Chapter 3

Multiple Regression Analysis - Estimation

Jim Bang

Multiple Regression

Wages and Education AND MORE!

- Using the wage1 data, regress wages on education (with an intercept).

Name this wage.lm1

- Using the wage1 data, regress wages on education and experience (with an intercept).
- Name this wage.lm2
- Summarize each regression object using "summary()"

```
wage.lm1 <- lm(wage ~ educ, data = wage1)
wage.lm2 <- lm(wage ~ educ + exper, data = wage1)
summary(wage.lm1)
```

```
##
## Call:
## lm(formula = wage ~ educ, data = wage1)
##
## 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
```

```
summary(wage.lm2)
```

```
##
## Call:
## lm(formula = wage ~ educ + exper, data = wage1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5532 -1.9801 -0.7071  1.2030 15.8370
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.39054    0.76657  -4.423 1.18e-05 ***
## educ         0.64427    0.05381  11.974  < 2e-16 ***
## exper        0.07010    0.01098   6.385 3.78e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.257 on 523 degrees of freedom
## Multiple R-squared:  0.2252, Adjusted R-squared:  0.2222
## F-statistic: 75.99 on 2 and 523 DF,  p-value: < 2.2e-16
```

Table Output for Regressions

- Display a *stargazer* object of each of the previous regressions using the *stargazer* function in the *stargazer* package.
- Display the output without assigning it to a named object
- Use a *text* output type. (In general, it's more useful to use *type* = 'html' and *out* = 'filename.html')

```
stargazer(wage.lm1, wage.lm2, type = 'text')
```

```
##
## ====================================================================
##                                  Dependent variable:
##                      -----------------------------------------------
##                                          wage
##                               (1)                       (2)
## --------------------------------------------------------------------
## educ                        0.541***                  0.644***
##                             (0.053)                   (0.054)
##
## exper                                                 0.070***
##                                                       (0.011)
##
## Constant                     -0.905                  -3.391***
##                             (0.685)                   (0.767)
##
## --------------------------------------------------------------------
## Observations                  526                       526
## R2                           0.165                     0.225
## Adjusted R2                  0.163                     0.222
## Residual Std. Error     3.378 (df = 524)          3.257 (df = 523)
## F Statistic         103.363*** (df = 1; 524) 75.990*** (df = 2; 523)
## ====================================================================
## Note:                                  *p<0.1; **p<0.05; ***p<0.01
```

Raw Stargazer HTML Output

```r
stargazer(wage.lm1, wage.lm2, type = 'html')
```

```
##
## <table style="text-align:center"><tr><td colspan="3" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:left"><
## <tr><td></td><td colspan="2" style="border-bottom: 1px solid black"></td></tr>
## <tr><td style="text-align:left"></td><td colspan="2">wage</td></tr>
## <tr><td style="text-align:left"></td><td>(1)</td><td>(2)</td></tr>
## <tr><td colspan="3" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:left">educ</td><td>0.541<sup>***</sup></
## <tr><td style="text-align:left"></td><td>(0.053)</td><td>(0.054)</td></tr>
## <tr><td style="text-align:left"></td><td></td><td></td></tr>
## <tr><td style="text-align:left">exper</td><td></td><td>0.070<sup>***</sup></td></tr>
## <tr><td style="text-align:left"></td><td></td><td>(0.011)</td></tr>
## <tr><td style="text-align:left"></td><td></td><td></td></tr>
## <tr><td style="text-align:left">Constant</td><td>-0.905</td><td>-3.391<sup>***</sup></td></tr>
## <tr><td style="text-align:left"></td><td>(0.685)</td><td>(0.767)</td></tr>
## <tr><td style="text-align:left"></td><td></td><td></td></tr>
## <tr><td colspan="3" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:left">Observations</td><td>526</td><td>5
## <tr><td style="text-align:left">R<sup>2</sup></td><td>0.165</td><td>0.225</td></tr>
## <tr><td style="text-align:left">Adjusted R<sup>2</sup></td><td>0.163</td><td>0.222</td></tr>
## <tr><td style="text-align:left">Residual Std. Error</td><td>3.378 (df = 524)</td><td>3.257 (df = 523)</td></tr>
## <tr><td style="text-align:left">F Statistic</td><td>103.363<sup>***</sup> (df = 1; 524)</td><td>75.990<sup>***</sup> (df = 2; 523)</td>
## <tr><td colspan="3" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:left"><em>Note:</em></td><td colspan="2"
## </table>
```

Formatted Stargazer HTML Output

You can directly insert a `stargazer` table in an `html` Rmarkdown document. One is to include the command that generates the `html` in an `R` chunk (with the option `results='asis'`) that itself is inside an `html` code chunk.

```
stargazer(
  wage.lm1,
  wage.lm2,
  title = "Education and Wages",
  type = "html",
  float = TRUE,
  no.space = TRUE,
  header = FALSE,
  covariate.labels = c("Education", "Experience")
)
```

Embedding Stargazer Output as HTML

Alternately, you might want to run your code from a `R` source file, save your pretty tables as html files, and include the content of those files in an `html` document.

One way to achieve this is to insert a code chunk with the `htmltools::includeHTML()` function:

```
includeHTML('output/wage.html')
```

Equivalently, you could do the same thing using the `xfun::file_string()` function:

```
file_string('output/wage.html')
```

Since this tutorial knits as an `html` document I needed to use some extra tricks. More commonly, you would knit your document as a `pdf`, and in this case you would want to leave out the `html` wrap, and set `type = "latex"`.

```
stargazer(
  wage.lm1,
  wage.lm2,
  title = "Education and Wages",
  type = "latex",
  float = TRUE,
  no.space = TRUE,
  header = FALSE,
  covariate.labels = c("Education", "Experience")
)
```

Table 1: Education and Wages

	Dependent variable:	
	wage	
	(1)	(2)
Education	0.541***	0.644***
	(0.053)	(0.054)
Experience		0.070***
		(0.011)
Constant	−0.905	−3.391***
	(0.685)	(0.767)
Observations	526	526
R^2	0.165	0.225
Adjusted R^2	0.163	0.222
Residual Std. Error	3.378 (df = 524)	3.257 (df = 523)
F Statistic	103.363*** (df = 1; 524)	75.990*** (df = 2; 523)
Note:		*p<0.1; **p<0.05; ***p<0.01

Implementing Regression

Assumptions of the Classical Regression Model

1. Linear in the parameters
2. Random Sampling: (X_i, Y_i) are independently and identically distributed
3. No Perfect Multicollinearity
4. $E[u|x] = 0$
5. Homoskedasticity: $u_i's$ have constant variance regardless of the value of x
6. Large Outliers are Unlikely: X and Y have finite fourth moments

$$E(X^4) < \infty$$

$$E(Y^4) < \infty$$

Deriving the OLS Estimator

One variable

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1})^2$$

Two variables

$$\min_{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2$$

k variables

$$\min_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik})^2$$

Taking the derivative with respect to each of the $\hat{\beta}$'s separately yields a system of equations equal to the number of $\hat{\beta}$'s $\{k+1)$.

Interpretating the Coefficients

Predicted values of y:

$$w\hat{a}ge = \hat{\beta}_0 + \hat{\beta}_1 educ + \hat{\beta}_2 exper$$

Changes in $\hat{y}$:

$$\Delta \hat{wage} = \hat{\beta}_0 + \hat{\beta}_1 \Delta educ + \hat{\beta}_2 \Delta exper$$

Ceteris paribus the predicted change in y in response to an observed change in x is:

$$\Delta \hat{wage} = 0 + \hat{\beta}_1 \Delta educ + \hat{\beta}_2 (0)$$

$$\Delta \hat{wage} = \hat{\beta}_1 \Delta educ$$

The predicted change in wage is $\hat{\beta}_1$ dollars per hour *for each additional year of education.*

Optimization by Hand Demo

The following code very closely replicates the method used in the *lm* function for minimizing the sum of squared residuals.

```
ssr <- function(b, y, x1, x2) {
  b1 <- b[1]
  b2 <- b[2]
  b0 <- b[3]
  sum((y - b0 - b1 * x1 - b2 * x2) ^ 2)
}
b.ols <-
  optim(
    par = c(mean(wage1$wage), 0, 0),
    fn = ssr,
    method = "BFGS",
    y = wage1$wage,
    x1 = wage1$educ,
    x2 = wage1$exper
  )
b.ols$par
```

```
## [1]  0.6442721  0.0700954 -3.3905395
```

```
sqrt(b.ols$value / (length(wage1$wage) - length(b.ols$par)))
```

```
## [1] 3.257044
```

Note: the value of the function at the minimum is the SSR, so $\hat{\sigma}^2 = b.ols\$value/(n-k)$

Partition Regression

Partialing Out Control Variables

Regress wages on experience and then education on experience.

Name the results *wage.p* and *educ.p*

```
wage.p <- lm(wage~exper, data = wage1)
educ.p <- lm(educ~exper, data = wage1)
```

Regressing the Residuals

Summarize the regression of the residuals of these regressions on each other and compare the effects of education to the summary of *wage.lm*2.

```
summary(lm(wage.p$resid~educ.p$residuals))
```

```
##
## Call:
## lm(formula = wage.p$resid ~ educ.p$residuals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5532 -1.9801 -0.7071  1.2030 15.8370
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.100e-16  1.419e-01    0.00        1
## educ.p$residuals 6.443e-01  5.375e-02   11.98   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.254 on 524 degrees of freedom
## Multiple R-squared:  0.2152, Adjusted R-squared:  0.2137
## F-statistic: 143.7 on 1 and 524 DF,  p-value: < 2.2e-16
```

```
summary(wage.lm2)
```

```
##
## Call:
## lm(formula = wage ~ educ + exper, data = wage1)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5532 -1.9801 -0.7071  1.2030 15.8370
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.39054    0.76657  -4.423 1.18e-05 ***
## educ         0.64427    0.05381  11.974  < 2e-16 ***
## exper        0.07010    0.01098   6.385 3.78e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.257 on 523 degrees of freedom
## Multiple R-squared:  0.2252, Adjusted R-squared:  0.2222
## F-statistic: 75.99 on 2 and 523 DF,  p-value: < 2.2e-16
```

Notice that the coefficients are identical! The standard error of regression (and hence the standard errors of the coeffients, t-statistics, and p-values) differ slightly. This is because R only takes into account the education variable when determining the degrees of freedom. The true standard error of the residual regression should take that into account and the SER should be scaled up by $(n-2)/(n-k)$, where $k=3$ in this example.

Properties of OLS Redux

- $\sum_{i=1}^{n} \hat{u}_i = 0$
- $\sum_{i=1}^{n} x_i \hat{u}_i = 0$
- The *Total* Sum of Squares, $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$
- The *Explained* Sum of Squares, $SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$
- The *Residual* Sum of Squares, $SSR = \sum_{i=1}^{n} (y_i - \hat{y})^2$
- SST = SSE + SSR (see section 2.3 for proof)
- Goodness of Fit, $R^2 = SSE/SST = 1 - SSR/SST$

We will discuss issues with using the (unadjusted) R^2 more in Chapter 6.