**RSCH 6120/8120: HW 4**

**Kristen Zhang | Nov 14**

This HW uses the eclsk.csv data. Context: We are going to investigate reading scores in the eclsk data and several student and school variables available. Review the eclsk codebook to understand the different variables available.

1. **Use graphical methods to check for relationships among the variables. ("read","math", "gen", "age", "income", "attend", "free.lunch"). Provide code and plot(s).**

**a) Identify a pair of variables with little to no relationship?**

**b) Identify a pair of variables that are positively correlated?**

**c) Identify a pair of variables that are negatively correlated?**

> cor(eclsk[,c("read", "math", "gen","age","income","attend","free.lunch")])

read math gen age income attend

read 1.0000000 0.6811514 0.4920539 0.14646540 0.27670170 0.1340969

math 0.6811514 1.0000000 0.6178206 0.24093458 0.27567338 0.1294169

gen 0.4920539 0.6178206 1.0000000 0.27826447 0.33271409 0.1246938

age 0.1464654 0.2409346 0.2782645 1.00000000 0.01876012 0.0468883

income 0.2767017 0.2756734 0.3327141 0.01876012 1.00000000 0.1173212

attend 0.1340969 0.1294169 0.1246938 0.04688830 0.11732116 1.0000000

free.lunch -0.2569902 -0.3144246 -0.4132580 -0.06578905 -0.40786668 -0.2785802

free.lunch

read -0.25699016

math -0.31442459

gen -0.41325801

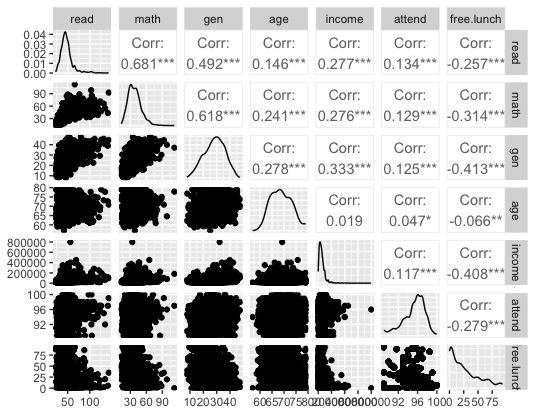
age -0.0657890**5**

income -0.40786668

attend -0.27858021

free.lunch 1.00000000

> ggpairs(eclsk[,c("read", "math", "gen","age","income","attend","free.lunch")])

****

1. A pair of variables with little/no relationships: age and income.
2. Positively correlated: read and math, read and gen, read and age, read and income, read and attend, math and gen, math and age, math and income, math and attend, gen and age, gen and income, gen and attend, age and attend, income and attend
3. Negatively correlated: read and free.lunch, math and free.lunch, gen and free.lunch, age and free.lunch, income and free.lunch, attend and free.lunch.

**2. Fit a multiple regression model to predict reading scores with at least three predictors. Remember: Better models generally have a higher . Provide code and results**

**a) Identify if each predictor is significant or not and interpret the coefficients as needed**

> model.multiple <- lm(read~math+gen+income, data=eclsk)

> summary(model.multiple)

Call:

lm(formula = read ~ math + gen + income, data = eclsk)

Residuals:

Min 1Q Median 3Q Max

-39.692 -5.278 -1.024 3.195 86.114

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 14.246601997 0.840007909 16.960 < 0.0000000000000002 \*\*\*

math 0.715896627 0.022305172 32.096 < 0.0000000000000002 \*\*\*

gen 0.176667873 0.035768020 4.939 0.000000837 \*\*\*

income 0.000024233 0.000004749 5.103 0.000000360 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.37 on 2431 degrees of freedom

Multiple R-squared: 0.4778, Adjusted R-squared: 0.4771

F-statistic: 741.3 on 3 and 2431 DF, p-value: < 0.00000000000000022

Interpretation: From the coefficients, we can tell that math scores, general intelligence scores, and income are all significantly positively related to the outcome: reading scores. That could be explained that, the increase in math scores, general intelligence scores, and income can lead to an increase in reading scores significantly. Increasing 1 in math scores will lead to 0.715896627 in reading scores; Increasing 1 in general intelligence scores will lead to 0.176667873 in reading scores; Increasing 1 dollar in family income will lead to 0.000024233 in reading scores.

**3. Model comparison: now add “female” to the multiple regression model you created for Question #2 and run the analysis.**

**a) Is female a significant predictor of reading scores?**

> model2 <- lm(read~math+gen+income+female, data = eclsk)

> summary(model2)

Call:

lm(formula = read ~ math + gen + income + female, data = eclsk)

Residuals:

Min 1Q Median 3Q Max

-38.418 -5.413 -1.155 3.316 87.389

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 15.455818658 0.855495812 18.067 < 0.0000000000000002 \*\*\*

math 0.715307493 0.022131707 32.320 < 0.0000000000000002 \*\*\*

gen 0.178673439 0.035490977 5.034 0.000000514661 \*\*\*

income 0.000024929 0.000004713 5.289 0.000000133699 \*\*\*

female -2.616320277 0.417335611 -6.269 0.000000000428 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.29 on 2430 degrees of freedom

Multiple R-squared: 0.4861, Adjusted R-squared: 0.4852

F-statistic: 574.6 on 4 and 2430 DF, p-value: < 0.00000000000000022

**b) Interpret the coefficients as needed.**

From the coefficients, we can tell that, besides math, gen, and income, female is also a significant predictor of reading scores. Considering that we labeled female (1), male (0), it means that the female will lead to lower reading scores.

**b) Compare the models (i.e., use anova() fuction in *R*). Are they significantly different? Which do you think is better and why?**

> anova(model.multiple,model2)

Analysis of Variance Table

Model 1: read ~ math + gen + income

Model 2: read ~ math + gen + income + female

Res.Df RSS Df Sum of Sq F Pr(>F)

1 2431 261475

2 2430 257313 1 4161.7 39.302 0.0000000004283 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The two models are significantly different (p = 0.0000000004283 \*\*\*). The second one (R2 = 0.4861) has larger explanatory power than the first one (R2 = 0.4778). I think the second one is better because it helps explain more variance in outcome.