Take\_Home\_Final\_Sai\_Giridhar

Sai Giridhar Rao Allada

12/16/2021

# Q1

IPLpoints <- read.csv("IPLpoints.csv")  
IPLpoints2 <- read.csv("IPLpoints2.csv")  
names(IPLpoints)[1]<-'Year'  
IPL14 <- subset(IPLpoints, Year %in% c(2008:2011, 2014:2021))  
p.win<-676/1372

## a

expected<-98 \* dbinom(0:14, 14, p.win)  
  
actual\_expected<-c(sum(expected[1:5]))  
actual\_expected<-c(actual\_expected,expected[6:10])  
actual\_expected<-c(actual\_expected,sum(expected[11:15]))  
cat("Expected wins for each category:",actual\_expected)

## Expected wins for each category: 9.70368 12.67504 18.46622 20.4978 17.42019 11.27974 7.957334

cat("\nVerifying if the Sum of all wins is 98:",sum(actual\_expected))

##   
## Verifying if the Sum of all wins is 98: 98

## b

possiblewins<-c(0,1,2,3,4,5,6,7,8,9,10,11,12,13,14)  
observed<-c()  
for (w in possiblewins){  
 observed<-c(observed,sum(IPL14$Wins == w))  
 #cat("\n Observed wins if total wins is ",w,":",sum(IPL14$Wins == w))  
}  
actual\_observed<-c(sum(observed[1:5]))  
actual\_observed<-c(actual\_observed,observed[6:10])  
actual\_observed<-c(actual\_observed,sum(observed[11:15]))  
cat("Observed wins for each category:",actual\_observed)

## Observed wins for each category: 12 8 18 24 13 16 7

## c

actual\_observed

## [1] 12 8 18 24 13 16 7

X2 <- sum((actual\_observed - actual\_expected)^2 / actual\_expected)  
cat("Pearson's Chi Squared statistic:", X2)

## Pearson's Chi Squared statistic: 6.089939

cat("P-Value:",1 - pchisq(X2, df=6))

## P-Value: 0.4131909

## d

The IPL is not completely random as there are strong teams and weak teams. The large pvalue confirms the null hypothesis but does not prove that the game is completely random. Some teams tend to perform better tha others.

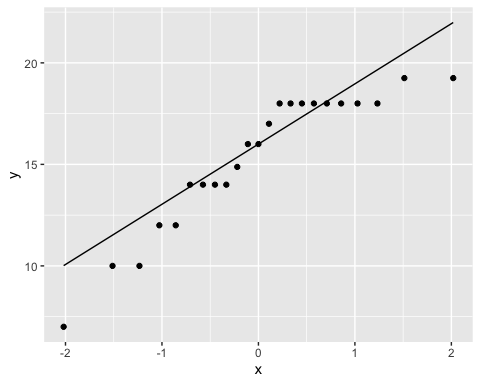
# Q2

library(ggplot2)  
temp=IPLpoints[IPLpoints$NZCoach==1,]  
temp

## Year Team NZCoach Wins Losses NoResult Points AdjPoints  
## 1 2021 CSK 1 9 5 0 18 18.000  
## 3 2021 KKR 1 7 7 0 14 14.000  
## 6 2021 RCB 1 9 5 0 18 18.000  
## 9 2020 CSK 1 6 8 0 12 12.000  
## 11 2020 KKR 1 7 7 0 14 14.000  
## 17 2019 CSK 1 9 5 0 18 18.000  
## 25 2018 CSK 1 9 5 0 18 18.000  
## 30 2018 RCB 1 6 8 0 12 12.000  
## 38 2017 RCB 1 3 10 1 7 7.000  
## 39 2017 RPS 1 9 5 0 18 18.000  
## 46 2016 RCB 1 8 6 0 16 16.000  
## 47 2016 RPS 1 5 9 0 10 10.000  
## 49 2015 CSK 1 9 5 0 18 18.000  
## 54 2015 RCB 1 7 5 2 16 16.000  
## 57 2014 CSK 1 9 5 0 18 18.000  
## 60 2014 MI 1 7 7 0 14 14.000  
## 62 2014 RCB 1 5 9 0 10 10.000  
## 65 2013 CSK 1 11 5 0 22 19.250  
## 68 2013 MI 1 11 5 0 22 19.250  
## 74 2012 CSK 1 8 7 1 17 14.875  
## 83 2011 CSK 1 9 5 0 18 18.000  
## 93 2010 CSK 1 7 7 0 14 14.000  
## 101 2009 CSK 1 8 5 1 17 17.000

ggplot(temp, aes(sample = temp$AdjPoints)) + stat\_qq() + stat\_qq\_line()

## Warning: Use of `temp$AdjPoints` is discouraged. Use `AdjPoints` instead.  
  
## Warning: Use of `temp$AdjPoints` is discouraged. Use `AdjPoints` instead.

 The points are not perfectly normal as seen in the above plot but are good enough to go ahead with hypothesis test.

t.test(temp$AdjPoints,mu=14,alternative = "greater")

##   
## One Sample t-test  
##   
## data: temp$AdjPoints  
## t = 1.9521, df = 22, p-value = 0.03188  
## alternative hypothesis: true mean is greater than 14  
## 95 percent confidence interval:  
## 14.16419 Inf  
## sample estimates:  
## mean of x   
## 15.36413

We perform a right tailed t-test to confirm our hypothesis that the Expected value of AdjPoints with a New Zealender coach to be greater than 14 hence why we consider the the null hypothesis to be less than or equal to 14. A low P-value of 0.03188 now proves that we can reject the null hypothesis. 95 % confidence interval of is [14.1619,Inf].

# Q3

IPLBig10 <- subset(IPLpoints, Team %in% c("CSK", "DC", "KKR", "HDC", "MI",  
"PBKS", "PWI", "RCB", "RR", "SRH"))  
iplm<-lm(IPLBig10$AdjPoints~IPLBig10$Team)  
anova(iplm)

## Analysis of Variance Table  
##   
## Response: IPLBig10$AdjPoints  
## Df Sum Sq Mean Sq F value Pr(>F)   
## IPLBig10$Team 9 348.61 38.734 2.8644 0.004743 \*\*  
## Residuals 101 1365.81 13.523   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The analysis of variance tells us that the performance of the IPL teams are all not similar. Here we assume the null hypothesis to be that all the teams have similar AdjPoints and the alternative to be that they don’t. We can reject the null hypothesis due to the really low p-value. The follow up analysis which could be done could include checking anova between different possible subsets of teams to distinguish the teams that are really good and those particularly aren’t. ## Q4 We assume that the distribution is normal and is of a sufficient sample size in Q2 in order to perform the t-test. In Q3 we assume that each sample is independent of the other which might not be the case exactly. The different samples of teams should also have the same variance of . The sample isn’t large enough either. # Q5 ## a

IPL\_pred <- lm(IPLpoints2$AdjPoints2~IPLpoints2$AdjPoints1)  
summary(IPL\_pred)

##   
## Call:  
## lm(formula = IPLpoints2$AdjPoints2 ~ IPLpoints2$AdjPoints1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.3955 -2.0228 0.2144 2.7400 7.9772   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 12.36230 1.43457 8.617 1.12e-13 \*\*\*  
## IPLpoints2$AdjPoints1 0.11861 0.09811 1.209 0.23   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.783 on 99 degrees of freedom  
## Multiple R-squared: 0.01455, Adjusted R-squared: 0.004595   
## F-statistic: 1.462 on 1 and 99 DF, p-value: 0.2296

$E[X]= slope AdjPointsIn2021 + Intercept $ \ Hence $E[SRH] = 0.11861 + 12.36230 = $

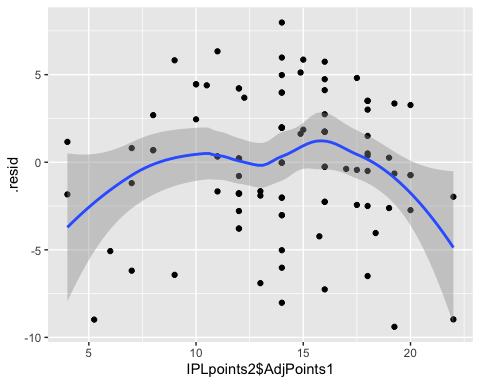
pred=0.11861\*6 +12.3620  
pred

## [1] 13.07366

## b

library(broom)  
lmdf <- augment(IPL\_pred)  
ggplot(lmdf, aes(IPLpoints2$AdjPoints1, .resid)) + geom\_point() + geom\_smooth()

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'

 It is not linear as the line should be horizontal. Our datapoints arent exactly independent as one team’s loss effect the other team’s win.