Tiered Polychotomous Regression: Ranking NFL Quarterbacks

Chris WHITE and Scott BERRY

Multinomial response modeling is still in debate, with numerous parametric and nonparametric methods being popular. We investigate a tiered logistic regression technique that can handle complex functional forms while still maintaining a parametric framework. Using our model, we find the value of different plays in the NFL. We apply these results to ranking NFL quarterbacks and compare our rankings to the rankings found using the NFL quarterback rating. The main application of this article is to a sports topic, but our model could be used in any polychotomous regression setting.

KEY WORDS: BIC; Logistic; Multinomial.

1. INTRODUCTION

In the National Football League, the official measure of a quarterback's performance is his "quarterback rating." This rating is used to rank quarterbacks and is so widely accepted that it is used in some contracts as an incentive. Quarterback Donovan McNabb has an incentive clause in his contract with the Philadelphia Eagles that pays him a bonus if his quarterback rating is 100 or above. Akili Smith, Tony Banks, and many others have also signed contracts that include bonuses if their quarterback ratings are above a certain threshold. It is by far the most widespread measure used to rank and differentiate quarterbacks.

The NFL quarterback rating is a linear combination of four categories: completion percentage, (average) yards per pass, (average) touchdowns per pass, and (average) interceptions per pass. It is the sum of the following four parts, multiplied by $\frac{100}{6}$:

1. $\frac{\text{Completion percentage}-0.3}{0.2}$
2. $\frac{\text{Yards per pass attempt}-3}{4}$
3. $\frac{\text{Touchdowns per pass attempt}}{0.05}$
4. $\frac{0.095-\text{Interceptions per pass attempt}}{0.04}$

Each of the above parts is truncated to be between 0 and 2.375. Thus, any negative values are set to 0 and any values greater than 2.375 are set to 2.375.

As an example, consider Randall Cunningham's 1998 season. He had 259 completions out of 425 attempts for 3,704 yards. He had 34 touchdown passes and 10 interceptions. For his season:

1. $\frac{\frac{259}{425}-0.3}{0.2} = 1.5471$
2. $\frac{\frac{3704}{425}-3}{4} = 1.4288$
3. $\frac{\frac{34}{425}}{0.05} = 1.6000$
4. $\frac{0.095-\frac{10}{425}}{0.04} = 1.7868.$

So Randall Cunningham's 1998–1999 NFL rating is $(1.5471 + 1.4288 + 1.6000 + 1.7868) * \frac{100}{6} = 106.045$. In this rating, all incompletions are equivalent, and all interceptions count equally against a quarterback. Similarly, all 20-yard touchdown passes are considered equal, regardless of the circumstance, and the above categories ignore sacks and fumbles.

To illustrate some of the drawbacks in more detail, consider the following examples from the 1998 NFL season. In week 3, Jeff George of the Oakland Raiders faced an intense pass rush on first down and 10 from the Denver Bronco 10-yard line and was sacked for an 11-yard loss. In week 10, Trent Dilfer of the Tampa Bay Buccaneers faced a first down and 10 from the Tennessee Oiler 10-yard line. He sensed a strong pass rush and threw the ball away for an incompletion, which is a much better alternative than being sacked for an 11-yard loss. The NFL quarterback rating penalizes quarterbacks for incompletions such as Dilfer's but does not penalize quarterbacks for sacks. Dilfer's team was better off because he threw the ball away for an incompletion, but the NFL rating penalizes him for this while George's rating is unaffected.

It would also be advantageous to fumble rather than throw an incompletion. In the 1998 week 15 San Diego Chargers versus Seattle Seahawks game, Ryan Leaf faced a second down and 14 and fumbled as he was being sacked. The fumble was returned by the defense for a touchdown, but the NFL rating does not penalize him. Leaf's team would have been much better off if he had simply thrown the ball away for an incompletion, but again the NFL rating would penalize him for this.

In week 10, the New Orleans Saints played the Minnesota Vikings and Brad Johnson of the Vikings faced a third down and 3 from the New Orleans five-yard line. He threw an interception that was returned 95 yards for a touchdown. In contrast, John Elway threw an interception on third down and 21 from his own 35-yard line in the week 6 Denver Broncos versus Seattle Seahawks game. The pass traveled 47 yards with no return on the interception. Both of these plays would count equally against each quarterback in the official quarterback rating, despite the fact that Brad Johnson severely hurt his team, while Elway's interception was better than most punts.

Chris White is Statistician at M. B. Flippen & Associates, 1199 Haywood, College Station, TX 77845 (E-mail: chris@leadershipsolutions.com). Scott Berry is Statistician at Berry Consultants, 1039 Wellington Court, Sycamore, IL 60178 (E-mail: scott@berryconsultants.com).

Finally, in the week 4 Minnesota Vikings versus Chicago Bears game, Erik Kramer—facing first down and 10—threw a 23-yard touchdown pass. In the week 2 Arizona Cardinals versus Seattle Seahawks game, Jake Plummer threw a 23-yard touchdown pass facing fourth down and 16. Fourth down and 16 is a much more difficult situation, given that you must get 16 yards in a single play, along with the fact that the defense is expecting a pass. Thus, should both of these plays count equally in measuring a quarterback or should Plummer be rewarded more?

We propose a method for rating quarterbacks (U. S. Patent Pending) that better measures the contribution to the team. We assign a value to each play and use these values to summarize a quarterback's effectiveness. A quarterback's ultimate goal is to win the football game, so one option would be to determine how much each play contributed toward or against the probability of winning. Because of multiple types of scores and a continuous clock it is very difficult to measure the probability of winning. The method we propose rewards or penalizes the quarterback according to the "expected points" gained or lost on each play. A football team is not necessarily trying to maximize the number of points scored on the given play, nor is a team attempting to maximize the number of points scored on the drive. If so, no team would ever punt. Teams are trying to maximize the number of points scored eventually, which is why teams punt. They hope to pin the other team back, not allow them to advance, and then get the ball back in good field position and be the next team to score. Since this is what teams are maximizing, we model eventual points scored as our response variable. This article uses the term expected points to refer to expected eventual points. Stern (1998) modeled eventual points on first down and 10 from various field positions using a least squares approach.

Assume that teams score an average of 3.9 eventual points on third down and 5 from the other team's 10-yard line. Then if the quarterback throws a touchdown pass, he basically "gained" 3.1 points for his team (7 points − 3.9 points). If he throws an interception that is run back for a touchdown, then he "lost" 10.9 points for his team (lost 3.9 points in addition to the 7-point touchdown). If he throws an incomplete pass, then his team faces a fourth down and 5 from the 10-yard line, where teams score an average of, say, 2.5 eventual points. Thus, an incompletion would cost his team 1.4 points (3.9 points − 2.5 points). Using this approach, quarterbacks are rewarded or penalized according to the value of the play, instead of having all 10-yard touchdown passes count equally, all interceptions count equally, and so on.

Numerous game situations affect expected points; the down, the yards to first down, the yard line, the time remaining, the score, the caliber of the defense, and so on. We focus on three explanatory variables: down (denoted "Down"), yards to go for a first down (denoted "ToGo"), and yards to the goal line (denoted "ToGoal"). To measure the differential points on each play, we create a model for the expected points as a combination of Down, ToGo, and ToGoal. Taking the difference between the expected points before and after a play measures the "value" of the play. We use the average value of all plays for each quarterback to rank them.

We fit a polychotomous regression model to estimate the expected number of points for each game situation. We use play-by-play data from the 1998 NFL season. Section 2 describes the data. Section 3 describes the tree-based polychotomous regression model. Section 4 describes how the model measures the impact of a play by reviewing each of the scenarios presented in the introduction. Section 5 presents quarterback rankings for 1998 and compares them to the NFL rating system. Concluding remarks are presented in Section 6.

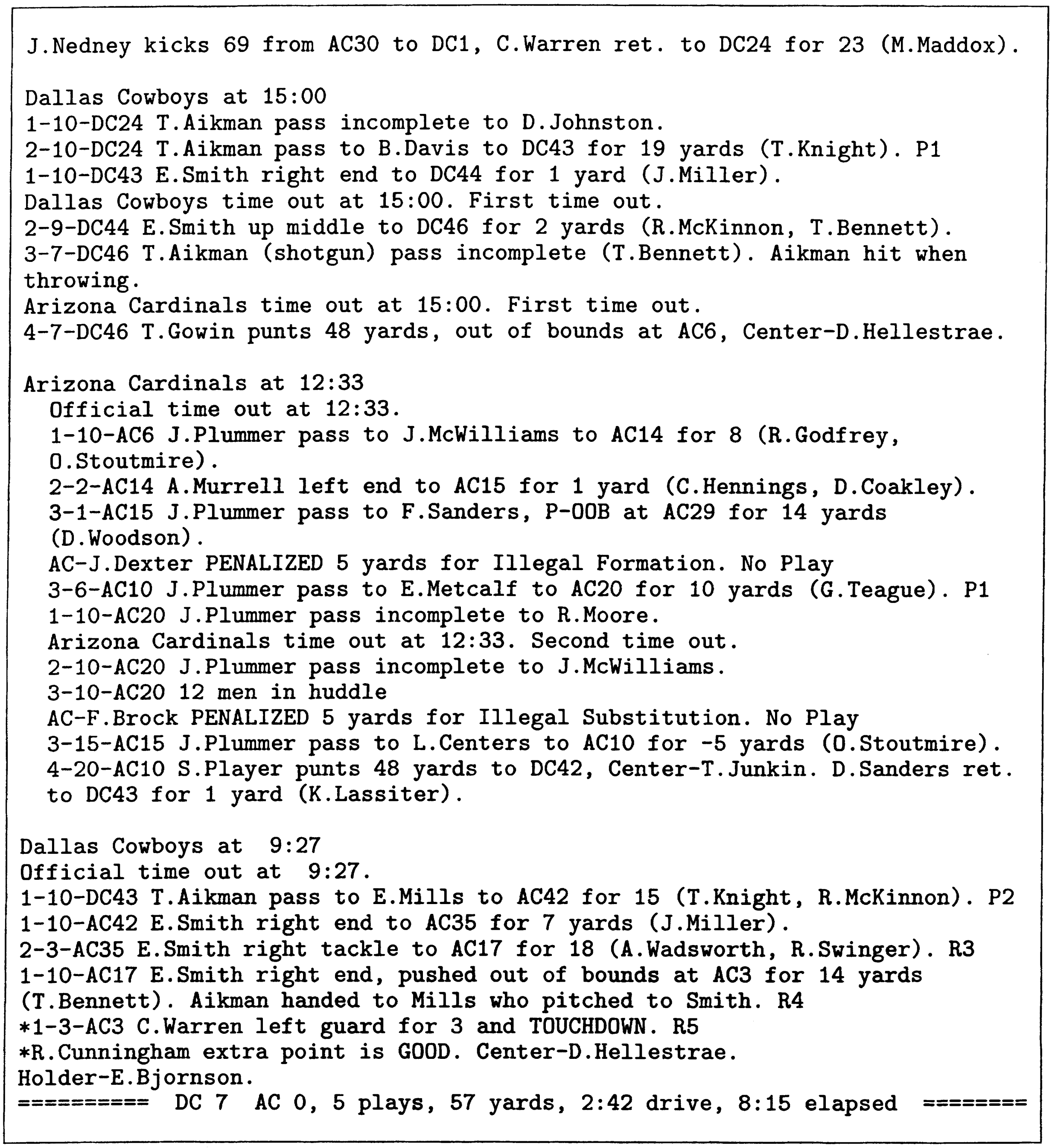
2. THE DATA

The original play-by-play data was in the format presented in Figure 1, using the first three possessions of the 1998–1999 Dallas Cowboys versus Arizona Cardinals game in week 11 of the season as an example. We wrote a program to create a structured dataset containing our variables of interest from this prose-based file. Our full dataset consisted of each play from the 1998–1999 season, which was more than 35,000 plays. The play-by-play data follow a common structure for each play which allows us to restructure the information into a dataset that includes only the relevant information. The Figure 1 data was restructured into the Figure 2 data. In the Figure 2 output, eventual points are in the last column. To find the eventual points for a given play, determine which team scored next. If the same team scored next, then the eventual points for the current play would be equal to how many points were scored. If the other team scored, eventual points would be equal to the negative of the amount that the other team scored. For example, in the Figure 1 output Dallas had the ball first and did not score on the drive. Arizona did not score on the subsequent drive. On the next drive, Dallas scored 7 points on a touchdown. Thus, each play on the scoring drive receives 7 eventual points. Each play on the Arizona drive receives −7 eventual points and each play on the first Dallas drive receives 7 eventual points.

We pooled eventual points that were −6 or −8 to be −7 and pooled points that were 6 or 8 to be 7. We want all touchdowns to count equally, without regard to the extra point. If neither team scored before the end of the first half or the end of the game, then eventual points for plays after the last score were recorded as zero.

3. EXPECTED POINTS MODEL

The three explanatory variables are Down, ToGo, and ToGoal. The possibilities for down are $\{1, 2, 3, 4\}$. We do not want to force the difference between first down and second down to be the same as the difference between third down and fourth down. This would be unrealistic, since moving from third down to fourth down is usually more costly in terms of expected points than moving from first down to second down. Thus, we use dummy variables for each down. Since down has four possibilities, we use three dummy variables. Thus, D2, D3, and D4 denote indicator variables for second, third, and fourth down, respectively. To model expected points from D2, D3, D4, ToGo, and ToGoal, we develop a model for the probability of each of the seven possible outcomes. From each situation, a team could eventually score a touchdown (+7), kick a field goal (+3), record a safety (+2), allow a safety (−2), allow a field goal (−3), allow a touchdown (−7), or no points could be scored. To estimate the expected points we average over the probabilities of each outcome to find the expected eventual points. Estimating the



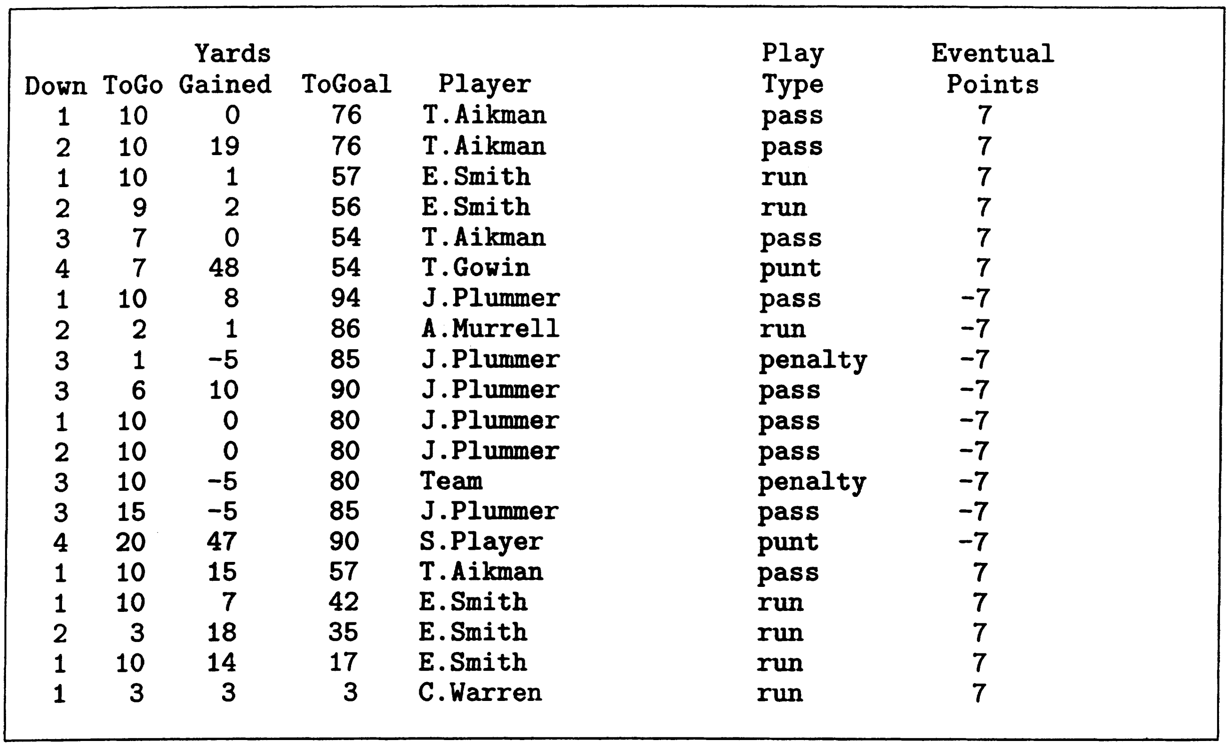
```
J.Nedney kicks 69 from AC30 to DC1, C.Warren ret. to DC24 for 23 (M.Maddox).

Dallas Cowboys at 15:00
1-10-DC24 T.Aikman pass incomplete to D.Johnston.
2-10-DC24 T.Aikman pass to B.Davis to DC43 for 19 yards (T.Knight). P1
1-10-DC43 E.Smith right end to DC44 for 1 yard (J.Miller).
Dallas Cowboys time out at 15:00. First time out.
2-9-DC44 E.Smith up middle to DC46 for 2 yards (R.McKinnon, T.Bennett).
3-7-DC46 T.Aikman (shotgun) pass incomplete (T.Bennett). Aikman hit when
throwing.
Arizona Cardinals time out at 15:00. First time out.
4-7-DC46 T.Gowin punts 48 yards, out of bounds at AC6, Center-D.Hellestrae.

Arizona Cardinals at 12:33
  Official time out at 12:33.
  1-10-AC6 J.Plummer pass to J.McWilliams to AC14 for 8 (R.Godfrey,
  O.Stoutmire).
  2-2-AC14 A.Murrell left end to AC15 for 1 yard (C.Hennings, D.Coakley).
  3-1-AC15 J.Plummer pass to F.Sanders, P-OOB at AC29 for 14 yards
  (D.Woodson).
  AC-J.Dexter PENALIZED 5 yards for Illegal Formation. No Play
  3-6-AC10 J.Plummer pass to E.Metcalf to AC20 for 10 yards (G.Teague). P1
  1-10-AC20 J.Plummer pass incomplete to R.Moore.
  Arizona Cardinals time out at 12:33. Second time out.
  2-10-AC20 J.Plummer pass incomplete to J.McWilliams.
  3-10-AC20 12 men in huddle
  AC-F.Brock PENALIZED 5 yards for Illegal Substitution. No Play
  3-15-AC15 J.Plummer pass to L.Centers to AC10 for -5 yards (O.Stoutmire).
  4-20-AC10 S.Player punts 48 yards to DC42, Center-T.Junkin. D.Sanders ret.
  to DC43 for 1 yard (K.Lassiter).

Dallas Cowboys at  9:27
Official time out at  9:27.
1-10-DC43 T.Aikman pass to E.Mills to AC42 for 15 (T.Knight, R.McKinnon). P2
1-10-AC42 E.Smith right end to AC35 for 7 yards (J.Miller).
2-3-AC35 E.Smith right tackle to AC17 for 18 (A.Wadsworth, R.Swinger). R3
1-10-AC17 E.Smith right end, pushed out of bounds at AC3 for 14 yards
(T.Bennett). Aikman handed to Mills who pitched to Smith. R4
*1-3-AC3 C.Warren left guard for 3 and TOUCHDOWN. R5
*R.Cunningham extra point is GOOD. Center-D.Hellestrae.
Holder-E.Bjornson.
==========  DC 7  AC 0, 5 plays, 57 yards, 2:42 drive, 8:15 elapsed  ========
```

Figure 1. Original play-by-play data.



Down	ToGo	Yards Gained	ToGoal	Player	Play Type	Eventual Points
1	10	0	76	T.Aikman	pass	7
2	10	19	76	T.Aikman	pass	7
1	10	1	57	E.Smith	run	7
2	9	2	56	E.Smith	run	7
3	7	0	54	T.Aikman	pass	7
4	7	48	54	T.Gowin	punt	7
1	10	8	94	J.Plummer	pass	-7
2	2	1	86	A.Murrell	run	-7
3	1	-5	85	J.Plummer	penalty	-7
3	6	10	90	J.Plummer	pass	-7
1	10	0	80	J.Plummer	pass	-7
2	10	0	80	J.Plummer	pass	-7
3	10	-5	80	Team	penalty	-7
3	15	-5	85	J.Plummer	pass	-7
4	20	47	90	S.Player	punt	-7
1	10	15	57	T.Aikman	pass	7
1	10	7	42	E.Smith	run	7
2	3	18	35	E.Smith	run	7
1	10	14	17	E.Smith	run	7
1	3	3	3	C.Warren	run	7

Figure 2. Structured data.

probability of $k > 2$ outcomes from explanatory variables is referred to as polychotomous regression.

The response variable expected points has the seven outcomes $\{-7, -3, -2, 0, 2, 3, 7\}$. McCullagh and Nelder (1989) discussed various models for polychotomous regression, including an example of tiered logistic regression, which they referred to as nested or hierarchical response modeling. Hosmer and Lemeshow (1989) described multiple logistic regression (multiple independent variables) and also briefly described the polychotomous regression setting. Kooperberg, Bose, and Stone (1997) described model selection procedures and introduced a polychotomous regression algorithm involving linear splines. They discussed maximum likelihood estimation, stepwise selection in model fitting, the Akaike information criterion, cross-validation, and the use of an independent test set in the model selection stage. The authors applied their procedure to a phoneme recognition dataset. Albert and Chib (1993) presented a Bayesian approach to polychotomous response data using the idea of data augmentation.

This article uses a tiered logistic regression model. The response variable in logistic regression has two possible outcomes, while in our problem the response variable, eventual points, has seven outcomes. To fit a tiered logistic regression model, we split the outcomes into $k-1$ tiers, where k is the number of outcomes. Here we have seven outcomes, so we split the outcomes into six pairs, structured as in Figure 3. There are numerous choices for how to split the outcomes so we chose one that was nicely organized. For a discussion of the effect of different tier structures see White (2000). At each dichotomous split we fit a logistic regression model. In Figure 3, each of the terminal nodes represents an outcome. To obtain the probability of a particular outcome, multiply the appropriate probabilities for the tiers leading to the terminal node of interest. For example, the outcome "no points" involves only the first tier. To find the probability of "no points," find the probability of 0 (i.e., no success) in the first-tier logistic regression. To find the probability of "negative safety," use tiers 1, 2, and 4. In the first tier and the fourth tier, find the probability of a success, and in the second tier find the probability of no success. Multiplying the three probabilities together gives us the probability of the next score being a safety for the other team.

We present the mathematical details for this model. Let Y_i be the outcome for play i, for $i = 1, \ldots, n$, and let the vector $\mathbf{X}_i$ be the independent variables for play i; thus, $\mathbf{X}_i = (\mathrm{D2}_i, \mathrm{D3}_i, \mathrm{D4}_i, \mathrm{ToGo}_i, \mathrm{ToGoal}_i)$. Now consider a tier indicator for the outcome of play i, denoted $\Phi_i = (\phi_{i1}, \phi_{i2}, \phi_{i3}, \phi_{i4}, \phi_{i5}, \phi_{i6})$, where

$$\phi_{it} = \begin{cases} 1 & \text{if tier } t \text{ is used in the outcome for play } i \\ 0 & \text{otherwise.} \end{cases}$$

For example, considering the "no points" outcome in Figure 3, the only tier involved is tier 1. Thus for all "no points" plays, $\Phi_i = (1, 0, 0, 0, 0, 0)$. For the seven outcomes,

$$\Phi_i = \begin{cases} (1,1,0,1,0,1) & \text{if} \quad Y_i = -7 \\ (1,1,0,1,0,1) & \text{if} \quad Y_i = -3 \\ (1,1,0,1,0,0) & \text{if} \quad Y_i = -2 \\ (1,0,0,0,0,0) & \text{if} \quad Y_i = 0 \\ (1,1,1,0,0,0) & \text{if} \quad Y_i = +2 \\ (1,1,1,0,1,0) & \text{if} \quad Y_i = +3 \\ (1,1,1,0,1,0) & \text{if} \quad Y_i = +7. \end{cases}$$

For each play i, let the vector $\Lambda_i = (\lambda_{i1}, \lambda_{i2}, \lambda_{i3}, \lambda_{i4}, \lambda_{i5}, \lambda_{i6})$ be the *left indicator*, where

$$\lambda_{ij} = \begin{cases} 1 & \text{if the outcome for play } i \text{ requires a left at tier } t \\ 0 & \text{otherwise.} \end{cases}$$

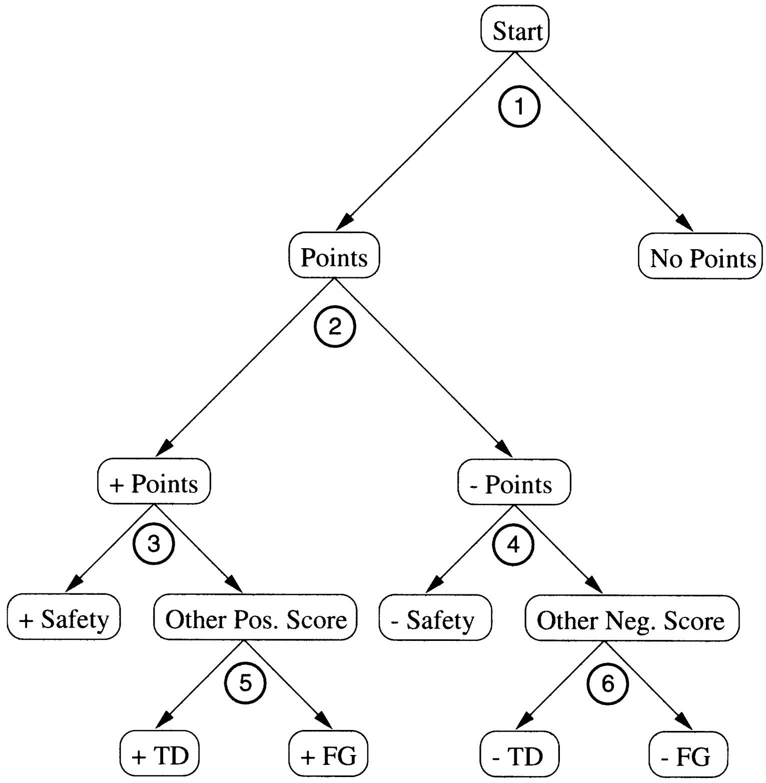


Figure 3. Tiers for the outcomes, with each tier numbered.

For the outcome "positive touchdown" in Figure 3 we need to go left at tiers 1, 2, and 5. Thus, for any "positive touchdown" play, $\Lambda_i = (1,1,0,0,1,0)$. Similarly, the seven outcomes are

$$\Lambda_i = \begin{cases} (1,0,0,0,0,1) & \text{if} \quad Y_i = \text{-}7 \\ (1,0,0,0,0,0) & \text{if} \quad Y_i = \text{-}3 \\ (1,0,0,1,0,0) & \text{if} \quad Y_i = \text{-}2 \\ (0,0,0,0,0,0) & \text{if} \quad Y_i = 0 \\ (1,1,1,0,0,0) & \text{if} \quad Y_i = +2 \\ (1,1,0,0,0,0) & \text{if} \quad Y_i = +3 \\ (1,1,0,0,1,0) & \text{if} \quad Y_i = +7. \end{cases}$$

This notation creates a convenient characterization of each outcome. Finally, let n_t denote the number of plays at each tier t, as delineated in Figure 3. For play i with independent variables $\mathbf{X}_i$, the probability of outcome Y_i is

$$P(Y_i|\mathbf{X}_i, \mathrm{L}_1, \ldots, \mathrm{L}_6) = \prod_{t=1}^{6} \left[\frac{\exp(\mathrm{L}_t)^{\lambda_{it}}}{1+\exp(\mathrm{L}_t)} \cdot \phi_{it} + (1-\phi_{it}) \right],$$

where L_t is the logit function at tier t. The likelihood function is

$$\begin{aligned} & f(\mathbf{Y}|\mathbf{X}_1, \ldots, \mathbf{X}_n, \mathrm{L}_1, \ldots, \mathrm{L}_6) \\ & \quad = \prod_{i=1}^{n} P(Y_i|\mathbf{X}_i) \\ & \quad = \prod_{i=1}^{n} \prod_{t=1}^{6} \left[\frac{\exp(\mathrm{L}_t)^{\lambda_{it}}}{1+\exp(\mathrm{L}_t)} \cdot \phi_{it} + (1-\phi_{it}) \right] \\ & \quad = \prod_{t=1}^{6} \prod_{\{i:\phi_{it}=1\}} \left[\frac{\exp(\mathrm{L}_t)^{\lambda_{it}}}{1+\exp(\mathrm{L}_t)} \right]. \end{aligned}$$

Table 1. Explanatory Variables Considered

Fit #	*Explanatory variables*
1	D2, D3, D4, ToGo, ToGoal
2	D2, D3, D4, sm(ToGo), ToGoal
3	D2, D3, D4, ToGo, sm(ToGoal)
4	D2, D3, D4, ToGo, ToGoal, ToGo*ToGoal
5	D2, D3, D4, sm(ToGo), sm(ToGoal)
6	D2, D3, D4, sm(ToGo), sm(ToGoal), ToGo*ToGoal
7	D2, D3, D4, sm(ToGo), sm(ToGoal), sm(ToGo*ToGoal)
8	sm(ToGo), sm(ToGoal), separately for each down
9	sm(ToGo), sm(ToGoal), sm(ToGo*ToGoal), separately for each down

The problem is now reduced to fitting $k - 1 = 6$ logistic regressions. One option would be to fit the usual linear logistic regression model at each tier. Another choice would be to fit a more complicated model at each tier, such as one that uses smoothed versions of the variables. We decided to determine the best model at each tier and then use the combination of the six best models as our overall model. We used S-Plus to fit models with various explanatory variables at each tier, shown in Table 1. In Table 1, sm($\cdot$) represents a nonlinear smoothed version of the variable in parentheses (estimated using cubic B-splines) and ToGo*ToGoal is the interaction term ToGo multiplied by ToGoal.

At each tier, we calculate the Bayesian information criterion (BIC) for each model and chose the model with the minimum

Table 2. Minimum BIC Models at Each Tier

Tier	Model with Minimum BIC
1: Points vs. No Points	D2, D3, D4, ToGo, sm(ToGoal)
2: Pos. Points vs. Neg. Points	D2, D3, D4, ToGo, sm(ToGoal)
3: Pos. Safety vs. Other Pos. Score	D2, D3, D4, ToGo, ToGoal
4: Neg. Safety vs. Other Neg. Score	D2, D3, D4, ToGo, sm(ToGoal)
5: Pos. TD vs. Pos. FG	sm(ToGo), sm(ToGoal), separately for each down
6: Neg. TD vs. Neg. FG	D2, D3, D4, ToGo, ToGoal

BIC as the best model. The BIC, due to Schwarz (1978), was discussed by Kass and Raftery (1995). It is beneficial in selecting between models because it takes into account the likelihood but also realizes that models with more terms will have higher likelihoods. Because of this, a "penalty" term is included in the BIC that penalizes models with more degrees of freedom. In our case, the BIC at a particular tier t is

$$\mathrm{BIC} = -2 \cdot \left[\log \left(\prod_{i:\phi_{it}=1} \frac{\exp(\mathrm{L}_t)^{\lambda_{it}}}{1+\exp(\mathrm{L}_t)} \right) - \frac{1}{2}\mathrm{d}_t \log(n_t) \right],$$

where

$$\mathrm{d}_t = \text{model degrees of freedom for } \mathrm{L}_t.$$

By using the BIC form as stated, one problem arises. Finding the minimum BIC at each tier does not necessarily assure us of ending up with the overall combination of six models that has overall minimum BIC. To remedy this issue, we minimize an alternative form of the BIC at each tier that uses $\log(n)$ instead of $\log(n_t)$. This alternative form is written as

$$\mathrm{BIC}^* = -2 \cdot \left[\log \left(\prod_{i:\phi_{it}=1} \frac{\exp(\mathrm{L}_t)^{\lambda_{it}}}{1+\exp(\mathrm{L}_t)} \right) - \frac{1}{2}\mathrm{d}_t \log(n) \right],$$

where the only difference is the last term. BIC* is a form of the BIC that penalizes even more for additional terms in the model. Using this form of the BIC at each tier, the overall BIC can simply be written $\mathrm{BIC} = \mathrm{BIC}_1^* + \mathrm{BIC}_2^* + \cdots + \mathrm{BIC}_6^*$. In this case, finding the minimum BIC* at each tier *does* ensure finding the model with overall minimum BIC. The proof of this result is presented in Appendix A.

The explanatory variables that had the minimum BIC at each tier are presented in Table 2. At the first tier (points vs. no points) we have the prediction equation

$$\begin{aligned}
&\mathrm{logit}(\hat{P}_{\mathrm{points}}) \\
&= 2.315 - 0.013\ \mathrm{D2} - 0.046\ \mathrm{D3} \\
&\quad -0.189\ \mathrm{D4} - 0.017\ \mathrm{ToGo} + \mathrm{sm(ToGoal)} \\
&= \hat{\mathrm{L}}_1
\end{aligned}$$

which we can convert from logit scale to probability scale using

$$\hat{P}_{\mathrm{points}} = \frac{\exp(\hat{\mathrm{L}}_1)}{1+\exp(\hat{\mathrm{L}}_1)}.$$

We estimated the logit equation using S-Plus, with the command

$$\begin{aligned}
\mathrm{logistic1} <& -\mathrm{gam}(Y1 \sim D2 + D3 + D4 \\
&+\mathrm{ToGo} + s(\mathrm{ToGoal}), \mathrm{family} = \mathrm{binomial}).
\end{aligned}$$

At the second tier (+ points vs. − points) we have

$$\begin{aligned}
\mathrm{logit}(\hat{P}_{+\mathrm{points}}) &= 3.074 - 0.177\ \mathrm{D2} \\
&\quad -0.439\ \mathrm{D3} - 0.831\ \mathrm{D4} \\
&\quad -0.009\ \mathrm{ToGo} + \mathrm{sm(ToGoal)} \\
&= \hat{\mathrm{L}}_2.
\end{aligned}$$

At the third tier (+ safety vs. other positive score) we have

$$\begin{aligned}
\mathrm{logit}(\hat{P}_{+\mathrm{safety}}) &= -5.558 + 0.118\ \mathrm{D2} \\
&\quad +0.465\ \mathrm{D3} + 0.667\ \mathrm{D4} - 0.016\ \mathrm{ToGo} \\
&\quad +0.011\ \mathrm{ToGoal} \\
&= \hat{\mathrm{L}}_3.
\end{aligned}$$

At the fourth tier (− safety vs. other negative score) we have

$$\begin{aligned}
\mathrm{logit}(\hat{P}_{-\mathrm{safety}}) &= -6.908 + 0.145\ \mathrm{D2} + 0.215\ \mathrm{D3} \\
&\quad -0.032\ \mathrm{D4} + 0.008\ \mathrm{ToGo} + \mathrm{sm(ToGoal)} \\
&= \hat{\mathrm{L}}_4.
\end{aligned}$$

At the fifth tier (+ touchdown vs. + field goal) we have

$$\begin{aligned}
\mathrm{logit}(\hat{P}_{+\mathrm{TD}}) &= (\mathrm{sm}_1(\mathrm{ToGo}) + \mathrm{sm}_2(\mathrm{ToGoal})) \cdot \mathrm{D1} \\
&\quad +(\mathrm{sm}_3(\mathrm{ToGo}) + \mathrm{sm}_4(\mathrm{ToGoal})) \cdot \mathrm{D2} \\
&\quad +(\mathrm{sm}_5(\mathrm{ToGo}) + \mathrm{sm}_6(\mathrm{ToGoal})) \cdot \mathrm{D3} \\
&\quad +(\mathrm{sm}_7(\mathrm{ToGo}) + \mathrm{sm}_8(\mathrm{ToGoal})) \cdot \mathrm{D4} \\
&= \hat{\mathrm{L}}_5,
\end{aligned}$$

where D1 is defined in the same manner as D2, D3, and D4, that is,

$$\mathrm{D1} = \begin{cases} 1 & \text{if down } = 1 \\ 0 & \text{if down } \neq 1. \end{cases}$$

The subscripts on the smoothed independent variables (e.g., $\mathrm{sm}_1(\mathrm{ToGo})$) were used to make it clear to the reader that separate smoothing functions were fit for each down. At the final tier (− touchdown vs. − field goal) we have

$$\begin{aligned}
\mathrm{logit}(\hat{P}_{-\mathrm{TD}}) &= 0.476 - 0.026 * \mathrm{D2} \\
&\quad -0.031 * \mathrm{D3} - 0.040 * \mathrm{D4} \\
&\quad +0.00063 * \mathrm{ToGo} - 0.00016 * \mathrm{ToGoal} \\
&= \hat{\mathrm{L}}_6.
\end{aligned}$$

As mentioned earlier, we determine the predicted probability of a particular outcome by multiplying the applicable probabilities from each tier. For example, to find the predicted probability of a touchdown while facing first down and 10 with 10 yards to the goal, we use the probabilities given in Figure 4.

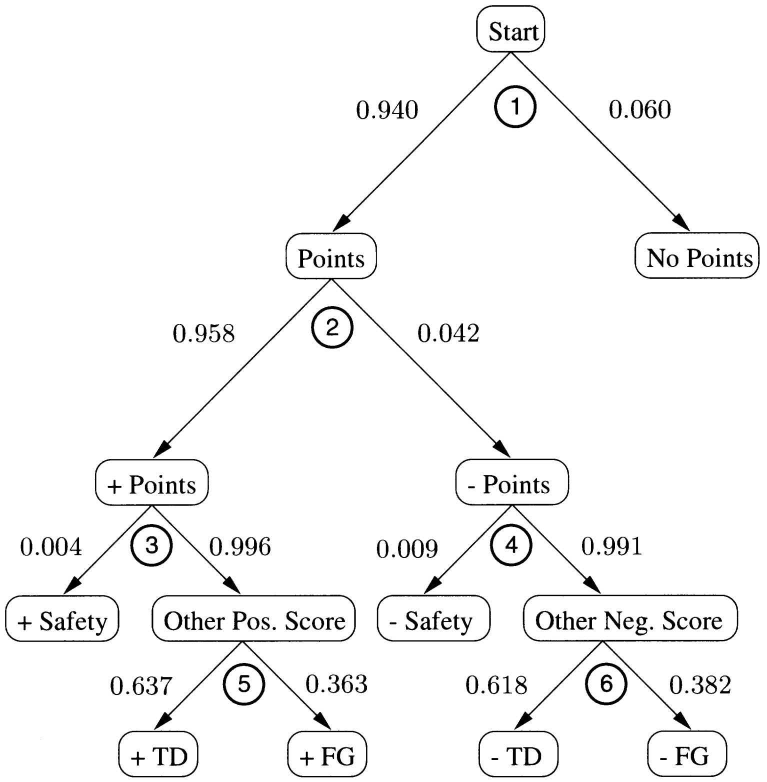


Figure 4. Logistic regression probabilities for first down and 10, 10 yards to goal.

The predicted probability of positive touchdown (vs. positive field goal) in the left bottom tier is $\hat{P}_{+\mathrm{TD}} = 0.637$. Continuing up the diagram, we find that the predicted probability of an "other positive score" (vs. positive safety) is $\hat{P}_{\text{other pos. score}} = 0.996$. The predicted probability of positive points (vs. negative points) is $\hat{P}_{+\text{points}} = 0.958$, and the predicted probability of points scored (vs. no points scored) is $\hat{P}_{\text{points}} = 0.940$. Multiplying these together, we get the predicted probability of a positive touchdown to be (0.637)(0.996)(0.958)(0.940) = 0.571. The predicted probabilities of the other six outcomes can be found in a similar manner, where the predicted probability of a positive field goal is 0.326, a positive safety is 0.00332, a negative safety is 0.000374, a negative field goal is 0.0150, a negative touchdown is 0.0243, and no points is 0.0598. Note that the probabilities do sum to 1.

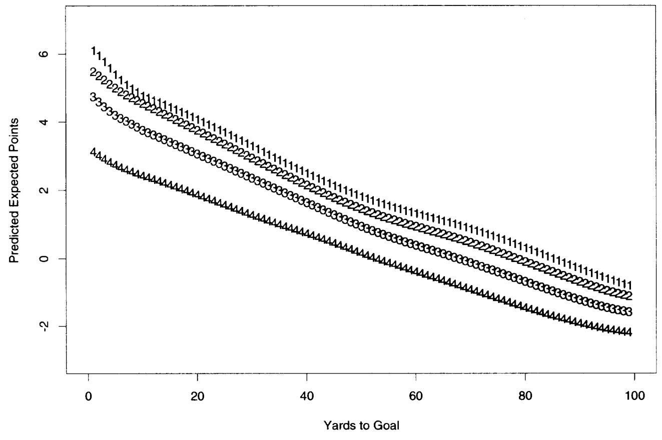


Figure 5. Downs one through four with 10 yards to go.

To find the predicted expected points, we simply weight each outcome by its probability, which results in 4.77 predicted expected points. Figure 5 shows the expected points for downs 1 through 4 with 10 yards to go. Figure 6 presents three-dimensional graphs of the seven outcome probabilities for first down. Each of the graphs in Figure 6 is consistent with conventional wisdom. The probability of 0 eventual points decreases as a team gets closer to the opponent's goal line. Similarly, the probability of the opponent scoring a touchdown or field goal also decreases as a team gets closer to the opponent's goal line. The probability of the opponent scoring a safety is relatively high when you are close to your own goal line, and the probability of a touchdown gets higher as a team gets closer to the opponent's goal line and as yards to first down decreases. It is worth mentioning again that we are displaying the probability of *eventual* outcomes—not outcomes on the next play. The probability of a field goal on the next play when 99 yards from the goal line is zero. In our graph, this probability is not zero. This is because we found the probability of three eventual points, meaning the probability that the *next* score in the game will be a field goal for the current offensive team.

We present outcome probability graphs for two fourth down outcomes in Figure 7. The probability of an eventual field goal increases steadily as a team crosses midfield until curving down slightly as a team gets within five yards of a touchdown. The probability of a touchdown increases as a team approaches midfield and then starts to decrease (because now a field goal is more likely) and then increases again when a team gets very close to the opponent's goal line.

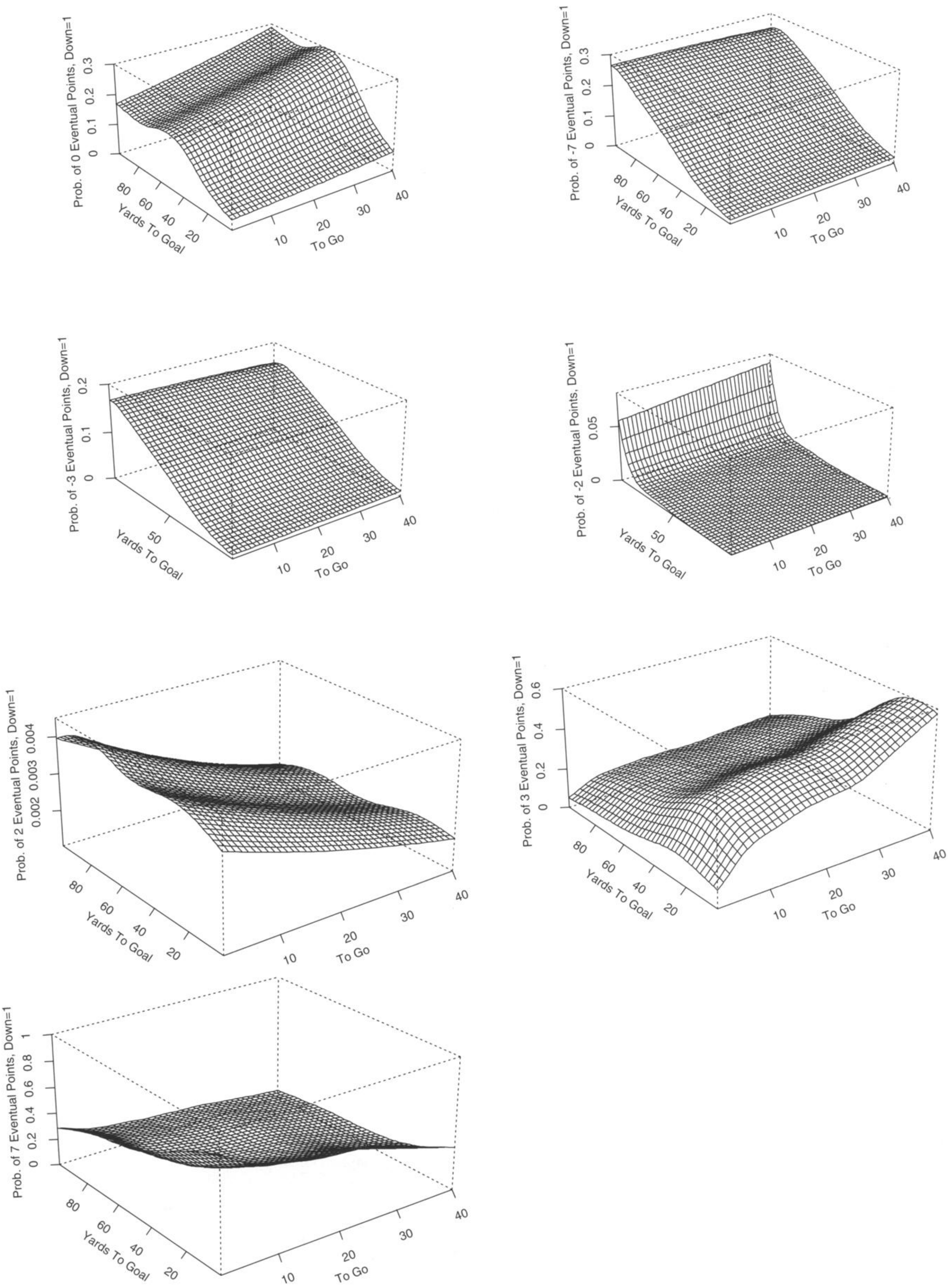


Figure 6. Predicted probabilities for each outcome, first down.

4. USING THE MODEL TO COMPARE PREVIOUSLY MENTIONED PLAYS

We now use our model to revisit the plays discussed in the introduction. In the first scenario, Jeff George faced an intense pass rush on first down and 10 at the Denver 10-yard line which gives an expected points of 4.77. After being sacked for an 11-yard loss, Oakland faced a second down and 21 from the Denver 21-yard line, which drops expected points to 3.68. Hence George (and the Oakland offense) lost 1.09 expected points on the play. In contrast, Trent Dilfer threw an incompletion from the same situation, resulting in second down and 10 from the 10. This puts his team in a position of 4.55 expected points, thus losing only 0.22 expected points.

When Ryan Leaf faced a second down and 14 with 57 yards to the goal line, the expected points were 0.95. By fumbling and then allowing Seattle to return the fumble for a touchdown, Leaf cost his team 7.95 points (0.95 + 7 points for touchdown). If he had simply thrown the ball away for an incompletion he would have only lost 0.53 expected points, but the NFL quarterback rating would penalize him for this incompletion while ignoring his fumble.

Erik Kramer faced a first down and 10 and threw a 23-yard touchdown pass, going from 3.85 expected points to 7, thus gaining 3.15 expected points. Similarly, Jake Plummer threw a 23-yard touchdown pass on fourth down and 16, going from 1.64 expected points to 7. He gained 5.36 expected points. In the last two examples, the outcomes were the same but the downs were different. This illustrates the fact that the down has a big impact on the importance and the difficulty of the play.

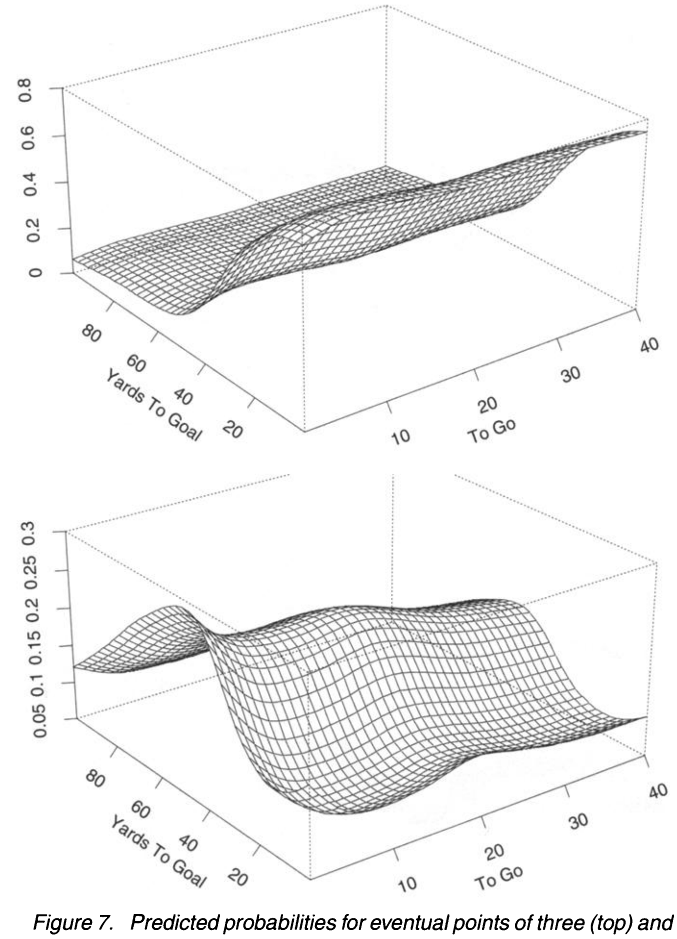


Figure 7. Predicted probabilities for eventual points of three (top) and seven (bottom), fourth down.

Finally, Brad Johnson threw an interception that was returned for a touchdown on third down and 3 from the 5-yard line. He lost 11.32 expected points, going from 4.32 to −7. John Elway faced a third down and 21 from his own 35-yard line and threw a 47-yard interception with no return, going from −0.20 expected points to −0.19, thus gaining 0.01 expected points. This case illustrates that all interceptions are not the same—some can even be beneficial!

5. USING THE MODEL TO RATE QUARTERBACKS

To rate quarterbacks, we decided to exclude plays in the last two minutes of each half. We did this for various reasons, one of the reasons being that the winning team will often play a "prevent defense" which gives up small gains in hopes of preventing a score, and we did not want this to confound our results. We also included defensive pass interference plays as completed passes since they are positive plays for the passing team. Finally, we counted intentional grounding penalties as sacks, since there is a loss of yardage and loss of down.

Although we refer to these as quarterback ratings they are certainly confounded with the ability of the quarterback's offense. A great quarterback without good blocking or without good receivers will have a low rating that might not be reflective of his true individual ability. Similarly, a quarterback who plays with a great running back is at an advantage since the defenses cannot concentrate on passes. Thus, it would be more accurate to say that we are ranking the offense's passing ability with the particular quarterback in control. When comparing quarterbacks on the same team, such as Flutie vs. Johnson or Elway vs. Brister, we can better control for the strength of the offense. The NFL rating also does not account for this limitation. We also do not account for the strength of the defense. One quarterback could have had a more difficult schedule and thus faced more talented defenses. The NFL rating similarly does not account for this.

To obtain our rankings, we use the expected point values previously discussed. For each play, we took the difference between the end-result point value and the initial point value to get the value of the play. Once we determine the point value of each play, we find the average point value over all plays for each quarterback. Each average, along with the corresponding rank of each average, is presented in Table 3. All quarterbacks from 1998–1999 who participated in a significant number of plays are included. We also include the NFL rating and corresponding rank for each quarterback in Table 3. Any differences between our ranking and the NFL ranking that are 10 or greater are denoted with a ⇑ (for a quarterback who moved up 10 or more positions) or a ⇓ (for a quarterback who moved down 10 or more positions).

We present three different rankings depending on which plays are included. We use the first ranking presented (our ranking, all plays) as our main ranking. It includes passes, sacks, interceptions, and runs. Our second ranking excludes runs, while our third ranking includes only passes and interceptions. These two supplemental rankings are presented to see if there are any unusual differences. For example, Rob Johnson is ranked 22nd

Table 3. Quarterback Rankings

Player	NFL rating		Our ranking All plays		Our ranking No runs		Our ranking No sacks, runs	
R. Cunningham	106.0	1	0.46	1	0.45	1	0.53	1
R. Johnson	102.9	2	0.11	⇓ 22	0.07	24	0.43	4
V. Testaverde	101.6	3	0.38	2	0.36	2	0.47	3
S. Young	101.1	4	0.25	9	0.20	10	0.36	7
C. Chandler	100.9	5	0.30	5	0.28	6	0.52	2
B. Brister	99.0	6	0.27	7	0.19	11	0.25	17
J. Elway	93.0	7	0.31	4	0.30	5	0.39	6
N. O'Donnell	90.2	8	0.00	⇓ 31	0.00	34	0.17	24
M. Brunell	89.9	9	0.23	10	0.24	8	0.35	10
B. Johnson	89.0	10	0.30	6	0.32	3	0.42	5
T. Aikman	88.5	11	0.26	8	0.28	7	0.31	11
S. Beuerlein	88.2	12	0.15	15	0.19	13	0.35	8
B. Favre	87.8	13	0.21	11	0.20	9	0.29	12
D. Flutie	87.4	14	0.32	⇑ 3	0.31	4	0.35	9
J. Garrett	84.5	15	0.15	16	0.19	12	0.28	13
C. Batch	83.5	16	0.12	21	0.09	22	0.26	15
B.J. Tolliver	83.1	17	0.12	20	0.15	16	0.21	21
E. Kramer	83.1	18	0.17	14	0.16	14	0.26	16
E. Zeier	82.0	19	−0.01	⇓ 32	−0.01	32	0.16	29
T. Green	81.8	20	0.05	26	0.04	25	0.20	23
D. Bledsoe	80.9	21	0.13	19	0.12	17	0.27	14
R. Gannon	80.1	22	0.14	17	0.10	20	0.21	20
S. McNair	80.1	23	0.17	⇑ 13	0.12	18	0.22	19
D. Marino	80.0	24	0.08	24	0.08	23	0.16	26
W. Moon	76.6	25	0.02	29	0.04	27	0.16	28
J. Plummer	75.0	26	0.07	25	0.04	26	0.17	25
T. Dilfer	74.0	27	0.05	27	0.02	29	0.13	30.5
J. Harbaugh	72.9	28	0.04	28	0.02	28	0.11	32
J. Kitna	72.3	29	0.17	⇑ 12	0.15	15	0.24	18
P. Manning	71.2	30	0.09	23	0.10	21	0.16	27
K. Graham	70.8	31	0.14	⇑ 18	0.11	19	0.21	21
K. Detmer	67.7	32	0.00	30	0.01	30	0.04	36
T. Banks	68.6	33	−0.04	35	−0.06	35	0.06	34
D. Kanell	67.3	34	−0.13	41	−0.14	41	−0.02	41
D. Wuerffel	66.3	35	−0.10	39	−0.13	40	0.09	33
G. Foley	64.9	36	−0.12	40	−0.12	39	0.00	39
R. Peete	64.7	37	−0.02	33	−0.03	33	0.13	30.5
K. Stewart	62.9	38	−0.05	36	−0.08	36	0.01	37
K. Collins	62.0	39	−0.09	38	−0.10	37	0.05	35
D. Hollas	60.6	40	−0.20	42	−0.17	42	−0.01	40
E. Grbac	53.1	41	−0.04	34	−0.05	34	0.01	38
C. Whelihan	48.0	42	−0.09	37	−0.11	38	−0.04	42
B. Hoying	45.6	43	−0.26	44	−0.29	44	−0.16	44
R. Leaf	39.0	44	−0.22	43	−0.23	43	−0.11	43

using our main rating but jumps to 4th if sacks and runs are excluded. This implies that Johnson was frequently sacked (he had 30 sacks relative to only 107 pass attempts!) and also explains how we ranked him 22nd, since we account for sacks, but the NFL rating ranked him 2nd. He is one of the best examples why the NFL rating is inadequate. Sacks are very costly to a team and Johnson was sacked much more than the average quarterback. As a comparison, fellow Buffalo Bills quarterback Doug Flutie had many more pass attempts (354), yet was sacked only 12 times. A quarterback rating should account for such costly plays.

Overall, the NFL ranking seems to place the highly regarded quarterbacks at the top and the lesser quarterbacks at the bottom. The rank correlation between the NFL ranking and our main ranking was 0.84. But there are definitely some deviations. First of all, the NFL rating ranks Bubby Brister of the Denver Broncos ahead of John Elway of the Denver Broncos. Any knowledgeable football enthusiast would wholeheartedly choose Elway over Brister. Our ranking does have Elway ahead of Brister. Another perplexing result is the fact that the NFL rating ranks Rob Johnson of the Buffalo Bills higher than Doug Flutie of the Buffalo Bills. Flutie had a terrific year, being named the starting quarterback ahead of Johnson and being selected to the Pro Bowl. Our ranking has Flutie ahead of Johnson.

Other rankings that differ significantly are for quarterbacks Neil O'Donnell and Eric Zeier. Neil O'Donnell is ranked 8th in the NFL rating but only 31st in our rating. This can be explained partially by his relatively high number of sacks (30 sacks relative to 343 pass attempts). O'Donnell also fumbled on three of the sacks, all of which were returned by the defense for touchdowns. Zeier's discrepancy can also be partially attributed to a relatively large number of sacks (18 sacks relative to 181 pass attempts).

Excluding runs had little effect on the ratings. Most of the

Table 4. Number of Plays With > 3 and 5 Expected Points Gained

# Plays > 3		# Plays > 5	
R. Cunningham	31	R. Cunningham	11
C. Chandler	21	S. Young	8
D. Bledsoe	20	B. Favre	8
S. Young	19	C. Chandler	7
J. Elway	19	T. Dilfer	6
S. Beuerlein	18	V. Testaverde	5
B. Favre	18	T. Green	4
V. Testaverde	16	D. Flutie	4
D. Marino	15	11 tied at	3
M. Brunell	14		
P. Manning	14		
J. Plummer	13		
T. Green	12		
R. Gannon	11		

quarterbacks that have a history of running did go down slightly (as expected) when runs were excluded, such as Steve Young, Steve McNair, and Kordell Stewart.

The play with the largest positive expected points gained belonged to Charlie Batch of the Detroit Lions. On third down and 10 with 98 yards to the goal line, Batch threw a touchdown pass. He went from -1.56 expected points to 7 points, thus gaining 8.56 expected points. The quarterback play with the largest negative expected points gained was mentioned earlier and belonged to Brad Johnson of the Minnesota Vikings. His interception on third down and 3 from the five-yard line was returned 95 yards for a touchdown, going from 4.32 expected points to -7 points, thus the expected points "gained" were -11.32.

We can also determine who had the most big plays. If we define a big play to be more than 3 gained expected points, then Randall Cunningham of the Minnesota Vikings had the most with 31 big plays (again we exclude plays in the last two minutes of each half). If we define a big play to be 5 gained expected points, then again Cunningham had the most with 11. The top quarterbacks with plays gaining more than 3 and 5 expected points are presented in Table 4. We can similarly determine who had the most bad plays. The quarterbacks with the highest number of plays gaining less than -3 and -5 expected points are presented in Table 5.

Table 5. Number of Plays With < −3 and −5 Expected Points Gained

# Plays < −3		# Plays < −5	
J. Plummer	19	B. Favre	6
B. Favre	18	V. Testaverde	5
T. Dilfer	16	K. Stewart	5
P. Manning	15	J. Plummer	5
S. Young	14	D. Hollas	5
D. Hollas	14	R. Leaf	4
T. Green	14	P. Manning	4
K. Stewart	13	D. Kanell	4
K. Collins	13	C. Chandler	4
R. Leaf	12	T. Dilfer	3
D. Bledsoe	11	S. Young	3
D. Marino	11	N. O'Donnell	3
T. Banks	10	G. Foley	3
V. Testaverde	9	E. Grbac	3
C. Whelihan	9	D. Marino	3
S. Beuerlein	9	D. Bledsoe	3
C. Chandler	9	11 tied at	2

Another interesting result is that many of the quarterbacks have negative average expected points. Thus, on average they contribute negatively with respect to expected points.

6. CONCLUDING REMARKS

Modeling a polychotmous response as a function of explanatory variables is a challenging statistical problem. This is especially true in the NFL example in which there are large deviations from monotonicity and lots of data. Our tiered logistic regression model adapts well to this scenario and the conditional structure provides extra insight into the data. Our model could be used in any multinomial response case and is especially useful in capturing complex functional forms.

The biggest hurdle that a ranking like ours must overcome is the complexity that is involved. The general public is used to simple formulas, so a ranking like ours can be intimidating. While acknowledging this limitation, we feel like the value added more than offsets the added complexity. The method we use is a major improvement through the fact that it does not force all plays with the same end-result to have the same value. Our expected point method of rating players could be extended to rating running backs, punters, offenses, defenses, and so on.

APPENDIX

The overall BIC can be expressed as

$$\begin{aligned}
\mathrm{BIC} &= -2\cdot\left[\log\left(\prod_{i=1}^{n}\prod_{t=1}^{T}\left[\frac{\exp(\mathrm{L}_t)^{\lambda_{it}}}{1+\exp(\mathrm{L}_t)}\cdot\phi_{it}+(1-\phi_{it})\right]\right)\right.\\
&\qquad\left.-\frac{1}{2}(\text{model degrees of freedom})\cdot\log(n)\right]\\
&= -2\cdot\left[\log\left(\prod_{t=1}^{T}\prod_{i:\phi_{it}=1}\left[\frac{\exp(\mathrm{L}_t)^{\lambda_{it}}}{1+\exp(\mathrm{L}_t)}\right]\right)\right.\\
&\qquad\left.-\frac{1}{2}(\mathrm{d}_1+\cdots+\mathrm{d}_T)\cdot\log(n)\right]\\
&= -2\cdot\log\left(\prod_{t=1}^{T}\prod_{i:\phi_{it}=1}\left[\frac{\exp(\mathrm{L}_t)^{\lambda_{it}}}{1+\exp(\mathrm{L}_t)}\right]\right)\\
&\qquad+(\mathrm{d}_1+\cdots+\mathrm{d}_T)\cdot\log(n)\\
&= -2\sum_{t=1}^{T}\sum_{i:\phi_{it}=1}\log\left(\frac{\exp(\mathrm{L}_t)^{\lambda_{it}}}{1+\exp(\mathrm{L}_t)}\right)\\
&\qquad+(\mathrm{d}_1+\cdots+\mathrm{d}_T)\cdot\log(n)\\
&= -2\left[\sum_{i:\phi_{i1}=1}\log\left(\frac{\exp(\mathrm{L}_1)^{\lambda_{i1}}}{1+\exp(\mathrm{L}_1)}\right)+\cdots\right.\\
&\qquad\left.+\sum_{i:\phi_{iT}=1}\log\left(\frac{\exp(\mathrm{L}_T)^{\lambda_{iT}}}{1+\exp(\mathrm{L}_T)}\right)\right]\\
&\qquad+(\mathrm{d}_1+\cdots+\mathrm{d}_T)\cdot\log(n)\\
&= -2\left[\sum_{i:\phi_{i1}=1}\log\left(\frac{\exp(\mathrm{L}_1)^{\lambda_{i1}}}{1+\exp(\mathrm{L}_1)}\right)+\mathrm{d}_1\log(n_1)+\cdots\right.
\end{aligned}$$

$$\begin{aligned}
&+ \sum_{i:\phi_{iT}=1} \log\left(\frac{\exp(\mathrm{L}_T)^{\lambda_{iT}}}{1+\exp(\mathrm{L}_T)}\right) + \mathrm{d}_T\log(n_T)\Bigg] \\
&+ \mathrm{d}_1\log(n) + \cdots + \mathrm{d}_T\log(n) - \mathrm{d}_1\log(n_1) \\
&- \cdots - \mathrm{d}_T\log(n_T) \\
= {} & -2\left[\sum_{i:\phi_{i1}=1} \log\left(\frac{\exp(\mathrm{L}_1)^{\lambda_{i1}}}{1+\exp(\mathrm{L}_1)}\right) + \mathrm{d}_1\log(n_1) + \cdots\right. \\
&\left. + \sum_{i:\phi_{iT}=1} \log\left(\frac{\exp(\mathrm{L}_T)^{\lambda_{iT}}}{1+\exp(\mathrm{L}_T)}\right) + \mathrm{d}_T\log(n_T)\right] \\
&+ \mathrm{d}_1 \cdot \log(\frac{n}{n_1}) + \cdots + \mathrm{d}_T \cdot \log(\frac{n}{n_T}) \\
= {} & \mathrm{BIC}_1 + \cdots + \mathrm{BIC}_T + \mathrm{d}_1 \cdot \log(\frac{n}{n_1}) + \cdots \\
&+\mathrm{d}_T \cdot \log(\frac{n}{n_T}).
\end{aligned}$$

Thus, minimizing the BIC at each tier does *not* ensure that the overall model will have overall minimum BIC. However, using a similar form of the BIC would ensure an overall minimum. This form uses $\log(n)$ instead of $\log(n_t)$ and is expressed as

$$\mathrm{BIC}_t^* = -2 \sum_{i:\phi_{it}=1} \log\left(\frac{\exp(\mathrm{L}_t)^{\lambda_{it}}}{1+\exp(\mathrm{L}_t)}\right) + \mathrm{d}_t\log(n).$$

[Received December 1999. Revised August 2000.]

REFERENCES

Albert, J. H., and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679.

Hosmer, D., and Lemeshow, S. (1989), *Applied Logistic Regression*, New York: Wiley.

Kass, R., and Raftery, A. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Kooperberg, C., Bose, S., and Stone, C. (1997), "Polychotomous Regression," *Journal of the American Statistical Association*, 92, 117–127.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*, New York: Chapman and Hall.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Stern, H. (1998), "American Football," in *Statistics in Sport*, ed. J. Bennett, London: Edward Arnold, pp. 3–23.

White, C. (2000), "Polychotomous Regression Applied to Futility Analysis and Expected Points Modeling," unpublished Ph.D. dissertation, Texas A&M University, Dept. of Statistics.