Intro til statistik

III Lineære modeller II

Claus Thorn Ekstrøm
Biostatistik, KU
ekstrom@sund.ku.dk

Tirsdag 12. maj 2020
Slides @ biostatistics.dk/puff/

Plan for i dag

- Multipel lineær regression
- Vekselvirkninger
- Modelkontrol

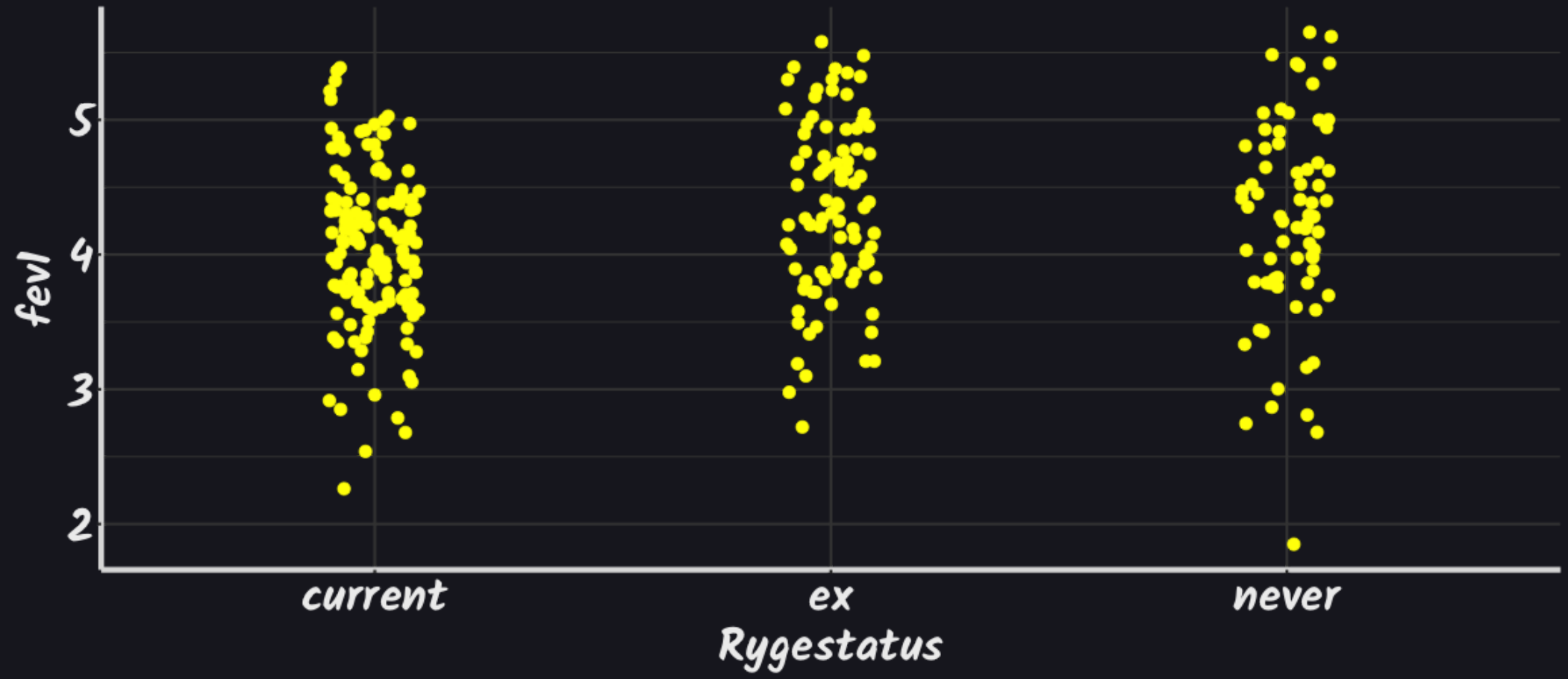
Krigsveteraners helbredsforhold

```
head(vet)
```

```
##   idnr alder bpright bpleft fev1 alkuge   ryger
## 1    1 37.33     120    119 5.23      0      ex
## 2    2 38.24     129    122 3.94      0 current
## 3    3 39.24     123    123 4.42      1 current
## 4    4 35.45     103     95 4.12      6 current
## 5    5 38.70     107    110 4.12      0      ex
## 6    6 38.62     121    124 3.21      0      ex
```

Sammenhæng mellem rygning og lungekapacitet.

Kategoriske forklarende variable



Kategoriske forklarende variable

Laver nye x'er:

$$x_1 = \begin{cases} 1 & \text{hvis "ex"} \\ 0 & \text{ellers} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{hvis "never"} \\ 0 & \text{ellers} \end{cases}$$

Og så multipel regression.

Multipel regression

Udvider den lineære model:

$$y_i = \beta_0 + \sum_{k=1}^{K} \beta_k \cdot x_{ki} + \varepsilon_i$$

Implicit antagelse: effekten af prædiktorerne, x_k, kan lægges sammen.

Fortolkning af estimaterne

Som for lineær regression:

> Hvad er den gennemsnitlige effekt på udfaldet, når variablen ændres *en* enhed (dog kun fra 0 til 1)

En kategori er "referencegruppen". Samme rolle som skæringen.

Krigsvetaraner

```
model <- lm(fev1 ~ ryger, data=vet)
library("broom")
model %>% tidy()
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    4.06     0.0553     73.4  4.90e-192
## 2 rygerex        0.290    0.0885      3.28 1.16e-  3
## 3 rygernever     0.169    0.0951      1.78 7.68e-  2
```

Krigsveteraner

```
confint(model)
```

```
##                   2.5 %    97.5 %
## (Intercept)  3.95042170 4.1679956
## rygerex      0.11610936 0.4643498
## rygernever  -0.01827333 0.3559124
```

Trick: `factor()` tvinger R til at opfatte en variabel som kategorisk.

Krigsveteraner

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

```
## Single term deletions
##
## Model:
## fev1 ~ ryger
##        Df Sum of Sq    RSS     AIC F value   Pr(>F)
## <none>              125.72 -253.05
## ryger   2    4.7373 130.46 -245.99  5.5769 0.004193 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Multipel regression

```
## `geom_smooth()` using formula 'y ~ x'
```

Multipel regression

Konfundering pga alder?

```
m2 <- lm(fev1 ~ ryger + alder, data=vet)
m2 %>% tidy()
```

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   5.60      0.585      9.59 4.02e-19
## 2 rygerex       0.315     0.0881     3.57 4.13e- 4
## 3 rygernever    0.198     0.0948     2.09 3.72e- 2
## 4 alder        -0.0406    0.0153    -2.65 8.41e- 3
```

Multipel regression

```
confint(m2)
```

```
##                    2.5 %      97.5 %
## (Intercept)  4.45278253  6.75351556
## rygerex      0.14127027  0.48790581
## rygernever   0.01181631  0.38482463
## alder       -0.07079412 -0.01049376
```

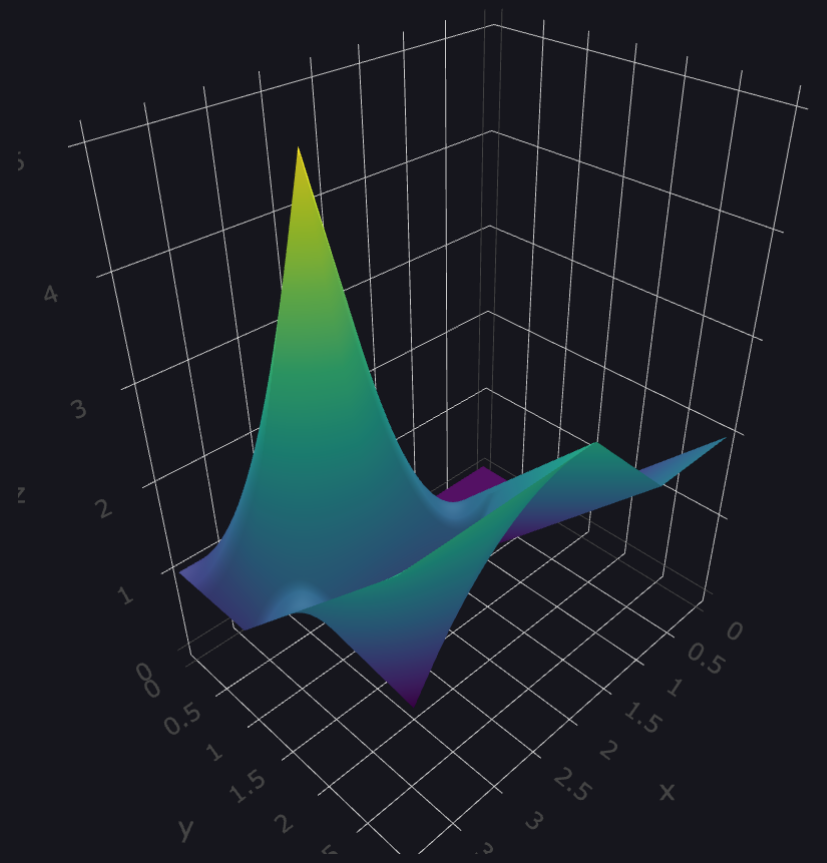
Prædiktioner

Forventet lungekapacitet for en 35-årig ex-ryger?

Forventet lungekapacitet for en 30-årig ryger?

Vekselvirkninger

En vekselvirkning (*interaction* eller *effektmodifikation*) er, når effekten af en variabel afhænger af værdien af en anden variabel.



Vekselvirkninger i R

```
m3 <- lm(fev1 ~ ryger*alder, data=vet)
m3 %>% tidy()
```

```
## # A tibble: 6 x 5
##   term             estimate std.error statistic  p.value
##   <chr>               <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)       5.66       0.818     6.93   2.75e-11
## 2 rygerex           0.212      1.35      0.157  8.75e- 1
## 3 rygernever        0.0295     1.63      0.0181 9.86e- 1
## 4 alder            -0.0423     0.0215   -1.97   5.01e- 2
## 5 rygerex:alder     0.00270    0.0350    0.0771 9.39e- 1
## 6 rygernever:alder  0.00439    0.0423    0.104  9.17e- 1
```

Prædiktioner

Forventet lungekapacitet for en 35-årig ex-ryger?

Forventet lungekapacitet for en 30-årig ryger?

Vekselvirkninger

```
confint(m3)
```

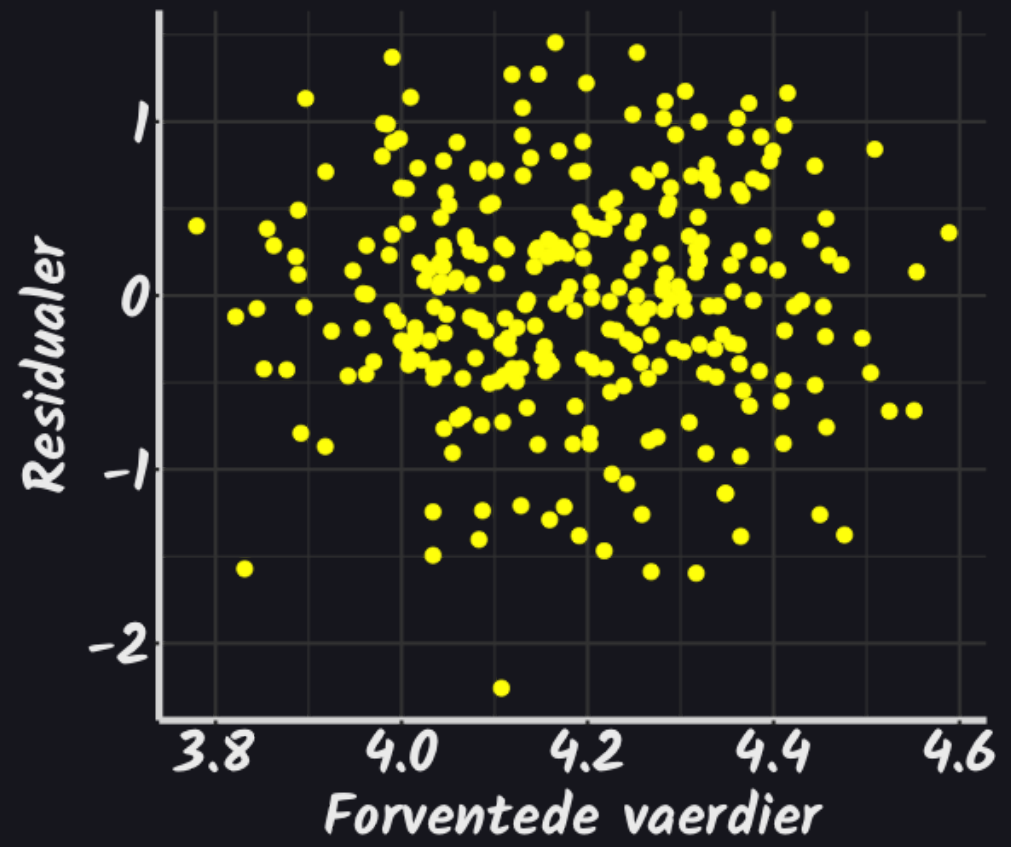
```
##                       2.5 %       97.5 %
## (Intercept)       4.05488981 7.274321e+00
## rygerex          -2.43575198 2.858776e+00
## rygernever       -3.18170532 3.240711e+00
## alder            -0.08454163 1.809803e-05
## rygerex:alder    -0.06616454 7.155742e-02
## rygernever:alder -0.07881932 8.760152e-02
```

Vekselvirkninger

```
drop1(m3, test="F")
```

```
## Single term deletions
##
## Model:
## fev1 ~ ryger * alder
##             Df Sum of Sq    RSS     AIC F value Pr(>F)
## <none>                   122.78 -254.11
## ryger:alder  2 0.0054464 122.79 -258.10  0.0065 0.9935
```

Modelkontrol - er antagelserne opfyldte?



1. Middelværdi ca. 0
2. Ingen systematiske afvigelser
3. Check for outliers
4. Varianshomogenitet
5. (Uafhængighed - kan ikke nødvendigvis ses)

Modelkontrol i R

```
library(MESS)
residualplot(m3)
wallyplot(m3)
```