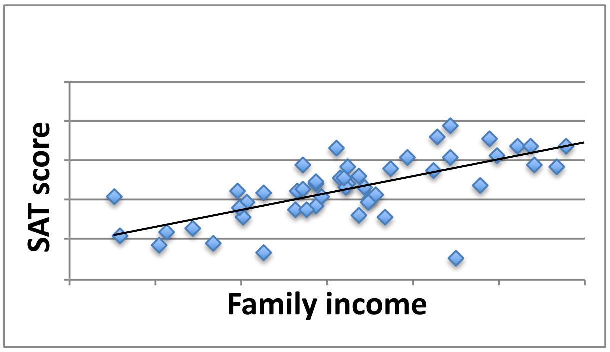
A warmup exercise

Education research has consistently shown that students from wealthier families tend to have higher SAT scores than those from poorer families. Given below are summary statistics on SAT score and family income for a random sample of 968 high school seniors.



	Mean	SD
SAT score (no units)	1222	123
Family income (in 1000 dollars)	75.8	9.2

A scatter plot of the sampled data suggests the relationship is approximately linear, and the correlation is $r = 0.68$. We want to construct a linear regression model to predict SAT scores from family income.

1. Identify the explanatory variable and the response variable.
2. Check whether the conditions for linear regression are met.
3. Find the equation of the regression line.
4. Find the R^2 value, and explain what it means in statistical terms.
5. Interpret what the slope indicates in this context.

Real-world extension of warmup

Here are some questions to think about:

- We're not really interested in the sample, right? What do we really want to know about SAT scores and family income?
- Regression model looks like: $\hat{y} = b_0 + b_1 x \quad$ (for some numbers b_0, b_1). It is the best fit for the **sampled data** – not for the **real world**.
- How (Can?) we find the best fit regression line for the real world?
- Assume the real world relationship is also approximately linear.
Q: Is that guaranteed just from the fact that the sample is linear?
- Suppose the best-fit linear model for the real world is: $\hat{y} = \beta_0 + \beta_1 x$.
- Note that β_0, β_1 are population parameters, while b_0, b_1 are estimates for those parameters found from a sample.
- Can we use b_0, b_1 in some way to find β_0, β_1? What further information would you need to make that possible?