Using Information Underlying Missing Data to Improve Estimation of NFL Field Goal Kicker Accuracy

Jarad Niemi with Dennis Lock, Dan Nettleton, and Casey Oliver

Iowa State University

November 19, 2016

Raw statistics

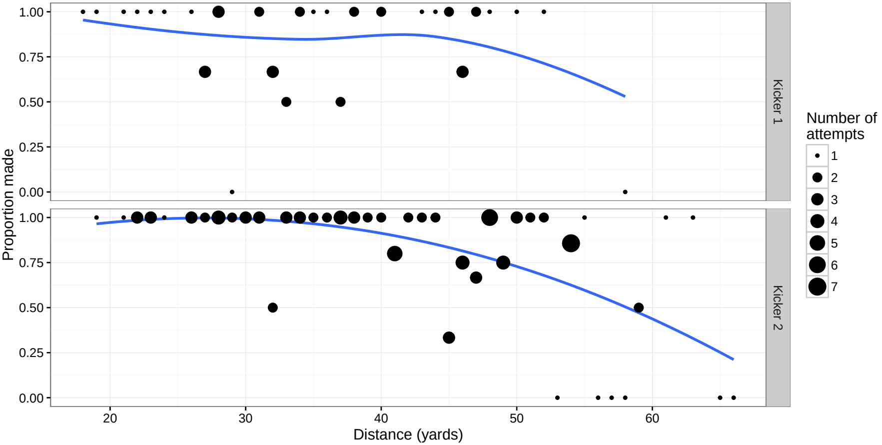
Let's compare two field goal kickers and their proportion of field goals made for the 2009-2011 seasons

Kicker	Proportion made
Kicker 1	0.84
Kicker 2	0.86

If we include the actual counts, we have some sense for uncertainty in these proportions:

Kicker	Number made	Number of attempts
Kicker 1	37	44
Kicker 2	90	105

Taking distance into account



Probit regression to account for explanatory variables

For attempt a for kicker k, let

- Y_{ak} be an indicator of success, i.e. 1 if successful and 0 otherwise, and
- x_{ak} be a vector of explanatory variables, e.g. distance, surface type, etc.

A probit regression model for each kicker k assumes

$$Y_{ak} \stackrel{ind}{\sim} Ber(\theta_{ak})$$

where $Y \sim Ber(\theta)$ indicates a Bernoulli distribution with

$$P(Y=1) = \theta \quad \text{and} \quad P(Y=0) = 1-\theta,$$

and the probability is determined by

$$\theta_{ak} = \Phi\left(\eta_{ak}\right) \qquad \eta_{ak} = x_{ak}^\top \beta_k$$

where Φ is the cumulative distribution function for a standrd normal.

Probit regression analysis

Explanatory variables included in this analysis:

- Distance
- Field surface: indicator for synthetic (as opposed to grass)
- Interaction between distance and field surface

In the model (dropping the k subscript):

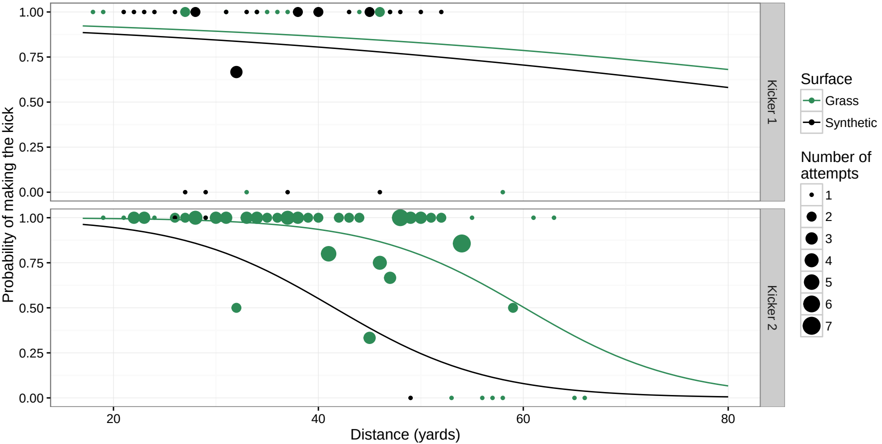
$$\begin{aligned} \eta_a &= \beta_0 + \beta_1 \text{Distance}_a + \beta_2 \text{Synthetic}_a \\ \eta_a &= \beta_0 + \beta_1 \text{Distance}_a + \beta_2 \text{Synthetic}_a + \beta_3 \text{Distance}_a \cdot \text{Synthetic}_a \end{aligned}$$

Number of observations:

Kicker	Grass	Synthetic
Kicker 1	16	28
Kicker 2	102	3

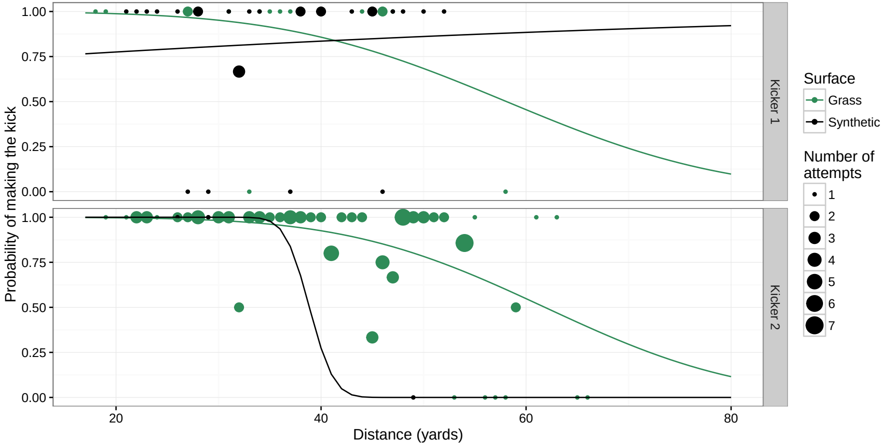
Probit regression analysis

$$\eta_a = \beta_0 + \beta_1 \text{Distance}_a + \beta_2 \text{Synthetic}_a$$



Probit regression analysis with an interaction

$$\eta_a = \beta_0 + \beta_1 \text{Distance}_a + \beta_2 \text{Synthetic}_a + \beta_3 \text{Distance}_a \cdot \text{Synthetic}_a$$



Hierarchical model to borrow information across kickers

A hierarchical probit regression model has the same initial structure:

$$Y_{ak} \stackrel{ind}{\sim} Ber(\theta_{ak}) \quad \theta_{ak} = \Phi\left(x_{ak}^{\top}\beta_k\right).$$

but then we assume a distribution for the β_k, e.g.

$$\beta_k \stackrel{ind}{\sim} N(\mu, \Sigma).$$

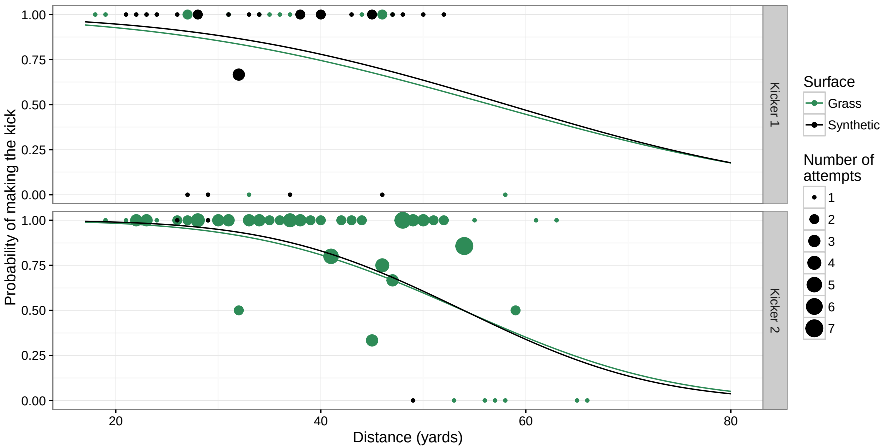
This distribution allows the data to inform us about the

- average effect of the explanatory variables across all kickers (μ) and
- the variability from kicker to kicker around this average (Σ).

If diagonal elements of Σ are estimated to be

- small then kickers are similar and we borrow a lot of information across kickers or
- large then kickers are dissimilar and we do not borrow much information.

Hierarchical probit regression model with interaction



Modeling attempts (or non-attempts)

For opportunity i for kicker k, let

- A_{ik} be an indicator of attempt, i.e. 1 if a kick was attempted and 0 otherwise, and
- w_{ik} be a vector of explanatory variables.

A probit regression model for each kicker k assumes

$$A_{ik} \stackrel{ind}{\sim} Ber(\pi_{ik}) \quad \pi_{ik} = \Phi(\zeta_{ik}) \quad \zeta_{ik} = w_{ik}^\top \alpha_k.$$

Two important explanatory variables for determining whether to take a kick are

- the probability of making the kick and
- will making this kick increase my chances of winning.

Explanatory variables for kick attempts

We already "know" the probability of making attempt i for kicker k, its θ_{ik}.

At any instant in the game, we can calculate a team's win probability using the method of Lock and Nettleton (2014). We use

$$\Delta_{ik} = WP_{ik}(\text{Successful kick}) - WP_{ik}(\text{Current}).$$

Final set of explanatory variables

$$\zeta_{ik} = \alpha_0 + \alpha_1 \Phi^{-1}(\theta_{ik}) + \alpha_2 \Delta_{ik} + \alpha_3 \Phi^{-1}(\theta_{ik}) \Delta_{ik}.$$

Informative missingness model

The full model is

$$\begin{aligned} Y_{ak} &\overset{ind}{\sim} Ber(\theta_{ak}) & \theta_{ak} &= \Phi\left(x_{ak}^\top \beta_k\right) & \beta_k &\overset{ind}{\sim} N(\mu, \Sigma) \\ A_{ik} &\overset{ind}{\sim} Ber(\pi_{ik}) & \pi_{ik} &= \Phi\left(\omega_{ik}^\top \alpha\ \right) \end{aligned}$$

$$\omega_{ik} = (w_{ik}, w_{ik}\Phi^{-1}(\theta_{ik})) = (w_{ik}, w_{ik}x_{ak}^\top \beta_k)$$

where

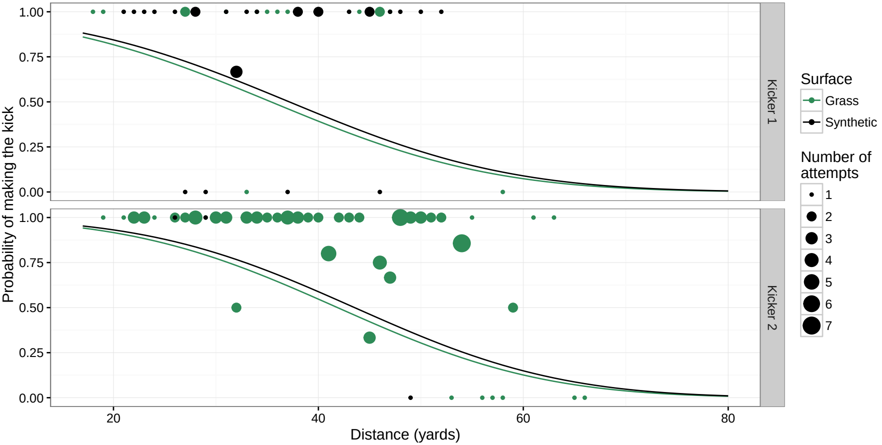
- x_{ak} is a set of explanatory variables that affect the probability of making a kick and
- w_{ik} is a set of explanatory variables that affect the probability of taking a kick.

Data for informative missingness model

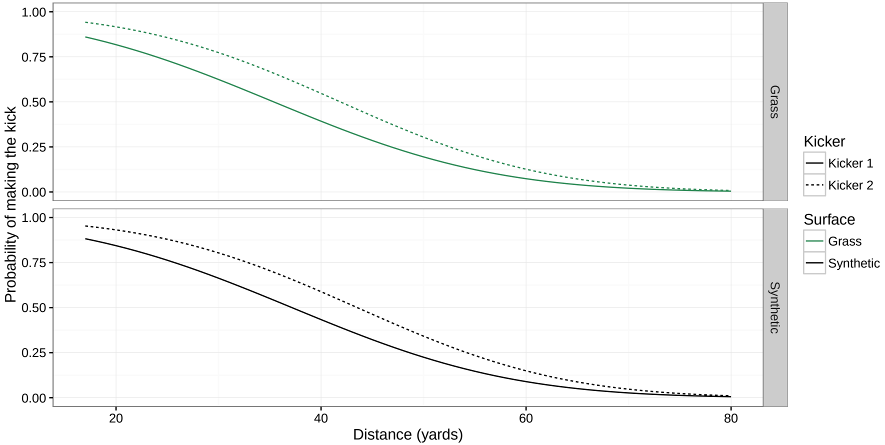
When considering what plays constitute an "opportunity" for a kicker, we considered 4th down plays when an attempted field goal would have been from a distance of no more than 76 yards, and

- a field goal was attempted or
- a field goal was not attempted even though making the field goal would have increased the team's win probability.

Informative missingness analysis



Kicker comparison



Summary

Constructed a hierarchical informative missingness model that

- borrowed information among the kickers
- incorporated information from non-attempts
- to estimate the probability of making a field goal
- as a function of explanatory variables, e.g. distance.

These slides are available at

- https://github.com/jarad/MwSAM2016/raw/master/JaradNiemi_MwSAM2016.pdf
- or on my website http://www.jarad.me/presentations.html.

Thank you!

Kicker name reveal