Prediction Intervals

CS1090A Introduction to Data Science

Pavlos Protopapas, Kevin Rader, and Chris Gumb

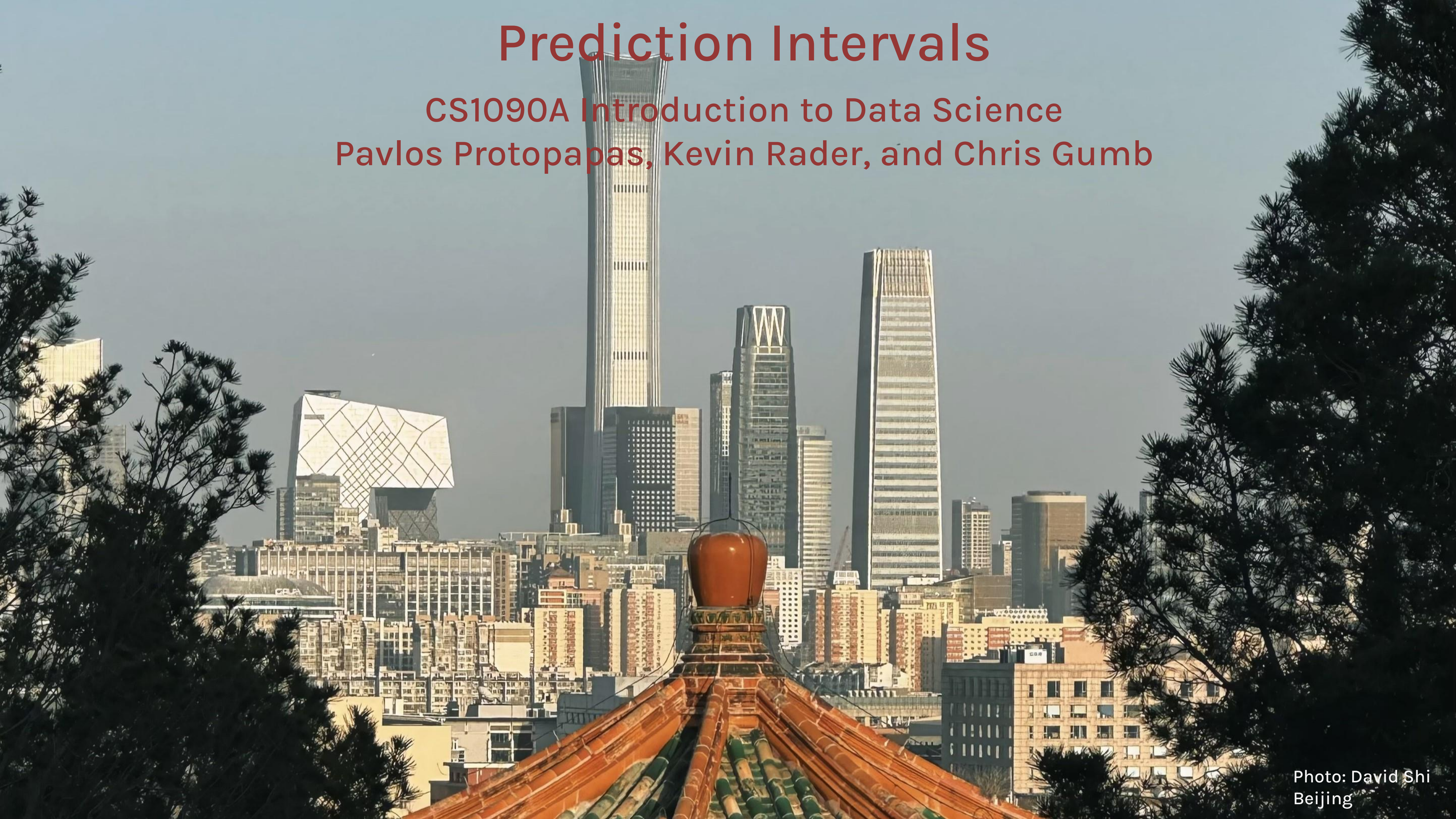


Photo: David Shi
Beijing

Outline

Part A and B: Assessing the Accuracy of the Coefficient Estimates

Bootstrapping and confidence intervals

Part C: How well do we know $\hat{f}$

The confidence intervals of $\hat{f}$

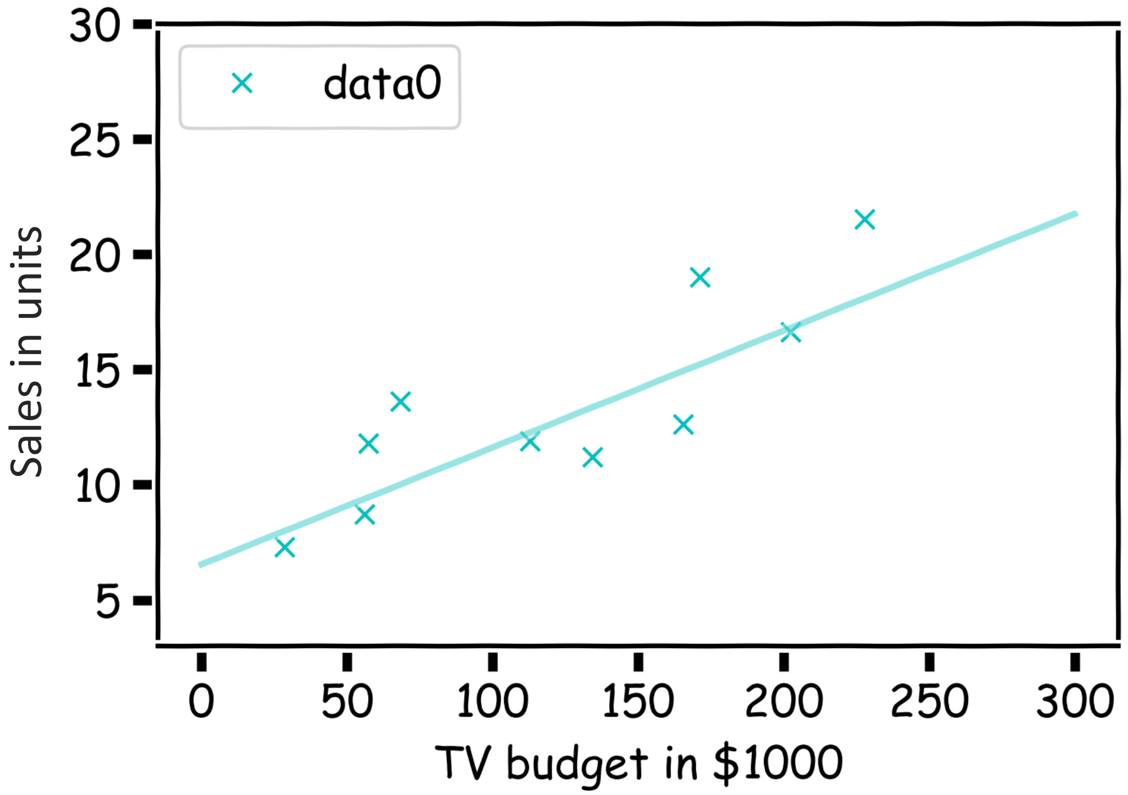
Part D: Evaluating Significance of Predictors

Does the outcome depend on the predictors?

Hypothesis testing

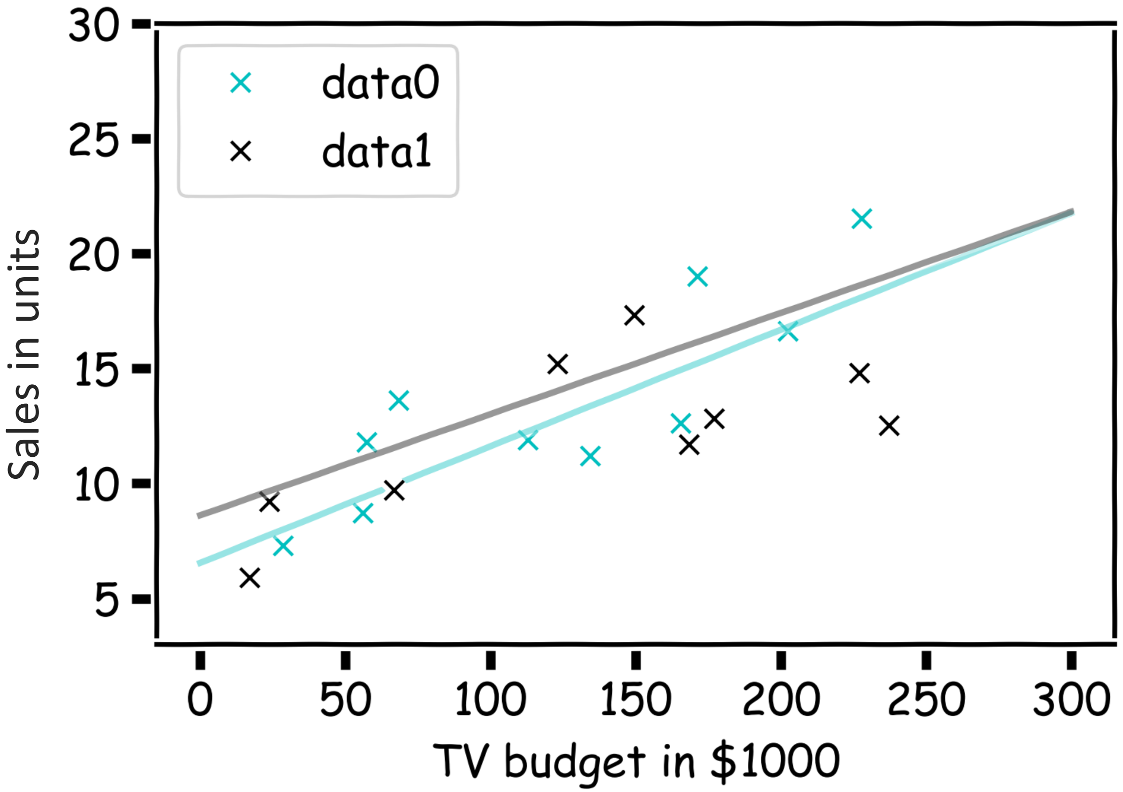
How well do we know $\hat{f}$?

Our confidence in f is directly connected with our confidence in βs. For each bootstrap sample, we have one β, which we can use to determine the model, $f(x) = X\beta$.



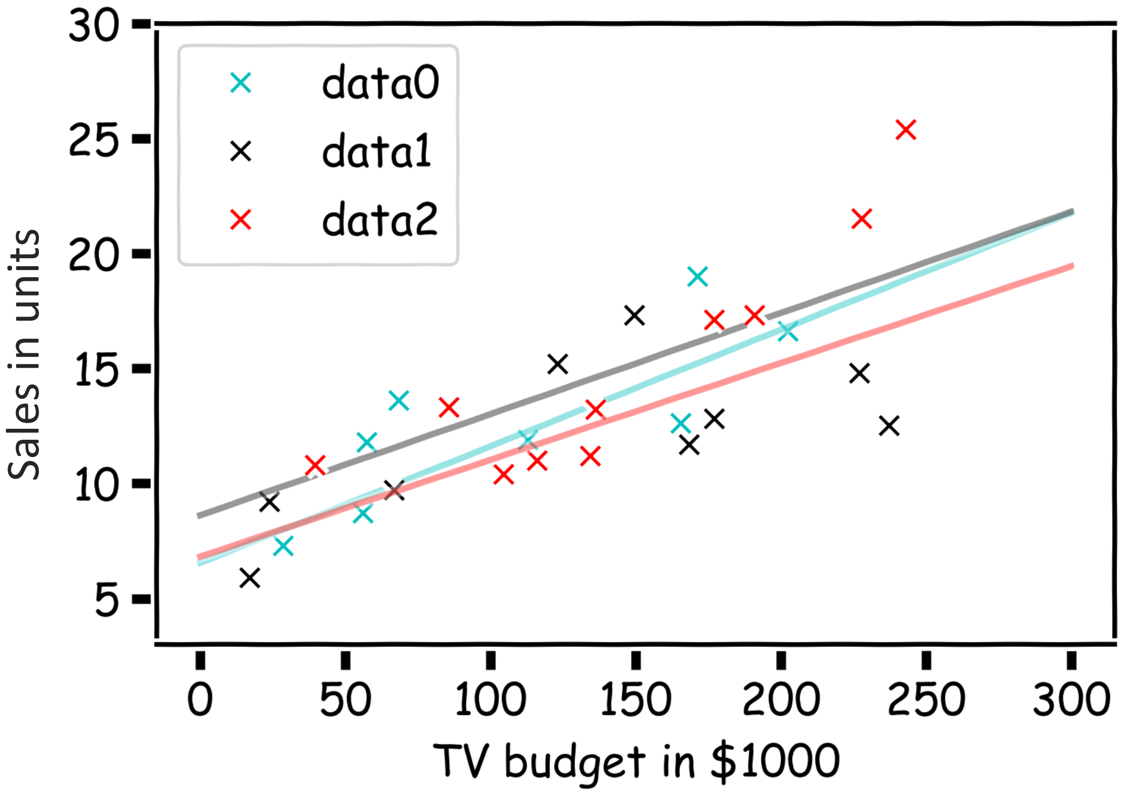
How well do we know $\hat{f}$?

Here we show two different models' predictions given the fitted coefficients, fit on two separate bootstrapped sets of data.



How well do we know $\hat{f}$?

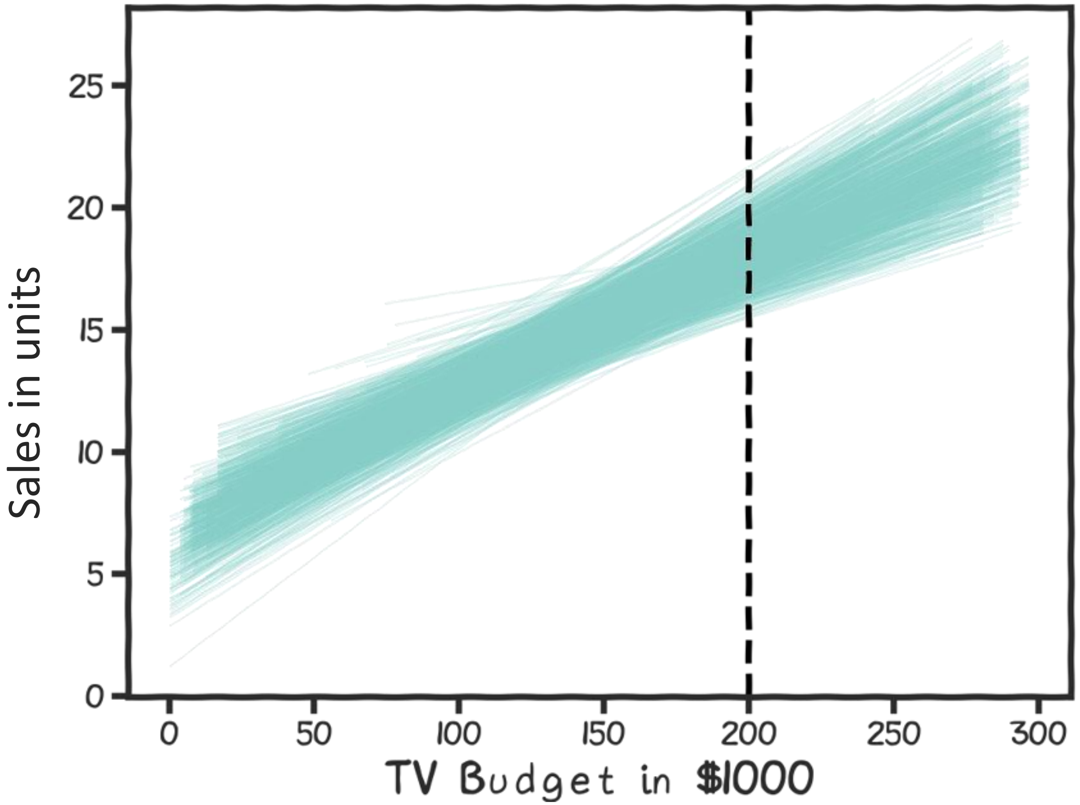
There is one such regression line for every bootstrapped sample.



How well do we know $\hat{f}$?

Below we show all regression lines for a thousand of such bootstrapped samples.

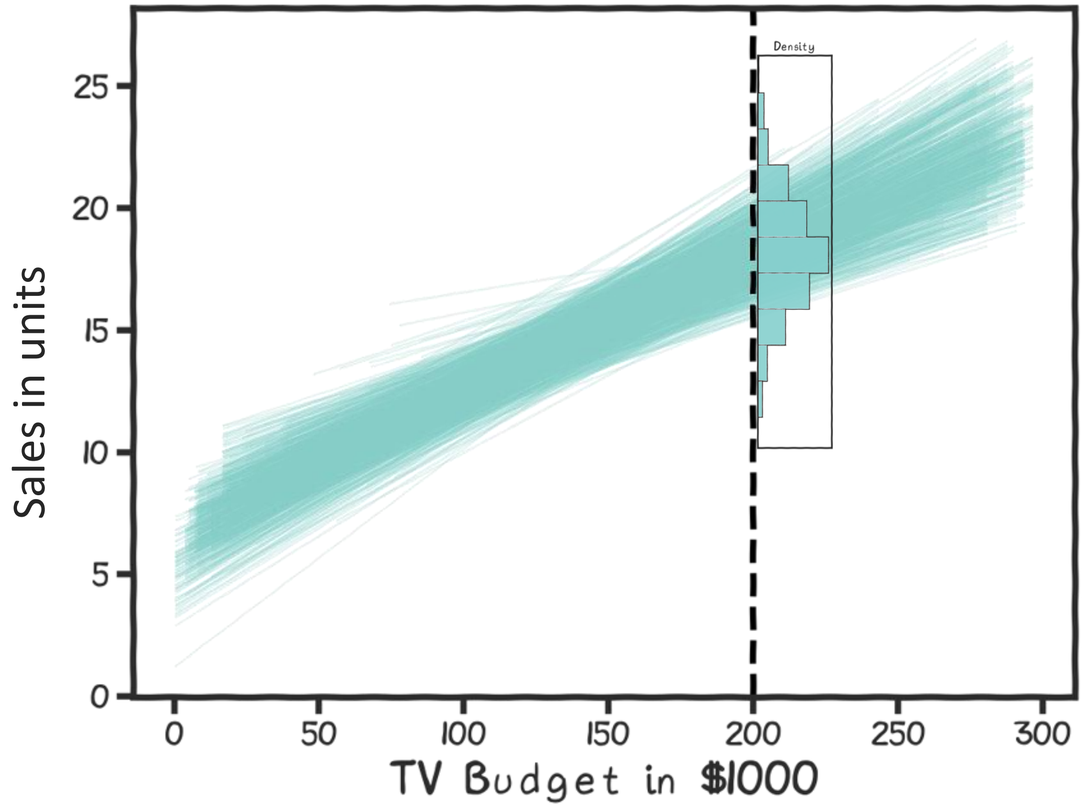
For a given x, we examine the distribution of f and determine the mean and standard deviation (or extract the desired quantiles).



How well do we know $\hat{f}$?

Below we show all regression lines for a thousand of such bootstrapped samples.

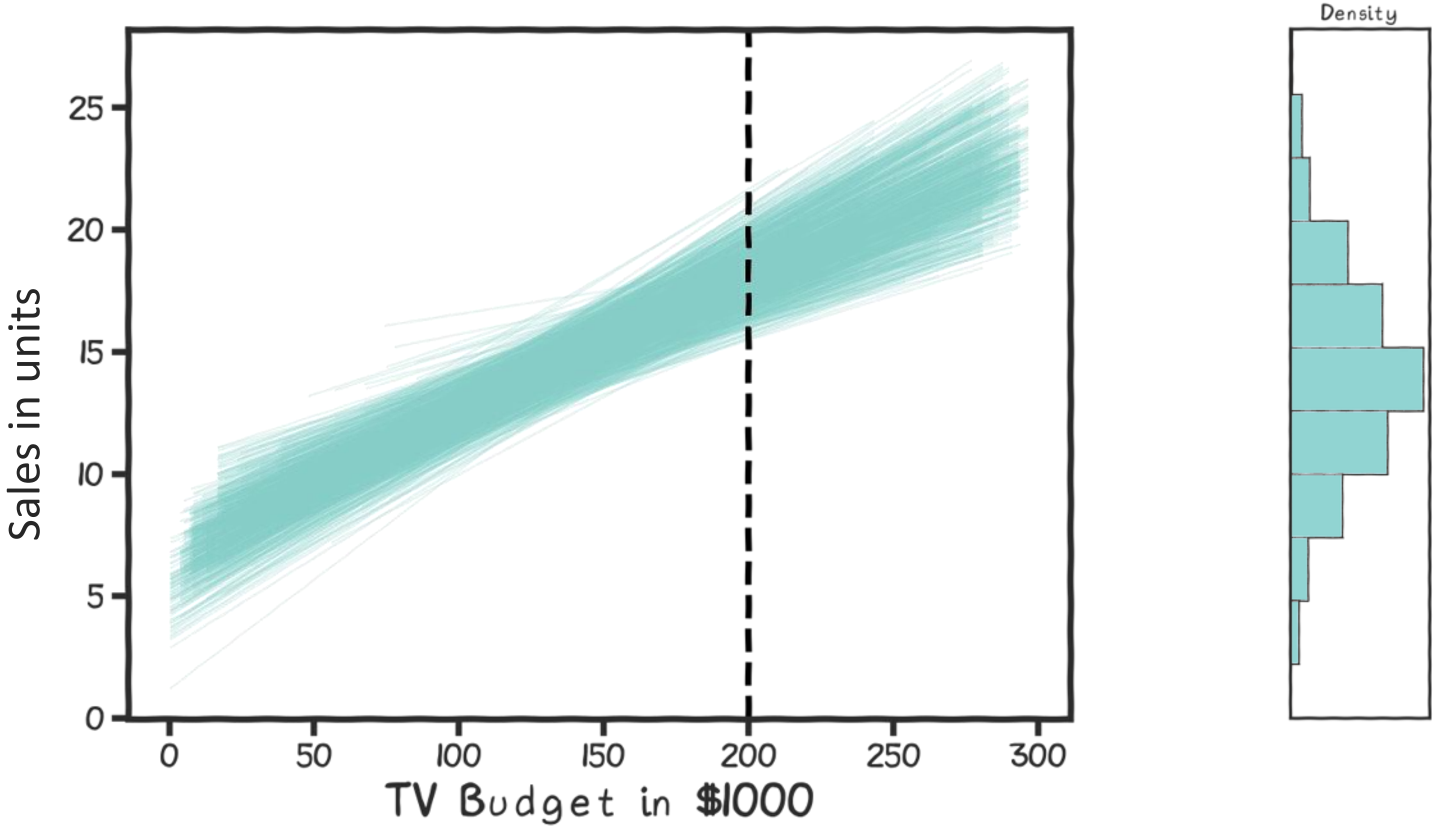
For a given x, we examine the distribution of f and determine the mean and standard deviation (or extract the desired quantiles).



How well do we know $\hat{f}$?

Below we show all regression lines for a thousand of such bootstrapped samples.

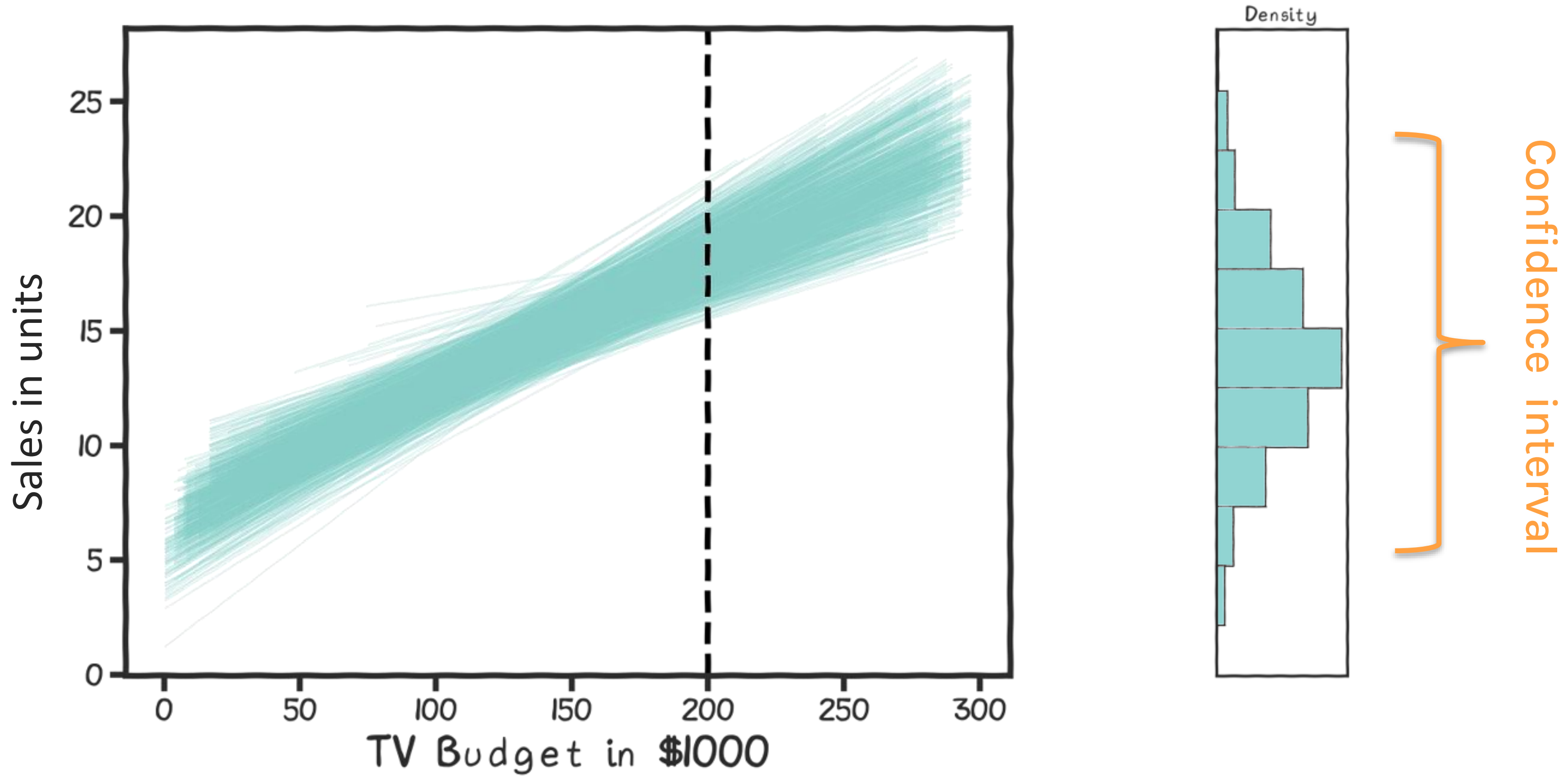
For a given x, we examine the distribution of f and determine the mean and standard deviation (or extract the desired quantiles).



How well do we know $\hat{f}$?

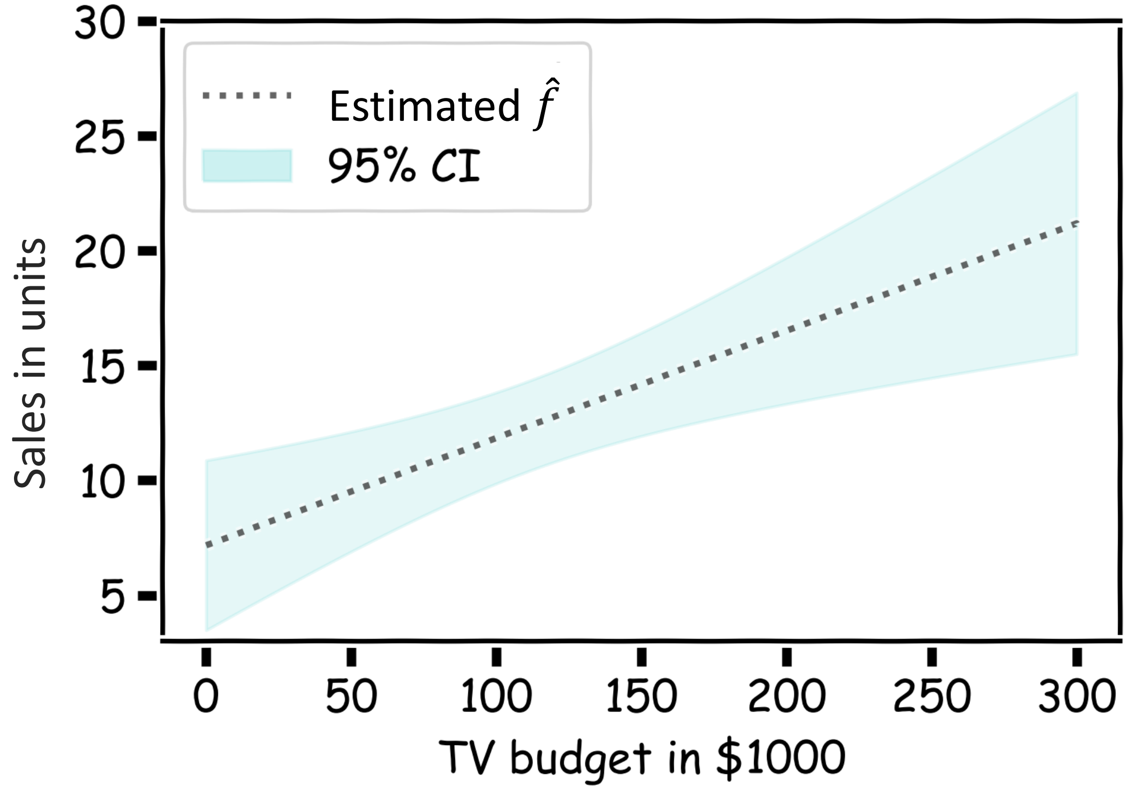
Below we show all regression lines for a thousand of such bootstrapped samples.

For a given x, we examine the distribution of f and determine the mean and standard deviation (or extract the desired quantiles).



How well do we know $\hat{f}$?

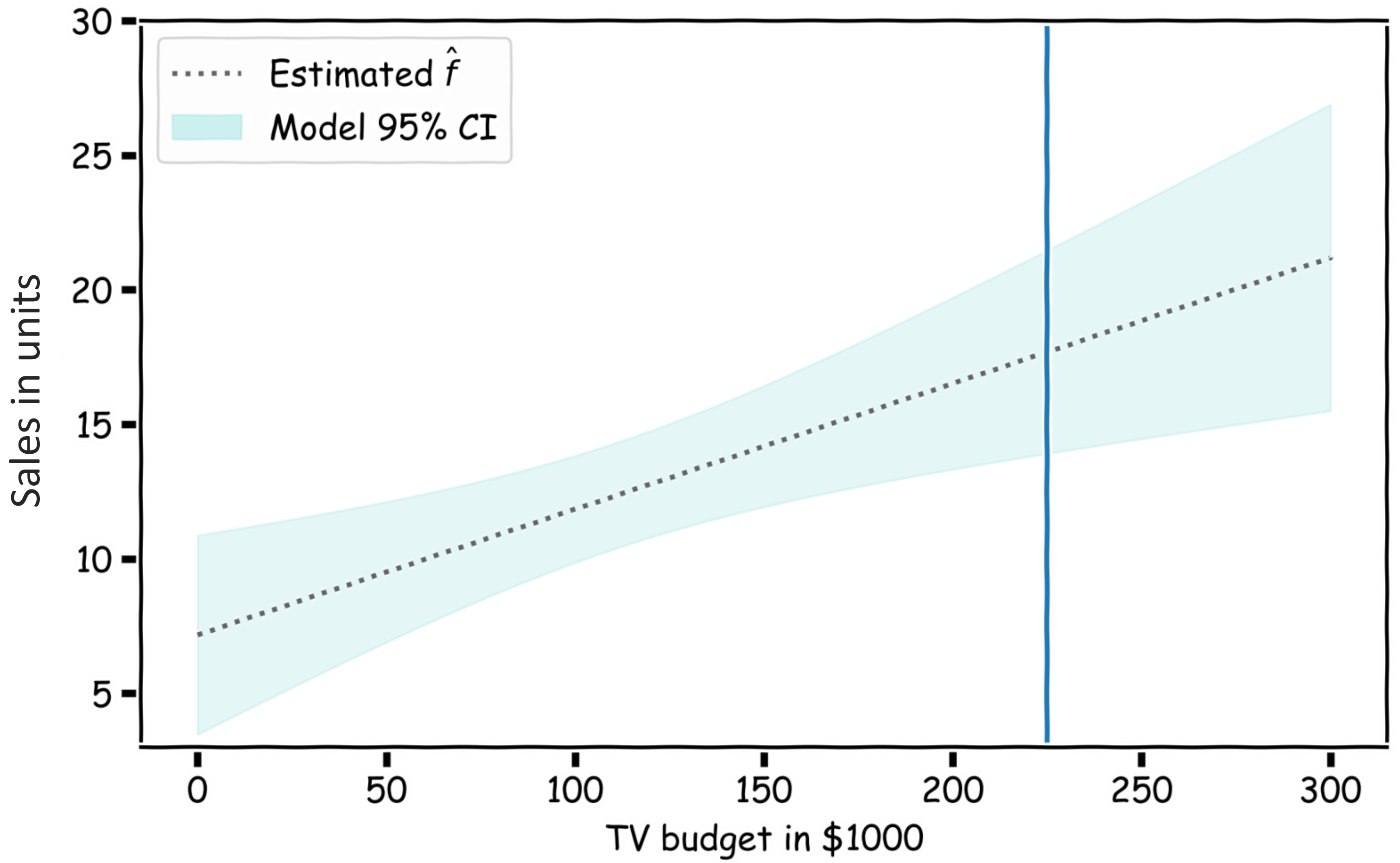
For every x, we calculate the mean of the models, $\widehat{\mu_f}$ (shown with dotted line) and the 95% CI of those models (shaded area).



Confidence in predicting $\hat{y}$

Even if we knew $f(x)$, the response value cannot be predicted perfectly because of the random error in the model (irreducible error).

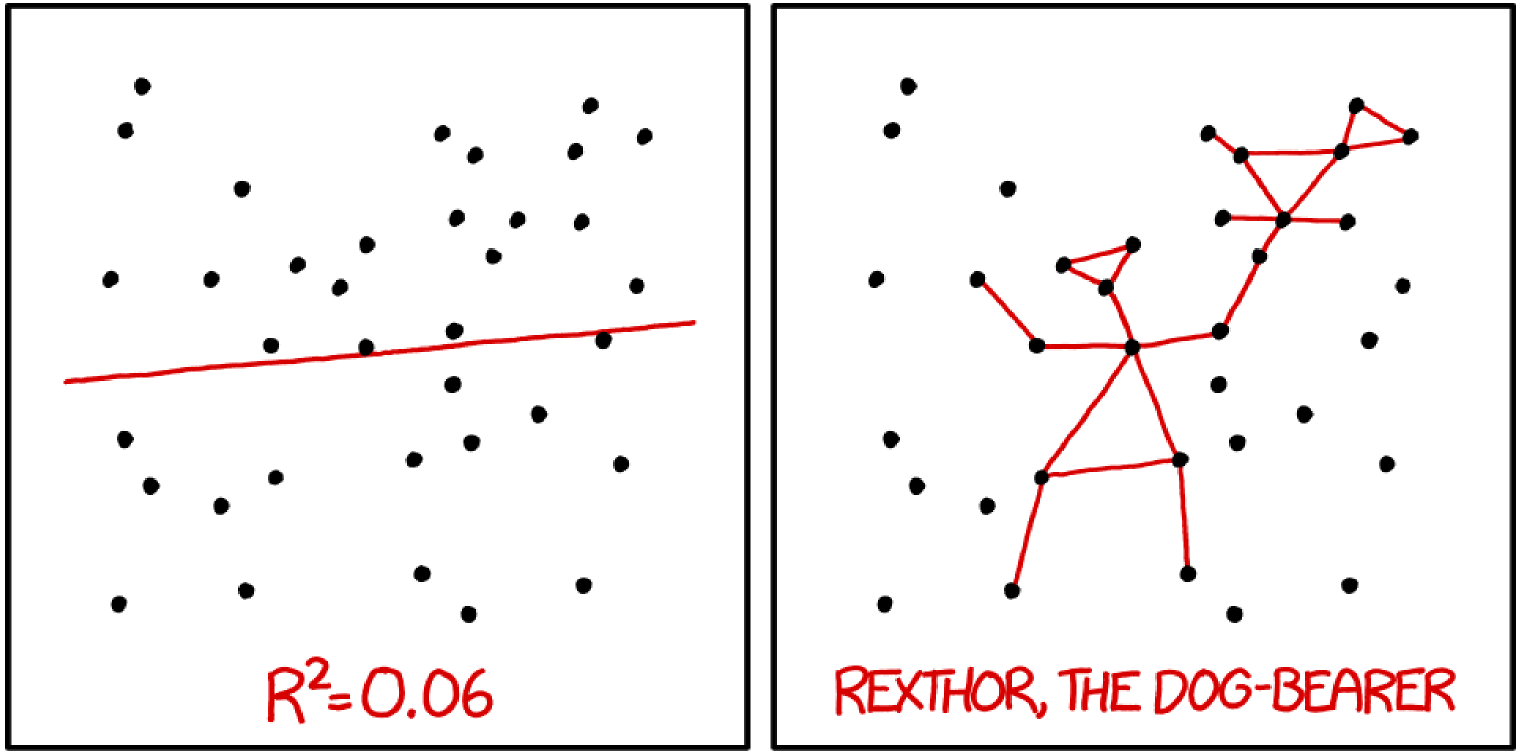
How much will Y vary from $\hat{Y}$? We use prediction intervals to answer this question.



Confidence in predicting $\hat{y}$

Even if we knew $f(x)$, the response value cannot be predicted perfectly because of the random error in the model (irreducible error).

How much will Y vary from $\hat{Y}$? We use prediction intervals to answer this question.

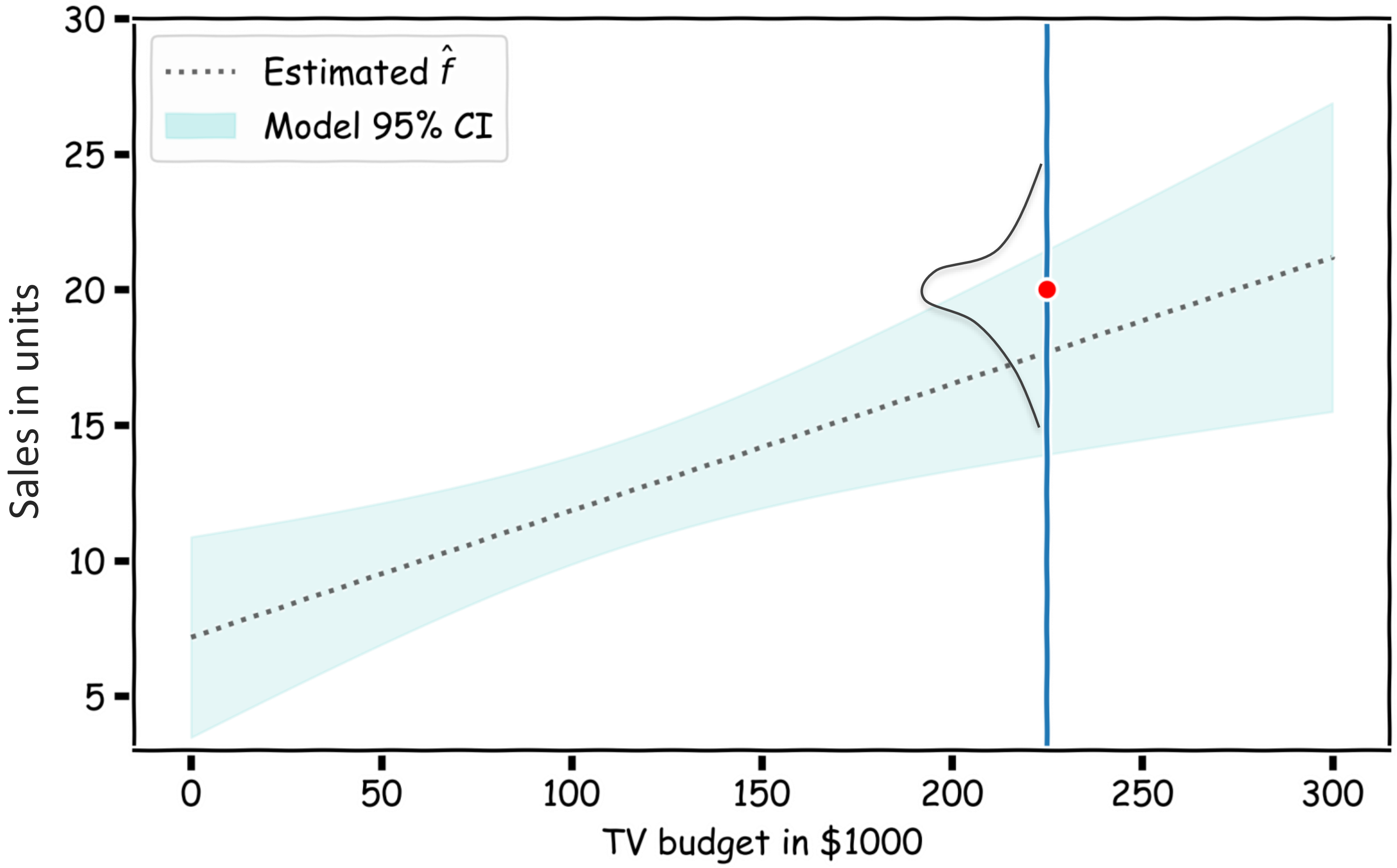


I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

Confidence in predicting $\hat{y}$

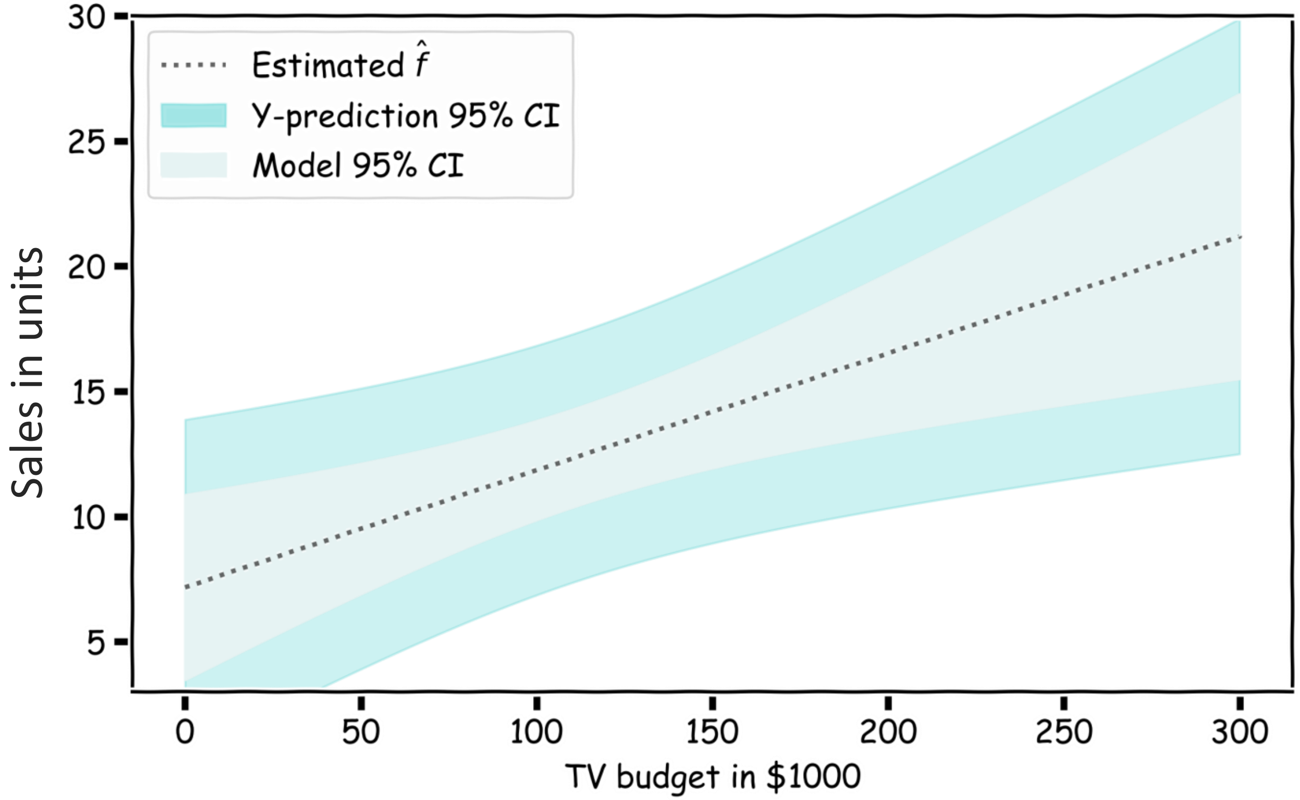
$f(x)$ is a random variable

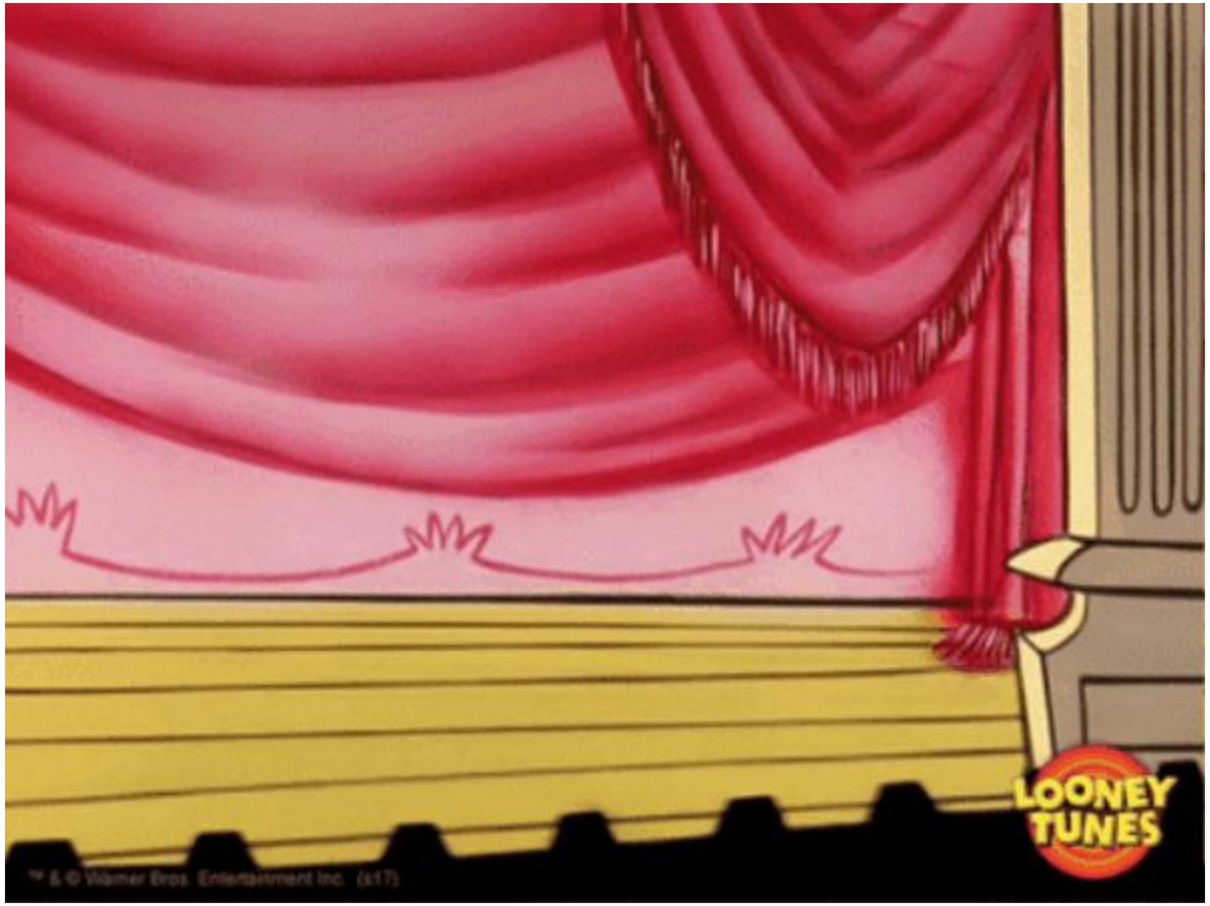
- For a given x, we have a distribution of models $f(x)$
- For each of these $f(x)$, the prediction for $y \sim N(f(x), \sigma_\epsilon)$



Confidence in predicting $\hat{y}$

- For a given x, we have a distribution of models $f(x)$
- For each of these $f(x)$, the prediction for $y \sim N(f(x), \sigma_\epsilon)$
- The prediction confidence intervals are then …





Thank you