Takehome final: The Indian Premier League

S520

Upload your typed answers as a PDF/Word/HTML document through the Assignments tab on Canvas by noon, Friday 17th December.

Instructions and warnings

- **You must not discuss this exam with anyone other than the instructor and the AIs until the due date has passed.**
- Write explanations for your answers. Answers alone will not get full credit. R code does not count as explanation.
- Write your answers in sentences. Do not just write your answering in R code comments.
- Round answers sensibly. Unrounded or inaccurately rounded answers may receive point deductions.
- Give both numerical results (e.g. P-values, confidence intervals) and substantive conclusions. For example:
  - "We reject the null hypothesis." — NOT MANY POINTS
  - "The P-value is 0.001. This means the data gives strong evidence that three-toed sloths have more toes than two-toed sloths." — LOTS OF POINTS

What may I ask the lecturer by email?

- If you think there is an error in the exam, notify the lecturer immediately.
- General questions about course material or help handling the data are better asked through the methods below.

What can I ask at office hours?

- General questions about course material.
- Help handling the data.

What can I ask on the course Q&A Community?

Same as office hours.

What can I ask other students?

Nothing.

About the Indian Premier League

The Indian Premier League (IPL) is a franchise tournament in Twenty20 cricket (a short form of the sport) that has been played every year since 2008. Since 2014 there have been eight teams each playing 14 games per regular season, though this has varied over time. (The plan for 2022 is to add two new teams, but to keep the number of games each team plays at 14.) From time to time the identity of the teams have changed. The teams that have played at least three seasons are:

- CSK: Chennai Super Kings (12 seasons)
- DC: Delhi Capitals (14 seasons)
- HDC: Hyderabad Deccan Chargers (5 seasons)
- KKR: Kolkata Knight Riders (14 seasons)
- MI: Mumbai Indians (14 seasons)
- PBKS: Punjab Kings (14 seasons)
- PWI: Pune Warriors India (3 seasons)
- RCB: Royal Challengers Bangalore (14 seasons)
- RR: Rajasthan Royals (12 seasons)
- SRH: Sunrisers Hyderabad (9 seasons)

A few other teams played one or two seasons.

I compiled data that includes the number of wins and points for every team for each season. We can use this data to try to answer questions like:

- Is the IPL just completely random?
- Do teams do better with New Zealand coaches?
- Are some teams systematically better than others?
- How good should we expect Sunrisers Hyderabad to be next year?

About the data

The data is on Canvas in two CSV files: `IPLpoints.csv` and `IPLpoints2.csv`. These two files are in somewhat different formats, but both of them can be read into R using `read.csv()`:

```
IPLpoints <- read.csv("IPLpoints.csv")
IPLpoints2 <- read.csv("IPLpoints2.csv")
```

In `IPLwins.csv`, each row represents one season for one team. Aside from the header, there are 16 rows ("team-seasons") in the file. It contains the following variables:

- `Year` gives the season (2008 to 2021.)
- `Team` gives the abbreviated name for the team.
- `NZCoach` is a binary variable that's 1 if the team's coach that season was a New Zealander and 0 otherwise.
- `Wins` gives the number of wins for that team that season (min 2, max 11.) Note: In these data sets, only regular season games are counted; playoff games are ignored.
- `Losses` gives the number of losses for that team that season (min 3, max 13.)
- `NoResult` gives the number of games with no result for that team that season (min 0, max 2.)
- `Points` gives the number of points the team earned that season. Teams earn 2 points for a win, 0 points for a loss, and 1 point for a game with no result.
- `AdjPoints` is equal to points *except* for the 2012 and 2013 seasons. In those years, teams played 16 games, so `AdjPoints` is `Points` $\times 14/16$ to put those seasons on the same scale as the others.

In `IPLwins2.txt`, each row represents a pair of consecutive seasons for one team. It contains the following variables:

- `Team` gives the abbreviated name for the team.
- `Year1` gives the first year of the pair.
- `AdjPoints1` gives the adjusted points in the first year.
- `Year2` gives the second year of the pair.
- `AdjPoints2` gives the adjusted points in the second year.

For questions 1, 2, and 3, use `AdjPoints1.txt`, treating all rows of the data as independent. For question 5, use `AdjPoints2.txt`.

Questions

1. (5 points.) An aging New Zealander who is used to cricket games that last five days says: "Is the IPL just completely random? If so, we could model each team's wins in a 14-game season as a Binomial random variable. Let's limit ourselves to the seasons where teams each played 14 games:

   ```
   IPL14 <- subset(IPLpoints, Year %in% c(2008:2011, 2014:2021))
   ```

   "That gives a data frame with 98 rows. In this data frame, the total of `Wins` is 676, while the total of `Wins`, `Losses`, and `NoResult` is 1372. So if we sample a game and randomly select one of the teams, the probability that team wins is $p = 676/1372 \approx 0.4927$. ($p$ is a bit less than 0.5 because there's some chance of no result.) So if the IPL is completely random, the distribution of `Wins` by team-season should be Binomial with $n = 14$ and $p = 676/1372$."

   We can test this null hypothesis model using a chi-squared test using the following categories for number of wins:

   0 to 4, 5, 6, 7, 8, 9, 10 to 14

   (a) **Under the Binomial null hypothesis, calculate the *expected* number of wins for each category (out of 98 team seasons) using `dbinom()` and/or `pbinom()`.** These expected values should add up to 98.

   (b) **Calculate and state the *observed* number of wins for each category.** For example, the observed number of teams that won 7 games is

   ```
   > sum(IPL14$Wins == 7)
   [1] 24
   ```

   (c) **Calculate a chi-squared statistic** (either Pearson's X^2 or the likelihood ratio G^2) **and get a P-value by comparing it to a chi-squared distribution with the appropriate number of degrees of freedom.** (Hint: The null hypothesis counts as a fully-specified probability model.)

   (d) The P-value you get should *not* be small, so we cannot reject the null hypothesis Binomial model. However, this does *not* prove that the IPL is completely random. **Explain why not to the aging New Zealander.**

2. (4 points.) Do teams do better with New Zealand coaches? The average value of `AdjPoints` is 14. If New Zealand coaches typically did the same as other coaches, we would expect their teams to average about 14 points a season.

   Your task: **Test the hypothesis that μ, the expected value of `AdjPoints` for team-seasons with New Zealand coaches, is 14. Say what kind of test you're doing and give a P-value and substantive conclusion. If you use a test that assumes normality, either justify the normal assumption (graphically or otherwise) or explain why normality is unnecessary. If you reject the null hypothesis, give a 95% confidence interval for μ. If you *don't* reject the null hypothesis, then guess whether this is because μ really is very close to 14, or if it's because your sample wasn't big enough.**

3. (3 points.) Are some teams consistently better than others? Let's limit the analysis to the ten teams that played at least three seasons.

```
IPLBig10 <- subset(IPLpoints, Team %in% c("CSK", "DC", "KKR", "HDC", "MI",
  "PBKS", "PWI", "RCB", "RR", "SRH"))
```

Perform an analysis of variance on the `IPLBig10` using `AdjPoints` as the numeric variable and `Team` as the response variable. What does this analysis of variance tell you? What precise hypothesis is being tested, and what does the ANOVA tell you about that hypothesis? What analysis would you want to do on the data to follow up on the ANOVA? Note: You don't actually have to do the follow-up analysis.

4. (2 points.) The analysis you did for all of the previous questions relies on an assumption that strictly speaking isn't true, or even that close to true. **What is this assumption?** (Hint: It's not normality.) **Why is it violated?**

5. (6 points.) Sunrisers Hyderabad (SRH) were pretty bad in the 2021 season, attaining only 6 points—remember that the average is 14. David predicts SRH will get 6 points again in the 2022 season, while Kane predicts SRH will get 14. We could try using linear regression to predict how well SRH will do next season. (Again the assumptions of simple linear regression might not be literally true, but in this case it might not matter as much.)

   (a) **Using the data set `IPLpoints2.csv`, find a regression line to predict `AdjPoints2` using `AdjPoints1`. Write down the equation for this line, and use it to make a prediction for the number of wins in 2022 for Sunrisers Hyderabad.** Is the prediction closer to 6 points or 14 points? **What does this tell you about the predictability of the IPL?**

   (b) **Check whether the assumptions of linear regression are met, supporting your answer with graphs.**

Strategic hint: Q1 might be the hardest technically, so feel free to move on to something else and come back to Q1 later if you get stuck.