Introduction

Our project is about some data which we need to make hypothesis test and regression for it. Our data explain briefly about baseball players so we will make relation between players’s ID and their salaries moreover we will make relation between IGID and year ID and the salaries . We are using regression with its kinds linear, multiple and poisson.

[**Null hypothesis**](https://en.wikipedia.org/wiki/Null_hypothesis)**(H0): -** A simple hypothesis associated with a contradiction to a theory one would like to prove.

[**Alternative hypothesis**](https://en.wikipedia.org/wiki/Alternative_hypothesis)**(H1): -** A hypothesis (often composite) associated with a theory one would like to prove.

**Linear regression** is an approach for modeling the relationship between a scalar [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable) *y* and one or more [explanatory variables](https://en.wikipedia.org/wiki/Explanatory_variable) (or independent variables) denoted *X*. The case of one explanatory variable is called [*simple linear regression*](https://en.wikipedia.org/wiki/Simple_linear_regression).

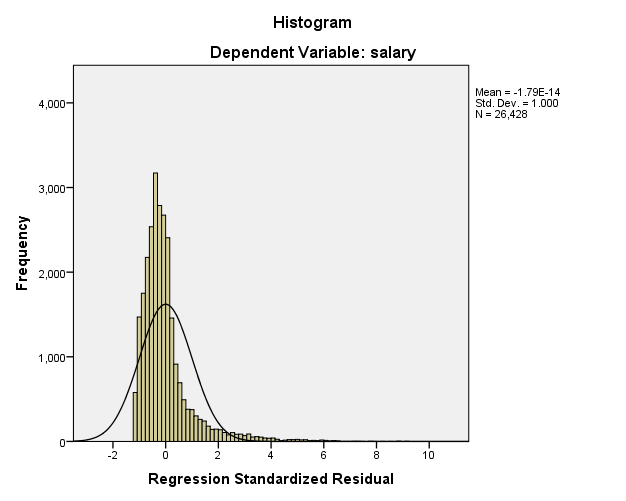
**Multiple regression** is an extension of simple linear **regression**. It is used when we want to predict the value of a variable based on the value of two or more other variables. The variable we want to predict is called the dependent variable

**Poisson regression** is a form of **regression** analysis used to model count data and contingency tables. **Poisson regression** assumes the response variable Y has a **Poisson** distribution, and assumes the logarithm of its expected value can be modeled by a linear combination of unknown parameters.

Abstract: -

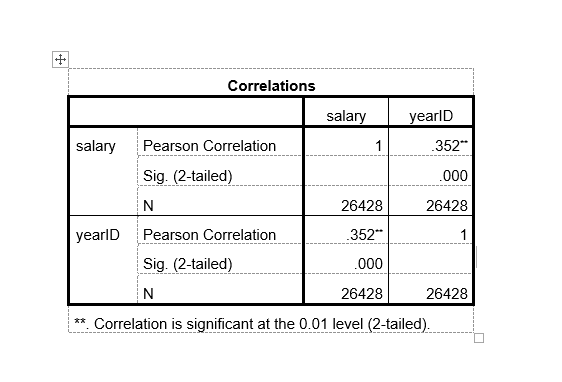
The project’s aim is to find a correlation between Salaries and year ID, IGID and player IDits aim to see the relation between them Questions: -

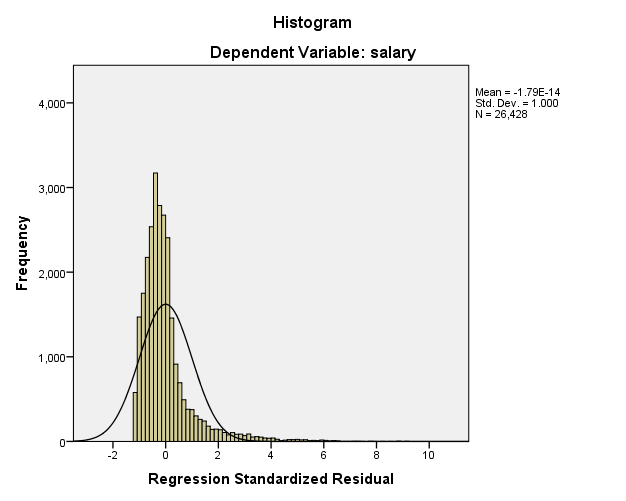
1. Is there any significant correlation between year ID and the salary ?

There is a postive correlaiton between the yea ID and saary. The correlation is postive because whenn the ge increase the salary increase. No signifcance as it is 0.01 so weak. 

1. Is there any significant correlation between IGID and salary?

There is no relation between IGID and the year ID. This explain that being in any team is not related to the the age of the player.



pa

Hypothesis Test: - it is a [hypothesis](https://en.wikipedia.org/wiki/Hypothesis) that is testable on the basis of [observing](https://en.wikipedia.org/wiki/Observable_variable) a process that is [modeled](https://en.wikipedia.org/wiki/Statistical_model) via a set of [random variables](https://en.wikipedia.org/wiki/Random_variable). A statistical hypothesis test is a method of [statistical inference](https://en.wikipedia.org/wiki/Statistical_inference). Commonly, two statistical data sets are compared, or a data set obtained by sampling is compared against a synthetic data set from an idealized model. A hypothesis is proposed for the statistical relationship between the two data sets, and this is compared as an [alternative](https://en.wikipedia.org/wiki/Alternative_hypothesis) to an idealized null hypothesis that proposes no relationship between two data sets. The comparison is deemed [*statistically significant*](https://en.wikipedia.org/wiki/Statistically_significant) if the relationship between the data sets would be an unlikely realization of the [null hypothesis](https://en.wikipedia.org/wiki/Null_hypothesis) according to a threshold probability—the significance level. Hypothesis tests are used in determining what outcomes of a study would lead to a rejection of the null hypothesis for a pre-specified level of significance. The process of distinguishing between the null hypothesis and the [alternative hypothesis](https://en.wikipedia.org/wiki/Alternative_hypothesis) is aided by identifying two conceptual types of errors [(type 1 & type 2)](https://en.wikipedia.org/wiki/Type_I_and_type_II_errors), and by specifying parametric limits on e.g. how much type 1 error will be permitted.

An alternative framework for statistical hypothesis testing is to specify a set of [statistical models](https://en.wikipedia.org/wiki/Statistical_model), one for each candidate hypothesis, and then use [model selection](https://en.wikipedia.org/wiki/Model_selection) techniques to choose the most appropriate model.[[2]](https://en.wikipedia.org/wiki/Statistical_hypothesis_testing" \l "cite_note-2)The most common selection techniques are based on either [Akaike information criterion](https://en.wikipedia.org/wiki/Akaike_information_criterion) or [Bayes factor](https://en.wikipedia.org/wiki/Bayes_factor).

Statistical hypothesis testing is sometimes called confirmatory data analysis. It can be contrasted with [exploratory data analysis](https://en.wikipedia.org/wiki/Exploratory_data_analysis), which may not have pre-specified hypotheses.

Regression: - it is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable) and one or more [independent variables](https://en.wikipedia.org/wiki/Independent_variable) (or 'predictors'). More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed. Most commonly, regression analysis estimates the [conditional expectation](https://en.wikipedia.org/wiki/Conditional_expectation) of the dependent variable given the independent variables – that is, the [average value](https://en.wikipedia.org/wiki/Average_value) of the dependent variable when the independent variables are fixed. Less commonly, the focus is on a [quantile](https://en.wikipedia.org/wiki/Quantile), or other [location parameter](https://en.wikipedia.org/wiki/Location_parameter) of the conditional distribution of the dependent variable given the independent variables. In all cases, the estimation target is a [function](https://en.wikipedia.org/wiki/Function_(mathematics)) of the independent variables called the **regression function**. In regression analysis, it is also of interest to characterize the variation of the dependent variable around the regression function which can be described by a [probability distribution](https://en.wikipedia.org/wiki/Probability_distribution).

Regression models involve the following variables:

* The **unknown parameters**, denoted as **β**, which may represent a [scalar](https://en.wikipedia.org/wiki/Scalar_(physics)" \o "Scalar (physics)) or a [vector](https://en.wikipedia.org/wiki/Euclidean_vector" \o "Euclidean vector).
* The **independent variables**, **X**.
* The **dependent variable**, *Y*.
* Conclusion: - the data which we have, we want to see if it is salary is affected by any of the other variables. First we are taking the probability in order to see if those affects the sample by chance. That’s why we are predicting the null hypothesis. As in null hypothesis we say that the apparent effects are not real. By null hypothesis we can the distribution. If it is true we will have the same distribution for both groups. Which will proof that our hypothesis is true. For example, the year ID which is the age of the player how it affects the salary.noreover how being in different team affect the player’s salary. Then we generated the p-value, so we can see if the null, hypothesis is true or not. The intercept of the results, we found that 0.01is the threshold of statistical significance. Which prove that the no significance If p-value is low, the effect is statistically significant, which means that it is unlikely to occur by chance. For null hypothesis, as I explained above there will be no correlation between IgID and year ID . We have followed spearman’s correlation. Where we got the distribution, in order to see the correlation. Also we generated 1000 random sample, then we had the absolute value of mean for 2 groups (year ID,sal ). Then we compared this correlation with the actual correlation. Therefore, if the random correlation is smaller than the actual correlation, then it is statistically significant. Then we had the slope by the linear least square fit, we had plotted the residuals. Also, we found the relation between dependent and independent variables which is the regression. Here are the outputs of the code.