Exercise 3 - solution

Paul Blanche

Question 1

We load the data and have a look a the first lines.

```
load(url("http://paulblanche.com/files/SCD.rda"))
d <- SCD # shorter name, just for convenience
head(d)
```

```
##   age sex SCD Pdias Psys height weight pulse  HCT Creat  Hb
## 1  26   2   1    62  120    164   57.0    68 21.5    37 7.7
## 2  33   2   1    60  115    165   58.0    64 21.6    55 7.8
## 3  43   2   1    66  117    152   57.2    67 29.3    68 9.2
## 4  23   1   1    67  129    177   61.0    62 25.0    66 8.9
## 5  37   1   1    47  106    167   50.0    65 23.1    66 8.0
## 6  46   2   1    65  134    163   68.0    67 21.0    69 7.7
```

We add the new variable MAP to he data and have a look a the first lines again, just to check.

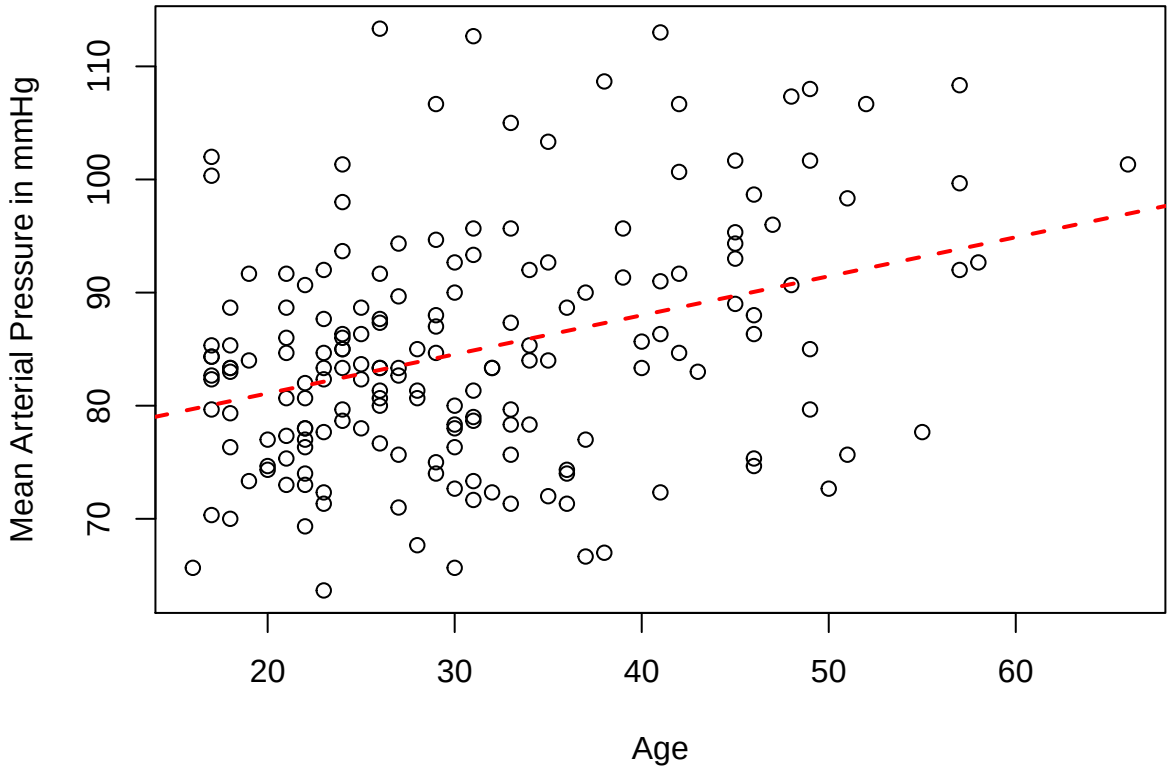
```
d$MAP <- d$Pdias + (1/3)*(d$Psys-d$Pdias)
head(SCD)
```

```
##   age sex SCD Pdias Psys height weight pulse  HCT Creat  Hb
## 1  26   2   1    62  120    164   57.0    68 21.5    37 7.7
## 2  33   2   1    60  115    165   58.0    64 21.6    55 7.8
## 3  43   2   1    66  117    152   57.2    67 29.3    68 9.2
## 4  23   1   1    67  129    177   61.0    62 25.0    66 8.9
## 5  37   1   1    47  106    167   50.0    65 23.1    66 8.0
## 6  46   2   1    65  134    163   68.0    67 21.0    69 7.7
```

Question 2

We make a scatter plot to visualize MAP versus age and add the fitted regression line.

```r
plot(MAP~age,
     ylab="Mean Arterial Pressure in mmHg",
     xlab="Age",
     data=d)
lm1 <- lm(MAP~age,data=d)
abline(lm1,col="red",lty=2,lwd=2)
```



Question 3

```r
summary(lm1)
```

```
##
## Call:
## lm(formula = MAP ~ age, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.3019  -6.2705  -0.4353   5.4795  30.1726
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 74.18829    2.29979  32.259  < 2e-16 ***
```

```
## age          0.34509    0.07054   4.892 2.26e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.864 on 174 degrees of freedom
## Multiple R-squared:  0.1209, Adjusted R-squared:  0.1159
## F-statistic: 23.93 on 1 and 174 DF,  p-value: 2.258e-06
```

```
confint(lm1)
```

```
##                  2.5 %     97.5 %
## (Intercept) 69.6492188 78.7273587
## age          0.2058731  0.4843152
```

The slope estimate of 0.345 means that, when comparing two persons, one being one year older than the other, we estimate that the older person has on average a MAP 0.345 mmHg higher than the younger person. This difference is significantly different from 0, with a p-value <0.001. A 95% confidence interval is (0.206,0.484). For this model the slope does not have a meaningful interpretation: the average arterial pressure (MAP) is not very well defined for someone age 0 year and anyway we do not observe data for subjects younger than 16 and we do not like to extrapolate that much.

Question 4

```
d$Zage <- (d$age - 30)/10
lm2 <- lm(MAP~Zage,data=d)
summary(lm2)
```

```
##
## Call:
## lm(formula = MAP ~ Zage, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.3019  -6.2705  -0.4353   5.4795  30.1726
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  84.5411     0.7460 113.329  < 2e-16 ***
## Zage          3.4509     0.7054   4.892 2.26e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.864 on 174 degrees of freedom
## Multiple R-squared:  0.1209, Adjusted R-squared:  0.1159
```

```
## F-statistic: 23.93 on 1 and 174 DF,  p-value: 2.258e-06
```

```
confint(lm2)
```

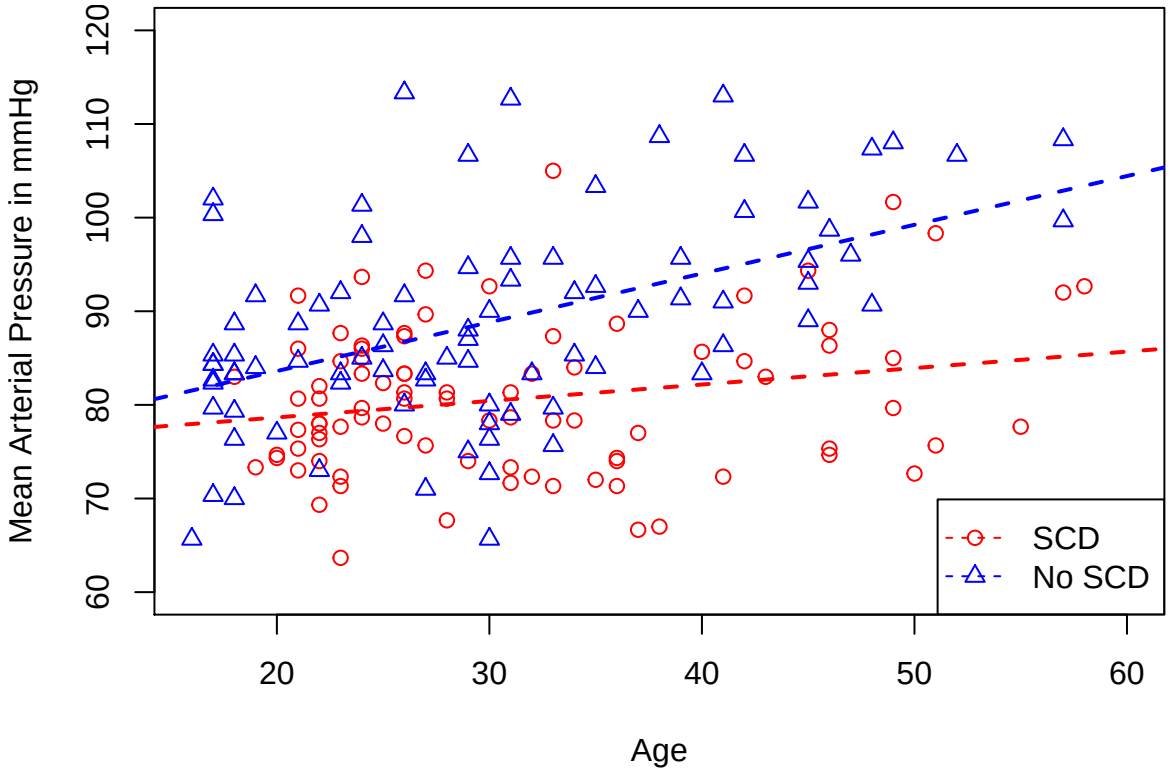
```
##                 2.5 %    97.5 %
## (Intercept) 83.068783 86.013443
## Zage         2.058731  4.843152
```

The interpretation of the slope is maybe more interesting. Now the interpretation is: when comparing two persons, one being 10 years older than the other, the older person has on average a MAP 3.45 mmHg higher than the younger person. One year difference is not much and the two persons are too similar (with respect to age) for us to see a **clinically** significant difference in average MAP, but with a ten year difference the results become clearer. Note that the range of ages that we observe is 16-66, so we do not extrapolate as long as we compare two ages ten years different that are both within this range. Note also that, unsurprisingly 10 times the previous estimate of the slope matches the new estimate. This is because the two linear models are mathematically identical, only the results are expressed using two different "units". This explains that the p-values for the slopes remain identical. The intercept (est.=84.5 mmHg, CI= 83.1-86.0) now has a meaningful interpretation: this is the estimated average MAP for a 30 year old person. Note that the (meaningless) p-value for the intercept has changed, which is not surprising. For the first model the (meaningless) null hypothesis was H0: "the MAP at age 0 is in average 0", whereas for the new model it is H0: "the MAP at age 30 is in average 0".

Question 5 and 6

```
# First plot red dots
# Trick: we use only the rows in the data SCD which
# correspond to SCD=1.
plot(MAP~age,
     data=d[d$SCD==1,],
     col="red",
     ylab="Mean Arterial Pressure in mmHg",
     xlab="Age",
     pch=1,           # we ask the dots on the plot to be circles
     ylim=c(60,120), # we set the range of the y-axis
     xlim=c(16,60))  # we set the range of the x-axis
# Second, we add the blue dots. Note that we now use only
# the rows in the data SCD which correspond to SCD=0.
points(d[d$SCD==0,"age"],d[d$SCD==0,"MAP"],
       col="blue",
       pch=2)  # we ask the dots to be triangles
# We fit the two linear models, one for each SCD yes/no group
lm11 <- lm(MAP~age,data=d[d$SCD==1,])
lm10 <- lm(MAP~age,data=d[d$SCD==0,])
```

```r
# We add the fitted lines for each group.
# Note that lty=2 is to draw dashed lines,
# The lwd option controls the thickness of the line.
abline(lm11,col="red",lty=2,lwd=2)
abline(lm10,col="blue",lty=2,lwd=2)
# We add a legend
legend("bottomright",c("SCD","No SCD"),pch=1:2,col=c("red","blue"),lty=2)
```

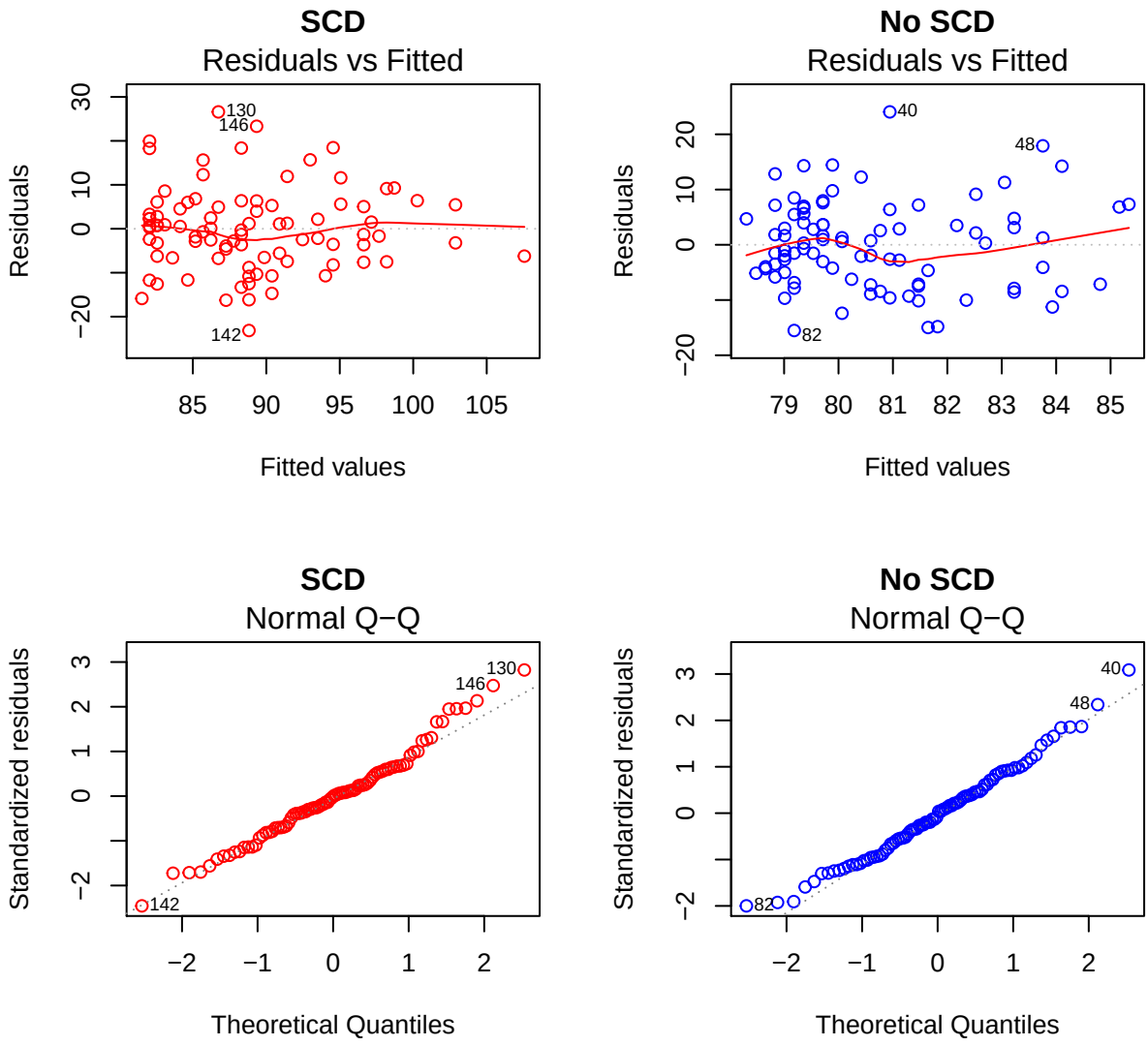


We can see that the association between age and MAP is estimated stronger for subjects without SCD (it increases more with age than for subjects with no SCD).

Note: the appendix contains alternative R code using `ggplot` and `dplyr` packages for creating the same plot.

Question 7

```r
par(mfcol=c(2,2)) # to split the plot area into 4 areas
plot(lm10,which=c(1,2),ask=FALSE,main="SCD",col="red")
plot(lm11,which=c(1,2),ask=FALSE,main="No SCD",col="blue")
```



Everything seems fine. For both models the residuals (top plots) seem randomly and symmetrically scattered around the zero-line with the same variability across the range of fitted values. "Symmetrically scattered around the zero-line" suggests that the mean of the residuals is zero for any fitted value (i.e. for any age) as it should. "Same variability across the range of fitted values (i.e. for all ages)" suggests that the standard deviation of the error term is the same for all fitted values/ages as it should.

The QQ-plots looks fine too. The dots are almost on the diagonal and the deviations do not seem to be larger that those we can reasonably expect from random variation with a moderate sample size (n=88).

Question 8

```
round(confint(lm11)[2,],3)
```

```
##  2.5 % 97.5 %
```

```
##  0.011  0.341
```

```
round(confint(lm10)[2,],3)
```

```
##  2.5 % 97.5 %
##  0.338  0.704
```

Here we are in the unfortunate situation in which we cannot easily conclude whether there is a significant differentrece between the two slopes. This is because the 95% confidence intervals overlap just a little bit, i.e. they overlap but the estimate of one group is not included in the 95% CI of the other group (see slides of course day 1).

Question 9

```
# compute the difference
DiffInSlope <- coef(lm11)[2]-coef(lm10)[2]
DiffInSlope
```

```
##        age
## -0.3448714
```

```
# compute the estimated standard error of the difference
seDiffInSlope <- sqrt(summary(lm11)$coef[2,2]^2 + summary(lm10)$coef[2,2]^2)
z <- DiffInSlope/seDiffInSlope
z
```

```
##       age
## -2.781157
```

```
# additional: compute a 2-sided p-value
2*(1-pnorm(abs(z)))
```

```
##         age
## 0.005416561
```

Yes, we can reject the null hypothesis H0:"the two slopes are equal" because |z|=2.78>1.96. Here the additional computation of the p-value gives P=0.005<5%.

Question 10

We first compute the 95% prediction intervals of the mean diastolic pressure given age in the two populations of subjects with and without SCD.

```
# 1. Create dataset with all possible ages (16 to 66)
d.new <- data.frame(age=seq(16,66,by=1))
# 2. Create dataset containing all the ages, predicted values and
#   limits of the prediction intervals for subjects with SCD.
```

```r
d.pi.SCD <- cbind(d.new, predict(lm11, d.new, interval='pred'))
# 3. Visualize the results
head(d.pi.SCD)
```

```
##   age      fit      lwr      upr
## 1  16 77.95582 62.08029 93.83135
## 2  17 78.13155 62.28151 93.98159
## 3  18 78.30728 62.48106 94.13351
## 4  19 78.48302 62.67891 94.28712
## 5  20 78.65875 62.87506 94.44243
## 6  21 78.83448 63.06951 94.59945
```

```r
# 3. Same for subjects without SCD.
d.pi.noSCD <- cbind(d.new, predict(lm10, d.new, interval='pred'))
head(d.pi.noSCD)
```

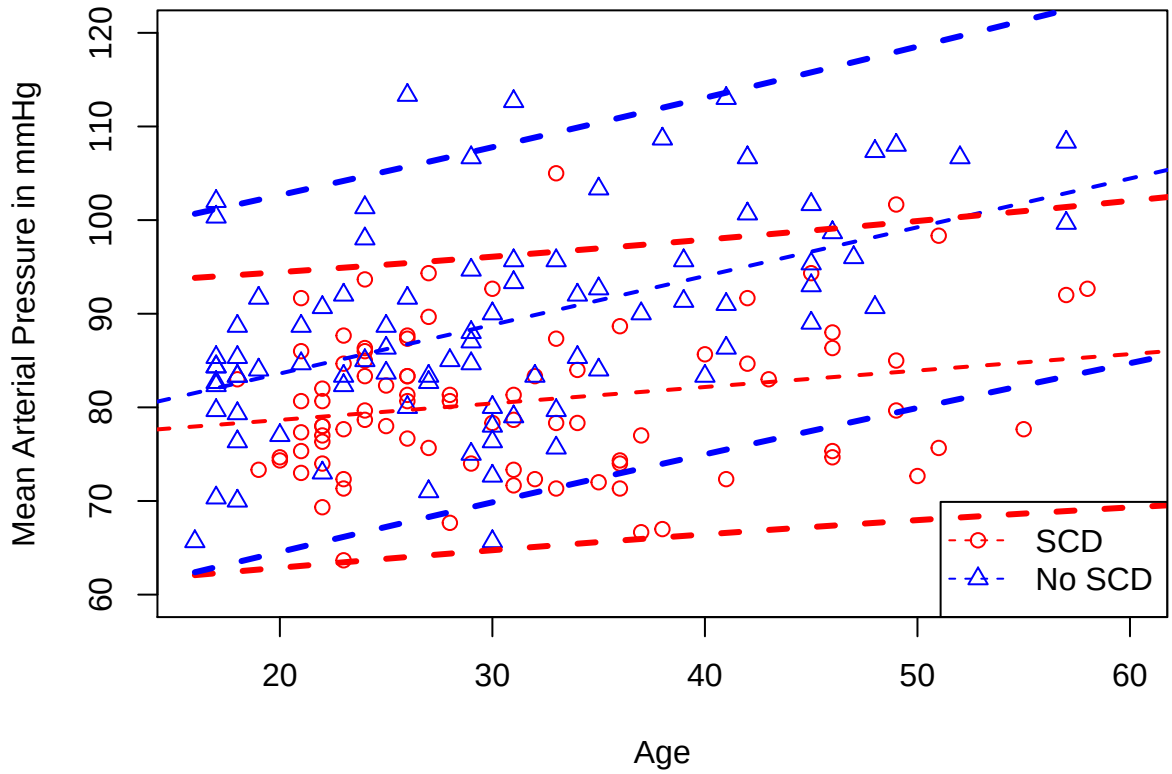
```
##   age      fit      lwr      upr
## 1  16 81.53312 62.39049 100.6758
## 2  17 82.05373 62.93540 101.1721
## 3  18 82.57433 63.47858 101.6701
## 4  19 83.09494 64.02003 102.1698
## 5  20 83.61554 64.55975 102.6713
## 6  21 84.13614 65.09773 103.1746
```

Note: the appendix contains alternative R code using `ggplot` and `dplyr` packages for creating the same plot.

We are now ready to produce the plot.

```r
# We first produce the same plot as before. Then we will add the prediction intervals.
par(mfcol=c(1,1)) # restore no split for the plot area
# First plot red dots, using only the rows in the data SCD which
# correspond to SCD=1.
plot(MAP~age,
     data=d[d$SCD==1,],
     col="red",
     ylab="Mean Arterial Pressure in mmHg",
     xlab="Age",
     pch=1,          # we ask the dots on the plot to be circles
     ylim=c(60,120), # we set the range of the y-axis
     xlim=c(16,60))  # we set the range of the x-axis
# Second, we add the blue dots. Note that we now use only
# the rows in the data SCD which correspond to SCD=0.
points(d[d$SCD==0,"age"],d[d$SCD==0,"MAP"],
       col="blue",
       pch=2)  # we ask the dots to be triangles
# We fit the two linear models, one for each SCD yes/no group
```

```r
lm11 <- lm(MAP~age,data=d[d$SCD==1,])
lm10 <- lm(MAP~age,data=d[d$SCD==0,])
# We add the fitted lines for each group.
# Note that lty=2 is to draw dashed lines,
# The lwd option controls the thickness of the line.
abline(lm11,col="red",lty=2,lwd=2)
abline(lm10,col="blue",lty=2,lwd=2)
# Add limits of the 95% prediction intervals
lines(d.pi.SCD$age, d.pi.SCD$upr, lty=2, lwd=3,col="red") # upper limit, SCD=1.
lines(d.pi.SCD$age, d.pi.SCD$lwr, lty=2, lwd=3,col="red") # lower limit, SCD=1.
lines(d.pi.noSCD$age, d.pi.noSCD$upr, lty=2, lwd=3,col="blue") # upper limit, SCD=0.
lines(d.pi.noSCD$age, d.pi.noSCD$lwr, lty=2, lwd=3,col="blue") # lower limit, SCD=0.
# We add a legend
legend("bottomright",c("SCD","No SCD"),pch=1:2,col=c("red","blue"),lty=2)
```



We describe the distribution of the age of subjects with and without SCD using median, first and third quartiles.

```r
summary(d$age[d$SCD==1])
```

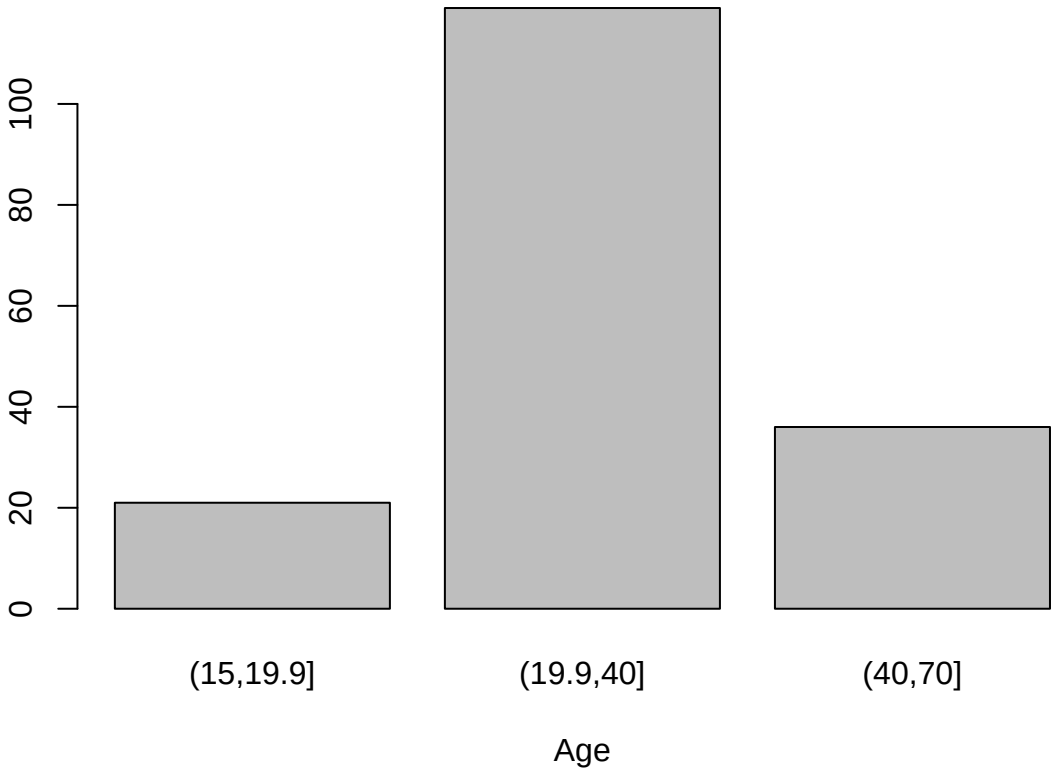
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   23.00   28.00   31.32   36.25   58.00
```

```r
summary(d$age[d$SCD==0])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   16.00   21.75   29.00   30.39   37.25   66.00
```

We note that "most" of the observations are from subjects from 20 to 40 years (in both groups). We have limited data about subjects outside this range, hence it would be wise to not overinterpret the results outside this range. The results are essentially driven and mostly appropriate for the age range 20 to 40. Due to limited data outside this range, it is not really possible to carefully check whether the modeling assumptions seem reasonable also outside the range 20-40.

```
barplot(table(cut(d$age,breaks=c(15,19.9,40,70))),xlab="Age")
```



Appendix

ggplot Q5 & Q6

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
```

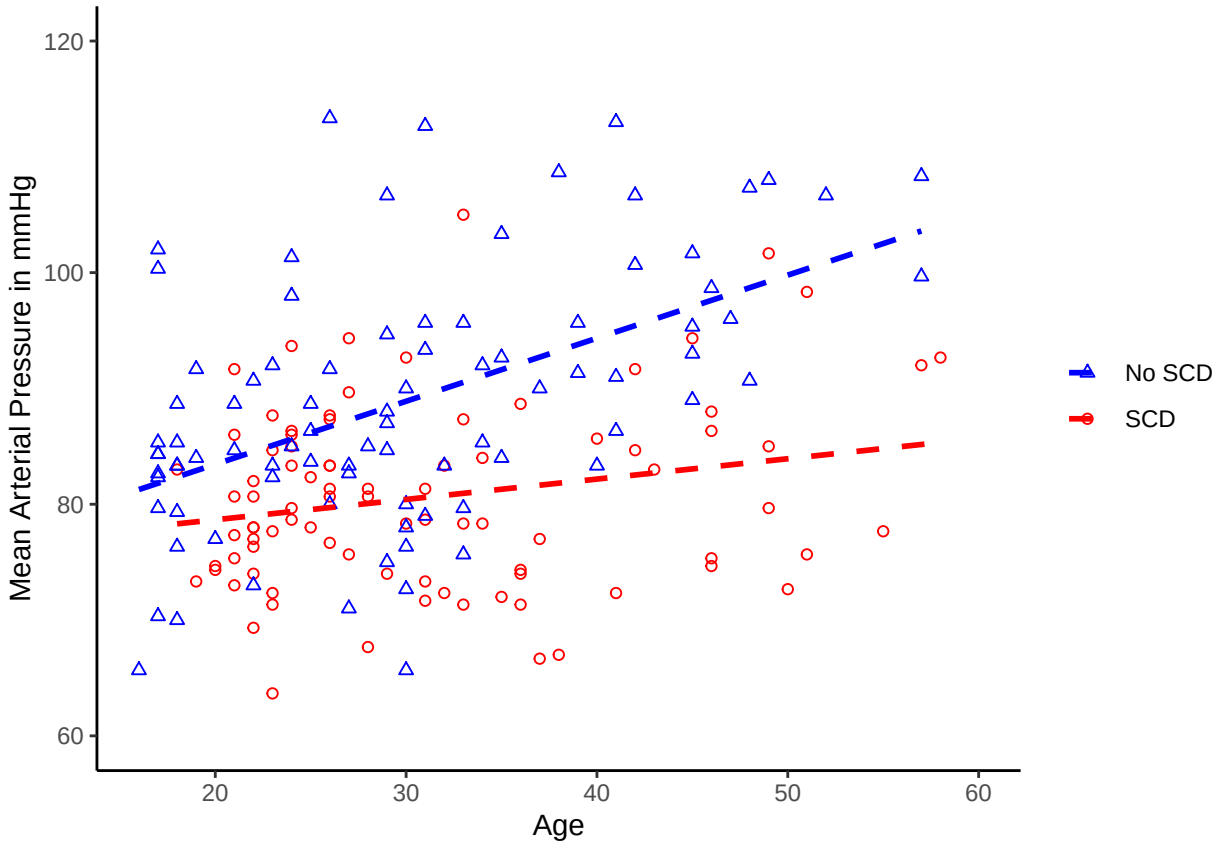
```
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
d %>%
        mutate(SCD = case_when(SCD == 1 ~ "SCD",
                               SCD == 0 ~ "No SCD")) %>%
        ggplot(., aes(y = MAP, x = age, colour = SCD, shape = SCD)) +
        geom_point() +
        scale_shape_manual(values = c(2,1)) +
        scale_colour_manual(values = c("blue", "red")) +
        geom_smooth(method = lm, se = FALSE, linetype = "dashed") +
        ylab("Mean Arterial Pressure in mmHg") +
        xlab("Age") +
        xlim(16,60) +
        ylim(60,120) +
        theme_classic() +
        theme(legend.title = element_blank())
```

```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



ggplot Q10

```
library(ggplot2)
library(dplyr)

d.pi.SCD <- d.pi.SCD %>% mutate(SCD = "SCD")
d.pi.noSCD <- d.pi.noSCD %>% mutate(SCD = "No SCD")

d %>%
        mutate(SCD = case_when(SCD == 1 ~ "SCD",
                               SCD == 0 ~ "No SCD")) %>%
        left_join(d.pi.SCD, by = c("age", "SCD")) %>%
        left_join(d.pi.noSCD, by = c("age", "SCD")) %>%
        ggplot(., aes(y = MAP, x = age, colour = SCD, shape = SCD)) +
        geom_point() +
        scale_shape_manual(values = c(2,1)) +
        scale_colour_manual(values = c("blue", "red")) +
        geom_smooth(method = lm, se = FALSE, linetype = "dashed") +
        ylab("Mean Arterial Pressure in mmHg") +
        xlab("Age") +
```

```
        xlim(16,60) +
        ylim(60,120) +
        theme_classic() +
        theme(legend.title = element_blank()) +
        geom_line(aes(y = lwr.x, x = age), colour = "red", linetype = "longdash") +
        geom_line(aes(y = upr.x, x = age), colour = "red", linetype = "longdash") +
        geom_line(aes(y = lwr.y, x = age), colour = "blue", linetype = "longdash") +
        geom_line(aes(y = upr.y, x = age), colour = "blue", linetype = "longdash")
```