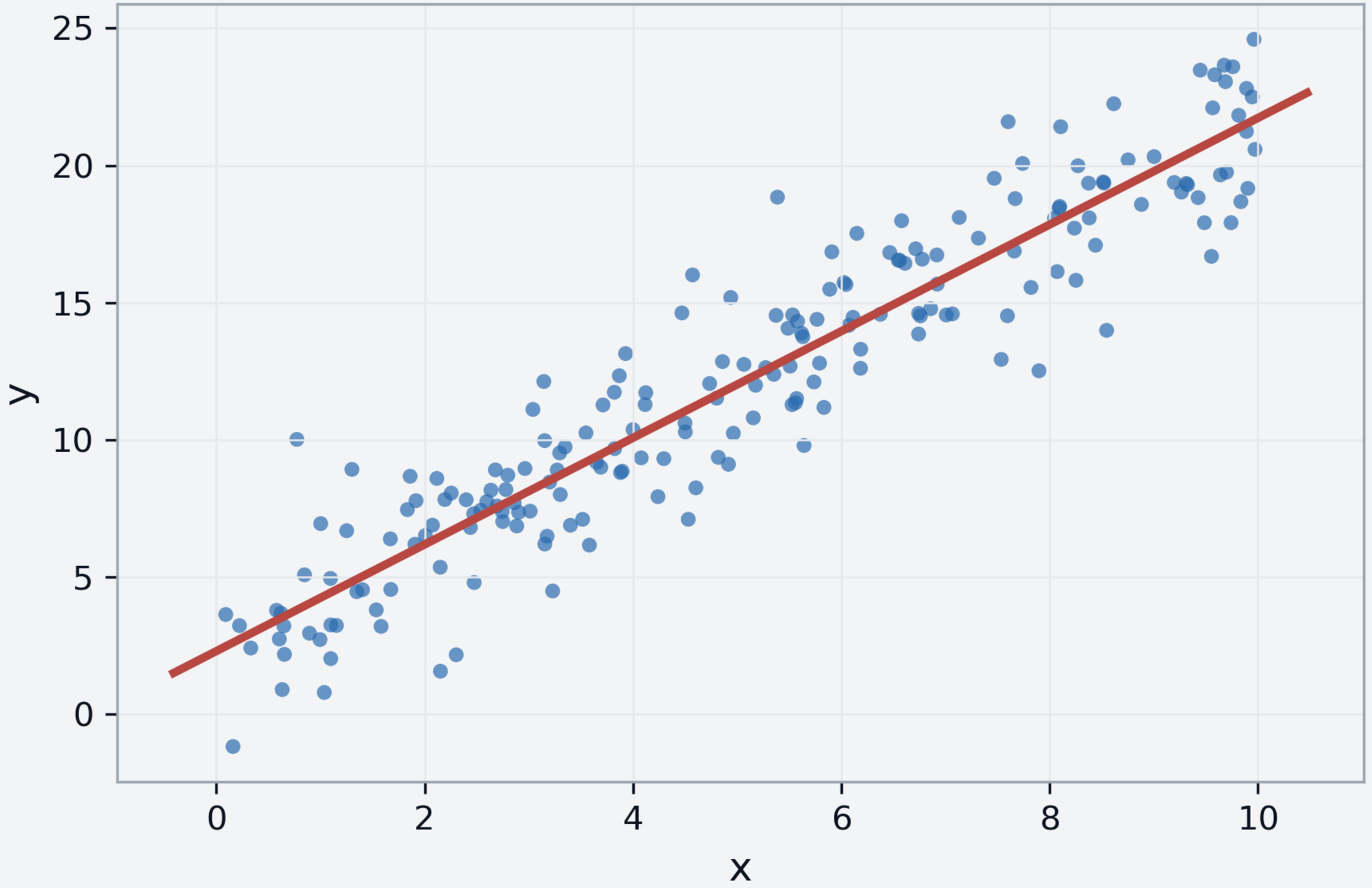
# Linear Regression Cont.

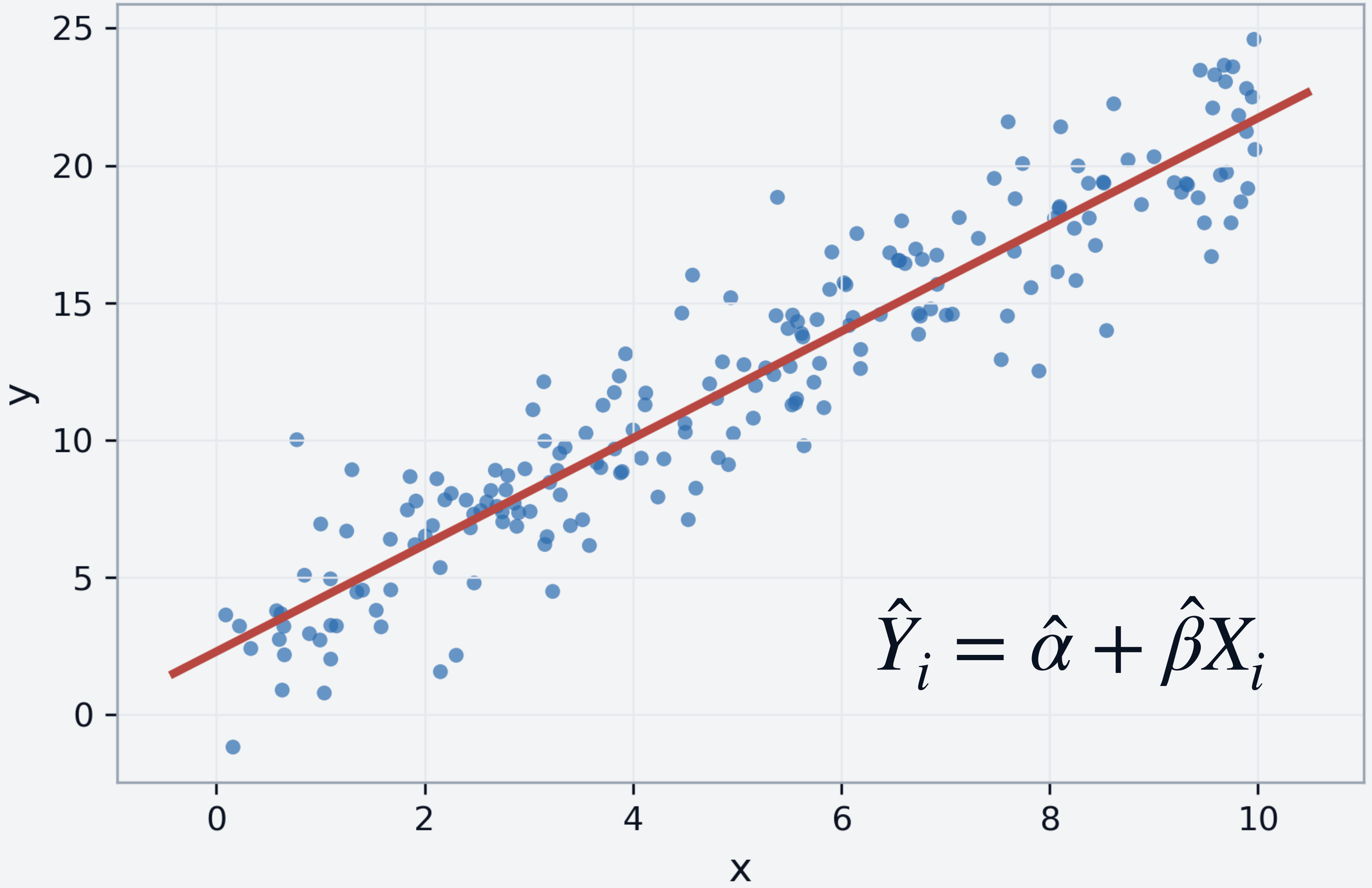
POLS 602

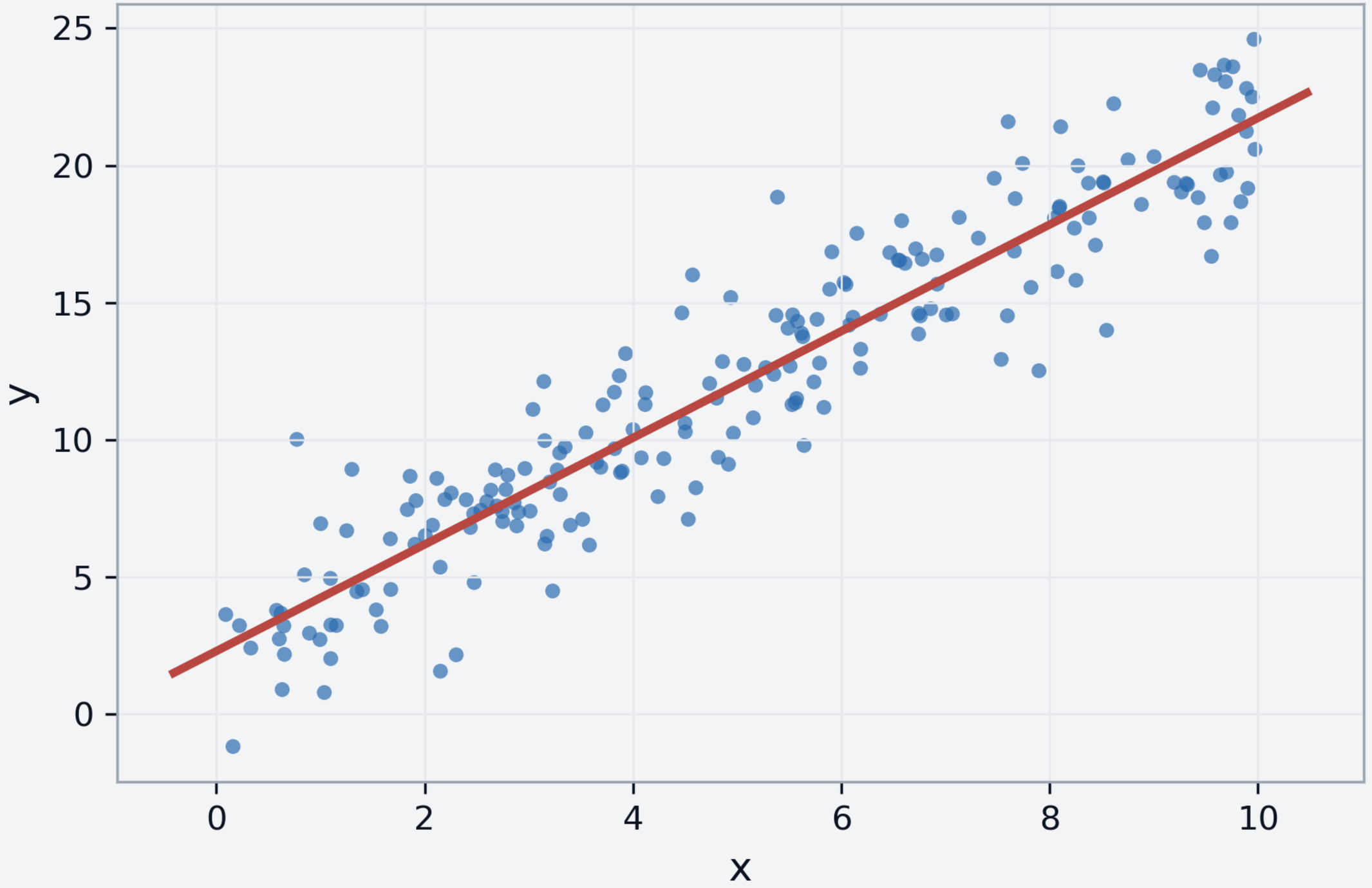
Dr. Mike Burnham
Texas A&M Political Science

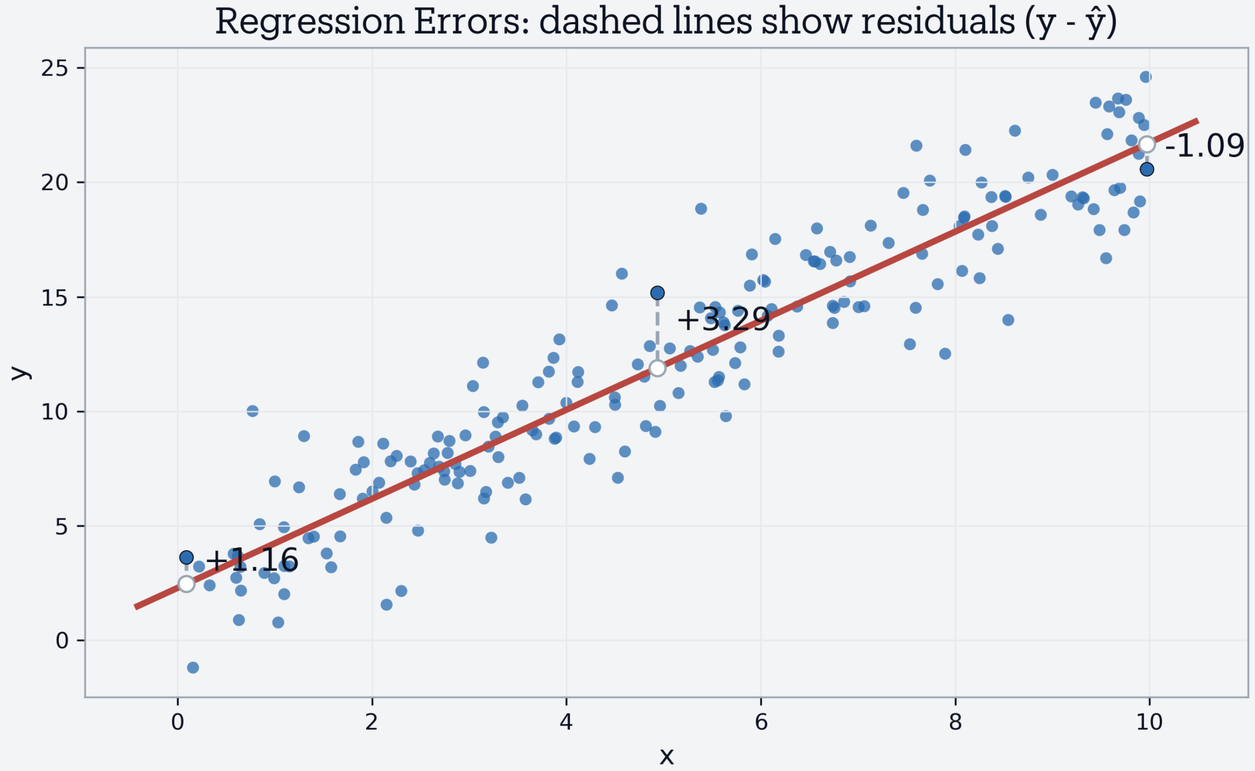
# Announcements

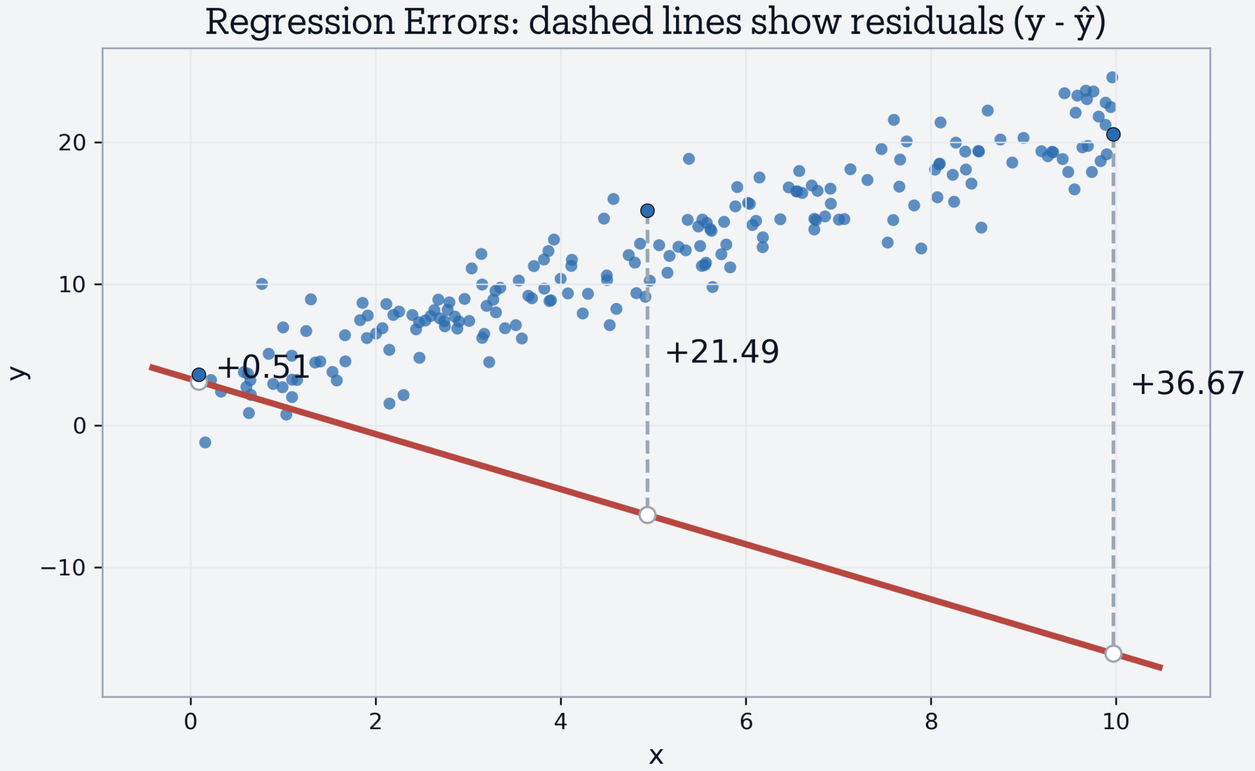
- Midterm next Thursday, study guide today
- Review next Tuesday
- Assignments and GitHub due Thursday


25
20
15
10
5
0
y
0
2
4
6
8
10
x


25
20
15
10
5
0
y
0
2
4
6
8
10
x
$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$


25
20
15
10
5
0
y
0
2
4
6
8
10
x




$$Y_i = \alpha + \beta X_i + \epsilon_i$$

# Model Fitting

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

# Sum of Squared Residuals (Residual Sum of Squares) ((Sum of Squared Errors))

$$SSR = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^{n} \epsilon_i$$

# Ordinary Least Squares (OLS) is BLUE

**B**est

**L**inear

**U**nbiased

**E**stimator

# Ordinary Least Squares

$$SSR = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial SSR}{\partial \beta_0} = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\partial SSR}{\partial \beta_1} = \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i)$$

# Ordinary Least Squares

$$SSR = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial SSR}{\partial \beta_0} = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial SSR}{\partial \beta_1} = \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

# Ordinary Least Squares

$$\frac{\partial SSR}{\partial \beta_0} = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

# Ordinary Least Squares

$$\frac{\partial SSR}{\partial \beta_1} = \sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\beta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

# Ordinary Least Squares

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\beta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

# Ordinary Least Squares

$$\beta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\beta_1 = \frac{\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

# Ordinary Least Squares

$$\beta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\beta_1 = \frac{\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \frac{Cov(X, Y)}{Var(X)}$$

# Assessing Model Fit

# Root Mean Squared Error

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2}$$

# $R^2$

$$SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$R^2 = 1 - \frac{SSR}{TSS}$$

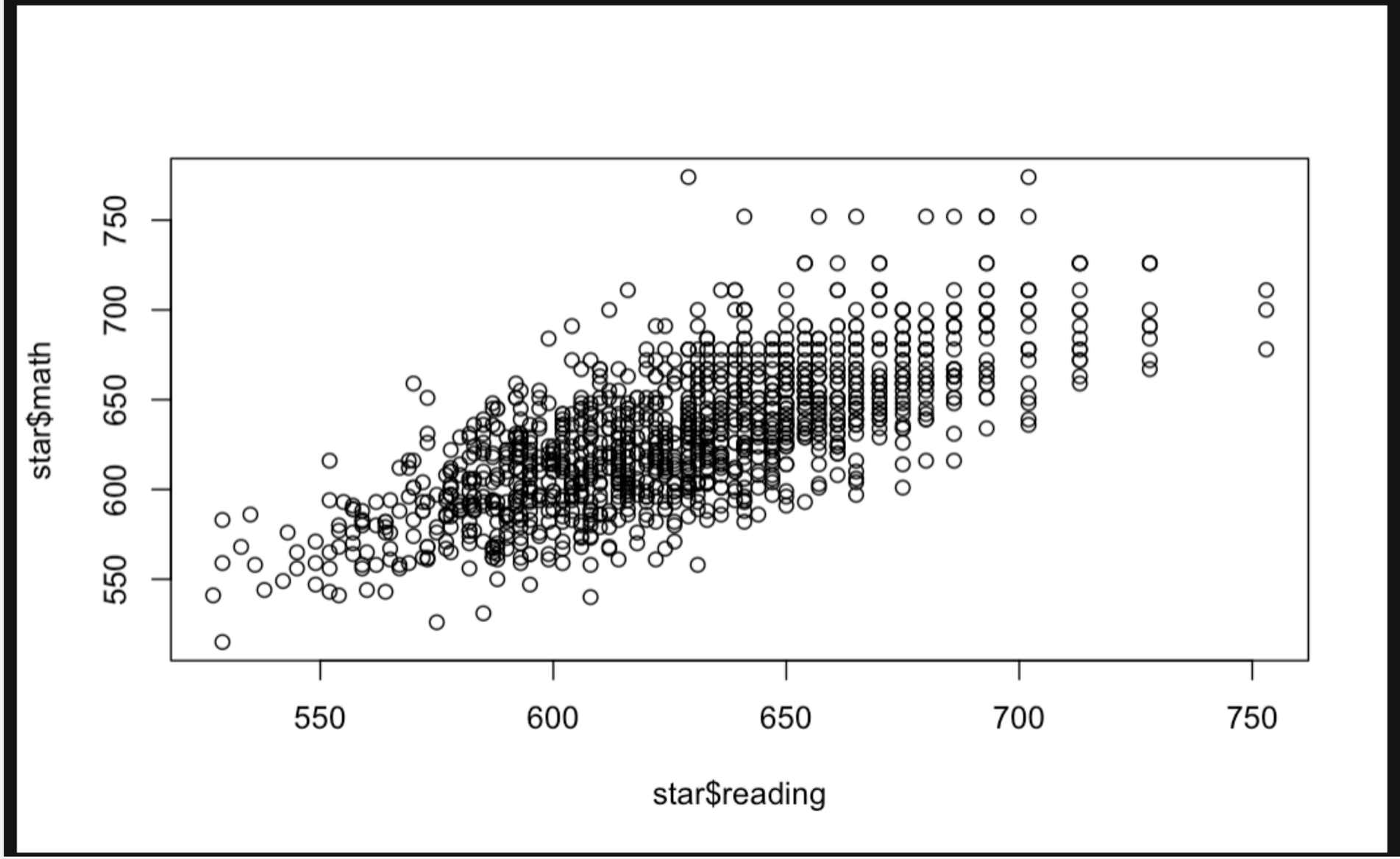
# Example with Continuous Predictor and Outcome

```{r}
star <- read.csv("https://raw.githubusercontent.com/MLBurnham/pols_602/refs/heads/main/data/STAR.csv")
head(star)
```

| | classtype <chr> | reading <int> | math <int> | graduated <int> |
|---|---|---|---|---|
| 1 | small | 578 | 610 | 1 |
| 2 | regular | 612 | 612 | 1 |
| 3 | regular | 583 | 606 | 1 |
| 4 | small | 661 | 648 | 1 |
| 5 | small | 614 | 636 | 1 |
| 6 | regular | 610 | 603 | 0 |

6 rows

```{r}
plot(star$reading, star$math)
```



```{r}
math_lm <- lm(math ~ reading, data = star)
summary(math_lm)
```

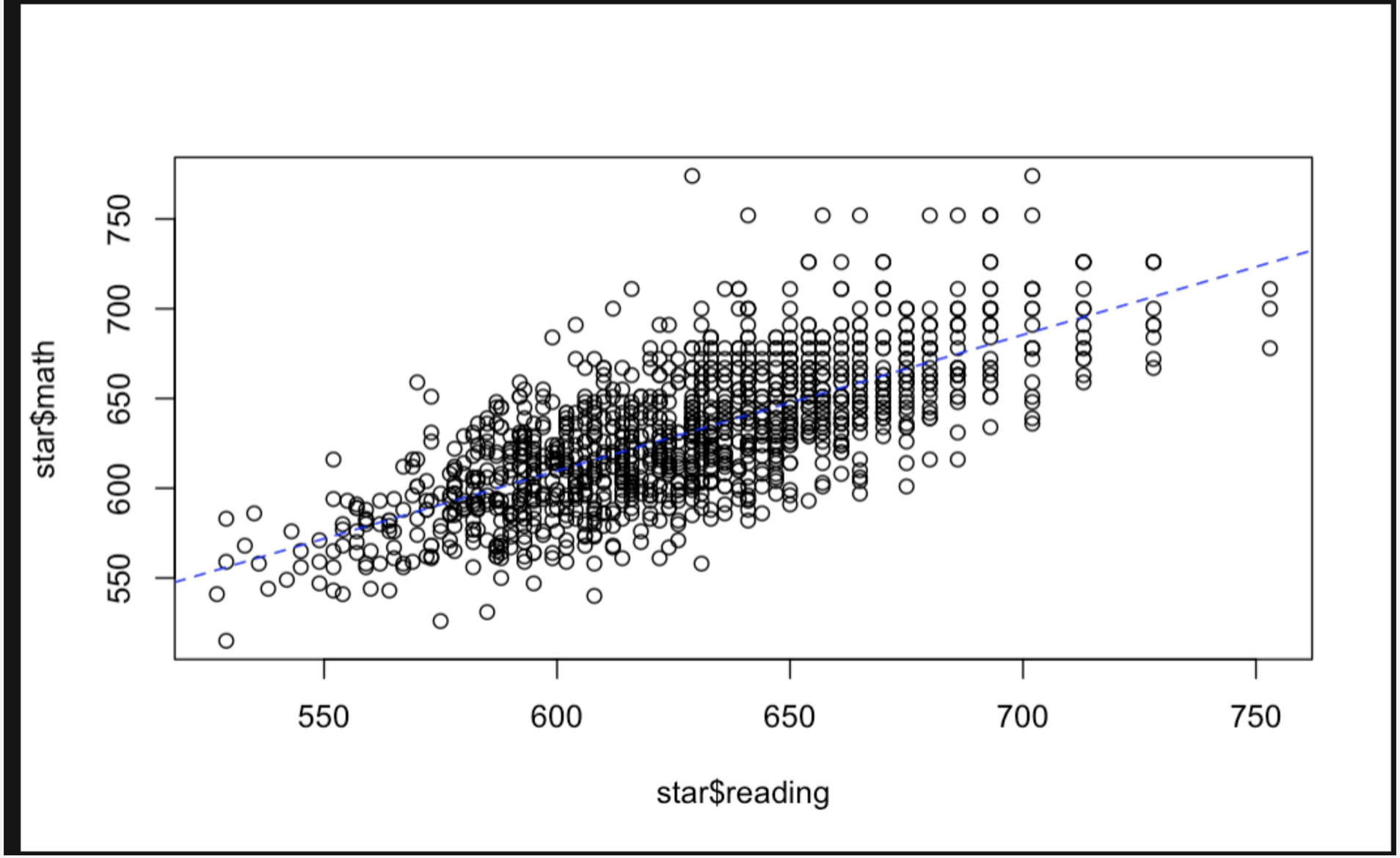
```
Call:
lm(formula = math ~ reading, data = star)

Residuals:
    Min      1Q  Median      3Q     Max
-75.835 -18.238  -1.462  17.243 142.263

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 155.43655   13.03448   11.93   <2e-16 ***
reading       0.75723    0.02069   36.59   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.12 on 1272 degrees of freedom
Multiple R-squared:  0.5128,    Adjusted R-squared:  0.5124
F-statistic:  1339 on 1 and 1272 DF,  p-value: < 2.2e-16
```

```{r}
math_lm <- lm(math ~ reading, data = star)
summary(math_lm)
```

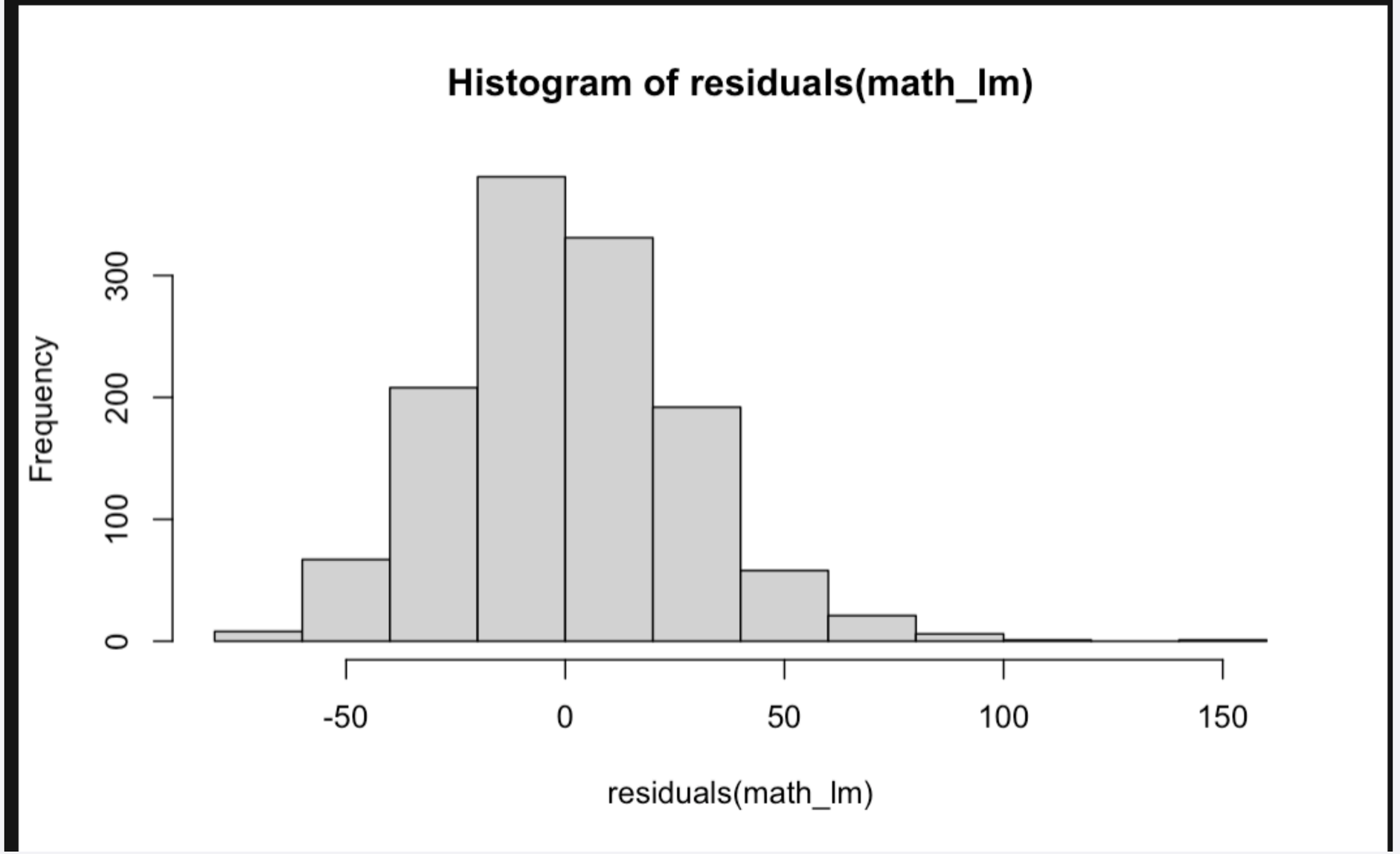
```
Call:
lm(formula = math ~ reading, data = star)

Residuals:
    Min      1Q  Median      3Q     Max
-75.835 -18.238  -1.462  17.243 142.263

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 155.43655   13.03448   11.93   <2e-16 ***
reading       0.75723    0.02069   36.59   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.12 on 1272 degrees of freedom
Multiple R-squared:  0.5128,    Adjusted R-squared:  0.5124
F-statistic:  1339 on 1 and 1272 DF,  p-value: < 2.2e-16
```

```
> math_lm$coefficients['reading']
  reading
0.7572346
```

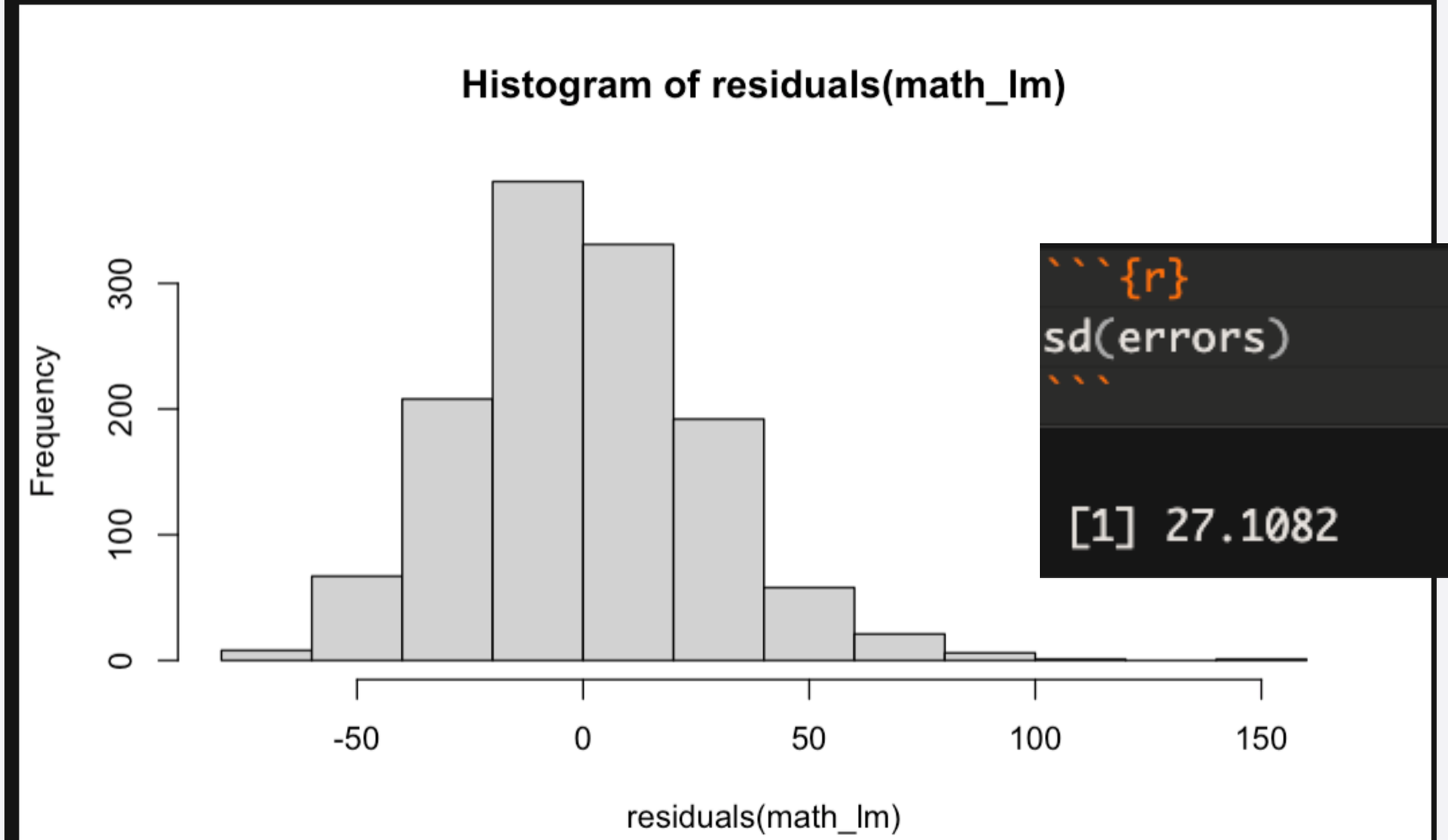
```{r}
plot(star$reading, star$math)
abline(math_lm, col = 'blue', lty='dashed')
```



```{r}
errors <- residuals(math_lm)
hist(errors)
```



```{r}
errors <- residuals(math_lm)
hist(errors)
```




```{r}
sd(errors)
```

```
[1] 27.1082
```

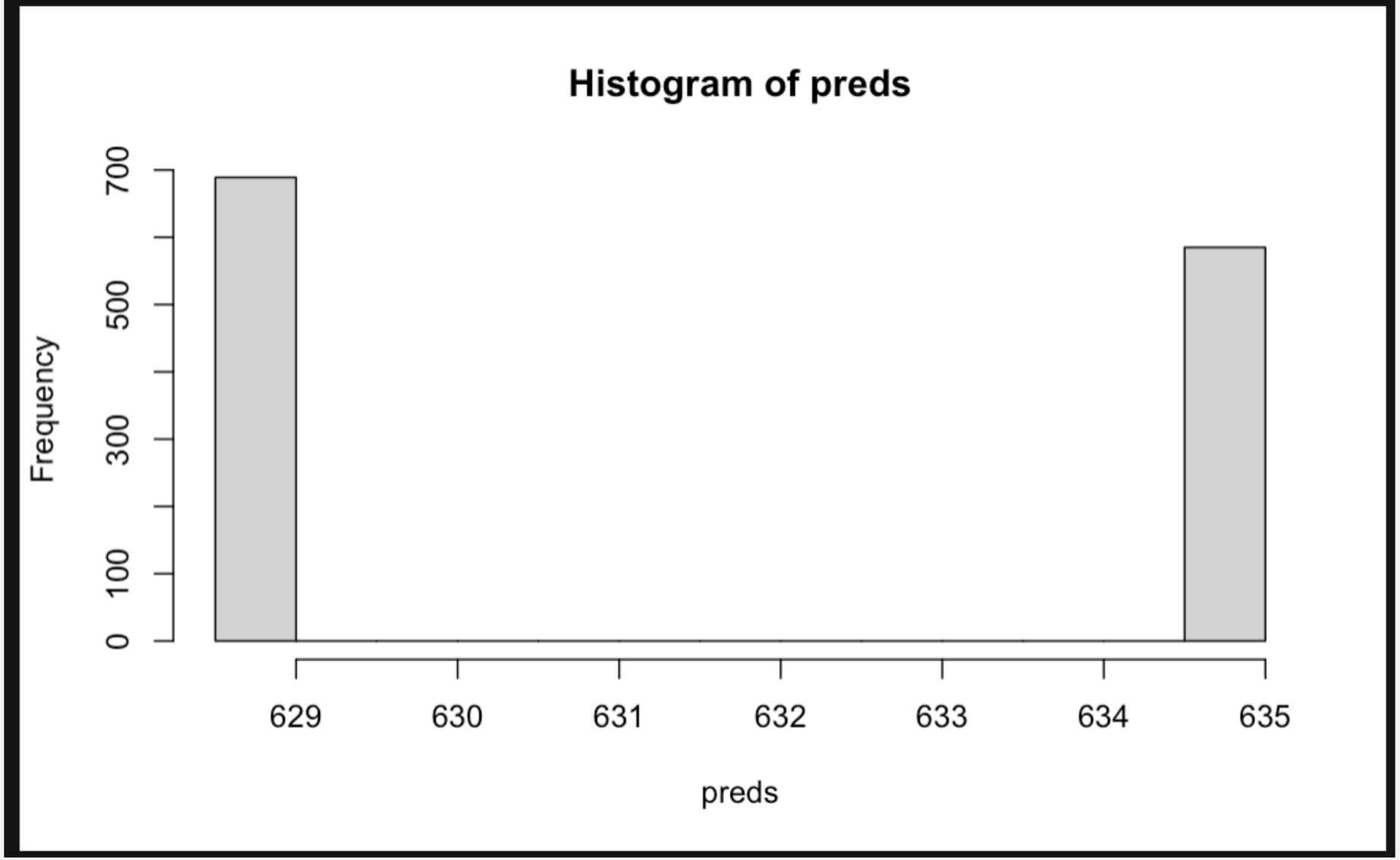
# Binary Predictor and Continuous Outcome

```{r}
math_lm2 <- lm(math~classtype, data = star)
summary(math_lm2)
```

```
Call:
lm(formula = math ~ classtype, data = star)

Residuals:
     Min       1Q   Median       3Q      Max
-119.827  -27.585   -0.827   26.163  145.163

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   628.837      1.476  426.09  < 2e-16 ***
classtypesmall    5.990      2.178    2.75  0.00604 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38.74 on 1272 degrees of freedom
Multiple R-squared:  0.005911,  Adjusted R-squared:  0.00513
F-statistic: 7.564 on 1 and 1272 DF,  p-value: 0.006039
```

```{r}
preds <- fitted.values(math_lm2)
hist(preds)
```

```{r}
preds <- fitted.values(math_lm2)
hist(preds)
```



```{r}
sc_math <- mean(star[star$classtype == 'small','math'])
rc_math <- mean(star[star$classtype == 'regular','math'])

round(sc_math - rc_math, 3)
```

```{r}
sc_math <- mean(star[star$classtype == 'small','math'])
rc_math <- mean(star[star$classtype == 'regular','math'])

round(sc_math - rc_math, 3)
```

```
[1] 5.99
```

# Example with Binary Outcomes

```{r}
math_lpm <- lm(graduated ~ math, data = star)
summary(math_lpm)
```

```
Call:
lm(formula = graduated ~ math, data = star)

Residuals:
     Min       1Q   Median       3Q      Max
-1.04098  0.03019  0.10784  0.16823  0.38176

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.4925316  0.1490061  -3.305 0.000975 ***
math         0.0021568  0.0002355   9.159  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3263 on 1272 degrees of freedom
Multiple R-squared:  0.06187,   Adjusted R-squared:  0.06114
F-statistic: 83.89 on 1 and 1272 DF,  p-value: < 2.2e-16
```

```{r}
plot(star$math, star$graduated)
abline(math_lpm, col = 'blue', lty='dashed')
```