Linear Regression Extensions Recap

Dr. Soliman

Wages, education and gender in 1985

1. Load the data `CPS1985` from the `AER` package. I will help you with this step. First, in the command line, type `install.packages("AER")`, and hit enter. Do you have to run `install.packages("AER")` every time or only once? Why? How about `library(AER)`?

No, you install just once and load the package each time you need it.

2. Look at the `help` to get the definition of each variable.

```
help(CPS1985)
```

3. We don't know if people are working part-time or full-time, but does it matter here? Answer yes or no and why.

It doesn't matter much since we are analyzing hourly wages.

4. Create a `log_wage` variable equal to the log of the variable `wage`, but assign the original `CPS1985` object you currently have in your environment to a new object/dataframe and call it something funny. Use this object/dataframe for the rest of the assignment.

```
cps_drsolimanistoocoolforschool <- CPS1985 %>%
    mutate(log_wage = log(wage))
```

5. Regress `log_wage` on `gender` and `education`, and save it as `reg1`. Interpret each coefficient.

```
reg1 <- lm(log_wage ~ gender + education, cps_drsolimanistoocoolforschool)
reg1
```

```
##
## Call:
## lm(formula = log_wage ~ gender + education, data = cps_drsolimanistoocoolforschool)
##
## Coefficients:
##  (Intercept)  genderfemale     education
##      1.16519      -0.23207       0.07685
```

This is a log-level regression so the coefficients are interpreted as a unit increase in x is associated with some percent change in y. However, first, on average, men with no education earn exp(1.16519) = 3.21 dollars an hour (or 1.16519 higher log hourly wage, but that doesn't sound as good…). Turning to the other coefficients, on average, women earn exp(-0.23207) = 0.79 percent of what men earn (i.e., 21% less), holding education constant. Note that since the coefficient is relatively small, you can approximate the interpretation directly from the coefficient: on average, women earn 23% (b1x100) less than men with the same level of education.

Holding gender constant, a 1-year increase in years of education is associated, on average, with 8% more in hourly earnings (exp(0.07685) = 1.08). Again, since the coefficient is quite small, you can use the approximation we saw before, i.e., controlling for gender, on average, a 1 year increase in education is associated with a 7.7% increase in hourly wage.

6. Regress the `log_wage` on `gender`, `education` and their interaction `gender*education`, save it as `reg2`. Interpret each coefficient. Does the gender wage gap decrease with education?

```
reg2 <- lm(log_wage ~ gender*education, cps_drsolimanistoocoolforschool)
reg2
```

```
##
## Call:
## lm(formula = log_wage ~ gender * education, data = cps_drsolimanistoocoolforschool)
##
## Coefficients:
##            (Intercept)           genderfemale              education
##                1.32352               -0.63315                0.06468
## genderfemale:education
##                0.03080
```

On average, men with no education earn exp(1.32352) = 3.76 dollars an hour. Women with no education, on average, earn about exp(-0.63315) = 0.53 percent of what men with no education earn (or 47% less); note that since this coefficient is quite large, we must use the exp() function first, not look directly at the coefficient. On average, for men, a 1-year increase in years of education is associated, on average, with a 6.4% increase in hourly wages (or more accurately: exp(0.06468) = 1.07, so 7%). The difference in the return to education between women and men is about exp(0.03080) = 1.03 percentage points, that is for each additional year of education, women's hourly wage increases by about exp(0.06468 + 0.03080) = 1.1, i.e. 10%. Graphically, this means that the slope for women is steeper than that for men. The gender pay gap therefore shrinks with years of education.

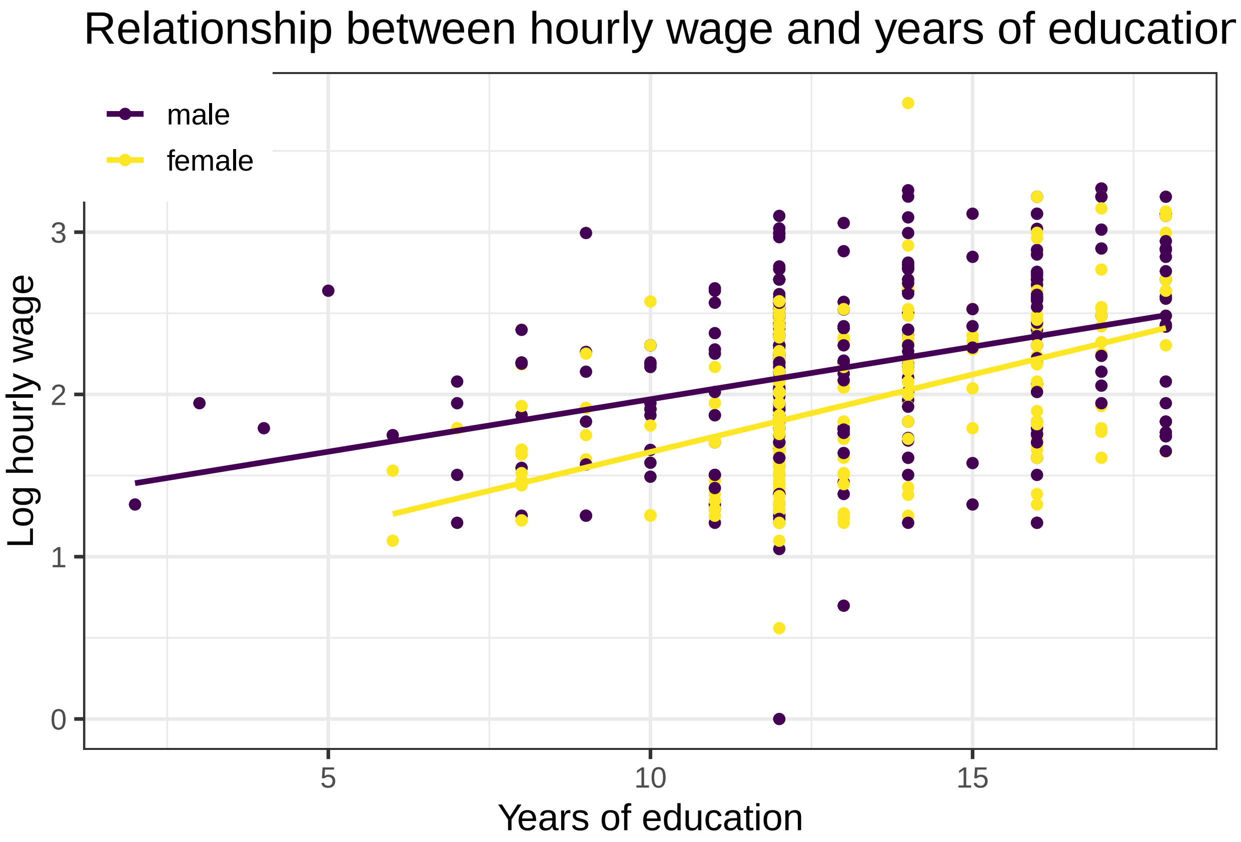
7. Create a plot showing this interaction. (*Hint:* use the `color = gender` argument in `aes` and `geom_smooth(method = "lm", se = F)` to obtain a regression line per gender.)

What we saw in the previous answer is confirmed here graphically.

```
ggplot(cps_drsolimanistoocoolforschool, aes(x = education, y = log_wage, col = gender)) +
    geom_point() +
    geom_smooth(method= "lm", se = F) +
    scale_color_viridis_d() +
    labs(x = "Years of education", y = "Log hourly wage",
         title = "Relationship between hourly wage and years of education by gender", color = NULL) +
    theme_bw(base_size = 14) +
    theme(legend.position = c(0,1),
          legend.justification = c(0,1))
```

```
## Warning: A numeric `legend.position` argument in `theme()` was deprecated in ggplot2
## 3.5.0.
## i Please use the `legend.position.inside` argument of `theme()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## `geom_smooth()` using formula = 'y ~ x'
```


Relationship between hourly wage and years of educatior
male
female
Log hourly wage
0
1
2
3
5
10
15
Years of education