Introduction to Empirical Methods

Lecture 4: Multiple Regression Model Specification

Interaction Terms I

- Sometimes you want the effect of one independent variable to be conditional on the value of another.
- For example, the effect of education on turnout might itself be affected by income or gender or ethnicity.
- To address this you can use an interaction term.

Interaction Terms II

- The simplest way to add an interaction is to multiply two variables and add the product as a new independent variable.
- $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$
- Make sure to add the constituent terms (those with X_1 and X_2) in addition to the product term!
- To interpret the effect of X_1, you consider the marginal effect of it on Y, which is given by the derivative $\frac{dY}{dX_1} = \beta_1 + \beta_3 X_2$.
- Note how the effect of X_1 on Y changes with the value of X_2.

Nonlinear effects

- Sometimes you might think that your data display nonlinear dependence of Y on one or more X_i.
- There are lots of ways to get at this.
- The simplest and most common is to add higher-order polynomials to your regression.
- For example, $Y = \alpha + \beta_1 X + \beta_2 X^2$.
- To interpret the effect of X, you consider the marginal effect of it on Y, which is given by the derivative $\frac{dY}{dX_1} = \beta_1 + 2\beta_2 X$.
- Note how the effect of X on Y changes with the value of X.

Time Series I

- Sometimes your variables display trends in time that can lead to spurious causation.
- For example, the growth of golf and divorce since WW II.
- Also, if you have variation in your data across time, then an event at time t might be dependent not only on contemporary independent variables, but also ones in the past.
- Including all past variables leads to issues with multicollinearity. It is also not clear how many lagged (i.e. from earlier times) variables one should include.

Time Series II

- Two solutions: differenced variables and lagged dependent variables.
- Differences: $\Delta Y = \alpha + \beta \Delta X$. $\Delta Y \equiv Y_t - Y_{t-1}$.
- Differences eliminate dependence beyond immediate changes, so need a theoretical reason to do this.
- Lagged DV: $Y_t = \lambda Y_{t-1} + \alpha + \beta X_t$.
- Lagged DV allows the DV at time t to depend on all previous X_{t-k} through their effect on the lagged DV. As long as $\lambda < 1$. Larger values of λ mean past lags have more of a continued effect.

Another Example

- What is the effect of partisanship on feelings towards the environment?
- How do we measure feelings towards the environment?
  - Which is closer to the way you feel?
    - Regulation to protect the environment already places too much of a burden on business.
    - Tougher regulations are necessary to protect the environment.

Variables

- Dependent Variable
  - Scale from 1-5: larger values more in favor of environmental regulation.
- Key Independent Variable
  - Partisanship

0	1	2	3	4	5	6
St.Dem.	Wk.Dem.	Ln.Dem	Ind.	Ln.Rep.	Wk.Rep.	St.Rep.

- Control Variables
  - Income
  - Education

Result

	Estimate	Standard Error
Party	-.111	.013
Education	.083	.026
Income	-.013	.017
Constant	3.59	.059

- Which variables are statistically significant?
- How do we interpret the coefficient estimates?
  - Partial effect of the variable.
  - Effect holding all other variables constant.

Predicted Values

- With several independent variables how do you calculate predicted values?
  - Vary one variable and hold others at some constant value.
    - Typically, the mean.
- What is this telling us?
  - The effect of partisanship... all else being equal.
  - We can compare a Democrat with an average education and an average income to a Republican with an average education and an average income.

Predicted Values

	Estimate	Mean
Party	-.111	2.73
Education	.083	1.92
Income	-.013	1.86
Constant	3.59	

- $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$
- $Y = 3.59 + -.111 * Party + .083 * Education + -.013 * Income$
- $Y = 3.59 + -.111 * Party + .083 * 1.92 + -.013 * 1.86$
- Strong Democrat
  - $3.59 + (-.111 * 0) + (.083 * 1.92) + (-.013 * 1.86) = 3.73$
- Independent
  - $3.59 + (-.111 * 3) + (.083 * 1.92) + (-.013 * 1.86) = 3.39$
- Weak Republican
  - $3.59 + (-.111 * 5) + (.083 * 1.92) + (-.013 * 1.86) = 3.17$

Hold On One Second!!

- Do we really believe the relationship between partisanship and feelings toward the environment is linear?
  - Do we really believe that for each unit change in party identification the effect is the same?
- How can we allow for different effects?
  - Dummy coding.
    - For simplicity only Democrats, Republicans and Independents
- Unlike with turnout, there are three categories instead of only two.
  - Two dummy variables and a reference category.

Dummy Coding for Variables with More than Two Categories

- Why do we need a reference category?
  - In order to estimate coefficients, variables cannot be "collinear."
    - i.e., they cannot be perfectly related.
  - If we include all three variables, they are perfectly related to each other.
- *A statistically significant effect for one of the dummy variables only tells us that category is different from the reference category.*
  - We cannot make any inferences about comparisons between the other categories.
- How do we select a reference category?
  - It should be driven by the most interesting comparisons, but it is really arbitrary.
    - Compare Democrats to Independents and Republicans.

New Results

	Estimate	Standard Error
Republican	-.531	.056
Independent	-.079	.084
Education	.089	.026
Income	-.011	0.61
Constant	3.48	.056

- What do these results tell us?
  - Republicans are different from Democrats.
  - Independents are not different from Democrats.
  - Are Republicans are different from Independents?
    - We cannot tell that from this table.

Calculating Predicted Values

- $Y = \alpha + \beta_1 * Republican + \beta_2 * Independent + \beta_3 * Education + \beta_4 * Income$
- $Y = \alpha + -.531 * Republican + -.079 * Independent + .089 * Education + -.011 * Income$
- Republican
  - $Y = 3.48 + -.531 * 1 + -.079 * 0 + .089 * 1.92 + -.011 * 1.86 = 3.099$
- Independent
  - $Y = 3.48 + -.531 * 0 + -.079 * 1 + .089 * 1.92 + -.011 * 1.86 = 3.551$
- Democrat
  - $Y = 3.48 + -.531 * 0 + -.079 * 0 + .089 * 1.92 + -.011 * 1.86 = 3.630$

What Do You Need to Know About Dummy Variables?

- They can help you overcome the linearity assumption for ordinal variables.
- You must use them to measure categorical variables.
- How many dummy variables do you include?
  - The number of categories minus one.
- You can only compare cases to the reference category.

Audience Participation

- Who knows more about politics?
- Dependent Variable
  - Political Knowledge
    - On a scale from 0-26
- Independent Variables:
  - Age
  - Education
  - Partisanship
    - Dummy variables for Democrats and Republicans with independents as the reference category.
- Write out the model...
  - $\mathrm{PK} = \alpha + \beta_1 \mathrm{Age} + \beta_2 \mathrm{Ed} + \beta_3 \mathrm{Dem} + \beta_4 \mathrm{Repub}$

The Results

Variable	Coefficient	T-Ratio
Age	0.06	7.73
Education	2.65	25.01
Democrat	3.24	7.55
Republican	4.21	9.59
Constant	-1.29	-2.13

- Which variables are statistically significant?
- According to the model, who knows more about politics:
  - A 70 year old or a 30 year old?
  - A high school graduate or a college graduate?
  - A Democrat or an Independent?
  - An Independent or a Republican?
  - A Democrat or a Republican?

Audience Participation

- Calculate the following predicted values
  - Holding education constant at 3.

Variable	Coefficient	T-Ratio
Age	0.06	7.73
Education	2.65	25.01
Democrat	3.24	7.55
Republican	4.21	9.59
Constant	-1.29	-2.13

- 30 year old Republican
- 60 year old Democrat
- 40 year old Independent

Answers

Variable	Coefficient	T-Ratio
Age	0.06	7.73
Education	2.65	25.01
Democrat	3.24	7.55
Republican	4.21	9.59
Constant	-1.29	-2.13

- 30 Year Old Republican
  - $-1.29 + 0.06 * 30 + 2.65 * 3 + 3.24 * 0 + 4.21 * 1 = 12.67$
- 60 Year Old Democrat
  - $-1.29 + 0.06 * 60 + 2.65 * 3 + 3.24 * 1 + 4.21 * 0 = 13.5$
- 40 Year Old Independent
  - $-1.29 + 0.06 * 40 + 2.65 * 3 + 3.24 * 0 + 4.21 * 0 = 9.06$

What kind of hypothesis test should we run?

- What if we have a binary dependent variable (e.g., vote or not) and a continuous independent variable (e.g., age)?
- What are the levels of measurement for our independent and dependent variables?

		I.V. Type	
		Categorical	**Continuous**
D.V. Type	**Categorical**	Tabular Analysis	**Probit/Logit**
	Continuous	Difference of Means	Correlation Coefficient

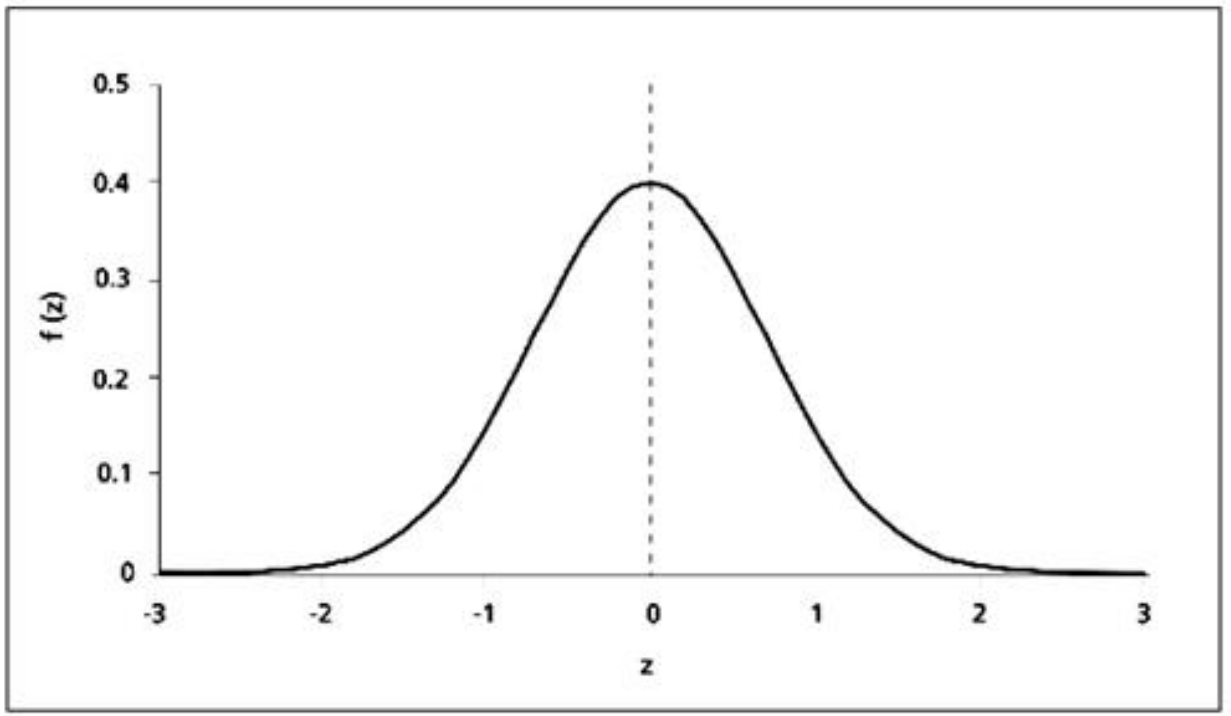
- There are two common types of probability models: logit and probit.
  - We are going to consider logit.

Assumptions of the O.L.S. Model

- Linearity
  - A straight line adequately represents the relationship in the population
  - Fitting a linear model to a nonlinear relationship results in biased estimates

Assumptions of the OLS Model

- Normality
  - The errors are normally distributed.
  - This also implies that the dependent variable is normally distributed.



Remember What that Last Assumption Means

- These assumptions restrict the types of dependent variables we can use.
- Our dependent variable must be:
  - Interval-ratio level
  - Continuous
  - Unbounded
- Very few variables meet these criteria.
  - Some come close enough.

O.L.S. with Dichotomous Outcomes

- One case where our model will likely fail to meet these assumptions is when our dependent variable is dichotomous.
  - Did the respondent vote?
  - Did the dyad go to war?
  - Did the country sign a treaty?
- You should use caution in using O.L.S. with a dichotomous dependent variable, as doing so violates assumptions.
  - But if you do, you will using a *linear probability model*.
  - We did this before with our vote choice examples.

Interpreting the L.P.M.

- So, we have a dichotomous dependent variable, but the model is just the same as before.
- $Y = \alpha + \beta X$
  - Dependent variable is coded one or zero.
  - For the case with a single independent variable.
- A one unit change in X results in a β unit change in the probability that $Y = 1$.

A Pedagogical Example

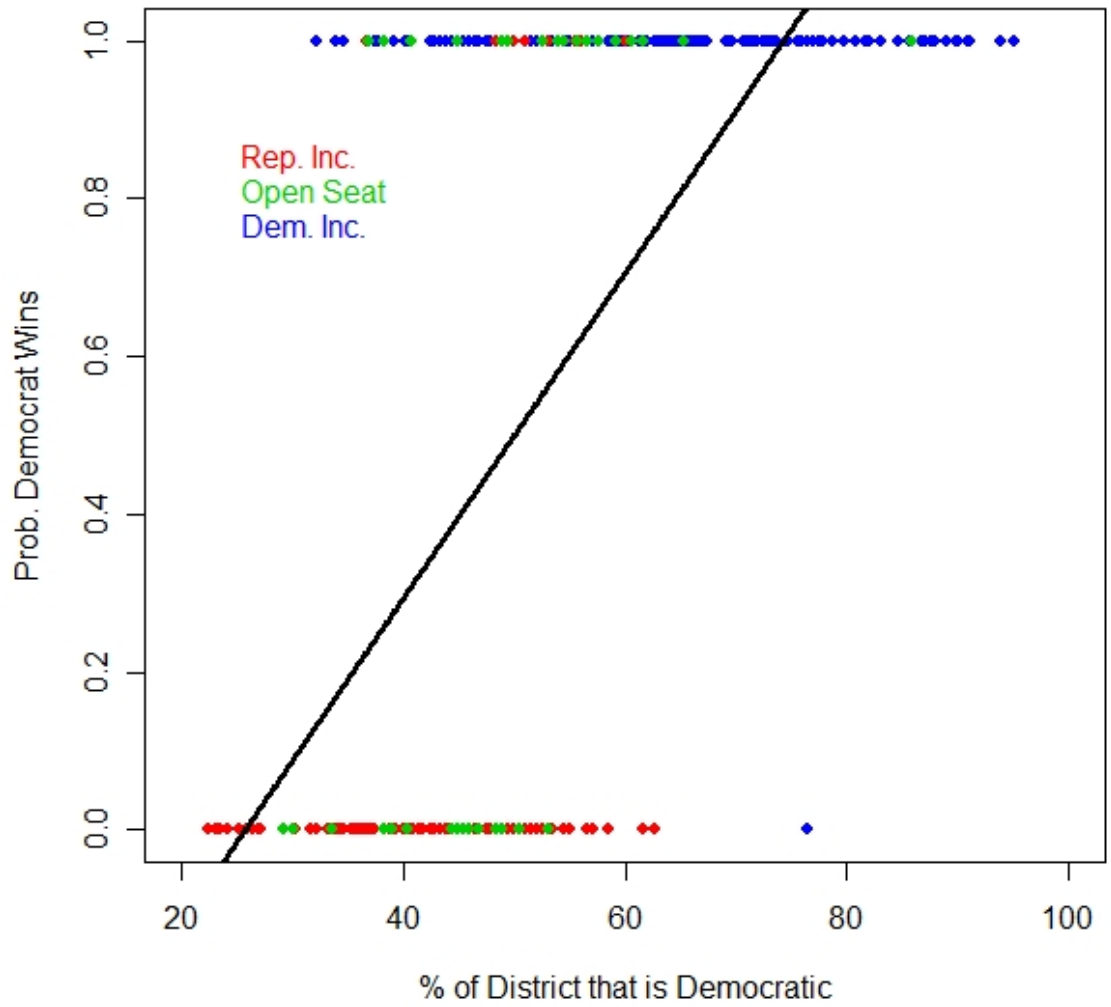
- Hypothesis
  - The more Democrats who live in a Congressional District, the more likely it is that the Democratic candidate will win.
- Data
  - 2008 Congressional Election Data
- Dependent Variable
  - 1=Democrat Won; 0=Republican Won
- Independent Variable
  - % of District that is Democratic

O.L.S. Results

	Estimate	Standard Error	T-Value
% Democratic	0.021	0.001	16.726
Constant	-0.532	0.070	-7.645

- Each percentage point increase in the % of the district that is Democratic results in a _____% increase in the probability that the Democrat wins.

OLS Prediction



Why We Shouldn't Have Done That

- There are a couple of obvious problems with the previous example:
  - Impossible Probabilities.
  - The straight line does not really fit the data.

When OLS Assumptions Fail

- We can transform the typical OLS model to estimate models that take into account the distribution of the dependent variable.
  - We can have many different types of models with different distributions of the dependent variable and we'll still be able to understand what the relationship is between the independent and dependent variables.
- For example, logistic regression.
  - A.K.A., logit.
- Logit and Probit are proper models with dichotomous dependent variables.

Logit

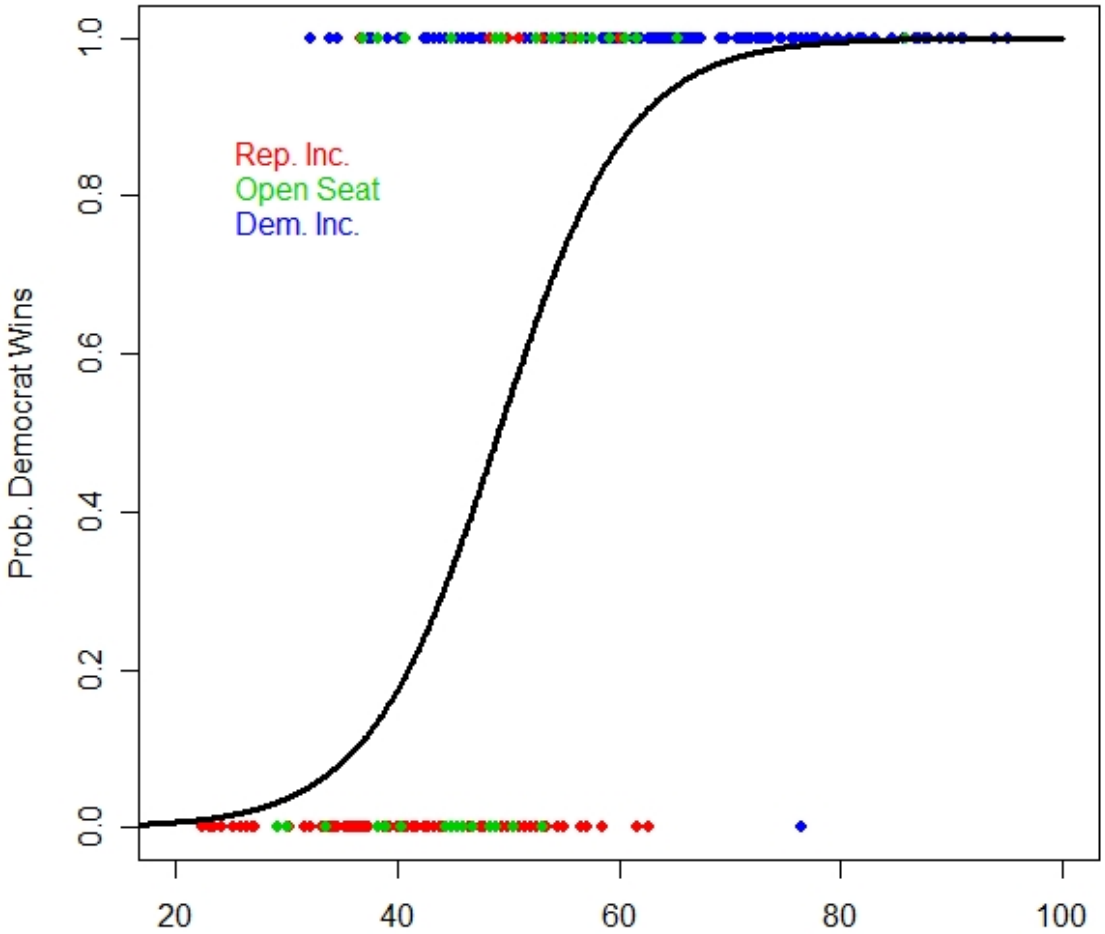
- $Y = \frac{1}{1+e^{-(\alpha+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_k X_k)}}$
- We can test for statistical significance just as we did before.
- This time we do a z-test instead of a t-test.
  - z and t distributions are the same with large samples sizes.
  - So these tests are the same with large sample sizes.
  - Should not use logit with small sample sizes.
- Critical Values at .05 Level:
  - One-Tailed Test: 1.645
  - Two-Tailed Test: 1.960

Logit Results

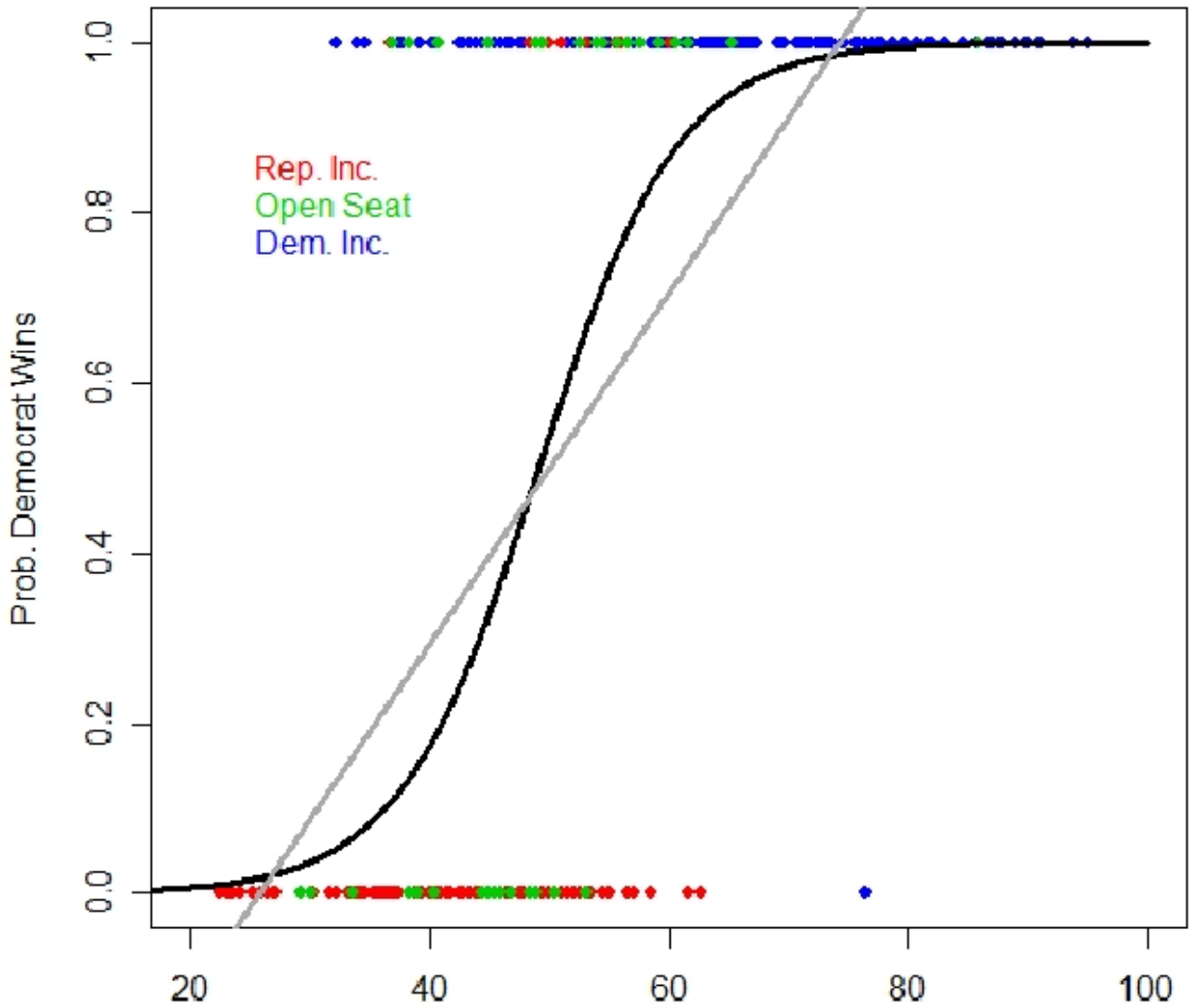
	Estimate	Standard Error	Z-Value
% Democratic	0.171	0.017	10.118
Constant	-8.410	0.849	-9.902

- % Democratic is still statistically significant.
- However, the coefficient does not mean what it used to.
  - A one unit increase in X does **not** produce a β change in the expected value of Y.

Predicted Line from the Logit Model



Comparing the Models



Another Example

- Dependent Variable
  - Did NES respondent vote?
    - Dichotomous
- Independent Variables
  - Education
    - Five category variable.
  - Gender
    - Dummy variable (0=Female, 1=Male)

Comparing OLS and Logit Models

OLS Model

	Estimate	Standard Error
Education	0.113	0.009
Gender	0.026	0.021
Constant	0.528	0.022

Logit Model

	Estimate	Standard Error
Education	0.709	0.062
Gender	0.176	0.128
Constant	-0.139	0.123

Comparing Predicted Values

Degree	OLS (Men)	OLS (Women)	Logit (Men)	Logit (Women)
No High School	55.40%	52.80%	50.90%	46.50%
High School	66.70%	64.10%	67.80%	63.90%
Some College	78.00%	75.40%	81.10%	78.20%
College Graduate	89.30%	86.80%	89.70%	87.90%
Advanced Degree	100.70%	98.10%	94.60%	93.70%

- What do you notice about the gaps between categories?
- In the Logit model, the effect of education depends on whether the person is a man or a woman.

Different Models, Similar Ideas

Dependent Variable	Model
Continuous	Ordinary Least Squares
Dichotomous	Logit
Dichotomous	Probit
Ordered Categories	Ordered Logit
Nominal Categories	Mulitnomial Logit
Counts of Events	Poisson
Counts of Events	Negative-Binomial
Time Until an Event	Cox Proportional Hazards

- We still use the coefficient estimate to perform hypothesis tests.
  - You can understand hypothesis tests in regression models you've never seen before.
    - NB: the substantive meaning of coefficients is different in each type of regression.

A Model You've Never Seen

- I want to understand how people viewed President Clinton at the end of his administration.
- Dependent Variable
  - Answers to the question: "Since President Clinton took office, has the president made the "moral climate" in this country better or worse or has it stayed the same?"
    - Three category ordered dependent variable.
  - Ordered Logit Model
- Independent Variables
  - Party
    - -1:Republican; 0:Independent; 1:Democrat
  - Highest degree attained.
  - Age in years
  - Regular church goer
    - Dummy Variable
  - Gender
    - Dummy Variable

The Results of the Model, I

Variable	Coef.	Z-Value
Party	0.780	13.31
Education	-0.245	-5.49
Age	-0.004	-1.20
Church	-0.389	-3.62
Female	-0.206	-1.91

τ_1	-1.462
τ_2	1.862

- What variables are statistically significant at the .05 and .10 levels?

The Results of the Model, II

Variable	Coef.	Z-Value
Party	0.780	**13.31**
Education	-0.245	**-5.49**
Age	-0.004	-1.20
Church	-0.389	**-3.62**
Female	-0.206	*-1.91*
τ_1	-1.462	
τ_2	1.862	

- Z-test Critical Values:
  - One-Tailed Test: *1.645*
  - Two-Tailed Test: **1.960**