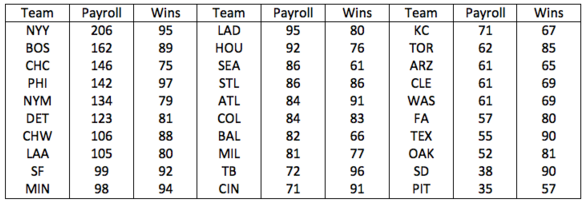
Unit 8 HW Solutions



Americans love their baseball. Even so, there is a concern that teams with more money to spend on players have more success than teams that have less money to spend. The payroll (X) (in millions of dollars) and the number of games won in the season (out of 162) (Y) are provided in the table below for the 30 major league teams. The numbers are from the 2010 regular season. We can use these data to illustrate statistical methods for drawing inferences about correlations.



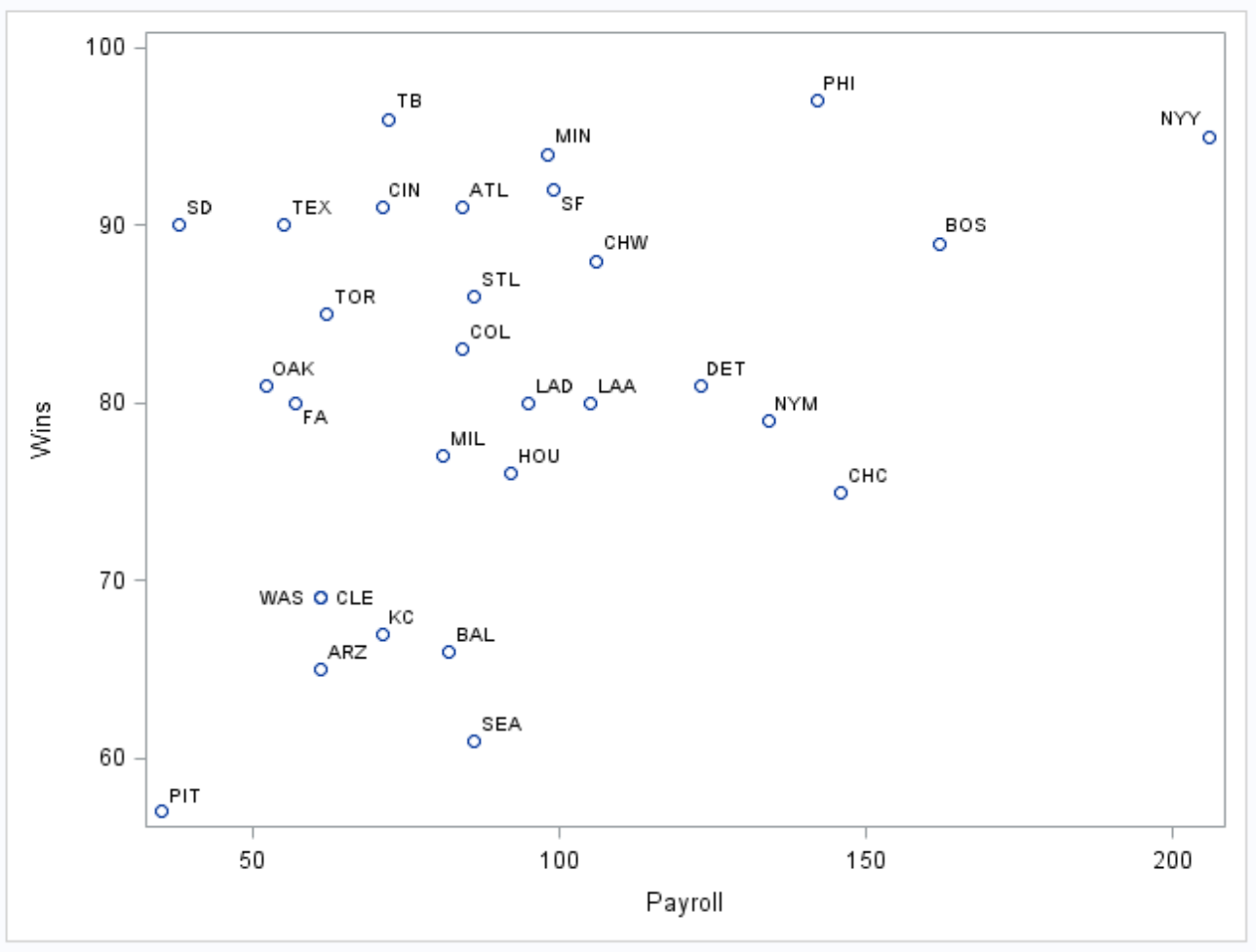
## Question 1 (30 points total)

Provide a scatterplot of the data using both SAS and R. Looking at the scatterplot, do you expect the correlation to be positive, negative, or close to 0? Why? Is the relationship between team payroll and number of wins strong, moderate or weak? Is the relationship linear? Take a guess of the value of the correlation coefficient.

*Note: there are multiple ways to do this in both SAS and R. One example of each is given below. As long as you have an equivalent representation, you should receive full points.*

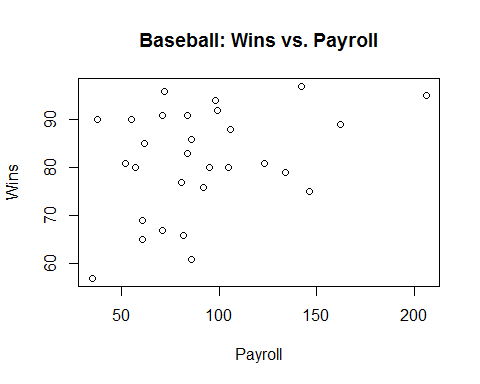
**SAS (10 points)**

\*Using sgplot (another potential option is sgscatter);  
proc sgplot data=baseball;  
scatter x=payroll y=wins / datalabel=team;  
run;



**R (10 points)**

##In R  
baseball <- read.csv('C:/Users/Charles/Documents/SMU/Online Teaching/MSDS 6371 - Statistical Foundations for Data Science/UNIT 8/Baseball\_Data.csv')  
  
plot(baseball$Payroll, baseball$Wins, xlab='Payroll', ylab='Wins', main='Baseball: Wins vs. Payroll')



**The relationship looks positive since the largest payrolls have mainly very many wins. However, many teams have equally as high wins with much lower payrolls. This leads us to believe the relationship is weak. There is a hint of evidence for a linear relationship, although the data are consistent with a curved trend as well. Our best guess of r is 0.2.**

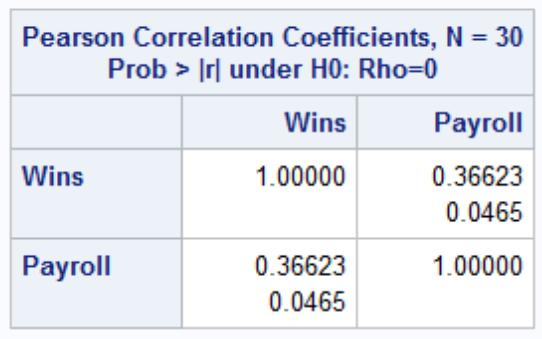
*Note: there can be flexibility in the responses to this question. The trend is definitely positive (3 points), but you may suggest the relationship is either moderate or weak (3 points). You may suggest the relationship is linear or curved (3 points), and your guess should be consistent with the rest of your answers (1 point). However, any guess larger than 0.5 is not reasonable and does not receive credit.*

## Question 2 (20 points total)

Find the correlation between team payroll and the number of wins. (No fair going back and changing your answer to the previous question! You should do this in both R and SAS.

**SAS (10 points)**

proc corr data=baseball;  
var wins payroll;  
run;



**R (10 points)**

##In R  
baseball <- read.csv('C:/Users/Charles/Documents/SMU/Online Teaching/MSDS 6371 - Statistical Foundations for Data Science/UNIT 8/Baseball\_Data.csv')  
  
cor(baseball$Payroll, baseball$Wins)

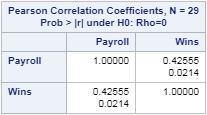
## [1] 0.366231

## Question 3 (20 points total)

San Diego (SD) has a payroll of 38 million, yet has 90 wins—more than Boston does 9 (with a payroll of $162 million). Delete SD from the data and rerun the analysis (scatter plot and correlation value). How does the correlation change? You may use your preference here, R or SAS.

**SAS**

proc corr data=baseball;  
where Team ne ‘SD’;  
var wins payroll;  
run;

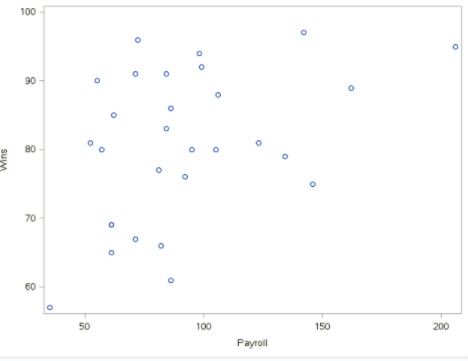


proc sgscatter data = baseball;

where team ne 'SD';

plot wins\*payroll;

run;

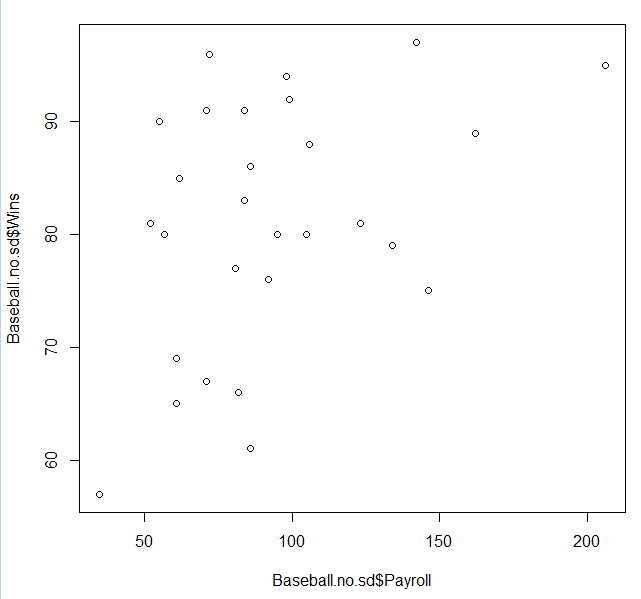


**R**

baseball.no.sd <- subset(baseball, TEAM != 'SD')  
cor(baseball.no.sd$Payroll, baseball.no.sd$Wins)

## [1] 0.4255494

plot(baseball.no.sd$Payroll, baseball.no.sd$Wins)



**The correlation increases with the deletion of the San Diego observation.**

*Note: you only need to do this once (in SAS or R, not both). The correlation is worth 10 points, the graph is worth 5 points, and commenting that is increases is worth 5 points.*

## Question 4 (15 points total)

The league commissioner notes that the Texas Rangers (TEX), with one of the lowest payrolls, won 90 games (and were the American League Champions) and the Chicago Cubs (CHC), with the highest payroll, won only 75 games. He argues that this proves that there is no advantage to teams with a higher payroll. Comment on his argument.

**Our analysis has only indicated that payroll is correlated with wins. We have not established a causation. There may be other factors besides the payroll that influence the number of wins. To establish causation, one idea would be to randomly assign payrolls (budgets) to different teams and observe the correlation. This gets complicated and is obviously unlikely.**

*Note: most reasonable answers should receive full points. There is no one particular correct answer.*

## Question 5 (15 points total)

What is the population for these data? Can these data be considered as a random sample from that population?

**There are two possible answers.**

**1) If the population is simply 2010 major league baseball teams, then this is a census. 2) If, however, the population is MLB teams in general, then these thirty teams are a sample from that population but do not form a random sample as only teams from 2010 were eligible to be selected into the survey.**

*Note: if you went with the first option and claim it is a census, this is worth the full 10 points. If you went with the second option, 5 points can be given for the statement on the population and 5 points can be given for the statement that they do not compose a* ***random*** *sample.*