International Journal of Scientific and Management Research
Volume 5 Issue 12 (December) 2022
ISSN: 2581-6888
Page: 150-163

Heteroscedasticity Detection in Cross-Sectional Diabetes Pedigree Function: A Comparison of Breusch-Pagan-Godfrey, Harvey and Glejser Tests

Omotayo Oluwatosin ILORI[1*], & Fatai Olalekan TANIMOWO[2]
[1,*]Department of Mathematics and Statistics, Austin Peay State University, Clarksville, Tennessee, **USA.**
[2]Department of Mathematics and Statistics, Austin Peay State University, Clarksville, Tennessee, **USA.**
DOI - http://doi.org/10.37502/IJSMR.2022.51211

Abstract

Diabetes is a serious defect that does not make the body to have enough insulin, and thereby allowing blood sugar to stay in the bloodstream more than the body requires and over time causes serious problems relating to health. So, predicting if a person has diabetes or not using the linear model surface, but a major challenge arises if there is heteroscedasticity in the model, which can make the least square estimates inefficient. So, there is a need to know the method that is best for detecting heteroscedasticity so as not to rely on inefficient model for predicting diabetes. This research therefore aimed at comparing the Breusch-Pagan-Godfrey (BPG), Harvey and Glejser tests for detecting heteroscedasticity in cross-sectional data. To achieve this, data were collected on Diabetes Pedigree Function (DPF), Plasma glucose concentration a 2 hours in an oral glucose tolerance test (G), 2-Hour serum insulin (mu U/ml) (I), and Triceps skin fold thickness (mm) (S) from National Institute of Diabetes and Digestive and Kidney Diseases (1990) comprising 768 observations. The data was divided into two, small sample and large sample. The result of the regression analysis showed that skin fold thickness is the most important factor that can predict diabetes in a patient, followed by plasma glucose concentration, and then by insulin. The result for heteroscedasticity showed that, heteroscedasticity is not present in small dataset using the three tests. However, for the large sample, both the Breusch-Pagan-Godfrey and Glejser detect heteroscedasticity, but Harvey did not. Hence, it is advisable to use either Breusch-pagan or Glejser tests because they are more sensitive to heteroscedasticity in diabetes patient data.

Keywords: Cross-sectional, Diabetes pedigree function, Heteroscedasticity, Ordinary Least square, Residual.

1. Introduction

Globally, according to the Center for Disease Control and Prevention (CDC), diabetes is a chronic (long-lasting) health conditions that affects how the body turns food into energy, thereby resulting to build-up of extra sugar in the bloodstream. When the blood sugar level in the body goes up, it signals the pancreas to release insulin. Insulin acts like a key to let the blood sugar into the body's cells for use as energy. With diabetes, the body does not make

enough insulin or cannot even use it as well as it should, and when the insulin in the body is not enough or when cells stop responding to insulin, too much blood sugar stays in the bloodstream. Over time, that can cause serious health problems, such as heart disease, vision loss, and kidney disease (CDC, 2022). Also, with diabetes, open wound are not easily healed. This can result to secondary disease or illness from the wound. Another problem arises when the pancreas can no longer breakdown sugar in the body system. The severity of diabetes depends on the type.

Type 1, Type 2 and gestational are the three types of diabetes. The stage before Type 2 diabetes, where the blood glucose levels are higher than normal but not high enough to be officially diagnosed with Type 2 diabetes (Wu et al., 2014) is called prediabetes. In the US, more than 90 million people above 18 years old account for about 30% prediabetes. More than 80% of them do not even know they have diabetes. Blood sugar level are higher than normal in prediabetes, but this level of sugar is not sufficient to be diagnosed of Type 2 diabetes. Prediabetes raises the risk for Type 2 diabetes, heart disease, and stroke, but there is good news. If an individual has prediabetes, a CDC-recognized lifestyle change program can be helpful for such individual to reverse the health issues, such as losing weight, eating healthy food, and being active (Zou et al., 2018). Most of these diagnosis and pharmaceutical, but there are non-pharmaceutical methods, which are mostly mathematical or statistical in nature where static, stochastic or dynamic model is used to predict whether a patient has diabetes or not using some predictor variables (Abdulhadi and Al-Mousa, 2021; Deysi, 2022).

Thus, the need to diagnose or predict a patient with diabetes, allows us to use the linear model. Literature estimation techniques using linear model are presented as those that allow us to take heterogeneity of variables into account, and provide more accuracy. The impact of some factors on diabetes may indeed be better identified when successive waves of the overall level of the factors are taken into account. The usual assumptions of homoscedasticity, disturbances and fixed coefficients may be violated in some applications of the linear model. When these assumptions are violated, loss of efficiency in using ordinary least squares (OLS) may be significant and, more importantly, the biases in standard errors estimated may lead to invalid references (Golfield and Quandt, 1965). This has caused researchers to propose models, which relax these conditions and to devise estimators for their more general specifications, example is Golfield and Quandt (1965) for heteroscedasticity and Hildreth and Houck (1968) for random coefficients. For more explanation on heteroscedasticity, see Alabi et al. (2020).

However, because the effect of introducing random coefficient variation is to assign a different variance to the dependent variable at each observation, models with this feature can be considered as particular heteroscedastic formulations for detecting departure from the standard linear model. Recall, the following assumptions have to hold for parameters estimates and regression inference to be correct. Firstly, the model must be correctly specified, secondly, the error term must have zero mean, error term mush have constant variance (homoscedasticity), the error terms should not be correlated (no autocorrelation), the predictor variables are fixed in repeated samples, and there should not be high linear relationship among the predictor variables (no multicollinearity) (Alabi et al., 2020). When assumption of homoscedasticity holds, the errors term in the regression model have constant variance, meaning there is no

heteroscedasticity, otherwise, there is heteroscedasticity (Kennedy, 1998). This paper is motivated as a result of detecting heteroscedasticity in a cross-sectional data, especially when predicting health related response variables; either for small or large sample is germane, in order not to be predicting health issues with inefficient model.

In this paper, a null hypothesis of homoscedasticity is stated against the alternative hypothesis of heteroscedasticity. If the test is significant at 5%, then we conclude that there is presence of heteroscedasticity in the dataset, otherwise we conclude that the data is free of heteroscedasticity. Breusch-Pagan-Godfrey, Harvey and Glejser tests are used to detect heteroscedasticity using the F-test and Chi-square statistic in the cross-sectional diabetes dataset. The dependent variable is the diabetes pedigree function used as a measure of diabetes. The predictors are Plasma glucose concentration a 2 hours in an oral glucose tolerance test (G), 2-Hour serum insulin (mu U/ml) (I), and Triceps skin fold thickness (mm) (S). This paper will be very useful to health and medical practitioners, and other users of statistics, especially in the area of linear modelling, epidemiology, biostatistics and public health researchers. It can also be applied when modelling infectious disease using static or stochastic models.

2. Materials and Methods

2.1 Data Description and Source

The data used for this research is a cross-sectional data harvested from National Institute of Diabetes and Digestive and Kidney Diseases (1990) comprising 768 observations. The important variables of interest are Diabetes Pedigree Function (DPF), Plasma glucose concentration a 2 hours in an oral glucose tolerance test (G), 2-Hour serum insulin (mu U/ml) (I), and Triceps skin fold thickness (mm) (S). The data is divided into two, small sample and large sample. The small sample comprises 20 observations selected randomly using uniform distribution from 768 observations, while the large sample is the 768 observations. The R version 4.2.1 was used in running the analysis. See Appendix I for the R code. Similar data can be used to replicate the process, as heteroscedasticity is not limited to diabetes data. The major method of estimating the parameters of a linear regression model, either simple or multiple is the Ordinary Least Square (OLS) method, because it is the best linear unbiased estimator among the class of unbiased estimators.

2.2 Ordinary Least Squares (OLS)

One of the methods of estimating the parameter of this model is and there are some assumptions underlying this model. One of the assumptions of the classical linear regression model is that the variance of the error term must be equal (Homoscedasticity). A violation of this assumption leads to a problem of Heteroscedasticity. The present of heteroscedasticity in a data set can be known using various tests in R. This set of tests allows for a range of specifications of heteroscedasticity in the residuals of a model. OLS parameters estimates are consistent in the presence of heteroscedasticity, meaning that the errors can still tend to zero as the sample size increase, but the calculated standard errors cannot be valid anymore, and thereby making the OLS estimates inefficient. If an evidence of heteroscedasticity is found in a model, then choose the robust standard errors to correct the standard errors or use Weighted Least Squares (WLS)

to model the heteroscedasticity to obtain more efficient estimates or use a generalized linear model (Ekum et al., 2013, 2015; Alabi et al., 2020).

The m^{th} variable linear model is specified as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im} + e_i, i = 1,2, \ldots n \quad (1)$$

where y_i is the response variable for the *i*th observation, x_{ij} is the *i*th observation of the *j*th predictor variable , β_j is the unknown parameters to be estimated for the *j*th predictor variable, *n* is the number of observations, *m* is the number of predictor variables and e_i is the residual term, which is normally distributed.

The model specified for this paper is given as

$$\mathrm{DPF}_i = \beta_0 + \beta_1 G_i + \beta_2 I_i + \beta_3 S_i + e_i, i = 1,2, \ldots n \quad (2)$$

where DPF_i is the response variable for the *i*th observation, G_i, I_i and S_i are the predictor variables for the *i*th observations, $n = 20$ for small sample, and $n = 768$ for large sample. The number predictor variables $m = 3$.

Using OLS, the parameter estimates are given as

$$\hat{B} = (X'X)^{-1}X'Y \quad (3)$$

where *B* is a column matrix of the βs, *Y* is the vector of the response variable, $(X'X)^{-1}$ is the inverse of the design matrix, and X is a $n \times m$ matrix.

2.3 Weighted Least Squares (WLS)

Detecting heteroscedasticity in a model will not affect the bias or consistency properties of OLS estimates, but OLS will no longer be efficient and the conventional parameters estimates of the standard errors are no more valid. Suppose there is known heteroscedasticity with known variances $\sigma_i^2 > 0$ up to a positive scale factor, then WLS can be used to obtain efficient estimates that support valid inference. The WLS estimator for the parameters, β minimizes the weighted sum of squares residuals with respect to the *m* dimensional vector of parameters β, where the weights are proportional to the inverse conditional variances. Equivalently, the regression of the square root weighted transformed data can be estimated. In R, it is possible to instruct a weight, such as cross-section seemingly uncorrelated regression estimates. This weight is a feasible specification that can correct cross-section heteroscedasticity and make the parameters estimates to be valid and efficient (Baltagi, 1980; Okunnu et al., 2017).

2.4 Heteroscedasticity Tests

One of the assumptions of the OLS method is that the residual must have equal variance (homoscedasticity), and the violation of this assumption leads to a problem called heteroscedasticity. The set of tests adopted in this research allows for different test of specifications of heteroscedasticity in the residuals models. In R, you have the opportunity of

specifying different heteroscedasticity tests. Every test involves performing an auxiliary regression by the use of residuals from the original equation. These tests are available for parameter estimated by OLS, two-stage least squares (2LS), and nonlinear least squares (NLS). We compared the efficiency of three of these tests. The Ordinary least square (OLS) method is adopted to obtain the Regression model for the diabetes pedigree function, while the weighted least square is used to obtain a better model with better coefficient of determination (R^2). The individual tests are Breusch-Pagan-Godfrey (BPG), Harvey, Glejser, ARCH LM Test, White's Heteroscedasticity Test, but on the first three are used in this paper.

2.4.1 Breusch-Pagan-Godfrey (BPG)

The Breusch-Pagan-Godfrey test (Breusch and Pagan, 1979, 1980) states a null hypothesis of homoscedasticity versus alternative hypothesis of heteroscedasticity, which has the form $\sigma_t^2 = \sigma^2 h(z_t'\sigma)$, where z_t is a vector of predictor variables. Usually this vector contains the regressors from the original least squares regression, but it is not necessary. The test is performed by completing an auxiliary regression of the squared residuals from the original equation on $(1, z_t)$. The explained sum of squares from the auxiliary regression is divided by $2\hat{\sigma}^4$ to give an LM statistic, which follows a χ^2-distribution with degrees of freedom equal to the number of variables in *z* under the null hypothesis of no heteroscedasticity. Koenker (1981) proposed a statistic to be used given by nR^2, which is easier to compute, where *n* is the number of observations and R^2 is from the auxiliary regression. Koenker's statistic is asymptotically χ^2 distributed with ν degrees of freedom, where ν is the number of variables in *z*. Apart from these two statistics, R displays an *F*-statistic for a test of redundant variable for the joint significance of the variables in *z* for the auxiliary regression.

Given a multiple linear regression of *k* predictor vriables

$$\log(y_1) = b_1 + b_2\log(ip) + b_3tb_3 \quad (4)$$

and it is believed that there was heteroscedasticity in the residuals that depended on a function of log(ip) and tb_3, then the following auxiliary regression could be performed $e^2 = b_1 + b_2\log(ip) + b_3tb_3$ and to formally test for heteroscedasticity, a Breusch-Pagan test can be performed. The null hypothesis (H_0) is homoscedasticity is present, meaning residuals have equal variance, as against alternative hypothesis (H_1) of Heteroscedasticity is present, meaning unequal variance of the residuals.

2.4.2 Harvey

Harvey (1976) proposed a test for the detection of heteroscedasticity, which is similar to that of Breusch-Pagan-Godfrey test. The null hypothesis for Harvey test is homoscedasticity against heteroscedasticity having the form of $\sigma_t^2 = \exp(z_t\alpha)$, where, z_t, is a vector of predictor variables. To test for this form of heteroscedasticity, an auxiliary regression of the log of the original equation's squared residuals on (1, z_t) is performed. The LM statistic is then explained sum of squares from the auxiliary regression divided by $\varphi(0.5)$, the derivative of the log gamma function evaluated at 0.5. This statistic is asymptotically χ^2 distributed with ν degrees of freedom, where ν is equal to the number of variables in z. In summary, Harvey's test is the

fitting of an auxiliary regression model in which the response variable is equal to the log of the vector of squared residuals from the original model and the design matrix Z ′Z consists of one or more exogenous variables that are suspected of being related to the error variance. When prior information on a possible choice of Z ′Z is absent, it is advisable to use the predictor variables from the original model. It is a right-tailed test (Mittelhammer et al., 2000).

2.4.3 Glejser

Glejser (1969) proposed a test statistic for fitting an auxiliary regression model in which the response variable is the absolute residual from the original model and the design matrix Z ′Z consists of one or more exogenous variables that are suspected of being related to the error variance. When prior information is absent on a possible choice of Z ′Z, then the use of explanatory variables from the original model is advised. The null hypothesis is homoscedasticity against the alternative hypothesis of heteroscedasticity. It is also similar to the Breusch-Pagan-Godfrey test. The for of the heteroscedasticity is $\sigma_t^2 = (\sigma^2 + z_t\, a)^m$ with m = 1, 2. The auxiliary regression that Glejser proposes regresses the absolute value of the residuals from the original equation upon (1, z_t). An LM statistic can be formed by dividing the explained sum of squares from this auxiliary regression by $\left((1 - 2/\pi)\hat{\sigma}^2\right)$. It is also asymptotically chi-squared distributed with ν degrees of freedom, where ν is the number of parameters. It is a right-tailed test.

3. Results

The relationship between diabetes pedigree function and the predictor variables are depicted on scatter plots for the small and large samples. The presence of heteroscedasticity is tested on the diabetes data using Breusch-Pagan-Godfrey, Harvey and Glejser and their results are compared favourably for both small and large samples.

3.1 Exploratory Data Analysis

Reporting the Exploratory Data Analysis (EDA) of a dataset is a preliminary analysis that helps to reveal some hidden features of the dataset, such as its spread, skewness, kurtosis, averages, minimum and maximum values (Iluno et al., 2021).

Table 1. Descriptive Summary of Ulcer Data

	Diabetes Pedigree Function	Glucose	Insulin (mu U/ml)	Skin Thickness (mm)
Min.	0.0780	0.0000	0.0000	0.0000
1st Qu.	0.2437	99.000	0.0000	0.0000
Median	0.3725	117.000	30.5000	23.0000
Mean	0.4719	120.900	79.8000	20.5400
3rd Qu.	0.6262	140.200	127.2000	32.0000
Max.	2.4200	199.000	846.0000	99.0000
Std. Dev.	0.3313	31.9730	115.2440	15.9522
Skewness	1.9162	0.1730	2.2678	0.1092
Kurtosis	8.5508	3.6288	10.1596	2.4755

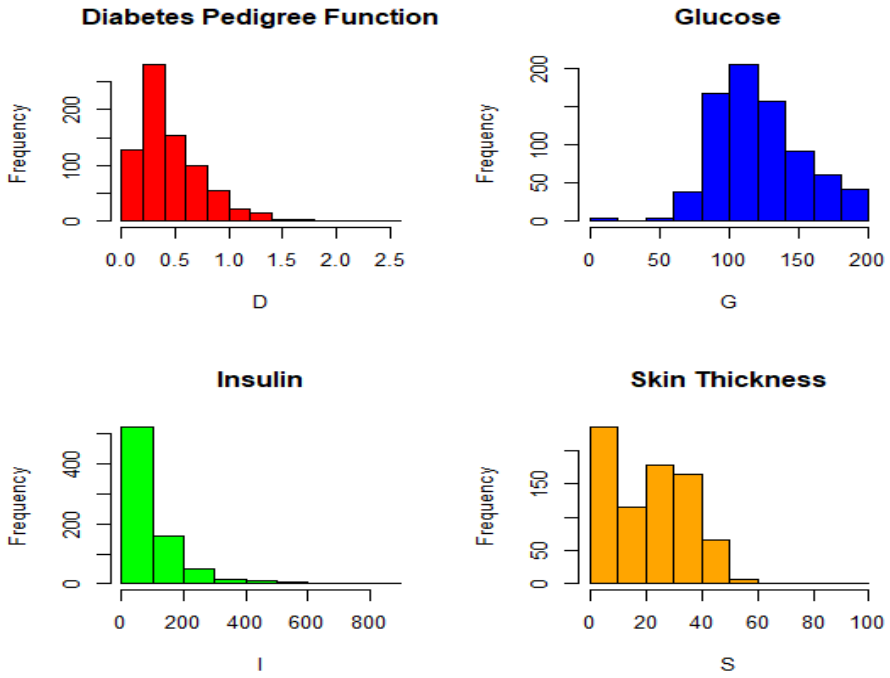

Figure 1. Histogram of the variables of interest

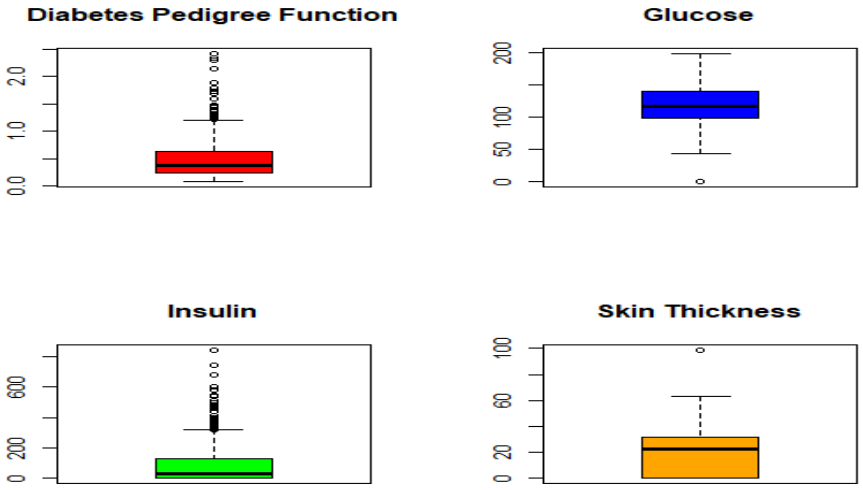

Figure 2. Boxplot of the variables of interest

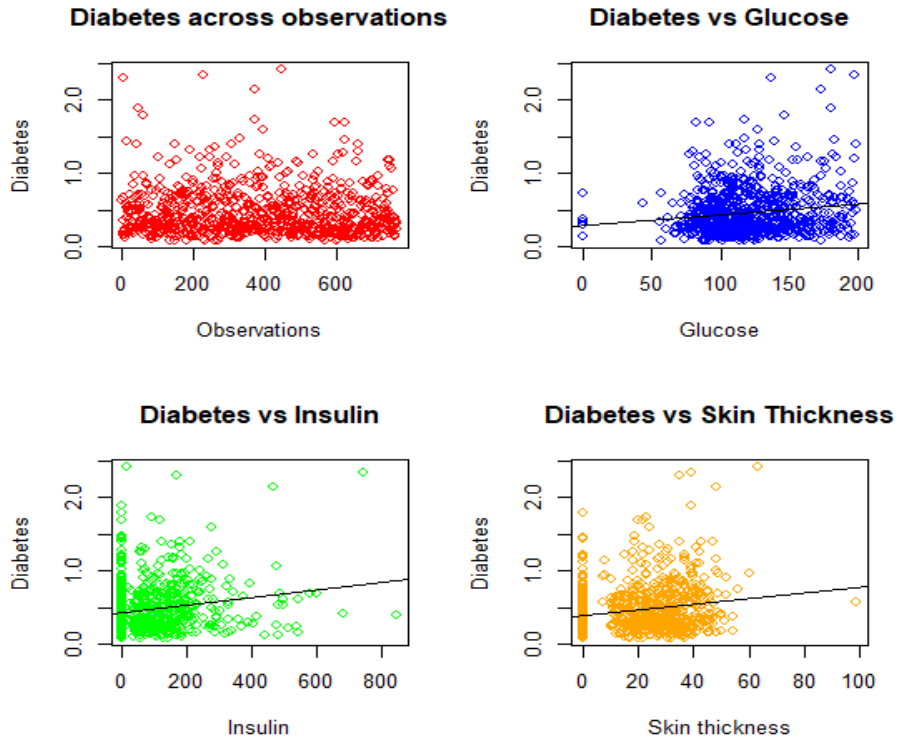

Figure 3. Scatter plot of the variables of interest

Table 1 shows that the minimum and maximum values of the diabetes pedigree function are 0.078 and 2.42 with a mean of 0.4719 and median of 0.3725, showing a positive skewness of 1.9162 and a leptokurtic of 8.5508. The plasma glucose concentration level can be as low as 0 and as high as 199. The insulin can be as low as 0 mu U/ml and as high as 846 mu U/ml; while the skin thickness can be as low as 0 mm and as high as 99 mm. Figure 1 shows that diabetes pedigree function and insulin are positively skewed, why glucose and skin thickness are near symmetric. Figure 2 depicts that diabetes pedigree function and insulin have many outliers at the top, glucose has only one outlier at the bottom and skin thickness has only one outlier as well but at the top. All the variables have at least one outlier. Figure 3 shows a line plot of diabetes pedigree function, which varied across individuals. The scatter plot of diabetes pedigree function against glucose shows a positive relationship. Figure 1 also shows that diabetes pedigree function is positively related to insulin and skin thickness.

3.2 Detection of Heteroscedasticity in Small Sample (n = 20)

The selected observations for the small sample are 32, 36, 123, 140, 155, 169, 200, 204, 233, 235, 244, 351, 390, 404, 584, 620, 623, 639, 703, 762, using uniform distribution from the total of 768 observations.

Table 2. Parameter estimate for small sample (n = 20) using OLS

	Estimate	St. Error	t value	p -value
(Intercept)	0.014025	0.378427	0.037	0.971
Glucose	0.003291	0.002693	1.222	0.239
Insulin	-0.0004	0.001065	-0.372	0.715
Skin	0.009091	0.008626	1.054	0.308

Residual standard error: 0.4204 on 16 degrees of freedom
Multiple R-squared: 0.1267, Adjusted R-squared: -0.0371
F-statistic: 0.7735 on 3 and 16 DF, p-value: 0.5257

Table 2 shows that all the three predictors for the small sample data do not have significant effect on the diabetes pedigree function at 5% level of significant. The coefficients show that glucose and skin thickness show positive coefficient while insulin shows negative coefficient. The model fitted is

$$\widehat{\mathrm{DPF}}_i = 0.014025 + 0.00329G_i - 0.0004I_i + 0.009091S_i + e_i, i = 1,2,\dots,20 \quad (5)$$

The model F-test shows that model 5 is not a good fit for the diabetes data. There is no enough evidence for the predictor variables to be able to predict diabetes in patients.

In order to carry out a diagnostic statistical testing of heteroscedasticity of general linear models for small sample, we use the Breusch-Pagan-Godfrey, Harvey and Glejser tests.

Table 3. Heteroscedasticity test in small sample (n = 20)

Test	Statistic	Degrees of freedom	p –value
Breusch-Pagan	5.43150	3	0.1428
Harvey	0.66107	15	0.5186
Glejser	4.72000	3	0.1930

Table 3 shows that all the three tests did not detect the presence of heteroscedasticity in the model for small sample (sample size = 20). This shows that all the three tests used in this paper are not sensitive to heteroscedasticity in a small dataset. There is no heteroscedasticity for small sample, but the sample is not large enough to show a relationship between diabetes and the predictors. Figure does not show varied variance, meaning no heteroscedasticity.

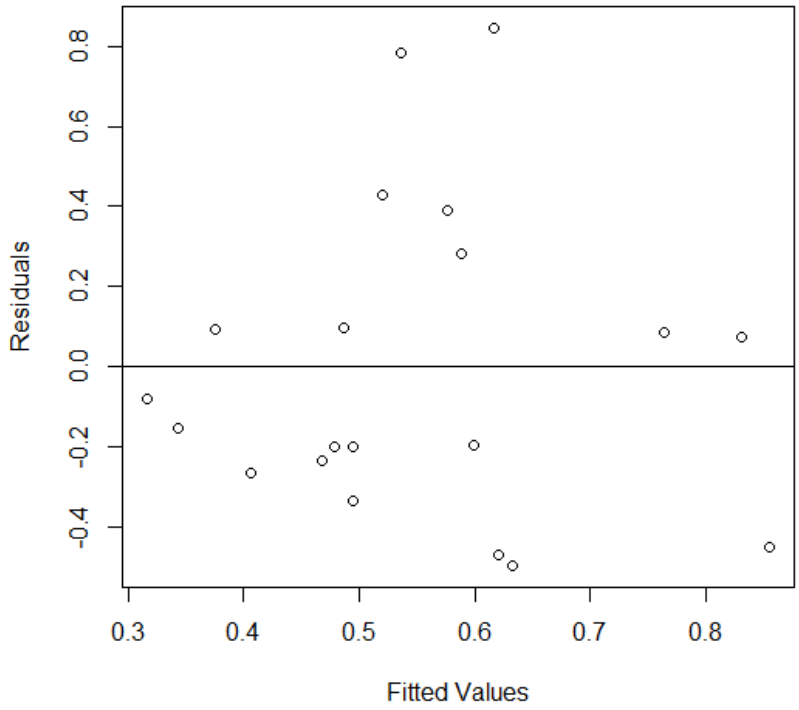

Figure 4. Residual against fitted values plot (n = 20)

3.3 Detection of Heteroscedasticity in Large Sample (n = 768)

In this large sample, the entire 768 observations in the dataset are used. In order to carry out a diagnostic statistical testing of heteroscedasticity of general linear models for large sample, we used the Breusch-Pagan-Godfrey, Harvey and Glejser tests.

Table 4: Parameter estimate for large sample (n = 768) using OLS

	Estimate	St. Error	t value	p –value
(Intercept)	0.2679163	0.0492339	5.442	0.0000000 11
G	0.0010254	0.0003881	2.642	0.008417000
I	0.0002644	0.0001195	2.213	0.027222000
S	0.0028680	0.0008159	3.515	0.000466000

Residual standard error: 0.3225 on 764 degrees of freedom
Multiple R-squared: 0.05601, Adjusted R-squared: 0.0523
F-statistic: 15.11 on 3 and 764 DF, p-value: 1.463e-09

Table 4 shows that all the three predictors for the large sample data have significant effect on the diabetes pedigree function at 5% level of significant. The coefficients show that glucose, insulin and skin thickness show positive coefficient. The model fitted for the large sample is given as

$$\widehat{\mathrm{DPF}}_i = 0.2679163 + 0.0010254G_i + 0.0002644I_i + 0.0028680S_i + e_i, i = 1,2,\ldots,768 \qquad (6)$$

The model F-test shows that model 6 is a good fit for the diabetes data. There is enough evidence for the predictor variables to be able to predict diabetes in patients.

In order to carry out a diagnostic statistical testing of heteroscedasticity of general linear models for small sample, we use the Breusch-Pagan-Godfrey, Harvey and Glejser tests.

Table 5. Heteroscedasticity test in large sample (n = 768)

Test	Statistic	Degrees of freedom	p -value
Breusch-Pagan	22.933	3	0.0000
Harvey	0.90756	15	0.3644
Glejser	35.3000	3	0.0000

Table 5 shows that both Breusch-Pagan-Godfrey and Glejser detected the presence of heteroscedasticity in the large sample model, while Harvey did not detect heteroscedasticity in the model for the large sample. This shows both Breusch-Pagan-Godfrey and Glejser are sensitive to heteroscedasticity in a large dataset, while Harvey is not sensitive to heteroscedasticity in large samples of a cross-sectional study. The model is a good fit but has heteroscedasticity presence. The result of Breusch-Pagan and Glejser tests show that the null hypothesis (H_0) of homoscedasticity is rejected and conclude that there is presence of heteroscedasticity, since the test is significant at 5% level, because the p-values (4.17e-05 and 1.044-07) is less than 0.05 for both tests.

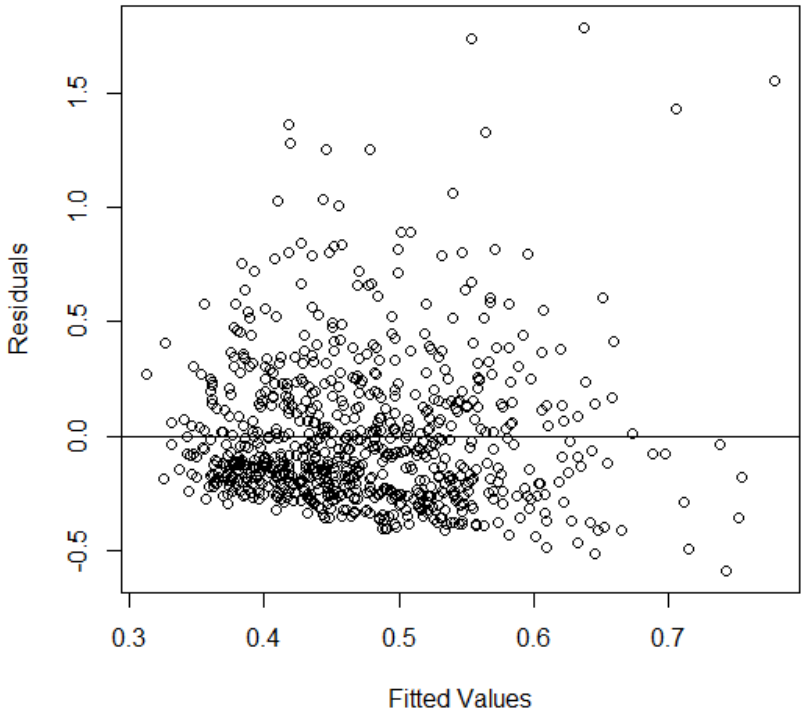

Figure 5. Residual against fitted values plot (n = 768)

It is easy to see from Figure 5 that the residuals exhibit a "cone" shape, and they are not distributed with equal variance throughout the plot.

4. Conclusion

In this paper, data on a cross-section of 768 patients were collected from National Institute of Diabetes and Digestive and Kidney Diseases (1990) to predict diabetes in patients. A patient is diabetic if the diabetes pedigree function is greater than 50. Variables that are suspected to predict diabetes are selected among others, they are plasma glucose concentration a 2 hours in an oral glucose tolerance test (G), 2-Hour serum insulin (mu U/ml) (I), and Triceps skin fold thickness (mm) (S). The diabetes pedigree function is positively skewed and highly leptokurtic and have some outliers. On the average, diabetes pedigree function is 0.4719 with a standard deviation of 0.3313. The maximum for the observations under study is 2.42 and the minimum is 0.078. A linear model is fitted using diabetes pedigree function as a response variable, while glucose, insulin and skin fold thickness as predictors. Most of the observations have diabetes pedigree function values less than 1. The maximum values of glucose, insulin and skin thickness are 3.6288, 10.1596 and 2.4755 respectively.

The linear model of the small sample shows that all the predictors cannot significantly predict diabetes because their p-values for individual t-test are greater than 0.05. However, the heteroscedasticity test using the three tests show that there is no heteroscedasticity in the model. The estimates are valid and reliable only that the data is not sufficient to fit a model for predicting diabetes. The short fall of the model is that there is no sufficient evidence (data) to show that the predictors can predict diabetes in patients. The model shows that only 12.67% of the variation in diabetes can be explained by the variations in the predictors. The model also shows that the higher the glucose level and skin fold thickness but lower insulin level, the more likely a patient is diabetic, only that the estimated parameters are not significantly different from zero and the model not a good fit for predicting diabetes.

However, when the large sample of 768 observations are used, all the three predictors are significant at 5% level. The most significant predictor is the skin fold thickness, followed by glucose level, then insulin level. The model shows that only 5.6% of the variation in diabetes can be explained by the variations in the predictors. Nevertheless, the model estimated parameters are significantly different from zero and the model is a good fit for predicting diabetes in patients. There is enough evidence to predict diabetes. This implies that the skin fold thickness, level of glucose and insulin in the body can be used to determine if a patient is diabetes or not. The low coefficient of determination can be attributed to the presence of heteroscedasticity in the model. Breusch-Pagan and Glejser tests show that there is heteroscedasticity in the model while Harvey test shows that there is no heteroscedasticity.

In conclusion, these changes in the level of diabetes pedigree function in individuals make the data to change for different individuals. It is scarcely possible to be certain about the nature of the heteroscedasticity in a regression model. It can be seen that, the comparison is clear, that in small sized samples, the three tests could not detect hereroscedasticity but Breusch-Pagan-Godfrey test is better than Glejser, and Glejser test is better tha Harvey test. More so, for large

samples, Breusch-Pagan-Godfrey and Glejser tests are more sensitive to heteroscedasticity than Harvey test, especially for cross-sectional diabetes pedigree function model. It is therefore recommended that attention should be given to plasma glucose concentration, insulin and triceps skin fold thickness in the body in order to reduce the probability of having diabetes. It is also recommended that as statisticians, one should use any one of Breusch-Pagan-Godfrey, Glejser or Harvey tests for small samples when trying to detect the presence of heteroscedasticity, but for large samples any of Breusch-Pagan-Godfrey and Glejser tests can perform well.

Furthermore, it is recommended that one should remove any heteroscedasticity caused by misspecification by removing (where possible) the source of that misspecification (e.g. correct omitted variables by including the appropriate variable). If there is still heteroscedasticity, it may not be harmful, and if any solution is sorted for, it should not distort regression model or the interpretation of coefficients. Finally, sufficiently large samples should be used when predicting diabetes in patients and the level of glucose, insulin and skin thickness can be used as predictors for diabetes prediction.

References

1) Abdulhadi, N., Al-Mousa, A. (2021). Diabetes Detection Using Machine Learning Classification Methods, International Conference on Information Technology (ICIT), 2021.
2) Alabi, O. O., Ayinde, K., Babalola, O. E., Bello, H. A., Okon, E. C. (2020). Effects of Multicollinearity on Type I Error of Some Methods of Detecting Heteroscedasticity in Linear Regression Model. Open Journal of Statistics, 2020.
3) Baltagi, B.H., (1980): "On seemingly unrelated regressions with error components", Econometrica 48, Pp 1547-1551.
4) Breusch, T.S. and Pagan, A.R. (1979): "A simple Test for Heteroscedasticity and Random Coefficient Variation". Econometrica, vol. 47, No. 5 (September, 1979).
5) Breusch, T. and Pagan, A.R (1980): "The LM Test and Its Applications to Model Specification in Econometrics". Review of Economic Studies, 47, 1980, Pp 239-254.
6) Center for Disease Control and Prevention (2022). Available at https://www.cdc.gov/diabetes/basics/diabetes. Accessed on 11th November, 2022.
7) Deysi, G. (2022). Type 2 Diabetes among Adult Latinas Living in California, California State University, Fresno.
8) Ekum, M.I., Farinde, D. A. and Ayoola, F.J. (2013): "Panel Data: The Effects of Some World Development Indicator (WDI) on GDP Per Capita of Selected African Union (AU) Countries (1981-2011)", International Journal of Science and Technology (IJST). Vol. 2 No. 12, December, 2013.
9) Ekum, M.I., Akinmoladun, O.M., Aderele, O.R. and Esan, O.A. (2015). Application of Multivariate Analysis on the effects of World Development Indicators on GDP per capita of Nigeria (1981-2013). International Journal of Science and Technology (IJST); Vol. 4, No. 12, December, 2015, Pp 254-534.

10) Glejser, H. (1969). A New Test for Heteroscedasticity. Journal of the American Statistical Association, 64 (325), 316-323.
11) Golfield, S.M. and Quandt, R.E. (1965). Some Test for Homoscedasticity. Journal of the American Statistical Association, 60, 539-547.
12) Harvey, A.C (1976). Estimating Regression Models with Multiplicative Heteroscedasticity. Econometrica, 44 (3), 461-465.
13) Hildreth, C. and Houck, J.P. (1968): Some Estimators for a Linear Model with Random Coefficients. Journal of the American Statistical Association, 63 (1968), Pp 584-595.
14) Iluno, C., Taylor, J. I., Akinmoladun, O. M., Aderele, O. R., and Ekum, M. I. (2021). Modelling the effect of Covid-19 mortality on the economy of Nigeria. Research in Globalization, 3, 100050.
15) Kennedy, P. (1998): A Guide to Econometrics. Chapters 5,6,7 and 9.
16) Koenker, R. (1981). A Note on Studentizing a Test for Heteroscedasticity. Journal of Econometrics, 17, 107-112.
17) Mittelhammer R. C., Judge, G. G., Miller, D. J. (2000). Econometric Foundations. Cambridge University Press, Cambridge.
18) National Institute of Diabetes and Digestive and Kidney Diseases (1990).
19) Okunnu, M. A., Ekum, M. I. and Aderele, O. R. (2017). The Effects of Microeconomic Indicators on Economic Growth of Nigeria (1970 - 2015). American Journal of Theoretical and Applied Statistics; 2017; 6(6): 325-334.
20) Wu Y, Ding Y, Tanaka Y, Zhang W (2014). Risk factors contributing to type 2 diabetes and recent advances in the treatment and prevention. Int J Med Sci. 11(11),1185-200. doi: 10.7150/ijms.10001.
21) Zou Q, Qu K, Luo Y, Yin D, Ju Y and Tang H (2018). Predicting Diabetes Mellitus With Machine Learning Techniques. Front. Genet. 9:515. doi: 10.3389/fgene.2018.00515.