

Lecture Outline for Sociology 413/513: Sociological Research

Methods II

Aaron Gullickson

March 31, 2008

[Note: These lecture notes are not a substitute for class attendance. They are provided by the instructor as a class aid, but the reader assumes all risk of errors, typos, and the like.]

Contents

1	Review of OLS Regression	3
1.1	Administrative/Review	3
1.2	Missing data	6
1.3	Weighting	8
1.4	Model Selection	9
2	Generalized Linear Models	14
2.1	Linear Probability Models and Generalized Least Squares	14
2.2	Generalized Linear Models	16
2.3	Maximum Likelihood Estimation	18
3	Models for Categorical Dependent Variables	24
3.1	Odds and Probabilities	24
3.2	Logistic Regression	27
3.3	Other types of link functions (probit)	34
3.4	Multinomial logit models	38
3.5	Ordered Logit Models	40

4	Event History Analysis	43
4.1	Events, rates, and relative risk	43
4.2	Parametric survival models	47
4.3	Semi-parametric models	53
4.4	Discrete time approximation	55
4.5	Event count poisson regression	58
5	Multilevel Models	60
6	Causal Modeling	65
7	Factor Analysis and Structural Equation Models	70

1 Review of OLS Regression

1.1 Administrative/Review

- Administrative

- Welcome back. Everyone introduce themselves as we have some new faces.
- My goal in this class is to give you breadth rather than depth. I think it is more important at this stage in your graduate career to be able to read and understand a variety of methods at a basic level, than to invest in knowing a particular one forwards and backwards. Each of the topics we discuss in this class are themselves the subjects of full-length book treatment.
- The book: Powers and Xie is my bible. For the most part, its explanations are very clearly written. Some of the material is too technical for our level (sometimes using calculus), but these portions of the text can be skimmed with little loss of information.
- Powers and Xie provide the datasets used in most of their examples on the book's website - often with STATA code showing how to do the example. The website is

www.la.utexas.edu/research/faculty/dpowers/book.

- This class will be organized very differently from the previous class.
- Course will be split into modules topic-based "modules" one to two weeks in length.
- There is no book homework. The only homework will be in the lab session.
- The paper
 - * At the conclusion of the class, each student should turn in an original research paper using a technique learned in class.
 - * The statistical technique used should be something we have learned in class. OLS regression is OK, but most of you will probably find your questions better addressed by other methods.
 - * Develop your substantive question and then ask what method to use to address it.
Avoid mindless empiricism.
 - * Quality over quantity - I don't expect the scope I would expect from a master's paper, but I do expect a nice, tight statistical analysis.
 - * Where to get your data?
 - If the data address your question, you can use one of the datasets we will be using this semester or last: The Public Use Microdata Samples, the General Social Survey, the NLSY79, and the National Election Study.

- Some other good data sets are: the Current Population Survey, the Panel Study of Income Dynamics, Demographic and Health Surveys, AddHealth
- Government and major NGOs (like the UN) are good sources for data.
- A good place to search for data is the ICPSR
- Data does not have to be survey data to be analyzed statistically - creativity is encouraged
- * Due dates
 - At the end of the second week of class, students will turn in a one-page summary of their research question in broad terms.
 - At the end of the fourth week, students will identify a dataset that they will use for their research project.
 - At the end of the seventh week, students will turn in a literature review portion of their paper.
 - On the last day of class, students will turn in the full paper.
- * The last week of class will provide an opportunity for each student to present their findings to the class.
- Any other administrative questions?
- Review of OLS Regression
 - Module Readings: Chapter 1, Chapter 2, and Appendix A (skim 2.2.2)
 - Attempting to understand how a set of **independent** variables affect a particular **dependent** variable:
observed=structural+stochastic
 - There are different ways to view the above schematic:
 - Causation:** observed=true process+disturbance
 - Prediction:** observed=predicted+error
 - Description:** observed=summary+residual
 - Formula is always the same, but way of thinking about what you are doing is not - we will generally prefer the last method as it involves the least amount of baggage.
 - The "observed" portion of the above equation is given by the observed values of the dependent variable, y_i .
 - The structural portion of the model is given by a linear equation for the expected value of the dependent variable given certain values of the independent variables:

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}$$

- We saw last semester how to easily incorporate categorical independent variables, interaction terms, and even polynomial terms into this linear equation.
- To get the full equation you have to add on the stochastic "tail".

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon_i$$

- Assumptions on the error term
 - * In practice, it is necessary to assume that the error term is uncorrelated with the independent variables in order to produce **unbiased** and **consistent** (estimate approaches parameter in large samples) parameter estimates.
 - * We also usually usually assume that the error terms are i.i.d. (homoskedasticity and no autocorrelation) which means that our OLS estimator is also the most **efficient** (smallest variance among unbiased estimators).
 - * Assuming error terms are also normally distributed is a handy, but unnecessary condition in samples of sufficient size.
- The beta values give the intercept and slope terms which give us information about the relationship between the dependent variable and each independent variable controlling for the other independent variables.
- We also estimate σ - the variance of the error term, to see how far the actual values of y are spread out around the predicted values.
- These beta values are parameters for some population, but we normally have to estimate them from a sample - thus all of our rules of statistical inference come into play.

$$[\beta_0, \beta_1, \dots, \beta_p, \sigma] \rightarrow [b_0, b_1, \dots, b_p, s]$$

- * The general form used to estimate these parameters can be represented in linear algebraic form as:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Where \mathbf{X} is the matrix of independent variables and \mathbf{y} is the vector of the dependent variable. This formula minimizes the sum of squared residuals:

$$(y_i - \mathbf{x}_i'\beta)^2$$

- * The b values are random variables because they will differ from sample to sample. In order to make statistical inferences, we need to know the sampling distribution

of these statistics. It turns out that the variance of these sampling distributions are given by:

$$\sigma_{\mathbf{b}}^2 = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

σ^2 is estimated by s^2 the observed variance of the residuals:

$$s^2 = \frac{\sum(y_i - \hat{y}_i)^2}{n - p - 1}$$

One can then use these estimates to construct a test statistic:

$$t = b_k / s_{b_k}$$

which, when the sample is sufficiently large, will roughly be distributed as a t-distribution with $n - p - 1$ degrees of freedom. Thus, we can run a hypothesis test to determine whether there is a relationship between x_k and y .

- Remember that there are three major assumptions of OLS regression. (go through the consequences)
 1. Linearity (poor fit)
 2. independence of dependent variables and error terms (not unbiased or consistent)
 3. i.i.d. of error terms (not efficient)

1.2 Missing data

- Up until now, we have ignored one of the major difficulties with observational data: it is quite common to have a substantial number of **missing values** for variables that you care about. In order to perform any statistical analysis, you will need to handle these missing values somehow.
- How you handle them depends to some extent on why they might be missing.
 1. The missing values are **missing completely at random** (MCAR). Missing values are randomly drawn from the sample and do not depend on other variables. This assumption can be roughly tested by running t-tests between the missing and non-missing group on other variables.
 2. The missing values are **missing at random** (MAR). In this case, the missing values are randomly distributed conditional on some other measured variables. In other words, missingness is only a function of the observed data.

3. The missing values are **missing not at random** (MNAR). Given the observed data, Missingness is still a function of the unobserved observations themselves.
- There are a variety of methods for handling missing data, but most of them are problematic.
 - **Deletion.** Delete cases with missing values. There are two possibilities:
 1. **Casewise deletion:** delete all cases with missing values on any of the variables you will use in the analysis.
 2. **Pairwise deletion:** delete cases for each statistical run, depending upon which variables are used in which run. This improves the sample size for most procedures, but it is very dangerous because differences across models can be the result of adding different variables or the differential deletion of cases.

This method assumes the data are MCAR. It has many problems:

- * It can produce biased results if even a moderate number of cases are missing
 - * If a large number of variables are used, small numbers of missing values per variable can add up to a lot of missing cases on the whole. This means a lot of non-missing data will be thrown out.
- **Mean Imputation:** Simply substitute the mean value for all missing values. This method also assumes MCAR. This is also an extremely problematic method:
 - * Variance will be underestimated.
 - * Typically, correlations and regression coefficients will also be underestimated.
 - **Imputation with dummy:** Substitute any constant value for all missing values. Also include a dummy variable into the regression model indicating where imputation occurred. This method also assumes MCAR.
 - * The coefficient on the dummy variable measures how far off your imputation was from what would be expected based on the non-missing values. (draw a picture)
 - * It shares the problem of underestimating variance with the previous method, but regression coefficients will be estimated based only on the set of non-missing values (as if casewise deletion had been used).
 - **Regression imputation:** Using this technique, other variables are used to predict the value of the variable where data are missing and then the fitted values are imputed. This method assumes MAR. It is less likely to produce bias, but still will underestimate variance.

- **Random Regression imputation:** This technique is like the previous one, but a random component is added to each imputation (usually based on the residuals of the model) in order to better estimate the variance. This technique
- **Multiple imputation:** The problem with all of the other techniques is they do not account for **imputation variability** - that is the uncertainty in our estimate due to the imputations used - this variability must be reflected in the standard errors of our estimates. Multiple imputation does this by running each statistical procedure multiple times with different imputed values each time. The results from each run are then pooled to produce final estimates which account for imputation variability.

1.3 Weighting

- As we discussed last term, many surveys use a stratified and/or cluster design. When the sample drawn is no longer a simple random sample, one has to take account of each observation's probability of being sampled.
- The inverse of the probability of being sampled give the population weights (p_i). Dividing the population weight by its mean gives sampling weights:

$$w_i = p_i / \bar{p}$$

- In order to produce unbiased estimates for the population, one must account for these sampling weights. In order to estimate the mean from weighted data, for example:

$$\bar{x} = \frac{\sum (w_i * x_i)}{\sum w_i}$$

- Weights can be used in regression as well to produce unbiased estimates. Instead of minimizing the sum of squared residuals, one wants to minimize the weighted sum of squared residuals. In matrix form, we want to minimize:

$$\sum_i w_i (y_i - \mathbf{x}_i' \beta)^2 = (\mathbf{y} - \mathbf{X}' \beta)' \mathbf{W} (\mathbf{y} - \mathbf{X}' \beta)$$

Where \mathbf{W} is an $n \times n$ matrix where the diagonal gives the weights for each individual. The solution for minimizing this formula is given by:

$$\mathbf{b} = (\mathbf{X}' \mathbf{W} \mathbf{X})^{-1} \mathbf{X}' \mathbf{W} \mathbf{y}$$

and

$$\mathbf{s}_b^2 = s^2 (\mathbf{X}' \mathbf{W} \mathbf{X})^{-1}$$

These weights can be implemented in STATA using the pweights command.

- Some cautions are in order:
 - If the variables on which the weighting occurs are properly specified in the model, then regular OLS will also produce unbiased estimates and its standard errors will be smaller. For example, if the weights came about simply by taking an oversample of blacks, then an indicator variable for being black would be sufficient to account for the weighting as long as there are no important interaction effects between black and other variables that are left unspecified.
 - In some cases, complex stratified and cluster sampling designs can lead to larger standard errors on the estimates than can be accounted for just by weighting. These kind of **design effects** can be handled in STATA with the svy* set of commands.

1.4 Model Selection

- When two models are “nested”, we can formally compare them using the F-test.
 - Recall the F-test for one model, The null hypothesis is that none of the independent variables are related to the dependent variable (all $B_j=0$).

$$F = MSM/MSE$$

If the null hypothesis is true, then this term is distributed as an F-distribution with p DF in the numerator and $n - p - 1$ in the denominator.

- Essentially the F-test is a test of the current model vs. the null model. We can also use the F-test to compare two non-null models. Consider two models, where model 2 is nested within model 1. then:

$$F = \frac{(SSE_1 - SSE_2)/g}{SSE_1/(n - k - 1)}$$

Where g is the number of parameters added in model 2 and k is the number of parameters in model 1. Under the null hypothesis that none of the variables added in the second model is related to the dependent variable, this term is distributed as an F-distribution with g DF in the numerator and $n - k - 1$ in the denominator.

- If model 2 adds only a single parameter then this F-test is equivalent to simply looking at the p-value for this new variable in model 2.
- this technique is not reliable in deciding among a possible set of predictors because the effect size and standard error of each variable will differ depending on what other variables are included. There is an uncertainty about the appropriate model that we need to account for.

- We are often interested in a comparison between models, or in including the right control variables to get the best estimate of the net effect of a particular independent variable. **Model selection** plays a large role in social science research - where is the right model between the **null model**

$$y_i = \beta_0 + \epsilon_i = \mu_y$$

and the **saturated model**?

- Example: state-level crime data from Ehrlich (1973). The question of interest is whether rational incentives like the probability of being caught and the length of imprisonment affect crime rates. In order to assess the effects of these variables we want a model that can control for other important predictors of crime rates. Our candidate variables are:

1. percent of males aged 14-24
2. south indicator variable
3. mean years of schooling
4. police expenditures in 1960
5. police expenditures in 1959
6. labor force participation rate
7. M/F sex ratio
8. population
9. nonwhites per 1000 population
10. unemployment rate of urban males 14-24
11. unemployment rate of urban males 35-39
12. GDP
13. income inequality
14. probability of imprisonment
15. average time served in state prisons

The full model with all variables and Ehrlich's chosen model are shown in the table. It is clear that the effects are much smaller in the full model than in the model Ehrlich chose, which makes us wonder if his effects are measured using the right model.

- We assess the goodness of fit of the OLS model generally by R^2 , but we have to balance goodness-of-fit with parsimony.

- We can always get a better fit (or at least not a worse fit) by adding more variables to the model, so R^2 is not helpful in and of itself. One technique for balancing parsimony is to use the **adjusted R^2** :

$$\text{adj.}R^2 = 1 - \left(\frac{n-1}{n-p-1}\right)\left(\frac{SSE}{SST}\right)$$

The expression before the sum of squares ratio will always be less than one when k is greater than zero, so this expression "penalizes" R^2 for including more variables into the model. We could try choosing the model which maximizes our adjusted R^2 .

- Another technique is to proceed in a stepwise fashion. There are three basic approaches:
 1. **Forward Selection.** Begin with the variable with the highest single R^2 value. Next search for the variable to add which will lead to the next greatest increment in R^2 , and continue this process until all possible variables are added or some stopping criterion is reached (typically that the next added value is not stat. sig. at some level).
 2. **Backward Selection.** Begin with the full model. Remove one variable based on some criterion (typically the smallest t-value). Continue this process until all remaining variables are above some pre-established threshold.
 3. **Stepwise Selection.** Begin like forward selection, but at the end of each step, remove any variables that have fallen below a certain threshold (typically on the t-value) before moving forward another step.

- These methods can sometimes give different results.
- Elimination is purely on statistical significance.
- What about things like interactions or polynomial terms?

- Bayesian Information Criterion (BIC)

- The mathematics behind BIC are fairly complex, but it is fairly intuitive to understand in practice - In addition it can be applied to a wide range of linear models besides OLS regression with the same interpretation, making it very versatile.
- For OLS regression, BIC' (the punctuation indicates our comparison is the null model) is calculated as:

$$BIC' = n \log(1 - R^2) + p \log(n)$$

This equation balances parsimony and goodness-of-fit.

- BIC' is implicitly compared to the null model. The smaller BIC' is, the better. If $BIC' < 0$, then the current model is preferred over the null model.

Variable	Full	Ehrlich	Adj. R^2	Stepwise	Forward	Backward
Intercept	-15.59 0.12	-19.56 0	-14.78 0.09	-24.38 0	-22.64 0	-24.38 0
% male 14-24	1.57 0	1.17 0.05	1.56 0	1.51 0	1.48 0	1.51 0
south	0.06 0.64					
years of education	2.15 0		2.21 0	2.39 0	2.22 0	2.39 0
police exp., 1960	0.82 0.3		0.79 0	0.91 0	0.85 0	0.91 0
police exp. 1959	-0.04 0.96					
labor force particip.	0.6 0.39	0.61 0.3				
M/F sex ratio	-2.39 0.21		-2.12 0.14			
population	-0.08 0.13		-0.08 0.11			
% nonwhites	0.11 0.02	0.2 0	0.12 0	0.08 0.03	0.11 0.01	0.08 0.03
unemp. rate, 14-24	-0.13 0.69	0.09 0.69				
unemp. rate, 35-39	0.45 0.06		0.3 0.03	0.32 0.02	0.29 0.03	0.32 0.02
GDP	0.67 0.11	1.74 0	0.72 0.07			
income inequality	1.59 0	0.93 0.03	1.68 0	1.23 0	1.24 0	1.23 0
imprisonment prob.	-0.3 0	-0.43 0	-0.3 0	-0.19 0.01	-0.31 0	-0.19 0.01
average time served	-0.27 0.13	-0.53 0.01	-0.25 0.13		-0.29 0.06	
R^2	0.870	0.700	0.863	0.827	0.842	0.827
Adj. R^2	0.806	0.637	0.820	0.795	0.809	0.795
BIC'	-37.9	-25.8	-51.2	-55.4	-55.9	-55.4

- BIC' can also be used to compare two non-null models, even if they are non-nested. The model with a smaller (generally, more negative) BIC' is preferred.
- The best approach depends on what you are doing with the model.
 - Its best to have a theoretical reason for fitting the model as you are, rather than just throwing in variables randomly and seeing what happens (monkeys and computers can do that—you are supposed to be the brains of the operation)
 - Generally we are interested in a particular variable or a small set of variables and the others are included as controls. Even if these controls look not statistically distinguishable from zero, it may still be good to include them in the equation. Show the full model and let your readers make their own decisions.
 - In some social science research, the key issue is a comparison of models which present competing explanations of the process. In these cases, model selection is very important, and you should carefully consider your options. Generally it is best here to present several techniques.

2 Generalized Linear Models

2.1 Linear Probability Models and Generalized Least Squares

- So far we know how to use both continuous and categorical variables to predict continuous variables. What if we want to predict categorical variables? (i.e. use them as the dependent variable) In sociology much of what we care about comes as categorical not continuous.
- Let's say we are interested in explaining why individuals fall into different categories of a dichotomous variable.

- Lets take our titanic example. Let y_i equal 1 if the person survived the titanic and 0 if the person did not. Let's look at how survivorship varied as a function of the fare (in british pounds) that a person payed:

Variable	Coefficient
Constant	0.3058 (19.27)
Fared paid (in british pounds)	0.0023 (9.10)

- What do the intercept and slope mean in this particular case? The slope is the the change in your probability associated with paying one more pound for your fare. In this case, for every extra pound paid, your probability of surviving goes up by about 0.22%. Those who paid no fare are estimated to have a probability of survivorship of 30.58%. This is called the **Linear Probability Model**.
- The IQR for fare payed was approximately 17 pounds and the maximum fared payed was 512 pounds. so it seems that fare had a fairly sizeable effect on survivorship, which is not surprising given that fare is a good proxy for the social class of passengers.
- There are two problems with the linear probability model.
 - * The i.i.d. assumption is violated. The variable y_i is distributed as a bernoulli variable. The variance of a bernoulli variable depends on the true probability of a "yes." Thus the variance of the error term will not be constant. In technical terms, this is the problem of **heteroskadasticity**.
 - * The linear probability model can produce fitted values that are outside the range of $[0,1]$. Take the passenger who paid the highest fare.

$$.3058 + .0023 * 512 = 1.48$$

- The first problem can be dealt with using the technique of **generalized least squares**.

- * Generalized least squares allows for a weighting matrix to be included into the estimation of β 's.

Before, we had:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Now, we have:

$$\mathbf{b} = (\mathbf{X}'\Psi^{-1}\mathbf{X})^{-1}\mathbf{X}'\Psi^{-1}\mathbf{y}$$

Ψ^{-1} is the **weighting matrix**. It has the effect of basically transforming the y and x variables. If we put the right values into our weighting matrix, we will recover our i.i.d. assumptions, even in the face of heteroskedasticity and autocorrelation.

- * The diagonals of the weighting matrix tell us how to adjust our variance to reflect heteroskedasticity. The off-diagonal cells tell us about the autocorrelation between error terms.
- * Let's just focus on the heteroskedasticity part for now - time-series models make use of GLS to estimate autocorrelation.
- * Because we only have sample data, we don't know exactly what values to put into our weighting matrix. We have to estimate them from our data. When we do that, we call it **feasible generalized least squares** (FGLS).
- * The values to put into the diagonal should be the inverse of the variation in our error terms (let's take an example).
- * In our binary case above, the variation in y_i is proportional to the underlying true p_i for the i th observation.

$$V(y_i) = p_i(1 - p_i)$$

- * If we put the inverse of these values into our weighting matrix, we will recover our i.i.d. assumption.

$$\Psi^{-1} = \begin{pmatrix} \frac{1}{p_1(1-p_1)} & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{p_2(1-p_2)} & 0 & \dots & 0 \\ 0 & 0 & \frac{1}{p_3(1-p_3)} & \dots & 0 \\ \vdots & \dots & \dots & \dots & \vdots \\ 0 & 0 & 0 & 0 & \frac{1}{p_n(1-p_n)} \end{pmatrix} \quad (1)$$

- * Of course, we don't know what the true p_i is for each observation, so we have to use an estimate. How do we get it? Run a linear probability model and use the resulting \hat{p}_i in place of p_i . Then we have FGLS.

* FGLS can usually be improved by iterating the result, when it becomes **Iteratively Reweighted Least Squares** (IRLS).

1. Run linear probability model using OLS. Get the predicted values of p_i from this run.
2. Use \hat{p}_i in the weighting matrix to run an FGLS. Get the predicted values of p_i from this run.
3. Use the updated \hat{p}_i in the weighting matrix of another FGLS. Get the predicted values of p_i from this run.
4. Repeat step 3 until the estimates of β no longer change very much.

* Let's try it out on our titanic data:

Coefficient	1	2	3	4	5	6	7	8
Intercept	0.30588	0.31292	0.3074	0.30987	0.30857	0.30921	0.30888	0.30905
Slope	0.00229	0.00186	0.00211	0.00199	0.00205	0.00202	0.00204	0.00203
Coefficient	9	10	11	12	13	14	15	
Intercept	0.30896	0.30901	0.30898	0.30899	0.30899	0.30899	0.30899	
Slope	0.00203	0.00203	0.00203	0.00203	0.00203	0.00203	0.00203	

* You can still run into problems with FGLS if your estimated values of p_i extend beyond the range of $[0,1]$ because it can lead to negative weights. In this example, I had to "hack" a few observations that were outside the range by substituting the maximum value observed within the range.

- GLS can resolve the problem of error terms which are not i.i.d., but it still can produce fitted values outside the range of the dependent variable. This second problem requires us to move to **Generalized linear models**.

2.2 Generalized Linear Models

- Our OLS regression equation is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon_i$$

- For this model, we assume:
 1. The relationship between the independent variables and dependent variables is a linear function.
 2. The errors are distributed identically with a constant variance.
- These assumptions become problematic for categorical dependent variables because:

1. Predictions can fall outside the range of the of the dependent variable (linearity problem)
2. The variance of the error term is often related to the expected value (\hat{y}_i) for each observation (true for both binomial and poisson distributed variables), meaning we have non-constant variance, or heteroskadasticity (i.i.d. problem). To put it more formally:

$$V(y|x) = V(\hat{y}) = f(\hat{y})$$

- The generalized linear model deals with this problem by two extensions of the linear regression model.

1. We have a **link function** that relates our linear function of the independent variables not to y_i directly, but rather to some function of y_i , given by $g(y_i)$. The idea of the link function is to expand the range of y from some limited range to the range of $(-\infty, \infty)$.
 - Take our titanic example. We are interested in modeling p_i , each individual's probability of survival. However, p_i is bounded by 0 and 1.
 - We can improve this by looking at the odds of survival rather than the probability of survival. (we will discuss odds in more depth in the next module)

$$o_i = \frac{p_i}{1 - p_i}$$

- While probabilities are bounded by 0 and 1, odds are bounded by 0 and ∞ . This is an improvement, but we still could get results outside of the lower bounds. We can resolve this problem by logging the odds:

$$\text{logit}(p_i) = \log(o_i) = \log(p_i(1 - p_i))$$
 - This is called the logit transformation. Rather than related our x variables linearly to p_i , we relate them linearly to the logit transformation of p_i .
- 2. Rather than assume i.i.d., we specify a particular **error distribution** for our errors. This distribution is based on what we think is driving the categorical distinctions we observe in y . (i.e. binomial for dichotomous variables)
 - Once again from our example. We believe each observation has an underlying p_i , but we actually only observe ones and zeroes (success and failures). Therefore, the error distribution from our predicted to actual values will be given by the binomial/bernoulli distribution.

y_i is given by $\text{binom}(p_i)$

$$V(p_i) = p_i(1 - p_i)$$

- Once we do this we have a slightly different form for the equation:

$$g(p_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}$$

- Generally, certain link and error functions go together.

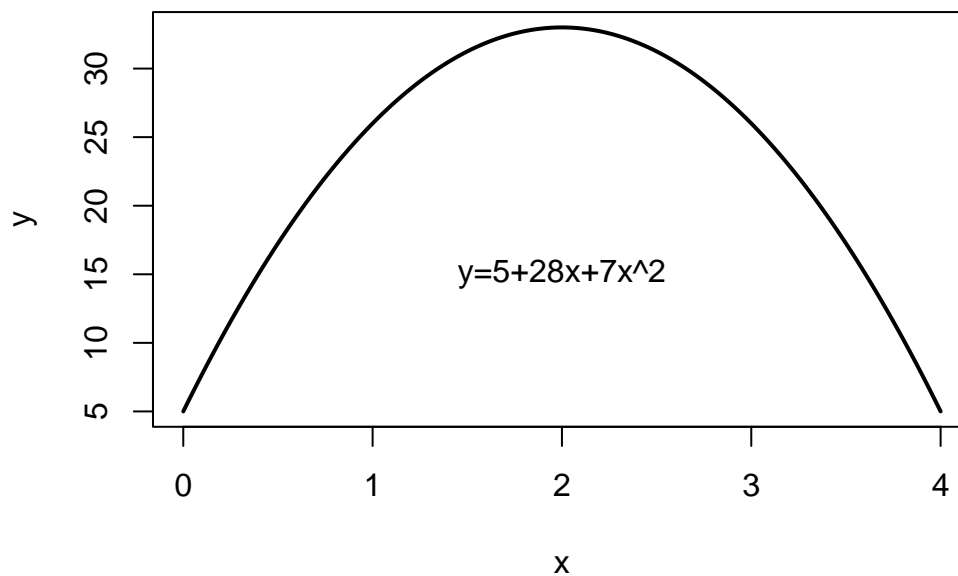
Type of category	link function	error dist.	name
continuous	identity	normal/gaussian	OLS regression
dichotomous	logit	binomial	logistic/logit regression
	probit	binomial	probit regression
discrete counts	log	poisson	log-linear models/poission regression
polytomous (unordered)	logit	multinomial	multinomial logit
polytomous (ordered)	logit	multinomial	ordered logit
survival times/counts of events	log	poisson	continuous-time hazard model

(discuss each of the items on the table)

2.3 Maximum Likelihood Estimation

- How do we actually estimate parameters for generalized linear models?
 - Take our titanic example. You might think that we could simply take the logit transformations of our y 's and then use GLS to correct for heteroskedastacity.
 - The problem here is that our y 's only take the values of one and zero which will both be undefined in the logit transformation.
 - Therefore, we need a new technique - this is where maximum likelihood estimation comes in.
- Crash couse in Calculus
 - It helps to know a few basic things from calculus.
 - The derivative
 - * Start with a function $f(x)$. This is some mathematical equation involving x .

$$f(x) = 5 + 28x - 7x^2$$



- * Every function has a derivative. The derivative is basically the exact slope of the function at x .

$$\frac{df(x)}{dx} = 28 - 14x$$

- * Local minimums and maximums of the function $f(x)$ will occur when the slope is zero. (show graphically).

$$0 = 28 - 14x$$

$$14x = 28$$

$$x = 2$$

- * You can tell if it is a local minimum or maximum by looking at the second derivative, which is the derivative of the derivative. If this is positive at that point, then it is a minimum. If it is negative, then it is a maximum.

$$\frac{d^2f(x)}{dx} = -14$$

- * So this is a maximum.

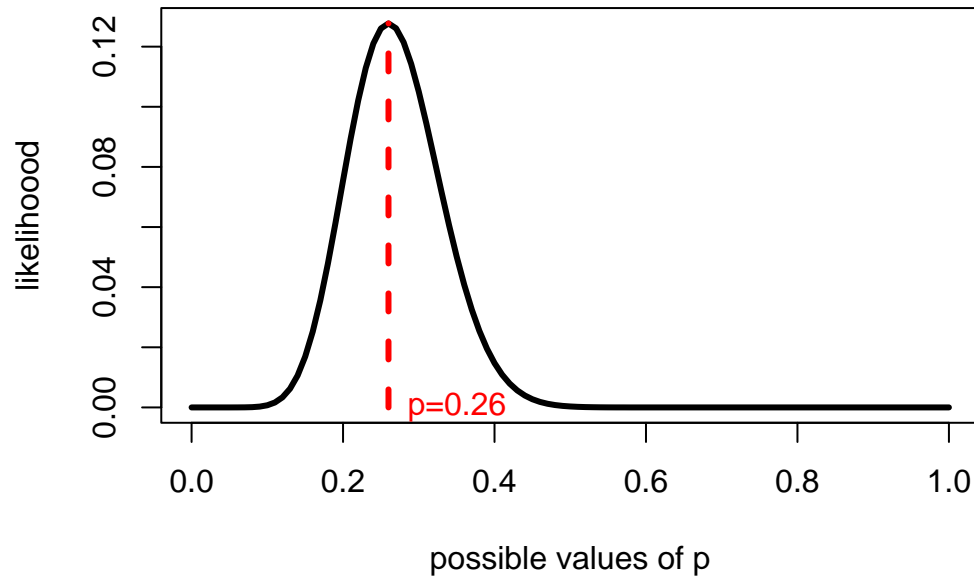
- The logic of maximum likelihood estimation is as follows:
 1. We have some process governed by some parameter(s), (say θ) which generates a set of observed data $(x_1, x_2, \dots, x_n$ or something).
 2. We ask what is the likelihood, given the process, that we observe the given data? This leads to a **likelihood function**, $L(\theta)$.
 3. The trick is we don't know θ . Rather it is usually the data we have and θ that we want to estimate. We choose a $\hat{\theta}$ such that it maximizes our likelihood function given the data.
- Example with a binomial proportion.
 - We have observed the outcome from a series of n coin tosses. From this coin toss, we have observed x heads. n and x are our observed data.
 - We know that the process of generating heads on coin tosses is a binomial process, because we only have two possible outcomes, and the probability p is the same on each coin toss. p (the probability of a heads on a single coin toss) is our unknown parameter.
 - Since this is a binomial process our likelihood of observing a particular number of heads in n trials is given by the following likelihood function:

$$L(p) = \binom{n}{x} p^x (1-p)^{n-x}$$

(review this formula a bit)

- We don't know the value of p , but we want to estimate it by choosing \hat{p} that maximizes $L(p)$ for the given n and x .
- For particular values of n and p , we could get \hat{p} by trial and error, but that might take a long time. A computer might simplify it.

Let's take the example, where $x = 13$ and $n = 50$. We can use the computer to calculate the likelihood functions for every p at intervals of 0.01 from 0 to 1.



The p which would maximize our likelihood of actually observing 13 heads on 50 tosses is about 0.26 - this is what we would expect by common sense.

- Even so, we would like to get a general result that can be expressed in the abstract terms n and p .
- Thus, we need to find the maximum of the likelihood function for p . We can do this using calculus. The point where the derivative of the likelihood function is zero and the second derivative is negative is the maximum.
- It turns out, that working with the log of the likelihood function is usually easier than working with the likelihood function itself. The results for the log-likelihood function hold for the likelihood function (the maximum is the same), so logging it does no damage.

$$\log L(p) = \log \binom{n}{x} + x \log(p) + (n - x) \log(1 - p)$$

- We take the derivative of this function with respect to p .

$$\frac{\partial \log L(p)}{\partial p} = \frac{x}{p} - \frac{n - x}{1 - p}$$

- When this function equals zero, we will have either a minimum or a maximum. So solve for p .

$$\begin{aligned}
 0 &= \frac{x}{p} - \frac{n-x}{1-p} \\
 \frac{n-x}{1-p} &= \frac{x}{p} \\
 p(n-x) &= x(1-p) \\
 pn - px &= x - px \\
 pn &= x \\
 p &= \frac{x}{n}
 \end{aligned}$$

- It turns out the second derivative at this point is negative (not shown), so this is indeed a maximum. Thus our best estimate of p is $\hat{p} = x/n$. The same as our intuition would have told us.
- This example was relatively straightforward. But we can use the same technique with the linear model.
- Our parameters are the β 's, which relate our x variables to our y variable.
- For some sample, we have observed data on the y variable and the set of x variables. Given this data, there are some β values which would have maximized our likelihood of observing it. By MLE, these values are our estimates of the β parameters.
- How do we figure out what β 's would have maximized our likelihood of observing the actual data? That is complicated. The remainder of appendix B provides a good introduction to the general technique, but it is highly technical. All we need to know for now is that computers are very good at finding such things. We just need to know intuitively what process guides their decision.
- The use of MLE gives us a new method of model selection.
 - We can think of the actual log-likelihood returned from maximizing the likelihood function as a measure of how well it fits.
 - However, we can't use the log-likelihood $\text{Log}L$ directly, because different sample sizes will have different likelihoods.
 - What we can use is what is called **deviance** or G^2 , which measures how much our current model deviates from the saturated model. If L_f is the likelihood of the saturated model,

then:

$$G^2 = -2\log(L_c)/L_f = -2(\log L_c - \log L_f)$$

The log of the likelihood will always be negative or zero, so multiplying it by a negative number will create a positive number. The closer this deviance is to zero, the closer our likelihood is to the saturated model.

- For most of the individual-level data we work with the likelihood of the saturated model is 1, so $\log L_f = 0$, so our deviance simplifies:

$$G^2 = -2\log L_c$$

- We can use deviances to compare two nested models using the Likelihood Ratio Test (LRT).
- Differences between nested models follow our friend the χ^2 distribution. Therefore, if we have two models n and u in which n is nested within u , the value $G_u^2 - G_n^2$ has a χ^2 distribution with DF of $df_u - df_n$, assuming model u is the correct model. This method can be used to compare two or more nested models.
- You can also use deviances to generate *BIC* statistics. To generate a BIC against the saturated model:

$$BIC = G^2 - df \log n$$

In most cases, however we will want BIC' relative to the null model. To calculate this first get the deviance for the null model (often reported in statistical software). Then:

$$BIC' = (G_0^2 - G^2) + p \log n$$

where p is the number of parameters fit in the current model.

- Although BIC and BIC' have different interpretations directly, when you compare the BIC of BIC' of two models, you should get the same difference.

3 Models for Categorical Dependent Variables

3.1 Odds and Probabilities

- Odds and probabilities

– If there is a p probability of something happening, then the odds can be considered the number of successes you expect to get for every failure on average. High odds correspond to high probabilities, low odds to low probabilities.

– To calculate the odds:

$$O = \frac{\text{proportion of successes}}{\text{proportion of failures}} = \frac{p}{1 - p}$$

– To go from odds back to probabilities:

$$p = \frac{O}{1 + O}$$

– While probabilities are bound to $[0,1]$, odds are bound to $[0,\infty)$

– Let's take the titanic example. What was the overall odds of survival on the Titanic. 38% of passengers survived the Titanic. Therefore:

$$\text{Odds of survival} = \frac{0.38}{1 - .38} = \frac{0.38}{0.62} = 0.61$$

How can we say this in English: "For every one death on the Titanic, there was on average 0.61 survivors."

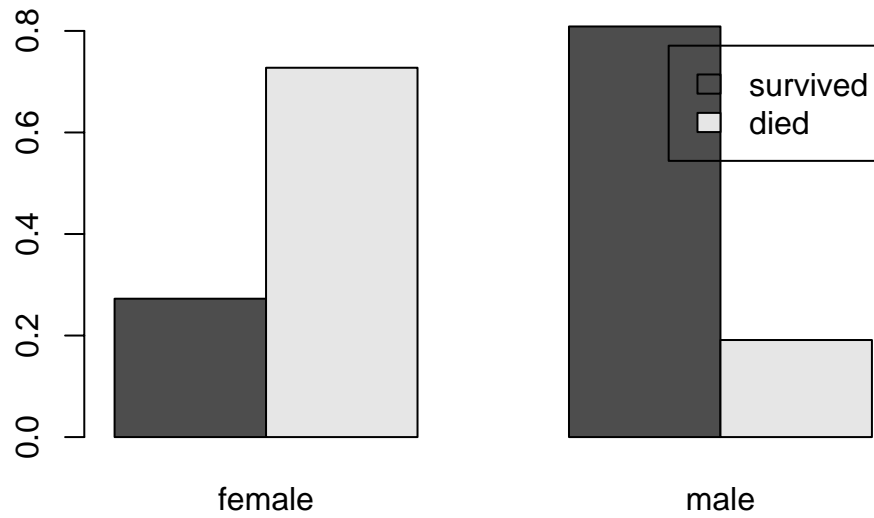
– Let's see how the odds change across the range of the probability.

p	O
0	0
0.01	.010
0.05	.052
0.10	.111
0.3333	0.5
0.40	0.6667
0.5	1
0.6	1.5
0.6667	2
0.75	3
0.9	9
0.95	19
0.999	999

- **Odds ratio:** the ratio of two odds. If one odds is twice as big as the other, then we say the odds ratio is 2. If it is half again as big, the odds ratio would be 1.5.

$$OR = O_1/O_2 = \frac{p_1}{1-p_1} / \frac{p_2}{1-p_2}$$

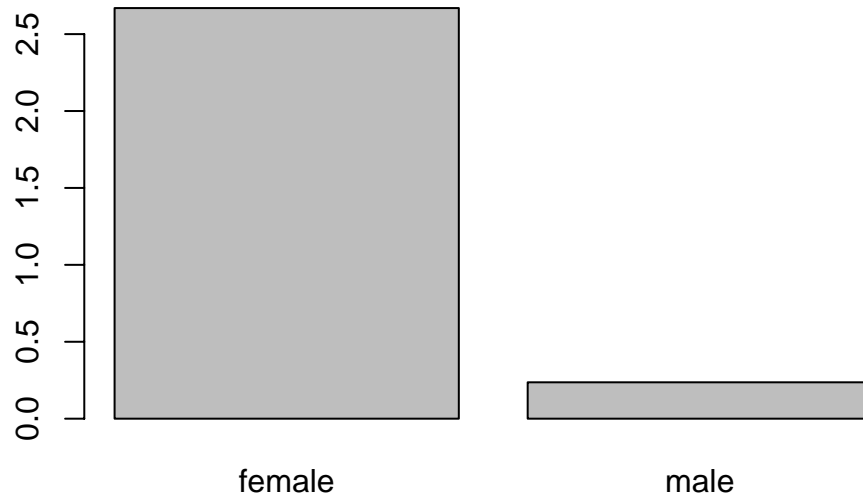
- Let's go back to the Titanic example. Let's look at the odds ratio between men and women on the Titanic. 19% of men survived while 73% of women survived. Let's look at this as a bar graph.



Now let's calculate the odds of survival for each group.

$$\text{Odds for women} = 0.73/(1 - 0.73) = 2.67$$

$$\text{Odds for men} = 0.19/(1 - 0.19) = 0.24$$



The odds of survival for women is 2.67 and the odds of survival for men is 0.24. The odds ratio is given by:

$$OR = 2.67/0.24 = 11$$

In English: "Women were eleven times as likely as men to survive the Titanic."

- Let's look at some of our previous examples, where we have an odds ratio of 2.

p_1	O_1	O_2	p_2	$\frac{p_2}{p_1}$
0.05	.052	.1004	0.912	18.24
0.1	.1111	.2222	0.1818	18.18
0.3333	0.5	1	0.5	1.5
0.5	1	2	0.6667	1.333
0.75	3	6	0.857	1.14
0.9	9	18	0.947	1.052
0.95	19	38	0.974	1.025
0.999	999	1998	0.9995	1.0005

- The odds ratios are constant here, but not the ratios of the probabilities directly. A doubling of the odds when the odds is already high doesn't have nearly the effect on the

probability of doubling the odds when the odds is low. The relationship is reversed if we halve the odds.

- Our interpretations of the parameters in the logistic regression model will be related to this odds ratio.

3.2 Logistic Regression

- Logistic regression

- When we have a dichotomous variable as the dependent variable, OLS regression won't work. The **Linear Probability Model** can be fit, but
 - * The relationship is non-linear because the probabilities are bound between 0 and 1.
 - * The error terms are heteroskedastic because the dependent variable is produced by a binomial process where the variance depends upon the underlying value.
- As we have learned we can correct these problems with a **generalized linear model**.
- We know that the error distribution is given by a binomial distribution. So, we only need to choose a link function. We know the identity link won't work because we have the non-linearity problem.
- There are several possible link functions, but the best one (or at least the easiest to interpret) is the **logit** function.
- The logit is the log of the odds:

$$\text{logit}(p) = \log \frac{p}{1-p}$$

- This function spreads the probabilities over the entire number range.
- So, our **logistic regression** model looks like:

$$\log \frac{\hat{p}_i}{1-\hat{p}_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}$$

- How do we interpret the β 's? Well, first let's relate this equation back to odds rather than the log-odds by exponentiating both sides.

$$\frac{\hat{p}_i}{1-\hat{p}_i} = e^{\beta_0} e^{\beta_1 x_{i1}} e^{\beta_2 x_{i2}} \dots e^{\beta_p x_{ip}}$$

- How does a one-unit change in x_{i1} affect the predicted odds?
- It increases the odds by a multiplicative factor of e^{β_1} .

- By exponentiating the β 's we get **odds ratios** - how much the odds increase multiplicatively with a one-unit change in the independent variable. For categorical variables, these can be interpreted directly as odds ratios between groups. For continuous variables they are the odds ratios between individuals who are identical on the other variables but differ by one unit on the variable of interest. (show an example of each)
- Therefore, the β 's themselves are **log-odds ratios**. Negative values indicate a negative relationship between the probability of "success" and the independent variable; positive values indicate a positive relationship.
- When you exponentiate them, the dividing line between a positive and negative relationship is 1 not 0.
- Let's take our titanic example. We have the equation:

$$\log \frac{\hat{p}_i}{1 - \hat{p}_i} = \beta_0 + \beta_1 x_i$$

Where x_i is an indicator variable for women vs. men. The results of this model are given by:

Variable	Coefficient
Constant	-1.44 (-16.46)
Gender	
Women	2.42 (17.82)
Men (ref.)	-

What is the intercept giving us here? The log-odds of survival for the reference group (Men). To convert this into odds take the exponential:

$$e^{-1.44} = 0.24$$

What is the slope giving us. The difference in the log-odds ratio of survival between men and women. To convert this into an odds ratio take the exponential:

$$e^{2.42} = 11.25$$

If we want to know the log-odds of survival for women then we have to add the relevant parameters:

$$\text{log-odds of a woman's survival} = -1.44 + 2.42 = 0.98$$

To get this as an odds, exponentiate:

$$e^{0.98} = 2.67$$

These numbers should look familiar. They are precisely what we got before by hand.

- The coefficients returned from a logistic regression model are **log-odds ratios**. They tell us how the log-odds of a "success" change with a one-unit change in the independent variable.
- Increasing the log-odds of a success means increasing the probability, and vice-versa decreasing the log-odds of a success means decreasing the probability. Therefore, the sign of the log-odds ratio indicates the direction of its relationship: + means a positive relationship between x_1 and the likelihood of a success, and - means a negative relationship.
- In order to get an intuitive sense of how much things are changing, we need to get the exponential of the log-odds ratio, which gives us the **odds ratio** itself.
- Let's return to our example from yesterday looking at survival of the Titanic by gender:

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_i$$

$$\log \frac{p_i}{1-p_i} = -1.44 + 2.42x_i$$

The positive coefficient indicates that women were more likely to survive the Titanic than men. But our coefficients are related to the log-odds of survival. Let's exponentiate both sides to see how they related to the odds of survival.

$$\frac{p_i}{1-p_i} = e^{-1.44} e^{2.42x_i} = (0.24)(11.25)^{x_i}$$

The odds for any individual is a multiplicative function of a "baseline" odds and "odds ratios" of their characteristics. The predicted odds for a man are:

$$\frac{p_i}{1-p_i} = (0.24)(2.42)^0 = 0.24$$

The odds for a woman are:

$$\frac{p_i}{1-p_i} = (0.24)(11.25)^1 = 0.24(11.25) = 2.67$$

- For the log-odds ratios, a negative value indicates a negative relationship. But all odds-ratios are positive values. The distinction regarding a positive or negative relationship in the odds ratios is given by which side of 1 they fall on. 1 indicates no relationship. Less than one indicates a negative relationship and greater than one indicates a positive relationship.
- The interpretation is similar with continuous variables. Let's take the case of predicting survival by fare paid.

$$\log \frac{p_i}{1-p_i} = -0.882 + 0.012x_i$$

Once again let's exponentiate this to get the results in terms of the odds of survival.

$$\frac{p_i}{1 - p_i} = e^{-0.882} e^{0.012x_i} = (0.414)(1.012)^{x_i}$$

Once again we have a multiplicative relationship. Let's take the three cases where the paid either zero, one, or two pound for his/her ticket.

$$\begin{aligned} \frac{p_1}{1 - p_1} &= (0.414)(1.012)^0 = 0.414 \\ \frac{p_2}{1 - p_2} &= (0.414)(1.012)^1 = 0.414(1.012) \\ \frac{p_3}{1 - p_3} &= (0.414)(1.012)^2 = 0.414(1.012)(1.012) \end{aligned}$$

For the first person, the odds of survival is simply given by the exponential of the intercept term, which in this case leads to an odds of 0.414. For the second person, the odds of survival increases by a factor of 1.012 because this person paid a pound more than the first. For the third person, the odds of survival increase a further factor of 1.012 because this person paid a pound more than the second.

- The exponential of the coefficient then gives the expected odds ratios between two individuals who only differ by one unit on the given independent variable.
- We can think of interactions in a similar way: they tell us how much the odds ratio related to one variable is different between groups. Lets now interact gender and fare in our Titanic example.

$$\log \frac{p_i}{1 - p_i} = -1.61 + 1.919x_{i1} + 0.006x_{i2} + 0.015x_{i1}x_{i2}$$

where x_1 is a female indicator and x_2 is the fare paid Exponentiate again:

$$\frac{p_i}{1 - p_i} = (0.20)(6.81)^{x_{i1}}(1.006)^{x_{i2}}(1.015)^{x_{i1}x_{i2}}$$

For men:

$$\frac{p_i}{1 - p_i} = (0.20)(1.006)^{x_{i2}}$$

For women:

$$\frac{p_i}{1 - p_i} = (0.20)(6.81)(1.006)^{x_{i2}}(1.015)^{x_{i2}} = (1.36)(1.021)^{x_{i2}}$$

The exponential of the gender effect (6.81) gives us the level odds ratio between genders, while the exponential of the interaction term tells us how much lower/higher in a multiplicative sense the odds ratio between survival and fare is for women than men. In this case, gender differences in survival increased with fare.

- Multivariate: change in the odds ratios holding all the other variables constant
- Let's look at a full example from the Titanic data.

Variable	Coefficient	SE	t-statistic	p-value
Intercept	-0.0514	0.2139	-0.2404	0.8101
Passenger class				
First Class (ref)				
Second Class	-1.6830	0.3286	-5.1224	0.0000
Third Class	-1.6235	0.2906	-5.5871	0.000
Gender				
Male				
Female	3.9976	0.5031	7.9452	0.000
Gender*Class				
Female*2nd Class	0.0611	0.6379	0.0958	0.9237
Female*3rd Class	-2.5360	0.5495	-4.6149	0.0000
Age (mean-centered)	-0.0454	0.0073	-6.2560	0.0000
Fare (mean-centered)	0.0004	0.0021	0.2003	0.8412
# Siblings/Spouses	-0.3349	0.1010	-3.3163	0.0009

- What is the intercept telling us? The reference group is a first-class man at the mean age and who payed the mean fare and had no siblings or spouse on board. This man had a log-odds of surviving of -0.0514. This translated into an odds of:

$$\text{odds of survival} = e^{-0.0514} = 0.9499$$

This translates into a probability of surviving of:

$$\text{probability of survival} = 0.9499 / (1 + .9499) = 0.487$$

- Let's plug in the following values: a second-class 35 year-old woman with a husband who payed 15 pounds for her fare. What are the predicted odds of a person with these values surviving? How about the probability.

$$e^{(-0.0514 - 1.6830 + 3.9976 + 0.0611 - 0.0454 * (35 - 29.85) + 0.0004 * (15 - 33.29) - 0.3349 * 1)} = e^{1.7483} = 5.745$$

This leads to a probability of 0.887.

- Let's change age to 25. How does this change the odds? how does it change the probabilities?

$$e^{(-0.0514-1.6830+3.9976+0.0611-0.0454*(25-29.85)+0.0004*(15-33.29)-0.3349*1)} = e^{2.202} = 9.043$$

This leads to a probability of 0.900. Note that the difference in the odds is much greater than the difference in the probability. The difference in the odds is a factor of $e^{0.0454*10} = 1.575$. This is because a one-unit change in age leads to a multiplicative change in the odds of $e^{0.0454}$.

- Be careful with the interaction term. Because it is in there, the effects of gender and passenger class alone only represent the effects for the reference group. So the passenger class variables give you the log-odds difference in survival between 1st and 2nd (-1.68) and 1st and 3rd (-1.62) **male** passengers.

To get the effects for female passengers, we need to add on the female interaction terms, so the log-odds difference in survival between female 1st and 2nd class passengers is (-1.68+0.061) and between female 1st and 3rd class passengers is (-1.62-2.536).

You would similarly need to calculate the difference between men and women within each class. 1st (3.998), 2nd (3.998+.061), 3rd (3.998-2.534). Here we see that the difference in survival between men and women dropped off considerably in the 3rd class.

- Statistical Inference

- Remember our friend, statistical inference? Our estimates in the logistic regression model are based on a sample and yet we want to measure the values for the population. Therefore, we need some measure of how secure we are in these numbers, just like for OLS regression.
- The GLM method for estimating the coefficients handily also returns estimated standard errors for their sampling distribution.
- Like OLS regression the sampling distribution of these coefficients is a t-distribution, so:

$$t = \frac{b_1}{s_{b_1}}$$

For all intents and purposes, our tests of statistical significance are identical in the case of logistic regression as they were for OLS regression, once we have these t-statistics:

- * If sample is large enough, then t-statistic is roughly normal ($|t| > 1.96$ means $p < .05$).

- * Like OLS regression our interest is in whether the parameter is distinguishable from zero.
 - * In this case , zero in the log-odds means one in the odds.
 - * We have to think about one-sided and two-sided tests in the same way.
 - * We can construct confidence intervals around our estimates.
- Let’s try a couple of examples.
- * Is the effect of fare paid ”statistically distinguishable” from zero at the 5% level?

$$t = 0.200$$

$$p - value = 0.84$$

There is an 84% chance that we would observe a value of 0.0004 or larger on a sample of this size just by random chance. We cannot distinguish the effect of fare from zero. Note that we found fare to be important earlier before we controlled for passenger class. The basic point is that most of the important information on survival contained in the fare variable is better picked up by knowing each passenger’s class.

- * Construct a 95% confidence interval for the odds ratio of the age variable

$$-.0454 \pm .0072 * 1.962 = (-.0595, -.0312)$$

$$(e^{-.0595}, e^{-.0312}) = (0.9422, 0.9692)$$

- Assessing model fit

- There is no R^2 measure for models with categorical dependent variables, because we can’t think about explaining the variation in a categorical dependent variable the same way we do for continuous dependent variables.
- People have created measures of ”pseudo- R^2 for these models, which basically measure the proportionate reduction in deviance of the current model over the null model.
- We can use McFadden’s version, D :

$$D = \frac{G_0^2 - G_c^2}{G_0^2}$$

Our titanic model produces a deviance of 919.27. The null model produces a deviance of 1413.57. Thus,

$$D = \frac{1413.57 - 919.27}{1413.57} = 0.35$$

- It is best to use "pseudo- R^2 with caution, if at all.
- We can use the likelihood ratio test to compare models however. In our previous example, we have reduced the deviance by $1413.57-919.27=494.3$ with the addition of 8 parameters. If the null model were accurate then that 494.3 would come from a χ^2 distribution with 8 degrees of freedom. Clearly the likelihood of such an extreme value is absurdly small. Thus, our model is a clear improvement over the null model.
- We can also compare our model using BIC' . Remember BIC' judges our model relative to the null model and is a tougher judge than the LRT test. STATA reports BIC which is judged relative to the saturated model – this is generally a meaningless term for individual level data.

$$BIC' = (G^2 - G_0^2) + p \log n$$

In our case:

$$BIC' = (919.27 - 1413.57) - 9 \log(1045) = -494.3 + 62.6 = -431.7$$

BIC' is negative indicating that in this case, it prefers our current model to the null model.

- How about the interaction term? Was this a valuable addition to the model. A model without the interaction between gender and class had a deviance of 969.97 on 1038 degrees of freedom.
 - * Well, we could just judge additions based on their significance level, but this technique doesn't work very well when our addition includes more than one variable.
 - * We could also do an LRT test. With two less degrees of freedom, the deviance was reduced from 919.27 to 969.97, for a change of:

$$969.97 - 919.27 = 50.7$$

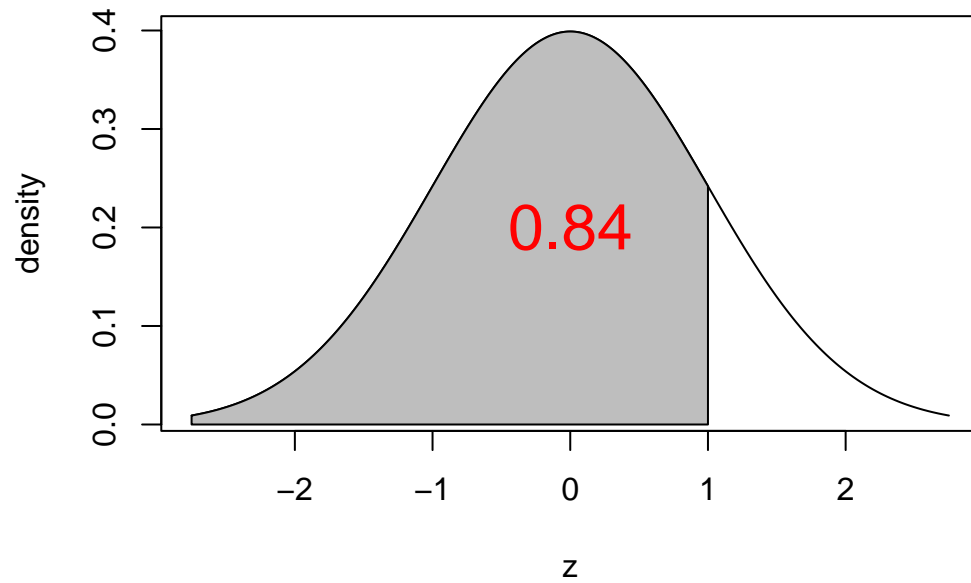
Is this change unlikely on a χ^2 distribution with two degrees of freedom? There is a very small probability of getting this much of a reduction in deviance by random chance.

- * We could also use the tougher BIC' . In this case, we would get a BIC' of -402.3. This BIC' is not as negative as our previous BIC' therefore we would prefer the previous model with the interaction term included.

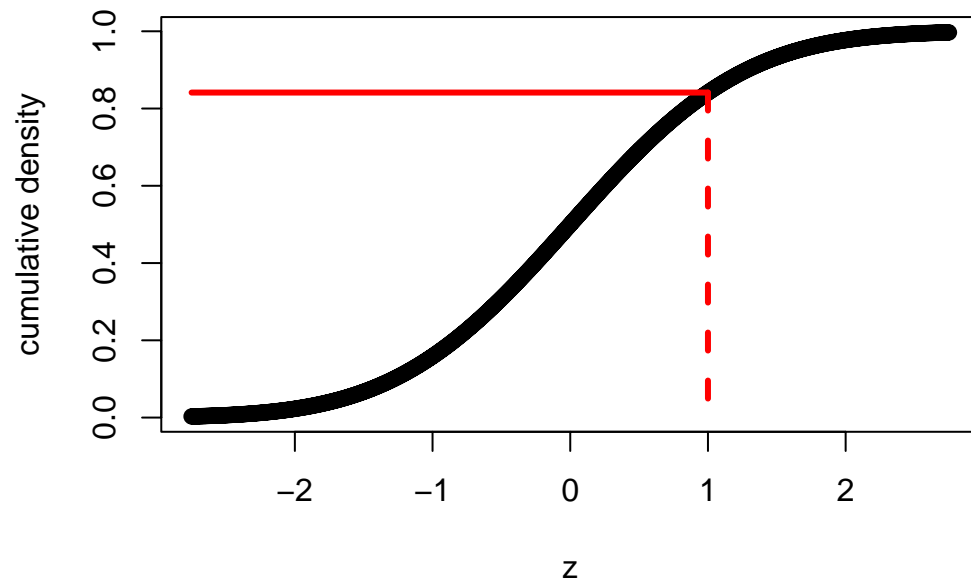
3.3 Other types of link functions (probit)

- The **probit** function

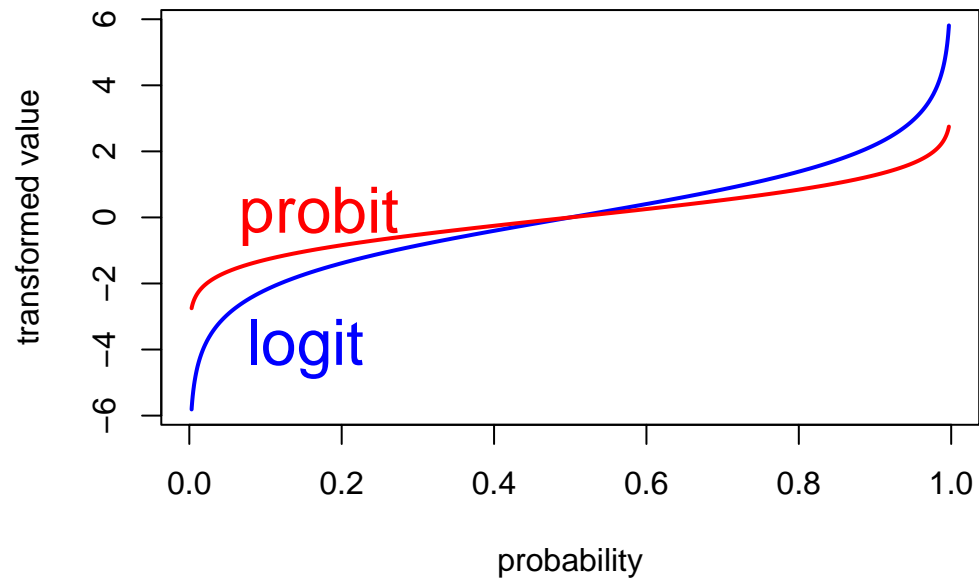
- The probit function is the inverse of the standard cumulative normal distribution. Straight-forward, right? let's break it down:
- What is the standard cumulative normal distribution. This is the area to the right of the value z on a standard normal distribution. Let's say we wanted to find the area to the right of $z = 1$ on the standard normal.



This can also be graphed as a **cumulative density function** of the normal.



- Taking the inverse is like saying, for what z-score is p% of the area of the standard normal below? Thus, the probit is something like a z-score.
- Like the logit, the probit stretches out from $(-\infty, \infty)$, and is symmetric around $p = .5$.



- Unlike the logit, the probit does not lend itself to easy interpretation. Generally, it will produce coefficients of similar size to the logit, and we can rely on directionality, but can't give a very precise definition. Therefore, there is little reason to prefer it over the logit link. (Economists use it because they think binary outcomes are a result of a latent variable which is continuous and normal).

Variable	OLS	IRWLS	Logit GLM	Probit GLM	Clog-log GLM
Intercept	0.434	0.444	-0.050	-0.061	-0.537
Passenger Class					
First Class (ref)					
Second Class	-0.265	0.026	-1.683	-0.977	-1.235
Third Class	-0.264	-0.138	-1.623	-0.932	-1.152
Gender					
Male (ref)					
Female	0.591	0.799	3.998	2.175	1.985
Gender*Class					
Female*2nd Class	0.144	-0.368	0.061	0.209	0.69
Female*3rd Class	-0.298	-0.661	-2.536	-1.293	-0.77
Age (mean-centered)	-0.006	-0.007	-0.045	-0.025	-0.023
# siblings/spouses	-0.046	-0.034	-0.335	-0.19	-0.207

3.4 Multinomial logit models

- Let's return to our example from last semester where we used years of education to predict abortion attitudes. Respondents were asked to respond whether they strongly disagree (1), disagree (2), had no opinion (3), agree (4), or strongly agree (5) that a woman should be able to seek an abortion for any reason. Previously we treated these responses as integer scores and used OLS regression to get:

Variable	Coefficient	SE
Intercept	1.912	0.188
Education	0.079	0.014

- We now realize that it is a bit crude to treat what is clearly a categorical response as a continuous response. We know how to put a dichotomous outcome on the left-hand side (logistic regression), and we know how to model association when both variables are categorical (log-linear models), but how do we deal with a polytomous dependent variable?
- What if we broke this question up into four different contrasts.
 1. For five categories, create four contrasts: (1) strongly disagree vs. disagree, (2) strongly disagree vs. no opinion (3) strongly disagree vs. agree, and (4) strongly disagree vs. strongly agree. (notice I kept one of the categories in each contrast the same).

2. For each contrast run a logistic regression on the individual-level data with the same independent variables: in our case years of education. Be sure to only use individuals who fall into one of the categories being contrasted.
3. The results will tell you how education affects the likelihood of being in each category vs. the reference (strongly disagree). Formally, you will have a set of β_j for the j the contrast. Then,

$$\log(p_{ij}/p_{i1}) = \beta_{0j} + \beta_{1j}x_i$$

- If we do this for our data we get the following values:

	Disagree vs. Strongly Disagree	No Opinion vs. Strongly Disagree	Agree vs. Strongly Disagree	Strongly Agree vs. Strongly Disagree
Intercept	0.498	-0.865	-0.728	-2.079
Education	-0.046	-0.003	0.063	0.139

- What are these parameters telling us?
 - The intercept tells us the expected odds of falling into the given category vs. the reference category when a person has zero years of education.
 - The slope tells us how the log-odds of falling into the given category vs. the reference changes with every year of education. ’
 - It is clear that the effect of education is not uniformly to increase one’s tolerance at all levels of the dependent variable. It also seems to create greater polarization.
- The model we have fit is not quite right because we don’t correctly specify the error distribution of the polytomous dependent variable. We can do this more effectively by specifying a **multinomial logit** model. In practice, it produces very similar results as before.

	Disagree vs. Strongly Disagree	No Opinion vs. Strongly Disagree	Agree vs. Strongly Disagree	Strongly Agree vs. Strongly Disagree
Intercept	0.542	-0.864	-0.708	-2.131
Education	-0.049	-0.003	0.062	0.143

- I have already shown you the form of this equation. We are fitting the log-odds of membership in each category of the dependent variable vs. some baseline category as a linear function of covariates:

$$\log(p_{ij}/p_{i1}) = \beta_{0j} + \beta_{1j}x_{i1} + \beta_{2j}x_{i2} + \dots + \beta_{pj}x_{ip}$$

where i is the i th individual and j is the j th category of the dependent variable. It is necessary to make one of the categories the baseline category ($j = 1$)

- There are two important cautions in interpreting coefficients from a multinomial model:
 1. In our previous models, each covariate had only one coefficient. Now each covariate will have $J - 1$ coefficients: one for each contrast.
 2. The decision about which category to set as baseline is arbitrary. It will not affect the overall fit of the model, but will affect interpretation. To get the coefficient for the contrast between j and j' :

$$\beta_j - \beta_{j'}$$

For example, let's say we were interested in the effect of education on the log-odds of being in the agree vs no opinion group.

$$0.062 - (-0.003) = 0.062 + .003 = 0.065$$

And the effect of education on the log-odds of being in the strongly agree vs. agree group is

$$0.143 - 0.062 = 0.081$$

This can also be done by rerunning the model with a different reference group.

3.5 Ordered Logit Models

- Ordered Logit Models
 - We have been thinking of attitudinal responses to abortion as a set of five unordered responses, but there is a very clear and intentional ordering to these responses: Strongly disagree, disagree, no opinion, agree, and strongly agree.
 - We call categorical variables that can be ordered **ordinal variables**. If we know that a category is ordinal then there are special models that tell us how independent variables relate to someone being higher or lower on the scale.
 - If we had a way of giving some underlying score to each level then we could fit the model using basic OLS regression. There are different approaches to **scoring techniques**.
 - * The simplest, but often problematic, method would be to use **integer scoring**. We assign the lowest score a 1 and then move up by an integer for every category. This method is very crude because we assume that the categories are equally spaced.

- * Another approach would be to use what's called a **normal score transformation**. Here we assume that the underlying categories are discrete realizations of some underlying continuous distribution of attitude. If we assume that this distribution is normally-distributed, we can assign each category a z-score based on the proportion of responses at its midpoint.

Category	Integer Score	Proportion	Cum. Prop.	Mid-point	Z-score
Strongly Disagree	1	0.235	0.235	0.117	-1.189
Disagree	2	0.216	0.45	0.342	-0.406
No Opinion	3	0.095	0.545	0.498	-0.006
Agree	4	0.262	0.807	0.676	0.456
Strongly Agree	5	0.193	1	0.903	1.301

- * Note that the distance between no opinion and agree and disagree is less than the distance between the strongly disagree/agree categories and the disagree/agree categories.

- * If we plug these scores into an OLS regression, we get

Coef	Integer Score Estimate	Z-score Estimate
Intercept	1.912	-0.595
Education	0.079	0.045

- We don't need to use this scoring approach, however. We can use the **ordered logit model** so that we can use the categories directly as our dependent variable.
- The ordered logit model depends upon the idea of the **cumulative logit**. This in turn relies on the idea of the **cumulative probability**. You can think of the cumulative probability C_{ij} as the probability that the i th individual is in the j th or higher category:

$$C_{ij} = Pr(y_i \leq j) = \sum_{k=1}^j Pr(y_i = k)$$

We can then turn this cumulative probability into the cumulative logit:

$$\text{logit}(C_{ij}) = \log(C_{ij}/(1 - C_{ij}))$$

- Our ordered logit model simply models the cumulative logit as a linear function of independent variables:

$$\text{logit}(C_{ij}) = \alpha_j - \beta x_i$$

- Note that there is a different intercept for each level of the cumulative logit, but that β does not vary by the level of the cumulative logit. Also note that β is subtracted rather than added. This means:

- * each α_j indicates the logit of the odds of being equal to or less than category j for the baseline group (when all independent variables are zero). Thus, these intercepts will increase over j . These intercepts are sometimes referred to as **cutpoints**.
- * The β tells us how a one-unit increase in the independent variable increases the log-odds of being higher than category j (due to the negative sign). Because, this β is not indexed by j we are assuming that the one unit increase affects the log-odds the same regardless of which cut-point we are considering.

– Let’s take our example with attitudes:

Coef	Estimate
Cutpoints	
Strongly Disagree	0.049
Disagree	1.045
No Opinion	1.440
Agree	2.708
Education	0.095

– What do the cutpoints tell us. They tell us about the expected cumulative distribution of answers for individuals with zero years of education. If we exponentiate these parameters, we will get the cumulative odds, and if take $O/(1 + O)$ we will get the cumulative probabilities.

	S. Disagree	Disagree	No Opinion	Agree	S. Agree
Cutpoints (α_j)	0.049	1.045	1.440	2.708	-
Cumulative Odds (e^{α_j})	1.05	2.84	4.22	15.00	-
Cumulative Probability	0.51	0.74	0.81	0.94	1
Distribution	0.51	0.23	0.07	0.13	0.06

– The coefficient for education tells us how the log-odds of each of these cutpoints **decreases** with a one year increase in education. The positive value indicates that one year of education increases the odds of being in a higher category.

What would be the predicted odds of having no opinion or a less tolerant view for someone with a high school degree:

$$e^{1.440-0.095*12} = 1.35$$

How about for someone with a college degree (16 years)

$$e^{1.440-0.095*16} = 0.92$$

We can reverse this and ask what are the odds they will have more than no opinion. To do this we simply take the inverse of these odds. So the odds of having a more tolerant attitude than "no opinion" are 1.09 for college graduates and 0.74 for high school graduates.

- Be careful. Because the ordered logit forces β to be the same across all cutpoints, differences in the importance of education across different boundaries will not be observed. Since we noted some important differences in the multinomial model, an ordered logit model might not be the best alternative in this case.

4 Event History Analysis

4.1 Events, rates, and relative risk

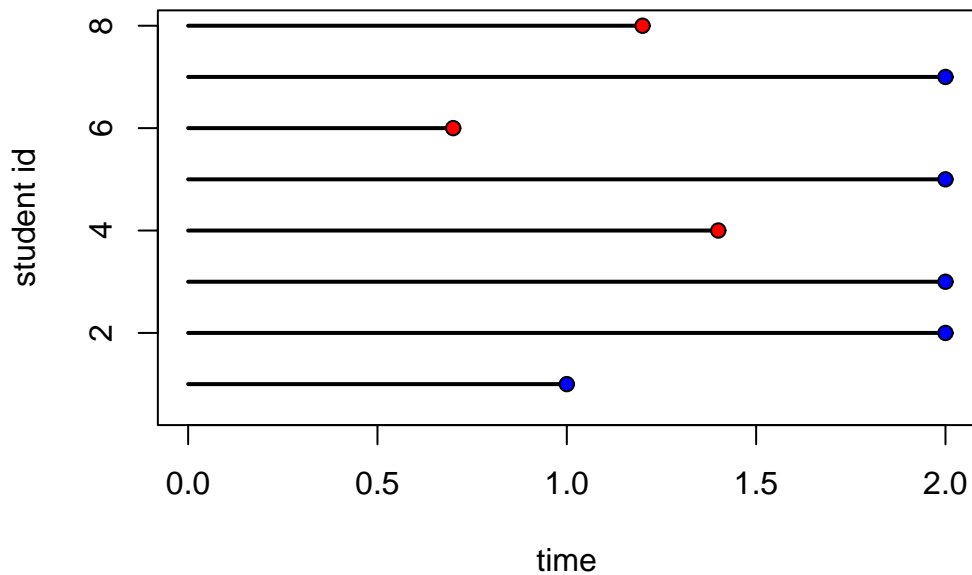
- When we have information on some units over a period of time (either from retrospective data or panel data), we may be interested in studying the transitions from one state to another.
 - from living to dying
 - from single to being married
 - the passage of a law
 - employment changes
 - high school drop out
- This kind of modeling is known by many names (hazard models, event-history models, duration analysis, etc.)
- The underlying idea is that we would like to know how the underlying **rate** of an event occurring is affected by covariates. (i.e. does being black increase your risk of becoming a high school drop out).
- To teach you event-history models, you must first learn a little demography.
- What is a rate? A RATE IS NOT A PROBABILITY. A true rate is the number of events that occur over the overall exposure to risk (as a unit of time).

$$rate = \frac{Event}{Exposure}$$

- Let's take a simple example. We have observed 8 high school students over 2 school years. 4 of these students are white and four students are black.

Student	Race	Exposure time	Dropped out
1	White	1	N
2	White	2	N
3	White	2	N
4	White	1.4	Y
5	Black	2	N
6	Black	0.7	Y
7	Black	2	N
8	Black	1.2	Y

- We have the exact time we observe each of these students in the system as well as an indicator variable indicating that they dropped out. For individuals that did not drop out, we have **censored observations**. Notice that student 1 did not drop out, but he does not have the full two years of exposure. Why might that be? (he graduated)



Red dots indicate the observation exited due to an event. Blue dots indicate that the observation was censored.

- Let's calculate the overall rate. How many events? How much exposure to risk?

$$\text{Exposure} = 1 + 2 + 2 + 1.4 + 2 + .7 + 2 + 1.2 = 12.3$$

$$\text{Rate} = 3/12.3 = 0.24$$

This is not a probability!!! It means we expect 0.24 drop-outs per **person-school year** spent in the system. It is possible to have rates greater than one.

- Now let's get a separate rate for the white and black students.

$$\text{Exposure (black)} = 2 + .7 + 2 + 1.2 = 5.9$$

$$\text{rate (black)} = 2/5.9 = 0.3390$$

$$\text{Exposure (white)} = 1 + 2 + 2 + 1.4 = 6.4$$

$$\text{rate (white)} = 1/6.4 = 0.1563$$

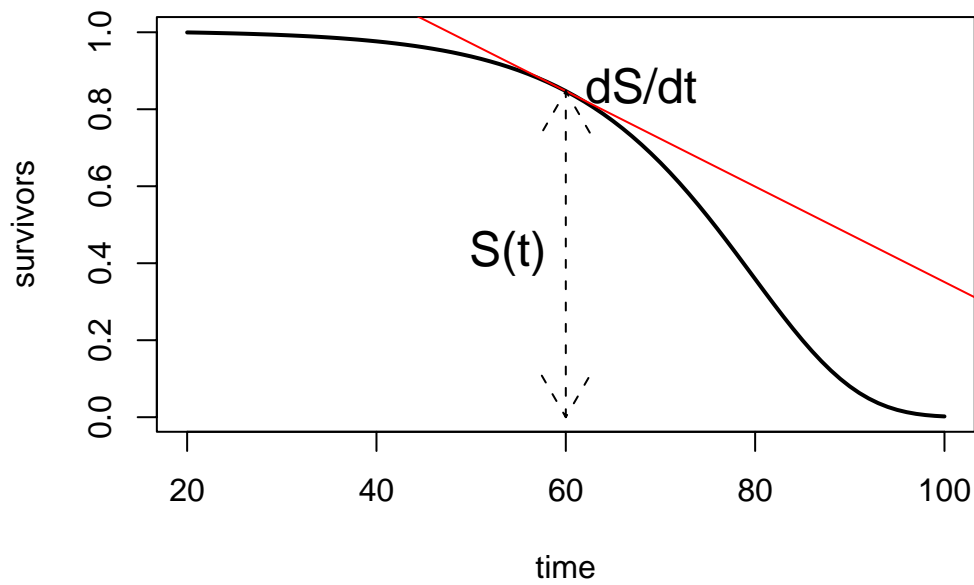
- To compare these two rates, we can calculate what is called the relative risk (RR) which is simply the ratio of the two rates:

$$RR = 0.3390/0.1563 = 2.17$$

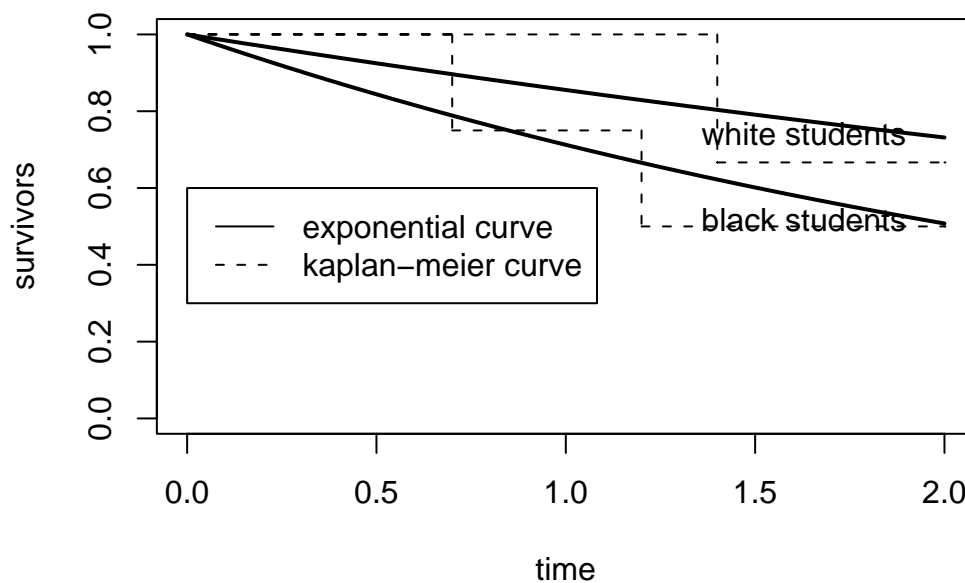
We see that the black students in this example are twice as likely to drop out as the white students.

- The rate of drop-out is a form of the **hazard rate** $\lambda(t)$. The hazard rate is derived from what is called the **survival curve** ($S(t)$). The survival curve gives the proportion of respondents who are still "alive" at time t . The hazard rate at exact time t is the instantaneous decrement at time t divided by the exposure ($S(t)$) at time t .

$$\lambda(t) = \frac{dS/dt}{S(t)}$$



- In our previous example, we were assuming that this hazard rate is constant over t . It can be shown from the previous equation, that if $\lambda(t) = \lambda$, then $S(t) = e^{-\lambda t}$ This is just the exponential curve.
- We can show these curves for the black and white students by plotting the two equations:
 black students: $S(t) = e^{-0.3390t}$
 white students: $S(t) = e^{-0.1563t}$



The dotted lines show what are called **Kaplan-Meier curves**. These lines plot a staircase function based on the actual decrements from the population. You can see that the exponential curve fits to the Kaplan-Meier curve as best it can.

4.2 Parametric survival models

- Our example from last time compared the drop-out rate between black and white high school students over a two-year window.
- This is an example of event history analysis. We are essentially asking how the drop-out rate is affected by a covariate (in this case race).
- We could express our relationship as follows:

$$\lambda_i(t) = \beta_0 + \beta_1 x_i$$

x_i is a dummy indicating that the student is black. This model says that the individual hazard rate is a function of race. White students will have a drop-out rate of β_0 and black students

will have a drop-out rate of $\beta_0 + \beta_1$. Since, we are assuming that the hazard rate is constant across time:

$$\lambda_i = \beta_0 + \beta_1 x_i$$

- The hazard rate will always be positive. Therefore, we will generally prefer a model that is linear in the log of the hazard rate:

$$\log(\lambda_i) = \beta_0 + \beta_1 x_i$$

In this format, the drop-out rate for white students is e^{β_0} and the relative risk for black students is e^{β_1} . We calculated these values by hand last session.

$$\beta_0 = \log(\text{white student drop-out rate}) = \log(0.1563) = -1.86$$

$$\beta_1 = \log(\text{relative risk for black students}) = \log(2.17) = 0.77$$

So,

$$\log(\lambda_i) = -1.86 + 0.77x_i$$

- From this model, we can develop a general model for the hazard rate:

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}$$

For categorical variables, the exponential of the coefficient gives the relative risk for the groups. For continuous variables, the exponential of the coefficient gives the relative risk between two individuals who only differ by one unit on the independent variable. β_0 gives the hazard rate for the baseline group.

- How do we estimate such models?
 - The secret lies in breaking λ_i into it's component pieces:

$$\lambda_i = \frac{\mu_i}{t_i}$$

Where μ_i is the expected number of events for observation i and t_i is the length of time observation i was observed.

- Plugging, this back into our equation:

$$\log\left(\frac{\mu_i}{t_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}$$

- By algebraic rules:

$$\log(\mu_i) - \log(t_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}$$

$$\log(\mu_i) = \log(t_i) + \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}$$

- We are trying to estimate the log of the mean number of events for the i th individual as a linear function of the independent variables. The only wrinkle is the $\log(t_i)$ on the right hand side. This is referred to as an **offset variable** - i.e. a variable whose coefficient we are assuming equals one.
- The only remaining trick is how to model μ_i . This is an expected number of events (i.e. counts). That's right, this is a straightforward GLM with a poisson error, log link, and an offset of the log of exposure time.
- So in order to model this we treat the number of events for each individual (in most cases with individual data, it will be one or zero) as the dependent variable and relate the independent variables to it using a log link, a poisson error distribution, and log-exposure as an offset. If we apply this to our school data, we get:

Variable	Coefficient (SE)
Intercept	-1.86 (1.00)
Black	0.77 (1.22)

This is precisely what we derived by hand earlier.

- Parametric models

- Our example treats the hazard rate as if it was constant over the duration time.

$$\lambda(t) = \lambda$$

This assumption leads to a an exponential survival curve:

$$S(t) = e^{-\lambda t}$$

- This is one form of a **parametric survival model**. We refer to it as parametric because we are assuming some underlying shape to the survival curve. The exponential model is the simplest form of parametric survival model.
- Most parametric survival models assume **proportional hazards**. That is they assume that there is an underlying hazard rate over time, and differences in the covariates simply

lead to differences in the relative hazard rate at a point in time. in other words, they assume no interaction between time and covariates. Mathematically,

$$\lambda_i(t) = h_0(t)e^{\beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}}$$

Where $h_0(t)$ is some time function of the hazard rate. Because the hazard rate depends on the length of times spent in the state, it is called **duration dependent**.

– Different specifications of this $h_0(t)$ lead to different parametric survival models.

* The Weibull model:

$$h_0(t) = \lambda p(\lambda t)^{p-1}$$

The Weibull distribution leads to hazard rates which either increase or decrease monotonically over time. If p is greater than one, the hazard rate is increasing, if p is less than one, the hazard rate is decreasing, and if $p = 1$, we have a constant hazard rate.

* The Gompertz model:

$$h_{0t} = \lambda e^{rt}$$

The hazard rate increases or decreases at an exponential rate. This is a useful model for mortality at ages past about 30, because mortality rates tend to follow this pattern.

– These models are all more complex to estimate than the exponential model because, they require modeling the parameters of the hazard function as well as the covariates, but intuitively they are similar. Choosing between them often requires a diagnostic analysis such as the Kaplan-Meier curve. (what do we think about our high school students).

Let's try fitting these parametric survival models to our data:

Model	log(RR)	RR
Exponential	0.77	2.17
Weibull	0.79	2.21
Gompertz	0.78	2.18

• Adding time-varying covariates

- One of the great advantages of longitudinal analysis is that we can observe not only changes over time in the state of interest, but also change over time in the characteristics which might predict the state
- In the drop out example, what are some covariates that might affect drop-out rates and change over time.

- * Grades last term
 - * Disciplinary problems
 - * Grade in school
 - * Romantic relationships
- We would like to be able to incorporate these changes into our estimates of the effect of the covariates. There is a straightforward technique for doing this. We have to change the format of our dataset into a **split-episode event** data set.
1. Break the overall interval for individual i into sub-intervals broken by each point where one of individual i 's covariates changed.
 2. record the value of each observed covariate at the start of each sub-interval.
 3. Treat each sub-interval as an observation in itself. If there are lots of changes, each individual will contribute many observations to the data. This will not artificially inflate the size of your sample.
- Let's take grades in math as an example for our high school students. Since grades are recorded each semester, we will get 2 observations per school year for each individual.

Student	Race	Previous Grade	Semester	Exposure time	Dropped out
1	White	A	Fall	.5	N
1	White	A	Spring	.5	N
2	White	B	Fall	.5	N
2	White	A	Spring	.5	N
2	White	B	Fall	.5	N
2	White	C	Spring	.5	N
3	White	B	Fall	.5	N
3	White	B	Spring	.5	N
3	White	A	Fall	.5	N
3	White	B	Spring	.5	N
4	White	D	Fall	.5	N
4	White	C	Spring	.5	N
4	White	D	Fall	.4	Y
5	Black	B	Fall	.5	N
5	Black	A	Spring	.5	N
5	Black	B	Fall	.5	N
5	Black	C	Spring	.5	N
6	Black	D	Fall	.5	N
6	Black	F	Spring	.2	Y
7	Black	B	Fall	.5	N
7	Black	B	Spring	.5	N
7	Black	C	Fall	.5	N
7	Black	B	Spring	.5	N
8	Black	C	Fall	.5	N
8	Black	C	Spring	.5	N
8	Black	D	Fall	.2	Y

- Now I can run a model which includes both a grade variable and a race variable. I could also include a semester variable to see whether drop-out rates are higher in the fall or spring. Let's include a quantitative grade variable which goes from 0 (F) to 4 (A).

Variable	Coefficient (SE)
Intercept	4.49 (2.29)
Black	-0.32 (1.41)
Letter Grade	-2.26 (0.95)

Once grades are controlled for, the higher dropout rate of black students goes away (of course this data is made up).

- In many cases, covariates will be recorded at regular intervals such as this, meaning that splitting the data will be easier.

4.3 Semi-parametric models

- The assumption of a parametric form to the survival curve is a serious one that can often be problematic. Two **semi-parametric methods** have been developed to reduce the importance of this assumption.
- Piecewise constant exponential model
 - This model assumes that the hazard is constant not over the whole range of time, but within certain specified intervals of time. So, the constant λ from the exponential model becomes λ_j where j is some interval of time.

$$\log(\lambda_{ij}) = \alpha_j + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}$$

Each α_j gives the constant hazard within time interval j for the baseline group.

- If the time units are small enough or the true hazard rate is not changing too dramatically, then the piecewise constant model often gives very acceptable results.
- The piecewise constant model approach allows the hazard rate to change in non-parametric ways.
- Implementing this model is straightforward. The time intervals are just treated as time-varying covariates. The researcher splits each observation into the J time intervals and estimates an exponential models with each time interval as a dummy variable.
- In our example, let's assume that each student begins the study in the 10th grade. We have reason to believe that the drop-out rate may differ in the 10th and 11th grade, so we run a piecewise constant model. Our split episode file looks like this:

Student	Race	Grade	Exposure time	Dropped out
1	White	10	1	N
2	White	10	1	N
2	White	11	1	N
3	White	10	1	N
3	White	11	1	N
4	White	10	1	N
4	White	11	.4	Y
5	Black	10	1	N
5	Black	11	1	N
6	Black	10	0.7	Y
7	Black	10	1	N
7	Black	11	1	N
8	Black	11	1	N
8	Black	11	.2	Y

– Now we just throw a grade dummy into our poisson GLM as well as a race dummy:

Coefficient	Constant	Piecewise Constant
Intercept	-1.86	-2.49
Grade		
10th (ref)		
11th		1.21
Race		
White (ref)		
Black	0.77	0.78

(go through and interpret these results)

- Cox proportional hazards model

- Probably the most popular model for estimating event history is the Cox proportional hazard model, also called **Cox regression**.
- Cox regression is popular because it makes no assumptions about the underlying hazard function. It simply assumes that relative risks are constant across all times (thus "proportional hazards").

$$\lambda(t) = h_0(t)e^{x'_{it}\beta}$$

Where $h_0(t)$ can be **any** hazard function.

- The set up for the Cox regression is similar to that for other models. The model requires the value of covariates, the status of the individual at the end of the interval, and the length of the interval. Time-varying covariates can be added in the normal way.
 - Estimation of the Cox regression model is complex and requires a technique called **partial likelihood** rather than MLE.
 - The basic idea is that Cox regression uses information on the rank-ordering of survival times to look at the likelihood that the i th individual will be the next to experience an event.
 - Caution: when there are many tied event times, Cox regression can give biased results.
- Here is a comparison of the relative risk for black students under various models

Model	log(RR)	RR
Exponential	0.77	2.17
Weibull	0.79	2.21
Gompertz	0.78	2.18
Piecewise Constant	0.78	2.18
Cox	0.83	2.30

4.4 Discrete time approximation

- Although it is nice to have information on the exact timing of events, we are often not so lucky. It is very common in longitudinal datasets to have information on an individual's status at regular intervals of time, but not information on when status changed in the interval for cases where it changed. Examples:
 - labor force status
 - school status
 - marital change
- These are cases of what is often called **discrete time**. In actuality, the data are **discrete time approximations** of a continuous time process. That is, we believe that there is some continuous hazard of a change in labor force status between our recording intervals, but all we know is whether the change did or did not occur during this interval.
- Because we don't have information on exact timing, we cannot directly calculate hazard rates, but we can calculate the probability of an event occurring in the interval.

- What is the relationship between the probability and the rate? If we assume that the hazard rate is constant over the interval, then the probability of an event for individual i in time period t is given by:

$$p_{it} = 1 - e^{-\lambda t_i}$$

Where t_i is the length of the time period, and λ is the constant hazard rate. This follows directly from the survival curve:

$$p_{it} = 1 - S(t_i) = 1 - e^{-\lambda t_i}$$

- If we relate our covariate x_{it} linearly to the log of λ as before, then

$$\lambda = e^{\beta_0 + \beta_1 x_{it}}$$

Then plugging into our previous equation:

$$p_{it} = 1 - e^{-e^{\beta_0 + \beta_1 x_{it}} t_i}$$

This is ugly. What we want to is to get a linear expression on the right. Let's log both sides:

$$\log(p_{it}) = -e^{\beta_0 + \beta_1 x_{it}} t_i$$

Getting better. Let's multiply both sides by -1 and log again.

$$\log(-\log(p_{it})) = \log(t_i) + \beta_0 + \beta_1 x_{it}$$

On the left we have a link function for the probability known as the **complementary log-log function**. On the right we have a linear function of the independent variables, plus an offset of the log of the length of the time period.

So the discrete time approximation can be run directly on the dichotomous status variable for each wave using a GLM with a cloglog link and an offset of $\log(t_i)$. In many cases, the data are collected once a year, so that the offset will be $\log(1) = 0$ and need not be explicitly added.

- Duration dependence can be incorporated directly into this type of model by treating time (measured as the wave of data) as a covariate. It can be treated as a set of dummies or it can be modeled using a more parametric approach.
- True discrete time data

- There are a few (rare) cases where the researcher may believe that the process is truly governed by a discrete time process. The classical example would be the passage of state legislation. Whether a bill gets passed early or late in legislative session may not really give us any information on the hazard of the bill being passed overall.
- When the process is a true discrete time process, then the appropriate way to model it is using a logit or probit transformation rather than a complementary log-log transformation. Because, we want to model the probability of transition directly, rather than the rate.
- Many textbooks do not make a distinction between true discrete time processes and continuous time processes approximated with discrete time data, and many researchers use the logit or probit transformation when they should be using the cloglog transformation - but you know better,
- When the intervals are short and even for everyone the difference between cloglog and logit is fairly trivial, but cloglog will always give closer estimates to what the continuous time process would have produced than logit.
- Let's take our example. Let's say we didn't know the exact time students dropped out, but just whether they had dropped out by the end of the school year. First, we make our split-episode file

Student	Race	Length of Inverval	Dropped out
1	White	1	N
2	White	1	N
2	White	1	N
3	White	1	N
3	White	1	N
4	White	1	N
4	White	1	Y
5	Black	1	N
5	Black	1	N
6	Black	1	Y
7	Black	1	N
7	Black	1	N
8	Black	1	N
8	Black	1	Y

- Now let’s run a discrete time approximation using first the logit and the cloglog. How do the relative risks compare to our exponential model?

Model	Constant	log(RR)
Exponential	-1.86	0.77
Discrete logit	-1.79	0.88
Discrete cloglog	-1.87	0.78

4.5 Event count poisson regression

- Sometimes researchers directly model the count of repeatable events that occur in evenly-divided intervals of time (often years).
- For example, in a recent *Social Forces* article, Debrah Minkoff examined the founding of Civil Rights organizations during and after the Civil Rights period.
- This method is straightforward. The dependent variable is the count of events for an observation-time period (i.e. state-years). The covariates can be contemporary or lagged variables applying to the observation.
- Since counts go from 0 to ∞ , the dependent variable is usually logged and the error term is given by a poisson. So we have a straightforward GLM:

$$\log(\mu_{ij}) = \beta_0 + \beta_1 x_{ij1} + \dots + \beta_p x_{ijp}$$

Where μ_{ij} is the expected count of events in the given time interval.

- Although it is not usually seen as such, this is a form of a hazard model. It appears from the model structure that we are modeling counts directly, but in fact we are modeling the rate. The question is what is the exposure?
- Using this technique we are really modeling the rate of events per observation-time unit (in the case of Minkoff, nation-state-year). To show this, let’s use our usual form for the hazard model where λ_{ij} is the rate of events per observation time-unit for observation i in time period j :

$$\log(\lambda_{ij}) = \beta_0 + \beta_1 x_{ij1} + \dots + \beta_p x_{ijp}$$

As before, we can break the rate into the number of events (μ_{ij}) and the exposure (E_{ij})

$$\log(\mu_{ij}/E_{ij}) = \beta_0 + \beta_1 x_{ij1} + \dots + \beta_p x_{ijp}$$

$$\log(\mu_{ij}) - \log(E_{ij}) = \beta_0 + \beta_1 x_{ij1} + \dots + \beta_p x_{ijp}$$

$$\log(\mu_{ij}) = \log(e_{ij}) + \beta_0 + \beta_1 x_{ij1} + \dots + \beta_p x_{ijp}$$

E_{ij} is the exposure count. What is it in the case of event count models? It is generally 1 unit of time. Therefore:

$$\log(\mu_{ij}) = \log(1) + \beta_0 + \beta_1 x_{ij1} + \dots + \beta_p x_{ijp}$$

$$\log(\mu_{ij}) = \beta_0 + \beta_1 x_{ij1} + \dots + \beta_p x_{ijp}$$

- Event count models implicitly include an offset of $\log(1)$ to account for one year's risk of exposure. Thus these models are modeling the hazard rate for a particular kind of event as a function of covariates.
- As an example, lets return to our Prussian horsekick fatality data. Remember our data consist of the count of fatal horsekicks to the head in 14 Prussian military corps over 20 years, giving us 280 corp-years of observations altogether. Let's look at how the rate of horsekick fatalities differed by corp and over time.

Variable	Coefficient	SE	t-stat	p-value
Intercept	-0.407	0.280	-1.455	0.146
Corp				
G (ref)	-	-	-	-
I	0.000	0.354	0.000	1.000
II	-0.288	0.382	-0.753	0.451
III	-0.288	0.382	-0.753	0.451
IV	-0.693	0.433	-1.601	0.109
V	-0.375	0.392	-0.957	0.339
VI	0.061	0.348	0.174	0.862
VII	-0.288	0.382	-0.753	0.451
VIII	-0.827	0.453	-1.824	0.068
IX	-0.208	0.373	-0.556	0.578
X	-0.065	0.359	-0.180	0.857
XI	0.446	0.320	1.394	0.163
XIV	0.405	0.323	1.256	0.209
XV	-0.693	0.433	-1.601	0.109
year (origin=1875)	0.019	0.012	1.510	0.131

The baseline represents the log-rate for corp G in 1875. The rate was $e^{-.407} = 0.67$ hosekicks/corp-year for corp G in 1875. The results don't give us any conclusive evidence of differences across

corps or of a trend across years, although there may have been an increase in the rate over time.

- The difference between event count models and the hazard models we were using previously is that the events in event count models are repeatable. Thus, the populations observed don't change states with the occurrence of events. They are simply at risk of more events happening to them (like the Prussian army is subject to more fatal horse kicks to the head).
- In most cases, the researcher has the exact date on which events occurred and so these event count models could be set up more like our other hazard models, where each record would be the interval between events or until censoring. This type of model provides a better estimate of the underlying hazard rate and can incorporate autocorrelation better, but many researchers use the event count model instead because of its simplicity.

5 Multilevel Models

- Much of the data that we analyze in the social sciences has a hierarchical or "nested" structure.
 - Students within a school
 - Individuals within a neighborhood
 - Workers within a labor market
 - Firms within an industry
 - children within a household
- Nested can be more than two levels such as students within a school within a school district or children within a household within a village.
- The nested nature of the data allow us to ask a question that has prime sociological significance: how does larger context affect individuals - i.e. relating macro and micro processes.
- Each of the nesting structures above has been used to address contextual effects:
 - Students within a school - school resources effect on educational outcomes
 - individuals within a neighborhood - effect of neighborhood dynamics on criminal activity
 - Workers within a labor market - the effect of racial composition on racial wage inequality
 - Firms within an industry - industrial context on organizational form?
 - children within a household - household health effects on child mortality.

- There are two ways to think about how context might matter:
 - Macro-level context might directly predict a particular micro-level dependent variable.
 - Macro-level context might mediate the relationship between a micro-level independent and dependent variable.
- The first formulation can be thought of as a direct effect of context on an outcome.
- The second formulation can be thought of as an interaction term.
- Intuitively, contextual effects are straightforward. We should be able to just add them to our regression model as independent variables and interaction terms.
- Let's take as an example. We will use a sample of data from the National Educational Longitudinal Study (1988) to predict scores on a standardized math test. The data consist of 519 students nested within 23 schools. We know that parental SES (here measured on an abstract scale) is important for school success as is non-white status, so let's fit a model predicting math scores from parental SES.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

Where x_1 is SES score and x_2 is an indicator variable for non-white status. So far we have a straightforward OLS regression model. It shows that being non-white has a substantial negative effect, while SES has a large positive effect.

Variable	Estimate
Intercept	52.63 (0.47)
Non-White	-3.61 (0.98)
SES	5.42 (0.48)

- Now, we would also like to know how the racial composition of the school, operationalized as the percent minority, affects math scores. This is a direct macro-level effect that operates apart from each individual student's race. A simple approach would be to simply append the percent minority of the school to each student's record and then include it in the model.

$$y_{ij} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{j3} + \epsilon_{ij}$$

We now have the added j as a subscript for each school.

Variable	Estimate
Intercept	53.00 (0.65)
Non-White	-2.90 (1.29)
SES	5.50 (0.49)
% non-white	-0.022 (0.026)

The point estimate suggests that being in a minority-concentrated school has a negative effect on math scores, but the results are not statistically distinguishable from zero.

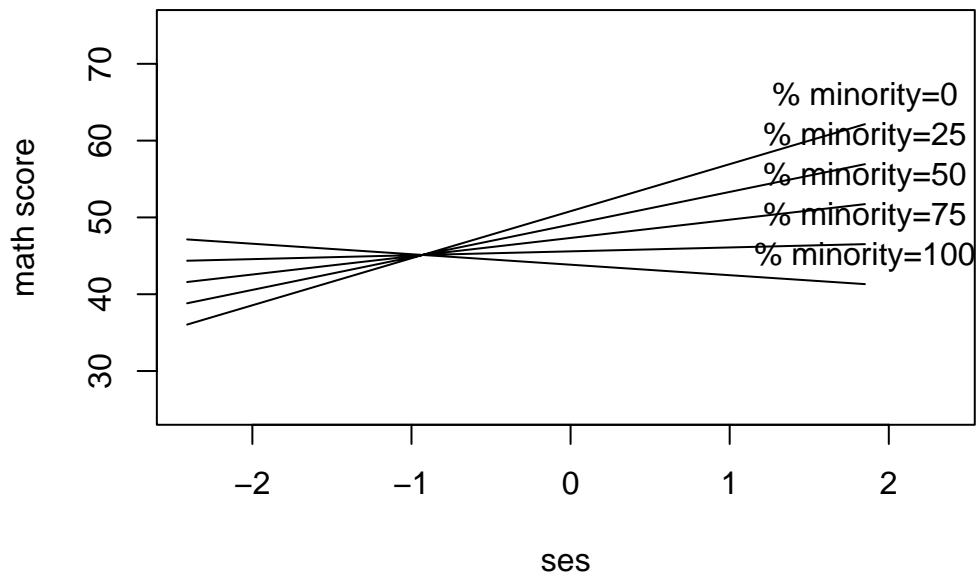
- Now let's take this one step further. We might suspect that the effect of each student's SES will differ depending upon the racial composition of the school. What if the affect of each student's SES differed by the average SES of the school? Then we need an interaction effect:

$$y_{ij} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{j3} + \beta_4 x_{i1} x_{j3} + \epsilon_{ij}$$

- Let's look at the parts of this model. Controlling for contextual effects, β_1 gives us the effect of each student's SES on math scores when minority percentage is zero. β_3 gives us the effect of minority composition on math scores when individual SES is zero. β_4 tells us the synergy between the two. A negative effect would indicate that individual SES is less important in high minority concentration schools (and vice versa).

Variable	Estimate
Intercept	53.58 (0.67)
Non-White	-3.48 (1.29)
SES	7.70 (0.83)
% non-white	-0.042 (0.026)
%non-white*SES	-0.079 (0.024)

These results suggest that the value of SES in achieving high math scores is reduced as the percentage minority in the school increases. This is shown graphically in the picture below:



- Make sure you understand this model. The extensions I will now introduce are largely for the purpose of generating more realistic standard errors - they don't change the underlying meaning of the regression.
- The problem with this model is that the students' percentage minority values are not individually drawn - they are clustered within sets of schools. Unless we take account of this clustering, our estimates of the standard error on these variables will be far too optimistic.
- So let's think about this a different way. Let me rewrite the micro-level model:

$$y_{ij} = \beta_{0j} + \beta_{1j}x_i + \beta_{2j}x_{i2} + \epsilon_i$$

Notice that this model only has the individual components, but I have subset the coefficients by j . This means that I expect different coefficients within each school. If I left the model in this form, it would be purely what is called a **random effects** model. These types of models don't assume that each individual has the same coefficient (fixed effect), but rather that the higher-level units draw their values from some distribution. In this case, students within the same school all draw the same value.

- In the random effects framework, the coefficients can be represented by school-level equations as well:

$$\beta_{0j} = \alpha_0 + \phi_j$$

$$\beta_{1j} = \alpha_1 + \omega_j$$

α_1 and α_2 give us the mean value for the intercept and SES effect, respectively. ϕ_j and ω_j give us the errors - the variance of these errors tells us how variable overall math achievement (β_{0j}) and the SES effect (β_{1j}) is across schools.

- In practice it is necessary to specify a distribution for the error terms in the school-level model (typically the normal distribution). This has important implications for the fitting of the model, because it forces the observed slopes across schools to conform to a pre-determined distribution. This can be useful because it will reduce the effect of outliers.
- This is all nice, but we want to expand on these equations by including school-level variables into our prediction of the β 's.

$$\beta_{0j} = \alpha_{00} + \alpha_{01}z_j + \phi_j$$

$$\beta_{1j} = \alpha_{10} + \alpha_{11}z_j + \omega_j$$

α_{01} tells us how minority percentage in a school affects overall math achievement. α_{11} tells us how minority percentage affects the effect of individual SES on overall math achievement.

- To see how this all fits together, plug these equations into our individual level equation.

$$y_{ij} = \alpha_{00} + \alpha_{01}z_j + \phi_j + \alpha_{10}x_i + \alpha_{11}x_iz_j + \omega_jz_j + \beta_2x_{i2} + \epsilon_i$$

$$y_{ij} = \alpha_{00} + \alpha_{10}x_i + \alpha_{01}z_j + \alpha_{11}x_iz_j + \beta_2x_{i2}(\phi_j + \omega_jz_j) + \epsilon_i$$

Don't be fooled by the fancy equations!! Except for the ϕ and ω terms, this equation is identical to our previous simplistic equation, $\beta_0 = \alpha_{00}$, $\beta_1 = \alpha_{10}$, $\beta_3 = \alpha_{01}$, $\beta_4 = \alpha_{11}$.

- The only difference is that this model includes random effects of the schools which we leave unaccounted for in the model. It is easy to see that if we leave off the error terms in our coefficient models:

$$\beta_{0j} = \alpha_{00} + \alpha_{01}z_j$$

$$\beta_{1j} = \alpha_{10} + \alpha_{11}z_j$$

We are right back to the fixed effect formulation. So the difference is that we are not assuming any longer that our fixed effect terms for the school district completely explain school-level context. By assuming this we will correct our standard errors. Let's compare our results:

Variable	OLS Estimate	Multilevel Estimate
Intercept	53.58 (0.67)	53.53 (1.20)
Non-White	-3.48 (1.29)	-2.72 (1.24)
SES	7.70 (0.83)	6.13 (0.96)
% non-white	-0.042 (0.026)	-0.070 (0.040)
%non-white*SES	-0.079 (0.024)	-0.075 (0.028)

Notice that our standard errors have gone up somewhat for the percent nonwhite variables. Also, to be accurate our t-tests for significance on these variables should be based on 21 degrees of freedom (23 schools - 2 parameters).

- We have two random effects terms in our model: a **random intercept** given by ϕ_j and a **random coefficient** given by ω_j . These random effects are combined with our fixed α effects. This is called a **mixed effects** model because it includes both random and fixed effects. It's estimation is tricky and we will not concern ourselves with it here. There are several specific computer programs designed explicitly to handle multilevel models (most notably HLM), but mixed effects model can also be estimated in most standard statistical packages as well. In stata this requires downloading the glamm package.

6 Causal Modeling

- We have not talked much about one of the major assumptions of the OLS regression technique. Let's say we have the following simple model.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

One major assumption of OLS regression is that the x_i values are uncorrelated with the error terms. Note that this is impossible to confirm because we only have estimates of the error terms and if correlation exists, then these estimates will be incorrect.

- What brings about this problem? Well, in general this problem is brought about by **omitted variable bias**. There is another variable which is correlated with both x and y so that after fitting the model above there is still a relationship with this other variable and the residuals.
- The omitted variable bias is the major difficulty of observational data. It is a major problem because we are generally interested in whether the model above represents a causal relationship between x and y . A frequent interpretation of the model above is that if we could manipulate x by raising it one unit, y would increase by β_1 units. This is a causal argument.

- Omitted variable bias is the most common illustration of what economists refer to as **endogeneity**. **endogenous variables** are variables determined by other variables in the system, while **exogenous variables** are variables which can be considered external shocks to the system (draw a picture).
- The other most important source of endogeneity is reverse causality.
- To truly be able to make a causal claim, we need a truly exogenous variable - that is, a variable which is not related to any of the other variables in the system, unobserved and observed. The problem with observational data is that there are an infinite number of unobserved variables which could render our observed relationship endogenous. This is the problem of **unobserved heterogeneity** in our sample.
- As an example, let's look at a simple question. Do private schools improve student's test performance? Let's say we had a sample of public and private school students' math test scores. We could look at the difference in the average score between groups. But it would be dangerous to assume that such a difference reflected the "treatment" of private schools, because it seems likely that more apt students are more likely to self-select into private schools.
- One standard solution is to control for all the observed measures that might lead to such self-selection which are available to us. The problem is that we are unlikely to effectively control for all of this selectivity, because some variables associated with the selection process are probably unobserved. Even if all the important variables were observed, we would only completely control for them if we correctly specified the functional form of their relationship to test scores.
- This problem has led to much wailing and gnashing of teeth among economists. Although aware of the problem, sociologists have been traditionally less concerned with the issue. I would argue this is due to a different conception of how arguments are presented and empirically tested in sociology. I would say the traditional model (for all empirical methods, not just statistical) follows this basic format:
 1. Make an argument about how and why things ARE AS THEY ARE.
 2. Show that the available empirical data are consistent with your argument.
 3. Demonstrate that the available empirical data are inconsistent with counter-arguments for how and why things ARE AS THEY ARE.
- The key issue here is the last one. The focus is on a debate between real concrete stories not on some generalized debate that some unspecified counter-story could plausibly exist.

- Although I actually tend to prefer this kind of conception of what we do, the problem of endogeneity is a real one and it behooves us to take a look at some of the ways in which people (partially) address it. The bottom line is that no method can perfectly recover causality from observational data, but in certain cases we can effectively reduce the range of plausible counter-stories. Let's focus on two common methods:

1. Fixed Effects Models

- Fixed effects models come primarily out of longitudinal data designs in which we have repeated observations on an individual over time. However, they can be applied more broadly than this.
- To continue with our example, lets say we had repeated observations on a set of high school student over their entire four-year high school period (to make it simple lets say none drop out or repeat a grade). We also have recorded test scores for each individual over this time period. Some students during this period have also migrated between public and private schools.
- Let's define some variables.
 - y_{it} is the test score for individual i in time period t .
 - s_{it} is the private school indicator variable for individual i in time period t .
 - \mathbf{x}_{it} is a set of observed time-varying variables for individual i in time period t .
 - \mathbf{x}_{it}^U is a set of unobserved time-varying variables for individual i in time period t .
 - \mathbf{z}_i is a set of time-constant observed variables for individual i .
 - \mathbf{z}_i^U is a set of time-constant unobserved variable for individual i .
- We could start with the same model we had above, but this time we will also control for any observed variables which may be confounding our relationship.

$$y_{it} = \beta_0 + \beta_1 s_{it} + \lambda x_{it} + \gamma z_i + \epsilon_{it}$$

This model estimates the average difference between private and public students' math scores with β_1 , controlling for all observed time-varying and time-constant covariates.

This model is problematic because it doesn't take account of the fact that we have repeated observations on the same students, which will likely lead to correlated error terms within students. Putting this issue aside, however, the model is still problematic in that it doesn't address the unobserved variables that may lead to self-selection into school type and better or worse test performance.

- We can take advantage of our longitudinal design to eliminate some of this unobserved heterogeneity (and to correct the issue with error terms). First, let's define a

set of dummy terms, D_i , which will be one if the observation comes from individual i and zero otherwise. Add these dummy terms to the models and we have:

$$y_{it} = \beta_0 + \beta_1 s_{it} + \lambda x_{it} + \alpha_i D_i + \epsilon_{it}$$

Or, more concisely

$$y_{it} = \beta_0 + \beta_1 s_{it} + \lambda x_{it} + \alpha_i + \epsilon_{it}$$

These dummy variables allow us to fit a term for every individual. Because we have multiple observations per individual, doing this will not saturate the model. Essentially, we are trying to explain variation within individuals. The α terms are our "fixed effects."

- Note that we are no longer explicitly including the observed z_i terms in the model. This is because our d_i terms explain **all** time-constant variation across individuals, so they supercede our z_i . In technical terms:

$$\alpha_i = \gamma_1 z_i + \gamma_2 z_i^U$$

So the fixed effects can account for both observed and unobserved time-constant variables. Thus we can be certain that our new estimate of β_1 is not the result of lurking variables that are constant across time.

- Another way of looking at this is that we are using the migrations of certain students as information about the effect of public and private schools. This is an improvement over the cross-sectional approach because we can rule out unobserved heterogeneity that is time-constant. However, we cannot rule out time-varying unobserved heterogeneity. In particular, we might think other events in students lives may be associated both with movements to and from private schools and test scores. A family disruption for example might reduce the resources to pay for private school and reduce test scores through stress and distraction.
- This approach can be applied to other data of a hierarchical structure. Longitudinal data are hierarchical data where time observations are nested within students. The fixed effects approach can be used on all data of this type to rule out any unobserved heterogeneity at the higher level. For example, if we had information on female siblings, we might use a fixed effects approach to rule out any family effects on the observed relationship between teen pregnancy and educational attainment.

2. Instrumental variables

- In some cases we may not be able to rule out that x is partially endogenous but we may have another variable z which we can be fairly certain has an effect on x but not on y . (draw a picture)
- The best situation is when we know that z has been completely randomized. Let's say for example that an experiment was done which randomly selected some families to receive private school vouchers. It is likely that these vouchers will induce some public school students to move to private school, thus there will be a relationship between z and x . However, we can be fairly certain (because we know assignment was random), that voucher assignment itself is not directly related to test scores.
- This situation is rare. The more common situation is a **natural experiment** in which for some historical reason we observe a shock to a system which can reasonably be treated as random. The most famous example here is the use of birthdates in the question of whether military service affected subsequent labor market experience. Because the draft was assigned on the basis of birthdates, it is highly correlated with military service, but unlikely to be correlated with labor market experience.
- If this relationship holds, then we can treat z as an **instrument** for inducement into the "treatment" of x .
- Mathematically, the basic reasoning is as follows. We have the basic relationship: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ But since x_i might be endogenous we cannot trust our estimate of β_1 . We can get an **instrumental variable** estimate of β_1 as:

$$b_1^{IV} = \frac{Cov(y, z)}{Cov(x, z)} = \frac{Cov(\beta_0 + \beta_1 x + \epsilon, z)}{Cov(x, z)} = \frac{Cov(x, z)\beta_1 + Cov(\epsilon, z)}{Cov(x, z)}$$

If we are correct in our assumptions about the instrument z , then $Cov(\epsilon, z) = 0$, and therefore:

$$b_1^{IV} = \frac{Cov(x, z)\beta_1 + Cov(\epsilon, z)}{Cov(x, z)} = \frac{Cov(x, z)\beta_1}{Cov(x, z)} = \beta$$

So that the IV estimator will be an unbiased estimator of β_1 . In essence we have used the exogenous shock of the instrument to "clean out" any endogenous relationship between x and y .

- The most common method for doing the actual estimation, two-stage least squares (2SLS), will help clarify this issue.
 - (a) As a first step, predict the value of x_i from z_i . If you plan on including other terms in your final model for y , say w_i , it is typical to include them at this stage as well:

$$\hat{x}_i = \alpha_0 + \alpha_1 z_i + \alpha_2 w_i$$

- (b) Now use the predicted value of x rather than its real value in an OLS regression predicting y

$$y_i = \beta_0 + \beta_1 \hat{x}_i + \beta_2 w_i + \epsilon_i$$

Because you are using the predicted value of x you are essentially leaving behind the residuals from the first equation. Since z is an exogenous shock on x , those residuals are the part of x which are potentially endogenous with y . You have stripped them away.

- In cases where you have strong reasons to believe in the exogeneity of z , the instrumental variables technique is quite clever. In practice, however, such instruments are often hard to find. This has led to the degradation of the method using what are called **weak instruments** which are only moderately correlated with x and where the assumption of no relationship between z and y may be dubious. In these cases, precise estimation is only possible when samples are large. Even then, IV generally doesn't solve the problem but rather re-focuses debate from the possible endogeneity of x to the validity of z as an instrument. Since this latter argument is likely to be more esoteric, the value of the IV approach becomes questionable.
 - My personal recommendation is that IV only be used in cases where the endogeneity of x is clear and unresolvable. In these cases, IV should be used as a complement to rather than a substitute for OLS.
- This field of what is called "causal modeling" or "causal inference" is very large and active, so I have only been able to touch on its surface. I have not even touched on a third major method, **propensity score matching**, nor have I discussed probably the major innovation of the last few years - the issue of **counterfactual causality** and **treatment heterogeneity**, which would be critical of all of these methods. If you are interested in it, I have provided some further readings on Courseworks.

7 Factor Analysis and Structural Equation Models

1. Factor Analysis

- Factor analysis is a part of a larger group of methods that differ substantially in their focus from the regression models we have been using so far.
- The key question in this set of methods is "how do things (observations/variables) go together?"

- Take as an example, the welfare state data provided by Esping-Andersen (pass out sheets). Esping-Andersen measures several characteristics of welfare states and argues that certain kinds of characteristics cluster together and define three types of welfare state: liberal, conservative, and social democratic.

	Corp.	Etatism	Means test	Priv. Pensions	Priv. Health	Universalism
Australia	1	0.7	3.3	30	36	33
Austria	7	3.8	2.8	3	36	72
Belgium	5	3	4.5	8	13	67
Canada	2	0.2	15.6	38	26	93
Denmark	2	1.1	1	17	15	87
Finland	4	2.5	1.9	3	21	88
France	10	3.1	11.2	8	28	70
Germany	6	2.2	4.9	11	20	72
Ireland	1	2.2	5.9	10	6	60
Italy	12	2.2	9.3	2	12	59
Japan	7	0.9	7	23	28	63
Netherlands	3	1.8	6.9	13	22	87
NewZealand	1	0.8	2.3	4	18	33
Norway	4	0.9	2.1	8	1	95
Sweden	2	1	1.1	6	7	90
Switzerland	2	1	8.8	20	35	96
UnitedKingdom	2	2	6.9	12	10	76
USA	2	1.5	18.2	21	57	54

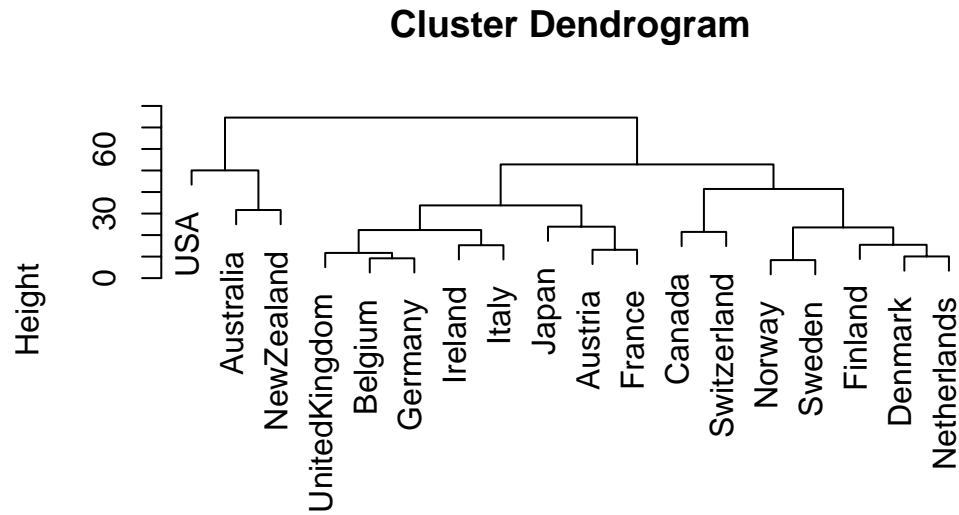
One can roughly look at the correlation matrix between these variables to provide evidence of such clustering.

	Corp.	Etatism	Means test	Priv. Pensions	Priv. Health	Universalism
Corp.	1	0.55	0.13	-0.39	-0.02	-0.02
Etatism	0.55	1	-0.09	-0.64	0.01	-0.03
Means test	0.13	-0.09	1	0.49	0.56	-0.03
Priv. Pensions	-0.39	-0.64	0.49	1	0.45	0
Priv. Health	-0.02	0.01	0.56	0.45	1	-0.28
Universalism	-0.02	-0.03	-0.03	0	-0.28	1

There is some evidence that the characteristics correlate as Esping-Andersen predicted

but it would be nice to have some more formal evidence.

- We could use the technique of **cluster analysis** to create a **dendrogram** which will create a branching tree based on observations that tend to share similar values vs. ones that are far apart. We won't go into technical details about how such trees are constructed, but here is an example of one:



`hclust (*, "complete")`

- But that's not necessarily what we want. The argument is really that the relationships between these variables are driven by underlying latent variables which measure the degree of social-democracy, liberalism, and corporatism in each country. This leads us to factor analysis.
- Let's start with an example with two variables and one common factor:
 - Let's turn our variables x_1 and x_2 into standardized variables, z_1 and z_2 . Now let's say that the actual observed values of these variables are determined by some common underlying factor (F) and a unique factor for each variable, y_1 and y_2 . (draw a picture)

$$z_{1i} = b_1 F_i + u_1 y_{1i}$$

$$z_{2i} = b_2 F_i + u_2 y_{2i}$$

- We will assume that F and both y 's are standardized and that they are all uncorrelated with one another.
- Let's ask what is the variance of z_1 . Since, z_1 is in standard form, this variance is simply given by $E(z_1^2)$, so:

$$E(z_1^2) = E(b_1^2 F^2 + u_1^2 y_1^2 + 2b_1^2 u_1^2 F y_1)$$

$$E(z_1^2) = b_1^2 E(F^2) + u_1^2 E(y_1^2) + 2b_1^2 u_1^2 E(F y_1)$$

$$\text{var}(z_1) = b_1^2 \text{var}(F) + u_1^2 \text{var}(y_1) + 2b_1^2 u_1^2 \text{covar}(F, y_1)$$

Assuming variances are one and covariance is zero:

$$\text{var}(z_1) = b_1^2 + u_1^2$$

$$1 = b_1^2 + u_1^2$$

Therefore, the variance in variable z_1 is determined by the contribution of the common factor and the unique factor.

- What about the covariance between z_1 and F :

$$\text{covar}(z_1, F) = E(z_1 F)$$

$$\text{covar}(z_1, F) = E((b_1 F + u_1 y_1) F)$$

$$\text{covar}(z_1, F) = (b_1 E(F^2) + u_1 E(y_1) F)$$

$$\text{covar}(z_1, F) = (b_1 \text{var}(F) + u_1 \text{covar}(y_1, F))$$

$$\text{covar}(z_1, F) = b_1$$

Since both F and z_1 are standardized, the covariance is the same as the correlation.

So r between z_1 and F is simply given by b_1 .

- How about the covariance between z_1 and z_2 :

$$\text{covar}(z_1, z_2) = E(z_1 z_2)$$

$$\text{covar}(z_1, z_2) = E((b_1 F + u_1 y_1)(b_2 F + u_2 y_2))$$

$$\text{covar}(z_1, z_2) = E(b_1 b_2 F^2 + b_1 u_1 F y_1 + u_1 u_2 y_1 y_2)$$

$$\text{covar}(z_1, z_2) = b_1 b_2 \text{var}(F) + b_1 u_1 \text{cov}(F, y_1) + u_1 u_2 \text{cov}(y_1, y_2)$$

$$\text{covar}(z_1, z_2) = b_1 b_2$$

$$r_{z_1, z_2} = b_1 b_2$$

So the correlation matrix between a set of variables is completely determined by their common factors.

- From the decomposition of the variance in z_1 , we can define the **communality** (h_j^2) of each variable as b_j^2 . This is the proportion of the variance explained by the common factor. The **uniqueness** of the variable is given by $1 - h_j^2$.
- This framework can be generalized to a set of j variables and m factors.

$$z_{ji} = b_{j1}F_{1i} + b_{j2}F_{2i} + \dots + b_{jm}F_{mi} + u_j Y_{ji}$$

(draw a picture)

With multiple common factors, communality is given by:

$$h_j^2 = \sum_{p=1}^m b_{jp}^2$$

This setup is called **common-factor analysis**. One can also assume that factors explain everything and there are no unique factors. In this case the number of factors will equal the number of variables.

$$z_{ji} = b_{j1}F_{1i} + b_{j2}F_{2i} + \dots + b_{jn}F_{ni}$$

The difference between the methods is that the common-factor approach only attempts to explain the covariation between observed variables, the principal-component approach attempts to explain all of the variation in the observed variables.

- The b 's are called **factor loadings**. They tell us the correlation coefficient between each factor and the observed variables.
- The values of each observation on the factor F_k are sometimes called the **factor scores**.
- The fundamental problem is that we never have the factors, we have the observed correlation matrix between variables and we **think** it may be produced by a set of common factors.
- If we are given a set of factors and their factor loadings for certain variables then we can reproduce the one and only one correlation matrix for these variables, **but** given a correlation matrix, we cannot deduce a single set of factor loadings.
- There are three related problems of indeterminacy in determining the factor structure from the correlation matrix
 - Competing causal structures: the relationship between x_1 and x_2 may be produced by their relationship to one another rather than by a common factor (draw a picture)
 - must have a theoretically-based model. This is the most serious problem.

- It is possible for the same covariance structure to be generated by differing numbers of factors. So how many factors are producing the process - we generally go for parsimony - decision is driven by goodness-of-fit.
- Even with the same number of factors, an infinite number of combinations of factor loadings could produce the same correlation matrix. This sounds like a serious problem, but it is actually not. Those infinite factor loadings actually all produce the same basic information, but they need to be **rotated** to be made more interpretable. (discuss US cities example)
- We won't go through how factor loadings are estimated, but suffice it to say they are initially estimated from the correlation matrix and then they need to be rotated to maximize the differences between the factors.
- The idea is to get factors which load highly on clusters of variables and have no loadings for other variables.
- Go through an example with esping-andersen data. The fitting of the factor analysis suggests that most of the variation in the observed variables can be explained by three latent factors - this is good because three is what we want.

Number of factors	1	2	3	4	5	6
Variance explained	0.39	0.66	0.83	0.91	0.97	1.00

Here are the rotated factor loadings for the three factors:

Variable	Factor 1	Factor 2	Factor 3	Uniqueness
Corporatism	0.85	0.15	0.07	0.25
Etatism	0.89	0.08	0.07	0.20
Means testing	-0.02	0.91	0.08	0.17
Private pensions	-0.68	0.63	0.05	0.14
Private health	0.02	0.80	-0.35	0.23
Average universalism	0.02	-0.05	0.97	0.06

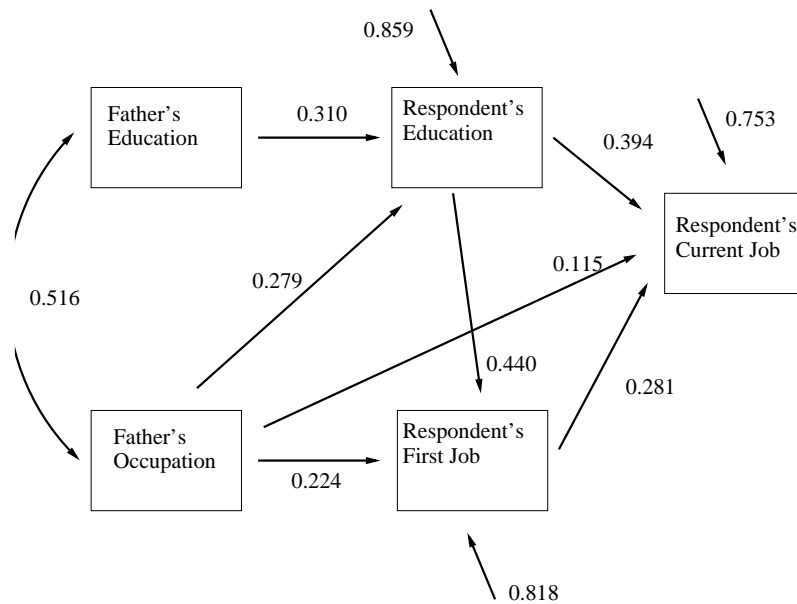
These factors correspond quite nicely to Esping-Andersen's typology(factor 1=conservative, factor 2=liberal, factor 3=social democratic). The only potential issues are the high negative factor loading on private pensions for factor 1 and the moderately negative factor on private health for factor 3, but these are not completely unexpected and in fact are somewhat understandable within Esping-Andersen's framework.

- I have already discussed the two major variants of factor analysis: common-factors and principal components - there is another way factor analysis varies:
 - exploratory factor analysis - seeing how things go together

- confirmatory factor analysis - setting an expectation about the underlying factors and their association with the observed variables and then testing that expectation.
- Factor analysis traditionally did not consider the issue of sampling variability and thus statistical inference.
- Factor analysis only works properly for continuous variables.

2. Structural Equation Models

- Structural equation models can be thought of as the model specifications for a path analysis diagram.
- Let's begin with the most famous path analysis of all time: Blau and Duncan, the occupational structure - a model was made showing how father's education and occupation affected respondent's current occupation through the intervening variables of education and first job.



- The numbers on the diagrams are partial correlation coefficients.
- There are two kinds of variables here : **exogenous** and **endogenous**.
- Endogenous variables are determined by other variables in the system. R's education, first job and occupation in 1962 are endogenous.
- Exogenous variables are those who enter in as determining variables but whose own determination is outside the system. Father's characteristics are exogenous here.

- This path diagram implies a set of equations that need to be estimated.

$$y_{1i} = \gamma_{10} + \gamma_{11}x_{1i} + \gamma_{12}x_{2i} + \zeta_{1i}$$

$$y_{2i} = \gamma_{20} + \gamma_{21}x_{1i} + \gamma_{22}x_{2i} + \beta_{21}y_{1i} + \zeta_{2i}$$

$$y_{3i} = \gamma_{30} + \gamma_{32}x_{2i} + \beta_{31}y_{1i} + \beta_{32}y_{2i} + \zeta_{3i}$$

- the γ 's are coefficients from exogenous to endogenous variables, and the β 's are coefficients between endogenous variables. (draw on board)
- Estimating this model is straightforward. This is an example of a **recursive model**, meaning:

- * There are no reciprocal paths or feedback loops in the path diagram.
- * The error terms are independent of one another (which is why they have no arrows).

When the system of equations is recursive, then each model can be estimated separately using OLS regression. In order to get our coefficients as partial correlation coefficients, we would simply standardize each of our variables first.

- This model can be used to determine the direct and indirect effects of different variables.
- Father's occupation has a direct effect on occupation of 0.115. But it also has a fairly significant indirect effect through its effect on getting a first job and through its association with father's education.

Through the first job:

$$(.224)(.281) = .063$$

Through father's education

$$(.516)(.310)(.394) + (.516)(.310)(.281) = .108$$

So the total indirect effect is:

$$.063 + .108 = .171$$

Thus, the total effect of father's occupation on occupation in 1962 is:

$$0.115 + 0.171 = 0.286$$

This is the correlation coefficient we would find between the two variables. The path analysis allows us to partition that total correlation into different paths.

- Non-recursive models create greater complications.
- Non-recursive models cannot be estimated as separate models because their error structures are correlated with one another. This can be done using a technique we will discuss later called **generalized least squares**. In economics, these kinds of models are called **simultaneous equation models**.
- SEM takes this issue one step further, by including latent unobserved variables into the path analysis. In effect, this method combines regression and factor analysis in one model.
 - The underlying idea is that our relationships in the path analysis are about some underlying concepts which are not measured perfectly by a set of observed variable.
 - Bollen uses the example of political democracy in industrialization. He creates a model where the level of political democracy in 1960 is related to the level of political democracy in 1965 and both of these are related to the level of industrialization in 1960.
 - The problem is that neither the level of political democracy or industrialization can be measured exactly. But we do have numerous observed variables which we think measure the underlying concept with some error.
 - * For political democracy, we have expert ratings of (1) freedom of the press, (2) freedom of political opposition, (3) fairness of elections, (4) and effectiveness of elected legislatures.
 - * For industrialization, we have GNP per capita, inanimate energy consumption per capita, and the percentage of the labor force in industry.
 - The model can be broken into two parts:
 - * first we have the **structural** model, which gives us the relationship between the underlying concepts. In our case:

$$\eta_1 = \gamma_{11}\xi_1 + \zeta_1$$

$$\eta_2 = \beta_{21}\eta_1 + \gamma_{21}\xi_1 + \zeta_2$$

Where η_1 and η_2 are the latent measures of political democracy in 1960 and 1965, respectively, and ξ_i is the latent measure of industrialization in 1960.

- * Then, we have what is called the **measurement** model, which measures the relationship between the latent variables and the observed variables.

We start by relating industrialization to the three observed measures

$$x_1 = \lambda_1\xi_1 + \delta_1$$

$$x_2 = \lambda_2 \xi_1 + \delta_2$$

$$x_3 = \lambda_3 \xi_1 + \delta_3$$

Now we relate the latent measure of political democracy to our observed measures of political democracy.

$$y_1 = \lambda_4 \eta_1 + \epsilon_1$$

$$y_2 = \lambda_5 \eta_1 + \epsilon_2$$

$$y_3 = \lambda_6 \eta_1 + \epsilon_3$$

$$y_4 = \lambda_7 \eta_1 + \epsilon_4$$

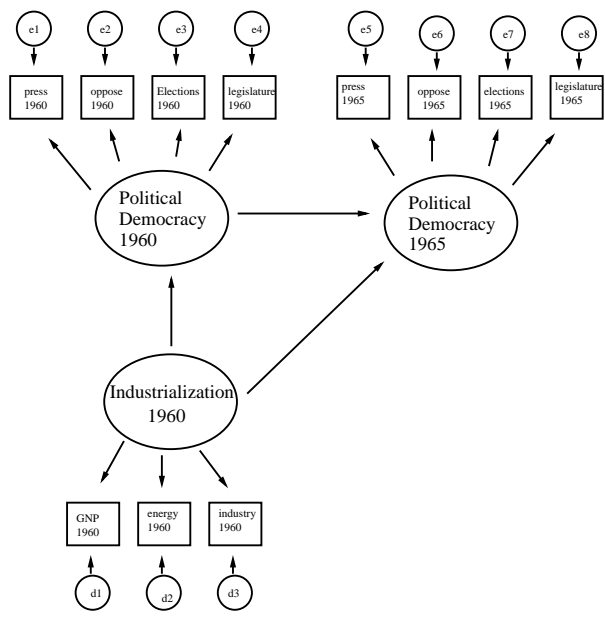
$$y_1 = \lambda_8 \eta_2 + \epsilon_5$$

$$y_2 = \lambda_9 \eta_2 + \epsilon_6$$

$$y_3 = \lambda_{10} \eta_2 + \epsilon_7$$

$$y_4 = \lambda_{11} \eta_2 + \epsilon_8$$

– This can all be shown in the following diagram.



- Now our only task is to estimate this set of simultaneous equations which involve both observed and latent variables :)
- We won't cover how this estimation works, but the intuitive ideas should be straightforward.