

VU LEARNING HUB

vulearninghub.xyz

HANDOUT

STA5204

Regression Analysis

STA5204 by VU Learning HUB

Prepared for VU students • Free study resources

<https://vulearninghub.xyz>

STA5204

Regression Analysis

Lecture-Wise Study Handout
Virtual University of Pakistan

11 lectures • prepared for exam study

Table of Contents

1. Introduction: Regression versus Correlation
2. Least Squares Estimation
3. Inference: Assumptions, BLUE, Tests and Intervals
4. ANOVA and Regression Diagnostics
5. Multiple Regression: Matrix Formulation
6. Multicollinearity, Ridge and Dummy Variables
7. Model Building and Selection
8. Nonlinear, Logistic and Generalized Linear Models
9. Time Series and Heteroscedasticity
10. Causality, Endogeneity and Measurement
11. Software Workflow and Reporting

About This Course

Regression analysis studies how a response variable depends on explanatory variables: how house prices depend on size and location, how crop yield depends on fertilizer and rain, how sales depend on advertising. It is the most widely used statistical modelling tool in science, business and policy.

This handout develops regression from simple linear regression through least squares, inference, diagnostics, multiple regression in matrix form, model building and extensions — with the key formulas explained in words, worked interpretations, and exam pointers in every lecture.

Lecture 1: Introduction: Regression versus Correlation

Learning Objectives

- Define regression
- Interpret the simple linear model

A regression model relates a dependent (response) variable Y to independent (explanatory) variables X : $Y = f(X) + \text{error}$. Regression (Galton's term for the tendency of extreme values to revert toward the mean) differs from correlation: correlation measures symmetric association, while regression models directional dependence — X explaining Y — and enables prediction.

The simple linear model is $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$: β_0 (intercept) is Y when $X = 0$, β_1 (slope) is the change in Y per unit X , and ε_i captures everything else — omitted variables, measurement error, inherent randomness. The error term is what makes it statistics rather than algebra.

Applications are everywhere: economics (demand curves), agriculture (yield vs fertilizer), medicine (dose–response), business (sales vs advertising). The exam's first distinction is always: regression models conditional expectation $E[Y|X]$ — the average Y at given X — not deterministic law.

Key Definitions

Regression model is defined as a model of the conditional expectation of Y given X : $Y = \beta_0 + \beta_1 X + \text{error}$

Key Points — very important for exams

Regression models Y 's dependence on X (directional, predictive); correlation measures symmetric association.

Simple linear model: $Y = \beta_0 + \beta_1 X + \varepsilon$ — intercept, slope, error.

Regression estimates $E[Y|X]$: average Y at given X , not exact law.

Review Questions

- ? Why is regression directional while correlation is symmetric?
- ? What does the error term represent?

Section Summary

Regression models how Y depends on X — the average relationship plus random error.

Exam Pointers

- Distinguish regression from correlation; interpret β_0 , β_1 and ε .

Lecture 2: Least Squares Estimation

Learning Objectives

- Compute OLS estimates
- Interpret R^2

Least squares chooses β_0 , β_1 to minimize the sum of squared residuals $\sum (Y_i - \hat{Y}_i)^2$ — penalizing large errors quadratically, which yields clean mathematics (differentiable) and the maximum-likelihood solution under normal errors. The normal equations give closed forms: $\hat{\beta}_1 = S_{xy}/S_{xx}$ (covariance over variance of X), $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ — the fitted line always passes through $(\bar{X}, \bar{Y})$.

Why squares, not absolute values? Squaring is differentiable (unique solution), emphasizes large errors (robust to none, sensitive to outliers — the trade-off), and connects to the normal distribution. Residuals $e_i = Y_i - \hat{Y}_i$ estimate the errors; their sum is zero and they are uncorrelated with X by construction.

Goodness of fit: $R^2 = 1 - SSE/SST$ is the fraction of Y's variation explained — 0.85 means 85% explained. But R^2 always rises when variables are added, so adjusted R^2 penalizes complexity. Very important: high R^2 never proves causation — it measures fit, not truth.

Key Definitions

Ordinary least squares (OLS) is defined as estimation minimizing $\sum (Y_i - \hat{Y}_i)^2$, yielding $\hat{\beta}_1 = S_{xy}/S_{xx}$, $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

R-squared is defined as the fraction of Y-variation explained by the model: $1 - SSE/SST$

Key Points — very important for exams

OLS minimizes squared residuals: $\hat{\beta}_1 = S_{xy}/S_{xx}$, $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$; line passes through $(\bar{X}, \bar{Y})$.

Squares: differentiable, MLE under normality; sensitive to outliers.

R^2 = explained fraction; adjusted R^2 penalizes complexity; fit $\neq$ causation.

Review Questions

- ? Why minimize squares rather than absolute deviations?
- ? If $R^2 = 0.92$, what can and cannot be concluded?

Section Summary

Least squares draws the best-fit line through the means — and R^2 scores its fit.

Exam Pointers

- Derive the OLS estimators; interpret R^2 and its limits.

Lecture 3: Inference: Assumptions, BLUE, Tests and Intervals

Learning Objectives

- State regression assumptions

- Test slopes and build intervals

Inference needs assumptions — the Gauss–Markov set: linearity, independent errors with mean zero, constant variance (homoscedasticity), and (for exact tests) normality. Under these, OLS is BLUE — the best linear unbiased estimator: minimum variance among linear unbiased estimators. Violations have named consequences, covered in diagnostics.

The sampling distributions: β_1 -hat is normal with standard error $SE = s/\sqrt{S_{xx}}$ — precision grows with X-spread and sample size. The t-test $(\beta_1\text{-hat} - \beta_1)/SE$ tests hypotheses like $H_0: \beta_1 = 0$ (no relationship); confidence intervals $\beta_1\text{-hat} \pm t^* \cdot SE$ quantify slope uncertainty. A slope significant at 5% means the data reject no-relationship.

Prediction has two intervals: confidence interval for the mean response at X (narrower) and prediction interval for an individual new Y (wider — adds individual variation). Confusing them is a classic error: predicting one house's price is harder than predicting average price at that size.

Key Definitions

BLUE (Gauss–Markov theorem) is defined as under Gauss–Markov assumptions, OLS has minimum variance among linear unbiased estimators

Key Points — very important for exams

Assumptions: linearity, independent zero-mean errors, homoscedasticity (+ normality for tests).

Gauss–Markov: OLS is BLUE; t-tests and CIs use $SE = s/\sqrt{S_{xx}}$.

Confidence interval (mean response) vs prediction interval (individual) — the latter wider.

Review Questions

- ? What does BLUE guarantee?
- ? Why is the prediction interval wider than the confidence interval?

Section Summary

Under Gauss–Markov, OLS is optimal — and t-tests plus intervals turn estimates into inference.

Exam Pointers

- State the Gauss–Markov assumptions; test $H_0: \beta_1 = 0$ and interpret the intervals.

Lecture 4: ANOVA and Regression Diagnostics

Learning Objectives

- Use ANOVA tables
- Diagnose assumption violations

ANOVA partitions total variation: $SST = SSR + SSE$ — total = explained + unexplained. The F-test (MSR/MSE) tests $H_0: \beta_1 = 0$ globally; in simple regression $F = t^2$, unifying the tests. The ANOVA table

organizes it: source, df, SS, MS, F — learn to read and complete one; exams love half-filled tables.

Diagnostics check assumptions via residuals. Plot residuals vs fitted: random scatter (good), funnel shape (heteroscedasticity), curve (non-linearity). Normal Q–Q plot checks normality; residuals vs X or time reveal missed structure. Assumptions are checked, not assumed — very important exam ethic.

Remedies follow diagnoses: transform Y or X (log for multiplicative effects), weighted least squares for heteroscedasticity, add polynomial terms for curvature. Influential points are flagged by leverage and Cook's distance — investigate, don't automatically delete.

Key Definitions

Regression ANOVA is defined as the decomposition $SST = SSR + SSE$ tested by $F = MSR/MSE$

Key Points — very important for exams

ANOVA: $SST = SSR + SSE$; $F = MSR/MSE$ tests overall significance; in simple regression $F = t^2$.

Residual plots diagnose: funnel = heteroscedasticity, curve = non-linearity, Q–Q = normality.

Remedies: transformations, weighted LS, polynomials; check influence via Cook's distance.

Review Questions

? What pattern signals heteroscedasticity?

? How do you remedy curved residual plots?

Section Summary

Partition variation with ANOVA; interrogate residuals — then fix what they reveal.

Exam Pointers

→ Complete an ANOVA table; diagnose heteroscedasticity and non-linearity from residual plots.

Lecture 5: Multiple Regression: Matrix Formulation

Learning Objectives

- Interpret multiple regression
- Use the matrix solution

Multiple regression $Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon$ models joint effects; each coefficient is the effect holding others constant — the ceteris paribus interpretation that makes multiple regression powerful and subtle. Omitted variable bias strikes when a correlated, relevant variable is left out: the included coefficients absorb its effect.

Matrix form $Y = X\beta + \varepsilon$ gives the compact solution $\hat{\beta} = (X'X)^{-1} X'Y$ — memorize this; it is the most quoted formula in the course. Geometrically, ■ is the projection of Y onto the column space of X; residuals are the orthogonal remainder. The hat matrix $H = X(X'X)^{-1}X'$ puts the hat on Y.

Overall F-tests the model; individual t-tests assess each variable — but with correlated predictors, t-tests can all be non-significant while F is significant (multicollinearity's signature, next lecture). Partial F-tests compare nested models: does adding X3, X4 significantly improve fit?

Key Definitions

Normal equations (matrix form) is defined as the OLS solution $\hat{\beta} = (X'X)^{-1} X'Y$

Key Points — very important for exams

Each β_j : effect holding others constant; omitting correlated variables biases estimates.

$\hat{\beta} = (X'X)^{-1}X'Y$ — projection of Y onto X's column space.

F tests the model; partial F compares nested models; t-tests assess individual variables.

Review Questions

? What is omitted variable bias?

? When would F be significant but no t-test is?

Section Summary

Multiple regression isolates ceteris-paribus effects through the projection $\hat{\beta} = (X'X)^{-1}X'Y$.

Exam Pointers

→ Interpret multiple coefficients ceteris paribus; state and interpret the matrix solution.

Lecture 6: Multicollinearity, Ridge and Dummy Variables

Learning Objectives

- Detect multicollinearity
- Use ridge and dummies

Multicollinearity — predictors correlated with each other — inflates standard errors: coefficients become unstable, t-tests fail, signs flip, yet R^2 stays high. Diagnose via correlation matrices and variance inflation factors ($VIF > 10$ signals trouble); the condition number summarizes $X'X$ ill-conditioning.

Remedies: drop redundant variables, combine them (indices, principal components), collect more data, or use ridge regression — which adds a penalty $\lambda||\beta||^2$, shrinking coefficients and stabilizing estimates at the cost of slight bias. Ridge's bias–variance trade-off is its exam moral: a little bias buys much precision.

Qualitative predictors enter via dummy variables: a binary D (0/1) shifts the intercept; interaction terms $D \cdot X$ allow different slopes per group — e.g. separate salary equations by gender. The dummy variable trap (including all categories plus intercept) creates perfect collinearity — always omit one category.

Key Definitions

Variance inflation factor (VIF) is defined as the factor by which a coefficient's variance is inflated by predictor collinearity

Key Points — very important for exams

Multicollinearity: correlated predictors → inflated SEs, unstable signs; diagnose with VIF.

Ridge: penalized fit shrinks coefficients — bias for precision.

Dummies encode categories; interactions allow group-specific slopes; avoid the dummy trap.

Review Questions

? Why does multicollinearity inflate standard errors?

? How do you model different slopes for two groups?

Section Summary

Correlated predictors destabilize estimates — diagnose with VIF, treat with ridge or respecification.

Exam Pointers

→ Diagnose multicollinearity and explain ridge regression and dummy variables.

Lecture 7: Model Building and Selection

Learning Objectives

- Select regression models
- Validate honestly

Model building chooses among candidate predictors. Criteria penalize complexity: adjusted R^2 , AIC, BIC (BIC penalizes harder — prefers simpler models), Mallows' Cp. Stepwise procedures (forward, backward, both) automate search but overfit and invalidate p-values — use for exploration, not inference.

Validation is honest: split data (train/validate/test) or cross-validate; the model winning on validation — not training — data is chosen. Best subsets with CV is the gold standard for small p; regularization paths (lasso) handle large p. Very important: selection uncertainty means reported p-values after stepwise are optimistic.

Practical workflow: start with subject-matter variables, check diagnostics, compare a few principled models by CV/AIC, validate on held-out data, and report the process transparently. The exam rewards this workflow over any single criterion.

Key Definitions

Information criteria is defined as penalized fit criteria (AIC/BIC) trading goodness-of-fit against parameter count

Key Points — very important for exams

Criteria: adjusted R^2 , AIC, BIC, Cp — penalize complexity; BIC prefers simpler.

Stepwise explores but invalidates inference; validate on held-out data or CV.

Workflow: theory → diagnostics → compare few models → validate → report transparently.

Review Questions

- ? Why are post-stepwise p-values misleading?
- ? When is BIC preferred to AIC?

Section Summary

Build models by penalized criteria and honest validation — not by training-data trophy hunting.

Exam Pointers

→ Compare model selection criteria and explain honest validation.

Lecture 8: Nonlinear, Logistic and Generalized Linear Models

Learning Objectives

- Fit curved relationships
- Model binary and count data

Polynomial regression fits curves with X , X^2 , ... terms — linear in parameters, so OLS applies; but high-degree polynomials wiggle wildly (Runge-like), so prefer low degrees or splines. Nonlinear models (exponential growth, logistic curves) need nonlinear least squares — iterative, sensitive to starting values.

Logistic regression handles binary Y : it models log-odds as linear, $\text{logit}(p) = X\beta$, fit by maximum likelihood. Coefficients are odds ratios — e^{β} is the multiplicative odds change per unit X . Classification tables, ROC and AUC evaluate it — linking back to the classification lecture's metrics.

Count data (Poisson regression), ordered categories (ordinal logit) and censored data (Tobit) extend the family: generalized linear models unify them — linear predictor + link function + exponential-family distribution. The exam theme: match the model to the response type.

Key Definitions

Logistic regression is defined as regression for binary outcomes modelling log-odds linearly, fit by maximum likelihood

Key Points — very important for exams

Polynomials: linear in parameters but wiggle — prefer low degree/splines.

Logistic: $\text{logit}(p) = X\beta$; e^{β} = odds ratio; evaluated by ROC/AUC.

GLMs unify: linear predictor + link + distribution — match model to response type.

Review Questions

- ? Interpret $e^{\beta} = 2.5$ in a logistic model.
- ? Why do high-degree polynomials misbehave?

Section Summary

Beyond the straight line: polynomials for curves, logistic for yes/no, GLMs for everything.

Exam Pointers

- Explain polynomial pitfalls, logistic odds ratios, and the GLM idea.

Lecture 9: Time Series and Heteroscedasticity

Learning Objectives

- Handle time-series regression
- Fix heteroscedasticity

Time-ordered data break independence: autocorrelation (today's error correlates with yesterday's) invalidates standard errors. The Durbin–Watson statistic tests AR(1) autocorrelation; plots of residuals vs time reveal it. Remedies: include lagged variables, Cochrane–Orcutt transformation, or Newey–West robust standard errors.

Time series regression models trend and seasonality explicitly: dummy variables for seasons, trend terms, and ARIMA errors. Spurious regression warns: two trending series correlate without any connection — always test for unit roots (Dickey–Fuller) and cointegration before trusting time-series regressions.

Heteroscedasticity (non-constant variance) is tested by Breusch–Pagan/White tests and visualized by funnel plots; weighted least squares or robust standard errors fix inference. The diagnostic ethic from simple regression scales up: every assumption gets a test and a plot.

Key Definitions

Autocorrelation is defined as correlation of regression errors across time, tested by Durbin–Watson

Key Points — very important for exams

Autocorrelation: DW test, residual-vs-time plots; fix with lags, Cochrane–Orcutt, robust SEs.

Spurious regression: trending series correlate spuriously — test unit roots/cointegration.

Heteroscedasticity: BP/White tests, funnel plots; weighted LS or robust SEs.

Review Questions

- ? Why is regressing two random walks dangerous?

? How do robust standard errors help?

Section Summary

Time breaks independence and variance breaks constancy — test both, then fix.

Exam Pointers

→ Detect and remedy autocorrelation and heteroscedasticity; warn against spurious regression.

Lecture 10: Causality, Endogeneity and Measurement

Learning Objectives

- Distinguish correlation from causation
- Address endogeneity

Regression is only as credible as its design. Randomized experiments identify causal effects: random assignment balances confounders, letting the treatment coefficient be causal. Observational data need extra care — control for confounders, but controlling for colliders or mediators biases estimates (the Table 2 fallacy: interpreting every coefficient causally).

Endogeneity (X correlated with error — omitted variables, simultaneity, measurement error) breaks OLS consistency; instrumental variables (correlated with X, affecting Y only through X) restore it — two-stage least squares. Regression discontinuity and difference-in-differences exploit quasi-experimental structure — the modern causal toolkit.

Measurement error in X attenuates slopes toward zero; in Y it merely adds noise. Missing data: complete-case analysis biases unless missing completely at random; multiple imputation is the principled fix. Very important: statistics can adjust for confounders, never for bad design.

Key Definitions

Instrumental variable is defined as a variable correlated with X but affecting Y only through X, enabling causal estimation

Key Points — very important for exams

Experiments identify causality; observational controls need causal graphs (avoid colliders).

Endogeneity breaks OLS; instruments (2SLS), RD and diff-in-diff restore credibility.

Measurement error in X attenuates; handle missing data by multiple imputation.

Review Questions

? What is the Table 2 fallacy?

? How does an instrument solve omitted-variable bias?

Section Summary

Regression describes; design plus instruments let it explain causally.

Exam Pointers

→ Explain endogeneity, instruments, and why design beats adjustment.

Lecture 11: Software Workflow and Reporting

Learning Objectives

- Run regression analyses
- Report results professionally

Statistical software implements the theory: R (lm, glm, diagnostics plots), Python (statsmodels, scikit-learn), SPSS/Stata menus. The workflow: import → explore (scatterplots, correlations) → fit → diagnose (residual plots, VIF, influence) → refine → validate → report with tables (coefficients, SEs, p-values, CIs) and plots.

Reporting standards: state the model equation, assumptions checked, estimation method, and interpret magnitudes with uncertainty — not just stars. Reproducibility (scripts, seeds, versions) is now required by journals; a regression nobody can rerun is a rumour.

Common software traps: categorical variables silently converted to numeric codes, default reference levels changing dummy interpretation, and stepwise p-values reported as valid. The analyst's judgment — which the course has trained — remains irreplaceable.

Key Definitions

Reproducibility is defined as the practice of sharing code, data and seeds so analyses can be independently rerun

Key Points — very important for exams

Workflow: explore → fit → diagnose → refine → validate → report (tables + plots).

Report magnitudes with uncertainty, assumptions checked, methods reproducible.

Traps: miscoded categoricals, reference levels, stepwise p-values — judgment required.

Review Questions

? What belongs in a regression report table?

? Why is reproducibility required?

Section Summary

Software computes; the analyst thinks — workflow, diagnostics and honest reporting.

Exam Pointers

→ Outline the applied regression workflow and reporting standards.

Glossary of Key Terms

Regression model — a model of the conditional expectation of Y given X : $Y = \beta_0 + \beta_1 X + \text{error}$

Ordinary least squares (OLS) — estimation minimizing $\sum (Y_i - \hat{Y}_i)^2$, yielding $\beta_1 = S_{xy}/S_{xx}$, $\beta_0 = \bar{Y} - \beta_1 \bar{X}$

R-squared — the fraction of Y -variation explained by the model: $1 - \text{SSE}/\text{SST}$

BLUE (Gauss–Markov theorem) — under Gauss–Markov assumptions, OLS has minimum variance among linear unbiased estimators

Regression ANOVA — the decomposition $\text{SST} = \text{SSR} + \text{SSE}$ tested by $F = \text{MSR}/\text{MSE}$

Normal equations (matrix form) — the OLS solution $\hat{\beta} = (X'X)^{-1} X'Y$

Variance inflation factor (VIF) — the factor by which a coefficient's variance is inflated by predictor collinearity

Information criteria — penalized fit criteria (AIC/BIC) trading goodness-of-fit against parameter count

Logistic regression — regression for binary outcomes modelling log-odds linearly, fit by maximum likelihood

Autocorrelation — correlation of regression errors across time, tested by Durbin–Watson

Instrumental variable — a variable correlated with X but affecting Y only through X , enabling causal estimation

Reproducibility — the practice of sharing code, data and seeds so analyses can be independently rerun