CARLETON UNIVERSITY

SCHOOL OF
MATHEMATICS AND STATISTICS

HONOURS PROJECT

TITLE: Modelling the Choice of Surgery for Rectal Cancer Using Logistic Regression and ROC Curves

AUTHOR: Duy Thai Doan

SUPERVISOR: Patrick Farrell

DATE: 2019-05-07

Carleton UNIVERSITY

CONSENT FOR DISCLOSURE OF FOUR YEAR HONOURS PROJECT

I authorize the

School of Mathematics and Statistics
Office/Program/Individual

To use my Four Year Honours Project

Submitted on

2019-05-07
Date submitted to Honours Coordinator

For the purpose of

Research and learning
State specific purpose of information release

In the period

Indefinite
State date range for which permission will exist

Full Name:	**Duy Thai Doan**
Student I.D. #:	**101011007**
Date:	**2019-05-07**

Signature D.D

Protection of Privacy
The personal information requested on this form is collected under the authority of Section 42 (R.S.O.) 1990, c. F.31) of the *Freedom of Information and Protection of Privacy Act* and will be protected under Part 3 of that *Act*. It will be used for the purpose of managing the consent for disclosure of personal information process. Direct any questions about this collection to*:* [contact position, full address, and business telephone number].

TABLE OF CONTENT

Title	Page No.
1. Introduction	2
- **Choice of topic**	
2. Logistic Regression	3
- **List of variables included in the analysis**	4
- **Deriving the model (Linear and Non-Linear)**	6
- **Deriving the maximum likelihood estimators**	10
3. Logistics Regression applications	14
- **Derived models using forward selection**	
4. Odds Ratio	23
- **Comparing explanatory variable**	
5. Receiver Operating characteristic curve	28
6. Conclusion	34
7. Citation	35

Introduction

Choice of topic

Throughout the last four years at Carleton University, I have learned various topics and concepts in the program of Math: Statistics and Economics that fit into my range of interest. One stands out concept that I really want to use as the topic for my honour project is data analysis using regression. Regression methods are an important component for resolving any data analysis concerns. The common denominator between all regression methods is that they are used to describe and analyze the relationship between a *response variable* (output) and one or continuous or discrete *explanatory variables* (input(s)). Realistically, for many cases in data analysis for real-life usage, the outcome variable is discrete taking on two or more input values. In this situation, the suitable regression method of analysis is the ***Logistic Regression Model (LRM)***. Dr. Patrick Farrell, who is my supervisor for the project, suggested the idea of *LRM* to me and I thought having it as the topic choice for the project would be enjoyable and informative. Around 2 months in the project, he also suggested the application of ROC curve that help to analyze how well a logistic model can perform and I decided to also include it in my project. We are going to describe characteristics and illustrate the applications of LRM using data on **(*)** ***Low Anterior Resection (LAR)*** observed from American College of Surgeons (ACOS).

(*) *LAR* is a surgery that is done to treat cancer of the rectum where the part of your rectum containing the cancer will be removed. The remaining part of the rectum is reconnected to the bowels in the usual way.

Logistic Regression

Let Y denote the *LAR* surgery type response variable. For instance, in this context of the data, Y reflects whether the patient should consider **(**) *Sphincter Sparing Procedure (SSP)*** or (*******) ***Abdominoperineal Resection (APR)*** as the method for *LAR* procedure. We let Y = 1 denotes the fact that *SSP* should be performed and Y = 0 denote that the patient should consider *APR*.

(**) *SSP* is for patients who have colorectal cancer in the lower part of colon, surgeons often remove both the rectum and the anus. This procedure allows for safe removal of the tumor and "spare the anal sphincter muscle".

(***) *APR* for many years was the treatment of choice for most patients with rectal cancer. With introduction of new advances in surgical techniques in modern age, there has been a marked increase in the rate of *SSP* operations while *APR* has decreased in return. However, for selected patients with special or different input factors, especially those with very distal tumors or poor sphincter function, it is still necessary for them to choose *APR* as the choice of method. Essentially, *APR* "completely removes the distal colon, rectum, and anal sphincter" complex using both incisions of anterior abdominal and perineal, resulting in an opening in the colon. Developed more than 100 years ago, *APR* is considered as a much older procedure compare to *SSP*.

List of variables included in the analysis:

HOSPITAL RELATED COVARIATES (based on hospital of j-th patient)

Average annual number of surgical cases for cancer of the rectum:

- *Average caseload (AVGCSLD)* = Quantitative variable

Indicator variables for type of hospital (Baseline: Comprehensive Cancer Centre):

- *HOSTYPE2* = 1 if Teaching Hospital
- *HOSTYPE3* = 1 if Health Maintenance Organization
- *HOSTYPE4* = 1 if hospital with ACOS* approved cancer program
- *HOSTYPE5* = 1 if hospital approved by ACOS*
- *HOSTYPE6* = 1 if hospital not approved by ACOS*
- *HOSTYPE7* = 1 if Veteran's Administration Hospital

  0 otherwise

*American College of Surgeons (ACOS)

PATIENT RELATED COVARIATES (for j-th patient at i-th hospital)

Gender:

- *SEX* = 1 if Male, 0 if Female

Age:

- *AGE* = Quantitative variable

Indicator variables for stage of Rectal Cancer (Baseline: Comprehensive Cancer Centre):

- *STAGE2* = 1 if regional disease - direct extension only
- *STAGE3* = 1 if regional disease – involved lymph nodes only

- *STAGE4* = 1 if regional disease with both direct extension and involved lymph nodes

  0 otherwise

Year of Surgery:

- *YEAR2*
- *YEAR3*
- *YEAR4*
- *YEAR5*

DEPENDENT/OUTCOME VARIABLE (for j-th patient at i-th hospital)

Type of Surgery:

$Y_{i,j}$ = 1 if SSP

0 if APR (requiring colostomy)

Deriving the model (Linear and non-linear)

To continue this example of cancer surgery type, suppose we only study the relationship between LAR and age where it will be treated as the only explanatory variable (continuous), let denote it X. In the data Dr. Farrell gave me, there exist 2005 observations taken in the sample yielding a result of:

Tabulated Statistics: Lar, Age

Rows: Lar Columns: Age

	19	20	21	26	27	28	29	30	31	32	34	35	36	37	38	39	40	41	42	43	44	45
0	1	1	2	1	0	1	1	0	3	2	2	1	2	4	5	4	7	5	7	5	10	9
1	0	0	0	0	1	1	2	2	3	3	3	4	5	4	5	3	6	5	7	6	5	2
All	1	1	2	1	1	2	3	2	6	5	5	5	7	8	10	7	13	10	14	11	15	11

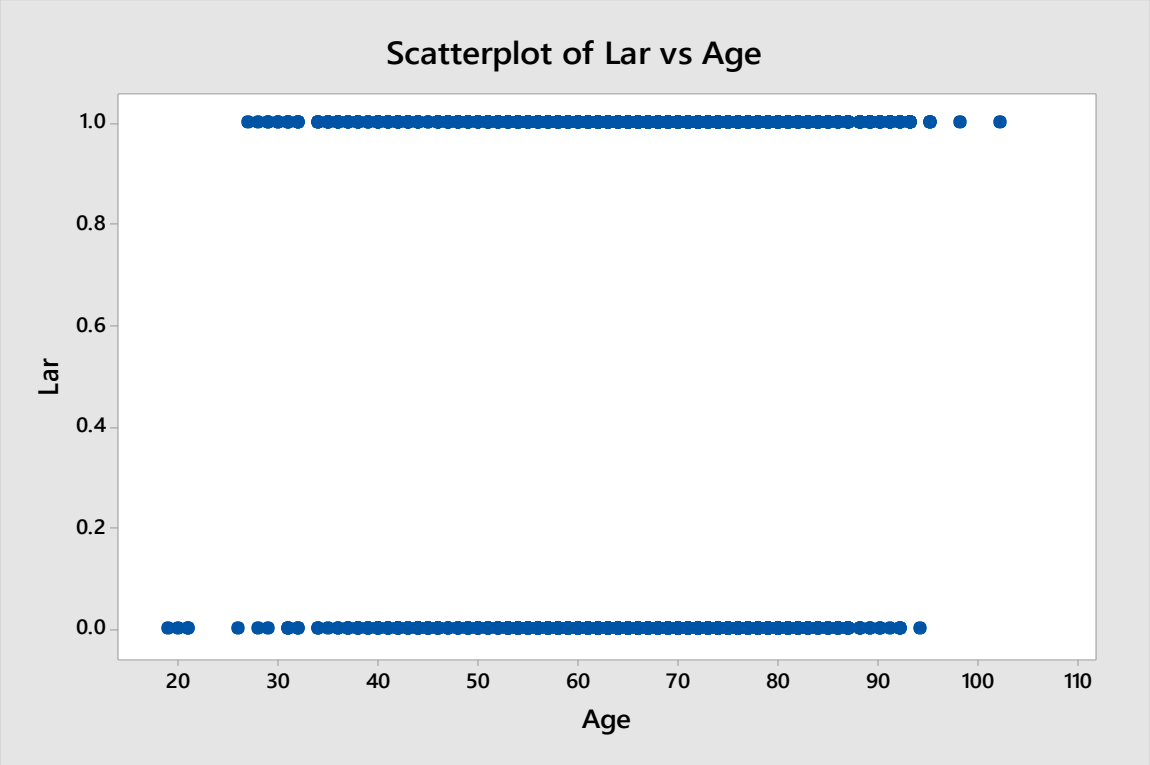
	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67
0	7	3	5	8	7	7	6	13	17	14	18	16	18	25	19	24	26	29	27	30	36	41
1	6	11	8	8	6	12	9	14	21	15	15	17	21	21	15	33	35	29	27	44	31	44
All	13	14	13	16	13	19	15	27	38	29	33	33	39	46	34	57	61	58	54	74	67	85

	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89
0	18	19	35	32	32	25	24	34	24	24	21	28	23	19	14	11	9	8	9	8	3	2
1	44	37	32	38	40	29	38	25	31	25	24	33	27	29	19	17	26	14	10	8	9	15
All	62	56	67	70	72	54	62	59	55	49	45	61	50	48	33	28	35	22	19	16	12	17

	90	91	92	93	94	95	98	102	All
0	2	1	3	0	1	0	0	0	898
1	6	6	5	6	0	3	1	1	1107
All	8	7	8	6	1	3	1	1	2005

Cell Contents
Count

Here's a representation of the relationship between *LAR* and *Age* in scatter plot:



According to the scatter plot, there is some tendency for patients with younger age to choose *APR* where patients with older age will more likely to choose *SSP* as a method of choice for the surgery. However, the scatter plot is unclear since the interval of age between ~25 and ~90 seems to be indifferent between *APR* or *SSP*. It is also difficult to describe the functional relationship between *LAR* and *age* since the variability is in surgery *LAR* at all ages is large. One way of removing variability while still maintaining the structure of the relationship between the response and explanatory variable is to create intervals for the explanatory variable and compute the mean of the response variable within each group interval.

We now set up a functional form to describe the relationship between *LAR* and *age*. The response variable Y here is a Bernoulli random variable with conditional mean:

$$E(Y \mid x) = 1[P(Y = 1)] + 0[P(Y = 0)] = P(Y = 1) = \pi(x)$$

We denote $P(Y = 1) = \pi(x)$ to reflect its dependence on the value of the explanatory variable.

In simple linear regression, we model the conditional mean of Y by $E(Y \mid x) = \beta_0 + \beta_1 x$. So, from here, we can say that one possibility for the function form to describe the relationship between LAR and age is:

$$E(Y \mid x) = \pi(x) = \beta_0 + \beta_1 x$$

With the binary response given, this model is called the ***Linear Probability Model (LPM)***

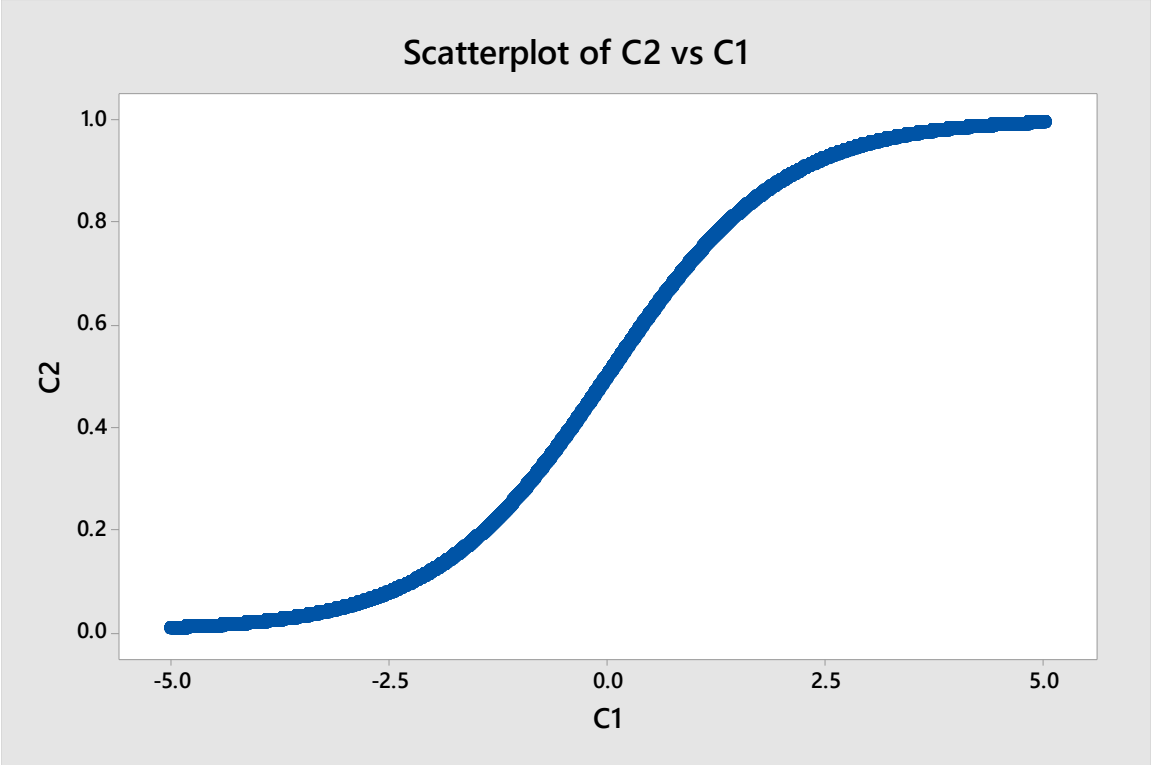
There are some problems with using the *LPM*. First being that the probabilities must fall between 0 and 1, whereas linear functions take values over the entire real line. Secondly, realistically we would usually expect a non-linear relationship between $\pi(x)$ and x. One other thing we should note is that a fixed change in $\pi(x)$ when $\pi(x)$ is near 0 and 1 is way less significant than when a fixed change occurs when $\pi(x)$ is in middle of its range. Take this example for instance, suppose that a family decides on buying a new car for their household and they are deciding on whether they should choose to buy a new or a used car. If we let $\pi(x)$ denote the probability of the family buying a new car and let annual family be x. We would expect an increase of $20,000 in income to have less effect when x = $10,000,000 (i.e. when $\pi(x)$ is near 1) than when x = $40,000. Essentially, when the family's income is high, the decision of considering a used car to save money is almost to non-existence.

We observe that the plot of proportion of LAR cases versus age seems to possess such feature. In turn, the plot of this relationship shows an S-shaped curved.

To model the relationship and the curve, we consider the logistic distribution of:

$$f(z) = \frac{1}{1+e^{z}} = \frac{e^{z}}{1+e^{z}} \text{ , where } -\infty \leq z \leq \infty$$

And the plot of this function can be shown as:



In turn, we propose the following functional form to describe the relationship between LAR and age:

$$E(Y \mid x) = \pi(x) = \frac{1}{1+e^{-(\beta_0 + \beta_{1x})}} = \frac{e^{\beta_0 + \beta_{1x}}}{1+e^{\beta_0 + \beta_{1x}}}$$

This form is known as the ***simple logistic regression model***, it can also be written as:

$$\log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \text{logit}[\pi(x)] = \beta_0 + \beta_1 x$$

The left-hand of the expression above is in the form of a **(*) *log odds***, which is defined as a ***logit***.

(*) Log odds are an alternate way of expressing probabilities, which simplifies the process of updating them with new evidence. The log odds are the log of the odds ratio. For example, the log odds of A are = $\left[\frac{P(A)}{1-P(A)}\right]$

Deriving the Maximum Likelihood Estimators

Since the data includes multiple explanatory variables, we wish to analyze the relationship between a categorical response that takes on two possible outcome and a set of *p* explanatory variables (where $p \geq 2$). To do this, suppose we take a sample of size *n*. For i-th individual, we let $\mathbf{x}_i$ be a vector of all explanatory variables required; in this case it would be *p* (i.e. for i-th individual, x_{i1} can be the *age* and x_{i2} can be the gender for example and so on) and augmented by constant one.

$$\mathbf{X}_i' = (1, x_{i1}, x_{i2}, x_{i3}, \ldots, x_{ip})$$

Like the single explanatory variable, the response variable Y_i in this case is the result of surgery type for individual i, where $Y_i = 1$ for i *SSP* and $Y_i = 0$ for *APR*. In so $\pi_i = \pi(\mathbf{x}_i) = P(Y_i = 1)$ and we know that:

$$\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_p)$$

We can now describe $\pi_i = \pi(\mathbf{x}_i) = P(Y_i = 1)$ with a multiple logistic regression model:

$$\pi_i = \pi(\mathbf{x}_i) = \frac{e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}}}$$

$$= \frac{e^{x_i' \beta}}{1 + e^{x_i' \beta}}$$

And equivalently equal to:

$$\log\{ \pi(\mathbf{x}_i)/ [1 - \pi(\mathbf{x}_i)]\} = \text{logit}[\pi(\mathbf{x}_i)] = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} = \mathbf{X}_i' \boldsymbol{\beta}$$

Now, we derive maximum likelihood estimates for the multiple logistic regression model $(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)$. For the i-th individual in the LAR data, let $Y_i = 1$ if the method of choice for individual i is *SSP* and let $Y_i = 0$ if the method of choice is *APR*. Therefore, the distribution of the data is:

$$\prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1 - y_i}, \qquad \text{where } \pi_i = \pi(\mathbf{x}_i) = \frac{e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}}} = \frac{e^{x_i' \beta}}{1 + e^{x_i' \beta}}$$

Since the log-likelihood of the data is

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \log(\pi_i) + \sum_{i=1}^{n} (1 - y_i) \log(1 - \pi_i)$$

And the maximum likelihood equations after take derivative of $L(\boldsymbol{\beta})$ and set it equal to zero are as follows

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_0} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \pi_i = 0 \quad \text{and} \quad \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij} y_i - \sum_{i=1}^{n} x_{ij} \pi_i = 0 \quad \text{for } j = 1, \ldots, p$$

With these equations combined, they can be solved to obtain an estimate for $\boldsymbol{\beta}$ by using the Newton-Raphson algorithm. Let

$$\boldsymbol{q}' = \left(\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_0}, \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_1}, \ldots, \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_p}\right)$$

$$= \left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \pi_i \,, \sum_{i=1}^{n} x_{i1}\, y_i - \sum_{i=1}^{n} x_{i1}\pi_i, \ldots, \sum_{i=1}^{n} x_{ip}\, y_i - \sum_{i=1}^{n} x_{ip}\pi_i\right.$$

$$\mathbf{H} = \begin{matrix} \frac{\partial^2 L(\boldsymbol{\beta})}{\partial {\beta_0}^2} & \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_0 \beta_1} & \cdots & \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_0 \beta_p} \\ \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_0 \beta_1} & \frac{\partial^2 L(\boldsymbol{\beta})}{\partial {\beta_1}^2} & \cdots & \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_1 \beta_p} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_0 \beta_p} & \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_1 \beta_p} & \cdots & \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_p^2} \end{matrix}$$

And so H becomes

$$\mathbf{H} = \begin{matrix} -\sum_{i=1}^{n} \pi_i(1-\pi_i) & -\sum_{i=1}^{n} x_{i1}\pi_i(1-\pi_i) & \cdots & -\sum_{i=1}^{n} x_{ip}\pi_i(1-\pi_i) \\ -\sum_{i=1}^{n} x_{i1}\pi_i(1-\pi_i) & -\sum_{i=1}^{n} x_{i1}^2\pi_i(1-\pi_i) & \cdots & -\sum_{i=1}^{n} x_{i1}x_{ip}\pi_i(1-\pi_i) \\ \vdots & \vdots & \vdots & \vdots \\ -\sum_{i=1}^{n} x_{ip}\pi_i(1-\pi_i) & -\sum_{i=1}^{n} x_{i1}x_{ip}\pi_i(1-\pi_i) & \cdots & -\sum_{i=1}^{n} x_{ip}^2\pi_i(1-\pi_i) \end{matrix}$$

Note: **H** is a pxp matrix

To estimate$\boldsymbol{\beta}$, we start with an initial guess, say $\boldsymbol{\beta}^{(0)}$and perform an iterative procedure. At the t-th step of this iterative process we obtain $\boldsymbol{\beta}^{(t+1)}$using

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - (\mathbf{H}^{(t)})^{-1}\mathbf{q}^{(t)}$$

Where $\mathbf{q}^{(t)}$ and $\mathbf{H}^{(t)}$ are equivalent to $\mathbf{q}$ and $\mathbf{H}$ evaluated at $\boldsymbol{\beta}^{(t)}$, the *t*-th guess for$\boldsymbol{\beta}$. The algorithm continues until successive estimates of $\boldsymbol{\beta}$ converge.

If the algorithm converges at iteration *T*, then $\boldsymbol{\beta} = \boldsymbol{\beta}^{(T)}$ is the maximum likelihood estimate for $\boldsymbol{\beta}$, and we can use it to determine estimates for $\pi_i = \pi(x_i)$ according to

$$\pi_i = \pi(\mathbf{x}_i) = \frac{e^{\widehat{\beta_0} + \widehat{\beta_1} x_{i1} + \widehat{\beta_2} x_{i2} + \cdots + \widehat{\beta_p} x_{ip}}}{1 + e^{\widehat{\beta_0} + \widehat{\beta_1} x_{i1} + \widehat{\beta_2} x_{i2} + \cdots + \widehat{\beta_p} x_{ip}}} = \frac{e^{x_i' \hat{\beta}}}{1 + e^{x_i' \hat{\beta}}}$$

In theory**, Logistic regression** is the appropriate regression analysis to conduct when the dependent variable is binary. Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. Think of it this way, combining all explanatory variables together into a logistic model that produces a value between 0 and 1. But if we set a cut off as 0.5, where any output that is greater than 0.5 will be set to 1 and any output that is smaller than 0.5 will be set to 0. This helps the form of logistic regression to stay consistent (i.e. yes vs. no).

Logistic Regression Applications

Deriving Model using Forward Selection

Now that we have got the theory of logistic regression out of the way, it is only fitting show how LR works in application. Using Minitab, we can fit a set of data using logistic model by 2 methods which are *forward selection* and *backward selection* where we would pick and add or eliminate one explanatory variable by one until we obtain the best model possible to predict *LA*. The way to decide whether we should add a particular variable to the model is to measure its impact individually and in the model in general via ***p-value*** and ***chi-square value***.

Firstly, we would measure the effect of every single explanatory variable in the data against *LAR* using the linear logistic model. We will choose the best variable by comparing the *p-value* and *chi-square value*.

DEVIANCE TABLE	**X^2-value**	**P-value**
Gender	*8.01*	*0.005*
Hospital Types	*11.34*	*0.079*
Surgery Year	*5.08*	*0.279*
Stage	*127.24*	*0.000*
Caseload	*17.84*	*0.000*
Age	*9.48*	*0.002*

(*Table 1.1)*

Initially, we would consider the p-value to see which variable produces the lowest p-value. In this case, we have **Stage** and **Caseload** both have the p-value of 0.000. Now to decide between

the two, we consider which variable has the higher Chi square-value where we clearly see **Stage** has the significant advantage. Hence, we choose **Stage** as the first explanatory variable for the model.

Add "**Stage**" to the model → **Degree of freedom (D.F)** = 3, $\mathbf{X^2}$ = 127.24

(i.e. X_{i1} = **Stage**)

Note: We will keep adding variables into the model until p-value of that particular variable exceeds p-value = 0.10 within the model.

Now, we will add a second explanatory variable.

DEVIANCE TABLE	**Δ D.F**	**Δ X^2**	**p-value**	**X^2**	**D.F**
Gender	*1*	*5.71*	*0.017*	*132.95*	*4*
Hospital Types	*6*	*7.95*	*0.2418*	*135.19*	*9*
Caseload	*1*	*17.79*	*0.000*	*145.03*	*4*
Age	*1*	*2.87*	*0.090*	*130.11*	*4*
Surgery Year	*4*	*4.68*	*0.3218*	*131.92*	*7*

(*Table 1.2)*

Caseload is the only variable here with p-value = 0.000 and is the lowest out of all, so we will add "**Caseload**" to the model.

→ **Degree of freedom (D.F)** = 4, $\mathbf{X^2}$ = 145.03

(i.e. X_{i1} = **Stage**, X_{i2} = **Caseload**)

DEVIANCE TABLE	**Δ D.F**	**Δ X^2**	**p-value**	**X^2**	**D.F**
Gender	*1*	*6.27*	*0.012*	*151.30*	*5*
Hospital Types	*6*	*9.94*	*0.1272*	*154.97*	*10*
Age	*1*	*3.51*	*0.061*	*148.54*	*5*
Surgery Year	*4*	*4.24*	*0.3745*	*149.27*	*8*

(*Table 1.3)*

Since **Gender** is the variable with the lowest p-value out of the bunch, we will add "**Gender**" to the Model.

→ **Degree of freedom (D.F)** = 5, **X^2** = 151.30

(i.e. X_{i1} = **Stage**, X_{i2} = **Caseload**, X_{i3} = **Gender**)

Let's continue and see if we could add more variable to the model. Note that adding more explanatory variables to the model whenever possible help the model become for sufficient and accurate at predicting the value of *LAR* (i.e. closest to value of 0 or 1).

DEVIANCE TABLE	**Δ D.F**	**Δ X^2**	**p-value**	**X^2**	**D.F**
Hospital Types	*6*	*10.11*	*0.1201*	*161.40*	*11*

Age	*1*	*3.18*	*0.074*	*154.48*	*6*
Surgery Year	*4*	*4.22*	*0.3771*	*155.52*	*9*

(*Table 1.4)*

As we can see, the p-value of adding **Age** to the model is still below p-value = 0.10 in which we will add "**Age**" to the model.

→ **Degree of freedom (D.F)** = 6, $\mathbf{X^2}$ = 154.48

(i.e. X_{i1} = **Stage**, X_{i2} = **Caseload**, X_{i3} = **Gender**, X_{i4} = **Age**)

Lastly, we will see if the current model can add the last 2 remaining variables.

DEVIANCE TABLE	**Δ D.F**	**Δ X^2**	**p-value**	**X^2**	**D.F**
Hospital Types	*6*	*9.15*	*0.1654*	*163.63*	*12*
Surgery Year	*4*	*4.23*	*0.3758*	*158.23*	*10*

(*Table 1.5)*

At this point, it is safe to say that the first model is completed with all the appropriate explanatory variables. Also note that in logistic regression, the order in which we select those variables does not matter.

The logistic regression for patient i in this model #1 is presented as:

$\pi_i = \pi(\mathbf{x}_i) = \frac{e^{Y_i}}{1+e^{Y_i}}$, where $Y_i = \beta_{10} + \beta_{11}X_{i1} + \beta_{12}X_{i2} + \beta_{13}X_{i3} + \beta_{14}X_{i4}$ & i = 1, 2, ..., 2005

$$= \beta_{10} + \beta_{11}\textbf{Stage} + \beta_{12}\textbf{Caseload} + \beta_{13}\textbf{Gender} + \beta_{14}\textbf{Age}$$

As you can see, **Hospital Type** and **Surgery Year** are two explanatory variables that have ***p-value*** significantly higher compare to the others. Note that a higher p-value implies less confident in the existence of relationship between the explanatory and response variables, or "more independent". Additionally, including the high p-value variable will have little to no contributions and benefits but instead creates more problems within the model overall. In contrast, low p-value guarantees or increases the chance of rejecting the null hypothesis which is the case where there's no relationship between the response variable and explanatory variable(s). Additionally, ***chi-square value*** is a special case of logistic regression use to analyze the model. With chi-square contingency analysis, the independent variable is dichotomous, and the dependent variable is dichotomous. The *chi-square value* comes from *Maximum Likelihood function* and it is based on the ability to predict y values with and without x. Furthermore, the *chi-square* test is intended to test how likely it is that an observed distribution is due to chance. It is also called a "goodness of fit" statistic, because it measures how well the observed distribution of data fits with the distribution that is expected if the variables are independent. To relate back to our data, throughout the process of forward selection, I have noticed that within the "Hospital Type" data set that includes data of 6 other hospitals (i.e. *HOSTYPE2, HOSTYPE3, HOSTYPE4, HOSTYPE5, HOSTYPE6,* and *HOSTYPE7*). One particular variable stood which is *HOSTYPE5,* where the p-value presented is significantly lower than other 5 hospitals within the dataset. This could be the case due to the fact that hospital #5 could be the biggest hospital that provides both methods of treatment, where all other hospitals only have one of them. The results of p-values can be seen below.

Deviance Table

Source	DF	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	12	163.63	13.6357	163.63	0.000
H2	1	0.17	0.1747	0.17	0.676
H3	1	1.26	1.2610	1.26	0.261
H4	1	0.59	0.5936	0.59	0.441
H5	1	5.88	5.8799	5.88	0.015
H6	1	1.47	1.4684	1.47	0.226
H7	1	0.56	0.5608	0.56	0.454
STG2	1	60.80	60.8018	60.80	0.000
STG3	1	21.84	21.8367	21.84	0.000
STG4	1	85.16	85.1626	85.16	0.000
Avgcsld	1	20.93	20.9333	20.93	0.000
Age	1	2.21	2.2139	2.21	0.137
Gender	1	6.16	6.1637	6.16	0.013
Error	1992	2594.07	1.3022		
Total	2004	2757.69			

Conveniently, within the **Surgery Year** data set, there also exists useful variable for the model that is *Y5* in which also contains a noticeable difference from other years' data in term of p-values. One may think if a patient is at the latest year of having the cancer, it would be easier for the model to predict which method to use that is suitable. A prominent factor is a tumor's distance to the anal sphincters. Because a wide distal margin has formerly been considered to be of particular importance, tumors less than 5 cm from the anal verge could not be operated on except by APR. More specifically, by logic, as the number year of the tumor increases, the size and length of the tumor will increase in response. So in application of the data, it is safe to assume that the response variable LAR will be predicted toward SSP if we choose to only consider the only *Y5* from the data in the model.

Deviance Table

Source	DF	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	10	158.23	15.8232	158.23	0.000
Y2	1	0.05	0.0462	0.05	0.830
Y3	1	0.01	0.0090	0.01	0.924
Y4	1	0.01	0.0102	0.01	0.920
Y5	1	2.41	2.4059	2.41	0.121
STG2	1	59.43	59.4338	59.43	0.000
STG3	1	22.38	22.3780	22.38	0.000
STG4	1	85.94	85.9379	85.94	0.000
Avgcsld	1	18.44	18.4428	18.44	0.000
Age	1	2.71	2.7117	2.71	0.100
Gender	1	5.96	5.9607	5.96	0.015
Error	1994	2599.46	1.3036		
Total	2004	2757.69			

Now let's test this hypothesis of only including *HOSTYPE5* and *Y5* in the model applying the same concept of forward selection like above.

DEVIANCE TABLE	**X^2-value**	**p-value**
Gender	*8.01*	*0.005*
Hospital #5	*3.33*	*0.068*
Year #5	*4.95*	*0.026*
Stage	*127.24*	*0.000*
Caseload	*17.84*	*0.000*
	9.48	*0.002*

(*Table 1.6)*

Like the last model, we will start by adding **"Stage"** as the initial explanatory variable.

Add "**Stage**" to the model → **Degree of freedom (D.F)** = 3, $\mathbf{X^2}$ = 127.24

(i.e. X_{i1} = **Stage**)

DEVIANCE TABLE	**Δ D.F**	**Δ X^2**	**p-value**	**X^2-value**	**D.F**
Gender	*1*	*5.71*	*0.017*	*132.95*	*4*
Hospital #5	*1*	*2.87*	*0.090*	*130.11*	*4*
Year #5	*1*	*4.61*	*0.032*	*131.85*	*4*
Caseload	*1*	*17.79*	*0.000*	*145.03*	*4*
Age	*1*	*2.87*	*0.090*	*130.11*	*4*

(*Table 1.7)*

Add "**Caseload**" to the model → **Degree of freedom (D.F)** = 4, $\mathbf{X^2}$ = 145.03

(i.e. X_{i1} = **Stage**, X_{i2} = **Caseload**)

DEVIANCE TABLE	**Δ D.F**	**Δ X^2**	**p-value**	**X^2-value**	**D.F**
Gender	*1*	*6.27*	*0.012*	*151.30*	*5*
Hospital #5	*1*	*7.53*	*0.006*	*152.56*	*5*
Year #5	*1*	*4.08*	*0.043*	*149.11*	*5*
Age	*1*	*3.51*	*0.061*	*148.54*	*5*

(*Table 1.8)*

Add "**Hospital #5**" to the model → **Degree of freedom (D.F)** = 5, $\mathbf{X^2}$ = 152.56

(i.e. X_{i1} = **Stage**, X_{i2} = **Caseload**, X_{i3} = **Hospital #5**)

DEVIANCE TABLE	**Δ D.F**	**Δ X^2**	**p-value**	**X^2-value**	**D.F**
Gender	*1*	*5.99*	*0.014*	*158.56*	*6*
Year #5	*1*	*3.85*	*0.050*	*156.42*	*6*
Age	*1*	*3.02*	*0.082*	*155.59*	*6*

(*Table 1.9)*

Add "**Gender**" to the model → **Degree of freedom (D.F)** = 6, **X^2** = 158.56

(i.e. X_{i1} = **Stage**, X_{i2} = **Caseload**, X_{i3} = **Hospital #5,** X_{i4} = **Gender**)

DEVIANCE TABLE	**Δ D.F**	**Δ X^2**	**p-value**	**X^2-value**	**D.F**
Year #5	*1*	*3.84*	*0.050*	*162.39*	*7*
Age	*1*	*2.74*	*0.098*	*161.29*	*7*

(*Table 1.10)*

Add "**Year #5**" to the model → **Degree of freedom (D.F)** = 7, **X^2** = 162.39

(i.e. X_{i1} = **Stage**, X_{i2} = **Caseload**, X_{i3} = **Hospital #5,** X_{i4} = **Gender,** X_{i5} = **Year #5**)

DEVIANCE TABLE	**Δ D.F**	**Δ X^2**	**p-value**	**X^2-value**	**D.F**
Age	*1*	*2.36*	*0.125*	*164.75*	*8*

(*Table 1.11)*

We will leave out **"Age"** variable as the p-value exceeded 0.10.

The logistic regression for patient i in this model #2 is represented as:

$\pi_i = \pi(\mathbf{x}_i) = \frac{e^{Y_i}}{1+e^{Y_i}}$, where $Y_i = \beta_{20} + \beta_{21}X_{i1} + \beta_{22}X_{i2} + \beta_{23}X_{i3} + \beta_{24}X_{i4} + \beta_{25}X_{i5}$ & i = 1, 2, ..., 2005

$= \beta_{(2)0} + \beta_{(2)1}$**Stage**$+ \beta_{(2)2}$**Caseload** $+ \beta_{(2)3}$**Hospital #5** $+ \beta_{(2)4}$**Gender** $+ \beta_{(2)5}$**Year #5**

Recall that Y_i from model #1 is:

$$Y_i = \beta_{(1)0} + \beta_{(1)1}\textbf{Stage} + \beta_{(1)2}\textbf{Caseload} + \beta_{(1)3}\textbf{Gender} + \beta_{(1)4}\textbf{Age}$$

Note: The first down subscript of the beta coefficients indicates the model number.

Odds Ratio

Comparing explanatory variables

Notice that beside the fact that both models included "Stage" and "Caseload" as the first 2 explanatory variables, the difference is clear in the number of explanatory variables in the models, where model #1 has 4 and model #2 has 5. Additionally, we also want to pay attention to the fact that model #2 includes "**Hospital #**5" and "**Year #**5" but model #1 doesn't even take any data regarding the "**Year**" and the "**Hospital**". We also know that a good *LRM* does not possess any unnecessary data because they may contribute into inaccurate predictions. With this information, we can assume that "**Hospital #2**, **Hospital #3, Hospital #4, Hospital #6, Hospital #7**" and "**Year #2, Year #3, Year #4**" are "unnecessary data" that explains why model #2 has explanatory variables in that way. To understand more about how to pick the necessary data for a model, we introduce the concept of *Odds Ratio (OR)* to compare explanatory variables in another

perspective. An *OR* is a measure of association between an exposure and an outcome. The *OR* represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the opposite of that exposure. When a logistic regression is calculated, the regression coefficient (i.e. let say β_{10}) is the estimated increase in the log odds of the *outcome per unit increase* in the value of the *exposure.* In other words, the exponential function of the regression coefficient ($e^{\beta 10}$) is the odds ratio associated with a one-unit increase in the exposure. In theory, *ORs* that are greater than 1 indicate that the event is more likely to occur as the predictor increases. In contrast, those that are less than 1 indicate that the event is less likely to occur as the predictor increases. Unlike percentages or probabilities, the exposure is not calculated compare to the result overall, instead it is one step more complex than that, where we take the probability of certain event and compare it to another event. For instance:

Let P(A) be the probability of some event A and P(B) be the probability of some even B.

Suppose that,

P(A) = 0.4 and P(B)= 0.1 in which implies that $P(A^c) = 0.6$ and $P(B^c) = 0.9$.

So the odds of A are:

0.6 / 0.4 = 1.5

& the odds of B are:

0.1 / 0.9 = 0.125

In which conclude that the odds of "some outcome" at A is 12 times (= 1.5 / 0.125) greater than

the odds of "some outcome" at B.

In application with data we are working on, here is the odds ratios table of all the hospitals:

ODDS RATIO for CONTINUOUS PREDICTORS (Individually)	**Odds Ratio**	**90% Confidence Interval**
Hospital #2	*0.6868*	*(0.4530, 1.0413)*
Hospital #3	*1.0793*	*(0.8771, 1.3281)*
Hospital #4	*1.1144*	*(0.9613, 1.2920)*
Hospital #5	*1.3877*	*(1.0298, 1.8700)*
Hospital #6	*0.8398*	*(0.6904, 1.0215)*
Hospital #7	*0.7023*	*(0.4558, 1.0820)*

(Table 2.1)

The odds ratios produced in the table come from running *LRM* with just one explanatory variable at once. Here, we see that the odds ratio of "**Hospital #5**" is the highest out of the list. In translation, the odds ratio basically implies that the odds of *SSP* (Y = 1) occurring at hospital #5 is approximately "1.3877" times greater than the odds of *SSP* at any other hospital. Additionally, we can conclude with 90% confidence that the odds of *LAR* = 1 at hospital #5 are anywhere from 1.03 to 1.87 times greater than the odds of *LAR* = 1 at any other hospital type. From this observation, it is also safe to assume that the odds of *APR* occurring at hospital #2 will be highest out the list since it has the lowest odds ratio in the table above.

ODDS RATIO for CONTINUOUS PREDICTORS (Combined)	**Odds Ratio**	**90% Confidence Interval**
Hospital #2	*0.9328*	*(0.5297, 1.6426)*
Hospital #3	*1.4314*	*(0.9264, 2.2118)*
Hospital #4	*1.4145*	*(0.9441, 2.2191)*
Hospital #5	*1.8207*	*(1.1203, 2.9590)*
Hospital #6	*1.1608*	*(0.7555, 1.7835)*
Hospital #7	*0.9522*	*(0.5343, 1.6970)*

(Table 2.2)

Now, this is the case when we run a *LRM* with more than one variable, in fact with all the hospitals together. The above table shows the odds ratios of each of the variable when there are more than one variable in the model. Again, hospital #5 provides the highest odds ratio or chance of achieving the result. But the difference between this table and the last table is that the values provided here are higher in comparison. This could be due to the fact that now the model has more information on all other hospitals it can rule in or out, hence the odds ratios are more accurate and higher.

Similarly, this is the same case for "Year #5" in the data set of years as we included it in the second model.

ODDS RATIO for CONTINUOUS PREDICTORS (Individually)	**Odds Ratio**	**90% Confidence Interval**
Year #2	*0.9136*	*(0.7638, 1.0926)*
Year #3	*0.9746*	*(0.8108, 1.1715)*
Year #4	*0.9476*	*(0.7826, 1.1474)*
Year #5	*1.2987*	*(1.0694, 1.5771)*

(Table 2.3)

And the combined variables in model odds ratios are:

ODDS RATIO for CONTINUOUS PREDICTORS (Combined)	**Odds Ratio**	**90% Confidence Interval**
Year #2	*0.9789*	*(0.7830, 1.2238)*
Year #3	*1.0293*	*(0.8197, 1.2925)*
Year #4	*1.0054*	*(0.7957, 1.2705)*
Year #5	*1.3022*	*(1.0280, 1.6495)*

(Table 2.4)

As we mentioned earlier, to produce the result of using SSP as the procedure of choice where Y = 1 with highest odds possible, we want to use explanatory variables whose odds ratios are

greater than 1 because those variables contribute in higher chance of the event to happen as the predictor increases.

From *Table 2.1*, notice that the odds ratios of hospital #3, 4, 5 are all greater than one in which implies that they are suitable variables to include in the model when we are trying to predict Y = 1. But as we seen earlier, data from hospital #3, 4 contribute in such high p-value within the model that they should not be included; hence, we only include the fifth hospital. From *Table 2.3*, since the odds ratio of year #5 is the only one that exceeded one, we shall only include it in the model. Hence, that is how model #2 was introduced.

We need to see if that picking variables by observing odds ratio works, so show this, let's compare the two models that were introduced earlier. Recall that the two models are:

1. $Y_i = \beta_{(1)0} + \beta_{(1)1}\mathbf{Stage} + \beta_{(1)2}\mathbf{Caseload} + \beta_{(1)3}\mathbf{Gender} + \beta_{(1)4}\mathbf{Age}$
2. $Y_i = \beta_{(2)0} + \beta_{(2)1}\mathbf{Stage} + \beta_{(2)2}\mathbf{Caseload} + \beta_{(2)3}\mathbf{Hospital\ \#5} + \beta_{(2)4}\mathbf{Gender} + \beta_{(2)5}\mathbf{Year\ \#5}$

Receiver Operating Characteristic curve (ROC curve)

One efficient way to determine if a model is good or to compare models is to use the idea of **Receiver Operating Characteristic curve** (ROC curve). ROC curve is a graphical plot that illustrates the diagnostic ability of a binary system that can be applied to the results logistic regression. The ROC curve is created by plotting the *true positive rate* (TPR) against the *false positive rate* (FPR) at various threshold settings (cut off rate that is between 0 and 1). The true-positive rate is also known as "*sensitivity"*, more specifically the "*probability of detection".* The false-positive rate is also known as the "*the probability of false alarm"* and it is denoted as (1 – "*specificity*"). The ROC curve can also be interpreted as a plot of the power as a function of the

Type I Error of the decision rule. The interpretation of formulas of *sensitivity* and (1 – *specificity*) can be seen as the followings:

Sensitivity:

$$\text{TPR} = \frac{TP}{P} = \frac{TP}{TP+FN} = 1 - \text{FNR}$$

Specificity:

$$\text{TNR} = \frac{TN}{N} = \frac{TN}{TN+FP} = 1 - \text{FPR}$$

1 – Specificity:

$$\text{FPR} = \frac{FP}{N} = \frac{FP}{TN+FP} = 1 - \text{TNR}$$

Where,

P = the number of real positive cases in the data (condition positive)

N = the number of real negative cases in the data (condition negative)

TP = true positive

TN = true negative

FP = false positive (Type I error)

FN = false negative (Type II error)

In general, the ROC curve can be generated by plotting the detection probability on the y-axis versus the probability of false-alarm probability on the x-axis (sensitivity vs. 1 - specificity). Additionally, ROC curve analysis provides tools to select the best or optimal models and to

discard the others non-optimal ones independently from the cost context. In general, ROC curve analysis is related in a direct way to cost/benefit of diagnostic decision making.

To put our set of data in application, we already have the predicted values of 2005 samples of each of the two models that were derived earlier. Let "*sensitivity"* equals the proportion of number of samples were predicted right for LAR = SSP (Y = 1) and "1 – specificity" be the proportion of number of samples that were predicted incorrectly for LAR = ARP (Y = 0). We now analyze the *sensitivity* versus *1 – specificity* under a certain "*cut-off probability"* that is between 0 and 1. We will consider 5 *cut-off probabilities,* which are 0.3, 0.4, 0.5, 0.6, and 0.7. Before we list all the results, let's do a demonstration. Suppose that we chose the *cut-off probability* of 0.5; model #1 predicted LAR on the first observation of the sample to be equal to ~0.4521 and model #2 predicted it to be equal to ~0.5111, if the *cut-off probability* is equal to 0.5 then model #1's prediction would be equal to 0, and 1 for model #2. We note that if the *cut-off probability* is low, sensitivity would be higher as there would be a larger margin for probability that is over 0.3 to be declared as *LAR = SSP*. One may ask what the optimal *cut-off probability* and there would be are many reasons to contribute in choosing it. SSP as mentioned before is a more advanced technique that is superior in comparison to APR as it is safer, has more flexibility range of treatments where APR should only be operated under certain circumstances; thus, this factor suggests that we should lean towards lower *cut-off probabilities* that is less than 0.5 (the optimal cut-off without any assumption).

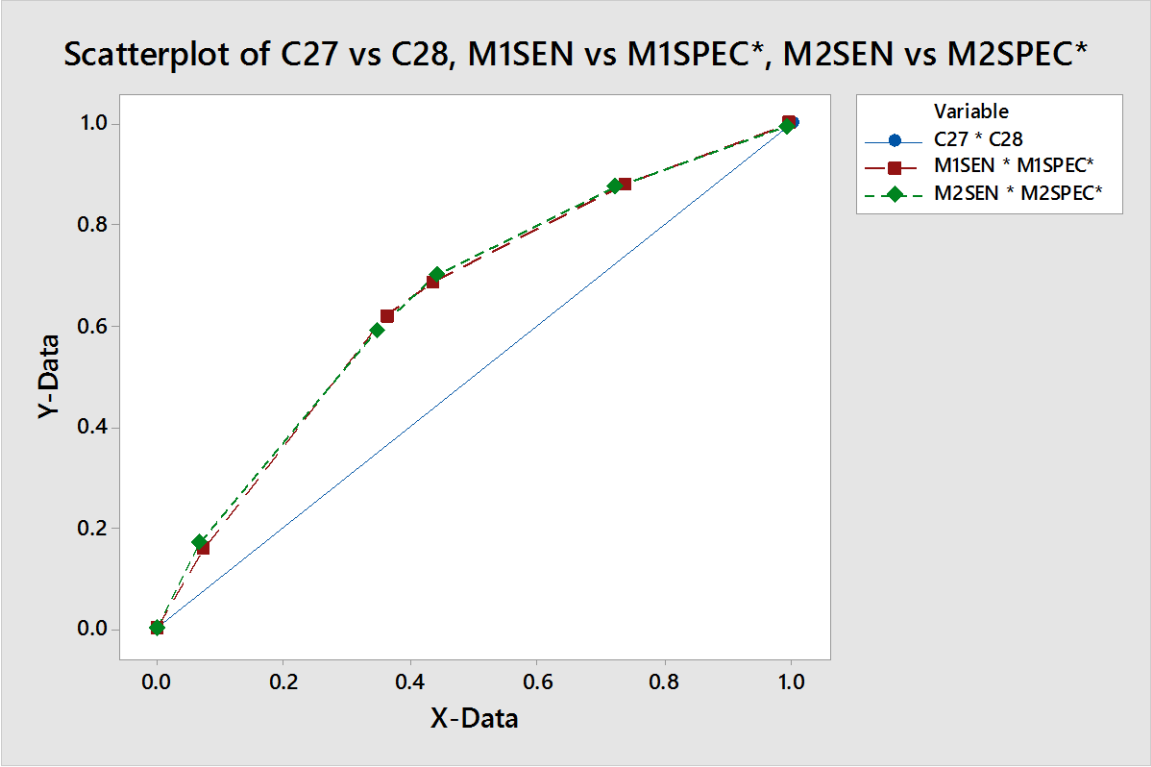
Here's a table of cross tabulation analysis between the two models:

	Model #1	**Model #2**
p (*cut-off probability)* = 0.3	Predicting 0:	Predicting 0:

	- 5/898 correct **-893/898 incorrect (1-specificity)** <u>Predicting 1:</u> **-1106/1107 correct (sensitivity)** -1/1107 incorrect	- 7/898 correct **-891/898 incorrect (1-specificity)** <u>Predicting 1:</u> **-1101/1107 correct (sensitivity)** -6/1107 incorrect
p = 0.4	<u>Predicting 0:</u> - 237/898 correct **-661/898 incorrect (1-specificity)** <u>Predicting 1:</u> **-974/1107 correct (sensitivity)** -133/1107 incorrect	<u>Predicting 0:</u> - 252/898 correct **-646/898 incorrect (1-specificity)** <u>Predicting 1:</u> **-966/1107 correct (sensitivity)** -141/1107 incorrect
p = 0.5	<u>Predicting 0:</u> - 507/898 correct **-391/898 incorrect (1-specificity)** <u>Predicting 1:</u> **-758/1107 correct (sensitivity)** -349/1107 incorrect	<u>Predicting 0:</u> - 503/898 correct **-395/898 incorrect (1-specificity)** <u>Predicting 1:</u> **-773/1107 correct (sensitivity)** -334/1107 incorrect

p = 0.6	Predicting 0: - 574/898 correct **-324/898 incorrect (1-specificity)** Predicting 1: **-684/1107 correct (sensitivity)** -423/1107 incorrect	Predicting 0: - 587/898 correct **-311/898 incorrect (1-specificity)** Predicting 1: **-653/1107 correct (sensitivity)** -454/1107 incorrect
p = 0.7	Predicting 0: - 832/898 correct **-66/898 incorrect (1-specificity)** Predicting 1: **-174/1107 correct (sensitivity)** -933/1107 incorrect	Predicting 0: - 838/898 correct **-60/898 incorrect (1-specificity)** Predicting 1: **-187/1107 correct (sensitivity)** -920/1107 incorrect

To observe, both models possessed very similar results in predictions for all 5 *cut-off probabilities.* This is the case due to the fact that both models include data of "Stage of cancer" and "Case closed" and as we seen earlier, these two sets of data contributed the most in the model in term of chi-square values and both possessed p-value of 0.0000. So the difference in the model didn't not make too much of a noticeable impact as we seen in the table. To illustrate more let's put these results into a ROC curve for further analysis.



The blue line in the middle is the divider that determines whether the model is good or not. Any model that is above the line to the left like the models we observed is considered "good". Essentially, further the distant from the divider, "better" the model is. As we can see, the two models in the scatter plot are almost identical and possessed the same amount of efficiency in predicting the cancer procedure.

Conclusion

We have only reached the surface of how useful the model of logistic regression can be in term of putting data in applications and analyzing them in various sub-applications that can be used like the ROC curve. As mentioned before, in almost all cases in real life that requires the need of regression models, they are problems that have many positive or negative characteristics (explanatory variables), that can be complicated to analyze overall. LRM helps to narrow those characteristics down to binary form and help to create the most efficient model possible under its application. I personally have learned a lot and I think this project was well time invested. This is going to be something that I will use in the future somehow whether in my job or to teach someone else. I want to thank Dr. Farrell for giving me an opportunity to take on this topic with him and also his time overall throughout the semester.

Citations:

1. Farrell, Patrick. “STAT 5602: Logistic Regression 1”
2. Farrell, Patrick. “STAT 5602: Logistic Regression 2”
3. Farrell, Patrick. “STAT 5602: Logistic Regression 3”
4. Narkhede, Sarang. “Understanding ROC Curve”. (2018). https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5
5. Szumilas M. Explaining odds ratios [published correction appears in J Can Acad Child Adolesc Psychiatry. 2015 Winter;24(1):58]. *J Can Acad Child Adolesc Psychiatry*. 2010;19(3):227–229.
6. Wikipedia. “Receiver operating characteristic”. https://en.wikipedia.org/wiki/Receiver_operating_characteristic
7. Bordeianou L, Maguire LH, Alavi K, Sudan R, Wise PE, Kaiser AM. Sphincter-sparing surgery in patients with low-lying rectal cancer: techniques, oncologic outcomes, and functional results. *J Gastrointest Surg*. 2014;18(7):1358–1372. doi:10.1007/s11605-014-2528-y