# Løsningsforslag øving 9, ST1301

**Oppgave 1** *Regresjon. Estimering av arvbarhet.*

**a)** Legg inn din egen høyde, din mors høyde, din fars høyde, og ditt kjønn via linken på fagets hjemmeside[1]. Last så ned dataene fra samme link og og les de inn i R ved bruk av `read.table` (bruk tilleggsargumentet `header=T` slik at kolonnene får navn). Lagre resultatet i en data.frame med navn `data`. Skriv så `attach(data)` for å gjøre variablene i data.framen direkte tilgjengelig tilgjengelige uten bruk av `$` notasjon.

Beregn såkalt "midparantal values", d.v.s., gjennomsnittlig høyde til hvert foreldrepar og lagre resultatet i en variabel med navn `midparent`.

Lag et spredningsplot av variablen `avkom` mot variabelen `midparent`. Velg forskjellig symboler for hvert kjønn i spredningsplottet ved å oppgi en vektor med elementer lik 1 og 2 tilsvarende hvert kjønn som tilleggsargumentet `pch`. Denne vektoren kan lages ved å konverte variabelen `kjonn` (en faktor) til en vektor av heltall med funksjonen `as.integer`. Argumentet `pch` er beskrevet på hjelpesiden til `par`.

Tilpass så en lineær modell ved hjelp av `lm` hvor du bruker høyde til avkom som responsvariabel og `midparent` og `kjonn` som forklaringsvariabler. Undersøk den tilpassede modellen med `summary`, `confint`, og eventuelt `dummy.coef`.

Hva blir estimatet av stigningstallet i modellen? Er regresjonskoeffisienten (stigningstallet) for variabelen `midparent` signifikant forskjellig fra 0?

Det kan virke rimelig at forventet høyde til avkom av et gitt foreldrepar er lik snittet av høyden til hver forelder? Dette vil i tilfelle innebære at regresjonskoeffisienten er lik 1. Utfør en test av denne hypotesen. (Hint: ta utgangspunktet i estimatet av koeffisienten og standardfeilen i utskriften du får når du bruker `summary` og finne $p$-verdien for testen med `pt`.)

Hva blir estimatet av og konfindensintervallet for gjennomsnittlig høydeforskjell mellom kjønnene? Er høydenforskjellen statistisk signifikant? Inneholder konfidensintervallet tallet 0?

Legg til to regresjonslinjer i spredningsplottet over som svarer til forventet høyde som funksjon av midparental value for hvert kjønn ved hjelp to kall til `abline`.

```
> data<-read.table("/home/jarlet/www/ST1301-2004v/hoyde.dat",
        header=T,as.is=5)
```

[1] `http://www.math.ntnu.no/~jarlet/ST1301-2004v/hoyde.imf?forste=1`

Leser vi inn 5'te kolonne "as is" forblir denne kolonnen av datatypen tekststreng og ikke faktor (det vil være mest praktisk, se nedenfor). Argumentet `colTypes` kan også brukes for å spesifisere valg av datatype for de ulike kolonnene i datafilen.

```
> attach(data)
> midparent <- .5*(mor+far)
```

Faktoren kjønn må konverteres til heltall for å spesifisere hvilke symbol vi vil ha i plottet.

```
> plot(midparent,avkom,pch=as.integer(kjonn))
> modell <- lm(avkom ~ midparent + kjonn)
> summary(modell)

Call:
lm(formula = avkom ~ midparent + kjonn)

Residuals:
     Min       1Q   Median       3Q      Max
-7.11798 -2.27418 -0.01789  2.55676 13.79089

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  86.3701    25.1371   3.436  0.00161 **
midparent     0.4772     0.1452   3.287  0.00241 **
kjonnM       12.2350     1.4799   8.267 1.51e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.323 on 33 degrees of freedom
Multiple R-Squared: 0.6869,     Adjusted R-squared: 0.6679
F-statistic:  36.2 on 2 and 33 DF,  p-value: 4.769e-09
```

Modellen vi har tilpasset kan skrives:

$$y = a + bx + c_i + \epsilon, \tag{1}$$

hvor $y$ er høyde på avkom, $x$ er gjennomsnittlig foreldrehøyde, $a$ er modellens intercept, $b$ er stigningstallet, og $c_1$ og $c_2$ er parameterne som spesifisere forskjellen i høyde mellom gruppe 1 og 2 (kvinner og menn). Fordi $c_1$ og $c_2$ ikke kan estimeres samtidig med $a$ innføres bibetingelsen $c_1 = 0$.

Merk at tolkningen av $c_2$ dermed blir forskjell i forventet høyde mellom kjønnene gitt ellers like betingelser (samme foreldresnitt).

Tabellen i utskriften over gir oss estimat av parameterne $a$, $b$, og $c_2$ med tilhørende standardfeil, testobservatorer og $p$-verdier for tester av nullhypotesene $a = 0$ (intercepten lik 0), $b = 0$ (ingen sammenheng med foreldresnittet), $c_2 = 0$ (ingen forskjell mellom kjønnene). Fra $p$-verdiene ser vi at alle disse nullhypotesene kan forkastes.

Parameteren $b$, stigningstallet i regresjonen, kalles arvbarhet eller arvegrad (heritability på engelsk) og er en viktig parameter i kvantitativ genetikk. Denne parameteren kan tolkes som andelen av den totale variansen i variabelen høyde som skyldes additive geneffekter. Denne skal derfor alltid ligge mellom 0 og 1. I vårt tilfelle er estimatet lik 0.48 som altså betyr at 48% av variasjonen i høyde (om vi ser bort fra kjønnsforskjeller) er genetisk betinget.

```
> confint(modell)
                 2.5 %      97.5 %
(Intercept) 35.2283472 137.5119447
midparent    0.1818073   0.7726254
kjonnM       9.2240374  15.2459120
```

Konfidensintervallene inneholder ikke null som stemmer med det faktum at alle parameterne er signifikant forskjellig fra 0.

`summary` gir oss ikke testresultatet for mer "sære" nullhypoteser som at $b = b_0 = 1$. For å teste denne hypotesen beregner vi $(\hat{b} - b_0)/s.e.(\hat{b})$ på grunnlag av tallene i tabellen over og finner signifikansverdien ved oppslag i kumulative $t$-fordeling. Husk også at vi har en to-sidig test:

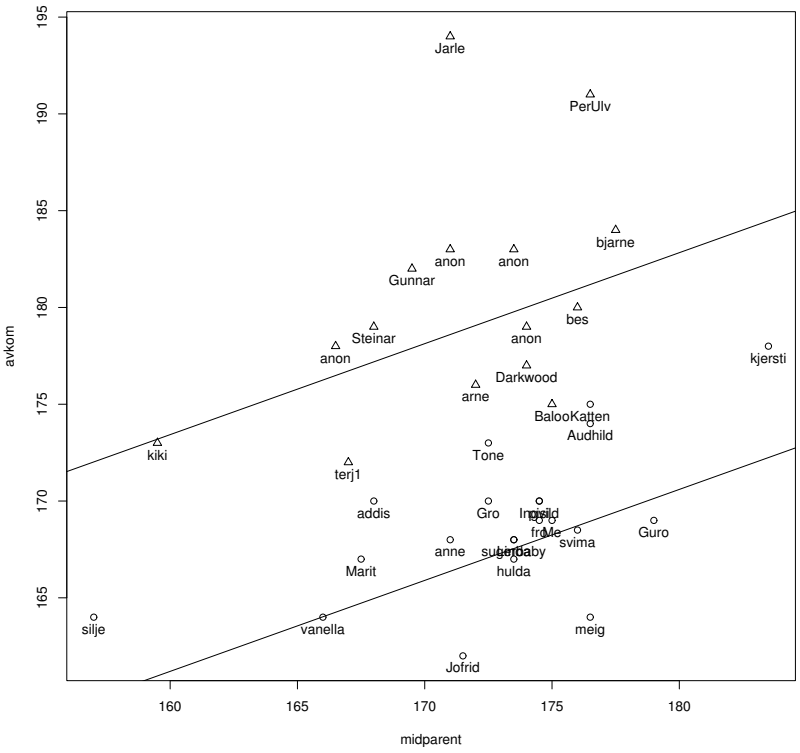
```
> 2*pt(q=(.4772-1)/0.1452,df=33)
[1] 0.001028957
>
```

Skal vi legge inn regresjonslinjer for kjønn 1 og 2 setter vi $i = 1$ og $i = 2$ i (1) og setter inn parameterestimatene slik at vi får ligningene

$$\begin{aligned} y &= 86.3 + 0 + 0.47x \\ y &= 86.3 + 12.2 + 0.47x \end{aligned} \tag{2}$$

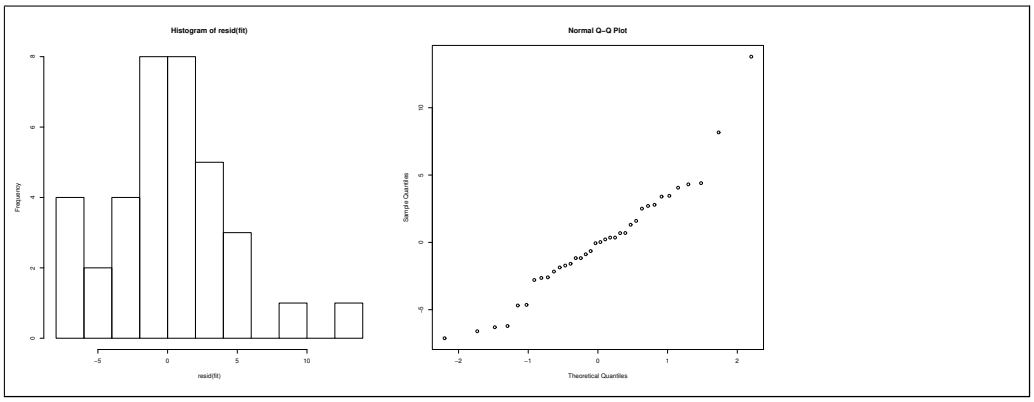
```
> abline(86,.47)
> abline(86+12.23,.47)
> text(midparent,avkom,nick,pos=1)
```

Kommandoen `text` legger til tekststrenger på koordinatene spesifisert av de to første argumentene. Uten argumentet `as.is=5` ved innlesing av dataene måtte vi her ha konvertert variabelen `nick` tilbake til datatypen tekststreng ved bruk av `as.character`.




**b)** Hent ut residualene fra den tilpassede modellen og undersøk om disse er normalfordelte ved å lage et histogram av disse. Lag også et Q-Q-plot med `qqnorm`.

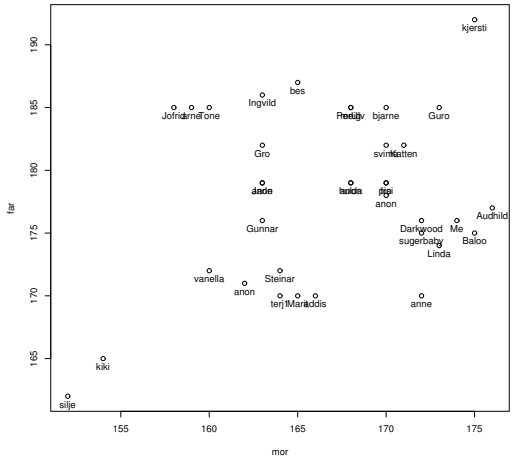
```
> hist(resid(modell),breaks=12)
> qqnorm(resid(modell))
```




**c)** I populasjonsgenetisk teori antas det ofte såkalt tilfeldig parring ("random mating") som betyr at egenskapene til et tilfeldig valgt foreldrepar er uavhengige tilfeldige variable. Dersom individene i en populasjon foretrekker partnere som er lik seg selv vil dette føre til at egenskapene blir positivt korrelerte. Dette kalles "assortative mating" og vil kunne ha evolusjonære konsekvenser, bl.a. vil det føre til en økning i den genetiske variansen i egenskapen som studeres (her høyde).

Undersøk om det er noen korrelasjon mellom høyden til fordeldrene ved å bruke funksjonen `cor.test`. Lag også et spredningsplot av høyde til mor versus høyde til far.

```
> plot(mor,far)
> text(mor,far,nick,pos=1)
```



```
> cor.test(mor,far,altern="greater")

        Pearson's product-moment correlation

data:  mor and far
t = 1.774, df = 34, p-value = 0.04251
alternative hypothesis: true correlation is greater than 0
95 percent confidence interval:
 0.01339168 1.00000000
sample estimates:
      cor
0.2910607
```

Gjør vi en en-sidig test med $H_1 : \rho > 0$ ser vi at vi kan forkaste $H_0 : \rho = 0$. Det er med andre ord slik at de høye foretrekker de høye og motsatt i populasjonen vi har trukket utvalget fra.