Chapter 9

The General Linear Model (GLM): A gentle introduction

9.1 Example with a single predictor variable.

Let's start with an example. Schizophrenics smoke a lot. They smoke between two and three times more than the general population and about 50% more than those with other types of psychopathology (**??**). Obviously, explicating the nature of this relationship might provide insights into the etiology of schizophrenia.

One early type of research into this area compared the density of cholingergic nicotinic receptors (nAChR) in the brains of schizophrenics and controls (**?**). The data set "Schizophrenia and nicotinic receptors" shown in Table 9.1. gives hypothetical data of such a study done in the past when analysis of post mortem brain specimens was the only way to examine this question.

For the moment, ignore the variables Age, Smoke and Cotinine and let us ask the simple question of whether schizophrenics have more or fewer nicotinic receptors in the brain area used in this study. The operative word in the general linear model (GLM) is "linear." That word, of course, implies a straight line. Hence, mathematically we begin with the equation for a straight line. In statisticalese, we write

$$\hat{Y} = \beta_0 + \beta_1 X \tag{9.1}$$

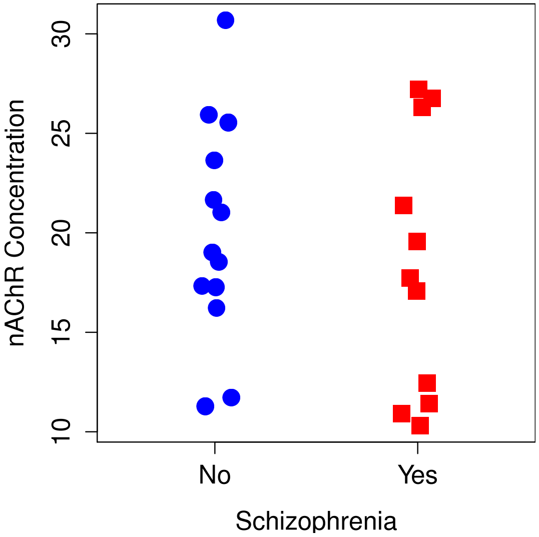
Read "the predicted value of the a variable ($\hat{Y}$) equals a constant or intercept (β_0) plus a weight or slope (β_1) times the value of another variable (X). Let's look at the data first by plotting Y (not $\hat{Y}$) as a function of X, or in the example, variable nAChR as a function of variable Schizophrenia (see Figure 9.1).

The purpose of a GLM is to fit a straight line through the points in Figure 9.1. Here is where the βs in Equation 9.1 come in. β_0 is the intercept for a straight line, i.e., the value of Y when X is 0. β_1 is the slope of the line. When $\beta_1 = 0$, then the predicted nAChR density for schizophrenics is the same as

Table 9.1: Data set on schizophrenia and brain density of nicotinic receptors.

	Schizophrenia	SzDummyCode	Age	Smoke	Cotinine	nAChR
1	No	0	55	No	2.00	18.53
2	No	0	83	No	9.03	11.73
3	No	0	52	?	5.6	19.01
4	No	0	74	No	2.00	25.93
5	No	0	61	No	2.00	21.66
6	No	0	56	?	103.11	25.54
7	No	0	80	?	5.27	11.28
8	No	0	84	?	4.85	16.22
9	No	0	49	Yes	85.19	30.69
10	No	0	87	Yes	78.54	21.03
11	No	0	74	?	72.33	23.65
12	No	0	44	?	4.40	17.27
13	No	0	94	No	2.00	17.34
14	Yes	1	91	?	4.69	11.41
15	Yes	1	70	?	100.70	10.90
16	Yes	1	58	?	65.50	21.38
17	Yes	1	61	?	78.89	12.45
18	Yes	1	42	Yes	84.64	27.20
19	Yes	1	70	?	66.74	17.08
20	Yes	1	69	?	108.62	26.77
21	Yes	1	30	?	74.00	19.56
22	Yes	1	70	Yes	90.08	17.73
23	Yes	1	40	Yes	113.77	26.30
24	Yes	1	91	No	2.00	10.30

Figure 9.1: Number of nicotinic receptors (nAChR) as a function of diagnosis.



that for controls. As the slope deviates from 0, in either a positive or negative direction, then there is more and more predictability.

At this point, you may rightly ask how one can have an intercept and a slope for a variable that has values of "No" and "Yes." We' will see the answer later, but for the time being let us create a numeric variable called SzDummyCode that has the numerical value of 1 for Schizophrenia = "Yes" and 0 otherwise.[1] Running the GLM gives these estimates: $\beta_0 = 19.99$ and $\beta_1 = -1.71$. Hence, for controls, the value of X in Equation 9.1 is 0, so the predicted nAChR concentration is

$$\hat{Y} = 19.99 - 1.71 * 0 = 19.99$$

and for the schizophrenics in the sample,

$$\hat{Y} = 19.99 - 1.71 * 1 = 18.28$$

One reason for calling the general linear model "general" is that it can handle an X that is *not* numerical as well as one that is numerical. Hence, there is no difference between performing a GLM analysis using Equation 9.1 with X is variable Schizophrenia with values of "No" and "Yes" and performing one where X is the numerical variable SzDummyCode with values of 0 and 1. Table 9.2 gives the results of GLMs in which the X variable is the numeric SzDummyCode (top) and in which the X variable is the qualitative variable Schizophrenia.

Notice that there are no differences in any value between the output for variable SzDummyCode and Schizophrenia. Notice also that there the bottom half of the table labels the variable "SchizophreniaYes" and not simply "Schizophrenia." This is a hint as to what is going on when the GLM handles a nonnumeric

[1] Dummy coding is described in Section X.X.

Table 9.2: GLM results using a numeric (SzDummyCode) and a nonnumeric (Schizophrenia) variable.

Numeric variable SzDummyCode				
Variable	Estimate	St. Error	t	p
Intercept	19.991	1.675	11.938	4E-11
SzDummyCode	-1.711	2.473	-0.692	.496

Nonnumeric variable Schizophrenia				
Variable	Estimate	St. Error	t	p
Intercept	19.991	1.675	11.938	4E-11
SchizophreniaYes	-1.711	2.473	-0.692	.496

X variable. All GLM programs change the nonnumeric variable into a numeric one so that they can solve the mathematical problem. After that is done, the GLM "translates" the numerical output back into the original categories. Hence, the "SchizophreniaYes" using the variable Schizophrenia signifies that one should add -1.711 to the value of the intercept to get the predicted value when the variable Schizophrenia = "Yes."

(A cautionary aside: Different GLM programs use different mechanisms for converting the categories in a nonnumeric variable into numbers. Also, a user can specify how to perform the conversion. Thus, the values of the βs can be different for different coding schemes for the same problem. The predicted values, however, for the groups will always remain the same).

Finally, look at the p value for the effect. It is .496 and definitely nonsignificant. One might be tempted to conclude that there is no difference in nAChR concentrations between schizophrenics and controls, but that would be unwise. To see why, we must combine substantive knowledge on neuroscience with statistics.

9.2 Example with more than one predictor variable.

Remember, schizophrenics smoke a lot. Most of you have already asked yourself about the effect of smoking on the nicotinic receptor density. Similarly, smoking is associated with early death, so any effect of age on nAChR concentration might also cloud the results. These are not trivial issues because there is evidence that the number of nicotinic receptors decrease with age (**?**) and that they are upregulated by the use of nicotine (**?**). The increase in nAChR from smoking and early death might have masked the differences between schizophrenics and controls in this hypothetical study.

Table 9.3: Results of the GLM predicting nAChR from Age and SzDummyCode

Variable	Estimate	Std.Error	t	p
Intercept	32.61	4.45	7.33	5E-07
Age	-0.18	0.06	-2.99	0.007
SzDummyCode	-2.77	2.14	-1.29	0.211

Ideally, one would like to have a control matched to each schizophrenic on age of death and smoking status at or near death. The practicalities of research with brain banks, however, make it difficult and expensive–perhaps even impossible–to pull that off. Smoking status at death is often not known, and even if it is known, there is wide variability in the amount of nicotine intake among smokers. Indeed, the data on variable Smoke (was the person a smoker at or near death?) in Table 9.1 has so many unknowns as to make the variable useless. One way to address this issue is to measure brain cotinine, a metabolite of nicotine, because it has a longer half-life than nicotine.

We now want to control for both age and cotinine levels. We could divide the specimens into groups by categorizing variables Age and Cotinine, but that approach is not recommended. In fact, it is downright stupid. If we used a cutoff of 65 on age for "young" versus "old," there would be no young schizophrenics with low cotinine values, and we would be comparing groups of size four with those of size two in other categories.

A GLM approach, however, avoids this. Suppose that we want to control for Age. We just add a second X variable to the right-hand side of Equation 9.1, or

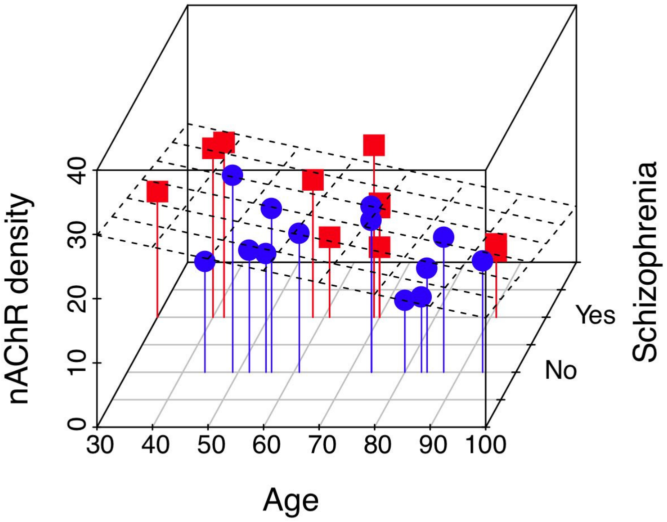
$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{9.2}$$

It is good practice to put any control variables into the equation before the variable of interest so X_1 denotes variable Age and X_2 is, as before, SzDummyCode (or Schizophrenia). Instead of a two dimensional plot as in Figure 9.1, the problem would now be visualized via a three dimensional plot. Variable nAChR would be axis equivalent to the height of the plot while Age and Schizophrenia would be the width and depth dimensions. With a single predictor variable, the predicted values form a straight line in a two-dimensional plot. With two predictor variables, the predicted nAChR levels form a plane in a three dimensional plot. Figure 9.2 gives an example.

From the prediction plane in the figure, age is associated with lower nAChR levels. Although it is difficult to tell from the plot, there is also a downward projection of the plane suggesting a decrease in the brains of schizophrenics. Would controlling for age now reveal a significant difference between controls and schizophrenics?

Table 9.3 gives the results of the GLM that predicts nAChR from Age and the dummy code for schizophrenia. . It is helpful to write the prediction equation

Figure 9.2: A scatterplot with two predictor variables.



twice, once for controls and the second time for schizophrenics:

$$\widehat{\text{nAChR}}_C = 32.61 - .18 * \text{Age}$$

$$\begin{aligned} \widehat{\text{nAChR}}_\text{S} &= 32.61 - .18 * \text{Age} - 2.77 \\ &= 29.84 - .18 * \text{Age} \end{aligned}$$

There are two salient aspects about the concept of *control* in the GLM. The first, arbitrarily called *predictive control* here, is evident by plugging any single value of age into both of the equations. No matter what value of age, schizophrenics will always be predicted to have 2.77 units of nicotinic receptors less than controls. Hence, we can use the following language to describe these results: "controlling for age, schizophrenics are predicted to have 2.77 fewer units of nAChR than controls."

The second type of control may be called *statistical control*, and it applies to the statistical significance of the results. From Table 9.3, the coefficient for age is significant while the coefficient for variable SzDummyCode is not. The statistics behind calculation of the p values are complicated, but their meaning is simple. For age, the meaning is equivalent to the following: "controlling for diagnosis, does age predict nAChR better than chance?" The answer here is "Yes."

For diagnosis, the relevant question is "controlling for any age differences between schizophrenics and controls, is the 2.77 unit difference between the two greater than chance?" Here, the answer is "No." It is logical to hypothesize that the excess early mortality associated with schizophrenia may have obscured

Table 9.4: Predicting nAChR from age, cotinine and diagnosis.

Variable	Estimate	Std.Error	t	p
Intercept	26.20	4.44	5.09	9E-06
Age	-0.12	0.06	-2.15	0.044
Cotinine	0.08	0.03	2.87	0.009
SzDummyCode	-5.70	2.11	-2.70	0.014

differences in nAChR density between them and controls in the initial analysis. The current GLM gives no support to that idea.

We now want to control for cotinine, so we enter that variable into the GLM. In "variable-ese" the equation is

$$\widehat{\text{nAChR}} = \beta_0 + \beta_1\text{Age} + \beta_2\text{Cotinine} + \beta_3\text{SzDummyCode}$$

or in statisticalese,

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Table 9.4 gives the results of this GLM.

Once again, write the equation for controls and the one for schizophrenics:

$$\widehat{\text{nAChR}}_C = 26.20 - .12\text{Age} + .08\text{Cotinine}$$

$$\begin{aligned}\widehat{\text{nAChR}}_S &= 26.20 - .12\text{Age} + .08\text{Cotinine} - 5.7(1) \\ &= 20.50 - .12\text{Age} + .08\text{Cotinine}\end{aligned}$$

Note again that if we substitute into both equations any single value for age and any single value for cotinine, then we predict that schizophrenics will have 5.7 fewer units of nAChr than controls. From Table 9.4, that difference is now significant!

The fact that all three variables in Table 9.4 are significant tells us that:

1. increases in age (regardless of, or controlling for, cotinine and diagnosis) predict lower nAChR levels better than chance;

2. that increases in cotinine (regardless of, or controlling for, age and diagnosis) predict higher nAChR levels better than chance;

3. that an "increase" in diagnosis or the presence of schizophrenia (regardless of, or controlling for, age and cotinine) predicts decreases nACHr density better than chance.

Why did we not find an association between schizophrenic an nAChR density in the first analysis? The answer is simple–schizophrenics smoke a lot.

Schizophrenics smoke more than controls. Because of the amount of missing data for smoking status at death, the initial brain samples could not be adequately matched for this important variable. Consistent with previous evidence, nicotine up regulated acetylcholine nicotinic receptors and, of course, results in high levels of its metabolite cotinine. This up regulation masked the difference in nAChR density between schizophrenic and control brains in the initial analysis.

Hence, the conclusion of this exercise is that schizophrenia is associated with decreases in nAChR number. Note carefully that the operative word is "*associated*". Synonyms would be "*correlated*" and "*predicted.*" Finally, note that any real life analysis would start with the third GLM that used age, cotinine and diagnosis as predictors. The order of presentation for the GLMs above was purely didactic.

9.3 GLM terminology

As in the vocabulary for any system that has evolved over time, GLM terminology can be confusing. As statistical theory grew, it was realized that several different techniques could be combined into a single, general technique. Hence, the term *general* in GLM. Also, the advent of digital computers permitted the mathematics behind the general approach to be implemented. Nevertheless, we are left with a legacy of terms derived from the old techniques as well as tables and short cuts used in hand calculations.

The first type of terminology applies to the variables in the GLM. The variable on the left hand side of the GLM equation (Y or nAChR in the example) is called the *dependent*, *predicted*, or *response* variable. The variables on the right side of the equation (the Xs or Schizophrenia, Age, and Cotinine) are called the *independent*, *predictor*, or *explanatory* variables. Usually, these terms are paired: dependent with independent, predicted with predictors, and response with explanatory.

An independent, predictor or explanatory variable that is measured with numbers is called a *numeric* or *quantitative* variable or a *covariate.* One that is not numeric (or uses numbers to indicate groups) is called a *factor.*[2] The specific groups within a factor are termed the *levels* of that factor. For example, the factor sex would have two levels–female and male.

The three classic statistical procedures that comprise the GLM are: (1) the analysis of variance or ANOVA; (2) the analysis of covariance or ANCOVA; and (3) regression. In ANOVA, all of the independent variables are factors (i.e., qualitative variables). An ANOVA with only one factor is called a *oneway ANOVA*. An ANOVA with more than one factor is called a *factorial ANOVA*. Often a factorial ANOVA is described by the number of levels in the factors. For example, if the first factor has two levels and the second factor has three levels, the model is called a "two by three" design or "two by three ANOVA."

[2] Sometimes nonnumeric variables are called qualitative variables.

A *regression* is GLM in which all of the variables are quantitative. When there is only one X or independent variable, the regression is called a *simple regression.* When there are two or more Xs, the regression is called a *multiple regression.*

An ANCOVA is a GLM with at least one qualitative and at least one quantitative predictor. Hence, ANCOVA is synonymous with GLM. Most statisticians today eschew the term ANCOVA and use GLM.

9.3.1 Orthogonal and non orthogonal designs

In generic statisticalese, the word *orthogonal* is a synonym for uncorrelated. Like most jargon in science, it was probably developed for two reasons: (1) lend an air of respectability to statistics as a science; and (2) deliberately confuse anyone trying to learn the field. When all independent variables of a GLM are uncorrelated with one another, then the model is *orthogonal.* When at least one pair of independent variables are correlated, the design is *non-orthogonal.* If the GLM has at least one continuous independent variable, then always regard it as non-orthogonal[3]. Hence, the term orthogonal only applies to classic ANOVA , i.e., when all independent variables are strictly categorical. An ANOVA is orthogonal when *each cell contains the same number of observations.* This condition is also termed a *balanced design.*

In an orthogonal design, there is one and only one mathematical way to estimate the parameters of the model and to perform the statistical tests. In non-orthogonal designs, however, there is more than one way to compute these statistics, so the user must make some assumptions about the best way to interpret the results.

Finally, orthogonality is not akin to falling off a cliff. A two by two ANOVA that has eight rats in three of its cells but seven in the fourth is so close to being orthogonal that the different ways of estimating the sums of squares will all yield the same substantive results. Hence, most designs in experimental neuroscience will be close to being orthogonal. The issue is much more salient for certain types of observational research. A random sample of, say, alcoholics or sociopaths will contain roughly three males for every female. Here, one must be very careful about which type of sums of squares to request and to interpret when a variable like gender is in the model. In general, *the more correlated the independent variables are, the more care must be taken in interpreting the results.*

9.4 The meaning of the betas

The general equation for GLM is

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \tag{9.3}$$

[3] There are exceptions to this rule but they are beyond the scope of this book.

The βs in a GLM are coefficients or weights assigned to the predictor variables, i.e. the Xs on the right had side of the prediction equation. Here, let us explore some properties of these coefficients.

The first β, β_0, is a constant. That it, it is the same for every observation regardless of any values on any of the Xs. In geometrical terms, β_0 is an intercept. Examine Equation 9.3 and let all of the Xs equal 0. β_0 is the predicted value of Y when all of the Xs equal 0. In terms of our example, it would be the predicted nAChr density for neonatal controls (age is 0) with no brain cotinine. (This prediction, however, is not sensible because it extends far beyond the age range of the observed data, See Section X.X).

The other βs are all associated with a variable. Because the variable is multiplied by the β, the β is a "weight" that determines how much the X contributes to prediction. If $\beta = 0$, then the associated variable does not predict individual differences in Y (once again, with the proviso that we are controlling for all the other variables). As an example, suppose that the β for age had been 0 in the nicotinic receptor data. Then if we picked a subject with a given diagnosis and cotinine value, then changing age would make no difference in the predicted nAChR level for that individual.

In more specific terms, a β gives the predicted change in Y for a one unit change in the X, keeping everything else constant. There is a very simple proof of this interpretation. Assume GLM equation of the form of Equation 9.3 and concentrate on the ith X. We can write this equation as

$$\hat{Y}_0 = \ldots + \beta_i X_i \tag{9.4}$$

where the ellipses $(\ldots)$ denote "everything else in the equation that is kept constant." Now change the value of $X_i$from X_i to $(X_i + 1)$. The predicted value is now

$$\hat{Y}_1 = \ldots + \beta_i (X_i + 1) \tag{9.5}$$

Subtracting Equation for $\hat{Y}_0$ from that for $\hat{Y}_1$ from X gives

$$\hat{Y}_1 - \hat{Y}_0 = \beta_i (X_i + 1) - \beta_i X_i = \beta_i$$

A β gives the predicted change in Y for a one unit increase in X.

Hence, the β for age (-.12) informs us that a one year increase in age is associated with a decrease of -.12 units of brain nAChR. The β for cotinine tells us that a one unit increase in cotinine predicts .08 units of increase in nAChR. Finally, an increase in one unit of diagnosis (in effect, a change from control to schizophrenia) predicts -5.7 units decrease in nAChR.

Note carefully that the actual magnitude of a βs is a function of the units of measurement of its X. Suppose X was measured in milligrams. The β would give the predicted change in $\hat{Y}$ for a one milligram increase in X. If we changed the scale of X from milligrams to micrograms, then the β in the new equation would give the change in $\hat{Y}$ from a one *micro*gram change in X. One can therefore arbitrarily make a β larger or smaller by simply changing the scale of its variable.

This scale property of β leads to one of the most important cautions in interpreting the results from a GLM: *never compare the* β*s across variables to determine the importance of the variables in prediction.* In our example, the β for a diagnosis of schizophrenia was -5.7 while the one for cotinine was .08. This does *NOT* imply that schizophrenia predicts nAChR much better than cotinine. Statistics other than the βs must be used to compare the effect sizes of the predictors.

> Never compare βs across variables to determine the importance of the variables in prediction.

9.4.1 Standardized betas

The type of betas (βs) that we have been dealing with are often called *raw* or *unstandardized* regression or GLM coefficients. These terms derive from the fact that the predictor variable are expressed in raw or unstandardized units. In some cases, it is helpful to examine *standardized* regression coefficients.

Suppose that we transformed the response variable, Y, to a new variable, Z_Y, with standard scores (see Section X.X). This means that the mean of Z_Y is 0 and the variance of Z_Y is 1.0. Suppose that we also standardized each of the predictor variables in the model to have means of 0 and standard deviations of 1. The GLM equation is

$$Z_{\hat{Y}} = \beta_0 + \beta_1 Z_{X_1} + \beta_2 Z_{X_2} + \ldots \; \beta_k Z_{X_k} \tag{9.6}$$

The βs in this equation are called *standardized* coefficients. They are the GLM coefficients from a model in which all variables have been standardized to have a mean of 0 and a standard deviation of 1.0.

> Standardized βs may be used to compare the relative predictive effects of the independent variables.

The interpretation of a standardized coefficient is the same as the one for a raw β but is expressed in terms of standard deviation units instead of raw units. Hence, if $\beta_1 = .09$, then we predict that a one standard deviation change in variable X_1 will result in a .09 standard deviation change in Y. Because all of the standardized predictor variables are the in the same units, standardized βs may be compared to assess the predictive effect of one variable versus another.

That is, if the standardized $\beta_1 = .12$ and standardized $\beta_2 = .07$, then X_1 is a better predictor of Y than X_2.

9.5 GLM and causality

It is essential to stress that even though we speak of “dependency”, “explanations” and “effects,” *causal interpretation* of a GLM depends on the design of the study. True experiments (i.e., direct experimental manipulation, random assignment, and strict control) permit inferences about causality. Given appropriate controls, if manipulation of variable A results in a change in the dependent variable, then in some way, shape or form–directly or indirectly–A has a causal influence on the response. How that causal influence comes about, whether the relationship is necessary and/or sufficient, and the mechanism(s) of causality cannot be answered by the statistical analysis of an experiment. Often, the answer to these questions depends on substantive issues coupled with the outcome of the experiment.

The smoking example is an excellent one for the discussion of causality. Cotinine predicts receptor density, but does it cause change in the number of receptors? Probably not. The most likely casual agent is nicotine. The nicotine up regulates receptors (**??**) and generates cotinine as a metabolite. Hence, cotinine is correlated with but has little causal effect on the number of receptors. Because of cotinine’s long half life (relative to nicotine), it works as a good control variable in the study.

Technically, a GLM applied to non-experimental observational research does not permit inferences about causality. But one must be reasonable here because interpretation of a GLM must be taken in the context of existing data and theory. There has never been, and never will be, a true experiment examining the health consequences of cigarette smoking in humans. It would be unethical–in fact, downright cruel–to randomly assign young adolescents to a smoking group and a non-smoking control group, compelling the former to smoke and the latter to abstain from cigarettes, until their health status could be ascertained 40 years later. Yet, all the observational, epidemiological data on humans agree so well with true experiments in animals and with mechanistic research into the cardiovascular and pulmonary effects of smoking that reasonable scientists infer a causal connection.