Emily Chu Assignment 7&8

Code

Assignment

Data frame for this assignment: empdat_cor (download from Canvas)

Part I Correlation

1. What is the appropriate measure(s) of association/correlation for current salary and education? Why?
Pearson's r - we are testing the strength between two variables.

2. Generate the appropriate measures(s) of association/correlation for current salary and education. Is the relationship significant? Why?

Hide

```
cor(empdat_cor$salary, empdat_cor$educ, method="pearson", use="complete.obs")
```

```
[1] 0.6605589
```

Hide

```
cor.test(empdat_cor$salary, empdat_cor$educ) # give us the p - value, is it significant?
```

```

    Pearson's product-moment correlation

data:  empdat_cor$salary and empdat_cor$educ
t = 19.115, df = 472, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6065810 0.7084748
sample estimates:
      cor
0.6605589
```

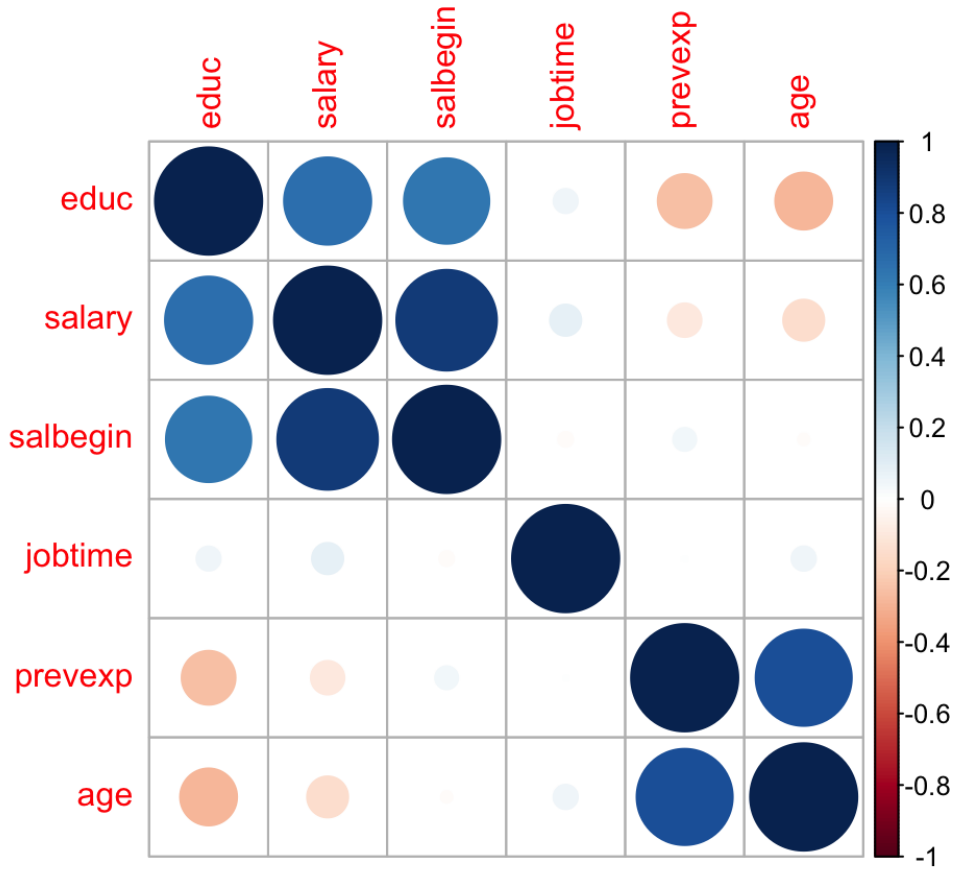
Hide

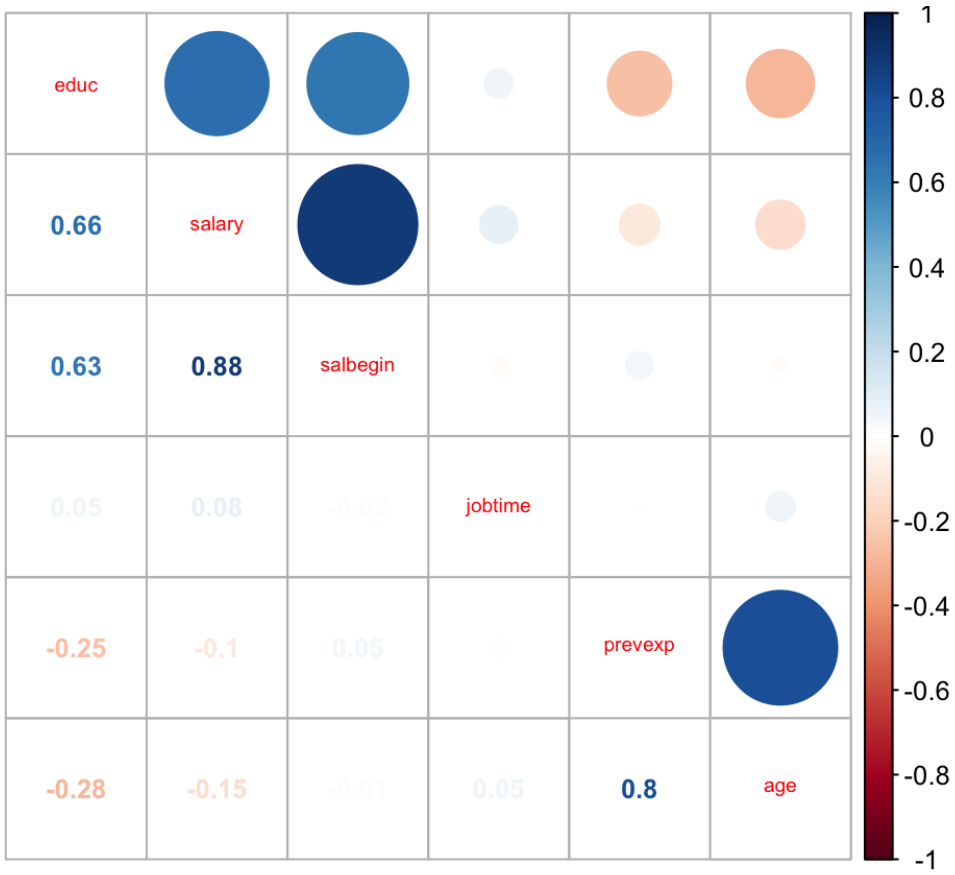
```
# What does R squared tell us?
CE <- cor(empdat_cor$salary, empdat_cor$educ, method="pearson", use="complete.obs")
CE^2
```

```
[1] 0.4363381
```

No, the relationship is not significant with p-value <2.2e-16. 44% of variation in salary is explained by education, but it is not enough to be considered statistically significant.

3. Choose one of the three visualization plots to show the correlation matrix for all the variables in the dataset empdat_cor. Which two variables have the strongest correlation relationship?

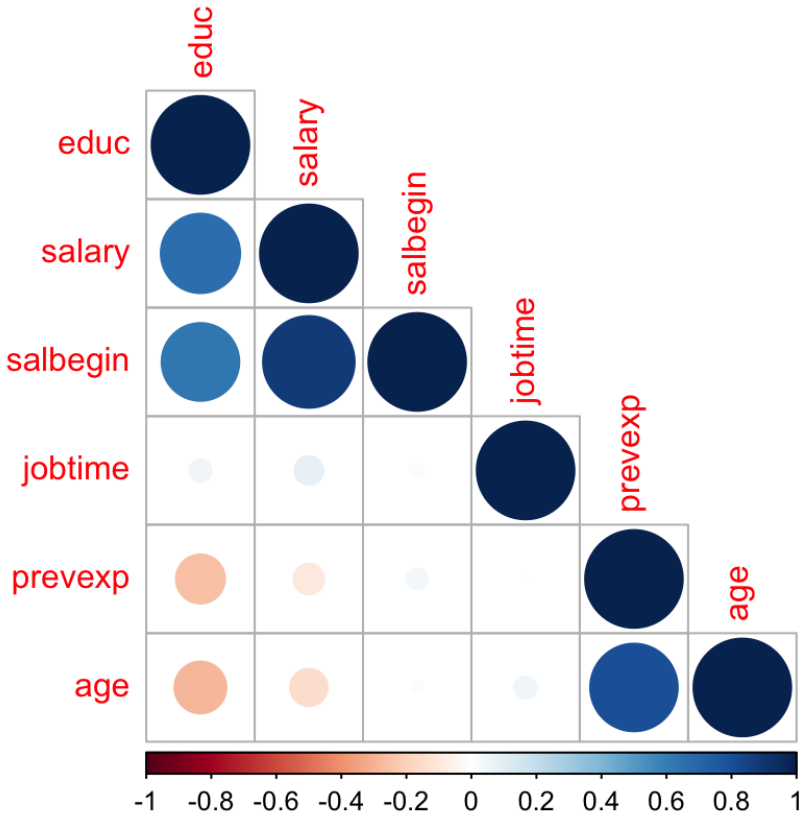

educ
salary
salbegin
jobtime
prevexp
age
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1


educ
0.66
salary
0.63
0.88
salbegin
0.05
0.08
-0.02
jobtime
-0.25
-0.1
0.05
prevexp
-0.28
-0.15
-0.01
0.05
0.8
age
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Hide

```
corrplot(corrmatrix, method="circle") # corrmatrix is the name of the correlation matrix
 we created above
corrplot.mixed(corrmatrix, number.cex = 0.8, tl.cex = 0.6)

#number.cex changes the size of the number fonts. tl.cex changes the size of the labels
corrplot(corrmatrix, type="lower")
```



Part II Simple Regression

1. Generate and interpret a simple regression model for the relationship between current salary and age. Using the 'sjPlot' package to show the regression output.

a. Show your regression model output.

Hide

```
regression_1 <- lm(salary ~ age, data = empdat_cor)
#R default regression output:
summary(regression_1)
```

```
Call:
lm(formula = salary ~ age, data = empdat_cor)

Residuals:
   Min     1Q Median     3Q    Max
-18950 -10175  -5892   2685 103190

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 41332.20    2296.37    18.0  < 2e-16 ***
age          -211.61      66.12    -3.2  0.00147 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16930 on 471 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.02128,   Adjusted R-squared:  0.0192
F-statistic: 10.24 on 1 and 471 DF,  p-value: 0.001466
```

```

*b. How much variance in salary can be explained by age?*
About 21% variance in salary can be explained by age (unadjusted)

*c. Is age a significant predictor of salary? How do you know?*
Yes. It is significant at the p<.01 level, which meets the standard p<.05 threshold.

*d. If someone's age increased by one year, by how much would you expect their salary to change?*
21%

*e. Is the model significant? Why?*
Yes, the model is significant because age is a significant predictor of salary.

*f. Speculate on the nature of the relationship between age and salary in this dataset.*
As age increases, salary increases.
```