SCIENCE ACADEMY
Excellence in Teaching & Learning

Logistic Regression

Hi guys, Its time to learn **Logistic Regression**.

In this section, we will discuss:

- Classification problem and the use of Logistic Regression to find the answer to our problem.
- How to interpret the results of the Logistic Regression through the confusion matrix.

✅ *Optional Readings and References:*
Introduction to Statistical Learning - Chapter 4
Machine Learning - A Probabilistic Perspective - section 1.4.6 & Chapter 8

Dr. Junaid S. Qazi PhD

General message: *Key concepts along with significant commentary / text is provided in the slides, so that they serve as a reference for the respective theory lecture. However, the suggested readings are recommended to explore more on the topic under discussion! Good luck!*

SCIENCE ACADEMY
Excellence in Teaching & Learning

Logistic Regression

In classification problem, rather than predicting a *continuous or quantitive* output value (e.g today's stock price) we are interested in non-numerical value, a ***categorical or qualitative*** output (e.g. will the stock index increase of decrease). In this section, our focus will be to learn **Logistic Regression** as a method for **Classification**. Today, Logistic Regression model is one of the most widely used binary models in the analysis of categorical data.

Optional Reading: The Regression Analysis of Binary Sequences by David R. Cox in Journal of the Royal Statistical Society. Vol. 20, No. 2 (1958), pp 215-242.

SCIENCE ACADEMY
Excellence in Teaching & Learning

Logistic Regression

In classification problem, rather than predicting a *continuous or quantitive* output value (e.g today's stock price) we are interested in non-numerical value, a ***categorical or qualitative*** output (e.g. will the stock index increase of decrease). In this section, our focus will be to learn **Logistic Regression** as a method for **Classification**. Today, Logistic Regression model is one of the most widely used binary models in the analysis of categorical data.

Common Examples of Binary Classification:

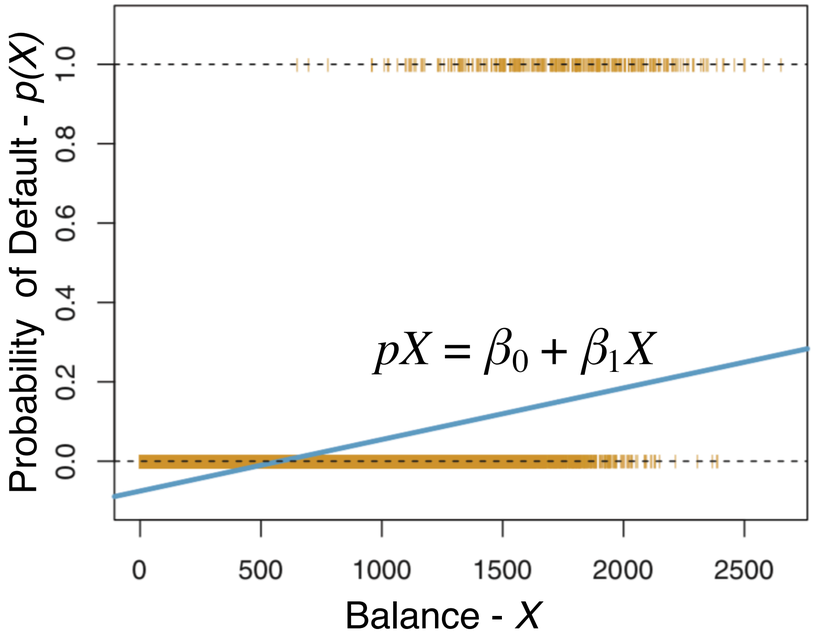
- win or loss
- pass or fail
- dead or alive
- spam or ham email
- Insurance or loan defaults (yes / no)
- healthy or sick (Yes / no)

The convention for binary classification is to have two classes 0 and 1

Optional Reading: The Regression Analysis of Binary Sequences by David R. Cox in Journal of the Royal Statistical Society. Vol. 20, No. 2 (1958), pp 215-242.

Linear regression is not appropriate in the case of a qualitative (classification problem) response. Why not?

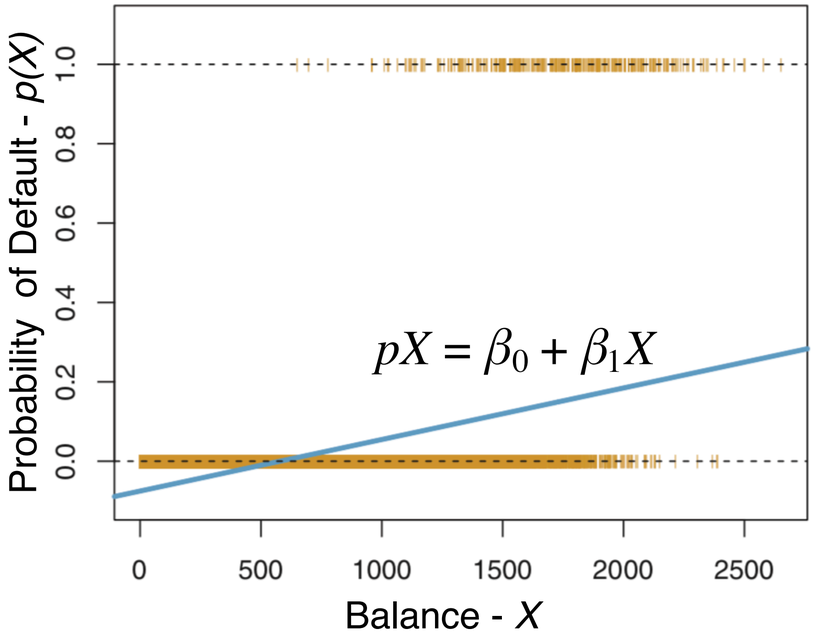
Let’s learn with a very common example!



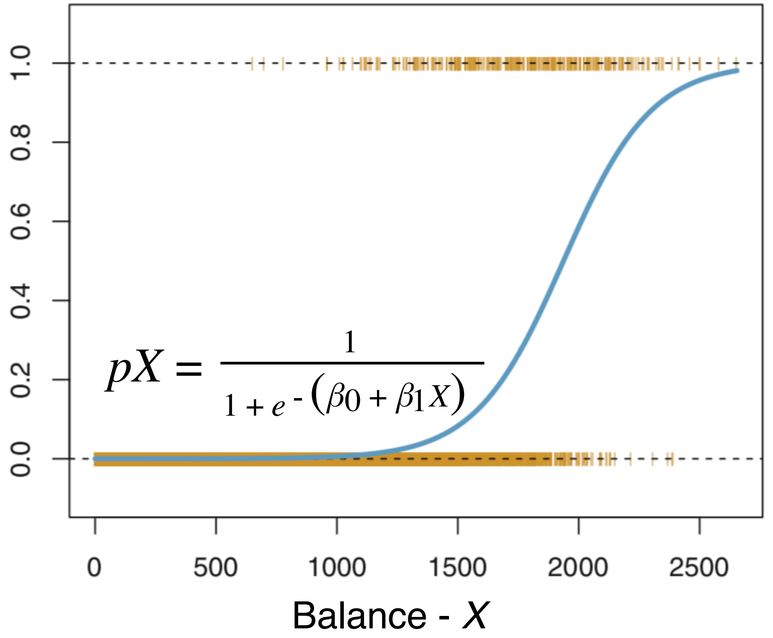
Linear Regression Model:
Estimated probabilities using linear regression (in blue). The orange ticks indicate the 0/1 values for No/Yes.

Shown is figure 4.2 (with some modification) from Introduction to Statistical Learning at page 131. (Fair Use policy)

SCIENCE ACADEMY
Excellence in Teaching & Learning



Linear Regression Model:
Estimated probabilities using linear regression (in blue). The orange ticks indicate the 0/1 values for No/Yes.

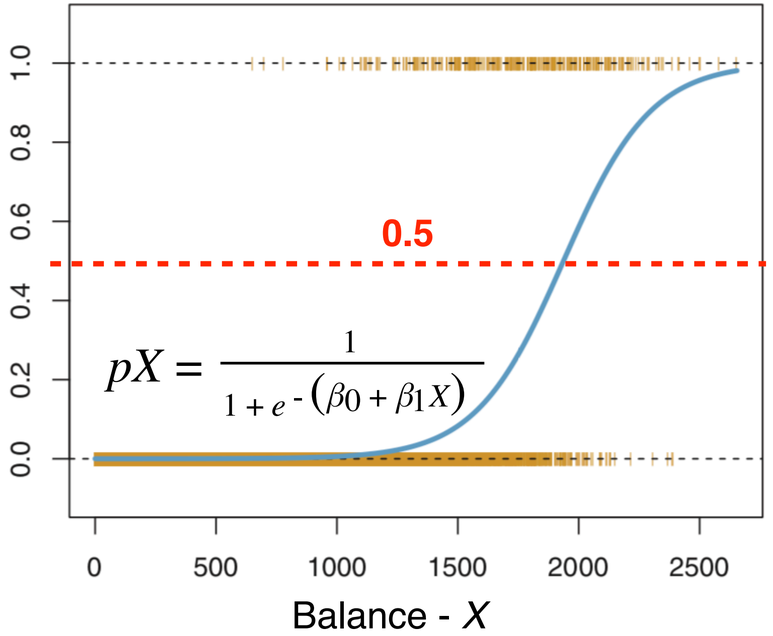


Logistic Regression Model:
Predicted probabilities using logistic regression (in blue). The orange ticks indicate the 0/1 values for No/Yes.
All probabilities lie between 0 and 1

Shown is figure 4.2 (with some modification) from Introduction to Statistical Learning at page 131. (Fair Use policy)

From the Logistic Regression, we have probabilities between 0 and 1 for our data, we set a cutoff point at 0.5:

- Anything below it results in class 0/No
- Anything above is class 1/Yes

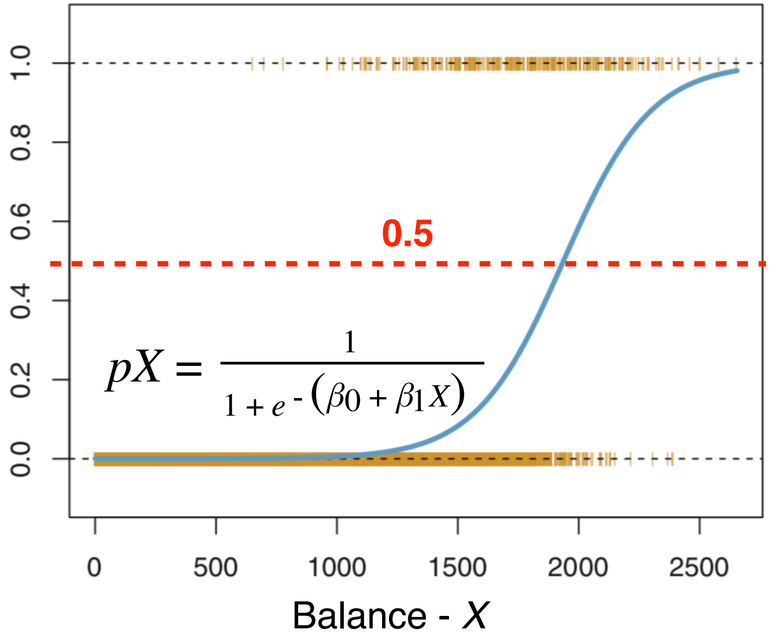


Shown is figure 4.2 (with some modification) from Introduction to Statistical Learning at page 131. (Fair Use policy)

SCIENCE ACADEMY
Excellence in Teaching & Learning

From the Logistic Regression, we have probabilities between 0 and 1 for our data, we set a cutoff point at 0.5:

- Anything below it results in class 0/No
- Anything above is class 1/Yes



Once, we train our Logistic Regression model on the train dataset, we deploy it on the test dataset. Its important to evaluate the model's performance to see how well the predictions are!
We can use Confusion Matrix to evaluate Classification models.

Shown is figure 4.2 (with some modification) from Introduction to Statistical Learning at page 131. (Fair Use policy)

SCIENCE ACADEMY
Excellence in Teaching & Learning

Logistic Regression

(Model Evaluation)

Confusion matrix:

(also known as error matrix) is often used to describe the performance of a classification model on a set of test data for which the true values are known.

SCIENCE ACADEMY
Excellence in Teaching & Learning

Logistic Regression

(Model Evaluation)

Confusion matrix:

(also known as error matrix) is often used to describe the performance of a classification model on a set of test data for which the true values are known.

Let's try to learn while testing number of persons for some disease!

n = 100	Predicted No	Predicted Yes
Actual No = 43	40	3
Actual Yes = 57	7	50
	47	53

SCIENCE ACADEMY
Excellence in Teaching & Learning

Logistic Regression

(Model Evaluation)

Confusion matrix:

n = 100	Predicted No	Predicted Yes
Actual No = 43	TN = 40	FP = 3
Actual Yes = 57	FN = 7	TP = 50
	47	53

Let's define the basic terminology:

- **True Negatives (TN):** Our model predicted no, and they don't have the disease (correct).
- **True Positives (TP):** Our model predicted yes, and they do have the disease (correct).
- **False Positives (FP):** Our model predicted yes, but they don't actually have the disease (wrong prediction - known as a "Type I error").
- **False Negatives (FN):** Our model predicted no, but they actually do have the disease (wrong prediction - known as a "Type II error").

SCIENCE ACADEMY
Excellence in Teaching & Learning

Logistic Regression

(Model Evaluation)

Confusion matrix:

n = 100	Predicted No	Predicted Yes
Actual No = 43	TN = 40	FP = 3
Actual Yes = 57	FN = 7	TP = 50
	47	53

Let's define the basic terminology:

- **True Negatives (TN):** Our model predicted no, and they don't have the disease (correct).
- **True Positives (TP):** Our model predicted yes, and they do have the disease (correct).
- **False Positives (FP):** Our model predicted yes, but they don't actually have the disease (wrong prediction - known as a "Type I error").
- **False Negatives (FN):** Our model predicted no, but they actually do have the disease (wrong prediction - known as a "Type II error").

SCIENCE ACADEMY
Excellence in Teaching & Learning

Logistic Regression

(Model Evaluation)

Confusion matrix:

n = 100	Predicted No	Predicted Yes
Actual No = 43	TN = 40	FP = 3
Actual Yes = 57	FN = 7	TP = 50
	47	53

Accuracy:

Overall, how often our model predicted correct?

Accuracy = correct predictions / total

Accuracy = (**TN** + **TP**) / total = 90 / 100 = **0.90**

Misclassification Rate / Error Rate:

Overall, how often our model predicted wrong?

Error Rate = wrong predictions / total

Error Rate = (**FP** + **FN**) / total = 10 / 100 = **0.10**

Specificity:

When it's actually No, how often does the model predicts No?

Specificity = **TN** / actual No = 40 / 43 = **0.93**

Precision:

When it predicts yes, how often the model is correct?

Precision = **TP** / predicted Yes = 50 / 53 = **0.94**

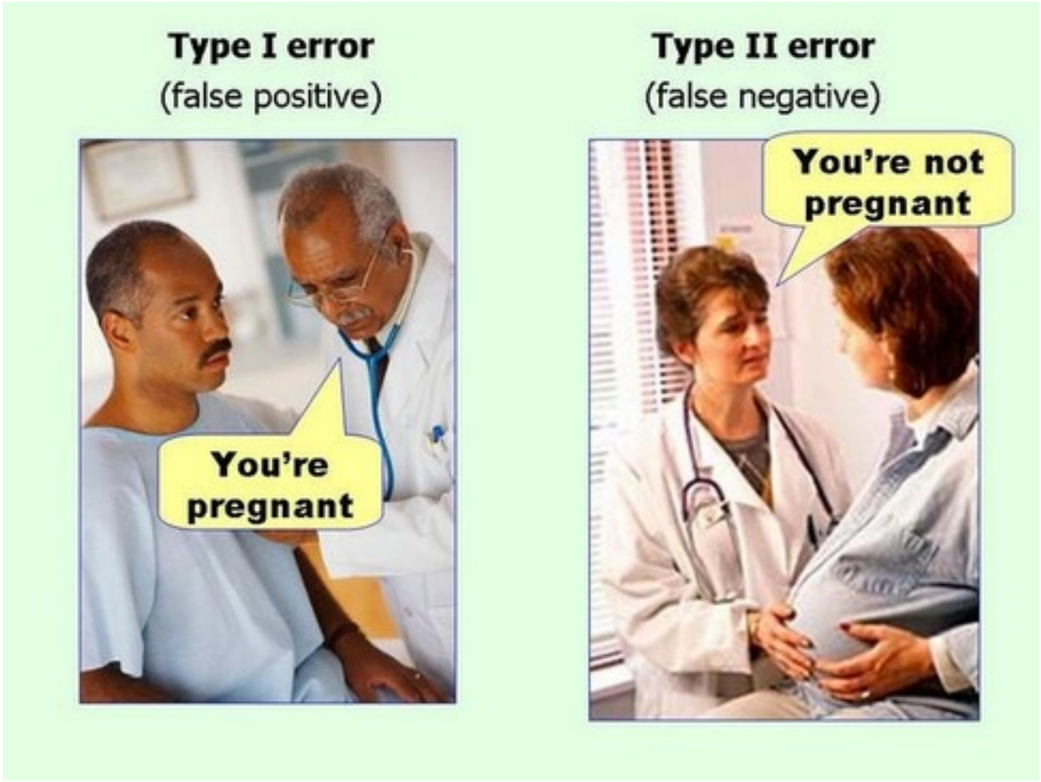
SCIENCE ACADEMY
Excellence in Teaching & Learning

Logistic Regression

(Model Evaluation)

Confusion matrix:

The example below may help to understand Type I and Type II error!



Let's move on to the jupyter notebook and learn by doing with data!