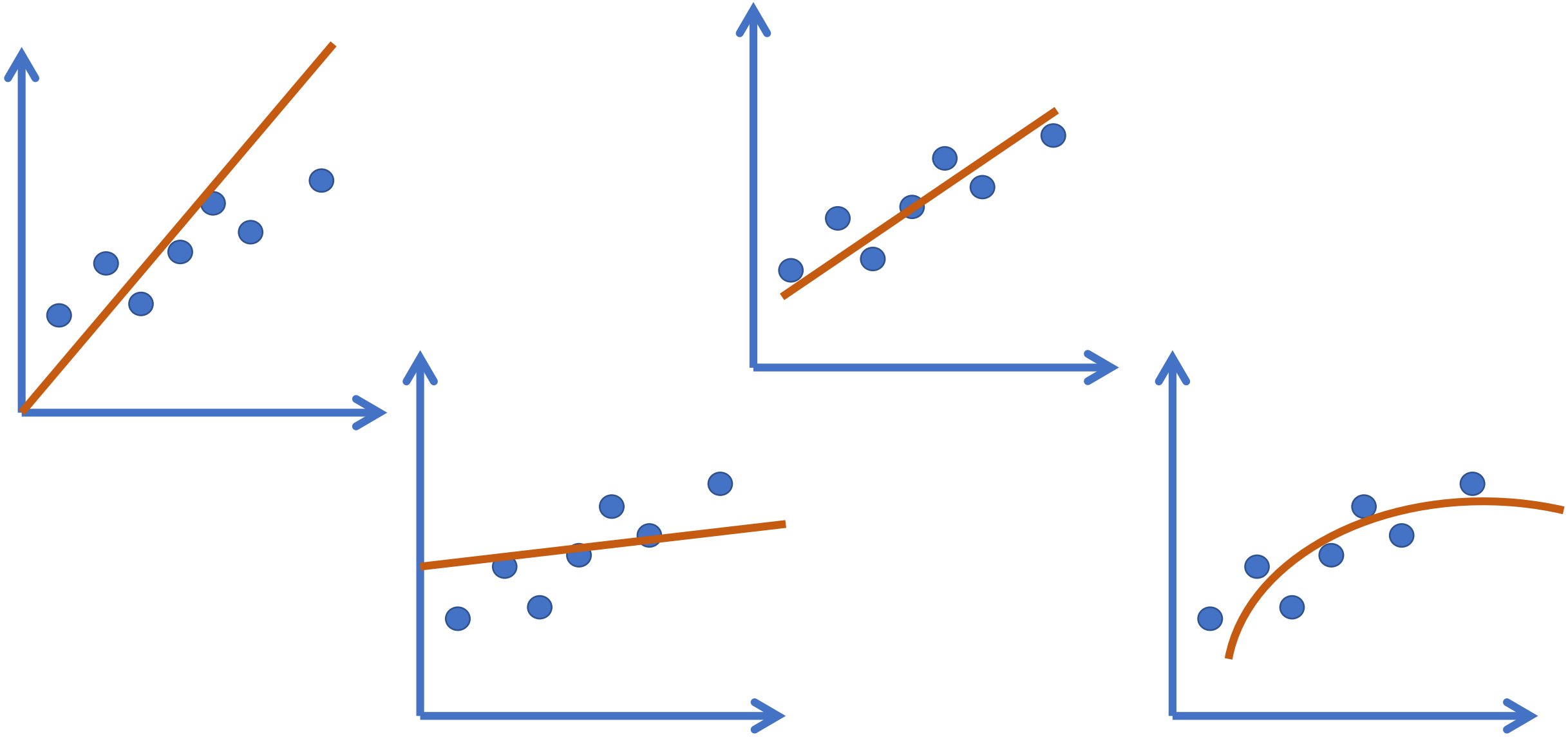
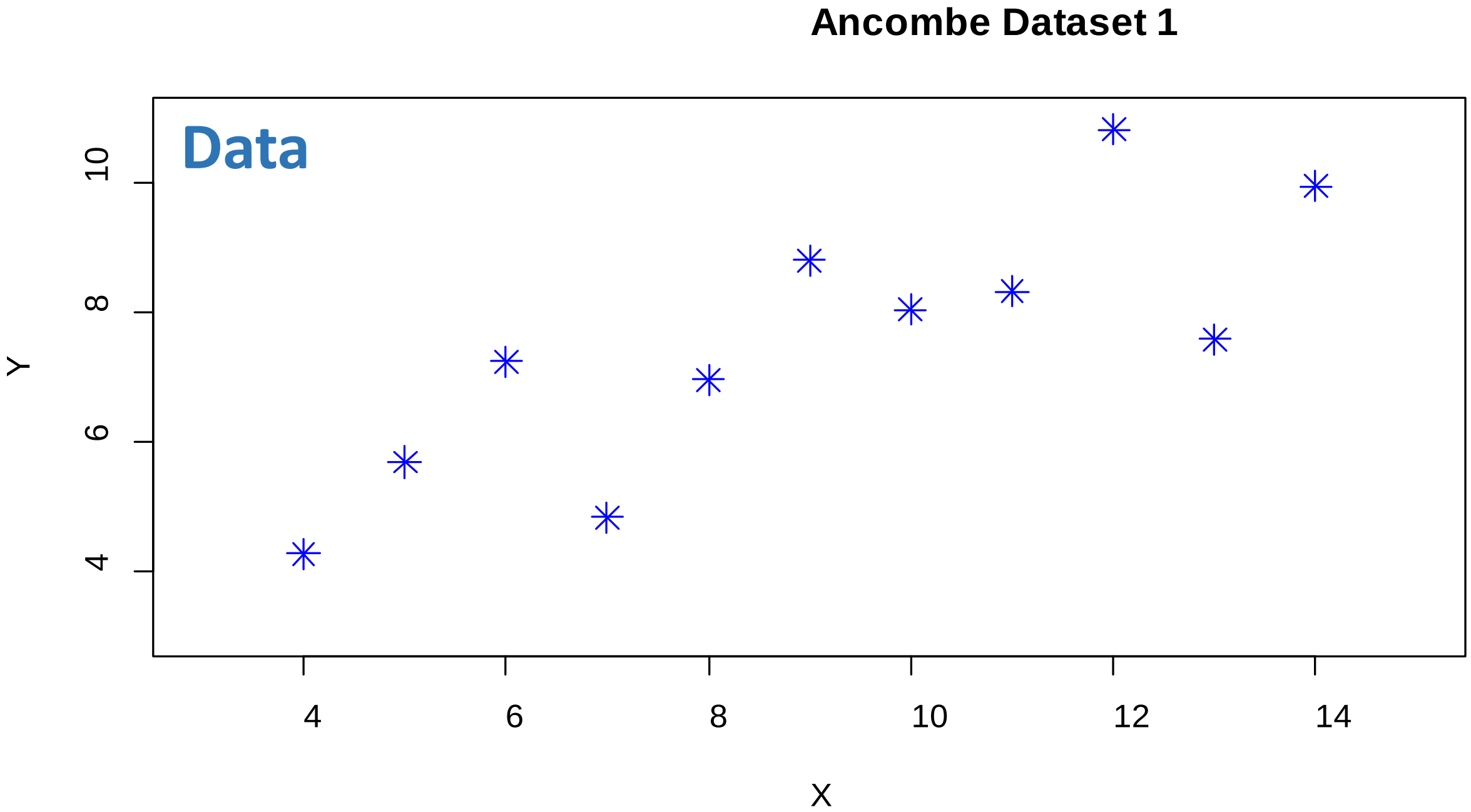
Modeling and Prediction Training and Testing

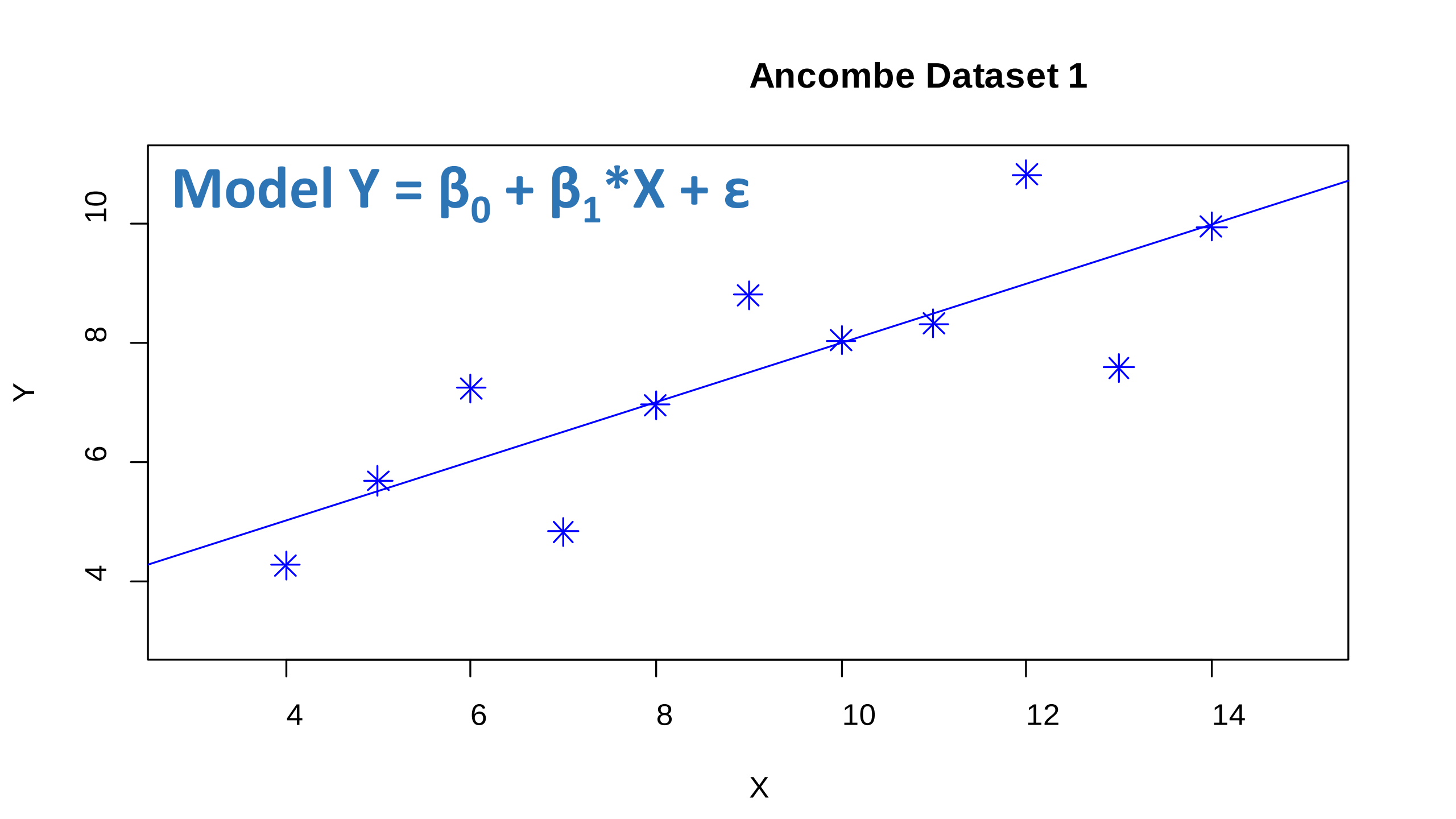
Melinda Higgins

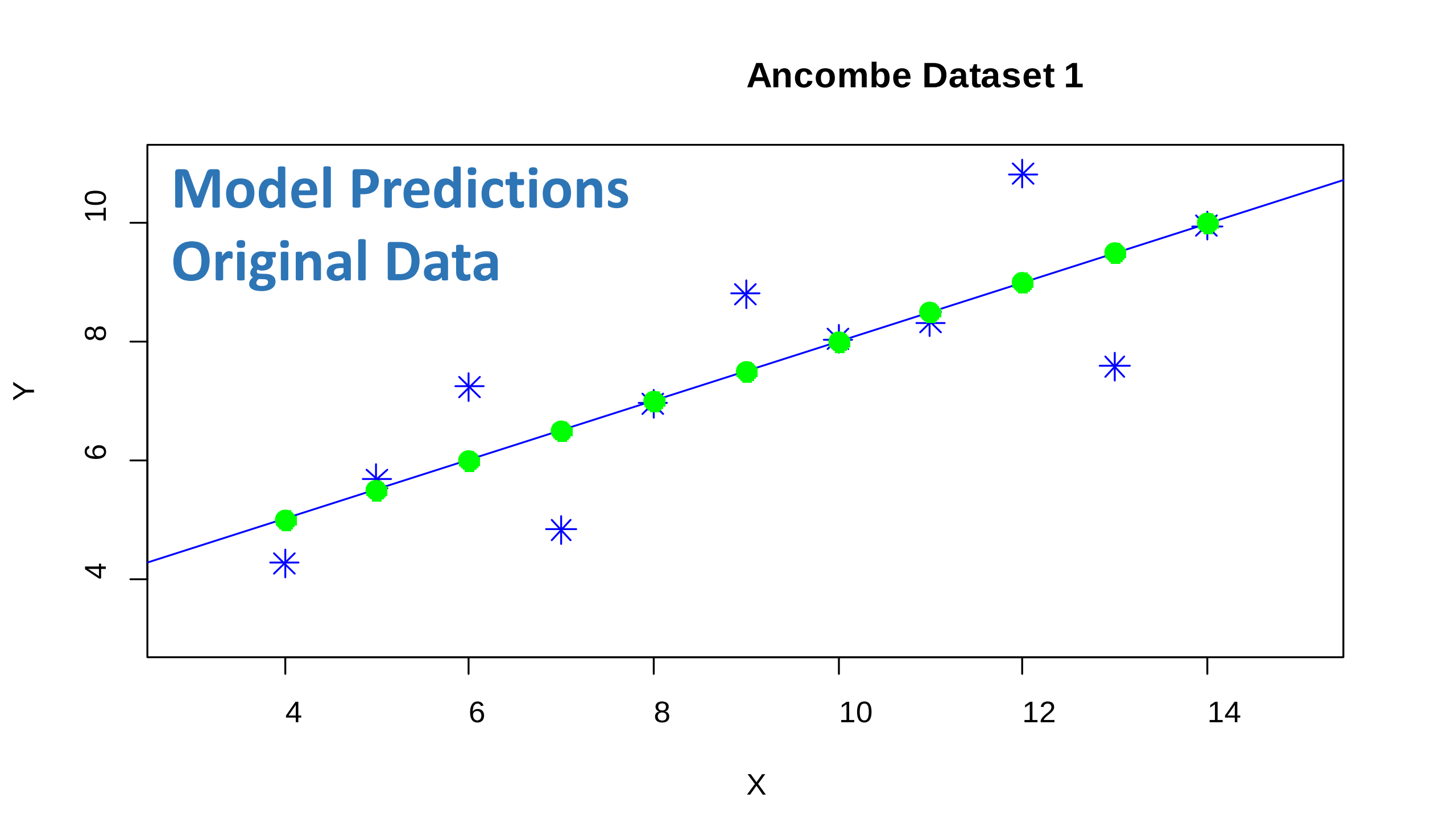
03/25/2020

Modeling – fit equation to data




Ancombe Dataset 1
Data
Y
10
8
6
4
4
6
8
10
12
14
X


Ancombe Dataset 1
Model Y = β0 + β1*X + ε
Y
X
4
6
8
10
12
14
4
6
8
10


Ancombe Dataset 1
Model Predictions
Original Data
Y
X
4
6
8
10
12
14
4
6
8
10

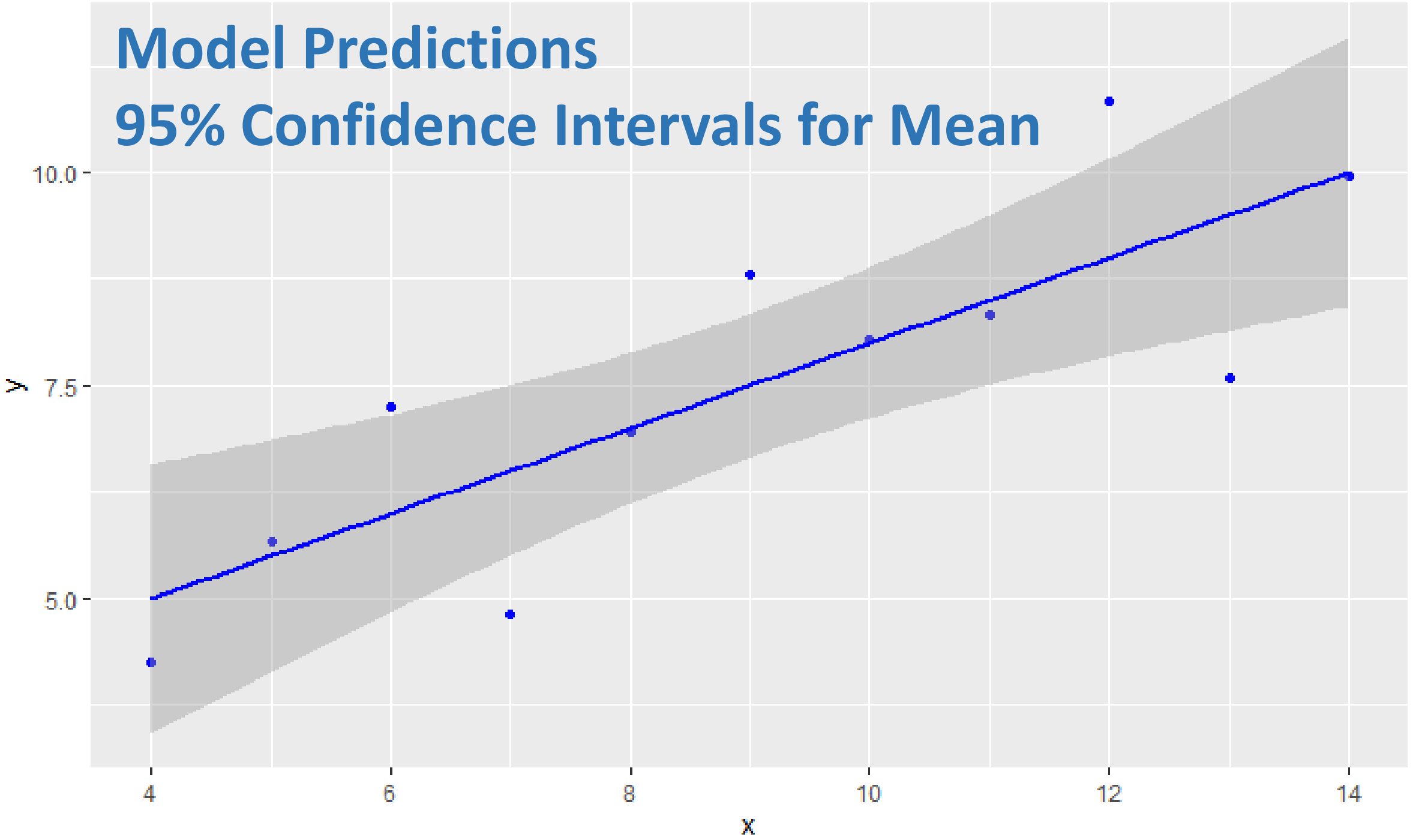
Model $Y = \beta_0 + \beta_1 * X + \varepsilon$

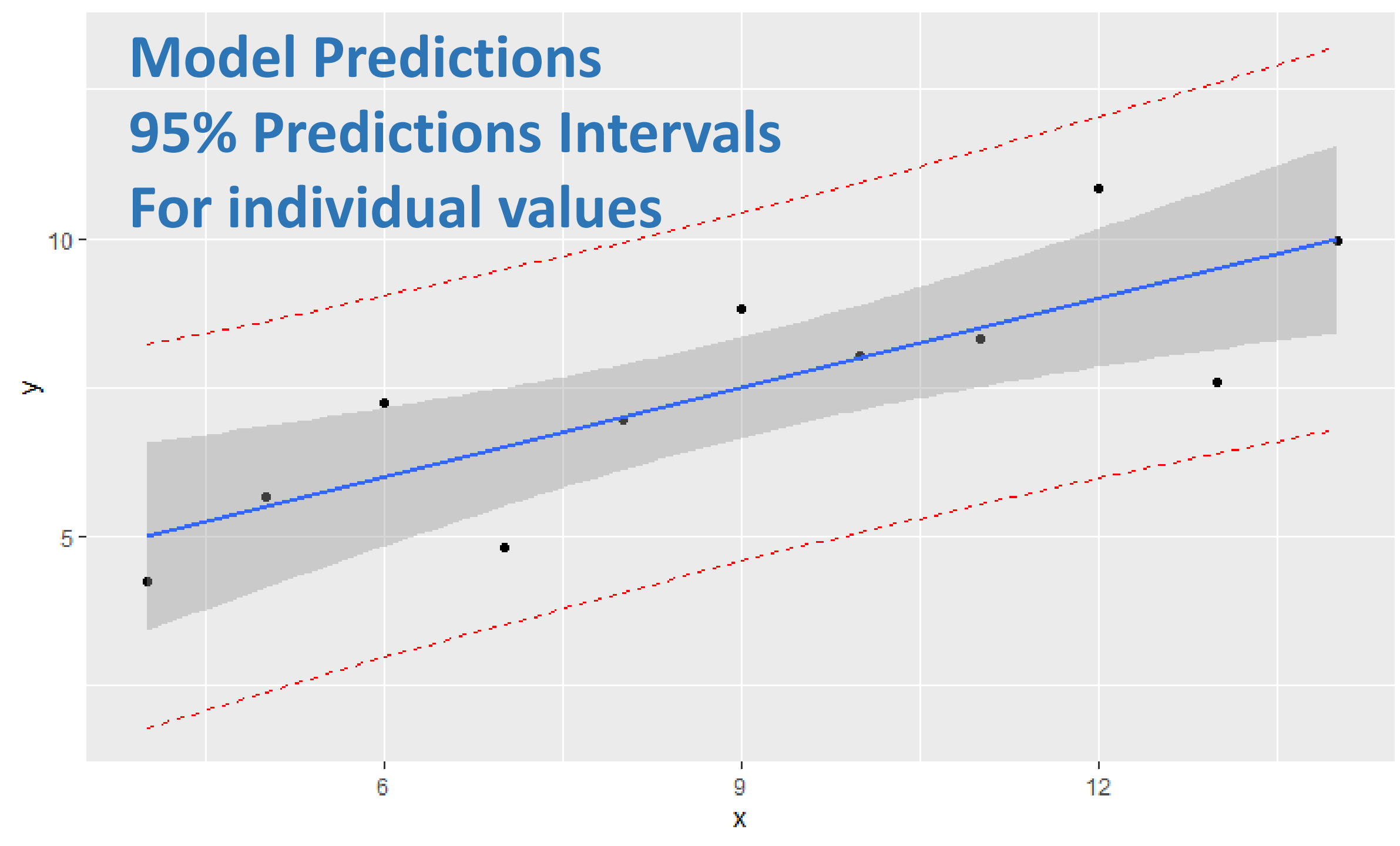
- Model “Errors” (aka. “residuals”)
  - Incorrect model → lack of fit
  - Measurement uncertainty
    - Assumed to only be in Y (outcome response)
    - But can also be in X (independent “fixed”)

Model Predictions

- Predict the expected (mean) response at a given value of X
- 95% Confidence Intervals are for the MEAN (average) response of Y at X
- 95% Prediction Intervals are for an INDIVIDUAL response of Y at X – always WIDER
- BOTH intervals will be narrower where there is “more data” or “higher concentration of data points” – more information – more confidence
- NEVER EXTRAPOLATE (outside of X’s)

Model Predictions
95% Confidence Intervals for Mean

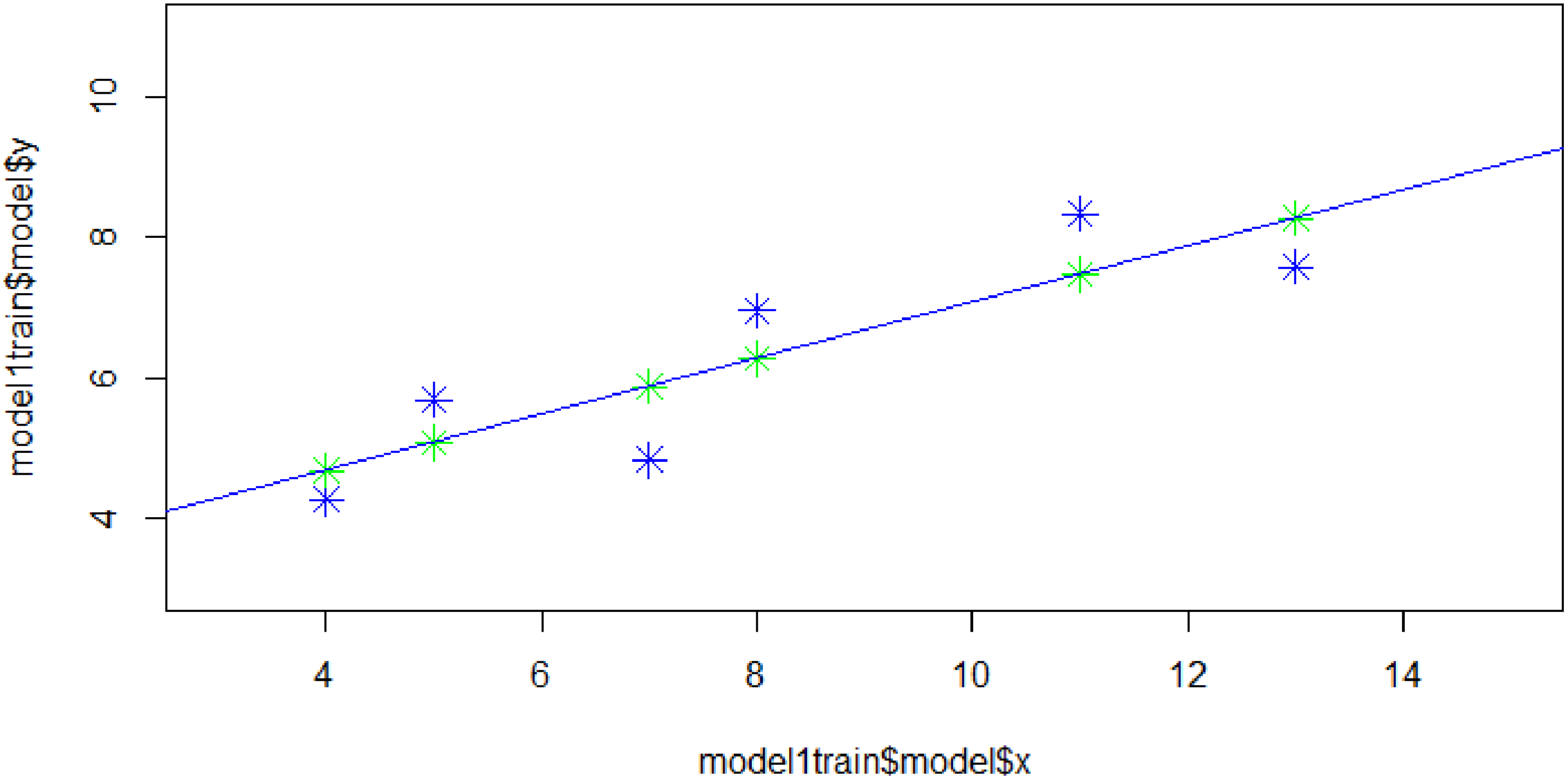



Model Predictions
95% Predictions Intervals
For individual values
10
5
y
6
9
12
x

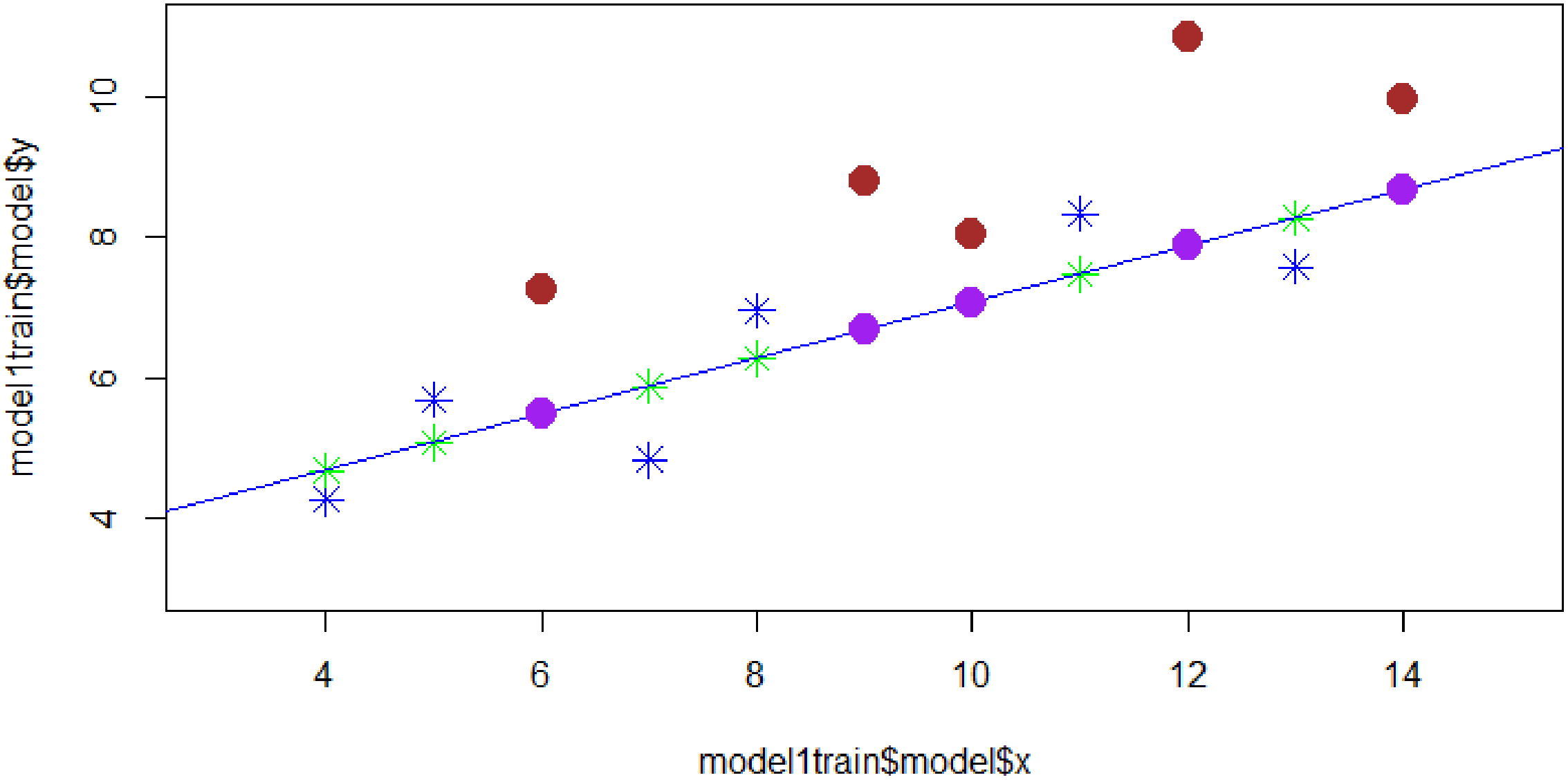
Training and Testing

- To get a good idea of how the model will perform with new (unknown, future) data
- Typically split the data usually 70/30 or 80/20 training, testing
- Sample at random (without replacement) 80% of the data
- Use these to TRAIN (create) the model
- Use the remaining 20% to TEST the model (prediction)

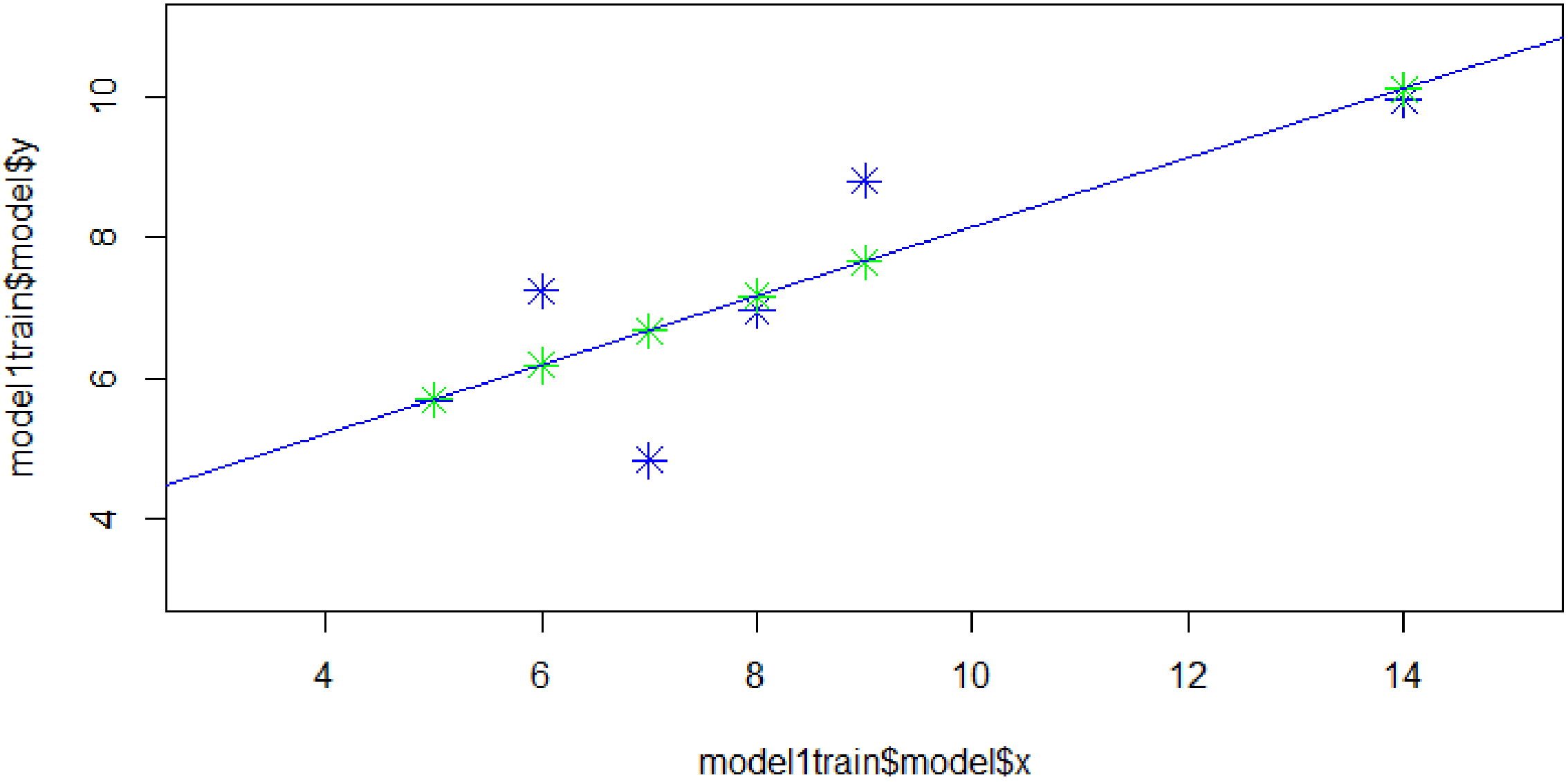
Anscome Data – Training – random set 1



Anscome Data – Test – random set 1



Anscome Data – Training – random set 2



Anscome Data – Test – random set 2