Introduction to widely used regression models in medical research using R, STATA, and SPSS: A tutorial

Mohammad Sadegh Loeloe
Center for Healthcare Data Modeling, Departments of Biostatistics and Epidemiology, School of public health, Shahid Sadoughi University of Medical Sciences, Yazd, Iran.

Farzan Madadizadeh (madadizadehfarzan@gmail.com)
Center for Healthcare Data Modeling, Departments of Biostatistics and Epidemiology, School of public health, Shahid Sadoughi University of Medical Sciences, Yazd, Iran.

Short Report

Keywords: Data analysis softwares, Regression Analysis, Types of variables, Medical Sciences

Posted Date: February 9th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2516586/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.
Introduction to widely used regression models in medical research using R, STATA, and SPSS: A tutorial

Mohammad Sadegh Loeloe¹, Farzan Madadizadeh*¹

¹ Center for Healthcare Data Modeling, Departments of Biostatistics and Epidemiology, School of public health, Shahid Sadoughi University of Medical Sciences, Yazd, Iran

*Corresponding author email address: madadizadefarzan@gmail.com

Abstract

Background: Regression models are often used to discover the relationship between variables. According to the type and scale of response variables, there are different type of regression models. Determining the type of regression is difficult for medical researchers.

Methods: In this tutorial, How to use different regressions models (ie, linear Regression, Ridge Regression, Polynomial Regression, Lasso Regression, Bayesian Linear Regression, Principal Components Regression, Partial Least Squares Regression, Elastic Net Regression, Support Vector Regression, Logistic Regression, Quantile Regression, Ordinal Regression, Poisson regression, Negative binomial regression, Quasi Poisson Regression, Fractional Regression, Cox Regression, Tobit Regression), their computational difficulties and their assumptions are shown. The application of each regression model is also specified with a medical example, then the model is implemented in different software (ie, SPSS, STATA, R) and its output is described.

Results: The models in this study are introduced according to the dependent variable’s type and scale; some models were used to increase the efficiency and improve the estimation of the relationship between specific types of variables.

Conclusion: Today, the use of regression models is growing and new statistical methods such as data mining basically use regression as a tool for forecasting. In this tutorial, we tried to help researchers to gain a better understanding of the computational implementation of different regression models by introducing widely used regression models, providing software instructions and interpretation.

Keywords: Data analysis softwares, Regression Analysis, Types of variables, Medical Sciences

Introduction:

Most people who deal with regression models use two or three general methods. On the other hand, regression models are much broader and are used in different ways to analyze qualitative and quantitative events. In general, regression models can be divided into linear and non-linear. Linear regression is a linear prediction function in which the dependent variable (the variable to be predicted) is predicted as a linear combination of independent variables. It means each independent variable is multiplied by the coefficient obtained for that variable in the estimation process [1].

In linear regression models, the basic assumption is that the error follow normal distribution. Sometimes this hypothesis is not acceptable; for example, the distribution of errors can be Poisson, Bernoulli, etc. These regression models are classified in a class called generalized
linear models (GLM) [2]. If the relationship between independent and dependent variables is in the form of a non-linear function regarding the parameters, the estimation of the model parameters can be obtained by non-linear regression [3]. In this case, the parameters can be estimated in two ways. One estimation method is to use Taylor expansion, and the other method is to use transformations that change the model under analysis to a linear regression model. Each regression method has its presuppositions determined by the characteristics of the Explanatory and the Response Variables. Note that explanatory variables are sometimes called independent variables, and response variables are called dependent variables. Of course, the terms Regressor or Predictor variables are also sometimes used for explanatory variables.

In this tutorial, first, we will introduce the hypotheses of the regression models’ accuracy. Then we will point to several widely used regression methods. Each of these methods will be placed in one of the above categories. We will present the codes for implementing models in the R programming language and how to implement them in STATA and SPSS software [4-6]. Knowing about these methods helps a researcher use the best strategy and model to analyze his data. As a result, the constructed models would have the highest efficiency and accuracy. The implementation of the models introduced in this study in software is based on the latest version of this software, i.e., version 4.2.0 for R, version 17 for STATA, and version 26 for SPSS software.

**Regression model assumptions**

**Troubleshooting outliers:** Since the presence of outlier points makes it impossible to estimate the parameters of the regression model correctly, we must ensure the absence of such data before performing calculations related to the regression model. When we encounter such observations, if possible, we replace them with appropriate values such as mean or median; otherwise, remove them from the data, then reconfigure. Outlier points here refer to observations with huge remnants compared to other points [3, 7].

**Independence of Explanatory Variables:** Explanatory variables in the regression model need to be independent. The model parameters would have a large variance if the explanatory variables are not independent and are linearly combined with other explanatory variables. Therefore, it is necessary to examine the explanatory variables in the regression model before implementing the model. If explanatory variables were not independent we will have collinearity or multicollinearity problem [7].

**Collinearity detection methods**

Collinearity detection can be done through the following methods.

1. Large changes in the estimation of regression coefficients when an explanatory variable is added or removed.

2. The hypothesis test for each of the regression coefficients in the multiple regression votes for their meaninglessness, while the analysis of the variance table finds the regression model appropriate. In other words, the assumption that the coefficients are zero in the F test is rejected.
3. There is a discrepancy between simple univariate and multiple regression results. In the test related to multiple regression, the assumption of zero coefficient of one explanatory variable is confirmed. In comparison, in simple regression of one variable, the test indicates the coefficient of that variable.

4. Use of Tolerance or Variance Inflation Factor, called VIF for short, for the regression model. In this case, the computational relationship for these two indicators is written as follows.

\[ Tolerance = 1 - R_j^2 \] (Eq.1)

\[ VIF = \frac{1}{Tolerance} \] (Eq.2)

Where \( R_j^2 \) is the coefficient of determination (R Square) of the regression model on the \( j \)th descriptive variable as the response variable with other explanatory variables as the independent variables. If the tolerance is less than 0.1 or the VIF is more than 10, it will indicate a Collinearity between the variables.

**Collinearity Troubleshooting**

One of the problems that alignment creates in the OLS regression model is Overfitting. Many variables are used to reduce the error rate to generate a regression model. But, the parameters will change drastically as soon as data is changed and re-fitting is performed. Thus, a completely different model is created. You can use the various solutions introduced in the list below to solve this situation [3].

1. Use the Stepwise Regression method to enter the variables with the highest correlation with the dependent variable [8].

2. Use Ridge Regression to reduce the number of variables and include the most dependent variables. It is done by creating a penalty parameter relative to the number of independent variables when estimating the parameters of the regression model [9].

3. Use other methods to estimate regression parameters, such as Lasso Regression, to avoid over-fitting the model [10].

4. Use more observations in the OLS method to reduce model error and variance of estimators.

5. Standardize the data before estimating the parameters of the regression model, i.e., using the standard score (Z-score) as follows:

\[ Z_i = \frac{X_i - \bar{X}}{S} \] (Eq.3)

In the above formula, \( Z_i \) is the standard score for data \( X_i \); \( \bar{X} \) is the mean; \( S \) is the standard deviation of the sample. This gives \( Z_i \) a mean of 0 and a variance of 1.
6. Use the Principle Components Analysis (PCA) technique to reduce the dimension of the problem and use the principal components to create a regression model. Since these components have the least correlation, these components will solve the Collinearity problem in the model.

When there is a strong correlation between explanatory variables, it is better to use Partial Least Square Regression (PLS) instead of Principal Component Regression (PCR). We also use Partial Least Squares regression (PLS) when the number of descriptive variables is large, and we want the most effective variables to be present in the model [11, 12].

7. Use the Lasso Regression method when the number of variables is large compared to the observations and the $X^TX$ matrix is not invertible [13].

8. The use of elastic net regression overcomes the disadvantages of Lasso and Ridge regressions by combining them, and it also is a reliable alternative to them. Thus, it is best to use elastic net regression to deal with a model whose descriptive variables are correlated [14].

1- **Linear Regression**

One of the simplest regression models is Linear Regression, in which the response variable or the dependent variable has numerical and continuous values. In this case, the relationship between the dependent and independent variables is linear in terms of model parameters. The simplest type of linear regression is Simple Linear Regression, which, unlike Multiple Linear Regression, has only one independent variable. Another type of linear regression is Multivariate Linear Regression, in which several dependent variables are predicted instead of predicting one dependent variable [1, 15].

If only one independent variable is used to identify and predict the dependent variable, the model is called Univariate Linear Regression. The form of a Simple Linear Regression model is as follows:

$$y = \beta_0 + \beta_1X + \epsilon \quad \text{(Eq.4)}$$

This relation is the equation of a line to which the error sentence or $\epsilon$ is added. The parameters of this linear model are intercept $\beta_0$ and the slope of $\beta_1$. Linear Slope in the univariate linear regression model shows how much the dependent variable changes relative to the independent variable. It means that by adding one unit to the value of the independent variable, how much the dependent variable will change. The intercept also represents the dependent variable value, which is calculated as zero for the components of the independent variable. Alternatively, a fixed value or the intercept can be considered the average value of a dependent variable if the independent variable is removed.

If several explanatory or independent variables are used in the regression model, the regression method is called Multiple Linear Regression.

Thus, if we have $n$ observations of the independent variable $p$ next to $X$ and want to establish a linear relationship with the response variable $y$, we can use the following Linear Regression Model:

$$y_i = \beta_0 + \beta_1x_{i1} + \ldots + \beta_px_{ip} + \epsilon_i, i = 1, 2, \ldots, n \quad \text{(Eq.5)}$$
Since the independent variable $X$ has a dimension $p$, we have replaced its value in each dimension with a one-dimensional independent variable. It is clear that index $i$ also shows the observation number. Finally, $\varepsilon$ is the error of the regression model. We should note that the relationship between the parameters is linear in the linear regression model. Thus a

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon$$

model is also a linear model in terms of parameters. In contrast, the

$$y_i = \beta_0 + \beta_1 x_i + \beta_1^2 x_i^2 + \varepsilon$$

relation will no longer be considered a linear relationship based on parameters.

In Simple Linear Regression, the relationship between the independent and dependent variables is expressed as the equation of a line. In Multiple Regression, if two independent variables are linearly related to a dependent variable, the relationship will take the form of a curve. If more than two independent variables are used in the linear regression model, the model appears as a Hyperplane [15].

In order to estimate Linear Regression Parameters, one of the standard methods is to use the Ordinary Least Square (OLS) model.

Software Guideline to do Linear Regression

The software code R for performing Linear Regression with an independent variable is:

- **R**:  
  ```R
  model = lm(y ~x , data =mydata)
  ```

And for a Multiple Model with more than one independent variable will be as follows.

- **R**:  
  ```R
  model = lm(y ~x1+x2+x3...+xp, data =mydata)
  summary(model)
  ```

As it turns out, the relationship between the dependent variable ($y$) and the other variables that play the role of independent variables is written by the `lm` function. Obviously, the dataset is also called `mydata`. In multiple regression, the interpretation of the estimated coefficients ($\beta_i$) is such that for one unit increase in the independent variable, provided that the other variables are constant, the dependent variable $y$ changes to the size $\beta_i$ of the unit.

The procedure for performing linear regression in STATA software is as follows.

- **STATA**: Statistics > Linear models and related > Linear Regression
  
  Or by using the `regress` function and with the following figure, for example, considering a dependent variable $y$ and two independent variables $x1$ and $x2$, it can be executed by typing the following command.

- **STATA command**: `regress y x1 x2`

And in SPSS software is as follows.

- **SPSS**: Analayze > Regression > Linear Regression
  
  Syntax:
  
  ```
  REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  ```
Example 1. A study to examine the factors affecting fertility and population growth, including fertility variables, the percentage of the total farmer population (Agriculture), the percentage of top people in the military exam (Examination), the percentage of people with higher education (Education), the percentage Consider Catholics and infant mortality rates.

The following output of R software results from a multiple linear regression model of the quantitative fertility variable as a dependent variable on other variables as explanatory variables.

Box 1. R output for example 1

```
Coefficients:
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept)  66.91518    10.70604    6.250       1.91e-07 ***
Agriculture -0.17211     0.07030    -2.448       0.01873 *
Examination -0.25801     0.25388     -1.016      0.31546
Education   -0.87094     0.18303    -4.758       2.43e-05 ***
Catholic     0.10412     0.03526     2.953       0.00519 **
Infant.Mortality 1.07705  0.38172     2.822       0.00734 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.165 on 41 degrees of freedom
Multiple R-squared: 0.7067, Adjusted R-squared: 0.671
F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

Based on the output of Box 1 and the value of R-squared, which is equal to 0.7067, about 70% of the changes in the dependent variable are expressed according to the model. Also in the column Pr (>|t|), which is the same as the probability value (p-Value), all parameters except the examination parameter can be considered significant. Being statistically significant here means that the parameter is opposite to zero. In this way, its presence in the model can be regarded as important. Note that the variable or statistically significant parameters (as opposed to zero) are marked with an * at the output of the lm function.

Sometimes a linear regression model can be used when the relationship between the dependent and independent variables is not in a straight line. Polynomial regression is one of the methods that model a relationship with a dependent variable using a polynomial in terms of an independent variable. In fact, sometimes, a Non-Linear model between explanatory and response variables can be defined as a k-degree polynomial. In such cases, the Polynomial Regression model can thoroughly express the relationship between predictive and dependent variables [16]. A relation or polynomial function of order K is written as follows:

\[ f(x) = a_0 + a_1x + a_2x^2 + \ldots + a_kx^k \] (Eq.6)
With the help of Support Vector Regression (sometimes known as SVR), we can also create linear and non-linear models and calculate their parameters. This is done by using a non-linear kernel function (such as polynomials). The parameters of this function are calculated in such a way to minimize the error, so the distance between the pages that create the act of separating the categories is maximized [3].

Another approach in linear regression is Bayesian linear regression, in which statistical analysis is performed in the framework of Bayesian inference. When the linear regression model errors follow a normal distribution, considering the previous distribution on the model parameters, the model prediction uses a later distribution obtained from Bayesian law [3, 6].

2. Generalized Linear Model

The Generalized Linear Model (GLM) is a generalized linear regression for data that does not have a normal distribution. For example, predicting the number of events, which is a discrete and counting quantity, or the waiting time, which is a positive quantity, can be done using a generalized linear model. This model consists of three components [1, 7]:

a) A distribution for the response variable $y$, which is usually selected from the exponential family of $d(\tau)$. Scattering is used to model high and low variances.

$$f_y(y|\tau, \theta) = h(y, \tau) \exp \left( \frac{b(\theta)^\top T(y) - A(\theta)}{d(\tau)} \right)$$ (Eq.7)

b) Linear prediction based on explanatory variable $X$

$$\eta = X \beta$$ (Eq.8)

c) The Link function $g$, which is a strictly uniform function and connects the above two components.

$$g(E[Y]) = X \beta$$ (Eq.9)

Logistic regression, binomial are Poisson are examples of the Generalized Linear Model.

2.1 Logistic Regression

In Logistic Regression, the dependent variable is Binary. This means that its values are classified into two categories, zero and one. In this case, the logistic regression model is written as follows [17, 18].

$$p(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}}$$ (Eq.10)

It is clear that in this regression model, the errors do not have a normal distribution and the dependent variable has a distribution of two or more polynomials. Therefore, a simple or linear regression model cannot be used in interpreting the coefficient estimation ($y = 1$) for the independent variable $x_i$ is equal to $e^{\beta_i}$ provided that the other independent variables are constant.
The types of logistic regression are as follows:

a. **Binary Logistic Regression**
   In this regression model, the variable is bilaterally dependent, so we will have only two groups.

b. **Multinomial Logistic Regression**
   In this regression model, the dependent variable is multidimensional (three groups and more). In fact, as with two-way logistic regression, we want to predict which of the groups a newer instance will fit into the model based on a series of predictor variables.

c. **Ordinal Logistic Regression**
   The Logistic Model can use these arrangements when the response variable is sequential. As a result, the results obtained from the sequential Logistics Model will be more interpretable and have more statistical power than the Multilevel Logistics Model (which considers the response variable as a Nominal variable).

**Software Guideline to do Logistic Regression**

The software code **R** for performing logistic regression is as follows.

- **R:**
  ```r
  model <- glm(y~X , data = Maydata, family = "binomial")
  summary(model)
  ```

The path of logistic regression in **STATA** software is as follows.

- **STATA:**
  - statistics > Generalized Linear models > Generalized Linear Model(GLM) > Binomial
  - **STATA command:** logit y x1 x2

And in **SPSS** software is as follows.

- **SPSS:**
  - Analyze > Generalized Linear models > Generalized Linear models > Binary logistic
  - **Syntax:**
    ```
    LOGISTIC REGRESSION VARIABLES y
    /METHOD=ENTER x1
    /CRITERIA=PIN(.05) POUT(.10) ITERATE(20) CUT(.5).
    ```

**Example 2.** The following output from R software examines the relationship between the number of cigarettes smoked per day (CD) and having cancer (LC) from the Breast Cancer Database (case2002).

**Box 2. R output for example 2**

```r
glm(formula = LC ~ CD, family = "binomial", data = case2002)
Coefficients:
  Estimate   Std. Error   z value Pr(>|z|)
(Intercept)  -1.53541  0.37707   -4.072  4.66e-05 ***
```
Based on the above output of Box 2, it can be said that increasing one cigarette a day increases the risk of breast cancer by $e^{0.05113} = 1.05$.

### 2.2 Quantile Regression

Quantile Regression can be considered an alternative to the Linear Regression model, which is resistant to problems caused by Outlier points, High Skewness, and Heteroscedasticity. In Linear Regression, the mean of the variable depends on the condition of the observations. In fact, the Linear Regression model considering the X matrix and the $\beta$ and $\alpha$ vectors is written as follows [3, 19].

$$E (Y | X ) = \alpha + \beta X \tag{Eq.11}$$

Since we have considered the average error sentence zero, the error value is not seen in this statement. Despite the dependent variable’s outlier, skewness, and heteroscedasticity data, the mean is not a good concentration indicator. Therefore, the Linear Regression method cannot describe the changes of such a variable well and create a suitable model to show the relationship between independent and dependent variables. Using multiples that are more resistant to the above conditions can provide a more complete and accurate regression model. In quantitative regression, the method is to estimate the quantities of the dependent variable (such as percentages, deciles, or quarters) as independent variables instead of estimating its mean. Note that in Multiple Regression, the dependent variable must be of a numerical type and from a scale type with continuous values to be able to calculate multiples.

If $Q_{Y|X} (p)$ is the quantile $p^{th}$ of the dependent variable Y with condition X, then the regression model will be as follows.

$$Q_{Y|X} (p) = \beta_p X \tag{Eq.12}$$

This model is very similar to Linear Regression. For this reason, Quantile Regression is sometimes referred to as the extended mode of Linear Regression. If we predict the regression model for different values of $p$ and draw the corresponding lines, we create a Multiple Regression. Note that the error minimization will be done according to the following equation.

$$p \left( \sum |e_i| \right) + (1 - p) \sum |e_i| \tag{Eq.13}$$

Thus, if the $p = 0.5$ value is quantile, it is converted to the median, and the regression model is called Median Regression.

Assume that the regression equation for the 25th percentile is obtained as follows:
\( Q_{y|x}(0.25) = 5 + 10X \) (Eq.14)

In this case, each unit of addition of the value of the independent variable (x) will increase by 10 units the value of the 25th percentile.

**Software Guideline to do Quantile Regression**

To use Quantile Regression in the R programming language, we will use the quantreg package. The following code specifies how to install and load this library.

- **R**: `install.packages("quantreg")`
  `library(quantreg)`

With the help of the `rq` function, it is possible to perform multiple regression. The code below performs multiple regression for the dependent variable y and the vector of the independent variables X from the `mydata` dataset.

- **R**: `model1 = rq(y~X,data = mydata,tau = 0.25)`
  `summary(model1)`

The value of \( p \) in the `rq` function is specified by the `tau` parameter. It is clear that the 25th percentile (first quantile) is used in the implemented model. If we consider the value of `tau` equal to 0.5, we will produce the median regression.

The procedure for performing multiple regressions in STATA software is as follows.

- **STATA**: `statistics > Linear models and related > Quantile Regression > Quantile Regression`
  Quantile regression of the 75th percentile of y on x1, x2

- **STATA command**: `qreg y x1 x2, quantile (.75)`
  Median regression of y on x1 and x2

- **STATA command**: `qreg y x1 x2`

And in SPSS software is as follows.

- **SPSS**: `Analyze > Regression > Quantile Regression`
  Syntax:

  QUANTILE REGRESSION y BY X1
  /CRITERIA QUANTILE=0.75 METHOD=AUTO IIDs=TRUE BANDWIDTH=BOFINGER
  TOL=0.000000000001 CONV=0.000001
  MAXITER=2000 CILEVEL=95
  /MISSING CLASSMISSING=EXCLUDE
  /MODEL INTERCEPT=TRUE
  /PRINT PARAMETER

  /PLOT PARAM_EST=50 PREDICTED_BY_OBSERVED=False MAX_CATEGORIES=10
  /PREDICT_EFFECTS NUM_TOP_EFFECTS=3.

**Example 3.** For the data in Example 1, we performed the 25th percentile regression again for the Fertility variable as a dependent variable and the other variables as discrete variables in the R software. The output of this example is shown below.

**Box 3. R output for example 3**
Based on the specified output bandwidth of Box 3, we can say that the hypothesis (that the variable coefficient is equal to zero) for the variables equal to zero in these ranges is not rejected. And the variable has no significant effect on the 25th percentile of the dependent variable, like the agriculture variable. Also, for the catholic variable, we can say that for a unit increase of this variable, the 25th percentile of the fertility variable increases by 0.06116, provided that other independent variables stay the same.

### 2.3 Ordinal Regression

Using the ordinal regression model would be better if the response (dependent) variable is Ordinal or Rank. Thus, if you encounter answers to a questionnaire using the Likert Scale, it is best to use sequential regression to model between this response variable and independent or explanatory variables. For example, you might want to measure how well patients’ pain improves with different drug doses based on a questionnaire whose answers consist of a 5-point Likert scale or spectrum. Such a model will be constructed using sequential regression.20

#### Software Guideline to do Ordinal Regression

The following command code is used to perform negative binomial regression in R. This is possible by calling the package or the ordinal library and using the clm function.

- **R:** library(ordinal)
  
  o.model <- clm( y~ x, data = mydata)

  summary(o.model)

The procedure for performing Ordinal Regression in STATA software is as follows.

- **STATA:** statistics > ordinal outco
  
  mes > Ordered logistic Regression

  Ordinal logit model of y on x1 and categorical variables a and b

  - **STATA command:** ologit y x1 i.a i.b

And in SPSS software is as follows.

- **SPSS:** Analayze > Regression > ordinal Regression

  Syntax:

  PLUM y BY x
  
  /CRITERIA=CIN(95) DELTA(0) LCONVERGE(0) MXITER(100) MXSTEP(5)
  PCONVERGE(1.0E-6) SINGULAR(1.0E-8)
Example 4. Consider the toxicity variable in a toxicology study that grouped response values into five categories in order. Temperature (temp) with two hot and cold levels is an explanatory variable.

Note that in regression models, when the independent variable is qualitative, we need Dummy variables. So for each variable with an m level, we must create an m-1 fictitious variable and define the model with those variables. For example, a fictitious variable is required for the temp variable in Example 4. Its value is equal to one if the temperature is hot; the value is equal to zero if the temperature is not hot. In this case, the cold temperature status is specified as a reference.

R output for example no. 4 is given below.

**Box 4. R output for example 4**

```r
o.model <- clm( toxicity~ temp, data =toxin )
summary(o.model)
```

| Coefficients: | Estimate | Std. Error | z value | Pr(>|z|) |
|---------------|----------|------------|---------|----------|
| tempwarm      | 3.683    | 1.465e+00  | 2.514   | 0.01194 * |
```

According to the output of Box 4, the variable temp affects the degree of toxicity. And the hot temperature increases the toxicity by 3.68 times compared to the cold.

2.4 Poisson regression

When the variable is numerically dependent, we use the Poisson regression model. For example, we use it to predict a. the number of telephone calls in a company based on sales b. the number of calls to the emergency room based on the occurrence of an accident, and c. the number of deaths due to air pollution [21, 22].

The dependent variable in this model must have the following conditions.

a. The dependent variable (y) has a Poisson distribution.

b. The dependent variable value should not be negative (because it is the result of counting).

c. The Poisson regression model should not be used when the values of the dependent variable do not belong to the set of natural numbers.

Software Guideline to do Poisson Regression

*The code you see below is for performing Poisson regression in R.*

- **R**:
  ```r
  pos.model<-glm(y~X, data=mydata , family=poisson)
  summary(pos.model)
  ```
In STATA software, it is as follows.

- **STATA**: statistics > Generalized Linear models > Generalized Linear Model(GLM) > poisson

Poisson regression of y on x

- **STATA command**: poisson y x

And the path of performing Poisson regression in SPSS software is as follows.

- **SPSS**: Analyze > Generalized Linear models > Generalized Linear models > poisson logliner

**Syntax:**

* Generalized Linear Models.

```
GENLIN y WITH x
/MODEL INTERCEPT=YES
/DISTRIBUTION=POISSON LINK=LOG
/Criteria METHOD=FISHER(1) SCALE=1 COVB=MODEL
/CRITERIA MAXITERATIONS=100 MAXSTEPHALVING=5
/PCONVERGE=1E-007 (ABSOLUTE) SINGULAR=1E-012
/ANALYSIS=TYPE=3(WALD) CILEVEL=95 CTYPE=WALD
/LIKELIHOOD=FULL
/MISSING CLASSMISSING=EXCLUDE
/PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION.
```

**Example 5.** To implement the Poisson regression model, we study a study that examines the effect of weather type on the number of accidents, the values of which are numerical.

**Box 5.** R output for example 5

```
pos.model<-glm(accident ~ weather, data=accident, family=poisson)
summary(pos.model)
```

| Coefficients: | Estimate | Std. Error | z value | Pr(>|z|) |
|---------------|----------|------------|---------|---------|
| (Intercept)   | 3.79674  | 0.04994    | 76.030  | < 2e-16 *** |
| dry           | -0.45663 | 0.08019    | -5.694  | 1.24e-08 *** |

According to Box 5, The rate of accidents in rainy weather increases by 58% compared to dry weather \( e^{0.45663} = 1.58 \).

**2.5 Negative binomial regression**

In a Poisson regression model, we may encounter the problem of overdispersion or under-dispersion. Negative binomial regression, which is another type of GLM model and is suitable for counting variables, is more resistant to Over-fitting and Under-fitting [22].

- **R**: `nb.model <- glm.nb(y~X, data=mydata)`

```
summary(nb.model)
```

As you can see here, we have used the glm.nb function to use the negative binomial regression model.
The procedure for performing Negative binomial Regression in STATA software is as follows.

- **STATA**: Statistics > Generalized Linear models > Generalized Linear model (GLM) > Negative binomial
  Negative binomial model of y on x1 and categorical variable a

- **STAT command**: `nbreg y x1 i.a`

The path of Negative binomial Regression in SPSS software is as follows.

- **SPSS**: Analyze > Generalized Linear models > Generalized Linear models > Negative binomial with log link
  Syntax:
  ```
  * Generalized Linear Models.
  GENLIN y WITH x
  /MODEL INTERCEPT=YES
  DISTRIBUTION=NEGBIN(1) LINK=LOG
  /CRITERIA METHOD=FISHER(1) SCALE=1 COVB=MODEL
  MAXITERATIONS=100 MAXSTEPHALVING=5
  PCONVERGE=1E-006 (ABSOLUTE) SINGULAR=1E-012
  ANALYSISTYPE=3 (WALD) CILEVEL=95 CITYPE=WALD
  LIKELIHOOD=FULL
  /MISSING CLASSMISSING=EXCLUDE
  /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION.
  ```

One of the regression methods that can be used as an alternative to Negative binomial regression is Quasi Poisson Regression. Although both methods produce the same answers, there are differences in the covariance matrix of the estimators. The variance of the Quasi Poisson model is a linear combination of the means, but the variance of the negative binomial regression model is a quadratic function of the mean [23].

**Example 6.** In R, we review a study that considers the number of students’ absence days due to illness (countable and dependent variable) according to their ethnicity, age, and gender. We only consider the gender variable as an independent variable in this example. The syntax and output code are as follows.

**Box 6. R output for example 6**

```r
nb.model <- glm.nb(Days ~ Sex, data = quine)
summary(nb.model)
Coefficients:

|            | Estimate | Std. Error | z value | Pr(>|z|) |
|------------|----------|------------|---------|---------|
| Intercept  | 3.01919  | 0.29706    | 10.163  | < 2e-16 *** |
| SexM       | -0.47541 | 0.39550    | -1.202  | 0.229355 |
```

According to the output of **Box 6**, gender is not statistically related to the number of absences of students due to illness.

### 2.6 Fractional regression
Dependent variables such as rates, quotas, and ratios, which take values between zero and one, exist in most sciences. Conventional regression methods cannot be used to analyze these variables. The problem with analyzing these variables is that the modeling has to be done so that the dependent variable is between zero and one. Relative response models (FRM) refer to models in which the dependent variable takes values greater than or equal to zero and less than and equal to one. The estimation of the parameters of these models is based on the quasi-maximum straight-line method (GMLE), which provides efficient and robust estimates under the conditions of general linear models [24, 25].

**Software Guideline to do Fractional Regression**

We will use the frm package or library to use Fractional Regression in the R programming language. The following code specifies how to install and load this library.

- **R**: install.packages("frm")
  ```
  library(frm)
  ```

With the help of the FRM function, it is possible to perform Fractional Regression. The code below performs fractional regression for the dependent variable y and the vector of the independent variables X from the mydata dataset.

- **R**: `modelfrm <- frm(y, X, linkfrac = 'logit')`
  ```
  summary(modelfrm)
  ```

The procedure for performing fractional regression in STATA software is as follows.

- **STATA**: statistics > Fractional outcomes > Fractional Regression
  Fractional probit model for y with values between 0 and 1 on continuous variable x1

- **STATA command**: fracreg probit y x1
  As above, but use logit distribution

- **STATA command**: fracreg logit y x1
  This type of regression is not yet applicable in SPSS software.

**Example 7.** The command and output code below a multiple model shows the fractional regression in R. The dependent variable part is a variable whose measured values are between zero and one.

**Box 7. R output for example 7**

```
library(frm)
y <- mydata$prate
x <- mydata[,c('mrate', 'age', 'sole', 'totemp1')]
myfrm <- frm(y, x, linkfrac = 'logit')

*** Fractional logit regression model ***

|           | Estimate | Std. Error | t value | Pr(>|t|) |
|-----------|----------|------------|---------|----------|
| Intercept | 1.074062 | 0.048902   | 21.963  | 0.000 ***|
| mrate     | 0.573443 | 0.079917   | 7.175   | 0.000 ***|
| age       | 0.030895 | 0.002788   | 11.082  | 0.000 ***|
| sole      | 0.363596 | 0.047595   | 7.639   | 0.000 ***|
| totemp1   | -0.057799| 0.011466   | -5.041  | 0.000 ***|

Number of observations: 4734
```
According to Box 7, all variables in the model are statistically significant. For example, it can be said that a unit increase in the age variable increases the dependent variable y by 

\[ e^{0.363596} = 1.44 \text{ times.} \]

### 2.7 Tobit regression

Tobit models, sometimes called sensitized or censored regressions, examine linear relationships in situations where a critical limit is observed to the right or left of a dependent variable. Hence it also refers to this particular type of censored regression prediction function. The presence of items above or below the critical limit in the dependent variable indicates a serious and oblique problem in the regression equation and requires using Tobit regression. In other words, this distribution level causes a serious error in using linear regression [3, 26].

**Software Guideline to do Tobit Regression**

*With the help of the Tobit function, it is possible to perform Tobit regression in R. The code below performs multiple regression for the dependent variable y and the vector of the independent variables X from the mydata dataset in R.*

- **R:**
  
  ```r
  modeltobit = tobit(y~X,data = mydata )
  summary(modeltobit)
  ```

*The procedure for performing Tobit regression in STATA software is as follows.*

- **STATA:** Statistics > Linear models and related > Censored regression > Tobit regression

  *Tobit regression of y on x1 and x2, specifying that y is censored at the minimum of y*

  - **STATA command:** Tobit y x1 x2, ll
    
    *As above, but where the lower-censoring limit is zero*

  - **STATA command:** Tobit y x1 x2, ll(0)
    
    *As above, but specify the lower- and upper-censoring limits*

  - **STATA command:** Tobit y x1 x2, ll(17) ul(34)
    
    *As above, but where lower and upper are variables containing the censoring limits*

  - **STATA command:** tobit y x1 x2, ll(lower) ul(upper)
    
    *This type of regression is not yet applicable in SPSS software.*
Example 8. The command and output code below show a multiple Tobit regression model in R. The dependent variable is affairs. Several observations in this dataset have been censored, so Tobit regression will be used.

Box 8. R output for example 8

```
tobit(formula = affairs ~ age + yearsmarried + religiousness + occupation + rating, right = 4, data = Affairs)
```

<table>
<thead>
<tr>
<th>Observations:</th>
<th>Total</th>
<th>Left-censored</th>
<th>Uncensored</th>
<th>Right-censored</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>601</td>
<td>451</td>
<td>70</td>
<td>80</td>
</tr>
</tbody>
</table>

| Coefficients:             | Estimate | Std. Error  | z value  | Pr(>|z|)  |
|---------------------------|----------|-------------|----------|-----------|
| (Intercept)               | 7.90098  | 2.80385     | 2.818    | 0.004834 ** |
| age                       | -0.17760 | 0.07991     | -2.223   | 0.026244 *  |
| yearsmarried              | 0.53230  | 0.14117     | 3.771    | 0.000163 ***|
| religiousness             | -1.61634 | 0.42440     | 3.809    | 0.000140 ***|
| occupation                | 0.32419  | 0.25388     | 1.277    | 0.201624    |
| rating                    | -2.20701 | 0.44983     | -4.906   | 9.28e-07 ***|
| Log(scale)                | 2.07232  | 0.11040     | 18.772   | < 2e-16 ***  |

According to Box 8, out of 601 observations, only 70 are uncensored. Other censored observations are left or right. Apart from the occupation variable, other variables are statistically significant. For example, with one unit increase in the age variable, if the other variables are constant, the affairs variable decreases by 0.017760 units.

2.8 Cox Regression

Cox regression is used for time-dependent data. To understand this better, consider the following examples:

a. The time elapsed from the time the user logs in to their account until they log out.
b. The duration of cancer treatment until the patient dies.
c. The time between the first and second heart attack.

As you can see, the Cox regression model is similar to the Logistic regression model. Except in logistic regression, the number of events is significant, but the length or time interval between events is considered in Cox regression.

Such analyses are known as Survival analyses. As it turns out, the dependent variable consists of two parts. In the first part, which is a constant value, the time or duration of an event is recorded. And in the second part, the occurrence or non-occurrence of that event is specified [27, 28].

Software Guideline to do Tobit Regression

Example 9. The following code in R is prepared to examine the relationship between the life expectancy of cancer patients (SurvObj) and other factors such as age, sex, and Karnofsky performance score, which can be seen with ph.karno, and weight loss (wt.loss).

Box 9. R output for example 9
As shown in Box 9, to perform Cox regression in R, you must first call the survival library and use the coxph function to perform this regression. The output of the above code is as follows.

The age and weight loss variables do not affect a person’s lifetime with cancer because they have a p-value greater than 0.05. For the gender variable, it can be said that the life expectancy of men with cancer is 0.598125 times that of women with cancer, provided that other independent variables are constant.

We use the following path or command code to perform Cox regression in STATA software.

- **STATA**: Statistics > Survival analysis > Regression models > Cox proportional hazards (PH) model
  Cox proportional hazards model with covariates x1 and x2.

- **STATA command**: stcox x1 x2
  To perform Cox regression in SPSS software, we use the following path.

- **SPSS**: Analyze > Survival > Cox regression
  Syntax:
  
  ```
  COXREG T
  /STATUS=C(1)
  /METHOD=ENTER X
  /PLOT SURVIVAL
  /CRITERIA=PIN(.05) POUT(.10) ITERATE(20).
  ```

**Discussion and conclusion:**

This study introduced several regression models widely used in various sciences, especially medical sciences. The types of models introduced in this study were presented according to the type and scale of the dependent variable. Some models were used to increase the efficiency and improve the relationship estimation between specific types of variables. This educational study aims to provide an implicit and non-systematic review of widely used models in medical sciences and how to implement these models in three software R, STATA, and SPSS. We have not involved the reader with purely theoretical and statistical content.
Also, the introduced models are used for response or dependent variables as univariate. If there is more than one measured variable for each test unit (Subject), you should refer to the generalization of multivariate models introduced in other books and articles.

According to what was stated, choosing the type of regression model to examine the relationships of independent and dependent variables in each study depends on the type of dependent variable and how it is measured. Therefore, medical researchers must learn enough about the study variables. After implementing the regression model, they should assess the hypotheses of the model’s accuracy stated in this study to increase the accuracy and efficiency of their model. In a nutshell, all the widely used regression models discussed here are summarized in Table 1.

<table>
<thead>
<tr>
<th>Type of Regression</th>
<th>Dependent variable</th>
<th>Type of relationship or application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple linear Regression</td>
<td>Continuous</td>
<td>Linear with one independent variable</td>
</tr>
<tr>
<td>Multiple linear Regression</td>
<td>Continuous</td>
<td>Linear with more than one independent variable</td>
</tr>
<tr>
<td>Ridge Regression Polynomial Regression</td>
<td>Continuous</td>
<td>Collinearity Troubleshooting Non-Linear</td>
</tr>
<tr>
<td>Lasso Regression</td>
<td>Continuous</td>
<td>Collinearity Troubleshooting</td>
</tr>
<tr>
<td>Bayesian Linear Regression Principal Components Regression</td>
<td>Continuous</td>
<td>Linear Collinearity Troubleshooting</td>
</tr>
<tr>
<td>Partial Least Squares Regression</td>
<td>Continuous</td>
<td>Collinearity Troubleshooting</td>
</tr>
<tr>
<td>Elastic Net Regression</td>
<td>Continuous</td>
<td>Collinearity Troubleshooting</td>
</tr>
<tr>
<td>Support Vector Regression</td>
<td>Continuous</td>
<td>Non-Linear</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>Binary</td>
<td>Non-Linear or GLM</td>
</tr>
<tr>
<td>Quantile Regression</td>
<td>Continuous</td>
<td>An alternative to the Linear Regression model, which is resistant to problems caused by Outlier points, High Skewness, and Heteroscedasticity</td>
</tr>
<tr>
<td>Ordinal Regression</td>
<td>Ordinal or Rank variable</td>
<td>Non-Linear or GLM</td>
</tr>
<tr>
<td>Poisson regression</td>
<td>Counting</td>
<td>Non-Linear or GLM</td>
</tr>
<tr>
<td>Negative binomial regression</td>
<td>Counting</td>
<td>An alternative to Poisson regression is that it is more resistant to over-fitting and under-fitting.</td>
</tr>
<tr>
<td>Quasi Poisson Regression</td>
<td>Counting</td>
<td>An alternative to Negative binomial regression</td>
</tr>
<tr>
<td>Fractional regression</td>
<td>Rates, Quotas, and Ratios</td>
<td>Linear and Non-Linear</td>
</tr>
<tr>
<td>Cox Regression</td>
<td>Time-dependent data</td>
<td>Survival analysis</td>
</tr>
<tr>
<td>Tobit regression</td>
<td>Censored</td>
<td>Linear</td>
</tr>
</tbody>
</table>

**Table 1. Summary of all the widely used regression models and their properties**

**Declarations**
Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Availability of data and materials
Not applicable

Competing interests
The authors declare that they have no competing interests.

Funding
The financial resources of this research will be provided by the support of Shahid Sadoughi University of Medical Sciences, Yazd.

Authors' contributions
Mohammad Sadegh Loeloe was the main author and Farzan Madidizadeh was the supervisor of this study.

References

6. Mitchell MN. Interpreting and visualizing regression models using Stata: Stata Press College Station, TX; 2012.