# Name:

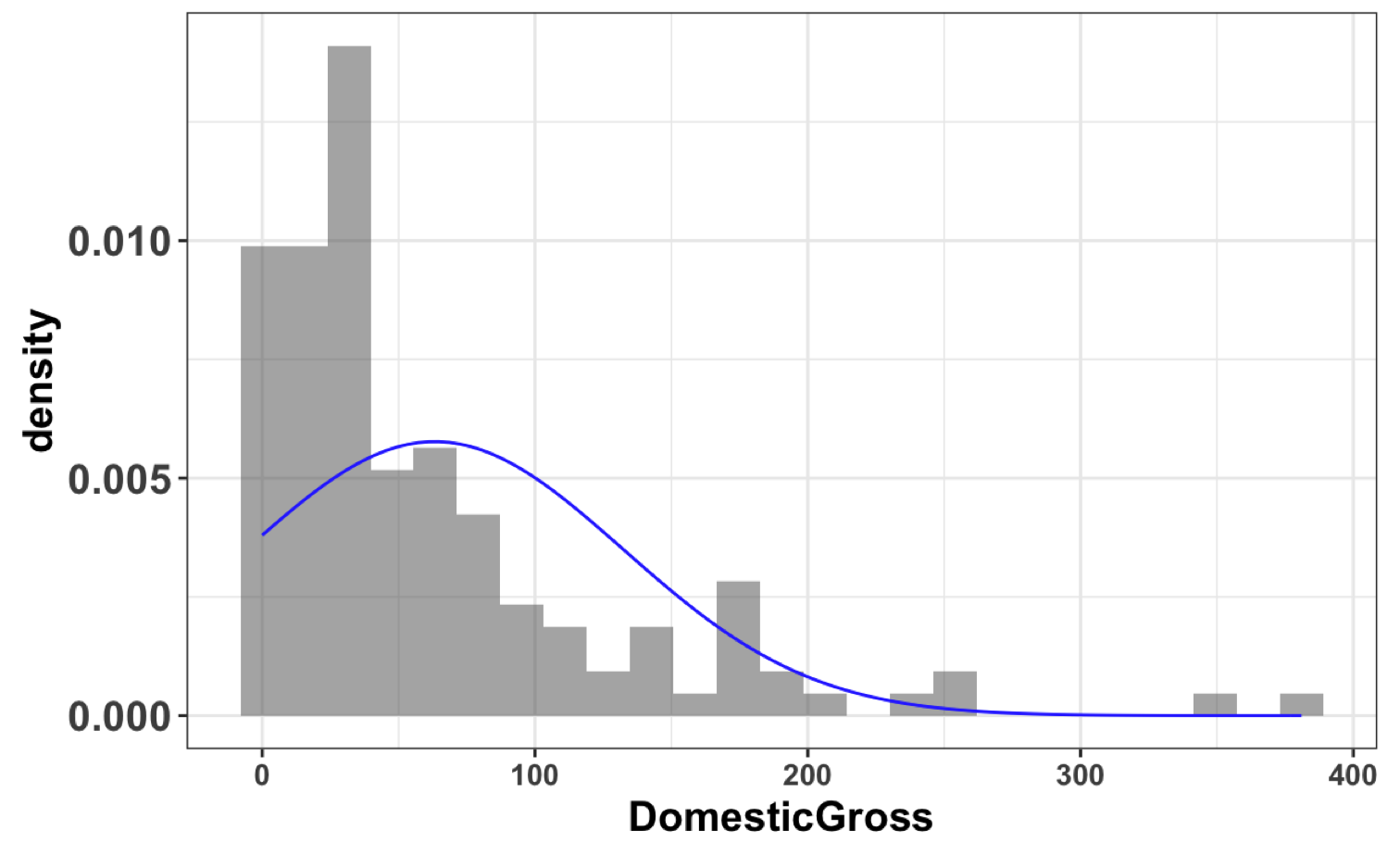
# Classwork 14: Do big budget movies make big dollars?

Let’s take a closer look at **HollywoodMovies2011**, a data frame that you have seen before in a practice quiz.

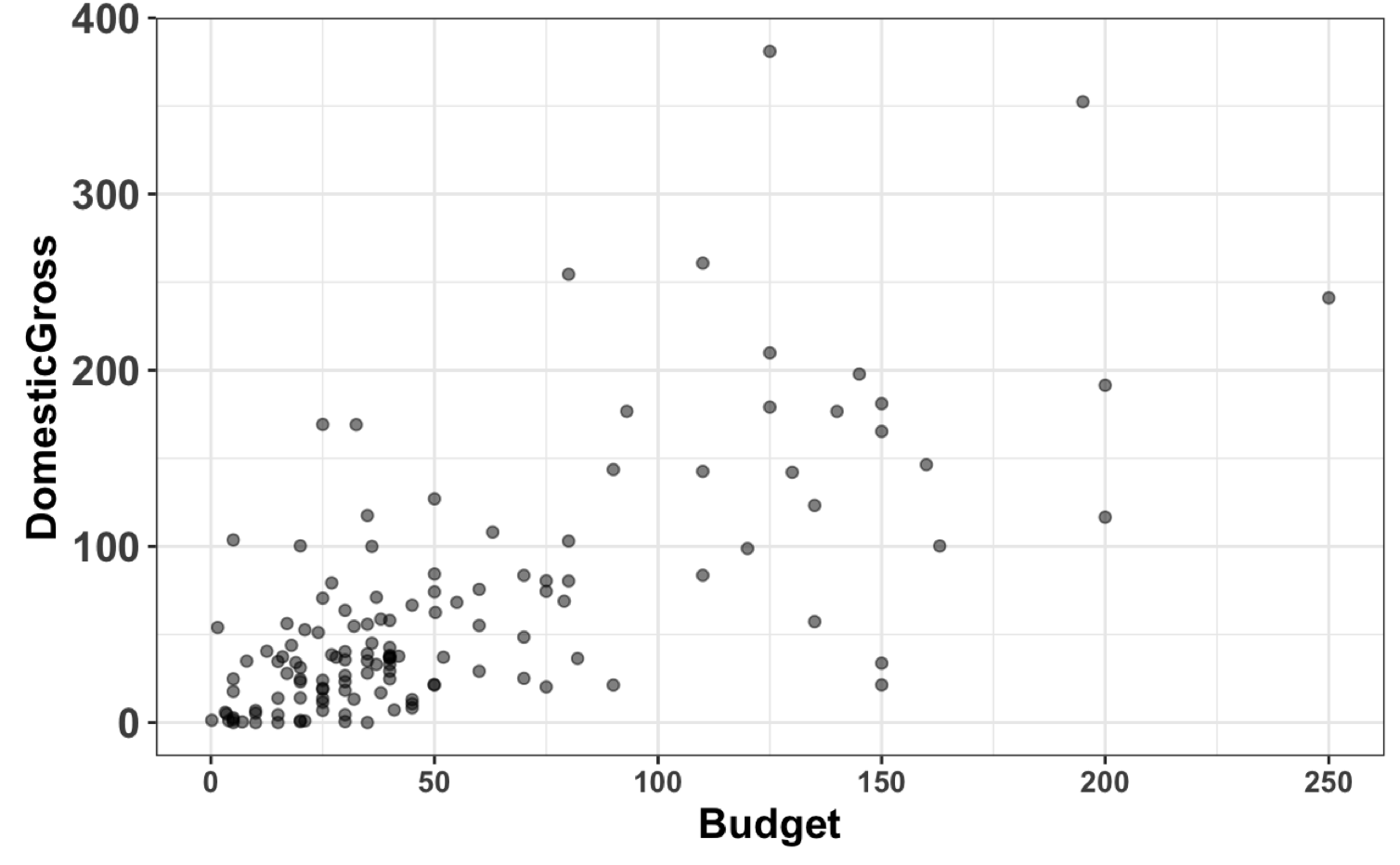
A data frame with 136 observations on the following 14 variables.

* Movie Title of movie
* LeadStudio Name of studio
* RottenTomatoes The Tomatometer score represents the percentage of professional critic reviews that are positive for a given film or television show. A Tomatometer score is calculated for a movie or TV show after it receives at least five reviews.
* AudienceScore The Audience Score is the percentage of users who have rated the movie or TV Show positively.
* Story General theme (there are 21 themes)
* Genre There are 9 categories Action Adventure Animation Comedy Drama Fantasy Horror Romance Thriller
* TheatersOpenWeek Number of screens for opening weekend.
* BOAverageOpenWeek Average box office income per theater – opening weekend
* DomesticGross Gross income for domestic viewers (in millions)
* ForeignGross Gross income for foreign viewers (in millions)
* WorldGross Gross income for all viewers (in millions)
* Budget Production budget (in millions)
* Profitability WorldGross divided by Budget
* OpeningWeekend Opening weekend gross (in millions)

1. Here is a histogram of **DomesticGross**. Do you think that the **Budget** might be able to explain some of the variation we see in **DomesticGross**? Why or why not?



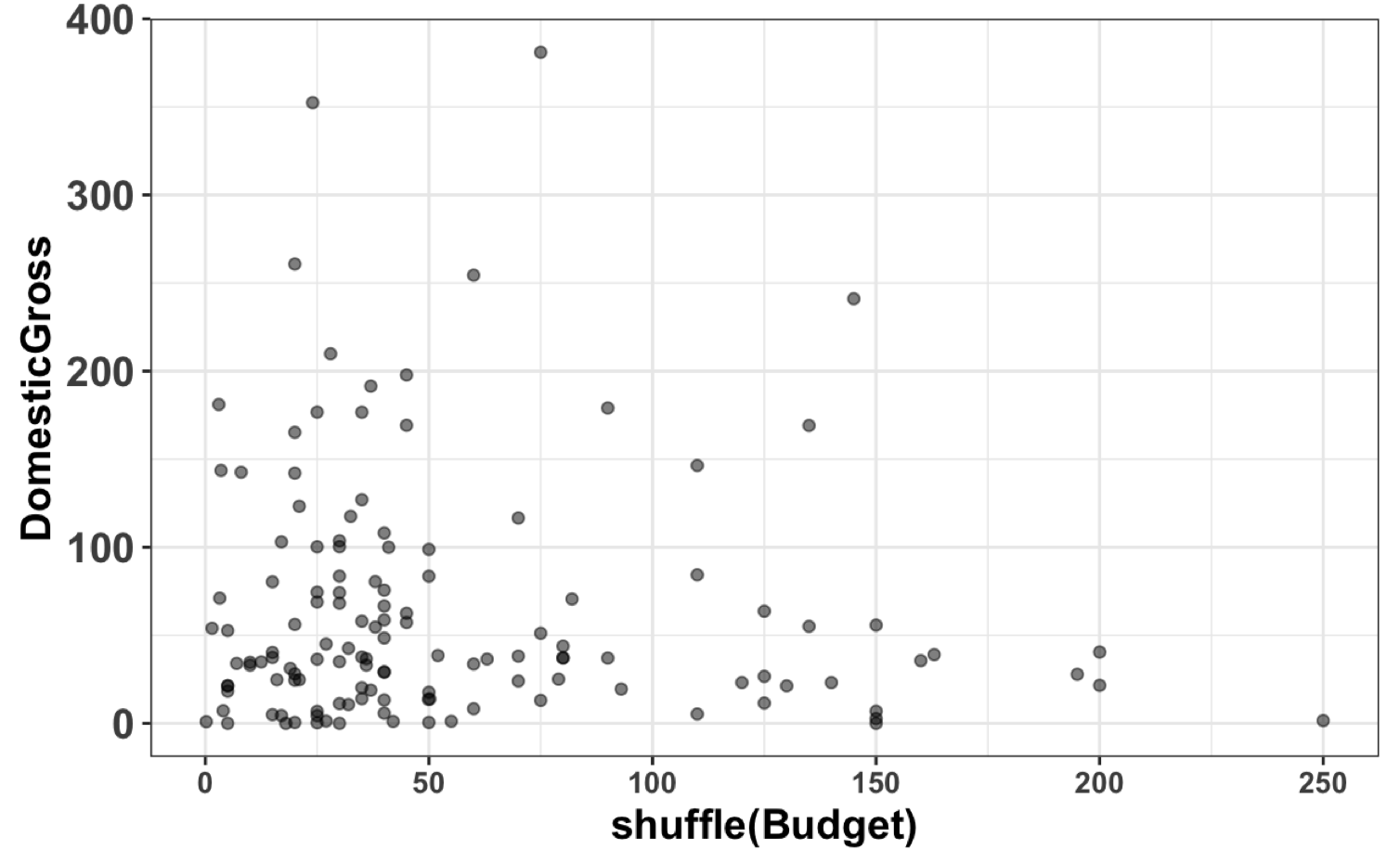
1. What does the curve represent?
2. Write a word equation for the idea that **Budget** might explain some of the variation seen in **DomesticGross**.
3. If we wanted to make a visualization to explore this idea, what would we make? Why that one?
4. Which variable should be depicted on the y-axis? Why?
5. Draw the empty model on this scatterplot. Also draw the standard deviation and variance. (Use blue if you can.)



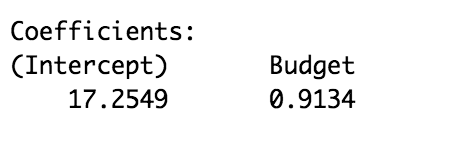
1. Even if you successfully make a scatterplot, you might get an error message. What do you think this error message mean? Can we do something to “clean” this situation up?
2. Try making a scatterplot again – have we “cleaned” our data frame up enough? [Learn a new function! na.omit()—this is not in book].

The following questions use the cleaned up data frame: **Movies**.

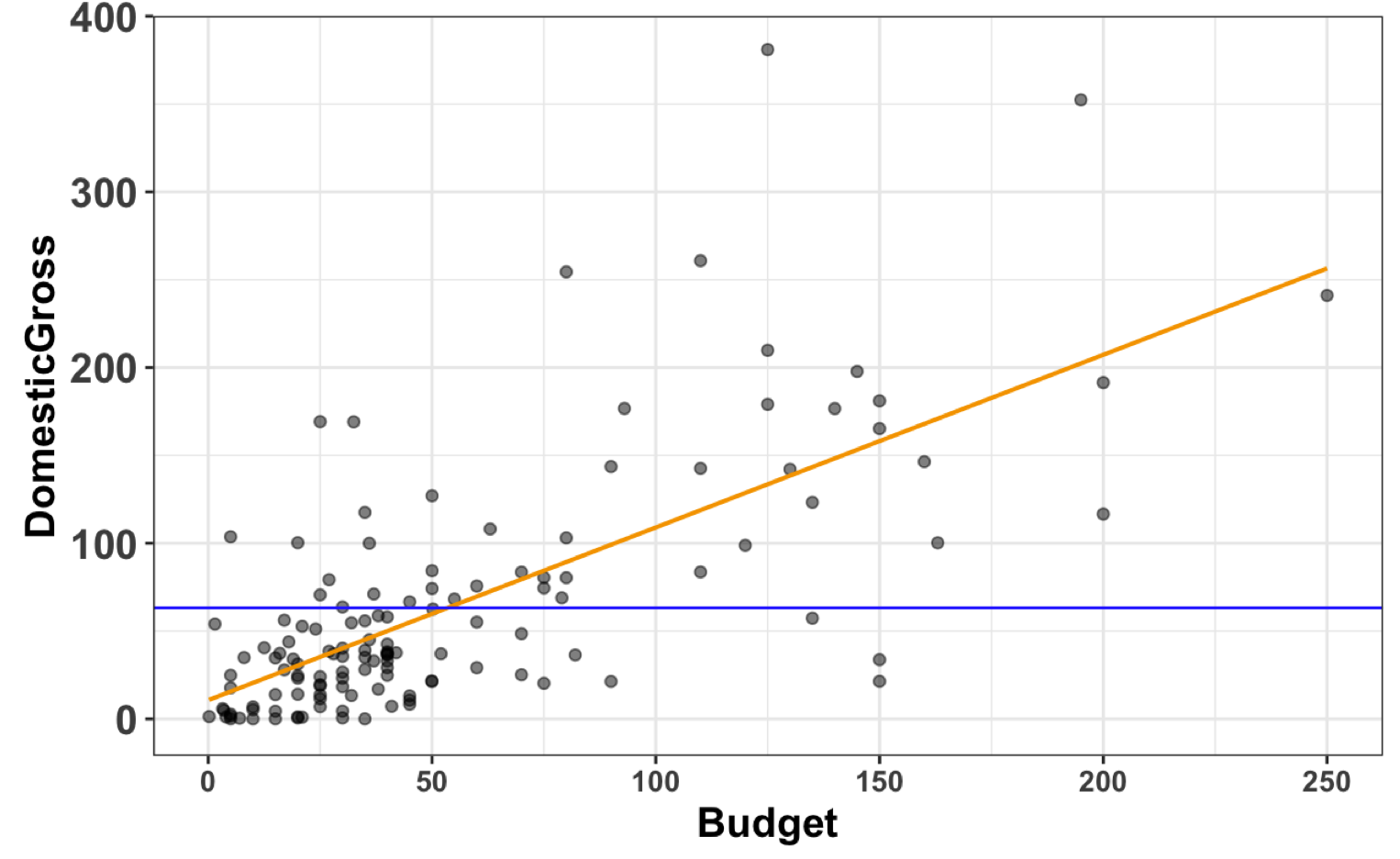
1. Describe the pattern (the “positive correlation”) that you see in your scatterplot.
2. What’s your intuition: Would it be easy for a random DGP (such as shuffle) to create this pattern?
3. Why is shuffle an example of a random DGP?
4. Here is a scatterplot where **Budget** values were shuffled. Why is it hard for a random DGP to produce a pattern like the empirical data?



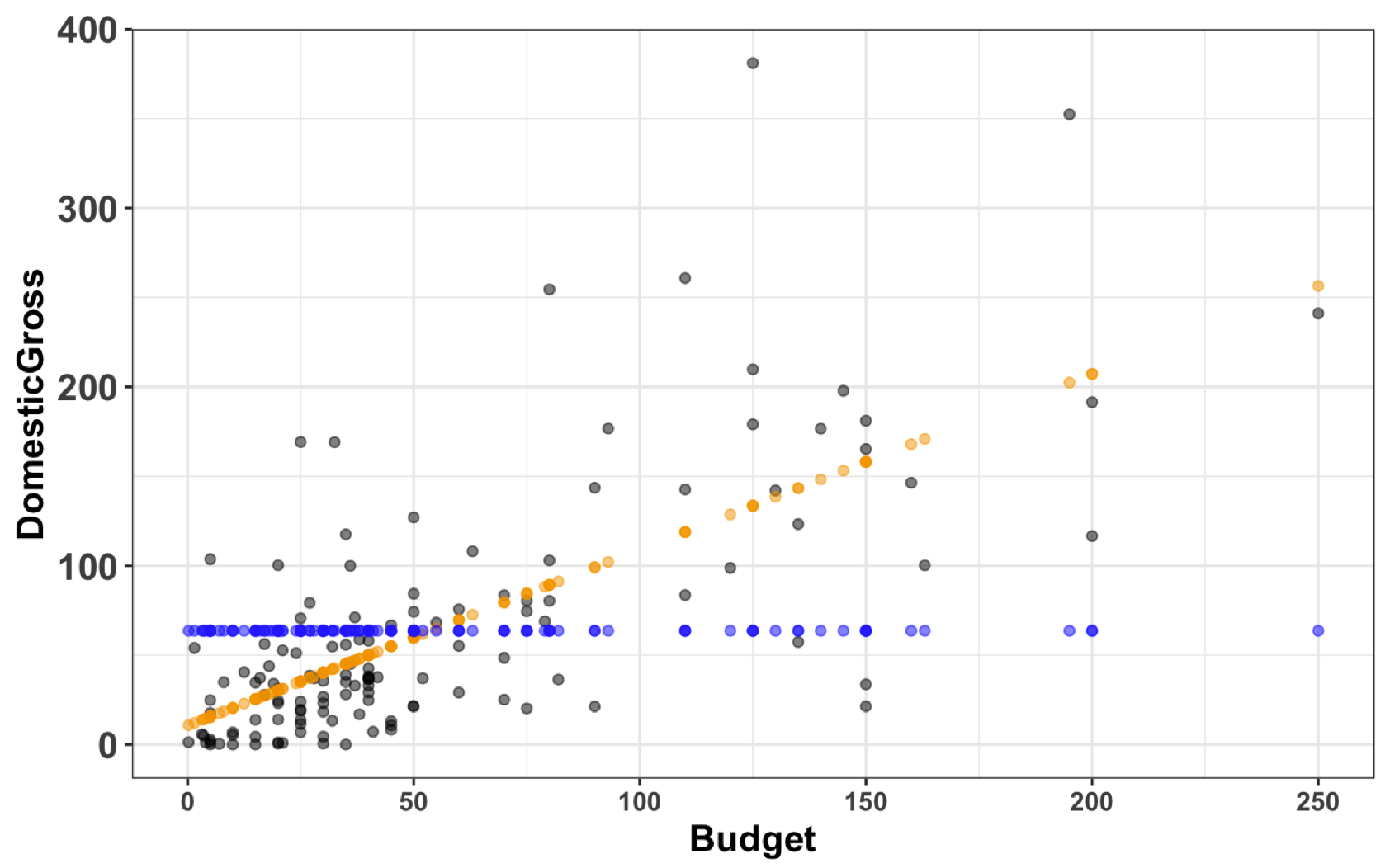
1. If you fit a model that predicts **DomesticGross** by including **Budget** as an explanatory variable, how many parameter estimates would you need? (You may want to write out the GLM notation… it might help!)
2. Let’s fit a model of **DomesticGross** predicted by **Budget** and call it **Budget.model**. Write the R code you would need to produce this output:



1. Write the GLM notation for this model.
2. If a movie had a Budget of 0 dollars (wow, super cheap!), what would the **Budget.model** predict would be its **DomesticGross**?
3. If a movie had a Budget of 1,000,000 dollars (still pretty cheap for as movie budgets go!), what would the **Budget.model** predict would be its **DomesticGross**?
4. What do each of the parts in the **Budget** model (written in GLM notation) mean?
5. If we use the *predict()* function on the Budget model, what would it do?
6. If we use the *resid()* function on the Budget model, what would it do?
7. Pixar’s Cars 2 had a budget of 200 million dollars and made 191 million in the domestic market. Would the Budget model predict Cars 2’s **DomesticGross** to be 191? Why or why not?
8. What is the error () from the Budget model for Cars 2? That is, what is the actual value?
9. Write the (additional) R code to put the Budget model on the scatterplot. (I also included the empty model.)



1. Now write the code to put the actual *predictions* of the Budget model on the scatterplot as well (like the picture below). I also included the predictions from the empty model. Why are the dots in a line?



1. How much of the error is explained by the Budget model?

