

Ø Ø Ø

[FAQ](#) | [Register Now](#) | [Sign In](#)

[HOME](#) | [PHYSICAL SCIENCES](#) | [EARTH SCIENCES](#) | [LIFE SCIENCES](#) | [MEDICINE](#) | [SOCIAL SCIENCES](#) | [CULTURE](#) | [VIDEO](#) | [CONTRIBUTORS](#)

[HOME](#) > [CULTURE](#) > [SPORTS SCIENCE](#) > [RANDOM WALK](#)



## How To Predict College Football Games

By Dan Gillick | December 30th 2011 07:49 PM | 8 comments | [Print](#) | [E-mail](#) | [Track Comments](#)

[RSS](#) [Share / Save](#) [f](#) [t](#) [g+](#) [...](#) [Tweet](#) [Like](#)



Dan Gillick

The inaugural season of intercollegiate football took place in 1869. It consisted of two games: Rutgers played Princeton, and then a week later, they played again. Each team won once, so the "national championship" (awarded retroactively) was split. And despite the schools' bitter rivalry, the Rutgers newspaper reported an "amicable feed together" after the contests. Since then, the business of selecting a national champion in college football has grown considerably more complex.



Controversy has long come with the territory. In 1932, for example, the national championship was awarded to Michigan via the "Dickinson System" (Frank Dickinson, a professor of economics, devised a ranking system for comparing teams with varying opponent strengths) even though both USC and Colgate went 9-0-0, and not a single point was scored against Colgate all season. To make matters worse, USC defeated 3rd ranked Pitt in the Rose Bowl since Michigan's conference, the "Big Nine", didn't allow postseason games.

By 1940, Dickinson retired as official ranker of football teams, ceding authority to the Associate Press (AP) poll of sportswriters, which had begun four years earlier. A competing coaches poll started in 1950. Today, a national champion is selected by the somewhat infamous Bowl Championship Series (BCS). The BCS ranks teams by averaging rankings derived from the AP poll (taken over by the Harris Interactive poll), the coaches poll (run by USA Today and ESPN), and 6 algorithmic ranking systems; the top two teams play a single game to determine, once and for all, the national champion.

Still, the BCS remains contentious. Or as Hal Stern, a statistician at UC Irvine says, "from a statistical point of view, the BCS is a bit of a sham." This year, while LSU is the clear choice for the top ranking,

### KNOW SCIENCE AND WANT TO WRITE?

[Register Now To Get Your Own Column!](#)

[Top Articles](#) [New Comments](#) [Events](#)

1. LAWS And The Autonomy Of Killer Robots
2. Columbia's Lame Oz Defense
3. Sorry Star Trek Fans, Tau Ceti Is Not The Next Earth
4. Screen For Celiac Disease In High-Rsk Groups Such As Diabetics, Anemics And Down's Syndrome
5. 5X: Alarming Rise In Cost Of Multiple Sclerosis Drugs Over 20 Years
6. Household Pets Can Transmit Infections To People
7. Are Insects The Future Of Food?



Chilesaurus diegosuarezi: Bizarre 'platypus' dinosaur discovered

We're wired to hate dieting

Oil or fat? Saturated fatty acids might directly damage heart

How to short circuit hunger

Neurons 'surgically' rewrite their DNA

[more](#)

[Links](#) [Hot Topics](#)

### Science 2.0 Links

Take a look at the best of Science 2.0 pages and web applications from around the Internet!

Alabama barely out-pollled Oklahoma State for the second spot. But, Alabama already lost (in a close game, 9–6) to LSU earlier in the season, so many fans feel that Oklahoma State deserves a chance.

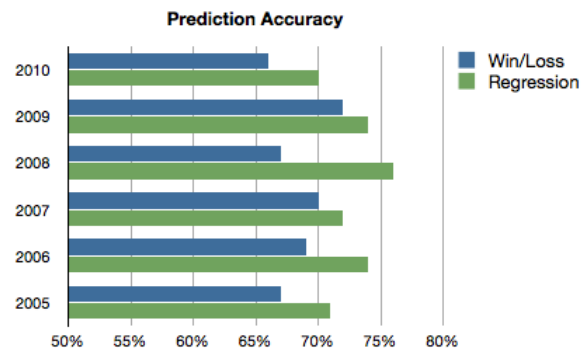
		BCS		Harris Poll			USA Today			Computer Rankings									
RK	TEAM	AVG	PVS	RK	PTS	%	RK	PTS	%	AVG	A&H	RB	CM	KM	JS	PW	%		
1	LSU	1.0000	1	1	2875	1.0000	1	1475	1.0000	1	1	1	1	1	1	1	1.000		
2	Alabama	.9419	2	2	2723	.9471	2	1399	.9485	3	3	3	3	3	2	2	.930		
3	Oklahoma State	.9333	3	3	2654	.9231	3	1367	.9268	2	2	2	2	2	3	3	.950		
4	Stanford	.8476	4	4	2504	.8710	4	1286	.8719	5	4	4	5	8	10	7	.800		
5	Oregon	.7901	9	5	2372	.8250	5	1232	.8353	8	12	5	8	10	9	6	.710		

All this controversy got me curious. What exactly are the ranking algorithms? How well do they work? Maybe I could do better.

If we're going to talk about assigning values to teams, we need to start by identifying a goal. Ours will be: maximize the prediction accuracy of future games. While this objective may diverge from that of the fans', as it may in the case of Alabama and Oklahoma State, at least it is easily measurable. The college football season is roughly 15 weeks long (including bowl games); we'll use the first 7 weeks to predict the results in week 8, then the first 8 weeks to predict week 9, and so on. This sort of rolling prediction, starting midseason, gives the algorithms enough data to make reasonable judgements.

Why can't we just use the win-loss records? That is, choose the team that has the higher value of *wins* / (*wins* + *losses*) as the predicted winner. We'll try this as a baseline, and see how much better we can do. But the main problem is that there are 120 division 1–A college football teams in 12 conferences, and each team's schedule is very different. This is the "strength of schedule" idea. If each team played every other 1–A team, the win-loss record would work just fine; but because they play a particular small subset (typically around 13 games), the win-loss record needs to be adjusted to reflect the quality of the wins and losses.

Let's start with some numbers. This chart shows how predictions made based on win-loss records compare to regression-based predictions (described below). Over 2051 games in six seasons, win-loss predictions average 68.5% accuracy, compared to 72.9% accuracy for regression. The gap is a bit more impressive for out-of-conference games, since the win-loss ratio better reflects a team's strength in its own conference: 73.7% for win-loss predictions vs. 79.1% for regression (measured across 344 games).



So how can we improve on the win-loss predictions? It appears that most (if not all) of the algorithms developed for the BCS start with the win-loss records and adjust them in various ways, sometimes mathematically justifiable (Sagarin and Colley, for example), and other times not (Billingsley says "I never had calculus. I don't even remember much about algebra"). But after the PAC-10 complained in 2001 that Oregon was unfairly denied a shot at the national championship, the BCS decided to preclude the use of "margin of victory" in algorithmic ranking. That is, all wins are to be treated equally, whether 8–7 or 50–7. This means that all the BCS algorithms are at a disadvantage.

I can measure the disadvantage by removing the margin from the model. To summarize:

Win-loss: 68.5%  
 Regression model without margin of victory: 70.2%  
 Regression model with margin of victory: 72.9%

How does this regression business work? Here's the idea in its most basic form. Suppose every team has an underlying strength such that the predicted point spread of a game is the difference between the teams' strengths. Say LSU's strength is 30 and Alabama's is 28, so the predicted spread would be  $30 - 28 = 2$  points.



**ScientificBlogging**  
 Like

8,392 people like ScientificBlogging.



Facebook social plugin

#### BOOKS BY WRITERS HERE

### REAL CLEAR SCIENCE

An Open Letter to Dr. Oz  
 Landing a Rocket on It's Tail  
 Making Oxygen on Mars  
 On Nuking Asteroids  
 Why Physics Needs Philosophy

more

SWAG Forum

This description is a model for how games actually play out. Of course in reality, there are players and coaches and strategies and plays and injuries and weather conditions that all effect the outcome of a game. I'm condensing all this into a single made-up thing called "strength". While this is a pretty drastic simplification, it has the advantage that we can actually do a good job of estimating strength values. This is called "training": choose the strength values that come closest to reproducing the actual point spreads observed during the season. The beauty of this optimization is that the strength values are derived from a compromise involving all the game outcomes simultaneously.

Here's a toy example. Suppose there are 3 teams, x, y, and z that play the following schedule:

week 1: x vs. y result: 21-3  
 week 2: y vs. z result: 15-18  
 week 3: x vs. z result: 10-42

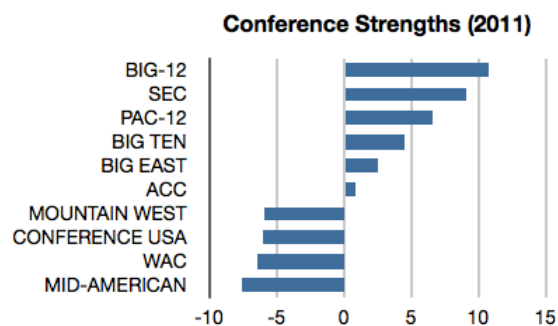
So we'd like to find strengths  $S_x$ ,  $S_y$ , and  $S_z$  such that this quantity is as small as possible:

$$[(S_x - S_y) - (21 - 3)]^2 + [(S_y - S_z) - (15 - 18)]^2 + [(S_x - S_z) - (10 - 42)]^2$$

Note that each term in this equation is the squared difference between a prediction and the actual result. The reason we use the squared differences is just because it makes the solution easier to find. This is a standard linear regression, sometimes called "ordinary least squares" because of the squared differences, and it has a straightforward algebraic solution. For the toy example, the solution is  $x = -4.7$ ,  $y = -7$ , and  $z = 11.7$ . The version of the equation we need to estimate all the division 1-A team strengths has 120 variables, one for each team, and about 700 terms, one for each game in the season. This is a relatively small learning problem, and my laptop solves it almost instantaneously.

Next, let's extend the model to include some other factors. We'll add a variable to represent Home Field Advantage (HFA). We could add an HFA variable for each team, as some teams may benefit more from being at home than others. But this turns out to work no better than the simple version (in terms of predictive accuracy), so I'll stick to a single HFA variable. Home field advantage, it turns out, is worth something like 3.5 points.

Let's also add a variable for each conference. This works as a kind of smoothing: there are far fewer conference variables than team variables, so their values should be better estimated. The overall strength is the sum of the conference strength and the leftover team strength. Here are the estimated conference strengths for 2011:



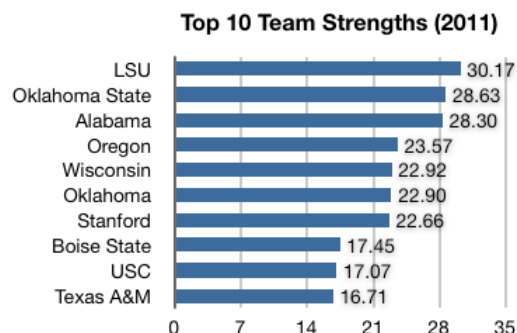
So far, the model assumes that teams either don't change (improve or get worse over the season) or all change at the same rate. Hal Stern and others have shown that the polls used by the BCS significantly over-value recent results, at least if what they care about is predictive accuracy. But that doesn't mean it isn't worthwhile to model change in some way. I found a small benefit from adding a variable for each team that indicates time. Perhaps the simplest way to do this is with a "week" counter. In the first week of the season, the counter is set to 1 for each team, then 2 during the second week, and so on. If a team improves during the season, its time variable will be positive, so that the team's overall strength increases as the season progresses; team's that suffer from injuries or get worse for other reasons will get a negative time variable. One small adjustment on this idea is to curve the week counter so that it starts at 0 and approaches 1 towards the end of the season, reflecting the notion that teams change more at the beginning of the season. These features reveal that while all the top 20 teams showed some improvement over the 2011 season, Kansas State and Texas improved very little, while Houston improved the most of any division 1-A team.

There's one last piece to the prediction story. At the end of November, nobody was quite sure what to make of Houston: they were 12-0, including some impressive blowout wins (56-3 against East Carolina, 73-17 against Tulane), but they're in the relatively weak Conference USA, and never had to

play a very strong opponent. So as far as the regression equation is concerned, Houston's strength value could well be larger than any other team's. This defies our intuition: they probably are not really better than once-defeated Alabama or Oklahoma State.

To address this issue we'll make one small change to the optimization: we'd like to find the set of strength values that reproduce the observed point spreads as closely as possible (as before), but we want those strength values to be as small as possible.

This is called "regularization". Now, Houston will have the smallest possible strength value that satisfies the regression equation, and in fact, they end up with a reasonable ranking. Here are the top 10, as of December 29th:



According to our model, then, Oklahoma State may indeed deserve a chance to play LSU for the national championship; LSU remains the favorite by about 1.5 points.

So how do my predictions compare with others? Todd Beck maintains a page that tracks over 50 different ranking/prediction systems, and the system I've described here performs close to the top of the list each year. Without doing any more careful analysis, I'd guess it's at least as good as the best existing systems (note that nothing I'm doing here is novel).

And since it's bowl season, I'll conclude with a few predictions:

The Vegas line favors Oregon over Wisconsin by 6.5 points in the Rose Bowl, but the model predicts Oregon by only half a point: I'd pick Wisconsin to beat the spread.

Public opinion has soured a bit on Houston since they were upset after going 12-0. The line favors them over Penn State in the Ticketcity Bowl by 5.5, while the model gives Houston a 10.5 point edge: I'd pick Houston to beat the spread.

As a Cal alum, I'd pick Oklahoma State to beat the 3.5 point spread against Stanford in the Fiesta Bowl since the model favors OK State by 6 points.

#### References:

On the first football game: <http://scarletknights.com/football/history/first-game.asp>

Dickinson system in 1932: <http://mvictors.com/?tag=dickinson-system>

BCS details: [http://en.wikipedia.org/wiki/Bowl\\_Championship\\_Series](http://en.wikipedia.org/wiki/Bowl_Championship_Series)

Current BCS standings: <http://espn.go.com/college-football/bcs>

NYT article by authors of "Death to the BCS":

<http://www.nytimes.com/2010/10/17/sports/ncaafootball/17score.html>

Hal Stern on college football: <http://pubs.amstat.org/doi/abs/10.1198/000313004X2098>

Performance of ranking systems: <http://www.thepredictiontracker.com/ncaareports.php?&orderby=wpct%20desc&year=11>

#### Random Walk

##### MORE ARTICLES

- How To Predict College Football Games
- Where Are All The Aliens?
- Predicting Movie Ratings: The Math that Won the Netflix Prize

[All Articles](#)

##### ABOUT

Wesleyan University → UC Berkeley → Research at Google I'm not sure where this walk will take me, but that's part of the point of walks....

[View Profile](#)

#### RELATED ARTICLES ON SCIENCE 2.0

[Are Americans Polarized Or Is Polling More Scientific?](#)  
[Statistics Suggest Brett Favre Will Not Win The Superbowl This Year](#)  
[Hey Football Players. We Can Get Hurt During The Super Bowl Too](#)  
[Science Of Baseball Playoffs: Giants To Win LDS, But Don't Bet The House](#)  
[Rhode Island team forfeits due to opposing football team size](#)

---

## COMMENTS

What happened to prediction accuracy as the number of games included in the regression increased? Was accuracy greater with more games to use as data or was there no pattern (which would be my prediction)?

Link

Bob (not verified) | 12/31/11 | 15:35 PM

Good question-- prediction accuracy does increase as the number of training games increases. For example, if you only use half the available training data, the average accuracy drops below the win/loss prediction accuracy. This is because there's no longer enough data to estimate all the parameters well. This also means that a longer season could better determine the top two teams.



Link

Dan Gillick | 12/31/11 | 16:12 PM

Did you calculate the strength for all teams, not just the top 10? If so, I would like to take your numbers and prepare a retrospective and prospective comparison for this bowl season. I am curious as to how you would stand up compared to Mr. Beck's prediction tracker for this bowl season. I reference his collection of predictions quite often. Sometime I even extract the top 10 predictions (by YTD performance) to create an additional, ASS-umably a more precise prediction for bowl season. I compare the two to determine any fallout or large variances.

Link

Anonymous (not verified) | 01/01/12 | 16:25 PM

Yes, I did. If you email me directly, I'll send them to you.



Link

Dan Gillick | 01/02/12 | 02:01 AM

Very interesting. What sticks out like a sore thumb though is A&M at number 10 -- a team that had six losses. However, a lot of them were close losses to good opponents, so I'm wondering if the flaw in your method is it gives too much credit for "quality" losses?

Link

Normalman (not verified) | 01/04/12 | 08:15 AM

This is a classic BCS-style question: shouldn't team X be ranked lower or higher given some particular qualitative observation?. I'm glad you asked it! At the outset, I said that my goal is to predict the outcomes of future games. You might disagree with this goal as a criteria for creating a ranking. But given my particular goal, I'd bet you would be hard-pressed to improve on the predictions by making adjustments to the output of the linear regression based on an observation about "quality losses".



Why do I say that? There's a mathematical reason: linear regression is optimal in the sense that if you accept my model of how football games are decided, you won't be able to choose better strength values for the teams.

There's also a more empirical reason: lots of people have worked on the problem of predicting football outcomes, so there's been a huge amount of the kind of twiddling you are suggesting. And none of it, not by BCS rankers like Billingsley, not by coders in the [ESPN winning formula competition](#), not by those who are indexed on Todd Beck's prediction tracker page, have been able to do any better than what I've presented here.

Link

Dan Gillick | 01/04/12 | 13:24 PM

Would a system that instead determined strength by regressing around (yard offense--yards defense) and/or turnovers instead of points scored provide a better measure? That might provide a better view of how good a team really is than score, which might just make teams that loose close look better than they are (though it'd also be more work).

I hope you'll post how your system worked in predicting bowl games against the spread. 30 bowl games ought to

4/27/2015

How To Predict College Football Games

give an idea whether this method is better than the "market" wisdom.

Link

Normalman (not verified) | 01/05/12 | 09:24 AM


It's certainly possible that using more descriptive features could give better estimates, but I haven't tried any of that. I think some companies (Accuscore?) do predictions by simulating games, which probably involves many more detailed statistics.


My system is better than 50% against the spread, but probably not better than 55%, so it would need to be improved to actually make a profit given 10% commissions.


Link


Dan Gillick | 01/05/12 | 15:25 PM

Who's Online?









440 guests

About Us | Contact Us | RSS | Terms | Privacy | Copyright and Removal | Advertise with Us

© 2015 ION Publications LLC

http://www.science20.com/random\_walk/how\_predict\_college\_football\_games-85816

6/6