Multiple Regression Analysis: Estimation

BS1802 Statistics and Econometrics

Jiahua Wu

Example 2.4

We first load the data set, and check out the data. Within all the data files I uploaded for this course, there will be a list called *desc*, which describes the variables within the data set.

```
load("wage1.RData")
ls()  # show data sets and functions you have defined
```

```
## [1] "data" "desc" "self"
```

```
desc
```

```
##    variable                          label
## 1      wage        average hourly earnings
## 2      educ             years of education
## 3     exper     years potential experience
## 4    tenure    years with current employer
## 5  nonwhite                 =1 if nonwhite
## 6    female                   =1 if female
## 7   married                  =1 if married
## 8    numdep           number of dependents
## 9      smsa             =1 if live in SMSA
## 10 northcen =1 if live in north central U.S
## 11    south   =1 if live in southern region
## 12     west    =1 if live in western region
## 13 construc  =1 if work in construc. indus.
## 14  ndurman  =1 if in nondur. manuf. indus.
## 15 trcommpu  =1 if in trans, commun, pub ut
## 16    trade    =1 if in wholesale or retail
## 17 services        =1 if in services indus.
## 18 profserv      =1 if in prof. serv. indus.
## 19  profocc    =1 if in profess. occupation
## 20  clerocc    =1 if in clerical occupation
## 21  servocc     =1 if in service occupation
## 22    lwage                      log(wage)
## 23  expersq                        exper^2
## 24  tenursq                       tenure^2
```

We rename the data frame with a more informative name :)

```
wage.data <- data
```

As the first step for any data analysis, we will need to get familiar with the data set by investigating summary statistics, univariate plot, and pairwise plot. For this example, we focus on the two variables of interest, *wage* and *educ*.

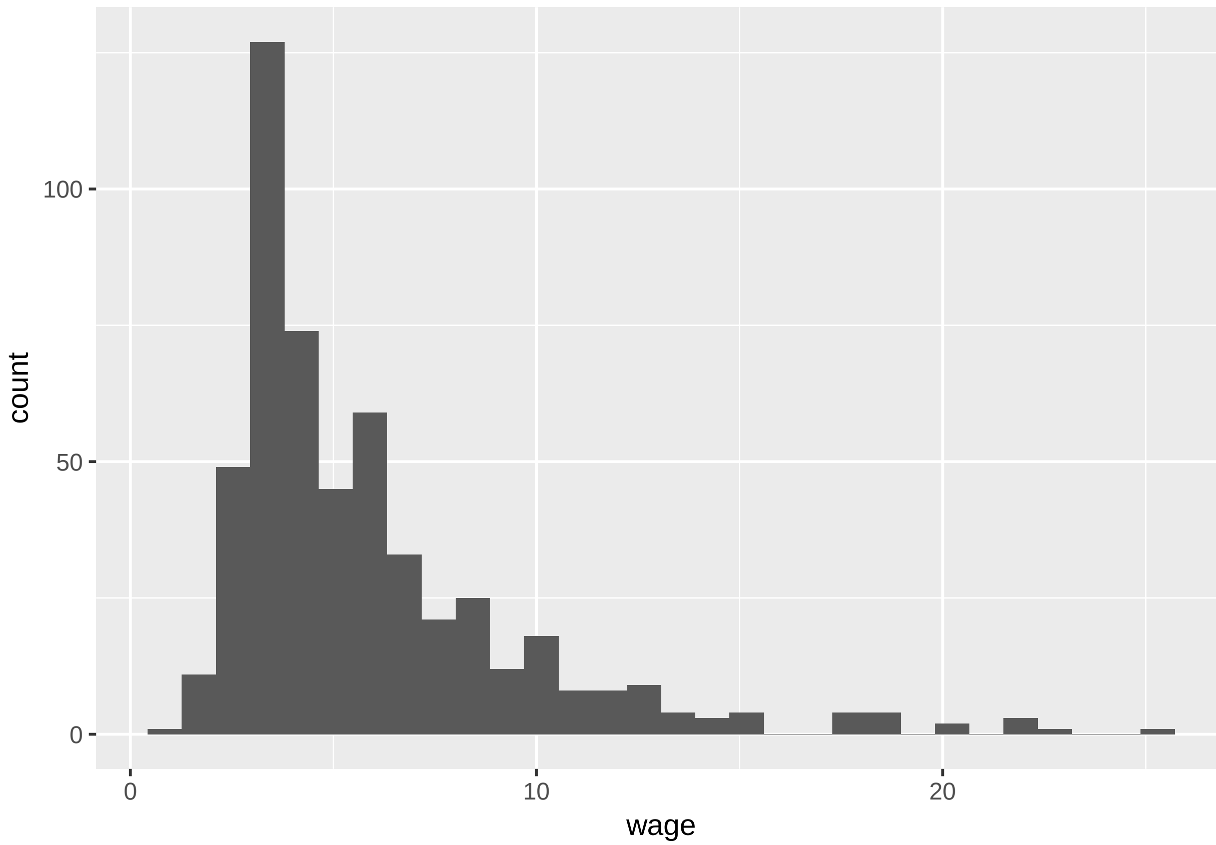
```
summary(wage.data[, 1:2])
```

```
##       wage             educ
##  Min.   : 0.530   Min.   : 0.00
```

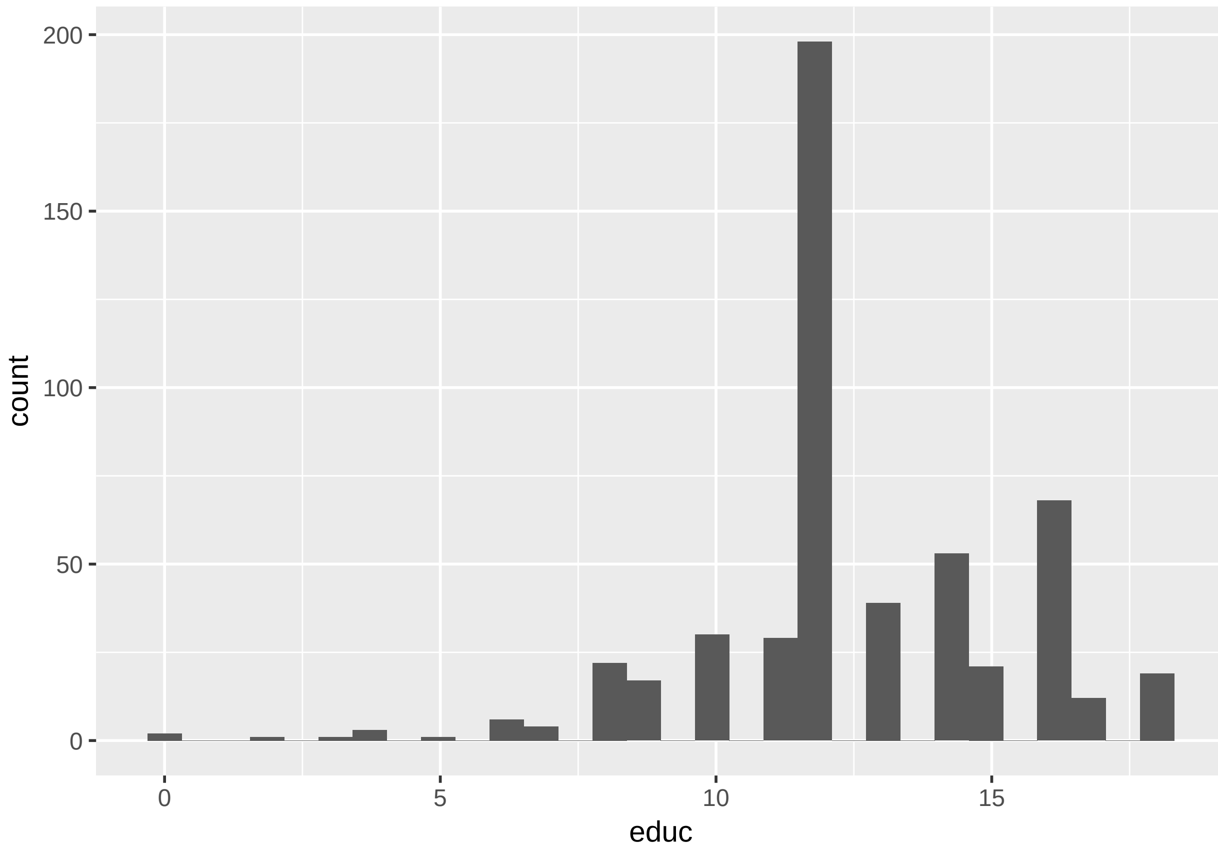
```
##   1st Qu.: 3.330   1st Qu.:12.00
##   Median : 4.650   Median :12.00
##   Mean   : 5.896   Mean   :12.56
##   3rd Qu.: 6.880   3rd Qu.:14.00
##   Max.   :24.980   Max.   :18.00
```

The summary statistics show that there is some discrepancy between median and mean of *wage*, which implies skewness in the distribution. Also from the summary statistics, we notice that *educ* is highly clustered, with more half of the samples are between 12 and 14. These findings are further confirmed with the histogram plots.

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

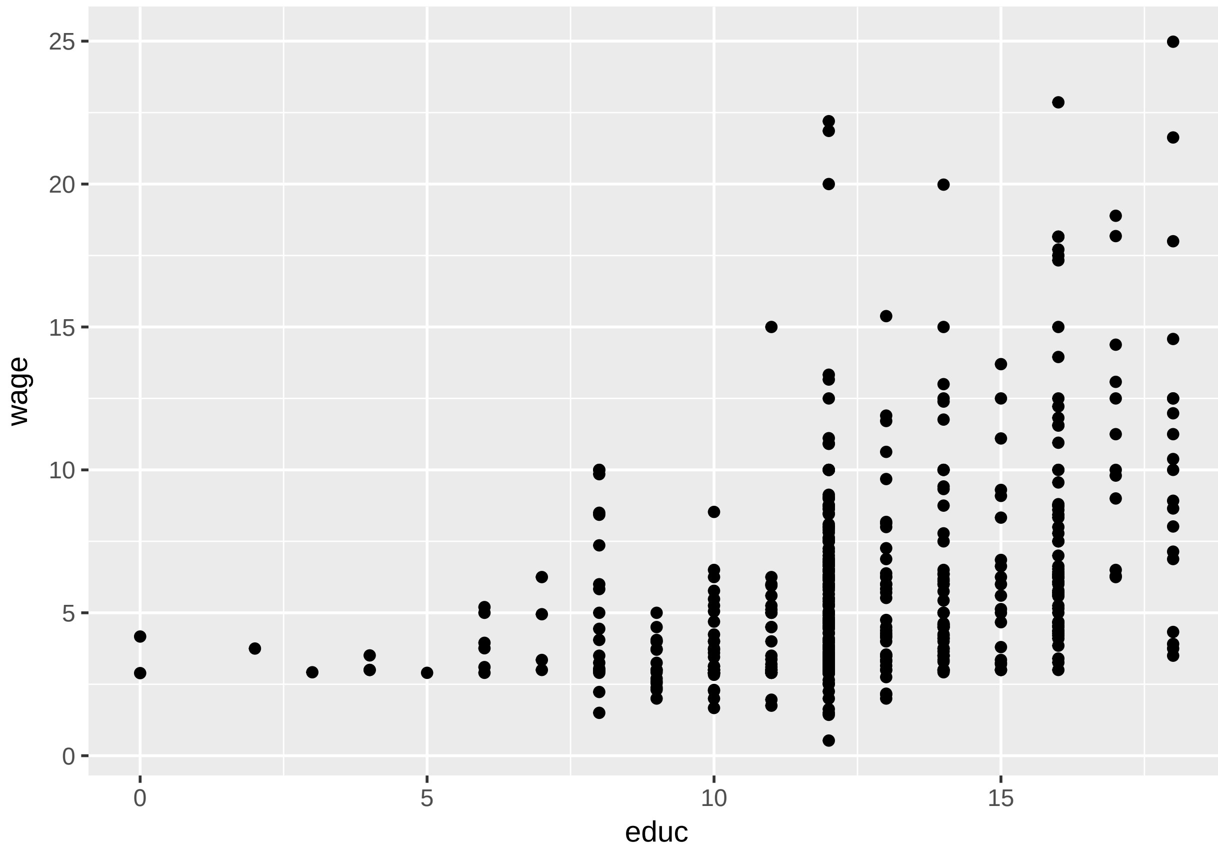



```
ggplot(data = wage.data, aes(x = educ)) + geom_histogram()
```



Next, we investigate the scatterplot of *wage* vs *educ*. It shows that *wage* generally increases in *educ*. Two things worth mentioning in this plot: (1) we have few observations with less than 5 years of education, which will impact the accuracy of prediction for lower levels of education; (2) the variation in *wage* generally increases in *educ*. This is a problem called heteroskedasticity in econometrics. We will discuss how to address it later in the course.

```
ggplot(data = wage.data, aes(x = educ, y = wage)) + geom_point()
```



Next we run regression and check out the output. The interpretation of OLS estimates is being dicussed on page 27 in the slide deck.

```
linear.m1 <- lm(wage ~ educ, data = wage.data)
summary(linear.m1)
```

```
##
## Call:
## lm(formula = wage ~ educ, data = wage.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3396 -2.1501 -0.9674  1.1921 16.6085
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.90485    0.68497  -1.321    0.187
## educ         0.54136    0.05325  10.167   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.378 on 524 degrees of freedom
## Multiple R-squared:  0.1648, Adjusted R-squared:  0.1632
## F-statistic: 103.4 on 1 and 524 DF,  p-value: < 2.2e-16
```

We can also calculate OLS estimates manually, using the matrix form of OLS estimates on page 15. *cbind*() is a R function that combines columns, and *rep*() is used to generate a vector of 1.

```r
# manual calculation of OLS estimates
X <- cbind(rep(1, nrow(wage.data)), wage.data$educ)
y <- wage.data$wage
OLS.est <- solve(t(X) %*% X, t(X) %*% y)
OLS.est
```

```
##            [,1]
## [1,] -0.9048516
## [2,]  0.5413593
```

We can also calculate R^2 manually using the formula on page 22.

```r
# manual calculation of R^2
y.hat <- linear.m1$fitted.values
R.sqrd <- sum((y.hat - mean(y.hat))^2) / sum((y - mean(y))^2)
R.sqrd
```

```
## [1] 0.1647575
```

```r
R.sqrd2 <- 1 - sum(linear.m1$residuals^2) / sum((y - mean(y))^2)
R.sqrd2
```

```
## [1] 0.1647575
```

Last, let us check out the fit of the OLS regression line by adding it to the scatterplot between *wage* and *educ*. It seems that the OLS regression line generally underestimates wage with low levels of education (< 5) and high levels of education (> 15). The plot suggests a nonlinear pattern between between *wage* and *educ*, which needs to be accounted for in the regression model.

```r
ggplot(data = wage.data, aes(x = educ, y = wage)) + geom_point() + stat_smooth(method = "lm")
```