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 the games and teams data into data frames games and teams

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

# Load the 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
# Display the first few rows of the games data
head(games)[, !(names(games) %in% c('neutralLocation', 'gameType'))]
```

`games` data contains:
- `date` column for the date of the game.
- `home` column for the name of the home team.
- `away` column for the name of the away team.
- `homeScore` column for the score of the home team.
- `awayScore` column for the score of the away team.

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
# Display the first few rows of the teams data
head(teams)
```
Finally, we make one vector containing all the team names, because the three columns do not perfectly overlap:

```
all.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?

### 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

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

(2)

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

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

(3)

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?

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 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, `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)

[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 $\beta_{\text{Stanford}} = 0$?

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?
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.

```r
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?

[Bonus Questions] 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)
}
```

```r
table <- schedule("wichita-state-shockers","NCAA")
```

<table>
<thead>
<tr>
<th></th>
<th>date</th>
<th>opponent</th>
<th>score</th>
<th>oppoScore</th>
<th>margin</th>
<th>oppoQuality</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<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>2</td>
<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>3</td>
<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>4</td>
<td>2013-03-30</td>
<td>ohio-state-buckeyes</td>
<td>70</td>
<td>66</td>
<td>4</td>
<td>7.8965499</td>
</tr>
<tr>
<td>5</td>
<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>

7. Model Rankings vs. Official Rankings

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

```r
cbind("Model Score" = homeAdv.coef,
       "Model Rank" = rank(-homeAdv.coef,ties="min"),
       "AP Rank" = teams$apRank,
       "USAT Rank" = teams$usaTodayRank)
```
Our rankings still differ significantly from the official rankings on several teams. To take one particularly glaring example, our model is highly confident that \textit{florida-gators} is an elite team, despite its relatively low ranking by the press. By contrast, our model doesn’t think too much of \textit{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.

8. The Value of Nerlens Noel

\textit{Nerlens Noel} was a star center for the University of Kentucky (\textit{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 \textit{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 \textit{player’s} quality?

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.