Predictors of Salaries of Head Coaches of US Collegiate Sports Teams

Group 5: Xin Jin, Reid Ginoza, Heidi Lovejoy

12/8/2019

Introduction

While it seems to be common knowledge that many of the highest paid US public employees are college sports teams head coaches (Michaels 2018), less is known about the salaries of college-level sports head coaches in general. Is head coaching a lucrative career for all or only a select few? What knowledge can we gain about head coaches in private schools or across other divisions and classifications outside of the NCAA Division I?

In this project, we examined institution-level data from the Equity in Athletics Survey, academic year 2017-2018, from the U.S. Department of Education Office of Postsecondary Education (2018). In this analysis we investigated potential predictors for salaries of sports team head coaches at four-year colleges and universities in the United States. This project focused on differences of salaries between head coaches of male teams and female teams; salary diversity between division classifications; and the effects of the number of student participants (i.e. athletes) on head coach salaries. Classifications were defined by the Equity in Athletics Survey according to the athletic association (such as NAIA, NCAA, etc) and divisions or subdivisions where applicable (e.g. NCAA Division I-FBS, NCAA Division I-FCS, NCAA Division I without football). Data was available from the Office of Postsecondary Education from academic year 2002-03 to 2017-18, and this study focused on data from the most recent academic year.

The full data set for 2017-18 listed completed surveys for 2079 institutions. Excluded were the 578 two-year colleges, the 120 four-year schools with co-ed teams, and 18 institutions with missing values for head coach salaries for male or female teams or both. Our working data set had 1363 four-year institutions without co-ed sports teams, and each institution had a classification, an average head coach salary for men's teams, an average head coach salary for women's teams, a participation count of men, and a participation count of women.

Hypotheses

To study predictors of head coach salaries, we performed several hypothesis tests. We started by investigating the gender of the team as a predictor. We first hypothesized that the variances of the head coach salaries were different between the coaches of male teams compared with the coaches of female teams.
H_0: $\sigma^2_{\text{men}} = \sigma^2_{\text{women}}$, H_1: $\sigma^2_{\text{men}} \neq \sigma^2_{\text{women}}$

We then hypothesized that the median of the salaries for coaches of male teams was greater than that of the female teams: H_0: $M_{\text{male}} \leq M_{\text{female}}$, H_1: $M_{\text{male}} > M_{\text{female}}$

The classification was our second predictor under investigation. We also expected there to be differences of salaries between classifications, but we first wanted to see if we could control for the effects of the gender of the team. We predicted that there would not be an interaction between classification and gender on head coach salaries: H_0: there is not an interaction between classification and gender, H_1: there is an interaction between classification and gender

As part of understanding head coach salaries in different classifications, we hypothesized that the variance of salaries between the classifications would not be homogeneous.
H_0: $\sigma^2_{\text{DivI-noFB}} = \sigma^2_{\text{DivI-FBS}} = \sigma^2_{\text{DivI-FCS}} = \ldots = \sigma^2_{n=18}$, H_1: at least one is different

We then predicted that out of all the 18 classifications, at least one of the median salary of head coaches would be different from that of coaches in other classifications:
H_0: $M_{\text{DivI-noFB}} = M_{\text{DivI-FBS}} = M_{\text{DivI-FCS}} = \ldots = M_{n=18}$, H_1: at least one is different

Lastly, number of participants was our third predictor under investigation. We hypothesized that a linear regression model with the number of participants would predict the head coach salaries: H_0: $\beta = 0$, H_1: $\beta \neq 0$

And to determine this model's usefulness, we hypothesized that the linear regression model would be appropriate: H_0: a linear regression model is appropriate, H_1: a linear regression model is not appropriate

Methods

Data was analyzed using R version 3.6.1 (2019-07-05) (R Core Team 2019). The data was described using counts (percentage) for categorical variables and mean (standard deviation) for continuous variables.

To assess the homogeneity of variances for salaries of head coaches of male teams versus female teams, we tested the ratio of the two variances using the F-distribution. If the normality assumptions were met on the samples, we would use the t-test to test the means of the head coach salaries of male teams vs female teams. Otherwise, the nonparametric Wilcoxon test would be used to test the difference in the medians.

To analyze the interactions of gender and classification, a two-way ANOVA was completed, and a test for interactions was performed with the F statistic. A Brown-Forsythe-Levene Test for homogeneity of variances was conducted on the variances of the classifications. After an examination of the residuals of the ANOVA, a Kruskal-Wallis was used to test if there are differences between the classifications, and the posthoc Kruskal-Wallis procedure was used for pairwise comparisons.

Due to the way the data was arranged, data for the linear regression model of the head coach salary and participation level used two cases for each school – one for the women participants and women's team coaches' salaries and one for those of the men. Thus the participation level used was more specifically participation by gender. A linear regression model was constructed, and a t-test was used to determine the significance of the model. Finally, we determined whether or not a linear regression model is appropriate by using the test for lack of fit in linear regression.

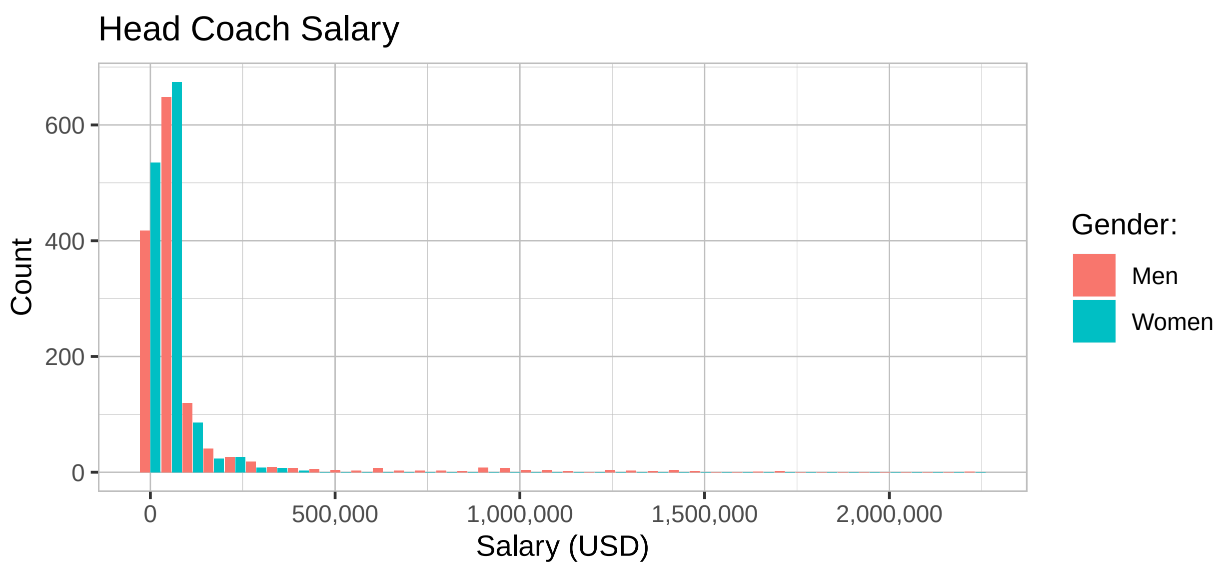
Results

Description of Data

Data are summarized in the tables below, but some highlights are included here. First, of all head coaches of both male and female teams the lowest salary for a head coach was $763, while the highest was $2,222,497. The mean was $77,955.90 ($69,698.10). Looking at participation by gender, the lowest participation was 1 student, and the highest was 728 students. The mean participation grouped by gender at an institution was 387.28 (207.90) student athletes. The classification with the largest number of institutions was NCAA Division III with football with 223 (16.4 %) institutions. Independent and NJCAA Division III were the smallest groups with 4 (0.3 %) institutions each.

Variable	Mean	SD
Head Coach–Men's Salary	107,801.45	230,718.91
Head Coach–Women's Salary	48,110.37	51,009.74
Head Coach–All Salary	77,955.91	169,698.1
Participation–Men	218.26	123.14
Participation–Women	169.02	94.32
Participation–Total	387.28	207.9

Classification	Count	Percentage	Salary Mean	Salary SD
Independent	4	0.3%	7,406.25	2,969.149
NAIA Division I	90	6.6%	30,560.88	11,576.56
NAIA Division II	102	7.5%	25,299.23	10,985.52
NCAA Division I without football	95	7.0%	115,310.71	91,262.84
NCAA Division I-FBS	117	8.6%	431,159.52	421,253.92
NCAA Division I-FCS	114	8.4%	100,266.23	62,701.30
NCAA Division II with football	160	11.7%	47,834.02	18,145.77
NCAA Division II without football	138	10.1%	41,016.35	20,359.58
NCAA Division III with football	223	16.4%	34,898.80	13,059.53
NCAA Division III without football	162	11.9%	24,589.05	14,933.11
NCCAA Division I	9	0.7%	20,168.89	12,021.08
NCCAA Division II	25	1.8%	12,522.12	10,128.83
NJCAA Division I	35	2.6%	35,547.23	20,867.87
NJCAA Division II	9	0.7%	15,033.67	8,274.244
NJCAA Division III	4	0.3%	16,895.75	10,753.45
NWAC	9	0.7%	6,894.611	1,444.328
Other	33	2.4%	17,607.74	31,676.31
USCAA	34	2.5%	11,719.66	12,993.22

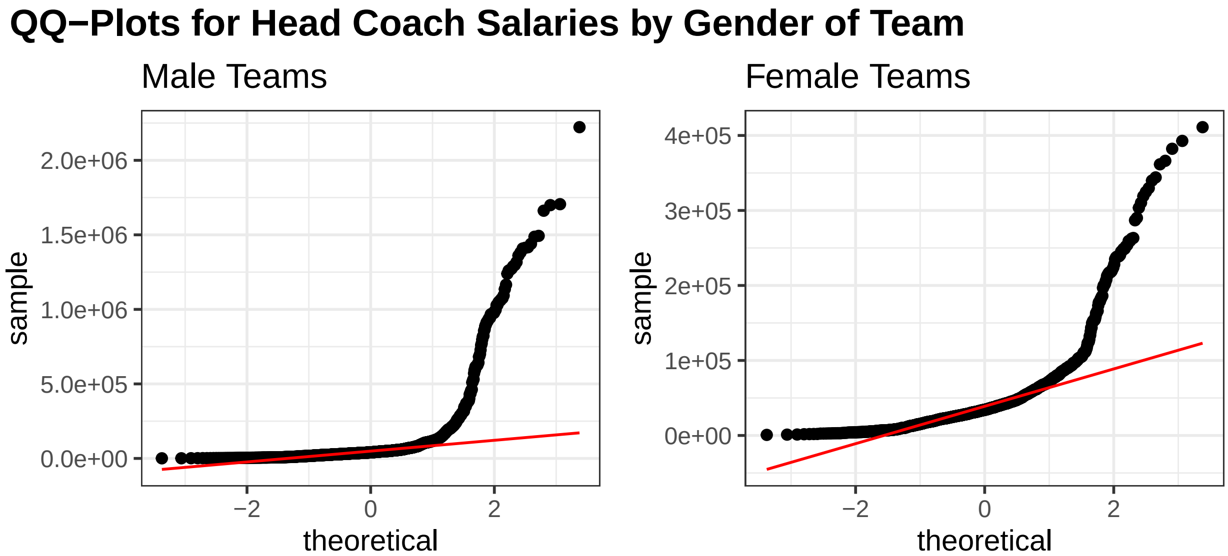


Analysis Results

```
gender.variance.test <- institution_data %$%
  var.test(HDCOACH_SALARY_MEN, HDCOACH_SALARY_WOMEN, alternative = "two.sided",
           conf.level = 0.95, ratio = 1)
gender.var.p.val <- p.value.string(gender.variance.test$p.value)
```

The test for equality of variances between salaries of head coaches of men's versus women's teams resulted in a test statistic of $F_0 = 20.46$ ($p < 0.0001$) and critical value of $F_{\alpha/2, df_{\text{men}}, df_{\text{women}}} = 1.081$, using $\alpha = 0.05$.

There was sufficient evidence to suggest a difference between the variances of head coach salaries of male teams and female teams.



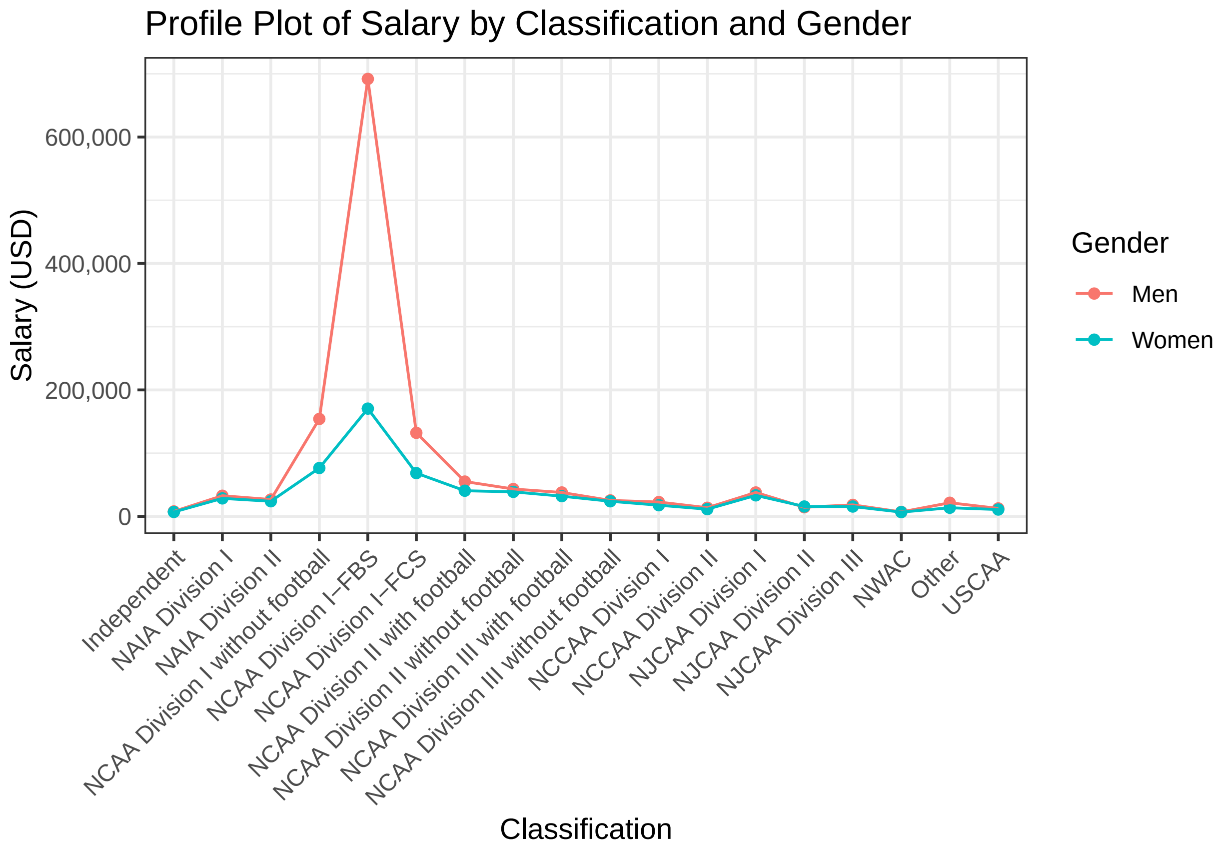
Upon visual inspection of the qq-plots, the data for salaries by gender of team could not be assumed to be normal.

```
gender.wilcoxon.test <- wilcox.test(institution_data$HDCOACH_SALARY_MEN,
                                    institution_data$HDCOACH_SALARY_WOMEN,
                                    alternative = "greater",
                                    paired = FALSE)
gender.wilcoxon.p.val <- p.value.string(gender.wilcoxon.test$p.value)
```

Testing to see if salaries of head coaches of men's teams tend to be greater than those of head coaches of women's teams resulted in a test statistic of $T_0 = 1.064179 \times 10^6$ ($p < 0.0001$). There was sufficient evidence to suggest that salaries for head coaches of men's teams are greater than those of women's.

```
classification.gender.interaction.results <- gender.separated_data %$%
  aov(Salary ~ Classification * Gender)
interaction.summary <- classification.gender.interaction.results %>% summary()
interaction.resid <- tibble(Residuals=(
  classification.gender.interaction.results %>% residuals()))
```

An ANOVA model with gender and classification factors resulted in a test statistic for interactions of $F_0 = 79.5$ ($p < 0.0001$). There was sufficient evidence to suggest that there are interaction effects between gender and classification.



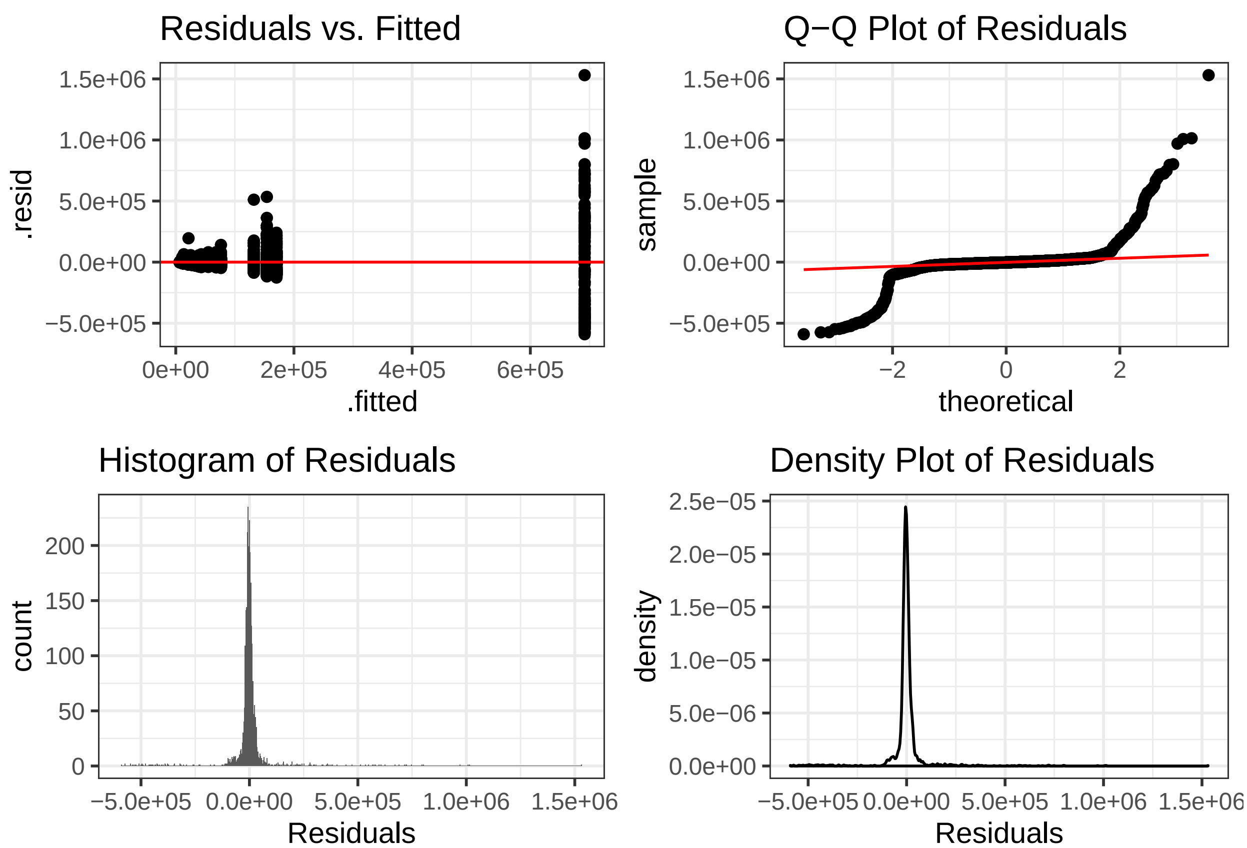
While this profile plot shows interaction effects, upon visual inspection, it seems to be limited to the three NCAA Division I classifications: NCAA Division I without football, NCAA Division I-FBS, and NCAA Division I-FCS. In those three cases, head coaches for men's teams have higher salaries, especially in the case of NCAA Division I-FBS.

To check the assumptions of the ANOVA model, we examined the variance across classifications to check for constant variance and normality of the residuals.

```
gender.separated_data %$%
  leveneTest(Salary ~ Classification, center = median) -> classification.variance.results
```

The Brown-Forsythe-Levene Test for Homogeneity of Variances resulted in a test statistic of $F_0 = 77.03$ ($p < 0.0001$), so there is sufficient evidence to suggest that at least one of the classifications' salaries have a variance different from the others.

Upon a visual examination of the residuals and the q-q plot, it is also clear that the normality assumption of the residuals was not satisfied.

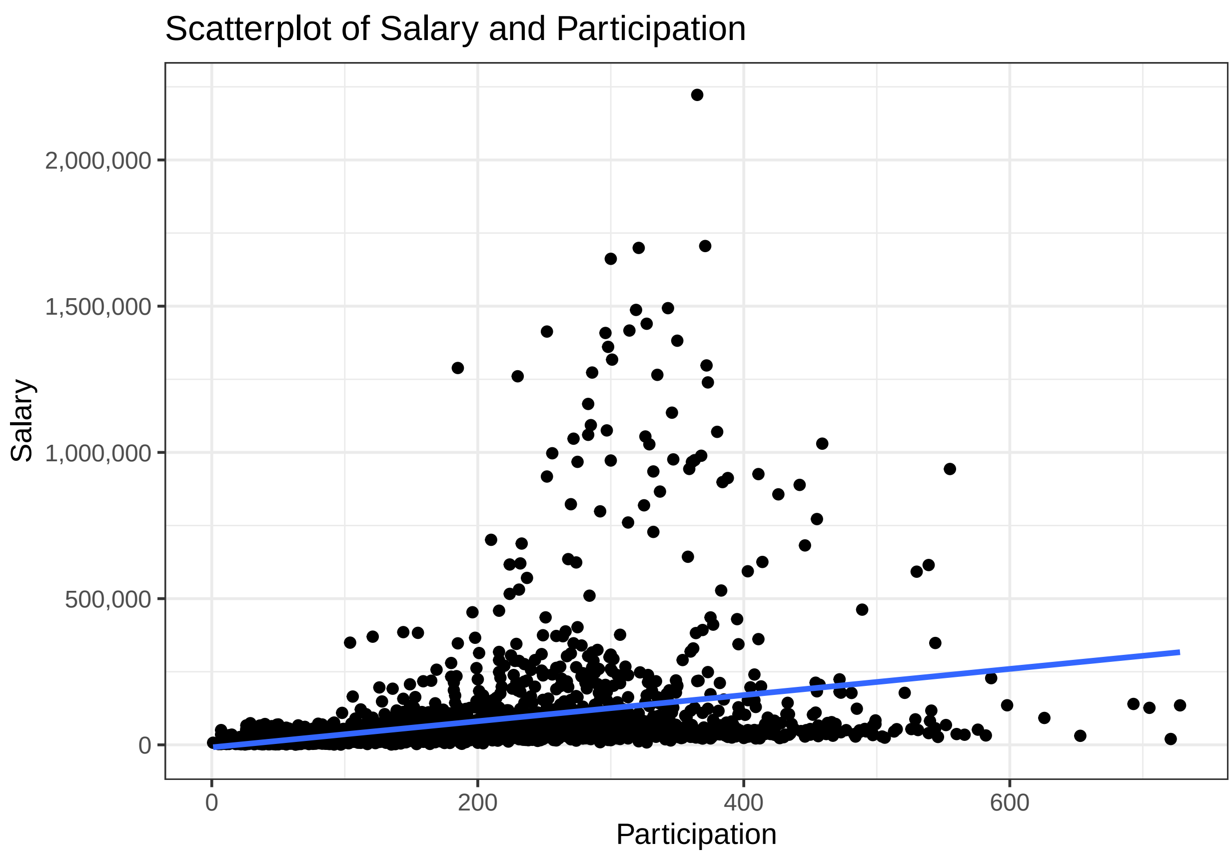



```r
classification.kw.test <- gender.separated_data %$%
  kruskal.test(Salary ~ Classification) %>% tidy()
```

The Kruskal-Wallis test on the salaries grouped by classifications resulted in a test statistic $H = 1752.51$ ($p < 0.0001$), so there is sufficient evidence to suggest that at least one classification is different from the others. Thus, the posthoc procedure for the Kruskal-Wallis test was completed to find pairwise differences between the classifications.

```r
classification.kw.posthoc <- gender.separated_data %$% kruskalmc(Salary ~ Classification)
```

From the 153 possible pairwise comparisons, 91 resulted in significant difference. Complete results are not included, but of particular interest is the NCAA Division I: NCAA Division I-FBS had a significant difference from all of the 17 other classifications, NCAA Division I-FCS had a significant different from 16 other classifications, NCAA Division I without football the only exception, and similarly, with NCAA Division I without football, the only non-significant difference was with NCAA Division I-FCS.




```
participation.model <- gender.separated_data %$% lm(Salary ~ Participation)
participation.coefficients <- participation.model %>% coefficients()
```

Shown above is a scatterplot of the salary and participation data with the line of best fit from the linear regression model given by the following:

$$\hat{y} = -8622.92 + 447.11x$$

```
participation.summary <- participation.model %>% summary()
participation.t.test <- participation.summary[[4]] %>% as_tibble()
```

A t-test was conducted to test the signifiance of the linear regression model. This resulted in a test statistics of $t_0 = 16.18$ ($p < 0.0001$). There was sufficient evidence to suggest that the regression line is significant.

```
participants.regression.appropriate <- participation.model %>%  pureErrorAnova()
```

The lack of fit ANOVA table is as follows:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Participation	1	6.880982e+12	6.880982e+12	292.468586	0
Residuals	2724	7.159205e+13	2.628196e+10	NA	NA
Lack of fit	479	1.877337e+13	3.919284e+10	1.665849	0
Pure Error	2245	5.281868e+13	2.352725e+10	NA	NA

Testing whether or not a linear regression model is appropriate resulted in test statistic $F_0 = 1.67$ ($p < 0.0001$). There is sufficient evidence to suggest that a linear regression model is not appropriate.

Conclusion

We sought to understand the factors of head coach salaries at the collegiate level, and while the highest salaries are commonly reported, not all head coaches have those headline-worthy salaries. The NCAA Division I had the highest salaries, and notably, the biggest difference between salaries of coaches for the men's teams and that of the women's teams. That the subdivisions of NCAA Division I were also shown to be different from one another further suggests that these differences in subdivisions strongly affect head coach salary.

There are different funding allowances between these subdivisions, but only between football teams. A future study might investigate the head coach salaries of collegiate football head coaches separately from the other sports or compare head coaches across classification and sport. Perhaps the sport of football is where the greatest gender disparity of salary exists.

This difference in head coach salary based on gender is alarming, and we urge our readers to investigate further. However, we wish to be specific: our data reports the gender of the athletes themselves, whereas we are calling attention to the position of head coach. Missing from our data from the EADA survey is the genders of the coaches themselves. Generally, men coach men's teams, but men and women both coach women's teams. Future analyses should investigate whether there are differences exist between male head coaches and female head coaches, regardless of the gender of the teams.

Further investigation is also needed to understand the relationship between participation numbers and head coach salary. We gave shown that generally, the salary will increase with the increasing of participation, but our current linear model was not appropriate and we urge our readers to continue to find a more appropriate model.

Our data set contained only the averages of head coach salaries of a given institution, so we were unable to capture any variance within an institution. Of particular interest would be variation between head coaches by type of sports, within and across institutions. Additionally, this study only used the surveys from one scholastic year, and future studies should investigate the salaries over multiple years.

This was an initial study of salaries of head coaches of collegiate sports teams. While there is more work to be done, we believe we have paved the way in identifying important predictors of their salaries, and we hope others will continue the investigation.

Session Info

```r
sessionInfo()
```

```
## R version 3.6.1 (2019-07-05)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS High Sierra 10.13.6
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] alr3_2.0.8       haven_2.1.1      magrittr_1.5     kableExtra_1.1.0
##  [5] knitr_1.24       broom_0.5.2      forcats_0.4.0    stringr_1.4.0
##  [9] dplyr_0.8.3      purrr_0.3.2      readr_1.3.1      tidyr_0.8.3
## [13] tibble_2.1.3     ggplot2_3.2.1    tidyverse_1.2.1  scales_1.0.0
## [17] pander_0.6.3     car_3.0-3        carData_3.0-2    pgirmess_1.6.9
## [21] cowplot_1.0.0
##
## loaded via a namespace (and not attached):
##  [1] nlme_3.1-140       sf_0.8-0           lubridate_1.7.4
##  [4] webshot_0.5.1      gmodels_2.18.1     httr_1.4.1
##  [7] tools_3.6.1        backports_1.1.4    rgdal_1.4-6
## [10] R6_2.4.0           KernSmooth_2.23-15 spData_0.3.2
## [13] rgeos_0.5-2        DBI_1.0.0          lazyeval_0.2.2
## [16] colorspace_1.4-1   withr_2.1.2        sp_1.3-1
## [19] tidyselect_0.2.5   splancs_2.01-40    curl_4.0
## [22] compiler_3.6.1     cli_1.1.0          rvest_0.3.4
## [25] expm_0.999-4       xml2_1.2.2         labeling_0.3
## [28] classInt_0.4-2     digest_0.6.20      foreign_0.8-71
## [31] rmarkdown_1.15     rio_0.5.16         pkgconfig_2.0.2
## [34] htmltools_0.4.0    rlang_0.4.0        readxl_1.3.1
## [37] rstudioapi_0.10    generics_0.0.2     jsonlite_1.6
## [40] gtools_3.8.1       spdep_1.1-3        zip_2.0.4
## [43] Matrix_1.2-17      Rcpp_1.0.2         munsell_0.5.0
## [46] abind_1.4-5        stringi_1.4.3      yaml_2.2.0
## [49] MASS_7.3-51.4      grid_3.6.1         maptools_0.9-8
## [52] gdata_2.18.0       crayon_1.3.4       deldir_0.1-23
## [55] lattice_0.20-38    splines_3.6.1      hms_0.5.1
## [58] zeallot_0.1.0      pillar_1.4.2       boot_1.3-22
## [61] LearnBayes_2.15.1  glue_1.3.1         evaluate_0.14
## [64] data.table_1.12.2  modelr_0.1.5       vctrs_0.2.0
## [67] cellranger_1.1.0   gtable_0.3.0       assertthat_0.2.1
## [70] xfun_0.9           openxlsx_4.1.0.1   e1071_1.7-2
## [73] coda_0.19-3        viridisLite_0.3.0  class_7.3-15
## [76] units_0.6-5
```

References

Michaels, Matthew. 2018. "College football and basketball coaches are the highest-paid public employees — here are the biggest paydays." *Business Insider*. https://www.businessinsider.com/highest-paid-public-job-every-state-college-football-basketball-2018-3.

Office of Postsecondary Education. 2018. "Equity in Athletics Data Analysis." U.S. Department of Education. https://ope.ed.gov/athletics/#/datafile/list.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.