Centering and Scaling

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)

> head(math)

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

Correlation

```
> 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, $r = 0.32$.

Unscaled Regression Coefficients

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

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  47.0316     1.6940  27.763  < 2e-16 ***
homework      1.9902     0.5952   3.344  0.00117 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.75 on 98 degrees of freedom
Multiple R-squared:  0.1024,	Adjusted R-squared:  0.09324
F-statistic: 11.18 on 1 and 98 DF,  p-value: 0.001173
```

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

What would have happened if we had measured the amount of time spent on homework in minutes instead of hours?

```
# Create new variable measuring homework in minutes
> math$homework_minutes = math$homework * 60
> head(math)

  homework achievement homework_minutes
1        2          54              120
2        0          53                0
3        4          53              240
4        0          56                0
5        2          59              120
6        0          30                0
```

```
> lm.b = lm(achievement ~ homework_minutes, data = math)
> summary(lm.b)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      47.03160    1.69404  27.763  < 2e-16 ***
homework_minutes  0.03317    0.00992   3.344  0.00117 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.75 on 98 degrees of freedom
Multiple R-squared:  0.1024,	Adjusted R-squared:  0.09324
F-statistic: 11.18 on 1 and 98 DF,  p-value: 0.001173
```

$$\widehat{\text{achievement}} = 47.03 + 0.03(\text{homework\_minutes})$$

Unit of Measurement	F	R^2	B_0	B_1
Hours	$F(1, 98) = 11.18$ $p = 0.0012$	0.1024	47.03 SE = 1.69 $p < 0.001$	1.99 SE = 0.60 $p = 0.0012$
Minutes	$F(1, 98) = 11.18$ $p = 0.0012$	0.1024	47.03 SE = 1.69 $p < 0.001$	0.03 SE = 0.01 $p = 0.0012$

The magnitude of the regression coefficients depends on the unit of measurement of the variables.

Centering and Scaling Variables

Centering a variable changes where the mean of that variable is located. Scaling a variable changes the standard deviation of that variable.

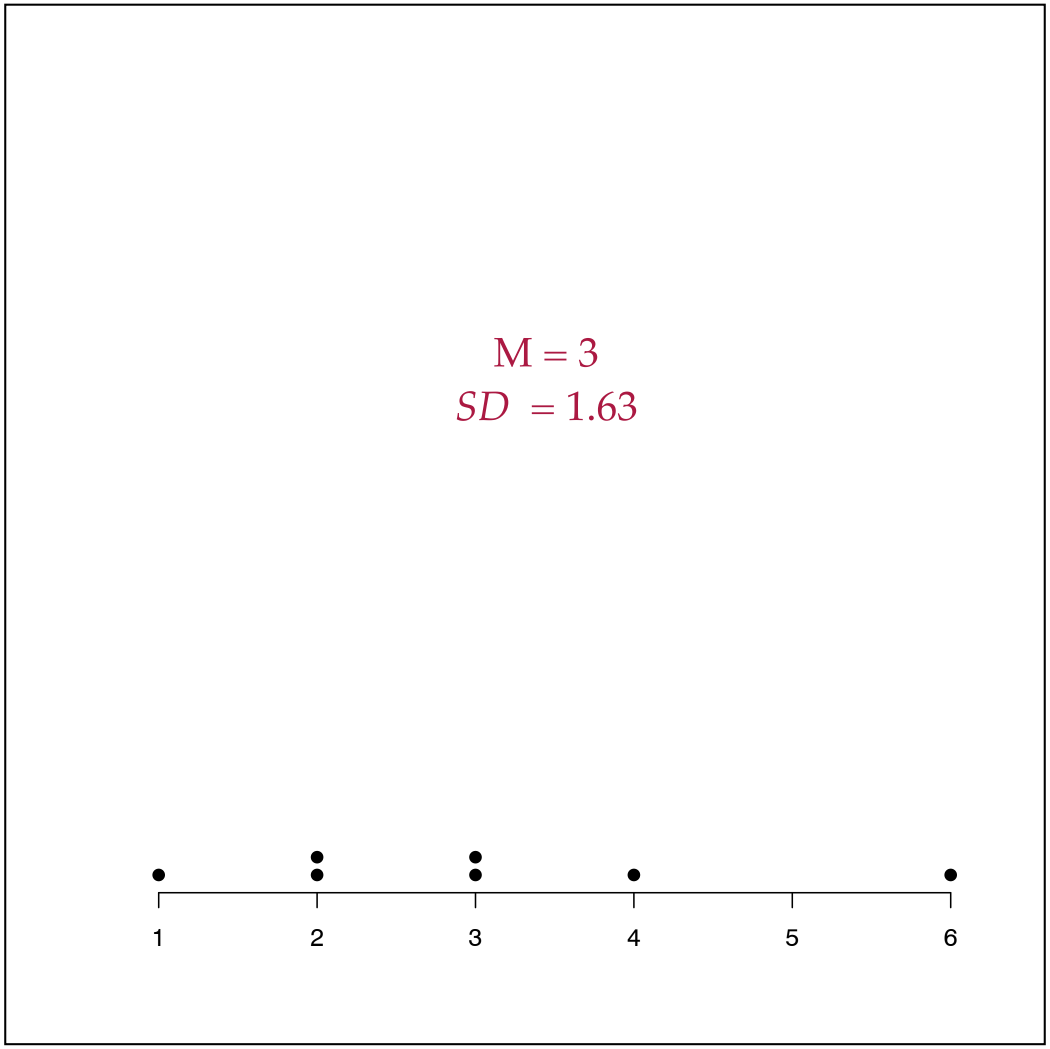
Subtracting the mean from each observation centers the variable (new mean = 0)

$$z = \frac{X - \bar{X}}{SD_X}$$

Dividing by the standard deviation scales the variable to have a new SD of 1

Simple Example

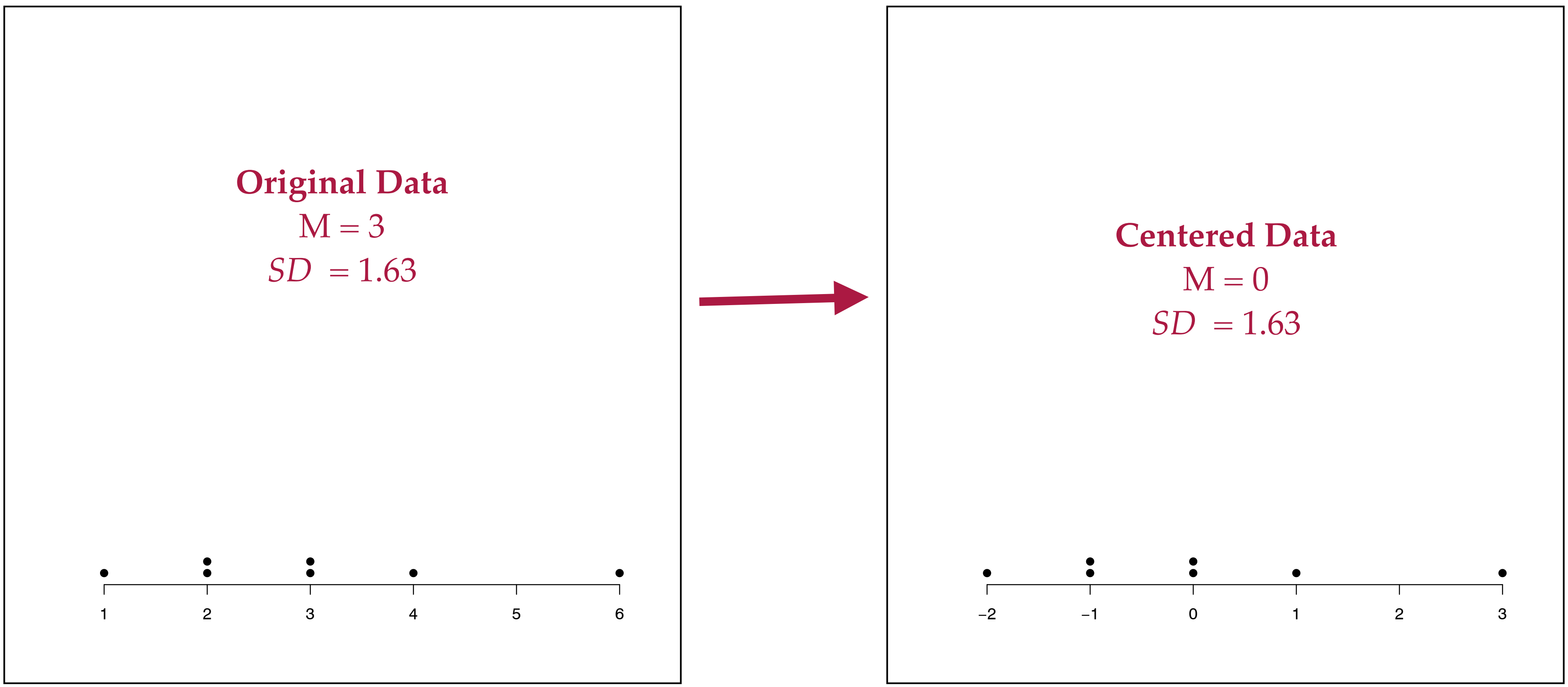
$$X = \{1, 2, 2, 3, 3, 4, 6\}$$



Centering

We center X by subtracting the mean of X from each observation.

$X - \text{mean}(X)$

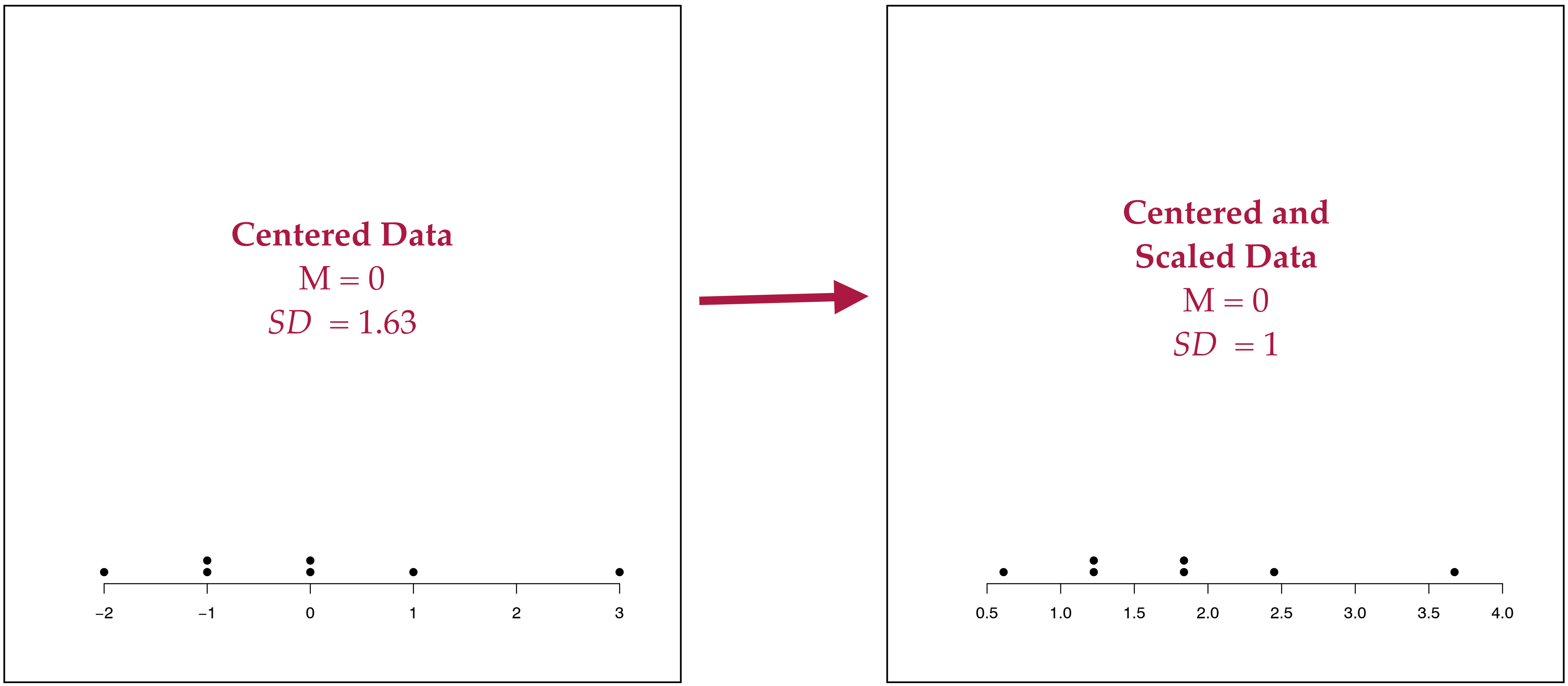


Scaling

We scale a variable X by dividing each observation by the SD of X.

$X \, / \, \mathrm{SD}(X)$

Here we scale the previously centered data.

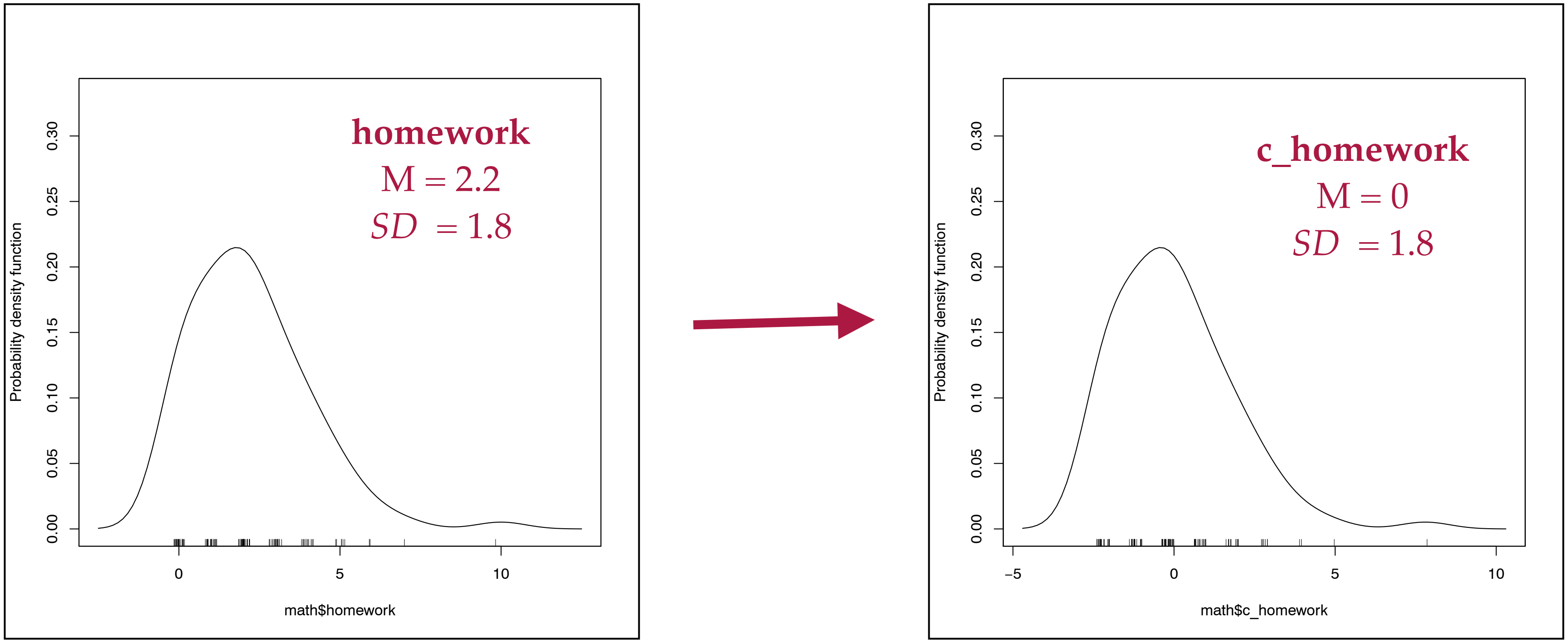


Centering the Predictors in a Regression

Prior to fitting the regression model, we can center the predictor(s).

$$\text{centered\_homework}_i = \text{homework}_i - 2.2$$

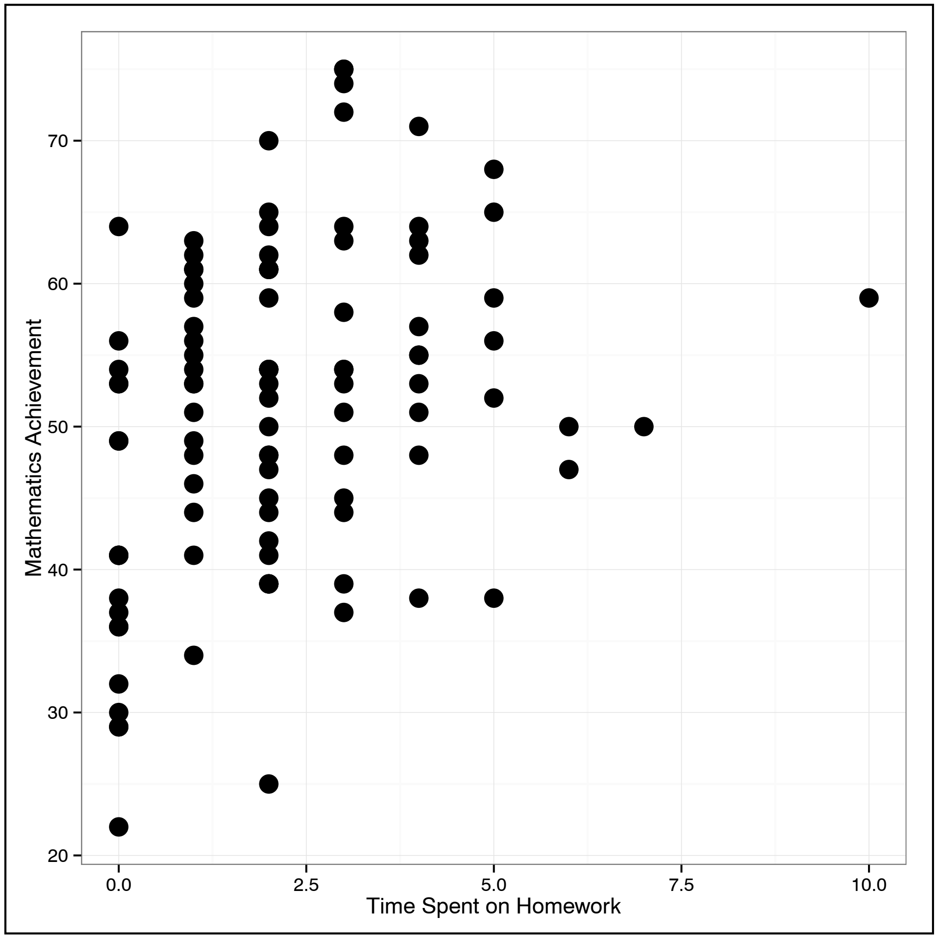
```
# Create centered predictor
> math$c_homework = math$homework - mean(math$homework)
> head(math)

  homework achievement c_homework
1        2          54       -0.2
2        0          53       -2.2
3        4          53        1.8
4        0          56       -2.2
5        2          59       -0.2
6        0          30       -2.2
```



Centering changes the mean of the distribution, but not the standard deviation.

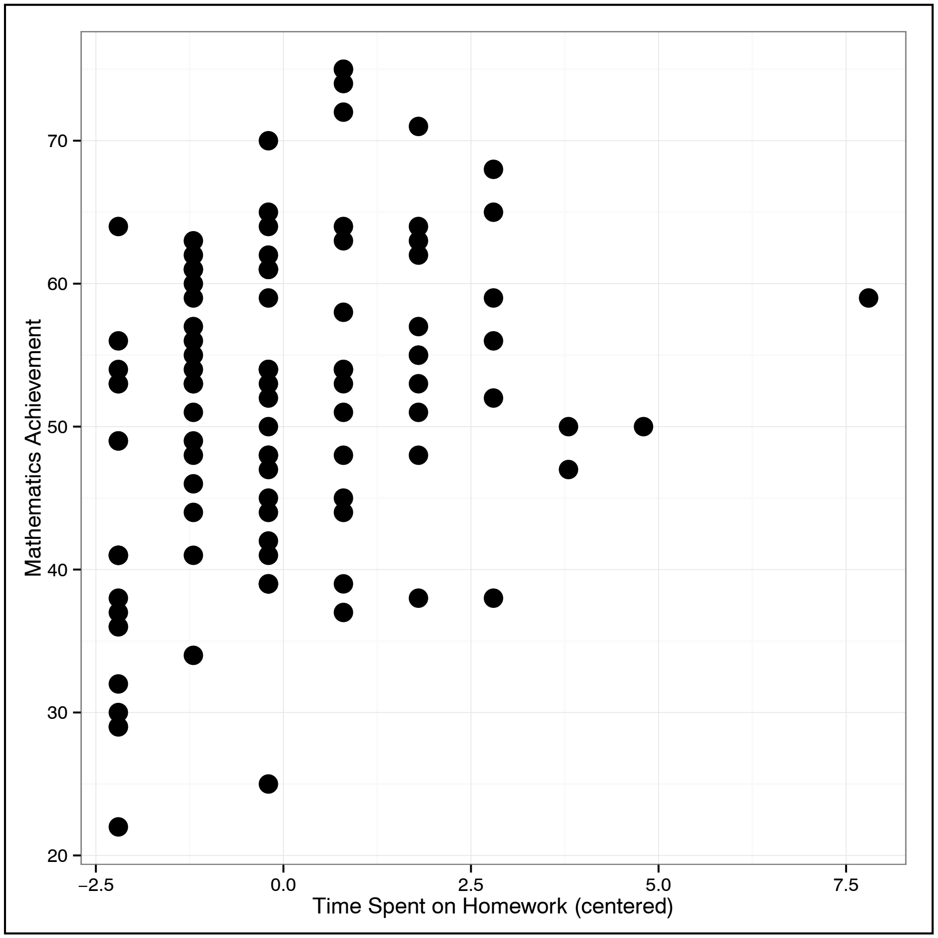
Relationship Between *X* and *Y*

Between achievement and homework



$r = 0.320$

Between achievement and c_homework



$r = 0.320$

Regression of Y on C_X

```
> lm.b = lm(achievement ~ c_homework, data = math)
> summary(lm.b)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  51.4100     1.0747  47.836  < 2e-16 ***
c_homework    1.9902     0.5952   3.344  0.00117 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.75 on 98 degrees of freedom
Multiple R-squared:  0.1024,    Adjusted R-squared:  0.09324
F-statistic: 11.18 on 1 and 98 DF,  p-value: 0.001173
```

The regression is statistically reliable, $F(1, 98) = 11.18$, $p = 0.001$. This suggests that differences in time spent on homework explain variation in mathematics achievement scores in the population ($R^2 = 0.102$).

Note the model-level output for the model with the centered predictor is exactly the same as the regression model-level output for the unscaled variables.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  51.4100     1.0747  47.836  < 2e-16 ***
c_homework    1.9902     0.5952   3.344  0.00117 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$$\widehat{\text{achievement}} = 51.41 + 1.99(\text{c\_homework})$$

The average mathematics achievement score for *all students* who have a *mean centered* score for mathematics homework of 0 (average time spent of mathematics homework) is predicted to be 51.41.

The difference in average mathematics achievement z-scores between students who have a one-unit difference in their mathematics homework scores is predicted to be 1.99.

The interpretation and tests for the slope are the same (we didn't change the scale of the distribution…only the location). The test for the intercept now examines whether the mean achievement score, in the population, is 0 (it is a one-sample t-test for Y).

What Happens if We Center Both the Predictor and Outcome?

```
# Create centered outcome
> math$c_achievement = math$achievement - mean(math$achievement)
> head(math)

  homework achievement c_homework c_achievement
1        2          54       -0.2          2.59
2        0          53       -2.2          1.59
3        4          53        1.8          1.59
4        0          56       -2.2          4.59
5        2          59       -0.2          7.59
6        0          30       -2.2        -21.41
```

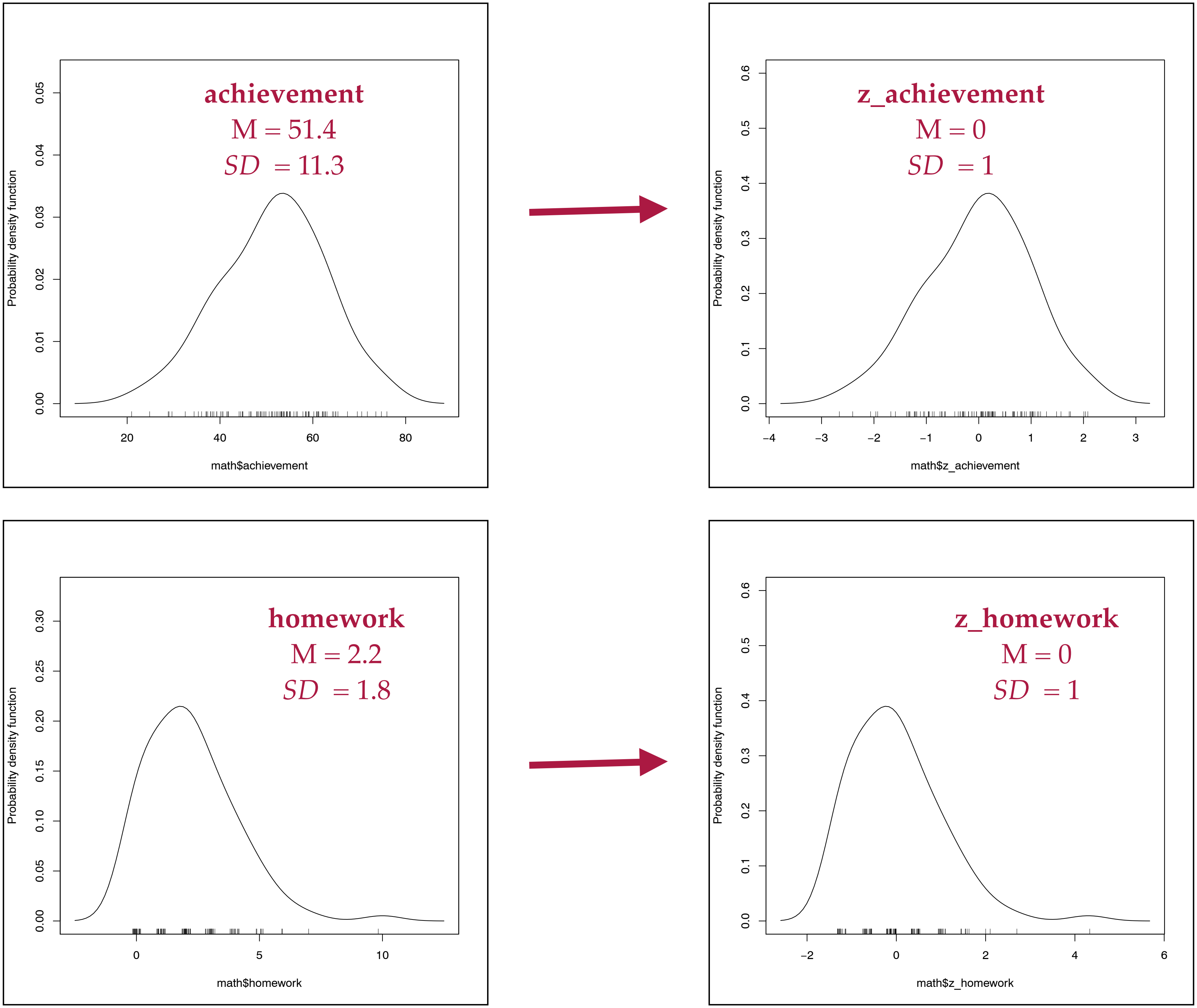
```
> lm.c = lm(c_achievement ~ c_homework, data = math)
> summary(lm.c)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.828e-15  1.075e+00   0.000  1.00000
c_homework  1.990e+00  5.952e-01   3.344  0.00117 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.75 on 98 degrees of freedom
Multiple R-squared:  0.1024,	Adjusted R-squared:  0.09324
F-statistic: 11.18 on 1 and 98 DF,  p-value: 0.001173
```

$$\widehat{\text{c\_achievement}} = 0 + 1.99(\text{c\_homework})$$

Centering and Scaling the Outcome and Predictor in a Regression

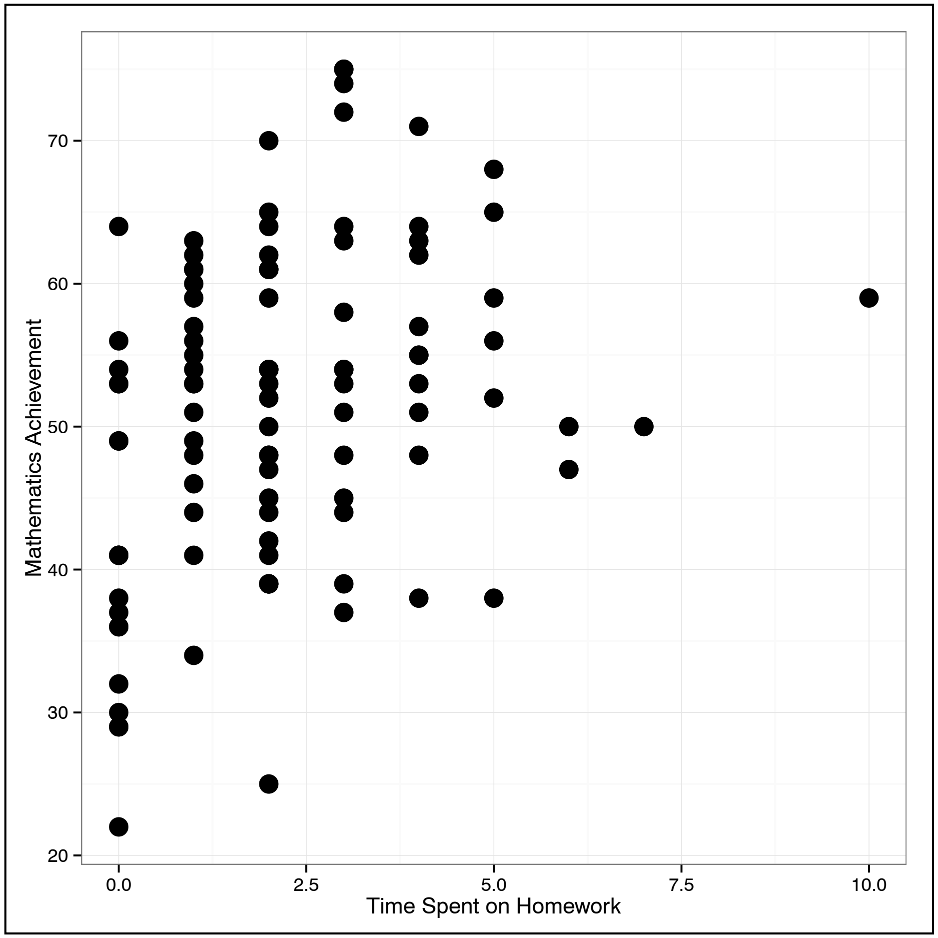
Prior to fitting the regression model, predictors are often centered and scaled.

```
# Create centered and scaled outcome
> math$z_achievement = (math$achievement - mean(math$achievement)) /
    sd(math$achievement)

# Create centered and scaled predictor
> math$z_homework = (math$homework - mean(math$homework)) /
    sd(math$homework)

> head(math)

  homework achievement homework_minutes z_achievement z_homework
1        2          54              120     0.2294862 -0.1102145
2        0          53                0     0.1408815 -1.2123597
3        4          53              240     0.1408815  0.9919306
4        0          56                0     0.4066956 -1.2123597
5        2          59              120     0.6725097 -0.1102145
6        0          30                0    -1.8970267 -1.2123597
```


achievement
M = 51.4
SD = 11.3
Probability density function
math$achievement
z_achievement
M = 0
SD = 1
Probability density function
math$z_achievement
homework
M = 2.2
SD = 1.8
Probability density function
math$homework
z_homework
M = 0
SD = 1
Probability density function
math$z_homework

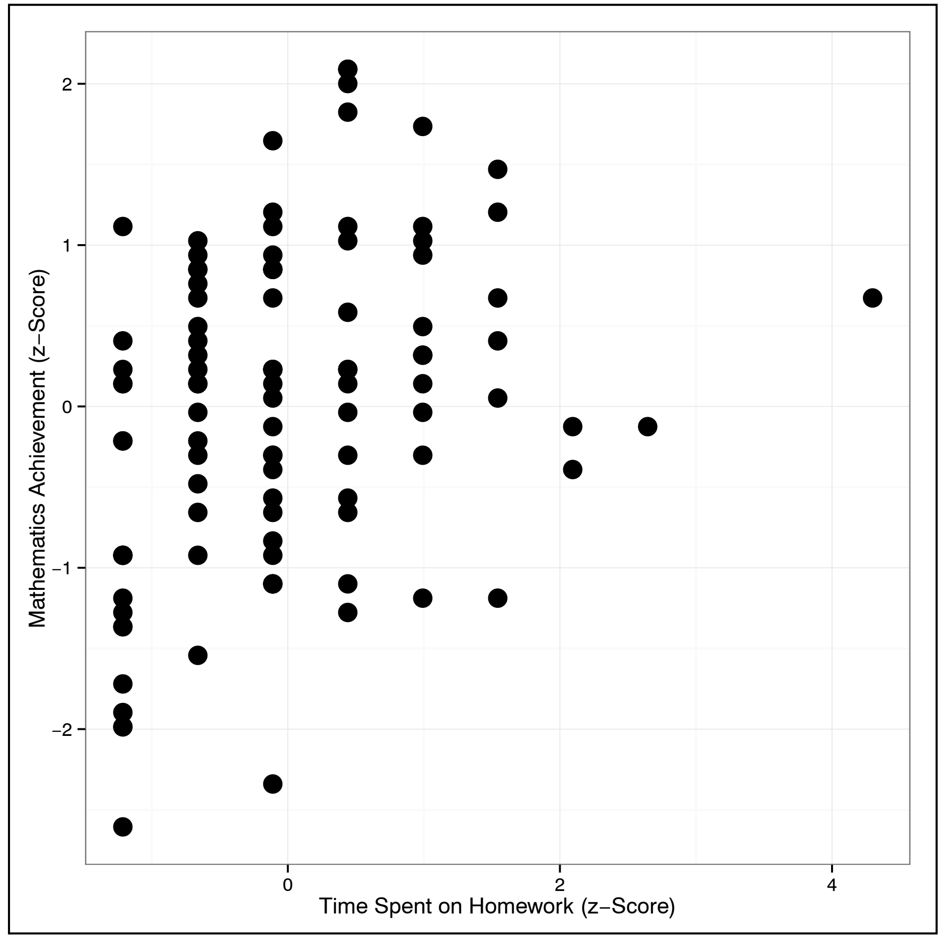
Relationship Between *X* and *Y*

Between achievement and homework



$r = 0.320$

Between z_achievement and z_homework



$r = 0.320$

Regression of Z_Y on Z_X

```
> lm.b = lm(z_achievement ~ z_homework, data = math)
> summary(lm.b)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.267e-16  9.522e-02   0.000  1.00000
z_homework  3.200e-01  9.570e-02   3.344  0.00117 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9522 on 98 degrees of freedom
Multiple R-squared:  0.1024,   Adjusted R-squared:  0.09324
F-statistic: 11.18 on 1 and 98 DF,  p-value: 0.001173
```

The regression is statistically reliable, $F(1, 98) = 11.18$, $p = 0.001$. This suggests that differences in time spent on homework explain variation in mathematics achievement scores in the population ($R^2 = 0.102$).

Note the model-level output for the centered and scaled variables are exactly the same as the regression model-level output for the unscaled variables.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.267e-16  9.522e-02   0.000  1.00000
z_homework  3.200e-01  9.570e-02   3.344  0.00117 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$$\widehat{\text{z\_achievement}} = 0 + 0.32(\text{z\_homework})$$

The average mathematics achievement *z*-score for *all students* who have a *z*-score for mathematics homework of 0 is predicted to be 0.

The difference in average mathematics achievement *z*-scores between students who have a one-unit difference in their mathematics homework *z*-scores is predicted to be 0.32.

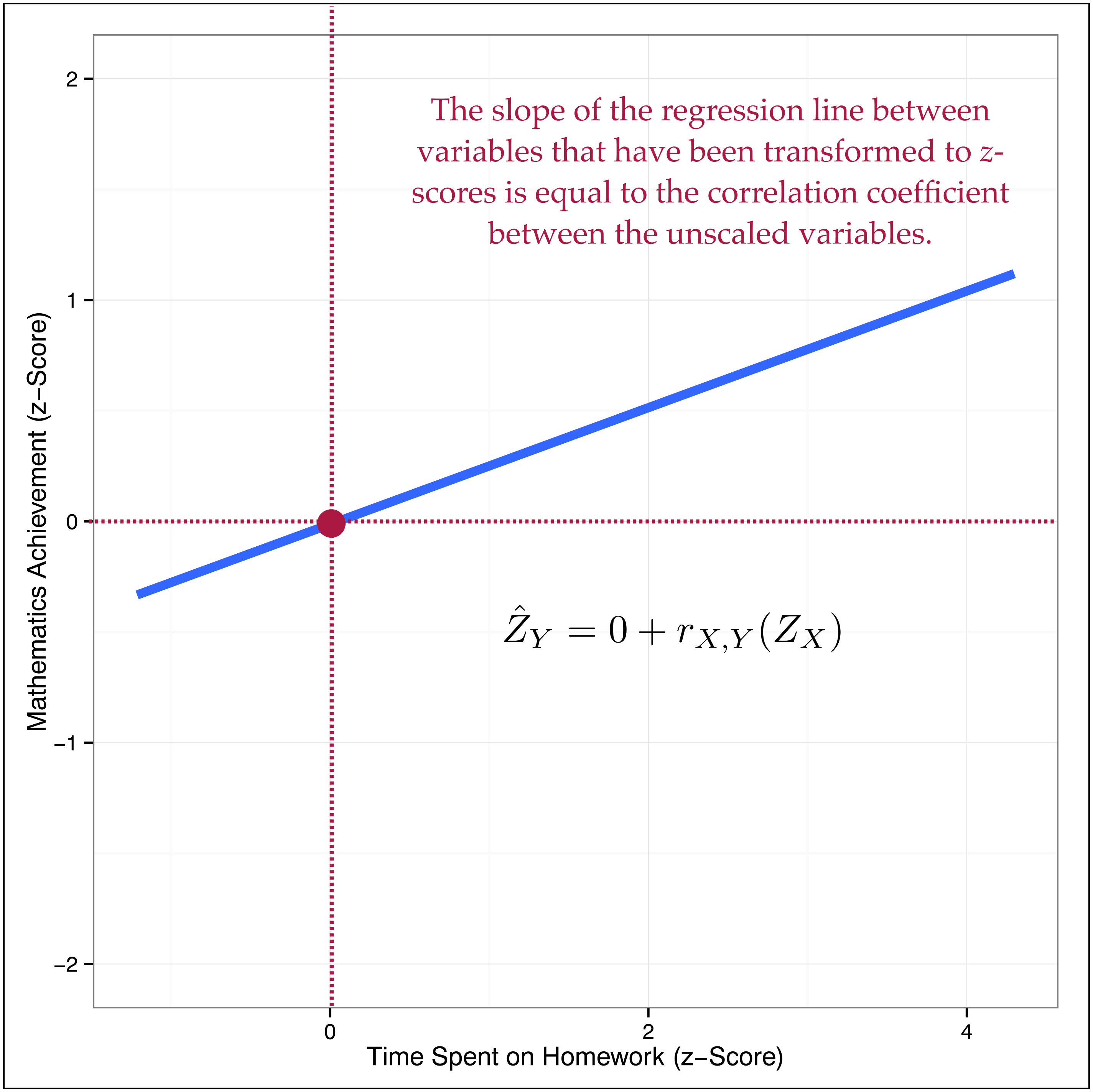
When both the outcome and predictor variables have been transformed to *z*-scores, the regression is often referred to as a **standardized regression**. The regression coefficients from a standardized regression are typically referred to as **beta weights** (not to be confused with the population parameters, e.g., β_0 and β_1).

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.267e-16  9.522e-02   0.000  1.00000
z_homework  3.200e-01  9.570e-02   3.344  0.00117 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$$\widehat{\text{z\_achievement}} = 0 + 0.32(\text{z\_homework})$$

The output for the test of the intercept (H_0: $\beta_0 = 0$) will always be non-significant at $p = 1$. This is because in a regression, the predicted value of Y for average values of X will always be the average Y.

The output for the test for the predictor (H_0: $\beta_1 = 0$) will always be identical to the test in the unstandardized regression. Note because the slope is equal to the correlation coefficient between the unstandardized variables, this is equivalent to testing, H_0: $\rho_{X,Y} = 0$.



The slope of the regression line between variables that have been transformed to *z*-scores is equal to the correlation coefficient between the unscaled variables.
$\hat{Z}_Y = 0 + r_{X,Y}(Z_X)$
Mathematics Achievement (z–Score)
Time Spent on Homework (z–Score)
2
1
0
−1
−2
0
2
4