Solution of assignment 2, ST2304

August 1, 2016

Problem 1

The data set is loaded and attached.

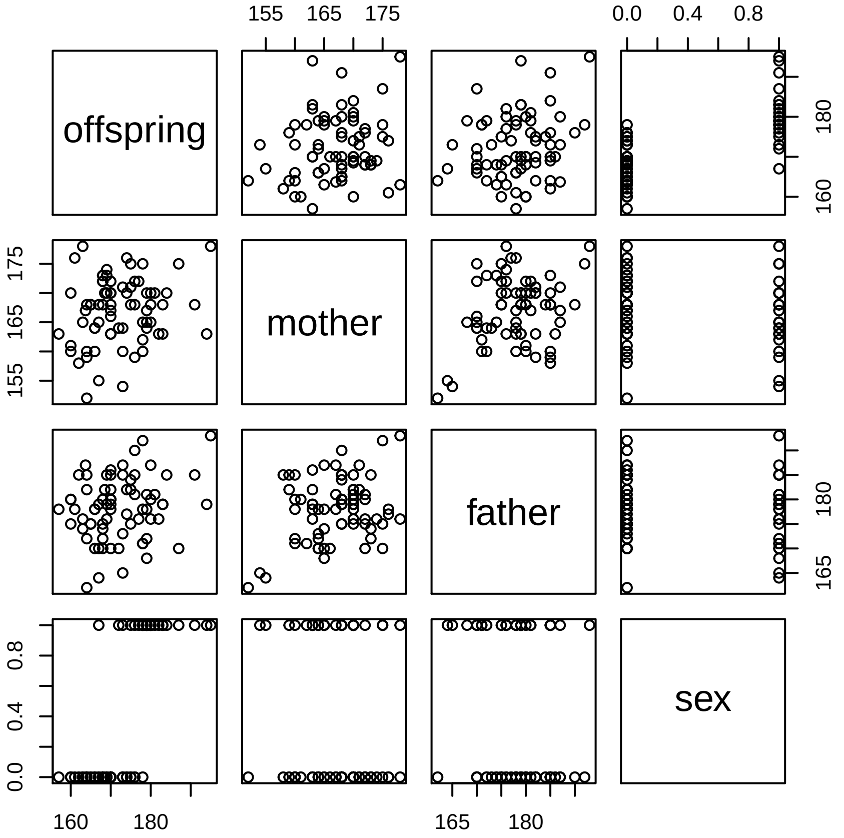
```
> heights <- read.table("https://www.math.ntnu.no/~jarlet/statmod/heights.dat",
+                        header=T)
> attach(heights)
```

There are four variables in the dataset with a total of 72 observations of each.

```
> str(heights)
'data.frame':	72 obs. of  4 variables:
 $ offspring: num  179 179 183 178 194 172 168 173 180 162 ...
 $ mother   : int  164 170 168 162 163 164 172 154 165 158 ...
 $ father   : int  172 178 179 171 179 170 175 165 187 185 ...
 $ sex      : int  1 1 1 1 1 1 0 1 1 0 ...
```

1. A matrix of scatterplots for all combinations of variables is easily made and provides a rapid overview of the data set.

```
> pairs(heights)
```



2. Sexual dimorphism in height is explored using a regression model.

```
> regsex <- lm(offspring˜sex)
> summary(regsex)

Call:
lm(formula = offspring ˜ sex)

Residuals:
     Min       1Q   Median       3Q      Max
-13.4000  -3.5191  -0.1383   2.1234  14.6000

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 167.8766     0.7882 212.983  < 2e-16 ***
sex          12.5234     1.3376   9.362 5.75e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.404 on 70 degrees of freedom
Multiple R-squared:  0.556,Adjusted R-squared:  0.5496
F-statistic: 87.65 on 1 and 70 DF,  p-value: 5.749e-14
```

The regression coeffiecient is the slope of the regression line, which is the difference in mean height between females and males. Sex has an significant effect on height, as the P-value is very small (5.75×10^{-14}). Hence, we can reject the null hypothesis ($H_0 : \beta$=0), and say that males on average are 12.52 cm higher than females.

We may also test for sexual dimorphism by applying a t-test directly. We treat the two variances (in female and male heigths) as being equal (`var.equal=TRUE`) calculating the pooled variance to estimate the variance.

```
> ttest <- t.test(offspring[sex==0],offspring[sex==1],var.equal=TRUE)
> ttest

Two Sample t-test

data:  offspring[sex == 0] and offspring[sex == 1]
t = -9.3623, df = 70, p-value = 5.749e-14
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -15.191256  -9.855553
sample estimates:
mean of x mean of y
 167.8766  180.4000
```

The t-test also reject the null hypothesis of equal means of the heights in the two sexes (H_0:$\mu_{female} - \mu_{male} = 0$) as the P-value is small (5.75×10^{-14}). We see that the difference between the means given in the t-test is equal to the estimate for the slope in the regression ($180.4 - 167.88 = 12.52$). Also, the P-values in the t-test and the regression are the same. The results from the two tests seems to support each other, as the regression clearly state that sex has an effect on height, and the t-test shows that two groups of sex have a significant different means of height.

3. Midtparent values are computed. `R` provides us with multiple ways in which this may be achieved (e.g. `?rowMeans, ?apply`). However, in this case simple vectorized computations are the best option.

```
> midparent<-(mother+father)/2
```

Then the midtparental values can be added to the regression model.

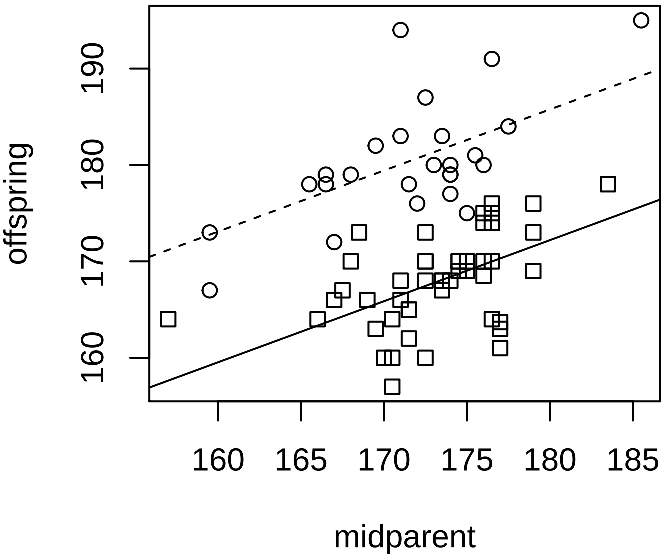
```
> regherit <- lm(offspring~sex + midparent)
> summary(regherit)

Call:
lm(formula = offspring ~ sex + midparent)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-9.3030 -2.5560  0.2545  2.5900 13.9421

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  58.1637    19.3822   3.001  0.00374 **
sex          13.5562     1.1280  12.018  < 2e-16 ***
midparent     0.6336     0.1119   5.664 3.14e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.497 on 69 degrees of freedom
Multiple R-squared:  0.6969,Adjusted R-squared:  0.6881
F-statistic: 79.32 on 2 and 69 DF,  p-value: < 2.2e-16
```

The heritability of height is here found to be 0.63, this seems to be of the same magnitude as found in the literature for stature of humans ($h^2 = 0.65$) (see *iGenetics, A Mendelian Approach* by P.J. Russel 2005). However, the genetic variance is the combined effect of additive genetic, dominance and epistatic effects, so finding the real relationship between additive genetic variance and phenotypic variance may be more complicated.

A scatterplot of student heights against midtparental values is shown below.

```
> plot(x=midparent, y=offspring, pch=sex)
> abline(a=coef(regherit)[1],
+        b=coef(regherit)[3]) # Females
> abline(a=coef(regherit)[1]+coef(regherit)[2],
+        b=coef(regherit)[3], lty=2) # Males
```



4. Including the midparent value in the regression, increased the estimated sex difference in heigth from 12.52 to 13.56, in addition the standard error decreased from 1.34 to 1.13 (midparent explain some of the variation in height). This may be further understood by examining the fitted model. For individuals $i = (1, ..., n)$ the fitted model takes the following form

$$\text{heigth}_i = \alpha + \beta_{\text{sex}}\text{sex}_i + \beta_{\text{midparent}}\text{midparent}_i + \epsilon_i \tag{1}$$

where α is the intercept, the regression coefficients ($\beta_{\text{midparent}}$ and β_{sex}) represent the independent contributions of each explanatory variable (midparent_i and sex_i) to the prediction of the response variable (height_i).

The reference value for sex is 0 (female) in the regression. The estimate of β_{sex} (13.56) is the difference in expected heights between the sexes given equal midparent values. Hence, if we insert our parameter estimates we get the following equations for females (top) and males (bottom)

$$\begin{aligned} heigth &= 58.16 + 13.56 \times 0 + 0.63 \times midparent \\ &= 58.16 + 0.63 \times midparent \\ heigth &= 58.16 + 13.56 \times 1 + 0.63 \times midparent \\ &= 71.72 + 0.63 \times midparent \end{aligned}$$

5. In the following model sex was removed when estimating heritability of height.

```
> regherit_no_sex <- lm(offspring~midparent)
> summary(regherit_no_sex)

Call:
lm(formula = offspring ~ midparent)

Residuals:
    Min      1Q  Median      3Q     Max 
-14.349  -4.889  -1.809   6.203  22.443 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) 100.3828    33.2833   3.016  0.00357 **
midparent     0.4162     0.1928   2.159  0.03425 * 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 7.852 on 70 degrees of freedom
Multiple R-squared:  0.06245,Adjusted R-squared:  0.04906 
F-statistic: 4.663 on 1 and 70 DF,  p-value: 0.03425 
```

Removing sex as a explanatory variable in the regression changed the estimate of heritability to 0.42. Hence, by not including sex as a explanatory variable a lower (downwards biased) estimate of heritability may result.

Problem 2

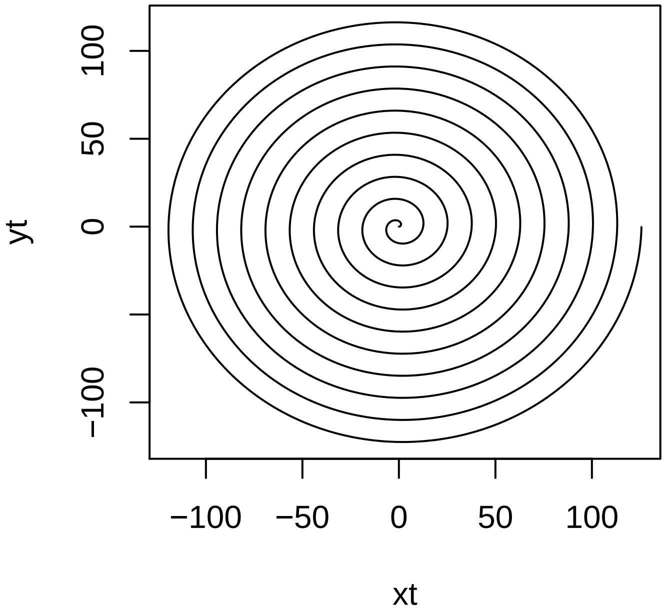
Spiral. a is a constant, we choose a=2.

$$x(t) = 2t\texttt{cos(t)} \tag{2}$$

$$y(t) = 2t\texttt{sin(t)} \tag{3}$$

Now we generate a sequence of t values sufficiently densely spaced, calculate *x(t)* and *y(t)* and plot these.

```
> t <- seq(from=0, to=10*2*pi, by=0.01)
> xt <- 2*t*cos(t)
> yt <- 2*t*sin(t)
> plot(x=xt, y=yt, type="l")
```



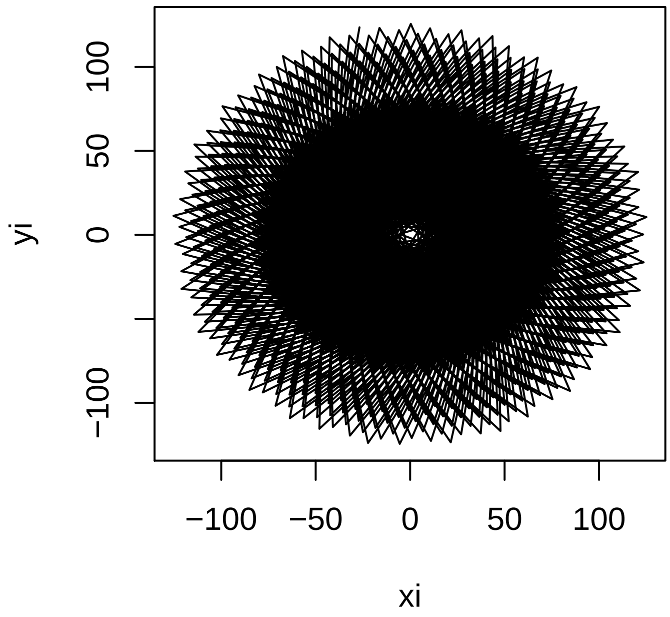
We see that $\cos(t)$ gives the angle direction of the spiral for each t and at is the distance from the centre, giving the x coordinates ($t\cos(t)$) and the y coordinates ($t\sin(t)$).

Sunflower. We recognize $r(i)$ and $\theta(i)$ as polar coordinates where $\theta(i)$ is the angle direction from the centre of seed number i, $\theta(i) = \pi(3-\sqrt{5})i$, and $r(i)$ is the distance from the centre of seed number i, $r(i) = a\sqrt{i}$. Following standard procedures these may be transformed into the Cartesian coordinates $x(i)$ and $y(i)$

$$\begin{aligned} x(i) &= r(i)\cos(\theta(i)) \\ &= a\sqrt{(}i)\cos(\pi(3-\sqrt{5})i) \\ y(i) &= r(i)\sin(\theta(i)) \\ &= a\sqrt{(}i)\sin(\pi(3-\sqrt{5})i) \end{aligned}$$

We choose a=4 and makes a sequence $\mathbf{i}$ = (1, ..., n), with n number of seeds. We can see that we can plot interesting things in `R`!

```
> a <- 4
> i <- seq(1:1000)
> theta <- pi*(3-sqrt(5))*i
> r <- a*sqrt(i)
> xi <- r*sin(theta)
> yi <- r*cos(theta)
> plot(x=xi, y=yi, type="l")
```


yi
100
50
0
−100
−100
−50
0
50
100
xi