FreeFellow

FORMULA SHEET

Exam MAS-II

CAS · Modern Actuarial Statistics II

151
FORMULAS

4
TOPICS

freefellow.org/cas-masii/formulas

Limited fluctuation full credibility standard (claim count)

$n_F = \left(\frac{z_{(1+P)/2}}{k}\right)^2$

where P = probability the estimate is within k of the expected value (Poisson claim counts). Example: P = 0.90, k = 0.05 gives $n_F = (1.645/0.05)^2 = 1082$.

Full credibility for pure premium

$n_F^{PP} = n_F^N \times (1 + CV_X^2)$

where n_F^N is the full credibility standard for claim count, X is severity per claim, $CV_X = \sigma_X/\mu_X$ is severity coefficient of variation. Pure premium adds severity-side variance to the claim-count standard.

Full credibility for aggregate losses (compound Poisson)

Same as pure premium because aggregate losses = N $\times$ X under compound Poisson with independence: $n_F^{Agg} = n_F^N(1 + CV_X^2)$.

The extra variance over count-only reflects severity variability. If severity is degenerate, formula reduces to the claim-count standard.

Partial credibility (limited fluctuation)

$Z = \sqrt{\frac{n}{n_F}}$

Capped at 1 (full credibility). Credibility-weighted estimate: $\hat{X} = Z\bar{X} + (1 - Z)M$ where M is the complement (e.g., manual rate, prior mean). Z grows as $\sqrt{n}$, not linearly.

Bühlmann credibility (greatest accuracy)

$Z = \frac{n}{n + k}$ where $k = \frac{\text{EPV}}{\text{VHM}}$

EPV = Expected Process Variance, VHM = Variance of Hypothetical Means.

Credibility-weighted estimate: $P_C = Z\bar{X} + (1 - Z)\mu$.

No upper-cap distortion (unlike LFC). Linear least-squares Bayes.

Expected process variance (EPV)

$\text{EPV} = E[\text{Var}(X|\Theta)]$

Average of conditional variances across risk types. Represents within-risk variation (pure noise that no amount of data can resolve). Computed as expected value of the variance for each parameter θ weighted by the prior on Θ.

Variance of hypothetical means (VHM)

$\text{VHM} = \text{Var}(E[X|\Theta])$

Variance of conditional means across risk types. Represents between-risk variation (the signal we want to estimate). Total variance: $\text{Var}(X) = \text{EPV} + \text{VHM}$ (law of total variance).

Bühlmann-Straub credibility (varying exposures)

For risks with exposure m_{ij} (year j of risk i): $Z_i = \frac{m_i}{m_i + k}$ where $m_i = \sum_j m_{ij}$, $k = \frac{v}{a}$, v = expected process variance per unit exposure, a = variance of hypothetical means. Allows different observation periods per risk.

Bayesian credibility (conjugate prior families)

Poisson-Gamma: count data X | λ is Poisson, λ is Gamma(α, β). Posterior is Gamma$(\alpha + \sum X_i, \beta + n)$. Posterior mean: $\frac{\alpha + \sum X_i}{\beta + n}$. Equivalent Bühlmann form: $Z = n/(n + \beta)$.

Bayesian credibility (Normal-Normal)

X | μ ~ N(μ, σ^2), μ ~ N(μ_0, τ^2). Posterior is Normal with mean $\frac{\sigma^2/n \cdot \mu_0 + \tau^2 \bar{X}}{\sigma^2/n + \tau^2}$.

Bühlmann form: $Z = \frac{n}{n + \sigma^2/\tau^2}$.

Empirical Bayes (non-parametric) credibility

Estimate EPV and VHM from data:

$\widehat{\text{EPV}} = \frac{1}{r}\sum_i s_i^2$ (avg within-risk sample variance)

$\widehat{\text{VHM}} = s_{\bar{X}}^2 - \widehat{\text{EPV}}/n$ (between-risk variance minus the within-risk contribution).

Empirical Bayes (semi-parametric) credibility

Assume process distribution is parametric (e.g., Poisson) and use the parametric variance form. For Poisson, $\text{Var}(X|\Theta) = \Theta$, so $\widehat{\text{EPV}} = \bar{X}$ (the grand mean). VHM estimated from cross-risk variation in observed rates. Less data-hungry than full non-parametric.

Linear least-squares (Bühlmann) credibility derivation

Choose Z to minimize $E[(Z\bar{X} + (1 - Z)\mu - E[X|\Theta])^2]$. Solution: $Z = \frac{a}{a + v/n}$

where a = VHM, v = expected process variance per observation, n = sample size. Same form as Bühlmann.

Credibility complement

When data are not fully credible (Z < 1), the complement (1 - Z) weight is given to a benchmark M:

$P_C = Z\bar{X} + (1 - Z)M$.

Common complements: overall manual rate, larger-class rate, prior-year experience adjusted for trend, present rates underlying current rates.

Loss-ratio credibility weighting

Indicated rate change $I = Z \cdot \frac{\text{Actual LR}}{\text{Target LR}} + (1 - Z) \cdot \text{Trended Current Rate Change}$.

Used in rate filings where indication blends company experience with a manual complement. Z derived per Bühlmann from EPV and VHM estimates.

Bühlmann credibility for severity

When estimating expected severity X (not pure premium), Bühlmann form uses claim count as exposure: $Z = \frac{N}{N + k}$ where N = observed claim count and k reflects EPV and VHM in the severity dimension. Pure-premium credibility combines count and severity.

Conjugate priors summary

Likelihood / Conjugate Prior / Posterior:

Poisson / Gamma / Gamma

Binomial / Beta / Beta

Normal (known σ) / Normal / Normal

Normal (unknown σ) / Normal-Inverse-Gamma / Normal-Inverse-Gamma

Exponential / Gamma / Gamma

Conjugacy gives closed-form posteriors; useful for sequential updating.

Credibility-weighted reserving (Cape Cod)

ELR (expected loss ratio) Cape Cod: $\widehat{\text{ELR}} = \frac{\sum C_i}{\sum(P_i \cdot f_i)}$ where C_i = paid losses, P_i = premium, f_i = expected reporting percent.

Applied to ultimate losses: $\text{Ult}_i = C_i + (1 - f_i) \cdot P_i \cdot \widehat{\text{ELR}}$.

Effect of trend and benefit changes on credibility

When experience period has different trend or benefit levels than the projection period, develop and trend experience to current level BEFORE applying credibility. Credibility weight Z is unchanged; the trended experience replaces $\bar{X}$.

Mahler credibility (changing risk parameters)

When risk parameter evolves over time (loss development, distribution drift), credibility on older data declines. Mahler's correlation: $\rho_{1,t}^2$ is the squared correlation between year 1 and year t hypothetical means. Older years receive lower weight in the credibility-weighted estimate.

Exposure-weighted risk average in Buhlmann-Straub

$\bar{X}_i = \frac{\sum_j m_{ij} X_{ij}}{m_i}$ — m_{ij} = exposure for risk i period j, X_{ij} = loss per unit exposure, m_i = total exposure

Bayesian predictive premium

$P_B = \int \mu(\theta)\, \pi(\theta \mid \mathbf{x})\, d\theta$ — μ(θ) = hypothetical mean, π(θ|x) = posterior density of Θ, x = observed data

Classical limited fluctuation reliability criterion

$\Pr(|\bar{X} - \mu| \le k\mu) \ge 1 - \alpha$ — X̄ = sample mean, μ = true mean, k = tolerance fraction, α = error probability

Poisson-Gamma posterior mean

$\frac{\alpha + \sum x_i}{\beta + n}$ — α, β = Gamma prior hyperparameters, Σx_i = sum of observed claim counts, n = number of exposure periods

Bayesian predictive mean

$P_B = \int \mu(\theta)\, \pi(\theta \mid \mathbf{x})\, d\theta$ — μ(θ) = hypothetical mean given θ, π(θ|x) = posterior density of θ given data x

Exposure-weighted risk average (Buhlmann-Straub)

$\bar{X}_i = \frac{\sum_j m_{ij} X_{ij}}{m_i}$ — m_ij = exposure in period j, X_ij = loss per exposure, m_i = total exposure for risk i

Credibility-weighted premium

$P_c = Z\bar{X} + (1 - Z)\mu$ — Z = credibility weight, X̄ = observed risk mean, μ = collective/prior/manual mean

Linear mixed model general form

$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}$

$\boldsymbol{\beta}$ = fixed-effect coefficients, $\mathbf{b}$ = random effects (with $\mathbf{b} \sim N(\mathbf{0}, \mathbf{G})$), $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{R})$.

Fixed effects vs random effects

Fixed: coefficients for ALL levels of interest (e.g., treatment vs control). Estimated directly.

Random: assume levels sampled from a population; estimate VARIANCE of the level effects rather than each effect.

Random intercept model

$Y_{ij} = (\beta_0 + b_{0i}) + \beta_1 x_{ij} + \varepsilon_{ij}$ where $b_{0i} \sim N(0, \sigma_b^2)$. Each group i gets its own intercept shifted by b_{0i} from the grand intercept. Common slope β_1 across groups.

Random slope model

$Y_{ij} = \beta_0 + (\beta_1 + b_{1i}) x_{ij} + \varepsilon_{ij}$ where $b_{1i} \sim N(0, \sigma_{b1}^2)$. Each group has its own slope. Random intercepts and slopes can co-occur with a correlation parameter ρ between them.

Intraclass correlation coefficient (ICC)

$$\text{ICC} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2}$$

Proportion of total variance attributable to between-group differences. ICC close to 0 = groups largely homogeneous. ICC close to 1 = strong clustering. High ICC justifies the random effect.

Best linear unbiased predictor (BLUP)

Predicted random effect: $\hat{\mathbf{b}} = \mathbf{G}\mathbf{Z}'(\mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R})^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$.

Shrinkage estimator: BLUP shrinks group means toward the grand mean. Equivalent to empirical Bayes / credibility weighting.

REML vs ML estimation

ML: maximizes full likelihood. Variance components biased downward in small samples (does not account for fixed-effect estimation).

REML (restricted ML): maximizes likelihood of contrasts orthogonal to fixed effects. Less biased variance estimates. Default in mixed-model software for variance-component inference.

Likelihood ratio test for variance components

Test $H_0 : \sigma_b^2 = 0$ against $H_a : \sigma_b^2 > 0$. Test statistic: $\Lambda = -2(\ell_0 - \ell_a)$.

Null distribution is mixture $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ (since $\sigma_b^2 = 0$ is on the boundary). Standard χ_1^2 p-value is conservative; divide by 2 for correct test.

LMM vs GEE (marginal models)

LMM: subject-specific (conditional) interpretation of β. Models the random effects explicitly.

GEE (generalized estimating equations): population-average (marginal) interpretation. Specifies only the mean and working correlation structure. Robust standard errors.

Variance component interpretation in actuarial context

Random territory effect: σ_b^2 measures cross-territory loss variation NOT explained by fixed-effect covariates. Higher σ_b^2 = more residual heterogeneity at the territory level = more value to including territory random effects (or a credibility step) in the rating plan.

BLUP shrinkage (credibility) factor for group i

$Z_i = \frac{n_i \sigma_u^2}{n_i \sigma_u^2 + \sigma_\varepsilon^2}$ — n_i = obs in group i, σ_u^2 = random-effect variance, σ_ε^2 = residual variance

Boundary-corrected p-value for variance-component LRT

$p = \frac{1}{2} P(\chi_1^2 \geq \text{LRT})$ — LRT = likelihood ratio statistic, p = tail probability under 50:50 χ_0^2/χ_1^2 mixture

Marginal distribution of LMM response

$y \sim N(X\beta,\ ZGZ^\top + R)$ — X = fixed-effect design, β = fixed effects, Z = random-effect design, G = Var(u), R = Var(ε)

Marginal covariance matrix V of LMM

$V = ZGZ^\top + R$ — Z = random-effect design, G = random-effect covariance, R = residual covariance; V is what REML and ML actually fit

Conditional variance of LMM response given random effects

$\text{Var}(y \mid u) = R$ — R = residual covariance matrix; conditional on u the LMM reduces to standard linear regression with covariance R

Conditional mean of LMM response given random effects

$E[y \mid u] = X\beta + Zu$ — X = fixed-effect design, β = fixed effects, Z = random-effect design, u = realized random-effect vector

BLUP shrinkage factor for hierarchical group j

$\lambda_j = \tau^2/(\tau^2 + \sigma^2/n_j)$ — τ² = between-group variance, σ² = within-group variance, n_j = group j sample size

Hierarchical partially pooled group rate

$\hat{\beta}_0 + \hat{u}_j = \hat{\beta}_0 + \lambda_j(\bar{y}_j - \hat{\beta}_0)$ — β₀ = grand intercept, λ_j = shrinkage factor, ȳ_j = group j sample mean

Boundary null distribution for variance-component LRT

$\text{LRT} \sim \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ under $H_0 : \tau^2 = 0$ — mixture because zero sits on parameter-space boundary

Marginal residuals in linear mixed model

$r^M = y - X\hat{\beta}$ — y = response vector, X = fixed-effect design matrix, β̂ = estimated fixed-effect coefficients

Likelihood ratio test statistic for nested LMMs

$LR = -2(\ell_R - \ell_F)$ — l_R = maximized log-likelihood of reduced model, l_F = maximized log-likelihood of full (nesting) model

Conditional residuals in linear mixed model

$r^C = y - X\hat{\beta} - Z\hat{b}$ — Z = random-effect design matrix, b̂ = predicted random effects (BLUPs), X, β̂ as in marginal residuals

Bias-variance decomposition

$E[(y - \hat{f}(x))^2] = (\text{Bias}[\hat{f}(x)])^2 + \text{Var}[\hat{f}(x)] + \sigma_\varepsilon^2$

Irreducible error σ_ε^2 cannot be reduced. Complex models lower bias but raise variance; simpler models do the opposite. Optimal complexity minimizes total MSE.

Training, validation, and test sets

Training: fit model parameters.

Validation: tune hyperparameters (e.g., λ in ridge, depth in trees). Select the model.

Test: estimate generalization error of the FINAL model on truly held-out data.

Common splits: 60/20/20 or 70/15/15. Never tune hyperparameters on test data (causes optimistic estimates).

k-fold cross-validation

Split data into k roughly-equal folds. For each fold i: train on the other k-1 folds, evaluate on fold i. CV error $= \frac{1}{k}\sum_{i=1}^{k} \text{MSE}_i$. Typical k = 5 or 10.

LOOCV: k = n. Low bias but high variance and computationally expensive (n model fits).

Bootstrap

Sample with replacement from the original n-row dataset to create B bootstrap samples (each of size n; ~63% of unique original rows). Compute statistic on each. Bootstrap standard error: $\widehat{\text{SE}}^* = \sqrt{\frac{1}{B-1}\sum(\hat{\theta}_b^* - \bar{\hat{\theta}}^*)^2}$.

OLS estimator (matrix form)

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

$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

Under Gauss-Markov (homoskedastic, uncorrelated errors), OLS is BLUE.

Multiple R-squared and adjusted R-squared

$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$. Never decreases when predictors are added.

$R^2_{adj} = 1 - \frac{SS_{res}/(n-p-1)}{SS_{tot}/(n-1)}$. Penalizes for extra predictors p. Adjusted R-squared can decrease when irrelevant variables are added.

AIC and BIC

$\text{AIC} = 2k - 2\ln L$ (k = parameters, L = max likelihood).

$\text{BIC} = k\ln n - 2\ln L$.

Lower is better. BIC has stiffer penalty than AIC for large n, so BIC favors smaller models. AIC targets predictive accuracy; BIC targets the true model under the assumption it is in the candidate set.

Ridge regression (L2 penalty)

$\hat{\boldsymbol{\beta}}_{\text{ridge}} = \arg\min \sum(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda\sum\beta_j^2$

Closed form: $\hat{\boldsymbol{\beta}}_{\text{ridge}} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$. Shrinks coefficients toward zero (never to zero).

Lasso regression (L1 penalty)

$\hat{\boldsymbol{\beta}}_{\text{lasso}} = \arg\min \sum(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda\sum|\beta_j|$

L1 penalty produces exact zeros: performs variable selection. No closed form (requires coordinate descent or LARS). Most useful when many predictors are irrelevant.

Elastic net

$\hat{\boldsymbol{\beta}} = \arg\min \sum(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda_1\sum|\beta_j| + \lambda_2\sum\beta_j^2$

Combines lasso variable selection with ridge stability. Useful when predictors are correlated (lasso would arbitrarily select one; elastic net keeps both with shrinkage).

Variable selection: forward, backward, stepwise

Forward: start empty; add the predictor that most improves fit (lowest p-value, highest R^2, lowest AIC).

Backward: start full; remove the least helpful predictor each step.

Stepwise: alternate forward and backward, allowing previously-added vars to be removed if a better predictor enters.

Logistic regression

$\Pr(Y = 1|\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{x}'\boldsymbol{\beta}}}$

Log-odds (logit): $\ln\frac{p}{1-p} = \mathbf{x}'\boldsymbol{\beta}$.

Fit by maximum likelihood. e^{β_j} is the odds ratio per unit change in x_j (others fixed).

Linear discriminant analysis (LDA)

Assumes each class is Gaussian with a SHARED covariance matrix. Decision boundary is linear.

$\delta_k(\mathbf{x}) = \mathbf{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \ln\pi_k$. Classify to k with highest δ_k.

Quadratic discriminant analysis (QDA)

Each class has its OWN covariance matrix $\boldsymbol{\Sigma}_k$. Decision boundary is quadratic.

More flexible than LDA but estimates more parameters; preferred when sample size is large and class covariances clearly differ. More variance, less bias.

k-Nearest Neighbors (KNN)

Classify a new point by majority vote among its k nearest training points (using Euclidean or another distance metric).

Low k: low bias, high variance (overfits).

High k: high bias, low variance.

No model fit. Sensitive to feature scaling (always standardize). Curse of dimensionality in high p.

Decision tree splitting criteria

Regression: minimize $\sum_R(y_i - \bar{y}_R)^2$ within each region.

Classification: Gini $= \sum_k p_k(1-p_k)$; Entropy $= -\sum_k p_k\ln p_k$; Classification error $= 1 - \max_k p_k$.

Gini and entropy are smooth; classification error is less sensitive. Trees built top-down by greedy recursive splitting.

Bagging (bootstrap aggregating)

Build B trees on bootstrap samples; average predictions (regression) or majority vote (classification).

Reduces variance versus a single tree. Each tree uses all p predictors. Out-of-bag (OOB) prediction = average of predictions from trees not built on a given observation; approximates CV error free of charge.

Random forest

Bagging plus random subset of predictors considered at each split (typical m = $\sqrt{p}$ for classification, m = p/3 for regression). Decorrelates trees, further reducing variance. OOB error is the standard out-of-sample estimate. Variable importance via mean decrease in Gini or permutation importance.

Boosting (AdaBoost and gradient boosting)

AdaBoost: sequentially fit weak learners (often stumps) to reweighted data (misclassified points up-weighted). Final classifier = weighted vote.

Gradient boosting: fit each new tree to the residuals (or negative gradient of loss) of the current ensemble.

Confusion matrix and classification metrics

TP / FP / FN / TN cells.

Accuracy = (TP + TN) / N

Sensitivity (recall, TPR) = TP / (TP + FN)

Specificity (TNR) = TN / (TN + FP)

Precision (PPV) = TP / (TP + FP)

F1 = harmonic mean of precision and recall = $\frac{2 \cdot P \cdot R}{P + R}$.

ROC curve and AUC

ROC plots sensitivity (TPR) vs 1 − specificity (FPR) across classification thresholds. Diagonal = random classifier.

AUC = area under ROC = probability that the model ranks a random positive higher than a random negative. AUC = 1 perfect; AUC = 0.5 random; AUC > 0.8 generally good.

Principal Components Analysis (PCA)

Find orthogonal directions of maximum variance in centered data $\mathbf{X}$. Loadings = eigenvectors of $\mathbf{X}'\mathbf{X}/(n-1)$. Variance explained by PC i = eigenvalue $\lambda_i / \sum_j \lambda_j$. Scores: project data onto loadings.

K-means clustering

Partition n points into K clusters minimizing within-cluster sum of squares $\sum_{k=1}^{K} \sum_{i \in C_k} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$. Algorithm: assign each point to nearest centroid; update centroids; repeat until convergence. Local minima possible (run with multiple random starts).

Hierarchical clustering linkage methods

Single linkage: distance between closest points across clusters (chains).

Complete linkage: distance between farthest points (compact, similar-size clusters).

Average linkage: mean pairwise distance.

Ward: minimize the increase in within-cluster variance from merging.

Silhouette score

For each point: $s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$ where a_i = mean distance to its own cluster, b_i = mean distance to nearest other cluster. Range -1 to $+1$. Closer to 1 = well-clustered; close to 0 = boundary point; negative = likely misclustered.

Naive Bayes classifier

Assumes predictors are independent given the class: $P(Y = k|\mathbf{x}) \propto P(Y = k) \prod_j P(x_j|Y = k)$. Despite the unrealistic independence assumption, performs well for text classification and other high-dimensional discrete problems.

Generalized linear models (GLM) actuarial workhorse

Three components: random component (exponential family distribution), systematic component ($\eta = \mathbf{x}'\boldsymbol{\beta}$), and link function g such that $g(\mu) = \eta$. Common: Poisson with log link for claim counts; Gamma with log link for severity; Tweedie for pure premium.

GLM deviance and goodness-of-fit

Deviance $D = -2(\ell_{\text{fitted}} - \ell_{\text{saturated}})$. For Gaussian, D = residual sum of squares. For Poisson, $D = 2\sum[y_i \ln(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)]$. Asymptotically χ^2_{n-p}. Likelihood ratio test: $\Delta D \sim \chi^2_{\Delta p}$ for nested model comparison.

Splines and polynomial regression

Polynomial: include $x, x^2, \ldots, x^d$ as predictors. Global; sensitive to boundary points.

Regression splines: piecewise polynomials joined at knots with continuity constraints. Cubic splines (degree 3) are common.

Natural splines: cubic in interior, linear past boundary knots (less variance at edges).

Generalized additive models (GAM)

$g(\mu) = \beta_0 + \sum_j f_j(x_j)$ where each f_j is a smooth function (spline) of one predictor. Captures nonlinearity without interactions. Interpretable: plot each f_j to see its effect. Fit by backfitting or penalized likelihood. Used in actuarial GLM extensions.

Gini index from Lorenz curve

$G = 2\int_0^1 (x - L(x))\,dx$ — G = Gini index, x = cumulative exposure share, L(x) = cumulative actual loss share at x

Root mean squared error (RMSE)

$\text{RMSE} = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2}$ — n = hold-out size, y_i = actual, ŷ_i = predicted for policy i

Double lift chart sort ratio

$R_i = \hat{y}_i^B / \hat{y}_i^A$ — R = sort key, ŷ^B = challenger prediction for policy i, ŷ^A = incumbent prediction for policy i

GLM deviance against saturated model

$D = 2\sum(\ell(y_i; y_i) - \ell(y_i; \hat{y}_i))$ — D = deviance, l(y;y) = saturated log-likelihood, l(y;ŷ) = fitted log-likelihood

Gradient boosting pseudo-residual

$r_{im} = -\left.\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right|_{f=F_{m-1}}$ — L = loss, y_i = target, F_{m-1} = current ensemble, r_{im} = stage-m pseudo-residual

Variance of bagged ensemble prediction

$\text{Var}(\bar{f}) = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$ — ρ = pairwise tree correlation, σ² = individual tree variance, B = number of trees

AdaBoost stage weight alpha

$\alpha_m = \log\big((1-\mathrm{err}_m)/\mathrm{err}_m\big)$ — err_m = weighted classification error at stage m, α_m = log-odds weight assigned to stump m

Bagged regression ensemble prediction

$\hat{f}_{\mathrm{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x)$ — B = number of bootstrap trees, $\hat{f}^{*b}$ = b-th tree fit on bootstrap sample, x = input vector

Gini coefficient from AUROC (binary classifier)

$\mathrm{Gini} = 2\cdot\mathrm{AUROC} - 1$ — AUROC = area under ROC curve; Gini ranges from 0 (random) to 1 (perfect).

AUROC as Mann-Whitney concordant-pair estimator

$\widehat{\mathrm{AUROC}} = \dfrac{\#\{\hat{p}_i > \hat{p}_j\} + 0.5\,\#\{\hat{p}_i = \hat{p}_j\}}{n_1 n_0}$ — i indexes positives, j negatives; n_1, n_0 counts; ties get 0.5 credit.

Lorenz-based Gini coefficient

$\mathrm{Gini} = 2\int_0^1 (x - L(x))\,dx$ — x = cumulative exposure share, L(x) = cumulative loss share at depth x.

Lift in a scored bin

$\mathrm{Lift}_b = \dfrac{\mathrm{positives}_b/n_b}{\mathrm{positives}_{\mathrm{total}}/N}$ — n_b = records in bin b, N = total records; positives counted in numerator bin and denominator overall.

KNN regression prediction (unweighted mean)

$\hat{f}(x_0) = \frac{1}{K}\sum_{i\in\mathcal{N}_K(x_0)} y_i$ — K = neighbor count, $\mathcal{N}_K(x_0)$ = K nearest training points to query x_0, y_i = neighbor response

Inverse-distance weighted KNN regression prediction

$\hat{f}(x_0) = \frac{\sum_{i\in\mathcal{N}_K} w_i y_i}{\sum_{i\in\mathcal{N}_K} w_i},\ w_i = \frac{1}{d(x_0,x_i)}$ — w_i = neighbor weight, $d(x_0,x_i)$ = distance from query to neighbor i, y_i = response

Euclidean distance between feature vectors

$d(x,y) = \sqrt{\sum_j (x_j - y_j)^2}$ — x_j, y_j = j-th standardized feature of points x and y, sum runs over all p features

KNN classifier posterior probability estimate

$\hat{P}(Y=j \mid X=x_0) = \frac{1}{K}\sum_{i\in\mathcal{N}_K(x_0)} \mathbf{1}\{y_i = j\}$ — K = neighbor count, $\mathcal{N}_K(x_0)$ = K nearest training points to query x_0, y_i = class label

Proportion of variance explained by a principal component

$\mathrm{PVE}_m = \lambda_m / \sum_{j=1}^{p}\lambda_j$ — λ_m = mth eigenvalue of sample covariance (or correlation) matrix, p = number of variables

First loading vector as variance maximizer

$\phi_1 = \arg\max_{\|\phi\|=1} \frac{1}{n}\sum_{i=1}^{n}(\phi^\top \tilde{x}_i)^2$ — ϕ = unit vector, $\tilde{x}_i$ = centered observation, n = sample size

Principal component score

$z_{im} = \phi_m^\top \tilde{x}_i = \sum_{j=1}^{p}\phi_{jm}\tilde{x}_{ij}$ — ϕ_m = mth unit loading vector, $\tilde{x}_i$ = centered observation, p = number of variables

Total variance identity in PCA

$\sum_{m=1}^{p}\lambda_m = \mathrm{tr}(S) = \sum_{j=1}^{p} s_j^2$ — λ_m = mth eigenvalue, S = sample covariance matrix, s_j^2 = sample variance of variable j

Proportion of variance explained by PC j

$\mathrm{PVE}_j = \lambda_j / \sum_{k=1}^{p}\lambda_k$ — λ_j = eigenvalue of PC j, p = number of variables

K-means total within-cluster sum of squares objective

$W = \sum_{k=1}^{K}\sum_{i\in C_k}\|x_i - \bar{x}_k\|^2$ — K = number of clusters, C_k = cluster k, x_i = observation, \bar{x}_k = cluster k centroid

Centroid linkage distance between two clusters

$d(A,B) = \|\bar{x}_A - \bar{x}_B\|$ — A, B = clusters, \bar{x}_A, \bar{x}_B = centroids of A and B, $\|\cdot\|$ = Euclidean norm

Eigenvalue from prcomp sdev output

$\lambda_j = \mathrm{sdev}_j^2$ — λ_j = variance captured by PC j, sdev_j = standard deviation of PC j from prcomp

Single linkage distance between two clusters

$d(A,B) = \min_{i\in A,\, j\in B} d(x_i, x_j)$ — A, B = clusters, x_i, x_j = points in A and B, d = pairwise distance metric

Unit-length constraint on PCA loadings

$\sum_{i=1}^{p}\phi_{ij}^2 = 1$ — ϕ_{ij} = loading of variable i on PC j, p = number of variables

PC score for a standardized observation

$z_{ij} = \sum_{k=1}^{p}\phi_{kj}(x_{ik} - \bar{x}_k)/s_k$ — ϕ_{kj} = loading, x_{ik} = raw value, $\bar{x}_k$ = mean, s_k = SD

Complete linkage distance between two clusters

$d(A,B) = \max_{i\in A,\, j\in B} d(x_i, x_j)$ — A, B = clusters, x_i, x_j = points in A and B, d = pairwise distance metric

Sigmoid activation function

$g(z) = \frac{1}{1+e^{-z}}$ — z = pre-activation, g(z) = predicted probability of class 1, range (0,1)

Permutation importance of feature j

$\mathrm{Imp}(j) = L(y, \hat{f}(x^{(\pi_j)})) - L(y, \hat{f}(x))$ — L = validation loss, π_j = permutation of column j, larger gap = more important variable

Partial dependence of prediction on variable j

$\mathrm{PD}_j(x_j) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}(x_j, x_{-j}^{(i)})$ — n = observations, x_{-j}^{(i)} = other features at observed values, average marginal effect

Feedforward neural network forward pass

$\hat{y} = g_L(W_L g_{L-1}(\cdots g_1(W_1 x + b_1)\cdots) + b_L)$ — x = input, W_k = weight matrix, b_k = bias, g_k = activation at layer k, L = number of layers

Regression tree residual sum of squares

$\mathrm{RSS}(T) = \sum_{m=1}^{M}\sum_{i\in R_m}(y_i - \hat{y}_{R_m})^2$ — R_m = terminal region m, $\hat{y}_{R_m}$ = regional mean, M = number of leaves

Regression tree training risk

$R(T) = \sum_{\text{leaves } t}\sum_{i\in t}(y_i - \bar{y}_t)^2$ — y_i = response in leaf t, ȳ_t = mean response in leaf t

Cost-complexity pruning criterion

$C_\alpha(T) = \mathrm{Loss}(T) + \alpha|T|$ — Loss(T) = tree RSS or weighted impurity, $|T|$ = number of terminal nodes, α = complexity parameter from CV

Gini index node impurity

$G_m = \sum_{k=1}^{K}\hat{p}_{mk}(1-\hat{p}_{mk}) = 1 - \sum_k \hat{p}_{mk}^2$ — $\hat{p}_{mk}$ = proportion of class k at node m, K = number of classes

One-standard-error rule threshold

$\mathrm{CV}(\alpha) \leq \mathrm{CV}_{\min} + \mathrm{SE}$ — CV(α) = K-fold CV error at penalty α, CV_min = minimum CV error, SE = standard error across folds; pick largest qualifying α

Weakest-link strength at internal node

$g(t) = \frac{R(t) - R(T_t)}{|T_t| - 1}$ — R(t) = risk if t is a single leaf, R(T_t) = risk of subtree rooted at t, |T_t| = leaves in that subtree

Cross-entropy node impurity

$D_m = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$ — $\hat{p}_{mk}$ = class k proportion at node m, K = classes, $0 \log 0 = 0$ by convention

Cost-complexity criterion for a subtree

$R_\alpha(T) = R(T) + \alpha |T|$ — R(T) = training risk (RSS or misclassification), α = complexity penalty per leaf, |T| = number of terminal nodes

Complete linkage inter-cluster distance

$d(A, B) = \max_{a \in A,\, b \in B} d(a, b)$ — A, B = clusters, d(a,b) = pairwise distance between points a in A and b in B

Average linkage inter-cluster distance

$d(A, B) = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a, b)$ — |A|, |B| = cluster sizes, d(a,b) = pairwise distance between a and b

K-means centroid update

$\bar{x}_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i$ — |C_k| = number of points in cluster k, x_i = data point assigned to cluster k, $\bar{x}_k$ = updated centroid

Within-cluster sum of squares (WCSS) for K-means

$\mathrm{WCSS} = \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - \bar{x}_k\|^2$ — K = number of clusters, C_k = cluster k, x_i = point i, $\bar{x}_k$ = centroid of cluster k

Stationarity (weak and strict)

Weak (covariance) stationarity: mean and variance constant over time; autocovariance $\gamma(t, t+h)$ depends only on lag h (not on t).

Strict stationarity: full joint distribution invariant under time shifts. Strict implies weak (with finite second moments) but not vice versa.

Autocovariance and autocorrelation functions

$\gamma_h = \mathrm{Cov}(X_t, X_{t+h})$

$\rho_h = \frac{\gamma_h}{\gamma_0}$ (autocorrelation function ACF).

$\rho_0 = 1$. For weakly stationary processes, ρ_h depends only on lag h. Sample ACF: $\hat{\rho}_h = \frac{\sum_{t=1}^{n-h}(X_t - \bar{X})(X_{t+h} - \bar{X})}{\sum_{t=1}^{n}(X_t - \bar{X})^2}$.

Partial autocorrelation function (PACF)

ϕ_{hh} = correlation between X_t and X_{t-h} after removing the linear effects of intermediate lags $X_{t-1}, \ldots, X_{t-h+1}$. For AR(p) process, PACF cuts off after lag p. For MA(q), PACF tails off. Used with ACF for Box-Jenkins model identification.

White noise and random walk

White noise: ε_t iid mean 0 variance σ^2. $\rho_h = 0$ for $h \neq 0$. Stationary.

Random walk: $X_t = X_{t-1} + \varepsilon_t$. Non-stationary; variance grows linearly with t. Differencing once produces white noise. Conditional forecast: $E[X_{t+h}|\mathcal{F}_t] = X_t$.

Autoregressive process AR(p)

$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t$

Stationary if all roots of $1 - \phi_1 z - \cdots - \phi_p z^p = 0$ lie OUTSIDE the unit circle. AR(1): $\rho_h = \phi^h$, so ACF decays geometrically. PACF cuts off at lag p.

Moving average process MA(q)

$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$

Always stationary (finite linear combination of white noise). Invertible if roots of $1 + \theta_1 z + \cdots + \theta_q z^q = 0$ lie OUTSIDE the unit circle. ACF cuts off at lag q; PACF tails off.

ARMA(p, q) and ARIMA(p, d, q)

ARMA: combines AR(p) and MA(q) for stationary series. ACF and PACF both tail off.

ARIMA: ARMA on the d-th difference $\Delta^d X_t$. Used when the original series is non-stationary but differencing produces stationarity. d = 1 for random walks; d = 2 for series with linear trends in slope.

Yule-Walker equations (AR estimation)

For AR(p): $\rho_h = \phi_1 \rho_{h-1} + \phi_2 \rho_{h-2} + \cdots + \phi_p \rho_{h-p}$ for $h \geq 1$.

Matrix form: $\boldsymbol{\rho} = \mathbf{R}\boldsymbol{\phi}$, solve $\hat{\boldsymbol{\phi}} = \mathbf{R}^{-1}\boldsymbol{\rho}$ using sample autocorrelations. Method of moments estimator.

Box-Jenkins methodology

1. Identification: plot the series, inspect ACF and PACF, difference if non-stationary, choose tentative (p, d, q).

2. Estimation: fit ARIMA by MLE or least squares.

3. Diagnostic checking: residuals should look like white noise (Ljung-Box test; ACF of residuals).

4.

Ljung-Box test for residual autocorrelation

$Q = n(n+2)\sum_{k=1}^{m} \frac{\hat{\rho}_k^2}{n-k}$

Under H_0 of white-noise residuals, $Q \sim \chi^2_{m-p-q}$ where p + q is the number of fitted ARMA parameters. Reject if Q exceeds critical value; indicates residual autocorrelation and model misspecification.

ARIMA forecast and forecast variance

One-step forecast: $\hat{X}_{n+1|n}$ is the conditional expectation given history. For AR(1) without drift: $\hat{X}_{n+h|n} = \phi^h X_n$.

Forecast error variance grows with horizon: for AR(1), $\mathrm{Var}(e_h) = \sigma^2(1 + \phi^2 + \cdots + \phi^{2(h-1)})$.

Seasonal ARIMA (SARIMA)

$\mathrm{ARIMA}(p,d,q)(P,D,Q)_s$. Adds seasonal AR, differencing, and MA terms at lag s (e.g., 12 for monthly).

Example: SARIMA(0, 1, 1)(0, 1, 1)$_{12}$ is the 'airline model' often used as a baseline for monthly series with both trend and seasonality.

AR(1) bounded h-step forecast variance

$\mathrm{Var}(\hat{Y}_{t+h|t}) = \sigma_\varepsilon^2(1 - \phi_1^{2h})/(1 - \phi_1^2)$ — σ_ε^2 = shock variance, ϕ_1 = AR coefficient, h = horizon

Long-run mean of stationary AR(p) process

$\mu = c/(1 - \phi_1 - \cdots - \phi_p)$ — c = printed intercept, ϕ_i = AR coefficients, μ = unconditional mean (not the intercept)

t-ratio for ARIMA coefficient significance

$t_i = \hat{\phi}_i / \mathrm{SE}(\hat{\phi}_i)$ — $\hat{\phi}_i$ = estimated coefficient, SE = standard error; significant when $|t_i| > 1.96$

AR(1) h-step ahead mean-corrected forecast

$\hat{Y}_{t+h|t} = \mu + \phi_1^h(Y_t - \mu)$ — μ = long-run mean, ϕ_1 = AR(1) coefficient, Y_t = last observation, h = horizon

Random walk with drift

$Y_t = \delta + Y_{t-1} + \varepsilon_t$ — δ = constant drift per period, Y_{t-1} = prior level, ε_t = iid shock with mean 0 and variance σ²

Random walk h-step forecast variance

$\mathrm{Var}(\hat{Y}_{n+h}) = h\sigma^2$ — h = forecast horizon, σ² = per-period shock variance; SE grows as $\sigma\sqrt{h}$

Deterministic linear trend regression

$Y_t = \beta_0 + \beta_1 t + \varepsilon_t$ — Y_t = series at time t, t = time index, β_1 = fixed per-period drift, ε_t = iid noise with mean 0 and variance σ²

Seasonal regression with dummy variables

$Y_t = \beta_0 + \beta_1 t + \sum_{j=1}^{s-1} \gamma_j D_{j,t} + \varepsilon_t$ — s = period, D_{j,t} = season-j indicator, γ_j = additive seasonal effect vs baseline

Invertibility condition for an MA(q) process

All roots of $\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q = 0$ satisfy $|z| > 1$ — θ_j = MA coefficients, q = MA order, z = complex root

Stationarity condition for an AR(p) process

All roots of $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p = 0$ satisfy $|z| > 1$ — φ_i = AR coefficients, p = AR order, z = complex root

White-noise confidence bands for sample ACF

$\pm 1.96/\sqrt{n}$ — n = sample size; sample autocorrelations inside the band are not significantly different from zero at the 5% level

Airline model for monthly time series

$(1-B)(1-B^{12})Y_t = (1+\theta_1 B)(1+\Theta_1 B^{12})\varepsilon_t$ — B = backshift, θ_1 = regular MA, Θ_1 = seasonal MA, ε_t = white noise

AR(1) h-step-ahead forecast

$\hat{Y}_{n+h} = \mu + \phi_1^h (Y_n - \mu)$ — μ = long-run mean, φ₁ = AR(1) coefficient, Y_n = most recent observation, h = horizon

AR(1) long-run mean

$\mu = \dfrac{c}{1-\phi_1}$ — c = intercept, φ₁ = AR(1) coefficient with |φ₁| < 1

MA(1) lag-1 autocorrelation

$\rho_1 = \dfrac{\theta_1}{1+\theta_1^2}$ — θ₁ = MA(1) coefficient with |θ₁| < 1 for invertibility; ρ_k = 0 for k ≥ 2

AR(1) unconditional variance

$\gamma_0 = \dfrac{\sigma^2}{1-\phi_1^2}$ — σ² = innovation variance, φ₁ = AR(1) coefficient with |φ₁| < 1