Standardized Regression Coefficients

R Tutorials for Applied Statistics

Note on required packages: The following code requires the package `tidyverse`, which actually contains many packages that allow you to organize, summarize, and plot data. If you have not already done so, download, install, and load the library with the following code:

```
# This only needs to be executed once for your machine
install.packages("tidyverse")

# This needs to be executed every time you load R
library("tidyverse")
```

---

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 (`MonthlyEarnings`) and a number of factors that may influence monthly earnings including including each person's IQ (`IQ`), a measure of knowledge of their job (`Knowledge`), years of education (`YearsEdu`), years experience (`YearsExperience`), and years at current job (`Tenure`).

The code below downloads a CSV file that includes data on the above variables from 1980 for 935 individuals, and assigns it to a data frame that we name `wages`.

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

We will estimate the following multiple regression equation using the above five explanatory variables:

$$y_i = b_0 + b_1 x_{1,i} + b_2 x_{2,i} + \ldots + b_k x_{k,i} + e_i,$$

yi=b0+b1x1,i+b2x2,i+...+bkxk,i+ei,

where y_i yi denotes the *income* of individual ii, each $x_{j,i}$ xj,i denotes the value of explanatory variable j j for individual ii, and $k = 5$k=5 is the number of explanatory variables.

We can use the `lm()` function to estimate the regression as shown in the R code below. We follow this with a call the `summary()` function to display the multiple regression results to the screen.

```
lmwages <- lm(MonthlyEarnings ~ IQ + Knowledge + YearsEdu + YearsExperience + Tenure,
              data=wages)
summary(lmwages)
```

```
##
## Call:
## lm(formula = MonthlyEarnings ~ IQ + Knowledge + YearsEdu + YearsExperience +
##     Tenure, data = wages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -826.33 -243.85  -44.83  180.83 2253.35
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -531.0392   115.0513  -4.616 4.47e-06 ***
## IQ                 3.6966     0.9651   3.830 0.000137 ***
## Knowledge          8.2703     1.8273   4.526 6.79e-06 ***
## YearsEdu          47.2698     7.2980   6.477 1.51e-10 ***
## YearsExperience   11.8589     3.2494   3.650 0.000277 ***
## Tenure             6.2465     2.4565   2.543 0.011156 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 365.4 on 929 degrees of freedom
## Multiple R-squared:  0.1878, Adjusted R-squared:  0.1834
## F-statistic: 42.97 on 5 and 929 DF,  p-value: < 2.2e-16
```

2 Comparing the magnitude of regression coefficients

Let’s suppose we want to compare the explanatory variables to each other in terms of how much they each impact the outcome variable, monthly earnings.

2.1 Variables have same sacle and have comparable magnitudes

If the explanatory variables that you wish to compare are measured on the same scale, and it makes intuitive sense to compare the magnitudes *of the variables* to each other, this can be as straight forward as comparing magnitude of the regression coefficients.

For example, suppose we wanted to determine which of the following has a bigger impact on monthly earnings: an additional year of experience in your field (i.e. the `YearsExperience` variable) or an additional year of experience with your current employer (i.e. the `Tenure` variable). Each of these variables are measured in years and it does make sense to compare these two.

The coefficient on `YearsExperience` is 11.86$11.86, and the coefficient on `Tenure` is 6.25$6.25. The return to an additional year of experience in your career, while holding constant `Tenure`, is estimated to be $11.86 in additional monthly earnings. The return to an additional year of experience at your current employer ,while holding constant total experience, `YearsExperience`, is $6.25 in additional monthly earnings. The additional year of career experience has a larger impact on monthly earnings.

Generally, this method is *not* appropriate. In particular, comparing the magnitudes of coefficients is irrelevant if one of the following are true:

1. The variables are measured on the same scale, but it does not make intuitive sense to compare the magnitudes.
2. The variables are not measured on the same scale.

2.2 Problems comparing variables on the same scale

For the first case, suppose one wanted to compare the relative importance of education and experience in determining monthly earnings. Both are measured in years, so the scale of measurement is identical. However, one additional year of education and one additional year of experience are very different things.

Let us look at some summary statistics for educational attainment:

```
table(wages$YearsEdu)
```

```
##
##   9  10  11  12  13  14  15  16  17  18
##  10  35  43 393  85  77  45 150  40  57
```

```
summary(wages$YearsEdu)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00   12.00   12.00   13.47   16.00   18.00
```

```
sd(wages$YearsEdu)
```

```
## [1] 2.196654
```

The call to `table()` provides a frequency distribution for years of education. The first row shows the values for `YearsEdu` in the sample and the second row reports how many observations there are at each level. We see that most of the individuals in our sample have education levels between 12 years (high school graduate) and 16 years (college graduate).

The call to `summary()` shows some summary statistics for `YearsEdu`. We can see that the median is 12 years of education, the mean is slightly higher at 13.5 years of education. Finally, the call to `sd()` shows the standard deviation is small at 2.2 years.

Given the median level of education equal to 12 years and the small standard deviation of 2.2 years, from these statistics we can see that a single additional year of education represents an economically meaningful increase in education.

Let us also look at some summary statistics for years experience:

```
table(wages$YearsExperience)
```

```
##
##  1  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
## 12  1 29 30 48 54 72 82 72 89 65 62 54 60 68 53 30 23 14 12  3  2
```

```
summary(wages$YearsExperience)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    8.00   11.00   11.56   15.00   23.00
```

```
sd(wages$YearsExperience)
```

```
## [1] 4.374586
```

Compared to years of education, there is a much larger range for years of experience, ranging from 1 to 23. The standard deviation is approximately twice as large, equal to 4.4. A single additional year of work experience does not represent as significant a step up in the distribution as does another year of education.

The solution is to **standardize the variables**, which scales them so that changes in magnitude are directly comparable. Standardizing a variable means to convert the observations from the raw data to z-scores. Instead of measuring each person's education and experience in years, measure each variable as the number of standard deviations above or below the mean.

The `scale()` function can be used to scale variables in any arbitrary way, but the default is to standardize them. The mean is subtracted from every observation, and the variable is scaled by the inverse of the standard deviation. That is, the scaled variable is equal to the z-score,

$$z = \frac{x - \bar{x}}{s}$$

$$z = \frac{x - \bar{x}}{s}$$

Consider the regression below with standardized values for `YearsExperience` and `YearsEdu`. Notice the calls to `scale()` in the regression formula.

```
lmwages <- lm(MonthlyEarnings ~ IQ + Knowledge + scale(YearsEdu) + scale(YearsExperience) + Te
              data=wages)
summary(lmwages)
```

```
##
## Call:
## lm(formula = MonthlyEarnings ~ IQ + Knowledge + scale(YearsEdu) +
##     scale(YearsExperience) + Tenure, data = wages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -826.33 -243.85  -44.83  180.83 2253.35
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             242.7433   102.3149   2.373 0.017870 *
## IQ                        3.6966     0.9651   3.830 0.000137 ***
## Knowledge                 8.2703     1.8273   4.526 6.79e-06 ***
## scale(YearsEdu)         103.8353    16.0313   6.477 1.51e-10 ***
## scale(YearsExperience)   51.8778    14.2148   3.650 0.000277 ***
## Tenure                    6.2465     2.4565   2.543 0.011156 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 365.4 on 929 degrees of freedom
## Multiple R-squared:  0.1878, Adjusted R-squared:  0.1834
## F-statistic: 42.97 on 5 and 929 DF,  p-value: < 2.2e-16
```

The output shows that a one standard deviation increase in years of education (which happens to be an additional 2.2 years) leads to a return of $103.84 of additional monthly earnings. A one standard deviation increase in years of experience (which happens to be 4.4 years) leads to a return of $51.88. We can see that increasing education has approximately twice the impact on monthly earnings as increasing experience.

Compare these coefficients to the non-scaled regression from Section 1 above. The non-scaled regression coefficients were equal to 47.27 and 11.86 for years of education and years of experience, respectively. Failing to standardize the explanatory variables would lead to an *incorrect conclusion* that education is approximately *four times* more valuable than experience.

Compare the remaining coefficients. You can see that all other coefficients, standard errors, and all p-values are identical. Linearly scaling a variable in the regression model does not change the results for other variables.

2.3 Problems comparing variables on different scales

Suppose we wanted to compare how education versus workplace knowledge affects monthly earnings. Education is measured in years, and knowledge is a workplace intelligence test score. These scales are not comparable.

Still, we can standardize each variable. Consider the following regression:

```
lmwages <- lm(MonthlyEarnings ~ IQ + scale(Knowledge) + scale(YearsEdu) + YearsExperience + Te
              data=wages)
summary(lmwages)
```

```
##
## Call:
## lm(formula = MonthlyEarnings ~ IQ + scale(Knowledge) + scale(YearsEdu) +
##     YearsExperience + Tenure, data = wages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -826.33 -243.85  -44.83  180.83 2253.35
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      401.2281   106.9480   3.752 0.000187 ***
## IQ                 3.6966     0.9651   3.830 0.000137 ***
## scale(Knowledge)  63.1751    13.9583   4.526 6.79e-06 ***
## scale(YearsEdu)  103.8353    16.0313   6.477 1.51e-10 ***
## YearsExperience   11.8589     3.2494   3.650 0.000277 ***
## Tenure             6.2465     2.4565   2.543 0.011156 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 365.4 on 929 degrees of freedom
## Multiple R-squared:  0.1878, Adjusted R-squared:  0.1834
## F-statistic: 42.97 on 5 and 929 DF,  p-value: < 2.2e-16
```

The regression output reveals that a one standard deviation increase in knowledge of work leads to an increase in monthly earnings equal to $63.18. A one standard deviation increase in education leads to an increase in monthly earnings equal to $103.84. We can conclude that education is relatively more valuable than knowledge of work in terms of increasing monthly earnings.

3 Standardized regression

A **standardized regression** is one in which all variables are standardized. In the call to `lm()` that follows, all explanatory variables and the outcome variable are standardized:

```
lmwages <- lm(scale(MonthlyEarnings) ~
                scale(IQ) + scale(Knowledge) + scale(YearsEdu) +
                scale(YearsExperience) + scale(Tenure),
              data=wages)
summary(lmwages)
```

```
##
## Call:
## lm(formula = scale(MonthlyEarnings) ~ scale(IQ) + scale(Knowledge) +
##     scale(YearsEdu) + scale(YearsExperience) + scale(Tenure),
##     data = wages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0435 -0.6031 -0.1109  0.4472  5.5726
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -5.308e-17  2.955e-02   0.000 1.000000
## scale(IQ)               1.376e-01  3.593e-02   3.830 0.000137 ***
## scale(Knowledge)        1.562e-01  3.452e-02   4.526 6.79e-06 ***
## scale(YearsEdu)         2.568e-01  3.965e-02   6.477 1.51e-10 ***
## scale(YearsExperience)  1.283e-01  3.515e-02   3.650 0.000277 ***
## scale(Tenure)           7.840e-02  3.083e-02   2.543 0.011156 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9036 on 929 degrees of freedom
## Multiple R-squared:  0.1878, Adjusted R-squared:  0.1834
## F-statistic: 42.97 on 5 and 929 DF,  p-value: < 2.2e-16
```

Any pair of coefficients can now be compared to each other to determine which has a relatively larger impact on the outcome variable. Because all variables are standardized, the interpretation for a coefficient is the increase in number of standard deviation of monthly earnings from a one standard deviation increase in the explanatory variables.

Some examples:

- A one standard deviation increase education leads to a 0.257 standard deviation increase in monthly earnings.
- A one standard deviation increase in workplace knowledge leads to a 0.156 standard deviation increase in monthly earnings.

- A one standard deviation increase in experience leads to a 0.128 standard deviation increase in monthly earnings.

Sometimes it is more useful to interpret coefficients when the outcome variable is not scaled. Consider the following regression where the explanatory variables are scaled but the outcome variable is not:

```
lmwages <- lm(MonthlyEarnings
              ~ scale(IQ) + scale(Knowledge) + scale(YearsEdu)
              + scale(YearsExperience) + scale(Tenure),
              data=wages)
summary(lmwages)
```

```
##
## Call:
## lm(formula = MonthlyEarnings ~ scale(IQ) + scale(Knowledge) +
##     scale(YearsEdu) + scale(YearsExperience) + scale(Tenure),
##     data = wages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -826.33 -243.85  -44.83  180.83 2253.35
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)               957.95      11.95  80.165  < 2e-16 ***
## scale(IQ)                  55.64      14.53   3.830 0.000137 ***
## scale(Knowledge)           63.18      13.96   4.526 6.79e-06 ***
## scale(YearsEdu)           103.84      16.03   6.477 1.51e-10 ***
## scale(YearsExperience)     51.88      14.21   3.650 0.000277 ***
## scale(Tenure)              31.70      12.47   2.543 0.011156 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 365.4 on 929 degrees of freedom
## Multiple R-squared:  0.1878, Adjusted R-squared:  0.1834
## F-statistic: 42.97 on 5 and 929 DF,  p-value: < 2.2e-16
```

The coefficients now have the interpretation as the increase in monthly earnings (measured in dollars) from a one-standard deviation increase in the explanatory variables.

Some examples:

- A one standard deviation increase education leads to a $103.84 increase in monthly earnings.

- A one standard deviation increase in workplace knowledge leads to a $63.18 increase in monthly earnings.
- A one standard deviation increase in experience leads to a $51.88 increase in monthly earnings.