Eco311 Optional Reading: Regression towards Mean

(Jing Li, Miami University)

1. The `GaltonFamilies` data in the `HistData` package show an interesting pattern: offspring of tall parents are, on average, not as tall as their parents. Meanwhile, offspring of short parents are, on average, not as short as their parents. In other words, the kids' heights tend to be closer to the population average than parents, a phenomenon called *Regression towards Mean or Mediocrity* (RTM) .

2. RTM appears everywhere. China has an old saying: "A rich family cannot keep being rich for three generations," or, kids born in a rich family tend to earn less money than their parents. In finance literature, it is routinely found that the performance of managers of a historically excellent mutual fund tends to worsen over time.

3. Let's duplicate the finding of RTM with `GaltonFamilies` data. The first six observations are

```
> library(HistData)
> data(GaltonFamilies)
> attach(GaltonFamilies)
> head(GaltonFamilies)
  family father mother midparentHeight children childNum gender childHeight
1    001   78.5   67.0           75.43        4        1   male        73.2
2    001   78.5   67.0           75.43        4        2 female        69.2
3    001   78.5   67.0           75.43        4        3 female        69.0
4    001   78.5   67.0           75.43        4        4 female        69.0
5    002   75.5   66.5           73.66        4        1   male        73.5
6    002   75.5   66.5           73.66        4        2   male        72.5
```

   (a) The data look like *panel data*—the first four observations belong to the first family that has one son and three daughters.

   (b) The dad and mom in that family are tall: father's height is 78.5, and mother's height is 67.0. The average parent height is $(78.5 + 67.0)/2 = 72.75$.

   (c) However, the average kid height is $73.2/2 + (69.2 + 69.0 + 69.0)/3/2 = 71.13333$. So, on average, the kids are not as tall as their parents $(71.13333 < 72.75)$—RTM applies to that family.

4. Across the whole sample, RTM still holds. For instance, focus on fathers and sons. The average father height is 69.19711. Using that number to divide the sample into tall and short dads, we see on average sons of tall dads are not that tall $70.08444 < 71.16586$, while sons of short dads are not that short $68.09333 > 66.8044$.

   ```
   > mean(father)
   [1] 69.19711
   > mean(father[father>69])
   [1] 71.16586
   > mean(childHeight[father>69&gender=="male"])
   [1] 70.08444
   > mean(father[father<69])
   [1] 66.8044
   > mean(childHeight[father<69&gender=="male"])
   [1] 68.09333
   ```

   The same pattern is found for mothers and daughters

   ```
   > mean(mother)
   [1] 64.08929
   > mean(mother[mother>64])
   [1] 65.97675
   > mean(childHeight[mother>64&gender=="female"])
   [1] 64.44886
   > mean(mother[mother<64])
   [1] 61.9095
   > mean(childHeight[mother<64&gender=="female"])
   [1] 63.53029
   ```

   To sum up, we confirm the RTM—kids tend to be closer to the mean than parents.

5. Next we use regression to illustrate RTM. Basically a simple regression model relates the parent height (x variable) to kid height (y variable)

```
> lm(childHeight[gender=="male"]~father[gender=="male"])$coef
(Intercept) father[gender == "male"]
38.3625810                0.4465226
> lm(childHeight[gender=="female"]~mother[gender=="female"])$coef
(Intercept) mother[gender == "female"]
43.6889690                  0.3182446
```

(a) The first regression looks into dads and sons. The slope coefficient is 0.4465226. The interpretation is, on average, if a dad is one inch taller (shorter) than the average, a son is only 0.4465226 inch taller (shorter) than the average. So sons move or regress towards the mean.

(b) The second regression uses mothers and daughters, and the slope coefficient is 0.3182446, implying that if a mother is one inch taller (shorter) than the average, a daughter is only 0.3182446 inch taller (shorter) than the average. So daughters move or regress towards the mean as well.

6. We find in both regressions that the slopes are less than one. This finding is no coincidence. Recall the formula for the OLS estimate of slope

$$\hat{\beta}_1 = \rho \frac{s_y}{s_x} \tag{1}$$

If the distribution of heights *remains stable across generations*, that is, if $s_y = s_x$ (fathers and sons, or mothers and daughters, have the same standard deviations), then it follows that

$$\hat{\beta}_1 = \rho < 1 \tag{2}$$

since correlation coefficient cannot exceed one.

7. Think about the opposite of RTM: if it did not hold, then sons of tall dads would be even taller, and sons of short dads would be even shorter. That implies a divergence in height that we did not observe in reality.