Solution in R: Exercises Day 2

PSY8003

Matthias Mittner

spring 2022

Exercise 1: Hierarchical regression

The second block of predictors improves the model significantly, the third does not. The coefficients are not much changed by including the additional predictors.

```
workout <- haven::read_dta("../data/workout.dta")

mod1 <- lm(whours ~ gender + age, data=workout)
mod2 <- lm(whours ~ gender + age + educ + marital, data=workout)
mod3 <- lm(whours ~ gender + age + educ + marital + health, data=workout)
anova(mod1,mod2,mod3)
```

```
Analysis of Variance Table

Model 1: whours ~ gender + age
Model 2: whours ~ gender + age + educ + marital
Model 3: whours ~ gender + age + educ + marital + health
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1    207 10070.6
2    205  9680.8  2    389.86 4.1212 0.0176 *
3    204  9649.0  1     31.72 0.6707 0.4138
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coef(mod1)
```

```
(Intercept)      gender         age
 17.2313102   2.2173022  -0.1339799
```

```r
coef(mod2)
```

```
(Intercept)      gender         age        educ     marital
19.53729471  2.05090395 -0.10505209 -1.64263656  0.03446252
```

```r
coef(mod3)
```

```
(Intercept)      gender         age        educ     marital      health
17.84223404  2.11800133 -0.11336462 -1.58764121  0.01942437  0.36534537
```

Exercise 2: Power analysis

a)

In G*Power:

- Test family=F-Test
- Statistical-Test = Linear multiple Regression: Fixed model, R^2 deviation from 0
- Type: A priori
- Input parameters: - "Determine" -> From correlation coefficient: Squared multiple correlation=0.1 - α err prob=0.05 - Power=0.95 - Number of predictors=9

Total sample size=221.

In R:

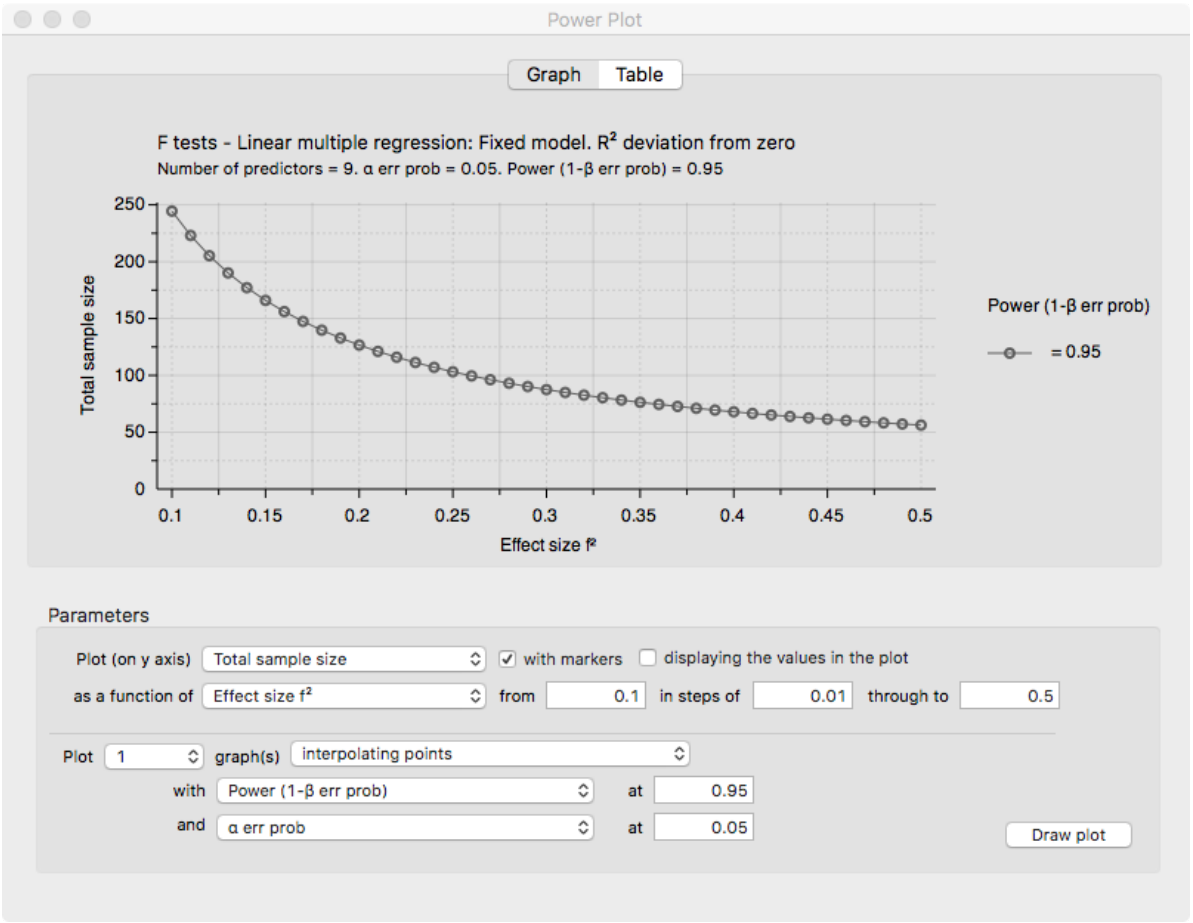
```r
library(pwr)
R2=0.1
k=9
f2=R2/(1-R2)
pwr.f2.test(u=k, f2=f2, power = 0.95, sig.level = 0.05)
```

```
     Multiple regression power calculation

              u = 9
              v = 210.9213
             f2 = 0.1111111
      sig.level = 0.05
          power = 0.95
```

```
N=211+k+1
N
```

```
[1] 221
```

b)



The critical f^2 for an $R^2 = 0.2$ is

$$f^2 = \frac{R^2}{1 - R^2} = 0.25$$

The graph "cuts" the $f^2 = 0.25$ line at about $N = 103$ (click on "Table" to get exact values).

c)

$N = 150$ corresponds to about $f^2 = 0.17$ which is

$$R^2 = \frac{f^2}{f^2 + 1} \approx 0.15$$

d)

In G*Power:

- Test family=F-Test
- Statistical-Test = Linear multiple Regression: Fixed model, R^2 increase
- Type: A priori
- Input parameters: - f^2=0.02 - α err prob=0.05 - Power=0.8 - Number of tested predictors=1 - Number of predictors=9 - Calculate -> $N = 395$

Exercise 3: Regression diagnostics

Two cases have large residuals >3; all of them occur for very high values of the dependent variables -> indicative of a problem with specification?

Leverage/Cooks-d pick up a few cases, but it does not look too severe (no Cooks'd close to 1).

No problem with multicollinearity.

There may be a slight misspecification (deviation from linearity), but its not severe.

Heteroscedasticity is present, variance seems to grow with the predicted value.

The residuals are definitely not normal, not surprising considering that this is a count-variable.

```
workout <- haven::read_dta("../data/workout.dta")

mod <- lm(whours ~ educ + gender + age + marital + health, data=workout)

# any residuals larger/smaller than expected?
workout |>
```

```
    mutate(resid=rstandard(mod)) |>
    filter(abs(resid)>2.5)
```

```
# A tibble: 5 x 7
  whours    gender   age               educ     marital             health resid
   <dbl> <dbl+lbl> <dbl>          <dbl+lbl>   <dbl+lbl>          <dbl+lbl> <dbl>
1     48 1 [men]      20 1 [secondary/high] 1 [single]  4 [4]                4.49
2     40 1 [men]      20 1 [secondary/high] 1 [single]  5 [5]                3.25
3     36 1 [men]      21 1 [secondary/high] 1 [single]  5 [5]                2.68
4     32 0 [women]    27 1 [secondary/high] 0 [married] 4 [4]                2.57
5     32 1 [men]      38 2 [university]     0 [married] 6 [6=Very importa~  2.55
```

```
  # corresponding leverage/cooks-d a problem?
  workout |>
    mutate(resid=rstandard(mod),
           leverage=hatvalues(mod),
           cooksd=cooks.distance(mod)) |>
    filter(abs(leverage)>(2*5+2)/210) |>
    arrange(desc(leverage))
```

```
# A tibble: 8 x 9
  whours    gender   age        educ  marital     health  resid leverage  cooksd
   <dbl> <dbl+lbl> <dbl>   <dbl+lbl> <dbl+lb>  <dbl+lbl>  <dbl>    <dbl>   <dbl>
1     12 0 [women]    17 1 [seconda~ 1 [sing~ 1 [1=Not ~ -0.414   0.0941 0.00297
2      8 0 [women]    43 2 [univers~ 0 [marr~ 1 [1=Not ~ -0.329   0.0899 0.00178
3     24 1 [men]      16 1 [seconda~ 1 [sing~ 1 [1=Not ~  1.08    0.0893 0.0189
4      4 0 [women]    28 3 [more th~ 1 [sing~ 1 [1=Not ~ -0.958   0.0887 0.0149
5     16 0 [women]    34 2 [univers~ 1 [sing~ 1 [1=Not ~  0.731   0.0879 0.00859
6     16 1 [men]      76 3 [more th~ 0 [marr~ 4 [4]       1.20    0.0757 0.0198
7      4 1 [men]      47 1 [seconda~ 1 [sing~ 3 [3]      -1.52    0.0581 0.0238
8     12 0 [women]    37 3 [more th~ 0 [marr~ 2 [2]       0.357   0.0574 0.00129
```

```
  # multicollinearity
  library(car)
```

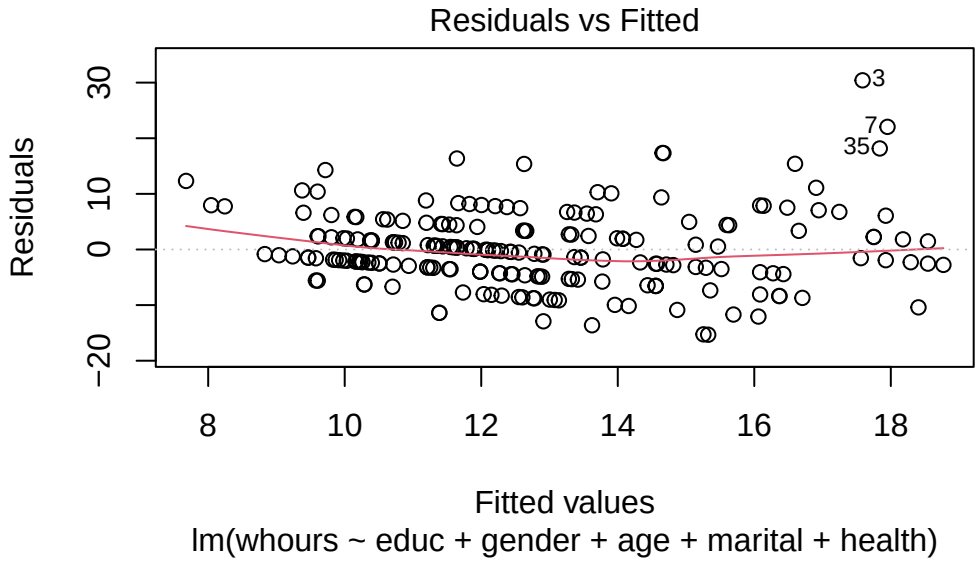
```
Loading required package: carData
```

```
Attaching package: 'car'

The following object is masked from 'package:dplyr':

    recode

The following object is masked from 'package:purrr':

    some
```

```r
vif(mod)
```

```
    educ   gender      age  marital   health 
1.100937 1.029322 1.507548 1.453946 1.070800 
```

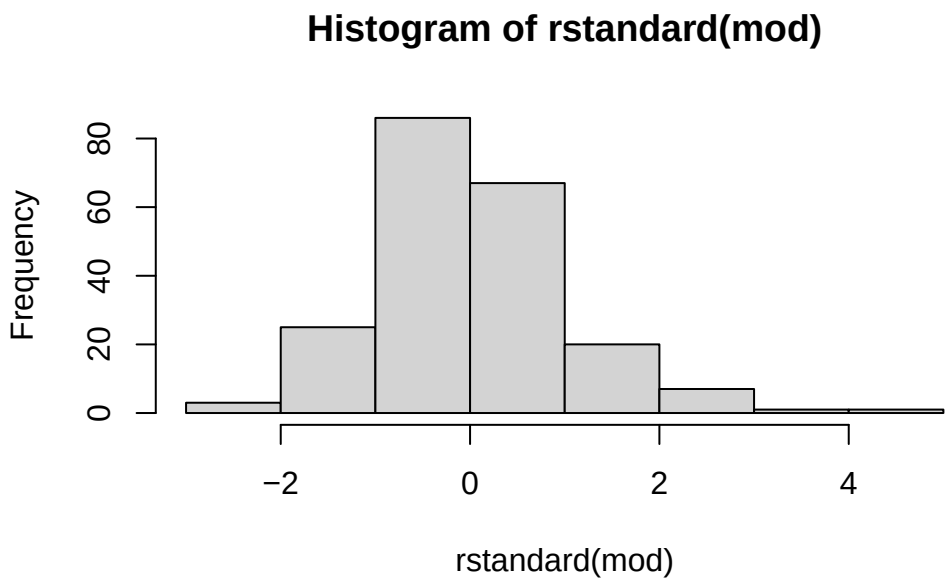
```r
# no problem here

# linearity: Fitted values vs. residuals plot
plot(mod, which = 1)
```



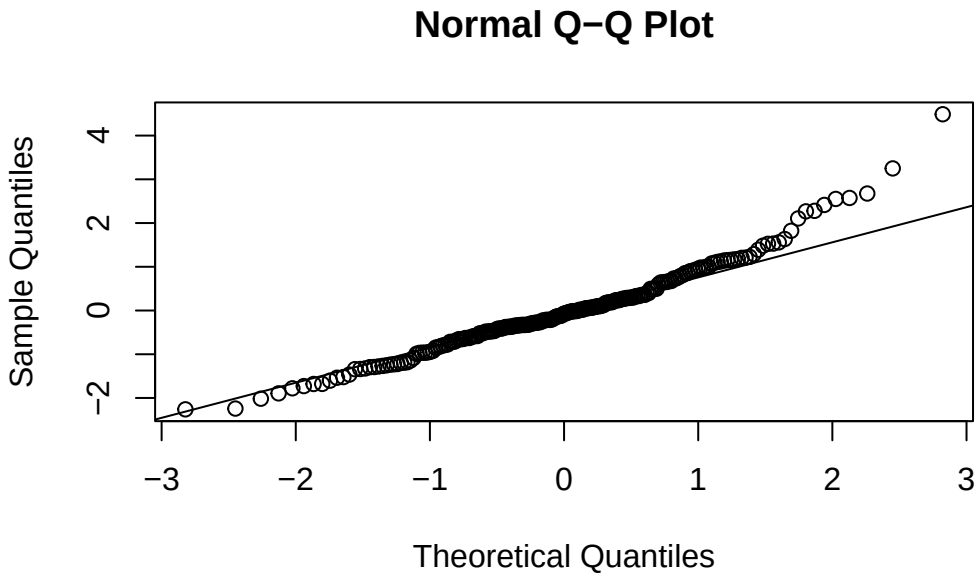
```r
# heteroscedasticity
ncvTest(mod) # Breusch-Pagan test
```

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 30.50699, Df = 1, p = 3.3267e-08
```

```r
# normality of residuals
hist(rstandard(mod))
```




```r
qqnorm(rstandard(mod))
qqline(rstandard(mod))
```




```
shapiro.test(rstandard(mod))
```

```
    Shapiro-Wilk normality test

data:  rstandard(mod)
W = 0.96414, p-value = 3.65e-05
```

```
# summary of plots/table
astatur::regression.diagnostics(mod)
```

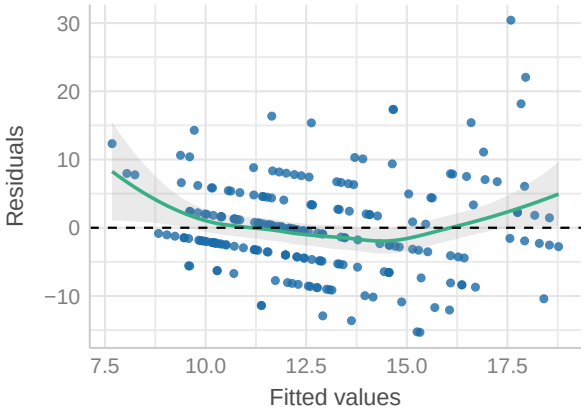
```
Tests of linear model assumptions
---------------------------------

6/13 (46.2 %) checks failed


Identified problems:
    heteroskedasticity
    normality
    model specification
    functional form
```

```
    outliers
Summary:
# A tibble: 13 x 8
   assumption          variable test   statistic  p.value  crit problem decision
   <chr>               <chr>    <chr>      <dbl>    <dbl> <dbl> <chr>   <chr>
 1 heteroskedasticity  global   stude~   17.9     3.05e-3  0.05 Problem -
 2 heteroskedasticity  global   Non-c~   30.5     3.33e-8  0.05 Problem -
 3 multicollinearity   educ     Varia~    1.10   NA        5    No Pro~ +
 4 multicollinearity   gender   Varia~    1.03   NA        5    No Pro~ +
 5 multicollinearity   age      Varia~    1.51   NA        5    No Pro~ +
 6 multicollinearity   marital  Varia~    1.45   NA        5    No Pro~ +
 7 multicollinearity   health   Varia~    1.07   NA        5    No Pro~ +
 8 normality           global   Shapi~    0.964   3.68e-5  0.01 Problem -
 9 model specification global   Stata~    0.231   5.65e-4  0.05 Problem -
10 functional form     global   RESET~    6.22    2.38e-3  0.05 Problem -
11 outliers            global   Cook'~    0.101  NA        1    No Pro~ +
12 outliers            global   Bonfe~    4.72    9.37e-4  0.05 Problem -
13 autocorrelation     global   Durbi~   -0.0243  7.26e-1  0.05 No Pro~ +

Outliers:
-----------
Cook's distance (criterion=1.00): No outliers
Outlier test (criterion=0.05):
  rstudent unadjusted p-value Bonferroni p
3 4.716255         4.4622e-06   0.00093707
```

```
library(performance)
check_model(mod)
```

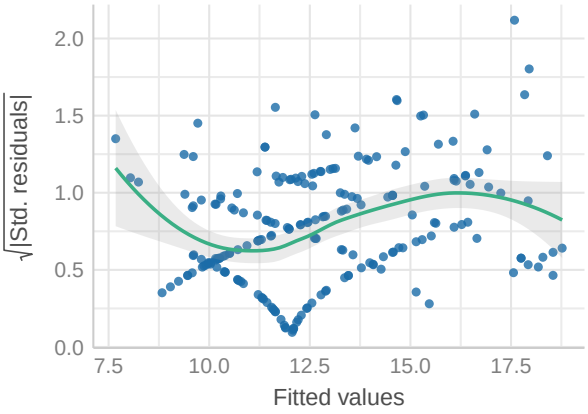
Linearity

Reference line should be flat and horizontal



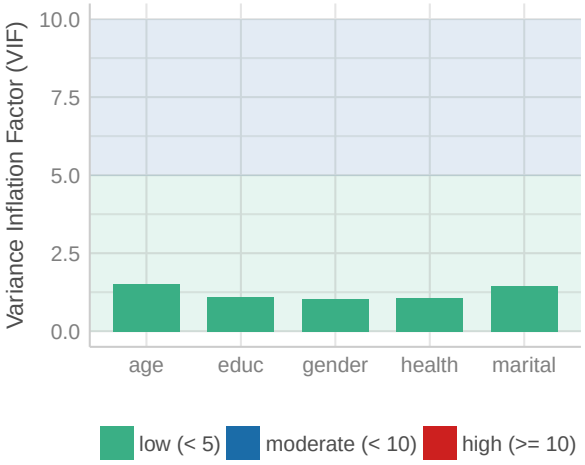
Homogeneity of Variance

Reference line should be flat and horizontal



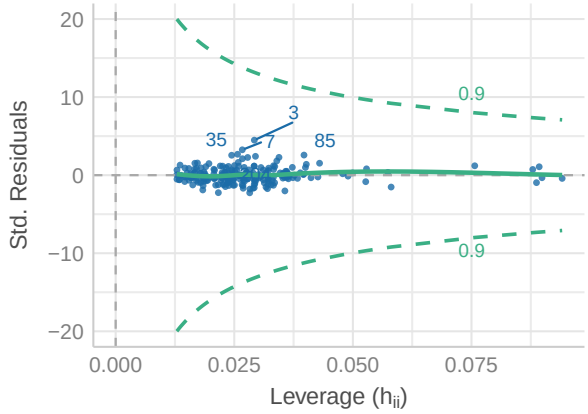
Collinearity

Higher bars (>5) indicate potential collinearity issues



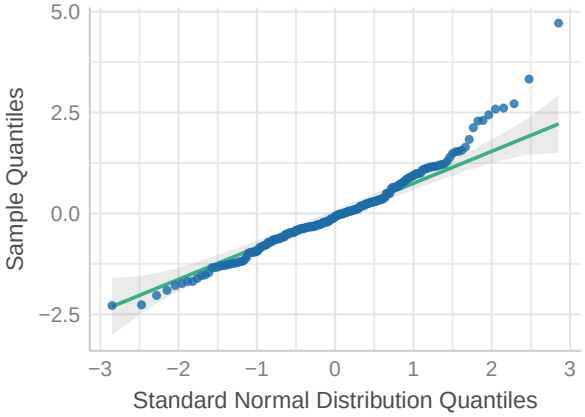
Influential Observations

Points should be inside the contour lines



Normality of Residuals

Dots should fall along the line



Normality of Residuals

Distribution should be close to the normal curve