Stat 106 Cullen MacNeil

General Analytic & Sports Metrics

Expected Points (EP)

- Average points scored given a specific situation or state in a game.
- Calculated based on historical play-by-play data.

$$EP = \text{Shot Value} \times \text{Probability of Success}$$

- **Ex:** A 2-point shot with 53.6% success chance: $EP = 2 \times 0.536 = 1.072$

Expected Points Added (EPA)

- Measure of value added by a play relative to expectation.
- Formula:

$$EPA = \text{Actual Points} - \text{EP}$$

- **Ex:** Player makes a 3pt shot with 35% chance:

$$EP = 3 \times 0.35 = 1.05, \quad EPA = 3 - 1.05 = +1.95$$

Win Probability (WP) & Win Probability Added (WPA)

- **Win Probability (WP)**: Probability of winning given the current game state.
- **Win Probability Added (WPA)**: Change in WP from before to after a specific play.
- Formula:

$$WPA = WP_{\text{after}} - WP_{\text{before}}$$

- **Ex:** If WP rises from 0.03 to 0.05: $WPA = 0.05 - 0.03 = +0.02$
- High leverage situations (e.g., late-game) significantly affect WPA.

Core Analytical Ideas

Stickiness, Leverage, Clutch-ness

- **Stickiness**: Stability of performance metrics over time.
- **Leverage**: Situational importance; high-leverage moments significantly influence outcomes.
- **Clutch-ness**: Ability to perform well in high-leverage situations.

Luck & Mean Reversion

- **Luck**: Random deviations from expected performance metrics.
- **Mean Reversion**: Tendency for extreme performance to return toward average levels over time.

Shrinkage Estimates

- Estimates adjusted towards a prior or mean to reduce variance.
- Useful for stabilizing performance estimates, particularly with limited data.
- Prevents overfitting and extreme predictions.

Regression Models

Linear Regression

- Model form: $Y = \beta_0 + \beta_1 X + \epsilon$.
- Interpretation: Coefficient β_1 is the average change in Y per unit change in X.
- Assumptions: linearity, independence, homoscedasticity, normality.

Logistic Regression

- Used for binary outcomes; models log-odds of an event.
- Model form: $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$.
- Coefficients represent log-odds; e^{β} gives odds ratios.

Linear vs Logistic

- **Linear**: Continuous response variable.
- **Logistic**: Binary response variable.
- Choose based on outcome type (continuous vs. categorical).

Transformations & Interactions

Log Transformations

- Used to stabilize variance or handle skewed predictors.
- Often applied to variables with exponential or multiplicative effects.

Polynomial Transformations

- Capture non-linear trends by including squared/cubic terms.
- Example: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$.
- ϵ: Random error term; represents unexplained variation in Y. Assumed to be normally distributed with mean 0 and constant variance.

Interaction Effects

- The effect of one variable depends on the level of another.
- Modeled with product terms: $Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 (XZ) + \epsilon$.

Log-Odds to Probability (Logistic Regression)

- Model: $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$
- The left-hand side is the **log-odds** of success.
- To convert to **odds**, exponentiate:

$$\text{odds} = e^{\beta_0 + \beta_1 X}$$

- To convert odds to probability:

$$p = \frac{\text{odds}}{1 + \text{odds}}$$

- **Example:** If log-odds = 0.8:

$$\text{odds} = e^{0.8} \approx 2.23$$

$$p = \frac{2.23}{1 + 2.23} \approx \frac{2.23}{3.23} \approx 0.69$$

Note: Odds $= p/(1-p)$ and probability $=$ odds$/(1+$ odds$)$

Simulation & Resampling

Bootstrapping

- Repeated resampling from observed data to estimate sampling distribution.
- Useful for confidence intervals, bias estimation, and variance reduction.

Permutation Testing

- Hypothesis testing by reshuffling labels to determine significance.
- No assumptions about underlying distribution required.

Rating & Ranking Systems

Bradley-Terry Models

- Assigns each team a latent strength λ_i.
- Win probability:

$$P(A \text{ beats } B) = \frac{1}{1 + e^{\lambda_B - \lambda_A}}$$

- Fit via logistic regression on match results.

- Ties not natively handled — average models treating ties as wins and losses.

Quantitative BT Extension:

- Predicts numeric outcomes (e.g., point differential) instead of binary wins.

Elo Models

- Dynamic update system — ratings evolve after each match.
- Win probability:

$$P(\text{A beats B}) = \frac{1}{1 + 10^{(R_B - R_A)/400}}$$

- Rating update:

$$R_{\text{new}} = R_{\text{old}} + k(\text{Outcome} - \text{Expected})$$

- Outcomes: 1 = win, 0.5 = draw, 0 = loss. k = sensitivity.

Elo Example:

- $R_A = 1500$, $R_B = 1600$, $k = 50$, A wins.
- $P = \frac{1}{1+10^{(1600-1500)/400}} \approx 0.36$
- $R_A = 1500 + 50(1 - 0.36) = 1532$, $R_B = 1600 - 50(0.64) = 1568$

Bradley-Terry vs. Elo

- **BT:** Static strengths from full dataset, fitted once.
- **Elo:** Dynamic updates over time, match-by-match.
- **BT:** Assumes independence; **Elo:** incorporates time ordering.

KenPom Efficiency

- Basketball-specific metrics evaluating team efficiency.
- Offensive and defensive ratings adjusted for opponent strength and pace.
- Higher efficiency indicates stronger overall performance.

Predictive Modeling & Validation

Train-Test-Validation Split

- Data partitioned into subsets:
  - **Training**: Model fitting.
  - **Validation**: Model selection and tuning.
  - **Test**: Evaluate out-of-sample predictive performance.

Cross-Validation

- Technique to assess model predictive performance.
- Data repeatedly split into training and validation subsets.
- Commonly used form: k-fold CV, data split into k subsets.

Prediction, Overfitting, Complexity

- **Overfitting**: Model captures noise, poor generalization.
- Balance complexity (number of parameters) vs. prediction accuracy.
- Cross-validation helps identify appropriate model complexity.

Model Selection & Regularization

Model Selection: AIC & BIC

- Used in sequential variable selection to balance fit and complexity.
- **AIC:**

$$AIC = 2(p + 1) - 2\ln(\hat{L})$$

- **BIC:**

$$BIC = 2\ln(n) - 2\ln(\hat{L})$$

- Both penalize model complexity; BIC does so more heavily.
- AIC $\rightarrow$ better prediction focus; BIC $\rightarrow$ better for finding simpler models.

Penalized Regression Overview

- Penalized regression minimizes an objective of the form:

$$\text{SSE} + \lambda \sum_j \text{Penalty}(\beta_j)$$

- λ: regularization strength (chosen via cross-validation).
- Predictors typically standardized before fitting.

Ridge Regression

- Minimizes: $\text{SSE} + \lambda \sum_j \beta_j^2$
- Shrinks coefficients smoothly toward 0 (none exactly 0).
- Keeps all predictors; useful with many small effects.
- Has a closed-form solution.

LASSO Regression

- Minimizes: $\text{SSE} + \lambda \sum_j |\beta_j|$
- Shrinks some coefficients exactly to 0 $\rightarrow$ performs variable selection.
- Sparse, interpretable models; no closed-form solution.

Choosing Between Methods

- **Sequential selection:** Good with few strong predictors.
- **Ridge:** Good with many weak/moderate predictors.
- **LASSO:** Best with mix of useful and useless predictors.
- Compare models using validation/test MSE, not just training fit.

Penalty Comparison Table:

Method	Penalty	Zero Coefs?	Closed Form?
OLS	–	No	Yes
Ridge	$\lambda \sum \beta_j^2$	No	Yes
LASSO	$\lambda \sum \lvert\beta_j\rvert$	Yes	No

Advanced Predictive Techniques

Random Forests

- Ensemble of decision trees built on bootstrapped samples.
- Reduces variance and improves prediction by averaging outcomes.
- Each split considers random subset of predictors.

Hyperparameter Tuning

- Process of optimizing model parameters not learned from data.
- Common methods: Grid Search, Random Search, Cross-Validation.
- Helps prevent overfitting and improves predictive performance.

Additional Quick Reference

Regression Assumptions

OLS Linear Regression:

- OLS = Ordinary Least Squares
- Linearity, independence, homoscedasticity, normality of residuals.

Logistic Regression:

- Independence, linearity in logit scale.

Important Metrics & Formulas

Pythagorean Wins:

$$\text{Wins} = \frac{\text{Runs Scored}^2}{\text{Runs Scored}^2 + \text{Runs Against}^2} \times \text{Games}$$

Confidence Interval (mean):

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

Z-Test Statistic:

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Conditional Distribution of $Y \mid X$ (Regression)

- In linear regression:

$$Y \mid X \sim N(\mu = \beta_0 + \beta_1 X, \sigma^2)$$

- Use this to estimate probabilities of outcomes:

$$P(Y > a \mid X = x) = P\left(Z > \frac{a - \mu}{\sigma}\right)$$

  where $\mu = \beta_0 + \beta_1 x$
- ***Note:*** a is the outcome threshold of interest (e.g., $a = 0$ if you're asking the probability that a team wins a game).

Example:

$$\beta_0 = 7.135, \quad \beta_1 = 3.224, \quad \hat{\sigma} = 14.46$$

$$\begin{aligned} P(Y > 0 \mid X = 1) &= P\left(Z > \frac{0 - (7.135 + 3.224 \cdot 1)}{14.46}\right) \\ &= P\left(Z > \frac{-10.36}{14.46}\right) = P(Z > -0.72) \\ &= P(Z < 0.72) \approx 0.7642 \end{aligned}$$

Model Interpretation

- Linear regression coefficient: change in Y per unit change in X.
- Logistic regression coefficient: log-odds, e^β for odds ratio.
- Adjusted R^2: accounts for number of predictors.

Coefficient Plot Comparison

Model	Path Summary
Seq. Select	Jumps at steps; adds/removes variables.
Ridge	Coefs shrink smoothly; asymptote to 0.
LASSO	Some coefs shrink exactly to 0 (sparse model).

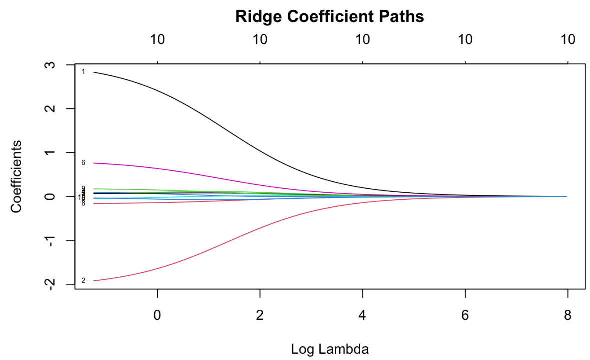

Figure 1: *

Ridge Coeff Paths: Coefficients shrink smoothly but do not hit zero.

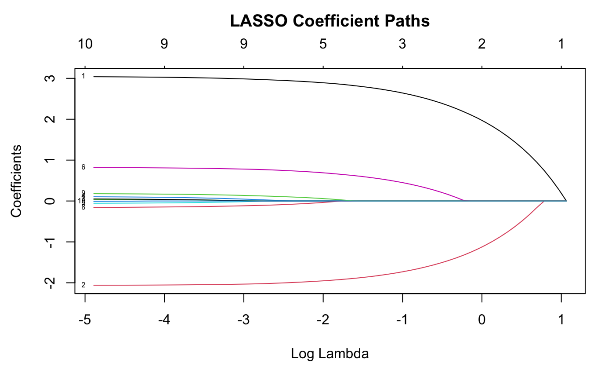

Figure 2: *

LASSO Coeff Paths: Some coefficients shrink exactly to 0 (variable selection).

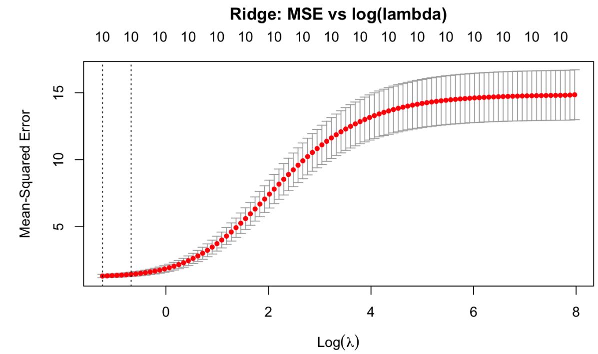

Figure 3: *

Ridge CV: MSE curve across log(λ); optimal λ balances bias-variance.

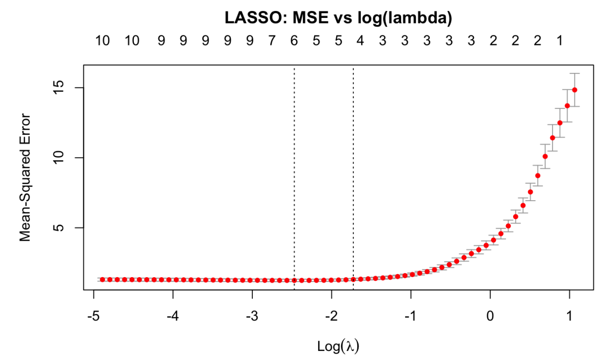

Figure 4: *

LASSO CV: MSE minimized where most unimportant coefficients are shrunk to 0.