Homework 10: Baseball Salaries

36-350, Fall 2011

Due at 11:59 pm on Friday, 9 December 2011

You know the rules by now for turning in work.

In this homework, you will look at the relationship between baseball team payrolls and performance between the years 1985 and 2010. The data come from the Baseball Databank `http://baseball-databank.org` and is based in part on Lahman's Baseball Database. Information on the attributes in the database can be found at `http://baseball1.com/files/database/readme58.txt`. You will need to download the SQLite database file `baseball.db` (located at `http://www.stat.cmu.edu/~cshalizi/statcomp/hw/10/baseball.db`) to your computer.

You will also need the R packages `DBI`, `RSQLite`, `fImport`, `plyr`, and `ggplot2` for this homework. There is an R package (`lahman`) containing data frames with all of this data. It is strictly off-limits for this homework.

1. Import payroll data from the database.

   (a) (5) Using DBI and RSQLite, setup a connection to the SQLite database stored in `baseball.db`. Use `dbListTables()` to list the tables in the database.

   (b) (10) Use the table that contains salaries and compute the payroll for each team in 2010. Do this using only `dbGetQuery()` and SQL. Which teams had the highest payrolls?

   (c) (10) Modify the SQL statement to compute the payroll for each team for each year from 1985 to 2010.

2. Use graphical techniques to study the change in payrolls over time and the relationship between payroll and performance. You will need to control for inflation (changing price levels). Use the following code snippet to get price levels (CPI, consumer price index) from the Federal Reserve Economic Data (FRED).

```
library(fImport)
cpi <- fredSeries('CPIAUCSL',
                  from = as.Date('1985-01-01'),
                  to = as.Date('2011-01-01'))
cpi <- cpi[months(as.Date(rownames(cpi))) == 'January']
cpi <- cpi / cpi[length(cpi)]
```

This produces a vector `cpi` containing consumer price indices from 1985 to 2011, normalized so that 1 = $1 in 2011. To convert x dollars in 1990 into y dollars in 2011 you would use the command

```
y <- x / cpi[1990 - 1985 + 1]
```

(a) (25) Modify the following code snippet to create a plot showing the change of inflation adjusted payrolls over time. Hint: Use `plyr` and `transform`.

```
### IMPORT DATA HERE ###
payrolls <- ### AND ADJUST FOR INFLATION HERE ###
library(ggplot2)
qplot(data = payrolls, x = yearID, y = payroll_adjusted,
      group = teamID, color = teamID, geom = 'line')
```

Log transformations are often useful for looking at prices — money tends to multiply and logarithms make multiplicative relationships linear. So you should also try the following code snippet:

```
qplot(data = payrolls, x = yearID, y = payroll_adjusted,
      group = teamID, color = teamID, geom = 'line', log = 'y')
```

Have team salaries kept up with inflation, fallen behind, or grown faster? Have certain teams always had top payrolls over the years?

(b) (25) Augment your SQL statement above to include the following team statistics for each season: number of games played and number of games won. Use the following code snippet to create a plot showing a scatterplot of inflation adjusted payrolls and proportion of games won.

```
### IMPORT DATA HERE ###
df <- ### ADJUST FOR INFLATION ###
      ### AND COMPUTE PROPORTION OF GAMES WON HERE ###
qplot(data = df, x = payroll_adjusted, y = win_prop, log = 'x')
```

Does there appear to be a relationship between payroll and winning?

(c) (25) Explore the change in the relationship between payroll and winning over time. To do this, stratify the data into 12 groups according to 3 year intervals by using the `cut()` function to add a factor variable to the data frame `df` indicating which group each row of `df` belongs to, and then look at a separate scatter plot for each group.

```
df <- ### CODE HERE ###
qplot(data = df, x = payroll_adjusted, y = win_prop,
```

```
      log = 'x', facets = ~ period)
```

The `facets = ~ period` argument tells qplot() to make a separate plot for each value of `period`. You should see 12 scatter plots, each corresponding to a different 3 year period and there should be a label identifying the 3 year period at the top of each scatter plot. What do the scatter plots suggest?