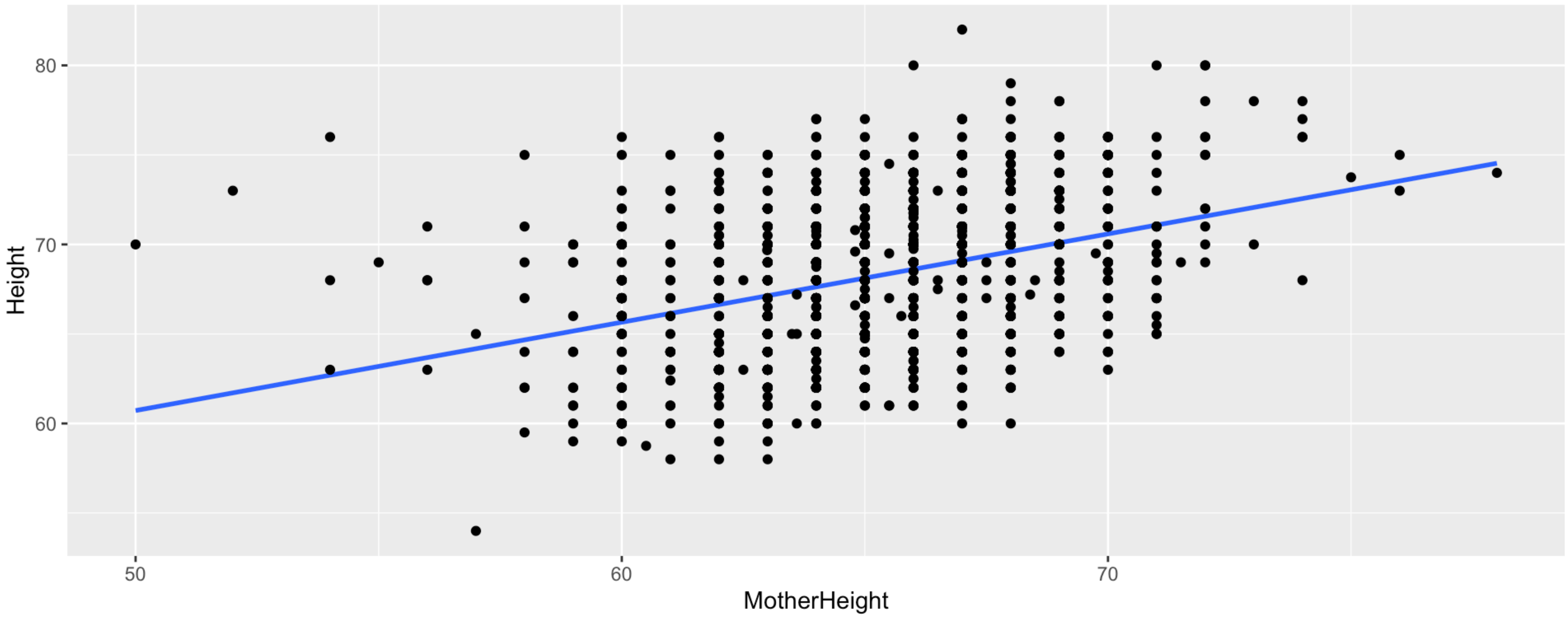
Simple Linear Regression - Prediction

Research Objective

Research Question: What is the average student height for students whose mother is 64 inches tall?

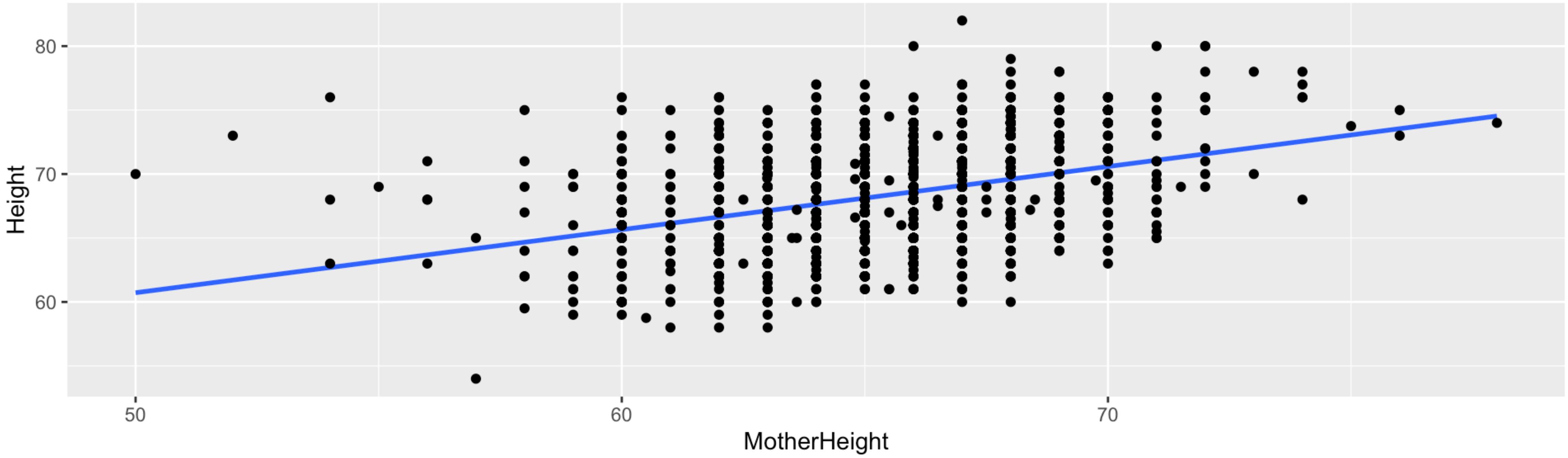


How would you figure this out?

Prediction in Regression

Research Question: What is the average student height for students whose mother is 64 inches tall?

Answer: Use the best fit regression line to tell you the answer.



$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 36.059 + 0.493 \times 64 = 67.611$$

Confidence Intervals for Averages

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

$$\hat{y} \pm t^{\star}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Is a confidence interval for the average value of y given an x (the population average student height for 64 inch tall mothers) where the value of $t^{\star}$ is determined by the confidence level.

For our analysis, this comes out to be (67.417, 67.838) for a 95% interval.

Notes:

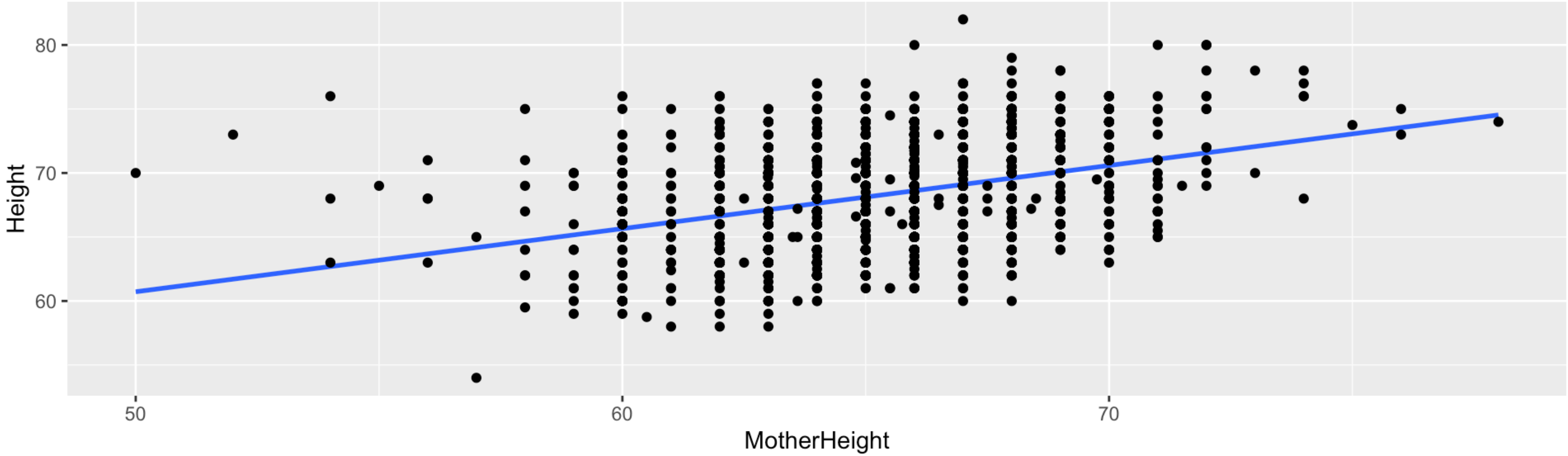
1. Don’t worry about the formula (computer will calculate this for you).
2. Interpetation: We are 95% confident that the average height of all students whose mothers are 64 inches tall is between 67.417 and 67.838.

Prediction in Regression

Research Question: Shaylee's mom is 64 inches tall, what will her height be?

Thought Questions:

1. Is this the same question as above? If not, what is the difference?
   - It's not the same. One is asking about an average while one is asking about a specific person.
   - The "average" is the line while specific people are the "dots".

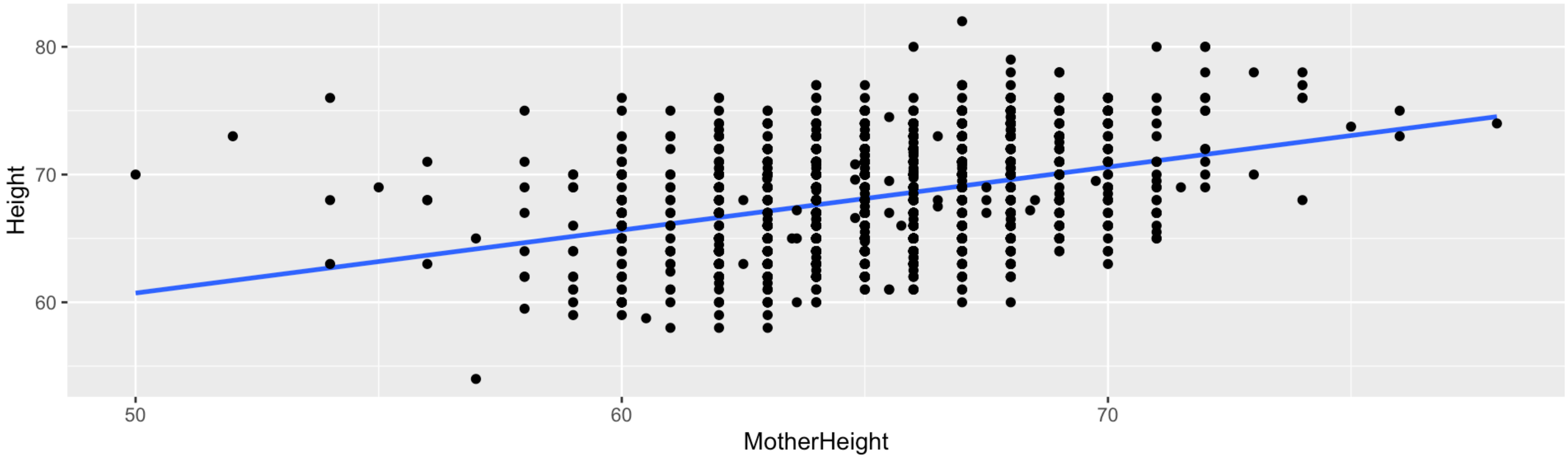


Prediction in Regression

Research Question: Shaylee’s mom is 64 inches tall, what will her height be?

Thought Questions:

2. Should our point prediction (1 number prediction) be the same or different?
    - The point prediction should be the same because “dots” could either fall above or below the line. In this case, we still think Shaylee’s height will be 67.611.

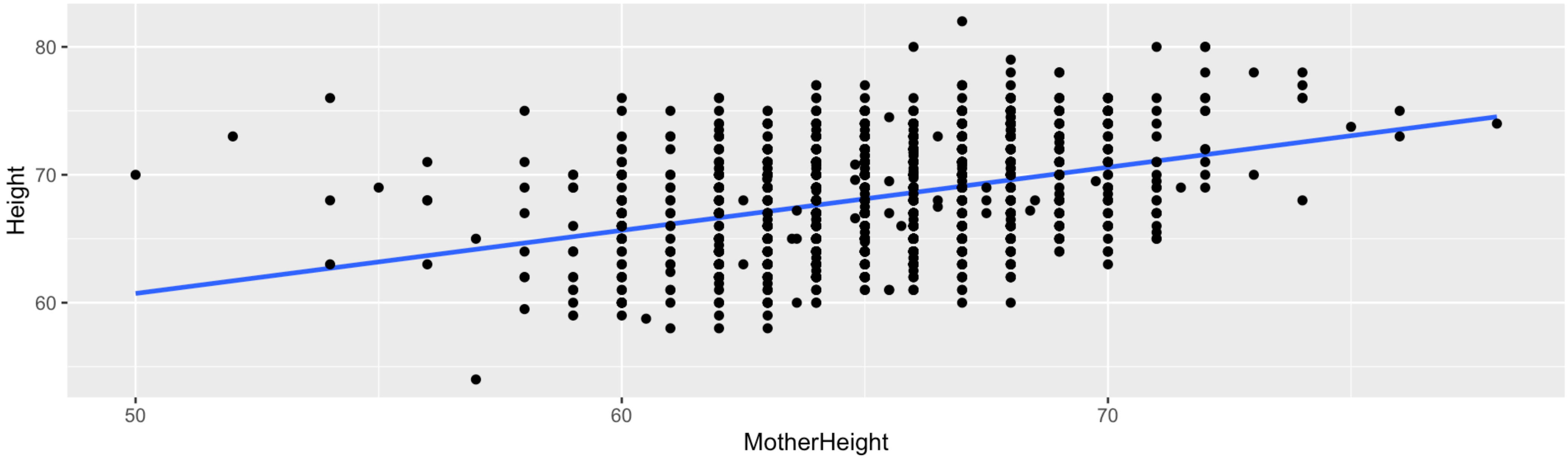


Prediction in Regression

Research Question: Shaylee’s mom is 64 inches tall, what will her height be?

3. Should our interval for the prediction be the same or different? Why or why not?

- It should be wider because heights vary a lot from person to person



Prediction Intervals for Individuals

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

$$\hat{y} \pm t^{\star}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

is a prediction interval for the value of y given an x (for example, Shaylee's height if her mom is 64 inches tall) where the value of $t^{\star}$ is determined by the confidence level.

For our analysis, this comes out to be (59.982, 75.273) for a 95% interval.

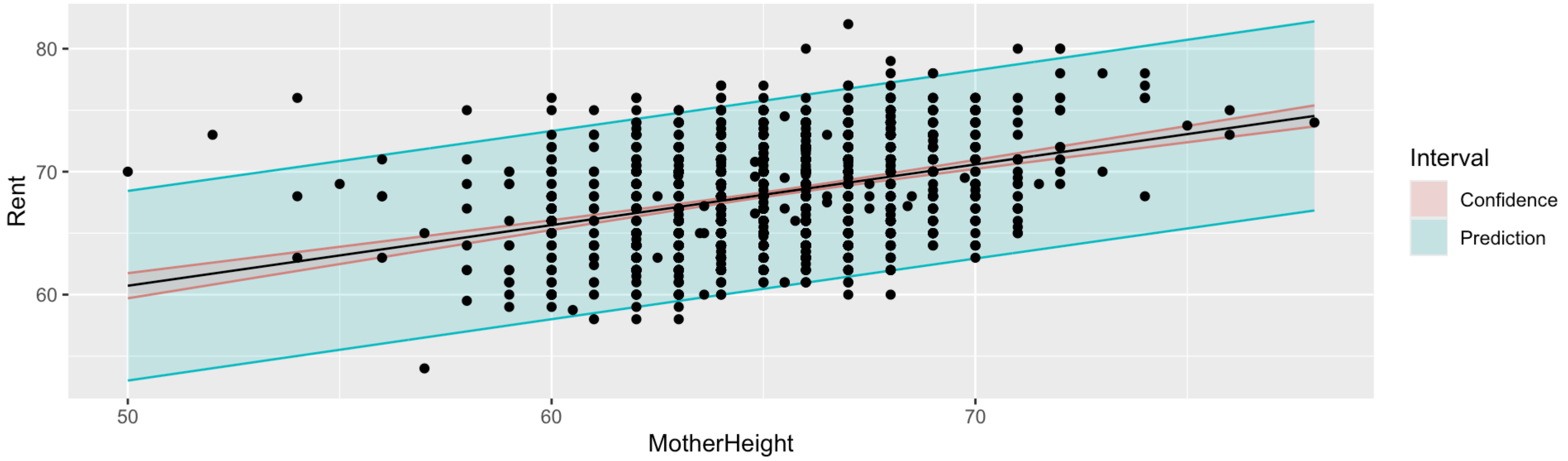
Notes:

1. Don't worry about the formula (computer will calculate this for you).
2. Interpetation is similar: We are 95% confident that Shaylee's height, given her mom is 64 inches tall, should be between 59.982 and 75.273.

Prediction vs Confidence Intervals

Confidence interval for prediction: An interval estimate for the average of y given an x.

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

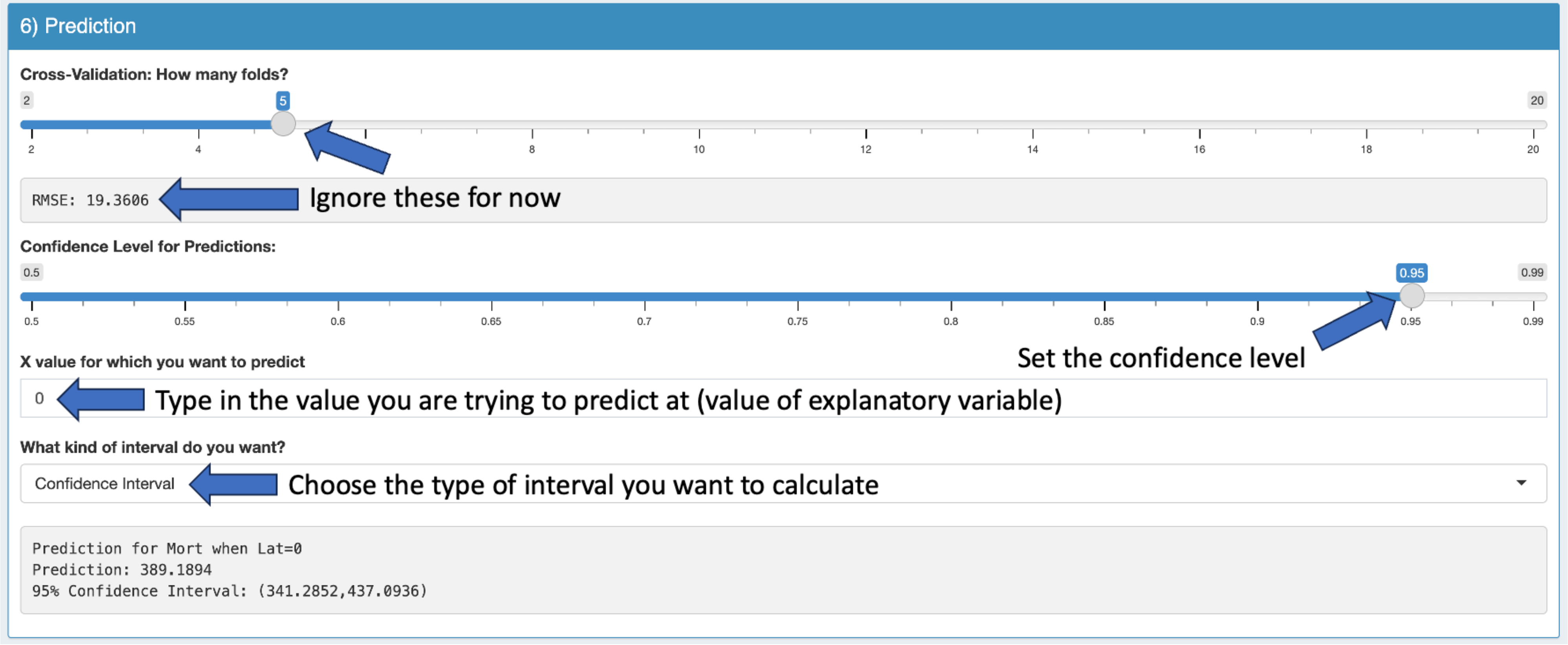


Prediction intervals are ALWAYS wider than confidence intervals. Why?

- There is more variability from student to student than with the average heights for students.

Using the Analysis Tool

All previous steps in the tool are the same as covered in previous lecture notes:



6) Prediction
Cross-Validation: How many folds?
2
5
20
2
4
8
10
12
14
16
18
20
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
X value for which you want to predict
0
What kind of interval do you want?
Confidence Interval

```
RMSE: 19.3606
```

Ignore these for now

Set the confidence level

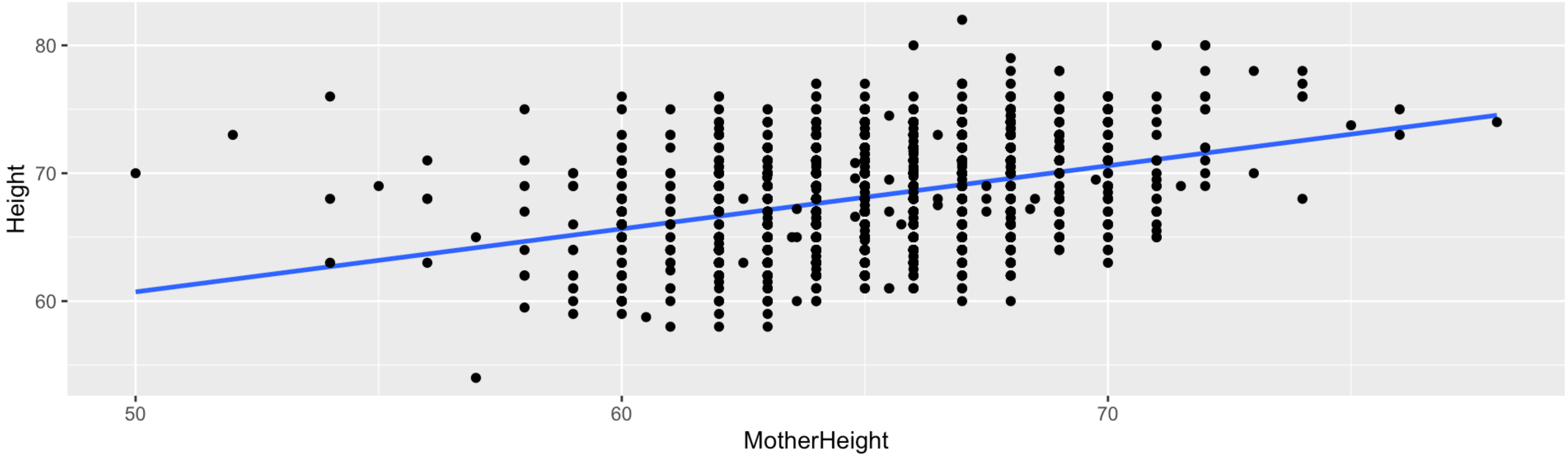
Type in the value you are trying to predict at (value of explanatory variable)

Choose the type of interval you want to calculate

```
Prediction for Mort when Lat=0
Prediction: 389.1894
95% Confidence Interval: (341.2852,437.0936)
```

Nuances of Predictions

Research Question: Lucy’s mom is 81 inches tall, what will her height be?

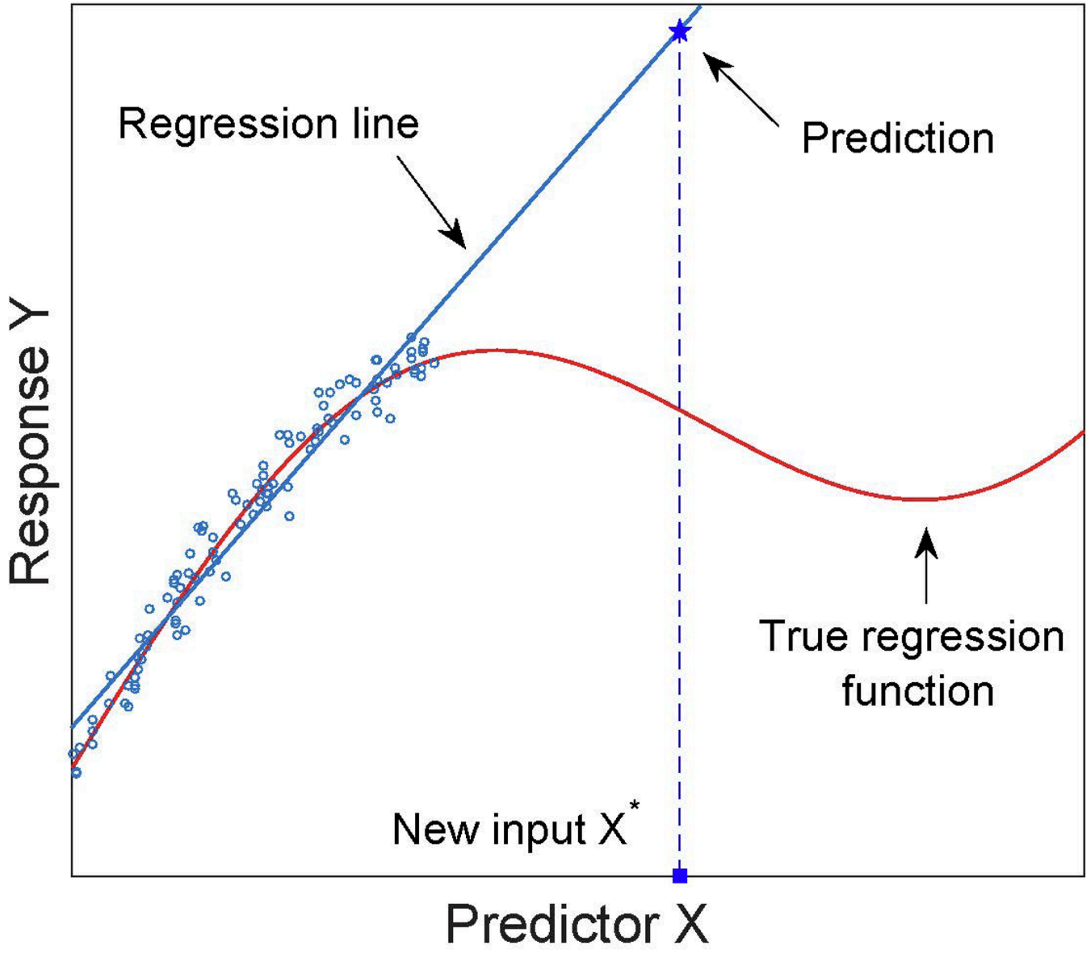


Answer:

- Don’t do the prediction because its 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.



Nuances of Predictions

2. How do we know if our predictions are any good? For example, how do we know if our prediction for Shaylee’s height was good or bad?
    - Issue: To evaluate how well we do at predicting, we essentially need to know the true answer of the thing we are predicting for.
    - Solution: Cross-validation

Principles of K-Fold CV

- **Purpose:** Assess how well your model does at predicting
- **General Idea:** Repeatedly fit your model to part of your data then see how well your model predicts the remainder of your data

Statistics 121

K-Fold Cross Validation

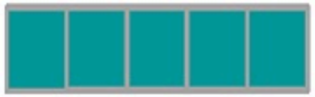
Step 1
Choose the number of folds, K (perhaps K=5). The class app will do the rest of the work. But I will explain it here so you know what it is doing.

Step 2
Randomly assign each data point from your sample into one of the 5 folds

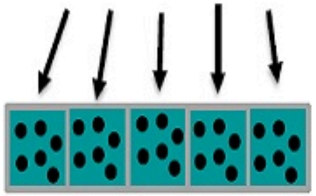
Step 3
Choose one fold to be your validation set. The other four folds are your training set.

Step 4
Fit a regression model using your training set

K=5 Folds
(Fold = group)

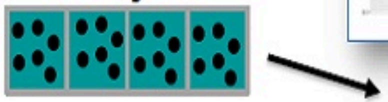


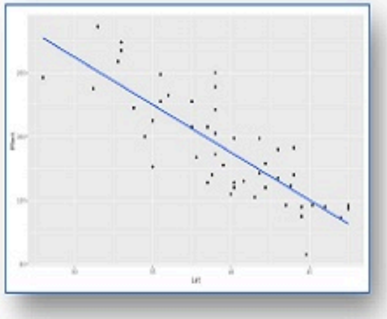
All data randomly assigned to one of the five folds



Validation data

Training data





$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Statistics 121

RMSE for K-Fold Cross Validation

Step 5
Use each x value in your validation set to predict y (i.e., plug each x value into the best fit regression equation and solve for $\hat{y}$)

Step 6
Calculate RMSE

Step 7
Repeat for all K folds

Actual measurements from our validation fold

Latitude	Mortality
33	219
34.5	160
35	170
37.5	182
39	149

➡

Predicted mortality for each latitude in our validation set

Predicted y
191.9274
182.9609
179.9721
165.0280
156.0616

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

From Validation data

From Training data

Next, choose a different fold to be your validation set. Use the remaining four folds as your training set. Fit a regression model to the training set. Predict y for each of the x values in your validation set. Calculate RMSE. Repeat again until you have performed the steps K times, changing the training and validation sets each time and calculating RMSE each time.

Melanoma Mortality Rates (per 10million) by Latitude

RMSE for K-Fold Cross Validation

Final Step
Calculate Average RMSE

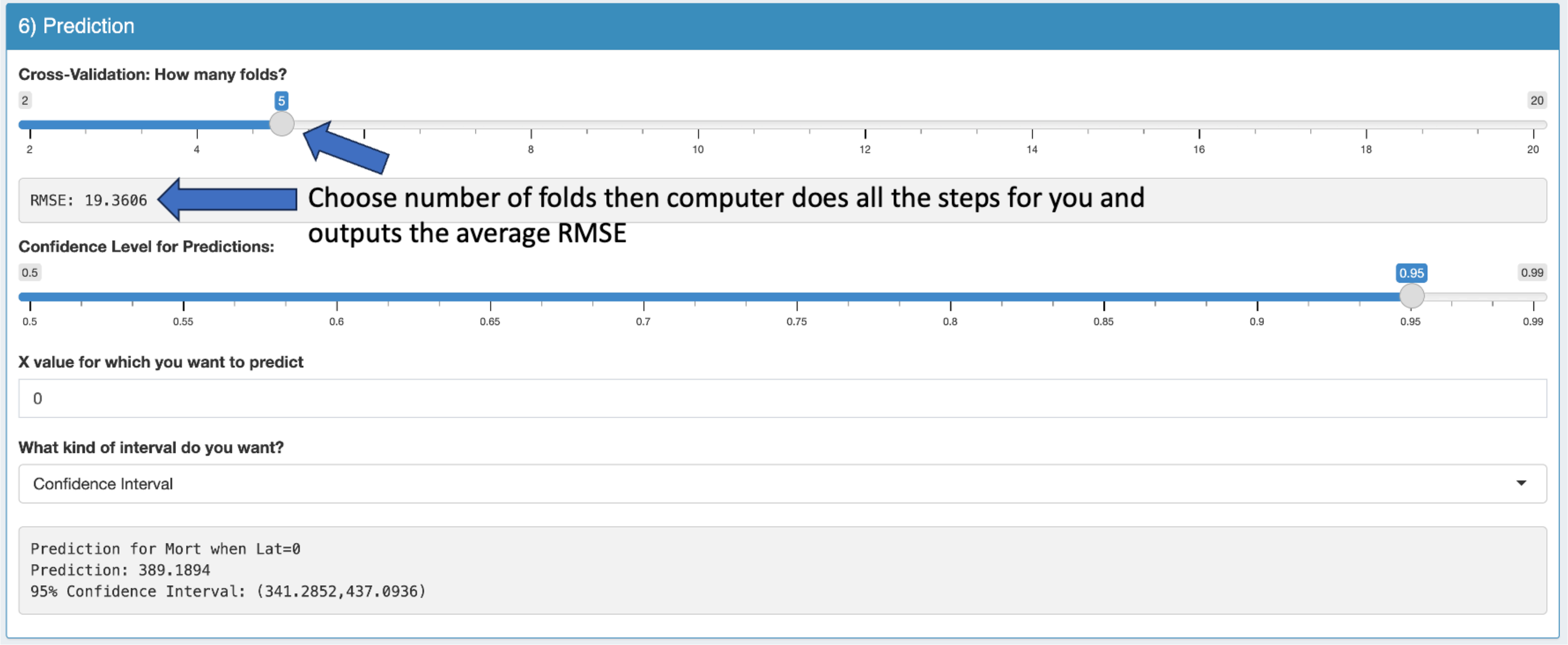
Take the average of your K RMSE values. That is the final assessment of your model fit. That composite RMSE value is the RMSE value reported in the course app.

The composite RMSE represents the average error in your model's ability to predict.

- If it is low, the model does a good job at prediction.
- If it is high, the model does not do a good job at prediction.
- What is considered "low" is subjective.

$$Composite\ RMSE = \frac{RMSE_1 + RMSE_2 + \cdots + RMSE_K}{K}$$

Using the Analysis Tool



6) Prediction

Cross-Validation: How many folds?

5

```
RMSE: 19.3606
```

Choose number of folds then computer does all the steps for you and outputs the average RMSE

Confidence Level for Predictions:

0.95

X value for which you want to predict

0

What kind of interval do you want?

Confidence Interval

```
Prediction for Mort when Lat=0
Prediction: 389.1894
95% Confidence Interval: (341.2852,437.0936)
```

Nuances of Cross Validation

1. Randomly split the data into folds $\longrightarrow$ every run of cross-validation will give slightly different results
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 across our predictions.
4. What constitutes a “small” RMSE is relative to the problem.

Additional Prediction Practice

Measuring possum head size can be difficult. However, measuring total possum length is easier. What is the relationship between possum length and head size? Use a simple linear regression model (and the course app) to answer the following questions:

1. Sydney found a huge 96 cm possum. What is a 95% interval for the head length for this possum?
    - 95% *prediction* interval is (92.431, 102.986).
2. Sydney found a huge 96 cm possum. What is a 95% interval for the average head length for possums of this size?
    - 95% *confidence* interval is (96.545, 98.872).
3. Sydney found a baby 70 cm possum. What is your predicted head length for this possum?
    - EXTRAPOLATION!
4. Is your model good or bad at predicting possum head sizes?
    - The RMSE of a 104 fold CV is 2.0132492.

Key Terminology

- Confidence Intervals for Averages
- Extrapoloation

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