STATISTICS FOR RESEARCHERS

Principles of Regression Analysis

NJ Gogtay, SP Deshpande, UM Thatte

Introduction

While correlation analysis helps in identifying associations or relationships between two variables,[1] the regression technique or regression analysis is used to "model" this relationship so as to be able to predict what will happen in a real-world setting. Let us understand this with something that happens in daily life. As children, we are often told by our parents that education is a vital part of our lives and a "good" education will help us get "good jobs" and thus "good wages"! If we were to explore the relationship between education and wages, we could think up two questions – a) Does a relationship exist between education and wages? And a related but more interesting and important question b) for *every extra year spent in education*, do wages commensurately increase [and if yes, by how much?]? The latter question is moving into the realm of prediction. Regression attempts to answer these and similar questions regarding relationships between variables.

Correlation Versus Regression- Understanding the Distinction

The difference between correlation analysis and regression lies in the fact that the former focuses on the strength and direction of the relationship between two or more variables *without making any assumptions about one variable being independent and the other dependent* [see below], but regression analysis assumes a dependence or causal relationship between one or more independent variables and the dependent variable. A correlation analysis with a scatter plot and a regression line[1] is however a prerequisite to regression and both analyses are often carried out together.

Dependent and Independent Variables

An important first step before carrying out a regression analysis is to understand the concept of independent and dependent variables. An independent variable, as the name suggests is a "stand alone" variable, and one that remains unaffected by other variables that are measured in a study. The "dependent" variable is the one that is usually of interest to the researcher and *alters in response to change/s in the independent variable*.

Let us understand this concept with the following two research questions – 1) "As age increases, does the risk of developing diabetes as measured by HbA1C or blood sugar levels increase? Or b) "Given that a patient has both diabetes and hypertension, what is his risk of developing coronary artery disease [CAD]"? In the former example, we are exploring the relationship between age and diabetes and in the latter, the relationship between two variables with a third variable– i.e., diabetes and hypertension with CAD. In the first example, age would be the independent variable while diabetes, which is *dependent on age* would become the "dependent" variable. In the second example, since CAD is associated with both diabetes and hypertension, CAD would become the dependent variable and diabetes and hypertension would be the two independent variables. If you will notice, diabetes in one example is a dependent variable and in the other, an independent variable. Thus, what is a dependent and what is an independent variable needs to be defined *à priori* by the researcher before carrying out the regression analysis. The choice of the variables would in turn be defined by the research question and the hypothesis being explored. Independent variables are often called predictor variables or exogenous variables and dependent variables are called prognostic or endogenous variables.

Types of Regression

Essentially in medical research, there are three common types of regression analyses that are used viz., linear, logistic regression and Cox regression. These are chosen depending on the type of variables that we are dealing with (Table 1). Cox regression is a special type of regression analysis that is applied to survival or "time to event "data and will be discussed in detail in the next article in the series.

Linear regression can be simple linear or multiple linear regression while Logistic regression could be Polynomial in certain cases (Table 1).

The type of regression analysis

Department of Clinical Pharmacology, Seth GS Medical College and KEM Hospital, Mumbai, Maharashtra
Received: 07.01.2017; Accepted: 15.01.2017

Table 1: Types of regression

Type of regression	Dependent variable and its nature	Independent variable and its nature	Relationship between variables
Simple linear	One, continuous, normally distributed	One, continuous, normally distributed	Linear
Multiple linear	One, continuous	Two or more, may be continuous or categorical	Linear
Logistic	One, binary	Two or more, may be continuous or categorical	Need not be linear
Polynomial (logistic) [multinomial]	Non-binary	Two or more, may be continuous or categorical	Need not be linear
Cox or proportional hazards regression	Time to an event	Two or more, may be continuous or categorical	Is rarely linear

to be used in a given situation is primarily driven by the following three metrics

- Number and nature of independent variable/s
- Number and nature of the dependent variable/s
- Shape of the regression line

A. Linear regression: Linear regression is the most basic and commonly used regression technique and is of two types *viz.* simple and multiple regression. You can use Simple linear regression when there is a single dependent and a single independent variable (e.g. for the research question described above "As age increases, does the risk of developing diabetes increase as measured by HbA1C or blood sugar levels?". Both the variables **must** be continuous (quantitative data[2]) and the line describing the relationship is a straight line (linear).

Multiple linear regression on the other hand can be used when we have one continuous dependent variable and two or more independent variables, for example when we want to answer the second question mentioned above, "Given that a patient has both diabetes and hypertension, what is his risk of developing coronary artery disease (CAD)". Importantly, the independent variables could be quantitative or qualitative. Both the independent variables here could be expressed either as continuous data (blood pressure or HbA1C values) or qualitative data (presence or absence of diabetes as defined by the ADA 2016 or hypertension as defined by JNC VIII).[3,4] A linear relationship should exist between the dependent and independent variables.

B. Logistic regression: This type of regression analysis is used when the dependent variable is binary in nature. For example, if the outcome of interest is death in a cancer study, any patient in the study can have only one of two possible outcomes- dead or alive. The impact of one or more predictor variables on this binary variable is assessed. The predictor variables can be either quantitative or qualitative. Unlike linear regression, this type of regression *does not* require a linear relationship between the predictor and dependent variables.

For logistic regression to be meaningful, the following criteria must be met/satisfied

- The independent variables must not be correlated amongst each other e.g. if the dependent variable is presence (or absence) of post-operative wound infection and the independent variables are duration of surgery, extent of blood loss and Hb levels at day 7, it is intuitive that all are correlated amongst each other and therefore will lead to erroneous results. What should be taken as the independent variable is only the duration of surgery.
- The sample size should be adequate

If the dependent variable is **non-binary** and has more than two possibilities, we use the multinomial or polynomial logistic regression.

Step Wise Regression

Stepwise regression is an automated tool that can be used in the exploratory stages of model building to identify a useful subset of predictor variables. The process systematically adds the most significant variable or removes the least significant variable during each step. The idea behind using such automated tools is to maximize the power of prediction with a minimum number of independent variables.

Steps in Conducting a Regression Analysis

Regression analysis is done in 3 steps:

1. Analyzing the correlation [strength and directionality of the data]
2. Fitting the regression or least squares line, and
3. Evaluating the validity and usefulness of the model.

Step 1: This has been described in the article on correlation analysis[1]

Step 2: *Fitting the regression line*

Conventionally, in mathematics, the equation of a line is given by the formula "y = mx+c", where, m is the "slope" or gradient of the line and "c" is where the line "cuts" the y-axis, also called as the "intercept", (Figure 1). In regression, this equation is given by

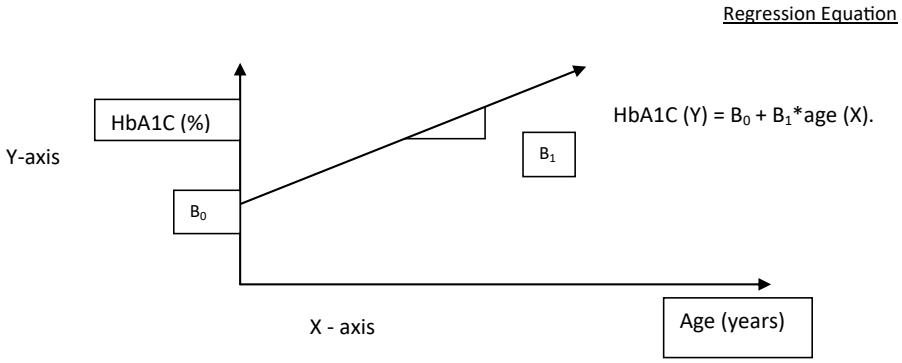

Fig. 1: Regression analysis exploring the relationship between Age and HbA1C

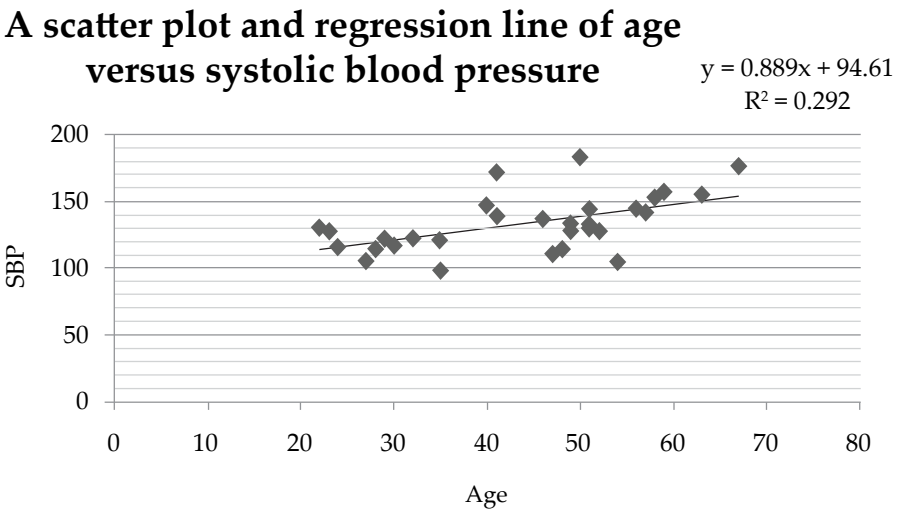

Fig. 2: Scatter plot and regression Line

$y=B_0 +B_1x$ which are the notations commonly used by statisticians. Here, B_0 represents the intercept, while B_1 represents the slope. In our example of relationship between age and HbA1C levels, the equation would be given by HbA1C levels (y) = B_0 + B_1*age (x) (Figure 2).

The intercept B_0 is that value of Y or the dependent variable (HbA1C, in this case), when the value of the predictor variable is zero (age). In reality age can never be zero, so this is the value of HbA1C that is present regardless of age. The equation $y = B_0 + B_1x$ classically describes simple linear regression. For Multiple linear regression, the equation would be $y = B_0 + B_1X_1 + B_2X_2 + B_3X_3$+……+$B_kX_k$ depending upon the number of predictor variables [X_1, X_2 and so on] and k is the kth predictor variable.

Step 3: Evaluating the validity of the model

Validation techniques can be broadly divided into two – numerical and graphical. An easy numerical technique is to look at the value of R^2. You will recollect that the coefficient of correlation (r)[1] is squared to obtain the Coefficient of Determination or R^2. This is the value that is used for predicting the extent of variability in the dependent variable that can be explained by the independent variable. In our example (Figure 1), the R^2 is 0.29 or 29% indicating that only 29% of the variability in SBP can be explained by age. If the value of R^2 is high, then we could assume that the model has a high predictive value. The value of 29% indicates that 71% of the variability still remains unaccounted for and thus model may not have a great predictive value. It is also useful to remember that the value of R^2 is affected by the choice of the independent variables and the presence of outliers[1] and hence researchers need to be careful while using R^2 alone for validation. Graphical analysis of residuals is a graphical technique for validation that uses graphs to visually inspect the data to assess its robustness.

Table 2: Age and Blood pressure in women (n=30)

Sr. No.	Age (years)	Systolic blood pressure (mm Hg)
1	22	131
2	23	128
3	24	116
4	27	106
5	28	114
6	29	123
7	30	117
8	32	122
9	35	99
10	35	121
11	40	147
12	41	139
13	41	171
14	46	137
15	47	111
16	48	115
17	49	133
18	49	128
19	50	183
20	51	130
21	51	133
22	51	144
23	52	128
24	54	105
25	56	145
26	57	141
27	58	153
28	59	157
29	63	155
30	67	176

Utility of Regression Analysis

The example given below highlights the utility of regression analysis. Table 2 has values of systolic blood pressure [SBP] for 30 women with age being presented in an ascending order. Age here would be the independent variables and SBP, the dependent variable.

We first make a scatter plot and eye ball the data and then

subsequently generate the regression line and the regression equation (Figure 1).

The regression equation here is y = 0.889x + 94.61

OR

SBP = 94.61 + 0.889 x age

The given data set if you will notice does not contain the age 60 or the corresponding SBP. Say we as asking the question, what would be the SBP of a 60-year old woman? We would "fit" 60 into the regression equation and calculate it as

SBP = 94.61 + 0.889 x 60 or 148 mm Hg.

It is important to note here that multiple lines can be drawn through the data set and hence we need to define the criterion for drawing the line. The method that is most commonly used is called the "least squares methods".

When we apply this equation to the population for making a prediction, we would really not be able to predict either the systolic blood pressure perfectly. Hence, we need to taken into account an "error" or "deviation" that is likely to occur when this equation is used. Thus, the equation would read as below

$$y = \underbrace{B0 + B1x}_{\text{linear component}} + \overbrace{e}^{\text{noise}}$$

Simply put, this could be interpreted as

Response = signal + noise

Or in medical research this would be

Outcome = prediction + deviation

Testing for Significance

Once, a regression equation is generated, the next logical step is to "test for significance" and generate a p value. There are three different ways in which this can be done (a) Carrying out an overall test of significance where all the predictor [X] variables [when there are multiple predictor variables] and all regression coefficients [B_1, B_2.....B_k] are tested together. (b) Testing a few X variables- here we can choose a few select X variables [these are the ones the researcher may think are maximally relevant] alone to be tested for their impact on the Y variable and finally (c) A test for individual significance where a single X variable is tested for its impact on the Y variable.

Applications of Regression Analysis

There are three major uses of regression analysis – attributing causality [cause and effect relationship], forecasting and prediction. These are explained with three examples.

Causality

Palmer KT and colleagues assessed working aged people from the general population in the United Kingdom to estimate the risks of occupational exposure to noise on self reported hearing difficulties and tinnitus using a validated questionnaire. The study showed that, in both sexes, after adjustment [see below] for age, the risk of severe hearing difficulty and persistent tinnitus rose with years spent in a noisy job indicative of a cause -effect relationship.[5]

Forecasting

Efficient management of patient flow in emergency departments (EDs) is a very important issue for hospital administrators. Marcilio I and colleagues[6] studied diverse models in an attempt to forecast the daily number of patients seeking emergency department [ED] services in a general hospital in Sao Paolo, Brazil, using calendar variables and ambient temperature reading as the independent variables. They found that the mean number of ED visits was 389 [166-613] with a seasonal distribution with the highest patient volumes seen on Mondays and lowest on weekends. Calendar variables rather than temperature were better at forecasting. They concluded that this data could be used for better allocation of personnel for the management of ED services.

Prediction

Brazer *et al*[7] used logistic regression to predict risk factors for colorectal cancer in a community practice where they studied 461 consecutive patients undergoing colonoscopy. Of these, 129 had adenomatous polyps (pre-cancerous) and 34 had colorectal cancer. They randomly chose 292 patients and evaluated the impact of several independent variables in a model that looked at prediction of occurrence of colorectal cancer. Five variables were identified to be predictive- the patient's age, sex, hematocrit, fecal occult blood test result and indication for colonoscopy. When this model was applied to the remainder of the 169 patients, it was found to be a reliable indicator of risk of colorectal neoplasia.

Understanding what Confounders are and the Concept of "Adjustment"

In study by Palmer and colleagues presented earlier,[6] the relationship between being in a noisy job and the extent of hearing difficulty and tinnitus was assessed. Intuitively, we do know that noise is not the only reason why hearing difficulty can occur. This may be related to stress, gender, tiredness, older age and many more factors. A confounder is defined as that variable that is "hidden" or "lurking" and was not accounted for or thought of initially and impacts the outcome being studied. Confounders [to the extent possible] need to be identified before the start of the study and addressed during analysis by a process called as "adjustment". This is a statistical technique to eliminate the influence of one or more confounders on

the treatment effect. Simply put, it can be understood as a process of "statistical correction" that is done once the data is gathered. In regression analysis, once a significant association between the independent variable/s and dependent variable has been found, it is important to see if this significance still persists after potential confounders have been adjusted for. In the study by Palmer, age was identified as a potential confounder *a priori* [since we know the hearing worsens with age] and adjusted for in the final analysis. Post adjustment, in both sexes, the relationship between years spent in a noisy job and severe hearing difficulty continued to remain significant [relative to those in non-noisy jobs].

Conclusion

In summary, regression analysis is a statistical tool that helps evaluate relationships between a dependent variable and one or more independent or predictor variables. More specifically, it helps us understand how the dependent variable changes with changes in the independent variable and thus finds its application in forecasting and predicting. The technique must however be used with clear understanding of the assumptions in each type of regression analysis, their limitations and the potential error that can occur when models are applied to a larger population.

References

1. Gogtay NJ, Deshpande S, Thatte UM. Principles of Correlation analysis. *J Assoc Phy Ind* 2017; 65:78-81.
2. Deshpande S, Gogtay NJ, Thatte UM. Data types. *J Assoc Phy Ind* 2016; 64:64-65.
3. Chamberlain JJ, Rhinehart AS, Shaefer CE, Neuman A. Diagnosis and management of diabetes:synopsis of the 2016 American Diabetes Assoication Standards of Medical Care in Diabetes. *Ann Intern Med* 2016; 164:542-552.
4. Hernandez-Vila E. A Review of the JNC 8 Blood Pressure Guideline. *Texas Heart Institute Journal* 2015; 42:226-228.
5. Palmer KT, Griffin MJ, Syddall HE, Davis A, Pannett B, Coggon D. Occupational exposure to noise and the attributable burden of hearing difficulties in Great Britain. *Occup Environ Med* 2002; 59:634-9.
6. Marcilio I, Hajat S, Gouveia N. Forecasting daily emergency department visits using calendar variables and Ambient temperature readings. *Academic Emergency Medicine* 2013; 20:769–777.
7. Brazer SR, Pancotto FS, Long TT 3rd, Harrell FE Jr, Lee KL, Tyor MP, Pryor DB. Using ordinal logistic regression to estimate the likelihood of colorectal neoplasia. *J Clin Epidemiol* 1991; 44:1263-70.