}^2(\alpha) = F_3^2(\alpha) = \begin{cases} 10.13 & \text{if } \alpha = 0.05 \\ 34.12 & \text{if } \alpha = 0.01 \end{cases}$

$H_0$  (= The entire regression equation is not significant) is rejected, and we conclude that the entire regression equation is significant.

MINITAB Solution to the Problem [1]

We enter the given data and issue the regression command as shown:

```
MTB > Set C1
DATA> 64 65 66 67 68
DATA> end
MTB > set C2
DATA> 130 145 150 165 170
DATA> end
MTB > Name C1 'X' C2 'Y'
MTB > REGRESS 'Y' 1 'X'
```

and obtain the MINITAB output:

```
Regression Analysis: Y versus X

The regression equation is
Y = - 508 + 10.0 X

Predictor      Coef      SE Coef      T          P
Constant     -508.00      66.02      -7.70      0.005
X              10.000       1.000       10.00      0.002
S = 3.162    R-Sq = 97.1%    R-Sq(adj) = 96.1%
```

```
Analysis of Variance

Source      DF      SS      MS      F      P
Regression    1    1000.0    1000.0    100.00    0.002
Residual      3      30.0     10.0
Total         4    1030.0
```

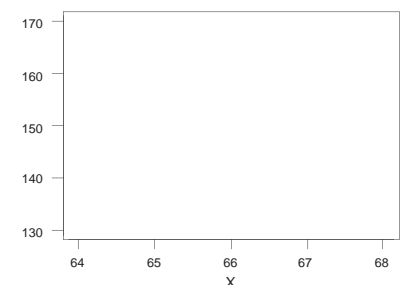
```
MTB > CORRELATE C2 C1
```

Correlations: Y, X

```
Pearson correlation of Y and X = 0.985
P-Value = 0.002
```

```
MTB > PLOT C2 * C1
```

Plot Y \* X



When we compare the MINITAB and hand solutions, they are identical. We obtain the same equation  $\hat{y} = -508 + 10X$ , the same standard deviations for a and b (under SE Coefficient) and the same t values, the same  $R^2$ , the same  $s = \hat{\sigma}$  and  $\hat{\sigma}^2 = 10$ . Notice also that an Analysis of Variance table provides the values for  $RSS_b$ ,  $ESS$ , and  $TSS$ . The only value missing is  $SS_a$ , which can be easily calculated from

$$SS_a = \frac{\left(\sum_{i=1}^n y_i\right)^2}{n}$$

The MINITAB solution also gives a p-value for each coefficient. The p-value is called the "Observed Level of Significance" and represents the probability of obtaining a value more extreme than the value of the test statistic. For example the p-value for the predictor X is calculated as  $p = 0.002$ , and it is given by:

$$p\text{-value} = P(t > t^* = 10) = \int_{10}^{\infty} f(t) dt = 0.002 \quad (28)$$

The p-value has the following connection to the selected  $\alpha$ -value.

- If  $p > \alpha$ , do not reject  $H_0$
- If  $p < \alpha$ , reject  $H_0$
- Since  $p = 0.002 < \alpha = 0.05$ ,  $H_0: \beta = 0$  will be rejected.

**Testing the MULTIVARIATE MODEL =**

$$\hat{y} = b_1 X_1 (=1) + b_2 X_2 + \dots + b_p X_p$$

Testing of this model consists of the following 3 steps:

a.) To test for the significance of each factor separately

The values of  $b_1, b_2, b_3, \dots, b_p$  are obtained from equation (11) and the values of  $\sigma(b_1), \sigma(b_2), \sigma(b_3), \dots, \sigma(b_p)$  from equations (15) and (16). Then, we test for the significance of each factor separately by either:

1.) Testing the hypotheses :  $H_0: \beta_i = 0$  vs.  $H_1: \beta_i \neq 0$  by calculating  $Z_i^* = \frac{b_i}{\sigma(b_i)}$  or  $t_i^* = \frac{b_i}{\sigma(b_i)}$ . Then,  $H_0: \beta_i = 0$  is rejected if  $Z_i^* > Z_{\alpha/2}$  (or if  $Z_i^* < -Z_{\alpha/2}$ ), when  $n \geq 30$  or if  $t_i^* > t_{n-p(\alpha/2)}$  (or if  $t_i^* < -t_{n-p(\alpha/2)}$ ), if  $n < 30$  or

2.) By constructing the confidence interval

$$P[b_i - Z_{\alpha/2} \sigma(b_i) \leq \beta_i \leq b_i + Z_{\alpha/2} \sigma(b_i)] = 1 - \alpha, \text{ if } n \geq 30$$

or

$$P[b_i - t_{n-p(\alpha/2)} \sigma(b_i) \leq \beta_i \leq b_i + t_{n-p(\alpha/2)} \sigma(b_i)] = 1 - \alpha, \text{ if } n < 30.$$

If the value  $\beta_i = 0$  is outside of these confidence intervals,

$H_0: \beta_i = 0$  is rejected.

b.) To test for the significance of the entire regression (including the constant)

The hypotheses being tested are:

- $H_0: \beta_1 = \beta_2 = \beta_3 = \dots = \beta_p = 0$  vs.  $H_1$ : The  $\beta_i$  are not all equal to 0
- or  $H_0$ : The entire regression (including the constant) is not significant
- vs.
- $H_1$ : The entire regression (including the constant) is significant.

It is carried out by calculating:

$$F_{\text{Total}}^* = \frac{RSS / \text{DOF}}{ESS / \text{DOF}} = \frac{b' X' Y / p}{(Y' Y - b' X' Y) / n - p} \text{ and comparing to } F_{n-p}^p(\alpha).$$

If  $F_{\text{Total}}^* > F_{n-p}^p(\alpha)$ ,  $H_0$  is rejected and we conclude that the entire regression (including the constant) is significant (to the calculation of the Y value).

c.) To test for the significance of the entire regression (excluding the constant)

The hypotheses being tested are:

- $H_0: \beta_2 = \beta_3 = \dots = \beta_p = 0$  vs.  $H_1: \beta_2, \beta_3, \dots, \beta_p$  are not all equal to 0
- or  $H_0$ : The entire regression (excluding the constant) is not significant
- vs.
- $H_1$ : The entire regression (excluding the constant) is significant.

It is carried out by calculating:

$$F_{\text{Total}-\beta_1}^* = \frac{(RSS - SS_a) / p - 1}{ESS / n - p} = \frac{(b' X' Y - SS_a) / p - 1}{(Y' Y - b' X' Y) / n - p}$$

and comparing it to  $F_{n-p}^{p-1}(\alpha)$ . If  $F_{\text{Total}-\beta_1}^* > F_{n-p}^{p-1}(\alpha)$ ,  $H_0$  is rejected and we conclude that the entire regression (excluding the constant) is significant.

**Determination of the final equation**

Any variable  $X_i$ , for which the hypothesis:

$H_0: \beta_i = 0$  (vs.  $H_1: \beta_i \neq 0$ ) is not rejected, is to be dropped from the regression equation.

The remaining terms are used to form the "Final Equation" which is then retained and used for Prediction/Forecasting purposes.

**A Multivariate Example**

The sales manager of a certain firm believes that Sales Ability depends on a salesperson's Verbal Reasoning Ability and Vocational Interest. The sales manager is interested in constructing a regression equation to use in future hiring, to predict a candidate's success as a salesperson, i.e. the sales manager wants to derive the regression equation.

$$Y = \text{Sales Ability} \\ = b_1 X_1 (=1) + b_2 X_2 (= \text{Verbal Reasoning}) + b_3 X_3 \\ (= \text{Vocational Interest})$$

To verify this belief, 10 salespersons are selected at random from the staff and given 2 tests: One for Verbal Reasoning Ability, the other for Vocational Interest. The results are shown below:

Salesperson	1	2	3	4	5	6	7	8	9	10
Y= Average sales in month	1	1	1	2	2	4	3	5	6	6
X <sub>2</sub> = Verbal Reasoning Ability	1	1	2	2	3	3	4	4	5	5
X <sub>3</sub> = Vocational Interest	2	1	1	3	2	4	3	5	4	6

- 1.) Find the estimated regression equation for this data set.
- 2.) Find TSS,  $RSS_b$ ,  $Q^* = ESS$ , and  $SS_a$
- 3.) Find  $\sigma(b_1), \sigma(b_2), \sigma(b_3)$
- 4.) Test the hypotheses: a)  $H_0: \beta_1 = 0$  vs.  $H_1: \beta_1 \neq 0$   
b)  $H_0: \beta_2 = 0$  vs.  $H_1: \beta_2 \neq 0$   
c)  $H_0: \beta_3 = 0$  vs.  $H_1: \beta_3 \neq 0$
- 5.) Construct 95% Confidence intervals for:  $\beta_1, \beta_2, \beta_3$
- 6.) Test the hypothesis:  $H_0$ : The entire regression equation (including the constant) is not significant (i.e.  $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ )

- 7.) Test the hypothesis:  $H_0$ : The entire regression equation (excluding the constant) is not significant (i.e.  $H_0: \beta_2 = \beta_3 = 0$ )
- 8.) Determine the "final regression" equation.

**Solutions:**

The matrices X, X' and their product X'X are given below. Note that the values of variable  $X_1$  are all equal to 1 because we want our regression equation to have a constant term.

$$X'X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 2 & 3 & 3 & 4 & 4 & 5 & 5 \\ 2 & 1 & 1 & 3 & 2 & 4 & 3 & 5 & 4 & 6 \end{bmatrix} = \begin{bmatrix} 10 & 30 & 31 \\ 30 & 110 & 111 \\ 31 & 111 & 121 \end{bmatrix}$$

Since  $X'X$  is not a diagonal matrix ( $A = \begin{pmatrix} a_1 & 0 & 0 \\ 0 & a_2 & 0 \\ 0 & 0 & a_3 \end{pmatrix}$  is a diagonal matrix and then  $A^{-1} = \begin{pmatrix} 1/a_1 & 0 & 0 \\ 0 & 1/a_2 & 0 \\ 0 & 0 & 1/a_3 \end{pmatrix}$ )

by inspection) we will use the Adjoint Matrix method to find the inverse of matrix  $(X'X)$  [3].

$$\text{The result is } (X'X)^{-1} = \frac{1}{1740} \begin{pmatrix} 989 & -189 & -80 \\ -189 & 249 & -180 \\ -80 & -180 & 200 \end{pmatrix}$$

$$\text{Also, } X'Y = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 2 & 3 & 3 & 4 & 4 & 5 & 5 \\ 2 & 1 & 1 & 3 & 2 & 4 & 3 & 5 & 4 & 6 \end{bmatrix} \begin{bmatrix} 31 \\ 118 \\ 124 \end{bmatrix}$$

Then:  $b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = (X'X)^{-1} \cdot (X'Y)$

$$= \frac{1}{1740} \begin{pmatrix} 989 & -189 & -80 \\ -189 & 249 & -180 \\ -80 & -180 & 200 \end{pmatrix} \begin{bmatrix} 31 \\ 118 \\ 124 \end{bmatrix} = \frac{1}{1740} \begin{bmatrix} -1563 \\ 1203 \\ 1080 \end{bmatrix}$$

and  $b_1 = \frac{-1563}{1740} = -0.8982$ ,  $b_2 = \frac{1203}{1740} = 0.6913$ ,  $b_3 = \frac{1080}{1740} = 0.6207$

Therefore, the regression equation is:

1.)  $\hat{y} = b_1 X_1 (=1) + b_2 X_2 + b_3 X_3 = -0.8982 + 0.6913X_2 + 0.6207 X_3$

2.)  $TSS = Y'Y = \sum_{i=1}^{10} y_i^2 = y_1^2 + y_2^2 + \dots + y_{10}^2 = 133.00$

$$RSS_b = b' X' Y = [-0.8982 \quad 0.6913 \quad 0.6207] \begin{bmatrix} 31 \\ 118 \\ 124 \end{bmatrix} = 130.7074$$

$$ESS = Q^* = Y'Y - b' X' Y = 133.0000 - 130.7047 = 2.2953$$

$$SS_a = \frac{\left(\sum_{i=1}^n y_i\right)^2}{n} = \frac{(y_1 + y_2 + \dots + y_{10})^2}{10} = \frac{(31)^2}{10} = \frac{961}{10} = 96.1$$

3.) From equation (16)  $\hat{\sigma}^2 = \frac{ESS}{n-p} = \frac{Q^*}{n-p} = \frac{2.2953}{10-3} = \frac{2.2953}{7} = 0.3279$

Then  $V(b_1) = \frac{989}{1740} \hat{\sigma}^2 = \frac{989}{1740} (0.3279) = 0.1864$ , and  $\sigma(b_1) = 0.4316$

$$V(b_2) = \frac{249}{1740} \hat{\sigma}^2 = \frac{249}{1740} (0.3279) = 0.0469$$
, and  $\sigma(b_2) = 0.2163$

$$V(b_3) = \frac{200}{1740} \hat{\sigma}^2 = \frac{200}{1740} (0.3279) = 0.0377$$
, and  $\sigma(b_3) = 0.1942$

4.) Test:  $H_0: \beta_1 = 0$  vs.  $H_1: \beta_1 \neq 0$ ,  $H_0: \beta_2 = 0$  vs.  $H_1: \beta_2 \neq 0$ ,  
 $H_0: \beta_3 = 0$  vs.  $H_1: \beta_3 \neq 0$ .

Calculate:  $t_1^* = \frac{b_1}{\sigma(b_1)} = \frac{-0.8982}{0.4316} = -2.081$

$t_2^* = \frac{b_2}{\sigma(b_2)} = \frac{0.6913}{0.2163} = 3.192$

$t_3^* = \frac{b_3}{\sigma(b_3)} = \frac{0.6207}{0.1942} = 3.196$

and compare each against:  
 $t_{n-p(\alpha/2)} =$   
 $t_{10-3(0.05/2)} =$   
 $t_{7(0.025)} = 2.365$

Since  $-2.365 < t_1^* = -2.081 < 2.365$ , do not reject:  $H_0: \beta_1 = 0$ ;

Therefore,  $\beta_1 = 0$

$$t_2^* = 3.192 > 2.365$$
, reject  $H_0: \beta_2 = 0$ ;

Therefore,  $\beta_2 \neq 0$

$$t_3^* = 3.196 > 2.365$$
, reject  $H_0: \beta_3 = 0$ ;

Therefore,  $\beta_3 \neq 0$

5.) Since  $n = 10 < 30$  the sampling distributions of  $b_1, b_2, b_3$  are  $t_{n-p} = t_7$ . Therefore the confidence intervals are to be obtained from:

$$P[b_1 - t_{7(\alpha/2)} \sigma(b_1) \leq \beta_1 \leq b_1 + t_{7(\alpha/2)} \sigma(b_1)] = 1 - \alpha$$

and they are (when the appropriate values are substituted in):

a)  $P[-1.919 \leq \beta_1 \leq 0.123] = 0.95$ ; since  $\beta_1 = 0$  falls *inside* this confidence interval, we do not reject  $H_0: \beta_1 = 0$ ; Therefore,  $\beta_1 = 0$

b)  $P[0.180 \leq \beta_2 \leq 1.202] = 0.95$ ; since  $\beta_2 = 0$  falls *outside* this confidence interval we reject  $H_0: \beta_2 = 0$ ; Therefore,  $\beta_2 \neq 0$

c)  $P[0.161 < \beta_3 < 1.080] = 0.95$ ; since  $\beta_3 = 0$  falls *outside* this confidence interval we reject  $H_0: \beta_3 = 0$ ; Therefore,  $\beta_3 \neq 0$

6.) We calculate:  
 $F_{Total}^* = \frac{(b' X' Y) / p}{(Y'Y - b' X' Y) / (n-p)} = \frac{130.7047 / 3}{(2.2953) / 7} = \frac{43.5682}{0.3279} = 132.87$

and compare to:  $F_{n-p}^p(\alpha) = F_7^3(\alpha) = \begin{cases} 4.35, & \text{if } \alpha = 0.05 \\ 8.45, & \text{if } \alpha = 0.01 \end{cases}$

Since  $F_{Total}^* > F_7^3(\alpha)$ , for both  $\alpha$  values,  $H_0$ : (The entire equation including the constant is not significant) is rejected, and we conclude that the entire equation is significant.

7.) We calculate:

$$F_{Total-\beta_1}^* = \frac{(b' X' Y - SS_a) / (P-1)}{(Y'Y - b' X' Y) / (n-p)} = \frac{(130.7047 - 96.1) / 2}{(2.2953) / 7} = \frac{34.6047}{2} = \frac{17.30235}{0.3279} = 52.77$$

and compare to:  $F_{n-p}^{p-1}(\alpha) = F_7^2(\alpha) = \begin{cases} 4.74, & \text{if } \alpha = 0.05 \\ 9.55, & \text{if } \alpha = 0.01 \end{cases}$

Since  $F_{Total-\beta_1}^* > F_7^2(\alpha)$ , for both  $\alpha$  values, we reject  $H_0$  ( $H_0$ : The entire equation (excluding the constant) is not significant) and conclude that the entire regression equation, excluding the constant, is significant.

8.) Because  $H_0: \beta_1 = 0$  is not rejected but  $H_0: \beta_2 = 0$  and  $H_0: \beta_3 = 0$  are rejected, the final regression equation is:  $\hat{y} = b_2 X_2 + b_3 X_3 = 0.6914 X_2 + 0.6207 X_3$  which should be retained and used to predict the sales ability of future candidates.

**MINITAB Solution to the Problem [1]**

We enter the given data and issue the regression command as shown:

```
MTB > Set C1
DATA> 1 1 1 2 2 4 3 5 6 6
DATA> end
MTB > Set C2
DATA> 1 1 2 2 3 3 4 4 5 5
DATA> end
MTB > Set C3
DATA> 2 1 1 3 2 4 3 5 4 6
DATA> end
MTB > print C1 C2 C3
```

**Data Display**

Row	C1	C2	C3
1	1	1	2
2	1	1	1
3	1	2	1
4	2	2	3
5	2	3	2
6	4	3	4
7	3	4	3
8	5	4	5
9	6	5	4
10	6	5	6

```
MTB > REGRESS c1 2 c2 c3;
SUBC> CONSTANT;
SUBC> BRIEF 2.
```

**Regression Analysis: C1 versus C2, C3**

The regression equation is  
 $C1 = -0.898 + 0.691 C2 + 0.621 C3$

Predictor	Coef	SE Coef	T	P
Constant	-0.8983	0.4320	-2.08	0.076
C2	0.6914	0.2168	3.19	0.015
C3	0.6207	0.1943	3.20	0.015

S = 0.5730      R-Sq = 93.8%      R-Sq(adj) = 92.0%

**Analysis of Variance**

Source	DF	SS	MS	F	P
Regression	2	34.602	17.301	52.69	0.000
Residual					
Error	7	2.298	0.328		
Total	9	36.900			

Source	DF	Seq SS
C2	1	31.250
C3	1	3.352

**Unusual Observations**

Obs	C2	C1	Fit	SE Fit	Residual	St Resid
9	5.00	6.000	5.041	0.359	0.959	2.15R

R denotes an observation with a large standardized residual

When we compare the MINITAB and hand solutions, they are identical, within the rounding of fractional numbers. From the p values calculated by MINITAB we see, once again, that  $H_0: \beta_1 = 0$  is not rejected because  $p = 0.076 > \alpha = 0.05$  but  $H_0: \beta_2 = 0$  and  $H_0: \beta_3 = 0$  are rejected because  $p = 0.015 < \alpha = 0.05$ . The F value of 52.69, shown in the MINITAB Analysis of Variance is almost identical to the hand value  $F_{Total-\beta_1}^* = 52.77$  obtained to test the hypothesis  $H_0$ : The entire

regression (excluding the constant) is not significant. Also, the values of  $\hat{\sigma}^2$  are almost identical.

### Conclusion

Reviewing our previous discussion we come to the following conclusions:

- The Linear Regression problem is relatively easy to solve and can be handled using algebraic methods.
- The Multivariate Regression problem is somewhat more complicated but can be handled efficiently using Matrix methods.
- Both problems can be solved easily using available statistical software, like MINITAB.
- Even though the solution to Regression problems can be obtained easily using MINITAB (or other statistical software), it is important to know what the hand methodology is and how it solves these problems before you can properly interpret and understand MINITAB's output.

### References

1. Black, K. *Business Statistics*. Wiley, 2004.
2. Canavos, G. C. *Applied Probability and Statistical Methods*. Little, Brown, 1984.
3. Childress, R. L., R. D. Gorsky and R. M. Witt. *Mathematics for Managerial Decisions*. Prentice Hall, 1989.
4. Draper, N. and H. Smith. *Applied Regression Analysis*. John Wiley & Sons, 1966.
5. Johnson, J. *Econometric Methods*. McGraw-Hill, 1963.
6. Pindyck, R. S. and D. L. Rubinfeld. *Econometric Models and Economic Forecasts, 2nd edition*. McGraw-Hill, 1981.

## Tobin College of Business launches **NEW** M.S. – Taxation to meet the increasing demand for specialists

As the demand for specialists in the field of taxation continues to increase, the need for targeted training almost becomes a requirement of future employment. With this need in mind, St. John's Master of Science in Taxation program was born. Its mission is to provide tax professionals with in-depth knowledge of the Internal Revenue Code, tax regulations, judicial decisions and Treasury rulings. Students learn to research tax questions, facilitate tax compliance and develop tax-planning strategies—all skills necessary for future career success.

- Our taxation faculty hold the highest academic credentials and have acquired professional experience in leading New York City accounting and law firms.
- Extensive course selection allows for a high degree of program focus.
- Taxation internships are available with "Big Five" firms for those not currently working full time.
- Classes are held Monday-Thursday evenings and Saturdays in convenient new York Metro locations to meet the needs of working tax professionals.
- All courses meet CPE credit requirements in NY and NJ.
- This program is accredited by AACSB International – The Association to Advance Collegiate Schools of Business.

### Master of Science in Taxation

This intensive program allows the candidate to select tax courses that are most applicable to his or her career. The degree requires 11 upper level graduate course (31 credits). Ten of these courses are in taxation and one may be a general business elective.

Required tax courses are:

- Research and Writing
- Corporations
- Partnerships
- Estates and Gifts
- Practice and Procedure
- Research Project

The first research course, (Tax Research and Writing) will be taken during the candidate's first semester. This course is designed to prepare the candidates for all subsequent tax courses by enabling them to research tax questions and to clearly communicate their findings.

The final research course, (Research Project), will allow the candidate, with guidance from the course instructor, to research a current tax topic and write a paper. The written work would be expected to be of publishable quality.

Elective tax courses include:

- Planning for High Net-Worth Individuals
- Income of Trusts and Estates
- Corporate Distribution/Liquidations/Reorganizations
- Consolidated Tax Returns
- Interstate Commerce
- Foreign Operations
- Compensation, Benefits and Retirement Plans
- Real Estate
- Financial Products
- Tax-Exempt Institutions
- Specialized Industries
- Tax Accounting
- Special Topics

The elective Business Course (three credits) allows the candidate to select one course from the offerings of the Graduate Division of the Tobin College of Business or one additional tax course.

### Admission Requirements

Candidates for admission must satisfy the following requirements:

- Possess an undergraduate or graduate degree in accounting or a related business field. Students must have completed core business courses or must take such courses before completing the degree.
- Successful completion of the GMAT (or an appropriate alternative examination). The successful completion of the uniform certified public accountants examination (CPA), the certified management accountants examination (CMA) or an equivalent examination could be used in lieu of the GMAT examination.

Please call for more information:

Adrian P. Fitzsimons, *Chairman*  
Department of Accounting  
and Taxation  
(718) 990-6461

*"An M.S. in Taxation can increase earning potential whether in public accounting, private companies, tax departments in municipal governments or the Internal Revenue Service."*

—Peter J. Tobin, *former Dean, The Peter J. Tobin College of Business former CFO of Chase Manhattan Bank*

# editorial review board

Suhail Abboushi  
Duquesne University

Raj Aggarwal  
Kent State University

Frederich Amling  
George Washington University

Rolph E. Anderson  
Drexel University

Karen Bahnik  
University of Wisconsin-  
Superior

Anthony Barbera  
SUNY Old Westbury

Nat R. Briscoe  
Northwestern State University

Bruce Buzby  
University of Connecticut-  
Stamford

Patrick Casabona  
St. John's University

John K.S. Chong  
University of Wisconsin –  
Parkside

James Cox  
Illinois State University

Randy F. Cray  
University of Wisconsin –  
Stevens Point

Dennis Duchon  
University of Texas –  
San Antonio

James Don Edwards  
University of Georgia

Fred Englander  
Fairleigh Dickinson University

P. Everett Fergenson  
Iona College

Allan Filley  
University of Wisconsin-  
Madison

Alan B. Flaschner  
University of Toledo

David Flynn  
Hofstra University

Frank Forman  
Department of Education

Eugene Garaventa  
CUNY Staten Island

Juan E. Gonzalez  
TVA

David J. Good  
Central Missouri State  
University

Donald Grunewald  
Iona College

David Hanson  
Duquesne University

Alfred C. Holden  
Fordham University

Gail Hudson  
Arkansas State University

Sharon Johnson  
Cedarville College

Tiffany Keller  
Purdue University

Bruce H. Kemelgor  
University of Louisville

Dominique Khactu  
University of North Dakota

James Kidney  
University of Southern  
Connecticut

Linda S. Kein  
University of Connecticut –  
Storrs

Deborah Kleiner  
St. John's University

Susan Kogler Hill  
Cleveland State University

Anthony C. Koh  
University of Toledo

Chidem Kurdas  
Penn State University – York

David Kurtz  
University of Arkansas

Kern Kwong  
California State University – LA

Jeffrey Lenn  
George Washington University

Edwin C. Leonard, Jr.  
Indiana University

William Lesch  
University of North Dakota

Aaron Liberman  
University of Central Florida

John E. Logan  
University of South Carolina

Dianne B. Love  
Houston – Clear Lake

F. Victor Lu  
St. John's University

W. Glynn Mangold  
Murray

J. Kenneth Matejka  
Duquesne University

Mary Maury  
St. John's University

Donald C. McCrory  
Memphis – Port Commission

Foster Morrison  
Turtle Hollow Association

Jay Nathan  
St. John's University

Maria Nathan  
Lynchburg College in Virginia

Robert Paul  
Kansas State University

George Peek  
Western Illinois University

Lucia Peek  
Western Illinois University

Robin T. Peterson  
New Mexico State University

James Poindexter  
Duquesne University

Simca Pollack  
St. John's University

Russ Ray  
University of Louisville

Denis Ridley  
Florida A & M

Lloyd C. Russow  
Philadelphia College

Anthony M. Sabino  
Lawyer – St. John's University

Charles S. Sherwood  
California State University –  
Fresno

Ronald R. Sims  
College of William and Mary

Lloyd Soobrian  
AT&T

M. Richard Sussman  
Central Michigan

John Thanopoulos  
University of Piraeus, Greece

John E. Triantis  
The Gillette Company

Nancy Upton  
Baylor University

Farok Vakil  
St. John's University

Philip M. Van Auken  
Baylor University

Iris Varner  
Illinois State University

Robert Whitis  
Arkansas State University

Matthew Wong  
St. John's University





Review of Business  
The Peter J. Tobin College of Business  
8000 Utopia Parkway  
Queens, NY 11439  
[www.stjohns.edu](http://www.stjohns.edu)

Non-Profit Org.  
U.S. Postage  
PAID  
St. John's University  
New York

- Remove from list.
- Change as shown.  
Please detach  
address label and  
mail to address  
shown above