Computational Problem

Yuhuan Huang

2025-01-21

We start with loading and preprocessing the data.

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

# Read .dta data
library(haven)
data <- read_dta("lalonde2.dta")
head(data)
summary(data)
```

```
# Data cleaning
processed_data <- na.omit(data) # dealing with missing values
processed_data <- processed_data[!duplicated(processed_data), ] # dealing with duplicate values
sapply(processed_data, is.numeric) # check whether the value of each variable is numerics
dim(processed_data)
summary(processed_data)
```

Q1

```
# Observe the treated dataset
treated_d <- subset(processed_data, treated == 1)
head(treated_d)
nontreated_d <- subset(processed_data, treated == 0)
head(nontreated_d)
```

```
##q1
avg_treated_RE_1978 <- mean(processed_data$re78[processed_data$treated == 1],na.rm = TRUE)
# na.rm = TRUE: ignore missing value
avg_nontreated_RE_1978 <- mean(processed_data$re78[processed_data$treated == 0],na.rm = TRUE)
print(avg_treated_RE_1978)
```

```
## [1] 5976.352
```

```
print(avg_nontreated_RE_1978)
```

```
## [1] 5090.048
```

```
difference <- avg_treated_RE_1978 - avg_nontreated_RE_1978
print(difference)
```

```
## [1] 886.3038
```

The sample average of Real Earnings 1978 for those who received treatment is 5976.35, whereas the sample average of Real Earnings 1978 for those who didn't receive the payment is 5090.05. We can see that those who received treatment earns higher, with an average amount of about 886.30.

Q2

```
##q2
model <- lm(re78 ~ 1 + treated, data = processed_data) #linear regression with interception
# the default regression contains intercept, so it is the same as:
# model <- lm(re78 ~ treated, data = processed_data)
summary(model)
```

```
##
## Call:
## lm(formula = re78 ~ 1 + treated, data = processed_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5976  -5090  -1519   3361  54332
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5090.0      302.8  16.811   <2e-16 ***
## treated        886.3      472.1   1.877   0.0609 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6242 on 720 degrees of freedom
## Multiple R-squared:  0.004872,   Adjusted R-squared:  0.003489
## F-statistic: 3.525 on 1 and 720 DF,  p-value: 0.06086
```

```
print(coefficients(model)["treated"])
```

```
##  treated
## 886.3038
```

```
temp <- coefficients(model)["treated"] - difference
print(temp)
```

```
##      treated
## 2.034994e-11
```

In the regression, the estimated value of β_1 is 5090.0, the estimated value of β_2, which is the coefficient of the variable *treated*, is 886.3. It is very close to the difference we get in question 1. The model also has a very low R-squared, which shows its poor explanatory effect.

It is consistent with the result in question 1. Because the variable *treated* works as an indicator (treated $=$ 1, nontreated $=$ 0) and it is also the only explanatory variable in the model. The meaning of its coefficient β_2 is that when we improve a unit of variable *treated*, on average the explained variable $re78$ would increase in β_2. For the indicator *treated*, it shows that those who have been treated ($treated = 1$) have on average about 886.3 dollars higher in their Real Earnings in 1978 then the nontreated, which is exactly what we get in question 1.

Q3

```
##q3

# predicted value and residuals:
pred_y <- predict(model)
head(pred_y, n=20)
```

```
##        1        2        3        4        5        6        7        8
## 5976.352 5976.352 5090.048 5976.352 5090.048 5090.048 5090.048 5090.048
##        9       10       11       12       13       14       15       16
## 5976.352 5090.048 5090.048 5090.048 5090.048 5090.048 5090.048 5090.048
##       17       18       19       20
## 5976.352 5976.352 5976.352 5976.352
```

```
resids <- residuals(model)
head(resids, n=20)
```

```
##          1          2          3          4          5          6          7
##  6441.7183  5680.1538 -4590.7910 10740.7691 25157.4518  -696.5253 11386.9713
##          8          9         10         11         12         13         14
## -5090.0482  9976.2476  7269.2614  1703.0289 -5090.0482 -4424.3108 11238.9157
##         15         16         17         18         19         20
##  1246.7345   358.7526  6827.6177 -4681.9430 -5876.7820 -5179.7923
```

The predicted values and residuals are as above. Then verify P1 to P5:

```
#check P1: y_bar = X_bar * b
y_bar <- mean(processed_data$re78)
X_mean <- c(1,mean(processed_data$treated)) # Contains the constant term
#print(X_mean)
b <- coefficients(model)
#print(b)
y_multiply <- sum(X_mean*b)
#print(y_bar)
#print(y_multiply)
dif <- y_bar - y_multiply
print(dif)
```

```
## [1] 6.366463e-12
```

Check P1: Regression passes through the mean of the data. We can see that the difference between $\bar{y}$ and $\bar{X}b$ is very small (6.366e-12). It verifies that $\bar{y} = \bar{X}b$.

```
#check P2: Sum of actual values equals sum of predicted values
sum_y <- sum(processed_data$re78)
sum_pred_y <- sum(pred_y)
sum_of_err <- sum_y - sum_pred_y
#print(sum_y)
#print(sum_pred_y)
print(sum_of_err)
```

```
## [1] 4.051253e-08
```

Check P2: Sum of actual values equals sum of predicted values. We can see that the difference between the sum of the real y and the sum of the predicted value $\hat{y}$ is very small (4.051e-08)

```
#check P3: Sum of residual is zero
sum_resids <- sum(resids)
print(sum_resids)
```

```
## [1] -1.237822e-09
```

Check P3: Residuals sum to zero. We can see that the sum of the residuals is indeed close to zero (-1.238e-09).

```
#check P4: Residuals are orthogonal to regressors
X <- model.matrix(model) # X matrix containing the intercept
#print(X)
rst <- t(X) %*% resids # compute X'e
print(rst)
```

```
##                      [,1]
## (Intercept) -1.237822e-09
## treated     -8.622010e-10
```

Check P4: Residuals are orthogonal to regressors. We compute $X'e$, and find that each element is close to zero (-1.237e-09 and -8.622e-10).

```
#check P5: Residuals are orthogonal to predicted values
rst2 <- t(pred_y) %*% resids # compute y_hat'e
print(rst2)
```

```
##               [,1]
## [1,] -1.860265e-06
```

Check P5: Residuals are orthogonal to regressors. We compute $X'e$, and find that each element is close to zero (-1.237e-09 and -8.622e-10).

Q4

```
model2 <- lm(re78 ~ 1 + treated + age + educ + black
             + married + nodegree + hisp + kids18
             + kidmiss + re74, data = processed_data)
#linear regression with interception
summary(model2)
```

```
##
## Call:
## lm(formula = re78 ~ 1 + treated + age + educ + black + married +
##     nodegree + hisp + kids18 + kidmiss + re74, data = processed_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -10466  -4334  -1495   3106  55107
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3543.7475  2611.0062   1.357 0.175137
## treated      796.5501   467.5145   1.704 0.088856 .
## age           16.0809    36.2045   0.444 0.657056
## educ         215.8654   179.5361   1.202 0.229628
## black      -1618.3498   799.0643  -2.025 0.043209 *
## married     -182.8522   683.0992  -0.268 0.789023
## nodegree    -346.2091   744.5919  -0.465 0.642099
## hisp         -24.9096  1047.0640  -0.024 0.981027
## kids18       512.0747   293.7879   1.743 0.081764 .
## kidmiss     -693.7150   726.4595  -0.955 0.339940
## re74           0.1334     0.0378   3.529 0.000444 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6139 on 711 degrees of freedom
## Multiple R-squared:  0.04961,    Adjusted R-squared:  0.03624
## F-statistic: 3.711 on 10 and 711 DF,  p-value: 7.239e-05
```

```
temp2 <- coefficients(model2)["treated"]
print(temp2)
```

```
##  treated
## 796.5501
```

In this multivariate model, the regression result shows that the estimated value of the coefficient of *treated* is 796.55. We can interpret it in this way: Keeping other explanatory variables (such as age, education, etc.) the same, on average, people who participate in a work experience program (treated) get about 796.55 higher in their Real Earnings in 1978 than those who didn't (nontreated).

The estimation value of β_2 we get here is different from (lower) than we get from the question2. Here are some possible explanation: Perhaps at first the impact of *treated* is over-estimated in the one variable model. As we add more reasonable variables to the regression function, other variables also contribute to the explanation of the explained variable. What's more, maybe there are some interaction effect between *treated* and other variables. We can also find an improvement in R-square in the new model.