CONFIDENCE & PREDICTION INTERVALS

LECTURE 3

SECTION 1

JUNE 5TH

INSTITUTE FOR APPLIED COMPUTATIONAL SCIENCE AT HARVARD UNIVERSITY

UNIVERSITY of RWANDA

UNCERTAINTY AND CONFIDENCE

HOW RELIABLE ARE MODEL INTERPRETATIONS?

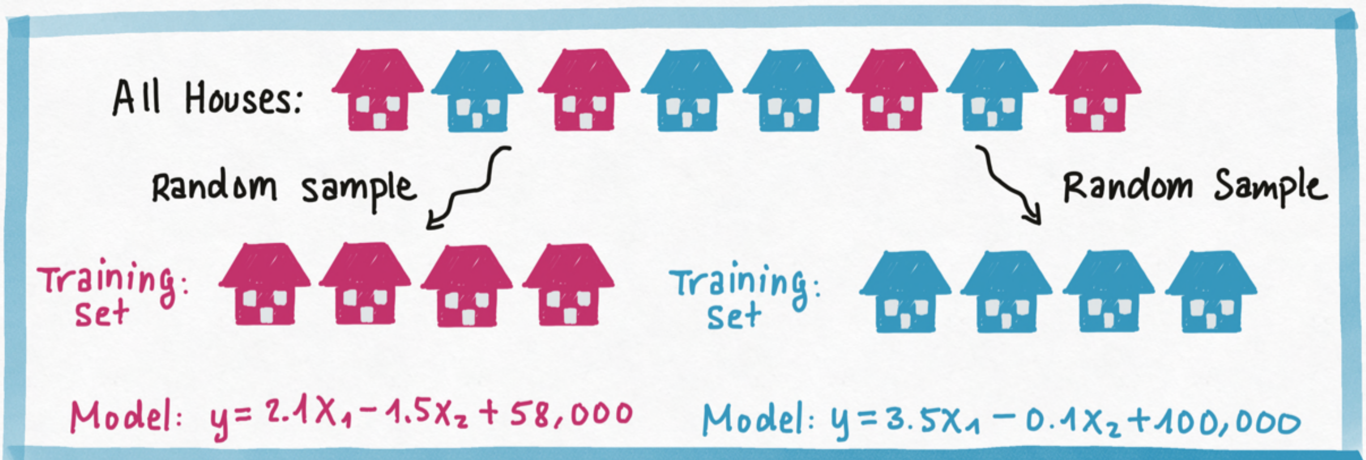
Suppose our model for house prices is:

$$y = 3.5x_1 - 0.1x_2 + 100{,}000$$

where y is price (\$), x_1 is size (m^2) and x_2 is distance to city center (m).

Interpretation: size (x_1) has bigger effect on price than x_2.

But how certain are we in our estimation of the coefficients 3.5, 0.1 and 100,000?

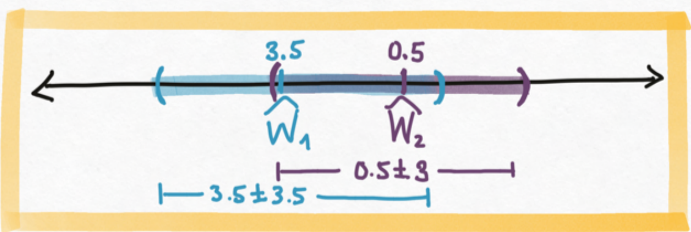


SAMPLING DISTRIBUTIONS & STANDARD ERRORS:

Our estimate $\hat{w}$ of the coefficients w of the regression model changes depending on the training data D sampled from the total population data, $D \sim P(D)$.

So $\hat{w}$ is a random variable. The distribution $P(\hat{w})$ is the sampling distribution. The standard deviation $\sigma_{\hat{w}}$ of $P(\hat{w})$ is called the standard error.

Example: In the housing model $y = 3.5x_1 - 0.5x_2 + 100{,}000$, $\hat{w}_1 = 3.5$, $\hat{w}_2 = -0.5$.
If we know that $\sigma_{\hat{w}_1} = 3.5$ and $\sigma_{\hat{w}_2} = 3.0$. Would we still be confident that x_1 is more important than x_2?



The estimated values of $\hat{w}_1$ & $\hat{w}_2$ fall within one standard error of each other.

Their difference could be simply due to random variations in the training set.

CONFIDENCE INTERVALS AND UNCERTAINTY:

The standard error gives us a sense of our uncertainty over our estimates. Typically, we express this uncertainty as a 95% confidence interval:

$$CI_{\hat{w}_i} = (\hat{w}_i - 2\sigma_{\hat{w}_i},\ \hat{w}_i + 2\sigma_{\hat{w}_i})$$

The 95% CI means that we are 95% sure that the true value of w_i lies in this interval.

<u>Example</u>: If $\hat{w}_1 = 3.5$, $\hat{w}_2 = -0.1$ and $\sigma_{\hat{w}_1} = 0.5$, $\sigma_{\hat{w}_2} = 3.5$, then the 95% CI's for $\hat{w}_1$ and $\hat{w}_2$ are:

$$CI_{\hat{w}_1} = (2.5,\ 4.5)\ ,\quad CI_{\hat{w}_2} = (-7.1,\ 6.9)$$

We see that the width of $CI_{\hat{w}_1}$ is 2 and that of $CI_{\hat{w}_2}$ is 14. So we are much more uncertain about our estimate of $\hat{w}_2$.

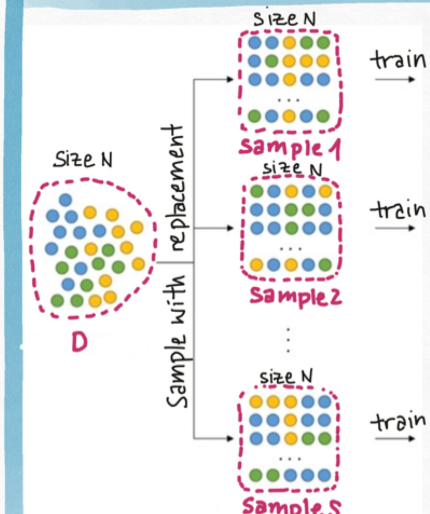
BOOTSTRAP UNCERTAINTY

HOW TO COMPUTE STANDARD ERROR & CI?

The standard error $\sigma_{\hat{w}_i}$ of an estimate $\hat{w}_i$ is the standard deviation of the sampling distribution $p(\hat{w}_i)$ of $\hat{w}_i$. The distribution $p(\hat{w}_i)$ is computed over all possible samples of training data from the population.

But we only get one training dataset D!

We simulate getting different training sets by randomly sampling D. This is called Bootstrapping.



Model 1: $\hat{y} = \hat{w}_1^{(1)} x_1 + \hat{w}_2^{(2)} x_2 + \hat{w}_0$

Model 2: $\hat{y} = \hat{w}_1^{(2)} x_1 + \hat{w}_2^{(2)} x_2 + \hat{w}_0$

$\vdots$

Model S: $\hat{y} = \hat{w}_1^{(S)} x_1 + \hat{w}_2^{(S)} x_2 + \hat{w}_0$

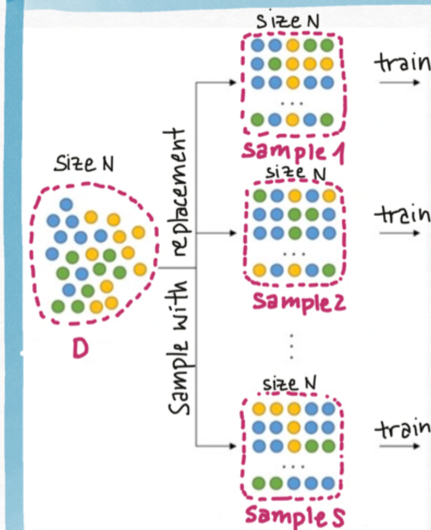
$$\sigma_{\hat{w}_i} = \sqrt{\frac{\sum_s \left(\hat{w}_i^{(s)} - \overline{\hat{w}_i}\right)^2}{S}}$$

PREDICTION INTERVALS:

Confidence intervals describe our uncertainty on the estimates $\hat{w}_i$.
we can also quantify our uncertainty on our predictions $\hat{y}$.

We compute the 95% prediction interval for an input x by bootstrapping S training sets from D, estimating $\hat{w}^{(s)}$ on each bootstrap sample, and making a prediction $\hat{y}^{(s)}$ for x.

The 95% interval of the predictions $\{\hat{y}^{(s)}\}$ is the 95% prediction interval.



Model 1: $\hat{y}^{(1)} = \hat{w}_1^{(1)} x_1 + \hat{w}_2^{(2)} x_2 + \hat{w}_0$

Model 2: $\hat{y}^{(2)} = \hat{w}_1^{(2)} x_1 + \hat{w}_2^{(2)} x_2 + \hat{w}_0$

$\vdots$

Model S: $\hat{y}^{(S)} = \hat{w}_1^{(S)} x_1 + \hat{w}_2^{(S)} x_2 + \hat{w}_0$

$I_{2.5}$ = 2.5 percentile of $\{\hat{y}^{(S)}\}$

$I_{97.5}$ = 97.5 percentile of $\{\hat{y}^{(S)}\}$

95% PI $= (I_{2.5}, I_{97.5})$

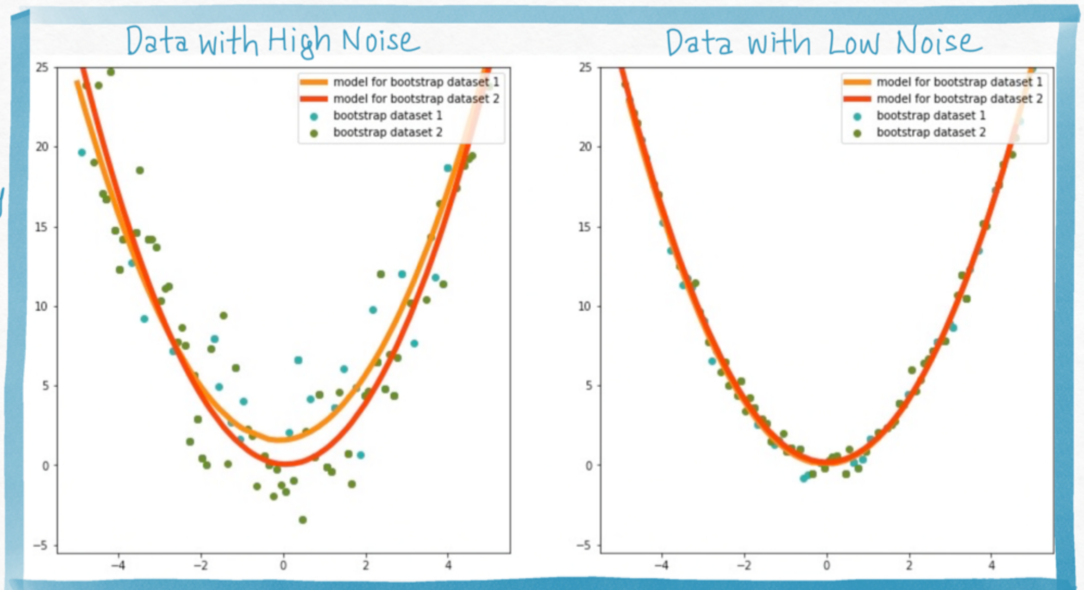
WHERE DOES UNCERTAINTY COME FROM

WHAT CONTRIBUTES TO UNCERTAINTY?

The confidence interval indicates how much our estimates of model parameter changes when we resample our training set.

When the data has a lot of noise, each bootstrap training set is very different, so each bootstrap model is very different. Thus, our parameter estimates will have high variation, hence the CI will be wide.

- Bootstrap datasets vary more.
- Bootstrap models vary more.



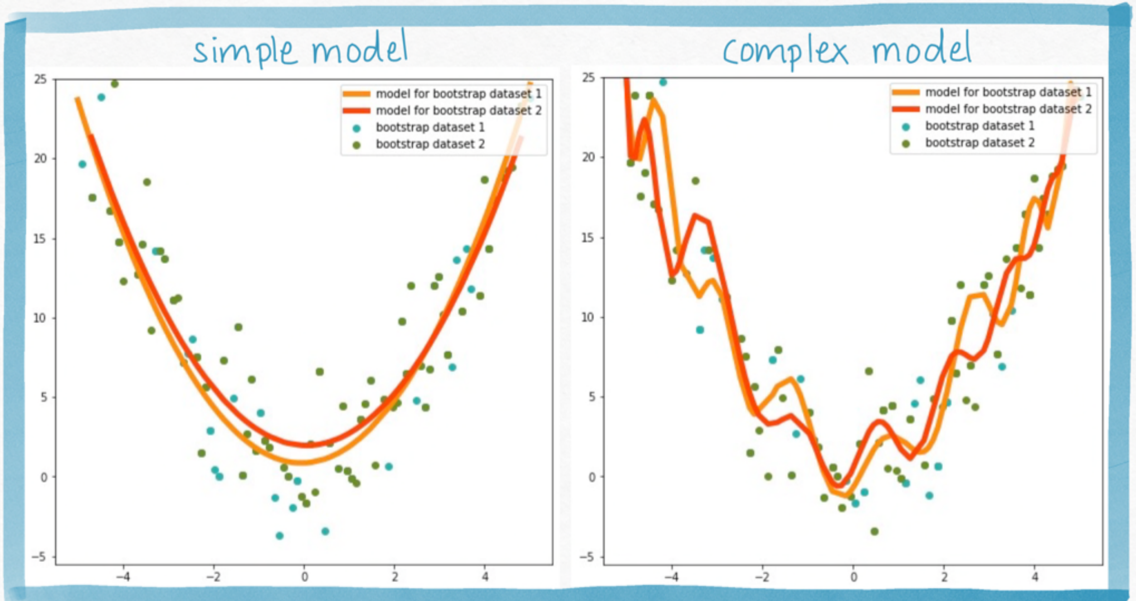
- Bootstrap datasets vary less.
- Boostrap models vary less.

WHAT CONTRIBUTES TO UNCERTAINTY?

The confidence interval indicates how much our estimates of model parameter changes when we resample our training set.

When the model is very complex and flexible, it can overfit to the noise in each bootstrap dataset. Then each bootstrap model will be very different.

- simple models do not fit to the noise in the data
- bootstrap models are similar



- Complex models can overfit to the noise in the data.
- bootstrap model are very different