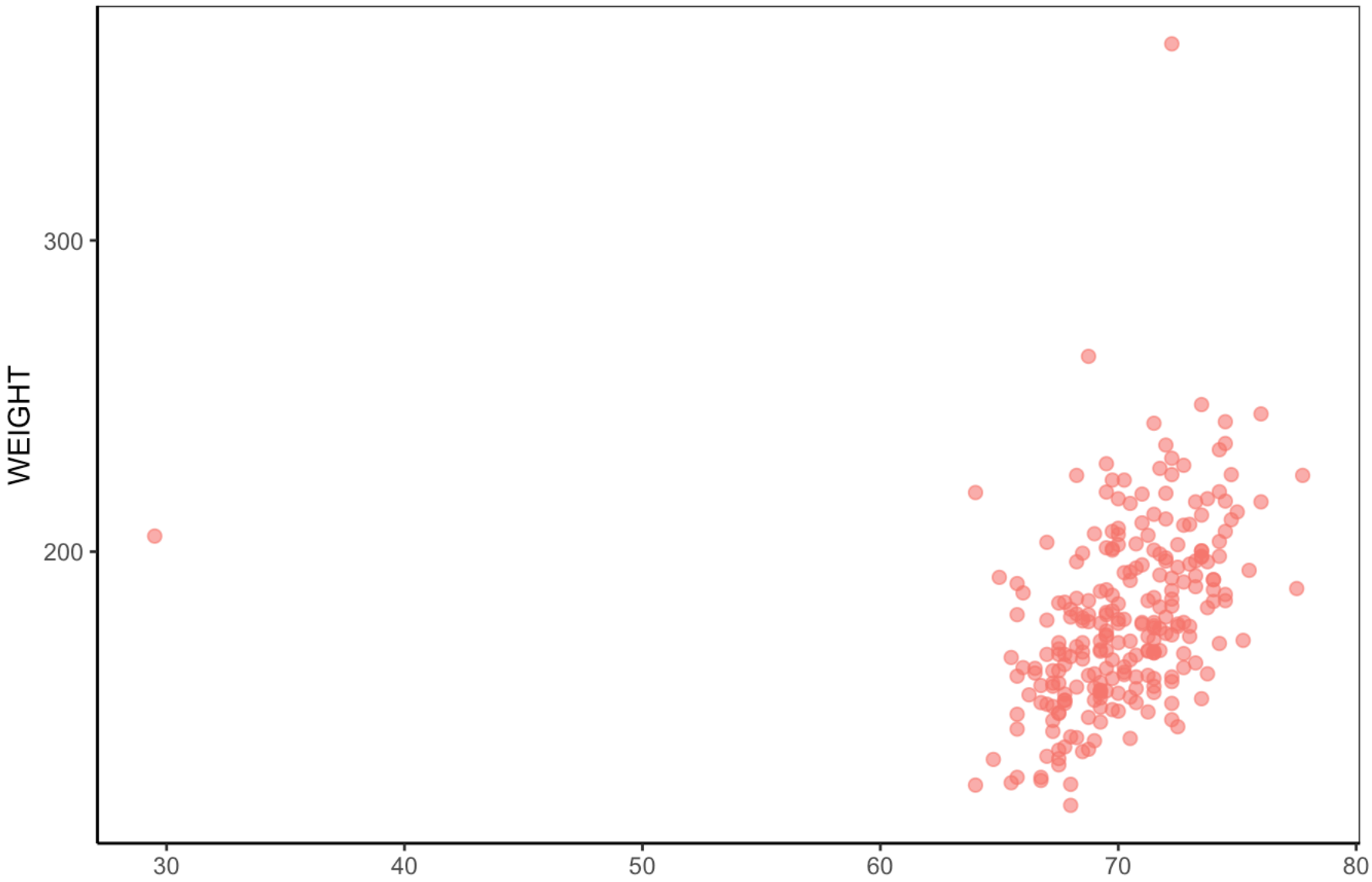
Lec3

HL

9/10/2019

Example of correlation coefficients using body fat data set

```
library(ggplot2)

##### scatter plots
Data_Fat <- read.table("BodyFat.txt", header=TRUE)
plotdata <- Data_Fat
fig <- ggplot(plotdata, aes(HEIGHT, WEIGHT))

fig <- fig + geom_point(aes(x = HEIGHT, y = WEIGHT,color= "blue"), alpha = 0.6, size
= 2)

fig <- fig + ggtitle(paste("" ))
fig <- fig + xlab("HEIGHT") + ylab("WEIGHT")
fig <- fig + theme_bw() + theme( panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(), axis.line = element_line(colour = "black"),le
gend.position = "none")
print(fig, newpage = FALSE)
```




```
which(Data_Fat$HEIGHT<40)
```

```
## [1] 42
```

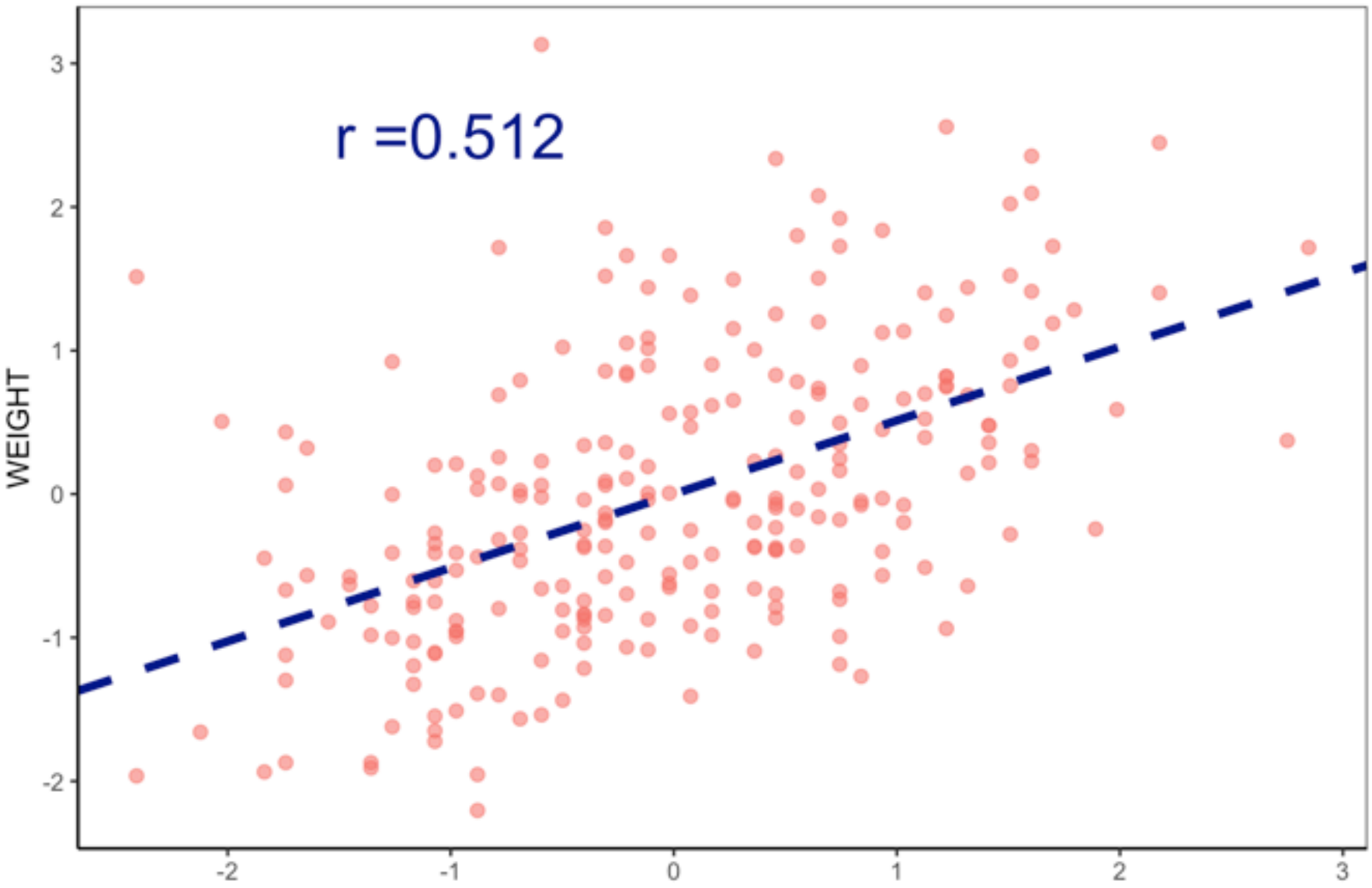
```
which(Data_Fat$WEIGHT >300)
```

```
## [1] 39
```

```
Data_Fat[42,]
```

```
##    IDNO BODYFAT DENSITY AGE WEIGHT HEIGHT ADIPOSITY NECK CHEST ABDOMEN
## 42   42    31.7   1.025  44    205   29.5      29.9 36.6   106   104.3
##      HIP THIGH KNEE ANKLE BICEPS FOREARM WRIST
## 42 115.5  70.6 42.5  23.7   33.6    28.7  17.4
```

```
Data_Fat <- Data_Fat[-c(39,42),]
plotdata <- Data_Fat
### normalize to standard coordinates

plotdata$HEIGHT <- plotdata$HEIGHT - mean(plotdata$HEIGHT)
plotdata$HEIGHT  <- plotdata$HEIGHT /sd(plotdata$HEIGHT)

plotdata$WEIGHT <- plotdata$WEIGHT - mean(plotdata$WEIGHT)
plotdata$WEIGHT  <- plotdata$WEIGHT /sd(plotdata$WEIGHT)

### calculate the correlation coefficient

r <- cor(plotdata$WEIGHT, plotdata$HEIGHT)
str1 <- as.character(floor(r*1000)/1000)
str2 <- paste0("r =",str1)

fig <- ggplot(plotdata, aes(HEIGHT, WEIGHT))

fig <- fig + geom_point(aes(x = HEIGHT, y = WEIGHT,color= "blue"), alpha = 0.6, size
= 2)

fig <- fig + ggtitle(paste("" ))
fig <- fig + xlab("HEIGHT") + ylab("WEIGHT")
fig <- fig + theme_bw() + theme( panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(), axis.line = element_line(colour = "black"),le
gend.position = "none")+ geom_abline(intercept = 0, slope = r, color="darkblue", line
type="dashed", size=1.5)+ggtitle("")+annotate("text",label=str2,x=-1, y=2.5,size=8, c
olor="darkblue")
print(fig, newpage = TRUE)
```




```
#dev.off()

### remove 2 more data points that might be outliers
ol1 <- which(plotdata$HEIGHT < (-2))

ol1
```

```
## [1]  29  36  72 214
```

```
ol2<- which(plotdata[ol1,"WEIGHT"]>1)
plotdata[ol1[ol2],]
```

```
##     IDNO BODYFAT DENSITY AGE   WEIGHT   HEIGHT ADIPOSITY NECK CHEST
## 216  216    45.1   0.995  51 1.513448 -2.40881      37.6 41.2 119.8
##     ABDOMEN   HIP THIGH KNEE ANKLE BICEPS FOREARM WRIST
## 216   122.1 112.8  62.5 36.9  23.6   34.7    29.1  18.4
```

```
ol3 <- which(plotdata$WEIGHT >3)
plotdata[ol3,]
```

```
##    IDNO BODYFAT DENSITY AGE   WEIGHT     HEIGHT ADIPOSITY NECK CHEST
## 41   41    33.1  1.0217  45 3.131691 -0.5935082      39.1 43.2 128.3
##    ABDOMEN   HIP THIGH KNEE ANKLE BICEPS FOREARM WRIST
## 41   126.2 125.6  72.5 39.6  26.6   36.4    32.7  21.4
```

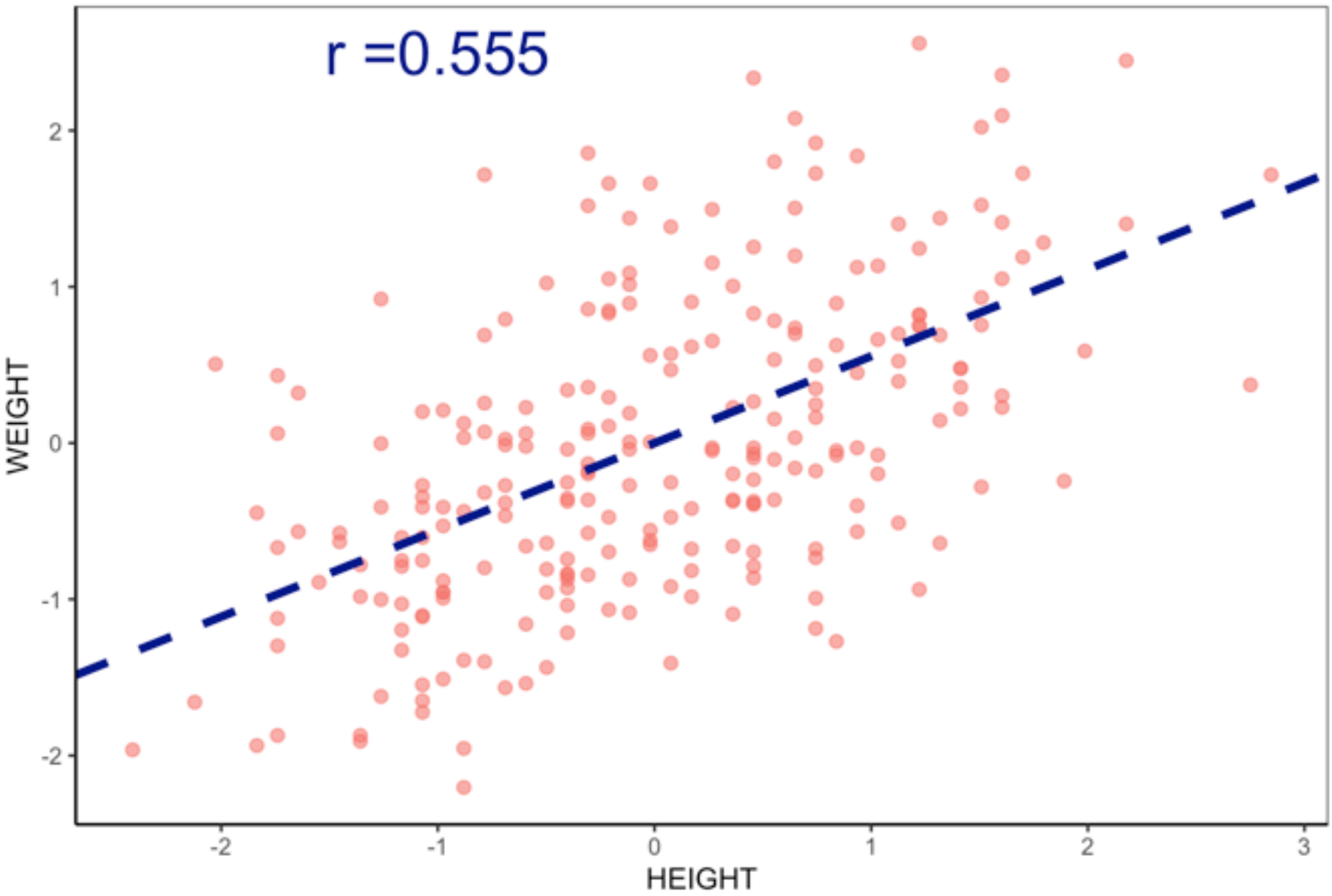
```
indx_o <- c(ol1[ol2],ol3)
nrow(plotdata)
```

```
## [1] 250
```

```
plotdata <- plotdata[-indx_o,]
nrow(plotdata)
```

```
## [1] 248
```

```
### recalculate the correlation coefficient
r <- cor(plotdata$WEIGHT, plotdata$HEIGHT)
str1 <- as.character(floor(r*1000)/1000)
str2 <- paste0("r =",str1)

fig <- ggplot(plotdata, aes(HEIGHT, WEIGHT))

fig <- fig + geom_point(aes(x = HEIGHT, y = WEIGHT,color= "blue"), alpha = 0.6, size
= 2)

fig <- fig + ggtitle(paste("" ))
fig <- fig + xlab("HEIGHT") + ylab("WEIGHT")
fig <- fig + theme_bw() + theme( panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(), axis.line = element_line(colour = "black"),le
gend.position = "none")+ geom_abline(intercept = 0, slope = r, color="darkblue", line
type="dashed", size=1.5)+ggtitle("")+annotate("text",label=str2,x=-1, y=2.5,size=8, c
olor="darkblue")
print(fig, newpage = TRUE)
```


r =0.555
WEIGHT
HEIGHT
2
1
0
-1
-2
-2
-1
0
1
2
3