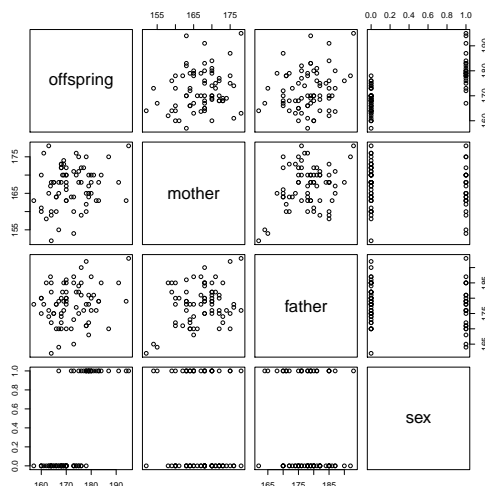# Solution of assignment 2, ST2304

Problem 1

The regression coeffiecient is the slope of the regression line, which is the difference in mean height between females and males.

```
 > 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
```

Sex has an significant effect on height, as the P-value (5.75e-14***) is small. The estimate of the regression coefficient (12.5234) is therefore significantly different from zero. Here we can reject the null hypothesis ($\beta=0$), and say that height has a significant effect on sex.

We treat the two variances (in female and male heigths) as being equal, setting *var.equal=TRUE* calculating the pooled variance to estimate the variance.

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

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 (5.749e-14). We see that the difference between the means given in the t-test (180.4-167.88) is equal to the estimate for the slope in the regression (12.52). Also, the p-values in the t-test and the regression are the same (which makes sense).

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 sexhave a significant different means of height.

```
> 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 heigth is here found to be 0.6336, this seems to be about the same found in the literature for stature of humans ($h^2 = 0.65$) (iGenetics, A Mendelian Approach by P.J. Russel 2005). (However, the genetic variance is the combined effect of additive genetic effects, dominance and epistasis effects, and environment, so finding the real relationship between additive genetic variance and phenotypic variance may be more complicated).

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

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

midparent$_i$ is the vector of midparentvalues, $i = 1, ...n$ and sex$_i$ is the vector of sexvalues (0 and 1). The reference value is the 0 (female) in the regression. The value of the sex estimate

(sex,13.5562) in the *summary*() is the difference in expected heigths between sexes given equal midparent values.

Setting in for

$$heigth = 58.1637 + 0.6336 * midparent + 0 \qquad (2)$$
$$heigth = 58.1637 + 0.6336 * midparent + 13.5562 * 1 \qquad (3)$$
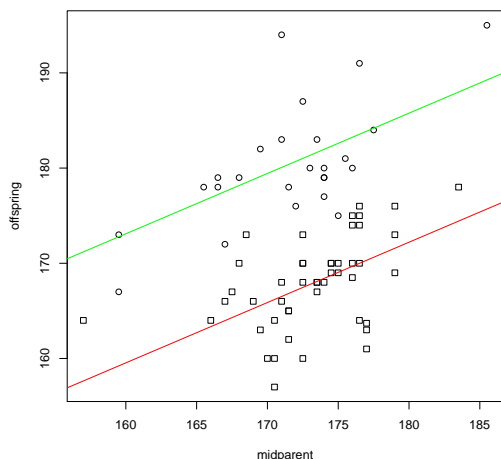
Including the midparent value in the regression, increased the estimated sex difference in heigth from 12.52 to 13.56, in addition the standard error decreased from 1.34 to 1.13 (midparent explain more of the height).

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

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 the sex as a explanatory variable in the regression changes the estimate of heritability to 0.4162. Not including sex as a explanatory variable can result in a lower estimate of heritability.

R-code

```
heights <- read.table("http://www.math.ntnu.no/~jarlet/statmod/heights.dat")
attach(heights)
#Make scatter plots of all variable combinations
pairs(heights)

#Estimate the difference in height between the sexes
regsex<-lm(offsring~sex)
summary(regsex)

# two-sample t-test of difference between the sexes
t.test(offspring[sex==0],offspring[sex==1],var.equal=TRUE)

#Compute the midparental value
 midparent<-rowMeans(heights[,2:3])
#or
 midparent<-(mother+father)/2

# add midparental as a second explanatory variable
regherit<-lm(offspring~sex+midparent)
summary(regherit)


 #scatter plot of the heights of the students versus their midparental values
  plot(midparent,offspring,pch=sex)

#lines representing the estimated expected response as function of midparental value for
abline(58.1637,0.6336,col="red")
abline(58.1637+ 13.5562,0.6336,col="green")


#removing \sex as a parameter in the regression
 regherit<-lm(offspring~midparent)
 summary(regherit)
```
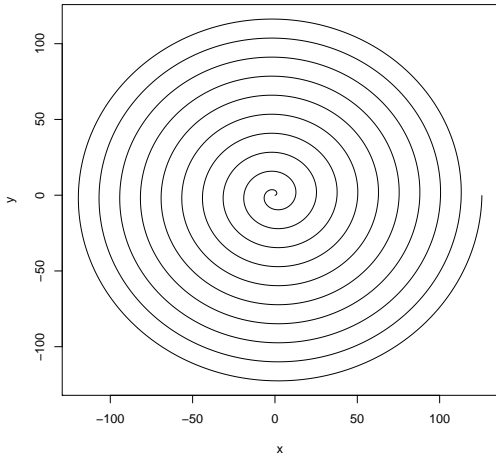
Problem 2 a is a constant, we choose $a=2$.

$$x(t) = 2t\cos(\texttt{t}) \tag{4}$$
$$y(t) = 2t\sin(\texttt{t}) \tag{5}$$

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, $t\texttt{cos}(\texttt{t})$ and the y coordinates, $t\texttt{sin}(\texttt{t})$.

$\theta(i)$ is the angle direction from the centre of seed number i, $\theta(i) = \pi(3 - \sqrt{5})i$. $r(i)$ is the distance from the centre of seed number i, $r(i) = a\sqrt{i}$.
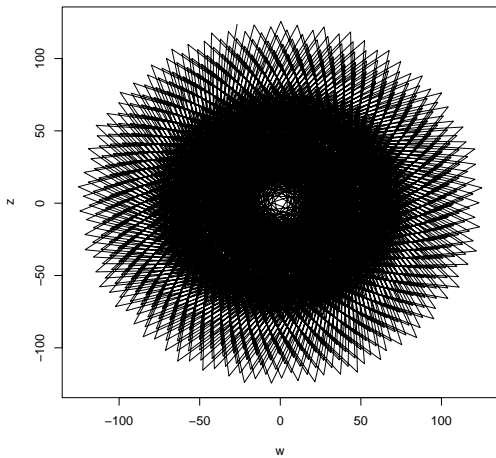
$$x(i) = r(i)\texttt{cos}(\theta(\texttt{i})) \tag{6}$$
$$y(i) = r(i)\texttt{sin}(\theta(\texttt{i})) \tag{7}$$

$$x(i) = a\sqrt{(i)}\texttt{cos}(\pi(3 - \sqrt{5})\texttt{i}) \tag{8}$$
$$y(i) = a\sqrt{(i)}\texttt{sin}(\pi(3 - \sqrt{5})\texttt{i}) \tag{9}$$

We choose a=4 and makes a sequence with n number of seeds, i=1,...,n. We can see that we can plot interesting things in R !!!

R code

```
#make a sequence of t values
   t<-seq(0,10*2*pi,by=0.01)
# make the x and y functions
```

5

```
 x=2*t*cos(t)
 y=2*t*sin(t)

#plot the vectors x,y
 plot(x,y,type="l")

#plot the sunflower
 a=4
    i<-seq(1:1000)
 theta=pi*(3-sqrt(5))*i
 r=a*sqrt(i)
 w=r*sin(theta)
 z=r*cos(theta)
 plot(w,z,type="l")
```