Regression Analysis

Results in Brusco et al. 2004

Table 3 Model Estimations of Vote Buying

	(1)	(2)	(3)	(4)	(5)
Dependent Variable	Patron	Puntero	Job	Gift	Influence
Model Estimated	*Logit*	*Logit*	*Logit*	*Logit*	*Ordered Logit*
Income	**-0.126**	0.005	-0.054	**-0.174**	**-0.207**
	(0.058	(0.055)	(0.037)	(0.074)	(0.070)
Education	-0.005	-0.050	**-0.197**	**-0.162**	**-0.185**
	(0.058)	(0.039)	(0.035)	(0.071)	(0.066)
Housing	-0.215	**-0.219**	**-0.133**	**-0.254**	**-0.294**
	(0.114)	(0.084)	(0.073)	(0.124)	(0.115)
Gender	-0.178	0.093	**0.208**	-0.092	0.153
	(0.166)	(0.118)	(0.103)	(0.181)	(0.171)
Age	-0.005	-0.001	**-0.022**	**-0.012**	**-0.016**
	(0.006)	(0.006)	(0.003)	(0.006)	(0.006)
Peronist sympathizer	**0.594**	**0.273**	**0.735**	**0.806**	**0.807**
	(0.192)	(0.187)	(0.119)	(0.202)	(0.189)
Radical sympathizer	**0.357**	0.041	0.146	-0.217	0.213
	(0.243)	(0.208)	(0.158)	(0.346)	(0.278)
Log population	**-0.361**	0.034	-0.035	**-0.108**	**-0.107**
	(0.044)	(0.042)	(0.029)	(0.047)	(0.043)
Constant	**3.254**	-0.437	1.879	0.911	
	(0.643)	(0.616)	(0.397)	(0.690)	
N observations	1114	1920	1920	1920	1920

NOTE: Cell entries are coefficients, standard errors in parentheses. Boldface indicates significance at the p=0.05 level or smaller. Models in columns 2 through 5 use five imputed datasets generated by *Amelia* program. (Responses to *Patron* depended on prior responses and reduced the relevant sample of respondents to 1,114; here we analyzed the original matrix and used listwise deletion.)

Explanation of variables. Refer to table 1 for question wording associated with these variables. *Patron, Puntero, Job, Gift*: coded yes=1. *Influence*: coded 1=did not receive goods; 2=received goods, no influence; 3=received goods, acknowledged influence. Based on responses to open-ended question. Other variables coded as explained in the note to table 2.

Relationship

Many questions we are interested in comparative politics are interested in relationships between between two variables, X and Y, e.g.:

X	**Y**
Income	Vote Buying
Party ID	Vote choice
Regime type	Number of parties
Ethnic diversity	Public goods provision

Regression

- A statistical tool to determine the relationship between two or more variables. Regression is primarily used for prediction and causal inference.
- Most widely used tool in the social sciences
- Knowing how to do/interpret a regression analysis is widely useful.

Notation

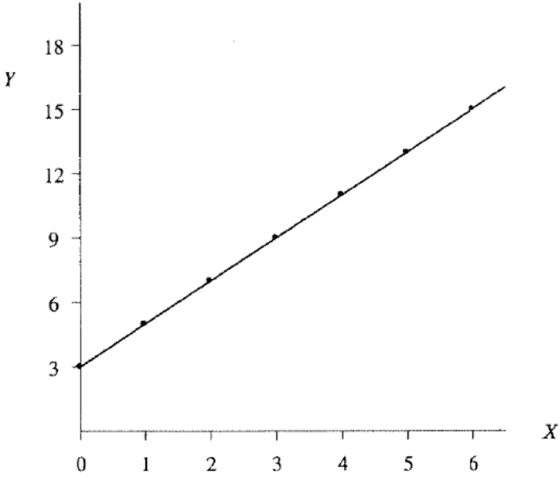
- **Y** is the outcome or "dependent" variable
- **X** is the explanatory or independent variable
- A simple linear regression equation is the function :

$$Y = \alpha + \beta X + \varepsilon \tag{1}$$

- where β is the slope; α is the intercept; ε is an error term that captures the amount of variation not predicted by the slope and intercept terms

Linear Function

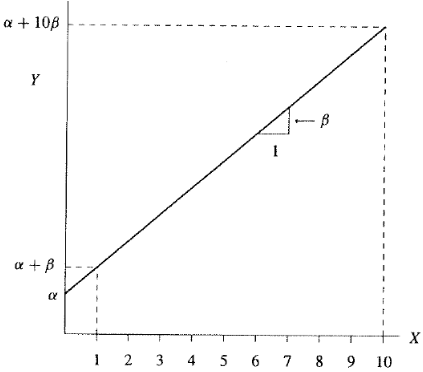
Graph of the Straight Line $Y = 3 + 2X$



Interpretation

Graph of the Straight Line $Y = \alpha + \beta X$

β is change in Y, given a one unit change in X



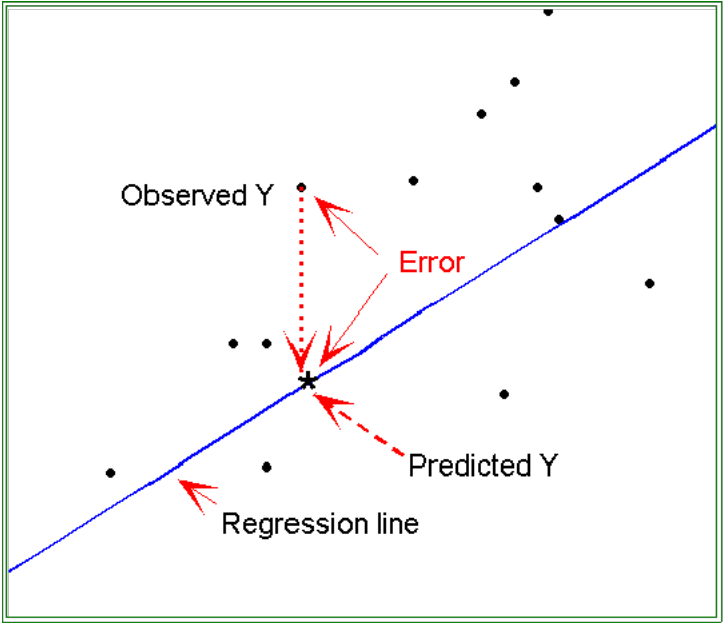
Interpretation

The sign of β determines the type of relationship between $\mathbf{Y}$ and $\mathbf{X}$

Simple Regression

- How the parameter estimates are generated:
- True value: $S = \alpha + \beta P + \varepsilon$
- Estimated Value: $\hat{S} = \hat{\alpha} + \hat{\beta} P$
- Estimated error $= S - \hat{S}$
- Regression analysis is to find the line that minimize the sum of squares of estimated errors.

Regression Line



Statistical Inference

- How do we know the coefficients from the regression analysis are reliable? How much confidence can we have for the coefficients?
- Fortunately, we have tools that discern the credibility of coefficients from the regression analysis

Statistical Inference

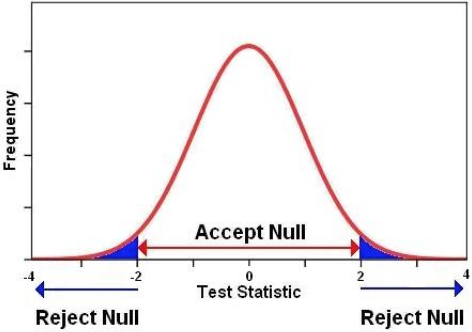
- $\hat{\beta}$ is a point estimate of β in the regression of Y on X. This is our best guess, but may be too high or too low
- Standard error s of $\hat{\beta}$ gives us an indication of how much the point estimate is likely to vary from the corresponding true parameter

Statistical Inference

- Population mean = μ, parameter mean = β, standard error = s
- t-statistics = $(\mu - \beta)/s$
- t-statistics is used in hypothesis testing when you want to figure out if you should accept or reject the null hypothesis
- Null hypothesis = there is no relationship between two variables

T-DISTRIBUTION

- The central region on this graph is the acceptance area and the tail is the rejection region
- The tail area is referred to as "p-value": what is the probability of observing this t-stat if the null hypothesis is true?



Statistical Inference

- A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so you can reject the null hypothesis ("Difference is statistically significant")
- A large p-value (typically ≥ 0.05) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis

STANDARD ERROR, T-STATS, P-VALUE

- Small standard error $\Rightarrow$ Large t-stats $\Rightarrow$ Small p-value: Coefficient of interest is statistically significant (***)

- Large standard error $\Rightarrow$ Small t-stats $\Rightarrow$ Large p-value: Coefficient of interest is not statistically significant