Linear Regression: Estimation

2023/6/22

Learning points:

- read csv file using `read.csv()`
- run OLS regression using `lm()`
- create plots using `plot()`
- read dta file using `read.dta()` in library `foreign`
- subset data using `subset()`
- the option of `na.action = na.exclude` in the `lm()` function

Running an OLS Regression

```r
# load csv data (library foreign not required)
exp <- read.csv("Activity1/lalondeexp.csv")
head(exp)
```

```
##   age education black hispanic married nodegree re74 re75     re78 u74 u75
## 1  37        11     1        0       1        1    0    0  9930.05   1   1
## 2  22         9     0        1       0        1    0    0  3595.89   1   1
## 3  30        12     1        0       0        0    0    0 24909.50   1   1
## 4  27        11     1        0       0        1    0    0  7506.15   1   1
## 5  33         8     1        0       0        1    0    0   289.79   1   1
## 6  22         9     1        0       0        1    0    0  4056.49   1   1
##   treat id
## 1     1  1
## 2     1  2
## 3     1  3
## 4     1  4
## 5     1  5
## 6     1  6
```

```r
# run OLS regression
OLS <- lm(re78 ~ treat, data=exp)
summary(OLS)
```

```
##
## Call:
## lm(formula = re78 ~ treat, data = exp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6349  -4555  -1829   2917  53959
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4554.8      408.0  11.162  < 2e-16 ***
## treat         1794.3      632.9   2.835  0.00479 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6580 on 443 degrees of freedom
## Multiple R-squared:  0.01782,    Adjusted R-squared:  0.01561
## F-statistic: 8.039 on 1 and 443 DF,  p-value: 0.004788
```

Note that R automatically includes an intercept. If you want to run an OLS without the intercept, add `- 1` at the end of the equation, just like:

```
# run OLS regression without intercept
OLS_no_intercept <- lm(re78 ~ treat - 1, data =exp)
summary(OLS_no_intercept)
```

```
##
## Call:
## lm(formula = re78 ~ treat - 1, data = exp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6349   -138   1109   5844  53959
##
## Coefficients:
##       Estimate Std. Error t value Pr(>|t|)
## treat   6349.1      546.9   11.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7439 on 444 degrees of freedom
## Multiple R-squared:  0.2328, Adjusted R-squared:  0.2311
## F-statistic: 134.8 on 1 and 444 DF,  p-value: < 2.2e-16
```

```
# get coefficients
coefficients <- coef(OLS)
coefficients
```

```
## (Intercept)       treat
##    4554.802    1794.343
```

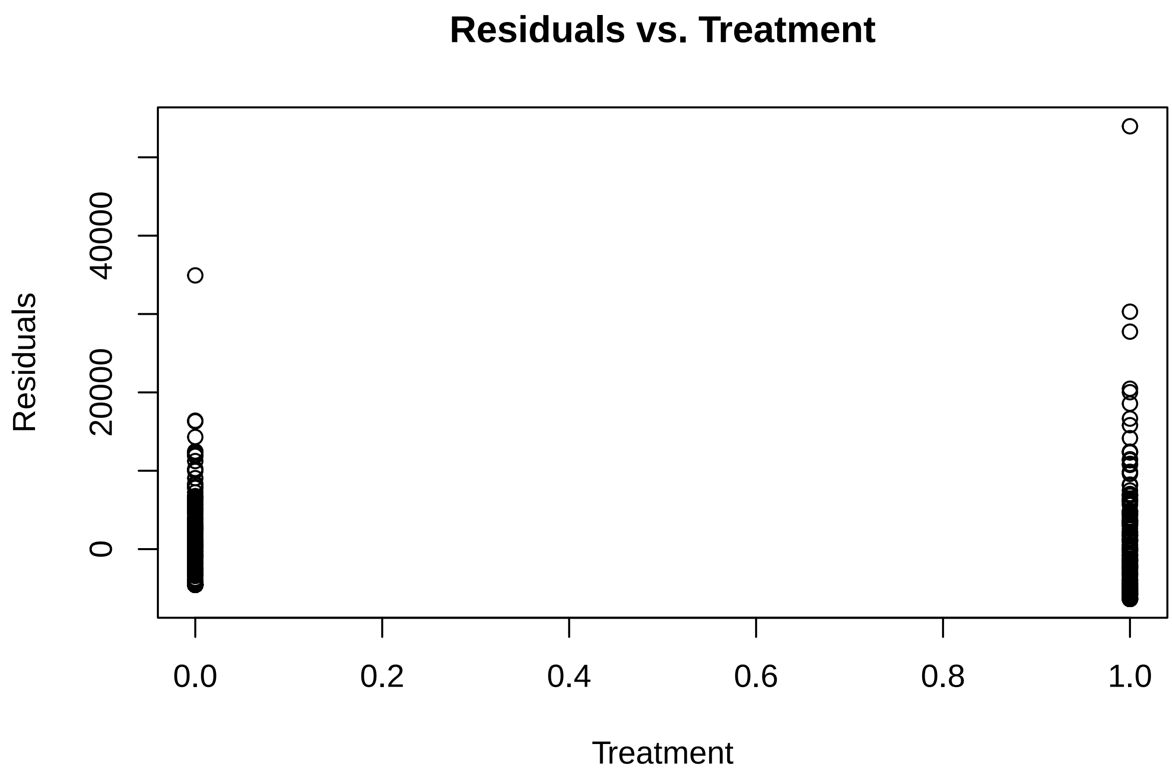
```
# get residuals
residuals <- residuals(OLS)
head(residuals)
```

```
##         1         2         3         4         5         6
##  3580.905 -2753.255 18560.355  1157.005 -6059.355 -2292.655
```

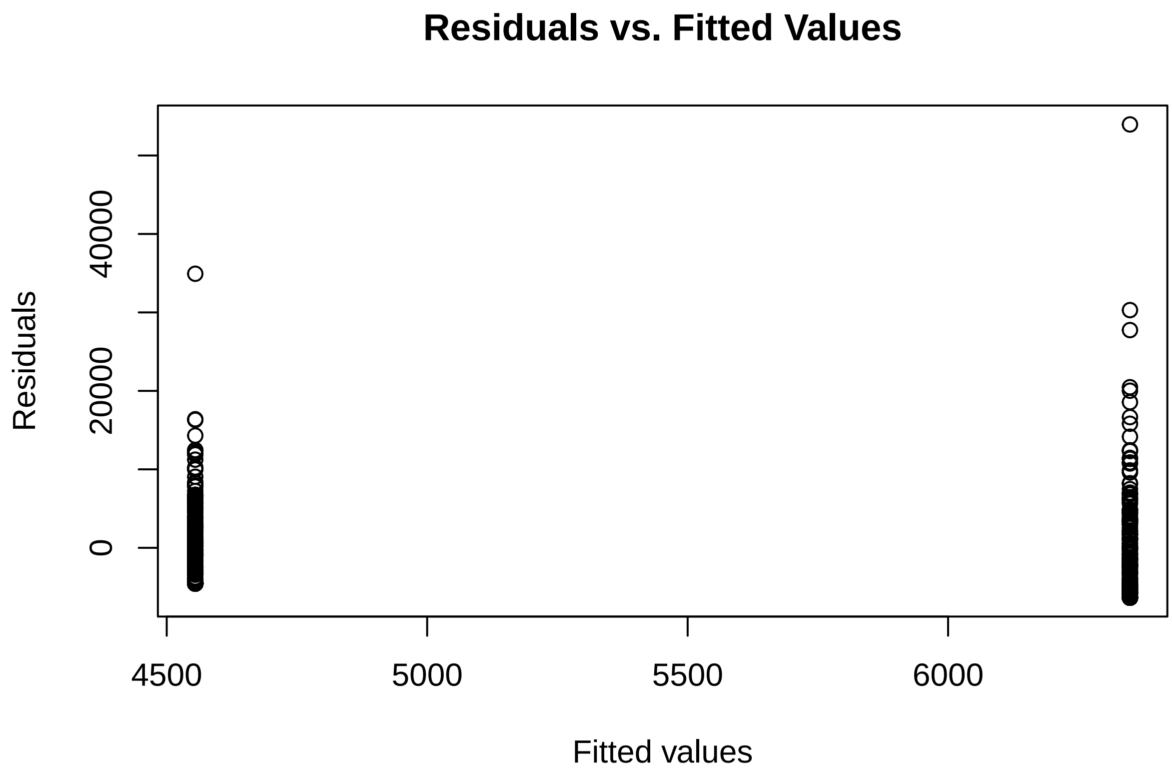
```r
# get fitted values
fitted_values <- fitted.values(OLS)
head(fitted_values)
```

```
##        1        2        3        4        5        6
## 6349.145 6349.145 6349.145 6349.145 6349.145 6349.145
```

```r
# plot the residuals against treatment variable
plot(exp$treat, residuals,
     xlab = "Treatment", ylab = "Residuals",
     main = "Residuals vs. Treatment")
```




```r
# plot the residuals against fitted values
plot(fitted_values, residuals,
     xlab = "Fitted values", ylab = "Residuals",
     main = "Residuals vs. Fitted Values")
```



Multiple Regression Model: Estimation

```
# Load the required package
library(foreign)
```

```
# Read the data
data <- read.dta("Activity2/CARD.dta")
```

The data has 34 variables and 3010 observations. We only care about a few of these, namely "wage", "educ", "IQ", "black", and "exper".

```
data_for_analysis <- subset(data, select = c(wage, educ, IQ, black, exper))
head(data_for_analysis)
```

```
##   wage educ  IQ black exper
## 1  548    7  NA     1    16
## 2  481   12  93     0     9
## 3  721   12 103     0    16
## 4  250   11  88     0    10
## 5  729   12 108     0    16
## 6  500   12  85     0     8
```

```r
# Create log wages
data_for_analysis$lwage <- log(data_for_analysis$wage)
```

Bivariate Regression

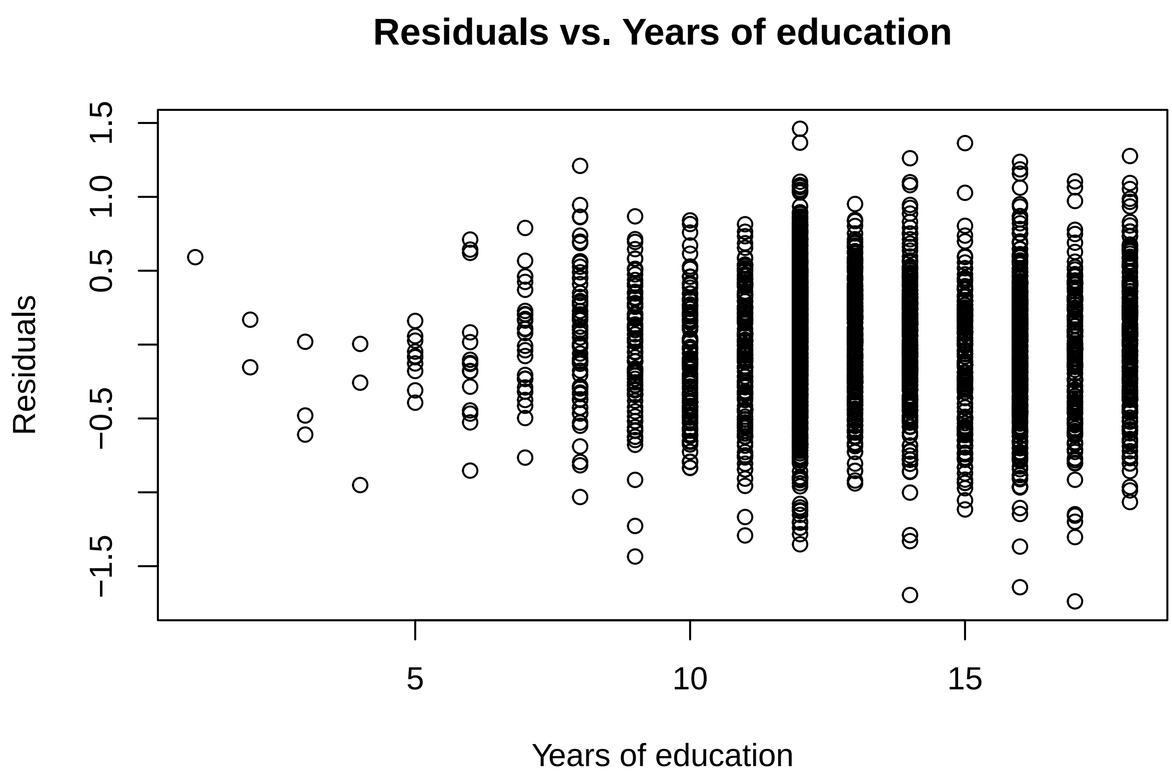
$$log(wage)_i = \beta_0 + \beta_1 educ_i + \epsilon_i$$

```r
# Run OLS regression
bivariate_OLS <- lm(lwage ~ educ, data=data_for_analysis)
summary(bivariate_OLS)
```

```
##
## Call:
## lm(formula = lwage ~ educ, data = data_for_analysis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.73799 -0.27764  0.02373  0.28839  1.46080
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.57088    0.03883  143.47   <2e-16 ***
## educ         0.05209    0.00287   18.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4214 on 3008 degrees of freedom
## Multiple R-squared:  0.09874,    Adjusted R-squared:  0.09844
## F-statistic: 329.5 on 1 and 3008 DF,  p-value: < 2.2e-16
```

The estimated coefficient on education is 0.052 implies that an additional year of schooling is associated with a 5.2 percent increase in wages.

```r
# plot residuals against education
residual <- residuals(bivariate_OLS)    # get the residuals
plot(data_for_analysis$educ, residual,
     xlab = "Years of education", ylab = "Residuals",
     main = "Residuals vs. Years of education")
```




```
# get R squared
summary(bivariate_OLS)$r.squared
```

```
## [1] 0.09873652
```

```
# get adjusted R squared
summary(bivariate_OLS)$adj.r.squared
```

```
## [1] 0.0984369
```

Multivariate Regressions

Adding one control

$$log(wage)_i = \gamma_0 + \gamma_1 educ_i + \gamma_2 exper_i + v_i$$

```
# run multiple OLS regression
multivar_OLS <- lm(lwage ~ educ + exper, data=data_for_analysis)
summary(multivar_OLS)
```

```
##
## Call:
## lm(formula = lwage ~ educ + exper, data = data_for_analysis)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.93442 -0.26396  0.02404  0.27287  1.42863
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.666034   0.063790   73.15   <2e-16 ***
## educ        0.093168   0.003612   25.80   <2e-16 ***
## exper       0.040657   0.002334   17.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4017 on 3007 degrees of freedom
## Multiple R-squared:  0.1813, Adjusted R-squared:  0.1808
## F-statistic:   333 on 2 and 3007 DF,  p-value: < 2.2e-16
```

Omitted variable bias formula

The fact that the coefficient on education changes when we include work experience as a control variable leads us to conclude that the estimate on education in the bivariate regression was (downward) biased. To see why this is, recall the **omitted variable bias formula**:

$$\hat{\beta}_1 = \hat{\gamma}_1 + \hat{\gamma}_2\hat{\pi}$$

where $\hat{\pi}$ is the OLS estimate from a regression of *exper* on *educ*. The bivariate and multivariate regressions we ran provide us with three of these coefficients, i.e.

$$.052 = .093 + .041\hat{\pi}$$

Solving for $\hat{\pi}$ we get $\hat{\pi} = -1$. This implies, that in a regression of *exper* on *edu*, the coefficient on *educ* should be -1. We can test this by running the following auxilliary regression

```r
# running auxilliary regression
aux_reg <- lm(exper ~ educ, data=data_for_analysis)
summary(aux_reg)
```

```
##
## Call:
## lm(formula = exper ~ educ, data = data_for_analysis)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.225 -3.081 -0.153  2.847  5.929
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 22.25545    0.28910   76.98   <2e-16 ***
## educ        -1.01024    0.02137  -47.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 3.137 on 3008 degrees of freedom
## Multiple R-squared:  0.4264, Adjusted R-squared:  0.4262
## F-statistic:  2236 on 1 and 3008 DF,  p-value: < 2.2e-16
```

Adding a second control: IQ

$$log(wage)_i = \gamma_0 + \gamma_1 educ_i + \gamma_2 IQ_i + \gamma_3 exper_i + v_i$$

```
# IQ has missing value
summary(data_for_analysis$IQ)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    50.0    93.0   103.0   102.4   113.0   149.0     949
```

We include the option of `na.action = na.exclude` in the `lm()` function to exclude the missing value. In fact, `R` automatically drops all missing observations in a regression. However, it is good practice to include option as it will force you to think about what observations you drop from your sample an possible selection issues.

```
# runing second multiple OLS regression
multivariate_OLS <- lm(lwage ~ educ + IQ + exper, data=data_for_analysis,
                       na.action=na.exclude)
summary(multivariate_OLS)
```

```
##
## Call:
## lm(formula = lwage ~ educ + IQ + exper, data = data_for_analysis,
##     na.action = na.exclude)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.59511 -0.23031  0.02295  0.25488  1.51582
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.5383780  0.0899992  50.427  < 2e-16 ***
## educ        0.0679058  0.0050183  13.532  < 2e-16 ***
## IQ          0.0045704  0.0006362   7.184 9.44e-13 ***
## exper       0.0453294  0.0028141  16.108  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3819 on 2057 degrees of freedom
##   (949 observations deleted due to missingness)
## Multiple R-squared:  0.1657, Adjusted R-squared:  0.1645
## F-statistic: 136.2 on 3 and 2057 DF,  p-value: < 2.2e-16
```

```
# get R squared
summary(multivariate_OLS)$r.squared
```

```
## [1] 0.1657152
```

```r
# get adjusted R squared
summary(multivariate_OLS)$adj.r.squared
```

```
## [1] 0.1644984
```

Frisch-Waugh theorem

The Frisch-Waugh theorem says that whether we run the multivariate regression as above leads to the same coefficient on education as if we (i) run an OLS regression of education on our controls and (ii) run an OLS regression of log wages on the residual obtained from the regression in (i). We already ran the multivariate regression above, so we can here implement the two step procedure to see if we get the same coefficient on education. To implement the procedure, run the following code:

```r
# first step OLs
step1 <- lm(educ ~ IQ + exper, data=data_for_analysis, na.action=na.exclude)
# get residuals for step 1
data_for_analysis$step1_residuals <- residuals(step1)
# second step OLS
step2 <- lm(lwage ~ step1_residuals, data=data_for_analysis, na.action=na.exclude)
summary(step2)
```

```
##
## Call:
## lm(formula = lwage ~ step1_residuals, data = data_for_analysis,
##     na.action = na.exclude)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.65300 -0.25188  0.03171  0.27415  1.26457
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6.335381   0.008857  715.28   <2e-16 ***
## step1_residuals 0.067906   0.005284   12.85   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4021 on 2059 degrees of freedom
##   (949 observations deleted due to missingness)
## Multiple R-squared:  0.07426,    Adjusted R-squared:  0.07381
## F-statistic: 165.2 on 1 and 2059 DF,  p-value: < 2.2e-16
```

Multicollinearity

We extend the model to include a dummy variable indicating whether an individual is black or not.

$$log(wage)_i = \gamma_0 + \gamma_1 educ_i + \gamma_2 IQ_i + \gamma_3 exper_i + \gamma_4 black_i + v_i$$

```r
# run OLS regression
multivariate_OLS2 <- lm(lwage ~ educ + IQ + exper + black, data=data_for_analysis,
                        na.action=na.exclude)
summary(multivariate_OLS2)
```

```
##
## Call:
## lm(formula = lwage ~ educ + IQ + exper + black, data = data_for_analysis,
##     na.action = na.exclude)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.60321 -0.22901  0.01811  0.25102  1.42742
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.7131687  0.0950585  49.582  < 2e-16 ***
## educ         0.0684151  0.0049851  13.724  < 2e-16 ***
## IQ           0.0030755  0.0006898   4.459 8.69e-06 ***
## exper        0.0443602  0.0028007  15.839  < 2e-16 ***
## black       -0.1424702  0.0263595  -5.405 7.24e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3793 on 2056 degrees of freedom
##   (949 observations deleted due to missingness)
## Multiple R-squared:  0.1774, Adjusted R-squared:  0.1758
## F-statistic: 110.9 on 4 and 2056 DF,  p-value: < 2.2e-16
```

Now create a new variable called `nblack`, which is "the opposit" of the variable `black`.

```
# create the variabel nblack
data_for_analysis$nblack = 1 - data_for_analysis$black
```

```
# run an extreme case of multicollinearity
multivariate_OLS3 <- lm(lwage ~ educ + IQ + exper + black + nblack, data=data_for_analysis,
                        na.action=na.exclude)
summary(multivariate_OLS3)
```

```
##
## Call:
## lm(formula = lwage ~ educ + IQ + exper + black + nblack, data = data_for_analysis,
##     na.action = na.exclude)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.60321 -0.22901  0.01811  0.25102  1.42742
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.7131687  0.0950585  49.582  < 2e-16 ***
## educ         0.0684151  0.0049851  13.724  < 2e-16 ***
## IQ           0.0030755  0.0006898   4.459 8.69e-06 ***
## exper        0.0443602  0.0028007  15.839  < 2e-16 ***
## black       -0.1424702  0.0263595  -5.405 7.24e-08 ***
## nblack              NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.3793 on 2056 degrees of freedom
##   (949 observations deleted due to missingness)
## Multiple R-squared:  0.1774, Adjusted R-squared:  0.1758
## F-statistic: 110.9 on 4 and 2056 DF,  p-value: < 2.2e-16
```

For perfect multicollinearity, R automatically omits the new regressor.

However, if we were to drop the intercept from the regression, R will estimate a coefficient for both variables.

```r
# run an extreme case of multicollinearity
multivariate_OLS3 <- lm(lwage ~ educ + IQ + exper + black + nblack - 1,
                        data=data_for_analysis, na.action=na.exclude)
summary(multivariate_OLS3)
```

```
##
## Call:
## lm(formula = lwage ~ educ + IQ + exper + black + nblack - 1,
##     data = data_for_analysis, na.action = na.exclude)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.60321 -0.22901  0.01811  0.25102  1.42742
##
## Coefficients:
##         Estimate Std. Error t value Pr(>|t|)
## educ   0.0684151  0.0049851  13.724  < 2e-16 ***
## IQ     0.0030755  0.0006898   4.459 8.69e-06 ***
## exper  0.0443602  0.0028007  15.839  < 2e-16 ***
## black  4.5706985  0.0895881  51.019  < 2e-16 ***
## nblack 4.7131687  0.0950585  49.582  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3793 on 2056 degrees of freedom
##   (949 observations deleted due to missingness)
## Multiple R-squared:  0.9964, Adjusted R-squared:  0.9964
## F-statistic: 1.151e+05 on 5 and 2056 DF,  p-value: < 2.2e-16
```