International Journal of Statistical Distributions and Applications
2017; 3(3): 32-37
http://www.sciencepublishinggroup.com/j/ijsda
doi: 10.11648/j.ijsd.20170303.12
ISSN: 2472-3487 (Print); ISSN: 2472-3509 (Online)

Case Studies of Statistical Analysis in Engineering

Osama Ahmed Marzouk[1, *], Ahmad Izzat Jamrah[2]

[1]Department of Mechanical Engineering, University of Buraimi, Al Buraimi, Sultanate of Oman

[2]College of Engineering (Dean), University of Buraimi, Al Buraimi, Sultanate of Oman

Email address:

osama.m@uob.edu.om (O. A. Marzouk), ahmad.j@uob.edu.om (A. I. Jamrah)

[*]Corresponding author

To cite this article:

Osama Ahmed Marzouk, Ahmad Izzat Jamrah. Case Studies of Statistical Analysis in Engineering. *International Journal of Statistical Distributions and Applications*. Vol. 3, No. 3, 2017, pp. 32-37. doi: 10.11648/j.ijsd.20170303.12

Received: February 18, 2017; **Accepted**: February 27, 2017; **Published**: October 31, 2017

Abstract: Statistical analysis finds a wide range of applications in scientific research, management, finance, and engineering. In this article, we aim at shedding some light on the significance of two tools of statistical analysis, namely ANOVA and regression, in civil (including environmental), mechanical, and architectural engineering disciplines through a short survey of recent research studies that utilized these tools to reach important engineering models or rules of interest to engineers for design or development. The studies represent diverse areas of traffic planning, residential energy, indoor air quality, and high-speed machining. We briefly present some sample selection rules and analysis outcomes in these studies.

Keywords: ANOVA, Regression, Engineering, Design, Modeling

1. Introduction

While trying to solve problems or provide an explanation for an observed phenomenon, engineers may find themselves overwhelmed by large amounts of data from sampling real or virtual environments. Statistical analysis allows systematic processing of these data to generate reduced-order analytical models relating variables of interest, as well as making inference about the impact of one or two categorical (discrete) variables. This enables simple interpretation of the data and thus making predictions, which in turn lead to improved processes or products. Statistical analysis tools include regression, ANOVA (analysis of variance), time series analysis, factor analysis, correlation, and response surface methodology (RSM) [1]. In this work, we aim at showing successful examples from recent research studies that utilized regression and/or ANOVA to find useful models or facts for civil, mechanical, or architectural engineers. These findings allow efficient and innovative design decisions by practitioners in the field.

In statistical contexts, regression refers to any procedure that tries to estimate the relationship between one or more independent variables (predictors) and a continuous dependent variable (response) through fitting a model to collected data. The objective of regression is either describing analytically the dependence among the variables, or forecasting values beyond the observed collection. Common regression models are the linear, polynomial, and logistic. Simple regression refers to the case where there is a single predictor, whereas multiple regression refers to the case where there are two or more. The predictors are typically continuous [2], but categorical ones are allowed, for example in a 1/0 binary form. The developed linear regression models are then judged quantitatively for their goodness of fit through statistics like the correlation coefficient (r), coefficient of determination (R-squared or R^2) for simple regression, and its adjusted version (adjusted R-squared or R^2_{adj}) for multiple regression. Although a high R-squared (or adjusted R-squared) approaching unity supports a strong relationship between the linear model and the dependent variable, it does not formally establish hypothesis testing for the relationship. Such hypothesis testing is conducted using the t-test for the significance and p-value of a single independent variable (single regression coefficient) at a time, or the F-test for the significance and p-value of the overall model, considering all independent variables (all regression coefficients) simultaneously. A low p-value (e.g., below 0.05) or a high F-statistic number (from tables or statistical software) supports rejecting the *null hypothesis* that the fit of the tested model is equal to the 'dummy' intercept-only model.

ANOVA indirectly compares the means of three or more samples through processing sets of variances, after the samples were subject to different treatments (explanatory variables or categorical factors). Estimates of treatment-based variation (between the samples) and chance-based variation (within the samples) are found separately and compared using an F-test, based on which a conclusion can be drawn about the hypothesized influence of the treatment, whether or not intentionally applied. ANOVA also helps deciding whether or not subsamples should be statistically analyzed differently (e.g., by applying a separate regression model to each subsample). A one-way (also called one-factor) ANOVA refers to the case where a single treatment is applied at any time. The t-test is viewed as a special case of the one-way ANOVA, as it is limited to comparing means of two groups (e.g., gender differences). Applying ANOVA to two samples yields same results as the t-test (with the relation between statistics: $F = t^2$). Multi-factor ANOVA refers to the case where two or more treatments (or different levels of the treatment) are present and the treatments are crossed. This procedure not only allows inference about the effects of each treatment, but also the interactions among them. It suits factorial experiments with fixed effects, because it is a relatively-simple type of experiments. If the treatments are not only crossed but also nested, or they are of combined categorical and quantitative types, or one of them is random; multi-factor ANOVA becomes inappropriate. In such situations with elevated complication, the General Linear Model (GLM) procedure should be used instead. It is an ANOVA procedure performing least-squares regression.

2. First Study: Models for Total Bus Stop Time

We start with the study of Arhin et al. [3], carried out in the busy urban area of Washington DC, the capital of the USA. They considered the total bus stop time (TBST), in seconds, defined as the sum of dwell time (time elapsed between opening and closing the doors for passengers) and the time a bus takes to effectively park at a bus stop and then to re-enter the traffic stream. The study used multiple linear regression (through the ordinary least squares method) to better understand the factors that would affect the TBST. The sampled data were collected on weekdays in 2014 at 60 bus stops that have high patronage during 3 selected peak periods: morning (7 a.m. to 10 a.m.), mid-day (12 p.m. to 2:30 p.m.) and evening (4 p.m. to 6 p.m.). The field data were collected manually by the participating team and recorded in data collection sheets.

These data include:

a) Bus stop ID number
b) Bus route number
c) *S1*: time the bus arrived to the bus pad
d) *X*: number of passengers boarding
e) *Y*: number of passengers alighting
f) *D1*: time door opens
g) *D2*: time door closes
h) *S2*: time bus pulls away from the bus pad after the doors closed
i) Presence of street parking adjacent to the bus stop
j) Number of lanes at the approach where the bus stop is located
k) Length of bus pad in inches

Although Washington Metropolitan Area Transit Authority (WMATA) had already installed on some buses automatic passenger counting (APC) and automated vehicle location (AVL) systems under a test phase, a quick comparison between the APC/AVL data and field data showed unexplained differences and thus the manual collection was adopted. For the regression model, the predictors were the dwell time in seconds (Dt), the number of passengers boarding (Pb), the presence of street parking (Pk), the number of approach lanes (Ln), the bus pad length in inches (Bp), and the number of passengers alighting (Pa). Thus, the model has the form

$$\text{TBST} = D_t\, k_1 + P_b\, k_2 + P_k\, k_3 + L_n\, k_4 + B_p\, k_5 + P_a\, k_6 + \varepsilon \quad (1)$$

Six regression models were developed because the data were split into a 2×3 array, with 3 choices for the time of the day (morning, mid-day, or evening) and 2 choices for the bus stops location (at intersections, or mid-block). The statistical significance was verified by the ANOVA tests (p-value < 0.05). The study used Microsoft Excel and Minitab [4] software tools. The proposed models can help civil engineers improve bus scheduling in similar urban areas.

3. Second Study: Downsizing Homes for Reduced Energy Consumption

The next study [5] we selected was conducted in the UK. With more and more emphasis on curbing energy consumption and the consequent greenhouse gas (GHG) emissions in modern societies, this study used statistical analysis to support housing downsizing (moving to a smaller dwelling) as a proposed method for reducing annual residential energy consumption (in MW-h). Downsizing of equipment (e.g., heaters and air conditioning units [6] and even automobile engines [7]) to save energy has already received earlier attention, whereas engineers should realize more the energy saving due to smaller living spaces and avoiding under-occupancy. The total sample size was N = 991 households which represent part of the 2011 Energy Follow-Up Survey (EFUS) by the British Department of Energy and Climate Change, and the 2011/2012 English Housing Survey (EHS). The EHS has data about the English housing stock as well as socio-demographics of the householders. On the other hand, the EFUS has data completed by the householders about their dwelling and

heating practices. The annual energy consumption was estimated from electricity and gas meter readings.

The EFUS had 2616 households, but trimming occurred by eliminating the following classes:

a) Dwellings with insufficient meter readings
b) Dwellings with an indicated change to them since the last EHS without sufficient description of the change and its time (thus, its influence on the energy consumption remains questionable)
c) Dwellings with an indicated change to the household structure since the last EHS without sufficient description of the change and its time (thus, its influence on the energy consumption remains questionable)
d) Outliers in terms of energy consumption (defined here as those with ±3 standard deviations from the sample mean)
e) Households using fuels other than gas or electricity for heating (to avoid having underrepresented subsamples)

Linear regression analysis tested the impact of different predictors on (log-transformed) annual residential energy consumption. These predictors included the household size and the floor area. Logistic regression was also used to predict householders who under-occupy their homes, with under-occupancy coded as '1' and non-under-occupancy coded as '0'. The predictors included the residency period in years, presence of a sick or disabled person in the household, employment status, age of the household reference person, presence of dependent children, ethnic origin, and tenure (e.g., rented or owned). ANOVA was used to show the impact of the numbers of bedrooms (1, 2, 3, or 4), which has nearly a linear relation with energy consumption as shown in Figure 1.

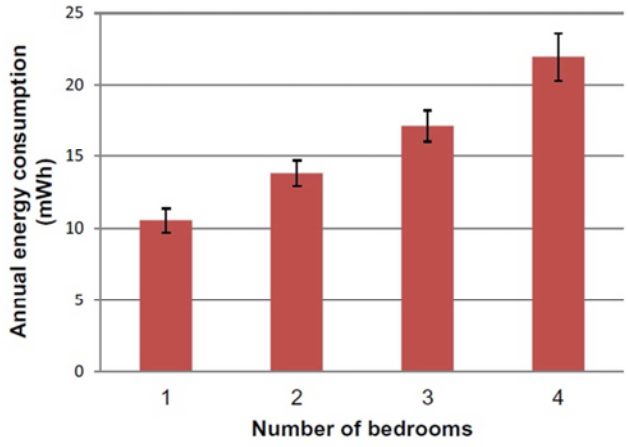

Figure 1. *Mean and standard error of the energy consumption (in megawatt-hour) per annum, for 1-, 2-, 3-, and 4-bedroom single-person household [5].*

The statistical analysis (we could not identify the software package used) provided strong support for a large potential for energy saving at the national level by promoting downsizing. The study is of special value to architectural engineers and building designers, as well as mechanical engineers who are dealing with residential heating and air conditioning. The study is also targeting the public at large, primarily those considering relocation to a new house and the elderly who continue to live alone in big family houses after their children have moved out.

4. Third Study: Influence of Construction Materials on Indoor Air Quality

Moving from residential energy consumption to residential air quality, the third study [8] considered here targets civil engineers and uses statistical analysis to quantify the relation between some construction materials and indoor ambience for office buildings. This subject is becoming important with the ongoing shift to urban societies where people spend more and more time in an indoor ambience. In 2014, 54% of the world's population lived in urban areas but this is expected to increase to 66% by 2050, and most of this shift seems to take place in Asia and Africa [9]. The National Human Activity Pattern Survey (NHAPS) in the USA conducted a survey where the respondents reported spending 87% of their time in enclosed buildings and other 6% in enclosed vehicles [10]. Thus, good indoor quality is vital for increased productivity as well as comfort.

The study was conducted for 15 office buildings in Yaoundé (the political capital of Cameroon in Central Africa region, see Figure 2).

Figure 2. *Political map of Cameroon and its neighborhood, showing the capital Yaoundé (source: http://motherearthtravel.com).*

Yaoundé is characterized by a sub-equatorial climate, and annually has two rainy seasons and 2 dry seasons. It is located about 300 km from Atlantic coast. The climate is overall moderate because this city is built on hills and has an average altitude of 750 m. Thus, the temperature is within

15 °C and 35 °C year-round. The dry seasons are hotter than the rainy ones. The relative humidity, however, is quite high and ranges between 55% and 75%. The wind is calm and the extreme wind speed is only 1.7 km/h.

The buildings are divided into groups based on the construction materials used in the walls. For the bricks, two categories exist

a) Compressed earth bricks: made from the local soil and clay, compressed by a press. This material was commonly used in old buildings (more than 25 years old).
b) Modern parent: imported masonry bricks

For the external coating, three categories exit

a) Plaster
b) Marble
c) Paint

The total sample consists of 15 office buildings, having same orientation and wall structure (0.6-cm external coating followed by a 1.1-cm external concrete layer, followed by 10-cm-thick bricks, followed by a 0.5-cm internal concrete layer, and finally a 1-cm internal coating of marble). Table 1 shows the number of buildings in each subsample divided based on the construction material.

Table 1. *Subsamples based on the construction material.*

	Bricks Type		**External Coating Type**		
	Earthen	**Imported**	**Plaster**	**Marble**	**Paint**
Size	6	9	8	2	3

During the study, the buildings were naturally ventilated, and any heating or cooling systems were deactivated. The buildings have large curtain-covered windows occupying more than half of the wall area, thereby blocking sunbeams. The occupants in each building are about 9 or more employees with sedentary activities (typical office work). The data used in the study came from two sources. The first is direct measurement of the indoor air temperature, relative humidity, and CO_2 concentration; taken as close as possible to the employees; and sampled at a frequency of 10-20 min over the working hours of 8am-5pm in each season. The second source is 218 questionnaires collected from the offices employees, with inquiries about thermal sensation, thermal preference, and acceptance of the thermal indoor environment, along with other items such as the gender, age, weight, and clothing. The study used the statistical software SPSS [11].

The study used the multiple linear regression to fit a model for the acceptability Acc (expressed as a percentage) as a comfort index. The predictors were the indoor air temperature T in °C and partial vapor pressure P_v (partial pressure of the water vapor laden by the moist air) in N/m^2. The model is

$$\text{Acc}\% = -106 + 4.6\,(30 - T) + 3.8\,(42.5 - P_v) + \varepsilon \quad (2)$$

The study then used one-way ANOVA to test the statistical significance of the construction materials when it comes to interpreting recorded differences in the indoor ambiences.

Although buildings with modern bricks showed higher mean in the temperature by 2.0 °C (25.3 °C compared to 23.3 °C) than those with earthen brick, this difference could be statistically differentiated only under a significance level of 0.1 (whereas the norm is 0.05). On the other, a difference in the mean relative humidity of 6% (73% for earthen bricks compared to 67% for modern bricks) was identified under a significance level of 0.05. Thus, it was decided that relative humidity is the better variable to use for recognizing each subsample. The regression model in (2) was thus shortened to

$$\text{Acc}\% = -106 + 3.8\,(42.5 - P_v) + \varepsilon \quad (3)$$

It was also found that an external coating of marble (which is more-impermeable) gives a higher indoor relative humidity than either the plaster or paint coating. Furthermore, no significant difference was identified between paint and plaster.

These findings can interpret well the observed low perceived indoor air quality index (PAQ) for buildings with earthen bricks or external marble coating were high in the dry seasons, where the elevated indoor relative humidity increases the comfort level, while PAQ was poor in the rainy seasons for these buildings because elevating the relative humidity makes it excessive.

5. Fourth Study: Parameter Identification for High-Speed Machining

We selected here a study [12] from the manufacturing arena, which is a main subfield of mechanical engineering. Statistical analysis was crucial to translate a large amount of data into key surface-roughness parameters of a machined product during die casting production, and then to determine which technological parameters influence them the most.

Die casting or die-cast manufacturing is a mass-production forming process where a metallic product is formed by injecting the molten metal under high pressure into a two-piece mold cavity made of hardened tool steel. The process is particularly used for aluminum, zinc, copper, and lead. Common products made by this process include propellers, gears, bushings, and automobile components (valves, pistons, cylinder heads, and engine blocks). In die casting, the produced castings may need a secondary process of machining to improve the surface finish. Unlike casting, parts (chips) of the product are removed during this machining. High-speed machining, HSM (also called high-speed cutting, HSC) is a relatively-modern concept where the attainable cutting speeds are significantly higher than those traditionally utilized for a particular material. The temperature of the surface at the interface with the cutting tool rises and approaches the melting point. Advantages of HSM include high material removal rate (thus shorter machining times), lower cutting force, improved surface quality, and increased dimensional accuracy [13]. However, HSM is not suitable if the machined surface has steep curvatures because the cutting

tool would have to progress slowly. Also, the tool life is shortened. Figure 3 distinguishes the cutting speeds of traditional and high-speed machining for different materials.

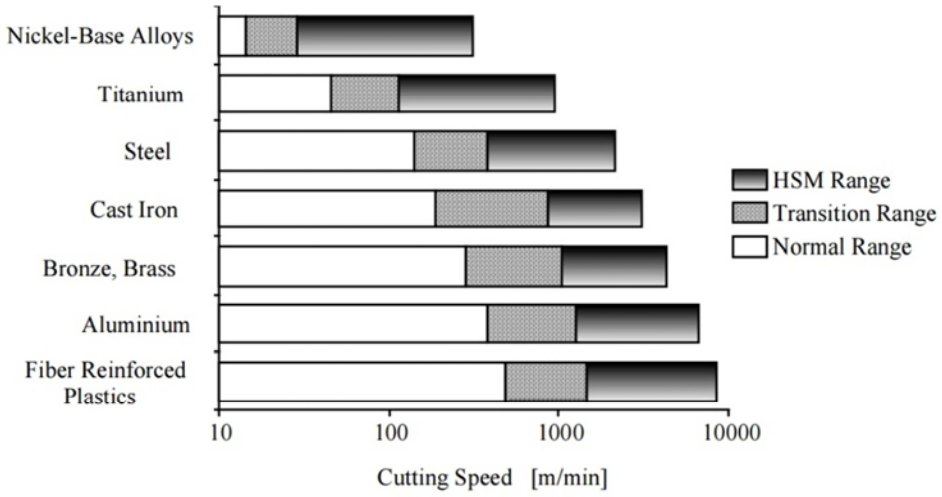

Figure 3. *Attainable cutting speeds for normal machining and high-speed machining (source: Schulz and Moriwaki, 1992 [14]).*

In the study, the surface roughness is measured by measuring the texture (z-coordinate at a 2D array of x and y coordinates) as shown in Figure 4.

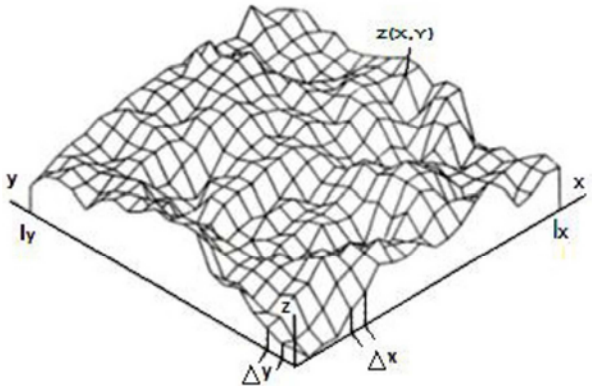

Figure 4. *Illustration of the 3D surface roughness measurements [12].*

Various roughness parameters are generated through processing these texture data. Each roughness parameter is a numerical value that gives an integral quantitative estimate of the level of the surface roughness. Examples include the arithmetic mean surface height (Sa) and the texture aspect ratio (Str). Calculating these roughness parameters from the 3D surface roughness measurements follows a defined standard, namely ISO 25178:2012. Aside from these roughness parameters, there are also technological parameters which are categorical variables for classifying the HSM operation, such as type of the machined material and the cutting depth.

The study aimed to identify

a) The most-important 3D surface roughness parameters
b) The most-critical technological parameters affecting each roughness parameter

To this end, samples of HSM surfaces were produced using a high-speed milling machine having a maximum spindle speed of 16,000 RPM (revolutions per minute), a maximum spindle power of 26 kW, and a maximum working feed rate of 20 m/min. Across all samples, the cutting depth was fixed at 0.3 mm. Other varied technological parameters are summarized in Table 2.

A total of 48 subsamples were arranged into groups appropriately.

Table 2. *Varied technological parameters.*

Technological parameter	Overlap (mm)	Feed rate (mm/min)	Cutting strategy	Material type	Milling mode
	0.05	2513	Linear path	Steel 1.2312	Up
Possible values		6283	Circular path	Steel 1.1730	
	0.1	1256	Two linear paths	Unalloyed titanium	Down

To achieve the first aim (identifying the most-significant roughness parameters), the study used the statistical analysis, and a correlation matrix was prepared using the Rcommander software [15] where a set roughness parameters were inserted into the matrix of n random variables, to obtain the most-significant roughness parameters for each group. The analysis identified 5 key roughness parameters, which are

a) Arithmetic mean surface height (Sa)
b) Kurtosis of the surface (Sku)
c) Height of the bearing area ratio (Stp)
d) Texture aspect ratio (Str)
e) Valley fluid retention index (Svi)

To achieve the second aim (identifying the most-critical technological parameters affecting each key roughness parameter), the study again resorted to the statistical analysis through the Rcommander software. The technological

parameters were replaced by factors, and the ANOVA multi-factor analysis was conducted on pairs of these factors, using each key roughness parameter as a response function, to see how each factor affects one or more key roughness parameters. The summary of results from this analysis is summarized in Table 3.

Table 3. *Influence of technological parameters on the roughness parameters.*

Key roughness parameter	Mostly affected by	ANOVA significance ratio
Sa	Cutting strategy	0.002253
Sku	Material type	0.01678
Stp	Material type	0.03999
Str	Feed rate	0.01621
Svi	Cutting strategy	0.04633

6. Conclusions

We selected 4 recent research studies in the architectural, civil, and mechanical engineering disciplines and used them to demonstrate how important regression and ANOVA are for engineers nowadays in order to allow interpreting data and extracting useful models and facts for later use in design and process development. The studies spanned topics in traffic planning, reduced residential energy consumption, indoor air quality, and high-speed machining for die casting. Regression and ANOVA in these studies enabled systematic development of analytical model, testing the statistical significance of one or more hypothesized influence of an independent variable, and having a quantitative measure of the relational dependence among variables. This work shows also that statistical software is an important skill for practical analysis that goes beyond basic classroom examples.

References

[1] S. Kowalski and G. G. Vining, Statistical Methods for Engineers, 3rd edition, Cengage Learning, 2010.

[2] D. T. Larose, Data Mining Methods and Models, John Wiley & Sons, 2006.

[3] S. Arhin, E. Noel, M. F. Anderson, L. Williams, A. Ribisso and R. Stinson, "Optimization of transit total bus stop time models," *Journal of Traffic and Transportation Engineering,* vol. 3, no. 2, p. 146-153, 2016.

[4] Minitab, Inc., "Minitab [Statistical Software]," [Online]. Available: www.minitab.com.

[5] G. M. Huebner and D. Shipworth, "All about size? – The potential of downsizing in reducing energy demand," *Applied Energy,* vol. 186, part 2, p. 226–233, 2017.

[6] A. Trianni, E. Cagno and A. De Donatis, "A framework to characterize energy efficiency measures," *Applied Energy,* vol. 118, p. 207–220, 2014.

[7] R. A. Simmons, G. M. Shaver, W. E. Tyner and S. V. Garimella, "A benefit-cost assessment of new vehicle technologies and fuel economy in the U.S. market," *Applied Energy,* vol. 157, p. 940–952, 2015.

[8] M. K. Nematchoua and J. A. Orosa, "Building construction materials effect in tropical wetand cold climates: A case study of office buildings in Cameroon," *Case Studies in Thermal Engineering,* vol. 7, p. 55–65, 2016.

[9] Department of Economic and Social Affairs - Population Division, "World Urbanization Prospects - The 2014 Revision," United Nations, New York, 2015.

[10] The National Human Activity Pattern Survey (NHAPS), "A Resource for Assessing Exposure to Environmental Pollutants," 2000.

[11] IBM Analytics, "IBM SPSS [Statistical Software]," IBM (International Business Machines Corp.), [Online]. Available: www.ibm.com/analytics/us/en/technology/spss/.

[12] A. Logins and T. Torims, "The influence of high-speed milling strategies on 3D surface roughness parameters," *Procedia Engineering,* vol. 100, p. 1253 – 1261, 2015.

[13] H. A. El-Hofy, Fundamentals of Machining Processes: Conventional and Nonconventional Processes, USA: CRC Press, 2006.

[14] H. Schulz and T. Moriwaki, "High-speed machining," *CIRP Annals - Manufacturing Technology,* vol. 41, no. 2, p. 637-643, 1992.

[15] Rcommander.com, "Rcommander [Graphical Interface for R]," Rcommander.com, [Online]. Available: www.rcommander.com.