

Vitalkapacitet

```
library(mosaic)
```

I denne øvelse analyseres et datasæt vedrørende vitalkapacitet, som er den maksimale luftmængde, der kan udåndes efter en maksimal indånding. Datasættet kan hentes via dette link, og gemmes i samme mappe som denne Rmd-fil, så du kan køre tingene helt på din egen computer. Alternativt kan du også bare indlæse data direkte fra web-adressen, hvor det så hentes hver gang du kører kommandoen (f.eks. ved at trykke “knit”):

```
vitcap <- read.delim("http://asta.math.aau.dk/dan/static/datasets?file=vitcap.txt")
vitcap <- vitcap %>% mutate(exposure = as.factor(exposure))
```

I datasættet findes variabelen `vital.capacity`, som er blevet målt for 84 arbejdere i cadmiumindustrien.

En yderligere variabel er faktoren `exposure`, som har 3 niveauer, der angiver, hvor længe den enkelte arbejder har været eksponeret for cadmium:

- A: Ingen eksponering
- B: Mindre end 10 år
- C: Mere end 10 år

Derudover inderholder datasættet også variabelen `age`, der angiver den enkelte arbejders alder. (Og yderligere dummy-variablene `z1` og `z2`, som er beskrevet i ekstraopgaven, hvor de benyttes.)

Opgave 1 (ANOVA-analyse af exposures effekt på vitalkapacitet)

Lav en analyse, der undersøger effekten af faktoren `exposure` på responsen `vital.capacity`. Dvs. en lineær model på formen:

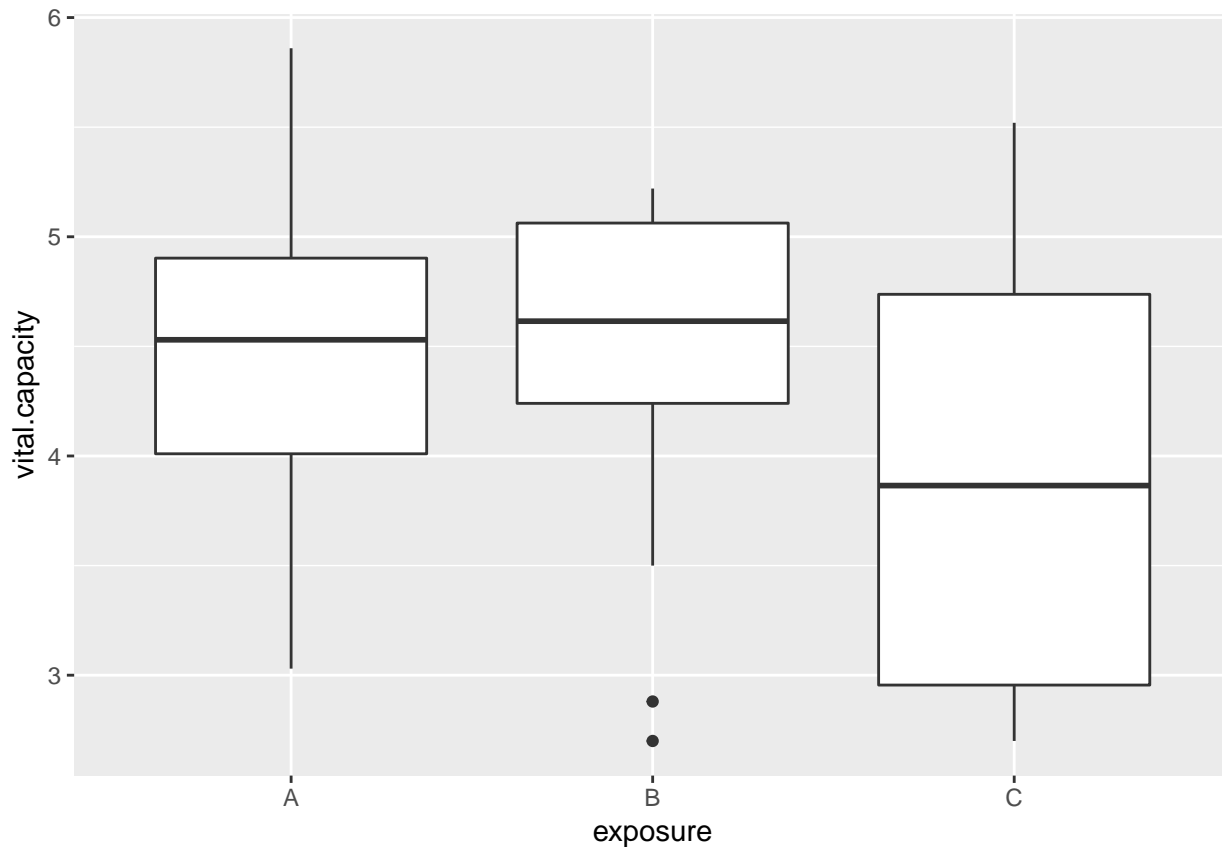
$$y = \alpha + \beta_1 z_1 + \beta_2 z_2 + \varepsilon$$

hvor z_1 og z_2 er dummy variable der indikerer om man er hhv. i gruppe B og C.

Du bør som minimum:

- Lave et relevant boxplot.

```
gf_boxplot(vital.capacity ~ exposure, data = vitcap)
```



- Lave summary af modellen.

```
modell1 <- lm(vital.capacity ~ exposure, data = vitcap)
summary(modell1)
```

```
##
## Call:
## lm(formula = vital.capacity ~ exposure, data = vitcap)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.77179 -0.45205  0.09808  0.51295  1.57083
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.46204    0.11223  39.757  <2e-16 ***
## exposureB    0.00974    0.17997   0.054  0.9570
## exposureC   -0.51288    0.24245  -2.115  0.0375 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7445 on 81 degrees of freedom
## Multiple R-squared:  0.05767,    Adjusted R-squared:  0.0344
## F-statistic: 2.478 on 2 and 81 DF,  p-value: 0.09021
```

- Opskriv prædiktionsligningen

$$\hat{y} = \hat{\alpha} + \hat{\beta}_1 z_1 + \hat{\beta}_2 z_2$$

hvor du indsætter tallene fra det tidligere summary for parameterestimerne.

$$\hat{y} = 4.46 + 0.01z_1 - 0.51z_2$$

- Hvad er den forventede forskel i lungekapacitet mellem personer i gruppe A og C? $\hat{\beta}_2 = -0.51$ dvs. vi forventer at lungekapaciteten er 0.51 **lavere** i gruppe C end i referencegruppen A.
- Gennemgå F-testen for ingen effekt af **exposure** (findes i slutningen af summary):
 - Hvad er nulhypotesen? ($H_0 : \beta_1 = \beta_2 = 0$, dvs. alle tre grupper har samme middel vitalkapacitet $\mu_A = \mu_B = \mu_C$. Bemærk at $\mu_A = \alpha$, $\mu_B = \alpha + \beta_1$ og $\mu_C = \alpha + \beta_2$).
 - Hvad er test-statistikken? ($F_{\text{obs}} = 2.48$)
 - Hvad er p-værdien? ($p = 0.09$)
 - Hvilken konklusion træffer vi om den samlede effekt af exposure på vitalkapaciteten? (Der er ikke nok evidens til at forkaste H_0 . Den samlede forskel mellem grupperne målt ved F-værdien er ikke større end det kan forklares ved tilfældig samplingvariation.)

Opgave 2 (Additiv model for alders og exposures effekt på vitalkapacitet)

Vi udvider nu analysen til også at omfatte arbejdernes alder (**age**) som en forklarende variabel for responsen `vital.capacity`.

Tilpas en model hvor **age** og **exposure** indgår additivt:

$$y = \alpha + \beta_1 z_1 + \beta_2 z_2 + \beta_3 x + \varepsilon$$

hvor x er alder og z_1 og z_2 angiver hhv. gruppe B og C som før.

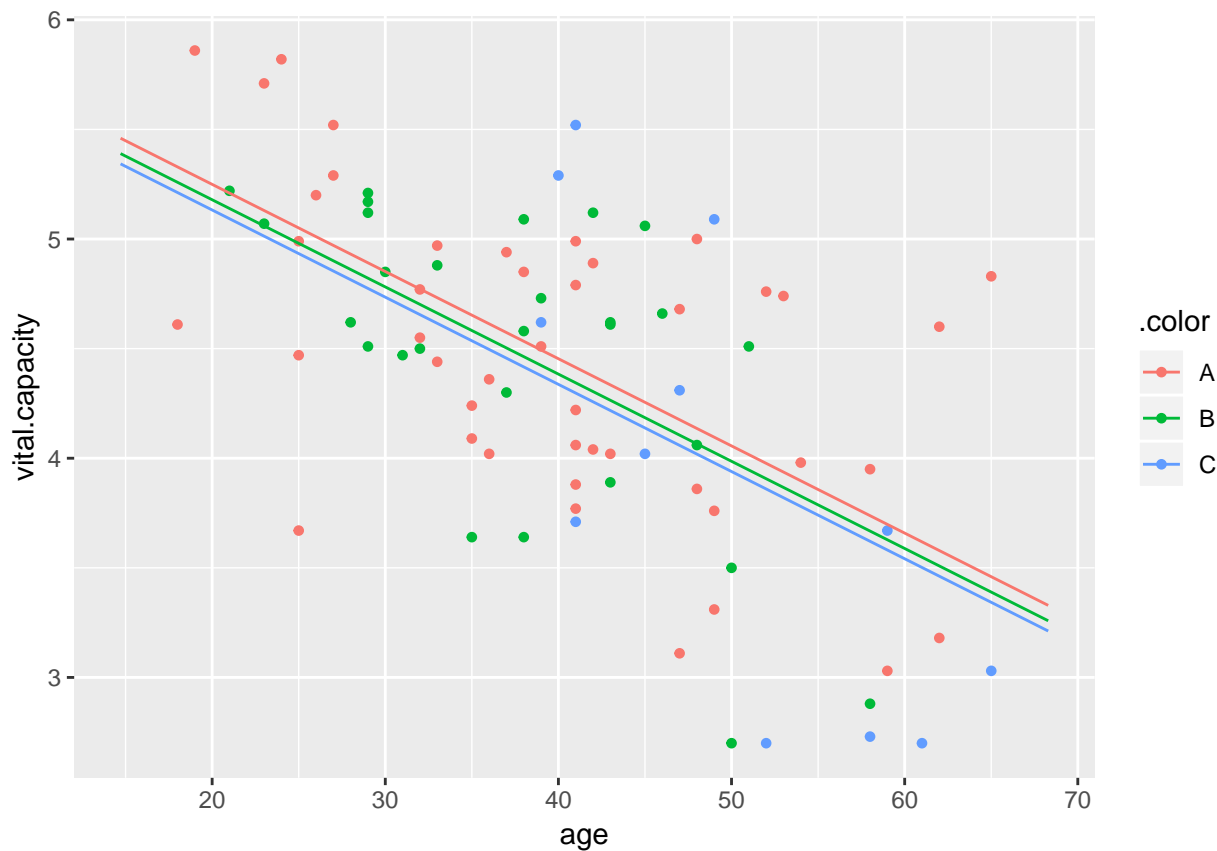
```
model2 <- lm(vital.capacity ~ age + exposure, data = vitcap)
summary(model2)
```

```
##
## Call:
## lm(formula = vital.capacity ~ age + exposure, data = vitcap)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.38054 -0.38050  0.01321  0.37909  1.37047
##
## 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 ***
## exposureB   -0.070198   0.148669  -0.472  0.638
## exposureC   -0.116935   0.209236  -0.559  0.578
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## 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
```

Gennemgå detaljerne for denne model:

- Illustrer modellen grafisk med f.eks. `plotModel()`. (Tip: Vær opmærksom på, at `plotModel` forventer, at den definerede model har den kontinuerte variabel `age` til at stå først i formeludtrykket. Så hvis man vil plote regressionslinjerne for den additive model, så skal man bruge formeludtrykket `vital.capacity ~ age + exposure`, når man tilpasser sin model, og IKKE `vital.capacity ~ exposure + age`. De to formeludtryk repræsenterer ganske vist præcis den samme additive model, men den bagvedliggende kode i funktionen `plotModel` virker desværre kun efter hensigten, hvis `age` står først.)

```
plotModel(model2)
```



- Med udgangspunkt i outputtet af et `summary` for modellen:
 - Indsæt talværdier i den samlede prædiktionsligning.

$$\hat{y} = \hat{\alpha} + \hat{\beta}_1 z_1 + \hat{\beta}_2 z_2 + \hat{\beta}_3 x$$

$$\hat{y} = 6.04 - 0.07z_1 - 0.12z_2 - 0.04x$$
 - Opskriv prædiktionsligningerne for hver af de tre grupper (hint: indsæt værdierne for dummyvariable og lad led med 0 udgå).

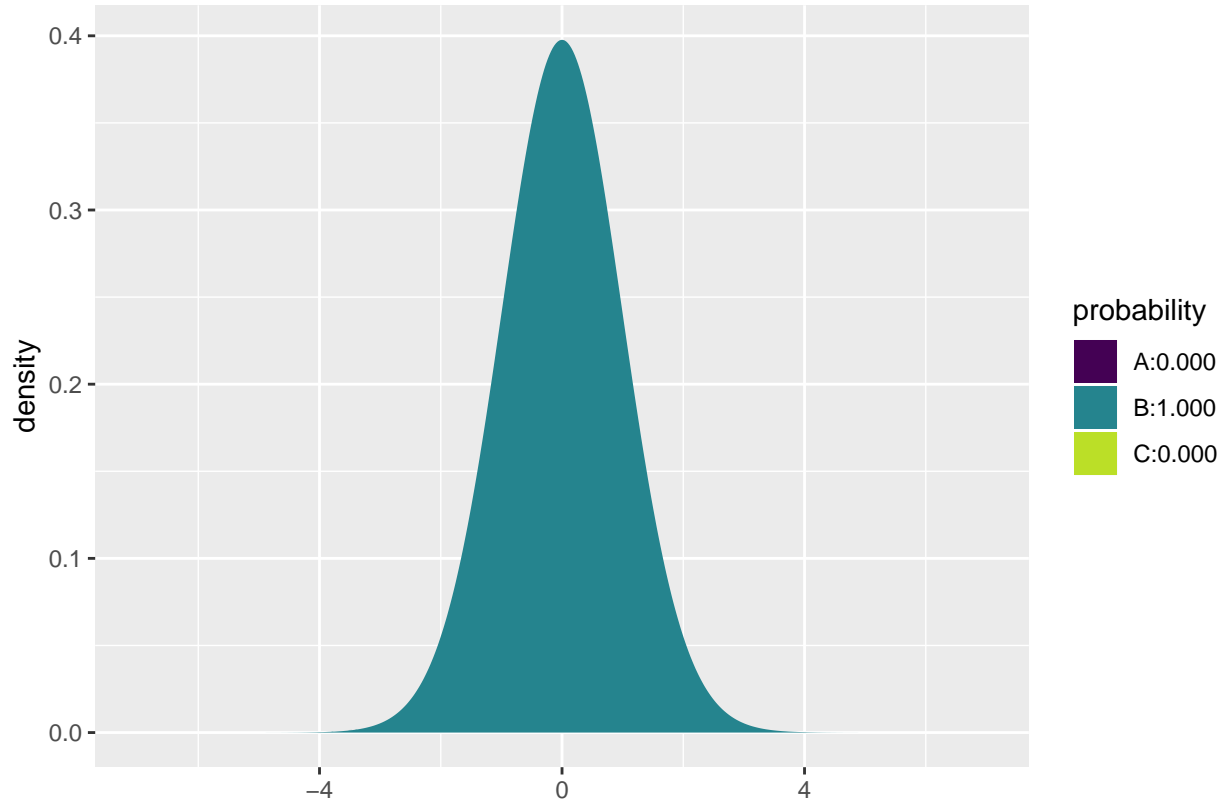
$$A : \hat{y} = 6.04 - 0.04x$$

$$B : \hat{y} = 6.04 - 0.07 - 0.04x = 5.97 - 0.04x$$

$$C : \hat{y} = 6.04 - 0.12 - 0.04x = 5.92 - 0.04x$$
 - Hvad er den forventede vitalkapacitet for en arbejder på 40 år i gruppe B? ($5.97 - 0.04 \cdot 40 = 4.38$ med afrunding.)
 - Hvor meget forventes vitalkapaciteten at ændre sig årligt for en arbejder i gruppe C? Hvad med for dem i gruppe A og B? (For gruppe C er den forventede ændring per år $\hat{\beta}_3 = -0.04$. Den er den samme for gruppe A og B da der ikke er interaktion.)

- Er effekten af alder signifikant? Gennemgå herunder teststørrelse og p-værdi og hvordan disse hænger sammen. (Ja, $t_{\text{obs}} = \hat{\beta}_3 / (se) = -0.04 / 0.0063 = -6.29$ giver en to-halet p-værdi i t-fordelingen med 80 frihedsgrader på 1.57×10^{-8} .)

```
pdist("t", df = 80, q = c(-6.26, 6.26), xlim = c(-7,7))
```



```
## [1] 8.988382e-09 1.000000e+00
```

+ Er der en signifikant forskel mellem gruppe A og B? Hvad med A og C?

Nej, parallelforskydningerne mellem A og B samt mellem A og C er ikke sigifikante som ses af p-værdierne

Opgave 3 (Interaktionsmodel for alders og exposures effekt på vitalkapacitet)

Vi udvider nu analysen til også at tillade en interaktion mellem alders og exposures effekt på vitalkapaciteten.

Tilpas en model hvor age og exposure indgår med interaktionseffekt

$$y = \alpha + \beta_1 z_1 + \beta_2 z_2 + \beta_3 x + \beta_4 z_1 x + \beta_5 z_2 x + \varepsilon$$

hvor x er alder og z_1 og z_2 angiver hhv. gruppe B og C som før.

```
model3 <- lm(vital.capacity ~ age * exposure, data = vitcap)
summary(model3)
```

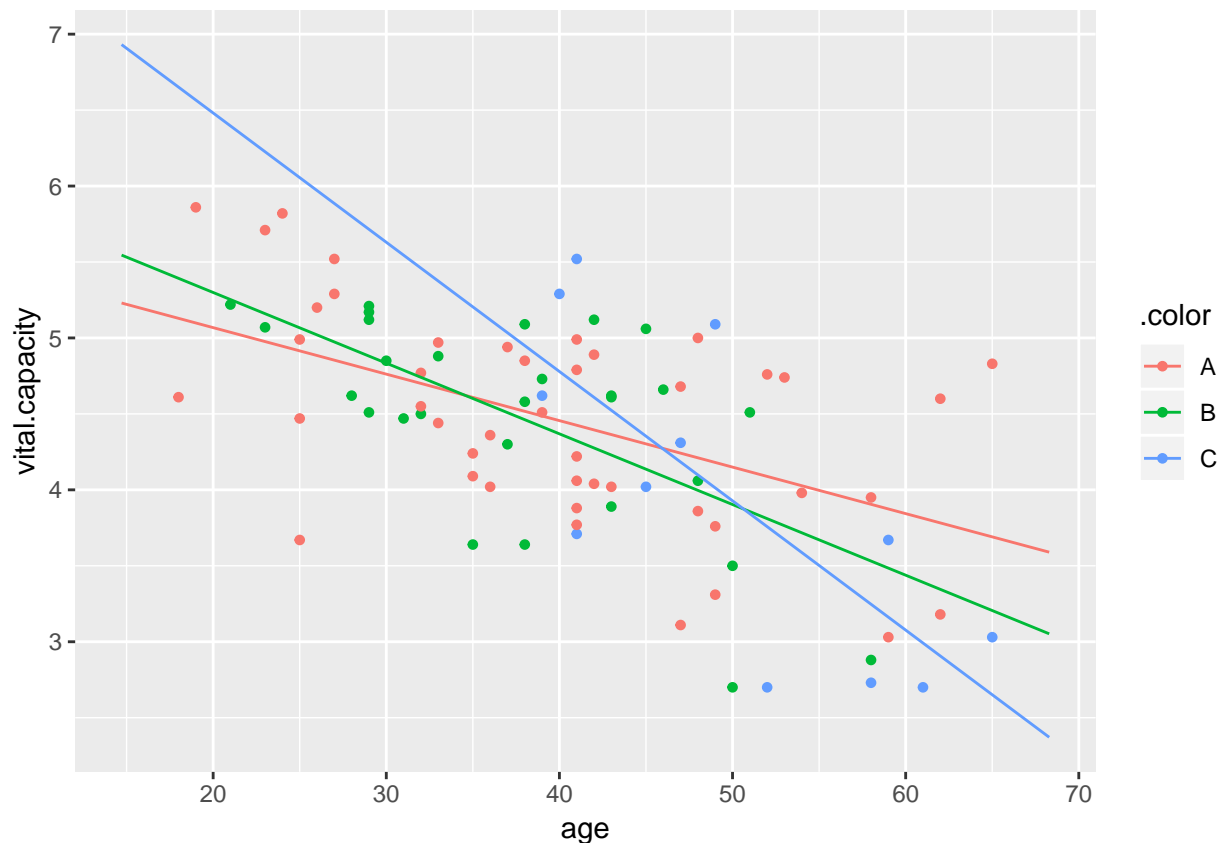
```
##
## Call:
```

```
## lm(formula = vital.capacity ~ age * exposure, data = vitcap)
##
## 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)   5.680291   0.313426  18.123 < 2e-16 ***
## age           -0.030613   0.007547  -4.056 0.000117 ***
## exposureB     0.549740   0.575884   0.955 0.342728
## exposureC     2.503148   1.041842   2.403 0.018655 *
## age:exposureB -0.015919   0.014547  -1.094 0.277170
## age:exposureC -0.054498   0.021070  -2.587 0.011554 *
## ---
## 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
```

Gennemgå detaljerne for denne model:

- Illustrer modellen grafisk med f.eks. `plotModel()` som før.

```
plotModel(model13)
```



- Med udgangspunkt i outputtet af et `summary` for modellen:
 - Indsæt talværdier i den samlede prædiktionsligning.

$$\hat{y} = \hat{\alpha} + \hat{\beta}_1 z_1 + \hat{\beta}_2 z_2 + \hat{\beta}_3 x + \hat{\beta}_4 z_1 x + \hat{\beta}_5 z_2 x$$

$$\hat{y} = 5.68 + 0.55z_1 + 2.50z_2 - 0.03x - 0.02z_1x - 0.05z_2x$$

- Opskriv prædiktionsligningerne for hver af de tre grupper (hint: indsæt værdierne for dummyvariable og lad led med 0 udgå).

$$A : \hat{y} = 5.68 - 0.03x$$

$$B : \hat{y} = 5.68 + 0.55 - 0.03x - 0.02x = 6.23 - 0.05x$$

$$C : \hat{y} = 5.68 + 2.50 - 0.03x - 0.05x = 8.18 - 0.08x$$

- Hvad er den forventede vitalkapacitet for en arbejder på 40 år i gruppe B? ($6.23 - 0.05 \cdot 40 = 4.37$ med afrunding)
- Hvor meget forventes vitalkapaciteten at ændre sig årligt for en arbejder i gruppe C? Hvad med for dem i gruppe A og B? (Gr C: $\hat{\beta}_3 + \hat{\beta}_5 = -0.03 - 0.05 = -0.08$. Gr A: $\hat{\beta}_3 = -0.03$. Gr B: $\hat{\beta}_3 + \hat{\beta}_4 = -0.03 - 0.02 = -0.05$)
- Er der signifikant forskel i effekten af alder mellem gruppe A og B? Hvad med mellem A og C? (Nej, ikke forskel mellem A og B pga. p-værdi på 0.277. Ja, forskel mellem A og C pga. p-værdi på 0.012)
- Brug `anova()` til at undersøge om der samlet set er en signifikant interaktion mellem effekten af exposure og alder på vitalkapaciteten.

```
anova(model2, model3)
```

```
## Analysis of Variance Table
##
## Model 1: vital.capacity ~ age + exposure
## Model 2: vital.capacity ~ age * exposure
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1      80 30.035
## 2      78 27.535  2    2.4995 3.5402 0.03376 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(Vi afviser $H_0 : \beta_4 = \beta_5 = 0$ med p-værdi på 0.03, dvs. der er signifikant interaktion mellem effekten af alder og exposure på vitalkapaciteten)

Opfølgende spørgsmål:

Hvordan kan det være, at det i opgavens første model (uden `age` som forklarende variabel) ser ud som om, at vi kan tillade os at slå de tre `exposure`-grupper sammen, mens det modellerne med `age` som forklarende variabel ikke ser sådan ud?

En måde at anskue fænomenet på er ved at starte med at betragte den mest komplekse model med interaktionen mellem `age` og `exposure`. Her er `age` meget signifikant og ser altså ud til at forklare `vital.capacity` bedre end `exposure`, men gennem forrige F-test af interaktionens signifikans, så ser det samtidig ud til, at der er en interaktion, hvorfor `exposure` også er en nødvendig variabel i modellen. Vælger vi nu at smide den meget signifikante variabel `age` ud af vores model, så får vi `model1`, der nødvendigvis har en ringere forklaringskraft - man kan sige, at dens forklaringskraft er så ringe, at den ikke bliver væsentligt dårligere af, at man slår alle tre `exposure`-grupper sammen til én gruppe.

Ekstraopgave (hvis du har tid)

Datasættet indeholder også dummy-variabler for faktoren `exposure`:

- `z1=1` hvis `exposure=B` ellers 0
- `z2=1` hvis `exposure=C` ellers 0

Betragt følgende to modeller:

```
model4 <- lm(vital.capacity ~ age*z2, data = vitcap)
model5 <- lm(vital.capacity ~ age*z1 + age*z2, data = vitcap)
```

- Brug en F-test til at vise, at der ikke er nogen signifikant forskel mellem `model4` og `model5` (Hint: bruge `anova()`).

```
anova(model4, model5)
```

```
## Analysis of Variance Table
##
## Model 1: vital.capacity ~ age * z2
## Model 2: vital.capacity ~ age * z1 + age * z2
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1      80 28.020
## 2      78 27.535  2   0.48468 0.6865 0.5064
```

- Giv en fortolkning af forskellen mellem de to modeller.

Ved hjælp af dummy-variablene `z1` og `z2` kan vi inddele data i `exposure`-grupperne, og `model4` svarer således til, at gruppe A og B er blevet slået sammen til én samlet gruppe, mens `model5` faktisk er identisk med vores tidligere `model13`. For at lave ovenstående F-test kunne vi således lige så godt have skrevet `anova(model4, model13)`, og hvis vi udskriver `summary(model5)`, så får vi nøjagtigt de samme estimater, som da vi udskrev `summary(model13)` (koefficienternes navne er dog ændret til dummy-variablenes navne).

I forbindelse med det sidste spørgsmål kan du lave følgende plot:

```
gf_point(vital.capacity ~ age, data = vitcap, color = ~ factor(z2)) %>%
  gf_smooth(method = "lm", se = F) %>%
  gf_labs(color = "Gruppe C")
```