MATH40031: Statistical Data Analysis and Visualisation

Submitted by: Gokila Sundaram (N1082709)

Contents

Introduction:

This report will explain the statistical analysis on a given dataset (“diabetes”) in R studio. Dataset has multiple predictor variables and one response variable. We will run different set of statistical tests to predict the relationship between variables and the factors which will influence diabetes. The goal of this report is to predict the binary outcome variable "Outcome" (whether a patient has diabetes or not) based on the other predictor variables.

a) Exploring the given dataset is the first step in Statistical Analysis.

The diabetes dataset has 8 predictor variables and one response variable (Outcome). Out of 8 Predictor variables, following variables (5) have missing values, represented by 0. Although Pregnancy variable has 0 values, those observations are valid data.

Total Number of Observations: 768
Number of Observations without missing values or complete cases: 392
Number of Observations with missing values or incomplete cases: 376
Number of Variables with Missing Data: 5

% of Missing Values in Insulin: 48%
% of Missing Values in Glucose: 0.65%
% of Missing Values in BMI:1.4%
% of Missing Values in BloodPressure:4.5%
% of Missing Values in Skin Thickness:30%

All missing values in the dataset is represented by zero. Overall 50% of data is missing in the given dataset, this will have significant impact on the accuracy of the statistical test results. Majority of values are missing from Insulin and Skin thickness. Whereas Glucose, BMI and Blood pressure has small number of values missing. Observed that the Pattern of missing data is random. As the percentage of missing data is statistically significant for Insulin and SkinThickness, ignoring missing data can lead to biased estimates of population parameters, such as means and variances, and inaccurate P values, and this is not generally recommended in statistics. It can make the study's analysis more difficult. And impacts reliability of the trials and produce false results. Mice Package is used in R Studio to understand the missing patterns of the dataset. It is feasible to impute missing values with mean or median to proceed further analysis. But insulin and skin thickness have large number of missing data, imputation with mean or median will impact the correlation between variables. Hence, we need analyse the dataset & understand what kind of statistical analysis we need to perform on the dataset. Accordingly, handling the missing data will provide valid results.

b) Cleaning the given dataset is the Second Step.

There are multiple methods available in statistics to clean the data. First, replace the missing values (zero) with NaN in order to find the central tendency of the dataset. Replace missing values in the respective variables one by one using the following syntax. Installed dplyr, tidyverse, tidyr packages to clean and filter data in R.

```
diabetes$Glucose[diabetes$Glucose==0] <- NaN
diabetes$BloodPressure[diabetes$BloodPressure==0] <- NaN
diabetes$BMI[diabetes$BMI==0] <- NaN
diabetes$Insulin[diabetes$Insulin==0] <- NaN
diabetes$SkinThickness[diabetes$SkinThickness==0] <- NaN
```

Result: All Missing Values in the dataset are replaced by NaN. Zero values from Pregnancy and Outcome variables are retained as is.

To understand the pattern of missing data:

```
install.packages("mice")
library(mice)
missing_pattern<- md.pattern(diabetes,rotate.names = TRUE)
View(missing_pattern)
```

Each row of the table represents a unique combination of missing values across the variables in the dataset. The columns of the table show the number of observations with that missing data pattern. Missing Pattern is random, there is no specific pattern identified on the missing values.

Figure 1. Missing Data Pattern

Identify duplicates in the dataset before moving to the next step.

```
Duplicates<- diabetes[duplicated(diabetes)]
View(Duplicates)
```

Result: There are no duplicates in the dataset.

c) Visualizing the given dataset is the next step in statistical Analysis.

Visualizing the dataset will help us to predict the distribution of dataset. There are multiple options available in R to visualize the distributions.

1.QQ Plots for each variable is generated from R studio.

Figure: QQ Plots

Analysis:

Variable	**QQ Plot**	**Confidence on Normality**
Pregnancies	Non-Linear Scatter Plot, Smallest values are not passing through the line	Low
Glucose	Values in the middle passing through the line, min and max values are away from the line.	Medium
BloodPressure	Most of the values passing through Linear plot	High
SkinThickness	Most of the values passing through Linear plot	High
Age	Non-Linear Scatter Plot	Low
Insulin	Non-Linear Scatter Plot	Low
BMI	Most of the values passing through Linear plot	High
DiabetesPedigreeFunction	Non-Linear Scatter Plot	Low

2.Boxplots for each predictor variable produced from R studio.

Figure: Boxplots

Variable	Boxplot
Pregnancies	Observed Outliers, Datapoints increasing in UpperWhisker. No Datapoints in Lower Whisker
Glucose	No Extremal values, 50% Data points are within Horizontal Bar
BloodPressure	Observed Outliers, Lower & Upper whisker length looks equal
SkinThickness	Observed Upper Whiskers, Lower & Upper whisker length looks equal
Age	Observed Outliers, Length of Upper Whisker is more
Insulin	Observed significant number of Outliers
BMI	Observed Outliers, Lower & Upper whisker length looks equal
DiabetesPedigreeFunction	Observed significant number of Outliers

Dataset has more outliers, hence it's difficult to determine the normality using plots.

3. Histogram for each predictor Variable from R Studio:
Figure : Histograms

Result: Bloodpressure & BMI has symmetric distribution. Age & Pregnancies, Insulin Variables are Right Skewed. Glucose & Skin thickness doesn't have proper bell-shaped curve. And DiabetesPedigreeFunction has longer tail. Its difficult to conclude that BMI and Bloodpressure is normally distributed as Histograms are sensitive to outliers.

4. **Kolmogorov Smirnov** Test: As the data set has more outliers and missing values, it's appropriate to run test to confirm the normality of predictor Variables. While running this test, NA values are excluded, by default.
Null hypothesis is that data is normally distributed. Except BMI all other variables return p value less than 0.05. P value for BMI is 0.05344. Only BMI is normally distributed in the given dataset.

Summary Statistics & Conclusion: Installed **psych Package** and used **describe** function to return the results in R Studio(Min, Max , Range, Skew..). Also used summary function to view count of NA values in each variable. There is evidence to say that only BMI is distributed normally. There is no evidence to prove that all other predictor variables are normally distributed.

Summary Statistics
Describe Statistics

d) Assuming all predictor variables are independent, will run statistical tests to assess differences in central tendencies with respect to Outcome Variable.

Visualization of differences in Central Tendencies

Executed statistical tests on 2 datasets, One dataset with MissingValues and another dataset with Median Imputed Values.

Missing Value Dataset 1:
Created 2 sample groups filtered by Outcome Variable. (Both groups have missing values represented by NaN).

Sample Groups: True_Diabetes: (People with Diabetes) & False_Diabetes: (People without Diabetes)

Group Name	Number of Observations	Outcome	Missing Data
True_Diabetes	268	1	None
False_Diabetes	500	0	None

Median Imputed dataset 2:

Sample Groups: ImputedData_true (People with Diabetes) & ImputedData_False(People without Diabetes)

Group Name	Number of Observations	Outcome	Missing Data
ImputedData_true	268	1	Imputed by Median
ImputedData_False	500	0	Imputed by Median

As we failed to prove the normality of the data using KS test & considering outliers in the dataset, (t-test is sensitive to outliers) Selected Mann-Whitney U test to identify the differences in the central tendencies (Medians) of sample groups, (as it makes no assumptions about distributions). By default, missing values will be excluded by Mann Whitney test.

Null Hypothesis: There is no difference in the levels (Median) of predictor variables between people with diabetes and people without diabetes.
Alternative Hypothesis(Two Sided Test): There is a significant difference in the levels (Median) of predictor variables between people with diabetes and people without diabetes.
Alternative Hypothesis(One Sided Test): Ranksum Value or Median Value of people without diabetes is greater than median value of people with diabetes.

Variables	Statistical Test	P Value (Sample Groups with Missing Data)
Glucose	Mann Whitney U test (Two sided)	< 2.2e-16
	Mann Whitney U test (One sided)	< 2.2e-16
BloodPressure	Mann Whitney U test (Two sided)	p-value = 1.629e-06
	Mann Whitney U test (One sided)	p-value = 8.143e-07
SkinThickness	Mann Whitney U test (Two sided)	p-value = 2.32e-09
	Mann Whitney U test (One sided)	p-value = 3.481e-10
BMI	Mann Whitney U test (Two sided)	< 2.2e-16
	Mann Whitney U test (One sided)	< 2.2e-16
Age	Mann Whitney U test (Two sided)	< 2.2e-16
	Mann Whitney U test (One sided)	< 2.2e-16
DiabetesPedigreeFunction	Mann Whitney U test (Two sided)	p-value = 1.197e-06
	Mann Whitney U test (One sided)	p-value = 5.983e-07
Insulin	Mann Whitney U test (Two sided)	p-value = 7.477e-14
	Mann Whitney U test (One sided)	p-value = 1.48e-14
Pregnancies	Mann Whitney U test (Two sided)	p-value = 3.745e-08
		p value=1.873e-08

Conclusion: With reference to the table above, P value is < 0.05, There is evidence against null hypothesis. Ranksum or Median difference of the sample groups are not equal to 0, there is a significant difference between central tendencies of sample groups. Median Value of people without diabetes is greater than median value of people with diabetes.

Observed that P value is <0.05 for the sample groups with imputed values. Hence running Mann Whitney test with Missing Values or Median Imputed Values doesn't affect the end results. Though the P value differs between these 2 sets of sample groups, the end result is same.

Effect Size: As the p-value is too small, verified the effect size of the differences in central tendency. Installed effect size package and used cohen.d function. (Refer R script for the results)
This returns a value which is **unbiased to sample sizes**, providing a **standardised way to assess distribution differences**. (95 Percent Confidence Interval)

Verified the effect size of sample groups with Missing Value & sample groups with Median imputed values. It shows that there is a significant difference in the effect size of Insulin. Effect size is Small for Insulin when the missing values are imputed with Medians. Median Imputation for Insulin produces biased results. Median imputation when large number of values are missing, impacts the accuracy of results.

Considering effect sizes & Mann Whitney test results, could say that predictor variables have an influence on the Diabetes Outcome. It's evident that difference in Glucose level have more influence on the diabetes Outcome.

Effect Size	**Estimate (Samples with Missing Values)**
Pregnancies	Medium
Glucose	Large
BloodPressure	Small
SkinThickness	Medium
Age	Medium
Insulin	Medium
BMI	Medium
DiabetesPedigreeFunction	Small

e) Correlation Coefficients test will help to understand the linear association between variables in a dataset:

Null Hypothesis: There is no linear relationship between variables in the given dataset. All predictor variables are independent.

Alternative Hypothesis: There is a linear relationship between predictor variables in the given dataset.

Excluded Outcome variable from diabetes dataset, Used cor() function to view the summary of correlation matrix between predictor variables. Used Plot() function to visualize the correlation between predictor variables. Used Corrplot to visualize the r values.

Spearman Correlation Test: As the normality test (kolmogorov smirnov test) witnessed p value <0.05, performed the non-parametric Correlation test.
Figure below shows Correlation Values on diabetes dataset excluding the Missing Values (392 Observations).

```
> corvalues
                          Pregnancies    Glucose BloodPressure SkinThickness   Insulin
Pregnancies                1.00000000 0.19048148    0.15241404    0.05475868 0.1231537
Glucose                    0.19048148 1.00000000    0.23660931    0.21583824 0.6589582
BloodPressure              0.15241404 0.23660931    1.00000000    0.25010618 0.1316389
SkinThickness              0.05475868 0.21583824    0.25010618    1.00000000 0.2411450
Insulin                    0.12315371 0.65895822    0.13163895    0.24114499 1.0000000
BMI                       -0.06555144 0.19907115    0.31742747    0.67439293 0.3009061
DiabetesPedigreeFunction   0.01171335 0.08934581   -0.02089194    0.09304947 0.1315153
Age                        0.63365655 0.35030470    0.32924408    0.24150672 0.2607474
                                 BMI DiabetesPedigreeFunction       Age
Pregnancies               -0.06555144               0.01171335 0.6336566
Glucose                    0.19907115               0.08934581 0.3503047
BloodPressure              0.31742747              -0.02089194 0.3292441
SkinThickness              0.67439293               0.09304947 0.2415067
Insulin                    0.30090608               0.13151530 0.2607474
BMI                        1.00000000               0.09624242 0.1669629
DiabetesPedigreeFunction   0.09624242               1.00000000 0.1029841
Age                        0.16696290               0.10298412 1.0000000
```

Click the link to view the Corrplot of r Values

With reference to the r Values, following variables have a High positive correlation.
In this case, Null hypothesis will be rejected, the values of below listed variables tend to increase together or decrease together in the same direction, but not as strongly as in the case of a perfect positive correlation. Hence, we can predict the values of one variable given the values of correlated variables.

- Age & Pregnancies : r Value: 0.63
- Glucose & Insulin: r Value : 0.66
- SkinThickness&BMI: r Value: 0.67

Below listed variables have medium positive correlation as the r value is greater than 0.30.

- Age & Glucose
- Age & BloodPressure
- BMI & BloodPressure

And there are few more variable the correlation value is greater than 0.20. (Small Positive Correlation)

- Glucose & BloodPressure,
- Glucose & Skinthickness,
- Bloodpressure & Skinthickness
- Insulin & Age
- Insulin & Skinthickness

P value is less than 0.05 for the correlated variables detailed above, We fail to prove Null Hypothesis.
P value is greater than 0.05 for cor test of DiabetedPedigreeFunction with Glucose, Skinthickness, Insulin and Age. We fail to Reject Null Hypothesis in this scenario.

Conclusion:

DiabetesPedigreeFunction is the only variable that has r value close to zero with most of the variable. And it has very weak relationship with Age & Insulin (0.1). There is a high probability of DiabetesPedigreeFunction to be an Independent Variable to predict Diabetes in Patients.
All other variables are correlated in the given dataset with strength of Correlation detailed earlier.
Above conclusion is based on 2 approaches,
Statistical analysis on sample population with no Missing Values (392 Observations).

Statistical analysis on the entire dataset (768 Observations), where missing values are Imputed using Regression Model. Both has given the same end results. Hence its evident that Regression Model imputation doesn't produce any adverse effects on the analysis.

Figure below shows Correlation Values of Imputed dataset.

```
> cor(regressionimputed_diabetes)
                          Pregnancies   Glucose BloodPressure SkinThickness    Insulin        BMI
Pregnancies                1.00000000 0.1309501   0.214620658     0.1071761 0.04912678 0.02430192
Glucose                    0.13095008 1.0000000   0.230760747     0.2432900 0.70043569 0.23703359
BloodPressure              0.21462066 0.2307607   1.000000000     0.2425273 0.14182082 0.29738677
SkinThickness              0.10717606 0.2432900   0.242527297     1.0000000 0.25192972 0.70825427
Insulin                    0.04912678 0.7004357   0.141820822     0.2519297 1.00000000 0.28010474
BMI                        0.02430192 0.2370336   0.297386770     0.7082543 0.28010474 1.00000000
DiabetesPedigreeFunction -0.03352267 0.1389691   0.001797845     0.1236115 0.15258211 0.15256820
Age                        0.54434123 0.2697443   0.334588613     0.1505202 0.23222009 0.02850615
Outcome                    0.22189815 0.4966410   0.174306122     0.2818919 0.34918122 0.31580231
                          DiabetesPedigreeFunction        Age   Outcome
Pregnancies                           -0.033522673 0.54434123 0.2218982
Glucose                                0.138969130 0.26974429 0.4966410
BloodPressure                          0.001797845 0.33458861 0.1743061
SkinThickness                          0.123611541 0.15052017 0.2818919
Insulin                                0.152582106 0.23222009 0.3491812
BMI                                    0.152568204 0.02850615 0.3158023
DiabetesPedigreeFunction               1.000000000 0.03356131 0.1738441
Age                                    0.033561312 1.00000000 0.2383560
Outcome                                0.173844066 0.23835598 1.0000000
>
```

f) To identify the predictor variables which have influence on Outcome variable, applied 2 methods.

As its evident that predictor variables are positively correlated, (High Correlation between Glucose & Insulin, Skinthickness & BMI. Refer screenshot in the previous section) fitted regression model to predict the influence on the outcome variable.
While fitting the model we should not sum up all correlated variables to predict the response variable, as it will produce Collinearity problems (due to correlation), and will not produce accurate P values or R^2 Values. (Author: Enders, Felicity Boyd, Year:2013).

Hence performed regression model analysis on each predictor variable against Outcome.

1. Linear Regression Model:
   Example: summary(lm(Outcome ~ Glucose. , data= diabetes))

2. Logistic regression Model:
   Example:model <- glm(Outcome ~ Glucose , family = binomial(link = "logit"), data = diabetes)

Null Hypothesis: there is no statistically significant relationship between the predictor variables and the Outcome variable.(Coefficient is equal to 0)
Alternative Hypothesis: there *is* a statistically significant relationship between predictor and response variable. (Coefficient is not equal to 0)

Followed 2 approaches:
Approach 1: Applied LM and GLM (Logistic Regression) models on Dataset with NaN Values.
Approach 2: Installed Mice Package and Imputed all missing values using Regression (norm.predict) method. And then applied LM & GLM Model on the imputed dataset.

Model	Dataset	P Values
Linear Regression Model	Data with Missing (NaN) Values	Glucose: P value <2e-16 BMI: P value: < 2e-16 DiabetesPedigreeFunction:P Value 1.25e-06 Pregnancies: P Value 5.07e-10 Age: p-value: 2.21e-11 SkinThickness: 8.96e-10 Insulin: 7.75e-10 BloodPressure:3.41e-06
Logistic Regression Model	Data with Missing (NaN) Values	Glucose: P value <2e-16 BMI: P value: 4.31e-16 DiabetesPedigreeFunction: P Value 3.7e-06 Pregnancies: P Value 2.15e-09 Insulin: 7.17e-08 SkinThickness: 8.02e-09 BloodPressure: 5.72e-06
Linear Regression Model	Imputed Data (Mice Imputation)	Pregnancies: 2.15e-09 Glucose: < 2e-16 BMI: P value: <2e-16 DiabetesPedigreeFunction :P Value 3.7e-06 Age: P value
Logistic Regression Model	Imputed Data (Mice Imputation)	Pregnancies: 5.07e-10 Glucose: < 2e-16 BMI: P value: <2e-16 DiabetesPedigreeFunction :P Value 1.25e-06 Age: P value 1.77e-10 Insulin: <2e-16 SkinThickness: 8.97e-14 BloodPressure: 2.15e-06

Click on link: Visualization on the linear regression models:
Visualisation on GLM Model:

P Values are too small, it could be due to the sample size. We fail to prove the NULL hypothesis. The Coefficient Values are greater than 0. Hence there is significant relationship between predictor and response variables.

Linear Regression Model shows R^2 values for each predictor variable as follows:

Glucose: Multiple R-squared: 0.2467(25%), Adjusted R-squared: 0.2457
Pregnancies: Multiple R-squared: 0.04924 (5%), Adjusted R-squared: 0.048
BMI: Multiple R-squared: 0.09973 (10%), Adjusted R-squared: 0.09856
DiabetesPedigreeFunction: Multiple R-squared: 0.03022 (3%), Adjusted R-squared: 0.02896
SkinThickness: Multiple R-squared: 0.06734 (6%), Adjusted R-squared: 0.06561
Age: Multiple R-squared: 0.05681, (5%) Adjusted R-squared: 0.05558
Insulin: Multiple R-squared: 0.1219, (12%) Adjusted R-squared: 0.1208

BloodPressure: Multiple R-squared: 0.03038, (3%) Adjusted R-squared: 0.02912

Overall, these values suggest that the model explains around 69% of the variance in the dependent variable, However the percentage of proportion of variance is not completely depends on Multiple R-squared values, it also depends on Residual Error values.

Logistic Regression Model:
Imputed dataset returns the Null &Residual deviance as follows: There is a significant difference between Null deviance and residual which suggests that the model has significantly improved the fit over the null model. However, this result alone is not sufficient to conclude whether the model is a good fit or not. It is important to also consider other measures.

Null deviance: 993.48 on 767 degrees of freedom
Residual deviance: 709.07 on 759 degrees of freedom

Conclusion: Both Approaches or models returns same end results.
Hence its evident that all the predictor variables have some influence on the diabetes Outcome. Glucose seems to have highest influence on the Outcome variable compare to other variables.

g) Assuming Glucose depends only on Age, used Linear Regression Model to predict the missing values in Glucose as follows.

Created 2 groups, one with Missing values in Glucose (Missing Data) and other one without Missing Values.(Complete Data)

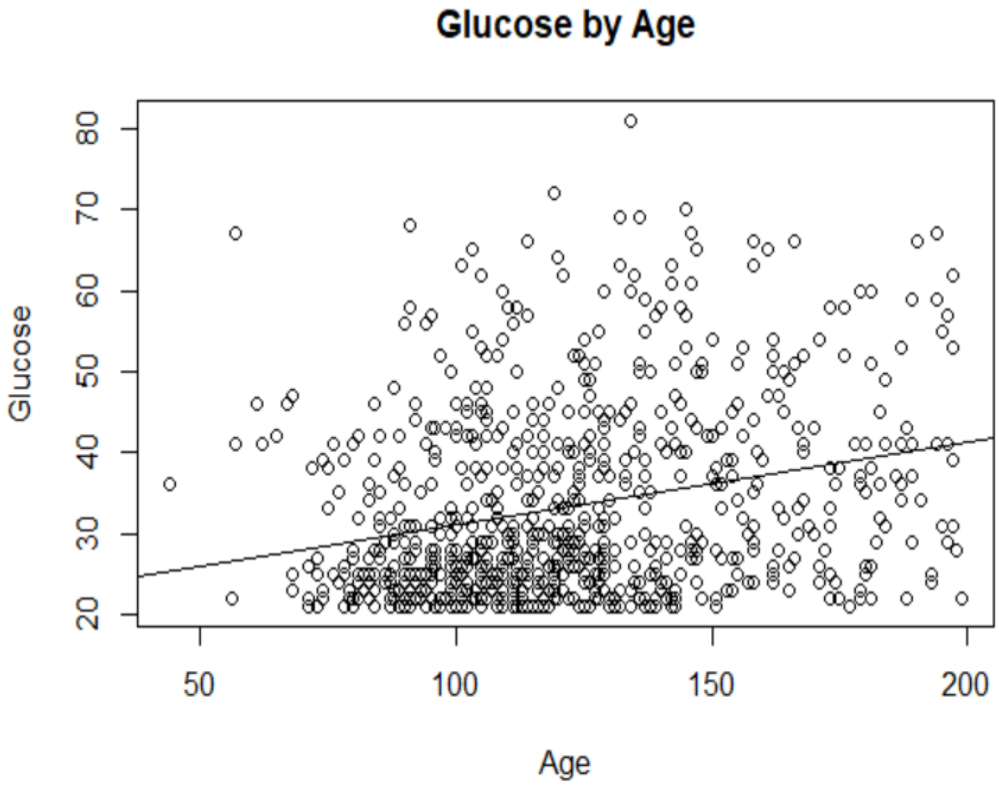
Used lm() function to predict the glucose values from Complete Data set.
Summary(lm(formula = Glucose ~ Age, data = Completedata_Glucose)

```
Call:
lm(formula = Glucose ~ Age, data = Completedata_Glucose)

Residuals:
    Min      1Q  Median      3Q     Max
-88.058 -21.310  -3.727  17.615  85.123

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 98.63245    3.19767  30.845  < 2e-16 ***
Age          0.69292    0.09061   7.647 6.21e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29.45 on 761 degrees of freedom
Multiple R-squared:  0.07136,   Adjusted R-squared:  0.07014
F-statistic: 58.48 on 1 and 761 DF,  p-value: 6.208e-14
```

(Intercept) 98.63245 : Value of Glucose when Age is 0.
P value 6.21e-14 is very small and its evident that Age is an influencing factor to predict Glu cose

Values. Estimate of Age 0.69 predicts the rate of increase in Glucose levels for every unit incr ease in Age. Using this prediction, we can find the missing values in Glucose. Considering Multiple R-Squared Value, approximately 7% variance in Glucose can be explained by Age With residual standar error 29.45.

Pass results of lm() function as an argument in Predict() function, and Missing Dataset. Predict()function returns a vector of predicted values based on a model that has been fitted to complete data.



Followed below steps to verify the accuracy of the predicted values:
Verify the mean of Original dataset with Missing Values & it remains almost same after imputation as well.

mean(diabetes$Glucose,na.rm=TRUE) ## mean of original dataset 121.6868
mean(ImputedGlucose$Glucose)##mean of imputed dataset 121.6654

Below are the predicted values of Glucose:

```
Row Number:       76 183 343 350 503
Predicted Values:114 113 114 124 127
```

##Confidence interval on predictions
confint(predictglucose)

```
              2.5 %      97.5 %
(Intercept) 92.3551519 104.9097386
Age          0.5150415   0.8707967
```

Conclusion:

The predicted values are higher than the values returned by 95% confidence interval.
It means that the predicted values may be outside the range of values that we are 95% confident the true values will fall within. This could indicate that the model is not accurately capturing the relationship between the predictor variables and the outcome variable. It could be due to the fact that Glucose values are correlated with other variables in the dataset and not only with Age.

References:

- *Corder, G.W.; Foreman, D.I. (2014). Nonparametric Statistics: A Step-by-Step Approach. Wiley. ISBN 978-1118840313.*
- *Thieme, Christian: (Mar 12, 2021):* Understanding Linear Regression Output in R. Refer: *https://towardsdatascience.com/understanding-linear-regression-output-in-r-7a9cbda948b3#:~:text=The%20Multiple%20R%2Dsquared%20value,model%20is%20fitting%20the%20data.*
- Enders, Felicity Boyd. "collinearity". Encyclopedia Britannica, 24 Dec. 2013, https://www.britannica.com/topic/collinearity-statistics. Accessed 16 March 2023.
- Hart, Anna (2001). *"Mann–Whitney test is not just a test of medians: differences in spread can be important". BMJ*. **323** (7309): 391–393. doi:10.1136/bmj.323.7309.391.
- Arnold, Taylor B.; Emerson, John W. (2011). *"Nonparametric Goodness-of-Fit Tests for Discrete Null Distributions" (*PDF). *The R Journal*. **3** (2): 34\[Dash]39. doi:10.32614/rj-2011-016.