431 Class 23

Thomas E. Love

2018-11-27

Today's Agenda

- The WOMAN-ETAC Trial
- Regression Analysis: The Fundamentals

Today's R Setup

```
library(knitr); library(kableExtra)
library(GGally); library(janitor); library(broom)
library(tidyverse) # always load tidyverse last

etac431 <- read_csv("data/etac431.csv") %>%
  clean_names()
```

Today's Data come from the WOMAN-ETAC trial

Effect of tranexamic acid on coagulation and fibrinolysis in women with postpartum hemorrhage (WOMAN-ETAC)

- Postpartum hemorrhage (PPH) is a leading cause of maternal death. Tranexamic acid has the potential to reduce bleeding and a large randomized controlled trial of its effect on maternal health outcomes in women with PPH (The WOMAN trial) is ongoing.
- WOMAN ETAC examined the effect of tranexamic acid on fibrinolysis and coagulation in a subset of WOMAN trial participants.
- Get a free account at https://ctu-app.lshtm.ac.uk/freebird/ to gain full access to the available data, as posted by the London School of Hygiene and Tropical Medicine on 2018-07-27.
- The PubMed link to the design paper (free full text) is https://www.ncbi.nlm.nih.gov/pubmed/28317031

We will use a subset of the data, stored in `etac431.csv`.

The WOMAN-ETAC Trial

The trial data come from legally adult women with clinically diagnosed postpartum hemorrhage following vaginal delivery of a baby or cesarean section, and where the responsible clinician is "substantially uncertain" about the appropriateness of tranexamic acid administration. This sub-study of WOMAN trial participants is meant to "provide information on the mechanism of action of tranexamic acid".

- The design was that 90 women would be randomized to the treatment group (`group` = TXA) and 90 more would be randomized to the placebo group (`group` = Placebo).
- Our data set (`etac431.csv`) includes a subset of these women.
- The subset is all women where complete baseline data are available on several predictors, and who have an outcome measurement at follow-up.
- Follow-up occurred at discharge, death or day 42, whichever is earlier.

Outcome, Predictors and Modeling Objective

- The outcome we'll study is an assessment of fibrinolysis, measured by D-dimer concentration in mg/L, which is available both at baseline (`dd_mgl_b`) and follow-up (`dd_mgl_f`).
- Generally, we are worried about an elevated D-dimer concentration in these subjects.
- Exposure `group` is either `TXA` or `Placebo`, assigned at random.
- Tranexamic axid is expected to help in the process of reducing D-dimer concentration.
- Key predictors, in addition to baseline D-dimer concentration, are international normalized ratio (`inr_hclots_b`), prothrombin time (`pt_hclots_b`) in seconds, and activated partial thromboplastin time (`aptt_hclots_b`), also in seconds. All are also available at follow-up.

What's the goal of our modeling?

Predict `dd_mgl_f` using `group` assignment and some baseline predictors.

Codebook (in Excel)

	A	B	C
1	Variable Name	type	Description
2	rand_id	character	randomization identification code
3	dd_mgl_f	numeric	D-dimer concentration in mg/L, at follow-up
4	group	factor	TXA or placebo (exposure group)
5	dd_mgl_b	numeric	D-dimer concentration in mg/L, at baseline
6	pt_hclots_b	numeric	Prothrombin time, in seconds, at baseline
7	aptt_hclots_b	numeric	Activated Partial Thromboplastin Time, (sec) at baseline
8	inr_hclots_b	numeric	International Normalized Ratio, at baseline
9	pt_hclots_f	numeric	Prothrombin time, in seconds, at follow-up
10	aptt_hclots_f	numeric	Activated Partial Thromboplastin Time, (sec) at follow-up
11	inr_hclots_f	numeric	International Normalized Ratio, at follow-up
12			

Codebook (into R)

```
etac_code <- read_csv("data/etac_codebook.csv")

etac_code %>% kable()
```

Variable Name	type	Description
rand_id	character	randomization identification code
dd_mgl_f	numeric	D-dimer concentration in mg/L, at follow-up
group	factor	TXA or placebo (exposure group)
dd_mgl_b	numeric	D-dimer concentration in mg/L, at baseline
pt_hclots_b	numeric	Prothrombin time, in seconds, at baseline
aptt_hclots_b	numeric	Activated Partial Thromboplastin Time, (sec) a
inr_hclots_b	numeric	International Normalized Ratio, at baseline
pt_hclots_f	numeric	Prothrombin time, in seconds, at follow-up
aptt_hclots_f	numeric	Activated Partial Thromboplastin Time, (sec) a
inr_hclots_f	numeric	International Normalized Ratio, at follow-up

The etac431 data, as parsed by read_csv

```
glimpse(etac431)
```

```
Observations: 95
Variables: 10
$ rand_id       <chr> "2281-21", "2281-22", "2281-...
$ dd_mgl_f      <dbl> 1.153, 3.024, 1.883, 3.131, ...
$ group         <chr> "TXA", "Placebo", "TXA", "Pl...
$ dd_mgl_b      <dbl> 1.591, 5.705, 2.361, 4.378, ...
$ pt_hclots_b   <dbl> 12.6, 15.7, 20.3, 14.9, 14.6...
$ aptt_hclots_b <dbl> 26.1, 22.2, 32.1, 28.8, 32.1...
$ inr_hclots_b  <dbl> 0.93, 1.18, 1.57, 1.11, 1.09...
$ pt_hclots_f   <dbl> 12.7, 15.0, 14.0, 15.7, 14.4...
$ aptt_hclots_f <dbl> 27.3, 22.6, 25.0, 30.5, 28.3...
$ inr_hclots_f  <dbl> 0.94, 1.12, 1.04, 1.18, 1.07...
```

Any Missing Values in our Outcome or Predictors?

```
colSums(is.na(etac431))
```

```
     rand_id      dd_mgl_f         group
           0             0             0
    dd_mgl_b   pt_hclots_b aptt_hclots_b
           0             0             0
inr_hclots_b   pt_hclots_f aptt_hclots_f
           0             2             3
inr_hclots_f
           2
```

Today's Plan (First Four Steps)

1. Assess the outcome variable `dd_mgl_f` to see if any cleaning or transformation is necessary to permit us to fit a linear regression model.
2. Partition the available data into a training sample and a test sample. Use the training sample exclusively for the next few steps (specifically steps 3-6).
3. Build a first regression model to predict our outcome on the basis of two predictors: the baseline level of D-dimer concentration and exposure `group`.
4. Assess the first model within the training sample, intepreting the coefficient estimates, and basic summaries of model fit. Describe the estimated effect of exposure to the treatment (vs. placebo) according to the first model.

Today's Plan (Last Four Steps)

5. Build a second model to include additional predictors. Evaluate the second model in a similar way to the first.
6. Compare models 1 and 2 in terms of in-sample predictive accuracy with summary measures like adjusted R^2 and the residual standard error.
7. Validate models 1 and 2 in terms of out-of-sample predictive accuracy by calculating prediction errors and assessing the MAPE, MSPE and maximum error in each case, as well as visualizing the distribution of errors from the two models. Select a winning model.
8. Use that "winning" model to fit the entire data set, describe the results, and then assess key regression assumptions.

Step 1. Do we need to re-express the outcome?

Step 1. Assessing the Outcome

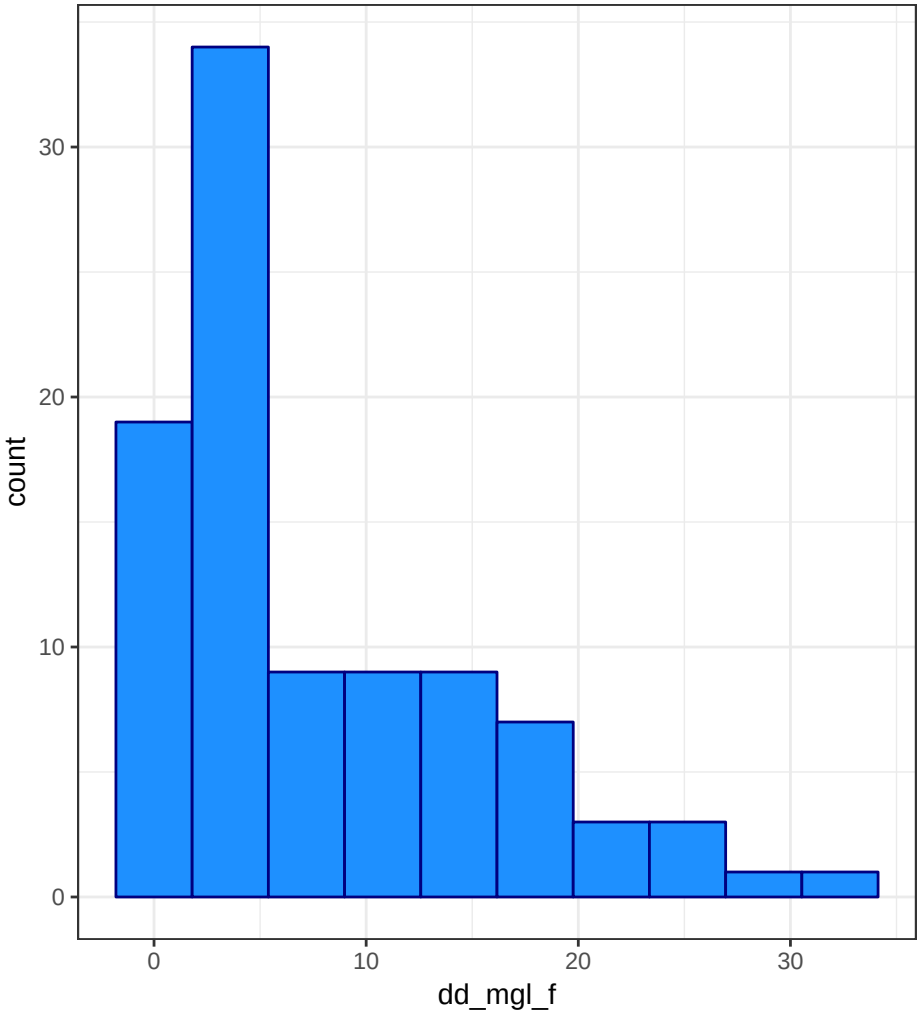
Our linear regression model will be more effective if the outcome variable is well approximated by a Normal distribution. Should we observe substantial skew in the data, it may be worthwhile to consider a transformation.

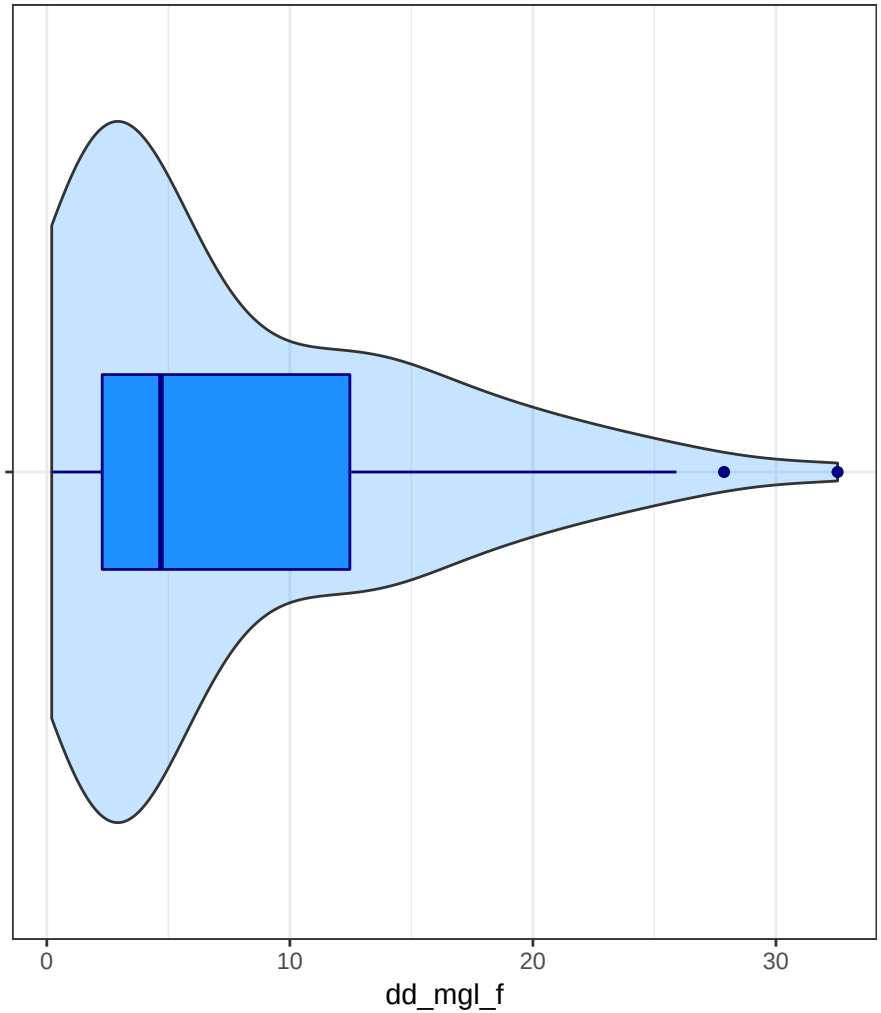
```
mosaic::favstats(~ dd_mgl_f, data = etac431) %>%
  knitr::kable(digits = 3)
```

	min	Q1	median	Q3	max	mean	sd	n	missing
	0.201	2.278	4.696	12.469	32.536	7.862	7.359	95	0

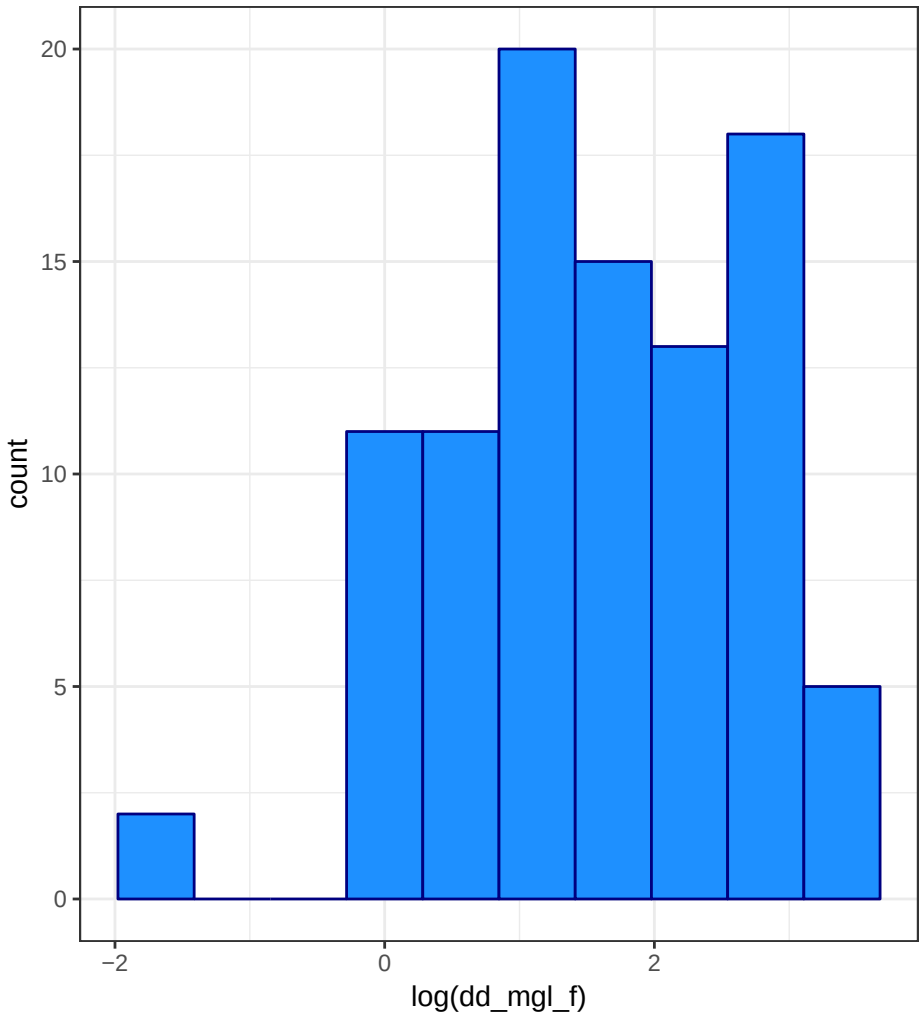
All of the values of `dd_mgl_f` are strictly positive, so our ladder of power transformations is an appealing option, if we need to transform. Of course, we need a picture. . .

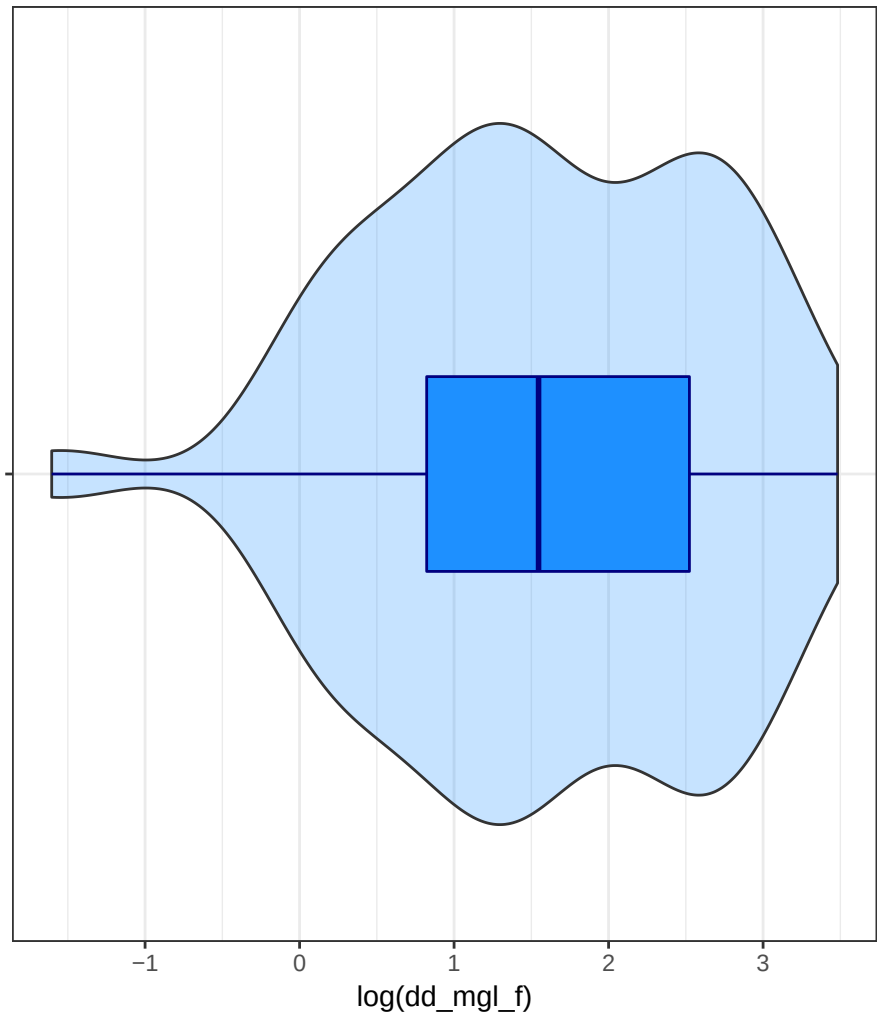
Distribution of dd_mgl_f





Distribution of log(dd_mgl_f)





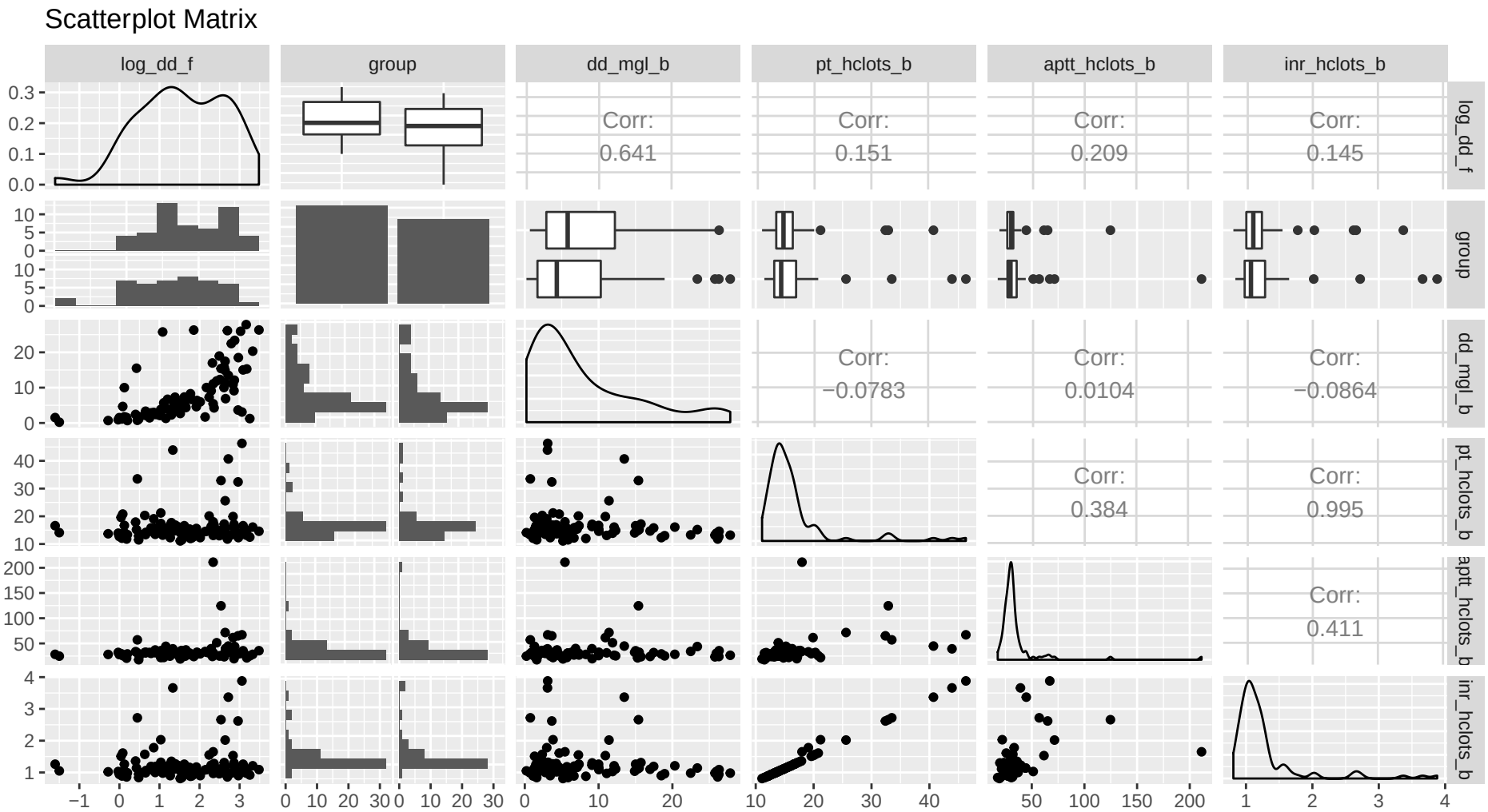
Scatterplot Matrix for etac431 (code)

```
etac431 <- etac431 %>%
  mutate(log_dd_f = log(dd_mgl_f))

etac431 %>%
  select(log_dd_f, group, dd_mgl_b,
         pt_hclots_b, aptt_hclots_b, inr_hclots_b) %>%
  ggpairs(., title = "Scatterplot Matrix",
       lower = list(combo = wrap("facethist", bins = 10)))
```

We could consider taking the logarithm of dd_mgl_b, as well, but I want to keep things simpler today.

Scatterplot Matrix for `etac431`



Step 2. Partition the data into separate training and test samples

Step 2. Partitioning the Data

In the etac431 data, we have 95 observations. Here, I'll split 70 of them into a training sample, and hold out the remaining 25 for testing.

```
set.seed(20181127)
etac431_train <- sample_n(etac431, size = 70)
etac431_test <- anti_join(etac431, etac431_train,
                          by = "rand_id")

dim(etac431_train)
```

```
[1] 70 11
```

```
dim(etac431_test)
```

```
[1] 25 11
```

Distribution of Exposure Group in our Partition

Did we get very unlucky with our partitioning?

```
etac431_train %>% tabyl(group)
```

```
   group  n   percent
 Placebo 40 0.5714286
     TXA 30 0.4285714
```

```
etac431_test %>% tabyl(group)
```

```
   group  n percent
 Placebo 11    0.44
     TXA 14    0.56
```

Step 3. Build Model 1: Predict our outcome using baseline D-dimer concentration and exposure group

Step 3. Building a Two-Predictor Model 1

```
m01 <- lm(log(dd_mgl_f) ~ group + dd_mgl_b,
          data = etac431_train)

tidy(m01) %>% select(term, estimate) %>%
  knitr::kable(digits = 3)
```

term	estimate
(Intercept)	1.132
groupTXA	-0.473
dd_mgl_b	0.086

summary(m01) (edited to fit in this space)

```
Call: lm(formula = log(dd_mgl_f) ~ group + dd_mgl_b,
                                   data = etac431_train)

Residuals:      Min       1Q   Median       3Q      Max
           -2.39418 -0.43385  0.04658  0.42628  2.01572

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.13219    0.15334   7.384 3.11e-10 ***
groupTXA      -0.47333    0.17727  -2.670  0.00951 **
dd_mgl_b       0.08565    0.01133   7.560 1.50e-10 ***
---
Sig. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7315 on 67 degrees of freedom
Multiple R-squared: 0.5037, Adjusted R-squared:  0.4889
F-statistic:   34 on 2 and 67 DF,  p-value: 6.406e-11
```

Step 4. Assess Model 1 in the Training Sample, and describe the effect of treatment exposure

Step 4. Assessing Model 1 in the Training Sample

```
tidy(m01, conf.int = TRUE, conf.level = 0.95) %>%
  knitr::kable(digits = 2)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	1.13	0.15	7.38	0.00	0.83	1.44
groupTXA	-0.47	0.18	-2.67	0.01	-0.83	-0.12
dd_mgl_b	0.09	0.01	7.56	0.00	0.06	0.11

The estimate for `groupTXA` is -0.47. How would you interpret the coefficient for `groupTXA` in this setting?

Interpreting the `tidy(m01)` output

```
tidy(m01, conf.int = TRUE, conf.level = 0.95) %>%
  knitr::kable(digits = 2)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	1.13	0.15	7.38	0.00	0.83	1.44
groupTXA	-0.47	0.18	-2.67	0.01	-0.83	-0.12
dd_mgl_b	0.09	0.01	7.56	0.00	0.06	0.11

Suppose we have two subjects with the same baseline D-dimer concentration (`dd_mgl_b`), and one receives TXA and one Placebo. Then the subject receiving TXA is predicted by Model 1 to have a final **log(D-dimer concentration)** that is 0.47 smaller than the subject receiving Placebo. Clinically, is this good or bad news for the treatment?

anova(m01) and glance(m01)

```
Analysis of Variance Table
Response: log(dd_mgl_f)
           Df Sum Sq Mean Sq F value    Pr(>F)
group       1  5.809  5.8094  10.856  0.001576 **
dd_mgl_b    1 30.584 30.5845  57.151 1.496e-10 ***
Residuals 67 35.855  0.5351
---
Sig. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
glance(m01) %>% select(r.squared, adj.r.squared, sigma) %>%
  knitr::kable(digits = 3)
```

	r.squared	adj.r.squared	sigma
value	0.504	0.489	0.732

Step 5. Build Model 2: Add other predictors at baseline

Step 5. Building a Larger Model 2

```
m02 <- lm(log(dd_mgl_f) ~ group + dd_mgl_b +
            inr_hclots_b + pt_hclots_b + aptt_hclots_b,
          data = etac431_train)

tidy(m02) %>% select(term, estimate) %>%
  knitr::kable(digits = 3)
```

term	estimate
(Intercept)	0.543
groupTXA	-0.477
dd_mgl_b	0.085
inr_hclots_b	-1.034
pt_hclots_b	0.097
aptt_hclots_b	0.009

summary(m02) (edited to fit this slide)

```
Residuals:      Min       1Q   Median       3Q      Max
            -2.36358 -0.45300  0.04045  0.45858  2.11522

Coefficients:  Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.543206   0.438376   1.239  0.21982
groupTXA      -0.476911   0.178101  -2.678  0.00941 **
dd_mgl_b       0.085452   0.011360   7.522 2.24e-10 ***
inr_hclots_b  -1.033624   2.034189  -0.508  0.61311
pt_hclots_b    0.096739   0.181138   0.534  0.59515
aptt_hclots_b  0.009111   0.008125   1.121  0.26633
---
Sig. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7268 on 64 degrees of freedom
Multiple R-squared:  0.5321,    Adjusted R-squared:  0.4956
F-statistic: 14.56 on 5 and 64 DF,  p-value: 1.619e-09
```

Step 6. Compare Models 1 and 2 in terms of in-sample predictive accuracy

Step 6. Models 1 and 2 in the Training Sample

Let's gather the key results for each model into a single tibble, which we'll display on the next slide.

```
results_1 <- glance(m01) %>% select(-logLik, -deviance) %>%
  round(digits = 3) %>% mutate(name = "m01")

results_2 <- glance(m02) %>% select(-logLik, -deviance) %>%
  round(digits = 3) %>% mutate(name = "m02")

comp_res <- bind_rows(results_1, results_2) %>%
  select(name, df, r.squared, adj.r.squared, sigma,
         AIC, BIC, p.value)
```

Comparing Model 1 to Model 2 (Training Sample)

```
comp_res %>% knitr::kable()
```

name	df	r.squared	adj.r.squared	sigma	AIC	BIC	p.value
m01	3	0.504	0.489	0.732	159.821	168.815	0
m02	6	0.532	0.496	0.727	161.695	177.434	0

What conclusions can you draw here? Which model looks like it fits the data in the training sample more effectively?

Step 7. Validate Models 1 and 2 by applying them to the Test Sample. Then select a winner

Step 7. Models 1 and 2 in the Test Sample

We use the `augment` function from `broom` to help calculate prediction errors. It's important to convert these back to the scale of the original D-dimers concentrations, and not their logarithms. To back-transform, we'll need to exponentiate.

```
test_m01 <- augment(m01, newdata = etac431_test) %>%
  mutate(modname = "m01",
         .resid = dd_mgl_f - exp(.fitted),
         .expfit = exp(.fitted)) %>%
  select(rand_id, modname, dd_mgl_f, .expfit, .resid,
         .fitted, group, everything())

head(test_m01, 2) %>% knitr::kable(digits = 2)
```

rand_id	modname	dd_mgl_f	.expfit	.resid	.fitted	group	dd_mg
2281-24	m01	1.88	2.37	-0.48	0.86	TXA	
2557-21	m01	21.28	2.52	18.76	0.92	TXA	

Gathering Prediction Errors

```
test_m02 <- augment(m02, newdata = etac431_test) %>%
  mutate(modname = "m02",
         .resid = dd_mgl_f - exp(.fitted),
         .expfit = exp(.fitted)) %>%
  select(rand_id, modname, dd_mgl_f, .expfit, .resid,
         group, everything())

test_comp <- union(test_m01, test_m02) %>%
  arrange(rand_id, modname)
```

Test Sample: First Few Results

```
head(test_comp) %>%
  select(rand_id, modname, dd_mgl_f, .expfit, .resid,
         .fitted, group) %>%
  knitr::kable(digits = 2)
```

rand_id	modname	dd_mgl_f	.expfit	.resid	.fitted	group
2281-24	m01	1.88	2.37	-0.48	0.86	TXA
2281-24	m02	1.88	2.46	-0.58	0.90	TXA
2557-21	m01	21.28	2.52	18.76	0.92	TXA
2557-21	m02	21.28	4.10	17.18	1.41	TXA
4081-41	m01	9.39	5.78	3.61	1.76	Placebo
4081-41	m02	9.39	5.99	3.41	1.79	Placebo

Boxplot of the Prediction Errors

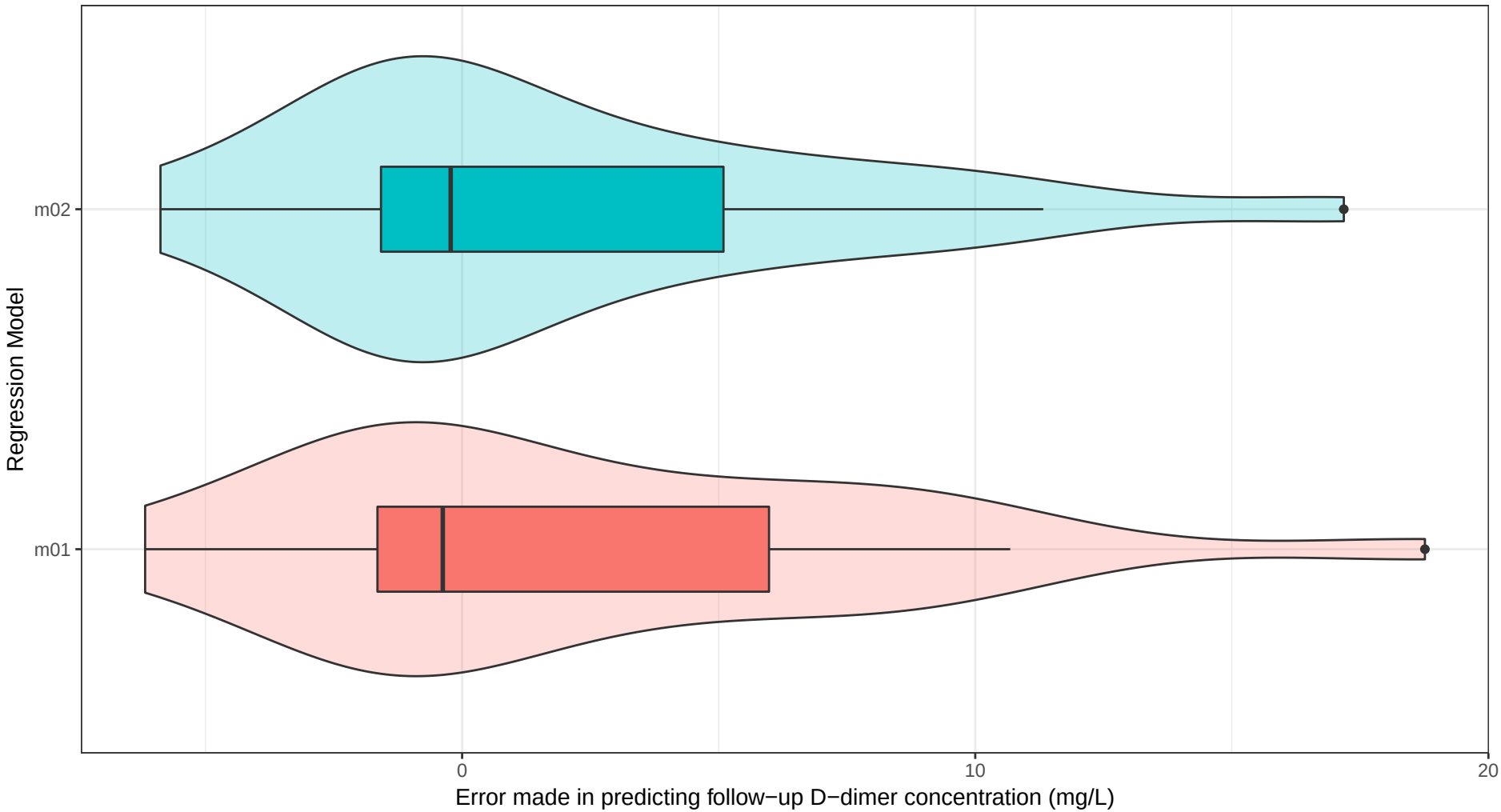


Table Comparing Model Predictions

```
test_comp %>%
  group_by(modname) %>%
  summarize(n = n(),
            MAPE = mean(abs(.resid)),
            MSPE = mean(.resid^2),
            max_error = max(abs(.resid))) %>%
  knitr::kable(digits = 3)
```

modname	n	MAPE	MSPE	max_error
m01	25	4.517	38.247	18.763
m02	25	3.967	32.119	17.182

And so now, which model looks like the winner?

Step 8. Fit the Winning Model to the Entire Data Set, and Assess Assumptions

Step 8. Winning Model in etac431

We'll use Model m02, since it did a bit better in the out-of-sample testing.

```
m_fin <- lm(log(dd_mgl_f) ~ group + dd_mgl_b +
              inr_hclots_b + pt_hclots_b + aptt_hclots_b,
            data = etac431)

tidy(m_fin, conf.int = TRUE) %>%
  select(term, estimate, conf.low, conf.high, p.value) %>%
  knitr::kable(digits = 3)
```

term	estimate	conf.low	conf.high	p.value
(Intercept)	0.111	-0.633	0.855	0.768
groupTXA	-0.446	-0.768	-0.125	0.007
dd_mgl_b	0.094	0.072	0.116	0.000
inr_hclots_b	-1.243	-4.278	1.791	0.418
pt_hclots_b	0.133	-0.130	0.396	0.318
aptt_hclots_b	0.009	0.001	0.017	0.032

summary(m_fin) (edited to fit in this slide)

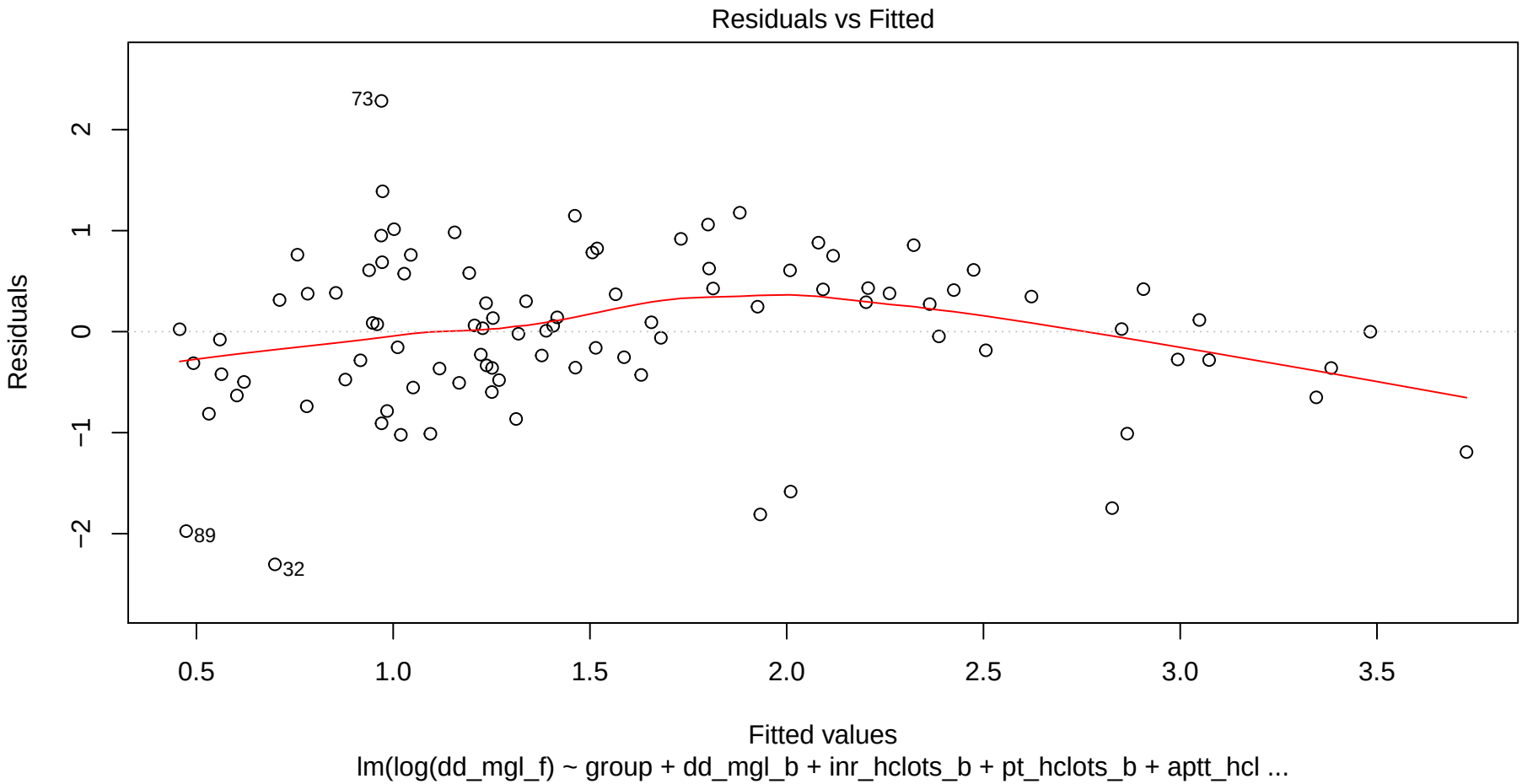
```
Residuals:      Min       1Q   Median       3Q      Max
            -2.30392 -0.39329  0.03371  0.42916  2.28403

Coefficients:  Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.110959   0.374291   0.296  0.76758
groupTXA      -0.446497   0.161819  -2.759  0.00703 **
dd_mgl_b       0.093874   0.011112   8.448  5.2e-13 ***
inr_hclots_b  -1.243175   1.527277  -0.814  0.41783
pt_hclots_b    0.132924   0.132453   1.004  0.31831
aptt_hclots_b  0.008853   0.004060   2.181  0.03186 *
---
Sig. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.779 on 89 degrees of freedom
Multiple R-squared:  0.5137,    Adjusted R-squared:  0.4864
F-statistic:  18.8 on 5 and 89 DF,  p-value: 1.032e-12
```

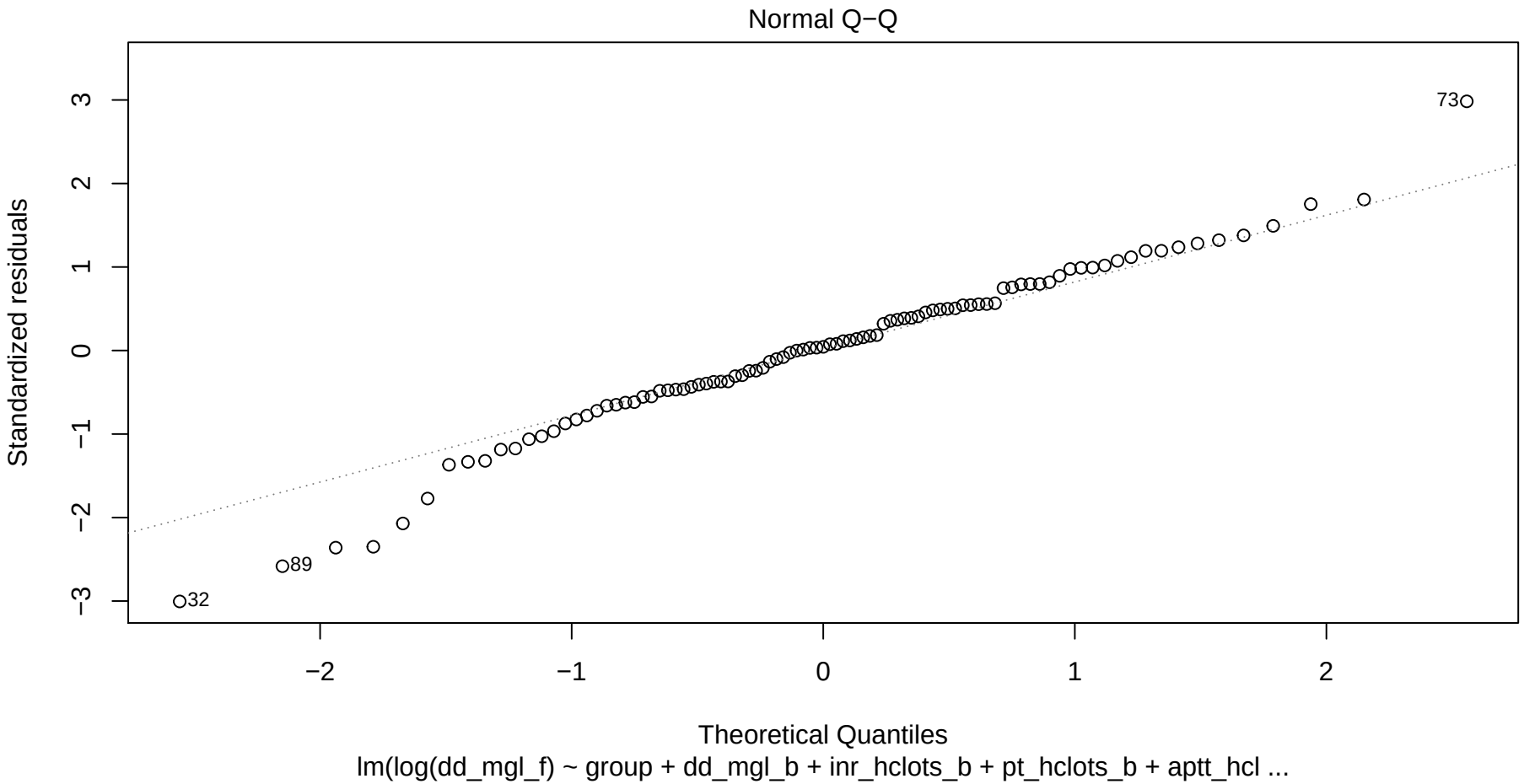
Checking Key Assumptions: Residual Plot 1

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



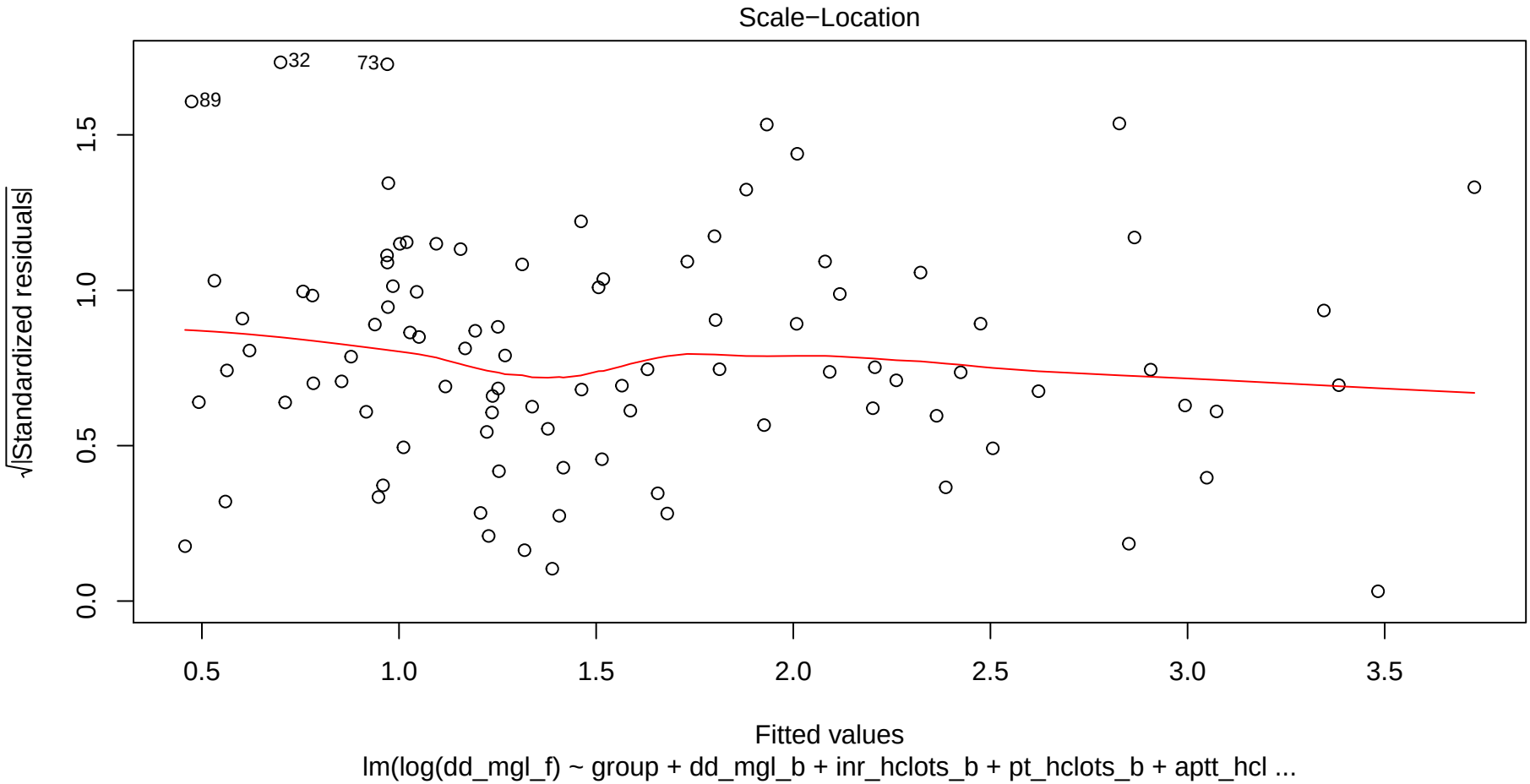
Checking Key Assumptions: Residual Plot 2

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



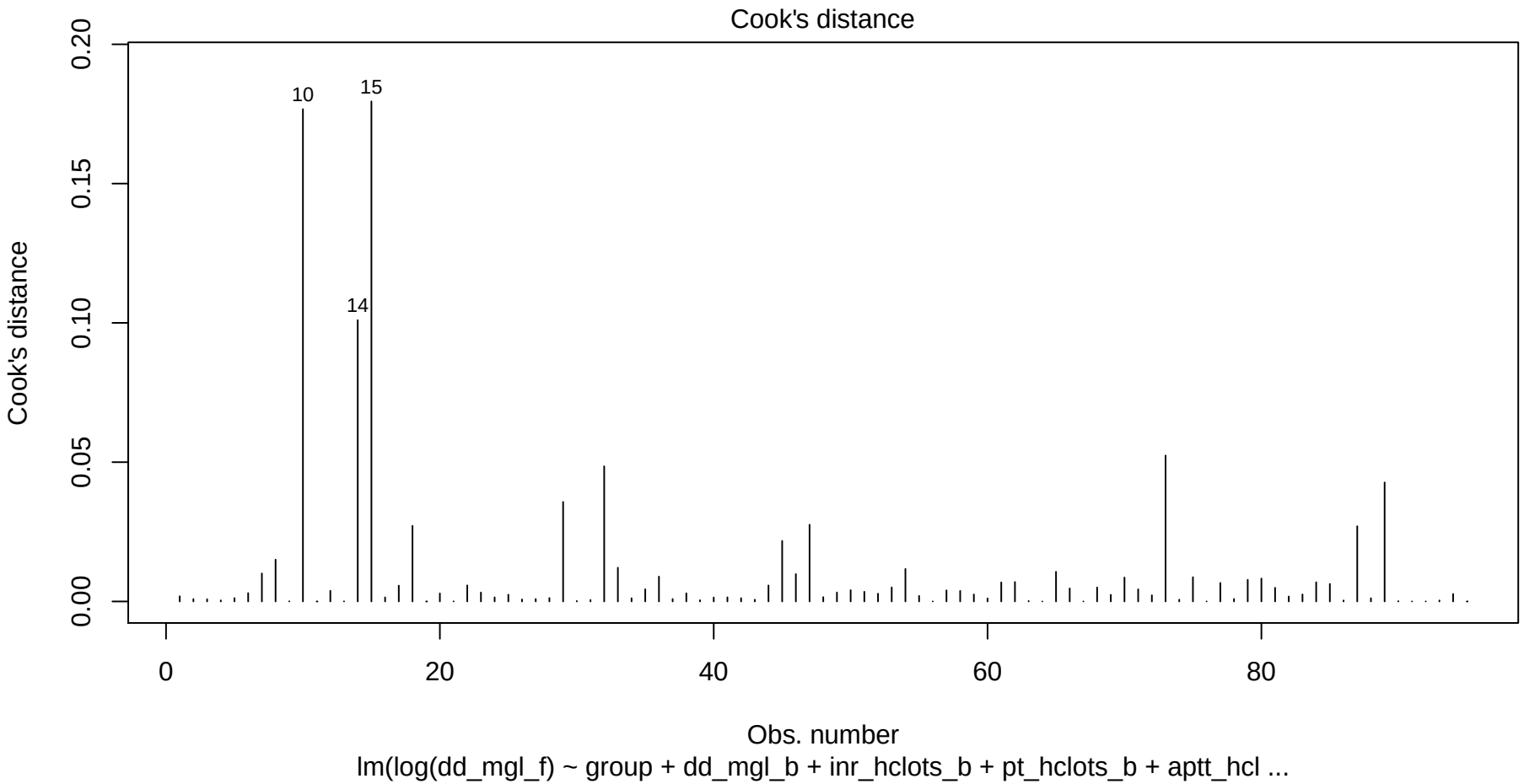
Checking Key Assumptions: Residual Plot 3

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



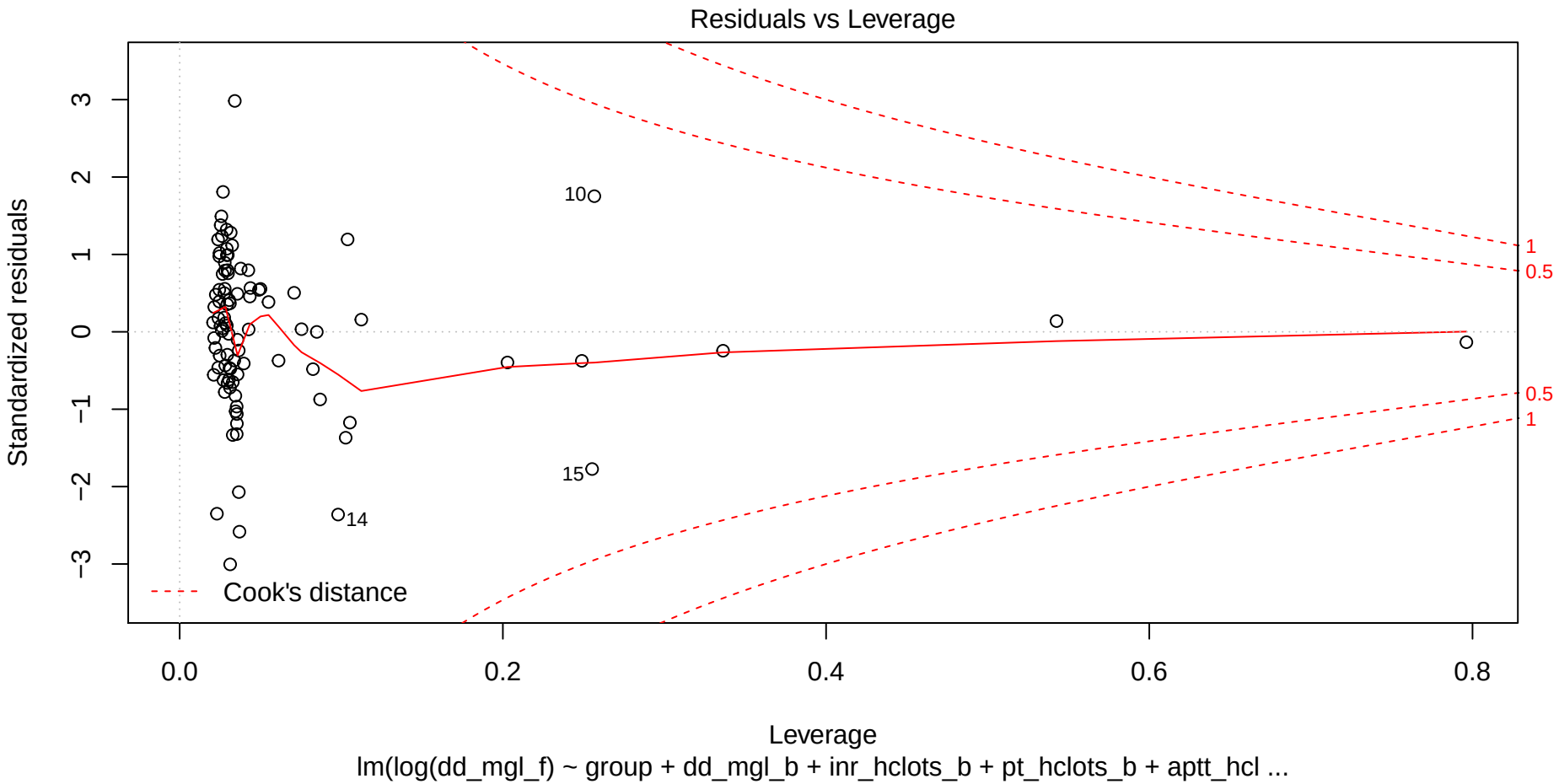
Checking Key Assumptions: Residual Plot 4

```
plot(m_fin, which = 4)
```



Checking Key Assumptions: Residual Plot 5

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



What Haven't We Done Today?

- Box-Cox approach to identifying sensible re-expressions
- Analysis of Variance to compare models
- Collinearity and the Variance Inflation Factor
- Stepwise Regression to help identify predictor sets
- Simple imputation and its impact on the model
- Making predictions / prediction vs. confidence intervals

And Maybe ...

- Standardizing our regression inputs / outputs
- Visualization of bootstrap estimated variation around our model
- Fake data simulation for model checking

plus many, many things that we'll do in 432.