Lab 1d. Linear Regression

Interaction between two quantitative variables

Yerim Oh

The purpose of this lab is to study the interaction between two quantitative variables. We will do this through a simulation study.

a) Use the following code to simulate data about income based on education and seniority.

```
set.seed(321)
n         <- 50
seniority <- trunc(rexp(n,.15)*12)/12
educ      <- trunc(runif(50,10,22)*2)/2
income    <- -30 + 4.5 * educ + 1.3 * seniority +  .2 * educ * seniority + rnorm(n,0, 1.5)
simdata   <- data.frame(income = income, educ = educ, seniority =seniority)
```

b) Examine the first few rows of the simdata dataframe.

```
head(simdata)
```

```
##     income educ seniority
## 1 20.68781 11.0  1.083333
## 2 64.40999 16.5  4.750000
## 3 70.79265 14.5  8.333333
## 4 38.39979 10.0  7.000000
## 5 97.05892 21.0  5.916667
## 6 86.96741 21.0  4.083333
```

- explanatory variable: `educ`, `seniority`
- response variable: `income`

$Y = \beta_0 + \beta_1 \times \text{seniority} + \beta_2 \times \text{educ} + \beta_3 \times \text{seniority} \times \text{educ} + \varepsilon$

c) Write down a model for the data that includes both variables and an interaction.

$Y = \beta_0 + \beta_1 \text{seniority} + \beta_2 \text{education} + \beta_3 \text{seniority} \times \text{education} + \varepsilon$

d) Fit the model for the model in part c) and print the summary().

```
model_fit  <- lm(formula = income ~ seniority + educ + seniority * educ, data = simdata)
summary(model_fit)
```

```
##
## Call:
## lm(formula = income ~ seniority + educ + seniority * educ, data = simdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2536 -0.8569  0.0452  0.7676  2.6937
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -30.812475   1.252004  -24.61  < 2e-16 ***
```

```
## seniority        1.327412   0.142574    9.31 3.76e-12 ***
## educ             4.557960   0.075138   60.66  < 2e-16 ***
## seniority:educ   0.196724   0.008367   23.51  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.269 on 46 degrees of freedom
## Multiple R-squared:  0.999,  Adjusted R-squared:  0.9989
## F-statistic: 1.519e+04 on 3 and 46 DF,  p-value: < 2.2e-16
```

e) Write the equation for the predicted value of Y based on the model fit in d).

$\hat{Y} = -30.81 + 1.33 \cdot \text{seniority} + 4.56 \cdot \text{education} + 0.20 \cdot \text{seniority} \times \text{education}$

$\hat{Y} = -30.81 + 1.33 \times \text{seniority} + 4.56 \times \text{educ} + 0.197 \times \text{seniority} \times \text{educ}$

f) Extract the estimated coefficients from the model fit using the function coef()

```
(coef_wi  <- coef(model_fit))
```

```
##    (Intercept)      seniority           educ seniority:educ
##    -30.8124753      1.3274115      4.5579600      0.1967239
```

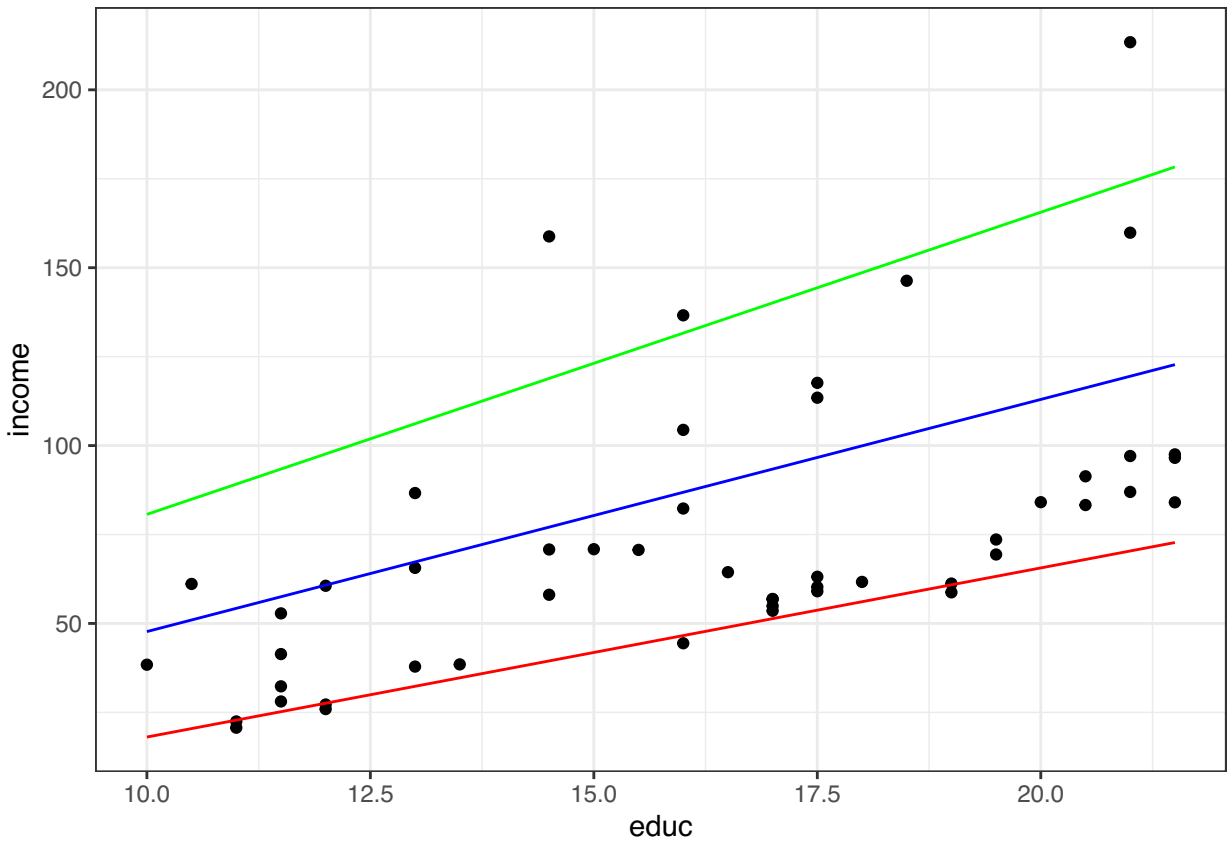
g) Using the estimated coefficients and the mutate function, add 3 variables to the simdata dataframe: the predicted values for the income based on the observed value for educ, but assuming

(i) seniority is equal to 1
(ii) seniority is equal to 10
(iii) seniority is equal to 20

```
simdata <- simdata %>%
  mutate(yhat_s1  = coef_wi[1] + coef_wi[2]*1 + coef_wi[3]*educ + coef_wi[4]*1*educ,
         yhat_s2  = coef_wi[1] + coef_wi[2]*10 + coef_wi[3]*educ + coef_wi[4]*10*educ,
         yhat_s3  = coef_wi[1] + coef_wi[2]*20 + coef_wi[3]*educ + coef_wi[4]*20*educ)
```

h) Plot the estimated lines of income vs educ for the three levels of seniority.

```
ggplot(simdata, aes(x=educ, income)) +
  geom_point() +
  geom_line(aes(y = yhat_s1), col="red") +
  geom_line(aes(y = yhat_s2), col="blue") +
  geom_line(aes(y = yhat_s3), col="green") +
  theme_bw()
```



i) Based on this exercise, explain in your own words why is the interpretation of the coefficients not straightforward when there is an interaction between two quantitative variables in the model.

We have a different slopes for education for every value of seniority (infinite number of slopes)