Star Power

RJ Cass

The NBA All-star game is a showcase basketball game focused on being a showcase of NBA talent. Given the NBA is a Professional League, this game is usually the best of the best showing off, and has the potential to be a spectale. This game is played around mid-February (middle of the NBA season). Being a professional league, the players do get paid, but the winning team also wins money for a charity of their choice. Because it's a showcase, the game tends to be more flashy, with players going for big moves (slam-dunks, etc.).

The players for the All-star game are picked via voting, with fans having 50% of the vote, and media and players each having 25% of the vote. The NBA is split into two conferences (East and West), and for the All-star game 12 players from each conference are selected. Previously voting was limited to players in specific positions, but starting in 2026 the voting will not be limited to positions.

Who was a 2003 All-Star?

#Available at this site https://www.basketball-reference.com/allstar/NBA_2003.html Michael Jordan, Shaq, Kobe

In the 2011 game, what team did LeBron James represent?

He was playing for the Heat at the time, and played for East Conference

What happened in 2016? Anybody from 2016 still playing?

First All-star game outside the US (in Canada) East Conference head coach would have been one guy, but he was fired Steph Curry, Lebron James, and others played, and are still playing

Other questions about this time period:

Impact of Michael Jordan retiring in 2003?

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.2     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```
library(car)
```

```
Loading required package: carData

Attaching package: 'car'

The following object is masked from 'package:dplyr':

    recode

The following object is masked from 'package:purrr':

    some
```

```
asg <- read_csv("https://grimshawville.byu.edu/BYUStat535/NBATVaudience.csv")
```

```
Rows: 16 Columns: 5
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (5): Year, TVaud, maxPER, best10PER, pointspread

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
summary(asg)
```

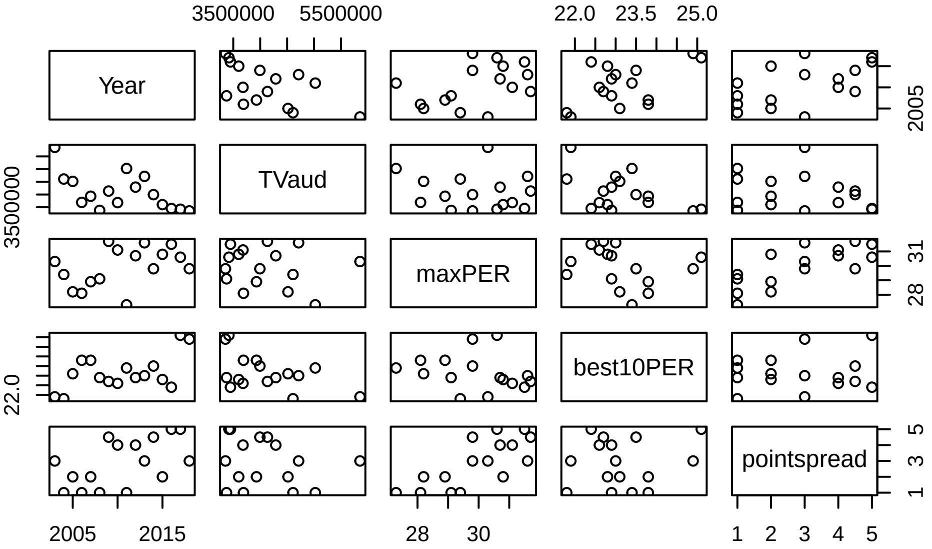
```
      Year          TVaud              maxPER         best10PER    
 Min.   :2003   Min.   :3355707   Min.   :27.30   Min.   :21.80  
 1st Qu.:2007   1st Qu.:3563823   1st Qu.:29.05   1st Qu.:22.68  
 Median :2010   Median :3963193   Median :30.05   Median :22.95  
 Mean   :2010   Mean   :4102464   Mean   :29.93   Mean   :23.16  
 3rd Qu.:2014   3rd Qu.:4539426   3rd Qu.:30.88   3rd Qu.:23.57  
 Max.   :2018   Max.   :5852653   Max.   :31.70   Max.   :25.10  
  pointspread   
 Min.   :1.000  
 1st Qu.:1.750  
 Median :3.000  
 Mean   :2.875  
 3rd Qu.:4.125  
 Max.   :5.000  
```

Nielsen TV Audience: uses devices in a sample of homes to track what's watched, then estimate to the population

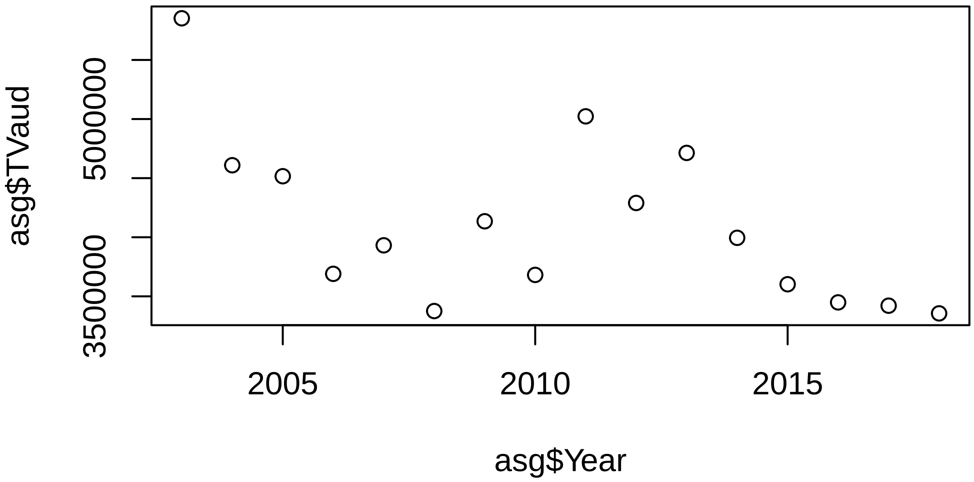
Hollinger's PER: Player Efficieny Rating. 15 is league average. (Positive Actions - Negative Actions)/minutes_player best10: the PER of the 10th - The provided PER values are calculated over the whole season

Point Spread: betting +5 means that your team can lose by up to 5 points and you still win. Smaller values indicates an expected close game, which draws viewership

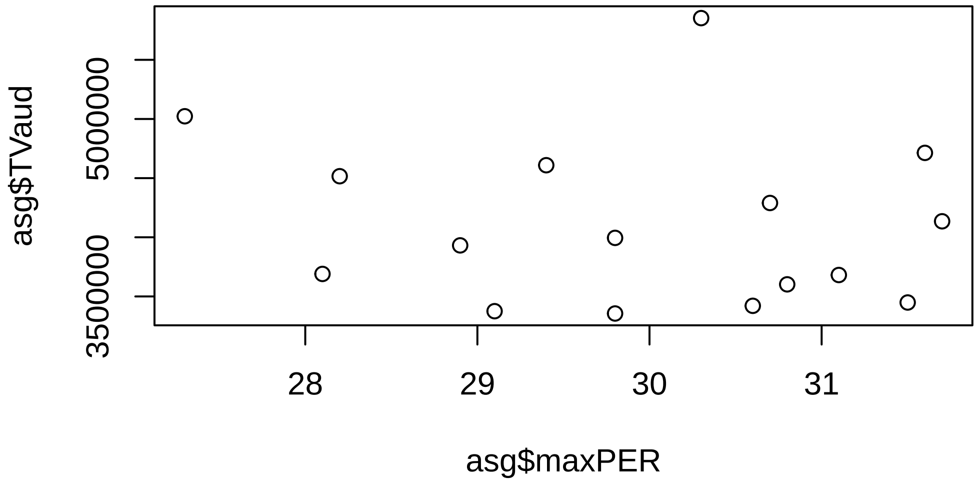
```r
# This install was not working
# library(GCally)
# install.packages('GCally')

pairs(asg)
```



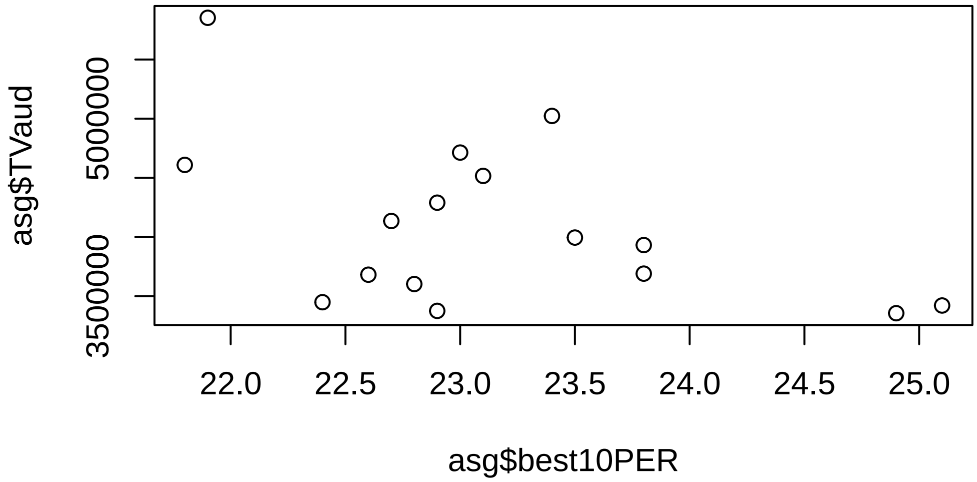
```r
plot(asg$Year, asg$TVaud) # Linearish, decreasing
```



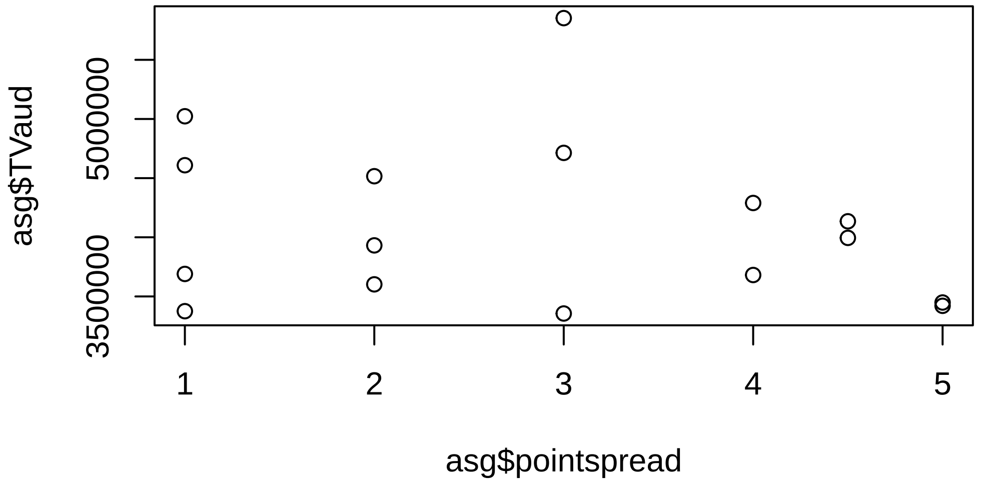
```r
plot(asg$maxPER, asg$TVaud) # Not really any relationship
```



```
plot(asg$best10PER, asg$TVaud) # Linearish, decreasing
```

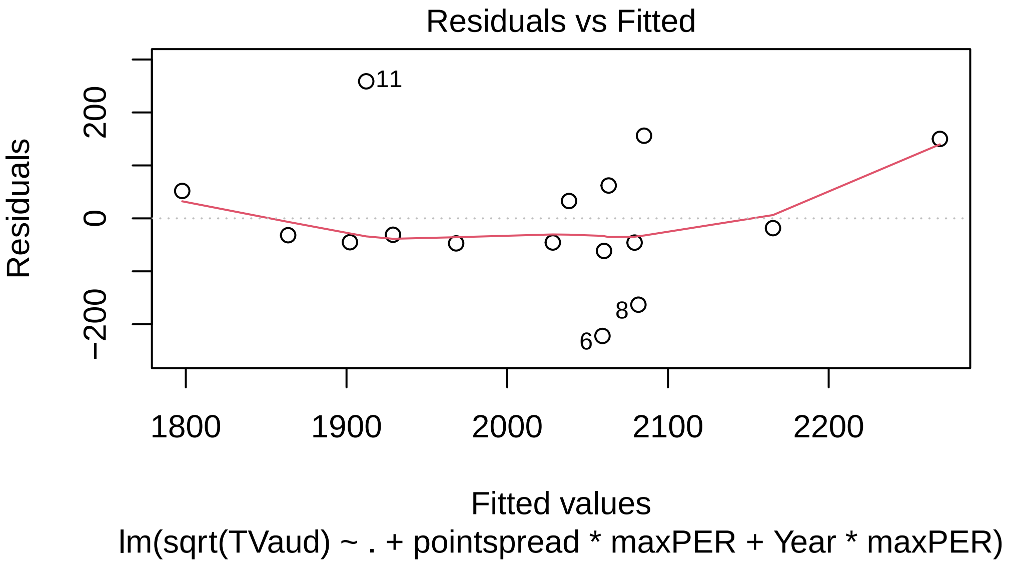


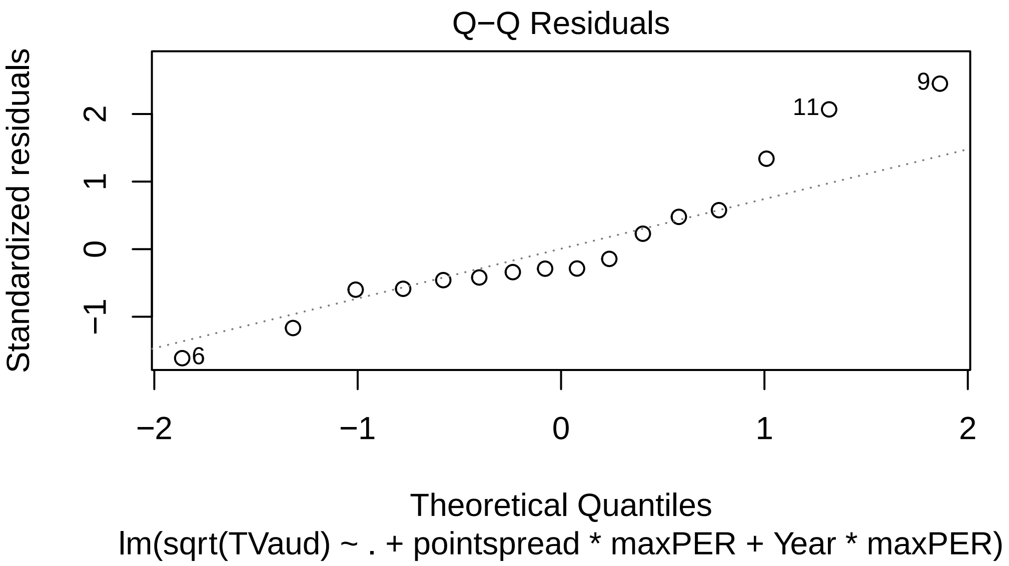
```
plot(asg$pointspread, asg$TVaud) # Linearish, weakly decreasing
```

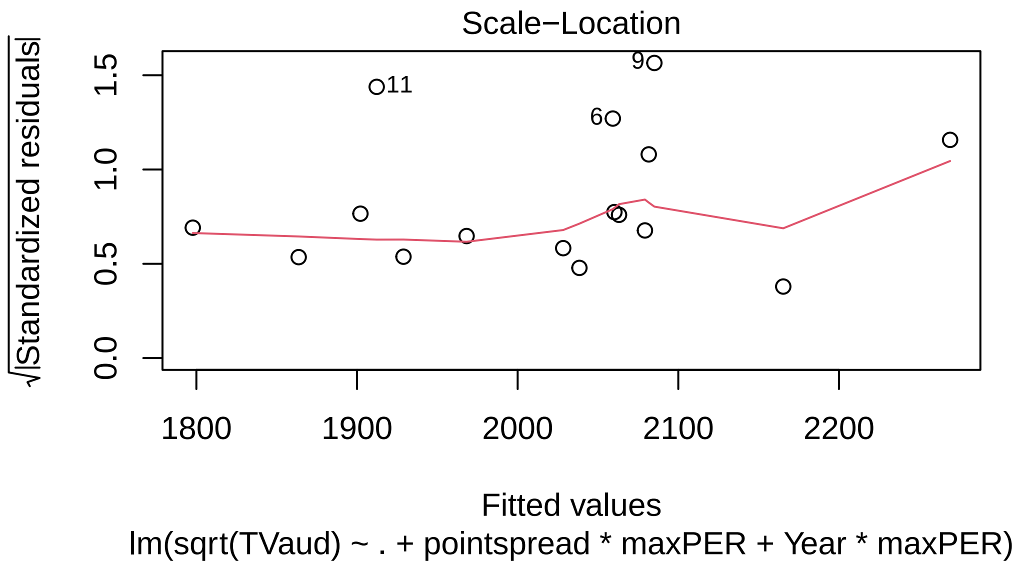


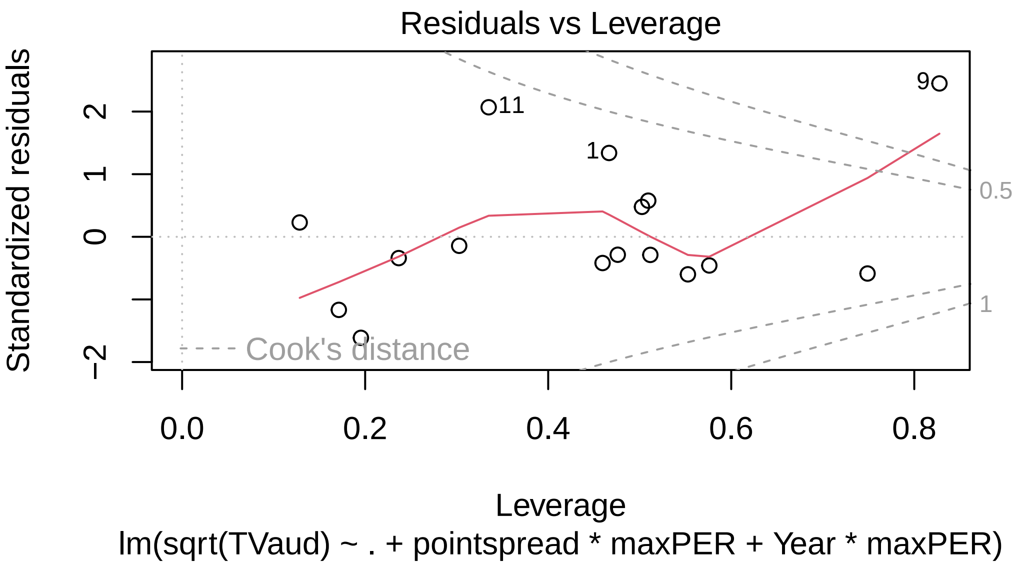
Let's build a model:

```
mod <- lm(sqrt(TVaud) ~ . + pointspread*maxPER + Year*maxPER, data = asg)
plot(mod)
```

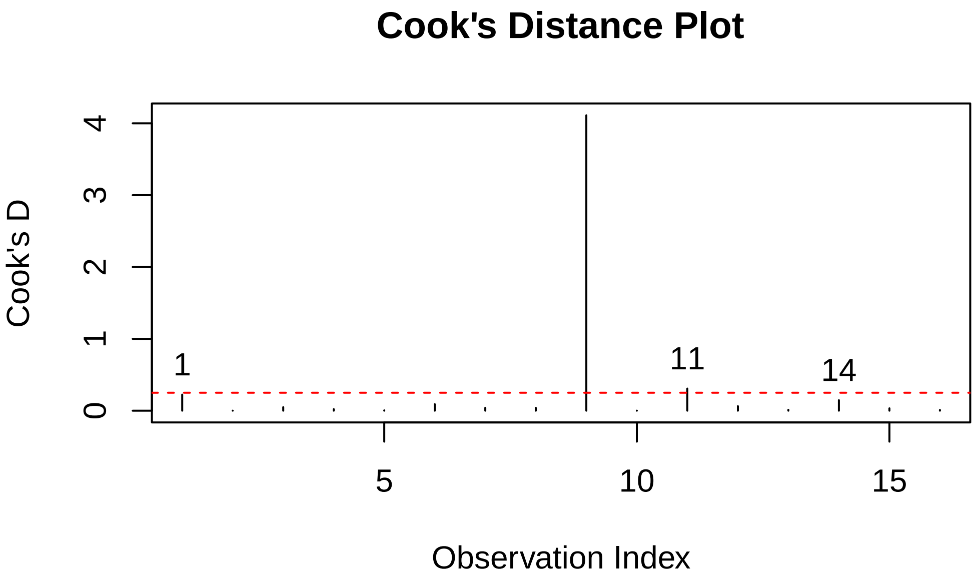



Q−Q Residuals
Standardized residuals
−1
0
1
2
−2
−1
0
1
2
Theoretical Quantiles
lm(sqrt(TVaud) ~ . + pointspread * maxPER + Year * maxPER)
9
11
6


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
1800
1900
2000
2100
2200
Fitted values
lm(sqrt(TVaud) ~ . + pointspread * maxPER + Year * maxPER)
9
11
6

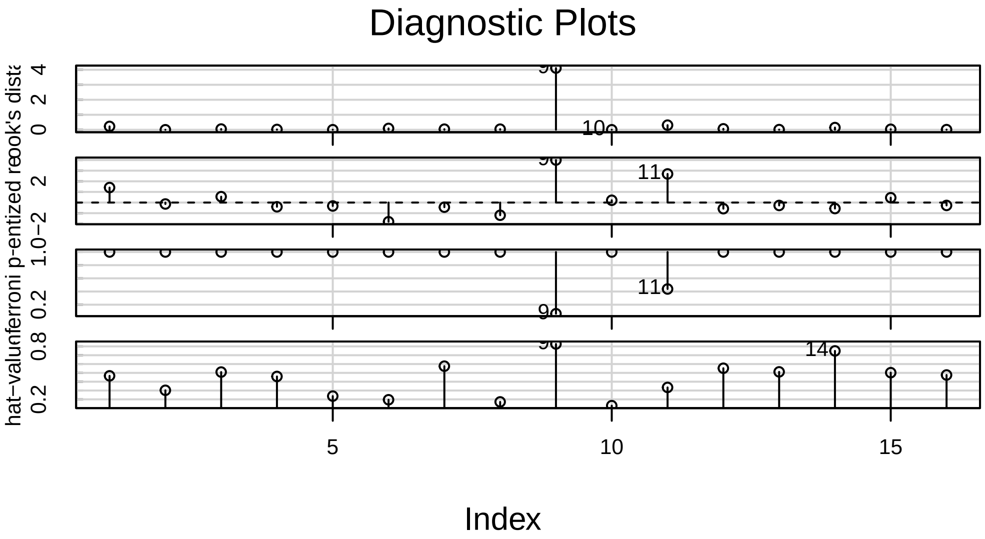


```
# Other things to try:
cooks_d <- cooks.distance(mod)
plot(cooks_d, type = "h", main = "Cook's Distance Plot", xlab = "Observation Index", ylab = "
abline(h = 4/length(cooks_d), col = "red", lty = 2) # Common threshold
text(which(cooks_d > 0.1), cooks_d[cooks_d > 0.1], labels = which(cooks_d > 0.1), pos = 3) #
```



```
# Index 9, Year 2011, seems to be an outlier/odd value
# Indeces 11, and 1 are also rough (14 is also interesting)
```

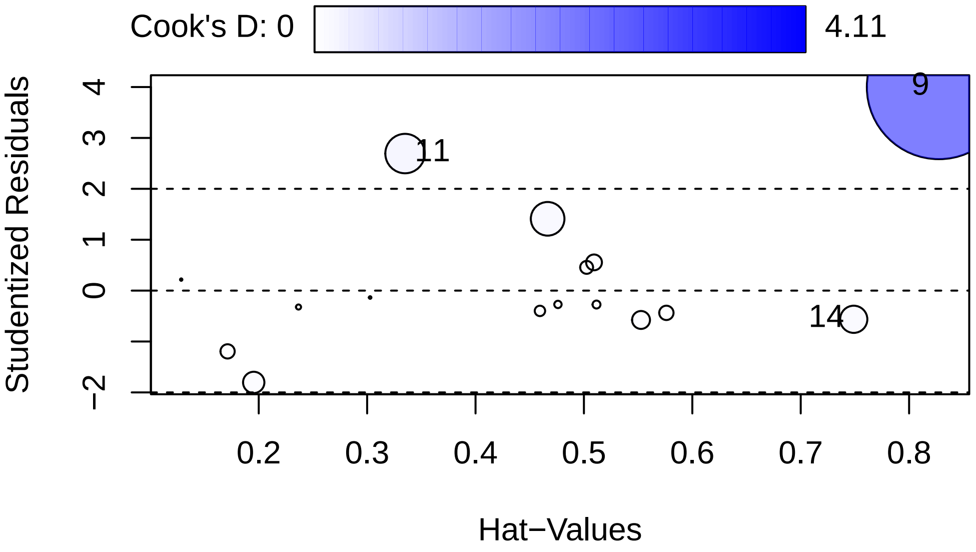
```r
# From CAR package
influenceIndexPlot(mod)
```



```r
outlierTest(mod)
```

```
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
  rstudent unadjusted p-value Bonferroni p
9 3.999546          0.0039523     0.063236
```

```r
influencePlot(mod)
```



```
      StudRes       Hat     CookD
9   3.9995460 0.8275105 4.1117931
11  2.6927028 0.3349665 0.3078867
14 -0.5636895 0.7489247 0.1465052
```

```
vif(mod)
```

```
there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif

              Year             maxPER          best10PER        pointspread
      1.406942e+03       5.018881e+05       2.593487e+00       1.130710e+03
maxPER:pointspread         Year:maxPER
      1.230235e+03       5.258635e+05
```

```
summary(mod)
```

```

Call:
lm(formula = sqrt(TVaud) ~ . + pointspread * maxPER + Year *
    maxPER, data = asg)

Residuals:
    Min      1Q  Median      3Q     Max
-222.02  -46.13  -31.40   54.31  258.79

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -715874.14  627106.10  -1.142    0.283
Year                   358.28     312.07   1.148    0.281
maxPER               24827.67   20910.31   1.187    0.265
best10PER              -78.27      69.55  -1.125    0.290
pointspread            334.92     894.61   0.374    0.717
maxPER:pointspread     -10.19      29.38  -0.347    0.737
Year:maxPER            -12.36      10.41  -1.188    0.265

Residual standard error: 153.4 on 9 degrees of freedom
Multiple R-squared:  0.4984,    Adjusted R-squared:  0.164
F-statistic: 1.491 on 6 and 9 DF,  p-value: 0.2832
```

Looking at outliers, years 2003 and 2011 seem to be outliers. We will remove them from the dataset.

```
# We're going to remove years 2003 and 2011
asg1 <- asg[c(-1, -9), ]
out <- lm(TVaud ~ maxPER + best10PER + pointspread + Year, data = asg1)
# How do things change after removing those points?

# Do the estimated coefficients have expected sign? Both Scatterplot and Theory
summary(out)
```

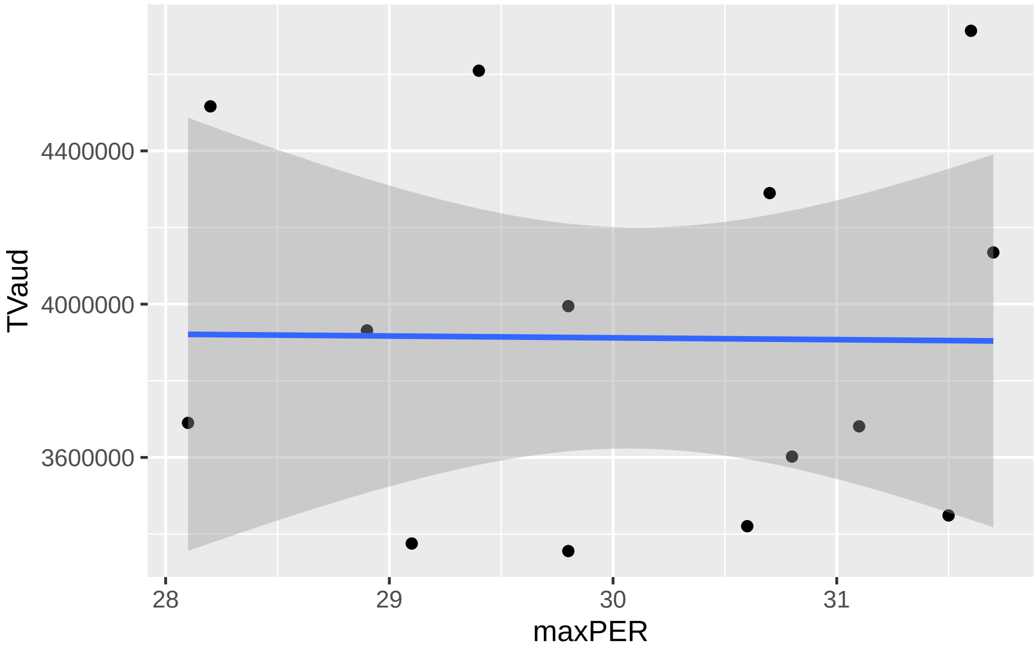
```
Call:
lm(formula = TVaud ~ maxPER + best10PER + pointspread + Year,
    data = asg1)

Residuals:
    Min      1Q  Median      3Q     Max
-589046 -239830  -85337  299121  748666

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 146441042   95615101   1.532    0.160
maxPER         125054     215310   0.581    0.576
best10PER      -39136     226346  -0.173    0.867
pointspread     26636     138089   0.193    0.851
Year           -72334      51185  -1.413    0.191

Residual standard error: 455100 on 9 degrees of freedom
Multiple R-squared:  0.367, Adjusted R-squared:  0.08572
F-statistic: 1.305 on 4 and 9 DF,  p-value: 0.3386
```

```
# What does 125054 mean?
```

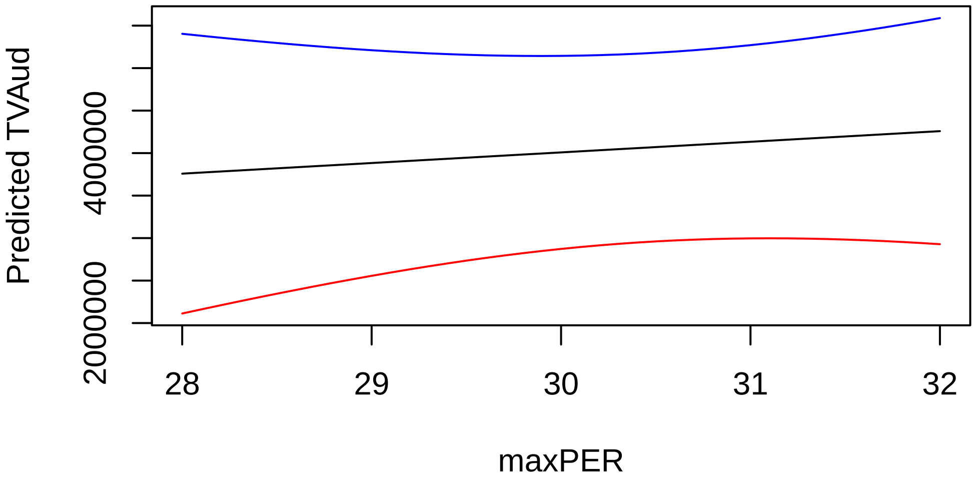
How do we look at the effect of maxPER?

```
# Pretty standard graph for effect:
ggplot(asg1, aes(x = maxPER, y = TVaud)) +
    geom_point() +
    geom_smooth(method = "lm")
```

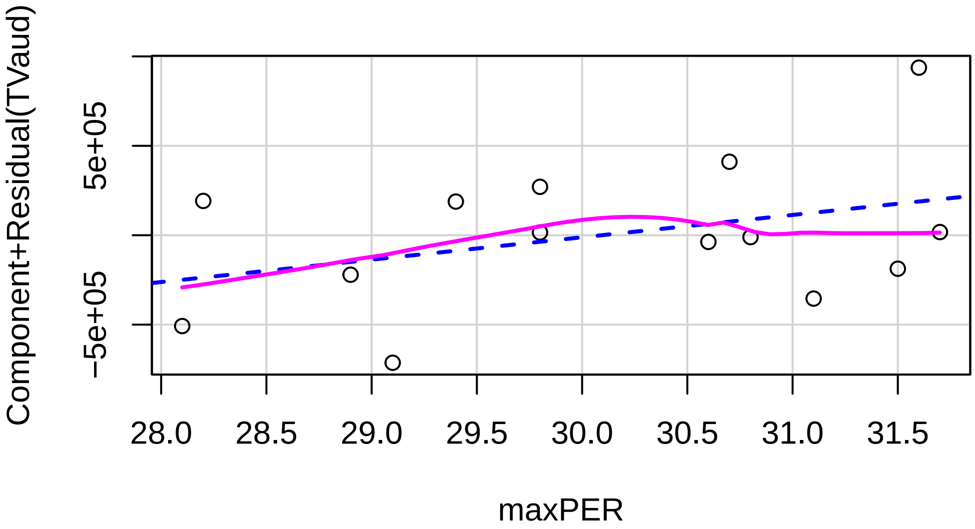
```
`geom_smooth()` using formula = 'y ~ x'
```



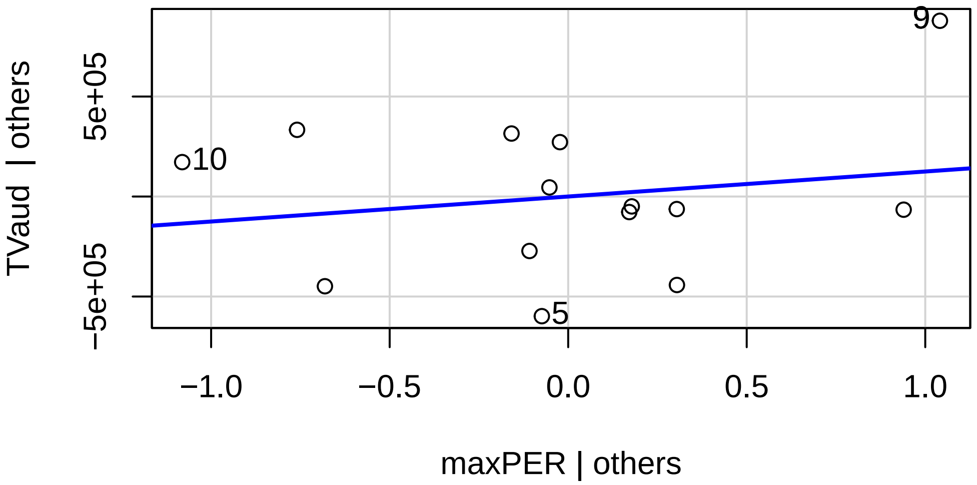
```
# What's good, bad, ugly?
# Good: Shows uncertainty
# Possibly confusing: points outside gray
# Misleading: Positive partial slope looks like 'no effect'


# Graph I made
# Uses the predictions
# What can we plot instead to show effect of maxPER on TVAud?
mPer <- seq(28, 32, by = .1)
botPer <- 23
year <- 2010
point_spread <- 4
data <- expand.grid(mPer, botPer, year, point_spread)
colnames(data) <- c('maxPER', 'best10PER', 'Year', 'pointspread')
p <- predict(out, data, interval = "prediction")
plot(mPer, p[, 1], type = 'l', ylim = c(min(p[, 2]), max(p[,3])), xlab = 'maxPER', ylab = 'P
lines(mPer, p[, 2], col = 'red', pch = 19)
lines(mPer, p[, 3], col = 'blue')
```



```
# Other plots people use:
# Component + Residual plots
crPlots(out, terms = ~ maxPER) # Learn some of the details of how this works
```



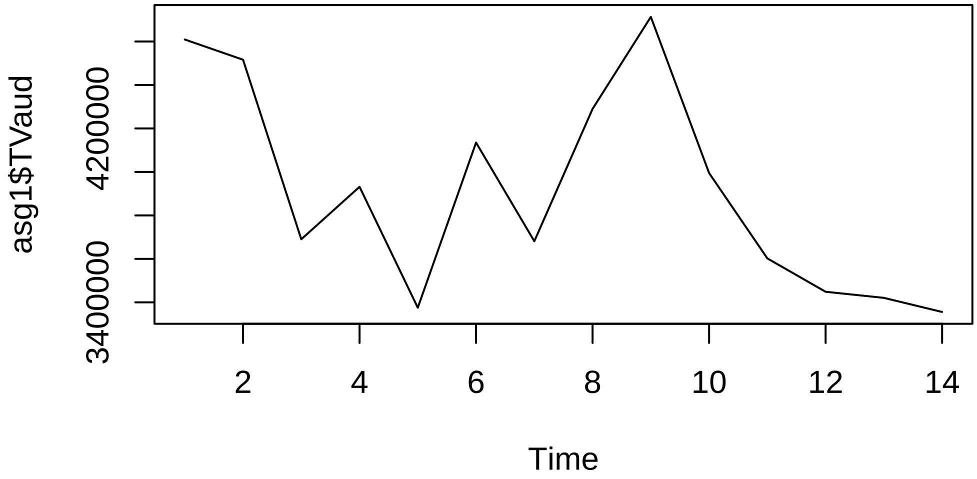
```
# Added Variable plot
avPlots(out, terms = ~ maxPER) # yAxis: If I make a model wihtout that var. xAxis: If I use
```



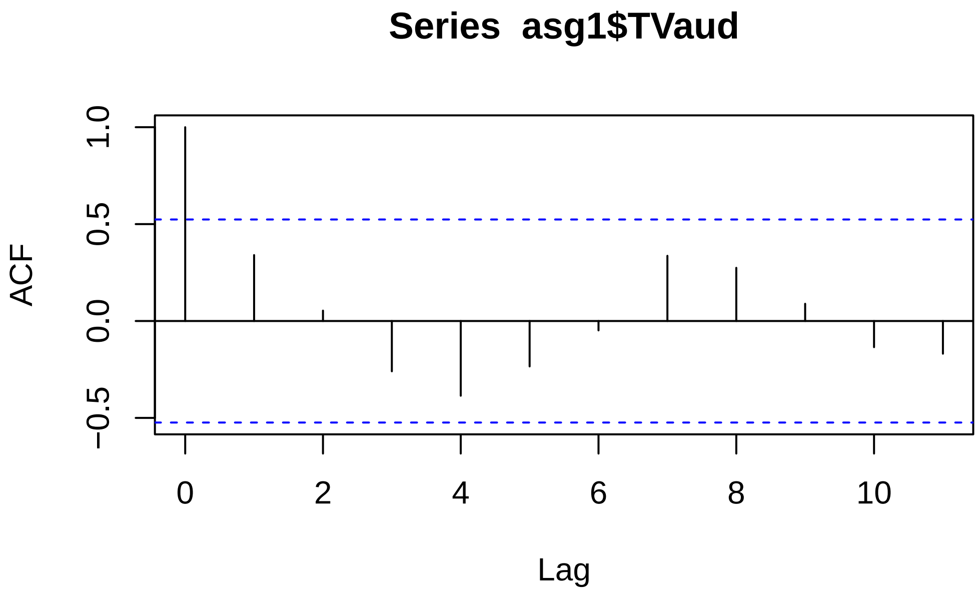
```r
# Other graph is actually exactly what I did above using the predictions
```

Thinking about other things we can do to the data:

```r
# Maybe log transform the response, consider correlation over time?

ts.plot(asg1$TVaud)
```



```r
acf(asg1$TVaud)
```



Examining the actual vectors involved in our mathematical representation of the model

```
out <- lm(I(TVaud / 10 ^ 6) ~ maxPER + best10PER + pointspread + Year,
          data = asg1,
          y = TRUE, x = TRUE)
t(t(out$y))
```

```
       [,1]
1  4.609195
2  4.516170
3  3.690126
4  3.931487
5  3.375371
6  4.135003
7  3.681147
8  4.289902
9  4.713407
10 3.994899
11 3.602214
12 3.448649
13 3.420655
14 3.355707
```

```
out$x
```

```
   (Intercept) maxPER best10PER pointspread Year
1            1   29.4      21.8         1.0 2004
2            1   28.2      23.1         2.0 2005
3            1   28.1      23.8         1.0 2006
4            1   28.9      23.8         2.0 2007
5            1   29.1      22.9         1.0 2008
6            1   31.7      22.7         4.5 2009
7            1   31.1      22.6         4.0 2010
8            1   30.7      22.9         4.0 2012
9            1   31.6      23.0         3.0 2013
10           1   29.8      23.5         4.5 2014
11           1   30.8      22.8         2.0 2015
12           1   31.5      22.4         5.0 2016
13           1   30.6      25.1         5.0 2017
14           1   29.8      24.9         3.0 2018
attr(,"assign")
[1] 0 1 2 3 4
```

Quadratic form of matrices: Something transposed times itself

WHat other factors are interesting?

```
# Import
asgMORE <- read_csv("https://grimshawville.byu.edu/BYUStat535/NBATVaudienceMORE.csv")
```

```
Rows: 16 Columns: 7
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (7): Year, TVaud, maxPER, best5PER, best10PER, meanPER, pointspread

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# EDA
summary(asgMORE)
```

```
      Year          TVaud             maxPER         best5PER    
 Min.   :2003   Min.   :3355707   Min.   :27.30   Min.   :23.70  
 1st Qu.:2007   1st Qu.:3563823   1st Qu.:29.05   1st Qu.:24.68  
 Median :2010   Median :3963193   Median :30.05   Median :26.00  
 Mean   :2010   Mean   :4102464   Mean   :29.93   Mean   :25.63  
 3rd Qu.:2014   3rd Qu.:4539426   3rd Qu.:30.88   3rd Qu.:26.32
 Max.   :2018   Max.   :5852653   Max.   :31.70   Max.   :27.40
   best10PER        meanPER       pointspread
 Min.   :21.80   Min.   :21.49   Min.   :1.000
 1st Qu.:22.68   1st Qu.:22.46   1st Qu.:1.750
 Median :22.95   Median :22.85   Median :3.000
 Mean   :23.16   Mean   :22.79   Mean   :2.875
 3rd Qu.:23.57   3rd Qu.:23.08   3rd Qu.:4.125
 Max.   :25.10   Max.   :23.89   Max.   :5.000
```

```
##### Explore, make some models, make a graphic (challenge is to get down to 1, but more is
```