Predictive Data Modeling: BoodleBox Review

Rento Saijo

February 11, 2026

Instructions

Use BoodleBox (LLAMA4) to better understand Chapter 3 on fitting a line (the linear regression model). You may ask it to fit a line to a dataset, explain what a "good fit" means for your data, or clarify ideas from the text. Try to be specific, and if you are not getting the answer you are looking for, refine your prompt using more technical terms from the chapter.

In this assignment, document your interactions by answering the following:

1. What did you ask?
2. Was it helpful initially, or were you able to refine your interaction to get a better response?
3. Were you able to have some discussion about line fitting?
4. Did you provide some data and get a model fit to that data with some explanation?
5. How much time did you spend on this effort?

1. **What did you ask?**

   I focused on two core topics from our Chapter 3 "fit a line" section: (i) why the standard objective is the *sum of squared errors* (SSE) rather than the *sum of absolute errors* (SAE), and (ii) how to think about and improve the randomized strategy we practiced in class (randomly choosing slope/intercept values under constraints, evaluating the error, and keeping the best). I also asked LLAMA4 to fit a line to a commonly known dataset and explain what "good fit" means using regression terms such as residuals, SSE, and R^2.

   Prompt 1 (conceptual): "In linear regression, why do we usually minimize sum of squared errors instead of sum of absolute errors? Please connect it to assumptions about noise and show how the math changes."

   Prompt 2 (random search strategy): "In class we did a randomized approach: sample random β_0, β_1 (with constraints), compute SSE, keep the best. How can we optimize this strategy? Please compare to gradient descent and the closed-form solution."

   Prompt 3 (fit a line to data): "Fit a line to the Boston Housing dataset using $x =$ `lstat` (percent lower status of the population) and $y =$ `medv` (median home value). Give the fitted model, explain the meaning of slope/intercept, and discuss goodness-of-fit (SSE and R^2)."

2. **Was it helpful initially, or were you able to refine your interaction to get a better response?**

   The initial answers were moderately helpful but a bit generic. When I refined my prompts using Chapter 3 terms like *objective function*, *residuals* $r_i = y_i - \hat{y}_i$, *SSE*, *convexity*, and *normal equations*, the responses became much more aligned with what we covered in class. For example, when I first asked about "squared vs absolute error," LLAMA4 gave high-level reasons ("squares penalize big errors more," which is more so what we discussed in class). After I explicitly asked it to connect to a noise model and to contrast how optimization changes, it responded in a more technical way (differentiability, closed-form

solution, and the Gaussian vs Laplace likelihood interpretation), which matched the chapter's emphasis on computation and fit criteria.

3. **Were you able to have some discussion about line fitting?**

   Yes. A useful part of the discussion was reframing line fitting as: choose parameters β_0, β_1 to minimize an objective over residuals. With the linear model

   $$\hat{y}_i = \beta_0 + \beta_1 x_i, \qquad r_i = y_i - \hat{y}_i,$$

   the usual least squares objective is

   $$\mathrm{SSE}(\beta_0, \beta_1) = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2 .$$

   LLAMA4 emphasized that SSE is a convex quadratic function in (β_0, β_1) (for ordinary linear regression), which implies there is a single global minimum and standard methods (normal equations or gradient descent) reliably find it. It also helped to talk about what "good fit" means in this context: small residuals, small SSE (or MSE), and a higher R^2 value when comparing to a baseline model that predicts the mean of y. We also discussed that R^2 is not the only metric (especially when comparing models or dealing with outliers), but it is a common summary for simple linear regression.

4. **Did you provide some data and get a model fit to that data with some explanation?**

   Yes. I used the commonly known Boston Housing dataset, fitting $y = \texttt{medv}$ as a function of $x = \texttt{lstat}$. I noted to LLAMA4 that `lstat` is the percentage of the population that is "lower status," and `medv` is the median home value. I asked for the fitted line and for interpretation in terms of residuals and fit quality. A standard simple linear regression fit for this dataset (with `medv` predicted from `lstat`) is:

   $$\widehat{\texttt{medv}} = 34.554 - 0.950\,\texttt{lstat}.$$

   Interpretation (as discussed with LLAMA4): the slope -0.950 means that increasing `lstat` by 1 percentage point is associated with about a 0.95 decrease in predicted `medv`, on average, under the linear model. The intercept 34.554 is the predicted `medv` when `lstat` $= 0$, which is an extrapolation beyond typical data values but still defines the best-fitting line in the least squares sense. For goodness-of-fit, LLAMA4 summarized that this model has an R^2 around:

   $$R^2 \approx 0.544,$$

   meaning `lstat` alone explains about 54% of the variation in `medv` (relative to the mean-only baseline) under this simple linear model. LLAMA4 also described how SSE is computed from the residuals $(y_i - \hat{y}_i)$ and that minimizing SSE is exactly what produced the fitted coefficients above. To make the "data provided" aspect concrete, I included a small sample of (`lstat`, `medv`) points in the prompt (I noted that the full dataset contains many more observations):

lstat	medv
4.98	24.0
9.14	21.6
4.03	34.7
2.94	33.4
5.33	36.2
10.26	22.9
8.77	27.1
12.43	16.5
19.15	18.9
29.93	11.8

5. **How much time did you spend on this effort?**

   I spent about 45 minutes total. Roughly 15 minutes were spent getting an initial response and noticing it was too generic, about 20 minutes were spent refining prompts using Chapter 3 terminology and asking follow-up questions about SSE vs SAE and the randomized strategy, and about 10 minutes were spent asking for and interpreting the Boston Housing line fit and goodness-of-fit summary. And admittedly too long on the LaTeX formatting, but it was good practice nonetheless.