7.1: General Linear Model

> *If you should say to a mathematical statistician that you have discovered that linear multiple regression analysis and the analysis of variance (and covariance) are identical systems, he would mutter something like, "Of course – general linear model," and you might have trouble maintaining his attention. If you should say this to a typical psychologist, you would be met with incredulity, or worse. Yet it is true, and in its truth lie possibilities for more relevant and therefore more powerful exploitation of research data. (Cohen 1968)*[1]

The above quote is from Cohen (1968) when he introduced psychology to the General Linear Models (GLM) using the language of regression analysis. So, what is GLM exactly? Before we delve into this topic, let's have a history lesson.

A bit of history – Pearson vs Fisher

There are three important statisticians that I would like to introduce to you – Francis Galton, Karl Pearson and Ronald Ayman Fisher. Some of these names have been mentioned previously in the book. Francis Galton introduced the statistical concepts of correlation and regression and was credited as the first person to apply statistical methods to study human differences and intelligence. Karl Pearson, a protege of Galton, built on his mentor's work and popularised various statistical analyses such as chi-squared tests. Both Galton and Pearson's work played a big role in the eugenic movements and was criticised as scientific racism.[2]

Fisher, on the other hand, is an experimentalist who primarily worked on agriculture. He proposed the concepts of *mean generalised* and *analysis of variance* (mostly known as ANOVA) using Galton's ideas on regression and Pearson's ideas on probability distribution. Fisher also published a journal article that criticised one of Pearson's formulas while introducing the idea of degrees of freedom (Salsburg, 2001).[3] Due to his work, RA Fisher is credited as the founder of modern statistics.

This brief history lesson provides background information on why people think that regression-based analysis and group-differences analysis (such as t-tests and ANOVAs) are different Regression analysis was popularised by figures like Galton and Pearson to explore *natural variations.* A couple of decades later, Fisher developed the analysis of variance and analysis of covariance to study artificial or controlled variations due to his experimentations in agriculture (Cohen, 1968).

Over the years, misconceptions about ANOVAs, t-tests, and regression analysis began to emerge. Some individuals held the mistaken belief that ANOVAs could establish causation while thinking that regression analysis could not. There was also a misconception that regression analysis was only suitable for numerical variables, while ANOVAs were somehow more appropriate for experiments. However, there are no arguments that the statistical methods are inherently related to each other. Some statisticians even suggest that ANOVA is *just a special case of the GLM.* Some psychology educators (me included), think that this division creates unnecessary stress for students as they are led to believe that the statistical analyses they learn are distinct and disconnected.

In saying that however, I do acknowledge that learning about ANOVAs and t-tests still has its place. For instance, a commentary on Twitter argued that learning about ANOVA "forces an understanding of and respect for degrees of freedom" and "treats experiments like actual effing experiments". They also argued that "people trained in ANOVA can correctly use regression to analyze experiments. People trained in regression but not ANOVA mostly can not, in my experience", (Brewer, 2023).[4]

Nevertheless, I believe that there are more advantages to teaching statistics using the GLM approach because it's easier for students to transition into more advanced topics when there's a framework that binds all statistical concepts together. Therefore, this framework will be the basis of the current chapter.

What is the General Linear Model?

Remember that early in the book we described the basic model of statistics:

$$data = model + error$$

Where our general goal is to find the model that minimises the error, subject to some other constraints (such as keeping the model relatively simple so that we can generalise beyond our specific dataset).

You have already seen the general linear model (GLM) in the earlier chapters where we modelled height in the NHANES dataset as a function of age. As you can see in Table 7.1.1, nearly all of the statistical analyses you will encounter in a psychology statistics

course can be framed in terms of the GLM or an extension of it.

A general linear model is one in which the model for the outcome variable (which is often referred to as *Y*) is composed of a *linear combination* of predictors (which is often referred to as *X*) that are each multiplied by a weight (which is often referred to as the Greek letter beta – β), which determines the relative contribution of that predictor variable to the model prediction.

Let's try a more intuitive explanation. A general linear model is like a recipe for making predictions. Imagine you're trying to predict something, like the price of a house. In this model, you have a bunch of ingredients (predictors), like the size of the house, the number of bedrooms, and so on. Each ingredient is given a number (the beta value), which tells you how important that ingredient is for making the prediction. The bigger the beta, the more impact that *ingredient* has on the prediction.

So, you take the size of the house, multiply it by its beta, and then do the same for the number of bedrooms and all the other *ingredients*. You then add all these numbers together, which gives you the final prediction. GLM helps you figure out the best combination of these *ingredients* to make the most accurate prediction.

Table 7.1.1. The differences between the GLM equation and the statistical tests procedure that are commonly taught in undergraduate statistics courses in psychology, adapted from Fife, 2022[5] and used under a CC BY-SA licence

Procedure	GLM Equation	Interpretation
one sample t test	$y = b_0$	b0 (the intercept) is the value we are testing against
independent-sample t test	$y = b_0 + b_1 \times Treatment$	b0 (the intercept) is the mean of the control group and b1 is the difference between treatment and control groups
within sample t test	$Time_2 Time_1 = B_0$	b0 (the intercept) is the average difference from Time 1 to Time 2
ANOVA	$y = b_0 + b_1 Treatment_A \times Treatment_B$	b0 (the intercept) is the mean of the control group, b1 is the difference between Treatment A and the control, and b2 is the difference between Treatment B and the control.
ANCOVA	$y = b_0 + b_1 Covariate + b_2 \times Treatment$	b0 (the intercept) is the mean of the control group, b1 is the slope of the covariate, and b2 is the difference between the Treatment and the control group.
Factorial ANOVA	$y = b_0 + b_1 \times Treatment + b_2 \times Female + b_3 \times Female \times Treatment$	b0 (the intercept) is the mean of the men in the control group, b1 is the difference between Treatment and control, b2 is the difference between Males and Females, and b3 is the difference between females in the treatment group and males in the control group.

Why is it important to learn about the GLM? Because it eliminates the need to commit complex decision trees with intricate rules to memory when determining which statistical model to use, as depicted in the image above. This simplification is incredibly beneficial. Essentially, all you need to know is which variables you aim to predict (i.e., your outcome variable) and which variables you use for prediction (such as IQ, SES, gender, or treatment/control status).

Group Differences Versus Predictors

When discussing different categories (e.g., treatment vs. control, males vs. females, freshmen vs. seniors), we often express our interest in estimating *differences between these groups*. For example, males and females *differ* in their hand grip strength. However, you could express group differences by stating that group membership *predicts* scores on the outcome variable. For example, gender *predicts* hand grip strength.

From a mathematical standpoint, there's absolutely no distinction between estimating group differences and predicting an outcome. It's purely a matter of terminology. Some may object to the use of this language, arguing that "it's not appropriate to refer to group membership as a predictor."

Well, in practical terms, it doesn't make a mathematical difference, so why engage in semantic disputes?

Moreover, adopting a consistent terminology can simplify the decision-making process in statistical analysis. Once more, it's simply a matter of distinguishing which variables act as predictors and which one serves as the outcome.

In the next few sections, we will apply the GLM framework using different questions. We will start with assessing group differences.

Chapter attribution

This chapter contains material taken and adapted from *The Order of the Statistical Jedi* by Dustin Fife, used under a CC BY-SA 4.0 licence.

---

1. Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin*, *70*(6, Pt.1), 426-443. doi.org/10.1037/h0026714 ↵
2. Nobles, M., Womack, C., Wonkam, A. & Wathuti, E. (2022, June 8). Science must overcome its racist legacy: *Nature*'s guest editors speak [Editorial]. *Nature*. https://www.nature.com/articles/d41586-022-01527-z ↵
3. Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the twentieth century*. Macmillan. ↵
4. Brewer, N. [@noelTbrewer]. (2023, August 14). *Yes and militant about it. ANOVA forces an understanding of and respect for degrees of freedoms (and familywise error), treats* [Post]. X. twitter.com/noelTbrewer/status/1690844693951631360 ↵
5. Fife, J. (2022). *The Order of the Statistical Jedi: Responsibilities, routines, and rituals*. QuantPysch. quantpsych.net/stats_modeling/the-general-linear-model.html ↵
