

1 Generalizations of the Regression Model (Soc 504)

As I said at the beginning of the semester, beyond the direct applicability of OLS regression to many research topics, one of the reasons that a full semester course on the linear model is warranted is that the linear model lays a foundation for understanding most other models used in sociology today. In these last sets of notes, I cover three basic generalizations of linear regression modeling that, taken as a whole, probably account for over 90% of the methods used in published research over the last few years. Specifically, we will discuss 1) generalized linear models, 2) multivariate models, and 3) time series and longitudinal methods. I will include discussions of fixed/random effects models in this process.

2 Generalized Linear Models

In sociological data, having a continuous outcome variable is rare. More often, we tend to have dichotomous, ordinal, or nominal level outcomes, or we have count data. In these cases, the standard linear model that we have been discussing all semester is inappropriate for several reasons. First, heteroscedasticity (and nonnormal errors) are guaranteed when the outcome is not continuous. Second, the linear model will often predict values that are impossible. For example, if the outcome is dichotomous, the linear model will predict scores that are less than 0 or greater than 1. Third, the functional form specified by the linear model will often be incorrect. For example, we should doubt that increases in a covariate will yield the same returns on the dependent variable at the extremes than would be obtained toward the middle.

2.1 Basic Setup of GLMs

Generalized linear models provide a way to handle these problems. The basic OLS model can be expressed as:

$$\begin{aligned} Y &\sim N(X\beta, \sigma_e) \\ Y &= X\beta + e \end{aligned}$$

Generalized linear models can be expressed as:

$$\begin{aligned} F(\mu) &= X\beta + e \\ E(Y) &= \mu \end{aligned}$$

That is, some function of the expected value of the expected value of Y is equal to the linear predictor with which we are already familiar. The function that relates $X\beta$ to μ is called the link function. The choice of link function determines the name of the model we are using. The most common GLMs used in sociology have the following link functions:

Link Function (F)	Model
μ	Linear Regression
$\ln\left(\frac{\mu}{1-\mu}\right)$	Logistic Regression
$\Phi^{-1}(\mu)$	Probit Regression
$\ln(\mu)$	Poisson Regression
$\ln(-\ln(1-\mu))$	Complementary Log-Log Regression

An alternate way of expressing this is in terms of probabilities. The logit and probit models are used to predict probabilities of observing a 1 on the outcome. Thus, we could write the model as:

$$p(y_i = 1) = F(X_i\beta).$$

In this notation, F is the link function. I will illustrate this with the probit regression model.

If our outcome variable is dichotomous, then the appropriate likelihood function for the data is the binomial distribution:

$$L(p | y) = p(y | p) \propto \prod_{i=1}^n p^{y_i} (1-p)^{1-y_i}$$

Our observed data constitute the y_i —if a person is a 1 on the dependent variable, then the second term in the likelihood drops out (for that individual); if a person is a 0, then the first term drops. We would like to link p to $X\beta$, but as discussed at the beginning, this is problematic because an identity link (i.e., $p = X\beta$) will predict illegitimate values for p . A class of functions that can map the predictor from the entire real line onto the interval $[0, 1]$ is cumulative distribution functions. So, for example, in the probit case, we allow $p = \Phi(X\beta)$, where Φ is the cumulative normal distribution function (i.e., $\int_{-\infty}^{X\beta} N(0, 1)$) Regardless of the value of $X\beta$, p will fall in the acceptable range. To obtain a logistic regression model, one would simply need to set $p = \frac{e^{X\beta}}{1+e^{X\beta}}$ (the cumulative logistic distribution function).

The approach discussed immediately above may seem different from what was presented in the table; however, the only difference is in how the link function is expressed—whether as a function of the expected value of Y , or in terms of the linear predictor. These are equivalent (just inverses of one another). For example, the logistic regression model could be written as:

$$\ln\left(\frac{\mu}{1-\mu}\right) = X\beta,$$

where $\mu = p$. Another way to think about GLMs is in terms of latent distributions. We could express the probit model as:

$$Y^* = X\beta + e,$$

using the link:

$$\begin{cases} Y = 1 & \text{iff } Y^* > 0 \\ Y = 0 & \text{iff } Y^* \leq 0 \end{cases}$$

Here, Y^* is a latent (unobserved) propensity. However, due to crude measurement, we only observe a dichotomous response. If the individual's latent propensity is strong enough, his/her propensity pushes him/her over the threshold (0), and we observe a 1. Otherwise, we observe a 0.

From this perspective, we need to rearrange the model somewhat to allow estimation. We can note that if $Y^* = X\beta + e$, then the expressions in the link equation above can be rewritten such that: If $Y = 1$ then $e > -X\beta$; if $Y = 0$ then $e < -X\beta$. If we assume a distribution for e (say normal), then we can say that:

$$p(Y = 1) = P(e > -X\beta) = P(e < X\beta) = \int_{-\infty}^{X\beta} N(0, 1).$$

Observe that this is the same expression we placed into the likelihood function above. If we assume a logistic distribution for the error, then we obtain the logistic regression model discussed above.

I will use this approach to motivate the generalization of the dichotomous probit model we've been discussing to the ordinal probit model. If our outcome variable is ordinal, rather than dichotomous, OLS is still inappropriate. If we assume once again that a latent variable Y^* underlies our observed ordinal measure, then we can expand the link equation above:

$$\begin{cases} Y = 1 & \text{iff} & -\infty = \tau_0 \leq Y^* < \tau_1 \\ Y = 2 & \text{iff} & \tau_1 \leq Y^* < \tau_2 \\ \vdots & & \vdots \\ Y = k & \text{iff} & \tau_{k-1} \leq Y^* < \tau_k = \infty \end{cases}$$

Just as before, this link, given a specification for the error term, implies an integral over the error distribution, but now the integral is bounded by the thresholds:

$$p(Y = j) = P(\tau_{j-1} - X\beta < e < \tau_j - X\beta) = \int_{\tau_{j-1} - X\beta}^{\tau_j - X\beta} N(0, 1).$$

2.2 Interpreting GLMs

GLMs are not as easy to interpret as the standard linear regression model. Because the link function is nonlinear, the model is now nonlinear, even though the predictor is linear. This complicates interpretation, because the effects of variables are no longer independent of the effects of other variables. That is, the effect of X_j depends on the effect of X_k . The probit model is linear in Z (standard normal) units. That is, given that $X\beta$ implies an increase in the upper limit of the integral of the standard normal distribution, each β can be viewed in terms of its effect on the Z score for the individual.

The logit model is linear in log-odds units. Recall that odds are computed as the ratio of $\frac{p}{1-p}$. The logistic link function, then, is a log-odds function. The coefficients from the model can be interpreted in terms of their linear effect on the log-odds, but this is not of much help. Instead, if we exponentiate the model, we obtain:

$$\exp\left(\ln\left(\frac{p}{1-p}\right)\right) = \exp(X\beta) = e^{\beta_0} e^{\beta_1 X_1} \dots e^{\beta_j X_j}$$

This says that the *odds* are equal to the multiple of the *exponentiated* coefficients. Suppose we had an exponentiated coefficient for gender (male) of 2. This would imply that the odds are twice as great for men as for women, net of the other variables in the model. The interpretation is slightly more complicated for a continuous variable, but can be stated as: the odds are multiplied by $\exp(\beta_j)$ for each unit increase in X_j . Be careful with this interpretation: saying the odds are multiplied by 2 does NOT mean that men are twice as likely to die as women. The word “likely” implies a ratio of probabilities and not odds.

The logistic regression model has become quite popular because of the odds-ratios interpretation. However, the unfortunate aspect of it is that this interpretation tells us nothing about the absolute risk (probability) of obtaining a ‘1’ response. In order to make this interpretation, we must compute the probabilities predicted by the model. It is in this process that we can see how the effect of one variable depends on the values of the other variables in the model. Below are a logistic regression and a probit regression of death on baseline age, gender (male), race (nonwhite), and education.

Variable	Logistic Reg. Parameter	$\text{Exp}(\beta)$	Probit Reg. Parameter
Intercept	-6.2107		-3.4661
Age	.1091	1.115	.0614
Male	.7694	2.158	.4401
Nonwhite	.5705	1.769	.3341
Education	-.0809	.922	-.0487

The results (for either model) indicated that age, being male, and being nonwhite increase one’s probability of death, while education reduces it. Although the coefficients differ between the two models, this is simply a function of the difference in the variances of the logistic and probit distributions. The variance of the probit distribution is 1 ($N(0, 1)$); the variance of the logistic distribution is $\frac{\pi^2}{3}$. The ratio of these variances ($\frac{L}{P}$) is 1.81, and this is also approximately the ratio of the coefficients—there is some slight deviation that is attributable to the slight differences in the shape of the distribution functions (the probit is steeper than the logit in the middle).

If we wanted to determine the difference in probability of mortality for a person with a high school diploma versus a college degree, we would need to fix the other covariates at some value, compute $X\beta$, and perform the appropriate transformation to invert the link function. Below are the predicted probabilities for 50 year-olds with different gender, race, and education profiles.

Profile			Probit	Logit
Sex	Race	Education	(Predicted probabilities)	
Male	White	12 yrs.	.29	.28
		16 yrs.	.23	.22
	Nonwhite	12 yrs.	.42	.40
		16 yrs.	.34	.33
Female	White	12 yrs.	.16	.15
		16 yrs.	.12	.11
	Nonwhite	12 yrs.	.26	.24
		16 yrs.	.20	.19

Notice that the estimated probabilities are very similar between the two models. For most data, the models can be used interchangeably. Notice also that the change in probability by altering one characteristic depends on the values of the other variables. For example, the difference in probability of death between 12 and 16 years of education is .06 for white males, .08 for nonwhite males, .04 for white females, and .06 for nonwhite females (all base on the probit model results). The odds ratio, however, does not vary. For example, take the odds ratio for white males with 12 versus 16 years of education ($OR = 1.38$) and the odds ratio for nonwhite males with 12 versus 16 years of education ($OR = 1.35$). The difference is only due to rounding of the probabilities.

3 Multivariate Models

For the entire semester, we have discussed univariate models—that is, models with a single outcome. Often, however, we may be interested in estimating models that have multiple dependent variables. Let's take a very simple model first.

Modeled

$$\begin{aligned}
 y_1 &= X\beta + e_1 \\
 y_2 &= Z\gamma + e_2 \\
 e_1 &\sim N(0, \sigma_{e_1}^2) \\
 e_2 &\sim N(0, \sigma_{e_2}^2)
 \end{aligned}$$

Not modeled, but true

$$\begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{e_1}^2 & \sigma_{e_1 e_2} \\ \sigma_{e_1 e_2} & \sigma_{e_2}^2 \end{bmatrix} \right)$$

This model says that outcomes y_1 and y_2 are functions of several covariates (X and Z) plus some error. The third and fourth components indicate that e_1 and e_2 are assumed to be uncorrelated. However, in fact, the last expression indicates that the errors across equations are correlated (as long as $\sigma_{e_1e_2}$ is nonzero). This model is sometimes called the “seemingly unrelated regression model.” The regressions seem as though they could be estimated independently, but if e_1 and e_2 are correlated, then it implies that there are variables (namely y_2 and possibly some Z) that are omitted from the model for y_1 (and vice versa for the model for y_2). Omitting relevant variables leads to ‘omitted variable bias,’ which means that our estimates for coefficients are incorrect. We could rewrite the model, either specifically incorporating the error covariance portion of the model, or specifying a joint distribution for y :

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim N \left(\begin{bmatrix} X\beta \\ Z\gamma \end{bmatrix}, \begin{bmatrix} \sigma_{e_1}^2 & \sigma_{e_1e_2} \\ \sigma_{e_1e_2} & \sigma_{e_2}^2 \end{bmatrix} \right)$$

This model is the same as the model above, just expressed explicitly as a multivariate model.

3.1 Multivariate Regression

So far, we have dealt with a number of univariate distributions, especially the univariate normal distribution:

$$f(Y) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{(Y - \mu)^2}{2\sigma^2} \right\}$$

The multivariate normal distribution is simply an extension of this:

$$f(Y) = 2\pi^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} \exp \left\{ [Y - \mu]^T \Sigma^{-1} [Y - \mu] \right\}$$

Here, μ is a vector of means, and Σ is the covariance matrix of Y . If the matrix is diagonal, then the distribution could be rewritten simply a set of univariate normal distributions; when Σ has off-diagonal elements, it indicates that there is some (linear) relationship between variables.

Just as with linear regression, we can assume the errors are (multivariately) normally distributed, allowing us to place $(Y - X\beta)$ into the numerator of the kernel of the density for maximum likelihood estimation of β . In the multivariate case, each element of the $[Y - \mu]$ vector is replaced with $[Y - X\beta(j)]$, where I’m using (j) to index the set of β parameters in each dimension of the model (i.e., X does not have to have the same effect on all outcomes). We can assume the X matrix is the same across equations in the model, and if an X does not influence one of the outcomes, then its β parameter is constrained to be 0.

3.2 Path Analysis

Some multivariate models can be called ‘path analysis’ models. The requirements for path analysis include that the variables must all be continuous, and the model must be recursive—that is, following the paths through the variables, one cannot revisit a variable. Below is an example of a path-analytic graph.

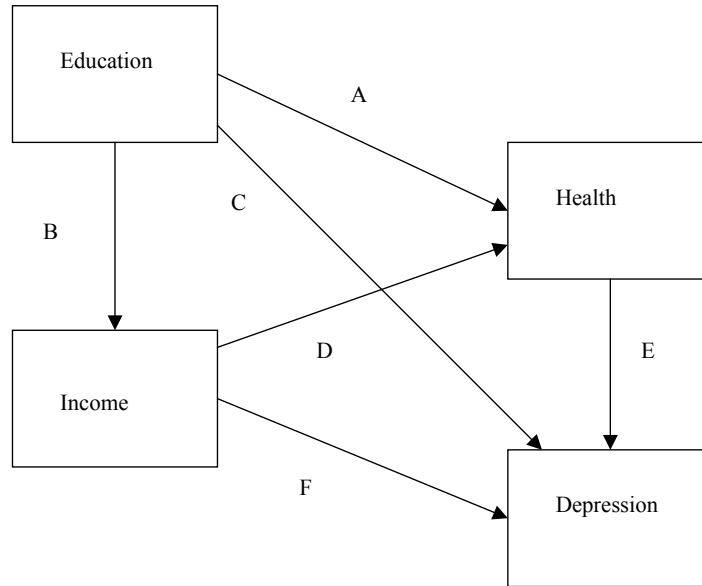


Figure: Path Diagram.

This path model says that depression is affected by physical health, education, and income; health is influenced by education and income; and income is influenced by education. If we estimated the model: $Depression = b_0 + b_1 Education$, we would find the total effect of education on depression. However, it is unlikely that education's effect is only direct. It is more reasonable that income and physical health also affect depression, and that education has direct effects (and possibly indirect effects) on income and health. Thus, the simple model would produce a biased effect of education if income and health were ignored. If we estimate the path model above, the coefficient for the direct effect of education (c) would most likely be reduced.

The path model above can be estimated using a series of univariate regression models:

$$\begin{aligned}
 Depression &= \beta_0 + \beta_1 education + \beta_2 health + \beta_3 income \\
 Health &= \gamma_0 + \gamma_1 education + \gamma_2 income \\
 Income &= \alpha_0 + \alpha_1 education
 \end{aligned}$$

The lettered paths in the diagram can be replaced as: $A = \gamma_1$, $B = \alpha_1$, $C = \beta_1$, $D = \gamma_2$, $E = \beta_2$, $F = \beta_3$.

Now the direct effect of education on depression is no longer equal to the total effect. Rather, the direct effect is simply C, while the total effect is:

$$\begin{aligned}
 Total &= Direct + Indirect \\
 &= c + (ae) + (bf) + (bde)
 \end{aligned}$$

As these expressions indicate, the indirect effects are simply the multiples of the paths that lead indirectly to depression from education.

3.3 Structural Equation Models

When nonrecursive (i.e., reciprocal effects are included) models are needed, variables are not measured on a continuous scale, and/or measurement error is to be considered, we can generalize the path analysis model above. Structural equation models provide a generalization. These models are often also called LISREL models (after the first software package to estimate them) and covariance structure models (because estimation is based on the covariance matrix of the data). Using LISREL notation, these models consist of 3 basic equations and 4 matrices:

$$\begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_j \end{bmatrix} = \begin{bmatrix} 0 & \beta_{12} & \dots & \beta_{1j} \\ \beta_{21} & 0 & \dots & \beta_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{j1} & \beta_{j2} & \dots & 0 \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_j \end{bmatrix} + \begin{bmatrix} \gamma_{11} & \gamma_{12} & \dots & \gamma_{1k} \\ \gamma_{21} & \gamma_{22} & \dots & \gamma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{j1} & \gamma_{j2} & \dots & \gamma_{jk} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_k \end{bmatrix} + \begin{bmatrix} \zeta_1 \\ \zeta_2 \\ \vdots \\ \zeta_j \end{bmatrix}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix} = \begin{bmatrix} \lambda_{y11} & \lambda_{y12} & \dots & \lambda_{y1j} \\ \lambda_{y21} & \lambda_{y22} & \dots & \lambda_{y2j} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{yp1} & \lambda_{yp2} & \dots & \lambda_{ypj} \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_j \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_p \end{bmatrix}$$

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} = \begin{bmatrix} \lambda_{x11} & \lambda_{x12} & \dots & \lambda_{x1k} \\ \lambda_{x21} & \lambda_{x22} & \dots & \lambda_{x2k} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{xp1} & \lambda_{xp2} & \dots & \lambda_{xpq} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_k \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_q \end{bmatrix}$$

- Φ = k-by-k covariance matrix of the ξ
- Ψ = j-by-j covariance matrix of the ζ
- θ_δ = q-by-q covariance matrix of the δ
- θ_ϵ = p-by-p covariance matrix of the ϵ

In this model, latent (unobserved) variables are represented by the Greek symbols η and ξ . The distinction between η and ξ is whether the variable is exogenous (ξ) or endogenous (η) in the model, where ‘endogenous’ means the variable is influenced by other variables in the model. The coefficients that relate the η to each other are β , while the coefficients that relate the ξ to the η are γ . The first equation is the ‘structural equation’ that relates the latent variables, with an error term ζ for each η . The second and third equations are measurement equations that show how the observed y and x are related to the latent variables η and ξ , respectively (via the λ coefficients). In these equations, ϵ and δ represent measurement errors—that is, the part of the observed variable that is unaccounted-for by the latent variable(s) which influence it.

The Φ matrix models the covariances of the exogenous latent variables, while the Ψ matrix models the covariances of the structural equation errors (allowing cross-equation error correlation to exist, which is something univariate regressions do not allow.) The two θ matrices allow correlation between errors in the measurement equations to exist (again, something that univariate regression cannot handle.)

This model is very general. If there is only one outcome variable (η), and all variables are assumed to be measured without error, then the model reduces to OLS regression. If we are uninterested in structural relations, but are only interested in the measurement portion of the model (and possibly in estimating simple correlations between latent variables), then we have a (confirmatory) factor analysis.

The model is estimated by recognizing that (1) the parameters are functions of the covariances (or correlations) of the variables and (2) a multivariate normal likelihood can be written in terms of these covariances. When some of the data are not continuous, but rather are ordinal, we can estimate something called ‘polychoric’ and ‘polyserial’ correlations between the observed variables, and the resulting correlation matrix can be used as the basis for estimation. The resulting model could then be called a ‘multivariate generalized linear model.’

Below is a graphic depiction of a relatively simple structural equation model. The equations for this SEM would be:

$$\eta_1 = \gamma_{11} + \zeta_1$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} \lambda_{y1} \\ \lambda_{y2} \\ \lambda_{y3} \end{bmatrix} \eta_1 + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} \lambda_{x1} \\ \lambda_{x2} \\ \lambda_{x3} \end{bmatrix} \xi_1 + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix}$$

$$\Phi = \phi_{11}, \Psi = \psi_{11}$$

$$\begin{bmatrix} \theta_{\delta_{11}} & 0 & 0 \\ 0 & \theta_{\delta_{22}} & 0 \\ 0 & 0 & \theta_{\delta_{33}} \end{bmatrix}$$

$$\begin{bmatrix} \theta_{\epsilon_{11}} & 0 & 0 \\ 0 & \theta_{\epsilon_{22}} & \theta_{\epsilon_{23}} \\ 0 & \theta_{\epsilon_{32}} & \theta_{\epsilon_{33}} \end{bmatrix}$$

Notice how most of the off-diagonal elements of the various covariance matrices are 0-this is because we have only specified 1 error correlation (between ϵ_2 and ϵ_3). The top and bottom portions of the figure constitute confirmatory factor analyses-the idea is that the observed x and y variables reflect underlying and imperfectly measured constructs (factors). We are really interested in examining the relationship between these constructs, but there is measurement error in our measures for them. Thus, with this model, γ_{11} is our estimate of

the relationship between the latent variables independent of any measurement error existing in our measures.

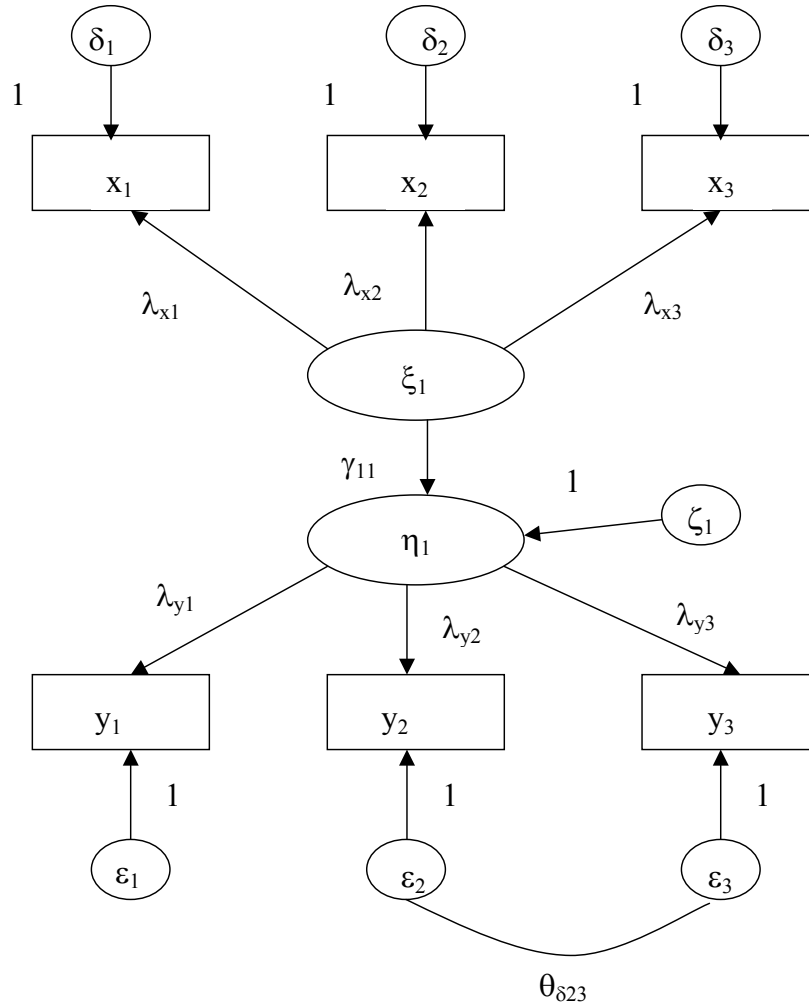


Figure: SEM Diagram.

3.4 Final note about multivariate models

Something you must remember with multivariate (and latent variable) models that isn't typically an issue in simple univariate models is the notion of *identification* or *identifiability*. We cannot estimate more parameters than we have pieces of information. In SEMs, the pieces of information are covariances and variances between variables—there are $\frac{(p+q)(p+q+1)}{2}$ of them. We need to be sure that we are not attempting to estimate too many parameters—if we are, then our model is 'under-identified,' which means loosely that there is no unique solution set for the parameters. The models we have dealt with to date have been 'just-identified,' meaning that there are exactly as many pieces of information as parameters. With SEMs, we often have 'over-identified' models, which means we have more than enough

information to estimate the model. This comes in handy when we would like to compare models.

4 Time Series and Longitudinal Methods

Analysis of time series and panel data is a very large area of methodology. Indeed, you can take entire courses on basic time series, event history models, or other methods of analysis for longitudinal data. In this brief introduction to these classes of models, I am exchanging depth for breadth in an attempt to point you to particular types of models that you may need to investigate further in doing empirical work.

I'll start with some basic terminology that's relevant to longitudinal data. First, the term 'longitudinal data' is somewhat vague. Generally, the term implies that one has panel data, that is, data collected on multiple units across multiple points in time (like the PSID). However, it is often also used to refer to repeated cross-sectional data, that is, data collected on multiple *different* units at multiple points in time (like the GSS). For the purposes of our discussion here, we will limit our usage to the first definition.

- Time Series typically refers to repeated observations on a single observational unit across time.
- Multiple Time Series means multiple observational units observed across time. This can also be considered a panel study, although the usage often differs depending on the unit of analysis. Generally, micro data is considered panel data, while macro data is considered time series data.
- Multivariate Time Series means that there are multiple outcomes measured over time.
- A Trend is a general pattern in time series data over time. For example, U.S. life expectancy shows an increasing trend over the last 100 years.
- Seasonality is a repeated pattern in time series data. For example, the sale of Christmas trees evidences considerable seasonality—sales are low except in November and December. No trend is necessarily implied, but the pattern repeats itself annually.
- Stationarity. A stationary time series evidences no trending. A stationary series may have seasonality, however. Stationarity is important for time series models, for reasons that will be discussed shortly.

4.1 Problems that Time Series/Longitudinal Data Present

There are really very few differences between the approaches that are used to analyze time series or panel data and the basic linear regression model with which you are already familiar. However, there are four basic problems that such data present that necessitate the expansion of the OLS model or the development of alternative models.

1. Error correlation within units across time. This problem requires alternative estimation of the linear model (e.g., GLS estimation), or the development of a different model (e.g., fixed/random effects models).
2. Spuriousness due to trending. When attempting to match two time series, it may appear that two aggregate time series with similar trends are causally related, when in fact, they aren't. For example, population size (which has an increasing trend) may appear to be related to increases in life expectancy (also with an increasing trend). However, this is probably not a causal relationship, because, in fact, countries with the fastest growth in population don't necessarily have the fastest increases (if any at all) in life expectancy. This may be an ecological fallacy or it may simply be that two similarly-trending time series aren't related.
3. Few units. Sometimes, time series data are relatively sparse. When dealing with a single time series, for example, we often have relatively few measurements. This makes it difficult to observe a trend and to include covariates/explanatory variables in models.
4. Censoring. Although our power to examine relationships is enhanced with time series and panel data, such data present their own problems in terms of "missing" data. What do we do with people who die or can't be traced for subsequent waves of study? What do we do with people who are missing at some waves but not others? What do we do when a person experiences an event that takes him/her out of the risk set? Etc.

4.2 Time Series Methods

Problems (1) and (2) above can often be resolved by placing some structure on the error term. There are many ways to do this, and such forms the basis for a smorgasboard of approaches to analyzing time series data.

The most basic models for time series are ones in which trends and seasonality are modeled by including time as a variable (or a collection thereof, e.g., like time dummies) in a model. For example, if we wanted to examine a trend in birth rates across time, and we observed birth rates on a monthly basis across, say, a 20 year span, we could first model birth rates as a function of time (years), examine the residual, and then include dummy variables (or some other type of variables) to capture seasonality. The key problem with this approach is that we must make sure that autocorrelation of the errors does not remain after detrending and deseasoning.

There are two basic domains of more complicated time series analysis: the time domain and the frequency domain. In frequency domain models, measures are seen as being a mixture of sine and cosine curves at different frequencies:

$$y_i \sim N((X\beta)_i + \sum_{j=1}^J (a_j \sin(w_j t_i) + b_j \cos(w_j t_i)), \sigma^2)$$

These models are often called 'spectral models.' I will not discuss these models, because a) they are not particularly common in sociology and demography and b) these methods are

ultimately related to time domain models (i.e., they are simply an alternate parameterization of time domain models).

More common in sociology are time domain models. Time domain models model a time t variable as a function of the outcome at time $t - 1$. In economics, these are called ‘dynamic models.’ A general class of models for time domain time series models are ARMA (AutoRegressive Moving Average) models, which can be represented as:

$$y_t = X_t\beta + \phi_1y_{t-1} + \phi_2y_{t-2} + \dots + \phi_my_{t-p} + e_t + \gamma_1e_{t-1} + \gamma_2e_{t-2} + \dots + \gamma_ny_{t-q}$$

In this equation, the autoregressive terms are the lagged y -values, while the moving average terms are the lagged errors. e_t is assumed to be $N(0, \sigma^2I)$ under this model. As specified, this model would be called an ARMA(p,q) model (although, technically, the classic ARMA model does not contain regressors). Very often, we do not need more than 1 AR term or one MA term to achieve stationarity of the error. (As a side note, a time series process in which only y_{t-1} is needed is called a Markov process). The model requires that ϕ be less than 1, or the model is considered ‘explosive.’ (this can be observed by repeated substitution).

To put the ARMA(p,q) model into perspective, we can view this model as placing structure on the error term in an OLS model. Recall from previous discussions that an assumption of the OLS regression model is that $e \sim N(0, \sigma_e I)$. In other words, the errors for observations i and j are uncorrelated for $i \neq j$. This assumption is violated by time series data, and so a natural extension of the OLS model is to use GLS estimation. GLS estimation involves estimating the parameters of the model with a modified estimator: $\hat{\beta}_{GLS} = (X^T \Sigma^{-1} X)^{-1} (X^T \Sigma^{-1} Y)$. Σ , however, is an $n \times n$ matrix, and all elements of this matrix cannot be estimated without some simplifying constraints. One such constraint that can be imposed is that the error correlations only exist between errors for adjacent time periods, and that all adjacent time periods have equal error correlation. In that case, we only need to estimate one additional parameter, say $\sigma_{i,i+1}$, $\forall i$, so our Σ matrix appears as:

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & 0 & \dots & 0 \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \ddots & \vdots \\ 0 & \sigma_{32} & \sigma_{33} & \sigma_{34} & 0 \\ \vdots & \ddots & \sigma_{43} & \sigma_{44} & \ddots \\ 0 & \dots & 0 & \ddots & \ddots \end{bmatrix},$$

where $\sigma_{11} = \sigma_{22} = \dots = \sigma_{TT}$, and $\sigma_{12} = \sigma_{21} = \sigma_{23} = \sigma_{32} = \dots$

An AR(1) model simplifies the error covariance matrix for GLS estimation by decomposing the original OLS error term into two components: $e_t = \rho e_{t-1} + v_t$. Here the error at one time point is simply a function of the error at the previous time point plus a random shock at time $t(v)$. Higher order AR models are obtained by allowing the error to be a function of errors at lags > 1 . Typically, autoregressive models are estimated by incorporating a lagged y variable into the model. So long as the absolute value of the coefficient for the lagged term(s) does not exceed 1, the series can be considered stationary.

An MA(1) model specifies structure on the random shocks: $e_t = \sigma v_{t-1} + v_t$. As with the AR models, higher order MA models can be obtained by adding additional lagged terms.

Moving average models are more difficult to estimate than autoregressive models, however, because the error term depends on the coefficients in the model, which, in turn, depend on the error.

How do we determine what type of ARMA model we need? Typically, before we model the data, we first construct an autocorrelation plot (also sometimes called a corellogram), which is a plot of the autocorrelation of the data (or errors, if a model was previously specified) at various lags. The function is computed as:

$$AC_L = \frac{m \sum (\theta_t - \bar{\theta})(\theta_{t-L} - \bar{\theta})}{(m - L) \sum (\theta_t - \bar{\theta})^2}$$

where m is the number of time series data points and L is the number of lags. The shape of this function across L tells us what type of model we may need, as we will discuss below.

When autocorrelation between errors cannot be removed with an ARMA(p,q) model, the next step may be to employ an ARiMA (Autoregressive integrated moving average) model. The most basic ARiMA model is one in which there are no autoregressive terms and no moving average terms, but the data are differenced once. That is, we simply take the difference in all variables from time $t - 1$ to t , so that $y_{diff} = y_t - y_{t-1}, \forall t$. We do the same for all the covariates. We then regress (using OLS) the differences in y on the differences in x . This model therefore ultimately relates change in x to change in y (this is why the term ‘integrated’ is used—from a calculus perspective, relating change to change is matching the first derivatives of x and y , thus, the original variable is ‘integrated’ relative to the differences).

Sometimes, a first differences approach is not sufficient to remove autocorrelation. In those cases, we may need to add autoregressive or moving average components, or we may even need to take second or higher order differences.

5 Methods for Panel Data

The time series methods just discussed are commonly used in economics, but are used somewhat less often in sociology. Over the last twenty years, panel data have become quite common, and sociologists have begun to use methods appropriate for multiple time series/panel data. In many cases, the methods that I will discuss here are very tightly related to the time series methods discussed above; however, they are generally presented differently than the econometric approach to time series. In this section, I will discuss two general types of panel methods: hazard/event history models and fixed/random effects hierarchical models (including growth models). The key feature that distinguishes these models is whether your outcome variable is an event that can occur versus simply repeated measures on individuals over time.

5.1 Hazard and Event History Models

Hazard models is another class of models that go by several names. One is “event history models.” In demography, “life table methods” accomplish the same goals. Other names

include “discrete time hazard models” and “continuous time hazard models.” Related approaches are called “survival models,” and “failure time models.” These related approaches are so-called because they model the survival probabilities ($S(t)$) and the distribution of times-to-event ($f(t)$), whereas hazard models model the hazard of an event. Hazard models are distinct from time series models, because time series models generally model an outcome variable that takes different values across time, but hazard models model the occurrence of a discrete event. If the time units in which the respondents are measured for the event are discrete, we can use ‘discrete time event history methods;’ if the time units are continuous (or very closely so), we can use ‘continuous time event history methods.’

A hazard is a probability of experiencing an event within a specified period of time, conditional on having survived to be at risk during the time interval. Mathematically, the hazard can be represented as:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{p(E(t, t + \Delta t) | S(t))}{\Delta t}$$

Here, $p(E(t, t + \Delta t))$ represents the probability of experiencing the event between time t and $t + \Delta t$, $S(t)$ indicates the individual survived to time t , and Δt is an infinitely small period of time. The hazard, unlike a true probability, is not bounded between 0 and 1, but rather has no upper bound.

If we examine the hazard a little further, we find that the hazard can be viewed as

$$h(t) = \frac{\# \text{ who experience event during the interval}}{t \times \# \text{ exposed in the interval}}$$

The numerator represents the number of persons experiencing the event, while the denominator is a measure of person-time-units of exposure. Given that individuals can experience the event at any time during the interval, each individual who experiences the event can only contribute as many time units of exposure as s/he existed in the interval. For example, if the time interval is one year, and an individual experiences the event in the middle of the interval, his/her contribution to the numerator is 1, while their contribution to the denominator is .5.

In a hazard model, the outcome to be modeled is the hazard. If the time intervals are sufficiently small (e.g, minutes, seconds, etc.), then we may use a “continuous time” hazard model; if the intervals are sufficiently large, then we may use a “discrete time” hazard model. There is no clear break point at which one should prefer a discrete time model to a continuous time model. As time intervals become smaller, discrete time methods converge on continuous time methods.

In this discussion, I will focus on hazard models rather than survival or failure time models, although all three functions are related. For example, the hazard ($h(t)$) is equal to:

$$h(t) = \frac{f(t)}{S(T) = 1 - F(t)},$$

the density function (indicating probability of the event at time t) conditional on (dividing by) survival up to that point. Notice that the survival function is represented as $1 - F(t)$, where $F(t)$ is the integral of the density function. The density function gives the probabilities of experiencing the event at each time t . So, if we want to know the probability that a person

will experience the event by time t , we simply need to know the area under the density function from $-\infty$ to t , which is $\int_{-\infty}^t uf(u)du = F(t)$. The survival probability beyond t , therefore, is $1 - F(t)$.

We will focus on hazard models because they are more common in sociology and demography. Survival analysis, beyond simple plotting techniques, is more common in clinical and biological research.

5.1.1 A Demographic Approach: Life Tables

The earliest type of hazard model developed was the life table. The life table generally takes as its input the hazard rate at each time (generally age), uses some assumptions to translate the hazard rate into transition (death) probabilities, and uses a set of flow equations to apply these probabilities to a hypothetical population to ultimately end up with a measure of the time remaining for a person at age a . A basic table might look like:

Table: Single Decrement Life Table

Age	l_a	$h() = \mu()$	q_a	d_a	L_a	e_a
20	$l_{20} = 100,000$	μ_{20}	q_{20}	$q_{20} \times l_{20}$	$\frac{l_{20}+l_{21}}{2}$	$\frac{\sum_{a=20}^{\omega} L_a}{l_{20}}$
21	$l_{21} = l_{20} - d_{20}$	μ_{21}	q_{21}	$q_{21} \times l_{21}$	$\frac{l_{21}+l_{22}}{2}$	$\frac{\sum_{a=21}^{\omega} L_a}{l_{21}}$
\vdots	\vdots	\vdots	\vdots	\vdots		

The columns in the life table are the following. The l column represents the number of individuals remaining alive in a hypothetical cohort (radix=100,000) at the start of the interval. The μ column represents the hazard for the time interval. The q column is the probability that an individual alive at the beginning of the interval will die before the end of the interval. An assumption is used to transform μ into q . For example, if we assume that individuals die, on average, in the middle of the interval, then the exposure (in person years) is simply the average of the number of individuals alive at the beginning and the end of the interval, and so:

$$\mu = \frac{q \times l}{\frac{1}{2}(l + (1 - q) \times l)}$$

Some rearranging yields:

$$q = \frac{\mu \times l}{l + \frac{1}{2}\mu l}$$

So, we can obtain the transition probabilities from the hazards. We then apply these probabilities to the population surviving to the beginning of the interval to obtain the count of deaths in the interval (d). We then have enough information to calculate the next l , and we can proceed through the table like this. When we are done, we then construct the L column, which is a count of the person years lived in the interval. Once again, if we assume individuals died in the middle of the interval, then the person years lived is simply the average of

the number of persons alive at the beginning and end of the interval (the denominator of the equation for μ). The next column, T , sums these person years lived from this interval through the end of the table. Finally, the last column, e , divides the number of individuals alive at the start of an interval (l) by the cumulative person years remaining to be lived to obtain an average of the number of years remaining for each person (life expectancy). The value at the earliest age in the table is an approximation of $\int_{-\infty}^{+\infty} t \times f(t)dt$, the expectation of the distribution of failure times.

The life table has been extended to handle multiple decrements (multiple types of exits), as well as reverse and repeatable transitions (the multistate life table). A key limitation of the life table has been the difficulty with which covariates can be incorporated in it and the difficulty in performing statistical tests for comparing groups. For this reason, researchers began using hazard regression models.

5.1.2 Continuous Time Hazard Models

The most basic continuous time hazard models include the exponential model (also called the constant hazard model), the Gompertz model, and the Weibull model. The difference between these models is their representation of the “baseline hazard,” which is the *hazard function when all covariate values are 0*. I emphasize this definition, because some may be misled by the name into thinking that the baseline hazard is the hazard at time $t = 0$, and that is not the case.

Suppose for a minute that there are no covariates, and that we assume the hazard does not change over time. In that case, the exponential model is:

$$\ln(h(t)) = a$$

The hazard is logged, because of the bounding at 0. The name “exponential” model stems from 1) the fact that if we exponentiate each side of the equation, the hazard is an exponential function of the constant, and 2) the density function for failure times that corresponds to this hazard is the exponential distribution.

If we assume that the hazard is constant across time, but that different subpopulations have different values of this constant, the exponential model is:

$$\ln(h(t)) = a + X\beta$$

In this specification, a is the baseline hazard, which is time-independent, and $X\beta$ is the linear combination of covariates thought to raise or lower the hazard.

Generally, the assumption of a constant hazard is unreasonable: instead, the hazard is often assumed to increase across time. The most common model representing such time-dependent hazards is the Gompertz model, which says that the log of the hazard increases linearly with time:

$$\ln(h(t)) = X\beta + bt.$$

In demography, we often see this model as:

$$h(t) = \alpha \exp(bt),$$

with $\alpha = \exp(X\beta)$.

Often, we do not think the log of the hazard increases linearly with time, but rather we believe it increases more slowly. Thus, the Weibull model is:

$$\ln(h(t)) = X\beta + b \times \ln(t).$$

Each of these models is quite common in practice. However, sometimes, we do not believe that any of these specifications for the baseline hazard is appropriate. In those cases, we can construct “piecewise” models that break the baseline hazard into intervals in which the hazard may vary in its form.

A special and very common model in which the baseline hazard remains completely unspecified, while covariate effects can still be estimated, is the Cox proportional hazard model. Cox’s great insight was that the likelihood function for a hazard model can be factored into a baseline hazard and the portion that contains the covariate effects. The Cox model looks like:

$$\ln(h(t)) = g(t) + X\beta$$

where $g(t)$ is an unspecified baseline hazard. Estimation of this model ultimately rests on the ordering of the event times. Thus, a problem exists whenever there are lots of “ties” in the data. This method, therefore, is generally most appropriate when the time intervals in the data are very small, and hence the probability of ties is minimal.

Final note on continuous time methods: Realize that the hazard, being an instantaneous probability, is ultimately always unobserved. Thus, estimation of these models thus requires special software/procedures and cannot be estimated with OLS or other standard techniques.

5.1.3 Discrete Time Models

When time intervals are discrete, we may use discrete time models. The most common discrete time method is the discrete time logit model. The discrete time logit model is the same logit model that we have already discussed. The only difference in application is the data structure to which the model is applied. The logit model is represented as:

$$\ln\left(\frac{p}{1-p}\right) = X\beta$$

where p is the probability that the event occurs to the individual in the discrete time interval. Construction of the data set for estimation involves treating each individual as multiple person-time records in which an individual’s outcome is coded ‘0’ for the time intervals prior to the occurrence of the event, and is coded ‘1’ for the time interval in which the event does occur. Individuals who do not experience the event over the course of the study are said to be censored, but they do not pose a problem for the analyses: they are simply coded ‘0’ on the outcome for all time intervals. To visualize the structure of the data, the first table shows 10 hypothetical individuals.

Standard format for data

ID	Time until event	Experienced event?
1	3	1
2	5	0
3	1	1
4	1	1
5	2	1
6	4	1
7	4	1
8	3	1
9	1	0
10	2	1

The study ran for 5 time intervals. Persons 2 and 9 did not experience the event and thus are censored. Person 2 is censored simply because s/he did not experience the event before the end of the study. Person 9 is censored due to attrition. The new data structure is shown in the second table.

Person-year format for data

Record	ID	Time Interval	Experienced event?
1	1	1	0
2	1	2	0
3	1	3	1
4	2	1	0
5	2	2	0
6	2	3	0
7	2	4	0
8	2	5	0
9	3	1	1
10	4	1	1
11	5	1	0
12	5	2	1
13	6	1	0
14	6	2	0
15	6	3	0
16	6	4	0
17	7	1	0
18	7	2	0
19	7	3	0
20	7	4	1
21	8	1	0
22	8	2	0
23	8	3	1
24	9	1	0
25	10	1	0
26	10	2	1

Now we have 26 records rather than the 10 in the original data set. We can compute the hazard, by observing the proportion of persons at risk who experience the event during each time period. The hazards are: $h(1) = \frac{2}{10}$, $h(2) = \frac{2}{7}$, $h(3) = \frac{2}{5}$, $h(4) = \frac{1}{3}$, $h(5) = \frac{0}{1}$. Now, when we run our logit model, the outcome variable is the logit of the hazard ($\ln \frac{h(t)}{1-h(t)}$). We can include time-varying covariates very easily, by simply recording the value of the variable for the respondent record for which it applies. As you can see, censoring is also handled very easily. We can also specify whatever form we would like for the baseline hazard. If we want the baseline hazard to be constant, we simply don't include a variable or function for time in the model. If we want complete flexibility-a completely piecewise linear model-we would simply include a dummy variable for every time interval (except one).

This model is very similar to the ones we have already discussed. As the time intervals get

smaller, the model converges on the continuous time hazard models. Which one it converges to is simply a matter of how we specify time in the model. For example, if we construct a variable equal to $\ln(t)$ and enter it into the model, we essentially have the Weibull model. If we just enter t as a variable, we have a Gompertz model.

The interpretation of the model is identical to that of the standard logit model, the only difference being that the outcome is the hazard rather than the probability.

5.2 Fixed/Random Effects Hierarchical Models

Hierarchical modeling is an approach to analyzing data that takes advantage of, and/or compensates for, nesting structure in data. The approach is used to compensate for multi-stage sampling, which induces dependence among errors and leads to biased standard errors. The approach is used to take advantage of hierarchical structuring of data by distinguishing between effects that occur at multiple levels. For example, the approach can be used to differentiate between within-individual change over time and between-individual heterogeneity. The approach can be used to distinguish between family-level effects and neighborhood-level effects on individual outcomes, etc. Thus, Fixed/Random Effects Hierarchical models are not exclusively used for panel data, but can be when the nesting structure for the data is individuals (level 2) measured across time (level 1).

As alluded to above, hierarchical modeling is a quite flexible and general approach to modeling data that are collected at multiple levels. Because of this flexibility and wide applicability, hierarchical modeling has been called many things. Here is a list of some of the terms that have been used in referring to these types of models:

- Hierarchical modeling
- Multilevel modeling
- Contextual effects modeling
- Random coefficient models
- Random/Fixed effects models
- Random intercept models
- Random effects models with random intercepts and slopes
- Growth curve modeling
- Latent curve analysis
- 2-level models, 3-level models, etc.
- Variance component models
- Mixed (effects) models
- Random effects ANOVA

This list is not exhaustive, but covers the most common labels applied to this type of modeling. In this brief discussion, I am not going to give an in-depth mathematical treatment of these models; instead, I will try to show how these names have arisen in different empirical research contexts, but are all part of the general hierarchical model.

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 z_j + e_{ij}$$

Here, Y_{ij} is the individual-level outcome for individual i in group j , β_0 is the intercept, β_1 is the effect of individual-level variable x_{ij} , β_2 is the effect of group-level variable z_j , and e_{ij} is a random disturbance.

If there is clustering within groups-e.g, you have all family members in a family, or you have repeated measures on an individual over time-this model is not appropriate, because e_{ij} are not independent (violating the OLS assumption that $e \sim N(0, \sigma^2 I)$).

Two simple solutions to this dilemma are 1) to pull out the structure in the error term (similar to ARMA models) by decomposing it into a group effect and truly random error and 2) to separate the intercept into two components: a grand mean and a group mean. The former approach leads to a random effects model; the latter a fixed effects model, but they look the same:

$$Y_{ij} = \beta_{00} + \beta_1 x_{ij} + \gamma_j + e_{ij}$$

Here, the subscript on the intercept has changed to denote the difference between this and the OLS intercept. γ_j denotes either a fixed effect (the decomposition of the intercept term) or random effect (the decomposition of the error term). I have eliminated z , because in a fixed effects approach, all fixed characteristics are not identifiable apart from the intercept.

If we treat the model as a random effects model, this model can be called a random intercept model. Note that there are two levels of variance in this specification-true within-individual variance (denoted σ_e^2) and between-individual (level 2) variance (denoted τ^2), the variance of the random effects. The total variance can be computed as: $\sigma_e^2 + \tau^2$.

We now have a model specification that breaks variance across two levels, and we can begin to bring in variables to explain variance at both levels. Suppose we allow the random intercept γ_j to be a function of group level covariates and residual group-specific random effects u_j , so that $\gamma_j = \gamma_0 + \gamma_1 z_j + u_j$. Then, substitution yields:

$$Y_{ij} = \beta_1 x_{ij} + (\gamma_0 + \gamma_1 z_j + u_j) + e_{ij}$$

Notice that I have eliminated the original intercept, β_{00} , as it is no longer identified as a parameter distinct from γ_0 , the new intercept after adjustment for group level differences contained in z_j . Now, every group has a unique intercept that is a function of a grand mean (γ_0), a fixed effect of a group-level variable (γ_1), and a random, unexplained component (u_j). τ^2 should shrink as group-level variables are added to account for structure in u , and measures of second-level model fit can be constructed from this information. Similarly, the addition of more individual-level measures (x_{ij}) should reduce σ_e^2 , and first-level model fit can be constructed from this information.

The next extension of this model can be made by observing that, if the intercept can vary between groups, so may the slope coefficients. We could, for example, assume that slopes vary according to a random group factor. So, in the equation:

$$Y_{ij} = (\gamma_0 + \gamma_1 z_j + u_j) + \beta_1 x_{ij} + e_{ij}$$

We could allow β_2 also to be a function of group level characteristics:

$$\beta_1 = \beta_0 + \beta_1 z_j + v_j$$

Substitution yields:

$$Y_{ij} = (\gamma_0 + \gamma_1 z_j + u_j) + (\beta_0 + \beta_1 z_j + v_j)x_{ij} + e_{ij}$$

Simplification yields:

$$Y_{ij} = (\gamma_0 + \gamma_1 z_j + u_j) + (\beta_0 x_{ij} + \beta_1 z_j x_{ij} + v_j x_{ij}) + e_{ij}$$

This is the full hierarchical linear model, also called a random coefficients model, a multilevel model, etc., etc. Notice that this model almost appears as a regular OLS model with simply the addition of a cross-level interaction between z_j and x_{ij} . Indeed, prior to the development of software to estimate this model, many people did simply include the cross-level interaction and estimate the model via OLS, possibly adjusting for standard error bias using some robust estimator for standard errors.

However, this model is NOT appropriately estimated by OLS, because we still have the random effect u_j and the term $v_j x_{ij}$. In a nutshell, this model now contains 3 sources of variance: within-individual (residual) variance, σ_e^2 , and two between-individual variances ($\tau_{intercept}^2$ and τ_{slope}^2).

We have now discussed the reason for many of the names for the hierarchical model, including multilevel modeling, hierarchical modeling, random/fixed effects modeling, random coefficient modeling, etc. We have not discussed growth curve modeling. I use growth curve modeling extensively in my research, and approach the hierarchical model from that perspective. I also approach the model from a probability standpoint, rather than a pure equation/residual variance standpoint. That being the case, this brief discussion of growth curve modeling will use a very different notation.

Until now, we have treated the two levels of analysis as individual and group. For growth curve modeling, the two levels are based on repeated measures on individuals across time. A basic growth model looks like:

$$\text{Within-Individual equation } \{ y_{it} \sim N(\alpha_i + \beta_i t, \sigma^2)$$

This equation says that time-specific individual measures are a function of an individual-specific intercept term and a slope term capturing change across time. The second level equation,

$$\text{Between-Individual equations } \left\{ \begin{array}{l} \alpha_i \sim N(\gamma_0 + \sum_{j=1}^J \gamma_j X_{ij}, \tau_\alpha^2) \\ \beta_i \sim N(\delta_0 + \sum_{k=1}^K \delta_k X_{ik}, \tau_\beta^2) \end{array} \right\}$$

says that there may be between-individual differences in growth patterns, and that they may be explained by individual-level characteristics. Realize that this model, aside from notation, is no different from the hierarchical model discussed above.