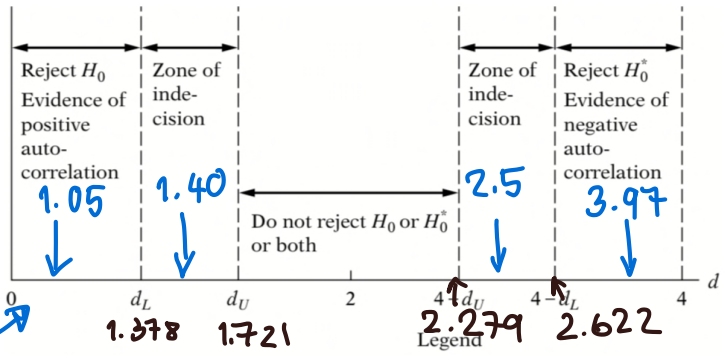


FIGURE 12.10 Durbin–Watson d statistic.

Now let us define

$$\hat{\rho} = \frac{\sum \hat{u}_t \hat{u}_{t-1}}{\sum \hat{u}_t^2} \qquad (12.6.9)$$

as the sample first-order coefficient of autocorrelation, an estimator of ρ. (See footnote 9.) Using (12.6.9), we can express (12.6.8) as

$$d \approx 2(1 - \hat{\rho}) \qquad (12.6.10)$$

But since $-1 \leq \rho \leq 1$, (12.6.10) implies that

$$0 \leq d \leq 4 \qquad (12.6.11)$$

These are the bounds of d; any estimated d value must lie within these limits.

← from Gujarati Chapter posted on Moodle.

a) Reject H_0 in favor of positive autocorrelation

b) in the zone of indecision. Cannot conclude if there is autocorr.

c)

d) Reject H_0^* in favor of negative autocorrelation.

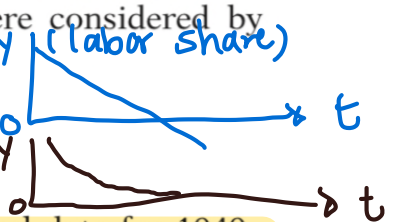
← look up the stat table!

12.2. Given a sample of 50 observations and 4 explanatory variables, what can you say about autocorrelation if (*a*) $d = 1.05$? (*b*) $d = 1.40$? (*c*) $d = 2.50$? (*d*) $d = 3.97$?

12.3. In studying the movement in the production workers' share in the value added (i.e., labor's share), the following models were considered by Gujarati*:

Model A: $Y_t = \beta_0 + \beta_1 t + u_t$

Model B: $Y_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + u_t$



where Y = labor's share and t = time. Based on annual data for 1949–1964, the following results were obtained for the primary metal industry:

Model A: $\hat{Y}_t = 0.4529 - \underset{(-3.9608)}{0.0041t}$ $\quad R^2 = 0.5284$ $\quad d = 0.8252$

n = 16 (periods)
k = 1

Model B: $\hat{Y}_t = 0.4786 - \underset{(-3.2724)}{0.0127t} + \underset{(2.7777)}{0.0005t^2}$

$R^2 = 0.6629$ $\quad d = 1.82$

n = 16
k = 2

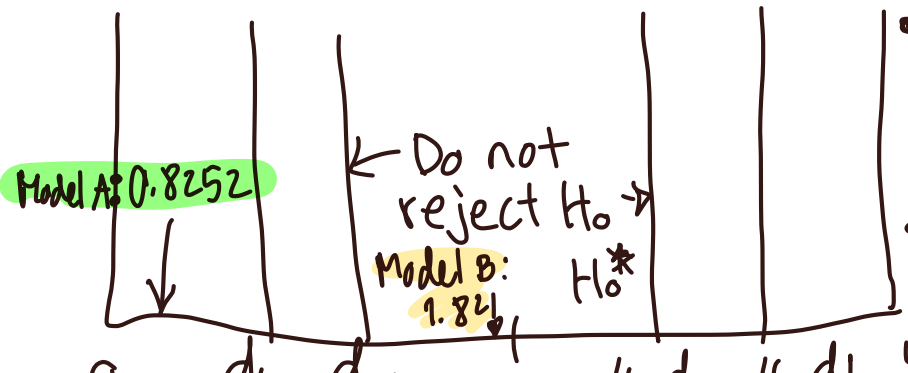
* Which model A or B do you think is better at explaining the data? Why?
⇒ Model B because higher R^2.

where the figures in the parentheses are t ratios.

a. Is there serial correlation in model A? In model B? → Obtain d_L, d_U Model A? Model B

b. What accounts for the serial correlation?

c. How would you distinguish between "pure" autocorrelation and specification bias?



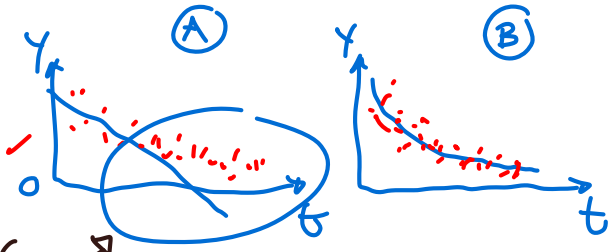
- In model A, we reject H_0 in favor of positive auto correlation.
- But in Model B, we do not reject H_0, H_0^*. So, cannot conclude that there is auto correlation (no auto correlation)

	d_L	d_u		4-du	4-dL
model A	1.106	1.371	2	2.629	2.894
model B	0.982	1.539		2.461	3.018

b) What accounts for serial correlation ??

Some causes?

- Inertia
- Incorrect functional form
- Cobweb Phenomenon



errors establishes autocorrelation pattern b/c the function is not correct!

c) How would you distinguish btw pure autor.. and spec. bias?

- Plot of data
- Compare autocorrelation test under many specifications.

Multicolinearity → when the X variables are so similar to each other, or so related to each other?

1 The Nature of Multicolinearity

- X variables in the model are highly correlated, though not perfectly correlated.

If this happens, STATA will drop either X_1 or X_2 anyway.

Perfectly correlated ($|\rho|=1$)

observation	x_1	x_2	$3x_1 - x_2$
1	6	18	0
2	12	36	0
3	7	21	0
4	-5	-15	0

Highly correlated ($|\rho|>0.8$)

observation	x_1	x_2	$3x_1 - x_2$
1	6	16	-2
2	12	45	9
3	7	18	-3
4	-5	-12	3

↑ STATA, OLS will not drop X_1 or X_2 BUT $Var(\hat{\beta}_{x_1})$, $Var(\hat{\beta}_{x_2})$ will be very large. So, the impacts of X_1 & X_2 will appear to be insignificant.

Causes

1) Data collection method, not enough sample variation.
2) Constraint on the model or on the sampled population.
3) Model specification (usually happens when there are too many polynomial terms, e.g. X^2, X^3, X^4).
4) There are too many X variables representing the same factor / concept.

2 Consequences of Multicolinearity

2.1 The OLS estimator will still be BLUE.

If MLR 1~5 are satisfied ⇒ $\hat{\beta}$ BLUE

2.2 The variances and covariances will be very large. This makes precise estimation difficult.

← even though they are correct in theory.

Recall $Var(\hat{\beta}_j) = \dfrac{\sigma^2}{SST_j(1-R_j^2)}$

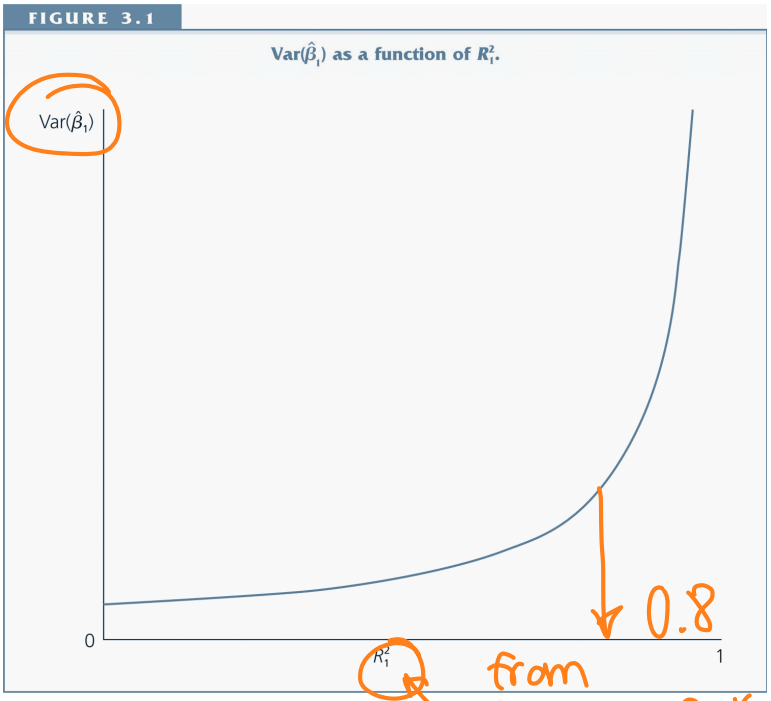
If X_j is highly correlated with other Xs, R_j^2 will be high ⇒ $Var(\hat{\beta}_j)$ to be high.

- SST_j is the SST from regressing X_j on all other Xs.

For example: SST_2 is SST from: $X_2 = \theta_0 + \theta_1 X_1 + \theta_2 X_3 + .. + error$.

R_2^2 is R^2 from regressing X_2 on all other Xs.

FIGURE 3.1



from $X_1 = \theta_0 + \theta_1 X_2 + \theta_2 X_3 + \ldots + \text{error}$.

3 Detection of multicolinearity

1. There is conflicting test between t- and F-test: if we find that the conclusion derived from the two tests are inconsistent, specifically R2 is high and F-test results in statistical overall significance; whereas, at least, one null hypothesis of some t-tests cannot be rejected, it is reasonable to suspect the multicollinearity problem.

2. Correlation of regressors is greater than 0.8: the higher the correlation, the higher the variance of estimators.

3. Variance inflation factor (VIF) is greater than 10: when the regressors face the multicollinearity problem, the value of VIF might be so high that the resulting high variance of estimators adversely affects the regression analysis.

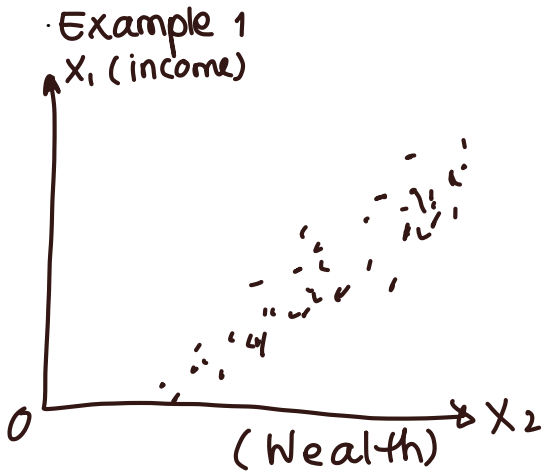
   - The VIF (variance inflation factor) to detect high multicolinearity:

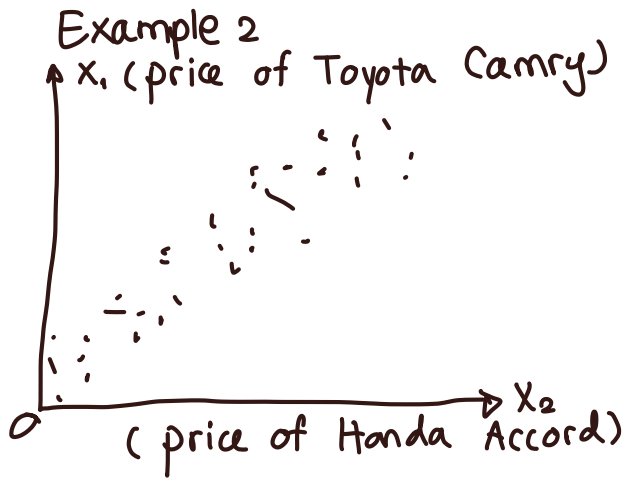
$$VIF_j = \frac{1}{1 - R_j^2}$$

← the higher VIF, the more severe the multi. col. problem.

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j} \cdot VIF_j$$

4. Scatter plot of two regressors is relatively linear: when we plot the value of on regressor against another and we find that both of them tend to change in the same way, this fact might suggest the existence of multicollinearity.





4 Remedial Measures

1. Do nothing → Because we still have BLUE
   - If the multicollinearity problem is not too severe.
   - ~~_____~~ among the control variables are not variables of interest.

2. Apply prior relationship among explanatory variables - $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$ (1)
   - Suppose "$X_2 = 0.7X_1$" → X_2 = total score of a university course. → X_1 = cummulative score before final exam of a university course.
   - Then (1) becomes $Y = \beta_0 + \beta_1 X_1 + 0.7\beta_2 X_1 + u$
     $= \beta_0 + \beta_1 (X_1 + 0.7X_1) + u$ → Can only determine the value of $\hat{\beta}_1$ not both $\hat{\beta}_1$ & $\hat{\beta}_2$.

3. Discard some explanatory variables - the removal of the variables could mitigate the problem; but, another problem, namely specification bias problem, might occur instead. For example, suppose we want to construct the model where the production is the explained variables; and labor and capital are the explanatory ones. If there is linear relationship between labor and capital, the elimination of one variable might assuage the multicollinearity problem, but might be contrary to economic reasoning. Hence, the decision of which variables will be disposed of should be based on economic theory.

4. Collect more observations - this practice will increase SST_j ↑ increase with sample variation., which is the component of the variances. As a result, the variances will be lower despite high correlation among explanatory variables.

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j} \cdot \frac{1}{(1 - R_j^2)}$$

R^2 if this ↑, $Var(\hat{\beta}_j)$ ↓

5. Transform the variables - although there is linear relationship among explanatory variables, it is not necessary that the first difference or ratio transformation of the variables will have that relationship

For example, first difference

$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t$ (1)

$Y_{t-1} = \beta_0 + \beta_1 X_{1,t-1} + \beta_2 X_{2,t-1} + u_{t-1}$ (2)

(1) − (2) gives

$Y_t - Y_{t-1} = \beta_0 + \beta_1 (X_{1t} - X_{1,t-1}) + \beta_2 (X_{2t} - X_{2,t-1}) + (u_t - u_{t-1})$