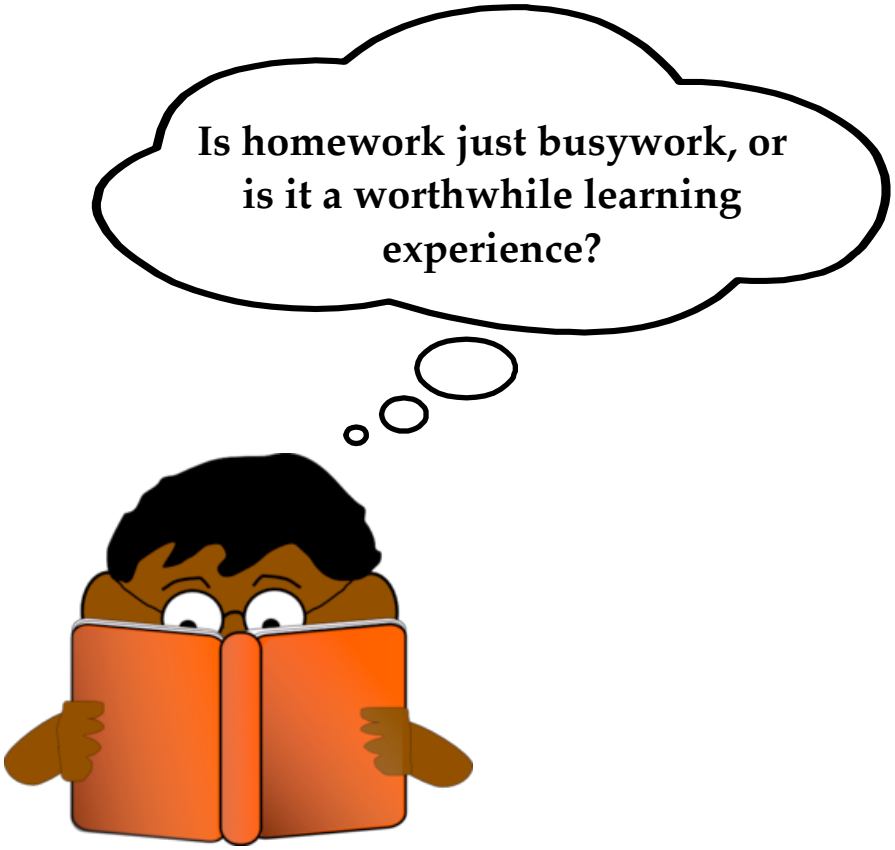
Simple Linear Regression



Data

- Eighth graders were asked to consider their mathematics homework over the last month, and respond to the survey question of approximately *how many hours of time they spend doing mathematics homework per week*.
- Scores on a standardized mathematics achievement test were also collected. (The test has a national mean score of 50 and a standard deviation of 10.)

Keith, T. Z. (2006). *Multiple regression and beyond*. Boston: Pearson.

Read in Data and Load Libraries

```
# Load the data (homework-achievement.csv)
> math = read.csv("EPSY-8262/data/homework-achievement.csv")

# Load libraries; Note: you may need to install them first
> library(ggplot2)
> library(psych)
> library(sm)

# The ggplot2 library will let us use the ggplot() function
# The psych library will let us use the describe() function
# The sm library will let us use the sm.density() function
```

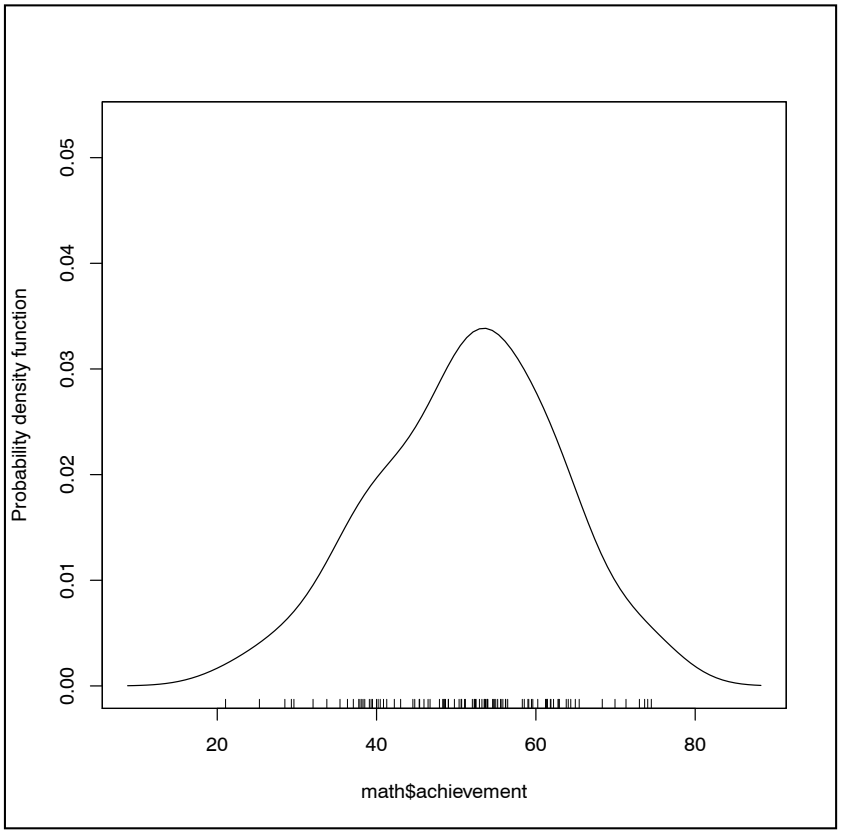
Examine the Data

```
> head(math)

  homework achievement
1        2          54
2        0          53
3        4          53
4        0          56
5        2          59
6        0          30

> tail(math)

    homework achievement
95         6          50
96         4          48
97         3          58
98         2          39
99         2          41
100        1          51
```

Examine the Outcome

```
> sm.density(math$achievement)

> describe(math$achievement)

  vars   n  mean    sd median trimmed   mad min max range  skew kurtosis   se
1    1 100 51.41 11.29     53   51.65 11.86  22  75    53 -0.22     -0.3 1.13
```

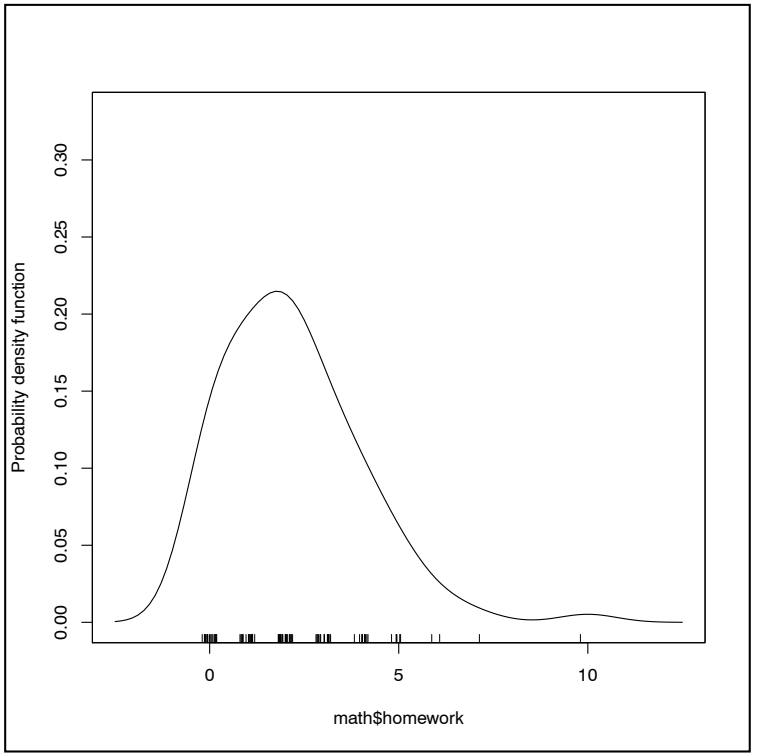
The marginal distribution of mathematics achievement scores is unimodal with a mean of 51. There is variation in these scores ($SD = 11$).



Examine the Predictor

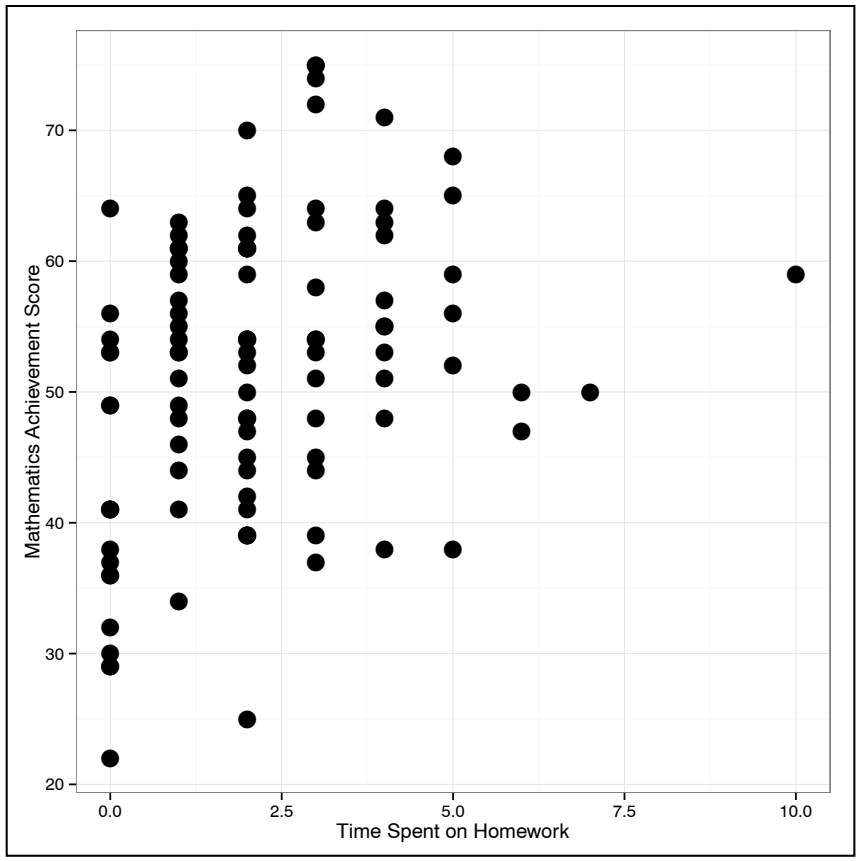
```
> sm.density(math$homework)

> describe(math$homework)

  vars   n mean   sd median trimmed  mad min max range skew kurtosis   se
1    1 100  2.2 1.81      2    2.01 1.48   0  10    10 1.11     2.19 0.18
```

The marginal distribution of time spent on homework is right-skewed with a mean of 2.2. There is a great deal of variation in the measurements ($SD = 1.8$).



Examining the Distribution of the Outcome Conditioned on the Predictor

```
> ggplot(data = math, aes(x = homework, y = achievement)) +
    geom_point() +
    xlab("Time Spent on Homework") +
    ylab("Mathematics Achievement Score") +
    theme_bw()
```



The plot suggests a relationship (in the *sample*) between time spent on mathematics homework and mathematics achievement scores.

- Functional form of the relationship?
- Direction?
- Strength?
- Weird observations?

Correlation

We use the `cor()` function to find the correlation. We give it an indexed data frame, `math[`*rows*`, `*columns*`]`, where *rows* is empty (all rows) and *columns* gives the names of the variables we want the correlation between.

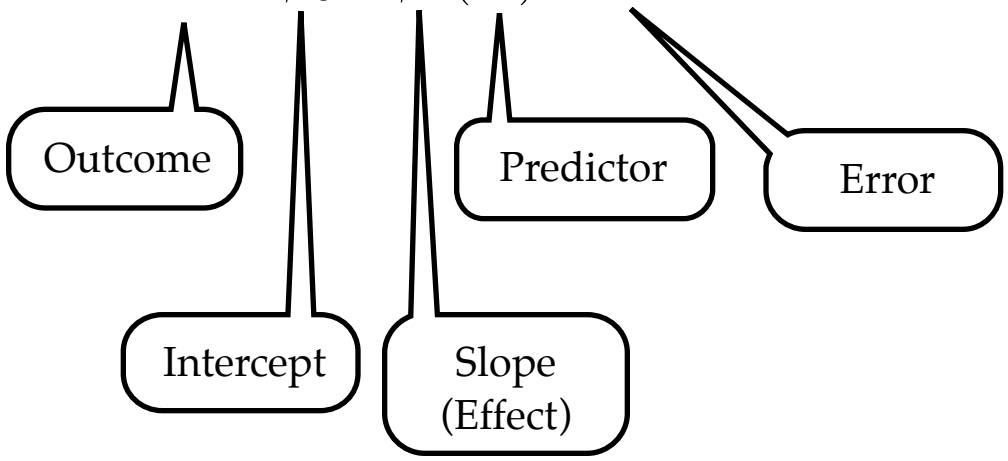
```
> cor(math[ , c("homework", "achievement")]

             homework achievement
homework    1.0000000   0.3199936
achievement 0.3199936   1.0000000
```

The Pearson correlation between time spent on mathematics homework and mathematics achievement suggests a moderate relationship between the variables, $r = 0.32$.

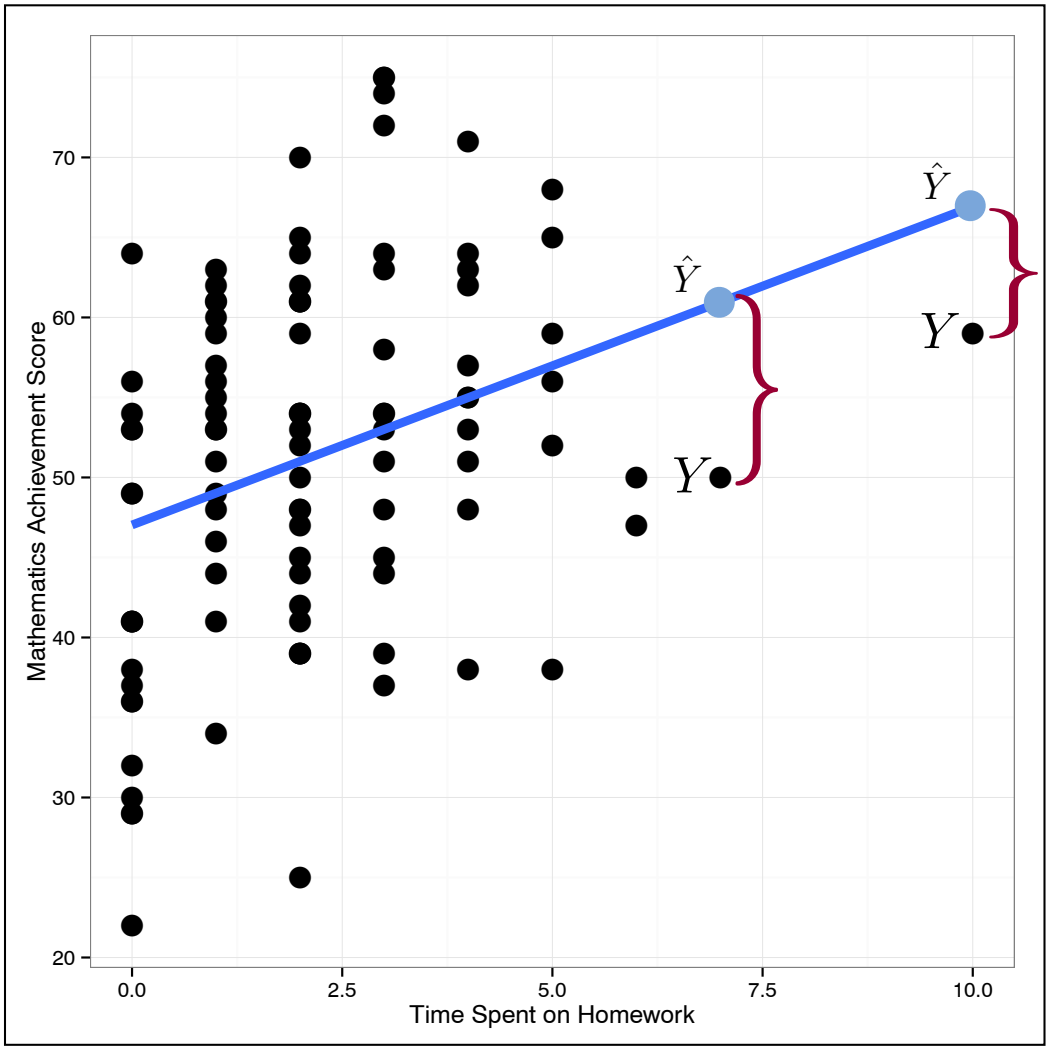
Fit a Model to the Data

Since the relationship's functional form seems reasonably linear, we will fit a linear model to the data.

$$Y = \beta_0 + \beta_1(X) + \epsilon$$



Why an Error Term?



We use a *single line* to describe the relationship between homework and achievement for *all* of the observations in the sample.

- The error allows for discrepancy between the line (predicted Y) and the observed Y
- For some observations the discrepancy is smaller than for others.

Regression (Fitted) Equation

The regression equation is the *systematic* part of the model that is fixed (the same) for all observations.

$$Y = \beta_0 + \beta_1(X) + \epsilon$$

Systematic (fixed) — Random (stochastic)

$$Y = \hat{Y} + \epsilon$$

$$\hat{Y} = \beta_0 + \beta_1(X)$$

One goal of regression analysis is to estimate the values of the regression parameters (i.e., intercept and slope).

Sample Estimates

The regression equation describes the linear relationship *in the population*.

$$\hat{Y} = \beta_0 + \beta_1(X)$$

We will use a *sample of data* (not the entire population) to approximate the parameters in this equation. The values we get for the intercept and slope are *estimates*.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1(X)$$

Hats are used to denote that the value is an estimate. Synonymously, a hat means *predicted value*.

Fitting the Regression Model Using R

```
> lm.a = lm(achievement ~ homework, data = math)
> lm.a

Call:
lm(formula = achievement ~ homework, data = math)

Coefficients:
(Intercept)     homework
      47.03         1.99
```

$$\hat{\beta}_0 = 47.0$$

$$\hat{\beta}_1 = 2.0$$

$$\hat{\text{achievement}} = 47.0 + 2.0(\text{homework})$$

Interpreting the Intercept

$$\widehat{\text{achievement}} = 47.0 + 2.0(\text{homework})$$

$$\hat{\beta}_0 = 47.0$$

The y-intercept gives the y-value where the line passes through the y-axis. It gives the predicted value of y when $x = 0$.

$$\widehat{\text{achievement}} = 47.0 + 2.0(0)$$

$$\widehat{\text{achievement}} = 47.0$$

If a student spends 0 hours per week on mathematics homework, we would predict that student to have a mathematics achievement score of 47.0.

Interpreting the Slope

$$\widehat{\text{achievement}} = 47.0 + 2.0(\text{homework})$$

$$\hat{\beta}_1 = 2.0$$

The *slope* describes the *predicted* change in y relative to the change in x.

$$\frac{\Delta \hat{Y}}{\Delta X} = \frac{2.0}{1}$$

Each one-unit difference in x is associated with a two-unit predicted difference in y.

For each additional hour spent on mathematics homework per week, we predict, *on average*, a 2.0-point difference in mathematics achievement score.

Consider three students…one who spends 2 hours per week on mathematics homework, one who spends 3 hours per week on mathematics homework, and another who spends 4 hours per week on mathematics homework.

2-hours per week

$$\hat{\text{achievement}} = 47 + 2.0(2) = 51$$

3-hours per week

$$\hat{\text{achievement}} = 47 + 2.0(3) = 53$$

4-hours per week

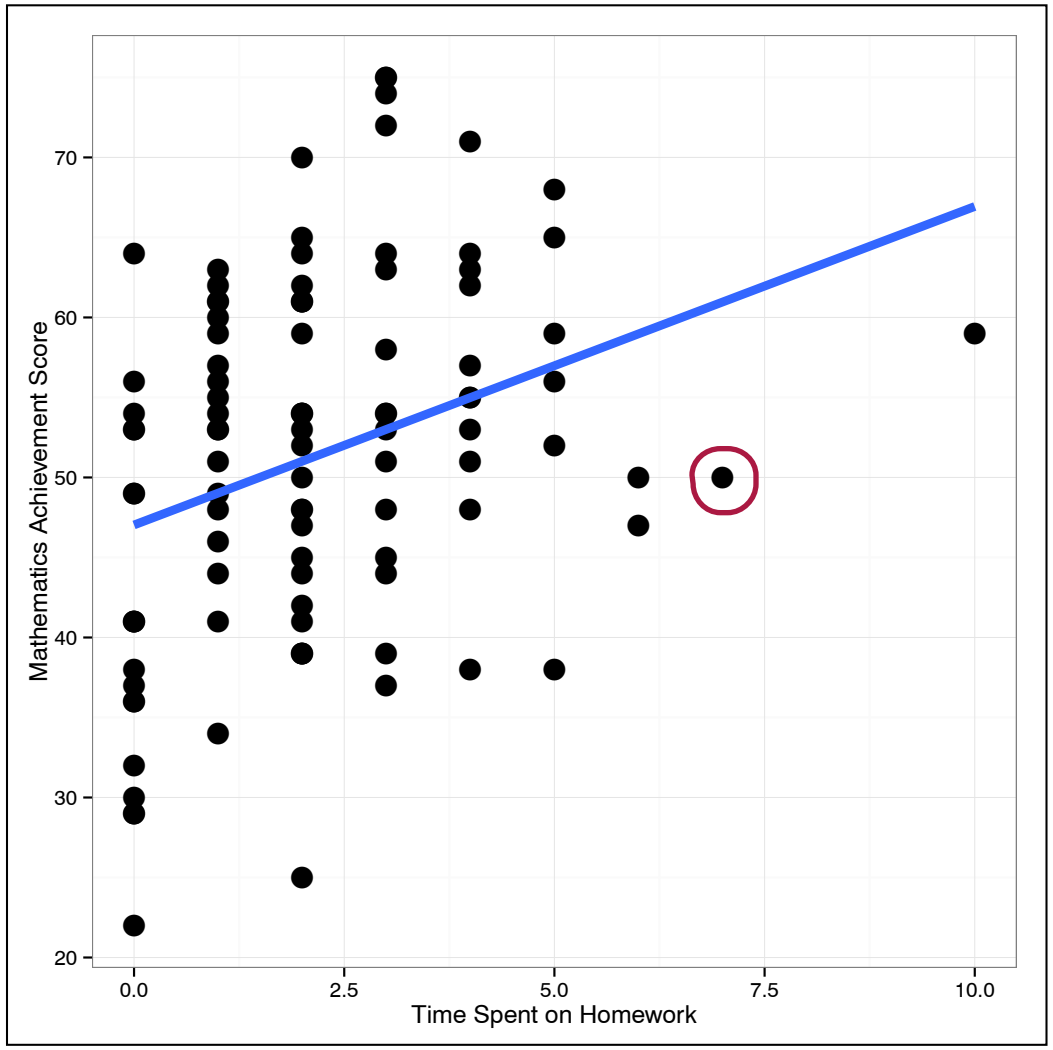
$$\hat{\text{achievement}} = 47 + 2.0(4) = 55$$

Each student's X-value differs by 1. The difference in predicted Y is 2.0.

Each difference of one hour spent on mathematics homework per week, is associated with a 2.0-point difference in mathematics achievement score, *on average*.

Observation, Prediction, and Error

$$\hat{\text{achievement}} = 47.0 + 2.0(\text{homework})$$



Observation 12

$$x = 7$$
$$y = 50$$

The student's observed mathematics achievement score is 50.

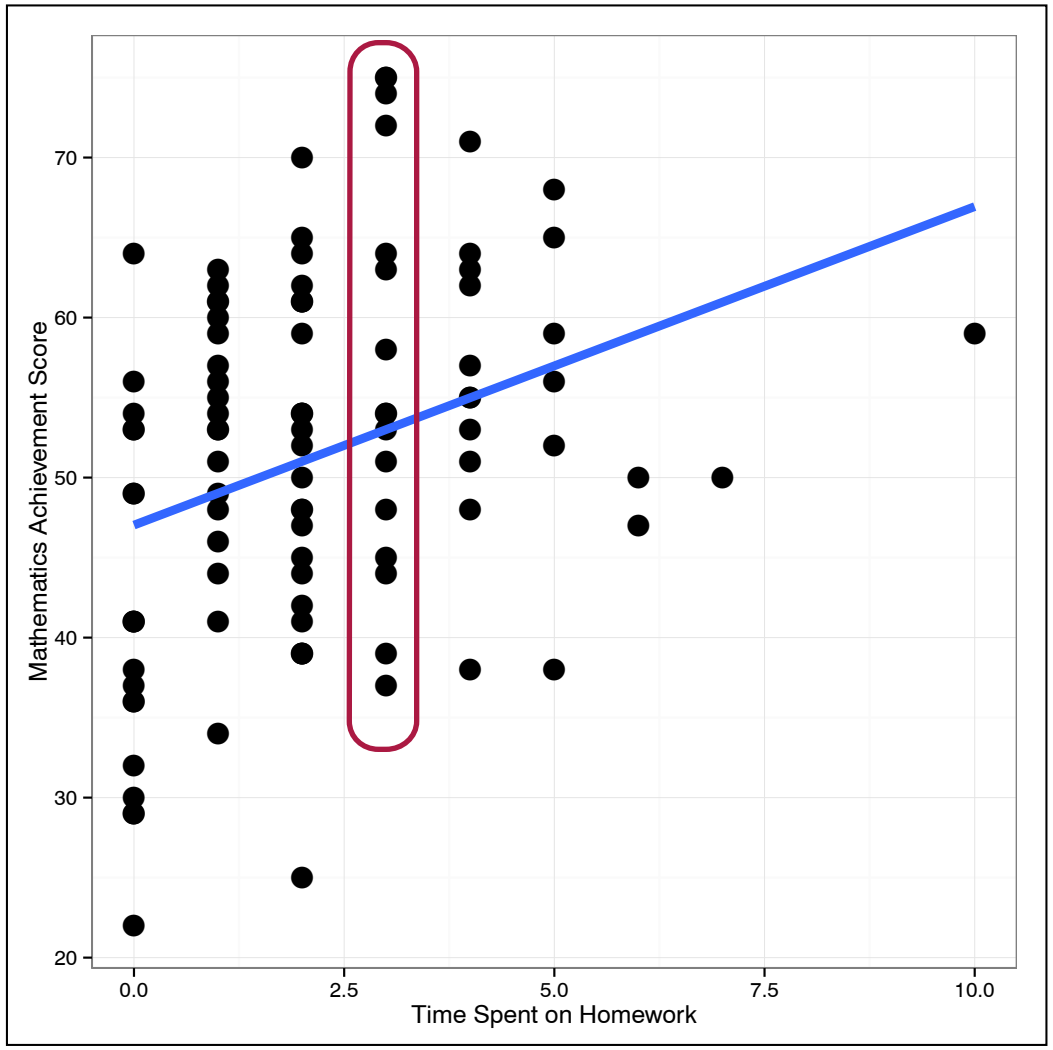
$$\hat{\text{achievement}} = 47 + 2.0(7)$$
$$= 61$$

The student's predicted mathematics achievement score is 61.

$$\hat{\epsilon} = 50 - 61$$
$$= -11$$

The student's observed mathematics achievement is 11 points lower than his/her predicted mathematics achievement score.

$$\hat{\text{achievement}} = 47.0 + 2.0(\text{homework})$$



All of these student's have the same X-value ($x = 3$)

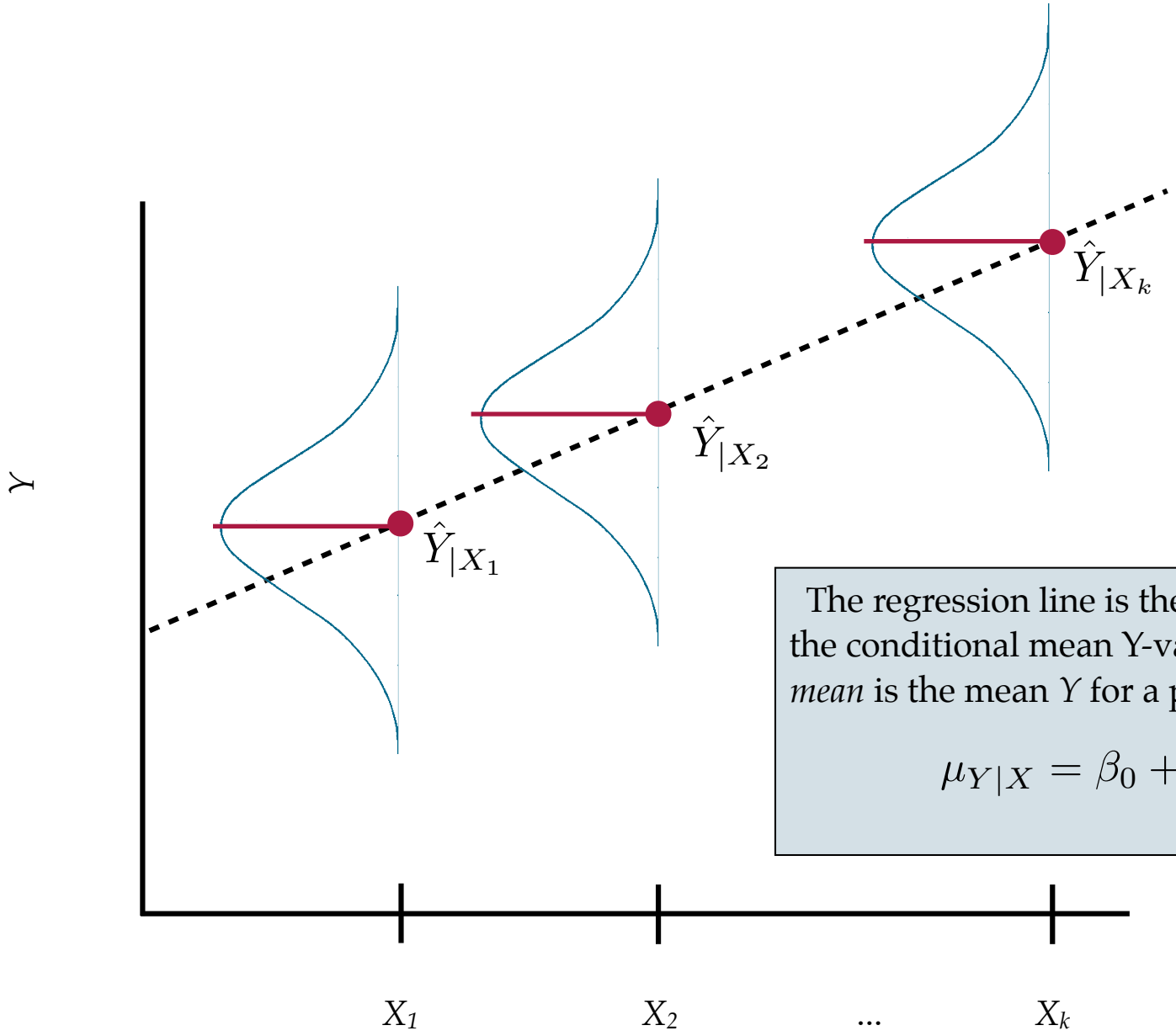
$$\begin{aligned}\hat{\text{achievement}} &= 47 + 2.0(3) \\ &= 53\end{aligned}$$

All of these student's have the same predicted mathematics achievement score of 53.

Since their observed Y varies, their error term will also vary.

Observations that have the same X-value will have the same predicted (fitted) value, despite possibly having different observed Y-values.

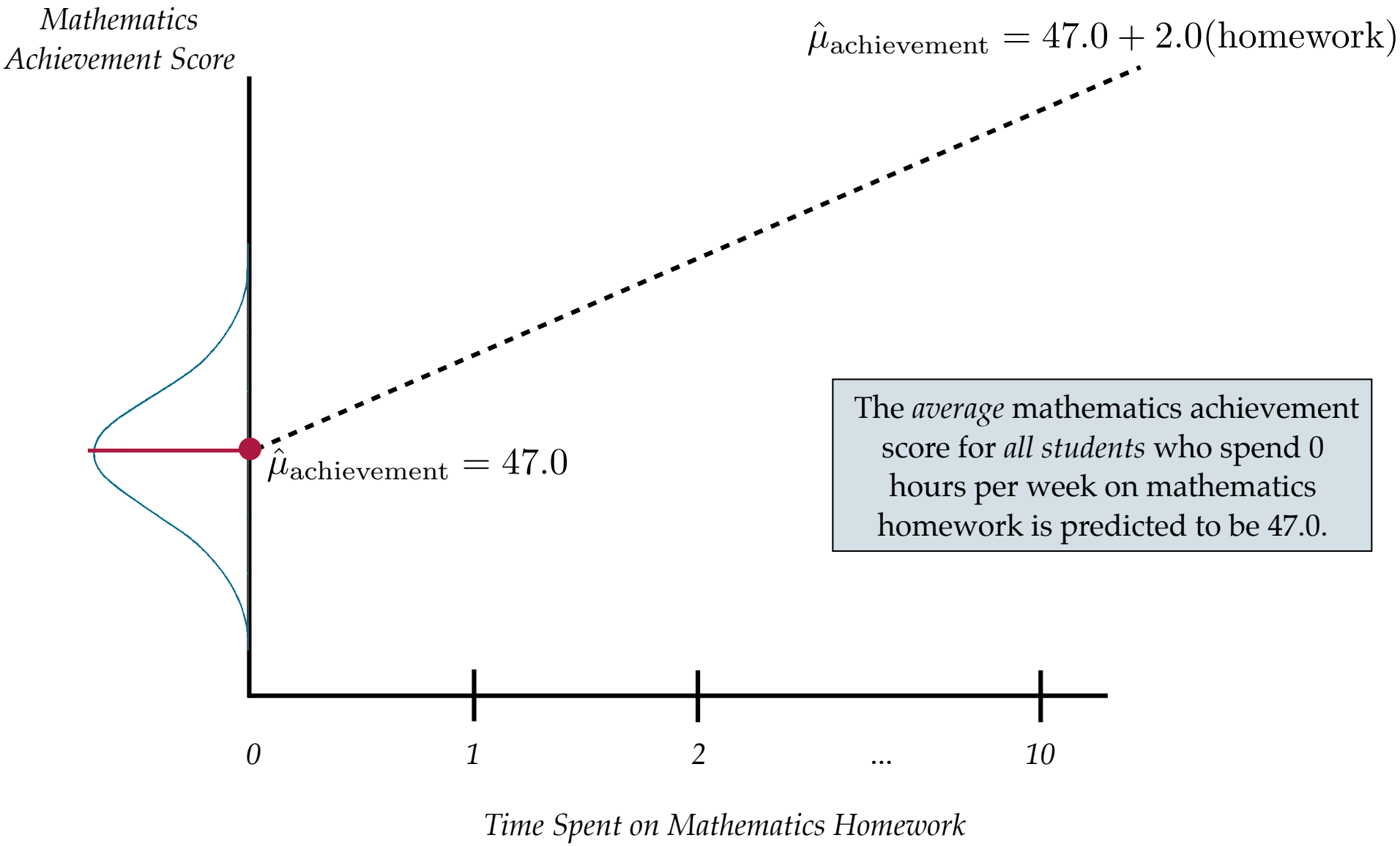
Predicted Values are Means



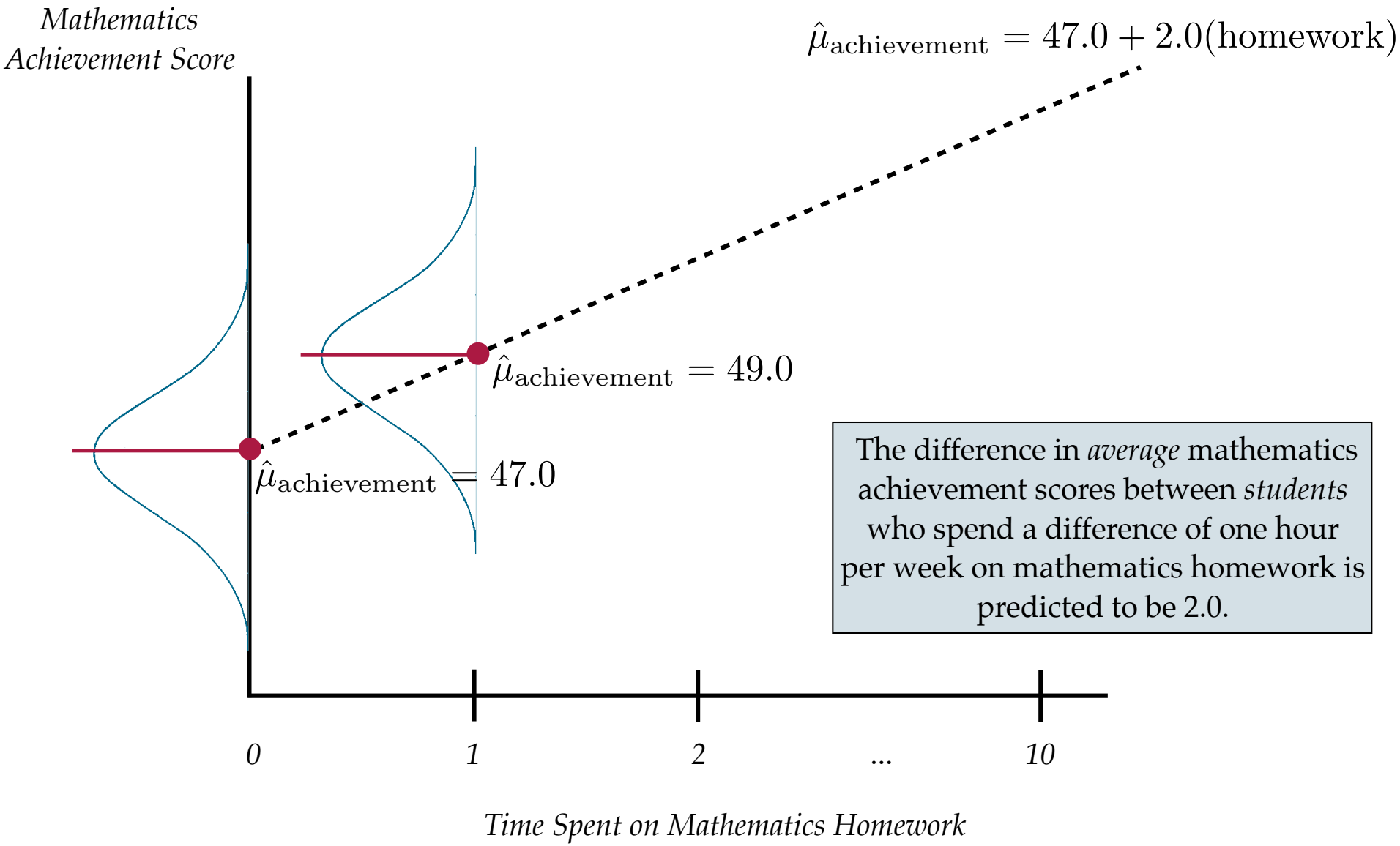
The regression line is the lie that connects the conditional mean Y-values. A *conditional mean* is the mean Y for a particular X-value.

$$\mu_{Y|X} = \beta_0 + \beta_1(X)$$

Interpretation of the Intercept (Revisited)



Interpretation of the Slope (Revisited)



Least Squares Estimation

$$\hat{\text{achievement}} = 47.0 + 2.0(\text{homework})$$

How do we get the values of 47.0 and 2.0 for the intercept and slope?

These values are based on an estimation method called *Least Squares*. Every estimation method requires two things:

- **Quantification of Model Fit:** We quantify how well (or not well) the estimated equation fits the data; and
- **Optimization:** We find the "best" equation based on that quantification. (this boils down to finding the equation that produces the biggest or smallest measure of fit.)

For most statistical models we quantify the model fit by examining the errors. Error is a measure of model misfit (i.e., bigger errors = worse fitting model).

Model A

$$\hat{\text{achievement}} = 30 + 1(\text{homework})$$

Model B

$$\hat{\text{achievement}} = 20 - 2(\text{homework})$$

Homework (*X*)	Achievement (*Y*)	Predicted Achievement	Error
3	63		
1	44		
3	64		
5	68		
2	25		

Model A

$$\hat{\text{achievement}} = 30 + 1(\text{homework})$$

Homework (*X*)	Achievement (*Y*)	Predicted Achievement	Error
3	63		
1	44		
3	64		
5	68		
2	25		

Model B

$$\hat{\text{achievement}} = 20 - 2(\text{homework})$$

Homework (*X*)	Achievement (*Y*)	Predicted Achievement	Error
3	63		
1	44		
3	64		
5	68		
2	25		