Comments for MEDB 5501, Week 14

Topics to be covered

- What you will learn
  - Interpretation of regression slope and intercept
  - Assumptions in linear regression
  - Calculation of sums of squares
  - Confidence intervals and hypothesis tests
  - Relationship to the correlation coefficient
  - Definition of residuals
  - Diagnostic plots
  - Interpretation with two independent variables

Bad joke, 1 of 4

Speaker notes

I borrowed this image from a movie poster. Does anyone know what movie this is?

https://en.wikipedia.org/wiki/Airplane!

So I am using this as a setting for a bad Statistics joke.

Two statisticians are on an airplane, flying from Miami to Seattle. Fiften minutes into the flight, they hear a loud …

Bad joke, 2 of 4

Speaker notes

… BANG! The pilot comes on the PA system and says “Excuse me, Ladies and Gentlemen. We’ve just had an engine explode. We’ll be just fine with three engines, but instead of a three hour flight, this will now be a four hour flight.”

The statisticians go back to talking, and fifteen minutes later, they hear another loud …

Bad joke, 3 of 4

Speaker notes

… BANG! The pilot comes back on and says, “I’m sorry to report that we’ve had another engine explode. We can still make it to Seattle, but it will now be a six hour flight. I apologize for the additional delay.”

The statisticians shrug and start talking again when fifteen minutes later, you guessed it, they hear a third loud …

Bad joke, 4 of 4

Speaker notes

… BANG! The pilot comes on again and says “I’m sorry to report that a third engine has exploded. Each engine on this jet is very powerful and we can still make it to Seattle, but it is now going to be a nine hour flight.”

At this point, one statistician says to the other, “Boy, I hope this last engine doesn’t fail …”

“… or we’ll be up here forever!”

This is an example of a dangerous extrapolation. The experience with three, two, and then only one engines may be consistent, but don’t expect that trend to continue with zero engines.

Algebra formula for a straight line

- $Y = mx + b$
- $m = \Delta y / \Delta x$
- m = slope
- b = y-intercept

Speaker notes

One formula in algebra that most people can recall is the formula for a straight line. Actually, there are several different formulas, but the one that most people cite is

Y = m X + b

where m represents the slope, and b represents the y-intercept (we'll call it just the intercept here). They can also sometimes remember the formula for the slope:

$$m = \increment y / \increment x$$

In English, we would say that this is the change in y divided by the change in x.

Linear regression interpretation of a straight line

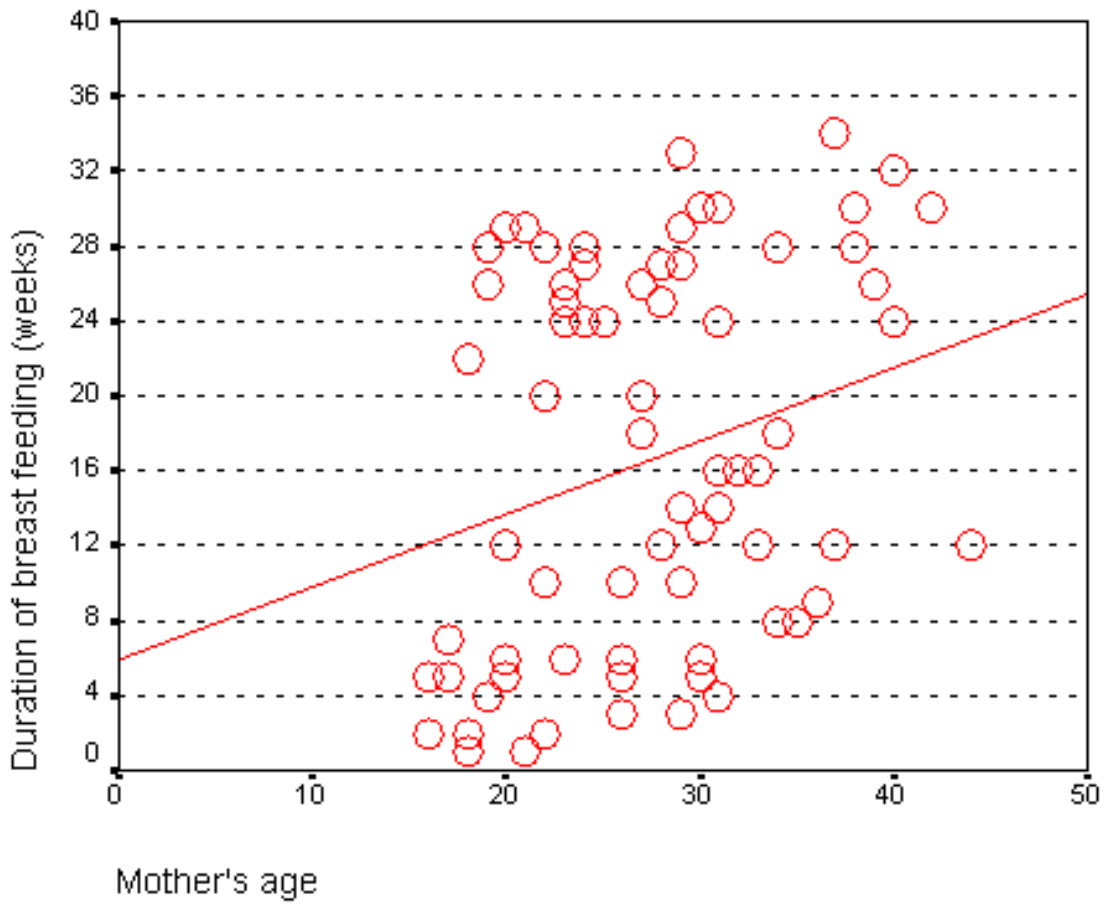
- The slope represents the estimated average change in Y when X increases by one unit.
- The intercept represents the estimated average value of Y when X equals zero.

Speaker notes

In linear regression, we use a straight linear to estimate a trend in data. We can’t always draw a straight line that passes through every data point, but we can find a line that “comes close” to most of the data. This line is an estimate, and we interpret the slope and the intercept of this line as follows:

Be cautious with your interpretation of the intercept. Sometimes the value X=0 is impossible, implausible, or represents a dangerous extrapolation outside the range of the data.

First regression example with interpretation



Speaker notes

The graph shown below represents the relationship between mother’s age and the duration of breast feeding in a research study on breast feeding in pre-term infants.

The regression coefficients are shown below. The intercept, 6, is represented the estimated average duration of breast feeding for a mother that is zero years old. This is an impossible value, so the interpretation is not useful. What is useful, is the interpretation of the slope, approximately 0.4. The estimated average duration of breast feeding increases by 0.4 weeks for every extra year in the mother’s age.

Output from SPSS

Parameter Estimates

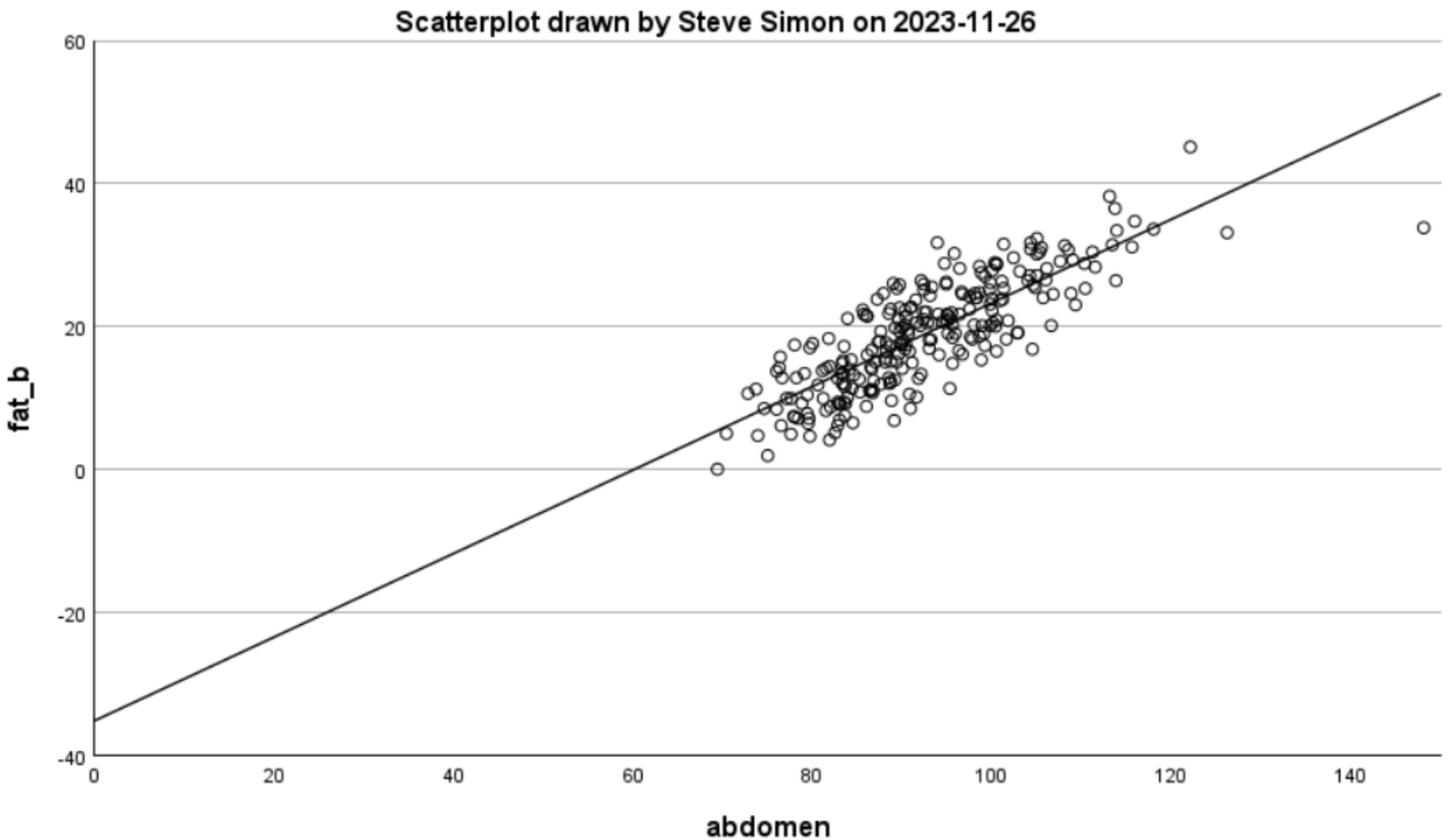
Dependent Variable: Duration of breast feeding (weeks)

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	5.920	4.580	1.292	.200	-3.195	15.035
MOM_AGE	.389	.162	2.399	.019	6.626E-02	.712

Speaker notes

Here is what the output from SPSS looks like.

Scatterplot of fat percentage and abdomen circumference



Speaker notes

Here’s a second example of linear regression. You’ve seen this dataset before, relating percentage of body fat to various circumference measures of the human body. Measuring body fat accurately is a difficult task. The best method involves dunking the entire body in a vat of water. Circumferences, on the other hand, are easy to measure. How well can you predict percentage body fat from these circumference measures.

You found out last week that abdomen circumference was the measure most strongly correlated with body fat. Here is a scatterplot showing the linear regression trend line.

Notice that the trend line goes negative for smaller values of body circumference. This represents a dangerous extrapolation, an extrapolation beyond the range of the data.

Correlating fat percentage with abdomen circumference, 1 of 3

Descriptive Statistics

	Mean	Std. Deviation	N
fat_b	18.938	7.7509	252
abdomen	92.556	10.7831	252

Speaker notes

Although it is not necessary to compute a correlation before running a linear regression, it does have some value in understanding what is going on.

The bivariate correlation dialog box in SPSS offers an optional table of descriptive statistics.

Correlating fat percentage with abdomen circumference, 2 of 3

Correlations

		fat_b	abdomen
fat_b	Pearson Correlation	1	.814
	Sig. (2-tailed)		<.001
	N	252	252
abdomen	Pearson Correlation	.814	1
	Sig. (2-tailed)	<.001	
	N	252	252

Speaker notes

The correlation shows a strong positive association between body fat and abdominal circumference.

Correlating fat percentage with abdomen circumference, 3 of 3

Confidence Intervals

	Pearson Correlation	Sig. (2-tailed)	95% Confidence Intervals (2-tailed)[a]	
			Lower	Upper
fat_b - abdomen	.814	<.001	.767	.852

a. Estimation is based on Fisher's r-to-z transformation.

Speaker notes

You can also get an optional confidence interval. This interval is narrow, mostly because of the large sample size and the strong relationship between the two variables.

R Square measure in linear regression

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.814[a]	.662	.661	4.5144

a. Predictors: (Constant), abdomen

b. Dependent Variable: fat_b

Speaker notes

By default, SPSS gives you a measure of R Square. The other statistics, R and adjusted R Square, are only useful for regression models with more than one independent variable.

ANOVA table in linear regression

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9984.086	1	9984.086	489.903	<.001[b]
	Residual	5094.931	250	20.380		
	Total	15079.017	251			

a. Dependent Variable: fat_b

b. Predictors: (Constant), abdomen

Speaker notes

Here is the ANOVA table. The total sum of squares is split unevenly with about 2/3 going to regression and only 1/3 going to error (residual). If the F-ratio is close to one, you should accept the null hypothesis that the regression slope is flat (equals zero). This is anything but the case here. The F-ratio is large, so you should definitely reject the null hypothesis. The small p-value also indicates that you should reject the null hypothesis.

Regression coefficients

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-35.197	2.462		-14.294	<.001
	abdomen	.585	.026	.814	22.134	<.001

a. Dependent Variable: fat_b

Speaker notes

The slope coefficient is 0.585 (round this to 0.6). This means that the estimated average body fat percentage increases by 0.6% for each one centimeter increase in abdomen circumference. The t-ratio is far from zero, so you should reject the null hypothesis that the population slope parameter equals zero. The small p-value indicates the same conclusion.

Do not try to interpret the intercept. It would be the estimated average body fat percentage in a person with a no abdomen (an abdomen circumference of zero).

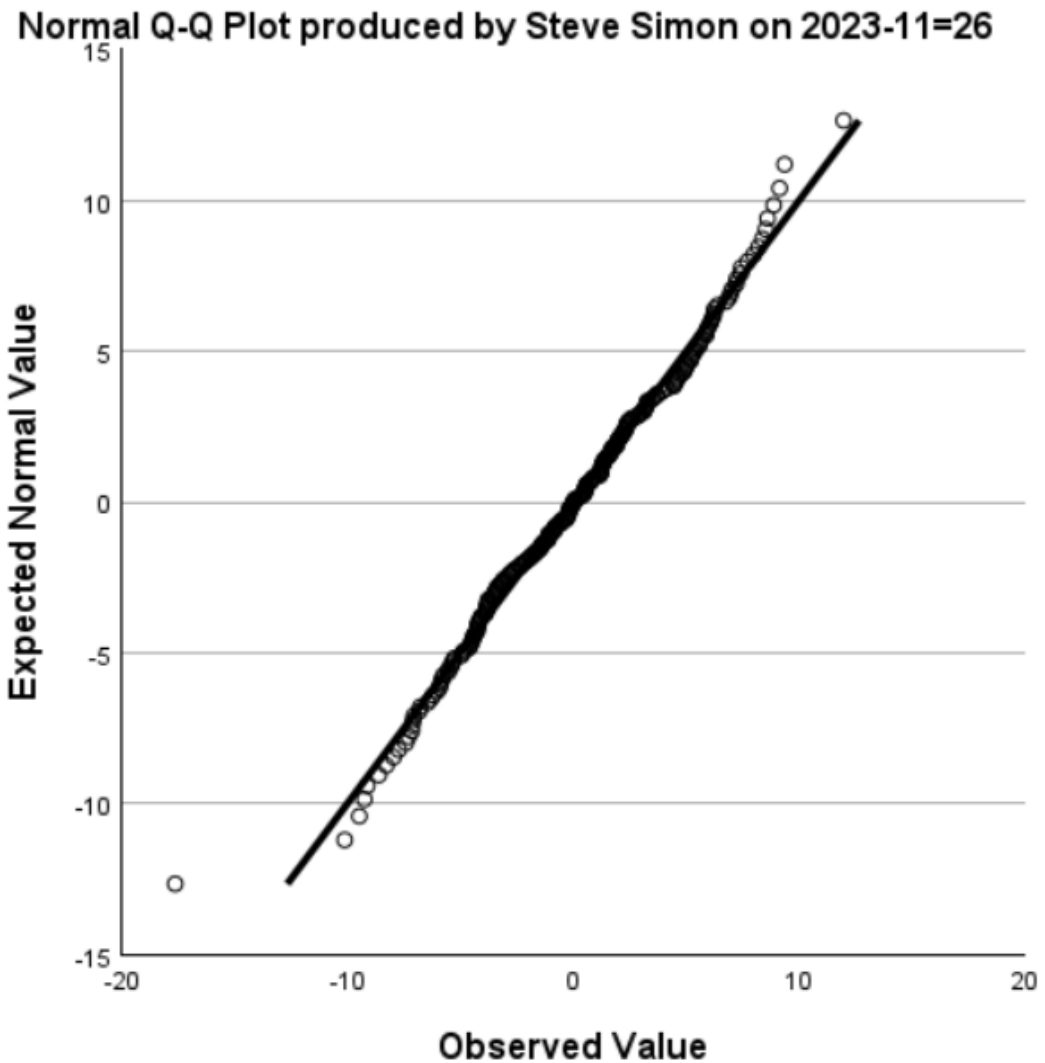
Predicting fat percentage from abdomen circumference (8/10)

95.0% Confidence Interval for B	
Lower Bound	Upper Bound
-40.046	-30.347
.533	.637

Speaker notes

The SPSS table is very wide, so I cut off and displayed the confidence intervals separately. You are 95% confident that the population slope lies between 0.53 and 0.64.

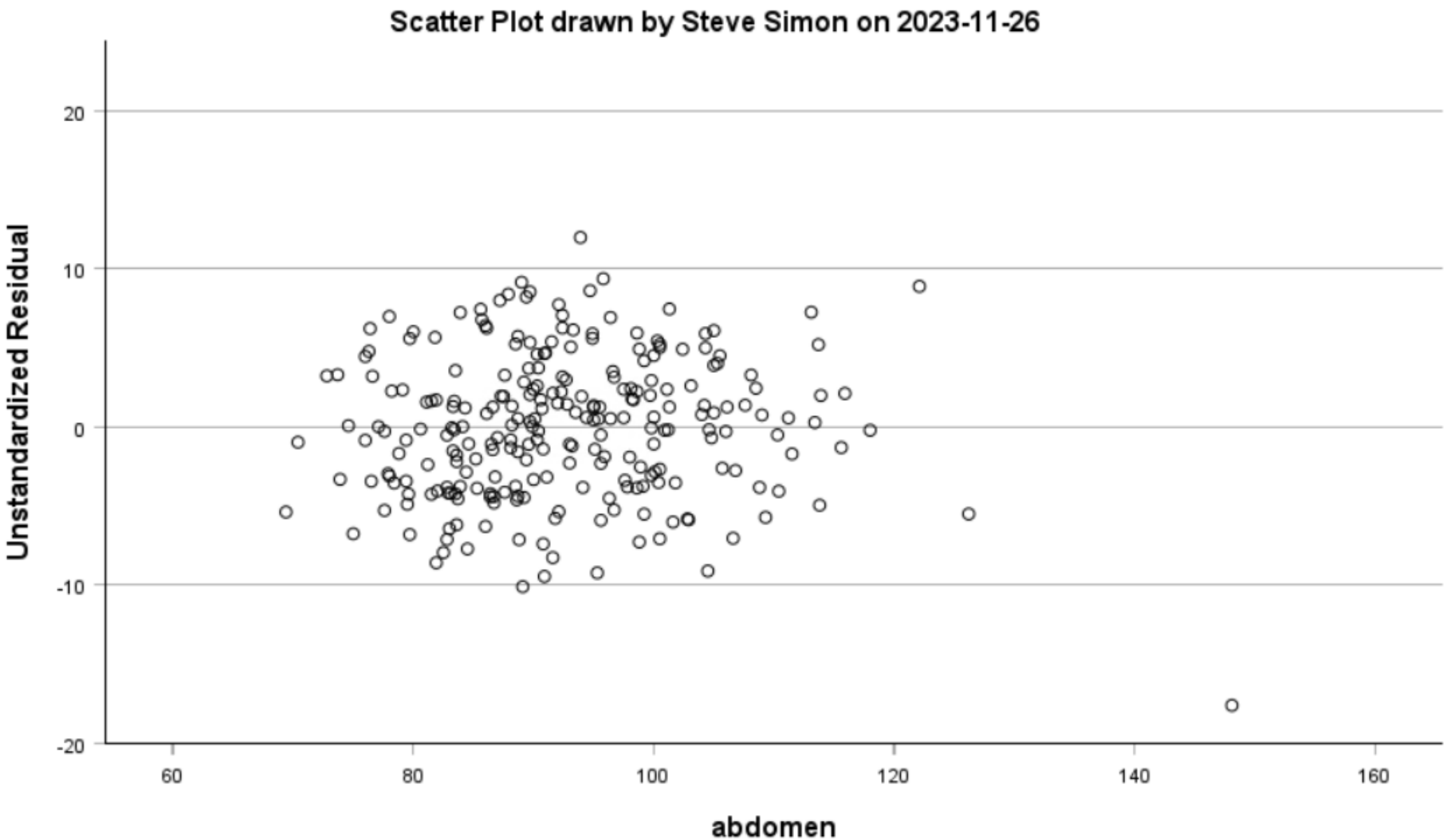
QQ plot of residuals



Speaker notes

The Q-Q plot of the residuals shows strong evidence of normality. There is, perhaps, an outlier on the low end, but this is not a serious issue.

Scatterplot of residuals



Speaker notes

The plot of residuals versus abdomen circumference shows no evidence of non-linearity and no evidence of heteroscedasticity.

Break #1

- What you have learned
  - Interpretation of regression slope and intercept
- What’s coming next
  - Assumptions in linear regression

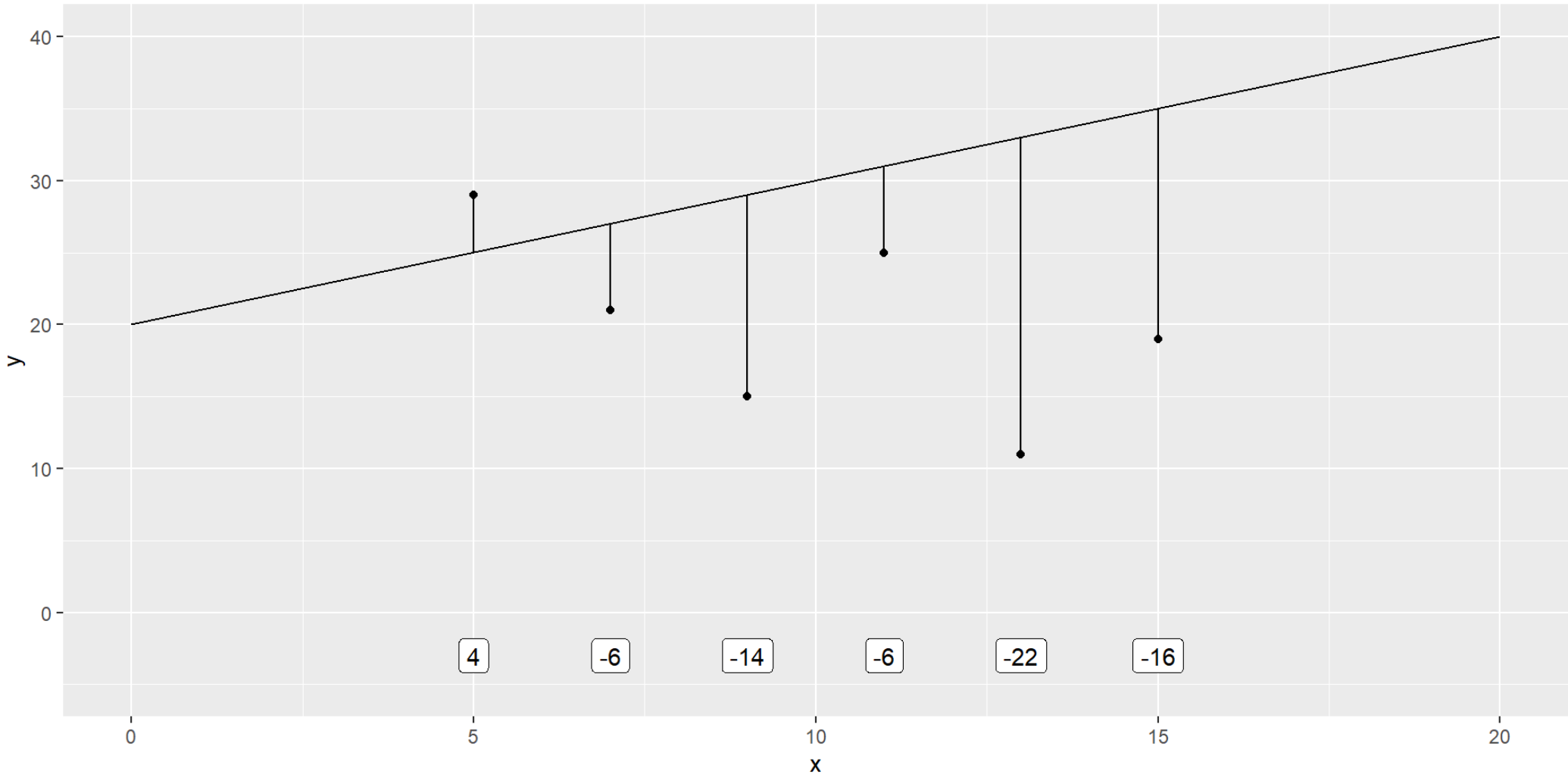
The population model

- $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,\ i = 1, \ldots, N$
  - ϵ_i is an unknown random variable
    - Mean 0, standard deviation, σ
    - Often assumed to be normal
  - β_0 and β_1 are unknown parameters
  - b_0 and b_1 are estimates from the sample

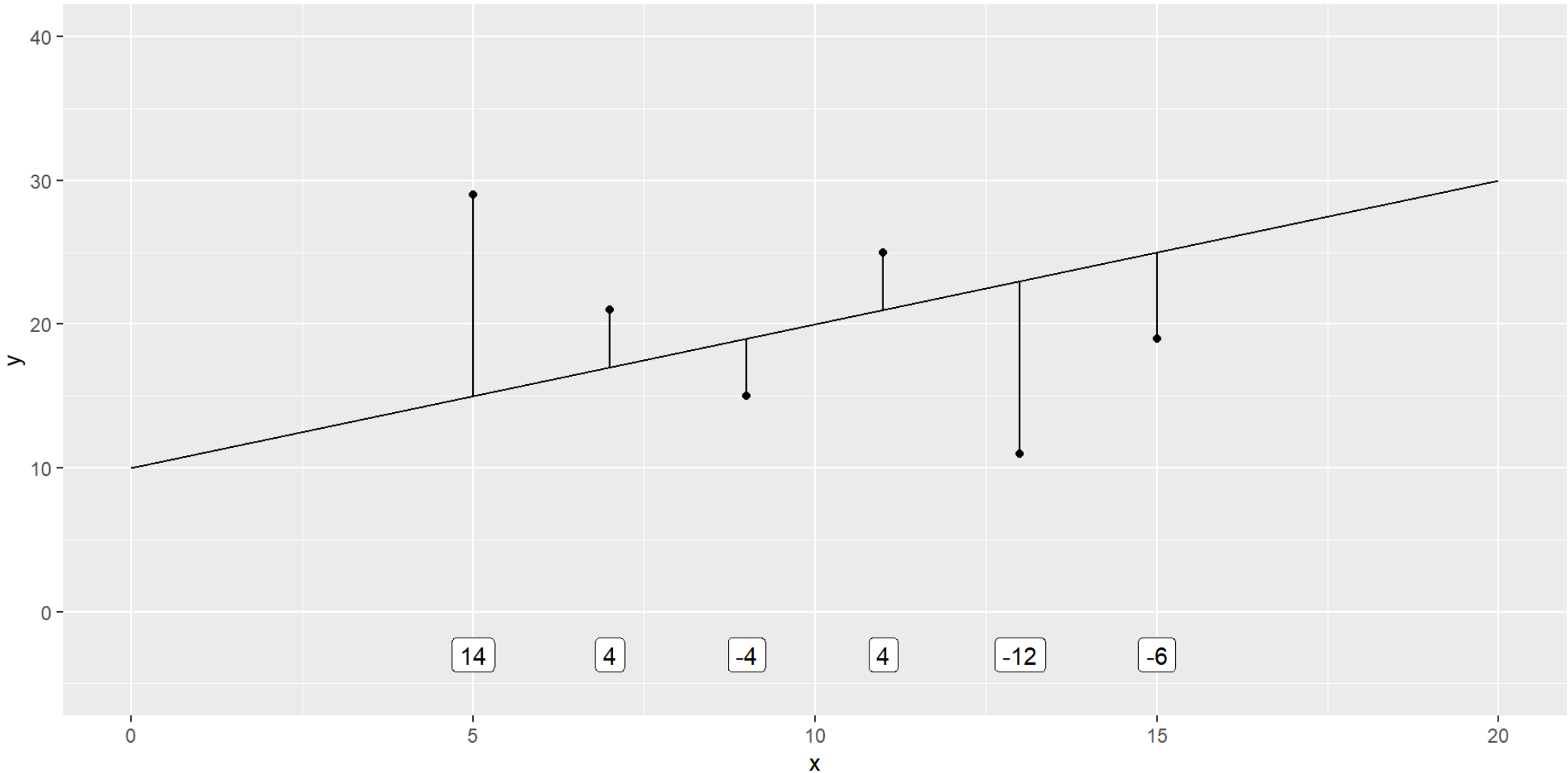
Speaker notes

Add note.

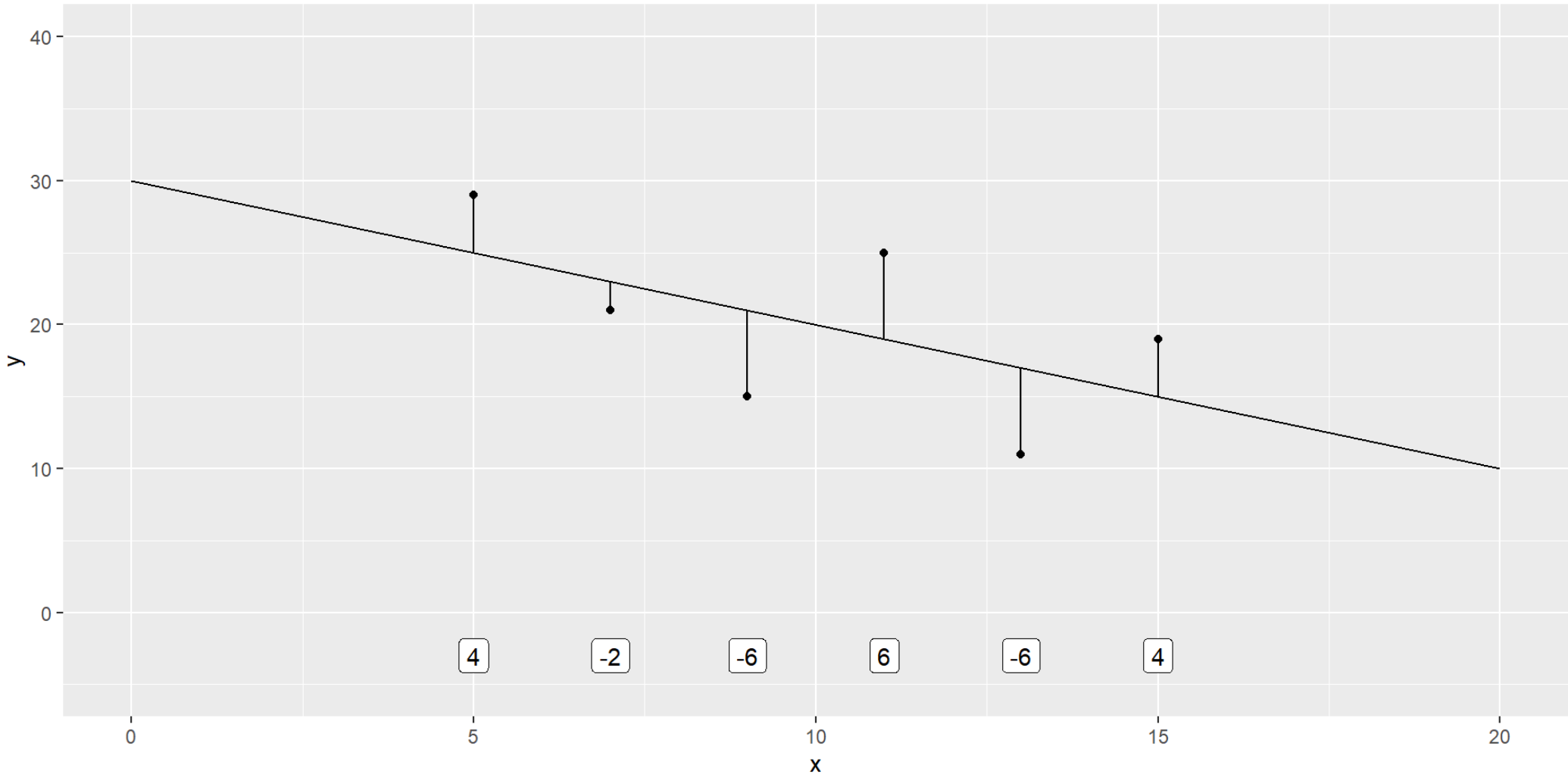
Least squares principle (1/3)



Least squares principle (2/3)



Least squares principle (3/3)



Violations of this model

- Nonlinearity
- Heterogeneity
- Non-normality
- Lack of independence

Break #2

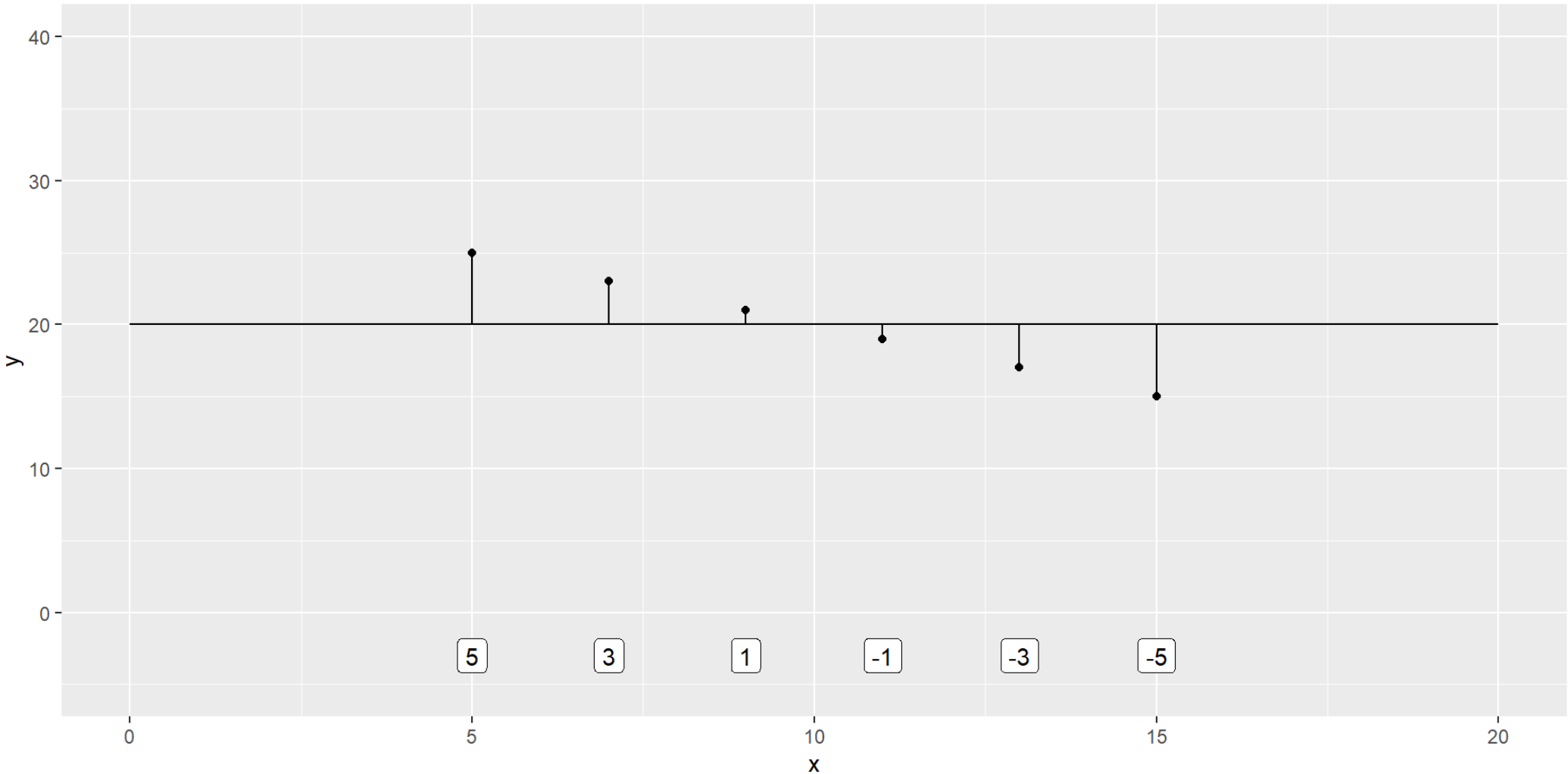
- What you have learned
  - Assumptions in linear regression
- What’s coming next
  - Calculation of sums of squares

Artificial data

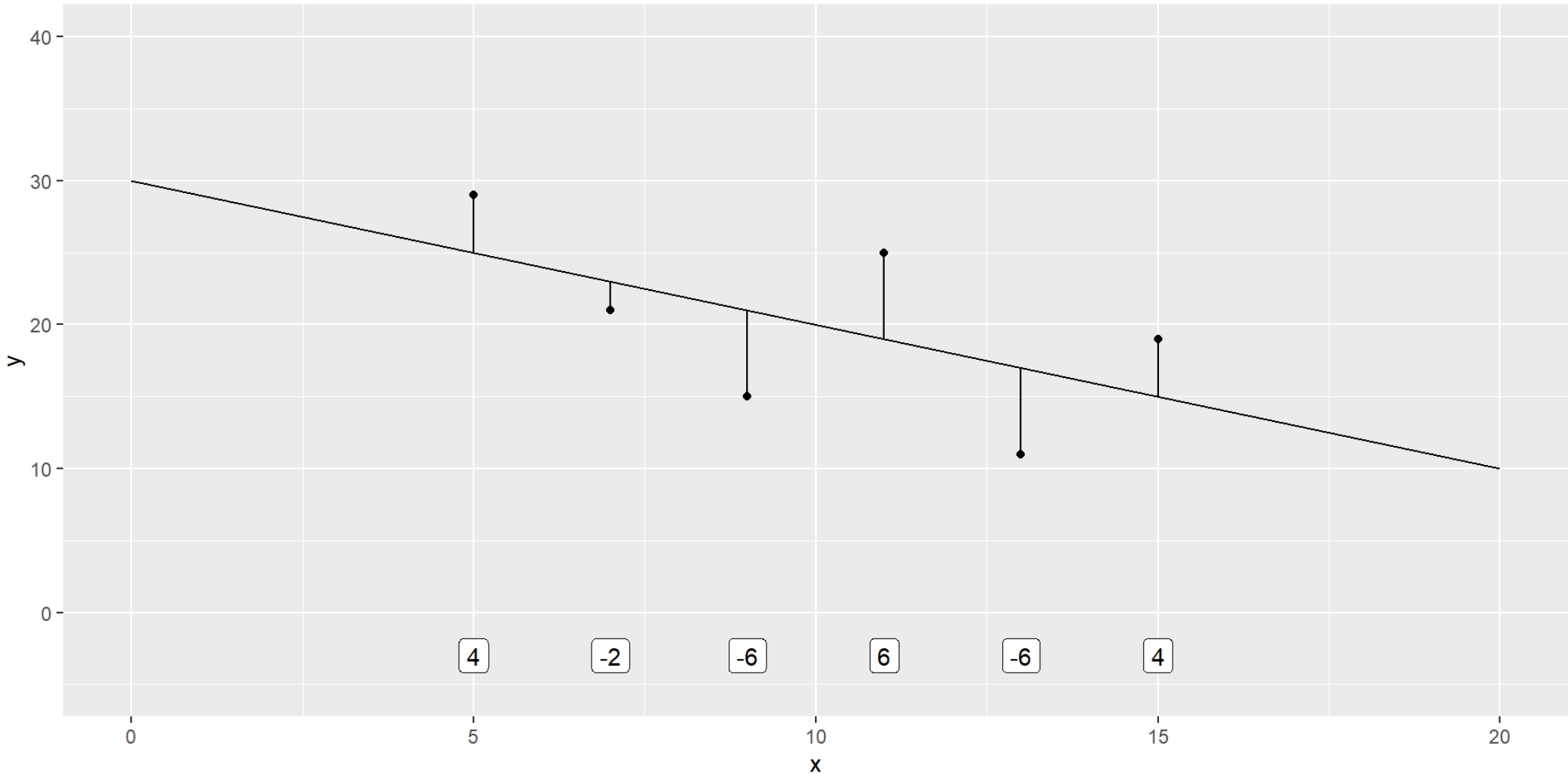
y	x
11	13
15	9
19	15
21	7
25	11
29	5

- $\bar{X}$
- X-bar = 10
- Y-bar = 20
- SD(X) = 3.7
- SD(Y) = 6.5

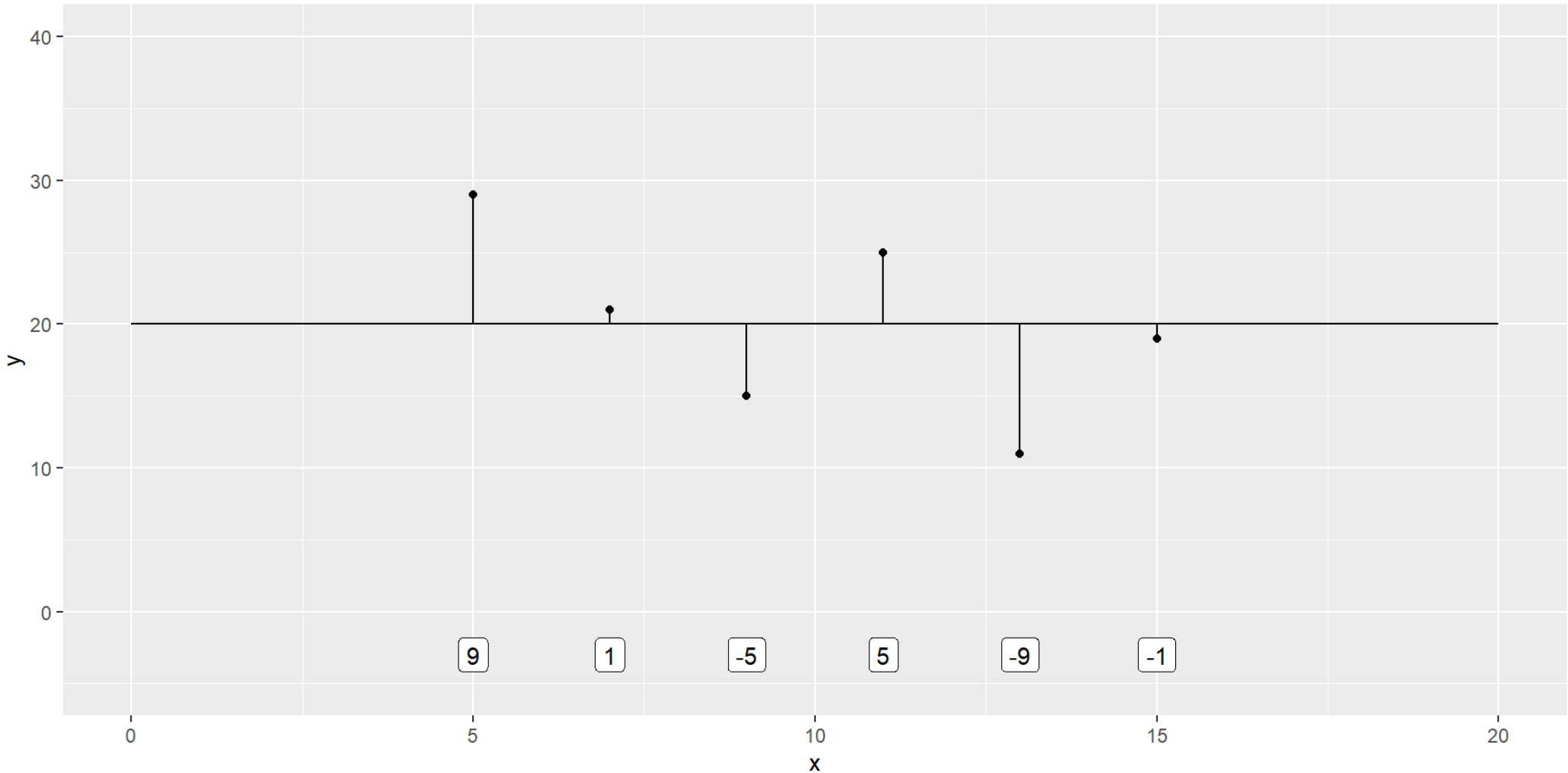
Sum of squares regression



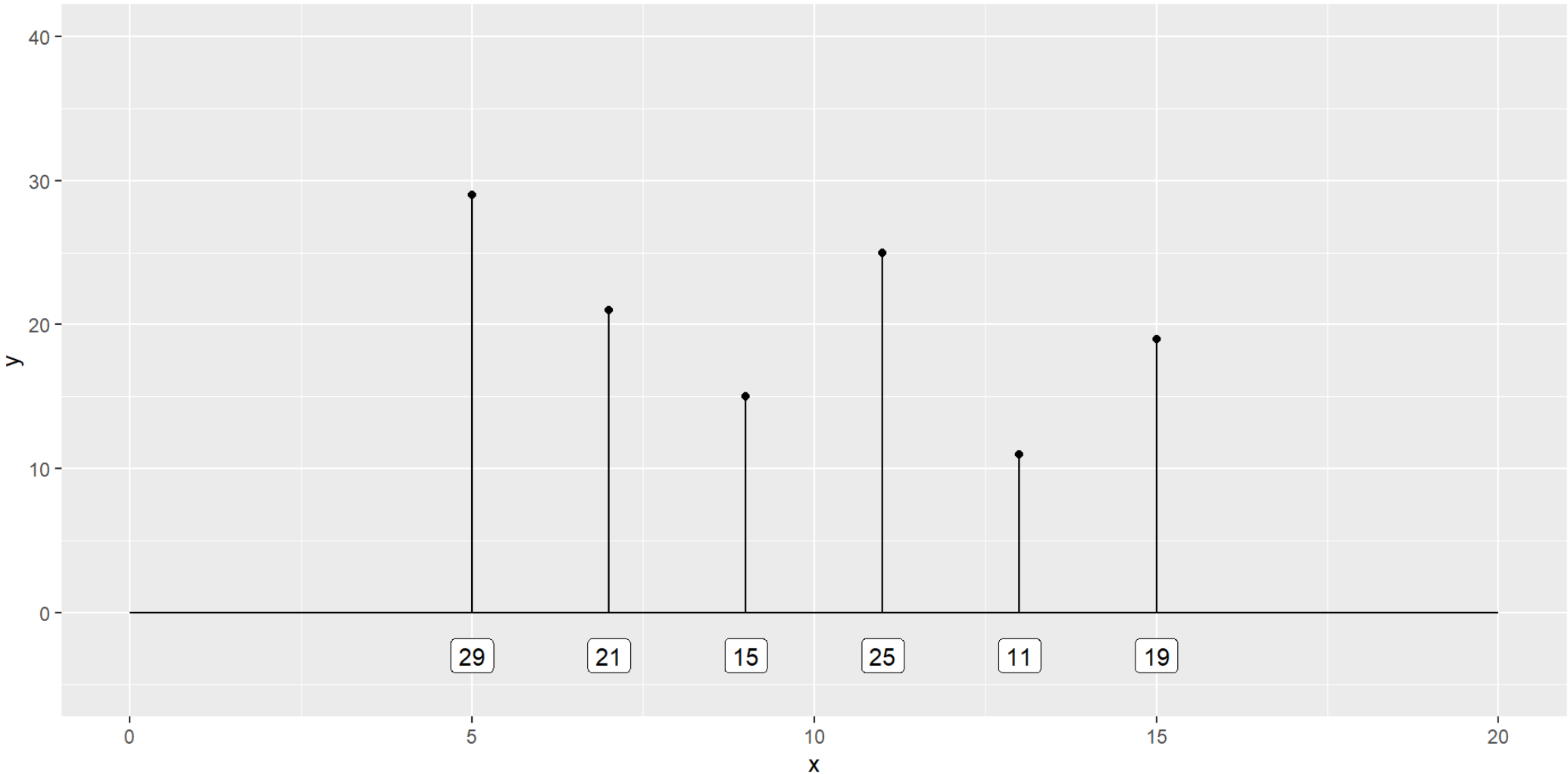
Sum of squares error



Sum of squares total / corrected total



Sum of squares total (uncorrected)



ANOVA table for linear regression

	SS	df	MS	$F-ratio$
$Regression$	SSR	1	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
$Error$	SSE	$n-2$	$MSE = \frac{SSE}{n-2}$	
$Total$	SST	$n-1$		

Review: ANOVA table for oneway ANOVA

	SS	df	MS	$F-ratio$
$Between$	SSB	$k-1$	$MSB=\frac{SSB}{k-1}$	$F=\frac{MSB}{MSW}$
$Within$	SSW	$n-k$	$MSW=\frac{SSW}{n-k-1}$	
$Total$	SST	$n-1$		

R-squared

- SST, total variation, is split into
  - SSR, explained variation, and
  - SSE, unexplained variation
- $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$
  - $0 < R^2 < 1$
  - Proportion of explained variation

SPSS linear regression

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	70.000	1	70.000	1.944	.236[b]
	Residual	144.000	4	36.000		
	Total	214.000	5			

a. Dependent Variable: y

b. Predictors: (Constant), x

- Row 1: SSR, df, MSR, F-ratio, p-value
- Row 2: SSE, df, MSE
- Row 3: SST, df

SPSS General Linear Model

Speaker notes

Ignore Row 2 completely.

Row 3 is the same as Row 1 when you have a single independent variable. With multiple independent variables, SPSS will divide the variation in row 1 across multiple rows, one for each independent variable.

Row 5 is often referred to as the uncorrected total. SPSS does it differently and refers to the uncorrected total as just plain “total” and then uses “Corrected total” where most researchers would use just plain “total”.

You may prefer the previous output as being less confusing and that’s fine. The big advantage of the general linear model lies in its ability to model a variety of different analyses.

Break #3

- What you have learned
  - Calculation of sums of squares
- What’s coming next
  - Confidence intervals and hypothesis tests

Confidence interval

- $b_1 \pm t(\alpha/2, n-2) s.e.(b_1)$
  - $s.e.(b_1) = \sqrt{\frac{MSE}{\Sigma(X_i - \bar{X})^2}}$
- How to get a narrower confidence interval
  - Decrease the noise (MSE)
  - Increase the sample size
  - Increase the spread of the X's

Speaker notes

The slope, β_1, in the entire population is an unknown parameter. You take a sample and estimate the slope from the sample. Any estimate based on a sample has sampling error.

If you can find a way to decrease the noise (as measured by MSE), that will produce a narrower confidence interval. A larger sample size will also work, because you will be computing a summation of $(X_i - \bar{X})^2)$ with more values, all of which are non-negative. Finally, you can increase the spread of the X's. Think of a table where all the legs are close to the center of the table. It would be very unstable. In contrast, if you spread the legs of a table out, you will produce much more stability. The same is true in regression. A wide spread of the independent variable improves precision.

Hypothesis test

- $H_0: \ \beta_1 = 0$ vs. $H_1: \ \beta_1 \neq 0$
- Compare $T = \frac{b_1}{s.e.(b_1)}$ to $t(1-\alpha/2; n-2)$
  - Accept H_0 if T is close to zero
  - Reject H_0 if T is large negative or large positive

Speaker notes

The hypotheses involve the population parameter, β_1. To test this hypothesis, compare the sample statistic, b_1, to its standard error. If that ratio is close to zero, then you would accept the null hypothesis. If you see extreme values, very large negative or very large positive values, then you should reject the null hypothesis.

Equivalent hypothesis test

- Compare $F = \frac{MSR}{MSE}$ to $F(1-\alpha; 1, n-2)$
  - Accept H_0 if F is close to one
  - Reject H_0 if T is large positive
- Note: $F = T^2$

Speaker notes

You get an identical result if you compute the F-ratio, MSR divided by MSE. If this value is close to 1, you would accept the null hypothesis. The signal (MSR) is comparable to the noise (MSE). If you see a large positive ratio, that implies that the signal is much stronger than the noise. A large positive ratio would cause you to reject the null hypothesis.

Two more equivalent tests

- Compute the p-value from T or F
  - Accept H_0 if $p - value > \alpha$
  - Reject H_0 if $p - value \leq \alpha$
- Compute the confidence interval (CI) for β_1
  - Accept H_0 if CI includes 0
  - Reject H_0 if CI does not include 0

Speaker notes

Recall that the p-value is the probability of observing the sample result or a result more extreme. If the p-value is large (greater than α), then you have little or no evidence against the null hypothesis. If the p-value is small (less than or equal to α), then you have lots of evidence against the null hypothesis.

Break #4

- What you have learned
  - Confidence intervals and hypothesis tests
- What’s coming next
  - Relationship to the correlation coefficient

Calculation of the regression slope and intercept

- $b_1 = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}$
- $b_0 = \bar{Y} - b_1 \bar{X}$

Relationship to the correlation coefficient

- Recall from the previous module
  - $Cov(X, Y) = \frac{1}{n-1}\Sigma(X_i - \bar{X})(Y_i - \bar{Y})$
  - $r_{XY} = \frac{Cov(X,Y)}{S_X S_Y}$
- This implies that
  - $b_1 = r_{XY}\frac{S_Y}{S_X}$

Important implications

- r_{XY} is unitless, b_1 is Y units per X units
- $r_{XY} > 0$ implies $b_1 > 0$
- $r_{XY} = 0$ implies $b_1 = 0$
- $r_{XY} < 0$ implies $b_1 < 0$
  - and vice versa

Break #5

- What you have learned
  - Relationship to the correlation coefficient
- What’s coming next
  - Definition of residuals

Predicted values

- For a new value of X
  - $\hat{Y}_{new} = b_0 + b_1 X_{new}$
- For an existing value in the data, X_i
  - $\hat{Y}_i = b_0 + b_1 X_i$

Why predict for a value you already have seen?

- Future Y may differ from previous Y
- Comparison of $\hat{Y}_i$ to existing Y_i.

Residual

- $e_i = Y_i - \hat{Y}_i$
  - Error in prediction
  - $\Sigma e_i = 0$
  - Estimate of ϵ_i

Speaker notes

The residual is the difference between what you observed (Y_i) and what the linear regression model would predict ($\hat{Y}_i$). If the residual is zero, you nailed it. A perfect prediction. That may happen once in a while, but you will almost never see perfect predictions for every patient in your study. The residuals will sum to zero because the linear regression uses the least squares principle to

Break #6

- What you have learned
  - Definition of residuals
- What’s coming next
  - Diagnostic plots

Testing the various assumptions

- Nonlinearity
  - Scatterplot of residuals vs. independent variable
- Heterogeneity
  - Scatterplot of residuals vs. independent variable
- Non-normality
  - Q-Q plot of residuals
- Lack of independence
  - Usually assessed qualitatively
  - Durbin-Watson test for serial correlation

Speaker notes

Add note.

Break #7

- What you have learned
  - Diagnostic plots
- What’s coming next
  - Interpretation with two independent variables

Model

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

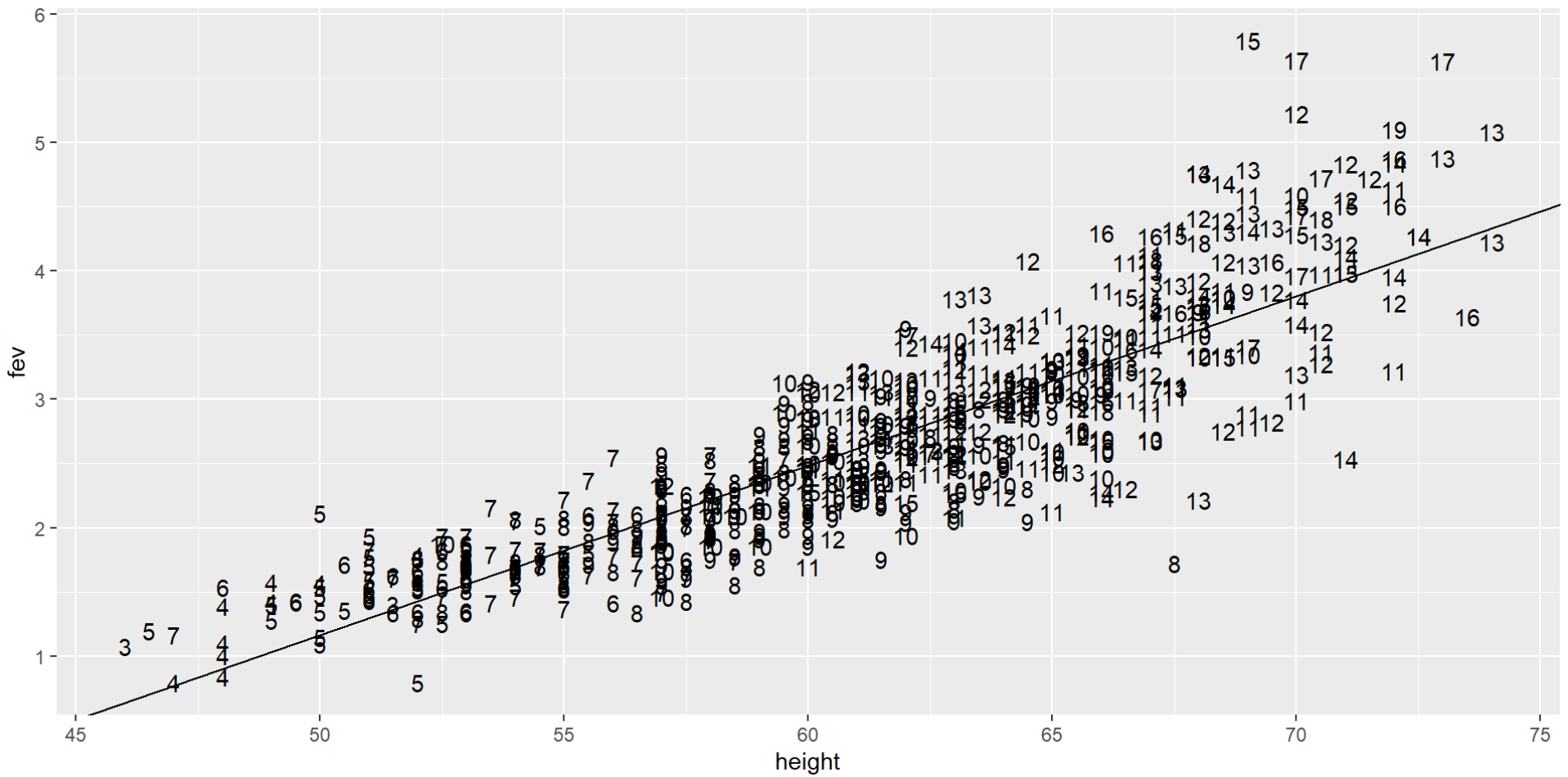
Speaker notes

Add note.

Interpretations

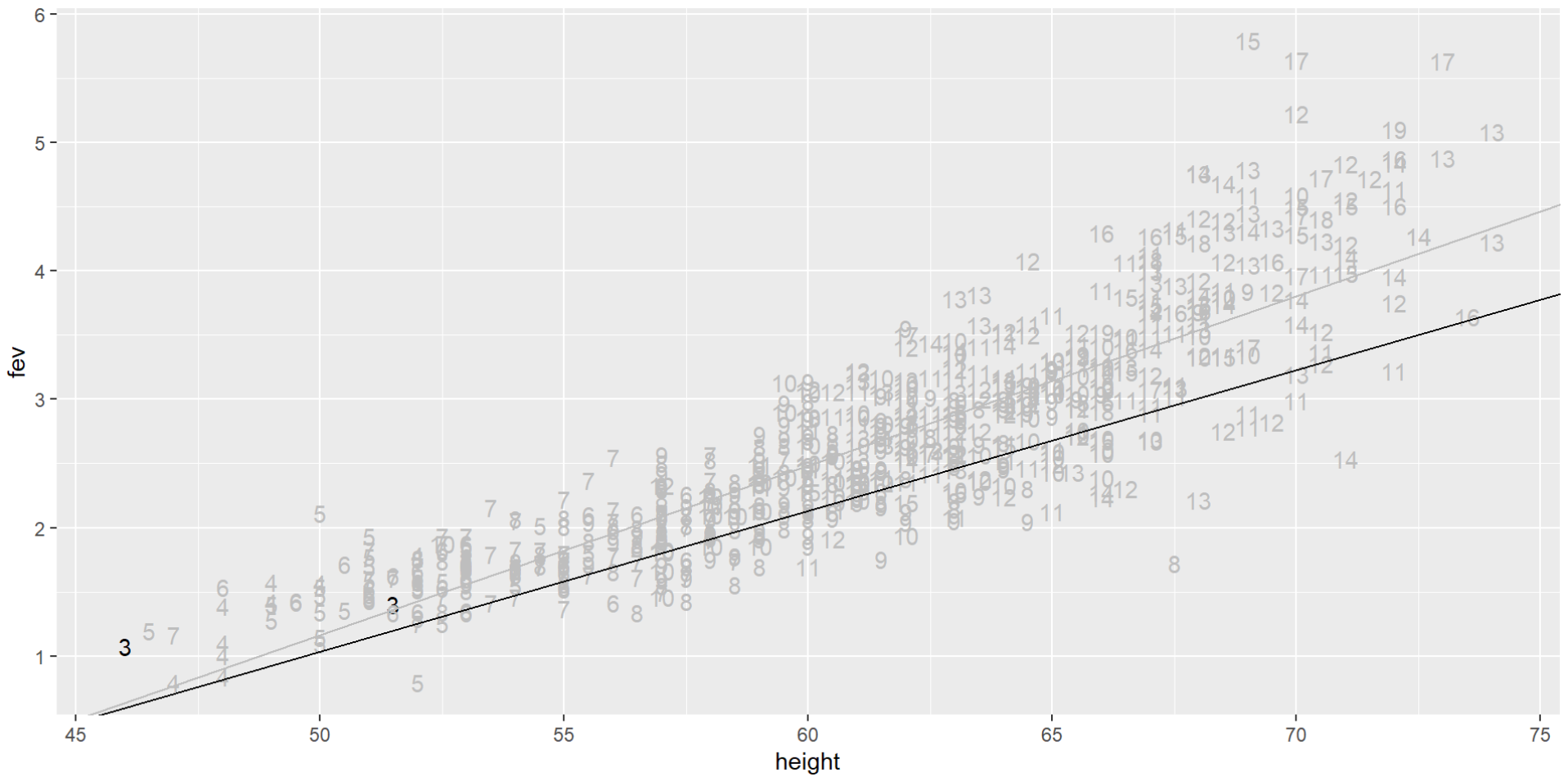
- b_0 is the estimated average value of Y when X1 and X2 both equal zero.
- b_1 is the estimated average change in Y
  - when X_1 increases by one unit, and
  - X_2 is held constant
- b_2 is the estimated average change in Y
  - when X_2 increases by one unit, and
  - X_1 is held constant

Unadjusted relationship between height and FEV

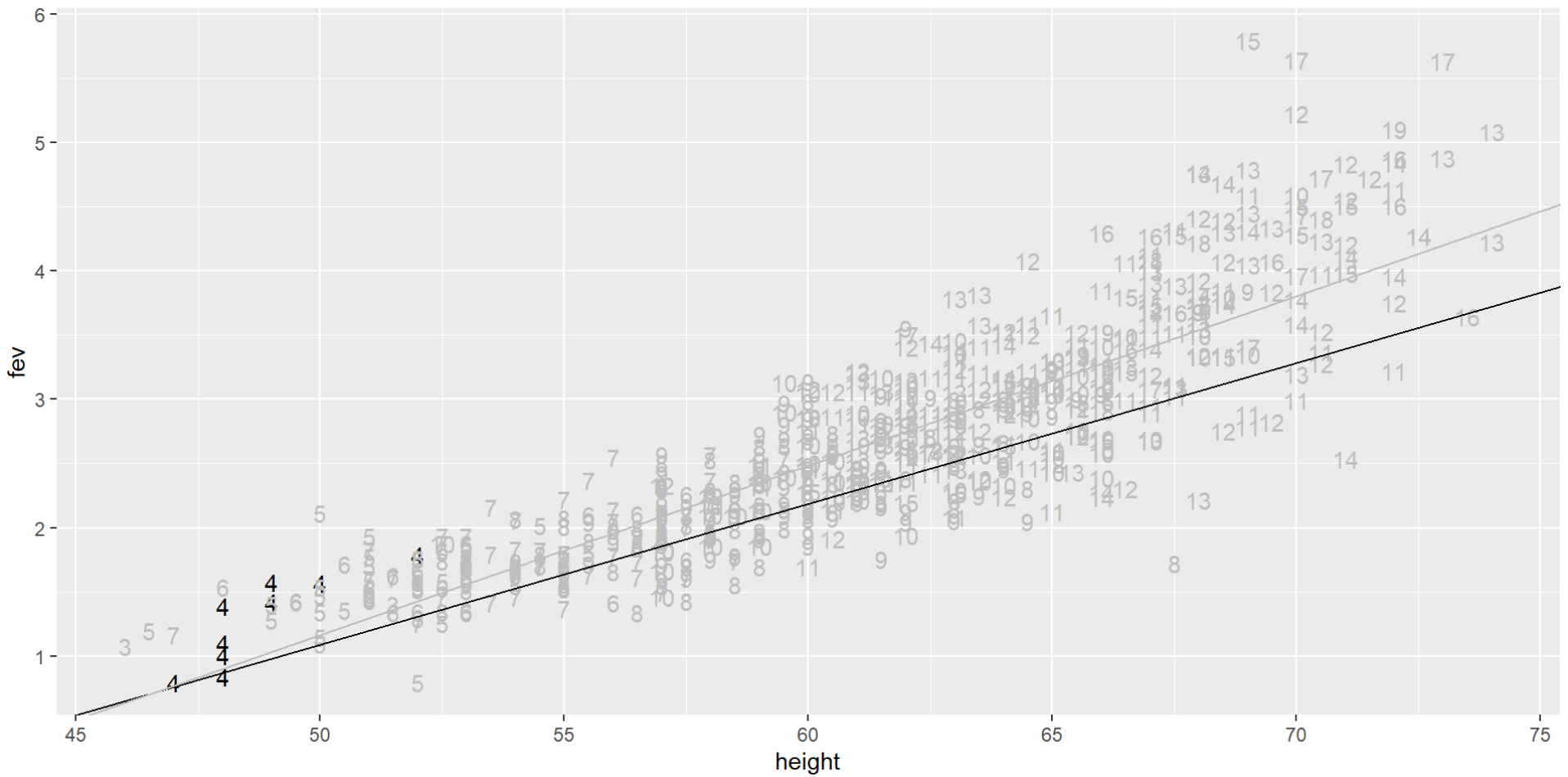


Relationship between height and FEV controlling at Age=3

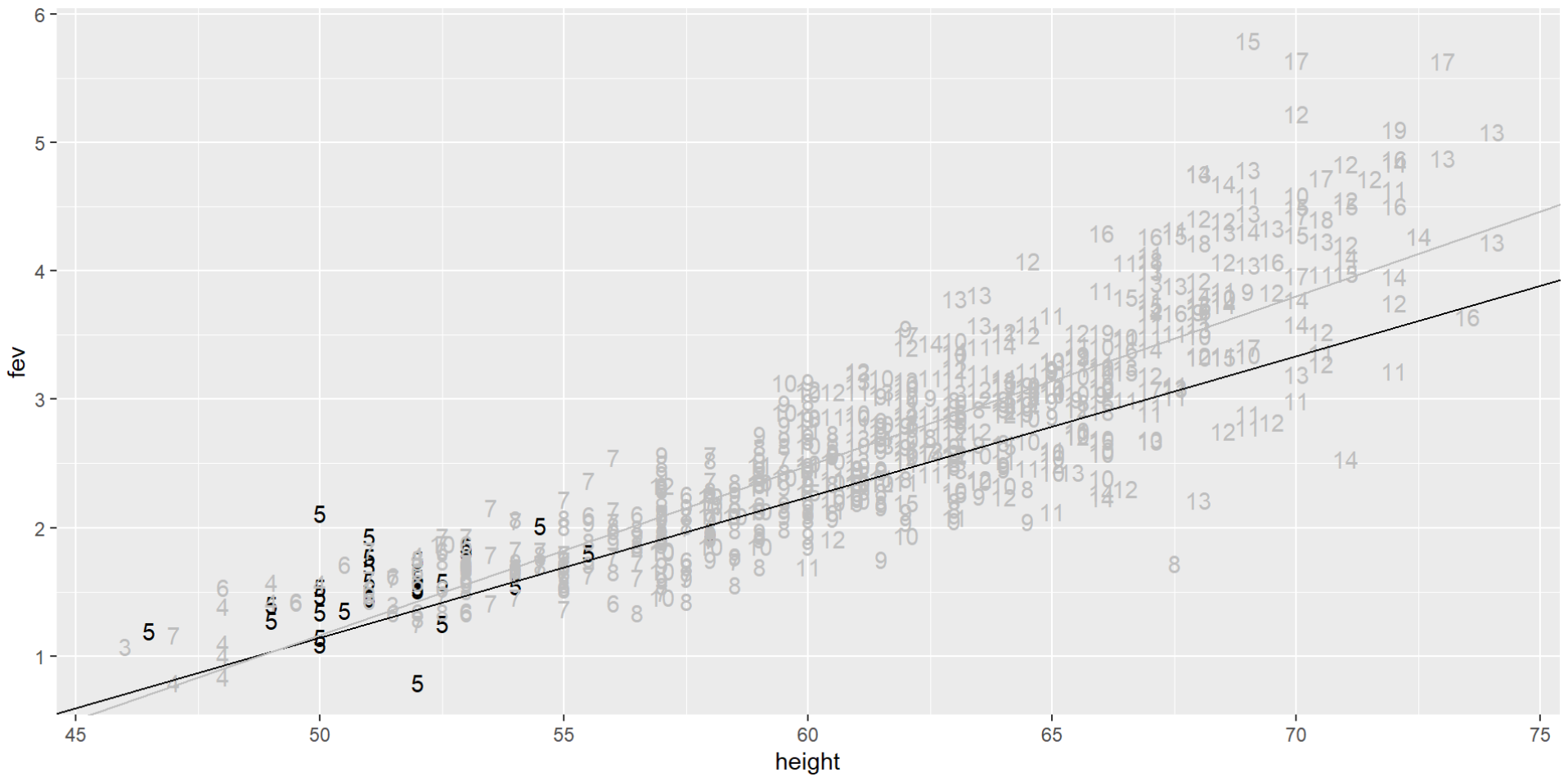
Relationship between height and FEV controlling at Age=3



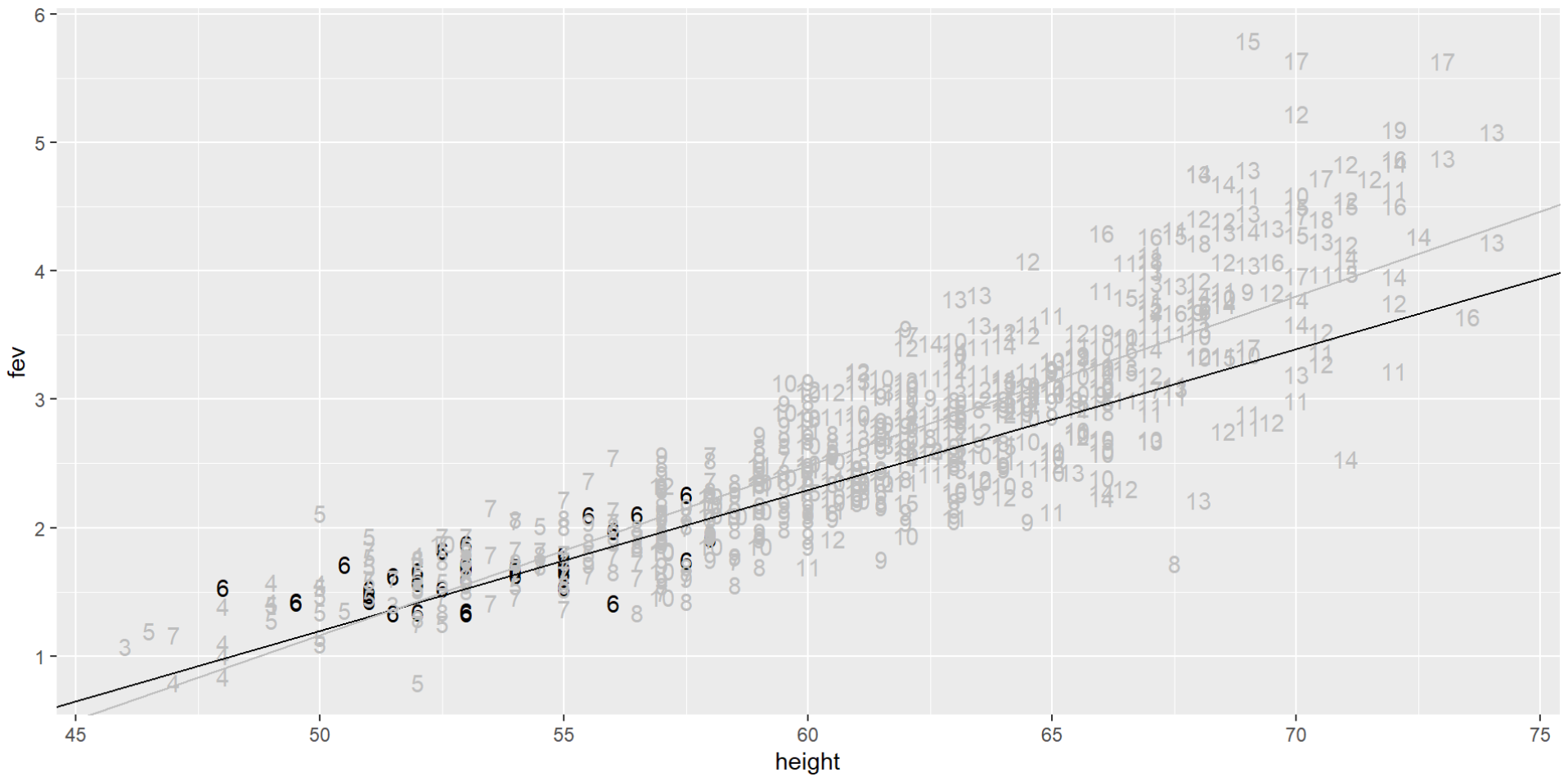
Relationship between height and FEV controlling at Age=4



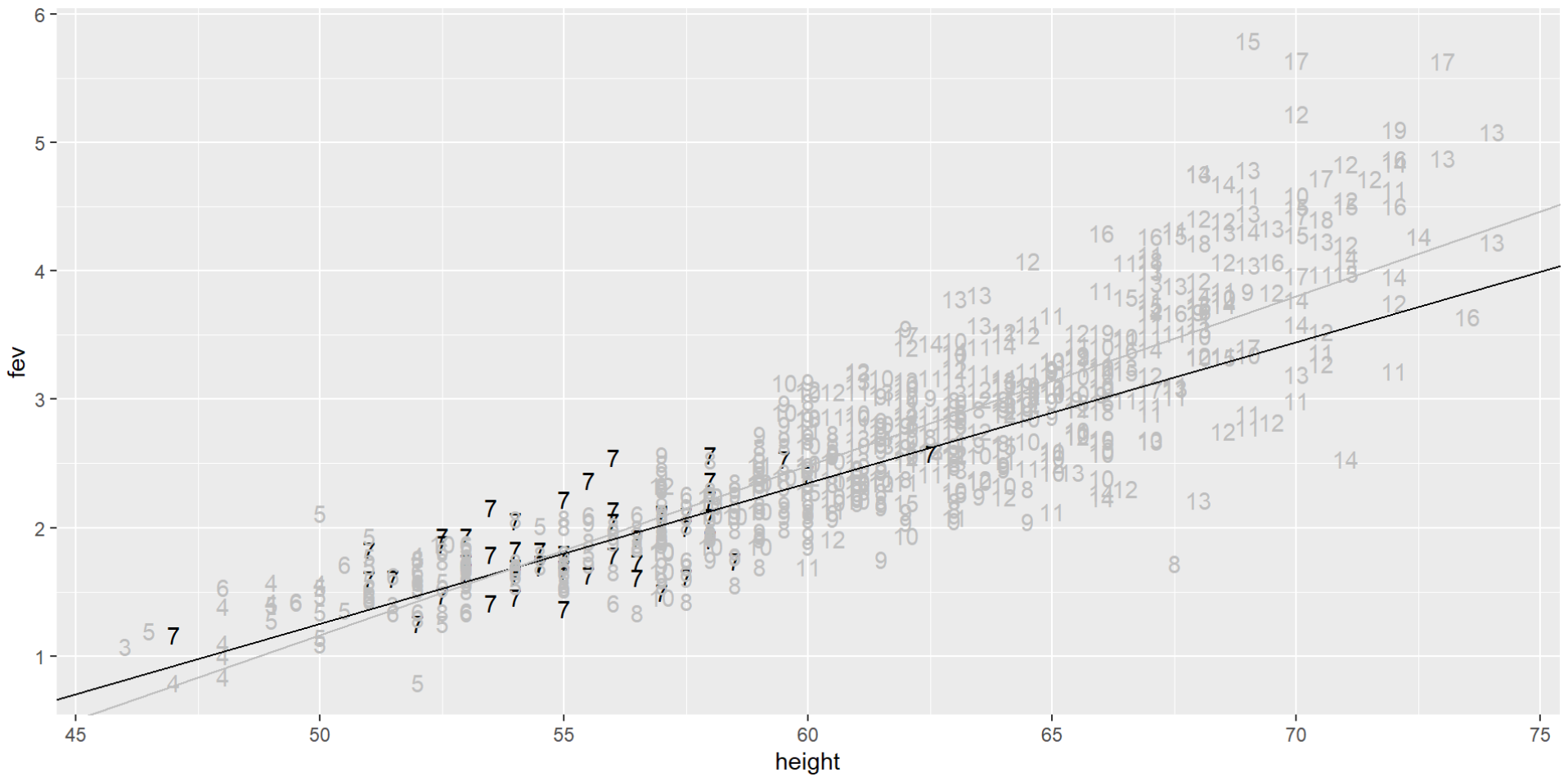
Relationship between height and FEV controlling at Age=5



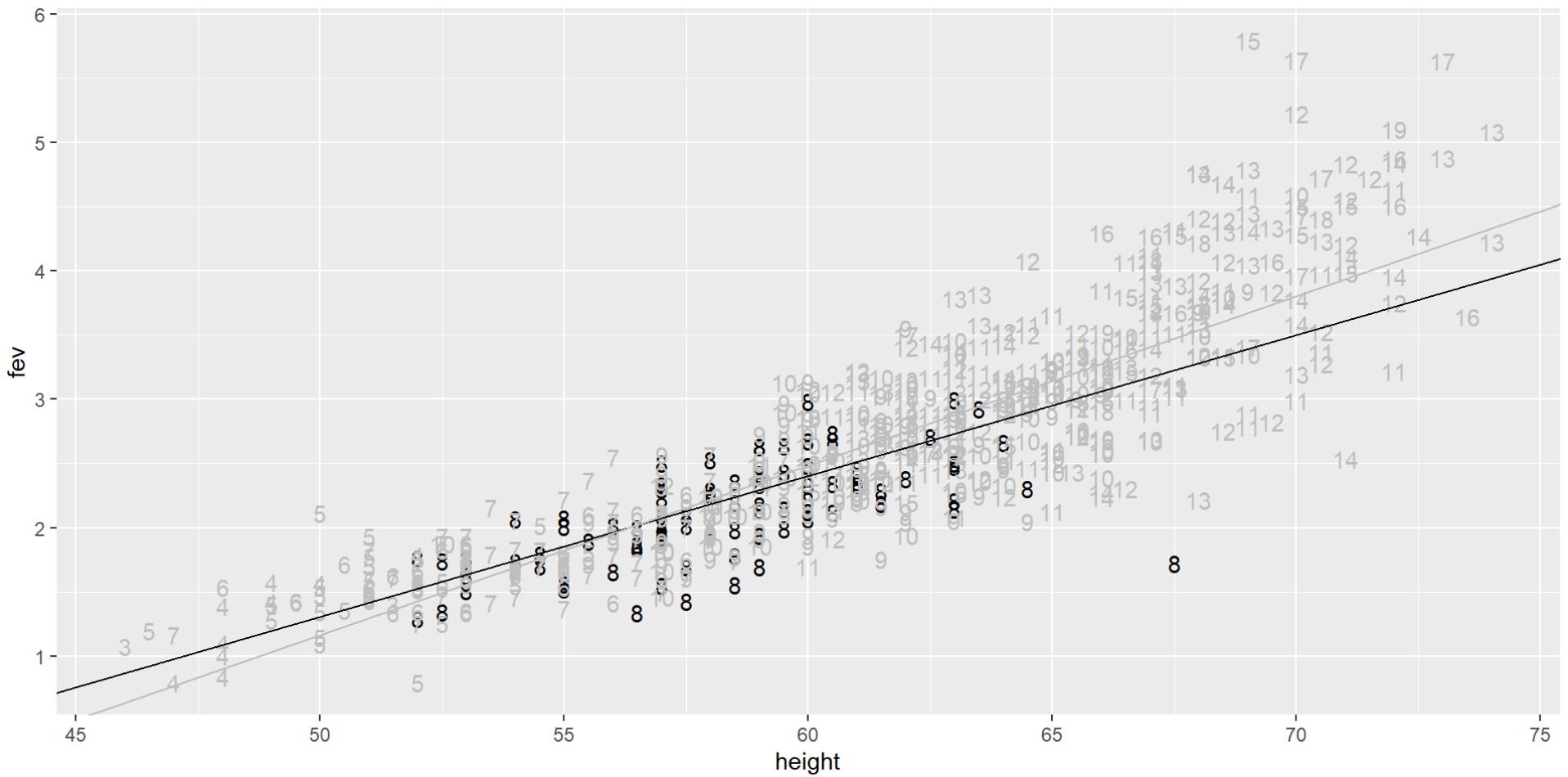
Relationship between height and FEV controlling at Age=6



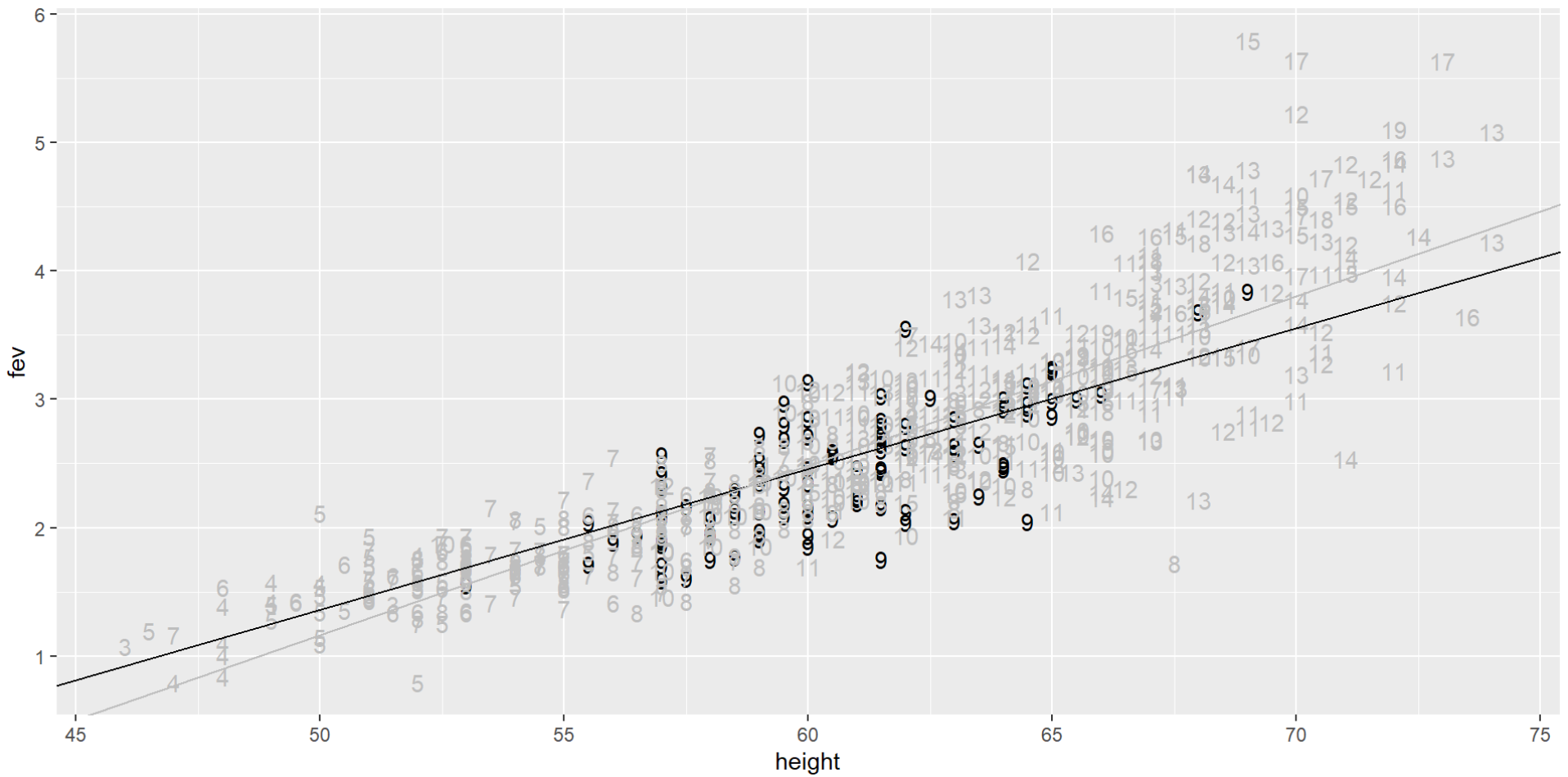
Relationship between height and FEV controlling at Age=7



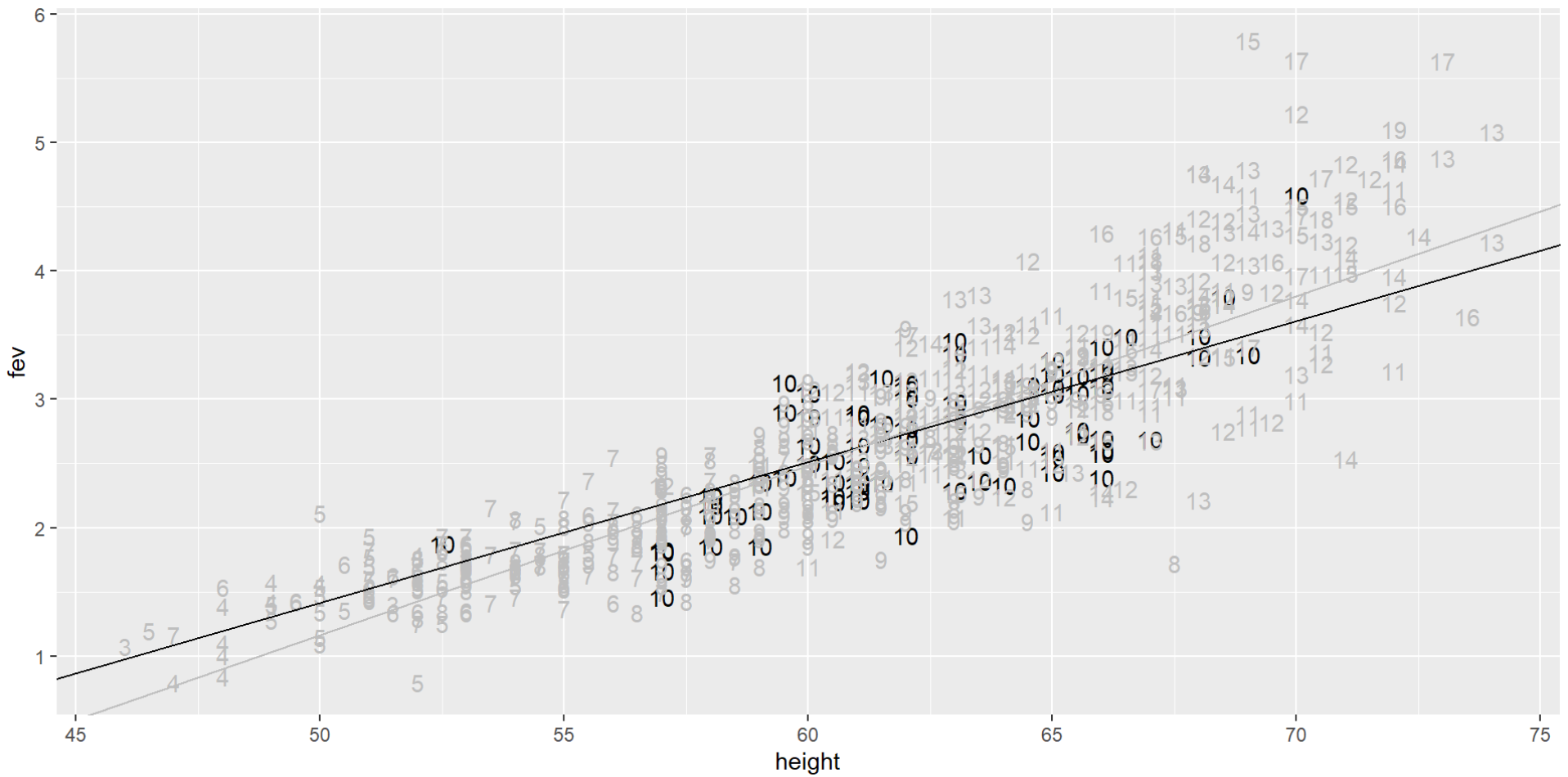
Relationship between height and FEV controlling at Age=8



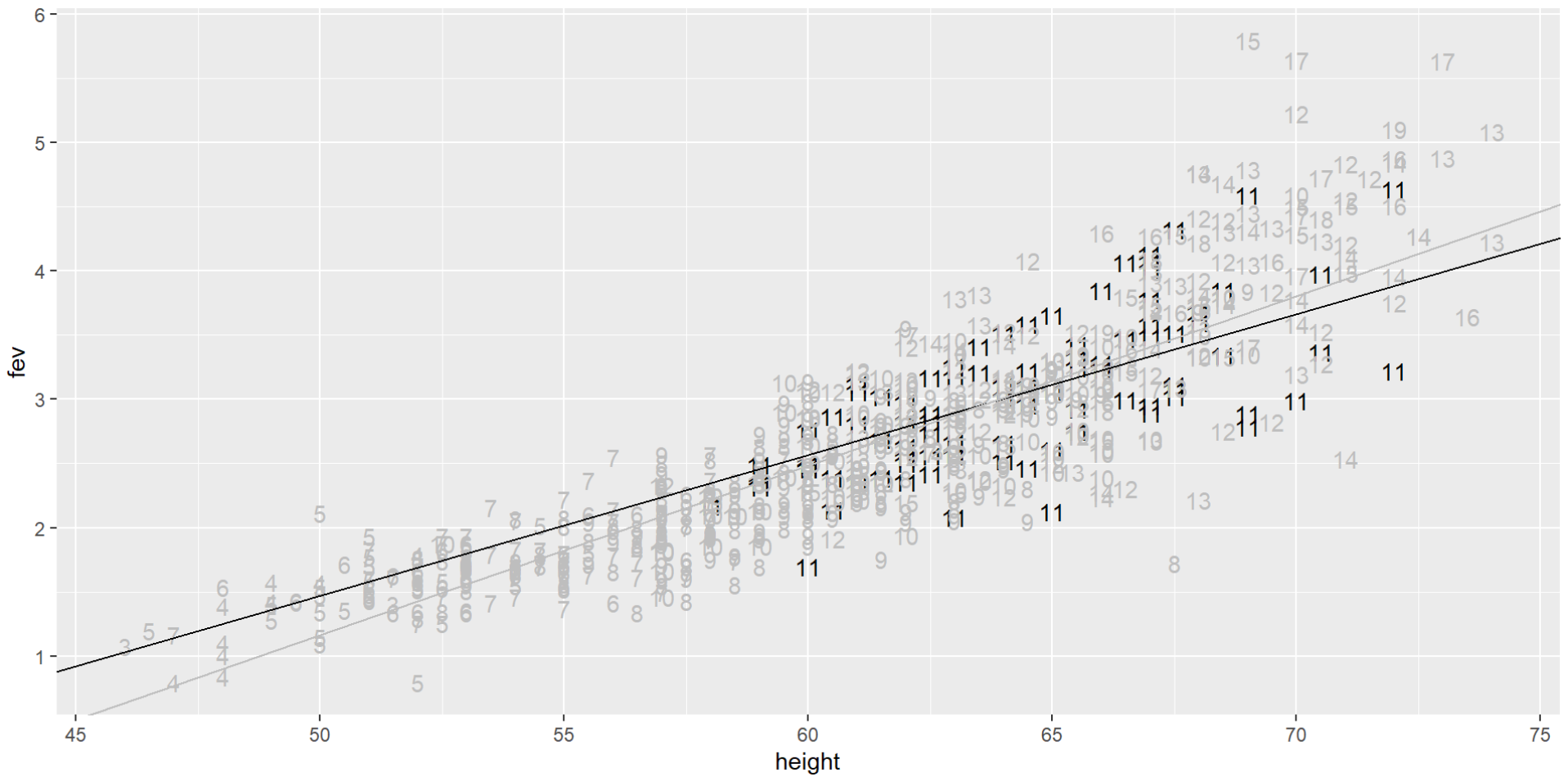
Relationship between height and FEV controlling at Age=9



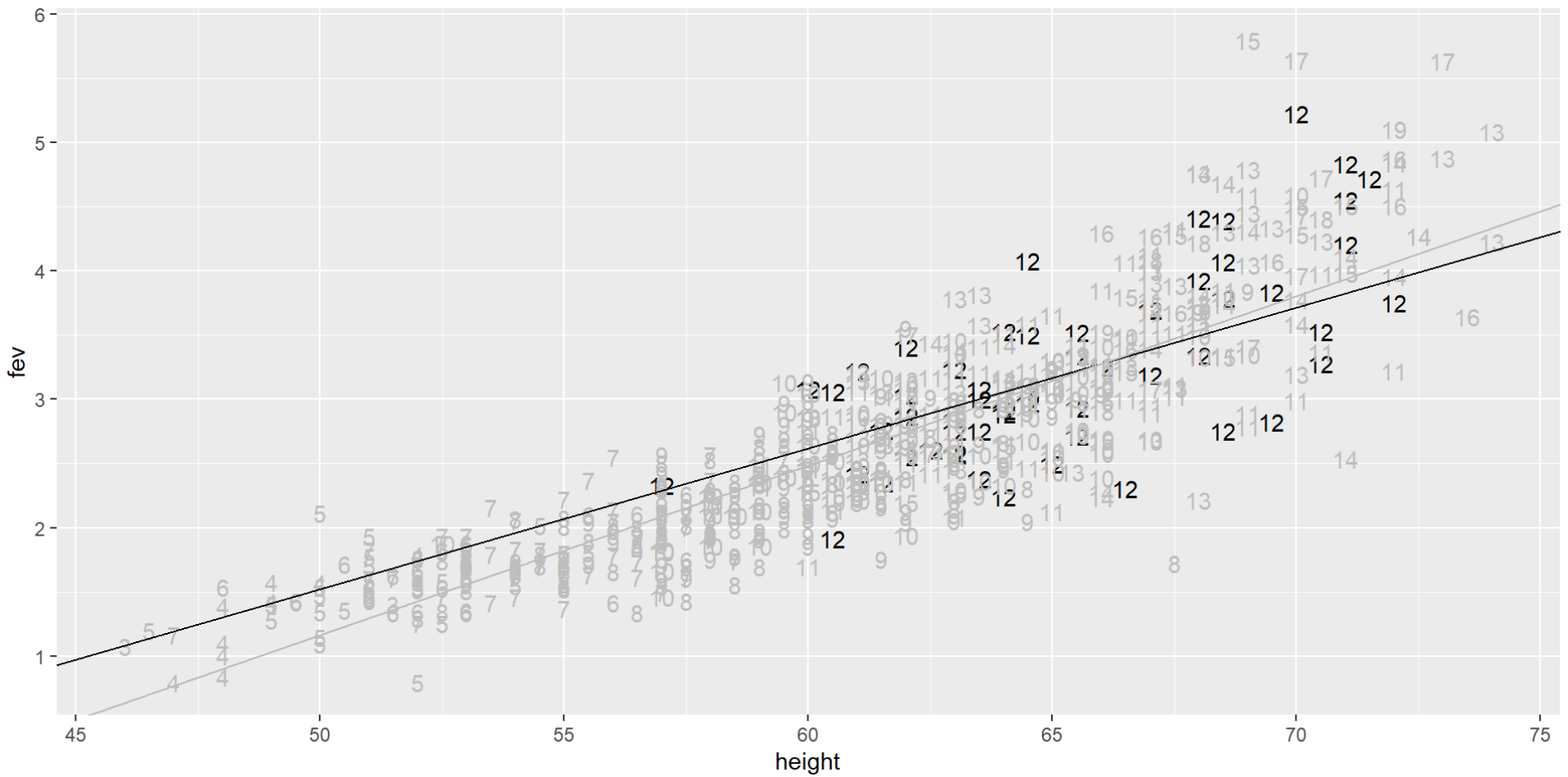
Relationship between height and FEV controlling at Age=10



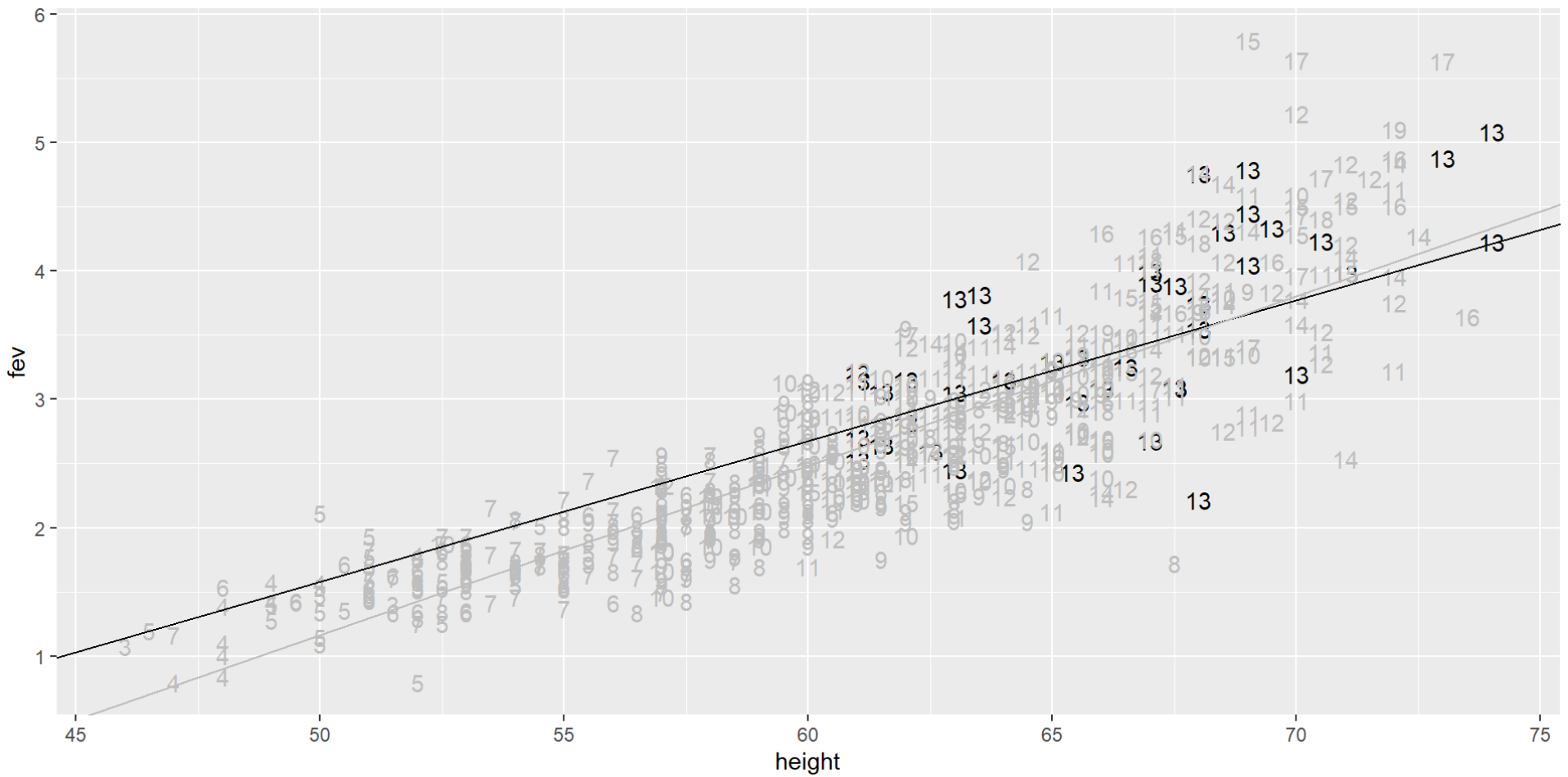
Relationship between height and FEV controlling at Age=11



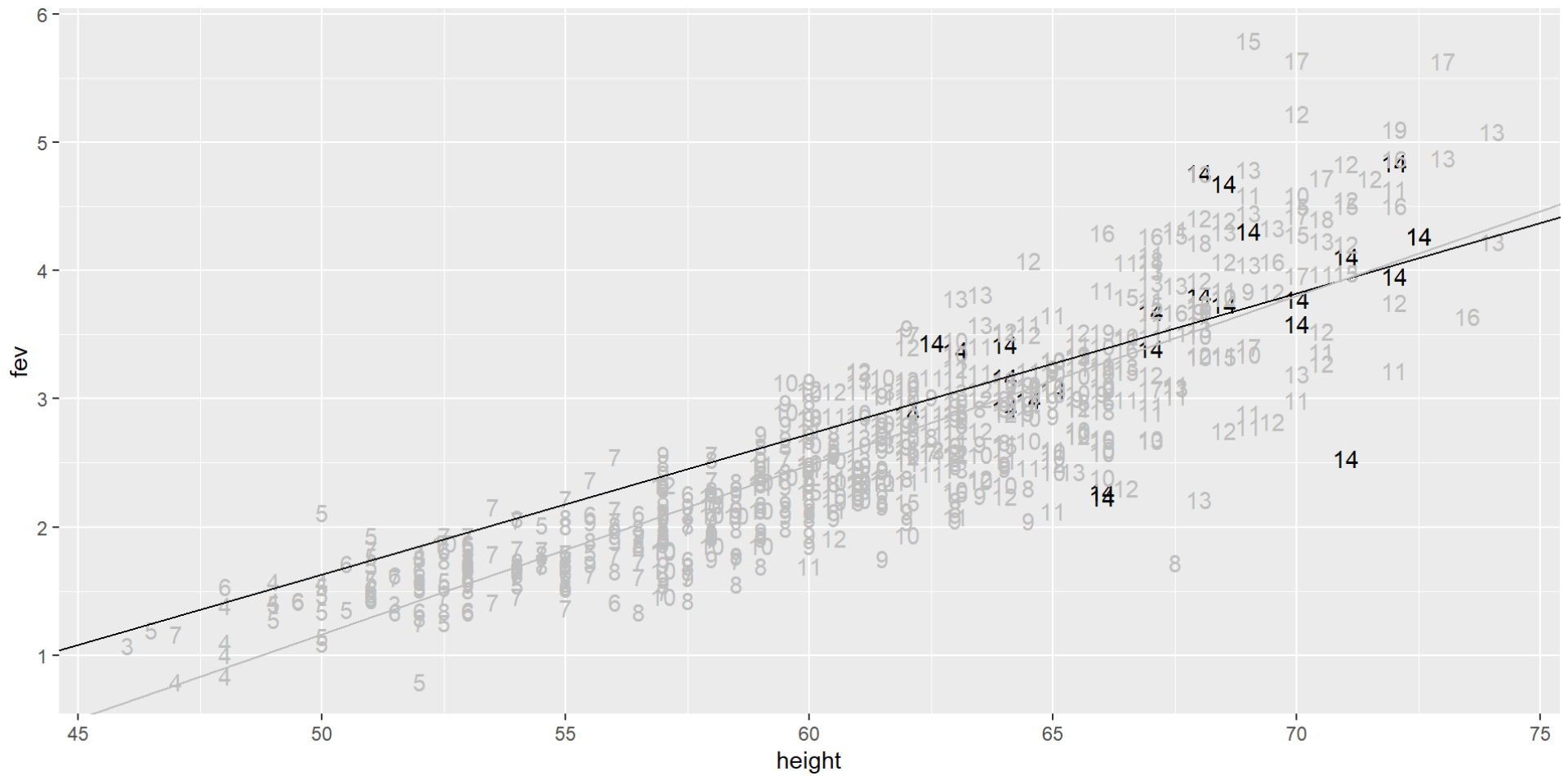
Relationship between height and FEV controlling at Age=12



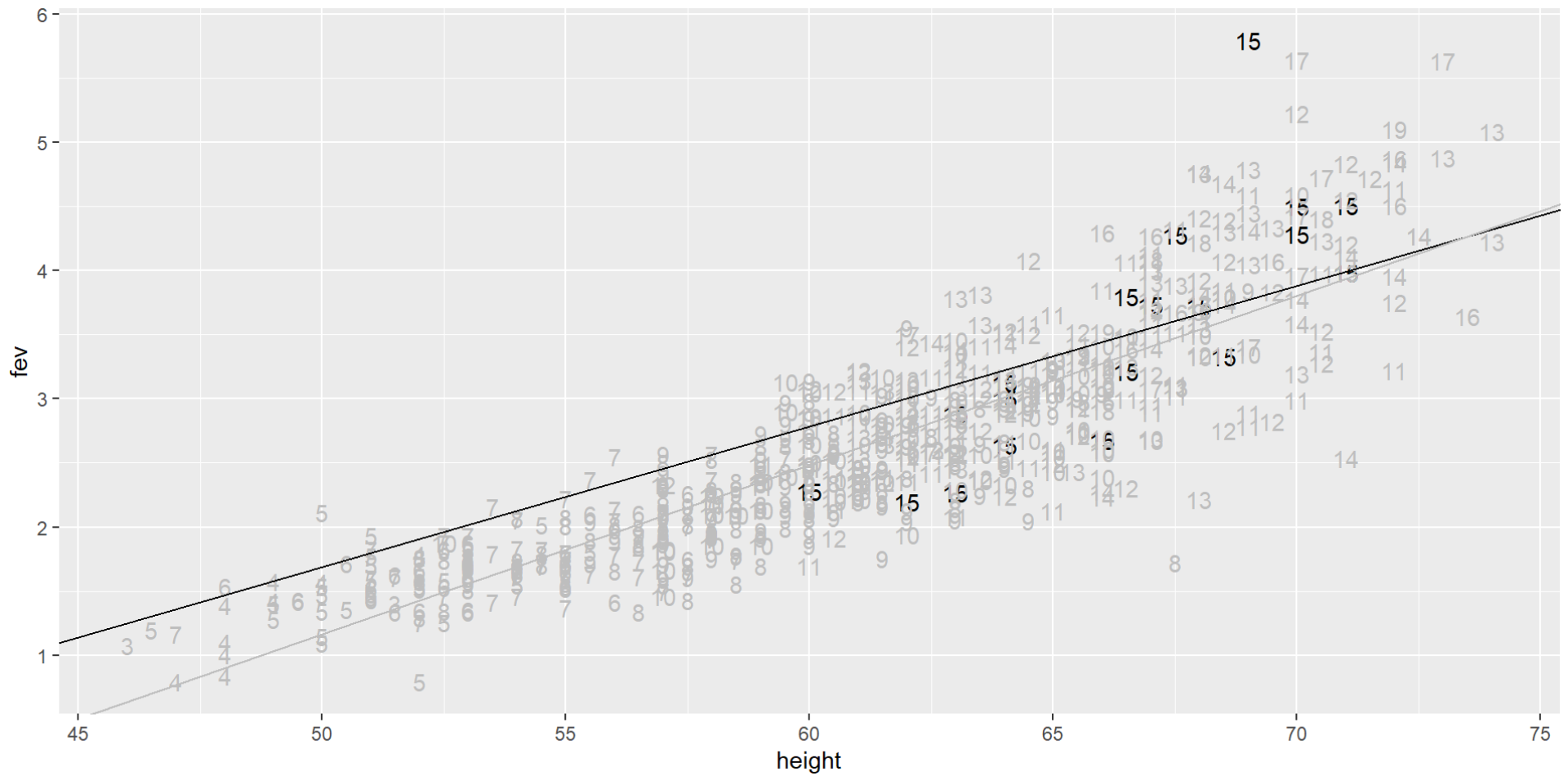
Relationship between height and FEV controlling at Age=13



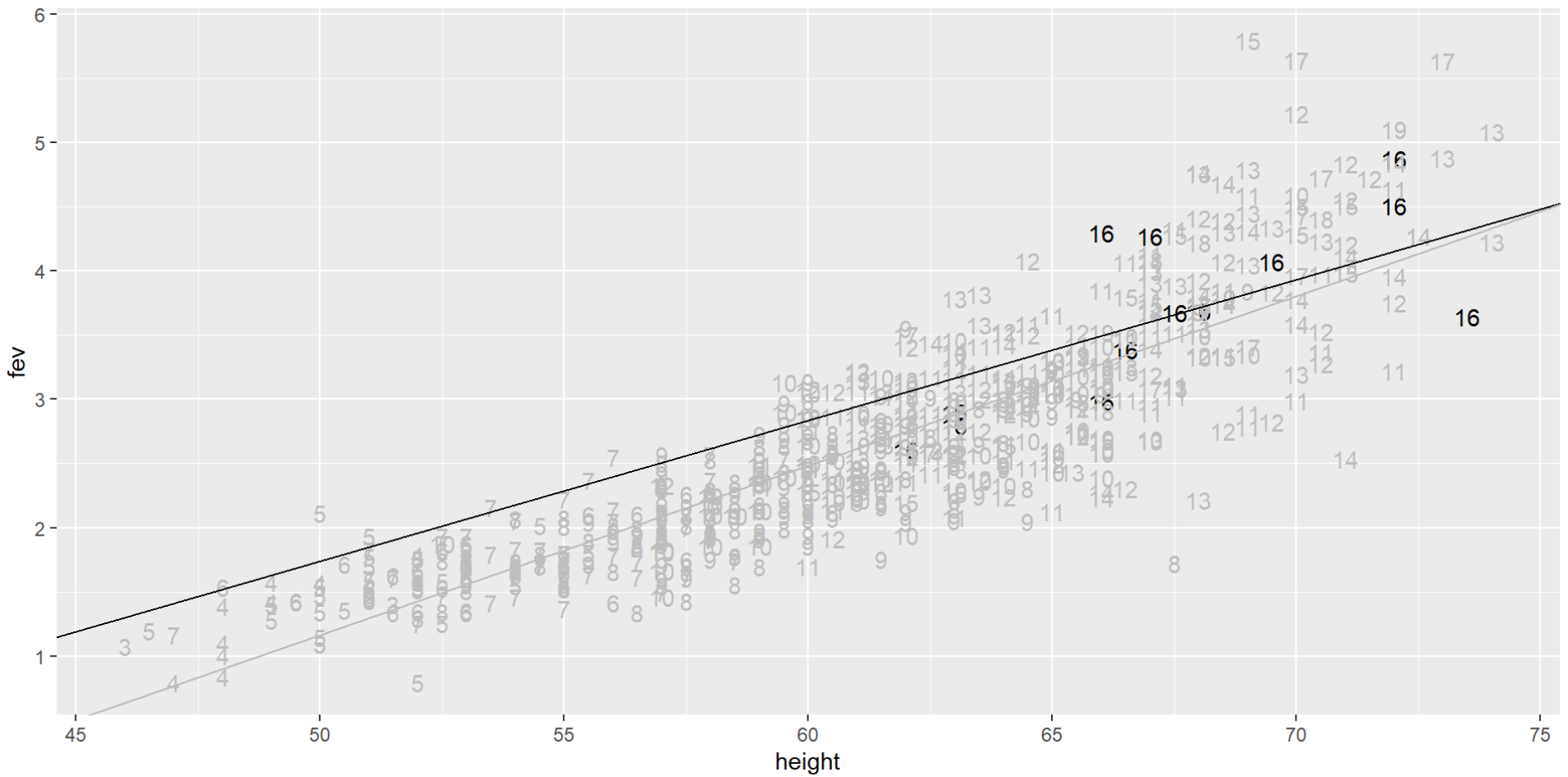
Relationship between height and FEV controlling at Age=14



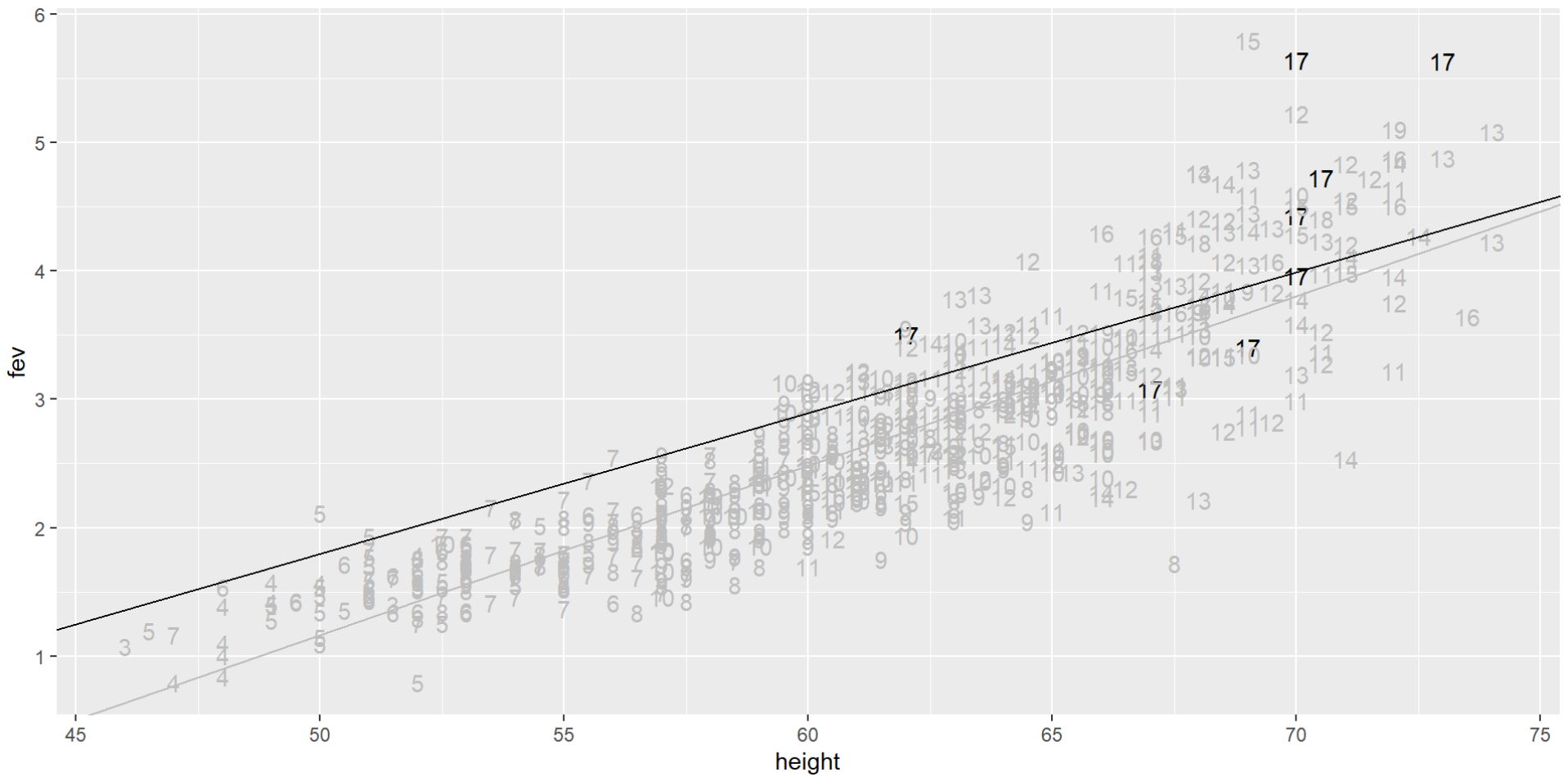
Relationship between height and FEV controlling at Age=15



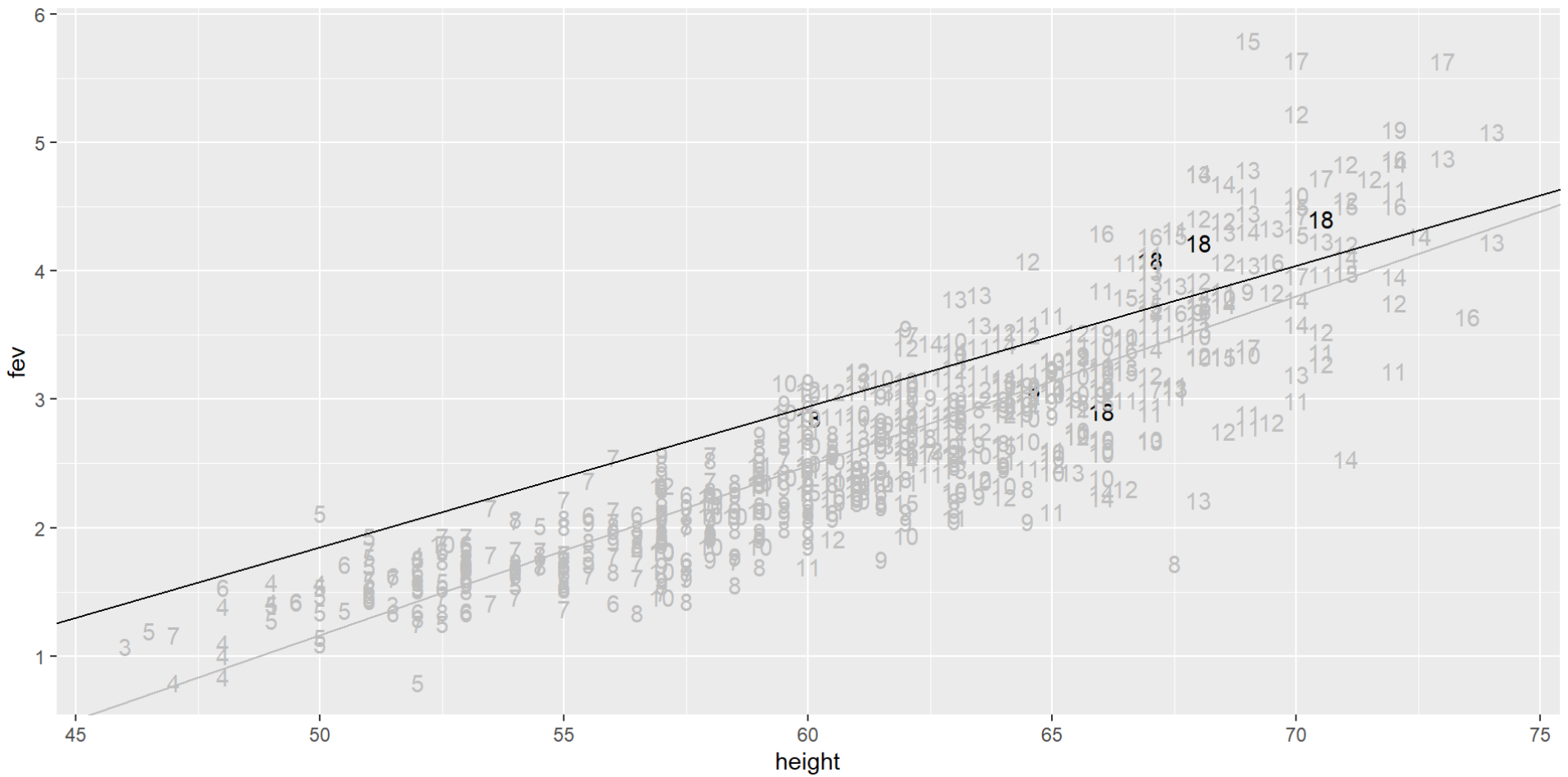
Relationship between height and FEV controlling at Age=16



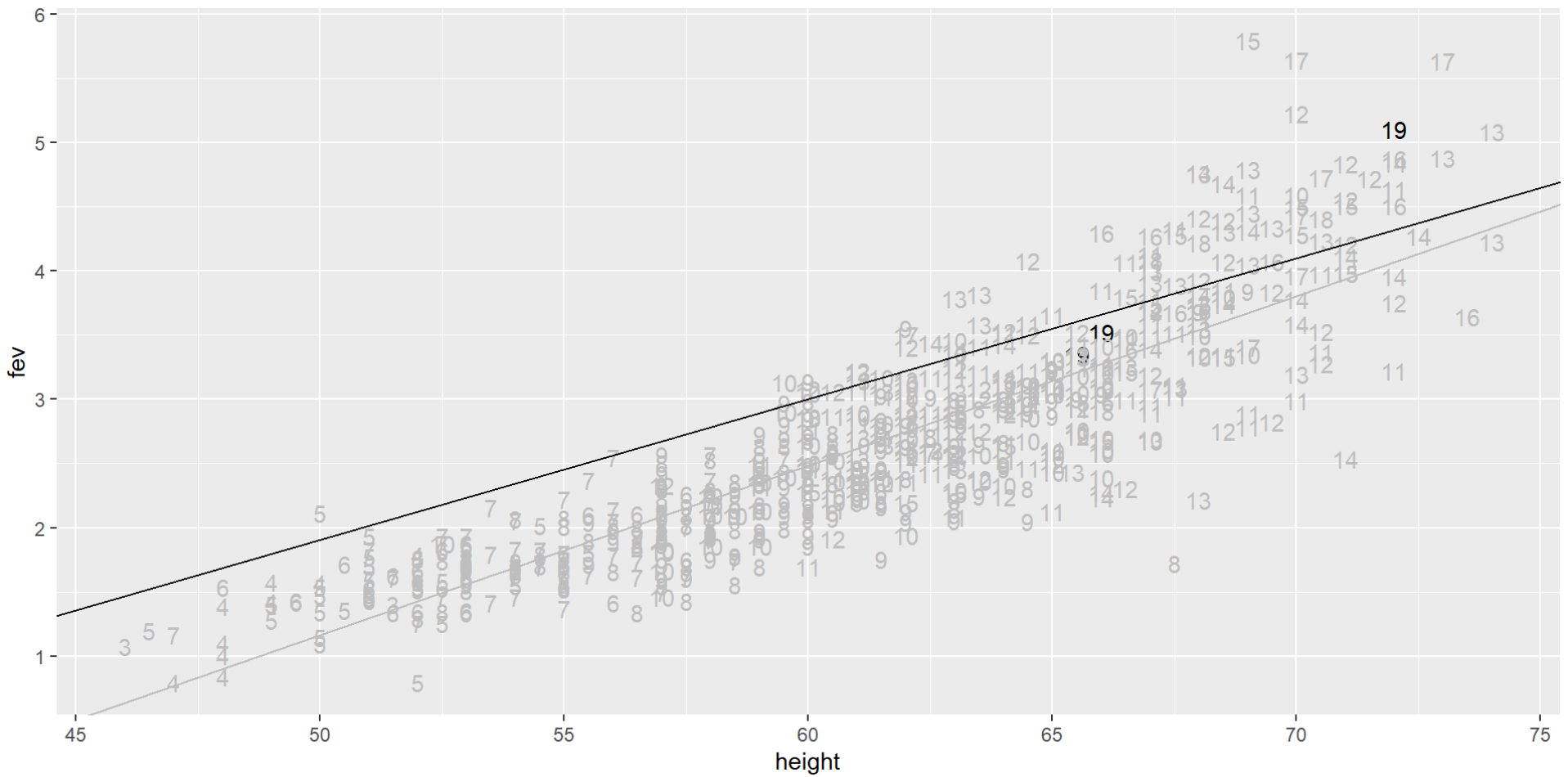
Relationship between height and FEV controlling at Age=17



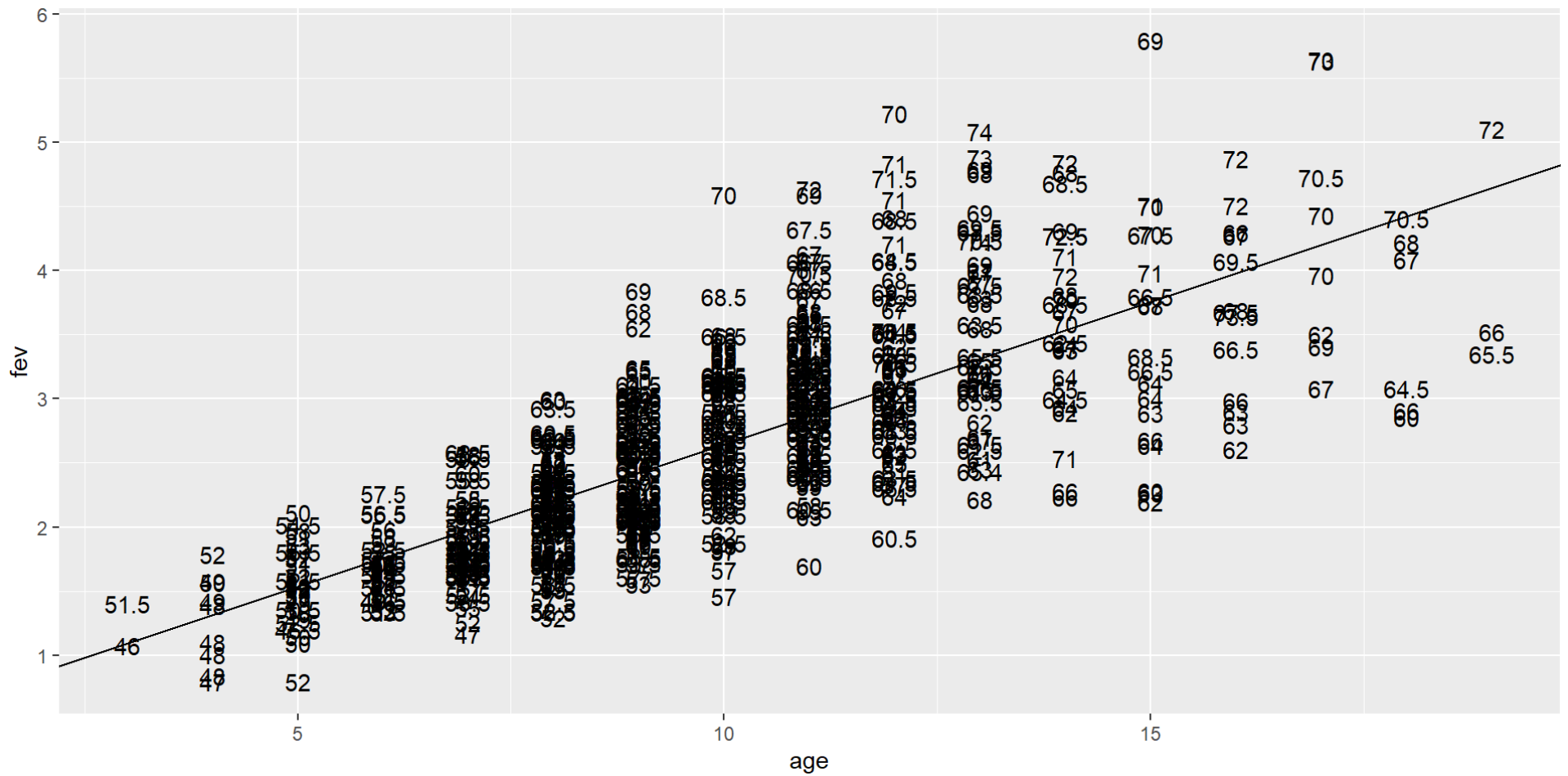
Relationship between height and FEV controlling at Age=18



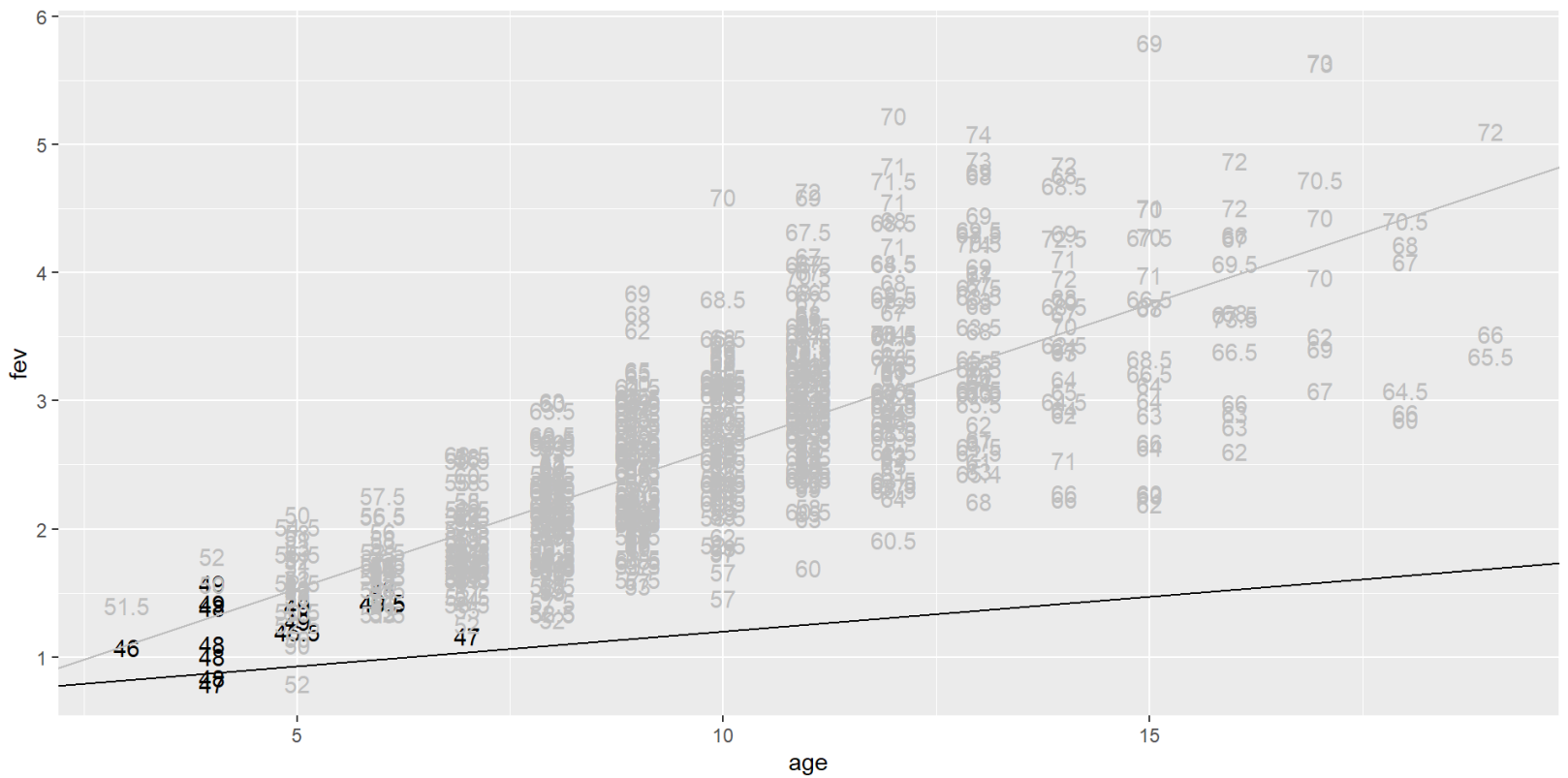
Relationship between height and FEV controlling at Age=19



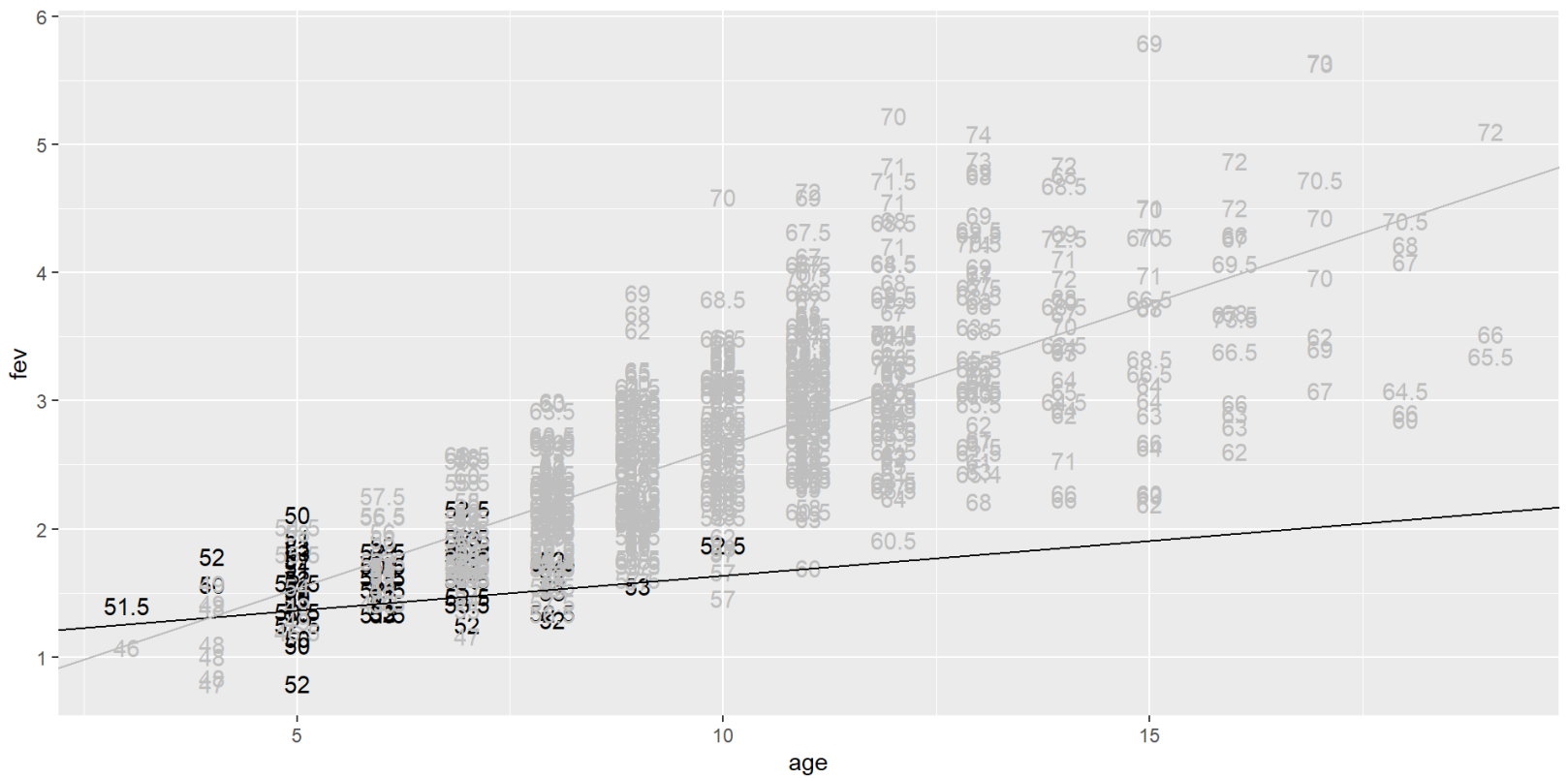
Unadjusted relationship between age and fev



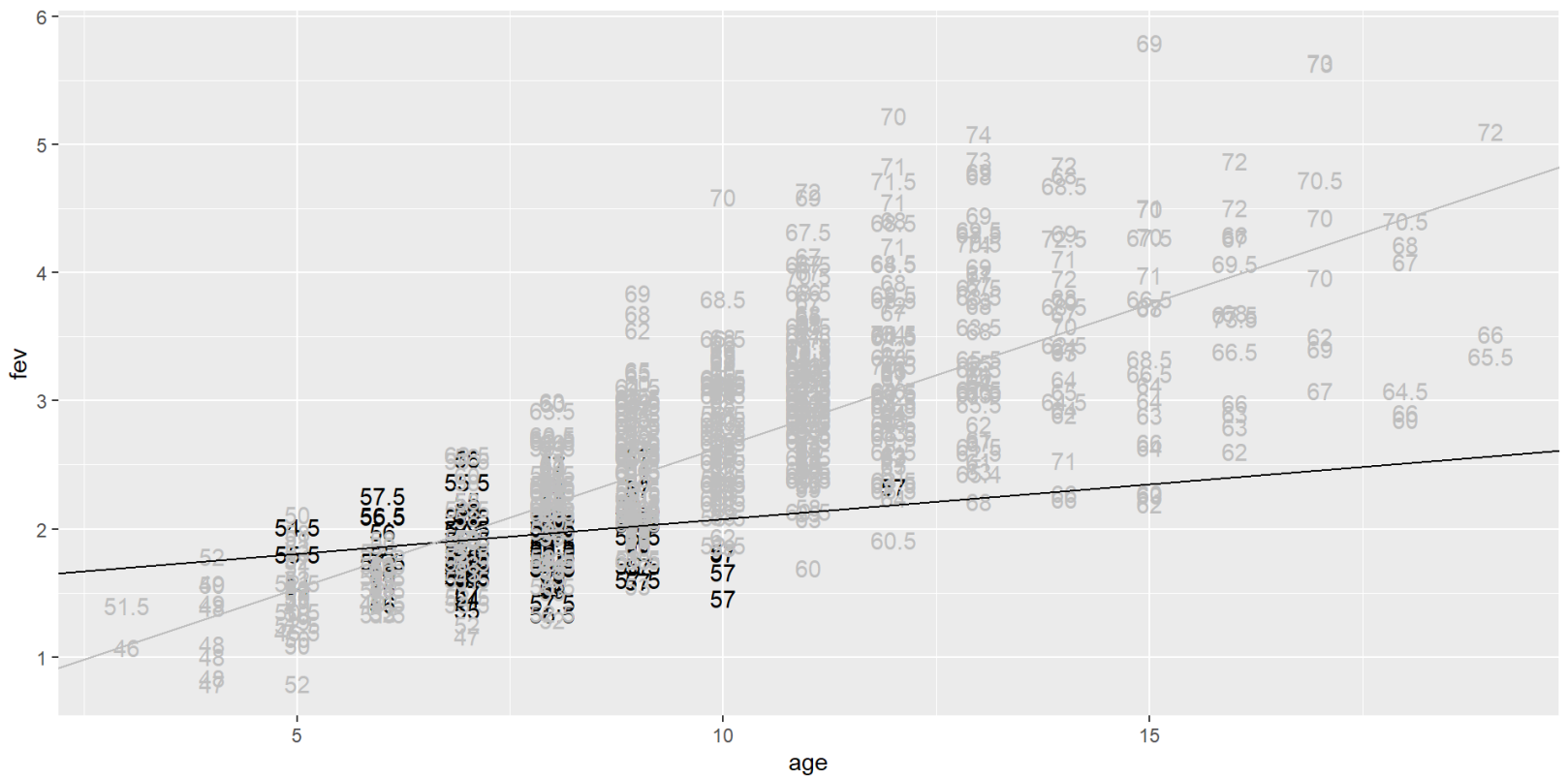
Relationship between age and FEV controlling for height between 46 and 49.5



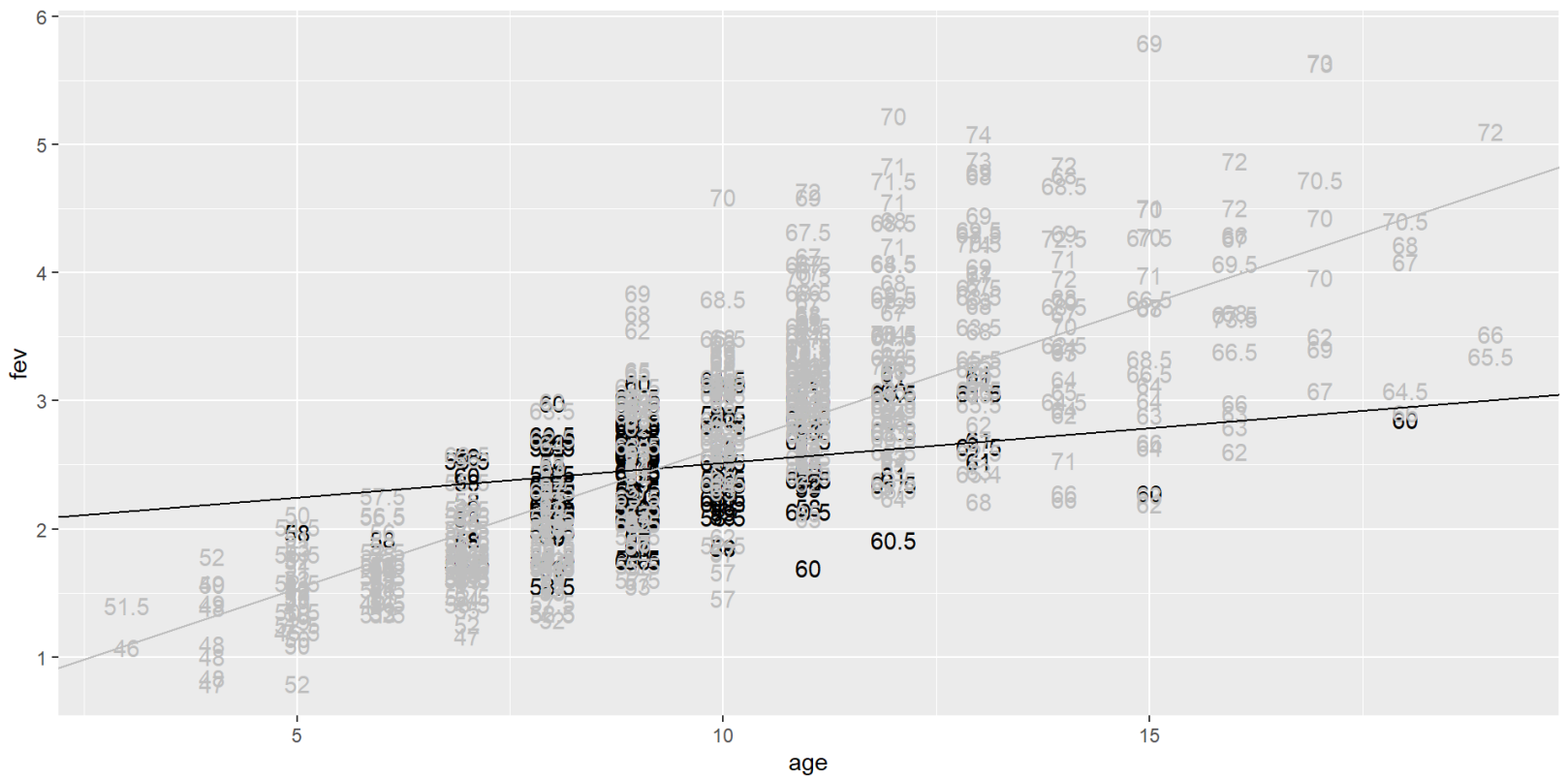
Relationship between age and FEV controlling for height between 50 and 53.5



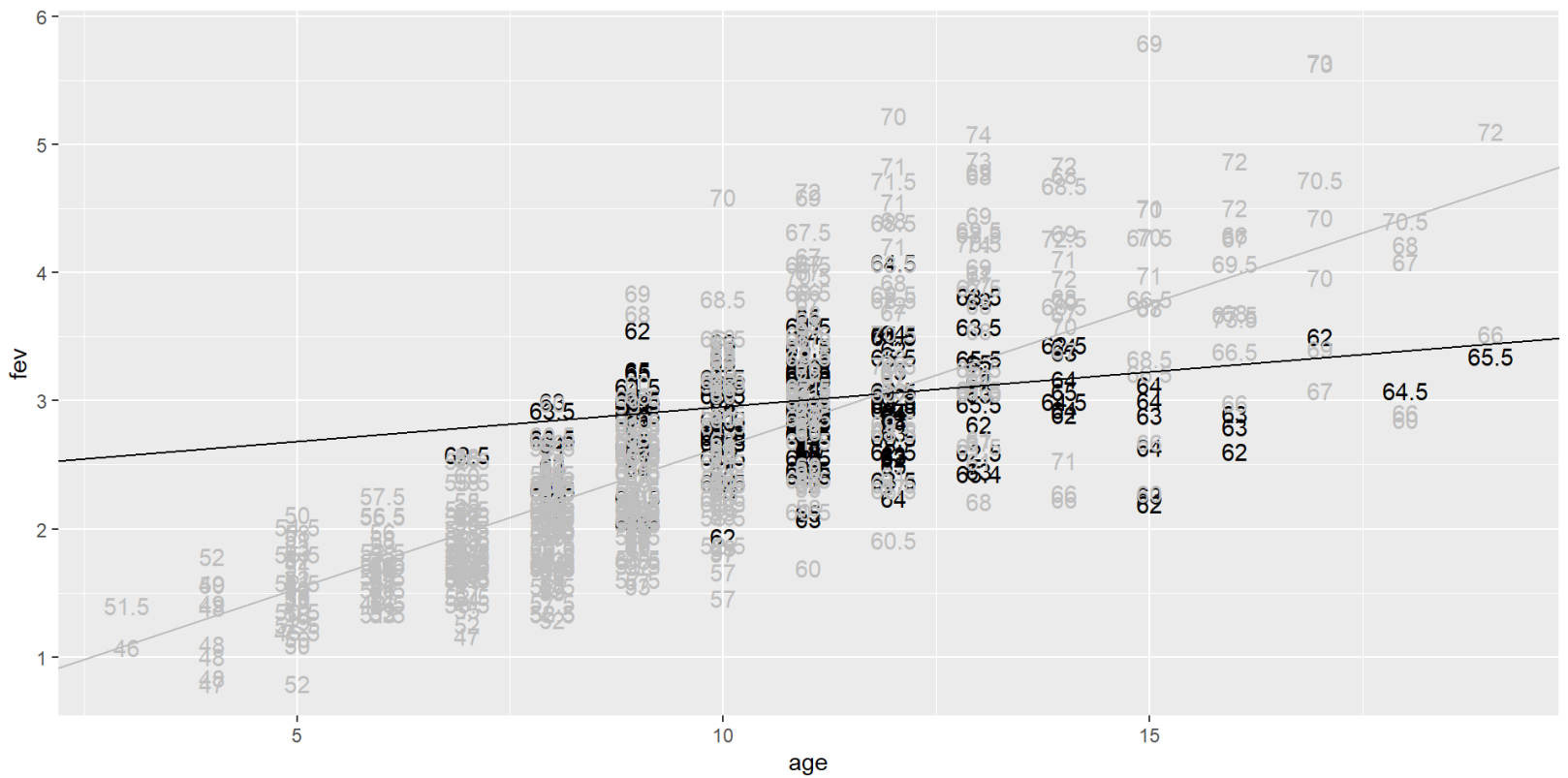
Relationship between age and FEV controlling for height between 54 and 57.5



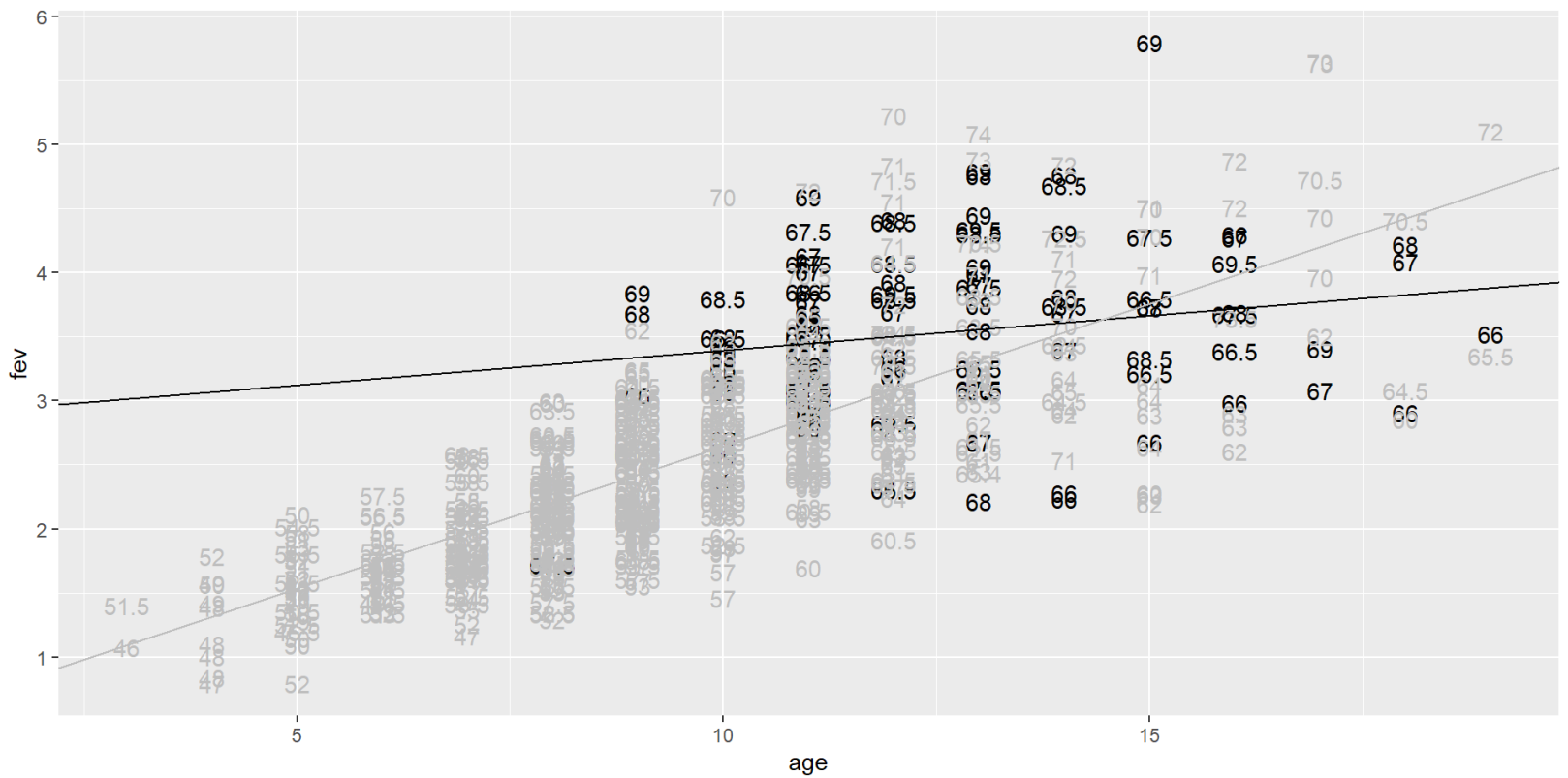
Relationship between age and FEV controlling for height between 58 and 61.5



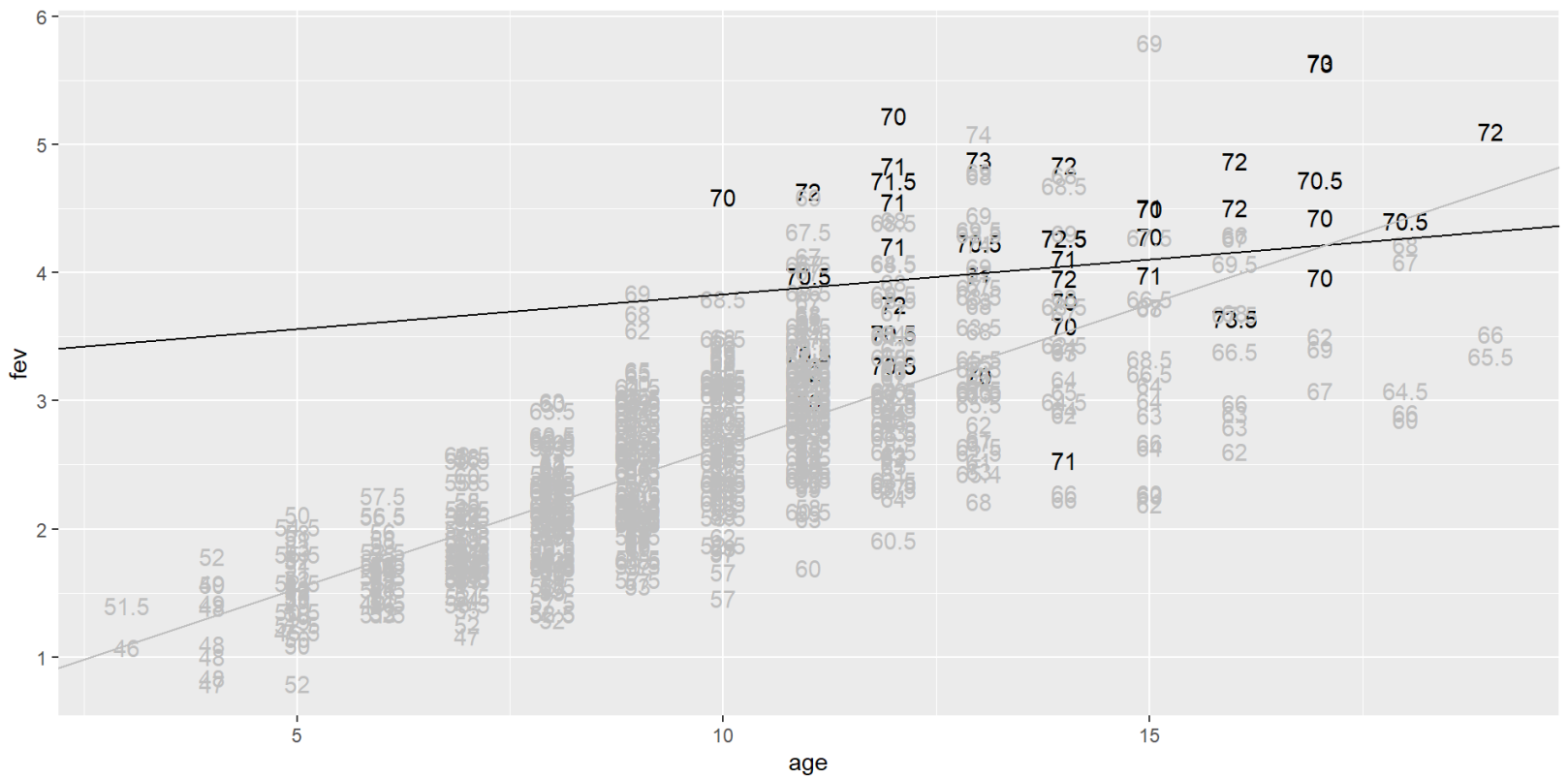
Relationship between age and FEV controlling for height between 62 and 65.5



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



Summary

- What you have learned
  - Interpretation of regression slope and intercept
  - Assumptions in linear regression
  - Calculation of sums of squares
  - Confidence intervals and hypothesis tests
  - Relationship to the correlation coefficient
  - Definition of residuals
  - Diagnostic plots
  - Interpretation with two independent variables

Additional topics??

Speaker notes

- Define Simple Linear Regression in terms of usage
- Define Predictor Variable
- Define Outcome Variable
- Define Line of Best Fit
- Define Slope
- Define Y intercept
- Define residuals
- Explain why the residuals are squared in the regression equation
- Define the Least Squares Method
- Identify the type of data used in a Linear Regression Analysis
- Explain the two main purposes of Linear Regression Analysis
- Identify the two assumptions that must be met for a Linear Regression Analysis to be valid
- Identify three additional assumptions that must be met for a Linear Regression Analysis to be run
- Explain Homoscedasticity of Residuals
- Define the assumption that must be met for a Linear Regression Analysis to be run
- Write the Null and Alternative Hypothesis for a Linear Regression Analysis
- Know the processes to test each assumption for a Linear Regression Analysis
- Define Autocorrelation
- Define the Durbin-Watson Test
- Define the interpretation of the Durbin-Watson Test
- Know the SPSS process for running a Linear Regression Analysis
- Run a Linear Regression Analysis using SPSS
- List all relevant output for a Linear Regression Analysis
- Define R
- Define R squared
- Define adjusted R squared
- Interpret R, R squared and adjusted R squared
- Know the mathematical equation for R squared
- Relate the significance test for ANOVA to Linear Regression Analysis
- Explain the model test for a Linear Regression Analysis
- Write the model equation using the Linear Regression Analysis output

- Write the model equation using the Linear Regression Analysis output
- Define B
- Define constant
- Understand the limits of Linear Regression Analysis
- Correctly and completely write up a Linear Regression Analysis