PREDICTION OF POLYCYSTIC OVARIAN SYNDROME USING LOGISTIC REGRESSION

PROJECT WORK-DSE-B2

BY NILANJANA DEY

UNDER THE GUIDANCE OF

Prof. DITHI BHATTACHARYA

CU REGISTRATION NO.-041-1211-0430-20
CU ROLL NO.- 203041-11-0076

Submitted for Bachelors of Science Course of Department of Statistics (Under CBCS) for the year 2023

UNIVERSITY OF CALCUTTA

BASANTI DEVI COLLEGE

DECLARATION

This declaration is made on the

…… 2023

Student's Declaration:

I -----NILANJANA DEY--------CU ROLL NUMBER------203041-11-0076----------- hereby declare that the work submitted for the module ------DSE-B2------is my original work. I have not copied from any other students' work or from any other sources except where due reference oracknowledgment is made explicitly, nor has any part been authored by another person.

Date submitted: _____________

Received for examination by:____University of Calcutta___ Date: ____________________________

Examiner's Signature: ______________

Examinee's Signature: ______________

CONTENTS

EXECUTIVE SUMMARY

Polycystic ovary syndrome (PCOS) is a common condition in women where multiple cysts are found in the ovaries. 1Experts estimate that as many as 20% of women in the world suffer from polycystic ovaries (PCO), without suffering from the full syndrome (PCOS), which affects between 5-10% of women and involves hormone imbalances and a range of concurrent symptoms.
Women suffering from full PCOS additionally have an imbalance of hormones, and a range of other symptoms that can include: unwanted hair growth, weight gain, acne, an irregular menstrual cycle and depression. The symptoms can be mild or severe, and in many cases, they tend to begin in adolescence.

In this project we will try to predict the PCOS condition of patients from a large variety of factors or symptoms. Mainly we shall try to fit the logistic regression model and check how accurate it is.

INTRODUCTION

Polycystic ovarian syndrome (PCOS) is the most common endocrine pathology in females of reproductive worldwide. Stein and Leventhal initially described it in 1935. The prevalence ranges between 5% and 15% depending on the diagnostic criteria applied. It is widely accepted among specialty society guidelines that the diagnosis of PCOS must be based on the presence of at least two of the following three criteria: chronic anovulation, hyperandrogenism (clinical or biological), and polycystic ovaries. It is a diagnosis of exclusion, and disorders that mimic clinical features of PCOS must be excluded. These include thyroid disease, hyperprolactinemia, and non-classical congenital adrenal hyperplasia. Selected patients may need more extensive workup if clinical features suggest other causes.

Despite its high prevalence, PCOS is underdiagnosed and frequently takes more than one visit or different physicians to get identified, and these usually occur in more than a one-year timeframe.

Symptoms

Symptoms of PCOS often start around the time of the first menstrual period. Sometimes symptoms develop later after you have had periods for a while.

The symptoms of PCOS vary. A diagnosis of PCOS is made when you have at least two of these:

- Irregular periods: Having few menstrual periods or having periods that aren't regular are common signs of PCOS. So is having periods that last for many days or longer than is typical for a period. For example, you might have fewer than nine periods a year. And those periods may occur more than 35 days apart. You may have trouble getting pregnant.
- Too much androgen: High levels of the hormone androgen may result in excess facial and body hair. This is called hirsutism. Sometimes, severe acne and male-pattern baldness can happen, too.
- Polycystic ovaries: Your ovaries might be bigger. Many follicles containing immature eggs may develop around the edge of the ovary. The ovaries might not work the way they should.
- Skin Darkening: PCOS also leads to acanthosis nigricans which is darkening of the skin which looks like purple or black in color in the

underarm, collar areas of the neck, inner thighs, around lips, under eyes and forehead. In many women the PCOS causes excessive dry or oily skin.

- Pimples: High levels of androgens are one of the most common symptoms of PCOS. Hyperandrogenism is the medical term for this condition. Androgens play a significant influence in PCOS acne formation. They induce the skin’s glands to create an excessive amount of sebum, an oily material.

- Weight Gain: PCOS makes it more difficult for the body to use the hormone insulin, which normally helps convert sugars and starches from foods into energy. This condition -- called insulin resistance -- can cause insulin and sugar -- glucose -- to build up in the bloodstream. Because the weight gain is triggered by male hormones, it is typically in the abdomen.

- Hair growth: PCOS causes both hirsutism and androgenic alopecia. Hirsutism refers to increased hair growth, typically on the face and body.

- AMH levels: AMH is a member of the transforming growth factor-β superfamily that is produced by growing ovarian antral follicles. Serum Anti-Müllerian Hormone (AMH) correlates with the total number of antral follicles over both ovaries, and therefore has been proposed as a biomarker for PCOS diagnosis.

- Thinning hair**:** People with PCOS may lose patches of hair on their head or start to bald.
- Infertility**:** PCOS is the most common cause of infertility in people AFAB. Not ovulating regularly or frequently can result in not being able to conceive.

Through this project we are going to analyse the most important factors that are fundamental in detecting PCOS in women.

METHODOLOGY

To predict the Polycystic Ovarian Syndrome (PCOS) , one needs to follow these steps:

1. **DATA COLLECTION**: Collection of the data is an important aspect in any Statistical analysis. The data for this project is collected from https://www.kaggle.com/.

2. **DATA CLEANSING**: All the missing observations have to be removed. Then we have to look out for outliers, high leverages and influential observations.

- ***Outliers***-In simple terms, an outlier is an extremely high or extremely low data point relative to the nearest data point and the rest of the neighbouring co-existing values in a dataset. To identify the outliers, we calculate the cook's distance which is a measure used in regression analysis to identify influential data points or outliers that may have a significant impact on the regression results. It measures the change in the regression coefficients when a particular observation is removed from the dataset.
- ***Influential observations***-In statistics, an influential observation is an observation for a statistical calculation whose deletion from the dataset would noticeably change the result of the calculation. In particular, in regression analysis an influential observation is one whose deletion has a large effect on the parameter estimates.

- ***High leverages***-A data point has high leverage if it has "extreme" predictor x values. With a single predictor, an extreme x value is simply one that is particularly high or low.

3. DATA REPRESENTATION: using bar graphs, line charts and contingency tables.

4. FITTING THE DATA INTO A LOGISTIC REGRESSION MODEL:

Logistic Regression : overview

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. The below image is showing the logistic function:

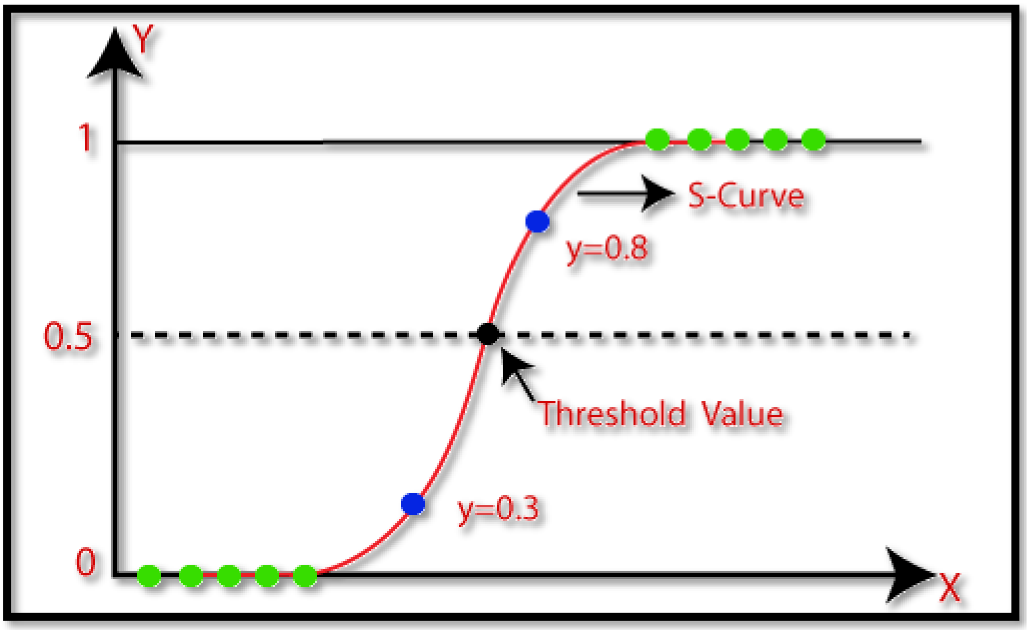


FIGURE 1: LOGISTIC CURVE

Logistic Function (Sigmoid Function):

- The sigmoid function is a mathematical function used to map the predicted values to probabilities.
- It maps any real value into another value within a range of 0 and 1.
- The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the Sigmoid function or the logistic function.
- In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

Logistic Regression Equation:

The Logistic regression equation can be obtained from the Linear Regression equation. The mathematical steps to get Logistic Regression equations are given below:

- We know the equation of the straight line can be written as:

$$y= b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

- In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

$$\frac{y}{1-y}\text{; 0 for y= 0, and infinity for y=1}$$

- But we need range between -∞ to +∞ then take logarithm of the equation it will become:

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

The above equation is the final equation for Logistic Regression.

Requirements for Logistic Regression

This model can work for all the datasets, but still, if you need good performance, then there will be some assumptions to consider,

1. The dependant variable in binary logistic regression must be binary.
2. The independent variables must be unrelated to one another. That is, there should be minimal or no multicollinearity in the model.
3. The log chances are proportional to the independent variables.
4. Large sample sizes are required for logistic regression.

5. **Check for Multicollinearity** :
   Multicollinearity is the occurrence of high intercorrelations among two or more independent variables in a multiple regression model. Multicollinearity can lead to skewed or misleading results when a researcher or analyst attempts to determine how well each independent variable can be used most effectively to predict or understand the dependent variable in a statistical model.

Detecting Multicollinearity

Correlation Matrix: A correlation plot can be used to identify the correlation or bivariate relationship between two independent variables.

We can also use statistical technique called the **variance inflation factor** (VIF) to detect and measure the amount of collinearity in a multiple regression model. VIF measures how much the variance of the estimated regression coefficients is inflated as compared to when the predictor variables are not linearly related. A VIF of 1 will mean that the variables are not correlated; a VIF between 1 and 5 shows that variables are moderately correlated, and a VIF between 5 and 10 will mean that variables are highly correlated.

5) ASSESSING THE GOODNESS-OF-FIT OF THE MODEL

❖ *Deviance:*

Using deviance, we can compare the current model with saturated model. A saturated model is that model which can provide the perfect fit for the data. Deviance is defined as

deviance = -2*(log-likelihood for the current model — log-likelihood for the saturated model)

In a saturated model, the number of parameters equals the sample size since it contains one parameter for each observation. The likelihood of a saturated model is 1 which mean a saturated model can provide perfect prediction from the predictor variables. Therefore, the log-likelihood for a saturated model is 0 and as a result deviance becomes -2*(log-likelihood for the current model).Here, in logistic regression, the comparable statistics to reduce the error variance is deviance. In other words, in logistic regression, the best model tries to reduce the deviance (i.e. the log-likelihood between the model under study and the saturated model).

❖ *Log-likelihood ratio :*

The overall significance of a logistic regression can be assessed with a likelihood ratio test where the null (constant only) model is compared to the current model including predictors. The larger the difference, the greater the evidence that the model is significant. The log of the likelihood ratio is the difference between these two log likelihoods. In practice we work with minus twice the log of the likelihood ratio as the log of the likelihoods are always negative. As with the G-test, where all frequencies are large, the natural log of L^2 (twice the log of the likelihood ratio) has an approximately chi-square distribution.

$$G = -2ln\left[\frac{L_{null}}{L_{full}}\right]$$

where,

- L_{null} is the log likelihood for the null model,
- L_{full} is the log likelihood for the current model

Hence the log likelihood ratio statistic is given by -2 log L_{null} - (-2 log L_{full}).

The number of degrees of freedom is equal to the difference between the number of β-parameters being fitted under the two models.

❖ ***Pseudo R²:***

More or less, we are all familiar with the R^2 interpretation in linear regression but in logistic regression, it's interpretation is different. The McFadden's Pseudo R^2 is defined as

McFadden's R^2= 1 — (deviance of the fitted model / deviance of the null model)

= 1 + (log-likelihood for the fitted model / log-likelihood for the null model)

When a model's likelihood value is small, the log-likelihood value becomes larger. If the null model is less likely and the fitted model is more likely, the second part of the above equation becomes smaller. In perfect case, this second portion becomes 0 and Pseudo R^2 value becomes 1. A value of 1 for Pseudo R^2 indicates that we can predict the probability of success or failure perfectly.

❖ ***AIC statistic:***

The likelihood ratio test and Pseudo R² are used to compare models which are nested. That means one model has less number of parameters than the other. In cases where the two models have different set of parameters, we cannot use likelihood ratio test and Pseudo R² to compare the models. That is when AIC (Akaike Information Criterion) statistic come into the picture. AIC is defined as:

AIC = -2*(Log-likelihood of the current model — k)

here, k = the total number of parameters in the model including intercept and n = sample size.

Likelihoods are between 0 and 1, so their log is less than or equal to zero. If a model is more likely, it's log-likelihood becomes smaller on negative side and "-2*log-likelihood" value becomes positive but small in value. AIC can be used to compare both nested and non-nested models. The model with lower AIC value provides the best fit.

❖ ***P-value***

It is a statistical metric that helps statisticians decide whether they should accept or reject the null hypothesis. P-values are traditionally used across many statistical techniques including ANOVA, t-tests, and regression. The p-value measures the probability there is no relationship between variables. A low p-value gives evidence against the null hypothesis. P-values provide evidence only in favour or against the null hypothesis. Specifically, a p-value < 0.05 is good. A p-value< 0.05 means there is a 5% probability that there is no relationship between the variables. Moreover, a small p-value can be interpreted as there is a small probability that the effect observed is due to chance.

❖ ***Confusion Matrix:***

Confusion matrix is used to evaluate classification models. Bear in mind, this is not used to evaluate linear regression models. Classification models categorize the outcome into two or more categories (e.g. whether an email is spam or non-spam), while linear regression predicts a number (e.g. predicting house price).

The Decision Matrix

		Predicted condition	
		Test (+)	**Test (-)**
True condition	Disease (+)	a	b
	Disease (-)	c	d

The receiver operating characteristic curve is drawn with the x-axis as 1 – specificity (false positive) and the y-axis as sensitivity.

❖ ***ROC curve:***

ROC curve is used to assess the overall diagnostic performance of a test and to compare the performance of two or more diagnostic tests. It is also used to select an optimal cut-off value for determining the presence or absence of a disease.

To understand the ROC curve, it is first necessary to understand the meaning of sensitivity and specificity, which are used to evaluate the performance of a diagnostic test. Sensitivity is defined as the proportion of people who actually have a target disease that are tested positive, and specificity is the proportion of people who do not have a target disease that are tested negative. FP refers to the proportion of people that do not have a disease but are incorrectly tested positive, while FN refers to the proportion of people that have the disease but are incorrectly tested negative.

DATA SUMMARY

DATA SOURCE: https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos

DATA DESCRIPTION:

The data is collected from 10 different hospitals across Kerala, India.

Instructions to be followed :
1) For every Yes/No questions *** , Indicate Yes = 1 ; No= 0
2) Blood Group indications **
A+ = 11, A- = 12, B+ = 13, B- = 14, O+ =15, O- = 16, AB+ =17, AB- = 18
3) RBS - Random glucose test

FSH-follicle stimulating hormone

PRL- prolactin

AMH- anti-müllerian hormone

LH-luteinizing hormone

BMI- Body Mass Index

TSH-Thyroid stimulating hormone

PRG- progesterone

4) Beta-HCG cases are mentioned as Case I and II.

The data file has 541 rows and 40 columns. It's not possible to accommodate the entire dataset in this project. Therefore, we will display only a part of it.

Since we will use logistic regression model, it is important to choose our predictor variables and the outcome variable.

Outcome variable- PCOS(Y/N)

Predictor variables: AMH.ng.mL. , Age..yrs. Skin.darkening..Y.N. ,Pimples.Y.N. ,Weight.gain.Y.N. ,hair.growth.Y.N. ,Weight..Kg. ,Height.Cm. ,Blood.Group, Pulse.rate.bpm. ,RR..breaths.min., Hb.g.dl. ,Cycle.R.I. Cycle.length.days., Marraige.Status..Yrs. ,Pregnant.Y.N. , No..of.aborptions , I...beta.HCG.mIU.mL. ,FSH.mIU.mL., LH.mIU.mL. , Fast.food..Y.N. ,Reg.Exercise.Y.N. ,BP._Systolic..mmHg., BP._Diastolic..mmHg. ,Follicle.No...L. ,Follicle.No...R. , Avg..F.size..L...mm. , Avg..F.size..R...mm., Endometrium..mm. , Hip.inch. ,Waist.inch., TSH..mIU.L., PRL.ng.mL.

DATA TABLE : The data table given below has 5 rows and 40 columns representing 38 predictor variables and the outcome variable.

TABLE 1: DISPLAYING 4 PATIENTS AND THEIR CORRESPONDING PCOS CONDITION ALONG WITH 38 FACTORS

Patient File No.	PCOS (Y/N)	Age (yrs)	Weight (Kg)	Height(Cm)	Blood Group	Pulse rate(bpm)	RR (breaths/min)	Hb(g/dl)	Cycle(R/I)
1	0	28	44.6	152	15	78	22	10.48	2
2	0	36	65	161.5	15	74	20	11.7	2
3	1	33	68.8	165	11	72	18	11.8	2
4	0	37	65	148	13	72	20	12	2

Cycle length(days)	Marriage Status (Yrs)	Pregnant(Y/N)	No. of abortions	I beta-HCG(mIU/mL)	II beta-HCG(mIU/mL)	FSH(mIU/mL)	LH(mIU/mL)
5	7	0	0	1.99	1.99	7.95	3.68
5	11	1	0	60.8	1.99	6.73	1.09
5	10	1	0	494.08	494.08	5.54	0.88
5	4	0	0	1.99	1.99	8.06	2.36

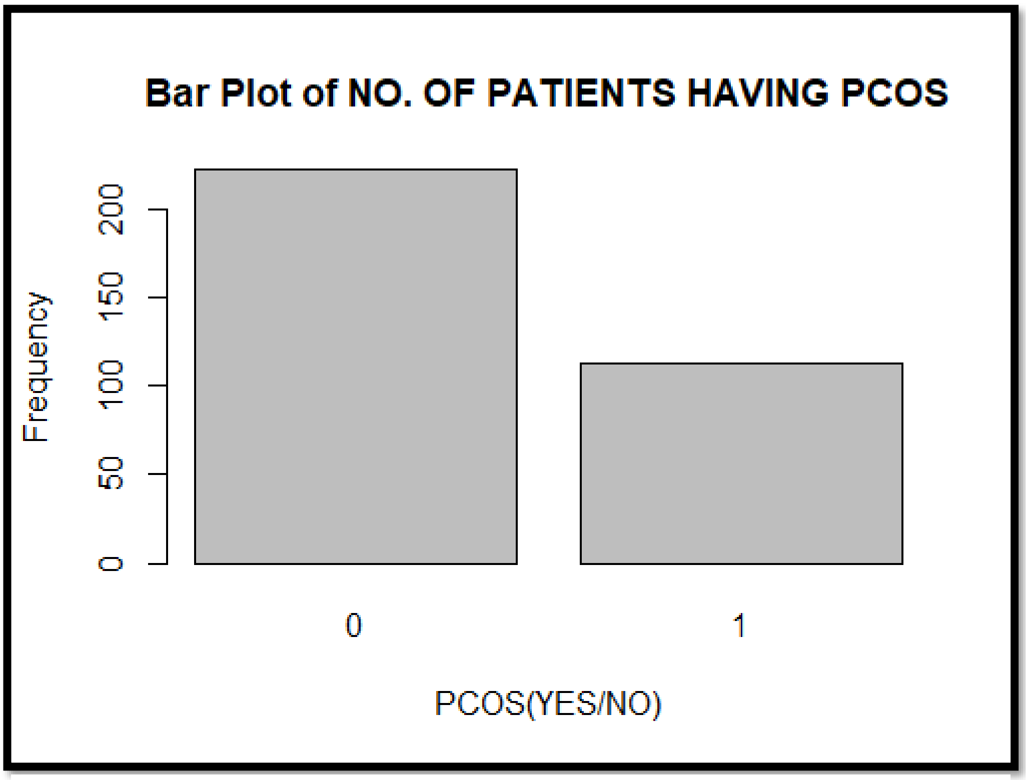
Hip(inch)	Waist(inch)	TSH (mIU/L)	AMH(ng/mL)	PRL(ng/mL)	Vit D3 (ng/mL)	PRG(ng/mL)	RBS(mg/dl
36	30	0.68	2.07	45.16	17.1	0.57	92
38	32	3.16	1.53	20.09	61.3	0.97	92
40	36	2.54	6.63	10.52	49.7	0.36	84
42	36	16.41	1.22	36.9	33.4	0.36	76

Weight gain(Y/N)	hair growth(Y/N)	Skin darkening (Y/N)	Hair loss(Y/N)	Pimples(Y/N)	Fast food (Y/N)	Reg.Exercise.Y.N	BP _Systoli (mmHg)
0	0	0	0	0	1	0	110
0	0	0	0	0	0	0	120
0	0	0	1	1	1	0	120
0	0	0	0	0	0	0	120

BP _Diastolic (mmHg)	Follicle No. (L)	Follicle No. (R)	Avg. F size (L) (mm)	Avg. F size (R) (mm)	Endometrium (mm)
80	3	3	18	18	8.5
70	3	5	15	14	3.7
80	13	15	18	20	10
70	2	2	15	14	7.5

RESULTS AND INTERPRETATION

The calculations have been done in R software and MS excel.

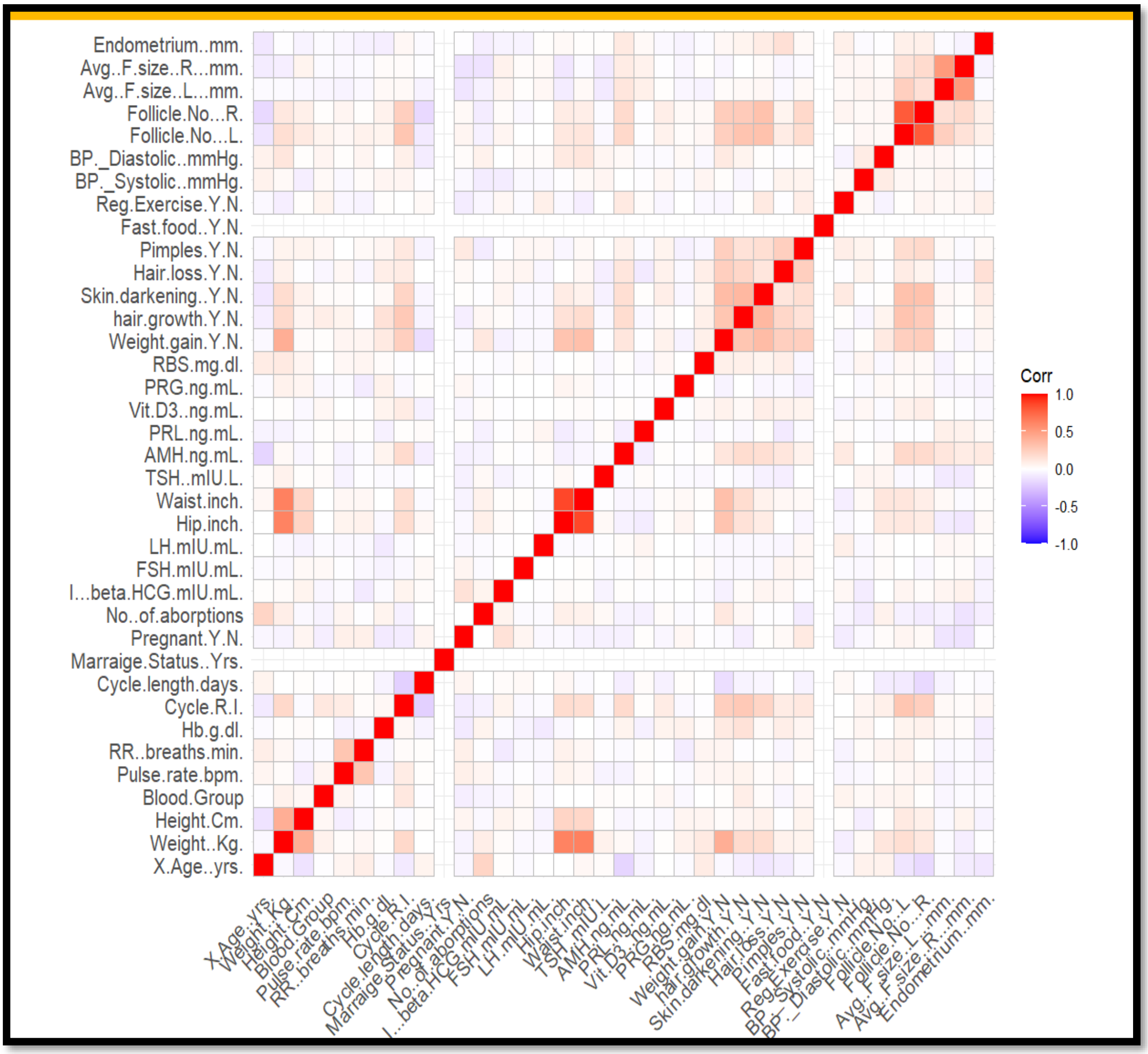


GRAPH 2 : NO. OF PATIENTS HAVING PCOS FROM THE GIVEN DATASET

Representation of the no. of PCOS patients using bar diagram. The figure below shows the frequency of occurrence of 1(which denotes that the patient has PCOS). The total no of patients having PCOS is 177.

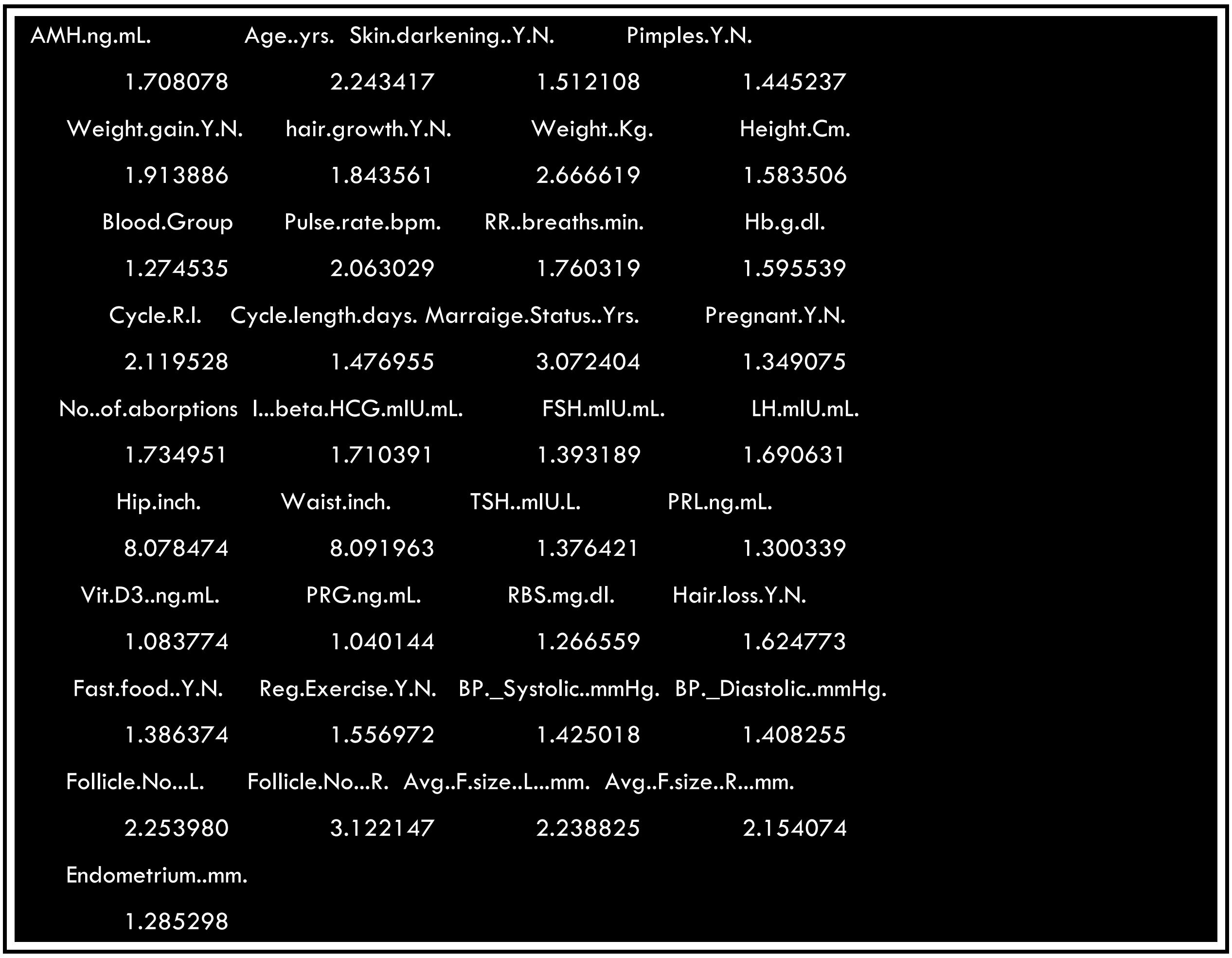
MULTICOLLINEARITY CHECKING:

- The **correlation matrix** given below gives an idea about the existence of multicollinearity among the factors given in the data set. The dak shades of red indicates the presence of multicollinearity. We can see that the factors " hip (inches)" and "waist (inches", "Follicle no.(R)" and "Follicle no.(L)" , "Avg F size(r) mm" and "Avg F size(l) mm" have high correlation. The row and column corresponding to "fast food(Y/N)" and " marriage status(yrs.)" remains blank because they don't take numeric values.



GRAPH 3: CORRELATION PLOT

- **VIF values** : As a rule of thumb, VIF values above 5 indicate severe multicollinearity. Since " hip (inches)" and "waist (inches" factors in our model have a VIF over 5, we can confirm that multicollinearity is an issue in our model. The results given below shows the VIF values for all the factors.

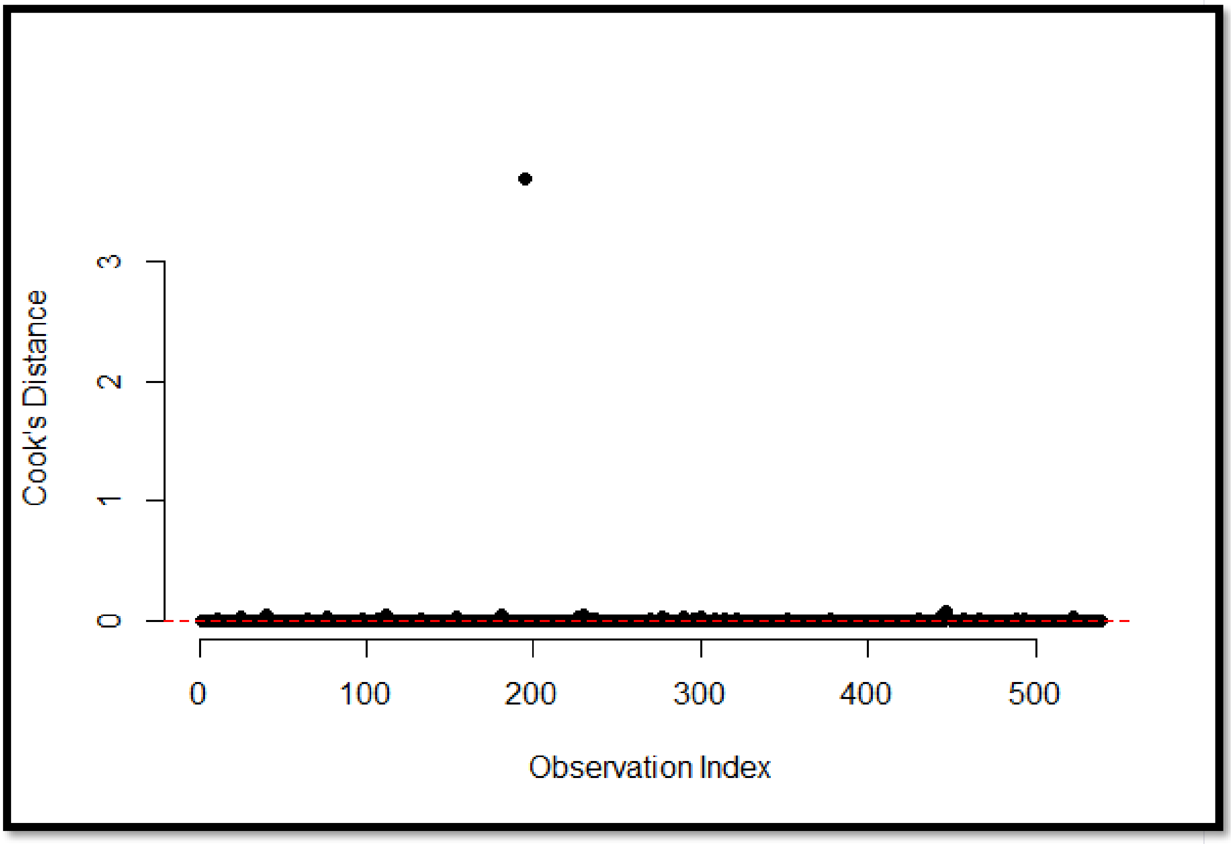


```
AMH.ng.mL.          Age..yrs.  Skin.darkening..Y.N.        Pimples.Y.N.
          1.708078          2.243417          1.512108          1.445237
   Weight.gain.Y.N.    hair.growth.Y.N.        Weight..Kg.         Height.Cm.
          1.913886          1.843561          2.666619          1.583506
       Blood.Group       Pulse.rate.bpm.     RR..breaths.min.          Hb.g.dl.
          1.274535          2.063029          1.760319          1.595539
        Cycle.R.I.    Cycle.length.days. Marraige.Status..Yrs.       Pregnant.Y.N.
          2.119528          1.476955          3.072404          1.349075
  No..of.aborptions  I...beta.HCG.mIU.mL.        FSH.mIU.mL.          LH.mIU.mL.
          1.734951          1.710391          1.393189          1.690631
         Hip.inch.          Waist.inch.         TSH..mIU.L.          PRL.ng.mL.
          8.078474          8.091963          1.376421          1.300339
     Vit.D3..ng.mL.          PRG.ng.mL.          RBS.mg.dl.       Hair.loss.Y.N.
          1.083774          1.040144          1.266559          1.624773
    Fast.food..Y.N.     Reg.Exercise.Y.N.   BP._Systolic..mmHg.  BP._Diastolic..mmHg.
          1.386374          1.556972          1.425018          1.408255
   Follicle.No...L.     Follicle.No...R.  Avg..F.size..L...mm.  Avg..F.size..R...mm.
          2.253980          3.122147          2.238825          2.154074
   Endometrium..mm.
          1.285298
```

Conclusion: We remove the factors " hip (inches)" and "waist (inches" and proceed for further calculations and assessment.

REMOVING INFLUENTIAL OBSERVATIONS, HIGH LEVERAGE VALUES AND OUTLIERS

❖ The **Cook's distance** can be visualized using a graph with the Cook's distance values plotted against the observation index (or observation number).



GRAPH 4: COOK'S DISTANCE VS NO. OF OBSERVATIONS PLOT

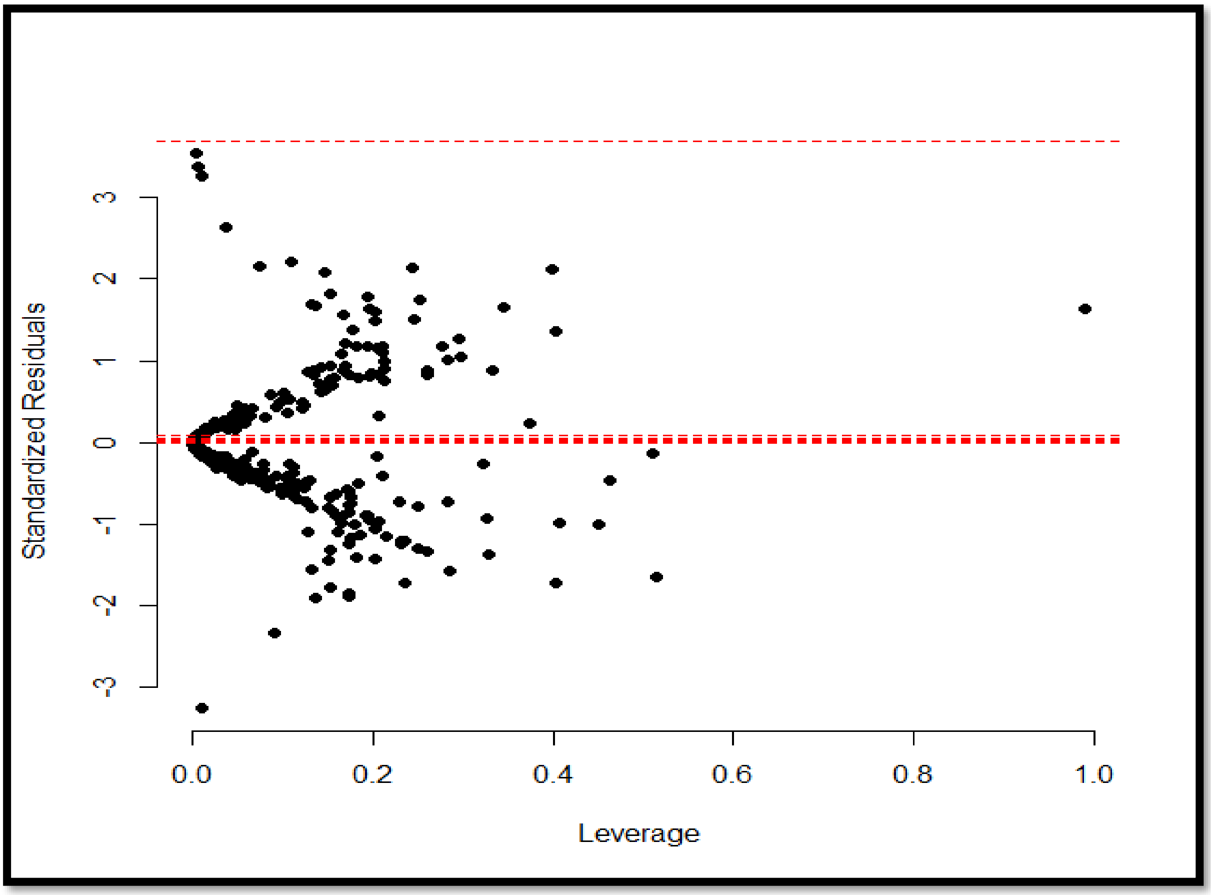
1. **Threshold Line**: The red dotted line is the threshold line on the graph to identify influential observations. The threshold line is typically set at a value of 4/(n-k-1), where "n" is the total number of observations(i.e., 541) and "k" is the number of predictors (i.e., 35) in the regression model. Observations that have Cook's distance values above this threshold line are considered potentially influential.

2. **Interpreting the Graph**: When interpreting the Cook's distance vs. observation index graph, we can see a couple of points crossing the

threshold line. There is an observation far away from all the other observations that are more or less around the threshold line. These observations are likely to have a significant impact on the regression results if removed from the analysis.

- If a point is well below the threshold line, it indicates that the corresponding observation has little influence on the regression model.
- If a point is close to or above the threshold line, it suggests that the corresponding observation has a higher influence on the regression model and may be considered an influential point or an outlier.

❖ The **leverage vs. standardized residuals** graph is a diagnostic tool used to assess the influence of individual observations in regression analysis. It helps identify observations that have high leverage (extreme values of predictors) and/or large standardized residuals (deviations from the expected model fit).



GRAPH 5: STANDARDIZED RESIDUALS VS LEVERAGE PLOT

INTERPRETATION OF THE GRAPH:

1. **Leverage Values**: The leverage values are plotted on the x-axis of the graph. Leverage measures the potential impact of an observation on the regression model.

2. **Standardized Residuals**: The standardized residuals are plotted on the y-axis of the graph. Standardized residuals represent the difference between the observed values and the predicted values, scaled by the standard error. They indicate the magnitude and direction of the deviations from the expected model fit.

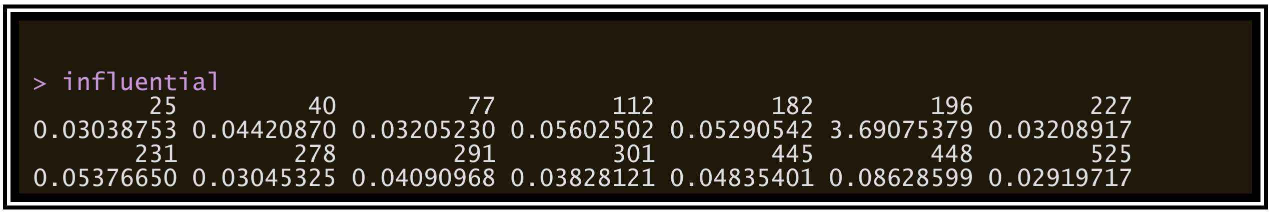
3. **Shape of the Graph**: In a leverage vs. standardized residuals graph, we are looking for specific patterns or outliers:
   - *Residual Distribution*: The standardized residuals are normally distributed around zero. Most of the points are clustered around zero, indicating a good fit of the model.
   - *Horizontal Line*: Typically, a horizontal line is drawn at zero on the y-axis to represent the expected mean of the standardized residuals. Points above or below this line indicate positive or negative deviations from the expected fit.

4. **High Leverage Points**: we focus on the observations with high leverage values, usually defined as having leverage greater than twice the average leverage (2/n, where "n" is the total number of observations=541). These observations can have a significant impact on the regression model due to their extreme predictor values. High leverage points are represented by points located on the far left or right of the graph.

5. **Outlying Residuals**: There are observations with large standardized residuals (outliers) that deviate significantly from the expected model fit. Outliers are represented by points located far above or below the horizontal line at zero.

- Identifying the Influential observations:

  The influential observations are identified and omitted from the dataset for a better model fitting. There are 14 influential observations in the dataset.



```
> influential
        25         40         77        112        182        196        227 
0.03038753 0.04420870 0.03205230 0.05602502 0.05290542 3.69075379 0.03208917 
       231        278        291        301        445        448        525 
0.05376650 0.03045325 0.04090968 0.03828121 0.04835401 0.08628599 0.02919717 
```

TRAINING AND TEST DATASET

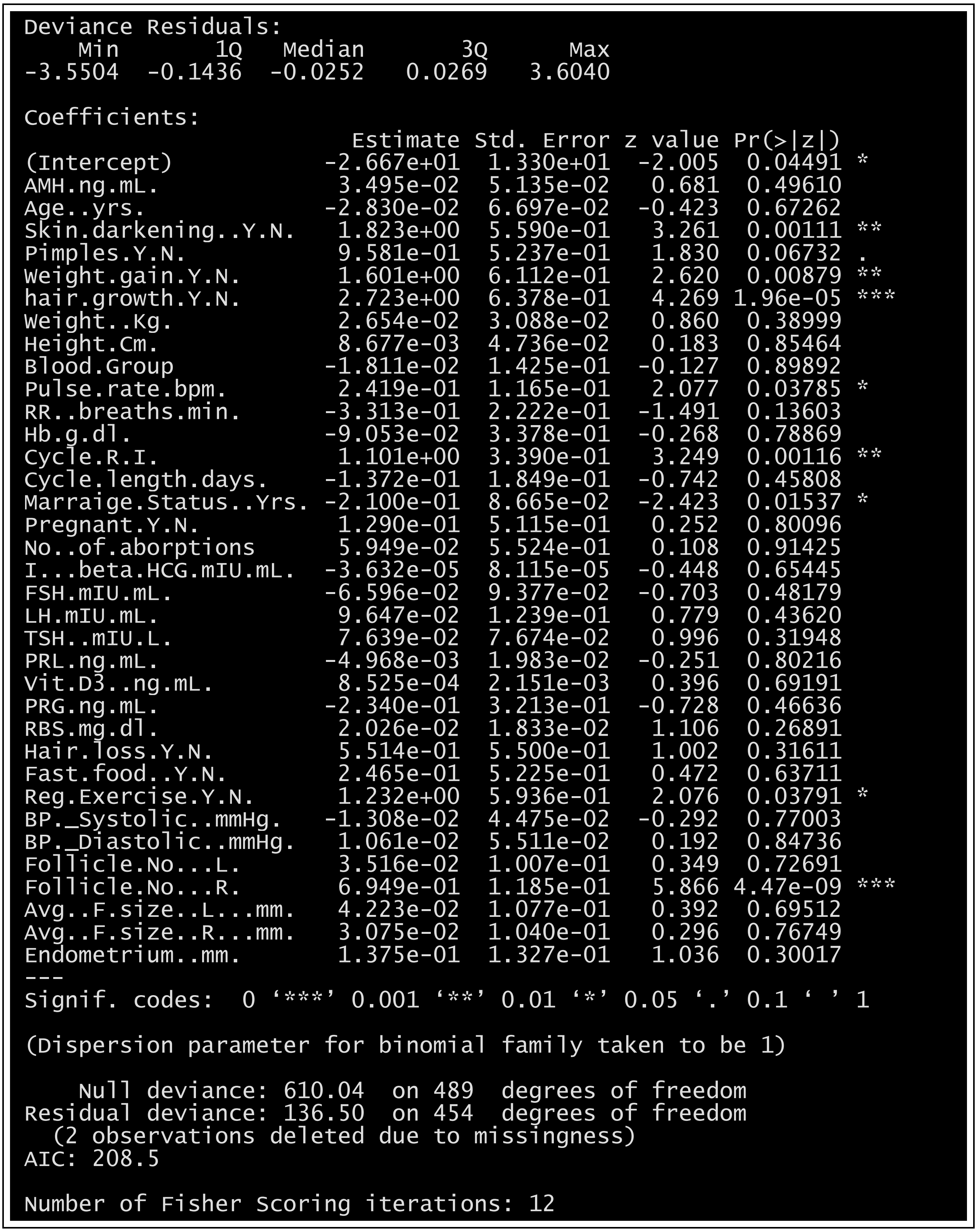
We divide the cleansed data set into training and test data set. The training dataset is used for model fitting and assessing its predictive capacity. The test dataset is used for model diagnostics. The training dataset contains 70% of the data and the test dataset contains 30% of the data.

MODEL FITTING

- **Model Summary**:

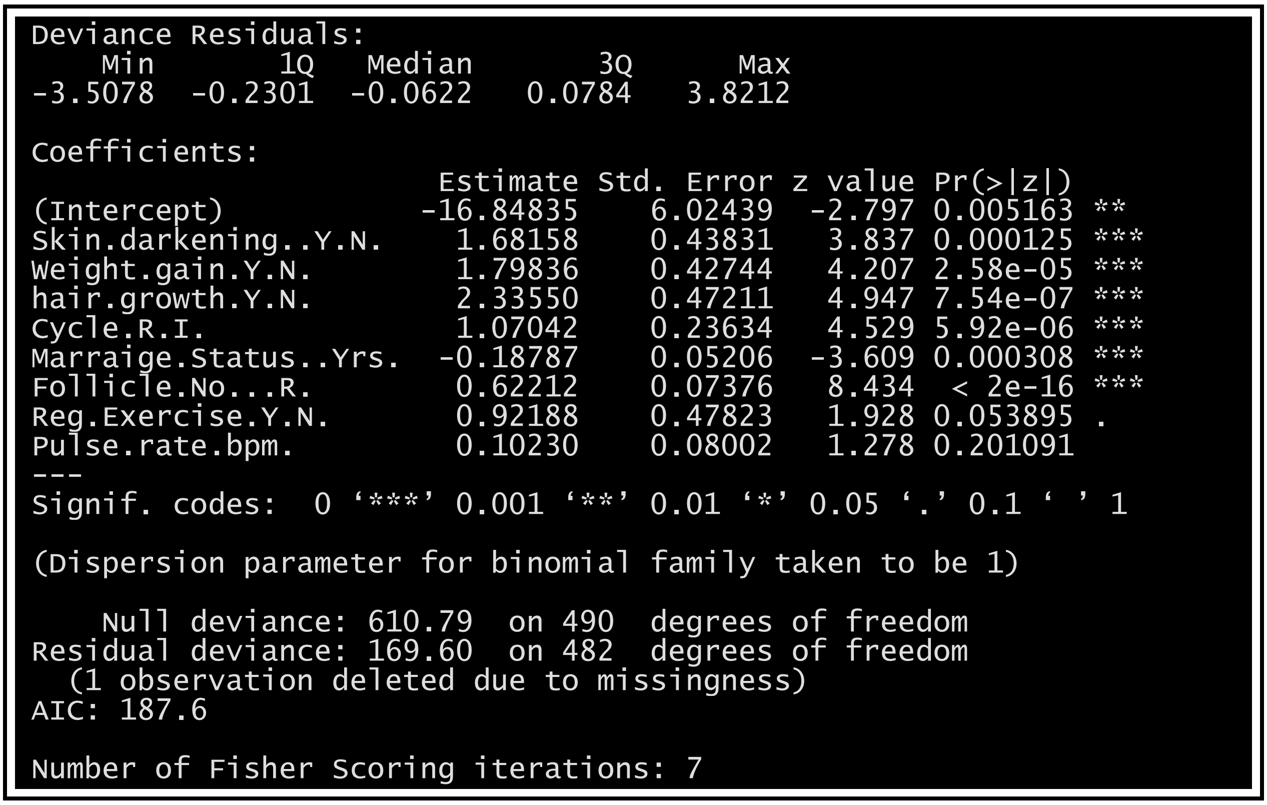
1. The factors with P-value < 0.05 are denoted by *, **, ***. These factors are the most significant for prediction of Polycystic Ovarian Syndrome.

   The factors are: Skin.darkening..Y.N. ,Weight.gain.Y.N. , hair.growth.Y.N., Pulse.rate.bpm. ,Cycle.R.I., Marraige.Status..Yrs., Reg.Exercise.Y.N., Follicle.No...R.



```
Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.5504  -0.1436  -0.0252   0.0269   3.6040  

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)    
(Intercept)           -2.667e+01  1.330e+01  -2.005  0.04491 *  
AMH.ng.mL.             3.495e-02  5.135e-02   0.681  0.49610    
Age..yrs.             -2.830e-02  6.697e-02  -0.423  0.67262    
Skin.darkening..Y.N.   1.823e+00  5.590e-01   3.261  0.00111 ** 
Pimples.Y.N.           9.581e-01  5.237e-01   1.830  0.06732 .  
Weight.gain.Y.N.       1.601e+00  6.112e-01   2.620  0.00879 ** 
hair.growth.Y.N.       2.723e+00  6.378e-01   4.269 1.96e-05 ***
Weight..Kg.            2.654e-02  3.088e-02   0.860  0.38999    
Height.Cm.             8.677e-03  4.736e-02   0.183  0.85464    
Blood.Group           -1.811e-02  1.425e-01  -0.127  0.89892    
Pulse.rate.bpm.        2.419e-01  1.165e-01   2.077  0.03785 *  
RR..breaths.min.      -3.313e-01  2.222e-01  -1.491  0.13603    
Hb.g.dl.              -9.053e-02  3.378e-01  -0.268  0.78869    
Cycle.R.I.             1.101e+00  3.390e-01   3.249  0.00116 ** 
Cycle.length.days.    -1.372e-01  1.849e-01  -0.742  0.45808    
Marraige.Status..Yrs. -2.100e-01  8.665e-02  -2.423  0.01537 *  
Pregnant.Y.N.          1.290e-01  5.115e-01   0.252  0.80096    
No..of.aborptions      5.949e-02  5.524e-01   0.108  0.91425    
I...beta.HCG.mIU.mL.  -3.632e-05  8.115e-05  -0.448  0.65445    
FSH.mIU.mL.           -6.596e-02  9.377e-02  -0.703  0.48179    
LH.mIU.mL.             9.647e-02  1.239e-01   0.779  0.43620    
TSH..mIU.L.            7.639e-02  7.674e-02   0.996  0.31948    
PRL.ng.mL.            -4.968e-03  1.983e-02  -0.251  0.80216    
Vit.D3..ng.mL.         8.525e-04  2.151e-03   0.396  0.69191    
PRG.ng.mL.            -2.340e-01  3.213e-01  -0.728  0.46636    
RBS.mg.dl.             2.026e-02  1.833e-02   1.106  0.26891    
Hair.loss.Y.N.         5.514e-01  5.500e-01   1.002  0.31611    
Fast.food..Y.N.        2.465e-01  5.225e-01   0.472  0.63711    
Reg.Exercise.Y.N.      1.232e+00  5.936e-01   2.076  0.03791 *  
BP._Systolic..mmHg.   -1.308e-02  4.475e-02  -0.292  0.77003    
BP._Diastolic..mmHg.   1.061e-02  5.511e-02   0.192  0.84736    
Follicle.No...L.       3.516e-02  1.007e-01   0.349  0.72691    
Follicle.No...R.       6.949e-01  1.185e-01   5.866 4.47e-09 ***
Avg..F.size..L...mm.   4.223e-02  1.077e-01   0.392  0.69512    
Avg..F.size..R...mm.   3.075e-02  1.040e-01   0.296  0.76749    
Endometrium..mm.       1.375e-01  1.327e-01   1.036  0.30017    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 610.04  on 489  degrees of freedom
Residual deviance: 136.50  on 454  degrees of freedom
  (2 observations deleted due to missingness)
AIC: 208.5

Number of Fisher Scoring iterations: 12
```

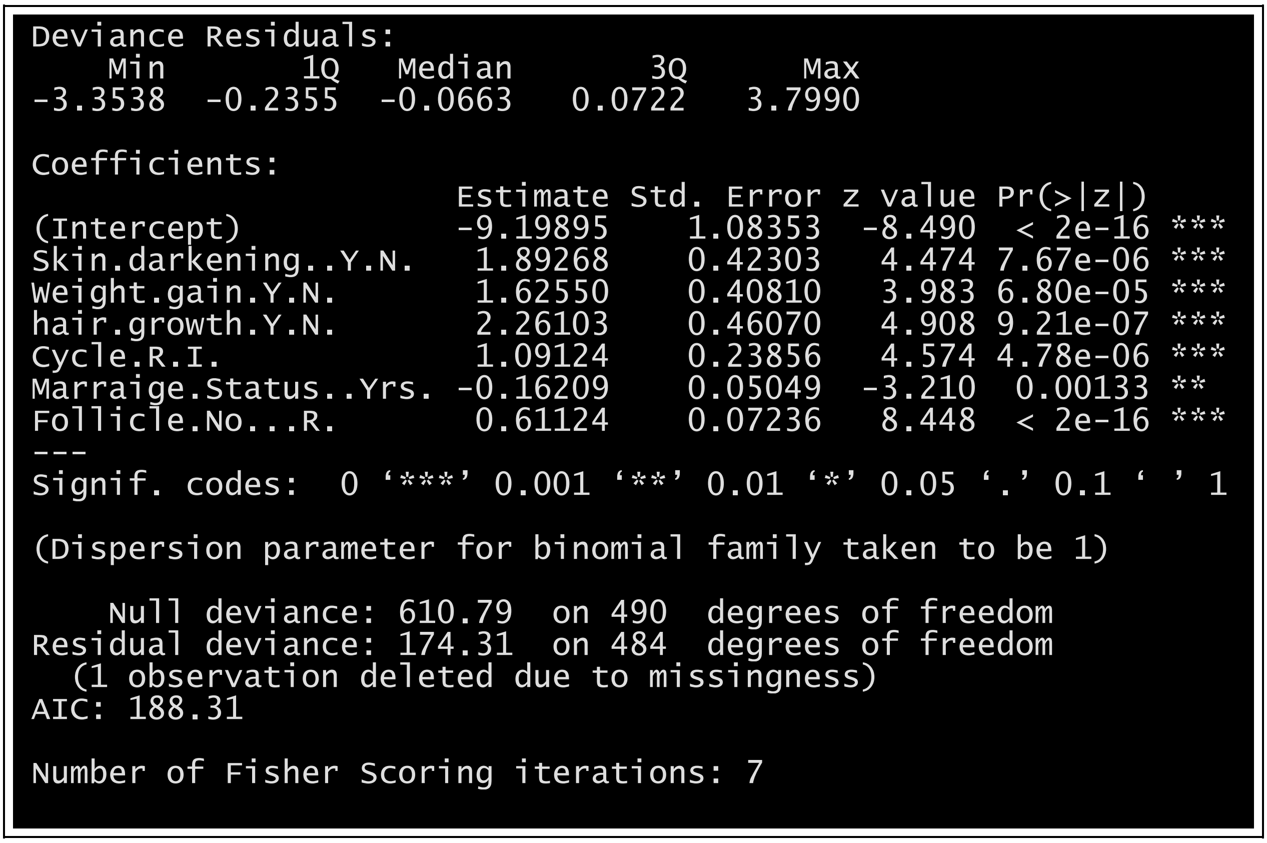
2. Again we fit the logistic regression model using the above 8 significant factors:



```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5078  -0.2301  -0.0622   0.0784   3.8212

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           -16.84835    6.02439  -2.797 0.005163 **
Skin.darkening..Y.N.    1.68158    0.43831   3.837 0.000125 ***
Weight.gain.Y.N.        1.79836    0.42744   4.207 2.58e-05 ***
hair.growth.Y.N.        2.33550    0.47211   4.947 7.54e-07 ***
Cycle.R.I.              1.07042    0.23634   4.529 5.92e-06 ***
Marraige.Status..Yrs.  -0.18787    0.05206  -3.609 0.000308 ***
Follicle.No...R.        0.62212    0.07376   8.434  < 2e-16 ***
Reg.Exercise.Y.N.       0.92188    0.47823   1.928 0.053895 .
Pulse.rate.bpm.         0.10230    0.08002   1.278 0.201091
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 610.79  on 490  degrees of freedom
Residual deviance: 169.60  on 482  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 187.6

Number of Fisher Scoring iterations: 7
```

Interpretation: After fitting the model, the factors REGULAR EXERCISE and PULSE RATE is not significant.

3. Again we fit the model with 6 factors.



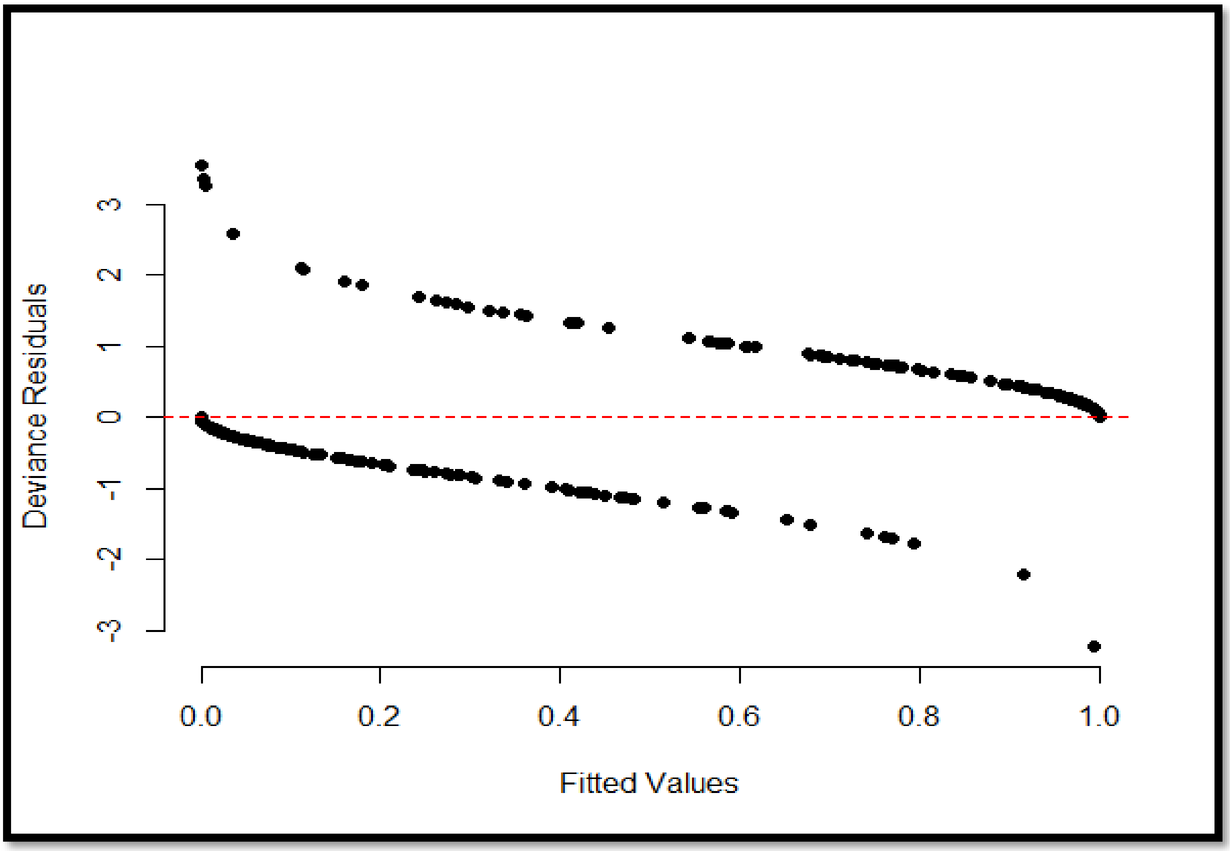
```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3538  -0.2355  -0.0663   0.0722   3.7990

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -9.19895    1.08353  -8.490  < 2e-16 ***
Skin.darkening..Y.N.   1.89268    0.42303   4.474 7.67e-06 ***
Weight.gain.Y.N.       1.62550    0.40810   3.983 6.80e-05 ***
hair.growth.Y.N.       2.26103    0.46070   4.908 9.21e-07 ***
Cycle.R.I.             1.09124    0.23856   4.574 4.78e-06 ***
Marraige.Status..Yrs. -0.16209    0.05049  -3.210  0.00133 **
Follicle.No...R.       0.61124    0.07236   8.448  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 610.79  on 490  degrees of freedom
Residual deviance: 174.31  on 484  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 188.31

Number of Fisher Scoring iterations: 7
```

Interpretation: The factors are all significant. We take these 6 factors for further assessment. Also, the **AIC values** are lower compared to the first fitting where it was 208.5. The smaller the AIC value, the better the model fit. The best-fit model according to AIC is the one that explains the greatest amount of variation using the fewest possible independent variables.

The coefficients in the output indicate the average change in **log odds** of a patient having PCOS. For example, a one unit increase in SKIN DARKENING is associated with an average increase of 1.89268 in the log odds of having PCOS.

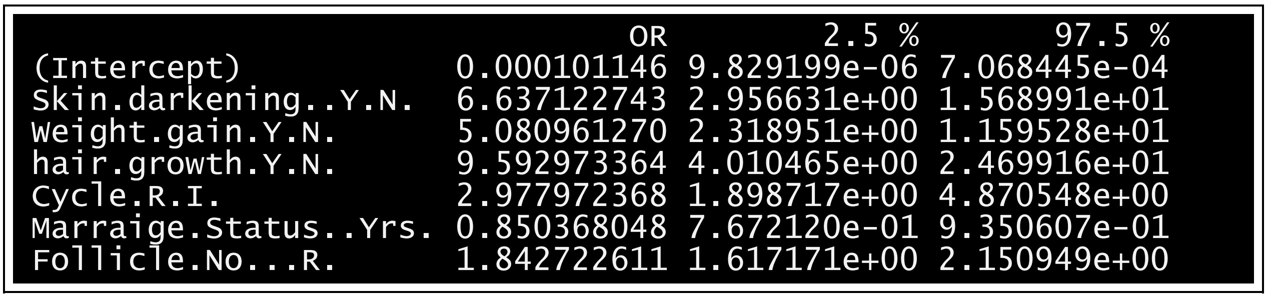
- The **deviance residuals vs. fitted values graph** is a diagnostic tool used to assess the goodness of fit of a logistic regression model. Here's how we can interpret Graph 6 for:

1. **Deviance Residuals**: The deviance residuals in logistic regression represent the difference between the observed outcomes (0 or 1) and the predicted probabilities from the model.
2. **Fitted Values**: The fitted values, also known as predicted probabilities, are calculated based on the logistic regression model's estimated coefficients and the predictor values of the observations.
3. **Shape of the Graph**: In the deviance residuals vs. fitted values graph, we're looking for specific patterns or outliers:
   - Symmetry around Zero: The majority of deviance residuals are symmetrically distributed around zero. This indicates that the model is capturing the underlying relationships well.
   - No Patterns: There is no noticeable patterns or trends in the residuals as the fitted values change. Patterns may indicate that the model is not capturing important features or that there is a violation of the logistic regression assumptions.



GRAPH 6: DEVIANCE RESIDUALS VS FITTED VALUES PLOT

- **Odds ratio and Confidence Interval**



```
                                 OR        2.5 %       97.5 %
(Intercept)           0.000101146 9.829199e-06 7.068445e-04
Skin.darkening..Y.N.  6.637122743 2.956631e+00 1.568991e+01
Weight.gain.Y.N.      5.080961270 2.318951e+00 1.159528e+01
hair.growth.Y.N.      9.592973364 4.010465e+00 2.469916e+01
Cycle.R.I.            2.977972368 1.898717e+00 4.870548e+00
Marraige.Status..Yrs. 0.850368048 7.672120e-01 9.350607e-01
Follicle.No...R.      1.842722611 1.617171e+00 2.150949e+00
```

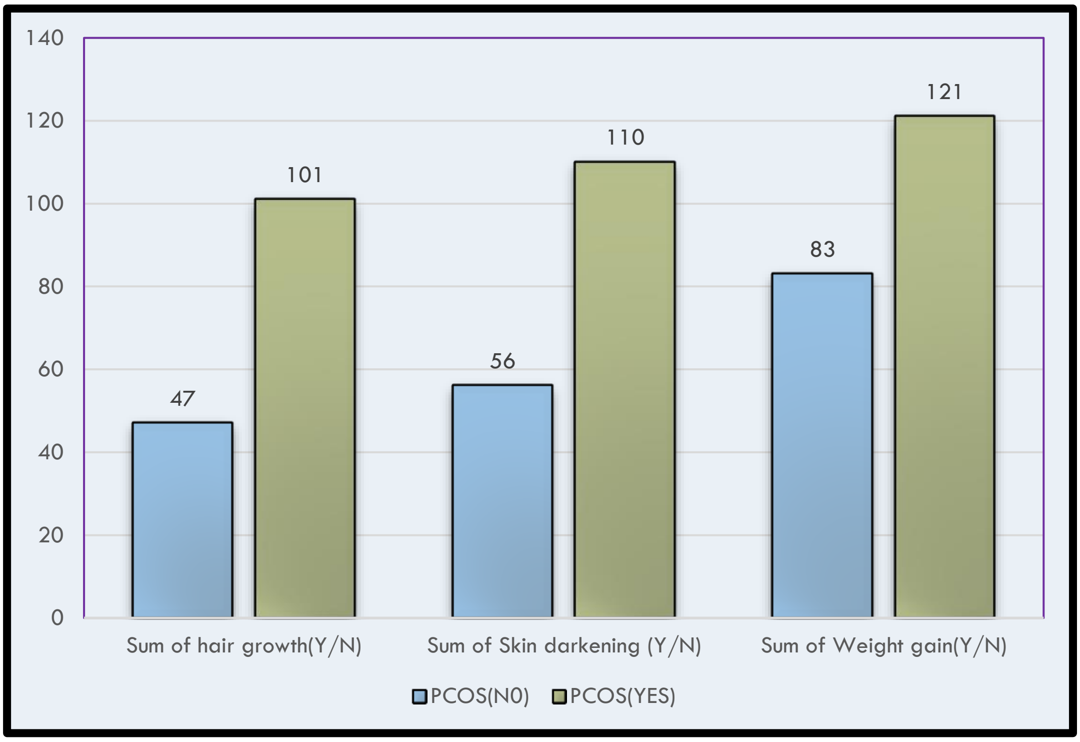
Interpretation: One unit increase in Hair growth, the odds of the patient having PCOS(versus not having PCOS) increase by factor of 9.59 approximately. Same happens with all other factors.

- **Diagrammatic representation of the significant factors from the dataset**

1) Representing the hair growth, skin darkening and weight gain factors using contingency table and their respective bar plots:

Table 2.1:

PCOS(Y/N)	0	1
Sum of hair growth(Y/N)	47	101
Sum of Skin darkening (Y/N)	56	110
Sum of Weight gain(Y/N)	83	121



GRAPH 7: COLUMN CHART SHOWING THE NO. OF PATIENTS WITH OR WITHOUT PCOS HAVING THE CONDITION OF SKIN DARKENING, HAIR GROWTH AND WEIGHT GAIN

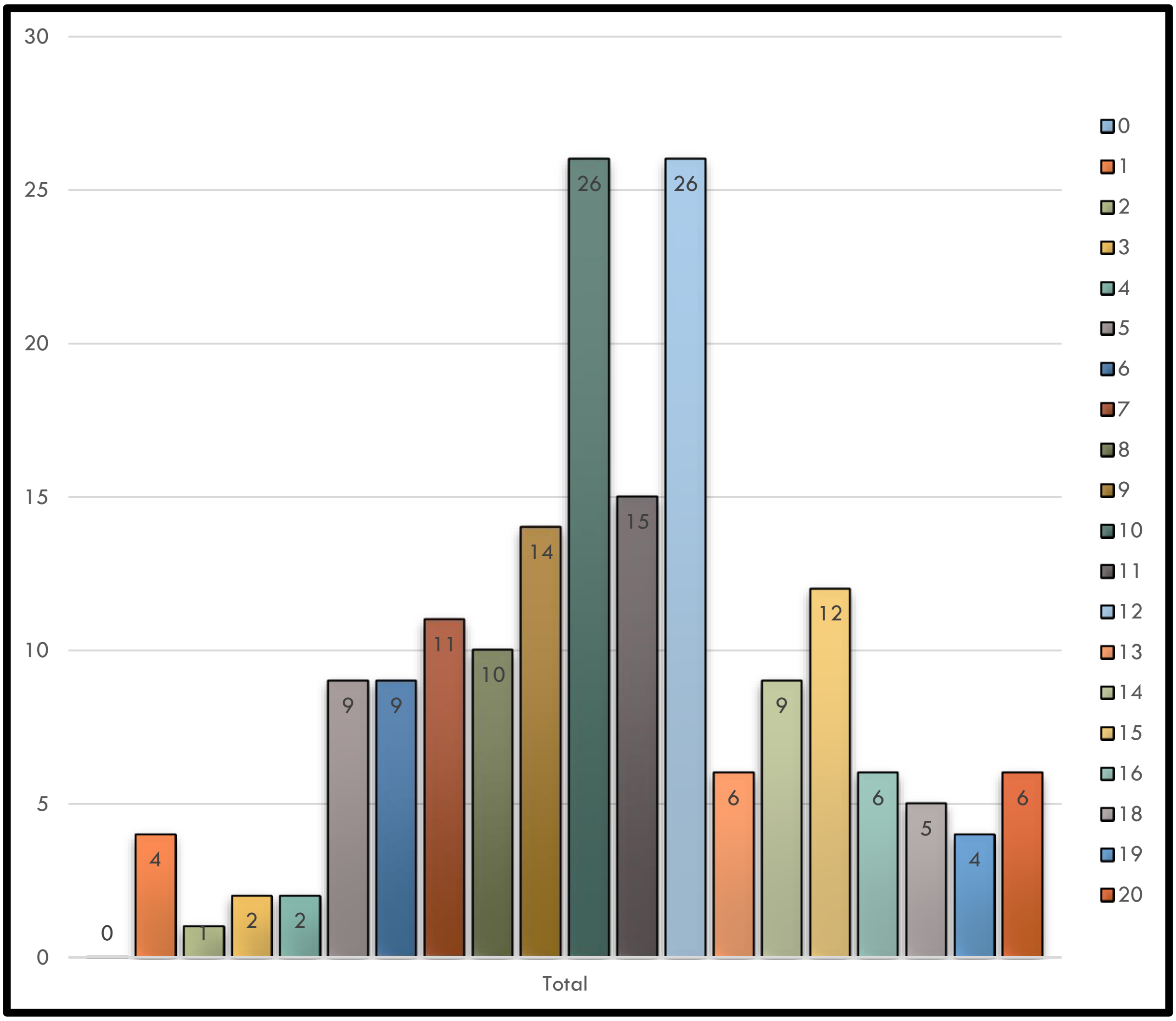
We can conclude that the patients having the skin darkening, hair growth and weight gain has higher chance of having PCOS.

2) Representation of follicle no.(R) factor using Bar plot:

The table below shows the follicle no. from the dataset and the total no. of patients having PCOS corresponding to that follicle no.

TABLE 3

Follicle No.(R)	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	18	19	20
Sum of PCOS (Y/N)	0	4	1	2	2	9	9	11	10	14	26	15	26	6	9	12	6	5	4	6



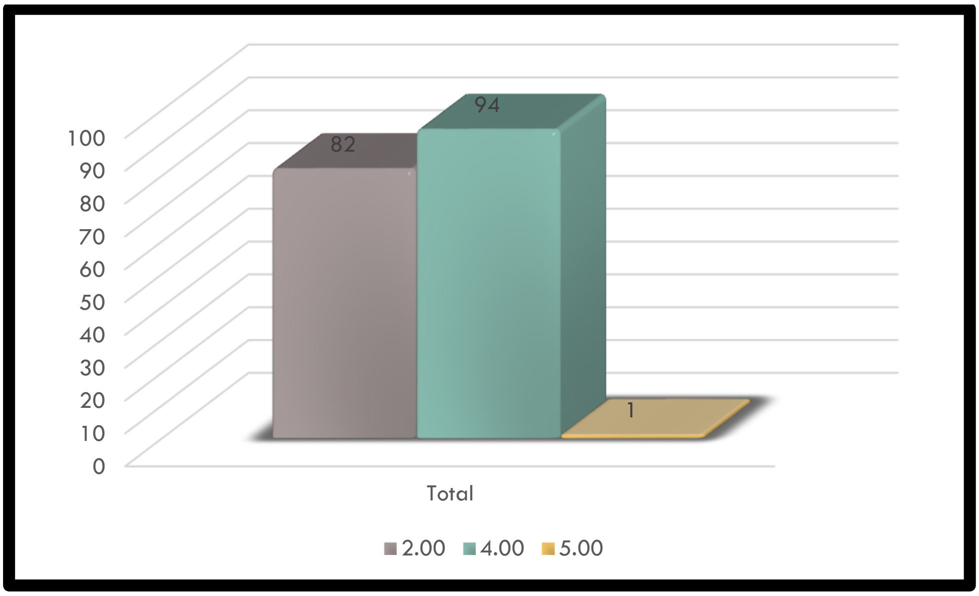
GRAPH 8: REPRESENTING THE DIFFERENT NO. OF FOLLICLES(R) AND THEIR CORRESPONDING PCOS(Y/N) CONDITION

We can conclude that the patients having follicle no.10 and 12 have the highest chance of having PCOS. Also, patients with follicle no. 7,8,9,11,15 also have a fair chance of having PCOS.

3) Representation of Cycle R.I factor using column chart:

TABLE 4

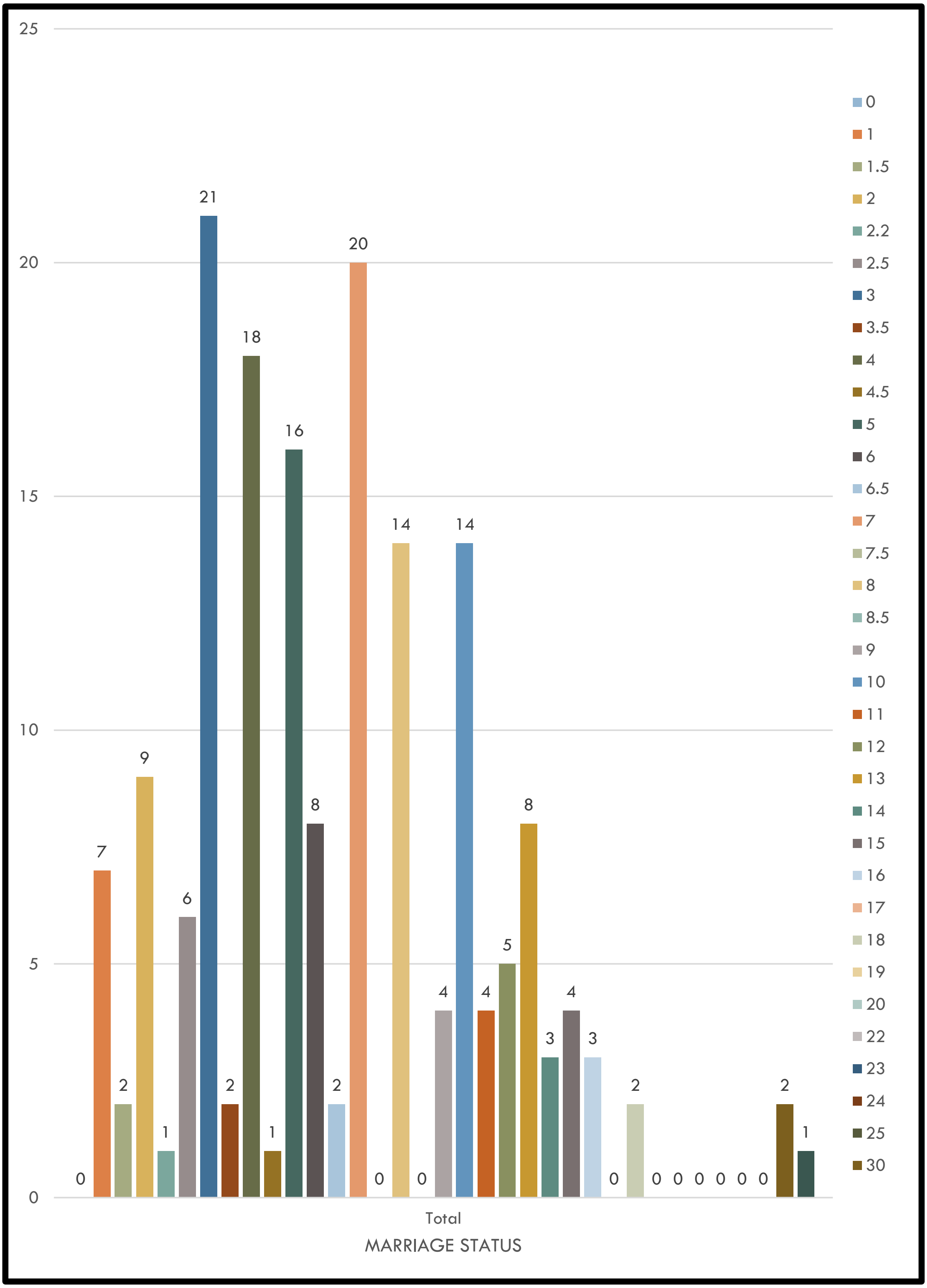
Cycle(R/I)		2	4	5	Grand Total
Sum of PCOS (Y/N)		82	94	1	177



GRAPH 9: CYCLE LENGTH AND THEIR CORRESPONDING PCOS CONDITION

4) Representation of Marriage Status(yrs) factor using bar chart :

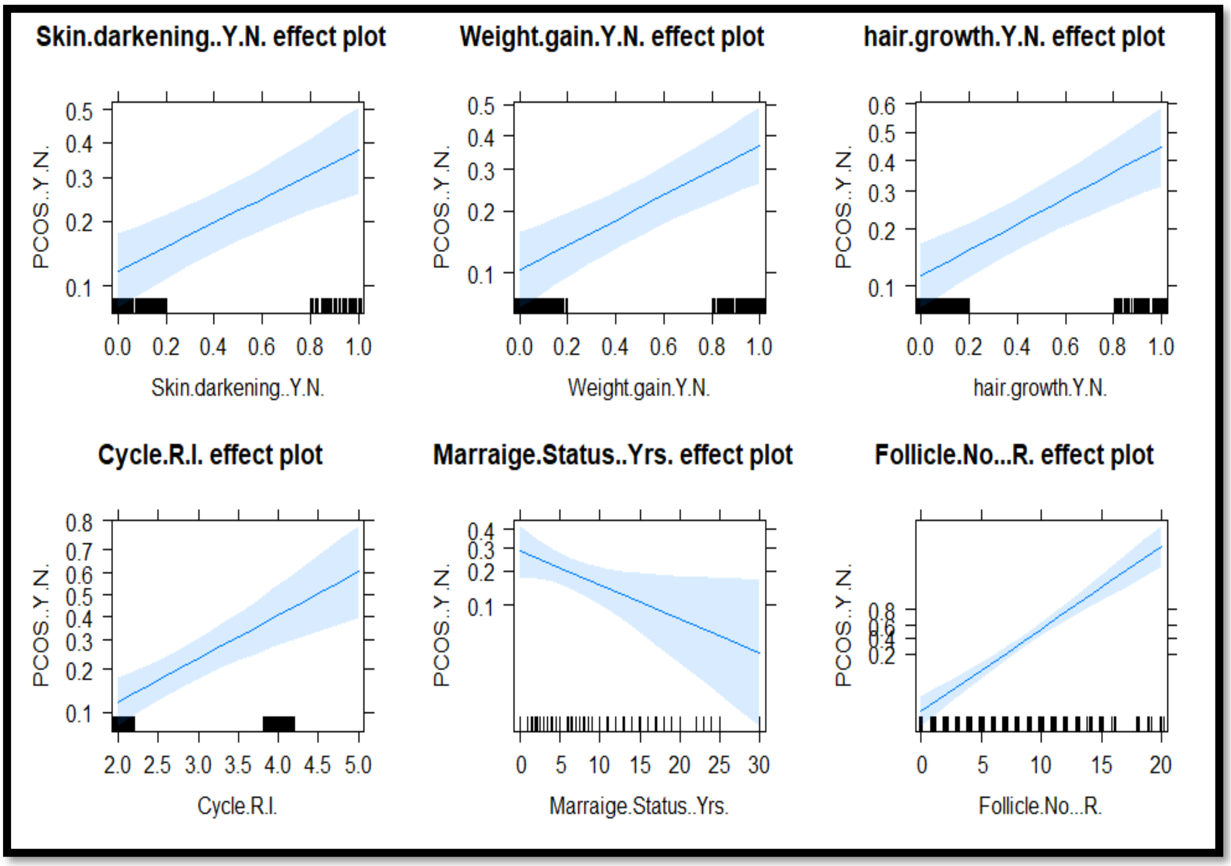
We can observe that the patients whose marriage status is 7 and 3 years have maximum PCOS, followed by patients whose marriage status is 4,5,8,10.



GRAPH 10: REPRESENTATING THE NO. OF YEARS A PATIENT HAS BEEN MARRIED AND THEIR CORRESPONDING PCOS CONDITION

4) Plotting each factor against PCOS variable to see their effects:

Statistical effect plots for the multivariable logistic regression model of PCOS with respect to the 6 significant factors is given below. Effect plots show the effect of varying the selected regressor variable on the response while holding all other regressors at their mean value. The tick marks along the x-axis of the plot show the distribution of samples. The light blue area surrounding the regression line indicates the 95% confidence interval of the model.

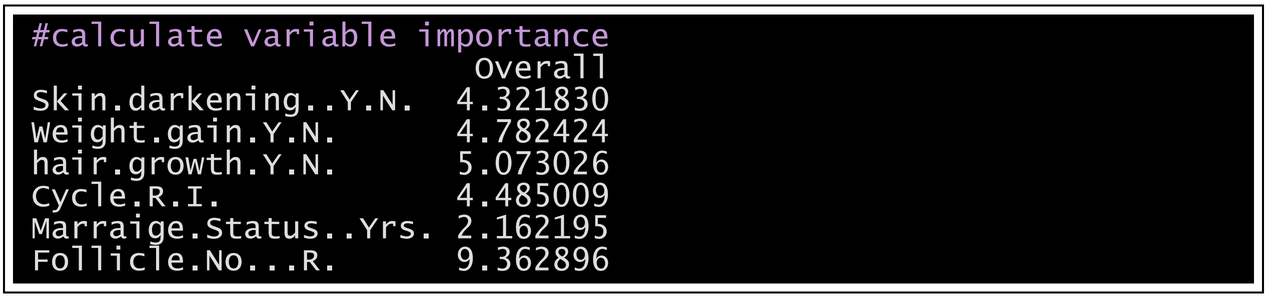


GRAPH 11: STATISTICAL EFFECTS PLOT

ASSESSING MODEL FIT

1) Variable Importance:

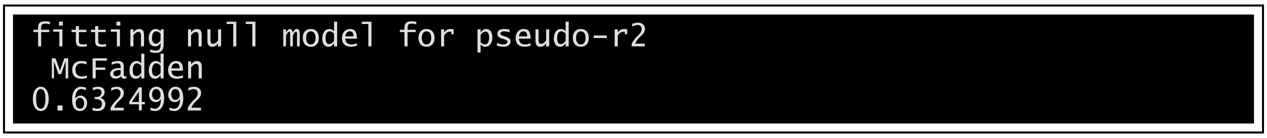
We can also compute the importance of each predictor variable in the model. Higher values indicate more importance. These results match up nicely with the p-values from the model. Follicle No. is by far the most important predictor variable, followed by hair growth and then Weight Gain, Cycle.R.I, Skin Darkening, Marriage Status.



```
#calculate variable importance
                          Overall
Skin.darkening..Y.N.   4.321830
Weight.gain.Y.N.       4.782424
hair.growth.Y.N.       5.073026
Cycle.R.I.             4.485009
Marraige.Status..Yrs.  2.162195
Follicle.No...R.       9.362896
```

2) Pseudo R^2:

In typical linear regression, we use R^2 as a way to assess how well a model fits the data. This number ranges from 0 to 1, with higher values indicating better model fit. However, there is no such R^2 value for logistic regression. Instead, we can compute a metric known as McFadden's R^2, which ranges from 0 to just under 1. Values close to 0 indicate that the model has no predictive power. In practice, values over 0.40 indicate that a model fits the data very well. Since the value for this model fit is 0.6 approximately, it can be considered a good fit.



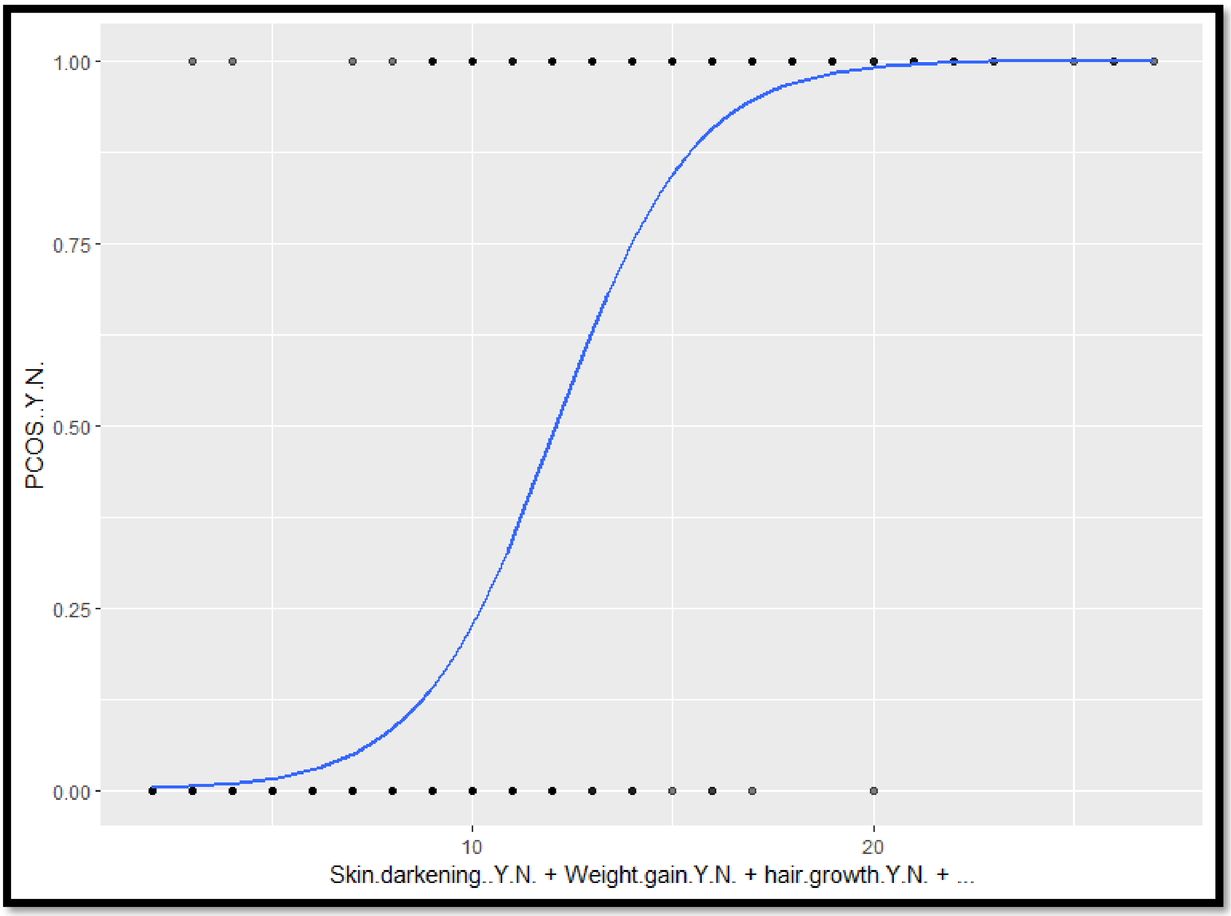
```
fitting null model for pseudo-r2
 McFadden
0.6324992
```

3) Fitting the logistic regression curve:

In Logistic Regression, we don’t directly fit a straight line to our data like in linear regression. Instead, we fit a S shaped curve, called *Sigmoid*, to our observations.

Firstly, Logistic Regression models are classification models; specifically binary classification models (they can only be used to distinguish between 2 different categories — like if a person has PCOS or not given its 6 significant factors). This means that our data has two kinds of observations (Category 1 and Category 2 observations) like we can observe in the figure.

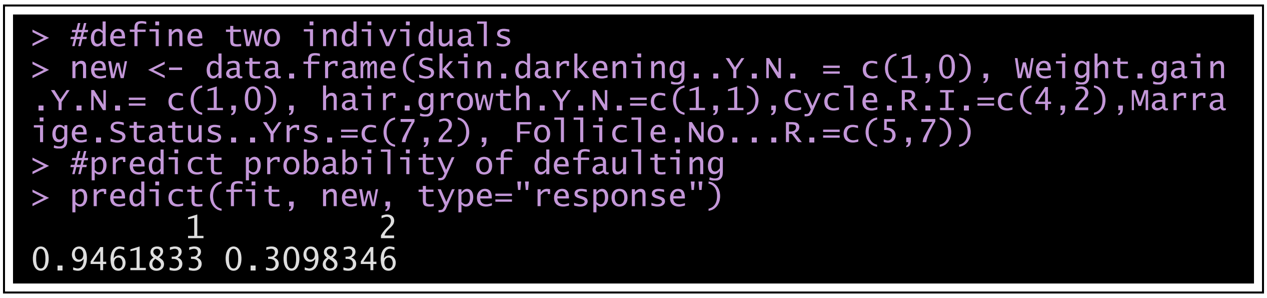
Secondly, as we can see, the Y-axis goes from 0 to 1. This is because the *sigmoid* function always takes as maximum and minimum these two values, and this fits very well our goal of classifying samples in two different categories. By computing the *sigmoid* function of X (that is a weighted sum of the input features), we get a probability (*between 0 and 1 obviously*) of an observation belonging to one of the two categories.



GRAPH 12: SIGMOID CURVE

4) Using the Model to Make Predictions:

Once we’ve fit the logistic regression model, we can then use it to make predictions about whether or not an individual will have PCOS based on their student Skin Darkening, hair growth, Weight Gain, Cycle.R.I, Marriage status, follicle no.(R):



```
> #define two individuals
> new <- data.frame(Skin.darkening..Y.N. = c(1,0), Weight.gain
.Y.N.= c(1,0), hair.growth.Y.N.=c(1,1),Cycle.R.I.=c(4,2),Marra
ige.Status..Yrs.=c(7,2), Follicle.No...R.=c(5,7))
> #predict probability of defaulting
> predict(fit, new, type="response")
        1         2 
0.9461833 0.3098346 
```

Interpretation:

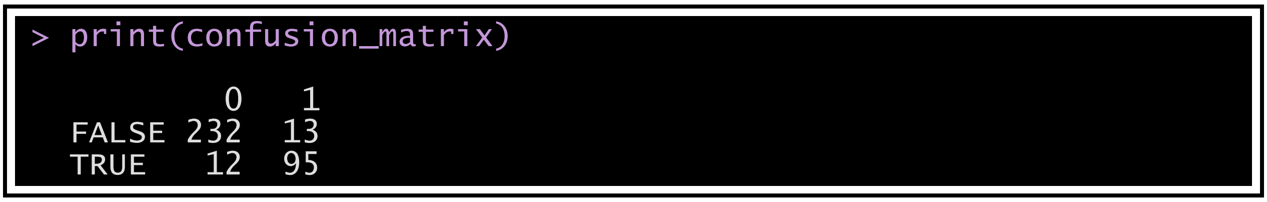
The probability of a patient , with darkened skin, having hair growth, gained weight, with cycle R.I 4 and marriage status of 7 years has a probability of having PCOS is 0.9461833. Similarly, an individual with no darkened skin, having hair growth, without weight gain, with cycle R.I 2 and marriage status of 2 years has a probability of having PCOS is 0.3098346.

MODEL DIAGNOSTICS

1) Confusion Matrix:

Lastly, we can analyse how well our model performs on the test dataset. By default, any individual in the test dataset with a probability of default greater than 0.5 will be predicted to have PCOS.

Using this threshold, we can create a confusion matrix which shows our predictions compared to the actual defaults:



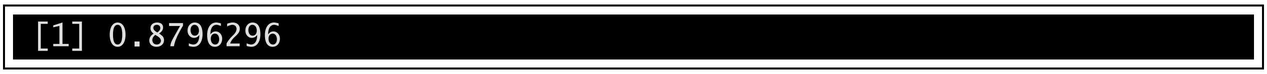
```
> print(confusion_matrix)

          0   1
  FALSE 232  13
  TRUE   12  95
```

Interpretation: It is a table with 4 different combinations of predicted and actual values.

4. True Positive: 95 patients are predicted to have PCOS and its true.
5. True Negative: 232 patients are predicted to be healthy but its false
6. False Positive: (Type 1 Error)12 patients are predicted to have PCOS but its false.
7. False Negative: (Type 2 Error)13 patients are predicted to be healthy but its false.

2) Sensitivity: measures how apt the model is to detecting events in the positive class. This means that about 87.96 % of the patients in the dataset were correctly predicted to have PCOS.

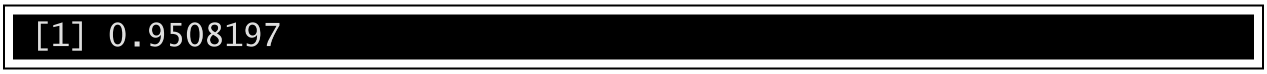
sensitivity = TP / (TP + FN)



```
[1] 0.8796296
```

3) Specificity: measures how exact the assignment to the positive class is, in this case, a patient having PCOS. The model reaches the specificity value of 0.9508197, so 5 % of all normal patients are predicted incorrectly to have PCOS.

Specificity= TN/(TN+FP)

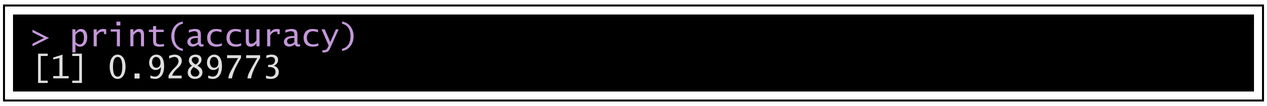


```
[1] 0.9508197
```

4) Accuracy of the model: It gives us the overall accuracy of the model, meaning the fraction of the total samples that were correctly classified by the classifier. The accuracy of the model is 0.9289,i.e., the model is 92.89% accurate to predict whether a patient has PCOS or not.

Accuracy= (TP+TN)/(TP+TN+FP+FN)

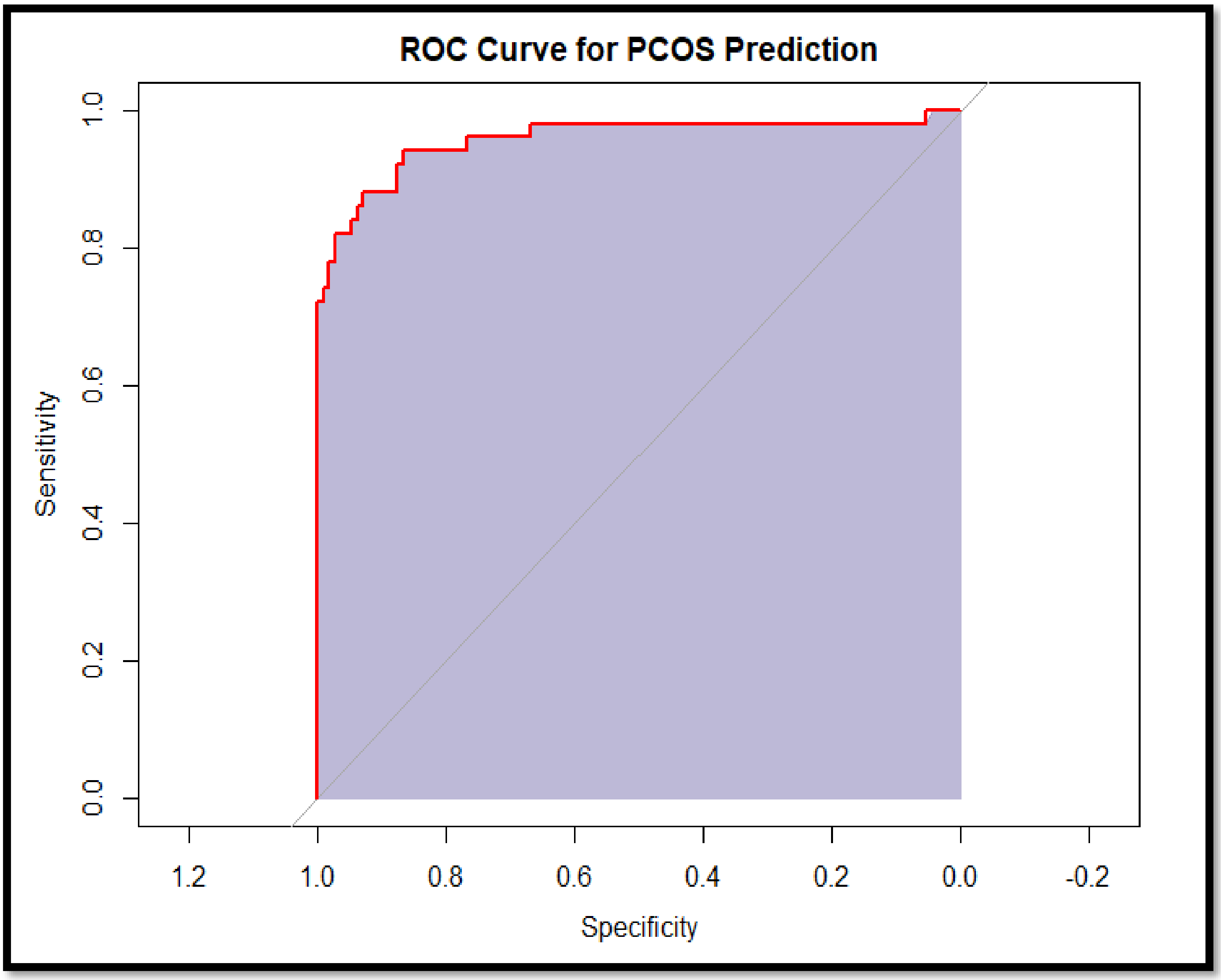
Also, the misclassification error is given by 1-accuracy,that is, 0.0710227. The total misclassification error rate is 7.1% for this model. In general, the lower this rate the better the model is able to predict outcomes, so this particular model turns out to be very good at predicting whether an individual has PCOS or not.



```
> print(accuracy)
[1] 0.9289773
```

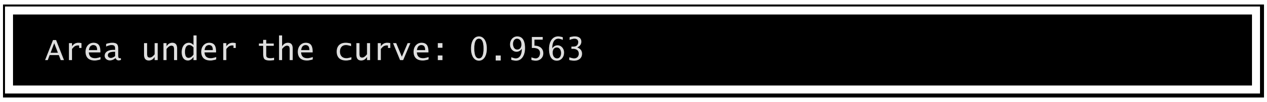
5) ROC curve:

Lastly, we can plot the ROC (Receiver Operating Characteristic) Curve which displays the percentage of true positives predicted by the model as the prediction probability cutoff is lowered from 1 to 0. The higher the AUC (area under the curve), the more accurately our model is able to predict outcomes. AUC provides an aggregate measure of performance across all possible classification thresholds.



GRAPH 13: ROC CURVE

Since AUC value is 0.9563, the model is 95.63% accurate to predict PCOS.



```
Area under the curve: 0.9563
```

CONCLUSION

Women needs to be aware about the changes taking place in their bodies. The delay in diagnosis of PCOS can lead to the progression of comorbidities. So, being aware of the causes and symptoms of PCOS can help a woman get early treatment and prevent further health complications, such as obesity, diabetes, heart disease, infertility, etc.

As we can see from the analysis of the training dataset, women need to watch out for the following symptoms:

Skin darkening, Weight Gain, Hair growth and no. of days their menstruation stay, and also take a test to find out the no. of follicles in their right ovary.

Also, from the test dataset we have come to the conclusion that our model has a AUC value 0.9563 which means that this dataset has the capability to predict the positive (having PCOS) patients as positive and the negative (not having PCOS) patients as negative with a probability of 0.9563.In other words, our model is 95.63% accurate to predict PCOS.

REFERENCES

- https://www.smart-academy.in/blog/pcos-its-awareness-and-prevention/
- https://www.ibm.com/topics/logistic-regression
- https://www.geeksforgeeks.org/understanding-logistic-regression/
- https://bookdown.org/egarpor/PM-UC3M/glm-diagnostics.html
- https://towardsdatascience.com/logistic-regression-explained-9ee73cede081
- https://www.statology.org/r-logistic-regression-odds-ratio/
- https://rstudio-pubs-static.s3.amazonaws.com/892643_8906d7f25fe94f9cafe9890f97f318e4.html
- https://www.kaggle.com/code/noelmat/pcos-data-cleaning-and-feature-importances/notebook
- https://www.statology.org/logistic-regression-in-r/
- https://www.hopkinsmedicine.org/health/conditions-and-diseases/polycystic-ovary-syndrome-pcos

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to all those who have contributed to the successful completion of my project titled "Prediction of Polycystic Ovarian Syndrome using Logistic Regression."

I would also like to extend my heartfelt appreciation to my guide and professor Dithi Bhattacharya for her support and encouragement in my project work.

I would like to express my gratitude to my friends for their help, support, and encouragement whenever I needed. Their belief in my abilities has been a driving force behind my academic pursuits.

In conclusion, this project has been a remarkable learning experience for me, and I am thankful to everyone who has played a part, directly or indirectly, in its successful completion. The knowledge and skills gained from this project will undoubtedly shape my future endeavors in the field of statistics and data analysis.

ANNEXURE

The following R codes have been used to get the results used in this project:

```
PCOS_DATA <- read.csv("PCOS_DATA.csv")
# Calculate the frequency table of a categorical variable
freq_table <- table(PCOS_DATA$PCOS..Y.N.)

# Create a bar plot of the frequency table
barplot (freq_table,
        xlab = "PCOS(YES/NO)", ylab = "Frequency",
        main = "Bar Plot of NO. OF PATIENTS HAVING PCOS")

#CORRELATION MATRIX
install.packages("corrr")
install.packages("ggcorrplot")
library(corrr)
library(ggcorrplot)
pcos_data=read.csv("PCAS.csv")
data_normalized <- scale(pcos_data)
head(data_normalized)
corr_matrix <- cor(data_normalized)
ggcorrplot(corr_matrix)


pcosraw <- read.csv("RAWPCOS.csv",header=TRUE)
glm.fit1 <- glm(PCOS..Y.N. ~
AMH.ng.mL.+Age..yrs.+Skin.darkening..Y.N.+Pimples.Y.N.+Weight.gain.Y.N.+
hair.growth.Y.N.+Weight..Kg.+Height.Cm.+Blood.Group+Pulse.rate.bpm.+RR..
breaths.min.+Hb.g.dl.+Cycle.R.I.+Cycle.length.days.+Marraige.Status..Yrs.+Pre
gnant.Y.N.+No..of.aborptions+I...beta.HCG.mIU.mL.++FSH.mIU.mL.+LH.mIU.
mL.+Hip.inch.+Waist.inch.+TSH..mIU.L.+PRL.ng.mL.+Vit.D3..ng.mL.+PRG.ng
.mL.+RBS.mg.dl.+Hair.loss.Y.N.+Fast.food..Y.N.+Reg.Exercise.Y.N.+BP._Systoli
c..mmHg.+BP._Diastolic..mmHg.+Follicle.No...L.+Follicle.No...R.+Avg..F.size..L.
..mm.+Avg..F.size..R...mm.+Endometrium..mm., family =
binomial(link="logit"), data = pcosraw)
```

```
summary(glm.fit1)

# Calculate Cook's distance
cooksd <- cooks.distance(glm.fit1)

# Identify observations with high Cook's distance
outliers <- which(cooksd > 30/length(cooksd))

# Load required libraries
library(dplyr)
library(stats)

# Load your PCOS dataset (replace "pcos_data.csv" with your dataset file name
and path)
pcos_data <- read.csv("rawpcos.csv", header = TRUE)

  # VIF VALUES
library(car)
car::vif(fit1)# multicollinearity check

# Calculate the deviance residuals
deviance_res <- residuals(fit, type = "deviance")



# Plot residuals vs. fitted values
plot(fit$fitted.values, deviance_res, pch = 19, frame = FALSE, xlab = "Fitted
Values", ylab = "Deviance Residuals")
abline(h = 0, col = "red", lty = 2)  # Add horizontal line at 0

# Identify influential observations using Cook's distance for logistic regression
cooksd <- cooks.distance(fit, type = "deviance")
cooksd
influential <- cooksd[(cooksd > (3 * mean(cooksd, na.rm = TRUE)))]
influential
```

```
# Plot Cook's distances
plot(cooksd, pch = 19, frame = FALSE, xlab = "Observation Index", ylab = "Cook's Distance")
abline(h = cutoff, col = "red", lty = 2)  # Add cutoff line
# Calculate standardized residuals
std_res <- rstandard(fit, type = "deviance")
std_res

# Calculate leverage values
leverage <- hatvalues(fit)
leverage
# Plot influence plot
plot(leverage,std_res, pch = 19, frame = FALSE, ylab = "Standardized Residuals", xlab = "Leverage")
abline(h = cutoff, col = "red", lty = 2)  # Add cutoff line
abline(h = cooksd, col = "red", lty = 2)  # Add cutoff line

# Determine the cutoff value for Cook's distance (e.g., 4/n, where n is the number of observations)
cutoff <- 4/nrow(pcos_data)  # Replace "pcos_data" with the name of your dataset

# Identify influential observations that exceed the cutoff
influential_obs <- which(cooksd > cutoff)
print(influential_obs)

# Remove influential observations from the dataset
cleaned_pcos_data <- pcos_data[-influential_obs, ]

# Print the cleaned dataset
print(cleaned_pcos_data)

#FITTING WITH SIGNIFICANT FACTORS
fit <- glm(PCOS..Y.N. ~
Skin.darkening..Y.N.+Weight.gain.Y.N.+hair.growth.Y.N.+Cycle.R.I.+Mar
raige.Status..Yrs.+Follicle.No...R., family = binomial(link="logit"), data =
cleaned_pcos_data)
summary(fit)

# ALL EFFECTS PLOT
install.packages("effects")
library(effects)
plot(allEffects(fit))


library(pscl)
pscl::pR2(fit)["McFadden"]# R square
caret::varImp(fit)# var importance


library(ggplot2)
#plot logistic regression curve
ggplot(fit,
aes(x=Skin.darkening..Y.N.+Weight.gain.Y.N.+hair.growth.Y.N.+Cycle.R.I.+Foll
icle.No...R., y=PCOS..Y.N.)) +
  geom_point(alpha=.5) +
  stat_smooth(method="glm", se=FALSE, method.args = list(family=binomial))


pcos_subset <- read.csv("rawpcos.csv",header=TRUE)
library(caret)
library(ggplot2)
library(lattice)
set.seed(123)
trainIndex <- createDataPartition(pcos_subset$PCOS..Y.N., p = 0.7, list =
FALSE)
train <- cleaned_pcos_data[trainIndex,]
test <- cleaned_pcos_data[-trainIndex,]


model <- glm(PCOS..Y.N. ~
Skin.darkening..Y.N.+Weight.gain.Y.N.+hair.growth.Y.N.+Cycle.R.I.+Marraige.
Status..Yrs.+Follicle.No...R. , data = test, family = "binomial")
exp(coef(model))
predicted <- predict(model, newdata = test, type = "response")
```

```
predicted

# Calculate the confusion matrix:
confusion_matrix <- table(predicted>0.5,train$PCOS..Y.N.)
print(confusion_matrix)
# Calculate the accuracy:
accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
print(accuracy)

# Calculate the AUC-ROC:
library(pROC)
roc_obj <- roc(test$PCOS..Y.N., predicted)
auc <- auc(roc_obj)
print(auc)

# Plot the ROC curve
plot(roc_obj, col = "red",type="shape" ,  auc.polygon = TRUE, auc.polygon.col=rgb(.35,0.31,0.61, alpha = 0.4),
auc.polygon.border=rgb(.35,0.31,0.61, 0.4), main = "ROC Curve for PCOS Prediction")

#define two individuals
new <- data.frame(Skin.darkening..Y.N. = c(1,0), Weight.gain.Y.N.= c(1,0), hair.growth.Y.N.=c(1,1),Cycle.R.I.=c(4,2),Marraige.Status..Yrs.=c(7,2),Follicle.No...R.=c(5,7))
#predict probability of defaulting
predict(fit, new, type="response")

# Calculate OR and confidence intervals
exp(cbind(OR=coef(fit),confint(fit)))
```