Centering and Scaling

2016-09-09

Introduction and Research Question

In this set of notes, you will continue your foray into regression analysis. To do so, we will again examine the question of whether education level is related to income. The data we will use in this set of notes comes from C. Lewis-Beck & Lewis-Beck (2016). The data contain five attributes collected from a random sample of $n = 32$ employees working for the city of Riverview, a hyopothetical midwestern city. The attributes include:

- `edu`: Years of formal education
- `income`: Annual income (in U.S. dollars)
- `senior`: Years of seniority
- `gender`: Sex (0 = Female, 1 = Male)
- `party`: Political party affiliation (0 = Democrat, 1 = Independent, 2 = Republican)

Preparation

```
# Read in data
city = read.csv(file = "~/Documents/data/Applied-Regression-Lewis-Beck/riverside_final.csv")
head(city)
```

```
  edu income senior gender party
1   8  26430      9      0     1
2   8  37449      7      1     0
3  10  34182     16      0     1
4  10  25479      1      0     2
5  10  47034     14      1     0
6  12  37656     14      1     0
```

```
# Load libraries
library(ggplot2)
library(sm)
```

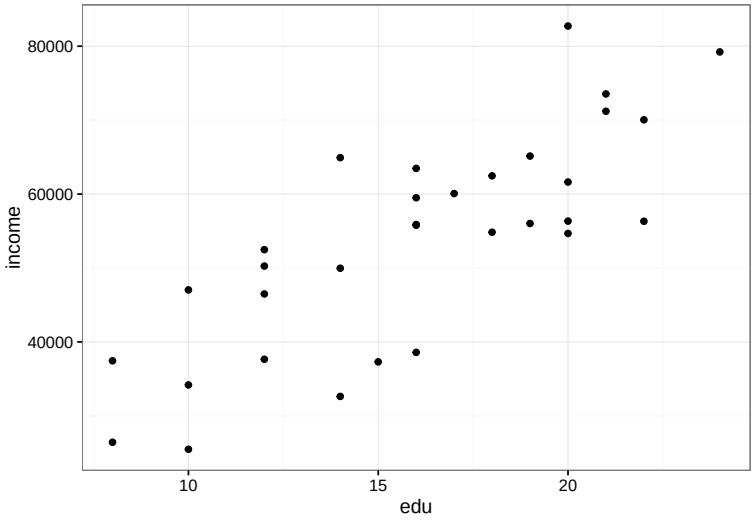
Correlation and Uncentered, Unscaled Regression Results

Here we compute the Pearson correlation coefficient between education level and income, and also examine the relationship via a scatterplot.

```
# Correlation
cor(city[, c("income", "edu")])
```

```
          income       edu
income 1.0000000 0.7947847
edu    0.7947847 1.0000000
```

```
# Scatterplot
ggplot(data = city, aes(x = edu, y = income)) + geom_point() + theme_bw()
```



We also fit the unscaled regression for reference.

```
lm.1 = lm(income ~ 1 + edu, data = city)
summary(lm.1)
```

```
Call:
lm(formula = income ~ 1 + edu, data = city)

Residuals:
   Min     1Q Median     3Q    Max
-15808  -5783   2088   5127  18379

Coefficients:
            Estimate Std. Error t value     Pr(>|t|)
(Intercept)  11321.4     6123.2   1.849       0.0743 .
edu           2651.3      369.6   7.173 0.0000000556 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8978 on 30 degrees of freedom
Multiple R-squared:  0.6317,    Adjusted R-squared:  0.6194
F-statistic: 51.45 on 1 and 30 DF,  p-value: 0.00000005562
```

Changing the Scale of Variables

What would have happened if we had measured education level in months instead of years?

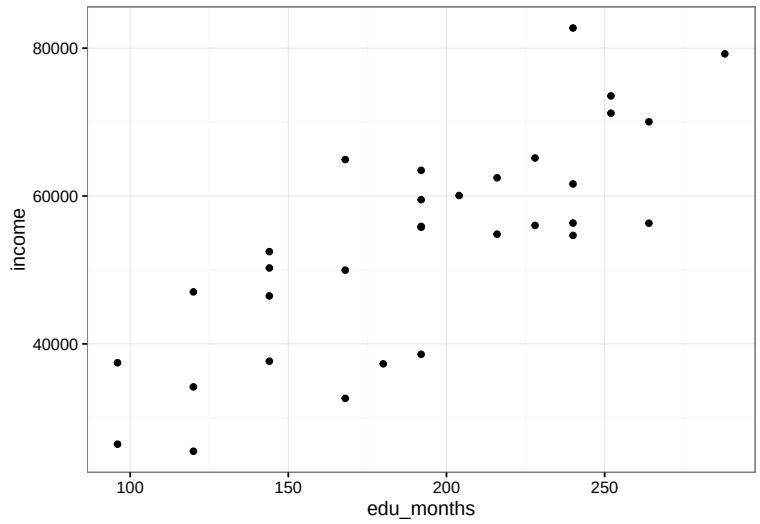
```
# Create months variable
city$edu_months = city$edu * 12
head(city)
```

```
  edu income senior gender party edu_months
1   8  26430      9      0     1         96
2   8  37449      7      1     0         96
3  10  34182     16      0     1        120
4  10  25479      1      0     2        120
5  10  47034     14      1     0        120
6  12  37656     14      1     0        144
```

```
# Correlation
cor(city[, c("income", "edu_months")])
```

```
              income edu_months
income     1.0000000  0.7947847
edu_months 0.7947847  1.0000000
```

```
# Scatterplot
ggplot(data = city, aes(x = edu_months, y = income)) + geom_point() + theme_bw()
```




```
# Fit regression with months as predictor
lm.2 = lm(income ~ 1 + edu_months, data = city)
summary(lm.2)
```

```
Call:
lm(formula = income ~ 1 + edu_months, data = city)

Residuals:
   Min     1Q Median     3Q    Max
-15808  -5783   2088   5127  18379

Coefficients:
            Estimate Std. Error t value     Pr(>|t|)
(Intercept)  11321.4     6123.2   1.849       0.0743 .
edu_months     220.9       30.8   7.173 0.0000000556 ***
---
```

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

Residual standard error: 8978 on 30 degrees of freedom
Multiple R-squared:  0.6317,    Adjusted R-squared:  0.6194
F-statistic: 51.45 on 1 and 30 DF,  p-value: 0.00000005562
```

How do the results between the two fitted regressions compare? At the model level, the R^2 values, the F-values, and model p-values are identical. Whether we measure in months or years, the relationship between educational-level and income are the same. Similarly, the correlation coeefficents are also the same.

At the coefficient level, some values are the same and others are different. The estimates for the intercept, the standard error of the intercept, as well as the t-value and p-value associated with the intercept are identical between the two analyses. This is because 0 months is the same as 0 years, so the predicted average income should be identical.

The estimates for the slope and the SE of slope are different. This is because the slope predicts the average difference in income for a one-unit difference in the predictor. For the first analysis, a one-unit difference is a one-year difference. For the second, it constitutes a one-month difference. The magnitude of the non-intercept regression coefficients *depends on the unit of measurement* of the variables.

Since the conversion from months to years is a linear transformation (e.g., can be expressed as $a + bx$), we can transform between the slope estimates quite easily. If we multiply the month slope by 12, we get the same estimate as that for year.

```
220.9 * 12
```

```
[1] 2650.8
```

The t-values and p-values associated with the slope are identical between the two analyses. Again, this is because they are both measuring the same thing, namely whether education level is statistically associated with income.

Centering

Centering a variable changes where the mean of that variable is located. To center a variable, we subtract (or add) a constant value to all the values in that variable. For example, consider centering the education level variable by subtracting 12 from each value.

```
city$edu_centered = city$edu - 12
head(city)
```

```
  edu income senior gender party edu_months edu_centered
1   8  26430      9      0     1         96           -4
2   8  37449      7      1     0         96           -4
3  10  34182     16      0     1        120           -2
4  10  25479      1      0     2        120           -2
5  10  47034     14      1     0        120           -2
6  12  37656     14      1     0        144            0
```

So now, if an employee had 12 years of formal education, their value on the centered predictor would be 0. Employees with a negative value on the centered predictor have fewer than 12 years of education, and those with positive values have more than 12 years of education.

```
# Examine means
mean(city$edu)
```

```
[1] 16
```

```
mean(city$edu_centered)
```

```
[1] 4
```

```
# Examine SDs
sd(city$edu)
```

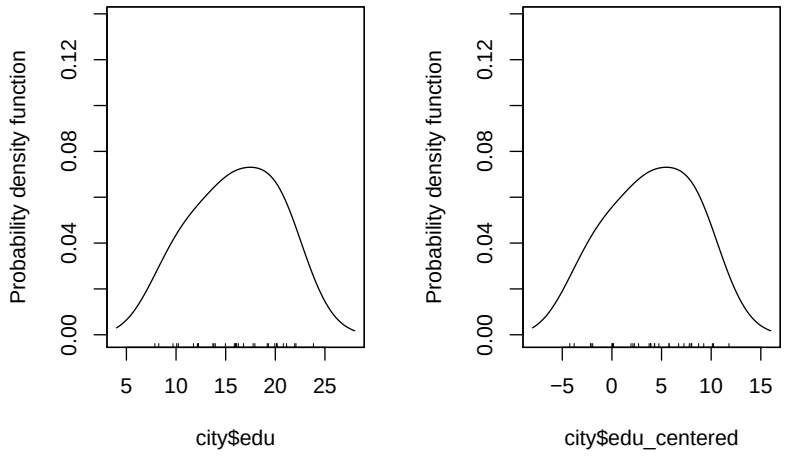
```
[1] 4.362598
```

```
sd(city$edu_centered)
```

```
[1] 4.362598
```

Note that the means for the two predictors are different (and in fact have a difference of 12)—this is the very definition of centering. The standard deviations, however, are identical. If we plot the distributions, we also see that the two distributions have the exact same shape.

```
sm.density(city$edu)
sm.density(city$edu_centered)
```

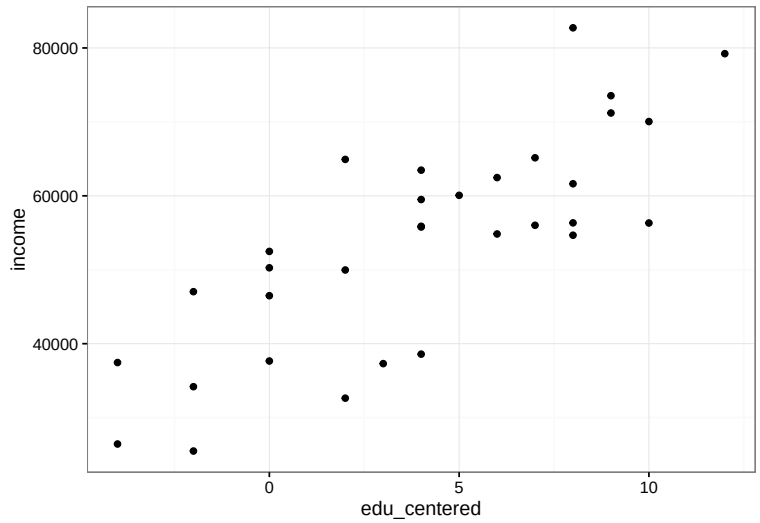


We also examine the correlation and scatterplot of income versus the centered predictor.

```
# Correlation
cor(city[, c("income", "edu_centered")])
```

```
               income edu_centered
income       1.0000000    0.7947847
edu_centered 0.7947847    1.0000000
```

```
# Scatterplot
ggplot(data = city, aes(x = edu_centered, y = income)) + geom_point() + theme_bw()
```



The correlation and the relationship depicted in the scatterplot is the same as that between the outcome and the uncentered predictor.

Regression with a Centered Predictor

```
lm.3 = lm(income ~ 1 + edu_centered, data = city)
summary(lm.3)
```

```
Call:
lm(formula = income ~ 1 + edu_centered, data = city)

Residuals:
   Min     1Q Median     3Q    Max
-15808  -5783   2088   5127  18379

Coefficients:
             Estimate Std. Error t value     Pr(>|t|)
(Intercept)   43136.9     2169.1  19.887      < 2e-16 ***
edu_centered   2651.3      369.6   7.173 0.0000000556 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8978 on 30 degrees of freedom
Multiple R-squared:  0.6317,    Adjusted R-squared:  0.6194
F-statistic: 51.45 on 1 and 30 DF,  p-value: 0.00000005562
```

- At the model level, the R^2 values, the F-values, and model p-values are identical.

- The intercept is different from the uncentered intercept. It is the predicted average income when the predictor is 0. In this analysis, it represents the predicted average income for employees with 12 years of education.
- The slope is the same as the uncentered slope. This is because the scales (SDs) for the two predictors are the same. A one-unit difference is the same in both predictors; it represents one year.

Centering a predictor changes the intercept. This can be useful if you want to test whether the average income for a particular education level is different than 0. For example, in the analysis we just performed, the t-value and p-value suggest that the average income for employees with 12 years of education is not likely to be zero; we reject the hypothesis that $\beta_0 = 0$. If you were interested in employees with 8 years of education (up through junior high), you could create another centered predictor, but subtract 8 rather than 12, and examine the information in the intercept row of the regression output.

Mean Centering

One common centering technique used by applied researchers is *mean centering*. As you might guess, this involves subtracting the mean from each value in a variable.

```
city$edu_mean_centered = city$edu - mean(city$edu)
head(city)
```

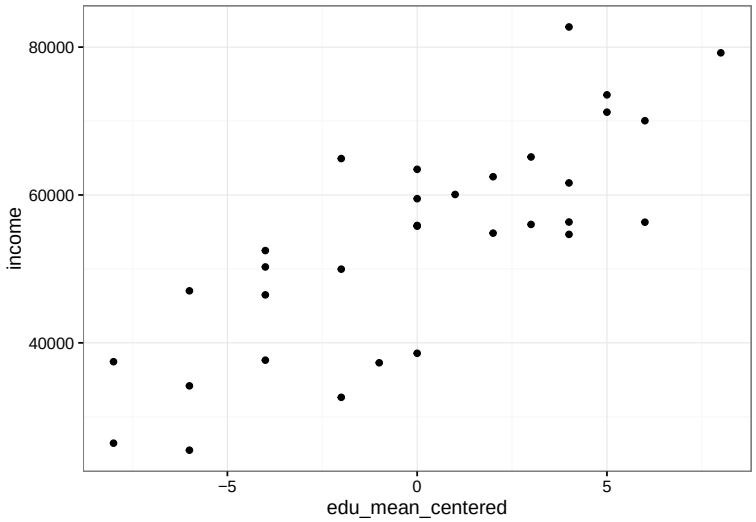
```
  edu income senior gender party edu_months edu_centered edu_mean_centered
1   8  26430      9      0     1         96           -4                -8
2   8  37449      7      1     0         96           -4                -8
3  10  34182     16      0     1        120           -2                -6
4  10  25479      1      0     2        120           -2                -6
5  10  47034     14      1     0        120           -2                -6
6  12  37656     14      1     0        144            0                -4
```

In this predictor, a value of 0 represents an employee with the average amount of education. Employees with a negative value on the mean centered predictor have an education level that is below average, and those with positive values have an education level that is above average.

```
# Correlation
cor(city[, c("income", "edu_mean_centered")])
```

```
                    income edu_mean_centered
income            1.0000000         0.7947847
edu_mean_centered 0.7947847         1.0000000
```

```
# Scatterplot
ggplot(data = city, aes(x = edu_mean_centered, y = income)) + geom_point() +
    theme_bw()
```




```
# Fit regression
lm.4 = lm(income ~ 1 + edu_mean_centered, data = city)
summary(lm.4)
```

```
Call:
lm(formula = income ~ 1 + edu_mean_centered, data = city)

Residuals:
   Min     1Q Median     3Q    Max
-15808  -5783   2088   5127  18379

Coefficients:
                  Estimate Std. Error t value     Pr(>|t|)
(Intercept)        53742.1     1587.1  33.861      < 2e-16 ***
edu_mean_centered   2651.3      369.6   7.173 0.0000000556 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8978 on 30 degrees of freedom
Multiple R-squared:  0.6317,    Adjusted R-squared:  0.6194
F-statistic: 51.45 on 1 and 30 DF,  p-value: 0.00000005562
```

Again, the correlation and scatterplot indicate the same relationship as before. The intercept now represents the predicted average income for an employee with an average education level. The hypothesis test is again to evaluate whether that average income is different than 0. Everything else in the analysis (model level and slope) are identical to the other analyses.

Scaling

Scaling a variable changes the standard deviation of that variable. To scale a variable, we divide (or multiply) each value in the variable by a fixed amount. The most common type of scaling used by applied researchers is to scale so that a variable has a SD of 1. To do this, we divide by the SD of the variable.

```
city$edu_scaled = city$edu/sd(city$edu)
head(city)
```

```
  edu income senior gender party edu_months edu_centered edu_mean_centered
1   8  26430      9      0     1         96           -4                -8
2   8  37449      7      1     0         96           -4                -8
3  10  34182     16      0     1        120           -2                -6
4  10  25479      1      0     2        120           -2                -6
5  10  47034     14      1     0        120           -2                -6
6  12  37656     14      1     0        144            0                -4
  edu_scaled
1   1.833770
2   1.833770
3   2.292212
4   2.292212
5   2.292212
6   2.750655
```

```
# Examine scaled predictor
mean(city$edu_scaled)
```

```
[1] 3.66754
```

```
sd(city$edu_scaled)
```

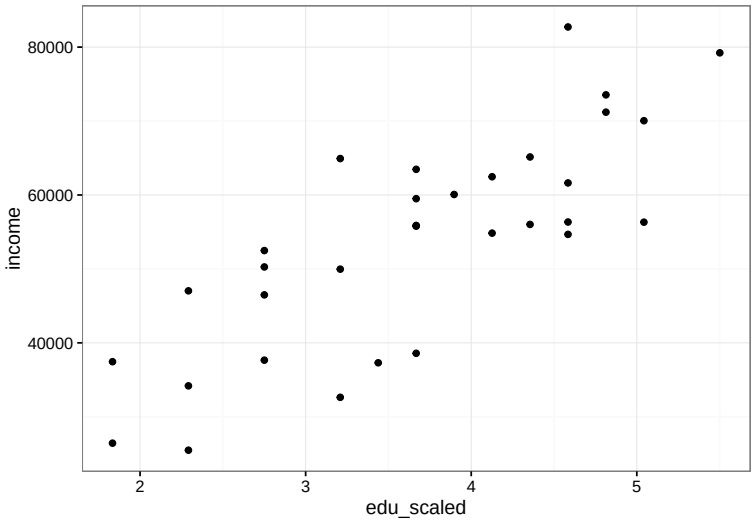
```
[1] 1
```

The SD of the scaled predictor is now 1. The mean also changes. It is the scaled mean...the old mean divided by the SD of the predictor.

```
# Correlation
cor(city[, c("income", "edu_scaled")])
```

```
              income edu_scaled
income     1.0000000  0.7947847
edu_scaled 0.7947847  1.0000000
```

```
# Scatterplot
ggplot(data = city, aes(x = edu_scaled, y = income)) + geom_point() + theme_bw()
```



Both the correlation and scatterplot, again indicate the same relationship between the variables as when the unscaled predictor was used.

Regression with a Scaled Predictor

```
lm.5 = lm(income ~ 1 + edu_scaled, data = city)
summary(lm.5)
```

```
Call:
lm(formula = income ~ 1 + edu_scaled, data = city)

Residuals:
   Min     1Q Median     3Q    Max
-15808  -5783   2088   5127  18379

Coefficients:
            Estimate Std. Error t value     Pr(>|t|)
(Intercept)    11321       6123   1.849       0.0743 .
edu_scaled     11566       1612   7.173 0.0000000556 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8978 on 30 degrees of freedom
Multiple R-squared:  0.6317,    Adjusted R-squared:  0.6194
F-statistic: 51.45 on 1 and 30 DF,  p-value: 0.00000005562
```

- At the model level, the R^2 values, the F-values, and model p-values are identical.
- The intercept is the same as the unscaled intercept. It is the predicted average income when the predictor is 0. If an employee had 0 years of education originally, dividing by the SD wouldnot change this. . . it would still be zero.
- The slope is different than the unscaled slope. A one-unit difference in the scaled predictor is a one-standard deviation difference in education level. So the slope represents the difference in income, on average, for employees whose education level differs by one standard deviation.

Scaling is typically used when the original metric is difficult to interpret, or when a one-unit difference doesn't make a lot of sense. For example, if GRE score was your predictor, talking about a one-unit difference may not make a lot of sense, so scaling migt ake the slope more interpretable.

Centering and Scaling the Outcome and Predictor: Standardized Regression

Sometimes, applied researchers will center and scale their variables. This is often referred to as standardizing variables. In your previous statistics courses, you might have learned about standardized scores, or z-scores. This is really what we are computing, z-scores for each of the variables. to do this,

$$z_x = \frac{x - \bar{x}}{SD_x}$$

```
city$z_edu = (city$edu - mean(city$edu))/sd(city$edu)
city$z_income = (city$income - mean(city$income))/sd(city$income)
head(city)
```

```
  edu income senior gender party edu_months edu_centered edu_mean_centered
1   8  26430      9      0     1         96           -4                -8
2   8  37449      7      1     0         96           -4                -8
3  10  34182     16      0     1        120           -2                -6
4  10  25479      1      0     2        120           -2                -6
5  10  47034     14      1     0        120           -2                -6
6  12  37656     14      1     0        144            0                -4
  edu_scaled      z_edu  z_income
1   1.833770 -1.8337699 -1.876729
2   1.833770 -1.8337699 -1.119568
3   2.292212 -1.3753274 -1.344057
4   2.292212 -1.3753274 -1.942076
5   2.292212 -1.3753274 -0.460943
6   2.750655 -0.9168849 -1.105344
```

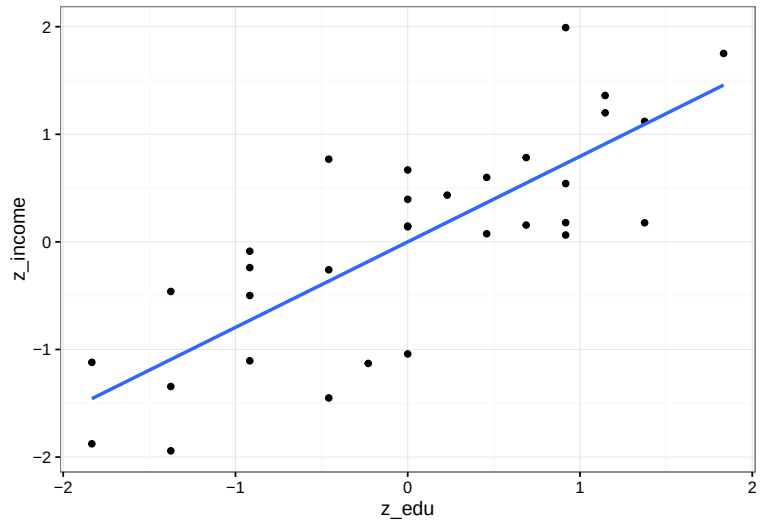
Examining the standardized predictor and outcome, we find that both variables have a mean of 0 and a standard deviation of 1. Positive values indicate that an employee is above the mean, and negative values indicate an employee below the mean. SInce the values were also scaled, the new metric is that of the standard deviation. FOr example the first employee has a standardized education level of -1.83. This indicates that that employee's education level is 1.83 standard deviations below the mean education level.

Let's also look at the scatterplot of the relationship between the standardized variables. The regression line has also been added to each plot.

```
# Correlation
cor(city[, c("z_income", "z_edu")])
```

```
          z_income     z_edu
z_income 1.0000000 0.7947847
z_edu    0.7947847 1.0000000
```

```
# Scatterplot
ggplot(data = city, aes(x = z_edu, y = z_income)) + geom_point() + geom_smooth(method = "lm",
    se = FALSE) + theme_bw()
```



The relationship is exactly the same as the others. Let's fit the OLS regression.

```
lm.6 = lm(z_income ~ 1 + z_edu, data = city)
summary(lm.6)
```

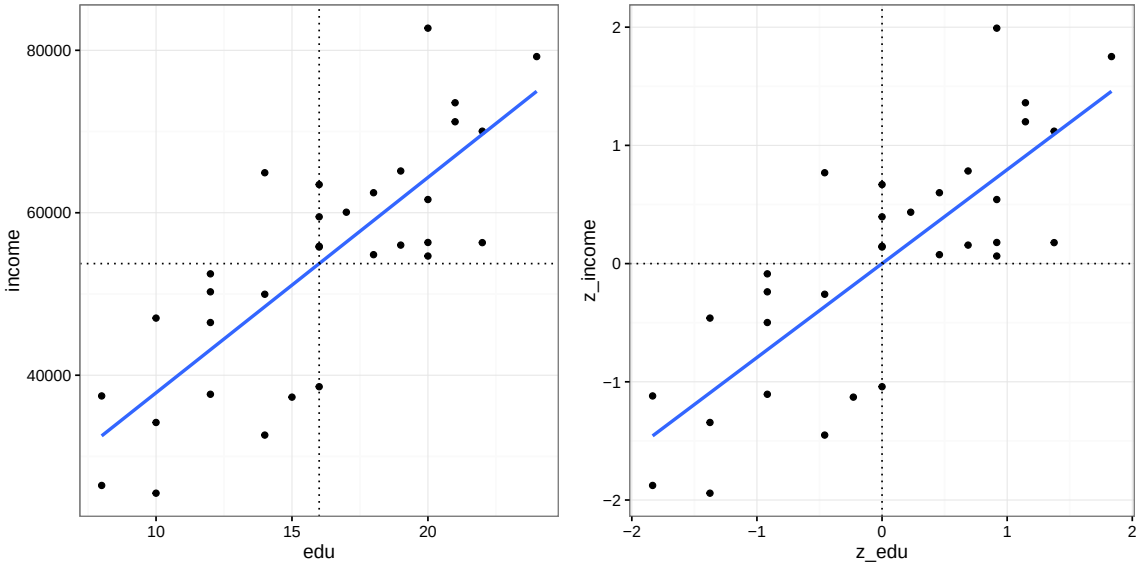
```
Call:
lm(formula = z_income ~ 1 + z_edu, data = city)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0863 -0.3974  0.1435  0.3523  1.2629

Coefficients:
             Estimate Std. Error t value     Pr(>|t|)
(Intercept) 1.097e-16  1.091e-01   0.000            1
z_edu       7.948e-01  1.108e-01   7.173 0.0000000556 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6169 on 30 degrees of freedom
Multiple R-squared:  0.6317,    Adjusted R-squared:  0.6194
F-statistic: 51.45 on 1 and 30 DF,  p-value: 0.00000005562
```

- At the model level, the R^2 values, the F-values, and model p-values are the same as before.
- The intercept estimate is 0. It is the predicted average centered and scaled income when the predictor is 0. Employees with an average education level, have, on average, an average income. Notice the t-value and p-value are results from testing whether $\beta_0 = 0$. Since $\hat{\beta}_0 = 0$, we completely fail to reject that hypothesis.
- The interpretation of slope is that, a one-standard deviation difference in education level is associated with a 0.794-standard deviation difference in income, on average. The t-value and p-values are again the same as the previous analyses.

One interesting thing is that the slope coefficient of 0.794 is the correlation coefficient we computed earlier between the original variables. In general, fitting a standardized regression results in an equation that takes the following form:

$$\hat{z}_Y = 0 + r_{X,Y}(z_X).$$

This leads us, perhaps, to a better understanding of correlation and regression. For example the plots below again show the relationships for the raw relationships and the standardized relationships. The regression line is again superimposed on the plot. We have also placed horizontal and vertical dotted lines at the mean values for X and Y in both plots.



1. The regression line will always go through the point $(\hat{\mu}_X, \hat{\mu}_Y)$. In other words, observations with an average value of X are predicted, on average, to have an average value of Y.

2. The slope for the standardized regression is the correlation coefficient between the raw variables. This implies that correlation and regression are very much linked. In unstandardized regression, there is also a relationship with the correlation, namely,

$$\hat{\beta}_1 = r_{X,Y}(\frac{SD_Y}{SD_X})$$

   The slope is a scaled multiple of the correlation coefficient, where the scaling depends on the variation in both X and Y. As such, both the correlation coefficient and slope provide information about the direction and "steepness" of the relationship. Note that "steepness" is not strength of a relationship, which is related to how tightly the observations cluster around the line, not how steep the line is. If you are interested in strength of the relationship, a measure such as R^2 would be more appropriate.

3. The direction of a correlation is a function of the observations' relationship to the mean. One way to compute correlation is,

$$r_{X,Y} = \frac{\sum(X_i - \hat{\mu}_X)(Y_i - \hat{\mu}_Y)}{n(SD_X)(SD_Y)}.$$

   The denominator is always a positive value since n and both standard deviations need to be positive (at least in order to compute a correlation). So, the numerator is what determines whether the correlation coefficient will be positive or negative. The sum constitutes products of an observations' X and Y deviations from the mean. Observations below the mean will have a negative deviation and those above the mean will have a positive deviation. So if an observation is above the mean on both X and Y,

or is below the mean on both variables, that observation's deviation product will be positive. If the observation is above the mean on one of the variables and below the mean on the other, the product of deviations will be negative. So long as the deviations for all the observations are similarly sized, the sum of the products (the value of the numerator) will depend on whether more observations are above or below the mean on both variables, or whether they are opposite on the two variables. Graphically, you can look at the four quadrants depicted by demarcating the mean values in scatterplot (as we did above). In our example, most of the observations were in the upper-right and the lower-left quadrants. These observations would have a positive product of deviations. As such, the correlation coefficient will be positive.

4. The size of a correlation is a function of the distance of the observations' to the mean. The numerator will get larger, which will make the correlation larger (all else equal) if an observation has a large deviation from one (or both) of the means. This means that an outlier, or outliers, in either direction can influence the relationship between variables. Lastly, although technically, the correlation will get smaller if the denominator gets larger, larger sample sizes and the more variation in X and Y are generally good qualities in a regression analysis.

References

Lewis-Beck, C., & Lewis-Beck, M. (2016). *Applied regression: An introduction* (2nd ed.). Sage.