L1: Introduction to Econometrics

Feng Li
feng.li@cufe.edu.cn

School of Statistics and Mathematics
Central University of Finance and Economics

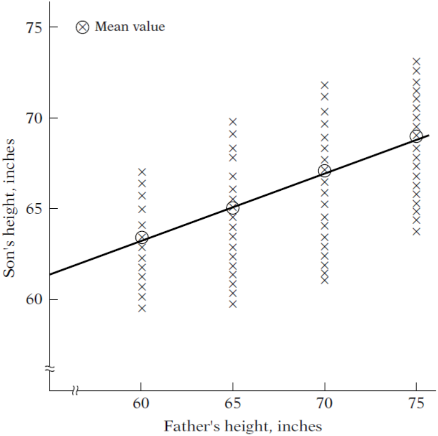
Revision: September 2, 2014

Today we are going to learn...

1. **What is Econometrics?**
2. **Two-variable regression**
   - The population regression
   - The sample regression

What is Econometrics?
↪ An Introduction

- Econometrics means: economic measurement.
- Regression means: the study of the **dependence** of one variable (the dependent variable), on one or more other variables (the explanatory variables), with estimating and/or predicting the population. See this example:



What is Econometrics?

↬ Regression,deterministic relationships, and causation

- Regression is not deterministic.
- deterministic: If I know how old you are. I known exactly when you were born.
- Regression relationships deal with random or stochastic. **Very important**!
- Regression deals with dependence but can never establish causal connection.

What is Econometrics?

↪ Terminology and notations

Dependent variable	Explanatory variable
⇕	⇕
Explained variable	Independent variable
⇕	⇕
Predictand	Predictor
⇕	⇕
Regressand	**Regressor**
⇕	⇕
Response	Stimulus
⇕	⇕
Endogenous	Exogenous
⇕	⇕
Outcome	Covariate
⇕	⇕
Controlled variable	Control variable

What is Econometrics?
↛ Types of data

- Time series data: China's GDP for the last 60 years.
- Cross-section data: The survey of diabetes study for 10 people in past ten years.
- Pooled data: Pooled or combined data are elements of both time series and cross-section data. Consumption Price Index for OECD countries from 1980–2000.
- **Question**: How accurate of those data?

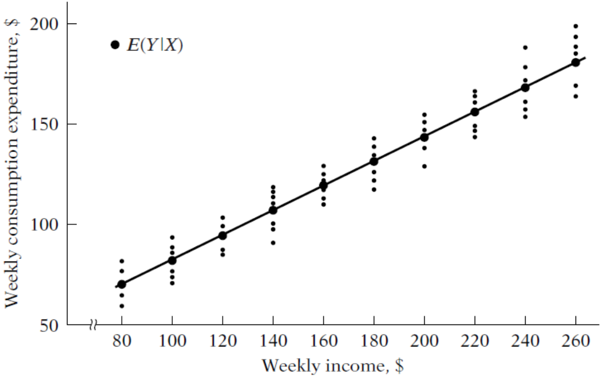
Population regression line
↪ The data

WEEKLY FAMILY INCOME X, \$

Y ↓ \ X →	80	100	120	140	160	180	200	220	240	260
Weekly family consumption expenditure Y, \$	55	65	79	80	102	110	120	135	137	150
	60	70	84	93	107	115	136	137	145	152
	65	74	90	95	110	120	140	140	155	175
	70	80	94	103	116	130	144	152	165	178
	75	85	98	108	118	135	145	157	175	180
	–	88	–	113	125	140	–	160	189	185
	–	–	–	115	–	–	–	162	–	191
Total	325	462	445	707	678	750	685	1043	966	1211
Conditional means of Y, $E(Y\|X)$	65	77	89	101	113	125	137	149	161	173

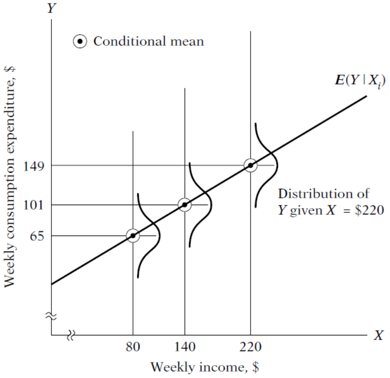
- The data refer to a total **population** of 60 families.
- The mean of consumption expenditure depends on the income. So we call the mean as conditional mean of Y, $\mathrm{E}(Y|X)$.
- The unconditional mean is written as $\mathrm{E}(Y)$.

Population regression line
↪ The regression line



- Scatter Plot the data and mart the conditional mean $\mathrm{E}(Y|X)$
- Join all the conditional mean we obtain the **Population regression line(curve)**.
- Simply we call it **regression of** Y **and** X.
- Again the **population** means we use all the information of 60 families.

Population regression line
↪ The first regression model



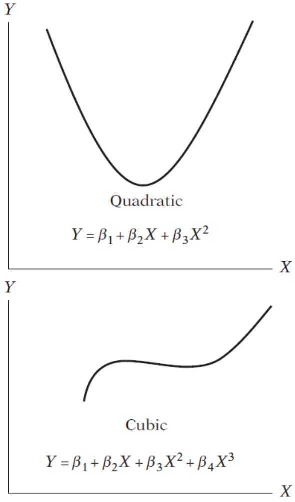
- The conditional mean is a function of income X_i, we write it as $E(Y|X_i) = f(X_i)$
- Our example can explained as $E(Y|X_i) = \beta_1 + \beta_2 X_i$, where β_1 is called **intercept**, and β_2 is called **slope coefficients**.
- Then we have some sort of **distribution** of expenditure given income.

Population regression line

↬ The term of linear

- Linearity in the variables: Our previous example, Y is linear expression of X, but $E(Y|X) = \beta_1 + \beta_2 X^2$ is not.
- Linearity in the parameters.
  - The conditional expectation of Y ($E(Y|X)$) is a linear function of the parameters.
  - What about $E(Y|X) = \beta_1 + \beta_2 X^2$ this time?
  - How about $E(Y|X) = \beta_1 + \beta_2^2 X$?

Population regression line
↛ The term of linear – more examples



Sample regression line

↪ Still the family income example

A RANDOM SAMPLE FROM THE POPULATION OF TABLE 2.1

Y	X
70	80
65	100
90	120
95	140
110	160
115	180
120	200
140	220
155	240
150	260

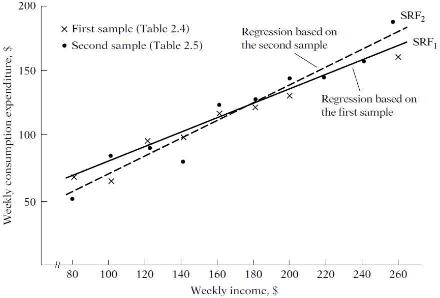
ANOTHER RANDOM SAMPLE FROM THE POPULATION OF TABLE 2.1

Y	X
55	80
88	100
90	120
80	140
118	160
120	180
145	200
135	220
145	240
175	260

- It is not common we can have the whole population information to use.
- Instead we may only have some **samples** from the population.

Sample regression line

↪ Still the family income example

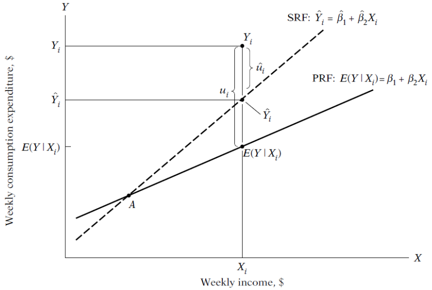


- We do the preceding procedures to obtain the regressions line as we did in the population example anyway.
- You may find each regression line differs from others with different sample we obtained.
- We have the **sample regression function** as $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$
    - $\hat{Y}_i$: estimator of $E(Y|X_i)$
    - $\hat{\beta}_1$: estimator of β_1
    - $\hat{\beta}_2$: estimator of β_2

Sample regression line

↪ Still the family income example

- Sample regression is more common in the regression analysis.
- Since we don't have the whole population information, our sample regression line can be different from the population regression line.
- We use $\hat{u}_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_i)$ to measure the difference.
- That means we can write $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$ **Why**?
- $\hat{u}$ is called the residual. Smaller $\hat{u}_i$ means our sample regression line is closer to the population regression line.
- **Question:** can we have $\hat{\beta}_1$ and $\hat{\beta}_2$ so that $\hat{u}$ to be a minimal?



Take home questions

- What are the relationships among the concepts: **random**, **deterministic**, and **disturbance term** ?
- There are some variables of interest in the one variable regression model, such as Y, X, β, $\hat{Y}$, $\hat{\beta}$, u and $\hat{u}$. What variables do you think are random?
- What are the reasons of appearing disturbance term u in the regression analysis?
- In the big data world, what are population and sample?