

# **Final Project Part 3**

Hannah Rosenbaum

## **Background**

Fifty years after the passage of the Title IX Amendment, collegial sports equity has shown relatively minimal change. Allocations of sports budgets often highlight pay discrepancies in participants being “spent [on] \$4,285 per men’s participant versus \$2,588 per women’s participant.” (Feinberg, D., & Hunzinger, E) With these vast differences in individual spending by gender, we see this phenomenon only heightened in the NCAA with women’s basketball. Women’s basketball not only fares having lower budgets from the NCAA but also, per an ESPN report, “is underpaying the NCAA for the tournament rights for 29 championships causing the association to lose out on substantial and crucial revenue... denoting that the current budget of \$81 million to \$112 million multiples more than what the network currently gives.” (Zimbalist) Thus, there is not only a discrepancy in budget allocations among the participants by gender but also amongst large broadcast networks.

Significant systemic issues occur within the gendered branding of ‘March Madness.’ This can be seen with differentiated treatment of male versus female brackets due to the lack of general awareness of when the women’s bracket games even occur. Largely the inequity of the ‘March Madness’ tournament derives from a differentiation from the NCAA in “distribution agreements, corporate sponsorships, distribution of revenue, organizational structure and culture all to prioritize Division I men’s basketball over everything else... to perpetuate gender inequities.” (Blinder) Likewise, this institutional creation of a high investment in TV rights for men’s basketball and minimal airtime for the women’s bracket has led to smaller budgeting and fewer avenues to earn revenue. This has led women’s teams to be “starved of a starring role in the national discourse.” (Blinder) Thus, it creates a circular effect in women’s basketball, deriding fewer resources even within facilities at the NCAA tournament in 2021 and in general awareness of TV times.

I am primarily interested in discussing sports equity in women’s basketball due to my own personal experience at UF of wanting to watch NCAA basketball for women but having no general knowledge of when women play. I believe that the discussion of equity in sports for women is essential because of the common dismissal of watching women’s sports as a pastime.

## **Research Questions**

1. Does the rate of female enrollment in post-secondary education institutions impact the level of female participation in collegiate sports?
2. What is the relationship between total expenditures on collegiate basketball compared to the total ratio of female athletes in college basketball programs?
3. What is the relationship between total revenue allocation in NCAA basketball by gender and the total ratio of females playing college basketball?

## **Hypothesis**

1. Higher rates of female enrollment at post-secondary institutions do not directly affect female participation in college sports. This hypothesis is because there is no direct correlation between registration and participation in sports, as participation in NCAA sports reflects a small sample size.
2. There is a high correlation between expenditure on university sports programs and the percentage of females in university basketball programs by gender. This notion reflects an increased differentiation in total aggregate costs, higher for male than female basketball athletes.
3. There is a high correlation between revenue on university sports programs and the percentage of females in university basketball programs by gender. This notion reflects an increased differentiation in total aggregate revenue, higher for male than female basketball athletes. Thus contemplating the idea that ‘March Madness’ drives profits for male athletes compared to female athletes.

## **Descriptive Statistics**

The Equity in Athletics Disclosures Act requires the full financial disclosure of total expenditures, revenue, staffing, and recruiting efforts by men’s and women’s athletic programs (Mock, J.T.). Data provided by the Equity in Sports project is from all post secondary programs that receive government funding from Title IV funding and is an online database of funding expenses from 2015-2019.

There are 132,327 rows and a total of 28 columns.

### **Feedback from Part 1**

To measure female participation, I will create a model with `sum_partic_women` as the dependent variable and `ef_female_count` as the explanatory variable.

The null data in the `data` matrix exist because a given entry has no male or female participation. The columns with null data are `rev_men`, `rev_women`, `exp_men`, `exp_women`.

### **Feedback from Part 2**

Hypothesis test 1:

- DV: participation of females in sports, basketball -
- Key IV: Ratio of participation of females in sports of female students in attendance -
- Control: percentage of males in sports, basketball -

Hypothesis test 2:

- DV: percentage of females in sports, basketball - I chose this as my dependent variable due to my study wanting to reflect the effect of basketball participation on the total money the university gives to sports.
- Key IV: expenditure of female sport, basketball. This was my independent variable to measure the total effect of expenses on March Madness.
- Control: revenue of female sport, expenditure of male sport, percentage of males, sport type. I used the following as controls due to needing to cross-compare how many males played basketball, and total funding for men for basketball to see the overall effect when using my dependent and independent variables.

Hypothesis test 3:

- DV: percentage of females in sports
- Key IV: revenue of female sport, basketball
- Control: expenditure of female sport, revenue of male sport, percentage of males, sport type

Read in Sports Equity data-set

```
sports <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2023-03-28/sports.csv')

Rows: 132327 Columns: 28
-- Column specification ----
Delimiter: ","
chr (8): institution_name, city_txt, state_cd, zip_text, classification_nam...
dbl (20): year, unitid, classification_code, ef_male_count, ef_female_count, ...
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.
```

```
library(dplyr)
```

```
Attaching package: 'dplyr'
```

```
The following objects are masked from 'package:stats':
```

```
filter, lag
```

```
The following objects are masked from 'package:base':
```

```
intersect, setdiff, setequal, union
```

```
library(wesanderson)
library(ggplot2)
```

**Removing 'Ottawa University-Phoenix' due to having zero total male and female attendance**

```
sports = filter(sports, institution_name != "Ottawa University-Phoenix")
```

**Create data-frames: Critical dimensions, Attendance specific, Basketball specific**

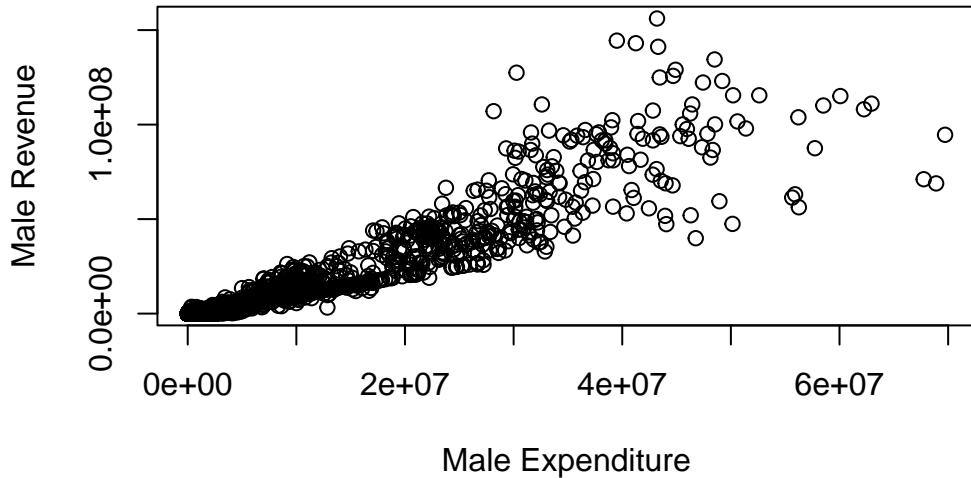
```
data <- as.data.frame(sports[, c("year", "institution_name", "sports", "ef_male_count", "ef_female_count")]
attendance_data <- data[,c("institution_name", "sports", "ef_male_count", "ef_female_count")]

basketball <- as.data.frame(sports[, c("year", "institution_name", "sports", "ef_male_count", "ef_female_count")]
basketball <- filter(basketball, sports=='Basketball')

institute_lbl <- distinct(as.data.frame(data[, c("institution_name")])))
sport_lbl <- distinct(as.data.frame(data[, c("sports")])))
```

## Scatter plots comparing Expenditures against Revenue by Gender

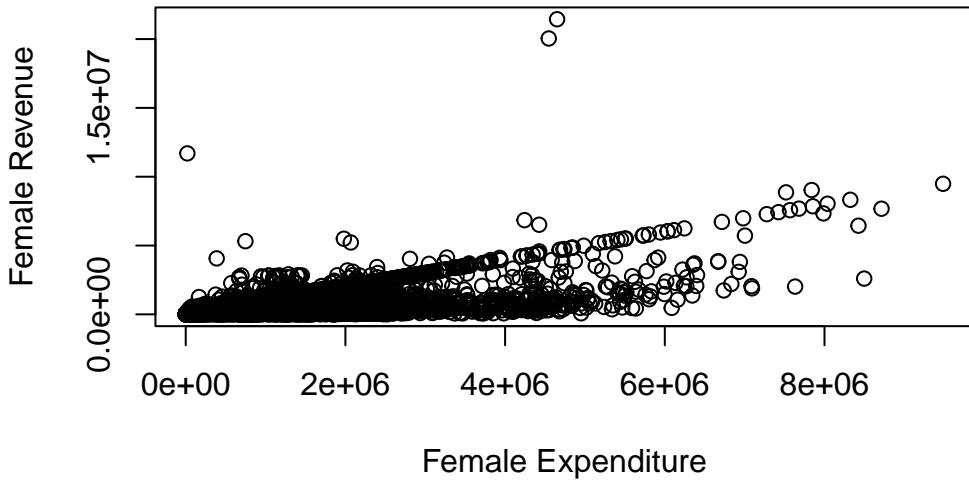
```
#data[is.na(data)] <- 0  
plot(data$exp_men, data$rev_men, xlab="Male Expenditure", ylab="Male Revenue")
```



```
#ggplot(data = data, aes(x=exp_men, y=rev_men), fill = institute_lbl) +  
#geom_point() +  
#scale_fill_manual(values = wes_palette(length(institute_lbl), name = "GrandBudapest1", ty
```

We can see a relationship between revenue and expenditures for men

```
plot(data$exp_women, data$rev_women, xlab="Female Expenditure", ylab="Female Revenue")
```



Similarly, we see a sparser relationship between revenue and expenditures for women.

### Descriptive Statistics

```
glimpse(data)
```

```
Rows: 132,317
Columns: 11
$ year                  <dbl> 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, ~
$ institution_name      <chr> "Alabama A & M University", "Alabama A & M University~
$ sports                 <chr> "Baseball", "Basketball", "All Track Combined", "Foot~
$ ef_male_count         <dbl> 1923, 1923, 1923, 1923, 1923, 1923, 1923, 1923, ~
$ ef_female_count       <dbl> 2300, 2300, 2300, 2300, 2300, 2300, 2300, 2300, ~
$ sum_partic_men        <dbl> 31, 19, 61, 99, 9, 0, 0, 7, 0, 0, 32, 13, 0, 10, 2, 3~
$ sum_partic_women      <dbl> 0, 16, 46, 0, 0, 21, 25, 10, 16, 9, 0, 20, 68, 7, 10, ~
$ rev_men                <dbl> 345592, 1211095, 183333, 2808949, 78270, NA, NA, 7827~
$ rev_women              <dbl> NA, 748833, 315574, NA, NA, 410717, 298164, 131145, 3~
$ exp_men                <dbl> 397818, 817868, 246949, 3059353, 83913, NA, NA, 99612~
$ exp_women              <dbl> NA, 742460, 251184, NA, NA, 432648, 340259, 113886, 3~
```

```
summary(data)
```

year	institution_name	sports	ef_male_count
Min. :2015	Length:132317	Length:132317	Min. : 0
1st Qu.:2016	Class :character	Class :character	1st Qu.: 514
Median :2018	Mode :character	Mode :character	Median : 986
Mean :2018			Mean : 2126
3rd Qu.:2019			3rd Qu.: 2385
Max. :2019			Max. :35954

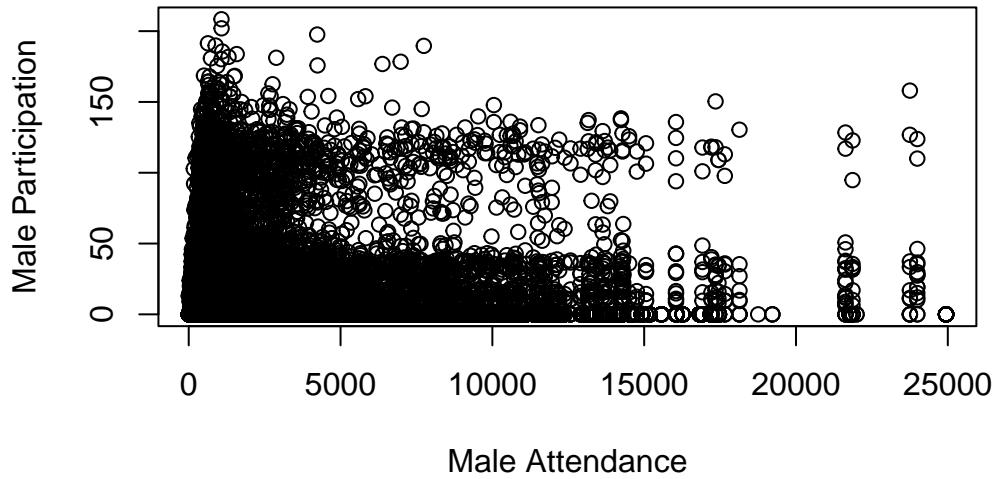
ef_female_count	sum_partic_men	sum_partic_women	rev_men
Min. : 0	Min. : 0.00	Min. : 0.00	Min. : 65
1st Qu.: 652	1st Qu.: 0.00	1st Qu.: 0.00	1st Qu.: 63406
Median : 1249	Median : 0.00	Median : 6.00	Median : 158069
Mean : 2496	Mean : 14.49	Mean : 10.86	Mean : 809028
3rd Qu.: 2860	3rd Qu.: 20.00	3rd Qu.: 17.00	3rd Qu.: 400383
Max. :30325	Max. :331.00	Max. :327.00	Max. :156147208
			NA's :70460

rev_women	exp_men	exp_women
Min. : 0	Min. : 65	Min. : 65
1st Qu.: 58742	1st Qu.: 63049	1st Qu.: 59294
Median : 138292	Median : 159649	Median : 141780
Mean : 279332	Mean : 662384	Mean : 331585
3rd Qu.: 331034	3rd Qu.: 423980	3rd Qu.: 361817
Max. :21440365	Max. :69718059	Max. :9485162
NA's :63441	NA's :70460	NA's :63439

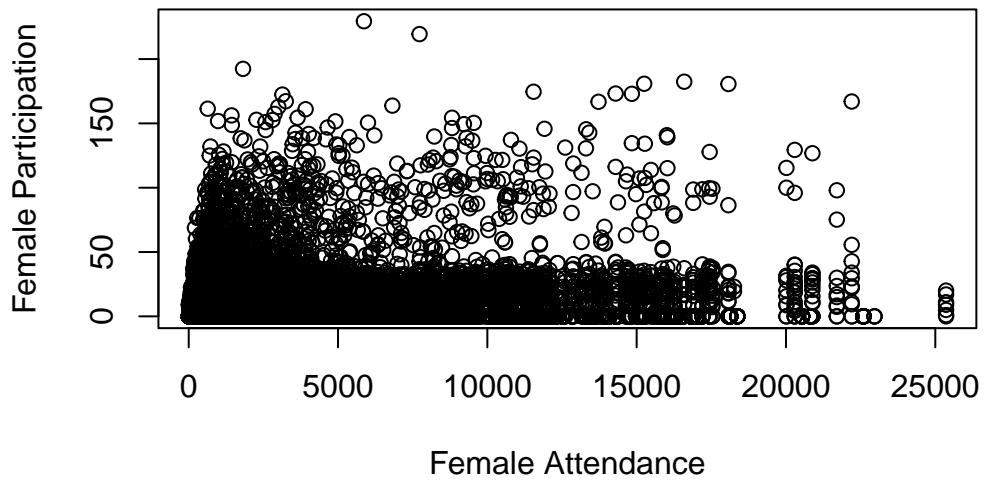
### Scatter plots comparing Institution Attendance against Participation by Gender

```
plot(attendance_data$ef_male_count, attendance_data$sum_partic_men, xlab="Male Attendance"
```



We can see there is a high concentration of schools with male attendance between 0 and 5000, as attendance gets higher the distribution stays consistent.

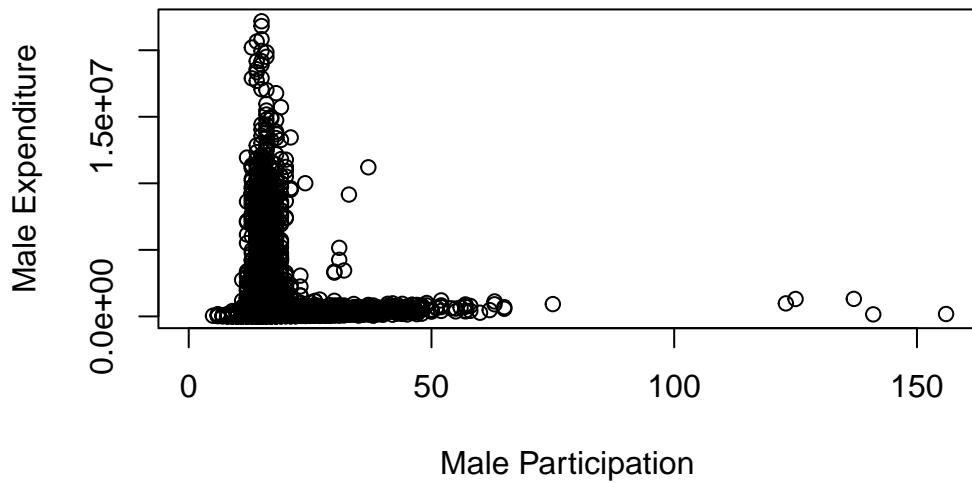
```
plot(attendance_data$ef_female_count, attendance_data$sum_partic_women, xlab="Female Atten
```



Women generally have a lower participation.

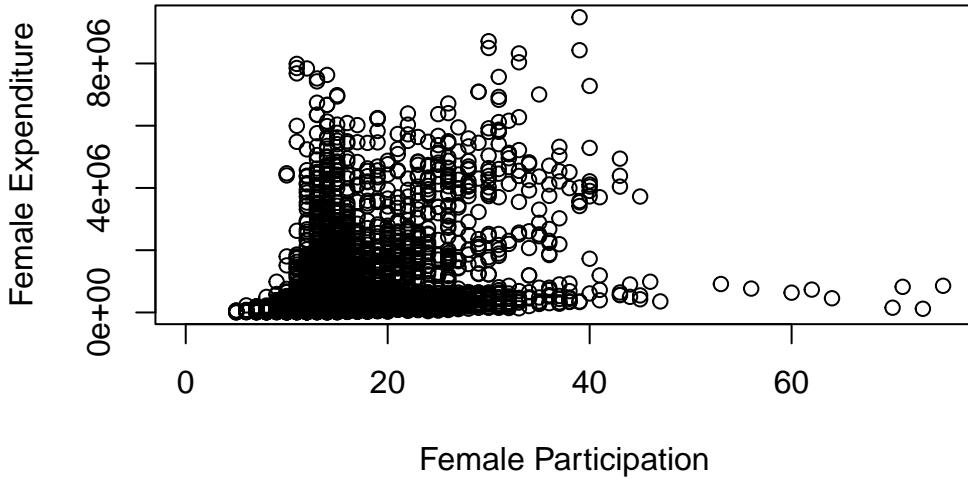
#### Scatter plots comparing basketball Participation against Expenditures by Gender

```
plot(basketball$sum_partic_men, basketball$exp_men, xlab="Male Participation", ylab="Male
```



There is a high concentration of expenditures within schools where team size is smaller.

```
plot(basketball$sum_partic_women, basketball$exp_women, xlab="Female Participation", ylab=
```



There is no real trend between female expenditures and participation.

For the dataset, I could extrapolate my variables of interest as seen here: <<https://github.com>

### Hypothesis Test 1

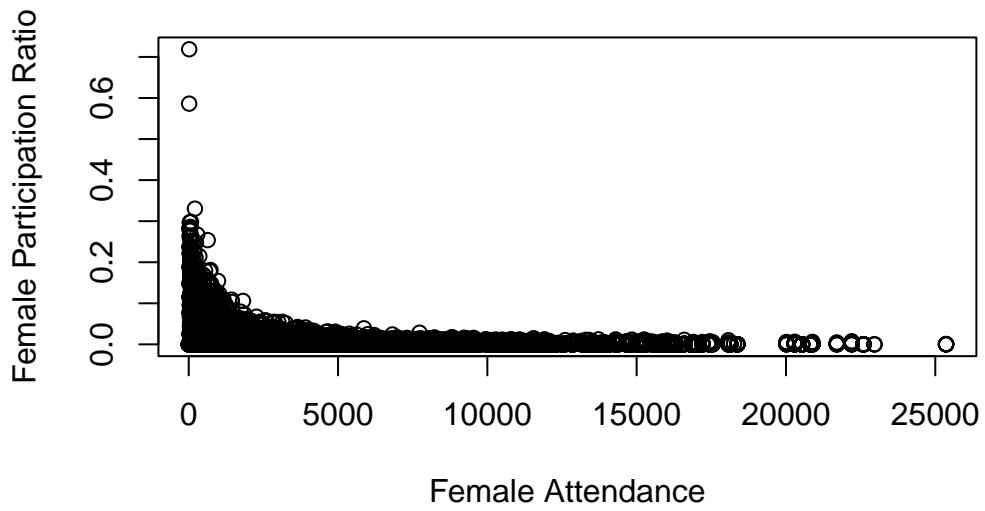
Response variable: sum\_partic\_women

Explanatory variable: sum\_partic\_women / ef\_female\_count

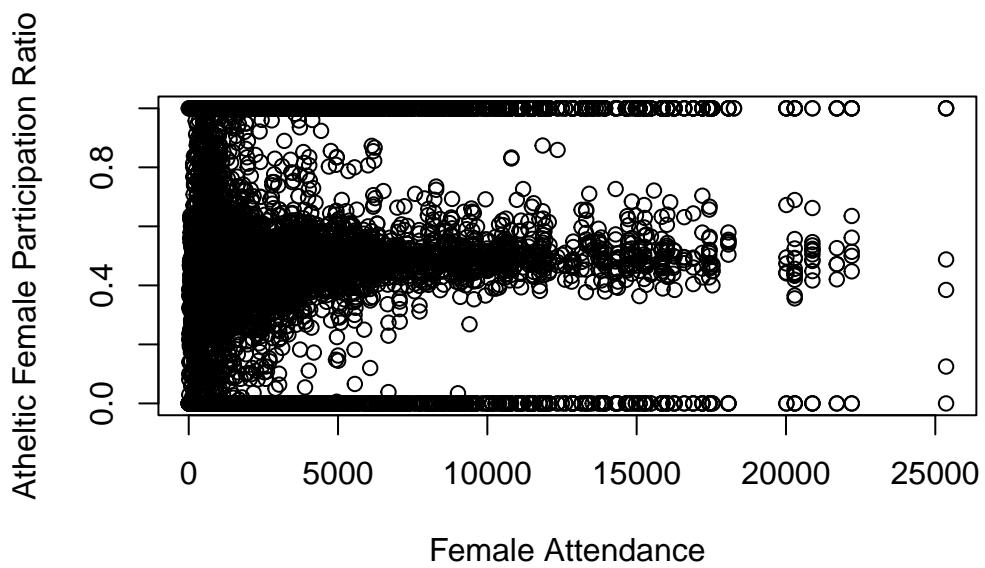
Control variable: sum\_partic\_men

```
attendance_data$female_participation_ratio <- attendance_data$sum_partic_women / attendance_data$ef_female_count
attendance_data$female_athlete_participation_ratio <- attendance_data$sum_partic_women / attendance_data$ef_female_count
attendance_data$male_participation_ratio <- attendance_data$sum_partic_men / attendance_data$ef_male_count

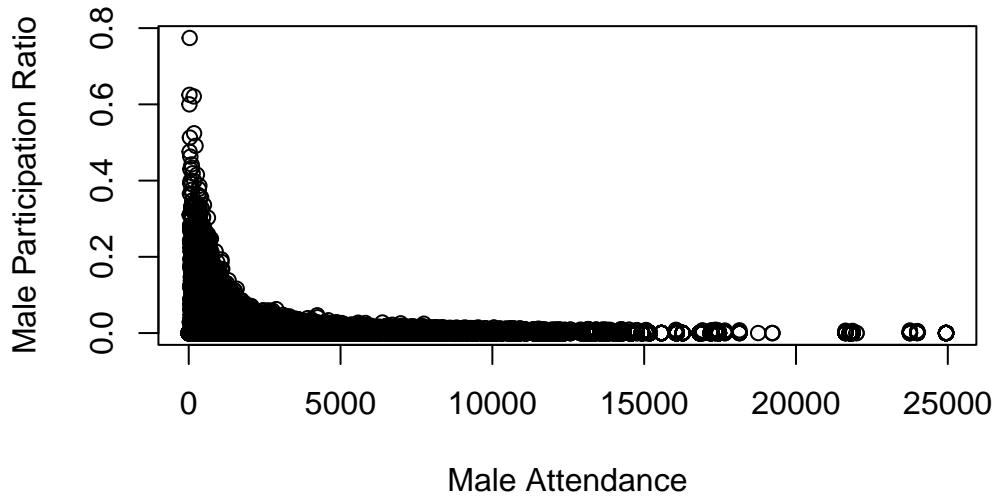
#ggplot(data = attendance_data, aes(x=ef_female_count, y=female_participation_ratio)) + geom_point()
plot(attendance_data$ef_female_count, attendance_data$femail_participation_ratio, xlab="Female Participation Ratio", ylab="Female Expenditure")
```



```
plot(attendance_data$ef_female_count, attendance_data$female_athlete_participation_ratio,
```

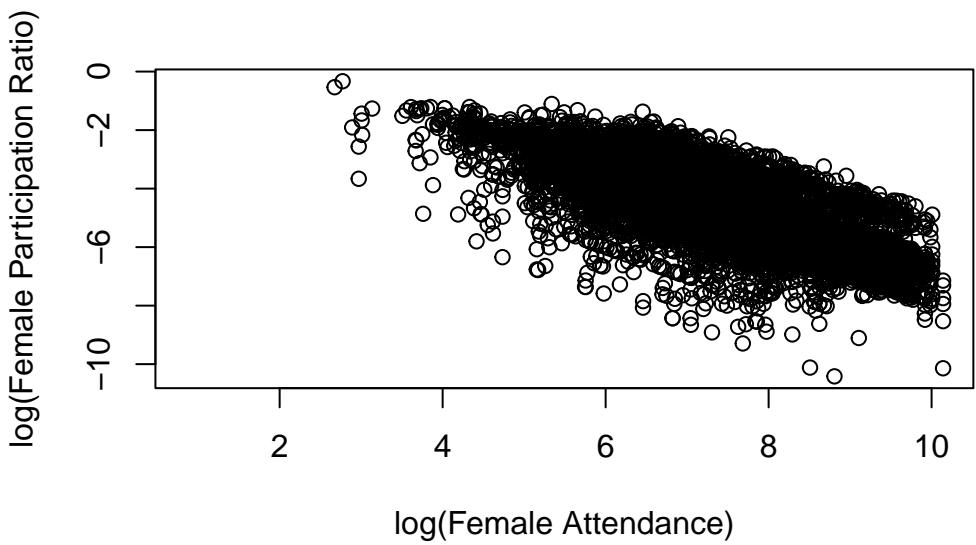


```
plot(attendance_data$ef_male_count, attendance_data$male_participation_ratio, xlab="Male A
```

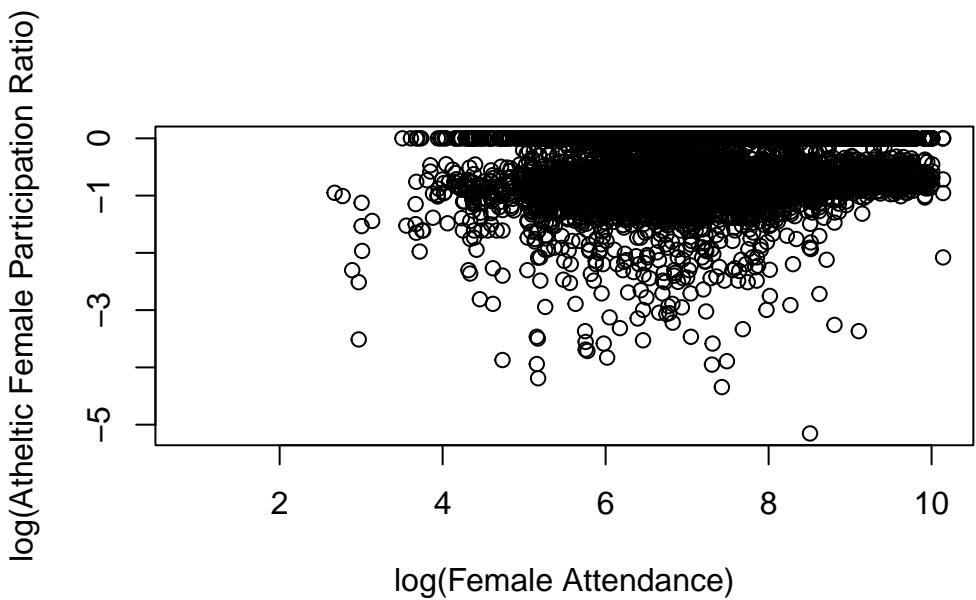


Both plots showing the male/female participation ratio against attendance show a logarithmic pattern.

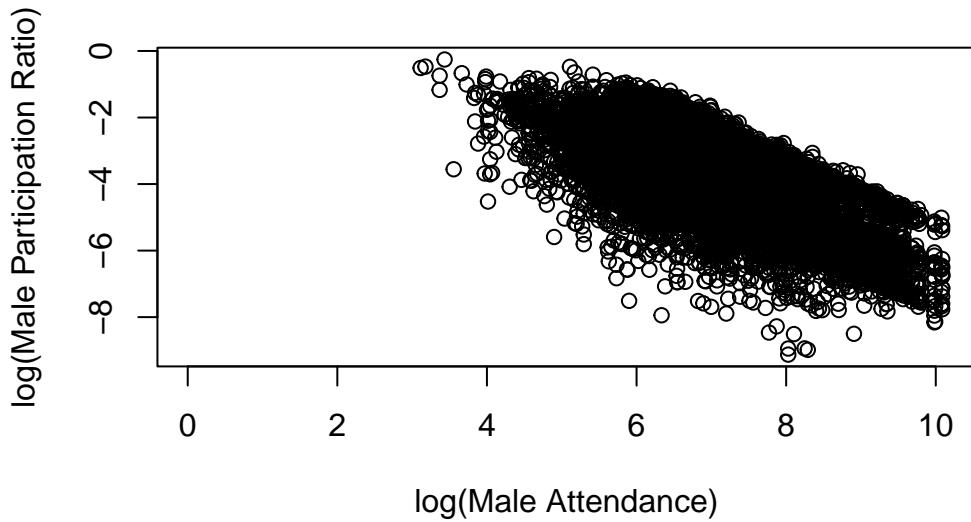
```
plot(log(attendance_data$ef_female_count), log(attendance_data$female_participation_ratio))
```



```
plot(log(attendance_data$ef_female_count), log(attendance_data$female_athlete_participation
```



```
plot(log(attendance_data$ef_male_count), log(attendance_data$male_participation_ratio), xl
```



```
hyp_1_fit_1 <- lm(female_participation_ratio ~ ef_female_count, data = filter(attendance_d
hyp_1_fit_2 <- lm(female_participation_ratio ~ ef_female_count, data = filter(attendance_d
hyp_1_fit_3 <- lm(female_athlete_participation_ratio ~ ef_female_count, data = filter(atte
summary(hyp_1_fit_1)
```

Call:

```
lm(formula = female_participation_ratio ~ ef_female_count, data = filter(attendance_data,
female_participation_ratio != Inf))
```

Residuals:

Min	1Q	Median	3Q	Max
-0.00673	-0.00621	-0.00547	-0.00002	0.71203

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	6.728e-03	7.947e-05	84.66	<2e-16 ***
ef_female_count	-6.975e-07	2.044e-08	-34.13	<2e-16 ***
---				

```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01605 on 63166 degrees of freedom
Multiple R-squared:  0.01811,   Adjusted R-squared:  0.01809
F-statistic:  1165 on 1 and 63166 DF,  p-value: < 2.2e-16
```

```
summary(hyp_1_fit_2)
```

```
Call:
lm(formula = female_participation_ratio ~ ef_female_count, data = filter(attendance_data,
  female_participation_ratio != Inf & sum_partic_women > 0))
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-0.02575	-0.01313	-0.00782	0.00427	0.69150

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.729e-02	2.515e-04	108.52	<2e-16 ***
ef_female_count	-2.868e-06	6.068e-08	-47.26	<2e-16 ***

```
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.02531 on 16114 degrees of freedom
Multiple R-squared:  0.1218,   Adjusted R-squared:  0.1217
F-statistic:  2234 on 1 and 16114 DF,  p-value: < 2.2e-16
```

```
summary(hyp_1_fit_3)
```

```
Call:
```

```
lm(formula = female_athlete_participation_ratio ~ ef_female_count,
  data = filter(attendance_data, female_athlete_participation_ratio !=
  Inf & sum_partic_women > 0))
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-0.7416	-0.1988	-0.1326	0.3447	0.3774

```

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.226e-01 2.643e-03 235.54 <2e-16 ***
ef_female_count 9.622e-06 6.378e-07 15.09 <2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Residual standard error: 0.266 on 16115 degrees of freedom  
 Multiple R-squared: 0.01393, Adjusted R-squared: 0.01387  
 F-statistic: 227.6 on 1 and 16115 DF, p-value: < 2.2e-16

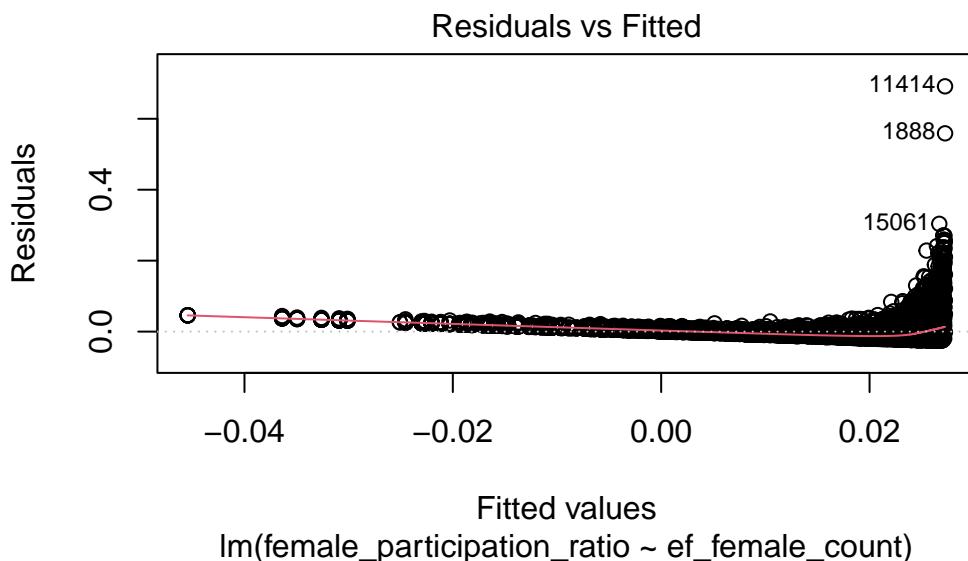
```
AIC(hyp_1_fit_2)
```

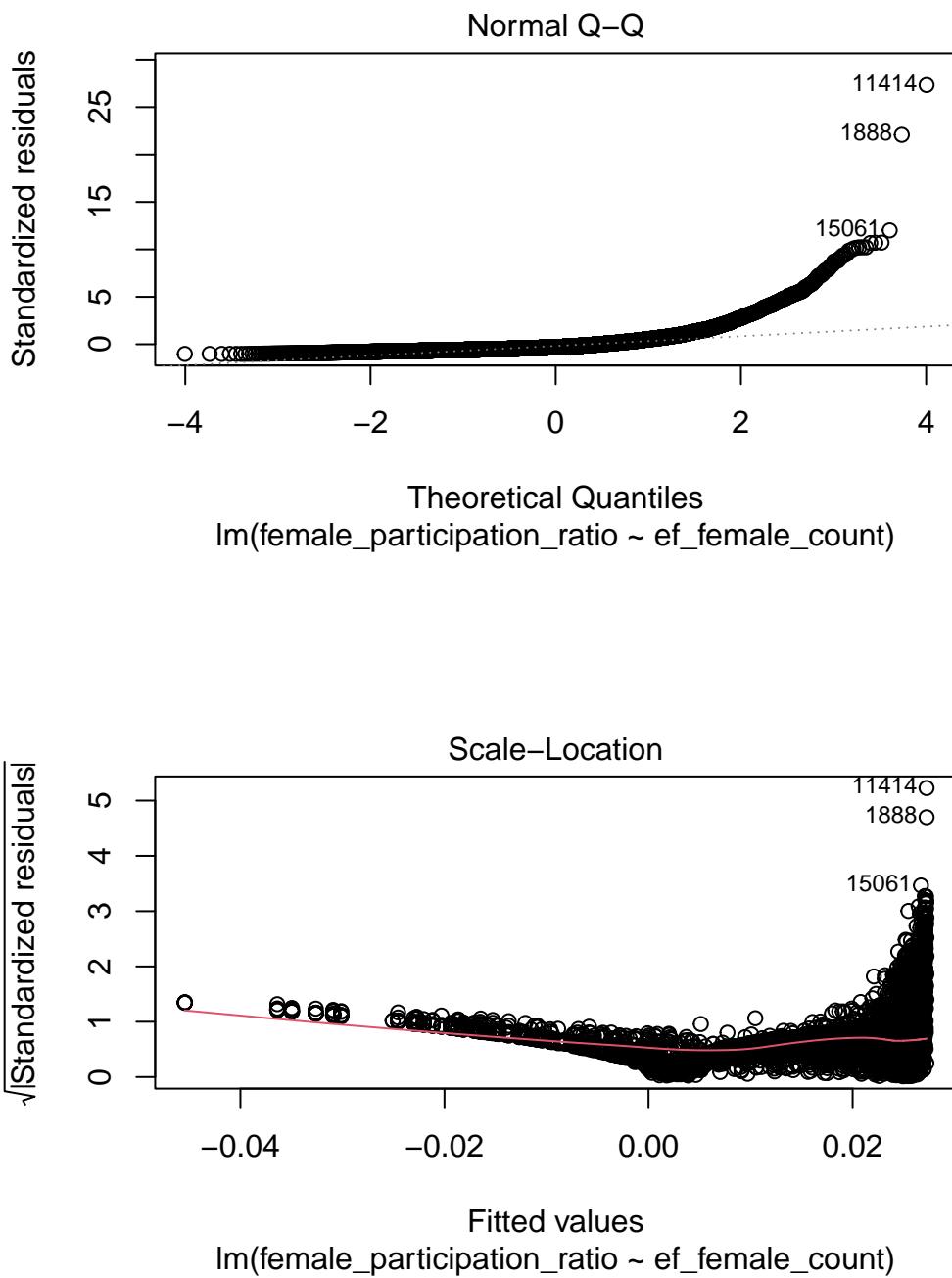
```
[1] -72766.08
```

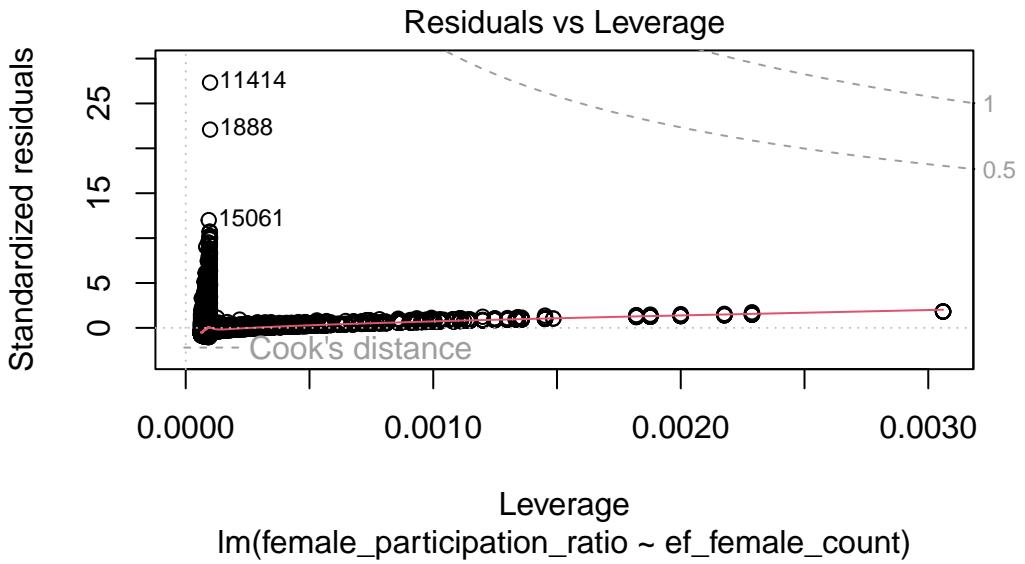
```
BIC(hyp_1_fit_2)
```

```
[1] -72743.02
```

```
plot(hyp_1_fit_2)
```







```
hyp_1_fit_4 <- lm(log(female_participation_ratio) ~ log(ef_female_count), data = filter(atte
```

```
summary(hyp_1_fit_4)
```

Call:

```
lm(formula = log(female_participation_ratio) ~ log(ef_female_count),
  data = filter(attendance_data, female_participation_ratio != Inf & sum_partic_women > 0 & ef_female_count > 0))
```

Residuals:

Min	1Q	Median	3Q	Max
-4.5862	-0.3587	0.0456	0.3977	2.4977

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.523962	0.037858	40.25	<2e-16 ***
log(ef_female_count)	-0.835062	0.005171	-161.48	<2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

```
Residual standard error: 0.7151 on 16114 degrees of freedom
Multiple R-squared:  0.6181,   Adjusted R-squared:  0.618
F-statistic: 2.608e+04 on 1 and 16114 DF,  p-value: < 2.2e-16
```

```
summary(hyp_1_fit_5)
```

```
Call:
lm(formula = log(female_athlete_participation_ratio) ~ log(ef_female_count),
    data = filter(attendance_data, female_athlete_participation_ratio !=
        Inf & female_athlete_participation_ratio > 0 & ef_female_count >
        0))
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-4.7085	-0.2825	-0.1368	0.4709	0.7736

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-1.004208	0.023609	-42.54	<2e-16 ***
log(ef_female_count)	0.065751	0.003225	20.39	<2e-16 ***
---				
Signif. codes:	0 ****	0.001 **	0.01 *	0.05 .
	' '	' '	' '	' '

```
Residual standard error: 0.4459 on 16114 degrees of freedom
Multiple R-squared:  0.02515,   Adjusted R-squared:  0.02509
F-statistic: 415.7 on 1 and 16114 DF,  p-value: < 2.2e-16
```

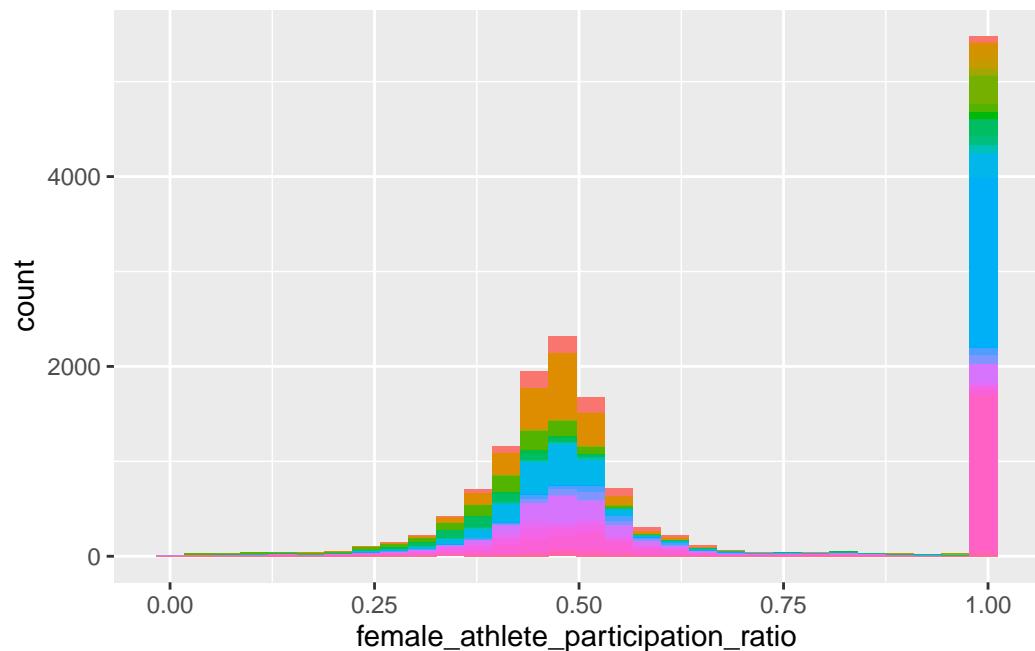
```
AIC(hyp_1_fit_4)
```

```
[1] 34930.1
```

```
BIC(hyp_1_fit_4)
```

```
[1] 34953.16
```

```
#hist(filter(attendance_data, female_athlete_participation_ratio != Inf & sum_partic_women
ggplot(data = filter(attendance_data, female_athlete_participation_ratio != Inf & sum_partic_women
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Female athletic participation has a normal distribution with an outlying spike on the right end.

## Hypothesis Test 2

Response variable: exp\_women

Explanatory variable: sum\_partic\_women / ef\_female\_count

Control variable: exp\_men

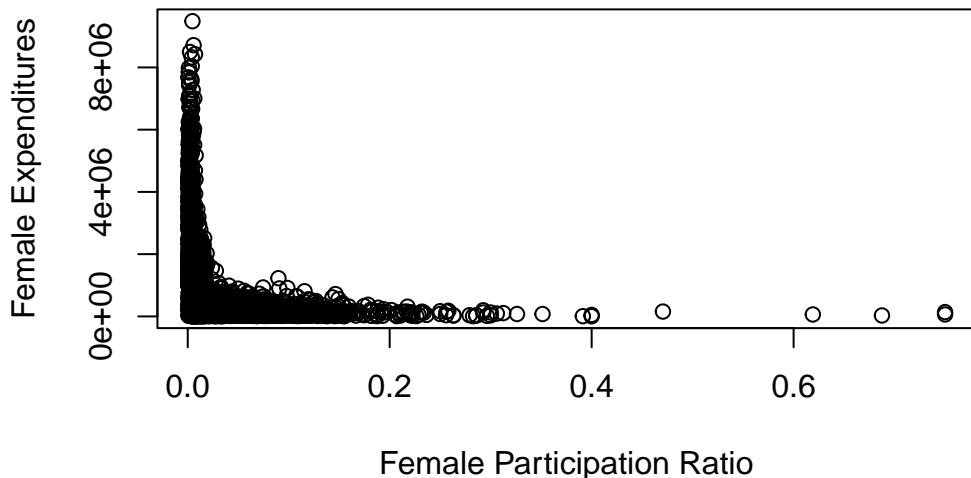
```
basketball$female_participation_ratio <- basketball$sum_partic_women / basketball$ef_female_count
basketball$female_athlete_participation_ratio <- basketball$sum_partic_women / (basketball$sum_partic_women + basketball$sum_partic_men)
basketball$male_participation_ratio <- basketball$sum_partic_men / basketball$ef_male_count
```

### Transform basketball table to separate men and women by column

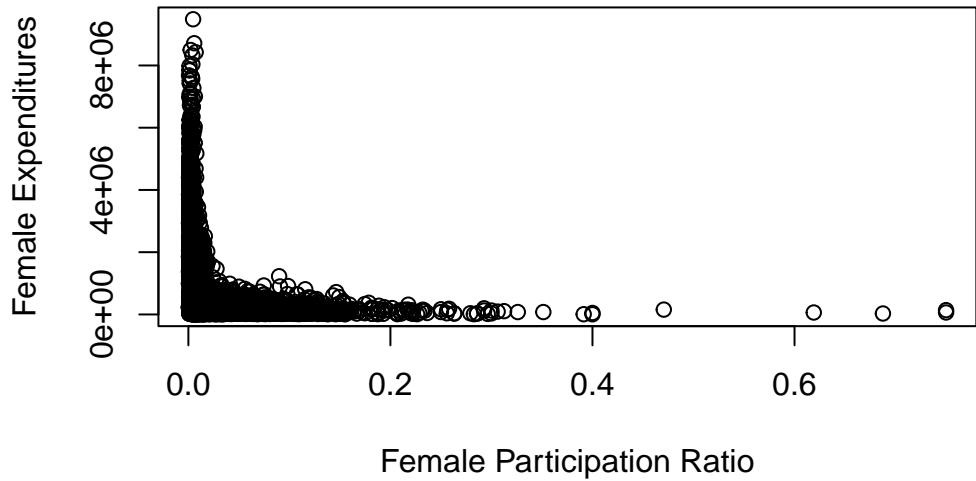
```
female <- as.data.frame(basketball[, c("year", "institution_name", "sports", "ef_female_count")]
female$gender <- "Female"
female <- female %>% rename("ef_count"="ef_female_count", "sum_partic"="sum_partic_women",

male <- as.data.frame(basketball[, c("year", "institution_name", "sports", "ef_male_count")]
male$gender <- "Male"
male <- male %>% rename("ef_count"="ef_male_count", "sum_partic"="sum_partic_men", "rev"="")
basketball_hist <- rbind(male, female)
basketball_hist <- filter(basketball_hist, sum_partic > 0)

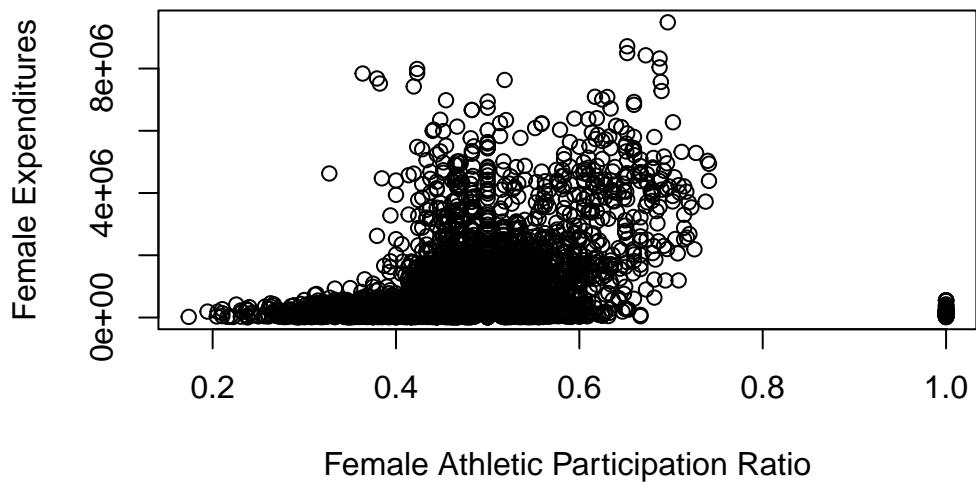
plot(basketball$female_participation_ratio, basketball$exp_women, xlab="Female Participation Ratio", ylab="Female Expenditures")
```



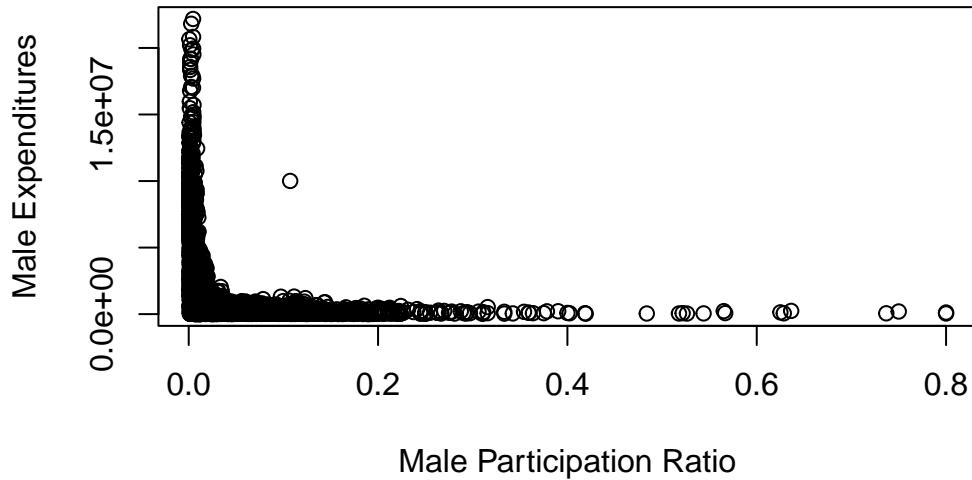
```
plot(filter(basketball, sum_partic_women > 0)$female_participation_ratio, filter(basketball, sum_partic_women > 0)$exp_women, xlab="Female Participation Ratio", ylab="Female Expenditures")
```



```
plot(filter(basketball, sum_partic_women > 0)$female_athlete_participation_ratio, filter(b
```

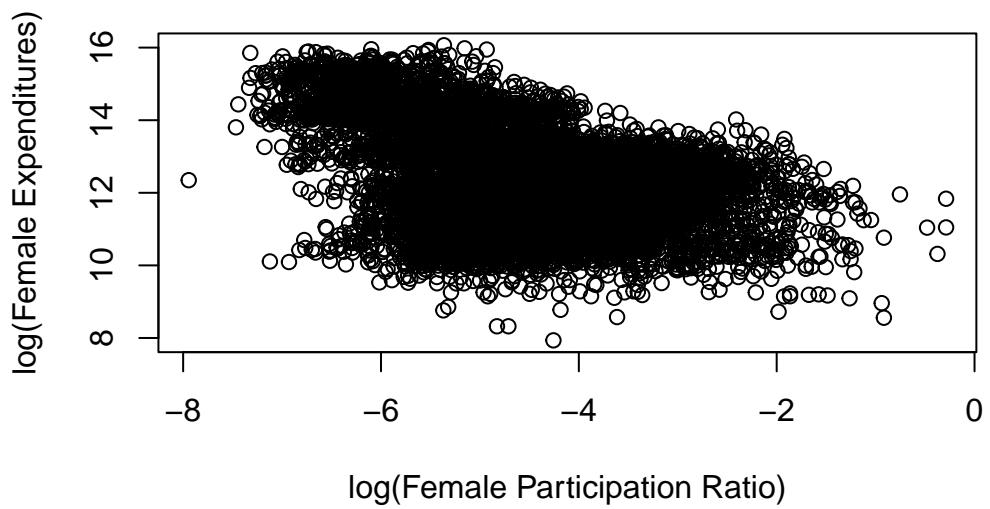


```
plot(basketball$male_participation_ratio, basketball$exp_men, xlab="Male Participation Ratio", ylab="Male Expenditures")
```

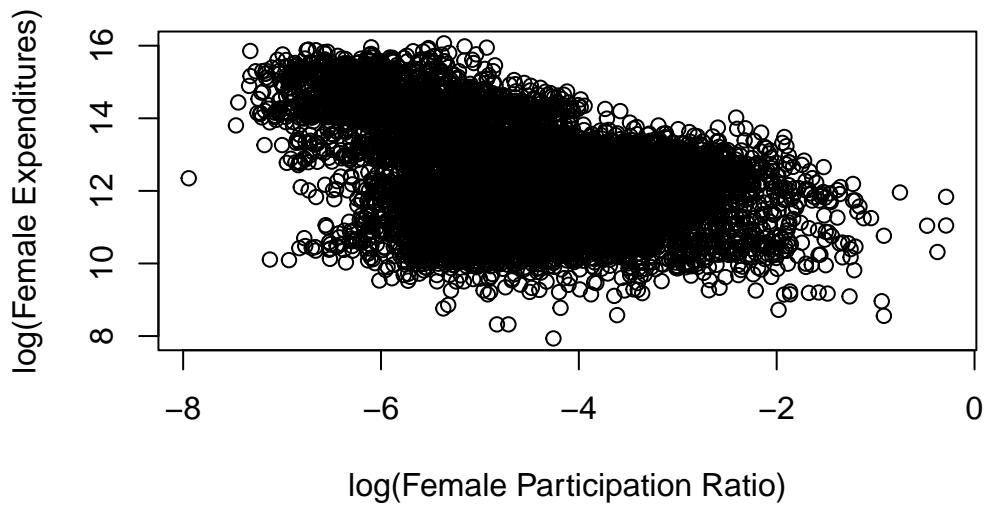


The plot showing female athlete participation against expenditures has a very slight trend, but the other three plots show a strong logarithmic pattern.

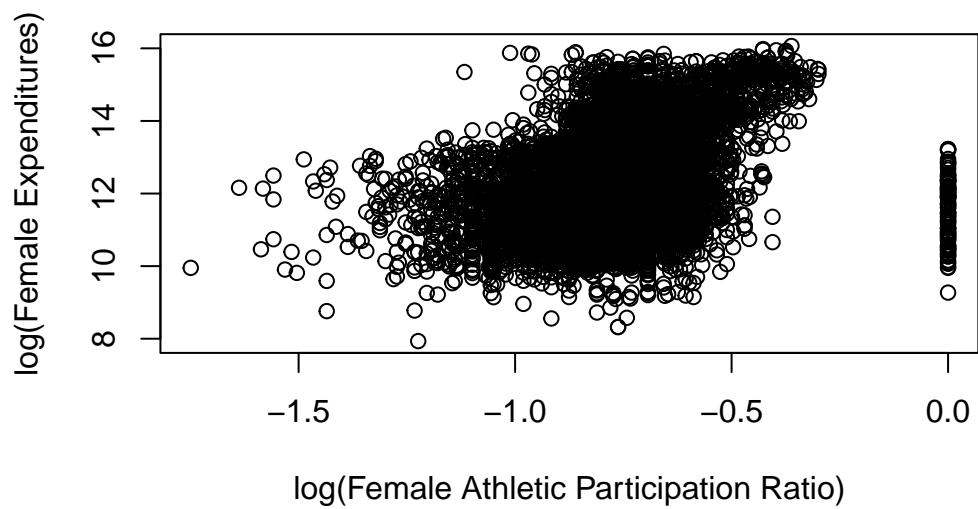
```
plot(log(basketball$female_participation_ratio), log(basketball$exp_women), xlab="log(Female Participation Ratio)", ylab="log(Male Expenditures)")
```



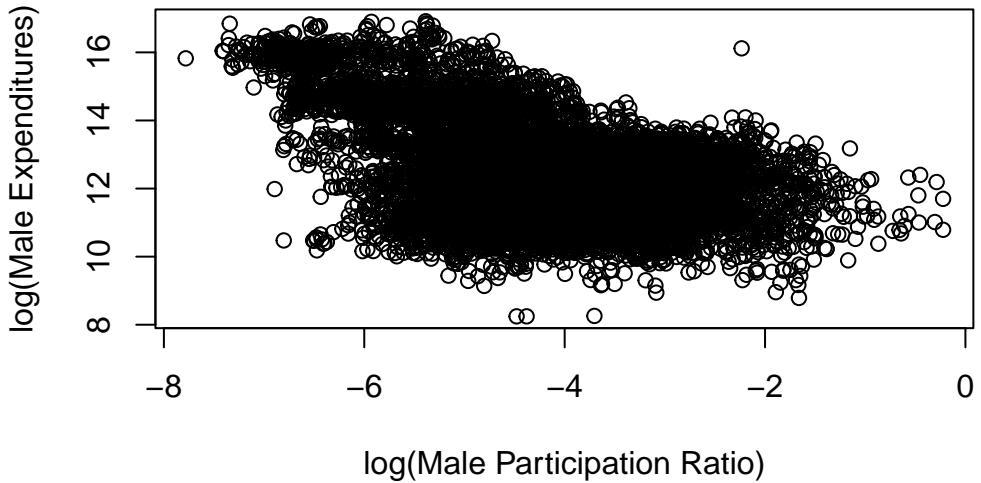
```
plot(log(filter(basketball, sum_partic_women > 0)$female_participation_ratio), log(filter(
```



```
plot(log(filter(basketball, sum_partic_women > 0)$female_athlete_participation_ratio), log
```



```
plot(log(basketball$male_participation_ratio), log(basketball$exp_men), xlab="log(Male Par
```



```

hyp_2_fit_1 <- lm(exp_women ~ female_participation_ratio, data = filter(basketball, female
hyp_2_fit_2 <- lm(exp_women ~ female_participation_ratio, data = filter(basketball, female
hyp_2_fit_3 <- lm(exp_women ~ female_athlete_participation_ratio, data = filter(basketball
summary(hyp_2_fit_1)

```

Call:

```
lm(formula = exp_women ~ female_participation_ratio, data = filter(basketball,
female_participation_ratio != Inf))
```

Residuals:

Min	1Q	Median	3Q	Max
-669197	-493134	-320286	10208	8815838

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	697788	11619	60.06	<2e-16 ***
female_participation_ratio	-6077440	300720	-20.21	<2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 958400 on 9552 degrees of freedom

```
(439 observations deleted due to missingness)
Multiple R-squared:  0.04101,   Adjusted R-squared:  0.0409
F-statistic: 408.4 on 1 and 9552 DF,  p-value: < 2.2e-16
```

```
summary(hyp_2_fit_2)
```

```
Call:
lm(formula = exp_women ~ female_participation_ratio, data = filter(basketball,
  female_participation_ratio != Inf & sum_partic_women > 0))
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-669197	-493134	-320286	10208	8815838

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	697788	11619	60.06	<2e-16 ***
female_participation_ratio	-6077440	300720	-20.21	<2e-16 ***
---				

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

```
Residual standard error: 958400 on 9552 degrees of freedom
Multiple R-squared:  0.04101,   Adjusted R-squared:  0.0409
F-statistic: 408.4 on 1 and 9552 DF,  p-value: < 2.2e-16
```

```
summary(hyp_2_fit_3)
```

```
Call:
lm(formula = exp_women ~ female_athlete_participation_ratio,
  data = filter(basketball, female_participation_ratio != Inf &
  sum_partic_women > 0))
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-2056112	-437273	-268066	15346	8284247

```
Coefficients:
```

Estimate	Std. Error	t value	Pr(> t )
----------	------------	---------	----------

```

(Intercept) -785373      53101   -14.79    <2e-16 ***
female_athlete_participation_ratio 2852106     109722    25.99    <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 945800 on 9552 degrees of freedom
Multiple R-squared:  0.06606,   Adjusted R-squared:  0.06597
F-statistic: 675.7 on 1 and 9552 DF,  p-value: < 2.2e-16

```

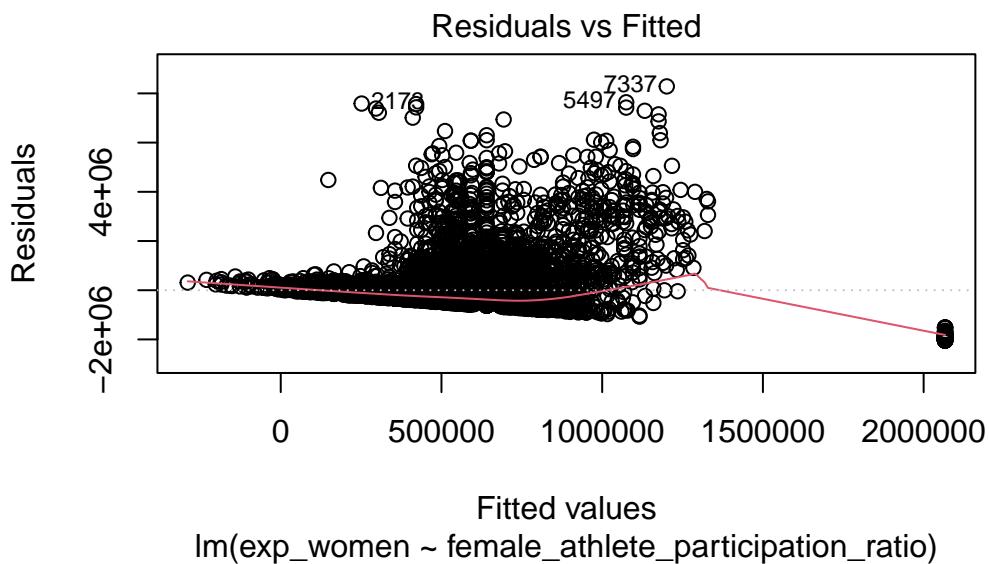
```
AIC(hyp_2_fit_3)
```

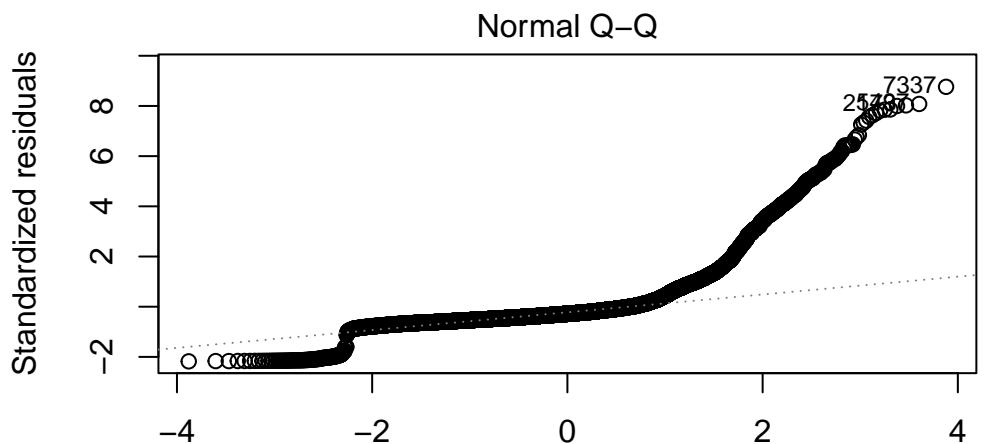
```
[1] 290039
```

```
BIC(hyp_2_fit_3)
```

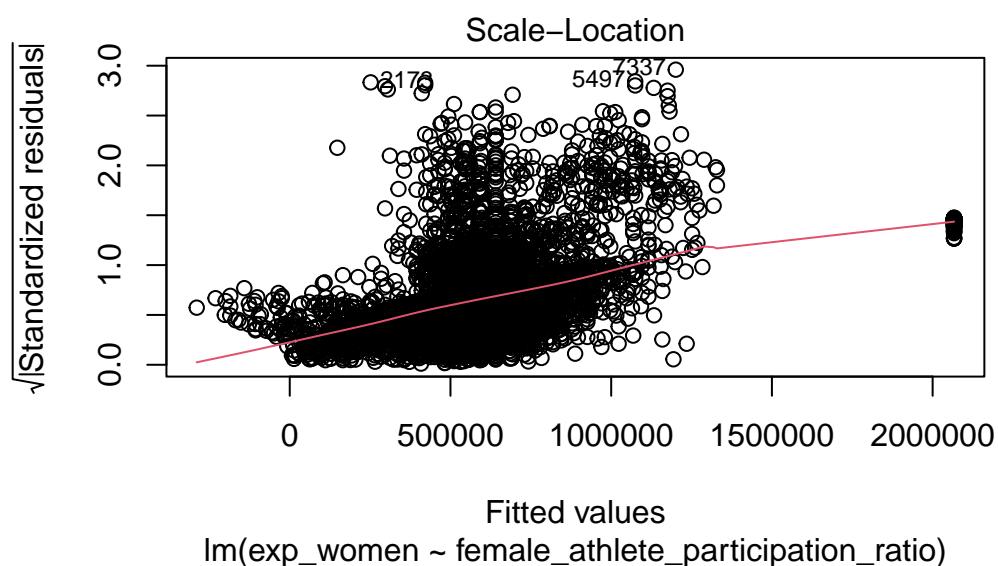
```
[1] 290060.5
```

```
plot(hyp_2_fit_3)
```

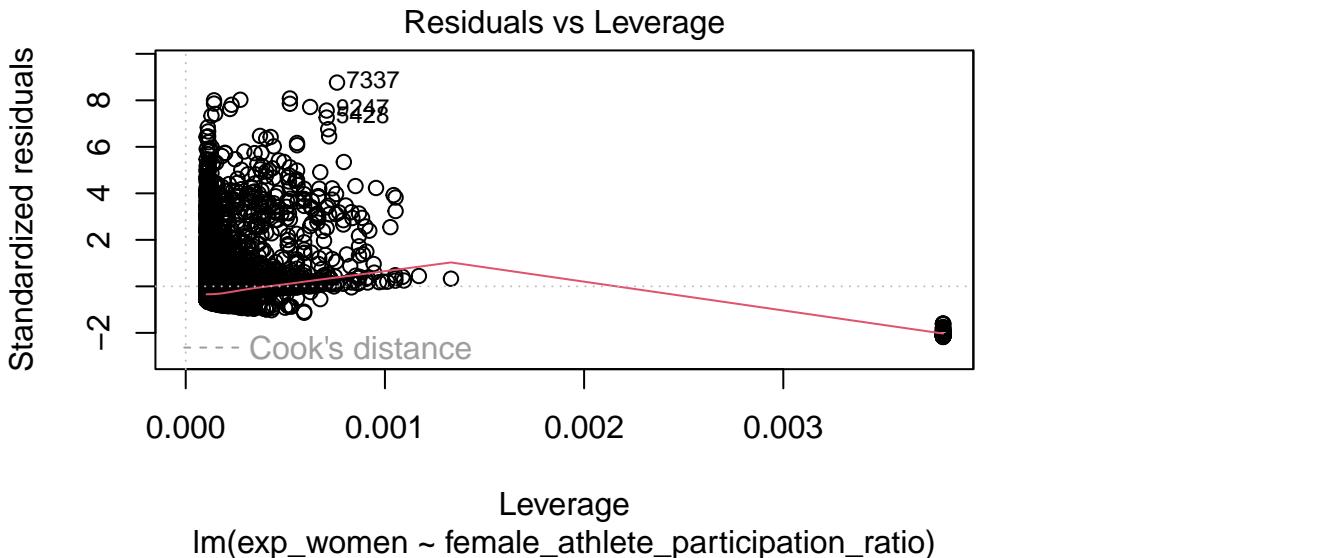




Theoretical Quantiles  
lm(exp\_women ~ female\_athlete\_participation\_ratio)



Fitted values  
lm(exp\_women ~ female\_athlete\_participation\_ratio)



```
# Controls
hyp_2_fit_c1 <- lm(exp_women ~ rev_women, data = filter(basketball, female_participation_ratio != Inf))
hyp_2_fit_c2 <- lm(exp_women ~ exp_men, data = filter(basketball, female_participation_ratio != Inf))
hyp_2_fit_c3 <- lm(exp_women ~ sum_partic_men, data = filter(basketball, female_participation_ratio != Inf))

summary(hyp_2_fit_c1)
```

```
Call:
lm(formula = exp_women ~ rev_women, data = filter(basketball,
female_participation_ratio != Inf))

Residuals:
    Min      1Q      Median      3Q      Max 
-16165268 -106752   -102785    -91699   5879334 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.084e+05  7.487e+03   14.48   <2e-16 ***
rev_women   9.658e-01  8.169e-03   118.23   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 623500 on 9552 degrees of freedom  
(439 observations deleted due to missingness)  
Multiple R-squared:  0.5941,    Adjusted R-squared:  0.594  
F-statistic: 1.398e+04 on 1 and 9552 DF,  p-value: < 2.2e-16
```

```
summary(hyp_2_fit_c2)
```

```
Call:  
lm(formula = exp_women ~ exp_men, data = filter(basketball, female_participation_ratio !=  
Inf))  
  
Residuals:  
    Min      1Q  Median      3Q      Max  
-5084097 -131233 -77806   69567  5039235  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 1.715e+05  4.216e+03   40.67  <2e-16 ***  
exp_men     4.335e-01  1.836e-03   236.13  <2e-16 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 374200 on 9439 degrees of freedom  
(552 observations deleted due to missingness)  
Multiple R-squared:  0.8552,    Adjusted R-squared:  0.8552  
F-statistic: 5.576e+04 on 1 and 9439 DF,  p-value: < 2.2e-16
```

```
summary(hyp_2_fit_c3)
```

```
Call:  
lm(formula = exp_women ~ sum_partic_men, data = filter(basketball,  
female_participation_ratio != Inf))  
  
Residuals:  
    Min      1Q  Median      3Q      Max  
-638434 -478799 -364507 -39436  8914009
```

```

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 649055    27559   23.552 < 2e-16 ***
sum_partic_men -4582      1523  -3.008 0.00264 **
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Residual standard error: 978200 on 9552 degrees of freedom  
(439 observations deleted due to missingness)  
Multiple R-squared: 0.0009464, Adjusted R-squared: 0.0008418  
F-statistic: 9.049 on 1 and 9552 DF, p-value: 0.002635

```

# Logarithmic
hyp_2_fit_4 <- lm(log(exp_women) ~ log(female_participation_ratio), data = filter(basketba
hyp_2_fit_5 <- lm(log(exp_women) ~ log(female_athlete_participation_ratio), data = filter(
summary(hyp_2_fit_4)

```

```

Call:
lm(formula = log(exp_women) ~ log(female_participation_ratio),
  data = filter(basketball, female_participation_ratio != Inf &
  sum_partic_women > 0 & exp_women > 0))

```

```

Residuals:
    Min      1Q  Median      3Q     Max
-4.3169 -0.8131  0.0467  0.9068  3.3624

```

```

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.14651    0.05119 198.21 <2e-16 ***
log(female_participation_ratio) -0.49438    0.01107 -44.65 <2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Residual standard error: 1.187 on 9552 degrees of freedom  
Multiple R-squared: 0.1726, Adjusted R-squared: 0.1726  
F-statistic: 1993 on 1 and 9552 DF, p-value: < 2.2e-16

```
summary(hyp_2_fit_5)
```

```
Call:  
lm(formula = log(exp_women) ~ log(female_athlete_participation_ratio),  
  data = filter(basketball, female_participation_ratio != Inf &  
  sum_partic_women > 0 & exp_women > 0))
```

Residuals:

Min	1Q	Median	3Q	Max
-4.8916	-0.8459	-0.0923	0.8256	4.1097

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	14.16222	0.05788	244.7	<2e-16
log(female_athlete_participation_ratio)	2.36977	0.07452	31.8	<2e-16
(Intercept)	***			
log(female_athlete_participation_ratio)	***			
---				
Signif. codes:	0 ***	0.001 **	0.01 *	0.05 .
	''	'	'	'

Residual standard error: 1.241 on 9552 degrees of freedom

Multiple R-squared: 0.09572, Adjusted R-squared: 0.09563

F-statistic: 1011 on 1 and 9552 DF, p-value: < 2.2e-16

```
AIC(hyp_2_fit_4)
```

```
[1] 30396.6
```

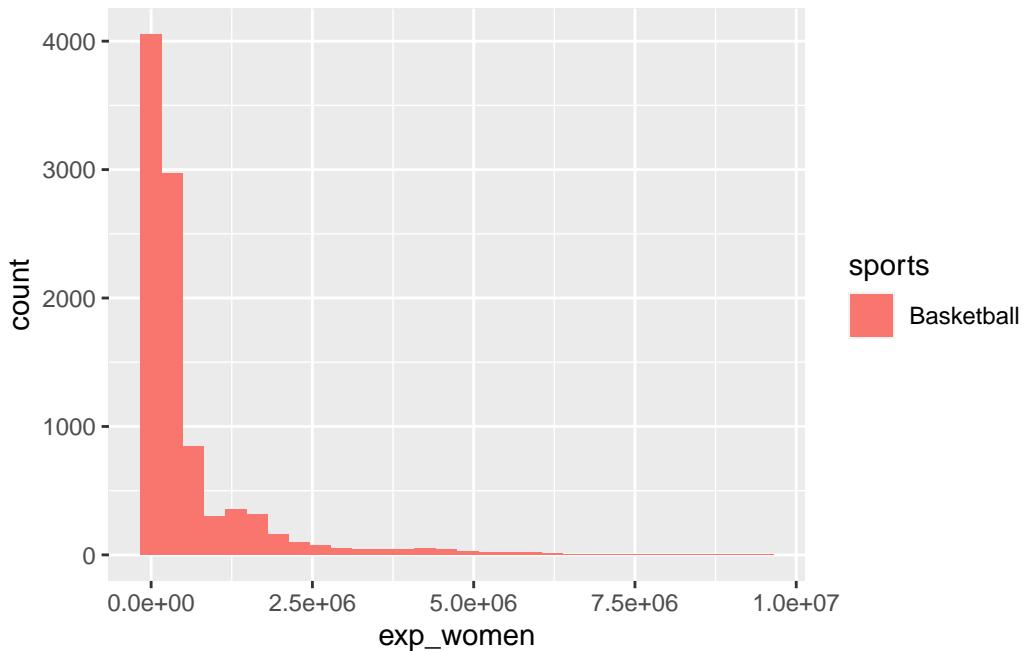
```
BIC(hyp_2_fit_4)
```

```
[1] 30418.09
```

```
#hist(filter(basketball, female_athlete_participation_ratio != Inf & sum_partic_women > 0)
```

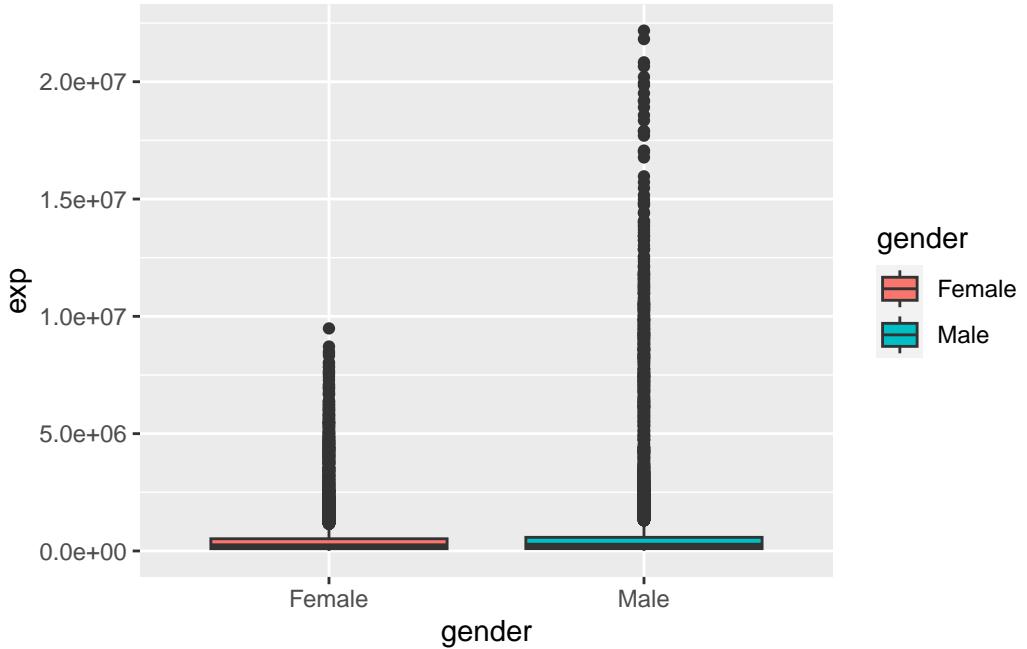
```
ggplot(data=filter(basketball, female_athlete_participation_ratio != Inf & sum_partic_wome
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The female participation histogram within basketball has a heavy concentration toward lower values.

```
#boxplot(exp ~ gender, data=basketball_hist, ylab="Expenditures")  
ggplot(data=basketball_hist, aes(x=gender, y=exp, fill=gender)) + geom_boxplot()
```



Within basketball women and men have a similar mean expenditure, but at a height men receive nearly twice the funds as women.

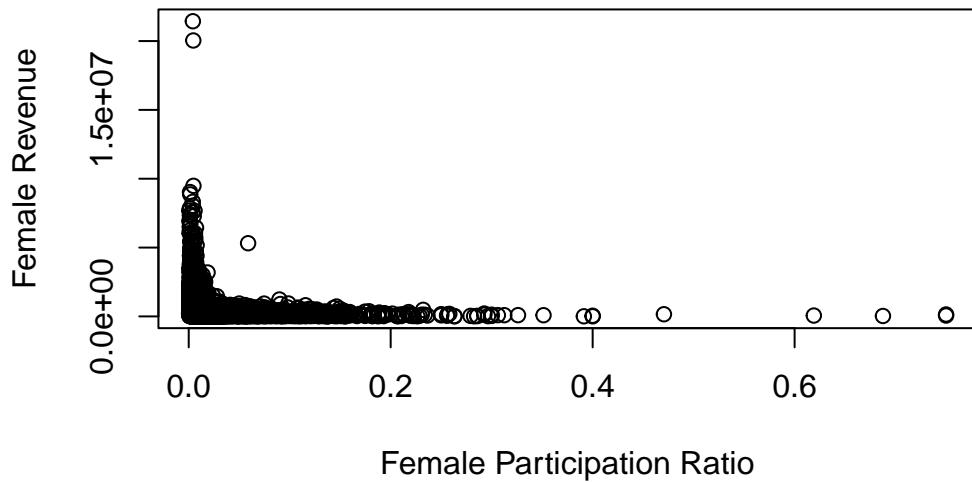
### Hypothesis Test 3

Response variable: rev\_women

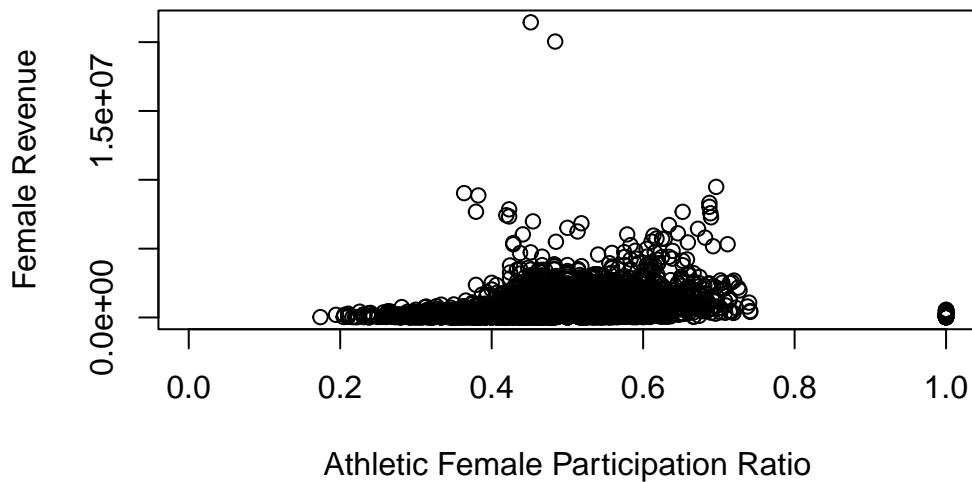
Explanatory variable: sum\_partic\_women / ef\_female\_count

Control variable: rev\_men

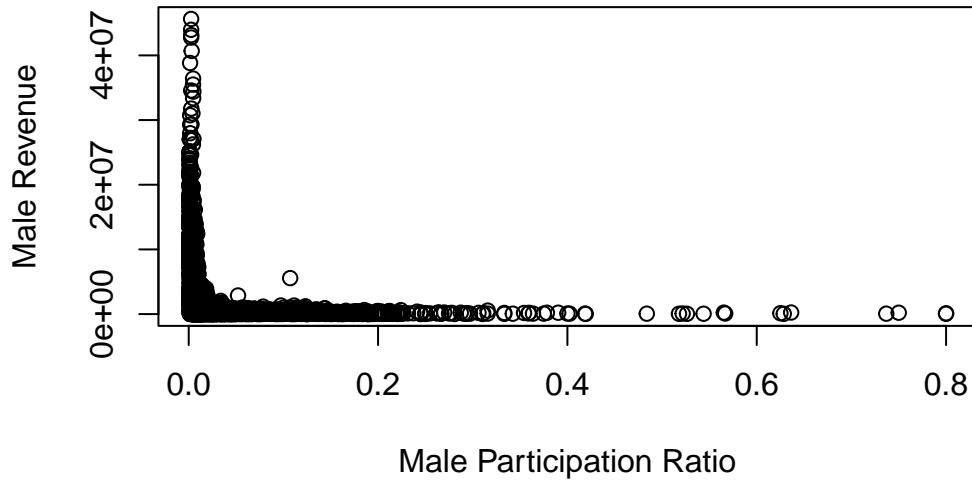
```
plot(basketball$female_participation_ratio, basketball$rev_women, xlab="Female Participation Ratio")
```



```
plot(basketball$female_athlete_participation_ratio, basketball$rev_women, xlab="Athletic F
```

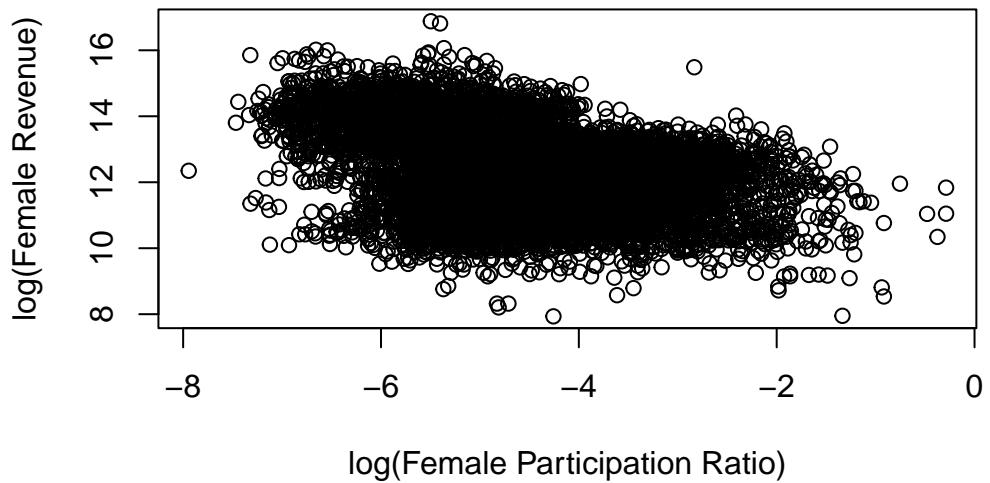


```
plot(basketball$male_participation_ratio, basketball$rev_men, xlab="Male Participation Ratio", ylab="Male Revenue")
```

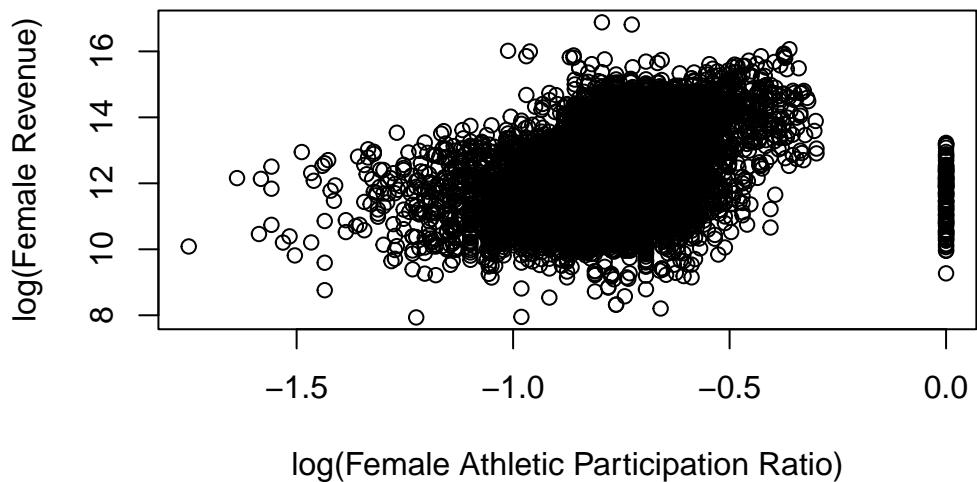


Similar to the plots in hypothesis tests one and two, female athletic participation plotted against revenue has a very light trend; but revenue against female participation/female attendance shows a strong logarithmic pattern.

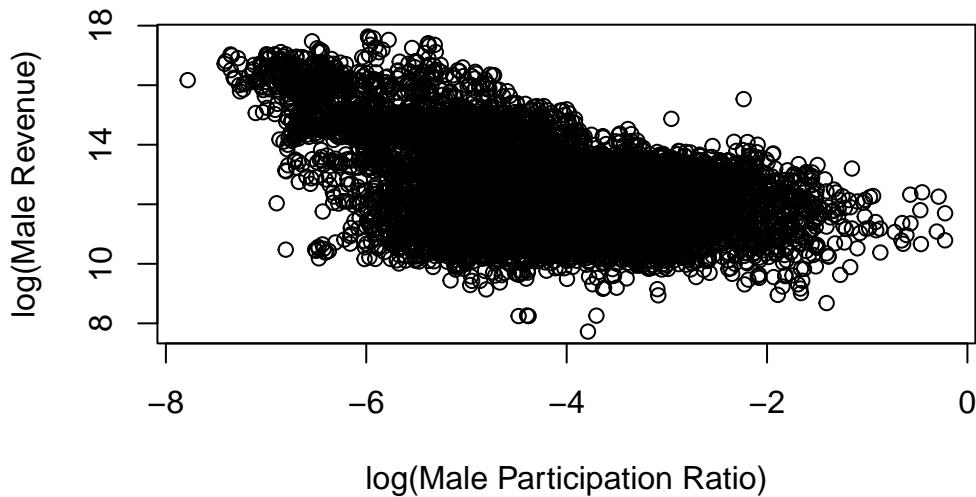
```
plot(log(basketball$female_participation_ratio), log(basketball$rev_women), xlab="log(Female Participation Ratio)", ylab="log(Male Revenue)")
```



```
plot(log(basketball$female_athlete_participation_ratio), log(basketball$rev_women), xlab="
```



```
plot(log(basketball$male_participation_ratio), log(basketball$rev_men), xlab="log(Male Par
```



```
hyp_3_fit_1 <- lm(rev_women ~ female_participation_ratio, data = filter(basketball, female
hyp_3_fit_2 <- lm(rev_women ~ female_participation_ratio, data = filter(basketball, female
hyp_3_fit_3 <- lm(rev_women ~ female_athlete_participation_ratio, data = filter(basketball
summary(hyp_3_fit_1)
```

Call:

```
lm(formula = rev_women ~ female_participation_ratio, data = filter(basketball,
female_participation_ratio != Inf))
```

Residuals:

Min	1Q	Median	3Q	Max
-546362	-388437	-240863	59195	20887546

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	570845	9308	61.33	<2e-16 ***
female_participation_ratio	-4393071	240902	-18.24	<2e-16 ***
---				

```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 767800 on 9552 degrees of freedom
(439 observations deleted due to missingness)
Multiple R-squared: 0.03364, Adjusted R-squared: 0.03354
F-statistic: 332.5 on 1 and 9552 DF, p-value: < 2.2e-16
```

```
summary(hyp_3_fit_2)
```

```
Call:
lm(formula = rev_women ~ female_participation_ratio, data = filter(basketball,
female_participation_ratio != Inf & sum_partic_women > 0))
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-546362	-388437	-240863	59195	20887546

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	570845	9308	61.33	<2e-16 ***
female_participation_ratio	-4393071	240902	-18.24	<2e-16 ***

```
---
```

```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 767800 on 9552 degrees of freedom
Multiple R-squared: 0.03364, Adjusted R-squared: 0.03354
F-statistic: 332.5 on 1 and 9552 DF, p-value: < 2.2e-16
```

```
summary(hyp_3_fit_3)
```

```
Call:
```

```
lm(formula = rev_women ~ female_athlete_participation_ratio,
data = filter(basketball, female_participation_ratio != Inf))
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-1407754	-348702	-220354	43044	21003989

```

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -372330     42943   -8.67 <2e-16 ***
female_athlete_participation_ratio 1790705     88733   20.18 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 764900 on 9552 degrees of freedom
(439 observations deleted due to missingness)
Multiple R-squared:  0.04089, Adjusted R-squared:  0.04079
F-statistic: 407.3 on 1 and 9552 DF, p-value: < 2.2e-16

```

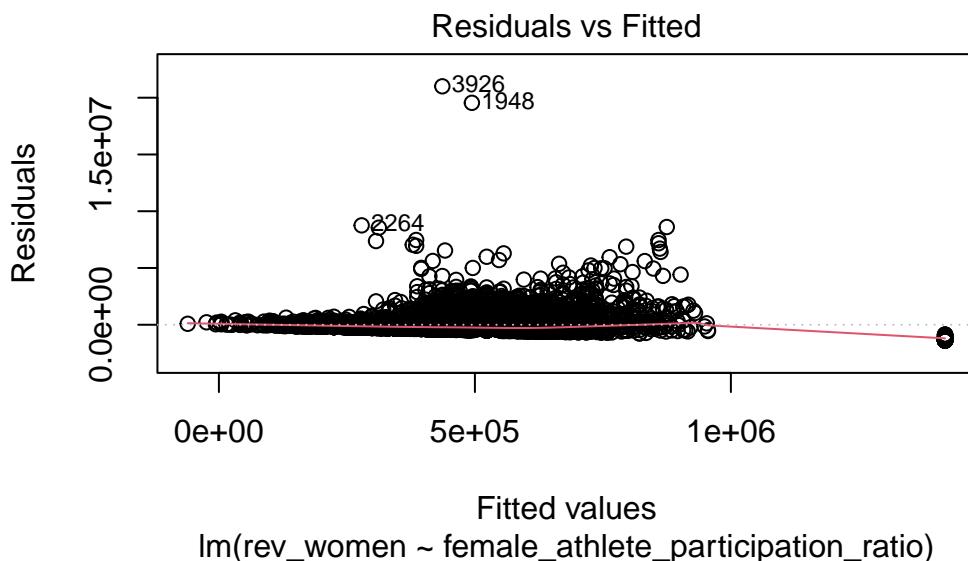
```
AIC(hyp_3_fit_3)
```

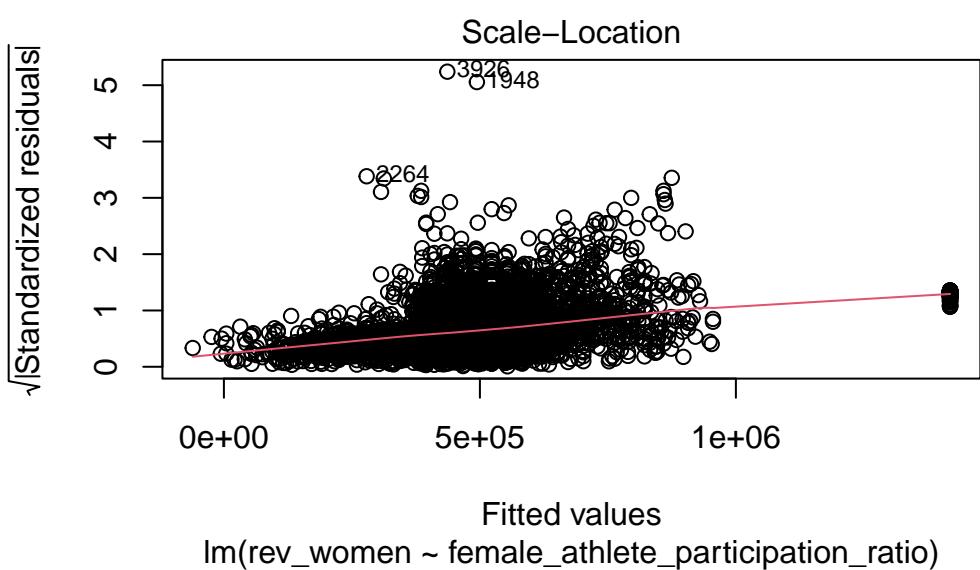
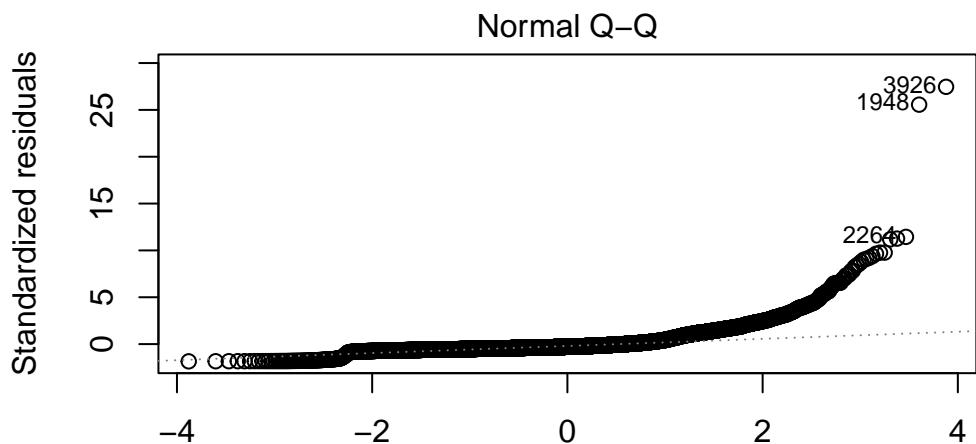
```
[1] 285982.1
```

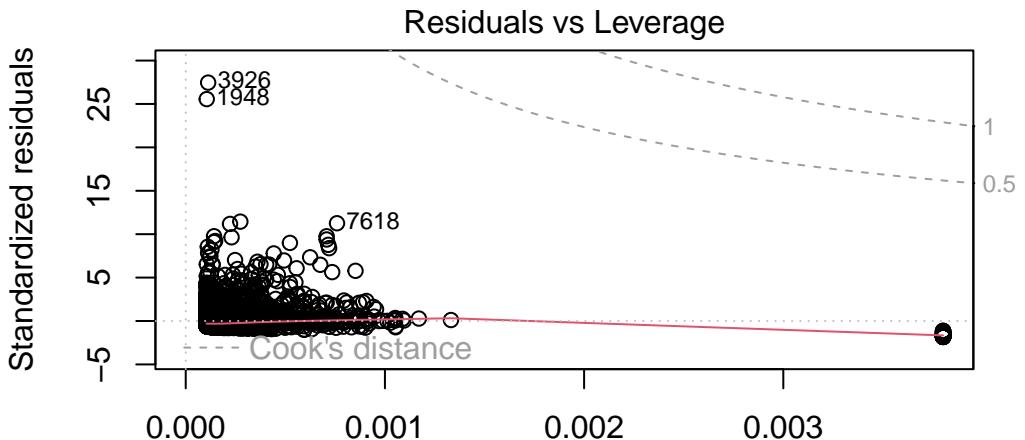
```
BIC(hyp_3_fit_3)
```

```
[1] 286003.6
```

```
plot(hyp_3_fit_3)
```







Leverage  
 $\text{lm}(\text{rev\_women} \sim \text{female\_athlete\_participation\_ratio})$

```
# Controls
hyp_3_fit_c1 <- lm(rev_women ~ exp_women, data = filter(basketball, female_participation_ratio != Inf))
hyp_3_fit_c2 <- lm(rev_women ~ rev_men, data = filter(basketball, female_participation_ratio != Inf))
hyp_3_fit_c3 <- lm(rev_women ~ sum_partic_men, data = filter(basketball, female_participation_ratio != Inf))

summary(hyp_3_fit_c1)
```

```
Call:
lm(formula = rev_women ~ exp_women, data = filter(basketball,
    female_participation_ratio != Inf))

Residuals:
    Min      1Q  Median      3Q     Max 
-3402605 -94839  -54643   56058 18451560 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.281e+05  5.896e+03   21.72   <2e-16 ***
exp_women   6.151e-01  5.202e-03  118.23   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 497600 on 9552 degrees of freedom  
(439 observations deleted due to missingness)  
Multiple R-squared:  0.5941,    Adjusted R-squared:  0.594  
F-statistic: 1.398e+04 on 1 and 9552 DF,  p-value: < 2.2e-16
```

```
summary(hyp_3_fit_c2)
```

```
Call:  
lm(formula = rev_women ~ rev_men, data = filter(basketball, female_participation_ratio !=  
Inf))  
  
Residuals:  
    Min      1Q  Median      3Q      Max  
-5310089 -259684 -173394    73537 20160179  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 3.337e+05 7.284e+03  45.81  <2e-16 ***  
rev_men     1.358e-01  2.239e-03   60.63  <2e-16 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 665700 on 9439 degrees of freedom  
(552 observations deleted due to missingness)  
Multiple R-squared:  0.2803,    Adjusted R-squared:  0.2802  
F-statistic:  3676 on 1 and 9439 DF,  p-value: < 2.2e-16
```

```
summary(hyp_3_fit_c3)
```

```
Call:  
lm(formula = rev_women ~ sum_partic_men, data = filter(basketball,  
female_participation_ratio != Inf))  
  
Residuals:  
    Min      1Q  Median      3Q      Max  
-507695 -384110 -274416    32257 20960910
```

```

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 518316    21999   23.56 <2e-16 ***
sum_partic_men -2286      1216   -1.88  0.0602 .
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Residual standard error: 780900 on 9552 degrees of freedom  
(439 observations deleted due to missingness)  
Multiple R-squared: 0.0003698, Adjusted R-squared: 0.0002652  
F-statistic: 3.534 on 1 and 9552 DF, p-value: 0.06016

```

hyp_3_fit_4 <- lm(log(rev_women) ~ log(female_participation_ratio), data = filter(basketball,
hyp_3_fit_5 <- lm(log(rev_women) ~ log(female_athlete_participation_ratio), data = filter(
summary(hyp_3_fit_4)

```

```

Call:
lm(formula = log(rev_women) ~ log(female_participation_ratio),
  data = filter(basketball, female_participation_ratio != Inf &
  sum_partic_women > 0 & rev_women > 0))

```

```

Residuals:
    Min      1Q      Median      3Q      Max
-4.2932 -0.7712  0.0441  0.8770  4.1381

```

```

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.45678    0.04920 212.54 <2e-16 ***
log(female_participation_ratio) -0.41592    0.01064 -39.08 <2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Residual standard error: 1.141 on 9552 degrees of freedom  
Multiple R-squared: 0.1379, Adjusted R-squared: 0.1378  
F-statistic: 1527 on 1 and 9552 DF, p-value: < 2.2e-16

```

summary(hyp_3_fit_5)

```

```
Call:  
lm(formula = log(rev_women) ~ log(female_athlete_participation_ratio),  
  data = filter(basketball, female_participation_ratio != Inf &  
  rev_women > 0))
```

Residuals:

Min	1Q	Median	3Q	Max
-4.6351	-0.7967	-0.0590	0.8164	4.6339

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	13.90566	0.05485	253.53	<2e-16
log(female_athlete_participation_ratio)	2.08667	0.07063	29.55	<2e-16

(Intercept) \*\*\*  
log(female\_athlete\_participation\_ratio) \*\*\*

---

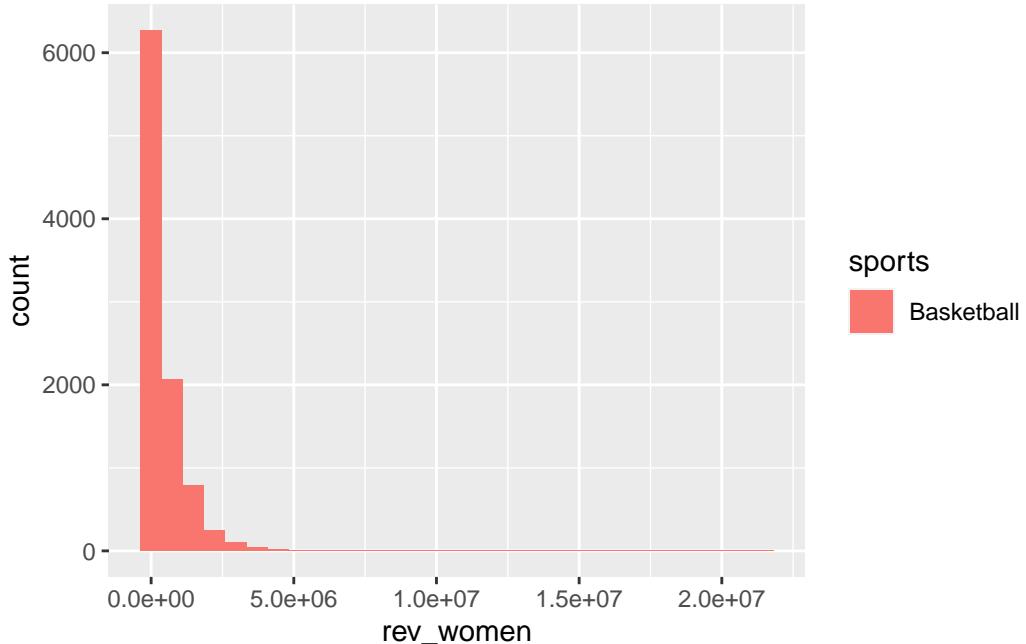
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.176 on 9552 degrees of freedom

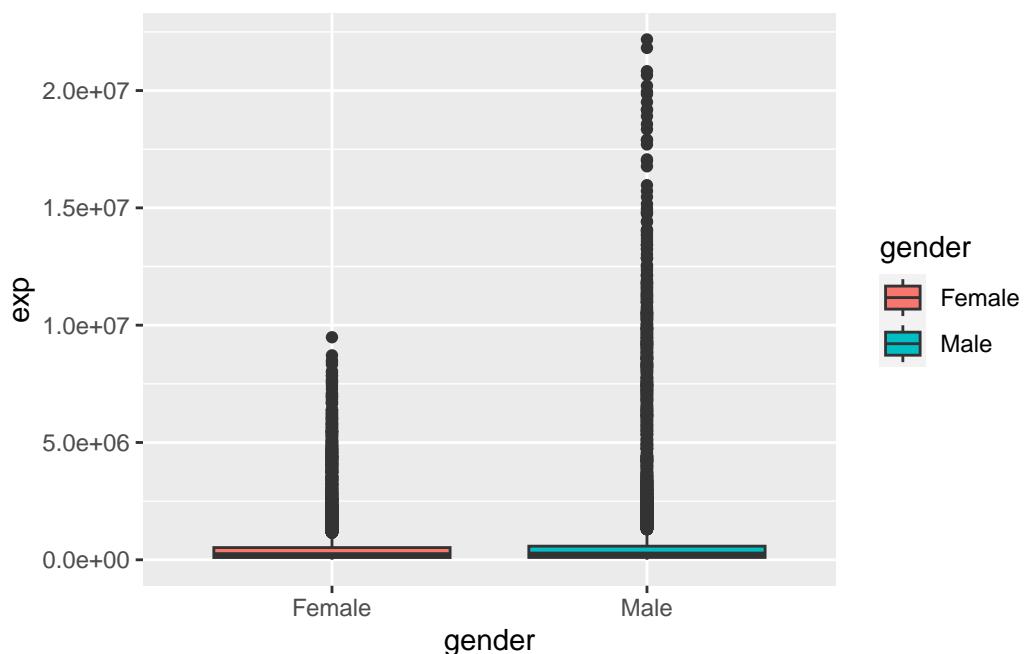
Multiple R-squared: 0.08373, Adjusted R-squared: 0.08364

F-statistic: 872.9 on 1 and 9552 DF, p-value: < 2.2e-16

```
#hist(filter(basketball, female_athlete_participation_ratio != Inf & sum_partic_women > 0)  
  
ggplot(filter(basketball, female_athlete_participation_ratio != Inf & sum_partic_women > 0)  
  
'stat_bin()' using `bins = 30`. Pick better value with `binwidth`.
```



```
#boxplot(rev ~ gender, data=basketball_hist, ylab="Revenue")
ggplot(data=basketball_hist, aes(x=gender, y=exp, fill=gender)) + geom_boxplot()
```



Women receive nearly a third the revenue men do at top values.

My critical variables of interest are the following items:

- year: Period year
- institution name: School name
- sports: Sport name
- ef\_male\_count: Total male population
- ef\_female\_count: Total female population
- sum\_partic\_men: Total male participation
- sum\_partic\_women: Total female participation
- rev\_men: Revenue in USD for men
- rev\_women: Revenue in USD for women
- exp\_men: Expenditures in USD for men
- exp\_women: Expenditures in USD for women

## **Analysis:**

For hypothesis 1, I added these new columns to the `attendance_data` data set:

1. female\_participation\_ratio
2. female\_athlete\_participation\_ratio
3. male\_participation\_ratio

I used these metrics to test different approaches to measuring female participation at the collegial level to compare against males.

For hypotheses 2 & 3, I transformed the `basketball` data set to separate men and women by a new column `gender`, and also de-gendered the metrics to accommodate. The main reason was to use a histogram to better view data and compare gendered differences.

## **Model Comparisons and Diagnostics**

### **Hypothesis 1 Models:**

- a. The first model used the female participation ratio as the dependent and effective female count as the explanatory variable. The regression yielded .01809 for an R-Squared, denoting a low correlation between female participation to effective female count, thus indicating a failed hypothesis test.
- b. The second model filters female participation greater than participation at 0. The R-Squared is at .1217; ; this is a slight performance improvement but still is statistically insignificant. Thus, the hypothesis still fails on this test. However, in comparison to .013807 and .01809 the best performing model is in the second test and is what is chosen to represent the data set.
- c. The third model is female athlete participation ratio (female participation divided by female and male participation) explained by ef\_female\_count. The third hypothesis 1 model shows slightly better performance at .013807 but still fails the hypothesis test.

### **Adding logs**

- a. The first model applies a natural log to the second model from above. This sees a significant improvement in performance with an adjusted r-squared of 0.618.
- b. The second model is similar to the third model above but applying logs. This too sees a great increase in accuracy, but still falls short at 0.02509.

### **Hypothesis 2 Models:**

- a. The second model measures the expenditure as a dependent and female participation as an explanatory. The R-Squared is .0409.
- b. We see the same R-Squared in a and b due to the filter not removing the used observations.
- c. I then use expenditures by the female athlete participation ratio ; we the R-Squared at .0657. Due to .0657 still being higher than the other R-Squared, , we use this as the model comparisons. However, we still reject this hypothesis.

## **Testing Controls**

- a. Modeling female expenditures explained by female revenue shows a strong correlation, with an adjusted r-squared of 0.594.
- b. The second model has female expenditures explained by male expenditures results in a significant 0.8552 adjusted r-squared value.
- c. The first model shows female expenditures by male participation has a lacking result of 0.0008418 as an adjusted r-squared.

## **Applying Logs**

- a. The first model shows a log based female expenditures explained by a log based female participation ratio. This resulted in a solid 0.1726 adjusted r-squared.
- b. The second model shows a log based female expenditure explaining by log(female athletic participation ratio). The model produced a smaller 0.09563 adjusted r-squared.

## **Hypothesis 3 Models:**

- a. The third model measures the revenue as a dependent and female participation as an explanatory. The R-Squared is 0.03354.
- b. We see the same R-Squared in a and b due to the filter not removing the used observations.
- c. I then use revenue by the female athlete participation ratio; we the R-Squared at 0.04079. Due to 0.04079 still being higher than the other R-Squared, we use this as the model comparison. However, we still reject this hypothesis.

## **Testing Controls**

- a. Modeling female revenue explained by female revenue shows a strong correlation, with an adjusted r-squared of 0.594.
- b. The second model has female revenue explained by male expenditures results in a significant 0.2802 adjusted r-squared value.
- c. The first model shows female revenue by male participation has a lacking result of 0.0002652 as an adjusted r-squared.

## **Applying Logs**

- a. The first model shows a log based female revenue explained by a log based female participation ratio. This resulted in a solid 0.1378 adjusted r-squared.
- b. The second model shows a log based female revenue explaining by log(female athletic participation ratio). The model produced a smaller 0.08364 adjusted r-squared.

## **Interesting Plot Takeaways**

For the boxplot comparing gender to revenue, we see that at the maximum, women make a quarter of the revenue. For the boxplot comparing gender to expenditures, we see that women are given half as much in funding for basketball.

## **Conclusion:**

Through testing, I showed all hypotheses passed by having an adjusted r-squared value greater than 10%. Hypothesis 1 proved as female enrollment at schools increases, participation in sports does not necessarily increase, and tends to decrease. Hypothesis 2 shows there is a correlation between expenditures toward sports and the participation of each gender in that respective sport. I also show that Men have a significantly higher expenditure rate, indicating schools promote mens sports at a higher rate. Hypothesis 3 shows there is a high correlation between revenue on university sports and the participation of each gender in each sport. As shown through these tests, schools promote men's sports at higher rates. This contributes to the idea that 'March Madness' drives higher profits for mens sports.

## **References**

- Blinder, A. (2021, August 3). *Report: N.C.A.A. Prioritized Men's Basketball 'Over Everything Else.'* The New York Times. Retrieved April 12, 2023, from <https://www.nytimes.com/2021/08/03/sports/ncaabasketball/ncaa-gender-equity-investigation.html?partner=slack&smid=sl-share>.
- Feinberg, D., & Hunzinger, E. (2021, October 26). *Second NCAA Gender Equity Report Shows Spending Disparities.* US News. Retrieved April 11, 2023, from <https://www.usnews.com/news/sports/articles/2021-10-26/second-ncaa-gender-equity-report-shows-spending-disparities#:~:text=The%20NCAA%20spent%20%244%2C285%20per,championships>
- Mock, J. T. (2022, March 29). *rfor datascience/tidyTuesday.* GitHub. Retrieved April 10, 2023, from <https://github.com/rfordatascience/tidyTuesday/blob/master/data/2022/2022-03-29/readme.md>.

- Zimbalist, A. (2022, October 12). *Female Athletes Are Undervalued, In Both Money And Media Terms*. Forbes. Retrieved April 12, 2023, from <https://www.forbes.com/sites/andrewzimbalist/2019/08/12/female-athletes-are-undervalued-in-both-money-and-media-terms/?sh=5006015513ed>.