ECONOMICS AND MACHINE LEARNING

Use of Data in Economics (Social Sciences)

- Descriptive tool for observed economic phenomenon

- Testing models and hypotheses
  - Observe economic outcomes – economic data
  - Economic models tells us what determines these outcomes i.e. the relationship between a given economic variable and other economic and non-economic variables
    - Estimands are obtained from these models e.g. elasticity
  - Economic Model → Econometric Model
    - Equations from economic models → equations of econometric models
    - Deterministic → probabilistic statement
    - How and where do you incorporate a variable with a distribution in the econometric model
    - Identify which parameters of the econometric model map to estimands
  - Choose appropriate estimation methodology - estimator
  - Test if model implications / hypotheses for estimands hold true for estimated parameters from a given sample (estimates)
  - Inference from estimates

- Prediction
  - Use estimates to predict

Data Used in Economic Analysis

- Data
  - Information : Economic and non-economic variables
  - Unit of observation
    - Individual, household, firm, industry, geography (village, state)
    - Time dimension : Cross-section, time series, panel data
  - Dataset is a large matrix : X_{ij}
  - Sample vs. population
  - Observational vs. experimental

- Examples
  - Data on annual spending on food categories (milk, vegetables, fruits etc.) by households in 2001
  - Data on unemployment rate for 2000-12
  - Data on monthly sales of different brands of cars in India between 2001 and 2010
  - Data on daily returns for all stocks in the NSE index for 2015
  - Data on total years of schooling for all children born in 1972

- How does this map to data typically used in ML applications?

Structure of Dataset

- Typically variables as columns and observations as rows
- ID variable
- Variable definition
- Sampling methodology

ID	Variable 1	Variable 2	….	Variable k
ID1	X11	X12	….	X1k
ID2	X21	X22		X2k
IDn	Xn1	Xn2	….	Xnk

Variables used in Economic Analysis

- Economic variables
  - Price: price of goods and services, interest rate, rental rate, wage
  - Quantity: consumption, sales, production, inventories, hours/days worked, investment, advertising
  - Income (labour and non-labour earning), wealth, savings
  - Size of firm, size and number of other firms in the market

- Non-economic variables?
  - Socio-economic information: Education, years of experience, parents education, religion, caste
  - Demographic information: Age, gender, race,
  - Location (rural-urban, city, state)
  - Time (year, month)

- Level of aggregation
  - Macro data : aggregate
    - For example, Time Series: GDP, inflation, unemployment
  - Micro data: at the level of the economic agent

Data Sources – Observational or Experimental

- Survey data
  - Surveys taken at regular intervals by government organizations, research organizations
    - For example: Census, NSS surveys, Annual Survey of Industry
  - One–off surveys to address particular issues
  - Ministry of statistics http://mospi.nic.in/

- Administrative Data
  - Data collected as part of administering an initiative e.g. data on participants in a government program
  - Data that is required to be submitted to the government as part of a regulatory requirement e.g. financial information for firms to MCA

- Publicly available data compiled by data collection firms
  - CMIE

- Data which is a by-product of digital interactions (Big Data)
  - Data from transactions on e-commerce websites
  - Data from use of websites
  - Scanner data

Survey Data

- Method of collecting survey data: surveyor can use paper forms or digital devices (Fitbit)

- Can choose to survey units such that the sample has desired properties
  - Representative sample
  - Random sample vs. weighted to oversample certain population groups
  - Contrast with some of the digital data

- Can use it to get qualitative information as well

- Challenges
  - Framing of questions appropriately to get the desired information
  - Accuracy, adequate cross checks
  - Completeness
  - Consistency across different datasets
  - Contrast with administrative data, particularly if it is digitally collected

Size of Datasets

- Two dimensions: Units of observation, Number of variables

Data points

- Anecdotal evidence: few data points
- Minimize cost of collecting data vs. reliability of results
  - Minimum sample size n=30?
  - Annual Survey of Industries (NSS): 63296 cases (factories) in 2014-15 survey
  - Consumer survey (NSS): 83600 cases (households) in 2014-15 (Household expenditure on Services and Durable goods)
  - Census: 1210854977 individuals, 1931119328 households in 2011 Census
  - Digital data: Price discrimination by Netflix studied using data for 60,000 computer users
    - Need for sampling with some types of big data e.g. searches on Google

Number of variables

- Take advantage of the opportunity to collect information on as many related and relevant variables as possible
  - Variables depend on hypotheses being tested, issues being assessed
  - Consumer survey: 156
  - Census: 40
  - Digital data: Price discrimination by Netflix studied using data on 5000 web-browsing variables

Big Data

- “computer-mediated transactions" generate huge amounts of data
  - All online behaviour is recoded: searches, clicks, sales
  - Text messages, cellphone usage, geo-locations
  - Offline sales data : scanner data
  - Employment records, electronic health records

- What is different?
  - Data is now available faster, sometimes in real time: how can this facet be used for policy analysis?
  - Has greater coverage and scope,
  - Includes new types of observations and measurements e.g. social network data
  - Structure is different e.g. entire shopping history for a consumer, data not rectangular

- Issues unique to big data
  - Large size of data (observations and variables) requires more powerful data manipulation tools.
  - More potential covariates than needed for estimation so variable selection methods become important.
  - Large datasets may allow one to consider more flexible/complex relationships than simple linear relationships

Data for Descriptive Analysis

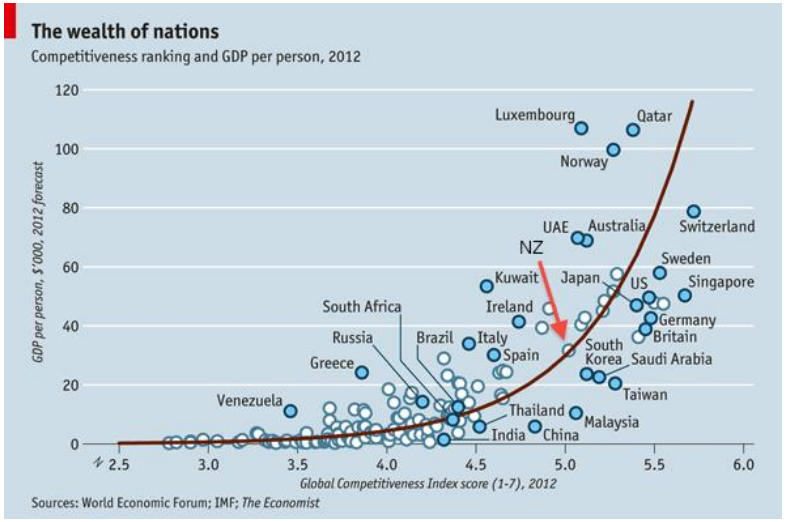
- Descriptive analysis to provide background for the issue being addressed: understand the data
  - Generate questions: Motivation for research
  - Context for answers

- Graphs
  - Scatter plots
  - Distribution: Histograms
  - Time trends
  - Plotting X vs. Y

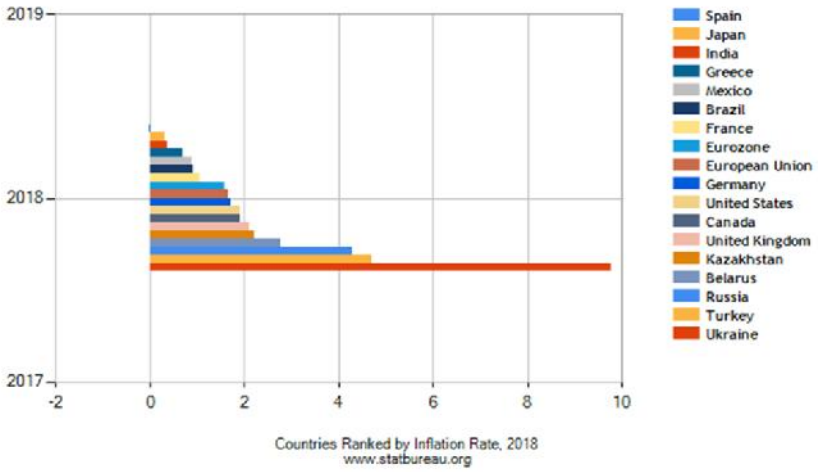
- Summary statistics to capture distribution: overall or in categories
  - Mean, median etc.
  - Variance
  - Measures of skewness etc., percentiles
  - Correlation
  - Regression?

Plots and Graphs

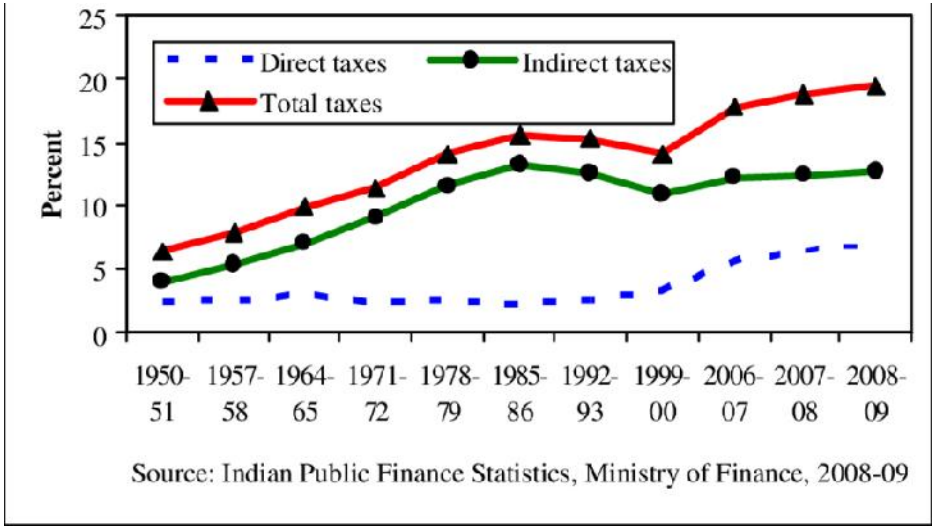
- Scatter Plot



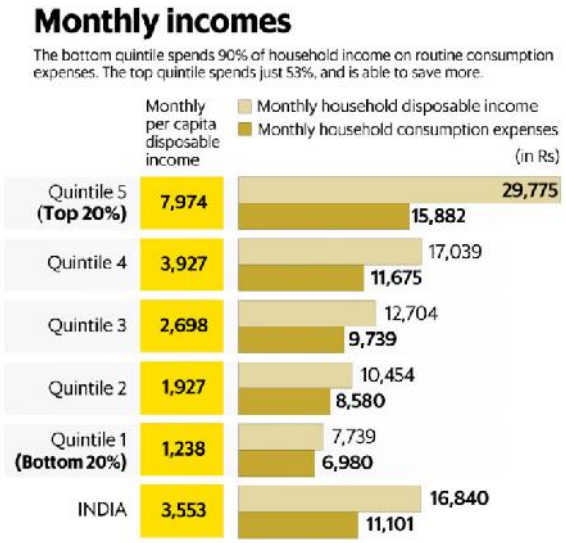
Bar chart



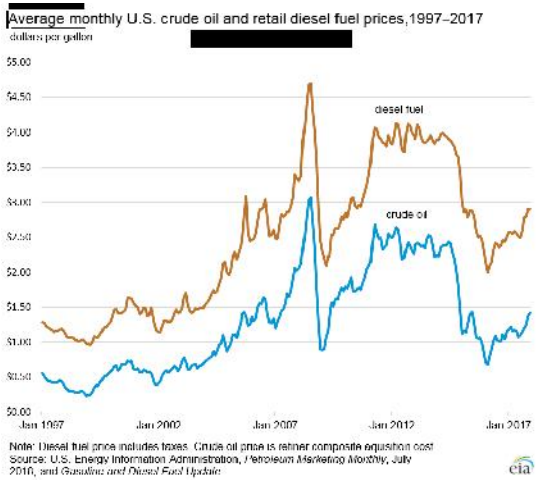
- Trend (India)

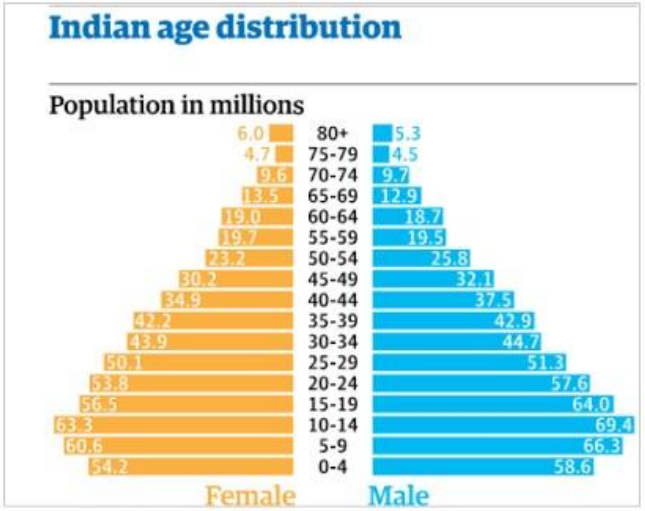


Income distribution (India)

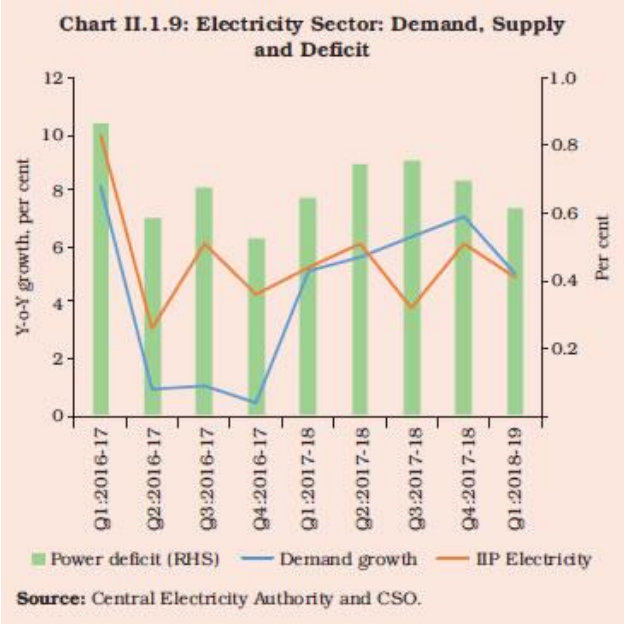


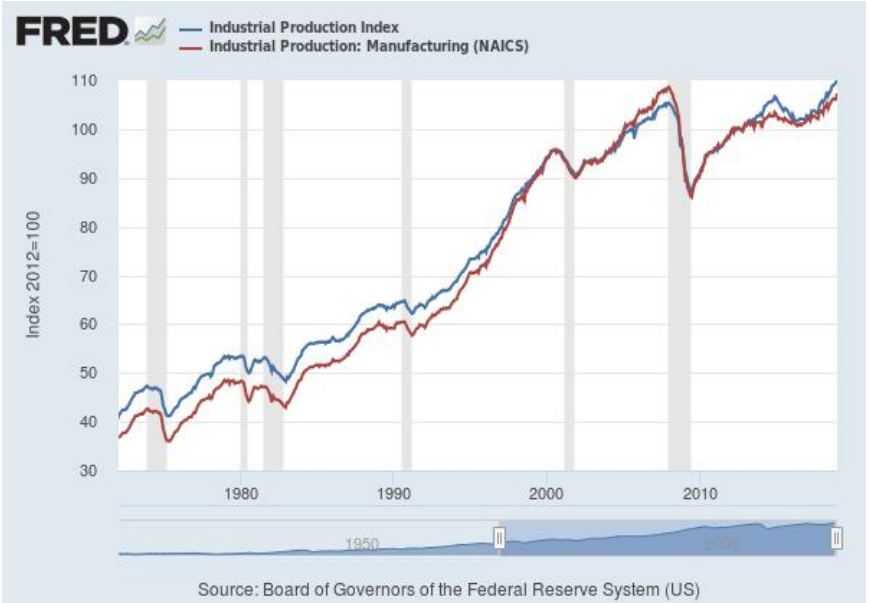
Plots and Graphs





Source: Economic Census 2011





Source: Board of Governors of the Federal Reserve System (US)

Summary Statistics

TABLE I

HOUSEHOLD ECONOMIC ACTIVITIES

	Landed	Landless
1. Percentage of sample	62	38
2. Percentage of households that engage in: (based on previous year)		
Agriculture	100	0
Grow Rice	70.9	
Animal Husbandry	63	29
Fishing	5	4
Manufacturing	9	8
Transport	2	3
Services	2	3
Trade	14	22
Other	1	2
Wage Earnings	48	73
3. Percentage of households with: (based on previous week)		
Male member earning wage off-farm	24.4	57.1
Female member earning wage off-farm	13.1	39.6
Male member in off-farm agricultural employment	13.4	32.3
Female member in off-farm agricultural employment	9.5	31.3
Male member in off-farm nonagricultural employment	18.4	46.7
Female member in off-farm nonagricultural employment	15.9	26.4
Male member working on own "farm"	70.6	6.7
Female member working on own "farm"	48.3	3.4
Unemployed member	4.6	3.2
Underemployed (self-reported) member	11.0	16.9
4. Percentage of individuals: (based on previous week)		
Male Labor Force Participation Rate[a]	70.7	70.4
Female Labor Force Participation Rate	33.3	39.8
Percentage of Male Labor Force Earning Wage	23.7	69.4
Percentage of Female Labor Force Earning Wage	27.4	78.1
Unemployment Rate	3.2	2.5
"Underemployment Rate"	8.7	16.9

TABLE V

DIFFERENT SUBSAMPLES OF YOUNG MEN FROM THE NATIONAL LONGITUDINAL SURVEY

MEANS AND STANDARD DEVIATIONS[a]

Variable	All	All Valid *IQ* Scores	Brothers (pairs) Total	Brothers (pairs) Within	Not Enrolled in 1969 All	Not Enrolled in 1969 With Valid *IQ* Scores
N	4601	3025	580		2026	1362
AGE 69	21.2	21.5	20.3		22.2	22.3
	3.2	3.0	2.3	1.4	3.2	3.2
EXSC[b]	13.8	14.4	14.8		12.7	13.4
	3.0	2.4	2.3	1.1	2.8	2.3
S69[c]					11.6	12.5
					2.4	1.9
S66[d]	10.7	11.5	11.3		10.8	11.6
	2.4	1.9	1.7	1.1	2.4	2.0
EXLOMY[e]	8.61	8.65	8.67		8.53	8.58
	.403	.386	.404	.270	.389	.366
LW69[f]					5.60	5.68
					.426	.398
KWW[g]	33.3	35.5	34.9		33.0	35.1
	8.6	7.6	7.7	4.5	9.0	7.9
IQ[h]		101.2	102.8			97.7
		15.9	15.9	7.5		15.3
FOMY14[i]	5120	5372	5418		4826	5095
	1951	1960	2179		1779	1777
BLACK	.27	.17	.20		.28	.19
CULTURE[j]	2.2	2.4	2.5		2.0	2.3
	.97	.80	.76		1.0	.9
SIBLINGS	3.3	2.9	3.6		3.6	3.1
	2.6	2.3	2.1		2.7	2.4
EXP69[k]					4.0	3.7
					3.1	2.8
XBT[l]					.70	.72
					.27	.18
SMSA[m]			.67		.61	.65
ROS[n] or *RNS*[o]	.34	.33	.32		.41	.33

[a] The lower number in a pair of numbers is the standard deviation.
[b] *EXSC* = Expected total schooling to be completed eventually, in years.
[c] *S69* = Schooling completed in 1969, in years.
[d] *S66* = Schooling completed in 1966, in years.
[e] *EXLOMY* = Logarithm of the 1959 median earnings (in dollars) of all males in the occupation expected (desired) at age 30.
[f] *LW69* = Logarithm of hourly earnings (in cents) on the current or last job in 1969.
[g] *KWW* = Score on the "knowledge of the world of work" test, administered in 1966.
[h] *IQ* = Score on IQ-type tests, collected from the high school last attended by the respondent.
[i] *FOMY14* = Occupation of father or head of household when respondent was 14, scaled by the median earnings of all United States males in this occupation in 1959.
[j] *CULTURE* = Index based on the availability of newspapers, magazines, and library cards in the respondents home.
[k] *EXP69* = Post-school work experience. Estimated on the basis of the work record (in weeks) since 1966 and the date of first job after school and the date stopped school. Truncated at age 14, if respondent started working earlier. In years.
[l] $XBT = e^{-0.1 \times EXP69}$
[m] *SMSA* = Respondent in *SMSA* in 1969.
[n] *ROS* = Respondent in South when 14, columns 1–3.
[o] *RNS* = Respondent in South now (1969), columns 6–7.

Econometric Modeling

- Economic models → relationships
  - Exogenous variables and parameters determine endogenous variables

- Examine patterns and relationships in data using econometric models
  - Descriptive econometric model
  - Derive econometric model from theoretical (economic) model
    - Help identify where individual heterogeneity should come in
    - Identify sources of simultaneity, selection bias

- Econometric models are used for
  - **Estimation** of effects (parameters/ estimands)
    - Coefficient estimates and standard errors
  - **Inference** from these estimates
    - Size, statistical significance
    - Interpretation: Can a causal inference be drawn?
  - **Prediction**
    - Predicting outcomes within sample
    - Predicting outcomes in a counterfactual scenario
    - Prediction when testing theories that are about prediction as in predicting stock returns

Demand Functions for a Representative Consumer

- Cobb-Douglas utility function
  - Demand function: $X= (\alpha / (\alpha + \beta)) (I / P_x)$
- Leontief utility function
  - Demand function: $X= \beta I / (\beta P_x + \alpha P_y)$
- CES utility function
  - Demand function: $X=(\alpha/P_x)^{\delta} (I/(\alpha^{\delta} P_x^{1- \delta} + \beta^{\delta} P_y^{1- \delta}))$
- Stone Geary Utility Function
  - Demand function: $X= \gamma_x + \alpha (I-\gamma_x P_x - \gamma_y P_y) / P_x$
- AIDS

$$w_i = \alpha_i + \sum_{j=1}^{n} \gamma_{ij} \ln p_j + \beta_i \ln(X / P)$$

Relationships and Parameters

- Labour supply curve
  - $L = \beta * w + \gamma * P_y + \delta * I$
- Return to schooling
  - Ln W = = α + β* schooling + γ* experience
- SS curve (perfect competition)
  - $mc = P_x = \alpha + \beta * w + \gamma * r$
- *These are deterministic models / relationships between variables and parameters*
  - Economic theory tells us that the RHS variables are **"causing"** the observed endogenous variable
  - Given a few data points for exogenous and endogenous variables maybe we can solve for parameters
  - But does the economic model perfectly explain the observed data?
  - Data is noisy – has a distribution

Causal Effect

- Consider the effect of X on Y
  - Economic theory examines effects under ceteris paribus i.e. other factors being equal
    - Effect of price on quantity demanded holding income, price of other goods constant determines causal effect of price
  - Policy impact assessment also requires ceteris paribus assumption
    - Impact of skill upgrading program on wages holding education, experience etc. constant determines causal effect of the policy
  - Challenge is to hold all other factors constant in the empirical study so that a causal inference can be made
  - Ideal way to determine a causal effect is an experiment as is done in physical sciences
    - Impact of schooling: Change their schooling in a perfectly-controlled environment, or change schooling randomly so that those with different levels of schooling would be otherwise comparable.
    - Difficult to implement an experiment in social sciences for most variables of interest
    - When carefully applied econometric methods can simulate a ceteris paribus experiment so that a causal inference can be made
  - Alternately think of outcomes Y for an economic agent under alternative values of X
    - For example, outcomes under hospitalization and no hospitalization
    - Differences in potential outcomes are causal effect of hospitalization

Data Distribution

- What process generates the observed distribution of (economic data) Y?
  - Is Y related to variable X?
    - Interested in conditional expectation of y given x

      $E(y \mid x) = y\int f(y|x)\, dy$
  - Predictive vs. causal modelling

- Statistical Structure: Model (and estimate) joint population density of observed economic data X and Y i.e. f(x,y) as any parametric statistical distributions or a non-parametric distribution that allows the joint density to be estimated flexibly
  - For example, descriptive analysis (econometric model) estimates the joint density of x and y as
    - Best linear predictor (BLP) of endogenous variable y given predetermined variable x (Predictive modelling)

      $BLP(y \mid x) = \alpha + \beta * X$
    - Under certain conditions, least squares regression yields consistent estimates of BLP(y|x)
    - Without an economic model, coefficient estimates from this regression are merely statistics.
    - No causal statement can be made about β
- Economic Structure (Causal modelling)
  - Consider E(y | x) as some function of X and Y e.g. from $Y = \alpha + \beta * X + \varepsilon$
    - Also characterizes joint distribution of economic data
    - Derived from an economic model : structural econometric model
    - **Under certain conditions causal effect can be estimated**

- BLP and E (y |x)
    - In general , $BLP(y \mid x)$ and $E(y \mid x)$ are not equivalent
    - In general, $\partial BLP(y \mid x)/\partial x \neq \partial E(y \mid x)/\partial x$.

      Non-linearity of conditional expectation

Structural Econometric Model

- Objective: Model joint distribution of observed data X and Y i.e. f(x,y) after incorporating economic model → Economic activities place restrictions on population joint distribution
    - For example, economic theory places restrictions on the joint distribution (relationship) of consumption, prices and income.

$U(Q_x, Q_y; \theta)$ and $x = h(P_x, P_y, I; \theta)$.
Objective is to estimate θ

This model does not fully explain observed purchases Q_x or Q_y
Enrich model or add "error" that reflects variables outside of the model $U = U(Q_x, Q_y, \varepsilon; \theta)$ and $x = h(P_x, P_y, I, \varepsilon; \theta)$

If ε is unobservable, add assumption about the joint population distribution for $(\varepsilon, P_x, P_y, I)$ e.g. a specific parametric distribution for random variable ε and hence the joint distribution

→ a joint distribution for the observed data $f(Q_x, P_x, P_y, I)$ or conditional distribution $f(Q_x \mid P_x, P_y, I)$
Solution produces a joint density of all observables that has positive support on all possible realizations of these variables

Two questions:
Is the econometric model consistent with observed joint density $f(Q_x, P_x, P_y, I)$
Can you get a consistent estimate θ from the structure of $f(Q_x, P_x, P_y, I)$?

Stochastic Element (Noise) in Econometric Models

- Economic model → Econometric model
  - Deterministic relationship → stochastic relationship

- Econometric models include unobservable variables / uncertainty
  - Unobservables account for the fact that the economic model does not perfectly fit observed data

- Unobservables could arise from
  - Stochastic element in the economic model
  - Unobserved or omitted variables in the model
    - Factors unobserved by the econometrician/researcher
    - Factors reflecting uncertainty on part of economic agents and researchers
  - Measurement error in X or Y

- With no unobservables (i.e. no stochastic variable) a small number of observations may be used to solve for parameters
  - No statistical analysis is needed

- With unobservables (i.e. models include stochastic variables) we need assumptions about distribution of stochastic elements
  - Distributional assumptions drive the statistical analysis
  - Source of uncertainty has implications for parameter estimation approaches

Source of Stochastic Element

- Uncertainty from the economic model e.g. through modeling of heterogeneity in parameters
  - Stone Geary Utility Function:

    $U= (x- \gamma_x)^{\alpha} (y- \gamma_y)^{\beta}$ where $\alpha + \beta = 1$

    $\alpha = \alpha_0 + \alpha_1$*age+ α_2*gender + α_3 *marital status + ε

    $\beta = \beta_0 + \beta_1$*age+ β_2*gender + β_3 *marital status + ω

  Where ε, ω are random variables which reflect distribution of unknown elements of preference

  Demand function: $X= \gamma_x + \alpha (I-\gamma_x P_x - \gamma_y P_y) / P_x$

  - Constant elasticity demand function

    $\ln X= \beta$ *$\ln P_x + \gamma$*$\ln P_y + \delta$ *I

    $\beta = \beta_0 + \beta_1$*age+ β_2*gender + β_3 *marital status + ε

    Where ε is a random variable which reflects distribution of unknown elements of preference i.e. unobserved heterogeneity

- Unobservable variables

  In Wage= $\alpha + \beta$ *Schooling + γ* Experience + ε

  Where ε is a random variable which reflects distribution of unobservable variables e.g. ability

  Two individuals with the same schooling and experience do not have the same wage

- Measurement error in variables

  $\ln X= \beta$ *$\ln P_x + \gamma$*$\ln P_y + \delta$ *I

  $P_x = \tilde{p}_x + \varepsilon$

  Where ε is a random variable which reflects measurement error in P_x

  $\ln X= \beta$ *$\ln \tilde{p}_x + \gamma$*$\ln P_y + \delta$ *I + ln ε

- Typically the stochastic element comes from more than one source

Steps to Estimation

- Having defined a stochastic model
  - Select functional forms (and make other assumptions about economic environment as needed)
    - Flexibility
    - Economically realistic: Don't assume a functional form that delivers the desired empirical result
    - Ease of estimation
    - Estimation transparency
  - Select distributional assumption
    - If we completely specify the joint distribution of the economic model errors (unobservables), the model implies a conditional distribution of the observed endogenous variables given the exogenous variables.

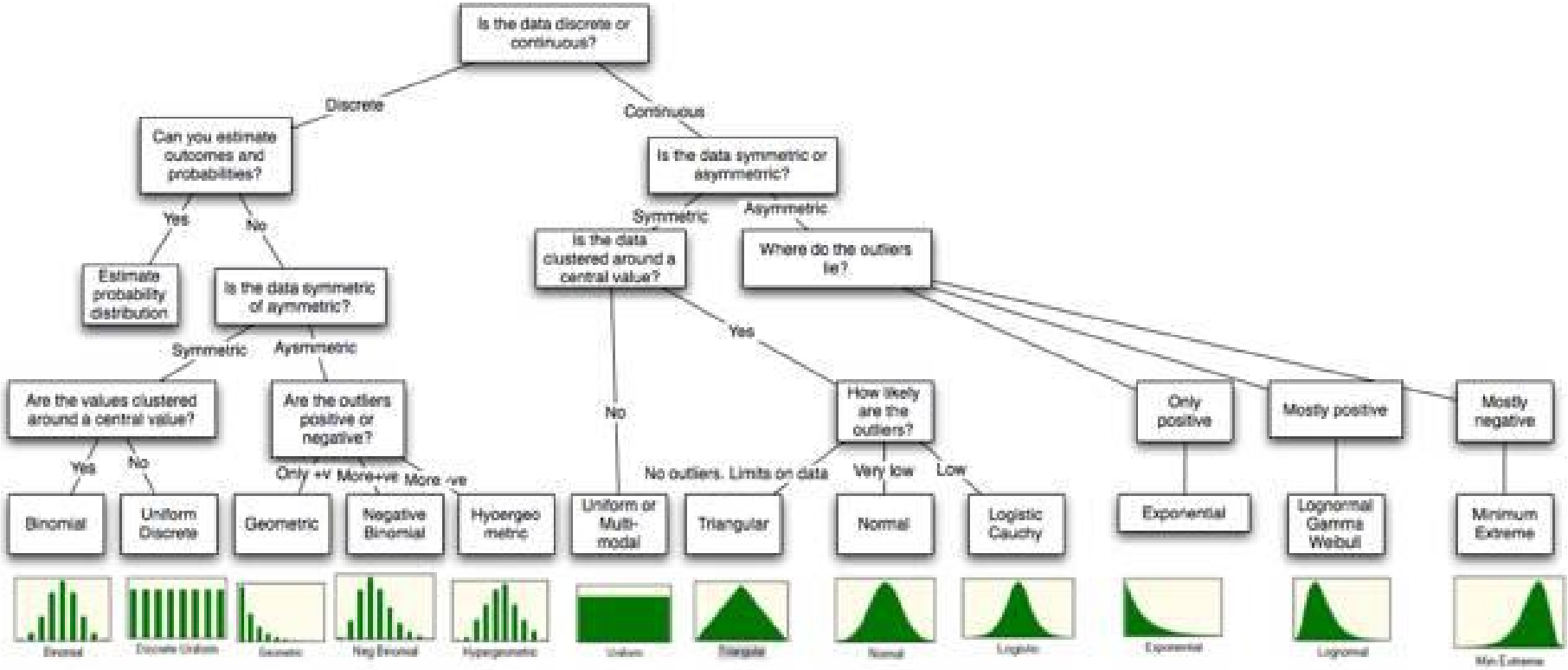
      ε_i is an independently identically distributed random variable (i.i.d.)

      Parametric assumptions : Normal, logistic, etc.

      Consider implication of distributions for the random variable: truncation, count vs. continuous data

    - Less structure on joint distribution e.g. only moment restrictions. Other forms of non-parametric distributions
  - Select an estimation technique
    - OLS
    - MLE
    - Method of Moments
  - Setting and economics should motivate specific probability models and estimation strategies

Statistical Distributions



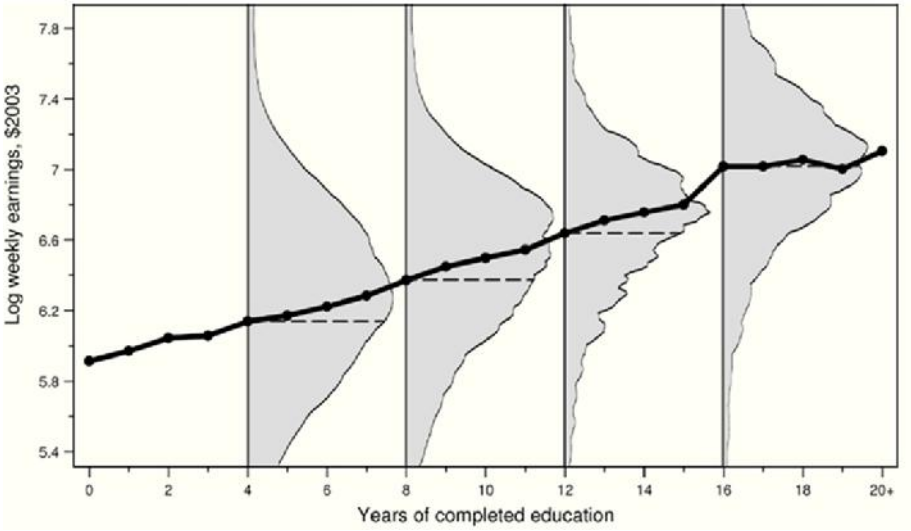
Conditional Expectation Function

- Y_i and X_i have a joint distribution
  - Goal is to estimate the density (parameters)

- Conditional expectation function
  - CEF = $E(Y_i/X_i)$
  - Effectively summarizes relationship between X and Y

$$Y_i = E[Y_i|X_i] + \varepsilon_i,$$

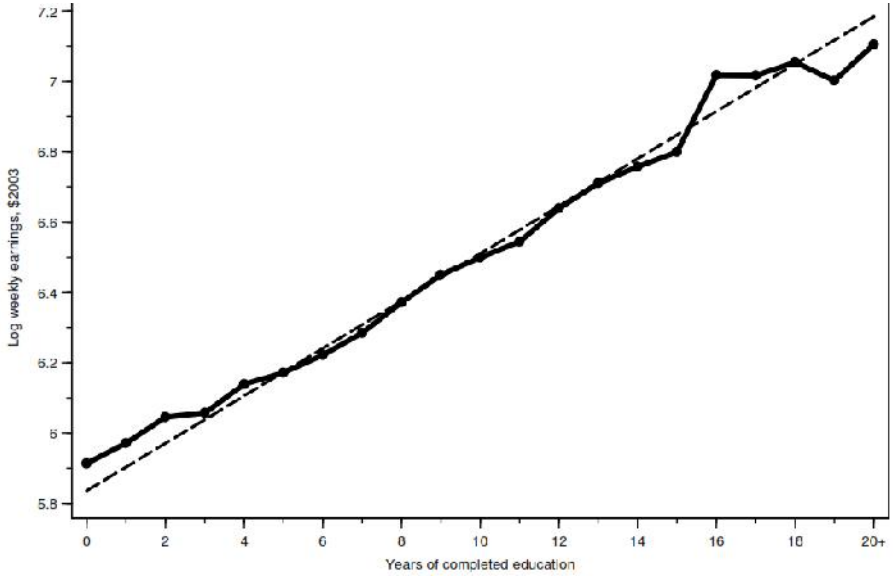
  - Population average: object of interest
  - CEF is random because X_i is random



- Any random variable (Y_i) can be expressed by a piece that is explained by X_i (i.e. CEF) and a piece that is uncorrelated with any function of X_i.
  - $E(Y_i) = E\{E(Y_i/X_i)\}$
  - And $V(Y_i) = V(E[Y_i|X_i]) + E[V(Y_i|X_i)]$
  - Properties hold in sample and population, no linearity assumption

Linear Regression

- *CEF is the best predictor of Y_i given X in class of all functions of X_i , MMSE*
- *Regression is the best linear predictor of Y_i given X*
  - Population regression
    - $Y_i = X_i'\beta + u_i$ where u_i is the "disturbance" or "error" or "unobserved"



- Motivating regression
  - If CEF is linear, CEF== population regression
    - CEF linear e.g. when Y,X jointly normal
  - Even if CEF is non-linear, population regression is the best linear approximation
- Regression is causal if CEF is causal

Linear Regression

- Population regression $Y_i = X_i'b + u_i$

$$\beta = \arg\min_b E\left[\left(Y_i - X_i'b\right)^2\right]$$

$$E\left[X_i\left(Y_i - X_i'b\right)\right] = 0.$$

$$\beta = E\left[X_i X_i'\right]^{-1} E\left[X_i Y_i\right].$$

$$\beta_k = \frac{Cov\left(Y_i, \tilde{x}_{ki}\right)}{V\left(\tilde{x}_{ki}\right)}$$

where $\tilde{x}_{ki}$ is the residual from a regression of x_{ki} on all the other covariates.

y	$x_1, x_2, \ldots, x_k$
Dependent variable	Independent variables
Explained variable	Explanatory variables
Response variable	Control variables
Predicted variable	Predictor variables
Regressand	Regressors

Regression Coefficients

- Don't know CEF or population regression
- Use samples to draw inferences about CEF or population regression coefficients
  - Repeated samples: independent draws of $[Y_i\ X_i]$
  - Law of large numbers: sample moments go to population moments
  - Can estimate regression coefficients using different estimation methods

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + u,$$

- $\beta_i = dy/dx_i$

- Ceteris paribus interpretation
  - β_i is effect of X_i after controlling for all other Xs
  - β_i is effect of X_i holding constant all other Xs
  - β_i is a partial effect of X_i after effect of X_j on X_i netted out

- Linear in parameters, can have non-linear X_i 's

Regression Coefficients

- β reflects correlation between Y and X and not necessarily causal effect of X on Y

- RHS: Continuous variables vs. discrete variables
  - Dummy variable e.g D_i =1 if male and = 0 if female
  - ln $wage_i$ = α + β * $schooling_i$ + γ * $experience_i$ + δ *D_i + u_i
  - Dummy variables affect intercept
  - Dummy variables interacted with continuous covariates
    - ln $wage_i$ = α + β * $schooling_i$ + γ * $experience_i$ + δ *D_i + θ *D_i * $schooling_i$ + u_i
    - Effect of one more year of schooling for males = β + θ

- When do these coefficients reflect causal effects?

Estimation

- Determining the parameter value using sample data
  - Mean / median is an estimate of a central tendency
  - Sample average is an estimate of the population mean
  - Parameters of the CEF or the linear regression can be estimated in multiple ways
  - Estimates evaluated in terms of the bias ($E\hat{\theta} - \theta$), variance.

Estimation- Ordinary Least Squares

- Minimizing mean square error in the population

$$\beta = \arg\min_b E\left[\left(Y_i - X_i'b\right)^2\right] \qquad \beta = E\left[X_i X_i'\right]^{-1} E\left[X_i Y_i\right].$$

- Sample analog: Ordinary Least Square (OLS)
  - Consider OLS estimates of β_0 and β_1 in a one variable regression

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

  - Choose β_0 and β_1 to minimize sum of squared residual

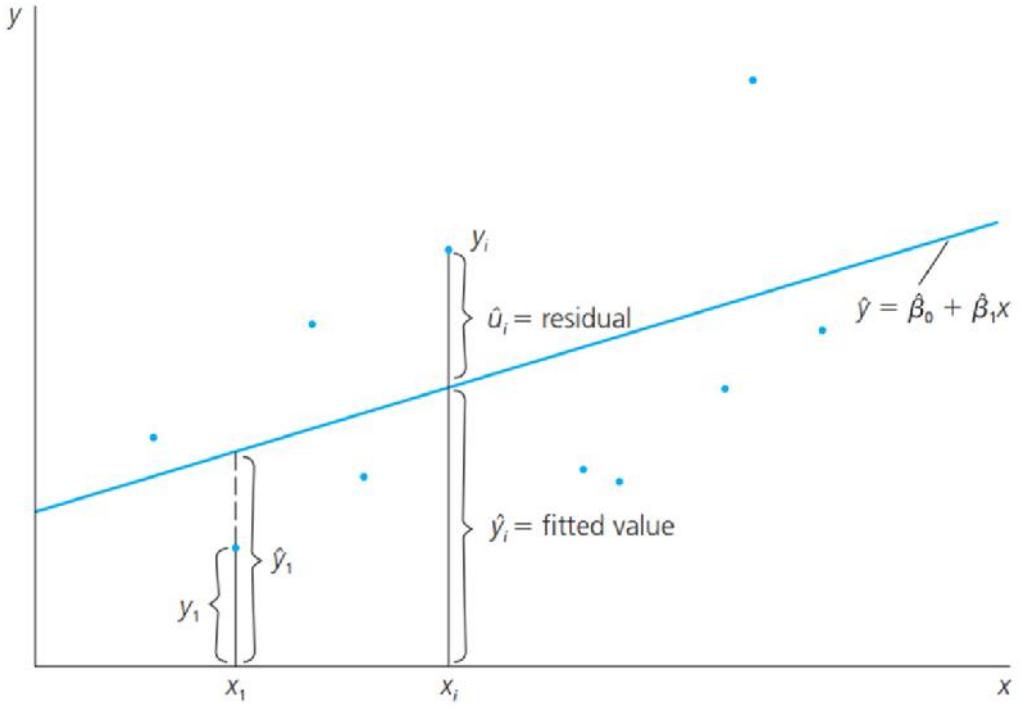
$$\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2,$$

  - FOC:

$$-2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$-2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0.$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}. \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$



OLS Estimates

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + u,$$

$$\hat{\beta} = \left[\sum_i \mathrm{X}_i \mathrm{X}_i'\right]^{-1} \sum_i \mathrm{X}_i \mathrm{Y}_i.$$

$\beta = (X'X)X'Y$

Model	Dependent Variable	Independent Variable	Interpretation of β_1
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
Level-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
Log-level	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

- Relationship between estimated coefficients
  - ln $wage_i$ = $\tilde{\alpha}$ + $\tilde{\beta}$ * $schooling_i$ + $\tilde{\gamma}$ * $experience_i$ + u_i
  - ln $wage_i$ = $\bar{\alpha}$+ $\bar{\beta}$ * $schooling_i$ + u_i
  - $\bar{\beta}$ = $\tilde{\beta}$ + $\tilde{\gamma}$ * λ where λ is co-efficient from regressing experience on schooling

OLS – Assumptions and Properties

- Assumptions
  - E(u/X)=0; E(u)=0
  - **E(X'u)=0;**
  - $E(u^2 \mid x)= \sigma^2$ (Homoscedasticity)
  - Cov (u_i , u_j) = 0 (No autocorrelation)
  - i.i.d.
  - Variation in X
  - Matrix X has full rank i.e. no multicollinearity
  - N>= k+1
  - Normality? (Enables us to get exact sampling distribution of parameters)

- Estimator Properties (Best Linear Unbiased Estimator)
  - Linear: OLS estimates are linear in y_i or error
  - Unbiased E($\hat{\beta}$) = β because E(u)=0 and E(X'u)=0
  - Best i.e. minimum variance

  - $\hat{\beta}$ is distributed with mean β and variance
  - Further assuming normality of error term, $\hat{\beta}$ is normally distributed

- For large N, OLS estimate is
  - Consistent
  - Asymptotically efficient
  - Asymptotically Normal (Central Limit Theorem)

Maximum Likelihood Estimation

- Maximum likelihood estimation
  - Y_i is random sample from population distribution f(y;θ)
  - Joint distribution of $Y_1, Y_2 \ldots Y_n$ is $f(y_1;\theta)*f(y_2;\theta)*f(y_3;\theta) \ldots f(y_n;\theta)$
  - Define Likelihood function $L(\theta; Y_1, Y_2 \ldots Y_n) = f(Y_1;\theta)*f(Y_2;\theta)*f(Y_3;\theta) \ldots f(Y_n;\theta)$
  - Maximum Likelihood Estimator (MLE) $\hat{\theta}$ maximizes L (or == Log L)
  - Econometric models use $f(Y_i \mid X_1 \ldots X_n;\theta_1 \ldots \theta_p)$

- Using normal distribution

$$L(\mu,\sigma^2;x_1,\ldots,x_n) = (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j-\mu)^2\right)$$

  - Econometric model $$= -\frac{n}{2}\log 2\pi - n\log s - \frac{1}{2s^2}\sum_{i=1}(y_i-(b_0+b_1x_i))^2$$
  - MLE equivalent to OLS

- Using other distributions
  - Bernoulli distribution (discrete random variable)
  - $X_1..X_n$ distributed Ber (p)

$$L(p|X_1,\cdots,X_n) = \prod_{i=1}^{n} P(X_i) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i}$$

Non Linear Models

- Limited dependent variable models
  - Endogenous variable is truncated e.g. wages are > 0
  - Endogenous variable takes on a limited number of values e.g. grade point average, number of children

- For example, random variable Y=1 or 0
  - P (y=1)= G(Xβ) where G is a function such that 0<G<1
  - Logit model if G is the logistic function: G(z) =exp(z)/ (1+exp(z))
  - Probit model if G is cdf of normal
  - Binary outcomes can arise from latent variable models: y* = Xβ +u; y=1 if y*> 0

- E (y/X) is non linear
  - E(y/X) = P(y=1| x) = G(Xβ)
  - OLS, WLS not applicable
  - MLE

Other Estimation Methods

- Method of Moments
  - Equate population and sample moments
  - $\theta = g(\mu)$ then $\hat{\theta} = g(\text{sample mean})$
- Non linear least squares
  - $Y = f(x; \theta)$; Minimize $\sum (Y - f(x; \theta))^2$
- Non-parametric estimation

Need estimation strategy with associated distribution theory

Statistical Inference

- Statistical Inference is distinct from interpretation of estimates
- Use sample estimates to draw inference about population parameters (distributions)

- Hypothesis testing
  - Testing significance of explanatory variables
  - Testing implications of models
  - Examples: $H_{0:}$ β=0; β=A; β <= 0;
  - T-statistic
    - t (α/2 , n-k) = $\hat{\beta}$-A/ std dev ($\hat{\beta}$)
  - Z-statistic
  - F-Stat for hypothesis involving more than one parameter: $H_0 : \beta_1 = \beta_2$;
  - Report results at 1%, **5%** and 10% significance level

- Confidence intervals (Interval estimates)
  - 100* (1- α/2) confidence interval
  - $\hat{\beta} \pm z_{\alpha/2}$ * std dev ($\hat{\beta}$)

Prediction

- Predictions are derived from estimates

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k.$$

- Prediction interval
  - Prediction error for an observation y^0

$$\hat{e}^0 = y^0 - \hat{y}^0 = (\beta_0 + \beta_1 x_1^0 + \ldots + \beta_k x_k^0) + u^0 - \hat{y}^0.$$

$$\mathrm{E}(\hat{e}^0) = 0.$$

  - Prediction error has mean 0 and standard deviation

$$\mathrm{se}(\hat{e}^0) = \{[\mathrm{se}(\hat{y}^0)]^2 + \hat{\sigma}^2\}^{1/2}.$$

- Goodness of fit measures

$$R^2 \equiv \mathrm{SSE}/\mathrm{SST} = 1 - \mathrm{SSR}/\mathrm{SST}.$$

$$\mathrm{SST} \equiv \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

$$\mathrm{SSE} \equiv \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2.$$

$$\mathrm{SSR} \equiv \sum_{i=1}^{n} \hat{u}_i^2.$$

Where SSE = explained sum of squares

SSR = residual sum of squares

SST= total sum of squares

Identification

- Model does not allow causal effect (or a particular parameter) to be identified
- Lack of identification because E (X'U) =/= 0
  - Simultaneous equation i.e. some X is also endogenous (Simultaneity bias)
    - P determines Q and Q determines P. Observed data is equilibrium P^* and Q^*
  - Unobserved variable correlated to observed (Omitted variable bias)
    - $\ln wage_i = \alpha + \beta * schooling_i + \gamma * experience_i + u_i$; Missing variable ability / motivation
    - $\ln wage_i = \alpha + \beta * schooling_i + \gamma * experience_i + \theta * Ability + u_i$. Cannot identify β from joint distribution of wages, schooling and experience alone . Need to know joint distribution of ability also.

- Relationship between estimated coefficients
  - $\ln wage_i = \bar{\alpha} + \bar{\beta} * schooling_i + u_i$
  - $\ln wage_i = \alpha + \beta * schooling_i + \gamma * experience_i + u_i$
  - $\bar{\beta} = \beta + \gamma * \lambda$ where λ is co-efficient from regressing experience on schooling

  - Selection bias
    - Sample selection based on Y
    - Self selection e.g. into a training program
  - Measurement error