DegreeLive

Bachelor of Commerce (B2001) | Level 1 • Part B Major | Monash University

Introduction to data analysis

Course Code: ETC1010 • Module 3 Notes

Module 3: Statistical Modelling, Linear & Logistic Regression & Model Assessment

3.1 Statistical Modelling Paradigm and Formula Interfaces

Exploratory data visualisation reveals empirical patterns, whereas statistical modelling formalises those patterns into compact mathematical structures separating systematic **signal** from stochastic **noise**. Following George Box's dictum that "all models are wrong, but some are useful," a statistical model does not replicate reality in exhaustive detail; instead, it extracts an interpretable, falsifiable approximation tailored to a specific analytical objective:

- **Descriptive Modelling:** Summarises conditional averages or covariation in a historical sample without claiming causal intervention or forecasting out-of-sample behaviour (for example, summarising median Victorian household electricity expenditure across income deciles).
- **Inferential (Explanatory) Modelling:** Estimates structural parameters and quantifies sampling uncertainty (standard errors, confidence intervals, hypothesis tests) to evaluate theoretical mechanisms or policy effects (for example, assessing whether an RBA cash rate increase shifts mortgage arrears).
- **Predictive Modelling:** Optimises out-of-sample forecasting accuracy on unseen observations, prioritising generalisation error metrics over individual parameter interpretability (for example, predicting credit default risk or weekly retail foot traffic).

In general, any supervised regression or classification model relates a response (outcome) variable y_i to a vector of p predictors $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})^\top$ through a systematic function $f(\mathbf{x}_i)$ and an unobservable random error term ε_i:

$$y_i = f(\mathbf{x}_i) + \varepsilon_i$$

Where:

- y_i is the observed response value for observation $i \in \{1, \dots, n\}$.
- $f(\mathbf{x}_i) = E[y_i \mid \mathbf{x}_i]$ is the systematic conditional expectation function mapping predictors to expected outcomes.
- ε_i is the irreducible stochastic disturbance capturing unmeasured factors and random noise with $E[\varepsilon_i \mid \mathbf{x}_i] = 0$.

Wilkinson-Rogers Symbolic Formula Interfaces and Design Matrices

Rather than requiring analysts to manually construct dummy columns, polynomial bases, and interaction products, modern modelling environments (R's `lm()`, `glm()`, and Python's `statsmodels/patsy`) adopt the **Wilkinson-Rogers symbolic formula grammar**. A formula `response ~ terms` instructs the modelling engine how to assemble the numerical **design matrix** $\mathbf{X} \in \mathrm{R}^{n \times k}$ passed to the estimation solver:

DegreeLive

Formula Expression	Expanded Linear Predictor	Design Matrix Effect
`y ~ x1 + x2`	$\beta_0 + \beta_1 x_1 + \beta_2 x_2$	Intercept column of 1s plus raw columns `x1`, `x2`
`y ~ x1 + x2 - 1`	$\beta_1 x_1 + \beta_2 x_2$	Suppresses the intercept column of 1s
`y ~ x1:x2`	$\beta_0 + \beta_1 (x_1 \cdot x_2)$	Adds only the product interaction term without main effects
`y ~ x1 * x2`	$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 \cdot x_2)$	Expands automatically to both main effects plus interaction
`y ~ x + I(x^2)`	$\beta_0 + \beta_1 x + \beta_2 x^2$	`I()` shields arithmetic operators from formula expansion
`y ~ . - id`	All columns except `y` and `id`	Convenience shorthand excluding non-predictive identifier keys

3.2 Simple and Multiple Linear Regression (OLS Estimation)

The **Multiple Linear Regression Model** specifies that the conditional expectation $f(\mathbf{x}_i)$ is linear in its unknown parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^\top$:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

Where:

- β_0 is the intercept, representing expected y_i when all predictors equal zero.
- β_j $(j = 1, \dots, p)$ is the partial slope coefficient measuring the expected change in y_i associated with a one-unit increase in x_{ij}, holding all other predictors in the model constant (*ceteris paribus*).
- ε_i is the zero-mean stochastic error term for observation i.

Given a random sample of n observations, **Ordinary Least Squares (OLS)** selects coefficient estimates $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)^\top$ that minimise the **Residual Sum of Squares (RSS)** (also denoted SSE):

$$\text{RSS}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij}\right)^2$$

Where:

- $\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij}$ is the fitted (predicted) value for observation i.
- $e_i = y_i - \hat{y}_i$ is the sample residual for observation i.

DegreeLive

Differentiating RSS with respect to $\hat{\boldsymbol{\beta}}$ and setting the gradient to zero yields the matrix normal equations $\mathbf{X}^\top \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{y}$. Provided the design matrix $\mathbf{X}$ has full column rank (no exact multicollinearity among predictors), the unique OLS solution is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

Where:

- $\mathbf{X}$ is the $n \times (p+1)$ design matrix containing a leading column of ones and p predictor columns.
- $\mathbf{y}$ is the $n \times 1$ column vector of observed responses.

In the special case of **simple linear regression** ($p = 1$), the closed-form scalar normal equations simplify to:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Where:

- $\bar{x}$ and $\bar{y}$ are the sample means of x_i and y_i, respectively.
- $\hat{\beta}_1$ is the sample covariance of (x, y) divided by the sample variance of x.

Worked Numerical Example: OLS Fit of Melbourne Apartment Prices on Floor Area

Consider $n = 5$ two-bedroom apartment sales in Melbourne's inner east, recording internal floor area x_i (in m^2) and sale price y_i (in thousands of Australian dollars, $\$1,000\text{s AUD}$):

Unit i	Area x_i (m^2)	Price y_i (AUD '000)	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	60	540	-20	-160	400	3200
2	70	630	-10	-70	100	700
3	80	690	0	-10	0	0
4	90	790	10	90	100	900
5	100	850	20	150	400	3000
Sum / Mean	$\bar{x} = 80$	$\bar{y} = 700$	**0**	**0**	**1000**	**7800**

Step 1 (Compute Slope and Intercept):

$\hat{\beta}_1 = 7800/1000 = 7.80$ (AUD 'ooo per additional m^2).

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 700 - 7.80(80) = 700 - 624 = 76.00$ (AUD 'ooo).

DegreeLive

Thus, the fitted regression line is $\hat{y}_i = 76.00 + 7.80\,x_i$. Each extra square metre of floor area is associated with an expected $\$7,800$ increase in sale price. Notice that the intercept $\hat{\beta}_0 = 76.00$ corresponds to $x = 0$ m^2, which lies far outside the observed data domain $[60, 100]$ m^2 and carries no literal physical interpretation—an important warning against blind extrapolation.

Step 2 (Fitted Values, Residuals, and RSS):

- $i = 1$: $\hat{y}_1 = 76 + 7.8(60) = 544$, residual $e_1 = 540 - 544 = -4$.
- $i = 2$: $\hat{y}_2 = 76 + 7.8(70) = 622$, residual $e_2 = 630 - 622 = +8$.
- $i = 3$: $\hat{y}_3 = 76 + 7.8(80) = 700$, residual $e_3 = 690 - 700 = -10$.
- $i = 4$: $\hat{y}_4 = 76 + 7.8(90) = 778$, residual $e_4 = 790 - 778 = +12$.
- $i = 5$: $\hat{y}_5 = 76 + 7.8(100) = 856$, residual $e_5 = 850 - 856 = -6$.

Summing residuals verifies the algebraic OLS property $\sum e_i = -4 + 8 - 10 + 12 - 6 = 0$. Squaring and summing gives $\text{RSS} = (-4)^2 + 8^2 + (-10)^2 + 12^2 + (-6)^2 = 16 + 64 + 100 + 144 + 36 = 360$.

Residual Diagnostics and Influential Points

Fitting a regression model is incomplete without inspecting residual plots:

- **Residuals vs. Fitted Values:** A systematic U-shape or curvature indicates omitted non-linear terms (such as quadratic age or logarithmic income); a fanning cone shape indicates heteroskedasticity (non-constant error variance).
- **Normal Q-Q Plot:** Compares ordered standardised residuals against theoretical standard normal quantiles to check heavy tails or extreme skew that affect finite-sample t and F inference.
- **Leverage and Cook's Distance:** An observation with an extreme predictor value exerts high *leverage* h_{ii}, capable of tilting the fitted plane. Cook's distance D_i combines residual magnitude and leverage to quantify how much all fitted values shift if observation i is deleted.

Statistical Inference, Collinearity, and Confidence vs. Prediction Intervals

In inferential applications, analysts assess whether an estimated partial slope $\hat{\beta}_j$ differs reliably from zero using its estimated standard error $\text{SE}(\hat{\beta}_j)$ and t-statistic:

$$t_j = \frac{\hat{\beta}_j - 0}{\text{SE}(\hat{\beta}_j)} \sim t_{n-p-1}, \qquad \text{VIF}_j = \frac{1}{1 - R_j^2}$$

Where:

- t_j is the Student-t test statistic under the null hypothesis $H_0 : \beta_j = 0$ with $n - p - 1$ residual degrees of freedom.
- VIF_j is the Variance Inflation Factor for predictor j, where R_j^2 is the R^2 obtained by regressing x_j on all other predictors.

DegreeLive

When two or more predictors are strongly correlated (for example, including both gross household income and after-tax disposable income in the same regression), VIF_j exceeds 5 or 10, inflating $\mathrm{SE}(\hat{\beta}_j)$ and rendering individual t-tests unstable even when the overall model F-test is highly significant. Crucially, collinearity does not invalidate out-of-sample predictions within the observed predictor space, but it precludes clean ceteris paribus interpretation of individual coefficients.

Finally, when forecasting at a new predictor profile $\mathbf{x}_0$, analysts must distinguish between a **confidence interval** for the conditional mean $E[y \mid \mathbf{x}_0]$ (reflecting uncertainty solely in $\hat{\boldsymbol{\beta}}$) and a wider **prediction interval** for an individual future outcome $y_0 = \mathbf{x}_0^\top \boldsymbol{\beta} + \varepsilon_0$ (which adds the irreducible error variance $\hat{\sigma}^2$ to estimation uncertainty).

Non-Linear Functional Forms in Linear Models

Linearity in OLS refers to linearity in the *parameters* $\boldsymbol{\beta}$, not necessarily linearity in raw predictors. Analysts routinely accommodate curved economic relationships within OLS by transforming variables prior to fitting:

Specification	Equation Form	Interpretation of β_1	Typical Commerce Use Case
Log-Level	$\ln(y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i$	$100 \cdot \beta_1\%$ approx. change in y per 1-unit Δx	Wage or property price growth per year of tenure
Level-Log	$y_i = \beta_0 + \beta_1 \ln(x_i) + \varepsilon_i$	$(\beta_1/100)$ unit change in y per 1% Δx	Diminishing returns to advertising spend
Log-Log	$\ln(y_i) = \beta_0 + \beta_1 \ln(x_i) + \varepsilon_i$	Constant elasticity: $\%\Delta y / \%\Delta x = \beta_1$	Price elasticity of electricity or retail demand
Quadratic	$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$	Marginal effect $\beta_1 + 2\beta_2 x_i$; extremum at $x^* = -\beta_1/(2\beta_2)$	Inverted-U age-earnings profiles

For example, if an estimated log-log demand model for Victorian domestic airfares yields $\widehat{\ln(Q)} = 8.40 - 1.25\ln(P)$, a 1% fare increase predicts a 1.25% decline in ticket volume ($\hat{\beta}_1 = -1.25$ elasticity). Notice that when predicting back onto the original y scale from a log-linear model, $\exp(\widehat{\ln y})$ estimates the conditional *median* rather than the conditional mean because $E[e^{\varepsilon}] > 1$; multiplying by $\exp(\hat{\sigma}^2/2)$ (or Duan's smearing factor) corrects for retransformation bias.

3.3 Categorical Predictors, Reference Levels, and Interaction Effects

Commercial datasets contain abundant categorical features—Australian states, industry sectors, loan types, and education levels. Regression engines represent a categorical predictor with K distinct levels using $K - 1$ binary **dummy (indicator) variables**. One category is designated the **reference (baseline) level**, absorbed into the intercept β_0, while each dummy coefficient measures the expected outcome difference relative to that baseline.

DegreeLive

Including all K indicator columns alongside an intercept creates exact linear dependence ($\sum_{k=1}^{K} D_{ik} = 1$), making $\mathbf{X}^\top\mathbf{X}$ singular and non-invertible—known as the **dummy variable trap**. Formula interfaces prevent this automatically by dropping one level (alphabetically first by default, though analysts routinely relevel via `fct_relevel()` to anchor comparisons against a large or policy-relevant baseline such as `"NSW"` or `"Control"`).

Additive vs. Interaction Specifications

When a model combines a continuous predictor x and a binary indicator $D \in \{0, 1\}$:

- **Additive (Parallel-Slopes) Model (`y ~ x + D`):** Assumes the marginal effect of x on y is identical across both groups (β_1), while the vertical intercept shifts by β_2:

$$E[y_i \mid x_i, D_i] = \beta_0 + \beta_1 x_i + \beta_2 D_i$$

- **Interaction (Varying-Slopes) Model (`y ~ x * D`):** Introduces the product term $x_i D_i$, allowing both intercept and slope to differ across categories:

$$E[y_i \mid x_i, D_i] = \beta_0 + \beta_1 x_i + \beta_2 D_i + \beta_3 (x_i \cdot D_i)$$

Where:

- β_0 is the baseline intercept for the reference group ($D_i = 0$).
- β_1 is the slope of x_i specifically within the reference group ($D_i = 0$).
- β_2 is the difference in intercept between group $D_i = 1$ and the reference group when $x_i = 0$.
- β_3 is the difference in slope with respect to x_i between group $D_i = 1$ and the reference group, so the total slope for group $D_i = 1$ equals $\beta_1 + \beta_3$.

Whenever an interaction term $x_i \cdot D_i$ is included, the **hierarchy principle** dictates retaining both constituent main effects (x_i and D_i) in the formula, even if one main effect fails an individual t-test, so that the model remains invariant under linear shifts of x. Interaction terms also extend naturally to two continuous predictors ($x_1 \cdot x_2$), where the partial effect of x_1 becomes a linear function of x_2: $\partial E[y]/\partial x_1 = \beta_1 + \beta_3 x_2$.

Worked Example: Graduate Salaries Across Australian Industries with Interaction

Suppose a fitted model estimates annual salary y (in \$$k$ AUD) for Melbourne commerce and IT graduates using years of experience x and industry indicator $D_{\text{Tech}} = 1$ for Technology (reference $D_{\text{Tech}} = 0$ for Finance):

$$\widehat{\text{Salary}} = 72.0 + 4.5\,x - 4.0\,D_{\text{Tech}} + 2.5\,(x \cdot D_{\text{Tech}}).$$

- **Finance Graduates ($D_{\text{Tech}} = 0$):** $\widehat{\text{Salary}} = 72.0 + 4.5\,x$. Starting salary at $x = 0$ is $\$72,000$, rising by $\$4,500$ per year of experience. At $x = 4$ years, expected salary is $72.0 + 4.5(4) = \$90,000$.

degreelive

- **Technology Graduates ($D_{\text{Tech}} = 1$):** $\widehat{\text{Salary}} = (72.0 - 4.0) + (4.5 + 2.5)\,x = 68.0 + 7.0\,x$. Starting salary at $x = 0$ is $\$68,000$ ($\$4,000$ below Finance), yet the annual experience slope is steeper at $\$7,000$/year. At $x = 4$ years, expected salary is $68.0 + 7.0(4) = \$96,000$, overtaking Finance by $\$6,000$.

Notice the crossover point where both sectors pay identically: $72.0 + 4.5\,x = 68.0 + 7.0\,x \implies 2.5\,x = 4.0 \implies x = 1.6$ years. An additive model lacking the interaction term would have forced parallel lines and completely missed this divergence.

Joint F-Tests for Multi-Level Categorical Factors

When a categorical predictor possesses $K \geq 3$ levels—for example, four Australian states (`VIC`, `NSW`, `QLD`, `WA`) represented by $q = K - 1 = 3$ dummy variables—individual *t*-tests only compare each state pairwise against the omitted reference level. Testing whether state as a whole improves explanatory power requires a nested **partial F-test** comparing the unrestricted model (including all q dummies, yielding residual sum of squares RSS_U and p_U total slopes) against the restricted model (omitting all q dummies, yielding RSS_R):

$$F = \frac{(\text{RSS}_R - \text{RSS}_U)/q}{\text{RSS}_U/(n - p_U - 1)} \sim F_{q,\, n-p_U-1}$$

Where:

- RSS_R and RSS_U are the Residual Sum of Squares of the restricted and unrestricted models, respectively.
- q is the number of linear parameter restrictions tested simultaneously ($K - 1$ dummy variables).
- $n - p_U - 1$ is the residual degrees of freedom in the unrestricted specification.

Suppose $n = 104$ regional stores are modelled using floor area ($p_R = 1$), yielding $\text{RSS}_R = 12{,}400$. Adding $q = 3$ state dummies ($p_U = 4$) reduces residual error to $\text{RSS}_U = 10{,}000$. Then $F = \frac{(12400-10000)/3}{10000/(104-4-1)} = \frac{800}{101.01} \approx 7.92$, far exceeding standard $F_{3,99}$ critical values near 2.70, confirming significant interstate level differences independent of which baseline state is selected.

3.4 Goodness-of-Fit, Error Metrics, and Model Complexity

Evaluating how accurately a fitted regression model captures variation in y begins with the fundamental ANOVA sum-of-squares decomposition (valid whenever an intercept is included in OLS):

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

Where:

- SST (Total Sum of Squares) measures total sample variation in y around its sample mean $\bar{y}$.
- SSR (Regression or Explained Sum of Squares) measures variation captured by the fitted predictors.
- SSE (Error or Residual Sum of Squares, identical to RSS) measures unexplained residual variation.

From these components, four primary numerical assessment metrics are computed:

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = \frac{\text{SSR}}{\text{SST}}, \qquad R^2_{\text{adj}} = 1 - \frac{\text{SSE}/(n-p-1)}{\text{SST}/(n-1)}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \qquad \text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Where:

- $R^2 \in [0, 1]$ is the coefficient of determination (proportion of variance in y explained by the predictors).
- R^2_{adj} is Adjusted R^2, which penalises adding p predictors via degrees of freedom $(n - p - 1)$.
- RMSE is Root Mean Squared Error, expressed in the original units of y and penalising large errors quadratically.
- MAE is Mean Absolute Error, also in the original units of y, weighting all errors linearly and resisting isolated outliers.

Worked Numerical Calculation of Fit and Error Metrics

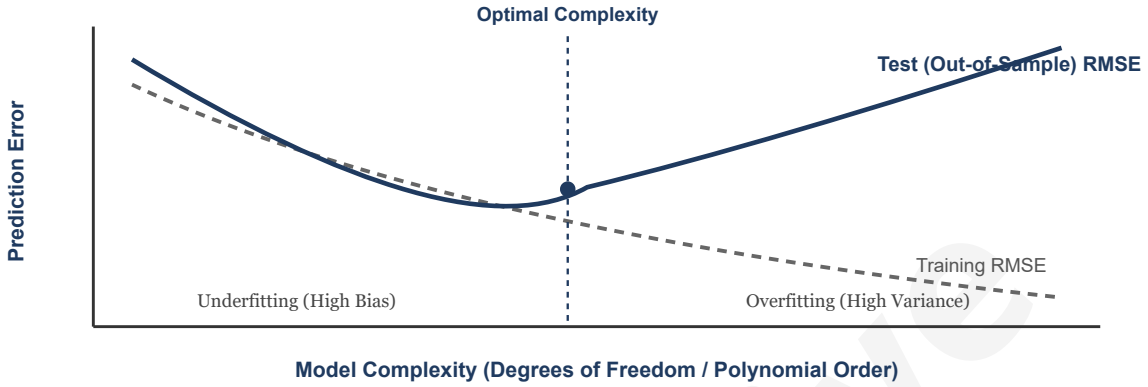
Returning to the Melbourne apartment OLS example from Section 3.2 ($n = 5$, $p = 1$ predictor, $\bar{y} = 700$, and $y = [540, 630, 690, 790, 850]$):

- **Total Sum of Squares:** $\text{SST} = (-160)^2 + (-70)^2 + (-10)^2 + 90^2 + 150^2 = 25600 + 4900 + 100 + 8100 + 22500 = 61200$.
- **Residual Sum of Squares:** Earlier we obtained $\text{SSE} = 360$, and residuals $e = [-4, +8, -10, +12, -6]$.
- **R^2 and Adjusted R^2:** $R^2 = 1 - (360/61200) = 1 - 0.005882 = 0.9941$ (99.41% of price variance explained). $R^2_{\text{adj}} = 1 - \frac{360/(5-1-1)}{61200/(5-1)} = 1 - \frac{120}{15300} = 1 - 0.007843 = 0.9922$.
- **RMSE and MAE:** $\text{RMSE} = \sqrt{360/5} = \sqrt{72} \approx 8.49$ (AUD 'ooo, or $\$8,485$). $\text{MAE} = (|-4| + |8| + |-10| + |12| + |-6|)/5 = 40/5 = 8.00$ (AUD 'ooo, or $\$8,000$).

Note that $\text{RMSE} \geq \text{MAE}$ always holds by Jensen's inequality, with the gap widening whenever large errors occur.

Bias-Variance Trade-Off and Model Complexity

Adding irrelevant predictors or high-degree polynomials never decreases in-sample R^2 and strictly shrinks training RMSE—yet eventually degrades performance on new data. Total expected prediction error at a test point decomposes into irreducible noise variance σ^2, squared estimation bias, and estimator variance:



Information criteria such as Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) formalise this parsimony penalty in-sample when comparing non-nested specifications on identical data, penalising log-likelihood by $2k$ or $k\ln(n)$ respectively, where k is the total number of estimated parameters.

3.5 Out-of-Sample Validation: Train-Test Splits and Cross-Validation

Because flexible models memorise idiosyncratic sample noise—inflating in-sample R^2 while failing on newly observed cases—predictive performance must be measured on held-out data that never participated in parameter estimation.

- **Holdout Train-Test Split:** Partitions the dataset randomly (commonly 75%–80% training set and 20%–25% test set) using a fixed random seed for reproducibility. For imbalanced binary outcomes, *stratified splitting* preserves identical class proportions across both partitions. For time-ordered commercial series (such as quarterly retail sales or ASX volatility), random splitting leaks future regime information into past training folds; instead, analysts apply *chronological (rolling-origin) splits* where training precedes testing strictly in time.
- **k-Fold Cross-Validation (CV):** A single holdout split can exhibit noisy test error on modest samples and wastes 20% of available data during training. k-fold CV partitions the training set into k roughly equal folds (typically $k = 5$ or $k = 10$). For each fold $m = 1, \ldots, k$, the model is trained on the remaining $k - 1$ folds and evaluated on fold m. Averaging fold errors yields the cross-validated RMSE:

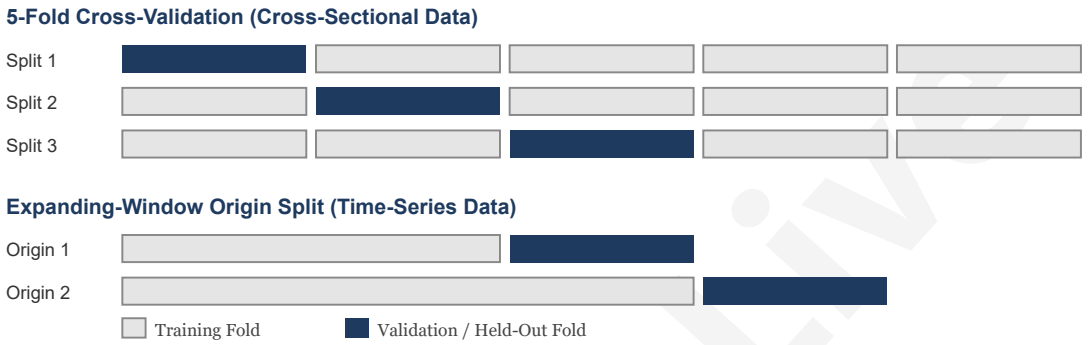
$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{m=1}^{k}\mathrm{RMSE}_m$$

DegreeLive

Where:

- k is the number of disjoint validation folds.
- RMSE_m is the Root Mean Squared Error computed exclusively on observations in held-out fold m.

Preventing Preprocessing Data Leakage: A pervasive methodological error occurs when analysts compute normalisation means/standard deviations, missing-value imputations, or PCA rotations across the *entire* dataset before performing the train-test split. Doing so leaks distributional information from test observations into training transformations, biasing validation error downward. All scaling, imputation, and feature selection parameters must be learned solely inside each training fold and applied deterministically to the corresponding validation fold.



Worked Numerical Example: Model Selection via 3-Fold Cross-Validation

Suppose an analyst compares two specifications predicting monthly store electricity bills across Victorian supermarkets using 3-fold CV:

- **Model A (Simple Linear):** Fold RMSEs $= [142.0, 138.0, 146.0]$ AUD; Training RMSE $= 135.0$ AUD.
- **Model B (Degree-5 Polynomial):** Fold RMSEs $= [168.0, 155.0, 172.0]$ AUD; Training RMSE $= 82.0$ AUD.

Computing average CV RMSE:

$$\text{CV}_{\text{Model A}} = (142.0 + 138.0 + 146.0)/3 = 426.0/3 = 142.0 \text{ AUD}.$$

$$\text{CV}_{\text{Model B}} = (168.0 + 155.0 + 172.0)/3 = 495.0/3 = 165.0 \text{ AUD}.$$

Although Model B achieves a substantially lower training RMSE ($82.0 < 135.0$), its cross-validated error jumps to 165.0 AUD due to high-variance overfitting. Model A ($142.0 < 165.0$) is selected as the superior generalising model before conducting one final evaluation on the untouched holdout test set.

3.6 Binary Outcomes and Logistic Regression

Many high-stakes commercial decisions involve binary outcomes $y_i \in \{0, 1\}$: whether a small-business borrower defaults, whether a customer churns next quarter, or whether a transaction is fraudulent. Applying OLS directly to a binary target—the **Linear Probability Model (LPM)**—violates key

probability axioms: linear predictions $\hat{y}_i = \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$ routinely escape the unit interval $[0, 1]$, binary residuals exhibit severe heteroskedasticity ($\text{Var}(y_i \mid \mathbf{x}_i) = p_i(1 - p_i)$), and linear marginal effects ignore saturation near 0% or 100%.

Logistic Regression resolves these defects by modelling the log-odds (logit) of $p_i = P(y_i = 1 \mid \mathbf{x}_i)$ as a linear function of predictors, mapping linear scores onto $(0, 1)$ via the S-shaped **logistic sigmoid function**:

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}}$$

Where:

- $p_i = P(y_i = 1 \mid \mathbf{x}_i) \in (0, 1)$ is the conditional probability of event occurrence.
- $p_i/(1 - p_i) \in (0, \infty)$ is the **odds** of the event occurring relative to not occurring.
- β_j represents the additive change in *log-odds* for a one-unit increase in x_{ij} holding other predictors constant, while e^{β_j} represents the multiplicative **odds ratio (OR)**.

Because closed-form OLS formulas do not apply to non-linear probability links, logistic regression coefficients are estimated via **Maximum Likelihood Estimation (MLE)**, which maximises the joint Bernoulli log-likelihood (equivalently minimising binary cross-entropy / log-loss):

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i)]$$

Where:

- $\ell(\boldsymbol{\beta})$ is the total sample log-likelihood evaluated at parameter vector $\boldsymbol{\beta}$.
- $y_i \in \{0, 1\}$ is the observed binary class label and p_i is the model-predicted probability.

Worked Numerical Example: Credit Default Probability and Classification Metrics

Suppose an Australian commercial lender fits a logistic regression predicting business loan default ($y = 1$) from debt-to-income ratio x (measured in multiples of annual EBITDA):

$$\text{logit}(\hat{p}) = -3.50 + 0.80\,x.$$

Step 1 (Odds Ratio Interpretation): A one-unit increase in debt-to-income ratio increases log-odds of default by 0.80, multiplying the odds of default by $e^{0.80} \approx 2.226$ (a 122.6% increase in odds).

DegreeLive

Step 2 (Predicted Probabilities):

- For a conservative borrower with $x = 2.0$: linear predictor $\eta = -3.50 + 0.80(2.0) = -1.90$, yielding $\hat{p} = \frac{1}{1+e^{1.90}} = \frac{1}{1+6.6859} \approx 0.1301$ (13.01% default probability).
- For a leveraged borrower with $x = 5.0$: linear predictor $\eta = -3.50 + 0.80(5.0) = +0.50$, yielding $\hat{p} = \frac{1}{1+e^{-0.50}} = \frac{1}{1+0.6065} \approx 0.6225$ (62.25% default probability).

Step 3 (Thresholding and Confusion Matrix Evaluation): Applying a decision threshold (for instance, classify $\hat{y}_i = 1$ if $\hat{p}_i \geq 0.50$) across 200 validation loans produces a 2×2 confusion matrix:

Predicted vs. Actual	Actual Default ($y = 1$)	Actual Non-Default ($y = 0$)	Row Total
Predicted Default ($\hat{y} = 1$)	True Positive (TP) = 32	False Positive (FP) = 18	50
Predicted Non-Default ($\hat{y} = 0$)	False Negative (FN) = 8	True Negative (TN) = 142	150
Column Total	**40**	**160**	**200**

Evaluating standard threshold-dependent classification metrics:

- **Accuracy:** $(\text{TP} + \text{TN})/N = (32 + 142)/200 = 174/200 = 87.0\%$. (Note that accuracy alone can mislead when classes are imbalanced: predicting "non-default" for everyone still achieves $160/200 = 80.0\%$ accuracy while missing all defaults.)
- **Sensitivity (Recall / True Positive Rate):** $\text{TP}/(\text{TP} + \text{FN}) = 32/40 = 80.0\%$ of actual defaults detected.
- **Specificity (True Negative Rate):** $\text{TN}/(\text{TN} + \text{FP}) = 142/160 = 88.75\%$ of sound borrowers correctly cleared.
- **Precision (Positive Predictive Value):** $\text{TP}/(\text{TP} + \text{FP}) = 32/50 = 64.0\%$ of flagged loans actually defaulting.

To evaluate discriminative ranking ability across *all* thresholds simultaneously without committing prematurely to 0.50, analysts plot the **Receiver Operating Characteristic (ROC)** curve (Sensitivity vs. $1 - \text{Specificity}$) and compute the **Area Under the Curve (ROC-AUC)**, where 0.50 reflects random guessing and 1.00 reflects perfect class separation.

Cost-Sensitive Thresholding and Probability Calibration

In operational credit risk or fraud detection, False Negatives and False Positives rarely carry equal financial penalties. Suppose approving a defaulting small-business loan (FN) costs $\$40,000$ in expected write-offs, whereas rejecting or manually auditing a sound applicant (FP) costs $\$2,000$ in foregone margin and underwriting staff time. Under Bayes decision theory, the expected-cost-minimising probability threshold p^* equals the cost ratio:

$$p^* = \frac{C_{\text{FP}}}{C_{\text{FP}} + C_{\text{FN}}} = \frac{2{,}000}{2{,}000 + 40{,}000} = \frac{2}{42} \approx 0.0476$$

Where:

- C_{FP} is the economic cost of a False Positive error.
- C_{FN} is the economic cost of a False Negative error.
- p^* is the optimal predicted probability cutoff above which an application is flagged for review or rejection.

In the validation matrix above at cutoff 0.50, total error cost equals $8(\$40,000) + 18(\$2,000) = \$320,000 + \$36,000 = \$356,000$. Lowering the cutoff closer to p^* increases False Positives moderately while sharply reducing costly False Negatives, demonstrating why threshold tuning must align with business economics rather than defaulting blindly to 0.50—provided predicted probabilities are well-calibrated against empirical default frequencies.

DegreeLive