References

Introduction to Econometrics
By James Stock and Mark Watson

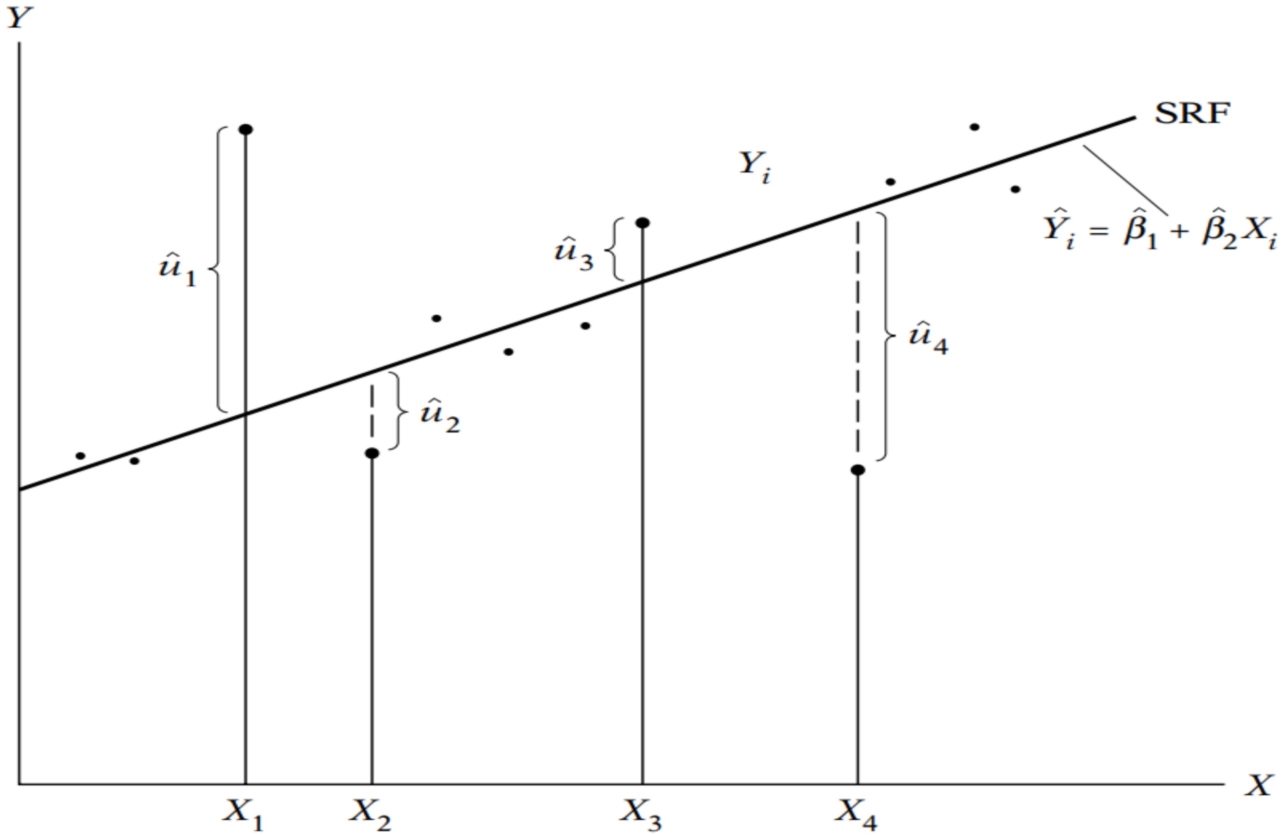
Basic Econometrics
By Damodar Gujarati

What are we looking at??

- We would be dealing with errors and variation.
- Objective is to minimize variation in errors and maximize variation in explained variables.
- Ideally one should look at minimum variation for all the moments condition.

Linear Regression

$$Y = \beta_1 + \beta_2 X + u$$



METHODOLOGY

- BETAs: Slope Linear
- X explanatory variables
- X can be both linear as well as non-linear
- Error is not related to another error.
- Heteroscedasticity and Autocorrelation

Multicollinarity

Errors and X-Variables

Errors are Y less Y-hat

X variables are non-stochastic

X variables can be linear or non-linear (depending upon model)

Beta are linear

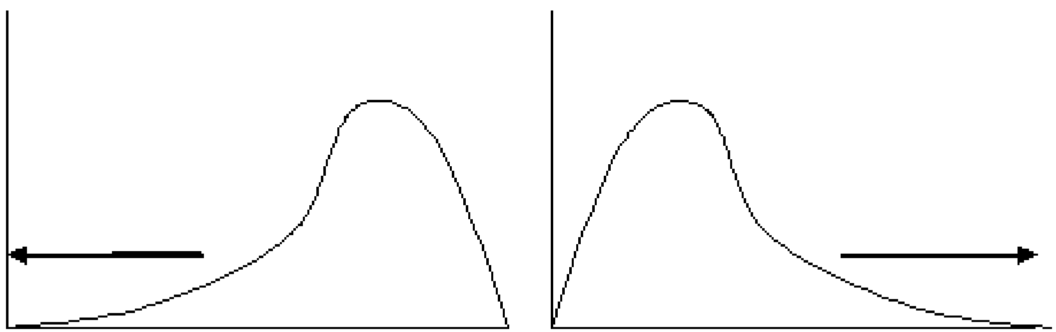
OLS assumes errors are normally distributed

Non-Parametric estimation does not assume any specific distribution
Function about the error terms

Generalized Methods of Moment is about minimizing all moments condition simultaneously.

Skewness

- For positive skewness: mean-income>median-income>mode-income
- Skewness of a normal density is zero
- Skewness is the third moment
- Negative skewness coefficient means it is a negative skewed distribution. Generally measure of skewness is 3(mean-mode)/sd.
- We are measuring income on the horizontal and probability of the population with that income on the vertical axis. When we have more people with less income we are saying the peak of the density is right/positive skewed.



Negative Skew
Elongated tail at the **left**
More data in the left tail than would be expected in a normal distribution

Positive Skew
Elongated tail at the **right**
More data in the right tail than would be expected in a normal distribution

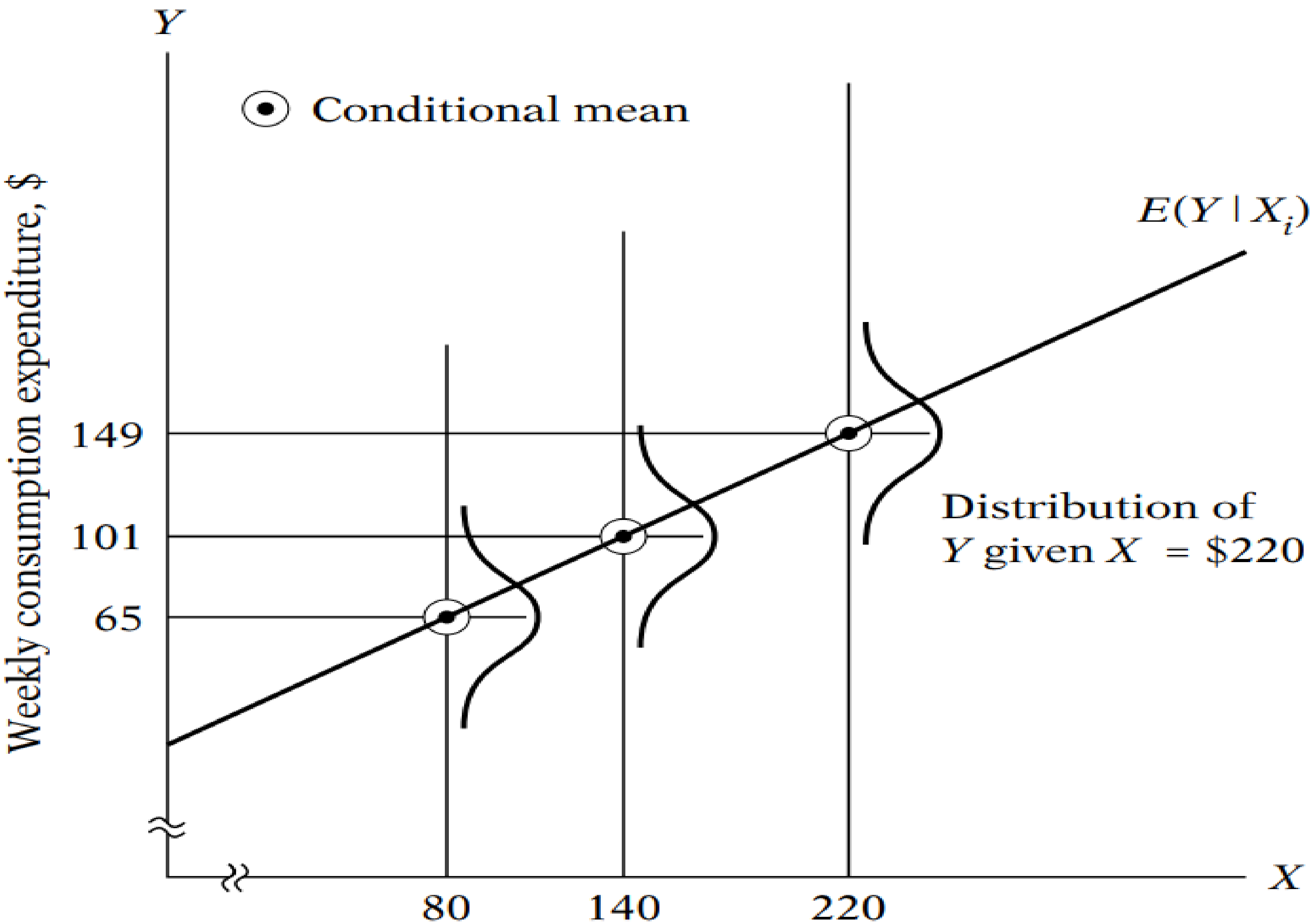
Kurtosis

- Fourth moment condition.
- Smaller variance means sharper, pointed, density function – also known as leptokurtic distribution.
- Leptokurtic has values greater than 3 and Platykurtic has values less than 3.
- Normal density has values equal to 3.

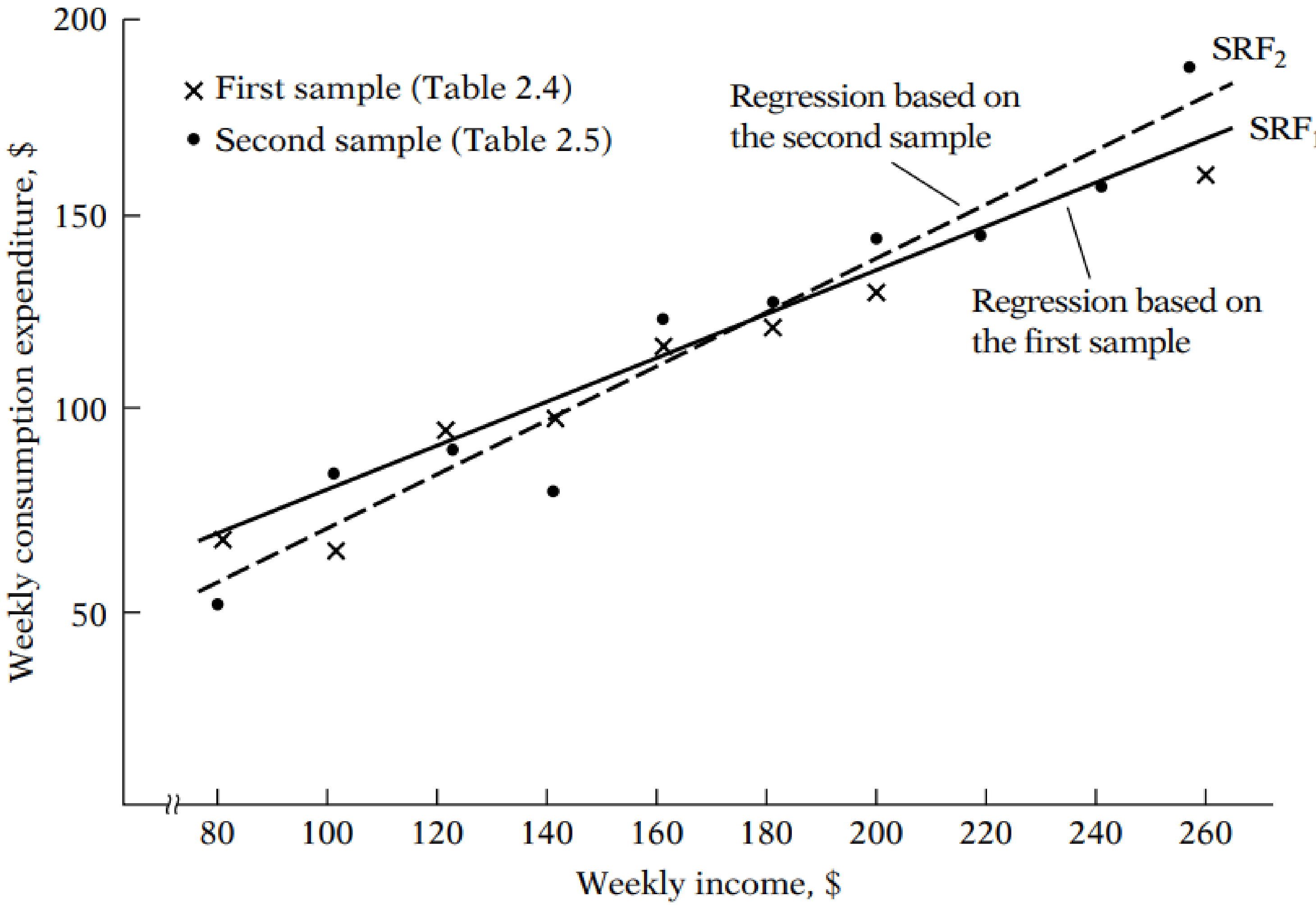
Weekly family Income X, $

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 Mean of Y, E(Y/X)	65	77	89	101	113	125	137	149	161	173

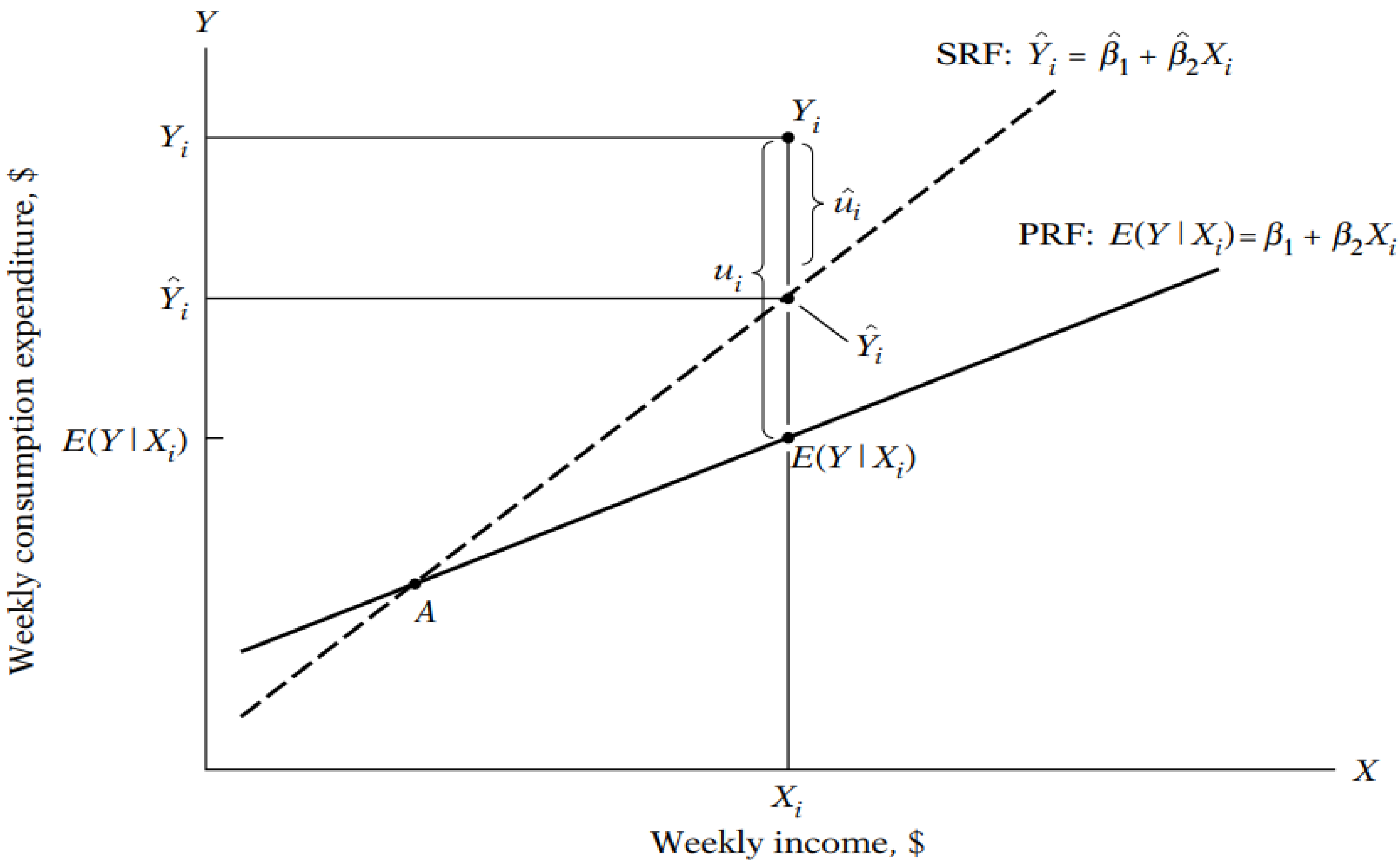
Conditional Mean



Sample Regression Function



ESS, RSS, and TSS



Important Formulas

$$\sum \hat{u}_i^2 = \sum (Y_i - \hat{Y}_i)^2$$
$$= \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$$

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

$$\hat{\beta}_2 = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - \left(\sum X_i\right)^2}$$
$$= \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$
$$= \frac{\sum x_i y_i}{\sum x_i^2}$$

Results (Cont.)

$$
\begin{aligned}
\textbf{cov}\,(u_i, X_i) &= E[u_i - E(u_i)][X_i - E(X_i)] \\
&= E[u_i(X_i - E(X_i))] \qquad \text{since } E(u_i) = 0 \\
&= E(u_iX_i) - E(X_i)E(u_i) \qquad \text{since } E(X_i) \text{ is nonstochastic} \\
&= E(u_iX_i) \qquad \text{since } E(u_i) = 0 \\
&= 0 \qquad \text{by assumption}
\end{aligned}
$$

Formula for one tail or two tail tests

$$\mathbf{t} = \frac{\bar{X} - \mu}{s / \sqrt{n}}$$

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n - 2}$$

Results (Cont.)

Mean: $E(u_i) = 0$

Variance: $E[u_i - E(u_i)]^2 = E(u_i^2) = \sigma^2$

cov (u_i, u_j): $E\{[(u_i - E(u_i)][u_j - E(u_j)]\} = E(u_i u_j) = 0 \quad i \neq j$

The assumptions given above can be more compactly stated as

$$u_i \sim N(0, \sigma^2)$$

$$t = \frac{\hat{\beta}_2 - \beta_2}{\text{se}(\hat{\beta}_2)} = \frac{\text{estimator} - \text{parameter}}{\text{estimated standard error of estimator}}$$

$$= \frac{(\hat{\beta}_2 - \beta_2)\sqrt{\sum x_i^2}}{\hat{\sigma}}$$

Some Intuition and Observations

- F test: *Y Hat* is significantly better than *Y Bar.*
- Regressing Y on *Y Bar* yields $R^2 = 0$.
- F test: Also on model components (whether or not add any additional variables, as in Xs).
- EXCEL report upper and lower confidence interval (CI) based on one-tail critical *t-statistics.*

Results (Cont.)

- For calculating CI, you need information about coefficients (β), standard errors of (β) and Critical *t.*

- Changing the scale, for instance, reporting 1 thousand as oppose to 1000, will increase the coefficient 1000 times.

- A *P* value of 0.05 means that null is correct only 5 times out of 100 times. Hence failing to accept the *Null*. *P* value is the probability of *Null* being correct.

Adjusted R^2 and F test

Adjusted R^2 = $$1-\frac{RSS/(n-k)}{TSS/(n-1)}$$

F test on model component =

$$F=\frac{(R_F^2-R_R^2)/(No.\ of\ Restrictions)}{(1-R_F^2)/(No.\ of\ observations-No.\ of\ Coefficients_F)}$$

Confidence Intervals:

$$\Pr\left[-t_{\alpha/2}\leq\frac{\hat{\beta}_2-\beta_2}{\mathrm{se}\,(\hat{\beta}_2)}\leq t_{\alpha/2}\right]=1-\alpha$$

Takeaways from Multiple Regression

1. Explain the Anova
2. Explain Mean Squares, ESS, RSS, TSS
3. Derivation of F both for model components, and Y-bar vis-à-vis Y-hat
4. How to calculate t-statistics
5. Check for endogeneity (that is relationship between X and Errors)
6. Corollary, covariance between Y-hat with Error
7. Prediction/Projection/Forecast