Anna Bruijn, Kiley Delaney, Wendy Franco
Christina Knudson
MATH 155
20 April 2016

I'll See You at the Movies?: Diversity and Profits in Hollywood

Introduction

After another year of #OscarsSoWhite trending on Twitter, diversity in Hollywood has been a hot topic. The Oscars are theoretically merit-based, and while it is highly debatable that no actors of color delivered performances that deserved recognition in 2015, commentators and a variety of people working in Hollywood noted that it is difficult to win awards for roles that don't exist. The big-budget movies Hollywood produces tend to feature predominantly white actors, and these are the movies that tend to be honored by the Oscars ceremony.

So, why doesn't Hollywood make a conscious effort to produce films that feature people of color? Studio executives often blame the box office. They like to say that films like these have a "niche" audience, and that while members of minority groups will watch movies about white people, the reverse is not true. We wanted to challenge this assumption through statistical analysis and explore the relationship between box office profits and the diversity of a film. We are doing this by looking at the diversity of actors within the movies. We define actors as the people in the movie with a speaking part. To get a more complete picture, we broadened our analysis to include female (another underrepresented group in Hollywood) actresses in our definition of "diversity," as well as the diversity (specifically gender and racial/ethnic) of directors in addition to actors. We chose to add diversity of directors into our analysis with the goal of discovering any connections between diversity and box office success of a film, due to directors having most significant creative impact on the final project. Through our analysis, we hoped we would be able to conclude that diversity among the cast and directors of a film has a neutral or positive effect on box office gross, and finally put the excuse cited above to rest. We can generalize our results to our population: all films that were in the top 100 highest-grossing films in the United States for every year between 2010 and 2014.

Methods

After discussing the variables we could explore, we decided on a final, specific question: Is there a relationship between a more diverse cast and creative team and greater profits at the box office for films with the biggest box office profits in the United States made in the last five years? We chose two categorical predictor variables (the gender and race of the film's director), two quantitative predictor variables (the percentage of speaking parts played by female actors and actors of color, in percentages), and one quantitative response variable (the domestic box office gross of a film, in dollars). We decided to conduct a retrospective observational study, because we did not have the time or resources to conduct an experiment that involved making movies in Hollywood, or to generate our own data.

To collect the data, we gathered a list of the 100 highest-grossing films in the United States from 2010 from Box Office Mojo. We compiled similar lists for the years 2011, 2012, 2013, and 2014, and these lists served as our population. From there, we randomly selected ten movies from each year's list to be our sample. To avoid any possible bias–potential bias being the selection of films we knew had a diversity within its cast directors or vise-versa–our data happened through random selection using a computer generator, choosing ten numbers at a time. These numbers corresponded with the rank in the top 100 list of highest grossing films. We repeated this process for all five years. So our sample consists of 50 movies, with 10 from the year 2010, 10 from 2011, 10 from 2012, 10 from 2013, and 10 from 2014. For each film in our sample, we found information on the gender and race of the film's director and calculated the percentages of speaking parts played by female actors and non-white actors using the listings posted on the Internet Movie Database. We also found the box office gross in the United States for each film from Box Office Mojo. No one was paid to obtain this information, or to work on this project. Because we randomly selected our sample from the list of the 100 highest-grossing films in the US for the years from 2010 to 2014, we are able to generalize our results to this population.

We believe that the data on which films made the most money in the United States and exactly how much money they made is trustworthy, because we obtained them from reliable sources that have no motivation, financial or otherwise, to misrepresent the data. Since we

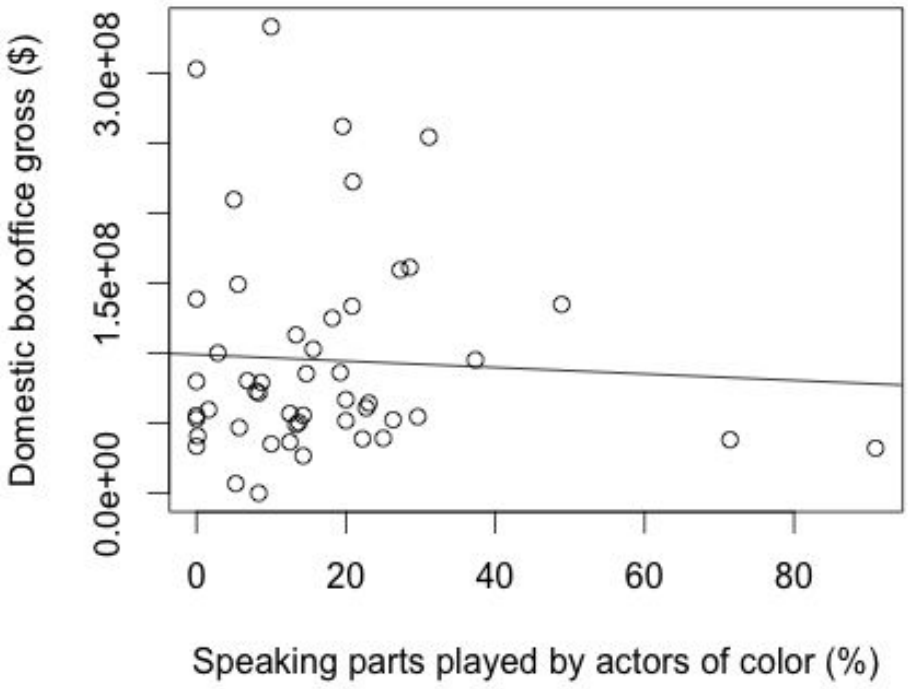
collected the data on race and gender ourselves, there is some room for error, simply because humans are imperfect and we may have missed an actor of color in the cast list or mislabeled a director, for example. However, we were careful in our collection, and we believe that the margin of error is small.

To conduct our analysis, we used R Studio (version R 3.2.3 GUI 1.66 Mavericks build). We choose a 0.05 significance level, because this is a standard level and we do not have a valid argument for our data or testing to change the significance level. In order to conduct our analysis, we had to create linear model regressions that would test the relationship between a film's box office gross (which will act as our response in a linear regression model) and the diversity within cast and directors (which will act as our predictors) , and put them into RStudio so that we could use them. After reading in our data with the functions **read.csv()** and **attach()**, the very first thing we did use the function **plot()** to see if there was any relationship between a director's race and a film's domestic box office gross in box plot where the race (white or non-white) of the film's director was the predictor and the film's domestic box office gross was the response. After analyzing our results for that particular box plot, we decided to focus on only diversity within cast and proceeded to name our regressions and read them as linear models into RStudio through the function **lm()**. For each of our regressions, we used the function **summary()** to obtain the a summary of the all the statistical information (such as strength of relationships between our predictor and our response for example) pertaining to our regression models. To get a visual the relationship between our model's data, we used the function **plot()** once again followed by the function **abline()** to plot our linear regression in order to see how well our linear model fits with our data.

Results

We noticed that, of the fifty films we had randomly chosen, only six had non-white directors and only one had a female director. This sample seemed too small to produce any truly relevant analysis, and we thought that any linear relationship we could generate between director race and/or sex and the box office gross of a film with this scant data would be unreliable. Thus, we decided to focus on modeling the relationship between actor diversity and box office gross.

To begin our analysis, we decided to work primarily with two linear models. The first, which we named **nonwhiteActors**, took the percentage of speaking parts played by nonwhite actors as the predictor in our model, and domestic box office gross as the response. When we plotted the data and fit the model **nonwhiteActors** to the plot, we produced the following:



This graph shows the percentage of speaking parts played by actors of color being the predictor of our model by being the horizontal axis of our graph, while the the vertical axis of our graph, our response, is the domestic box office of a film. This graphs shows a majority of the dots present bunching up at around 0%-30% 'Speaking parts played by actors of color'. This already tells us that the majority of the speaking parts (or cast) in the films in our data are not people of color. We have two lone dots (which we would visually consider two outliers) that go beyond 50%: *47 Ronin* (2013) had a cast consisting of approximately 70% people of color, and the cast of *Kevin Hart: Let Me Explain* (2013) was around 90% nonwhite. The two dots that represent these films are relatively low on our response scale, telling us that these two films also has relatively low box office gross incomes, while the dots on the other side vary their positions on the response scale representing an array of different gross incomes. Our linear regression line

also slopes down our graph, telling us that as the percentage of people of color within a film's cast (our predictor) increases, domestic box office gross (our response) decreases.

Looking at the **summary** output for this model, we can see that the regression equation for this model is **expected domestic box office gross = 99113551 - 230898(percentage of speaking parts played by actors of color)**. The coefficient -230898 indicates that, according to the model, there is a $230,898 decrease in box office gross for every 1% increase in the percentage of speaking parts played by actors of color. This is also visually shown in our graph above where our linear regression line goes down as the percentage of our predictor goes up. In the context of our study, this would appear to mean that these faceless studio heads are correct: movies featuring more actors of color do not perform as well at the box office. However, we must also consider the low concentration of nonwhite actors in the cast of almost every film in our data set. Given this, we cannot declare causation and conclude that a greater proportion of actors of color in the cast of a film causes a decrease in box office profits. Instead, this analysis supports our initial suspicion that the powers behind expensive, eagerly anticipated movies (which will often make the most money at the box office) tend not to "risk" anticipated profits by casting people of color.

However, we must also consider whether the linear relationship between these two variables is significant enough to draw that conclusion. Based on the summary output, we can perform a hypothesis test to determine if this model is useful. We will take a grand mean model as our null hypothesis (which will state "There is not a significant linear relationship between box office gross and the percentage of speaking parts played by actors of color") and the **nonwhiteActors** model as our alternative hypothesis (which will state:"There is a significant linear relationship between box office gross and the percentage of speaking parts played by actors of color"). Let 0.05 be the significance level, because this is a standard, and we do not have a reason to change the standard. In order to test the significance of the hypothesis test, we look at the p-value in the summary output of our regressions. We can see in the summary output for **nonwhiteActors** that the p-value for this hypothesis test is 0.718. Since the p-value is greater than the significance level, we fail to reject the null hypothesis in favor of the alternative hypothesis. This could be due to the fact that "There is a significant linear relationship between

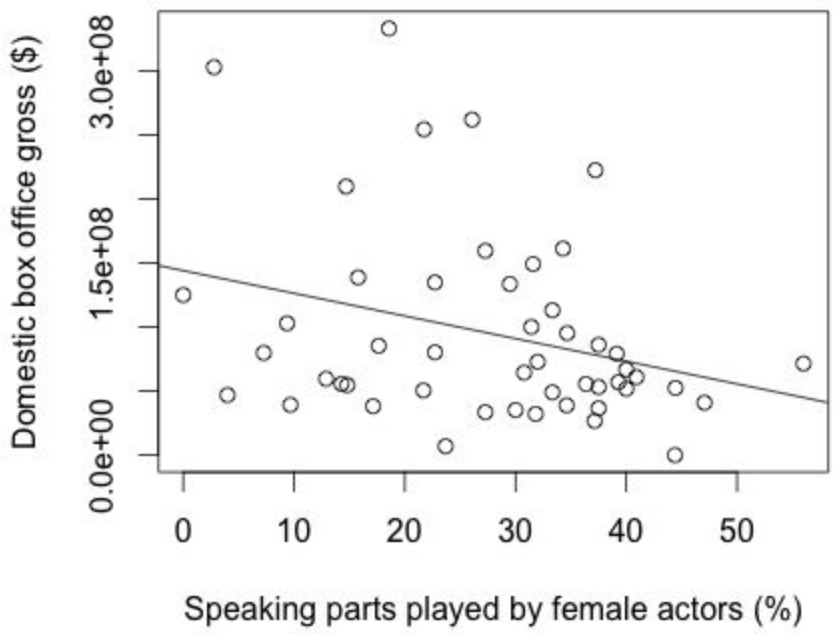
box office gross and the percentage of speaking parts played by actors of color" is truly wrong, or we might just not have enough evidence to prove it. Given the small sample size we are working with, this is a legitimate possibility.

Even without performing a hypothesis test, there are hints in the summary that this model might not be the best tool for predicting the box office success of a film. The standard error on the coefficient is $635,150, which indicates that the model has a 95% confidence interval (which we are using because our level of significance is 5%) of (-1501198, 1039402). We obtained this interval by doubling the standard error and adding/subtracting that number to to the mean (the previously mentioned coefficient) of the regression. In context, this means that we are 95% confident that every 1% increase in the percentage of speaking parts played by nonwhite actors is associated with a change in the domestic box office gross of a film that is between a $1,501,198 decrease and a $1,039,402 increase. This is a huge range of error, so it is not surprising that we were unable to definitively declare that a significant linear relationship between these two variables exists using this model and this data set.

Finally, we can also see in the summary output for **nonwhiteActors** that the r-squared value is 0.002804. The r-squared value shows how close the data are to the regression line so in our data this r-squared value indicates that the linear relationship with the predictor explains little of the response's variability due to it being such a small number. In context, this means that the percentage of nonwhite actors in a film is not a reliable way of predicting a film's success at the box office, just as we have concluded by conducting a hypothesis test and looking at the confidence interval.

Next, we looked at the percentage of speaking parts played by female actors to determine if it was more significant as a predictor variable. The second linear model, which we called **femaleActors**, took the percentage of speaking parts played by female actors as the predictor and domestic box office gross as the response. When we plotted the data and fit the model to the plot,

we produced the following:



This particular graph shows a lot of dots centering at around 25%-42% of our predictor, which is the percentage of speaking parts played by women in a film. While the dots are a little scattered around this plot, they mostly remain relatively low on our response scale with around 5 to 6 films reaching higher gross incomes. These particular dots, however, all have less 40% female cast in our 50% scale. Our scale alone tells us that the majority of highest grossing films have no more than 50% female cast. Our linear regression line shows a downwards growing slope, telling as that as the percentage of our predictor increases, the domestic box office gross decreases. 2

Looking at the **summary** output for this model, we can see that the regression equation for this model is **expected domestic box office gross = 144129274 - 1771376(percentage of speaking parts played by female actors)**. The coefficient -1771376 indicates that, according to the model, there is a $1,771,376 decrease in box office gross for every 1% increase in the percentage of speaking parts played by female actors. In the context of our study, this would mean that movies featuring more women do not perform as well at the box office. In fact,

according to this model, featuring more female actors is associated with a decrease of how much money a film makes.

However, once again, we must also consider whether the linear relationship between these two variables is significant enough to draw that conclusion. Based on the summary output, we can perform a hypothesis test to determine if this model is useful. We will take a grand mean model as our null hypothesis (which states "There is not a significant linear relationship between box office gross and the percentage of speaking parts played by female actors") and the **femaleActors** model as our alternative hypothesis (that states "There is a significant linear relationship between box office gross and the percentage of speaking parts played by female actors"). Let 0.05 be the significance level once again. We can see in the summary output for **femaleActors** that the p-value for this hypothesis test is 0.0386. Since the p-value is less than the significance level, we reject the null hypothesis in favor of the alternative hypothesis. That is, we conclude that there is a significant linear relationship between the percentage of speaking parts played by female actors and box office gross.

However, we should once again look the standard errors in the summary output in order to consider the reliability of this model. The standard error on the coefficient is 832519, which indicates that the model has a 95% confidence interval of (-3436414, -106338). In context, this means that we are 95% confident that every 1% increase in the percentage of speaking parts played by female actors is associated with a decrease in the domestic box office gross of a film between $106,338 and $3,436,414. Again, this is a huge range of error, so while the model may have passed our hypothesis test, it might not be the best method of predicting a film's box office gross.

Finally, we can also see in the summary output for **femaleActors** that the r-squared value is 0.08786. This indicates that the linear relationship with the predictor explains little of the response's variability due to its relatively small number. In context, this means that the percentage of female actors in a film is not a reliable way of predicting a film's success at the box office. Thus, despite the results of our hypothesis test, this model is not entirely reliable and would not be a good tool for predicting domestic box office gross.

Discussion

Ultimately, we found ourselves unable to reach any definitive conclusions about the relationship between the creative diversity behind a film and its success at the box office. We believe that our data set is just too small for us to declare anything with a great deal of accuracy, and furthermore there is too little diversity in our random sample for us to make any definitive conclusions about its effect on box office results for our population of the 100 highest grossing box office movies in the US for every year from 2010 to 2014. We were dismayed to find when collecting our data how few of the movies we randomly selected had casts with a percentage of women or nonwhite actors that were representative of the American population, and, as we discussed in the results section above, we found so few movies in our dataset that had directors who were not white men that we decided we did not have enough information to analyze any relationship that the data might produce.

This is disheartening, and it illustrates the Catch-22 of disproving Hollywood elites' argument that movies with diverse creative teams and casts don't make money with statistics: because Hollywood largely tends not to make movies heavily featuring women and actors of color, it is almost impossible to prove with any reliability that these movies tend to make just as much money, because there just isn't enough data there. The same is true for movies directed by anyone not a white man.

In addition, there are other confounding variables to consider. Some movies are seen as "slam dunks" in that they are pretty much guaranteed to make huge profits before the trailers have even been released. Superhero movies and sequels to hugely successful films often fall into this category. In these cases, studios are even more reluctant to hand off these projects to people who don't fit the established white, male director mold. As a result, female and nonwhite directors are perhaps more likely to make independent movies with smaller budgets, which we did not include in our population. Because these projects don't have the huge marketing pushes that come with being financed by a major studio, they usually make less money at the box office. The same can be said for projects that heavily feature women or actors of color, unfortunately. However, the lack of box office success for these films isn't necessarily attributable to the fact that audiences don't want to see them. It is highly possible, and even likely, that many people are

hungry to see themselves reflected on screen but don't have access to the small handful of theatres that these movies are playing in.

It is also important to note that we could be making a type II error, meaning that we fail to reject the null hypothesis, where we should have rejected the null hypothesis. Although we do not believe that the chance of a type II error is very likely, because the sample size consists of ten percent of the population size and is therefore big enough.

In conclusion, we think that diversity is a serious problem that needs to be addressed right now, but we don't think that this sort of statistical analysis is the best tool for proving why, simply because there isn't enough data right now. Hopefully, in the future there will be many, many movies with the kind of diverse creative teams we had hoped to see when we began collecting data, and at that point we might be able to draw conclusions about the relationship between diversity and box office success. However, that also means that the Hollywood elite we discussed in our introduction shouldn't be able to say that diverse films don't make money at the box office, because, in our opinion, there isn't enough data to support that assertion either. Right now, we just don't have the information we would need to define any relationship between the on- and offscreen diversity of a movie and the money it makes at the box office. Future research could include populations of lower grossing box office movies, so that we might be able to find some correlation between diversity and box office income not just in relation to the highest grossing movies. This would give us the opportunity to assemble enough diversity datasets.

Appendix

Anna Bruijn, Kiley Delaney, Wendy Franco

April 20, 2016

Required mosaic package to import some methods we will use later

```
require(mosaic)
```

```
## Loading required package: mosaic
```

```
## Loading required package: 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
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
## Loading required package: car
```

```
## Loading required package: mosaicData
```

```
##
## Attaching package: 'mosaic'
```

```
## The following object is masked from 'package:car':
##
##     logit
```

```
## The following objects are masked from 'package:dplyr':
##
##     count, do, tally
```

```
## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cov, D, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var
```

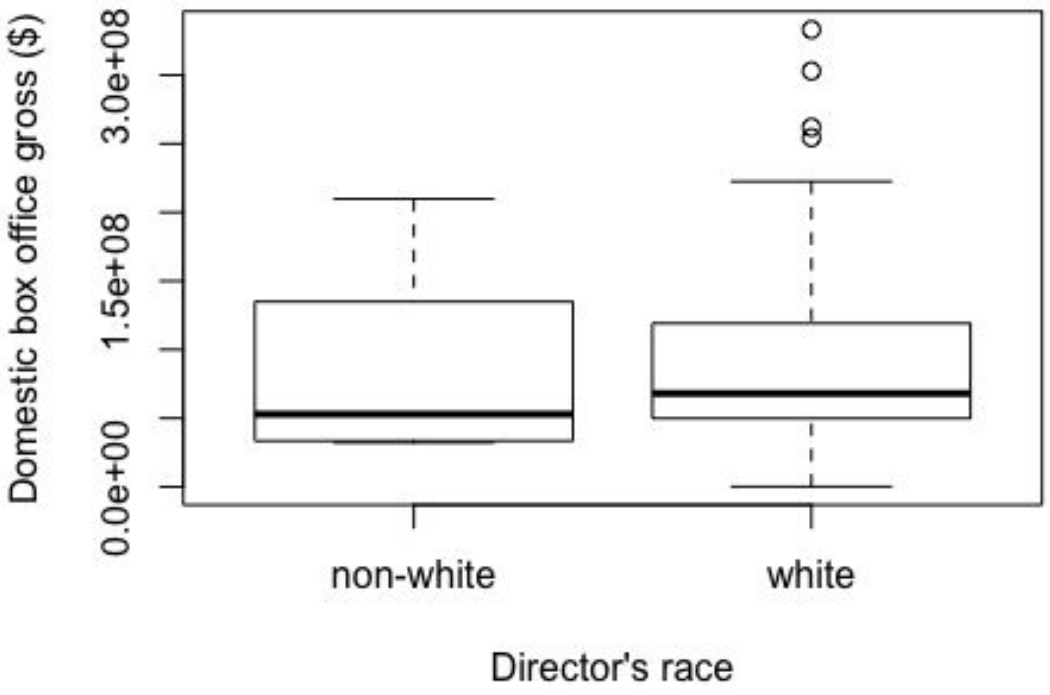
```
## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

Read the data we collected in from the original csv file, named it movieData, and attached it so that we would not have to specify data=movieData in each call

```
movieData <- read.csv("Project Data.csv")
attach(movieData)
```

Made a boxplot with the race (white or non-white) of the film's director as the predictor and the film's domestic box office gross as the response

```
plot(Director.race, Domestic.box.office.gross, xlab="Director's race",
ylab="Domestic box office gross ($)")
```



Defined directorDiversity to be a linear model with the race (white or non-white) of the film's director as the predictor and the film's domestic box office gross as the response, then called summary(directorDiversity) to get the intercept and coefficient estimates, as well as other information about the model

```
directorDiversity <- lm(Domestic.box.office.gross~Director.race)
summary(directorDiversity)
```

```
##
## Call:
## lm(formula = Domestic.box.office.gross ~ Director.race)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -94894732 -46813557 -26975593  27096428 238150360
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        86054222   30974957   2.778  0.00778 **
## Director.racewhite  8972018   33019415   0.272  0.78700
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 75870000 on 48 degrees of freedom
## Multiple R-squared:  0.001536,   Adjusted R-squared:  -0.01927
## F-statistic: 0.07383 on 1 and 48 DF,  p-value: 0.787
```

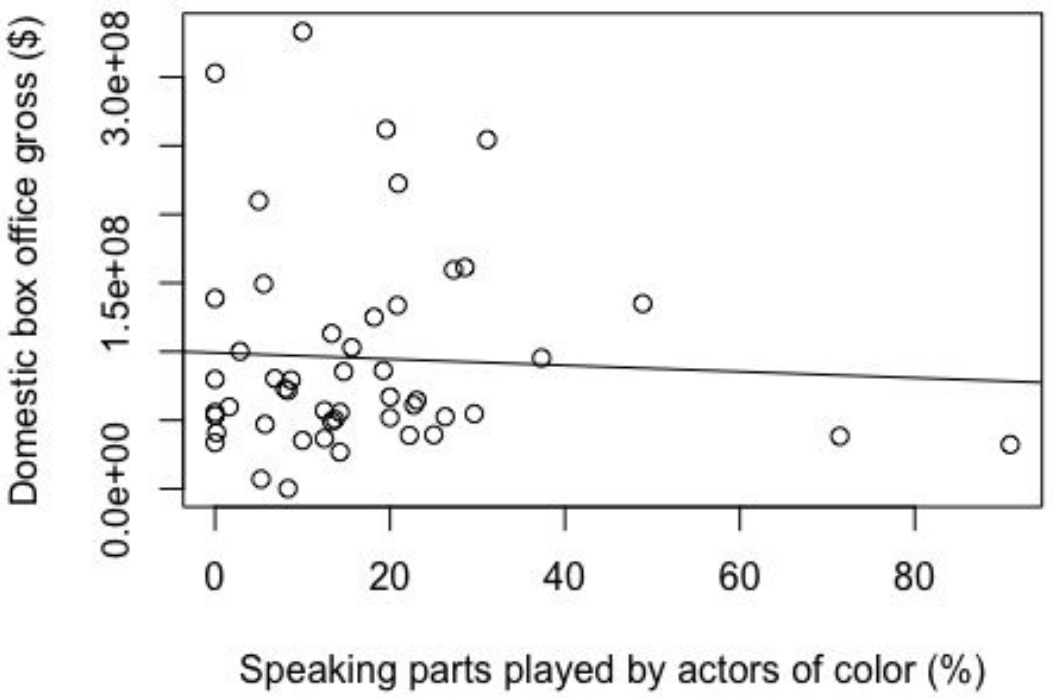
Made a boxplot with the percentage of speaking parts played by non-white actors in a film as the predictor and the film's domestic box office gross as the response, then fitted a linear model to this relationship and added that line to the plot. We also called the summary for this model to get information about it, like its coefficient and intercept estimates.

```
plot(Speaking.parts.played.by.actors.of.color....,
Domestic.box.office.gross, xlab = "Speaking parts played by actors of color
(%)",  ylab="Domestic box office gross ($)")
summary(lm(Domestic.box.office.gross~Speaking.parts.played.by.actors.of.col
or....))
```

```
##
## Call:
## lm(formula = Domestic.box.office.gross ~
Speaking.parts.played.by.actors.of.color....)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -97057895 -45851607 -27541447  30180191 236372026
```

```
##
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                  99113551   15330926   6.465
## Speaking.parts.played.by.actors.of.color....  -230898     635150  -0.364
##                                              Pr(>|t|)
## (Intercept)                                  5.31e-08 ***
## Speaking.parts.played.by.actors.of.color....    0.718
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 76110000 on 47 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.002804,   Adjusted R-squared:  -0.01841
## F-statistic: 0.1322 on 1 and 47 DF,  p-value: 0.7178
```

```
abline(99113551, -230898)
```

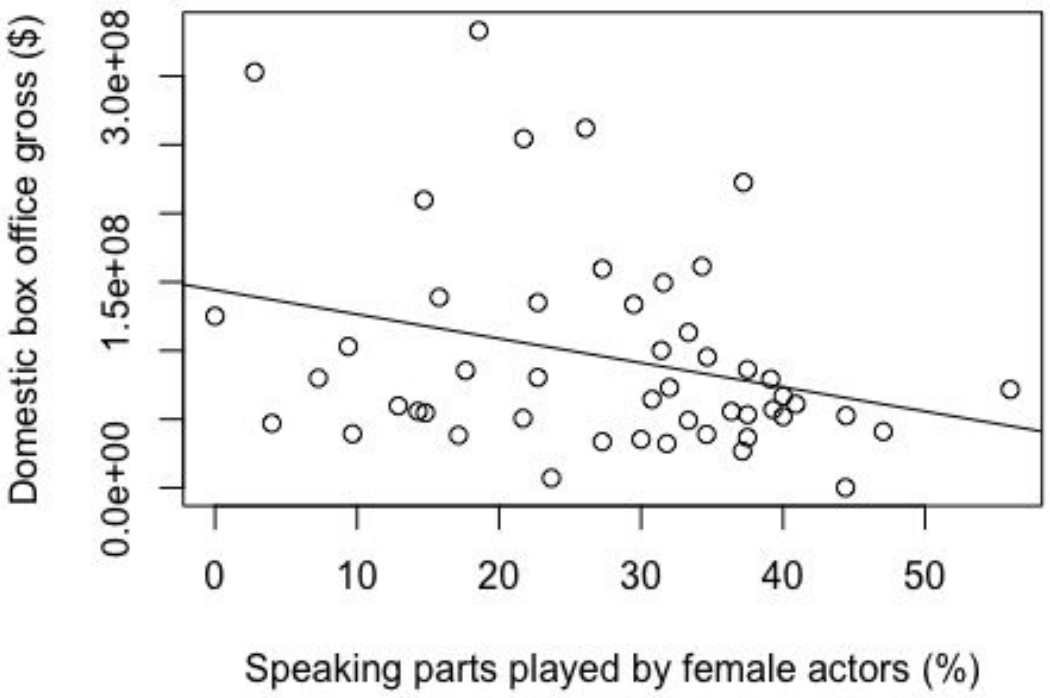


Made a boxplot with the percentage of speaking parts played by female actors in a film as the predictor and the film's domestic box office gross as the response, then fitted a linear model to this relationship and added that line to the plot. We also called the summary for this model to get information about it, like its coefficient and intercept estimates.

```
plot(Speaking.parts.played.by.female.actors...., Domestic.box.office.gross,
xlab = "Speaking parts played by female actors (%)",  ylab="Domestic box
office gross ($)")
summary(lm(Domestic.box.office.gross~Speaking.parts.played.by.female.actors
....))
```

```
##
## Call:
## lm(formula = Domestic.box.office.gross ~
Speaking.parts.played.by.female.actors....)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -95143627 -51569096 -19780349  26695961 221944308
##
## Coefficients:
##                                            Estimate Std. Error t value
## (Intercept)                                144129274   25244513   5.709
## Speaking.parts.played.by.female.actors....  -1771376     832519  -2.128
##                                            Pr(>|t|)
## (Intercept)                                7.39e-07 ***
## Speaking.parts.played.by.female.actors....   0.0386 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 72790000 on 47 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.08786,    Adjusted R-squared:  0.06845
## F-statistic: 4.527 on 1 and 47 DF,  p-value: 0.03863
```

```
abline(144129274, -1771376)
```



Defined GMmod to be the grand mean model for domestic box office gross (in US dollars), then called summary(GMmod) to get estimated value of the mean domestic box office gross, as well as other information about the model

```
GMmod <- lm(Domestic.box.office.gross~1)
summary(GMmod)
```

```
##
## Call:
## lm(formula = Domestic.box.office.gross ~ 1)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -93818090 -46416696 -28235966  28173070 239227002
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 93949598   10628150    8.84 1.02e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 75150000 on 49 degrees of freedom
```

Defined nonwhiteActors to be a linear model with percentage of speaking parts played by non-white actors in a film as the predictor and the film's domestic box office gross as the response, then called summary(nonwhiteActors) to get the intercept and coefficient estimates, as well as other information about the model

```
nonwhiteActors <-
lm(Domestic.box.office.gross~Speaking.parts.played.by.actors.of.color....)
summary(nonwhiteActors)
```

```
##
## Call:
## lm(formula = Domestic.box.office.gross ~
Speaking.parts.played.by.actors.of.color....)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -97057895 -45851607 -27541447  30180191 236372026
##
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                  99113551   15330926   6.465
## Speaking.parts.played.by.actors.of.color....  -230898     635150  -0.364
##                                              Pr(>|t|)
## (Intercept)                                  5.31e-08 ***
## Speaking.parts.played.by.actors.of.color....    0.718
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 76110000 on 47 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.002804,   Adjusted R-squared:  -0.01841
## F-statistic: 0.1322 on 1 and 47 DF,  p-value: 0.7178
```

Defined femaleActors to be a linear model with percentage of speaking parts played by female actors in a film as the predictor and the film's domestic box office gross as the response, then called summary(femaleActors) to get the intercept and coefficient estimates, as well as other information about the model

```
femaleActors <-
lm(Domestic.box.office.gross~Speaking.parts.played.by.female.actors....)
summary(femaleActors)
```

```
##
## Call:
## lm(formula = Domestic.box.office.gross ~
Speaking.parts.played.by.female.actors....)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -95143627 -51569096 -19780349  26695961 221944308
##
## Coefficients:
##                                            Estimate Std. Error t value
## (Intercept)                                144129274   25244513   5.709
## Speaking.parts.played.by.female.actors....  -1771376     832519  -2.128
##                                            Pr(>|t|)
## (Intercept)                                7.39e-07 ***
## Speaking.parts.played.by.female.actors....   0.0386 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 72790000 on 47 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.08786,    Adjusted R-squared:  0.06845
## F-statistic: 4.527 on 1 and 47 DF,  p-value: 0.03863
```