ALLAMA IQBAL OPEN UNIVERSITY

USMAN NOTES
www.usmannotes.com

Name:

Registration No.:

Course Code: 807

Program: M.Sc

Semester: Autumn 2026

ALLAMA IQBAL OPEN UNIVERSITY, ISLAMABAD

Course: Basic Econometrics (807) • Level: M.Sc

Semester: Autumn 2026

Assignment No. 2 (Units 5–9)

Total Marks: 100 • Pass Marks: 40 • Credit Hours: 03

SOLVED ASSIGNMENT GUIDE

(For Study & Reference)

USMAN NOTES www.usmannotes.com

Note: This guide is prepared for study and reference purposes. Students are advised to read the relevant units of the course book, understand the concepts, and write the assignment in their own words. Do not copy it verbatim.

Question 1: Distributed Lag Models and Autoregressive Models — and the Difference Between Them (20 Marks)

Write down the relationship between two variables in the form of a distributed lag model and auto regressive model. What is the difference between the two models?

Introduction

Economic action rarely responds instantly and completely to its causes. A change in income changes consumption gradually as habits adjust; an advertising campaign affects sales over many months; an investment induced by a fall in interest rates is planned, built and brought into production over years. Static regression, which relates Y in period t only to X in period t, cannot represent such behaviour. Time-series econometrics answers with two closely related

but conceptually distinct families of models. In a distributed lag model, the dependent variable responds to current and past values of the explanatory variable, the effect being distributed over time. In an autoregressive model, the dependent variable depends on its own past values: the past of Y itself carries the dynamics. This answer writes down both models in general and specific (Koyck) forms, explains their properties, derives the bridge between them, and then sets out their differences systematically.

The Distributed Lag Model

Let Y denote, for example, consumption or sales and X income or advertising expenditure. The finite distributed lag model of order q is:

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \ldots + \beta_q X_{t-q} + u_t$$

The coefficients β_i are the lag weights. β_0 is the impact (short-run) multiplier — the effect of a unit change in X on Y within the same period; $\beta_0 + \beta_1 + \ldots + \beta_q$ is the long-run (total distributed lag) multiplier — the full effect after all lags have worked through; and the partial sums $\beta_0 + \ldots + \beta_j$ are the interim multipliers. Dividing each weight by the total gives the pattern of adjustment over time, from which the mean lag, $\Sigma(i \cdot \beta_i)/\Sigma\beta_i$, and the median lag summarise how fast the response occurs. If q is infinite the model is an infinite distributed lag, which cannot be estimated directly because it contains infinitely many parameters; structure must be imposed on the weights. The Almon (polynomial) lag assumes the weights lie on a polynomial in i, while the Koyck (geometric) lag assumes the weights decline geometrically, $\beta_i = \beta_0 \lambda^i$ with $0 < \lambda < 1$, so that distant causes matter progressively less — the pattern shown in Figure 1.

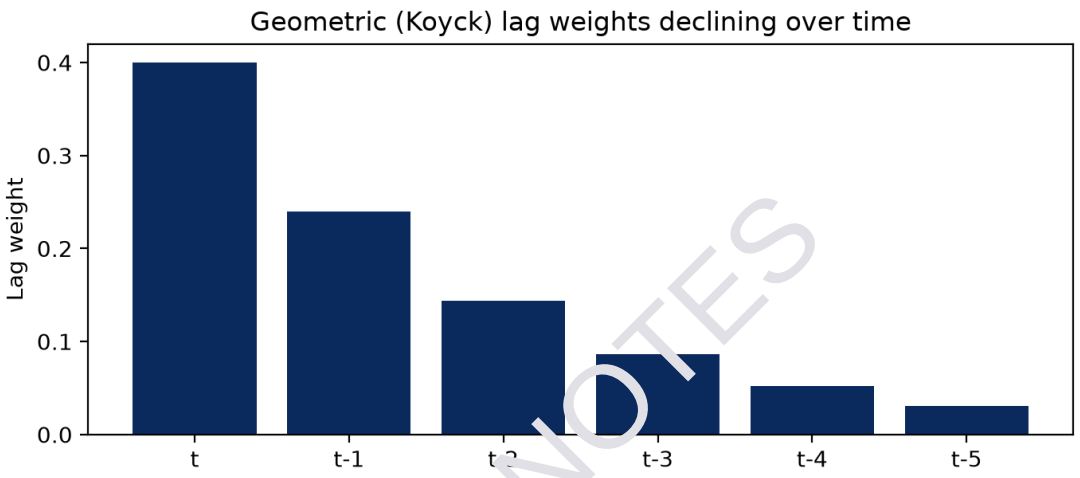


Figure 1: Geometric (Koyck) lag weights with λ = 0.6: the largest effect is immediate and later weights decline by the constant ratio λ.

Unrestricted estimation of a long distributed lag is impractical: each lag consumes a degree of freedom and a sample observation, and successive lags of X are typically highly correlated, producing severe multicollinearity, imprecise and often wrongly-signed weights. Structured lag schemes (Almon, Koyck) exist precisely to replace many free parameters with a few.

The Autoregressive Model

An autoregressive model of order p, AR(p), explains Y by its own past:

$$Y_t = \alpha + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p\, Y_{t-}p + u_t$$

The simplest and most used case is AR(1), $Y_t = \alpha + \phi Y_{t-1} + u_t$ with $|\phi| < 1$ for stationarity. Autoregression embodies inertia, habit, expectations and partial adjustment: today's consumption resembles yesterday's because habits persist; today's price level reflects yesterday's. The canonical economic derivation is the partial adjustment model: if desired consumption is $Y^*_t = \alpha + \beta X_t$ and actual consumption adjusts each period by a fraction δ of the

gap, $Y_t - Y_{t-1} = \delta(Y^*_t - Y_{t-1})$, the estimable equation is $Y_t = \delta\alpha + \delta\beta X_t + (1 - \delta)Y_{t-1} + u_t$ — an autoregressive (more precisely autoregressive distributed lag, ADL) equation in which the coefficient on Y_{t-1} measures the speed of adjustment. Adaptive expectations models (Cagan, Friedman, Nerlove) generate the same form: expectations are revised by a fraction of the last forecast error, again producing a lagged dependent variable.

The Bridge: The Koyck Transformation

The two families meet in the Koyck transformation. Start from the geometric distributed lag $Y_t = \alpha + \beta_o \Sigma \lambda^i X_{t-i} + u_t$. Lag the equation one period, multiply by λ, and subtract: $Y_t - \lambda Y_{t-1} = \alpha(1 - \lambda) + \beta_o X_t + (u_t - \lambda u_{t-1})$, that is,

$$Y_t = \alpha(1 - \lambda) + \beta_o X_t + \lambda Y_{t-1} + v_t, \quad v_t = u_t - \lambda u_{t-1}$$

An infinite distributed lag has been rewritten as a simple autoregressive equation with one current regressor and the lagged dependent variable. The long-run multiplier is recoverable as $\beta_o/(1 - \lambda)$. But the transformation creates two estimation problems that dominate the practical literature: Y_{t-1} is correlated with the composite error v_t (since Y_{t-1} depends on u_{t-1}), so OLS is biased and inconsistent; and v_t is serially correlated by construction, invalidating the usual tests and making the Durbin–Watson statistic biased toward 2 (Durbin's h test or the Breusch–Godfrey test must be used instead). Instrumental variables or maximum likelihood estimation is therefore required for such models.

The Differences Between the Two Models

Table 1: Distributed lag model versus autoregressive model

Feature	Distributed lag model	Autoregressive model

Regressors	Current and lagged values of the explanatory variable X	Lagged values of the dependent variable Y itself (often plus current X in ADL form)
Source of dynamics	The response to an external cause is spread over time	Inertia within Y: habit, partial adjustment, adaptive expectations
Typical equation	$Y_t = \alpha + \Sigma\beta_i X_{t-i} + u_t$	$Y_t = \alpha + \Sigma\phi_i Y_{t-i} (+ \beta X_t) + u_t$
Multipliers	Impact, interim and long-run multipliers $\Sigma\beta_i$ read directly from the lag weights	Long-run effect computed as $\beta/(1 - \Sigma\phi_i)$; adjustment speed from the ϕ coefficients
Main estimation problem	Too many parameters; multicollinearity among lagged X's	Lagged dependent variable correlated with (transformed) disturbance; OLS biased/inconsistent; DW invalid
Data cost	Loses q initial observations for the lags of X	Loses p initial observations for the lags of Y

Beyond the table, three differences deserve emphasis. First, interpretation: in a distributed lag model the dynamics describe how the world impinges on Y with delay, whereas in an autoregressive model they describe Y's own memory — a shock persists because Y carries it forward, with the effect of a disturbance at time t decaying as ϕ, ϕ^2, ϕ^3, … in AR(1). Second, forecasting: an autoregressive model can forecast Y from its own history alone, which is why AR

processes anchor time-series forecasting (the Box–Jenkins ARIMA family generalises them), while a distributed lag model needs future values of X before it can forecast Y. Third, nesting: the general ADL model $Y_t = \alpha + \phi Y_{t-1} + \beta_0 X_t + \beta_1 X_{t-1} + u_t$ contains both as special cases ($\phi = \beta_1 = 0$ gives a short distributed lag; $\beta_0 = \beta_1 = 0$ gives pure AR(1)), and modern practice often estimates the general form and tests down, rather than imposing either restriction blindly.

Conclusion

A distributed lag model writes Y_t as a function of current and past X with weights β_i whose sum is the long-run multiplier; an autoregressive model writes Y_t as a function of its own past values, capturing inertia and partial adjustment. They differ in regressors, interpretation, forecasting requirements and estimation pitfalls — multicollinearity and parameter proliferation on one side, correlation between the lagged dependent variable and the disturbance on the other — yet the Koyck transformation shows they are two faces of the same dynamics: a geometric distributed lag is exactly an autoregressive equation in disguise.

Question 2: Dummy Variables and Interaction Effects (20 Marks)

What is a dummy variable? Explain interaction effects using dummy variables?

Introduction

Not every influence on an economic variable is naturally measured in numbers. Gender, season, region, marital status, membership of a trade union, possession of a degree, peace versus war, a policy in force versus not in force — these are qualitative attributes, yet they plainly shift intercepts and slopes in regression models. The dummy variable is the device that admits such attributes into regression analysis without abandoning its quantitative machinery. This answer defines dummy variables and their rules of use, including the dummy variable trap, then turns

to the question's main demand: interaction effects, in which a dummy variable is multiplied by another variable (qualitative or quantitative) so that the effect of one variable is allowed to depend on the category to which an observation belongs.

What Is a Dummy Variable?

A dummy variable (also called an indicator, binary or dichotomous variable) is an artificial variable that takes the value 1 when a specified attribute is present and 0 when it is absent: D = 1 if female, 0 if male; D = 1 if the quarter is a festival quarter, 0 otherwise. Despite its artificiality it is treated in estimation exactly like any other regressor, and its coefficient obeys the usual t-tests. With a single dummy entering additively, $Y_i = \beta_1 + \beta_2 D_i + \beta_3 X_i + u_i$, the coefficient β_2 is a differential intercept: for D = 0 the intercept is β_1, for D = 1 it is $\beta_1 + \beta_2$, so β_2 measures the vertical shift of the regression line between the two groups, holding X constant, and its t-test answers whether the two groups' lines differ in level at all. The group coded 0 is the base (reference, benchmark) category, against which all comparisons are made.

Rules and Cautions: The Dummy Variable Trap

- Number of dummies: for an attribute with m categories, include only m − 1 dummy variables when an intercept is present. Including all m dummies plus an intercept creates perfect multicollinearity — the dummies sum to the intercept column — the celebrated dummy variable trap, under which OLS cannot be computed. The alternatives are to drop one category (making it the base) or to drop the intercept.
- Interpretation is relative: every dummy coefficient is measured against the omitted base category; changing the base changes the printed coefficients but not the model, its fit, or any test of group differences.

- Dummies are not only for intercepts: a dummy may also shift a slope (see interaction effects below), may capture outliers or structural breaks (a pulse dummy for a strike year, a step dummy after a policy reform, the basis of the Chow-type dummy test for stability), and seasonal dummies (three for quarterly data, eleven for monthly) are the standard way to model seasonality in time series.
- Care with coding: the 0/1 coding is conventional but not mandatory; what matters is consistent definition and correct interpretation. A dummy must never be treated as if its values were a genuine quantitative scale.

Interaction Effects: When Effects Depend on the Group

The additive dummy model forces the two groups to have parallel regression lines: X has the same slope for both groups, and the dummy shifts only the intercept. Frequently this is exactly what theory denies. The return to a year of schooling may differ for men and women; the consumption response to income may differ between urban and rural households; a training programme may raise the productivity of educated workers more than of uneducated ones. An interaction effect exists when the effect of one variable on Y depends on the value of another variable. To model it, add the product of the two variables as an additional regressor.

Dummy–quantitative interaction. Let D be a dummy (say 1 = female, 0 = male) and X years of schooling in a wage equation. The interacted model is:

$$Y_i = \beta_1 + \beta_2 D_i + \beta_3 X_i + \beta_4 (D_i \cdot X_i) + u_i$$

Setting D = 0 gives the male equation $Y = \beta_1 + \beta_3 X$; setting D = 1 gives the female equation $Y = (\beta_1 + \beta_2) + (\beta_3 + \beta_4)X$. Thus β_2 is the intercept difference, β_3 is the slope for the base group, β_4 is the

slope difference — the interaction effect — and $\beta_3 + \beta_4$ is the slope for the D = 1 group. The t-test on β_4 tests directly whether the slopes differ; the model nests the parallel-lines case ($\beta_4 = 0$) and the equal-lines case ($\beta_2 = \beta_4 = 0$, testable by an F-test, which is the dummy-variable form of the Chow test for structural difference between groups). Figure 1 illustrates the geometry.

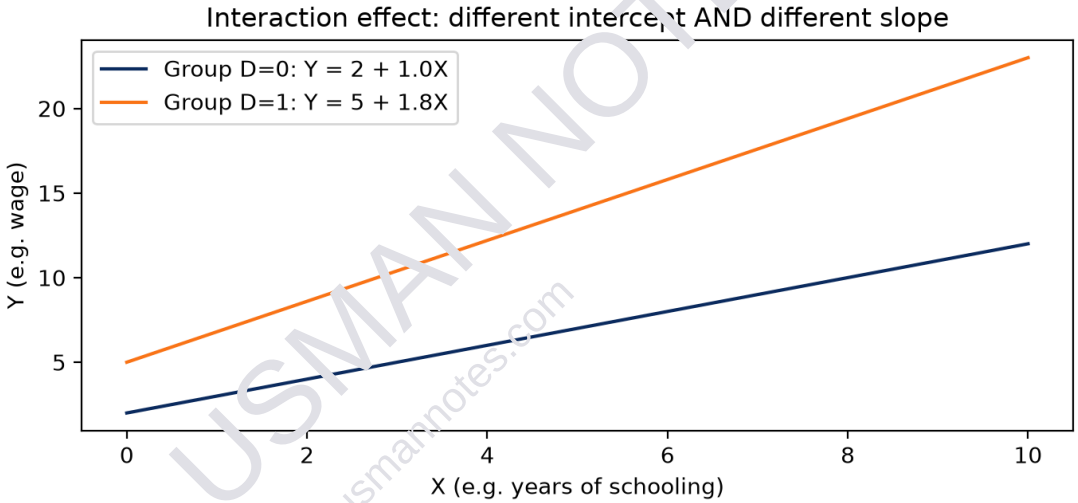


Figure 1: An interaction between a dummy D and a quantitative variable X produces lines with different intercepts and different slopes.

Dummy–dummy interaction. When both variables are qualitative, the product term captures the extra effect of belonging to both categories at once, beyond the sum of the separate effects. Example: wage = f(female dummy, union dummy, female × union): the interaction coefficient measures whether the union premium (or penalty) for women differs from that for men — equivalently, whether the gender gap differs between union and non-union workers. In a 2 × 2 layout the interaction is a difference of differences: $(\bar{Y}_{11} - \bar{Y}_{10}) - (\bar{Y}_{01} - \bar{Y}_{00})$. Table 1 works such a case in full.

Table 1: Dummy–dummy interaction — mean wages (illustrative figures, Rs '000 per month)

Group	Non-union (U=0)	Union (U=1)	Union effect
Male (D=0)	80	95	+15
Female (D=1)	68	78	+10
Gender gap	−12	−17	Interaction = 10 − 15 = −5

Reading the table into regression form with male non-union as base: intercept 80; female dummy −12; union dummy +15; interaction −5. The check is exact: female union mean = 80 − 12 + 15 − 5 = 78. The interaction of −5 says the union premium is 5 thousand rupees smaller for women than for men (or, symmetrically, the gender penalty is 5 thousand larger in union jobs). Without the interaction term the model would force one common union premium and one common gender gap, misrepresenting both. (The wage figures are illustrative.)

Quantitative–quantitative interaction, such as X·Z in $Y = \beta_1 + \beta_2 X + \beta_3 Z + \beta_4 XZ + u$, follows the same logic: the marginal effect of X becomes $\partial Y/\partial X = \beta_2 + \beta_4 Z$, varying with Z. Two practical rules apply to all interactions. First, never include a product term while omitting its constituent main effects, or the product will absorb level effects and be misinterpreted. Second, interpret main effects conditionally: in the wage model β_3 is the return to schooling for the base group only, not an 'average' return. Centring continuous variables before forming products can ease computation and interpretation and reduces the (harmless but alarming-looking) collinearity between a variable and its product.

Conclusion

A dummy variable is a 0/1 regressor that brings qualitative attributes into regression, subject to the m − 1 rule that avoids the dummy variable trap, with coefficients interpreted relative to the omitted base category. Used additively it shifts intercepts; multiplied by another variable it creates an interaction effect that lets slopes — and, in dummy–dummy models, group premiums — differ across categories. Interaction terms are how regression expresses the economically vital idea that the effect of a variable depends on context, and their coefficients are tested by the ordinary t and F machinery.

Question 3: ARCH and GARCH Models (20 Marks)

Explain Autoregressive Conditional Heteroskedasticity (ARCH) and Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models. [Note: the question paper prints 'GARH', which is a typographical error for GARCH.]

Introduction

Classical regression assumes homoscedasticity: a constant error variance. Financial time series mock that assumption. Returns on shares, exchange rates and interest rates pass through long calm stretches punctuated by storms of turbulence — the phenomenon Mandelbrot summarised as 'large changes tend to be followed by large changes, of either sign, and small changes tend to be followed by small changes'. This volatility clustering means the variance of the disturbance, conditional on recent history, is itself changing over time and, crucially, is partly predictable. Robert Engle's ARCH model (1982), extended by his student Tim Bollerslev into GARCH (1986), turned that insight into the standard framework for modelling and

forecasting volatility — work recognised by Engle's Nobel Prize in 2003. This answer develops both models, their properties, estimation and testing, and their extensions.

The ARCH Model

Let the mean equation be $y_t = x_t'\beta + \varepsilon_t$, with $\varepsilon_t = \sigma_t z_t$, $z_t \sim$ i.i.d. $(0, 1)$ (often standard normal or Student-t). In the ARCH(q) model the conditional variance σ_t^2 depends on the squares of the past q disturbances:

$$\sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + \ldots + \alpha_q \varepsilon_{t-q}^2, \quad \omega > 0, \alpha_i \geq 0$$

A large shock yesterday (big ε_{t-1}^2) raises today's conditional variance, making a large movement today more likely — volatility clustering is built in. The ARCH(1) case, $\sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2$, shows the mechanics: the unconditional variance is $\omega/(1 - \alpha_1)$, which is finite only if $\alpha_1 < 1$ (the stationarity condition); the model is covariance stationary while its conditional variance moves. Properties worth stressing: (i) the disturbances remain serially uncorrelated in level — $E(\varepsilon_t) = 0$ and no autocorrelation — so OLS on the mean equation stays consistent, but they are dependent through their squares; (ii) even with normal z_t, the unconditional distribution of ε_t is fat-tailed (leptokurtic), matching the excess kurtosis of financial returns; (iii) variance forecasts revert to the unconditional variance at a speed governed by α_1. Detection uses Engle's ARCH-LM test: regress the squared residuals $\hat{e}_t^2$ on a constant and q of their own lags and test nR^2 against $\chi^2(q)$; rejection signals ARCH effects. Squared-residual autocorrelations (Ljung–Box on squares) serve the same purpose.

The GARCH Model

ARCH has a practical drawback: capturing persistent volatility may require many lags q, hence many parameters, with non-negativity constraints awkward to impose. Bollerslev's GARCH(p, q) lets the conditional variance depend on its own past as well:

$$\sigma_t^2 = \omega + \Sigma\, \alpha_i\, \varepsilon_{t-i}^2 \ \text{(ARCH terms, i = 1..q)} \ + \ \Sigma\, \beta_j\, \sigma_{t-j}^2 \ \text{(GARCH terms, j = 1..p)}$$

The workhorse GARCH(1,1), $\sigma_t^2 = \omega + \alpha_1\varepsilon_{t-1}^2 + \beta_1\sigma_{t-1}^2$, captures most empirical volatility dynamics with just three variance parameters. It is equivalent to an ARCH(∞) with geometrically declining weights: yesterday's variance summarises the whole past. The sum $\alpha_1 + \beta_1$ measures volatility persistence — shocks to variance decay at the rate $(\alpha_1 + \beta_1)^h$ over horizon h — and in financial data it is typically 0.95–0.99, indicating highly persistent but (if the sum is below 1) still stationary volatility; the unconditional variance is $\omega/(1 - \alpha_1 - \beta_1)$. If $\alpha_1 + \beta_1 = 1$ the model is integrated (IGARCH) and variance shocks are permanent. Figure 1 shows a simulated series exhibiting exactly the clustering these models describe.

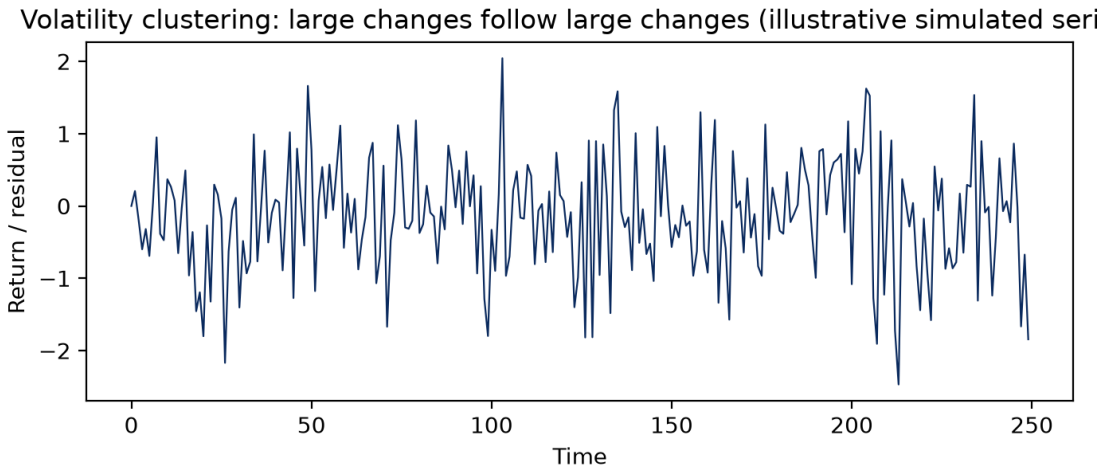


Figure 1: Volatility clustering in an illustrative simulated GARCH(1,1) series: turbulent and calm episodes alternate, while the series itself shows no pattern in level.

Estimation, Forecasting and Use

ARCH/GARCH models are estimated by maximum likelihood, jointly with the mean equation: assuming a conditional distribution for z_t (normal, or Student-t and GED to accommodate fat tails), the log-likelihood $\Sigma[-\tfrac{1}{2}\ln \sigma_t^2 - \varepsilon_t^2/(2\sigma_t^2)]$ (plus constants, in the normal case) is maximised numerically, typically with robust (Bollerslev–Wooldridge) standard errors. Diagnostics then check the standardised residuals $\hat{z}_t = \hat{\varepsilon}_t/\hat{\sigma}_t$ and their squares for remaining autocorrelation and ARCH effects. Multi-step variance forecasts follow by recursion: for GARCH(1,1), $E_t[\sigma_{t+h}^2]$ converges geometrically from σ_{t+1}^2 toward the unconditional variance $\omega/(1 - \alpha_1 - \beta_1)$ as h grows. The applications explain the model's ubiquity: value-at-risk and risk management, option pricing (where volatility is the key unobservable input), portfolio allocation, and the construction of time-varying hedge ratios — anywhere the second moment, not the first, is the object of interest. Table 1 contrasts the two models.

USMAN NOTES www.usmannotes.com

Table 1: ARCH versus GARCH

Feature	ARCH(q)	GARCH(p, q)
Variance depends on	Past q squared disturbances only	Past squared disturbances AND past conditional variances
Parameters for persistent volatility	Many (large q)	Few — GARCH(1,1) usually suffices
Persistence measure	$\Sigma\alpha_i < 1$ required for stationarity	$\Sigma\alpha_i + \Sigma\beta_j < 1$; the sum measures persistence
Equivalent form	—	ARCH(∞) with geometrically

		declining weights
Estimation	Maximum likelihood	Maximum likelihood

Extensions: Asymmetry and Beyond

Standard GARCH treats positive and negative shocks symmetrically, yet in equity markets bad news raises volatility more than good news of the same size (the leverage effect). The main asymmetric extensions are: EGARCH (Nelson, 1991), which models $\ln \sigma_t^2$ — guaranteeing a positive variance without parameter constraints — with a term distinguishing the sign of the shock; GJR-GARCH / threshold GARCH (Glosten–Jagannathan–Runkle, 1993), which adds $\alpha_2 \varepsilon_{t-1}^2 \cdot I(\varepsilon_{t-1} < 0)$ so negative shocks get extra weight; and APARCH, which nests several of these. Further branches include GARCH-M, where the conditional variance enters the mean equation as a risk premium; component and FIGARCH models for multi-scale and long-memory persistence; and multivariate GARCH (BEKK, DCC) for time-varying correlations between assets. All remain children of the same insight: model the conditional variance as a predictable, autoregressive process.

Conclusion

ARCH makes today's conditional variance a function of recent squared shocks, generating volatility clustering and fat tails while leaving the mean equation's residuals uncorrelated. GARCH adds lagged conditional variances, achieving parsimonious, highly persistent volatility dynamics — usually with the three-parameter GARCH(1,1) — and is estimated by maximum likelihood after ARCH effects are confirmed by the ARCH-LM test. Together they transformed the treatment of heteroscedasticity from a nuisance violating textbook assumptions into a

structured, forecastable feature of the data, with central applications in finance and risk management.

Question 4: What to Do When You Find a Problem of Autocorrelation (20 Marks)

What to do when you find a problem of Autocorrelation?

Introduction

Autocorrelation (serial correlation) means the disturbance terms of a regression are correlated across observations — in time series, typically that u_t is correlated with u_{t-1}. It is endemic in time-series work because economic variables are persistent and because misspecification often masquerades as correlated errors. Finding autocorrelation in a fitted model is not the end of the analysis but a fork in it: the correct response depends on diagnosing why it is present. This answer first recalls the consequences that make action necessary, then sets out a disciplined sequence: verify the detection, hunt for misspecification, and only then apply the appropriate remedy — transformation and GLS-type estimation where the error process is genuinely autoregressive, robust standard errors where only inference needs repair, or respecification where the 'autocorrelation' is really an omitted dynamic. Remedies are matched to causes, because the wrong remedy can do more harm than the disease.

Why Action Is Necessary: The Consequences

With autocorrelated disturbances (but exogenous regressors), OLS coefficient estimates remain unbiased and consistent but cease to be efficient — OLS is no longer BLUE — and, more dangerously for practice, the usual OLS standard errors are biased. With the positive autocorrelation typical of economic data, the conventional standard errors usually understate

the true ones, so t-statistics are inflated, confidence intervals too narrow, and insignificant variables appear significant: the researcher is misled with false confidence. R^2 is likewise overstated. Prediction intervals inherit the same defect. In the special case where a lagged dependent variable is among the regressors, autocorrelation additionally makes OLS biased and inconsistent — the one configuration in which the coefficients themselves cannot be trusted. Detection rests on the Durbin–Watson d statistic ($d \approx 2(1 - \hat{\rho})$; valid only without a lagged dependent variable and with an intercept), the Breusch–Godfrey LM test (regress residuals on the regressors plus lagged residuals; $nR^2 \sim \chi^2(p)$; valid generally and for higher orders), the Ljung–Box test on residuals, and the simple but eloquent plot of residuals against time, where runs and waves betray the correlation.

Step 1: Check Whether It Is Misspecification in Disguise

Residual autocorrelation is a symptom, and the first duty is diagnosis. What looks like serially correlated errors is frequently an omitted variable whose own persistence is dumped into the residuals, a wrong functional form (fitting a line to a curve leaves systematic runs of positive and negative residuals), omitted dynamics (the true model needs lagged Y or lagged X), or a structural break splitting the sample. The remedies at this stage are respecification, not correction of the errors: add the omitted variable or lags, transform the functional form (logs, quadratics), test for breaks (Chow tests), and re-run the diagnostics — Breusch–Godfrey after respecification, plus RESET for functional form. Only autocorrelation that survives a well-specified model should be treated as a property of the disturbance process itself. Textbook warnings are blunt on this point: mechanically 'correcting' autocorrelation that is really misspecification buries information the data are trying to deliver.

Step 2: If the Errors Are Genuinely Autoregressive — GLS and Feasible GLS

Suppose the disturbances follow a first-order autoregression $u_t = \rho u_{t-1} + \varepsilon_t$ with ε_t white noise and $|\rho| < 1$. If ρ were known, generalised least squares would restore efficiency by quasi-differencing the data: transform every variable as $Y^*_t = Y_t - \rho Y_{t-1}$, $X^*_t = X_t - \rho X_{t-1}$, which converts the model into one with white-noise errors, estimable by OLS on the transformed data. Since ρ is unknown, feasible GLS estimates it from the OLS residuals, $\hat{\rho} = \Sigma \hat{e}_t \hat{e}_{t-1} / \Sigma \hat{e}_{t-1}^2$ (equivalently $\hat{\rho} \approx 1 - d/2$ from the Durbin–Watson statistic), and iterates if desired. The named procedures differ in their treatment of the first observation:

- Cochrane–Orcutt: applies the quasi-difference to observations 2…n and drops the first observation; simple, and iterated to convergence it approximates maximum likelihood.
- Prais–Winsten: retains the first observation with the special transformation $\sqrt{(1 - \rho^2)} Y_1$ and $\sqrt{(1 - \rho^2)} X_1$, which is the theoretically correct GLS treatment and performs better in small samples.
- Hildreth–Lu: searches over a grid of ρ values and selects the one minimising the residual sum of squares of the transformed regression — a useful check on the estimated $\hat{\rho}$.
- Maximum likelihood: estimates ρ and the regression coefficients jointly; with modern software this and Prais–Winsten are the default choices. For higher-order error processes (AR(2), AR(p)) the same GLS logic applies with the corresponding multi-term quasi-differences, the order being chosen from the residual correlogram / Breusch–Godfrey tests at several lags.

Step 3: The Alternative Strategy — Keep OLS, Repair the Inference (HAC Standard Errors)

Often the coefficients from OLS are acceptable — OLS remains consistent under autocorrelation when there is no lagged dependent variable — and only the standard errors are wrong. The modern, increasingly default response is to keep the OLS point estimates and compute heteroscedasticity-and-autocorrelation-consistent (HAC) standard errors, above all the Newey–West estimator, which weights residual autocovariances at a chosen number of lags (bandwidth) with Bartlett weights that decline with lag length. Its advantages are robustness: it does not require committing to an exact AR(1) form for the errors, and it handles heteroscedasticity simultaneously. Its costs are efficiency (no gain in the precision of the coefficients themselves) and sensitivity to the bandwidth choice, which should grow slowly with sample size. The pragmatic rule many econometricians follow: respecify first; use HAC errors when the dynamics are a nuisance rather than the object of interest; use FGLS/ML when the error process is credibly known and efficiency matters, as in small samples.

Special and Related Cases

- Lagged dependent variable present: OLS is inconsistent under autocorrelation, so neither HAC errors nor simple corrections suffice; instrument the lagged dependent variable or estimate the dynamic model by maximum likelihood / IV, and test with Breusch–Godfrey or Durbin's h rather than Durbin–Watson.
- Panel and cross-section data: use cluster-robust standard errors (clustered by the unit or time period) to absorb arbitrary within-cluster correlation, the panel analogue of HAC.

- Seasonal autocorrelation: in quarterly or monthly data, check and model correlation at the seasonal lag (4 or 12) with seasonal dummies or seasonal error processes.
- What not to do: do not difference the data mechanically — first-differencing when ρ is small over-differences and induces a unit-root moving-average error; do not add a time trend as a ritual cure; and do not report Durbin–Watson alone as proof of a cure — always retest the transformed/corrected model's residuals with Breusch–Godfrey.

Conclusion

When autocorrelation is found: confirm it with appropriate tests (Breusch–Godfrey, residual plots, Durbin–Watson where valid); first interrogate the specification, since omitted variables, wrong functional form and missing dynamics commonly manufacture it; if genuine autoregressive errors remain, estimate by feasible GLS — Cochrane–Orcutt, Prais–Winsten or maximum likelihood — using ρ̂ from the residuals; alternatively retain OLS with Newey–West HAC standard errors when only inference needs repair; and finally re-diagnose the corrected model. The sequence — detect, diagnose, respecify, then correct — is what distinguishes handling autocorrelation from merely hiding it.

Question 5: Steps in Principal Component Analysis (PCA) (20 Marks)

Outline the steps you would follow in the Principal Component Analysis (PCA).

Introduction

Principal Component Analysis, introduced by Karl Pearson (1901) and developed by Harold Hotelling (1933), is the classic technique for simplifying a data set with many correlated variables. Its idea is elegant: replace the original p variables $X_1, \ldots, X_p$ — tangled, overlapping and hard to visualise — with a smaller number of new, mutually uncorrelated variables, the

principal components, each a linear combination of the originals, chosen in turn to capture the maximum possible remaining variance. The first component is the single direction along which the data vary most; the second is the direction of greatest remaining variation at right angles to the first, and so on. If the first few components capture most of the total variance, the data can be summarised, plotted, indexed and fed into further analysis (as principal component regression) with little genuine loss. This answer outlines, in order, the steps one would actually follow, with the decisions and checks each step requires.

Step 1: Define the Purpose and Assemble the Data Matrix

Begin by stating why PCA is being done — data reduction, construction of a composite index (for example a socio-economic status or financial-conditions index), visualisation, removing multicollinearity before regression, or noise reduction — because the purpose governs later choices such as how many components to retain. Assemble the n × p data matrix X (n observations, p variables), check for missing values and gross errors or outliers (PCA, being variance-based, is sensitive to outliers, which can seize a whole component), and confirm the variables are quantitative and meaningfully related: PCA only pays when variables are substantially correlated. A preliminary inspection of the correlation matrix is essential — if the variables are nearly uncorrelated, the components will merely reproduce the original variables and nothing is gained. Formal gatekeepers at this step are Bartlett's test of sphericity (tests whether the correlation matrix is an identity matrix; rejection is desired) and the Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy (values above 0.6, ideally above 0.8, indicate the correlations are compact enough for PCA).

Step 2: Standardise the Variables

Decide the basis of the analysis. If the variables are measured in different units or on very different scales (income in rupees alongside years of schooling), the large-variance variables would dominate the components purely through their units. Standard practice is therefore to standardise each variable to mean zero and unit variance, $z_{ij} = (x_{ij} - \bar{x}_j)/s_j$, which is equivalent to performing PCA on the correlation matrix. Use the covariance matrix (unstandardised) only when the variables share a common unit and their natural variances are substantively meaningful. Record the choice: component loadings and the variance shares depend on it.

Step 3: Compute the Correlation (or Covariance) Matrix

From the standardised data compute the $p \times p$ correlation matrix R (or the covariance matrix S in the unstandardised case). This symmetric matrix is the entire input of the analysis: PCA is the eigen-decomposition of R. Verify it is well behaved — determinant not effectively zero (extreme multicollinearity or an exactly redundant variable inflates and destabilises the small components) — and eyeball the pattern of correlations to anticipate how many strong dimensions the data contain.

Step 4: Extract the Eigenvalues and Eigenvectors

Solve the characteristic equation $|R - \lambda I| = 0$ for the p eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$, and find the corresponding eigenvectors $a_1, \ldots, a_p$. Each eigenvalue measures the variance carried by its component: with standardised variables the total variance equals p, so the proportion explained by component j is λ_j/p. The principal components themselves are the linear combinations $PC_j = a_j'z = a_{j1}z_1 + a_{j2}z_2 + \ldots + a_{j}_p\ z_p$, with coefficients from the eigenvector (usually rescaled so they are the loadings — correlations between component and variables —

obtained by multiplying eigenvector elements by $\sqrt{\lambda_j}$). By construction the components are mutually uncorrelated, and they are ordered: PC_1 captures the most variance, PC_2 the most of what remains, and so on. Table 1 gives an illustrative extraction for six standardised variables.

Table 1: Illustrative PCA extraction — six standardised variables (total variance = 6)

Component	Eigenvalue	% of variance	Cumulative %	Decision (Kaiser: $\lambda > 1$)
PC1	2.85	47.5	47.5	Retain
PC2	1.32	22.0	69.5	Retain
PC3	0.71	11.8	81.3	Drop ($\lambda < 1$)
PC4	0.48	8.0	89.3	Drop
PC5	0.34	5.7	95.0	Drop
PC6	0.30	5.0	100.0	Drop

USMAN NOTES www.usmannotes.com

Step 5: Decide How Many Components to Retain

Apply several criteria together rather than any single rule mechanically. (i) The Kaiser criterion: with correlation-matrix PCA, retain components with eigenvalue above 1, since a component should explain more than a single original variable. (ii) The scree test (Cattell): plot eigenvalues against component number (Figure 1) and retain those before the 'elbow' where the curve flattens into rubble (scree). (iii) The cumulative variance rule: retain enough components to reach a pre-set share, commonly 70–90% of total variance. (iv) Interpretability and parallel analysis: a component that cannot be interpreted, or whose eigenvalue barely exceeds what

random data of the same size would produce (Horn's parallel analysis), is a candidate for dropping. In the illustration, Kaiser and the scree elbow agree on two components carrying 69.5% of the variance; if the purpose demanded 80%, the third component would be added and the trade-off reported openly. (These eigenvalues are illustrative.)

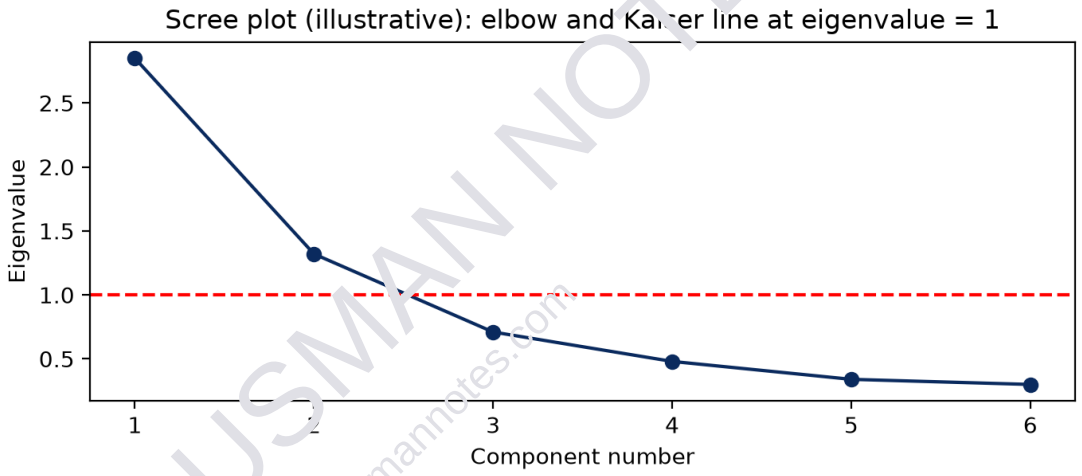


Figure 1: Scree plot for the illustrative extraction: retain the components before the elbow, cross-checked against the Kaiser line at eigenvalue 1.

Step 6: Interpret the Components from Their Loadings

Interpretation is the step that turns arithmetic into knowledge. Examine the loadings matrix: variables with high loadings (say |loading| > 0.5, or > 0.4 for a softer threshold) on a component define its meaning, and the pattern of signs shows contrasts. A component loading strongly on income, education, occupation and housing variables is naturally named a 'socio-economic status' dimension; one loading positively on some variables and negatively on others measures a trade-off or contrast between those groups. If interpretation is muddy, an orthogonal rotation (varimax) of the retained components may sharpen the loading pattern — noting that rotated components sacrifice the strict successive-maximum-variance property and

the exact variance ordering. Name components cautiously and report the loadings table in full so readers can judge the naming.

Step 7: Compute Component Scores and Use Them

Compute each observation's score on each retained component by applying the component coefficients to its standardised values, $score_{ij} = a_j'z_i$. The scores are the deliverable for most applications: plot observations on the PC_1–PC_2 plane to reveal clusters and outliers invisible in p dimensions; use PC_1 as a composite index when it is a 'general factor' (all loadings same sign); enter a few components instead of many collinear originals into a regression (principal component regression) to stabilise estimates; or feed scores into clustering and classification procedures. Validate the solution before relying on it: check the reproduced correlation matrix (original minus reproduced correlations — the residuals should be small), split the sample or bootstrap to confirm the loading pattern is stable, and state clearly what share of variance was discarded with the dropped components, since that information is genuinely lost.

Conclusion

The steps of PCA form a single disciplined chain: define the purpose and assemble clean, suitably correlated data; standardise unless a common scale justifies covariances; compute the correlation matrix and verify adequacy (Bartlett, KMO); extract eigenvalues and eigenvectors; retain components by the combined evidence of the Kaiser criterion, the scree plot, cumulative variance and interpretability; interpret through the loadings; and compute scores for use in plotting, indexing or regression, with validation of the reproduced correlations and stability. Followed in this order, PCA converts an unwieldy set of correlated variables into a few

uncorrelated, interpretable dimensions with a known, honestly reported cost in discarded variance.

USMAN NOTES
www.usmannotes.com