University of Edinburgh, School of Mathematics
Biostatistics (MATH11230), 2021/2022

Vanda Inácio

In this supplement I reproduce the results in the slides about confounding.

```
require(readxl)
data_wcgs <- read_excel("wcgsdata.xls")

smoking_binary <- ifelse(data_wcgs$Ncigs0 == 0, 0, 1)
data_wcgs$smoking_binary <- as.factor(smoking_binary)
res_1 <- glm(Chd69 ~ smoking_binary, family = "binomial",
             data = data_wcgs)
summary(res_1)
```

```
##
## Call:
## glm(formula = Chd69 ~ smoking_binary, family = "binomial", data = data_wcgs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4731  -0.4731  -0.3497  -0.3497   2.3769
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -2.7636     0.1042  -26.54  < 2e-16 ***
## smoking_binary1   0.6299     0.1337    4.71 2.47e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1781.2  on 3153  degrees of freedom
## Residual deviance: 1758.4  on 3152  degrees of freedom
## AIC: 1762.4
##
## Number of Fisher Scoring iterations: 5
```

```
exp(confint.default(res_1))
```

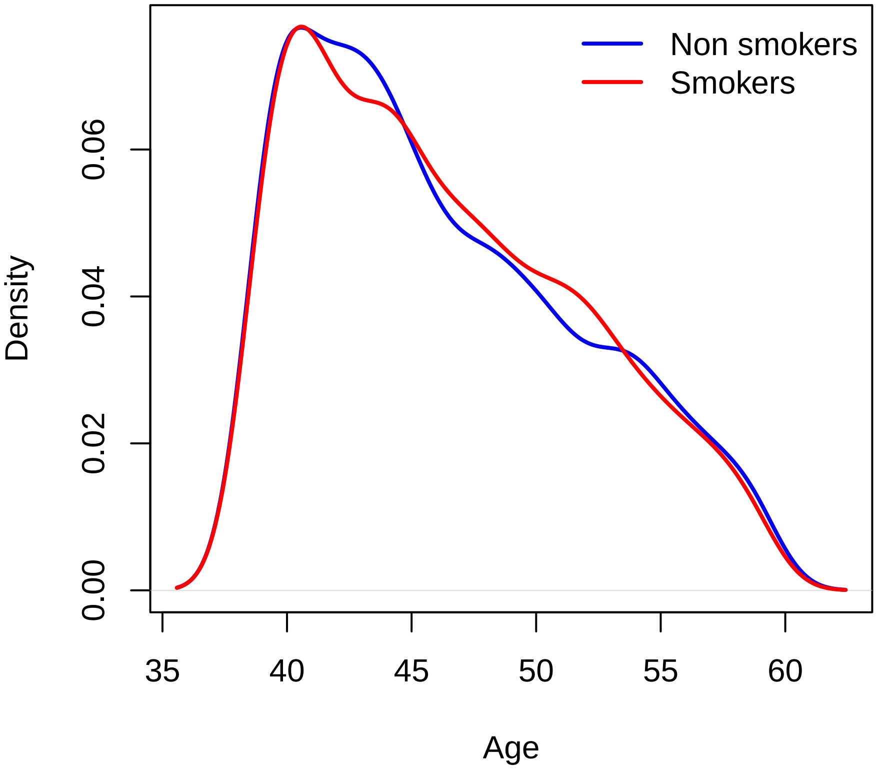
```
##                     2.5 %     97.5 %
## (Intercept)     0.0514188 0.07734428
## smoking_binary1 1.4445172 2.43988522
```

```
Age_non_smokers <- data_wcgs$Age0[data_wcgs$smoking_binary == 0]
Age_smokers <- data_wcgs$Age0[data_wcgs$smoking_binary == 1]
```

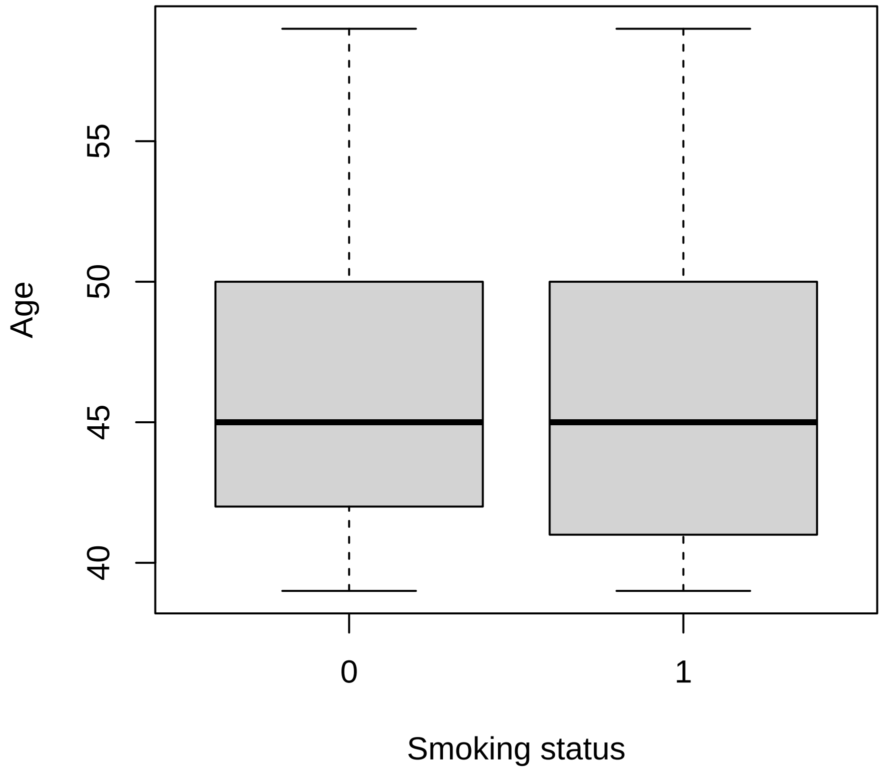
```
t.test(Age_non_smokers, Age_smokers, alternative = "two.sided", var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  Age_non_smokers and Age_smokers
## t = -0.26738, df = 3132.6, p-value = 0.7892
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4385357  0.3332829
## sample estimates:
## mean of x mean of y
##  46.25363  46.30626
```

```
plot(density(Age_non_smokers), col =  "blue2", lwd = 2,
     xlab = "Age", ylab = "Density", main = "")
lines(density(Age_smokers), col = "red", lwd = 2)
legend("topright", legend = c("Non smokers", "Smokers"),
       lwd = c(2, 2), col = c("blue2", "red"), bty = "n")
```



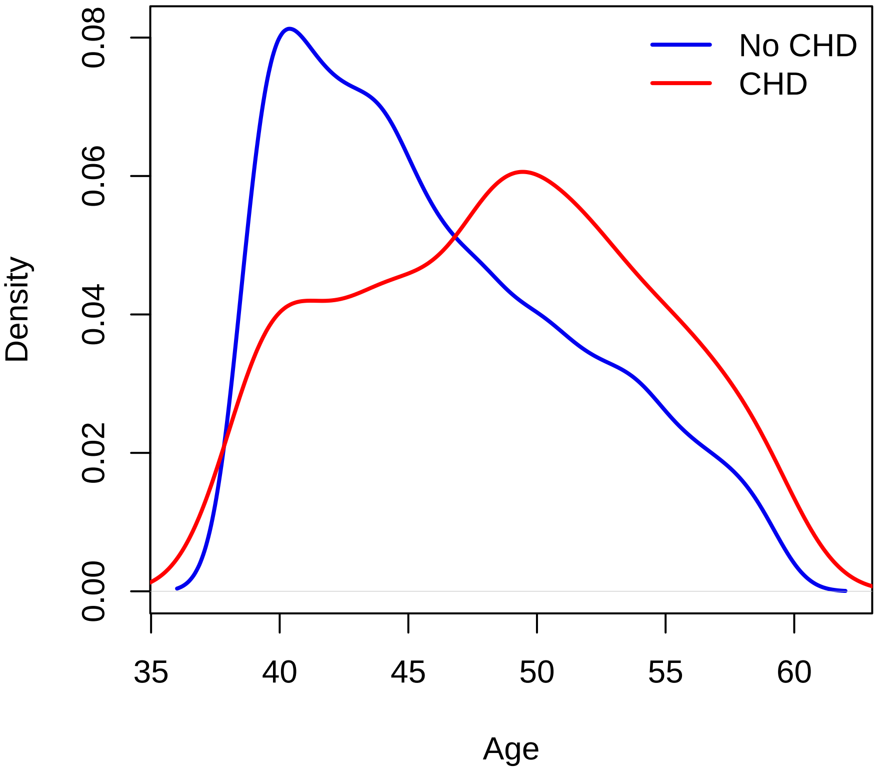
```
boxplot(data_wcgs$Age0 ~ data_wcgs$smoking_binary,
        ylab = "Age", xlab = "Smoking status")
```



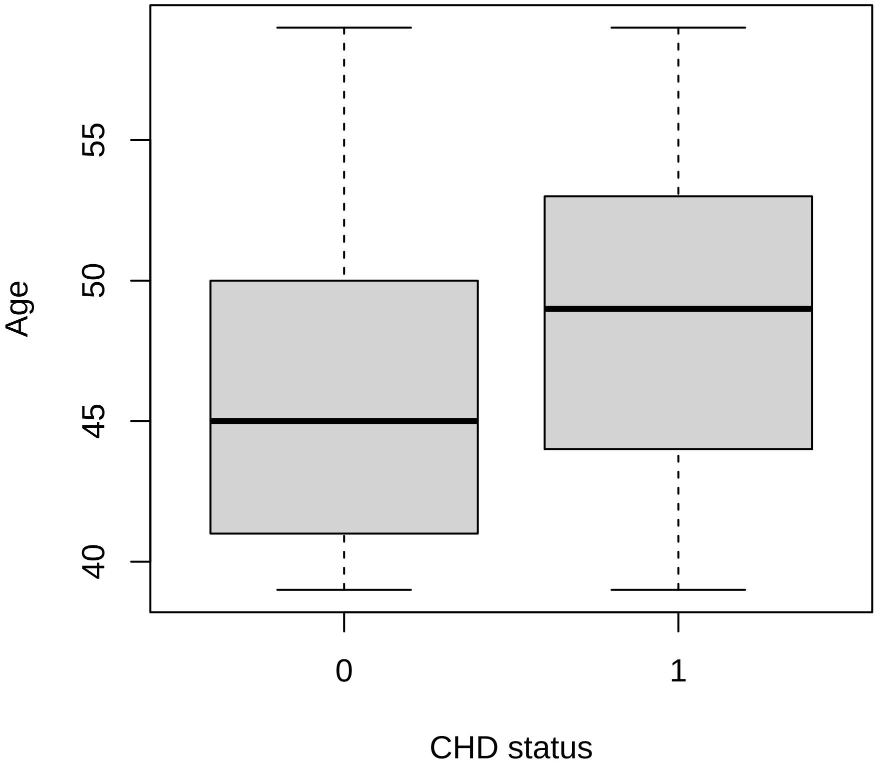
```r
Age_no_CHD <- data_wcgs$Age0[data_wcgs$Chd69 == 0]
Age_CHD <- data_wcgs$Age0[data_wcgs$Chd69 == 1]

t.test(Age_no_CHD, Age_CHD, alternative = "two.sided", var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  Age_no_CHD and Age_CHD
## t = -6.4067, df = 297.6, p-value = 5.799e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.147372 -1.668175
## sample estimates:
## mean of x mean of y
##  46.08250  48.49027
```

```r
plot(density(Age_no_CHD), col =  "blue2", lwd = 2,
     xlab = "Age", ylab = "Density", main = "")
lines(density(Age_CHD), col = "red", lwd = 2)
legend("topright", legend = c("No CHD", "CHD"),
       lwd = c(2, 2), col = c("blue2", "red"), bty = "n")
```



```
boxplot(data_wcgs$Age0 ~ data_wcgs$Chd69,
        ylab = "Age", xlab = "CHD status")
```



```
res_2 <- glm(Chd69 ~ smoking_binary + Age0, family = "binomial",
             data = data_wcgs)
```

```r
summary(res_2)
```

```
##
## Call:
## glm(formula = Chd69 ~ smoking_binary + Age0, family = "binomial",
##     data = data_wcgs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7106  -0.4452  -0.3717  -0.2985   2.6163
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -6.32132    0.56144 -11.259  < 2e-16 ***
## smoking_binary1   0.63816    0.13472   4.737 2.17e-06 ***
## Age0              0.07518    0.01139   6.599 4.13e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1781.2  on 3153  degrees of freedom
## Residual deviance: 1715.2  on 3151  degrees of freedom
## AIC: 1721.2
##
## Number of Fisher Scoring iterations: 5
```

```r
#rule of thumb
as.numeric((exp(res_1$coefficients[2]) - exp(res_2$coefficients[2]))/exp(res_1$coefficients[2]))
```

```
## [1] -0.008331309
```

```r
#information criteria
AIC(res_1); AIC(res_2)
```

```
## [1] 1762.381
```

```
## [1] 1721.234
```

```r
BIC(res_1); BIC(res_2)
```

```
## [1] 1774.494
```

```
## [1] 1739.404
```

```r
#Likelihood ratio test
dif <- res_1$deviance - res_2$deviance
dif
```

```
## [1] 43.14663
```

```r
pchisq(dif, df = 1, lower = FALSE)
```

```
## [1] 5.07874e-11
```

```r
#alternative and easier way
anova(res_1, res_2, test = "Chisq")
```

```
## Analysis of Deviance Table
##
```

```
## Model 1: Chd69 ~ smoking_binary
## Model 2: Chd69 ~ smoking_binary + Age0
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      3152     1758.4
## 2      3151     1715.2  1   43.147 5.079e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```