# Multiple Linear Regression

POLS 602
Fall 2025

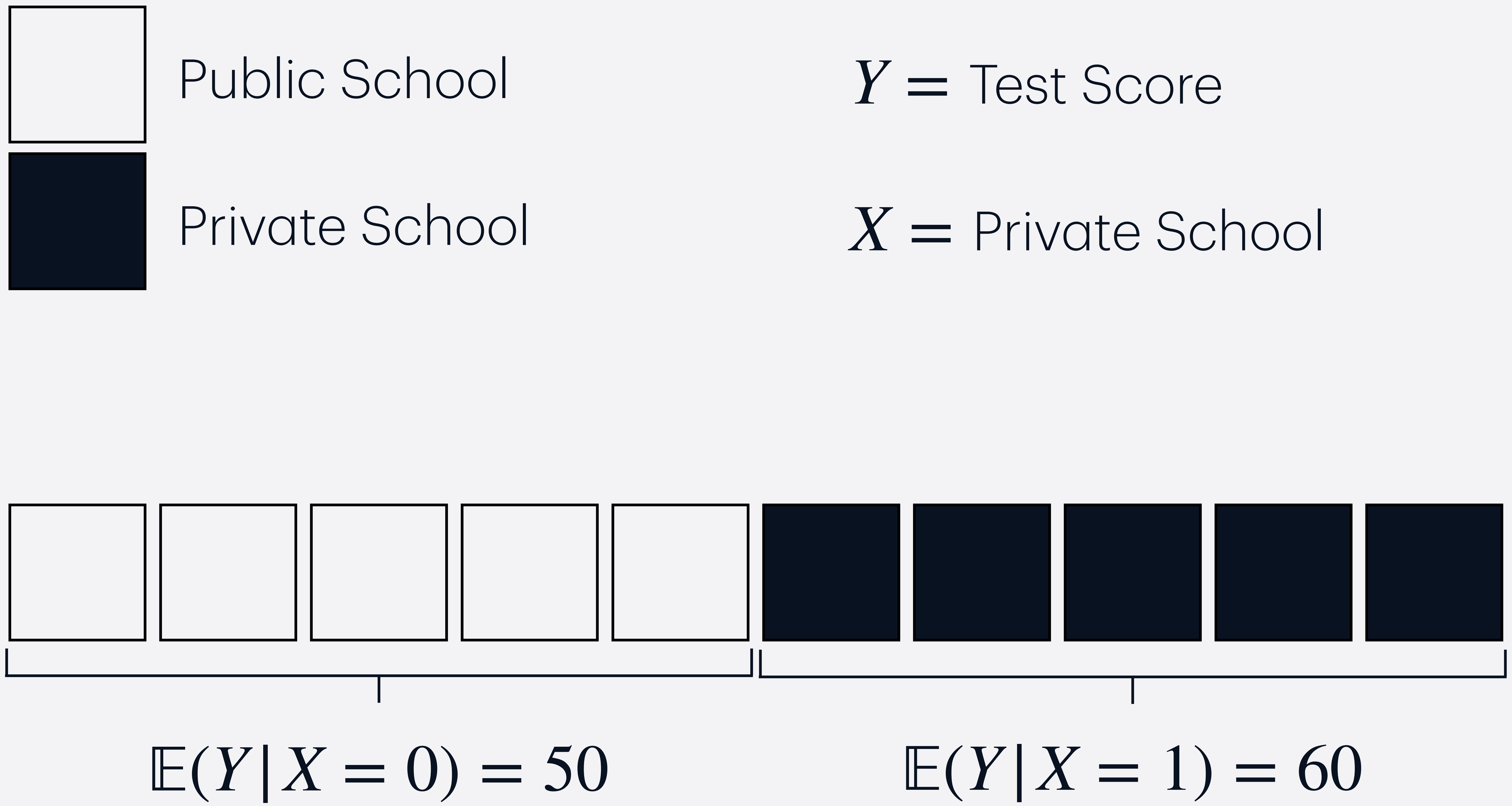
Dr. Mike Burnham
Texas A&M Political Science

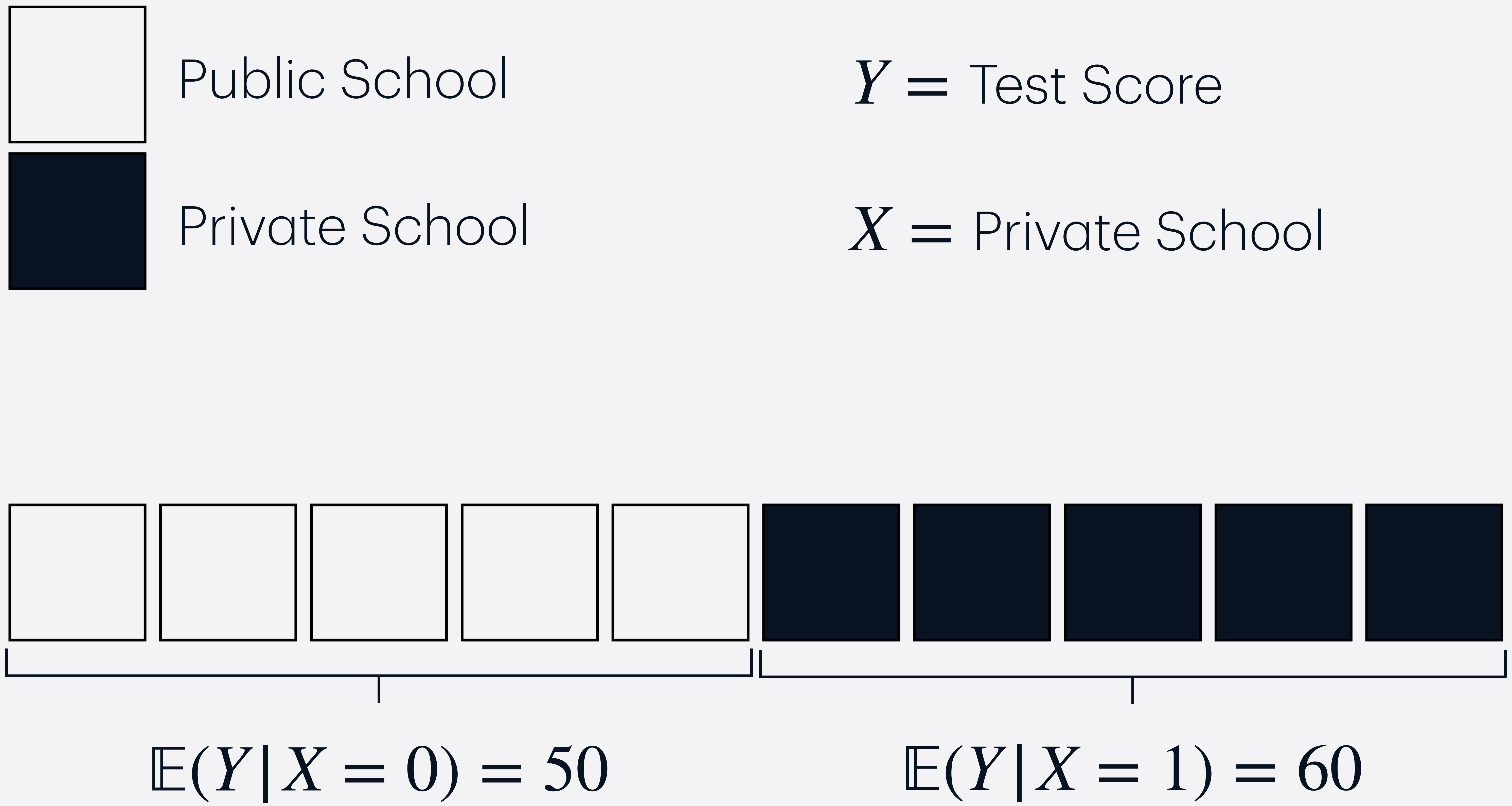
Public School

Private School

$Y =$ Test Score

$X =$ Private School


Public School
Private School
$Y =$ Test Score
$X =$ Private School
$\mathbb{E}(Y|X=0) = 50$
$\mathbb{E}(Y|X=1) = 60$


Public School
Private School
$Y =$ Test Score
$X =$ Private School
$\mathbb{E}(Y|X=0) = 50$
$\mathbb{E}(Y|X=1) = 60$

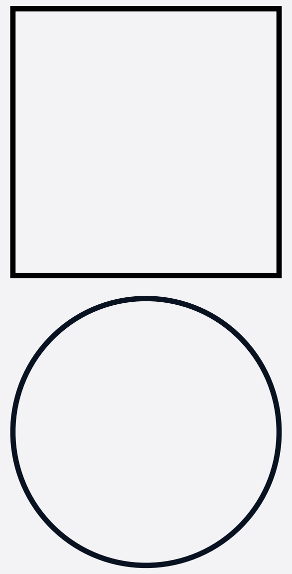
$$\beta_1 = 10$$

□ Not Wealthy

○ Wealthy

$Y =$ Test Score

$Z =$ Wealthy

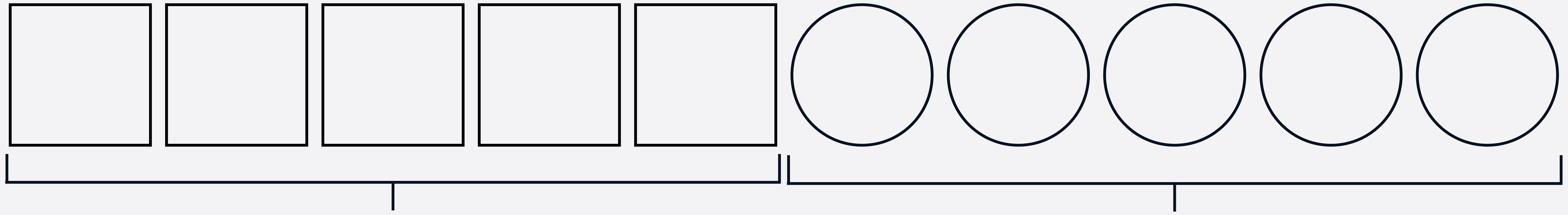


Not Wealthy

Wealthy

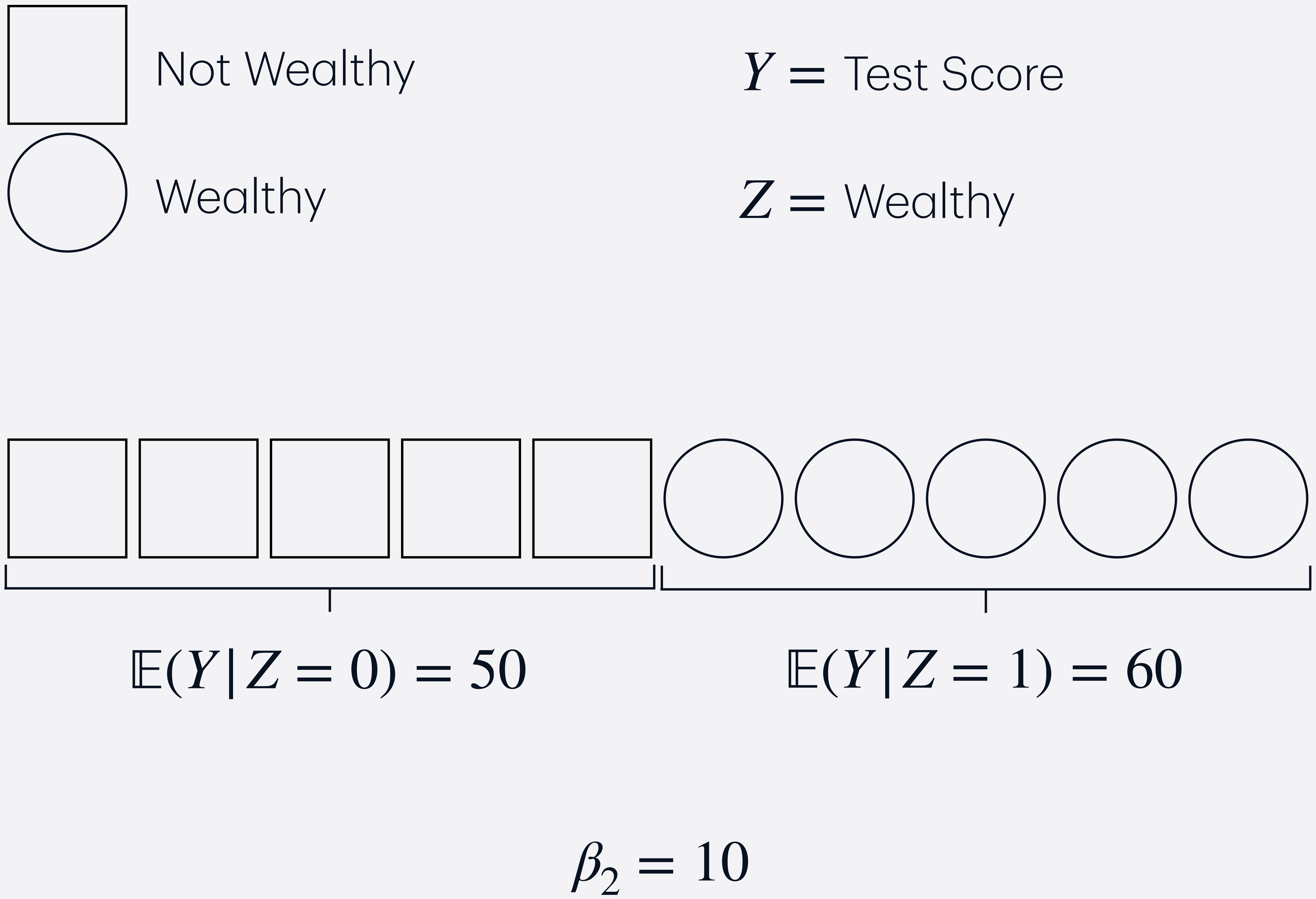
$Y =$ Test Score

$Z =$ Wealthy



$\mathbb{E}(Y|X = 0) = 50$

$\mathbb{E}(Y|X = 1) = 60$


Not Wealthy
Wealthy
$Y =$ Test Score
$Z =$ Wealthy
$\mathbb{E}(Y | Z = 0) = 50$
$\mathbb{E}(Y | Z = 1) = 60$
$\beta_2 = 10$

Not Wealthy

Wealthy

Public School

Private School

$Y =$ Test Score

$X =$ Private School

$Z =$ Wealthy

$\beta_0 = 50$

$\beta_1 = 10$

$\beta_2 = 10$

Not Wealthy

Wealthy

Public School

Private School

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i$$

$Y =$ Test Score

$X =$ Private School

$Z =$ Wealthy

$\beta_0 = 50$

$\beta_1 = 10$

$\beta_2 = 10$


$Y =$ Test Score

$X =$ Private School

$Z =$ Wealthy

$\beta_0 = 50$

$\beta_1 = 10$

$\beta_2 = 10$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i$$


$Y =$ Test Score

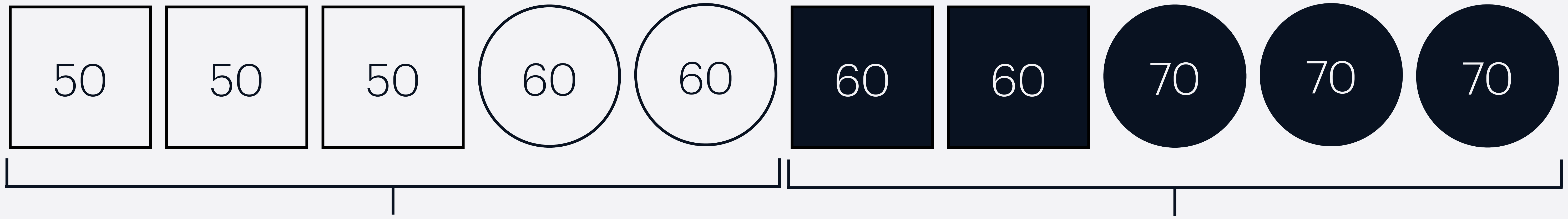
$X =$ Private School

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\epsilon}_i$$


$Y =$ Test Score

$X =$ Private School

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\epsilon}_i$$

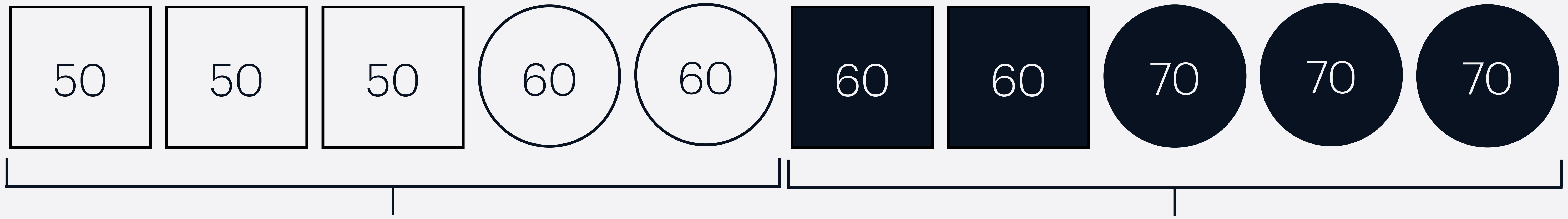



$\bar{Y}_{x=0} = 54$ $\bar{Y}_{x=1} = 66$


$Y =$ Test Score

$X =$ Private School

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\epsilon}_i$$




$\bar{Y}_{x=0} = 54$

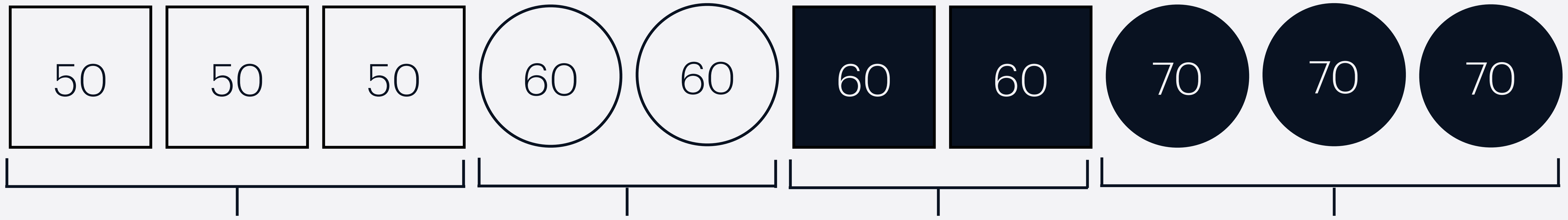
$\bar{Y}_{x=1} = 66$

$$\hat{\beta}_1 = 12$$


$Y =$ Test Score

$X =$ Private School

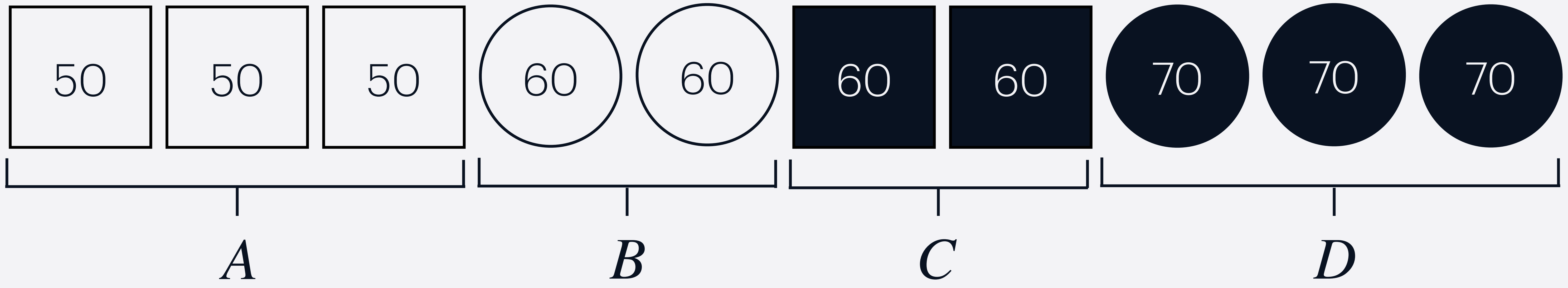
$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 z_i + \hat{\epsilon}_i$$




$Y =$ Test Score

$X =$ Private School

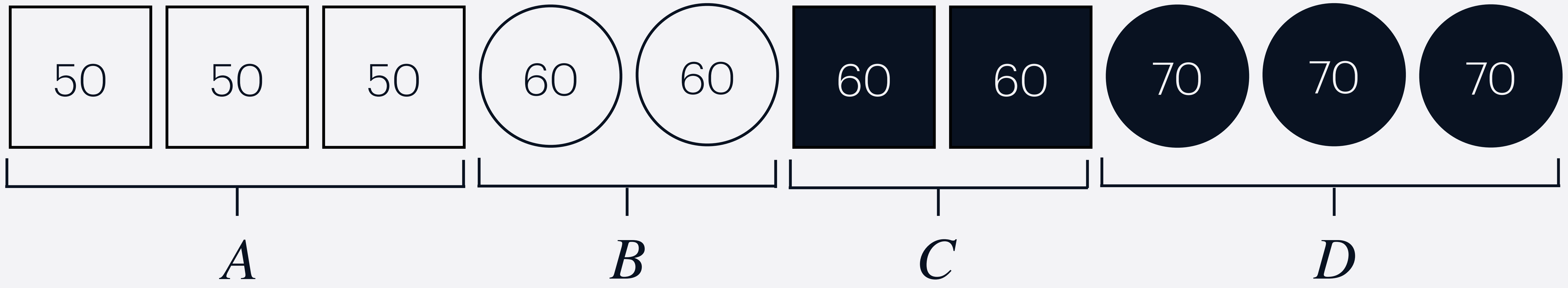
$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 z_i + \hat{\epsilon}_i$$




$Y =$ Test Score

$X =$ Private School

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 z_i + \hat{\epsilon}_i$$




$$\hat{\beta}_1 \approx \frac{[\bar{C} - \bar{A}] + [\bar{D} - \bar{B}]}{2} \approx 10$$

# Multiple Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

# Linear Algebra Detour

# Matrix formulation

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

# Matrix formulation

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_1 + \epsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_2 + \epsilon_2 \\
&\vdots \\
y_n &= \beta_0 + \beta_1 x_n + \epsilon_n
\end{aligned}$$

# Matrix formulation

$$
\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_1 + \epsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_2 + \epsilon_2 \\
&\vdots \\
y_n &= \beta_0 + \beta_1 x_n + \epsilon_n
\end{aligned}
$$

$$
\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}
$$

# Matrix formulation

$$
\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_1 + \epsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_2 + \epsilon_2 \\
&\vdots \\
y_n &= \beta_0 + \beta_1 x_n + \epsilon_n
\end{aligned}
$$

$$
\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}
$$

# Matrix formulation

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$Y = X\beta + \epsilon$$

# Matrix formulation

$$X\beta = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \vdots \\ \beta_0 + \beta_1 x_n \end{bmatrix}$$

# Matrix formulation

$$b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix} = (X'X)^{-1}X'Y$$

# Categorical Variables

```{r}
social <- read.csv("https://raw.githubusercontent.com/MLBurnham/pols_602/refs/heads/main/d
ata/social.csv")

head(social)
```

| | sex | yearofbirth | primary2004 | messages | primary2006 | hhsize |
|---|---|---|---|---|---|---|
| | <chr> | <int> | <int> | <chr> | <int> | <int> |
| 1 | male | 1941 | 0 | Civic Duty | 0 | 2 |
| 2 | female | 1947 | 0 | Civic Duty | 0 | 2 |
| 3 | male | 1951 | 0 | Hawthorne | 1 | 3 |
| 4 | female | 1950 | 0 | Hawthorne | 1 | 3 |
| 5 | female | 1982 | 0 | Hawthorne | 1 | 3 |
| 6 | male | 1981 | 0 | Control | 0 | 3 |

6 rows

```{r}
# Convert to a factor variable
social$messages <- as.factor(social$messages)
# Check categories
levels(social$messages)
```

```
[1] "Civic Duty" "Control"    "Hawthorne"  "Neighbors"
```

````
```{r}
fit <- lm(primary2006 ~ messages, data = social)
summary(fit)
```
````

```
Call:
lm(formula = primary2006 ~ messages, data = social)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.3780 -0.2966 -0.2966  0.6776  0.7034 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)        0.314538   0.002367 132.909  < 2e-16 ***
messagesControl   -0.017899   0.002592  -6.905 5.03e-12 ***
messagesHawthorne  0.007837   0.003347   2.341   0.0192 *  
messagesNeighbors  0.063411   0.003347  18.944  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4627 on 305862 degrees of freedom
Multiple R-squared:  0.003283,	Adjusted R-squared:  0.003273 
F-statistic: 335.8 on 3 and 305862 DF,  p-value: < 2.2e-16
```

```{r}
# Adjust factors so the control is in the intercept
social$messages <- factor(social$messages, levels = c("Control", "Civic Duty",
"Hawthorne", "Neighbors"))
# Fit new model
fit <- lm(primary2006 ~ messages, data = social)
summary(fit)
```

```
Call:
lm(formula = primary2006 ~ messages, data = social)

Residuals:
    Min      1Q  Median      3Q     Max
-0.3780 -0.2966 -0.2966  0.6776  0.7034

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.296638   0.001058 280.393  < 2e-16 ***
messagesCivic Duty 0.017899   0.002592   6.905 5.03e-12 ***
messagesHawthorne  0.025736   0.002593   9.927  < 2e-16 ***
messagesNeighbors  0.081310   0.002593  31.360  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4627 on 305862 degrees of freedom
Multiple R-squared:  0.003283,	Adjusted R-squared:  0.003273
F-statistic: 335.8 on 3 and 305862 DF,  p-value: < 2.2e-16
```

```{r}
# Add aditional variables
fit <- lm(primary2006 ~ messages + sex + primary2004, data = social)
summary(fit)
```

```
Call:
lm(formula = primary2006 ~ messages + sex + primary2004, data = social)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.4747 -0.3221 -0.2417  0.6000  0.7707 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)        0.229297   0.001492 153.661  < 2e-16 ***
messagesCivic Duty 0.018051   0.002558   7.057 1.71e-12 ***
messagesHawthorne  0.025296   0.002558   9.888  < 2e-16 ***
messagesNeighbors  0.080358   0.002558  31.409  < 2e-16 ***
sexmale            0.012447   0.001651   7.540 4.73e-14 ***
primary2004        0.152632   0.001684  90.636  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4565 on 305860 degrees of freedom
Multiple R-squared:  0.02954,	Adjusted R-squared:  0.02952 
F-statistic:  1862 on 5 and 305860 DF,  p-value: < 2.2e-16
```

# Interaction Terms

# Interaction terms

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

- Multiplying two variables in our model together
- Why use them?
    - Heterogeneous treatment effects
    - Joint treatment effects

```{r}
# Let's subset our treatment to a single message
neighbors <- social[social$messages == 'Control' | social$messages == 'Neighbors',]
# Fit a new model with an interaction term
fit_primary <- lm(primary2006 ~ messages + primary2004 + messages*primary2004, data =
neighbors)
# Alternatively...
fit_primary <- lm(primary2006 ~ messages*primary2004, data = neighbors)
summary(fit_primary)
```

```
Call:
lm(formula = primary2006 ~ messages * primary2004, data = neighbors)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.4823 -0.3064 -0.2371  0.6142  0.7629 

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                   0.237110   0.001345 176.322  < 2e-16 ***
messagesNeighbors             0.069296   0.003310  20.934  < 2e-16 ***
primary2004                   0.148695   0.002125  69.963  < 2e-16 ***
messagesNeighbors:primary2004 0.027229   0.005198   5.239 1.62e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4554 on 229440 degrees of freedom
Multiple R-squared:  0.03078,	Adjusted R-squared:  0.03076 
F-statistic:  2428 on 3 and 229440 DF,  p-value: < 2.2e-16
```

```{r}
neighbors$age = 2006 - neighbors$yearofbirth
fit_age <- lm(primary2006 ~ age + messages + age*messages, data = neighbors)
summary(fit_age)
```

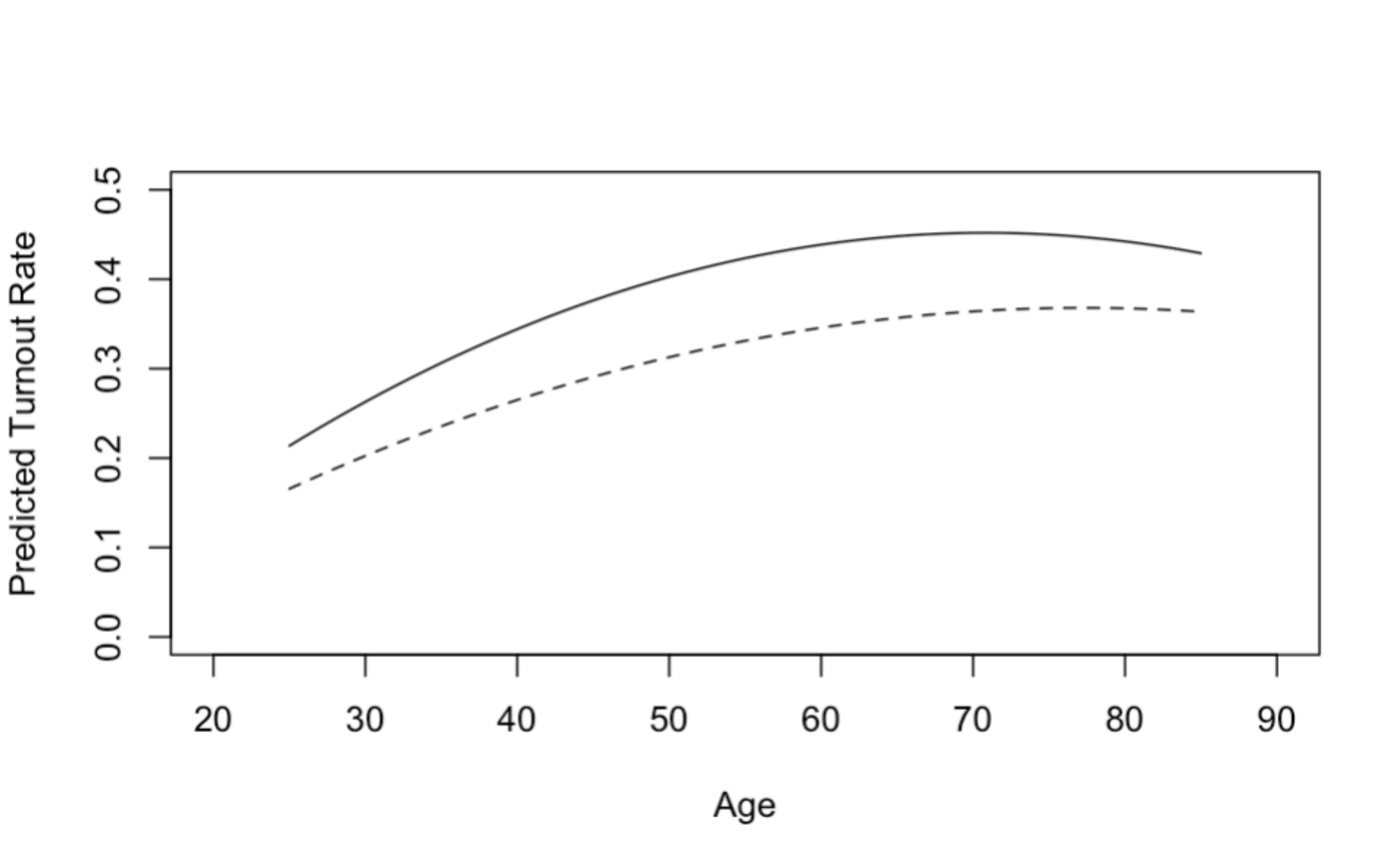
```
Call:
lm(formula = primary2006 ~ age + messages + age * messages, data = neighbors)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.6146 -0.3214 -0.2654  0.6227  0.8226 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          0.0974733  0.0037603  25.922  < 2e-16 ***
age                  0.0039982  0.0000725  55.145  < 2e-16 ***
messagesNeighbors    0.0498294  0.0091519   5.445 5.19e-08 ***
age:messagesNeighbors 0.0006283  0.0001762   3.565 0.000364 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4577 on 229440 degrees of freedom
Multiple R-squared:  0.02081,	Adjusted R-squared:  0.02079 
F-statistic:  1625 on 3 and 229440 DF,  p-value: < 2.2e-16
```

````
```{r}
fit_agesq <- lm(primary2006 ~ age + I(age^2) + messages + age*messages +
messages*I(age^2), data = neighbors)

summary(fit_agesq)
```
````

```
Call:
lm(formula = primary2006 ~ age + I(age^2) + messages + age *
    messages + messages * I(age^2), data = neighbors)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4519 -0.3344 -0.2758  0.6334  0.8749

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -7.385e-02  9.164e-03  -8.058 7.78e-16 ***
age                         1.143e-02  3.696e-04  30.914  < 2e-16 ***
I(age^2)                   -7.389e-05  3.605e-06 -20.494  < 2e-16 ***
messagesNeighbors          -4.330e-02  2.232e-02  -1.940   0.0523 .
age:messagesNeighbors       4.646e-03  8.989e-04   5.169 2.36e-07 ***
I(age^2):messagesNeighbors -3.961e-05  8.744e-06  -4.529 5.92e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4571 on 229438 degrees of freedom
Multiple R-squared:  0.02346,   Adjusted R-squared:  0.02344
F-statistic:  1102 on 5 and 229438 DF,  p-value: < 2.2e-16
```

```{r}
# Predicted turnout rate for the treatment group
yhat_t <- predict(fit_agesq, newdata = data.frame(age = 25:85, messages = "Neighbors"))
# Predicted turnout rate for the control group
yhat_c <- predict(fit_agesq, newdata = data.frame(age = 25:85, messages = "Control"))
```

```{r}
plot(x = 25:85, y = yhat_t, type = 'l', xlim = c(20,90), ylim = c(0, 0.5),
     xlab = 'Age', ylab = 'Predicted Turnout Rate')
lines(x = 25:85, y = yhat_c, lty = 'dashed')
```



```{r}
plot(x = 25:85, y = yhat_t - yhat_c, type = 'l', xlim = c(20,90), ylim = c(0, 0.1),
     xlab = 'Age', ylab = 'Estimated ATE')
```