Estimating a relative change using a log-transformation of the outcome

Brice Ozenne Brice Ozenne

December 5, 2020

1 Interpretation of the regression coefficient after log-transformation

Let's denote by Y the outcome and by G a binary group variable. We are interested in the relative change in Y between the groups. We decide to model the group effect on the log scale:

$$\log(Y) = Z = \alpha + \beta G + \varepsilon \text{ where } \mathbb{E}\left[\varepsilon\right] = 0 \text{ and } \mathbb{E}\left[\varepsilon\right] = \sigma^2$$

We claim that:

$$\frac{\mathbb{E}\left[Y|G=1\right] - \mathbb{E}\left[Y|G=0\right]}{\mathbb{E}\left[Y|G=0\right]} = e^{\beta} - 1$$

1.1 Proof: re-writting the model as a multiplicative model

We can re-write the model as:

$$Y = e^{\alpha+\beta G} e^{\varepsilon} \text{ where}$$

So for $g \in \{1,2\}$:

$$\mathbb{E}\left[Y|G=g\right] = e^{\alpha+\beta g}\mathbb{E}\left[e^{\varepsilon}\right]$$

Then:

$$\begin{aligned}\frac{\mathbb{E}\left[Y|G=1\right] - \mathbb{E}\left[Y|G=0\right]}{\mathbb{E}\left[Y|G=0\right]} &= \frac{e^{\alpha+\beta}\mathbb{E}\left[e^{\varepsilon}\right] - e^{\alpha}\mathbb{E}\left[e^{\varepsilon}\right]}{e^{\alpha}\mathbb{E}\left[e^{\varepsilon}\right]} \\ &= \frac{e^{\alpha+\beta} - e^{\alpha}}{e^{\alpha}} = e^{\beta} - 1\end{aligned}$$

1.2 Proof: using a Taylor expansion

Using a second order Taylor expansion of $\exp(Z)$ around $\mu(G) = \alpha + \beta G$ and assuming that the first moments of Z are finite and the remaining moments are neglectable regarding the factorial of the moment order (i.e. $\forall i \geq 1$, $\frac{1}{i!}\mathbb{E}\left[\varepsilon^i\right] < +\infty$ and $\sum_{i=1}^{\infty} \frac{1}{i!}\mathbb{E}\left[\varepsilon^i\right] < +\infty$), we get:

$$\begin{aligned}
Y = e^Z &= e^{\mu} + \sum_{i=1}^{\infty} \frac{1}{i!}(Z-\mu)^i \frac{\partial^i e^{\mu}}{(\partial\mu)^i} \\
&= e^{\alpha+\beta G} + \sum_{i=1}^{\infty} \frac{1}{i!}(Z-\alpha-\beta G)^i e^{\alpha+\beta G} \\
\mathbb{E}\left[Y|G=g\right] &= e^{\alpha+\beta G} + \sum_{i=1}^{\infty} \frac{1}{i!}\mathbb{E}\left[(Z-\alpha-\beta g)^i\right] e^{\alpha+\beta G} \\
&= e^{\alpha+\beta G}\left(1 + \sum_{i=1}^{\infty} \frac{1}{i!}\mathbb{E}\left[\varepsilon^i\right]\right)
\end{aligned}$$

where we used that the distribution of ε is independent of g. We can now express our parameter of interest:

$$\begin{aligned}
\Delta_G &= \frac{\mathbb{E}\left[Y|G=1\right] - \mathbb{E}\left[Y|G=0\right]}{\mathbb{E}\left[Y|G=0\right]} = \frac{\mathbb{E}\left[Y|G=1\right]}{\mathbb{E}\left[Y|G=0\right]} - 1 \\
&= \frac{e^{\alpha+\beta}\left(1 + \sum_{i=1}^{\infty} \frac{1}{i!}\mathbb{E}\left[\varepsilon^i\right]\right)}{e^{\alpha}\left(1 + \sum_{i=1}^{\infty} \frac{1}{i!}\mathbb{E}\left[\varepsilon^i\right]\right)} - 1 \\
&= e^{\beta} - 1
\end{aligned}$$

2 Note for power calculation

2.1 Recall: delta-method for normally distributed variables

Theory: we recall that for a random variable Y with finite first two moments, the delta method applied around the mean for a transformation f is:

$$f(Y) = f(\mu_Y) + f'(\mu_Y)(Y - \mu_Y) + \frac{1}{2}f''(\mu_Y)(Y - \mu_Y)^2 + \frac{1}{6}f'''(\mu_Y)(Y - \mu_Y)^3 + o\left((Y - \mu_Y)^2\right)$$

where $\mu_Y = \mathbb{E}[Y]$. Introducing $\sigma_Y^2 = \mathbb{V}ar[Y]$ and using that for a normal distribution $\mathbb{E}[(Y - \mu_Y)^3] = 0$, we have:

$$\begin{aligned}\mathbb{E}[f(Y)] =& f(\mu_Y) + f'(\mu_Y)(\mathbb{E}[Y] - \mu_Y) + \frac{1}{2}f''(\mu_Y)\mathbb{E}\left[(Y - \mu_Y)^2\right] \\ &+ \frac{1}{6}f'''(\mu_Y)\mathbb{E}\left[(Y - \mu_Y)^3\right] + o\left(\mathbb{E}\left[(Y - \mu_Y)^3\right]\right) \\ =& f(\mu_Y) + \frac{\sigma_Y^2}{2}f''(\mu_Y) + o\left(\mathbb{E}\left[(Y - \mu_Y)^3\right]\right)\end{aligned}$$

Similarly using that for a normal distribution $\mathbb{E}[(Y - \mu_Y)^4] = 3\sigma_Y^4$:

$$\begin{aligned}\mathbb{V}ar[f(Y)] =& \left(f'(\mu_Y)\right)^2 \mathbb{V}ar[\mathbb{E}[Y] - \mu_Y] + f'(\mu_Y)f''(\mu_Y)\mathbb{E}\left[(Y - \mu_Y)^3\right] \\ &+ \left(\frac{f'(\mu_Y)f'''(\mu_Y)}{3} + \frac{\left(f''(\mu_Y)\right)^2}{4}\right)\mathbb{E}\left[(Y - \mu_Y)^4\right] + o\left(\mathbb{E}\left[(Y - \mu_Y)^4\right]\right) \\ =& \left(f'(\mu_Y)\right)^2 \sigma_Y^2 + 3\sigma_Y^4\left(\frac{f'(\mu_Y)f'''(\mu_Y)}{3} + \frac{\left(f''(\mu_Y)\right)^2}{4}\right) + o\left(\mathbb{E}\left[(Y - \mu_Y)^4\right]\right)\end{aligned}$$

and introducing X with mean μ_X, variance σ_X^2, and correlation ρ with Y:

$$\begin{aligned}\mathbb{C}ov[f(X), f(Y)] =& f'(\mu_X)f'(\mu_Y)\mathbb{C}ov[X - \mu_X, Y - \mu_Y] \\ &+ \frac{1}{4}f''(\mu_X)f''(\mu_Y)\mathbb{C}ov\left[(X - \mu_X)^2, (Y - \mu_Y)^2\right] + o\left(\mathbb{C}ov\left[(X - \mu_X)^2, (Y - \mu_Y)^2\right]\right)\end{aligned}$$

Application: exponential transformation ($f = \exp$)

Using that $\mathbb{C}ov[(X - \mu_X)^2, (Y - \mu_Y)^2] \approx 2\rho^2\sigma_X^2\sigma_Y^2$:

$$\begin{aligned}\mathbb{E}[\exp(Y)] &\approx \exp(\mu_Y)\left(1 + \frac{\sigma_Y^2}{2}\right) \\ \mathbb{V}ar[\exp(Y)] &\approx \exp(2\mu_Y)\left(\sigma_Y^2 + \frac{7}{4}\sigma_Y^4\right) \\ \mathbb{C}ov[\exp(X), \exp(Y)] &\approx \exp(\mu_X + \mu_Y)\left(\rho\sigma_X\sigma_Y + \frac{1}{2}\rho^2\sigma_X^2\sigma_Y^2\right)\end{aligned}$$

⚠ these approximations are precise when the mean and variance are small

Illustration: We consider a normally distributed outcome with expectation 0.1 and variance 0.1. What is its expectation and variance after exp-transformation?

```
set.seed(10); n <- 1e4
mu <- 0.1; sigma2 <- 0.1

## first order method
mu.exp1 <- exp(mu)
var.exp1 <- exp(2*mu)*sigma2

## second order method
mu.exp2 <- exp(mu)*(1+sigma2/2)
var.exp2 <- exp(2*mu)*(sigma2 + (7/4)*sigma2^2)

## empirical value
X.exp <- exp(rnorm(n, mean = mu, sd = sqrt(sigma2)))
mu.expGS <- mean(X.exp)
var.expGS <-  var(X.exp)
```

Comparison mean:

```
rbind(value = c(first.order = mu.exp1,
        second.order = mu.exp2,
        truth = mu.expGS),
      bias = c(mu.exp1,mu.exp2,mu.expGS)-mu.expGS,
      relative.bias = (c(mu.exp1,mu.exp2,mu.expGS)-mu.expGS)/mu.expGS)
```

```
              first.order second.order    truth
value           1.1051709   1.38146365 1.425308
bias           -0.3201366  -0.04384390 0.000000
relative.bias  -0.2246088  -0.03076101 0.000000
```

Comparison variance:

```
rbind(value = c(first.order = var.exp1,
        second.order = var.exp2,
        truth = var.expGS),
      bias = c(var.exp1,var.exp2,var.expGS)-var.expGS,
      relative.bias = (c(var.exp1,var.exp2,var.expGS)-var.expGS)/var.
   expGS)
```

```
              first.order second.order   truth
value           0.6107014    1.1450651 1.35949
bias           -0.7487890   -0.2144253 0.00000
relative.bias  -0.5507865   -0.1577248 0.00000
```

The second order estimate is much more accurate, especially for the variance.

We now consider a bivariate normally distributed outcome with expectation 0.1, variance 0.1, and correlation 0.5. What is the correlation after exp-transformation?

```
set.seed(10); n <- 1e4
mu <- c(0.1,0.1); sigma2 <- c(0.1,0.1); rho <- 0.5
Sigma <- matrix(c(sigma2[1],
          rho*sqrt(prod(sigma2)),
          rho*sqrt(prod(sigma2)),
          sigma2[2]),
        2,2)
XY <- mvtnorm::rmvnorm(n, mean = mu, sigma = Sigma)
X <- XY[,1] ; Y <- XY[,2]

cov(exp(X),exp(Y))
exp(mean(X)+2*mean(Y)) * (cor(X,Y)*sd(Y)*sd(X) + 0.5*cor(X,Y)^2*var(Y)*
   var(X))
```

```
[1] 0.06839007
[1] 0.06846545
```

2.2 Two independent groups - normal distribution

Theory: consider two groups $G = 0$ and $G = 1$ for which we want to compare the percentage difference in outcome Y. We are willing to assume that on the log-scale Y is normally distributed. Our parameter of interest is:

$$\frac{\mathbb{E}[Y|G=1] - \mathbb{E}[Y|G=0]}{\mathbb{E}[Y|G=0]} = \gamma$$

we denote $\alpha = \mathbb{E}[Y|G=0]$ and we assume that on the original scale:

$$\mathbb{V}ar[Y|G=1] = \mathbb{V}ar[Y|G=0] = \sigma^2$$

How should be parametrized the gaussian distribution of $\log(Y)|G = 0$ and $\log(Y)|G = 1$ to satisfy $(\alpha, \gamma, \sigma^2)$? In other words we want to find m_0, m_1, s_0, s_1 such that:

$$\begin{aligned} Z_0 = \log(Y)|G = 0 &\sim \mathcal{N}\left(m_0, s_0^2\right) \\ Z_1 = \log(Y)|G = 1 &\sim \mathcal{N}\left(m_1, s_1^2\right) \end{aligned}$$

We can use the delta method to identify these parameters since:

$$\begin{aligned} \alpha &= \mathbb{E}[\exp(Z_0)] = \exp(a_0)\left(1 + \frac{s_0^2}{2}\right) \\ \sigma^2 &= \mathbb{V}ar[\exp(Z_0)] = \exp(2a_0)\left(s_0^2 + \frac{7}{4}s_0^4\right) \\ \alpha(\gamma+1) &= \mathbb{E}[\exp(Z_1)] = \exp(a_1)\left(1 + \frac{s_1^2}{2}\right) \\ \sigma^2 &= \mathbb{V}ar[\exp(Z_1)] = \exp(2a_1)\left(s_1^2 + \frac{7}{4}s_1^4\right) \end{aligned}$$

i.e.

$$\begin{aligned} \frac{\alpha^2}{\sigma^2} &= \frac{\left(1 + \frac{s_0^2}{2}\right)^2}{s_0^2 + \frac{7}{4}s_0^4} && \rightarrow \text{gives } s_0 \\ a_0 &= \frac{1}{2}\log\left(\frac{\sigma^2}{\left(s_0^2 + \frac{7}{4}s_0^4\right)}\right) && \rightarrow \text{gives } a_0 \\ \frac{\alpha^2(\gamma+1)^2}{\sigma^2} &= \frac{\left(1 + \frac{s_1^2}{2}\right)^2}{s_1^2 + \frac{7}{4}s_1^4} && \rightarrow \text{gives } s_1 \\ a_1 &= \frac{1}{2}\log\left(\frac{\sigma^2}{\left(s_1^2 + \frac{7}{4}s_1^4\right)}\right) && \rightarrow \text{gives } a_1 \end{aligned}$$

The first and third equation can be solved numerically.

Illustration: We consider two groups having a 10% difference in their baseline value ($\alpha = 1.15$) and a variance of $\sigma^2 = 0.15$. What are the parameters of the corresponding normal distribution on the log-scale and the standardized effect size?

```
alpha <- 1.15
sigma2 <- 0.15
gamma <- 0.1
```

Solve the equations:

```
s0 <- uniroot(function(x){alpha^2/sigma2 - (1+x/2)^2/(x+x^2*7/4)},
          interval = c(0,1))$root
a0 <- log(sigma2/(s0+s0^2*7/4))/2
s1 <- uniroot(function(x){alpha^2*(gamma+1)^2/sigma2 - (1+x/2)^2/(x+x^2
    *7/4)},
          interval = c(0,1))$root
a1 <- log(sigma2/(s1+s1^2*7/4))/2
c(a0 = a0, s0 = s0, a1 = a1, s1 = s1)
```

```
        a0          s0          a1          s1
0.08802784 0.10608948 0.19175319 0.08851048
```

We can check that `uniroot` converged correctly:

```
c(exp(a0)*(1+s0/2) - alpha,
  exp(2*a0)*(s0+s0^2*7/4) - sigma2,
  exp(a1)*(1+s1/2) - alpha*(1+gamma),
  exp(2*a1)*(s1+s1^2*7/4) - sigma2)
```

```
[1] -5.563198e-05  0.000000e+00 -1.895835e-05  0.000000e+00
```

and the variables have the appropriate distribution:

```
Z0 <- exp(rnorm(1e4, mean=a0, sd = sqrt(s0)))
Z1 <- exp(rnorm(1e4, mean=a1, sd = sqrt(s1)))
c(alpha = mean(Z0),
  gamma = (mean(Z1)-mean(Z0))/mean(Z0),
  sigma2 = var(Z0),
  sigma2 = var(Z1))
```

```
    alpha     gamma    sigma2    sigma2
1.1435272 0.1090391 0.1473705 0.1507638
```

For a power calculation we would use:

```
pwr.t.test(d = (a1-a0)/sqrt(s0/2+s1/2), sig.level = 0.05, power = 0.8)
## dvmisc::power_2t_unequal(n = 143, d = a1-a0, sigsq1 = s0, sigsq2 =
    s1, alpha = 0.05)
```

```
     Two-sample t test power calculation

              n = 142.9312
              d = 0.3325282
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

2.3 Two independent groups - log-normal distribution

An alternative approach is to use a log-normal distribution. Random variables with log normal distribution have their logarithm equal to a specific value a and their standard deviation equal to a specific value s. So we want to get:

$$\begin{aligned}
\alpha &= \exp(a_0 + \frac{1}{2}s_0^2) \\
\sigma^2 &= \exp(2 * a_0 + s_0^2) * (\exp(s_0^2) - 1) \\
\alpha(1+\gamma) &= \exp(a_1 + \frac{1}{2}s_1^2) \\
\sigma^2 &= \exp(2 * a_1 + s_1^2) * (\exp(s_1^2) - 1)
\end{aligned}$$

So

$$\begin{aligned}
s_0 &= \log\left(1 + \frac{\sigma^2}{\alpha^2}\right) \\
a_0 &= \log(\alpha) - \frac{s_0^2}{2} \\
s_1 &= \log\left(1 + \frac{\sigma^2}{\alpha * (1+\gamma)^2}\right) \\
a_1 &= \log(\alpha * (1+\gamma)) - \frac{s_1^2}{2}
\end{aligned}$$

Illustration: We still consider two groups having a 10% difference in their baseline value ($\alpha = 1.15$) and a variance of $\sigma^2 = 0.15$. What are the parameters of the corresponding normal distribution on the log-scale and the standardized effect size?

```
alpha <- 1.15
sigma2 <- 0.15
gamma <- 0.1
```

We identify the parameters of the log-normal distributions:

```
s0 <- log(1+sigma2/alpha^2)
a0 <- log(alpha) - s0/2
s1 <- log(1+sigma2/(alpha*(1+gamma))^2)
a1 <- log(alpha*(1+gamma)) - s1/2
c(a0 = a0, s0 = s0, a1 = a1, s1 = s1)
```

```
        a0         s0         a1         s1
0.08604307 0.10743775 0.19027207 0.08960011
```

We can check that the variables have the appropriate distribution:

```
Z0 <- rlnorm(1e4, mean=a0, sd = sqrt(s0))
Z1 <- rlnorm(1e4, mean=a1, sd = sqrt(s1))
c(alpha = mean(Z0),
  gamma = (mean(Z1)-mean(Z0))/mean(Z0),
  sigma2 = var(Z0),
  sigma2 = var(Z1))
```

```
    alpha     gamma    sigma2    sigma2
1.1480725 0.1019535 0.1455856 0.1510286
```

For a power calculation we would use:

```
pwr.t.test(d = (a1-a0)/sqrt(s0/2+s1/2), sig.level = 0.05, power = 0.8)
## dvmisc::power_2t_unequal(n = 143, d = a1-a0, sigsq1 = s0, sigsq2 =
    s1, alpha = 0.05)
```

```

     Two-sample t test power calculation

              n = 143.3238
              d = 0.3320693
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

3 Moments of the normal distribution

Denote X and Y two normally distributed variables, with mean μ_X,μ_Y and variance σ_X^2,σ_Y^2. Then:

- $\mathbb{E}\left[X^3\right] = 3\mu_X\sigma_X^2 + \mu_X^3$
- $\mathbb{E}\left[X^4\right] = 3\left(\sigma_X^2\right)^2 + 6\sigma_X^2\mu_X^2 + \mu_X^4$
- $\mathbb{C}ov\left[X^2, X\right] = 2\mu_X\sigma_X^2$
- $\mathbb{C}ov\left[X^2, Y\right] = 2\mu_X\rho\sigma_X\sigma_Y$
- $\mathbb{E}\left[X^2 * Y^2\right] = (\sigma_X^2 + \mu_X^2)(\sigma_Y^2 + \mu_Y^2) + 2\rho^2\sigma_X^2\sigma_Y^2 + 4\rho\sigma_Y\sigma_X\mu_X\mu_Y$
- $\mathbb{C}ov\left[\left(X - \mu_X\right)^2, \left(Y - \mu_Y\right)^2\right] = 2\rho^2\sigma_X^2\sigma_Y^2$