MEDB 5501, Module07

2024-10-01

Topics to be covered

- What you will learn
  - Categorical independent variables
  - R code for categorical independent variables
  - Multiple linear regression
  - R code for multiple linear regression
  - Diagnostic plots and multicollinearity
  - R code for diagnostic plots and multicollinearity
  - Your homework

Categorical independent variables, 1

- Regression equation
  - $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
- How do you modify this if X_i is categorical?
  - Indicator variables
- Examples
  - Treatment: active drug=1, placebo=0
  - Second hand smoke: exposed=1, not exposed=0
  - Gender: male=1, female=0
- To be discussed later: three of more category levels

Speaker notes

The regression equation expects a numerical value for both X_i and Y_i. What if X_i is a categorical variable like treatment group, second-hand smoke exposure, or gender? You can't plug a category like "active drug" or "placebo" into this equation.

The trick is to convert your categorical variable into an indicator variable. An indicator variable is equal to 1 for a particular category and 0 for the other category.

It is a bit arbitrary which category gets the 1 and which gets the 0. I like to visualize the choice as 0 representing the absence of a quality and 1 representing the presence of a quality. So I always choose 0 for the placebo group and 1 for the active drug. I choose 0 for the unexposed group and 1 for the group with exposure.

So for gender, I always use 0 for females and 1 for males. This represents absence or presence of the y-chromosome.

Categorical independent variables, 2

- If $X_i = 0$
  - $Y_i = \beta_0 + \beta_1(0) + \epsilon_i$
  - $Y_i = \beta_0 + \epsilon_i$
- If $X_i = 1$
  - $Y_i = \beta_0 + \beta_1(1) + \epsilon_i$
  - $Y_i = \beta_0 + \beta_1 + \epsilon_i$

Speaker notes

When X is equal to either zero or one, the equation simplifies. For the “zero category”, Y is just equal to beta0 plus epsilon. For the “one category”, Y is equal to beta0 plus beta1 plus epsilon.

Categorical independent variables, 3

- Intperetation
  - b_0 is the estimated average value of Y when X equals the "zero category"
  - b_1 is the estimated average change in Y when X changes from the "zero category" to the "one category."

Speaker notes

The intepretation changes, but only slightly, when X is an indicator variable.

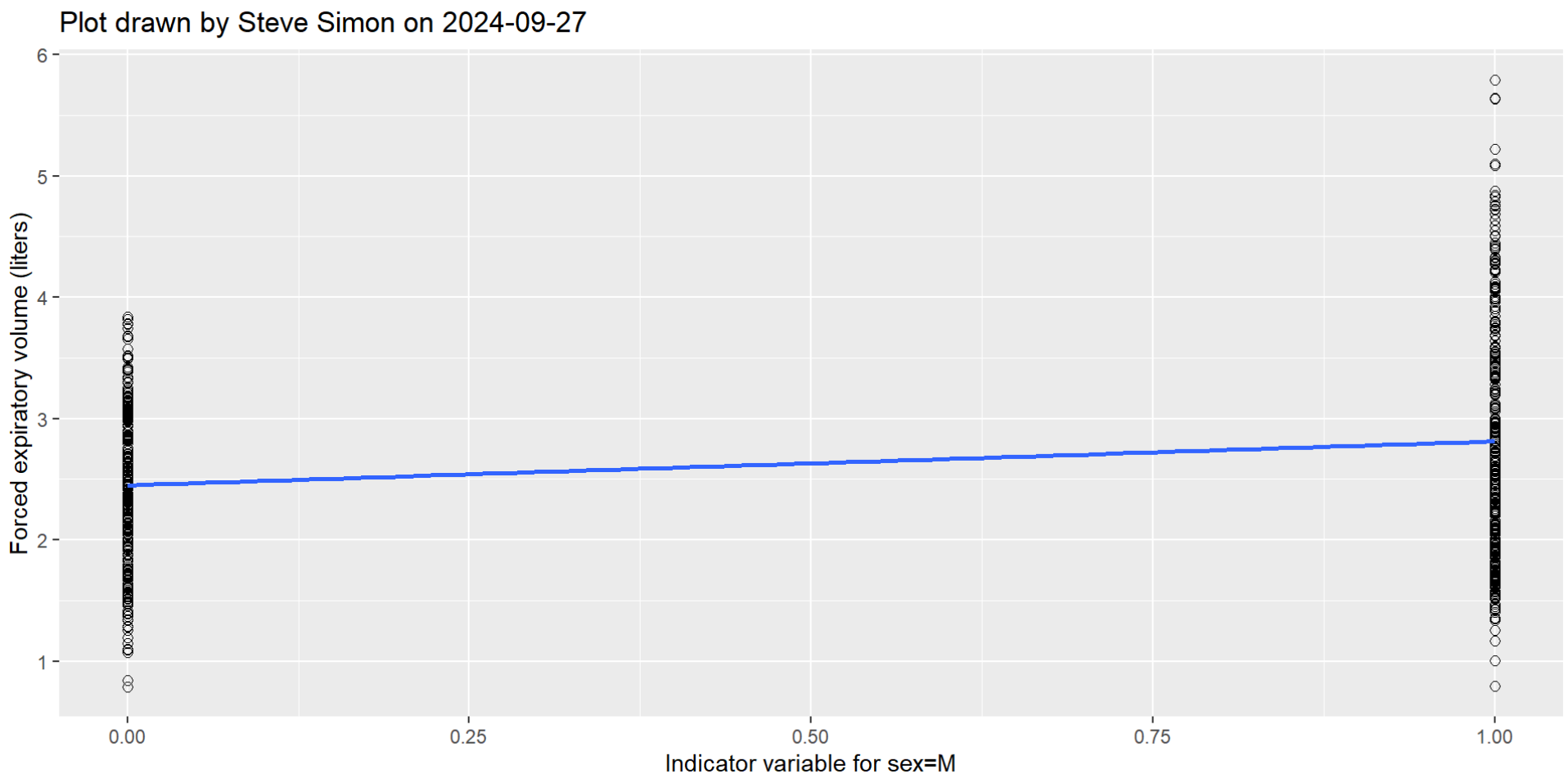
Creating an indicator variable

```
    fev sex sex_male
1 1.708   F        0
2 1.724   F        0
3 1.720   F        0
4 1.558   M        1
5 1.895   M        1
6 2.336   F        0
```

Speaker notes

Here is a small piece of the fev dataset with an indicator variable, sex_male, added.

Graphical display using the indicator variable



Speaker notes

It’s a bit hard to read this graph, but it looks like the line is around 2.4 when X equals zero. That would be the intercept. The line does show an increase . At X equals one, the line appears to be around 2.8. This is a 0.4 unit increase in Y for a one unit increase in X.

Linear regression using the indicator variable

```
Call:
lm(formula = fev ~ sex_male, data = fev_a)

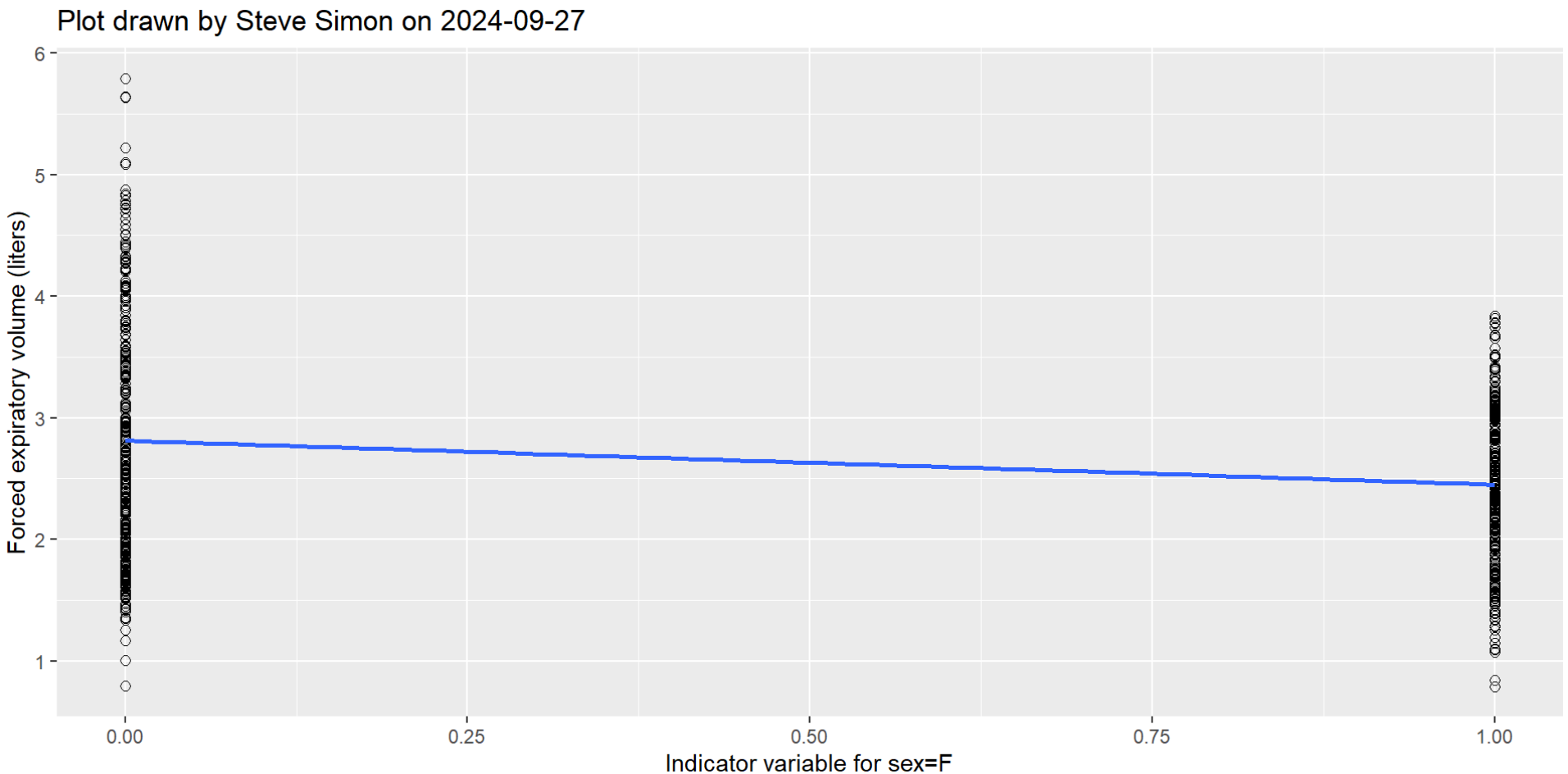
Coefficients:
(Intercept)     sex_male
     2.4512       0.3613
```

The estimated average fev value is 2.45 liters for females. The estimated average fev value is 0.36 liters larger for males.

Speaker notes

The intercept represents the estimated average value of Y when X equals zero. In this case, it represents the estimated average fev for female.s The slope represents the estimated average value of Y when X increases by one unit. In this case, it represents how much larger the estimated average fev is for males compared to females.

Graphical display using alternate indicator variable



Speaker notes

The choice of 1 for males was arbitrary, and you could have just as easily designated 1 as the female category. When you do, the graph flips. The intercept is a bit larger (2.8) and the slope is negative.

Linear regression using alternate indicator variable

```
Call:
lm(formula = fev ~ sex_female, data = fev_b)

Coefficients:
(Intercept)   sex_female
     2.8124      -0.3613
```

Letting your software create the indicator variable

- Different rules for different software
  - SPSS, SAS: first alphabetical category=1, second=0
  - R: second alphabetical category=1, first=0
- Always compare your output to the descriptive statistics

Speaker notes

You don’t have to create the indicator variable yourself. Most statistical software will do it for you. Just be careful because the software has to make an arbitrary choice. SPSS and SAS choose the first category that appears when you put the data in alphabetical order. So they would choose females as 1 because the “F” code for sex appears alphabetically before the “M” code for sex. R does the opposite. If you ask R to create indicator variables automatically, it codes males as 1.

It is easy to get confused about this, so you should always orient yourself by looking at the graphs and simple descriptive statistics before trying to interpret the output from a linear regression model.

Break #1

- What you have learned
  - Categorical independent variables
- What’s coming next
  - R code for categorical independent variables

fev data dictionary, 1

```
data_dictionary: fev (.csv, sas7bdat, .sav, .txt)

copyright: |
  The author of the jse article holds the copyright, but does not list
conditions under which it can be used. Individual use for educational purposes
is probably permitted under the Fair Use provisions of U.S. Copyright laws.

description: |
  Forced Expiratory Volume (FEV) in children. The data was collected  in
Boston in the 1970s.

additional_description: https://jse.amstat.org/v13n2/datasets.kahn.html
```

fev data dictionary, 2

```
download_url: https://www.amstat.org/publications/jse/datasets/fev.dat.txt
format:
  csv: comma delimited
  sas7bdat: proprietary (SAS)
  sav: proprietary (SPSS)
  txt: fixed width
varnames: not included
missing_value_code: not needed
size:
  rows: 654
  columns: 5
```

fev data dictionary, 3

```
age:
  scale: ratio
  range: positive integer
  unit: years
fev:
  label: Forced Expiratory Volume
  scale: ratio
  range: positive real
  unit: liters
ht:
  label: Height
  scale: positive real
  unit: inches
```

fev data dictionary, 4

```
sex:
  value:
    F: Female
    M: Male

smoke:
  value:
    'N': Nonsmoker
    'Y': Smoker
```

simon-5501-07-fev.qmd, 1

```
---
title: "Linear regression modules using the fev dataset"
author: "Steve Simon"
format:
  html:
    embed-resources: true
date: 2024-09-25
---

There is a [data dictionary][dd] that provides more details about the data.
The program was written by Steve Simon on 2024-09-02 and is placed in the
public domain.

[dd]: https://github.com/pmean/datasets/blob/master/fev.yaml
```

Speaker notes

The first few lines are the documentation header

simon-5501-07-fev.qmd, 2

````
## Libraries

You should always load the tidyverse library. The broom library provides the glance, tidy, and augment functions that help you with computations of linear regression models. The car library provides the vif function for measuing collinearity.

```{r setup}
#| message: false
#| warning: false
library(broom)
library(car)
library(tidyverse)
```
````

simon-5501-07-fev.qmd, 3

````
## List variable names

Since the variable names are not listed in the data file itself, you need to
list them here.

```{r names}
pulmonary_names <- c(
    "age",
    "fev",
    "height",
    "sex",
    "smoke")
```
````

simon-5501-07-fev.qmd, 4

````
## Reading the data

Here is the code to read the data and show a glimpse.

```{r read}
pulmonary <- read_csv(
  file="../data/fev.csv",
  col_names=pulmonary_names,
  col_types="nnncc")
glimpse(pulmonary)
```
````

simon-5501-07-fev.qmd, 5

```
## m1: Linear regression model using sex to predict fev

Is there a relationship between sex and fev? Do males tend to have larger fev
values than females. This section (labelled m1) shows some simple descriptive
and graphical summaries followed by a linear regression model.
```

simon-5501-07-fev.qmd, 6

````
## m1: Descriptive stastics for sex

```{r sex}
pulmonary |>
    count(sex) |>
    mutate(total=sum(n)) |>
    mutate(pct=100*n/total)
```

There are slightly more males (51%) than females in this sample.
````

simon-5501-07-fev.qmd, 7

````
## m1: Descriptive statistics for fev

```{r fev}
pulmonary |>
  summarize(
    fev_mn=mean(fev),
    fev_sd=sd(fev),
    fev_min=min(fev),
    fev_max=max(fev))
```
````

The average fev value, 2.6, seems reasonable. The standard deviation, 0.87, indicates a fair amount of variation. The minimum and maximum values both appear to be reasonable.

simon-5501-07-fev.qmd, 8

```
## m1: Tabular summary of the relationship between sex and fev, 1

```{r sex-and-fev-1}
pulmonary |>
  group_by(sex) |>
    summarize(
        fev_mn=mean(fev),
        fev_sd=sd(fev))
```

The average fev is 0.36 liters higher for males than females. There is very slightly more variation in the males (the standard deviation is 1.00 versus 0.65).
```

simon-5501-07-fev.qmd, 9

````
## m1: Graphical summary of the relationship between sex and fev, 2

```{r sex-and-fev-2}
pulmonary |>
  ggplot(aes(sex, fev)) +
    geom_boxplot() +
    coord_flip() +
    ggtitle("Graph drawn by Steve Simon on 2024-09-26") +
      xlab("Sex") +
      ylab("Forced Expiratory Volume (liters)")
```

The boxplot shows the same pattern slightly larger fev values for males
compared to females and slightly more variation.
````

simon-5501-07-fev.qmd, 10

````
## m1: Fit the linear regression model

```{r m1-model}
m1 <- lm(fev ~ sex, data=pulmonary)
m1
```

The estimated average fev is 2.45 for females and 0.36 inches higher for
males.
````

simon-5501-07-fev.qmd, 11

```
## m1: Analysis of variance table

```{r m1-anova}
anova(m1)
```

The F-statistic, 29.6, is large, and the p-value is very small (< 0.001).
Reject the null hypothesis and conclude that the average fev values are
different between males and females.
```

simon-5501-07-fev.qmd, 12

```
## m1: R-squared

```{r m1-r-squared}
glance(m1)$r.squared
```

Sex is a very weak predictor of fev. Only 4% of the variation in fev values
can be accounted for by a patient's sex.
```

simon-5501-07-fev.qmd, 13

````
## m1: Confidence intervals

```{r m1-ci}
confint(m1)
```
````

We are 95% confident that the difference in fev values is between 0.23 and 0.49. This is a positive difference for all values in the confidence interval, demonstrating that the average fev values are larger for males than for females. This interval is narrow indicating that there is very little sampling error. Hardly a surprise with such a large dataset (n=654).

simon-5501-07-fev.qmd, 14

````
## m1: T-tests

```{r m1-t-tests}
tidy(m1)
```

The t-statistic, 5.4, is not close to zero. Conclude that there is a difference in average fev values between males and females. Since the t-statistic is positive, conclude also that the average fev value is larger for males.
````

simon-5501-07-fev.qmd, 15

```
## m1: Normal probability plot of residuals, 1

Note: I learn something new everyday. You do not have to use qqnorm to create
a normal probability plot. You can do it using the ggplot and stat_qq
functions. This looks nicer, is consistent with other visualizations in R, and
allows you to put in a title using ggtitle. In your homework, you are welcome
to use this approach (ggplot and stat_qq) or you can use qqnorm.
```

simon-5501-07-fev.qmd, 16

````
## m1: Normal probability plot of residuals, 2

```{r m1-qq-plot}
r1 <- augment(m1)
r1 |>
    ggplot(aes(sample=.resid)) +
      stat_qq() +
    ggtitle("Graph drawn by Steve Simon on 2024-09-30")
```

The straight lines indicates that the residuals are normally distributed.
````

simon-5501-07-fev.qmd, 17

````
## m1: Histogram of residuals

```{r m1-histogram}
r1 |>
    ggplot(aes(.resid)) +
      geom_histogram(
        binwidth=0.2,
        color="black",
        fill="white") +
    ggtitle("Graph drawn by Steve Simon on 2024-09-26") +
      xlab("Residuals from m1 regression model")
```

The histogram also indicates that the residuals are normally distributed.
````

simon-5501-07-fev.qmd, 18

m1: Influential data points

Both leverage and Cook's distance make little sense for a regression model using a categorical independent variable. The studentized deleted residual is still useful.

```{r m1-studentized-deleted-residual}
r1 |>
    filter(abs(.std.resid) > 3)
```

There are three outliers on the high end, corresponding to fev values of 5.6, 5.6, and 5.8 liters in males. There are no outliers among the female patients.

simon-5501-07-fev.qmd, 19

```
## m2: Linear regression model using smoke to predict fev

Is there a relationship between smoke and fev? This section (labeled m2) shows
some simple descriptive and graphical summaries followed by a linear
regression model.
```

simon-5501-07-fev.qmd, 20

````
## m2: Descriptive statistics for smoke

```{r smoke}
pulmonary |>
    count(smoke) |>
    mutate(total=sum(n)) |>
    mutate(pct=100*n/total)
```

There are very few smokers (65 or 10%) in this sample. Descriptive statistics
for fev were shown earlier.
````

simon-5501-07-fev.qmd, 21

```
## m2: Relationship between smoke and fev, 1

```{r smoke-and-fev-1}
pulmonary |>
  group_by(smoke) |>
    summarize(
        fev_mn=mean(fev),
        fev_sd=sd(fev))
```
```

Smokers have an average fev value that is 0.7 units higher than non-smokers. The standard deviations (0.85 and 0.75) demonstrate roughly the same amount of variation in the two groups.

simon-5501-07-fev.qmd, 22

````
## m2: Relationship between smoke and fev, 2

```{r smoke-and-fev-2}
pulmonary |>
  ggplot(aes(smoke, fev)) +
    geom_boxplot() +
    coord_flip() +
    ggtitle("Graph drawn by Steve Simon on 2024-09-26") +
      xlab("Did the patient smoke?") +
      ylab("Forced Expiratory Volume (liters)")
```

The boxplot shows the same pattern as noted above.
````

simon-5501-07-fev.qmd, 23

````
## m2: Fit the linear regression model

```{r m2-model}
m2 <- lm(fev ~ smoke, data=pulmonary)
m2
```
````

The estimated average fev is 2.57 liters for non-smokers and 0.71 liters higher on average for smokers.

Normally, you would follow this up with various functions (anova, confint, tidy), assess various diagnostic plots using the residuals, and identify influential data points.

Break #2

- What you have learned
  - R code for categorical independent variables
- What's coming next
  - Multiple linear regression

Model

- $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i,\ i = 1, \ldots, N$
- Least squares estimates: $b_0,\ b_1,\ b_2$

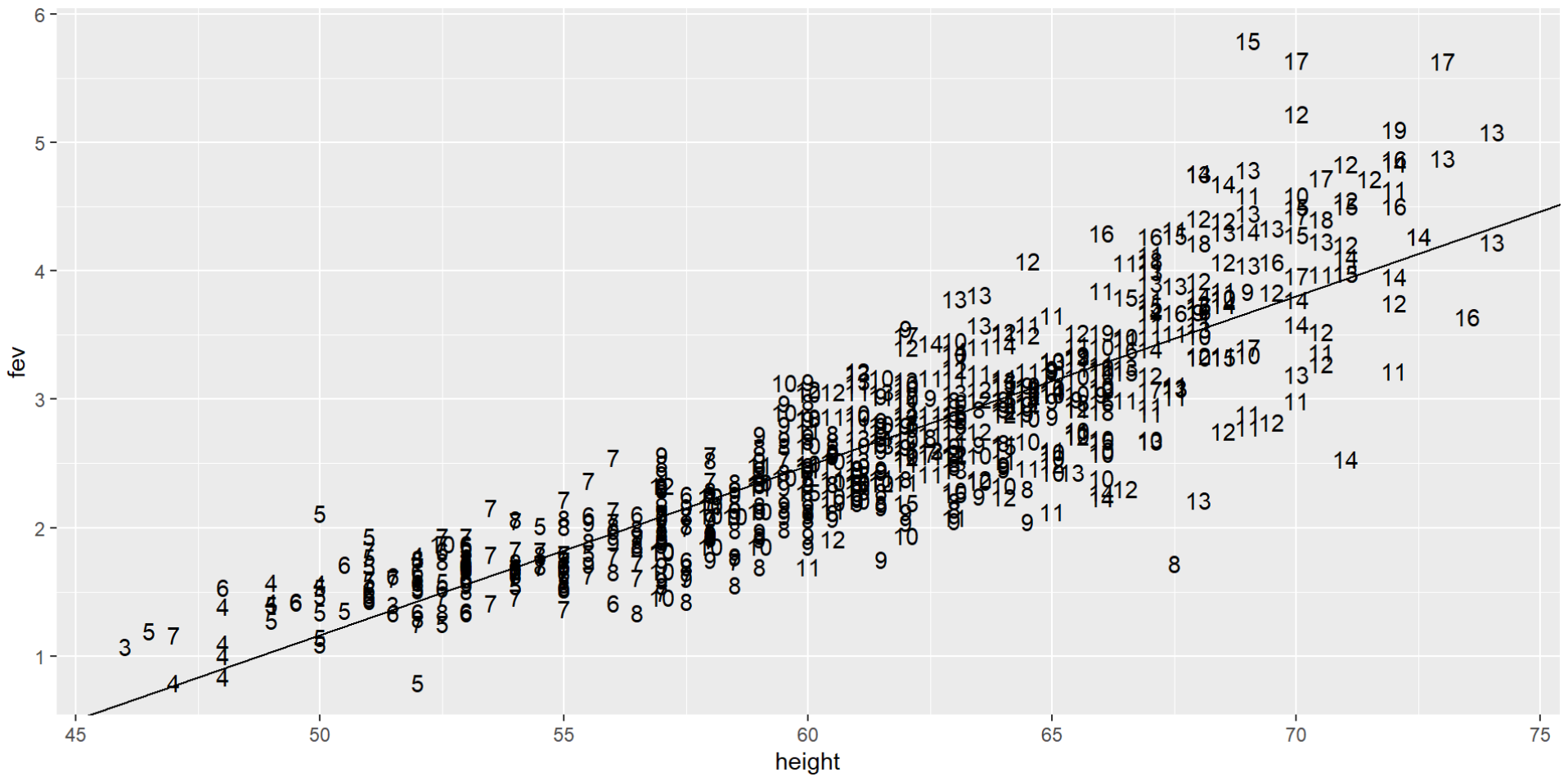
Speaker notes

Add note.

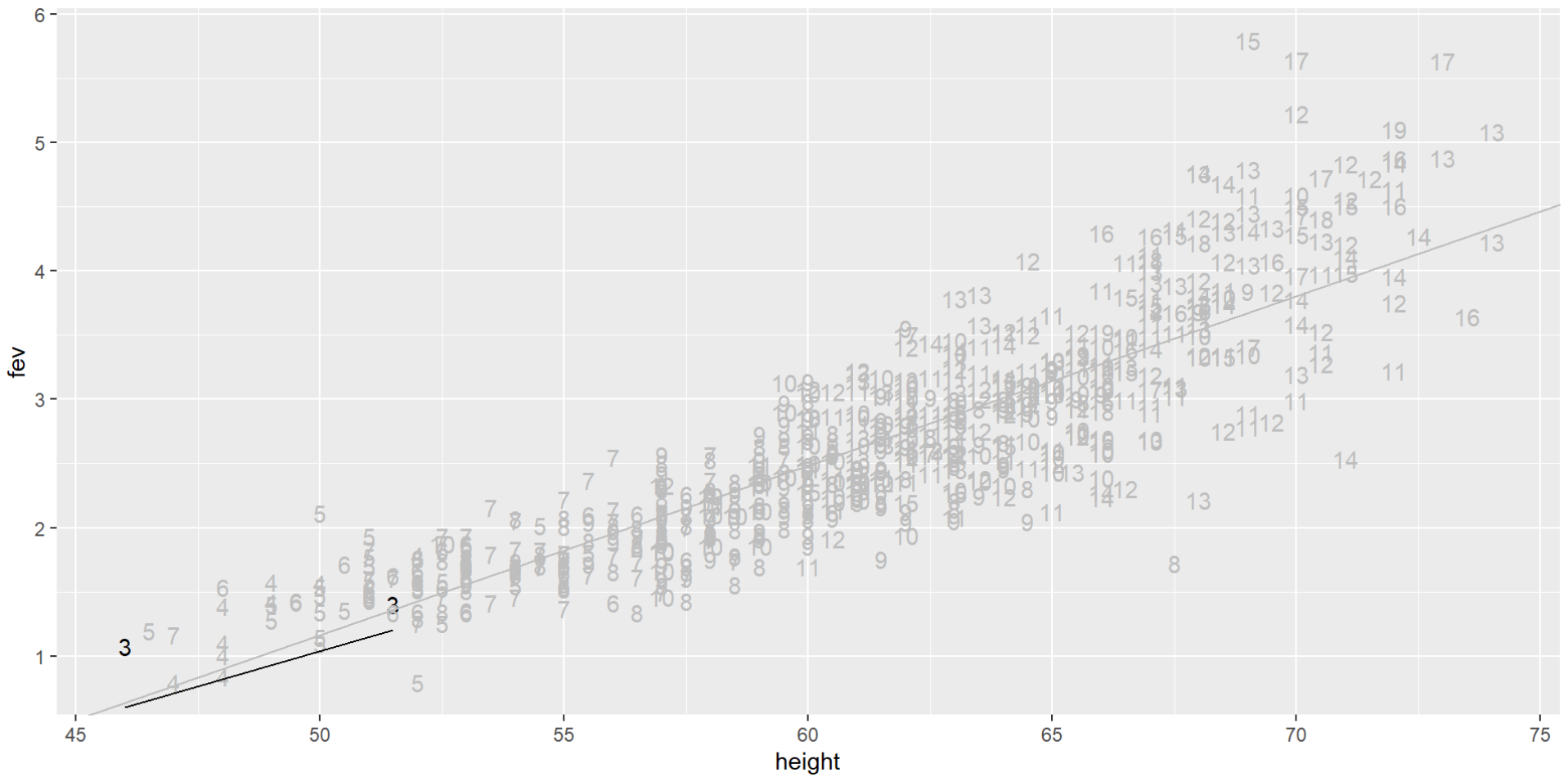
Interpretations

- b_0 is the estimated average value of Y when X1 and X2 both equal zero.
- b_1 is the estimated average change in Y
  - when X_1 increases by one unit, and
  - X_2 is held constant
- b_2 is the estimated average change in Y
  - when X_2 increases by one unit, and
  - X_1 is held constant

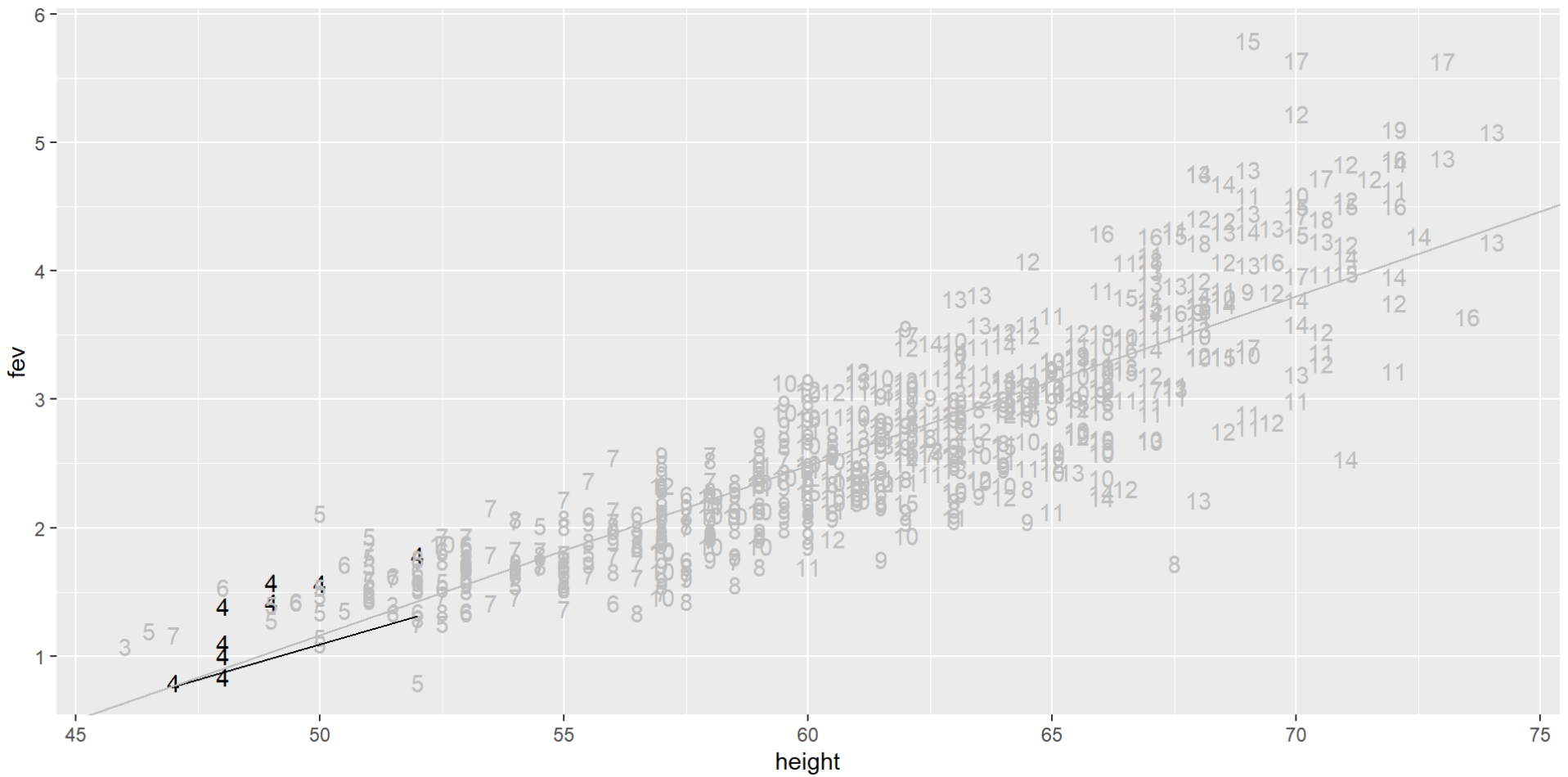
Unadjusted relationship between height and FEV



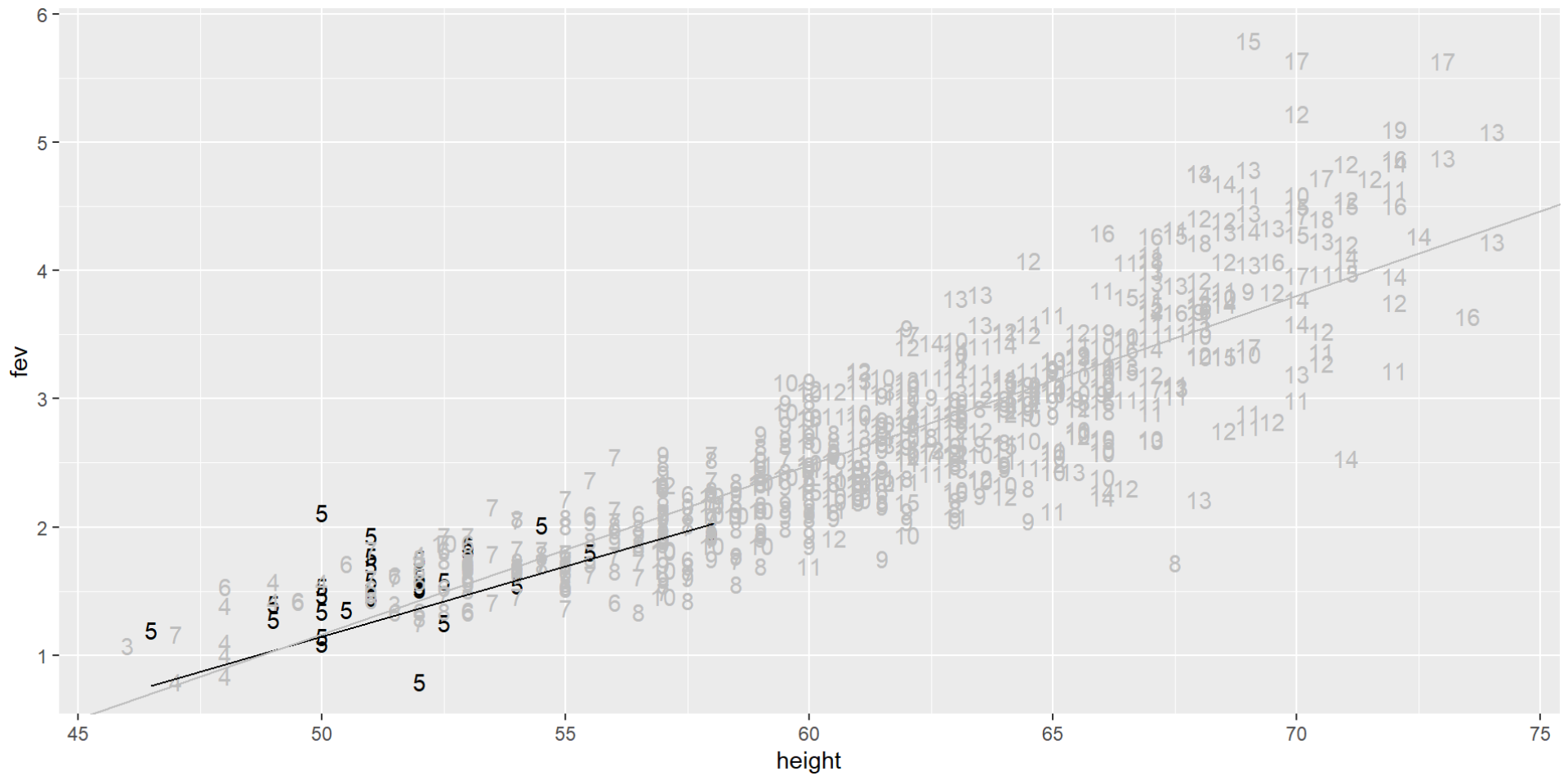
Relationship between height and FEV controlling at Age=3



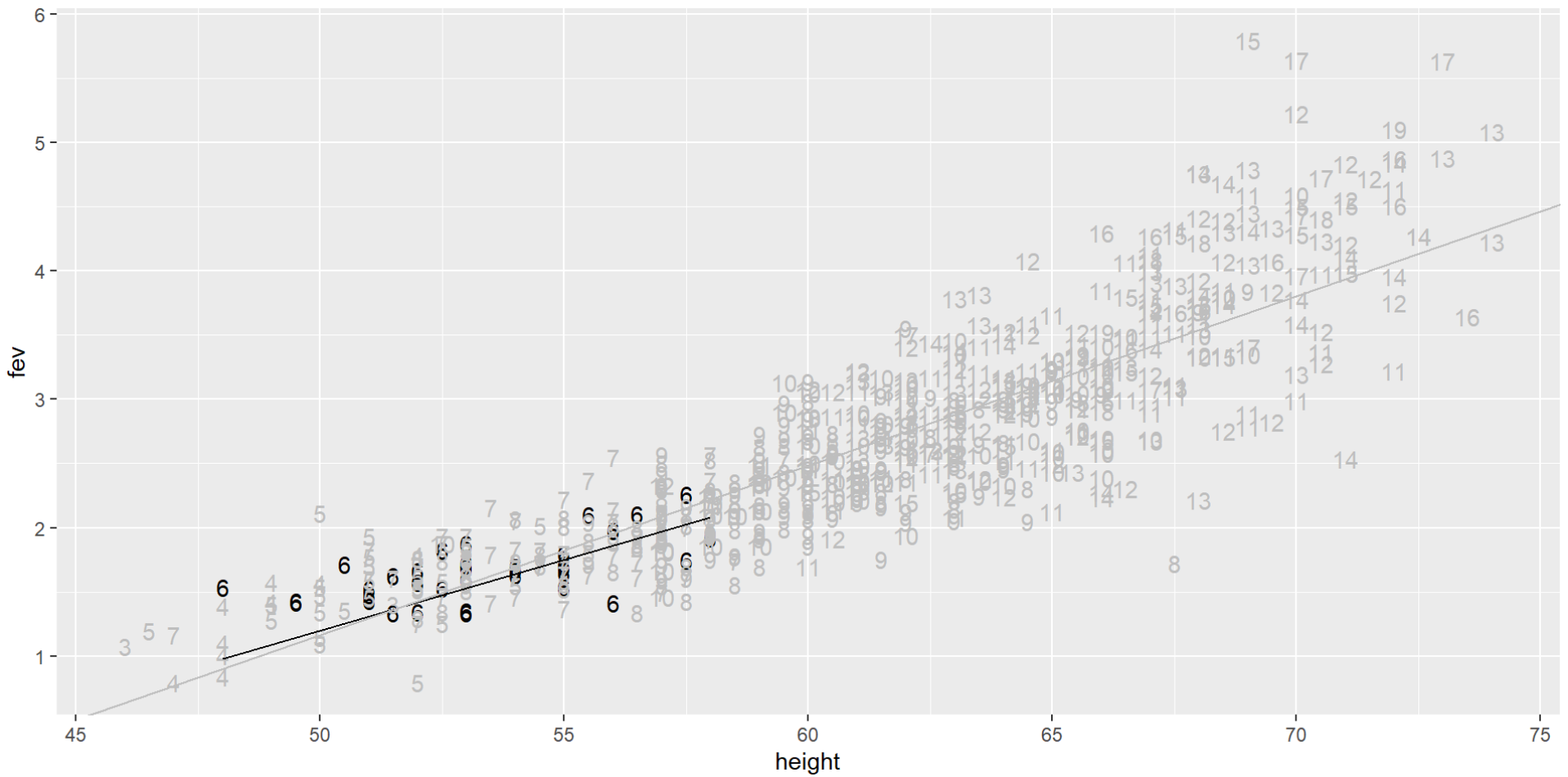
Relationship between height and FEV controlling at Age=4



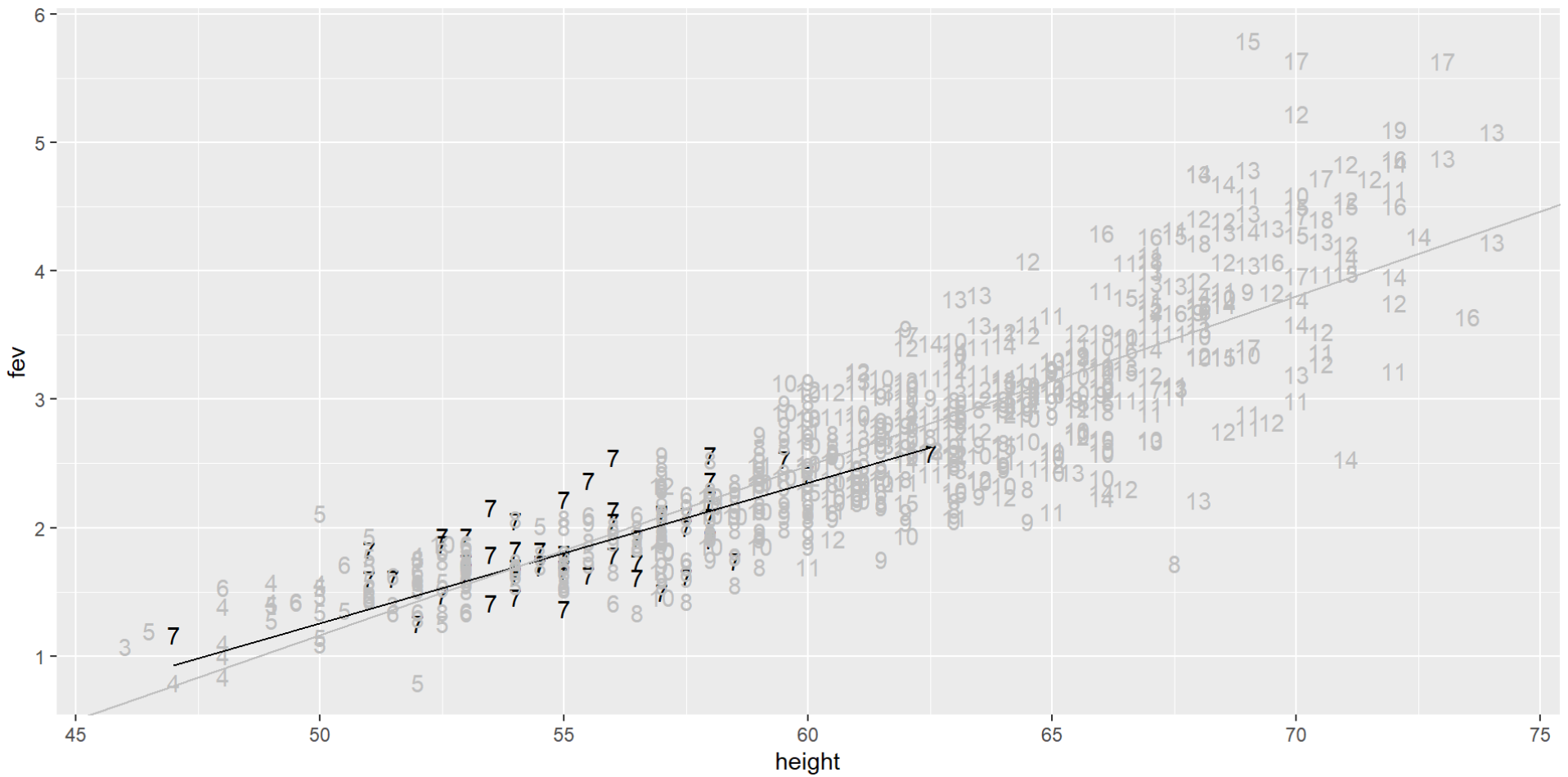
Relationship between height and FEV controlling at Age=5



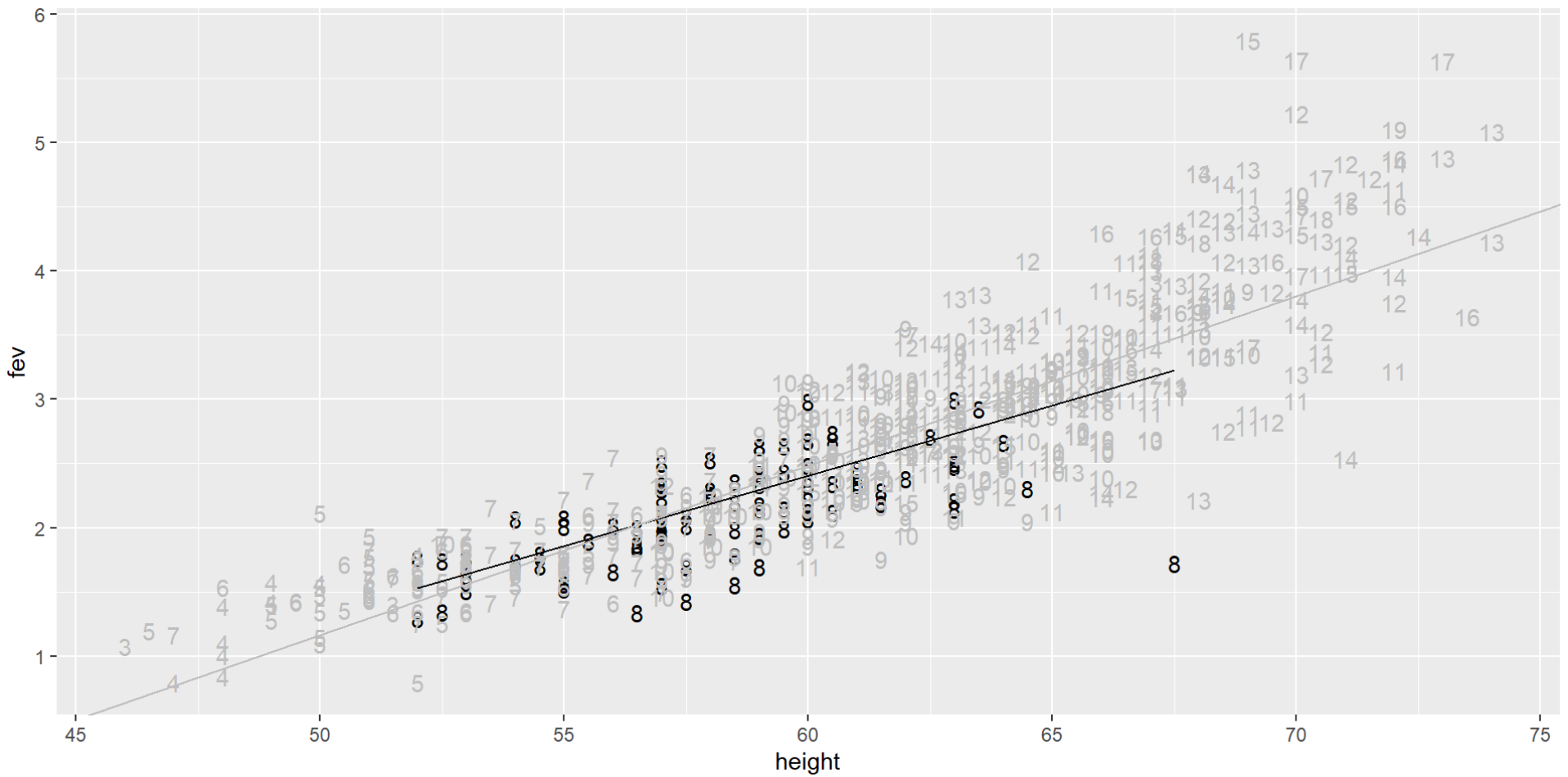
Relationship between height and FEV controlling at Age=6



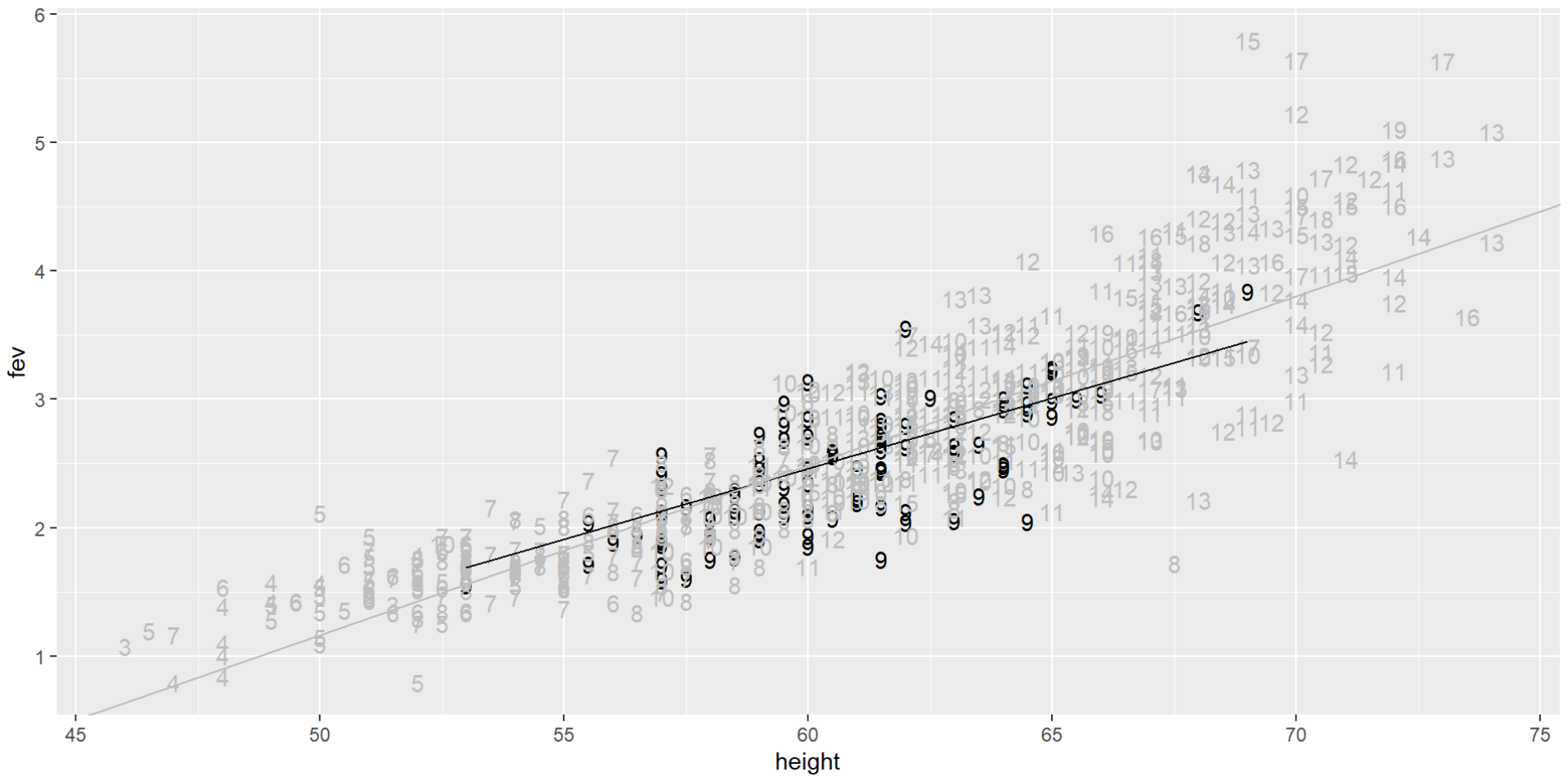
Relationship between height and FEV controlling at Age=7



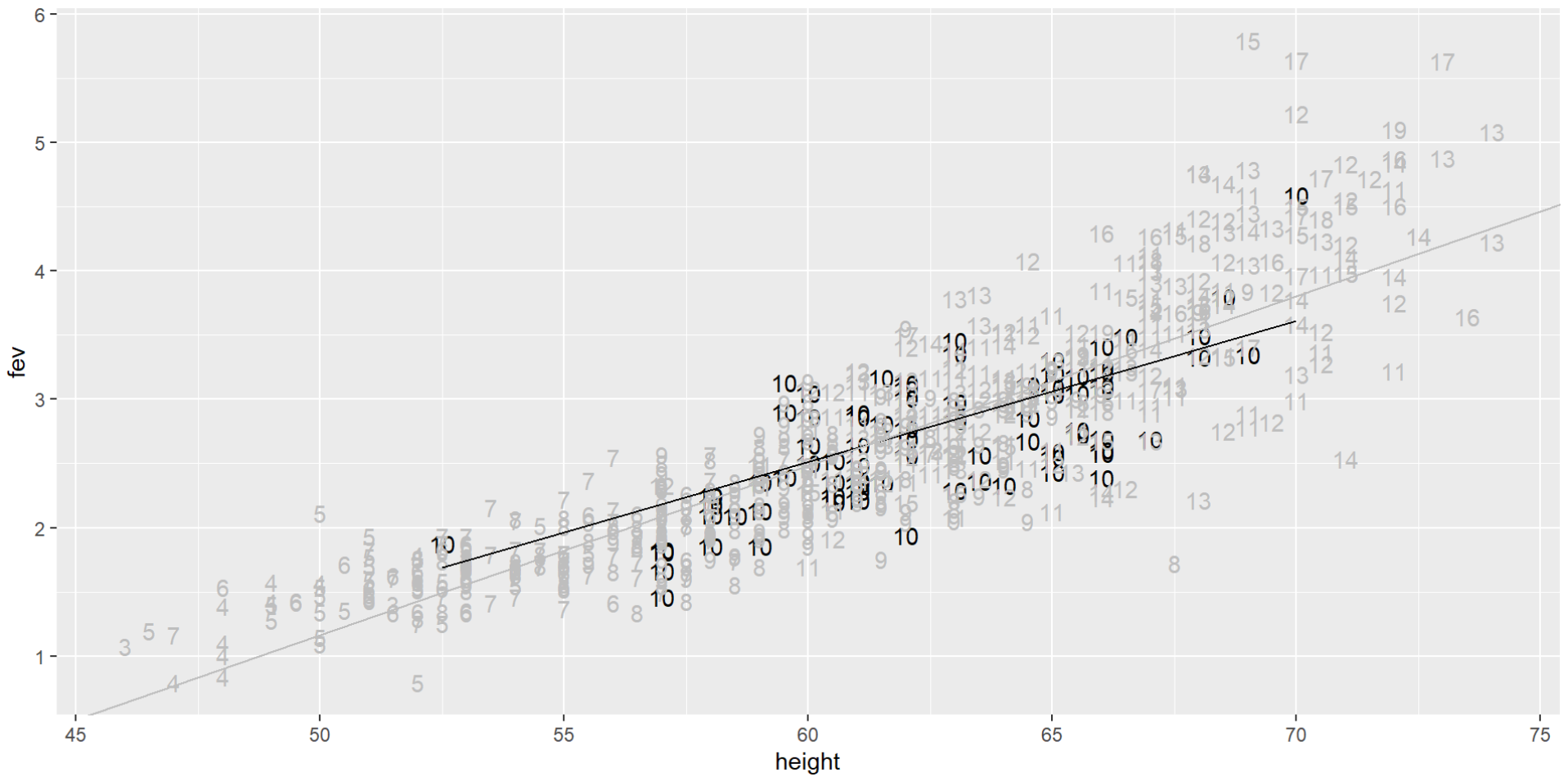
Relationship between height and FEV controlling at Age=8



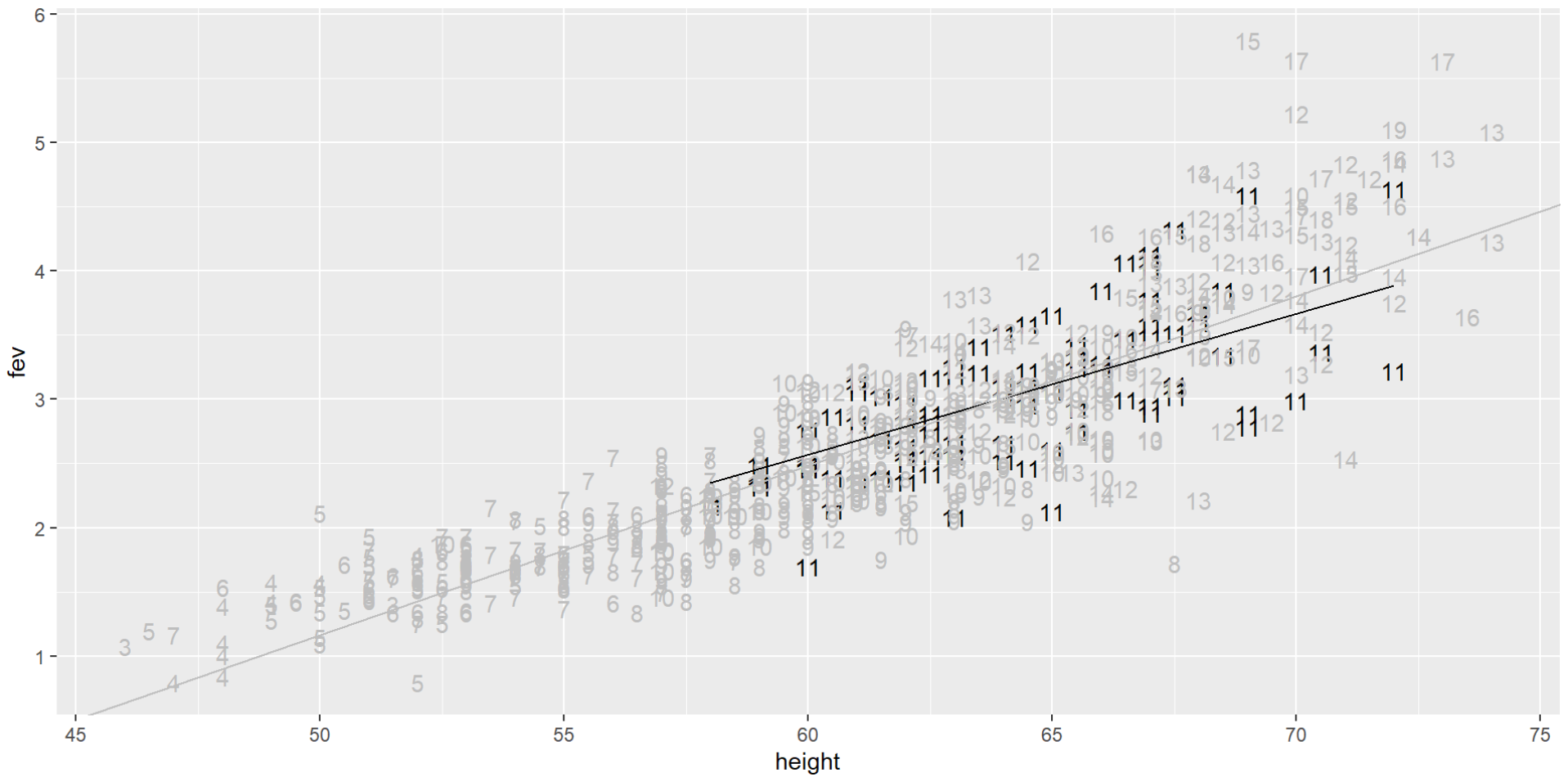
Relationship between height and FEV controlling at Age=9



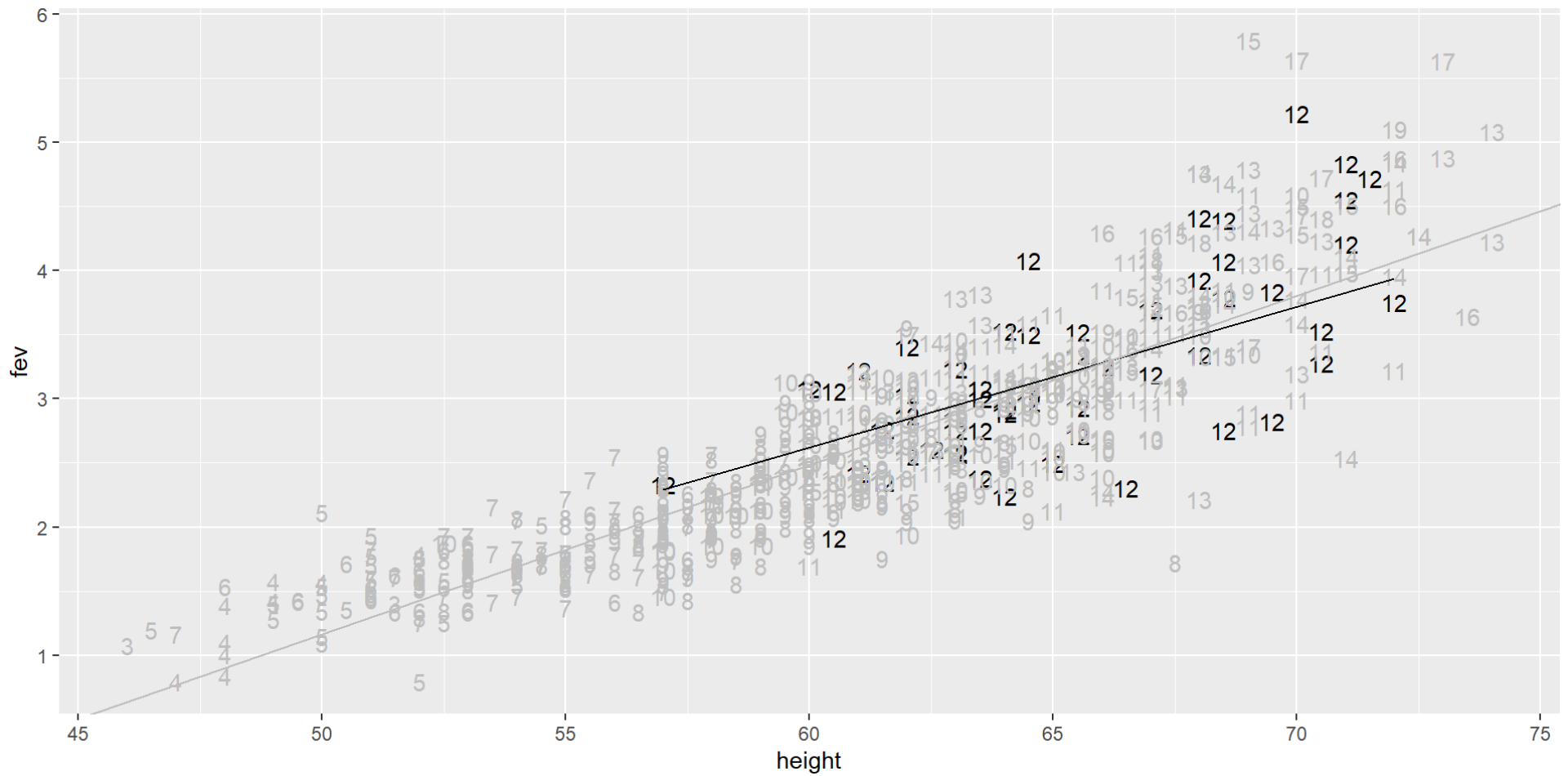
Relationship between height and FEV controlling at Age=10



Relationship between height and FEV controlling at Age=11

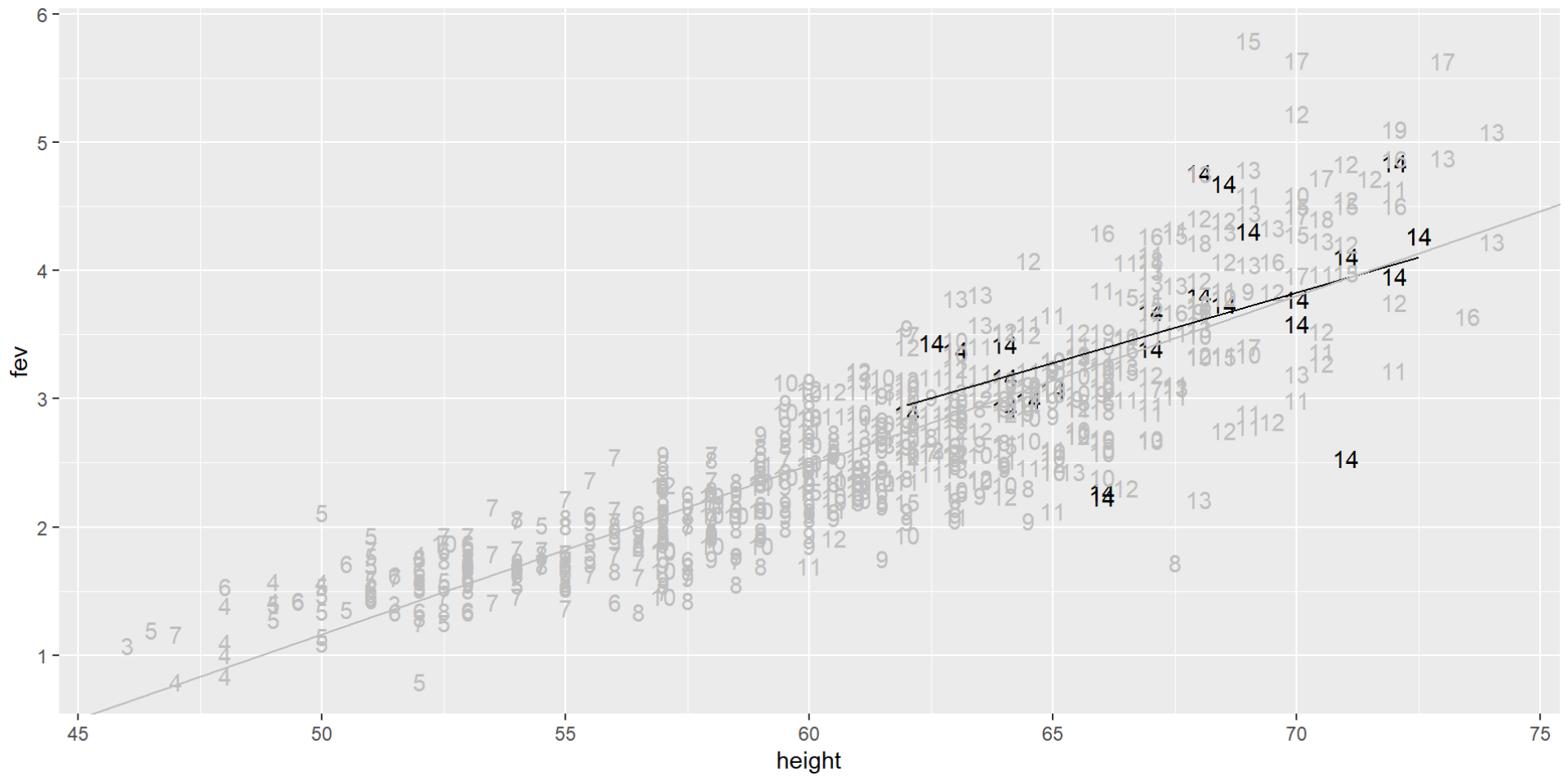


Relationship between height and FEV controlling at Age=12

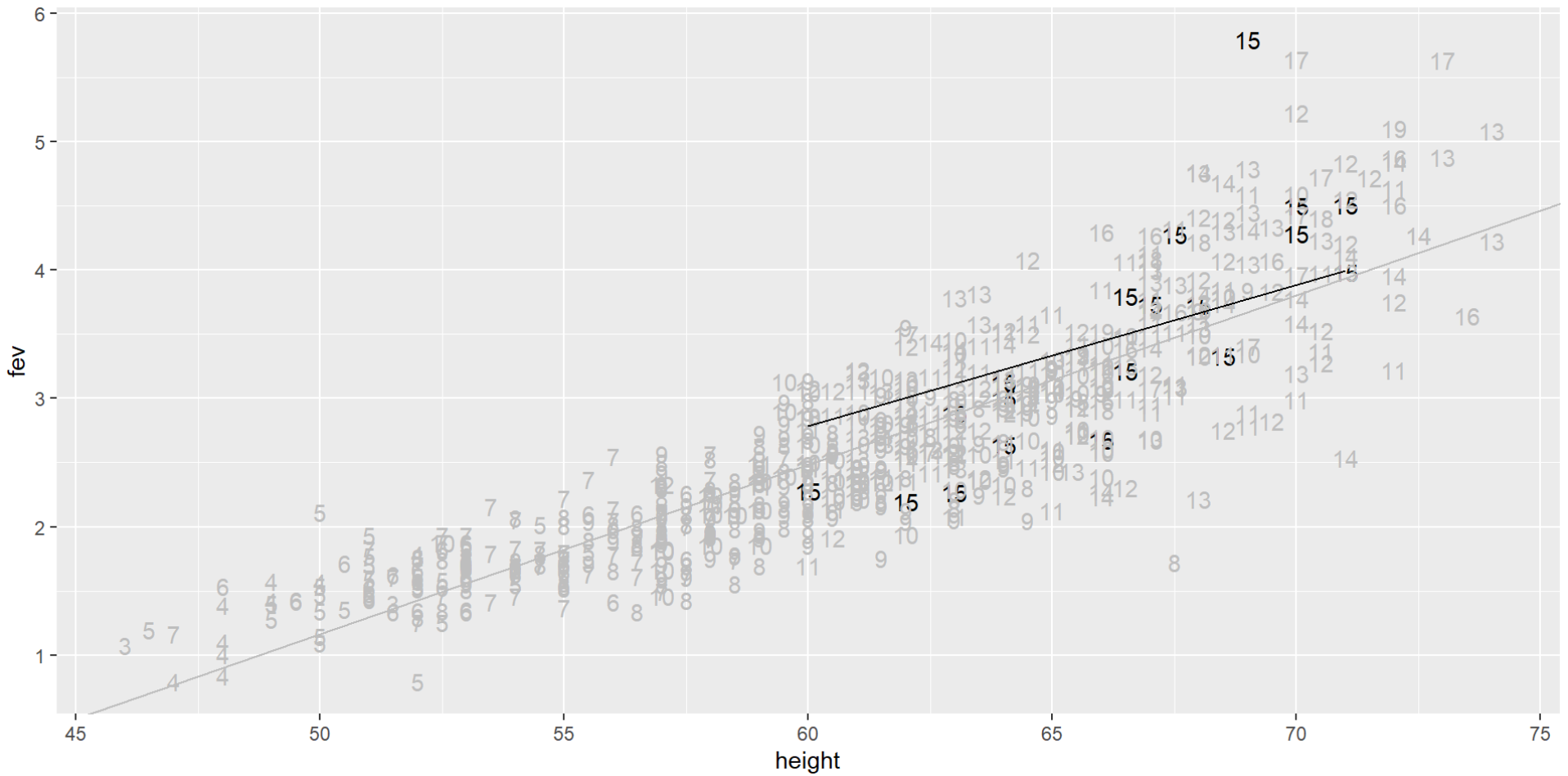


Relationship between height and FEV controlling at Age=13

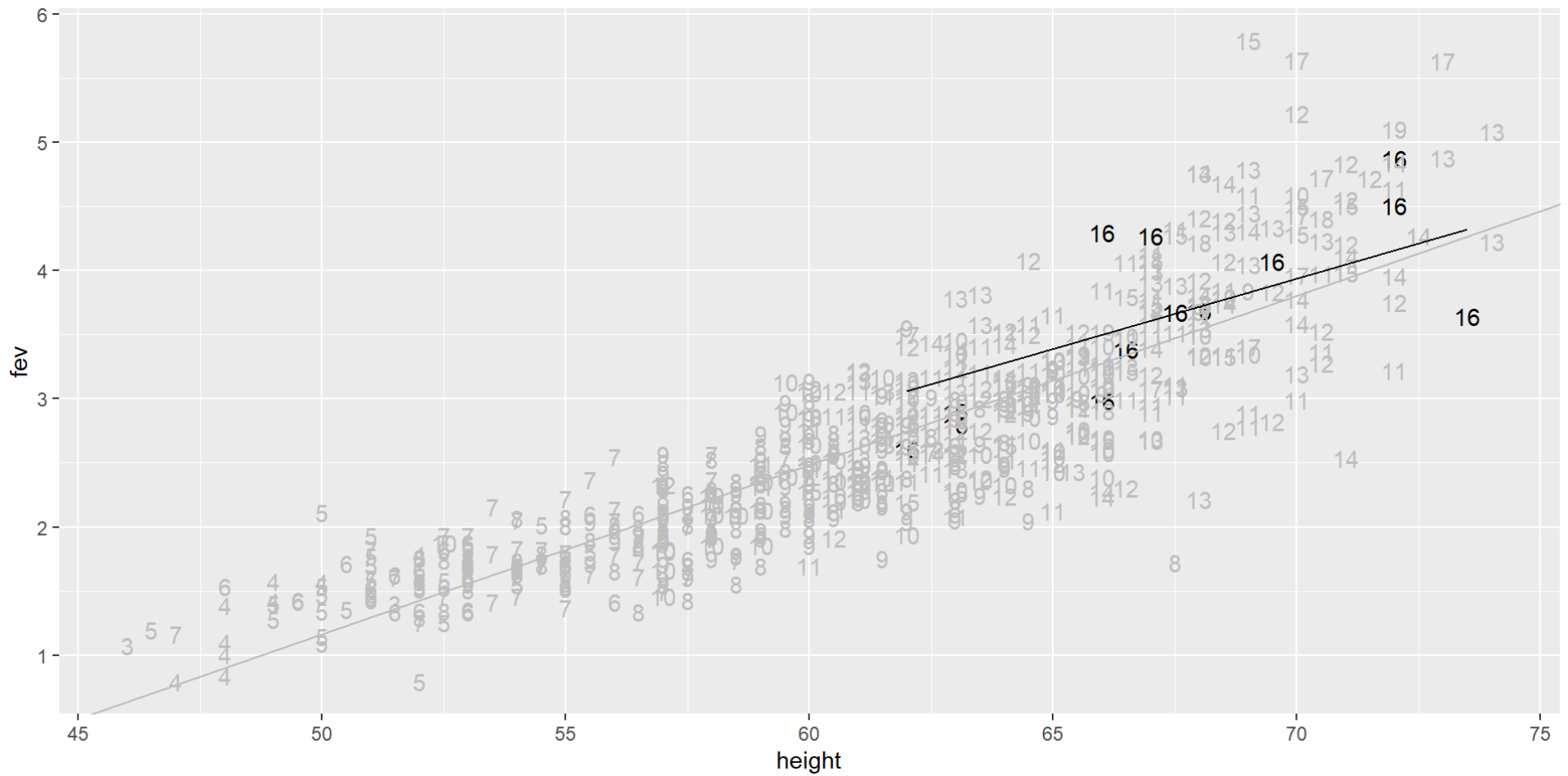
Relationship between height and FEV controlling at Age=14



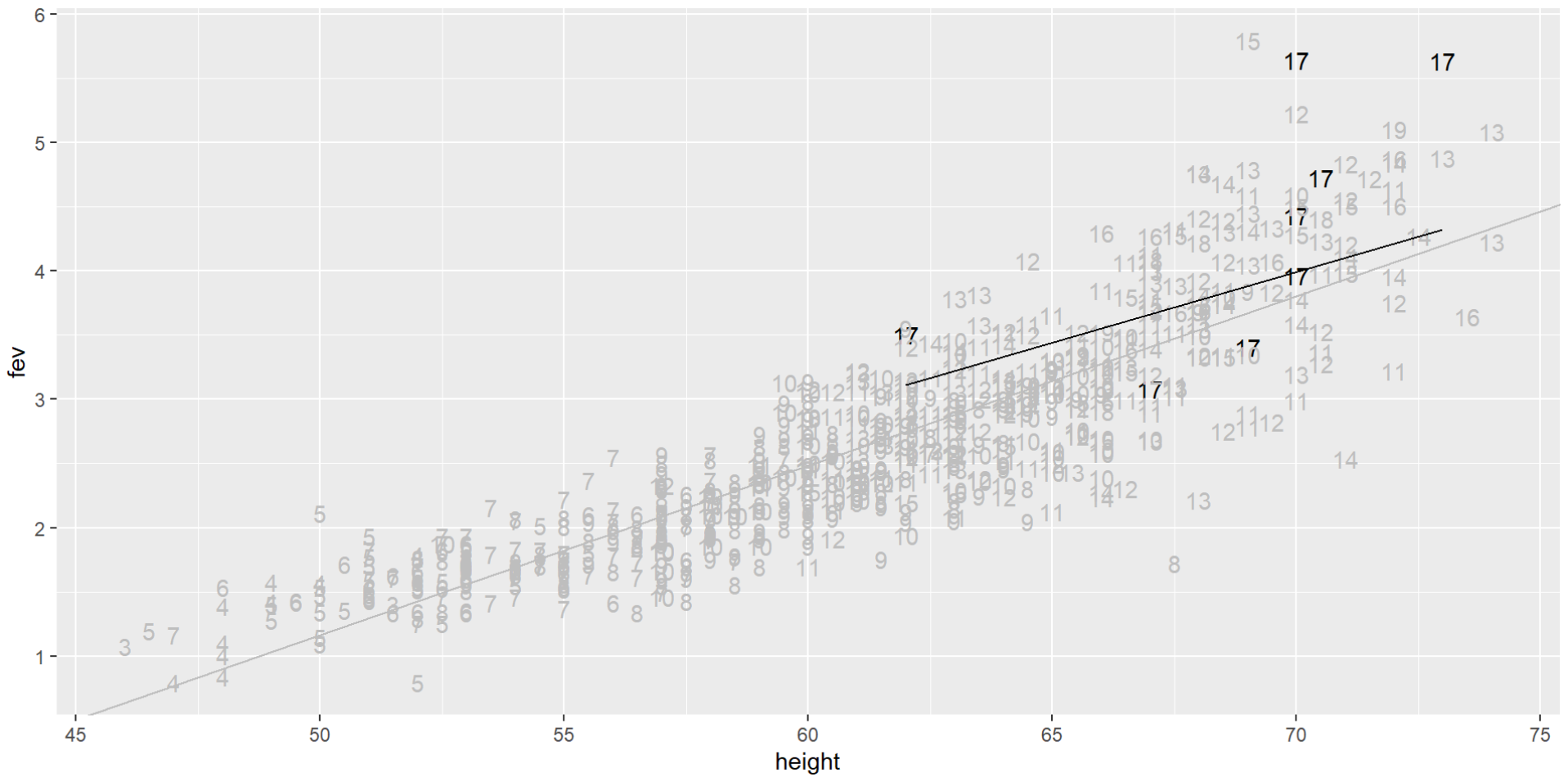
Relationship between height and FEV controlling at Age=15



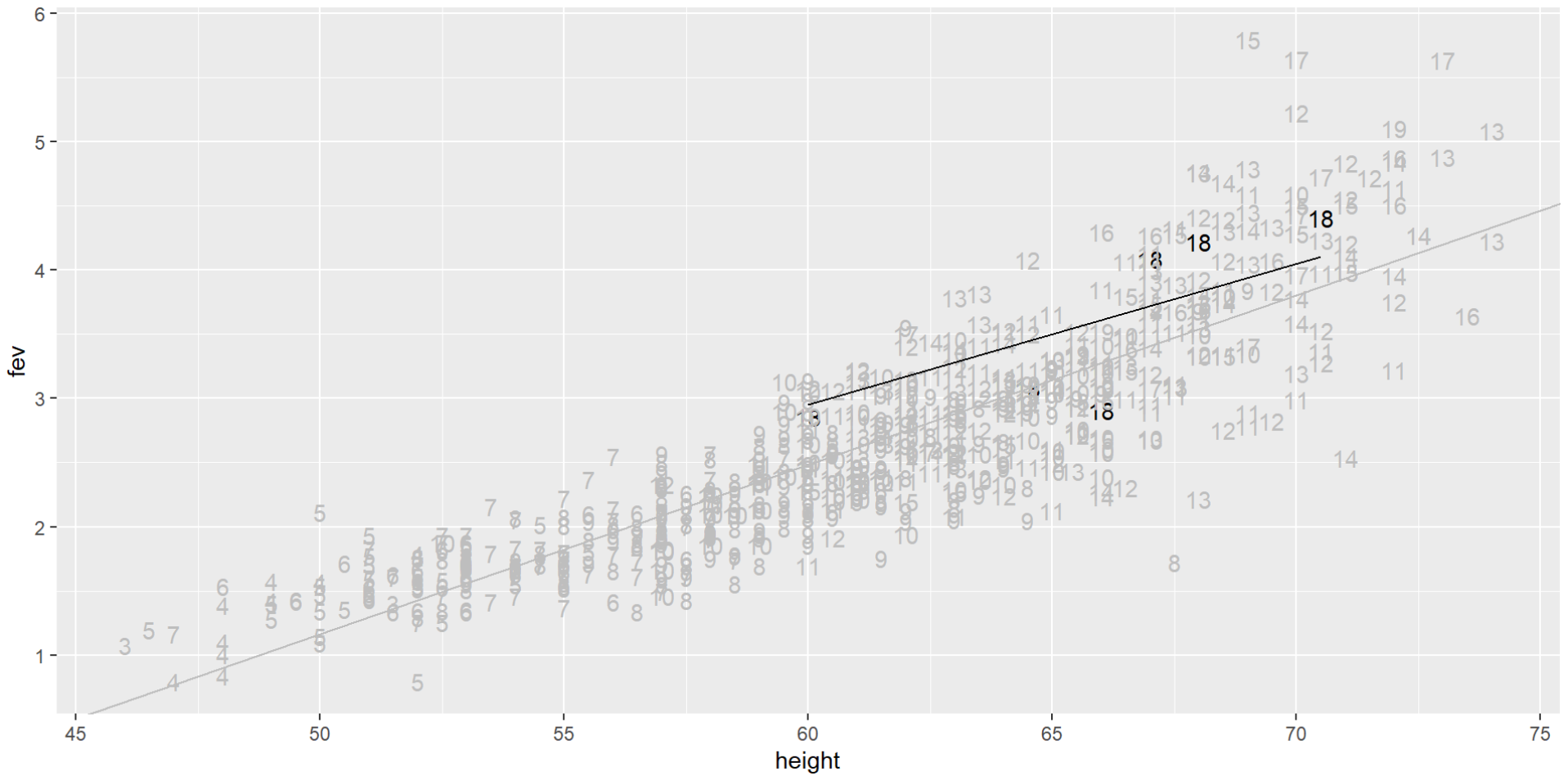
Relationship between height and FEV controlling at Age=16



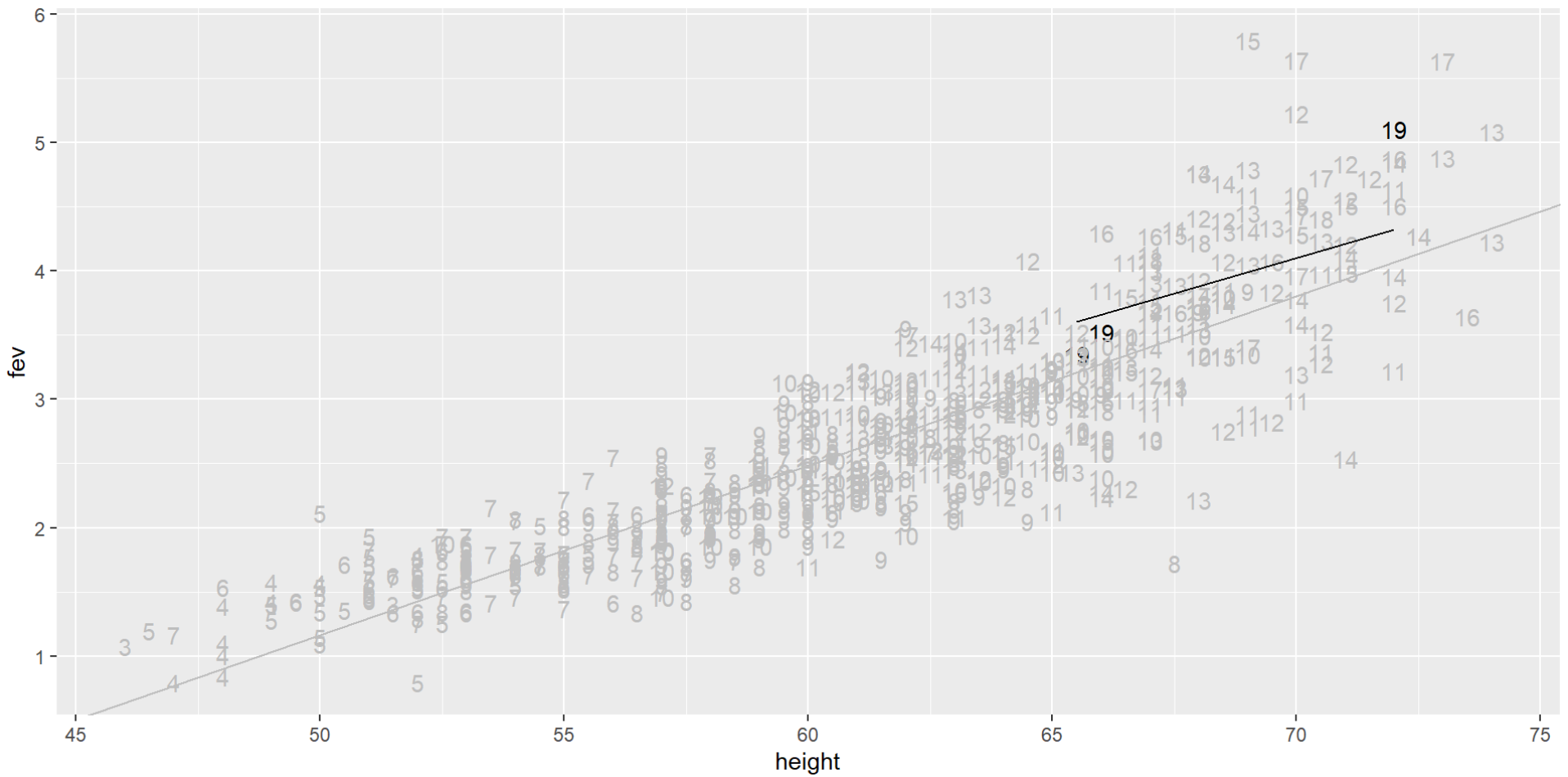
Relationship between height and FEV controlling at Age=17



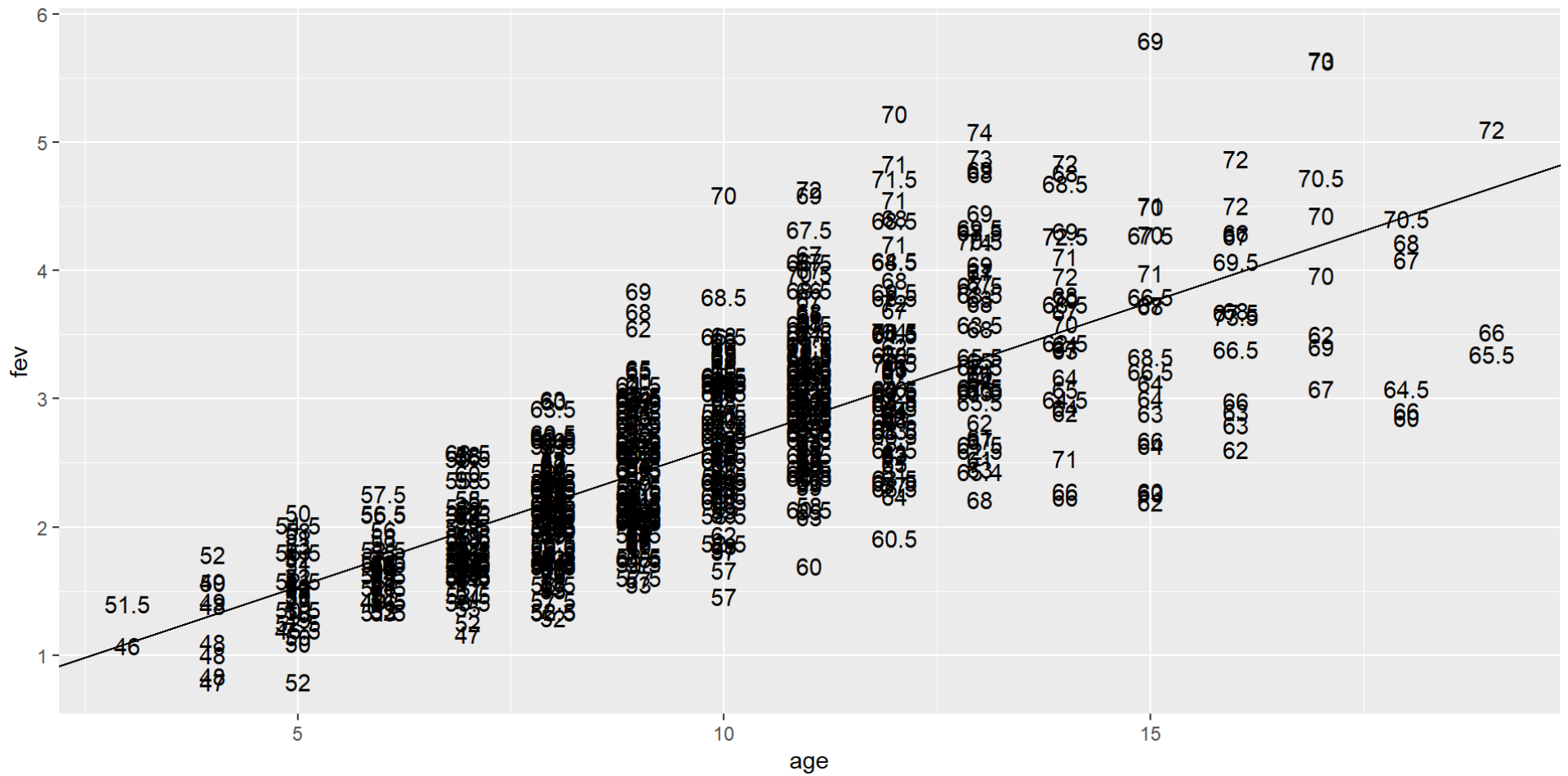
Relationship between height and FEV controlling at Age=18



Relationship between height and FEV controlling at Age=19

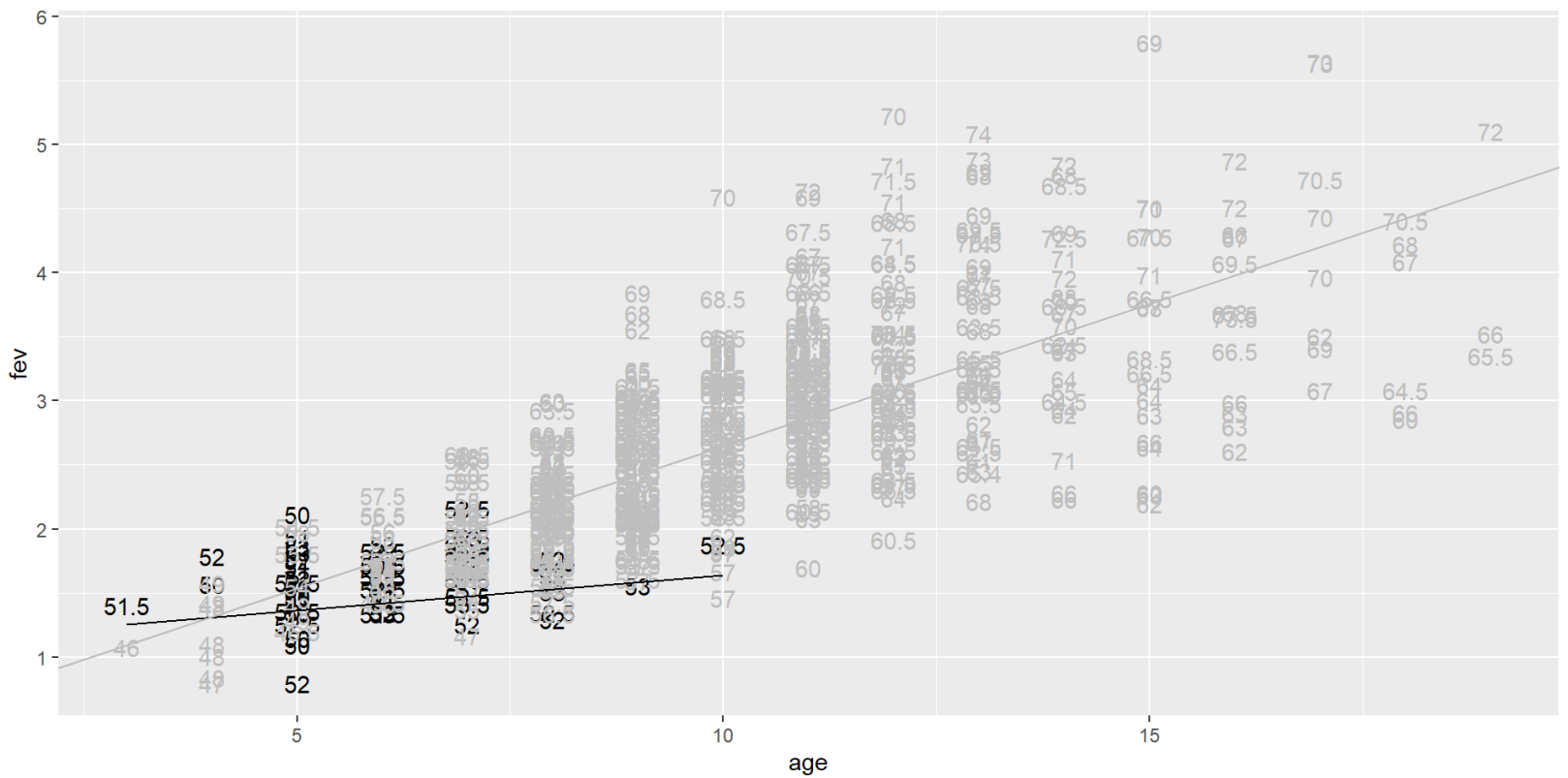


Unadjusted relationship between age and fev

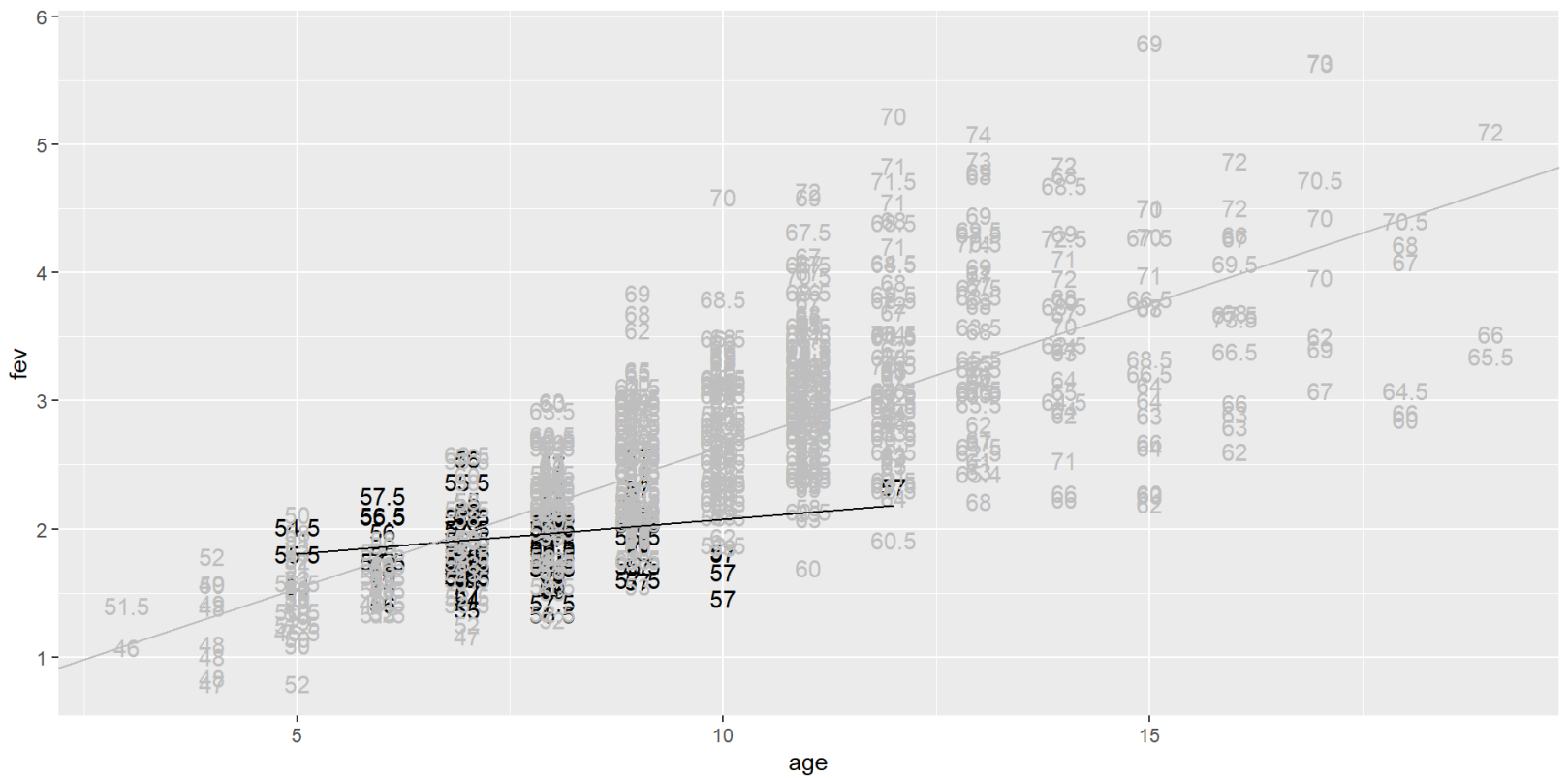


Relationship between age and FEV controlling for height between 46 and 49.5

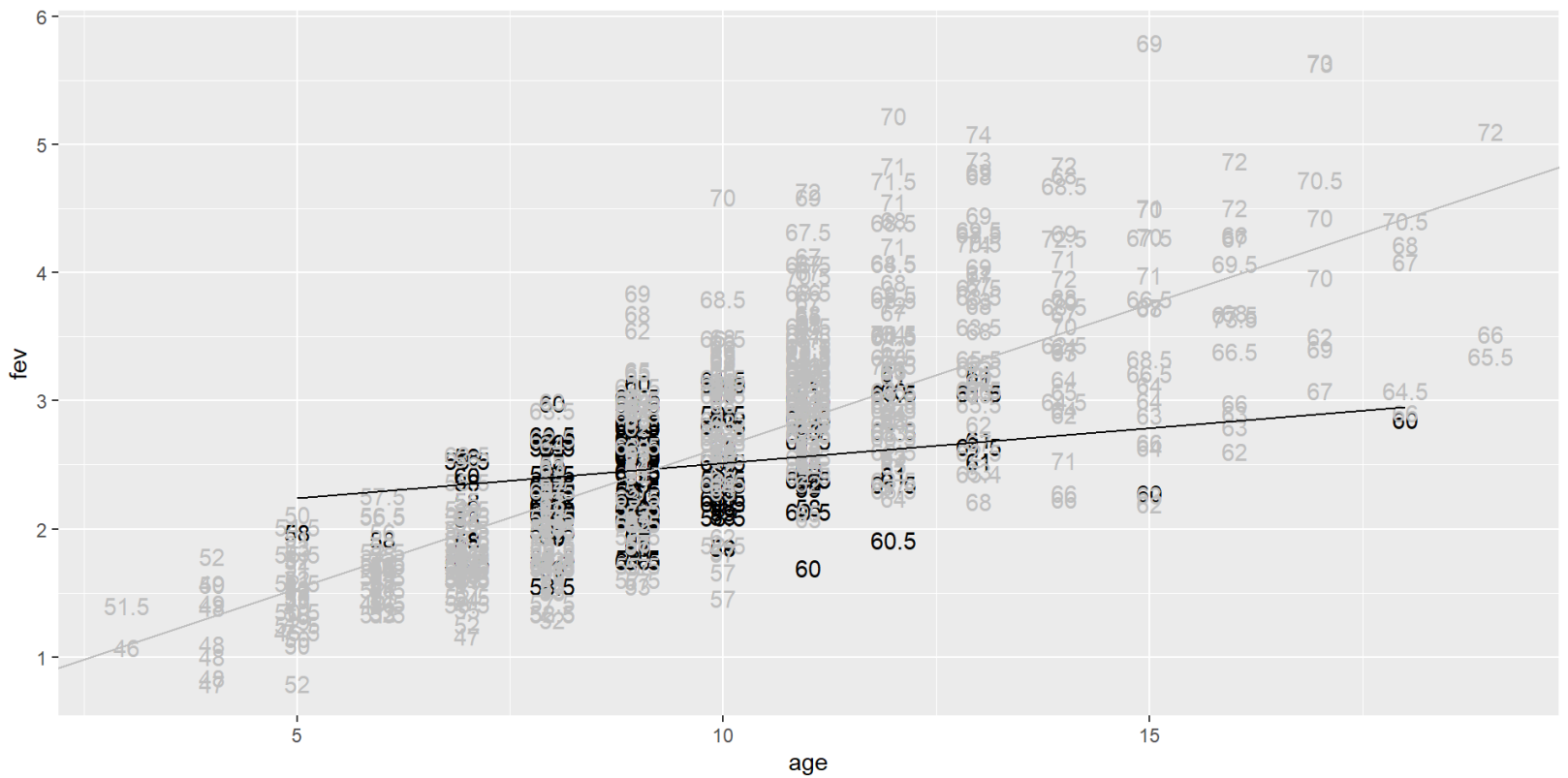
Relationship between age and FEV controlling for height between 50 and 53.5



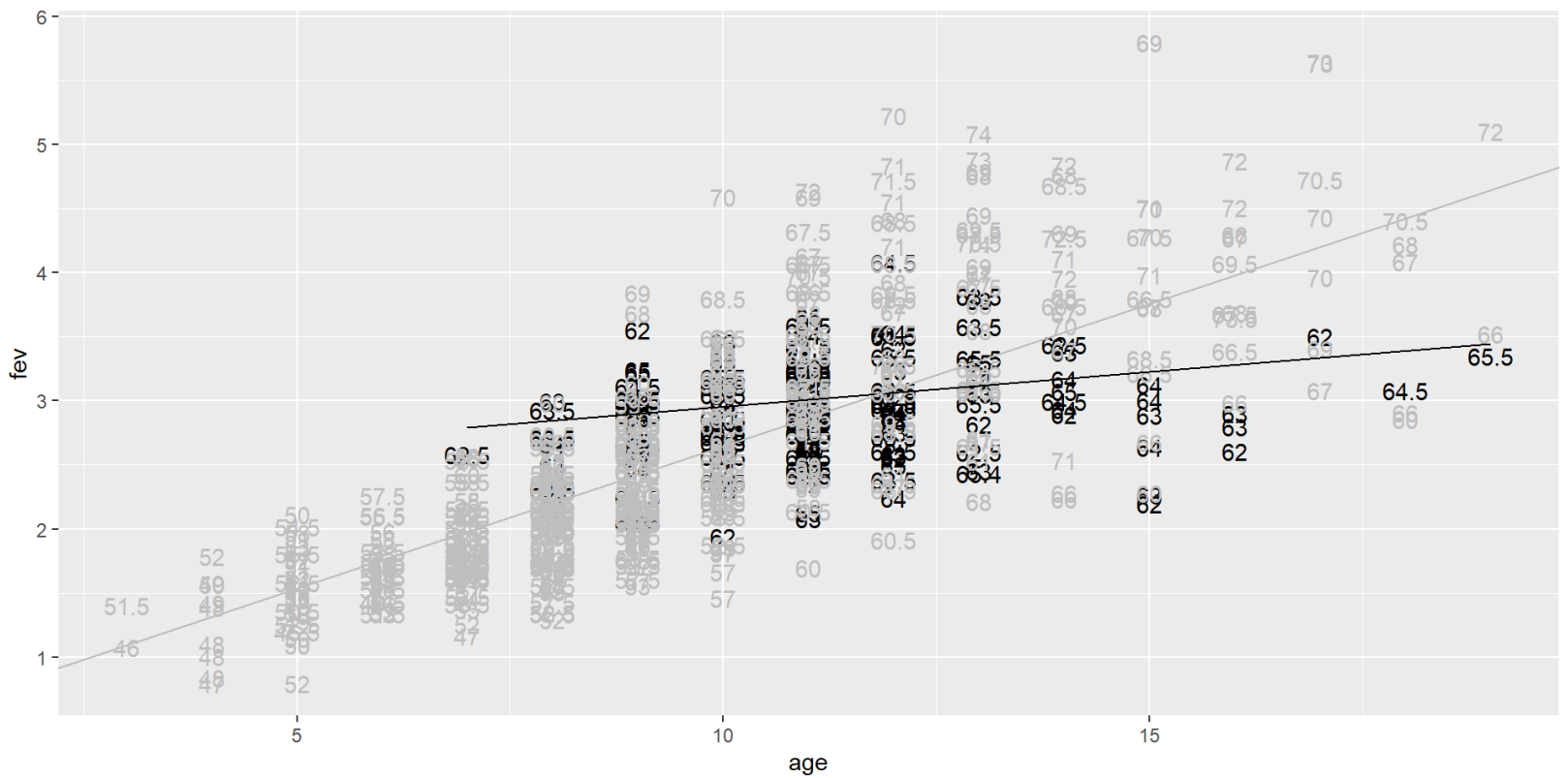
Relationship between age and FEV controlling for height between 54 and 57.5



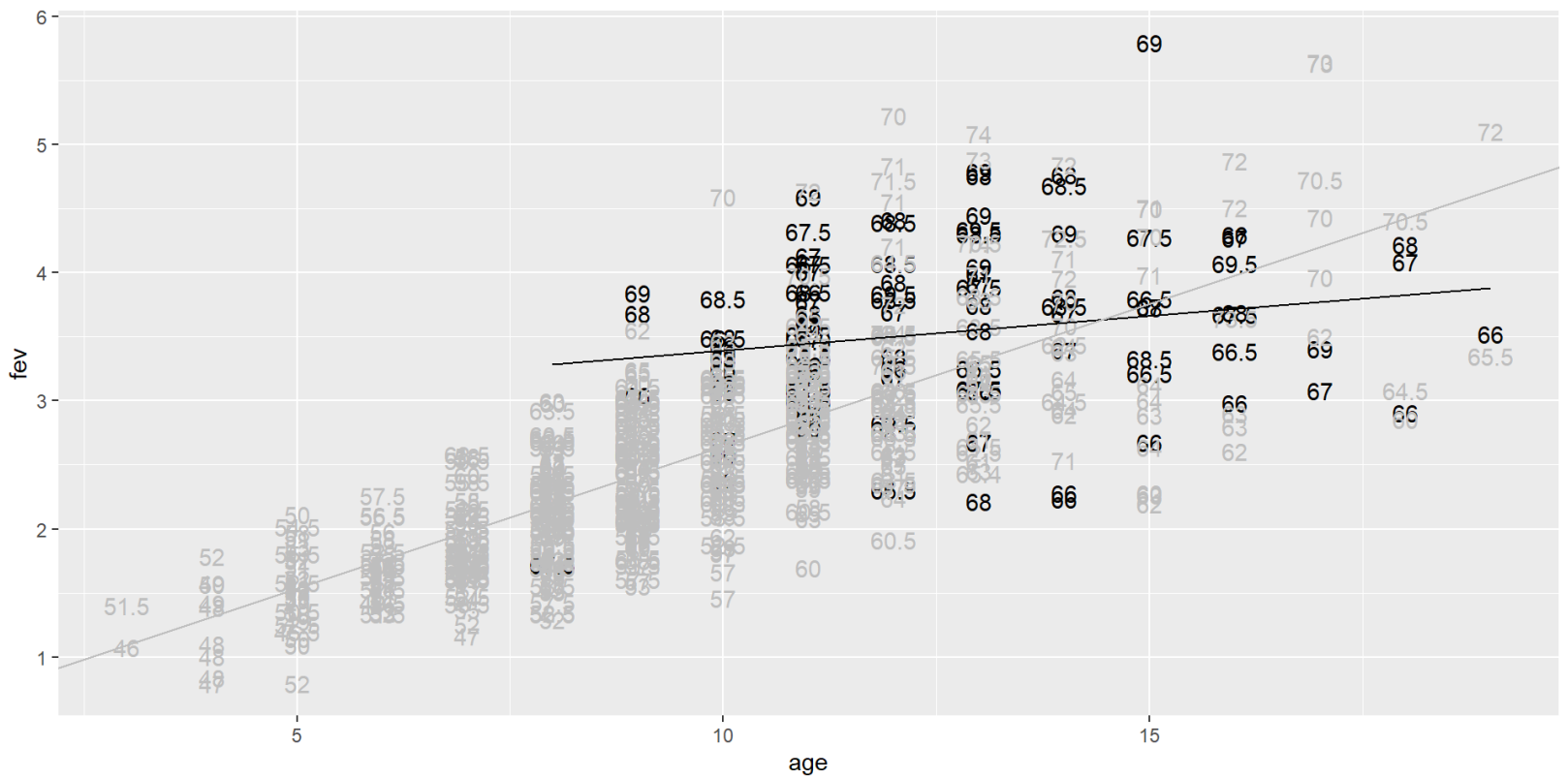
Relationship between age and FEV controlling for height between 58 and 61.5



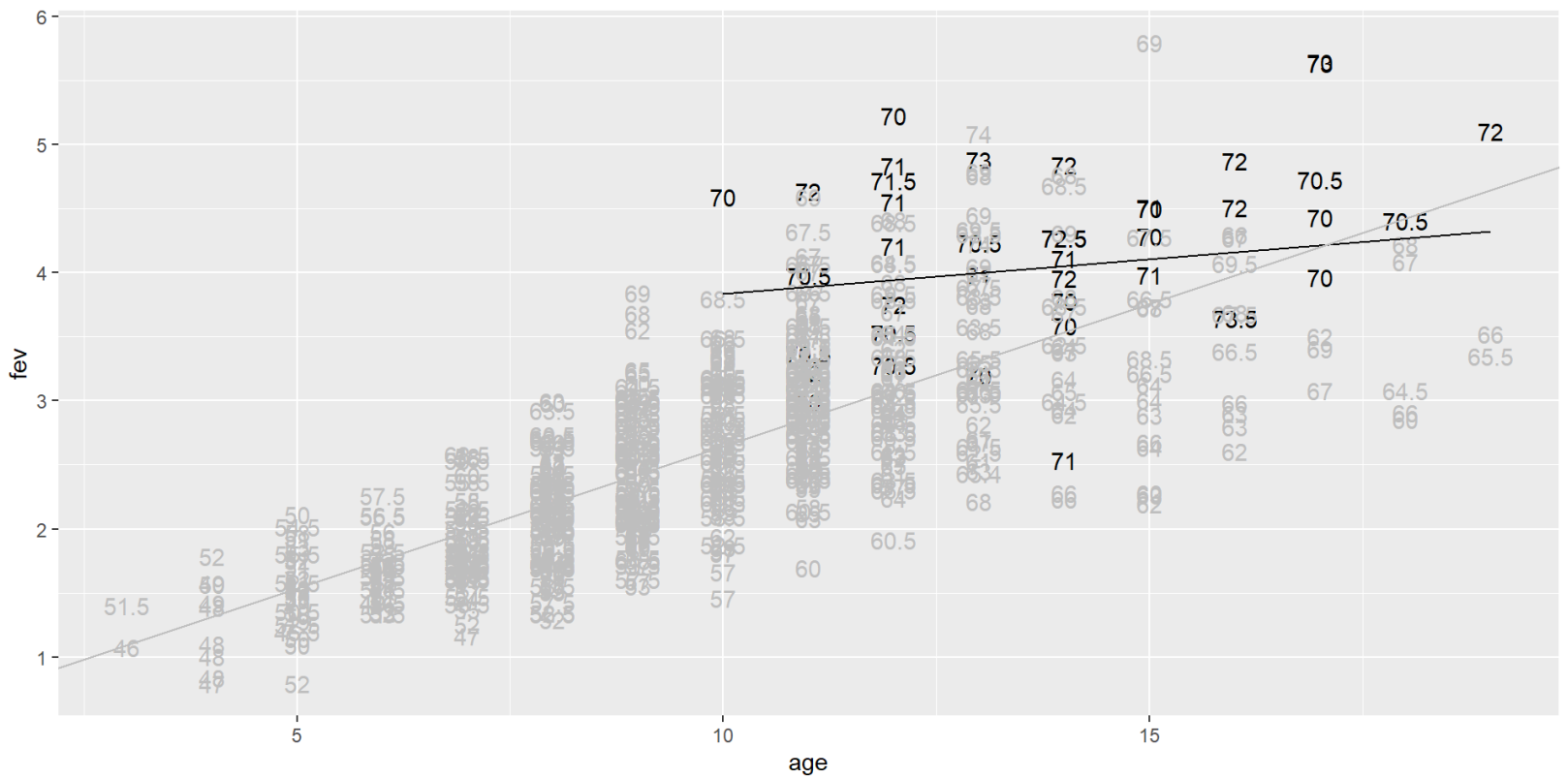
Relationship between age and FEV controlling for height between 62 and 65.5



Relationship between age and FEV controlling for height between 66 and 69.5



Relationship between age and FEV controlling for height between 70 and 73.5



Break #3

- What you have learned
  - Multiple linear regression
- What’s coming next
  - R code for multiple linear regression

simon-5501-07-fev.qmd, 24

m3: Linear regression model using age and height to predict fev

Previous analysis has shown that age by itself is a strong predictor of fev and height by itself is a strong predictor of fev. A multiple linear regression model including both age and height should do an even better job in predicting fev. This model will also allow you to compare the relative importance of the two variables. Is fev most strongly associated with how big a child is or how old that child is?

You should always start with descriptive statistics and graphs. With two or more independent variables, you should also examine the correlations between the independent variables and their correlations with the dependent variable.

simon-5501-07-fev.qmd, 25

````
## m3: Descriptive statistics for age

```{r age}
pulmonary |>
  summarize(
    age_mn=mean(age),
    age_sd=sd(age),
    age_min=min(age),
    age_max=max(age))
```
````

The descriptive statistics are consistent with a pediatric study. The average age is 9.9 years. The standard deviation, 3.0, shows a large amount of variation in age (large at least for a pediatric study). The range, 3 years to 19 years, also demonstrates a large amount of variation.

simon-5501-07-fev.qmd, 26

````
## m3: Descriptive statistics for height

```{r height}
pulmonary |>
  summarize(
    ht_mn=mean(height),
    ht_sd=sd(height),
    ht_min=min(height),
    ht_max=max(height))
```

The average height, 61 inches, is reasonable for a group of children with an average age of around 10 years. The standard deviation, 5.7 inches, and the range, 46 inches to 74 inches, show a moderate amount of variation.
````

simon-5501-07-fev.qmd, 27

````
## m3: Correlations

Reduce the pulmonary data frame to just the first three columns before
computing correlations.

```{r corr}
pulmonary |>
  select(height, age, fev) |>
  cor()
```

The two independent variables, age and ht, are both strongly correlated with
fev (r=0.76 and 0.87). They are also strongly correlated with one another
(r=0.79).
````

simon-5501-07-fev.qmd, 28

````
## m3: Predicting fev using age and ht

```{r m3}
m3 <- lm(fev ~ age + height, data=pulmonary)
m3
```
````

The estimated average fev value increases by 0.05 liters for each increase of one year in age, holding height constant. The estimated average fev value increases by 0.11 liters for every increase of one inch in height. The estimated average fev is -4.6 for a patient of age zero with a height of zero inches. This is clearly and extrapolation beyond the range of the data.

simon-5501-07-fev.qmd, 29

```
## m3: Confidence intervals, 1

The use of [2, ] (and of [3, ]) isolates the individual rows of the data frame
produced by confint. This is not really necessary, but is done to fit the
information easily on separate slides.

```{r m3-ci-1}
m3
confint(m3)[2, ]
```

We are 95% confident that the estimated average fev value increases between
0.036 and 0.072 liters for each increase of one year of age holding height
constant. Conclude that there is a positive relationship between fev and age.
```

simon-5501-07-fev.qmd, 30

````
## m3: Confidence intervals, 2

```{r m3-ci-2}
confint(m3)[3, ]
```

We are 95% confident that the estimated average fev value increases between 0.10 and 0.12 liters for each increase of one inch in patient's height holding age constant. Conclude that there is a positive relationship between height and fev.

Do not interpret the confidence interval for the intercept.
````

simon-5501-07-fev.qmd, 31

```
## m3: Analysis of variance table

```{r m3-anova}
anova(m3)
```
```

The sum of squares total (SST) is 280.9 + 95.3 + 114.7 = 490.9. Only a small portion of the variation (114.7) is unexplained variation. The F-ratio is much larger than 1, and the p-value is less than 0.001. You can conclude that the combination of age and height helps significantly in predicting fev.

Speaker notes

The first few lines are the documentation header

simon-5501-07-fev.qmd, 32

````
## m3: R-squared

```{r m3-r-squared}
glance(m3)$r.squared
```

Roughly 77% of the variation in fev can be explained by the combination of the age and height of the patients.
````

simon-5501-07-fev.qmd, 33

````
## m3: t-tests

```{r}
tidy(m3)
```
````

Break #4

- What you have learned
  - R code for multiple linear regression
- What's coming next
  - Diagnostic plots and multicollinearity

Assumptions

- Population model
  - $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i,\ i = 1, \ldots, N$
- Assumptions about ϵ_i
  - Normal distribution
  - Mean 0
  - Standard deviation sigma
  - Independent

Speaker notes

The population model requires that you have access to the entire population. The size of the population, N, is almost always a large number. It is a number so large that you have to rely on a much smaller subset of the population, a sample.

Because N is so large, β_0 and β_1 are unknown constants and ϵ_i is also unknown.

Residuals

- $\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i}$
- $e_i = Y_i - \hat{Y}_i$
  - Behavior of e_i helps evaluate assumptions about ϵ_i

Speaker notes

The residuals from the sample help you assess assumptions about the epsilons.

Assessing normality assumption

- Normal probability plot
- Histogram

Speaker notes

The assessment of normality is exactly the same.

Assessing heterogeneity, nonlinearity

- Plot e_i versus $\hat{Y}_i$
  - Composite of X_1 and X_2
  - Look for differences in variation
  - Look for curved pattern

Speaker notes

To assess heterogeneity and nonlinearity, you could look at the residuals versus each independent variable. That may be fine with just 2 independent variables, but is not tenable when you have 20 independent variables. The fitted value is a composite of the 2 or 20 independent variables. If you see nonlinearity or heterogeneity with any of the independent variables, it would be very likely to also show up in the plot using fitted values.

Independence is always assessed qualitatively

Speaker notes

Independence is assessed qualitatively. Look at the conditions under which the data were collected. Do the individual data values group into clusters. Measurements within a family or within a clinic could be correlated. Also, does proximity influence the outcome. An infectious disease, for example, might produce correlated results for people who are geographically close to one another.

Influential values

- Leverage
  - Compare to 3*(k+1)/n
    - k is number of independent variables
- Studentized deleted residual
  - Compare to ± 3
- Cook's distance
  - Compare to 1

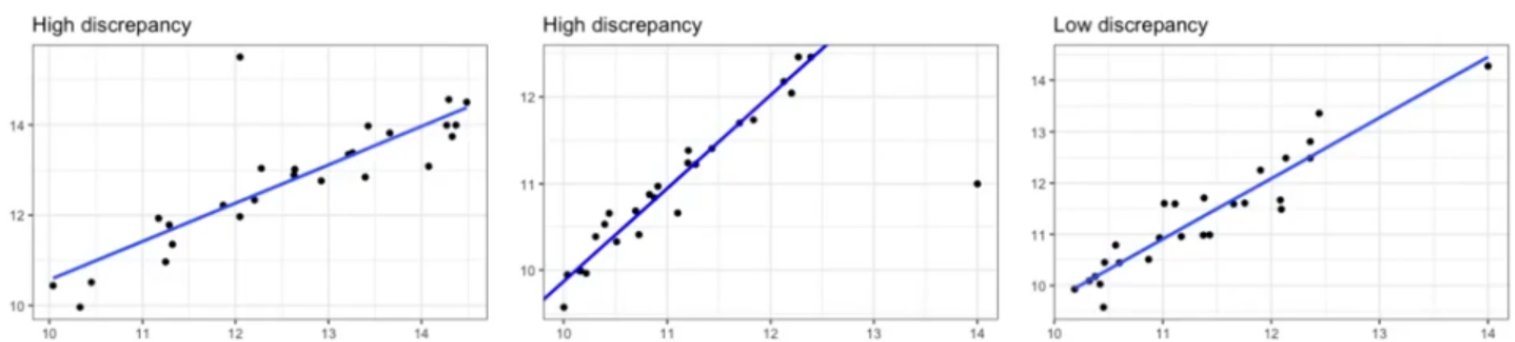
Speaker notes

Influential data values are extreme values which can have an undue influence on the location of the regression equation. Leverage is a measure of how extreme a data value is with respect to the two independent variables. It can be extreme with respect to the first, with respect to the second, or with respect to both. Odd combinations, such as a small birthweight associated with a large gestational age could be influential.

The studentized deleted residual is a measure of how far the dependent variable is from the predicted value. It is standardized and unitless and any value larger than plus or minus 3 is considered extreme.

Cook's distance is a composite measure and it largest when a high leverage value is associated with an extreme residual.

In a two-variable data set, individual points may be unusual in their x-values, their y-values, or both. Technically, an **outlier** in such a set is a point with high **discrepancy**, that is, one whose y-value is far from the general trend of the data. In practice, the word *outlier* is often used more loosely.



Each of these plots includes an unusual observation. Technically only the first two show outliers.

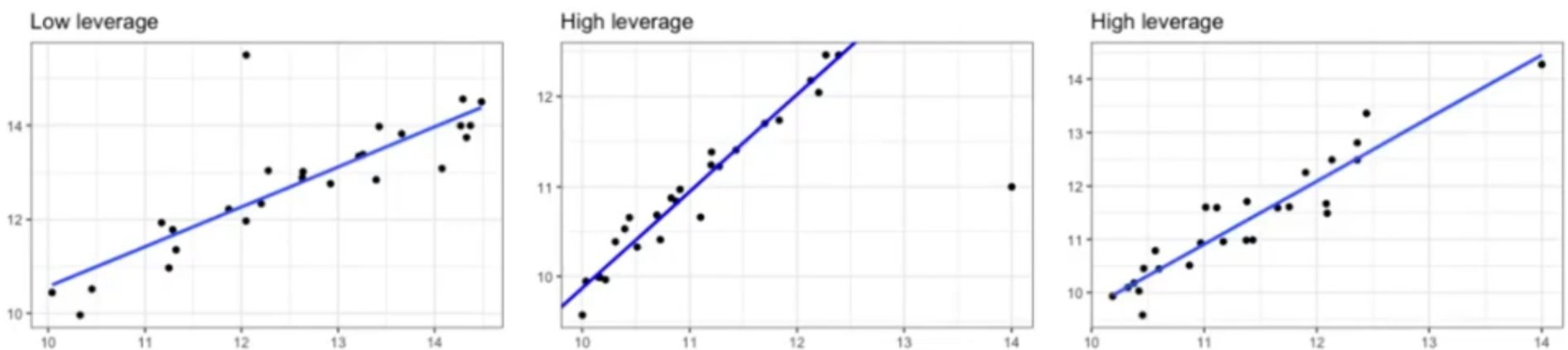
Leverage, 1

```
# A tibble: 15 × 9
     fev   age height .fitted  .resid   .hat .sigma  .cooksd .std.resid
   <dbl> <dbl>  <dbl>   <dbl>   <dbl>  <dbl>  <dbl>    <dbl>      <dbl>
 1  1.16     7   47     0.926  0.239  0.0148  0.420 0.00165       0.574
 2  2.91    18   66     3.61  -0.702  0.0200  0.419 0.0194       -1.69
 3  5.10    19   72     4.32   0.782  0.0171  0.419 0.0205        1.88
 4  3.52    19   66     3.66  -0.143  0.0262  0.420 0.00107      -0.345
 5  3.34    19   65.5   3.61  -0.262  0.0274  0.420 0.00376      -0.633
 6  3.08    18   64.5   3.44  -0.361  0.0231  0.420 0.00598      -0.870
 7  2.90    16   63     3.17  -0.267  0.0150  0.420 0.00208      -0.641
 8  4.22    18   68     3.83   0.393  0.0167  0.420 0.00506       0.944
 9  3.5     17   62     3.11   0.386  0.0228  0.420 0.00672       0.929
10  2.61    16   62     3.06  -0.452  0.0170  0.420 0.00679      -1.09
11  4.09    18   67     3.72   0.369  0.0183  0.420 0.00487       0.887
12  4.40    18   70.5   4.10   0.303  0.0141  0.420 0.00251       0.726
```

Speaker notes

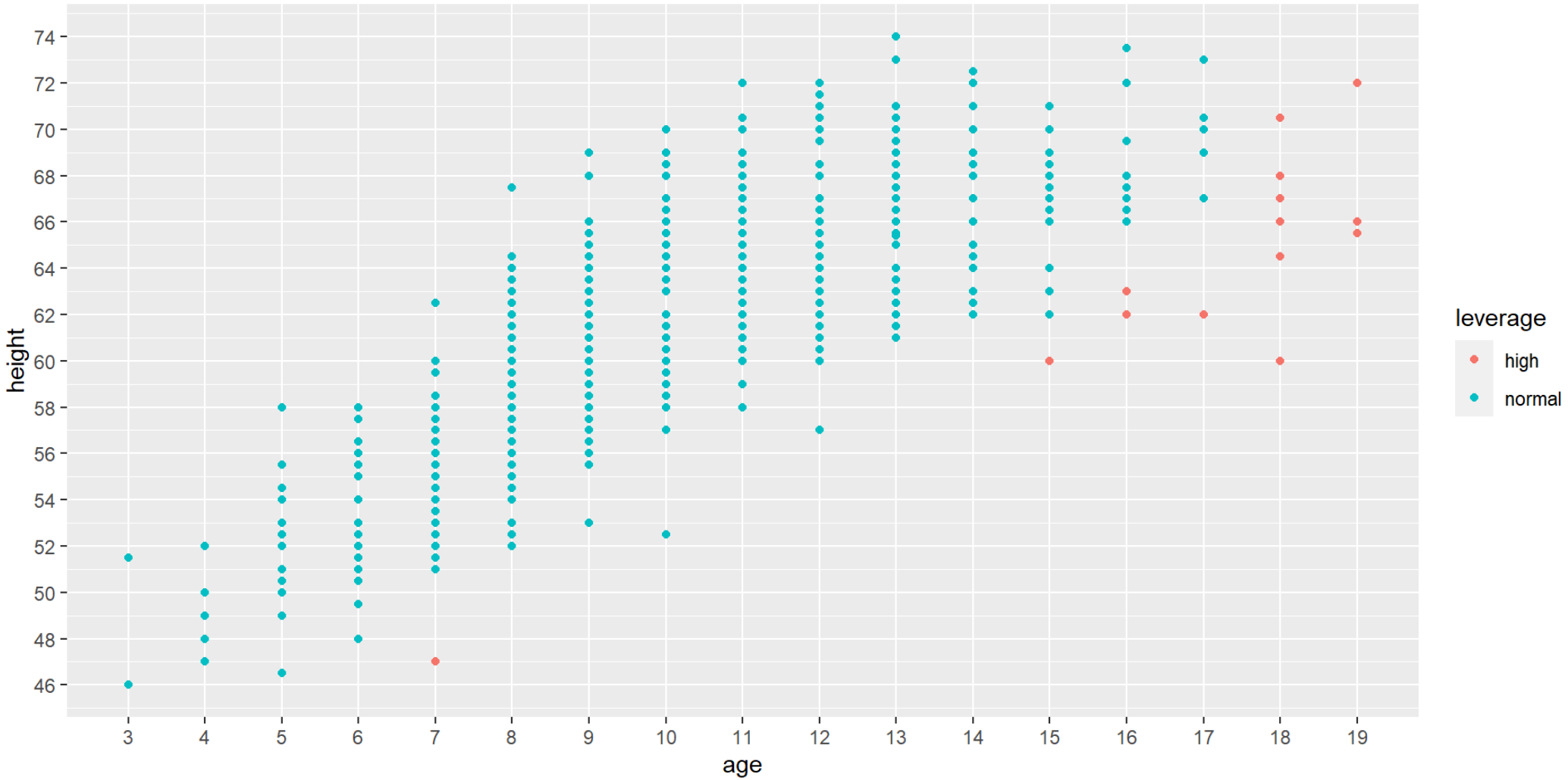
Here are the leverage values for a regression model using both hieght and age to predict fev. There are quite a few values and they are a bit tricky to interpret. This is typical for two independent variables. Finding out why a data point has high leverage gets even harder when there are three or more independent variables.

Observations with unusual x-values have greater potential to affect the fit of the model. Such observations are said to have high **leverage**. The leverage of a point only depends on its x-value.



Only the second two plots show observations with high leverage.

Leverage, 2


height
74
72
70
68
66
64
62
60
58
56
54
52
50
48
46
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
age
leverage
high
normal

Speaker notes

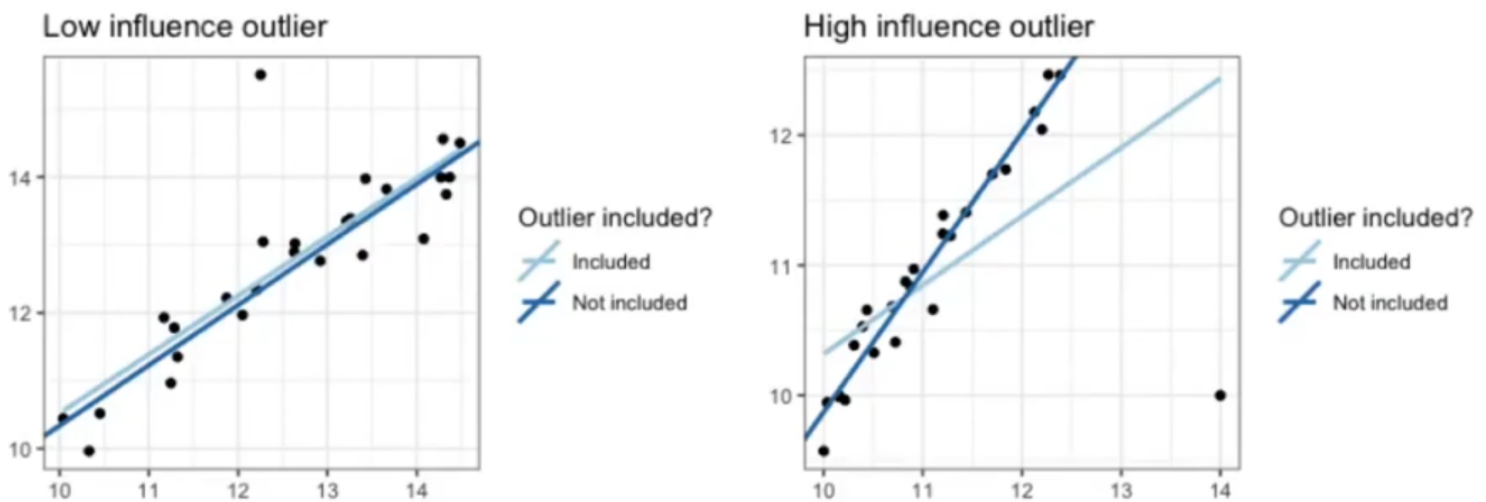
Here a graph can help a bit. In this graph, the two independent variables, age and height are displayed and high leverage points are highlighted in a different color.

One of the leverage points is a 7 year old with a height of ony 47 inches, close to the smallest height but not at all close to the youngest age.

Other high leverage points are the handful of patients aged 18 and 19 years, plus a few 15 and 16 year olds who are very short for their ages.

These values should be investigated, but they do not seem so extreme as to warrant their possible exclusion from the regression model.

An observation that substantially changes the fit of a model is said to have high **influence**.



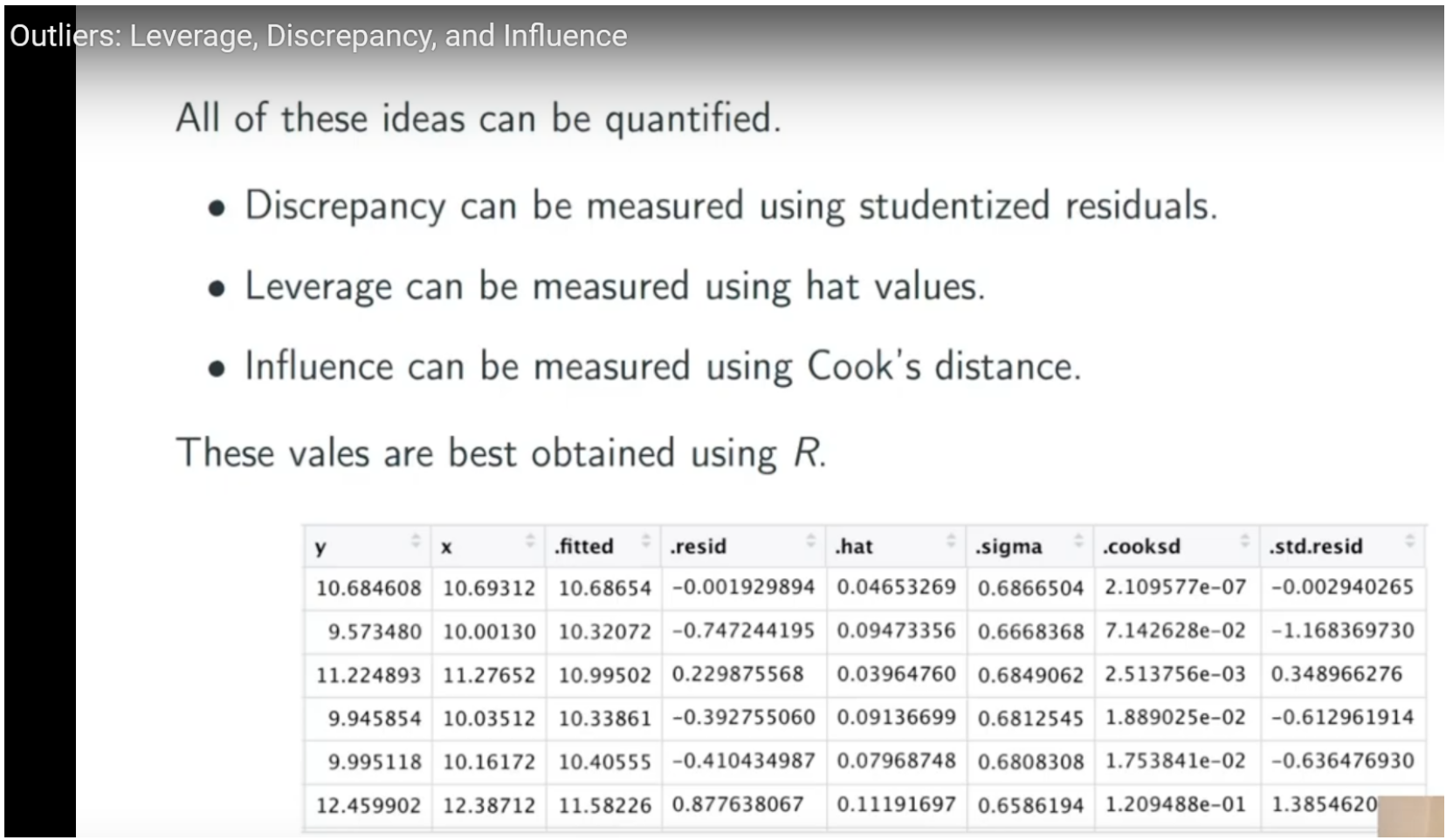
Typically, an observation must have high discrepancy *and* high leverage to be influential.

Studentized residuals, 1

```
# A tibble: 7 × 9
    fev   age height .fitted .resid    .hat .sigma .cooksd .std.resid
  <dbl> <dbl>  <dbl>   <dbl>  <dbl>   <dbl>  <dbl>   <dbl>       <dbl>
1  1.72     8   67.5    3.23  -1.51 0.0131   0.416  0.0578      -3.61
2  5.22    12   70      3.72   1.50 0.00637  0.416  0.0276       3.59
3  2.54    14   71      3.94  -1.40 0.00610  0.416  0.0229      -3.35
4  2.22    13   68      3.56  -1.34 0.00377  0.417  0.0129      -3.20
5  5.79    15   69      3.77   2.02 0.00604  0.412  0.0472       4.83
6  5.63    17   73      4.32   1.31 0.0104   0.417  0.0347       3.14
7  5.64    17   70      3.99   1.65 0.0108   0.415  0.0565       3.94
```

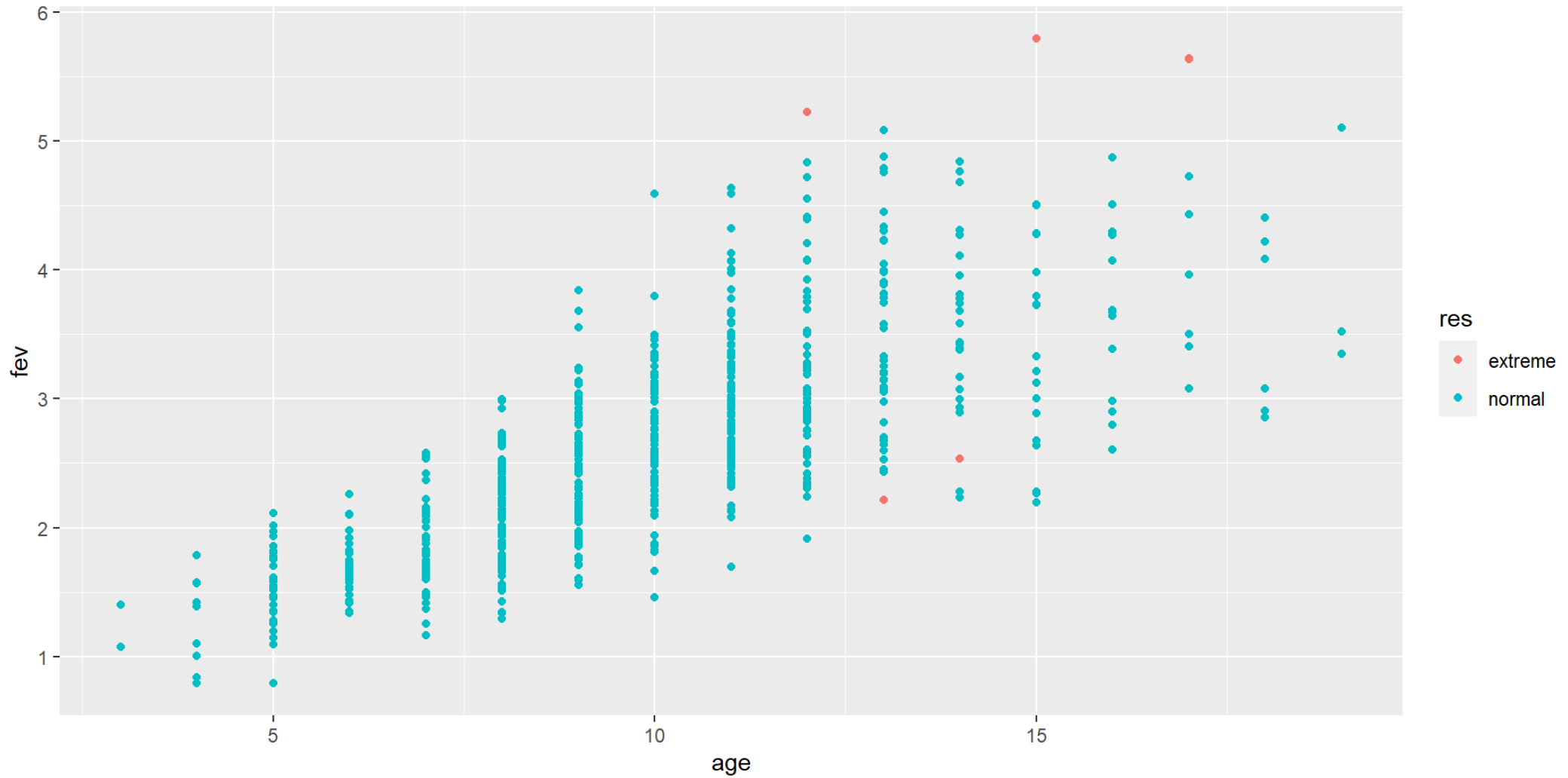
Speaker notes

Again, the interpretation is a bit tricky.



y	x	.fitted	.resid	.hat	.sigma	.cooksd	.std.resid
10.684608	10.69312	10.68654	-0.001929894	0.04653269	0.6866504	2.109577e-07	-0.002940265
9.573480	10.00130	10.32072	-0.747244195	0.09473356	0.6668368	7.142628e-02	-1.168369730
11.224893	11.27652	10.99502	0.229875568	0.03964760	0.6849062	2.513756e-03	0.348966276
9.945854	10.03512	10.33861	-0.392755060	0.09136699	0.6812545	1.889025e-02	-0.612961914
9.995118	10.16172	10.40555	-0.410434987	0.07968748	0.6808308	1.753841e-02	-0.636476930
12.459902	12.38712	11.58226	0.877638067	0.11191697	0.6586194	1.209488e-01	1.3854620

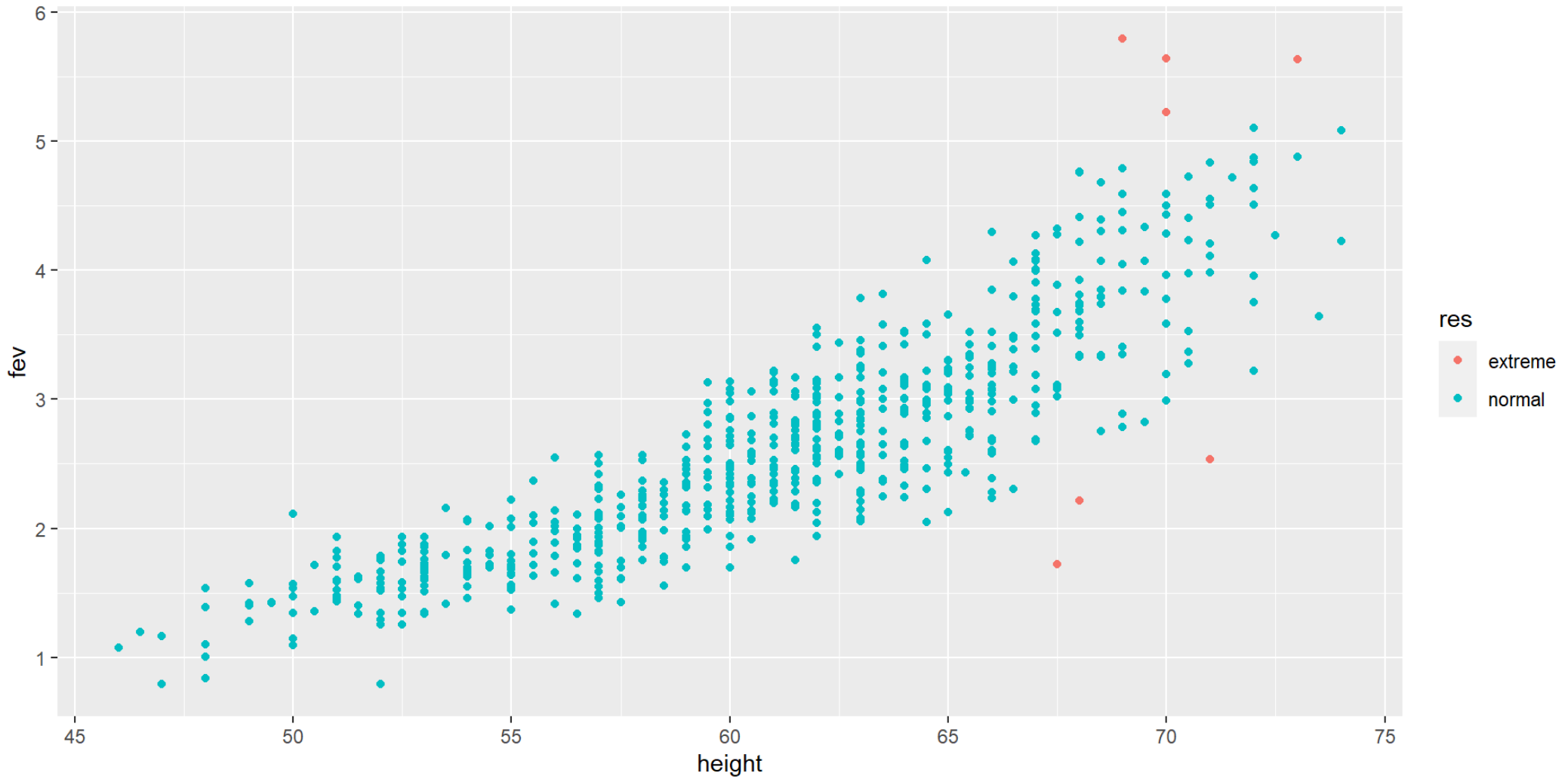
Studentized residuals, 2



Speaker notes

The large studentized residuals are associated with older patients, but the pattern is not consistent.

Studentized residuals, 3



Speaker notes

Here the pattern is a bit more consistent. Taller patients with very low or with very high fev values are flagged as exteme based on the studentized residuals.

Cook's distance

In the pulmonary database, no combination of high leverage and extreme studentized residuals is going to cause concern.

Note: interpreting influential values gets tricky with two independent variables.

Speaker notes

Assessing influential data points is not easy.

You can’t always just look at the values to see what is causing the extreme values. In this example, a graph helps. Sometimes even a graph doesn’t help. It gets even harder with three or more independent variables.

Don’t get too anxious about this. It’s not easy, but if everything in Statistics was easy, you’d be getting the minimum wage for your work when you graduate.

You will see more discussions about the complexities associated with multiple independent variables in MEDB 5502, Applied Biostatistics, II.

Multicollinearity

- Synonyms:
  - Collinearity
  - Ill-conditioning
  - Near collinearity
- When two variables are correlated
- When three or more variables add up to nearly a constant

Speaker notes

In a regression model with more than one independent variable, you need to assess whether the data exhibits multicollinearity.

Multicollinearity occurs when the two independent variables are highly correlated. It can also occur in more complex models when three or more variables add up to a value that is nearly constant.

Problems caused by multicollinearity

- Intepretation
  - What does “holding one variable constant” mean?
  - Difficult to disentangle the individual impacts
- Inflated standard errors
  - Very wide confidence intervals
  - Loss of statistical power
- Note: multicollinearity is NOT a violation of assumptions

Speaker notes

Multicollinearity makes interpretation difficult. It also leads to a loss of precision and power.

Variance inflation factor (VIF)

- How much precision is lost due to multicollinearity
- Values larger than 10 are cause for concern

Speaker notes

The variance inflation factor is a measure of how much precision is lost due to multicollinearity.

Break #5

- What you have learned
  - Diagnostic plots and multicollinearity
- What's coming next
  - R code for diagnostic plots and multicollinearity

simon-5501-07-fev.qmd, 34

````
## m3: Diagnostic plots, 1

```{r diagnostic-1}
r3 <- augment(m3)
r3 |>
    ggplot(aes(sample=.resid)) +
      stat_qq() +
    ggtitle("Graph drawn by Steve Simon on 2024-09-30")
```
````

simon-5501-07-fev.qmd, 35

````
## m3: Diagnostic plots, 2

```{r diagnostic-2}
r3 |>
  ggplot(aes(.resid)) +
    geom_histogram(
      binwidth=0.2,
      color="black",
      fill="white") +
    xlab("Residuals from m3 regression") +
    ggtitle("Graph drawn by Steve Simon on 2024-09-25")
```
````

simon-5501-07-fev.qmd, 36

````
## m3: Diagnostic plots, 3

```{r diagnostic-3}
r3 |>
  ggplot(aes(.fitted, .resid)) +
    geom_point() +
    xlab("Predicted values from m3 regression") +
    ylab("Residuals from m3 regression") +
    ggtitle("Graph drawn by Steve Simon on 2024-09-25")
```
````

simon-5501-07-fev.qmd, 37

````
## m3: Influential data points, 1

```{r influence-1}
n <- nrow(r3)
r3 |> filter(.hat > 3*3/n)
```
````

simon-5501-07-fev.qmd, 38

````
## m3: Influential data points, 2

```{r influence-2}
r3 |>
  filter(abs(.std.resid) > 3)
```
````

simon-5501-07-fev.qmd, 39

````
## m3: Influential data points, 3

```{r influence-3}
r3 |>
  filter(.cooksd > 1)
```
````

simon-5501-07-fev.qmd, 40

````
## m3: Variance inflation factor

```{r vif}
library(car)
vif(m3)
```
````

Break #6

- What you have learned
  - R code for diagnostic plots and multicollinearity
- What’s coming next
  - Your homework

simon-5501-07-directions.qmd, 1

```
---
title: "Directions for 5501-07 programming assignment"
format:
  html:
    embed-resources: true
date: 2024-09-27
---

This file was written by Steve Simon on 2024-09-27 and is placed in the public
domain.
```

simon-5501-07-directions.qmd, 2

```
## Program

-   Download [simon-5501-07-fev.qmd][tem]
    -   Store it in your src folder
-   Modify the file name
    -   Use your last name instead of "simon"
-   Modify the documentation header
    -   Add your name to the author field
    -   Optional: change the copyright statement

[tem]: https://github.com/pmean/classes/blob/master/biostats-1/07/src/simon-
5501-07-fev.qmd
```

simon-5501-07-directions.qmd, 3

```
## Data

-   Download [breast-feeding-preterm.csv][dat]
    -   Store it in your data folder
    -   Refer to the [data dictionary][dic] if needed

[dat]: https://github.com/pmean/datasets/blob/master/breast-feeding-
preterm.csv
[dic]: https://github.com/pmean/datasets/blob/master/breast-feeding-
preterm.yaml
```

simon-5501-07-directions.qmd, 4

```
## Question 1

Change the program so that it reads in the breast-feeding-preterm.csv file.
Show a glimpse of the data and verify that you have properly read in all 82
rows and 31 columns. No intepretation is necessary for this question.

## Question 2

Compute descriptive statistics (counts and percentages) for feed_type.
Interpret these values.

## Question 3

Compute descriptive statistics (mean, standard deviation, minimum, and
maximum) for age_stop. Interpret these values.
```

simon-5501-07-directions.qmd, 5

```
## Question 4

Draw a boxplot comparing age_stop for each level of feed_type. Interpret this
plot.

## Question 5

Calculate the means and standard deviations for each level of feed_type.
Interpret these numbers.

## Question 6

Compute a linear regression model predicting age_stop using feed_type. What
value does R assign to 0 and what value does R assign to 1? Interpret the
slope and intercept for this linear regression model.
```

simon-5501-07-directions.qmd, 6

Question 7

Compute R-squared for this regression model. Interpret this number.

Question 8

Draw a normal probability plot and a histogram for the residuals from this regression model. Is the assumption of normality satisfied?

Question 9

Calculate descriptive statistics (mean, standard deviation, minimum, and maximum) for mom_age and para. Interpret these values.

simon-5501-07-directions.qmd, 7

```
## Question 10

Calculate the correlations between mom_age, para, and age_stop. Interpret
these values.

## Question 11

Draw a scatterplot with mom_age on the x-axis and age_stop on the y-axis.
Repeat this with para on the x-axis. Interpret these plots.

## Question 12

Compute a linear regression model using mom_age and para to predict age_stop.
Interpret the regression coefficients.
```

simon-5501-07-directions.qmd, 8

Question 13

Compute R-squared for this regression model. Interpret this number.

Question 14

Draw a normal probability plot and a histogram of the residuals. Interpret these plots.

Question 15

Draw a plot with the predicted values on the x-axis and the residuals on the y-axis. Is there any evidence of heterogeneity or non-linearity?

simon-5501-07-directions.qmd, 9

```
## Question 16

Display any extreme values for leverage (greater than 3*2/n), studentized
deleted residuals (absolute value greater than 3), and for Cook's distance
(greater than 1). Explain why these values are extreme.

## Your submission

-   Save the output in html format
-   Convert it to pdf format.
-   Make sure that the pdf file includes
    -   Your last name
    -   The number of this course
    -   The number of this module
-   Upload the file
```

simon-5501-07-directions.qmd, 10

```
## If it doesn't work

If your program has any errors or fails to produce the output that you desire
and you can't resolve the problem, upload the program file along with the pdf
file to help us figure out what went wrong. You will get a chance to resubmit
the assignment if needed.
```

Summary

- What you have learned
  - Categorical independent variables
  - R code for categorical independent variables
  - Multiple linear regression
  - R code for multiple linear regression
  - Diagnostic plots and multicollinearity
  - R code for diagnostic plots and multicollinearity
  - Your homework