Infant Weight Prediction and Analysis

**INTRODUCTION**

The purpose of this study is to determine whether an accurate prediction of birth weight is possible using predictor variables: Race, pregnant woman’s participation in a food education program, age of the newborn from the first day of the last menstruation, mother's age, and sex of the newborn. Multivariate linear regressions are deployed to determine which of these variables are significant for prediction. The model is then used to predict the weight of babies at birth. The efficacy of the prediction models is assessed and compared on basis of errors in the predicted weights. The resulting model can explain 55% of weight variation in data. We also check for the difference in medians (Due to non-normalcy of data) for dependent variable weight between males and females and also for the difference between participation in the food education program. R computation platform was used for modeling and computation.

Keywords: Multiple Linear Regression, Linear regression, Wilcoxon Rank sum Test

**ABOUT R SOFTWARE**

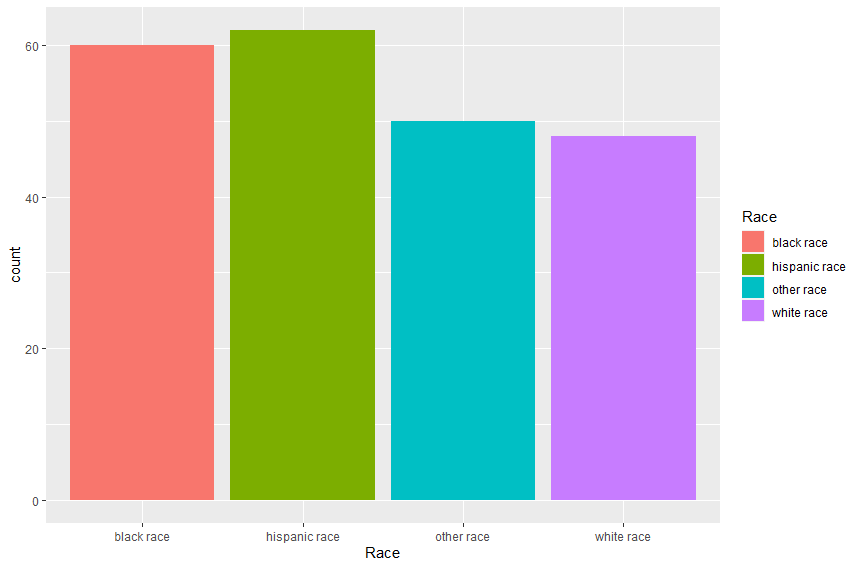
**R** is a programming language and free software developed by Ross Ihaka and Robert Gentleman in 1993. R possesses an extensive catalog of statistical and graphical methods. It includes machine learning algorithms, linear regression, time series, and statistical inference to name a few. Most of the R libraries are written in R, but for heavy computational tasks, C, C++, and Fortran codes are preferred.

R is not only entrusted by academics, but many large companies also use R programming language, including Uber, Google, Airbnb, Facebook, and so on.

Data analysis with R is done in a series of steps; programming, transforming, discovering, modeling, and communicating the results.

**BASIC ANALYSIS**

**Bar plot for Race**

****

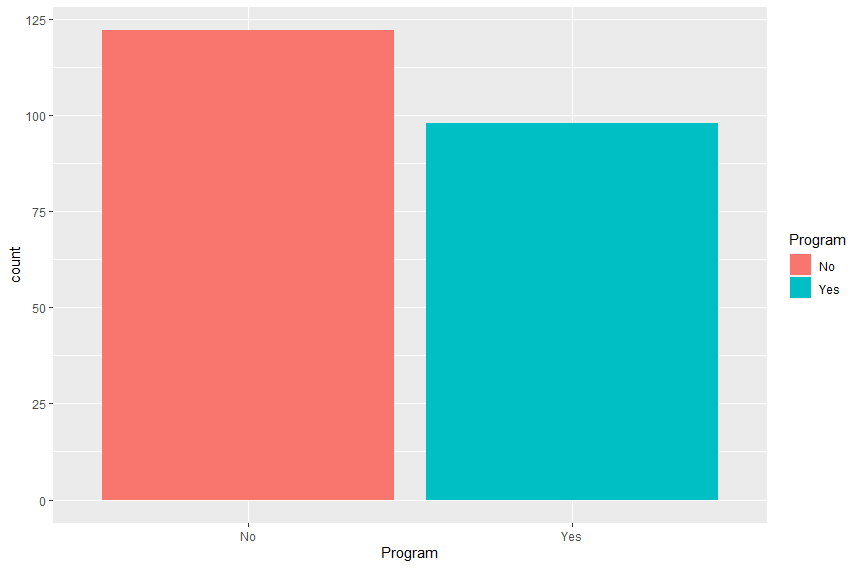
48

50

62

60

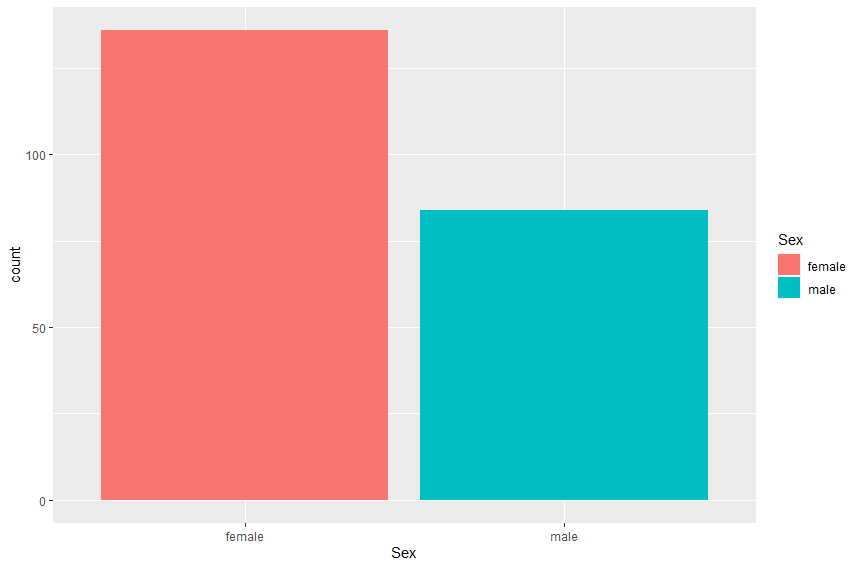
**Bar plot for program**

****

98

122

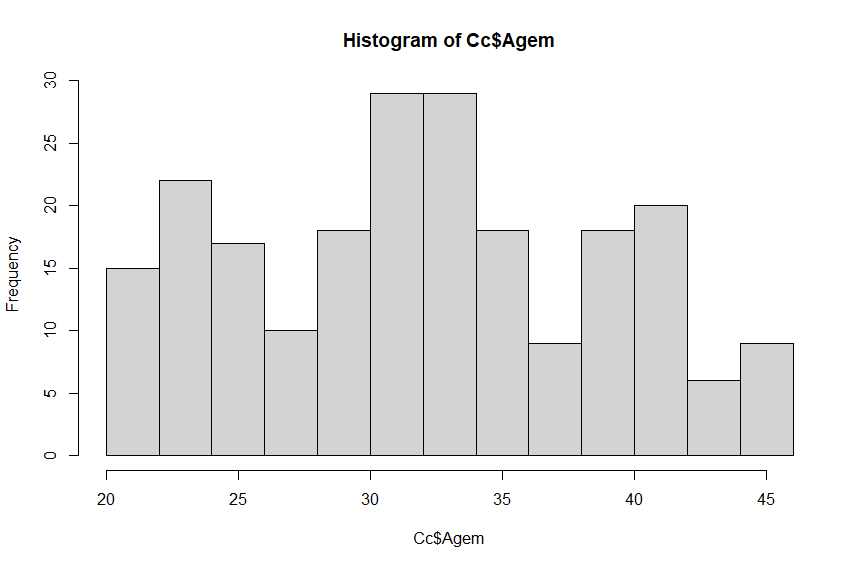
**Bar plot for Sex of**

****

84

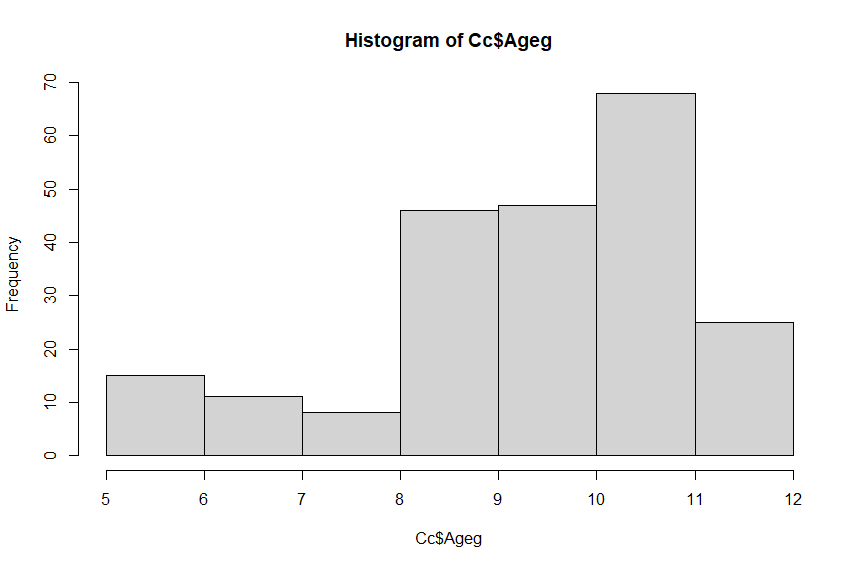
136

**Histogram for Mother’s Age**

****

Mother’s Age

**Histogram for gestational Age**



Gestational Age

**SIMPLE LINEAR REGRESSION**

We use a simple linear regression model to explain the value of the variable "Weight" as a function of the variable gestational age (Ageg).

Regression analysis is a very widely used statistical tool to establish a relationship model between two variables. One of these variables is called the predictor variable, whose value is gathered through experiments. The other variable is called the response variable whose value is derived from the predictor variable. In Linear Regression these two variables are related through an equation, where the exponent (power) of both these variables is 1. Mathematically a linear relationship represents a straight line when plotted as a graph. The general mathematical equation for linear regression is

Y=βx+c, where, Y is the response variable, x is the predictor variable. β and c are constants which are called the coefficients. The error term is also known as residual. The "Residual" term represents the deviations of the observed values y from their means, which are normally distributed.

On running the linear regression model, we get the following results,

**Residuals**:

The residuals are the difference between the actual values and the predicted values. We can generate these same values by taking the actual values of infant weight and subtracting them from the predicted values of the model. The following table shows the descriptive statistics for the residual values.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Min | 1Q | Median | 3Q | Max |
| -2083.17 | -287.52 | 9.67 | 339.32 | 1476.78 |

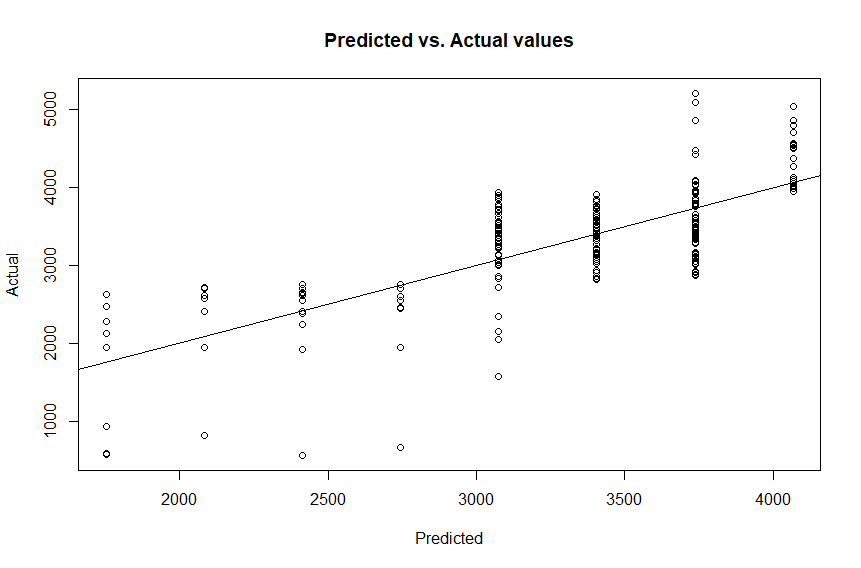
**Coefficients:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| Intercept | 98.68 | 201.24 | 0.49 | 0.624 |
| Ageg | 330.69 | 20.24 | 16.34 | <2e-16 |

Based on the above table, our linear regression model is,

**Y = 98.68 + 330.69\*X**, where Y is the response variable and X is the predictor variable.

The p-value is calculated using the t-statistic from the T distribution. The p-value, in association with the t-statistic, helps us to understand how significant our coefficient is to the model. In practice, any p-value below 0.05 is usually deemed as significant. When we see the p-value we see that the intercept is **not statistically significant** in predicting the weight, however, the Ageg (gestational age) is **statistically significant** in predicting the infant’s weight.



**Residual standard error:**

The residual standard error is a measure of how well the model fits the data.

Here we have Residual standard error **= 515 on 218 degrees of freedom**. If we look at the least-squares regression line, we notice that the line doesn’t perfectly flow through each of the points and that there is a “residual” between the point and the line. The residual standard error tells us the **average** amount that the actual values of Y (the dots) differ from the predictions (the line) in units of Y. In general, we want the smallest residual standard error possible, because that means our model’s prediction line is very close to the actual values, on average.

**Multiple R-squared: 0.5505, Adjusted R-squared: 0.5485**

The Multiple R-squared value is most often used for simple linear regression (one predictor). It tells us what percentage of the variation within our dependent variable the independent variable is explaining. In other words, it’s another method to determine how well our model is fitting the data. Here the R-squared value explains 55.05% of the variation within infant weight, our dependent variable.

The Adjusted R-squared value shows what percentage of the variation within our dependent variable that all predictors are explaining. Here it explains 54.85% of the variation.

**F-statistic: 267 on 1 and 218 DF, p-value: < 2.2e-16**

When running a regression model, either simple or multiple, a hypothesis test is being run on the global model. The null hypothesis is that there is no relationship between the dependent variable and the independent variable and the alternative hypothesis is that there is a relationship.  The F-statistic and overall p-value help us determine the result of this test.

We can see from our model, that the F-statistic is very large and our p-value is so small it is zero. **This would lead us to reject the null hypothesis and conclude that there is strong evidence that a relationship does exist between infant weight and gestational age.**

**LINEAR CORRELATION COEFFICIENT BETWEEN AGEG AND INFANT WEIGHT**

Correlation is a statistical method to measure the relationship between two quantitative variables. The positive values of *r* indicate the positive relationship and vice versa. The higher the absolute value of *r*, the stronger the correlation. If the value of *r* is 0, it indicates that there is no relationship between the two variables.

Here we can see that the correlation coefficient (r) is **0.74199** which means that there is a high positive correlation between gestational age and weight.

**DIFFERENCE IN MEANS FOR THE DEPENDENT VARIABLE WEIGHT BETWEEN MALE AND FEMALE**

**Wilcoxon Rank-Sum Test (Mann Whitney U):**

In statistics, the T-test is one of the most common tests which is used to determine whether the mean of the two groups is equal to each other. However, the assumption for the test is that both groups are sampled from a normal distribution with equal fluctuation.

Now here, if we conduct a Shapiro-test of normality we see that,

* **Data: Male**

W = 0.9464, p-value = 0.001596

* **Data: Female**

W = 0.92203, p-value = 8.56e-07

In both cases, the p-value is significantly smaller than 0.05 which indicates that the groups are **not normally distributed**.

Therefore, we have to use **Wilcoxon rank-sum test (Mann-Whitney U test)** which is a general test to compare two distributions in independent samples. It is a commonly used alternative to the two-sample t-test when the assumptions are not met.

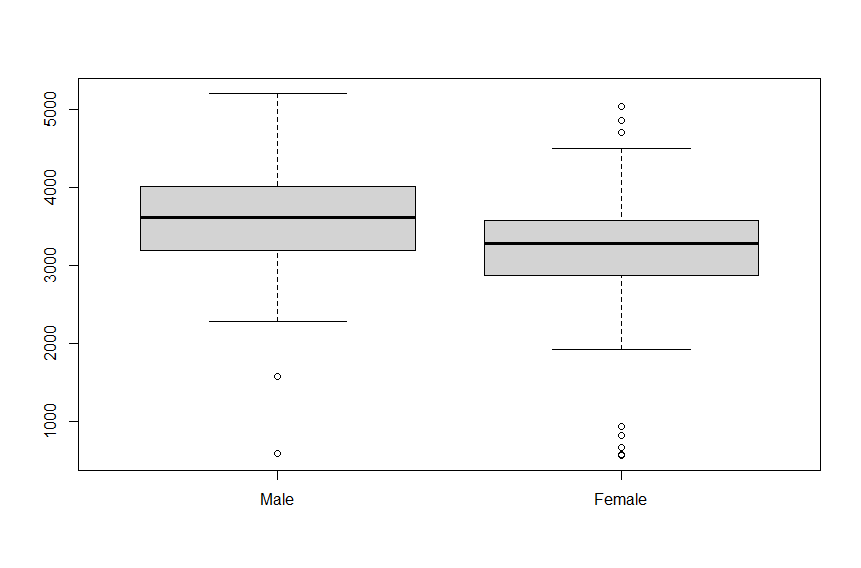
We are interested in testing the following hypothesis,

Null Hypothesis (H0)= Median weight of male infant = Median weight of female infant

Alternative Hypothesis (H1): Median weight of male infant ≠ Median weight of female infant

On conducting a **Wilcoxon rank-sum test** at the level of significance α = 0.05, we see that,

W = 2469, p-value = 0.002302, which concludes that the alternative hypothesis is true, i.e., location shift is not equal to 0. Hence the median weight of male infant and median weight of female infant has a significant difference.

**DIFFERENCE IN MEANS FOR THE DEPENDENT VARIABLE WEIGHT BETWEEN PARTICIPATION OR NOT IN FOOD PROGRAM**

Infant weight

On conducting a Shapiro-test of normality we see that,

* **Data: Program taken**

W = 0.92288, p-value = 2.405e-05

* **Data: Program not taken**

W = 0.93832, p-value = 2.851e-05

In both cases, the p-value is significantly smaller than 0.05 which indicates that the groups are **not normally distributed**.

Therefore, we have to use **Wilcoxon rank-sum test (Mann-Whitney U test)** which is a general test to compare two distributions in independent samples. It is a commonly used alternative to the two-sample t-test when the assumptions are not met.

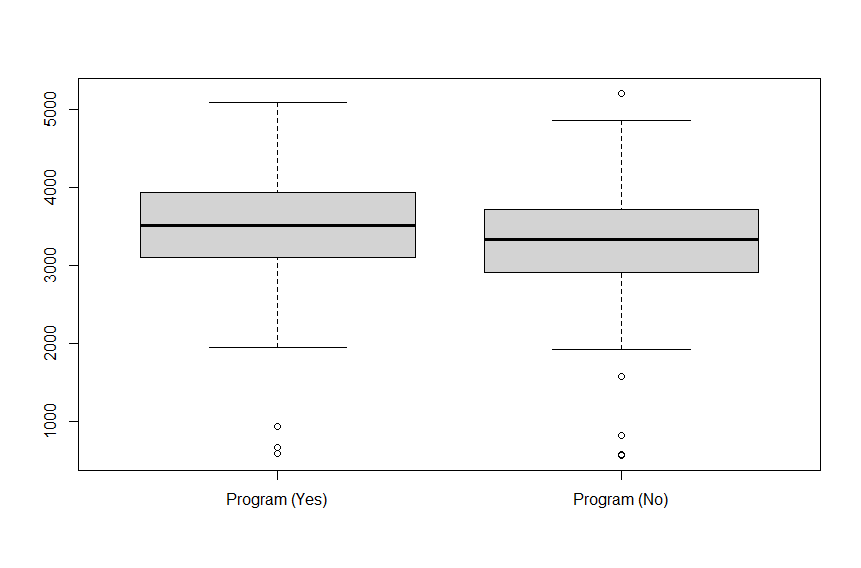
We are interested in testing the following hypothesis,

Null Hypothesis (H0): Median weight of infants whose mothers had taken part in the food program = Median weight of infants whose mothers had not taken part in the food program

Alternative Hypothesis (H1): Median weight of infants whose mothers had taken part in the food program ≠ Median weight of infants whose mothers had not taken part in the food program

On conducting a **Wilcoxon rank-sum test** at the level of significance α = 0.05, we see that,

W = 3059, p-value = 0.02489, which concludes that the alternative hypothesis is true, i.e., location shift is not equal to 0. Hence the median weight of infants whose mothers had taken part in the food program and median weight of infants whose mothers had not taken part in the food program are significantly different.



Infant weight

**MULTIPLE LINEAR REGRESSION**

In the case of P independent variables, for predicting the dependent variable, the population regression line for p explanatory variables x1, x2, ..., xp can be defined as:

Actual data = Fitted model+ Error, where the fitted model is represented as:

Y= c+β1x1+ β2x2+ β3x3+ β4x4+………………..+ βpxp

Let us now interpret the Multiple Linear regression model,

**Residuals**:

The following table shows the descriptive statistics for the residual values.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Min | 1Q | Median | 3Q | Max |
| -2107.08 | -200.62 | 0.86 | 279.62 | 1172.46 |

**Coefficients**:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| Intercept | 1550.645 | 321.666 | 4.821 | 2.71e-06 |
| Ageg | 271.826 | 20.483 | 13.271 | <2e-16 |
| Agem | -31.032 | 5.208 | -5.959 | 1.03e-08 |
| Race | 15.711 | 29.908 | 0.525 | 0.59992 |
| Program | 50.797 | 64.630 | 0.786 | 0.43276 |
| Sex | 182.137 | 66.169 | 2.753 | 0.00642 |

Based on the above table, our linear regression model is,

**Y = 1550.645 + 271.826\*X1 + (-32.032) \*X2 + 15.711\*X3 + 50.797\*X4 + 182.137\*X5,** where Y is the response variable and X1, X2, X3, X4 and X5 are the predictor variable.

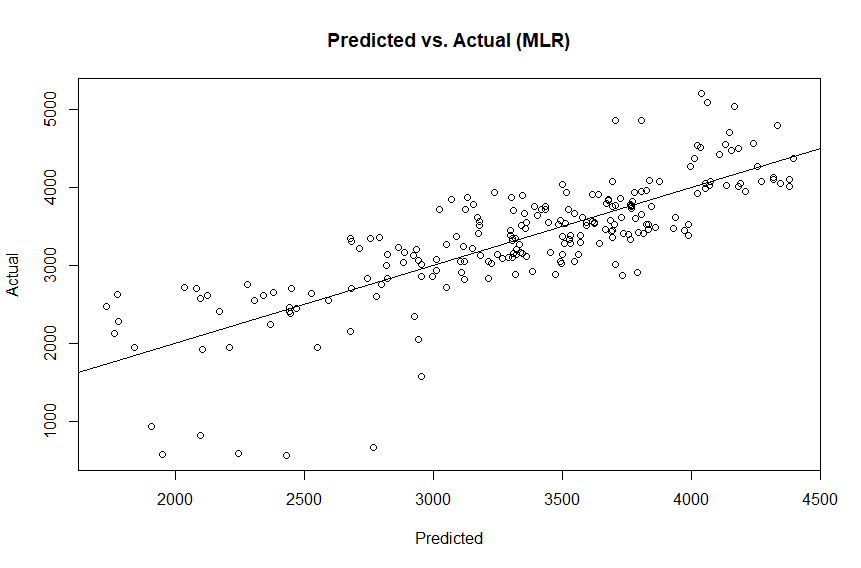
X1 = Gestational Age (**statistically significant** in predicting the infant’s weight)

X2 = Mother’s Age (**statistically significant** in predicting the infant’s weight)

X3 = Race (**not statistically significant** in predicting the weight)

X4 = Participation in the food education Program (**not statistically significant**)

X5 = Sex (**statistically significant** in predicting the infant’s weight)



We see that the **Residual standard error: is 468.3 on 214 degrees of freedom**

**Multiple R-squared: 0.6353, Adjusted R-squared: 0.6267**

Here the R-squared value explains 55.05% of the variation within infant weight, our dependent variable. The Adjusted R-squared value explains 54.85% of the variation

**F-statistic: 74.54 on 5 and 214 DF, p-value: < 2.2e-16**

We can see from our model, that the F-statistic is very large and our p-value is so small it is zero. **This would lead us to reject the null hypothesis and conclude that there is strong evidence that a relationship does exist between infant weight and predictor variables X1, X2, X3, X4, and X5.**

**CONCLUSION:**

The objective of the current study was to propose a simple and efficient mathematical model for the prediction of infant weight, based on 220 observations. Race, pregnant woman’s participation in a food education program, age of the newborn from the first day of the last menstruation, mother's age, and sex of the newborn were used as predictor variables to estimate the delivery weight of the babies. A multi-linear regression model was employed to assess the impact of several predictors.

The food education program does not prove to be a significant predictor for the multiple linear regression model as the p-value is greater than 0.05, however, the difference in medians of infant weight for the program takers and non-takers were significantly different which was proved by the **Wilcoxon Rank Sum Test**. This shows that even though there was a significant difference in the medians of infant weights, the part taking in the food education program does not influence the regression model. This can also be shown by the correlation coefficient between infant weight and program which comes out to be 0.1447664 which is very small.

The limitations of the study are as follows:

1. Sample size is extremely small. We should use a larger data set if we want better predictions
2. Minimize categorical variables and use more numerical variables. In this data set, we use 3 categorical variables, which is more than enough given that we only have 5 predictor variables.