Estimating team strength in the NFL

Mark E. Glickman Hal S. Stern

April 28, 2016

For NFL football fans, comparing team abilities and speculating on the outcomes of team matchups is one of the most enjoyable aspects of following the game. The comparison of teams can be a source of intense argument, with discussions commonly peppered with anecdotes and cherry-picked information to support the contention of one team being better than another. Various web sites promote a semblance of scientific rigor by computing power rankings, numerical team ratings, and other related quantities that are ostensibly reliable measures of team strength. But if interest centers on a serious study of NFL team strength, how can we appeal to a solid statistical foundation to measure ability and predict game outcomes?

This chapter covers basic statistical approaches to measure NFL team strength from historical game results. In Section 1, we formally lay out the problem and describe some basic procedures for estimating team strength. This is followed in Section 2 with a presentation of the statistical model for game outcomes along with common approaches to perform statistical inference of team strengths. The basic model is extended in Section 3 to account for the possibility that team strengths may be evolving over time. In Section 4, we explain the role of covariates in our models from the two previous sections. We then introduce some recent approaches to modeling game outcomes in Section 5. In Section 6 we demonstrate some of the methods in the chapter on NFL data, and we conclude the chapter in Section 7 with some discussion of future directions of NFL game outcome modeling.

1 Football team strength – the basics

From a brief examination online, it is clear that many methods are in use to rate football teams. Sites such as `http://masseyratings.com/cf/compare.htm` regularly report NCAA football team rankings from no fewer than 115 different methods of rating team strengths. As demonstrated on Massey's site, these methods produce rankings that are highly correlated despite the variety of methods implemented. Many of these approaches are principled and based on sound statistical reasoning, but plenty are ad hoc algorithms. We demonstrate in this section a particular principled approach to rate football teams.

The fundamental assumption for most models and rating methods for football team strengths is that the j-th team's ability, $j = 1, \ldots, J$, ($J = 32$ currently for the NFL) can be represented as a scalar parameter θ_j, and that game outcomes between teams j and k depend on the team strengths only through their difference, $\theta_j - \theta_k$. The main goal of a statistical or rating procedure for football is to estimate $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)$ from game results. The resulting estimates can then be used as the basis for measuring relative strengths of teams, ranking teams, and predicting outcomes of new games.

Consider a set of n games to analyze, and let y_i be an outcome of game i played between teams j_i and k_i (the choice of assigning team labels within a pair is arbitrary). Let $\boldsymbol{y}$ denote the vector of all n game outcomes, $(y_1, \ldots, y_n)$. The outcome y_i could be defined in a variety of ways. Common choices are the final score difference in the game from the perspective of team j_i, or the trinary outcome corresponding to 1 if team j_i wins, 0.5 if the game is a tie, and 0 if j_i loses. In the former case, a natural choice for estimating the θ_j is to minimize the sum of squared differences

$$\mathrm{SS}(\boldsymbol{\theta} \mid \boldsymbol{y}) = \sum_{i=1}^{n} \left(y_i - (\theta_{j_i} - \theta_{k_i})\right)^2. \tag{1}$$

One way to interpret θ_j is through team j's mean margin of victory plus the mean strength of the opponents. Larger θ_j correspond to better (higher relative scoring) teams and smaller θ_j correspond to worse teams. The objective function in (1) is optimized solely based on game outcomes. We consider incorporating covariate information, such as which team is playing on its home field, in Section 4. This basic setup for football team strength was described in Stern (1995).

The objective function in (1) can be more conveniently represented in matrix form. Let $\boldsymbol{X}$ be an $n \times J$ matrix which encodes the identities of teams competing across the n games in the following manner. The i-th row of X encodes information about game i. In row i, the j_i-th element is set to 1, the k_i-the element is set to -1, and the remaining elements are set to 0. An example of $\boldsymbol{X}$ might appear as

$$\begin{pmatrix} 1 & 0 & -1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & -1 & \cdots & 0 \\ -1 & 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & & & & & \ddots & \vdots \\ 0 & -1 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

where the first game involves team 1 playing team 3, the second game involves team 3 playing team 5, and so on. Then the n-vector formed from $\boldsymbol{X\theta}$ is the vector of ability parameter differences for each game, that is, $(\theta_{j_1} - \theta_{k_1}, \theta_{j_2} - \theta_{k_2}, \ldots, \theta_{j_n} - \theta_{k_n})$. In the example above,

$\boldsymbol{X\theta}$ is given by

$$\begin{pmatrix} 1 & 0 & -1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & -1 & \cdots & 0 \\ -1 & 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & & & & & \ddots & \vdots \\ 0 & -1 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_J \end{pmatrix} = \begin{pmatrix} \theta_1 - \theta_3 \\ \theta_3 - \theta_5 \\ \theta_5 - \theta_1 \\ \vdots \\ \theta_J - \theta_2 \end{pmatrix}$$

Equation (1) can then be re-expressed in matrix notation as

$$\mathrm{SS}(\boldsymbol{\theta} \mid \boldsymbol{y}) = (\boldsymbol{y} - \boldsymbol{X\theta})'(\boldsymbol{y} - \boldsymbol{X\theta}). \tag{2}$$

Recognizing that (2) is the sum of squared residuals for linear regression with design matrix $\boldsymbol{X}$, the unique least-squares estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is given by $\hat{\boldsymbol{\theta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ only if $\boldsymbol{X}'\boldsymbol{X}$ is invertible. The problem is that the columns of $\boldsymbol{X}$ are linearly dependent because the sum of elements across every row is 0, so $\boldsymbol{X}'\boldsymbol{X}$ is not of full rank and therefore not invertible.

Before describing two common approaches to address the linear dependence in the columns of $\boldsymbol{X}$, it is worth commenting on the structure of the $J \times J$ matrix $\boldsymbol{X}'\boldsymbol{X}$. Any least-squares estimate of $\boldsymbol{\theta}$, regardless of whether it is unique, satisfies the normal equations

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{X}'\boldsymbol{y}. \tag{3}$$

The j-th diagonal element counts the total number of times team j has competed among the n games in the data set. Each off-diagonal element, say the (j, k) element of $\boldsymbol{X}'\boldsymbol{X}$, is the negative of the number of times teams j and k have competed among the n games in the data set. The sum of elements in each row of $\boldsymbol{X}'\boldsymbol{X}$ therefore is 0, which also highlights that this square matrix is not invertible. Meanwhile, $\boldsymbol{X}'\boldsymbol{y}$ is the vector of J elements in which the j-th element is the sum of the score differences for the games team j won less the sum of the score differences for the games team j lost. So, for the j-th element of the left-hand and right-hand side of (3), we have

$$\left(\sum_{k\neq j} n_{jk}\right)\theta_j - \sum_{k\neq j} n_{jk}\theta_k = \sum_{i:\ j \text{ wins}} |y_i| - \sum_{i:\ j \text{ loses}} |y_i| \tag{4}$$

where n_{jk} is the number of times j and k have competed. Solving (4) for θ_j, we obtain

$$\theta_j = \frac{\sum_{k\neq j} n_{jk}\theta_k}{\sum_{k\neq j} n_{jk}} + \frac{\sum_{i:\ j \text{ wins}} |y_i| - \sum_{i:\ j \text{ loses}} |y_i|}{\sum_{k\neq j} n_{jk}}, \tag{5}$$

which demonstrates that θ_j can be understood as the average of the opponents' strengths plus the average margin of victory.

1.1 Incorporating linear constraints

An approach to address the linear dependence in $\boldsymbol{X}$ is to impose a linear constraint on $\boldsymbol{\theta}$. Such a constraint does not change the essential solution to the optimization problem; it only forces the selection of one solution among the infinite number in the unconstrained version. The most direct method to address the linear dependence is to re-express $\boldsymbol{X\theta}$ as $\boldsymbol{X}^*\boldsymbol{\theta}_{-J}$, where $\boldsymbol{\theta}_{-J}$ is the vector of length $J-1$ of team strengths with θ_J removed, and $\boldsymbol{X}^*$ is the $n \times (J-1)$ design matrix that accounts for the linear constraint assumed on $\boldsymbol{\theta}$. This can be accomplished by creating a $J \times (J-1)$ contrast matrix $\boldsymbol{W}$ such that $\boldsymbol{X}^* = \boldsymbol{XW}$. This construction also implies that

$$\boldsymbol{X}^*\boldsymbol{\theta}_{-J} = \boldsymbol{XW}\boldsymbol{\theta}_{-J} = \boldsymbol{X\theta}, \tag{6}$$

so that $\boldsymbol{W}\boldsymbol{\theta}_{-J} = \boldsymbol{\theta}$.

For example, if the linear constraint assumed for identifiability was to set $\theta_J = 0$, then $\boldsymbol{X}^*$ would be simply the first $J-1$ columns of $\boldsymbol{X}$. With $\boldsymbol{X}^* = \boldsymbol{X}\boldsymbol{W}$, this is equivalent to assuming

$$\boldsymbol{W} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}.$$

If instead the linear constraint assumed was $\sum_{j=1}^{J} \theta_j = 0$, then $\boldsymbol{X}$ can be replaced with $\boldsymbol{X}^* = \boldsymbol{X}\boldsymbol{W}$ where $\boldsymbol{W}$ is given by

$$\boldsymbol{W} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -1 & -1 & -1 & \cdots & -1 \end{pmatrix}.$$

This latter approach of obtaining the least-squares solution with the "sum contrast" linear constraint on $\boldsymbol{\theta}$ is sometimes credited in the context of college football to Kenneth Massey (Massey, 1997). In either case, the least-squares estimate $\hat{\boldsymbol{\theta}}_{-J}$ of $\boldsymbol{\theta}_{-J}$ is given by

$$\hat{\boldsymbol{\theta}}_{-J} = (\boldsymbol{X}^{*\prime}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*\prime}\boldsymbol{y}.$$

It is worth noting that the inclusion of a linear constraint does not itself guarantee a solution to the least-squares problem, let alone a unique solution. The extra requirement is that every team must be involved in at least one game in the data set so that the diagonal entries of $\boldsymbol{X}'\boldsymbol{X}$ are all positive. This is trivially satisfied by including team abilities in $\boldsymbol{\theta}$ only if they have game results in the data set.

1.2 Regularization

An alternative approach to address the linear dependence among the columns of $\boldsymbol{X}$ is to modify the objective function in (1) so that it penalizes certain types of potential choices of $\boldsymbol{\theta}$. This approach to penalizing an objective function is sometimes known as regularization. A common type of penalty increases the objective function based on the sum of squared differences of the θ_j from a fixed set of values. A general form of the penalized sum of squared differences can be expressed as

$$\mathrm{PSS}(\boldsymbol{\theta} \mid \boldsymbol{y}) = \sum_{i=1}^{n} (y_i - (\theta_{j_i} - \theta_{k_i}))^2 + \lambda \sum_{j=1}^{J} (\theta_j - \gamma_j)^2 \tag{7}$$

where $\lambda \geq 0$ and the γ_j are values set in advance of optimization. The addition of the second term in (7) ensures that the optimizing values of the θ_j are centered at the γ_j. The value of λ attenuates the degree to which the θ_j are "shrunk" to the γ_j. For example, when $\lambda \to \infty$, the first term involving the score differences plays no role in the optimization and (7) is optimized at $\theta_j = \gamma_j$ for all j. When λ is set to a small value, such as $\lambda = 0.0001$, the impact of the second term is merely to break the tie among the infinite set of equally optimal solutions that satisfy optimizing only the first term. The method of adding a penalty term of this type for least-squares estimation is commonly known as ridge regression (Hoerl and Kennard, 1970).

It is straightforward to show from differential calculus that $\hat{\boldsymbol{\theta}}$, the optimizing value of $\boldsymbol{\theta}$ in (7), is equivalent to computing

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X}'\boldsymbol{X} + \lambda \mathrm{I})^{-1}(\boldsymbol{X}'\boldsymbol{y} + \lambda\boldsymbol{\gamma}) \tag{8}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_J)$ and I is the identity matrix (of dimension J). Adding λ to each diagonal element of the rank $J-1$ matrix $\boldsymbol{X}'\boldsymbol{X}$ in (8) ensures the non-singularity of $(\boldsymbol{X}'\boldsymbol{X} + \lambda \mathrm{I})$ so that its inverse is uniquely defined.

The choice of values for λ and $\boldsymbol{\gamma}$ can be argued based on their role in (7). The parameter $\boldsymbol{\gamma}$ behaves as the center of $\boldsymbol{\theta}$. To ensure objectivity of the procedure, the elements of $\boldsymbol{\gamma}$ are commonly set to the same value. The specific choice of the common value is arbitrary, and for example can be set to 0. The choice of λ on the other hand can be understood as a tuning parameter in a regularization problem. A principled procedure to choose λ is through cross-validation. A particular form of cross-validation, namely K-fold cross-validation ($K = 10$ is a conventional choice), proceeds in the following manner.

- Divide the data set of n games at random into K mutually exclusive subsets. Let Y_k denote the k-th subset of data and let Y_{-k} denote its complement. Assume $\boldsymbol{\gamma}$ is already selected.
- Suppose λ^* is a candidate value of λ.
  - For $k = 1, \ldots, K$ determine $\hat{\boldsymbol{\theta}}$ from (8) with $\lambda = \lambda^*$ based on the data subset Y_{-k}. Then compute SS_k, the predicted sum of squared differences in (1) with $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ and restricting the sum to the validation data Y_k.
  - Compute $\mathrm{SS}^* = \sum_{k=1}^{K} \mathrm{SS}_k$. This is the sum of squared deviations of the predicted differences and the withheld observed score differences.

This process results in a predictive discrepancy for each candidate choice of λ. Optimization could proceed either by selecting a preset range of choices of λ, or preferably by using an automated optimization procedure such as the Nelder-Mead algorithm (Nelder and Mead, 1965) to determine the optimizing value of λ. Once the optimized value of λ is chosen through cross-validation, it is fixed and then $\hat{\boldsymbol{\theta}}$ can be determined from (8) using the entire data set.

An interesting special case of the penalized sums of squared differences is credited to Wesley Colley (Colley, 2002). While not described as the solution to a regularized sums of squares problem, his approach involves applying (8) for particular choices of the variables. In the Colley setup, $\lambda = 2$, and $\gamma_j = 0.5$ for all j. Furthermore, for each $i = 1, \ldots, n$, $y_i = 0.5$ if team j_i wins and $y_i = -0.5$ if team k_i wins. Games in which teams tie in Colley's original system are ignored, but in fact they can be included in the penalized sum of squares framework letting $y_i = 0$ for such games.

2 Probabilistically modeling football outcomes

The approaches described in Section 1 are limited in that they only provide point estimates of team abilities. Such methods do not acknowledge the uncertainty of the ability estimates, and do not provide a mechanism to forecast game prediction distributions. They also do not provide a means to assess the significance or importance of the magnitude of differences in teams' strengths. In this section, we consider common probability models for football outcomes as a function of team strengths which permit addressing the above issues.

2.1 Modeling wins versus losses

A common approach to modeling football game outcomes is to ignore the final scores and record only which team won, or whether the game resulted in a tie. For the development below, we will ignore the possibility of a tie, but will address this issue at the end of this section. Ties, which occur in football only when a game is tied at the end of regulation play and then if the game remains tied in overtime, occur typically about once per season. An argument for considering modeling game results as binary outcomes as opposed to actual scores is that teams are only incentivized to win games and not necessarily run up the score. For example, if a team consistently wins, but plays conservatively so that the team rarely wins by a large margin, then modeling the result of the game as a binary outcome would capture the team's winning tendencies without factoring in the margin of victory.

Assume teams j_i and k_i are involved in game i, with $i = 1, \ldots, n$. Let

$$y_i = \begin{cases} 1 & \text{if team } j_i \text{ wins} \\ 0 & \text{if team } k_i \text{ wins.} \end{cases} \tag{9}$$

We can then model the probability of a game result as

$$p_i = \mathrm{P}(y_i = 1) = F(\theta_{j_i} - \theta_{k_i}) \tag{10}$$

where θ_{j_i} and θ_{k_i} are the strengths of teams j_i and k_i, respectively, and where F is a monotonic function increasing from 0 to 1. The most conventional choices of F are

$$F(x) \equiv \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{w^2}{2}\right) dw \tag{11}$$

and

$$F(x) \equiv \frac{1}{1 + \exp(-x)}. \tag{12}$$

The model in (11) is known as the Thurstone-Mosteller model (Mosteller, 1951; Thurstone, 1927) and the model in (12) is known as the Bradley-Terry model (Bradley and Terry, 1952). These models are often presented in the form of the distribution of a binary response (win versus loss) conditional on the strength parameters, but both models have an alternative representation as latent variable models. Suppose Z_{j_i} and Z_{k_i} are independent unobserved continuous variables that can be understood as the latent performance for game i for teams j_i and k_i. If $Z_{j_i} > Z_{k_i}$, indicating that team j_i outperforms k_i, then $y_i = 1$ is observed; otherwise $y_i = 0$ is observed. Under the assumption that $Z_{j_i} \sim \mathrm{N}(\theta_{j_i}, 1/2)$ and $Z_{k_i} \sim \mathrm{N}(\theta_{k_i}, 1/2)$ independently, the distribution of $Z_{j_i} - Z_{k_i}$ is $\mathrm{N}(\theta_{j_i} - \theta_{k_i}, 1)$, and

$$\mathrm{P}(y_i = 1) = \mathrm{P}(Z_{j_i} - Z_{k_i} > 0) = \Phi(\theta_{j_i} - \theta_{k_i}) \tag{13}$$

which is the Thurstone-Mosteller model. Similarly, if Z_{j_i} follows a Gumbel distribution with location parameter θ_{j_i}, that is

$$p_Z(z) = \exp(-(z - \theta_{j_i} + e^{-(z-\theta_{j_i})})), \tag{14}$$

and if independently Z_{k_i} has a Gumbel distribution with location parameter θ_{k_i}, then the distribution of the difference $Z_{j_i} - Z_{k_i}$ can be shown to follow a logistic distribution with density function

$$p_{Z_j - Z_k}(z) = \frac{\exp(z - (\theta_{j_i} - \theta_{k_i}))}{(1 + \exp(z - (\theta_{j_i} - \theta_{k_i})))^2} \tag{15}$$

so that

$$\mathrm{P}(y_i = 1) = \mathrm{P}(Z_{j_i} - Z_{k_i} > 0) = \int_0^\infty p_{Z_j - Z_k}(z)\, dz = \frac{1}{1 + \exp(-(\theta_{j_i} - \theta_{k_i}))} \tag{16}$$

which is the Bradley-Terry model. The representations as latent variable models permit increased flexibility in modeling beyond the standard paired comparison models, such as modeling the latent performance as a function of game-specific covariates.

For the ordinary Bradley-Terry and Thurstone-Mosteller models, estimates of $\boldsymbol{\theta}$ can be obtained through maximum likelihood by maximizing

$$L(\boldsymbol{\theta}|\boldsymbol{y}) \propto \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \tag{17}$$

coupled with a linear constraint (as described in Section 1.1) on $\boldsymbol{\theta}$ to ensure identifiability. The covariance of the maximum likelihood estimates, $\hat{\boldsymbol{\theta}}$, that incorporate the linear constraint can be obtained through the usual method: evaluating the Hessian of the log-likelihood at the maximum likelihood estimate, and then inverting. Standard errors are the square roots of the diagonal entries of this matrix. See the discussion at the end of Section 2.2 for computing the covariance matrix of the full vector of J parameter estimates.

To ensure identifiability of $\boldsymbol{\theta}$ in (17), an additional condition, originally described by Ford (1957), must be met. For every possible partition of the J teams into two non-empty groups, some team in one group must have defeated a team in the other group. This condition can sometimes be difficult to satisfy in football, especially in the analysis of a partially completed regular season. The condition implies that no team has won every game or lost every game in the data set being analyzed. Thus, for example, analyzing the 2007 regular NFL season via maximum likelihood would not be possible because the Patriots won every game. In this situation, the likelihood in (17) would continue to increase as the ability parameter for the Patriots, θ_{NE}, increased.

Two other approaches are sometimes used in addressing concerns about uniquely identifying $\boldsymbol{\theta}$. One approach is to add ϵ to both y_i and $1-y_i$ in (17), where ϵ is a small positive constant set in advance. For example, the value $\epsilon = 0.2$ has been suggested (David, 1988). The effect of this change is two-fold: First, because each game is essentially a win plus a fractional or a partial loss, Ford's separability condition automatically holds. Second, the addition of ϵ to game outcomes has the effect of shrinking team abilities slightly towards the mean.

The other approach to uniquely identifying $\boldsymbol{\theta}$ is to explicitly shrink the elements of $\boldsymbol{\theta}$ to a constant via a penalty factor. For example, rather than maximizing the expression in (17), one could maximize

$$L(\boldsymbol{\theta}|\boldsymbol{y}) \propto \exp(-\frac{1}{2c}\boldsymbol{\theta}'\boldsymbol{\theta}) \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}, \tag{18}$$

which shrinks the $\boldsymbol{\theta}$ to 0. This approach is identical to the Bayesian problem of finding the mode of the posterior distribution of $\boldsymbol{\theta}$ for $\boldsymbol{\theta}$ having a normal prior distribution with mean 0 and variance c. The shrinkage in this approach and the previous one can actually have improved predictive performance.

While ties are rare in football, they should be acknowledged in the context of a binary response model. Models for ties in the context of the Thurstone-Mosteller model include Glenn and David (1960), and for the Bradley-Terry model include Davidson (1970) and Rao and Kupper (1967). Rather than explicitly modeling the probability of a tie, which is rare, a more efficient approach is to act as though ties do not occur but incorporate them into the likelihood in (17) in a principled manner. Viewing a tie as equivalent information to half a win and half a loss, a factor in the likelihood for a tie could appear as

$$p^{1/2}(1-p)^{1/2}$$

where p is the probability of a win for the first team. Maximizing the likelihood with terms for tied games can proceed normally. This approach was used in Glickman (1999).

2.2 Modeling score differences

A more common approach to inferring NFL team strength is to model score differences rather than binary outcomes (with ties). Despite the arguments in Section 2.1 in favor of modeling outcomes as binary outcomes, the flip side of the coin is that score differences may convey greater detail about team strengths. In particular, games won by large margins generally indicate better team strength than games won by, say, an overtime scoring event.

Even though score differences are integer-valued, the most common model for score differences is a normal distribution centered at the difference in team strength parameters. Choosing a normal distribution does not imply that the modeler believes that score differences are continuously distributed. Instead, this assumption is viewed as a continuous approximation to the true (discrete) probability distribution of score differences that might otherwise be too difficult to model in a convenient way. The normal model permits evaluating probabilities of ranges of score differences, including the probability one team would defeat another, using the normal distribution as a tool for probability calculations. Stern (1991) supports the use of a normal distribution around the expected score difference. While normal distributions are the most common continuous approximation for the true model of score differences, other continuous distributions could be explored, such as t distributions or other unimodal continuous and real-valued distributions.

Letting y_i be the score difference for game i involving teams j_i and k_i, the normal model assumes

$$y_i \sim \mathrm{N}(\theta_{j_i} - \theta_{k_i},\ \sigma^2) \tag{19}$$

where σ^2 is the residual variance. The model can be expressed more compactly as

$$\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\theta},\ \sigma^2 \mathrm{I}) \tag{20}$$

with $\boldsymbol{X}$ defined as in Section 1 and where I is the identity matrix of dimension n. Least-squares estimates $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ can be obtained accounting for the collinearity in $\boldsymbol{X}$ using the adjustment described in Section 1.1. Once $\hat{\boldsymbol{\theta}}$ is obtained, an unbiased estimate of σ^2 is given by

$$\hat{\sigma}^2 = \frac{1}{n-(J-1)}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}) \tag{21}$$

where J is the number of teams, and $J-1$ is the number of linearly independent columns of $\boldsymbol{X}$. The estimated covariance matrix for J-dimensional $\hat{\boldsymbol{\theta}}$ can be computed as

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}) = \widehat{\mathrm{Var}}(\boldsymbol{W}\hat{\boldsymbol{\theta}}_{-J}) = \boldsymbol{W}\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}_{-J})\boldsymbol{W}' = \hat{\sigma}^2\boldsymbol{W}(\boldsymbol{X}^{*\prime}\boldsymbol{X}^{*})^{-1}\boldsymbol{W}' \tag{22}$$

with $\boldsymbol{W}$ defined in Section 1.1. Standard errors are the square roots of the diagonal elements of $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})$ in (22). Note that the approach of obtaining the covariance matrix for $\hat{\boldsymbol{\theta}}$ through the application of contrast matrices also applies to binary response models described in Section 2.1.

3 Dynamic models for score differences

The approaches in Section 2 are appropriate if team abilities are not changing appreciably over time. We investigate in this section models for time-varying abilities. Because NFL teams typically make most of their player trades, acquire new players, make coaching staff changes, and lose players to retirement between seasons, it is natural to assume teams undergo ability changes mostly from season to season. The emphasis of this section is on state space models for changes in ability, though we consider alternative approaches at the end of this section.

Let t index a regular season, where $t = 1, \ldots, T$. We now assume that team j during season t has a strength θ_{jt} which may vary across seasons. As before, we assume that a score difference for game i during season t depends on current team strengths,

$$y_{it} \sim \mathrm{N}(\theta_{j_{it}t} - \theta_{k_{it}t},\ \sigma^2) \tag{23}$$

for teams j_{it} and k_{it}. We further assume that teams' abilities evolve through an autoregressive process. For team j and all seasons t, we assume

$$\theta_{j,t+1} \sim \mathrm{N}(\rho\theta_{jt}, \tau^2) \tag{24}$$

where τ^2 is an innovation variance and $|\rho| < 1$ is an autoregressive parameter. The restriction on ρ is so that the stochastic process on θ_{jt} is stationary. This particular model assumes that the θ_{jt} for fixed t are distributed around 0 so that the impact of ρ is to regress the θ_{jt} towards the mean over time. This model is an example of a normal linear state-space model (West, 1996), and is also an example of the Kalman filter (Kalman and Bucy, 1961; Meinhold and Singpurwalla, 1983). This approach has been applied to football score differences by Harville (1980) and Glickman and Stern (1998). The model can be extended to ensure that $\sum_{j=1}^{J} \theta_{jt} = 0$ for all t. For example, the model can be specified multivariately as

$$\boldsymbol{\theta}_{t+1} \sim \mathrm{N}(\rho\boldsymbol{\theta}_t,\ \tau^2 \left(\mathrm{I} - \frac{1}{J}\mathbf{1}\mathbf{1}'\right)) \tag{25}$$

where $\boldsymbol{\theta}_t$ is the vector of J team strengths at time t, I is the $J \times J$ identity matrix, and $\mathbf{1}$ is a J-vector with 1 as each element.

The likelihood for the set of T vectors of team strength parameters $(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_T)$ along with σ^2, τ^2 and ρ is an alternating product of multivariate normal densities for the game score differences, $\boldsymbol{y}_t$, and multivariate normal densities for the $\boldsymbol{\theta}_t$ that represent the innovations in strength over time. The full likelihood expression for the model in (23) and (24) can be written as

$$\begin{aligned} & L(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_T,\sigma^2,\tau^2,\rho \mid \boldsymbol{y}_1,\ldots,\boldsymbol{y}_T) \\ \propto \;\; & \mathrm{N}(\boldsymbol{y}_1|\boldsymbol{X}_1\boldsymbol{\theta}_1,\sigma^2\mathrm{I}_{n_1})\prod_{t=2}^{T}\mathrm{N}(\boldsymbol{\theta}_t|\rho\boldsymbol{\theta}_{t-1},\tau^2\mathrm{I}_J)\mathrm{N}(\boldsymbol{y}_t|\boldsymbol{X}_t\boldsymbol{\theta}_t,\sigma^2\mathrm{I}_{n_t}), \end{aligned} \tag{26}$$

where $\boldsymbol{X}_t$ is the design matrix of team pairings during season t, I_m is the identity matrix of dimension m, and the notation $\mathrm{N}(\cdot|\boldsymbol{\mu},\boldsymbol{\Sigma})$ is the multivariate normal density of its argument with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Inference for this model can be performed through maximum likelihood (Harville, 1980; Sallas and Harville, 1988), though our experience is that taking a Bayesian approach to inference permits greater flexibility in extending the basic model. To perform Bayesian inference for the model in (23) and (24), a prior distribution needs to be assumed on the two variance parameters, the autoregressive parameter, and the team strengths during the first season, $\boldsymbol{\theta}_1$. Particular choices are described in Glickman and Stern (1998). Inference may be implemented through Markov chain Monte Carlo (MCMC) simulation from the posterior distribution, a standard approach to inference in the Bayesian setting. The model can be implemented in a straightforward manner using standard Bayesian software that implements MCMC simulation, such as JAGS (Plummer, 2003).

The state-space model in (23) and (24) can be extended in a number of ways. For example, instead of assuming that team strength remains constant throughout a season, the model could be extended to allow for an autoregressive process for week-to-week variation in team abilities. Such a model was considered by Glickman and Stern (1998). Allowing for weekly variation can account for short-term effects such as injuries by franchise players. Furthermore, the evolution of team ability over time can be extended to more complex models including those applied in financial applications. Glickman (2001) considered a time component for the parameters in a model of binary football game outcomes that followed a stochastic volatility process (Kim et al., 1998; Jacquier et al., 2002). Instead of the autoregressive model in (24), the stochastic volatility extension assumes

$$\theta_{j,t+1} \sim \mathrm{N}(\rho\theta_{jt}, \tau^2_{t+1}) \tag{27}$$

along with a stochastic process on the τ^2_t,

$$\log \tau^2_{t+1} \sim \mathrm{N}(\log \tau^2_t, \omega^2). \tag{28}$$

This model can account for sudden changes in team ability that are not well captured by the normal state-space model. Consideration of other processes for time variation is an open research question.

Assuming a stochastic process on team strengths is not the only approach to account for changes in team strength. In the context of binary outcomes, Baker and McHale (2014) used barycentric rational interpolation, a particular type of scatterplot smoother. Their approach assumes that a team's strength follows a non-parametric regression as a function of time. While the focus of their work was on applications to binary outcomes, the approach can apply to score differences. In general, modeling team strengths as a non-parametric function of time is an under-explored avenue, and is another opportunity for further development.

4 The inclusion of covariates

The models considered thus far rely only on the identity of the teams involved in a game. Incorporating game-specific or team-specific covariate information has the potential for substantially improving game predictions. We illustrate in this section an approach to incorporating covariate information into our football outcome models.

Covariate information can be divided into two types: Factors that are endogenous to team strength, and those that are exogenous. Endogenous factors are ones that are intrinsic to describing a team's strength. These would include covariate data like summaries of team performance, player-level factors, or any components of performance that describes some aspect of a team's ability. Exogenous factors are external to a team's make-up, and might include location of the game (home versus away), weather-related information, time of the day a game was played, schedule-related factors (e.g., whether a team played after a "bye" week), and so on. The division of covariates into these two types has implications not so much for modeling, but for summarizing team strengths as we describe below.

Probably the most compelling exogenous covariate information is whether a team played on its home field. Various authors have shown that playing on the home field conveys roughly a 3-point advantage in the final score (Acker, 1997; Glickman and Stern, 1998; Harville, 1980). The basic model in (19) can be extended to include a home field term. With teams j_i and k_i competing in game i, assume

$$y_i \sim \mathrm{N}(\theta_{j_i} - \theta_{k_i} + h_i\delta,\ \sigma^2) \tag{29}$$

where

$$h_i = \begin{cases} -1 & \text{if team } k_i \text{ plays on its home field} \\ 1 & \text{if team } j_i \text{ plays on its home field} \\ 0 & \text{if the game is played on a neutral field} \end{cases} \qquad (30)$$

and the parameter δ is the effect of a team playing on its home field. Glickman and Stern (1998) investigated an extension of (29) in the context of the dynamic normal linear model in which each team had its own home-field advantage. Thus, rather than considering a single δ, their model was parameterized with $J = 32$ parameters, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_J)$.

A variety of covariates endogenous to overall team strength can be incorporated into the linear predictor for the score difference model. Typical information included in these models consists of season-to-date summaries (both offensive and defensive) of passing and rushing yards, rates of fumbles, and rates of interceptions. Ad hoc rules for including such measures early in a season have involved incorporating averages from the previous season. For game i involving teams j_i and k_i, it seems desirable to include separate variables for each measure specific to team j_i and k_i; for example, season-to-date average rushing yards for team j_i as one variable, and season-to-date average rushing yards for team k_i as a second variable. However, because the ordering of teams j_i and k_i is arbitrary, it would not be sensible to have effects of the two variables appear as the effect of team j_i and of team k_i. Instead, a more principled approach is to transform the two variables into their difference and their average, and use these two variables in the model. Thus, to include season-to-date average rushing yards for each team separately into a model, it would be more prudent to include the difference between the average rushing yards for team i and for team j, and the average rushing yards.

Let $\boldsymbol{u}_i$ be the vector of additional endogenous covariates to incorporate into the basic model, and let $\boldsymbol{v}_i$ be the vector of exogenous covariates. The $\boldsymbol{u}_i$ would include differences and averages of performance measures between the two teams involved in game i, and $\boldsymbol{v}_i$ would include h_i as a component if home field is recorded. Then the model in (29) can be extended to

$$y_i \sim \mathrm{N}(\theta_{j_i} - \theta_{k_i} + \boldsymbol{u}_i\boldsymbol{\beta}_u + \boldsymbol{v}_i\boldsymbol{\beta}_v,\ \sigma^2), \tag{31}$$

where $\boldsymbol{\beta}_u$ and $\boldsymbol{\beta}_v$ are the vector of effects of $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$, respectively. Because all the parameters are included as linear terms, equation (31) can be expressed more compactly in matrix form as

$$\boldsymbol{y} \sim \mathrm{N}((\boldsymbol{X} \mid \boldsymbol{U} \mid \boldsymbol{V})\begin{pmatrix} \boldsymbol{\theta} \\ \boldsymbol{\beta}_u \\ \boldsymbol{\beta}_v \end{pmatrix},\ \sigma^2) \tag{32}$$

where $\boldsymbol{U}$ is the matrix with rows $\boldsymbol{u}_i$ and $\boldsymbol{V}$ is the matrix with rows $\boldsymbol{v}_i$. Recognizing that (32) is a normal linear model, inference for the parameters can be obtained as usual through least-squares regression, accounting for the collinearity in $\boldsymbol{X}$ by replacing it with $\boldsymbol{X}^*$ as described in Section 1.1. This results in least-squares estimates $\hat{\boldsymbol{\theta}}_{-J}$, $\hat{\boldsymbol{\beta}}_u$ and $\hat{\boldsymbol{\beta}}_v$, as well as the covariance matrix of this collection of estimates. To obtain the covariance matrix of estimates for $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\beta}}_u$ and $\hat{\boldsymbol{\beta}}_v$, similarly to Equation (22), an analogous computation is required. Let

$$\hat{\boldsymbol{\eta}} = \begin{pmatrix} \hat{\boldsymbol{\theta}} \\ \hat{\boldsymbol{\beta}}_u \\ \hat{\boldsymbol{\beta}}_v \end{pmatrix}, \quad \text{and} \quad \hat{\boldsymbol{\eta}}_{-J} = \begin{pmatrix} \hat{\boldsymbol{\theta}}_{-J} \\ \hat{\boldsymbol{\beta}}_u \\ \hat{\boldsymbol{\beta}}_v \end{pmatrix}. \tag{33}$$

Furthermore, let

$$\tilde{\boldsymbol{W}} = \begin{pmatrix} \boldsymbol{W} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \mathrm{I}_u & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \mathrm{I}_v \end{pmatrix}, \tag{34}$$

where I_u and I_v are identity matrices having dimensions equal to the number of variables in $\boldsymbol{u}$ and $\boldsymbol{v}$, respectively. Then

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\eta}}) = \widehat{\mathrm{Var}}(\tilde{\boldsymbol{W}}\hat{\boldsymbol{\eta}}_{-J}) = \tilde{\boldsymbol{W}}\widehat{\mathrm{Var}}(\hat{\boldsymbol{\eta}}_{-J})\tilde{\boldsymbol{W}}', \tag{35}$$

and the standard errors can be determined as the square root of the diagonal entries.

Summarizing team strength in the presence of game-specific covariates is not as simple as summarizing inferences about the θ_j. Because the model in (32) adjusts the team parameters for endogenous covariate information, the θ_j have the interpretation of the effects beyond those of the covariates. If teams' covariate information correlates strongly with game score differences, then it is possible that the θ_j will be lower for stronger teams because the evidence of team strength will be encoded in $\boldsymbol{U}\boldsymbol{\beta}_u$. It would be more sensible to summarize inferences about

$$\theta_j + \boldsymbol{u}^*\boldsymbol{\beta}_u \tag{36}$$

rather than θ_j alone, where $\boldsymbol{u}^*$ is a vector of specific covariate values that represent the team and a typical opponent. Because the variables $\boldsymbol{v}$ are exogenous to team strength, they may be omitted when describing team strength. The quantity in (36) can be estimated by replacing the parameters with their least-squares estimates. Note that this interpretation is specific to a particular set of covariate values $\boldsymbol{u}$, and is not an overall description of the strength of team j.

This approach can be generalized slightly. Rather than describing team strength in (36) by selecting particular values of $\boldsymbol{u}^*$, one could consider a distribution of likely values of elements of $\boldsymbol{u}^*$ and summarize the distribution. For example, one could perform a Monte Carlo simulation of values of $\boldsymbol{u}^*$ from a specified joint distribution, and then summarize the Monte Carlo distribution of $\hat{\theta}_j + \boldsymbol{u}^*\hat{\boldsymbol{\beta}}_u$ for team j. To account for the variability of the estimates $\hat{\theta}_j$ and $\hat{\boldsymbol{\beta}}_u$, these parameters could also be simulated from their respective approximate normal distributions.

Another commonly used approach to address the difficulties with estimating team strength in the presence of endogenous covariates is simply to fit models that do not explicitly include strength parameters. Many authors, including David et al. (2011), Warner (2010) and Uzoma et al. (2015) analyze game outcomes as a function of covariate information but do not include team-specific parameters. This general approach assumes that team strength is sufficiently captured by the included covariates, and that game outcome predictions are sufficiently described by the information in these covariates and not on the team identities. A disadvantage of this approach is that the evolution of team strength is difficult to capture only through covariate information.

The state-space model in (23) and (24) can be extended to include covariate information in a straightforward manner. As above, the mean in (23) can include linear terms corresponding to covariate information, typically with coefficients that are non-dynamic. For example, the inclusion of a home field variable as in (30) would likely have an effect that would not be expected to vary over time. In the context of an MCMC posterior simulation, the non-dynamic parameters can be updated via conditional posterior sampling as a block within one iteration of MCMC. Glickman and Stern (1998) demonstrate model-fitting for the inclusion of non-dynamic home field advantage parameters (one per team) in their state-space model.

5 Other approaches

Models that include team-specific parameters θ_j are not the only approach to predicting NFL game scores and as a method to assess team abilities. In this section, we describe several other approaches that have been considered for game prediction.

An active area of NFL game prediction is to model game outcomes using algorithmic approaches, such as through machine learning methods. Some examples include modeling game outcomes through neural network models (David et al., 2011), nearest-neighbor models (Uzoma et al., 2015), and Gaussian process models (Warner, 2010). In all instances, game outcomes (whether as a categorical outcome, or as a score difference) are predicted from a possibly large set of covariate data. For example, in Warner (2010), 47 possible covariates were considered for game outcome prediction, including previous winning percentage, average points per game scored and allowed, total yards per game gained and allowed, rushing yards gained and allowed, temperature, and so on. Because team parameters are not included in these models, the distinction between endogenous and exogenous factors to team strength is not relevant.

In any of these algorithmic approaches, modeling game outcomes involves a combination of covariate selection along with constraining the model estimation from over-fitting, the latter typically through cross-validation. Because machine learning is a highly active area in the field of computer science and statistics, applications to NFL prediction and beating the Vegas point spread is likely to attract continued attention.

A novel approach to game score prediction was developed by Baker and McHale (2013). Their approach involved modeling the exact (discrete) game score outcome by viewing a game as a continuous time stochastic process composed of a series of birth processes.

Each "birth" event corresponds to one of ten different scoring events. These include for each of the home and away teams a touchdown without a conversion, a touchdown with a 1-point conversion, a touchdown with a 2-point conversion, a field goal, and a safety. The time to occurrence between events is modeled as a proportional hazards model that depends on team characteristics, home field advantage, and an effect that depends on the type of scoring event. In their analysis, the authors worked with only final game scores, so they made simplifying assumptions in order to account for the multitude of ways in which the final score could be achieved from the point process. However, data on the exact timing of scoring events in NFL games is more easily accessible today, so analyses that recognize actual timings could be performed making use of some of the ideas in their paper. In particular, more precise modeling of inter-arrival times beyond proportional hazards may be worth exploring.

It is worth noting that neither of the above methods directly address measuring the ability of teams, but rather focus on modeling game outcomes. Ability measures could be derived quantities from these analyses, such as inserting team specific covariate data as a proxy for estimating a team strength parameter. A potential area for future work is to extend the above approaches to include team strengths as explicitly defined parameters, though the endogenous factors would need to be addressed by an approach analogous to that presented in Equation (36) to address confounding.

6 Application to NFL game data

In this section, we demonstrate the application of some of the methods described in the chapter to NFL football outcomes. Game outcomes can be scraped directly from `www.nfl.com`, though we obtained our data from `http://www.aussportsbetting.com/data/` which provided all NFL game data from the 2006 season onward conveniently in one spreadsheet.

We first demonstrate least-squares modeling from Section 2 to game outcome data from the 2014-2015 regular season. With $J = 32$ teams each playing 16 games during the regular season, the number of games in our analyses is 256. We fit a least-squares model as in (19) with the inclusion of a home-field advantage parameter as in (29). Our data included information on which team played on their home field. Out of the 256 games, four were played on a neutral field so that $h_i = 0$ for these four games.

The results of the model fits are summarized in Table 1 with the teams rank ordered according to their fitted strengths. We incorporated the linear constraint $\sum_{j=1}^{J} \theta_j = 0$ for parameter identifiability using the methods in Section 2.2, and determined the $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})$ using the computation in (22). The standard error of any $\hat{\theta}_j$ to two decimal places was 3.58. The standard error of the home field parameter was 0.886. The residual standard deviation, $\hat{\sigma}$, was about 14, suggesting large variation in score differences around the mean. Stern (1991) obtained a similar result.

The home field advantage parameter was estimated as 2.443, suggesting that playing at home versus away conveyed roughly a 4.8 point advantage in the 2014-2015 regular season. The range in strengths spanned from -11.8 (Titans) to 10.9 (Patriots), suggesting that on a neutral field the Patriots would outscore the Titans on average by $10.9 - (-11.8) = 22.7$ points. The order of strength estimates in Table 1 arguably has face validity. The teams in the Super Bowl (the Patriots and the Seahawks) were two of the top three teams listed, and teams that had poor regular season records were at the bottom of the rank order.

The variance of $\hat{\theta}_j - \hat{\theta}_k$ for teams j and k in the above analyses can differ due to the different covariances of the estimates. For example, because the Patriots and the Jets, who are in the same division, play twice during the regular season, the estimated variance of the difference in strength is lower than most pairs of teams. From the model fit, we have

$$\begin{aligned}
\widehat{\text{Var}}(\hat{\theta}_{\text{NE}} - \hat{\theta}_{\text{NYJ}}) &= \widehat{\text{Var}}(\hat{\theta}_{\text{NE}}) + \widehat{\text{Var}}(\hat{\theta}_{\text{NYJ}}) - 2\widehat{\text{Cov}}(\hat{\theta}_{\text{NE}}, \hat{\theta}_{\text{NYJ}}) \\
&= 12.83 + 12.83 - 2(1.80) = 22.06
\end{aligned}$$

which corresponds to a standard error of $\sqrt{22.06} = 4.70$. In contrast, the same calculation to compute the estimated variance of the difference in strength between the Patriots and the Seahawks, who did not play during the regular season, is

$$\begin{aligned}
\widehat{\text{Var}}(\hat{\theta}_{\text{NE}} - \hat{\theta}_{\text{SEA}}) &= \widehat{\text{Var}}(\hat{\theta}_{\text{NE}}) + \widehat{\text{Var}}(\hat{\theta}_{\text{SEA}}) - 2\widehat{\text{Cov}}(\hat{\theta}_{\text{NE}}, \hat{\theta}_{\text{SEA}}) \\
&= 12.83 + 12.83 - 2(-1.04) = 27.74
\end{aligned}$$

corresponding to a larger standard error of $\sqrt{27.74} = 5.27$. These standard errors are not appreciably different, and for the NFL they do not vary considerably. However, in a larger less-connected league (e.g., NCAA college football) where teams mostly compete against each other in the same division one might expect larger variation in the standard error of the team strength differences between divisions.

We also fit a dynamic model of team strengths described in Section 3 that includes a home-field advantage parameter as a non-dynamic component of the model. Our model assumed that team j had a strength parameter θ_{jt} during season t, where $t = 2006, 2007, \ldots, 2014$, that evolved according to an autoregressive process presented in (24). We fit our model via MCMC simulation using the Bayesian software JAGS called from within R (R Core Team, 2015) using the Rjags package. We ran two parallel chains for 30,000 iterations, including a burn-in period of 10,000 iterations. Convergence diagnostics (Gelman et al., 2014) revealed that the chains converged, and parameter summaries were based on every 5th sampled value beyond the burn-in period to lessen the autocorrelation among successive draws but also to save on space resources. Thus 8,000 simulated values from the posterior distribution for each parameter were saved based on the MCMC simulation.

Summaries of the team strengths for the 2014-2015 season based on the dynamic model, along with other model parameters, are summarized in Table 2. The team strengths are ordered according to the posterior means. The posterior mean strengths range from 8.42 (Seahawks) to –9.41 (Jaguars). The posterior mean of τ, the innovation standard deviation, equal to 4.28 indicates a non-trivial amount of change in team abilities from season to season, though with autoregressive parameter $\rho \approx 0.67$ top teams and bottom teams are regressing substantially towards the mean team ability.

The comparison of the results in Table 2 to those in Table 1 reveal some interesting differences. First, the home field advantage is inferred to be nearly the same for both sets of models, about 2.4 points relative to playing on a neutral field. The residual standard deviations are also of similar magnitude, with the dynamic model estimate of 13.1 compared to the least-squares estimate of 14.0. The Bayesian equivalent of the standard errors of the team strengths are all around 2.66, which is smaller than the corresponding value of 3.58 in the least-squares fit. The lower standard error is explainable by the use of historical game outcome data to aid in inferring 2014-2015 team strengths. The spread in estimated team strengths is slightly narrower for the dynamic model. This may again be due to incorporating previous seasons' game results, which recognize that teams' strengths regress towards the mean between seasons. The order of team strengths for the 2014-2015 season is similar between the two analyses (e.g., the top five and bottom five teams are the same, though in different orders), but some notable exceptions exist. For example, the Saints in the dynamic model are estimated to be nearly 3 points higher in strength than in the least-square fit. This may reflect the much better performances by the Saints in seasons previous to 2014 which factor more directly into the dynamic model. Similarly, the Bucs are inferred by the dynamic model to be a substantially better team at the end of the 2014-2015 season than through the least-squares estimates. Again, this may be due to the Bucs having a particularly poor performance during the 2014-2015 season. The Seahawks are inferred to be slightly better than the Patriots in the dynamic model compared to the least-squares model, and this may be carry-over from the superb 2013-2014 season in which the Seahawks won the Super Bowl.

Trajectories of team strengths over time can be understood by examining the posterior mean strengths across multiple seasons. The estimated strengths from all 32 teams is displayed in Figure 1 organized by division (each consisting of four teams who are in competition with each other to earn a playoff spot). Several features of these plots are worthy of mention. The AFC East plot demonstrates how dominant the Patriots have been over the last decade in their division, maintaining a posterior mean strength higher than the Bills, Jets and Dolphins all nine years. Only the Green Bay Packers come close to demonstrating this level of dominance in their division. Several teams have shown particularly quick improvement over the last few years. The Seahawks, for example, were a sub-par team between 2008-2010, but with the hiring of Pete Carroll in 2010 and the signing of quarterback Russell Wilson in 2012, they have improved to be the best in their division by the 2013-14 and 2014-15 seasons. Similarly, the Denver Broncos were a below-average team through the end of the 2011-12 season. With the acquisition of Peyton Manning from the Colts, the Broncos from 2012 onward were the strongest team in the AFC West, and eventually won the Super Bowl at the end of the 2015-16 season.

A closer examination of the Broncos team strength over time illustrates the comparison between inferences from the dynamic linear model and the least-squares estimates obtained on a season-by-season basis. In Figure 2, 95% central posterior intervals and 95% confidence intervals for $\theta_{DEN,t}$ are plotted side by side for each season. Because the dynamic model pools information across all seasons, inferences about team strengths in any given year

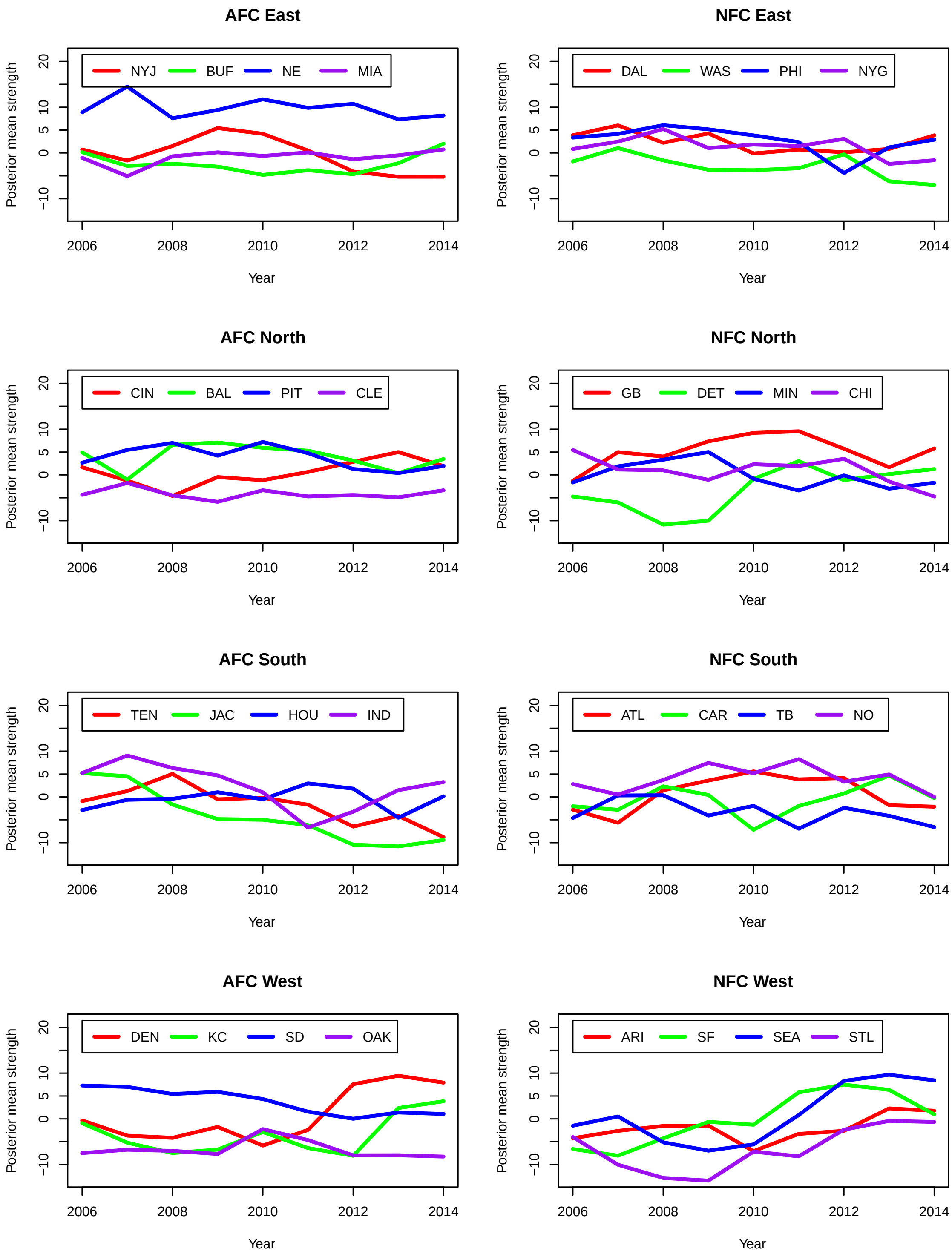

Figure 1: Posterior mean team strengths from the Bayesian dynamic linear model for all 32 NFL teams across nine seasons grouped according to division.

have less posterior uncertainty compared to the least-squares model that estimates strength for each season separately. This explains why the dynamic model interval estimates are narrower than the corresponding least-squares intervals. Furthermore, the dynamic model estimates tend to shrink more to the adjacent years' estimates relative to the corresponding least-squares estimates. For example, in 2009 and 2011, the posterior mean estimates of the Broncos were -1.74 and -2.41. The least-squares estimate in 2010 was -9.03, much lower than the posterior means in 2009 and 2011. The dynamic model in effect averages the -9.03 least squares estimate with the neighboring strengths to produce a posterior mean estimate in 2010 of -5.83, somewhat higher than the least-squares estimate. When the Broncos' results in the 2012-13 season was a large improvement over the previous season, the least-squares estimate was 10.10, but the shrinkage due to the neighboring estimates pulled the dynamic model estimate down slightly to 7.58.

7 Conclusions

This chapter described the basics of estimating NFL team strength based on game outcomes within a season, game outcomes recorded across seasons, and the inclusion of potential covariate information to sharpen inferences. These basic approaches rely on establishing a relationship between game outcomes and differences in team strength parameters. This foundation has served the statistical community well for many years, and is a solid basis for team rankings and predicting game outcomes.

Plenty of room exists for exploring and developing new ideas for measuring NFL team strength. For example, representing NFL team strength by a scalar parameter within a linear model is arguably a limiting approach. Such a simple linear model, which assumes a strict ordering in strengths, does not acknowledge the existence of circular triads in which

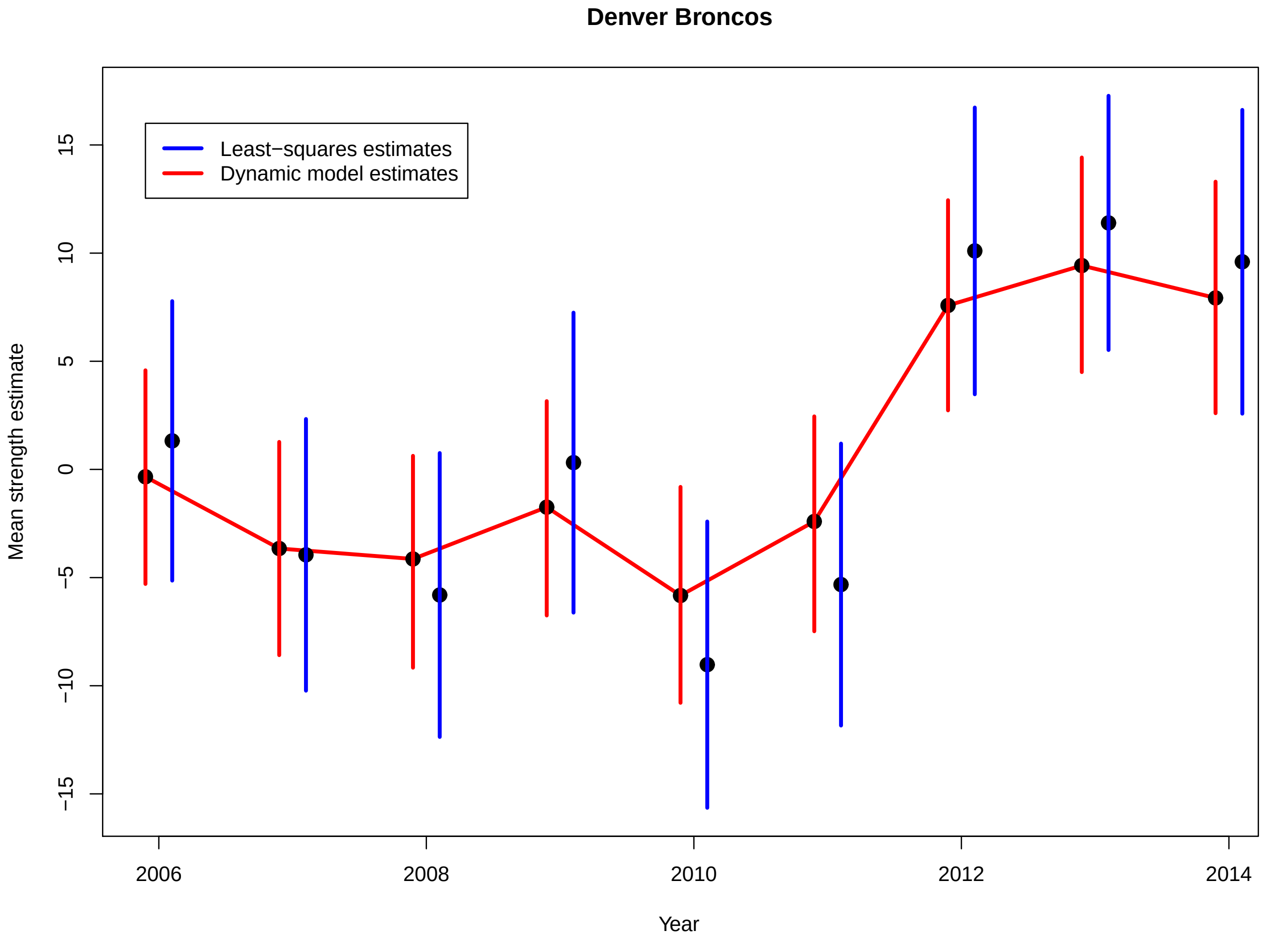

Figure 2: Point and interval estimates for the strength parameter of the Denver Broncos over time. The red segments are the 95% central posterior intervals from the dynamic linear model; the blue segments are the 95% confidence intervals based on the least-squares estimates. The points in the center of each interval are the corresponding posterior means and least-squares estimates.

team A is better than B which is better than C, but that team C is better than A. The development of such models, which surely require vectors of team strengths rather than scalar strengths, is an open area. Furthermore, now that detailed information is available for individual players, including player information from play-by-play data and the increased availability of player tracking data, possibilities exist for constructing team strength models based on player contributions that incorporate synergistic relationships among players (e.g., strength of quarterback/receiver pairs). Improvements in team strength estimation derived from these novel approaches may lead to new ways of conceptualizing ability measurement in NFL team strength, and competitive ability in sports more generally.

References

Acker, J. C. (1997): "Location variations in professional football," *Journal of Sport Behavior*, 20, 247.

Baker, R. D. and I. G. McHale (2013): "Forecasting exact scores in National Football League games," *International Journal of Forecasting*, 29, 122–130.

Baker, R. D. and I. G. McHale (2014): "A dynamic paired comparisons model: Who is the greatest tennis player?" *European Journal of Operational Research*, 236, 677–684.

Bradley, R. A. and M. E. Terry (1952): "Rank analysis of incomplete block designs: I. The method of paired comparisons," *Biometrika*, 324–345.

Colley, W. N. (2002): "Colley's bias free college football ranking method: The Colley matrix explained," *Princeton University.*

David, H. A. (1988): *The Method of Paired Comparisons*, New York: Oxford University Press.

David, J. A., R. D. Pasteur, M. S. Ahmad, and M. C. Janning (2011): "NFL prediction using committees of artificial neural networks," *Journal of Quantitative Analysis in Sports*, 7.

Davidson, R. R. (1970): "On extending the Bradley-Terry model to accommodate ties in paired comparison experiments," *Journal of the American Statistical Association*, 65, 317–328.

Ford, L. R. (1957): "Solution of a ranking problem from binary comparisons," *American Mathematical Monthly*, 28–33.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2014): *Bayesian Data Analysis*, Taylor & Francis, 3rd edition.

Glenn, W. and H. David (1960): "Ties in paired-comparison experiments using a modified Thurstone-Mosteller model," *Biometrics*, 16, 86–109.

Glickman, M. E. (1999): "Parameter estimation in large dynamic paired comparison experiments," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48, 377–394.

Glickman, M. E. (2001): "Dynamic paired comparison models with stochastic variances," *Journal of Applied Statistics*, 28, 673–689.

Glickman, M. E. and H. S. Stern (1998): "A state-space model for National Football League scores," *Journal of the American Statistical Association*, 93, 25–35.

Harville, D. (1980): "Predictions for National Football League games via linear-model methodology," *Journal of the American Statistical Association*, 75, 516–524.

Hoerl, A. E. and R. W. Kennard (1970): "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, 12, 55–67.

Jacquier, E., N. G. Polson, and P. E. Rossi (2002): "Bayesian analysis of stochastic volatility models," *Journal of Business & Economic Statistics*, 20, 69–87.

Kalman, R. E. and R. S. Bucy (1961): "New results in linear filtering and prediction theory," *Journal of Fluids Engineering*, 83, 95–108.

Kim, S., N. Shephard, and S. Chib (1998): "Stochastic volatility: likelihood inference and comparison with arch models," *The Review of Economic Studies*, 65, 361–393.

Massey, K. (1997): "Statistical models applied to the rating of sports teams," Technical report, Bluefield College.

Meinhold, R. J. and N. D. Singpurwalla (1983): "Understanding the Kalman filter," *The American Statistician*, 37, 123–127.

Mosteller, F. (1951): "Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations," *Psychometrika*, 16, 3–9.

Nelder, J. A. and R. Mead (1965): "A simplex method for function minimization," *The Computer Journal*, 7, 308–313.

Plummer, M. (2003): "JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling," in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003). March*, 20–22.

R Core Team (2015): *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, URL `https://www.R-project.org/`.

Rao, P. and L. L. Kupper (1967): "Ties in paired-comparison experiments: A generalization of the Bradley-Terry model," *Journal of the American Statistical Association*, 62, 194–204.

Sallas, W. M. and D. A. Harville (1988): "Noninformative priors and restricted maximum likelihood estimation in the Kalman filter," .

Stern, H. (1991): "On the probability of winning a football game," *The American Statistician*, 45, 179–183.

Stern, H. S. (1995): "Who's number 1 in college football?... and how might we decide," *Chance*, 8, 7–14.

Thurstone, L. L. (1927): "A law of comparative judgment." *Psychological review*, 34, 273.

Uzoma, A. O., E. Nwachukwu, et al. (2015): "A hybrid prediction system for american NFL results," *International Journal of Computer Applications Technology and Research*, 4, 42–47.

Warner, J. (2010): "Predicting margin of victory in NFL games: Machine learning vs. the las vegas line," Technical report, Pennsylvania State University.

West, M. (1996): *Bayesian Forecasting*, Wiley Online Library.

Team	Least-squares estimate
New England Patriots	10.92
Denver Broncos	9.60
Seattle Seahawks	9.54
Green Bay Packers	8.25
Kansas City Chiefs	5.66
Dallas Cowboys	5.23
Buffalo Bills	5.03
Baltimore Ravens	4.64
Indianapolis Colts	4.62
Philadelphia Eagles	3.91
Miami Dolphins	2.39
Pittsburgh Steelers	2.26
Arizona Cardinals	1.98
Detroit Lions	1.88
San Diego Chargers	1.86
Houston Texans	1.77
Cincinnati Bengals	0.77
St. Louis Rams	–0.84
San Francisco 49ers	–0.96
Minnesota Vikings	–1.69
New York Giants	–1.70
New Orleans Saints	–2.90
Carolina Panthers	–3.05
Atlanta Falcons	–3.70
Cleveland Browns	–3.86
New York Jets	–5.21
Chicago Bears	–6.76
Washington Redskins	–8.74
Oakland Raiders	–8.85
Tampa Bay Buccaneers	–9.77
Jacksonville Jaguars	–10.48
Tennessee Titans	–11.83
Residual standard deviation ($\hat{\sigma}$)	14.05
Home field ($\hat{\delta}$)	2.44

Table 1: Least-squares estimates of $\boldsymbol{\theta}$ for score difference model based on NFL game outcomes during the 2014-2015 regular season.

Team	Posterior Mean	Posterior Std Dev
Seattle Seahawks	8.42	2.65
New England Patriots	8.19	2.65
Denver Broncos	7.93	2.71
Green Bay Packers	5.79	2.68
Kansas City Chiefs	3.86	2.65
Dallas Cowboys	3.85	2.63
Baltimore Ravens	3.48	2.66
Indianapolis Colts	3.25	2.66
Philadelphia Eagles	2.91	2.66
Buffalo Bills	2.02	2.68
Pittsburgh Steelers	1.97	2.62
Cincinnati Bengals	1.94	2.65
Arizona Cardinals	1.78	2.68
Detroit Lions	1.28	2.70
San Diego Chargers	1.08	2.65
San Francisco 49ers	1.00	2.66
Miami Dolphins	0.77	2.70
Houston Texans	0.14	2.67
New Orleans Saints	–0.01	2.68
Carolina Panthers	–0.25	2.67
St. Louis Rams	–0.66	2.67
New York Giants	–1.59	2.68
Minnesota Vikings	–1.71	2.68
Atlanta Falcons	–2.14	2.61
Cleveland Browns	–3.36	2.67
Chicago Bears	–4.71	2.68
New York Jets	–5.19	2.68
Tampa Bay Buccaneers	–6.59	2.68
Washington Redskins	–6.98	2.66
Oakland Raiders	–8.23	2.69
Tennessee Titans	–8.79	2.65
Jacksonville Jaguars	–9.41	2.70
Residual standard deviation (σ)	13.12	0.20
Home field (δ)	2.41	0.27
Innovation standard deviation (τ)	4.28	0.36
Autoregressive factor (ρ)	0.67	0.07

Table 2: Posterior means and standard deviations of 2014-2015 team strengths and other model parameters based on a Normal dynamic linear model for NFL game outcomes recorded for the period 2006–2014.