Regression Analysis: Dummy Variables

Statistics and Econometrics

Jiahua Wu

The example on page 23 - Interaction between a dummy and a quatitative variable

In this example, we seek to determine whether there exists any gender wage gap. As such, we construct a flexible model, where both the intercept and slope of *educ* (which measures returns to education) are allowed to differ between males and females.

```r
load("wage1.RData")
wage.m1 <- lm(log(wage) ~ female + educ + female:educ, data = data)
wage.m2 <- lm(log(wage) ~ female*educ, data)
stargazer(wage.m1, wage.m2, header = FALSE, type = 'latex',
          title = "A model to study gender differences in wage")
```

Table 1: A model to study gender differences in wage

	Dependent variable:	
	log(wage)	
	(1)	(2)
female	-0.360^{*}	-0.360^{*}
	(0.185)	(0.185)
educ	0.077^{***}	0.077^{***}
	(0.009)	(0.009)
female:educ	-0.0001	-0.0001
	(0.015)	(0.015)
Constant	0.826^{***}	0.826^{***}
	(0.118)	(0.118)
Observations	526	526
R^2	0.300	0.300
Adjusted R^2	0.296	0.296
Residual Std. Error (df = 522)	0.446	0.446
F Statistic (df = 3; 522)	74.649^{***}	74.649^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01	

$female : educ$ in $lm()$ indicates that the model includes an interaction term between $female$ dummy and $educ$ as an independent variable. Based on t tests, $female$ is only weakly significant, and the interaction term is insignificant at all. But we knew that variables could be individually insignificant, but jointly they might have an impact on the dependent variable. To test whether there is any gender gap in wages, we want to test the null hypothesis $H_0 : \delta_0 = 0, \delta_1 = 0$ (that is we want to test the null hypothesis: Expected wages are the same for men and women who have the same level of education).

```
linearHypothesis(wage.m1, c("female = 0", "female:educ = 0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## female = 0
## female:educ = 0
##
## Model 1: restricted model
## Model 2: log(wage) ~ female + educ + female:educ
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    524 120.77
## 2    522 103.80  2    16.971 42.673 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the F test, we observe that the two variables are jointly very significant. This is typically a consequence of multicollinearity, i.e., two variables are highly correlated. We verify this conjecture by investigating the correlation between the two, and the result indeed shows that the correlation is as high as 0.96.

```
data$female.educ <- data$female * data$educ
cor(data$female.educ, data$female)
```

```
## [1] 0.9635661
```

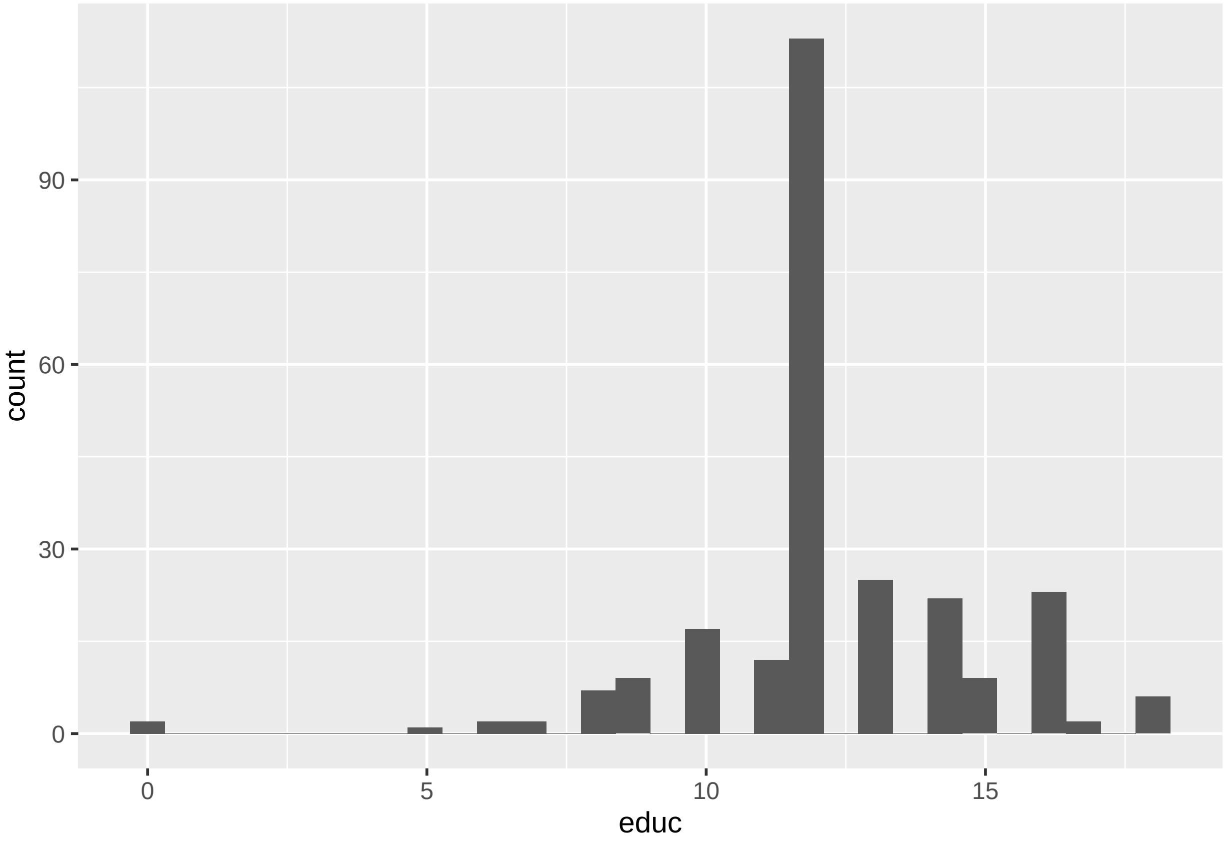
Remember that one variable is a dummy, and the other one is the interaction of this dummy with another quantitative variable. Correlation between the two depends on variations in females' educational levels in the sample. The less variation in females' educational levels, the higher the correlation between the two variables. Inspecting summary statistics and histogram, we observe that more than half of female individuals have 12-13 years of education. The distribution of females' educational level is highly clustered in the sample.

```
data.female <- data %>% filter(female == 1)

# distribution of females' education in the sample
summary(data.female$educ)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   12.00   12.00   12.32   13.00   18.00
```

```
ggplot(data = data.female, aes(x = educ)) + geom_histogram()
```


count
90
60
30
0
0
5
10
15
educ