OLS Regression Project

David Purucker

Introduction

- two paragraphs - first: state research question and research importance • second: data source and methods

What is the relationship between adolescent academic achievement and popularity?

The plots do not show evidence of heteroscedasticity in the relationship between in-degree and GPA and in-degree and honor society membership. Heteroscedasticity is [define], and _____. A cone-shaped relationship between the fitted values and residuals would indicate a problem with heteroscedasticity.

```
##                   GVIF Df GVIF^(1/(2*Df))
## pseudoGPA     1.229370  1        1.108770
## honorsociety  1.129777  1        1.062910
## nsports       1.118302  1        1.057498
## race          1.136052  5        1.012838
## sex           1.104993  1        1.051187
## grade         1.116401  1        1.056599
## alcoholuse    1.154255  1        1.074363
## smoker        1.180082  1        1.086316
## bandchoir     1.079270  1        1.038879
## academicclub 1.078208  1        1.038368
## parentinc     1.101546  1        1.049546
```

VIF (variance inflation factor) applied to a test model with all variables shows low values for all the variables in this dataset. All VIF values are well below 4, indicating that there is no problem with multicollinearity between these variables.

```
addhealth$inc.cuberoot = addhealth$parentinc^(1/3)
```

This fixes the problem of $0 values

```
lm_svy_mi <- function(formula, imputations) {

  #setting up null objects allows us to easily add results
  #later
  b <- se <- R2 <- NULL

  #now loop through our imputations and run the model
  for(i in 1:imputations$m) {
    #grab the complete dataset
    imputation <- complete(imputations, i)
    #create the design effect object
    imputation.svy <- svydesign(ids=~cluster, weight=~sweight,
                                data=imputation)
    #run the model
    model <- svyglm(formula, design=imputation.svy)
    #collect the results
    b <- cbind(b, coef(model))
    se <- cbind(se, summary(model)$coef[,2])
    #We should get R squared too. Sadly, svyglm won't give
    #it to us by default, but we can get it from some of the
```

```
    #slots in the model output
    SSR <- sum((model$residuals)^2)
    SSY <- sum((model$y-mean(model$y))^2)
    R2 <- c(R2,1-SSR/SSY)
  }

  #now pool the results
  b.pool <- apply(b, 1, mean)
  between.var <- apply(b, 1, var)
  within.var <- apply(se^2, 1, mean)
  se.pool <- sqrt(within.var+between.var+between.var/imputations$m)
  t.pool <- b.pool/se.pool
  pvalue.pool <- (1-pnorm(abs(t.pool)))*2
  coefficients <- data.frame(b.pool, se.pool, t.pool, pvalue.pool)

  #lets take the mean R2 value
  r.squared <- mean(R2)
  #we can also grap n and p from the last model since
  #they should be the same across all iterations
  n <- nobs(model)
  p <- length(model$coefficients)-1
  #go ahead and calculate BIC.null
  bic.null <- n*log(1-r.squared)+p*log(n)

  #return everything in a list
  return(list(coef=coefficients,
              n=n,
              r.squared=r.squared,
              bic.null=bic.null))
}
convertModel <- function(model) {
  tr <- createTexreg(
    coef.names = rownames(model$coef),
    coef = model$coef$b.pool,
    se = model$coef$se.pool,
    pvalues = model$coef$pvalue.pool,
    gof.names = c("R2","BIC (null)","N"),
    gof = c(model$r.squared, model$bic.null, model$n),
    gof.decimal = c(T,F,F)
  )
}
```

```
## Warning: Number of logged events: 1
```

Section Header

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document.

Here is an embedded code chunk for stargazer output in pdf form:

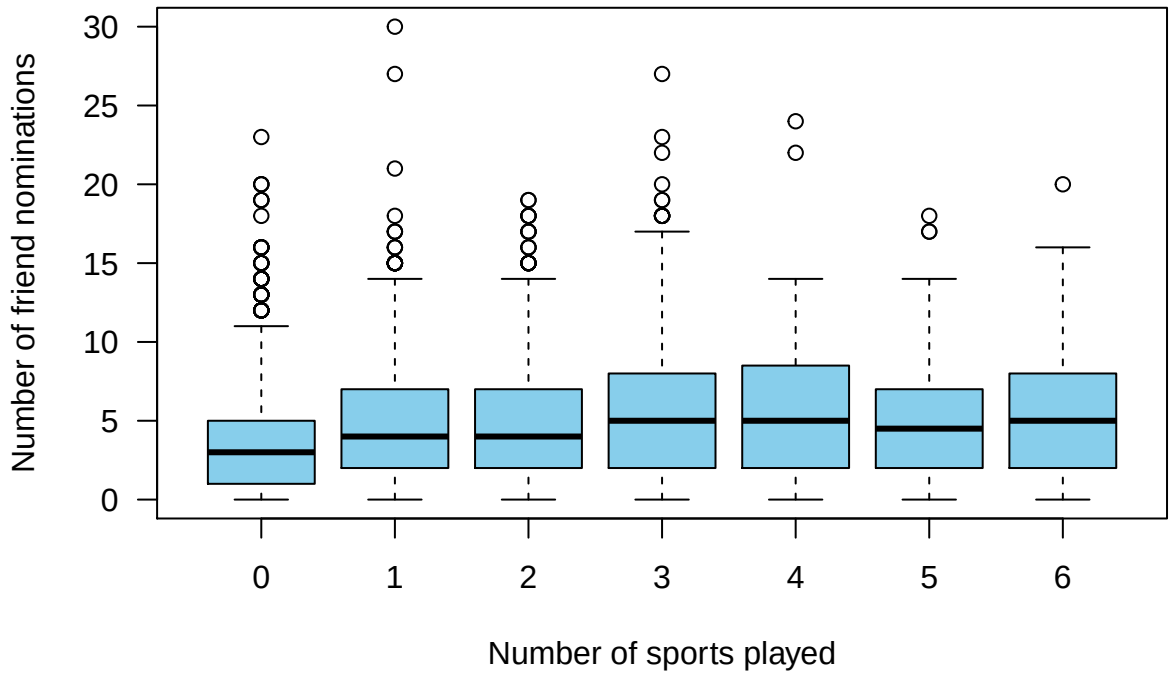

Figure 1: A Lovely Caption Goes Here

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	2.42 ***	0.81	1.44 ***	2.27 ***	2.53 ***	4.45
	(0.38)	***(0.42)***	***(0.41)***	***(0.32)***	***(0.32)***	***(0.12)***
parental income (cube root) 0.66	0.46 ***	0.44 ***				
	(0.11)	(0.11)	(0.11)			
number of sports played	0.46 ***	0.48 ***				
	(0.07)	(0.07)				
pseudo-GPA	0.63 ***	0.47 ***	0.83 ***	0.70 ***		
	(0.11)	(0.11)	(0.11)	(0.11)		
member of honor society	1.06 ***	1.31 ***	1.80	***(0.31)***	***(0.32)***	***(0.32)***
Male -0.51						
	(0.15)					
R2	0.02	0.06	0.07	0.03	0.04	0.02
BIC (null)	-86	-259	-286	-116	-146	-70
N	4397	4397	4397	4397	4397	4397

*** p < 0.001, ** p < 0.01, * p < 0.05

evaluate each model with adjusted R^2, F statistics, and BIC

- higher R2 good, higher F good, lower BIC good

choose the best models and apply the function to them

findings

The OLS regression models show that adolescent popularity is