Journal of Advances in Mathematics and Computer Science

34(6): 1-18, 2019; Article no.JAMCS.48814
ISSN: 2456-9968
(Past name: British Journal of Mathematics & Computer Science, Past ISSN: 2231-0851)

Modelling Factors Affecting Lung Capacity

Wanyonyi Maurice[1*]

[1]*Department of Mathematics, Computing and Information Technology, University of Embu, Kenya.*

Author's contribution

The sole author designed, analysed, interpreted and prepared the manuscript.

Article Information

DOI: 10.9734/JAMCS/2019/v34i630229
Editor(s):
(1) Dr. Dariusz Jacek Jakóbczak, Assistant Professor, Chair of Computer Science and Management in this Department, Technical University of Koszalin, Poland.
Reviewers:
(1) Abdullah Sonmezoglu, Bozok University, Turkey.
(2) Anthony Spiteri Staines, University of Malta, Malta.
Complete Peer review History: http://www.sdiarticle4.com/review-history/48814

Original Research Article

Received: 25 February 2019
Accepted: 30 April 2019
Published: 16 January 2020

Abstract

This study aims to evaluate modelling factors affecting lung capacity using linear Regression model. The study employed multiple regression models which were used to fit the factors affecting lung capacity. The factors affect lung capacity includes the following; age, gender, smoking and height. The objectives of the study were; fitting regression model on factors affecting lung capacity, determining the relationship between age and height with lung capacity. The study aim also includes predicting the value of lung capacity using the fitted model.

The data used in this study was a secondary source which was obtained from Marin [1]. The dataset is publicly available on their website. The data had 725 observations. Since multiple linear regression model was employed in this study, the model was of the form;

$$Lung\ capacity = \beta_0 + \beta_1 Age + \beta_2 Height + \varepsilon_i$$

Where;

Lung capacity is the dependent variable, $\beta_0, \beta_1 and \beta_2$ are the coefficients (parameters) to be estimated, Age and Height are the independent variables while ε_i is the random error component. The methods of parameter estimation discussed under this study include; maximum likelihood estimator and the least square estimator.

The data for this study were analyzed using SPSS and R software which are statistical software used for data analysis. From the analysis of variance table, a p-value of 0.00 was recorded which is less than alpha (alpha= 0.05). This implies that the overall model is significant.

From the model formulated, it was concluded that height and age greatly affect lung capacity. The model

**Corresponding author: E-mail: mauricewanyonyi5@gmail.com;*

formulated can be used to predict the value of lung capacity provided the values of Age and Height are known. Also from the descriptive statistics, it is deduced that gender and smoking greatly affect lung capacity.

Keywords: Model; lung capacity; multiple regression; maximum likelihood estimator; least square estimator.

Abbreviations

SPSS : Statistical Package for Social Sciences
ANOVA : Analysis of Variance
MLE : Maximum Likelihood Estimator
F : Fishers statistics
LSE : Least Square Estimator
r : Pearson correlation coefficient.
δ^2 *: Population variance.*
Pdf : Probability Density Function.
BIC : Bayesian Information Criterion.
AIC : Akaike Information Criterion.

1 Introduction

1.1 Background of the study

The study aims at modelling factors affecting lung capacity using the regression model. This is done by fitting a linear regression model on the factors affecting lung capacity. Some of the factors that affect lung capacity include; Age, Height, Smoking and Gender.

The regression model is one of the most widely used statistical technique that is used for investigating and modelling the relationship between variables. There are two types of linear regression model, that is simple linear regression and multiple linear regression model.

i. Simple linear regression model is a model that contains only on the independent variable. Here the model is linear in parameters β_0 β_1 and The model can be expressed as

$y = \beta_0 + \beta_1 x_1 + \varepsilon_i$ Where the intercept β_0 and the slope β_1 are the unknown constants which can be estimated by the method of least squares that is;

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{S_{XY}}{S_{XX}}$$

And ε_i is the random error component.

ii. Multiple linear regression model-This is the model where the dependent variable is related to more than one independent variables. The model can be expressed as $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon_i$.

In a multiple regression model, the parameter β is estimated by the method of least square estimator and the maximum likelihood estimators. Factors affecting lung capacity are fitted using a multiple linear regression model. Which is defined as a multivariate technique for determining the correlation between a response variable and some combination of two or more predictor variables?

Multiple regression analysis is useful in building econometric models. Which is a socioeconomic phenomenon that is influenced by the combined action of several factors?

The purpose of the econometric analysis is to estimate and predict the average value of the variable y based on the known or fixed values of the explanatory variables. Multiple regression analysis allows estimating the parameters of the econometric model, analyzing correlations between variables and testing the significance of the explanatory variables. In this study, a multiple linear regression model was used.

Multiple linear regression model has the following assumptions;

a) The random error term ε_i has an expected value of zero and a constant. Variance σ^2. That is, $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \delta^2$ for each recorded value of the dependent variable *Y*
b) The error components are uncorrelated with one another.
c) The regression coefficients $\beta_0, \beta_1, ..., \beta_k$ are parameters (and hence constant).
d) The independent (predictor) variables $X_1, X_2, ..., X_K$ are known.
e) The random error term ε_i is a normally distributed random variable, with an expected value of zero and a constant variance σ^2, by assumption (a). That is, $\varepsilon_i \sim N(0, \delta^2)$ under this additional assumption, the error components are not only uncorrelated with one another, but also necessarily independent.

Since the study is based on multiple linear regression model, the dependent variable is the lung capacity and the independent/explanatory variables are height and age.

Thus, the need to come up with a linear model to fit these factors and use the model to determine the relationship between age and height with lung capacity. The model will also be used to predict the value of lung capacity provided the value of age and height are known.

1.2 Objectives

i. To fit a regression model on factors affecting lung capacity.
ii. To determine the relationship between smoking, age, gender and height with lung capacity.
iii. To use the fitted model to predict the value of the dependent variable that is lung capacity.

1.3 Statement of the problem

There are several studies and researches which have been conducted by different researchers on different study topics using the regression model. Basing on my study topic; that is modelling factors affecting lung capacity, there are different studies which have been conducted on factors affecting lung capacity using different methods and approaches some of which includes the use of cross-sectional studies on the individuals under study in order to identify how factors like; age, height and gender affects lung capacity.

Since there are no studies which have been conducted on modelling factors affecting lung capacity using linear regression model, this motivated me to come up with a linear model using regression analysis to fit factors affecting lung capacity and consequently use the model to determine the relationship between age and height with lung capacity. The model will also be used to predict the value of the lung capacity.

1.4 Significance of the study

The significance of the study is to model factors affecting lung capacity using a linear regression model. Consequently, determines the relationship between smoking, age, gender and height with lung capacity. Also, to use the model to predict individuals lung capacity.

1.5 Research questions

i. Is there a relationship between age, gender, height and smoking with lung capacity?
ii. Is it possible to come up with a model that can be used to fit the factors affecting lung capacity?

1.6 Research hypothesis

There is no relationship between age and lung capacity.
There is no relationship between height and lung capacity.
Mean lung capacity of smokers = Mean lung capacity of non-smokers.

1.7 Justification of the study

There is a need to carry out this study considering the fact that there are no studies which have been done on modelling factors affecting lung capacity. Different researchers have used different methods and researches to determine factors affecting lung capacity. This study will try to fit a linear regression model on factors affecting lung capacity and also determines the relationship between Age, Smoking, Height and Age with Lung capacity.

I intend to use a linear regression model. This is because linear regression attempts to model the relationship between two or more variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.

This model will be used to determine the effect of age, gender, height and smoking on lung capacity. The model will also be used to determine the relationship between age, smoking, height and gender with lung capacity. Finally, the model will be used to predict the value of the dependent variable provided the independent variables are known.

2 Literature Review

A considerable amount of literature is available on modelling using linear regression model. Different researchers and writers have modelled using linear regression model. This chapter will give an insight into the various researches and work that have been done using a linear regression model in my topic of study or other different areas of studies. The listed below are some of the literature of the studies of recent past done using linear regression model.

Vlasta Bahovec [2], They researched on Regression analysis of individual financial performance. They used cross-section linear regression model to estimate how gender as a dummy variable and financial literacy as an ordinal /categorical variable impact financial performance. They also used ordinary least squares method to estimate the cross section linear regression model.

Plotts [3], He investigated A Multiple Regression Analysis of Factors Concerning Superintendent Longevity and Continuity Relative to Student Achievement. He used a multiple regression model to find out if the superintendent's tenure, longevity, and continuity have an impact on student academic achievement.

Shakil [4], He used A Multiple Linear Regression Model to Predict the Student's Final Grade in a Mathematics Class. According to him he defined multiple linear regression as multiple linear regression is defined as a multivariate technique for determining the correlation between a response variable *Y* and some combination of two or more predictor variables, *X*.

He based his study on multiple linear regression and the model he came up with was as the one below.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_K X_K + \varepsilon_i$$

Where;

Y = response variable, *X* =predictor variables, β *k* = the population regression coefficients, and ε_i is a random error.

The objective of the study was to develop an appropriate multiple linear regression model to relate the student's final examination score (considered as the dependent or response variable *Y*) to the student's scores in tests, quizzes, etc. (considered as the independent or predictor variables *X)*. It examined how well the scores in tests, quizzes, etc. could be used to predict the student's final grade.

Gibbs et al. [5], They Used regression model to establish the relationship between home environment and reading achievement. They came up with a model which they used to determine the relationship. They adopted a simple regression model of the form;

$$Y = a + bX$$

Where;

Y= dependent variable; X = independent variable; a = slope of line; b = intercept on y–axis and also;

$$a = \frac{\sum y - b\sum x}{n} \text{ And } b = \frac{\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$$

They used Students t test to assess variations between values of different samples at a level of significance of $p<0.05$ and $p<0.01$.

Sunthornjittanon [6], He used linear regression analysis to investigate net income of Agrochemical Company. The methods and models used, were created based on the company's available data and regression analysis methods, using the company's sales numbers and profits for the past nine years (2005-2013). The final model is selected using Stepwise Regression Methods. A linear regression line and equation for the model are generated to help observe and predict future trends.

Flavia et al. [7], They studied the economic performance of an organization using multiple regression model. The analysis they used includes; regression analysis, correlation analysis, ANOVA and time series analysis. They came up with the following model which they used in fitting their data.

$$y = f\left(x_1, x_2, ..., x_n\right) + \varepsilon_i$$

Where; y – dependent variable (output), randomized; $x_1, x_2, ..., x_n$ are the independent variables (factorial), non-randomized; ε_i – random variable error or residue. Random variable ε_i summarizes the influences of variables not included in the model that influences the variable y.

Catherine et al. [8], They used regression models to predict the length patients' stays in the hospitals. They proposed a prediction model based on logistic regression. The authors compared the predictive accuracy of regression trees with that of logistic regression models for predicting in-hospital mortality in patients hospitalized with heart failure. Their main conclusion is that Logistic regression predicted in-hospital mortality in patients hospitalized with heart failure is more accurate than that of regression trees.

Luigi et al. [9], They used regression analysis in marketing research. Their main objective was to illustrate the applicability of the linear multiple regression model within marketing research based on primary, quantitative data. The theoretical background of the developed regression model is the value-chain concept of relationship marketing. They generated a regression model which they used in their study.

Byerly [10], He used multiple regression and path analysis in analyzing the success in Journalism. Anita and Pooja [11], They used correlation and regression analysis in assessing lentic water quality. Study was conducted to determine the levels of water quality indicators and to study statistical interrelationships amongst them. An attempt has also been made to establish regression equations to provide a prediction of water quality prior to detailed investigation. Multiple linear regression model was used in this study.

Jelena et al. [12], They used a regression model to predict business result. The study was based on planning and prediction of business results in insurance when calculating premium trend by use of linear and nonlinear regression.

Having studied all the literature available on researches that have been done using a linear regression model, thus the need to come up with a model using multiple regression analysis to fit the factors affecting lung capacity. The model will also be used to determine the relationship between the dependent and the independent variable. The model can also be used to predict individual lung capacity provided the independent variables basing on the model are known.

3 Research Methodology

3.1 Introduction

This chapter deals with research method that was adopted and analysis of data collected. This study entailed descriptive survey design, population, sample design, data collection and analysis.

3.2 Research Design

In this study descriptive survey was used. A descriptive study was undertaken in order to ascertain and describe the characteristics of variables of interest under consideration.

3.3 Population

A population is an entire group of individuals or objects having common characteristics that conform to given specifications. Since the data used for this study was secondary data, the total population was not indicated from where the data was obtained. The data for this study was obtained from the website [1].

3.4 Sample

Sample is the collection of sampling units from a sampling frame whereas; sampling frame is the set of sampling units that constitute the set of the sample.

Sampling is the process of selecting a sufficient number of the right elements from the population (Groves, 2010). According to Marin [1], the sample number that was used for this study was 725 observations that is they sampled 725 individuals whom the required data was collected from. $(n = 725)$.

3.5 Data collection

Data collection is gathering empirical evidence in order to gain new insight about a situation and answer questions that prompt undertaking of the research (Flick, 2009). The data for this study was Secondary data which was obtained from Marin [1] which is publicly available from their website.

3.6 Data analysis

The data was analyzed using Statistical Program for Social Sciences (SPSS) and R. A linear regression model analysis was developed and used. Linear regression attempts to model the relationship between two variables by fitting a linear equation to the observed data. One variable is considered to be an independent variable, and the other is considered to be a dependent variable. Before attempting to fit a linear model to observed data, a modeller should first determine whether or not there is a relationship between the variables of interest. A scatterplot can be a helpful tool in determining the strength of the relationship between two variables. If there appears to be no association between the proposed explanatory and dependent variables (i.e., the scatterplot does not indicate an increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model. A valuable numerical measure of association between two variables is the correlation coefficient, which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables.

The model employed was based on multiple linear regression model:

Consider a dependent variable y that is linearly related to p independent variable $X_1, X_2, ..., X_P$ through $\beta_0, \beta_1, ..., \beta_P$ the parameters $\beta_0, \beta_1, ..., \beta_p$ are the regression coefficient. Assuming the multiple regression model is given by;

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_P X_P + \varepsilon_i$. This model can be summarized into;

$$Y = X\beta + \varepsilon_i$$

Where y is the dependent variable, β is the parameter to be determined whereas X is the independent variable whereas ε is the normally distributed error term.

Thus, the multiple linear regression model based on my study will be represented as: -

$$Lung\ \ capacity = \beta_0 + \beta_1 Age + \beta_2 Height + \varepsilon_i$$

Where;

Lung capacity is the dependent variable; Age and Height are the independent variable, $\beta_0, \beta_1 and \beta_2$ are the parameters to be estimated and ε_i is the normally distributed error term.

3.7 Methods of parameter estimation

There are two different methods that are used to estimate the parameter that is, the use least square method and the use of maximum likelihood estimator.

i. Least Square Estimator (LSE)

Consider a multiple linear regression model $y = x\beta + \varepsilon_i$. We find the parameter estimates of β by minimizing the sum of squares due to errors. Therefore, sum of squares is minimized by finding the partial derivative with respect to β and equating to zero.

$$S(\beta) = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \varepsilon_i' \varepsilon_i = (y - x\beta)'(y - x\beta)$$

When simplified it results to

$$S(\beta) = y'y + \beta' x' x'\beta - 2\beta' x' y$$

We then differentiate $S(\beta)$ with respect to β and equate to zero.

$$\frac{\partial S(\beta)}{\partial \beta} = 2x'x\beta - 2x'y = 0$$

And thus, the estimate of the parameter β is given by

$$\hat{\beta} = (x'x)^{-1}(x'y)$$

The properties of least square estimators include;

- Estimation Error- The estimation error for $\hat{\beta}$ is given as;

  $$\hat{\beta} - \beta = (x'x)^{-1} x'y - \beta$$

  But we have that $y = x\beta + \varepsilon_i$

  $$\hat{\beta} - \beta = (x'x)^{-1} x'(x\beta + \varepsilon_i) - \beta = (x'x)^{-1} x'x\beta + (x'x)^{-1} x'\varepsilon_i - \beta$$

  $$\hat{\beta} - \beta = \beta + (x'x)^{-1} x'\varepsilon_i - \beta = (x'x)^{-1} x'\varepsilon_i$$

- Bias- Since X is assumed to be non-stochastic and $E(\varepsilon_i) = 0$ then

  $$E(\hat{\beta} - \beta) = E\left[(x'x)^{-1} x'\varepsilon_i\right] = (x'x)^{-1} x' E(\varepsilon_i) = 0$$

  Thus, the least square estimator of $\hat{\beta}$ is unbiased estimator of β

- Covariance matrix- The covariance matrix of $\hat{\beta}$ is given by;

  $$\text{var}(\hat{\beta}) = E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)\right]$$

  $$= (x'x)^{-1} x' E\left(\varepsilon_i' \varepsilon_i\right) x (x'x)^{-1}$$

  But $E(\varepsilon_i \varepsilon_i) = \delta^2$

  $$\text{var}(\hat{\beta}) = \delta^2 (x'x^{-1}) x'x (x'x)^{-1} = \delta^2 (x'x)^{-1}$$

ii. Method of Maximum Likelihood Estimator (MLE)

In the model $y = x\beta + \varepsilon_i$ it is assumed that the errors are normally distributed with a constant variance δ^2. The pdf of errors is assumed to be;

$f(\varepsilon_i) = \frac{1}{\delta\sqrt{2\pi}} \exp\left[-\frac{1}{2\delta^2}(x-\mu)^2 \right]$ The likelihood function of $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ joint pdf is given by;

$$L(\beta, \delta^2) = \prod_{i=1}^{n} f(\varepsilon_i) = \frac{1}{(2\pi\delta^2)^{n/2}} \exp\left[-\frac{1}{2\delta^2} \sum_{i=1}^{n} \varepsilon_i^{\,2} \right]$$

$$= \frac{1}{(2\pi\delta^2)^{n/2}} \exp\left(-\frac{1}{2\delta^2} \sum_{i=1}^{n} \varepsilon_i' \varepsilon_i\right) \frac{1}{(2\pi\delta^2)^{n/2}} \exp\left[-\frac{1}{2\delta^2}(y-x\beta)'(y-x\beta) \right]$$

The log likelihood function is given by $LnL(\beta, \delta^2) = -\frac{n}{2} Ln(2\pi\delta^2) - \frac{1}{2\delta^2}(y-x\beta)'(y-x\beta)$

The MLE of β is obtained by first getting the first derivative of $LnL(\beta, \delta^2)$ with respect to β and equating to zero, that is;

$$\frac{\partial LnL(\beta, \delta^2)}{\partial \beta} = \frac{1}{2\delta^2} x'(y - x\beta) = 0$$

This implies that $x'(y - x\beta) = 0$ and thus $x'y - x'x\beta = 0$ this yield $x'x\beta = x'y$ making β the subject of the formula, yields;

$$\hat{\beta} = (x'x)^{-1} x'y$$

Correlation coefficients (r) is determined by the mathematical formula as given below. Let x and y be any two variables and n = number of observations. Then the correlation coefficient (r), between the two variables x and y is given by the relation.

$$r = \frac{n\sum(xy) - \sum x \sum y}{[f(x)f(y)]^{0.5}}$$

Where;

$$f(x) = n\sum(x^2) - \left(\sum x\right)^2 \text{ And } f(y) = n\sum(y^2) - \left(\sum y\right)^2$$

If the numerical value of the correlation coefficient between two variables x and y is fairly large, it implies that these two variables are highly correlated.

The following are methods of model selection that is used in regression analysis to determine the best model to use. They include;

a) Forward selection-This procedure begins with the assumption that there is no regressor in the model other than the intercept. The goal is to find an optimal subset by inserting regressors into the model one at a time.
b) Backward elimination-This procedure is the opposite approach from the forward selection. First, we begin with a full model with k-candidate regressors. Then partial F statistics are computed for each

regressor. If the regressor with the smallest partial F value is less than the preselected F value, that regressor is removed from the model.

c) Stepwise regression-It is a method that allows moves in either direction, dropping or adding variables at the various steps. It combines both the forward selection and backward elimination. We perform two steps in the forwarding selection and a backward step. Then perform another forward step and another backward step. We continue until no action can be taken in either direction.

d) AIC- is an estimator of the relative quality of a statistical model for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. The decision rule is that the smaller the AIC the better the model. AIC is given by;

$$AIC = 2k - 2\ln\left(\hat{L}\right)$$

Where;

k-is the number of parameters estimated by the model.

$\hat{L}$ -is the maximized value of the likelihood function.

e) BIC- This is a criterion for model selection among a finite set of models; the model with the lowest BIC is preferred.

$$BIC = \ln(n)k - 2\ln\left(\hat{L}\right)$$

Where;

k-is the number of parameters estimated by the model.

$\hat{L}$ -is the maximized value of the likelihood function.

n-is the number of data points.

The data was run in SPSS and R to come up with a model which was used to determine the relationship between variables and also to fit the factors that affect lung capacity.

Since the data was secondary, the researcher did not collect any invalid or unreliable data as such it was not necessary to conduct tests of validity and reliability.

4. Data Analysis and Interpretation

4.1 Introduction

This chapter presents data analysis and interpretation. The objective of the study was to fit a regression model on factors affecting lung capacity so as to determine the relationship between age, height, gender and smoking with lung capacity. Secondary data sources were used and they were obtained from Marin [1]. The dataset is publicly available on their website.

4.2 Descriptive statistics

Box plot is used to compare the distribution of numerical variable for different groups that are formed by a categorical variable. From the box plot below, it can be seen clearly that Males have a higher lung capacity compared to Females.

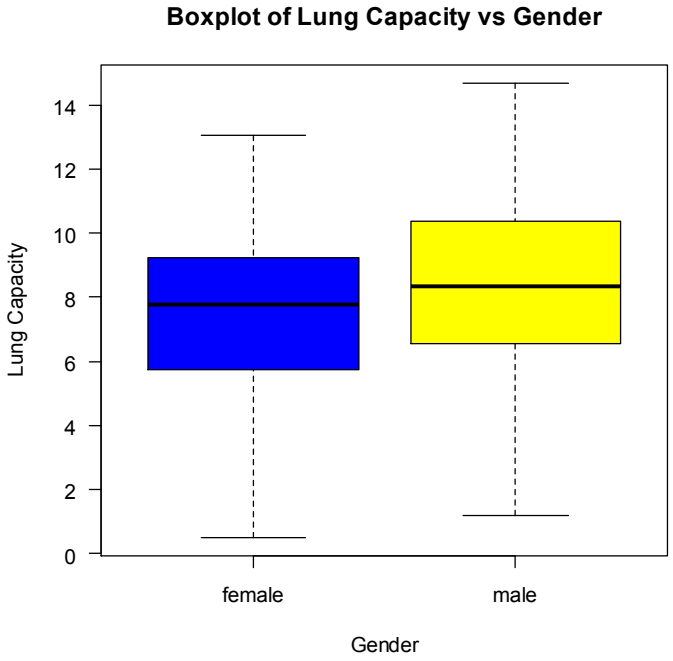

Fig. 1. Boxplot of lung capacity against gender

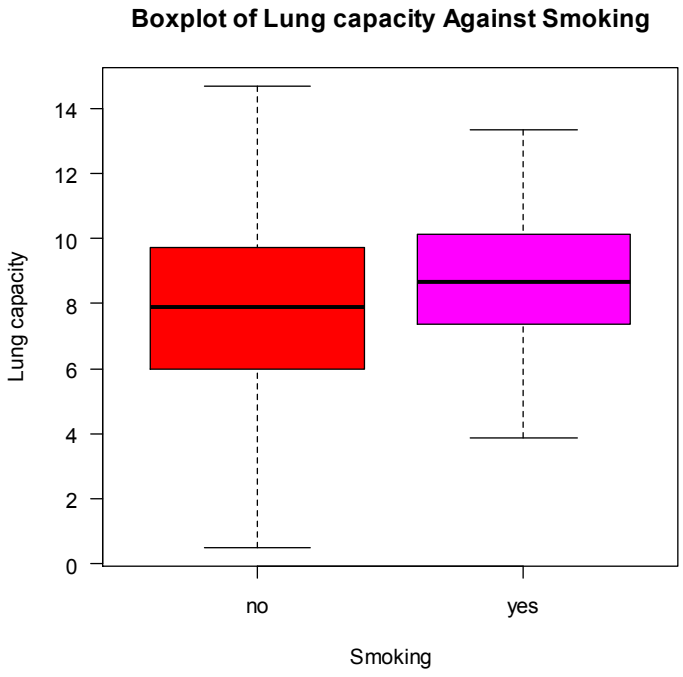

Fig. 2. Boxplot of lung capacity against smoking

From the boxplot above, it can be seen that individuals who are non-smokers have a higher lung capacity compared to those who are smokers.

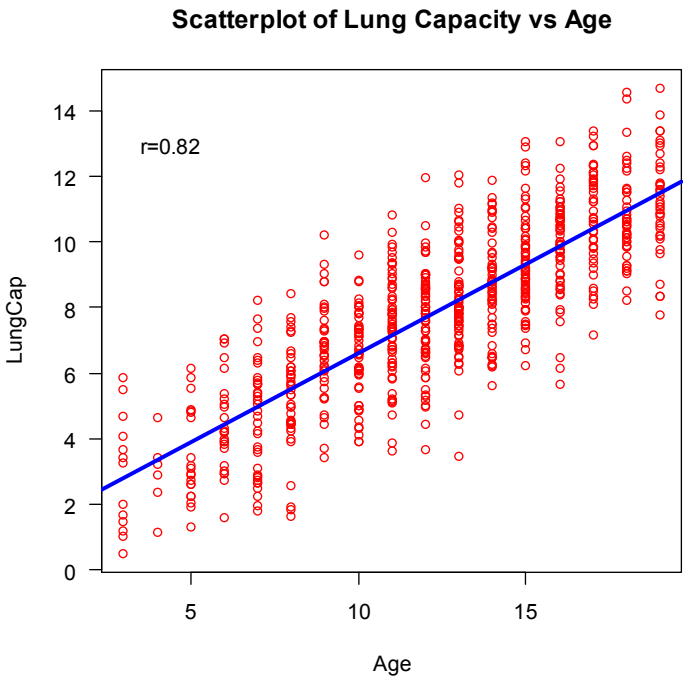

Fig. 3. Scatterplot of lung capacity vs age

Scatter plot is used to examine the relationship between two numerical variables. Thus, from the scatter plot above, we can observe that there is a linear relationship between lung capacity and Age. Lung capacity and Age have a strong positive correlation of 0.82, that is (r=0.82). Implying that an increase in age increases lung capacity and vice versa.

Table 1. Correlation matrix

Correlations

		LungCap	Height	Age
LungCap	Pearson Correlation	1	.912**	.820**
	Sig. (2-tailed)		.000	.000
	N	725	725	725
Height	Pearson Correlation	.912**	1	.836**
	Sig. (2-tailed)	.000		.000
	N	725	725	725
Age	Pearson Correlation	.820**	.836**	1
	Sig. (2-tailed)	.000	.000	
	N	725	725	725

**. Correlation is significant at the 0.01 level (2-tailed).

4.3 Correlation analysis

To quantify the strength of the relationship between variables, the study used Karl Pearson's coefficient of correlation. The Pearson product-moment correlation coefficient is a measure of the strength of a linear association between two variables and is denoted by (r). The Pearson correlation coefficient (r) ranges from -1 to +1. A value 0 indicates that there is no relationship between two variables. A value greater than 0

indicates a positive association whereas a value less than 0 indicates a negative association. Pearson's Correlation Coefficient was carried out and the results obtained are presented in Table 1.

From the correlation matrix above, we can observe a significant strong positive relationship between Lung capacity and Age $(r = 0.82)$ and Lung capacity and Height ($r = 0.912$) and their p-value which is 0.01 is less than 0.05, implying that both Age and Height affects Lung capacity.

4.4 Regression analysis

Regression Analysis is a statistical technique that identifies the relationship between two or more quantitative variables. A dependent variable whose value is to be predicted and independent (explanatory variables), about which knowledge is available. The technique is used to find the equation that represents the relationship between the variables. The relation between variables can be illustrated using an equation. The study adopted multiple linear Regression guided by the following model:

$$y = x\beta + \varepsilon_i$$

Where: -

$y = Lung\ capacity$
$x =$ The independent variable,

While β is the unknown parameter which determines the relationship between independent variable x and dependent variable y.

$\varepsilon_i =$ Normally distributed error term

Table 2. Model summary

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.918[a]	.843	.843	1.056287

a. Predictors: (Constant), Height, Age

Here, the coefficient of determination (percentage of variation in the dependent variable being explained by changes in the independent variables). R^2 equals 0.843, that is, Age and Height explain 84.3% of the variance in the Lung capacity.

Table 3. Analysis of variance

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4324.906	2	2162.453	1938.129	.000[b]
	Residual	805.566	722	1.116		
	Total	5130.472	724			

a. Dependent Variable: LungCap

b. Predictors: (Constant), Height, Age

Here, the significance value of the F statistics is 0.00 $(p-value=0.00)$ indicating that the predictor variable (Age and Height) explains a variation in Lung capacity and that the overall model is significant. This is because $p-value=0.00<0.05$.

4.4.1 Regression equation

Table 4. Coefficients

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-11.747	.477		-24.632	.000
	Age	.126	.018	.190	7.079	.000
	Height	.278	.010	.753	28.051	.000

a. Dependent Variable: LungCap

Based on regression coefficients results, the regression equation can be written as follows;

$$Lung\ \ capacity = -11.749 + 0.126Age + 0.278Height + \varepsilon_i$$

Regression analysis reveals the extent to which Age and Height significantly predict Lung capacity. The supremacy in prediction is determined by Beta coefficients of 0.126 and 0.278. The finding reveals that Lung capacity is greatly influenced by Age and Height. Thus, the individual lung capacity can be predicted provided Age and Height are known based on the regression model above.

4.5 Independent two sample t-test

Independent two-sample t-test is a parametric method that is used examining the difference in means for two populations. It is also used for examining the relationship between a numeric outcome variable (Y) and a categorical /explanatory variable (X, with two levels).

Suppose we want to test the following null hypothesis; mean lung capacity of smokers = mean lung capacity of non-smokers.

The following is the output of data analyzed using independent two-sample t-test.

Table 5. Independent two sample t-test

```
        Welch Two Sample t-test

data:  LungCap by Smoke
t = -3.6498, df = 117.72, p-value = 0.0003927
Alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.3501778 -0.4003548
Sample estimates:
 Mean in group no mean in group yes
        7.770188         8.645455
```

From the table above we can observe that the p-value which is 0.0003927 is less than alpha (0.05). This implies that we can reject the null hypothesis and conclude that there is a difference in means between lung capacity of smokers and lung capacity of non-smokers.

4.6 Summary and interpretation of findings

The study examines and model factors affecting lung capacity. From the findings, Males have a higher lung capacity compared to females as shown in the boxplot of lung capacity and gender above. This implies that gender affects lung capacity. Also, from the boxplot of lung capacity and smoking, it can be seen that individuals who are non-smokers have a higher lung capacity compared to smoking individuals thus it can be concluded that smoking affects lung capacity.

From the scatterplot above, it can be seen that there is a linear relationship between lung capacity and gender. Thus, there is a strong positive association between the two variables. That is there is an association between Lung capacity and Age.

From the correlation matrix above, it can be seen that there is a strong positive correlation between Lung capacity and Age $(r = 0.82)$ and there is also a strong positive correlation between Lung capacity and Height $(r = 0.912)$. This implies that there is a linear relationship between Age and Height with Lung capacity.

From the Analysis of variance table, the significance value of the F statistics is 0.00 $(p - val = 0.00)$ indicating that the predictor variable (Age and Height) explains a variation in Lung capacity and that the overall model is significant. This is because $p - val = 0.00 < 0.05$.

From the model summary above, the coefficient of determination (percentage of variation in the dependent variable being explained by changes in the independent variables). R^2 equals 0.843, that is, Age and Height explain 84.3% of the variance in the Lung Capacity.

The regression equation based on the model is given by: -

$$Lung\ \ capacity = -11.749 + 0.126Age + 0.278Height + \varepsilon_i$$

The regression analysis reveals the extent to which Age and Height significantly predict Lung capacity. The supremacy in prediction is determined by Beta coefficients of 0.126 and 0.278. Thus, one can predict lung capacity provided Age and Height are known using the regression equation above. Suppose that we want to predict the lung capacity of an individual whose Age is 19 years and the height is 78 centimeters. The value of the lung capacity is determined by;

$$Lung\ \ capacity = -11.749 + 0.126\,Age + 0.278\,Height$$

By replacing the value of age and height, we have the following;

$Lung\ \ capacity = -11.749 + 0.126(19) + 0.278(78)$ =12.329. Which is the predicted lung capacity. Thus, aged and taller individuals have a higher lung capacity compared to young and shorter individuals.

Finally, the findings from the results above reveals that Lung capacity is greatly affected by Age, Gender, Smoking and Height.

5. Summary, Conclusion and Recommendation

5.1 Summary

This chapter presents a summary of findings, conclusion and recommendation. The aim of the study was to model factors affecting lung capacity. The empirical literature on factors affecting lung capacity shows that factors like age, gender, height and smoking seriously affects lung capacity. But those studies did not incorporate the use of modelling like regression model rather they used cross-sectional studies. This forms the motivation of the study. It becomes important to model factors affecting lung capacity so as to be able to determine the relationship between age, height and gender with lung capacity.

From the findings, Males have a higher lung capacity compared to females as shown in the boxplot of lung capacity and gender above. This implies that gender affects lung capacity. Also, from the boxplot of lung capacity and smoking, it can be seen that individuals who are non-smokers have a higher lung capacity compared to smoking individuals thus it can be concluded that smoking affects lung capacity.

From the scatterplot above, it can be seen that there is a linear relationship between lung capacity and gender. Thus, there is a strong positive association between the two variables. That is there is an association between Lung capacity and Age.

From the correlation matrix above, it can be seen that there is a strong positive correlation between lung capacity and age $(r = 0.82)$ and there is also a strong positive correlation between lung capacity and height $(r = 0.912)$. This implies that there is a linear relationship between age and height with lung capacity.

From the Analysis of variance table, the significance value of the F statistics is 0.00 $(p - val = 0.00)$ indicating that the predictor variable (age and height) explains a variation in Lung capacity and that the overall model is significant. This is because $p - val = 0.00 < 0.05$.

From the model summary above, the coefficient of determination (percentage of variation in the dependent variable being explained by changes in the independent variables). R^2 equals 0.843, that is, age and height explain 84.3% of the variance in the lung capacity.

The regression equation based on the model is given by: -

$$Lung\ \ capacity = -11.749 + 0.126Age + 0.278Height + \varepsilon_i$$

The regression analysis reveals the extent to which age and height significantly predict lung capacity. The supremacy in prediction is determined by Beta coefficients of 0.126 and 0.278. Thus, one can predict lung capacity provided age and height are known using the regression equation above.

Finally, the findings from the results above reveal that lung capacity is greatly affected by age, gender, smoking and height.

5.2 Conclusion

The purpose of the study was to model factors affecting lung capacity. The study concluded that age, gender, smoking and height affect Lung capacity of individual as shown from the findings of the results above and also from the fitted model. From the research hypothesis, we had the following null hypothesis;

There is no relationship between age and lung capacity and there is no relationship between height and lung capacity.

Since the p-value obtained above from the analysis of variance table is 0.00 which is less than alpha 0.05 that is $(0.00 < 0.05)$, therefore we reject the null hypothesis and conclude that there is significant evidence that there is a relationship between age and lung capacity and there is also a relationship between height and lung capacity. This implies that age and height affects lung capacity.

5.3 Recommendation

From my findings, I recommend that smoking individuals should stop smoking because smoking lowers lung capacity. I also recommend that linear regression model be used when determining factors affecting lung capacity and also when predicting lung capacity.

Competing Interests

Author has declared that no competing interests exist.

References

[1] Marin M. MarinStatsLectures; 2015.
Available:https://www.statslectures.com/r-stats-videos-tutorials

[2] Vlasta Bahovec, Dajana Barbić, Irena Palić. The regression analysis of individual financial performance: Evidence from Croatia; 2017.

[3] Plotts T. A multiple regression analysis of factors concerning superintendent longevity and continuity relative to student; 2011.

[4] Shakil DM. A multiple linear regression model to predict the student's final grade in a mathematics class. Florida; 2006.

[5] Gibbs Y. Kanyongo, Janine Certo, Brown I. Launcelot. Using regression analysis to establish the relationship between home environment and reading achievement: A case of Zimbabwe. International Education Journal. 2006;632-641.

[6] Sunthornjittanon S. Linear regression analysis on net income of an Agrochemical Company in Thailand; 2015.

[7] Flavia Fechete, Anisor Nedelcu. Analysis of the economic performance of an organization using multiple regression model. International Conference of Scientific Paper; 2014.

[8] Catherine Combes, Farid Kadri, Sondès Chaabane. Predicting hospital length of stay using regression models; 2004.

[9] Luigi Dumirescu, Oana Stanciu (Duralia), Mihai Tichindelean, Simona Vinerean. The use of regression analysis in marketing research; 2012.

[10] Byerly RL. The use of multiple regression and path analysis in analyzing success in journalism at Iowa State; 1979.

[11] Anita Bhatnagar, Pooja Devi. Applications of correlation and regression analysis in assessing lentic water quality: A case study at Brahmsarovar Kurukshetra, India. International Journal of Environmental Sciences. 2012;3(2).

[12] Jelena Rusov, Mirjana Misita, Dragan D. Milanovic, Dragan Lj. Milanovic. Applying regression models to predict business results; 2015.

APPENDIX: R Commands

```
data<-read.csv (file.choose (), header=T, sep=",")
attach (data)
boxplot(LungCap~Gender,main="Boxplot of Lung Capacity vs Gender",las=1,col=c(4,7),xlab="Gender",ylab="Lung Capacity")
```

```
data<-read.csv(file.choose(),header=T,sep=",")
attach(data)
cor(LungCap,Age,method="pearson")
[1] 0.8196749
plot(LungCap~Age,main="Scatterplot of Lung Capacity vs Age",col=2,las=1)
abline(lm(LungCap~Age),lwd=3,col=4)
text(x=3.5,y=13,adj=0,label="r=0.82")
```

```
data<-read.csv(file.choose(),header=T,sep=",")
attach(data)
boxplot(LungCap~Smoke,main="Boxplot of Lung capacity Against Smoking",las=1,xlab="Smoking",ylab="Lung capacity",col=c(2,6))
```

```
data<-read.csv(file.choose(),header=T,sep=",")
attach(data)
t.test(LungCap~Smoke,mu0=0,Alt="two.sided",conf=0.95,var.eq=F,paired=F)
```


Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
http://www.sdiarticle4.com/review-history/48814