Predicting NFL Field Goal Conversions: A Logistic Random Effects Approach

Haben Michael (haben.michael)

CS 229 Report
December 13, 2013

Abstract

We apply logistic regression methods to predict the success of a field goal. Unable to find satisfactory prediction rates, we identify situations in which prediction is sounder. One of these situations is when the data is less overdispersed. We pursue this observation by applying a prior in the regression model, with modest improvement. We write software routines, where none to our knowledge is available, to estimate the parameters of this variant logistic regression model.

1 Introduction

The field goal is a key offensive play in American football. In the past 20 years of games in the National Football League, an average of 4 field goals were attempted per game, with an average 80% success rate. Of those games, 24% were decided by 3 or fewer points, the value of a successful field goal [9]. Even when a field goal does not directly decide the outcome of a game, the course of in-game strategy, both offensive and defensive, is affected by judgments on the likelihood of a field goal. Therefore, predicting the success of a field goal is of interest to team coaches and management, the casual fan, and participants in the multi-billion dollar NFL betting market.

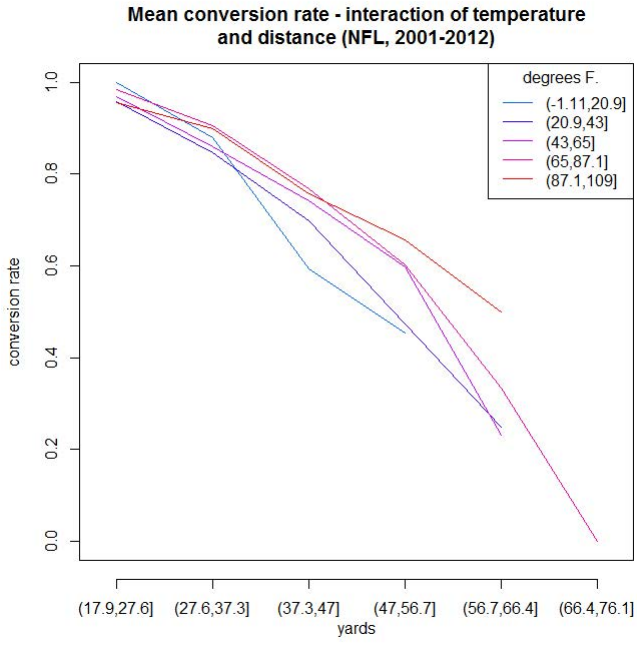

Figure 1: Distance from goal and temperature are two possible predictors of field goal success.

The nature of the play invites statistical analysis. Like the foul shot in basketball or penalty shot in soccer, the field goal is a relatively controlled, repetitive play within the game. To successfully score a field goal, a kicker must place kick the ball between the uprights and above the 10' crossbar of a goal structure. A key predictor of success is the distance from the goal, and relating field goal kick success to distance from the goal is a textbook application of logistic regression, e.g., [3]. Other predictors that may be of interest include the identity of the kicker, the offensive line, or weather conditions (Fig. 1).

Nevertheless, no models known to us were designed with a view toward prediction. They instead focus on identifying statistically significant explanatory variables. Even those that make recommendations to bettors or coaching staff fail to validate their models [1]. The goal of this project is to build a model to predict the success of a field goal kick in the NFL, with an emphasis on logistic regression.

2 Data

The data consists of 9,331 observations corresponding to kicks in the NFL regular and post-seasons years 2001-2012. These observations were obtained from a larger data set by excluding kicks in domed or otherwise covered stadiums, to permit an analysis of the effect of weather. (The results that follow, apart from those specifically referring to weather, were consistent with results obtained on the entire data set.) For each kick, available covariates include: the kicker, distance to goal, teams involved, season, date, and time, game clock, game score, game duration, and precipitation.

These data were obtained second-hand from participants at the MIT Sloan Sports Analytics Conference, March 1-2, 2013. We confirmed the accuracy of the data by scraping the same data for the years 2008-2012 from profootballreference.com.

The data set was augmented with weather data from the National Climatic Data Center via API accessible from *.edu domains. The weather data includes hourly temperature, humidity, wind-speed, and a "feels like" measure. In order to match the hourly weather data to a kick during a game, we assume that the game duration is divided evenly into four quarters. We then take the weather at a given kick to be the weather at the hour in which the corresponding quarter lies, or the end of regulation in the case of a kick during overtime.

We factorize temperature, humidity, and wind-speed into 5, 5, and 2 levels, respectively. We code precipitation into a binary variable, presence or not of any one of rain, light rain, light snow, flurries, or snow.

3 Logistic Regression Models Have Little Predictive Value

In the logistic regression model, the binary outcomes $K_{ij}, \ldots, K_{in_i}$ are grouped by level of the covariates x_i. The sums $Y_i = \sum_{j=0}^{n_i} K_{ij}, K_{ij} \in \{0, 1\}$, are modeled as independent binomial variables with success probabilities π_i depending on some linear combination of the covariate level [6]:

$$Y_i \overset{indep.}{\sim} binomial(n_i, \pi_i) \qquad (1)$$
$$\pi_i = logit^{-1}(\beta^T x^{(i)})$$

Maximum likelihood estimates of the coefficients β_i may be obtained by Newton-Raphson iteration.

We use logistic regression to model field goal success by the distance from the goal, ambient temperature, wind speed, precipitation, and whether the kick occurs in a postseason game. Truncated output from R is given in Fig. 2. The output flags many predictors as statistically significant under an asymptotic test. Even if the data do not meet the distributional assumptions underlying the significance conclusion, more general nonparametric tests also suggest a number of explanatory variables. For example, a rank sum test identifies the effect on conversion rate of whether Matt Stover vs. Sebastian Janikowski is kicking the ball, precipitation in the stadium, and a distance greater than 35 yards as significant at the 7.1%, 1.9%, and infinitesimal levels, respectively. These conclusions support the impression left by Fig. 1 that there is much information contained in the available predictors.

These promising results are not borne out when we use the model to make predictions. Initially, we make predictions on new observations by rounding the success probability at new data under the model, i.e., predicting success if and only if the predicted probability is greater than or equal to 0.5. We test the model with 10-fold cross-validation and find misclassification rates between 19% and 19.5%. This rate is disappointing because 19.16% of the field goals in the full data set were successful. Nearly all errors are errors of specificity: the model predicts too many kicks as successful. A simple constant term model ignoring all the predictors ("$success \sim 1$") would perform similarly.

When trained on evenly balanced subsets of the data (e.g., randomly select 1,500 made field goals, 1,500 missed), the logistic model's misclassification rates are in the range of 35-40%. The model thus does perform better than a constant term model according to cross-validation on this balanced subset.

```
Call:
glm(formula = good ~ dist + temp + wspd + postseason + precip,
    family = binomial(link = "logit"), data = fg)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7472   0.2604   0.4275   0.6809   1.6398

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     5.290323   0.192162  27.530  < 2e-16 ***
dist           -0.105941   0.003454 -30.672  < 2e-16 ***
temp            0.008396   0.001857   4.521 6.16e-06 ***
wspd           -0.022820   0.005267  -4.333 1.47e-05 ***
postseasonTRUE  0.049668   0.101991   0.487  0.62627
precipTRUE     -0.271290   0.102081  -2.658  0.00787 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 2: R output of logistic regression.

When we use this model to predict on the remaining (unbalanced) data set, however, misclassification rates are at 29.8%. This rate is worse than a constant term model on the remaining data set, which would misclassify 15.4%. Training on a balanced set appears only to have overfit to the balanced set.

We were unable to move beyond this impasse by common transformations or variable selection methods. Specifically, adding higher-order and interaction terms was ineffective, as were forward and backward elimination (on the main effects) using the likelihood ratio/chi-square and Wald statistics. We also tried link functions other than the logit. Though not the focus of this project, we note in passing that we attempted to use an off-the-shelf SVM learner with common kernels as well as LDA (the latter using the original non-factorized, continuous variables), without improvement to the cross-validation results.

4 Identifying Situations with Improved Prediction Rates

Without success improving prediction rates of the logistic model on the full data set, we shift focus and look to identify narrower circumstances in which the logistic model may perform well. We found this path to be treacherous for two principal reasons:

- First, we do not know how to test whether an improvement in the misclassification rate is significant or not.

For example, misclassification rates are lowest when data is restricted to temperatures in the 50s F. Some range of temperatures will have the minimum misclassification rate, we do not know the one we found is minimum randomly, and if we select it as a principle in making predictions we may overfit (where the point of CV in the first place was to avoid overfitting).

- Second, when we exclude certain sets of data, the conversion rate for the remaining data set may be higher than the already high 80% conversion rate of the entire data set. In that case, the misclassification rate may go down, but, as discussed above, the same can be said for a constant term model.

For example, it may seem natural to check whether the model makes better predictions if we focus on kickers for whom we have many observations. Fig. 3 displays the error from fitting and predicting on each kicker separately, plotted against the number of observations for that kicker. We notice a downward trend, emphasized in the figure by a linear regression line. However, it would probably be wrong to conclude that our model performs better when we only make predictions involving kickers with, say, 250 kicks in the data set. Fig. 4 shows that the kickers with more kicks in the data set also have higher conversion rates. Thus even a constant-term model would improve when we focus on kickers for whom we have more data.

We might have expected that kickers who are more reliable would be sent out more frequently to attempt kicks, over longer playing careers. A perhaps subtler, but more damning, example of the second bullet is to try to improve the model's performance by focusing on more recent years. This may be plausible since, as mentioned in Sec. 2, the data for years 2008-2012 was validated from 2 sources, or perhaps the

NFL's or NCDC's recording methods have improved over the last 10 years. As Fig. 5 shows, we do find that the logistic model makes better predictions in more recent years. Unfortunately, as the mean conversion rates per season overlaid in Fig. 5 show, that improvement is just the improvement one gets for free from surprisingly improved conversion rates in recent years. Again, a model that predicted success all the time would do about as well.

With these caveats in mind, two ways of improving prediction rates are given below. The first does not involve subsetting the data before modeling, so it avoids most of these issues. The second does involve subsetting the data, but the change in mean conversion rates is small compared to the improvement in misclassification error.

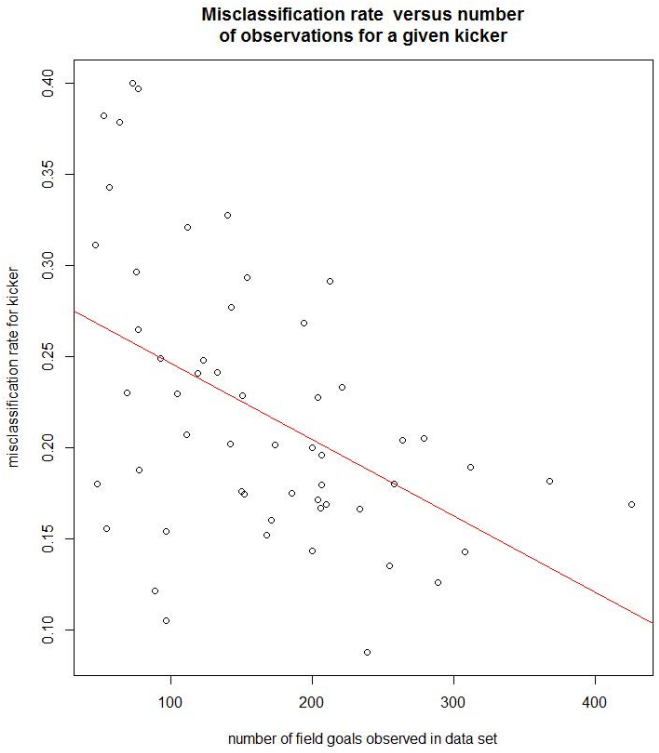

Figure 3: Prediction appears to improve when trained on kickers with more data points.

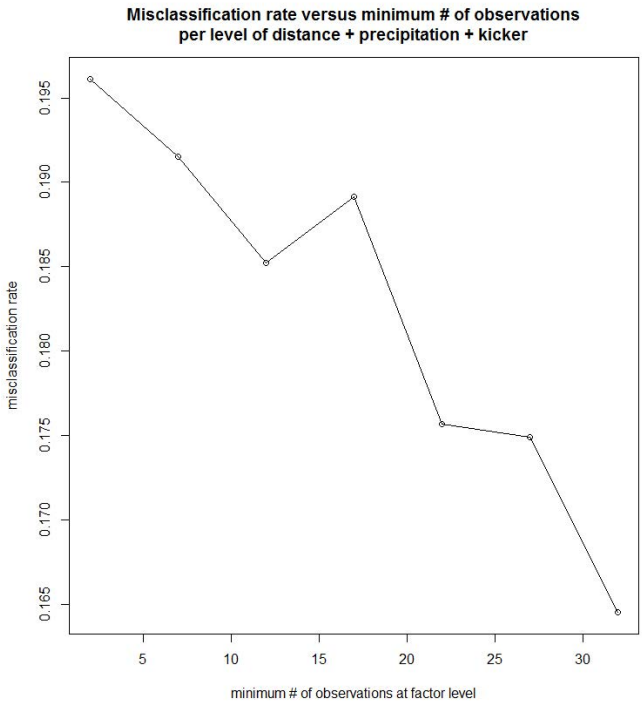

Figure 4: Kickers with more data points are the better kickers, explaining the improved prediction in Fig. 3.

4.1 Stricter Threshold for Making a Prediction

We may refrain from making a prediction except when the logistic model predicts success or failure with a certain level of certainty. Predictions discussed in Sec. 3 were made by predicting success if and only if the predicted probability on the new data was greater than or equal ot 0.5. We may instead make no prediction unless the predicted probability is more than a certain distance q from 0.5, $q \in (0, .5)$. Doing so cuts misclassification to arbitrarily small levels, at the expense of having no prediction model in cases with less certain predictions. Fig. 6 shows the misclassification rate when predictions are only made at levels of certainty corresponding to a range of q between 0 and 0.45, along with the proprtion of test cases where the that level of certainty is achieved. From the figure, we can cut misclassification rates in half to 10% if we refrain from making predictions in about 35% of cases.

4.2 Excluding Overdispersed Observations

Another way to improve misclassification rates is to exclude results that violate the logistic regression model (1) in some way. Model (1) implies a relationship between the moments of the Y_i, namely, $Var(Y_i) = n_i \mathbb{E}(Y_i)(1 - \mathbb{E}(Y_i))$. To estimate $Var(Y_i)$ and $\mathbb{E}(Y_i)$ we use respectively the sample mean $K_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} K_{ij}$ and sample variance

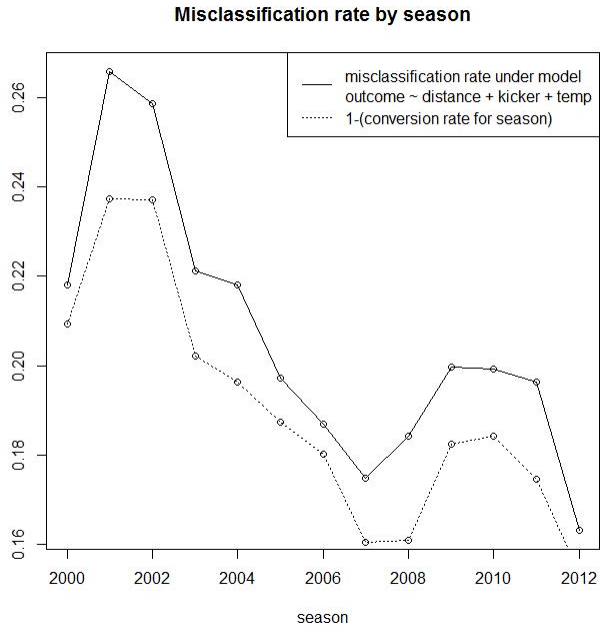

Figure 5: Misclassification rates in lockstep with overall conversion rates in the training set.

$\frac{n_i}{n_i-1}\sum_{j=1}^{n_i}(K_{ij}-K_{i\cdot})^2$, where as before $Y_i=\sum_{j=1}^{n_i}K_{ij}$. We can then exclude observations Y_i for which this observed variance estimate deviates too far from the predicted variance estimate $n_iK_{i\cdot}(1-K_{i\cdot})$. Fig. 7 shows the results. First, we order the Y_i by the excess of the observed variance estimate over the predicted. Then we exclude an upper quantile of the Y_i under this ranking, and fit and test the model on the remaining data. The misclassification error is plotted against the size of the upper quantile excluded, achieving 2.5-3% improvement.

Overdispersion, when the variances among the success probabilities exceed the binomial variance $n\pi(1-\pi)$ implied by the logistic model, is commonly encountered in logistic regression [6]. Among its causes are correlation among the bernoulli events and a grouping of the data missed by the model. For example, we compute $Var(Y_i)=Var(\sum K_j)=\sum_{j=1}^{n_i}Var(K_j)+\sum_{j\neq k}Cov(K_i,K_j)=n_i\pi(1-\pi)+\sum_{j\neq k}Cov(K_i,K_j)$. Thus, if the bernoulli events have nonzero covariance, the observed variance will exceed a binomial variance. One might expect such covariance in the NFL data if a kicker learns from his experience kicking under given conditions.

A common technique to handle overdispersion is a quasi-binomial model [6]. This model postulates a scaling factor σ^2 as an additional degree of freedom in (1) to tune the variance: $Var(_i)=n_i\pi_i(1-\pi_i)\sigma^2$. This model does not correspond to a real distribution for the observations Y_i; it is a binomial model but with a modified variance. One proceeds formally to estimate σ^2 from the data. This approach alters the standard errors of the logistic model coefficient estimates but the coefficient estimates are the same. Therefore, we did not find the quasi-likelihood model directly useful for making predictions. It is useful for avoiding conclusions that covariates are statistically significant when they are not, but a logistic model might suggest that they are (cf. Sec. 3).

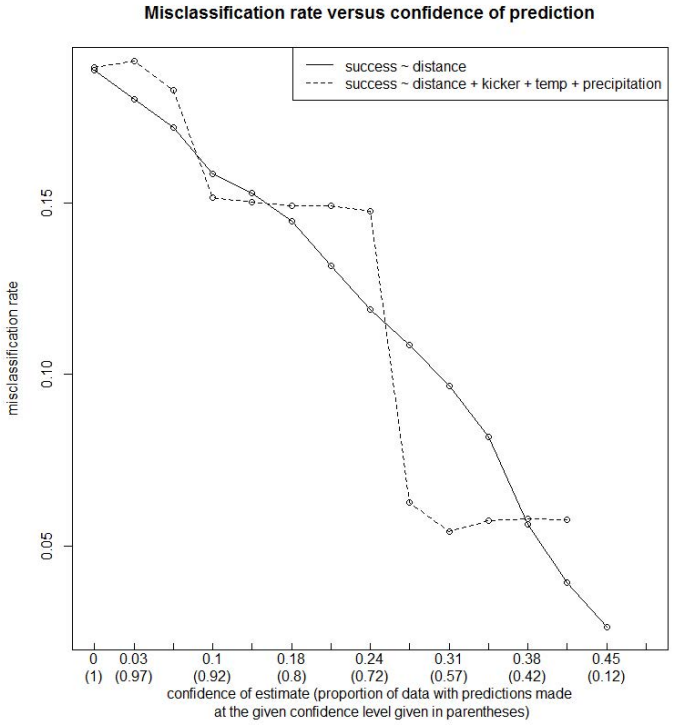

Figure 6: Improving prediction rates by ignoring new data on which the model predicts a success probability too close to 0.5.

5 Beta-binomial Model

We attempt a beta-binomial model to more accurately model the variances of the data. We modify (1) by taking the success probabilities to be random with a beta distribution,

$$\begin{aligned}\Pi_i &\sim Beta(\gamma\alpha_i,\gamma(1-\alpha_i)) \\ \mathbb{P}[\Pi_i=p] &\propto p^{\alpha\gamma-1}(1-p)^{(1-\alpha)\gamma-1}.\end{aligned} \tag{2}$$

The mean parameter of the beta distribution, rather than the success probabilities, is then estimated as the inverse logit of a linear combination of the feature observations.

$$\begin{aligned}Y_i|\Pi_i &\overset{indep.}{\sim} binomial(n_i,\Pi_i) \\ \Pi_i &\sim Beta(\gamma\alpha_i,\gamma(1-\alpha_i)) \\ \alpha_i &= logit^{-1}(\beta^T x^{(i)}).\end{aligned} \tag{3}$$

The marginal distribution of Y_i is then a beta-binomial distribution,

$$\mathbb{P}[Y_i=y]=\binom{n}{y}\frac{B(y+\alpha\gamma,n-y+(1-\alpha)\gamma)}{B(\alpha\gamma,(1-\alpha)\gamma)},$$

where $B(u,v)=\frac{\Gamma(u+v)}{\Gamma(u)\Gamma(v)}$ is the Beta function, with mean and variance

$$\begin{aligned}\mathbb{E}[Y_i] &= \alpha_i=\mathbb{E}[\Pi_i] \\ Var[Y_i] &= n_i\alpha_i(1-\alpha_i)\frac{\gamma+n_i}{\gamma+1}.\end{aligned} \tag{4}$$

The beta distribution is conjugate to the binomial and the posterior distribution of Π_i is beta as well: $\Pi|Y_i\sim Beta(\gamma\alpha+Y_i,\gamma(1-\alpha)+n-Y_i)$ with posterior mean $\frac{\gamma\alpha+Y_i}{\gamma+n}$.

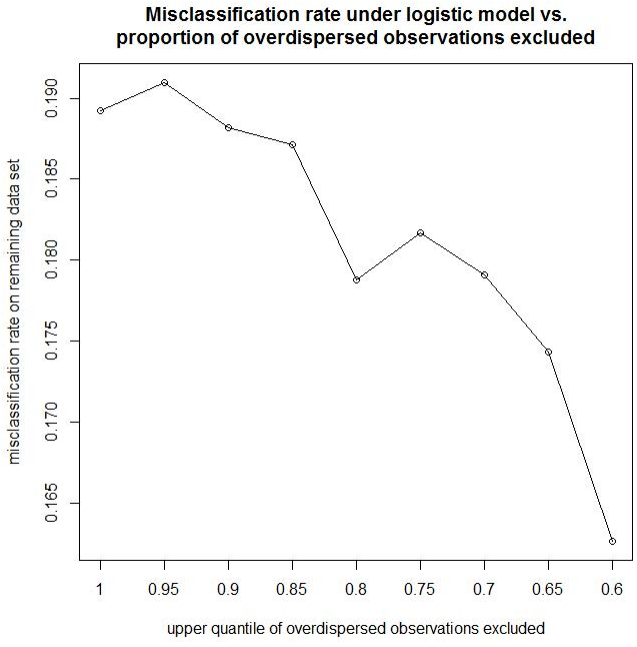

Figure 7: Improving prediction rates by excluding from the data set cases of overdispersion.

One may interpret this model as a "random effects" model of logistic regression [6]. Doing so may be sensible a priori as we expect there to be causes of variation in a kicker's ability unaccounted for by the model. From (4) the beta prior on π_i also provides a degree of freedom to account for the non-binomial variance that was encountered in the data set.

We can fit the model (3) as follows. First, we compute the MLEs of the free parameters, the regression coefficients $\beta_1, ..., \beta_p$ and the new parameter γ. Second, at each observation, we use the parameter estimates to compute the posterior mean of the the kick success probability Π_i. The second step is straightforward as we have the posterior mean. For the first step, we use Newton-Raphson to maximize the log-likelihood of the data. Dropping subscripts to focus on a single observation Y, applying the identity $\Gamma(t+1) = t\Gamma(t)$, and simplifying, we have as log likelihood,

$$l(\gamma, \alpha) = \sum_{k=0}^{y-1} log(\gamma\alpha + k) + \sum_{k=0}^{n-y-1} log(\gamma(1-\alpha) + k) - \sum_{k=0}^{n-1} log(\gamma + k) + const.,$$

where $const.$ is constant with respect to γ, β.

We compute the score function to be

$$\frac{\partial l}{\partial \gamma} = \sum_{k=0}^{y-1} \frac{\alpha}{\gamma\alpha + k} + \sum_{k=0}^{n-y-1} \frac{1-\alpha}{\gamma(1-\alpha)+k} - \sum_{k=0}^{n-1} \frac{1}{\gamma + k}$$

$$\frac{\partial l}{\partial \beta_i} = \left(\sum_{k=0}^{y-1} \frac{\gamma}{\gamma\alpha + k} + \sum_{k=0}^{n-y-1} \frac{-\gamma}{\gamma(1-\alpha)+k} \right) \frac{\partial \alpha}{\partial \beta_i}$$

$$\frac{\partial \alpha}{\partial \beta_i} = \sigma(-x^T\beta)(1 - \sigma(-x^T\beta))(-x_i), \sigma := logit^{-1}.$$

We compute the Hessian to be

$$\frac{\partial^2 l}{\partial \gamma^2} = \sum_{k=0}^{y-1} \frac{-\alpha^2}{\gamma\alpha + k} + \sum_{k=0}^{n-y-1} \frac{-(1-\alpha)^2}{\gamma(1-\alpha)+k} + \sum_{k=0}^{n-1} \frac{1}{(\gamma + k)^2}$$

$$\frac{\partial^2 \beta}{\partial \gamma \partial \beta_i} = \sum_{k=0}^{y-1} \left(\frac{-\alpha\gamma}{(\gamma\alpha + k)^2} + \sum_{k=0}^{n-y-1} \frac{\gamma(1-\alpha)^2}{(\gamma(1-\alpha)+k)^2} \right) \frac{\partial \alpha}{\partial \beta_i}$$

$$\frac{\partial^2 \beta}{\partial \beta_i \partial \beta_j} = \sum_{k=0}^{y-1} \left(\frac{-\gamma^2}{(\gamma\alpha + k)^2} + \sum_{k=0}^{n-y-1} \frac{-\gamma^2}{(\gamma(1-\alpha)+k)^2} \right) \frac{\partial \alpha}{\partial \beta_j} \frac{\partial \alpha}{\partial \beta_j} + \left(\sum_{k=0}^{y-1} \frac{\gamma}{\gamma\alpha + k} + \sum_{k=0}^{n-y-1} \frac{-\gamma}{\gamma(1-\alpha)+k} \right) \frac{\partial}{\partial \beta_j} \left(\frac{\partial \alpha}{\partial \beta_i} \right)$$

$$\frac{\partial}{\partial \beta_j} \left(\frac{\partial \alpha}{\partial \beta_i} \right) = \frac{1}{2} x_i x_j (1 - e^{-x^T\beta})(1-\alpha).$$

Our Newton-Raphson update rule is $(\gamma_{i+1}, \beta_{i+1}) := (\gamma_i, \beta_i) - Hessian(\gamma_i, \beta_i)^{-1} score(\gamma_i, \beta_i)$.

Although overdispersion is a commonly encountered issue in binomial models (6), and although the beta-binomial model is a commonly cited solution, its use in a logistic regression setting is less common. There are apparently no software routines publicly available to estimate the parameters (2). We wrote R routines to perform the above parameter estimation and make predictions using the estimates.

On 10-fold cross-validation we find a modest improvement of 1.5-2% in misclassification rate. On many of the training sets randomly selected by our CV routine we found that the estimate of γ was negative, which is outside the parameter space of the beta distribution. We clipped these estimates at 0. From (4), a negative γ corresponds to a smaller than beta-binomial variance, calling into question our conclusion that the data was overdispersed.

6 Conclusion

We were unable to use logistic regression to assist in making predictions on field goal rates in general. We did find that logistic regression made better predictions when 1) the predicted success probability was farther from 0.5 and 2) when overdispersion was not present. We applied a random effects model made a definite but small improvement to misclassification rates. Throughout, we did not find that including environmental covariates or the identity of the kicker improved prediction rates beyond the model regressing against only distance to goal.

The random effects model did not provide the hoped-for improvement. The logistic models attempted seemed to have a persistent difficulty with cases close to the decision boundary $\pi = 0.5$. An avenue of further work would be classifiers with more flexible decision boundaries.

R routines `bbr` and `bbr.predict` to predict the parameters of a logistic beta-binomial regression model are available at `www.stanford.edu/ habnice/cs229`, as well as Python scripts to retrieve the from the NCDC and profootballreference.com most of the data used.

7 References

[1] T.K. Clark, A.W. Johnson & A.J. Stimpson. Going for Three: Predicting the Likelihood of Field Goal Success with Logistic Regression. MIT Sloan Sports Analytics Conference, March 1-2, 2013.

[2] C. Czado. *Regression Models Lecture Notes*, TÜ Munchen, 2004.

[3] B. Efron. Maximum Likelihood and Decision Theory. *The Annals of Statistics* 10(2): 340-356, 1982.

[4] B. Efron. *Stats 306A Lecture Notes*, Stanford University, 2009.

[5] M.J. Kahn & A.E. Raftery. Discharge Rates of Medicare Stroke Patients to Skilled Nursing Facilities: Bayesian Logistic Regression with Unobserved Heterogeneity. Tech. Report 241, Aug. 1992, Statistics Dept., Univ. of Washington.

[6] P. McCullagh & J.A. Nelder. Generalized Linear Models, 2d ed. 1989.

[7] A. Ng. *CS 229 Lecture Notes*, Stanford University, 2013.

[8] www.ncdc.noaa.gov/cdo-web/webservices/, *accessed 12/13/2013*.

[9] www.pro-football-reference.com, *accessed 12/13/2013*.