Yiduo Feng

CS 555

project

Date: 8/18/2022

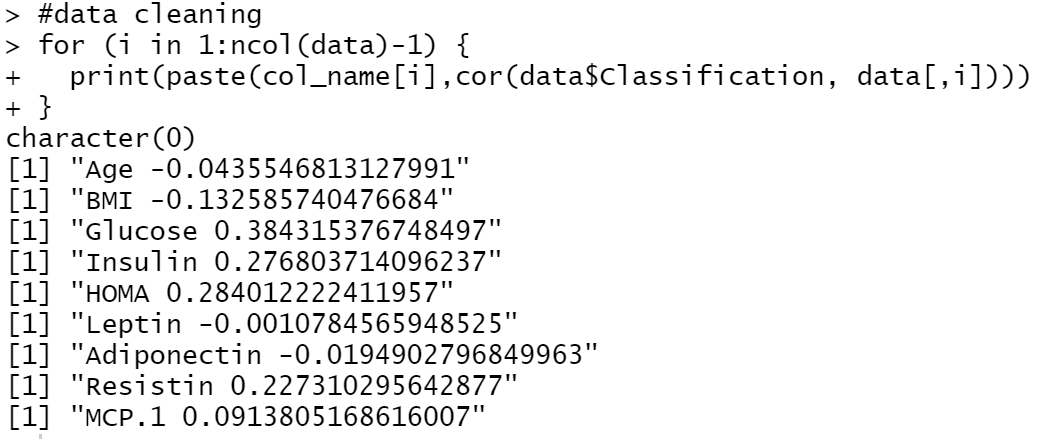
**Analysis of Breast Cancer Coimbra Data**

CS 555 Final Project (Prof. Heather Shappell, Summer 2022)

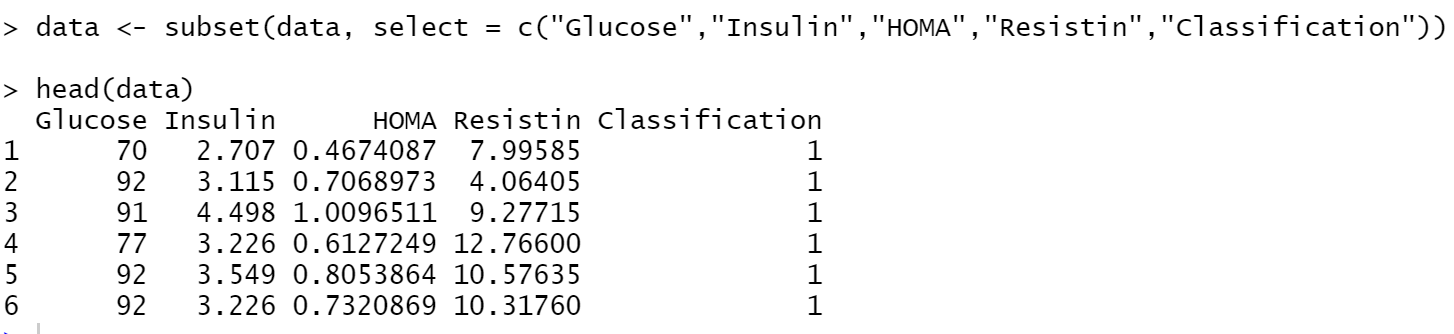
**Dataset Details**

Due to the high cost of hospitals, long queues and the lack of professional medical personnel, the difficulty of seeking medical treatment has become a huge problem in contemporary society. At the same time, with the development of machine learning, reducing human waste and life automation is the main development direction at present. Therefore, I want to help people reduce the pressure of doctors by learning knowledge, and let patients directly predict their condition through machine learning. This data set provide 64 patients with breast cancer and 52 healthy controls(1=Healthy controls, 2=Patients), and there are 1 column for class/label and the other 9 predictors. The predictors include Age (years), BMI (kg/m2), Glucose (mg/dL) , Insulin (µU/mL), HOMA, Leptin (ng/mL), Adiponectin (µg/mL), Resistin (ng/mL), MCP-1(pg/dL).

* **Data cleaning**



For this analysis, I plan to use the top 4 predictors with strongest correlation with class. According to the results above, I will only keep Glucose, Insulin, HOMA and Resistin. After cleaning data, we get the data shown below.

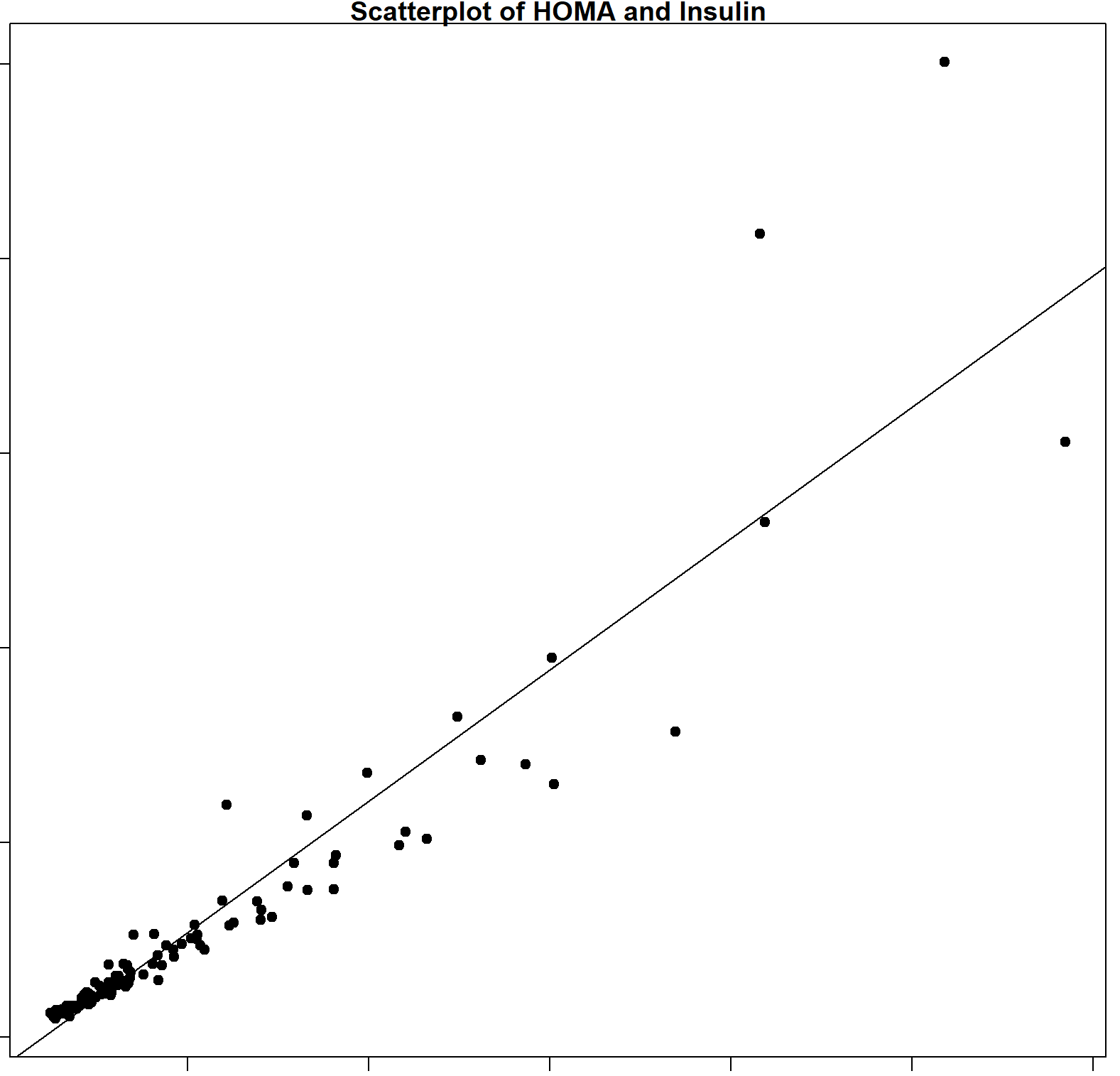


**Objective and Question**

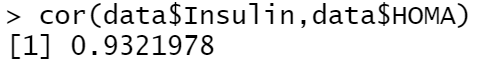
Breast cancer is one of the diseases that is difficult to detect in advance because the accompanying symptoms are not obvious. And non-medical professionals struggle to link Glucose, Insulin, HOMA, and Resistin to breast cancer through physical Resistin data alone. In this project I will study the extent to which these three factors affect breast cancer, and how they affect breast cancer.

**Graphical analysis**

By reading the data roughly in csv, we can find that the people with cancer, they will have high Insulin and high HOMA at the same time.



Though looking at the scatter plot above, we can see a upward trend in the graph. The Insulin increase HOMA, so the direction is positive. The form is linear, and it is strong relationship, and we can see that there is few outlier at the right top.



By using R, the correlation coefficient is 0.9321978. The greater the absolute value of the correlation coefficient, the stronger the correlation, the closer the correlation coefficient is to 1 or -1, the stronger the correlation, the closer the correlation coefficient is to 0, the weaker the correlation. 0.9321978 shows a large correlation coefficient and close to 1, so this correlation is strong.

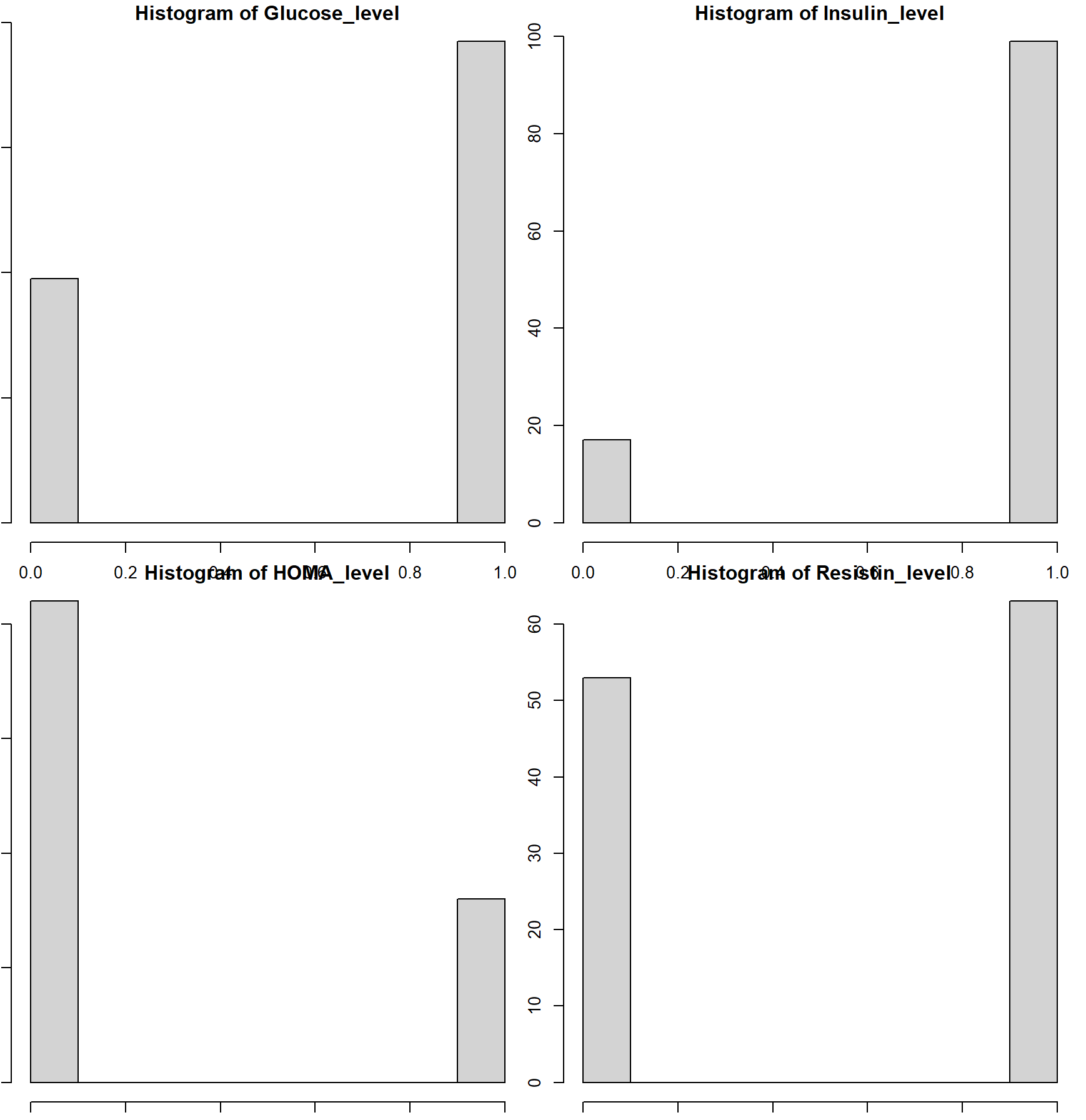
According to the normal level standard for 4 predictors.

Normally, the serum concentration of resistin in humans ranges from 7 to 22 ng·mL−1. In humans, resistin is primarily produced by cell populations other than adipocytes, which include peripheral blood mononuclear cells (PBMCs), macrophages and bone marrow cells (Fain et al., 2003; Patel et al., 2003).

HOMA-IR is less than 1.

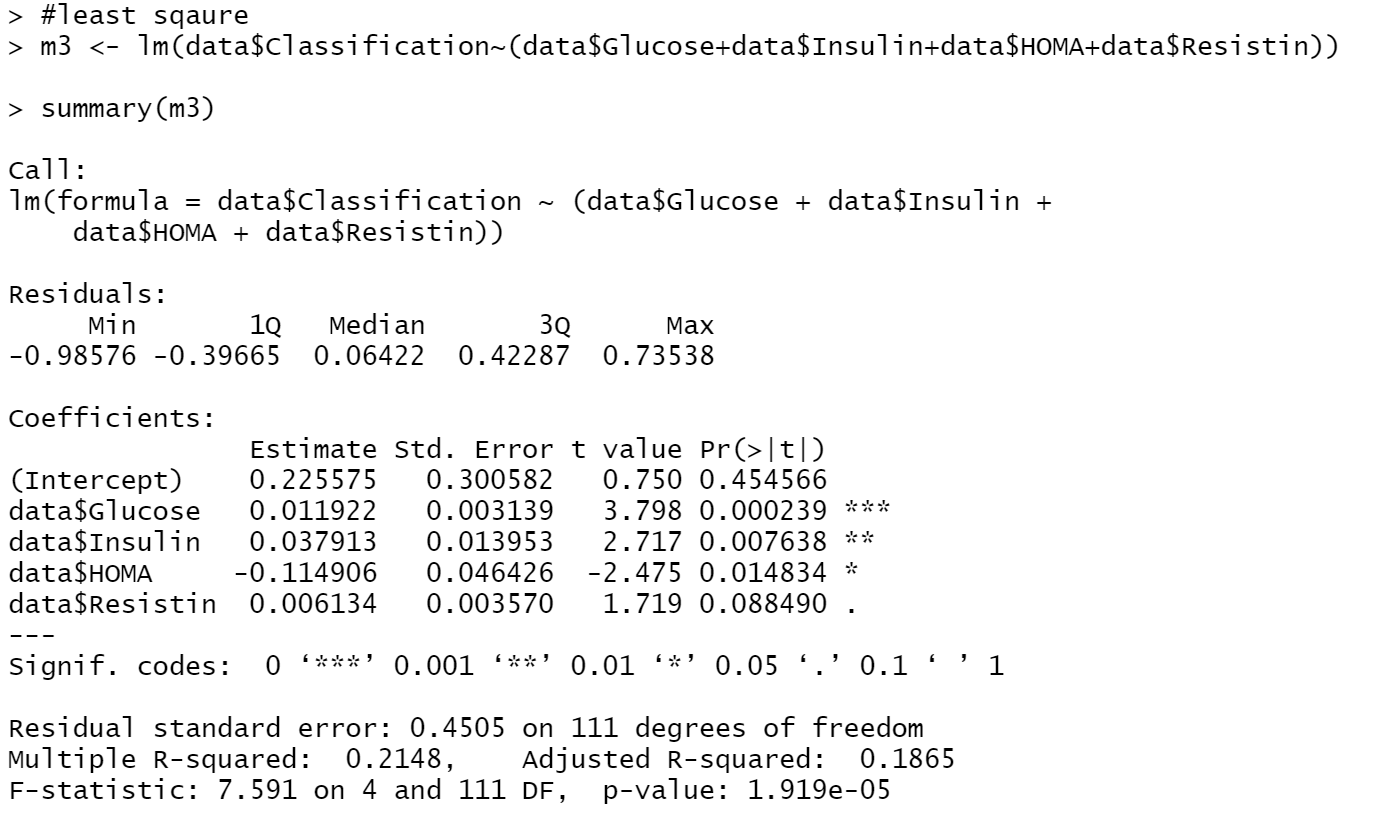
A normal measurement of free insulin is less than 17 mcU/mL.

Then we can get histogram for each of them:



**least squares regression**

question: whether the set of predictors are associated with cancer at the α = 0.05 level



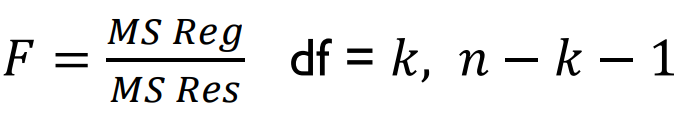
1. Set up the hypotheses and select the alpha level

H0 : 𝛽i= 0 (the set of these predictors are not associated with cancer)

H1 : 𝛽i≠ 0 (the set of these predictors are associated with cancer)

𝛼 = 0.01

1. Select the appropriate test-statistic



1. State the decision rule

Decision Rule: Reject H0 if 𝑝 ≤ 𝛼. Otherwise, do not reject H

1. Compute the test statistic

According to the code above,



P < 0.05

1. Conclusion

Reject H0

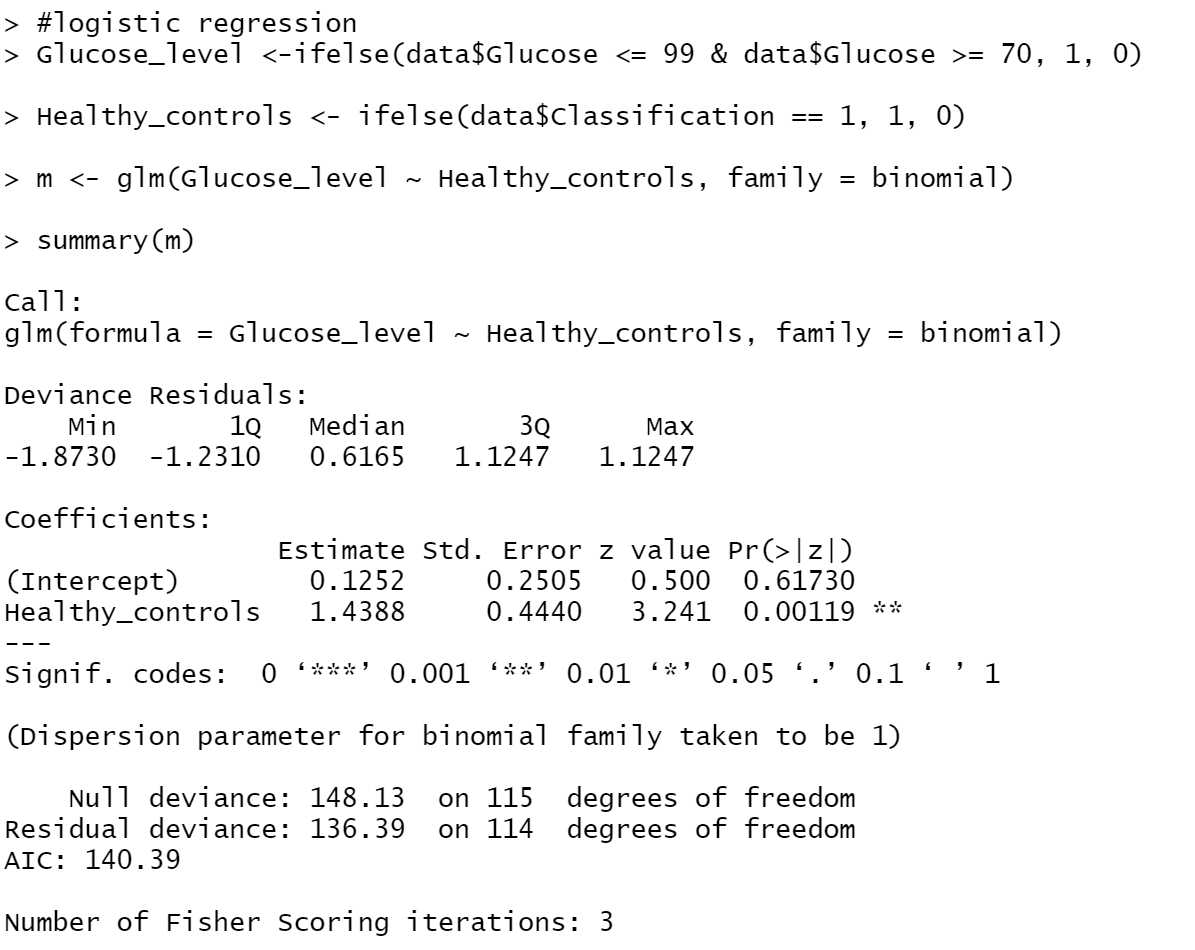
since 𝑝 ≤ 𝛼. We have significant evidence at the 𝛼 = 0.01 level

that these predictors when taken together are predictive of cancer.

That is, there is evidence of a linear association between breast cancer and Glucose, Insulin, HOMA, and Resistin. (here, p < 0.001).

**Logistic Regression**

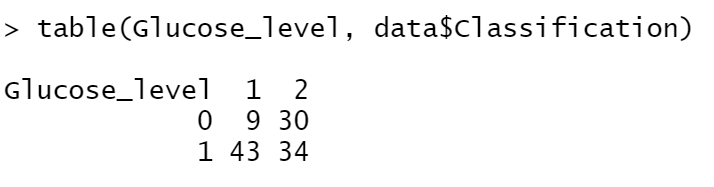
According to the standard of Glucose from Cleveland clinic website, ”Normal blood glucose level (while fasting) range within 70 to 99 mg/dL (3.9 to 5.5 mmol/L).” So we will find if At first, we perform a logistic regression with Glucose.



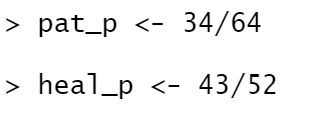
For question if the odds of having a glucose in normal level is the same between patients and health controls. We can see the z value is 3.241 and p value is 0.00119, for alpha = 0.05 level, p is smaller than alpha. In 5-step method, we will reject H0 and have significant evident that the beta 1 is not zero, and there is sufficient evident that there is association between glucose and cancer.

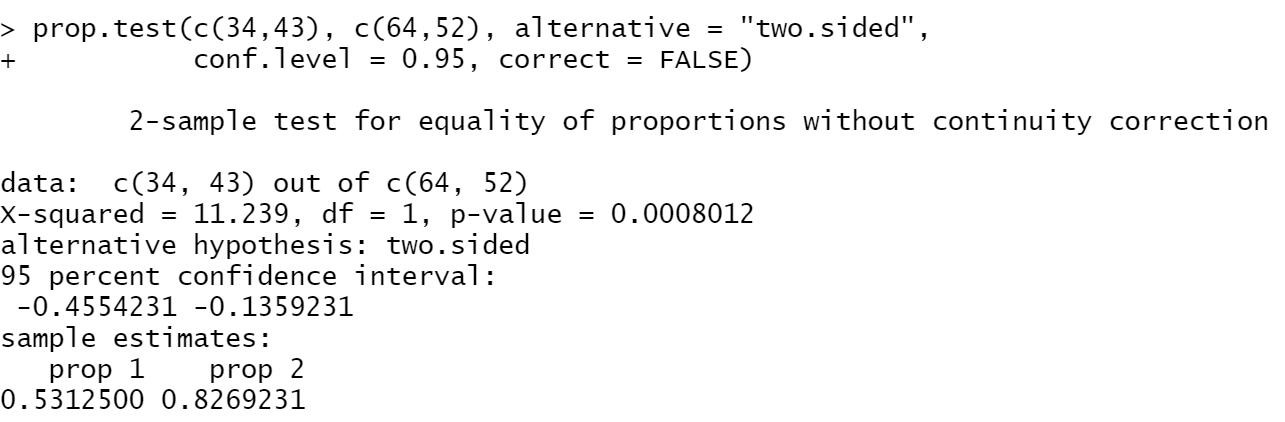
* **Risk difference**

Question: whether the proportion of people with normal glucose level is the same between patients and health controls.



So we can get the proportion for patient and the healthy:





So **risk difference** = 0.8269231 - 0.5312500= 0.2956731

5-step method:

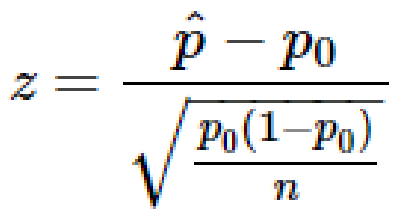
1. Set up the hypotheses and select the alpha level

H0 : 𝛽1= 0 (the proportion of people with normal glucose level is the same between patients and health controls..)

H1 : 𝛽1≠ 0 (the proportion of people with normal glucose level is not the same between patients and health controls..)

𝛼 = 0.05

1. Select the appropriate test-statistic



1. State the decision rule

Decision Rule: Reject H0 if 𝑝 ≤ 𝛼. Otherwise, do not reject H

1. Compute the test statistic

According to the code above,



P < 0.05

1. Conclusion

Reject H0

since 𝑝 ≤ 𝛼. We have significant evidence at the 𝛼 = 0.05 level

that the proportion of people with normal glucose level is not the same between patients and health controls. That is, there is evidence that the proportion of people with normal glucose level is not the same between patients and health controls.

**Conclusion**

Glucose, Insulin, HOMA, and Resistin show no relation to breast cancer in most people, but the analysis above shows that: There is evidence of a linear association between breast cancer and Glucose, Insulin, HOMA, and Resistin. And these predictors also have strong direct correlations, such as Insulin and HOMA. Their direct correlation is as high as 93 percent. In addition, I also found that after breast cancer, the substance indicators in the body are generally higher, which should be caused by the hormone and cell activity promoted by cancer cells. Because people's physical conditions are different, it is not excluded that some patients have other diseases at the same time, so this also makes our analysis somewhat limited.

**References**

[1]<https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra>

[2]<https://my.clevelandclinic.org/health/diagnostics/12363-blood-glucose-test#:~:text=A%20blood%20glucose%20test%20is,indicate%20pre%2Ddiabetes%20or%20diabetes>.

[3] Fain JN, Cheema PS, Bahouth SW, Lloyd Hiler M. Resistin release by human adipose tissue explants in primary culture. Biochem Biophys Res Commun. 2003;300:674–678. [PubMed] [Google Scholar]

[4]<https://labs.selfdecode.com/blog/homa-ir/#:~:text=Generally%2C%20you%20have%20optimal%20insulin,2.9%20signal%20significant%20insulin%20resistance>.

[5]https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=167&contentid=insulin\_total\_free