DegreeLive

Bachelor of Commerce (B2001) | Level 1 • Part A Core | Monash University

Business and economic statistics

Course Code: ETC1000 • Module 4 Notes

Module 4: Correlation, Simple and Multiple Linear Regression Analysis and Business Forecasting

4.1 Bivariate Relationships, Covariance, and Pearson Correlation

Business analysts and economists routinely investigate how one numerical variable moves in tandem with another—for example, how retail revenue responds to digital advertising spend, how mortgage arrears co-move with RBA cash-rate increases, or how residential electricity demand varies with ambient temperature. Quantifying bivariate linear association begins with visual inspection via a scatter plot, followed by covariance and correlation metrics.

Sample Covariance and Pearson Correlation Coefficient (r)

Given n paired numerical observations $(x_1, y_1), \ldots, (x_n, y_n)$, the **sample covariance (s_{xy})** averages the cross-products of deviations around each variable's mean:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{SS_{xy}}{n-1}$$

Where: s_{xy} is the sample covariance, $SS_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ is the sum of cross-deviations, $\bar{x}$ and $\bar{y}$ are the sample means, and n is the number of paired observations. Positive covariance indicates that above-average X tends to accompany above-average Y, whereas negative covariance indicates an inverse association. However, because s_{xy} depends directly on measurement units (e.g., AUD^2 versus $\text{AUD} \times \text{kilograms}$), its raw magnitude cannot measure association strength.

Dividing covariance by the product of the two sample standard deviations removes measurement units entirely, yielding the dimensionless **Pearson product-moment correlation coefficient (r)**:

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\cancel{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}} = \frac{SS_{xy}}{\cancel{\sqrt{SS_{xx} \cdot SS_{yy}}}}$$

Where: r is the sample Pearson correlation ($-1 \le r \le +1$), s_x and s_y are the sample standard deviations of X and Y, $SS_{xx} = \sum(x_i - \bar{x})^2$, and $SS_{yy} = \sum(y_i - \bar{y})^2$. Key properties of r include: (1) $r = +1$ or $r = -1$ occurs if and only if all points lie exactly on a non-horizontal straight line; (2) r is symmetric ($r_{xy} = r_{yx}$) and invariant to positive linear rescaling (e.g., converting AUD to USD or Celsius to Fahrenheit); and (3) r measures *linear* association only—a strong parabolic relationship (such as inventory holding plus stockout cost against order size) can yield $r \approx 0$.

Correlation versus Causation and Testing Population Correlation (ρ)

A high sample correlation never by itself proves that X causes Y. Apparent correlation may reflect reverse causality, coincidental trending over time (spurious correlation), or an omitted **confounding variable** driving both X and Y simultaneously (for example, warm summer weather simultaneously raising both ice-cream sales and coastal surf rescues). When assessing ordinal data or monotonic nonlinear relationships distorted by outliers, analysts compute **Spearman's rank correlation (r_s)** by applying Pearson's formula to the within-sample ranks of X and Y.

To test whether an observed sample correlation r reflects genuine linear association in a bivariate normal population ($H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$), the exact test statistic is:

$$t_{\text{stat}} = r\sqrt{\frac{n-2}{1-r^2}} \sim t_{n-2} \text{ under } H_0$$

Where: r is the sample Pearson correlation, n is the number of paired observations, and $n - 2$ is the degrees of freedom.

DegreeLive

Worked Numerical Example 1: Advertising Spend and Weekly E-Commerce Revenue

A Melbourne online retailer tracks digital advertising expenditure (X, in A$ thousands) and gross weekly revenue (Y, in A$ thousands) across $n = 8$ campaign weeks. From preliminary totals: $\bar{x} = 23.0$, $\bar{y} = 80.5$, $SS_{xx} = 530$, $SS_{yy} = 2,638$, and $SS_{xy} = 1,179$. Compute the covariance, Pearson correlation, and test $H_0 : \rho = 0$ at $\alpha = 0.05$.

$$s_{xy} = \frac{1,179}{8-1} = 168.43, \qquad r = \frac{1,179}{\sqrt{530 \times 2,638}} = \frac{1,179}{\sqrt{1,398,140}} = \frac{1,179}{1,182.43} = +0.9971$$

$$t_{\text{stat}} = 0.9971\sqrt{\frac{8-2}{1-(0.9971)^2}} = 0.9971\sqrt{\frac{6}{0.00579}} = 32.09$$

Where: $s_{xy} = 168.43$ is the sample covariance, $r = +0.9971$ indicates a near-perfect positive linear association, and $t_{\text{stat}} = 32.09$ with $df = 6$ far exceeds the two-tailed 5% critical value $t_{0.025,\,6} = 2.4469$ ($p < 0.0001$), confirming a statistically significant linear relationship.

Spearman's Rank Correlation Worked Example

When evaluating ordinal rankings or heavily skewed financial metrics where extreme outliers distort Pearson's r, Spearman's rank correlation (r_s) replaces raw values with their integer ranks $(1, 2, \ldots, n)$. If no rank ties occur, r_s simplifies to:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

Where: $d_i = \text{Rank}(x_i) - \text{Rank}(y_i)$ is the rank difference for observation i and n is the number of pairs. Suppose $n = 6$ South Australian Shiraz vintages are ranked from lowest to highest by retail bottle price (1, 2, 3, 4, 5, 6) and independently ranked by a wine critic (2, 1, 3, 5, 4, 6). The rank differences are $d_i = (-1, +1, 0, -1, +1, 0)$, giving $\sum d_i^2 = 1 + 1 + 0 + 1 + 1 + 0 = 4$. Thus $r_s = 1 - \frac{6(4)}{6(36-1)} = 1 - \frac{24}{210} = +0.8857$, demonstrating strong monotonic agreement between bottle price and critic ranking.

4.2 Simple Linear Regression (SLR) and Ordinary Least Squares (OLS)

Whereas correlation treats X and Y symmetrically, **regression analysis** models how a single **dependent (response) variable (Y)** depends directionally on one or more **independent (explanatory or predictor) variables (X)**, enabling both marginal causal/structural quantification and quantitative prediction.

Population Regression Function versus Sample OLS Regression Line

The **Simple Linear Regression (SLR) population model** expresses Y_i as a linear function of X_i plus an unobservable random error (disturbance) term ε_i capturing all other omitted factors and inherent randomness:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

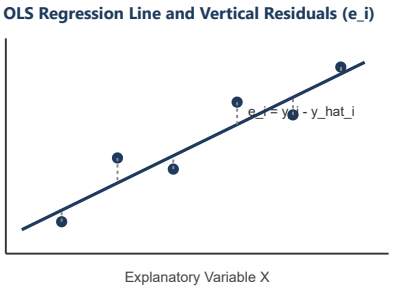
Where: Y_i is the response for observation i, X_i is the explanatory predictor, β_0 is the unknown population intercept, β_1 is the unknown population slope (the expected change in Y per one-unit increase in X), and ε_i is the random error term with $E(\varepsilon_i \mid X_i) = 0$.

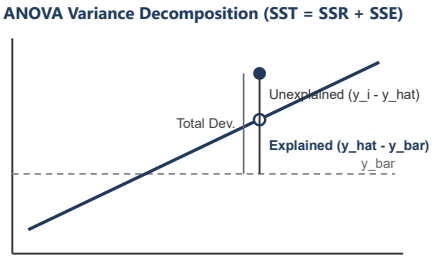
Because β_0 and β_1 are unknown, they are estimated from sample data $(x_i, y_i)_{i=1}^{n}$ by fitting a sample regression line $\hat{y}_i = b_0 + b_1 x_i$. For each observation, the vertical discrepancy between the actual outcome and the fitted value is the **residual ($e_i = y_i - \hat{y}_i$)**. The **Ordinary Least Squares (OLS)** criterion chooses b_0 and b_1 to minimise the Sum of Squared Errors (Residuals), $\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$. Setting partial derivatives with respect to b_0 and b_1 to zero yields the OLS normal equations and closed-form estimators:

DegreeLive

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{SS_{xy}}{SS_{xx}} \qquad \text{and} \qquad b_0 = \bar{y} - b_1\bar{x}$$

Where: b_1 is the OLS estimated slope, b_0 is the OLS estimated intercept, SS_{xy} is the sum of cross-deviations, SS_{xx} is the sum of squared x-deviations, and $(\bar{x}, \bar{y})$ is the sample centroid through which every OLS regression line always passes. By the first-order conditions $\frac{\partial \text{SSE}}{\partial b_0} = -2\sum(y_i - b_0 - b_1 x_i) = 0$ and $\frac{\partial \text{SSE}}{\partial b_1} = -2\sum x_i(y_i - b_0 - b_1 x_i) = 0$, every OLS fit guarantees two algebraic orthogonality identities: $\sum_{i=1}^{n} e_i = 0$ and $\sum_{i=1}^{n} x_i e_i = 0$ (meaning residuals have zero sample mean and zero sample correlation with X).





Interpretation and Extrapolation Risks

The slope b_1 measures the predicted change in Y associated with a one-unit increase in X. The intercept b_0 represents the predicted Y when $X = 0$; however, b_0 only carries substantive economic meaning if $X = 0$ is physically feasible and lies near the observed sample domain. Predicting Y for values of X far outside the observed range $[\min(x_i), \max(x_i)]$—known as **extrapolation**—is hazardous because linearity may break down beyond observed market regimes.

Worked Numerical Example 2: Full OLS Calculation Table

Continuing the Melbourne e-commerce dataset ($n = 8$ weeks) from Section 4.1, the full observation-level table for advertising spend (x_i, A$ '000s) and weekly revenue (y_i, A$ '000s) is constructed below ($\bar{x} = 23, \bar{y} = 80.5$):

Week (i)	Ad Spend (x_i)	Revenue (y_i)	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$	Fitted ($\hat{y}_i$)	Residual (e_i)
1	10	52	-13	-28.5	169	370.5	51.58	+0.42
2	14	61	-9	-19.5	81	175.5	60.48	+0.52
3	18	68	-5	-12.5	25	62.5	69.38	-1.38
4	22	79	-1	-1.5	1	1.5	78.28	+0.72
5	25	84	+2	+3.5	4	7.0	84.95	-0.95
6	28	93	+5	+12.5	25	62.5	91.62	+1.38
7	32	98	+9	+17.5	81	157.5	100.52	-2.52
8	35	109	+12	+28.5	144	342.0	107.19	+1.81
Sum	**184**	**644**	**0**	**0.0**	**530**	**1,179.0**	**644.00**	**0.00**

$$b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{1,179}{530} = 2.2245, \qquad b_0 = \bar{y} - b_1\bar{x} = 80.5 - (2.224528)(23) = 29.3358$$

$$\hat{y} = 29.34 + 2.2245x$$

Degreelive

Where: $b_1 = 2.2245$ indicates that each additional A$1,000 in weekly digital advertising spend is associated with an estimated A$2,224.50 increase in gross weekly revenue, and $b_0 = 29.34$ represents baseline organic revenue (A$29,340) at zero spend—interpreted cautiously since $x = 0$ lies below the observed range $[10, 35]$. Notice also that OLS residuals always sum to zero ($\sum e_i = 0$).

4.3 Goodness of Fit, ANOVA Decomposition, and Standard Error of the Estimate

Fitting an OLS line always produces numerical values for b_0 and b_1, even if X has negligible explanatory power. Analysts must evaluate how tightly the regression line fits the observed data.

Sum of Squares Decomposition and Coefficient of Determination (R^2)

Total variation in Y around its sample mean partitions orthogonally into explained variation (due to regression) plus unexplained residual variation:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad \Leftrightarrow \qquad \text{SST} = \text{SSR} + \text{SSE}$$

Where: $\text{SST} = SS_{yy}$ is the Total Sum of Squares ($df = n - 1$), SSR is the Regression Sum of Squares ($df = 1$ in SLR), and SSE is the Error (Residual) Sum of Squares ($df = n - 2$ in SLR). Dividing SSR by SST defines the **coefficient of determination (R^2)**, the proportion of total sample variation in Y explained by linear variation in X:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}, \qquad 0 \le R^2 \le 1$$

Where: R^2 is the coefficient of determination (in simple linear regression only, $R^2 = r^2$). While R^2 measures relative fit as a dimensionless percentage, managers also need an absolute measure of typical prediction error in the original units of Y, supplied by the **standard error of the estimate (s_e)**:

$$s_e = \sqrt{\text{MSE}} = \sqrt{\frac{\text{SSE}}{n - 2}}$$

Where: s_e estimates the population error standard deviation σ_ε, $\text{MSE} = \text{SSE}/(n-2)$ is the Mean Square Error, and $n - 2$ reflects two degrees of freedom lost in estimating b_0 and b_1.

Worked Numerical Example 3: Regression ANOVA Table

For the e-commerce advertising regression ($n = 8$, $\text{SST} = SS_{yy} = 2,638$), $\text{SSR} = b_1 \cdot SS_{xy} = (2.224528)(1,179) = 2,622.7189$ and $\text{SSE} = 2,638 - 2,622.7189 = 15.2811$. These populate the Regression ANOVA table:

Source of Variation	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Square ($MS = SS/df$)	F-Statistic
Regression (Model)	1	2,622.72	$\text{MSR} = 2,622.72$	$F = 2,622.72/2.5469 = 1,029.79$
Residual (Error)	$n - 2 = 6$	15.28	$\text{MSE} = 2.5469$	—
Total	$n - 1 = 7$	**2,638.00**	—	—

$$R^2 = \frac{2,622.72}{2,638.00} = 0.9942 \ (99.42\%) \qquad \text{and} \qquad s_e = \sqrt{2.5469} = 1.5959 \ (\text{AUD } 1,596)$$

DegreeLive

Where: $R^2 = 99.42\%$ implies that 99.42% of weekly revenue variation is explained linearly by digital advertising spend, and $s_e = \text{AUD } 1,596$ measures the typical residual deviation around the regression line.

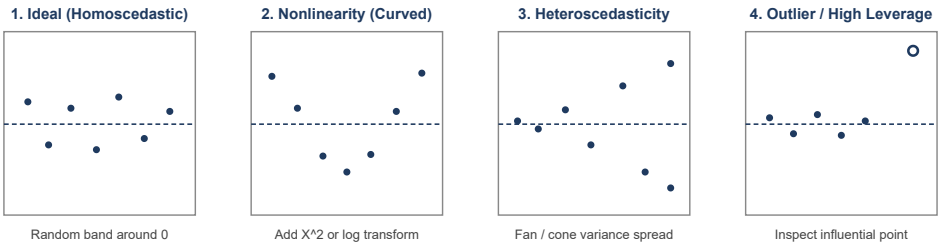
Interpreting R^2 Responsibly in Business and Economics

Analysts must avoid evaluating regression quality solely on R^2. First, in cross-sectional micro-econometric studies of individual consumer spending or employee wages, idiosyncratic human heterogeneity routinely produces low R^2 values (0.10 to 0.25) even when the estimated slope b_1 is highly statistically significant ($p < 0.001$) and policy-relevant. Second, regressing two unrelated time series that share a common upward drift (such as Australian nominal GDP and global smartphone ownership) can yield a deceptive **spurious** $R^2 > 0.95$ without any causal link. Third, R^2 values can only be compared across models sharing the *exact same dependent variable* Y: comparing R^2 from a linear model in Y against R^2 from a log-linear model in $\ln(Y)$ is mathematically invalid because SST is measured on different scales.

4.4 Classical Assumptions, Residual Diagnostics, and Inference in SLR

Under the **Gauss-Markov (LINE) classical linear regression assumptions**, OLS estimators b_0 and b_1 are Best Linear Unbiased Estimators (BLUE)—having the minimum variance among all linear unbiased estimators—and standard t and F inference is exact:

1. **Linearity:** $E(Y_i \mid X_i) = \beta_0 + \beta_1 X_i$ (the true conditional mean is linear in parameters).
2. **Independence:** Errors ε_i are statistically independent across observations ($\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$). In time-series regressions, serial correlation (autocorrelation) violates independence and can be flagged via the Durbin-Watson statistic.
3. **Normality:** Errors are normally distributed, $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ (essential for exact small-sample t and F tests; relaxed asymptotically for large n).
4. **Equal Variance (Homoscedasticity):** Conditional error variance is constant across all levels of X, $\text{Var}(\varepsilon_i \mid X_i) = \sigma_\varepsilon^2$. Non-constant variance (e.g., fanning variance in household expenditure as income grows) is termed **heteroscedasticity**, which leaves b_1 unbiased but invalidates conventional standard errors unless heteroscedasticity-robust (White) standard errors are used.



Hypothesis Testing on the Slope (β_1) and Prediction Intervals

To test whether X exerts a statistically significant linear effect on Y ($H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$), compute the standard error of b_1 and t-statistic ($df = n - 2$):

$$\text{SE}(b_1) = \frac{s_e}{\sqrt{SS_{xx}}}, \qquad t_{\text{stat}} = \frac{b_1 - 0}{\text{SE}(b_1)} \sim t_{n-2}, \qquad \text{CI for } \beta_1 : \; b_1 \pm t_{\alpha/2,\, n-2}\, \text{SE}(b_1)$$

Where: $\text{SE}(b_1)$ is the standard error of the slope, s_e is the standard error of the estimate, and $SS_{xx} = \sum (x_i - \bar{x})^2$. Notice that in SLR, $t_{\text{stat}}^2 = F_{\text{ANOVA}}$ identically.

When forecasting at a specific predictor value x^*, analysts distinguish between estimating the *average* response $E(Y \mid x^*)$ across all units versus predicting a *single individual* future outcome $Y_{\text{new}} \mid x^*$. The **Prediction Interval (PI)** for an individual outcome is strictly wider than the **Confidence Interval (CI)** for the mean response because it includes both estimation uncertainty in the line and an extra +1 variance term for individual error ε_{new}:

$$\text{CI for } E(Y \mid x^*): \quad \hat{y}^* \pm t_{\alpha/2,\,n-2}\, s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{SS_{xx}}}$$

$$\text{PI for } Y_{\text{new}} \mid x^*: \quad \hat{y}^* \pm t_{\alpha/2,\,n-2}\, s_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{SS_{xx}}}$$

Where: $\hat{y}^* = b_0 + b_1 x^*$ is the point prediction at x^* and $(x^* - \bar{x})^2 / SS_{xx}$ captures rising uncertainty as x^* moves away from the sample mean $\bar{x}$.

Worked Numerical Example 4: Slope Inference and Interval Forecasting

From the $n = 8$ e-commerce dataset ($s_e = 1.5959$, $SS_{xx} = 530$, $df = 6$, $t_{0.025,6} = 2.4469$):

1. **Slope Test & 95% CI:** $\text{SE}(b_1) = 1.5959/\sqrt{530} = 0.0693$. Thus $t_{\text{stat}} = 2.2245/0.0693 = 32.09$ (note $32.09^2 = 1,029.8 \approx F_{\text{ANOVA}}$), rejecting $H_0: \beta_1 = 0$. The 95% CI for β_1 is $2.2245 \pm 2.4469(0.0693) = [2.0549,\ 2.3941]$.
2. **95% CI vs. 95% PI at $x^* = 24$ (A$24,000 ad spend):** Point prediction $\hat{y}^* = 29.3358 + 2.224528(24) = 82.7245$ (A$82,725).

$$\text{CI Half-Width} = 2.4469(1.5959)\sqrt{\frac{1}{8} + \frac{(24-23)^2}{530}} = 3.9050 \times 0.3562 = 1.391 \implies [81.33,\ 84.12]$$

$$\text{PI Half-Width} = 2.4469(1.5959)\sqrt{1 + \frac{1}{8} + \frac{(24-23)^2}{530}} = 3.9050 \times 1.0615 = 4.145 \implies [78.58,\ 86.87]$$

Where: $[81.33, 84.12]$ estimates mean weekly revenue at A$24,000 ad spend, whereas $[78.58, 86.87]$ (in A$ thousands) bounds revenue for a single individual week.

Nonlinear Functional Forms via Variable Transformations

Linearity in parameters allows OLS to estimate curved economic relationships—such as diminishing marginal returns or constant-elasticity demand curves—by transforming Y, X, or both using natural logarithms ($\ln$) or quadratic terms (X^2) prior to estimation:

Specification	Model Equation	Economic Interpretation of β_1	Typical Business Application
Level–Level	$Y = \beta_0 + \beta_1 X + \varepsilon$	$\Delta Y \approx \beta_1 \Delta X$ (unit change)	Ad spend (A$) vs. revenue (A$)
Log–Level (Semi-Log)	$\ln(Y) = \beta_0 + \beta_1 X + \varepsilon$	$\%\Delta Y \approx (100\beta_1)\Delta X$	Wage growth per year of experience
Level–Log	$Y = \beta_0 + \beta_1 \ln(X) + \varepsilon$	$\Delta Y \approx (\beta_1/100) \times (\%\Delta X)$	Diminishing returns to marketing reach
Log–Log (Constant Elasticity)	$\ln(Y) = \beta_0 + \beta_1 \ln(X) + \varepsilon$	$\%\Delta Y \approx \beta_1 (\%\Delta X)$ (elasticity)	Price elasticity of retail demand

Comprehensive Regression Diagnostic Checklist

Before deploying any estimated regression model for business forecasting or policy recommendations, analysts verify assumptions systematically across residuals and predictors:

Diagnostic Tool	Target Assumption	Warning Pattern / Threshold	Corrective Action
Residuals vs. Fitted Plot	Linearity & Homoscedasticity	U-shape curve or fanning cone spread	Add quadratic/log terms; use White robust SEs

DegreeLive

Diagnostic Tool	Target Assumption	Warning Pattern / Threshold	Corrective Action
Normal Q–Q Plot of Residuals	Normality of ε_i	Systematic S-curve or heavy-tail deviations	Transform Y ($\ln Y$); rely on large-n CLT
Durbin–Watson (DW) Statistic	Independence (Time Series)	DW far below 2 (positive autocorrelation)	Include lagged variables or trend/seasonality
Cook's Distance / Leverage	Influence & Outliers	Isolated high-leverage point distorting b_1	Audit data entry; estimate with/without point
Variance Inflation Factor (VIF)	No Perfect Multicollinearity	$\text{VIF}_j = 1/(1 - R_j^2) > 10$	Drop redundant predictor or combine index

4.5 Multiple Linear Regression (MLR) Foundations

Real economic outcomes rarely depend on a single driver. Extending OLS to k explanatory variables yields the **Multiple Linear Regression (MLR) model**:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

Where: β_0 is the intercept and each **partial slope coefficient (β_j)** measures the expected change in Y for a one-unit increase in X_j, *holding all other included explanatory variables constant* (the *ceteris paribus* principle). Controlling for confounding variables simultaneously prevents omitted-variable bias that distorts simple regressions.

Model Fit (R^2_{adj}), Global F-Test, Dummy Variables, and Multicollinearity

- **Adjusted R^2 (R^2_{adj}):** Because adding *any* predictor—even random noise—never lowers ordinary R^2, analysts compare competing MLR specifications using Adjusted R^2, which penalises unnecessary predictors ($df = n - k - 1$):

$$R^2_{\text{adj}} = 1 - (1 - R^2)\left(\frac{n-1}{n-k-1}\right)$$

  Where: n is the sample size and k is the number of slope predictors.

- **Global F-Test versus Partial t-Tests:** The global ANOVA F-test evaluates $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$ (whether the model as a whole explains significant variation, $df_1 = k$, $df_2 = n - k - 1$). Individual partial t-tests ($t_j = b_j / \text{SE}(b_j)$ with $df = n - k - 1$) test each predictor's incremental contribution after controlling for all other predictors.
- **Dummy (Indicator) Variables and Interactions:** Qualitative attributes with m categories enter MLR via $m - 1$ binary $\{0, 1\}$ dummy variables (omitting one reference category to avoid perfect collinearity, the **dummy variable trap**). Multiplying two predictors (e.g., $X_1 \times D_1$) creates an **interaction term** allowing the slope of X_1 to vary across subgroups.
- **Multicollinearity:** When two or more predictors are strongly correlated with each other, OLS coefficients remain unbiased, but standard errors inflate severely, obscuring individual t-tests and flipping signs. Analysts diagnose collinearity via the **Variance Inflation Factor ($\text{VIF}_j = 1/(1 - R_j^2)$)**, where R_j^2 is obtained by regressing X_j on the remaining predictors; $\text{VIF}_j > 10$ signals serious multicollinearity.

Worked Numerical Example 5: Interpreting Melbourne Residential Property Regression Output

A Melbourne property economist models auction sale price (Y, in A$ thousands) across $n = 120$ houses using internal floor area (X_1, m^2), radial distance from Melbourne GPO (X_2, km), and a heritage-overlay indicator ($D_1 = 1$ if heritage-listed, o otherwise). Software produces the following regression summary ($R^2 = 0.784$, $R^2_{\text{adj}} = 0.778$, $s_e = 112.4$, Global $F = 140.4$, $p < 0.0001$):

Term	Coefficient (b_j)	Std. Error ($\text{SE}(b_j)$)	t-Statistic	p-Value
Intercept (b_0)	685.20	74.10	9.25	< 0.0001

Term	Coefficient (b_j)	Std. Error ($\mathrm{SE}(b_j)$)	t-Statistic	p-Value
Floor Area (m^2, X_1)	4.85	0.38	12.76	< 0.0001
Distance from CBD (km, X_2)	-34.60	3.25	-10.65	< 0.0001
Heritage Overlay (D_1)	142.50	46.80	3.04	0.0029

Interpretation: Holding CBD distance and heritage status fixed, each additional square metre of floor area adds an estimated A$4,850 to sale price ($p < 0.0001$). Holding floor area and heritage status fixed, each extra kilometre away from Melbourne's CBD lowers expected price by A$34,600 ($p < 0.0001$). Holding size and distance fixed, heritage-listed dwellings command an average premium of A$142,500 ($p = 0.0029$). Together, the three predictors explain 78.4% of price variation (77.8% adjusted for $k = 3$ predictors).

Extending to Interaction Terms and Multicollinearity Diagnostics: If an analyst suspects that the marginal price per square metre differs between heritage and non-heritage homes, adding an interaction product $X_1 \cdot D_1$ yields $\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3D_1 + b_4(X_1D_1)$. For non-heritage homes ($D_1 = 0$), the floor-area slope is b_1; for heritage homes ($D_1 = 1$), the floor-area slope becomes $(b_1 + b_4)$. Conversely, if a model includes both floor area (m^2) and total bedroom count, high correlation between them can cause $\mathrm{VIF} > 10$, producing a highly significant global F-statistic ($p < 0.0001$) alongside deceptive, individually insignificant partial t-statistics—a classic symptom of multicollinearity.

Omitted Variable Bias (OVB) in Business Econometrics

Omitting a relevant explanatory variable Z from a regression does not merely increase residual variance—it actively biases the estimated slope on included variable X whenever Z both affects Y ($\beta_2 \neq 0$) and correlates with X ($\mathrm{Cov}(X, Z) \neq 0$). Formally, if the true population model is $Y = \beta_0 + \beta_1X + \beta_2Z + \varepsilon$, regressing Y on X alone yields an expected slope of:

$$E(b_{1,\text{short}}) = \beta_1 + \beta_2 \cdot \frac{\mathrm{Cov}(X, Z)}{\mathrm{Var}(X)}$$

Where: β_1 is the true partial effect of X, β_2 is the true partial effect of the omitted variable Z, and $\mathrm{Cov}(X, Z)/\mathrm{Var}(X)$ is the slope from regressing Z on X. For example, estimating the effect of store staff hours (X) on retail sales (Y) without controlling for store floor-area or CBD foot-traffic (Z) induces upward bias ($\beta_2 > 0$ and $\mathrm{Cov}(X, Z) > 0$), overstating labour productivity unless Z is explicitly included in a multiple regression.

4.6 Business Forecasting and Time-Series Foundations

Budgeting, inventory replenishment, and macroeconomic policy require projecting historical sequential data ($Y_1, Y_2, \ldots, Y_T$) into future periods ($T + 1, T + 2, \ldots$). Classical time-series analysis decomposes an observed series Y_t into four structural components: **Trend (T_t)** (long-run secular direction), **Cyclical (C_t)** (multi-year macroeconomic expansions and recessions around trend), **Seasonal (S_t)** (regular calendar fluctuations within a year, such as Q4 Australian retail surges), and **Irregular (I_t)** (unpredictable shocks). Two standard decompositions relate these components: **additive** ($Y_t = T_t + C_t + S_t + I_t$, appropriate when seasonal amplitude stays constant in dollars) and **multiplicative** ($Y_t = T_t \times C_t \times S_t \times I_t$, appropriate when seasonal swings grow proportionally with trend level).

Smoothing Methods and Regression-Based Trend and Seasonality

- **Moving Averages (MA):** A k-period simple moving average forecasts $F_{t+1} = \frac{1}{k}\sum_{i=0}^{k-1} Y_{t-i}$. For quarterly ($k = 4$) or monthly ($k = 12$) seasonal data, centred moving averages eliminate seasonal swings cleanly to isolate the underlying trend-cycle.
- **Simple Exponential Smoothing (SES):** Assigns exponentially decaying weights to older observations using a smoothing constant $\alpha \in (0, 1)$ (initialised typically at $F_1 = Y_1$):

$$F_{t+1} = \alpha Y_t + (1 - \alpha)F_t = F_t + \alpha(Y_t - F_t)$$

DegreeLive

Where: F_{t+1} is the one-step-ahead forecast, Y_t is the actual observation at time t, F_t is the prior forecast, and α controls responsiveness (larger α reacts faster to recent shifts; smaller α filters noise more aggressively).

- **Trend and Seasonal Dummy Regression:** When a series exhibits steady growth and stable quarterly seasonality, analysts regress Y_t on a time index $t = 1, 2, \dots, T$ and three quarterly dummy indicators (Q_{2t}, Q_{3t}, Q_{4t}, with Quarter 1 as the baseline):

$$\hat{Y}_t = b_0 + b_1 t + b_2 Q_{2t} + b_3 Q_{3t} + b_4 Q_{4t}$$

Where: b_1 estimates constant trend growth per quarter, and b_2, b_3, b_4 estimate the average seasonal shift of Quarters 2, 3, and 4 relative to Quarter 1.

Forecast Accuracy Metrics

Forecasting models are validated out-of-sample (or across m holdout periods) using error $e_t = Y_t - F_t$ via four standard accuracy metrics:

$$\text{MAE} = \frac{1}{m}\sum_{t=1}^{m}|e_t|, \qquad \text{MSE} = \frac{1}{m}\sum_{t=1}^{m}e_t^2, \qquad \text{RMSE} = \sqrt{\text{MSE}}, \qquad \text{MAPE} = \frac{100\%}{m}\sum_{t=1}^{m}\left|\frac{e_t}{Y_t}\right|$$

Where: $e_t = Y_t - F_t$ is the forecast error in period t, m is the number of evaluated forecast periods, MAE is Mean Absolute Error, RMSE is Root Mean Squared Error (penalising large errors more heavily in original units), and MAPE is Mean Absolute Percentage Error (scale-free percentage accuracy).

Worked Numerical Example 6: Exponential Smoothing and Forecast Accuracy

Quarterly deseasonalised revenue (A$ millions) for an Australian hardware chain across six quarters is $Y_1 = 110$, $Y_2 = 118$, $Y_3 = 115$, $Y_4 = 124$, $Y_5 = 130$, $Y_6 = 128$. Apply Simple Exponential Smoothing with $\alpha = 0.30$ and initial forecast $F_1 = 110.00$, then compute accuracy metrics across periods $t = 2, \dots, 6$ ($m = 5$) and forecast F_7.

Quarter (t)	Actual (Y_t)	Forecast (F_t)	Error ($e_t = Y_t - F_t$)	$\lvert e_t \rvert$	e_t^2	$\lvert e_t/Y_t \rvert \times 100\%$
1	110	110.00	—	—	—	—
2	118	$0.3(110) + 0.7(110) = 110.00$	+8.00	8.00	64.00	6.78%
3	115	$0.3(118) + 0.7(110) = 112.40$	+2.60	2.60	6.76	2.26%
4	124	$0.3(115) + 0.7(112.40) = 113.18$	+10.82	10.82	117.07	8.73%
5	130	$0.3(124) + 0.7(113.18) = 116.43$	+13.57	13.57	184.14	10.44%
6	128	$0.3(130) + 0.7(116.43) = 120.50$	+7.50	7.50	56.25	5.86%
Sum ($m = 5$)	—	—	—	**42.49**	**428.22**	**34.07%**

$$F_7 = 0.30(128) + 0.70(120.50) = 38.40 + 84.35 = \text{AUD } 122.75\text{M}$$

$$\text{MAE} = \frac{42.49}{5} = \text{AUD } 8.50\text{M}, \qquad \text{RMSE} = \sqrt{\frac{428.22}{5}} = \sqrt{85.64} = \text{AUD } 9.25\text{M}, \qquad \text{MAPE} = \frac{34.07\%}{5} = 6.81\%$$

Where: $F_7 = \text{AUD } 122.75\text{M}$ is the Quarter 7 forecast. Because actual revenues trend steadily upward from 110 to 128, every forecast error e_t is positive—illustrating a critical property of Simple Exponential Smoothing: SES lags behind a persistent upward trend, signalling that a trend-adjusted model (such as Holt's linear exponential smoothing or linear trend regression) would

DegreeLive

eliminate systematic under-prediction in future quarters.

Benchmarking Against a 3-Quarter Moving Average & Holt's Trend Extension: Applying a 3-quarter Moving Average ($k = 3$) to the same series yields $F_4 = (110 + 118 + 115)/3 = 114.33$, $F_5 = (118 + 115 + 124)/3 = 119.00$, $F_6 = (115 + 124 + 130)/3 = 123.00$, and $F_7 = (124 + 130 + 128)/3 = \text{AUD } 127.33\text{M}$. Across $t = 4, 5, 6$, the 3-quarter MA produces errors $(+9.67, +11.00, +5.00)$ with $\text{MAE} = 8.56\text{M}$ and $\text{RMSE} = 8.93\text{M}$, responding faster to recent level growth than $\alpha = 0.30$ SES. To track persistent growth trends systematically without lag, **Holt's Double Exponential Smoothing** updates both a level equation $L_t = \alpha Y_t + (1 - \alpha)(L_{t-1} + T_{t-1})$ and a slope trend equation $T_t = \beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1}$, generating h-step-ahead forecasts $F_{t+h} = L_t + h\,T_t$.

Worked Numerical Example 7: Quarterly Trend-and-Seasonal Dummy Forecast

To capture both secular growth and regular quarterly seasonality in Australian department-store turnover (A$ millions), an analyst estimates an OLS regression over 16 historical quarters ($t = 1, \ldots, 16$) with Quarter 1 as the omitted reference season:

$$\hat{Y}_t = 420.0 + 6.5\,t - 28.0\,Q_{2t} - 15.0\,Q_{3t} + 62.0\,Q_{4t}$$

Where: t is the sequential quarter index, Q_{2t}, Q_{3t}, Q_{4t} are binary seasonal dummies, +6.5 is the underlying quarterly growth trend (A$6.5M per quarter), and +62.0 captures the Q4 Christmas trading surge relative to Q1. Projecting the four quarters of Year 5 ($t = 17, 18, 19, 20$) yields:

- **Year 5, Q1 (**$t = 17$, $Q_2 = Q_3 = Q_4 = 0$**):** $\hat{Y}_{17} = 420.0 + 6.5(17) = \text{AUD } 530.50\text{M}$.
- **Year 5, Q2 (**$t = 18$, $Q_2 = 1$**):** $\hat{Y}_{18} = 420.0 + 6.5(18) - 28.0 = \text{AUD } 509.00\text{M}$.
- **Year 5, Q3 (**$t = 19$, $Q_3 = 1$**):** $\hat{Y}_{19} = 420.0 + 6.5(19) - 15.0 = \text{AUD } 528.50\text{M}$.
- **Year 5, Q4 (**$t = 20$, $Q_4 = 1$**):** $\hat{Y}_{20} = 420.0 + 6.5(20) + 62.0 = \text{AUD } 612.00\text{M}$.

Worked Numerical Example 8: Centred 4-Quarter Moving Averages and Seasonal Ratios

When seasonal swings grow proportionally with sales volume, analysts isolate the trend-cycle (TC_t) using a **centred 4-quarter moving average (CMA)**—averaging two consecutive 4-quarter moving averages so the smoothed estimate aligns precisely with an observed quarter—and divide Y_t/CMA_t to extract multiplicative seasonal-irregular ratios. Consider eight quarters of Australian specialty retail revenue (A$ millions):

Period (t)	Quarter	Revenue (Y_t)	4-Quarter MA	Centred MA (CMA_t)	Seasonal Ratio (Y_t/CMA_t)
1	Y1 Q1	80	—	—	—
2	Y1 Q2	72	$(80 + 72 + 84 + 124)/4 = 90.0$ (between Q2–Q3)	—	—
3	Y1 Q3	84	$(72 + 84 + 124 + 88)/4 = 92.0$ (between Q3–Q4)	$(90.0 + 92.0)/2 = 91.0$	$84/91.0 = 0.9231$
4	Y1 Q4	124	$(84 + 124 + 88 + 80)/4 = 94.0$ (between Q4–Q1)	$(92.0 + 94.0)/2 = 93.0$	$124/93.0 = 1.3333$
5	Y2 Q1	88	$(124 + 88 + 80 + 92)/4 = 96.0$ (between Q1–Q2)	$(94.0 + 96.0)/2 = 95.0$	$88/95.0 = 0.9263$
6	Y2 Q2	80	$(88 + 80 + 92 + 136)/4 = 99.0$ (between Q2–Q3)	$(96.0 + 99.0)/2 = 97.5$	$80/97.5 = 0.8205$
7	Y2 Q3	92	—	—	—
8	Y2 Q4	136	—	—	—

The centred moving average removes all quarterly calendar noise, exposing a steady underlying trend-cycle growing from 91.0 to 97.5 million, while the seasonal ratio of 1.3333 in Q4 reveals that December-quarter turnover runs 33.3% above the annual trend baseline whereas Q2 sits 17.95% below trend (0.8205). Dividing historical revenue by normalised seasonal indices yields

seasonally adjusted figures—the same adjustment methodology published monthly by the Australian Bureau of Statistics for retail trade and labour force data.

DegreeLive