POLS201 Spring 2019

Dichotomous Dependent Variables: Predicting Probabilities

April 8

Agenda: We have ten sessions left

- Three new topics:
  - binary dependent variables
  - robustness checks
  - interactions of independent variables
- Two labs
  - both give you practice running regressions
- Four days of presentations

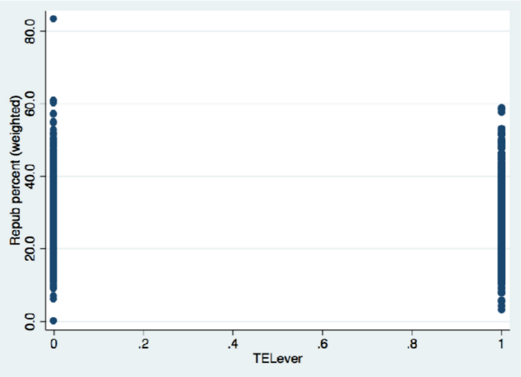
You need to know one function to create an OLS linear model: lm()

- Remember the format: your_model <- lm(data = your_dataset, dv ~ iv1 + iv2 + iv3_etc)
- Then you can run summary(your_model) to get your coefficients.

We know that a linear estimate works for binary independent variables

POLS201
Spring 2019

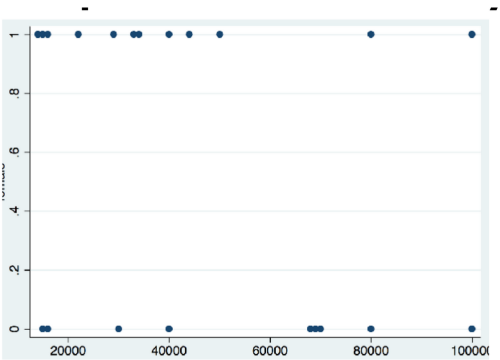
- We fit a line between the two dummy variables and we have a prediction about the incremental effect of increasing the IV from 0 to 1.



Regressions with dummy variables as IVs using OLS works fine.

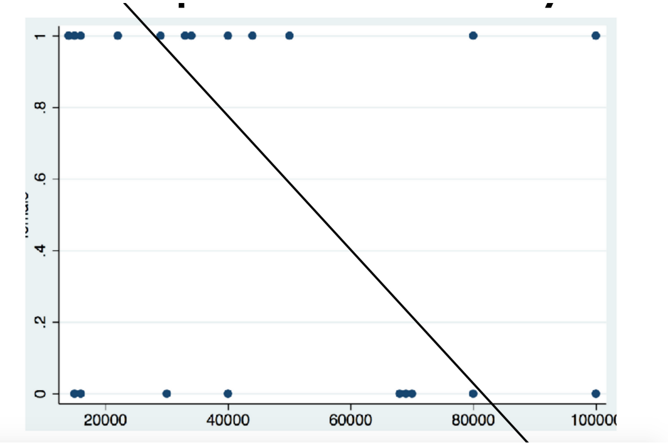
Scatter Plot with Binary DV

-But a linear prediction fails when the dependent variable is binary



See how a linear estimate fails to match the data

- The most obvious kind of dichotomous variable: vote choice



We deal with this problem by running a different kind of model

POLS201
Spring 2019

- We don't try to fit a straight line. Instead, we transform the DV into a continuous variable
- The new line describes the probability that Y=1, given the value of our IV's.
- I created some data shows the relationship between "feelings about immigration" and Trump vote.

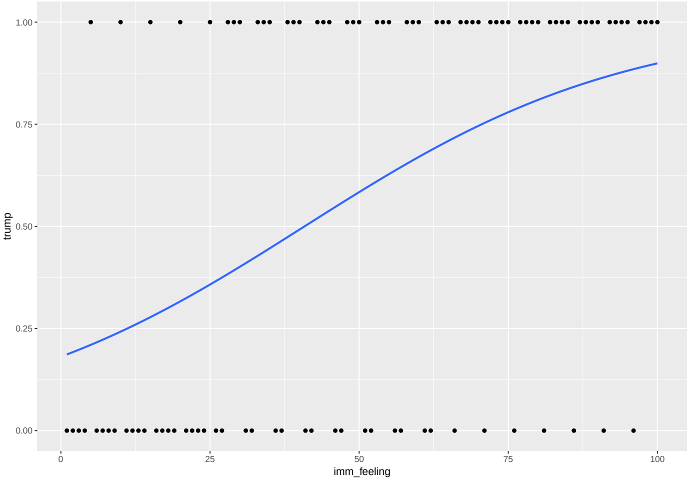
```
library(readxl)
suppressMessages(library(ggplot2))

prob_ex <- read_xlsx("probit_example.xlsx")

fit <- lm(trump ~ imm_feeling, data=prob_ex )
gfit <- glm(trump ~ imm_feeling, data=prob_ex, family
summary(gfit)
```

Suppose we have an immigration feeling thermometer (1-100) and Trump vote

```
ggplot(prob_ex, aes(y = trump, x = imm_feeling)) + ge
```



Dichotomous DV's

- Because "Linear Probability Models" often make predictions outside the interval 0 and 1 (and a few other reasons. . .), most political scientists will choose another regression option. Two are widely used (and we don't need to dwell on the choice: either is fine for us)
- **Logit** is one popular example
- **Probit** is the most popular
- Instead of a linear predsiction of y is transformed into a prediction about the probability of 1 vs 0.
- We transform y into a continuinous variable using a link function.
- The result produces a prediction between zero and one for every x value.

Using a Logit or Probit Regression: Problems

- Unlike a linear prediction, the coefficient doesn't give you a simple prediction of the marginal increase.
- Notice how the line curves around. The marginal increase depends on your x value.
- The best way to adapt: compute the result for specific x values, or draw a picture.

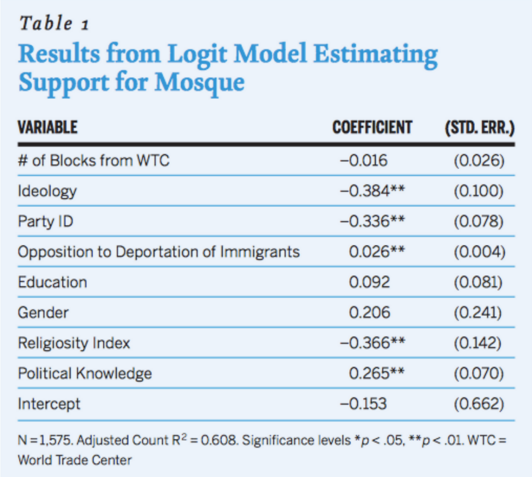


Table 1

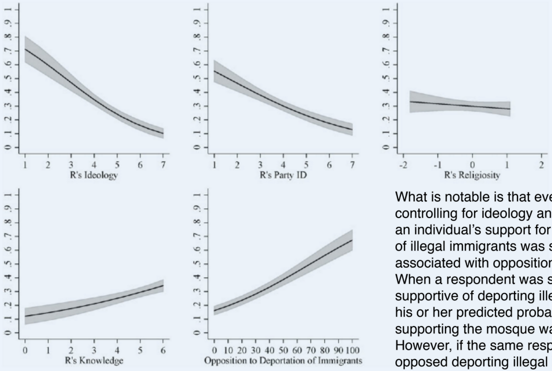
Results from Logit Model Estimating Support for Mosque

VARIABLE	COEFFICIENT	(STD. ERR.)
# of Blocks from WTC	−0.016	(0.026)
Ideology	−0.384**	(0.100)
Party ID	−0.336**	(0.078)
Opposition to Deportation of Immigrants	0.026**	(0.004)
Education	0.092	(0.081)
Gender	0.206	(0.241)
Religiosity Index	−0.366**	(0.142)
Political Knowledge	0.265**	(0.070)
Intercept	−0.153	(0.662)

N = 1,575. Adjusted Count R^2 = 0.608. Significance levels *$p < .05$, **$p < .01$. WTC = World Trade Center

Figure 5

Relationship between Independent Variables and Predicted Probability of Supporting Mosque



What is notable is that even after controlling for ideology and partisanship, an individual's support for the deportation of illegal immigrants was strongly associated with opposition to the mosque. When a respondent was strongly supportive of deporting illegal immigrants, his or her predicted probability of supporting the mosque was less than .2. However, if the same respondent strongly opposed deporting illegal immigrants, the predicted probability of supporting the mosque increased to .75.

Consider the Hillygus Articles Argument and its Model

- The structure of the paper is now familiar:
- Hypothesis
- Structure of Literature Review
- Unit of Analysis
- DV, IV, Confounds

Three related but distinct explanations of link between education and political participation

- **Civic Education Hypothesis:** suggests that additional years of education can continue to equip citizens with political information that further ceases the costs of political engagement
- **Social Networks:** More educated individuals may have friends who engage them in political conversation; and they are more likely to be mobilized by campaigns or candidates. Limited number of seats at the table explains null relationship over time.
- **Intellectual Meritocracy:** Put simply, this hypothesis suggests that intelligence begets educational attainment, not the other way around. Correlation not causal effect.

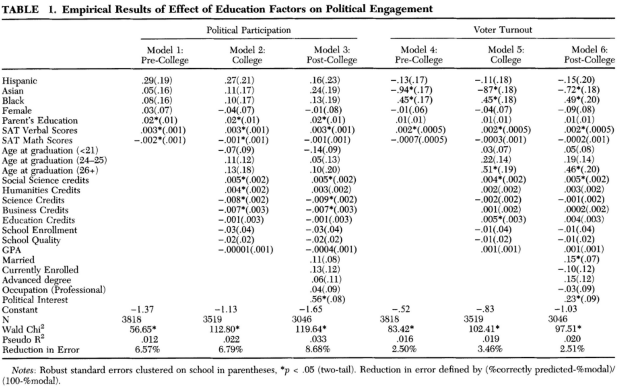


TABLE 1. Empirical Results of Effect of Education Factors on Political Engagement

	Political Participation			Voter Turnout		
	Model 1: Pre-College	Model 2: College	Model 3: Post-College	Model 4: Pre-College	Model 5: College	Model 6: Post-College
Hispanic	.29(.19)	.27(.21)	.16(.23)	−.13(.17)	−.11(.18)	−.15(.20)
Asian	.05(.16)	.11(.17)	.24(.19)	−.94*(.17)	−87*(.18)	−.72*(.18)
Black	.08(.16)	.10(.17)	.13(.19)	.45*(.17)	.45*(.18)	.49*(.20)
Female	.03(.07)	−.04(.07)	−.01(.08)	−.01(.06)	−.04(.07)	−.09(.08)
Parent's Education	.02*(.01)	.02*(.01)	.02*(.01)	.01(.01)	.01(.01)	.01(.01)
SAT Verbal Scores	.003*(.001)	.003*(.001)	.003*(.001)	.002*(.0005)	.002*(.0005)	.002*(.0005)
SAT Math Scores	−.002*(.001)	−.001*(.001)	−.001(.001)	−.0007(.0005)	−.0003(.001)	−.0002(.001)
Age at graduation (<21)		−.07(.09)	−.14(.09)		.03(.07)	.05(.08)
Age at graduation (24–25)		.11(.12)	.05(.13)		.22(.14)	.19(.14)
Age at graduation (26+)		.13(.18)	.10(.20)		.51*(.19)	.46*(.20)
Social Science credits		.005*(.002)	.005*(.002)		.004*(.002)	.005*(.002)
Humanities Credits		.004*(.002)	.003(.002)		.002(.002)	.003(.002)
Science Credits		−.008*(.002)	−.009*(.002)		−.002(.002)	−.001(.002)
Business Credits		−.007*(.003)	−.007*(.003)		.001(.002)	.0002(.002)
Education Credits		−.001(.003)	−.001(.003)		.005*(.003)	.004(.003)
School Enrollment		−.03(.04)	−.03(.04)		−.01(.04)	−.01(.04)
School Quality		−.02(.02)	−.02(.02)		−.01(.02)	−.01(.02)
GPA		−.00001(.001)	−.0004(.001)		.001(.001)	.001(.001)
Married			.11(.08)			.15*(.07)
Currently Enrolled			.13(.12)			−.10(.12)
Advanced degree			.06(.11)			.15(.12)
Occupation (Professional)			.04(.09)			−.03(.09)
Political Interest			.56*(.08)			.23*(.09)
Constant	−1.37	−1.13	−1.65	−.52	−.83	−1.03
N	3818	3519	3046	3818	3519	3046
Wald Chi^2	56.65*	112.80*	119.64*	83.42*	102.41*	97.51*
Pseudo R^2	.012	.022	.033	.016	.019	.020
Reduction in Error	6.57%	6.79%	8.68%	2.50%	3.46%	2.51%

Notes: Robust standard errors clustered on school in parentheses, *$p < .05$ (two-tail). Reduction in error defined by (%correctly predicted-%modal)/(100-%modal).

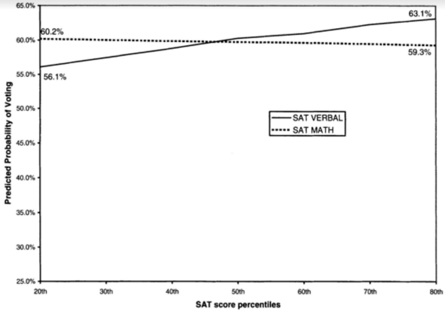
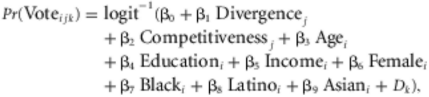


FIG. 2. Predicted probability of voting by SAT percentiles.

A positive effect of verbal proficiency is also found for vote turnout, illustrated in Fig. 2, with the predicted probability increasing from 56% to 63% from the 20th to 80th percentile.

Notation

$$P(Y_i = 1) = \Lambda(\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i)$$



$$\begin{aligned} Pr(\text{Vote}_{ijk}) = \text{logit}^{-1}(&\beta_0 + \beta_1 \text{ Divergence}_j \\ &+ \beta_2 \text{ Competitiveness}_j + \beta_3 \text{ Age}_i \\ &+ \beta_4 \text{ Education}_i + \beta_5 \text{ Income}_i + \beta_6 \text{ Female}_i \\ &+ \beta_7 \text{ Black}_i + \beta_8 \text{ Latino}_i + \beta_9 \text{ Asian}_i + D_k), \end{aligned}$$