

# 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 ( $\sigma$ ))

**Sample:** A convenience sample of 1727 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:

$$\hat{y} = 23.26 + 0.28 \times \text{MotherHeight}_i + 0.21 \times \text{FatherHeight}_i + 0.35 \times \text{Sports}_i + 3.19 \times \text{Sex}_i + 1.06 \times \text{ShoeSize}_i$$

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:

$$\hat{y} = 23.26 + 0.28 \times \text{MotherHeight}_i + 0.21 \times \text{FatherHeight}_i + 0.35 \times \text{Sports}_i + 3.19 \times \text{Sex}_i + 1.06 \times \text{ShoeSize}_i$$

How would you use the model to figure this out?

$$\begin{aligned}\hat{y} &= 23.26 + 0.28 \times 64 + 0.21 \times 68 + 0.35 \times 0 + 3.19 \times 1 + 1.06 \times 9 \\ &= 68.419\end{aligned}$$

# Prediction in Regression

**Thought Question:** Is our prediction of  $\hat{y} = 68.419$  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:

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

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  $t^*$  is determined by the confidence level.

For our analysis, this comes out to be (68.177, 68.665) for a 95% interval.

Notes:

1. Don't worry about the formula (computer will calculate this for you).
2. Interpretation: 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 68.177 and 68.665.

# 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:

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

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

For our analysis, this comes out to be (64.928, 71.914) for a 95% interval.

Notes:

1. Don't worry about the formula (computer will calculate this for you).
2. Interpretation is similar: We are 95% confident that Eddie's height will be between 64.928 and 71.914.

# 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...

**6) Prediction**

**Cross-Validation: How many folds?**  
2 5 20  
2 4 6 8 10 12 14 16 18 20  
Ignore this for now (but you probably already know what it is)

RMSE: 17.1519  
Ignore this for now (but you probably already know what it is)

**Confidence Level for Predictions:**  
0.5 0.95 0.99  
0.5 0.55 0.6 0.65 0.7 0.75 0.8 0.85 0.9 0.95 0.99  
Choose the confidence level

**What kind of interval do you want?**  
Prediction Interval  
Choose the type of interval you want to use

**Select number for Lat**  
39.5

**Select number for Ocean**  
0

**Select number for Long**  
89.5

Enter ALL of the explanatory variable values

Prediction: 143.0952  
Lower: 109.2024  
Upper: 176.988  
Displays the point prediction and interval

# Nuances of Predictions

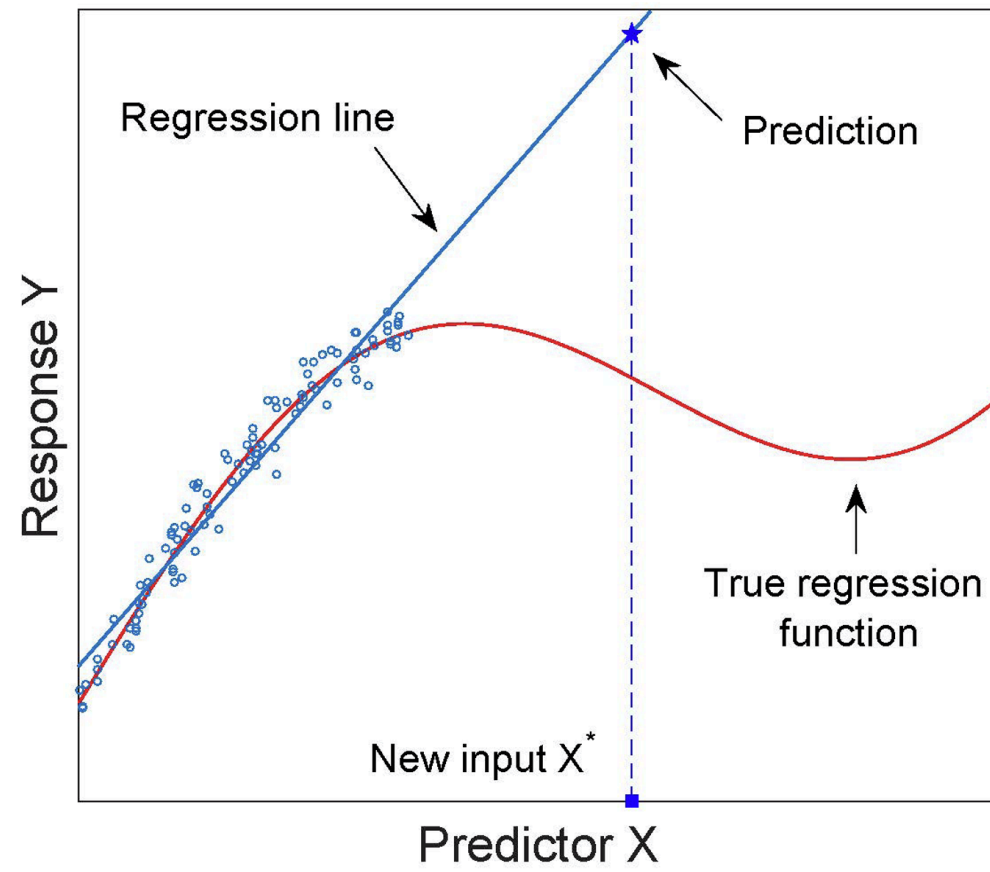
**Research Question:** Jordan is a male student who has a 77 inch tall mom, 85 inch tall dad, did play sports in HS and wears a size 12 shoe. What will his height be?

**Answer:**

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

# Nuances of Predictions

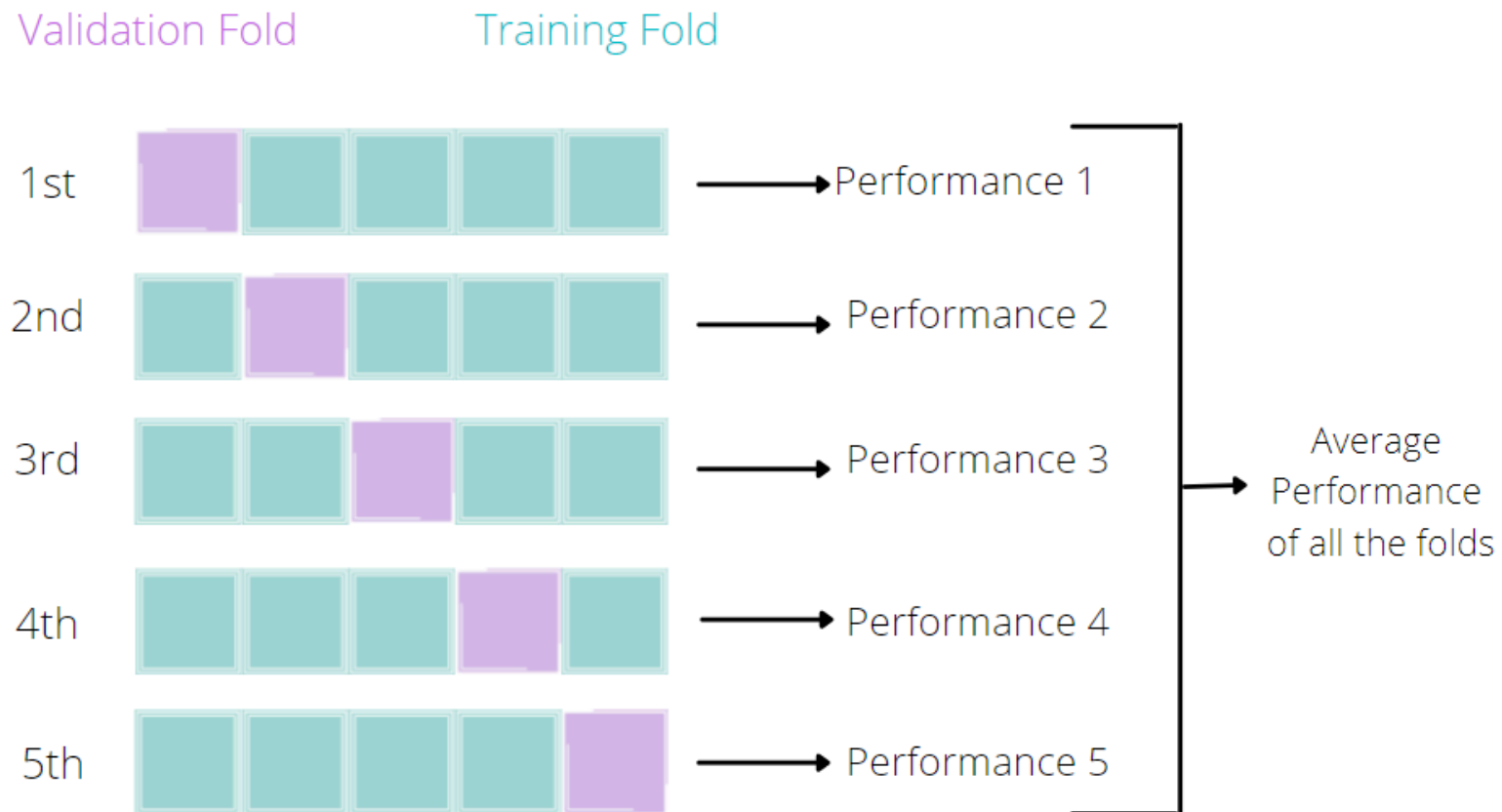
1. Extrapolation - trying to predict outside of the range of the data.
  - In multiple linear regression, we have several ways to extrapolate. If ANY of the explanatory values are outside the range of the data, we shouldn't do the prediction.



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