AIP

by Tarno Tarno

RESEARCH ARTICLE | MAY 16 2023

Modelling of relationship between fasting glucose level with respect to age using linear and nonlinear models to predict prediabetes

Sudarno Sudarno; Tarno Tarno

AIP Conference Proceedings 2683, 040009 (2023)
https://doi.org/10.1063/5.0125557

Modelling of Relationship Between Fasting Glucose Level with Respect to Age Using Linear and Nonlinear Models to Predict Prediabetes

Sudarno Sudarno[a)], Tarno Tarno[b)]

Department of Statistics, Faculty Sciences and Mathematics, Diponegoro University, Semarang, Indonesia

a) Corresponding author: dsghani@gmail.com
b) tarno.stat@gmail.com

Abstract. Humans have two kinds of variables related to their health. These variables are age, and fasting glucose level. It turns out that glucose levels change with time. This study wanted to find model of relationship between age and fasting glucose level. The models tested were linear model, logarithm model and logarithm linear regression model. The model used is a significant model. Each model was measured for error accuracy using the mean square error (MSE) and Akaike information criterion (AIC). Criteria for selecting the best model using the smallest MSE and AIC values. The best model is the logarithm linear model. The fastest prediction time for someone suffering from prediabetes is 51 years. Therefore, starting from that age a person needs to be careful to maintain his glucose levels in order to avoid diabetes.

INTRODUCTION

There are data on people with four variables are body mass index, age, cholesterol, and fasting glucose level [1, 2]. Based on data processing, it turns out that fasting glucose level is affected by age. Therefore, this study wants to find the right model to find the distribution between age and fasting glucose level. The models used are linear and nonlinear models. For linear model using the simple linear regression model, while nonlinear models using the logarithm model and their combinations.

A simple linear regression model is a regression model that has only one predictor. The logarithm model is a model with a base function in the form of a logarithm function, while the combined model is a combination of a logarithm model with a simple linear regression model. All of these models were used to predict the value of fasting glucose level toward age. The fasting glucose level value can be used to predict when a person will be diagnosed with prediabetes. The person is diagnosed with prediabetes, if the fasting glucose level is between 100-125 mg/dL [3]. Prediabetes is a disease where a person has an abnormality in metabolism of glucose level. If it be treated, it is possible to become a normal or healthy person. But if it is not guarded, it can become diabetes.

The mean square error (MSE) and Akaike information criterion (AIC) methods are used to measure the accuracy of model to predict the response value [4]. The mean square error method is a method of measuring the result of computational error. The Akaike information criteion is the measurement result of the difference between the number of parameters and the logarithm of the loglikelihood function. If the value is small, the model is considered good for predicting response values, and vice versa. Data processing and computation using R software.

Some related research namely research about effect of age and gender on blood lipids and glucose. The result that the mean differences for fasting blood glucose across age groups were observed to rise from 21-30 years age group and in subjects having ages between 50-60 years. However, later in life the mean fasting blood glucose were observed to be lower. Gender differences were not significant between groups based upon glycemia and lipidemia. Fasting glycemia and lipidemia were observed to show significant differences among different age groups [5]. Relationship between age at menarche and risk of glucose metabolism disorder: a systematic review and dose-response meta-analysis concluded that older age at menarche (range 8-18 years old) is associated with reduced risk of glucose

The 6th International Conference on Energy, Environment, Epidemiology and Information System (ICENIS) 2021
AIP Conf. Proc. 2683, 040009-1–040009-9; https://doi.org/10.1063/5.0125557
Published by AIP Publishing. 978-0-7354-4517-8/$30.00

metabolism disorder. The strongest reduction in risk of gestational diabetes mellitus is observed at menarche age of 14.5 years [6].

Association between fasting glucose and all-cause mortality according to sex and age: a prospective cohort study resulted that higher glucose levels were generally associated with higher body mass index, systolic blood pressure, and total cholesterol values [7]. Diabetes mellitus in elderly conclude that aging is characterized by high prevalence of associated co-morbidities and risk of frailty. Therefore, it is important to provide high quality and specific care for old diabetic patients. Any treatment should be based on elderly classification and individualization to avoid iatrogenic complications, especially dehydration and hypoglycemias [8]. Diabetes in eldery recommended that healthy elderly people with diabetes should be treated to achieve thesame glycemic, blood pressure and lipid targets as younger people with diabetes [9].

The aim of this research was to find a significant model to determine the relationship between fasting glucose level with respect to age. The models tested were linear regression model with two parameters, logarithm model with one parameter, and logarithm linear regression model with two parameters. The accuracy of models be measured using MSE and AIC. The model is said to be the best if it has the smallest MSE and AIC values compared to other models. The best model is used to predict the value of fasting glucose level with respect to age. This value is used to predict when a person is diagnosed early with prediabetes.

MATERIAL AND METHODS

Linear Regression Model with Two Parameters

Linear regression model is the regression equation has one or more predictors where each predictor powers one. Linear regression model with two parameters is a regression equation with one predictor to the power of one and having two parameters [10, 11]. Linear regression model with two parameters has equation

$$y = a + bx + e \tag{1}$$

where a : intercept, b : slope, x : predictor, y : response, and e : error.

It is necessary to estimate the value of parameters a and b to get this model. The parameters a and b could be estimated using the sum squares error. The sum squares error is derived with respect to those parameters, respectively. Step of searching for parameters are as follow. First, the error is determined by

$$e = y - a - bx$$

Then the square error is determined with

$$e^2 = (y - a - bx)^2$$

If the square error to a set of paired observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. The mathematical expression for the sum square error becomes

$$Sr = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \tag{2}$$

The equation Sr is necessary differentiated with respect to these parameters to get an estimate of a and b values. The partial derivative of Sr to parameter a is as follow

$$\frac{\partial Sr}{\partial a} = -2\left(\sum_{i=1}^{n} y_i - an - b\sum_{i=1}^{n} x_i\right)$$

The partial derivative of Sr toward parameter b is

$$\frac{\partial Sr}{\partial b} = -2\left(\sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2\right)$$

The partial derivative of Sr with respect to a and b in order to be minimal, they must be equal zero. The result of the simplification is obtained

$$an + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

and

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

The two equations produce a normal equation. Based on the normal equation, the parameter values can be obtained by

$$b = \left(n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i \right)\left(n\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i \right)^{-1} \quad (3)$$

and

$$a = \bar{y} - b\bar{x} \quad (4)$$

where $\bar{y}$ and $\bar{x}$ are the means of response and predictor variables, respectively.

The best model is

$$\hat{y} = a + bx \quad (5)$$

This model is said to be the best model because it has no error.

Logarithm Model with One Parameter

The next discussion is the logarithm model with one parameter. The logarithm model with one parameter is a function of natural logarithm with parameter as coefficient [12-14]. The model is written with

$$y = a\ln x + e \quad (6)$$

where a: parameters of function, x : predictor, y : response, $\ln x$: natural logarithm function of x, and e : error. The sum square error method will be tried to estimate the value of the parameters a. The steps are as follows. First, the error is determined, that is

$$e = y - a\ln x$$

Next, the square error is determined by

$$e^2 = (y - a\ln x)^2$$

If it use the square error to a set of paired observations: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. The equation of the sum square error is

$$Sr = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a\ln x_i)^2 \quad (7)$$

Equation Sr is necessary to differentiate with respect to its parameter to get an estimate of the a value.

$$\frac{\partial Sr}{\partial a} = -2\left(\sum_{i=1}^{n} y_i \ln x_i - a\sum_{i=1}^{n} \ln^2 x_i \right)$$

In order to equation Sr is minimum, then take this derivatives must equal to zero. If this is done, the equations can be expressed as follow

$$\sum_{i=1}^{n} y_i \ln x_i = a\sum_{i=1}^{n} \ln^2 x_i$$

Based on the normal equation above, the parameter value can be generated, namely

$$a = \left(\sum_{i=1}^{n} y_i \ln x_i \right)\left(\sum_{i=1}^{n} \ln^2 x_i \right)^{-1} \quad (8)$$

The best model that is produced with no errors, namely

$$\hat{y} = a\ln x \quad (9)$$

Logarithm Linear Regression Model with Two Parameters

The extension of the linear and natural logarithmic functions is the logarithm linear regression model with two parameters [10-15]. The model is an equation

$$y = a + b\ln x + e \tag{10}$$

where a and b : parameters of model, x : predictor, y : response, e : error.

The sum square error is used to determine parameter estimates. It is partially derivative for each of their parameters. The steps for this process are as follows. First the error is determined by

$$e = y - a - b\ln x$$

Then the square error is determined by

$$e^2 = (y - a - b\ln x)^2$$

If it use the square error to a set of paired observations: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. The formula of the sum square error is

$$Sr = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - b\ln x_i)^2 \tag{11}$$

The equation Sr is necessary to differentiate with respect to their parameters to get the parameter values a and b. The description of the partial derivative of Sr with respect to each parameter is summarized as follow

$$\frac{\partial Sr}{\partial a} = -2\left(\sum_{i=1}^{n} y_i - an - b\sum_{i=1}^{n} \ln x_i\right)$$

and

$$\frac{\partial Sr}{\partial b} = -2\left(\sum_{i=1}^{n} y_i \ln x_i - a\sum_{i=1}^{n} \ln x_i - b\sum_{i=1}^{n} \ln^2 x_i\right)$$

The modeling principle aims to obtain minimal error. Therefore, the two equations must be equal to zero. If this is done, the expression could be presented by

$$na + b\sum_{i=1}^{n} \ln x_i = \sum_{i=1}^{n} y_i$$

and

$$a\sum_{i=1}^{n} \ln x_i + b\sum_{i=1}^{n} \ln^2 x_i = \sum_{i=1}^{n} y_i \ln x_i$$

Based on the above normal equation, the parameter estimates can be determined, i.e.

$$a = \left(\sum_{i=1}^{n} y_i \sum_{i=1}^{n} \ln^2 x_i - \sum_{i=1}^{n} \ln x_i \sum_{i=1}^{n} y_i \ln x_i\right)\left(n\sum_{i=1}^{n} \ln^2 x_i - \left(\sum_{i=1}^{n} \ln x_i\right)^2\right)^{-1} \tag{12}$$

and

$$b = \left(n\sum_{i=1}^{n} y_i \ln x_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} \ln x_i\right)\left(n\sum_{i=1}^{n} \ln^2 x_i - \left(\sum_{i=1}^{n} \ln x_i\right)^2\right)^{-1} \tag{13}$$

The best model that is obtained in the form

$$\hat{y} = a + b\ln x \tag{14}$$

This model is considered to have no error.

Age and Prediabetes

Human age determines his abilities. If humans get older, their abilities will decrease. For example, if you get older, your vision becomes more and more blurred. According to World Health Organization (WHO), the age of 45-59 years is called the middle age, while according to the Ministry of Health of the Republic of Indonesia (2021), the age of 46-55 years is called early advanced age. Humans have this age that has begun to decrease in their abilities. So it needs vigilance in activities so that health is maintained. Reduced ability of the body can lead to prediabetes. Prediabetes is a disease in which fasting glucose levels exceed normal limit. The fasting blood sugar level for normal people is 70-99 mg/dL. Symptoms of prediabetes are fatigue, blurred vision, frequent thirst and hunger, frequent urination, and wound that doesn't heal.

RESULT AND DISCUSSION

The discussion of this paper is to apply three kind functions, that are the linear regression model with two parameters, the logaritm model with one parameter, and the logaritm linear regression model with two parameters. Data processing using sofware R, while output analysis using statistical methods. The parameter results are visualized in the form of table and scatter diagram.

The data that be investigated were 51 people with four variables, namely age (year), body mass index or BMI (kg/m^2, cholesterol (mg/dL), and fasting glucose level (mg/dL). The result of descriptive analysis are summarized in Table 1 below [16]:

TABLE 1. Descriptive Statistics of Sample.

Variable	Mean	Minimum	Maximum
Age	47.20	21	76
BMI	29.81	19.28	47.87
Cholesterol	185.96	119	283
Glucose	98.14	72	136

Furthermore, searching for a model of the relationship between one variable and another. There is result that age is related to fasting glucose level. Therefore, the variable age is used as a predictor of x, while the variable glucose is the response to y. It want to find the best model between fasting glucose level toward age. The best model can be used to predict when a person will develop prediabetes. Several models that can be made to determine the relationship between the variables of age and fasting glucose level are described below.

Application of Linear Regression Model with Two Parameters

The equation (5) needs to find the values of parameters intercept and slope. The computation of the data obtained results as shown in the Table 2, that is [16, 17]:

TABLE 2. Result of parameters of linear regression model.

Coefficients	Estimate	t-value	Pr(>|t|)	Sign
Intercept	83.6380	15.580	< 2e-16	***
x	0.3072	2.848	0.00642	**

According to the Table 2, it means that the intercept parameter a = 83.6380. It has sign three asterics. This parameter gets meaning that significant by 0.001 level. The slope parameter b = 0.3072 with sign is two asterics. This parameter gets meaning that significant by 0.01 level. The estimated value of two parameters are very appropriate in building the model. The following figure to see the accuracy of model and data distribution [17-20].

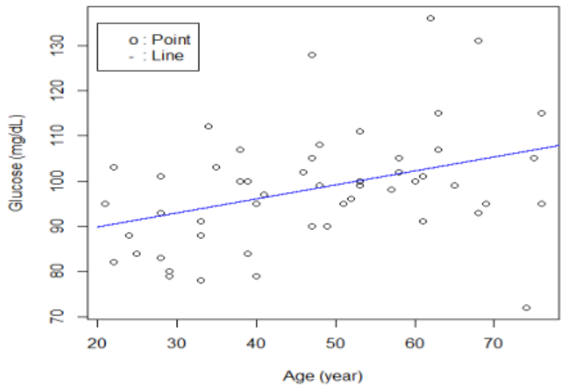

FIGURE 1. Distribution of data and linear regression model.

In Figure 1, there are points that describe the successive pairing between age in year and fasting glucose level in mg/dL. The linear line as model that describes the linear relationship between level of fasting glucose toward age. We can construct the linear regression model with two parameters as

$$\hat{y} = 83.6380 + 0.3072\,x \tag{15}$$

This function is useful for predicting the level of glucose by age. The glucose level can be used to predict that someone has prediabetes. If this model is used to find fasting glucose level, a person can be diagnosed with prediabetes for at least 53.3 or 53 years old.

Application of Logarithm Model with One Parameter

The equation (9) needs to determine the value of a multiplication parameter. The processed results of the data obtained value as presented in Table 3 by [16, 17]:

TABLE 3. Result of logarithm model parameter.

Parameter	Estimate	t-value	$\Pr(>\lvert t \rvert)$	Sign
a	25.7672	55.35	< 2e-16	***

According to the Table 3, it means that the parameter $a = 25.7672$. It has sign three asterics. This parameter could be interpreted as having significance to 0.001 level. The estimated value of that parameter is very appropriate in making the model. The following figure to see the accuracy of model and data distribution [16-19].

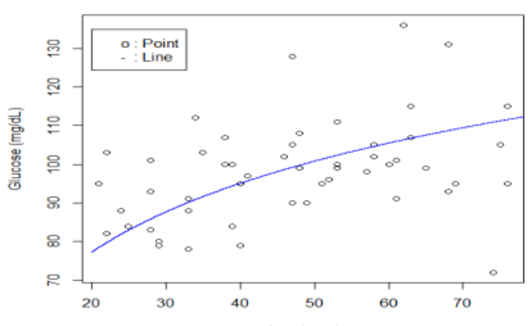

FIGURE 2. Distribution of data and logarithm model.

In Figure 2, there are points that describe the successive pairing between age in year and fasting glucose level in mg/dL. The curve is model that describes the nonlinear relationship between level of fasting glucose toward age. It could be made the logarithm model with one parameter that is

$$\hat{y} = 25.7672 \ln x \tag{16}$$

This function is useful to predict the level of fasting glucose by age. The fasting glucose level can be used to predict that someone has prediabetes. If this model is used to find fasting glucose levels, a person can be diagnosed with prediabetes for at least 48.5 or 49 years old.

Application of Logarithm Linear Regression Model with Two Parameters

The equation (14) can be determined by looking for the values for a and b parameters. The results of data processing obtained values as summarized in Table 4 as below [16, 17]:

TABLE 4. Result of parameters of logarithm linear regression model.

Parameters	Estimate	t-value	$Pr(>\|t\|)$	Sign
a	44.707	2.502	0.0158	*
b	14.086	3.003	0.0042	**

The Table 4 means that the parameter a = 44.707. It has sign one asteric. This parameter gets meaning that significant by 0.05 level. The parameter b = 14.086 with sign is two asterics. This parameter gets meaning that significant by 0.01 level. The estimated value of two parameters are very appropriate in construction this model. The following figure to see the accuracy of model and data distribution [16-19].

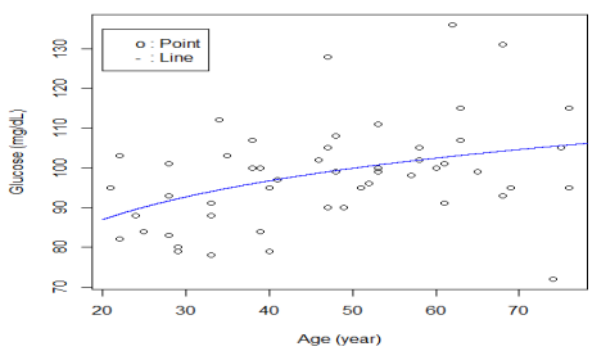

FIGURE 3. Distribution of data and logarithm linear regression model.

Figure 3 shows the distribution of pair points between age in year and fasting glucose level in mg/dL. The curve is model that describes the nonlinear relationship between level of fasting glucose with respect to age. It could be made the logarithm linear regression model with two parameters as

$$\hat{y} = 44.707 + 14.086 \ln x \tag{17}$$

This function is useful to predict the fasting glucose level wiht age. The fasting glucose level can be used to predict that someone has prediabetes. If this model is used to find fasting glucose levels, a person be said to have symptoms of prediabetes for at least 50.7 or 51 years old. If it wants to see the comparison of data distribution and all models graphically shown in Figure 4 in form [16-19]:

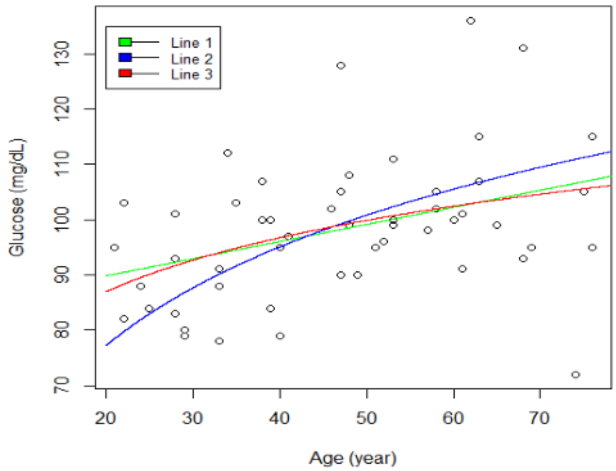

FIGURE 4. Distribution of data and all models.

Figure 4 shows distribution of points and all models. Graph of the linear regression model is almost the same as the logarithm linear regression model, while graph of the logarithm model is different with the others. This can also be seen from the AIC and MSE values of model. The AIC and MSE values of the linear regression model are almost the same as the logarithm linear regression model, but for the logarithm model is different. Graphically, the logarithm model has a high slope.

Measure of Accuracy of Model

The model accuracy measure is the value of error function. It use mean sqare error (MSE) and Akaike information criterion (AIC) methods [4]. They have the property that the smaller the value, the better the model for predicting that problem. Formula the mean square error and AIC, respectively, are

$$MSE = \overline{e^2} \quad (18)$$

where $\overline{e^2}$ = the mean of square error
and

$$AIC = 2\left(p - \ln \hat{L}\right) \quad (19)$$

where $\hat{L}$: likelihood function, p : number of model parameter.
The MSE and AIC values from the processed data using each model are summarized in Table 5 below

TABLE 5. MSE and AIC values of models.

Models	MSE	AIC
Linear regression model	141.7229	403.3793
Logarithm model	157.3179	406.7034
Logarithm linear regression model	139.5022	402.5738

Actually, these three models are suitable models to be used as data distribution models regarding the relationship between fasting glucose levels with respect to age. If it is seen from the MSE and AIC values of each model. If the MSE value is small, the AIC value is also small, and vice versa. The best model for this case of the relationship between age and fasting glucose level is the logarithm linear regression model. Because this model has the smallest MSE and AIC values, when compared with the MSE and AIC values of the other models.

CONCLUSION

Prediabetes is a dangerous disease. A person is said to have prediabetes if the fasting glucose level is at least 100 mg/dL. The level of fasting glucose can be affected by age. The model used to predict fasting glucose level with respect to age are linear regression model, logarithm model and, logarithm linear regression model. The linear regression model predicts that a person is suspected of having prediabetes for at least 53 years old. The logarithm model predicts that a person is diagnosed with prediabetes for at least 49 years old, meanwhile the logarithm linear regression model predicts that a person is considered to have prediabetes for at least 51 years old. Based on the smallest MSE and AIC values, the best model is obtained. The best model is the logarithm linear regression model with a value of MSE = 139.5022 and a value of AIC = 402.5738. The best models has equation $\hat{y} = 44.707 + 14.086 \ln x$. Because glucose level depends on age, then at that age or more, a person is expected to maintain his health to standard or not high glucose levels.

REFERENCES

1. C. Pérez, M. Guzmán, A.P. Ortiz, M. Estrella, Y. Valle, E. Suárez, Ethn. Dis. 18 (2008)
2. E. Suarez, C.M. Perez, R. Rivera, M.N. Martinez, Applications of regression models in epidemiology. John Wiley & Sons, Inc. New Jersey (2017)
3. Alodokter, Diabetes. Kementerian Kesehatan Republik Indonesia (2021)
4. D. Collett D, Modelling Survival Data in Medical Research. Chapman & Hall Press New York (2015)
5. S.H. Khan, U. Masood, M.S. Hanif, Rawal Med. J. 37, 4 (2012)
6. R. Yongcheng, Z. Haiyin, Z. Dongdong, H. Dongsheng, Menopause J. 27, 7 (2020)
7. S. Yi, S. Park, Y. Lee, Science Report J. 7, 19 (2017)
8. F. Chentli, S. Azzoug, S. Mahgoun, Indian J. Endocrinol Metab. 19, 6 (2015)
9. G.S. Meneilly, D. Tessier, Canadian J. of Diabetes 37: 184-190 (2013)
10. N.R. Draper, H. Smith, Applied Regression Analysis. John Wiley & Sons Inc. New York (1998)
11. D.C. Montgomery, G.C. Runger, Applied Statistics and Probability for Engineers. John Wiley & Sons Inc. New York (2007)
12. S. Makgakga, P. Sepeng, 2013 J. of Social Sciences. Rome-Italy 4, 13 (2013)
13. J. Stewart, Calculus. Division of International Thomson Publishing Inc. (1998)
14. D. Varberg, E.J. Purcell, S.E. Rigdon, Calculus. Pearson Education Inc. (2007)
15. D. Rasch, D. Schott, Mathematical Statistics. John Wiley & Sons Ltd . (2018)
16. M.J. Crawley, The R Book. John Wiley & Sons Ltd . United Kingdom (2007)
17. P. Dalgaard, Introductory Statistical with R. Springer Science. New York (2008)
18. M. Gardener, Beginning R The Statistical Programming Language. John Wiley & Sons Inc. (2012)
19. G.J. Kerns, Introduction to Probability and Statistics Using R. (2011)

AIP

ORIGINALITY REPORT

18%	17%	8%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.researchgate.net	Internet Source	2%
2	pubmed.ncbi.nlm.nih.gov	Internet Source	2%
3	www.ncbi.nlm.nih.gov	Internet Source	1%
4	epdf.pub	Internet Source	1%
5	ndl.ethernet.edu.et	Internet Source	1%
6	edoc.pub	Internet Source	1%
7	vdoc.pub	Internet Source	1%
8	garuda.kemdikbud.go.id	Internet Source	1%
9	Submitted to Zespol Szkol Ogolnoksztalcacych nr 2 w Tarnowie	Student Paper	1%

10	library.binus.ac.id Internet Source	1 %
11	www.sagelink.ca Internet Source	1 %
12	dokumen.pub Internet Source	1 %
13	www.health.harvard.edu Internet Source	<1 %
14	homepage.smc.edu Internet Source	<1 %
15	www.cse.unr.edu Internet Source	<1 %
16	ebin.pub Internet Source	<1 %
17	O. Carisse, V. Morissette-Thomas, H. Van der Heyden. "Lagged Association Between Powdery Mildew Leaf Severity, Airborne Inoculum, Weather, and Crop Losses in Strawberry", Phytopathology, 2013 Publication	<1 %
18	Submitted to ACT Information Consulting Student Paper	<1 %
19	www.nature.com Internet Source	<1 %
	www.scribd.com	

20	Internet Source	<1 %
21	Submitted to Padjadjaran University Student Paper	<1 %
22	escholarship.org Internet Source	<1 %
23	scholar.sun.ac.za Internet Source	<1 %
24	www.nccohio.org Internet Source	<1 %
25	www.stata.com Internet Source	<1 %
26	bmcmedresmethodol.biomedcentral.com Internet Source	<1 %
27	doczz.biz.tr Internet Source	<1 %
28	www.jmi.nic.in Internet Source	<1 %
29	berlmathges.de Internet Source	<1 %
30	inis.iaea.org Internet Source	<1 %
31	www.digitalimagecorp.de Internet Source	<1 %

32	www.econ.au.dk Internet Source	<1 %
33	www.slideshare.net Internet Source	<1 %
34	Ing-Guey Jiang. "On the Mass-Period Distributions and Correlations of Extrasolar Planets", The Astronomical Journal, 11/2007 Publication	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		

AIP

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

Instructor

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10