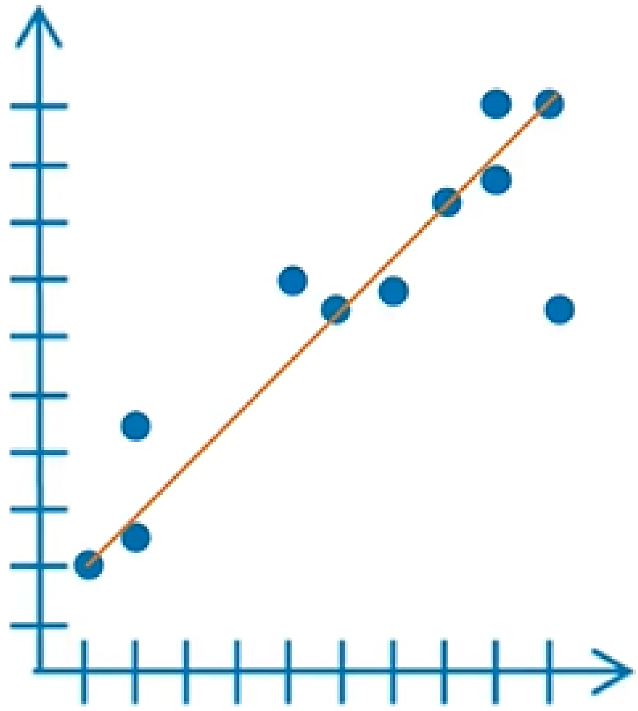
Linear regression (review)

Demand/supply, Y, for a service is dependent on:

$$Y = \beta_0 + \beta_1 \text{ (Price)} + \beta_2 X_2 + \beta_3 X_3 + \ldots + \varepsilon$$



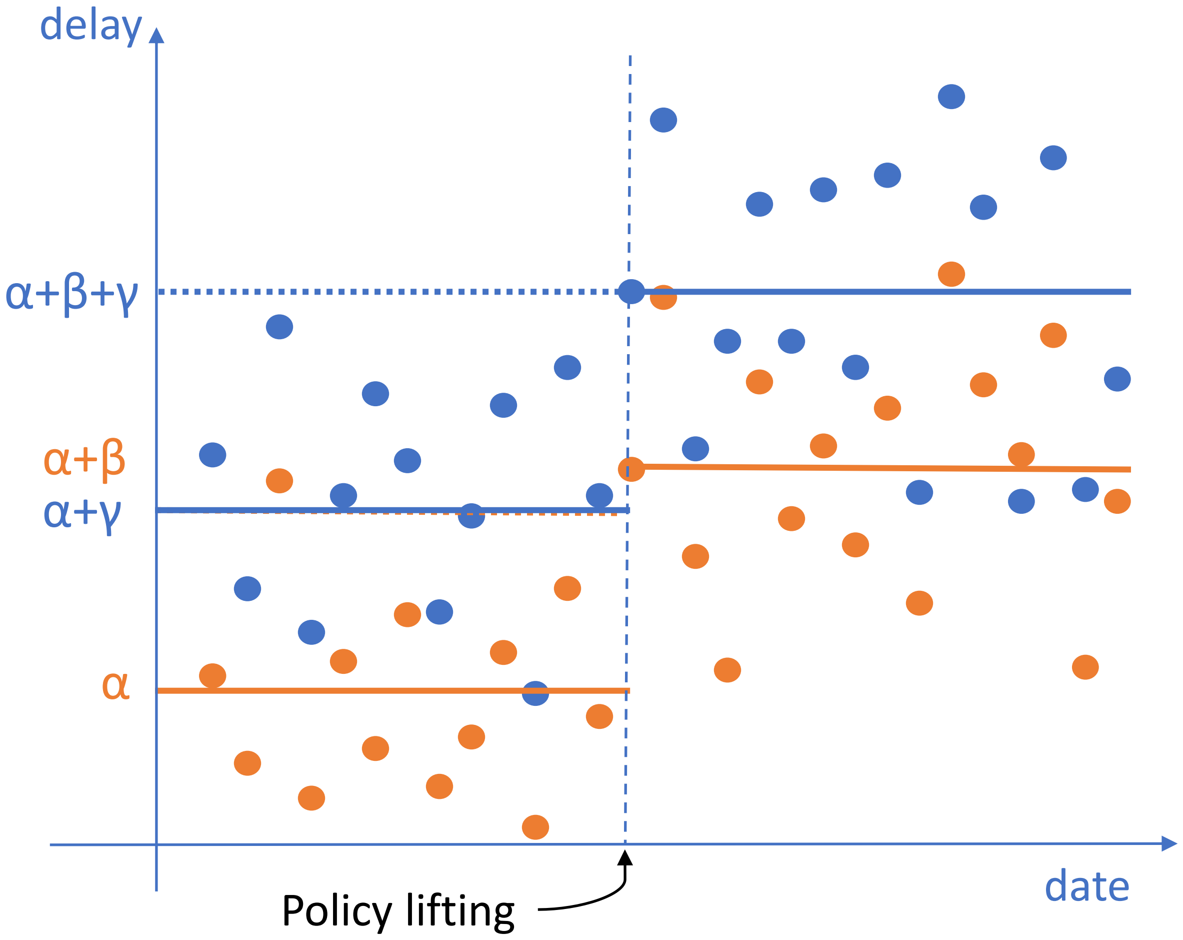
- Explanatory variables: Price, X_2 , X_3, …
  - That we observe
- Coefficients: β_0, β_1, β_2, …
  - Unknown parameters of interest
- Random error term ε
  - that are unobservable/“unpredictable” to us
- If we have data on the dependent and explanatory variables, we can “estimate” the coefficients that would best “fit” the data
  - i.e., choose coefficients to minimize distance between actual data points and prediction

Hanna, Kreindler, and Olkein (2017)

Is an "event study"

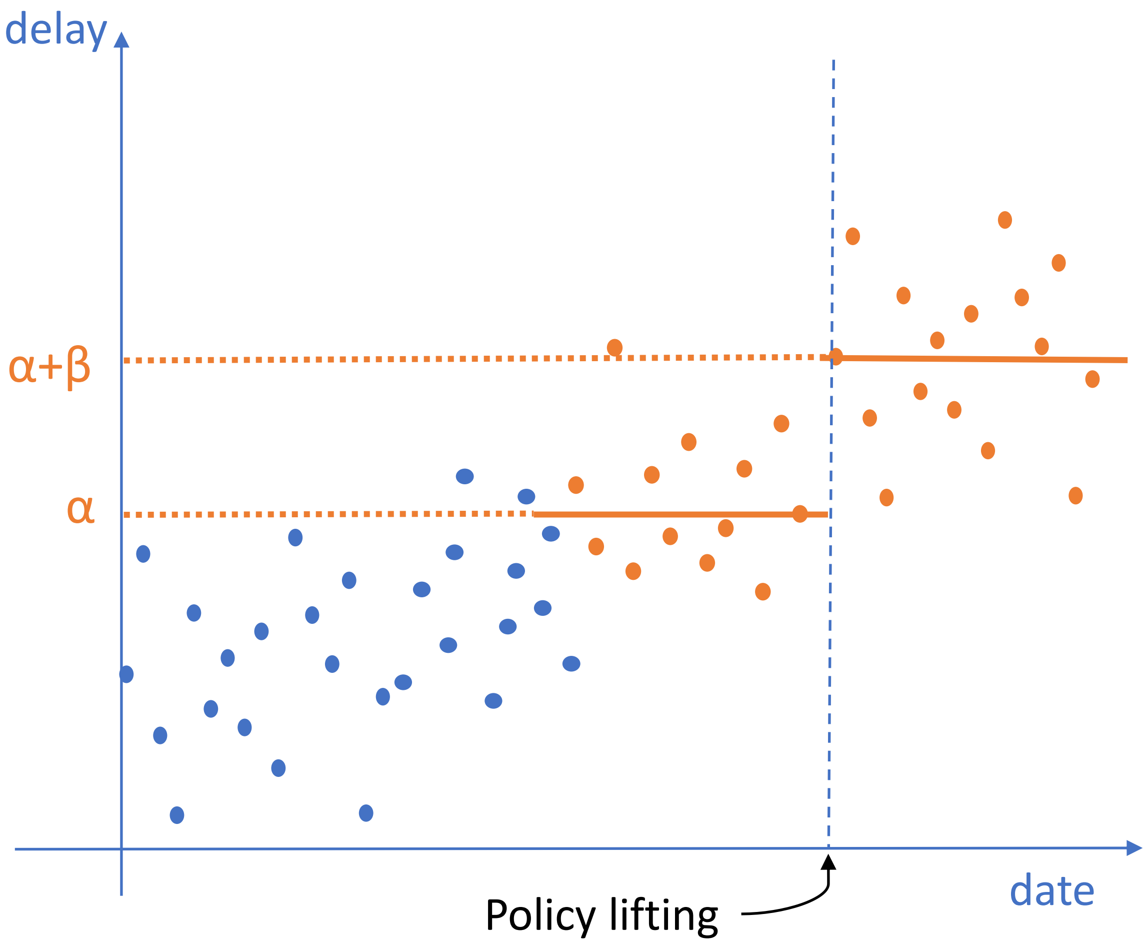
$$delay_{idh} = \alpha + \beta \cdot post_d + \gamma \cdot north_i + \varepsilon_{idh}$$

- Dependent/outcome variable: travel delay on segment *i*, on date *d* and departure hour *h*
- Independent/explanatory variable of interest: indicator for whether date *d* is after the policy lifting
  - $post_d$ = 0 before policy lifting ("control" group)
  - $post_d$ = 1 after policy lifting ("treatment" group)
- 'Conditional' on direction "fixed effect"
  - north or not
  - $north_i$ = 1 if heading north, and =0 otherwise



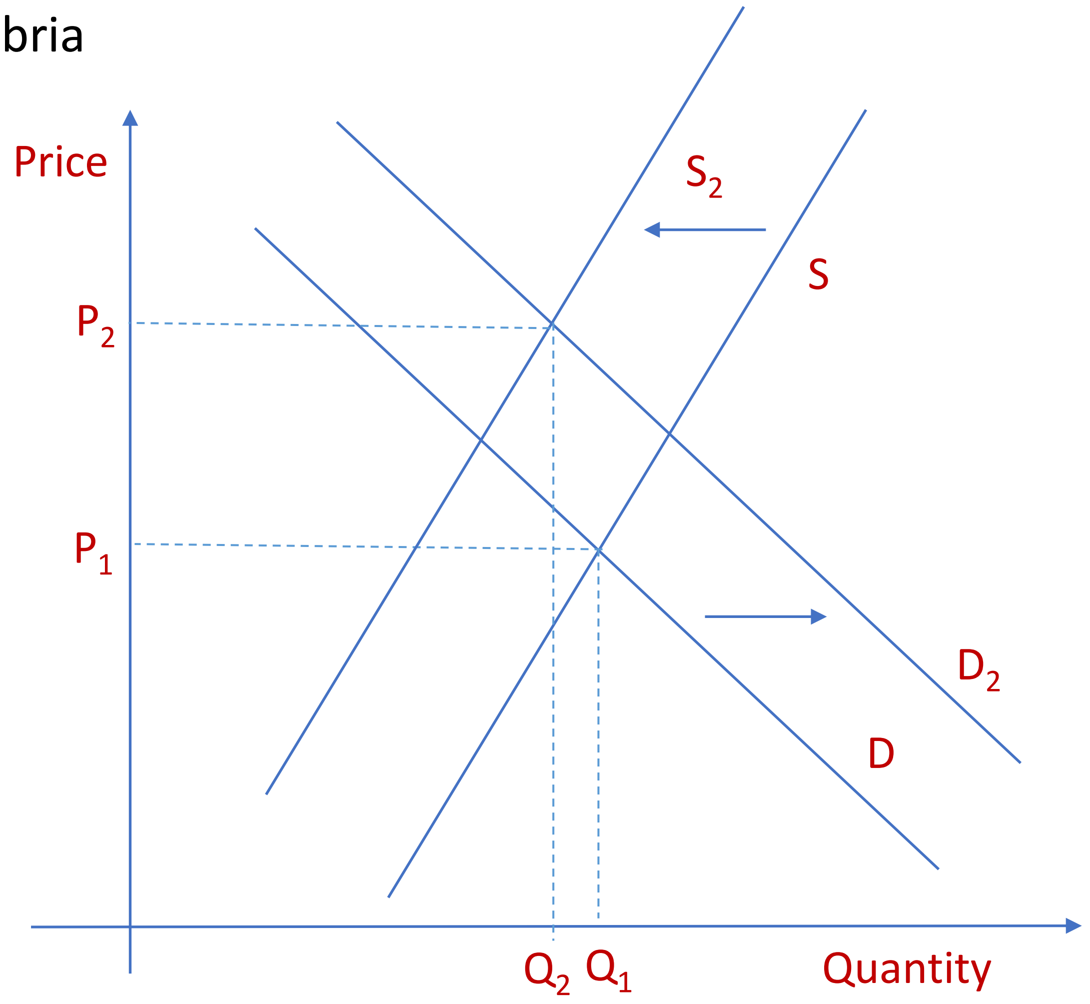
‘Identification’ of causal effect in event studies

- What if the timing of event is intended to coincide with the changes in outcomes?
  - As opposed to the changes being caused by the event?
  - Assumption: Event is uncorrelated with trends in outcomes
- What would outcomes have looked like in the absence of the policy?
  - Would the average delay have stayed at α?
  - Assumption: ‘Treated’ observations would resemble ‘control’ observations in the absence of the event



Linear regression (review)

- Most of the time, we only observe equilibrium prices and quantities.
  - But many things may have changed between two equilibria
- To estimate the supply curve, we need a shift in demand **only**.
  - So we observe points along the supply curve
- To estimate the demand curve, we need a shift in supply **only**.
  - So we observe points along the demand curve
- **Or** we need to "condition on" one of the shifts

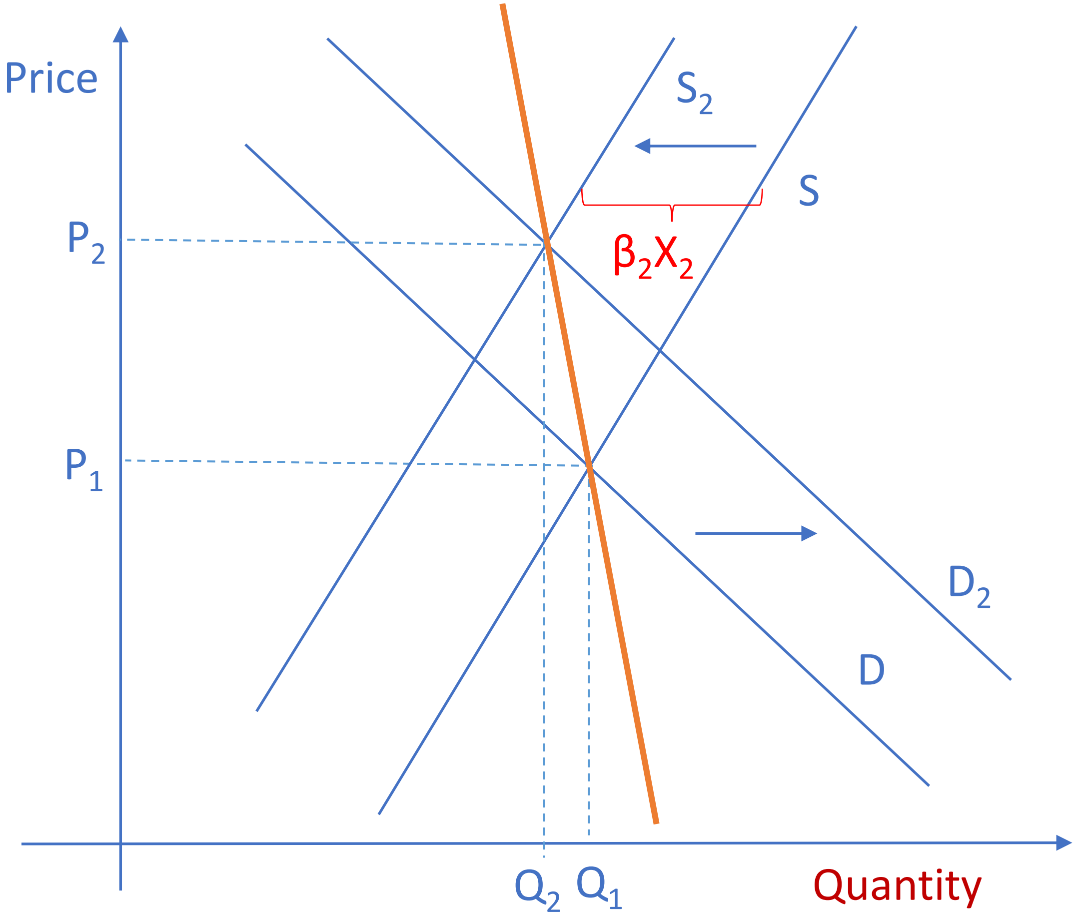


Omitted Variable Bias

e.g., if supply shift is caused by X_2: conditional on the effect of X_2 on quantity, the relationship between price and quantity lets us estimate the slope of the supply curve:

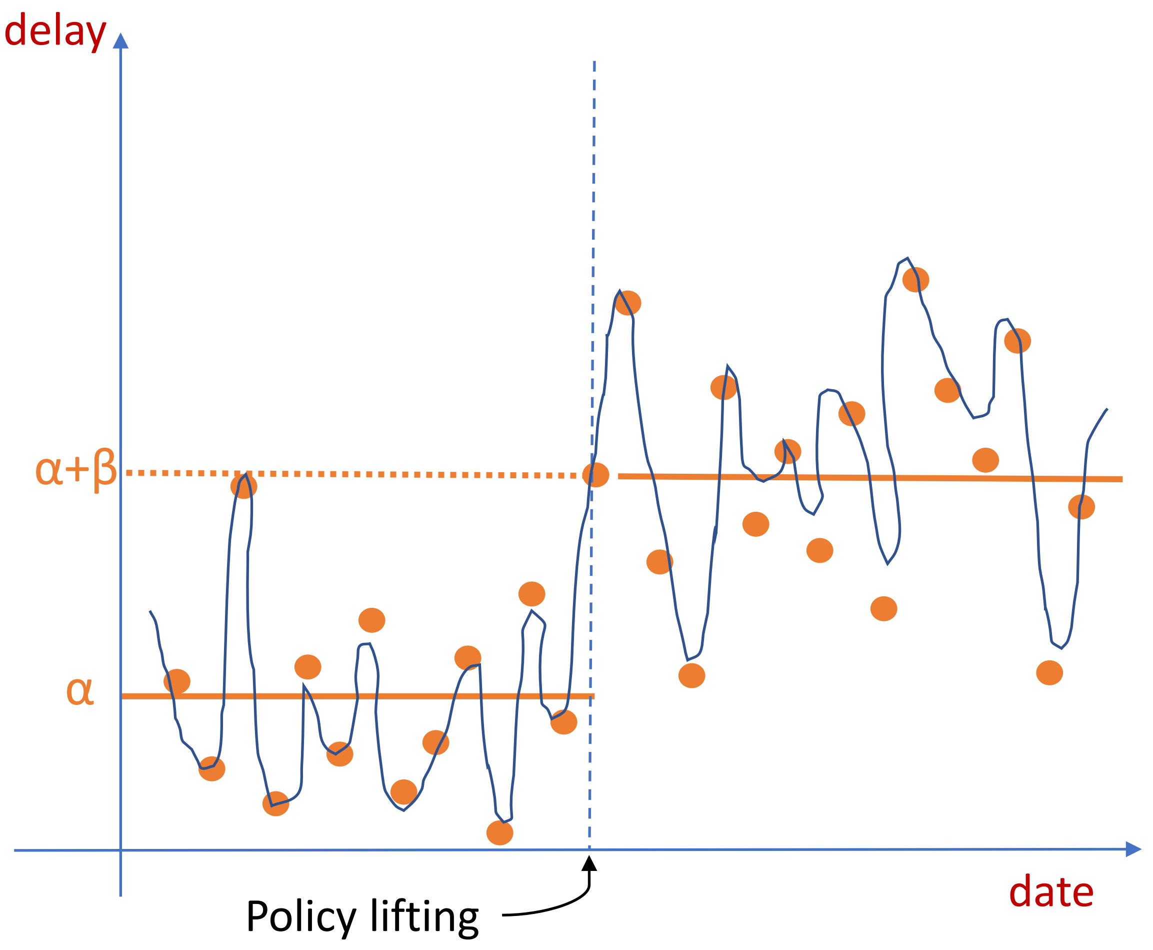
$$Ln(Quantity) = \beta_0 + \beta_1 Ln(Price) + \beta_2 X_2 + \varepsilon$$

Not including the variable X_2 in the regression can **bias** our estimate of β_1.



But we don't want to overfit the data

- We could try to fit more complicated models to the data
  - The world is usually more complicated than a single linear shift
- But our model and estimates won't be very generalizable
- Linear regressions are the most popular estimation techniques.



Linear regression estimation on Stata

Using data on individual automobiles, we want to fit the model:

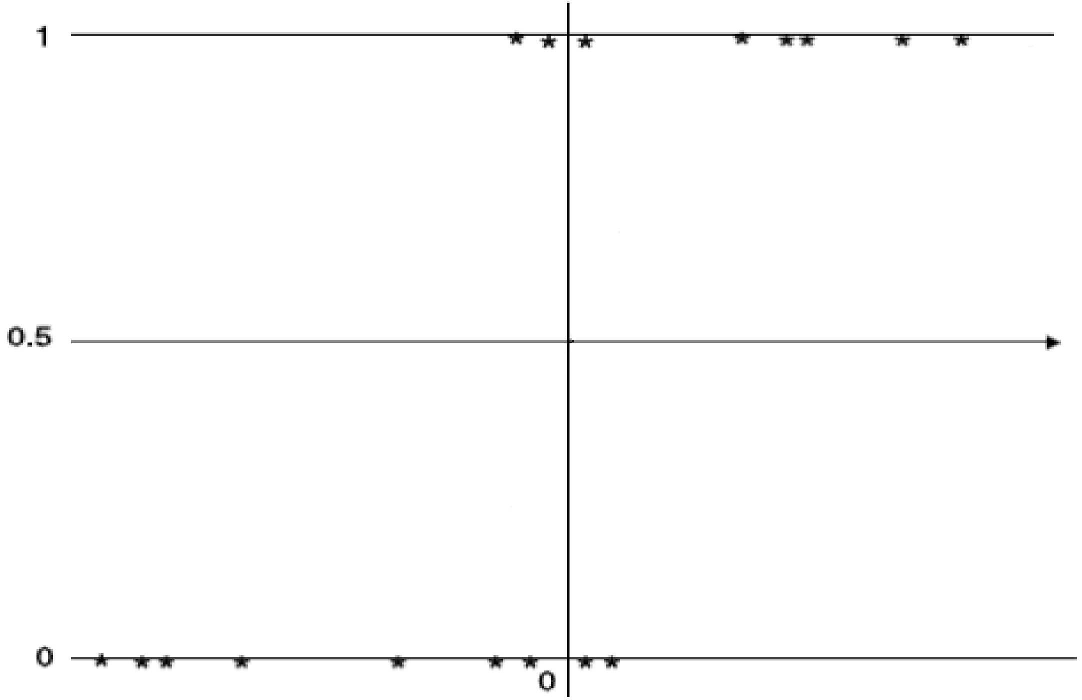
$$\texttt{mpg} = \beta_0 + \beta_1 \texttt{weight} + \beta_2 \texttt{foreign} + \epsilon$$

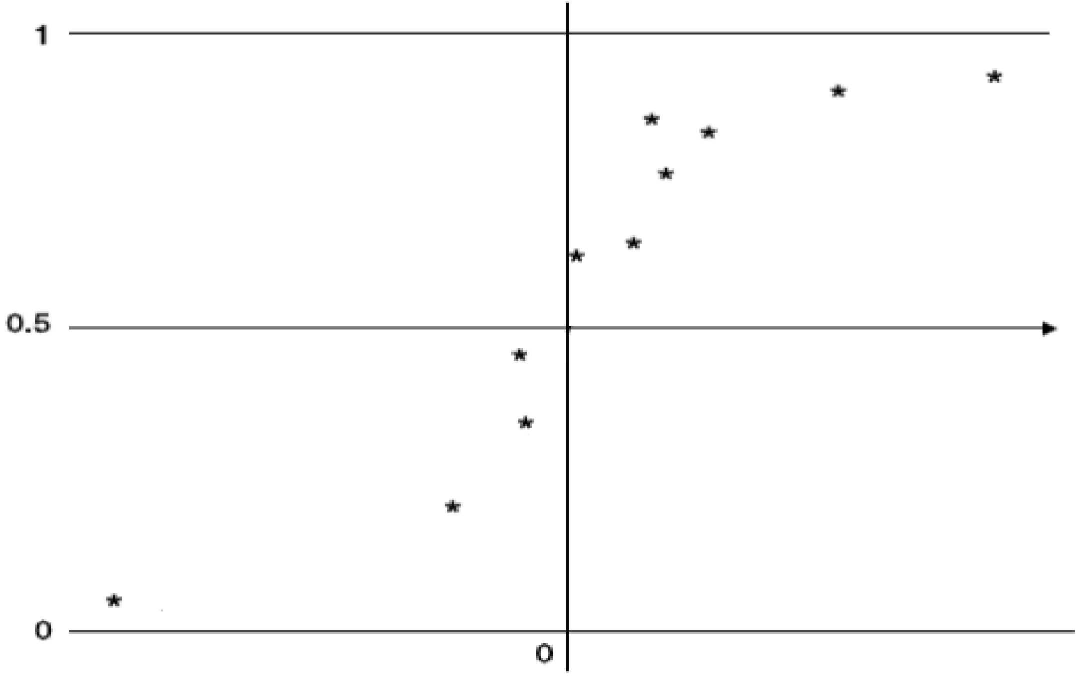
Discrete Choices

- Outcome/dependent variables need not be continuous
  - (e.g., mpg or travel time delay)

- …could represent discrete choices
  - e.g., choice of travel mode (bus vs car vs rail vs …) from surveys of individuals

- …could be bounded
  - e.g., modal / market shares, average age, … from aggregate data





Discrete Choice Models

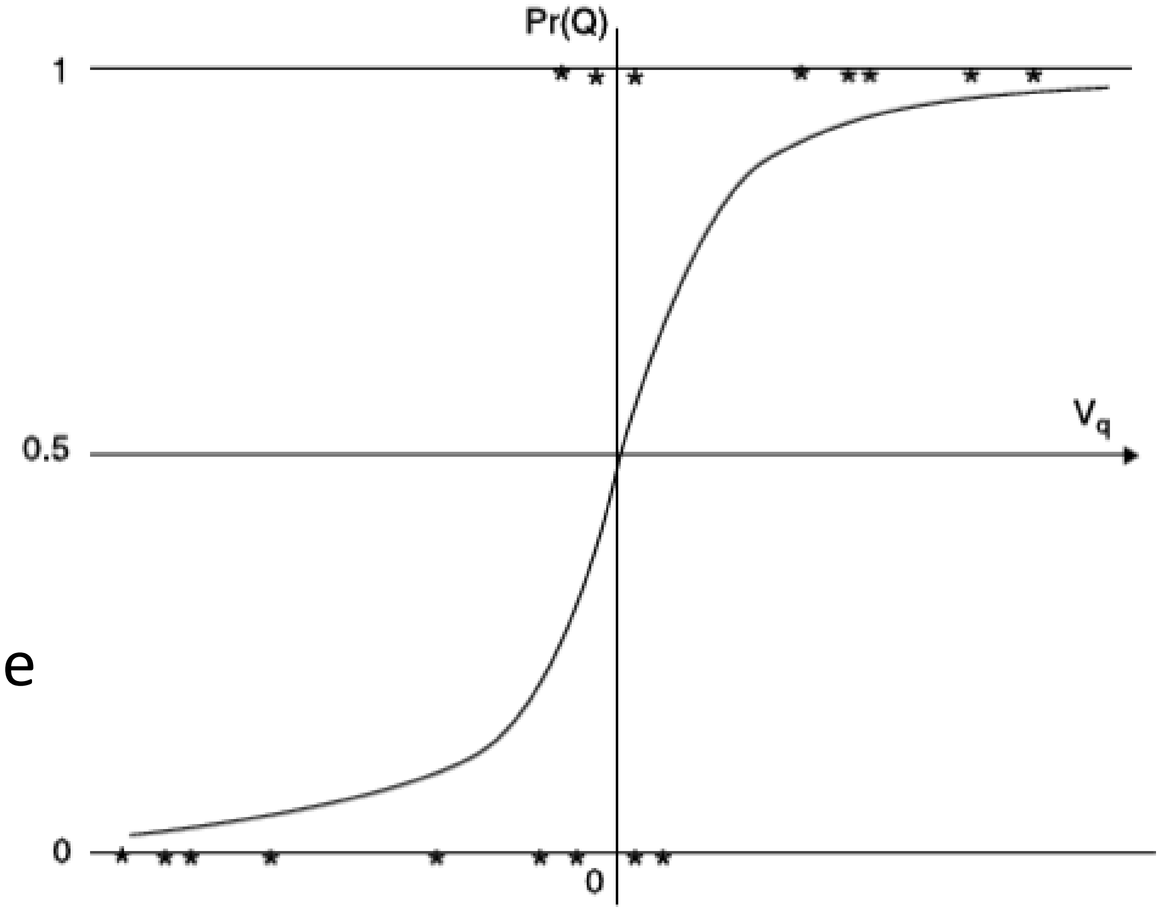
- Faced with J different alternatives (e.g., bus, car, etc.), a decision maker i chooses the one that maximizes their utility U:

$$U_{ij} = V_{ij} + \varepsilon_{ij}$$

where V_{ij} is a function of observable characteristics of the decision maker i and the alternative j

- ε is randomly distributed with an 'extreme value distribution': the probability of choosing alternative Q is

$$Pr(Q) = Pr(U_{iQ} \geq U_{ij} \text{ for all } j) = \frac{e^{V_{iQ}}}{\sum_j e^{V_{ij}}}$$



Logit regression estimation on Stata

Using data on individual automobiles,

The model that we wish to fit is

$$\Pr(\texttt{foreign} = 1) = F(\beta_0 + \beta_1 \texttt{weight} + \beta_2 \texttt{mpg})$$

where $F(z) = e^z/(1 + e^z)$ is the cumulative logistic distribution.