Name: Chang Ni

UmichID: channi

What dataset are you working with: tennis\_events\_time, fifa\_audience, elo\_blatter

List 3 questions that you can ask with your dataset.

Q1:Is there any difference between seconds added per point for events on hard and clay court?

Q2: Is a country’s population related to its world cup TV audience share?

Q3: Is a country’s ELO in 2015 related to its ranking in 1998

List the associated null hypothesis for each question:

Q1: There is no difference between seconds added per point on hard court and clay court.

Q2: I don’t think building a linear regression model requires null hypothesis…If there have to be one, I guess it’s “The slope of the line is not different than 0 (A country’s share of world cup TV audience is not significantly related to its population share.)”  
Q3: A country’s ELO ranking in 2015 is not significantly related to the ELO ranking in 1998

What statistical test(s) will you use to answer each of the questions:

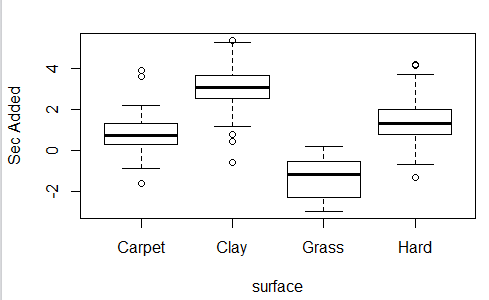
Q1: t-test

Q2: linear regression

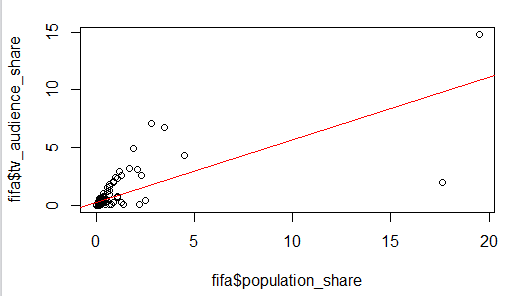
Q3: linear regression

Make a visual plot showing the relationship that you will analyze statistically (e.g. boxplot for t-test or ANOVA; scatterplot for regression; table for chi-square).

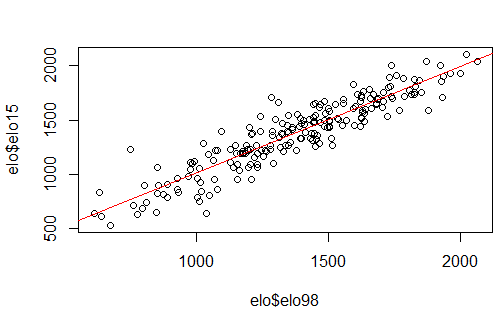
Q1: I plotted this first than decided to test the difference between clay and hard court. From my knowledge about tennis, grass is the fastest court. However, there tend to be a good deal of debate about the speed of hard and clay courts, as some of the slow hard courts seem to have speed similar to clay. Among top tournaments carpet is rare, and I’m not really interested…



Q2:



Q3:



Do your data meet the assumptions required for the statistical test you want to run? Please state the assumptions you examined and whether or not your data meet those assumptions:

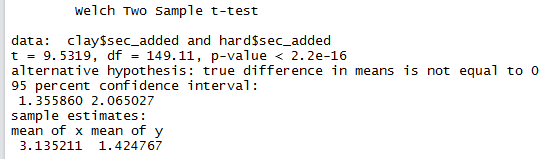
Q1: sample size>=30 (met), equal variance (met, var.test), independent observation (met)

Q2: Linear relationship (met), homoscedasticity (confirmed by plot), statistical independence of errors (met, plot of residuals), normality of error distribution (not met, confirmed by shapiro test of residuals, but I ran the test all the same…)

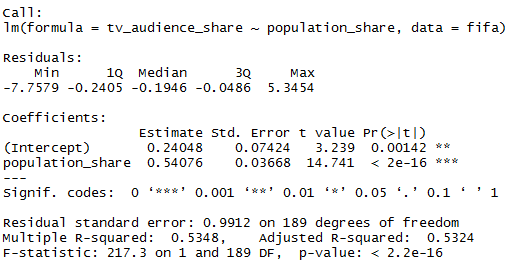
Q3: Linear relationship (met), homoscedasticity (confirmed by plot), statistical independence of errors (met, plot of residuals), normality of error distribution (met by plotting histogram of residuals)

Run the statistical test! Put your results here:

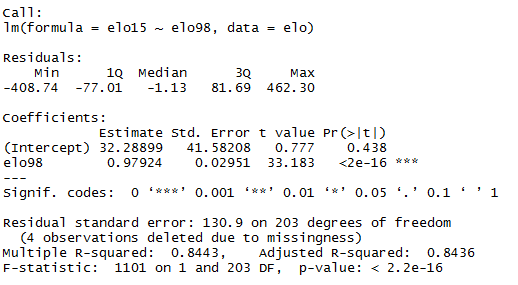
Q1:



Q2:



Q3:



Interpret your results!

Q1: p-value<0.05, so we reject the null hypothesis. The seconds added per point on hard court and clay court are significantly different, the former one significantly smaller than latter. Even with slow hard courts in account, hard courts seem significantly faster than clay.

Q2: The slope of the line is significantly greater than 0 (p<0.05), so a country’s world cup TV audience share is significantly and positively related to its population share. For each 1% increase in the total population share, the share of audience increase by 0.5%

Q3: The slope of the line is significantly greater than 0 (p<0.05), so a country’s ELO rating in 2015 is significantly and positively related to its ELO rating in 1998. For each point higher in 1998, the country score 0.98 points higher in 2015. The relative strength between countries does not seem to have changed much during the years.