Prediction with different types of regression

As you have been learning, key regression techniques that you will encounter in your work as a data professional include linear regression, hypothesis testing, and logistic regression. When your goal is to make predictions with data, it is important to consider these different approaches and think about which approach will best help you achieve your task. In this reading, you will learn more about how to choose the most relevant regression technique for a project, based on the question you want to answer, the outcome variable, and how it is measured.

How to choose a regression technique

When choosing a regression technique, it is important to consider the data you are working with and the question you want to address.

Things to consider

1. What is the question you want to answer? In other words, what do you want to predict?
2. Which variable in your data can be the outcome variable?
3. How is the outcome variable measured? If the outcome variable is continuous, it is more likely that either linear regression or hypothesis testing will be most appropriate. However, if the outcome variable is binary, you will find logistic regression to be more useful.

Example contexts for regression

The following examples demonstrate how the questions about prediction, outcome variable, and measurement can be navigated in order to choose a regression technique.

Example context: User engagement

In your work as a data professional, imagine that you are interested in making predictions about user engagement for a mobile app.
First, you might ask, what is the question you want to answer?
One possible question could be “How much does each in-app feature influence user engagement?”
The in-app features might include a live chat with customer support, an FAQ section that updates weekly, and a community space to connect with other users. Next, you might ask, which variable in

your data can be the outcome variable? If you have access to data about users' session lengths (in other words, how long users spend in the app each time they open it), the outcome variable can be session length. Your next question might be: how is the outcome variable measured? Session length can be measured by number of minutes, which is continuous. Because the outcome variable is continuous, and you are interested in how much each feature influences the outcome variable, you could proceed with linear regression and check the relevant model assumptions. If there is only one feature of interest, you would build a simple linear regression model. If there are multiple features of interest, you would build a multiple linear regression model.

Another question of interest could be "Does a dynamic landing page versus a static landing page make a difference in user engagement?" The outcome variable can be session length, measured by number of minutes, for this example, too. Since the outcome variable is continuous, and the target question is about whether there is a difference in user engagement when one type of landing page is used over the other, you could proceed with hypothesis testing. You can then frame the hypotheses, which could be the following:

- Null hypothesis (H_0): Users spend approximately the same amount of time in the app when the landing page is dynamic versus when it is static.
- Alternative hypothesis (H_1): Users do NOT spend approximately the same amount of time in the app when the landing page is dynamic versus when it is static.

Another question you might be interested in is "Will a user engage with the new line of products in-app?" Next, you might ask, which variable in your data can be the outcome variable? If you have access to data about whether a user clicks to view the new line of products, that could be the outcome variable. The next question is: how is the outcome variable measured? Whether a user clicks to view that content can be represented as a binary variable, with 1 indicating they clicked to view the content and 0 indicating that they did not click to view that content. Since this outcome variable is binary, you could proceed with binomial logistic regression.

Example context: Patient response

Now imagine that you are tasked with making predictions about patient responses to medical treatments.

You can start by asking, what is the question you want to answer?

A possible question could be "How much does each factor influence a patient's response to a medical treatment?" If the goal of the treatment is to improve white blood cell (WBC) count and you have access to that data, WBC count can be the outcome variable. The outcome variable is a continuous measure, and you could use linear regression to address this task.

Another question of interest could be “Will Treatment A, Treatment B, or Treatment C have a stronger impact on a patient’s WBC count?” The outcome variable in this case would also be WBC count, which is continuous. Since the target question is about comparing different treatments, it would be best to proceed with hypothesis testing. You can then form the hypotheses, which could be the following.

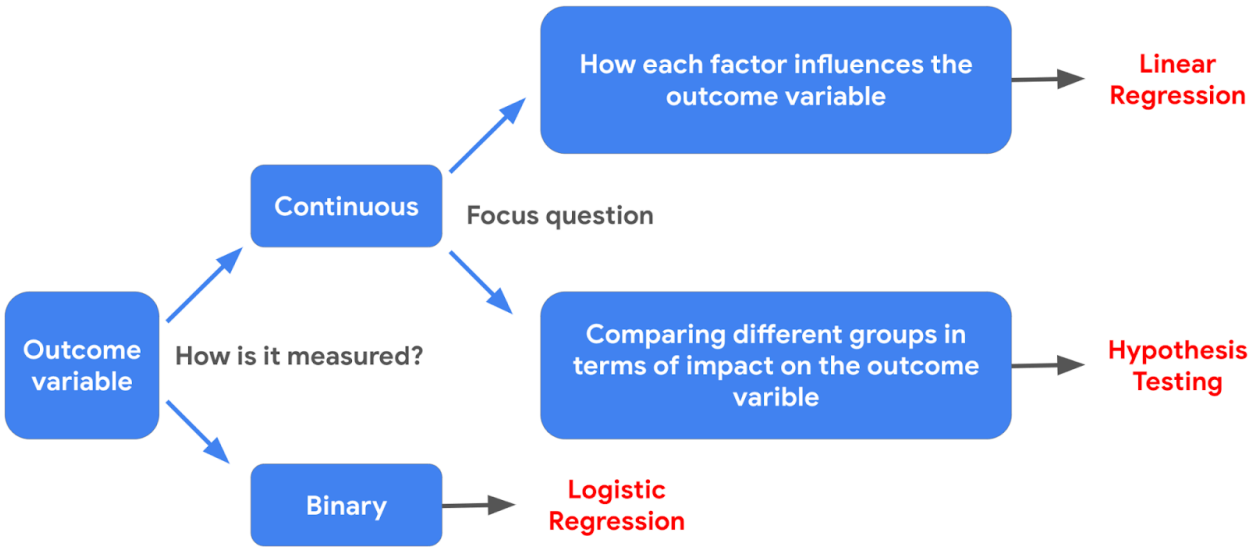
- Null hypothesis (H_0): Patients have approximately the same white blood cell count with each treatment.
- Alternative hypothesis (H_1): Patients do NOT have approximately the same white blood cell count with each treatment.

A different question you might be interested in: “With Treatment A, will a patient’s WBC count reach the ideal range?” If you have access to the associated data, the outcome variable would be whether a patient’s WBC count reaches the ideal range or not, which is a binary variable: 1 indicating that their WBC count falls within the ideal range and 0 indicating that it does not. You could build a logistic regression model to make predictions in this scenario.

Key takeaways

- Consider the question you want to answer and the data you have access to when choosing a regression technique for making predictions.
- Identifying the outcome variable of interest and how it is measured will help you decide which regression technique is most suitable for your task.

The following flowchart captures a high-level approach for choosing a regression technique, starting from the outcome variable, as discussed in this reading. Also note that hypothesis testing is connected to regression analysis. For example, in linear regression, the process of testing whether there is a correlation between two variables (in other words, determining if the coefficients are statistically significant in the linear model) involves a hypothesis test.


Outcome variable
How is it measured?
Continuous
Binary
Focus question
How each factor influences the outcome variable
Comparing different groups in terms of impact on the outcome varible
Linear Regression
Hypothesis Testing
Logistic Regression