Linear Models

Anders Stockmarr
Course developers: Anders Stockmarr, Elisabeth Wreford Andersen

DTU Department of Applied Mathematics and Computer Science
Section for Statistics and Data Analysis
Technical University of Denmark
anst@dtu.dk

January 8th, 2025

DTU Compute
Department of Applied Mathematics and Computer Science

Plan for this week

Monday	Statistical inference, and the t-test
Tuesday	Simple and Multiple regression
Wednesday	ANOVA, ANCOVA, and linear models
Thursday	Categorical data, statistical report writing, logistic Regression
Friday	Introduction to repeated measures , Principal Component Analysis

Overview

1. The Linear Model
   - ANCOVA
2. Vital Capacity and Cadmium
3. Exercises

Terminology

For continuous outcomes (e.g. birth weight)

- Regression: The covariates are also continuous.
  - Simple (linear) regression: Just one covariate.
  - Multiple (linear) regression: Two or more covariates.
- Variance analysis: Covariates are categorical (grouped, factors).
  - One-way analysis of variance: Just one covariate (factor).
  - Two-way analysis of variance: Two covariates (factors).
- General linear model: Both types of covariates in the same model.
  - Analysis of covariance: Exactly one continuous and one categorical covariate.

The General Linear Model (GLM)

Y_i is the outcome for person i and $(X_{i1}, \ldots, X_{ik})$ are explanatory covariates e.g. age of person i, or a "dummy"variable:

$$X_{ij} = \begin{cases} 1 & \text{if person } i \text{ is from group } j \\ 0 & \text{if person } i \text{ is not from group } j \end{cases}$$

E.g. $X_{i1} = 1$ if person i a boy and $X_{i1} = 0$ if person i a girl.
Model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon_i$$

Where $\varepsilon_i \sim N(0, \sigma^2)$ and independent.
The predicted values are called $\hat{Y}_i$.

Model Reduction in GLM

* In a general linear model we can split the variation.

$$SS_{total} = SS_{model} + SS_{residual}$$

Model Reduction in GLM

* In a general linear model we can split the variation.

$$SS_{total} = SS_{model} + SS_{residual}$$

Model Sum of Squares $SS_{model} = \sum(\hat{Y}_i - \bar{Y})^2$

- Explained variation
- How much do the predicted values vary?
- Large is good

Model Reduction in GLM

* In a general linear model we can split the variation.

$$SS_{total} = SS_{model} + SS_{residual}$$

Model Sum of Squares $SS_{model} = \sum(\hat{Y}_i - \bar{Y})^2$

- Explained variation
- How much do the predicted values vary?
- Large is good

Residual Sum of Squares $SS_{residual} = \sum(Y_i - \hat{Y}_i)^2$

- Variation not explained by model.
- How large are the differences between observed and predicted?
- Small is good.

Model Reduction - F test

* We want to compare two models.
  The original (no. 1) and a simplified (the hypothesis, no. 2).

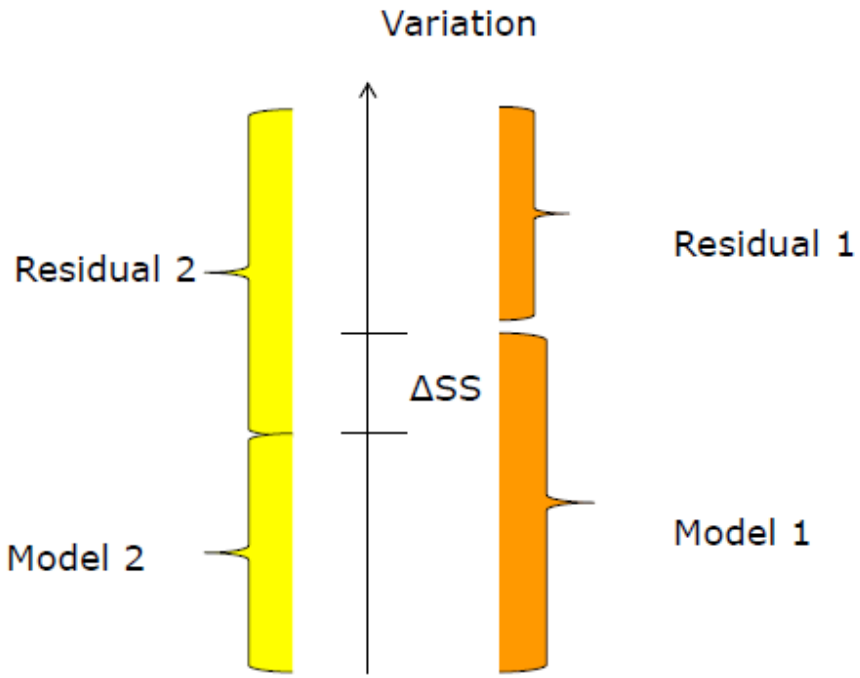
* Is it ok to use the simplified model? Is it good enough?

* Note the models must be nested, i.e. you get one from the other by setting parameters to zero ("remove effects").

* We look at changes in model sum of squares:
  How much less is explained by the simpler model?

$$\Delta SS = SS_{model1} - SS_{model2}$$

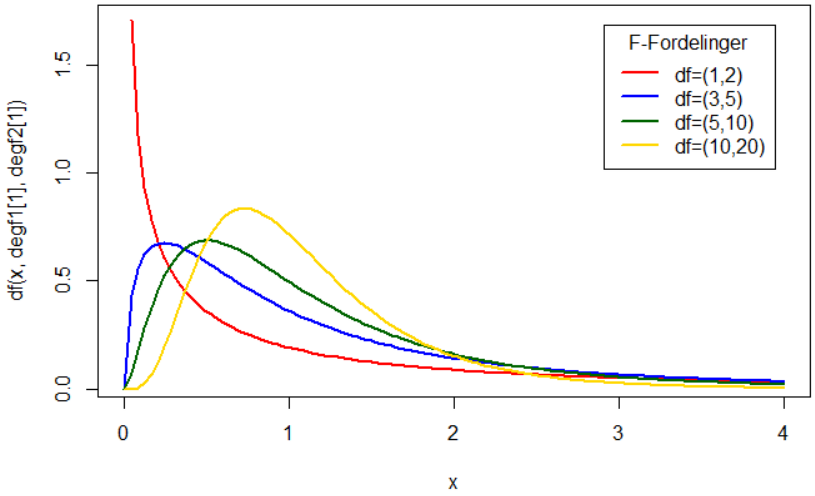
Model Reduction - contd.



- More parameters can explain (a little) more variation $\Delta SS > 0$.
- How much more?
- How large ΔSS before test significant?

F-test

- The size of ΔSS is seen together with the reduction in parameters $\Delta DF = Df_1 - Df_2$.
- ΔSS is compared to the residual variation from the larger model.



$$F = \frac{\Delta SS/\Delta Df}{SS_{residual}/Df_1} \sim F(\Delta Df, Df_1)$$

The R^2 Statistic

- The R^2 statistic is given as

$$R^2 = \frac{SS_{model}}{SS_{total}}$$

- Often referred to as the *coefficient of determination*.
- Measures how much of the variation that the model explains, large is good. Is found in the summary output from lm.
- A high R^2 gives a model that explains a lot; but says absolutely **nothing** about whether it is a *sensible* explanation.
- Whether the explanations are *sensible* in modelling terms, is decided from the model control.

The Adjusted R^2 Statistic

- The R^2 automatically increases when you add explanatory variables to the model. This is not always sensible.
- To correct for this phenomenon, one often uses the *adjusted* R^2, $\overline{R}^2$ instead:

$$MS_{model} = SS_{model}/df_{model}; MS_{res} = SS_{res}/df_{res};$$

$$MS_{total} = SS_{model}/df_{total}$$

$$\overline{R}^2 = 1 - \frac{MS_{res}}{MS_{total}}$$

- Also found in the summary output of lm.

Analysis of Covariance - The Simplest GLM

* A (historical) term for a model with exactly one categorical covariate (group, factor) and exactly one continuous covariate.
* What could be the aim of such an analysis?

Analysis of Covariance - The Simplest GLM

* A (historical) term for a model with exactly one categorical covariate (group, factor) and exactly one continuous covariate.
* What could be the aim of such an analysis?
  - To study the two covariates.

Analysis of Covariance - The Simplest GLM

* A (historical) term for a model with exactly one categorical covariate (group, factor) and exactly one continuous covariate.
* What could be the aim of such an analysis?
  - To study the two covariates.
  - Remove bias, e.g. correct for height differences when comparing lung capacity of smokers and non-smokers.

Analysis of Covariance - The Simplest GLM

* A (historical) term for a model with exactly one categorical covariate (group, factor) and exactly one continuous covariate.
* What could be the aim of such an analysis?
  - To study the two covariates.
  - Remove bias, e.g. correct for height differences when comparing lung capacity of smokers and non-smokers.
  - Increase the power in a randomized clinical trial by reducing the unexplained part of the variance, e.g. by including age as a covariate.

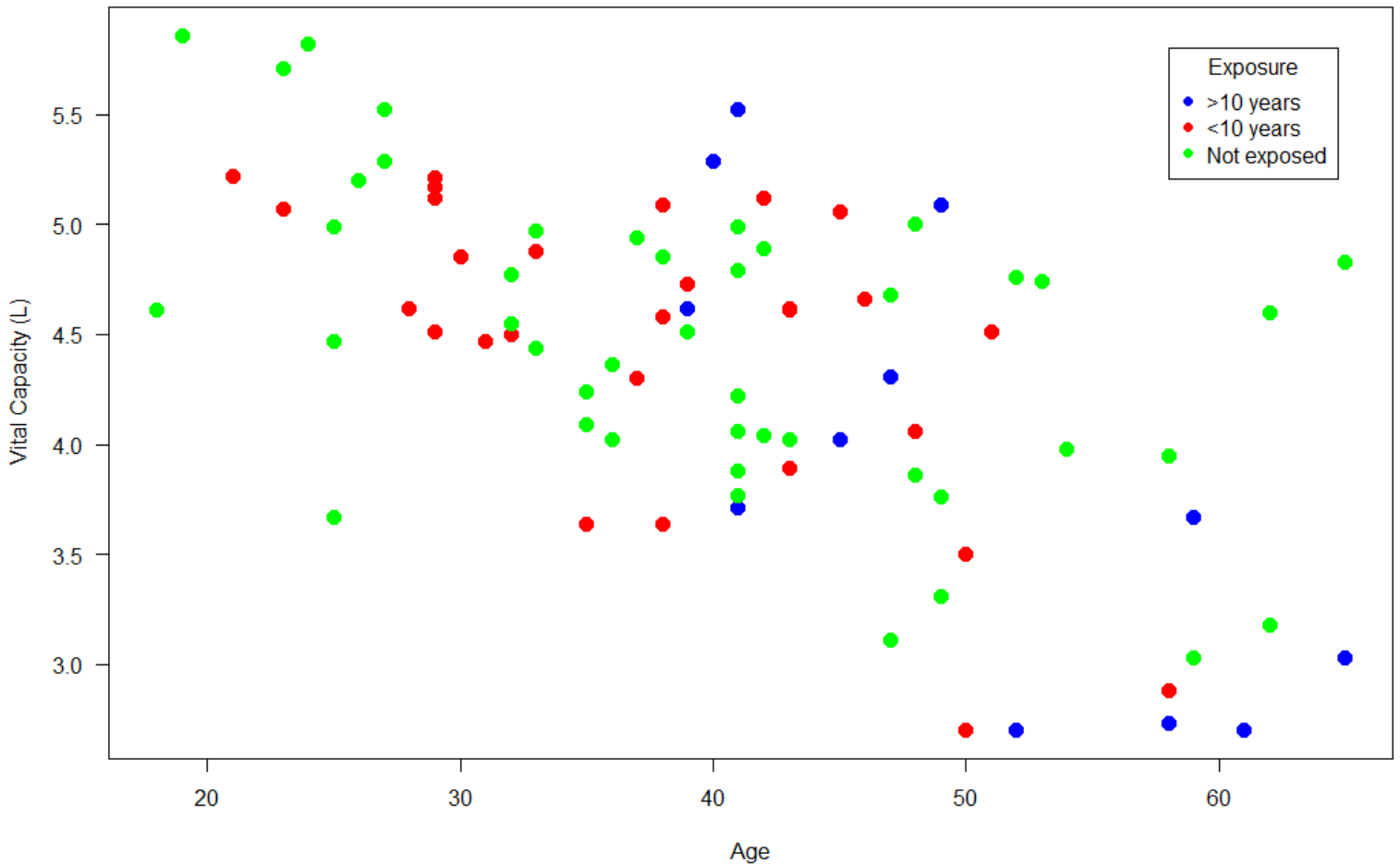
Example: Vital Capacity and Cadmium

- We have data from a study of the effect of exposure to cadmium on the vital capacity. (From P. Armitage & G. Berry: Statistical methods in medical research. 2nd ed. Blackwell 1987)
- Vital capacity is the maximum amount of air a person can expel from the lungs after a maximum inhalation.
- We have measurements of vital capacity (L), age and exposure to cadmium ($>$ 10 years, $<$ 10 years, not exposed).
- Start by plotting the data!

Rcode for plots, scatter plot

```
CADdata <- read.csv("cadmium.txt", sep="")
CADdata$group <- as.factor(CADdata$group)

#PLOT THE DATA WITH DIFFERENT COLOURS IN EXPOSURE GROUPS
plot(CADdata$age, CADdata$vitcap,
     col = c("blue","red","green")[CADdata$group],
     xlab = "Age", ylab = "Vital Capacity (L)",
     las = 1, cex = 1.5, pch = c(16,16,16))
legend(55,5.8, c(">10 years","<10 years", "Not exposed"),
       col = c("blue","red","green"),
       pch = c(16,16,16), title = "Exposure")
```

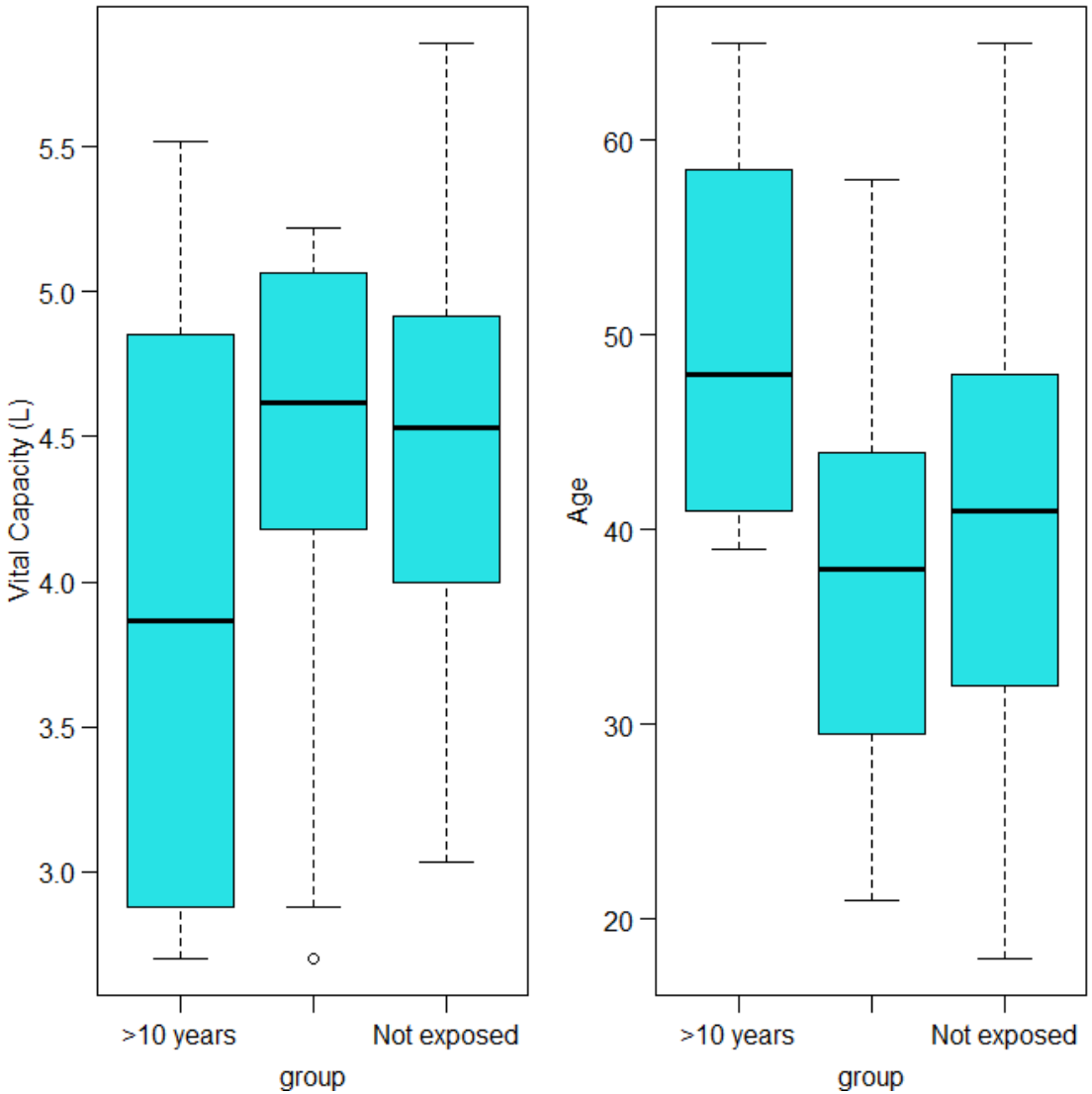
Scatter Plot



Rcode for plots, boxplot

```
#TWO PLOTS NEXT TO EACH OTHER
par(mfrow = c(1,2), mgp = c(2,0.7,0), mar = c(3,3,1,1))

boxplot(vitcap ~ group, data = CADdata, ylab =
        'Vital Capacity (L)',las = 1, xaxt = "n", col = 5)
axis(1, at = c(1,2,3),
     labels = c(">10 years","<10 years", "Not exposed"))


boxplot(age ~ group, data = CADdata ,ylab = 'Age',
        las = 1, xaxt = "n", col = 5)
axis(1, at = c(1,2,3),
     labels = c(">10 years", "<10 years", "Not exposed"))
#BACK TO ONE PLOT
par(mfrow = c(1,1))
```

Boxplots


Vital Capacity (L)
5.5
5.0
4.5
4.0
3.5
3.0
>10 years
Not exposed
group
Age
60
50
40
30
20
>10 years
Not exposed
group

Comparing Groups

Comparing groups that are not quite comparable (e.g. cadmium exposure).
Confounder: A variable that

- Has an effect on the outcome.
- Is associated to group (different ages in groups)

This can cause bias.

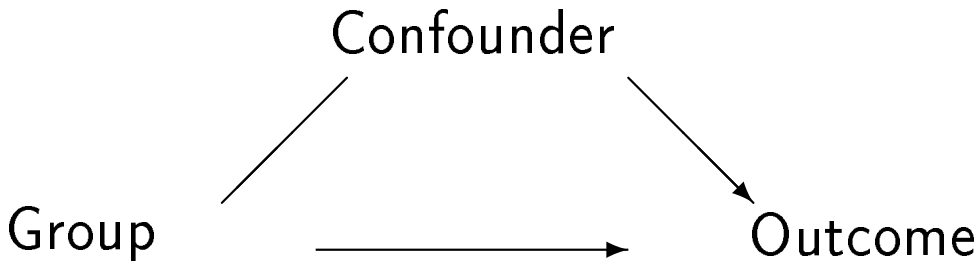
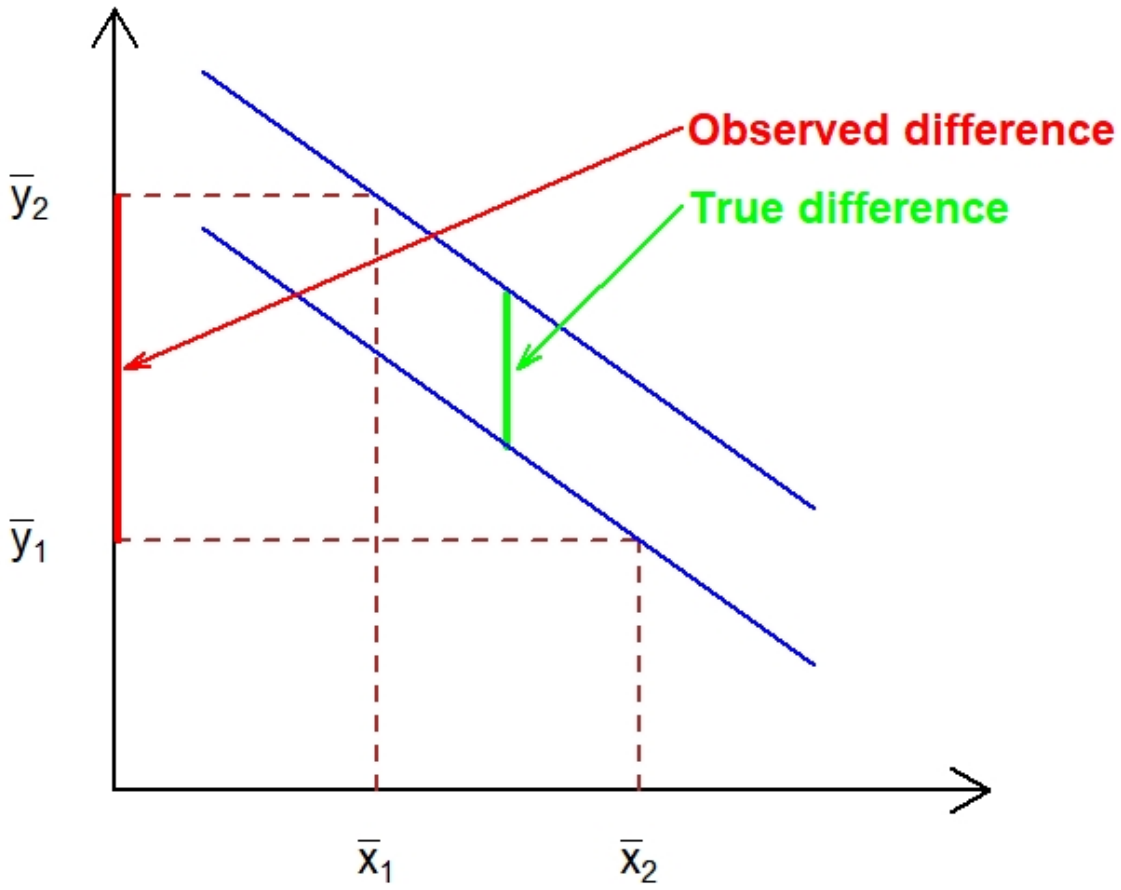
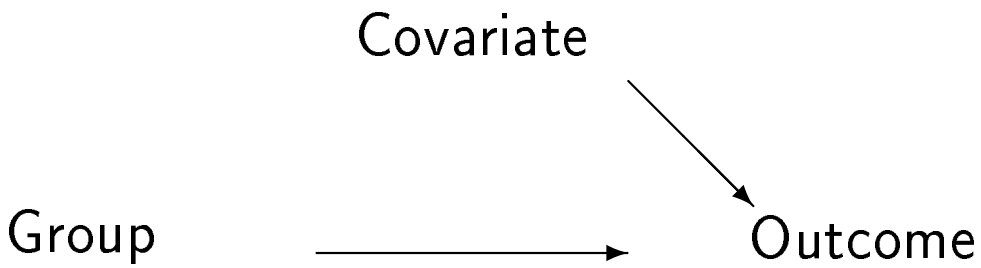


Illustration of Confounding and ANCOVA



Adjustment

Even if the distribution of the covariate is the same in the groups, then it can reduce the variation.



- This gives greater power.
- But remember that we are answering a different scientific question (which one?).

Trying to Avoid Bias

Matching Choose individuals so they are similar for important disturbing covariates. Remember to include the matching variables as covariates. Otherwise one can create bias due to unmeasured confounding. Do not interpret the effect

Randomization Draw lots between intervention groups.

Adjust Include the skew covariate in the model.

Overview

1 The Linear Model
- ANCOVA

2 Vital Capacity and Cadmium

3 Exercises

Vital Capacity and Cadmium

The model for vital capacity

$$Y_i = \beta_0 + \beta_{>10} X_{i,>10} + \beta_{<10} X_{i,<10} + \beta_{age} X_{i,age} + \epsilon_i$$

Here
$X_{i,>10} = 1$ if person i is exposed > 10 years 0 otherwise.
$X_{i,<10} = 1$ if person i is exposed < 10 years 0 otherwise.
$X_{i,age} =$ age of person i.

Exercise:

- Work in pairs. Online: Work with yourself ☺.
- Draw a sketch of how you envision the above model on a piece of paper.

Vital Capacity and Cadmium

We have a model with three parallel lines:

β_{age}	Common slope.
β_0	Intercept for not exposed
$\beta_0 + \beta_{<10}$	Intercept for exposed < 10 years
$\beta_0 + \beta_{>10}$	Intercept for exposed > 10 years

Model Check

- Normally distributed residuals $(y - \hat{y})$ (qq-plot).
- Independent observations.
- Variance homogeneity (residual plot).
- Linear effects (residual plots).

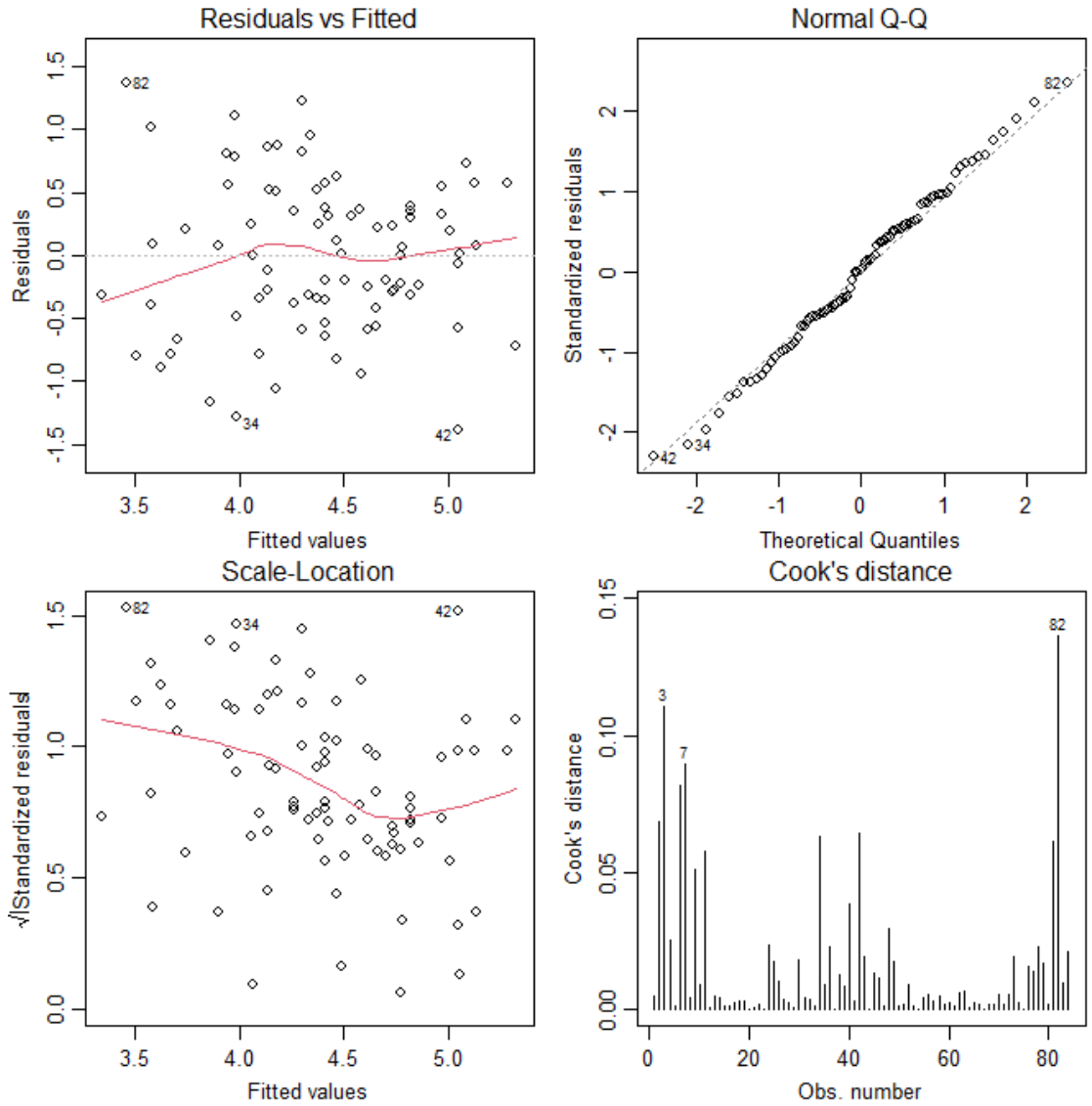
Assumption about Independence

A simple assessment: "Random sample", "Each individual only sampled once"

Model Check, using built in plot

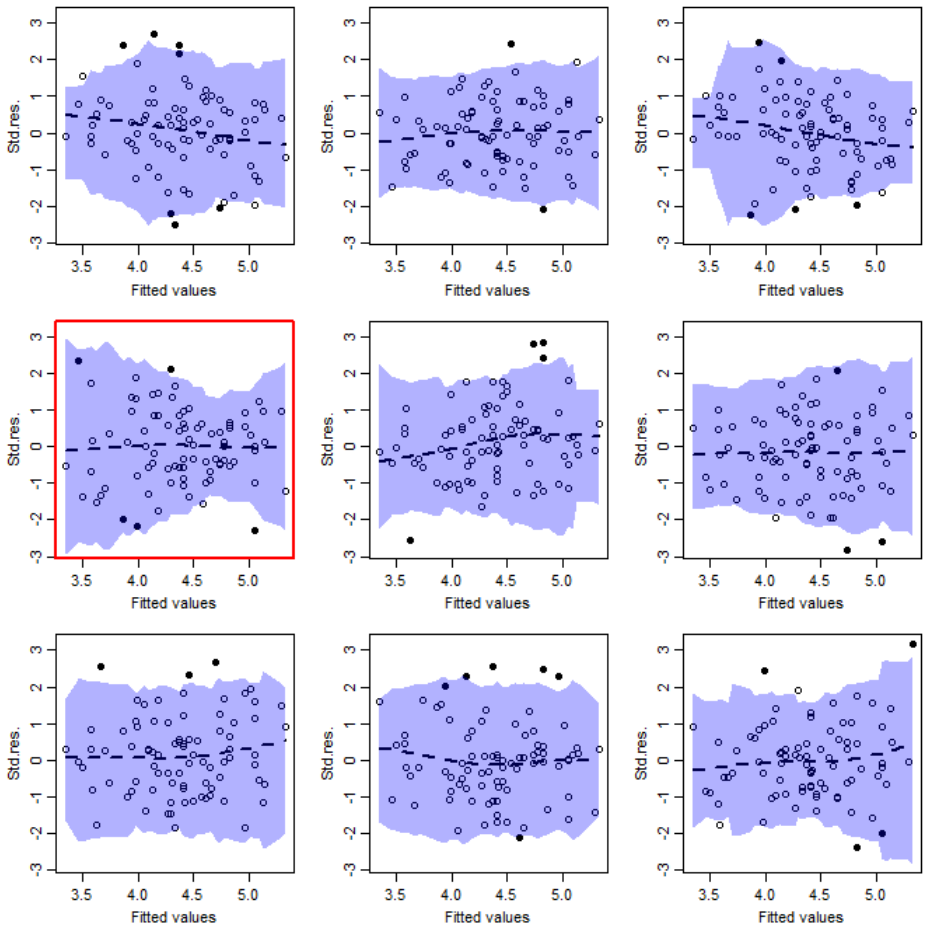
```
#EXPO WHERE NOT EXOPOSED 1, <10 IS EXPO==2, >10 is EXPO==3
CADdata$expo[CADdata$group==3] <- 1
CADdata$expo[CADdata$group==2] <- 2
CADdata$expo[CADdata$group==1] <- 3

#DECLARE EXPO AS A FACTOR
CADdata$expo<-as.factor(CADdata$expo)

#Initial model
Model1<-lm(vitcap ~ age + expo, data = CADdata)

#Model check
par(mfrow = c(2,2), mgp = c(2,0.7,0), mar = c(3,3,1.5,1))
plot(Model1, which = 1:4)
par(mfrow = c(1,1))
```

Model Check, using built in plot



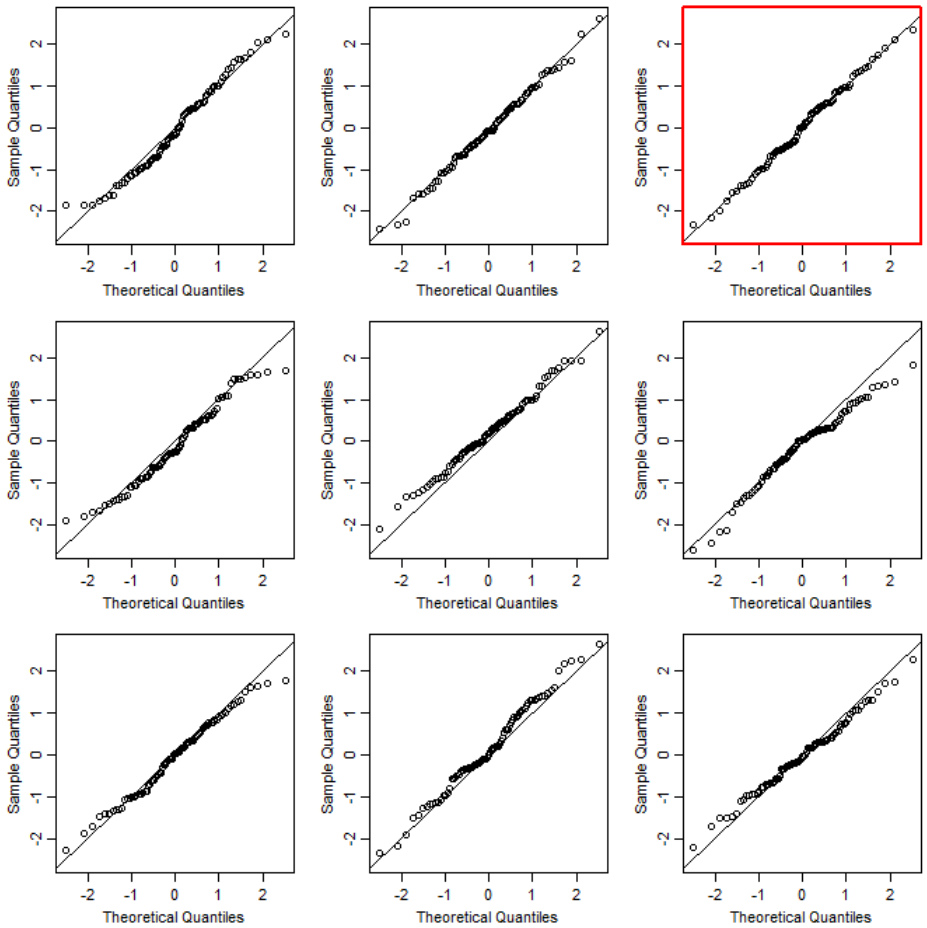
Extra plots if in doubt: Plot to check variance homogeneity

```
library(MESS)
wallyplot(Model1)
```



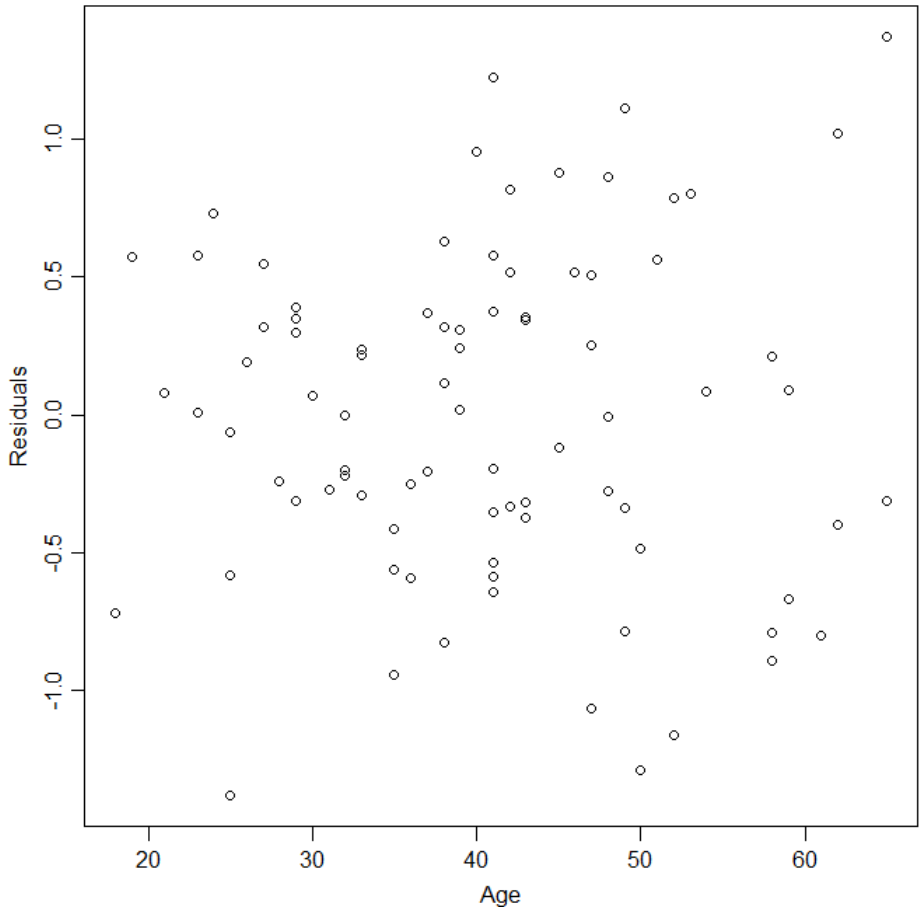
Plot to check normal residuals

```
qqwrap <- function(x, y, ...) {qqnorm(y,main="",...); abline(a=0, b=1)}
wallyplot(Model1, FUN=qqwrap)
```



Plot to check linearity of age

```
plot(CADdata$age, Model1$residuals, xlab = 'Age',
ylab = 'Residuals')
```

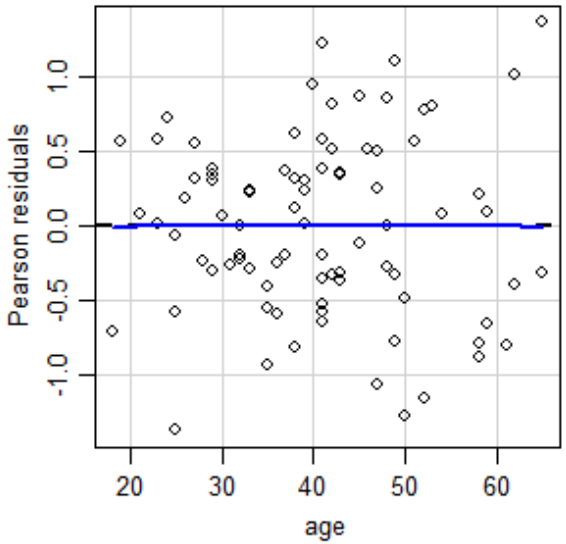


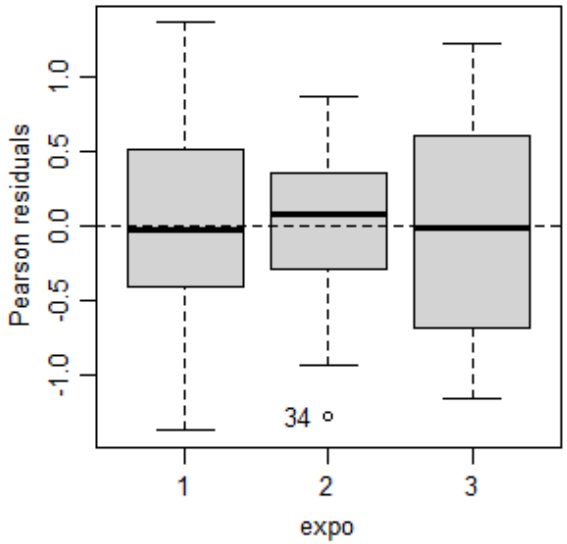
Plot to check linearity of age using library(car)

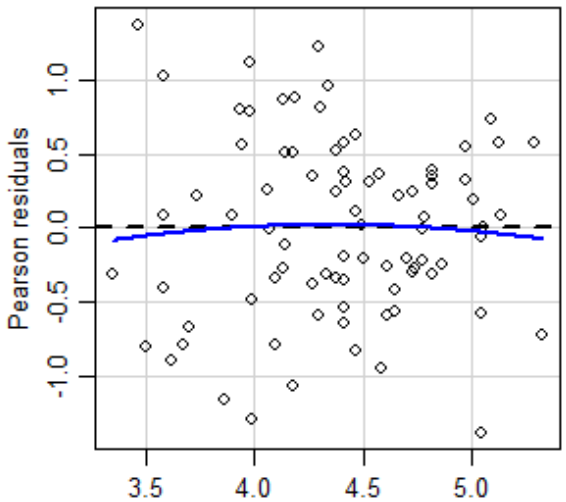
```
library(car)
residualPlots(Model1)
```

```
##               Test stat  Pr(>|t|)
## age             -0.0535    0.9575
## expo
## Tukey test      -0.3918    0.6952
```

Plot to check linearity of age using library(car)







The Model check went well

- Normally distributed residuals $(y - \hat{y})$ (qq-plot) (straight line).
- Independent observations. (Cannot check, have to assume).
- Variance homogeneity (residual plot, no trumpet).
- Linear effects (residual plots, looks random).
- We could also look for influential observations looking at Cook's distance.

Estimates

```
Model1 <- lm(vitcap ~ age + expo, data = CADdata)
summary(Model1)

## Coefficients:
##              Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)  6.044917    0.268025   22.554   < 2e-16 ***
## age         -0.039775    0.006322   -6.291  1.57e-08 ***
## expo2       -0.070198    0.148669   -0.472     0.638
## expo3       -0.116935    0.209236   -0.559     0.578
##
## Residual standard error: 0.6127 on 80 degrees of freedom
## Multiple R-squared:  0.3696,Adjusted R-squared:  0.3459
## F-statistic: 15.63 on 3 and 80 DF,  p-value: 4.323e-08
```

Table of results in R

```
confint(Model1)
```

```
##                   2.5 %      97.5 %
## (Intercept)  5.51153040  6.57830307
## age         -0.05235723 -0.02719313
## expo2       -0.36605755  0.22566252
## expo3       -0.53332814  0.29945819
```

```
# Nice table
tab <- cbind(coef(summary(Model1))[ , 1:2], "Lower" = confint(Model1)[ , 1],
             "Upper" = confint(Model1)[ , 2])

# Nice table with p-values
data.frame(round(tab, 2),
           "p-value" = format.pval(coef(summary(Model1))[ , 4], digits = 3, eps = 1e-3))
```

```
##             Estimate Std..Error Lower Upper p.value
## (Intercept)     6.04       0.27  5.51  6.58  <0.001
## age            -0.04       0.01 -0.05 -0.03  <0.001
## expo2          -0.07       0.15 -0.37  0.23   0.638
## expo3          -0.12       0.21 -0.53  0.30   0.578
```

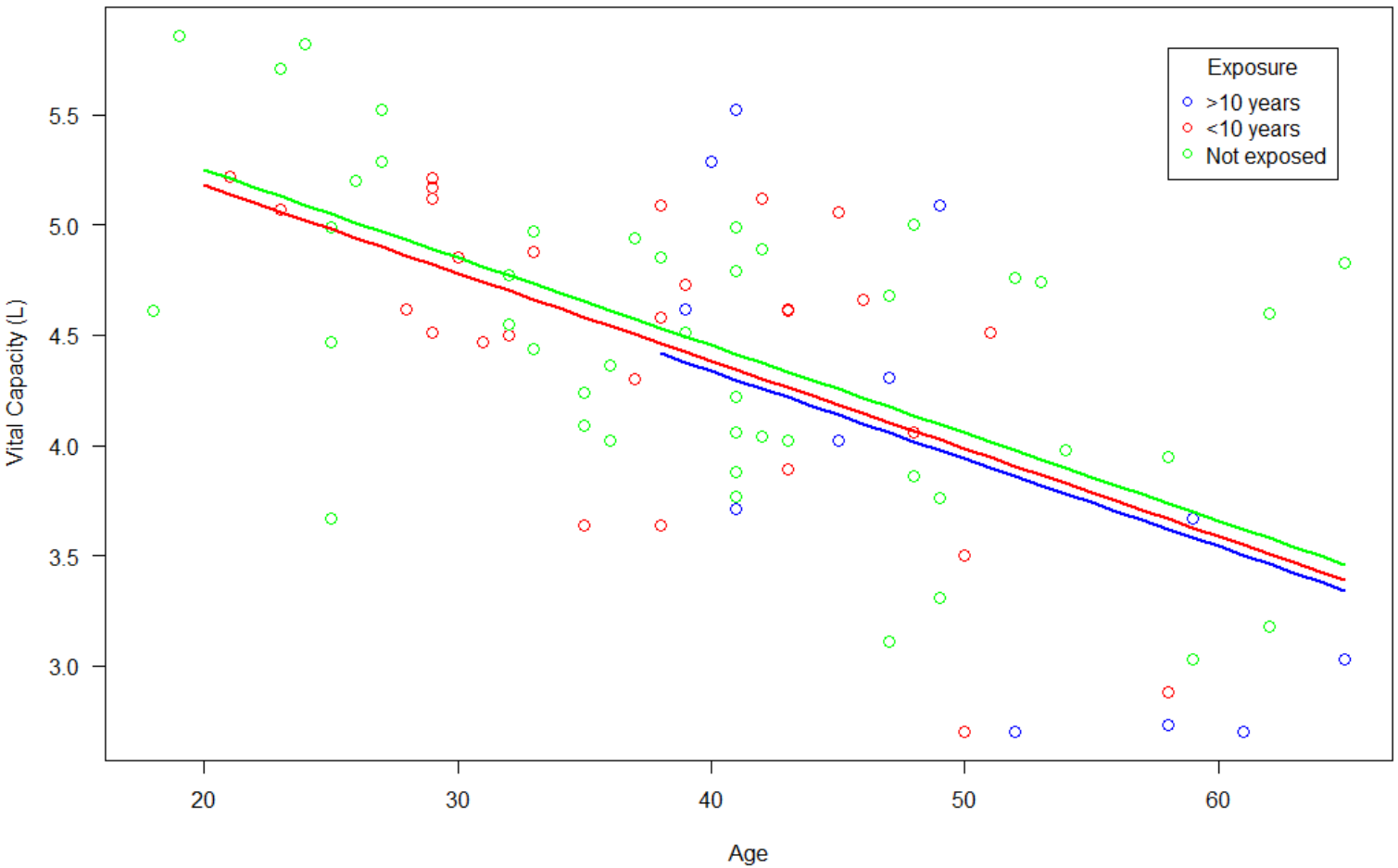
Estimates from the output

From the **R** output we got:

$\hat{\beta}_{age}$	$-0.04\ (-0.05; -0.03)$	(Common slope)
$\hat{\beta}_0$	$6.04\ (5.51; 6.58)$	(Intercept for not exposed)
$\hat{\beta}_{<10}$	$-0.07\ (-0.37; 0.23)$	(Extra intercept for exposed $<$ 10 years)
$\hat{\beta}_{>10}$	$-0.12\ (-0.53; 0.30)$	(Extra intercept for exposed $>$ 10 years)

And the variance $\sigma^2 = 0.613^2 = 0.376$.

Fitted Lines



Interaction

- The vital capacity decreases with -0.04 L per year.
- Is it reasonable that the vital capacity decreases with the same rate in all three exposure groups?
- Allow different slopes in the three groups $\rightarrow$ Include an interaction between age and group.

Estimates, from model with interaction

```
Model2 <- lm(vitcap ~ age + expo + age:expo, data = CADdata)
summary(Model2)

## Coefficients:
##              Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)  5.680291    0.313426   18.123   < 2e-16 ***
## age         -0.030613    0.007547   -4.066  0.000117 ***
## expo2        0.549740    0.575884    0.955  0.342728
## expo3        2.503148    1.041842    2.403  0.018655 *
## age:expo2   -0.015919    0.014547   -1.094  0.277170
## age:expo3   -0.054498    0.021070   -2.587  0.011554 *
##
## 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
```

Test Interaction

```
drop1(Model2, test = "F")

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

Test Interaction

```
drop1(Model2, test = "F")

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

So the interaction is statistically significant $0.03376 < 0.05$ and we need this more complex model.

The same model different parametrization

- We want to be able to get the three intercepts and slopes directly from the output.
- Notice the '0', says not to have common intercept.
- This parametrization not for testing the interaction but for understanding.

```
Model2B<-lm(vitcap ~ 0 + expo + age:expo, data = CADdata)
```

The same model different parametrization

```
summary(Model2B)

## Coefficients:
##             Estimate  Std. Error  t value  Pr(>|t|)
## expo1       5.680291    0.313426   18.123   < 2e-16 ***
## expo2       6.230031    0.483122   12.895   < 2e-16 ***
## expo3       8.183438    0.993579    8.2436 3.28e-12 ***
## expo1:age -0.030613    0.007547   -4.056  0.000117 ***
## expo2:age -0.046532    0.012436   -3.742  0.000347 ***
## expo3:age -0.085111    0.019672   -4.327  4.44e-05 ***
##
## Residual standard error: 0.5942 on 78 degrees of freedom
## Multiple R-squared:  0.9835,Adjusted R-squared:  0.9822
## F-statistic:  774.5 on 6 and 78 DF,  p-value: <2.2e-16

#confint(Model2B)
```

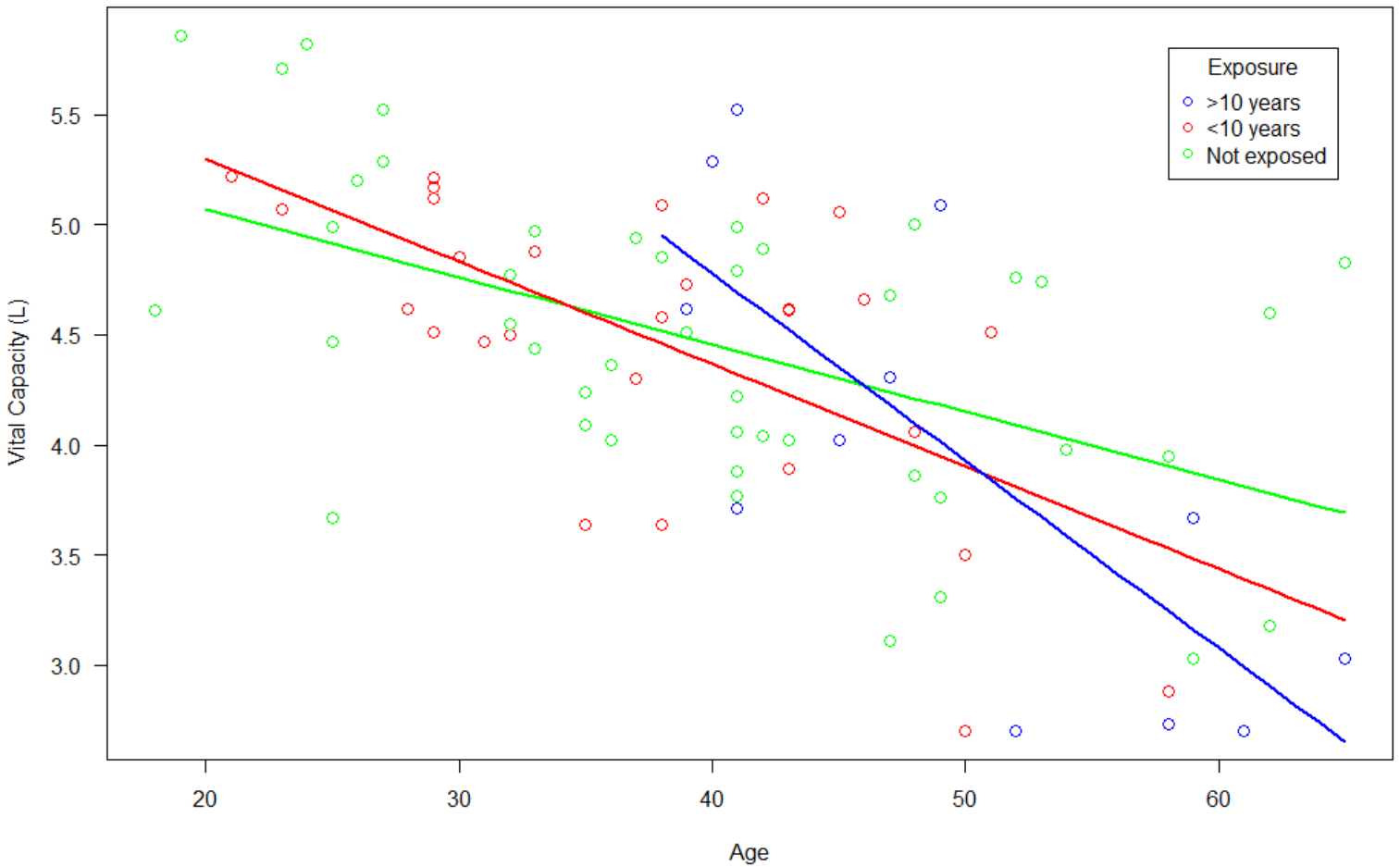
Estimates from the output with interaction

From the **R** output we got:

Intercept for not exposed	$5.68\ (5.06; 6.30)$
Slope for not exposed	$-0.03\ (-0.05; -0.02)$
Intercept for exposed $<$ 10 years	$6.23\ (5.27; 7.19)$
Slope for exposed $<$ 10 years	$-0.05\ (-0.07; -0.02)$
Intercept for exposed $>$ 10 years	$8.18\ (6.21; 10.16)$
Slope for exposed $>$ 10 years	$-0.09\ (-0.12; -0.05)$

And the variance $\sigma^2 = 0.594^2 = 0.353$.

Fitted Lines Interaction



Overview

1 The Linear Model
- ANCOVA

2 Vital Capacity and Cadmium

3 Exercises

Exercises

- Exercise 4: Prostate Cancer
- Exercise 5: Birth weight and gestation week