Regression Inference Recap Answers

Dr. Soliman

1. Load the data `CPS1985` from the `AER` package and look back at the help to get the definition of each variable: `?CPS1985`

2. Create the `log_wage` variable equal to the log of `wage`.

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

3. Regress `log_wage` on `gender` and `education`, and save it as `reg1`.

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

```
##
## Call:
## lm(formula = log_wage ~ gender + education, data = cps)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.08737 -0.35123  0.03087  0.33134  1.78649
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.165192   0.106130  10.979   <2e-16 ***
## genderfemale -0.232067   0.041254  -5.625    3e-08 ***
## education     0.076848   0.007867   9.768   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.475 on 531 degrees of freedom
## Multiple R-squared:  0.1928, Adjusted R-squared:  0.1898
## F-statistic: 63.42 on 2 and 531 DF,  p-value: < 2.2e-16
```

- Interpret each coefficient.

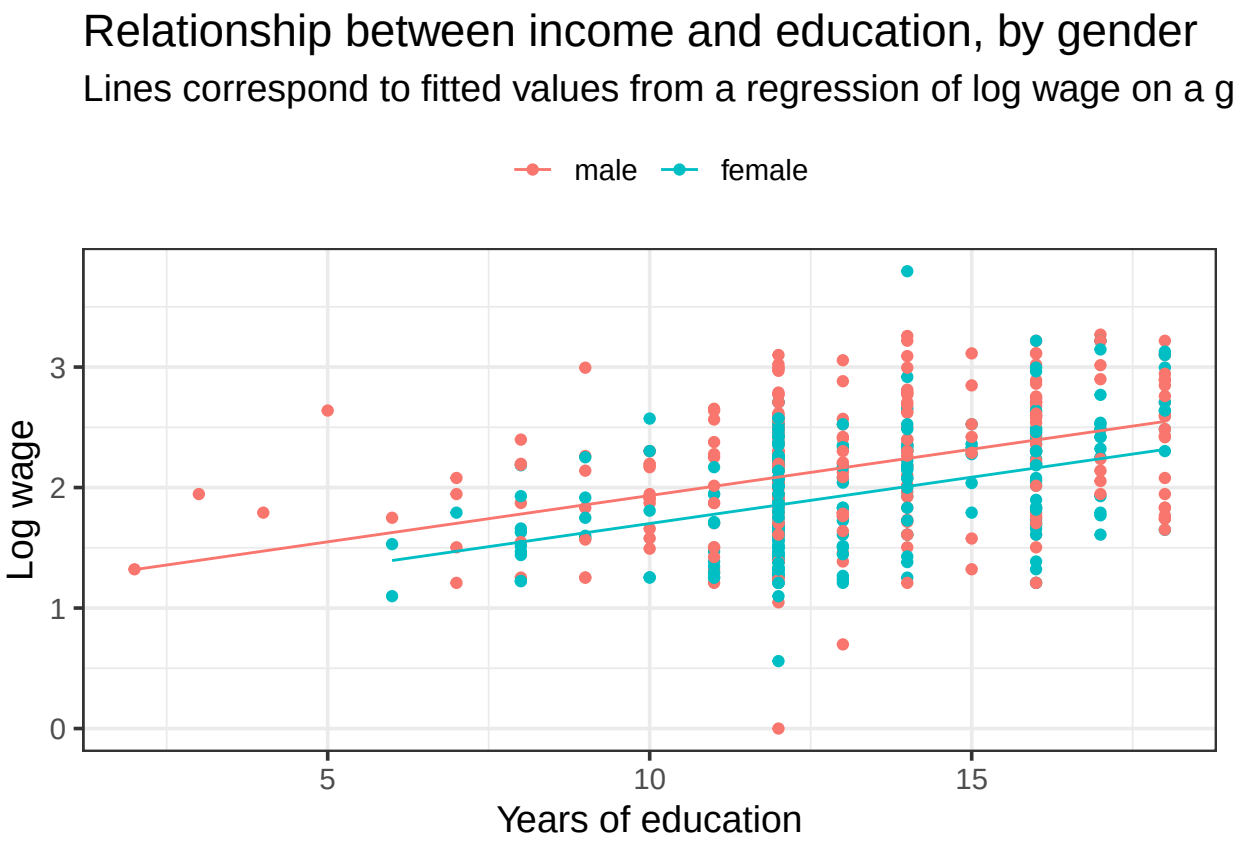
Intercept: Expected log wage of men without any education.
`genderfemale`: On average, women earn about 23% less than men controlling for education level.
`education`: An additional year of education is associated, on average, with a 7.68% increase in wages controlling for gender.

One way to see this regression is as follows:

```
reg1_aug <- broom::augment(reg1)

ggplot(cps, aes(x = education, y = log_wage, color = gender)) +
  geom_point() +
  geom_line(data = reg1_aug %>% filter(gender == "male"), aes(x = education, y = .fitted)) +
  geom_line(data = reg1_aug %>% filter(gender == "female"), aes(x = education, y = .fitted)) +
```

```
  labs(
    x = "Years of education",
    y = "Log wage",
    color = "",
    title = "Relationship between income and education, by gender",
    subtitle = "Lines correspond to fitted values from a regression of log wage on a gender dummy and e
  ) +
  theme_bw(base_size = 14) +
  theme(legend.position = "top")
```



Both lines have the same slope. This slope is equal to the `education` coefficient from the previous regression. The distance between the two lines is equal to the coefficient on `gender`. Given a certain level of education, on average, the difference between men and women's wages will be 23%. The intercept is equal to the value of the y-axis for the red line when years of education is equal to 0.

- Are the coefficients statistically significant? At which significance level?

All the coefficients are statistically significant at the 1% level since their p-values are all very close to 0.

4. Regress the `log_wage` on `gender`, `education` and their interaction `gender*education`, save it as `reg2`.

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

```
##
## Call:
## lm(formula = log_wage ~ gender * education, data = cps)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.09971 -0.36313  0.03421  0.33156  1.76830
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.32352    0.13422   9.861  < 2e-16 ***
## genderfemale            -0.63315    0.21303  -2.972  0.00309 **
## education                0.06468    0.01009   6.411 3.19e-10 ***
## genderfemale:education   0.03080    0.01605   1.919  0.05554 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4738 on 530 degrees of freedom
## Multiple R-squared:  0.1984, Adjusted R-squared:  0.1938
## F-statistic: 43.72 on 3 and 530 DF,  p-value: < 2.2e-16
```

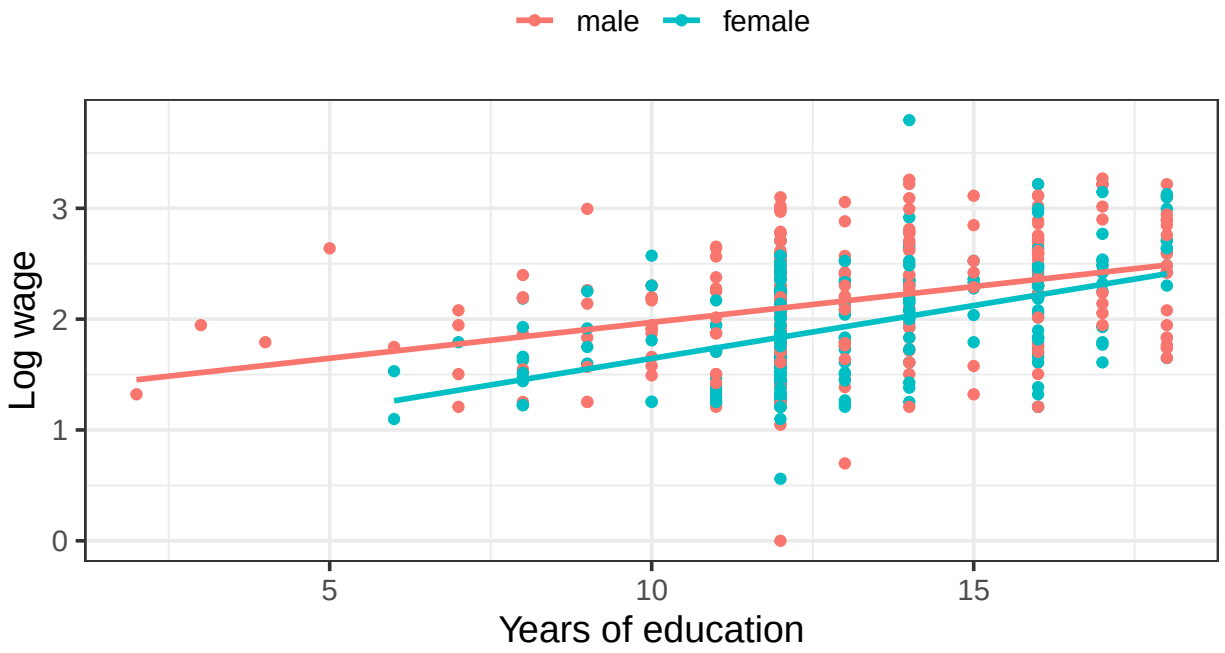
- How do you interpret the coefficient associated to `female*education`?

One additional year of education is associated, on average, with an additional 3% increase in wages for women relative to men. In other words, one additional year of education is associated for women, on average, with a 9.5% increase in wages, while it is *only* 6.5% for men.

One way to see this regression is as follows:

```
ggplot(cps, aes(x = education, y = log_wage, color = gender)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    x = "Years of education",
    y = "Log wage",
    color = "",
    title = "Relationship between income and education, by gender",
    subtitle = "Lines correspond to fitted values from a regression of log wage on a gender dummy, educa
  ) +
  theme_bw(base_size = 14) +
  theme(legend.position = "top")
```

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



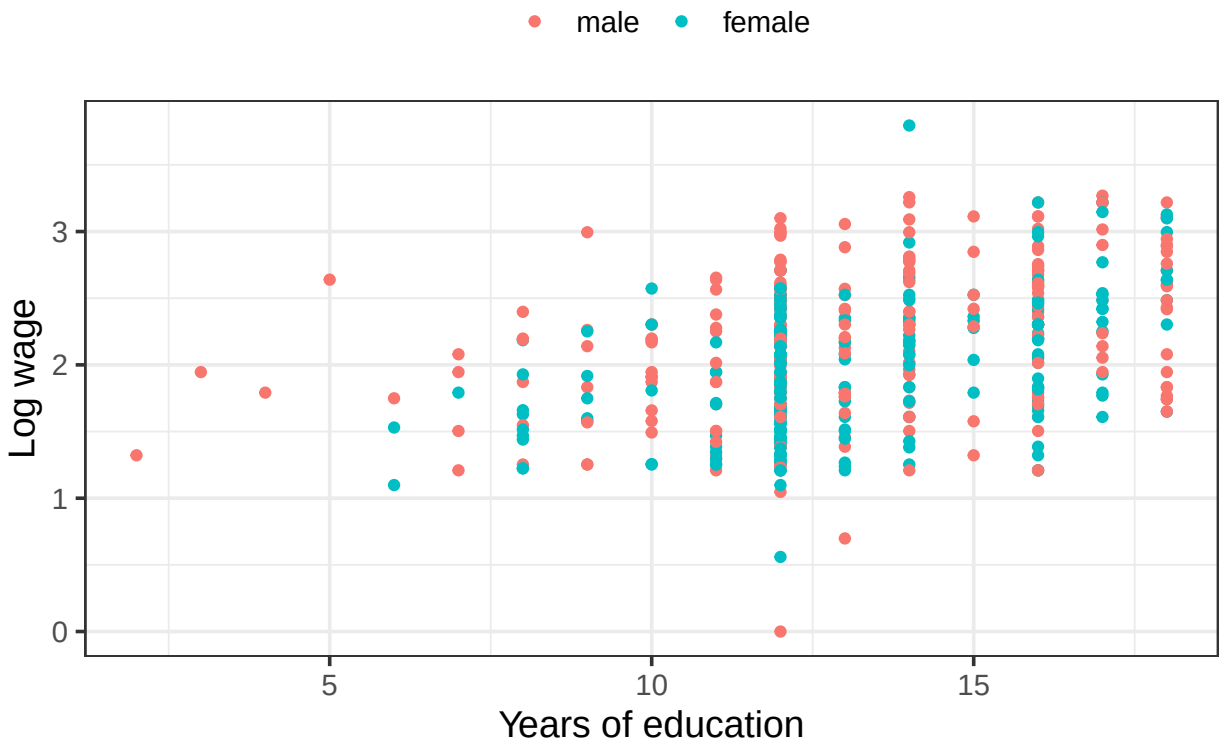
Each gender has its own slope: for men it is equal to the coefficient on `education` while for women it is equal to `education` + `gender*female`. An interaction term precisely allows for different slopes (if one of the two variables is continuous). The intercept is equal to the expected wage for men without any education while the coefficient on `gender` corresponds to the expected difference in incomes between men and women without any education.

- Can we reject the nullity of this coefficient at the 5% level? At 10%?

The interaction term coefficient's p-value is equal to 0.055 so it just fails to be significant at the 5% level but is significant at the 10% level.

1. Produce a scatterplot of the relationship between the log wage and the level of education, by gender.

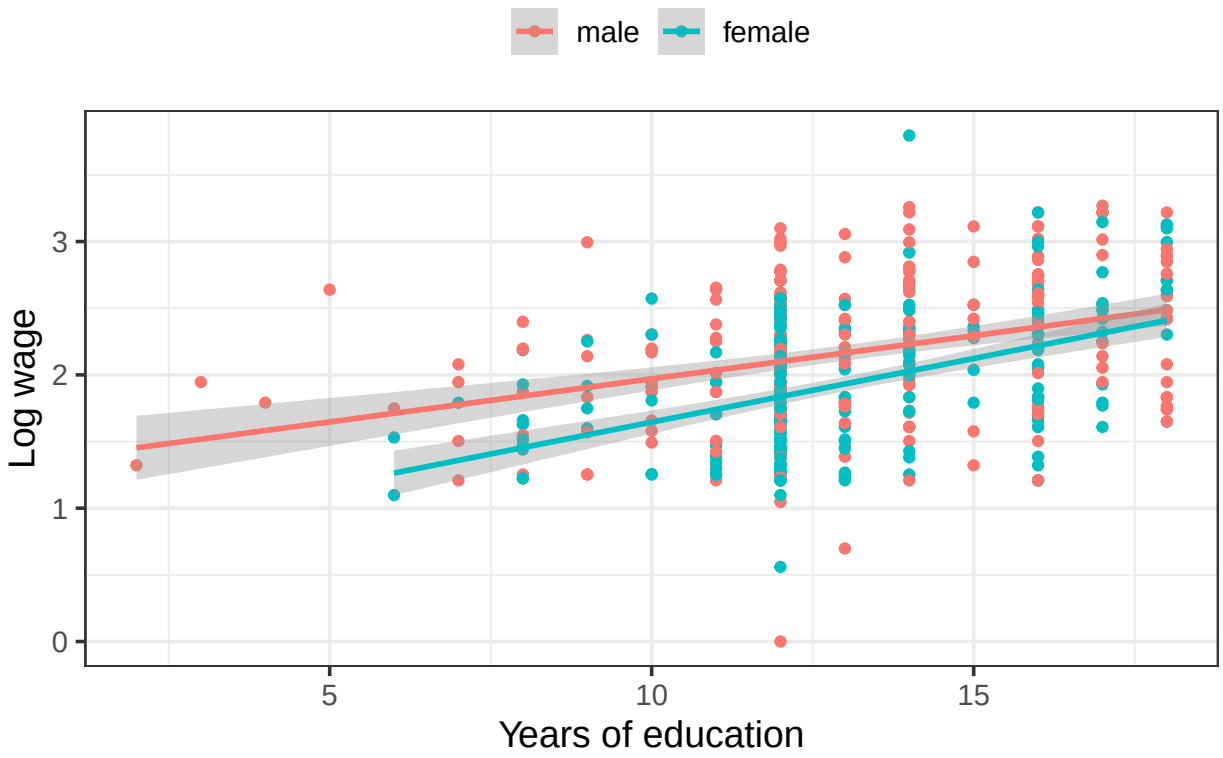
```
ggplot(cps, aes(x = education, y = log_wage, color = gender)) +
geom_point() +
labs(
  x = "Years of education",
  y = "Log wage",
  color = "",
  title = "Relationship between income and education, by gender") +
theme_bw(base_size = 14) +
theme(legend.position = "top")
```



2. Add the regression line with `geom_smooth`. What does this line represents?

```
ggplot(cps, aes(x = education, y = log_wage, color = gender)) +
geom_point() +
geom_smooth(method = "lm") +
labs(
  x = "Years of education",
  y = "Log wage",
  color = "",
  title = "Relationship between income and education, by gender") +
theme_bw(base_size = 14) +
theme(legend.position = "top")
```

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



The lines represent the fitted values from the regression of log wage on a gender dummy, education and their interaction (the one we ran in the previous task). The shaded area corresponds to the 95% confident interval for the fitted line itself (not the coefficients).

3. Let's illustrate what the shaded area stands for.

1. Draw one bootstrap sample from our `cps` data.

```
cps_boot <- cps %>%
  rep_sample_n(reps = 1, size = nrow(cps), replace = TRUE)
```

2. Regress the `log_wage` on `gender`, `education` and their interaction `gender*education`, save it as `reg_bootstrap`.

```
reg_bootsrap <- lm(log_wage ~ gender*education, cps_boot)
summary(reg_bootsrap)
```

```
##
## Call:
## lm(formula = log_wage ~ gender * education, data = cps_boot)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.14058 -0.35813  0.05393  0.30510  1.77331
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.435907   0.122797  11.693  < 2e-16 ***
## genderfemale            -0.769959   0.203259  -3.788 0.000169 ***
## education                0.058723   0.009132   6.431 2.84e-10 ***
```

```
## genderfemale:education  0.038151   0.015360   2.484 0.013311 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4539 on 530 degrees of freedom
## Multiple R-squared:  0.2291, Adjusted R-squared:  0.2247
## F-statistic: 52.49 on 3 and 530 DF,  p-value: < 2.2e-16
```

3. From `reg_bootstrap` extract and save the value of the intercept for men as `intercept_men_bootstrap` and the value of the slope for men as `slope_men_bootstrap`. Do the same for women.

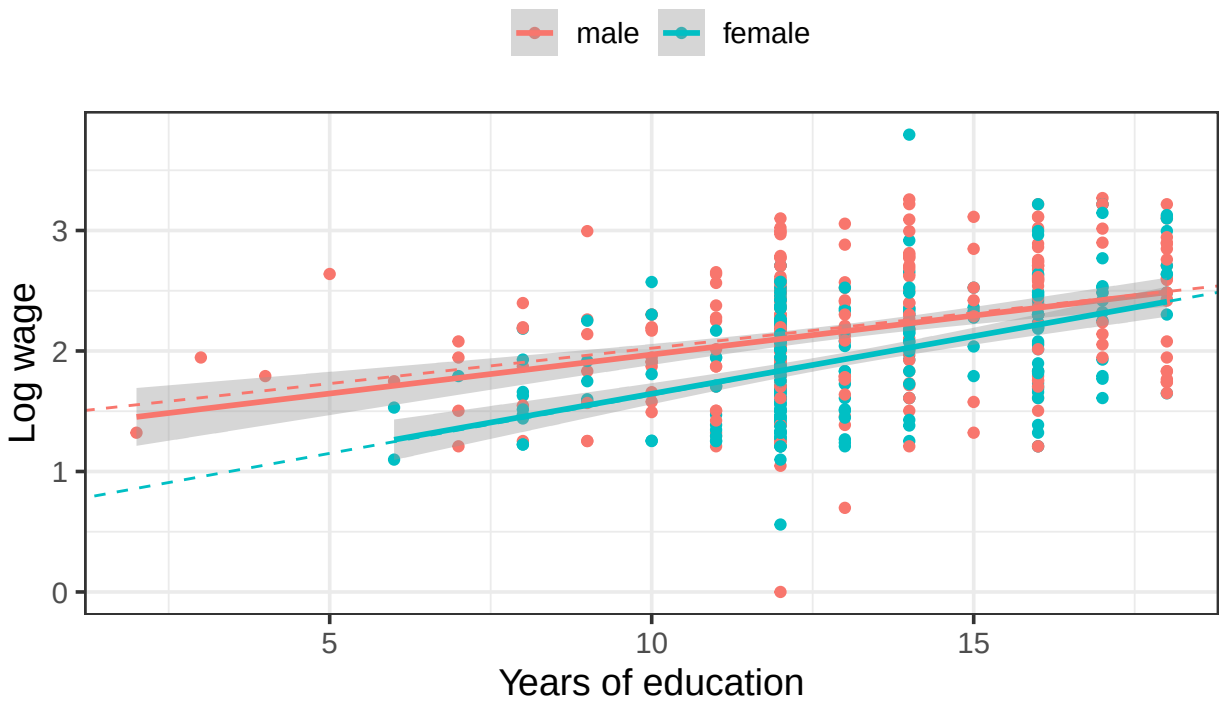
```
intercept_men_bootstrap = reg_bootsrap$coefficients[1]
slope_men_bootstrap = reg_bootsrap$coefficients[3]

intercept_women_bootstrap = reg_bootsrap$coefficients[1] + reg_bootsrap$coefficients[2]
slope_women_bootstrap = reg_bootsrap$coefficients[3] + reg_bootsrap$coefficients[4]
```

4. Add both predicted lines from this bootstrap sample to the previous plot (Hint: use `geom_abline` (x2))

```
  ggplot(cps, aes(x = education, y = log_wage, color = gender)) +
  geom_point() +
  geom_smooth(method = "lm") +
  geom_abline(slope = slope_men_bootstrap, intercept = intercept_men_bootstrap, linetype = "dashed", col
  geom_abline(slope = slope_women_bootstrap, intercept = intercept_women_bootstrap, linetype = "dashed"
  labs(
    x = "Years of education",
    y = "Log wage",
    color = "",
    title = "Relationship between income and education, by gender",
    subtitle = "Sample regression lines (solid), bootstrap sample regression lines (dashed)") +
  theme_bw(base_size = 14) +
  theme(legend.position = "top")
```

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



If we were the repeat this procedure 100 times, at each location of the shaded area, on average, 95% of the lines would lie within the shaded area and 5% would lie outside.