

Linear Regression Example

Mariam Tanou

March 2, 2020

Linear Regression

No Restriction on Standard Error

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

Set Up Data

```
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.

```
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.

```
equationomatic::extract_eq(fit1)
```

$$\text{score} = \alpha + \beta_1(\text{STR}) + \epsilon$$

```
equationomatic::extract_eq(fit1, use_coefs = TRUE)
```

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

We can include further controls in the equation.

```
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

equatiomatic::extract_eq(fit2)
```

$$\text{score} = \alpha + \beta_1(\text{STR}) + \beta_2(\text{english}) + \beta_3(\text{income}) + \epsilon$$

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

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

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.

```
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
##
```

