Lab: Models do what they're told

1. Expected Points in American Football

Expected points added (EPA) is a popular tool for evaluating NFL team offenses, team defenses, and quarterbacks. The success of a play can be captured by EPA by the difference between expected points (EP) before and after a play. The Expected points of a game-state x attempts to capture: what is the expected value of the net points of the next score in the half for the team possessing the ball?

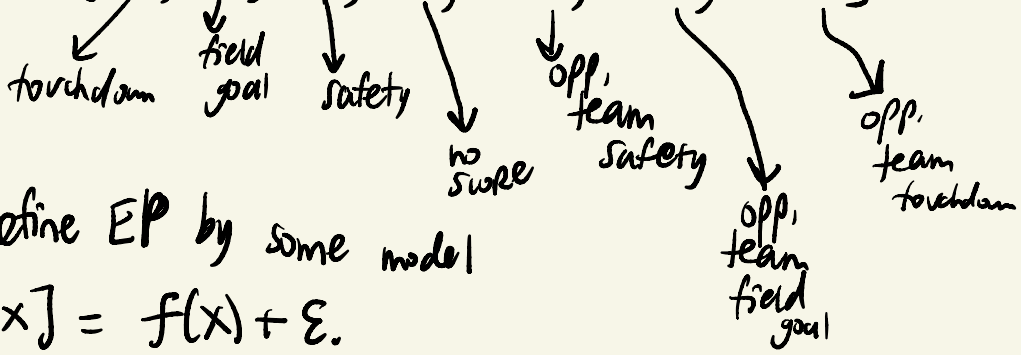
Expected points depends on many variables that comprise the game-state, such as yardline, down, yards to go until reaching a first down, time remaining in the half, and measure(s) of team strength or relative team strength (e.g., pre-game point spread relative to the team with possession.

For simplicity, let's consider just these 5 covariates.

This is not observable and is defined by a Model.

Letting y = net pts of the next score

$$\in \{7, 3, 2, 0, -2, -3, -7\}$$



we define EP by some model

$$\mathbb{E}[y|x] = f(x) + \varepsilon.$$

We want to estimate the conditional mean function x from data.

One way to do this is using multivariable linear regression, $y = x^T\beta + \varepsilon$, so $EP(x) := \mathbb{E}(y|x) = x^T\beta$.

Another way is multinomial logistic Regression,

$$\log \frac{\mathbb{P}(y=K)}{\mathbb{P}(y=0)} = x^T\beta_k \quad \text{for } K \in \{7,3,2,-2,-3,-7\}$$

or equivalently

$$\begin{cases} \mathbb{P}(y=K) = \dfrac{1}{1+e^{-x^T\beta_k}} & \text{if } K \neq 0 \\ \mathbb{P}(y=0) = 1 - \sum\limits_{K\neq 0} \mathbb{P}(y=K) \end{cases}$$

and then $EP(x) := \sum\limits_K K \cdot \mathbb{P}(y=K)$.

* **Building an additive model sequentially:**

Both of the aforementioned models are additive in that they have a $x^T\beta = x_1\beta_1 + x_2\beta_2 + \cdots + x_p\beta_p$ term.

Use multinomial logistic regression (in R: nnet package, multinom function) to model EP:

- as a purely linear function of yardline,

  $k \neq 0$, $\log \frac{\mathbb{P}(y=k)}{\mathbb{P}(y=0)} = \beta_{k0} + ydl \cdot \beta_{k1}$

  Plot EP vs. yardline.

  What's wrong with this model?

- use a spline on yardline to capture nonlinearities (in R: splines package, bs function)

  plot EP vs. yardline

- EP should differ by down. Should we make down a numeric

or categorical) variable? Why?

Make the right choice and model EP as a function of yardline and down. Plot EP (y axis) vs. yardline (x axis) and down (color).

- adjust for yards to go.
  Consider using a spline or log transform.
  Plot EP (y axis) versus yardline (x axis) and yards to go (color) and down (facet).
- adjust for time remaining in the half.
  Try a linear term $\beta \cdot$ (time) and a spline. Plot EP (y axis) versus yardline (x axis) and time (color) on 1st down and 10.
- With this modeling process, we see that models do what they're told.
  They capture the trends that we tell them to capture!

- Public EP models essentially end here and do not adjust for team quality (e.g., via pre-game point spread). They argue that there is no need to adjust for team quality since for team/player evaluation they want EP for an average offense facing an average defense. What's wrong with that?

- Let's call the best EP model you've just made M. Now make model M' which adjusts for pre-game point spread (e.g. using a linear term). Is EP from M' with pointspread = 0 different from EP from M?

- What we see here is Selection Bias, a massive bedeviler across all of sports analytics, a truly pervasive issue.

Note that good teams have more plays and good teams score more points.

Visualize these two ideas.

How does this explain that EP from M differs from EP from M' with point spread = 0?

Because of this selection bias, EP for randomly drawn teams differs from EP for average teams!

- ARe the following quantities the same or different and why:

(A) The percentage of all 3 point attempts in the NBA this year that were successful

(B) The "true" 3 point percentage of an average NBA player