Lecture 5 - General Linear Model (GLM) I

Caspar J. van Lissa

2023-08-12

General Linear Model

What is the GLM?

A family of models to analyze the relationship between one outcome and one or more predictors

- Today we introduce bivariate linear regression, one member of the family
- Describes a linear relationship between a continuous outcome variable and a continuous predictor
- Other family members can handle:
  - Predictors of any measurement level (continuous or categorical)
  - More than one predictor
  - Transformations of Y
  - Transformations of X
  - Other error distributions than the normal distribution
  - Etc.

Linear regression

Thought experiment…

If I told you that last year's average exam grade was:

$\bar{Y} = 6.1$

What grade would you expect to get for this year's exam?

Thought experiment…

If I additionally told you that hours studied is strongly associated with the exam grade
And you know that you studied far more than average
Does that change your expectation for your grade?

What does this demonstrate?

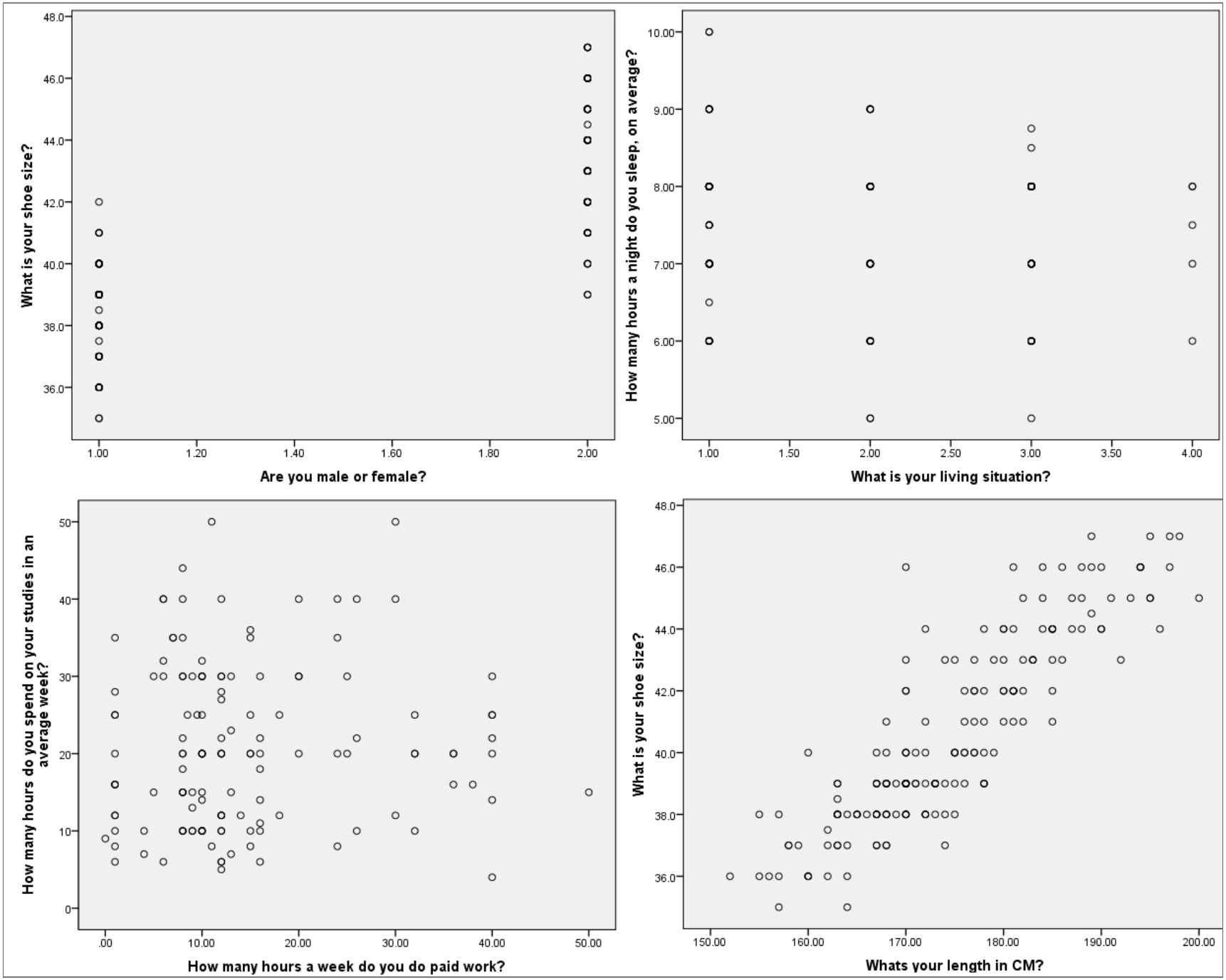
1. The mean is the best predictor (expected value) when there's no further relevant information
2. If you DO have information about other variables that are **associated** with the outcome, you can use that information to improve your predictions

   This is regression.

Scatterplot

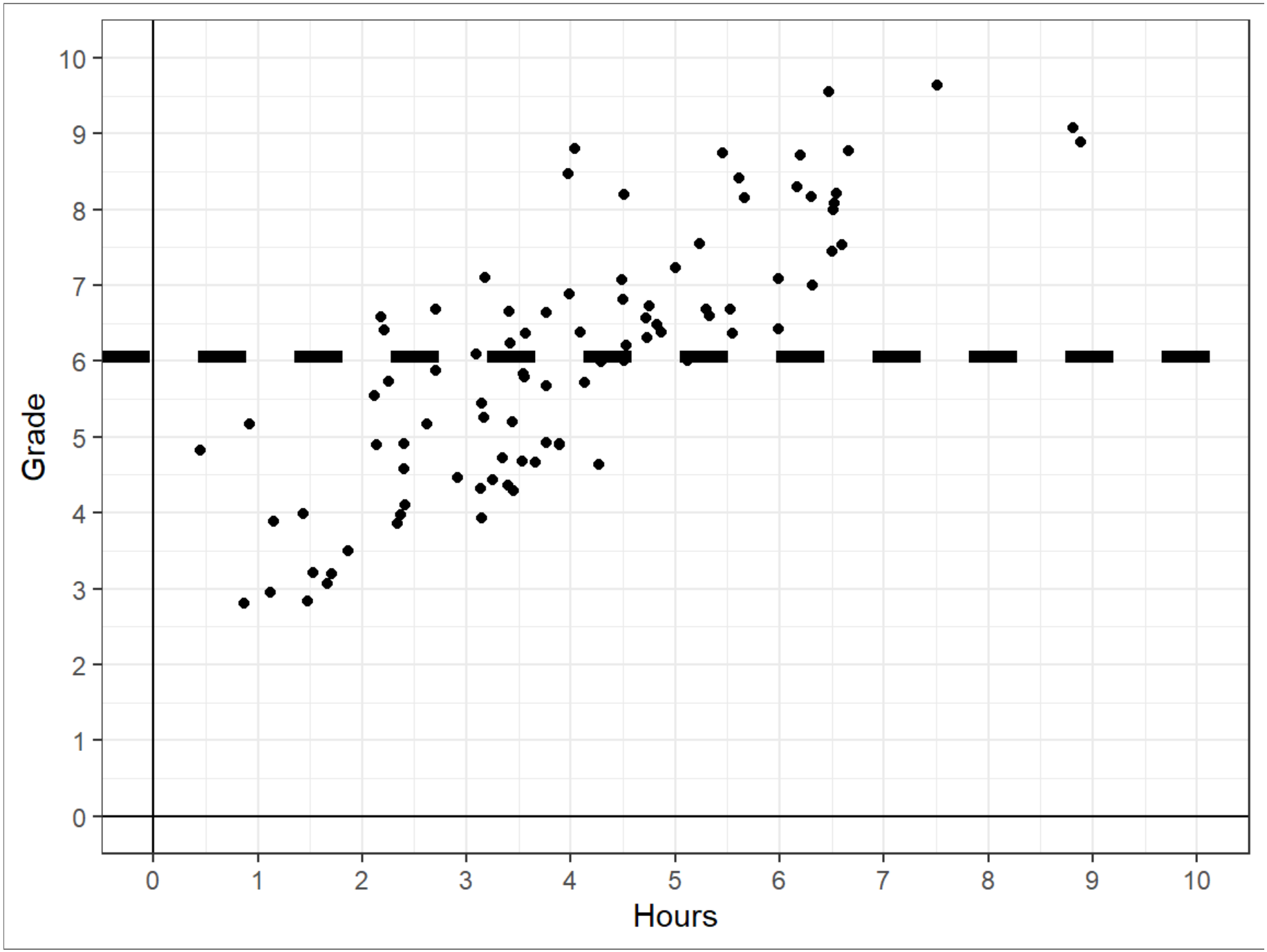
- Visualization for two variables
  - To show associations, both should be at least of ordinal measurement level
- Sex and shoe size?
- Travel time to TiU and monthly phone bill?
- Hours of work and hours of study?
- Hours of study and grade?

Scatterplot examples


What is your shoe size?
48.0
46.0
44.0
42.0
40.0
38.0
36.0
1.00
1.20
1.40
1.60
1.80
2.00
Are you male or female?
How many hours a night do you sleep, on average?
10.00
9.00
8.00
7.00
6.00
5.00
1.00
1.50
2.00
2.50
3.00
3.50
4.00
What is your living situation?
How many hours do you spend on your studies in an average week?
50
40
30
20
10
0
.00
10.00
20.00
30.00
40.00
50.00
How many hours a week do you do paid work?
What is your shoe size?
48.0
46.0
44.0
42.0
40.0
38.0
36.0
150.00
160.00
170.00
180.00
190.00
200.00
Whats your length in CM?

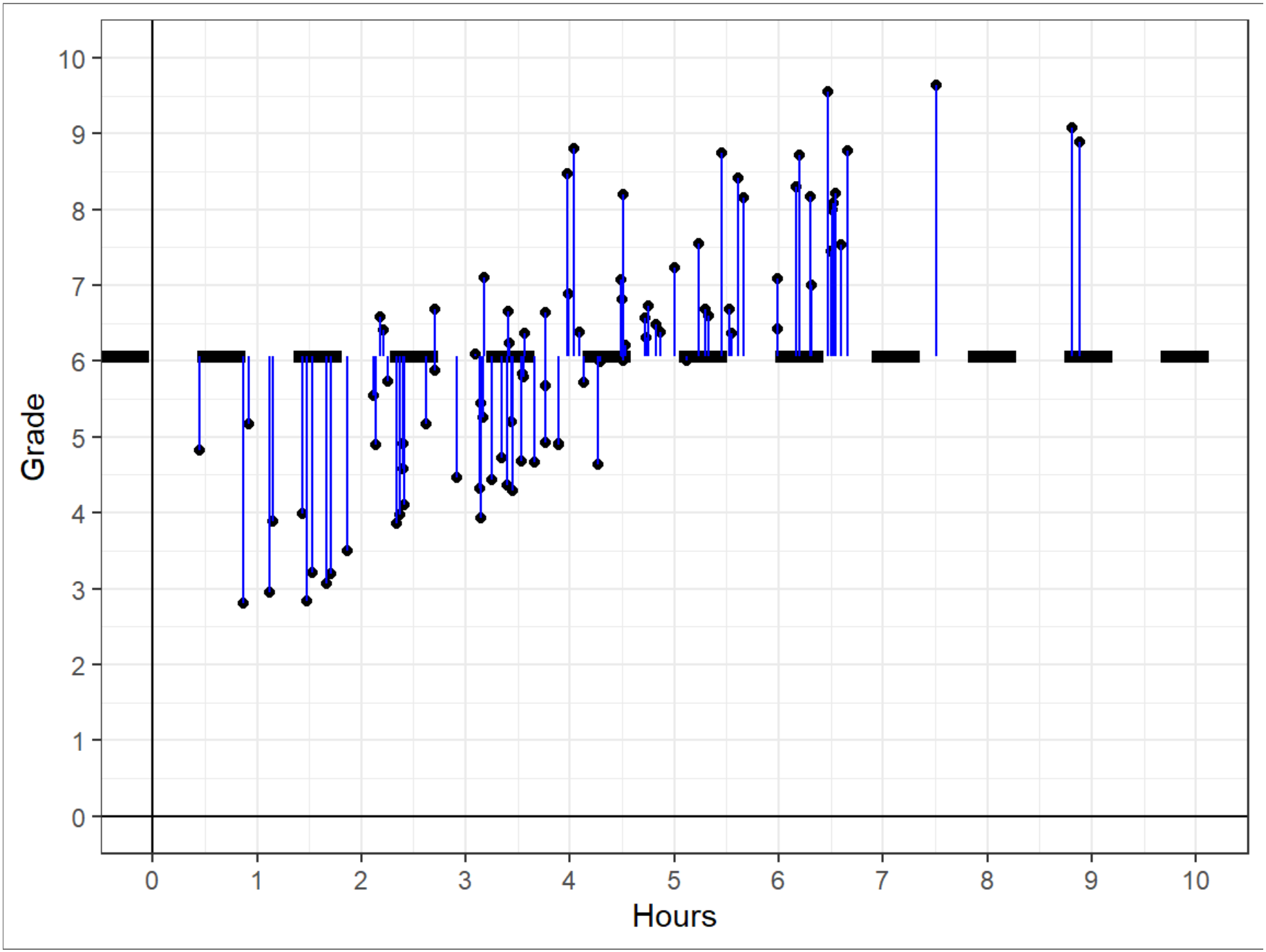
Null model

If there were NO association, the mean $\bar{Y}$ would be the best prediction for each student:


Grade
Hours
0
1
2
3
4
5
6
7
8
9
10

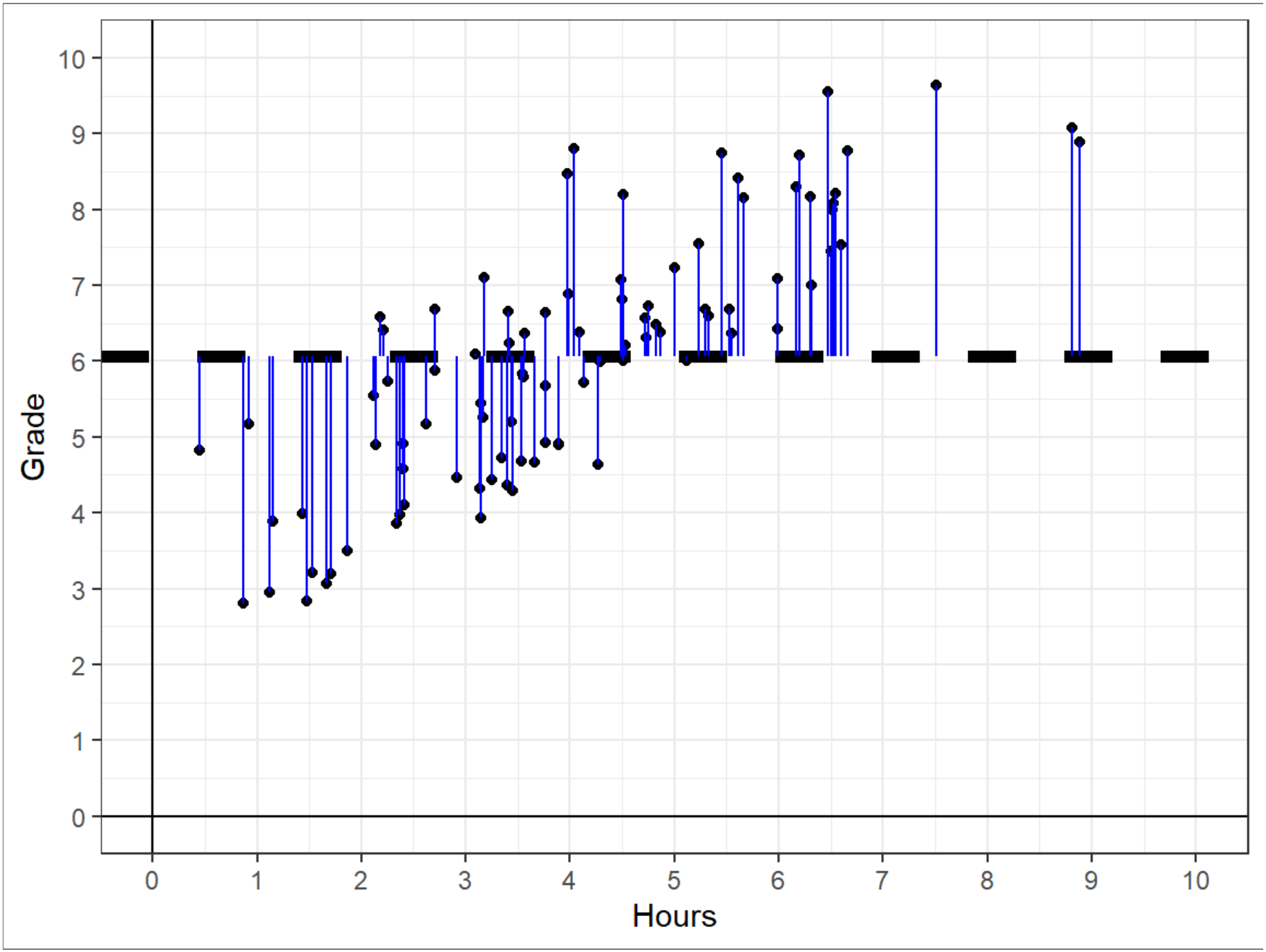
Prediction error

This prediction will be a bit wrong for every student:


Grade
Hours
0
1
2
3
4
5
6
7
8
9
10

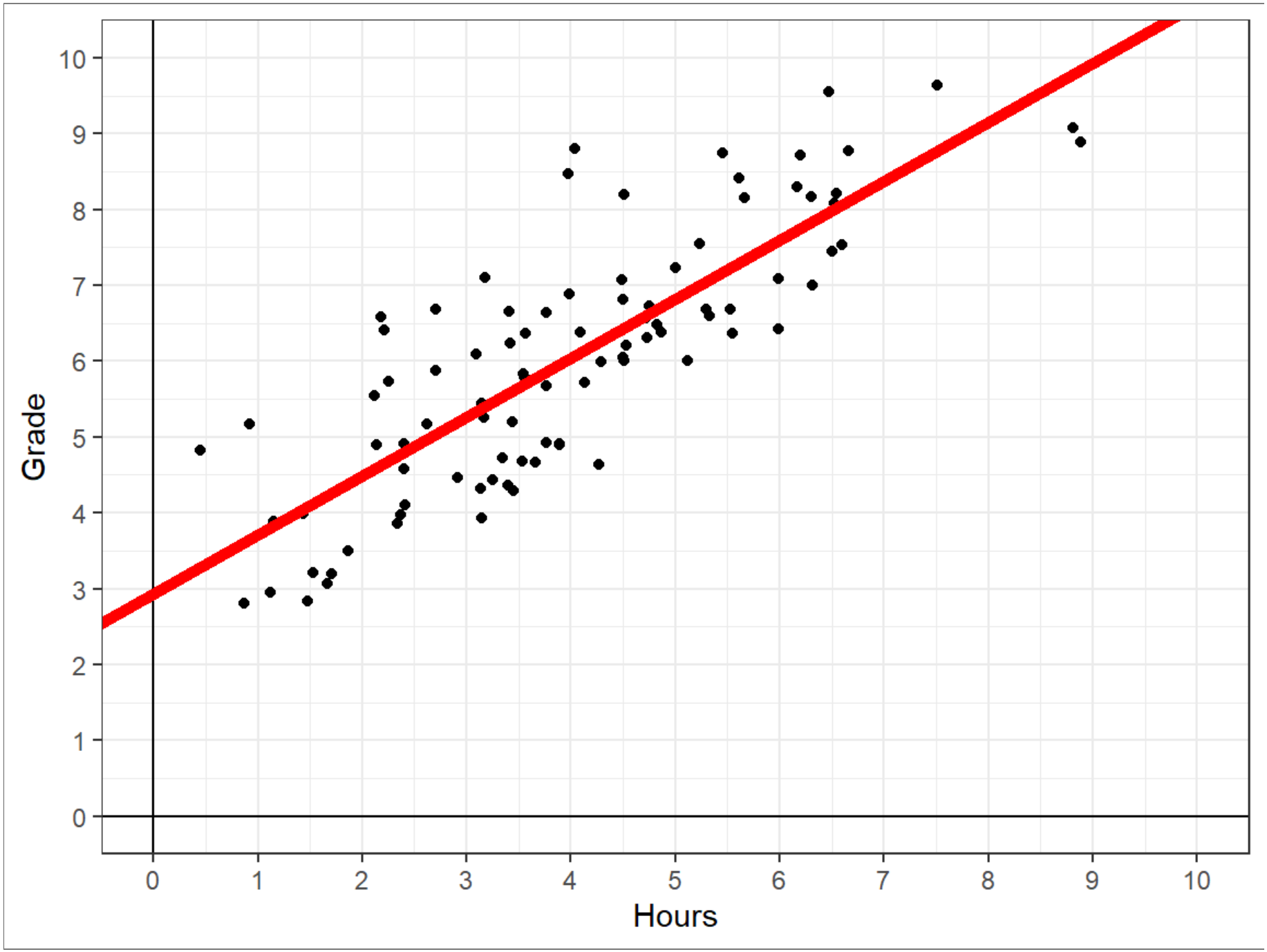
Link with standard deviation

The mean of those prediction errors (squared) is the variance of “grade”
Its square root is the SD of grade


Grade
Hours
0
1
2
3
4
5
6
7
8
9
10

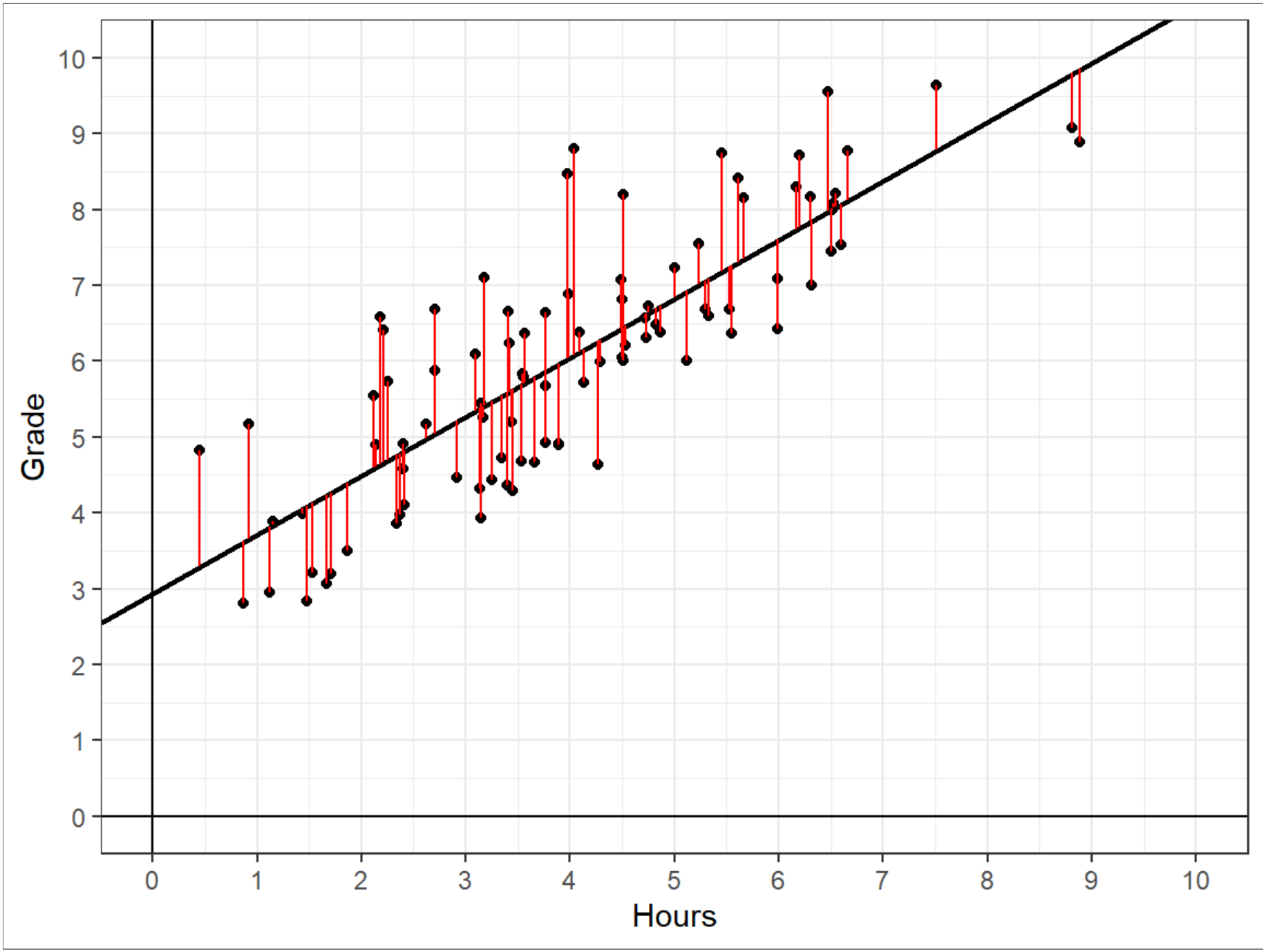
Association

The points appear to follow a diagonal upward line, rather than the straight line of the mean:


Grade
Hours
0
1
2
3
4
5
6
7
8
9
10

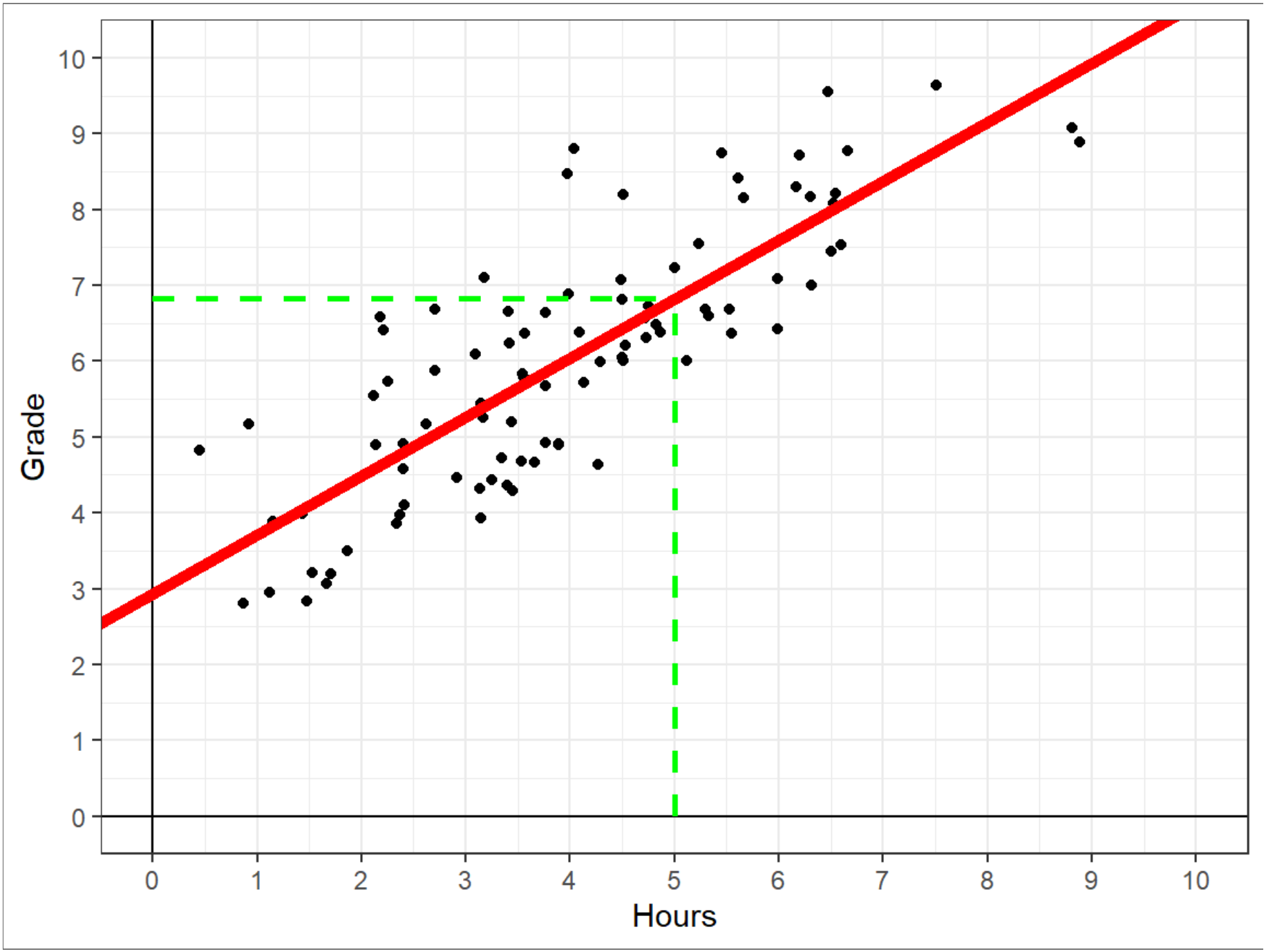
Diagonal line

The distances of points from a diagonal line are obviously smaller than from the straight line of the mean:


Grade
Hours
0
1
2
3
4
5
6
7
8
9
10

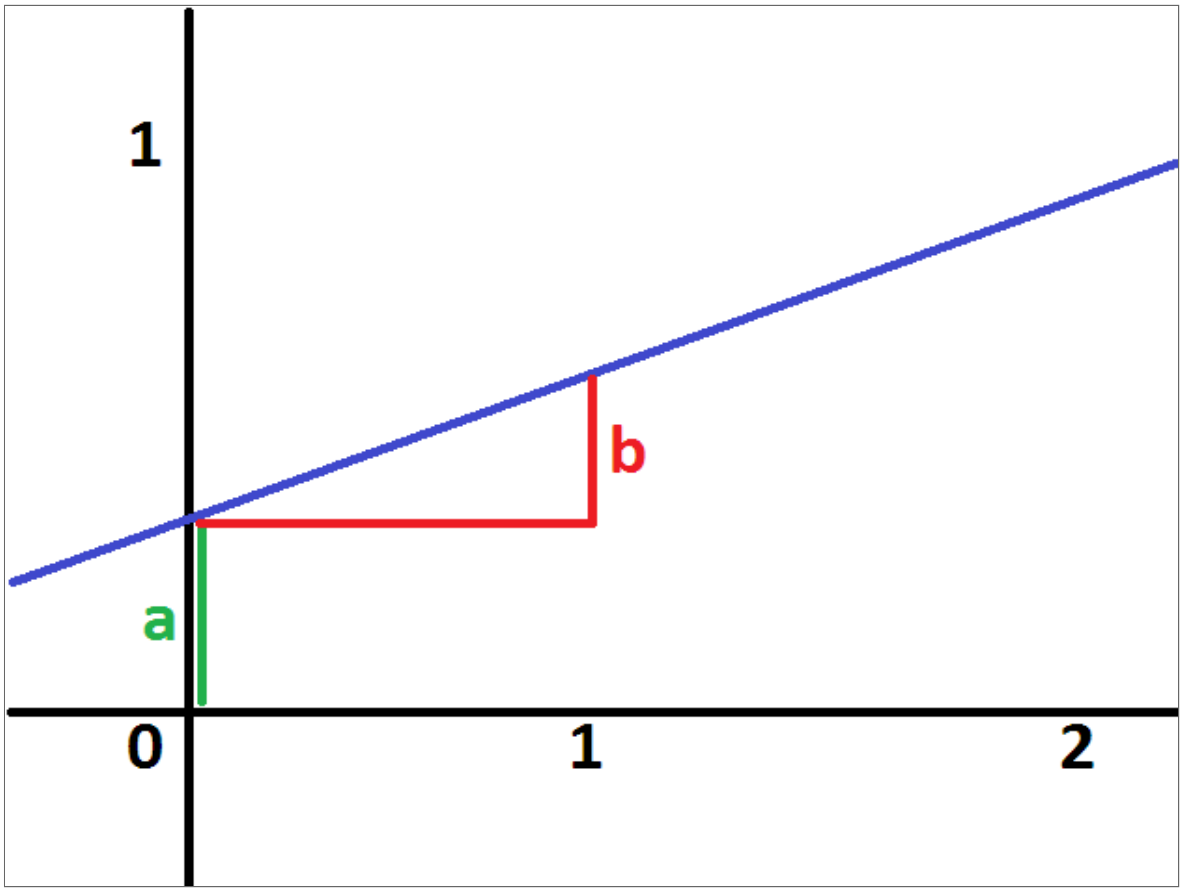
Prediction

By following the line you can kind of guess what grade you might expect for a specific number of hours studied. These predictions are better than just using the mean:


Grade
Hours
0
1
2
3
4
5
6
7
8
9
10

Formula

As you might remember from high school, a diagonal line is described by:
$Y = a + bX$



Coefficients

The formula for a line is:
$Y = a + bX$
a is the intercept, where the line crosses the Y-axis

- This is the predicted value when X equals 0

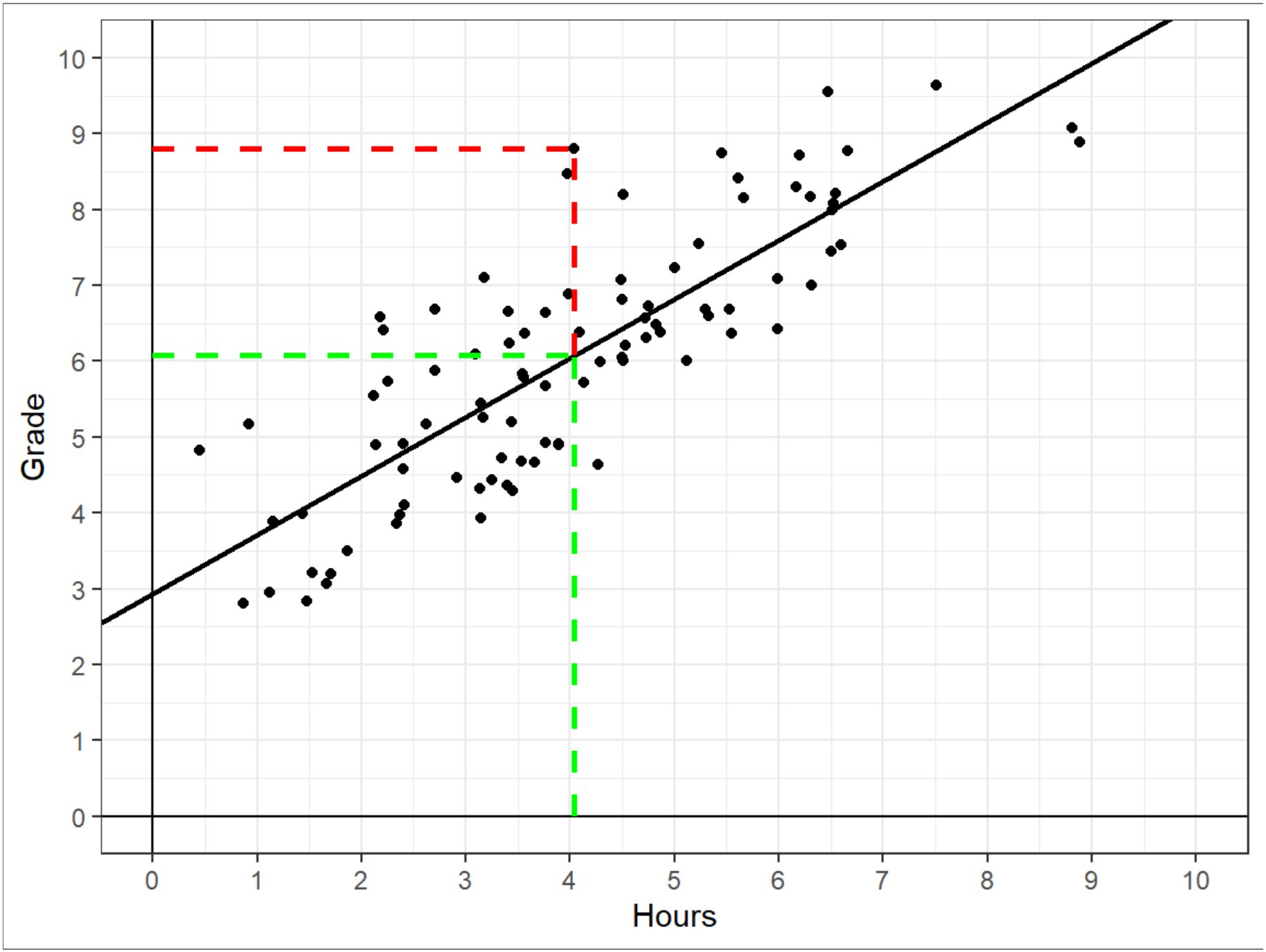
b is the slope, how steeply the line in/decreases

- Y increases by b when X increases by 1

Prediction error

We can use the line to predict values of Y for individuals $_i$

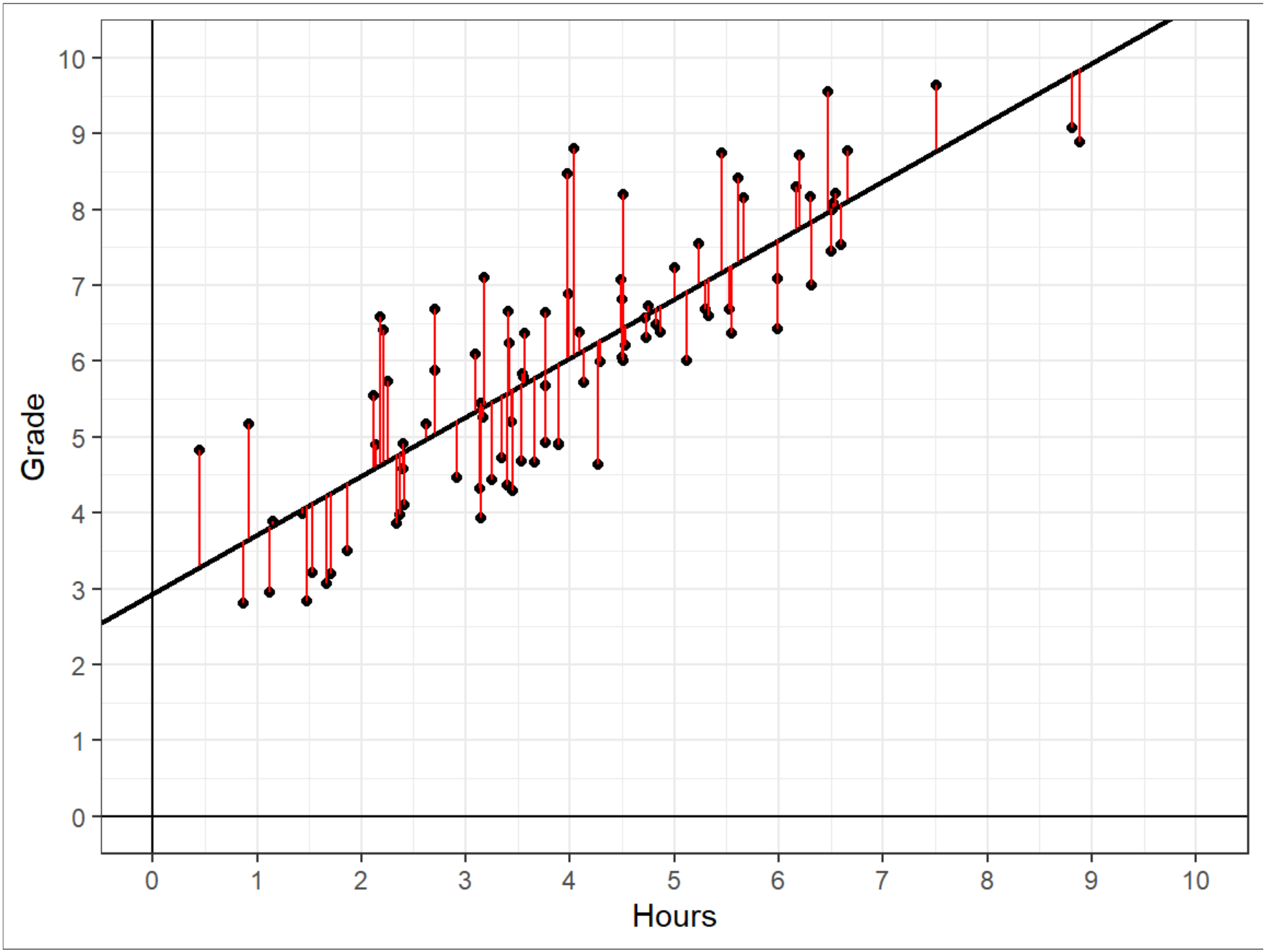
- This prediction $\hat{Y}_i$ is never identical to the observed value of that individual, Y_i
- There is always some *prediction error*, $Y_i - \hat{Y}_i$


Grade
Hours
0
1
2
3
4
5
6
7
8
9
10

Ordinary Least Squares

We want to obtain the line that gives us the best possible predictions

- There exist values of a and b that give us a line with the best possible predictions
- These can be calculated using matrix algebra (not part of this course)
- The resulting line goes *exactly through the middle* of the cloud of datapoints, which is why $e_i \sim N(\mathbf{0}, \hat{\sigma}_e^2)$
- This is called *ordinary least squares* regression
  - Will be explained further, but squares refers to squared prediction errors


Grade
Hours
0
1
2
3
4
5
6
7
8
9
10

Numeric example

Substituting numeric values for the coefficients, the function to predict grade based on hours is:

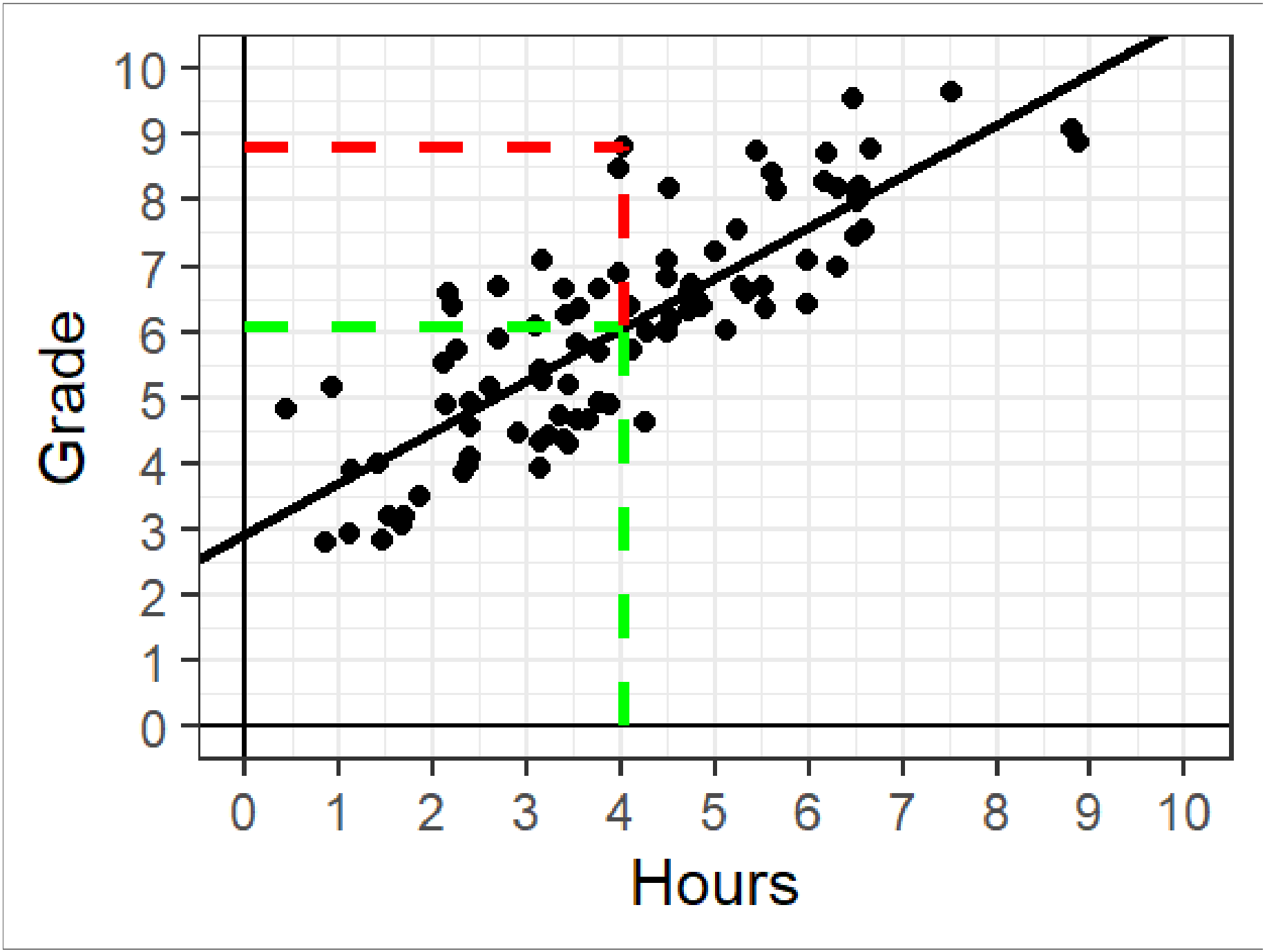
$\hat{Y}_i = 2.9 + 0.8 * X_i$

Student 71 studies 4 hours, so the predicted grade $\hat{Y}_{71}$ is:

$\hat{Y}_{71} = 2.9 + 0.8 * 4 = 6.1$

In reality student 71' grade was 8.8, so the prediction error was

$Y_i - \hat{Y}_i = 8.8 - 6.1 = 2.7$


Grade
Hours
0
1
2
3
4
5
6
7
8
9
10

Complete regression formula

The formula $Y = a + bX$ describes the diagonal line
It does not yet describe the prediction error
The *linear regression* model expands the formula to include prediction error
$Y_i = a + b * X_i + e_i$
e_i refers to the individual prediction error

- We assume prediction errors are normally distributed around the regression line
  - $e_i \sim N(0, \hat{\sigma}_e^2)$

Bivariate regression formula

$Y_i = a + b * X_i + e_i$

Symbol	Interpretation
Y_i	Individual i's score on dependent variable Y
a	Coefficient, intercept of the regression line
b	Coefficient, slope of the regression line
X_i	Individual i's score on independent variable X
e_i	Individual i's prediction error

Bivariate regression formula

$Y_i = a + b * X_i + e_i$

In words, this formula says:

“The individual values on variable Y are equal to the intercept, plus the slope times the individual values on the predictor X, plus individual prediction error.”

Other notations

$Y_i = a + b * X_i + e_i$

Symbol	Interpretation
Y_i	Outcome, dependent variable (DV)
a	b_0, β_0
b	b_1, β_1
X_i	Predictor, independent variable (IV)
e_i	e_i

Observed and predicted value

$Y_i = a + b * X_i + e_i$

“The individual values on variable Y are equal to the intercept, plus the slope times the individual values on the predictor X, plus individual prediction error.”

And also:

“The individual values on variable Y are equal to the predicted values, plus individual prediction errors”

$Y_i = \hat{Y}_i + e_i$

The predicted value is the value on the regression line:

$\hat{Y}_i = a + b * X_i$

Coefficients

Testing coefficients

You can perform hypothesis tests on the coefficients a and b

- We use the t-test from last week
- Most software by default uses a two-sided test with $H_0 : \beta = 0$
- But you can perform custom tests if you want

Testing coefficients

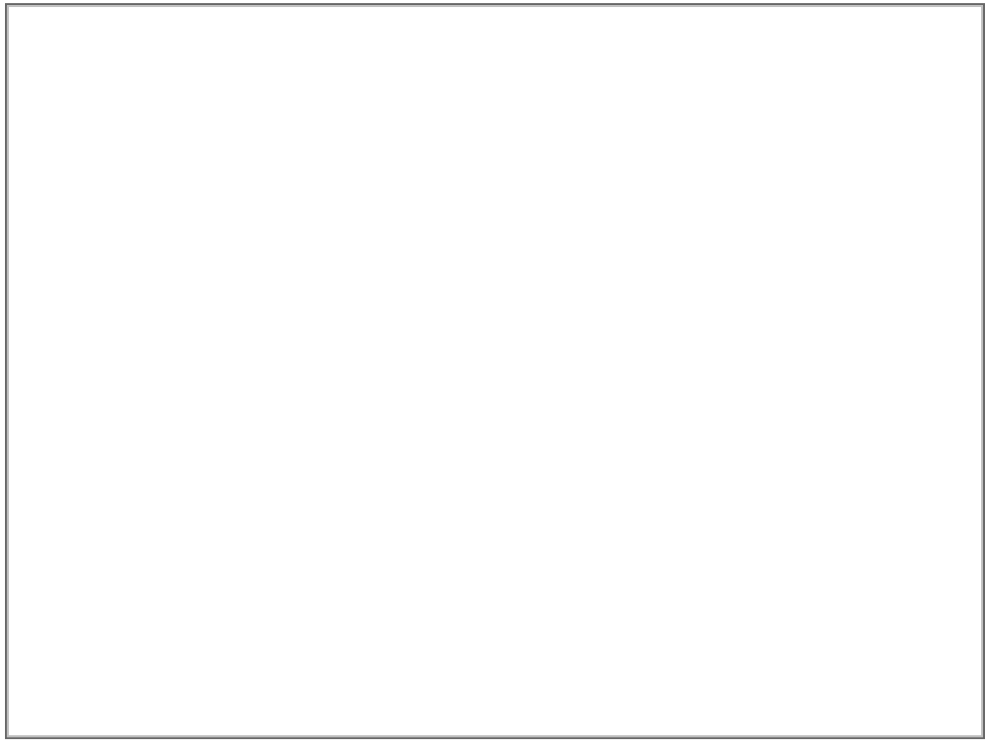
Remember: hypotheses are statements about the population, so we use symbols for population parameters

- Intercept is zero: $H_0 : \beta_0 = 0$
- Slope is zero: $H_0 : \beta_1 = 0$

  We use the t-distribution because we typically don't know population variance of X or Y
- Account for additional uncertainty, as explained last week
- For samples of $n > 30$, this is approximately the same as the Z-distribution

  $t = \frac{b}{SE_b}$
- df: n - p, where p is the number of parameters (2: intercept & slope)

Visualization



Testing coefficients SPSS

- This tests the intercept and slope
- Default null-hypothesis for the intercept is $H_0 : \beta_0 = 0$
  - Is the intercept significantly different from zero? No.
- Default null-hypothesis for the intercept is $H_0 : \beta_1 = 0$
  - Is the slope significantly different from zero? Yes.

Custom tests using SPSS

Let's conduct a one-sided hypothesis, $H_0 : \beta_1 \leq 0$

- Our true belief, in this case, is that the effect will be positive, $H_a : \beta_1 > 0$
- If the effect is in the direction of H_a, we can divide the p-value by 2 for a one-sided test
- $p = .001/2 = .0005$ which we round to $p < .001$, so no practical difference

---

Custom tests using SPSS

Or you might wonder: if I would study 0 hours, should I expect a passing grade? $H_0 : \beta_0 \leq 5.5$

- $t = \frac{b_0 - \beta_0}{SE_b} = \frac{1.31 - 5.5}{.97} = -4.32$
- The effect is not in the direction of H_a; instead, it's congruent with H_0, so we will never reject it

---

Reporting

The effect of hours studied on exam grade was significant, $b = 0.78, t(90) = 13.55, p < .001$. This means that for every additional hour studied, the expected grade increased by 0.78 points.

Assumptions

Assumptions

- A model is only valid if its assumptions are met
- Otherwise it misrepresents the data
  - Tests are misleading
  - Inferences not justified
- We can (try) to check assumptions
  - But no evidence of violation is not the same as evidence of no violation

Assumption checks

Dictionary definition: “something that you accept as true without question or proof”

- Assumptions are statements about the population
  - So we never really know if they’re true
- Assumption checks are limited to the sample
- Tailoring your analysis based on assumption checks in the sample risks overfitting the model

Assumptions of linear regression

1. Model is correctly specified, which includes
    - Linearity of relationship between X and Y
    - Normality of residuals
    - Direction of causality (if you want to interpret your model causally)
2. Homoscedasticity
    - Residuals are equally distributed for all values of the predictor
3. Independence of observations

Linearity

- Visual check
  - Scatterplot: do the points follow a straight line?

Linearity 2

- Residual plots

Violations of linearity

- Curvilear
- Outlier

Normal residuals

Why are residuals normally distributed?

- Because they are the result of many random processes

https://www.youtube.com/watch?v=6YDHBFVIvIs&feature=youtu.be&t=6

Assessing normality

- Histogram, “Normal P-P plot” or “Q-Q plot”
- Kolmogorov-Smirnov / Shapiro-Wilkes test

Normality: Visual inspection

Homoscedasticity

- Distribution of residuals ϵ_i
- equal variance for all predicted scores
- Residual plot:
  - Y: *standardized* residuals
  - X: *standardized* predicted values

Homoscedasticity plots

Independent observations

- Scores should be independent of one another
  - Every observation conveys unique information
- If your observations are dependent, your “effective sample size” is lower
  - This causes you to underestimate uncertainty about estimates
- Assumption is satisfied with simple random sampling.

Independent observations

- Examples of violations:
  - Cheating on an exam
  - Married couples
  - Children within the same class (same teacher, background, etc)
- Dependent observations are more similar than randomly sampled ones
- So each conveys slightly less *unique information*
  - E.g., 10 couples who are highly similar to each other
  - If you compute standard errors by dividing by $\sqrt{20}$, they will be too small
  - You will underestimate the uncertainty of your estimates
- If you know *why* cases are dependent, there are solutions (not part of this course)

Dealing with assumption violations

- Linear effects
  - Transform variable (square, square root)
  - Include quadratic term
- Normality of residuals
  - Increase sample size
  - Use different outcome distribution (e.g., binomial)
  - Use non-parametric approach
  - Remove outliers
- Homoscedasticity
  - Account for source of heteroscedasticity
  - Use non-parametric approach
- Independent observations
  - Account for group membership (not part of course)