**Project 2**

**SOCI556 Spring 2020 Name Nicole Dennis**

This assignment is due on Thursday, March 5th.

To complete this assignment, please use the dataset Learning.sav.

Learning.sav contains 5 variables and a sample of 436 cases. The dataset contains demographic information of adult learners and their perception of the learning environment and their learning experiences.

Variables: sex—learner’s sex (1=male; 2=female)

age—learner’s age

learning—learner’s perceived learning (achievement)

ppres—learner’s perceived teaching presence

isolation—learner’s perceived isolation in their learning settings

**Question 1**

Your first concern is that “teaching presence” (interaction with instructors, etc.) may affect learner’s perception of how much they have learned; and you decide to use a simple linear regression model to estimate the relation between the two variables.

1. **Obtain a scatterplot of “learning” and “ppres.” Does the scatter plot warrant a linear regression model? What is your evidence? (2 point)**

Learninga<-read.spss("Learning.sav", header=TRUE,

use.value.labels = FALSE,

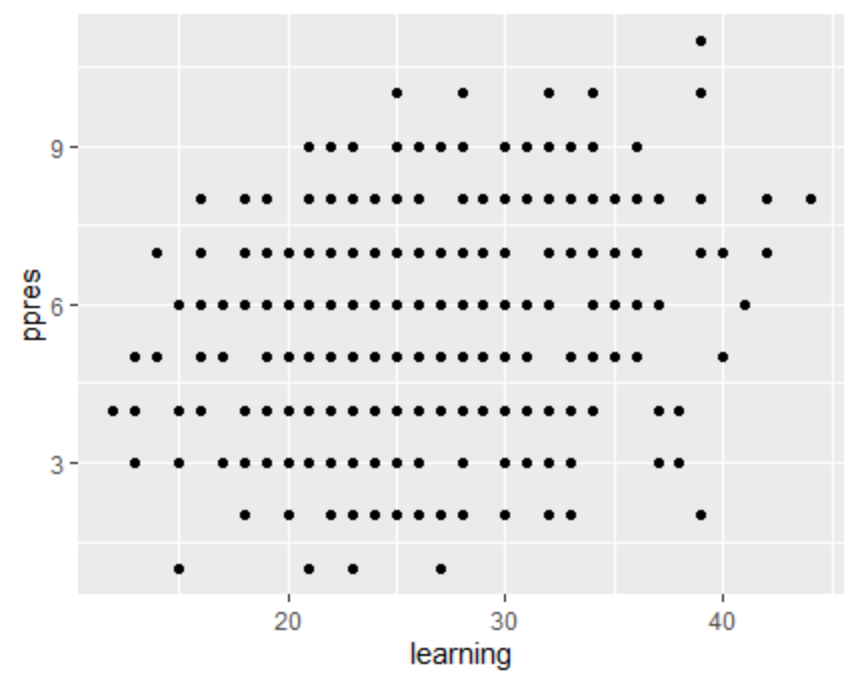
to.data.frame = TRUE, use.missings = TRUE)

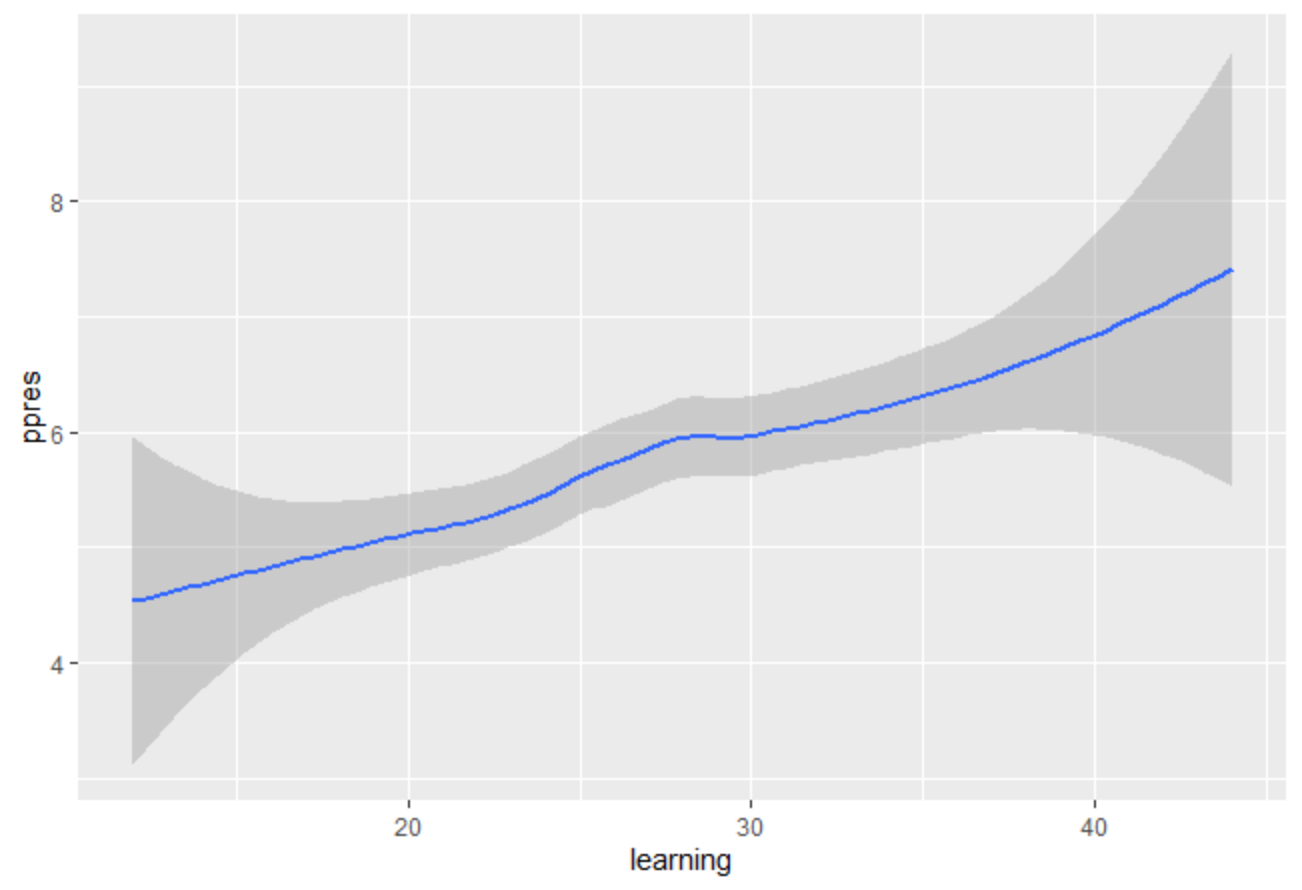
Learninga <- Learninga %>% mutate(sex=recode\_factor(sex, '1'="Male", '2'="Female"))

glimpse(Learninga)

ggplot(Learninga, aes(x=learning, y=ppres))+geom\_point()

2nd graph: Learninga%>%ggplot(mapping=aes(x=learning, y=ppres))+geom\_smooth(method=loess)





**This scatter plot does warrant a linear regression model because there is a visual relationship between the two variables learning and ppres.**

1. **Estimate the model and present the model’s “Summary,” “ANOVA,” and “Coefficients” tables. (1 points)**

mod1<-lm(ppres~learning, Learninga)

summary(mod1)

Browse[2]> summary(mod1)

Call:

lm(formula = ppres ~ learning, data = Learninga)

Residuals:

Min 1Q Median 3Q Max

-4.7352 -1.3798 0.1037 1.4260 4.4260

Coefficients:

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

(Intercept) 3.55973 0.45729 7.784 5.44e-14 \*\*\*

learning 0.08057 0.01680 4.795 2.26e-06 \*\*\*

---

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

Residual standard error: 1.979 on 423 degrees of freedom

(11 observations deleted due to missingness)

**Multiple R-squared: 0.05155,** Adjusted R-squared: 0.04931

F-statistic: 22.99 on 1 and 423 DF, p-value: 2.259e-06

coef(mod1)

(Intercept) learning

3.55973430 0.08057242

anova(mod1)

Analysis of Variance Table

Response: ppres

Df Sum Sq Mean Sq F value Pr(>F)

learning 1 90.04 90.039 22.991 2.259e-06 \*\*\*

Residuals 423 1656.61 3.916

---

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

1. **Interpret the intercept and the slope. (3 points)**

The number corresponding to the y intercept is 3.55973430 and the slope of the equation is 0.08057242.

To put this in the form of an equation y=0.08057242\*x + 3.55973430)

1. **What percentage of variance in the dependent variable is explained (or predicted) by the model? (1 point)**

5% of the variance in the dependent variable is explained by the model.

1. **Obtain two residual plots: the Normal Q-Q plot and the scatterplot against the predicted values. Is there obvious evidence showing that the residuals are not normally distributed? Why, or why not? It there detectable evidence that the variances are not constant? Why or why not? (2 points)**

**Question 2**

As a researcher, your main concern is how much adult learners feel isolated in their learning settings. You hypothesize that sex, age, and perception of teaching presence affect their experience of isolation. In order to examine the relations between “isolation” and the other three variables, you decide to use a multiple regression model. Please take the following steps and answer questions related to these steps:

1. **Estimate the linear regression model. (2 points)**
2. **What variables have significant effects on the dependent variables? At what levels of significance? (2 points)**
3. **Interpret the coefficients (or parameter estimates) that are significant? (5 points)**
4. **How much variance in the dependent variable is explained by your model? What is your evidence? (1 points)**
5. **What is the largest value of the variance inflation factor (VIF)? Is multicollinearity a problem in your model? (2 points)**
6. **Obtain a residual plot against the predicted values. Is heteroscedasticity an issue to be resolved? Why or why not? (2 points)**

**Question 3**

(Continued from problem 2) When one plots the variable “isolation” against “age,” one has reason to believe that the relation between the two variables is curvilinear (a quadratic relation is detectible).

1. **Obtain a scatterplot for the relation between “isolation” and “age.” Does the plot justify a model with a quadratic term? (2 points)**
2. **Estimate a model to include all the independent variables in problem 2, plus a quadratic term of “age.” Is this quadratic term significant? (Please make sure you center the variable “age” to avoid issues of multicollinearity) (3 points)**
3. **Compare the two sets of “Summary” table produced with and without the quadratic term, what changed from the first set of tables to the second set of tables? (2 points)**
4. **Check whether multicollinearity is a problem. Show your evidence. (1 point)**