Vital capacity - with addon

Lasse Engbo Christiansen

15 jan 2019

Vital capacity

Vital capacity for workers in cadmium factory three groups of workers (+10 , 0-10 and 0 years of work at the factory)

```
# You might have to install this package
library(ISwR)

# Load data
vit <- vitcap2
summary(vit)
```

```
##      group            age        vital.capacity
##  Min.   :1.000   Min.   :18.00   Min.   :2.700
##  1st Qu.:2.000   1st Qu.:32.00   1st Qu.:3.935
##  Median :3.000   Median :41.00   Median :4.530
##  Mean   :2.381   Mean   :40.55   Mean   :4.392
##  3rd Qu.:3.000   3rd Qu.:48.00   3rd Qu.:4.947
##  Max.   :3.000   Max.   :65.00   Max.   :5.860
```

```
str(vit)
```

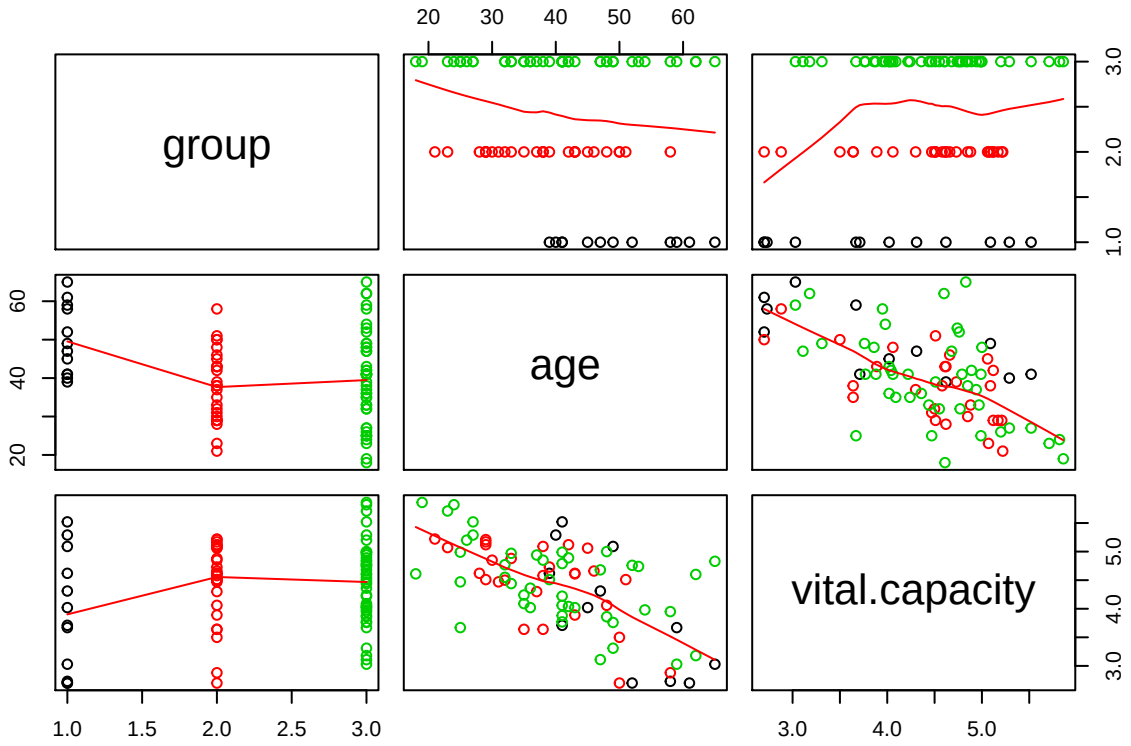
```
## 'data.frame':    84 obs. of  3 variables:
##  $ group         : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ age           : int  39 40 41 41 45 49 52 47 61 65 ...
##  $ vital.capacity: num  4.62 5.29 5.52 3.71 4.02 5.09 2.7 4.31 2.7 3.03 ...
```

```
# Make group a factor
vit$group <- factor(vit$group, labels = c("+10","0-10","Not"))
summary(vit)
```

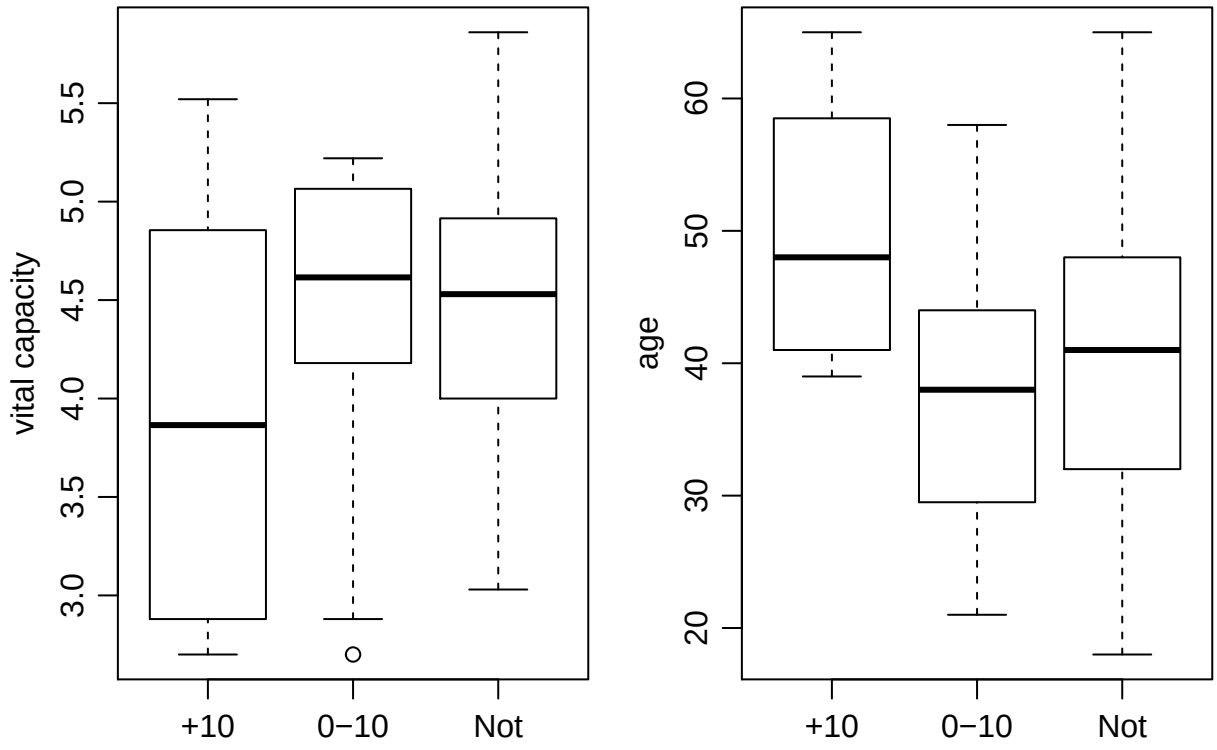
```
##   group         age        vital.capacity
##  +10 :12   Min.   :18.00   Min.   :2.700
##  0-10:28   1st Qu.:32.00   1st Qu.:3.935
##  Not :44   Median :41.00   Median :4.530
##            Mean   :40.55   Mean   :4.392
##            3rd Qu.:48.00   3rd Qu.:4.947
##            Max.   :65.00   Max.   :5.860
```

Plotting the data

```
pairs(vit, panel = panel.smooth, col = vit$group)
```

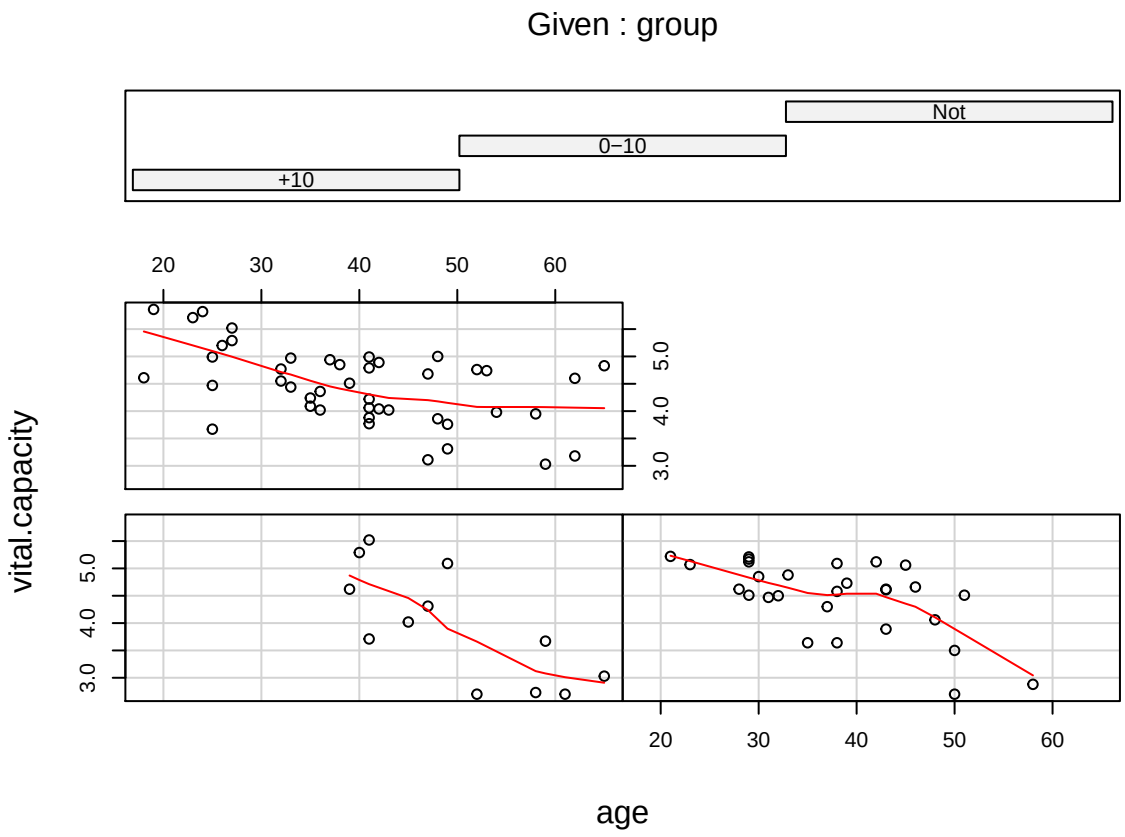



```
par(mfrow = c(1,2), mar=c(3,3,2,1), mgp=c(2, 0.7,0))
boxplot(vital.capacity ~ group, vit, ylab = "vital capacity")
boxplot(age ~ group, vit, ylab = "age")
```



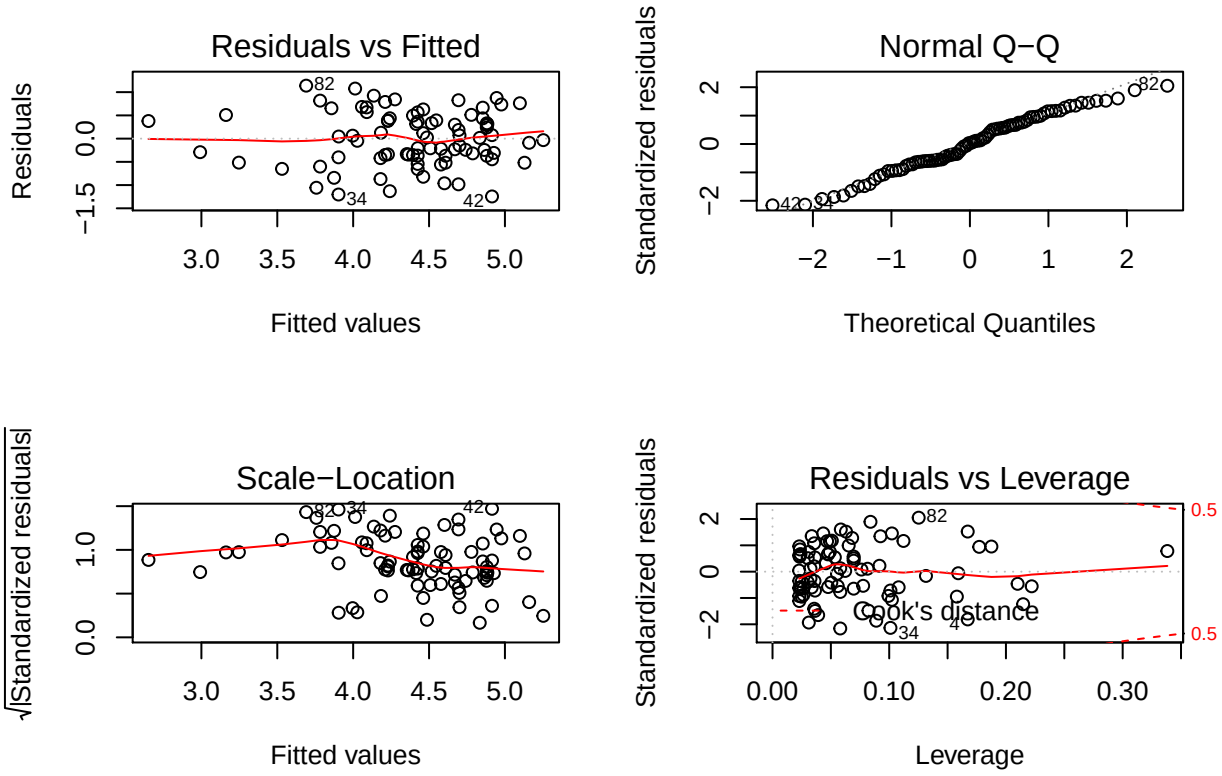
Visualize potential interaction between group and age

```
coplot(vital.capacity ~ age | group, vit, panel = panel.smooth)
```



ANCOVA including interaction between group and age

```
lm1 <- lm(vital.capacity ~ group*age, vit)
par(mfrow = c(2,2))
plot(lm1)
```




```
# Assumtions are not rejected
drop1(lm1, test = "F")
```

```
## Single term deletions
##
## Model:
## vital.capacity ~ group * age
##           Df Sum of Sq    RSS     AIC F value  Pr(>F)
## <none>                 27.535 -81.689
## group:age  2    2.4995 30.035 -78.391  3.5402 0.03376 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The interaction is significant so the model should not be reduced.

Interpret results

Summary as raw R-output

```
summary(lm1)
```

```
##
## Call:
## lm(formula = vital.capacity ~ group * age, data = vit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.24497 -0.36929  0.01977  0.43681  1.13953
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     8.18344    0.99358   8.236 3.28e-12 ***
## group0-10      -1.95341    1.10481  -1.768   0.0810 .
## groupNot       -2.50315    1.04184  -2.403   0.0187 *
## age            -0.08511    0.01967  -4.327 4.44e-05 ***
## group0-10:age   0.03858    0.02327   1.658   0.1014
## groupNot:age    0.05450    0.02107   2.587   0.0116 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5942 on 78 degrees of freedom
## Multiple R-squared:  0.422,  Adjusted R-squared:  0.385
## F-statistic: 11.39 on 5 and 78 DF,  p-value: 2.871e-08
```

But it could also be in table

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.1834	0.9936	8.236	0.000
group0-10	-1.9534	1.1048	-1.768	0.081
groupNot	-2.5031	1.0418	-2.403	0.019
age	-0.0851	0.0197	-4.327	0.000
group0-10:age	0.0386	0.0233	1.658	0.101
groupNot:age	0.0545	0.0211	2.587	0.012

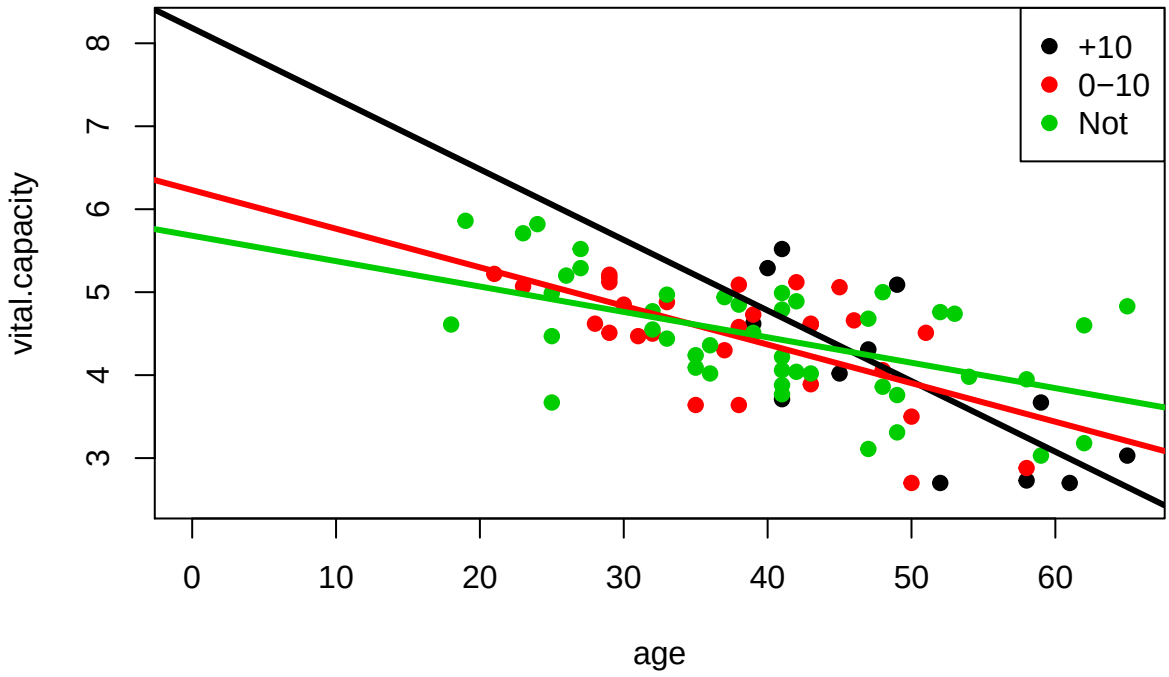
Read (https://haozhu233.github.io/kableExtra/awesome_table_in_html.html) if you want to see more examples on how to format tables for Markdown and HTML from R.

Plot results

Graphical interpretation is nicer :)

First the simple

```
par(mfrow = c(1,1))
plot(vital.capacity ~ age, vit, pch = 19, col = vit$group, xlim = c(0,65), ylim = c(2.5,8.2))
legend("topright", legend = c("+10","0-10","Not"), pch = 19, col = 1:3)
co <- coef(lm1)
abline(a = co[1], b = co[4], lwd = 3)
abline(a = co[1]+co[2], b = co[4]+co[5], lwd = 3, col = 2)
abline(a = co[1]+co[3], b = co[4]+co[6], lwd = 3, col = 3)
```

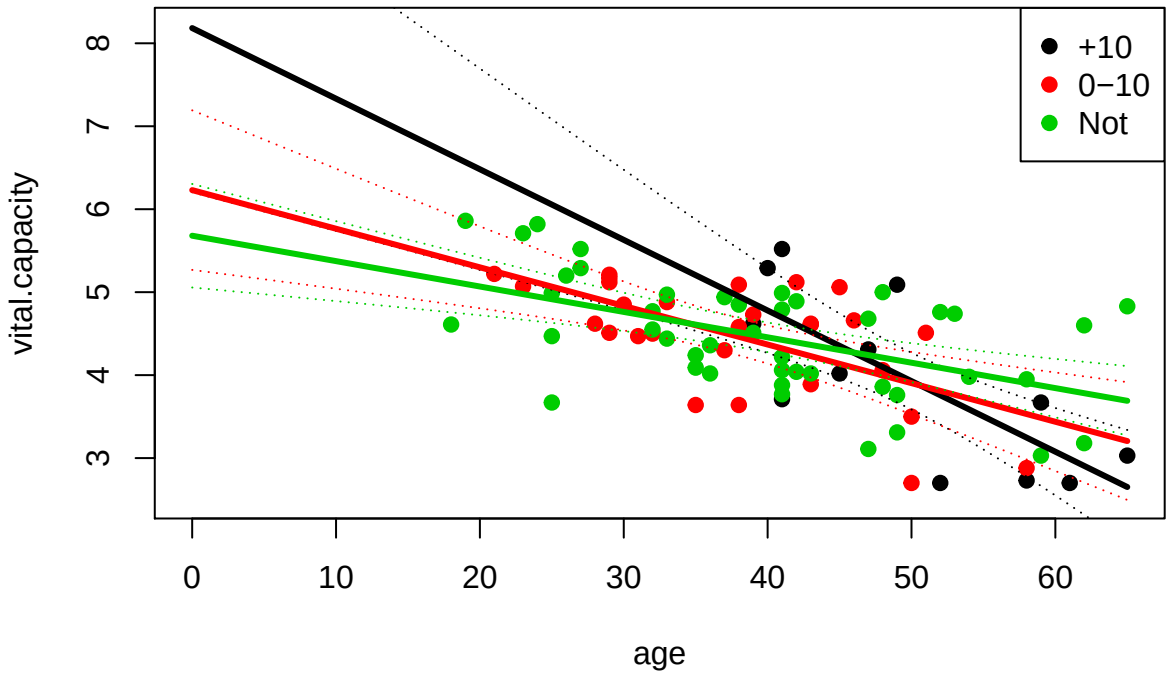


But some mearsure of uncertainty should be added. So redoing with 95% confidence intervals

```
new.data <- data.frame("group" = rep(c("+10","0-10","Not"), each = 100),
                       "age" = seq(0,max(vit$age), length = 100))
conf.int <- predict(lm1, new.data, interval = "confidence")

plot(vital.capacity ~ age, vit, pch = 19, col = vit$group, xlim = c(0,65), ylim = c(2.5,8.2))
legend("topright", legend = c("+10","0-10","Not"), pch = 19, col = 1:3)

matlines(new.data$age[new.data$group=="+10"], conf.int[new.data$group=="+10",],
         lty = c(1,3,3), lw = c(3,1,1), col = 1)
matlines(new.data$age[new.data$group=="0-10"], conf.int[new.data$group=="0-10",],
         lty = c(1,3,3), lw = c(3,1,1), col = 2)
matlines(new.data$age[new.data$group=="Not"], conf.int[new.data$group=="Not",],
         lty = c(1,3,3), lw = c(3,1,1), col = 3)
```



Addon: Uncertainty on combinations of parameters

When a factor is involved the parameters associated with the factor, e.g. 'group' are describing how these levels are different from the reference level. And the Std. error that is associated is also for that difference. We cannot just square the relevant standard errors and add them as the parameters are most likely correlated. In the particular case the correlation matrix for the estimated parameters is:

```
lm1s <- summary(lm1, correlation = TRUE)
kable_styling( kable(lm1s$correlation, digits = 3), full_width = FALSE, bootstrap_options = c("striped"
```

	(Intercept)	group0-10	groupNot	age	group0-10:age	groupNot:age
(Intercept)	1.000	-0.899	-0.954	-0.985	0.833	0.920
group0-10	-0.899	1.000	0.858	0.886	-0.976	-0.827
groupNot	-0.954	0.858	1.000	0.939	-0.794	-0.980
age	-0.985	0.886	0.939	1.000	-0.845	-0.934
group0-10:age	0.833	-0.976	-0.794	-0.845	1.000	0.789
groupNot:age	0.920	-0.827	-0.980	-0.934	0.789	1.000

In the particular case the correlations are quite large. If we want to calculate the individual intercepts and their standard deviation then we should just recap the theory.

Let us define the vector of estimated parameters as $\hat{\theta}$ with $V[\hat{\theta}] = \Sigma_\theta$. Given a matrix A then the variance of $A\hat{\theta}$ is

$$V[A\hat{\theta}] = AV[\hat{\theta}]A^T = A\Sigma_\theta A^T$$

In the particular case where we want the three intercepts A should be:

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Leading to the following estimates of the individual intercepts

```
A <- cbind(diag(3),0,0,0)
A[,1] <- 1
est <- A %*% lm1s$coefficients[,1]
var_est <- A %*% lm1s$cov.unscaled %*% t(A) * lm1s$sigma^2
coef <- data.frame(Group=levels(vit$group), Intercept = est, sd.error=sqrt(diag(var_est)))
kable_styling(kable(coef, digits = 3), full_width = FALSE, bootstrap_options = c("striped", "hover"))
```

Group	Intercept	sd.error
+10	8.183	0.994
0-10	6.230	0.483
Not	5.680	0.313

Similarly for the slopes

```
A <- cbind(0,0,0,diag(3))
A[,4] <- 1
est <- A %*% lm1s$coefficients[,1]
var_est <- A %*% lm1s$cov.unscaled %*% t(A) * lm1s$sigma^2
coef <- data.frame(Group=levels(vit$group), Slope = est, sd.error=sqrt(diag(var_est)))
kable_styling(kable(coef, digits = 4), full_width = FALSE, bootstrap_options = c("striped", "hover"))
```

Group	Slope	sd.error
+10	-0.0851	0.0197
0-10	-0.0465	0.0124
Not	-0.0306	0.0075

It is noticed that the uncertainty is highest in the group with those that worked in the cadmium industry for 10+ years. There are two reasons for that: It is the group with the fewest observations (12 vs 28 and 44) and the span of the ages is the smallest.