Baseball Viewership Analysis

Carly Lundgreen

September 8, 2020

Introduction

The MLB All-Star Game is a popular program within the local television market in Detroit, Michigan. Viewership of the game varies annually within the Detroit region, however. Various aspects of the local MLB team's roster construction may contribute to this variation. This analysis will seek to understand how the local Detroit TV audience viewership is affected by characterisitcs of the Detroit Tigers' roster. We will also estimate local Detroit TV viewership numbers for various scenarios. The procedure for this analysis will begin with an exploratory analysis of the data. We will then fit a multiple linear regression model, evaulate whether our data meets the assumptions of our model, and finally we will use the model to answer our research questions. We will also set the level of significance for this analysis as $\alpha = 0.05$.

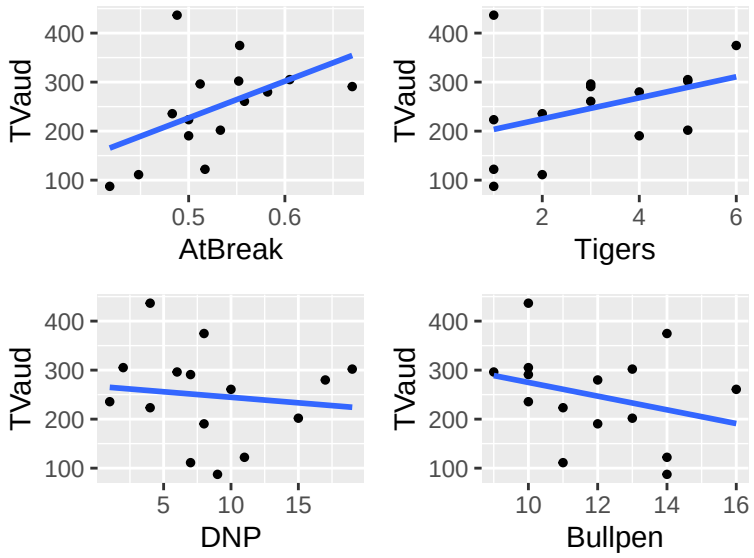
Exploratory Data Analysis

These data include a 15 year summary (2004-2018) of Detroit TV audience numbers ('TVAud') for the All-Star Game along with four variables relating to the Tigers' roster. Those four variables include: the number of Tigers named All-Stars that year ('Tigers'), the winning proportion of the team at the All-Star break ('AtBreak'), the number of players who are named All-Stars but do not play in the game that year ('DNP'), and finally the number of bullpen pitchers ('Bullpen').

Note that the first two variables ('Tigers' and 'AtBreak') are often perceived as a measures of the success and/or failure of the team for a given year. For example, when more Tigers are chosen for the All-Star game, our hypothesis is that the team has performed well that year, so TV viewership in Detroit will increase, and vice-versa. The variables 'DNP' and 'Bullpen' are explanatory variables that are often 'blamed' for local lack of interest in the All-Star game. For example, we hypothesize that the more All-Star Tigers that do not play in the game, the more people in Detroit will not be interested in watching the game. Our response variable for this analysis will be 'TVaud' (Detroit TV audience numbers in 1,000 households).

Exploratory Plots

The scatterplots below illustrate the relationships between each of our explanatory variables and our response variable (TVaud). The line of best fit for each of these plots is overlaid in blue. The patterns seen in these plots roughly follow the patterns discussed in the previous section. It appears that, on average, as 'AtBreak' (the Tigers' winning proportion) increases, the TV audience numbers tend to increase as well. We observe a similar trend between 'Tigers' (the number of Tigers selected for the All-Star Game) and TV viewership. The opposite trend is observed between viewership and the number of All-Star Tigers that do not play in the Game–it appears that viewership decreases on average as 'DNP' increases. Again, a similar trend is observed between viewership and 'Bullpen' (number of relief pitchers).



Summary Statistics

The table below provides the correlations between each of our covariates (excluding the Year of each observation) and our response variable. We can also see the mean and standard deviation for each of the covariates, as well as the response.

	TVaud	Tigers	AtBreak	DNP	Bullpen
Correlation	1	0.38	0.49	-0.12	-0.29
Mean	247870	3070.00	530.00	8530.00	11930.00
SD	96270	1710.00	60.00	5240.00	2020.00

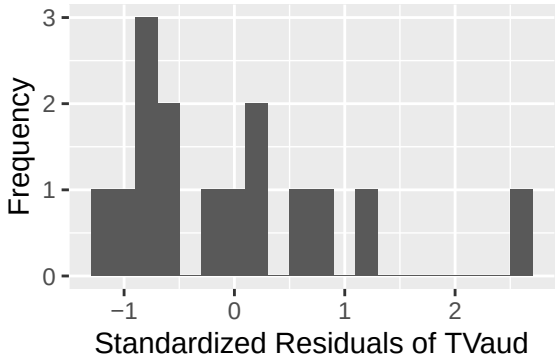
This table again illustrates the negative correlation between the more 'boring' aspects of roster construction (relief pitchers and non-playing All-Stars) and TV viewership. We also see positive correlations between our other two covariates and our response. However, we can see that the strength of these linear relationships is moderately weak, with the strongest correlation being 0.49: the correlation coefficient between the Tigers' Winning proportion and Detroit TV audience numbers. The table above also shows that the mean number of All-Star Game viewers in the Detroit TV market (for these 15 years) is 247870, with a standard deviation of 96270.

Fit Model and Verify Assumptions

We will fit the following model for this analysis:

$$\hat{TVaud} = \beta_0 + \beta_1 Tigers + \beta_2 AtBreak + \beta_3 DNP + \beta_4 Bullpen + \epsilon$$

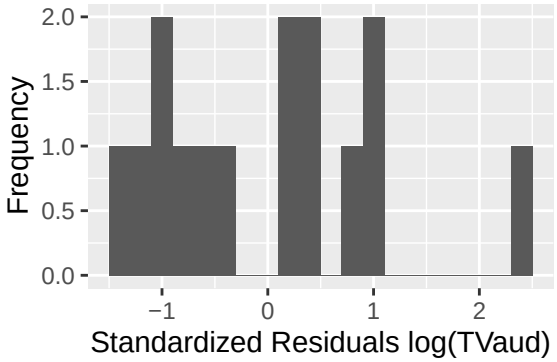
In fitting this model, we first assume linearity between each of our covariates and our response. Looking at the scatterplots in the previous section, we concluded that the plots appeared sufficiently linear to proceed. We also assume that each year's observations of TV viewership are independent from the last. In this situation it is possible that viewership is affected by previous years viewership. For example, those who chose not to watch the All-Star Game one year may be less likely to watch the next year. We will proceed with the analysis and assume that the observations are sufficiently independent, however we are mindful that there may be some year-to-year correlation between TV audience numbers. We also assume normality of the residuals, and looking at a histogram of the residuals below, it appears to have a bit of a right-skewness to it.

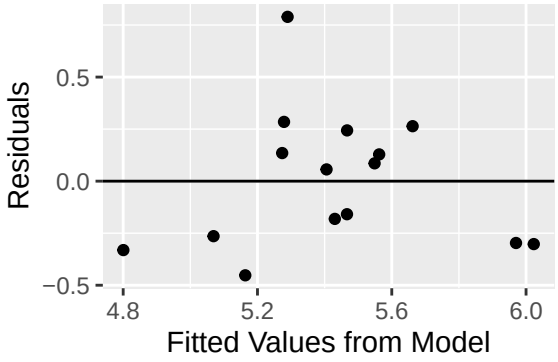


Because of this skewness, we will utilize the log transformation of the response variable, giving the new model:

$$\hat{log(TVaud)} = \beta_0 + \beta_1 Tigers + \beta_2 AtBreak + \beta_3 DNP + \beta_4 Bullpen + \epsilon$$

Below, we can see the updated residuals do appear to be more normal than the residuals from our untransformed model. As a final check, we plotted the fitted values from our transformed model against the residuals to verify the condition of equal variance. No concerning divergent patterns appear in this plot, so we conclude that the condition of equal variance of residuals is satisfied and this model is sufficient for our analysis.





Analysis

The table below displays the estimates for each β coefficient in our model, along with a 95% confidence interval for each of those coefficients. Note that we have 'untransformed' our estimates and intervals to counteract the log-transform we fit the model with originally. This makes each estimate and interval more interpretable.

	Estimate	Lower	Upper
(Intercept)	74.835	4.370	1281.604
Tigers	1.093	0.920	1.297
AtBreak	20.070	0.228	1766.468
DNP	0.986	0.931	1.044
Bullpen	0.949	0.825	1.093

From the table above, we can begin to understand the effects that our covariates have on Detroit TV audiences. For example, assuming all other covariates remain constant, on average the TV audiences increase by $e^{\beta_{Tigers}}$ (per thousand households) for every one additional Detroit tiger added to the All-Star team. Based on our model, we are 95% confident that the true effect of the number of Tigers named All-Stars on TV

audiences is between 0.92 and 1.297. We are also 95% confident that the true effect of bullpen pitchers on TV audiences in Detroit is between 0.825 and 1.093. Note that zero is contained in both of these intervals, which means that we don't have sufficient evidence to say there is a significant effect of either of these covariates on TV audience numbers.

We also wanted to know if how well the Tigers are playing has a significant effect on the popularity of the game. This would be a test of the following hypothesis:

$$H_o: \ \beta_{Tigers} = \beta_{AtBreak} \ = 0$$
$$H_a: \ At\ least\ one\ of\ the\ two\ \beta_i \neq 0$$

The p-value for that ANOVA test was 0.065, which is rather close to our significance level $\alpha = 0.05$, but is still not sufficient evidence to say that those two covariates have a significant effect on TV audiences.

To know whether the more 'boring' aspects of roster construction have a significant effect on TV audiences, we tested the following hypotheses in an ANOVA test:

$$H_o: \ \beta_{DNP} = \beta_{Bullpen} \ = 0$$
$$H_a: \ At\ least\ one\ of\ the\ two\ \beta_i \neq 0$$

The p-value for that test was 0.384, which causes us to conclude that the more boring aspects of roster construction do not have a significant effect on the numbers of Detroit TV audiences.

Finally, we utilized our model to find an estimate for the mean TV audience if the Tigers have five All-Stars and a 0.600 winning proportion at the All-Star Break. Assuming there are seven total All-Star players that do not play in the game, and 11 relief pitchers, the average number of TV audience households in that scenario is 359624.4 households with a 95% confidence interval of (138118.6,936367.2). Keeping the number of players who do not play and the number of relief pitchers the same, if the Tigers have one All-Star and a 0.400 winning proportion, the mean TV audience households is found to be 138543.5 households with a 95% confidence interval of (49647.9,386608.7).

Conclusions

The Tigers' All-Star for 2019 (a year that is not included in our data for this analysis) was Shane Greene, RP. It's interesting that a relief pitcher was chosen for the All-Star team that year, since we utilized relief pitchers as a 'boring' aspect of roster construction in this analysis. Jack Morris was the only BYU player to be inducted into the Baseball Hall of Fame and was a 5-time All-Star (4 times as a member of the Detroit Tigers team). Also, the rank of the Detroit DMA (Designated Market Area) was 14 out of 210 in 2019 – according to the Nielsen TV Universe estimates for that year, so our analysis did focus on a relatively large TV market area in the US.

While we did not find statistically significant results in testing the effects of our covariates, there are some considerations to take into mind. First, we did not take interactions between covariates into account in this analysis. It is possible that the winning proportion could have an effect on the number of All-Stars chosen from the Tigers' team (a better playing team could yield more All-Stars). Including that interaction in our model could have changed our results. Other interactions could have been a consideration as well, and there are likely many more factors that effect TV viewership than the four we analyzed. A further analysis might include expanding the covariate selection to improve the model's fit. A final comment is that while we assumed independence between television audience numbers for each respective year, it is possible that there is an element of temporal correlation that we overlooked by making that assumption. A further recommendation would be to investigate whether any significant collinearities exist among our covariates.

R Code Appendix

```
knitr::opts_chunk$set(fig.pos = "!h")
knitr::opts_chunk$set(tidy.opts = list(width.cutoff = 60), tidy = TRUE)
options(scipen = 999)  #suppress scientific notation
## Read in Data
views <- read.csv(file = "C:/Users/Carly/Documents/Grad_School/535/mlbDetroit.csv",
    sep = ",")
library(ggplot2)
library(MASS)
library(lmtest)
library(gridExtra)
library(grid)
library(car)
library(MASS)
library(bestglm)
library(dplyr)
library(kableExtra)
library(gridExtra)
views$Year <- as.factor(views$Year)  #'Year' needs to be a factor
# draw a scatterplot of viewership numbers by the winning
# proportion of the team at the all-star break
g1 <- ggplot(data = views, mapping = aes(x = AtBreak, y = TVaud)) +
    geom_point(size = 1) + geom_smooth(method = "lm", se = FALSE)

# draw a scatterplot of viewership numbers by the number of
# Tigers named all-stars that year
g2 <- ggplot(data = views, mapping = aes(x = Tigers, y = TVaud)) +
    geom_point(size = 1) + geom_smooth(method = "lm", se = FALSE)

# draw a scatterplot of viewership numbers by the number of
# players who are named all-stars but do not play in the game
# (DNP)
g3 <- ggplot(data = views, mapping = aes(x = DNP, y = TVaud)) +
    geom_point(size = 1) + geom_smooth(method = "lm", se = FALSE)

# draw a scatterplot of viewership numbers by number of
# bullpen/relief pitchers
g4 <- ggplot(data = views, mapping = aes(x = Bullpen, y = TVaud)) +
    geom_point(size = 1) + geom_smooth(method = "lm", se = FALSE)

grid.arrange(g1, g2, g3, g4, ncol = 2, nrow = 2)
# Get correlations
corrs <- cor(views[, -1])  #calculate correlations, exclude year (first column)
Correlation <- round(corrs, 2)[1, ]
# we only need the first row (this is correlation of the
# variable w/ TVaud)
Mean <- colMeans(views[, -1]) %>% round(2) * 1000  #per 1k households
SD <- apply(views[, -1], 2, sd) %>% round(2) * 1000
vals <- rbind(Correlation, Mean, SD)

# create table
kable(vals)
```

```r
# fit a MLR
views.lm <- lm(TVaud ~ ., data = views[, -1])
# fit model with log-transform of response
views.loglm <- lm(log(TVaud) ~ ., data = views[, -1])

# histogram of standardized residuals
std.resids <- stdres(views.lm)
ggplot() + geom_histogram(mapping = aes(x = stdres(views.lm)),
    binwidth = 0.2) + labs(x = "Standardized Residuals of TVaud",
    y = "Frequency") + theme(plot.title = element_text(hjust = 0.5))
# residuals appear to have potential right-skew without a
# transformation of tv viewership

# histogram of standardized residuals for log model
ggplot() + geom_histogram(mapping = aes(x = stdres(views.loglm)),
    binwidth = 0.2) + labs(x = "Standardized Residuals log(TVaud)",
    y = "Frequency") + theme(plot.title = element_text(hjust = 0.5))

## create a scatterplot of fitted values vs. residuals
ggplot(data = views, mapping = aes(x = fitted(views.loglm), y = resid(views.loglm))) +
    geom_point() + geom_abline(slope = 0, intercept = 0) + labs(x = "Fitted Values from Model",
    y = "Residuals") + theme(plot.title = element_text(hjust = 0.5))

# get intervals and estimates
intervals <- confint(views.loglm, lvl = 0.95) %>% exp()  #confidence intervals
beta_hats <- summary(views.loglm)$coefficients[, "Estimate"] %>%
    exp()  #estimates

# create table of intervals and estimates
results <- as.data.frame(cbind(beta_hats, intervals)) %>% round(3)
colnames(results) <- c("Estimate", "Lower", "Upper")
kable(results)

# what effect does the number of Tigers named All-Stars have
# on the audience?
t_est <- results["Tigers", "Estimate"] * 1000
t_low <- results["Tigers", "Lower"]
t_up <- results["Tigers", "Upper"]

# what effect does the number of bullpen pitchers have on the
# audience?
b_est <- results["Bullpen", "Estimate"]
b_low <- results["Bullpen", "Lower"]
b_up <- results["Bullpen", "Upper"]

# does how well the tigers are doing have a significant
# effect on the popularity of the game?
full.lm <- views.loglm
reduced.lm <- lm(log(TVaud) ~ DNP + Bullpen, data = views[, -1])
f_test <- anova(full.lm, reduced.lm)
p_val1 <- (f_test$`Pr(>F)` %>% round(3))[2]  #extract p-value

# test Ho: boring aspects of roster construction have no
```

```r
# effect on the Detroit TV audience-
full.lm <- views.loglm
reduced.lm <- lm(log(TVaud) ~ Tigers + AtBreak, data = views[,
    -1])
f_test <- anova(full.lm, reduced.lm)
p_val2 <- (f_test$`Pr(>F)` %>% round(3))[2]  #extract p-value

####### Predictions ############

# Viewership estimates for 'Successful Tigers':
dframe <- data.frame(Tigers = 5, AtBreak = 0.6, DNP = 7, Bullpen = 11)
p1 <- predict.lm(views.loglm, newdata = dframe, interval = "prediction",
    level = 0.95)

# un-transform prediction
real.pred <- exp(p1)[, "fit"] * 1000
real.low <- exp(p1)[, "lwr"] * 1000
real.up <- exp(p1)[, "upr"] * 1000

# Viewership estimates for 'Tanking Tigers':
dframe2 <- data.frame(Tigers = 1, AtBreak = 0.4, DNP = 7, Bullpen = 11)
p2 <- predict.lm(views.loglm, newdata = dframe2, interval = "prediction",
    level = 0.95)

# un-transform prediction
real.pred2 <- exp(p2)[, "fit"] * 1000
real.low2 <- exp(p2)[, "lwr"] * 1000
real.up2 <- exp(p2)[, "upr"] * 1000
knitr::opts_chunk$set(tidy.opts = list(width.cutoff = 70), tidy = TRUE)
```