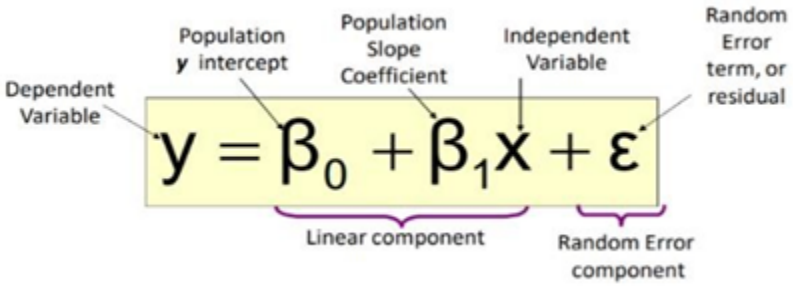
Regression Analysis

In this analysis, we are using **earning by bachelor degree** and **earning by advanced degree as our variables**, where we will test whether the value of **earning by advanced degree** depend on the value of **earning by bachelor degree**, using earning by bachelor degree as the independent variable(x) and earning by advanced degree as the dependent variable(y). Our regression model is a linear model, hence simple linear regression is used. The changes in the values of earning by advanced degree are assumed to be caused by the changes in the values of bachelor degree.

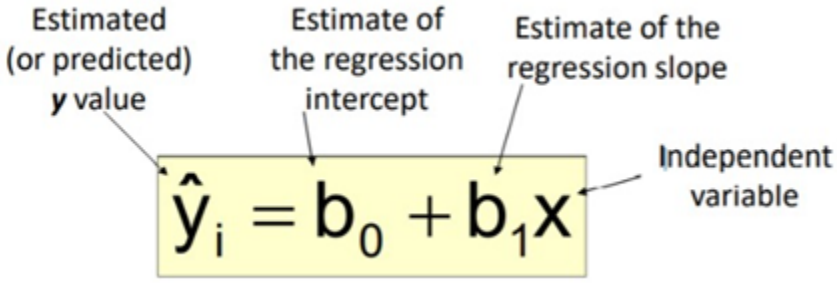
The mathematical equation for Population Linear Regression:



We assume that:

- Error values(Ɛ) are statistically independent, and normally distributed for any *x*
- The probability distribution of errors has constant variance
- The underlying relationship between variable *x* and variable *y* is linear

1. Estimated Regression Model:



From the equation above, b_0 is the estimated average value of *y* (earning by advanced degree) when the value of *x*(earning by bachelor degree) is zero. Whereas b_1 is the estimated change in the average value of *y* due to a one-unit change in *x*.

- Find least squares criterion:

From the above formula, we can find the values of b_0 and b_1 by:

$$b_0 = \bar{y} - b_1\bar{x}$$

$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

```
> mean(edu.wages$x)
[1] 40.55545
> mean(edu.wages$y)
[1] 53.08
> n<-11
> sum(edu.wages$x)
[1] 446.11
> sum(edu.wages$y)
[1] 583.88
> sum(edu.wages$x^2)
[1] 18130.32
> sum(edu.wages$x*edu.wages$y)
[1] 23723.29
>b1<-(sum(edu.wages$x*edu.wages$y)-(sum(edu.wages$x)*sum(edu.wages$y/n))/(sum(edu.wages$x^2)-((sum(edu.wages$x)^2)/
>b0<-mean(edu.wages$y)-(b1*mean(edu.wages$x)
```

By using RStudio, we get b_1 = 1.1479, b_0 = 6.5248

Substitute the values of b_0 and b_1 into the regression model equation:

$$\hat{y_i} = 6.5248 + 1.1479x$$

Find Coefficient of Determination, $\boldsymbol{R^2}$, by:

$$R^2 = \frac{SSR}{SST}$$

```
> R2 <- SSR/SST
```

By using RStudio, we get:

Coefficient of Determination, $\boldsymbol{R^2}$ = 0.9074

This shows that is 90% of the earning by advanced degree is explained by the value of earning by bachelor degree.

Find Standard Error of Estimate by:

$$s_\varepsilon = \sqrt{\frac{SSE}{n-k-1}}$$

```
> k <- 1
> Se <- sqrt(SSE/(n-k-1))
```

By using RStudio, we get Standard Error of Estimate, s_ε = 0.7546

Find Standard Deviation of Regression Slope by:

$$s_{b_1} = \frac{s_\varepsilon}{\sqrt{\sum(x-\bar{x})^2}} = \frac{s_\varepsilon}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

```
> Sb1 <- Se/(sqrt(sum((x-mean(x))^2)))
```

By using RStudio, we get Standard Deviation of Regression Slope, s_{b1}= 0.1222 Inference about the Slope: **t-Test**

Hypothesis Statement:

H_0: $\beta_1 = 0$ (no linear relationship)

H_1: $\beta_1 \neq 0$ (linear relationship does exist)

Find critical value, using α = 0.05, *df* = n-2 = 9

From *t*-table, since this is a two-tailed test, there are two critical values:

Lower tail critical value $-t_{\alpha/2=0.025,\ df=9}$ = -2.262,Upper tail critical value $t_{\alpha/2=0.025,\ df=9}$ = 2.262

From RStudio, we also get *p-value* = 6.008e-06 .Hence, we reject H_0 if test statistics > 2.262 / test statistics < -2.262

Calculate test statistic by:

$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$

```
> t <- (b1-0)/Sb1
```

By using RStudio, we get test statistic *t* = 9.394

State the decision

Since test statistics t = 9.394> upper tail critical value $t_{\alpha/2=0.025,\ df=9}$ =2.262,we **reject** the null hypothesis. There is sufficient evidence that wages for advanced degree holder affect by wages for bachelor degree.

∴ Linear Regression Model: $\hat{y_i} = 6.5248 + 1.1479x$

To perform linear regression in RStudio, we use the **lm()** function: we can see the values of intersection coefficient (Intercept) and slope coefficient (x).Finally, we can plot a scatter plot using the **plot()** function, and add the linear regression model into the plot using **abline()** function:

```
> wages.regression<-lm(y~x, data=edu.wages)
> summary(wages.regression)

Call:
lm(formula = y ~ x, data = edu.wages)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2629 -0.4895  0.3272  0.5115  0.9835

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.5248     4.9613   1.315    0.221
x             1.1479     0.1222   9.394 6.01e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7546 on 9 degrees of freedom
Multiple R-squared:  0.9074,    Adjusted R-squared:  0.8972
F-statistic: 88.24 on 1 and 9 DF,  p-value: 6.008e-06

> abline(wages.regression, col="red")
```

When we view the summary of our linear regression model, we can get the values of intersection coefficient b_0 = 6.5248 slope coefficient b_1 =1.1479, Standard Deviation of Regression Slope, s_{b_1} = 0.1222 , Standard Error of Estimate, s_ε =0.7546 , df = 9 , Coefficient of Determination, R^2 = 0.9074 and ***p-value*** = 6.008e-06