Introduction to Logistic Regression

Introduction

Recall that **linear regression** can be used to describe the relationship between two or more variables where the target variable is numeric.

For categorical targets, we can use **logistic regression**.

Example - ICU Admission

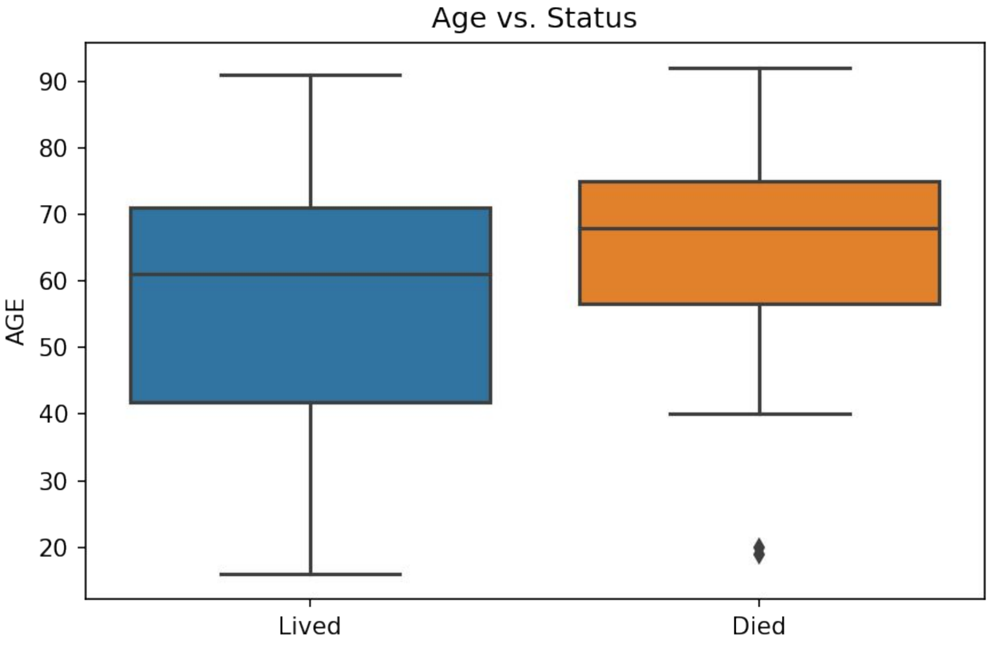
Let’s say we want to study survival rates of patients admitted to an intensive care unit. We gather several variables:

- Status: lived or died
- Age
- Sex
- Systolic Blood Pressure
- Type of Admission (Elective or Emergency)
- etc.

Example - ICU Admission

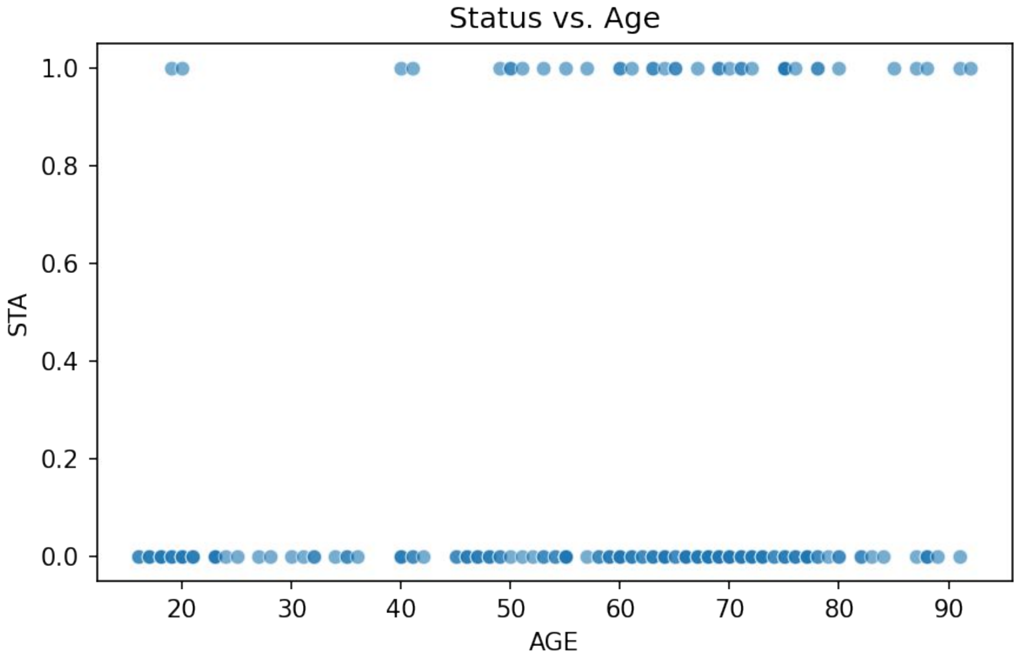
We might start by examining age vs. status.

It appears that those that died tended to be older.



Example - ICU Admission

We could also plot this as a scatterplot, where we encode status numerically, with lived = 0 and died = 1.



Example - ICU Admission

How can we build a model to describe the relationship between age and status?

Example - ICU Admission

Can we use a linear regression model?

Linear regression: The distribution of *Y*, given *X* is normal with mean

$$\mu = \beta_0 + \beta_1 X$$

Example - ICU Admission

Can we use a linear regression model?

Linear regression: The distribution of *Y*, given *X* is normal with mean

$$\mu = \beta_0 + \beta_1 X$$

Would it make sense to use this model
for our target here (lived or died)?

Example - ICU Admission

Can we use a linear regression model?

Linear regression: The distribution of *Y*, given *X* is normal with mean

$$\mu = \beta_0 + \beta_1 X$$

Would it make sense to use this model for our target here (lived or died)?

Idea: Instead of using 0/1 as our target, let's make our target a *probability*.

Recall: Bernoulli Distribution

Setup: An experiment with exactly two outcomes, labeled "success" (denoted by 1) and "failure" (denoted by 0).

> Probability of success = p
> Probability of failure = $1 - p$

Example: A marketing company knows that historically, search ads have a click-through rate of 1.5%.
We can view each interaction as a Bernoulli trial with $p = 0.015$

Example - ICU Admission

Can we use a linear regression model?

Linear regression: The distribution of *Y*, given *X* is **normal** with mean

$$\mu = \beta_0 + \beta_1 X$$

Logistic regression: The distribution of *Y*, given *X* is **Bernoulli** with probability of success (mean)

$$p = \beta_0 + \beta_1 X$$

Example - ICU Admission

Can we use a linear regression model?

Linear regression: The distribution of *Y*, given *X* is **normal** with mean

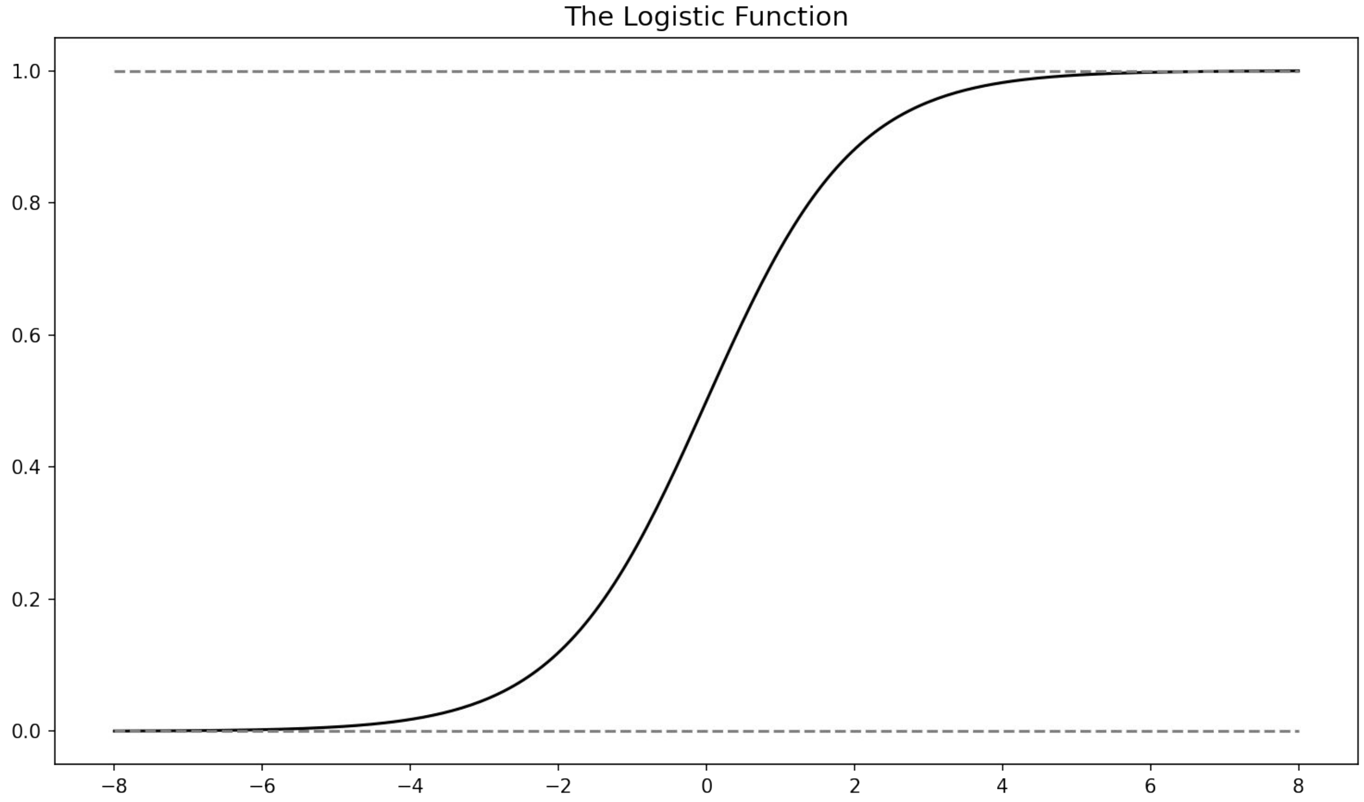
$$\mu = \beta_0 + \beta_1 X$$

Logistic regression: The distribution of *Y*, given *X* is **Bernoulli** with probability of success (mean)

$$p = \beta_0 + \beta_1 X$$

But wait, a probability must be between 0 and 1, and there is no guarantee that this expression will be.

The logistic function: $f(x) = \dfrac{1}{1 + e^{-x}}$



Example - ICU Admission

Can we use a linear regression model?

Linear regression: The distribution of *Y*, given *X* is **normal** with mean

$$\mu = \beta_0 + \beta_1 X$$

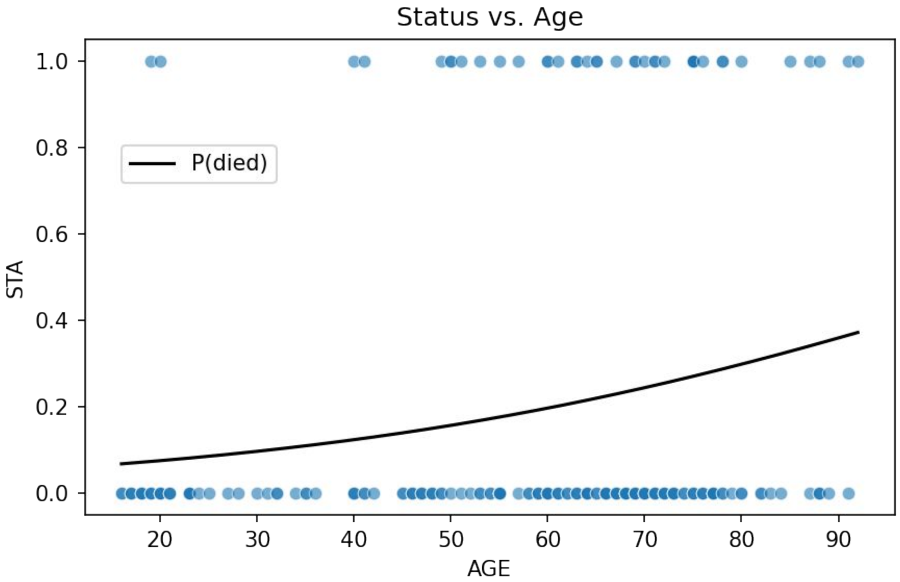
Logistic regression: The distribution of *Y*, given *X* is **Bernoulli** with probability of success (mean)

$$p = \text{logistic}(\beta_0 + \beta_1 X)$$

Example

If we fit a logistic regression model to the ICU data, using age, we get

$$P(\text{died}) = \text{logistic}(-3.0585 + 0.0275(\text{age}))$$



Inference for Logistic Regression Models

Types of questions that we can ask:

- How precise is our estimate of the coefficient associated with age?
- Is the coefficient associated with age statistically significant?
- If I add additional predictor variables, are their coefficients statistically significant, after controlling for age?

Inference vs. Prediction

When building a statistical model, there are a number of possible objectives:

- **inference:** identifying key explanatory variables and understanding the relationship between these variables and the target
- **prediction:** predicting the outcome on new observations

Predictive analytics typically focuses on model-building for prediction rather than inference, and the techniques you can use in each differ.

Predictions on a New Observation

Once we have build a logistic regression model, we can use it to generate predictions on new observations.

To do this, we much translate our predicted probabilities π_i's into predictions.

A simple rule that can be used is to predict 1 if π_i > 0.5 and 0 otherwise.

Predictions on a New Observation

There are a number of metrics that can be used to evaluate predictions of a model on the basis of True Positives, False Positives, True Negatives, and False Negatives.

When evaluating the performance of a predictive model, it should be done by separating out a test set of data which the model is not fit on.

Logistic Regression

Let’s see all of this in action in a Jupyter notebook.