Regression to the mean

André Meichtry

May 7, 2024

> Shrinkage of results can be seen to be a necessary fact of life.
>
> *(Stephen Senn)*

1 Introduction

Regression to the mean is a statistical phenomenon wherein extreme observations tend to move closer to the mean upon subsequent measurements. This phenomenon was first observed by Sir Francis Galton in the late 19th century (Galton 1889: *Regression toward mediocrity*) and has since been studied extensively in statistics and related fields.

2 Mathematical Definition

Consider two random variables, X and Y, where X represents the initial measurement and Y represents the subsequent measurement. Let μ be the mean of X, and σ^2 be the variance of X.

According to the principle of regression to the mean, if X is extreme (far from μ) on its first measurement, Y will tend to be closer to μ on its second measurement. Mathematically, the conditional expectation of Y given X is expressed as:

$$E(Y|X) = \mu_Y + \underbrace{\frac{\sigma_{X,Y}}{\sigma_X^2}}_{\beta}(X - \mu_X), \tag{1}$$

where μ_Y is the mean of Y, σ_X^2 is the variance of X, and $\sigma_{X,Y} = \rho\sigma_X\sigma_Y$ is the covariance between X and Y, β is the regression coefficient. This can be written in a standardized form as

$$\frac{E(Y|X) - \mu_Y}{\sigma_Y} = \rho\frac{X - \mu_X}{\sigma_X}. \tag{2}$$

$$Z_{Y|X} = \rho Z_X \tag{3}$$

If $\rho < 1$, then (X, Y) shows regression toward the mean (by this definition).

3 Placebo versus Regression to the mean

Placebo effects can often be interpreted as a purely statistical – not a psychological – phenomenon.

See R package for illustration: https://mcdr65.github.io/RegToMeanExample/

```
library(RegToMeanExample)
```

Assuming no true change.

We simulate correlated pre-post diastolic blood pressure data assuming *no change* from baseline to follow-up: simulations from parameters: $\rho_{BL,FU} = 0.76, \mu_{BL} = \mu_{FU} = 90, \sigma_{BL} = \sigma_{FU} = 8$. Then let us look at the *subgroup* of "hypertensive at baseline" only. We have regression to the mean, since $\beta_{FU|BL} = \frac{\sigma_{FU,BL}}{\sigma^2_{BL}} = r < 1$.

Paired t-test for all and for extreme group:

```
DBP.RTM(n=200,show.plot = FALSE,show.out=TRUE)$ttestall
```

```
##
##  Paired t-test
##
## data:  X[, 1] and X[, 2]
## t = -3.6e-16, df = 199, p-value = 1
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.7728  0.7728
## sample estimates:
## mean difference
##      -1.421e-16
```

```
DBP.RTM(n=200,show.plot = FALSE,show.out=TRUE)$ttestextrem
```

```
##
##  Paired t-test
##
## data:  X2[, 1] and X2[, 2]
## t = 3.2, df = 52, p-value = 0.002
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  0.9363 4.0209
## sample estimates:
## mean difference
##           2.479
```

```
args(DBP.RTM)

## function (mu = 90, sigma = 8, r = 0.76, n = 1000, limit = 95,
##     TrueChange = 0, show.plot = TRUE, show.out = FALSE)
## NULL

DBP.RTM(n=200)
```

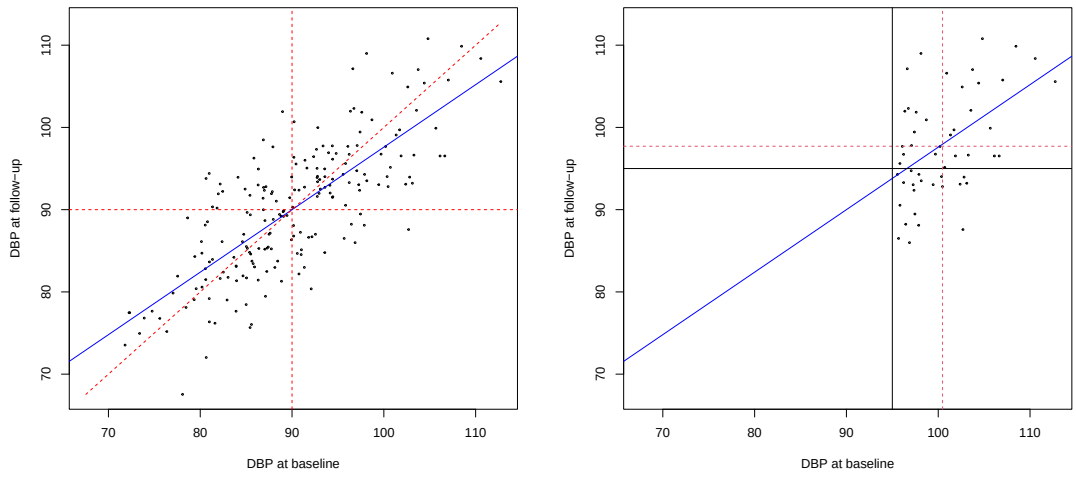
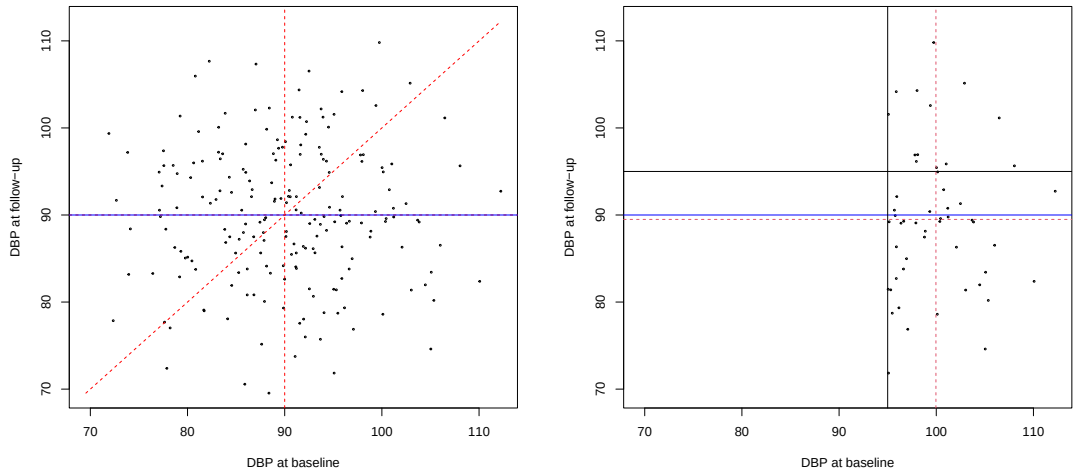


Figure 1: Simulation of diastolic blood pressure data. Simulations from parameters: $\rho_{BL,FU} = 0.76, \mu_{BL} = \mu_{FU} = 90, \sigma_{BL} = \sigma_{FU} = 8$. Left panel: Baseline versus Follow-up for diastolic blood pressure: no change in the mean. Right panel: Baseline versus Follow-up for "hypertensive at baseline" only. We observe an apparent change due to regression to the mean (Solid line: Regression of follow-up on baseline-measure (that is, by fixing baseline)). Dashed lines: mean values and equality lines.

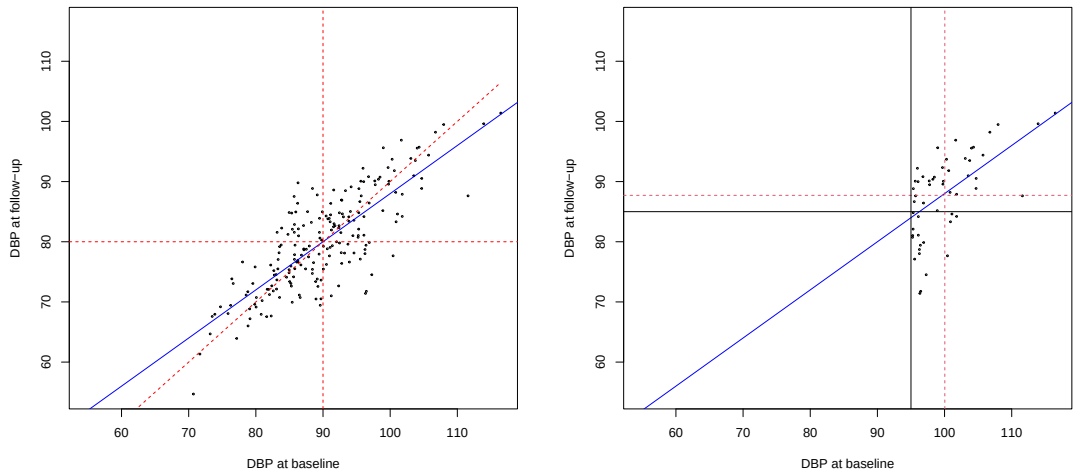
Extreme case: ρ=0

```
DBP.RTM(n=200,r=0)
```



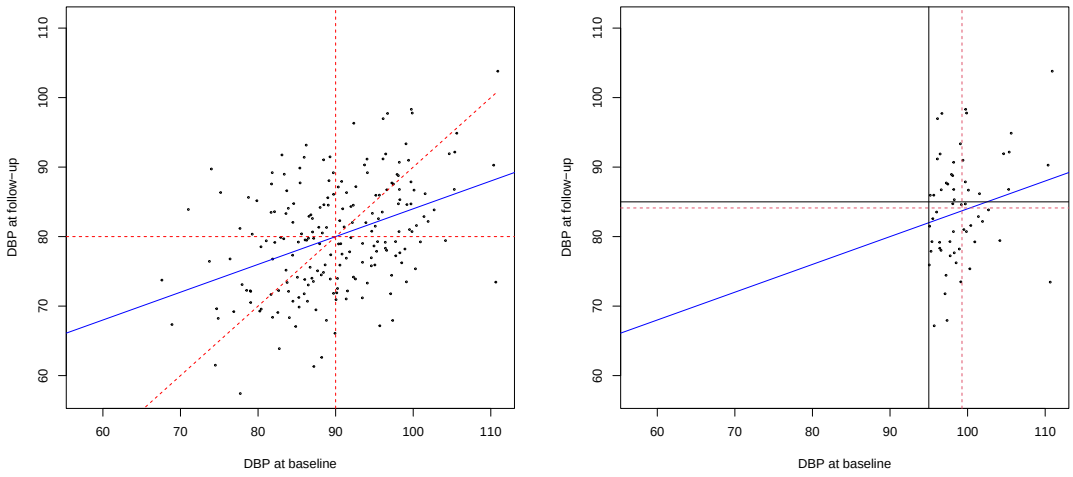
Including a true change of -10 and ρ=.8

```
DBP.RTM(n=200,r=0.8,TrueChange=-10)
```



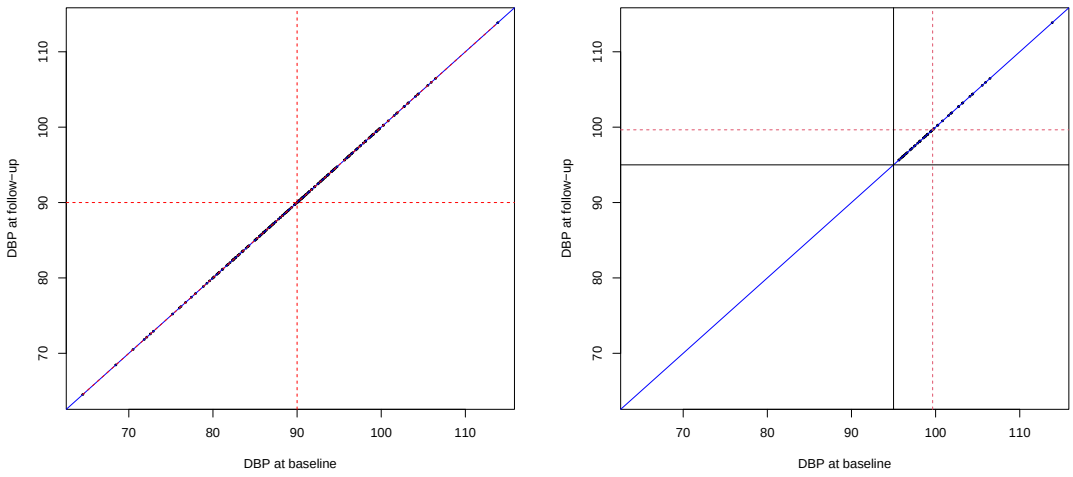
Including a true change of -10 and ρ=.4

```
DBP.RTM(n=200,r=.4,TrueChange=-10)
```



ρ=1

```
DBP.RTM(n=200,r=1,TrueChange=0)
```



4 True and observed

Assume true diastolic blood pressure, τ, at baseline is measured with error ϵ so that

$$X = \tau + \epsilon \tag{4}$$

is the *observed* blood pressure.

Let the *true* mean difference between patients be Δ and the *observed* mean difference D, then the expectation of D is

$$\mathrm{E}(D \mid \Delta = \delta) = \delta, \tag{5}$$

However, the contrary is not true. We have for the expectation of Δ, given an observed difference d,

$$\boxed{|\mathrm{E}(\Delta \mid D = d)| < |d|}, \tag{6}$$

and we have *regression to the mean*[1].

Reliability as upper bound The maximal possible correlation between Δ and D is $\sqrt{rel_D}$.

5 Myths

- *Regression to the mean leads to diminished variance*: Variance remains constant since regression to the mean works in both directions. Extreme values on "post"-measure have less-extreme values on "pre"-measure. Regression to the mean is **not** a directed or temporal effect.
- *Pre-Post changes are biased by Regression to the mean*: This is only true for an "extreme" subgroup.
- *Regression to the mean is only induced by the reliability of the measure*: For example, body heights of mothers and daughters could be measured with perfect reliability, but there is still regression to the mean, because there is still random variation in the observed height.

[1] In a Bayesian approach, shrinking is *natural* and we have *inverse unbiasedness*. Most bayesians are rather unconcerned about unbiasedness (in the formal sampling-theory sense) of their estimates. For example, Gelman et al (1995) write: "From a Bayesian perspective, the principle of unbiasedness is reasonable in the limit of large samples, but otherwise it is potentially misleading. Unbiasedness as conventionally understood is not a necessary property of good inferences". Assume without loss of generality $\mathrm{E}(\Delta) = 0$ and $\widehat{\Delta}$ an unbiased estimate of a given effect Δ and $\widehat{\Delta}_{shrunk}$ a shrunk estimate. Although $\widehat{\Delta}_{shrunk}$ is not unbiased in the classic forward sense, $\mathrm{E}(\widehat{\Delta}_{shrunk} \mid \Delta = \delta) \neq \delta$ it is unbiased in the Bayesian backward sense: $\mathrm{E}(\Delta \mid \widehat{\Delta}_{shrunk} = \widehat{\delta}_{shrunk}) = \widehat{\delta}_{shrunk}$.

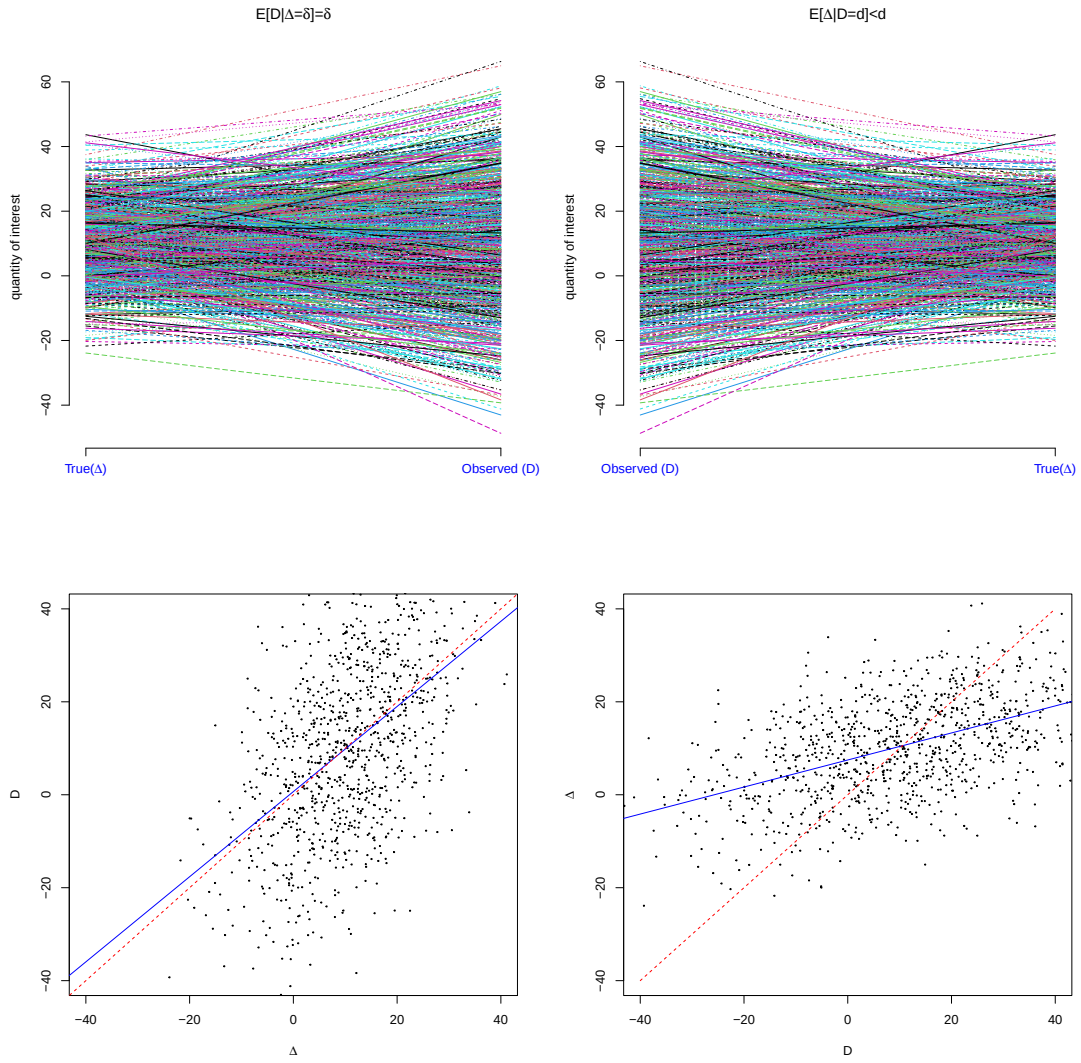


Figure 2: Regression of observed on true (left) and of true on observed (right). Dashed line: equality, Solid line: regression line.