Variables which effects on scores

kiyana mirbaghestan

2023-01-26

در سال 1980 آموزش و پرورش داده های مربوط به 11000 دبیرستان را جمع آوری کرده و براساس متغیرهایی مانند سطح درآمد و.... بهترین مدل را تفسیر کند که کدام دانش آموزان دارای کدام ویژگی نمره بیشتر و کدام نمره کمتری در امتحانات دریافت کرده است.

متغیرها به صورت زیر هستند:

Gender: جنسیت

ethnicity:قومیت

score:نمره (متغیر پاسخ)

Fcolleg :آیا پدر به کالج رفته؟

Mcollege:آیا مادر به کالج رفته؟

Home: آیا خانواده مالک خانه خود هستند؟

Urban: آیا مدرسه در منطقه شهری واقع شده است؟

Unemp:نرخ بیکاری در آن سال

Wage:دستمزد ساعتی دولت

Distance:فاصله از کالج طی4سال

Tuition:متوسط شهریه 4 ساله

Education:تعداد سال های تحصیل

Income:آیا درامد خانواده بالای 25000در سال هست؟

Region:نقطه جغرافیایی منطقه

دیتا را فراخوانی کرده ونوع داده را تشخیص می دهیم.

college<-read.csv("C:/Users/asus/Documents/CollegeDistance.csv")  
#preprocessing  
college<-as.data.frame(college)  
str(college)

## 'data.frame': 4739 obs. of 15 variables:  
## $ X : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ gender : chr "male" "female" "male" "male" ...  
## $ ethnicity: chr "other" "other" "other" "afam" ...  
## $ score : num 39.2 48.9 48.7 40.4 40.5 ...  
## $ fcollege : chr "yes" "no" "no" "no" ...  
## $ mcollege : chr "no" "no" "no" "no" ...  
## $ home : chr "yes" "yes" "yes" "yes" ...  
## $ urban : chr "yes" "yes" "yes" "yes" ...  
## $ unemp : num 6.2 6.2 6.2 6.2 5.6 ...  
## $ wage : num 8.09 8.09 8.09 8.09 8.09 ...  
## $ distance : num 0.2 0.2 0.2 0.2 0.4 ...  
## $ tuition : num 0.889 0.889 0.889 0.889 0.889 ...  
## $ education: int 12 12 12 12 13 12 13 15 13 15 ...  
## $ income : chr "high" "low" "low" "low" ...  
## $ region : chr "other" "other" "other" "other" ...

ستون اول که به نوعی ایدی است حذف می کنیم.

college<-college[,-1]  
summary(college)

## gender ethnicity score fcollege   
## Length:4739 Length:4739 Min. :28.95 Length:4739   
## Class :character Class :character 1st Qu.:43.92 Class :character   
## Mode :character Mode :character Median :51.19 Mode :character   
## Mean :50.89   
## 3rd Qu.:57.77   
## Max. :72.81   
## mcollege home urban unemp   
## Length:4739 Length:4739 Length:4739 Min. : 1.400   
## Class :character Class :character Class :character 1st Qu.: 5.900   
## Mode :character Mode :character Mode :character Median : 7.100   
## Mean : 7.597   
## 3rd Qu.: 8.900   
## Max. :24.900   
## wage distance tuition education   
## Min. : 6.590 Min. : 0.000 Min. :0.2575 Min. :12.00   
## 1st Qu.: 8.850 1st Qu.: 0.400 1st Qu.:0.4850 1st Qu.:12.00   
## Median : 9.680 Median : 1.000 Median :0.8245 Median :13.00   
## Mean : 9.501 Mean : 1.803 Mean :0.8146 Mean :13.81   
## 3rd Qu.:10.150 3rd Qu.: 2.500 3rd Qu.:1.1270 3rd Qu.:16.00   
## Max. :12.960 Max. :20.000 Max. :1.4042 Max. :18.00   
## income region   
## Length:4739 Length:4739   
## Class :character Class :character   
## Mode :character Mode :character   
##   
##   
##

head(college,10)

## gender ethnicity score fcollege mcollege home urban unemp wage distance  
## 1 male other 39.15 yes no yes yes 6.2 8.09 0.2  
## 2 female other 48.87 no no yes yes 6.2 8.09 0.2  
## 3 male other 48.74 no no yes yes 6.2 8.09 0.2  
## 4 male afam 40.40 no no yes yes 6.2 8.09 0.2  
## 5 female other 40.48 no no no yes 5.6 8.09 0.4  
## 6 male other 54.71 no no yes yes 5.6 8.09 0.4  
## 7 female other 56.07 no no yes no 7.2 8.85 0.4  
## 8 female other 54.85 no no yes no 7.2 8.85 0.4  
## 9 male other 64.74 yes no yes yes 5.9 8.09 3.0  
## 10 female other 56.06 no no yes yes 5.9 8.09 3.0  
## tuition education income region  
## 1 0.88915 12 high other  
## 2 0.88915 12 low other  
## 3 0.88915 12 low other  
## 4 0.88915 12 low other  
## 5 0.88915 13 low other  
## 6 0.88915 12 low other  
## 7 0.84988 13 low other  
## 8 0.84988 15 low other  
## 9 0.88915 13 low other  
## 10 0.88915 15 low other

آیا داده گمشده داریم؟خیر!

length(which(!complete.cases(college)))

## [1] 0

به جای اینکه جداگانه متغیرهای کاراکتری را به فکتور تبدیل کنیم از تابع زیر استفاده میکنیم، مثلا متغیر قومیت در این تابع به دو دسته تقسیم می شود.(1 به معنی اسپانیایی بودن یا افریقایی-امریکایی بودن است)

در جنسیت1 به معنی مرد بودن است.

پدر یا مادر به کالج رفته اند؟ بله1 خیر0

برای مالکیت خانه هم 1 به معنی مالک خانه بودن و 0 نبودن است.

منطقه شهری بودن یا نبودن 1 به معنای بله و0 به معنای خیر.

درامد1 به معنی زیاد و0 به معنی کم است.

منطقه1 به معنی غرب و 0 به معنی نواحی دیگر است.

#we can write function which can convert chr to factors  
cleanData = function(data) {  
 # convert the ethnicity label and remove the old one  
 data$hispanic = 1 \* (data$ethnicity == "hispanic")  
 data$afam = 1 \* (data$ethnicity == "afam")  
 data = data[-1 \* which(names(data) == "ethnicity")]  
   
 # convert the rest of the labels with automation  
 convert\_labels = c(c("fcollege", "yes"), c("mcollege", "yes"), c("home", "yes"),  
 c("urban", "yes"), c("income", "high"), c("region", "west"),  
 c("gender", "male"))  
   
 # loop through each label  
 for (label\_index in seq(1, length(convert\_labels), 2)) {  
   
 # get the column name and positive label name  
 col\_name = convert\_labels[label\_index]  
 positive\_label = convert\_labels[label\_index + 1]  
   
 # convert the label appropriately  
 data[col\_name] = 1 \* (data[col\_name] == positive\_label)  
 }  
   
 # return the data  
 return(data)  
}

حال داده ها را با تابع مان تست می کنیم:

collegedistance = cleanData(college)  
head(collegedistance)

## gender score fcollege mcollege home urban unemp wage distance tuition  
## 1 1 39.15 1 0 1 1 6.2 8.09 0.2 0.88915  
## 2 0 48.87 0 0 1 1 6.2 8.09 0.2 0.88915  
## 3 1 48.74 0 0 1 1 6.2 8.09 0.2 0.88915  
## 4 1 40.40 0 0 1 1 6.2 8.09 0.2 0.88915  
## 5 0 40.48 0 0 0 1 5.6 8.09 0.4 0.88915  
## 6 1 54.71 0 0 1 1 5.6 8.09 0.4 0.88915  
## education income region hispanic afam  
## 1 12 1 0 0 0  
## 2 12 0 0 0 0  
## 3 12 0 0 0 0  
## 4 12 0 0 0 1  
## 5 13 0 0 0 0  
## 6 12 0 0 0 0

گرد شده ماتریس همبستگی:

college\_cor = round(cor(collegedistance), 2)  
  
# summary of the correlation  
head(college\_cor)

## gender score fcollege mcollege home urban unemp wage distance  
## gender 1.00 0.08 0.04 0.02 0.04 0.01 -0.03 0.03 0.00  
## score 0.08 1.00 0.25 0.19 0.13 -0.09 -0.03 0.12 -0.07  
## fcollege 0.04 0.25 1.00 0.43 0.08 -0.05 -0.10 0.03 -0.11  
## mcollege 0.02 0.19 0.43 1.00 0.06 -0.03 -0.09 0.02 -0.08  
## home 0.04 0.13 0.08 0.06 1.00 -0.10 0.01 0.07 0.02  
## urban 0.01 -0.09 -0.05 -0.03 -0.10 1.00 -0.05 -0.03 -0.29  
## tuition education income region hispanic afam  
## gender 0.01 0.01 0.06 -0.01 0.02 -0.04  
## score 0.13 0.47 0.18 -0.03 -0.16 -0.28  
## fcollege 0.03 0.28 0.35 0.03 -0.11 -0.10  
## mcollege 0.04 0.23 0.25 -0.01 -0.11 -0.02  
## home 0.00 0.10 0.14 0.00 -0.06 -0.13  
## urban -0.02 -0.02 -0.07 -0.05 0.08 0.18

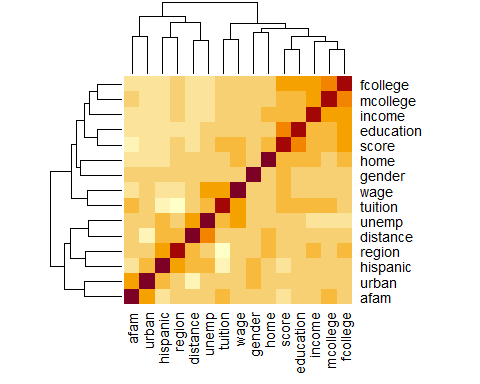
#install.packages("ggheatmap")  
library(ggheatmap)

## Warning: package 'ggheatmap' was built under R version 4.2.2

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.2.2

heatmap(college\_cor)

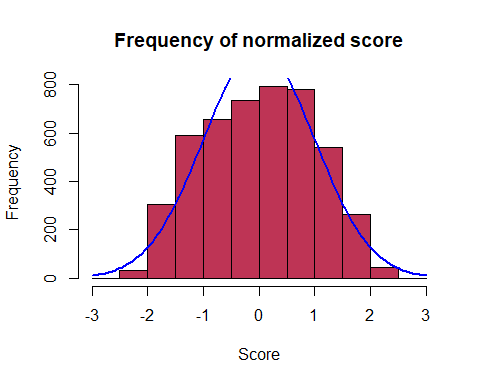


نمودار بالا به نوعی ماتریس همبستگی را نشان می دهد.

آیا توزیع متغیر وابسته مان نرمال است یا نه؟که جواب تقریبا مثبت است و متغیر وابسته نیازی به تبدیل ندارد.

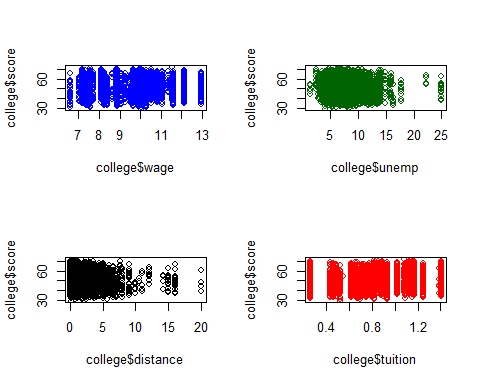
#Histogram  
f.hist<-hist((collegedistance$score-mean(collegedistance$score))/sd(collegedistance$score),col="#BE3455",xlab = "Score",main ="Frequency of normalized score")#col for 2023  
multiplier = f.hist$counts / f.hist$density  
x = seq(-3, 3, length.out = length(collegedistance$score-mean(collegedistance$score))/sd(collegedistance$score))  
curve(multiplier \* dnorm(x), col = "blue", lwd=2, add=TRUE, yaxt="n")

## Warning in multiplier \* dnorm(x): longer object length is not a multiple of  
## shorter object length



آیا هر تبدیل عددی برای متغیر مستقل ضروری است یا نه؟

#Scatter plot  
par(mfrow=c(2,2))  
plot(college$wage, college$score, col = "blue")  
plot(college$unemp, college$score, col = "darkgreen")  
plot(college$distance, college$score, col = "black")  
plot(college$tuition, college$score, col = "red")



از نمو دارهای بالا میفهمیم متاسفانه داده ها کاملا تصادفی اند و روندی بین آنها نداریم و تغییری هم نیاز نیست. به جای معیارهای نمودار به معیارهای آماری تکیه خواهیم کرد.

حال مدلی ساده برازش می دهیم فرض نرمالیتی رد اما فرض برابر بودن میانگین ها با آزمون ویلکاکسون قبول است.

simple\_add = lm(score ~ ., data = collegedistance)  
shapiro.test(resid(simple\_add))

##   
## Shapiro-Wilk normality test  
##   
## data: resid(simple\_add)  
## W = 0.99801, p-value = 8.988e-06

wilcox.test(resid(simple\_add))

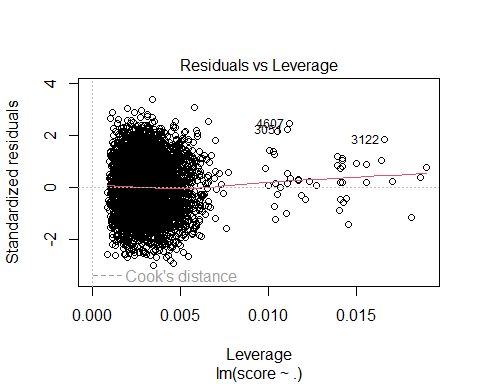
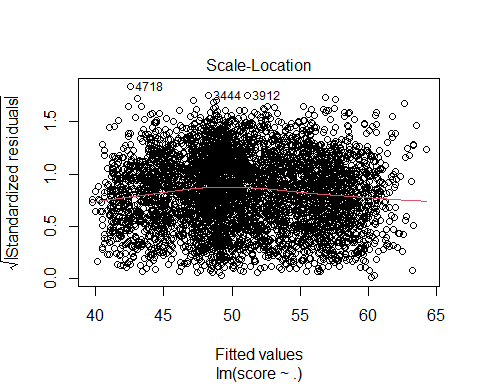
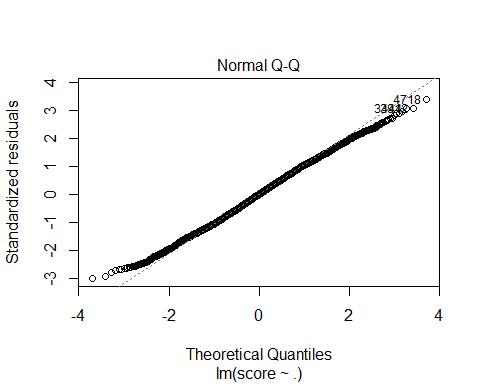
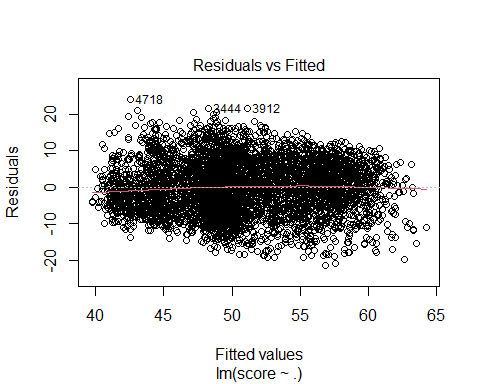
##   
## Wilcoxon signed rank test with continuity correction  
##   
## data: resid(simple\_add)  
## V = 5613392, p-value = 0.9803  
## alternative hypothesis: true location is not equal to 0

# Start with a model with all predictors.  
fullscoremodel = lm(score ~ ., data = collegedistance)  
summary(fullscoremodel)

##   
## Call:  
## lm(formula = score ~ ., data = collegedistance)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.2725 -5.0863 0.0051 5.0579 23.9730   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 21.48313 1.17176 18.334 < 2e-16 \*\*\*  
## gender 1.09852 0.20808 5.279 1.36e-07 \*\*\*  
## fcollege 1.31305 0.30043 4.371 1.27e-05 \*\*\*  
## mcollege 1.05432 0.33665 3.132 0.00175 \*\*   
## home 0.70834 0.27526 2.573 0.01010 \*   
## urban -0.34560 0.26250 -1.317 0.18805   
## unemp -0.07889 0.04195 -1.881 0.06007 .   
## wage 0.19602 0.08499 2.306 0.02113 \*   
## distance -0.11499 0.05030 -2.286 0.02230 \*   
## tuition 2.22984 0.41516 5.371 8.20e-08 \*\*\*  
## education 1.93818 0.06142 31.558 < 2e-16 \*\*\*  
## income -0.16482 0.25088 -0.657 0.51123   
## region 0.68419 0.32437 2.109 0.03497 \*   
## hispanic -3.60741 0.29186 -12.360 < 2e-16 \*\*\*  
## afam -6.29323 0.30112 -20.899 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.096 on 4724 degrees of freedom  
## Multiple R-squared: 0.3369, Adjusted R-squared: 0.335   
## F-statistic: 171.5 on 14 and 4724 DF, p-value: < 2.2e-16

مقدار آر2 وآر2 اجاست هردو کم است لذا مدل قابل قبولی نداریم.

plot(fullscoremodel)



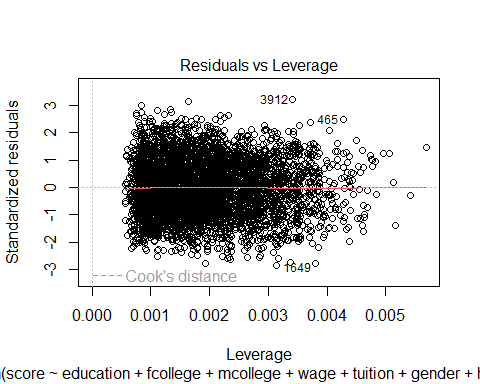
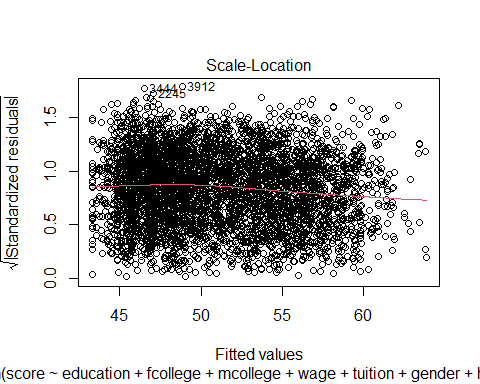
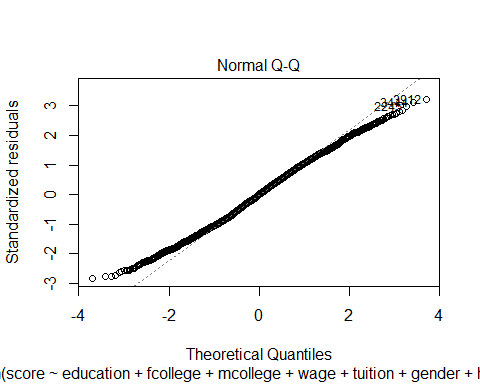
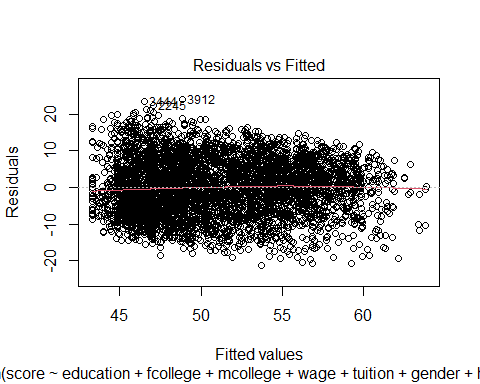
حال مدل را با استفاده از متغیرهای مهم برازش می دهیم(متغیرهای مهم در نمودار مربوط به همبستگی ها آشکارند)

fullimportantmodel<-lm(score~education + fcollege + mcollege + wage + tuition +gender + home + income,data=collegedistance)  
summary(fullimportantmodel)

##   
## Call:  
## lm(formula = score ~ education + fcollege + mcollege + wage +   
## tuition + gender + home + income, data = collegedistance)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.2044 -5.7175 0.0191 5.4571 23.9723   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 14.81236 1.16800 12.682 < 2e-16 \*\*\*  
## education 2.00346 0.06453 31.048 < 2e-16 \*\*\*  
## fcollege 2.07041 0.31449 6.583 5.10e-11 \*\*\*  
## mcollege 1.01711 0.35443 2.870 0.00413 \*\*   
## wage 0.44084 0.08582 5.137 2.90e-07 \*\*\*  
## tuition 2.20413 0.33866 6.508 8.39e-11 \*\*\*  
## gender 1.13651 0.21930 5.183 2.28e-07 \*\*\*  
## home 1.49600 0.28766 5.201 2.07e-07 \*\*\*  
## income 0.41389 0.26274 1.575 0.11525   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.492 on 4730 degrees of freedom  
## Multiple R-squared: 0.26, Adjusted R-squared: 0.2587   
## F-statistic: 207.7 on 8 and 4730 DF, p-value: < 2.2e-16

آر2 وآر2 اجاست همچنان مقادیر کوچکی دارند.حتی نسبت به قبل کمتر شده پس مدل بهبود نیافته است.

plot(fullimportantmodel)



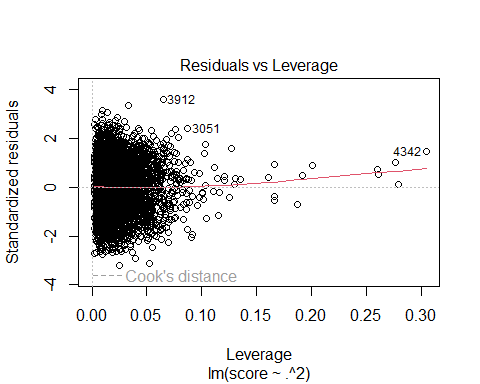
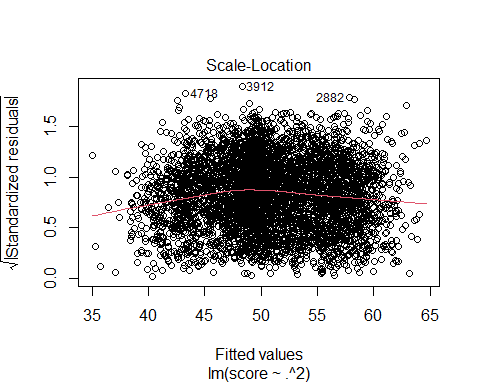
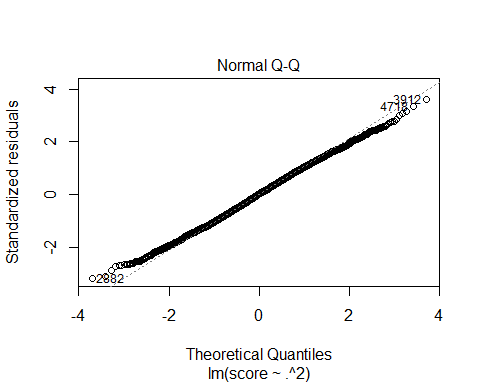
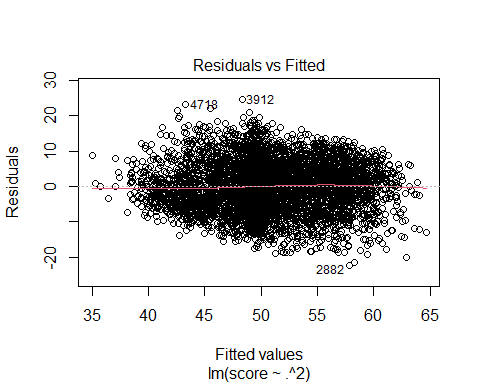
پس مدل با اثرات متقابل را بررسی می کنیم.

fullintmodel<-lm(score ~ .^2, data = collegedistance)  
summary(fullintmodel)

##   
## Call:  
## lm(formula = score ~ .^2, data = collegedistance)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -22.2167 -5.0308 0.1524 4.9649 24.5201   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.204e+01 8.332e+00 2.645 0.008188 \*\*   
## gender -9.079e+00 2.352e+00 -3.860 0.000115 \*\*\*  
## fcollege 3.122e-01 3.451e+00 0.090 0.927919   
## mcollege 3.419e+00 3.918e+00 0.873 0.382922   
## home 2.520e+00 3.199e+00 0.788 0.430979   
## urban -1.135e+00 3.061e+00 -0.371 0.710887   
## unemp -5.489e-02 4.900e-01 -0.112 0.910806   
## wage 7.496e-01 8.612e-01 0.870 0.384152   
## distance -9.844e-01 6.150e-01 -1.601 0.109534   
## tuition -3.050e-01 5.641e+00 -0.054 0.956880   
## education 2.070e+00 5.087e-01 4.069 4.81e-05 \*\*\*  
## income 6.654e-02 2.905e+00 0.023 0.981730   
## region 9.873e-01 3.822e+00 0.258 0.796171   
## hispanic -4.722e+00 3.809e+00 -1.240 0.215190   
## afam -1.534e+01 3.375e+00 -4.545 5.62e-06 \*\*\*  
## gender:fcollege 5.151e-01 6.076e-01 0.848 0.396579   
## gender:mcollege 4.253e-01 6.896e-01 0.617 0.537469   
## gender:home -7.959e-01 5.597e-01 -1.422 0.155085   
## gender:urban 9.332e-01 5.335e-01 1.749 0.080326 .   
## gender:unemp 6.097e-02 8.576e-02 0.711 0.477144   
## gender:wage 2.769e-01 1.720e-01 1.610 0.107537   
## gender:distance 1.072e-01 1.046e-01 1.025 0.305410   
## gender:tuition 1.122e+00 8.437e-01 1.330 0.183584   
## gender:education 4.116e-01 1.243e-01 3.310 0.000939 \*\*\*  
## gender:income -2.751e-01 5.058e-01 -0.544 0.586593   
## gender:region 8.417e-01 6.560e-01 1.283 0.199533   
## gender:hispanic 1.265e+00 5.928e-01 2.135 0.032834 \*   
## gender:afam 1.357e+00 6.158e-01 2.203 0.027660 \*   
## fcollege:mcollege -4.038e-02 7.216e-01 -0.056 0.955377   
## fcollege:home 3.406e-01 8.730e-01 0.390 0.696424   
## fcollege:urban -9.868e-02 7.872e-01 -0.125 0.900253   
## fcollege:unemp -4.061e-02 1.221e-01 -0.332 0.739567   
## fcollege:wage 1.380e-01 2.529e-01 0.546 0.585366   
## fcollege:distance 5.878e-02 1.620e-01 0.363 0.716678   
## fcollege:tuition -1.279e+00 1.276e+00 -1.002 0.316157   
## fcollege:education -8.578e-03 1.754e-01 -0.049 0.961001   
## fcollege:income 9.486e-01 6.279e-01 1.511 0.130877   
## fcollege:region -8.400e-01 9.725e-01 -0.864 0.387753   
## fcollege:hispanic 1.287e+00 9.469e-01 1.359 0.174077   
## fcollege:afam 1.283e-02 9.963e-01 0.013 0.989726   
## mcollege:home -2.876e-01 1.024e+00 -0.281 0.778829   
## mcollege:urban 9.112e-01 8.771e-01 1.039 0.298923   
## mcollege:unemp 5.298e-02 1.483e-01 0.357 0.720924   
## mcollege:wage 2.670e-02 2.798e-01 0.095 0.923985   
## mcollege:distance -1.490e-01 1.979e-01 -0.753 0.451568   
## mcollege:tuition -2.547e-01 1.486e+00 -0.171 0.863886   
## mcollege:education -2.100e-01 2.006e-01 -1.047 0.295103   
## mcollege:income 1.597e+00 7.204e-01 2.217 0.026661 \*   
## mcollege:region -8.507e-01 1.145e+00 -0.743 0.457404   
## mcollege:hispanic -1.446e+00 1.165e+00 -1.241 0.214593   
## mcollege:afam -3.069e-01 9.949e-01 -0.308 0.757769   
## home:urban 1.450e+00 6.715e-01 2.159 0.030864 \*   
## home:unemp -1.869e-01 1.186e-01 -1.576 0.115123   
## home:wage -4.892e-02 2.373e-01 -0.206 0.836727   
## home:distance 1.985e-01 1.315e-01 1.510 0.131194   
## home:tuition 1.608e+00 1.134e+00 1.417 0.156464   
## home:education -1.646e-01 1.684e-01 -0.977 0.328406   
## home:income 8.633e-01 7.653e-01 1.128 0.259370   
## home:region 1.198e+00 8.878e-01 1.349 0.177367   
## home:hispanic 3.596e-01 7.404e-01 0.486 0.627271   
## home:afam 2.068e-01 7.316e-01 0.283 0.777453   
## urban:unemp -1.719e-01 1.395e-01 -1.232 0.217941   
## urban:wage 1.826e-01 2.299e-01 0.794 0.426991   
## urban:distance 2.945e-01 3.786e-01 0.778 0.436642   
## urban:tuition 4.307e-01 1.105e+00 0.390 0.696660   
## urban:education -1.055e-01 1.582e-01 -0.667 0.504773   
## urban:income -7.150e-01 6.639e-01 -1.077 0.281565   
## urban:region -2.524e-01 8.894e-01 -0.284 0.776618   
## urban:hispanic -4.032e-01 7.346e-01 -0.549 0.583092   
## urban:afam -7.518e-01 7.066e-01 -1.064 0.287382   
## unemp:wage -1.183e-02 3.457e-02 -0.342 0.732092   
## unemp:distance 2.762e-02 1.205e-02 2.293 0.021882 \*   
## unemp:tuition -1.294e-02 1.816e-01 -0.071 0.943176   
## unemp:education 7.587e-03 2.602e-02 0.292 0.770619   
## unemp:income -5.600e-02 1.047e-01 -0.535 0.592642   
## unemp:region -3.870e-05 1.433e-01 0.000 0.999784   
## unemp:hispanic 1.590e-01 1.216e-01 1.307 0.191287   
## unemp:afam 1.538e-01 1.503e-01 1.023 0.306302   
## wage:distance 3.309e-02 4.523e-02 0.732 0.464365   
## wage:tuition -1.738e-01 4.527e-01 -0.384 0.701075   
## wage:education -3.668e-02 5.213e-02 -0.704 0.481724   
## wage:income -5.057e-02 2.097e-01 -0.241 0.809473   
## wage:region 3.697e-01 2.822e-01 1.310 0.190265   
## wage:hispanic -3.107e-01 2.736e-01 -1.136 0.256165   
## wage:afam 5.278e-01 2.513e-01 2.100 0.035742 \*   
## distance:tuition 1.499e-01 2.329e-01 0.643 0.519943   
## distance:education 7.924e-03 3.225e-02 0.246 0.805887   
## distance:income -1.359e-01 1.330e-01 -1.022 0.306951   
## distance:region 1.548e-01 1.318e-01 1.175 0.240222   
## distance:hispanic -2.475e-01 1.439e-01 -1.719 0.085640 .   
## distance:afam -3.500e-01 1.929e-01 -1.815 0.069613 .   
## tuition:education 1.790e-01 2.556e-01 0.700 0.483747   
## tuition:income 1.301e+00 1.076e+00 1.210 0.226416   
## tuition:region -5.524e+00 1.366e+00 -4.044 5.34e-05 \*\*\*  
## tuition:hispanic 1.965e+00 1.175e+00 1.672 0.094538 .   
## tuition:afam 2.441e-01 1.297e+00 0.188 0.850730   
## education:income -1.318e-01 1.460e-01 -0.903 0.366824   
## education:region -2.214e-01 2.060e-01 -1.075 0.282556   
## education:hispanic 3.536e-02 1.747e-01 0.202 0.839586   
## education:afam 1.945e-01 1.902e-01 1.022 0.306676   
## income:region 1.319e+00 8.338e-01 1.582 0.113746   
## income:hispanic 1.444e+00 7.826e-01 1.845 0.065118 .   
## income:afam 7.929e-01 8.061e-01 0.984 0.325346   
## region:hispanic 6.329e-01 8.244e-01 0.768 0.442682   
## region:afam 4.102e-01 1.321e+00 0.311 0.756157   
## hispanic:afam NA NA NA NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.05 on 4634 degrees of freedom  
## Multiple R-squared: 0.3581, Adjusted R-squared: 0.3437   
## F-statistic: 24.86 on 104 and 4634 DF, p-value: < 2.2e-16

باز هم شاهد آر2 وآر2 اجاست بسیار کوچکی هستیم.

plot(fullintmodel)



حال دو مدل را با یکدیگر توسط جدول آنووا تست می کنیم.

anova(fullscoremodel, fullintmodel)

## Analysis of Variance Table  
##   
## Model 1: score ~ gender + fcollege + mcollege + home + urban + unemp +   
## wage + distance + tuition + education + income + region +   
## hispanic + afam  
## Model 2: score ~ (gender + fcollege + mcollege + home + urban + unemp +   
## wage + distance + tuition + education + income + region +   
## hispanic + afam)^2  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 4724 237888   
## 2 4634 230293 90 7595 1.6981 5.067e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

با دیدن ستون اخر یعنی 5.067 با ارقام اعشار میفهمیم هیچ کدام از مدل ها ذخوب نیستند.مدل دوم را انتخاب می کنیم تا پیش بینی ها را محدود کنیم.

با روش پیشرو پارامترها را جستجو و محدود خواهیم کرد.

distanceaic = step(fullintmodel, direction = "backward", trace = 0)  
coef(distanceaic)

## (Intercept) gender fcollege mcollege   
## 27.51578466 -8.45366574 0.81384297 3.99964110   
## home urban unemp wage   
## 0.18519034 -2.03672698 -0.06018998 -0.10246390   
## distance tuition education income   
## -0.44747309 1.37291771 1.80866472 -0.97737050   
## region hispanic afam gender:home   
## 2.97861345 -5.12992820 -13.38382092 -0.86720151   
## gender:urban gender:wage gender:education gender:hispanic   
## 0.77074554 0.38898937 0.43096415 1.14001836   
## gender:afam fcollege:income mcollege:education mcollege:income   
## 1.27223670 0.84145397 -0.25943771 1.72731246   
## home:urban home:unemp home:distance home:tuition   
## 1.48638416 -0.19985687 0.18255690 1.69659493   
## home:region unemp:distance wage:region wage:afam   
## 1.36196167 0.03505865 0.42560855 0.76479061   
## distance:hispanic distance:afam tuition:region tuition:hispanic   
## -0.18692067 -0.34256326 -5.12187834 1.54255097   
## education:region income:hispanic   
## -0.35646928 1.40776578

حال که ضرایب اثگذار خود را داریم می توانیم مدل بزرگتری ایجاد و ای آی سی را اجرا کنیم.

distance\_larger = lm(score ~ .^2 + I(gender^2) + I(fcollege^2) + I(mcollege^2) +I(home^2) + I(urban^2) + I(unemp^2) + I(wage^2) +  
 I(distance^2) + I(tuition^2) + I(education^2) +  
 I(hispanic^2) + I(afam^2),   
 data = collegedistance)  
distanceaic2 = step(distance\_larger, direction = "backward", trace = 0)  
coef(distanceaic2)

## (Intercept) gender fcollege mcollege   
## -6.92257098 -8.81842187 0.75287995 0.24920943   
## home urban unemp wage   
## 0.16661740 -0.70409981 0.03838828 -0.08707851   
## distance tuition education income   
## -0.41932121 1.02692976 6.77157103 -0.99153378   
## region hispanic afam I(distance^2)   
## 1.36351911 -5.84466691 -13.55114414 0.02622444   
## I(education^2) gender:home gender:urban gender:wage   
## -0.17663289 -0.79186265 0.76748589 0.28024332   
## gender:tuition gender:education gender:region gender:hispanic   
## 1.25447229 0.43641796 1.05006853 1.25172546   
## gender:afam fcollege:income mcollege:income home:urban   
## 1.31295054 0.89990994 1.65491986 1.45738018   
## home:unemp home:distance home:tuition home:region   
## -0.20758448 0.22831047 1.64485020 1.29236644   
## urban:unemp unemp:hispanic wage:region wage:afam   
## -0.19133902 0.22898793 0.54209956 0.76591562   
## distance:hispanic distance:afam tuition:region education:region   
## -0.26260115 -0.33830105 -5.28140769 -0.35341008   
## income:hispanic   
## 1.55536706

anova(distanceaic,distanceaic2, test = "F")

## Analysis of Variance Table  
##   
## Model 1: score ~ gender + fcollege + mcollege + home + urban + unemp +   
## wage + distance + tuition + education + income + region +   
## hispanic + afam + gender:home + gender:urban + gender:wage +   
## gender:education + gender:hispanic + gender:afam + fcollege:income +   
## mcollege:education + mcollege:income + home:urban + home:unemp +   
## home:distance + home:tuition + home:region + unemp:distance +   
## wage:region + wage:afam + distance:hispanic + distance:afam +   
## tuition:region + tuition:hispanic + education:region + income:hispanic  
## Model 2: score ~ gender + fcollege + mcollege + home + urban + unemp +   
## wage + distance + tuition + education + income + region +   
## hispanic + afam + I(distance^2) + I(education^2) + gender:home +   
## gender:urban + gender:wage + gender:tuition + gender:education +   
## gender:region + gender:hispanic + gender:afam + fcollege:income +   
## mcollege:income + home:urban + home:unemp + home:distance +   
## home:tuition + home:region + urban:unemp + unemp:hispanic +   
## wage:region + wage:afam + distance:hispanic + distance:afam +   
## tuition:region + education:region + income:hispanic  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 4701 232156   
## 2 4698 231051 3 1104.7 7.487 5.36e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

با بررسی ضرایب، ای آی سی و جداول در نهایت به مدل پیچیده زیر میرسیم. اما هنگام استفاده از آنوای مجدد و غیره ازیک سری از پارامترها فاکتور خواهیم گرفت.

# create the model with large parameters  
large\_param\_model = lm(score ~ gender + fcollege + fcollege:mcollege + home +   
 urban + wage + distance + I(distance^(0.25)) +  
 tuition + education + income + hispanic + afam +  
 I(distance^2) + I(tuition^2) + I(education^2) +  
 gender:home + gender:wage + gender:education +  
 gender:hispanic + gender:afam + fcollege:income +  
 mcollege:income + home:urban + home:unemp +  
 home:distance + unemp:distance +  
 unemp:hispanic + wage:hispanic + wage:afam +  
 distance:hispanic + distance:afam +   
 tuition:education + income:hispanic + I(tuition^3) +  
 log(tuition) + I(education^3)+ log(education) +  
 I(tuition^0.25) + I(education^0.25) + I(wage^2) +  
 gender:I(wage^2) + gender:I(log(wage)) +  
 hispanic:I(wage^2) + hispanic:I(log(wage)) +  
 home:I(unemp^2) + home:I(unemp^0.25) +  
 home:I(log(unemp)) + home:I(distance^2) +  
 home:I(distance^0.25) + unemp:I(distance^2) +  
 unemp:I(distance^0.25) + hispanic:I(distance^2) +  
 hispanic:I(distance^0.25) + afam:I(distance^2) +  
 afam:I(distance^0.25) + region,   
 data = collegedistance)  
summary(large\_param\_model)

##   
## Call:  
## lm(formula = score ~ gender + fcollege + fcollege:mcollege +   
## home + urban + wage + distance + I(distance^(0.25)) + tuition +   
## education + income + hispanic + afam + I(distance^2) + I(tuition^2) +   
## I(education^2) + gender:home + gender:wage + gender:education +   
## gender:hispanic + gender:afam + fcollege:income + mcollege:income +   
## home:urban + home:unemp + home:distance + unemp:distance +   
## unemp:hispanic + wage:hispanic + wage:afam + distance:hispanic +   
## distance:afam + tuition:education + income:hispanic + I(tuition^3) +   
## log(tuition) + I(education^3) + log(education) + I(tuition^0.25) +   
## I(education^0.25) + I(wage^2) + gender:I(wage^2) + gender:I(log(wage)) +   
## hispanic:I(wage^2) + hispanic:I(log(wage)) + home:I(unemp^2) +   
## home:I(unemp^0.25) + home:I(log(unemp)) + home:I(distance^2) +   
## home:I(distance^0.25) + unemp:I(distance^2) + unemp:I(distance^0.25) +   
## hispanic:I(distance^2) + hispanic:I(distance^0.25) + afam:I(distance^2) +   
## afam:I(distance^0.25) + region, data = collegedistance)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -22.051 -5.059 0.010 4.956 23.690   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.718e+06 2.357e+06 -0.729 0.466213   
## gender -1.986e+02 9.744e+01 -2.038 0.041581 \*   
## fcollege 8.158e-01 4.429e-01 1.842 0.065551 .   
## home -8.661e+01 1.837e+02 -0.471 0.637349   
## urban -1.812e+00 5.775e-01 -3.137 0.001717 \*\*   
## wage 1.031e-01 1.304e+00 0.079 0.937013   
## distance -2.260e-01 6.981e-01 -0.324 0.746152   
## I(distance^(0.25)) -1.392e+00 2.313e+00 -0.602 0.547257   
## tuition 2.248e+03 1.439e+03 1.563 0.118168   
## education -3.969e+04 5.106e+04 -0.777 0.437004   
## income -1.005e+00 3.184e-01 -3.157 0.001603 \*\*   
## hispanic -2.138e+02 2.147e+02 -0.996 0.319312   
## afam -1.122e+01 2.633e+00 -4.263 2.06e-05 \*\*\*  
## I(distance^2) -2.267e-03 4.408e-02 -0.051 0.958982   
## I(tuition^2) -6.348e+02 4.274e+02 -1.485 0.137587   
## I(education^2) 6.274e+02 7.435e+02 0.844 0.398849   
## I(tuition^3) 1.081e+02 7.806e+01 1.385 0.166166   
## log(tuition) 1.525e+03 9.518e+02 1.602 0.109122   
## I(education^3) -6.484e+00 7.112e+00 -0.912 0.361983   
## log(education) -5.329e+05 7.480e+05 -0.712 0.476222   
## I(tuition^0.25) -1.131e+04 7.082e+03 -1.597 0.110274   
## I(education^0.25) 1.853e+06 2.543e+06 0.729 0.466321   
## I(wage^2) -7.315e-03 6.751e-02 -0.108 0.913729   
## region -3.617e-01 5.034e-01 -0.718 0.472547   
## fcollege:mcollege 1.318e-01 5.836e-01 0.226 0.821302   
## gender:home -9.017e-01 5.487e-01 -1.643 0.100379   
## gender:wage -5.680e+01 2.753e+01 -2.063 0.039132 \*   
## gender:education 4.247e-01 1.161e-01 3.659 0.000256 \*\*\*  
## gender:hispanic 1.107e+00 5.514e-01 2.007 0.044778 \*   
## gender:afam 1.555e+00 5.879e-01 2.644 0.008210 \*\*   
## fcollege:income 7.244e-01 5.675e-01 1.277 0.201831   
## mcollege:income 1.716e+00 5.961e-01 2.878 0.004015 \*\*   
## home:urban 1.511e+00 6.452e-01 2.342 0.019202 \*   
## home:unemp -1.832e+00 4.696e+00 -0.390 0.696423   
## home:distance -5.751e-01 5.268e-01 -1.092 0.275037   
## distance:unemp 1.164e-01 5.681e-02 2.049 0.040563 \*   
## hispanic:unemp 1.931e-01 1.139e-01 1.695 0.090133 .   
## wage:hispanic -7.188e+01 6.025e+01 -1.193 0.232865   
## wage:afam 5.472e-01 2.178e-01 2.512 0.012024 \*   
## distance:hispanic -9.860e-01 5.629e-01 -1.752 0.079898 .   
## distance:afam -6.726e-01 5.763e-01 -1.167 0.243199   
## tuition:education 2.148e-01 1.728e-01 1.243 0.213985   
## income:hispanic 1.467e+00 6.884e-01 2.131 0.033114 \*   
## gender:I(wage^2) 1.512e+00 7.149e-01 2.115 0.034477 \*   
## gender:I(log(wage)) 2.653e+02 1.307e+02 2.030 0.042377 \*   
## hispanic:I(wage^2) 1.999e+00 1.561e+00 1.281 0.200400   
## hispanic:I(log(wage)) 3.151e+02 2.868e+02 1.099 0.271987   
## home:I(unemp^2) 8.740e-03 6.205e-02 0.141 0.887991   
## home:I(unemp^0.25) 9.062e+01 1.927e+02 0.470 0.638106   
## home:I(log(unemp)) -2.546e+01 5.152e+01 -0.494 0.621225   
## home:I(distance^2) 4.131e-02 2.852e-02 1.448 0.147557   
## home:I(distance^0.25) 2.594e+00 2.020e+00 1.284 0.199126   
## I(distance^2):unemp -4.791e-03 3.492e-03 -1.372 0.170095   
## unemp:I(distance^0.25) -2.581e-01 1.529e-01 -1.689 0.091352 .   
## hispanic:I(distance^2) 4.752e-02 2.869e-02 1.656 0.097730 .   
## hispanic:I(distance^0.25) 1.784e+00 2.270e+00 0.786 0.431824   
## afam:I(distance^2) 5.279e-02 3.782e-02 1.396 0.162772   
## afam:I(distance^0.25) -2.063e-01 2.099e+00 -0.098 0.921700   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.006 on 4681 degrees of freedom  
## Multiple R-squared: 0.3597, Adjusted R-squared: 0.3519   
## F-statistic: 46.13 on 57 and 4681 DF, p-value: < 2.2e-16

باز هم بدلیل کم بودن آر2 و آر2 اجاست شکست خوردیم.

آنووا را تعریف میکنیم.

Anova(distanceaic, distanceaic2, test = “F”)

## Analysis of Variance Table  
##   
## Model 1: score ~ gender + fcollege + mcollege + home + urban + unemp +   
## wage + distance + tuition + education + income + region +   
## ispanic + afam + gender:home + gender:urban + gender:wage +   
## gender:education + gender:ispanic + gender:afam + fcollege:income +   
## mcollege:education + mcollege:income + home:urban + home:unemp +   
## home:distance + home:tuition + home:region + unemp:distance +   
## wage:region + wage:afam + distance:ispanic + distance:afam +   
## tuition:region + tuition:ispanic + education:region + income:ispanic  
## Model 2: score ~ gender + fcollege + mcollege + home + urban + unemp +   
## wage + distance + tuition + education + income + region +   
## ispanic + afam + I(distance^2) + I(education^2) + gender:home +   
## gender:urban + gender:wage + gender:tuition + gender:education +   
## gender:region + gender:ispanic + gender:afam + fcollege:income +   
## mcollege:income + home:urban + home:unemp + home:distance +   
## home:tuition + home:region + urban:unemp + unemp:ispanic +   
## wage:region + wage:afam + distance:ispanic + distance:afam +   
## tuition:region + education:region + income:ispanic  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 4701 232156   
## 2 4698 231051 3 1104.7 7.487 5.36e-05 \*\*\*  
## ---  
## Signif. Codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ‘ 1

آنووا با بزرگترین مدل:

anova(distanceaic2, large\_param\_model, test = "F")

## Analysis of Variance Table  
##   
## Model 1: score ~ gender + fcollege + mcollege + home + urban + unemp +   
## wage + distance + tuition + education + income + region +   
## hispanic + afam + I(distance^2) + I(education^2) + gender:home +   
## gender:urban + gender:wage + gender:tuition + gender:education +   
## gender:region + gender:hispanic + gender:afam + fcollege:income +   
## mcollege:income + home:urban + home:unemp + home:distance +   
## home:tuition + home:region + urban:unemp + unemp:hispanic +   
## wage:region + wage:afam + distance:hispanic + distance:afam +   
## tuition:region + education:region + income:hispanic  
## Model 2: score ~ gender + fcollege + fcollege:mcollege + home + urban +   
## wage + distance + I(distance^(0.25)) + tuition + education +   
## income + hispanic + afam + I(distance^2) + I(tuition^2) +   
## I(education^2) + gender:home + gender:wage + gender:education +   
## gender:hispanic + gender:afam + fcollege:income + mcollege:income +   
## home:urban + home:unemp + home:distance + unemp:distance +   
## unemp:hispanic + wage:hispanic + wage:afam + distance:hispanic +   
## distance:afam + tuition:education + income:hispanic + I(tuition^3) +   
## log(tuition) + I(education^3) + log(education) + I(tuition^0.25) +   
## I(education^0.25) + I(wage^2) + gender:I(wage^2) + gender:I(log(wage)) +   
## hispanic:I(wage^2) + hispanic:I(log(wage)) + home:I(unemp^2) +   
## home:I(unemp^0.25) + home:I(log(unemp)) + home:I(distance^2) +   
## home:I(distance^0.25) + unemp:I(distance^2) + unemp:I(distance^0.25) +   
## hispanic:I(distance^2) + hispanic:I(distance^0.25) + afam:I(distance^2) +   
## afam:I(distance^0.25) + region  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 4698 231051   
## 2 4681 229734 17 1317.2 1.5788 0.06098 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

بعد از این نتیجه خوب به سمت نهایی کردن مدل خواهیم رفت.قبل از آن نقاط غیر عادی را شامل داده پرت و....حذف می کنیم(این مرحله بسیار مهم است یکبار از این مرحله صرف نظر کردم و متغیرهای اثرگذاری که در اخر بدست می آمد

تفاوت چشمگیری با نتیجه اکنون داشت.

# finds the unusual points (leverage, residual, influential) of a certain model  
getUnusualPoints = function(model) {  
 # calculate leverage  
 num\_high\_leverage = sum(hatvalues(model) > 2 \* mean(hatvalues(model)))  
   
 # calculate large residuals  
 num\_large\_residual = sum(abs(rstandard(model)) > 2)  
   
 # calculate influential  
 cooks\_dist = cooks.distance(model)  
 num\_influential = sum(cooks\_dist > 4 / length(cooks\_dist))  
   
 # return the values  
 return(list(leverage = num\_high\_leverage,  
 residual = num\_large\_residual,  
 influential = num\_influential))  
}  
  
unusual\_point\_data = getUnusualPoints(large\_param\_model)  
print(unusual\_point\_data)

## $leverage  
## [1] 267  
##   
## $residual  
## [1] 182  
##   
## $influential  
## [1] 206

206نقطه با نفوذ بسیار بالا و 182تا نقطه با باقیمانده بسیار بالا داریم.این روند تا زمانی که داده ها به 2000تا برسد

ادامه دارد تا اگر چیزی الزام به حذف است،حذف شود.

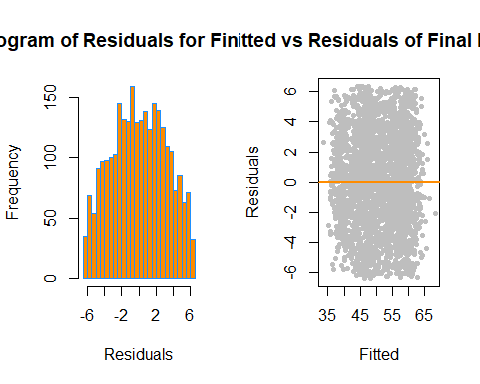
# initialize our final model and cleaned dataset  
final\_model = large\_param\_model  
collegedistance\_cleaned = collegedistance  
  
while(nrow(collegedistance\_cleaned) > 2000) {  
   
 # calculate influence indexes  
 infl\_indexes = which(cooks.distance(final\_model) > 4 /  
 length(cooks.distance((final\_model))))  
   
 # calculate residual indexes  
 resid\_indexes = which(abs(rstandard(final\_model)) > 2)  
   
 # get all unique subset indexes  
 subset\_indexes = unique(c(infl\_indexes, resid\_indexes))  
   
 # break cases  
 if (nrow(collegedistance\_cleaned) - length(subset\_indexes) < 2000) {  
 break  
 }  
 if (length(subset\_indexes) == 0) {  
 break  
 }  
   
 # adjust the data set  
 collegedistance\_cleaned = collegedistance\_cleaned[-1 \* subset\_indexes,]  
   
 # recalculate the final model  
 final\_model = lm(score ~ gender + fcollege + fcollege:mcollege + home +   
 urban + wage + distance + I(distance^(0.25)) +  
 tuition + education + income + hispanic + afam +  
 I(distance^2) + I(tuition^2) + I(education^2) +  
 gender:home + gender:wage + gender:education +  
 gender:hispanic + gender:afam + fcollege:income +  
 mcollege:income + home:urban + home:unemp +  
 home:distance + unemp:distance +  
 unemp:hispanic + wage:hispanic + wage:afam +  
 distance:hispanic + distance:afam +   
 tuition:education + income:hispanic + I(tuition^3) +  
 log(tuition) + I(education^3)+ log(education) +  
 I(tuition^0.25) + I(education^0.25) + I(wage^2) +  
 gender:I(wage^2) + gender:I(log(wage)) +  
 hispanic:I(wage^2) + hispanic:I(log(wage)) +  
 home:I(unemp^2) + home:I(unemp^0.25) +  
 home:I(log(unemp)) + home:I(distance^2) +  
 home:I(distance^0.25) + unemp:I(distance^2) +  
 unemp:I(distance^0.25) + hispanic:I(distance^2) +  
 hispanic:I(distance^0.25) + afam:I(distance^2) +  
 afam:I(distance^0.25) + region,   
 data = collegedistance\_cleaned)  
}

حال نمودارهای مورد نظر را میکشیم تا بصری همه چیز را ملاحظه کنیم.

# create the fitted versus residuals scatter plot  
createFittedResidualsPlot = function(model) {  
 plot(fitted(model), resid(model), col = "grey", pch = 20,  
 xlab = "Fitted", ylab = "Residuals",   
 main = "Fitted vs Residuals of Final Model")  
 abline(h = 0, col = "darkorange", lwd = 2)  
}  
# creates a histogram of residuals  
createResidualHistogram = function(model) {  
 hist(resid(model),  
 xlab = "Residuals",  
 main = "Histogram of Residuals for Final Model",  
 col = "darkorange",  
 border = "dodgerblue",  
 breaks = 20)  
}  
  
# creates a Q-Q plot  
createQQPlot = function(model) {  
 qqnorm(resid(model), main = "Normal Q-Q Plot", col = "darkgrey",   
 cex.main = 2, cex.axis = 2, cex.lab = 2)  
 qqline(resid(model), col = "dodgerblue", lwd = 2)  
}  
  
par(mfrow = c(1, 2))  
createResidualHistogram(final\_model)  
createFittedResidualsPlot(final\_model)

باقی مانده ها توزیع نسبتا نرمال دارند البته بیشتر به توزیع تی شباهت دارند.هیچ روند کلی در نمودار پراکنگی وجود

ندارد.



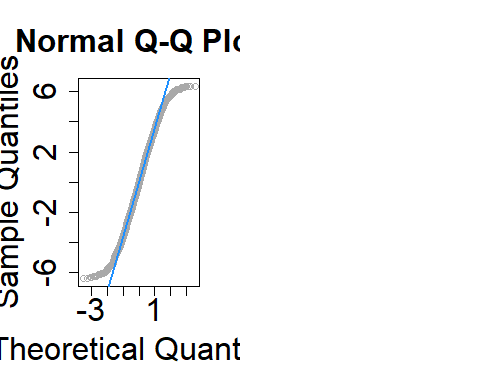
کیوکیو پلات را برای مدل نهایی میکشیم.

createQQPlot(final\_model)

تعداد نقاط غیرمعمول در مدل:

final\_unusual = getUnusualPoints(final\_model)  
print(final\_unusual)

## $leverage  
## [1] 144  
##   
## $residual  
## [1] 0  
##   
## $influential  
## [1] 0



ما با مجموعه داده ای از 4739 مشاهده از 15 متغیر شروع کردیم. ابتدا داده های خود را پاک کردیم و مقادیر گمشده را بررسی کردیم. برای بهبود بیشتر مدل امیدوارکننده‌مان، انواع تغییرات پاسخ و پیش‌بینی‌کننده را اعمال کردیم و نقاط غیرمعمول را حذف کردیم و در نهایت «بهترین» مدل خود را به ما داد.

پکیج های مورد نظر را فراخوانی می کنیم:

library(MASS)   
library(reshape2)

## Warning: package 'reshape2' was built under R version 4.2.2

library(reshape)

## Warning: package 'reshape' was built under R version 4.2.2

##   
## Attaching package: 'reshape'

## The following objects are masked from 'package:reshape2':  
##   
## colsplit, melt, recast

هدف اصلی این مطالعه این بود که ببینیم چگونه عوامل مختلف اجتماعی و اقتصادی می‌توانند بر توانایی دانش‌آموزان در مدرسه تأثیر بگذارند. در این قسمت به تفسیر ضرایب مدل خود می پردازیم.

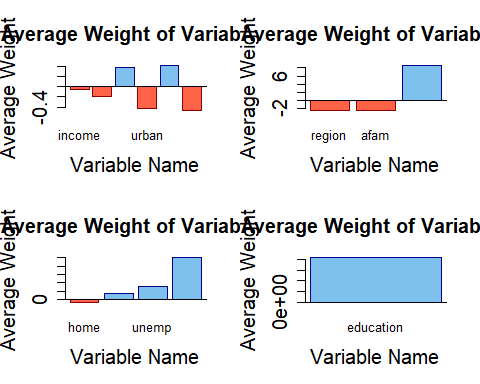
ابتدا دیتافریم را با داده هایی که نیاز داریم ایجاد می کنیم.

داده ها را بصورت تصویری ارائه می کنیم.

# initialize the dataframe  
coef\_df = data.frame(matrix(ncol = length(names(collegedistance)) - 1, nrow = 1))  
score\_index = which(names(collegedistance) == "score")  
colnames(coef\_df) = names(collegedistance)[-1 \* score\_index]  
  
# iterate through each name  
for (name in names(collegedistance)[-1 \* score\_index]) {  
   
 # get the sum of coefficients  
 coef\_indexes = grepl(name, names(coef(final\_model)))  
 coef\_avg = mean(coef(final\_model)[coef\_indexes])  
   
 # add to the dataframe  
 coef\_df[name] = coef\_avg  
}  
  
# melt into our data  
coef\_data = melt(coef\_df)

## Using as id variables

# sort by value  
coef\_data = coef\_data[order(abs(coef\_data$value)),]  
  
createWeightPlot = function(start, end) {  
 barplot(height = coef\_data[start:end,2],  
 names.arg = coef\_data[start:end,1],  
 col = ifelse(coef\_data[start:end,2] < 0, "tomato", "skyblue2"),  
 border = ifelse(coef\_data[start:end,2] < 0, "darkred", "darkblue"),  
 xlab = "Variable Name",  
 ylab = "Average Weight",  
 main = "Average Weight of Variables",  
 cex.lab = 1.5,  
 cex.axis = 1.5,  
 cex.main = 1.5)  
 abline(h = 0)  
}  
  
par(mfrow = c(2, 2))  
createWeightPlot(1, 6)  
createWeightPlot(7, 9)  
createWeightPlot(10, 13)  
createWeightPlot(14, 14)



ما می‌توانیم وزن ترکیبی هر متغیر را به صورت بصری ببینیم (نمودار به 4 بخش تقسیم می‌شود تا وزن‌های ضرایب مختلف محاسبه شود). نمودارهای آبی نشان دهنده ضرایب میانگین مثبت هستند، در حالی که نمودارهای قرمز نشان دهنده ضرایب میانگین منفی هستند. با این داده ها، اکنون می توانیم متغیرهای خود را تفسیر کنیم:

درآمد: درآمد اساساً صفر است، به این معنی که داشتن خانواده با درآمد بالا یا کم به طور جالبی بر نمرات آزمون تأثیری ندارد.

فاصله: هرچه کالج چهار ساله دانش‌آموز نزدیک‌تر باشد، در امتحان بدتر می‌شوند. این منطقی است زیرا بسیاری از نزدیک‌ترین کالج‌ها به دانش‌آموزان، کالج‌های با عملکرد ضعیفتر خواهند بود.

mcollege: نتایج نشان می‌دهد که اگر مادر دانش‌آموز فارغ‌التحصیل کالج باشد، در مورد نمره آزمون از مزیت بیشتری برخوردار خواهد بود.

Fcollege: مثل بالایی اما یکم دارای اثر بیشتر.

شهری:دانش آموزان در این مناطق در آزمونهای استاندارد بدتر عمل خواهند کرد.

جنسیت:بطور متوسط مردان بدتر عمل کردند.

منطقه:به غیر از غرب بقیه نسبتا خوب عمل کردند.

afam: دانش آموزان آفریقایی-آمریکایی به طور کلی در آزمون در مقایسه با دانش آموزان غیر اسپانیایی و غیرآمریکایی-آمریکایی بدتر عمل کردند، اما این می تواند با عوامل مختلفی مرتبط باشد.

Hispanic:دانش‌آموزان اسپانیایی تبار در آزمون به طور کلی در مقایسه با دانش‌آموزان غیر اسپانیایی و غیرآفریقایی-آمریکایی بهتر عمل کردند، اما این می‌تواند با عوامل مختلفی مرتبط باشد.

خانه: دانش‌آموزان صاحب خانه در آزمون‌های استاندارد شده عملکرد بدتری داشتند.

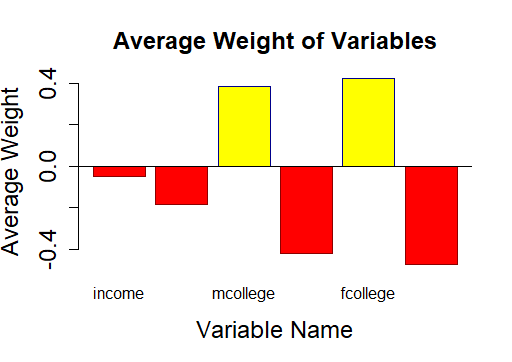
دستمزد: هر چه دستمزد ساعتی در محل دانش آموز بیشتر باشد، در امتحان بهتر عمل خواهند کرد. این منطقی است زیرا دستمزد بالاتر می تواند منجر به آموزش با بودجه بهتر شود.

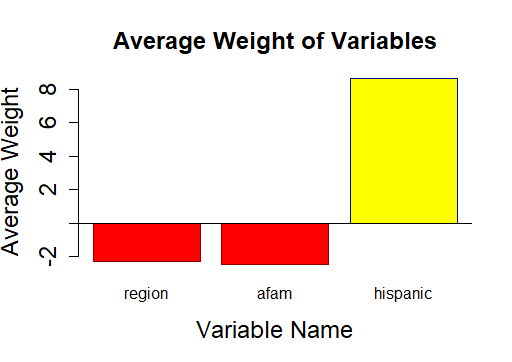
unemp: هر چه نرخ بیکاری در شهرستان بیشتر باشد، دانش آموزان در امتحان بهتر عمل خواهند کرد. این نتیجه بسیار عجیب است.

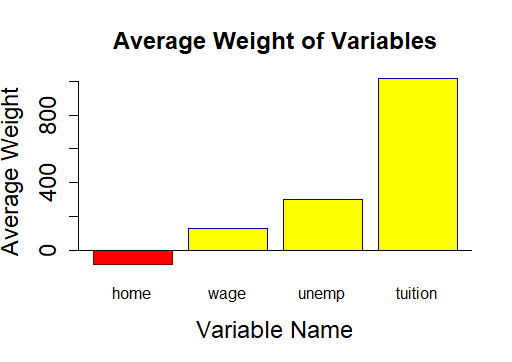
شهریه: دانشجویانی که شهریه بیشتری پرداخت می‌کردند، به طور میانگین نمرات آزمون بالاتری داشتند. این برای همان استدلال دستمزد منطقی است زیرا هرچه شهریه بیشتری پرداخت شود امکانات بیشتری در اختیار دانش آموزان قرار می گیرد.

تحصیلات: دانش آموزانی که سطوح بالاتر تحصیلی را دنبال کردند، نمرات آزمون بسیار بالاتری داشتند. این تعجب آور نیست، زیرا ما انتظار داریم دانش آموزانی که تحصیلات بیشتری را دنبال می کنند، علاقه بیشتری به مدرسه داشته باشند و در نتیجه در امتحان بهتر عمل کرده باشند.

چون نمودار های بالا چندان واضح نیست بطور جدا هم خروجی میگیریم تا واضح تر باشد.

**



**

