# Regression Analysis 1 - Wages

**[Purpose]:** I want to see how annual wage is related to years of education, years of work experience and age. In my guess:

* More education more wage
* More work experience more wage
* Older age more wage

**[Analysis]:**

1. Define independent variable and dependent variable.

Wage: dependent variable

Years of education, years of work experience and age: independent variables

1. Run the multiple regression model.

Megastat not allow to include “age” in Regression Analysis.

1. So I re-defined the independent variables and dependent variables as bellow:

Wage: dependent variable

Years of education, years of work experience, female and occupation: independent variables.

I think male makes more money than the female.

I think people with management and sales make more money than the others.

1. Re-run the multiple regression model.



* Global test of the model P-Value is less α =.05; it means at least one out of four independent variables is related to dependent variable – wage.
* Occupation P-Value is 0.799, more than α =.05; we need to drop this independent variable.

1. Drop the independent variable - “Occupation”, and run the regression model.



* Global test of the model P-Value is less α =.05; it means at least one out of three independent variables is related to dependent variable – wage.
* All three independent variables’ P-Value is less than α =.05; so each independent variable is related to wage.

1. Interpretation

* Global test of the model P-Value 4.17E-09 is less α =.05; it means at least one out of three independent variables is related to dependent variable – wage.
* Adjusted R2 = 0.332: 33.3% percent of wages (dependent variable) can be explained by years of education, years of work experience, gender (independent variables).
* **For every unit increase** in years of education, **holding** the other two independent variables **constant**, the annual wage **is expected to increase by** $3,160.
* **For every unit increase** in years of work experience, **holding** the other two independent variables **constant**, the annual wage **is expected to increase by** $396.
* **For the female**, **holding** the other two independent variables **constant**, the annual wage **is expected to decrease by** $11,149.

1. Test the assumptions.

* From bellow list, VIF <10, so there is no Multicollinearity problem.



* Residuals by predicted



* Normal Probability plot of residuals



**[Conclusion]:** the annual wage is related to years of education, years of work experience and gender. It is not strong relation though.About 33% percent of wages can be explained by years of education, years of work experience, gender. Years of education contribute more to the annual wage. So if you want to increase your income, it is better to get more education. In a household, it is better to let husband work instead of wife. Because the husband will contribute more income to the family.

# Regression Analysis 2 - International

**[Purpose]:** I want to see how GDP is related to member of G-20, petroleum, Percent of population aged 65 years and over, Literacy and Labor force. In my guess:

* Member of G-20 has more GDP
* Country with petroleum has more GDP
* Higher Percent of population aged 65 years and over, the less GDP
* Higher literacy more GDP
* More labor force more GDP

**[Analysis]:**

1. Define independent variable and dependent variable.

GDP: dependent variable

G-20, petroleum, Percent of population aged 65 years and over, Literacy and Labor force: independent variables

1. Run the multiple regression model.

Since ‘India’ has no labor force data, so I have deleted this data row.



* Global test of the model P-Value is less α =.05; it means at least one out of five independent variables is related to dependent variable – GDP.
* Other than Literacy, all independent variable’s P-Value is more than α =.05; we drop the largest one – G-20.

1. Drop the independent variable - “G-20”, and run the regression model.



* Global test of the model P-Value is less α =.05; it means at least one out of four independent variables is related to dependent variable – GDP.
* Other than Literacy and “65&over”, all independent variable’s P-Value is more than α =.05; we drop the largest one – “Petroleum”.

1. Drop the independent variable - “Petroleum”, and run the regression model.



* Global test of the model P-Value is less α =.05; it means at least one out of three independent variables is related to dependent variable – GDP.
* All independent variable’s P-Value is more than α =.05; we drop the largest p-value one – “labor force”.

1. Drop the independent variable - “labor force”, and run the regression model.



* Global test of the model P-Value is less α =.05; it means at least one out of two independent variables is related to dependent variable – GDP.
* “65 and over” P-Value is more than α =.05; we drop the independent variable – “65 and over”.

1. Drop the independent variable - “65 and over”, and run the regression model.



* Global test of the model P-Value is less α =.05; it means at least one independent variables is related to dependent variable – GDP.
* The independent variable – Literacy’s P-Value is less than α =.05; so only one independent variable - Literacy is related to GDP.

1. Interpretation

* r=0.657, it indicates that GDP is positively strongly dependent on independent variable - Literacy.
* r2=0.432, about 43% GDP can be explained by Literacy.
* Slope=0.479, for every additional literacy percentage, the GDP goes up 479 CAP.

1. Test the assumptions.

* Residuals by predicted



* Normal Probability plot of residuals



**[Conclusion]:** according the analysis based on provided data,the GDP is not related all following variables: member of G-20, petroleum, Percent of population aged 65 years and over, Labor force. It is only related to ‘Literacy’. It means education is just so important for a country development.

# Regression Analysis 3 – Baseball2012

**[Purpose]:** I want to see how Wins is related to Attendance, BA, ERA, HR, Errors and Salary. In my guess:

* More attendance more wins
* More Batting Average, the more wins
* More ERA, more wins
* Less errors, more wins
* More salary more wins

**[Analysis]:**

1. Define independent variable and dependent variable.

Wins: dependent variable

Attendance, BA, ERA, HR, Errors and Salary: independent variables

1. Run the multiple regression model.



* Global test of the model P-Value is less α =.05; it means at least one out of six independent variables is related to dependent variable – Wins.
* Other than BA and ERA, all independent variable’s P-Value is more than α =.05; we drop the largest p-value one – Attendance.

1. Drop the independent variable - “Attendance”, and run the regression model.



* Global test of the model P-Value is less α =.05; it means at least one out of five independent variables is related to dependent variable – Wins.
* Other than BA and ERA, all independent variable’s P-Value is more than α =.05; we drop the largest p-value one – “Errors”.

1. Drop the independent variable - “Errors”, and run the regression model.



* Global test of the model P-Value is less α =.05; it means at least one out of four independent variables is related to dependent variable – Wins.
* Other than BA and ERA, all independent variable’s P-Value is more than α =.05; we drop the largest p-value one – “HR”.

1. Drop the independent variable - “HR”, and run the regression model.



* Global test of the model P-Value is less α =.05; it means at least one out of three independent variables is related to dependent variable – Wins.
* The independent variable - Salary’s P-Value is more than α =.05; we drop this variable.

1. Drop the independent variable - “Salary”, and run the regression model.



* Global test of the model P-Value is less α =.05; it means at least one independent variables is related to dependent variable – Wins.
* The independent variables – BA and ERA’s P-Value is less than α =.05; so these independent variables are related to Wins.

1. Interpretation

* Global test of the model P-Value 6.92E-08 is less α =.05; it means at least one out of two independent variables is related to dependent variable – Wins.
* Adjusted R2 = 0.683: 68.3% percent of wins (dependent variable) can be explained by BA and ERA (independent variables).
* **For every unit increase** in team BA, **holding** the other independent variable - ERA **constant**, the team wins **are expected to increase by** 305.
* **For every unit increase** in team ERA, **holding** the other independent variable - BA **constant**, the team wins are **expected to decrease by** 20.

1. Test the assumptions.

* From bellow list, VIF <10, so there is no Multicollinearity problem.



* Residuals by predicted



* Normal Probability plot of residuals



**[Conclusion]:** according to the analysis based on provided Baseball 2012,the team wins are not related all following variables: Attendance, HR, Errors and Salary. Instead, the team wins are related to BA (Batting Average) and ERA (team earned run average). The team BA contributes more to team wins positively. But team ERA is negatively related to team wins.