Multiple Linear Regression - Prediction

Research Objective

Research Question: How well can we use the explanatory variables to predict height?

Population: All BYU students.

Parameters of Interest: The regression line parameters (all slopes and spread (σ))

Sample: A convenience sample of 1575 BYU students who are in Stat 121.

Research Objective

Research Question: What is the average height for male students who have a 64 inch tall mom, 68 inch tall dad, did not play sports in HS and wears a size 9 shoe?

Fitted Model:

$$\hat{y} = 23.94 + 0.23 \times \text{MotherHeight}_i + 0.24 \times \text{FatherHeight}_i + 0.12 \times \text{Sports}_i + 3.02 \times \text{Sex}_i + 1.12 \times \text{ShoeSize}_i$$

How would you use the model to figure this out?

Research Objective

Research Question: What is the average height for male students who have a 64 inch tall mom, 68 inch tall dad, did not play sports in HS and wears a size 9 shoe?

Fitted Model:

$$\hat{y} = 23.94 + 0.23 \times \text{MotherHeight}_i + 0.24 \times \text{FatherHeight}_i + 0.12 \times \text{Sports}_i + 3.02 \times \text{Sex}_i + 1.12 \times \text{ShoeSize}_i$$

How would you use the model to figure this out?

$$\begin{aligned}\hat{y} &= 23.94 + 0.23 \times 64 + 0.24 \times 68 + 0.12 \times 0 + 3.02 \times 1 + 1.12 \times 9 \\ &= 68.38\end{aligned}$$

Prediction in Regression

Thought Question: Is our prediction of $\hat{y} = 68.38$ of the average height for male students who have a 64 inch tall mom, 68 inch tall dad, did not play sports in HS and wears a size 9 shoe a sample estimate or population parameter?

- Sample estimate!
- We would rather build an interval for the population parameter.

Confidence Intervals for Averages

Using similar principles as we have used in the past to build confidence intervals:

$$\hat{y} \pm t^{\star} \mathrm{SE}(\hat{\beta}_0 + \hat{\beta}_1 \mathrm{MH} + \cdots + \hat{\beta}_5 \mathrm{Shoe})$$

Is a confidence interval for the average value of y given an x (the population average height for male students who has a 64 inch tall mom, 68 inch tall dad, did not play sports in HS and wears a size 9) where the value of $t^{\star}$ is determined by the confidence level.

For our analysis, this comes out to be (68.113, 68.646) for a 95% interval.

Notes:

1. Don't worry about the formula (computer will calculate this for you).
2. Interpetation: We are 95% confident that the average height for all male students who has a 64 inch tall mom, 68 inch tall dad, did not play sports in HS and wears a size 9 is between 68.113 and 68.646.

Prediction Intervals for Individuals

Research Question: Eddie is a male student who has a 64 inch tall mom, 68 inch tall dad, did not play sports in HS and wears a size 9. What will his height be?

Using similar principles as we have used in the past to build confidence intervals:

$$\hat{y} \pm t^{\star}\text{SE}(\hat{\beta}_0 + \hat{\beta}_1\text{MH} + \cdots + \hat{\beta}_5\text{Shoe} + \hat{\epsilon})$$

is a **prediction** interval for the value of y given an x (for example, Eddie's height) where the value of $t^{\star}$ is determined by the confidence level.

For our analysis, this comes out to be (64.822, 71.937) for a 95% interval.

Notes:

1. Don't worry about the formula (computer will calculate this for you).
2. Interpetation is similar: We are 95% confident that Eddie's height will be between 64.822 and 71.937.

Prediction vs Confidence Intervals

Confidence interval for prediction: An interval estimate for the average of y given the $x's$.

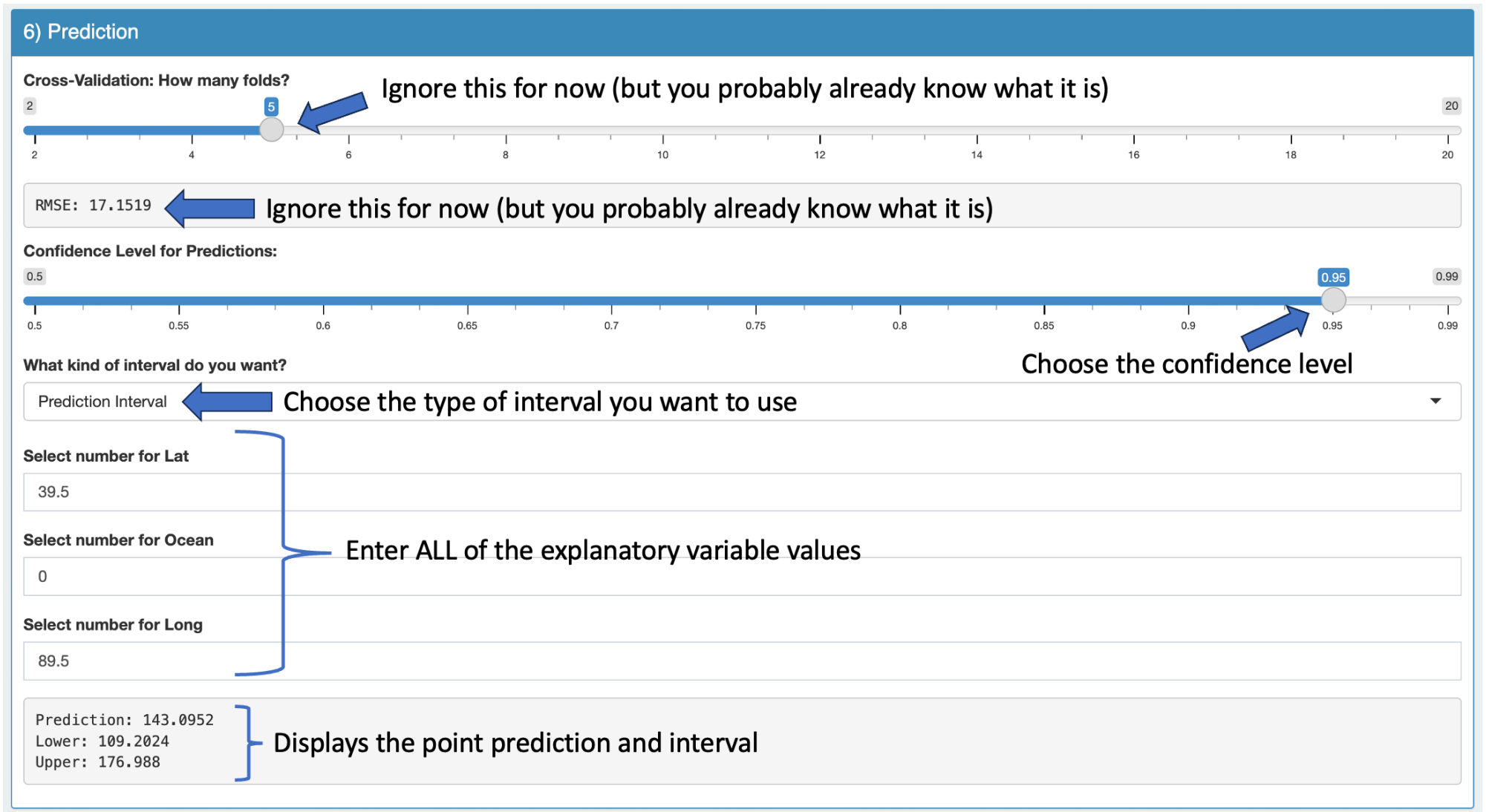
Prediction interval for prediction: An interval estimate for the value of a single y given the $x's$.

- Prediction intervals are ALWAYS wider than confidence intervals. Why?

- There is more variability from student to student than the average of all students

Using the Analysis Tool

All of the steps are the same as in previous lectures…



Nuances of Predictions

Research Question: Jordan is a male student who has a 77 inch tall mom, 85 inch tall dad, did play sports in HS and wears a size 12 shoe. What will his height be?

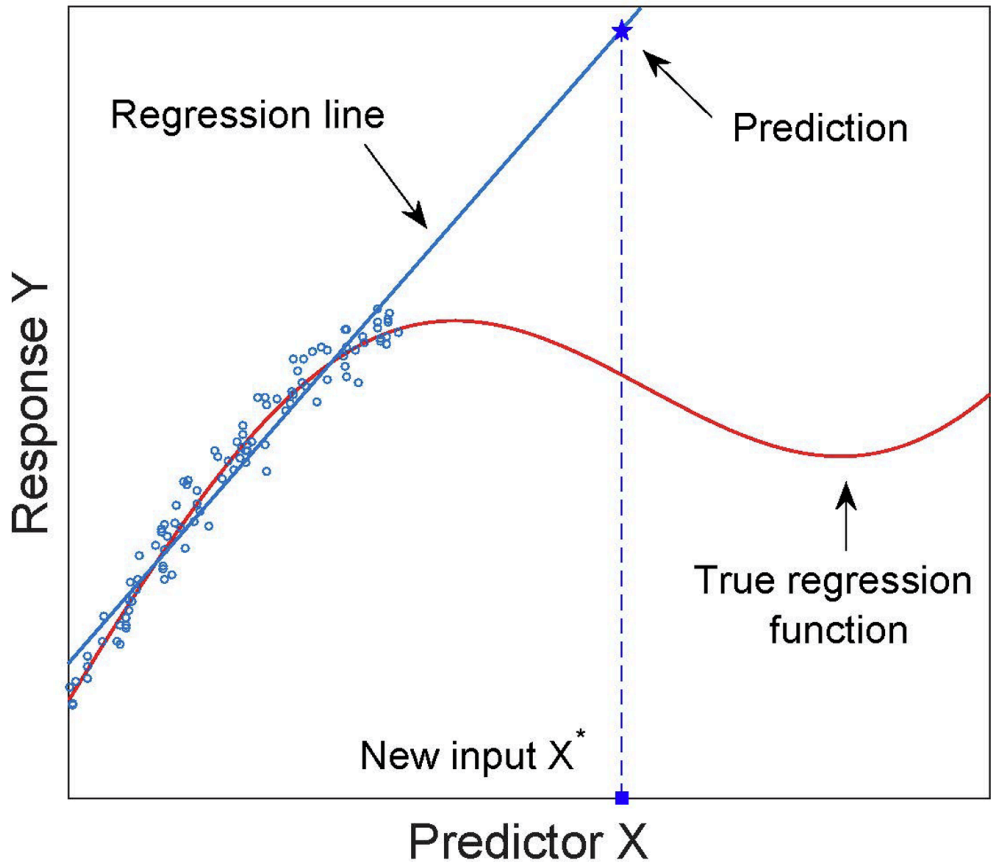
Answer:

- Don’t do the prediction because it’s outside of the data range! This is referred to as **extrapolation**.

Nuances of Predictions

1. Extrapolation - trying to predict outside of the range of the data.

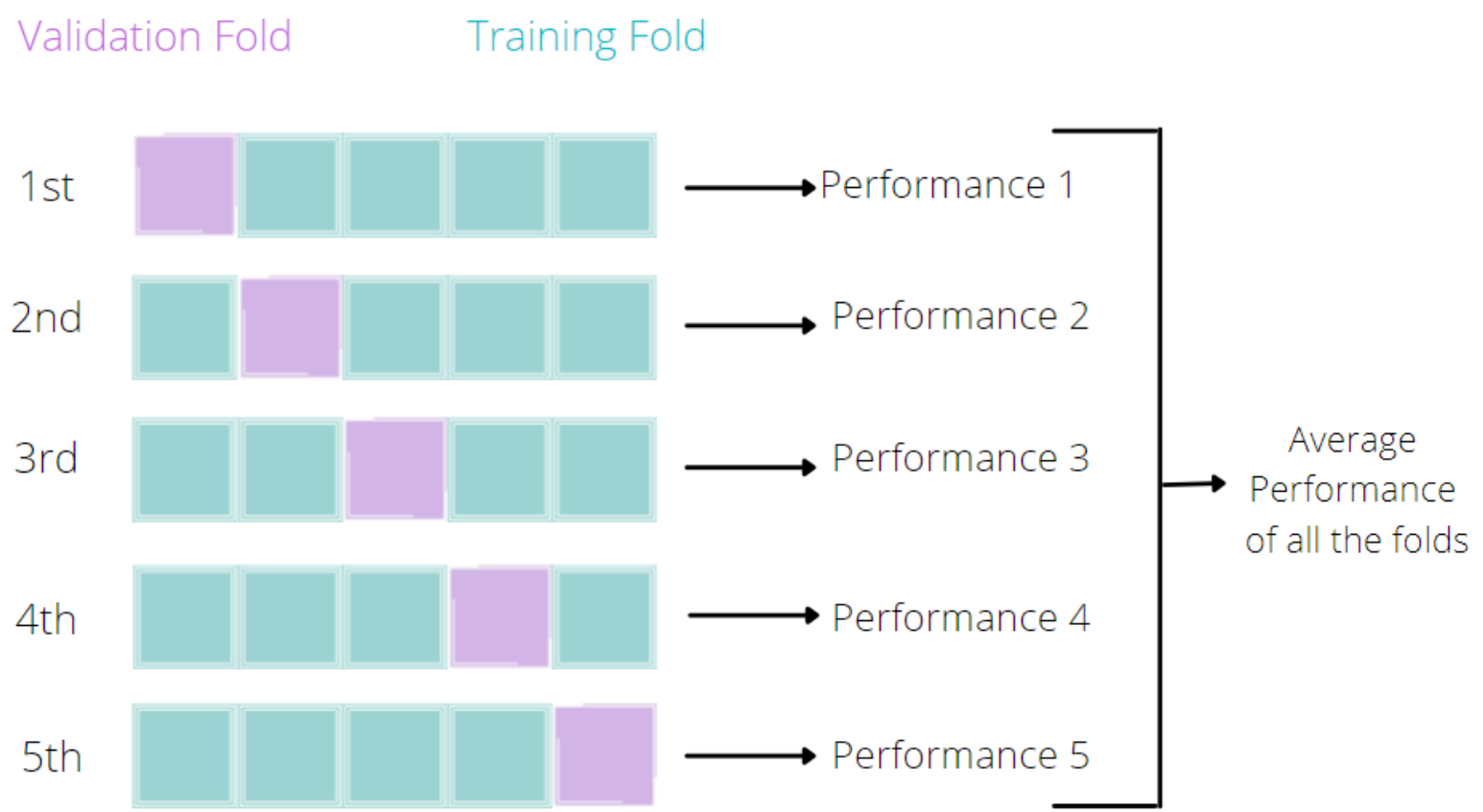
- In multiple linear regression, we have several ways to extrapolate. If ANY of the explanatory values are outside the range of the data, we shouldn’t do the prediction.


Regression line
Prediction
Response Y
True regression function
New input X*
Predictor X

Nuances of Predictions

How do we know if our predictions are any good?

- Use K-fold Cross Validation to see how well you are predicting.



Nuances of Predictions

Notes:

1. Randomly split the data into validation folds. Each “fold” gets a turn to be predicted.
2. Lots of performance metrics but most common is **root mean square error**

$$\text{RMSE} = \sqrt{\frac{1}{n_{\text{validation}}} \sum_{i=1}^{n_{\text{validation}}} (y_i - \hat{y}_i)^2}$$

   where y_i is an observation in the validation set and $\hat{y}_i$ is the corresponding prediction.
3. The intuitive interpretation of RMSE is the average error in our prediction.

Additional Prediction Practice

Measuring possum head size can be difficult. However, various other factors can be used to predict head size? Use a multiple linear regression model (and the course app) to answer the following questions:

1. Hyrum found a huge (96 cm total, male, 7 years, 68cm skull, 42 length tail) possum, What is your predicted head length for this possum?
    - 101.7571379 with a 95% *prediction* interval is (97.278, 106.237).
2. Hyrum found a huge (96 cm total, male, 7 years, 68cm skull, 42 length tail) possum. What is the average head length for possums of this size?
    - 101.7571379 with 95% *confidence* interval is (100.021, 103.493).
3. Hyrum found a baby (70 cm total, male, 0.5 years, 42cm skull, 28 length tail) possum. What is your predicted head length for this possum?
    - EXTRAPOLATION
4. Is your model good or bad at possum head sizes?
    - The RMSE of a 10 fold CV is 1.6442366.

Key Terminology

- Confidence Intervals for Averages
- Extrapolation
- Prediction Intervals for Individuals
- Cross validation
- Root mean square error (RMSE)