Review EE 325

- Review the material that we covered before the midterm exam on your own ☺
- You are allowed to bring the following items for final exam:
  - 1 A4 paper (can write anything both sided)
  - 1 Calculator (financial or non-financial calculator)
  - Blue or Black ink pens

Functional Forms of Regression Models

- The log-linear model
- Semilog models
- Reciprocal models
- The logarithmic reciprocal model

The log-linear model

The exponential regression model:

$$Y_i = \beta_1 X_i^{\beta_2} e^{u_i}$$

Which may be expressed alternatively as

$$\ln Y_i = \ln \beta_1 + \beta_2 \ln X_i + u_i$$

$$\ln Y_i = \alpha_1 + \beta_2 \ln X_i + u_i$$

$$where\ \alpha = \ln \beta_1$$

The log-linear model

- The slope coefficient β_2 measures the elasticity of Y with respect to X, that is, the percentage change in Y for a given percentage change in X
- Example Price elasticity of demand

Semilog Models

- Log-Lin model

$$\ln Y_i = \beta_1 + \beta_2 X_i + u_i$$

- Lin-Log model

$$Y_i = \beta_1 + \beta_2 \ln X_i + u_i$$

The Log-Lin model

$$\ln Y_i = \beta_1 + \beta_2 X_i + u_i$$

$$\beta_2 = \frac{relative\ change\ in\ regressand}{absolute\ change\ in\ regressor}$$

β_2 *is known as the semielasticity of Y with respect to X*

If we multiply the relative change in Y by 100, it will give the percentage change in Y for an absolute change in X

$100 * \beta_2$ *is known as the semielasticity of Y with respect to X*

Lin-Log models

The absolute change in Y for a percentage change in X

$$Y_i = \beta_1 + \beta_2 \ln X_i + \mu_i$$

$$\beta_2 = \frac{Change\ in\ Y}{relative\ change\ in\ X}$$

$$= \frac{\Delta Y}{\Delta X / X}$$

The absolute change in Y for a percentage change in X

If $\Delta X / X$ changes by 0.01 unit or 1%, the absolute change in Y is

$$0.01 * \beta_2$$

Reciprocal models

$$Y_i = \beta_1 + \beta_2 \left(\frac{1}{X_i} \right) + u_i$$

As X increases indefinitely, in term $\beta_2 \left(\frac{1}{X_i} \right)$ approaches zero and Y approaches the limiting or asymptotic value β_1

The slope

$$\frac{dY}{dX} = -\beta_2 \left(\frac{1}{X^2} \right)$$

if β_2 is positive, the slope is negative throughout

β_2 is negative, the slope is positive throughout

Choice of Functional Form

TABLE 6.6

Model	Equation	Slope $\left(=\frac{dY}{dX}\right)$	Elasticity $\left(=\frac{dY}{dX}\frac{X}{Y}\right)$
Linear	$Y=\beta_1+\beta_2 X$	β_2	$\beta_2\left(\frac{X}{Y}\right)^*$
Log–linear	$\ln Y=\beta_1+\beta_2 \ln X$	$\beta_2\left(\frac{Y}{X}\right)$	β_2
Log–lin	$\ln Y=\beta_1+\beta_2 X$	$\beta_2 (Y)$	$\beta_2 (X)^*$
Lin–log	$Y=\beta_1+\beta_2 \ln X$	$\beta_2\left(\frac{1}{X}\right)$	$\beta_2\left(\frac{1}{Y}\right)^*$
Reciprocal	$Y=\beta_1+\beta_2\left(\frac{1}{X}\right)$	$-\beta_2\left(\frac{1}{X^2}\right)$	$-\beta_2\left(\frac{1}{XY}\right)^*$
Log reciprocal	$\ln Y=\beta_1-\beta_2\left(\frac{1}{X}\right)$	$\beta_2\left(\frac{Y}{X^2}\right)$	$\beta_2\left(\frac{1}{X}\right)^*$

Note: * indicates that the elasticity is variable, depending on the value taken by X or Y or both. When no X and Y values are specified, in practice, very often these elasticities are measured at the mean values of these variables, namely, $\bar{X}$ and $\bar{Y}$.

Assumptions

- Linear regression model, or linear in the parameters
- Fixed X values or X values independent of the error term. We require zero covariance between u_i and each X variables

$$\operatorname{cov}(u_i, X_{2i}) = \operatorname{cov}(u_i, X_{3i}) = 0$$

- Zero mean value of disturbance u_i

$$E(u_i \mid X_{2i}, X_{3i}) = 0, \; for \; each \; i$$

- Homoscedasticity or constant variance of u_i

$$\operatorname{var}(u_i) = \sigma^2$$

- No autocorrelation, or serial correlation, between the disturbances

$$\operatorname{cov}(u_i, u_j) = 0, \quad i \neq j$$

- The numbers of observations n must be greater than the number of parameters to be estimated
- There must be variation in the values of the X variables
- No exact collinearity between the X variables
- There is no specification bias

The Three-Variable Model

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

The coefficients β_2 and β_3 are called the **partial regression coefficients**

Interpretation of Multiple Regression Equation

$$E(Y_i \mid X_{2i}, X_{3i}) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i}$$

The conditional mean or expected value of Y conditional upon the given or fixed values of $X_2 \; and \; X_3$

The Meaning of Partial Regression Coefficients

The coefficients β_2 and β_3 are called the **partial regression coefficients**

The meaning of partial regression coefficient is as follows

β_2 measures the change in the mean value of, Y, E(Y), per unit change in X_2, holding the value of X_3 constant.

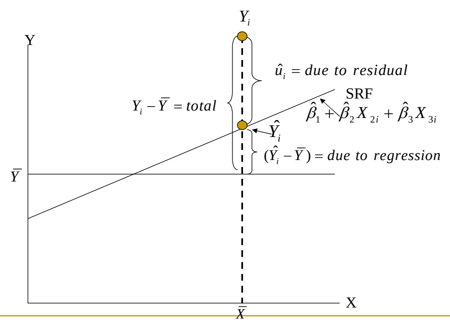
β_3 measures the change in the mean value of, Y, E(Y), per unit change in X_3, holding the value of X_2 constant.

The Multiple Coefficient of Determinant (R^2)

- Measure the goodness of fit of the regression equation
- We would like to know the proportion of the variation of Y explained by the variables jointly.
- The quantity that gives this information is known as the **multiple coefficient of determinant**

The Multiple Coefficient of Correlation (R)

- Measure of the degree of association between Y and all the explanatory variables jointly



$$TSS = ESS + RSS$$

$$\sum y_i^2 = \sum (Y_i - \overline{Y})^2$$ **Total variation of the actual Y values about their sample mean**

(Total Sum of Squares, TSS)

$$\sum \hat{y}_i^2 = \sum (\hat{Y}_i - \hat{\overline{Y}})^2$$

Variation of the estimated Y values about their mean (Explained Sum of Squares, ESS)

$$\sum \hat{u}_i^2$$ **Residual or unexpected variation of the Y values about the regression line (Residual Sum of Squares, RSS)**

R-squared

$$R^2 = \frac{ESS}{TSS}$$
$$= 1 - \frac{RSS}{TSS}$$
$$= 1 - \frac{\sum \hat{u}_i^2}{\sum y_i^2}$$

R-squared

- Lies between 0 and 1
- Nondecreasing function of the number of explanatory variables

$Adjusted \quad R^2$

$$\bar{R}^2 = 1 - \frac{\sum \hat{u}_i^2 /(n-k)}{\sum y_i^2 /(n-1)}$$

k = the number of parameters in the model including the intercept term

$$\bar{R}^2 = 1 - (1 - R^2)\frac{(n-1)}{(n-k)}$$

- For k>1, Adjusted R-Squared < R-Squared implies that as the number of X variables increases, the adjusted R-Squared increases less than the unadjusted R-Squared
- Adjusted R-Squared can be negative

Hypothesis Testing in Multiple Regression

1. Hypothesis Testing about Individual Regression Coefficients
2. Testing the Overall Significance of the Sample Regression
3. Testing the Equality of Two Regression Coefficients
4. Restricted Least Squares: Testing Linear Equality Restrictions
5. Testing for Structural or Parameter Stability of Regression Models: The Chow Test

Hypothesis Testing about Individual Regression Coefficients

- State the hypothesis $H_0 : \beta_j = 0$
  $H_1 : \beta_j \neq 0$
- t-value
  $$t = \frac{(\hat{\beta}_j - \beta_j)}{se(\hat{\beta}_j)}, df = n-3$$
- critical value $t > t_{\alpha/2} \quad or \quad t < t_{\alpha/2}$
- Conclusion
  Reject the null hypothesis if $t > t_{\alpha/2}$
  Not Reject the null hypothesis if $t < t_{\alpha/2}$

1) $H_0: \beta_2 = 0 \quad H_1: \beta_2 \neq 0$

2) $t = \dfrac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)} = \dfrac{-0.0056}{0.0020} = -2.8187$

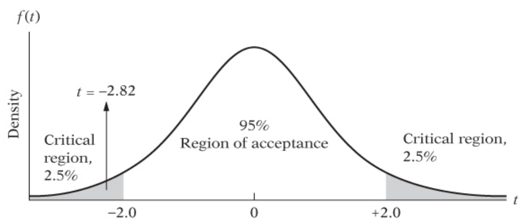
3) $df = 64 - 3 = 61$

The critical t value is 2 for a two-tail test $(\alpha = 0.05)$

4) *Since the computed t value of* 2.8187 *exceeds the critical t value of* 2

5) *We can reject the* H_0 *that PGNP has no effect on child mortality*

The female literacy rate held constant, per capita GNP has a significant negative effect on child mortality



Testing the Overall Significance of the Sample Regression-The F Test

Given the k-variable regression model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + u_i$$

$$H_0: \beta_2 = \beta_3 = \ldots = \beta_k = 0$$

$H_1:$ Not all slope coefficients are simultaneously zero

$$F = \frac{ESS/df}{RSS/df} = \frac{ESS/(k-1)}{RSS/(n-k)}$$

$$If \;\; F > F_\alpha(k-1, n-k), reject \; H_0$$

If F > critical region $F_{(1-\alpha);k-1,n-k}$ Reject H_0

If F < critical region $F_{(1-\alpha);k-1,n-k}$ Not reject H_0

Testing the Overall Significance of a Multiple Regression in Terms of R-Squared

Given the k-variable regression model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + \mu_i$$

$$H_0: \beta_2 = \beta_3 = \ldots = \beta_k = 0$$

$H_1:$ Not all slope coefficients are simultaneously zero

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

$$If \;\; F > F_\alpha(k-1, n-k), reject \; H_0$$

The "incremental" or Marginal contribution of an explanatory variable

In most empirical investigations the researcher may be completely sure whether it is worth adding an X variable to the model knowing that several other X variables are already present in the model

One does not wish to include a variable (s) that contributes very little toward ESS.
One does not want to exclude a variable (s) that substantially increases ESS

How does one decide whether an X variable significantly reduces RSS?

$$F = \frac{(ESS_{new} - ESS_{old}) / number\ \ of\ \ new\ regressors}{RSS_{new} / df\,(= n - number\ \ of\ \ parameters\ \ in\ \ the\ new\ Model)}$$

$$F = \frac{196{,}912.9}{1742.8786} = 112.9814$$

$$F > critical\ \ F$$

F value is highly significant, suggesting that the addition of FLR to the model significantly increases ESS and hence the R-square value

Source	SS	df	MS
Model	60449.4605	1	60449.4605
Residual	303228.539	62	4890.78289
Total	363678	63	5772.66667

Number of obs =	64
F(1, 62) =	12.36
Prob > F =	0.0008
R-squared =	0.1662
Adj R-squared =	0.1528
Root MSE =	69.934

cm	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
pgnp	-.0113645	.0032325	-3.52	0.001	-.0178262	-.0049027
_cons	157.4244	9.845583	15.99	0.000	137.7434	177.1055

Source	SS	df	MS
Model	257362.373	2	128681.187
Residual	106315.627	61	1742.87913
Total	363678	63	5772.66667

Number of obs =	64
F(2, 61) =	73.83
Prob > F =	0.0000
R-squared =	0.7077
Adj R-squared =	0.6981
Root MSE =	41.748

cm	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
pgnp	-.0056466	.0020033	-2.82	0.006	-.0096524	-.0016408
flr	-2.231586	.2099472	-10.63	0.000	-2.651401	-1.81177
_cons	263.6416	11.59318	22.74	0.000	240.4596	286.8236

$$F = \frac{(R^2_{new} - R^2_{old}) / number\ \ of\ \ new\ regressors}{(1 - R^2_{new}) / df\,(= n - number\ \ of\ \ parameters\ \ in\ \ the\ new\ Model)}$$

$$F = \frac{(0.7077 - 0.1662)/1}{(1 - 0.7077)/61} = 113.05$$

$$F > critical\ \ F_{\alpha}$$

Testing the Equality of Two Regression Coefficients

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$$

$$H_0 : \beta_3 = \beta_4\ \ or\ \ (\beta_3 - \beta_4) = 0$$

$$H_1 : \beta_3 \neq \beta_4\ \ or\ \ (\beta_3 - \beta_4) \neq 0$$

$$t = \frac{(\hat{\beta}_3 - \hat{\beta}_4) - (\beta_3 - \beta_4)}{se(\hat{\beta}_3 - \hat{\beta}_4)}$$

Degree of freedom = n-k

$$se(\hat{\beta}_3 - \hat{\beta}_4) = \sqrt{\operatorname{var}(\hat{\beta}_3) + \operatorname{var}(\hat{\beta}_4) - 2\operatorname{cov}(\hat{\beta}_3, \hat{\beta}_4)}$$

$$t = \frac{(\hat{\beta}_3 - \hat{\beta}_4)}{\sqrt{\operatorname{var}(\hat{\beta}_3) + \operatorname{var}(\hat{\beta}_4) - 2\operatorname{cov}(\hat{\beta}_3, \hat{\beta}_4)}}$$

Example

TABLE 7.4
Total Cost (Y) and Output (X)

Output	Total Cost, $
1	193
2	226
3	240
4	244
5	257
6	260
7	274
8	297
9	350
10	420

$$\hat{Y}_i = 141.7667 + 63.4777X_i - 12.9615X_i^2 + 0.9396X_i^3$$
$$se = (6.3753) \quad (4.7786) \quad (0.9857) \quad (0.0591)$$
$$COV(\hat{\beta}_3, \hat{\beta}_4) = -0.0576$$
$$R^2 = 0.9983$$

$$H_0 : \beta_3 = \beta_4$$
$$H_1 : \beta_3 \neq \beta_4$$
$$t = \frac{\hat{\beta}_3 - \hat{\beta}_4}{\sqrt{\mathrm{var}(\hat{\beta}_3) + \mathrm{var}(\hat{\beta}_4) - 2\,\mathrm{cov}(\hat{\beta}_3, \hat{\beta}_4)}}$$
$$= \frac{-12.9615 - 0.9396}{\sqrt{(0.9867)^2 + (0.0591)^2 - 2(-0.0576)}} = \frac{-13.9011}{1.0442} = -13.3130$$

Degree of freedom = n-k=10-4=6   Check critical value

Reject the null hypothesis

Restricted Least Squares: Testing Linear Equality Restrictions

Example

$$Y_i = \beta_1 X_{2i}^{\beta_2} X_{3i}^{\beta_3} e^{\mu_i}$$
$$\ln Y_i = \beta_0 + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + \mu_i$$
โดยที่ $\beta_0 = \ln \beta_1$

Is this restriction valid?

$$\beta_2 + \beta_3 = 1$$

The t-Test Approach

$$t = \frac{(\hat{\beta}_2 + \hat{\beta}_3) - (\beta_2 + \beta_3)}{se(\hat{\beta}_2 + \hat{\beta}_3)}$$
$$= \frac{(\hat{\beta}_2 + \hat{\beta}_3) - (\beta_2 + \beta_3)}{\sqrt{\mathrm{var}(\hat{\beta}_2) + \mathrm{var}(\hat{\beta}_3) + 2\,\mathrm{cov}(\hat{\beta}_2, \hat{\beta}_3)}}$$

The F-Test Approach: Restricted Least Squares

$\sum \hat{u}_{UR}^2$ RSS of the unrestricted regression

$\sum \hat{u}_{R}^2$ RSS of the restricted regression

m Number of linear restrictions

k Number of parameters in the unrestricted regression

n Number of observations

$$F = \frac{(RSS_R - RSS_{UR})/m}{RSS_{UR}/(n-k)} = \frac{(\sum \hat{u}_R^2 - \sum \hat{u}_{UR}^2)/m}{\sum \hat{u}_{UR}^2/(n-k)}$$

$$\sum \hat{u}_{UR}^2 \le \sum \hat{u}_R^2$$

F distribution with degree of freedom m, n-k

$$F = \frac{(R_{UR}^2 - R_R^2)/m}{(1 - R_{UR}^2)/(n-k)}$$

$$R_{UR}^2 \ge R_R^2$$

F distribution with degree of freedom m, n-k

Example

TABLE 8.8 Real GDP, Employment, and Real Fixed Capital—Mexico

Source: Victor J. Elias, *Sources of Growth: A Study of Seven Latin American Economies*, International Center for Economic Growth, ICS Press, San Francisco, 1992. Data from Tables E5, E12, and E14.

Year	GDP*	Employment†	Fixed Capital‡
1955	114043	8310	182113
1956	120410	8529	193749
1957	129187	8738	205192
1958	134705	8952	215130
1959	139960	9171	225021
1960	150511	9569	237026
1961	157897	9527	248897
1962	165286	9662	260661
1963	178491	10334	275466
1964	199457	10981	295378
1965	212323	11746	315715
1966	226977	11521	337642
1967	241194	11540	363599
1968	260881	12066	391847
1969	277498	12297	422382
1970	296530	12955	455049
1971	306712	13338	484677
1972	329030	13738	520553
1973	354057	15924	561531
1974	374977	14154	609825

*Millions of 1960 pesos.
†Thousands of people.
‡Millions of 1960 pesos.

$$Y_i = \beta_1 X_{2i}^{\beta_2} X_{3i}^{\beta_3} e^{\mu_i}$$

$$\widehat{\ln GDP_t} = -1.6524 + 0.3397 \ln Labor_t + 0.8460 \ln Capital_t$$

$$t \quad = (-2.7259) \quad (1.8295) \qquad (9.0625)$$

$$p\ value = (0.0144) \quad (0.0849) \qquad (0.0000)$$

$$R^2 = 0.9951 \qquad RSS_{UR} = 0.0136$$

As you can see, the output/labor elasticity is about 0.34 and the output/capital elasticity is about 0.85. If we add these coefficients, we obtain 1.19, suggesting that perhaps the Mexican economy during the stated time period was experiencing increasing returns to scale.

Let us impose the restriction of constant returns to scale

$$\widehat{\ln(GDP/Labor)_t} = -0.4947 + 1.0153 \ln(Capital/Labor)_t$$

$$t = (-4.0612) \quad (28.1056)$$

$$p\ value = (0.0007) \qquad (0.0000)$$

$$R_R^2 = 0.9777 \qquad RSS_R = 0.0166$$

$$F = \frac{(RSS_R - RSS_{UR})/m}{RSS_{UR}/(n-k)}$$
$$= \frac{(0.0166 - 0.0136)/1}{0.0136/(20-3)} = 3.75$$

F-distribution with degree of freedom 1, 17

F-value is not significant at the 5% level

The conclusion is that the Mexican economy was probably characterized by constant returns to scale over the sample period and therefore there may be no harm in using the restricted regression

Testing for Structural or Parameter Stability of Regression Models: The Chow Test

- "Structural Change" mean that the values of parameters of the model do not remain the same through the entire period

- Divide sample data into two time periods:
- 1970-1981 and 1982-1995

Three possible regressions:

Time period 1970-1981: $Y_t = \lambda_1 + \lambda_2 X_t + \mu_{1t} \quad n_1 = 12$

Time period 1982-1995: $Y_t = \gamma_1 + \gamma_2 X_t + \mu_{2t} \quad n_2 = 14$

Time period 1970-1995: $Y_t = \alpha_1 + \alpha_2 X_t + \mu_t \quad n = 26$

$$\hat{Y}_t = 1.0161 + 0.0803 X_t$$
$$t = (0.0873) \ (9.6015)$$
$$R^2 = 0.9021 \quad RSS_1 = 1785.032 \quad df = 10$$

$$\hat{Y}_t = 153.4947 + 0.0148 X_t$$
$$t = (4.6922) \ (1.7707)$$
$$R^2 = 0.2971 \quad RSS_2 = 10{,}005.22 \quad df = 12$$

$$\hat{Y}_t = 62.4226 + 0.0376 X_t + \ldots$$
$$t = (4.8917) \ (8.8937) + \ldots$$
$$R^2 = 0.7672 \quad RSS_3 = 23{,}248.30 \quad df = 24$$

The mechanics of the Chow test

1. Estimate $Y_t = \alpha_1 + \alpha_2 X_t + u_t \quad n = 26$, which is appropriate if there is no parameter instability, and obtain RSS_3 with df = $(n_1 + n_2 - k)$. We call RSS_3 the restricted residual sum of squares (RSS_R)
2. Estimate $Y_t = \lambda_1 + \lambda_2 X_t + u_{1t} \quad n_1 = 12$ and obtain its residual sum of squares, RSS_1, with df = $(n_1 - k)$
3. Estimate $Y_t = \gamma_1 + \gamma_2 X_t + u_{2t} \quad n_2 = 14$ and obtain its residual sum of squares, RSS_2, with df = $(n_2 - k)$

4. Since the two sets of samples are deemed independent, we can add RSS_1 and RSS_2 to obtain what may be called the **unrestricted residual sum of squares** (RSS_{UR})

$$RSS_{UR} = RSS_1 + RSS_2 \quad with\ df = (n_1 + n_2 - 2k)$$
$$RSS_{UR} = (1785.032 + 10{,}005.22) = 11{,}790.252$$

5. If there is no structural change, then the RSS_R and RSS_{UR} should not be statistically different.

$$F = \frac{(RSS_R - RSS_{UR})/k}{(RSS_{UR})/(n_1+n_2-2k)} \sim F_{[k,(n_1+n_2-2k)]}$$

then the Chow has shown that under the null hypothesis the regression $Y_t = \lambda_1 + \lambda_2 X_t + u_{1t}$ $n_1 = 12$ and $Y_t = \gamma_1 + \gamma_2 X_t + u_{2t}$ $n_2 = 14$ are statistically the

6. We find that for 2, 22 df the 1 percent critical F value is 5.72.

$$F = \frac{(23,248.30 - 11,790.252)/2}{(11,790.252)/22} = 10.69$$

Therefore, the probability of obtaining F value of as much as or greater than 10.69. We reject the null hypothesis of parameter stability and conclude that the regressions $Y_t = \lambda_1 + \lambda_2 X_t + u_{1t}$ $n_1 = 12$ and $Y_t = \gamma_1 + \gamma_2 X_t + u_{2t}$ $n_2 = 14$ are different

Dummy Variable Regression Models

$$Y_i = \beta_1 + \beta_2 X_i + \alpha D_i + u_i$$

$$Y_i = \beta_1 + \beta_2 X_i + \alpha D_i + u_i$$

$Y_i =$ total consumption (thousand baht)

$X_i =$ total income (thousand baht)

$D_i = \begin{cases} 1 & \text{male} \\ 0 & \text{female} \end{cases}$

$u_i =$ residual term

$$E(Y_i \mid D_i = 0) = \beta_1 + \beta_2 X_i + \alpha(0) = \beta_1 + \beta_2 X_i$$

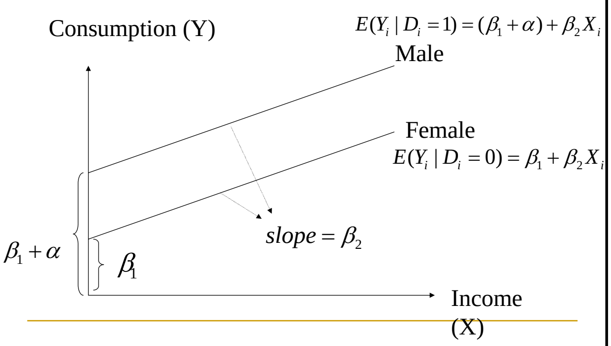
$$E(Y_i \mid D_i = 1) = \beta_1 + \beta_2 X_i + \alpha(1) = (\beta_1 + \alpha) + \beta_2 X_i$$

$$E(Y_i \mid D_i = 1) - E(Y_i \mid D_i = 0) = \alpha$$

The difference between mean total consumption of male and female is equal α

$\alpha > 0$ when the mean total consumption of male is more than female

$\alpha < 0$ when the mean total consumption of male is less than female

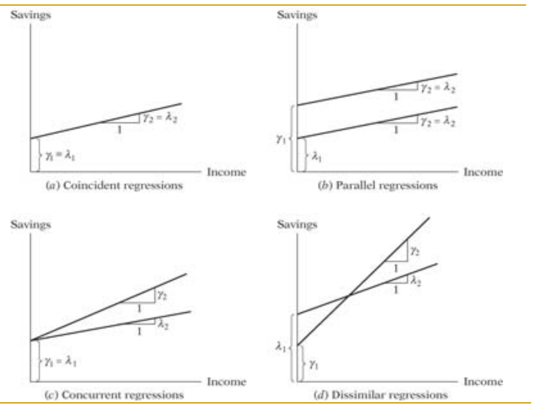


The Dummy Variable Alternative to Chow Test

Chow Test – examine the structural stability of a regression model

- **Coincident regressions** – both the intercept and the slope coefficients are the same in the two regressions
- **Parallel regressions**- only the intercepts in the two regressions are different but the slopes are the same

- **Concurrent regressions** – the intercepts in the two regressions are the same, but the slopes are different
- **Dissimilar regressions** – both the intercepts and slopes in the two regressions are different



Example

$$Y_t = \alpha_1 + \alpha_2 D_t + \beta_1 X_t + \beta_2 (D_t X_t) + u_t$$

Y = savings

X = income

t = time

D =1 for observations in 1982-1995

=0, otherwise

$$E(u_i) = 0$$

Mean savings function for 1970-1981

$$E(Y_t \mid D_t = 0, X_t) = \alpha_1 + \beta_1 X_t$$

Mean savings function for 1982-1995

$$E(Y_t \mid D_t = 1, X_t) = (\alpha_1 + \alpha_2) + (\beta_1 + \beta_2) X_t$$

Source	SS	df	MS
Model	88079.8327	3	29359.9442
Residual	11790.2539	22	535.920634
Total	99870.0867	25	3994.80347

Number of obs =	26
F(3, 22) =	54.78
Prob > F =	0.0000
R-squared =	0.8819
Adj R-squared =	0.8658
Root MSE =	23.15

savings	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	.0803319	.0144968	5.54	0.000	.0502673	.1103964
dum	152.4786	33.08237	4.61	0.000	83.86992	221.0872
incdum	-.0654694	.0159824	-4.10	0.000	-.098615	-.0323239
_cons	1.016115	20.16483	0.05	0.960	-40.80319	42.83542

$$Y_t = \alpha_1 + \alpha_2 D_t + \beta_1 X_t + \beta_2 (D_t X_t) + u_t$$

$$\hat{Y}_t = 1.0161 + 152.4786 D_t + 0.0803 X_t + 0.0655 (D_t X_t)$$
$$se = (20.1648) \ (33.0824) \quad (0.0144) \quad (0.0159)$$
$$t = (0.0504)^{**} (4.6090) \quad (5.5413)^{*} \quad (-4.0963)^{*}$$
$$R^2 = 0.8891$$

where * indicates p values less than 5 percent

** indicates p values greater than 5 percent

Savings function for 1970-1981

$$\hat{Y}_t = 1.0161 + 0.0803 X_t$$

Savings function for 1982-1995

$$\hat{Y}_t = (1.0161 + 152.4786) + (0.0803 - 0.0655) X_t$$
$$= 153.4947 + 0.0148 X_t$$

- The Chow test does not explicitly tell us which coefficient, intercept, or slope is different or whether both are different in two periods.
- That is, one can obtain a significant Chow test because the slope only is different or the intercept only is different or both are different

- Multicollinearity
- Heteroscedasticity
- Autocorrelation
- Type of Specification Error

Theoretical Consequences of Multicollinearity

- The OLS estimators are **unbiased**. But unbiasedness is a multisample or repeated sampling property

  Keeping the values of the variables X fixed, if one obtains repeated samples and computes the OLS estimators for each of these samples, the average of the sample values will converge to the true population values of the estimators as the number of sample increases

- Collinearity does not destroy the property of minimum variance
- In the class of all linear unbiased estimators, the OLS estimators have minimum variance- they are **efficient**
- But this does not mean that the variance of an OLS estimator will necessarily be small

- Multicollinearity is essentially a sample (regression) phenomenon, even if the X variables are not linearly related in the population, they may be so related in the particular sample
- When we postulate the theoretical or population regression function (PRF), we believe that all the X variables included in the model have a separate or independent influence on the dependent variable Y.

$$Consumption_i = \beta_1 + \beta_2 Income_i + \beta_3 Wealth_i + u_i$$

Two variables may be highly , if not perfectly, correlated: Wealthier people generally tend to have higher incomes.

To assess the individual effects of wealth and income on consumption expenditure we need a sufficient number of sample observations of wealthy individuals with low income, and high income individuals with low wealth

- OLS estimators are **BLUE** despite multicollinearity

Detection of Multicollinearity

- High R-Squared but few significant t-ratios
- High pair-wise correlations among regressors
- Examination of partial correlations
- Auxiliary regressions
- VIF
- Scatter plot

Example

Source	SS	df	MS
Model	8565.55407	2	4282.77704
Residual	324.445926	7	46.349418
Total	8890	9	987.777778

Number of obs =	10
F(2, 7) =	92.40
Prob > F =	0.0000
R-squared =	0.9635
Adj R-squared =	0.9531
Root MSE =	6.808

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x2	.9415373	.8228983	1.14	0.290	-1.004308	2.887383
x3	-.0424345	.0806645	-0.53	0.615	-.2331757	.1483067
_cons	24.77473	6.7525	3.67	0.008	8.807609	40.74186

Example

Correlation between income (x2) and wealth (x3)

	x2	x3
x2	1.0000	
x3	0.9990	1.0000

variance-inflating factor (VIF)

VIF shows how the variance of an estimator is inflated by the presence of multicollinearity

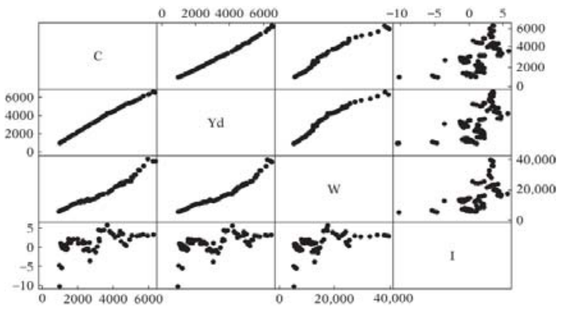
$$VIF = \frac{1}{(1 - r_{23}^2)}$$

$$r_{23}^2 \to 1, VIF \to \infty$$

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2} VIF$$

$$\text{var}(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2} VIF$$

Scatter plot



Remedial Measures

- Do Nothing
- Rule-Thumb Procedures

1. A priori information
2. Combining cross-sectional and time series data
3. Dropping a variable (s) and specification bias
4. Adding or new data
5. Transformation of variables

Multicollinearity may not pose a serious problem

- When R-squared is high and the regression coefficients are individually significant as revealed by the higher t values

OLS Estimation in the Presence of Heteroscedasticity

$\hat{\beta}_2$ *is best linear unbiased estimator (BLUE)*

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

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

$$\text{var}(\hat{\beta}_2) = \frac{\sum x_i^2 \sigma_i^2}{(\sum x_i^2)^2}$$

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2} \quad \textit{Homoscedasticity}$$

Is $\hat{\beta}_2$ still BLUE when we drop only homoscedasticity assumption and replace it with the assumption of heteroscedasticity?

$\hat{\beta}_2$ is **no longer best and the minimum variance** given by

$$\text{var}(\hat{\beta}_2) = \frac{\sum x_i^2 \sigma_i^2}{(\sum x_i^2)^2}$$

Detection of Heteroscedasticity

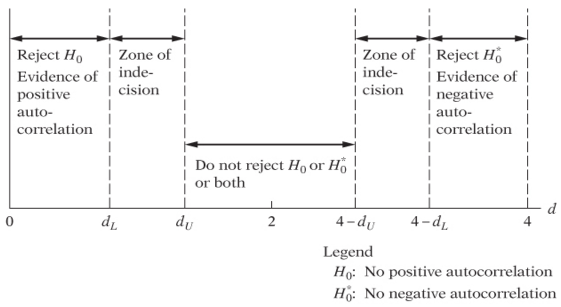
- Formal methods
  - **White's General Heteroscedasticity Test**

As in the case of **heteroscedasticity** in the presence of **autocorrelation** the OLS estimators are still linear unbiased as well as consistent and asymptotically normally distributed, but they are **no longer efficient (i.e., minimum variance)**

Detecting Autocorrelation

- **Durbin-Watson d Test**

Durbin-Watson d statistic



Correcting for (Pure) Autocorrelation

1. **The method of Generalized Least Squares (GLS) method**

2. The Newey West method – to obtain standard errors of OLS estimators that are corrected for autocorrelation