MEDB 5501, Module07

2025-10-07

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.

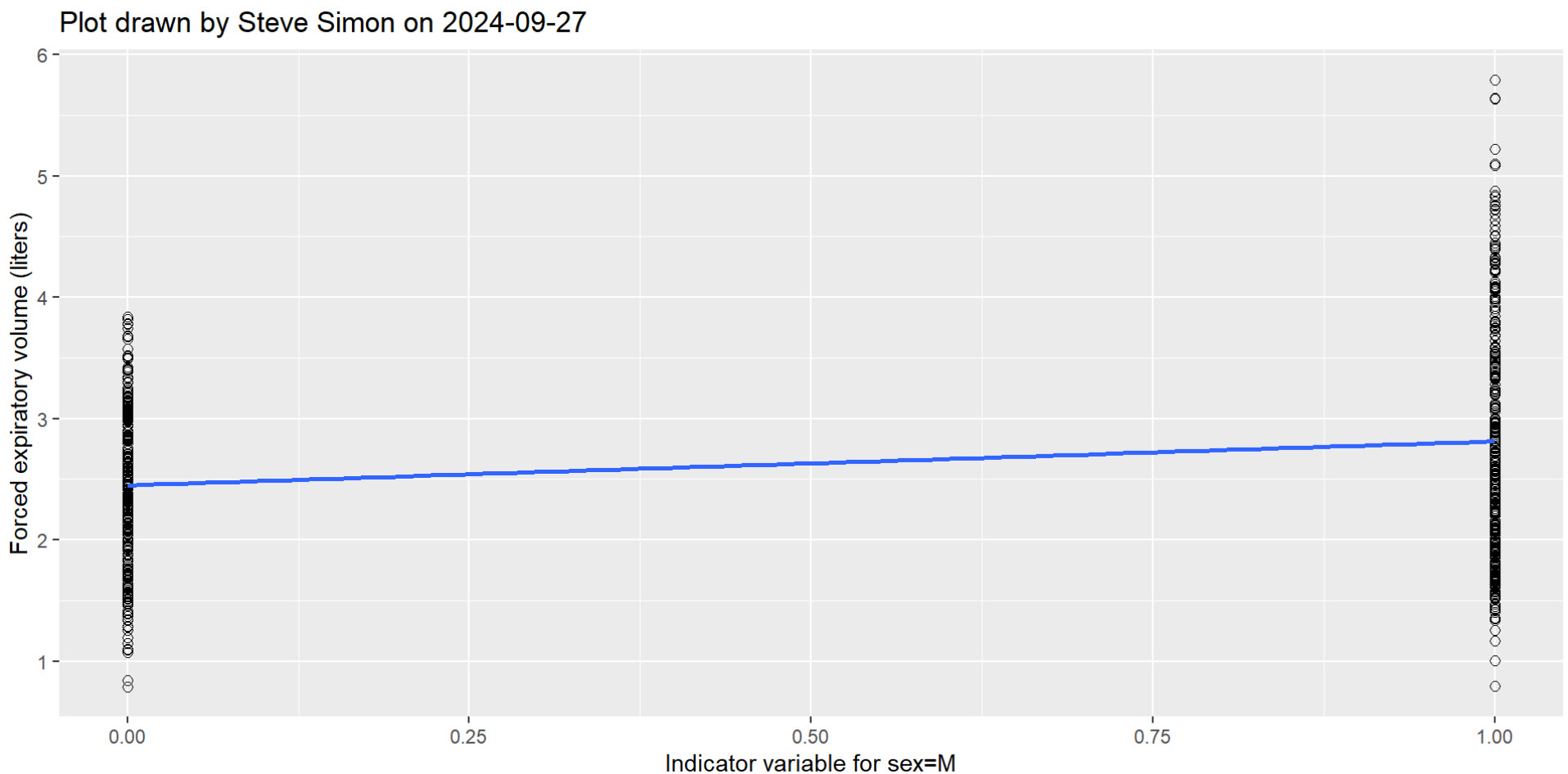
Creating an indicator variable

```
    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

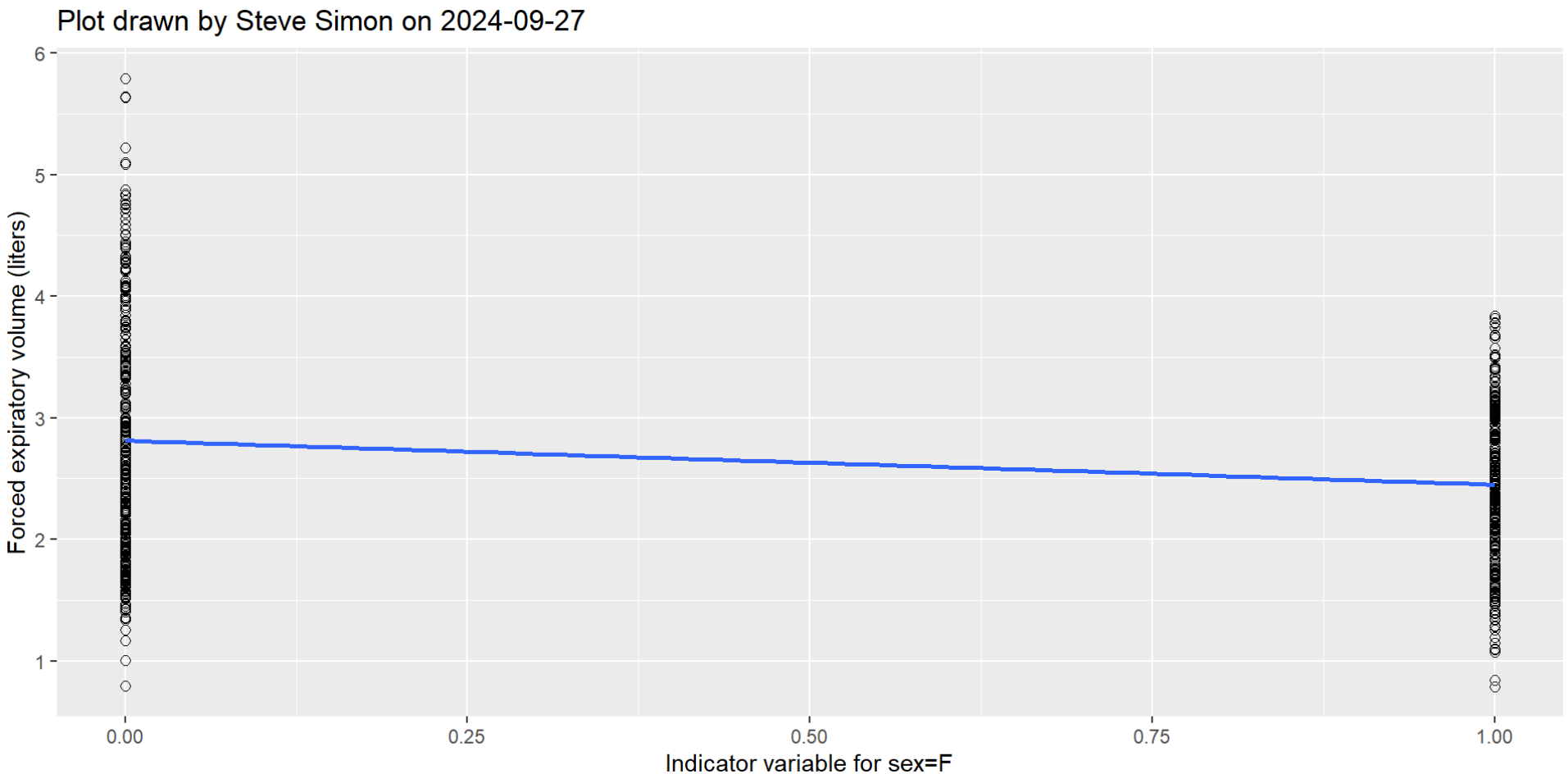
```
Call:
lm(formula = fev ~ sex_male, data = fev_a)

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

Refer to the data dictionary.

simon-5501-07-fev.qmd

Refer to part 1 of my code.

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: $b_0,\ b_1,\ b_2$

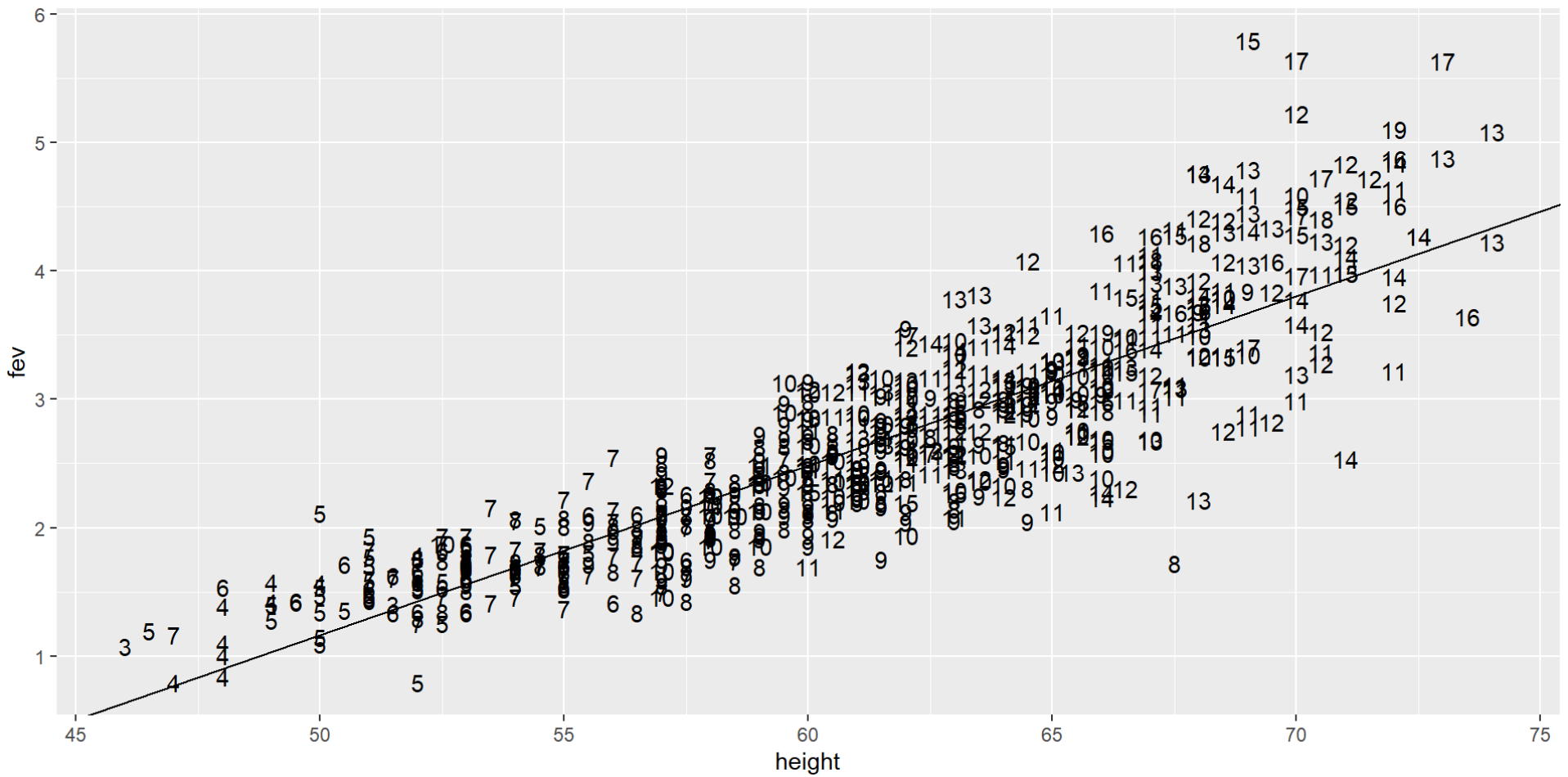
Speaker notes

Add note.

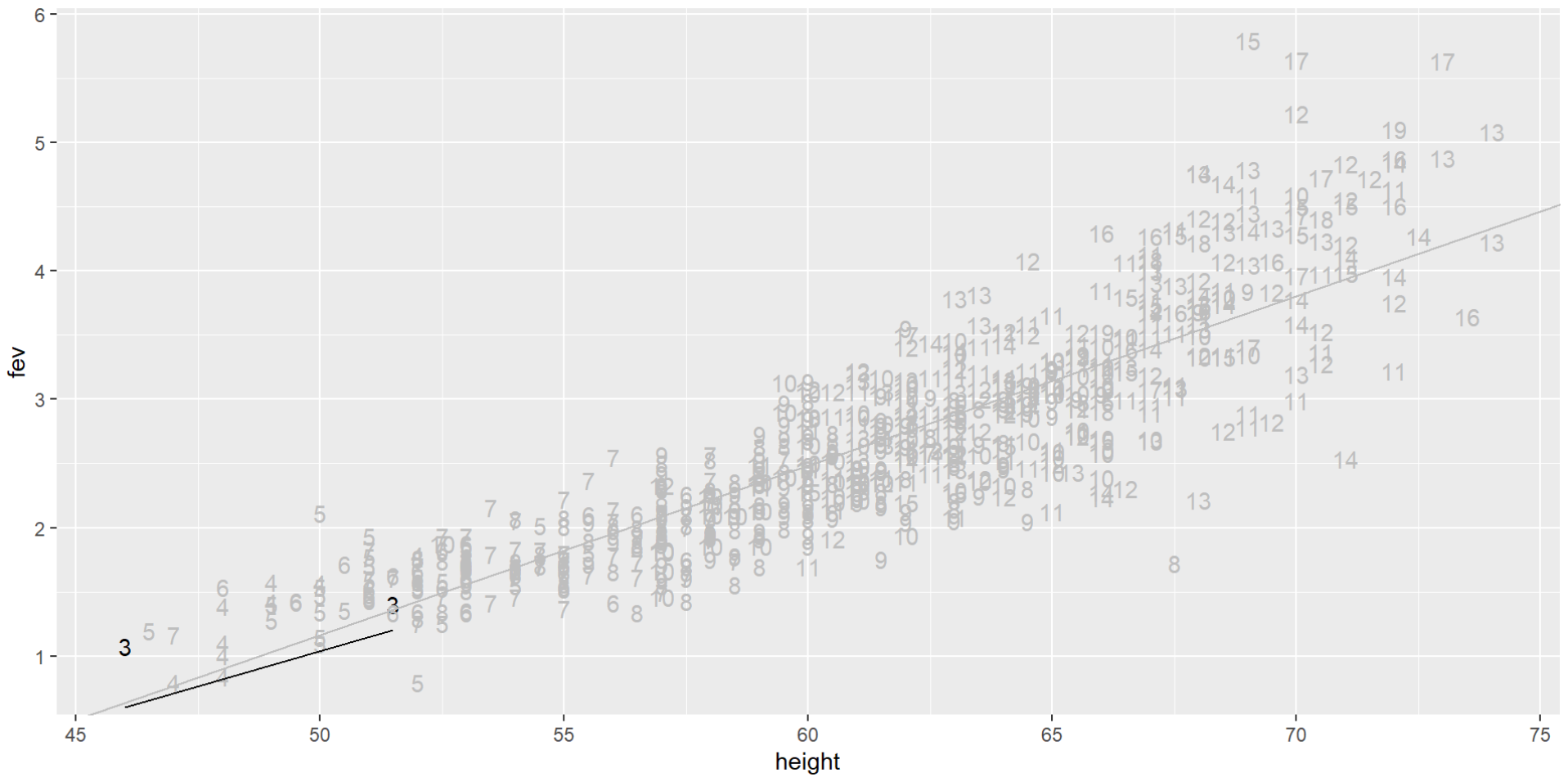
Interpretations

- 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

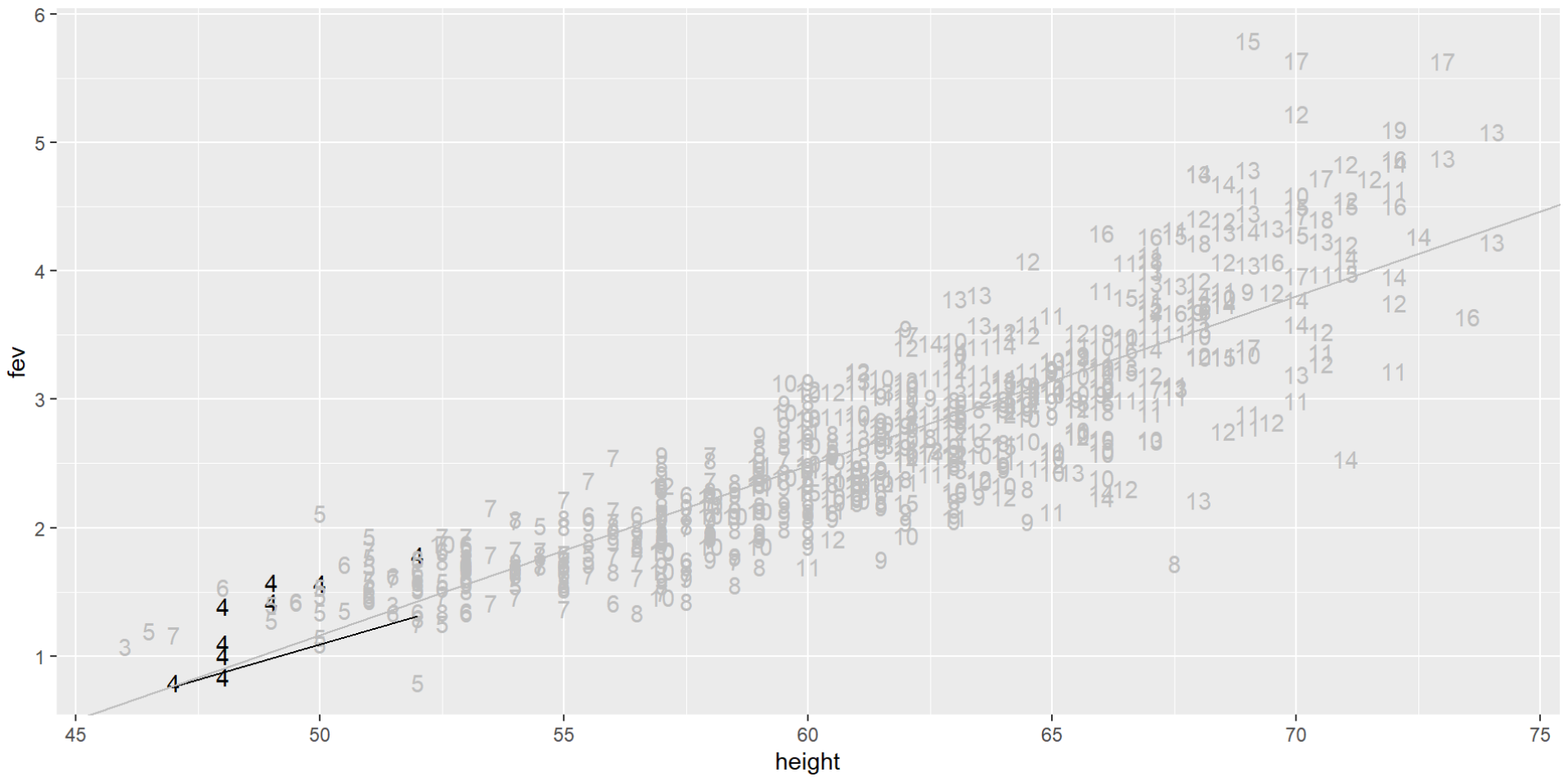
Unadjusted relationship between height and FEV



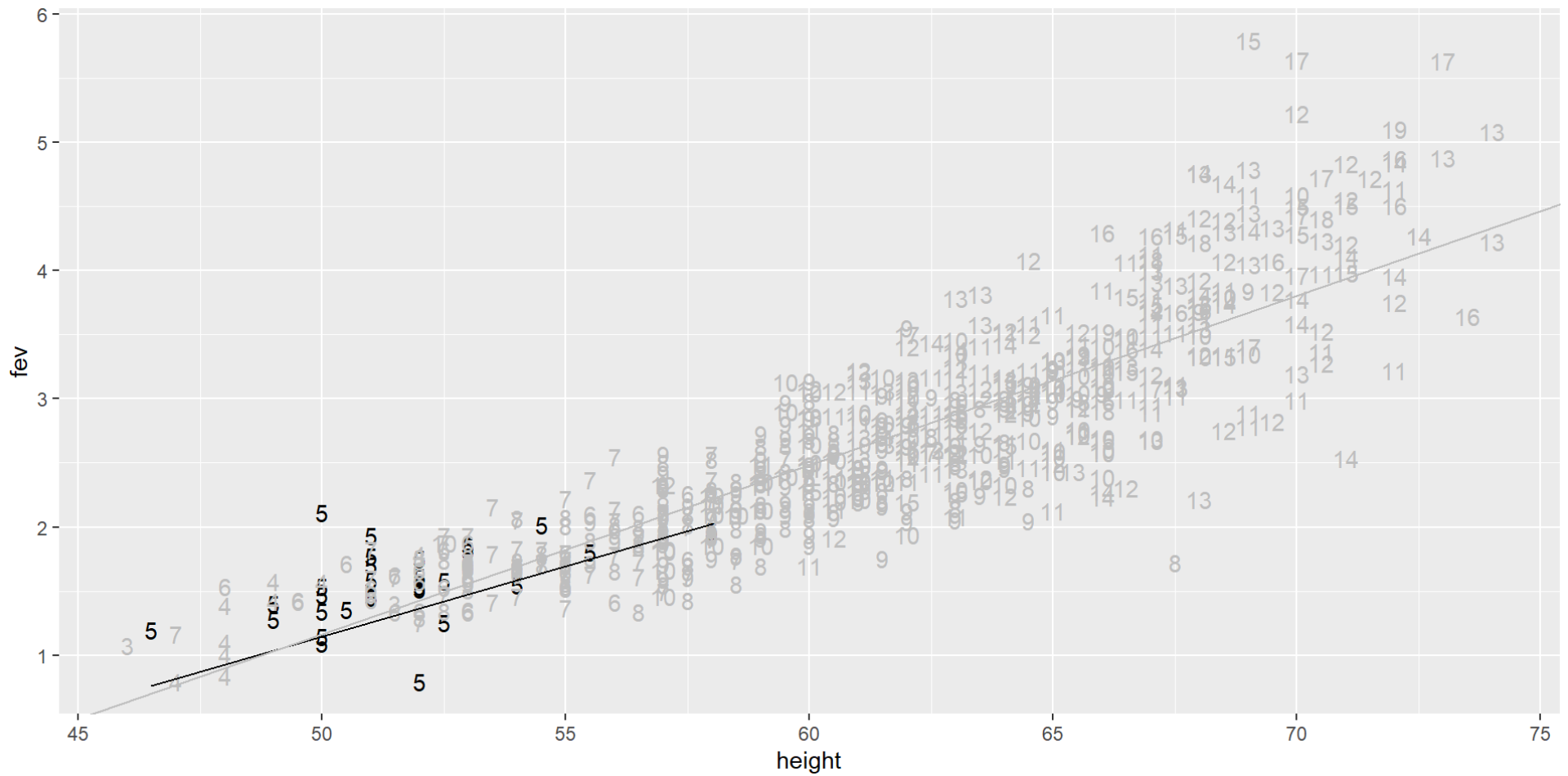
Relationship between height and FEV controlling at Age=3



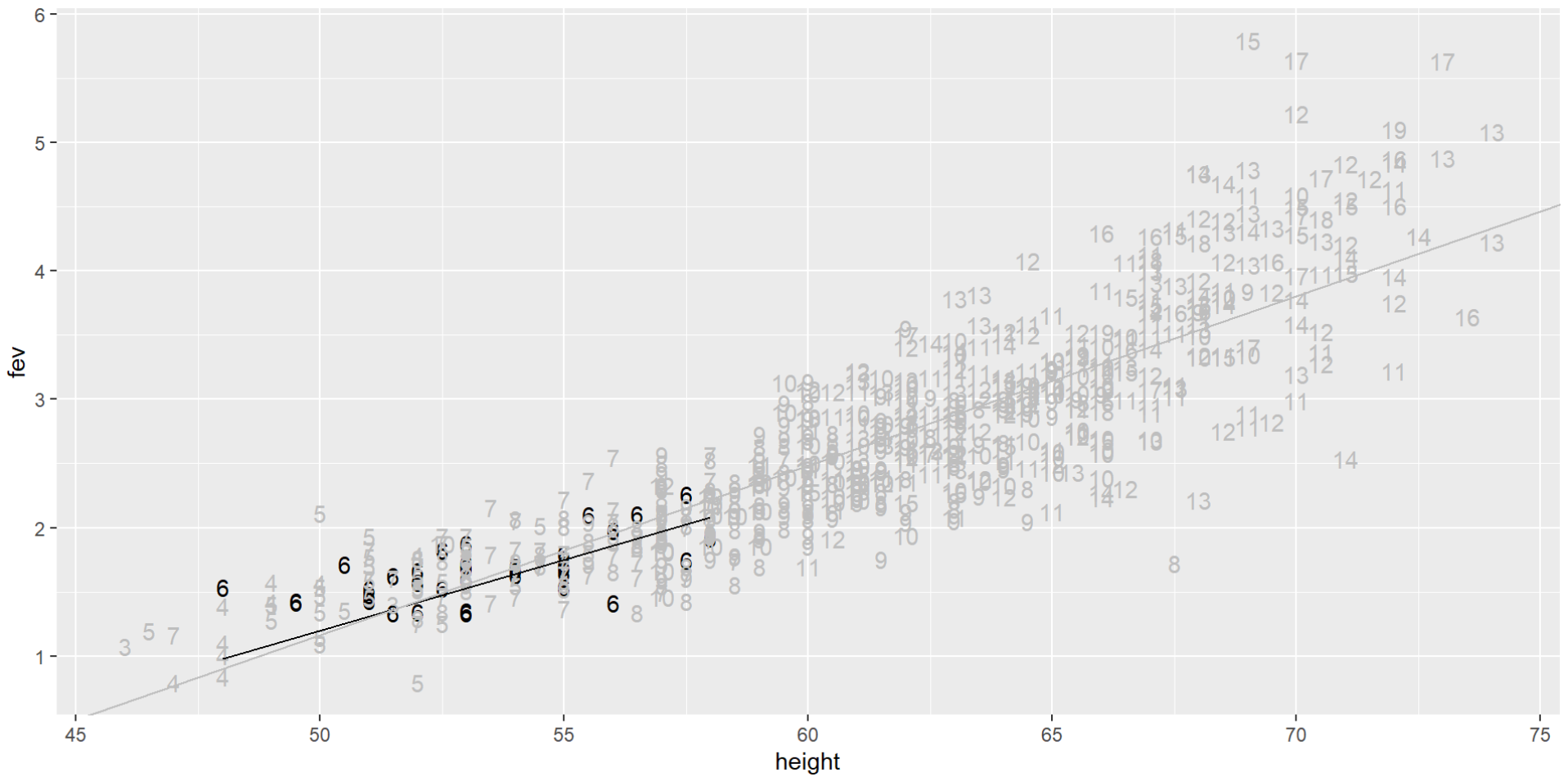
Relationship between height and FEV controlling at Age=4



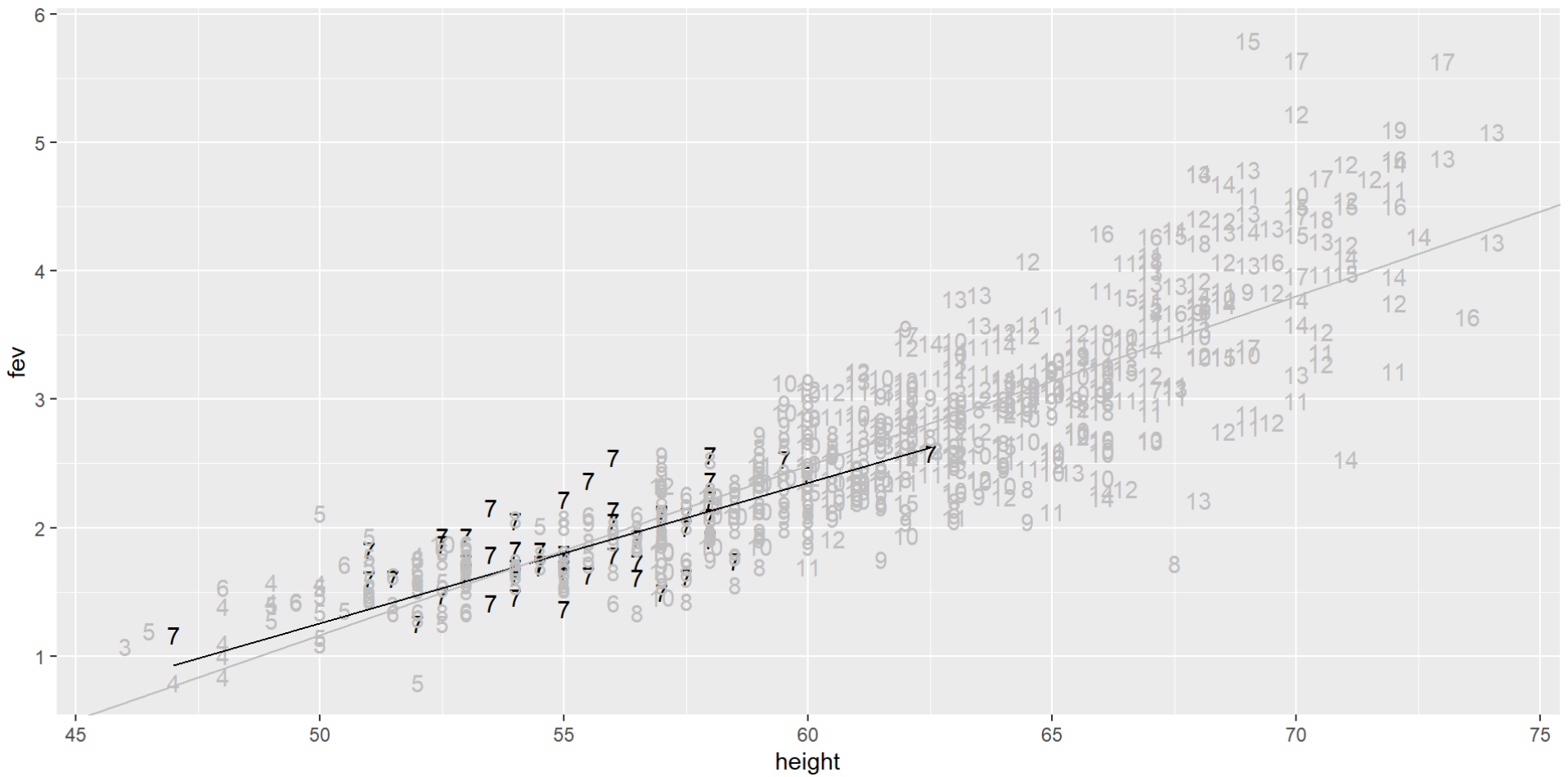
Relationship between height and FEV controlling at Age=5



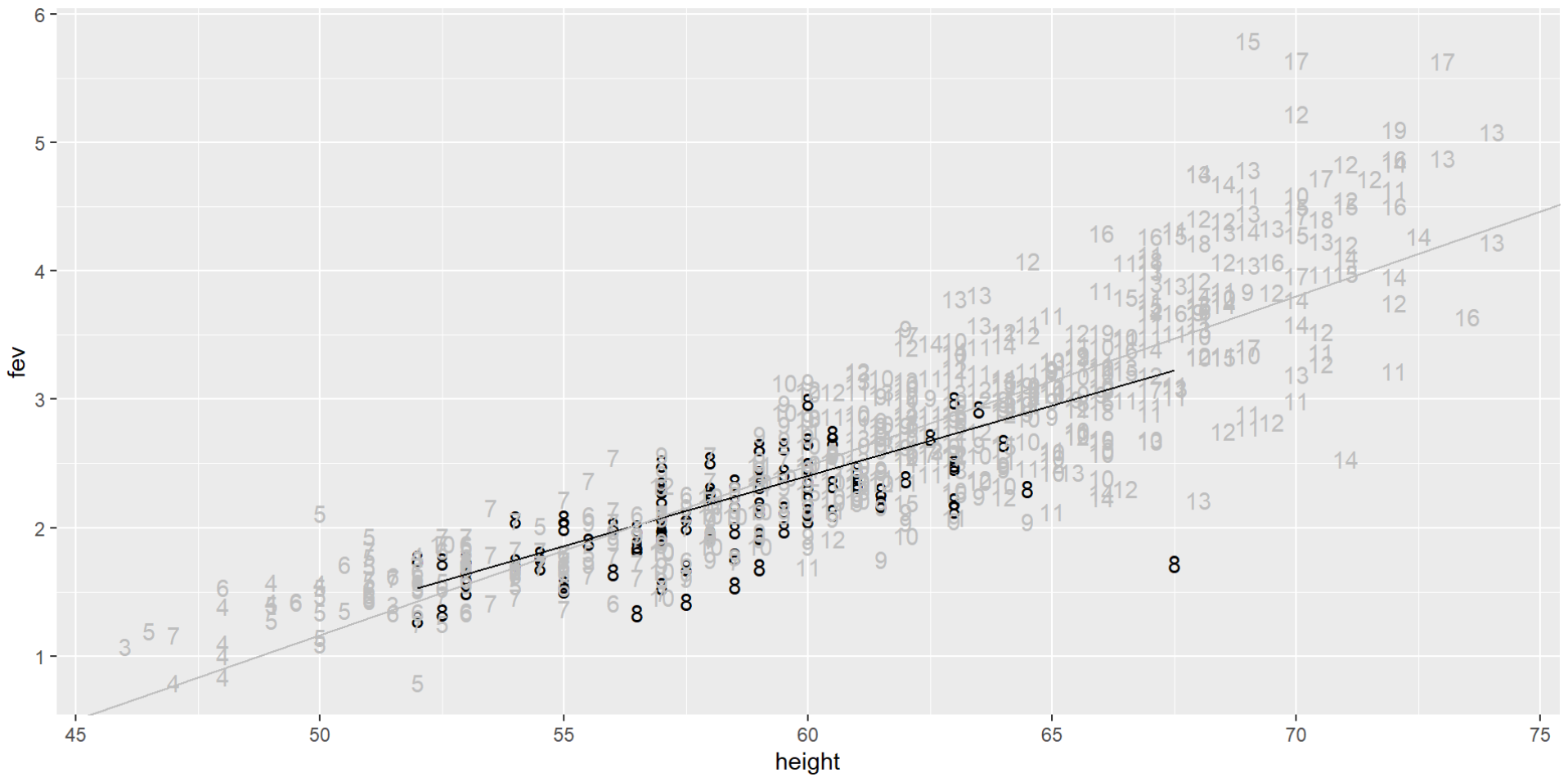
Relationship between height and FEV controlling at Age=6



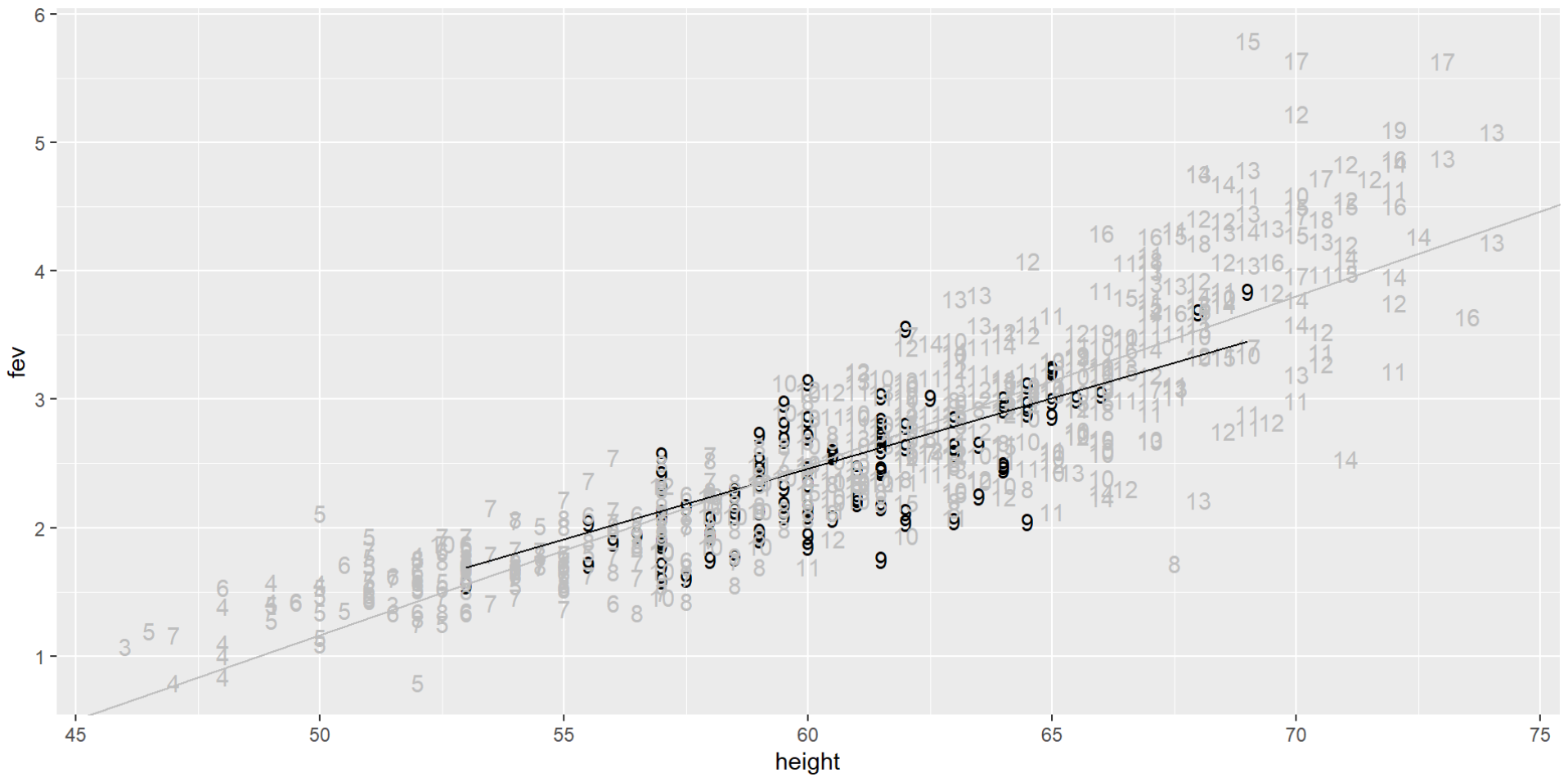
Relationship between height and FEV controlling at Age=7



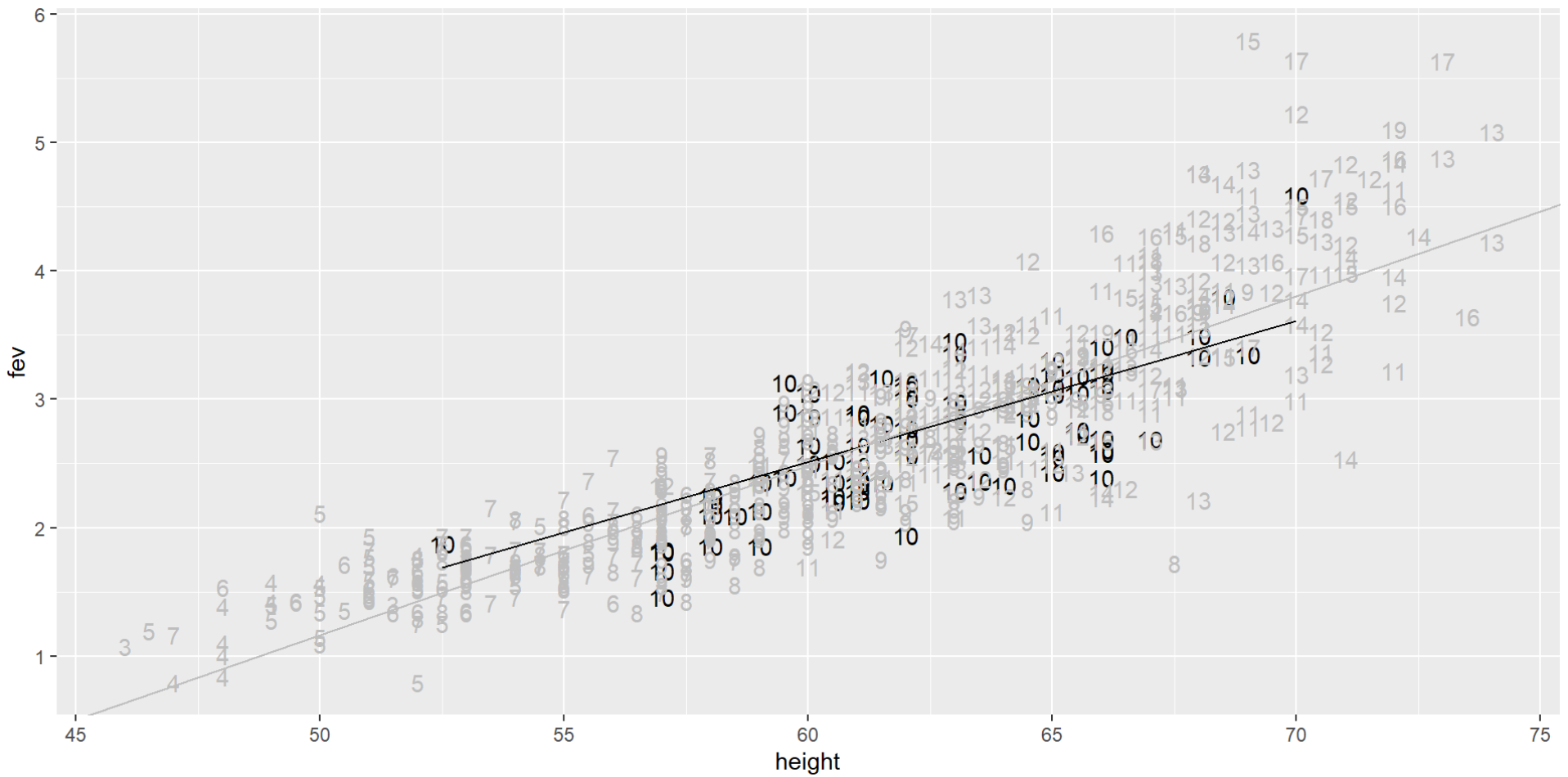
Relationship between height and FEV controlling at Age=8



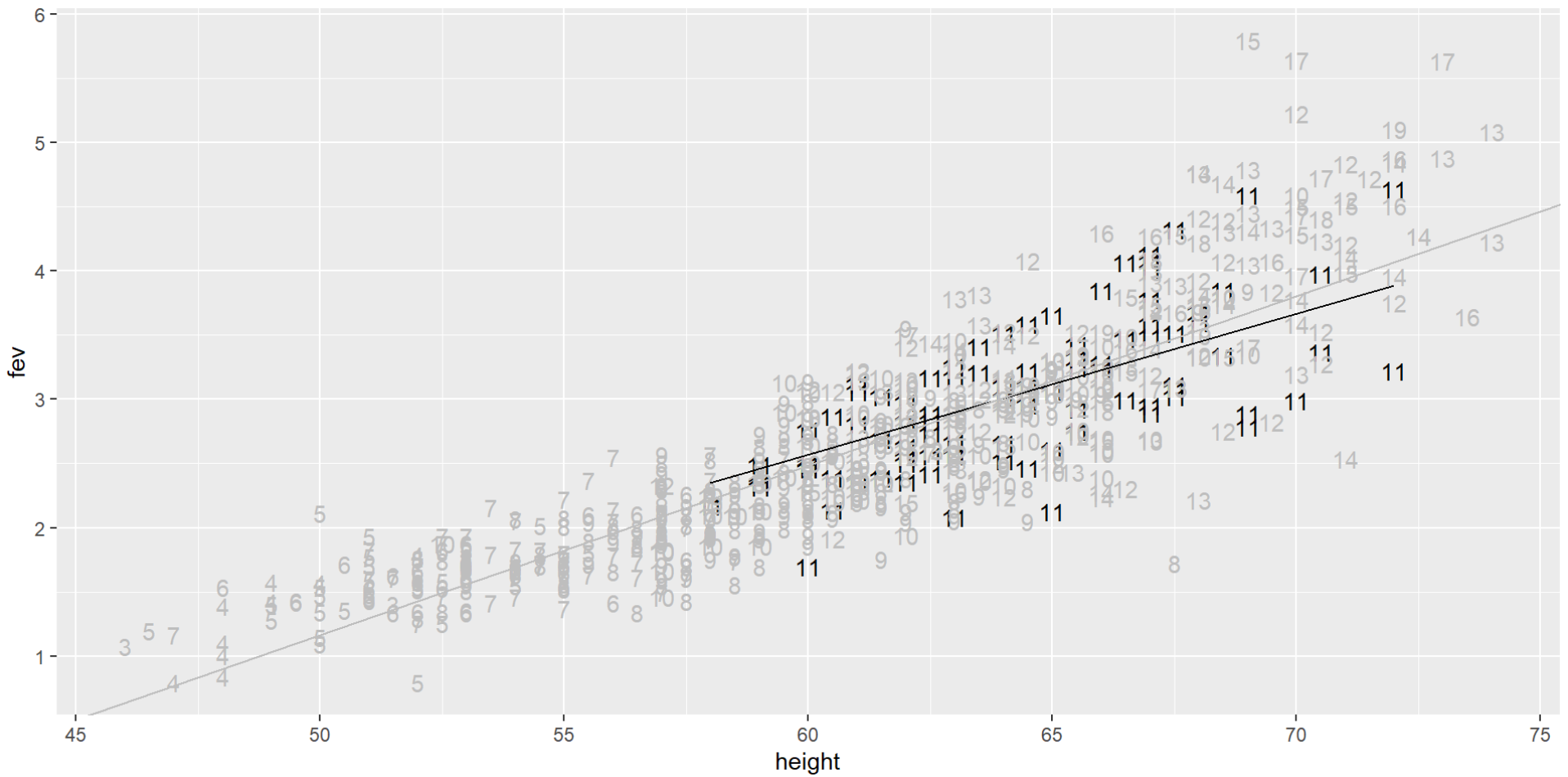
Relationship between height and FEV controlling at Age=9



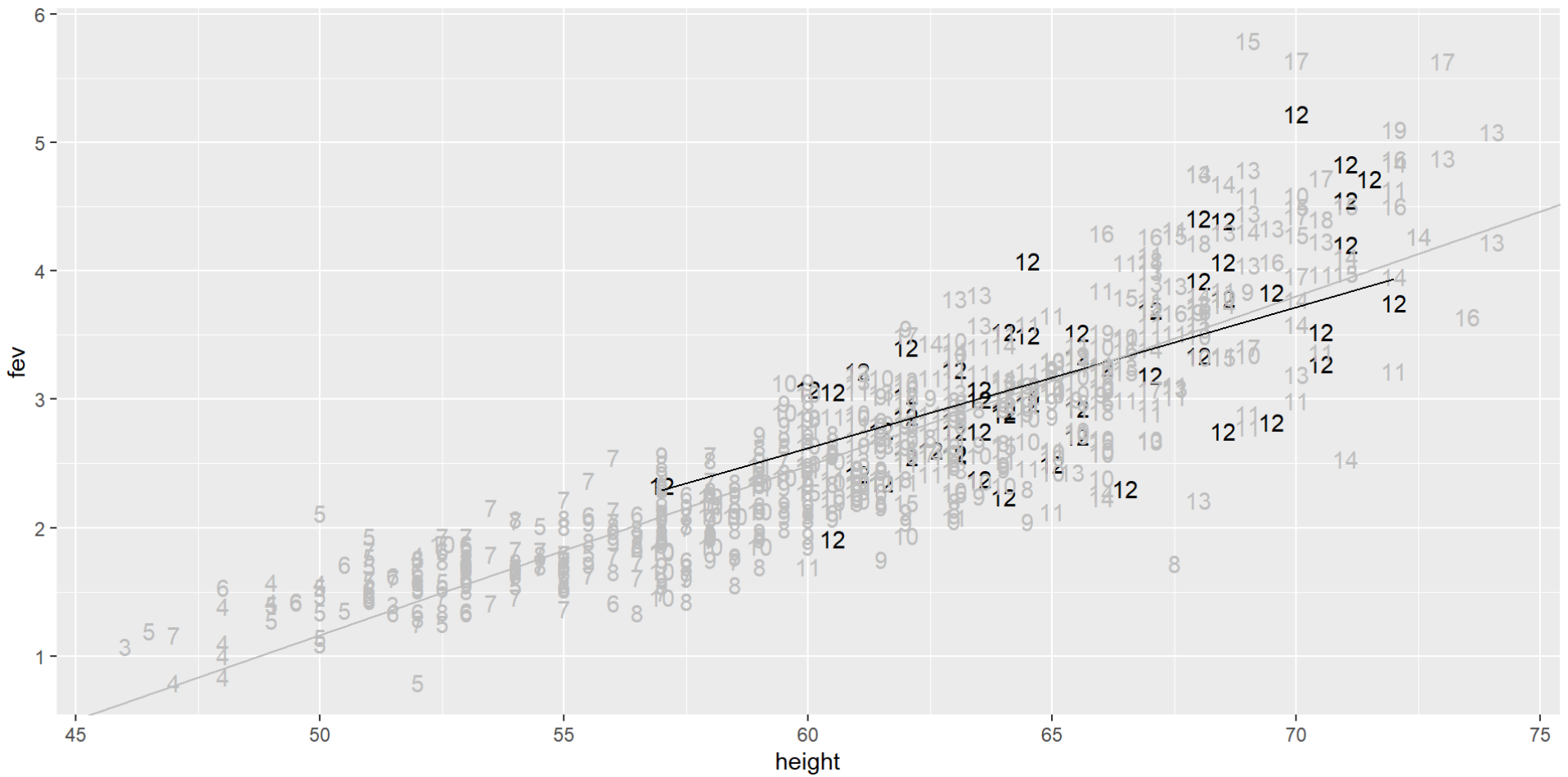
Relationship between height and FEV controlling at Age=10



Relationship between height and FEV controlling at Age=11

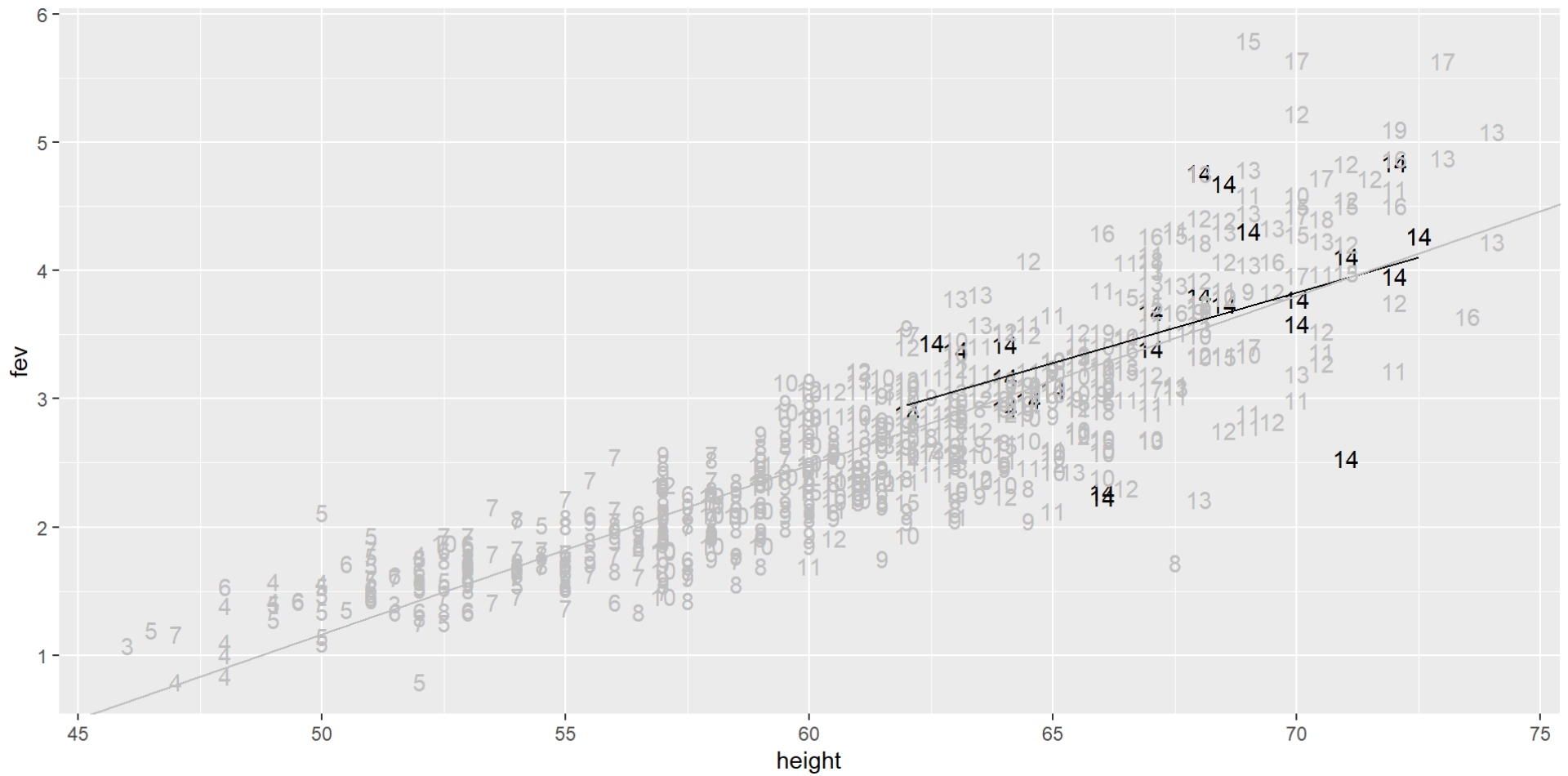


Relationship between height and FEV controlling at Age=12

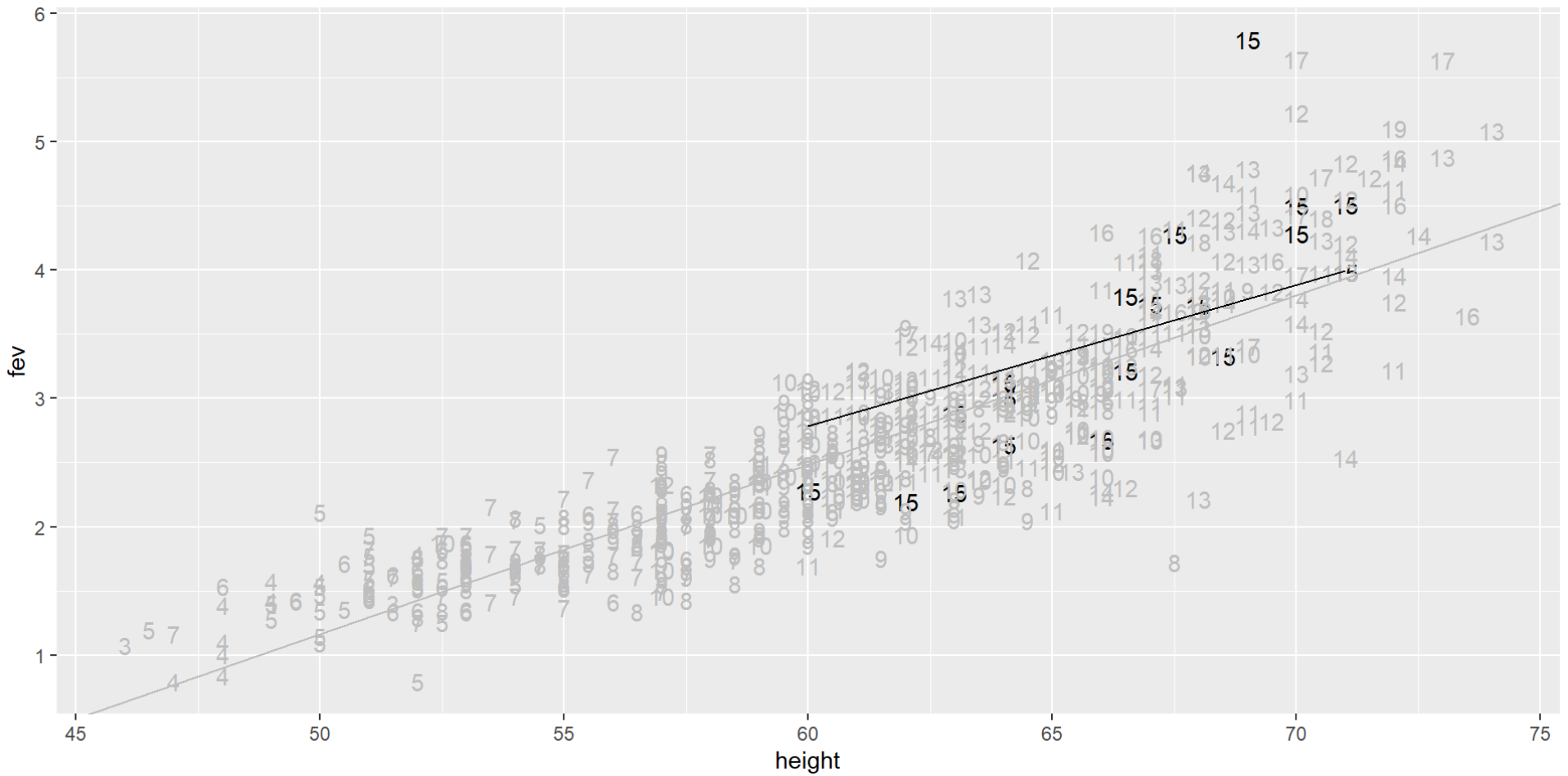


Relationship between height and FEV controlling at Age=13

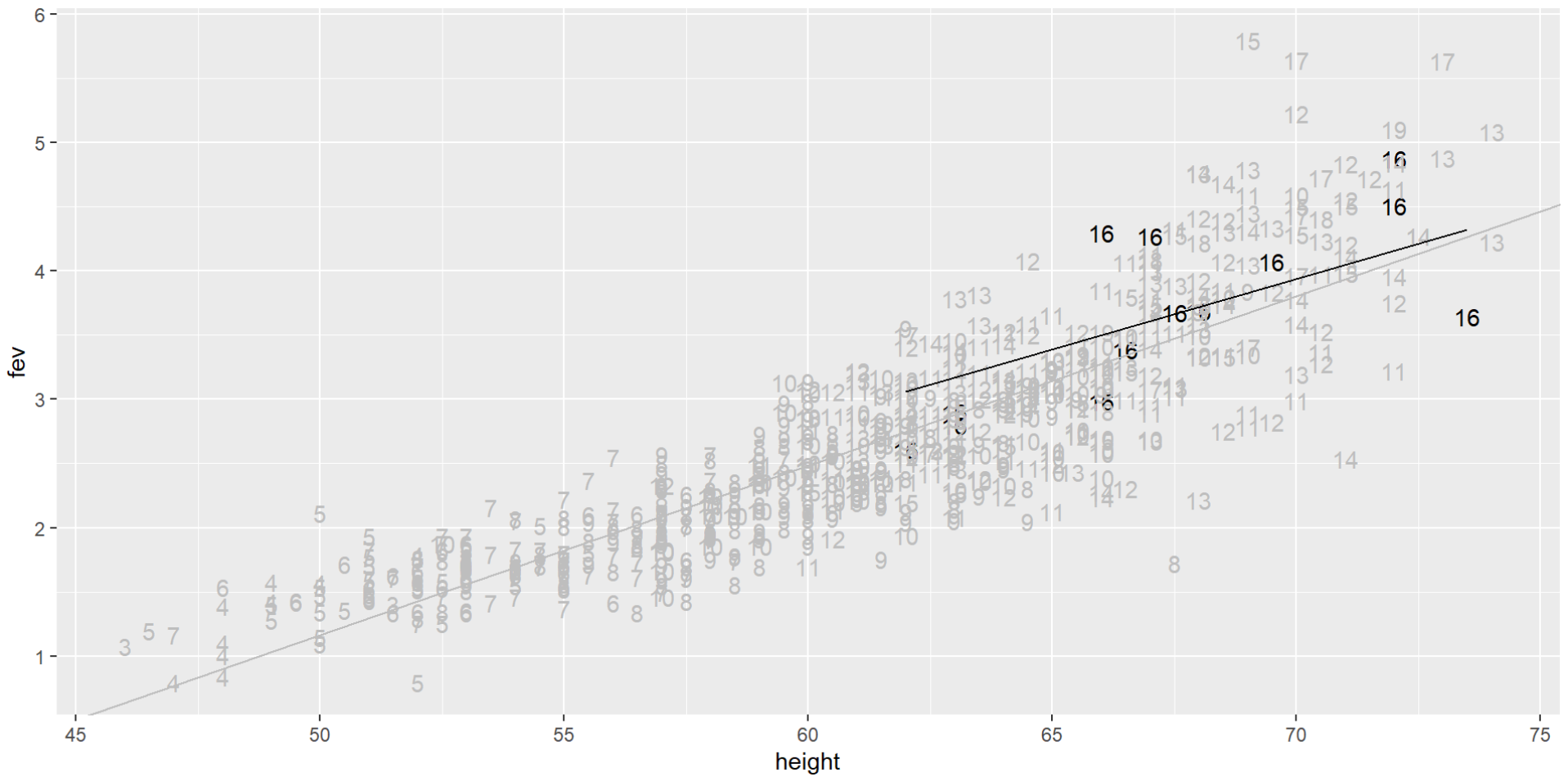
Relationship between height and FEV controlling at Age=14



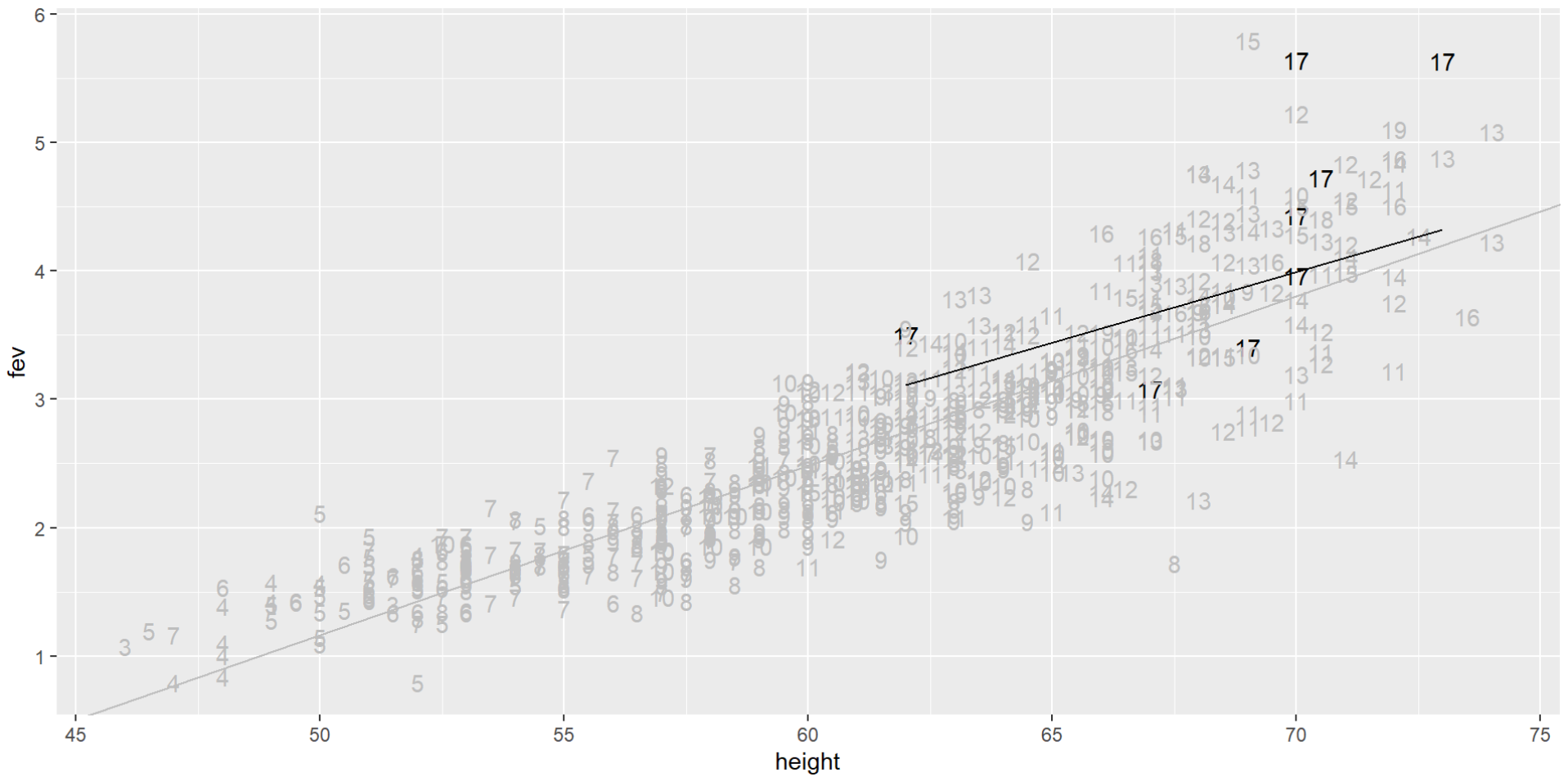
Relationship between height and FEV controlling at Age=15



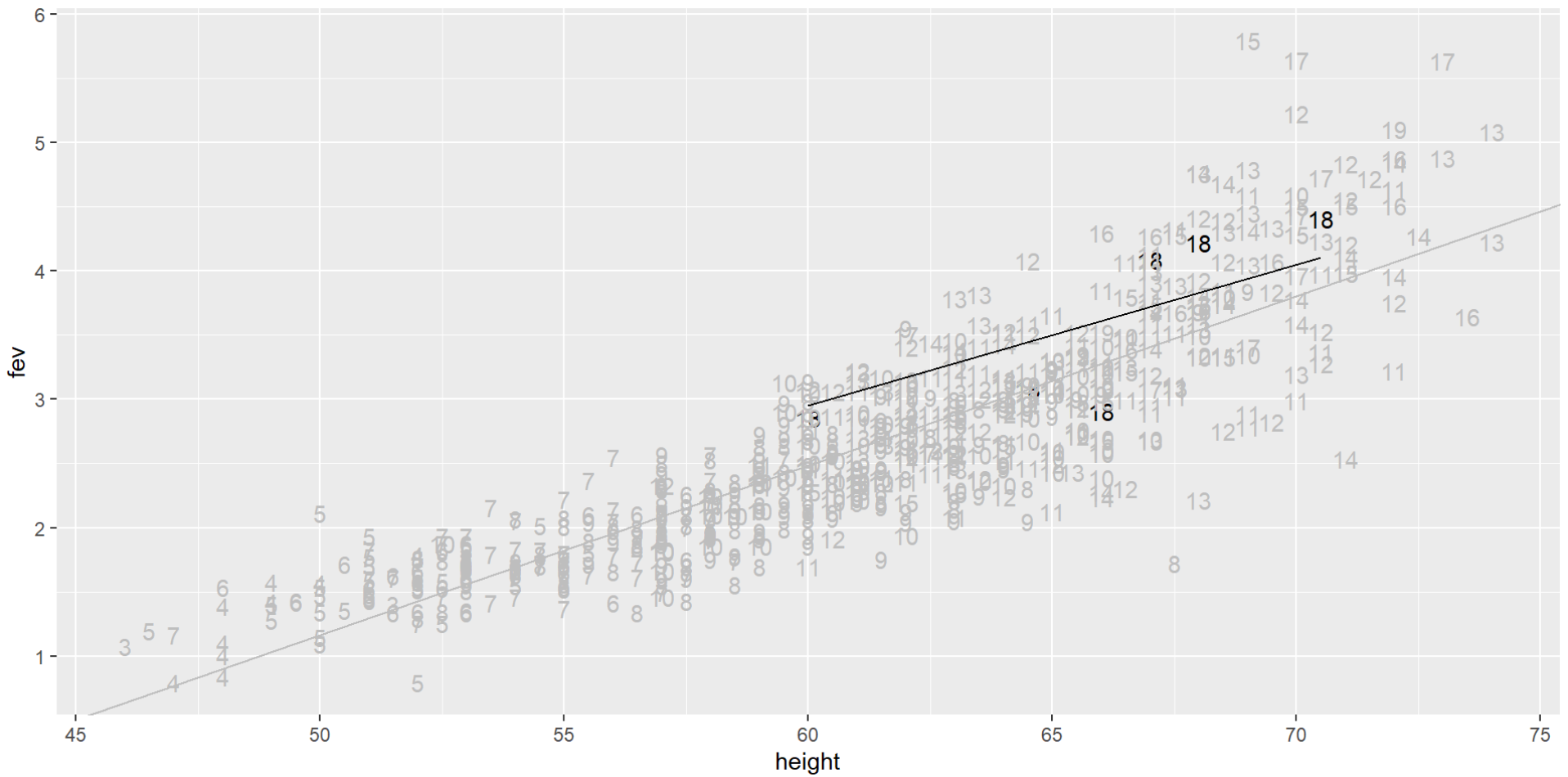
Relationship between height and FEV controlling at Age=16



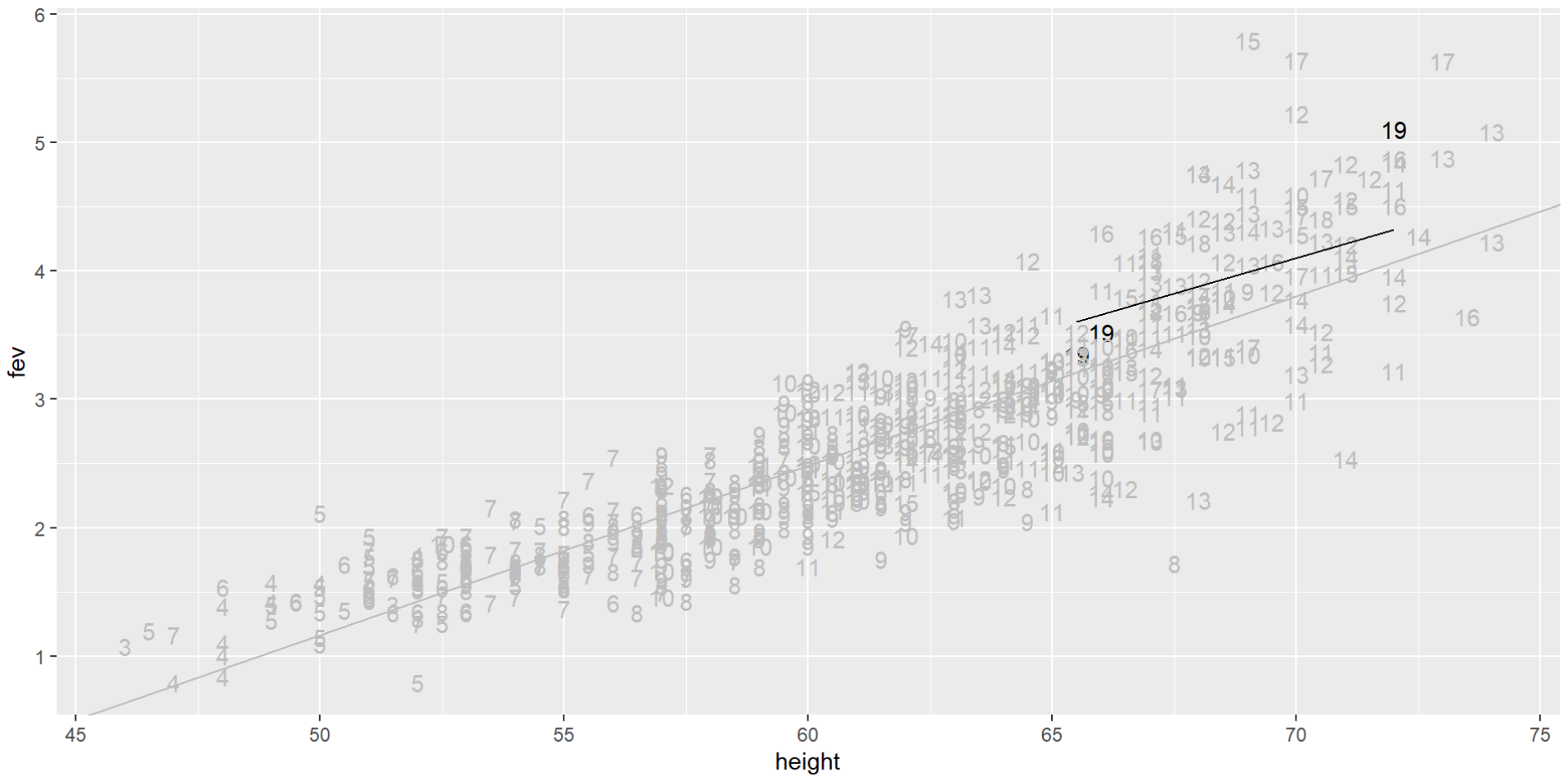
Relationship between height and FEV controlling at Age=17



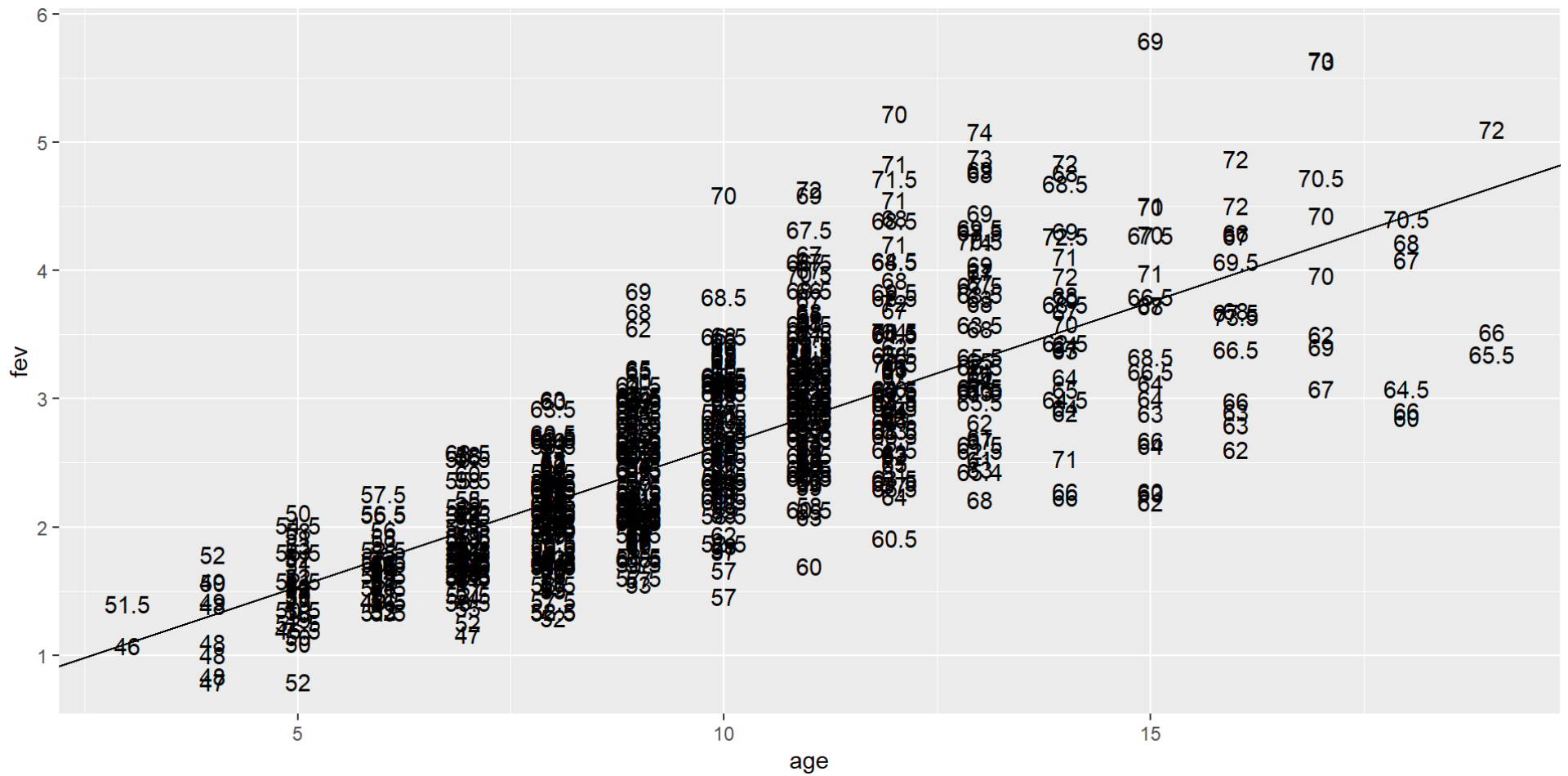
Relationship between height and FEV controlling at Age=18



Relationship between height and FEV controlling at Age=19

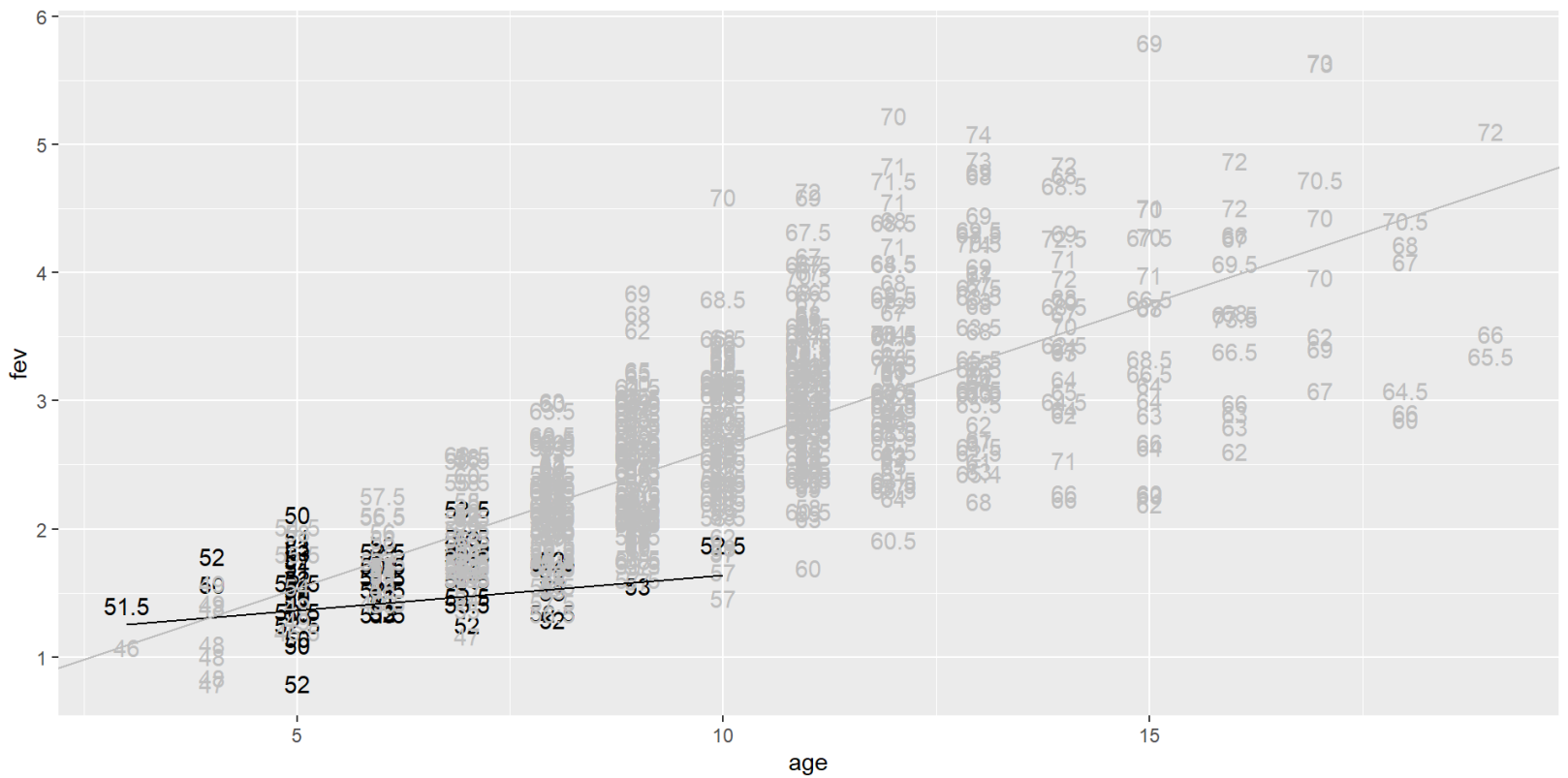


Unadjusted relationship between age and fev

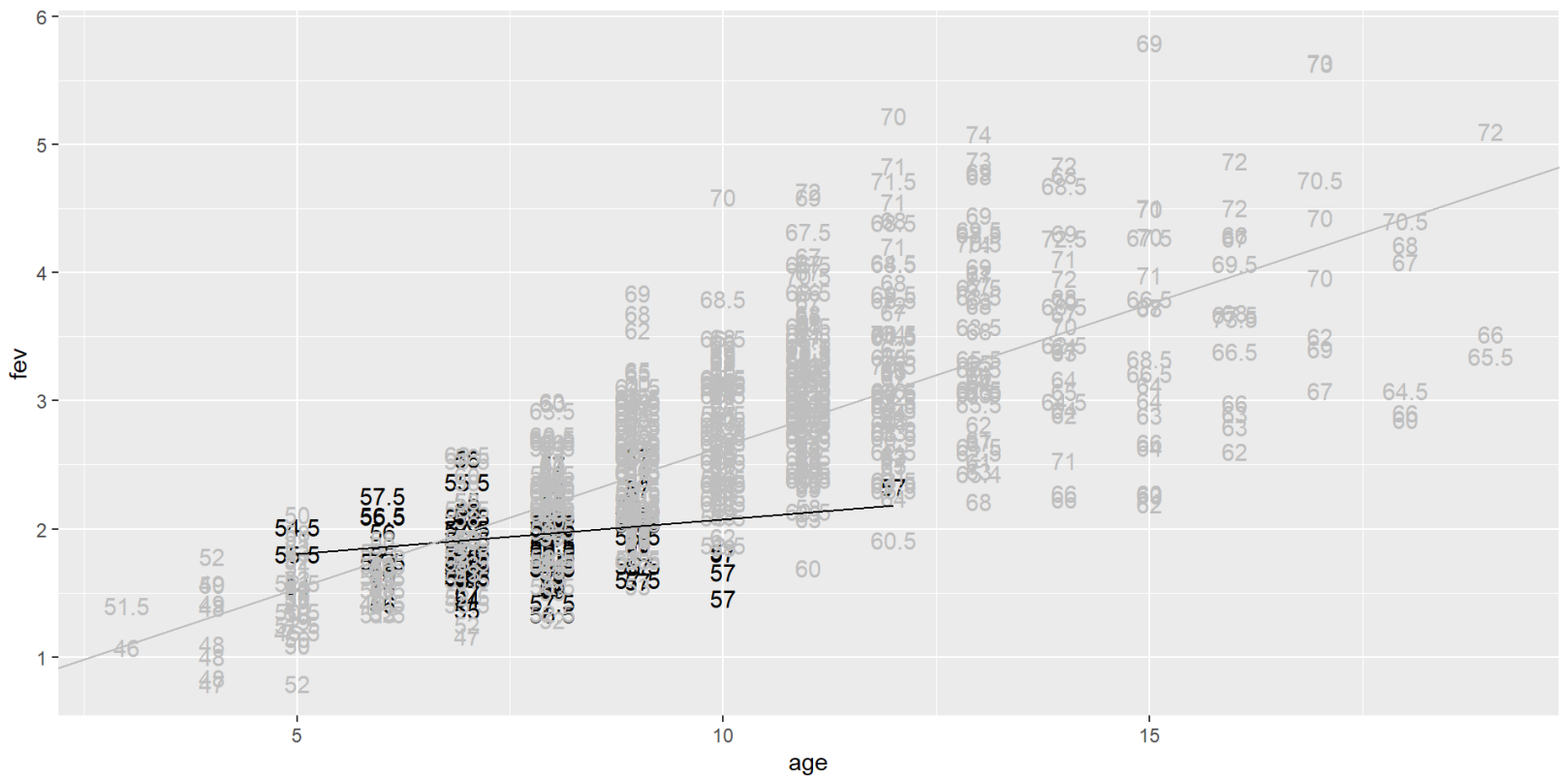


Relationship between age and FEV controlling for height between 46 and 49.5

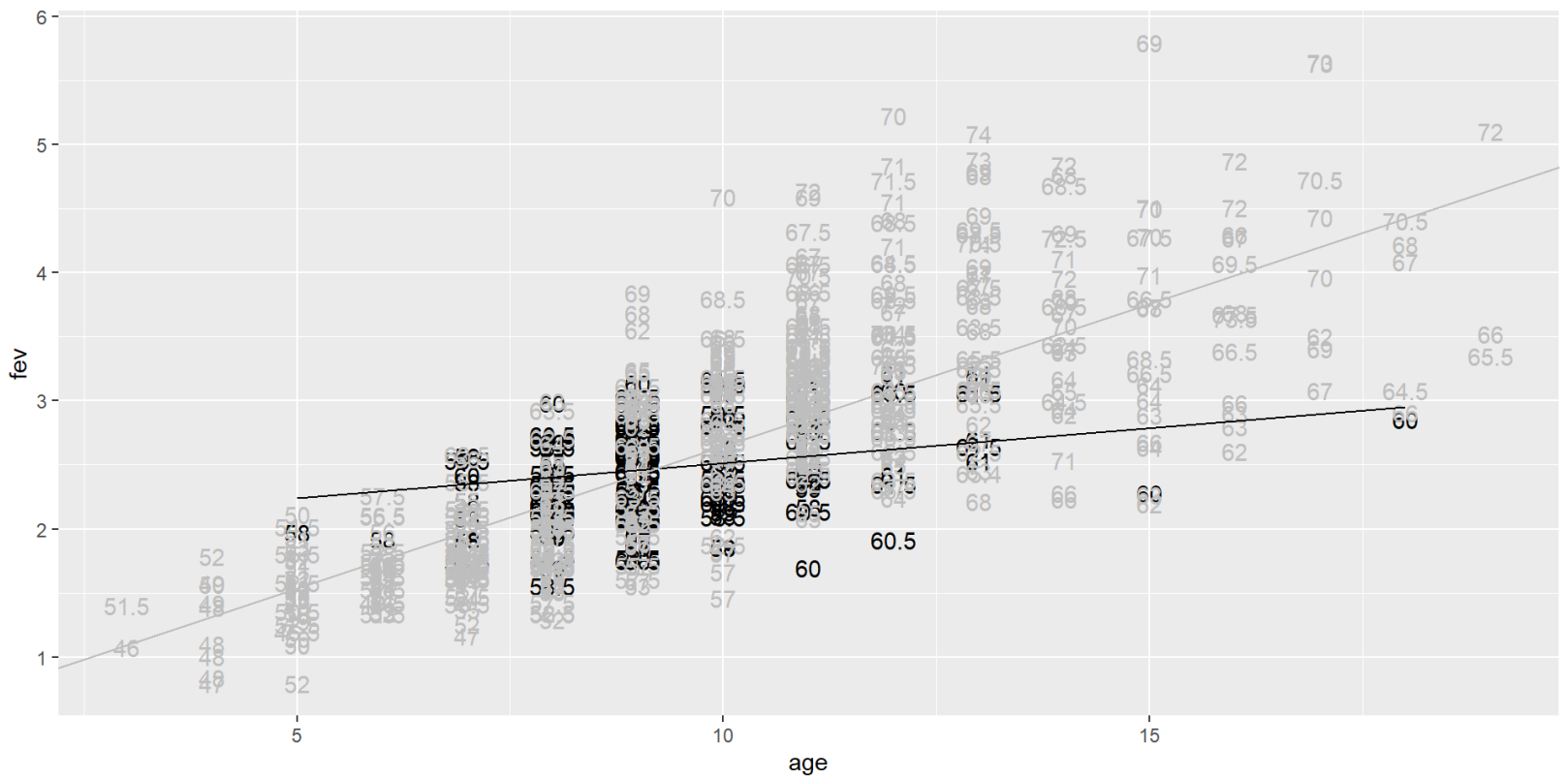
Relationship between age and FEV controlling for height between 50 and 53.5



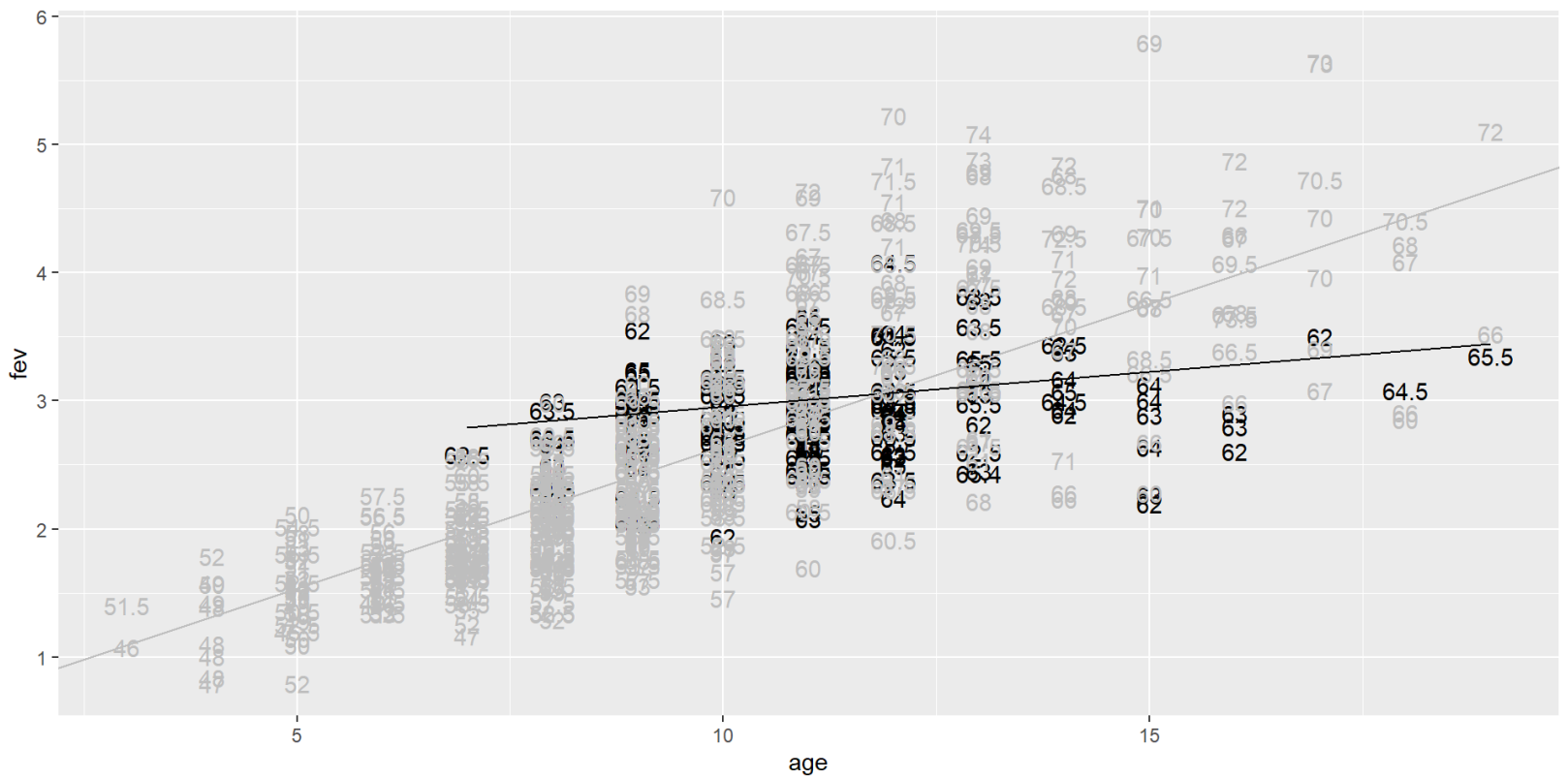
Relationship between age and FEV controlling for height between 54 and 57.5



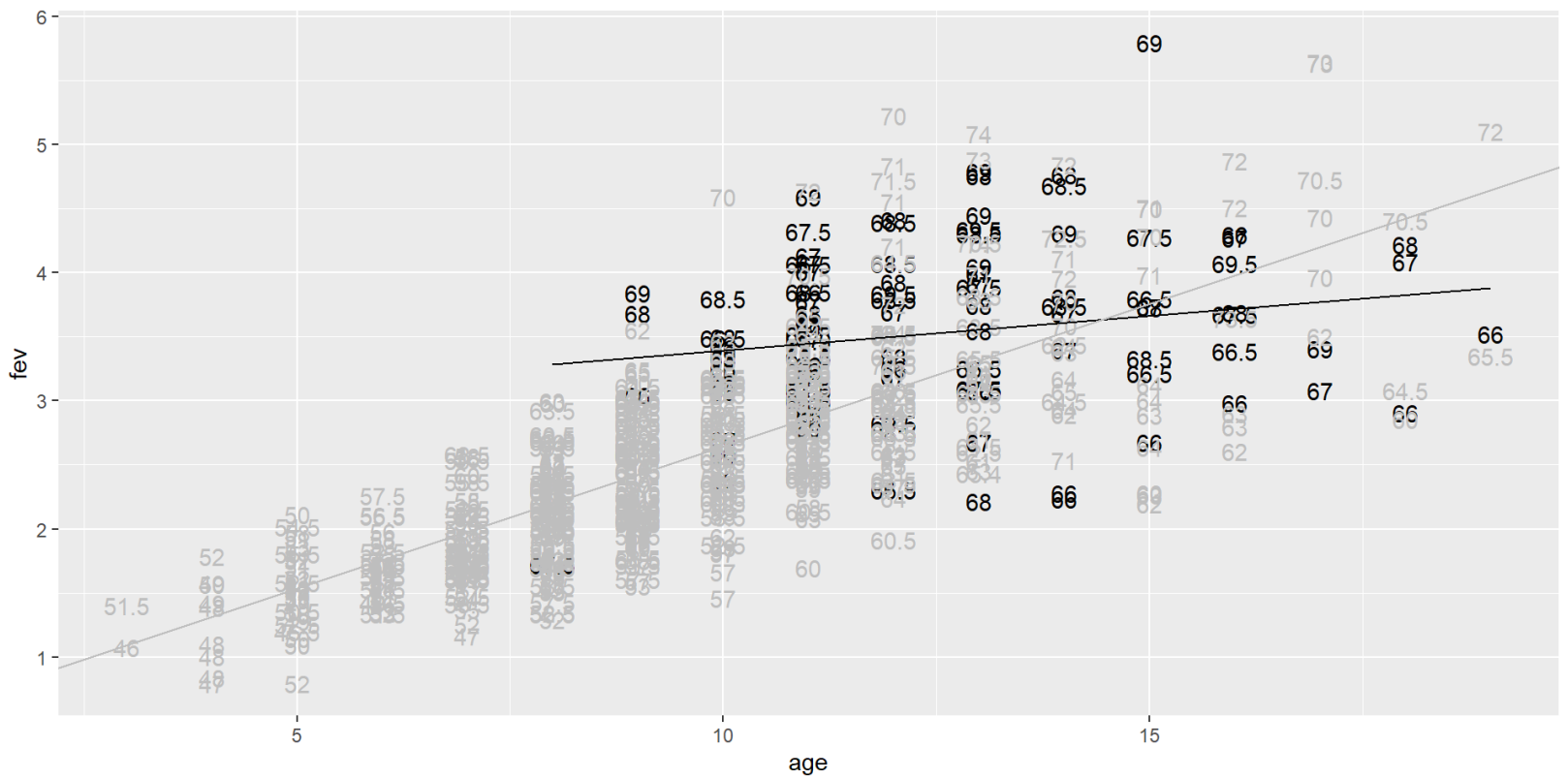
Relationship between age and FEV controlling for height between 58 and 61.5



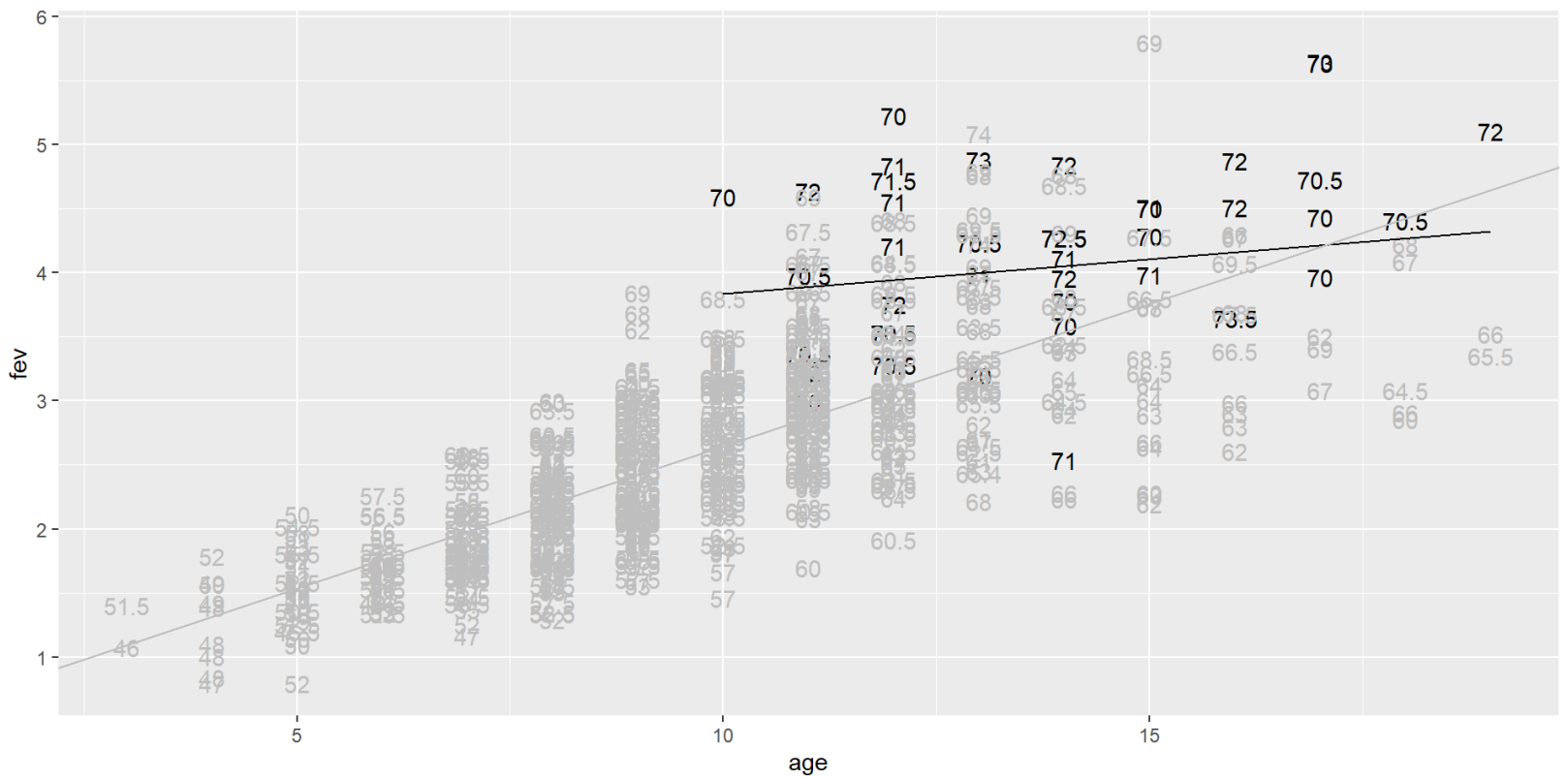
Relationship between age and FEV controlling for height between 62 and 65.5



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

Refer to part 2 of my code.

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.

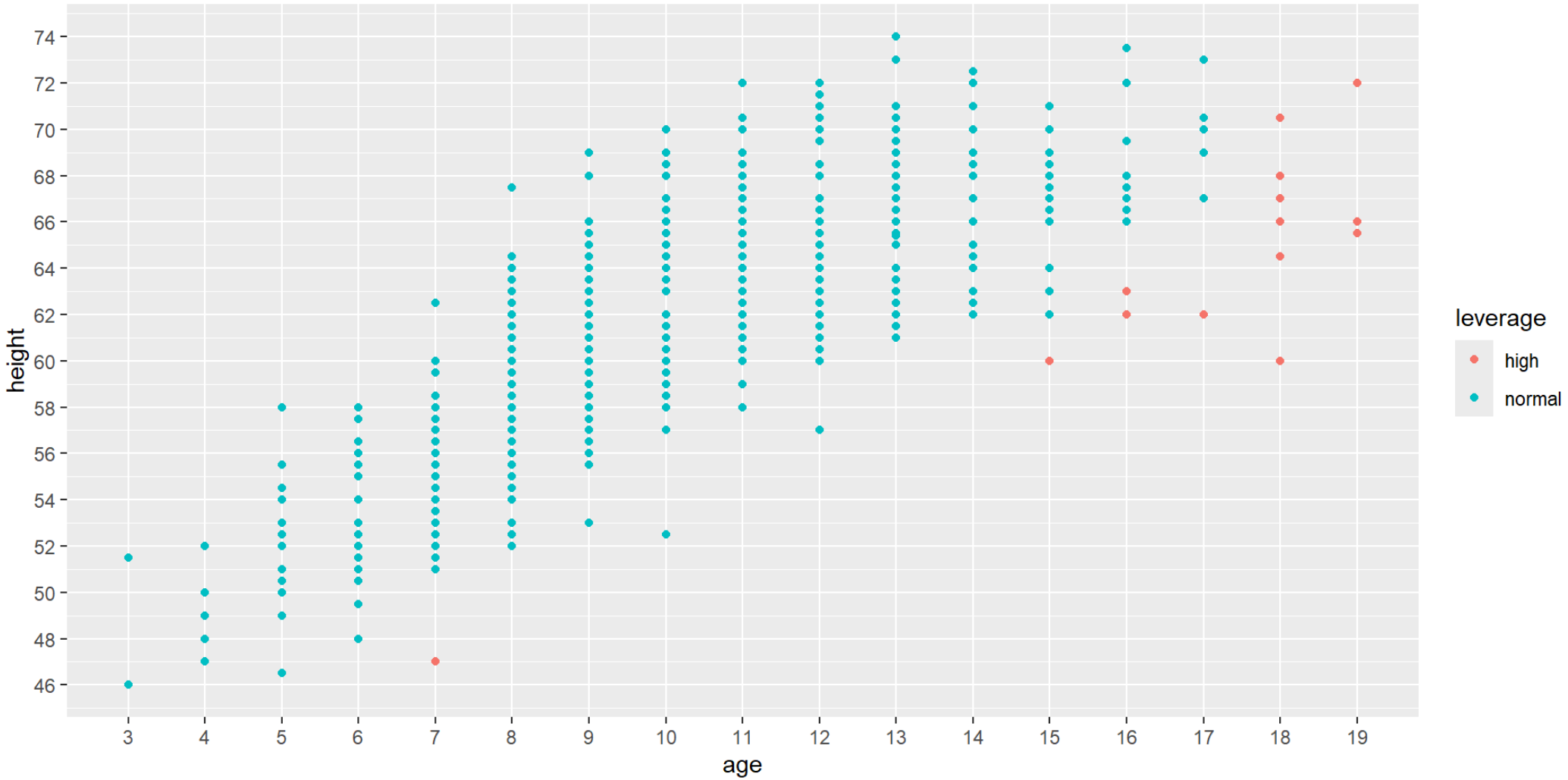
Leverage, 1

```
# 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
```

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.

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.

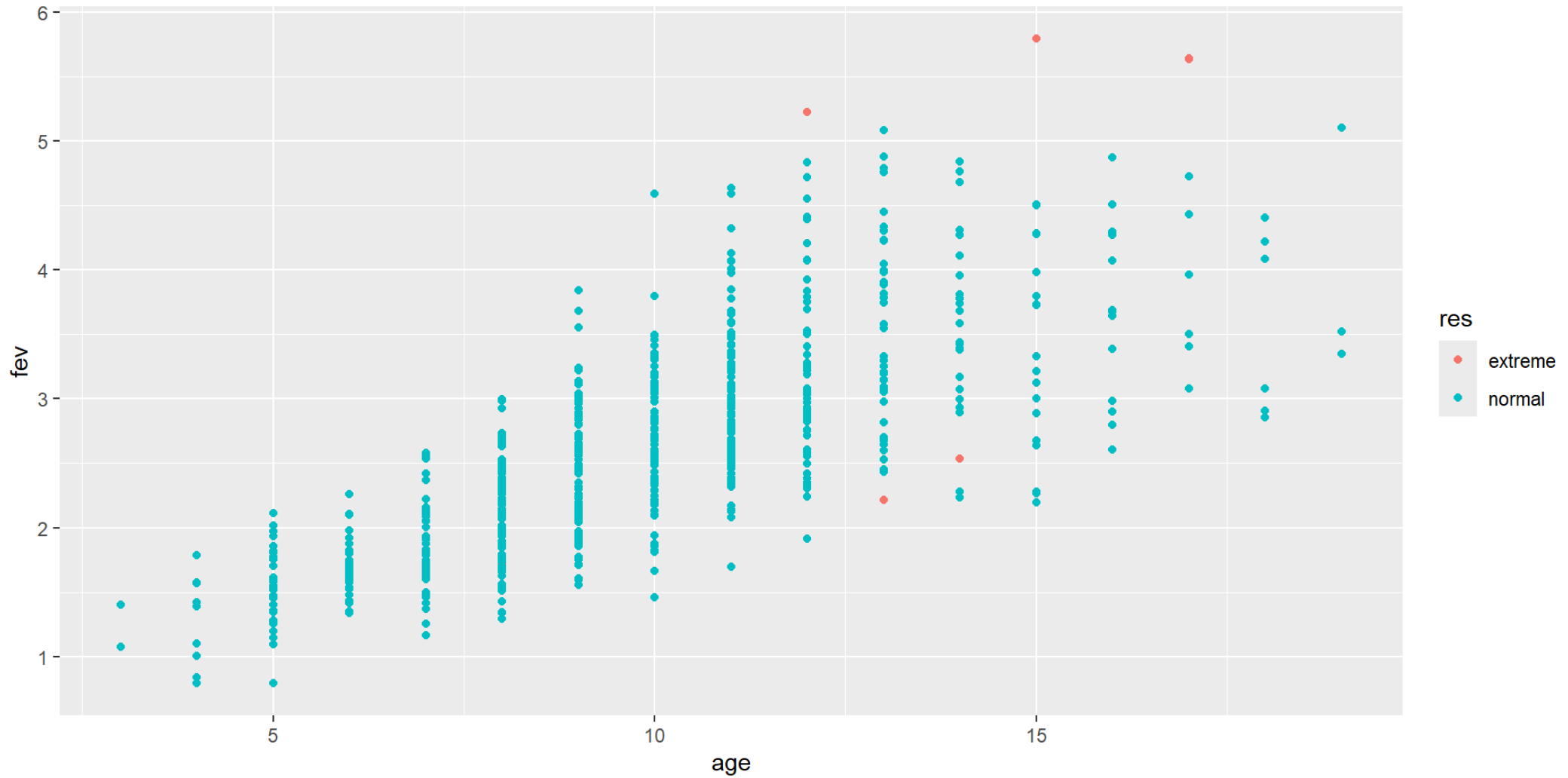
Studentized residuals, 1

```
# 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.

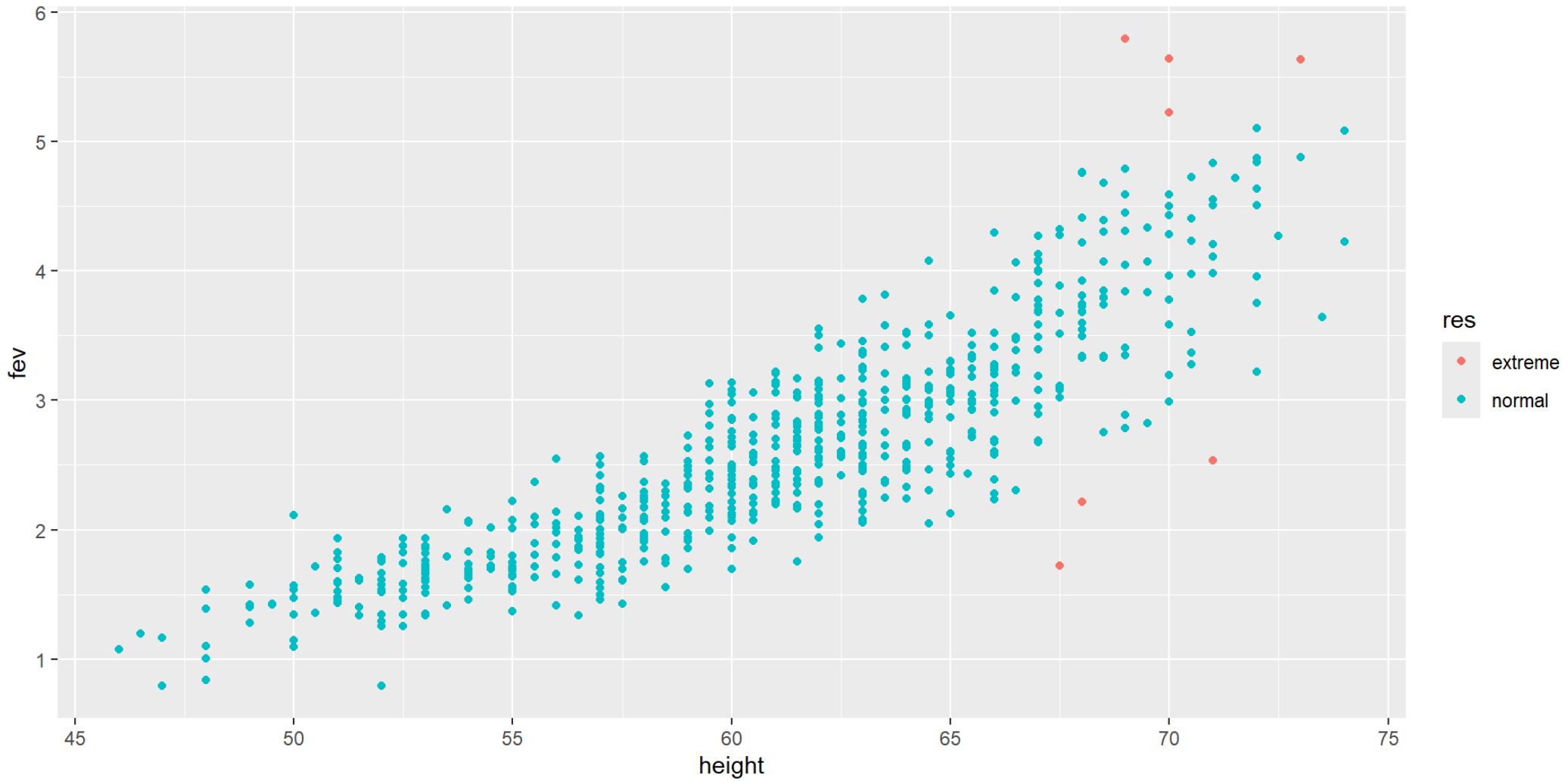
Studentized residuals, 2



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

Refer to part 3 of my code.

Break #6

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

simon-5501-07-directions.qmd

Refer to the programming assingment on my github site.

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