Methods of Multivariate Analysis

There are three categories of analysis:

- **Univariate analysis**, which looks at just one variable
- **Bivariate analysis**, which analyzes two variables simultaneously
- **Multivariate analysis**, which looks at more than two variables simultaneously

Univariate and Bivariate analyses can be treated as a particular case of Multivariate analysis.

Multivariate analysis techniques can be divided into two categories:

- **Dependence techniques:** A dependence technique may be defined as one in which a variable or set of variables is identified as the dependent variable to be predicted or explained by other variables known as independent variables.
  Examples: **Generalized Linear Model** (includes *Simple Linear Regression, Multiple Linear Regression, Logistic Regression,* Poisson Regression, Negative Binomial Regression), Principal Component Analysis, Canonical correlation analysis, Analysis of Variance (ANOVA), Multivariate Analysis of Variance (MANOVA), Log-linear model for contingency table, etc.

- **Interdependence techniques:** An interdependence technique is one in which no single variable or group of variables is defined as independent or dependent. Instead, the procedure involves the simultaneous analysis of all variables in the set.

  Examples: Factor Analysis, Cluster Analysis, Multidimensional Scaling (MDS), Correspondence Analysis, etc.

Regression Analysis (dependence technique)

Regression analysis measures the probable movement of the Dependent (Response or Endogenous) variable for a unit increase of the independent variable.

Regression analysis is used for one of two purposes:

- predicting the value of the **dependent** variable when information about the independent variables is known
- predicting the effect of an **independent** variable on the dependent variable.

Definition

Dependent variable: the variable we wish to explain or predict.

Independent/exogenous/explanatory variable: the variable we use to explain or predict the dependent variable.

Types of Regression Models

There are numerous regression analysis approaches available for making predictions. Various parameters, including the number of independent variables, the form of the regression line, and the type of dependent variable, determine the choice of technique for regression analysis.

1. **Simple Linear Regression** (Linear relationship exists between dependent and independent variables)
2. **Multiple Linear Regression** (linear relationship exists between dependent and more than one independent variable)
3. **Binary Logistic Regression** (dependent variable is binary)

4. **Multicategory Logistic regression** (dependent variable is multicategory)
5. **Polynomial Regression:** (nonlinear relationship between dependent and independent variables)
6. **Ridge Regression**: (independent variables are highly correlated)
7. **Quantile Regression** (when outliers, high skewness and heteroscedasticity exist in the data.).
8. **Bayesian Linear Regression**
9. **Principal Components Regression** (for many independent variables or multicollinearity exist in data).

10. **Partial Least Squares Regression** (many independent variables with a high probability of multicollinearity between the variables).

10. **Elastic Net Regression** (suitable for strongly correlated data.

11. **Support Vector Regression** (suitable for linear and nonlinear models).

12. **Ordinal Regression** (for ordinal dependent variable)

13. **Poisson Regression** (when the dependent variable has count data)

14. Negative Binomial Regression (used for overdispersed count dependent data)

15. **Quasi-Poisson Regression** (used for overdispersed count-dependent data)

16. **Tobit Regression** (when censoring exists in the dependent variable)

17. **Jackknife regression** (a resampling procedure)
18. **Ecological Regression** (used to study predicted human behaviour within a population data set), etc.

1. Simple linear regression model:

(Simple – the model contains only one independent variable, linear – the power of the regression coefficient is 1).

The population Simple Linear Regression Model can be stated as

$$y = \alpha + \beta X + \varepsilon,\ \ \varepsilon(\text{the random error}) \sim N\left(0, \sigma^2\right)$$

Thus, $y \sim N\left(X\beta, \sigma^2\right)$ and *X* is fixed for a particular analysis.

Before estimating the model, we must visualize the data to check the

- the linear relationship between X and Y,
- normality of Y
- presence of an outlier.

If Y is not normally distributed, we can use **Box-cox transformation** to transform Y to normal.

Using the Method of least square principle, we can estimate the model as

$$\hat{y} = a + bX\text{, where } a = \overline{y} - b\overline{x}\text{, and } b = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}$$

[Least square principle: Minimize the sum of squares of errors (SSE), $\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n}\left(y_i - (a + bx)\right)^2$ with respect to *a* and *b*, where $\hat{y}$ is an estimate of $E(y / X = x) = \alpha + \beta x$, the conditional mean of y given X=x (fixed).]

The accuracy of the estimated model can be evaluated by the adjusted coefficient of determination $R^2\left(\overline{R}^2\right)$, which varies from 0 to 1.

(When should we use a multiple linear regression model?)

If the coefficient of determination is unsatisfactory (low), we incorporate/add meaningful and relevant independent variables into the model to create the Multiple Linear Regression Model (MLRM).

2. **Multiple Linear Regression Model** (MLRM)

The population multiple regression model with k independent variables can be written as

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + +\beta_3 X_3 + \cdots + \beta_k X_k + \varepsilon, \text{where } \varepsilon \sim N\left(0, \sigma^2\right)$$

In matrix form,

$y = X\beta + \varepsilon$, $\left[y_{n\times 1} = X_{n\times k}\beta_{k\times 1}\right]$, n is the sample size or number of observations.

Using the Method of least square, the estimated regression model can be written as

$$\hat{y} = b_0 + b_1 X_1 + b_2 X_2 + +b_3 X_3 + \cdots + b_k X_k$$

In matrix form, $\hat{y} = X\hat{\beta}$ $\quad$ (or, $\hat{y} = Xb$)

where, $b = \hat{\beta} = \left(X'X\right)^{-1} X'y$

The total variation of y can be expressed in terms of variation due to Regression and Error. Mathematically,

$$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2 + \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

$$\text{Total SS (TSS)} = \text{Regression SS (RSS)} + \text{Error SS (ESS)}$$

The above expression is beneficial in constructing the ANOVA table and Testing hypotheses.

Steps testing the validity or accuracy of the estimated model:

Test whether all independent variables have a simultaneous influence on the target variable or not (Global Test).

We require the following ANOVA table to test

$H_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0$ against

$H_1:$ At least one regression coefficient (β_i) is non-zero

Under H_0 the test statistic is

$$F = \frac{MSSR/k}{MSSE/(n-k-1)} \sim F(k, n-k-1)$$

ANOVA table

SV	df	SS	MSS	F
Regression	k	RSS= $\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2$	MSSR=RSS/k	$F = \dfrac{MSSR/k}{MSSE/(n-k-1)}$
Error	n-(k+1)	ESS= $\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$	MSSE=ESS/(n-k-1)	
Total	n-1	TSS= $\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$		

Abbreviation: SV – Source of variation, df – degrees of freedom, SS – Sum of Squares, MSS – Mean Sum of Squares, RSS – Regression Sum of Squares, ESS – Error Sum of Squares, TSS – Total Sum of squares.

We will not proceed anymore if H_0 is accepted. We continue the investigation/analysis if at least one regression coefficient is non-zero, i.e., at least one independent variable has a linear influence on the dependent variable. (use the p-value of the ANOVA table to make a decision).

The dependent variable is continuous in Multiple Linear Regression because of the normality of the error term.

Note that a scientific calculator can estimate a simple linear regression model and correlation coefficient, whereas multiple regression and logistics regression require a computer.

We test each regression coefficient or a subset of the coefficients if the above global Test is rejected.

To test

$H_0 : \beta_i = 0$ (X_i has no linear influence on y) against

$H_a : \beta_i \neq 0$ (X_i has a linear influence on y),

Under H_0 the test statistic is

$$t = \frac{b_i - 0}{se(b_i)} \sim t(n - k - 1)$$

We do not reject H_0 if p-value>significance level. (p-value can be obtained directly from computer output but not from a statistical table).

Please study the file SimpleLinearRegEx1.pptx for a better understanding.

Example 1 (Simple Linear Regression Analysis):

A real estate agent wishes to examine the relationship between a home's selling price (measured in $) and its size (measured in square feet).

House Price in $1000s (X)	Square Feet (Y)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Questions:

a) Identify the dependent and independent variables.
b) Construct a scatter diagram and comment on it.
c) Determine and interpret the correlation coefficient and coefficient of determination.
d) Estimate the regression equation of the selling price of a home on its size.
e) Predict the price for a house with 2000 square feet.
f) Test the significance of the linear relationship between price and size at a 5% significance level.
g) Construct a 95% confidence interval for the population correlation coefficient.
h) Test the linear influence of size on price at a 5% significance level.

Solution: (A scientific calculator can be used for this simple linear regression problem)

The majority of the above questions can be answered from the following computer output:

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

The regression equation is:

$$\widehat{\text{house price}} = 98.24833 + 0.10977 \text{ (square feet)}$$

$$\hat{y} = 98.24833 + 0.10977\, X$$

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Please study the file MultipleLinearRegEx2.pptx for a better understanding.

Example 2 (Multiple Linear Regression Analysis):

A distributor of frozen desert pies wants to evaluate factors thought to influence demand. Data are collected for 15 weeks on Pie sales (units per week), price (in $) and Advertising cost($100's)

Week	Pie Sales (Y)	Price ($) (X1)	Advertising ($100s) (X2)
1	350	5.5	3.3
2	460	7.5	3.3
3	350	8	3
4	430	8	4.5
5	350	6.8	3
6	380	7.5	4
7	430	4.5	3
8	470	6.4	3.7
9	450	7	3.5
10	490	5	4
11	340	7.2	3.5
12	300	7.9	3.2
13	440	5.9	4
14	450	5	3.5
15	300	7	2.7

a) Identify the dependent and independent variables.
b) Construct scatter diagrams to check linearity and the presence of outliers.
c) Estimate the regression equation of pie sales on price and advertisement cost. Interpret the estimated coefficients.
d) Predict the pie sales when the price per unit is $6.35, and the advertisement cost is $4100.

e) Test the significance of the linear relationship between price and size at a 5% significance level.
f) Test the joint influence of price and advertisement on pie sales.
g) Construct a 95% confidence interval for the population regression coefficient of pie sales on price per unit.

The majority of the above questions can be answered from the following computer output:

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$R^2 = \frac{SSR}{SST} = \frac{29460.0}{56493.3} = .52148$$

52.1% of the variation in pie sales is explained by the variation in price and advertising

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

The basic framework of the above regression analyses (SLRM and MLRM) is the Classical Regression Model (**CLRM).** The CLRM is based on a set of assumptions. Gujarati (2008) outlined about ten different assumptions. One of the assumptions is that each error term ε_i is independently and normally distributed with mean 0 and constant variance σ^2 (identical). That is,

$$\varepsilon_i \sim NIID\left(0, \sigma^2\right)$$

(NIID means normal, independent and identical distribution)
The above assumption implied that the dependent variable is also normally distributed as follows

$$y \sim NIID(X\beta, \sigma^2),$$

where, $X\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$.
In case of violating at least one assumption of normality, independence, and constant variance, we use a Generalized linear model (GLM) (Agresti, 2019, chapter 3). Simple Linear Regression and Multiple Linear Regression are particular cases of GLM.

We use a generalized linear model for different choices of the dependent variable. A short list is given below:

Types of Generalized Linear Models		
Type/distribution of dependent variable/Random Component	Independent variable/Systematic Component	Model
Continuous/Normal	Continuous (one independent var)	Simple Linnear Regression
Continuous/Normal	Continuous (more than one independent var)	Multiple Linnear Regression
Continuous/Normal	Categorical	Analysis of Variance
Continuous/Normal	Mixed	Analysis of Covariance
Binary/Binomial	Mixed	Binary LOGISTIC REGRESSION
Multicategory/Multinomial	Mixed	Multinomial logistic Regression
Count/Poisson	Mixed	Loglinear
Count with overdispersion /Negative Binomial	Mixed	Negative Binomial Regression
Ordinal/cumulative normal	Mixed	Cumulative Logistic Model

The Simple, Multiple Linear Regression, and binary Logistic Regression models are particular cases of GLM.

BINARY LOGISTIC REGRESSION

Binary dependent outcomes violets the assumption of normality. In this situation, we use binary logistic Regression, as per the above discussion.

Despite its name, logistic Regression is a method of **classification**. It is conceptually similar to linear Regression.

Examples of binary dependent outcomes:

- The patient survives the operation or does not.
- The accused is convicted or is not.
- The customer makes a purchase or does not.
- The marriage lasts at least five years or does not.

Using the technique of GLM, the logistic regression model with k explanatory variables can be expressed in the following form:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k = X\beta.$$

$$\text{Thus,}\quad p = P(Y=1/x) = E(Y=1/x) = \frac{e^{X\beta}}{1+e^{X\beta}} = \frac{1}{1+e^{-X\beta}}.$$

Note that, Logistic Regression is nonlinear in regression coefficients.

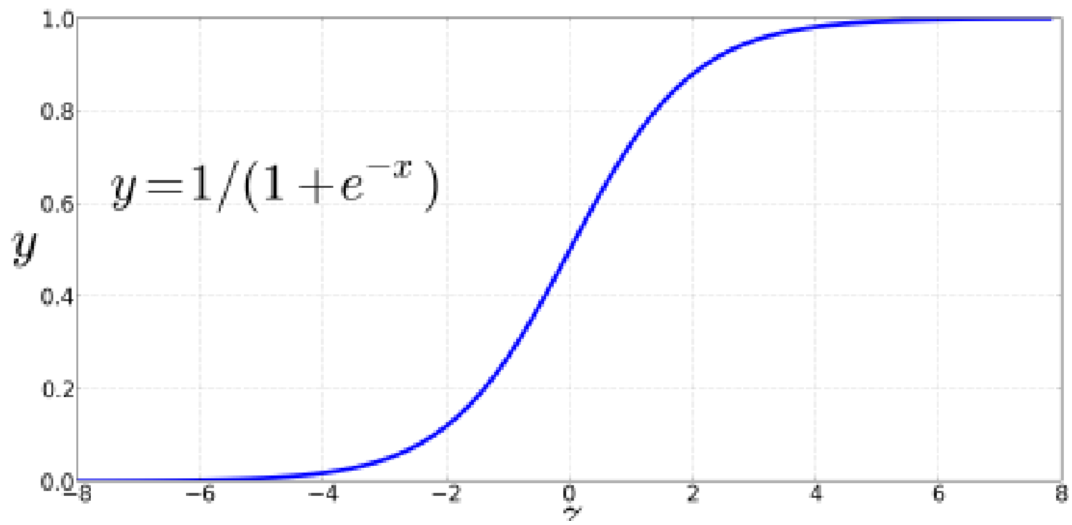
$$\text{Odds} = \frac{p}{1-p} = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)$$

$$= \exp(\beta_0) \times \exp(\beta_1 x_1) \times \exp(\beta_2 x_2) \times \cdots \times \exp(\beta_k x_k)$$

where, $p = \Pr(\mathrm{Y}=1/\mathrm{x})$ and $(1-p) = \Pr(Y=0/x)$

Defn: Odds is the ratio of the probability of happening to the probability of not happening of an event.

The higher the probability, the greater the odds.



The sigmoid function $y = \dfrac{1}{1+e^{-X\beta}}$ (so named because it looks like an s) is also called the logistic function. It takes a real value and maps it to the range [0, 1]. It is nearly linear around 0, but outlier values get squashed toward 0 or 1.

For given values of x's, we can predict p. If p>0.5, the observation(s) belongs to group 1; otherwise, it belongs to group 0.

Note that the Method of least squares is used in MLR and Maximum Likelihood in Logistic Regression. We require a Computer to estimate each model listed in the GLM table.

In terms of log odds, Logistic Regression is like regular Regression

- The exponential function of the logistic regression coefficients are *odds ratios*
- When x_k is increased by one unit and all other independent variables are held constant, the odds of Y=1 are multiplied by e^{β_k}.
- Another way of writing e^{a+bX} is $e^a(e^b)^X$. That means that a one-unit increase in X multiplies the odds by e^b.

The goodness of fit and accuracy of the Binary Logistic Regression model

In Linear Regression, we check adjusted R^2, F Statistics, MAE, and RMSE to evaluate model fit and accuracy.

Logistic Regression employs different sets of metrics. In this case, we deal with probabilities and categorical values. The following are the evaluation metrics used for Logistic Regression:

1. Akaike Information Criteria (AIC): The model with the lowest AIC will be relatively better.
2. Null Deviance and Residual Deviance: The deviance of an observation is computed as -2 times the log-likelihood of that observation. The larger the difference between null and residual deviance, the better the model. (Also, whichever model has a lower null deviance, the model explains deviance pretty well and is better. The lower the residual deviance, the better the model.

Practically, AIC is always given preference above deviance to evaluate model fit.

3. Confusion Matrix

Confusion matrix is the most crucial metric commonly used to evaluate classification models. The skeleton of a confusion matrix looks like this:

	1 (Predicted)	0 (Predicted)
1 (Actual)	True Positive	False Negative
0 (Actual)	False Positive	True Negative

The confusion matrix avoids "confusion" by measuring the actual and predicted values in a tabular format. The table above shows the Positive class = 1 and the Negative class = 0. Following are the metrics we can derive from a confusion matrix:

$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$ (It determines the overall predicted accuracy of the model)

True Positive Rate, $TPR = \frac{TP}{TP+FN}$. (It indicates how many positive values, out of all the positive values, have been **correctly predicted**.

`TPR = 1 - False Negative Rate`. It is also known as **Sensitivity** or **Recall**.

False Positive Rate: $FPR = \frac{FP}{TP+TN}$ **,** It indicates how many negative values, out of all the Negative values, have been **incorrectly predicted**.

`FPR = 1 - True Negative Rate`.

True Negative Rate, $TNR = \dfrac{TN}{TN + FP}$ **,** It indicates how many negative values, out of all the negative values, have been **correctly predicted**. It is also known as **Specificity**.

False Negative Rate, $FNR = \dfrac{FN}{FN + TP}$ **,** It indicates how many positive values, out of all the positive values, have been incorrectly predicted.

$\text{Precision} = \dfrac{TP}{TP + FP}$ **,** It indicates how many values, out of all the predicted positive values, are positive.

F-score is the harmonic mean of precision and recall. It lies between 0 and 1. The higher the value, the better the model. It is formulated as

$$F - score = \frac{2}{\dfrac{1}{\text{Precision}} + \dfrac{1}{\text{Recall}}} = 2(\text{precision*recall}) / (\text{precision+recall}).$$

Example (Daniel: pp.573, EXAMPLE 11.4.2):

Cardiac rehabilitation programs offer "information, support, and monitoring for return to activities, symptom management, and risk factor modification."

The researchers conducted a study to identify factors associated with participation in such programs among women.

The following data (Table 11.4.3 in the text) are the ages of 185 women discharged from a hospital in Australia who met eligibility criteria involving discharge for myocardial infarction, artery bypass surgery, angioplasty, or stent.

We wish to use these data to develop a model (Binary Logistic Regression Model) regarding the relationship between age in years (independent variable,

X) and participation in a cardiac rehabilitation program (dependent variable,

Y) (ATT=1 if participated, and ATT=0 if not).

We also wish to know if we may use the results of our analysis to predict the likelihood of participation by a woman if we know her age:

TABLE 11.4.3 Ages of Women Participating and Not Participating in a Cardiac Rehabilitation Program

age	att		age	att		age	att		age	att		age	att		age	att
50	0		71	0		73	0		75	0		41	1		69	1
59	0		69	0		68	0		68	0		64	1		66	1
42	0		78	0		72	0		81	0		46	1		57	1
50	0		69	0		59	0		74	0		65	1		60	1
34	0		74	0		64	0		65	0		50	1		63	1
49	0		86	0		78	0		81	0		61	1		63	1
67	0		49	0		68	0		62	0		64	1		56	1

44	0		63	0		67	0		85	0		59	1		70	1
53	0		63	0		55	0		84	0		73	1		70	1
45	0		72	0		71	0		39	0		73	1		63	1
79	0		64	0		80	0		52	0		65	1		63	1
46	0		72	0		75	0		67	0		67	1		65	1
62	0		79	0		69	0		82	0		60	1		67	1
58	0		75	0		80	0		84	0		69	1		68	1
70	0		70	0		79	0		79	0		61	1		84	1
60	0		73	0		71	0		81	0		79	1		69	1
67	0		66	0		69	0		74	0		66	1		78	1
64	0		75	0		78	0		85	0		68	1		69	1
62	0		73	0		75	0		92	0		61	1		79	1
50	0		71	0		71	0		69	0		63	1		83	1
61	0		72	0		69	0		83	0		70	1		67	1
69	0		69	0		77	0		82	0		68	1		47	1
74	0		76	0		81	0		85	0		59	1		57	1
65	0		60	0		78	0		82	0		64	1		66	1
80	0		79	0		76	0		80	0		62	1			
69	0		78	0		84	0		74	1		74	1			
77	0		62	0		74	0		50	1		61	1			
61	0		73	0		59	0		55	1		69	1			
72	0		46	0		81	0		66	1		76	1			
67	0		57	0		74	0		49	1		71	1			
73	0		53	0		77	0		55	1		61	1			
75	0		40	0		59	0		73	1		46	1			

(Data file: Logisticdata1.xlsx)

Partial SPSS output

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	229.520[a]	.037	.051

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Confusion/Classification Table[a]

	Predicted		
	ATT		Percentage
Observed	0	1	Correct

Step 1	ATT	0	111	10	91.7
		1	58	5	7.9
	Overall Percentage				63.0

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	-.038	.015	6.710	1	.010	.963
	Constant	1.875	.981	3.653	1	.056	6.519

From the above SPSS output, we can write the **estimated Binary Logistic Regression Model as**

$$\hat{y}_i = \ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \hat{\alpha} + \hat{\beta}x_i = 1.875 - 0.038x_i$$

The predicted probability of attending cardiac rehabilitation for a woman aged x_i is

$$\hat{p}_i = \frac{1}{1+e^{-(1.875-0.038x_i)}}$$

For $x = 57$, $\hat{p} = \frac{1}{1+e^{-(1.875-0.038\times 57)}} = 0.427759$

$\hat{p}_{57} = 0.427759 < 0.50$, Thus, a 57-year-old woman did not participate in the program.

For $x = 37$, $\hat{p} = \frac{1}{1+e^{-(1.875-0.038\times 37)}} = 0.615147$

$\hat{p}_{37} = 0.615147 > 0.50$, Thus, a 37-year-old woman participated in the program.

Test: We can check the adequacy of the logistic model by testing the null hypothesis that the slope of the regression line/coefficient of age (x) is zero. That is, we test the null hypothesis

$H_0 : \beta = 0$ versus the two-sided alternative $H_a : \beta \neq 0$.

Under the null hypothesis, the test statistic is

$$W = \left(\frac{\hat{\beta}}{se(\hat{\beta})} \right)^2 \sim \chi_1^2$$ (distributed as Chi-square with 1 degree of freedom)

.

From computer output, $W = 6.710$ with p-value 0.01<0.05.

Thus, we reject the null hypothesis at a 5% significance level.

The logistic regression coefficient is significant, and hence, the logistic regression model is adequate. That is, the age of a woman influences her participation in the program.

Accuracy can be observed from the **Confusion matrix,** and various accuracy measures can be obtained from this confusion matrix.

It is observed from the Confusion/Classification table that only 63% of the data were correctly reclassified, with those participating in the rehabilitation program much more poorly classified than those who did not attend the program. The frequency distribution shows the large number of ATT=1 subjects who were misclassified as ATT=0 based on the model.

References:

Agresti, A. (2019) An Introduction to categorical Analysis (Chapter 4), Wiley & Sons.

Agresti, A. (2019). AN INTRODUCTION TO CATEGORICAL DATA ANALYSIS, chapter 3, Wiley.

Daniel, W. W. (2013). BIOSTATISTICS: A Foundation for Analysis in the Health Sciences, (Chapter 9-11).

Gujarati, D. (2014). Econometrics by Example. Chapter 1, Palgrave.

Hair, F. F. (2019). Multivariate Data Analysis, (Chapter 1), Cengage Learning.

Newbold, P. (2023). Statistics for Business and Economics, Pearson (