Chapter (3)

Comparison Between Logistic and Calibration linear Regression

3.1 Introduction:

In this chapter we are going to handle the notion of calibration.

Here is a random sample of 120 people, 100 of them are infected with blood cancer, the rest 20 are fit , given that Logistic regression Model has been established using SPSS , and for the same data calibration regression model has been made ready using STATA applying the order vec (Robust) .

Our aim is to estimate calibration standard errors for both models patterns to be compared.

3.2 Theoretical frame work:

3.2.1 Regression calibration:

The regression calibration method is a simple approach wherein we need only develop and fit the calibration model for the regression of the unknown covariates X_U on (X_Z, X_W). This is accomplished using replication, validation, or instrumental data in place of the unknown X_U. This first stage regression results in a calibration function for estimating X_U. The unobserved covariates are then replaced by their predicted values from the calibration model in a standard analysis. Finally, the standard errors are adjusted to account for the estimation of the unknown covariates. The typical approach is to calculate standard errors using bootstrap or sandwich methods, but asymptotic standard errors have been derived by one of the authors (Carroll) and are included in the associated software.

With replicate data, the measurement error variance may be estimated by

$$\sum_{uu} = \frac{1}{\sum_{i=1}^{n} k_i - 1} \sum_{i=1}^{n} \sum_{j=1}^{k_t} (\mathrm{Wij} - \overline{\mathrm{Wi}}\cdot)(\mathrm{Wij} - \overline{\mathrm{Wi}}\cdot)^{\mathrm{T}}$$

Alternatively, the user may specify this variance matrix if it is known or estimated externally (to the dataset in use). We need information to substitute for X_U and we know that the best linear approximant

to X_U given (Z,W) is:

$$\hat{X}_U \approx \mu X_U + \begin{pmatrix} \sum_{xx} \\ \sum_{zx} \end{pmatrix} + \begin{pmatrix} \sum_{xx} + \sum_{uu} / k & \sum_{xz} \\ \sum_{zx} & \sum_{zz} \end{pmatrix}^{-1} \begin{pmatrix} \overline{W} - \mu_W \\ Z - \mu_Z \end{pmatrix} \quad \text{.....(3.1)}$$

In order to operationalize this linear approximant, we make the usual substitutions for unknown quantities. First, we use $\hat{\mu}_W = \mu X_U$; substituting the mean of the replicate values for the mean of the unknown covariates. In addition,

$$\sum_{zz} = \frac{1}{n-1} \sum_{i=1}^{n} (Z_i - \overline{Z})(Z_i - \overline{Z})^T$$

as the usual analysis of variance estimate. We use

$$\sum_{xz} = \frac{1}{\nu} \sum_{i=1}^{n} K_i (\overline{W}_i - \hat{\mu}_W)(Z_i - \overline{Z})^T$$

$$\sum_{xx} = \frac{1}{\nu} \left[\sum_{i=1}^{n} K_i (\overline{W}_i - \hat{\mu}_W)(\overline{W}_i - \hat{\mu}_W)^T \right] - \frac{n-1}{\nu} \sum_{uu}$$

Where $\nu = \sum_i K_i - \frac{\sum_i K_i}{\sum_i K^2{}_i}$. The estimated variance matrix for the unknown X_U is seen in two components due to the variance of X_U and the measurement error variance. Armed with these estimates, it is straightforward to mechanically derive the estimated values for the unknown X_U, produce estimates $\hat{X}_U$ using equation (3.1), and proceed with a standard analysis. Obviously, calculation of a variance matrix for the estimated coefficients will have to address the additional parameters from this susbstitution .

3.2.2 Logistic regression:

Logistic regression assumes that the dependent variable is a stochastic event. That is, for instance if we analyze a pesticides kill rate the outcome event is either killed or alive. Since even the most resistant bug can only be either of these two states, logistic regression thinks in likelihoods of the bug getting killed. If the likelihood of killing the bug is > 0.5 it is assumed dead, if it is < 0.5 it is assumed alive.

The outcome variable – which must be coded as 0 and 1 – is placed in the first box labeled Dependent, while all predictors are entered into the Covariates box (categorical variables should be appropriately dummy coded). SPSS predicts the value labeled 1 by default, so careful attention

should be paid to the coding of the outcome (usually it makes more sense to examine the presence of a characteristic or "success."

Mathematically logistic regression estimates a multiple linear regression function defined as:

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\beta}_0 + \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2 - \hat{\beta}_3 X_3 + \ldots \hat{\beta}_p X_p \ldots\ldots\ldots (3.1)$$

$$i = 1,2,\ldots,n$$

When selecting the model for the logistic regression analysis another important consideration is the model fit. Adding independent variables to a logistic regression model will always increase its statistical validity, because it will always explain a bit more variance of the log odds (typically expressed as R^2). However, adding more and more variables to the model makes it inefficient and over fitting occurs.

Nevertheless, many people want an equivalent way of describing how good a particular model is, and numerous pseudo-R^2 values have been developed. These should be interpreted with extreme caution as they have many computational issues which cause them to be artificially high or low. A better approach is to present any of the goodness of fit tests available; Hosmer-Lemeshow is a commonly used measure of goodness of fit based on the Chi-square test (which makes sense given that logistic regression is related to cross tabulation).(www.Statistic solutions.com).

3.3 Stata programming:

Stata is general-purpose statistical software that is developed and sold by Stata Corp LP, located in College Station, Texas. Currently in version 11 which began distribution in July 2009, Stata is known for its wide range of statistical routines, ease of data management, and custom publication-quality graphics. Stata is available on virtually all computer platforms—including Windows, Macintosh, and Unix/Linux—and is designed to function identically across all Stata may be used in an interactive mode, and those learning the package may wish to make use of the menu and dialog system. But when you execute a command from a pull-down menu, it records the command that you could have typed in the Review window, and thus you may learn that with experience you could type that command (or modify it and re-submit it) more quickly than by use of the menus.

This approach allows for all previous work to be readily reproduced through the use of batch files, known as 'do-files' and 'log-files'. It also

allows for the continual expansion of Stata's capabilities. The vast majority of Stata commands are written in Stata's own programming language – the 'ado-file' language. Hence, Stata users can also write commands that will work just like official commands.

Let us consider the form of Stata commands. One of Stata's great strengths, compared with many statistical packages, is that its command syntax typically follows strict rules: in grammatical terms, there are few irregular verbs. This implies that when you have learned the way a few key commands work, you will be able to use many more without extensive study of the manual or even the on-line help. The fundamental syntax of all Stata commands follows a template. Not all elements of the template are used by all commands, and some elements are only valid for certain commands. But where an element appears, it will appear in the same place, following the same grammar. Like Unix or Linux, Stata is case sensitive. Commands must be given in lower case. For best results, keep all variable names in lower case to avoid confusion.

The general syntax of a Stata command is as follows:

[prefix_cmd:] cmdname [varlist] [=exp] [if exp] [in range]
[weight] [using. . .] [,options]

where elements in square brackets are optional for some commands. In some cases, only the cmdname itself required. For example, describe without arguments gives a description of the current contents of memory (including the identifier and timestamp of the current data set), while summarize without arguments provides summary statistics for all numeric variables. Both may be given with a varlist specifying the variables to be considered.

3.3.1 Syntax and options:

There are two or more parenthesized equations specified to the command. The first parenthesized equation is the dependent variable Y, followed by an equal sign, followed by the list of covariates measured without error X_Z. Subsequent parenthesized equations list information on the measurement error variables .

For each unobserved variable, an equation is specified with a label, followed by a colon, followed by the list of replicate variables for that unobserved variable. Hardin and Carroll (2003) explain that these lists must all be the same length and the order
matters when specifying the replicate variables for the unobserved variable. If the measurement error equations all have a single replicate, the user must specify the variance for the measurement error variance. This is specified with the suuinit() option which names a matrix communicating the measurement error variance. This matrix should be square and of dimension the number of measurement error equations. For example,

rcal $Y = (Z_1\ Z_2\ Z_3)\,(X_1 : W_{11}\ W_{12}\ W_{13})\,(X_2 : W_{21}\ W_{22}\ W_{23})$
specifies a model with dependent variable y, X_Z given by $[Z_1\ Z_2\ Z_3]$, and includes information for two unobserved variables. The labels x1 and X_2 are restricted to not coincide with any variable names that exist in the dataset. The unobserved variable represented by the x1 label will be estimated by the 3 replicates stored in the W_{11}, W_{12} and W_{13} variables. Likewise, we have a similar description for the unobserved variable represented by the X_2 label. Utilizing the notation laid out in Hardin and Carroll (2003), we have X_U represented by $[X_1, X_2]$. The replicate observed variables are X_W specified by $[(W_{11}\ W_{12}\ W_{13}),\ (W_{21}\ W_{22}\ W_{23})]$. Writing the X_W matrix in this manner makes it clear that there are two unknown covariates so that the dimension of the measurement error variance is (2 × 2). Since there are three possible replicates for each observation for the two different unobserved variables in this particular model, the user is not required to specify the measurement error variance. However, this does not preclude the user from providing this estimate if it is available. An additional consideration of user-specified measurement error variances is the confidence of the estimate. If the measurement error variance is specified, then the variability of obtaining this estimate is ignored in the calculation of the standard errors

3.3.2 Computational note on calculating standard errors:

The calculation of the asymptotic and sandwich standard errors are very computationally intensive. Since much time was spend optimizing the algorithm, the results will generally be available within seconds for small to moderate size data sets. For large data sets, however, the time required to compute the standard errors can be very long.

Should rcal require more than 30 seconds to provide an answer, it will display an estimate of the time it will take to complete the calculation.

The rcal command implements a fast internal bootstrap which for large data sets is significantly faster than the default asymptotic standard errors (when using the default or a reasonable number of bootstrap replicates). The larger the data set the more of a speed advantage the bootstrap will have over asymptotic standard errors. In one case a hundred fold speed differential was observed for a data with of 100,000 observations.

It is also possible to calculate the naive variance regression calibration estimate using the naive option. Naive standard errors are calculated by generating the missing covariates and assuming them to be true. In general this option is for pedagogical and diagnostic purposes only since the standard errors it calculates are incorrect. One advantage naive estimates do have is that they require no special computational effort.

That makes them useful when testing the rcal command and options on a large data set. Another use would be if no standard errors are desired, say if one wanted to use the rcal command to calculate the jackknife standard errors

3.3.3 Robust estimation of variance:

3.3.3.1The vce() command:

The vce() option causes Stata to change the way standard error is calculated. The vce option has three major types of variance estimators: likelihood-based, replication-based and sandwich estimators.

The two likelihood estimator subcommands are vce(oim) short for observed information matrix and vce(opg) short for outer product of the gradient vectors. Both refer to the matrices and math which underly the procedure.

The two replication-based estimators are vce(bootstrap) and vce(jackknife) To oversimplify, bootstrapping takes a series of random samples from the sample and uses this constant sampling with replacement to calculate standard error. This makes it useful for populations whose normality is uncertain. A brief explanation (with minimal math) of bootstrapping can be found here. Jackknifing is a somewhat similar procedure but where bootstrapping does relatively infinite sampling with replacement, Jackknifing does resampling equal to n and each iteration takes exactly one person out of the sample and recalculates the desired statistic.

The two sandwich estimator subcommands are , vce(robust) which uses a Huber/Whites/sandwich estimator and , vce (cluster [cluster variable]. Using the ,vce (cluster [cluster variable] command negates the need for independent observations, requiring only that from cluster to cluster the observations are independent. Additionally, the Stata User's Guide [U] has a subsection specifically on robust variance estimates and the logic behind them. Both of these adjustments alter the precise interpretation of your data, so be aware of the implications if you use them.

3.4 Blood cancers

Blood cancers affect the production and function of blood cells. Most of these cancers start in bone marrow where blood is produced. Stem cells in bone marrow mature and develop into three types of blood cells: red blood cells, white blood cells, or platelets. In most blood cancers, the normal blood cell development process is interrupted by uncontrolled growth of an abnormal type of blood cell. These abnormal

blood cells, or cancerous cells, prevent blood from performing many of its functions, like fighting off infections or preventing serious bleeding.

There are three main types of blood cancers:

Leukemia, a type of cancer found in blood and bone marrow, is caused by the rapid production of abnormal white blood cells. The high number of abnormal white blood cells are not able to fight infection, and they impair the ability of the bone marrow to produce red blood cells and platelets.

Lymphoma is a type of blood cancer that affects the lymphatic system, which removes excess fluids from body and produces immune cells. Lymphocytes are a type of white blood cells that fight infection. Abnormal lymphocytes become lymphoma cells, which multiply and collect in lymph nodes and other tissues. Over time, these cancerous cells impair your immune system.

Myeloma is a cancer of the plasma cells. Plasma cells are white blood cells that produce disease- and infection-fighting antibodies in body. Myeloma cells prevent the normal production of antibodies, leaving body's immune system weakened and susceptible to infection. (www.hematology.org)

3.5 variables of Research:

3.5.1 Age:

Cancer is primarily a disease of older people, with incidence rates increasing with age for most cancers

Children aged 0-14, and teenagers and young adults aged 15-24, each account for less than one per cent of all new cancer cases in the UK (2011-2013) for example Adults aged 25-49 contribute a tenth (10%) of all new cancer cases, with twice as many cases in females as males in this age group. Adults aged 50-74 account for over half (53%) of all new cancer cases, and elderly people aged 75+ account for over a third (36%), with slightly more cases in males than females in both age groups. There are more people aged 50-74 than aged 75+ in the population overall, hence the number of cancer cases is higher in 50-74s, but incidence rates are higher in 75+. (www.cancerresearchuk.org)

3.5.2 Pcv:

The Pcv test is used to measure the amount of cells in the blood; blood is made up of cells and plasma. The amount of cells in the blood is

expressed as a percentage of the total volume of blood; for example, a Pcv measurement of 50% means that there are 50 millilitres of cells per 100 millilitres of blood.

The Pcv measurement may increase or decrease depending on the individual's health; if they are dehydrated, the measurement will rise and the measurement will decrease if the individual has a condition, such as anaemia.

The Pcv test is usually ordered as part of the series of tests that make up the full blood count. The test is used to diagnose and monitor conditions including anaemia, polyctyhaemia and dehydration. The test may also be used to determine whether an individual is fit to have a blood transfusion; the test may also be repeated regularly to check whether the transfusion has worked effectively.

The test is usually ordered to monitor the condition of people who have been diagnosed with anaemia; it may also be used to monitor those with dehydration and persistent bleeding.

The test is done by taking a sample of blood from the patient; in most cases, the sample is taken from a vein in the patient's arm. A needle is inserted into the vein and the blood is drawn out and collected in a syringe. Once the sample has been collected, it will be bottled, labelled with the patient's name and sent off to a laboratory for testing.

In children, a sample may be collected from the fingertip; in infants it may be collected from the heel. The samples are obtained by pricking the finger or the heel with a needle.

A decreased Pcv result usually indicates anaemia. A low Pcv count may also indicate vitamin or mineral deficiencies, liver cirrhosis and bleeding episodes.

Increased Pcv results are usually associated with dehydration; in most cases, the Pcv result will return to normal once the individual has increased their fluid intake.

High Pcv results may also be caused by polycythaemia vera, a condition which occurs when an individual has too many red blood cells; this is caused by a problem with the function of the bone marrow.

Living at high altitude usually increases Pcv. Pregnancy usually causes Pcv results to be slightly lower than normal. (www.medic8.com)

3.5.3 Mch:

Mch is the initialism for Mean Corpuscular Hemoglobin. Taken from Latin, the term refers to the average amount of hemoglobin found in red blood cells.

A CBC (complete blood count) blood test can be used to monitor Mch levels in blood. Lab Tests Online explains that the Mch aspect of a CBC test "is a measurement of the average amount of oxygen-carrying hemoglobin inside a red blood cell. Macrocytic RBCs are large so tend to have a higher Mch, while microcytic red cells would have a lower value." Mch levels in blood tests are considered high if they are 35 or higher. A normal hemoglobin level is considered to be in the range between 26 and 33 picograms per red blood cell.

High Mch levels can indicate macrocytic anemia, which can be caused by insufficient vitamin B12. Insufficient folic acid can be another cause of macrocytic anemia.

Alcohol abuse can be a contributing factor, and should be disclosed in the diagnostic process to better enable accuracy in diagnosis and in treatment determination.

A simple calculation is used to determine the mean corpuscular hemoglobin level in blood. According to Med Friendly, the total amount of hemoglobin in the sample is multiplied by ten and then divided by the number of red blood cells present.

The method recommended to treat a high Mch level will depend upon what is causing it in the patient. The treatment will also depend upon other medical conditions and any medications the patient may be taking. Allergies will also be taken into account.

Any person with a high Mch level should carefully discuss treatment with his physician and follow the directions carefully. Any dietary supplements and over-the-counter medications should be disclosed in order to prevent any negative results.

If the cause is macrocytic anemia, the treatment could involve adding liver to the diet or adding more vitamin B12.(www.brighthub.com)

3.6 Application Aspect:

The application aspect includes to what explained in the theoretical aspect and depending on Statistical software "SPSS & STATA", we would describe the data, estimate parameters models and comparison between calibration and logistic models.

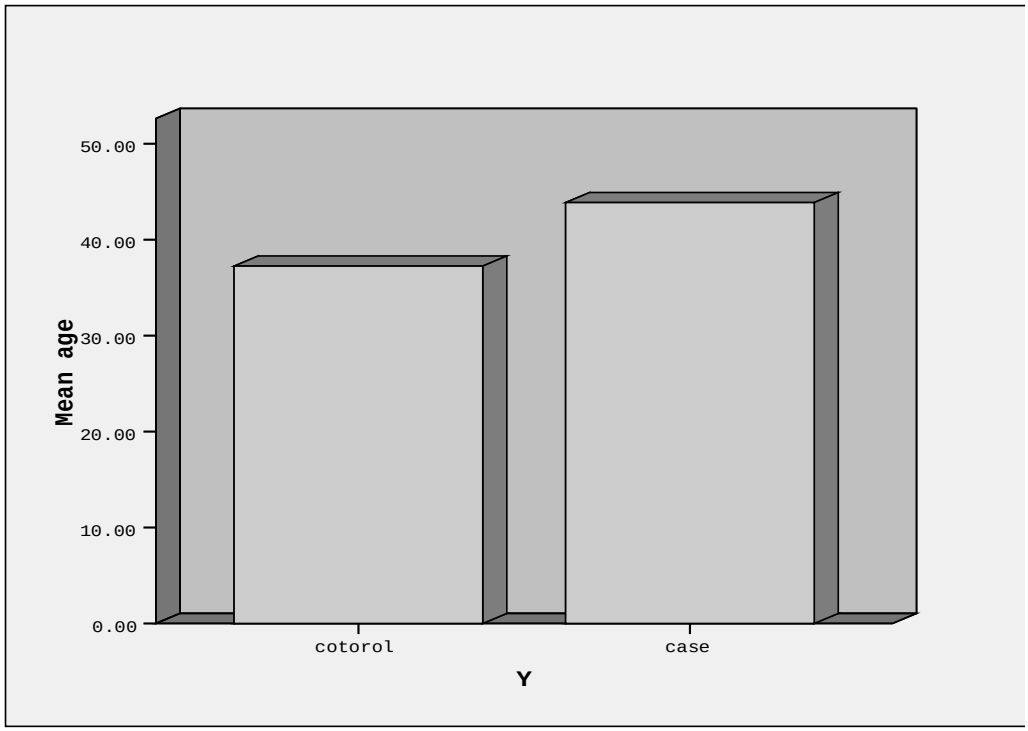
(Table 3.1): Describtives

	N	*Mean*	*Std.Devision*	*Std.Erorr*

Age	**control**	20	37.250	6.307	1.410
	case	100	43.870	10.556	1.056
Pcv	**control**	20	38.050	1.791	0.400
	case	100	31.985	5.454	0.545
Mch	**control**	20	28.550	1.234	0.276
	caserr	100	27.030	2.401	0.240

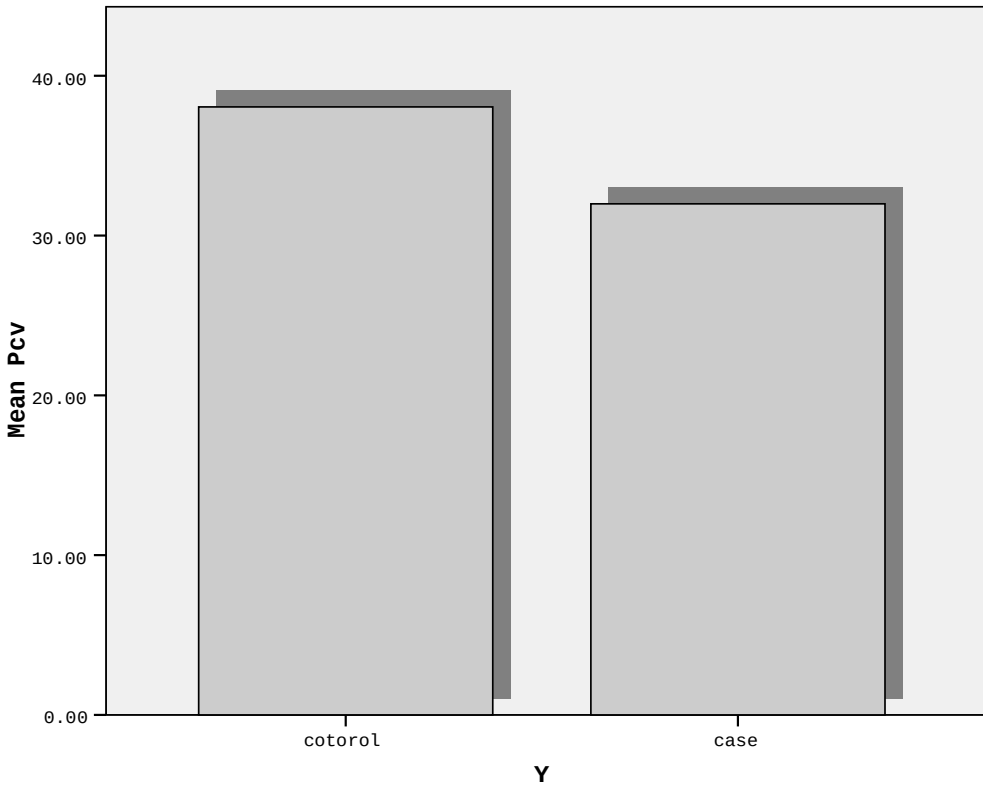
Source: The researcher from applied study, SPSS Package, 2016

Figure(3.1): bar chart explain variable Y and mean of age



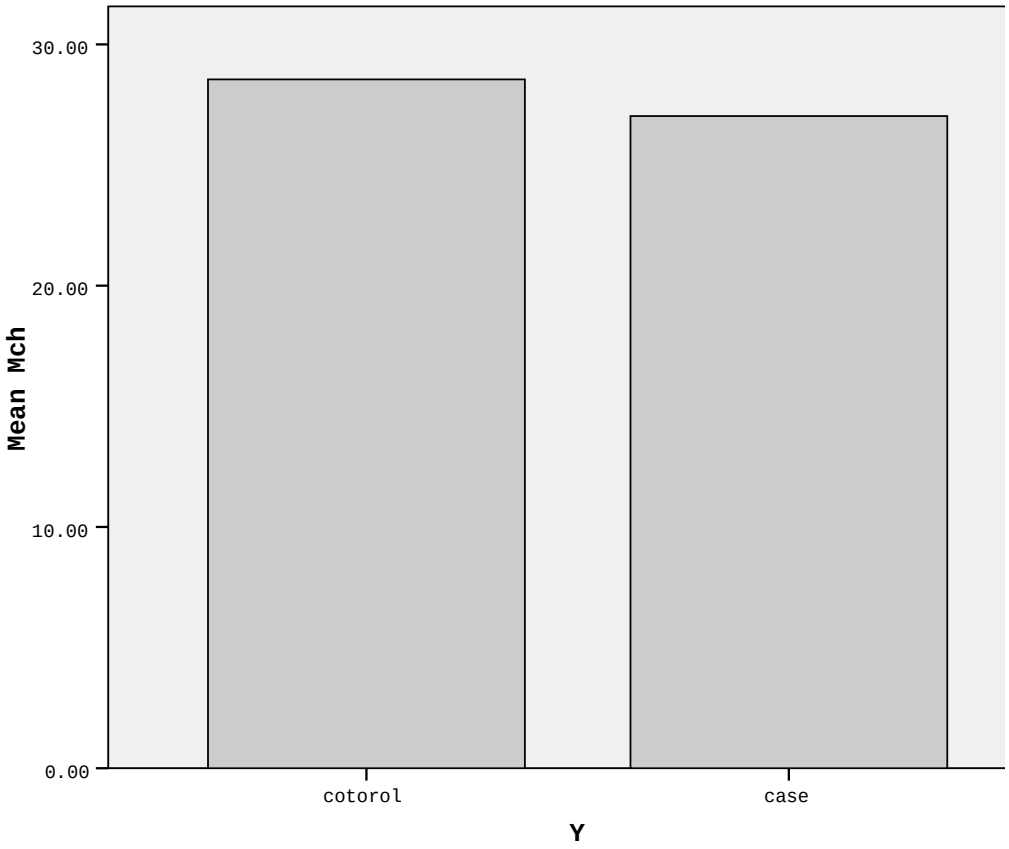
Source: The researcher from applied study, SPSS Package, 2016

Figure(3.2): bar chart explain variable Y and mean of pcv



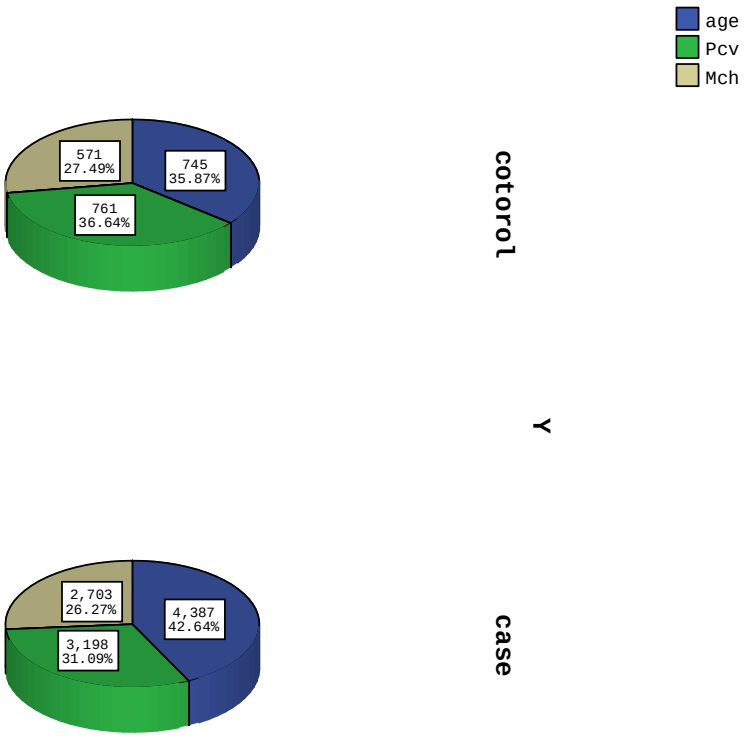
Source: The researcher from applied study, SPSS Package, 2016

Figure(3.3): bar chart explain variable Y and mean of mch



Source: The researcher from applied study, SPSS Package, 2016

Figure(3.4): pie chart explain variable Y and variables age ,pcv and mch



Source: The researcher from applied study, SPSS Package, 2016

From the table(3.1) and the figure (3.1), (3.2), (3.3) and (3.4) in the variable Age we find the mean of the infected 37.250 while the fit 43.870 year with standard error 6.307and 10,556 respectively .for the variable Pcv the average of the infected is 38.050 and the fit 31.985 years with standard error 1.791 and 5.454 .

The average of Mch for the fit is 27.030 year and the infected 28.550

Table(3.2): Variables in Equation

	β	*S.E*	*Wald*	$d.f$	*sig*	$Exp(\beta)$	0.95% C.I for Exp(β)	
							Lower	*Upper*
Age	0.121	0.044	7.572	1	0.006	1.128	1.035	1.229
Pcv	-0.288	0.084	11.617	1	0.001	0.750	0.636	0.885
Mch	-0.344	0.160	4.611	1	0.032	0.709	0.518	0.970
Constant	16.591	5.690	8.503	1	0.004	16046912		

Source: The researcher from applied study, SPSS Package, 2016

The table (3.2) includes all the models parameters in addition to some statistics like standard errors of the model , Wald statistics , degrees of freedom , and the final two columns are the exponential function added to both confidence interval columns. The null hypotheses assumption that is to be tested to know if the model parameters are influencing the dependent variable or not ($H_0 = 0$). And we notice that all parameters are significant which means the non existence assumption is rejected meaning that the variables age pcv and mch affect cancer resilience. Logistics inclination model can be written as follows:

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 16.591 + 0.121\text{Age} - 0.288\text{Pcv} - 0.344\text{Mch}$$

Where $\hat{p}$ refers to blood cancer infection and these estimations clarify the relation between the dependent variable (infection / non infection) and the independent variable using ((log it)) units yet we noticed that the independent variable is most affected by the Pcv variable with a parameter -0.288 and the value of $sig = 0.001$ and on the lower degree of affecting the independent variable comes the Age variable with a parameter of 0.121 and $sig = 0.006$. On the final level comes the variable Mch with a parameter of -0.344 and a probable value of 0.032.

Table (3.3):Omnibus Tests of Model Coefficients

	Chi – square	$d.f$	*sig*
Model	38.450	3	0

Source: The researcher from applied study, SPSS Package, 2016

The table (3.3) tests the quality of the logistic model being used , we found the value of χ^2 test equals 38,450 and is significant at $\alpha = 0.05$ and the probable value equals zero which asserts the previous result , meaning that the null hypothese assumption has been rejected which means the model is significant.

Table(3.4): Hosmer and Lemeshow test

Chi – square	*d.f*	*sig*
7.287	8	0.506

Source: The researcher from applied study, SPSS Package, 2016

The table (3.4) clarifies that χ^2 test equals 7.285 with a degree of freedom equal 8 and probable value 0.506 which asserts the goodness of the method as a whole.

Table(3.5):Classification Table

Observed		predicted		
		Y		Percentage Correct
		Control	Case	45.0
Y	Control	9	11	94.0
	Case	6	94	85.8
Over all precentage				

Source: The researcher from applied study, SPSS Package, 2016

The table(3.5) clarifies the percentage of the right classification which is 85.8% and the percentage of wrong classification 14.2% which means the model presents the data very well.

Table(3.6): standard errors by using vce(robust)

Y	*Coef.*	*Robust* Std.Error	*t*	$P>\|t\|$	0.95% C.I	
Age	0.009	0.003	3.39	0.001	0.004	0.015
Pcv	-0.024	0.006	-4.13	0	-0.356	-0.012
Mch	-0.028	0.012	-2.29	0.024	-0.051	-0.004
Constant	1.987	0.282	7.04	0	1.428	2.546

Source: The researcher from applied study, STATAPackage, 2016

The table (3.6) clarifies regress of the dependent variable (y) (infected / not infected) on the dependent variables mentioned before , the aim behind establishing this model is to find standard errors , known here by (

Robust Std. Error) then comparing it with standard errors column in table (2).

We notice here that the value of trusted calibration errors is less from standard errors in logistic regression model table (3.2). And it is known that the less errors in a model , the more accurate and more capable of forecast it becomes . Meaning that by using calibration linear regression model , error were made less with more accuracy. However logistic was well tested in table (3.5) and (3.6) yet calibration regression was noticeably better.