Stats216: Session 2

Linear Regression Analysis of NCAA Basketball Data

In this in-class session, we will analyze a data set containing the outcomes of every game in the 2012-2013 regular season, and the postseason NCAA tournament. There are 5541 games and 347 teams. Our goals will be to:

- Estimate the quality of each team, to obtain objective rankings
- Explain the outcomes in the regular season and test hypotheses that may interest us
- Predict the winner and margin of victory in future games

1. Loading in the data

First, read in the files games.csv and teams.csv into the data frames games and teams. We will load them “as is” (i.e. strings not converted to factors).

```r
# Load data

# Load games data
games <- read.csv("http://statweb.stanford.edu/~jgorham/games.csv",as.is=TRUE)

# Load teams data
teams <- read.csv("http://statweb.stanford.edu/~jgorham/teams.csv",as.is=TRUE)
```

The games data has an entry for each game played, and the teams data has an entry for each Division 1 team (there are a few non-D1 teams represented in the games data).

```r
head(games)[, !(names(games) %in% c('neutralLocation', 'gameType'))]
```

```r
##    date              home             away
## 1 2012-11-09 albany-great-danes duquesne-dukes
## 2 2012-11-11 ohio-state-buckeyes albany-great-danes
## 3 2012-11-13 washington-huskies albany-great-danes
## 4 2012-11-17 albany-great-danes umkc-kangaroos
## 5 2012-11-18 albany-great-danes loyola-(md)-greyhounds
## 6 2012-11-20 south-carolina-state-bulldogs albany-great-danes
```

Two columns have been omitted above: the column neutralLocation is 1 if the game was not a real home game for the nominal home team, and 0 otherwise. gameType is REG for regular-season games, NCAA for NCAA tournament games, and POST for other postseason games (not the NCAA tournament).

The teams data has an entry for each team coding its name, conference, whether it made the NCAA tournament, and its AP and USA Today ranks at the end of the regular season (1-25 or NA if unranked).

```r
head(teams)
```
### team conference inTourney apRank usaTodayRank

<table>
<thead>
<tr>
<th></th>
<th>team</th>
<th>conference</th>
<th>inTourney</th>
<th>apRank</th>
<th>usaTodayRank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>stony-brook-seawolves</td>
<td>america-east</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>2</td>
<td>vermont-catamounts</td>
<td>america-east</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>3</td>
<td>boston-university-terriers</td>
<td>america-east</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>4</td>
<td>hartford-hawks</td>
<td>america-east</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>5</td>
<td>albany-great-danes</td>
<td>america-east</td>
<td>1</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6</td>
<td>maine-black-bears</td>
<td>america-east</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

Finally, we make one vector containing all the team names, because the three columns do not perfectly overlap:

```r
call.teams <- sort(unique(c(teams$team, games$home, games$away)))
```

### 2. How to Rank the Teams?

Now, spend a few minutes to come up with a way to rank the teams, based on all of the regular-season games. Try your method, or a couple of different methods, and compare the ranks you get to the official end-season rankings. Which rankings do you find more credible?

**[Bonus Question]** Think about the following bonus question if you have extra time. I will not be going over the answers to bonus questions in class, but you can ask the circulating staff.

- Does your method account for a team’s strength of schedule (the quality of a teams’ opponents)? If not, how might you modify it so that it does?

**Answers:**

One simple way to rank teams is by comparing their average score to their opponents’ average score.

```r
## Function to compute a team's total margin of victory
total.margin <- function(team) {
  with(games,
    sum(homeScore[home==team]) + sum(awayScore[away==team])
    - sum(homeScore[away==team])
    - sum(awayScore[home==team]))
}

## Function to compute the number of games a team played in
number.games <- function(team) {
  with(games, sum(home==team) + sum(away==team))
}

## Compute total margin and number of games for each team
margins <- sapply(teams$team, total.margin)
number.games <- sapply(teams$team, number.games)
## check: make sure names line up
mean(names(margins) == names(number.games))

## [1] 1

## compute average
margin.per.game <- margins / number.games
```
```r
rank.table <- cbind("Margin (Avg)" = margin.per.game,
                    "Margin Rank" = rank(-margin.per.game,ties="min"),
                    "AP Rank" = teams$apRank,
                    "USAT Rank" = teams$usaTodayRank)

margin.top25 <- order(margin.per.game,decreasing=TRUE)[1:25]
rank.table[margin.top25,]
```

<table>
<thead>
<tr>
<th>Margin (Avg)</th>
<th>Margin Rank</th>
<th>AP Rank</th>
<th>USAT Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>florida-gators</strong></td>
<td>17.027027</td>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td><strong>indiana-hoosiers</strong></td>
<td>16.527778</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td><strong>gonzaga-bulldogs</strong></td>
<td>16.294118</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td><strong>louisville-cardinals</strong></td>
<td>15.675000</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td><strong>kansas-jayhawks</strong></td>
<td>13.243243</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td><strong>pittsburgh-panthers</strong></td>
<td>13.181818</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td><strong>weber-state-wildcats</strong></td>
<td>12.972222</td>
<td>7</td>
<td>NA</td>
</tr>
<tr>
<td><strong>middle-tennessee-blue-raiders</strong></td>
<td>12.676471</td>
<td>8</td>
<td>NA</td>
</tr>
<tr>
<td><strong>virginia-commonwealth-rams</strong></td>
<td>12.416667</td>
<td>9</td>
<td>NA</td>
</tr>
<tr>
<td><strong>michigan-wolverines</strong></td>
<td>11.923077</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td><strong>duke-blue-devils</strong></td>
<td>11.888889</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td><strong>syracuse-orange</strong></td>
<td>11.725000</td>
<td>12</td>
<td>16</td>
</tr>
<tr>
<td><strong>stephen-f.-austin-lumberjacks</strong></td>
<td>11.448276</td>
<td>13</td>
<td>NA</td>
</tr>
<tr>
<td><strong>belmont-bruins</strong></td>
<td>11.322581</td>
<td>14</td>
<td>NA</td>
</tr>
<tr>
<td><strong>creighton-bluejays</strong></td>
<td>11.277778</td>
<td>15</td>
<td>22</td>
</tr>
<tr>
<td><strong>saint-mary’s-gaels</strong></td>
<td>11.205882</td>
<td>16</td>
<td>NA</td>
</tr>
<tr>
<td><strong>davidson-wildcats</strong></td>
<td>11.176471</td>
<td>17</td>
<td>NA</td>
</tr>
<tr>
<td><strong>ohio-state-buckeyes</strong></td>
<td>10.864865</td>
<td>18</td>
<td>7</td>
</tr>
<tr>
<td><strong>ole-miss-rebels</strong></td>
<td>10.277778</td>
<td>19</td>
<td>NA</td>
</tr>
<tr>
<td><strong>saint-louis-billikens</strong></td>
<td>10.028571</td>
<td>20</td>
<td>13</td>
</tr>
<tr>
<td><strong>arizona-wildcats</strong></td>
<td>9.828571</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td><strong>stony-brook-seawolves</strong></td>
<td>9.718750</td>
<td>22</td>
<td>NA</td>
</tr>
<tr>
<td><strong>memphis-tigers</strong></td>
<td>9.555556</td>
<td>23</td>
<td>19</td>
</tr>
<tr>
<td><strong>missouri-tigers</strong></td>
<td>9.411765</td>
<td>24</td>
<td>NA</td>
</tr>
<tr>
<td><strong>north-dakota-state-bison</strong></td>
<td>9.218750</td>
<td>25</td>
<td>NA</td>
</tr>
</tbody>
</table>

Some of the highly-ranked teams look about right (Louisville, Indiana, Gonzaga), but some, like Weber State, who went 26-6 in the Big Sky conference and are ranked 7th, appear to be rather overrated by this metric.

Looking at the page for Weber State’s 2012-2013 season, it appears that they benefited from an overall easy schedule with no ranked opponents, including one 65-point victory against the Southwest Minnesota State Mustang.

In the next section, we will see how to modify this method to account for strength of schedule.

### 3. A Linear Regression Model for Ranking Teams

Statistical modeling is a powerful tool for learning from data. Our meta-strategy is to define some statistical model whose parameters correspond to whatever quantities we want to estimate.

Our response variable for today is the margin of victory (or defeat) for the home team in a particular game. That is, define

\[ y_i = (\text{home score} - \text{away score}) \text{ in game } i \] (1)
Now, we want to define a linear regression model that explains the response, \( y_i \), in terms of both teams’ merits. The simplest such model will look something like

\[
y_i = \text{quality of home}(i) - \text{quality of away}(i) + \text{noise}
\]  

where home\((i)\) and away\((i)\) are the home and away teams for game \( i \).

To formulate this model as a linear regression in standard form, we need to come up with a definition for the predictors \( x \) and coefficients \( \beta \) so that estimating \( \beta \) amounts to estimating the quality of each team. That is, we want a definition for \( x_{ij} \) and \( \beta_j \) for which

\[
y_i = \sum_j x_{ij} \beta_j + \varepsilon_i 
\]

Now, with your group, try to formulate our model as a linear regression. How many predictor variables are there? How many coefficients? How is \( x_{ij} \) defined?

**[Bonus Questions]** If you have time, consider the following questions:

- In our new model, suppose that a small high school plays only one game, against Weber State, and loses by a very wide margin. Can you prove that this game will have no effect whatsoever on the fitted values of \( \beta \)?
- (Challenging) Our model now controls for the quality of a team’s opponents. As a result, I claim that scheduling many difficult opponents (or many easy opponents) will neither hurt nor help a team’s ranking. Can you back up this claim in a precise mathematical way?

**Answers:**

Define one predictor variable for each team, which is a sort of “signed dummy variable.” For game \( i \) and team \( j \), let

\[
x_{ij} = \begin{cases}  
+1 & \text{if } j \text{ is home}(i) \\
-1 & \text{if } j \text{ is away}(i) \\
0 & \text{if } j \text{ didn’t play}
\end{cases}
\]

For example, if game \( i \) consists of team 1 visiting team 2, then \( x_i = (-1, 1, 0, 0, \ldots, 0) \).

Now we can check that

\[
\sum_j x_{ij} \beta_j = \beta_{\text{home}(i)} - \beta_{\text{away}(i)}
\]

so the coefficient \( \beta_j \) corresponds exactly to the quality of team \( j \) in our model.

We can generate \( y \) and \( X \) as follows:

```r
## Construct a data frame of the right dimensions, with all zeros
X0 <- as.data.frame(matrix(0, nrow(games), length(all.teams)))
names(X0) <- all.teams

## Fill in the columns, one by one
for(tm in all.teams) {
  X0[[tm]] <- 1*(games$home==tm) - 1*(games$away==tm)
}
```

```r
ey <- with(games, homeScore-awayScore)
```

Now, we are in good shape, because we can use all our tools from linear regression to estimate \( \beta \), carry out hypothesis tests, construct confidence intervals, etc.

4. An Identifiability Problem

When we fit our model, we will ask R to find the best-fitting \( \beta \) vector. There is a small problem, however: for any candidate value of \( \beta \), there are infinitely many other values \( \tilde{\beta} \) that make exactly the same predictions. So the “best \( \beta \)” is not uniquely defined.

For any constant \( c \), suppose that I redefine \( \tilde{\beta}_j = \beta_j + c \). Then for every game \( i \),

\[
\tilde{\beta}_{\text{home}(i)} - \tilde{\beta}_{\text{away}(i)} = \beta_{\text{home}(i)} - \beta_{\text{away}(i)}
\]

so the distribution of \( y \) is identical for parameters \( \tilde{\beta} \) and \( \beta \), no matter what \( c \) is. We can never distinguish these two models from each other, because the models make identical predictions no matter what. In statistical lingo, this is called an \textit{identifiability} problem. It very often arises with dummy variables.

To fix it, we need to add some linear constraint on \( \beta \) to resolve the ambiguity. For example \( \sum_j \beta_j = 0 \), or we can pick some “special” baseline team \( j \) and require that \( \beta_j = 0 \). We will use the latter strategy, and we will take Stanford’s team as the baseline. Now, with your team, figure out how to modify the \( X \) matrix to implement this.

(Actually, \texttt{lm} is smart enough to fix this automatically for you by arbitrarily picking one team to be the baseline. But let’s not blindly rely on that)

\[\textbf{Bonus Questions}\] If you have time, consider the following questions:

- Suppose that we had chosen a different team as our baseline.
  - How would the estimates be different?
  - Would we obtain identical rankings?
  - Would we obtain identical standard errors?
- Under what circumstances would we still have an identifiability problem even after constraining \( \beta_{\text{Stanford}} = 0 \)?

\textbf{Answers}:

We can effectively force \( \beta_j = 0 \) for the \( j \) corresponding to Stanford by eliminating that column from the predictor matrix.

\[
X \leftarrow X[, names(X) != "stanford-cardinal"]
\]

\[
\text{reg.season.games} \leftarrow \text{which}(\text{games$gameType=="REG"})
\]

Now, let’s fit our model. There is no intercept in the model, so we explicitly exclude it from the formula.

\[
\text{mod} \leftarrow \text{lm}(y \sim 0 + ., \text{data}=X, \text{subset=reg.season.games})
\]

\[
\text{head(coef(summary(mod)))}
\]

| # | Estimate | Std. Error | t value | Pr(>|t|) |
|---|----------|------------|---------|----------|
| `air-force-falcons` | -6.687934 | 2.897768 | -2.307960 | 2.104205e-02 |
| `akron-zips` | -2.557678 | 2.841221 | -0.900203 | 3.680552e-01 |
| `alabama-a&m-bulldogs` | -30.590013 | 2.908702 | -10.516721 | 1.338001e-25 |
| `alabama-crimson-tide` | -2.851010 | 2.802443 | -1.017330 | 3.090456e-01 |
| `alabama-state-hornets` | -29.958427 | 2.849296 | -10.514327 | 1.371671e-25 |
| `albany-great-danes` | -13.334555 | 2.813834 | -4.731345 | 2.291935e-06 |
It looks like our model explains about half of the variability in basketball scores.

5. Interpreting the Model

Next, let’s try to interpret the model that we just fit. With your group, take a few moments to answer the following questions:

5.1. Based on this model, what would be a reasonable point spread if Alabama (alabama-crimson-tide) played against Air Force?

5.2. Can we be confident that Stanford is better than Alabama? Better than Air Force?

5.3. How can we test whether Alabama is better than Air Force?

5.4. Does the dataset and model support the notion of home field advantage? How many points per game is it? Is it statistically significant?

Answers:

5.1. If Alabama were the home team, then the expected score difference is

$$\hat{\beta}_{\text{Alabama}} - \hat{\beta}_{\text{Air Force}}$$

(7)

So we can answer the question by comparing their coefficients:

```r
coef(mod)["\text{alabama-crimson-tide}\"] - coef(mod)["\text{air-force-falcons}\"]
```

## alabama-crimson-tide
## 3.836924

so a fair point spread would be about 3.84 in favor of Alabama.

5.2. The hypothesis that Stanford is better than Alabama is equivalent to the hypothesis that $\beta_{\text{Alabama}} < 0$. If we use a two-sided test, our t statistic needs to be bigger than 2 in absolute value. It is not for Alabama, but it is for Air Force; hence the answers are “no” and “yes.”

5.3. We cannot answer this question based only on the coefficient table printed above. To test the null hypothesis that

$$\theta = \beta_{\text{Alabama}} - \beta_{\text{Air Force}} = 0$$

(8)

we would need to know the standard error of

$$\hat{\theta} = \hat{\beta}_{\text{Alabama}} - \hat{\beta}_{\text{Air Force}}$$

(9)

The easiest way to find this out would be to simply re-fit the model with Alabama (or Air Force) as the baseline.

5.4 We can assess home field advantage by including another column, which is 1 if and only if that game was played in a non-neutral location:
homeAdv <- 1 - games$neutralLocation
Xh <- cbind(homeAdv=homeAdv, X)
homeAdv.mod <- lm(y ~ 0 + ., data=Xh, subset=reg.season.games)
head(coef(summary(homeAdv.mod)), 1)

## Estimate Std. Error t value Pr(>|t|)
## homeAdv 3.528499 0.1568782 22.49197 8.654579e-107

It looks like home field advantage is worth about 3.5 points per match and is clearly significant.

### 6. Predicting Wins and Losses

If we wanted to run a bookie business, we would not only have to set point spreads, but also set odds in advance of each game. In this section we will see that our model can give us odds as well as point spreads.

We made an assumption that the errors were normal, so we should also check whether the residuals are normal. Let’s check out the residuals from the last model (with home-court advantage) and see if they look reasonably normal.

homeAdv.coef <- coef(homeAdv.mod)[paste("",teams$team,"",sep="")] names(homeAdv.coef) <- teams$team
resids <- homeAdv.mod$resid
par(mfrow=c(1,2))
hist(resids,freq=FALSE)
curve(dnorm(x,mean(resids),sd(resids)),add=TRUE,col="blue",lwd=2) qqnorm(resids)

Since our normal-errors assumption passes this basic sanity check, let’s take our model seriously. Assuming the errors are truly normal, how can we use our model to predict the win/loss outcome of a particular game?
If you have time, consider the following questions:

- **Wichita State** (wichita-state-shockers to us) made an improbable run in the NCAA tournament, beating a #1 seed (Gonzaga) and a #2 seed (Ohio State), and eventually losing to another #1 seed (Louisville) in the Final Four. According to the model we just fit, just how improbable was Wichita State’s run? That is, what was the probability they would have beaten the first four opponents that they faced?

- The actual tournament winner was louisville-cardinals. Given the six opponents they faced, what was the probability they would win the tournament?

You can use this function to get data in a slightly more manageable form:

```r
schedule <- function(team, game.type) {
  home.sch <- with(games, games[home==team & gameType==game.type,c(1,3,4,5)])
  away.sch <- with(games, games[away==team & gameType==game.type,c(1,2,5,4)])
  names(home.sch) <- names(away.sch) <- c("date","opponent","score","oppoScore")
  sch <- rbind(home.sch,away.sch)
  sch$margin <- with(sch, score-oppoScore)
  sch$oppoQuality <- homeAdv.coef[as.character(sch$opponent)]
  sch <- sch[order(sch$date),]
  rownames(sch) <- NULL
  return(sch)
}

schedule("wichita-state-shockers","NCAA")
```

<table>
<thead>
<tr>
<th>date</th>
<th>opponent</th>
<th>score</th>
<th>oppoScore</th>
<th>margin</th>
<th>oppoQuality</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013-03-21</td>
<td>pittsburgh-panthers</td>
<td>73</td>
<td>55</td>
<td>18</td>
<td>7.3562577</td>
</tr>
<tr>
<td>2013-03-23</td>
<td>gonzaga-bulldogs</td>
<td>76</td>
<td>70</td>
<td>6</td>
<td>9.9875221</td>
</tr>
<tr>
<td>2013-03-28</td>
<td>la-salle-explorers</td>
<td>72</td>
<td>58</td>
<td>14</td>
<td>-0.7725049</td>
</tr>
<tr>
<td>2013-03-30</td>
<td>ohio-state-buckeyes</td>
<td>70</td>
<td>66</td>
<td>4</td>
<td>7.8865499</td>
</tr>
<tr>
<td>2013-04-06</td>
<td>louisville-cardinals</td>
<td>68</td>
<td>72</td>
<td>-4</td>
<td>12.0810415</td>
</tr>
</tbody>
</table>

Answers:

Our model says that $y_i \sim N(\mu_i, \sigma^2)$, where $\sigma^2$ is the noise variance and $\mu_i$ is the home-away quality difference (possibly adjusted for home-court advantage).

To calculate the probability the home team loses ($y_i < 0$), we can exploit the fact that $(y_i - \mu_i)/\sigma \sim N(0,1)$. If $\Phi$ denotes the cumulative distribution function of a $N(0,1)$ variable, then:

$$P(y_i < 0) = P\left(\frac{y_i - \mu_i}{\sigma} < -\mu_i/\sigma\right)$$

$$= \Phi(-\mu_i/\sigma)$$

We can estimate $\sigma$ by calculating the residual variance (technically we should make an adjustment for degrees of freedom but let’s ignore that).

```r
(sigma <- sd(resids))
```

```
## [1] 9.821427
```
So, for example, we can compute the probability that Stanford would beat Berkeley (a slightly worse team, according to our model), at Berkeley:

```r
mu <- coef(homeAdv.mod)["homeAdv"] + coef(homeAdv.mod)["california-golden-bears"] - 0
pnorm(-mu/sigma)
```

```
## homeAdv
## 0.4204275
```

According to our model, that game would be roughly a toss-up.

Next, let’s find out how improbable Wichita State’s run was. Note that all the tournament games were played at neutral locations.

```r
wst.schedule <- schedule("wichita-state-shockers","NCAA")
mu <- wst.schedule$oppoQuality - homeAdv.coef["wichita-state-shockers"]
names(mu) <- wst.schedule$opponent
(p.wst.win <- pnorm(-mu/sigma))
```

```
## pittsburgh-panthers gonzaga-bulldogs la-salle-explorers
## 0.2915299 0.2070124 0.6097760
## ohio-state-buckeyes louisville-cardinals
## 0.2729443 0.1515074
```

So the probability that Wichita State would have won its first four games (conditional on the opponents it faced) is about 1%.

```r
prod(p.wst.win[1:4])
```

```
## [1] 0.0100444
```

By contrast, the probability of Louisville winning the whole tournament (conditional on their opponents) was about 25%:

```r
lou.schedule <- schedule("louisville-cardinals","NCAA")
mu <- lou.schedule$oppoQuality - homeAdv.coef["louisville-cardinals"]
prod(pnorm(-mu/sigma))
```

```
## [1] 0.2509396
```

### 7. Model Rankings vs. Official Rankings

Finally, let’s inspect the rankings given by our model.

```r
rank.table <- cbind("Model Score" = homeAdv.coef,
   "Model Rank" = rank(-homeAdv.coef,ties="min"),
   "AP Rank" = teams$apRank,
   "USAT Rank" = teams$usaTodayRank)
rank.table[order(homeAdv.coef,decreasing=TRUE)[1:25],]
```
## Model Score Model Rank AP Rank USAT Rank

<table>
<thead>
<tr>
<th>Team</th>
<th>Model Score</th>
<th>Model Rank</th>
<th>AP Rank</th>
<th>USAT Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>indiana-hoosiers</td>
<td>13.255228</td>
<td>1</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>florida-gators</td>
<td>12.638550</td>
<td>2</td>
<td>14</td>
<td>12</td>
</tr>
<tr>
<td>louisville-cardinals</td>
<td>12.081041</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>gonzaga-bulldogs</td>
<td>9.987522</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>duke-blue-devils</td>
<td>9.393678</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>kansas-jayhawks</td>
<td>8.631876</td>
<td>6</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>ohio-state-buckeyes</td>
<td>7.896550</td>
<td>7</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>michigan-wolverines</td>
<td>7.505373</td>
<td>8</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>pittsburgh-panthers</td>
<td>7.356258</td>
<td>9</td>
<td>20</td>
<td>22</td>
</tr>
<tr>
<td>syracuse-orange</td>
<td>7.069567</td>
<td>10</td>
<td>16</td>
<td>18</td>
</tr>
<tr>
<td>wisconsin-badgers</td>
<td>6.831338</td>
<td>11</td>
<td>18</td>
<td>17</td>
</tr>
<tr>
<td>michigan-state-spartans</td>
<td>6.176226</td>
<td>12</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>creighton-bluejays</td>
<td>5.283583</td>
<td>13</td>
<td>22</td>
<td>21</td>
</tr>
<tr>
<td>miami-(fl)-hurricanes</td>
<td>5.191585</td>
<td>14</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>virginia-commonwealth-rams</td>
<td>4.957174</td>
<td>15</td>
<td>NA</td>
<td>23</td>
</tr>
<tr>
<td>arizona-wildcats</td>
<td>4.684883</td>
<td>16</td>
<td>21</td>
<td>20</td>
</tr>
<tr>
<td>minnesota-golden-gophers</td>
<td>4.421583</td>
<td>17</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>georgetown-hoyas</td>
<td>4.258965</td>
<td>18</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>missouri-tigers</td>
<td>3.962353</td>
<td>19</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>oklahoma-state-cowboys</td>
<td>3.690893</td>
<td>20</td>
<td>17</td>
<td>19</td>
</tr>
<tr>
<td>saint-mary's-gaels</td>
<td>3.372377</td>
<td>21</td>
<td>NA</td>
<td>25</td>
</tr>
<tr>
<td>colorado-state-rams</td>
<td>3.152205</td>
<td>22</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>new-mexico-lobos</td>
<td>3.004408</td>
<td>23</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>north-carolina-tar-heels</td>
<td>2.715207</td>
<td>24</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>ole-miss-rebels</td>
<td>2.644823</td>
<td>25</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

Our rankings still differ significantly from the official rankings on several teams. To take one particularly glaring example, our model is highly confident that **florida-gators** is an elite team, despite its relatively low ranking by the press. By contrast, our model doesn’t think too much of **miami-(fl)-hurricanes** despite the opinion of the press that the 'Canes were elite.

See if you can figure out why our model might beg to differ with the press.

**Answers:**

From the schedules above, we see that Florida's 7 losses are all close, with only one loss by more than 6 points, whereas all of their 26 wins are by at least 10 points.

Miami has a better overall win-loss record (27-6) against a somewhat stronger group of opponents, and finished the season strong by winning the ACC tournament. However, 10 of their wins were by less than 10 points, and they lost three games by wide margins.

Below, we plot each team’s margin of victory distribution for its 33 regular season games.

```r
hist(schedule("miami-(fl)-hurricanes", "REG")$margin, col="gray",border="gray",breaks=seq(-20,50,5), xlab="Margin of Victory", main="Margins of Victory for Miami (gray) and Florida (red)",)
hist(schedule("florida-gators", "REG")$margin,add=TRUE, border="red",breaks=seq(-20,50,5),col="red",density=20)
abline(v=0)
```
Miami did have slightly better opponents, but not by much

\[\text{mean(schedule("miami-(fl)-hurricanes", "REG")$oppoQuality)}\]

\[\text{mean(schedule("florida-gators", "REG")$oppoQuality)}\]

Notice that in our model, going from a -1 margin of victory to a +1 margin has the same impact on a team’s rating as going from +6 to +8, or from +40 to +42. So, we might be giving Florida a little too much credit for running up the score against its opponents.

Conversely, sports journalists typically care a lot about whether a team wins or loses, because they believe that scoring a buzzer-beater to win a game proves that a team possesses the “heart of a champion.” This elusive quality is thought to outweigh the more prosaic ability to score a lot of points.

8. The Value of Nerlens Noel

Nerlens Noel was a star center for the University of Kentucky (kentucky-wildcats in our data set). He was projected to be the #1 overall pick in the 2013 NBA draft before he tore his ACL in a February 12 game (and was eventually the #6 overall pick despite his injury).

Statistical analysts employed by sports teams are especially interested in evaluating individual players. There are various methods, usually requiring granular minute-by-minute data on scoring and substitutions. We do not have such granular data at our disposal, but for an injured player like Noel, we can estimate how much better Kentucky was with him than without him.
Re-fit your model with one additional predictor variable, to get an estimate and confidence interval for Noel’s contribution to Kentucky (i.e., how much better Kentucky is with him than without him).

**Bonus Questions** If you have time, consider the following questions:

- Give an estimate and confidence interval for Kentucky’s quality with Noel, and for its quality without Noel. Where would Kentucky be ranked if Noel had not been injured?
- Suppose that instead of the games data, we had a data point for every one-minute long period during the course of every game. During each minute, suppose we also had a record of which five players were playing for each team, as well as the number of points scored by each team during that minute (for simplicity, assume substitutions only happen in between the one-minute periods). Can you come up with a model, similar to the one we fit today, that we could use to estimate each player’s quality?

**Answers:**

Let \( n_i \) be +1 if Noel played for the home team (i.e., all UK home games before February 12), -1 if he played for the away team (UK away games before Feb 12), and 0 if he did not play. Then we can simply fit the model:

\[
y_i = \beta_{-1} n_i + \beta_0 h_i + \beta_{\text{home}(i)} - \beta_{\text{away}(i)} + \varepsilon_i
\]  

(12)

```r
nerlens <- rep(0, nrow(games))
nerlens[games$home=="kentucky-wildcats" & games$date<"2013-02-12"] <- 1
nerlens[games$away=="kentucky-wildcats" & games$date<"2013-02-12"] <- -1
nerlens.mod <- lm(y ~ 0 + nerlens + ., data=Xh, subset=reg.season.games)
head(coef(summary(nerlens.mod)))
```

|             | Estimate | Std. Error | t value | Pr(>|t|)  |
|-------------|----------|------------|---------|-----------|
| nerlens     | 11.4503609 | 4.1033060  | 2.7905208 | 5.282328e-03 |
| homeAdv     | 3.5264822  | 0.1567733  | 22.4941548 | 8.294412e-107 |
| `air-force-falcons` | -5.3069737  | 2.7602474  | -1.9226442 | 5.458163e-02 |
| `akron-zips` | -0.9369568  | 2.7066503  | -0.3461684 | 7.292308e-01 |
| `alabama-a&m-bulldogs` | -27.1891732 | 2.7742456 | -9.8005644 | 1.784244e-22 |
| `alabama-crimson-tide` | -2.6585470  | 2.6688528  | -0.9961385 | 3.192312e-01 |

```r
nerlens.estimate <- coef(nerlens.mod)["nerlens"]
nerlens.se <- coef(summary(nerlens.mod))[,2]
(nerlens.CI <- c("CI lower"=nerlens.estimate - 2*nerlens.se, 
                 "CI upper"=nerlens.estimate + 2*nerlens.se))
```

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CI lower.nerlens</td>
<td>CI upper.nerlens</td>
<td></td>
</tr>
<tr>
<td>3.243749</td>
<td>19.656973</td>
<td></td>
</tr>
</tbody>
</table>

So Noel was worth about 11 points to Kentucky, but we cannot place too much confidence in this figure.
If we have minute-level data, then we can assign a coefficient $\beta_j$ to each player. Then, letting $y_i$ be the home-away scoring margin for minute $i$, we can set

$$X_{ij} = \begin{cases} 
+1 & j \text{ played for the home team in minute } i \\
-1 & j \text{ played for the away team in minute } i \\
0 & j \text{ didn’t play in minute } i 
\end{cases}$$

This is known as the adjusted plus/minus model, and it is one of the tools used by statisticians in the NBA to evaluate players.

9. A Non-Regression Based Ranking Approach

There are many other ways to rank teams; we'll see one more that can be cast as a random walk on a graph. We setup the problem as follows. Place $M$ counters initially on each team. We'll then complete 100 rounds of the following procedure:

1. Randomly permute all the teams
2. For each team $X$, grab all their counters and then for each counter:
   - Select a random game that $X$ played in and then flip a coin that comes up heads with probability $p$. If the coin is heads, put the counter on the winning team (otherwise move it to the losing team).

After many rounds have been completed, the teams with the most counters are deemed the highest ranked. Let’s see an example to make things clear. Suppose that we only had three teams–A, B and C–with 2, 0, and 1 counters respectively. Let’s also say there were three games played, A beat B, A beat C, and B beat C. Then if we were to update the counters in one round, we’d start with A. Since A has two counters, we’d grab two random games; suppose they were both the game with B. Since A won that game, we’d flip two coins that come up heads with probability $p$. Say one came up heads and the other tails; this means we’d put one counter with A and one with B for the next round.

Since B doesn’t have any counters in this round, we don’t have any work. Finally, as C has one counter, we’d randomly choose a game they played; let’s pick the game with B. Since B won the game, we’d move to counter to B with probability $p$. Let’s say we do; then the counters for the next round would be A with 1, B with 2, and C with 0.

Try coding this up, using $p = 0.8$, $M = 100$, and 50 rounds cycling through all the teams. Then see if you can compare these to the previously used rankings.

**Answers:**

First, let’s set the parameters and write a helper function to do most of the work (i.e. Step 2 above). Given a current set of counters, we’ll write a function that takes a team and number of counters they currently have, and then redistributes them according to the games they played and coin flips.

```r
p <- 0.8
M <- 100
counters <- rep(M, length(all.teams))
names(counters) <- all.teams

move.counters <- function(team, num.counters) {
  team.games <- subset(games, home == team | away == team)
  num.games <- nrow(team.games)
  # Implement the logic to redistribute counters
  # based on the games and coin flips
  # Example: Update counters for team
  # Final step: Update counters for the next round
}
```
```r
games.to.move.idx <- sample(1:num.games, num.counters, replace=TRUE)
games.to.move <- team.games[games.to.move.idx, ]
# now get the winners
home.won <- with(games.to.move, homeScore - awayScore) > 0
coin.flips <- sample(c(0,1), num.counters, replace=TRUE, prob=c(p, 1-p))
# the coin flips only change the winner if 1
move.to.home <- (home.won + coin.flips) %% 2

next.counters <- c(
  games.to.move[move.to.home == 1, 'home'],
  games.to.move[move.to.home == 0, 'away']
)
table(next.counters)
}

Now, let’s write the main for loop that moves the counters. We’ll check the results to make sure it looks sane.

```r
set.seed(7)
for (round in 1:50) {
  # initialize the new counters after this round is complete
  new.counters <- rep(0, length(all.teams))
names(new.counters) <- all.teams
  perm.teams <- all.teams[sample(length(all.teams))]
  for (team in perm.teams) {
    num.counters <- counters[team]
    moved.counters <- moveCounters(team, num.counters)
    # move the counters!
    new.counters[names(moved.counters)] <- (new.counters[names(moved.counters)] + moved.counters)
  }

  counters <- new.counters
}

counters <- sort(counters, decreasing=TRUE)
head(as.data.frame(counters),10)

```r

## counters
## louisville-cardinals  632
## duke-blue-devils  548
## michigan-wolverines  477
## syracuse-orange  461
## gonzaga-bulldogs  447
## kansas-jayhawks  438
## ohio-state-buckeyes  437
## indiana-hoosiers  436
## miami-(fl)-hurricanes  434
## michigan-state-spartans  391

Cool, things look good. Let’s finally compare them to previous ranks we used.
```r
# now put in the rank table
counter.rank <- rank(-counters, ties="min")
rank.table.idx <- sapply(row.names(rank.table), function (team) {
  which(team == names(counter.rank))
})
rank.table.idx <- unlist(rank.table.idx)
rank.table <- cbind(rank.table, 
  Counter Rank = counter.rank[rank.table.idx])
head(rank.table[order(rank.table[, 'AP Rank']), ], 25)
```

<table>
<thead>
<tr>
<th>Model Score</th>
<th>Model Rank</th>
<th>AP Rank</th>
<th>USAT Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.9875221</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>12.0810415</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>8.6318763</td>
<td>6</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>13.2552278</td>
<td>1</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5.1915853</td>
<td>14</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>9.3936777</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>7.8965499</td>
<td>7</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>4.2589645</td>
<td>18</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>6.1762258</td>
<td>12</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>7.5053732</td>
<td>8</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>3.0044080</td>
<td>23</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>1.6525402</td>
<td>36</td>
<td>12</td>
<td>14</td>
</tr>
<tr>
<td>2.6372882</td>
<td>26</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>12.6385504</td>
<td>2</td>
<td>14</td>
<td>12</td>
</tr>
<tr>
<td>2.2985477</td>
<td>31</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td>7.0695669</td>
<td>10</td>
<td>16</td>
<td>18</td>
</tr>
<tr>
<td>3.6908930</td>
<td>20</td>
<td>17</td>
<td>19</td>
</tr>
<tr>
<td>6.8313376</td>
<td>11</td>
<td>18</td>
<td>17</td>
</tr>
<tr>
<td>1.0143519</td>
<td>42</td>
<td>19</td>
<td>15</td>
</tr>
<tr>
<td>7.3562577</td>
<td>9</td>
<td>20</td>
<td>22</td>
</tr>
<tr>
<td>4.6848829</td>
<td>16</td>
<td>21</td>
<td>20</td>
</tr>
<tr>
<td>5.2835829</td>
<td>13</td>
<td>22</td>
<td>21</td>
</tr>
<tr>
<td>1.9049621</td>
<td>35</td>
<td>23</td>
<td>NA</td>
</tr>
<tr>
<td>1.0059603</td>
<td>43</td>
<td>24</td>
<td>NA</td>
</tr>
<tr>
<td>0.6666812</td>
<td>45</td>
<td>25</td>
<td>24</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model Score</th>
<th>Model Rank</th>
<th>AP Rank</th>
<th>USAT Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>gonzaga-bulldogs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>louisville-cardinals</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>kansas-jayhawks</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>indiana-hoosiers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>miami-(fl)-hurricanes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>duke-blue-devils</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>ohio-state-buckeyes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>georgetown-hoyas</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>michigan-state-spartans</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>michigan-wolverines</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>new-mexico-lobos</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>kansas-state-wildcats</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>saint-louis-billikens</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>florida-gators</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>marquette-golden-eagles</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>syracuse-orange</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Counter Rank
## gonzaga-bulldogs 5
## louisville-cardinals 1
## kansas-jayhawks 6
## indiana-hoosiers 8
## miami-(fl)-hurricanes 9
## duke-blue-devils 2
## ohio-state-buckeyes 7
## georgetown-hoyas 13
## michigan-state-spartans 10
## michigan-wolverines 3
## new-mexico-lobos 11
## kansas-state-wildcats 25
## saint-louis-billikens 17
## florida-gators 15
## marquette-golden-eagles 14
## syracuse-orange 4
We see that these rankings look relatively closer to the AP Rank than the margin based rankings. In fact, only the Oklahoma State Cowboys, the Pittsburgh Panthers, and the UCLA Bruins show up in the AP top 25 but are not the counter-based rankings.