Student t using correlation coefficient

Context

The purpose of this post is to show one of the problems of using the p-value as a criterion with large sample sizes (and frequentist statistics more generally). We will see how easy it is to find false positives when testing the significance of a coefficient in linear regression.

We want to show that, for a low level of correlation, it is enough to have a large enough sample size to obtain a significant p-value (which is exactly the same as having a sufficiently high t statistic).

Even if the demonstration is not really important here, the conclusion is important and should be known by all those who use linear regressions.

Main idea

We can prove that the t statistics of a given coefficient can be write as follow:

$$t = \frac{\hat{\beta} - 0}{\hat{\sigma}_{\hat{\beta}}} \approx cor(x, y) \times \sqrt{n - k}$$

If we set a Type I error threshold at 5%, it is (more or less) equivalent to rejecting the null hypothesis that $\beta = 0$ if $t > 1.96$. The equation above is fairly self-explanatory: whatever the level of correlation, if our sample size is large enough, our test will tell us that the variable associated with β has a significant impact on the predicted variable.

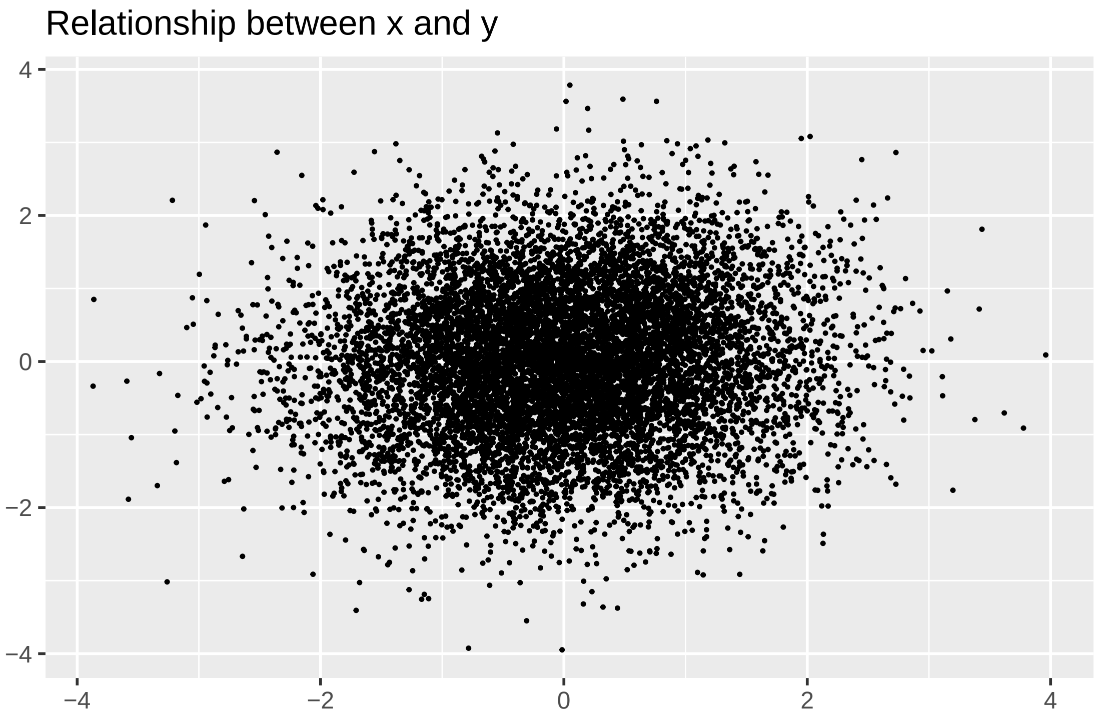
Another thing to keep in mind is that a sample size that is too small is also a problem, especially for the error-following law. With frequentist statistics, the sample size really matters and has some important limitations. Always remember that the effect size is not optional in statistics and that the p-value is not a complete result.

One could argue that a larger sample size decreases Type I and Type II errors, and increases the power of the statistic, and this is true. The problem I am talking about is the fact that the effect size needed to detect an effect does not have to be large to be significant.

Example

In this section, we will create data with a low correlation level and show that we still can find them significantly related.

```r
#plot their relationship
library(ggplot2)
library(hrbrthemes)
ggplot(data, aes(x=x, y=y)) +
  geom_point(size=0.3) +
  ggtitle("Relationship between x and y") + xlab("") + ylab("")
```



As you can see, it is difficult to see anything other than 2 variables with no particular correlation. Let's check this by using the Pearson correlation coefficient:

```
cor(x,y)
```

```
## [1] 0.06258228
```

Their correlation is actually not very important. But if we try to predict y with x and have a threshold at 5%, we will have a different conclusion:

```
regression = lm(y ~ x)
summary(regression)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9532 -0.6725  0.0051  0.6708  3.7748
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.006643   0.009951   0.668    0.504
## x           0.063035   0.010054   6.270 3.76e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9951 on 9998 degrees of freedom
## Multiple R-squared:  0.003917,   Adjusted R-squared:  0.003817
## F-statistic: 39.31 on 1 and 9998 DF,  p-value: 3.762e-10
```

The p-value is in fact well below 0.05, independently of the size effect of x on y. One way to see this is to check the R^2 which is (when there is only one explanatory variable) equal to $cor^2(x, y)$ and which is an overall quality of fit of a model to the data. Let's check it quickly:

```r
summary(regression)$r.squared
```

```
## [1] 0.003916542
```

```r
cor(x,y)^2
```

```
## [1] 0.003916542
```

It is easy to see here that the model is very bad and does not fit the data well. But if we only use the p-value as a criterion, we might think otherwise.

Proof

To prove it, we have to start from the sum of the squares of the residuals and, step by step, add the result to another equation until we reach t. This is not very complicated but it requires several steps.

We assume the following:
- $y = x\beta + \varepsilon$
- $\tilde{x}_i = \bar{x} - x_i$
- $\beta = \frac{cov(x,y)}{Var(x)}$

Step 1

$$\begin{aligned}
\sum r_i^2 &= \sum(\tilde{y}_i - \tilde{x}_i\beta)^2 \\
&= \sum(\tilde{y}_i - \tilde{x}_i\frac{cov(x,y)}{Var(x)})^2 \\
&= \sum(\tilde{y}_i^2 - 2\tilde{y}_i\tilde{x}_i\frac{cov(x,y)}{Var(x)} + \tilde{x}_i^2\frac{cov^2(x,y)}{Var^2(x)}) \\
&= \sum\tilde{y}_i^2 - 2\sum\tilde{y}_i\tilde{x}_i\frac{cov(x,y)}{Var(x)} + \sum\tilde{x}_i^2\frac{cov^2(x,y)}{Var^2(x)} \\
&= nVar(y) - 2n\frac{cov^2(x,y)}{Var(x)} + n\frac{cov^2(x,y)}{Var(x)} \\
&= nVar(y) - n\frac{cov^2(x,y)}{Var(x)} \\
&= nVar(y)(1 - cor^2(x,y))
\end{aligned}$$

Step 2

$$\begin{aligned}
Var(\varepsilon) = \sigma_\varepsilon^2 &= \frac{\sum r_i^2}{n-k} \\
&= \frac{nVar(y)(1 - cor^2(x,y))}{n-k}
\end{aligned}$$

Step 3

$$\begin{aligned}
Var(\hat{\beta}) &= \frac{\sigma_\varepsilon^2}{nVar(x)} \\
&= \frac{nVar(y)(1 - cor^2(x,y))}{nVar(x)(n-k)} \\
\sqrt{Var(\hat{\beta})} &= \frac{\sigma(y)}{\sigma(x)} \times \frac{\sqrt{1 - cor^2(x,y)}}{\sqrt{n-k}}
\end{aligned}$$

Step 4

$$t = \frac{\hat{\beta}}{\sigma(\hat{\beta})} = \frac{\sigma(y)}{\sigma(x)} \times cor(x,y) \times \frac{\sigma(x)}{\sigma(y)} \times \frac{\sqrt{n-k}}{\sqrt{1-cor^2(x,y)}}$$
$$= cor(x,y) \times \frac{\sqrt{n-k}}{\sqrt{1-cor^2(x,y)}}$$

For a “low” level of correlation, we can rewrite:

$$\approx cor(x,y) \times \sqrt{n-k}$$

The above assumption about the low level of correlation is not *costly* since we are talking specifically about the low level of correlation. When the correlation is high, it is not a problem that our t is significant. It is important to keep in mind that t is decreasing with k, the number of explanatory variables, since n is most of the time much larger than k, it does not really matter here. I don’t know if there is a similarity or a difference when working on high dimensional data.

I hope this post has helped to understand why higher is not always better for the sample size. The main idea to remember is that there is a better range for the sample size that has an upper bound.