

# Inferential Statistics on the Regression Coefficient

---

## 1. Example: Monthly Earnings and Years of Education

In this tutorial, we will focus on an example that explores the relationship between total monthly earnings and years of education. We will estimate the following regression equation:

$$y_i = b_0 + b_1x_i + e_i$$

where  $y_i$  denotes the *income* of individual  $i$  and  $x_i$  denotes the number of *years of education* of individual  $i$ .

The code below downloads a CSV file that includes data from 1980 for 935 individuals on variables including their total monthly earnings (`MonthlyEarnings`) and a number of variables that could influence income, including years of education (`YearsEdu`) and assigns it to a dataset that we call `wages`.

```
wages <- read.csv("http://murraylax.org/datasets/wage2.csv");
```

We estimate the simple regression with the following call to `lm()` and store the output in an object we call `lmwages`:

```
lmwages <- lm(wages$MonthlyEarnings ~ wages$YearsEdu)
```

We can print a summary of the results with the following call to the `summary()` function:

```
summary(lmwages)
```

```
##
## Call:
## lm(formula = wages$MonthlyEarnings ~ wages$YearsEdu)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -877.38 -268.63  -38.38  207.05 2148.26
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    146.952     77.715   1.891  0.0589 .
## wages$YearsEdu  60.214       5.695  10.573 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 382.3 on 933 degrees of freedom
## Multiple R-squared:  0.107, Adjusted R-squared:  0.106
## F-statistic: 111.8 on 1 and 933 DF, p-value: < 2.2e-16
```

These 'Estimate' column of the coefficients table implies the equation for the best fitting line is given by,

$$\hat{y}_i = 146.95 + 60.21x_i.$$

## 2. Hypothesis Testing on Coefficients (Two-tailed)

The regression coefficient on years of education,  $b_1 = 60.21$ , implies that in our sample each additional year of education is associated with \$60.21 higher monthly earnings. Suppose we wanted to test the hypothesis that having more years of education is associated with a change in monthly earnings. The null and alternative hypotheses would be as follows:

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \neq 0$$

The null hypothesis says that in the population the slope is equal to zero. This implies that the changing the explanatory variable, years of education, is associated with zero change on the outcome variable, monthly earnings.

The alternative hypothesis says that in the population the slope is different than zero. This implies that changing the explanatory variable, years of education, is associated with a change in the outcome variable, monthly earnings.

Results from this hypothesis test are reported in the summary above. The p-values *for a two-tailed test* are in the column labeled,  $\Pr(>|t|)$ . In our case, the p-value on `YearsEdu` is equal to  $2 \times 10^{-16}$ . Since this is far below a significance level equal to 5%, we reject the null hypothesis and conclude having more years of education is associated with a difference in monthly earnings.

## 2. Hypothesis Testing on Coefficients (One-tailed)

Suppose instead we have reason to believe that more education should result in higher average income, so we want to test the hypothesis that having *more education* is associated with *higher monthly earnings*. For this, we will conduct the following one-tailed hypothesis test on the coefficient:

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 > 0$$

Again the null hypothesis says that in the population the slope is equal to zero, therefore years of education is associated with zero change for monthly earnings.

The alternative hypothesis says that in the population the slope is greater than zero. This implies that higher years of education is associated with higher average monthly earnings.

We can use the same output from above. If the estimated coefficient is indeed positive as stated in our alternative hypothesis, we can use the two-tailed p-value from the table, but divide it by two, so as to only include the area in the right side tail.

Our p-value is therefore equal to  $1 \times 10^{-16}$ , which is far below the 5% significance level, so we reject the null hypothesis and conclude having more years of education is associated with higher average monthly earnings.

## 3. Confidence Interval Inference on Coefficients

Our sample evidence suggests that a single additional year of education is associated with an additional \$60.21 in monthly earnings. A 95% confidence interval can give us an interval estimate for our belief of the size of this impact, based on an estimate for the margin of error due to random sampling. We can compute a 95% confidence interval with the following call to `confint()`:

```
confint(lmwages, level=0.95)
```

```
##                2.5 %    97.5 %  
## (Intercept)  -5.56393 299.46881  
## wages$YearsEdu 49.03783 71.39074
```

From the result from the row labeled `wages$YearsEdu`, we can say with 95% confidence that one additional year of education is associated with higher average monthly earnings within the range of \$49.04 and \$71.39.