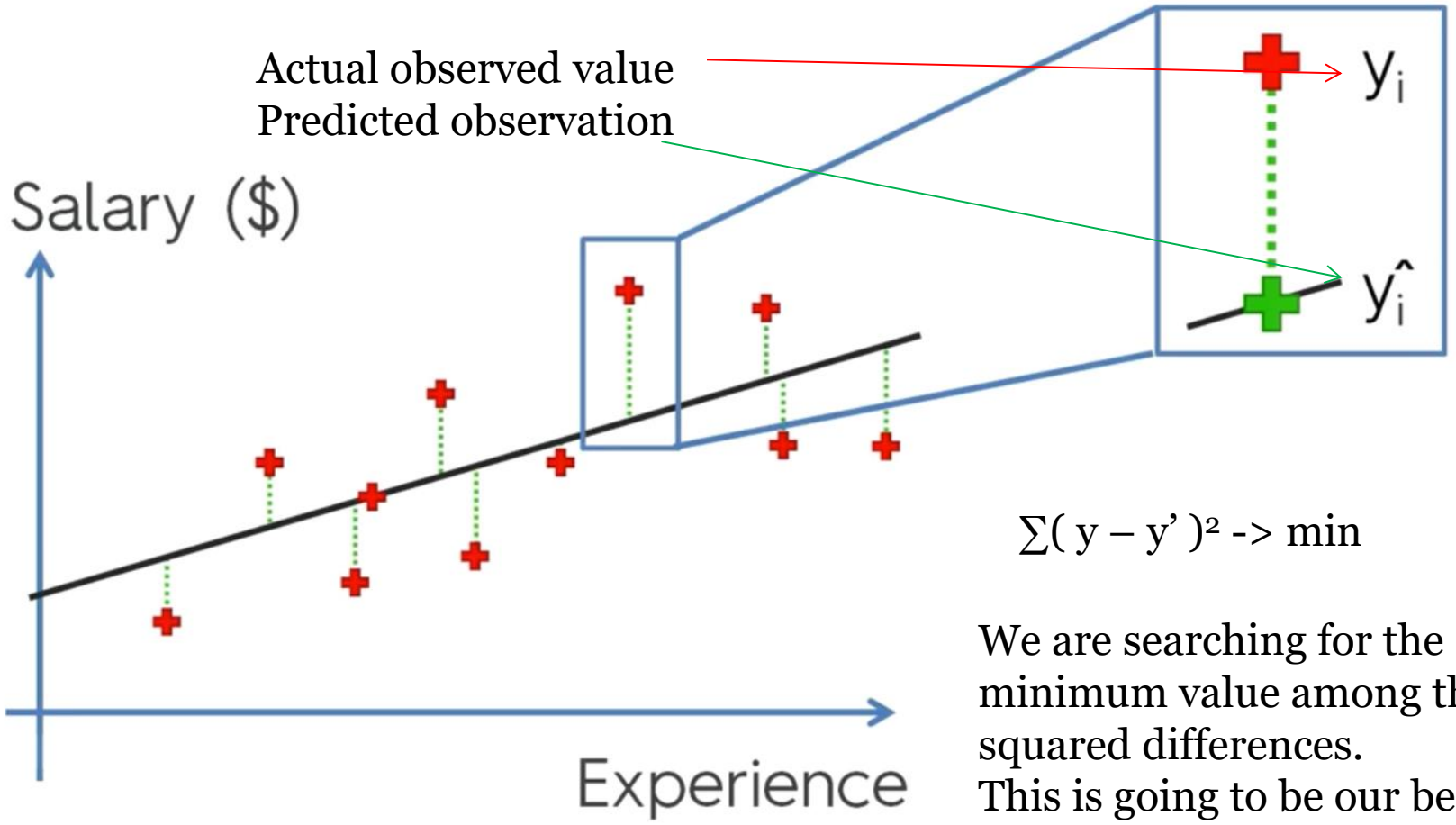
Introduction to Machine Learning. Lec.8 Evaluating Regression Models Performance

Aidos Sarsembayev, IITU, 2018

MEAN SQUARED ERROR

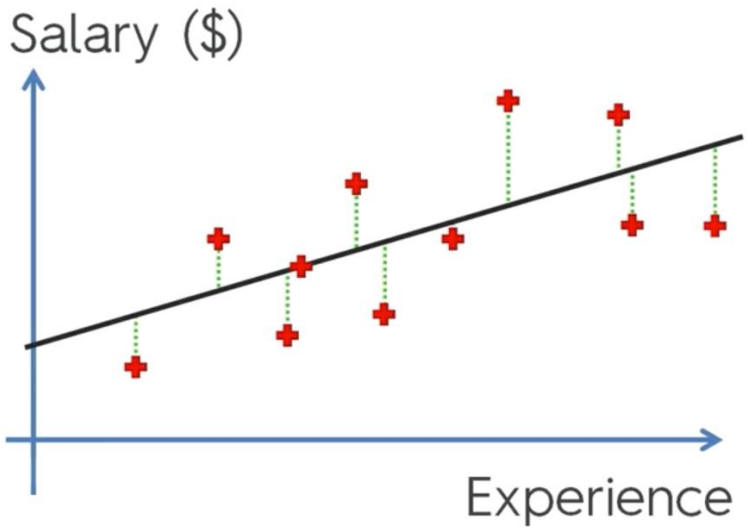
SLR



$$\sum(y - y')^2 \rightarrow \min$$

We are searching for the minimum value among these squared differences.
This is going to be our best fitting line

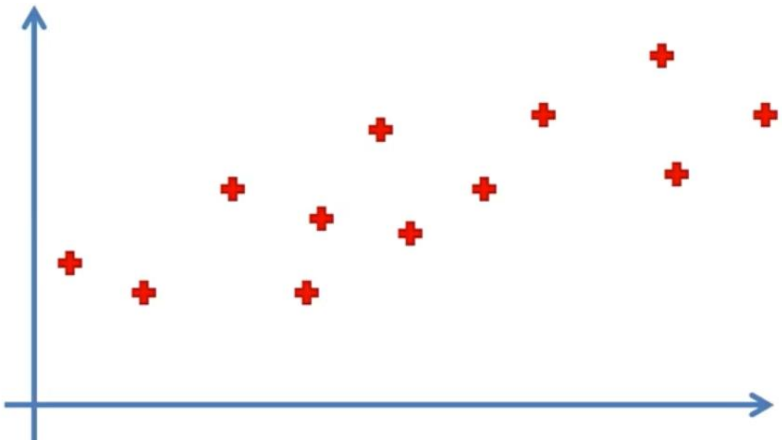
MSE



This is the sum of squares of residuals

$SS_{res} = \sum(y - y')^2$

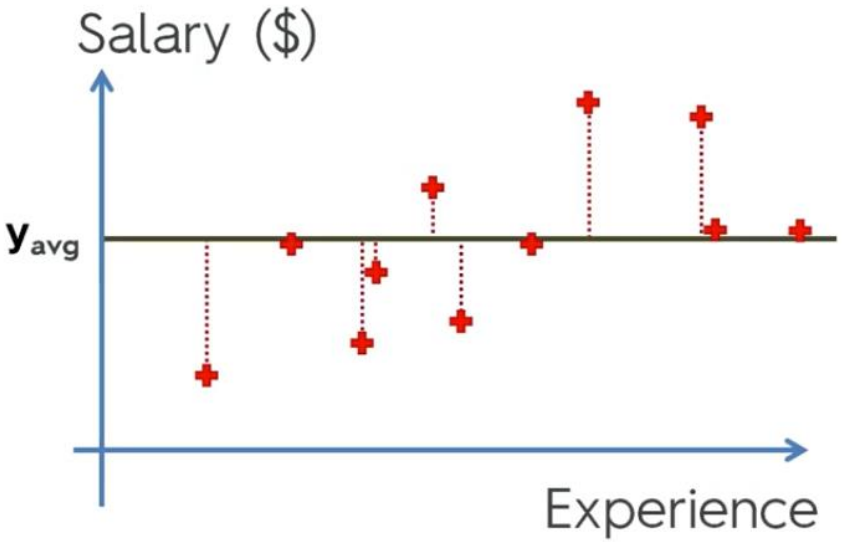
MSE



This is the sum of squares of residuals

$SS_{res} = \sum(y - y')^2$

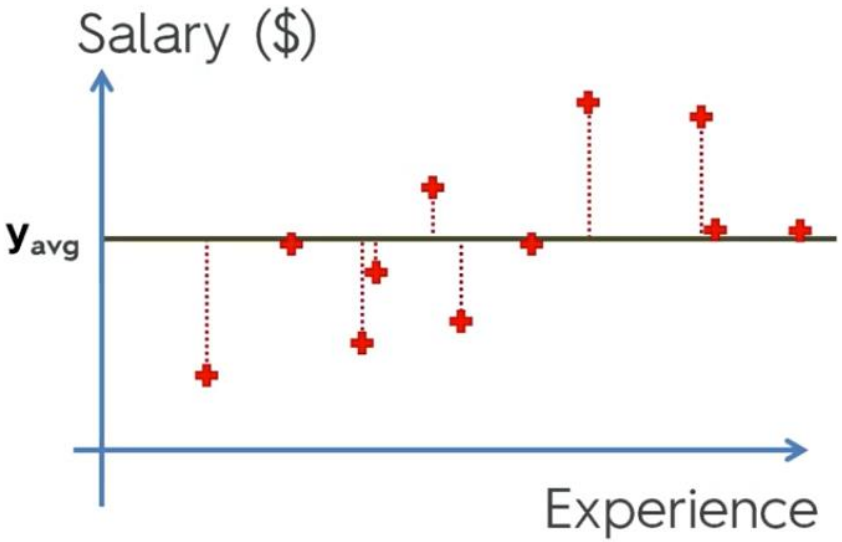
MSE



This is the sum of squares of residuals

$$SS_{res} = \sum(y_i - y_i')^2$$

MSE



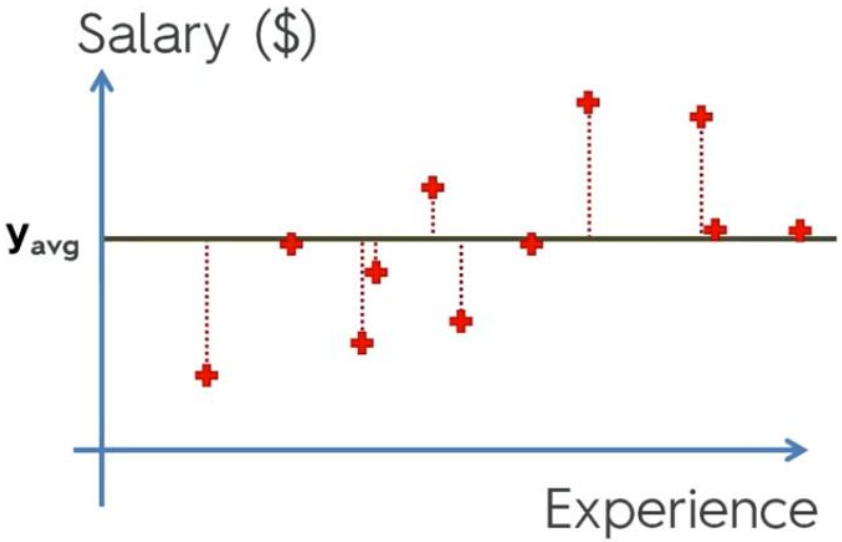
This is the sum of squares of residuals

$$SS_{res} = \sum(y_i - y_i')^2$$

The total sum of squares

$$SS_{tot} = \sum(y_i - y_{avg})^2$$

MSE



This is the sum of squares of residuals

$SS_{res} = \sum(y_i - y_i')^2$

The total sum of squares

$SS_{tot} = \sum(y_i - y_{avg})^2$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

MSE

- The total sum of squares will be the same value for a particular dataset

MSE

- The total sum of squares will be the same value for a particular dataset
- However, SS_{res} is changeable

MSE

- The total sum of squares will be the same value for a particular dataset
- However, SS_{res} is changeable
- In fact, you want to minimize it (in other words – find the best fitting line)

MSE

- The total sum of squares will be the same value for a particular dataset
- However, SS_{res} is changeable
- In fact, you want to minimize it (in other words – find the best fitting line)
- R^2 is telling us – how good is your line compared to the average line

MSE

- When you are minimizing the SS_{res} it becomes smaller

MSE

- When you are minimizing the SS_{res} it becomes smaller
- R^2 on it's turn becomes greater

MSE

- When you are minimizing the SS_{res} it becomes smaller
- R^2 on it's turn becomes greater
- Ideally, if your SS_{res} is zero (normally never happens), your R^2 will be equal to one.

MSE

- When you are minimizing the SS_{res} it becomes smaller
- R^2 on it's turn becomes greater
- Ideally, if your SS_{res} is zero (normally never happens), your R^2 will be equal to one.
- The closer R^2 to one, the better

MSE

- When you are minimizing the SS_{res} it becomes smaller
- R^2 on it's turn becomes greater
- Ideally, if your SS_{res} is zero (normally never happens), your R^2 will be equal to one.
- The closer R^2 to one, the better

$$\uparrow R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \downarrow$$

MSE

- When you are minimizing the SS_{res} it becomes smaller
- R^2 on it's turn becomes greater
- Ideally, if your SS_{res} is zero (normally never happens), your R^2 will be equal to one.
- The closer R^2 to one, the better
- Can R^2 be negative? Yes, it can. It happens if you SS_{res} fits your data worse than SS_{tot}. It's hard to do, but you can try =)

Adjusted MSE

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

$y = b_0 + b_1 * x_1$

$y = b_0 + b_1 * x_1 + b_2 * x_2$

Adjusted MSE

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

$y = b_o + b_1 * x_1$

$y = b_o + b_1 * x_1 + b_2 * x_2$

$y = b_o + b_1 * x_1 + b_2 * x_2 + \ldots + b_n * x_n$

Adjusted MSE

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

R^2 is biased!!!

$y = b_0 + b_1 * x_1$

$y = b_0 + b_1 * x_1 + b_2 * x_2$

$y = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_n * x_n$

Problem

Adjusted MSE

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

$y = b_o + b_1 * x_1$

$y = b_o + b_1 * x_1 + b_2 * x_2$

$y = b_o + b_1 * x_1 + b_2 * x_2 + \ldots + b_n * x_n$ Problem

$Ss_{res} \rightarrow \min$

Adjusted MSE

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

$y = b_0 + b_1 * x_1$

$y = b_0 + b_1*x_1 + b_2*x_2$

$y = b_0 + b_1*x_1 + b_2*x_2 + \ldots + b_n*x_n$

Problem

Ss_{res} -> min

R^2 will never decrease

Adjusted MSE

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

$y = b_0 + b_1 * x_1$

$y = b_0 + b_1*x_1 + b_2*x_2$

$y = b_0 + b_1*x_1 + b_2*x_2 + \ldots + b_n*x_n$

Problem

Ss_{res} -> min

R^2 will never decrease
It will either increase, or remain the same

Adjusted MSE

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

$y = b_0 + b_1 * x_1$

$y = b_0 + b_1*x_1 + b_2*x_2$

$y = b_0 + b_1*x_1 + b_2*x_2 + \ldots + b_n*x_n$

Ss_{res} -> min

Problem

R^2 will never decrease
It will either increase, or remain the same
Because the coefficient b_n will never be equal to zero.

Adjusted MSE

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

This is why we need the adjusted R^2

$y = b_0 + b_1 * x_1$

$y = b_0 + b_1*x_1 + b_2*x_2$

$y = b_0 + b_1*x_1 + b_2*x_2 + \ldots + b_n*x_n$

Ss_{res} -> min

Problem

R^2 will never decrease
It will either increase, or remain the same
Because the coefficient b_n will never be equal to zero.

Adjusted MSE

$$R_{adj}{}^2 = 1 - (1 - R^2)\,\frac{n - 1}{n - p - 1}$$

n – is a number of sample
p – is a number of regressors

Adjusted MSE

$$R_{adj}{}^2 = 1 - (1 - R^2)\frac{n - 1}{n - p - 1}$$

- Adjusted MSE penalizes you for adding independent variables that don't improve your model
- There will be a battle between p and R^2

```
Call:
lm(formula = Profit ~ R.D.Spend + Administration + Marketing.Spend +
    State, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-33504  -4736     90   6672  17338

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.008e+04  6.953e+03   7.204 5.76e-09 ***
R.D.Spend        8.060e-01  4.641e-02  17.369  < 2e-16 ***
Administration  -2.700e-02  5.223e-02  -0.517    0.608
Marketing.Spend  2.698e-02  1.714e-02   1.574    0.123
State2           4.189e+01  3.256e+03   0.013    0.990
State3           2.407e+02  3.339e+03   0.072    0.943
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9439 on 44 degrees of freedom
Multiple R-squared:  0.9508,    Adjusted R-squared:  0.9452
F-statistic: 169.9 on 5 and 44 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Profit ~ R.D.Spend + Administration + Marketing.Spend,
    data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-33534  -4795     63   6606  17275

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.012e+04  6.572e+03   7.626 1.06e-09 ***
R.D.Spend        8.057e-01  4.515e-02  17.846  < 2e-16 ***
Administration  -2.682e-02  5.103e-02  -0.526    0.602
Marketing.Spend  2.723e-02  1.645e-02   1.655    0.105
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9232 on 46 degrees of freedom
Multiple R-squared:  0.9507,    Adjusted R-squared:  0.9475
F-statistic:   296 on 3 and 46 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Profit ~ R.D.Spend + Marketing.Spend, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-33645  -4632   -414   6484  17097

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.698e+04  2.690e+03  17.464   <2e-16 ***
R.D.Spend       7.966e-01  4.135e-02  19.266   <2e-16 ***
Marketing.Spend 2.991e-02  1.552e-02   1.927     0.06 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9161 on 47 degrees of freedom
Multiple R-squared:  0.9505,    Adjusted R-squared:  0.9483
F-statistic: 450.8 on 2 and 47 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Profit ~ R.D.Spend, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-34351  -4626   -375   6249  17188

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.903e+04  2.538e+03   19.32   <2e-16 ***
R.D.Spend   8.543e-01  2.931e-02   29.15   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9416 on 48 degrees of freedom
Multiple R-squared:  0.9465,    Adjusted R-squared:  0.9454
F-statistic: 849.8 on 1 and 48 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Profit ~ R.D.Spend + Administration + Marketing.Spend +
    State, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-33504  -4736     90   6672  17338

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.008e+04  6.953e+03   7.204 5.76e-09 ***
R.D.Spend        8.060e-01  4.641e-02  17.369  < 2e-16 ***
Administration  -2.700e-02  5.223e-02  -0.517    0.608
Marketing.Spend  2.698e-02  1.714e-02   1.574    0.123
State2           4.189e+01  3.256e+03   0.013    0.990
State3           2.407e+02  3.339e+03   0.072    0.943
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9439 on 44 degrees of freedom
Multiple R-squared:  0.9508,    Adjusted R-squared:  0.9452
F-statistic: 169.9 on 5 and 44 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Profit ~ R.D.Spend + Administration + Marketing.Spend,
    data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-33534  -4795     63   6606  17275

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.012e+04  6.572e+03   7.626 1.06e-09 ***
R.D.Spend        8.057e-01  4.515e-02  17.846  < 2e-16 ***
Administration  -2.682e-02  5.103e-02  -0.526    0.602
Marketing.Spend  2.723e-02  1.645e-02   1.655    0.105
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9232 on 46 degrees of freedom
Multiple R-squared:  0.9507,    Adjusted R-squared:  0.9475
F-statistic:   296 on 3 and 46 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Profit ~ R.D.Spend + Marketing.Spend, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-33645  -4632   -414   6484  17097

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.698e+04  2.690e+03  17.464   <2e-16 ***
R.D.Spend       7.966e-01  4.135e-02  19.266   <2e-16 ***
Marketing.Spend 2.991e-02  1.552e-02   1.927     0.06 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9161 on 47 degrees of freedom
Multiple R-squared:  0.9505,    Adjusted R-squared:  0.9483
F-statistic: 450.8 on 2 and 47 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Profit ~ R.D.Spend, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-34351  -4626   -375   6249  17188

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.903e+04  2.538e+03   19.32   <2e-16 ***
R.D.Spend   8.543e-01  2.931e-02   29.15   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9416 on 48 degrees of freedom
Multiple R-squared:  0.9465,    Adjusted R-squared:  0.9454
F-statistic: 849.8 on 1 and 48 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Profit ~ R.D.Spend + Administration + Marketing.Spend +
    State, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-33504  -4736     90   6672  17338

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.008e+04  6.953e+03   7.204 5.76e-09 ***
R.D.Spend        8.060e-01  4.641e-02  17.369  < 2e-16 ***
Administration  -2.700e-02  5.223e-02  -0.517    0.608
Marketing.Spend  2.698e-02  1.714e-02   1.574    0.123
State2           4.189e+01  3.256e+03   0.013    0.990
State3           2.407e+02  3.339e+03   0.072    0.943
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9439 on 44 degrees of freedom
Multiple R-squared:  0.9508,    Adjusted R-squared:  0.9452
F-statistic: 169.9 on 5 and 44 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Profit ~ R.D.Spend + Administration + Marketing.Spend,
    data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-33534  -4795     63   6606  17275

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.012e+04  6.572e+03   7.626 1.06e-09 ***
R.D.Spend        8.057e-01  4.515e-02  17.846  < 2e-16 ***
Administration  -2.682e-02  5.103e-02  -0.526    0.602
Marketing.Spend  2.723e-02  1.645e-02   1.655    0.105
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9232 on 46 degrees of freedom
Multiple R-squared:  0.9507,    Adjusted R-squared:  0.9475
F-statistic:   296 on 3 and 46 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Profit ~ R.D.Spend + Marketing.Spend, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-33645  -4632   -414   6484  17097

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.698e+04  2.690e+03  17.464   <2e-16 ***
R.D.Spend       7.966e-01  4.135e-02  19.266   <2e-16 ***
Marketing.Spend 2.991e-02  1.552e-02   1.927     0.06 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9161 on 47 degrees of freedom
Multiple R-squared:  0.9505,    Adjusted R-squared:  0.9483
F-statistic: 450.8 on 2 and 47 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Profit ~ R.D.Spend, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-34351  -4626   -375   6249  17188

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.903e+04  2.538e+03   19.32   <2e-16 ***
R.D.Spend   8.543e-01  2.931e-02   29.15   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9416 on 48 degrees of freedom
Multiple R-squared:  0.9465,    Adjusted R-squared:  0.9454
F-statistic: 849.8 on 1 and 48 DF,  p-value: < 2.2e-16
```

Interpreting coeffs

```
Call:
lm(formula = Profit ~ R.D.Spend + Marketing.Spend, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-33645  -4632   -414   6484  17097

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.698e+04  2.690e+03  17.464   <2e-16 ***
R.D.Spend       7.966e-01  4.135e-02  19.266   <2e-16 ***
Marketing.Spend 2.991e-02  1.552e-02   1.927     0.06 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9161 on 47 degrees of freedom
Multiple R-squared:  0.9505,    Adjusted R-squared:  0.9483
F-statistic: 450.8 on 2 and 47 DF,  p-value: < 2.2e-16
```

Interpreting coeffs

```
Call:
lm(formula = Profit ~ R.D.Spend + Marketing.Spend, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-33645  -4632   -414   6484  17097

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.698e+04  2.690e+03  17.464   <2e-16 ***
R.D.Spend       7.966e-01  4.135e-02  19.266   <2e-16 ***
Marketing.Spend 2.991e-02  1.552e-02   1.927     0.06 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9161 on 47 degrees of freedom
Multiple R-squared:  0.9505,    Adjusted R-squared:  0.9483
F-statistic: 450.8 on 2 and 47 DF,  p-value: < 2.2e-16
```

Interpreting coeffs

```
Call:
lm(formula = Profit ~ R.D.Spend + Marketing.Spend, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-33645  -4632   -414   6484  17097

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.698e+04  2.690e+03  17.464   <2e-16 ***
R.D.Spend       7.966e-01  4.135e-02  19.266   <2e-16 ***
Marketing.Spend 2.991e-02  1.552e-02   1.927     0.06 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9161 on 47 degrees of freedom
Multiple R-squared:  0.9505,    Adjusted R-squared:  0.9483
F-statistic: 450.8 on 2 and 47 DF,  p-value: < 2.2e-16
```

If the **sign** in positive, the independent variable is correlated with the output. This means that if you will increase it, the output will increase as well

Interpreting coeffs

```
Call:
lm(formula = Profit ~ R.D.Spend + Marketing.Spend, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-33645  -4632   -414   6484  17097

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.698e+04  2.690e+03  17.464   <2e-16 ***
R.D.Spend       7.966e-01  4.135e-02  19.266   <2e-16 ***
Marketing.Spend 2.991e-02  1.552e-02   1.927     0.06 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9161 on 47 degrees of freedom
Multiple R-squared:  0.9505,    Adjusted R-squared:  0.9483
F-statistic: 450.8 on 2 and 47 DF,  p-value: < 2.2e-16
```

Magnitude is a bit tricky feature comparing to sign.
Obviously, 7.966e-01 is a higher magnitude than 2.991e-02

Interpreting coeffs

```
Call:
lm(formula = Profit ~ R.D.Spend + Marketing.Spend, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-33645  -4632   -414   6484  17097

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.698e+04  2.690e+03  17.464   <2e-16 ***
R.D.Spend       7.966e-01  4.135e-02  19.266   <2e-16 ***
Marketing.Spend 2.991e-02  1.552e-02   1.927     0.06 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9161 on 47 degrees of freedom
Multiple R-squared:  0.9505,    Adjusted R-squared:  0.9483
F-statistic: 450.8 on 2 and 47 DF,  p-value: < 2.2e-16
```

However, what if I would say that in the first case 2.991e-02 was in dollars and when I recalculate it in cents, the magnitude will increase a 100 times, overwhelming R.D.Spend

Interpreting coeffs

```
Call:
lm(formula = Profit ~ R.D.Spend + Marketing.Spend, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-33645  -4632   -414   6484  17097

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.698e+04  2.690e+03  17.464   <2e-16 ***
R.D.Spend       7.966e-01  4.135e-02  19.266   <2e-16 ***
Marketing.Spend 2.991e-02  1.552e-02   1.927     0.06 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9161 on 47 degrees of freedom
Multiple R-squared:  0.9505,    Adjusted R-squared:  0.9483
F-statistic: 450.8 on 2 and 47 DF,  p-value: < 2.2e-16
```

You should analyze magnitude as this: R.D.Spend has a greater impact on profit **per unit of R.D.Spend**, than Marketing has **per unit of Marketing spend**.