Stat 245: Modeling Diversity and PhD programs

Trevor VanVeldhuisen

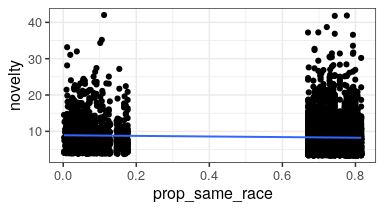
September 16, 2021

pi <- read\_csv('https://sldr.netlify.app/data/phd\_innovation.csv')

For this data set I will be looking at novelty, which is a measure of innovation and the proportion of same race. The response variable of interest is novelty, while the predictor variable is prop\_same\_race

slr <- lm(novelty ~ prop\_same\_race, data = pi)

gf\_point(novelty ~ prop\_same\_race, data = pi) %>%  
 gf\_lm()

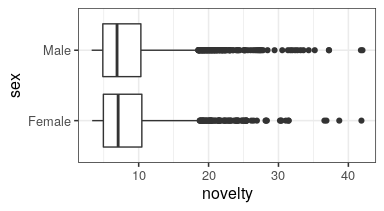


This model I chose to see if there was a relationship on a satter plot between novelty and prop\_same\_race. As we can see from the graph and the coefficients below, there is a relationship.

coef(slr)

## (Intercept) prop\_same\_race   
## 8.9330446 -0.8387549

gf\_boxplot(sex ~ novelty, data = pi)



For this box plot, I wanted to see if there was a difference in novelty between the two sexes. As we observe from the boxplot, the medians and both quartiles are similar if not the same, with just some variation in the outliers. This shows that novelty does not tend to be greater in either one sex.

pi <- pi %>%  
 mutate(pred = predict(slr))

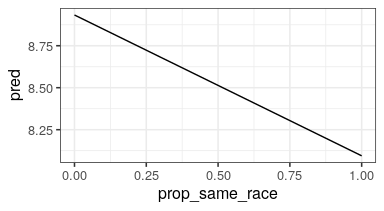
pi\_simple <- pi %>%  
 select(prop\_same\_race, novelty) %>%  
 mutate(preds = predict(slr))

pi <- pi %>%  
 select(prop\_same\_race) %>%  
 na.omit() %>%  
 mutate (pred = predict (slr))

fake\_data <- expand.grid(prop\_same\_race = seq(from = 0,  
 to = 1,  
 by = 0.1))  
fake\_data <- fake\_data %>%  
 mutate(pred = predict(slr, newdata = fake\_data))   
glimpse(fake\_data)

## Rows: 11  
## Columns: 2  
## $ prop\_same\_race <dbl> 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0  
## $ pred <dbl> 8.933045, 8.849169, 8.765294, 8.681418, 8.597543, 8.513…

gf\_line(pred ~ prop\_same\_race, data = fake\_data)



preds <- predict(slr, newdata = fake\_data, se.fit = TRUE)  
glimpse(preds)

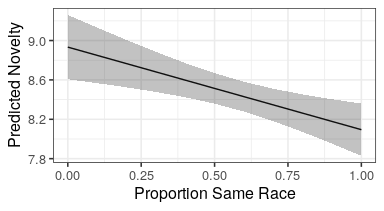
## List of 4  
## $ fit : Named num [1:11] 8.93 8.85 8.77 8.68 8.6 ...  
## ..- attr(\*, "names")= chr [1:11] "1" "2" "3" "4" ...  
## $ se.fit : Named num [1:11] 0.1664 0.144 0.1228 0.1039 0.0886 ...  
## ..- attr(\*, "names")= chr [1:11] "1" "2" "3" "4" ...  
## $ df : int 4193  
## $ residual.scale: num 4.98

fake\_data <- fake\_data %>%  
 mutate(pred = preds$fit,  
 pred.se = preds$se.fit)

fake\_data <- fake\_data %>%  
 mutate(CI\_lower = pred - 1.96\*pred.se,  
 CI\_upper = pred + 1.96\*pred.se)  
glimpse(fake\_data)

## Rows: 11  
## Columns: 5  
## $ prop\_same\_race <dbl> 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0  
## $ pred <dbl> 8.933045, 8.849169, 8.765294, 8.681418, 8.597543, 8.513…  
## $ pred.se <dbl> 0.16644194, 0.14396612, 0.12284926, 0.10392312, 0.08860…  
## $ CI\_lower <dbl> 8.606818, 8.566996, 8.524509, 8.477729, 8.423881, 8.358…  
## $ CI\_upper <dbl> 9.259271, 9.131343, 9.006078, 8.885107, 8.771204, 8.668…

gf\_line(pred ~ prop\_same\_race,  
 data = fake\_data) %>%  
 gf\_labs(y = 'Predicted Novelty',  
 x = 'Proportion Same Race') %>%  
 gf\_ribbon(CI\_lower + CI\_upper ~ prop\_same\_race)



Looking at this model, it shows a negative correlation between novelty and the proportion of the same race. From this data we can conclude that a more diverse group tends to be more innovative and ranks higher in novelty. Just the regression line was not enough, and to account for residual error, I incorporated confidence intervals at 95% as shown with the ribbons. This reinforces the conclusion of the graph, as the ribbons downtrend as well.