Experiences of Pearson formula in analysis regression

I Maryati*, T S Sumartini and D Sofyan

Mathematics Education Program of Institut Pendidikan Indonesia (IPI), Jl. Terusan Pahlawan 32, Garut Indonesia

*iyammaryati@institutpendidikan.ac.id

Abstract. Regression is a statistical analysis method used to see the effect between two or more variables. The benefits of regression analysis are used in almost all fields of life, both in the fields of economics, industry and employment, history, government, environmental sciences, etc. In this article reveals what assumptions must be met for an ideal regression to be obtained. How it is applied in real life can be demonstrated in this research.

1. Introduction
Regression is a statistical analysis method used to see the effect between two or more variables. Regression analysis studies the relationship between one or more variables / independent variables (X) with one dependent variable (Y). In the study of free variables (X), usually variables that are determined by researchers are free, for example drug dosage, storage time, preservative levels, livestock age and so on. Besides, the free variable can also be an independent variable, for example in measuring body length and body weight of a cow, because body length is easier to measure, body length is included in the free variable (X), while bodyweight is included in the dependent variable (Y). While the dependent variable (Y) in the study was the measured response due to the treatment/independent variable (X). for examples the number of red blood cells due to treatment with a certain dose, the number of microbial meat after being stored for several days, the weight of the chicken at a certain age and so on.

Linear regression is one of the types of forecasting or prediction analysis that is often used on quantitative scale data (intervals or ratios). The objectives of doing linear regression include: Is a set or set of predictor variables significant in predicting response variables? Which predictor variable is significant in explaining the response variable? This is indicated by the regression estimation coefficient. This estimation coefficient will later form the regression equation. Regression problems are problems studying the input-output relationship of data when an output variable is a real number [1]. Linear regression models a dependent variable Y in terms of a linear combination of p independent variables X=[X1,…,Xp] and estimates the coefficients of the combination using independent observations (Xi,Yi),i=1,…,n [2].

In its simplest form, one independent variable (X) with one independent variable (Y) has the following equation: Y =a +bx. Here a is called intercept and b is the direction coefficient or beta coefficient. In the understanding of the equation function of the lines Y + a + bx there is only one that can be formed from two points with different coordinates, namely (X1, Y1) and X2, Y2). This means we can make a lot of equation lines in other forms through two points for different coordinates / not coincide. in a linear equation, the relationship between two variables when drawn graphically, all X
and Y values corresponding to the equation $Y = a + bX$ will fall in a straight line. The line is called the regression line.

There are two types of relationship between the two variables: a) functional relationship; and b) regression relationships [3]. There is said to be a functional relationship if there is a true value of $Y$ for every possible $X$ value and vice versa. This relationship is mostly found in natural science, whereas regression relationships occur if there is no true value of $Y$ for each $X$ value and vice versa. For each $X$ value there are many $Y$ values, as long as $Y$ is not completely determined by $X$. and vice versa. Regression relationships are found in social science. Linear equations have many uses and are important not only because there are many relationships in this form, but also because they are often used in approaches to complex relations.

For example, for phenomena that include rice yields and the volume of fertilizer used, instead the independent variable or predictor $X = $ fertilizer volume is taken and the dependent variable or response $Y = $ yield. For three variables including bacterial growth, types of intermediaries where bacteria live and time, a response can be taken $Y = $ bacterial growth, predictor $X_1 = $ kinds of intermediates and predictor $X_2 = $ time. But two variables about weight and height, one can be chosen as an independent variable.

Statistical inference aims to infer population which in general by using the results of sample data analysis. Specially for regression, trying to determine the functional relationships that are expected to apply to the population-based on sample data taken. The functional relationship will be written in a mathematical equation called a regression equation [4]. Regression models or equations for the general population are written:

$$
\mu_{y|x_1,x_2,x_3,\ldots, x_k} = ^q(X_1, X_2, X_3, \ldots, x_k | \theta_1, \theta_2, \ldots, \theta_m)
$$

(1)

with $\theta_1, \theta_2, \ldots, \theta_m$ the parameters in the regression. Example of simple regression for population with an independent variable known as simple linear regression with the model.

$$
\mu_{y|x} = \theta_1, \theta_2 x
$$

(2)

with $\theta_1, \theta_2, \ldots, \theta_m$ as a parameters.

How can the regression equation be determined if the observations have been obtained? There are two types of methods known, namely the freehand method and the least-squares method. In this study, the least-squares method was used.

2. Methods

The data in this study were obtained from observations of many people who came (X) and many people who borrowed books (Y) in a school library for 30 days [5]. This research method uses quantitative methods [6]. The least-squares method is used to get parameter estimators and estimates. This method stems from the phenomenon that the square power (squared) rather than the distance between the points with the regression line being sought must be as small as possible.

3. Result and discussion

This study obtained a lot of visitors and borrowers of books in a school library within 30 days. For phenomena consisting of an independent variable $X$ and an independent variable $Y$ with a linear regression model for the population that can be predicted, it is necessary to estimate the regression parameters so that the regression equation is obtained. So that the linear regression model population as follows:

$$
\mu_{y|x} = \theta_1, \theta_2 x
$$

(3)

Will be estimated by the prices $\theta_1$ and $\theta_2$ by a and b so that the regression equation is obtained using the sample area, which is $\hat{Y} = a + bX$. Data from observations of many visitors and book borrowers in a school library is shown in the following table.
Table 1. Many visitors and book borrowers in a school library for 30 Days.

| Visitors | Book Borrowers | Visitors | Book Borrowers |
|----------|---------------|----------|---------------|
| 40       | 35            | 36       | 34            |
| 34       | 32            | 39       | 35            |
| 32       | 31            | 36       | 32            |
| 34       | 30            | 34       | 30            |
| 39       | 36            | 41       | 37            |
| 40       | 37            | 33       | 32            |
| 42       | 35            | 37       | 32            |
| 38       | 36            | 38       | 34            |
| 30       | 29            | 34       | 32            |
| 40       | 33            | 37       | 34            |
| 35       | 32            | 36       | 30            |
| 42       | 36            | 42       | 38            |
| 34       | 31            | 37       | 33            |
| 33       | 31            | 40       | 36            |
| 40       | 38            | 32       | 30            |

It will be determined the regression equation $Y$ for $X$ which is estimated to best fit the state of the data obtained. So that the scatter diagram shows that the location of the points is around a straight line. In making scatter diagrams using Ms. Excel is as follows [7,8]:

Figure 1. Make 2 arrays.

Figure 2. Fill with quantitative data. Quantitative means that the data is in the form of numbers or also called numeric data. A little review, that the Pearson product-moment test is a correlation test that is used on two numerical data scales (intervals or ratios).
Figure 3. Pearson's formula in excel, there is already a built-in feature to make it easier for Excel users to do the Pearson test, namely the "Pearson" formula. Examples are as follows: Create the code as follows in cell D2: = Pearson(A2: A20; B2: B20).

Figure 4. Interpretation or how to read excel: Pearson r value > 0: This means there is a positive relationship. Approaching 1: Means a strong relationship. Next, we will do the Pearson test using a scatter diagram of how to Make a Scatter Diagram: On the menu, click INSERT, click Scatter, select Scatter with Smooth Markers.

Figure 5. Right-click, block array 1 and 2 (visitors and borrowers).
Interpretation of Pearson's Scatter Test Diagram:
- Trendline lines: Tilt to the right: The nature of a positive relationship
- Markers/plots in blue: form a narrow straight line: very close relationship
- R² = 0.7962: regression value (square of r Pearson) = 0.7962
- Y = 0.7174X + 6.8698: the number of borrowers can be predicted by 0.7174 times the number of visitors plus a constant of 6.8698.
- Constant = 6.8698 means that many borrowers, without being influenced by visitors can change by +6.8698.

Now we use the formula [9,10]:

\[ a = \frac{\sum Y(\sum X^2) - \sum X \sum XY}{n \sum X^2 - (\sum X)^2} \quad (4) \]

\[ b = \frac{n \sum XY - \sum X \sum Y}{n \sum X^2 - (\sum X)^2} \quad (5) \]

So that the formula can be used, we calculate the units needed and the data should be arranged as the following table:
Table 2. List of values for calculating linear regression coefficients.

| X  | Y  | X.Y | Y.Y | X.X |
|----|----|-----|-----|-----|
| 40 | 35 | 1400 | 1225 | 1600 |
| 34 | 32 | 1088 | 1024 | 1156 |
| 32 | 31 | 992  | 961  | 1024 |
| 34 | 30 | 1020 | 900  | 1156 |
| 39 | 36 | 1404 | 1296 | 1521 |
| 40 | 37 | 1480 | 1369 | 1600 |
| 42 | 35 | 1470 | 1225 | 1764 |
| 38 | 36 | 1368 | 1296 | 1444 |
| 30 | 29 | 870  | 841  | 900  |
| 40 | 33 | 1320 | 1089 | 1600 |
| 35 | 32 | 1120 | 1024 | 1225 |
| 42 | 36 | 1512 | 1296 | 1764 |
| 34 | 31 | 1054 | 961  | 1156 |
| 33 | 31 | 1023 | 961  | 1089 |
| 40 | 38 | 1520 | 1444 | 1600 |
| 36 | 34 | 1224 | 1156 | 1296 |
| 39 | 35 | 1365 | 1225 | 1521 |
| 36 | 32 | 1152 | 1024 | 1296 |
| 34 | 30 | 1020 | 900  | 1156 |
| 41 | 37 | 1517 | 1369 | 1681 |
| 33 | 32 | 1056 | 1024 | 1089 |
| 37 | 32 | 1184 | 1024 | 1369 |
| 38 | 34 | 1292 | 1156 | 1444 |
| 34 | 32 | 1088 | 1024 | 1156 |
| 37 | 34 | 1258 | 1156 | 1369 |
| 36 | 30 | 1080 | 900  | 1296 |
| 42 | 38 | 1596 | 1444 | 1764 |
| 37 | 33 | 1221 | 1089 | 1369 |
| 40 | 36 | 1440 | 1296 | 1600 |
| 32 | 30 | 960  | 900  | 1024 |
| 1101 | 996 | 36816 | 33304 | 40773 |

To determine the linear regression between visitors (X) and book borrowers (Y), in this case, we can predict Y if X is known. The required values are:

\[
X = 1101; \quad Y = 996; \quad X.Y = 36816; \quad X^2 = 40773; \quad Y^2 = 33304; \quad n = 30
\]

The coefficients a and b obtained for the regression of Y over X are:

\[
a = \frac{(996)(40773) - (1101)(36816)}{30(40773) - (1101)^2} = 6.87; \quad b = \frac{30(36816) - (1101)(996)}{30(40773) - (1101)^2} = 0.72
\]

Then we get the coefficient value of a = 6.87 and b = 0.72. Thus the linear regression equation Y for X for the above problem is:

\[
\hat{Y} = a + b \cdot X = 6.87 + 0.72 \cdot X
\]

The independent variable Y in regression is symbolized by \(\hat{Y}\) to express the value of Y obtained from the regression, not from the observations. The graph of the regression is shown in table 1 above. The coefficient b is called the linear regression direction coefficient and states the average change in the variable Y for each change in variable X by one unit. This change is an increase if the coefficient b is positive and a reduction if it is negative. From this problem the value of b = 0.72 is positive so we can say that for each visitor (X) increases with a book borrower, the average book borrower (Y) increases by 0.72 people or rounded up to 1 person. Linear regression equation using either the formula or with Ms. Excel shows the same linear regression equation.
4. Conclusion
Regression analysis obtained from this problem is then used to predict if the value of the independent variable is known. For example, X = 30 then by substituting into the linear regression equation obtained 28.5. So, the average borrower of books for every 30 visitors to the school library.

Acknowledgments
Thanks to Fakultas Ilmu Terapan dan Sains Institut Pendidikan Indonesia Garut

References
[1] Pak S and Oh T 2010 Correlation and simple linear regression. J. Vet. Clin. 27 427–34
[2] Ranganathan S, Nakai K and Schonbach C 2018 Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics (Elsevier)
[3] Olive D J 2017 Multiple linear regression Linear Regression (Springer) pp 17–83
[4] Barbur V A 1994 Introduction to Linear Regression Analysis. J. R. Stat. Soc. Ser. D (The Stat. 43
[5] Garfield J 2002 The challenge of developing statistical reasoning J. Stat. Educ. 10
[6] Santoso S 2003 Statistik Deskriptif: Konsep dan Aplikasi dengan Microsoft Exel dan SPSS Yogyakarta, Andi Offset
[7] Maryati I, Sumartini T S, Noordyana M A, Sofyan D and Madio S S 2019 Experiences of analysis ToolPak in Matrix Algebra Journal of Physics: Conference Series vol 1402 (IOP Publishing) p 77087
[8] Sahid 2018 Analisis Data Statistik dengan MS Excel Juridikmat FMIPA UNY
[9] Siregar S 2014 Statistika Parametrik untuk penelitian kuantitatif: dilengkapi dengan perhitungan manual dan aplikasi versi 17
[10] Sarmanu 2017 Dasar Metodologi Penelitian Kuantitatif Kualitatif dan Statistika (Surabaya: Airlangga University Press)