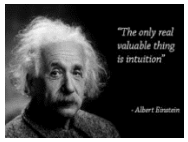
Transformation #7:
Standardization



Standardization: Intuition

- It is sometimes useful to **divide** a **centered** variable by the variable's **standard deviation**
- This produces a transformed variable with $\bar{x} = 0$ and $\sigma = 1$, often called a **"standard score"** or **"z-score"**: $y^* = \frac{y-\bar{y}}{\sigma_y}$ $x^* = \frac{x-\bar{x}}{\sigma_x}$
- This is very useful when you want to **compare dissimilar scales** (e.g., is weight larger than height?) or when the **effect size** of an unstandardized variable has **no meaning**.
- For example, in **survey** studies, we often see rating questions (e.g., rate your satisfaction from 1 to 7). So, what is the meaning of the effect from increasing the response by 1 scale point (e.g., 4 to 5)? It has no meaning.
- Standardizing x and y in $y^* = \beta_0 + \beta_1 x^* + \varepsilon$ the **interpretation** is:

  x ↑ 1 standard deviation → y ↑ β_1 standard deviations

- **Fun fact:** in a simple regression model like the one above, the resulting **standardized coefficient** is identical to the **correlation** between y and x

Standardized Regression

- If you need to standardize **1 or a few variables**, you can compute the transformed variables as explained in the prior slide
- However, a **simpler** and more efficient alternative is to ask the regression routine to report **"standardized coefficients"**, also referred to as **"beta coefficients"**.
- The regression output will thus **show both**, the regular regression coefficients and the standardized coefficients.
- A standardized coefficient for x is $\boldsymbol{\beta_x^* = \beta_x \frac{\sigma_x}{\sigma_y}}$
- So, it can be **computed directly** by multiplying the unstandardized coefficient by the ration of the standard deviations of x and y
- It can be shown mathematically that the standardized coefficients of a regular regression model are **identical** to the coefficients that would be obtained if each of the variables were standardized.
- Like with standardized variables, the standardized coefficient of x in a regression has this **interpretation**:

x ↑ 1 standard deviation → y ↑ β_1^* standard deviations

Tips

```
Standard.data = data.frame(scale(dataName,
                center=FALSE, scale=TRUE))
```
→ Same function as for centering, but using `scale=TRUE` to standardize the entire data frame with numeric values

```
dataName$variableName = scale(Boston$medv,
                        center=FALSE, scale=TRUE)
```
→ Same function as for centering, but using `scale=TRUE` to standardize a specific numeric column in the data set

An alternative to standardizing variables is to run a linear model with the original variables and then extract standardized coefficients:

`lm.beta{lm.beta}` → Useful package to extract standardized coefficients

`lm.fit = lm(y~x1+x2,data=dataName)` → Run linear model first

`lm.standardized = lm.beta(lm.fit)` → To create a linear model object that contains standardized coefficients

KOGOD SCHOOL
of
BUSINESS