

ECN 3972
Lab 8

Make sure you call up mosaic.

Then, we are again going to use a built-in data set called CPS85.

In R:
`data(CPS85)`

But, we want to see what all data is here. So, let's look at the first few rows.

In R:
`head(CPS85)`

The variables we will be using are:

wage = workers hourly wage rate
educ = years of education
sex = M if male; = F if female
exper = years of work experience
union = Union if worker is in labor union; = Not if not
age = age of worker

Let's try two different models, one without the variable age and one with the variable age:

Model 1

$$Wage = \beta_0 + \beta_1 Exper + \beta_2 Educ + \beta_3 Sex + \beta_4 Union + e$$

In R:
`model1 = lm(wage~exper + educ + sex + union, data=CPS85)`
`summary(model1)`

Model 2

$$Wage = \beta_5 + \beta_6 Age + \beta_7 Exper + \beta_8 Educ + \beta_9 Sex + \beta_{10} Union + e$$

In R:
`model2 = lm(wage~age + exper + educ + sex + union, data=CPS85)`
`summary(model2)`

What happens? Why?

Let's check correlations between our continuous independent variables.

In R:

```
cor(CPS85$age, CPS85$exper)
cor(CPS85$age, CPS85$educ)
cor(CPS85$exper, CPS85$educ)
```

This is fine with only three independent variables but would get tedious with many more. So, it might be best to get a matrix of correlations of our variables.

But, before we can do so, we need to reduce the data to only numeric columns. (You can't do correlations with string variables.) So, we need to create a subset of the data containing only the columns that are numeric. You need to know which column numbers this is. (That's why we did the head earlier – we can see that we need columns 1, 2, 8, and 10.)

In R:

```
numdata <- CPS85[, c(1, 2, 8, 10)]
```

Now to get the correlation matrix:

In R:

```
cor(numdata)
```

We can also see these relationships using scatterplots.

In R:

```
xyplot(age~exper, data=CPS85)
xyplot(age~educ, data=CPS85)
xyplot(exper~educ, data=CPS85)
```

Last, let's calculate a variance inflation factor for our 3 continuous variables.

First step is to run a model with each of our independent variables (age, educ, and exper) as the dependent variable and a function of the other variables.

In R:

```
model3a = lm(age~educ + exper + sex + union, data=CPS85)
summary(model3a)
```

```
model3b = lm(educ~age + exper + sex + union, data=CPS85)
summary(model3b)
```

```
model3c = lm(exper~age + educ + sex + union, data=CPS85)
summary(model3c)
```

Then we can calculate a VIF for each variable:

$$\text{VIF} = \frac{1}{(1 - R_i^2)}$$

What if we just dropped “age” from our models? Let’s recheck the VIF for our other two variables.

In R:

```
model4a = lm(educ~ exper + sex + union, data=CPS85)  
summary(model4a)
```

```
model4b = lm(exper~ educ + sex + union, data=CPS85)  
summary(model4b)
```

Then we can calculate the VIF again. Much better, right?