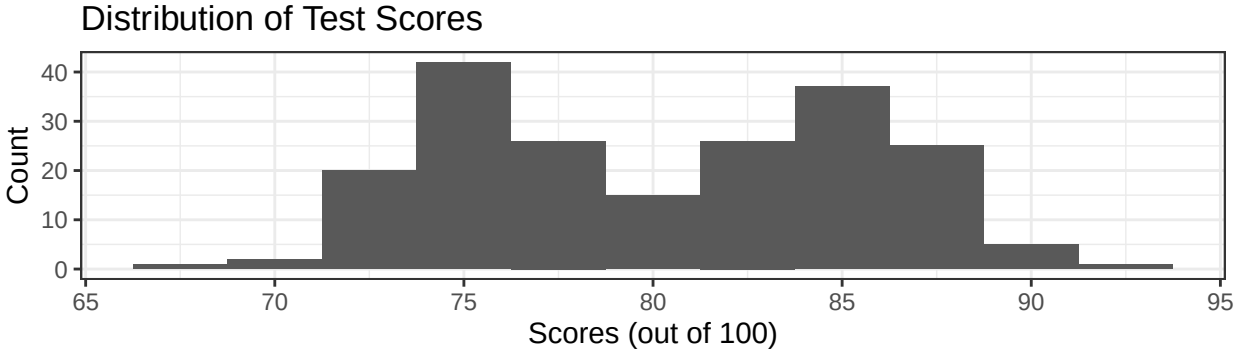
AM STAT. Imagine you are a researcher trying to understand the performance of a school district on a particular standardized test. You obtain the scores of 200 random students in the school district. The distribution of the scores for the students in your sample is displayed below, as are some summary statistics that relate to the distribution:



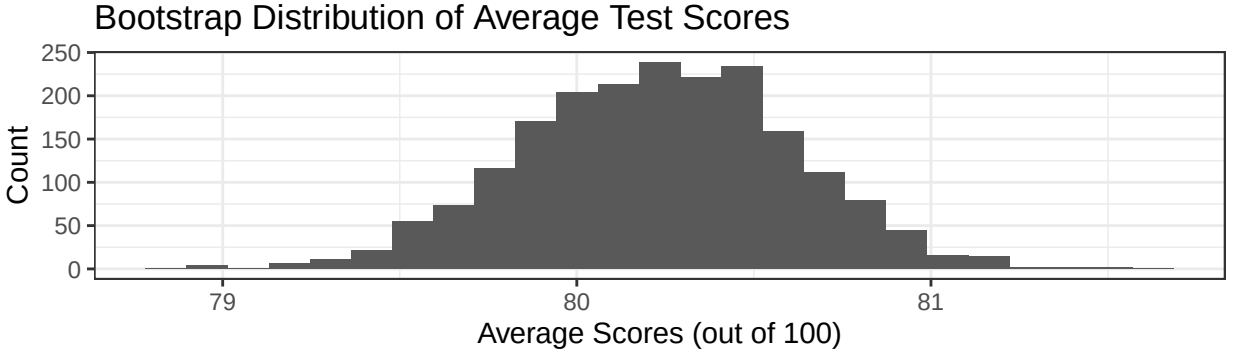
Mean	SD	Median	IQR
80.22	5.43	80.56	9.71

Quantiles:	0.025	0.05	0.10	0.20	0.80	0.90	0.95	0.975
	71.4	72.09	73.46	74.7	85.42	86.99	87.99	89.07

(a) Describe the above distribution of test scores using statistical terminology related to <u>each</u> of the following: (i) **shape**, (ii) **center**, and (iii) **spread**.

(b) Imagine you are particularly interested in knowing the *average* test score in the *entire* school district. In this setting, what is the **population**? What is the **sample**? What is the **parameter**? What is the **statistic**?

You decide to create a confidence interval for the average test score in the district using a bootstrap. You collect 2000 bootstrap samples, and calculate the average test score in each bootstrap sample. The distribution of your bootstrap statistics is displayed below, as are some summary statistics of the bootstrap distribution:



Mean	SD	Median	IQR
80.23	0.38	80.24	0.52

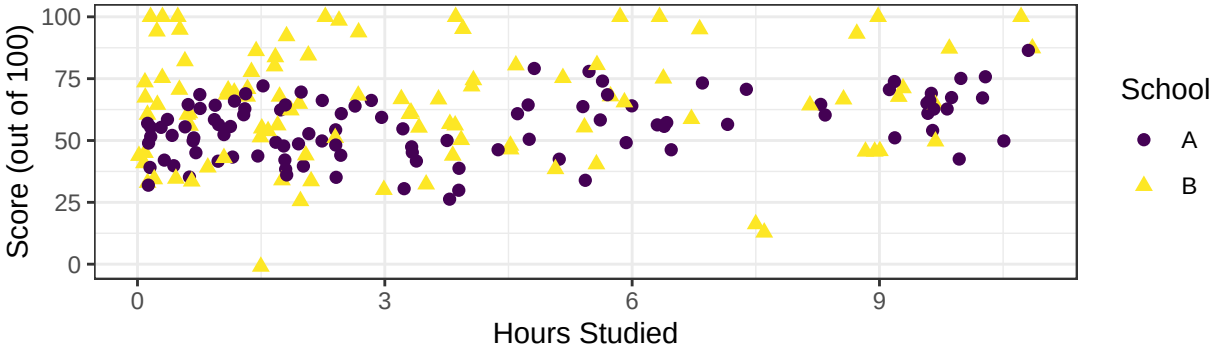
Quantiles:	0.025	0.05	0.10	0.20	0.80	0.90	0.95	0.975
	79.5	79.6	79.75	79.9	80.55	80.71	80.84	80.94

(d) Calculate a 95% confidence interval for your parameter of interest using the **standard error method**.
(e) Interpret your confidence interval from part (d) in context.
(f) Now create confidence intervals using the **percentile method**: create a 95% confidence interval, a 90% confidence interval, and an 80% confidence interval.
(g) In this setting, is there any reason why you should prefer the standard error method versus the percentile method when creating your confidence interval(s)? Why, or why not?

PM STAT. Imagine you are a student trying to decide how much you should study for a particular standardized test. You obtain the following information from 200 random students in the school district from the previous year:

- **the test score** (`Score`) of the student, out of 100 possible points,
- **the hours studied** (`Hours`) by the student before taking the test, and
- **the school attended** (`School`) by the student in the year they took the test (School A, or School B).

Below is a visualization of your data:



You decide to construct a linear regression model for `Score` by `Hours` and `School` of the form:

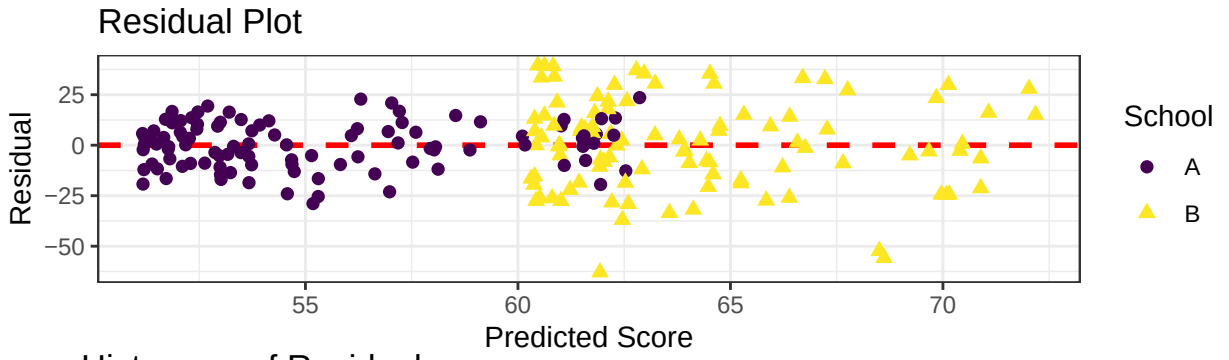
$$\widehat{\texttt{Score}} = \hat{\beta}_0 + \hat{\beta}_1 * \texttt{Hours} + \hat{\beta}_2 * \texttt{School\_B}$$

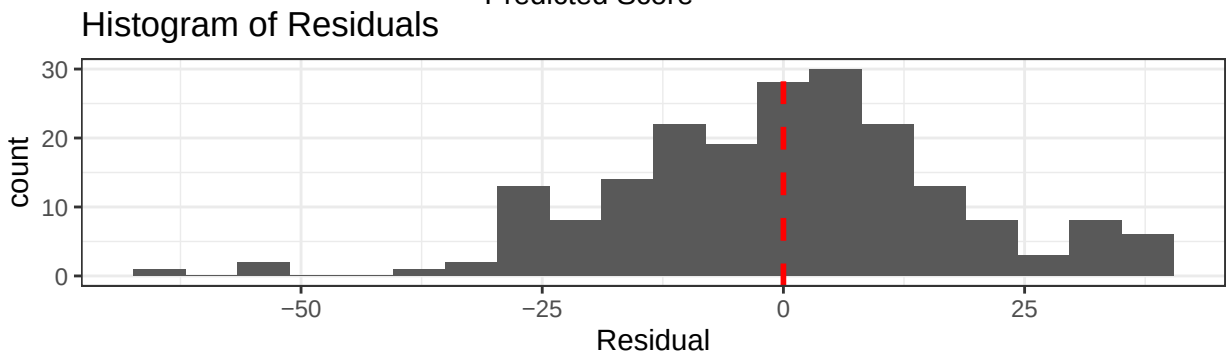
(a) Based on the regression table below, write down the formulas for (i) the full model, (ii) the model for School A, and (ii) the model for School B. *(You should include actual coefficient values from the table below in your model formulas.)*

```
## # A tibble: 3 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept    51.0      2.36      21.6    0       46.4      55.7
## 2 Hours         1.10     0.395      2.77   0.006    0.316     1.87
## 3 School: B     9.26     2.53       3.67   0        4.28     14.2
```

(b) For the regression table in part (a), carefully interpret, in context, (i) the intercept coefficient and (ii) the slope coefficient for `Hours`.

(c) Considering the scatterplot from the beginning of this question, as well as the two plots below, discuss whether or not each of the four common linear regression conditions holds for your model: (i) **Linearity**, (ii) **Independence**, (iii) **Normality**, and (iv) **Equal Variance**.





(d) You are interested in conducting a hypothesis test with the following null and alternative hypotheses:

$$H_0 : \beta_1 = 0 \qquad H_a : \beta_1 \neq 0$$

Explain, in context, (i) what a **Type I Error** would mean, and (ii) what a **Type II Error** would mean.

(e) Regardless of your answers to parts (c) and (d), use the regression table given in part (a) to conduct the hypothesis test from part (d), using a significance level of 0.05. Use this result to make a final determination: given the information available to you, should you study for this standardized test?