Solution in R: Exercises Day 3

PSY8003

Matthias Mittner

spring 2022

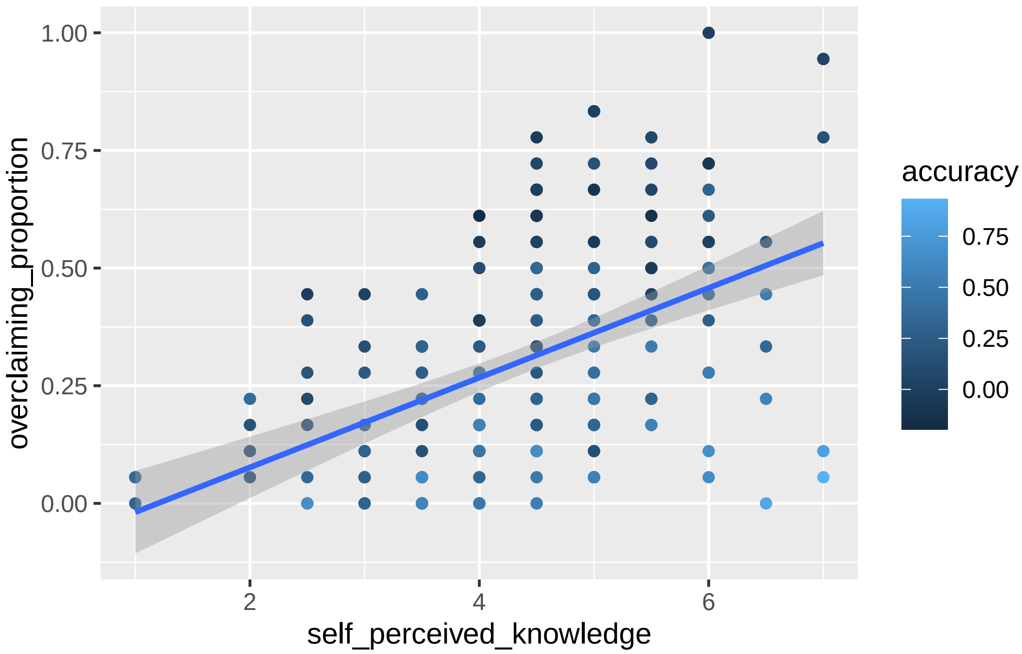
Exercise 1: Interactions

1.

There is a significant positive effect of self-perceived knowledge on overclaiming (β=0.1) and a negative effect of accuracy on overclaiming (β=-0.75).

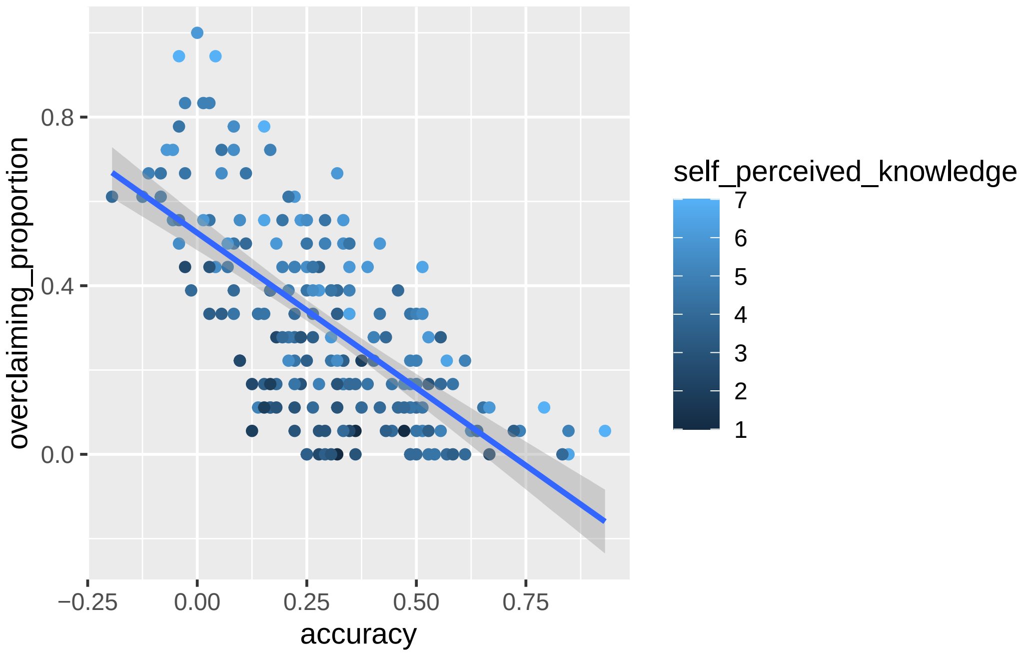
```r
atir2015 <- haven::read_dta("../data/atir2015.dta")
atir2015 |>
  ggplot(aes(x=self_perceived_knowledge, y=overclaiming_proportion,
             color=accuracy))+
  geom_point()+geom_smooth(method="lm")
```

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




```
atir2015 |>
  ggplot(aes(x=accuracy, y=overclaiming_proportion,
             color=self_perceived_knowledge))+
  geom_point()+geom_smooth(method="lm")
```

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




```r
atir2015 |>
  select(self_perceived_knowledge,overclaiming_proportion,accuracy) |>
  cor()
```

```
                         self_perceived_knowledge overclaiming_proportion
self_perceived_knowledge               1.00000000               0.4811502
overclaiming_proportion                0.48115020               1.0000000
accuracy                               0.03261025              -0.6720098
                            accuracy
self_perceived_knowledge  0.03261025
overclaiming_proportion  -0.67200976
accuracy                  1.00000000
```

```r
summary(mod <- lm(overclaiming_proportion ~ accuracy+self_perceived_knowledge,
                  data=atir2015))
```

```
Call:
lm(formula = overclaiming_proportion ~ accuracy + self_perceived_knowledge,
    data = atir2015)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.37190 -0.09280 -0.00796  0.09369  0.31250 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)               0.088910   0.036737    2.42   0.0164 *  
accuracy                 -0.753986   0.042195  -17.87   <2e-16 ***
self_perceived_knowledge  0.099765   0.007632   13.07   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1268 on 199 degrees of freedom
Multiple R-squared:  0.7049,	Adjusted R-squared:  0.702 
F-statistic: 237.7 on 2 and 199 DF,  p-value: < 2.2e-16
```

2.

- Comparing mean overclaiming between the two possible orderings results in a significant difference in the (M1=0.34, M2=0.27).

```r
atir2015 |> group_by(order_of_tasks) |>
  summarise(mean(overclaiming_proportion))
```

```
# A tibble: 2 x 2
                              order_of_tasks `mean(overclaiming_proportion)`
                                   <dbl+lbl>                           <dbl>
1 1 [Self-Perceived Knowledge Measured First]                          0.344
2 2 [Overclaiming Measured First]                                      0.272
```

```r
summary(mod<-lm(overclaiming_proportion ~ order_of_tasks, data=atir2015))
```

```
Call:
lm(formula = overclaiming_proportion ~ order_of_tasks, data = atir2015)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.34378 -0.17712 -0.03025  0.15622  0.72772 
```

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.41529    0.05118   8.115 4.87e-14 ***
order_of_tasks -0.07151    0.03237  -2.209   0.0283 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.23 on 200 degrees of freedom
Multiple R-squared:  0.02382,   Adjusted R-squared:  0.01894
F-statistic: 4.881 on 1 and 200 DF,  p-value: 0.0283
```

3.

- the effect is present for `order_of_tasks`=1: $\beta = 0.11$
- the effect is present for `order_of_tasks`=2: $\beta = 0.08$
- the effect is present when modulation by `order_of_task` is allowed, $beta = 0.11$
- the interaction is not significant ($p = .06$) but almost so. The interpretation is that the association between self-perceived knowledge and overclaiming is reduced by 0.03 when the order of presentation of the tests is switched

```r
# subset for order_of_task=1
summary(mod1<-lm(overclaiming_proportion ~ accuracy+self_perceived_knowledge,
                 data=subset(atir2015, order_of_tasks==1)))
```

```
Call:
lm(formula = overclaiming_proportion ~ accuracy + self_perceived_knowledge,
    data = subset(atir2015, order_of_tasks == 1))

Residuals:
     Min       1Q   Median       3Q      Max
-0.38711 -0.09364  0.01148  0.09499  0.24362

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)               0.03213    0.05787   0.555     0.58
accuracy                 -0.82338    0.06176 -13.332  < 2e-16 ***
self_perceived_knowledge  0.11609    0.01181   9.828 2.86e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.1309 on 98 degrees of freedom
Multiple R-squared:  0.721, Adjusted R-squared:  0.7153 
F-statistic: 126.6 on 2 and 98 DF,  p-value: < 2.2e-16
```

```r
# subset for order_of_task=2
summary(mod2<-lm(overclaiming_proportion ~ accuracy+self_perceived_knowledge,
                 data=subset(atir2015, order_of_tasks==2)))
```

```

Call:
lm(formula = overclaiming_proportion ~ accuracy + self_perceived_knowledge, 
    data = subset(atir2015, order_of_tasks == 2))

Residuals:
     Min       1Q   Median       3Q      Max 
-0.23398 -0.07864 -0.01642  0.07955  0.35685 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)               0.12564    0.04749   2.646   0.0095 ** 
accuracy                 -0.69036    0.05709 -12.092  < 2e-16 ***
self_perceived_knowledge  0.08625    0.01018   8.469  2.5e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1206 on 98 degrees of freedom
Multiple R-squared:  0.6875,    Adjusted R-squared:  0.6811 
F-statistic: 107.8 on 2 and 98 DF,  p-value: < 2.2e-16
```

```r
# model including the interaction term
summary(mod3 <- lm(overclaiming_proportion ~ accuracy+
                     self_perceived_knowledge*factor(order_of_tasks),
                   data=atir2015))
```

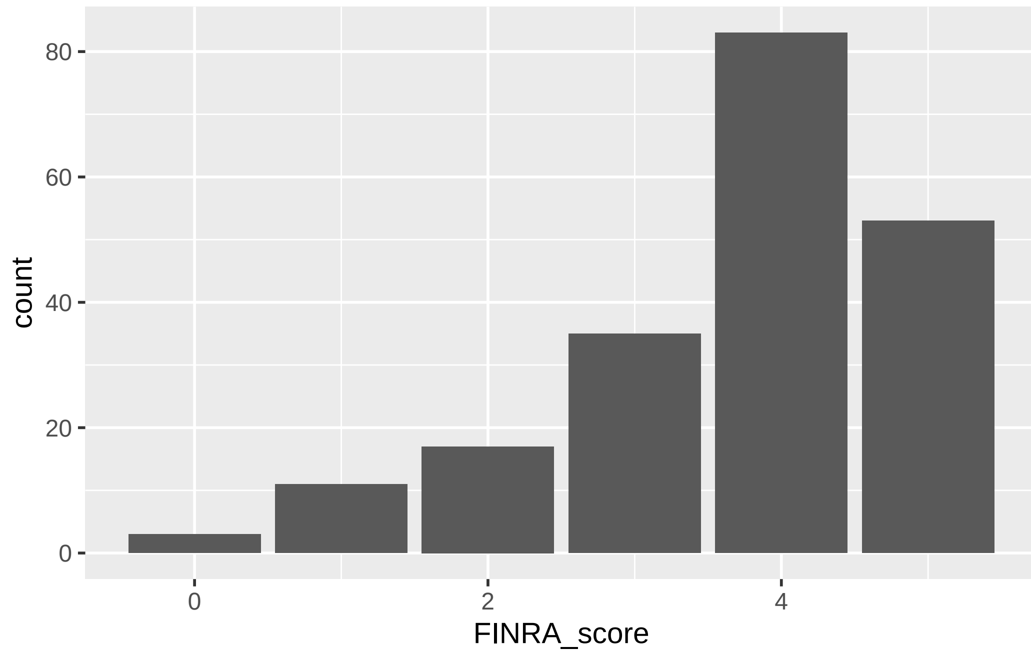
```

Call:
lm(formula = overclaiming_proportion ~ accuracy + self_perceived_knowledge * 
    factor(order_of_tasks), data = atir2015)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.37736 -0.09409 -0.00809  0.09121  0.33591 

Coefficients:
                                             Estimate Std. Error t value
(Intercept)                                   0.01874    0.05521   0.339
accuracy                                     -0.75706    0.04222 -17.932
self_perceived_knowledge                      0.11498    0.01138  10.102
factor(order_of_tasks)2                       0.12702    0.07171   1.771
self_perceived_knowledge:factor(order_of_tasks)2 -0.02859    0.01559  -1.833
                                             Pr(>|t|)    
(Intercept)                                    0.7346    
accuracy                                       <2e-16 ***
self_perceived_knowledge                       <2e-16 ***
factor(order_of_tasks)2                        0.0781 .  
self_perceived_knowledge:factor(order_of_tasks)2   0.0683 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1263 on 197 degrees of freedom
Multiple R-squared:  0.7099,	Adjusted R-squared:  0.704 
F-statistic: 120.5 on 4 and 197 DF,  p-value: < 2.2e-16
```

4.

- FINRA has a mean of 3.7 and an SD of 1.9. Pretty high scores given that 5 is max.
- When controling for actual knowledge, the effect of self-perceived knowledge on overclaiming is still present but slightly reduced, $\beta = 0.09$
- there is also a weak effect of actual knowledge on overclaiming, $\beta = 0.018$

```r
atir2015 |> ggplot(aes(FINRA_score))+geom_bar()
```




```r
atir2015 %>% summarise(mean(FINRA_score), sd(FINRA_score))
```

```
# A tibble: 1 x 2
  `mean(FINRA_score)` `sd(FINRA_score)`
                <dbl>             <dbl>
1                3.70              1.19
```

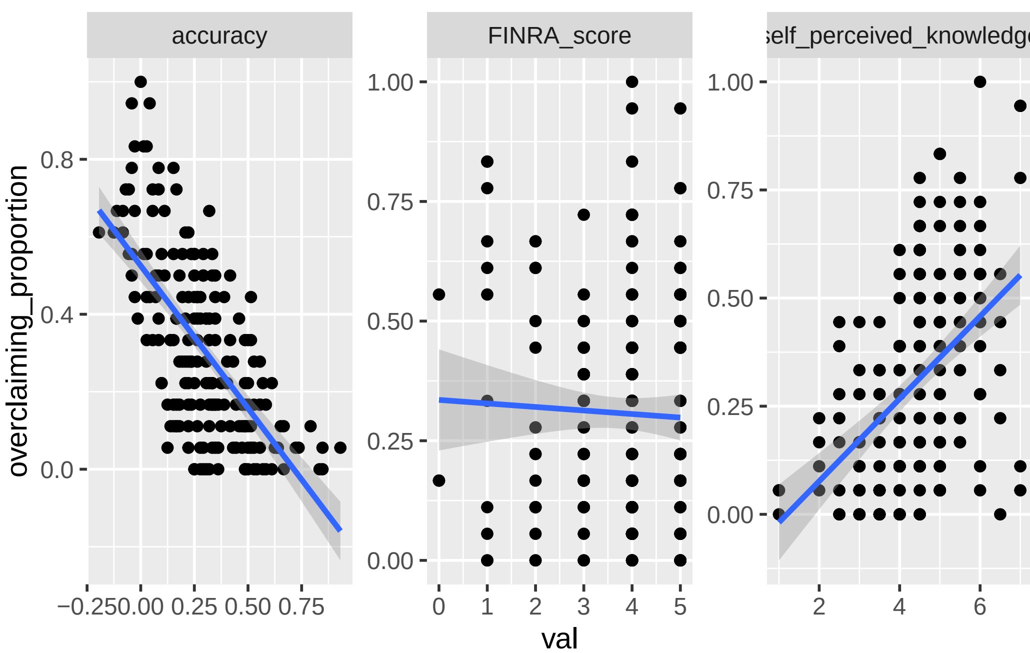
```r
summary(mod <- lm(overclaiming_proportion ~ accuracy+self_perceived_knowledge+
                    FINRA_score, data=atir2015))
```

```
Call:
lm(formula = overclaiming_proportion ~ accuracy + self_perceived_knowledge +
    FINRA_score, data = atir2015)

Residuals:
     Min       1Q   Median       3Q      Max
-0.38033 -0.08672 -0.01418  0.08808  0.30886

Coefficients:
```

```
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                0.057787   0.039203   1.474   0.1421
accuracy                  -0.793219   0.045655 -17.374   <2e-16 ***
self_perceived_knowledge   0.094069   0.008018  11.732   <2e-16 ***
FINRA_score                0.018370   0.008576   2.142   0.0334 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1256 on 198 degrees of freedom
Multiple R-squared:  0.7116,    Adjusted R-squared:  0.7073
F-statistic: 162.9 on 3 and 198 DF,  p-value: < 2.2e-16
```

```r
atir2015 |> select( overclaiming_proportion, accuracy,
                    self_perceived_knowledge, FINRA_score) |>
  gather(var, val, -overclaiming_proportion) |>
  ggplot(aes(val, overclaiming_proportion))+
  geom_point()+facet_wrap(~var, scales="free")+geom_smooth(method="lm")
```

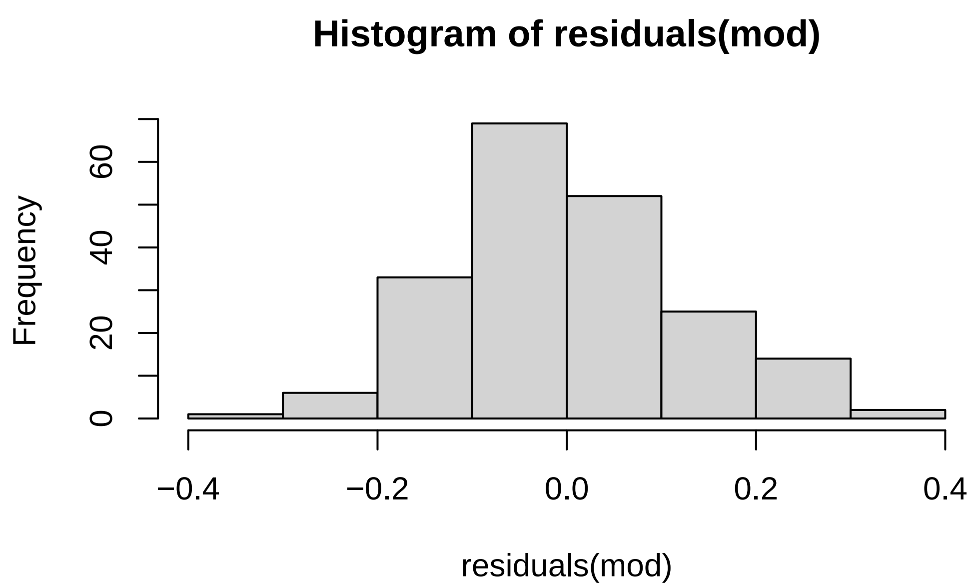
```
Warning: attributes are not identical across measure variables;
they will be dropped

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



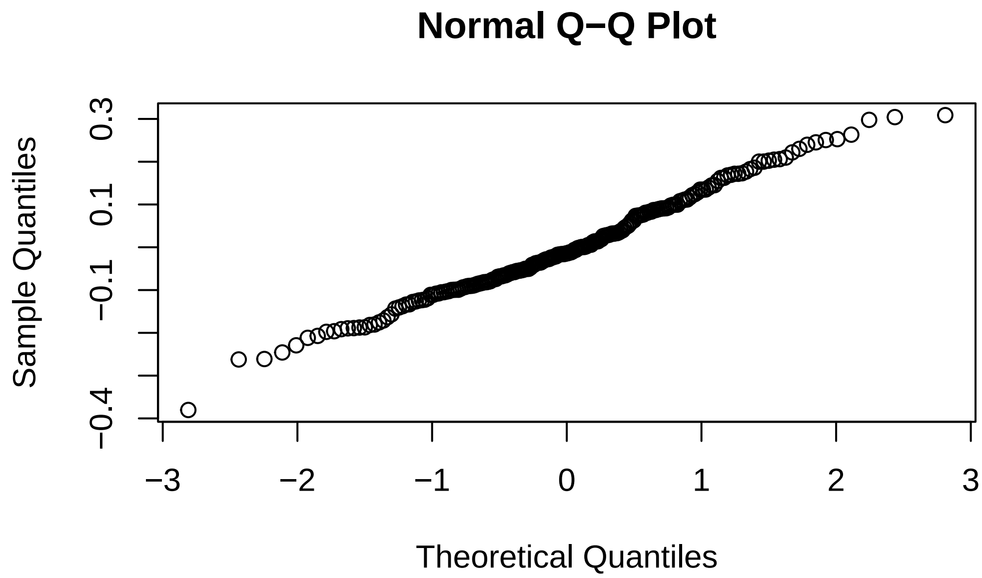
5.

- the histogram of the residuals does not show a strong departure from the normal distribution
- nor does the QQ-plot
- the predicted vs. residuals plot shows some heterogeneity in variance (increasing with predicted value)
- the "stripe"-structure comes from the discrete nature of the `overclaiming_proportion` variable

```
hist(residuals(mod))
```

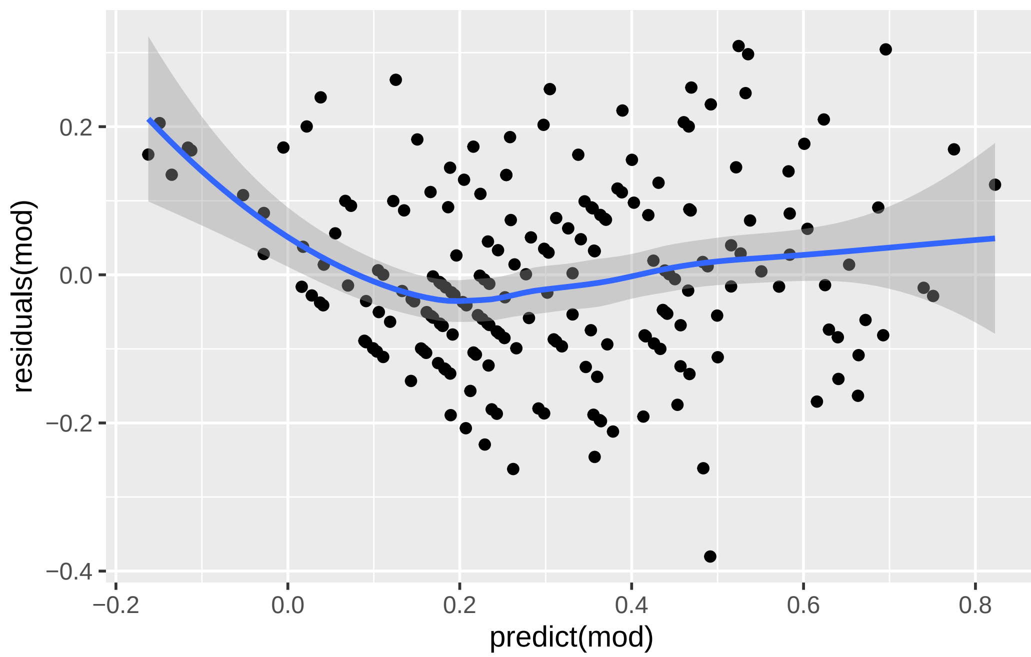



```
qqnorm(residuals(mod))
```

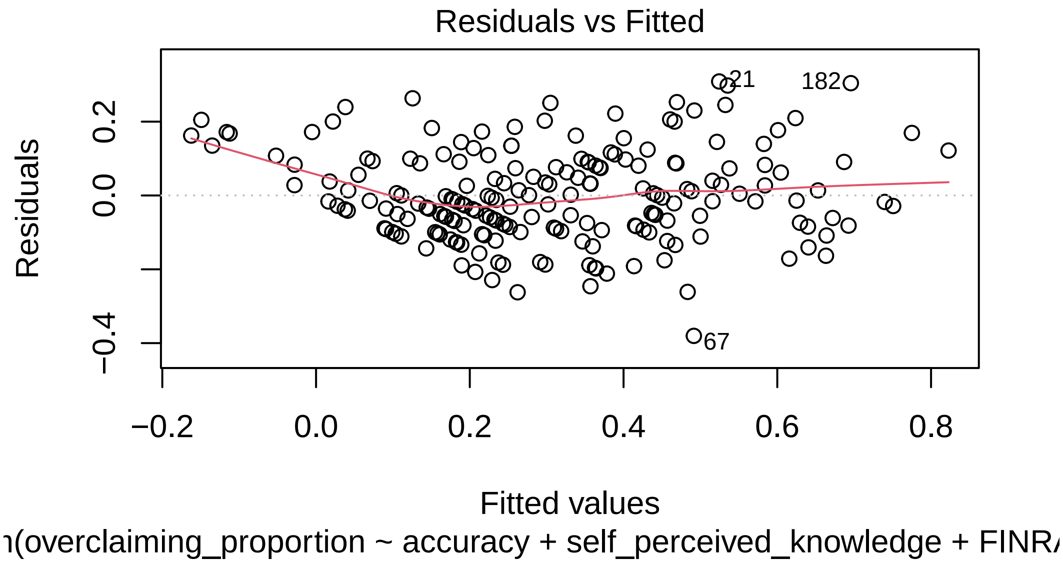


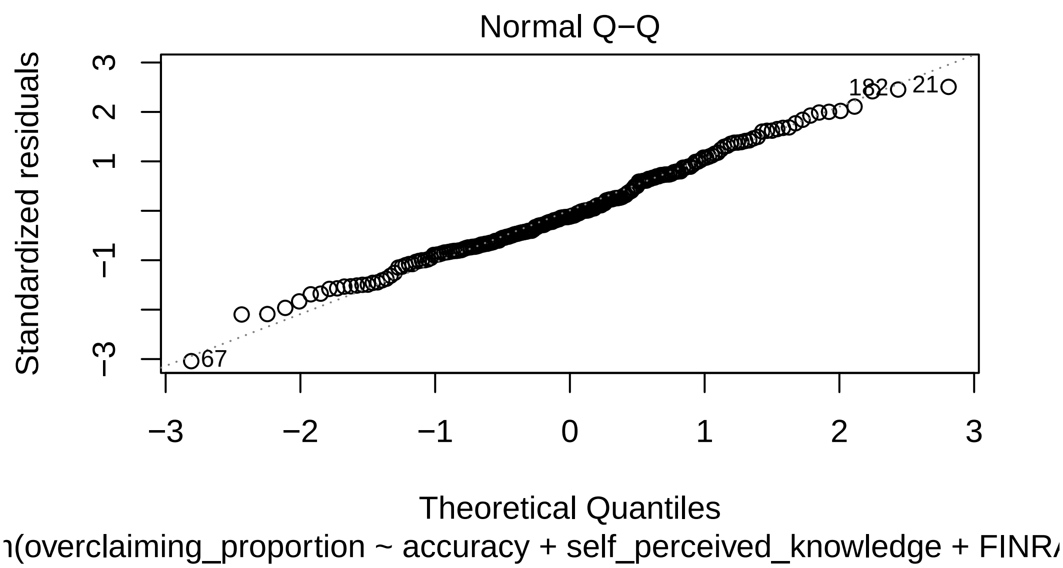

```r
qplot(predict(mod), residuals(mod))+geom_smooth()
```

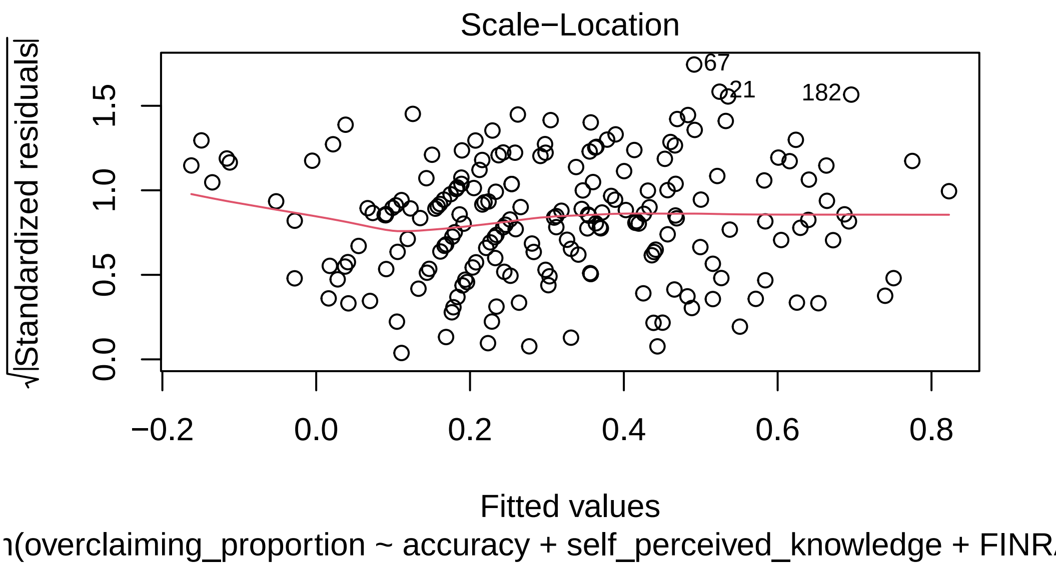
```
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

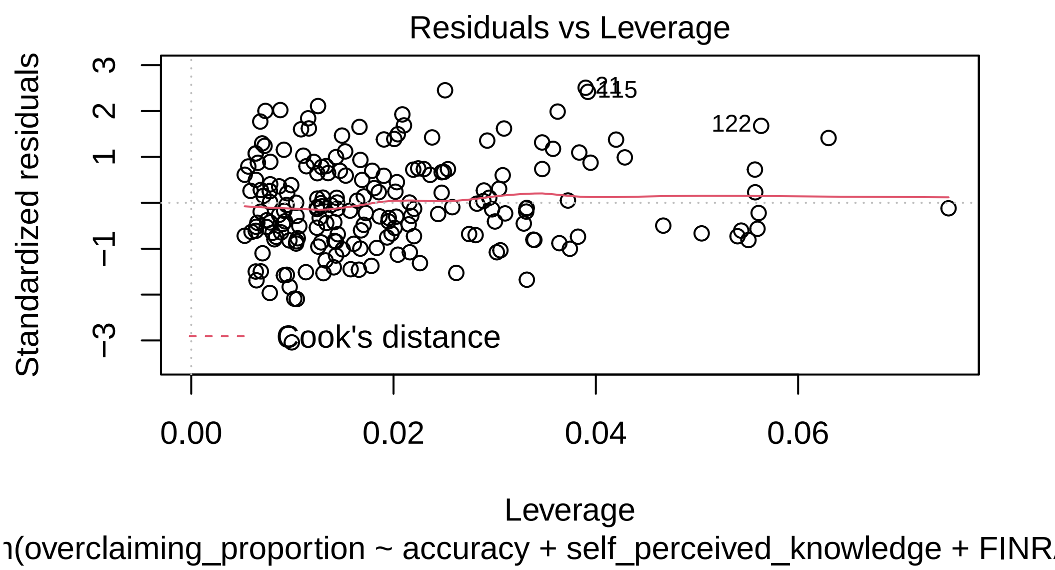


```
plot(mod)
```


Residuals vs Fitted
Residuals
Fitted values
lm(overclaiming_proportion ~ accuracy + self_perceived_knowledge + FINRA
21
182
67


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(overclaiming_proportion ~ accuracy + self_perceived_knowledge + FINRA
182
21
67






```r
library(car)
```

```
Loading required package: carData
```

```
Attaching package: 'car'
```

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

    some


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

    recode
```

```
outlierTest(mod)
```

```
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
    rstudent unadjusted p-value Bonferroni p
67 -3.107901          0.0021629       0.4369
```

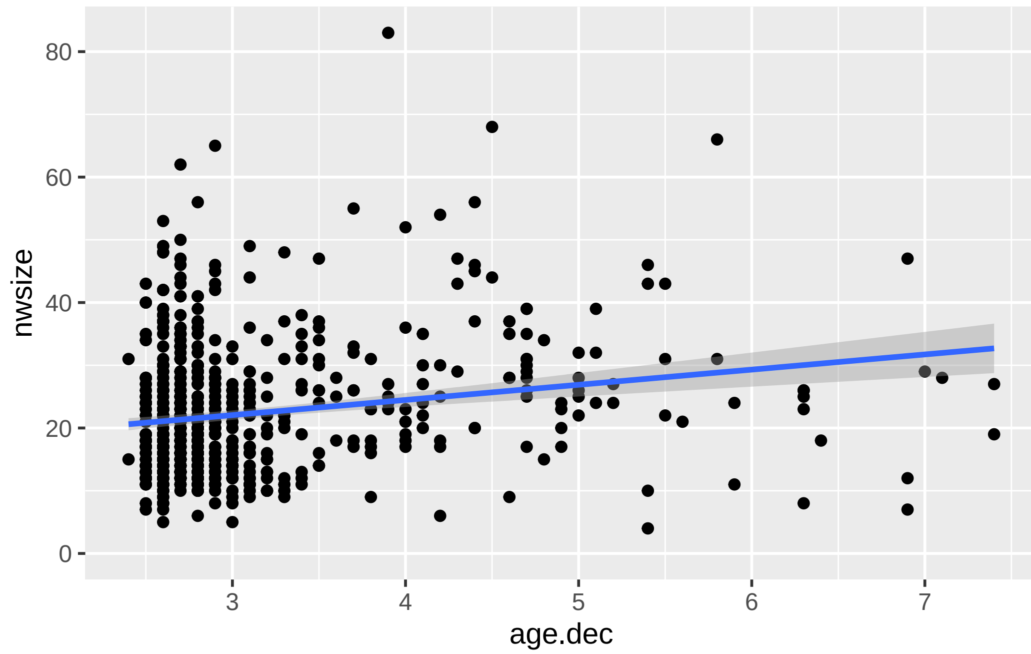
```
cooks.distance(mod) %>% sort %>% rev %>% head
```

```
        21        115        122        182         99        140
0.06378176 0.05967856 0.04202365 0.03872156 0.03714847 0.03353071
```

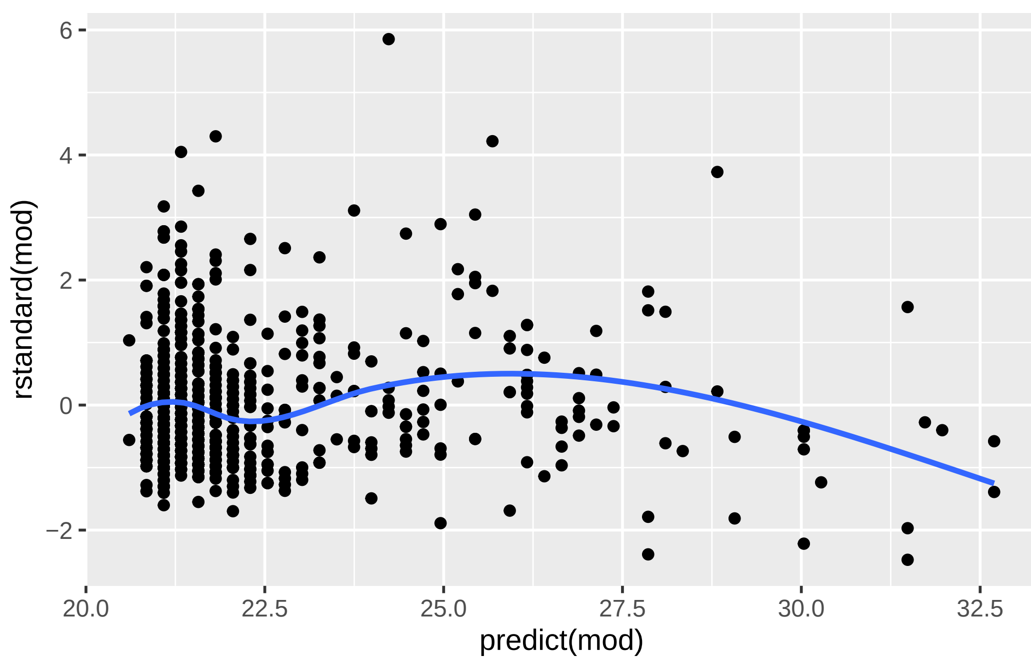
Exercise 2: Nonlinear regression

```
library(dplyr)
penguin <- haven::read_dta("../data/explorepenguin.dta")
penguin <- penguin |>
  mutate(age.years=2022-age,
         age.dec=age.years/10)

penguin |>
  ggplot(aes(x=age.dec, y=nwsize))+
  geom_point()+
  geom_smooth(method="lm")
```

```
Warning: Removed 48 rows containing non-finite values (stat_smooth).


Warning: Removed 48 rows containing missing values (geom_point).
```




```
mod <- lm(nwsize ~ age.dec, data=penguin)
qplot(predict(mod),rstandard(mod))+geom_smooth(se=F)
```



```r
library(mfp)
mod.fp <- mfp(nwsize ~ fp(age.dec), data=penguin, verbose=T)
```

```
    Variable     Deviance     Power(s)
------------------------------------------------
Cycle 1
    age.dec
                74677.2
                71868.63    1
                71412.42    -1
                70564   3 3


Tansformation
        shift scale
age.dec     0     1

Fractional polynomials
        df.initial select alpha df.final power1 power2
age.dec          4      1  0.05        4      3      3


Transformations of covariates:
                                       formula
age.dec I(age.dec^3)+I(age.dec^3*log(age.dec))


Deviance table:
         Resid. Dev
Null model   74677.2
Linear model     71868.63
Final model  70564
```

```r
print(mod.fp)
```

```
Call:
mfp(formula = nwsize ~ fp(age.dec), data = penguin, verbose = T)
```

```
Deviance table:
         Resid. Dev
Null model   74677.2
Linear model     71868.63
Final model  70564

Fractional polynomials:
        df.initial select alpha df.final power1 power2
age.dec          4      1  0.05        4      3      3


Transformations of covariates:
                                       formula
age.dec I(age.dec^3)+I(age.dec^3*log(age.dec))

Rescaled coefficients:
Intercept  age.dec.1  age.dec.2
  15.7564     0.5224    -0.2577

Degrees of Freedom: 712 Total (i.e. Null);  710 Residual
Null Deviance:      74680
Residual Deviance: 70560    AIC: 5307
```

```r
p1=2
p2=3
penguin <-
  penguin |> mutate(age.dec.2=(age.dec)^2,
                    age.dec.3=(age.dec)^3)

summary(mod.fp2<-lm(nwsize ~ age.dec.2 + age.dec.3, data=penguin))
```

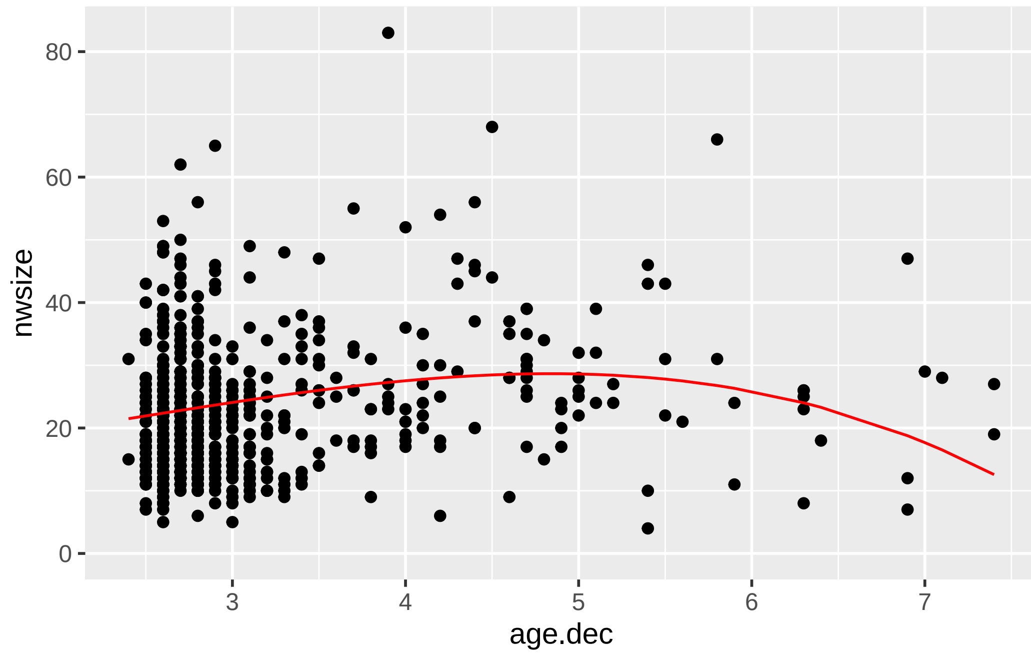
```
Call:
lm(formula = nwsize ~ age.dec.2 + age.dec.3, data = penguin)

Residuals:
    Min      1Q  Median      3Q     Max
-25.091  -7.039  -1.966   4.108  56.865

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.56408    1.78694   7.031 4.82e-12 ***
age.dec.2    1.73841    0.35072   4.957 8.98e-07 ***
age.dec.3   -0.21697    0.05076  -4.275 2.17e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.972 on 710 degrees of freedom
  (48 observations deleted due to missingness)
Multiple R-squared:  0.05453,   Adjusted R-squared:  0.05186
F-statistic: 20.47 on 2 and 710 DF,  p-value: 2.268e-09
```

```
# create new variable for model predictions
penguin |> mutate(
  predicted=14.49 + 1.815*age.dec^2 - 0.25*age.dec^3,
  predict.p1=14.49 + 1.815*age.dec^2,
  predict.p2=14.49 + - 0.25*age.dec^3
) -> penguin


# plot prediction and scatter
penguin |>
  ggplot(aes(age.dec, nwsize))+geom_point()+
    geom_line(aes(y=predicted), color="red")
```

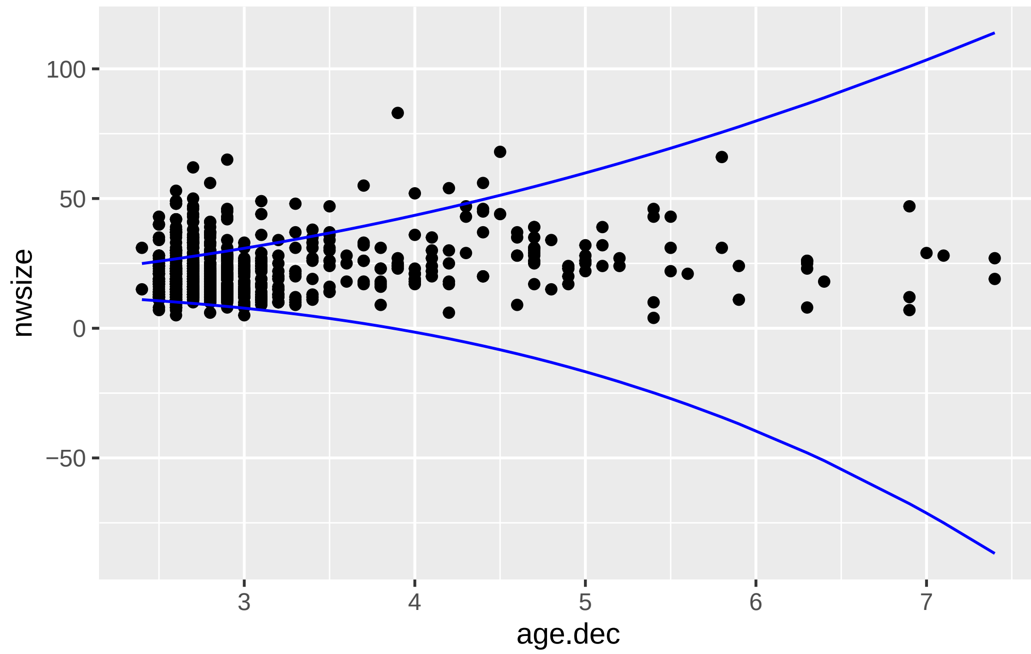
```
Warning: Removed 48 rows containing missing values (geom_point).
```

```
Warning: Removed 22 row(s) containing missing values (geom_path).
```

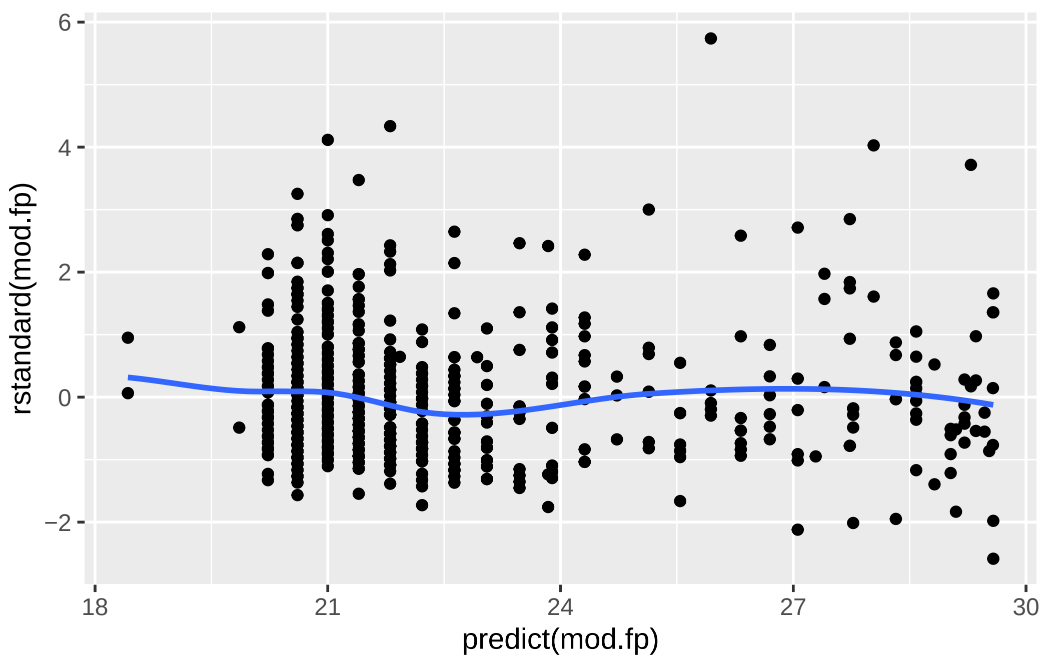



```
# plot  the individual components
penguin %>%
  ggplot(aes(age.dec, nwsize))+geom_point()+
    geom_line(aes(y=predict.p1), color="blue")+
    geom_line(aes(y=predict.p2), color="blue")
```

```
Warning: Removed 48 rows containing missing values (geom_point).
Removed 22 row(s) containing missing values (geom_path).
Removed 22 row(s) containing missing values (geom_path).
```




```
qplot(predict(mod.fp),rstandard(mod.fp))+geom_smooth(se=F)
```



Exercise 2: Splines

- there is no "correct" solution for the parameter settings at this point
- going up with knots shows an earlier peak in the data (early twenties) which might reflect university/educational setting which then goes down before the "final" social network is established
- it's hard/impossible to interpret the regression coefficients properly

```r
library(splines)
nknots=8
degree=3
mod.spline <- lm( nwsize ~ ns(age.dec, df=nknots), data=penguin)
summary(mod.spline)
```

```
Call:
lm(formula = nwsize ~ ns(age.dec, df = nknots), data = penguin)

Residuals:
    Min      1Q  Median      3Q     Max 
-25.301  -6.837  -1.837   4.410  55.357 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)                26.8617     6.4277   4.179  3.3e-05 ***
ns(age.dec, df = nknots)1   3.4778     4.2475   0.819 0.413178    
ns(age.dec, df = nknots)2 -12.8564    12.3585  -1.040 0.298563    
ns(age.dec, df = nknots)3  -2.6801     4.9762  -0.539 0.590353    
ns(age.dec, df = nknots)4  -9.0839     7.1992  -1.262 0.207447    
ns(age.dec, df = nknots)5  -3.2963     6.4396  -0.512 0.608895    
ns(age.dec, df = nknots)6  14.1437     3.7020   3.821 0.000145 ***
ns(age.dec, df = nknots)7 -12.4423    17.4606  -0.713 0.476334    
ns(age.dec, df = nknots)8  -0.6589     4.1463  -0.159 0.873782    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.893 on 704 degrees of freedom
  (48 observations deleted due to missingness)
Multiple R-squared:  0.07731,   Adjusted R-squared:  0.06683 
F-statistic: 7.373 on 8 and 704 DF,  p-value: 1.896e-09
```