Introduction to Econometrics

Hormoz Ramian
Imperial College London
Department of Finance

Wednesday 25th September, 2019

Overview

Hormoz Ramian

- Email: `h.ramian@imperial.ac.uk`
- Office Hours (Please email for appointments)

References for Today

- Probability, Statistics and Econometrics, Linton

Coursework for Today

- Handout includes some practice exercises
- These are not assessed but if you hand them in, you receive feedback

Overview

MRes/PhD Econometrics 1: Introduction

1. Linear Algebra
2. Econometrics Background Revision (Today)
3. Elements of Probability and Statistics, Modes of Convergence, etc. (Monday)

Motivation

- Study interaction between economic variables (correlation or causation)
- Statistical inference and economic significance

Motivating Example

Suppose you wish to study gender inequality

- average wage difference across genders
- test averages across two subsamples

Gender is a random assignment

- attribute difference to genders
- unclear at disaggregated level

Motivating Example

Measure its breakdown across

- age, tenure, experience
- hours worked, full-time or part-time
- occupational capacity (directors, professionals, non-professionals)
- sector (finance, manufacturing, health, education, PA)

Across time

- recession, expansion
- administrations, policies

Motivating Example

Suppose you observe $i = 1, \ldots, N$ persons:

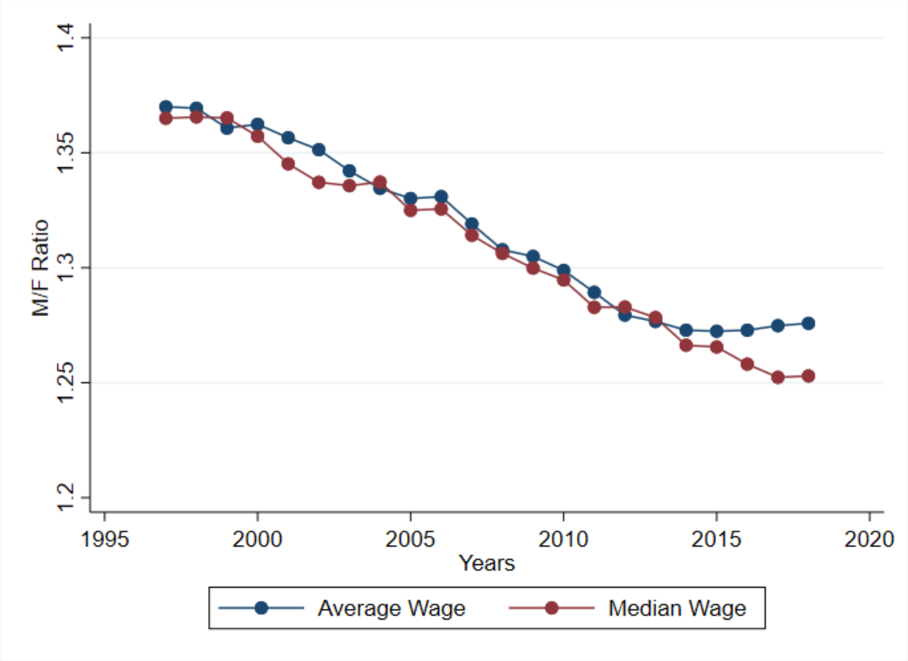
- wage y_i (\$)
- gender x_i (categorical)
- tenure (not experience) z_{1i}
- occupational capacity (ordinal $1, 2, \ldots, 9$) z_{2i}

$$\text{Wage}_i = f\left(x_i, z_{1i}, z_{2i}, \text{ other variables}\right) + \text{error}_i$$

Disaggregation matters

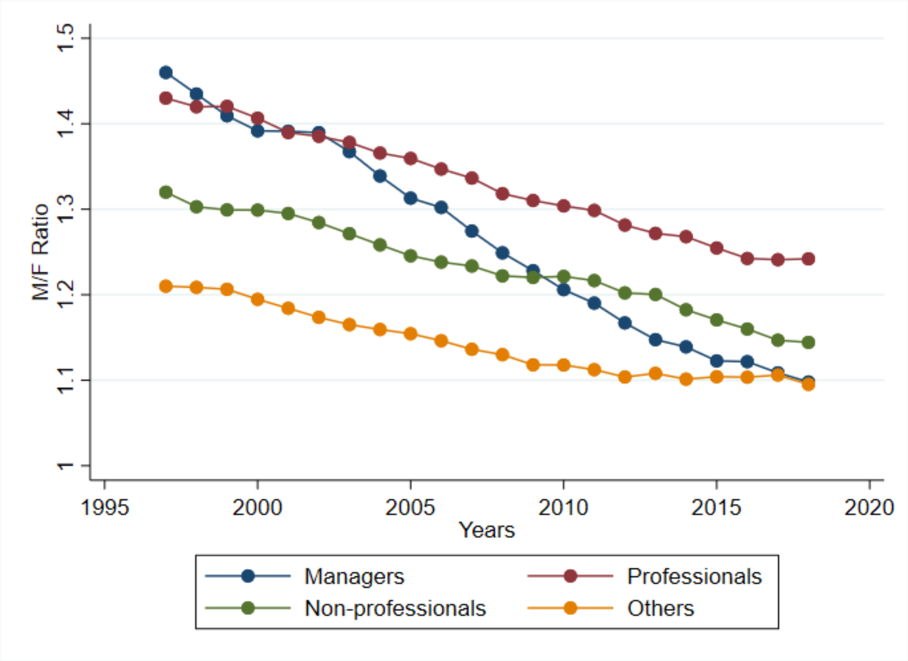
- how is each subgroup affected
- how does gender gap respond to each variable
- how can a policy help to close the gap

Descriptive Statistics



Descriptive Statistics

Breakdown



Assumptions

$$\begin{aligned}
y_i &= \beta_0 + \beta_1 x_i + u_i && (1)\\
u_i &\sim \mathcal{N}(0, \sigma^2) && (2)\\
\mathbb{E}[u_i|x_i] &= 0,\ \mathbb{E}[u_i x_i] = 0,\ \mathbb{E}[u_i] = 0 && (3)\\
\text{cov}[u_i, u_j] &= 0 && (4)
\end{aligned}$$

Assumptions

- regression equation itself, linearity, specification, etc.
- distributional assumption (normality)
- zero mean, exogeneity or zero conditional mean
- homoskedastic, zero cross-correlations
- full rank, collinearity

Importance varies:

- Some can be dropped with minimal problem
- Some can be problematic to drop but there are remedies to overcome the repercussions
- Some are critical

Variables

- Dependent (regressand, outcome): object of interest (effect)
- Independent (regressor, covariate): source but not always the cause
  - control: Similar to independent but not particularly interesting
  - but related to the dependent variable
  - need to remove (account for) their effects from the equation
  - confounder
- Error: Beyond control! particularly when working with non-experimental data

Contents

- Continuous (ordered set)
- Discrete
  - Categorical: order has no meaning (50 US states)
  - Ordinal: order matters (tax brackets)

Equation vs. Identity

The left hand side is the right hand side:

$$\begin{aligned}\text{weight}_t &= \text{weight}_{t-1} + (\text{intake}_t - \text{expenditure}_t)\\ \text{capital}_i &= \text{assets}_i - \text{liabilities}\end{aligned}$$

There's nothing to explain. If we know the RHS, we know the LHS (biologically or financially).

The left hand side co-moves (linearly or nonlinearly) with the right hand side:

$$\begin{aligned}\text{weight}_t &= \psi_0 + \psi_1(\text{intake}_t - \text{exercise hours}_t) + \text{error}_t\\ \text{borrowing rate}_i &= \theta_0 + \theta_1\text{leverage}_i + \text{error}_i\\ \text{income}_i &= \phi_0 + \phi_1\text{education}_i + \phi_2\text{experience}_i + \text{error}_i\\ \text{longevity}_i &= \beta_0 + \beta_1\text{bmi}_i + \beta_2\text{bmi}_i^2 + \text{error}_i\end{aligned}$$

RHS fundamentally matter to the LHS.

Population vs. Sample

Population:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + u_i \\ \boldsymbol{y} &= \boldsymbol{X\beta} + \boldsymbol{u} \end{aligned}$$

Sample:

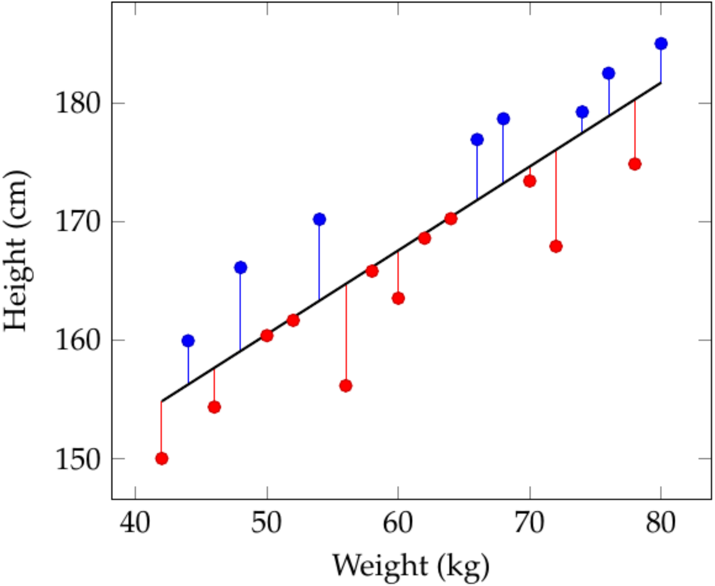
$$\begin{aligned} y_i &= \widehat{\beta}_0 + \widehat{\beta}_1 x_i + e_i \\ \boldsymbol{y} &= \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{e} \end{aligned}$$

Regression equation

- fitted value
- regression line
- conditional expectation

$$\begin{aligned} \widehat{y}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 x_i \\ \widehat{\boldsymbol{y}} &= \boldsymbol{X}\widehat{\boldsymbol{\beta}} \end{aligned}$$

Recall that $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$


Height (cm)
180
170
160
150
40
50
60
70
80
Weight (kg)

Multivariate

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + u_i$$

Each row: *collection of characteristics of the same observation*

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

$$\underset{n\times 1}{\boldsymbol{y}} = \begin{bmatrix} \underset{1\times k}{\boldsymbol{x}_1}' & \underset{1\times k}{\boldsymbol{x}_2}' & \ldots & \underset{1\times k}{\boldsymbol{x}_{k+1}}' \end{bmatrix}' \boldsymbol{\beta} + \underset{n\times 1}{\boldsymbol{u}}$$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$$

Each column: *collection of observations of the same characteristic*

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

$$\underset{n\times 1}{\boldsymbol{y}} = \begin{bmatrix} \underset{n\times 1}{\boldsymbol{1}} & \underset{n\times 1}{\boldsymbol{x}_1} & \underset{n\times 1}{\boldsymbol{x}_2} & \ldots & \underset{n\times 1}{\boldsymbol{x}_k} \end{bmatrix} \boldsymbol{\beta} + \underset{n\times 1}{\boldsymbol{u}}$$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$$

Properties I

Unbiasedness ($\mathbb{U}$)

$$\begin{aligned}
\mathbb{E}[\widehat{\boldsymbol{\beta}}|\boldsymbol{X}] &= \mathbb{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}] \\
&= \mathbb{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{u})|\boldsymbol{X}] \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbb{E}[\boldsymbol{u}|\boldsymbol{X}]
\end{aligned}$$

Variance

$$\begin{aligned}
\mathrm{var}[\widehat{\boldsymbol{\beta}}|\boldsymbol{X}] &= \mathrm{var}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{u})|\boldsymbol{X}] \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\mathrm{var}[\boldsymbol{u}|\boldsymbol{X}]\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}
\end{aligned}$$

if assume a spherical distribution $\mathrm{var}[\boldsymbol{u}|\boldsymbol{X}] = \sigma^2\boldsymbol{I}$ then,

$$\mathrm{var}[\widehat{\boldsymbol{\beta}}|\boldsymbol{X}] = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

Suppose $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1)'$, then standard errors:

$$\sqrt{\mathrm{diag}\left(\mathrm{var}[\widehat{\boldsymbol{\beta}}|\boldsymbol{X}]\right)} = \sigma \begin{bmatrix} \sqrt{\frac{\sum_i x_i^2}{\sum_i (x_i - \overline{x})^2}} \\ \sqrt{\frac{1}{\sum_i (x_i - \overline{x})^2}} \end{bmatrix}$$

Exercise: Show this.

Properties II

Definition (efficiency): The OLS estimators yield the minimum variance ($\mathbb{B}$) with the class of linear unbiased estimators.

- The variance of u is the variance of $\boldsymbol{Y}$ conditional on $\boldsymbol{X}$. Less variation of $\boldsymbol{Y}$ around the regression line yields greater precision.
- N is the number of observations. It shows up, implicitly, inside $\boldsymbol{X}'\boldsymbol{X}$, which for x_i drawn from some density has an expectation that increases linearly with N, thus the variance moves inversely proportional with N.
- $\boldsymbol{X}'\boldsymbol{X}$ is related to the covariance matrix of the vectors x_i', $i = 1, \ldots, N$. If each column of $\boldsymbol{X}$ is mean-zero, then $\boldsymbol{X}'\boldsymbol{X}$ is the covariance matrix of the K columns of $\boldsymbol{X}$. For this reason, if $\boldsymbol{X}$ has a lot of variance, then $\boldsymbol{X}'\boldsymbol{X}$ is bigger, so $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ is smaller, so variance of $\widehat{\boldsymbol{\beta}}$ is smaller and the estimate $\widehat{\boldsymbol{\beta}}$ is more precise.

Exercise: cov$[\widehat{\boldsymbol{\beta}}_0, \widehat{\boldsymbol{\beta}}_1]$=?

Linearity

EXERCISE: Specify which of the following specifications are linear ($\mathbb{L}$):

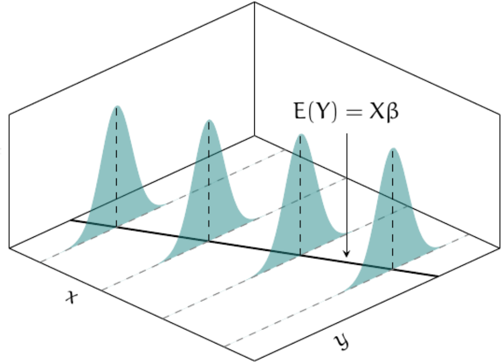
1. $y_i = \alpha + \beta x_i + \gamma x_i^2 + u_i$
2. $y_i = \alpha + \beta x_i + \gamma / x_i + u_i$
3. $y_i = \alpha + \beta x_i + x_i / \gamma + u_i$
4. $y_i = \alpha + \beta x_i + z_i / \gamma + u_i$
5. $y_i = \alpha + \beta x_i z_i + u_i$
6. $y_i = \alpha + \beta \gamma x_i + \gamma z_i + u_i$
7. $y_i = \alpha + \beta x_i + (1 - \beta) z_i + u_i$

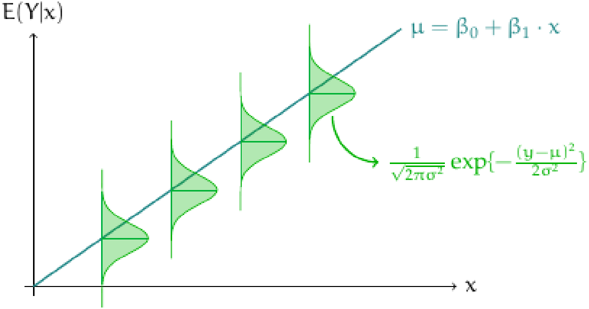
Conditional Expectation

Given a value of $\boldsymbol{X}$, the average value of the outcome variable is:

$$\begin{aligned} \mathbb{E}[y_i|\boldsymbol{X}] &= \widehat{\beta}_0 + \widehat{\beta}_1 x_i = \widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} \\ y_i|\boldsymbol{X} &\sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2) \end{aligned}$$

Statistically:




E(Y|x)
$\mu = \beta_0 + \beta_1 \cdot x$
$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(y-\mu)^2}{2\sigma^2}\}$
x

Gauss-Markov Theorem: OLS estimators are the best linear unbiased estimators ($\mathbb{BLUE}$), as they have the smallest variances among the class of all linear unbiased estimators.

1. Linearity (specification must be linear in its parameters): otherwise, ordinary least squares problem is inapplicable (why?).
2. Full column rank regressors: otherwise, $X'X$ is singular (non-invertibile) or *perfect* collinearity.
3. $\mathbb{E}[u] = 0$: otherwise, biased or inconsistent estimator(s).
4. $\mathbb{E}[u_i^2] = \sigma^2 \ \forall_i$: otherwise, heteroskedasticity ($\text{var}\,(u_i^2) = \sigma_i^2$ with $\sigma_i^2 \neq \sigma_j^2, \ \forall_{i \neq j}$) or cross-correlation/autocorrelation ($\text{cov}(u_i, u_j) \neq 0, \ \forall_{i \neq j}$), or both, and hence estimators are not efficient (still consistent and unbiased). This implies that statistical inference is no longer appropriate.
5. Non-stochastic regressors (exogeneity $\mathbb{E}[X_i u_i] = 0$): otherwise, endogeneity which causes biased estimators and reverse causality.

First Order Conditions

$$\begin{aligned} \min_{\beta} \qquad & \sum_{i=1}^{N} \left(y_i - \beta_1 - \sum_{j=2}^{K} \beta_j x_{ij} \right)^2 \\ & (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \\ & \boldsymbol{Y}'\boldsymbol{Y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{Y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \end{aligned}$$

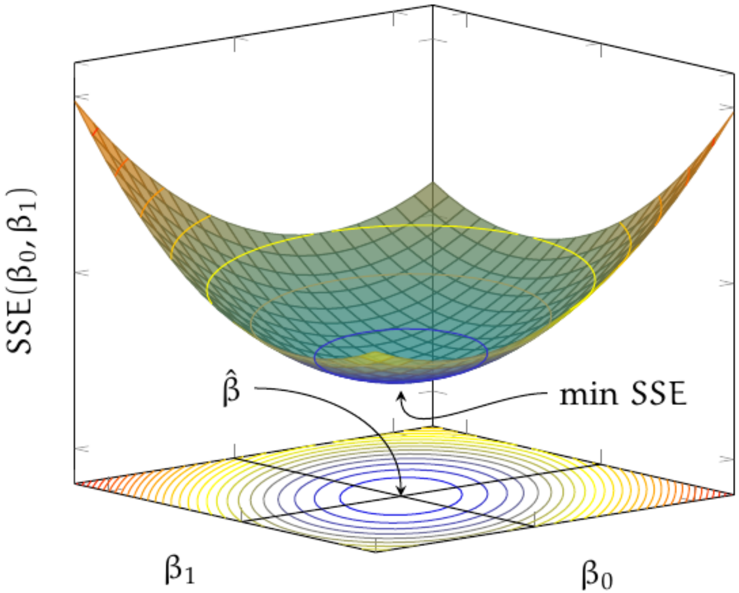
1×1 scalar value sum of squared residuals

- Note that $-\boldsymbol{Y}'\boldsymbol{X}\boldsymbol{\beta}$ and $-\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{Y}$ are equal scalars

Differentiating with respect to $\boldsymbol{\beta}$:

$$\frac{\partial(\boldsymbol{Y}' - \boldsymbol{\beta}'\boldsymbol{X}')'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}$$

$K \times 1$ (or $1 \times K$) vector


SSE(β_0, β_1)
$\hat{\beta}$
min SSE
β_1
β_0

Interpretation

Statistical significance as evidence in support of the (alternative) hypothesis

$$\begin{aligned} H_0 &: \beta_j = 0 \\ H_A &: \beta_j \neq 0 \end{aligned}$$

Interpretations

- units
- on average
- within confidence level
- keeping everything else constant

Measurement Units

- simple units (dollars, counts)
- log-units (log wage, log tenure)
- per cent (shares, equity to asset)
- percentage point (units within percentages e.g. interest rate)
- differences (median wage across gender, sector)

Consider the following four models:

$$\begin{aligned} &\text{Model 1} \qquad && y = \beta_0 + \beta_1 x + u && (5)\\ &\text{Model 2} && y = \beta_0 + \beta_1 \ln x + u && (6)\\ &\text{Model 3} && \ln y = \beta_0 + \beta_1 x + u && (7)\\ &\text{Model 4} && \ln y = \beta_0 + \beta_1 \ln x + u && (8) \end{aligned}$$

1. What are models 2–4 called in comparison to the first linear model?
2. Derive elasticity of $\mathbb{E}[Y|X]$ with respect to x.
3. Which model has a constant elasticity?
4. Interpret β_1 in each model

- Linear, linear-log, log-linear and log-log, respectively
- Elasticity of Y with respect to X:

$$\frac{d\mathbb{E}[Y|X]}{dX} \times \frac{X}{\mathbb{E}[Y|X]} \quad = \quad \frac{d\ln\mathbb{E}[Y|X]}{d\ln X}$$

$$\begin{aligned}
\text{Elasticity w.r.t. } x\text{: Model 1} &= \beta_1 x/(\beta_0 + \beta_1 x) \\
\text{Elasticity w.r.t. } x\text{: Model 2} &= \beta_1/(\beta_0 + \beta_1 \ln x) \\
\text{Elasticity w.r.t. } x\text{: Model 3} &= \beta_1 x \\
\text{Elasticity w.r.t. } x\text{: Model 4} &= \beta_1
\end{aligned}$$

Only the log-log transformation has a constant elasticity. The first three, elasticities depend on the value of X and are varying.

Parametric Specification

Specification

- economic theory (structural or reduced form)
- stylized facts (empirical evidence)
- heuristics
- other scientific foundations (e.g. BMI and life expectancy are quadratically related)

Features

- minimalistic approach (parsimonious)
- testable (falsifiable)
- interpretable
- suitable to dataset (e.g. curse of dimensionality)

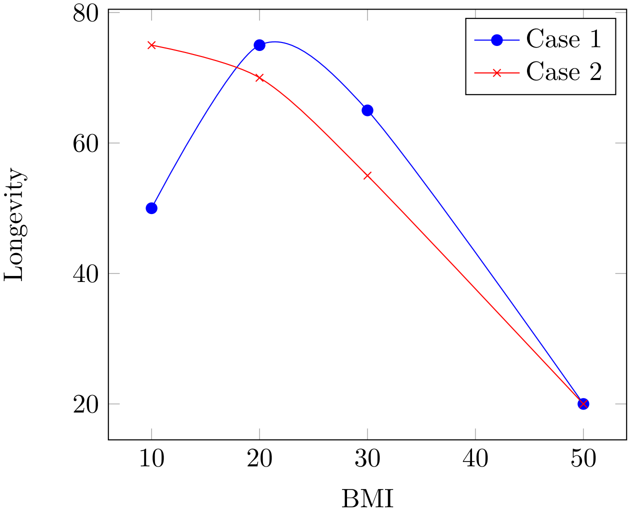
Specification itself is an assumption

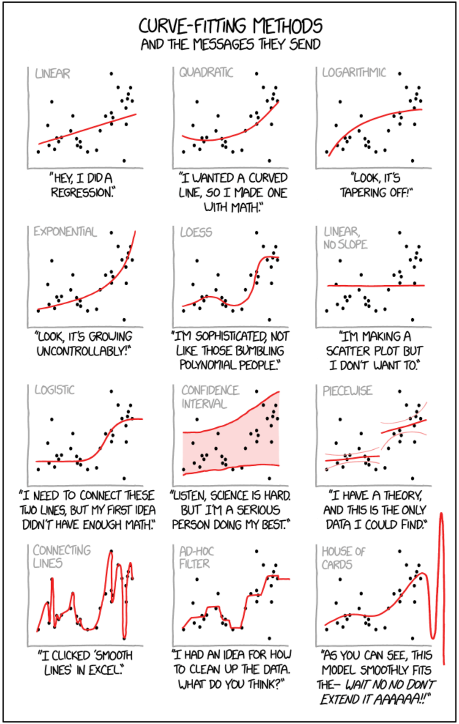
- testable (e.g. nonparametrics)
- harder to defend or invalidate

Life Expectancy and BMI

Which one of the cases below does present a more meaningful relationship between two variables?

$$\begin{aligned}\text{longevity}_i &= \beta_0 + \beta_1 \text{bmi}_i + \beta_2 \text{bmi}_i^2 + \text{error}_i \\ \text{longevity}_i &= \beta_0 + \beta_1 \text{bmi}_i + \text{error}_i\end{aligned}$$




CURVE-FITTING METHODS
AND THE MESSAGES THEY SEND
LINEAR
"HEY, I DID A REGRESSION."
QUADRATIC
"I WANTED A CURVED LINE, SO I MADE ONE WITH MATH."
LOGARITHMIC
"LOOK, IT'S TAPERING OFF!"
EXPONENTIAL
"LOOK, IT'S GROWING UNCONTROLLABLY!"
LOESS
"I'M SOPHISTICATED, NOT LIKE THOSE BUMBLING POLYNOMIAL PEOPLE."
LINEAR, NO SLOPE
"I'M MAKING A SCATTER PLOT BUT I DON'T WANT TO."
LOGISTIC
"I NEED TO CONNECT THESE TWO LINES, BUT MY FIRST IDEA DIDN'T HAVE ENOUGH MATH."
CONFIDENCE INTERVAL
"LISTEN, SCIENCE IS HARD. BUT I'M A SERIOUS PERSON DOING MY BEST."
PIECEWISE
"I HAVE A THEORY, AND THIS IS THE ONLY DATA I COULD FIND."
CONNECTING LINES
"I CLICKED 'SMOOTH LINES' IN EXCEL."
AD-HOC FILTER
"I HAD AN IDEA FOR HOW TO CLEAN UP THE DATA. WHAT DO YOU THINK?"
HOUSE OF CARDS
"AS YOU CAN SEE, THIS MODEL SMOOTHLY FITS THE– WAIT NO NO DON'T EXTEND IT AAAAAA!!"

not easy to interpret!

Identification

Suppose asking to hand in optional coursework and the following,

$$y = \beta d + \epsilon \qquad d = \begin{cases} 1 & \text{handed in} \\ 0 & \text{otherwise} \end{cases}$$

at the beginning of a term (bad identification):

$$\begin{aligned} H_0 &: \text{too much econometrics before or was busy} \\ H_A &: \text{committed to eonometrics!} \end{aligned}$$

during the break period (better identification):

$$\begin{aligned} H_0 &: \text{too much econometrics before or was busy} \\ H_A &: \text{committed to eonometrics!} \end{aligned}$$

Good identification should be able to isolate the variation

- single out cause and effect
- variations in outcome variable should be ideally only attributable to the independent variable

NB Turns out 'identification' has two meanings within the same subject!

Identification: Parameter identification problem is the inability to estimate each individual parameter in regression model separately. When parameters are only jointly obtained, then a regression specification amounts to unidentified parameters.

Correlation vs. Causation

Regression

- establishes a correlation or joint distribution between variables
- relationship can be refined to establish a causal relationship

Causality Methodology

1. random assignment or natural experiment
2. instrumental variables
3. regression discontinuity
4. DiD

Random Assignment:

- observations inside a sample are observed partly because of a pattern that is not measurable or not observable
- sample is not representative of the true population

Self-selection bias in marketing

Case Study

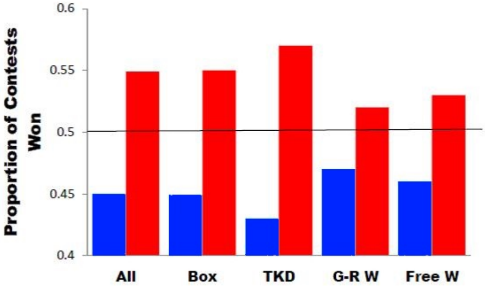
Customer Satisfaction Surveys (CSS's) are a useful tool for identifying strengths and weaknesses in a commercial organisation providing goods or services. However, these types of surveys suffer from a self-selection bias that can negatively influence the quality and reliability of the results and can lead to an inaccurate evaluation of customer satisfaction. These errors occur because whether or not a survey is completed is not an entirely random event and the people who choose to respond to a survey may be systematically different from those who do not respond.

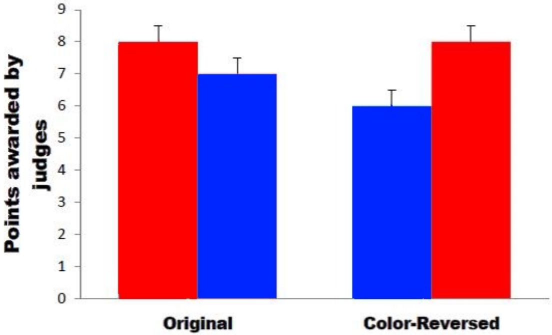
When a survey is conducted by calling a random sample of publicly available telephone numbers during a weekday, it will not include the responses of people with unlisted telephone numbers or those people who are unable to answer the phone because they are at work. Instead, it is likely to include a disproportionate number of respondents who have traditional land-line telephone services, and who are at home during normal working hours (for example it is likely that the elderly subsample is over-represented). In terms of regression, this implies that our estimates are downward or upward biased which can subsequently cause unreliable inference.

Correlation vs. Causation

In the 2004 Athens Olympics, contestants in four combat sports (Boxing, Taekwondo, Greco-Roman wrestling, and Freestyle wrestling) were randomly assigned red or blue outfits:

- Colour has pervasive effects on human psychology, judges' perceptions, and physical performance.
- Red Enhances Human Performance in Contests
- When all factors are equal, red tips the balance





Statistical vs. Economic Significance

Statistical Significance:

- rejection of a null in favour of alternative
- deviation from a given value beyond confidence level

Economic Significance:

- economically meaningful relationship
- supported by data but not necessarily statistically significant
- e.g. increasing a certain tax changes GDP growth by 10^{-18} dollars

Important to have both (optional reading, Karlan and Zinman (Econometrica, 2009))

EMPIRICAL TESTS OF HIDDEN INFORMATION AND HIDDEN ACTION: FULL SAMPLE

	OLS							
Dependent Variable:	*Monthly Average Proportion Past Due*		*Proportion of Months in Arrears*		*Account in Collection Status*		*Standardized Index of Three Default Measures*	
Mean of Dependent Variable:	0.09	0.09	0.22	0.22	0.12	0.12	0	0
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Contract rate (Hidden Action Effect 1)	0.005	0.002	0.006*	0.002	0.001	−0.001	0.014	0.004
	(0.003)	(0.004)	(0.003)	(0.004)	(0.005)	(0.005)	(0.011)	(0.013)
Dynamic repayment incentive dummy (Hidden Action Effect 2)	−0.019*	−0.000	−0.028**	0.004	−0.025**	−0.004	−0.080**	−0.000
	(0.010)	(0.017)	(0.011)	(0.021)	(0.012)	(0.020)	(0.032)	(0.057)
Dynamic repayment incentive size		−0.005		−0.009**		−0.006		−0.023*
		(0.004)		(0.004)		(0.005)		(0.013)
Offer rate (Hidden Information Effect)	0.005	0.004	0.002	0.002	0.007	0.007	0.015	0.015
	(0.003)	(0.003)	(0.003)	(0.004)	(0.005)	(0.005)	(0.011)	(0.012)
Observations	4348	4348	4348	4348	4348	4348	4348	4348
Adjusted R-squared	0.08	0.08	0.14	0.15	0.06	0.06	0.10	0.11
Probability(both dynamic incentive variables = 0)		0.06		0.00		0.06		0.01
Probability(all 3 or 4 interest rate variables = 0)	0.0004	0.0005	0.0003	0.0012	0.0006	0.0016	0.0000	0.0001

*significant at 10%; **significant at 5%; ***significant at 1%. Each column presents results from a single OLS model with the RHS variables shown and controls for the randomization condtions: observable risk, month of offer letter, and branch. Adding loan size and maturity as additional controls does not change the results. Robust standard errors in parentheses are corrected for clustering at the branch level. "Offer rate" and "Contract rate" are in monthly percentage point units (7.00% interest per month is coded as 7.00). "Dynamic repayment incentive" is an indicator variable equal to one if the contract interest rate is valid for one year (rather than just one loan) before reverting back to the normal (higher) interest rates. "Dynamic repayment incentive size" interacts the above indicator variable with the difference between the lender's normal rate for that individual's risk category and the experimentally assigned contract interest rate. A positive coefficient on the Offer Rate variable indicates hidden information, a positive coefficient on the Contract Rate or Dynamic Repayment Incentive variables indicates hidden action (moral hazard).

The dependent variable in columns (7) and (8) is a summary index of the three dependent variables used in columns (1)–(6). The summary index is the mean of the standardized value for each of the three measures of default.

Further Remarks

Discuss the advantages and dis-advantages of the following approaches:

- **Structural Form vs. Reduced Form:** An structural form (theoretical model) is the result of a normative framework, whereas a reduced form (statistical model) is the result of a positive framework.
- **Frequentist vs Bayesian:** Suffices to know we're in the frequentist world when the true value of each parameter is a fixed (unknown) value. Whereas in the Bayesian world, the true parameters are themselves distributions. The true value of inflation at each point in time is a fixed (unknown) number, or itself is a range (probabilistic) of numbers?
- **Parametric, semi-parametric and non-parametric:** A parametric form uses coefficients to summarize the data, where as a non-parametric form relies only on an empirical distribution of the data.
- **Misspecification:** The regression equation is inconsistent with the true relationship, e.g. an important explanatory variable is missing (omitted variable), or the functional form is unfitting, etc.

Practice Problems

Consider the categorical variables $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$ each a vector of size $n \times 1$, and that $\boldsymbol{d}_2 = \boldsymbol{\iota}_n - \boldsymbol{d}_1$ with $n = n_1 + n_2$ (n_1: number of men and n_2: number of women) such that:

$$d_{1,i} = \begin{cases} 1 & \text{if man} \\ 0 & \text{if woman} \end{cases}$$

for $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$, suppose:

$$\boldsymbol{y} = \boldsymbol{d}_1 \widehat{\gamma}_1 + \boldsymbol{d}_2 \widehat{\gamma}_2 + \widehat{\boldsymbol{u}}$$

1. Show that $(\widehat{\gamma}_1,\ \widehat{\gamma}_2)' = (\overline{y}_1,\ \overline{y}_2)'$.
2. Compare $\widehat{\gamma}_1$ and $\widetilde{\gamma}_1$ from two OLS regressions:

$$\begin{aligned} \widehat{\boldsymbol{y}} &= \boldsymbol{d}_1 \widehat{\gamma}_1 + \boldsymbol{d}_2 \widehat{\gamma}_2 && (9) \\ \widehat{\boldsymbol{y}} &= \boldsymbol{d}_1 \widetilde{\gamma}_1 && (10) \end{aligned}$$