Computational Problem

Yuhuan Huang

2025-02-12

```
#getwd()
#setwd("./Documents/R-for-Econometrics")

library(haven)
library(dplyr)
library(broom)
library(ggplot2)
library(segmented)
data <- read_dta("lalonde2.dta")
```

Q1

```
covariates <- c("age", "educ", "black", "married", "nodegree", "hisp", "kids18", "kidmiss", "re74")

data <- data %>%
  mutate(group = case_when(
    treated == 1 & sample == 1 ~ "Treated_NSW",
    treated == 0 & sample == 1 ~ "Compar_NSW",
    sample == 2 ~ "Compar_CPS",
    TRUE ~ "Other"
  ))

table <- data %>%
  group_by(group) %>%
  filter(group != "Other") %>%
  summarise(across(all_of(covariates), list(
    mean = \(x) mean(x, na.rm = TRUE),
    sd = \(x) sd(x, na.rm = TRUE)
  )))
print(table)
```

```
## # A tibble: 3 x 19
##   group       age_mean age_sd educ_mean educ_sd black_mean black_sd married_mean
##   <chr>          <dbl>  <dbl>     <dbl>   <dbl>      <dbl>    <dbl>        <dbl>
## 1 Compar_CPS      33.2  11.0       12.0    2.87     0.0735    0.261        0.712
## 2 Compar_NSW      24.4   6.59      10.2    1.62     0.8       0.400        0.158
## 3 Treated_NSW     24.6   6.69      10.4    1.82     0.801     0.400        0.168
## # i 11 more variables: married_sd <dbl>, nodegree_mean <dbl>,
## #   nodegree_sd <dbl>, hisp_mean <dbl>, hisp_sd <dbl>, kids18_mean <dbl>,
```

```
## #   kids18_sd <dbl>, kidmiss_mean <dbl>, kidmiss_sd <dbl>, re74_mean <dbl>,
## #   re74_sd <dbl>
```

The result(table) is shown above. Compare treated NSW group with comparison NSW group, we can see that the metrics(mean and standard deviation) are very similar, while the treated group has a slightly higher mean in those variables. Compare treated NSW group with comparison CPS group, we can see that the groups have relatively large difference! The comparison CPS data has a larger average age and larger average education, etc.

Q2

```
data_filtered <- data %>%
  filter(group %in% c("Treated_NSW", "Compar_CPS"))

data_filtered$treated[is.na(data_filtered$treated)] <- 0 ## CPS data "treated" should be zero.

model_no_cov <- lm(re78 ~ 1 + treated, data = data_filtered)
model_with_cov <- lm(re78 ~ 1 + treated + age + educ + black + married + nodegree +
                       + hisp + kids18 + kidmiss + re74
                       ,data = data_filtered)

summary(model_no_cov)
```

```
##
## Call:
## lm(formula = re78 ~ 1 + treated, data = data_filtered)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -14847  -8917   1469  10718  54332
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14846.66      75.95  195.48   <2e-16 ***
## treated     -8870.31     562.48  -15.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9605 on 16287 degrees of freedom
## Multiple R-squared:  0.01504,    Adjusted R-squared:  0.01498
## F-statistic: 248.7 on 1 and 16287 DF,  p-value: < 2.2e-16
```

```
summary(model_with_cov)
```

```
##
## Call:
## lm(formula = re78 ~ 1 + treated + age + educ + black + married +
##     nodegree + +hisp + kids18 + kidmiss + re74, data = data_filtered)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -25702  -4239   1523   4243  57170
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.939e+03  4.777e+02  12.432  < 2e-16 ***
## treated     -1.569e+03  4.747e+02  -3.305 0.000952 ***
## age         -1.045e+02  6.164e+00 -16.953  < 2e-16 ***
## educ         2.144e+02  2.990e+01   7.172 7.72e-13 ***
## black       -9.742e+02  2.204e+02  -4.419 9.96e-06 ***
## married      4.735e+02  1.482e+02   3.195 0.001401 **
## nodegree     3.094e+02  1.852e+02   1.670 0.094918 .
## hisp        -1.593e+02  2.274e+02  -0.700 0.483637
## kids18       1.619e+00  3.524e+01   0.046 0.963354
## kidmiss      3.496e+02  7.626e+02   0.458 0.646670
## re74         6.742e-01  7.025e-03  95.977  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7318 on 16278 degrees of freedom
## Multiple R-squared:  0.4285, Adjusted R-squared:  0.4281
## F-statistic:  1220 on 10 and 16278 DF,  p-value: < 2.2e-16
```

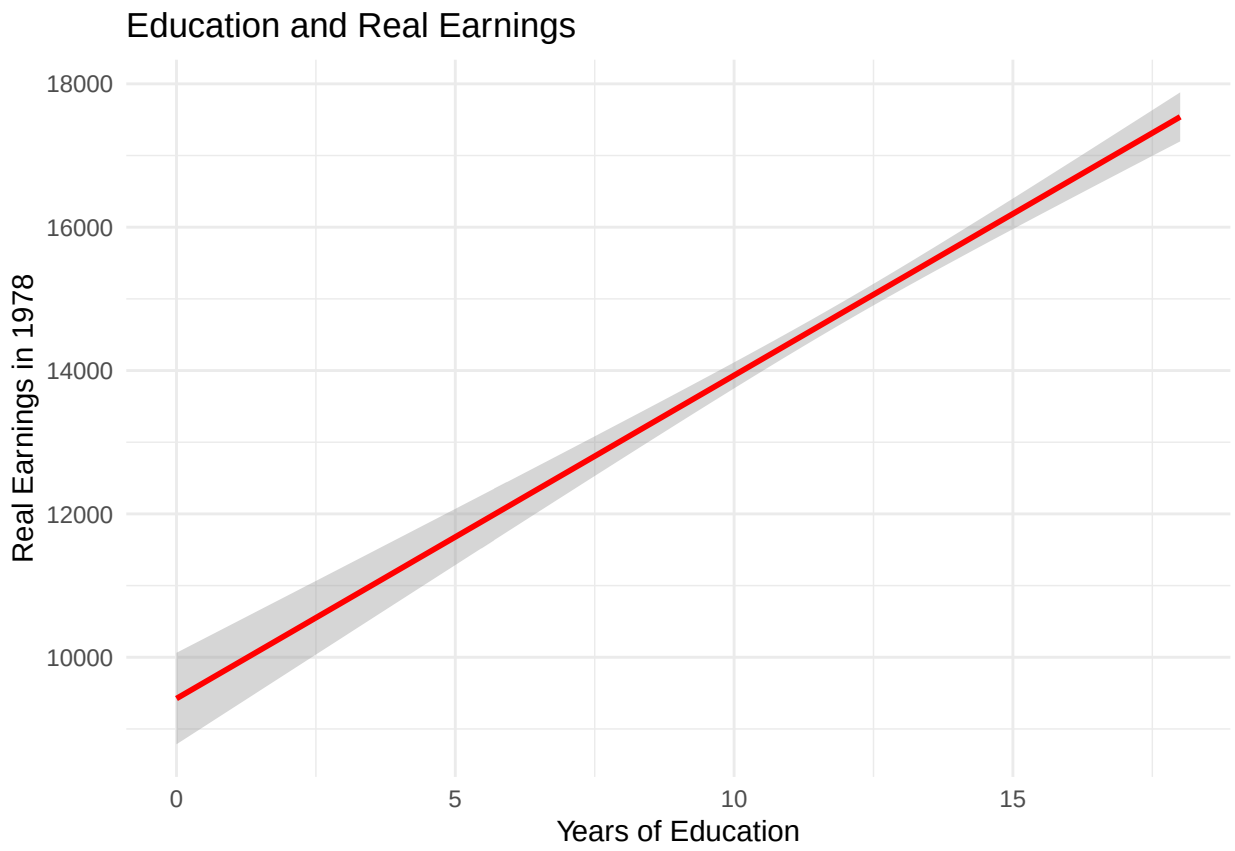
It is very different from PS1's result. When using the data made up of the treated group from NSW data and the comparison data from the CPS group, the coefficient of the variable "treated" is significantly different from what we get from the regression using both treated and non-treated NSW data. Also we can see that the difference between the with covariance model and without covariance model is different from what we get in PS1. A possible explanation of this is that, the ones from the CPS is not a good comparison group of the treated ones in NSW. This also reflects what we get from Q1: That the other dimensions of the treated group and its comparison (other variables such as age, educ, and black) are very different.

Q3

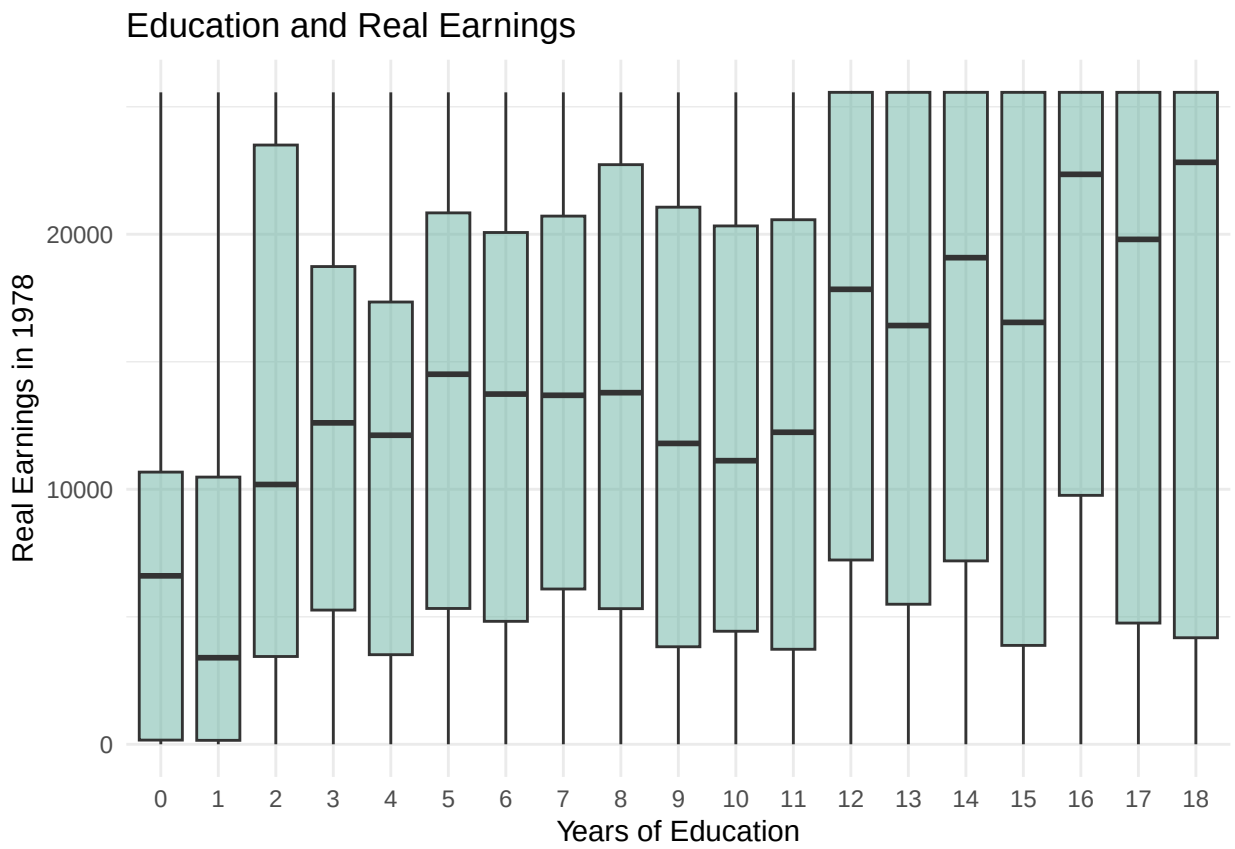
```
data_cps <- data %>%
  filter(group == "Compar_CPS")


ggplot(data_cps, aes(x = educ, y = re78)) +
  geom_smooth(method = "lm", color = "red", se = TRUE) +  # Regression line with confidence interval
  theme_minimal() +
  labs(title = "Education and Real Earnings",
       x = "Years of Education",
       y = "Real Earnings in 1978")
```

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




```
ggplot(data_cps, aes(x = factor(educ), y = re78)) +  #
  geom_boxplot(outlier.shape = NA, fill = "#69b3a2", alpha = 0.5) +
  theme_minimal() +
  labs(title = "Education and Real Earnings",
       x = "Years of Education",
       y = "Real Earnings in 1978")
```



From the regression plot and the box plot, we can see that, the real earnings in 1978 is positively correlated with years of education. And according to the box plot, there is a comparatively significant difference between education years of 12-18, 3-11, and 1-2.

Q4

```r
# Create piecewise terms
data_cps$edu_above12 <- pmax(0, data_cps$educ - 12)
data_cps$edu_above16 <- pmax(0, data_cps$educ - 16)

model <- lm(re78 ~ educ + edu_above12 + edu_above16, data = data_cps)
summary(model)
```

```
##
## Call:
## lm(formula = re78 ~ educ + edu_above12 + edu_above16, data = data_cps)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -16864  -8778   1763   9144  16759
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8805.41     463.81  18.985  < 2e-16 ***
```

```
## educ          509.97      42.50  11.999  < 2e-16 ***
## edu_above12   -25.31      84.66  -0.299  0.76500
## edu_above16  -706.39     254.84  -2.772  0.00558 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9557 on 15988 degrees of freedom
## Multiple R-squared:  0.01875,    Adjusted R-squared:  0.01857
## F-statistic: 101.9 on 3 and 15988 DF,  p-value: < 2.2e-16
```

```
# Plot the results
# Get model coefficients
coeffs <- coef(model)
intercept <- coeffs[1]
slope1 <- coeffs[2]  # Slope before 12 years
slope2 <- slope1 + coeffs[3]  # Slope between 12 and 16 years
slope3 <- slope2 + coeffs[4]  # Slope after 16 years

cat("slope1 =", slope1, ", slope2 =", slope2, " ,slope3 = ", slope3, "\n")
```

```
## slope1 = 509.9693 , slope2 = 484.6621  ,slope3 =  -221.7233
```

We set two dummies: education years above 12 and education years above 16. The result of the model means that: The slope of the first regression is the the coefficient of educ plus the coefficient of edu_above12, which is 509.97; the slope of the second regression is the coefficient of educ plus the coefficient of edu_above12 plus the coefficient of edu_above16, which is 484.66; the slope of the third regression is the coefficient of educ plus the coefficient of edu_above12 plus the coefficient of educ_above16, which is -221.72. It means that, when you have education less than 12 years, keep other things the same, on average 1 year's increase in education would bring about 509 dollars in salary; when you have years of education between 12 to 16, keep other things the same, on average 1 year's increase in education would bring about 484.66 salary increase; But if you already have 16 years and above (have a bachelor's degree already), keep other things the same, an increase in a year of education would decrease the average salary by about 221.72 dollars.

We also draw the graph below. Since the slope of the first and the second segment are close, it is not easy to see the two regression lines on the graph.

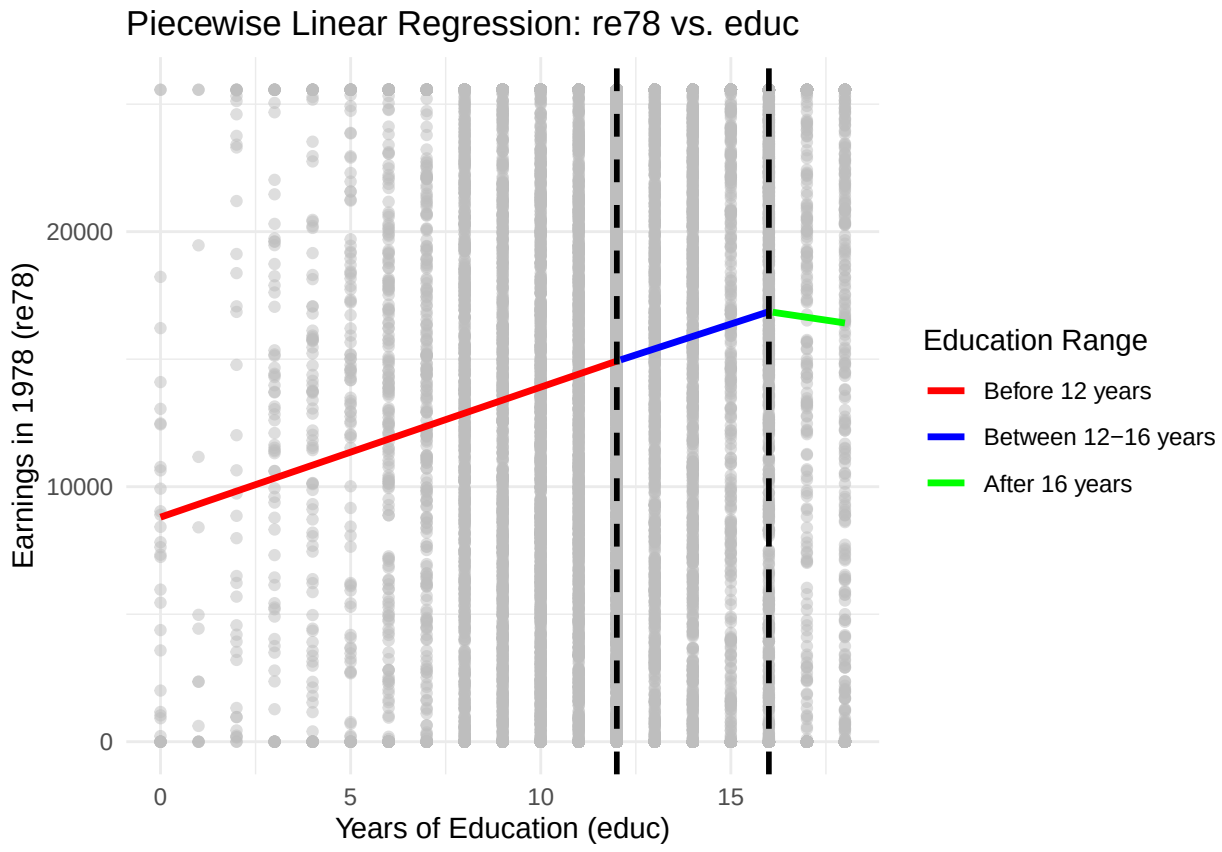
```
###### Plot
educ_seq <- seq(min(data_cps$educ), max(data_cps$educ), by = 0.1)

# Create a data frame for segmented predictions
pred_df <- data.frame(
  educ = educ_seq,
  segment = factor(
    case_when(
      educ_seq <= 12 ~ "Before 12 years",
      educ_seq > 12 & educ_seq <= 16 ~ "Between 12-16 years",
      educ_seq > 16 ~ "After 16 years"
    ),
    levels = c("Before 12 years", "Between 12-16 years", "After 16 years")
  ),
  re78 = case_when(
    educ_seq <= 12 ~ intercept + slope1 * educ_seq,
    educ_seq > 12 & educ_seq <= 16 ~ (intercept + slope1 * 12) + slope2 * (educ_seq - 12),
    educ_seq > 16 ~ (intercept + slope1 * 12 + slope2 * 4) + slope3 * (educ_seq - 16)
```

```
  )
)

# Plot the segmented piecewise regression with different colors
ggplot(data_cps, aes(x = educ, y = re78)) +
  geom_point(alpha = 0.5, color="grey") +  # Scatter plot of data points
  geom_line(data = pred_df, aes(x = educ, y = re78, color = segment), size = 1.2) +  # Color-coded piec
  geom_vline(xintercept = c(12, 16), linetype = "dashed", color = "black", size = 1) +  # Threshold mar
  scale_color_manual(values = c("Before 12 years" = "red",
                                "Between 12-16 years" = "blue",
                                "After 16 years" = "green")) +  # Custom segment colors
  labs(title = "Piecewise Linear Regression: re78 vs. educ",
       x = "Years of Education (educ)",
       y = "Earnings in 1978 (re78)",
       color = "Education Range") +
  theme_minimal()
```

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```



I also used package "segmented" to automatically select breakpoints. The breakpoints selected are 11 and 12. The automatic model indeed has a slightly higher R-square than the model using breakpoints 12 and 16.

```r
########## Use Package "Segmented": automatic choose breakpoints


base_model <- lm(re78 ~ educ, data = data_cps)
seg_model <- segmented(base_model, seg.Z = ~educ, psi = c(12, 16))
```

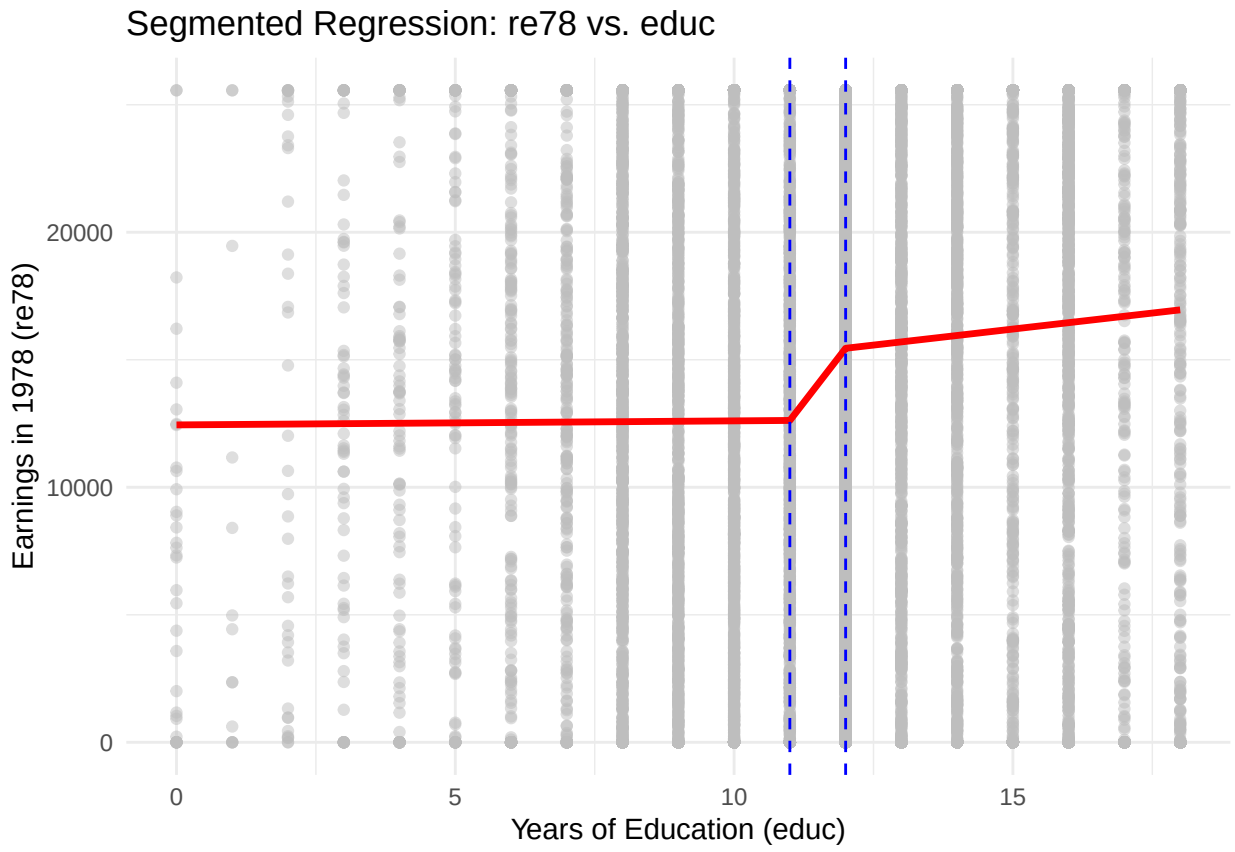
```
## Warning: Breakpoint estimate(s) outdistanced to allow finite estimates and
## st.errs
```

```r
summary(seg_model)
```

```
##
##  ***Regression Model with Segmented Relationship(s)***
##
## Call:
## segmented.lm(obj = base_model, seg.Z = ~educ, psi = c(12, 16))
##
## Estimated Break-Point(s):
##            Est. St.Err
## psi1.educ   11  0.141
## psi2.educ   12  0.121
##
## Coefficients of the linear terms:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12445.95     651.48  19.104   <2e-16 ***
## educ           15.81      77.35   0.204    0.838
## U1.educ      2814.00     325.94   8.633       NA
## U2.educ     -2578.77     327.07  -7.885       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9529 on 15986 degrees of freedom
## Multiple R-Squared: 0.0248,  Adjusted R-squared: 0.02449
##
## Boot restarting based on 10 samples. Last fit:
## Convergence *not* attained in 0 iterations (rel. change 10)
```

```r
data_cps$predicted <- predict(seg_model)
breakpoints <- seg_model$psi[, "Est."]

ggplot(data_cps, aes(x = educ, y = re78)) +
  geom_point(alpha = 0.5, color = "grey") +
  geom_line(aes(y = predicted), color = "red", size = 1.2) +
  geom_vline(xintercept = breakpoints, linetype = "dashed", color = "blue") +
  labs(title = "Segmented Regression: re78 vs. educ",
       x = "Years of Education (educ)",
       y = "Earnings in 1978 (re78)") +
  theme_minimal()
```


Segmented Regression: re78 vs. educ
Earnings in 1978 (re78)
20000
10000
0
0
5
10
15
Years of Education (educ)