STA 9750

Basic Software Tools for Data Analysis

**Introduction:**

A student's GPA (Grade Point Average) is a commonly used criteria that colleges and employers look at to evaluate a person's potential to succeed in their future endeavours in academia or professional lives. GPA is an indicator as to how hard a student has worked in their educational courses, how much he or she knows in specific topics, and where his or her strengths and weaknesses lie. Whether a student is in high school, applying for colleges, or looking for work after graduation, their GPA will likely affect their future prospects. This attribute is derived from various other attributes. For example, how busy the student’s schedule is, how actively they apply themselves towards their learning as well as personal factors such as family, relationship status, and their social life can all affect their GPAs.

Data:

Responses are from students in an introductory statistics class at the University of Tennessee in 2010. Our goal is to predict a student’s college GPA from some of their characteristics.

* A data frame has 607 observations. The following are 18 variables in the data set:
  + CollegeGPA, Gender, HSGPA, ACT, APHours, JobHours, School, LanguagesSpoken", HSHonorsClasses, SmokeInHS, PayCollegeNoLoans, ClubsInHS, JobInHS, Churchgoer, Height, Weight, Family, Pet
  + We chose CollegeGPA as our response Y variable, (independent) and 4 X variable (independent) as our predictive variables. We are going to examine each of the X variable's relationship with CollegeGPA as well as potential correlations.

Summary of each X variable:

## CollegeGPA HSGPA ACT Weight Gender   
## Min. :1.890 Min. :2.340 Min. :17.00 Min. : 86.0 Female:338   
## 1st Qu.:3.000 1st Qu.:3.500 1st Qu.:24.00 1st Qu.:130.0 Male :269   
## Median :3.300 Median :3.800 Median :26.00 Median :150.0   
## Mean :3.271 Mean :3.709 Mean :26.14 Mean :154.4   
## 3rd Qu.:3.635 3rd Qu.:3.960 3rd Qu.:28.00 3rd Qu.:175.0   
## Max. :4.000 Max. :5.000 Max. :35.00 Max. :298.0   
## School ## Private:122 ## Public :485

Observations based on the summary above:

* CollegeGPA's max is 4 which is the max value of GPA range.
* HSGPA (High School GPA) nowadays has a scale of 1-5 which might lead to some issues.
* Weight has a big range of data points than others.
* Gender: There are more female data points collected than male.
* School: There are more Public school data points collected than Private ones.

Dataset is large enough sample so not overly concerned with our Gender and School sampling.

We converted certain binary variables such as "Gender" and "School" by converting them to 1 or 0 so that we can test out against CollegeGPA

Colinearity Matrix

From the matrix below, we note CollegeGPA seems positively correlated to HSGPA and ACT and negatively correlated to Weight. Further examination needed for potential collinearity amongst the variables. We do observe that correlations mentioned do not seem overly strong.

## CollegeGPA HSGPA ACT Weight Gender\_b School\_b  
## CollegeGPA 1.000 0.376 0.250 -0.174 0.207 0.050  
## HSGPA 0.376 1.000 0.259 -0.172 0.263 -0.037  
## ACT 0.250 0.259 1.000 -0.023 -0.013 0.053  
## Weight -0.174 -0.172 -0.023 1.000 -0.595 0.037  
## Gender\_b 0.207 0.263 -0.013 -0.595 1.000 -0.041  
## School\_b 0.050 -0.037 0.053 0.037 -0.041 1.000

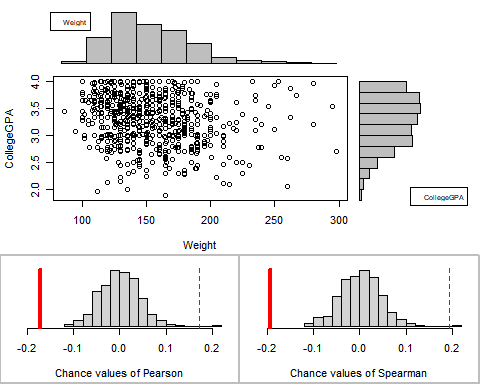
To confirm that we picked the best variables we sorted all variables in the data set to be sure. See below.

## var1 var2 correlation pval  
## 1 CollegeGPA HSGPA 0.3760406 7.982618e-22  
## 2 CollegeGPA ACT 0.2503677 3.960260e-10  
## 3 CollegeGPA Gender\_b 0.2071693 2.611048e-07  
## 4 CollegeGPA Weight -0.1736529 1.690644e-05  
## 5 CollegeGPA School\_b 0.0501086 2.176601e-01

## Association between HSGPA (numerical) and CollegeGPA (numerical)  
## using 607 complete cases

associate(CollegeGPA~Weight,data=edu3) # both Weight and College GPA look skewed toward right. Spearman correlation looks approapriate, it shows negative correlation as we have seen in the scatter plot.

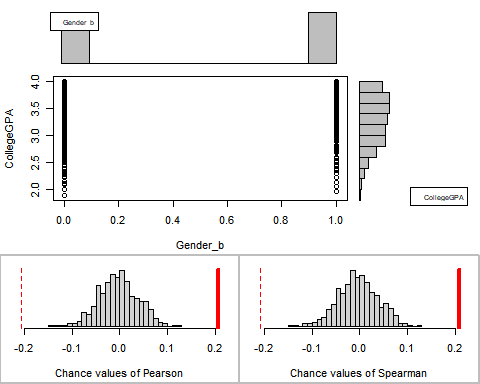
## Association between Weight (numerical) and CollegeGPA (numerical)  
## using 607 complete cases



## Permutation procedure:  
## Value Estimated p-value  
## Pearson's r -0.1736529 0.002  
## Spearman's rank correlation -0.1934833 0.002  
## With 500 permutations, we are 95% confident that:  
## the p-value of Pearson's correlation (r) is between 0 and 0.011   
## the p-value of Spearman's rank correlation is between 0 and 0.011   
## Note: If 0.05 is in this range, increase the permutations= argument.  
##   
##   
##   
## Advice: If stream of points is well described by an ellipse, use Pearson's r.  
## Otherwise, as long as stream is monotonic, use Spearman's rank correlation  
## or try logs, e.g. associate( log10(y)~log10(x) )

associate(CollegeGPA~Gender\_b,data=edu3) # both Spearman and Pearson correlation look normality distributed

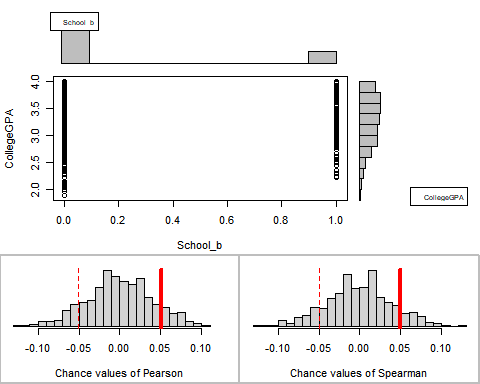
## Association between Gender\_b (numerical) and CollegeGPA (numerical)  
## using 607 complete cases



## Permutation procedure:  
## Value Estimated p-value  
## Pearson's r 0.2071693 0  
## Spearman's rank correlation 0.2086050 0  
## With 500 permutations, we are 95% confident that:  
## the p-value of Pearson's correlation (r) is between 0 and 0.007   
## the p-value of Spearman's rank correlation is between 0 and 0.007   
## Note: If 0.05 is in this range, increase the permutations= argument.  
##   
##   
##   
## Advice: If stream of points is well described by an ellipse, use Pearson's r.  
## Otherwise, as long as stream is monotonic, use Spearman's rank correlation  
## or try logs, e.g. associate( log10(y)~log10(x) )

associate(CollegeGPA~School\_b,data=edu3) # Pearson Correlation looks appropriate, Public school has more data points as we have seen earlier in the summary table

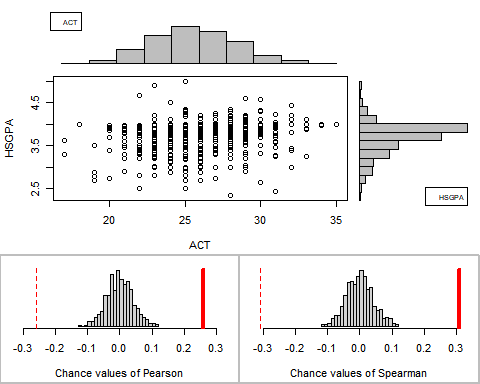
## Association between School\_b (numerical) and CollegeGPA (numerical)  
## using 607 complete cases



## Permutation procedure:  
## Value Estimated p-value  
## Pearson's r 0.05010860 0.226  
## Spearman's rank correlation 0.04993425 0.214  
## With 500 permutations, we are 95% confident that:  
## the p-value of Pearson's correlation (r) is between 0.19 and 0.265   
## the p-value of Spearman's rank correlation is between 0.179 and 0.253   
## Note: If 0.05 is in this range, increase the permutations= argument.  
##   
##   
##   
## Advice: If stream of points is well described by an ellipse, use Pearson's r.  
## Otherwise, as long as stream is monotonic, use Spearman's rank correlation  
## or try logs, e.g. associate( log10(y)~log10(x) )

# Association  
associate(HSGPA~ACT,data=edu3) # ACT and HSGPA seem positively correlation, Pearson looks normally distributed.

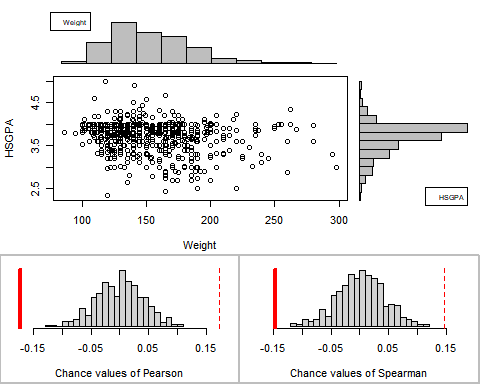
## Association between ACT (numerical) and HSGPA (numerical)  
## using 607 complete cases



## Permutation procedure:  
## Value Estimated p-value  
## Pearson's r 0.2593769 0  
## Spearman's rank correlation 0.3105118 0  
## With 500 permutations, we are 95% confident that:  
## the p-value of Pearson's correlation (r) is between 0 and 0.007   
## the p-value of Spearman's rank correlation is between 0 and 0.007   
## Note: If 0.05 is in this range, increase the permutations= argument.  
##   
##   
##   
## Advice: If stream of points is well described by an ellipse, use Pearson's r.  
## Otherwise, as long as stream is monotonic, use Spearman's rank correlation  
## or try logs, e.g. associate( log10(y)~log10(x) )

associate(HSGPA~Weight,data=edu3) # HSGPA and Weight seem negatively correlated, due to the uneven variance, Spearman correlation works better

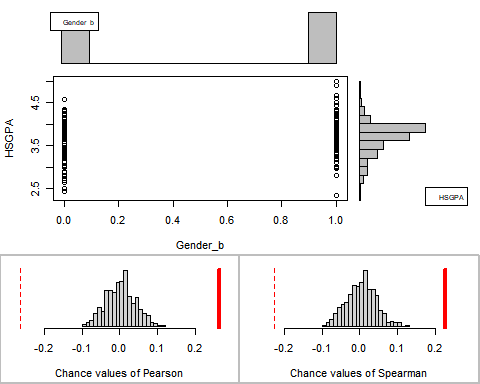
## Association between Weight (numerical) and HSGPA (numerical)  
## using 607 complete cases



## Permutation procedure:  
## Value Estimated p-value  
## Pearson's r -0.1724826 0  
## Spearman's rank correlation -0.1459553 0  
## With 500 permutations, we are 95% confident that:  
## the p-value of Pearson's correlation (r) is between 0 and 0.007   
## the p-value of Spearman's rank correlation is between 0 and 0.007   
## Note: If 0.05 is in this range, increase the permutations= argument.  
##   
##   
##   
## Advice: If stream of points is well described by an ellipse, use Pearson's r.  
## Otherwise, as long as stream is monotonic, use Spearman's rank correlation  
## or try logs, e.g. associate( log10(y)~log10(x) )

associate(HSGPA~Gender\_b,data=edu3) # Gender and HSGPA seem having positive correlation

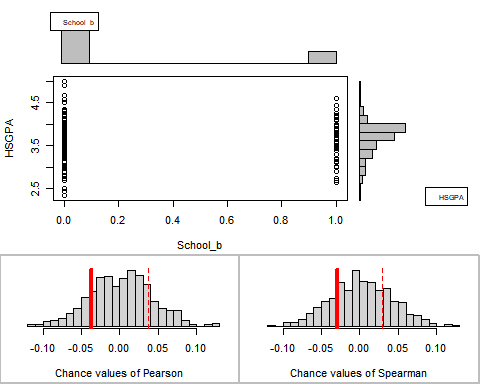
## Association between Gender\_b (numerical) and HSGPA (numerical)  
## using 607 complete cases



## Permutation procedure:  
## Value Estimated p-value  
## Pearson's r 0.2631413 0  
## Spearman's rank correlation 0.2262502 0  
## With 500 permutations, we are 95% confident that:  
## the p-value of Pearson's correlation (r) is between 0 and 0.007   
## the p-value of Spearman's rank correlation is between 0 and 0.007   
## Note: If 0.05 is in this range, increase the permutations= argument.  
##   
##   
##   
## Advice: If stream of points is well described by an ellipse, use Pearson's r.  
## Otherwise, as long as stream is monotonic, use Spearman's rank correlation  
## or try logs, e.g. associate( log10(y)~log10(x) )

associate(HSGPA~School\_b,data=edu3) # School and HSGPA seem having negative correlation. This association is still statistically significant but their p-value is higher than other associations.

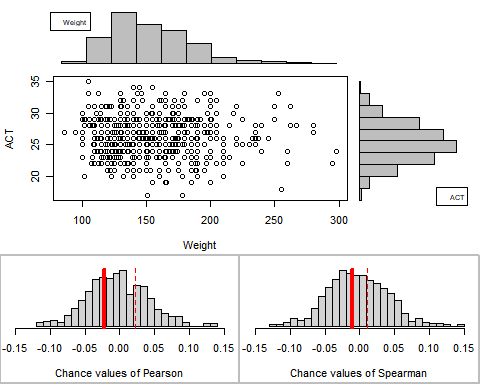
## Association between School\_b (numerical) and HSGPA (numerical)  
## using 607 complete cases



## Permutation procedure:  
## Value Estimated p-value  
## Pearson's r -0.03714554 0.340  
## Spearman's rank correlation -0.02965885 0.458  
## With 500 permutations, we are 95% confident that:  
## the p-value of Pearson's correlation (r) is between 0.299 and 0.383   
## the p-value of Spearman's rank correlation is between 0.414 and 0.503   
## Note: If 0.05 is in this range, increase the permutations= argument.  
##   
##   
##   
## Advice: If stream of points is well described by an ellipse, use Pearson's r.  
## Otherwise, as long as stream is monotonic, use Spearman's rank correlation  
## or try logs, e.g. associate( log10(y)~log10(x) )

# Association  
associate(ACT~Weight,data=edu3) # this association is not statistically significant, p-value is higher than 0.5

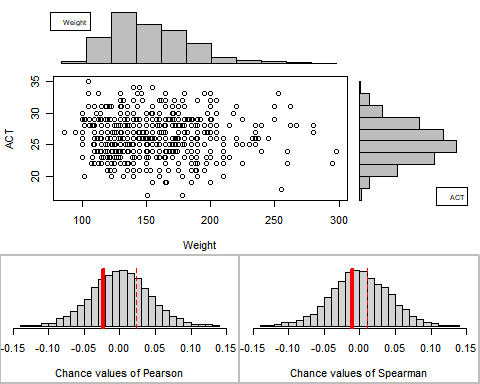
## Association between Weight (numerical) and ACT (numerical)  
## using 607 complete cases



## Permutation procedure:  
## Value Estimated p-value  
## Pearson's r -0.02269029 0.604  
## Spearman's rank correlation -0.01104548 0.810  
## With 500 permutations, we are 95% confident that:  
## the p-value of Pearson's correlation (r) is between 0.56 and 0.647   
## the p-value of Spearman's rank correlation is between 0.773 and 0.843   
## Note: If 0.05 is in this range, increase the permutations= argument.  
##   
##   
##   
## Advice: If stream of points is well described by an ellipse, use Pearson's r.  
## Otherwise, as long as stream is monotonic, use Spearman's rank correlation  
## or try logs, e.g. associate( log10(y)~log10(x) )

associate(ACT~Weight,data=edu3,permutations = 5000) # tried with higher number of permutation, still p-value is higher than 0.5

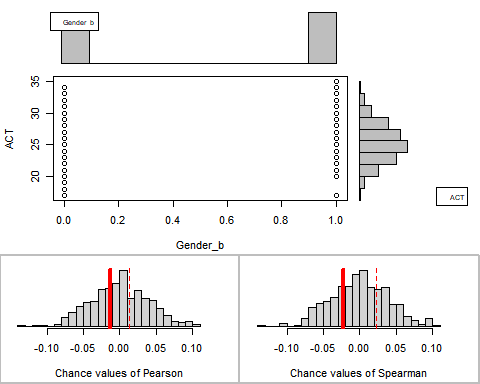
## Association between Weight (numerical) and ACT (numerical)  
## using 607 complete cases



## Permutation procedure:  
## Value Estimated p-value  
## Pearson's r -0.02269029 0.5708  
## Spearman's rank correlation -0.01104548 0.7786  
## With 5000 permutations, we are 95% confident that:  
## the p-value of Pearson's correlation (r) is between 0.557 and 0.585   
## the p-value of Spearman's rank correlation is between 0.767 and 0.79   
## Note: If 0.05 is in this range, increase the permutations= argument.  
##   
##   
##   
## Advice: If stream of points is well described by an ellipse, use Pearson's r.  
## Otherwise, as long as stream is monotonic, use Spearman's rank correlation  
## or try logs, e.g. associate( log10(y)~log10(x) )

associate(ACT~Gender\_b,data=edu3) # this association is not statistically significant, p-value is higher than 0.5

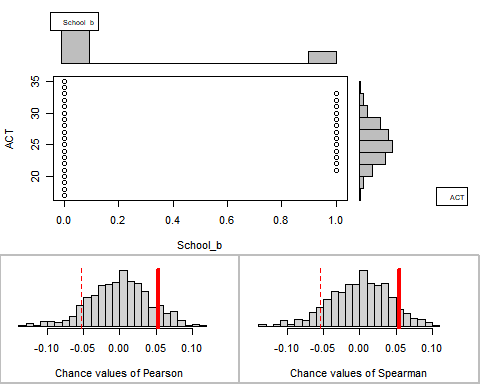
## Association between Gender\_b (numerical) and ACT (numerical)  
## using 607 complete cases



## Permutation procedure:  
## Value Estimated p-value  
## Pearson's r -0.01258852 0.732  
## Spearman's rank correlation -0.02280386 0.560  
## With 500 permutations, we are 95% confident that:  
## the p-value of Pearson's correlation (r) is between 0.691 and 0.77   
## the p-value of Spearman's rank correlation is between 0.515 and 0.604   
## Note: If 0.05 is in this range, increase the permutations= argument.  
##   
##   
##   
## Advice: If stream of points is well described by an ellipse, use Pearson's r.  
## Otherwise, as long as stream is monotonic, use Spearman's rank correlation  
## or try logs, e.g. associate( log10(y)~log10(x) )

associate(ACT~School\_b,data=edu3) # correlation is very weak, statistically significant though

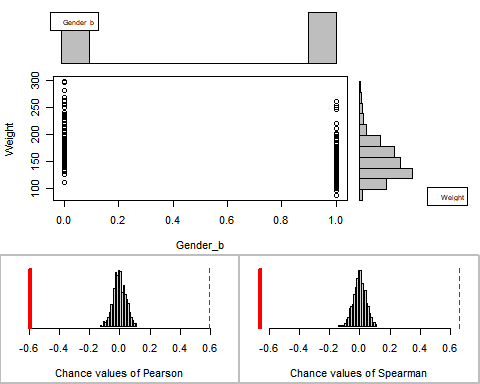
## Association between School\_b (numerical) and ACT (numerical)  
## using 607 complete cases



## Permutation procedure:  
## Value Estimated p-value  
## Pearson's r 0.05316812 0.194  
## Spearman's rank correlation 0.05525071 0.172  
## With 500 permutations, we are 95% confident that:  
## the p-value of Pearson's correlation (r) is between 0.16 and 0.231   
## the p-value of Spearman's rank correlation is between 0.14 and 0.208   
## Note: If 0.05 is in this range, increase the permutations= argument.  
##   
##   
##   
## Advice: If stream of points is well described by an ellipse, use Pearson's r.  
## Otherwise, as long as stream is monotonic, use Spearman's rank correlation  
## or try logs, e.g. associate( log10(y)~log10(x) )

associate(Weight~Gender\_b,data=edu3) # statistically significant but negatively correlated

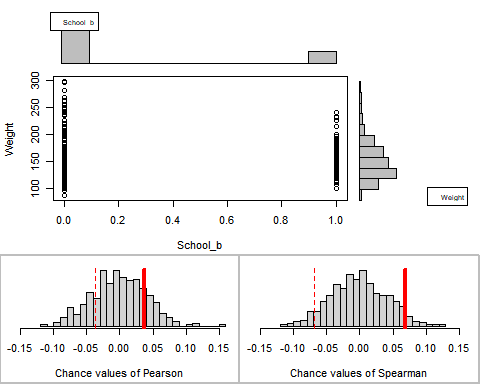
## Association between Gender\_b (numerical) and Weight (numerical)  
## using 607 complete cases



## Permutation procedure:  
## Value Estimated p-value  
## Pearson's r -0.5948267 0  
## Spearman's rank correlation -0.6581910 0  
## With 500 permutations, we are 95% confident that:  
## the p-value of Pearson's correlation (r) is between 0 and 0.007   
## the p-value of Spearman's rank correlation is between 0 and 0.007   
## Note: If 0.05 is in this range, increase the permutations= argument.  
##   
##   
##   
## Advice: If stream of points is well described by an ellipse, use Pearson's r.  
## Otherwise, as long as stream is monotonic, use Spearman's rank correlation  
## or try logs, e.g. associate( log10(y)~log10(x) )

associate(Weight~School\_b,data=edu3) # statistically significant but the two correlation has a large gap

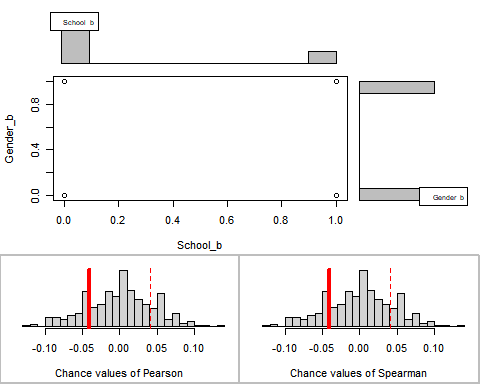
## Association between School\_b (numerical) and Weight (numerical)  
## using 607 complete cases



## Permutation procedure:  
## Value Estimated p-value  
## Pearson's r 0.03675362 0.37  
## Spearman's rank correlation 0.06859103 0.10  
## With 500 permutations, we are 95% confident that:  
## the p-value of Pearson's correlation (r) is between 0.328 and 0.414   
## the p-value of Spearman's rank correlation is between 0.075 and 0.13   
## Note: If 0.05 is in this range, increase the permutations= argument.  
##   
##   
##   
## Advice: If stream of points is well described by an ellipse, use Pearson's r.  
## Otherwise, as long as stream is monotonic, use Spearman's rank correlation  
## or try logs, e.g. associate( log10(y)~log10(x) )

associate(Gender\_b~School\_b,data=edu3) # the two categorical variables have negative correlation

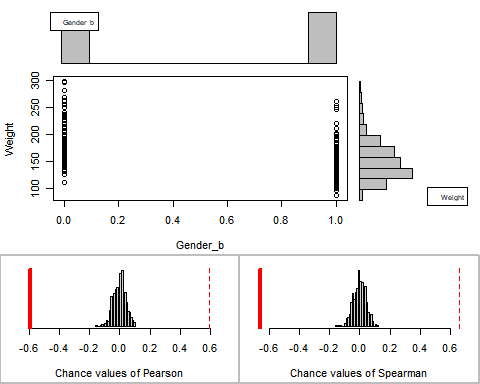
## Association between School\_b (numerical) and Gender\_b (numerical)  
## using 607 complete cases



## Permutation procedure:  
## Value Estimated p-value  
## Pearson's r -0.04083302 0.398  
## Spearman's rank correlation -0.04083302 0.398  
## With 500 permutations, we are 95% confident that:  
## the p-value of Pearson's correlation (r) is between 0.355 and 0.442   
## the p-value of Spearman's rank correlation is between 0.355 and 0.442   
## Note: If 0.05 is in this range, increase the permutations= argument.  
##   
##   
##   
## Advice: If stream of points is well described by an ellipse, use Pearson's r.  
## Otherwise, as long as stream is monotonic, use Spearman's rank correlation  
## or try logs, e.g. associate( log10(y)~log10(x) )

# Association  
associate(Weight~Gender\_b,data=edu3)

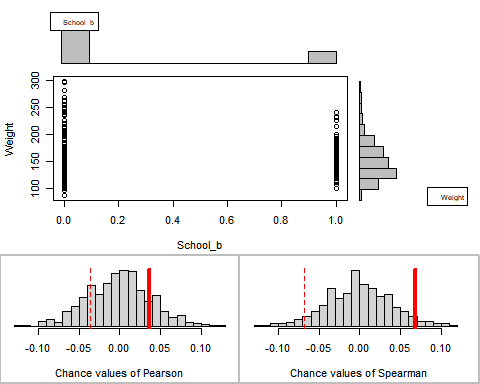
## Association between Gender\_b (numerical) and Weight (numerical)  
## using 607 complete cases



## Permutation procedure:  
## Value Estimated p-value  
## Pearson's r -0.5948267 0  
## Spearman's rank correlation -0.6581910 0  
## With 500 permutations, we are 95% confident that:  
## the p-value of Pearson's correlation (r) is between 0 and 0.007   
## the p-value of Spearman's rank correlation is between 0 and 0.007   
## Note: If 0.05 is in this range, increase the permutations= argument.  
##   
##   
##   
## Advice: If stream of points is well described by an ellipse, use Pearson's r.  
## Otherwise, as long as stream is monotonic, use Spearman's rank correlation  
## or try logs, e.g. associate( log10(y)~log10(x) )

associate(Weight~School\_b,data=edu3)

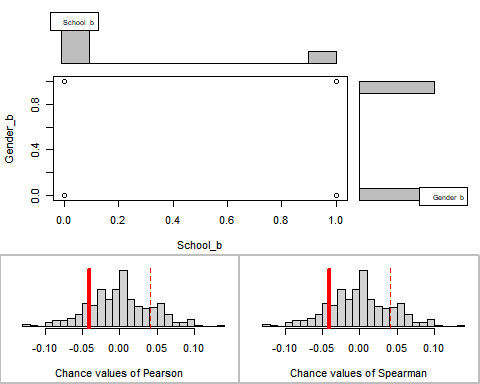
## Association between School\_b (numerical) and Weight (numerical)  
## using 607 complete cases



## Permutation procedure:  
## Value Estimated p-value  
## Pearson's r 0.03675362 0.37  
## Spearman's rank correlation 0.06859103 0.08  
## With 500 permutations, we are 95% confident that:  
## the p-value of Pearson's correlation (r) is between 0.328 and 0.414   
## the p-value of Spearman's rank correlation is between 0.058 and 0.107   
## Note: If 0.05 is in this range, increase the permutations= argument.  
##   
##   
##   
## Advice: If stream of points is well described by an ellipse, use Pearson's r.  
## Otherwise, as long as stream is monotonic, use Spearman's rank correlation  
## or try logs, e.g. associate( log10(y)~log10(x) )

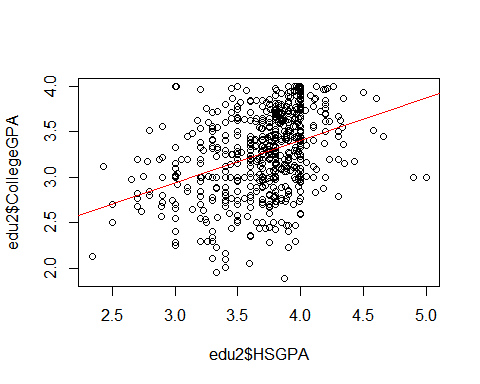
# Association  
associate(Gender\_b~School\_b,data=edu3)

## Association between School\_b (numerical) and Gender\_b (numerical)  
## using 607 complete cases



## Permutation procedure:  
## Value Estimated p-value  
## Pearson's r -0.04083302 0.344  
## Spearman's rank correlation -0.04083302 0.344  
## With 500 permutations, we are 95% confident that:  
## the p-value of Pearson's correlation (r) is between 0.302 and 0.387   
## the p-value of Spearman's rank correlation is between 0.302 and 0.387   
## Note: If 0.05 is in this range, increase the permutations= argument.  
##   
##   
##   
## Advice: If stream of points is well described by an ellipse, use Pearson's r.  
## Otherwise, as long as stream is monotonic, use Spearman's rank correlation  
## or try logs, e.g. associate( log10(y)~log10(x) )

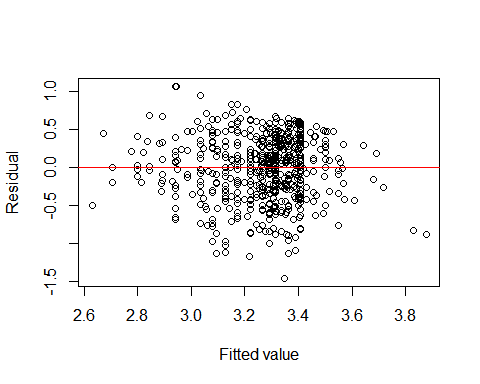
# regressions between collegeGPA vs X variables  
  
### comment: we can start off modeling the highest correlated pair: collegeGPA and HSGPA  
plot(edu2$CollegeGPA~edu2$HSGPA); m\_gpas = lm(CollegeGPA~HSGPA, data=edu2); abline(m\_gpas,col="red")



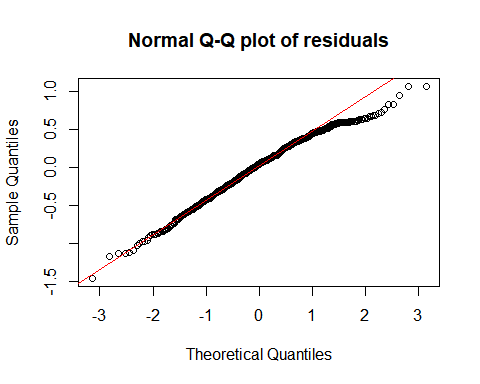
summary(m\_gpas) # HSGPA is statstically significant

##   
## Call:  
## lm(formula = CollegeGPA ~ HSGPA, data = edu2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.45672 -0.28533 0.03922 0.32606 1.06079   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.53399 0.17485 8.773 <2e-16 \*\*\*  
## HSGPA 0.46841 0.04693 9.982 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4151 on 605 degrees of freedom  
## Multiple R-squared: 0.1414, Adjusted R-squared: 0.14   
## F-statistic: 99.64 on 1 and 605 DF, p-value: < 2.2e-16

### comment: How does residual look like? They are spread around 0  
plot(m\_gpas$residuals~m\_gpas$fitted.values,xlab="Fitted value",ylab="Residual"); abline(h=0,col='red')



### comment: The residuals seem following normality  
qqnorm(m\_gpas$residuals,main="Normal Q-Q plot of residuals"); qqline(m\_gpas$residuals,col='red')



# Add additional variables  
  
# Stepwise approach: start with a regression model including all variables  
all\_m = lm(CollegeGPA~HSGPA+ACT+Weight+Gender\_b+School\_b, data=edu3); summary(all\_m)

##   
## Call:  
## lm(formula = CollegeGPA ~ HSGPA + ACT + Weight + Gender\_b + School\_b,   
## data = edu3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.53924 -0.26949 0.03823 0.30842 1.13072   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.3225499 0.2274760 5.814 9.91e-09 \*\*\*  
## HSGPA 0.3713417 0.0493609 7.523 1.96e-13 \*\*\*  
## ACT 0.0244659 0.0055211 4.431 1.11e-05 \*\*\*  
## Weight -0.0008350 0.0005982 -1.396 0.1633   
## Gender\_b 0.0857549 0.0422058 2.032 0.0426 \*   
## School\_b 0.0651724 0.0411493 1.584 0.1138   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4049 on 601 degrees of freedom  
## Multiple R-squared: 0.1883, Adjusted R-squared: 0.1816   
## F-statistic: 27.89 on 5 and 601 DF, p-value: < 2.2e-16

# Weight and School\_b seem not statistically significant  
# Weight's coefficient is negative  
  
no\_weight\_m = lm(CollegeGPA~HSGPA+ACT+Gender\_b+School\_b, data=edu3); summary(no\_weight\_m) # school seems not statistically significant

##   
## Call:  
## lm(formula = CollegeGPA ~ HSGPA + ACT + Gender\_b + School\_b,   
## data = edu3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.53125 -0.27228 0.04184 0.31105 1.12498   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.165416 0.197824 5.891 6.38e-09 \*\*\*  
## HSGPA 0.372032 0.049397 7.531 1.85e-13 \*\*\*  
## ACT 0.024730 0.005522 4.478 9.01e-06 \*\*\*  
## Gender\_b 0.119817 0.034463 3.477 0.000544 \*\*\*  
## School\_b 0.064191 0.041176 1.559 0.119532   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4052 on 602 degrees of freedom  
## Multiple R-squared: 0.1857, Adjusted R-squared: 0.1803   
## F-statistic: 34.32 on 4 and 602 DF, p-value: < 2.2e-16

no\_school\_m = lm(CollegeGPA~HSGPA+ACT+Weight+Gender\_b, data=edu3); summary(no\_school\_m) # both Weight and Gender seem not statistically significant

##   
## Call:  
## lm(formula = CollegeGPA ~ HSGPA + ACT + Weight + Gender\_b, data = edu3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.5562 -0.2652 0.0401 0.3083 1.1155   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.3317176 0.2276871 5.849 8.12e-09 \*\*\*  
## HSGPA 0.3679598 0.0493765 7.452 3.21e-13 \*\*\*  
## ACT 0.0250175 0.0055170 4.535 6.97e-06 \*\*\*  
## Weight -0.0008188 0.0005989 -1.367 0.1720   
## Gender\_b 0.0849577 0.0422556 2.011 0.0448 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4054 on 602 degrees of freedom  
## Multiple R-squared: 0.185, Adjusted R-squared: 0.1795   
## F-statistic: 34.15 on 4 and 602 DF, p-value: < 2.2e-16

# so let's examine the regression model we found three variables; HSGPA, ACT, Gender statistically significant against College GPA  
# Regression without School and Weight  
no\_w\_s\_m = lm(CollegeGPA~HSGPA+ACT+Gender\_b, data=edu3); summary(no\_w\_s\_m);

##   
## Call:  
## lm(formula = CollegeGPA ~ HSGPA + ACT + Gender\_b, data = edu3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.54811 -0.25729 0.04619 0.31192 1.11006   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.177449 0.197908 5.949 4.56e-09 \*\*\*  
## HSGPA 0.368687 0.049409 7.462 2.99e-13 \*\*\*  
## ACT 0.025268 0.005518 4.579 5.67e-06 \*\*\*  
## Gender\_b 0.118381 0.034492 3.432 0.00064 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4057 on 603 degrees of freedom  
## Multiple R-squared: 0.1824, Adjusted R-squared: 0.1784   
## F-statistic: 44.85 on 3 and 603 DF, p-value: < 2.2e-16

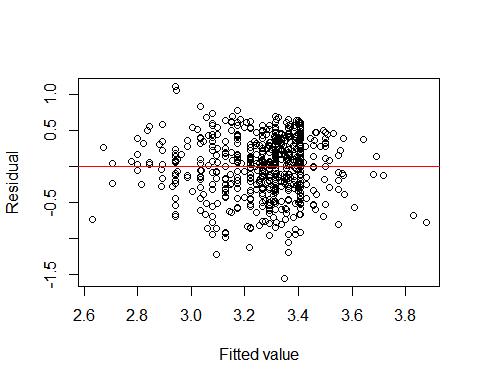
# is this linear model practical?   
# how well does the linear model explain the actual data?  
# Explain linear model interpretation, significance of the relationship, model fitting performance   
### RMSE (Root Mena Sqaure Error) -> this could be subjective so we could use R^2 -> evaluate model performance  
  
# R-squared is 0.1824 -> this means this model can explain 18% of variance in the data when we regress HSGPA, ACT, Gender againsts College GPA  
  
### F-test: ANOVA  
anova(no\_w\_s\_m) # All three variables show statistical significance!!!

## Analysis of Variance Table  
##   
## Response: CollegeGPA  
## Df Sum Sq Mean Sq F value Pr(>F)   
## HSGPA 1 17.168 17.1683 104.29 < 2.2e-16 \*\*\*  
## ACT 1 3.040 3.0404 18.47 2.013e-05 \*\*\*  
## Gender\_b 1 1.939 1.9391 11.78 0.0006399 \*\*\*  
## Residuals 603 99.263 0.1646   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

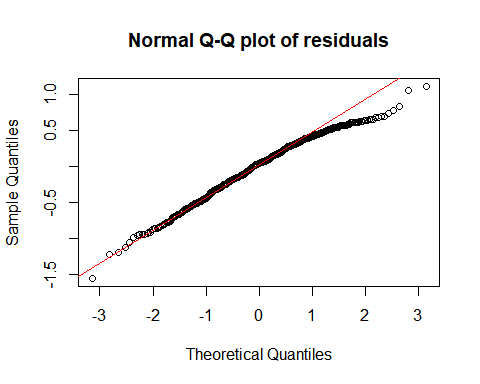
#Confidence refers to the long run fraction of time that the procedure yields a correct interval.  
#If 0 is inside the 95% confidence interval for the slope, the regression is not statistically significant.  
# We are 95% confidence that the difference in the average value of y between two groups of individuals whose values of x differ by 1 unit is between L and U.  
confint(no\_w\_s\_m,level=0.95)

## 2.5 % 97.5 %  
## (Intercept) 0.78877739 1.56612103  
## HSGPA 0.27165186 0.46572185  
## ACT 0.01443119 0.03610473  
## Gender\_b 0.05064171 0.18611966

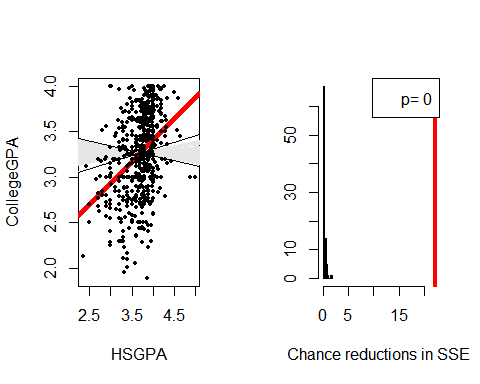
### comment: How does residual look like? They are spread around 0  
plot(no\_w\_s\_m$residuals~m\_gpas$fitted.values,xlab="Fitted value",ylab="Residual"); abline(h=0,col='red')



### comment: There is slight deviation at the high end tail but the residuals seem following normality,   
qqnorm(no\_w\_s\_m$residuals,main="Normal Q-Q plot of residuals"); qqline(m\_gpas$residuals,col='red')



# Permutation Simulation  
possible\_regressions(no\_w\_s\_m)



### The red line is placed far right in the chart on the righthand side, it shows the evidnece of significant association  
### Grey area (left chart) shows permuted data sets, red line is actual data set