431 Class 24

Thomas E. Love

2017-11-28

Today's Agenda

- What Happened Over the Break
- Regression Modeling using the National Youth Fitness Survey data
  - Box-Cox to identify potential outcome transformations
  - Adjusted R^2 and when that indicates severe overfitting
  - Testing/Pruning a Kitchen Sink Model
    - Hypothesis testing approaches, t and F tests
    - Using `step` for variable selection using `AIC`
  - Identification of meaningful collinearity with `vif`
  - Checking Assumptions with Residual Plots
  - Summarizing a Model: Drawing Conclusions

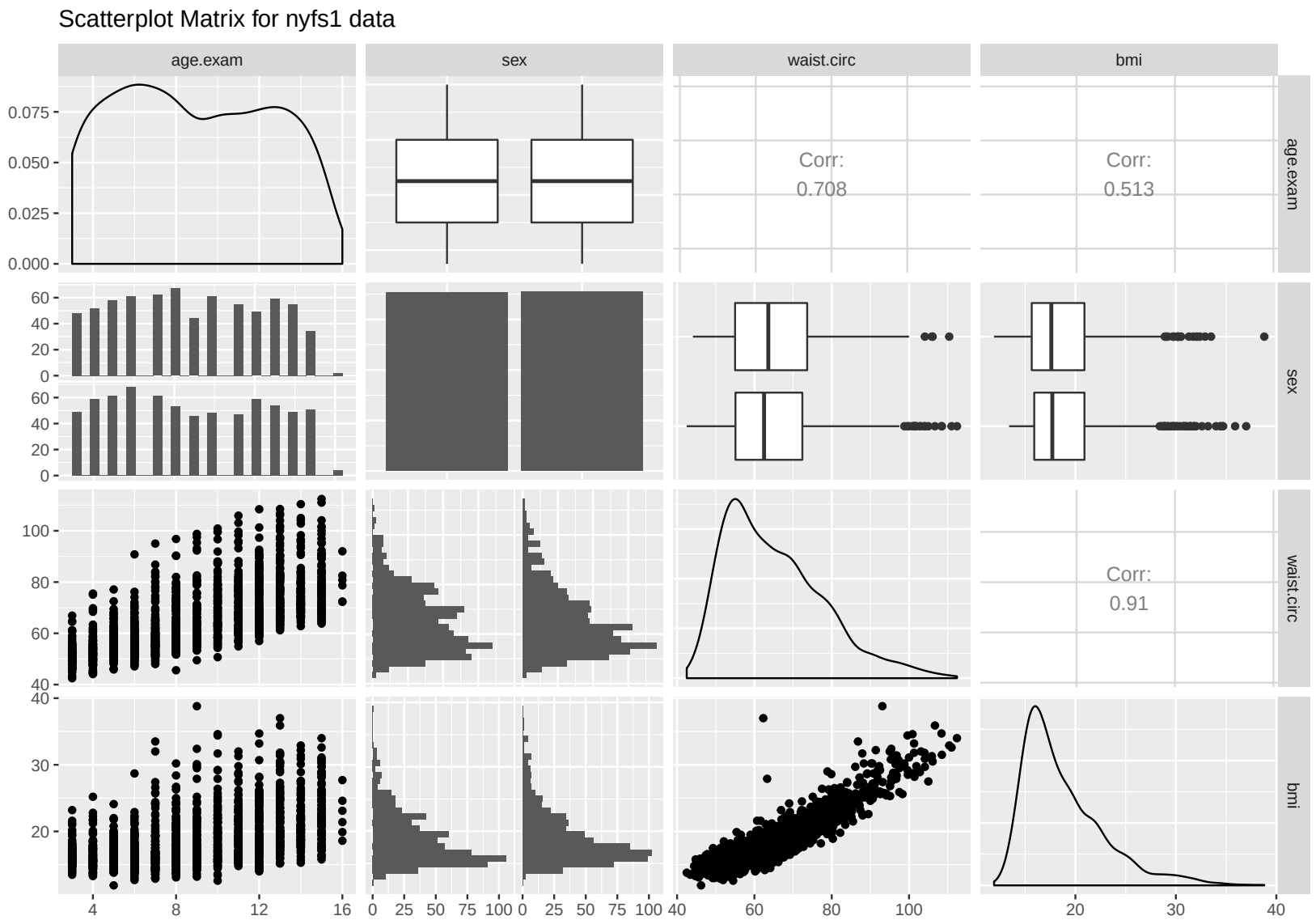
Today's R Setup and Data Set

```
library(car); library(magrittr)
library(broom); library(tidyverse)
```

```
-- Attaching packages ---------------------------------------

v ggplot2 2.2.1     v purrr   0.2.4
v tibble  1.3.4     v dplyr   0.7.4
v tidyr   0.7.2     v stringr 1.2.0
v readr   1.1.1     v forcats 0.2.0

-- Conflicts ------------------------------------------------
x tidyr::extract()   masks magrittr::extract()
x dplyr::filter()    masks stats::filter()
x dplyr::lag()       masks stats::lag()
x dplyr::recode()    masks car::recode()
x purrr::set_names() masks magrittr::set_names()
x purrr::some()      masks car::some()
```

Model `m2` for the `nyfs1` data: Predicting `bmi` using `waist.circ` as well as `age.exam` and `sex`

Scatterplot Matrix for some NYFS1 variables



Building Model 2

```
m2 <- nyfs1 %$%
  lm(bmi ~ waist.circ + age.exam + sex)

m2
```

```
Call:
lm(formula = bmi ~ waist.circ + age.exam + sex)

Coefficients:
(Intercept)   waist.circ     age.exam      sexMale
    -1.4452       0.3484      -0.2932       0.1881
```

Summary of Model m2 (rearranged a little)

```
Call: lm(formula = bmi ~ waist.circ + age.exam + sex)

Multiple R-squared:  0.8635,    Adjusted R-squared:  0.8633
F-statistic:  2979 on 3 and 1412 DF,  p-value: < 2.2e-16

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.445179   0.222018  -6.509 1.05e-10 ***
waist.circ    0.348370   0.004421  78.801  < 2e-16 ***
age.exam     -0.293247   0.015440 -18.992  < 2e-16 ***
sexMale       0.188094   0.080208   2.345   0.0192 *
---
Sig. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residuals:     Min      1Q  Median      3Q     Max
           -4.0232 -0.8879 -0.0802  0.7836 20.3659
Residual standard error: 1.509 on 1412 degrees of freedom
```

Summary of `m1` (for reference, rearranged)

```
Call: lm(formula = bmi ~ waist.circ)

Multiple R-squared:  0.8282,    Adjusted R-squared:  0.828
F-statistic:  6814 on 1 and 1414 DF,  p-value: < 2.2e-16

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.06646    0.23292  -0.285    0.775
waist.circ     0.28893    0.00350  82.548   <2e-16 ***
---
Sig. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residuals:     Min      1Q  Median      3Q     Max
           -4.2343 -1.0941 -0.0744  0.9254 19.0664
Residual standard error: 1.692 on 1414 degrees of freedom
```

95% Confidence Intervals for m2 Coefficients

```
confint(m2, level = 0.95)
```

```
                   2.5 %     97.5 %
(Intercept) -1.88069873 -1.0096587
waist.circ   0.33969754  0.3570418
age.exam    -0.32353535 -0.2629586
sexMale      0.03075409  0.3454346
```

Tidying the Coefficients (with `broom::tidy`)

Places the coefficient summary into a tibble.

```
tidy(m2) ## from broom package
```

```
         term   estimate   std.error  statistic
1 (Intercept) -1.4451787 0.222017690  -6.509295
2  waist.circ  0.3483697 0.004420855  78.801422
3    age.exam -0.2932470 0.015440287 -18.992327
4     sexMale  0.1880943 0.080208291   2.345073
        p.value
1 1.046338e-10
2 0.000000e+00
3 8.257321e-72
4 1.916119e-02
```

Model Summaries, at a glance (with broom::glance)

```
glance(m2) ## also from broom
```

```
  r.squared adj.r.squared    sigma statistic p.value
1 0.8635437     0.8632538 1.509103  2978.545       0
  df   logLik      AIC      BIC deviance df.residual
1  4 -2589.92 5189.84 5216.118 3215.678        1412
```

```
glance(m1) ## for comparison
```

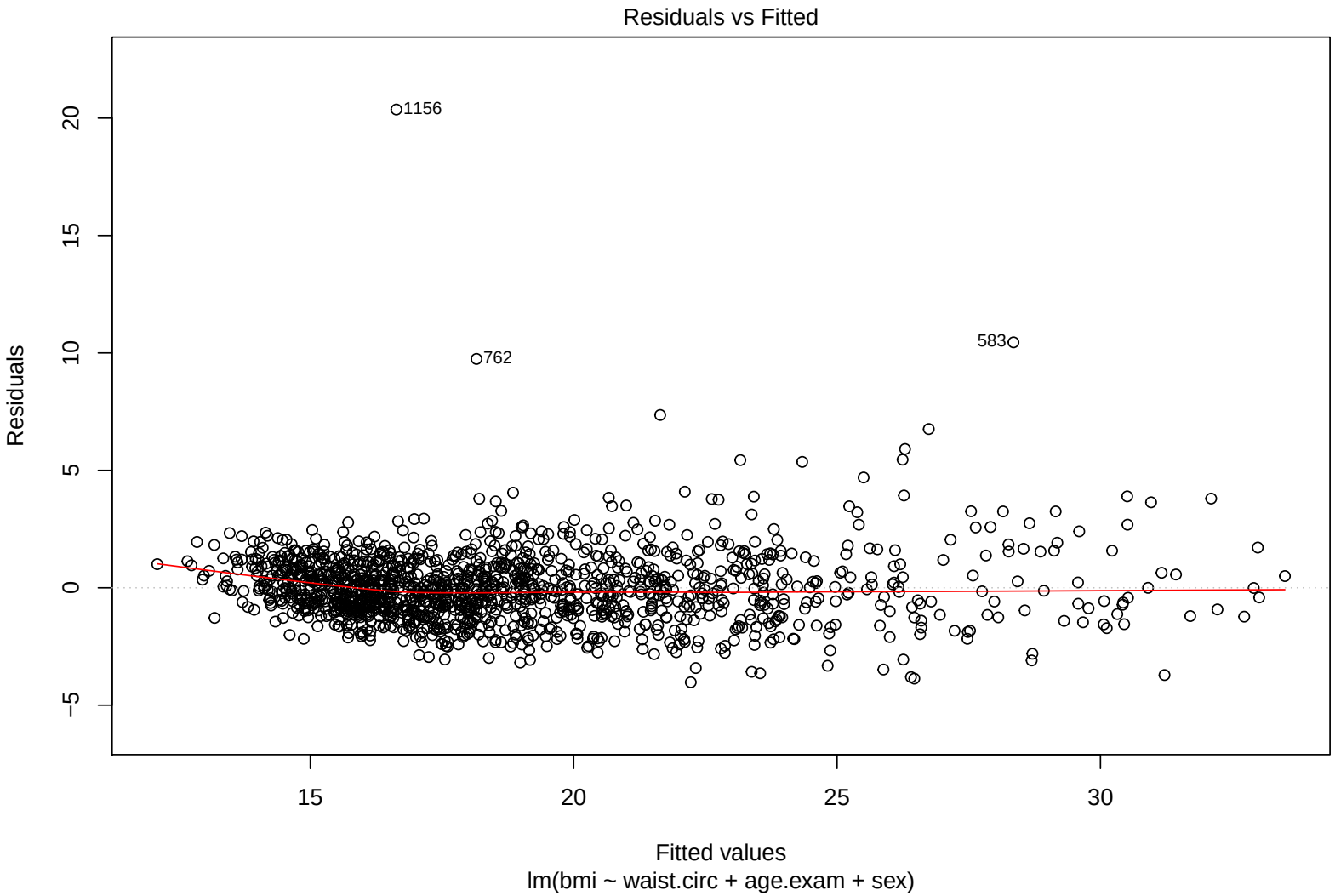
```
  r.squared adj.r.squared    sigma statistic p.value
1 0.8281523     0.8280307 1.692336  6814.214       0
  df    logLik       AIC      BIC deviance df.residual
1  2 -2753.188 5512.376 5528.143   4049.7        1414
```

Augmenting the Data with Model Results (broom::augment) run with warning = FALSE

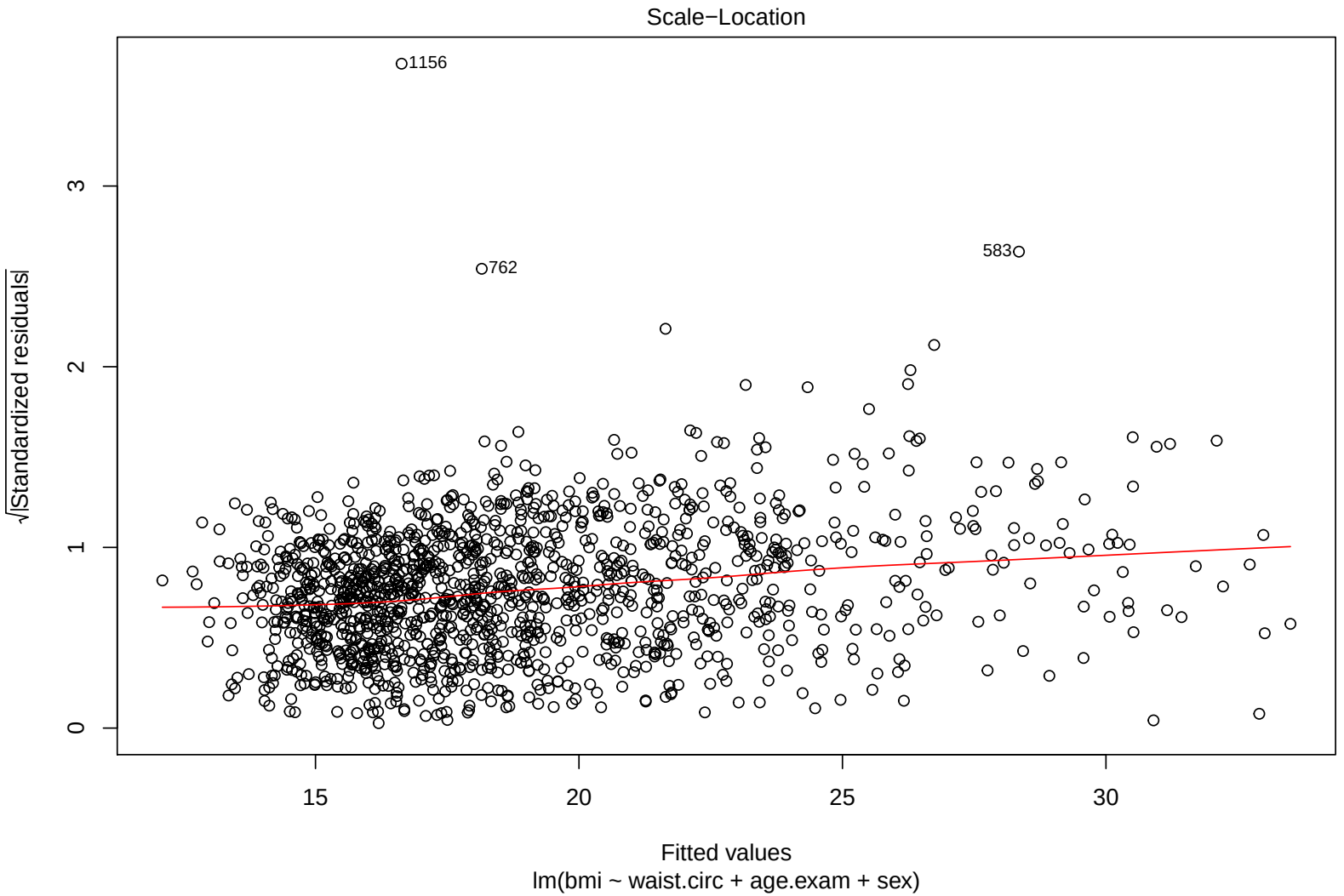
```
newdat2 <- augment(m2) ## again from broom
head(newdat2,3)
```

```
   bmi waist.circ age.exam    sex  .fitted    .se.fit
1 22.3        71.9        8 Female 21.25663 0.06927752
2 19.8        79.4       14 Female 22.10992 0.08007933
3 15.2        46.8        3   Male 14.16688 0.08724867
     .resid         .hat    .sigma       .cooksd
1  1.043374 0.002107399 1.509382 0.0002529070
2 -2.309917 0.002815807 1.508381 0.0016586209
3  1.033124 0.003342564 1.509387 0.0003942708
  .std.resid
1  0.6921161
2 -1.5328151
3  0.6857415
```

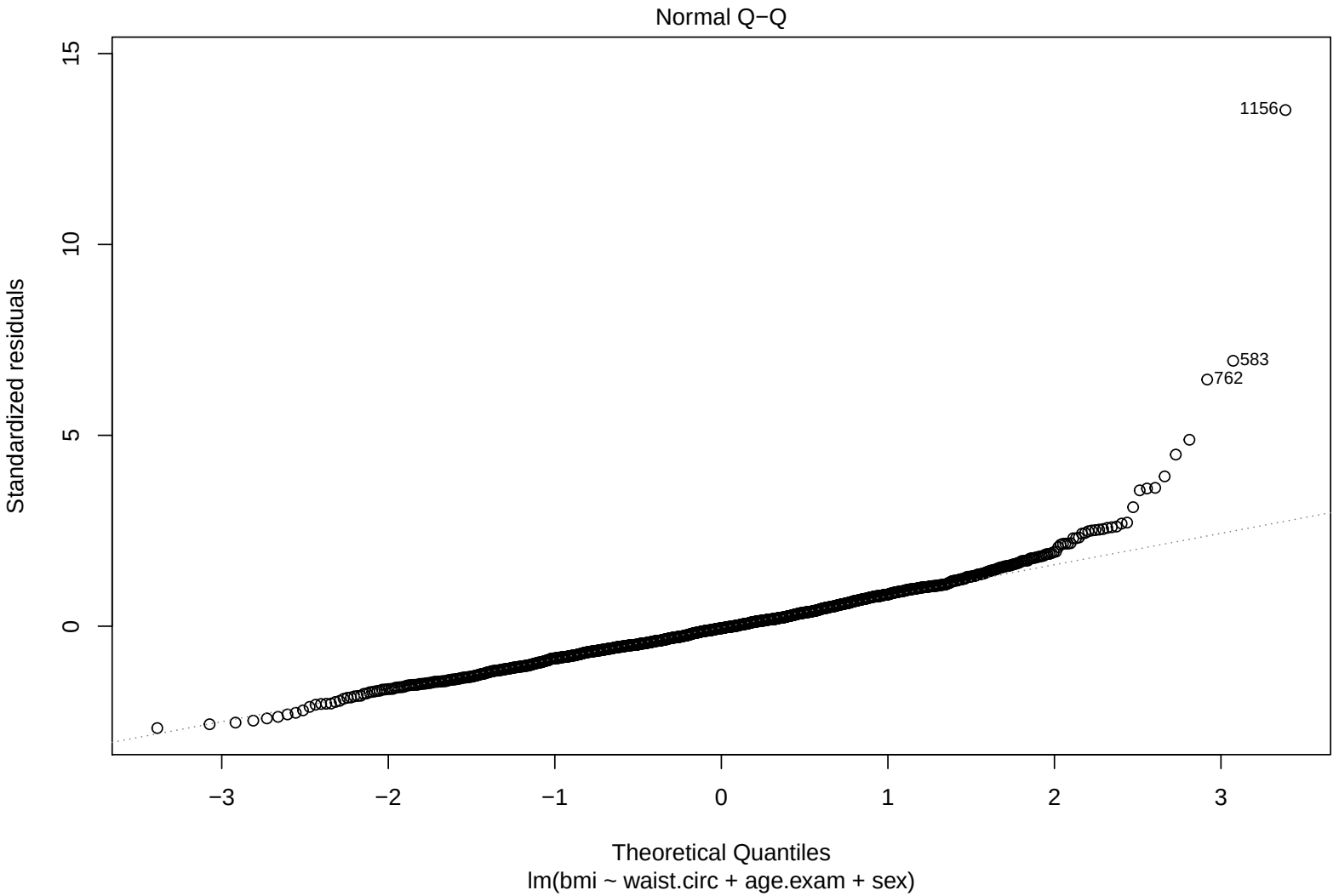
Residuals vs. Fitted Values `plot(m2, which = 1)`



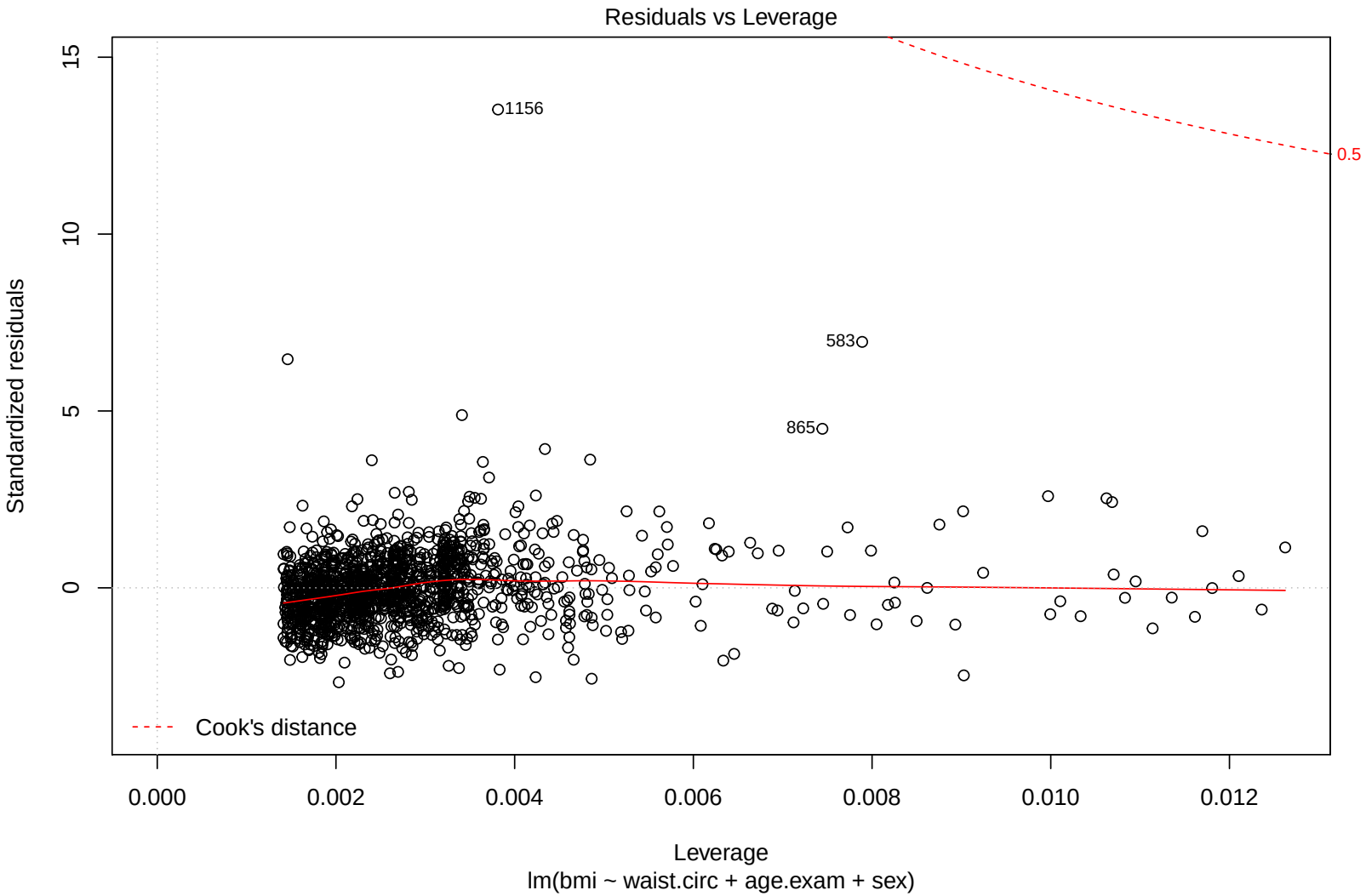
Scale-Location Plot `plot(m2, which = 3)`



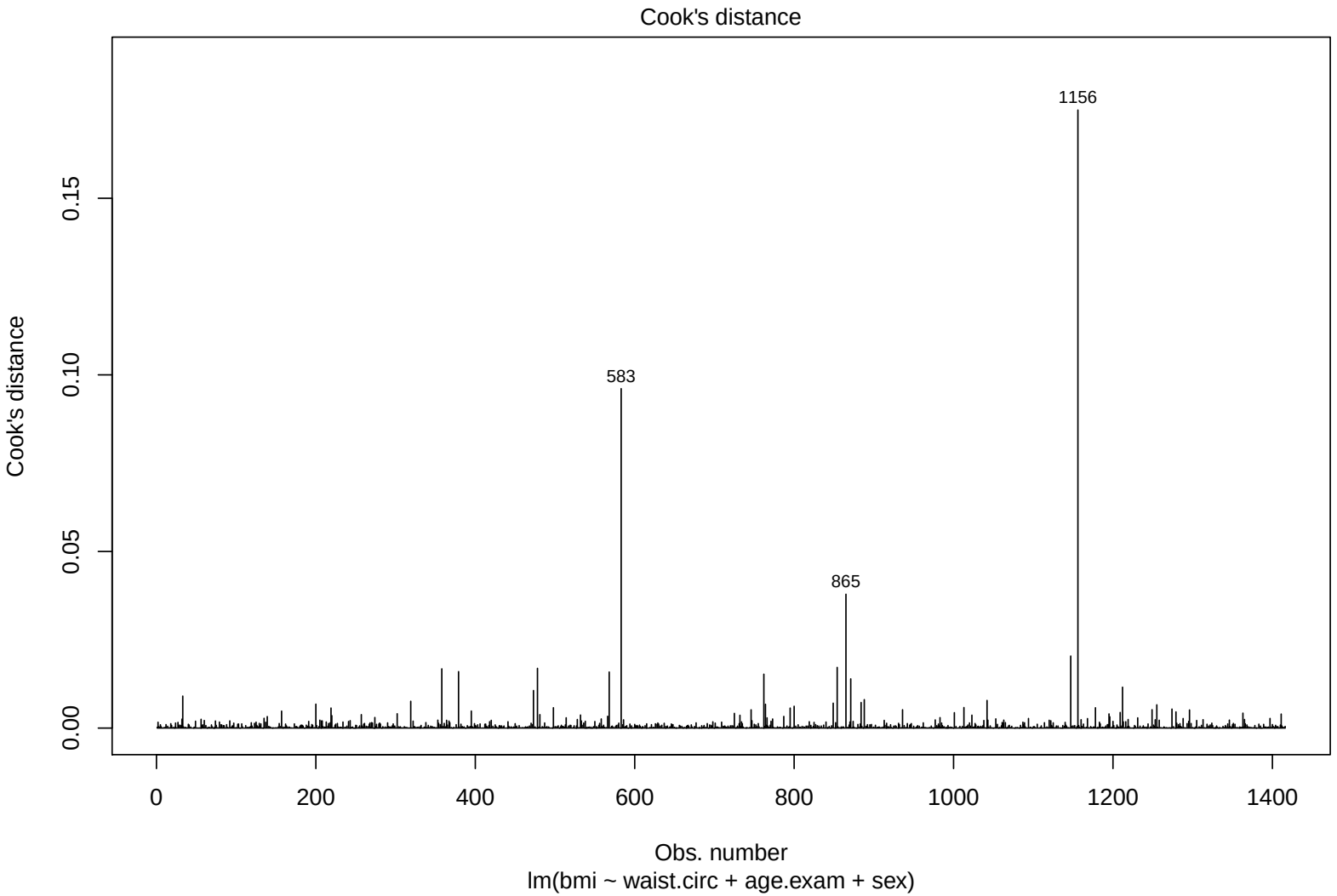
Standardized Residuals `plot(m2, which = 2)`



Residuals, Leverage, Influence `plot(m2, which = 5)`



Cook's Distance Index Plot `plot(m2, which = 4)`



Collinearity in Model m2

```
vif(m2)
```

```
waist.circ    age.exam         sex
  2.006275    2.006268    1.000007
```

- If two predictors (A and B) are highly correlated (collinear) with one another, then the predictive value of the second one into the model (B) will be masked by its strong correlation with A if A is already in the model.
- If we see a variance inflation factor above 5 (certainly above 10) this will indicate that we have some highly correlated predictors and we might be better off including one or the other in our final model.
- In this case, while `waist.circ` and `age.exam` have some correlation with each other ($r = 0.7$ from our scatterplot matrix), it's not enough to worry us on this score.

Conclusions from `m2`?

Model `m2` includes three inputs to predict `bmi`:

```
bmi = - 1.45
      + 0.35 waist.circ
      - 0.29 age.exam
      + 0.19 if sex = Male
```

- $R^2 = 0.8635$ (adjusted $R^2 = 0.8633$)
- global F test is highly significant ($p < 0.0001$)
- some issues with our residual plots
- no sign of important collinearity

Should We Transform our Outcome? And, if so, how?

Transforming / Re-expressing our Outcome?

We can use the Box-Cox family of transformations to isolate specific choices of the power transformation parameter λ for re-expressing our quantitative outcome which might lead to a more effective (yet still interpretible) model.

This approach is appropriate for strictly **positive** outcomes. If our minimum value is -14, we might add 15 to each observation before using Box-Cox.

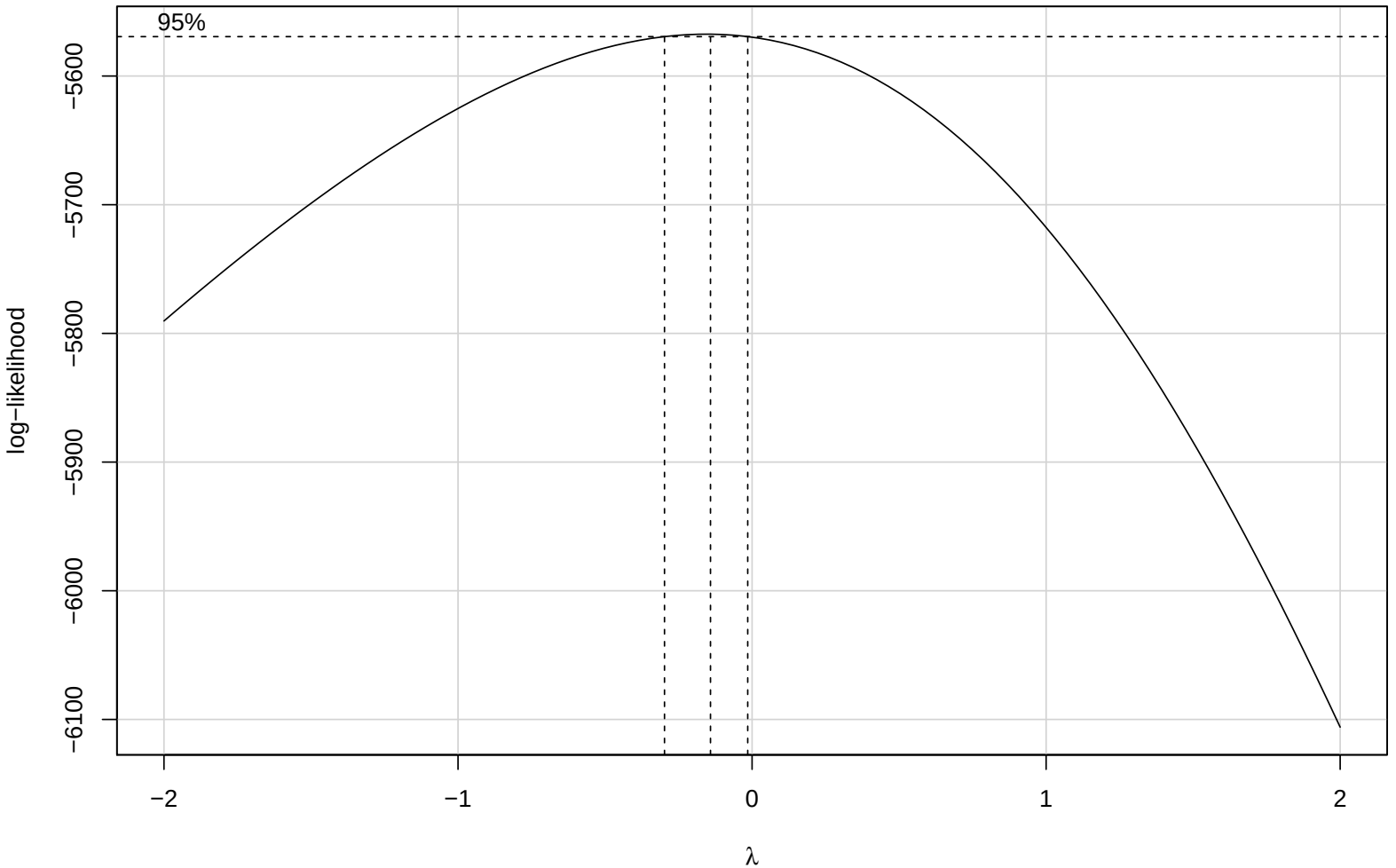
Ladder of Power Transformations

Power (λ)	Transformation
2	y^2
1	y (untransformed)
0.5	$\sqrt{y}$
0	$\log y$
-1	$\frac{1}{y}$

From the `car` package, we use `boxCox` and `powerTransform`.

The Box-Cox Transformation Plot for our model m2

```
boxCox(lm(bmi ~ waist.circ + age.exam + sex, data = nyfs1))
```



Power Transformation details for our model m2

```
powerTransform(lm(bmi ~ waist.circ + age.exam + sex,
                  data = nyfs1))
```

```
Estimated transformation parameters
        Y1
-0.1543853
```

What re-expression of the bmi data does the Box-Cox plot suggest?

Model `m3`: log(`bmi`) as a function of waist circumference, age at exam and sex

Model m3

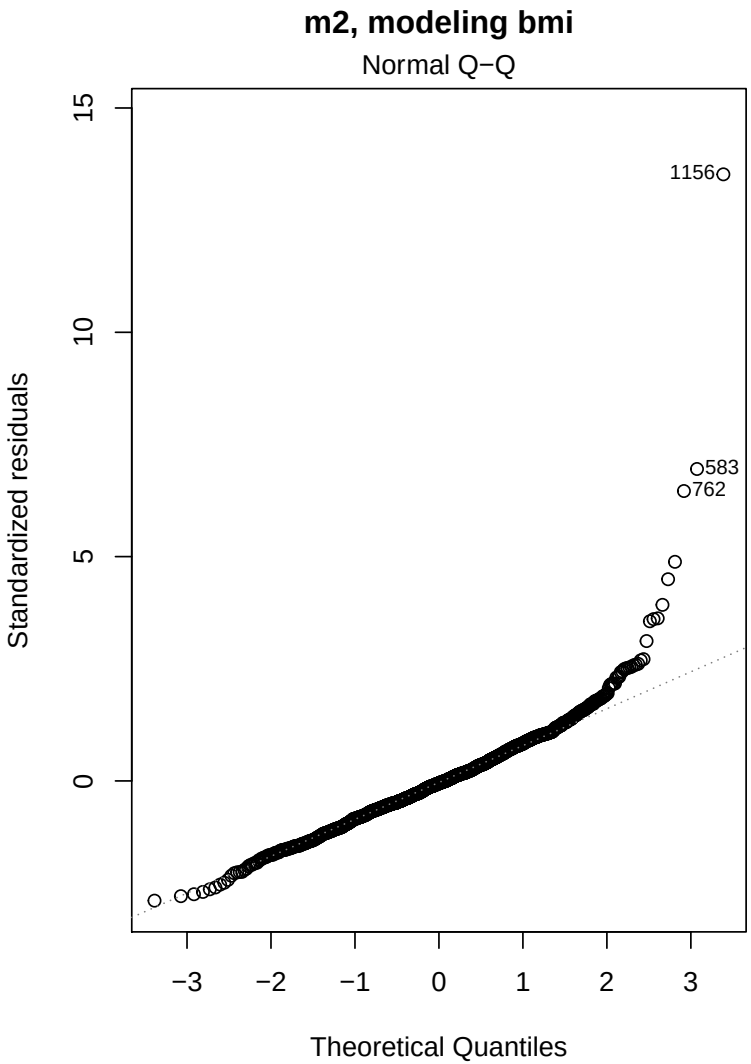
```
m2 <- nyfs1 %$% lm(bmi ~ waist.circ + age.exam + sex)
m3 <- nyfs1 %$% lm(log(bmi) ~ waist.circ + age.exam + sex)

glance(m2)
```

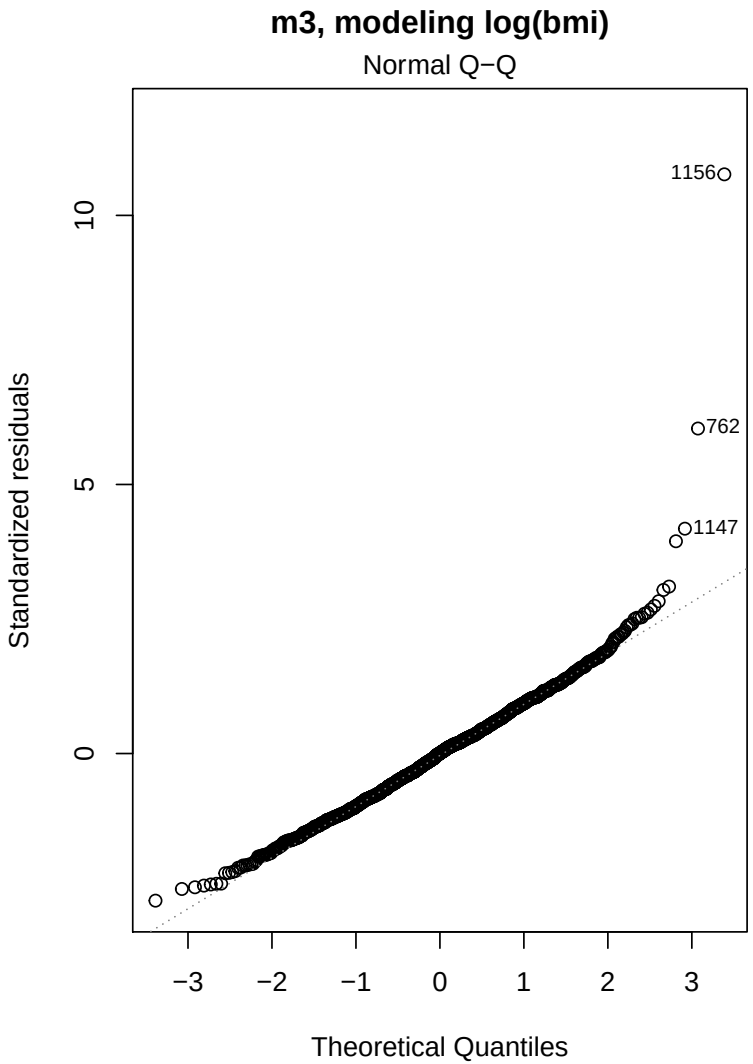
```
  r.squared adj.r.squared    sigma statistic p.value
1 0.8635437     0.8632538 1.509103  2978.545       0
  df   logLik      AIC      BIC deviance df.residual
1  4 -2589.92 5189.84 5216.118 3215.678        1412
```

```
glance(m3)
```

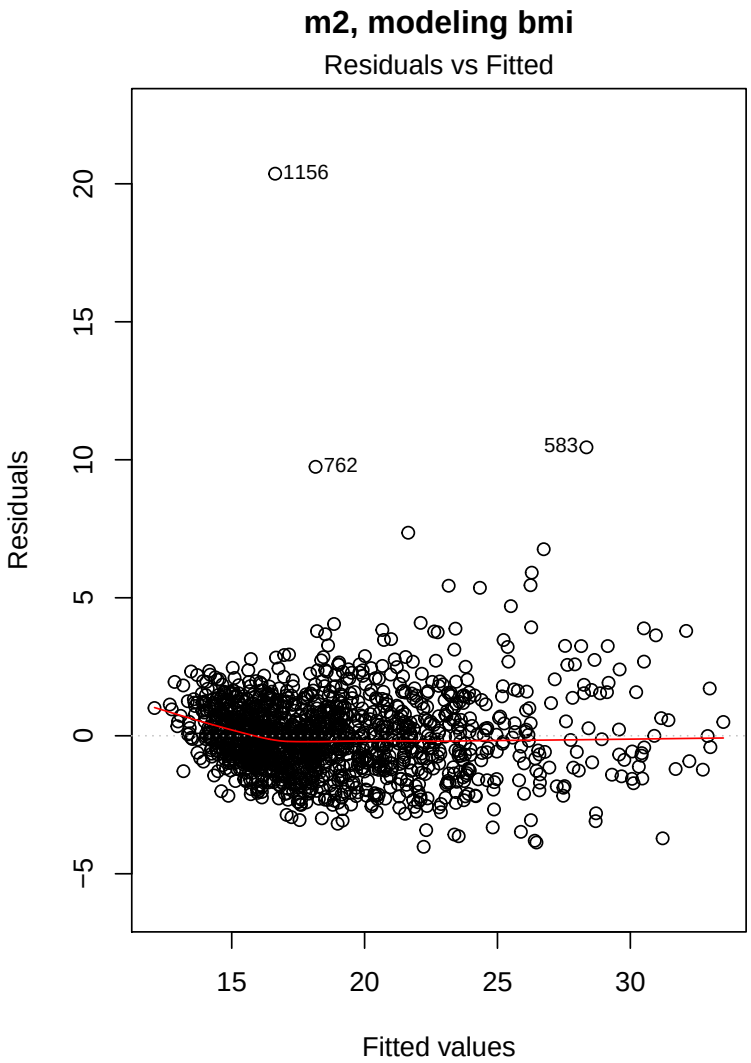
```
  r.squared adj.r.squared      sigma statistic p.value
1 0.8624454     0.8621531 0.07384071  2951.005       0
  df   logLik       AIC       BIC deviance df.residual
1  4 1682.663 -3355.325 -3329.047 7.698859        1412
```

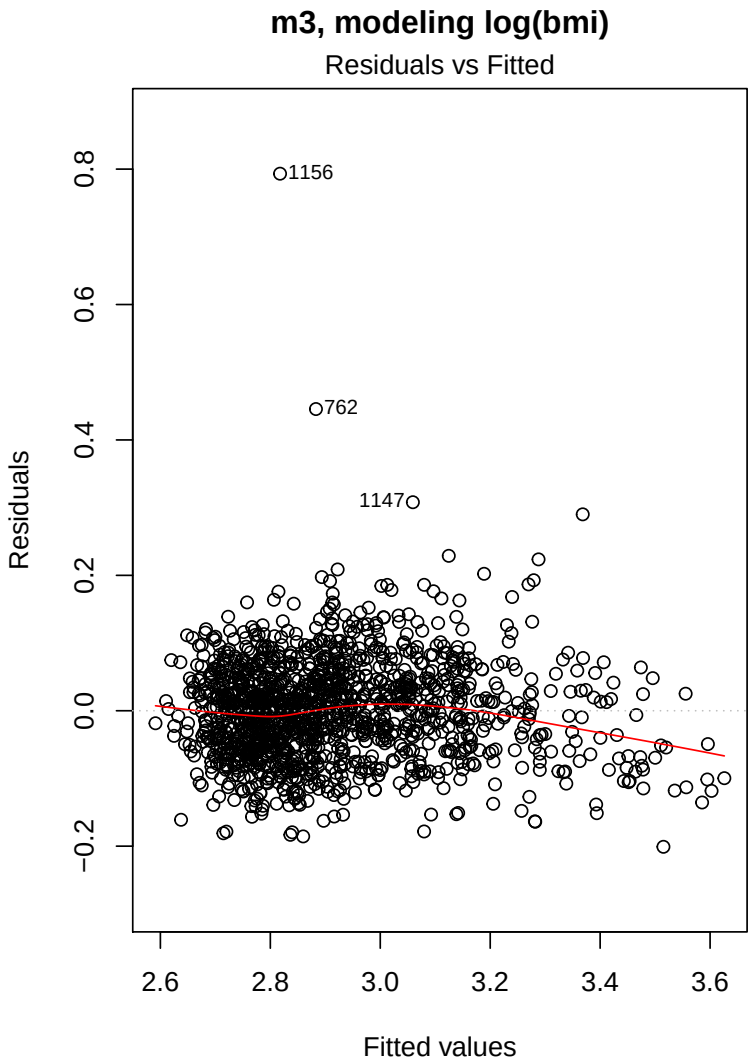
Does the Normality problem improve with `m3`?



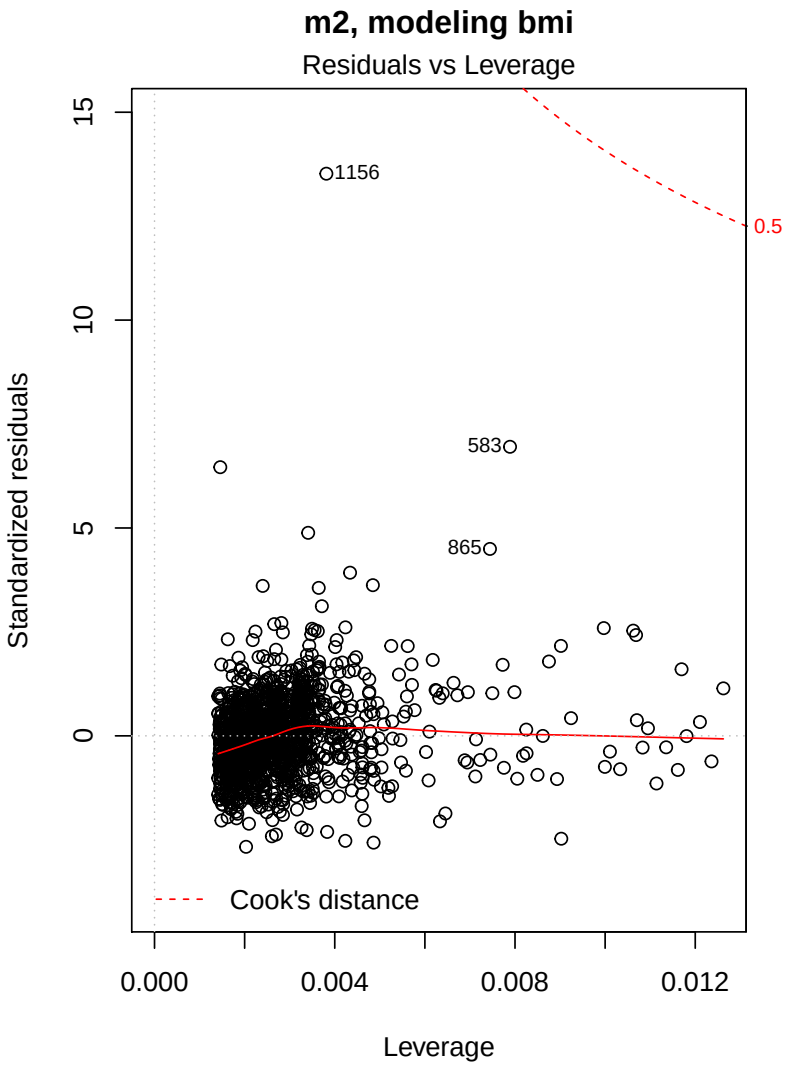


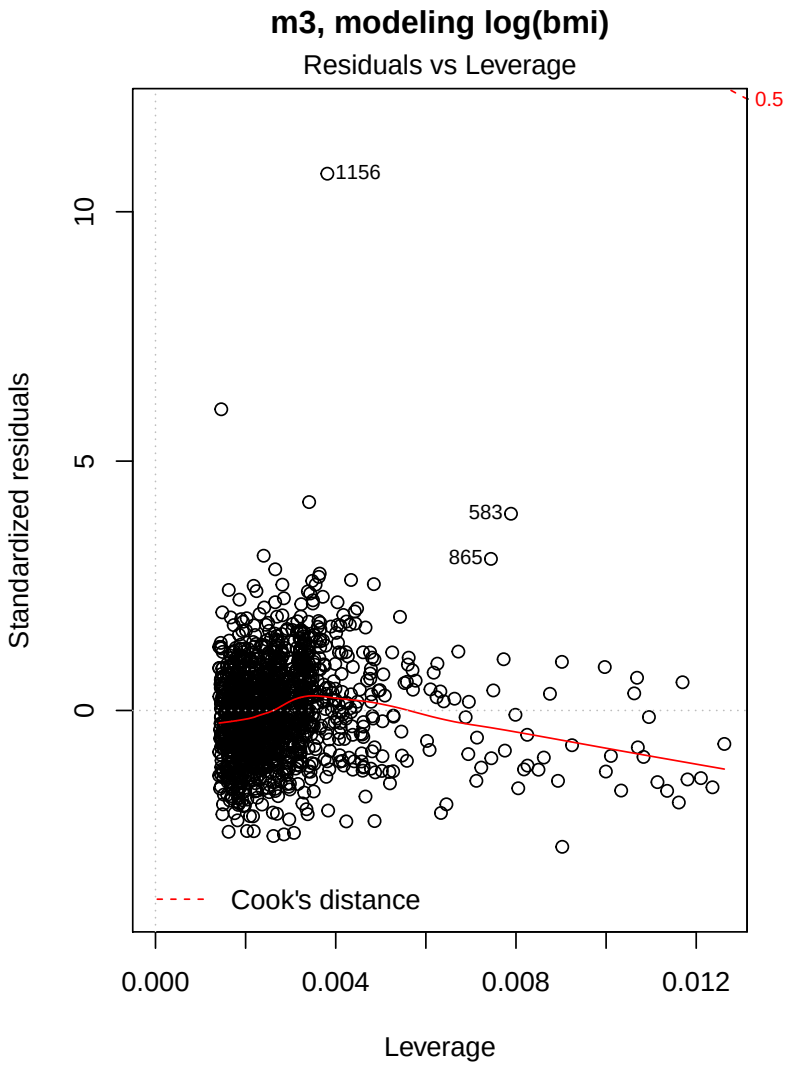
More of a "fuzzy football" with log(bmi): m3?





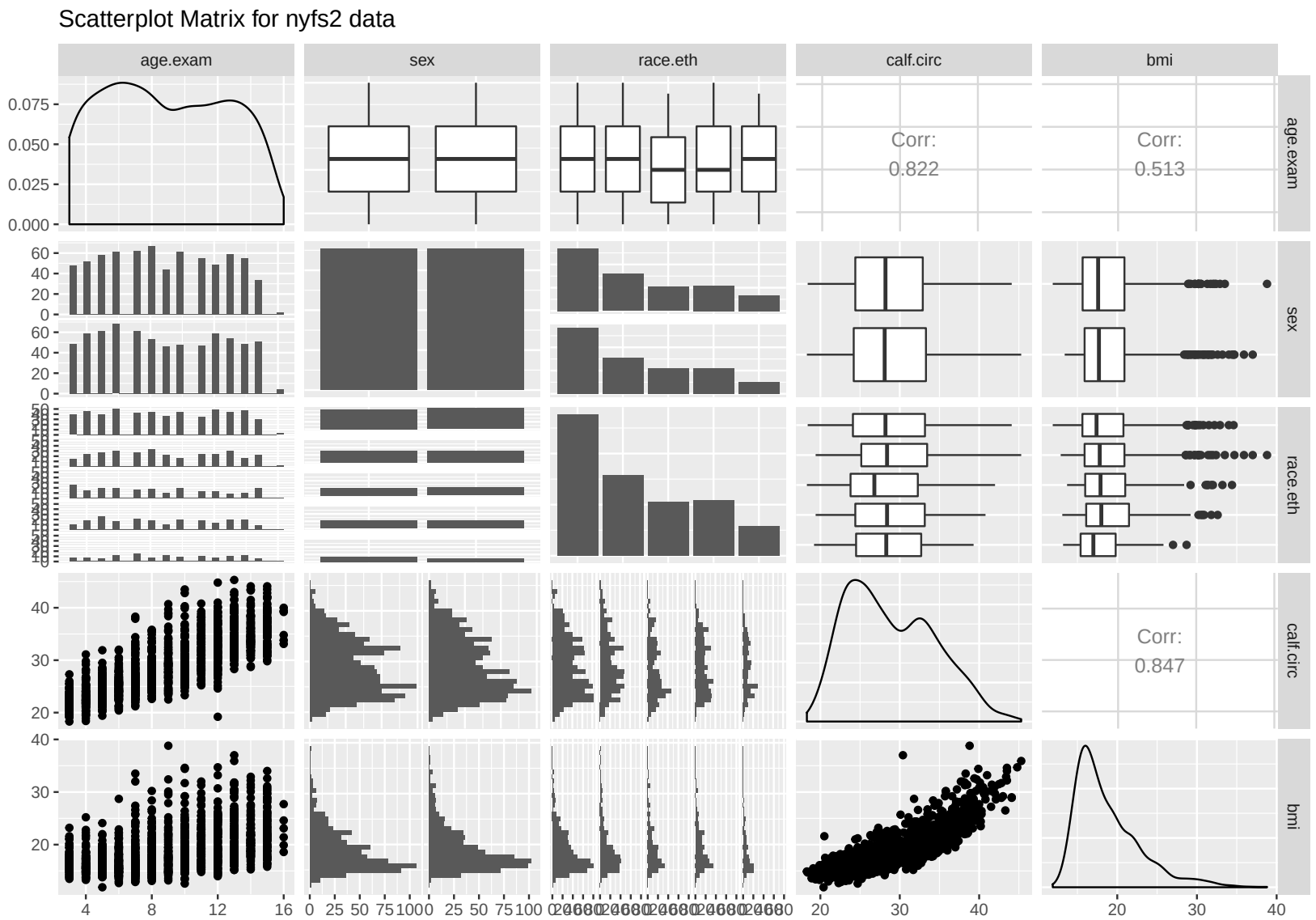
Residuals/Leverage/Influence comparing models





Model `m4` for the expanded `nyfs2` data set: Predicting `bmi` using `calf.circ`, `age.exam`, `sex` and `race.eth`

(Part of) the expanded `nyfs2` data



The Kitchen Sink Model (m4)

```
m4 <- lm(bmi ~ calf.circ + age.exam + race.eth + sex,
         data = nyfs2)

glance(m4)
```

```
  r.squared adj.r.squared    sigma statistic p.value
1 0.8300404     0.8291954 1.686596  982.3316       0
  df    logLik      AIC      BIC deviance df.residual
1  8 -2745.366 5508.732 5556.033 4005.205        1408
```

Summary of m4 (output lightly edited)

```
bmi ~ calf.circ + age.exam + race.eth + sex, data = nyfs2
Multiple R-squared:   0.83, Adjusted R-squared:  0.8292
F-statistic: 982.3 on 7 and 1408 DF,  p-value: < 2.2e-16

Coefficients:                      Estimate    SE      t    p
(Intercept)                          -4.005 0.291 -13.77 < 2e-16 ***
calf.circ                             0.972 0.014  67.83 < 2e-16 ***
age.exam                             -0.625 0.021 -29.10 < 2e-16 ***
race.eth2 Non-Hispanic Black         -0.026 0.119  -0.22   0.825
race.eth3 Mexican American            0.997 0.136   7.31 4.4e-13 ***
race.eth4 Other Hispanic              0.375 0.135   2.78   0.006 **
race.eth5 Other or Multi-Race        -0.130 0.172  -0.76   0.450
sexMale                               0.021 0.090   0.24   0.812

Residuals:    Min      1Q     Med      3Q     Max       SE
           -4.539 -0.994 -0.127   0.796 19.595    1.687
```

Can We/Should We Simplify the Model?

1. t tests for individual predictors as "last predictor in"
2. F tests for groups of predictors (order matters)
3. Stepwise Variable Selection with `step` using AIC

t test for `sex` looks NOT significant

Parsimony is an attractive feature in a model. We often want to reduce the number of regression inputs to our model, particularly if some of those inputs add no significant predictive value given the other regression inputs.

- Would the model be better without sex?
  - t test p value is 0.812
  - t test uses "Last Predictor In" approach
- Would the model be better without race.eth, too?
  - Here we have four t tests, some sig, some NS
  - What we need is a test of whether `race.eth` as an input is significant or not

anova(m4) Results (lightly edited)

```
Analysis of Variance Table

Response: bmi
             Df  Sum Sq Mean Sq   F value    Pr(>F)
calf.circ     1 16917.2 16917.2 5947.1300 < 2.2e-16 ***
age.exam      1  2451.2  2451.2  861.7043 < 2.2e-16 ***
race.eth      4   191.8    48.0   16.8576 1.653e-13 ***
sex           1     0.2     0.2    0.0566     0.812
Residuals  1408  4005.2     2.8
```

ANOVA tests whether each regression input adds significant value given that the preceding inputs are already in the model. So, order matters here.

ANOVA for m4 built in a different order

```
anova(lm(bmi ~ calf.circ + sex + age.exam + race.eth,
         data = nyfs2))
```

```
Analysis of Variance Table

Response: bmi
             Df  Sum Sq Mean Sq   F value    Pr(>F)
calf.circ     1 16917.2 16917.2 5947.1300 < 2.2e-16 ***
sex           1     2.8     2.8    0.9997    0.3176
age.exam      1  2448.7  2448.7  860.8399 < 2.2e-16 ***
race.eth      4   191.6    47.9   16.8379 1.715e-13 ***
Residuals  1408  4005.2     2.8
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Stepwise (Backwards Elimination) Variable Selection

The rest of the output follows on the next slide. . .

```
step(m4)
```

```
Start:  AIC=1488.3
bmi ~ calf.circ + age.exam + race.eth + sex

            Df Sum of Sq     RSS    AIC
- sex        1       0.2  4005.4 1486.4
<none>                    4005.2 1488.3
- race.eth   4     191.6  4196.8 1546.5
- age.exam   1    2408.5  6413.7 2153.0
- calf.circ  1   13086.9 17092.1 3540.9

Step:  AIC=1486.36
bmi ~ calf.circ + age.exam + race.eth
```

Second Part of the `step(m4)` output

```
Step:  AIC=1486.36
bmi ~ calf.circ + age.exam + race.eth

             Df Sum of Sq     RSS    AIC
<none>                    4005.4 1486.4
- race.eth    4     191.8  4197.2 1544.6
- age.exam    1    2410.3  6415.6 2151.4
- calf.circ   1   13098.2 17103.6 3539.9

Call:
lm(formula = bmi ~ calf.circ + age.exam + race.eth,
     data = nyfs2)
```

Model m5: Leaving out sex from m4

```
m4 <- lm(bmi ~ calf.circ + age.exam + race.eth + sex,
         data = nyfs2)

m5 <- lm(bmi ~ calf.circ + age.exam + race.eth,
         data = nyfs2)

select(glance(m4), r.squared, adj.r.squared, AIC)
```

```
  r.squared adj.r.squared      AIC
1 0.8300404     0.8291954 5508.732
```

```
select(glance(m5), r.squared, adj.r.squared, AIC)
```

```
  r.squared adj.r.squared      AIC
1 0.8300336     0.8293098 5506.789
```

Model m5 coefficients (output edited)

Coefficients:	Estimate	SE	t	p
(Intercept)	-3.996	0.288	-13.87	< 2e-16
calf.circ	0.972	0.014	67.88	< 2e-16
age.exam	-0.625	0.021	-29.12	< 2e-16
race.eth2 Non-Hispanic Black	-0.027	0.119	-0.23	0.822
race.eth3 Mexican American	0.997	0.136	7.31	4.3e-13
race.eth4 Other Hispanic	0.375	0.135	2.78	0.006
race.eth5 Other or Multi-Race	-0.132	0.172	-0.77	0.443

anova(m5)

```
Analysis of Variance Table

Response: bmi
             Df  Sum Sq Mean Sq  F value    Pr(>F)
calf.circ     1 16917.2 16917.2 5951.115 < 2.2e-16 ***
age.exam      1  2451.2  2451.2  862.282 < 2.2e-16 ***
race.eth      4   191.8    48.0   16.869 1.618e-13 ***
Residuals  1409  4005.4     2.8
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Signs of Meaningful Collinearity in model m5?

```
vif(m5)
```

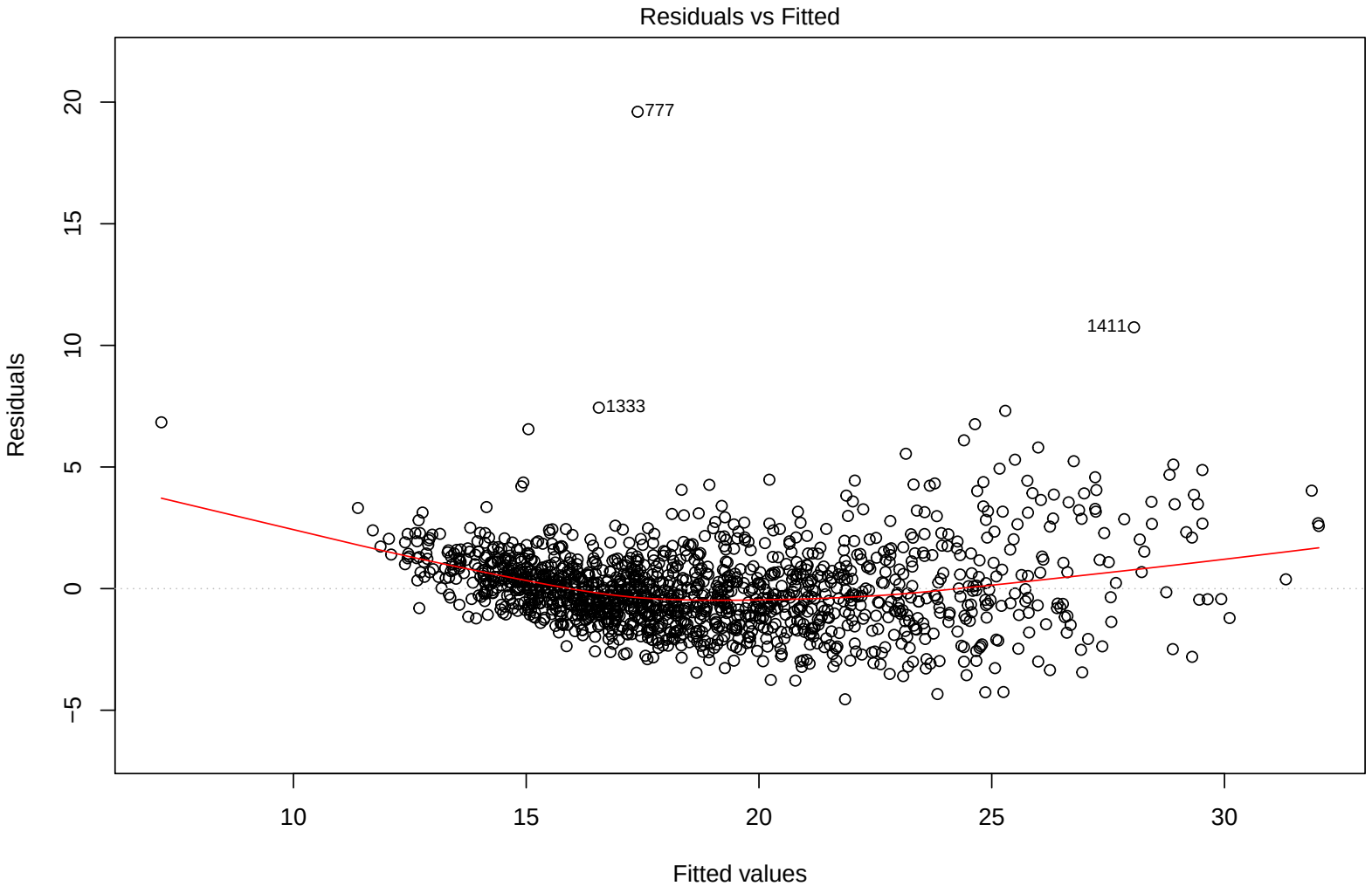
```
                GVIF Df GVIF^(1/(2*Df))
calf.circ   3.114603  1        1.764824
age.exam    3.105121  1        1.762135
race.eth    1.013989  4        1.001738
```

Note the use of a generalized variance inflation factor here. This will be used if any of the regression inputs are associated with more than one degree of freedom, usually because of indicator variables representing a multi-categorical variable.

As none of these values exceed 5 (let alone 10), again, we don't have any serious concerns.

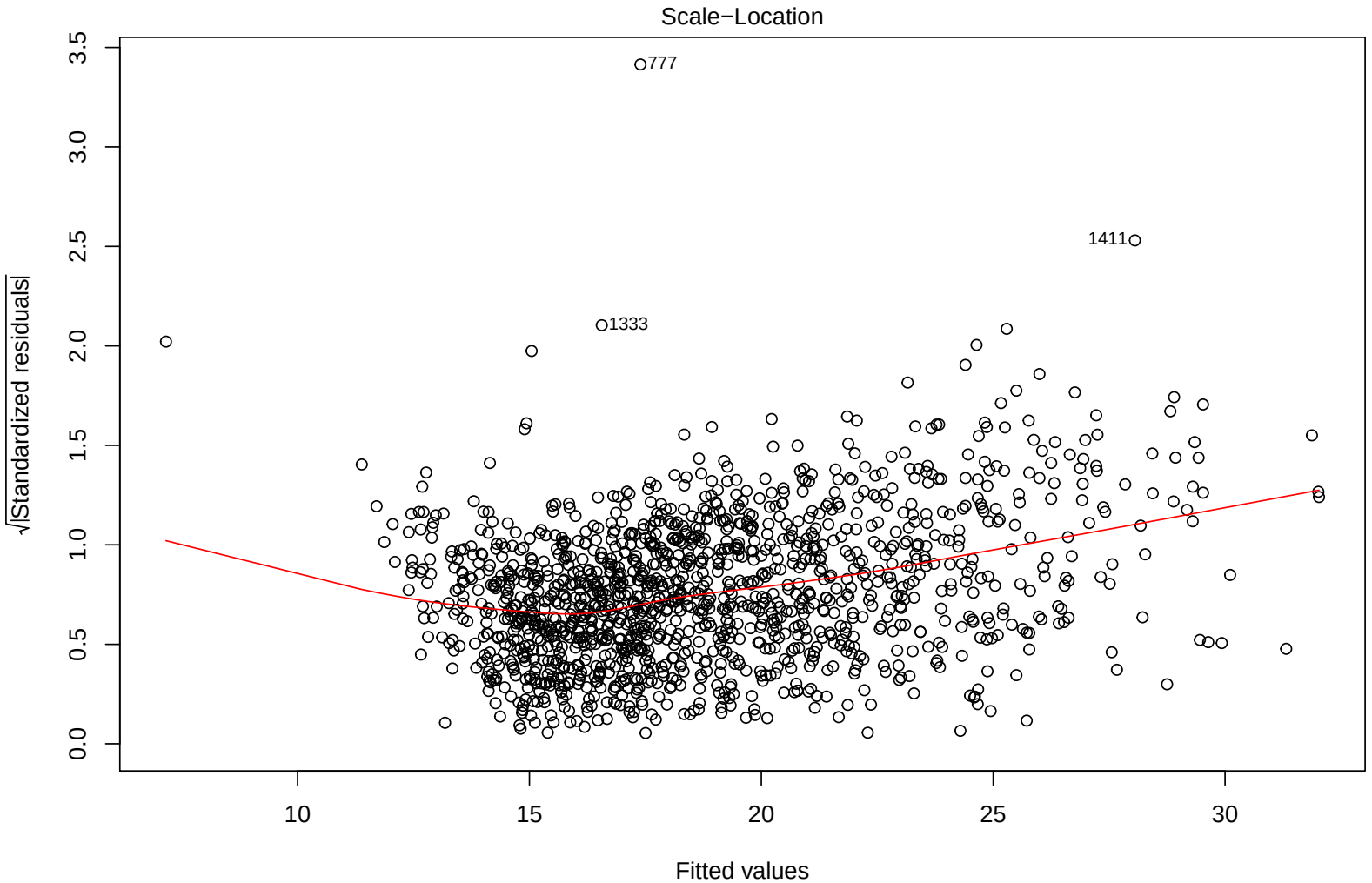
Check Assumptions via Residuals for m5

```
plot(m5, which = 1)
```



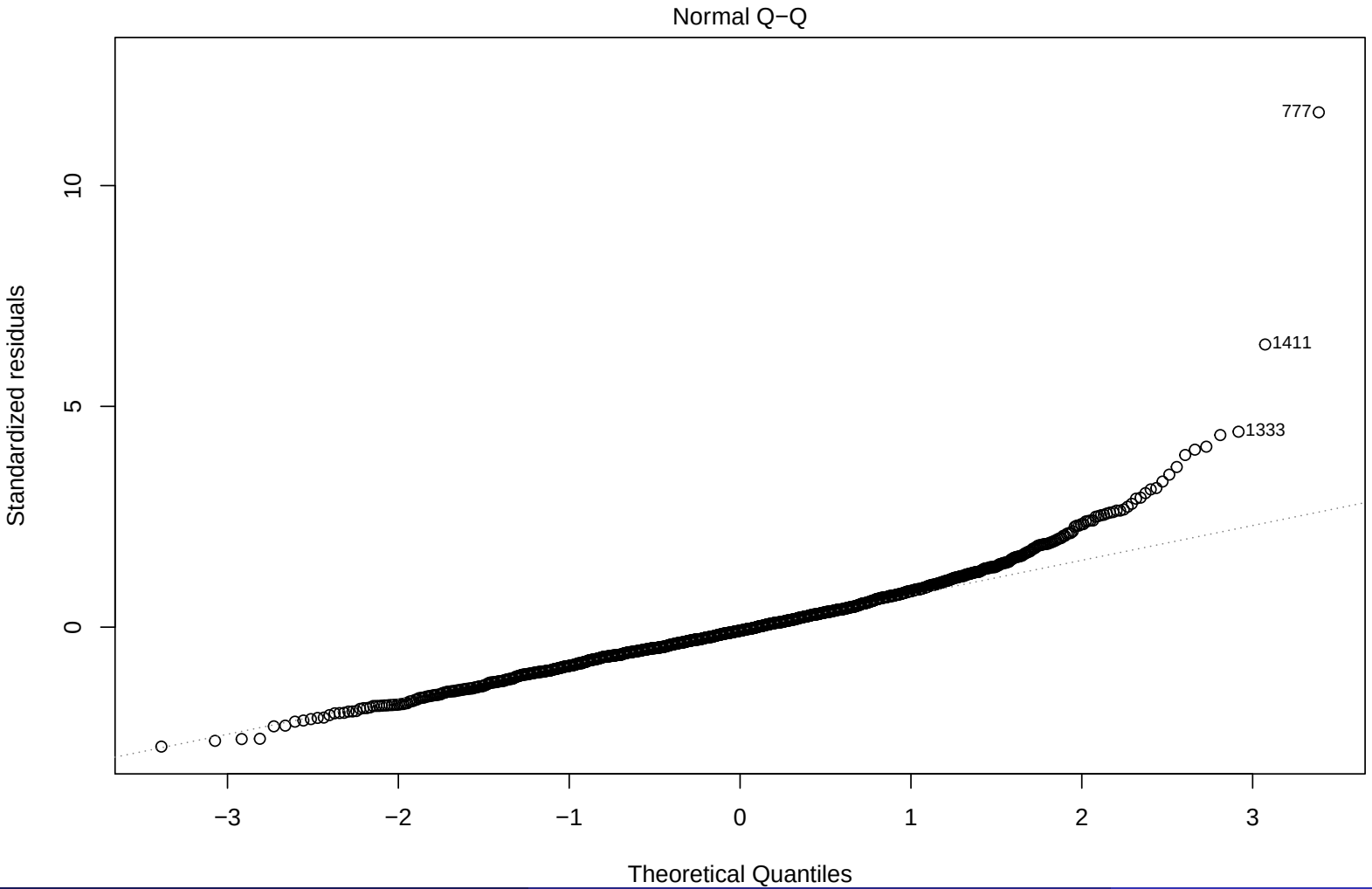
Problems with the "constant variance" assumption?

```
plot(m5, which = 3)
```



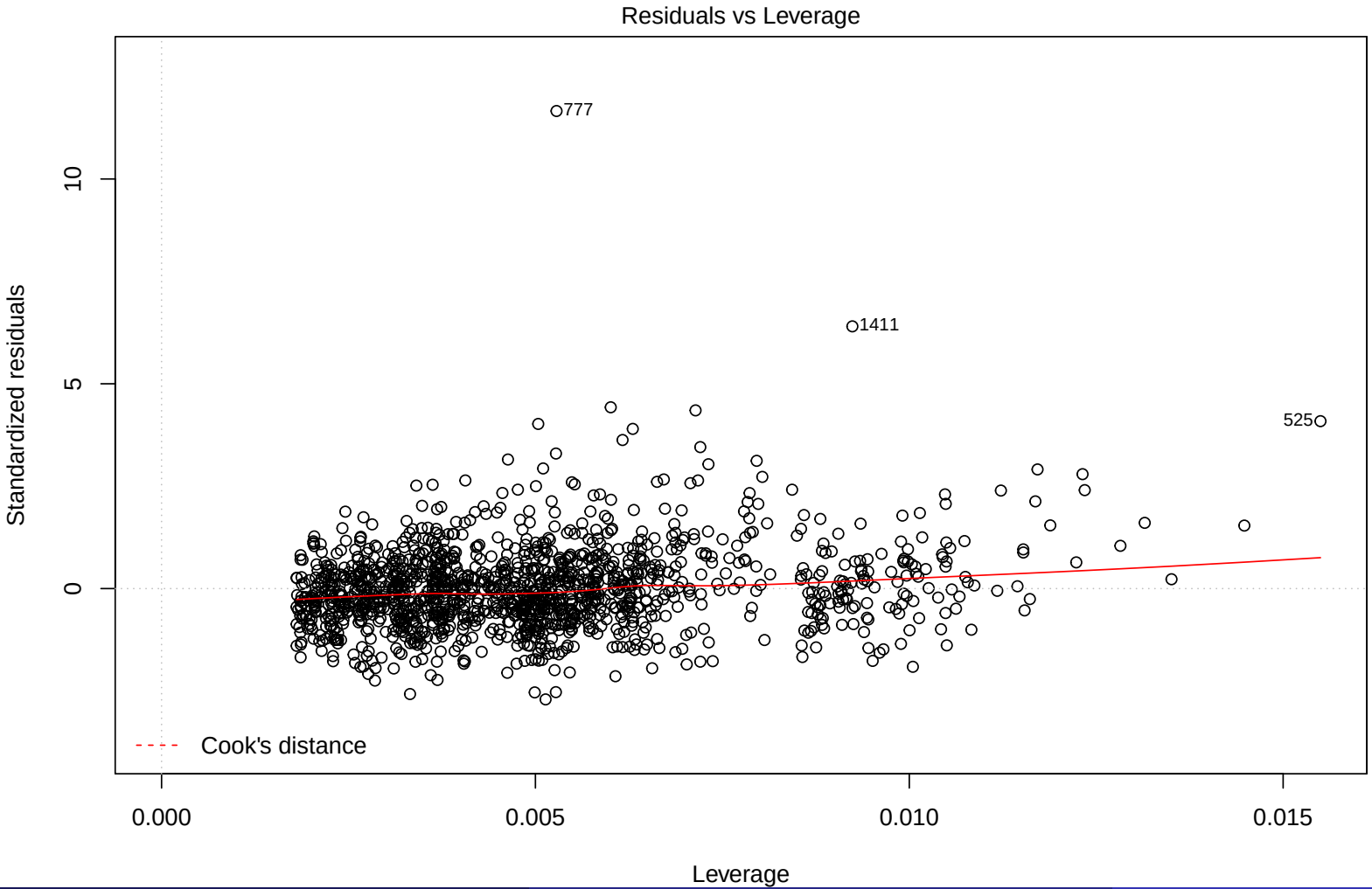
Are the m5 residuals Normally distributed?

```
plot(m5, which = 2)
```



Any influential points in m5?

```
plot(m5, which = 5)
```



Conclusions from `m5`?

Model `m5` includes three inputs to predict `bmi`:

```
bmi = - 4.00 + 0.97 calf.circ - 0.62 age.exam
      - 0.03 if race.eth = Non-Hispanic Black
      + 1.00 if race.eth = Mexican American
      + 0.37 if race.eth = Other Hispanic
      - 0.13 if race.eth = Other or Multi-Race
```

- $R^2 = 0.83$ (adjusted $R^2 = 0.829$)
- global F test is highly significant ($p < 0.0001$)
- still some issues with our residual plots
- no signs of meaningful collinearity

What Did We Discuss Today?

- Use `eda.ksam` instead of `eda.2sam`
- Box-Cox to identify potential outcome transformations
- Adjusted R^2 and when that indicates severe overfitting
- Testing/Pruning a Kitchen Sink Model
  - Hypothesis testing approaches, t and F tests
  - Using `step` for variable selection using `AIC`
- Identification of meaningful collinearity with `vif`
- Checking Assumptions with Residual Plots
- Summarizing the Model: Drawing Conclusions