

Regression Variable Transformations| R programming language

Preliminaries

It involves set up steps like `rm` and `install.packages` commands (for `tidyverse`, `stargazer` and `magrittr` packages).

Regression with quadratic term

[1]- Linear in variables

[2]- Non-Linear in variables

Script Editor

```
wage1 <- read.csv("C:\\Users\\amalz\\OneDrive\\Desktop\\wage1.csv")

model_1 <- lm(wage ~ educ, wage1)
summary(model_1)
wage1 %<>% mutate(wagehat1 = fitted(model_1))
ggplot(data = wage1, mapping = aes(x = educ)) +
  theme_bw() +
  geom_point(mapping = aes(y = wage, col = 'wage')) +
  geom_point(mapping = aes(y = wagehat1, col = 'linear prediction'))

wage1 %<>% mutate(educsq = educ^2)
model_2 <- lm(wage ~ educ + educsq, wage1) <-- wage = beta0 + beta1*educ + beta2*educsq + u
summary(model_2)
wage1 %<>% mutate(wagehat2 = fitted(model_2))
ggplot(data = wage1, mapping = aes(x = educ)) +
  theme_bw() +
  geom_point(mapping = aes(y = wage, col = 'wage')) +
  geom_point(mapping = aes(y = wagehat2, col = 'linear prediction'))
```

These commands are self explanatory.

Console and Plots pane

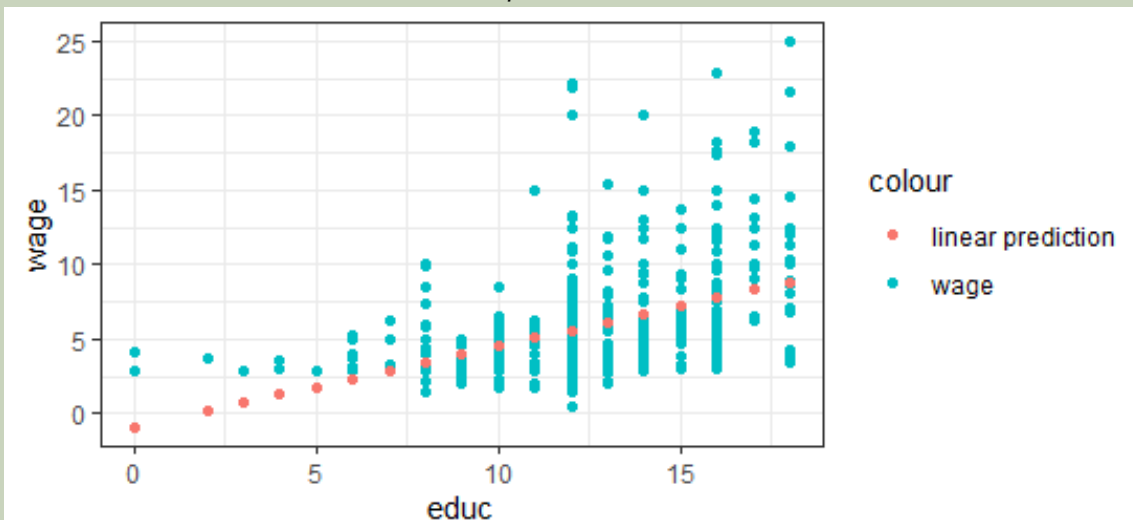
```
Call:
lm(formula = wage ~ educ, data = wage1)

Residuals:
    Min       1Q   Median       3Q      Max
-5.3396 -2.1501 -0.9674  1.1921 16.6085

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.90485    0.68497  -1.321   0.187
educ         0.54136    0.05325  10.167 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.378 on 524 degrees of freedom
Multiple R-squared:  0.1648,    Adjusted R-squared:  0.1632
F-statistic: 103.4 on 1 and 524 DF,  p-value: < 2.2e-16
```

Note that the predicted line is linear



Regression Variable Transformations| R programming language

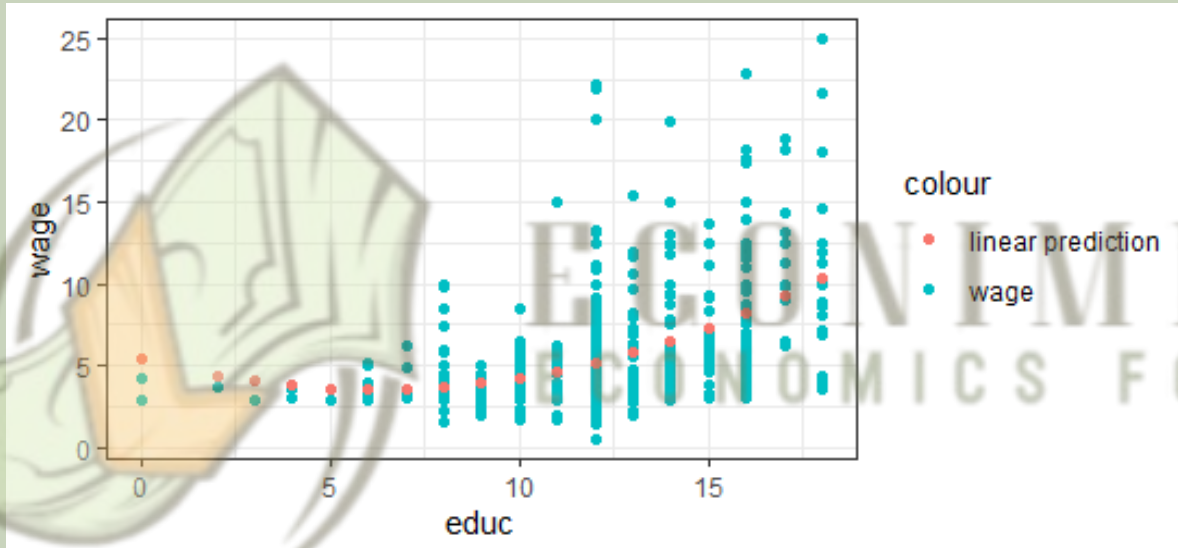
```
Call:
lm(formula = wage ~ educ + educsq, data = wage1)

Residuals:
    Min       1Q   Median       3Q      Max
-6.8722 -2.0002 -0.7472  1.2642 17.0159

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.40769    1.45886   3.707 0.000232 ***
educ        -0.60750    0.24149  -2.516 0.012181 *
educsq       0.04907    0.01007   4.872 1.46e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.307 on 523 degrees of freedom
Multiple R-squared:  0.201,    Adjusted R-squared:  0.198
F-statistic: 65.79 on 2 and 523 DF,  p-value: < 2.2e-16
```

Note that the predicted line is non-linear



Regression with quadratic term: Calculating partial effect

Note- $wage = \beta_0 + \beta_1 \cdot educ + \beta_2 \cdot educsq + u$

Script Editor

```
b_2      <- coef(model_2)
b_educ   <- b_2["educ"]
b_educsq <- b_2["educsq"]
-b_educ / (2*b_educsq)

b_educ + 2*b_educsq*5
b_educ + 2*b_educsq*6.18
b_educ + 2*b_educsq*10
b_educ + 2*b_educsq*15

(mean_educ <- mean(wage1$educ))
(pem_educ <- b_educ + 2*b_educsq*mean_educ)

pe_educ <- b_educ + 2*b_educsq*wage1$educ
(ape_educ <- mean(pe_educ))
```

*Calculate min or max point for partial effect,
educ*=-beta1/2*beta2*

Partial effect of educ on wage = $\beta_1 + 2 \cdot \beta_2 \cdot educ$

Calculate partial effect at the mean

Calculating average partial effect

[2]

Regression Variable Transformations| R programming language

Console

```
> # Calculate min or max point for partial effect, educ = -beta1/2*beta2
> b_2 <- coef(model_2)
> b_educ <- b_2["educ"]
> b_educsq <- b_2["educsq"]
> -b_educ / (2*b_educsq)
educ
6.189827
> # Partial effect of educ on wage = beta1 + 2*beta2*educ
> b_educ + 2*b_educsq*5
educ
-0.1167755
> b_educ + 2*b_educsq*6.18
educ
-0.0009644896
> b_educ + 2*b_educsq*10
educ
0.373949
> b_educ + 2*b_educsq*15
educ
0.8646734
> # Calculate partial effect at the mean
> (mean_educ <- mean(wage1$educ))
[1] 12.56274
> (pem_educ <- b_educ + 2*b_educsq*mean_educ)
educ
0.6254686
> # Calculating average partial effect
> pe_educ <- b_educ + 2*b_educsq*wage1$educ
> (ape_educ <- mean(pe_educ))
[1] 0.6254686
```

Regression with interaction term

- [1]- Regression without interaction term i.e. $\text{wage} = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + \beta_3 \text{tenure}$
- [2]- Generate interaction term
- [3]- Regression with interaction term i.e. $\text{wage} = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + \beta_3 \text{tenure} + \beta_4 \text{educ} \times \text{exper}$
- [4]- Calculate partial effect of education on wage at several levels of experience ($= \beta_1 + \beta_4 \text{exper}$)

Script Editor

```
model_5 <- lm(wage ~ educ + exper + tenure, wage1)
summary(model_5)
wage1 %<>% mutate(wagehat5 = fitted(model_5))
ggplot(wage1, aes(x = educ)) +
  theme_bw() +
  geom_point(aes(y = wage, col = 'wage')) +
  geom_point(aes(y = wagehat5, col = 'linear predictor'))
[1]

wage1 %<>% mutate(educXexper = educ*exper,
  experXtenure = exper*tenure)
[2]

model_6 <- lm(wage ~ educ + exper + tenure + educXexper, wage1)
summary(model_6)
wage1 %<>% mutate(wagehat6 = fitted(model_6))
ggplot(wage1, aes(x = educ)) +
  theme_bw() +
  geom_point(aes(y = wage, col = 'wage')) +
  geom_point(aes(y = wagehat6, col = 'linear predictor'))
[3]

b_6 <- coef(model_6)
b_educ <- b_6["educ"]
b_educexper <- b_6["educXexper"]
b_educ + b_educexper*10
b_educ + b_educexper*17
b_educ + b_educexper*30
[4]
```

Regression Variable Transformations| R programming language

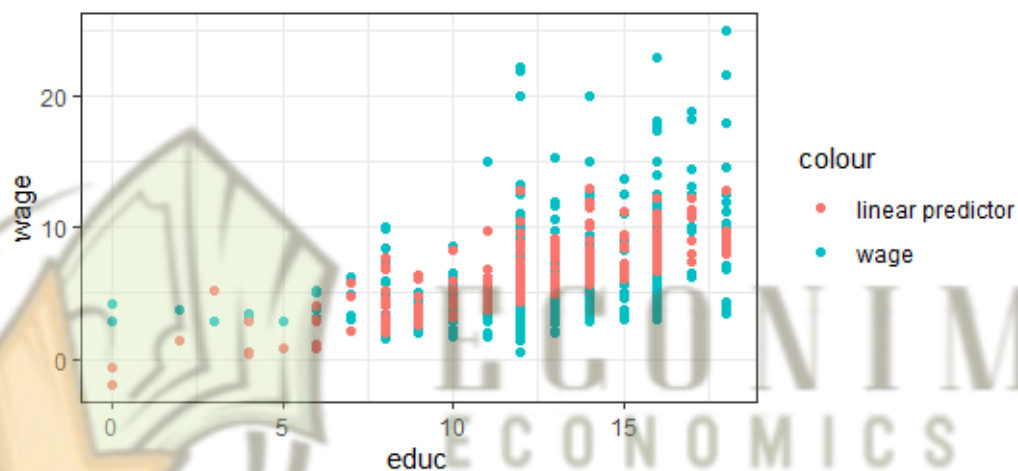
Console and Plots pane

```
Call:
lm(formula = wage ~ educ + exper + tenure, data = wage1)

Residuals:
    Min       1Q   Median       3Q      Max
-7.6068 -1.7747 -0.6279  1.1969 14.6536

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.87273    0.72896  -3.941 9.22e-05 ***
educ         0.59897    0.05128  11.679 < 2e-16 ***
exper        0.02234    0.01206   1.853  0.0645 .
tenure       0.16927    0.02164   7.820 2.93e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.084 on 522 degrees of freedom
Multiple R-squared:  0.3064,    Adjusted R-squared:  0.3024
F-statistic: 76.87 on 3 and 522 DF,  p-value: < 2.2e-16
```



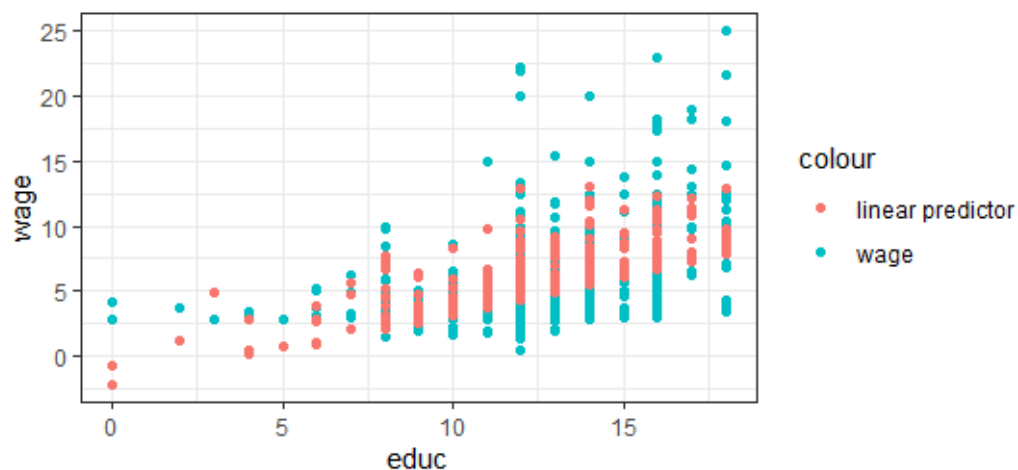
```
> summary(model_6)

Call:
lm(formula = wage ~ educ + exper + tenure + educXexper, data = wage1)

Residuals:
    Min       1Q   Median       3Q      Max
-7.6112 -1.7751 -0.6349  1.2313 14.6414

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.6534311    1.1191203  -2.371  0.0181 *
educ         0.5813908    0.0852000   6.824 2.47e-11 ***
exper        0.0123227    0.0405945   0.304  0.7616
tenure       0.1690064    0.0216877   7.793 3.58e-14 ***
educXexper    0.0008555    0.0033102   0.258  0.7962
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.087 on 521 degrees of freedom
Multiple R-squared:  0.3065,    Adjusted R-squared:  0.3012
F-statistic: 57.57 on 4 and 521 DF,  p-value: < 2.2e-16
```



Regression Variable Transformations| R programming language

Console ctd...

```
> b_educ + b_educexper*10
educ
0.5899455
> b_educ + b_educexper*17
educ
0.5959339
> b_educ + b_educexper*30
educ
0.607055
```

Regression with rescaled variables

Script Editor

```
CEOSAL2 <- read.csv(paste0(directory, "CEOSAL2.csv")) # CEO salary example
CEOSAL2 %>% mutate(salary_d = salary*1000) # Rescale sales from millions of dollars to thousands of dollars
CEOSAL2 %>% mutate(sales_k = sales*1000) # Rescale salary from thousands of dollars into dollars
CEOSAL2 %>%
  select(salary, salary_d, sales, sales_k, profits) %>% # Descriptive statistics
  stargazer(type = "text", digits = 0)
lm(salary ~ sales + profits, CEOSAL2) %>% summary
lm(salary_d ~ sales + profits, CEOSAL2) %>% summary # Regression with rescaled variables
lm(salary ~ sales_k + profits, CEOSAL2) %>% summary
lm(salary_d ~ sales_k + profits, CEOSAL2) %>% summary
```

Console

```
> CEOSAL2 %>%
+ select(salary, salary_d, sales, sales_k, profits) %>%
+ stargazer(type = "text", digits = 0)

=====
Statistic N      Mean    St. Dev.    Min      Max
-----
salary    177      866      588      100    5,299
salary_d  177 865,864 587,589 100,000 5,299,000
sales     177    3,529    6,089     29    51,300
sales_k   177 3,529,463 6,088,654 29,000 51,300,000
profits   177     288     404    -463    2,780
=====
Call:
lm(formula = salary ~ sales + profits, data = CEOSAL2)

Residuals:
    Min       1Q   Median       3Q      Max
-818.8 -313.2 -108.7   241.6  4485.1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  728.87174    47.01705   15.502 <2e-16 ***
sales         0.01749     0.01109    1.578  0.1164
profits      0.36206     0.16689    2.169  0.0314 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 539.3 on 174 degrees of freedom
Multiple R-squared:  0.1671,    Adjusted R-squared:  0.1575
F-statistic: 17.45 on 2 and 174 DF,  p-value: 1.235e-07
Call:
lm(formula = salary_d ~ sales + profits, data = CEOSAL2)

Residuals:
    Min       1Q   Median       3Q      Max
-818758 -313217 -108699  241586 4485057

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 728871.74   47017.05   15.502 <2e-16 ***
sales        17.49      11.09    1.578  0.1164
profits     362.06     166.89    2.169  0.0314 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 539300 on 174 degrees of freedom
Multiple R-squared:  0.1671,    Adjusted R-squared:  0.1575
F-statistic: 17.45 on 2 and 174 DF,  p-value: 1.235e-07
Call:
lm(formula = salary ~ sales_k + profits, data = CEOSAL2)

Residuals:
    Min       1Q   Median       3Q      Max
-818.8 -313.2 -108.7   241.6  4485.1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.289e+02  4.702e+01   15.502 <2e-16 ***
sales_k      1.749e-05  1.109e-05    1.578  0.1164
profits     3.621e-01  1.669e-01    2.169  0.0314 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 539.3 on 174 degrees of freedom
Multiple R-squared:  0.1671,    Adjusted R-squared:  0.1575
F-statistic: 17.45 on 2 and 174 DF,  p-value: 1.235e-07
```


Regression Variable Transformations| R programming language

Console ctd...

```
Call:
lm(formula = salary_d ~ sales_k + profits, data = CEOSAL2)

Residuals:
    Min       1Q   Median       3Q      Max
-818758 -313217 -108699  241586 4485057

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.289e+05  4.702e+04  15.502  <2e-16 ***
sales_k       1.749e-02  1.109e-02   1.578   0.1164
profits       3.621e+02  1.669e+02   2.169   0.0314 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 539300 on 174 degrees of freedom
Multiple R-squared:  0.1671,    Adjusted R-squared:  0.1575
F-statistic: 17.45 on 2 and 174 DF,  p-value: 1.235e-07
```

When variables are rescaled, the coefficients are rescaled.

Regression with logged rescaled variables

Script Editor

```
# Level-level regression
lm(salary ~ sales + profits, CEOSAL2) %>% summary
# Log-level regression
lm(lsalary ~ sales + profits, CEOSAL2) %>% summary
# Level-log regression
lm(salary ~ lsales + profits, CEOSAL2) %>% summary
# Log-Log regression
lm(lsalary ~ lsales + profits, CEOSAL2) %>% summary
# Rescale salary from thousands of dollars into dollars
CEOSAL2 %>% mutate(lsalary_d = log(salary_d))
# Rescale sales from millions of dollars to thousands of dollars
CEOSAL2 %>% mutate(lsales_k = log(sales_k))
# Descriptive statistics
CEOSAL2 %>%
  select(lsalary, lsalary_d, lsales, lsales_k, profits) %>%
  stargazer(type = "text", digits = 1)
# Regressions with original logged variables and rescaled logged variables
lm(lsalary ~ lsales + profits, CEOSAL2) %>% summary
lm(lsalary_d ~ lsales + profits, CEOSAL2) %>% summary
lm(lsalary ~ lsales_k + profits, CEOSAL2) %>% summary
lm(lsalary_d ~ lsales_k + profits, CEOSAL2) %>% summary
```

Console

```
> CEOSAL2 <- read.csv("C:\\Users\\amalz\\OneDrive\\Desktop\\CEOSAL2.csv")
> CEOSAL2 %>% mutate(salary_d = salary*1000)
> CEOSAL2 %>% mutate(sales_k = sales*1000)
> CEOSAL2 %>%
+   select(salary, salary_d, sales, sales_k, profits) %>%
+   stargazer(type = "text", digits = 0)
```

```
=====
Statistic  N      Mean    St. Dev.    Min      Max
-----
salary     177      866      588       100     5,299
salary_d   177  865,864  587,589  100,000 5,299,000
sales      177    3,529    6,089     29     51,300
sales_k    177 3,529,463 6,088,654 29,000 51,300,000
profits    177     208     404      -463    2,700
-----
```

```
> lm(salary ~ sales + profits, CEOSAL2) %>% summary
```

```
Call:
lm(formula = salary ~ sales + profits, data = CEOSAL2)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-818.8 -313.2 -108.7  241.6 4485.1
```

Regression Variable Transformations | R programming language

Console ctd...

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	728.87174	47.01705	15.502	<2e-16 ***
sales	0.01749	0.01109	1.578	0.1164
profits	0.36206	0.16689	2.169	0.0314 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 539.3 on 174 degrees of freedom

Multiple R-squared: 0.1671, Adjusted R-squared: 0.1575

F-statistic: 17.45 on 2 and 174 DF, p-value: 1.235e-07

```
> lm(salary_d ~ sales + profits, CEOSAL2) %>% summary
```

Call:

```
lm(formula = salary_d ~ sales + profits, data = CEOSAL2)
```

Residuals:

Min	1Q	Median	3Q	Max
-818758	-313217	-108699	241586	4485057

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	728871.74	47017.05	15.502	<2e-16 ***
sales	17.49	11.09	1.578	0.1164
profits	362.06	166.89	2.169	0.0314 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 539300 on 174 degrees of freedom

Multiple R-squared: 0.1671, Adjusted R-squared: 0.1575

F-statistic: 17.45 on 2 and 174 DF, p-value: 1.235e-07

```
> lm(salary ~ sales_k + profits, CEOSAL2) %>% summary
```

Call:

```
lm(formula = salary ~ sales_k + profits, data = CEOSAL2)
```

Residuals:

Min	1Q	Median	3Q	Max
-818.8	-313.2	-108.7	241.6	4485.1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	7.289e+02	4.702e+01	15.502	<2e-16 ***
sales_k	1.749e-05	1.109e-05	1.578	0.1164
profits	3.621e-01	1.669e-01	2.169	0.0314 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 539.3 on 174 degrees of freedom

Multiple R-squared: 0.1671, Adjusted R-squared: 0.1575

F-statistic: 17.45 on 2 and 174 DF, p-value: 1.235e-07

```
> lm(salary_d ~ sales_k + profits, CEOSAL2) %>% summary
```

Call:

```
lm(formula = salary_d ~ sales_k + profits, data = CEOSAL2)
```

Residuals:

Min	1Q	Median	3Q	Max
-818758	-313217	-108699	241586	4485057

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	7.289e+05	4.702e+04	15.502	<2e-16 ***
sales_k	1.749e-02	1.109e-02	1.578	0.1164
profits	3.621e+02	1.669e+02	2.169	0.0314 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Regression Variable Transformations| R programming language

Console ctd...

```
Residual standard error: 539300 on 174 degrees of freedom
Multiple R-squared:  0.1671,    Adjusted R-squared:  0.1575
F-statistic: 17.45 on 2 and 174 DF,  p-value: 1.235e-07
> # Level-level regression
> lm(salary ~ sales + profits, CEOSAL2) %>% summary

Call:
lm(formula = salary ~ sales + profits, data = CEOSAL2)

Residuals:
    Min       1Q   Median       3Q      Max
-818.8 -313.2 -108.7  241.6 4485.1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  728.87174    47.01705   15.502  <2e-16 ***
sales         0.01749     0.01109    1.578   0.1164
profits       0.36206     0.16689    2.169   0.0314 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 539.3 on 174 degrees of freedom
Multiple R-squared:  0.1671,    Adjusted R-squared:  0.1575
F-statistic: 17.45 on 2 and 174 DF,  p-value: 1.235e-07
> # log-level regression
> lm(lsalary ~ sales + profits, CEOSAL2) %>% summary

Call:
lm(formula = lsalary ~ sales + profits, data = CEOSAL2)

Residuals:
    Min       1Q   Median       3Q      Max
-2.00995 -0.35343  0.02414  0.38979  2.04681

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.433e+00  4.808e-02  133.796  <2e-16 ***
sales         2.556e-05  1.134e-05    2.255   0.0254 *
profits       2.872e-04  1.707e-04    1.683   0.0942 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5515 on 174 degrees of freedom
Multiple R-squared:  0.1813,    Adjusted R-squared:  0.1719
F-statistic: 19.26 on 2 and 174 DF,  p-value: 2.769e-08
> # Level-log regression
> lm(salary ~ lsales + profits, CEOSAL2) %>% summary

Call:
lm(formula = salary ~ lsales + profits, data = CEOSAL2)

Residuals:
    Min       1Q   Median       3Q      Max
-906.2 -278.8 -124.8  214.1 4391.0

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -102.2816    239.4181  -0.427 0.669755
lsales       125.1623     34.6779   3.609 0.000401 ***
profits       0.3036      0.1228    2.472 0.014379 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 523.9 on 174 degrees of freedom
Multiple R-squared:  0.214,    Adjusted R-squared:  0.205
F-statistic: 23.69 on 2 and 174 DF,  p-value: 7.955e-10
```


Regression Variable Transformations| R programming language

Console ctd...

```
> # log-log regression
> lm(lsalary ~ lsales + profits, CEOSAL2) %>% summary

Call:
lm(formula = lsalary ~ lsales + profits, data = CEOSAL2)

Residuals:
    Min       1Q   Median       3Q      Max
-2.1407 -0.3143 -0.0193  0.3225  1.9013

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.1450930   0.2347317   21.919  < 2e-16 ***
lsales       0.1936986   0.0339991    5.697  5.1e-08 ***
profits      0.0001786   0.0001204    1.483    0.14
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5137 on 174 degrees of freedom
Multiple R-squared:  0.2898,    Adjusted R-squared:  0.2817
F-statistic: 35.51 on 2 and 174 DF,  p-value: 1.17e-13
```

```
+ select(lsalary, lsalary_d, lsales, lsales_k, profits) %>%
+ stargazer(type = "text", digits = 1)

=====
Statistic  N  Mean  St. Dev.  Min   Max
-----
lsalary    177  6.6    0.6      4.6    8.6
lsalary_d  177 13.5    0.6     11.5   15.5
lsales     177  7.2    1.4      3.4   10.8
lsales_k   177 14.1    1.4     10.3   17.8
profits    177 207.8  404.5   -463  2,700
=====
```

```
> lm(lsalary_d ~ lsales_k + profits, CEOSAL2) %>% summary

Call:
lm(formula = lsalary_d ~ lsales_k + profits, data = CEOSAL2)

Residuals:
    Min       1Q   Median       3Q      Max
-2.1407 -0.3143 -0.0193  0.3225  1.9013

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.071e+01  4.676e-01  22.917  < 2e-16 ***
lsales_k     1.937e-01  3.400e-02   5.697  5.1e-08 ***
profits      1.786e-04  1.204e-04   1.483    0.14
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5137 on 174 degrees of freedom
Multiple R-squared:  0.2898,    Adjusted R-squared:  0.2817
F-statistic: 35.51 on 2 and 174 DF,  p-value: 1.17e-13
```

When Logged variables are rescaled, the coefficients do not change.