Linear Regression

Takaaki Kishida

February 12, 2020

Linear Regression

No Restriction on Standard Error

First we conduct regression without specifying the types of the standard error.

Set Up Data

```r
data(CASchools)

# student teacher ratio
CASchools$STR <- CASchools$students / CASchools$teachers

# average test score
CASchools$score <- (CASchools$read + CASchools$math)/2
```

This is same as: "reg score STR" in Stata. In R language we always need to specify which data will be used.

```r
fit1 <- lm(score ~ STR, data = CASchools)
summary(fit1)
```

```
##
## Call:
## lm(formula = score ~ STR, data = CASchools)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -47.73 -14.25   0.48  12.82  48.54
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)   698.93       9.47   73.82 < 0.0000000000000002 ***
## STR            -2.28       0.48   -4.75            0.0000028 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.6 on 418 degrees of freedom
## Multiple R-squared:  0.0512, Adjusted R-squared:  0.049
## F-statistic: 22.6 on 1 and 418 DF,  p-value: 0.00000278
```

Present the equations.

```r
equatiomatic::extract_eq(fit1)
```

$$\operatorname{score} = \alpha + \beta_{1}(\operatorname{STR}) + \epsilon$$

```r
equatiomatic::extract_eq(fit1, use_coefs = TRUE)
```

$$\operatorname{score} = 698.93 - 2.28(\operatorname{STR}) + \epsilon$$

We can include further controls in the equation.

```r
fit2 <- lm(score ~ STR + english + income, data = CASchools)
summary(fit2)
```

```
##
## Call:
## lm(formula = score ~ STR + english + income, data = CASchools)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -42.80  -6.86   0.27   6.59  31.20
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) 640.3155     5.7749  110.88 <0.0000000000000002 ***
## STR          -0.0688     0.2769   -0.25                 0.8
## english      -0.4883     0.0293  -16.67 <0.0000000000000002 ***
## income        1.4945     0.0748   19.97 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.3 on 416 degrees of freedom
## Multiple R-squared:  0.707,  Adjusted R-squared:  0.705
## F-statistic:  335 on 3 and 416 DF,  p-value: <0.0000000000000002
```

```r
equatiomatic::extract_eq(fit2)
```

$$\operatorname{score} = \alpha + \beta_{1}(\operatorname{STR}) + \beta_{2}(\operatorname{english}) + \beta_{3}(\operatorname{income}) + \epsilon$$

```r
equatiomatic::extract_eq(fit2, use_coefs = TRUE)
```

$$\operatorname{score} = 640.32 - 0.07(\operatorname{STR}) - 0.49(\operatorname{english}) + 1.49(\operatorname{income}) + \epsilon$$

Cluster Robust Standard Error

In empirical work we always deal with correlation within a group by clustering SE. Above lm code conduct regression under the assumption of homoskedasticity just like reg and without robust option in Stata. We now use the estimatr package.

```r
fit3 <- estimatr::lm_robust(score ~ STR + english + income,
                            clusters = county, se_type = "stata",
                            data = CASchools)
summary(fit3)
```

```
##
## Call:
## estimatr::lm_robust(formula = score ~ STR + english + income,
##     data = CASchools, clusters = county, se_type = "stata")
##
## Standard error type:  stata
##
```

```
## Coefficients:
##             Estimate Std. Error t value
## (Intercept) 640.3155     6.2866 101.855
## STR          -0.0688     0.2898  -0.237
## english      -0.4883     0.0309 -15.779
## income        1.4945     0.1023  14.604
##                                                             Pr(>|t|)
## (Intercept) 0.00000000000000000000000000000000000000000000000000000697
## STR         0.81352732429668439539938162852195091545581817626953125000
## english     0.00000000000000000009930987090849287295359433197705020740
## income      0.00000000000000000175194413389987415730081432545853544527
##             CI Lower CI Upper DF
## (Intercept)  627.646  652.985 44
## STR           -0.653    0.515 44
## english       -0.551   -0.426 44
## income         1.288    1.701 44
##
## Multiple R-squared:  0.707 , Adjusted R-squared:  0.705
## F-statistic:  301 on 3 and 44 DF,  p-value: <0.0000000000000002
```

We estimated the same model as fit2, but clustered SE at the county level (we do not know whether this is the best unit). Clustered SE for STR increased from 0.277 to 0.299.