Galton Regression

Gustavo & Ana

2025-08-28

Reading Galton's Data

```
DATA=read.csv('https://raw.githubusercontent.com/gdlc/STAT_COMP/master/DATA/GALTON.csv',header=T)
```

Adding the parental average (PA)

```
DATA$PA=(DATA$Father+DATA$Mother)/2
```

Regressions

```
fmMother=lm(Height~Mother,data=DATA)
fmFather=lm(Height~Father,data=DATA)
fmPA=lm(Height~PA,data=DATA)
fmPA2=lm(Height~PA+Gender,data=DATA)

summary(fmMother)
```

```
##
## Call:
## lm(formula = Height ~ Mother, data = DATA)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5474 -2.6346 -0.1079  2.8688 11.9526
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 46.69077    3.25874  14.328  < 2e-16 ***
## Mother       0.31318    0.05082   6.163 1.08e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.511 on 896 degrees of freedom
## Multiple R-squared:  0.04066,    Adjusted R-squared:  0.03959
## F-statistic: 37.98 on 1 and 896 DF,  p-value: 1.079e-09
```

```
summary(fmFather)
```

```
##
## Call:
```

```
## lm(formula = Height ~ Father, data = DATA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2683  -2.6689  -0.2092   2.6342  11.9329
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.11039    3.22706  12.120   <2e-16 ***
## Father       0.39938    0.04658   8.574   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.446 on 896 degrees of freedom
## Multiple R-squared:  0.07582,    Adjusted R-squared:  0.07479
## F-statistic: 73.51 on 1 and 896 DF,  p-value: < 2.2e-16
```

```
summary(fmPA)
```

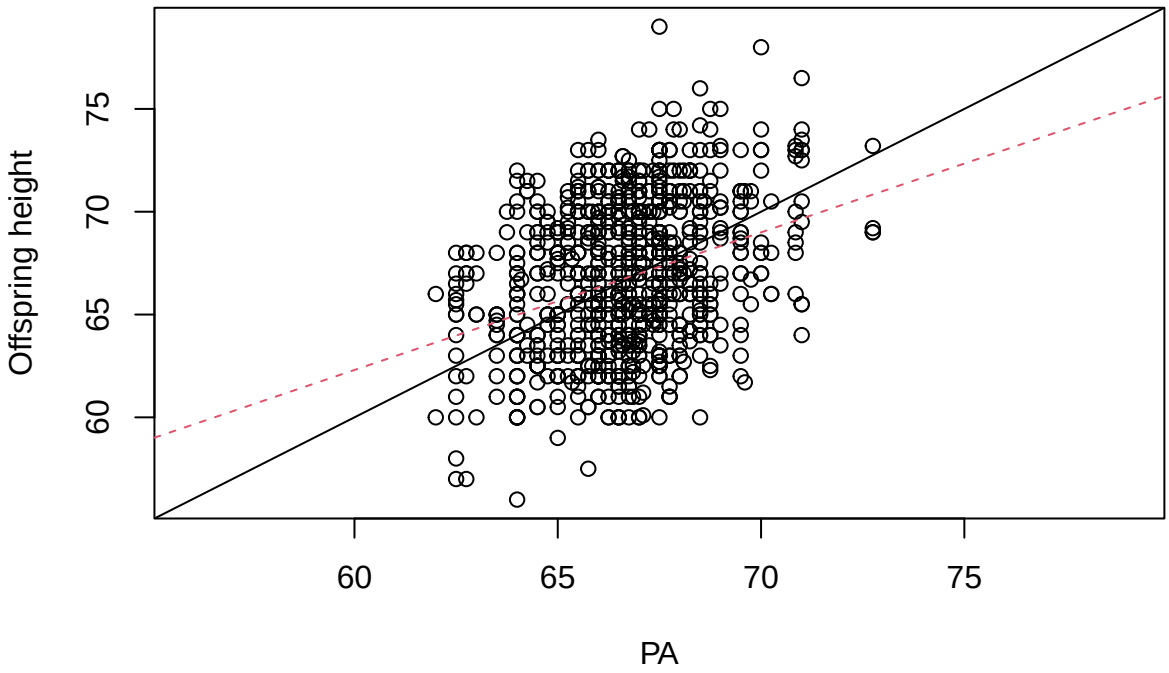
```
##
## Call:
## lm(formula = Height ~ PA, data = DATA)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9814 -2.6604 -0.1642  2.7795 11.6762
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.1488     4.3076   5.142 3.34e-07 ***
## PA            0.6693     0.0646  10.360  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.388 on 896 degrees of freedom
## Multiple R-squared:  0.107,  Adjusted R-squared:  0.106
## F-statistic: 107.3 on 1 and 896 DF,  p-value: < 2.2e-16
```

Beyond the In-class assigment

What follows was not part of the assignement.

Galton's plot

```
tmp=range(c(DATA$Height,DATA$PA))
plot(Height~PA,data=DATA,xlim=tmp,ylim=tmp,xlab='PA',ylab='Offspring height');
abline(0,1);
abline(a =coef(fmPA)[1],b=coef(fmPA)[2],col=2,lty=2)
```



The slope of the regressions and trait heritability

Fisher's infinitesimal model can be described as follows

$$Y_i = G_i + E_i$$

Where Y_i is a phenotype measurement on the ith subject, G_i is a genetic (random) effect (capturing the collective effects of all the loci affecting the trait), and E_i is an environmental effect, capturing all the non-genetic effects.

Because of the Central Limit Theorem, if the number of genes influencing the trait is large, $= G_i$ follows (approximately) a normal distribution.

Using the above model we can decompose the phenotypic variance $V_p = Var(Y_i)$ as follows

$$Var(Y_i) = Var(G_i) + Var(E_i) + 2Cov(G_i, E_i)$$

In absence of genetic-by-environmental covaraince (i.e., if the environmental effects do not depend on indiviudals' genotypes), we have

$$Var(Y_i) = Var(G_i) + Var(E_i)$$

Or

$$V_P = V_G + V_E$$

The trait heritability is the proportion of variance of the phenotypes explained by genetic factors, that ids

$$h^2 = \frac{V_G}{V_G + V_E}$$

Slopes

The slope on a linear regression of Y on X is $b = Cov(X, Y)/Var(X)$.

In the Mother's regression we have

$$b_M = Cov(Y_o, Y_m)/Var(Y_m)$$

For the offspring equation we can write

$$Y_o = (0.5 \times G_f + 0.5 \times G_m + \Psi) + E_o$$

where, G_f and G_m are the *genetic values* of the father and mother, the genetic value of the offspring is

$$G_o = (0.5 \times G_f + 0.5 \times G_m + \Psi)$$

the sum of half of the genetic value of the parents plus a term, Ψ, called mendelian sampling, that captures the randomness resulting from the sampling of alleles at the meiosis.

Using the above, and using the equation for mothers' phenotype, $Y_m = G_m + E_m$, we have

$$Cov(Y_o, Y_m) = 0.25 \times V_g + 0.25 \times V_g = 0.5 \times V_g$$

Therefore, the slope of the regression is

$$b = \frac{0.5 \times V_g}{V_p} = 0.5 \times h^2$$

The heritability of height is ~0.8, if we adjust by sex, the estimated slope is ($b = 0.5 \times h^2$), the estimated slope is

```
fmMother=lm(Height~Mother+Gender,data=DATA)
summary(fmMother)
```

```
##
## Call:
## lm(formula = Height ~ Mother + Gender, data = DATA)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4036 -1.6024  0.1528  1.5890  9.4199
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 41.44952    2.20949   18.76   <2e-16 ***
## Mother       0.35314    0.03439   10.27   <2e-16 ***
## GenderM      5.17669    0.15867   32.62   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.374 on 895 degrees of freedom
## Multiple R-squared:  0.5618, Adjusted R-squared:  0.5608
## F-statistic: 573.7 on 2 and 895 DF,  p-value: < 2.2e-16
```

```
confint(fmMother)
```

```
##                  2.5 %     97.5 %
## (Intercept) 37.1131324 45.7859147
## Mother       0.2856502  0.4206241
## GenderM      4.8652824  5.4881075
```

Under the model we discussed, the slope for Fathers' regression is the same ($b = 0.5 \times h^2$), from the data we have

```
fmMother=lm(Height~Father+Gender,data=DATA)
summary(fmMother)
```

```
##
## Call:
## lm(formula = Height ~ Father + Gender, data = DATA)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3708 -1.4808  0.0192  1.5616  9.4153
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.46113    2.13628   16.13   <2e-16 ***
## Father       0.42782    0.03079   13.90   <2e-16 ***
## GenderM      5.17604    0.15211   34.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.277 on 895 degrees of freedom
## Multiple R-squared:  0.5971, Adjusted R-squared:  0.5962
## F-statistic: 663.2 on 2 and 895 DF,  p-value: < 2.2e-16
```

```
confint(fmMother)
```

```
##                  2.5 %     97.5 %
## (Intercept) 30.2684263 38.6538353
## Father       0.3674023  0.4882411
## GenderM      4.8775165  5.4745684
```

In both cases, 0.4 (1/2 of h^2_{height}) is within the CI.

Question: How do we calculate the slope for the regression involving PA?