bachelors.R

fridc

Sun Dec 16 17:21:25 2018

library(SDSRegressionR)

## Loading required package: plyr

## Loading required package: tidyverse

## -- Attaching packages ----------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.0.0 v purrr 0.2.5  
## v tibble 1.4.2 v dplyr 0.7.8  
## v tidyr 0.8.1 v stringr 1.3.1  
## v readr 1.1.1 v forcats 0.3.0

## -- Conflicts -------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::arrange() masks plyr::arrange()  
## x purrr::compact() masks plyr::compact()  
## x dplyr::count() masks plyr::count()  
## x dplyr::failwith() masks plyr::failwith()  
## x dplyr::filter() masks stats::filter()  
## x dplyr::id() masks plyr::id()  
## x dplyr::lag() masks stats::lag()  
## x dplyr::mutate() masks plyr::mutate()  
## x dplyr::rename() masks plyr::rename()  
## x dplyr::summarise() masks plyr::summarise()  
## x dplyr::summarize() masks plyr::summarize()

library(psych)

##   
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

Bach <- read.csv("mbk\_Bachelors1.csv")  
  
names(Bach)

## [1] "ï..Characteristic" "Race.ethnicity"   
## [3] "Years" "PercentageBA"   
## [5] "Standard.Error.on.Percentage" "Math"   
## [7] "math.error" "violent\_crime"   
## [9] "vc\_err" "High\_School"   
## [11] "Drugs" "Drugs\_Err"

table(Bach$Race.ethnicity)

##   
## Asian, non-Hispanic Black Hispanic   
## 14 14 14   
## White, non-Hispanic   
## 14

levels(Bach$Race.ethnicity)

## [1] "Asian, non-Hispanic" "Black " "Hispanic"   
## [4] "White, non-Hispanic"

#descriptive stats  
mean(Bach$PercentageBA)

## [1] 9.316071

sd(Bach$PercentageBA)

## [1] 5.951647

mean(Bach$Math)

## [1] 238.75

sd(Bach$Math)

## [1] 15.90226

mean(Bach$High\_School)

## [1] 27.92321

sd(Bach$High\_School)

## [1] 6.871099

mean(Bach$violent\_crime)

## [1] 4.641071

sd(Bach$violent\_crime)

## [1] 1.848614

mean(Bach$Drugs)

## [1] 20.12143

sd(Bach$Drugs)

## [1] 5.0231

Bach <- Bach %>%  
 mutate(Race = factor(Race.ethnicity, levels=c("White, non-Hispanic","Black ",   
 "Hispanic","Asian, non-Hispanic"),   
 labels = c("White","Black","Hispanic","Asian")))  
Bach$Race

## [1] White White White White White White White   
## [8] White White White White White White White   
## [15] Black Black Black Black Black Black Black   
## [22] Black Black Black Black Black Black Black   
## [29] Hispanic Hispanic Hispanic Hispanic Hispanic Hispanic Hispanic  
## [36] Hispanic Hispanic Hispanic Hispanic Hispanic Hispanic Hispanic  
## [43] Asian Asian Asian Asian Asian Asian Asian   
## [50] Asian Asian Asian Asian Asian Asian Asian   
## Levels: White Black Hispanic Asian

model <- lm(PercentageBA ~ Race + Math + violent\_crime + Drugs + High\_School, data=Bach)  
summary(model)

##   
## Call:  
## lm(formula = PercentageBA ~ Race + Math + violent\_crime + Drugs +   
## High\_School, data = Bach)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.9898 -0.4347 -0.0360 0.5612 2.3474   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.28815 6.75115 -0.635 0.52833   
## RaceBlack -3.30706 0.77096 -4.290 8.61e-05 \*\*\*  
## RaceHispanic -5.51805 0.64563 -8.547 3.33e-11 \*\*\*  
## RaceAsian 7.07085 1.50097 4.711 2.14e-05 \*\*\*  
## Math 0.06484 0.02153 3.012 0.00414 \*\*   
## violent\_crime 0.07623 0.17766 0.429 0.66978   
## Drugs 0.02704 0.06436 0.420 0.67625   
## High\_School -0.08368 0.08617 -0.971 0.33638   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.056 on 48 degrees of freedom  
## Multiple R-squared: 0.9725, Adjusted R-squared: 0.9685   
## F-statistic: 242.9 on 7 and 48 DF, p-value: < 2.2e-16

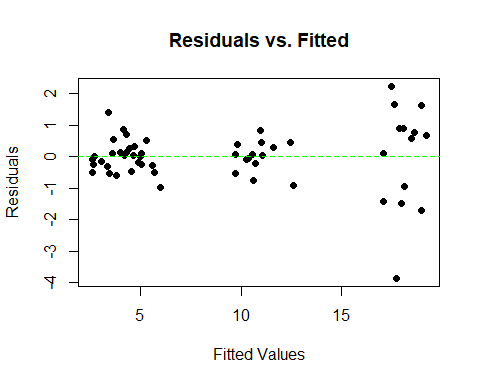
corr.test(Bach$violent\_crime, Bach$Drugs, use = "pairwise", method = "pearson")

## Call:corr.test(x = Bach$violent\_crime, y = Bach$Drugs, use = "pairwise",   
## method = "pearson")  
## Correlation matrix   
## [1] 0.56  
## Sample Size   
## [1] 56  
## Probability values adjusted for multiple tests.   
## [1] 0  
##   
## To see confidence intervals of the correlations, print with the short=FALSE option

#run and test interactions  
model\_int =lm(PercentageBA ~ Race + Math + violent\_crime + Drugs + High\_School  
 + Race\*High\_School ,   
 data=Bach)  
summary(model\_int)

##   
## Call:  
## lm(formula = PercentageBA ~ Race + Math + violent\_crime + Drugs +   
## High\_School + Race \* High\_School, data = Bach)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.8594 -0.3578 0.0396 0.4717 2.2265   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.82949 9.48192 1.459 0.15164   
## RaceBlack -23.16088 8.22032 -2.818 0.00716 \*\*  
## RaceHispanic -31.90625 11.08481 -2.878 0.00610 \*\*  
## RaceAsian -3.49702 5.97977 -0.585 0.56160   
## Math 0.03422 0.02374 1.441 0.15648   
## violent\_crime 0.05500 0.17582 0.313 0.75584   
## Drugs 0.11784 0.07039 1.674 0.10108   
## High\_School -0.52538 0.20760 -2.531 0.01495 \*   
## RaceBlack:High\_School 0.64170 0.26550 2.417 0.01977 \*   
## RaceHispanic:High\_School 0.86217 0.35791 2.409 0.02016 \*   
## RaceAsian:High\_School 0.38075 0.23954 1.590 0.11895   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.01 on 45 degrees of freedom  
## Multiple R-squared: 0.9765, Adjusted R-squared: 0.9712   
## F-statistic: 186.6 on 10 and 45 DF, p-value: < 2.2e-16

#check for homoskedasticity and outliers  
residFitted(model\_int)



library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:psych':  
##   
## logit

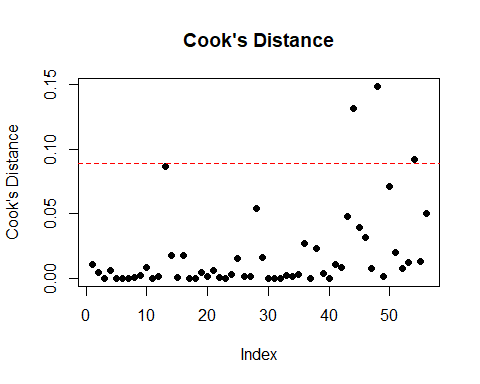
## The following object is masked from 'package:dplyr':  
##   
## recode

## The following object is masked from 'package:purrr':  
##   
## some

vif(model\_int)

## GVIF Df GVIF^(1/(2\*Df))  
## Race 5.429427e+07 3 19.459231  
## Math 7.693743e+00 1 2.773760  
## violent\_crime 5.700462e+00 1 2.387564  
## Drugs 6.747160e+00 1 2.597530  
## High\_School 1.098048e+02 1 10.478777  
## Race:High\_School 3.368617e+07 3 17.971126

cooksPlot(model\_int, key.variable="Years", print.obs=TRUE,  
 sort.obs=TRUE, save.cutoff=TRUE)



## Years PercentageBA Race Math violent\_crime Drugs High\_School  
## 1 2005 13.9 Asian 249 1.5 11.5 17.5  
## 2 2001 19.7 Asian 238 4.1 11.3 17.7  
## 3 2011 17.3 Asian 266 0.7 14.2 14.9  
## Predicted\_Y Cooks\_Distance F\_per  
## 1 17.75937 0.14851458 8.250912e-04  
## 2 17.47349 0.13135009 4.615516e-04  
## 3 18.99129 0.09162551 7.943108e-05

threeOuts(model\_int, key.variable="Years")

## Years PercentageBA Race Math violent\_crime Drugs High\_School  
## 1 2005 13.9 Asian 249 1.5 11.5 17.5  
## 2 2001 19.7 Asian 238 4.1 11.3 17.7  
## 3 2011 17.3 Asian 266 0.7 14.2 14.9  
## Predicted\_Y Student\_Resid Hat\_Values Cooks\_Distance F\_per inThree  
## 1 17.75937 -4.950251 NA 0.14851458 8.250912e-04 NA  
## 2 17.47349 2.609127 NA 0.13135009 4.615516e-04 NA  
## 3 18.99129 NA NA 0.09162551 7.943108e-05 NA

Anova(model\_int, type = "III")

## Anova Table (Type III tests)  
##   
## Response: PercentageBA  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 2.168 1 2.1273 0.151645   
## Race 18.027 3 5.8958 0.001752 \*\*  
## Math 2.117 1 2.0768 0.156479   
## violent\_crime 0.100 1 0.0979 0.755841   
## Drugs 2.856 1 2.8021 0.101078   
## High\_School 6.527 1 6.4044 0.014950 \*   
## Race:High\_School 7.616 3 2.4909 0.072268 .   
## Residuals 45.865 45   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#simple slopes time  
library(emmeans)  
ref\_grid(model\_int)

## 'emmGrid' object with variables:  
## Race = White, Black, Hispanic, Asian  
## Math = 238.75  
## violent\_crime = 4.6411  
## Drugs = 20.121  
## High\_School = 27.923

means <- emmeans(model\_int, "High\_School", at=list(High\_School = c(0,1)), by="Race")  
means

## Race = White:  
## High\_School emmean SE df lower.CL upper.CL  
## 0 24.625440 5.932662 45 12.676444 36.574435  
## 1 24.100058 5.725792 45 12.567722 35.632394  
##   
## Race = Black:  
## High\_School emmean SE df lower.CL upper.CL  
## 0 1.464564 5.233250 45 -9.075743 12.004872  
## 1 1.580885 5.081522 45 -8.653827 11.815596  
##   
## Race = Hispanic:  
## High\_School emmean SE df lower.CL upper.CL  
## 0 -7.280813 8.135866 45 -23.667288 9.105662  
## 1 -6.944030 7.884711 45 -22.824653 8.936594  
##   
## Race = Asian:  
## High\_School emmean SE df lower.CL upper.CL  
## 0 21.128420 2.234067 45 16.628779 25.628061  
## 1 20.983787 2.113430 45 16.727120 25.240453  
##   
## Confidence level used: 0.95

pairs(means, reverse = TRUE)

## Race = White:  
## contrast estimate SE df t.ratio p.value  
## 1 - 0 -0.5253818 0.2076048 45 -2.531 0.0150  
##   
## Race = Black:  
## contrast estimate SE df t.ratio p.value  
## 1 - 0 0.1163204 0.1523105 45 0.764 0.4490  
##   
## Race = Hispanic:  
## contrast estimate SE df t.ratio p.value  
## 1 - 0 0.3367835 0.2514278 45 1.339 0.1871  
##   
## Race = Asian:  
## contrast estimate SE df t.ratio p.value  
## 1 - 0 -0.1446333 0.1272529 45 -1.137 0.2617

slopes <- pairs(means, reverse = TRUE)  
  
  
#let's plot  
  
mns <- summary(emmeans(model\_int, "High\_School",   
 at=list(High\_School = seq(0,40,1)), by="Race"))  
simpleScatter(Bach, x=High\_School, y=PercentageBA, ptalpha = 0,  
 title="High School/GED and Bachelors degree percenage",  
 subtitle = "by Race") +  
 geom\_line(data=mns, aes(x=High\_School, y=emmean, color=Race)) +  
 ##geom\_ribbon(data=mns, aes(y=emmean, ymin=lower.CL, ymax=upper.CL, group=Race),   
 ## alpha=0.3) +   
 #Change to your group names and number of groups  
 scale\_colour\_manual(name = "Groups",   
 values =c("blue", "red", "green","yellow"),   
 #IMPORTANT: Same order below as the factor()  
 labels = c("White","Black","Hispanic","Asian"))

