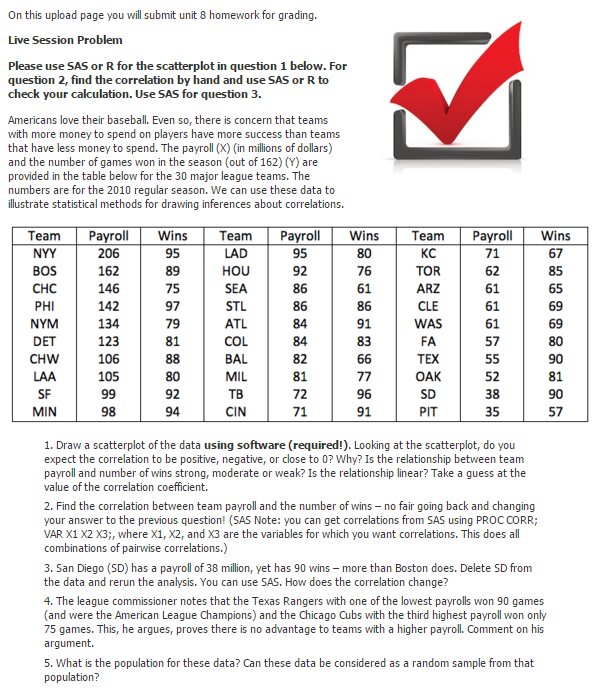
UNIT 8 HW



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 all of 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.



1. 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.

Code:

PROC IMPORT OUT= WORK.BeerData

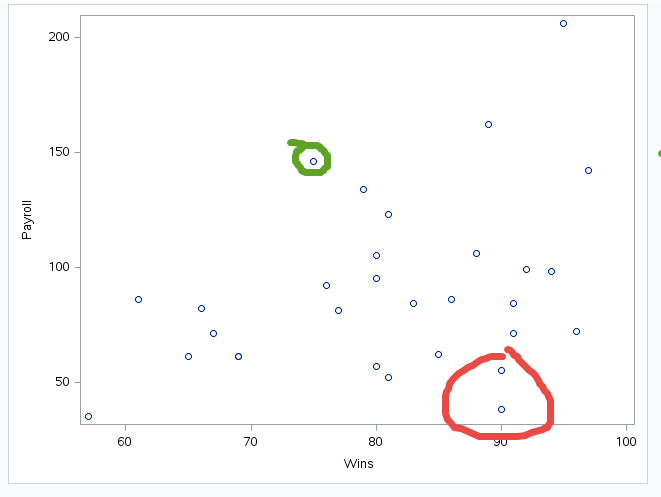
DATAFILE= "/home/marinfamily1010/sasuser.v94/Data/BeerPriceData.xlsx"

DBMS=xlsx REPLACE;

GETNAMES=YES;

DATAROW=2;

RUN;



There seems to be a relationship between payroll and wins though there are a few outliers circled in red and one questionable outlier in green.

R Code:

baseball <- read.csv("Baseball\_Data.csv")

correlation <- as.data.frame(cor(baseball))

n <- as.numeric(nrow(baseball))

r <- correlation[1,2]

criticalValue <- as.data.frame(qt(c(.025,.975), df = n-2))

cv <- criticalValue[2,1]

testStat <- (r\*sqrt(n-2))/(sqrt(1-r^2))

cortest <- cor.test(baseball$Payroll, baseball$Wins)

pvalue <- cortest$p.value

cortest

g <- ggplot(baseball, aes(Wins, Payroll ))

# Scatterplot

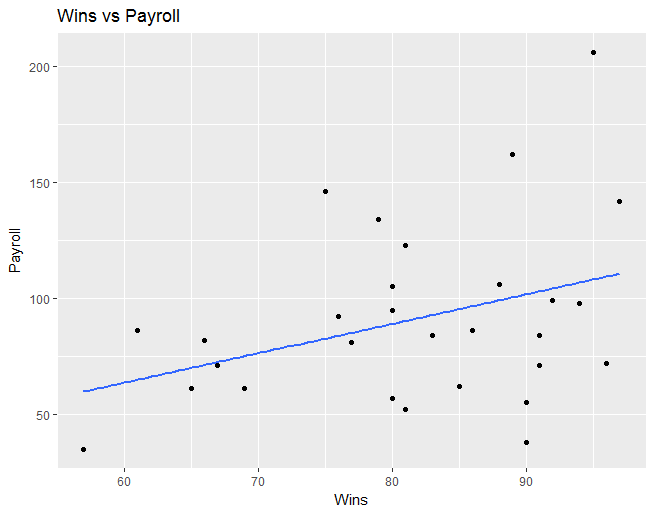
g + geom\_point() +

geom\_smooth(method="lm", se=F) +

labs(y="Payroll",

x="Wins",

title="Wins vs Payroll")



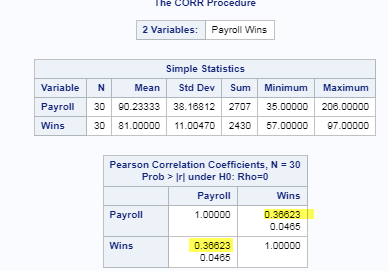
With the GGPlot R Code, I have added a regression line showing a positive correlation between wins and payroll.

1. 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.

Code:

proc corr data = Work.Baseball;

run;



R Code:

baseball <- read.csv("Baseball\_Data.csv")

cortest <- cor.test(baseball$Payroll, baseball$Wins)

cortest

data: baseball$Payroll and baseball$Wins

t = 2.0826, df = 28, p-value = 0.04654

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.00686799 0.64181770

sample estimates:

cor

0.366231

The correlation is .366 in SAS and R.

1. San Diego (SD) has a payroll of $38 million, yet SD has 90 wins – more than Boston does (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.

library(sqldf)

baseball <- read.csv("Baseball\_Data.csv")

baseball <- sqldf("select \* from baseball where Team <> 'SD'")

cortest <- cor.test(baseball$Payroll, baseball$Wins)

cortest

data: baseball$Payroll and baseball$Wins

t = 2.4435, df = 27, p-value = 0.02136

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.06995422 0.68518874

sample estimates:

cor

0.4255494

The correlation is stronger and improved from .36 to .43 which is 7% higher after deleting was appeared to be the outlier of one team, SD.

1. 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 third 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.

Well, let’s do a test:

If we don’t exclude what is considered to be an outlier, the results of the hypothesis test are a little different. For example, with the Texas Rangers included at an alpha level of .05 we have the following:

|  |  |
| --- | --- |
| Outlier of TX Include | Outlier of TX Excluded |
| Step 1: Ho:  Ha:  Step 2: Critval: 2.048407  Step 3: T-stat: 2.0826  Step 4: P- Value: .0465  Step 5: Well, if we round the p-value, we can make the argument that we can Fail to reject the null. If we don’t round, we can technically reject the null. I usually round in this case though and will do so to be consistent.  Step 6: Conclusion:  Without excluding the TX data, the data suggests that there is no evidence to suggest with an alpha of .05 that the data is linearly correlated (p-value = .05) and we fail to reject the null hypothesis. | Step 1: Ho:  Ha:  Step 2: Critval: 2.051831  Step 3: T-stat: 2.44  Step 4: P- Value: .02136  Step 5: Reject the null    Step 6: Conclusion:  If we exclude TX from the sample, there is strong  evidence to suggest with an alpha of .05 that the  data are linearly correlated (p-value = . 0.02136)  with a 95% confidence interval of [0.06995422,  0.68518874]  Given that by excluding TX we get pretty significant different results, I would conclude that TX is an outlier. |

The league commissioner’s statement is not valid in regards to disapproving that there is no

relationship between Payroll and Wins and by reviewing the scatterplot and doing a hypothesis test with and without the data, we can conclude that TX is an outlier.

Had the commission stated that outliers do exist in baseball and a team can strive to be an

outlier with a low payroll and a lot of wins, that statement will be valid given TX being

evidence.

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

According to the problem statement, this population applies to major league baseball and all teams have been included within the major leagues, so I don’t believe this dataset is a random sample in regards to teams, but it is a random sample in regards to the year since only 2010 data has been provided. This observational study that only applies to 2010 data.