

VU LEARNING HUB

vulearninghub.xyz

HIGHLIGHTED HANDOUT

STAT404

Regression and Correlation

STAT404 by VU Learning HUB

Prepared for VU students • Free study resources

<https://vulearninghub.xyz>

COMPREHENSIVE STUDY HANDOUT

STAT404

Regression and Correlation Analysis

Lecture-wise study notes prepared for Virtual University of Pakistan students.

Covers all major topics with objectives, key definitions, summaries and exam pointers. 

9 study units

Table of Contents

Placeholder for table of contents

0

Unit 1: Introduction: Regression vs Correlation

Learning Objectives

- Objective 1:** Distinguish regression from correlation
- Objective 2:** Classify variables: dependent and independent
- Objective 3:** List regression applications

Regression and correlation both study relationships between variables, but ask different questions. Correlation measures the strength and direction of linear association symmetrically ($r_{xy} = r_{yx}$); regression estimates how a dependent variable changes with independent variables, asymmetrically, for prediction and explanation. Correlation is a single number; regression is a model.

Variables divide into dependent (response, y : what we explain or predict) and independent (predictors, x : what we use). In economics, consumption depends on income; in biostatistics, blood pressure on dosage. Correct classification matters: regressing y on x differs from regressing x on y , while correlation is symmetric.

Applications span demand forecasting, dose-response modeling, quality control, finance (CAPM's beta is a regression slope), and sports analytics. Very important: the course's first distinction, regression-asymmetric versus correlation-symmetric, recurs throughout; examiners test it directly and through numerical choices. 

Key Definitions

Regression is defined as modeling a dependent variable as a function of independent variables for prediction and explanation.

Correlation is defined as the symmetric measure of linear association strength and direction.

Dependent variable is defined as the response y being explained or predicted.

Key Points to Remember

- Correlation: symmetric strength/direction; regression: asymmetric predictive model.
- Dependent (y) vs independent (x); regressing y on $x \neq x$ on y .
- Applications: forecasting, dose-response, quality, finance (beta), analytics.
- The asymmetric/symmetric distinction recurs throughout.

Unit Summary

Correlation symmetrically measures association; regression asymmetrically models dependence for prediction, with correct variable classification essential.

Exam Pointer

Very important: distinguish regression from correlation precisely; classify variables in applications.

Unit 2: Scatterplots and the Correlation Coefficient

Learning Objectives

Objective 1: Interpret scatterplots

Objective 2: Compute Pearson's r

Objective 3: Explain correlation properties and pitfalls

Every analysis begins with a scatterplot: it reveals form (linear or not), direction, strength, and outliers. A correlation computed without plotting is dangerous; Anscombe's quartet shows identical r values for wildly different patterns. Nonlinearity, heteroscedasticity (fanning spread), and influential outliers all show visually first.

Pearson's $r = \text{Cov}(x,y)/(s_x s_y)$ ranges from -1 to 1: sign gives direction, magnitude strength. Computationally, $r = [n \sum xy - \sum x \sum y] / \sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}$. r is unitless and unaffected by linear rescaling; it measures linear association only.

Pitfalls: correlation is not causation (confounders, reverse causation); restricted range attenuates r ; outliers distort it; ecological correlations (group-level) may not hold individually. Very important: r^2 as proportion of variance explained is the single most interpreted quantity; state it whenever reporting r .

Key Definitions

Pearson's r is defined as $\text{Cov}(x,y)/(s_x s_y)$; the unitless linear correlation in $[-1,1]$.

Anscombe's quartet is defined as four datasets with identical r but different patterns; plot first.

Heteroscedasticity is defined as non-constant spread of y across x .

Key Points to Remember

- Plot first: form, direction, strength, outliers; Anscombe's warning.
- r : sign = direction, $|r|$ = strength; unitless; linear only.
- r^2 = proportion of y -variance explained by x .
- Pitfalls: causation, range restriction, outliers, ecological fallacy.

Unit Summary

Scatterplots diagnose relationships before Pearson's r quantifies linear association, with r^2 interpreted as variance explained and causation never assumed.

Exam Pointer

Very important: compute r from summary statistics and interpret r^2 ; list correlation pitfalls.

Unit 3: Simple Linear Regression: Fitting the Line

Learning Objectives

Objective 1: Derive least-squares estimators

Objective 2: Interpret slope and intercept

Objective 3: Make predictions with the regression line

The simple model $y = \beta_0 + \beta_1 x + \epsilon$ is fitted by least squares: minimizing sum of squared residuals. The normal equations yield $b_1 = r(s_y/s_x) = [n \sum xy - \sum x \sum y] / [n \sum x^2 - (\sum x)^2]$.

- $(\sum x)^2$] and $b_0 = \bar{y} - b_1 \bar{x}$. The fitted line always passes through $(\bar{x}, \bar{y})$; residuals sum to zero.

Interpretation: b_1 is the expected change in y per unit change in x ; b_0 is the expected y when $x = 0$ (meaningful only if $x = 0$ is in range). Units matter: if y is rupees and x kilograms, b_1 is rupees per kilogram. The slope's sign matches r 's sign.

Prediction plugs x into the fitted line, but only within the observed range; extrapolation beyond it is unreliable (the relationship may change). Very important: least-squares numericals are guaranteed exam material; show the sums table, compute b_1 then b_0 , and warn against extrapolation in the interpretation.

Key Definitions

Least squares is defined as fitting by minimizing the sum of squared residuals.

Residual is defined as the observed minus fitted value; $e_i = y_i - \hat{y}_i$.

Extrapolation is defined as prediction outside the observed x -range; unreliable.

Key Points to Remember

- $b_1 = r(s_y/s_x)$; $b_0 = \bar{y} - b_1 \bar{x}$; line passes through $(\bar{x}, \bar{y})$.
- Slope = expected y -change per x -unit; intercept = y at $x=0$ (if meaningful).
- Residuals sum to zero; prediction only within range.
- Show the sums table in numericals; warn against extrapolation.

Unit Summary

Least squares fits $y = b_0 + b_1 x$ through the mean point, with the slope as the marginal effect used for in-range prediction.

Exam Pointer

Very important: fit a regression line from raw data showing all sums; interpret slope with units.

Unit 4: Assessing Fit: R-squared, Residuals and Inference

Learning Objectives

Objective 1: Decompose variation: SST, SSR, SSE

Objective 2: Use residual plots for diagnostics

Objective 3: Test the slope's significance

Total variation splits: $SST = SSR + SSE$, where SSR is explained by the regression and SSE is residual. $R^2 = SSR/SST = r^2$ is the proportion explained; adjusted R^2 penalizes added predictors (previewing multiple regression). High R^2 with patterned residuals still signals a wrong model.

Residual plots (residuals vs fitted) diagnose: random scatter means assumptions hold; curves mean nonlinearity (transform or add terms); fanning means heteroscedasticity (weighted least squares or transforms); outliers stand out. Normality of residuals is checked via Q-Q plots for inference validity.

Inference on the slope: $t = b_1/SE(b_1)$ tests $H_0: \beta_1 = 0$ (no linear relationship); confidence intervals $b_1 \pm t^*SE(b_1)$ quantify the effect. Assumptions: linearity, independent errors, constant variance, normal errors (LINE). Very important: residual-plot diagnosis is the most applied skill; examiners show a plot and ask what is wrong.

Key Definitions

R-squared is defined as SSR/SST ; the proportion of y-variation explained by the model.

Residual plot is defined as residuals vs fitted values; the primary diagnostic display.

Standard error of the slope is defined as $SE(b_1)$; the slope estimate's sampling variability.

Key Points to Remember

- $SST = SSR + SSE$; $R^2 = SSR/SST = r^2$ (simple case).
- Residual plots: random = good; curve = nonlinearity; fan = heteroscedasticity.
- t-test on slope: $H_0 \beta_1=0$; CI quantifies effect.
- Assumptions: linearity, independence, constant variance, normality (LINE).

Unit Summary

Fit is assessed by variance decomposition (R^2), residual diagnostics, and slope inference under the LINE assumptions.

Exam Pointer

Very important: diagnose a residual plot and test slope significance; state the LINE assumptions.

Unit 5: Multiple Regression

Learning Objectives

- Objective 1:** Specify and interpret multiple regression
- Objective 2:** Explain partial slopes and adjusted R-squared
- Objective 3:** Detect multicollinearity

Multiple regression $y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \epsilon$ models several predictors. Each partial slope β_j is the effect of x_j holding others constant, the ceteris paribus interpretation that makes multiple regression powerful for observational data. Matrix form $y = X\beta + \epsilon$ gives $b = (X'X)^{-1}X'y$.

R^2 always rises when predictors are added, so adjusted $R^2 = 1 - (1-R^2)(n-1)/(n-k-1)$ penalizes complexity for model comparison. Overall F-tests check whether any predictor matters; individual t-tests check each one. A significant F with no significant t hints at multicollinearity.

Multicollinearity (correlated predictors) inflates standard errors, making slopes unstable and signs unreliable; variance inflation factors ($VIF > 10$ as a rule of thumb) diagnose it. Remedies: drop redundant predictors, combine them, or use ridge regression. Very important: the ceteris paribus interpretation of partial slopes is the conceptual core; multicollinearity diagnosis is the applied core.

Key Definitions

Partial slope is defined as a predictor's effect holding other predictors constant.

Adjusted R-squared is defined as R^2 penalized for the number of predictors; for model comparison.

Multicollinearity is defined as correlated predictors inflating standard errors and destabilizing slopes.

Key Points to Remember

- Partial slopes: ceteris paribus effects; matrix solution $b = (X'X)^{-1}X'y$.
- Adjusted R^2 for comparing models; F overall vs t individual tests.
- Significant F + insignificant t's signals multicollinearity.
- VIF diagnoses; remedies: drop, combine, or ridge.

Unit Summary

Multiple regression estimates ceteris paribus effects, compared via adjusted R^2 and guarded against multicollinearity's unstable slopes.

Exam Pointer

Very important: interpret a partial slope ceteris paribus; diagnose multicollinearity from F/t patterns.

Unit 6: Dummy Variables and Model Specification

Learning Objectives

- Objective 1:** Use dummy variables for categories
- Objective 2:** Interpret interaction terms
- Objective 3:** Address specification errors

Categorical predictors enter via dummy (0/1) variables: for c categories use $c-1$ dummies, with the omitted category as reference. Coefficients measure differences from the reference group. The dummy trap (including all c dummies plus intercept) creates perfect multicollinearity; always omit one.

Interactions let effects differ by group: $y = b_0 + b_1 x + b_2 D + b_3 (x \cdot D)$ gives different slopes and intercepts per group. Testing b_3 checks whether the x -effect differs across groups, e.g. whether a training program's effect differs by gender. Center variables to reduce nonessential collinearity in interactions.

Specification errors: omitting relevant variables biases slopes (omitted-variable bias, direction depending on correlations); including irrelevant ones merely costs efficiency. Functional-form misspecification (fitting linear when curved) is caught by residual plots. Very important: dummy-trap avoidance and interaction interpretation are standard exam items.

Key Definitions

Dummy variable is defined as a 0/1 variable encoding category membership.

Dummy trap is defined as perfect multicollinearity from including all category dummies with an intercept.

Interaction term is defined as a product term letting one variable's effect depend on another.

Key Points to Remember

- c categories $\rightarrow c-1$ dummies; coefficients = differences from reference.
- Dummy trap: never all c dummies + intercept.
- Interactions: different slopes per group; test the product term.
- Omitted variables bias; irrelevant ones cost efficiency; check form via residuals.

Unit Summary

Dummies encode categories against a reference, interactions allow group-specific effects, and correct specification avoids bias.

Exam Pointer

Very important: set up dummies avoiding the trap; interpret an interaction coefficient.

Unit 7: Nonlinear Regression and Transformations

Learning Objectives

Objective 1: Linearize nonlinear relationships

Objective 2: Fit polynomial regression

Objective 3: Choose between transformation and nonlinear fitting

Many relationships are nonlinear but linearizable. Exponential growth $y = a e^{bx}$ becomes linear via $\ln y = \ln a + bx$; power laws $y = a x^b$ via $\ln-\ln$; reciprocal and logit transforms handle other forms. Fit the transformed model by least squares, then back-transform predictions (noting bias corrections for expectations).

Polynomial regression adds x^2 , x^3 terms for curves: quadratic for U-shapes, cubic for S-shapes. It stays linear in parameters, so least squares applies directly. Danger: high-degree polynomials overfit, oscillating wildly; keep degree low and theory-guided. Orthogonal polynomials reduce collinearity among powers.

Choice: transform when theory suggests the form (economics: log-linear models give elasticities directly); fit polynomials for smooth unknown curves; use genuinely nonlinear least squares (iterative) when neither linearizes. Very important: log-linear interpretation (b as proportional change) is heavily used in economics; practice it.

Key Definitions

Linearization is defined as transforming variables so a nonlinear model becomes linear in parameters.

Polynomial regression is defined as adding powers of x as predictors; linear in parameters.

Overfitting is defined as fitting noise with excessive complexity; poor out-of-sample prediction.

Key Points to Remember

- Log transforms: exponential -> log-linear; power -> log-log (elasticities).
- Polynomials for smooth curves; low degree, theory-guided; overfitting danger.
- Back-transform predictions carefully (bias on re-transformation).
- Economics: log-linear slopes are proportional effects/elasticities.

Unit Summary

Nonlinearity is handled by linearizing transforms or polynomial terms, chosen by theory, with overfitting guarded against.

Exam Pointer

Very important: linearize an exponential model and interpret its slope; explain overfitting with polynomials.

Unit 8: Regression Diagnostics and Remedies

Learning Objectives

- Objective 1:** Run a full diagnostic checklist
- Objective 2:** Handle heteroscedasticity and autocorrelation
- Objective 3:** Manage outliers and influential points

A complete diagnostic pass checks: (1) linearity via residual plots, (2) constant variance (homoscedasticity), (3) normality of residuals (Q-Q plot), (4) independence (Durbin-Watson for time series), (5) multicollinearity (VIF), (6) outliers and influence (leverage, Cook's distance). Each violation has a remedy; diagnostics without remedies is half the skill.

Heteroscedasticity remedies: weighted least squares, variance-stabilizing transforms (log y), or robust standard errors. Autocorrelation (time-series errors correlated) is tested by Durbin-Watson and treated with AR models or Newey-West errors. Ignoring it understates uncertainty.

Outliers are unusual y; leverage points unusual x; influential points change the fit (Cook's distance combines both). Investigate first (data error? special case?), then consider robust regression rather than silent deletion. Very important: the diagnostic-remedy pairing is the applied capstone; examiners present symptoms and expect both diagnosis and cure.

Key Definitions

Homoscedasticity is defined as constant error variance across x; its absence is heteroscedasticity.

Cook's distance is defined as a measure combining leverage and residual to flag influential points.

Durbin-Watson is defined as a test statistic for first-order autocorrelation in residuals.

Key Points to Remember

- Checklist: linearity, variance, normality, independence, collinearity, influence.
- Heteroscedasticity: WLS, transforms, robust SEs; autocorrelation: DW test, AR/Newey-West.

- Outlier (y) vs leverage (x) vs influential (changes fit); investigate before deleting.
- Every diagnosis pairs with a remedy.

Unit Summary

Systematic diagnostics check each regression assumption, pairing every violation (heteroscedasticity, autocorrelation, influence) with its remedy.

Exam Pointer

Very important: given symptoms, diagnose and prescribe the remedy; explain Cook's distance components.

Unit 9: Rank Correlation and Applied Regression

Learning Objectives

- Objective 1:** Compute Spearman and Kendall correlations
- Objective 2:** Choose Pearson vs rank methods
- Objective 3:** Apply regression across disciplines

When data are ordinal or assumptions fail, rank correlations help. Spearman's rho applies Pearson's formula to ranks, measuring monotonic association; Kendall's tau counts concordant versus discordant pairs, with better small-sample properties. Both resist outliers that would distort Pearson's r .

Choice rules: Pearson for linear relationships in continuous, roughly normal data; Spearman/Kendall for ordinal data, monotonic-but-nonlinear patterns, or outlier-prone data. Report which was used and why; mixing them up is a common methodological error in student projects.

Applications integrate the course: economics (demand estimation, elasticities), business (sales forecasting), biostatistics (dose-response), agriculture (yield models), social sciences (attitude scales via ordinal methods). The workflow is constant: plot, choose method, fit, diagnose, interpret cautiously. Very important: the Pearson-vs-rank decision rule and one full applied walkthrough are the capstone exam expectations.

Key Definitions

Spearman's rho is defined as Pearson correlation computed on ranks; monotonic association.

Kendall's tau is defined as concordant-minus-discordant pairs over total pairs; robust small-sample rank measure.

Monotonic is defined as consistently increasing or decreasing, not necessarily linear.

Key Points to Remember

- Spearman: ranks, monotonic; Kendall: pair concordance, small samples.
- Pearson: linear, continuous, near-normal; ranks: ordinal/nonlinear/outliers.
- Always state which correlation and why.
- Universal workflow: plot -> choose -> fit -> diagnose -> interpret.

Unit Summary

Rank correlations handle ordinal and non-ideal data; method choice follows data type, applied through the universal plot-fit-diagnose-interpret workflow.

Exam Pointer

Very important: choose Pearson vs Spearman for a scenario with justification; walk through a complete applied analysis.