Advanced Topics in Research Methods and Design

Jingru Zhang, Lesheng Xu, Yi Chen

Columbia University

Exercise One

Advanced Topics in Research Methods and Design

# Exercise in Chapter 2

Identify the appropriate statistical test for each question by using both decision-making tools.

## Question 1

## To what degree do SAT scores predict college freshmen GPAs?

*Answer*: Bivariate Correlation and/or Regression.

## Question 2

Does ethnicity significantly affect reading achievement, math achievement, and overall achievement among sixth grade students?

*Answer*: One-way MANOVA or One-way MANCOVA (if other important covariates available)

## Question 3

What are the causal effects (direct and indirect) among number of school absences due to illness, reading ability, semester GPA, and total score on the Iowa Test of Basic Skills among eighth grade students?

*Answer*: Path analysis

## Question 4

Do males and females have significantly different SAT scores?

*Answer*: t Test, One-way analysis of variance (ANOVA).

## Question 5

What is the relationship between SAT scores and college freshmen GPAs?

*Answer*: Bivariate Correlation and/or Regression.

## Question 6

Which risk-taking behaviors (amount of alcohol use, drug use, sexual activity, violence) distinguish suicide attempters from non-attempters?

*Answer*: logistic regression.

## Question 7

## Do adolescents from low, middle, and high socioeconomic status families have different literacy test scores after adjusting for family type?

*Answer*: One-way ANOVA or One-way ANCOVA (if other important covariates available)

## Question 8

Does ethnicity significantly affect reading achievement, math achievement, and overall achievement among sixth grade students after adjusting for family income?

*Answer*: One-way MANCOVA (family income is the covariate)

## Question 9

## Which combination of risk-taking behaviors (amount of alcohol use, drug use, sexual activity, and violence) best predicts the amount of suicide behavior among adolescents?

*Answer*: Discriminate Analysis if the amount of suicide behavior among adolescents is defined as categorical variable. Multiple regression if amount of suicide behavior among adolescents is defined as continuous/count variable.

## Question 10

Do preschoolers of low, middle, and high socioeconomic status have different literacy test scores?

*Answer*: One-way ANOVA or One-way ANCOVA (if other important covariates available)

## Question 11

Do ethnicity and learning preference significantly affect reading achievement, math achievement, and overall achievement among sixth grade students?

*Answer*: Factorial MANOVA or Factorial MANCOVA (if other important covariates available)

## Question 12

To what extent do certain risk-taking behaviors (amount of alcohol use, drug use, and sexual activity, and the presence of violent behavior) increase the odds of a suicide attempt occurring?

*Answer*: Logistic Regression

## Question 13

## Do ethnicity and learning preference significantly affect reading achievement, math achievement, and overall achievement among sixth grade students after adjusting for family income?

*Answer*: Factorial ANCOVA (family income is the covariate)

## Question 14

What underlying structure exists among the following variables: amount of alcohol use, drug use, sexual activity, school misconduct, cumulative GPA, reading ability, and family income?

*Answer*: Factor Analysis (Theoretical) or Principal Component Analysis (Empirical)

# Exercise in Chapter 3

## Question 1

You are interested in investigating if being above or below the median income (medloinc) impacts ACT means (act94) for schools. Complete the necessary steps to examine univariate grouped data in order to respond to the questions below. Although deletions and/or transformations may be implied from your examination, all steps will examine original variables.

* How many participants have missing values for medloinc and act94?

*Answer*: There is no missing value for medlonic and act94.

* Is there a severe split in frequencies between groups?

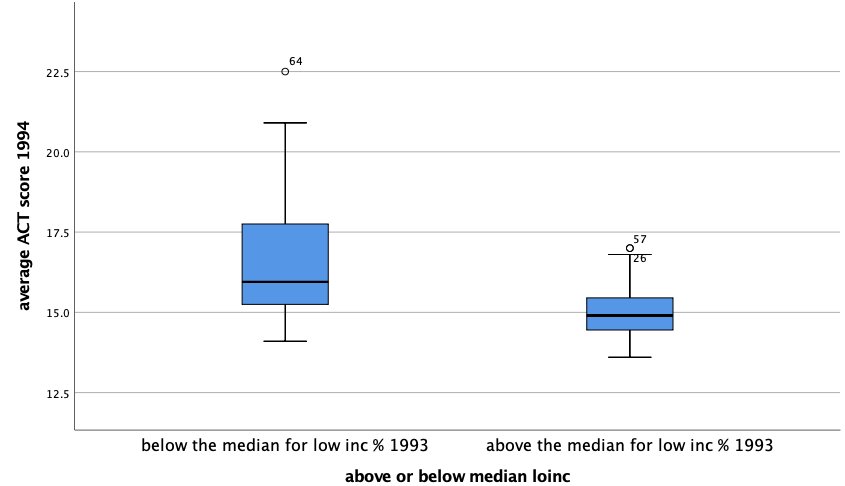
*Answer*: No. Both group (below and above the median) have the same frequency of 32. But in terms of the frequency in ACT score, based on the plot 1, the distributions are different.

* What are the cutoff values for outliers in each group?

*Answer*:

1. for below the median group
   1. Upper bound of act94: 17.475
   2. Lower bound of act94: 15.925
2. for above the median group
   1. Upper bound of act94: 15.337
   2. Lower bound of act94: 14.707

* Which outlying cases should be deleted for each group?

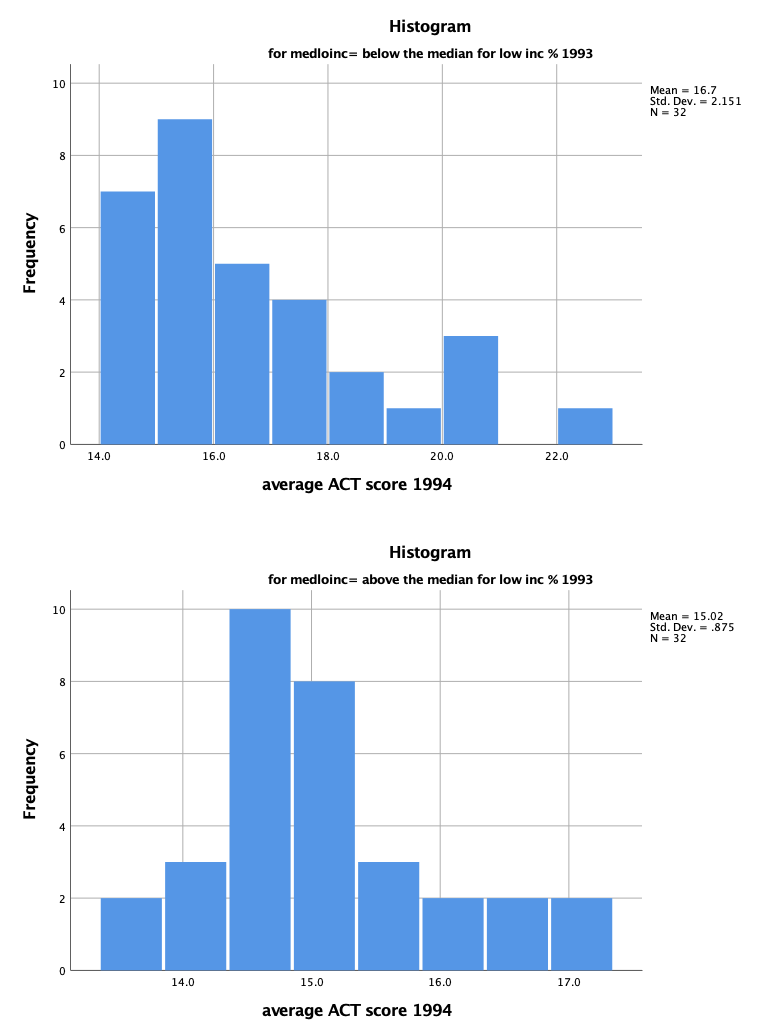
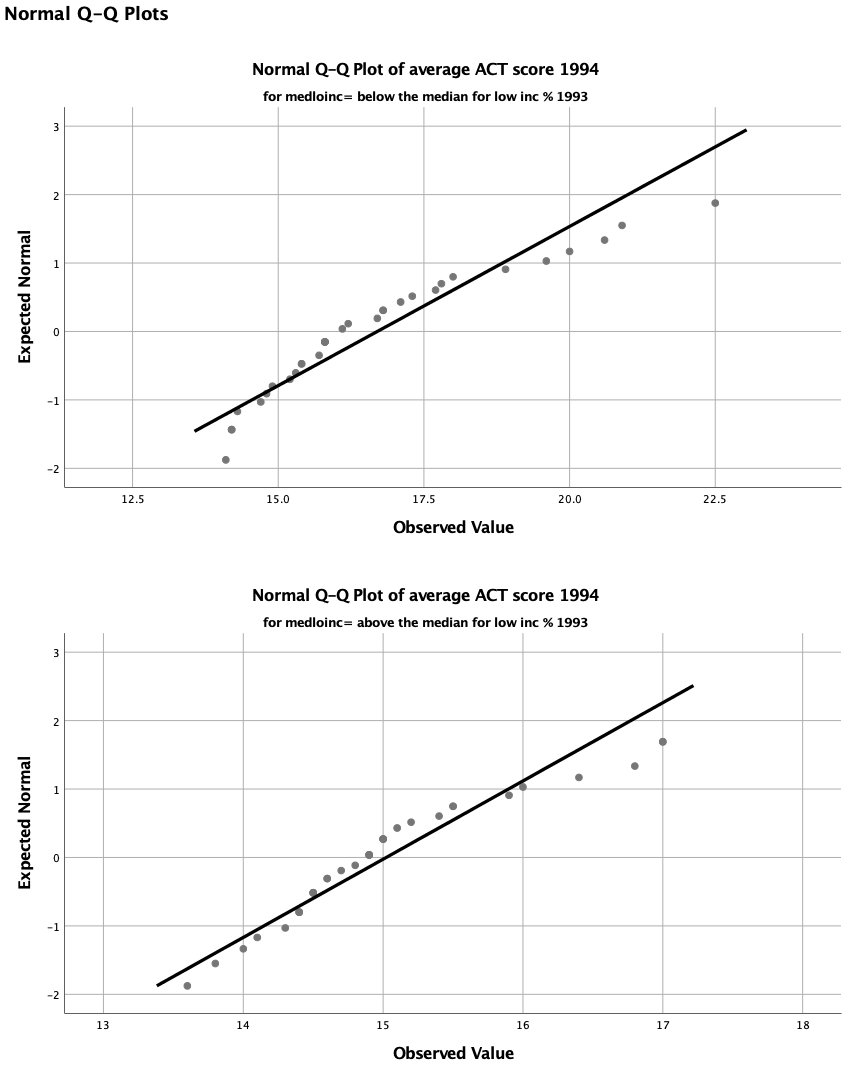
*Answer*: For the group of below median: the case 64 should be deleted. For the group of above median: case 26 and 57 should be deleted.

*Plot 1* distribution of DV for each group

* Analyzing histograms, normal Q-Q plots, and tests of normality, what is your conclusion regarding normality? If a transformation is necessary, which one would you use?

*Answer*: Based on the evidence from histogram, QQ plot, we see the normality assumption does not hold. Further, this judgement can be reinforced by the evidence from Kolmogorov-Smirnov and Shapiro-Wilk tests since both tests fail to reject the null hypothesis that the distribution under each group follows normal distribution.

Since the distributions differs only moderately from normal, we can choose a square root transformation initially. If after the transformation, the new distributions still do not follow normal, we can then try a log transformation.



*Plot 2* histograms (left) and QQ plot (right) for each group

* Do the results from Levene’s test for equal variances indicate homogeneity of variance? Explain.

*Answer*: No, the evidence from Leven’s test reject the null hypothesis of homogeneity of variance, no matter based on mean, median, trimmed mean, or median with adjusted degree of freedom. This conclusion can also be proved by plot 1, which show the significant difference in their distribution.

## Question 2

Examination of the variable of scienc93 indicates a substantial to severe positively skewed distribution. Transform this variable using the two most appropriate methods. After examining the distributions for these transformed variables, which produced the better alteration?

*Answer*: I choose the log transformation and inverse transformation since the original data is positively skewed. Based on the evidence from Kolmogorov-Smirnov and Shapiro-Wilk tests inverse transformation produce a better alternative.

## Question 3

You are interested in studying predictors (math94me, loinc93, and read94me) of the percentage graduating in 1994 (grad94).

* Examine univariate normality for each variable. What are your conclusions about the distributions? What transformations should be conducted?

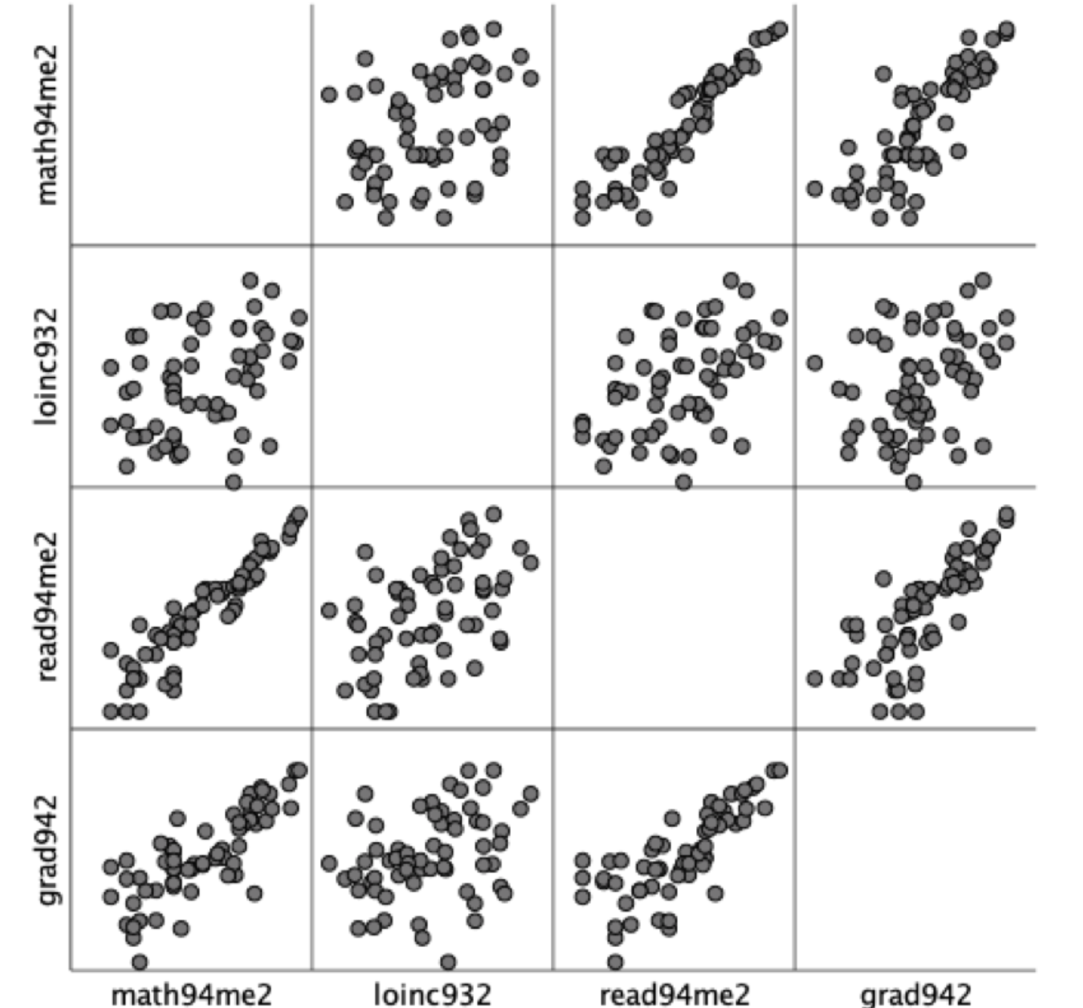
*Answer*: According to the Kolmogorov-Smirnov and Shapiro-Wilk tests, math94me and read94me does not follow normal distribution. Both distributions are positive skewed, while math94me differ more significantly from normal distribution. Consequently, I choose to use the log transformation for math94me and read94me. After transformation, both variable pass the normality check.

While, loinc93 cannot reject the normal distribution assumption. Consequently, there is no need for transformation. grad94 is the dependent variable, which do not need to follow the normal distribution.

* After making the necessary transformations, examine multivariate outliers using Mahalanobis distance. Which cases should be deleted?

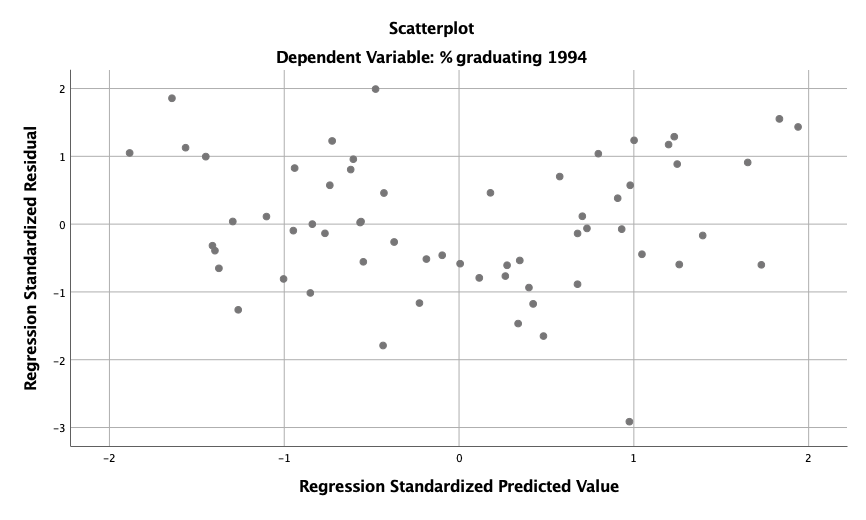
*Answer*: According to the Mahalanobis distance, the case 55 should be deleted.

* After deleting the multivariate outliers, examine multivariate normality and linearity by creating a Scatterplot Matrix.

*Answer*: The plot 3 shows that most combination shows certain level of linearity and some combination displays elliptical shape. However, the overall normality and linearity in this model is low.

*Plot 3* Scatterplot Matrix

* Examine the variables for homoscedasticity by creating a residuals plot (standardized vs. predicted values). What are your conclusions about homoscedasticity?

*Answer*: Based on the evidence from plot, the residual creates an approximate rectangular distribution which center at 0. This means, in general, the assumptions of linearity, normality, and homoscedasticity are met.

*Plot 3* Scatterplot of standardized vs. predicted values

# Exercise in Chapter 7

## Question 1

The following output was generated from conducting a forward multiple regression to identify which IVs (urban, birthrat, lnphone, and lnradio) predict lngdp. The data analyzed were from the SPSS country-a.sav data file.

* Evaluate the tolerance statistics. Is multicollinearity a problem?

*Answer*: The first proposed model is a simple regression model with only one independent variable. Thus, there is no need to worry about multicollinearity and the tolerance is 1. For the second one, Allison (1999) notes that there is no strict cutoff of tolerance, but suggest to tolerance of below 0.4 is cause of concern. Meanwhile, Norusis (1998) also notes that 0.1 can be a cutoff point for tolerance. Consequently, it is not a big problem of multicollinearity in general.

* What variables create the model to predict *lngdp*? What statistics support your response?

*Answer*: Based on the forward selection algorithm, two models are proposed. I choose the second model with the independent variable as inphone and birthrat. Mainly, it is because adjusted R square is higher since this statistic measure the overall model fit with consideration of number of parameters as well. (It will be good if AIC and BIC are available for these nested models, but I expect the result will be the same).

* Is the model significant in predicting *lngdp*? Explain.

*Answer*: Yes. First, the two variable all pass the t test which means their effect are significant. Secondly, the R square (Adjusted R square) indicates a high fit of model since both statistics are bigger than 0.88. Third, the F test in two-way ANOVA also indicates that the effect from two variables cannot be 0 simultaneously (F (1, 111) = 896.968, p < .001).

* What percentage of variance in *lngdp* is explained by the model?

*Answer*: Based on R square, we can say that 89% of the observed variance in dependent variable can be explained by the model.

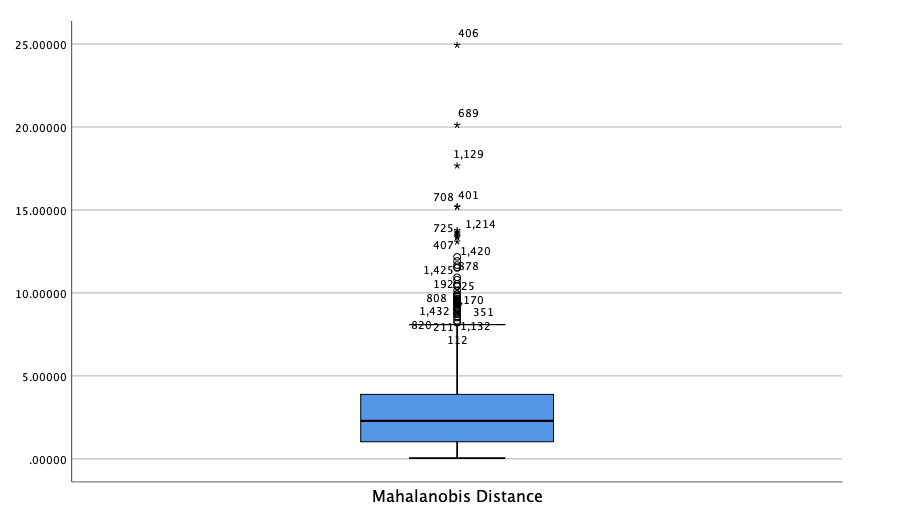
* Write the regression equation for *lngdp*.

*Answer*:

## Question 2

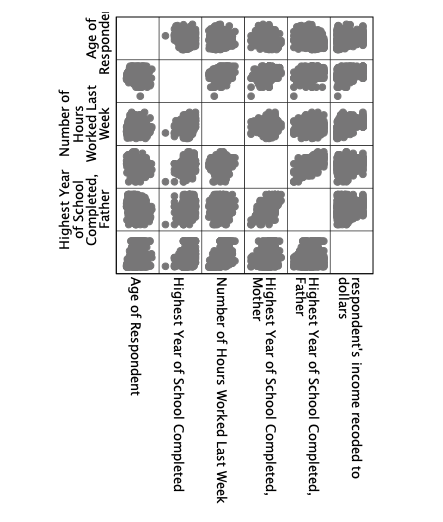
You are interested in examining whether the variables shown here in brackets [years of age (*age*), hours worked per week (*hrs1*), years of education (*educ*), years of education for mother (*maeduc*), and years of education for father (*paeduc*)] are predictors of individual income (*rincmdol*). Complete the following steps to conduct this analysis

* Using *profile-a.sav*, conduct a preliminary regression to calculate Mahalanobis distance. Identify the critical value for chi-square. Conduct **Explore** to identify outliers. Which cases should be removed from further analysis?

*Answer*: Based on the forward selection, all variables are kept. The critical value for chi-square is 12.7, and there are 37 cases should be deleted.

*Plot 4* Distribution of Mahalanobis Distance

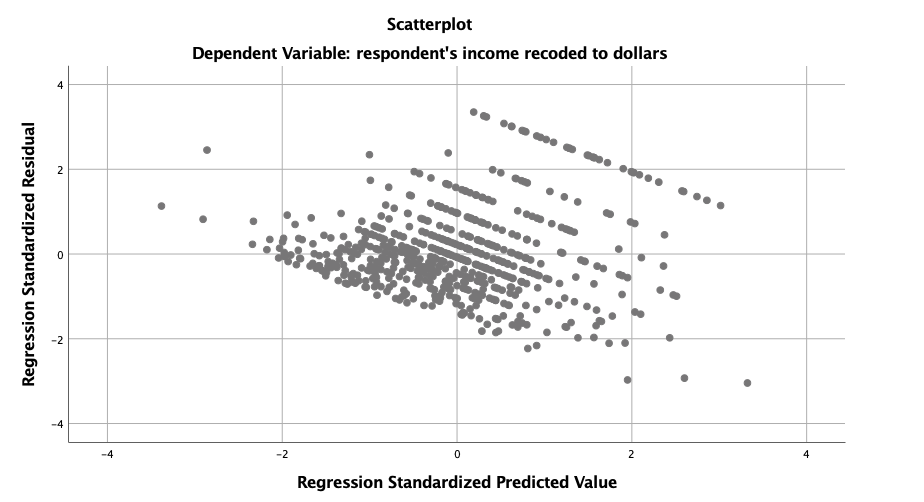
* Create a scatterplot matrix. Can you assume linearity and normality?

*Answer*: Based on the evidence from plot, the assumption of linearity and normality is rejected since most plot do not display elliptical shapeand the linear relationship is also unclear.

*Plot 5* Scatterplot Matrix

* Conduct a preliminary regression to create a residual plot. Can you assume normality and homoscedasticity?

*Answer*: based on the evidence from plot 6, the normality and homoscedasticity does not hold.

*Plot 6* Scatterplot of standardized vs. predicted values

* Conduct multiple regression using the Enter method. Evaluate the tolerance statistics. Is multicollinearity a problem?

*Answer*: Based on the enter method, all variables are kept. The tolerances are bigger than 0.1 for all variables. Thus, the multicollinearity is not a problem.

* Does the model significantly predict *rincmdol*? Explain.

*Answer*: AVOVA (F (5, 605) = 74.567, p < .0005) also indicates that the coefficients in the regression cannot be zero simultaneously, which shows a certain level of fit. However, the overall R square is 0.381, which indicate a relative limited model fit.

* Which variables significantly predict *rincmdol*? Which variable is the best predictor of the DV?

*Answer*: years of age (*age*), hours worked per week (*hrs1*), years of education (*educ*) pass the t test. Among these three, *hrs1* has the highest t value and highest standardized coefficient which mean it is the best predictor.

* What percentage of variance in *rincmdol* is explained by the model?

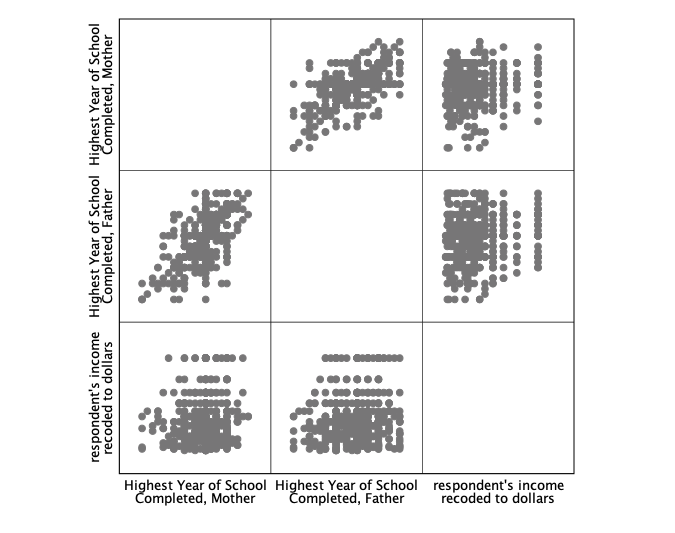
*Answer*: According to the R square statistics, 38.1% of the variance in the dependent variable has been explained by the model.

* Write the regression equation for the standardized variables.

*Answer*:

* Explain why the variables of mother’s and father’s education are not significant predictors of *rincmdol*.

*Answer*: *Maeduc* and *paeduc* do not pass the t test which means we cannot say that the effects from these two variables are significantly different from zero. We can see from plot7 that, there does not exist strong linear correlation between *rincmdol* and these two variables. Meanwhile, the normality assumption is also weak. Thus, it is not supervising to see their effects are not significant.

*Plot 7* Scatterplot Matrix