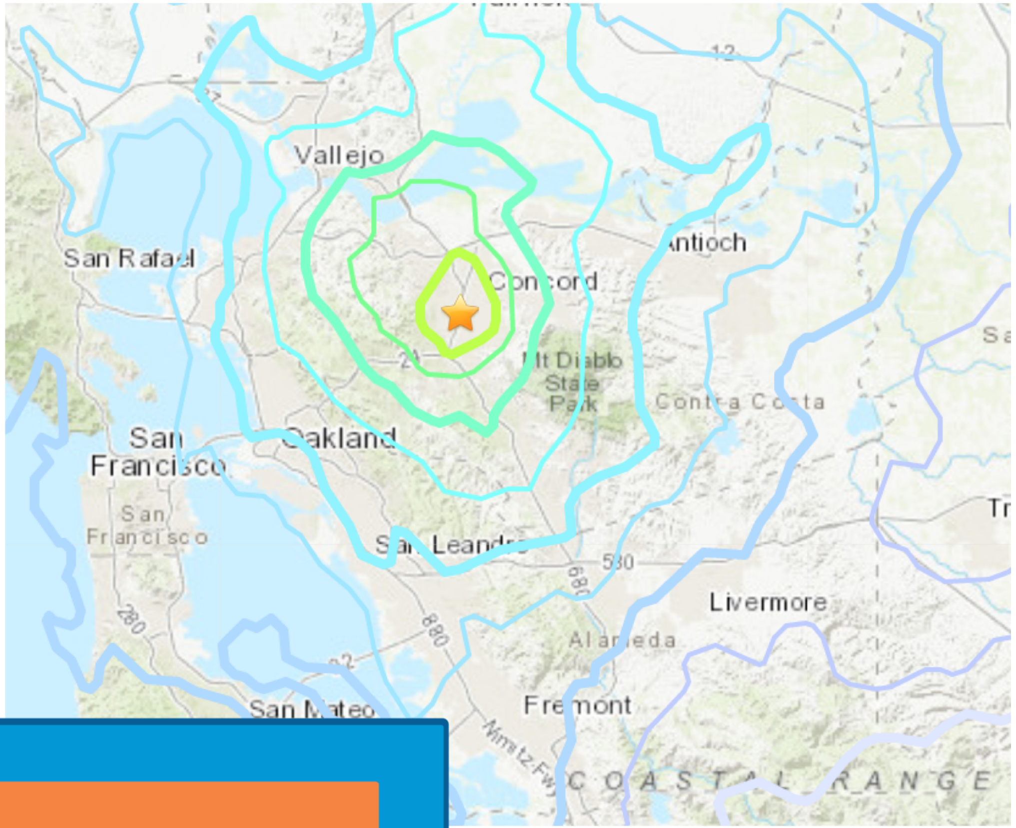
Simple Linear Regression

Today's Topics

- Review simple linear regression, including
  - Least squares
  - Correlation
  - Prediction
  - Inference
  - Hypothesis testing
- Connect regression to L_2 loss minimization
- Case Studies

Great ShakeOut Earthquake Drill 10/17 @ 10:17



Cancer Magister aka Dungeness Crab

All crab photos courtesy of Oregon Fish and Wildlife

Fishing Regulations

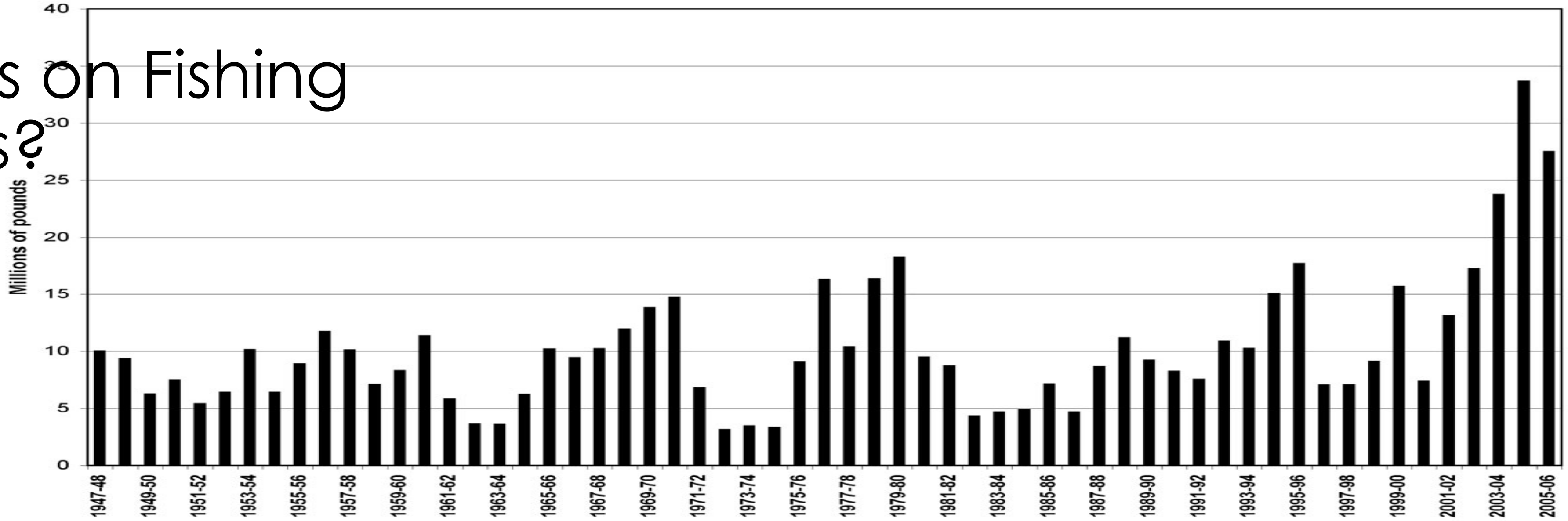
Male crabs only

No fishing in mating season

Limits on the numbers caught

Lift Restrictions on Fishing Female Crabs?

Dungeness crab landings 1947-2006



General Problem

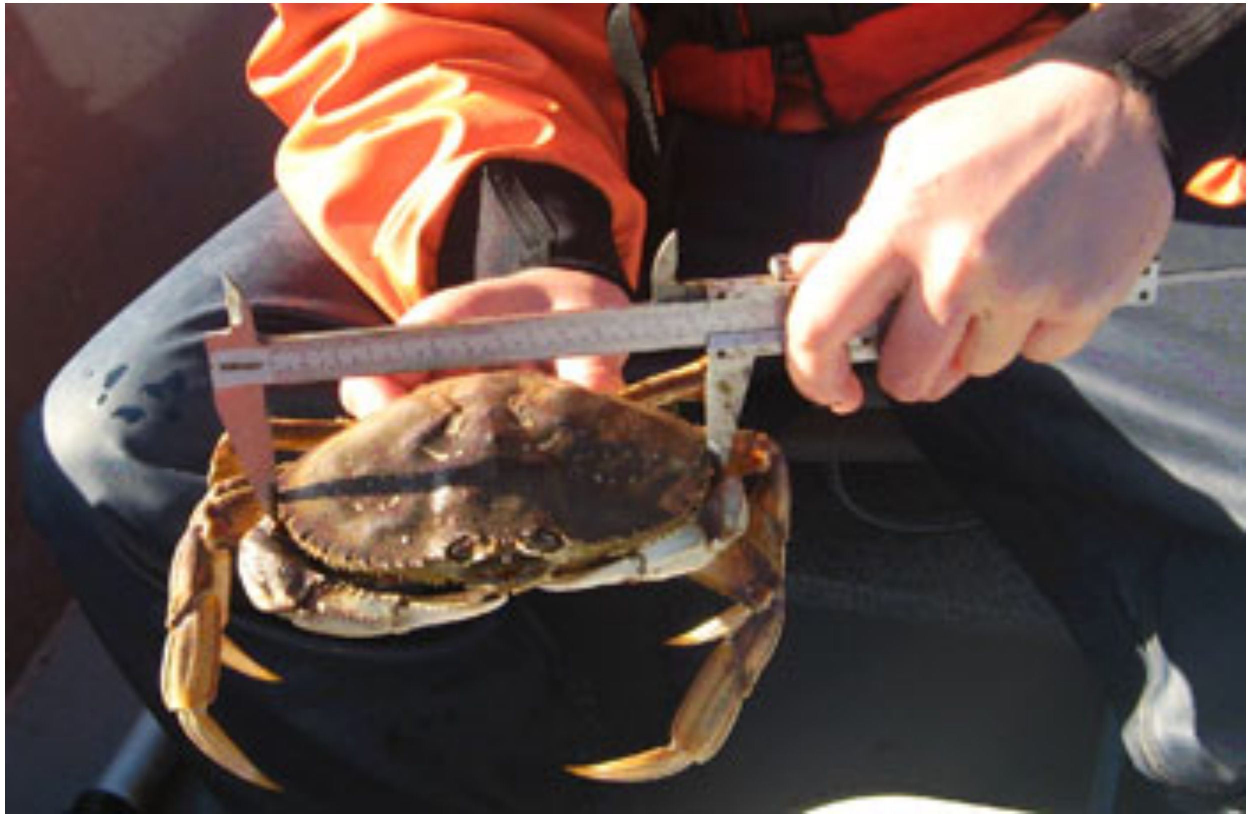
- Want to be sure that females have an opportunity to produce offspring for a few years before fished
- Can we use size to tell how old the crab is?
- Crabs has exoskeletons, which they shed every year - This makes it hard to estimate the age of a crab

Answerable Question:
Given a crab’s postmolt size,
Estimate how much it grew?

With this tool,
researchers can
estimate the age
of a crab.

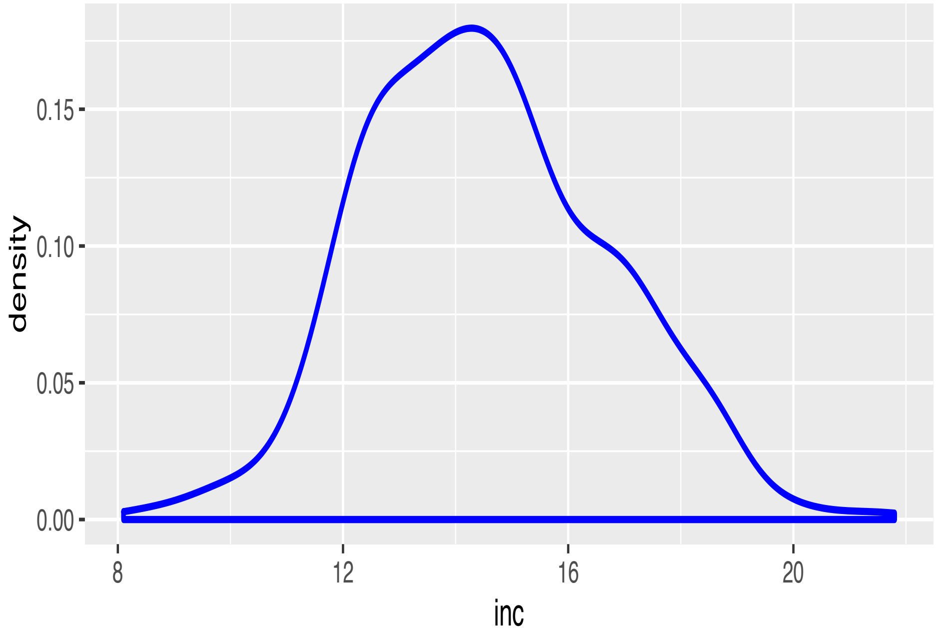
Data Collection Methods

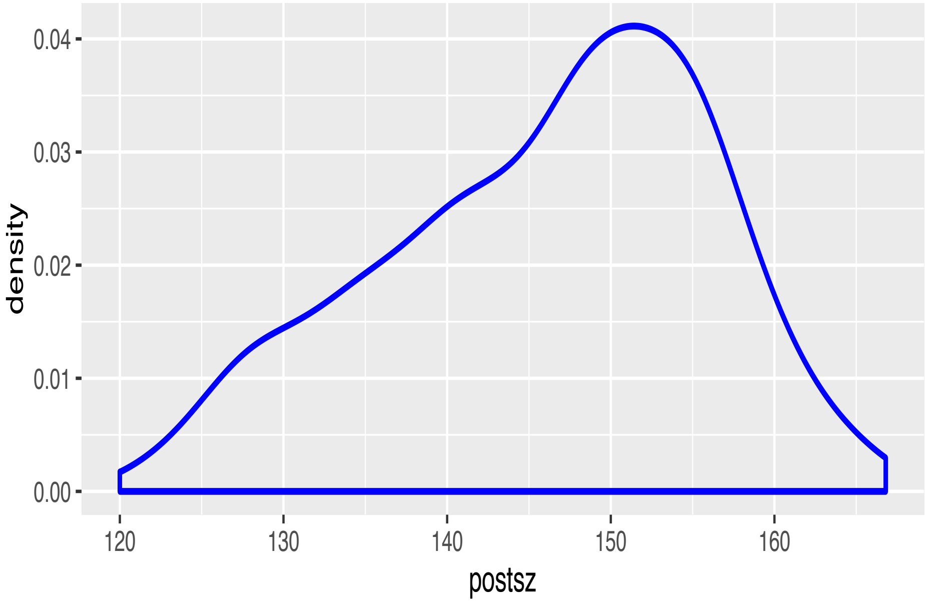
- Crabs were caught in mating embrace,
- Females measured before and after molting
- 452 crabs
- Variables
  - Premolt size (mm)
  - Postmolt size (mm)
  - Increment (mm)



Univariate Distributions

But what is their joint relationship?

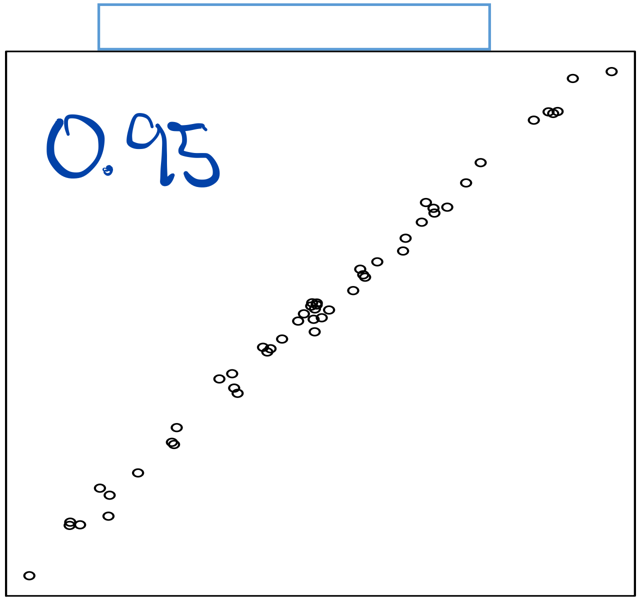




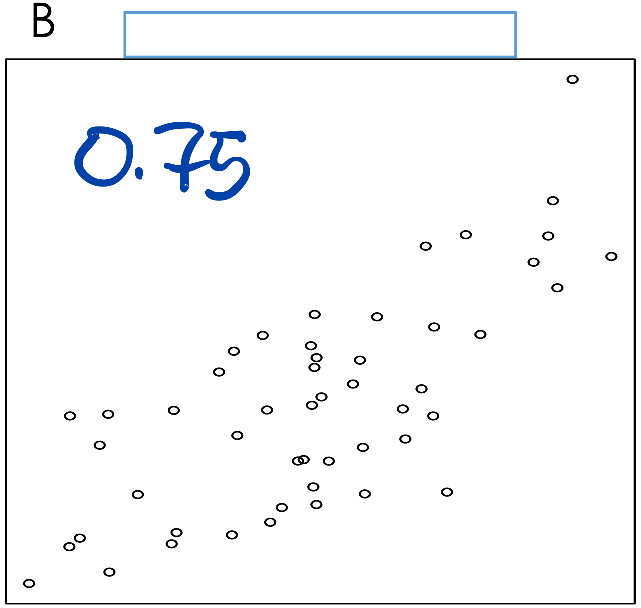
We see that postmolt size and increment are both unimodal and somewhat skewed. Growth increment is right skewed and postmolt size is left skewed.

Guess what the correlation is like

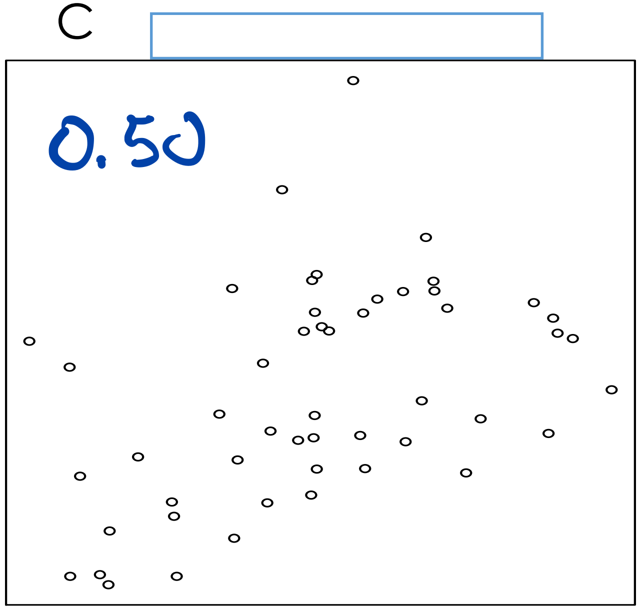
A



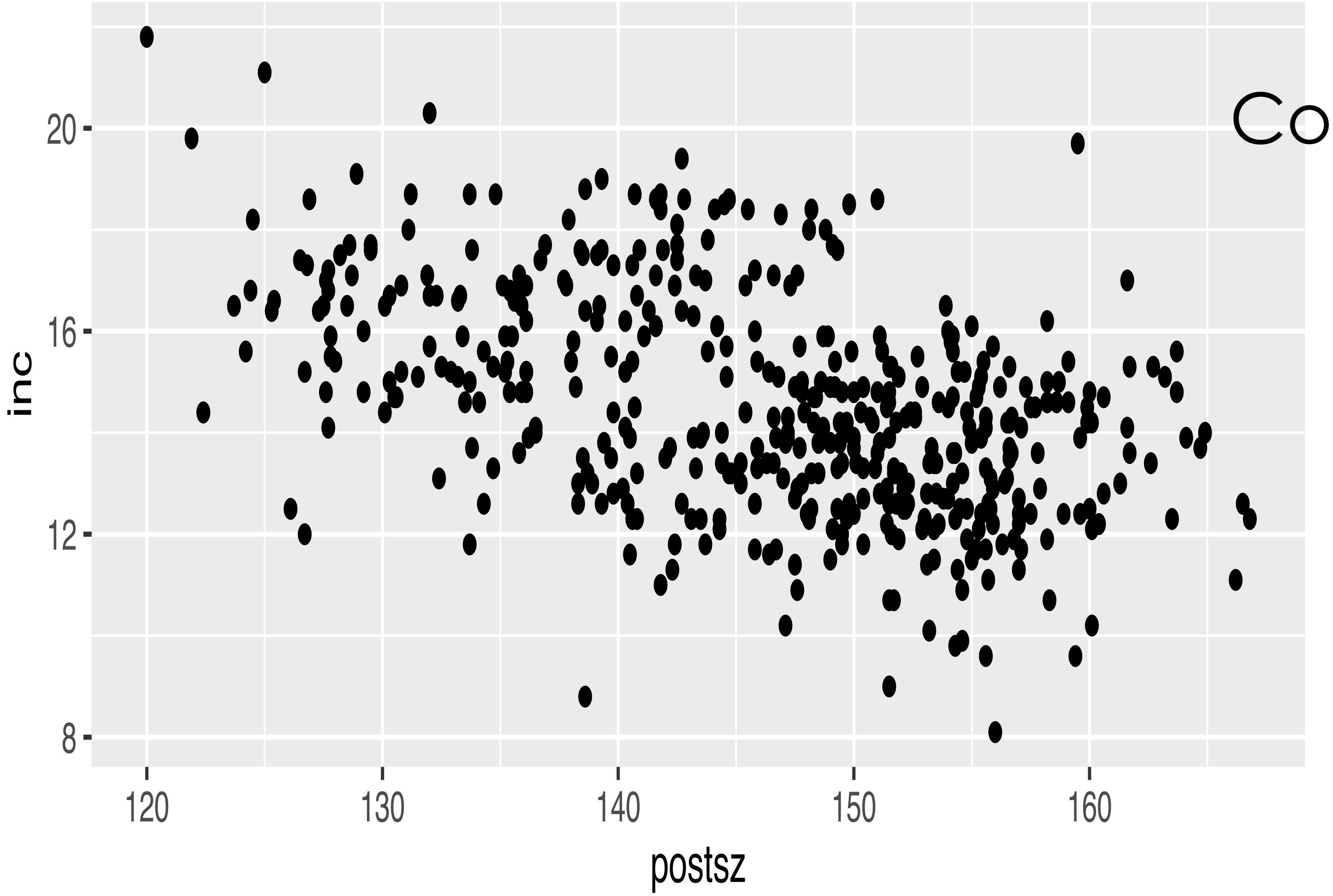
B



C



Relationship: postmolt & increment



Correlation = -0.77

Theres a negative correlation!

Rough linear association

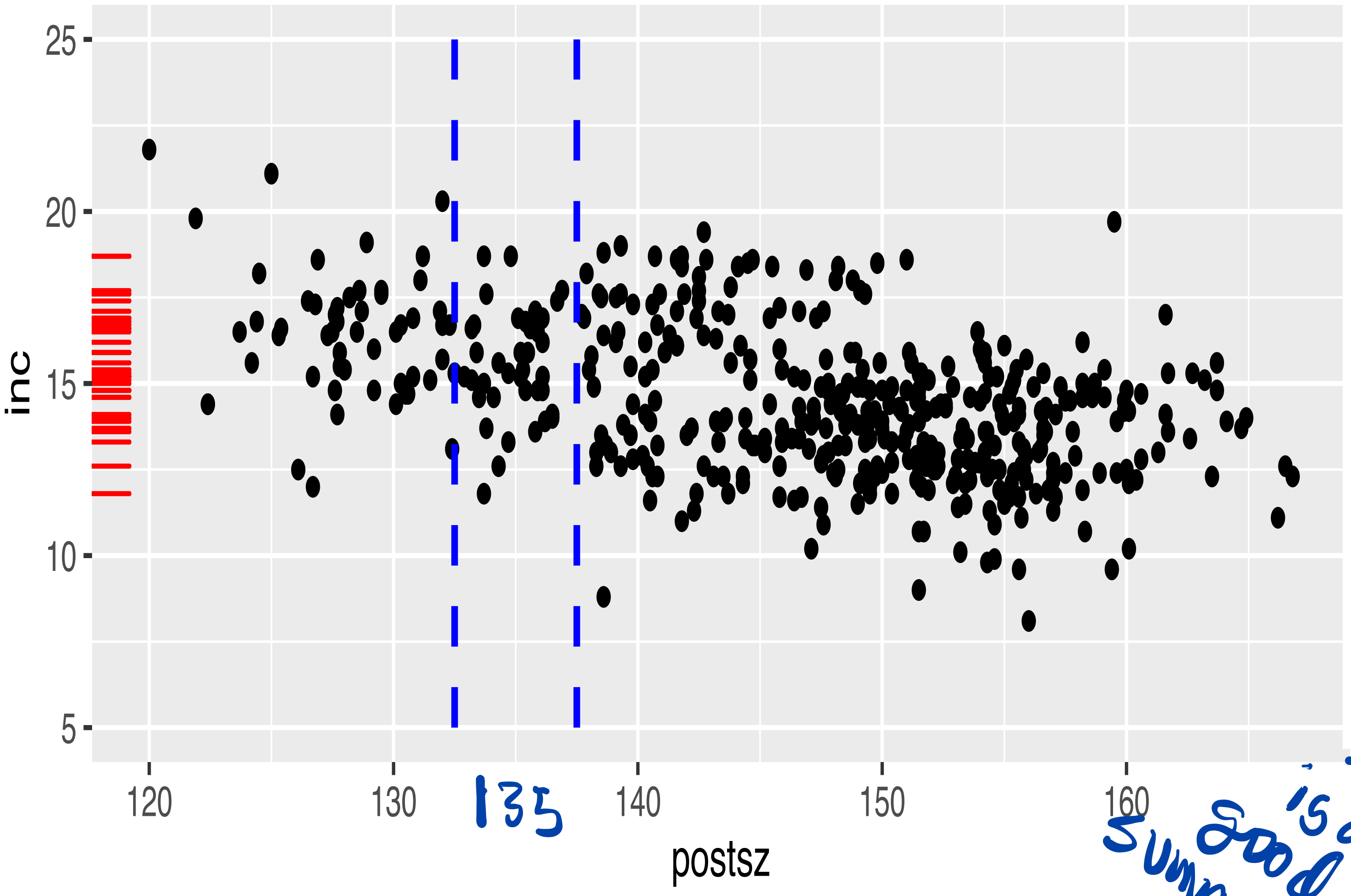
Larger crabs tend to have smaller increments

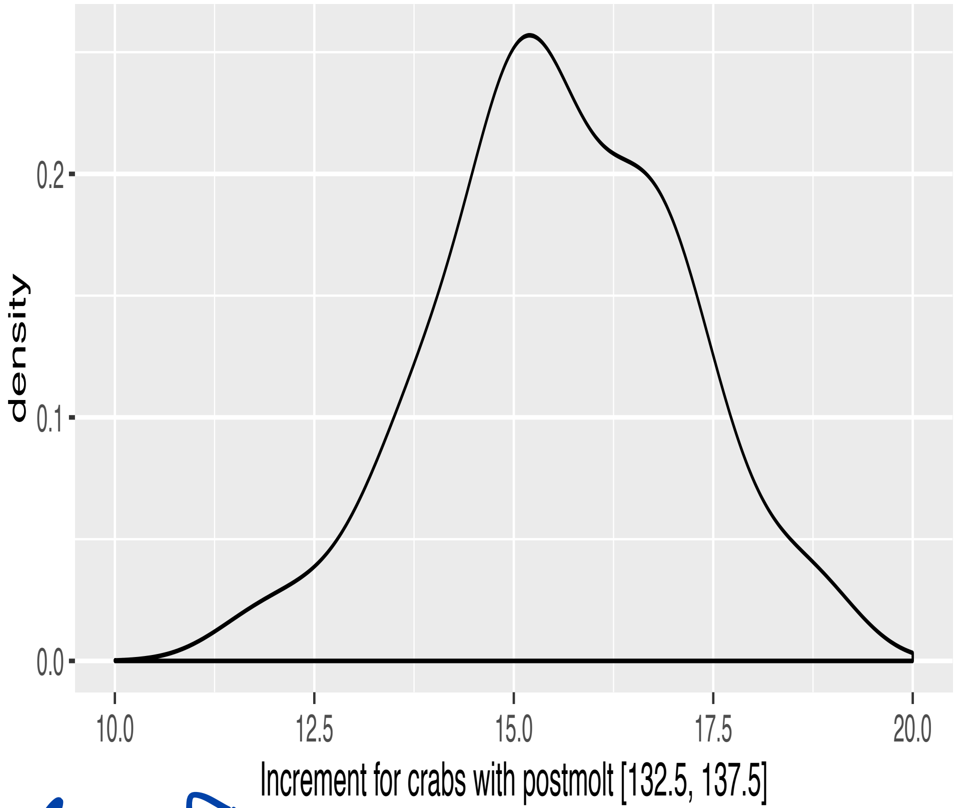
How can we use postmolt carapace size to predict the growth increment?

e.g., what do we predict for growth increment of a crab with 135 mm postmolt carapace?

Crabs with postmolt size 135

(to the nearest 2.5 mm)



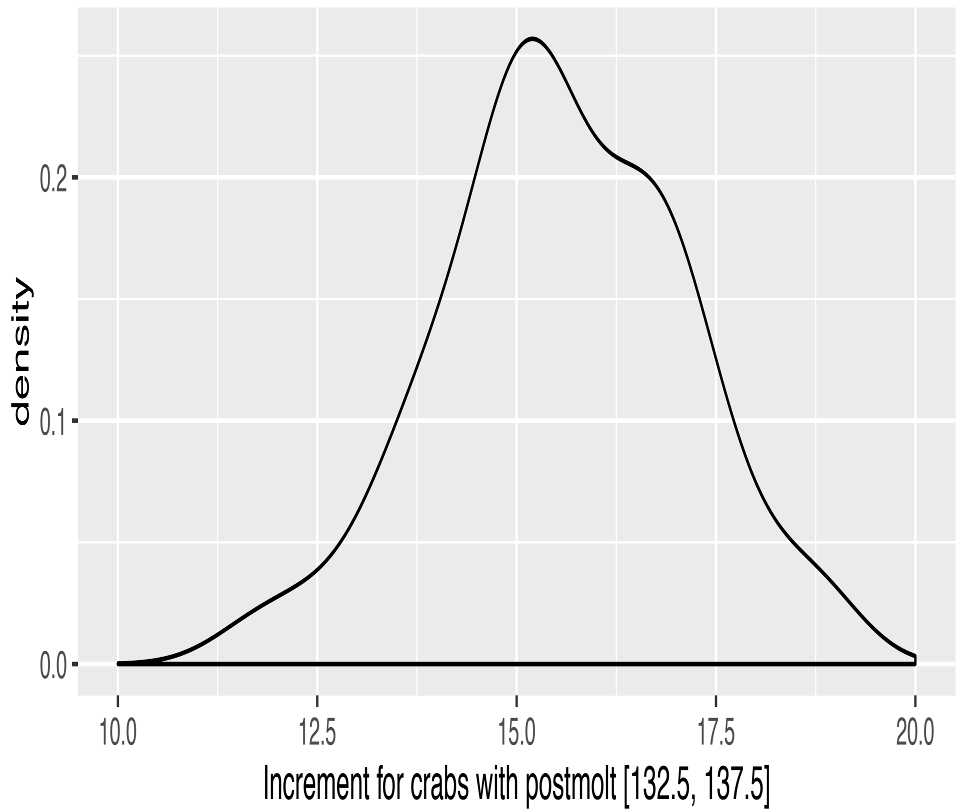


AVG is a good summary

unimodal
not too asymetric
reasonable tails

Crabs with postmolt size 135

(to the nearest 2.5 mm)

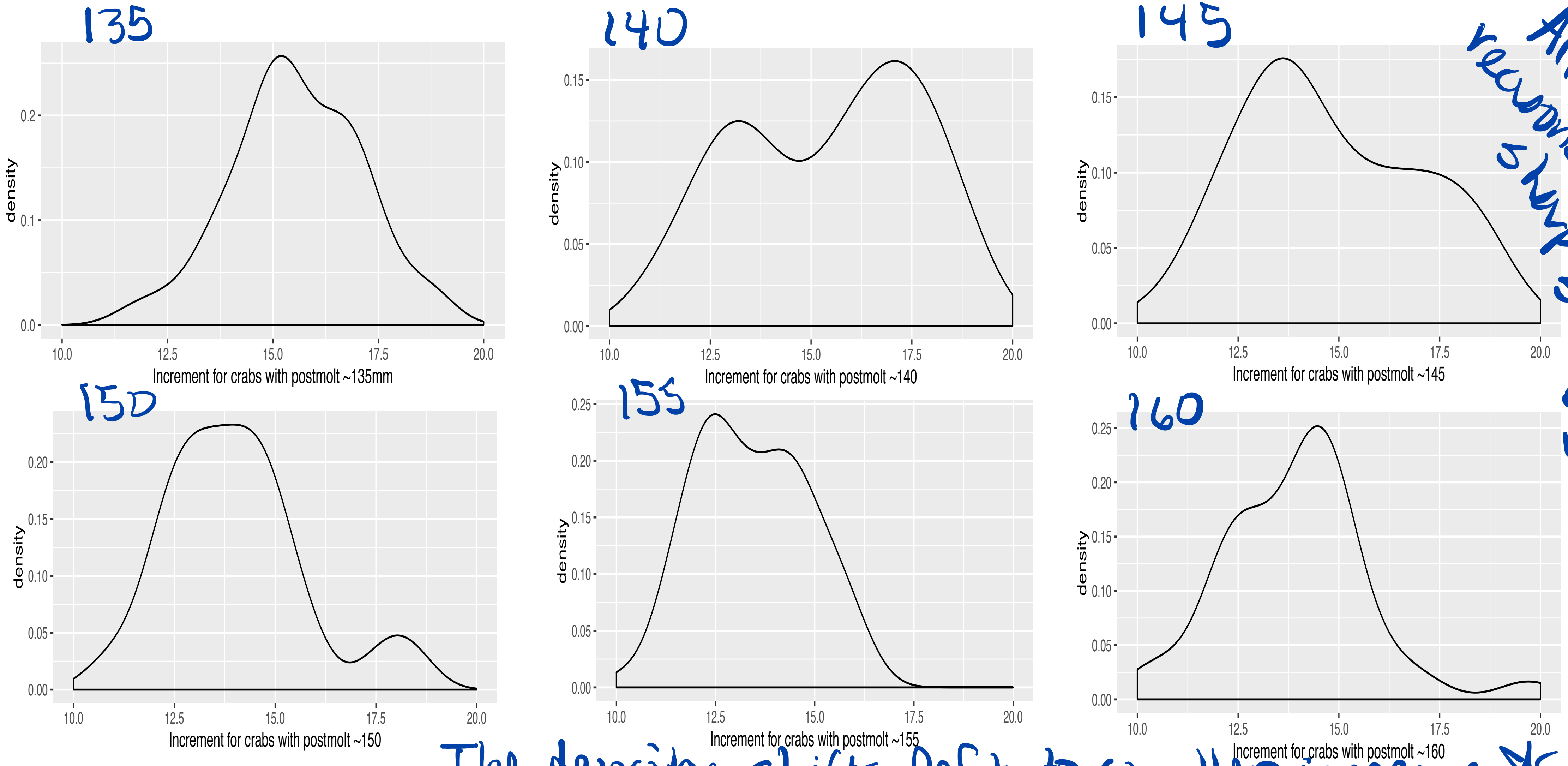


How would we summarize the growth increment for this subgroup?

$$\min_c \sum_{i:x_i \approx 135} (y_i - c)^2$$

$$\hat{c} = 15.6$$

Increment Distribution for fixed Post size



All are reasonable shapes to summarize by means

The density shifts left to smaller increments

For each bin of crabs

Let (x_i, y_i) represent the i^{th} crab's

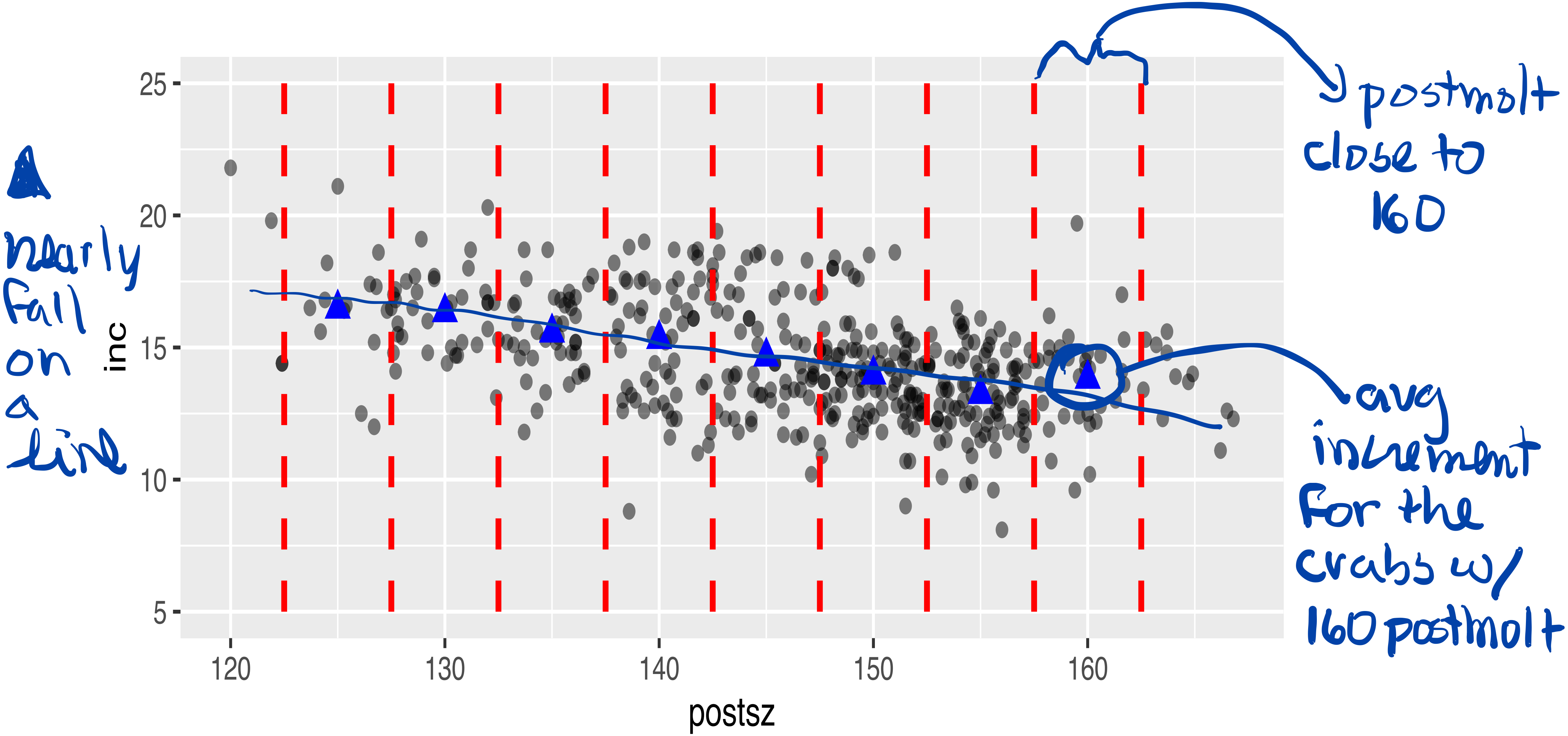
(postmolt size, growth increment)

For a bin of crabs with same postmolt size, predict increment

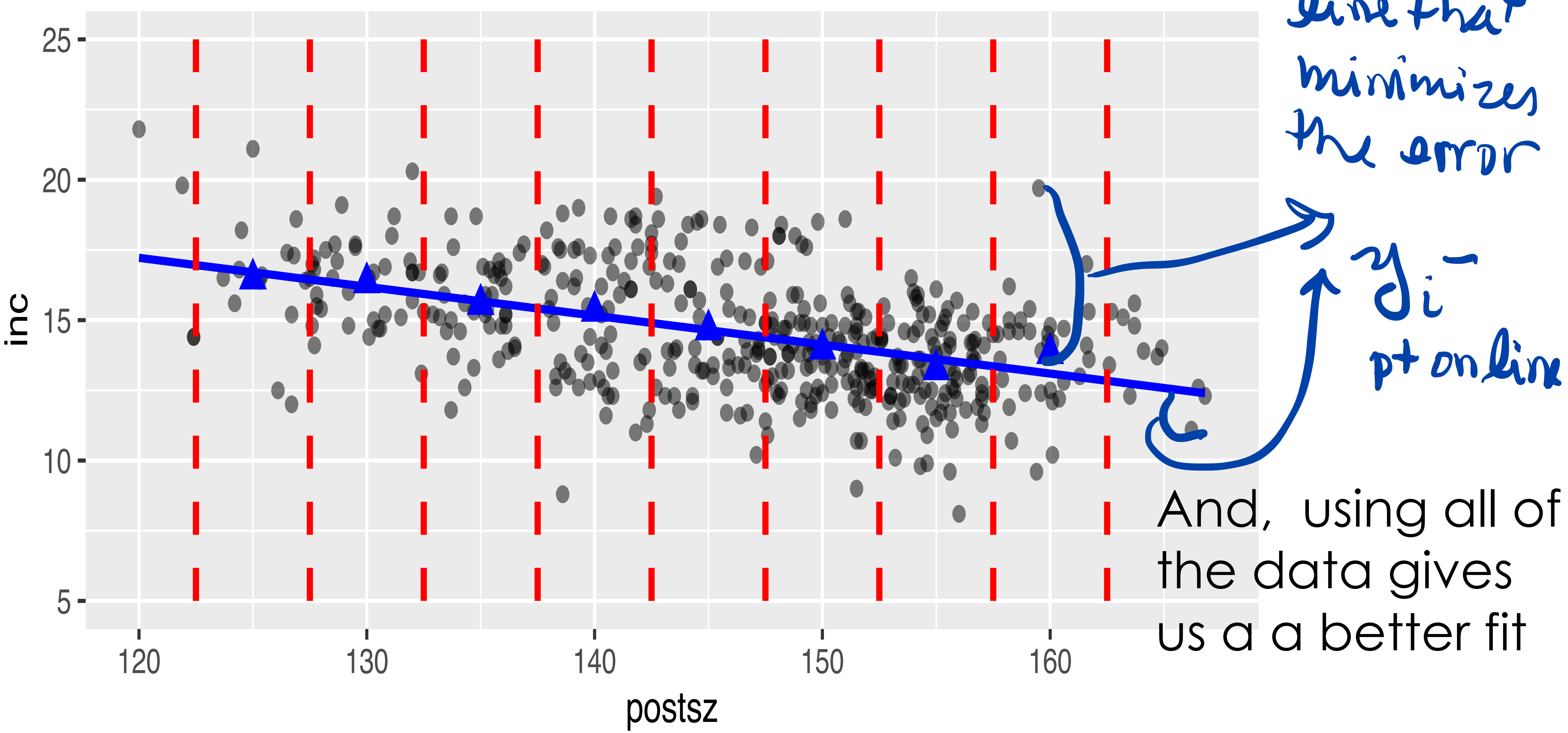
$$\min_{c} \sum_{i:x_i \in bin} [y_i - c]^2$$

We find the constant that minimizes L_2 empirical risk for the growth increment of crabs in a bin

Avg Increment for each postmolt bin



Averages Roughly fall on a line



And, using all of the data gives us a a better fit

Average Empirical Risk

Our data

(x_i, y_i) pairs

postmolt increment

For all of the data together:

Minimize empirical risk for estimating crab increment by a linear function of postmolt size

$$\min_{a,b} \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

point on line

$$min_{a,b} \sum_i (y_i - (a + bx_i))^2$$

First Term $0 = \sum y_i - \sum \hat{a} - \hat{b} x_i$

$\bar{y} n - n\hat{a} - \hat{b} \bar{x} n$

$\hat{a} = \bar{y} - \hat{b} \bar{x}$

- Derivative with respect to *a*

$$-2 \sum_i (y_i - a - bx_i)$$

- *Derivative with respect to b*

$$-2 \sum_i (y_i - a - bx_i) x_i$$

- *Set to 0 and solve for a and b*

Minimization:

$$\hat{a} = 30$$

$$\hat{b} = -0.10$$

Nice interpretation:
Predict growth increment to be
30 mm less 10% of the postmolt size

For a 135 mm postmolt crab, we predict its increment was
30 – 0.1x135 = 16.5 mm

Our binned mean was 15.6 mm

Which is better?

* If the relationship is roughly linear, then using all of the data to fit the line gives a better prediction

Fitted parameters:

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

Regression line: $\hat{y} = \hat{a} + \hat{b}x$

$$\hat{b} = r\frac{SD_y}{SD_x}$$

Rearrange terms:

$$\hat{y} = \bar{y} + rSD_y\frac{(x - \bar{x})}{SD_x}$$

← x in std units (subtract mean and divide by SD)

For an **x** that is, say 2 standard units above/below average, the regression line estimates y to be 2r standard units above/below average.

Least Squares Regression

Some Important Concepts

Correlation

Correlation measures the strength of linear association between x and y

- Correlation is a measure for two quantitative variables
- Need to plot the data to check if the relationship is linear

mm   mm

$$r(x,y) = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - \bar{x})}{SD_x} \times \frac{(y_i - \bar{y})}{SD_y}$$

mm

x is measured in mm

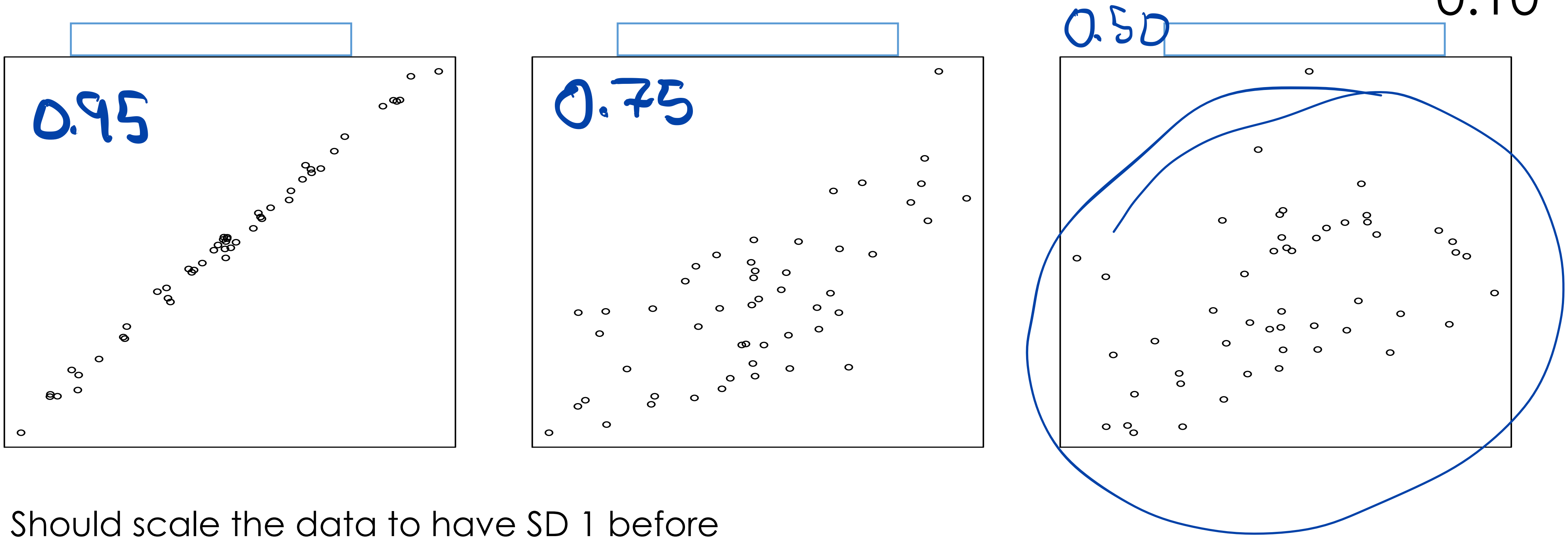
$\frac{x_i - \bar{x}}{SD_x}$ ← mm / ← mm cancel so unitless

Correlation is unitless

$$SD(x)^2 = Var(x)$$

Example Correlations for data with positive linear association (SDs = 1)

0.95
0.75
0.50
0.30
0.10

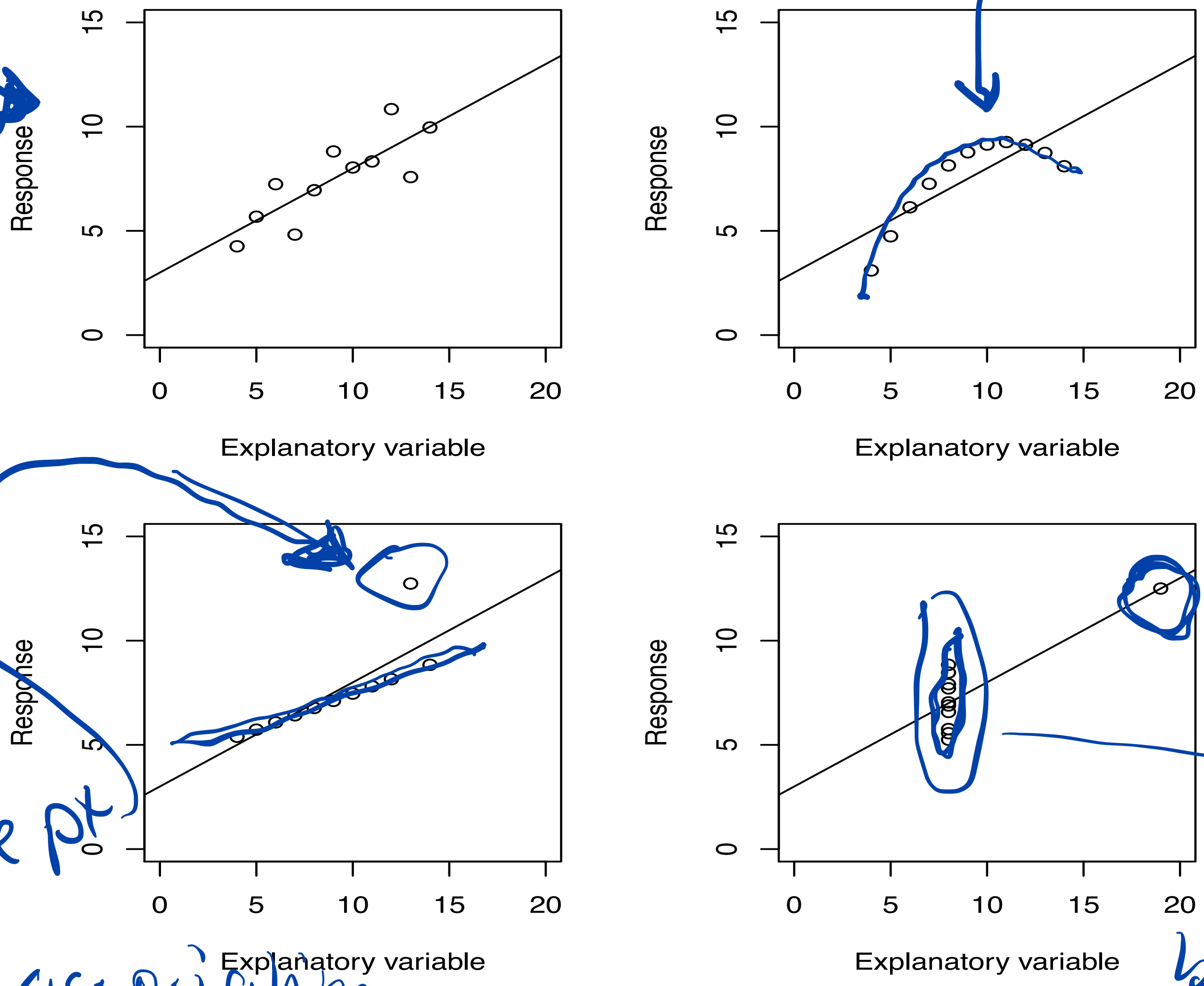


Should scale the data to have SD 1 before visually assessing the linear association

SAME regression line and correlation

points have a linear relationship

Graphical methods are important for assessing linearity



Pts have a perfect non linear association

With the exception of one pt The points have a perfect linear association

this 1 pt impacts the relationship

These points show no relation btw x & y

Correlation does not imply Causation

- Since this is not an experiment where we controlled the size of the postmolt size of the crab and observed its growth, we can not make any causal conclusions
- With observational studies we can observe and describe relationships.
- We can make predictions, but we need to be careful about the interpretation of the models that we build.

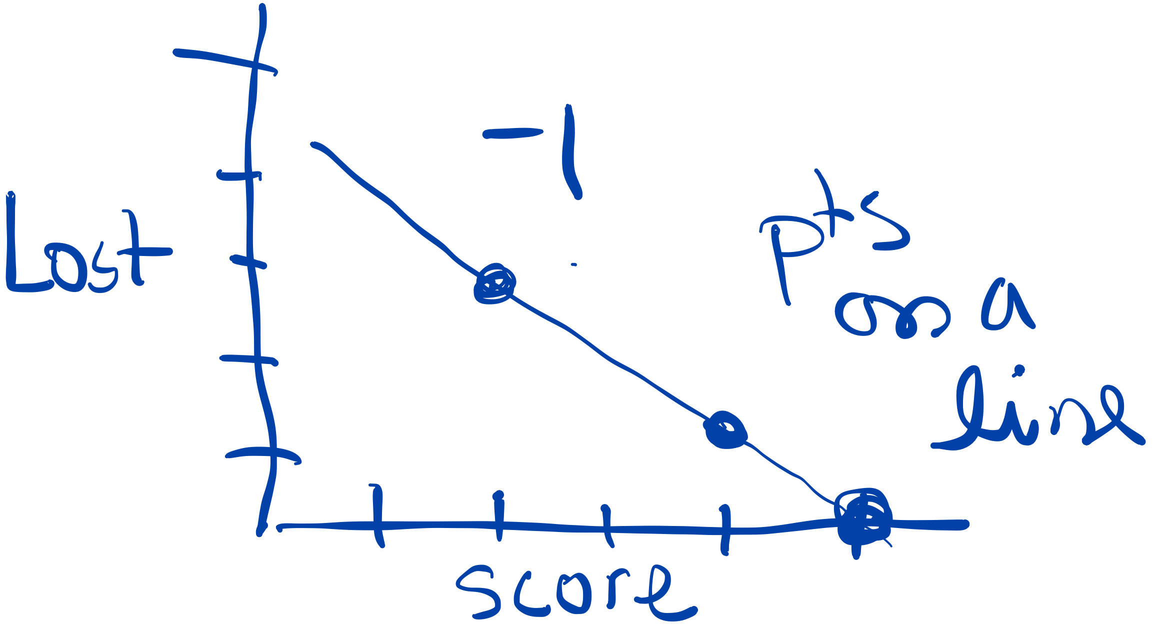
Correlation does not imply Causation

- Consider other variable(s) that is highly correlated with x.
- Correlation is still informative, even if we can't assign causality.
- .

An example of perfect correlation

- *score* on quiz (out of 25 points)
- *points_lost* on quiz
- The scatter plot of (*score, points_lost*) shows all the points fall on a line
- What's the correlation between the *score* and *points_lost*?

score	Points lost
25	0
20	5
22	3
15	10
25	0



Lost = 25 − score

↓ y ↓ x

$\bar{x}$ SD(x)

$$\bar{y} = \sum y_i / n = \sum \frac{25 - x_i}{n} = 25 - \bar{x}$$

$$\text{Var}(y) = \frac{1}{n}\sum (y_i - \bar{y})^2 = \frac{1}{n}\sum \left(25 - x_i - (25 - \bar{x})\right)^2$$

$$= \text{Var}(x)$$

$$r_{x,y} = \frac{1}{n}\sum\left(\frac{x_i-\bar{x}}{SD_x}\right)\left(\frac{y_i-\bar{y}}{SD_y}\right)$$

$$= \frac{1}{n}\sum\left(\frac{x_i-\bar{x}}{SD_x}\right)\left(\frac{25-x_i-(25-\bar{x})}{SD_x}\right)$$

$$\underbrace{\frac{25-x_i-(25-\bar{x})}{SD_x}}_{\to\ -\frac{(x_i-\bar{x})}{SD_x}}$$

25s cancel

$$= \frac{-1}{n}\frac{\sum(x_i-\bar{x})^2}{SD_x^2}$$

$$= -1$$

Find the correlation

$$y_i = 25 - x_i$$

$$\bar{y} = \frac{1}{n}\sum_i (25 - x_i) = 25 - \bar{x}$$

$$Var(y) = \frac{1}{n}\sum_i [25 - x_i - (25 - \bar{x})]^2 = Var(x)$$

$$r = \frac{1}{n}\sum_i \frac{x_i - \bar{x}}{SD(x)} \frac{y_i - \bar{y}}{SD(y)}$$

$$= \frac{1}{n}\sum_i \frac{x_i - \bar{x}}{SD(x)} \frac{\bar{x} - x_i}{SD(x)} \qquad = -1$$

In general, with a perfect linear association

$$y_i = a + bx_i \quad \text{for } i = 1, \ldots, n$$

$$\bar{y} = a + b\bar{x}$$

$$Var(y) = b^2 Var(x)$$

$$r = 1 \text{ if } b > 0 \qquad\qquad r = -1 \text{ if } b < 0$$

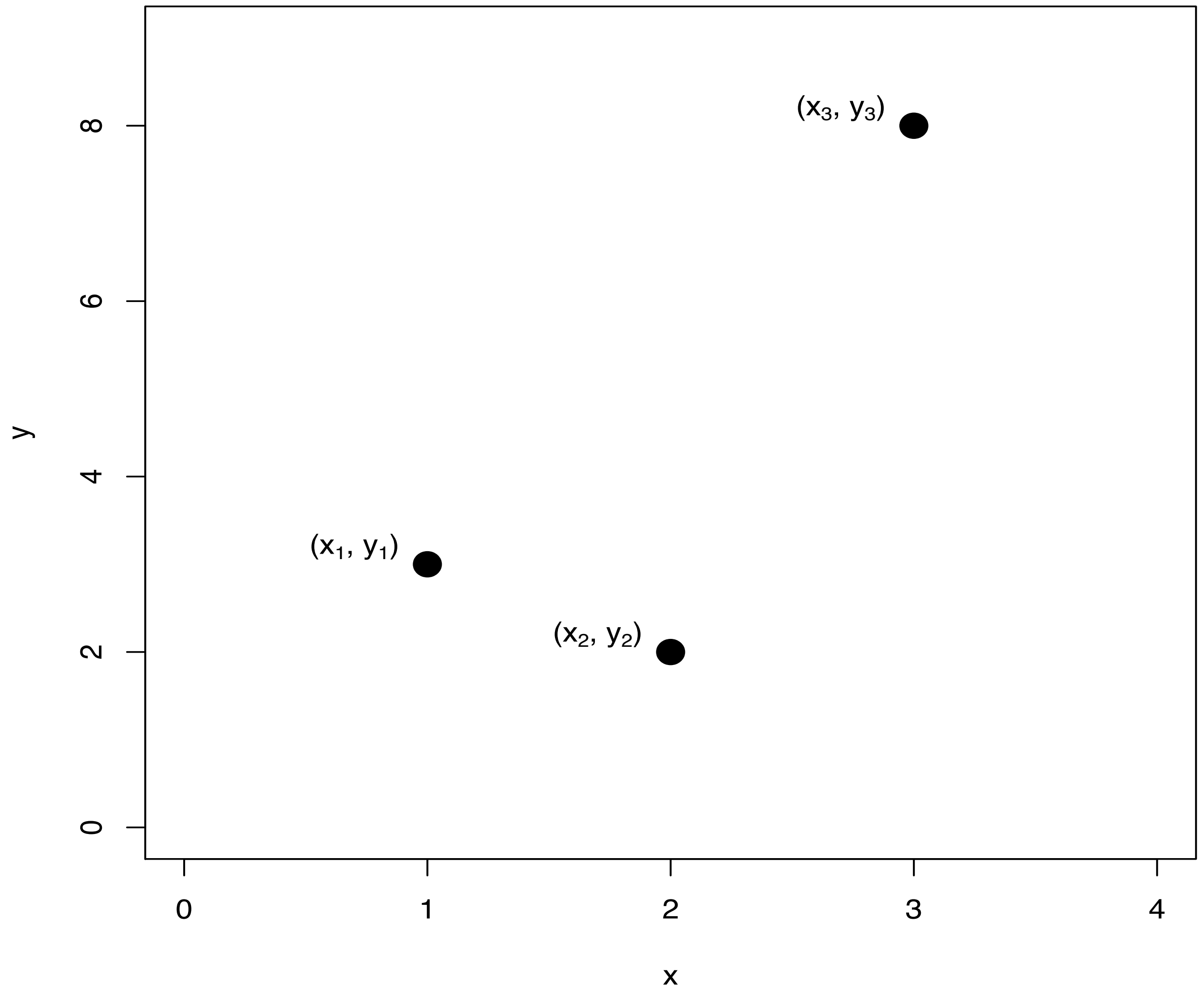
Fitted Values and Residuals

Values on line

$\hat{y}$

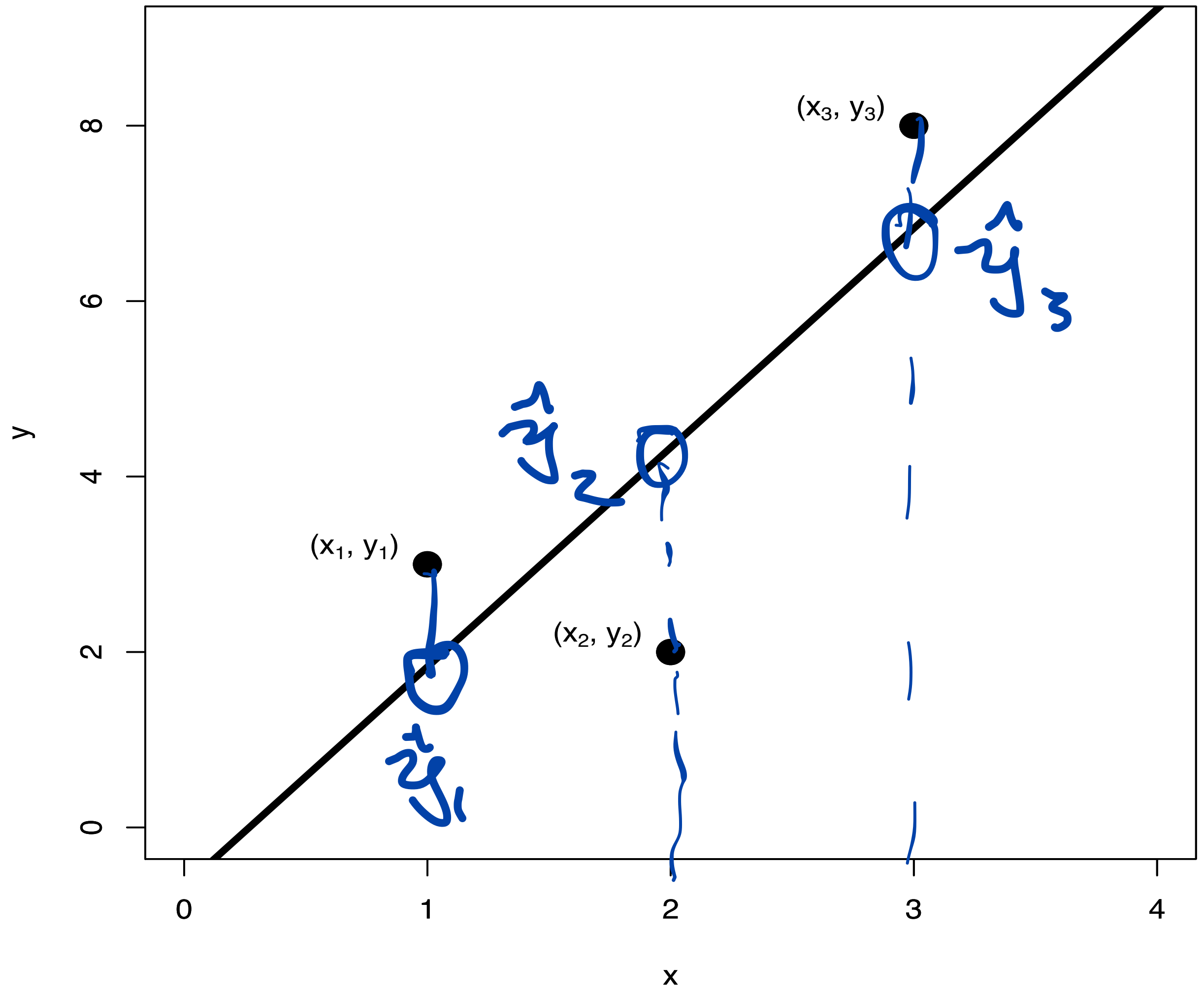
What's left over

$y - \hat{y}$



Data:

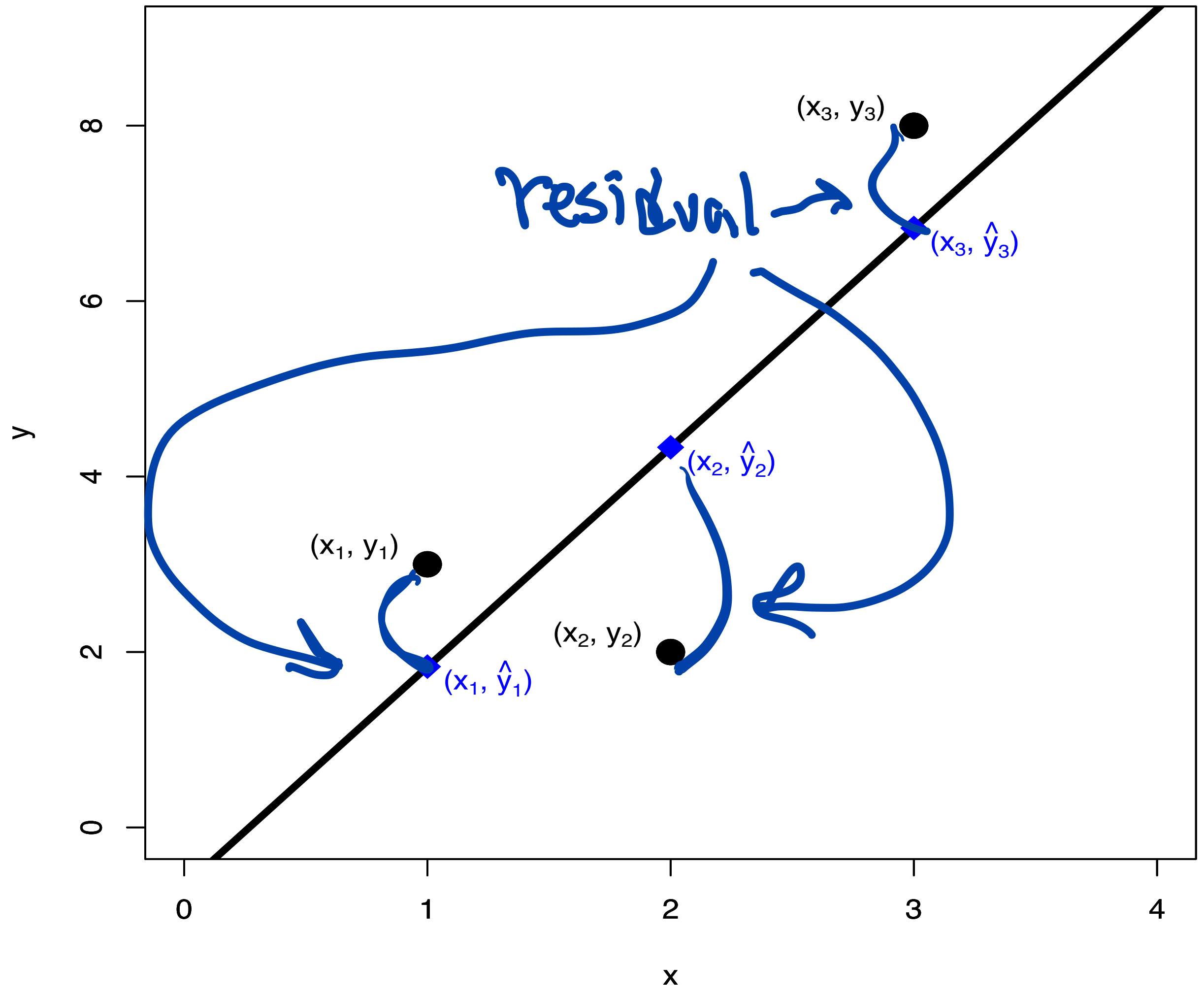
$$(x_i, y_i)$$



Regression Line minimizes the L_2 loss between y_i and $a+bx_i$

$$\min_{a,b} \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

Fitted Values



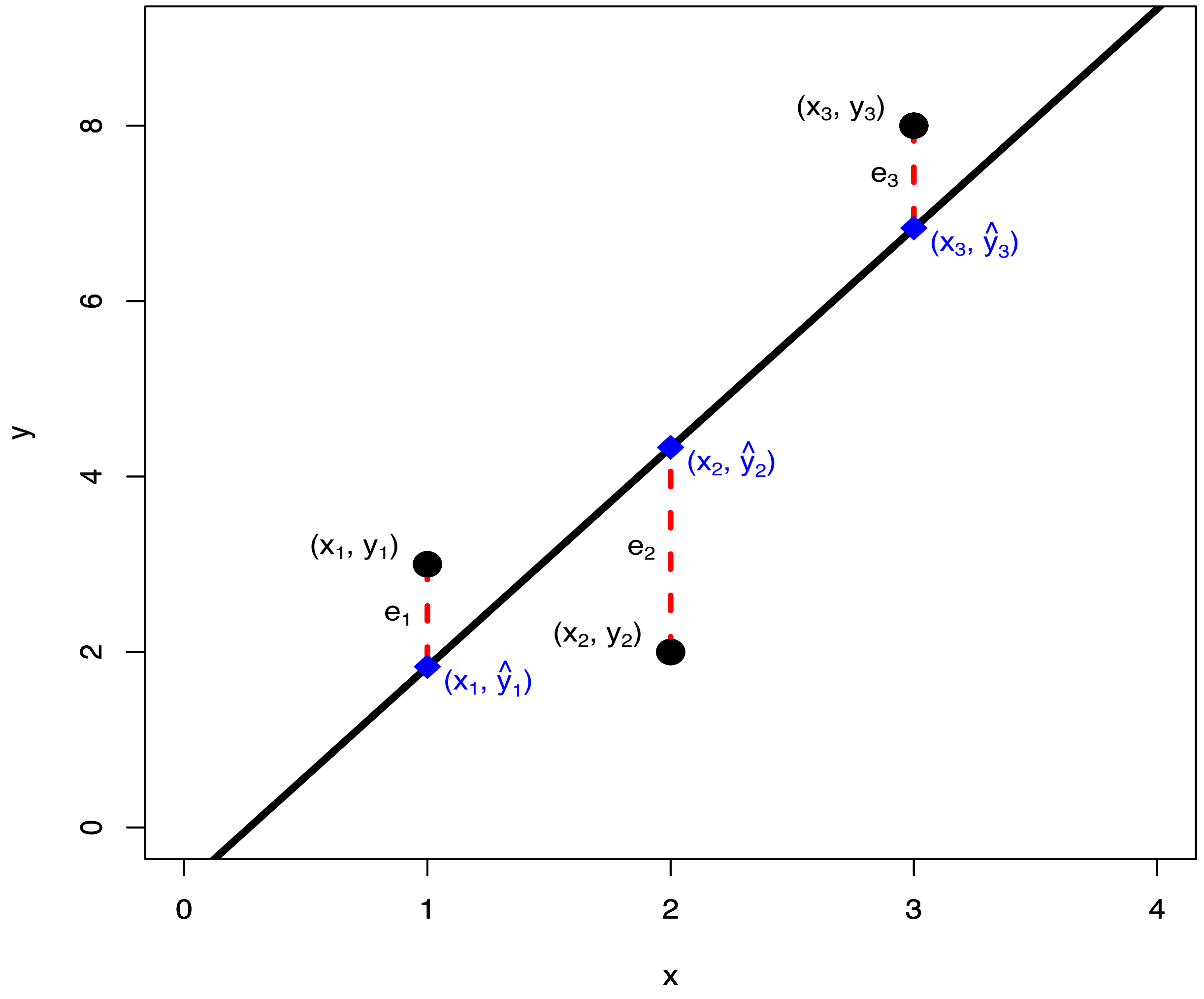
Predictions are points on the line

$$\hat{y} = \hat{a} + \hat{b}x$$

Given an x value, what is the prediction for y?

Errors AKA Residuals

$\Sigma (y_i - (a+bx_i))^2$



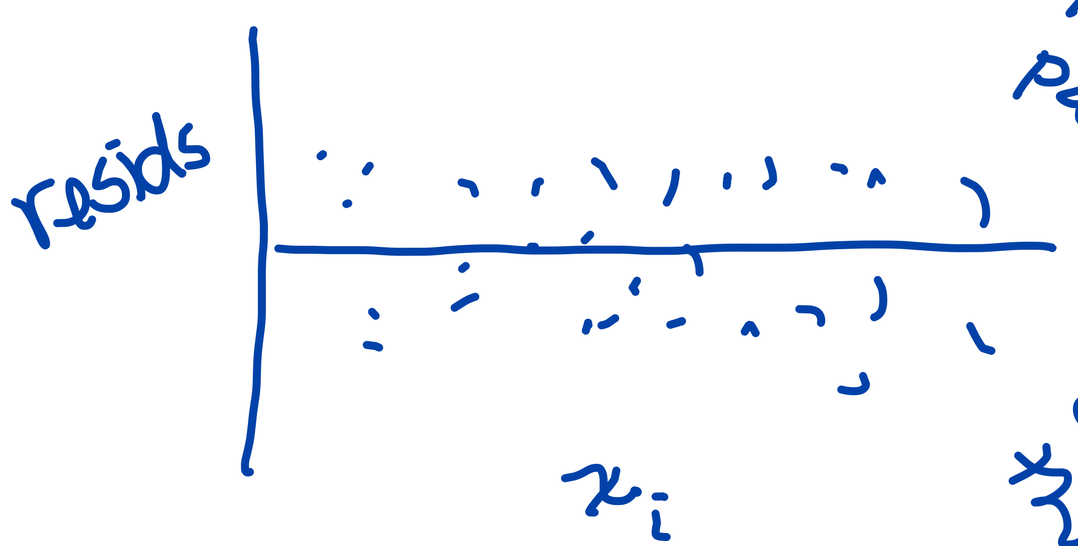
The errors (AKA residuals) in our prediction

$$e_i = y_i - \hat{y}_i$$

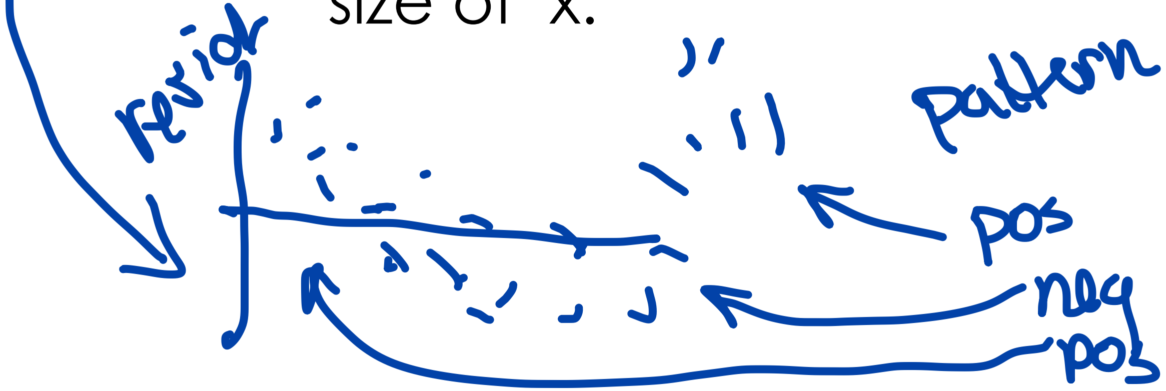
Note that these errors are vertical distances between the line and the points

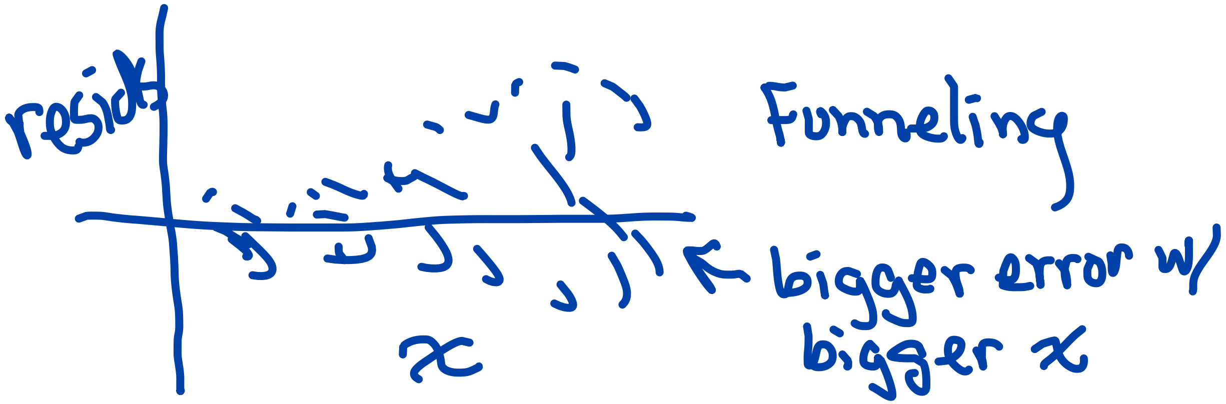
Residual plots

- Plot the pairs (x_i, e_i)
- Plot the pairs $(\hat{y}_i, e_i)$



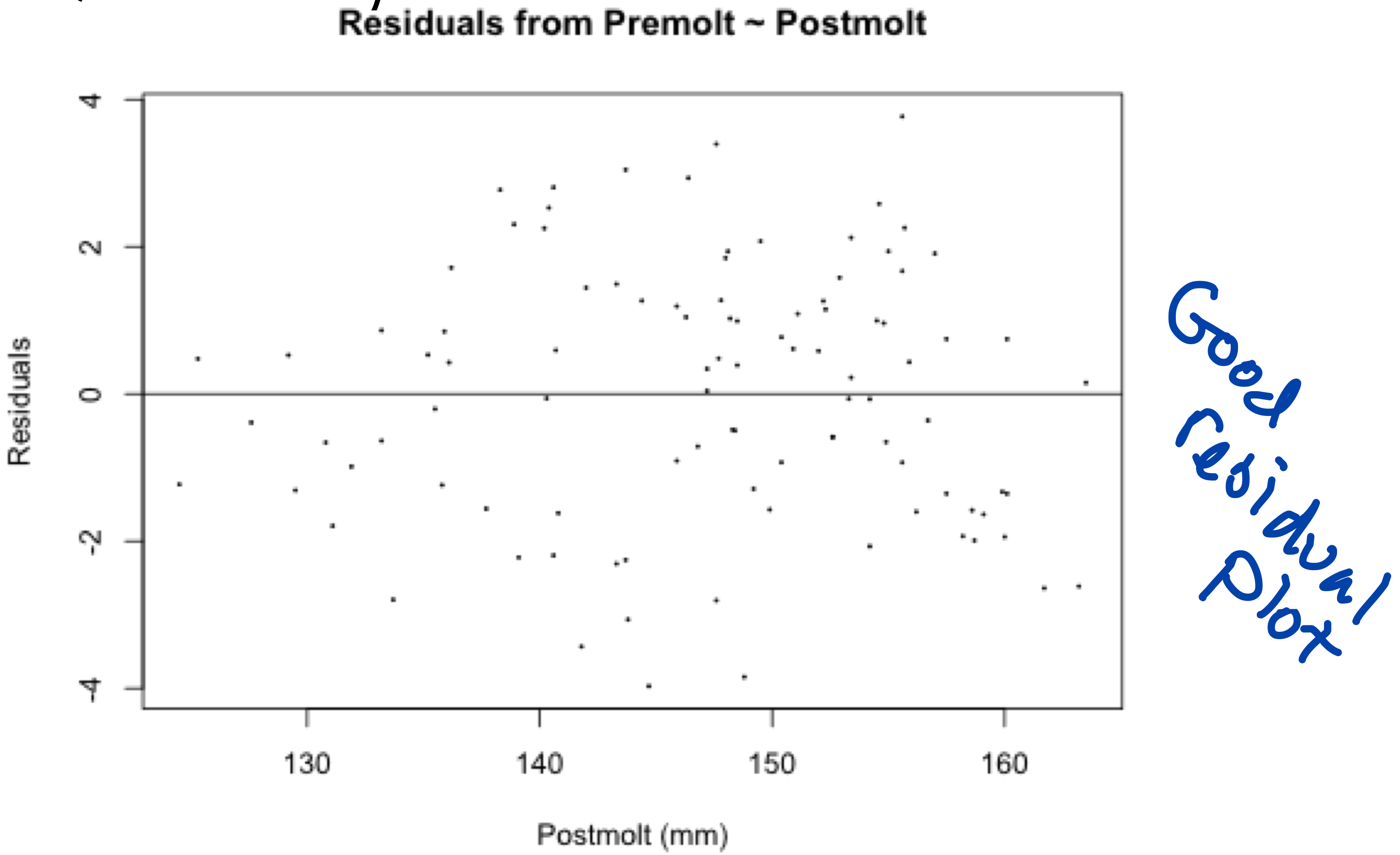
- Look for patterns in the residual plots
  - See no pattern – the relationship is well represented by a line
  - Curve – transformation or additional variable may be needed
  - Funneling – the accuracy of the regression line varies with the size of x.





Residuals

Plot the pairs
(postmolt size, residual)



Variation – Explained and Unexplained

Total Variation. AKA Sum of Squares

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\underbrace{y_i - \hat{y}_i}_{} + \underbrace{\hat{y}_i - \bar{y}}_{})^2$$

Tot SS

$$= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$+ \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

$+$ 2 cross-product ↙ 0

Variation – Explained & Unexplained

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$$

Total
Variation

$$= \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2$$

e_i^2

noise

Unexplained
Variation

Explained
Variation

line

Regression from the Scatter Plot Perspective

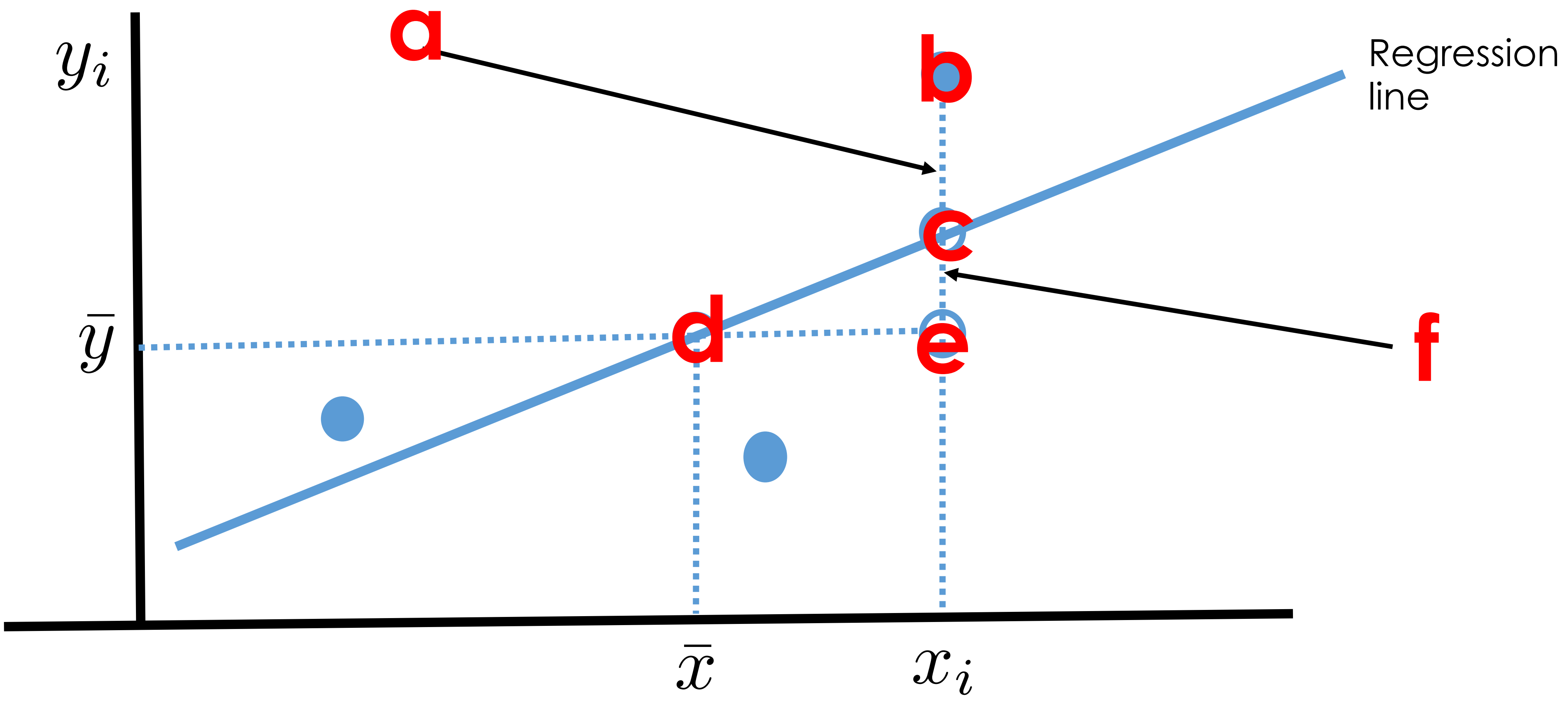
$(x_i, \hat{y}_i)$

(x_i, y_i)

$(x_i, \bar{y})$

$(\bar{x}, \bar{y})$

e_i



Regression from the Scatter Plot Perspective

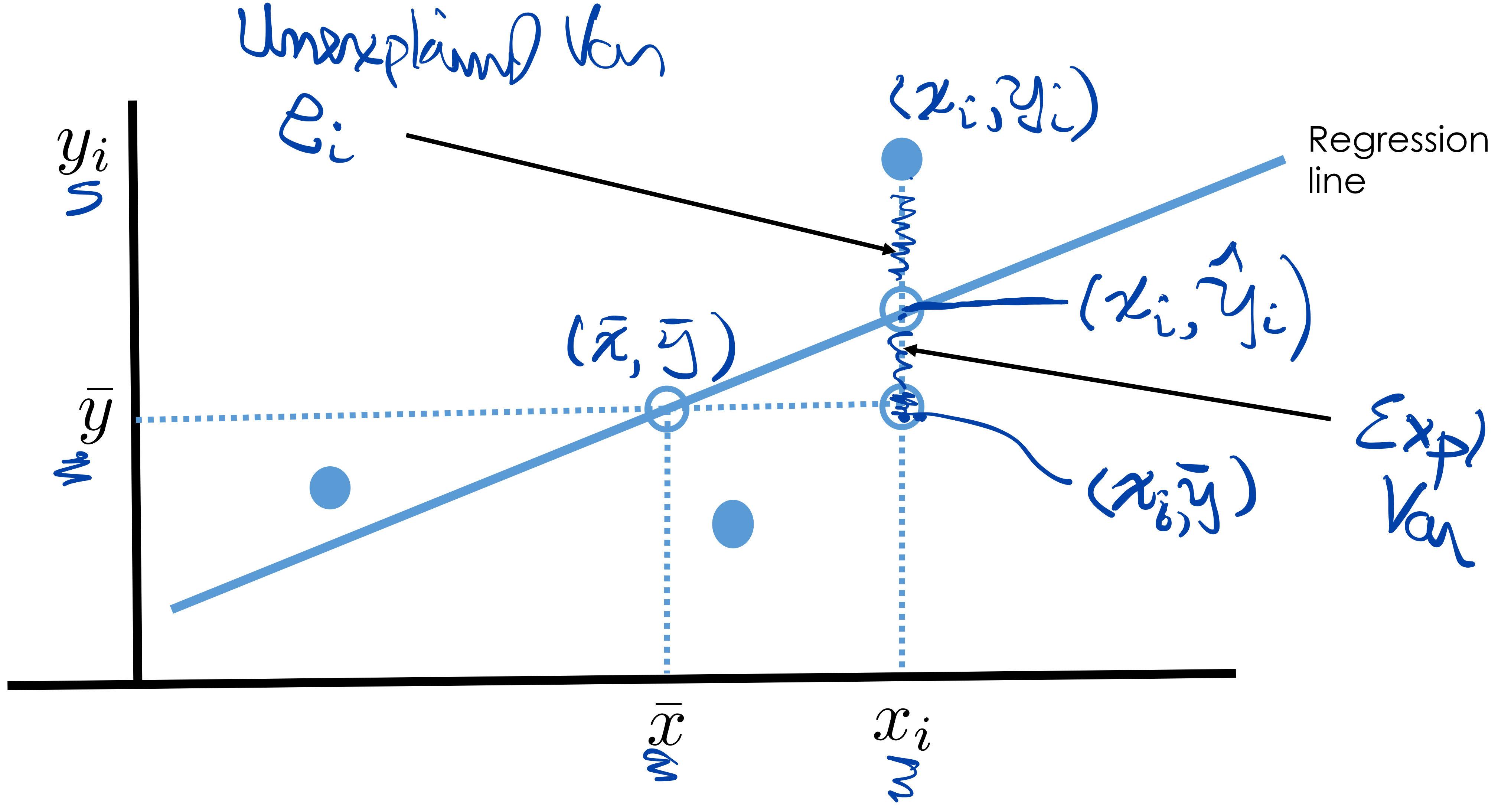
$(x_i, \hat{y}_i)$

(x_i, y_i)

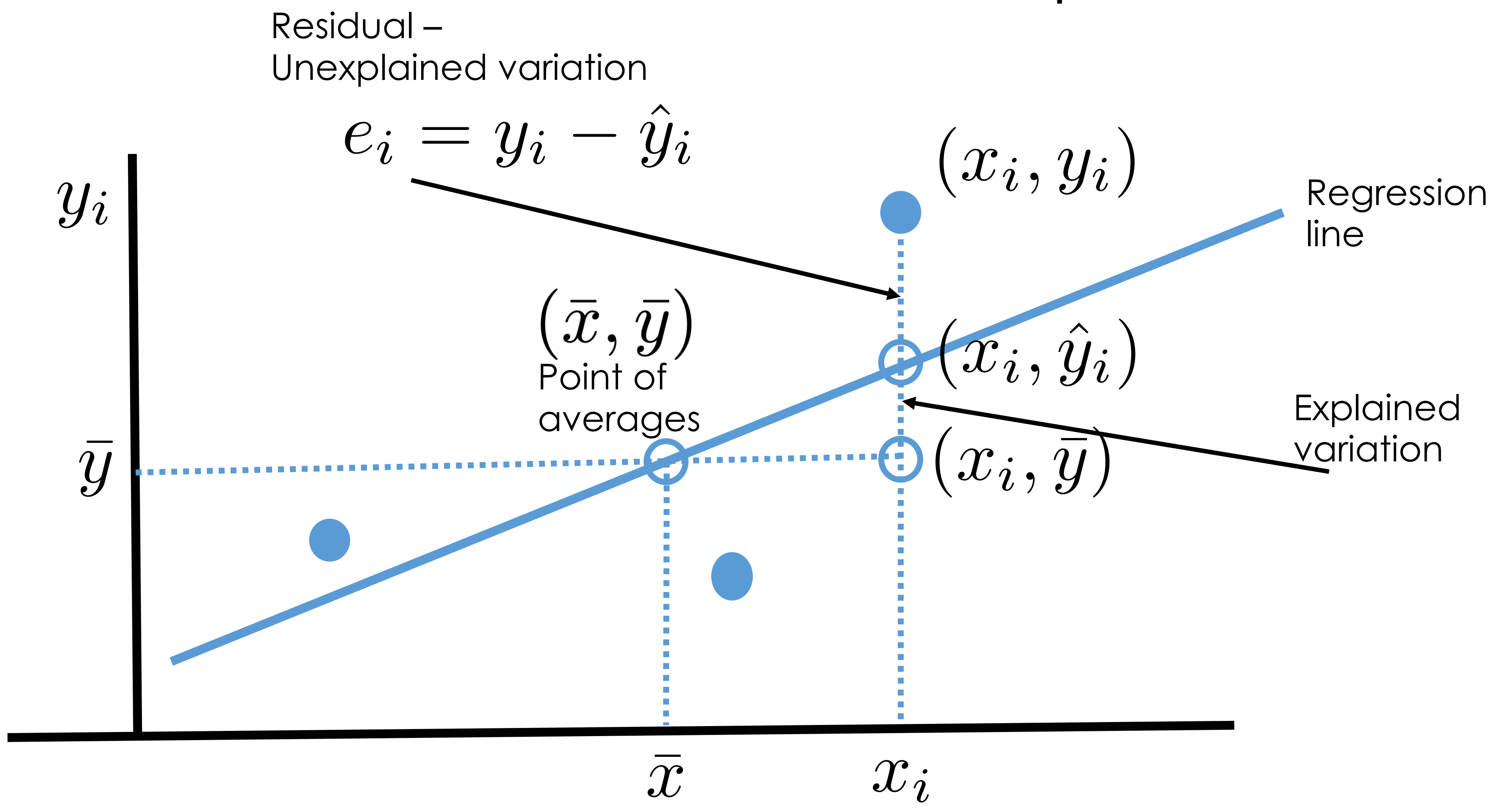
$(x_i, \bar{y})$

$(\bar{x}, \bar{y})$

e_i



Regression from the Scatter Plot Perspective



Regression & Inference

Question: Do 720 5-kg cats produce more heat than 1 3600 kg elephant?

Or, the story of the spherical cat

Kleiber's Equation

- Does a horse produce more heat per day *per kilogram* of body mass than a rat?
- This is a question studied by Kleiber (1947), Clarke (2010)
- Metabolic Rate: kilocalories per day
- Mass in kg
- He measured 19 animals (mouse, dog, cat, goat, man, cow, elephant…)

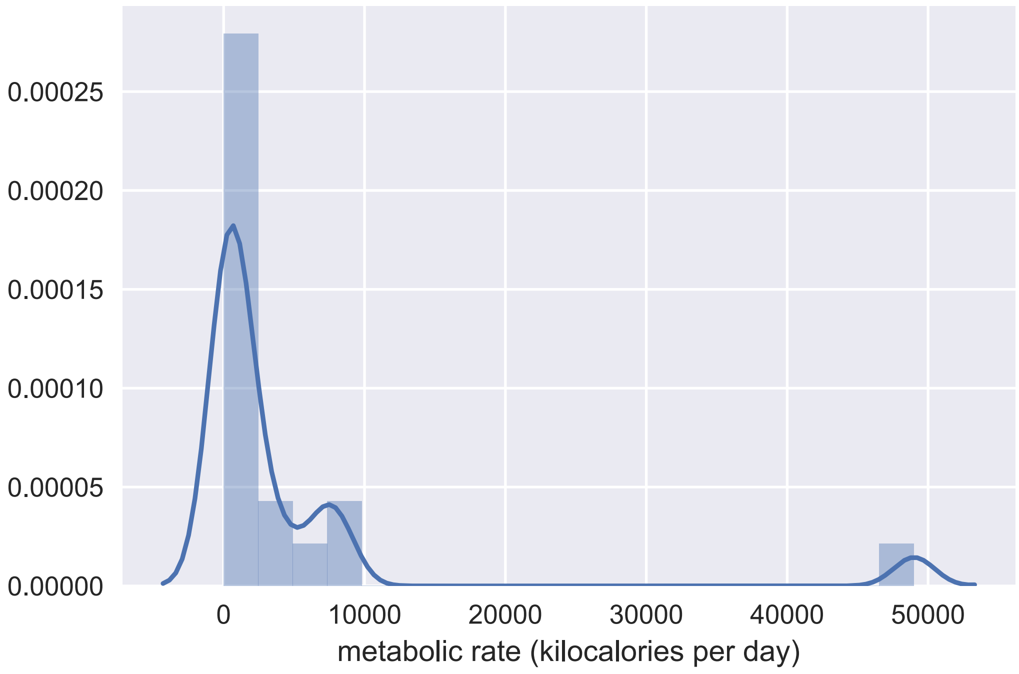
Kleiber's Data

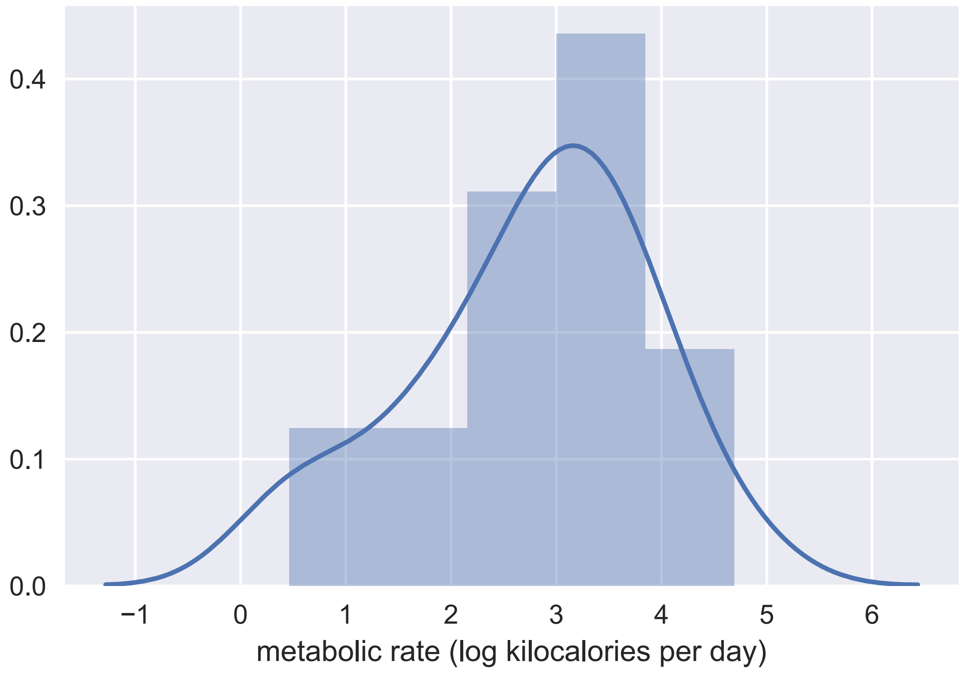
- Population – a typical "mammal"
- Sampling Frame - - an experiment is not possible here
- How were the subjects obtained? From a population, a random sample, or a sample of convenience?

  Sample of convenience

Metabolic Rate is highly skewed

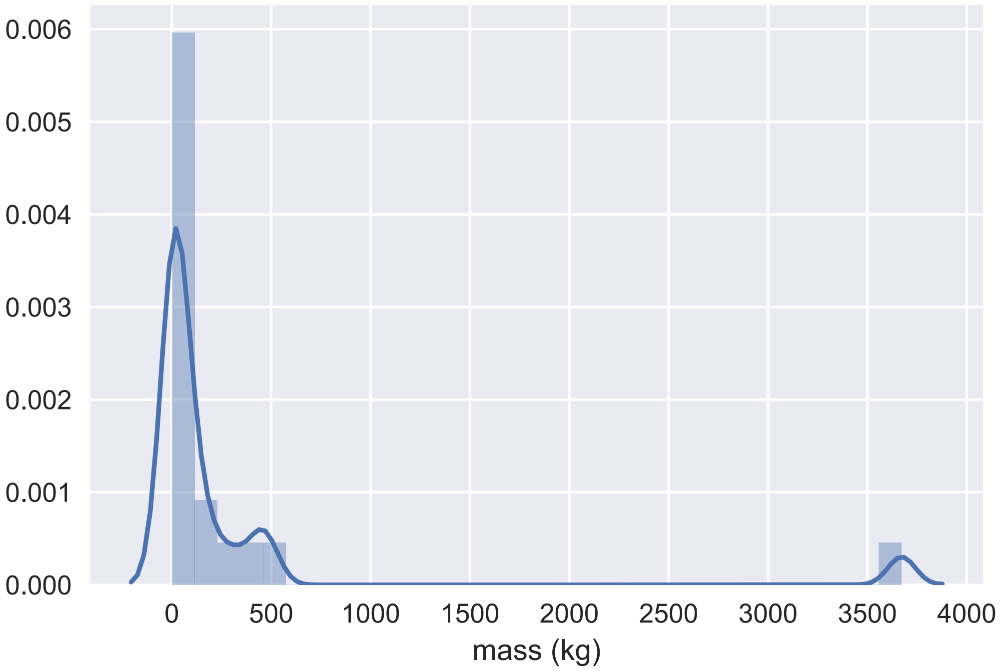
Log Metabolic Rate is less skewed.



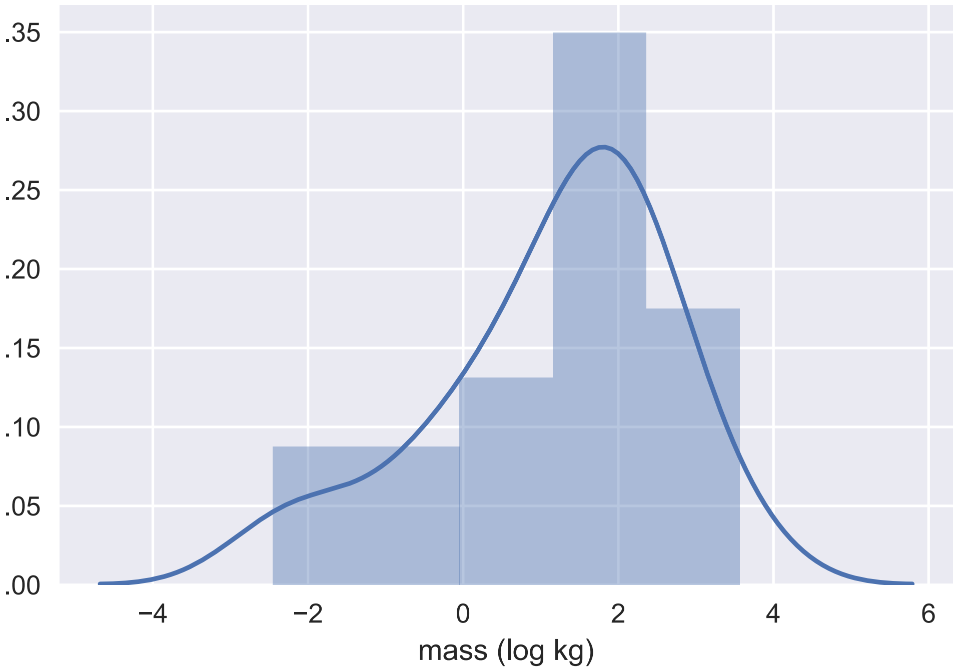


transformation → more stable estimate

Mass is also highly skewed



Log Mass is less skewed.
The skew is in the other direction



How do these two quantities vary together?

Response & Explanatory Variables

- Y is the response variable aka dependent variable
- X is the explanatory variable aka independent variable aka feature

Which is which in our example?

Y - Metabolic Rate

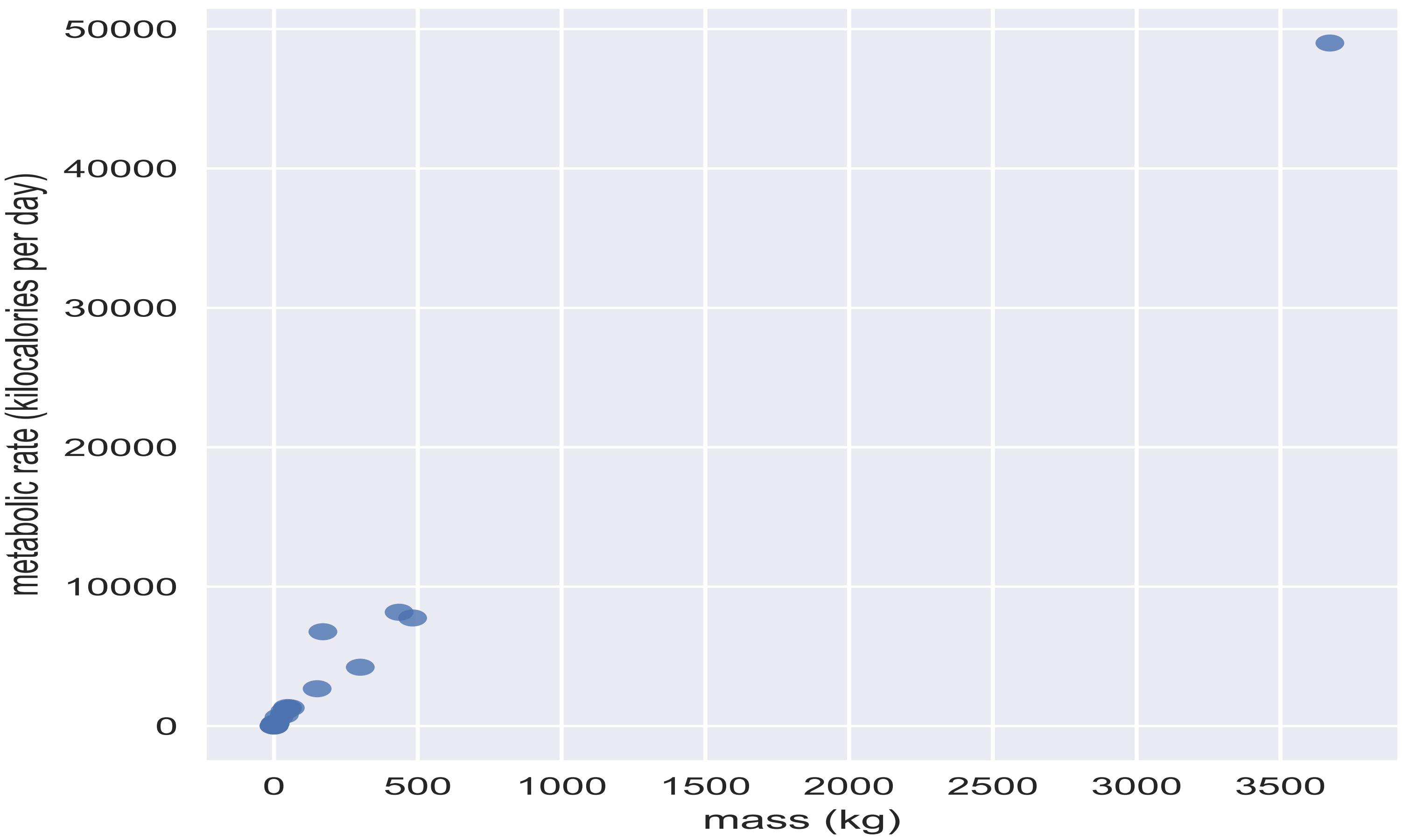
X – Mass

Because Kleiber's question is to explain metabolic rate in terms of mass

Examine the Joint Distribution

The histograms do not give us information about how the two variables vary together

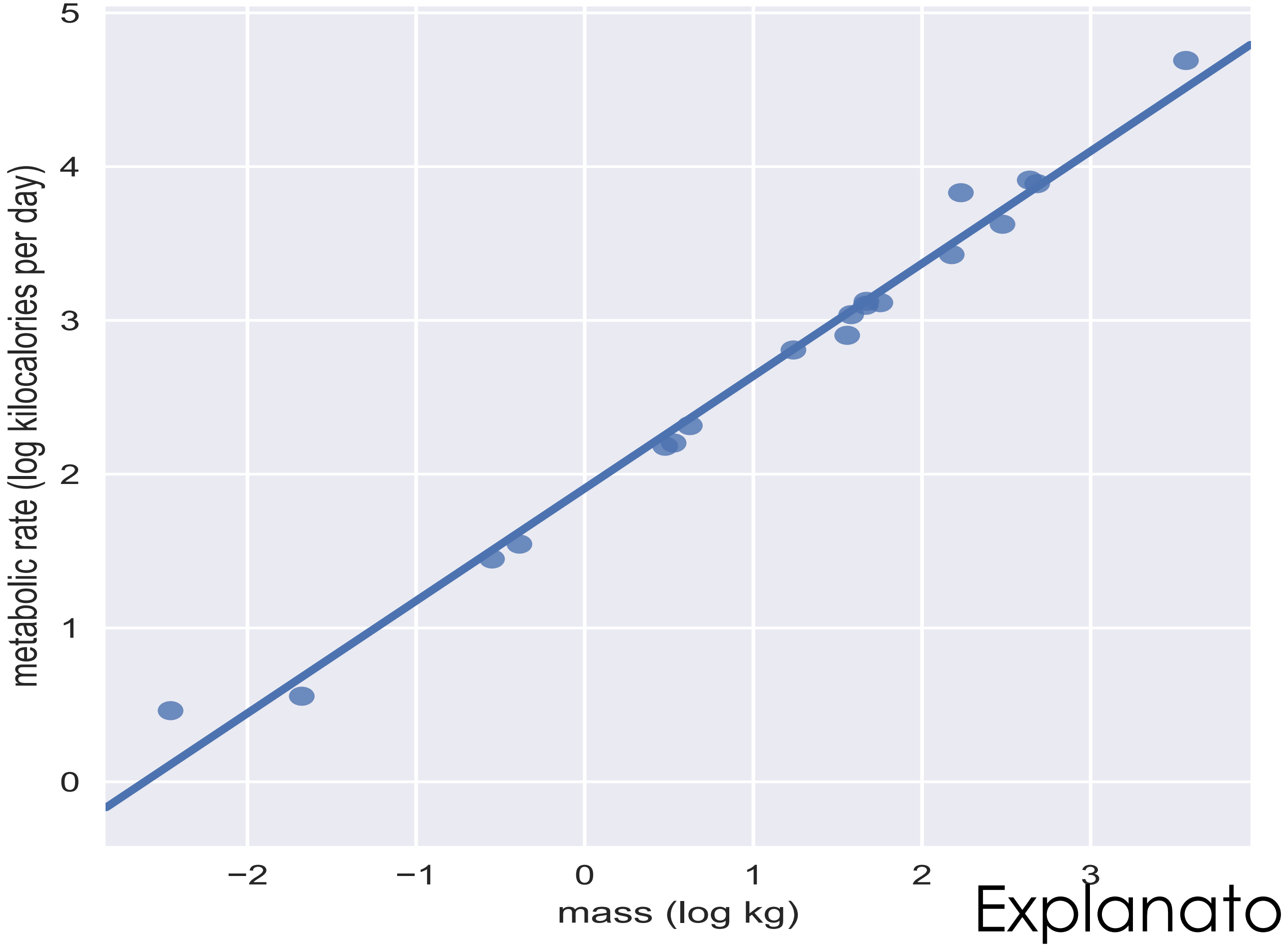
Kleiber's Data



One point makes it difficult to see the relationship between these variables

Deviations of the observed metabolic rate from the regression line

Response



Explanatory

The error about the regression line is the root mean square error loss.
It is like an SD of the regression line.

A Log-Log Relationship

Linear relationship
between log(x) & log(y)

$$\log(y) = a + b\log(x)$$

A Log-Log Relationship

Linear relationship between log(x) & log(y)

$$\log(y) = a + b\log(x)$$

Intercept

Slope

$$y = cx^b$$

Same b as above

We typically use "log" to represent the natural log.
The base does not impact the shape of the relationship.

A Log-Log Relationship - interpretation

$$\log(y) = a + b\log(1.5x)$$

50% increase in x

$$y = c1.5^b x^b$$

corresponds to a
1.5^b % change in y

Log-log relationships are usually expressed in terms of %change in x and y

Method of least squares

Minimize the average squared loss (L_2 loss) when predicting log(rate) from log(mass)

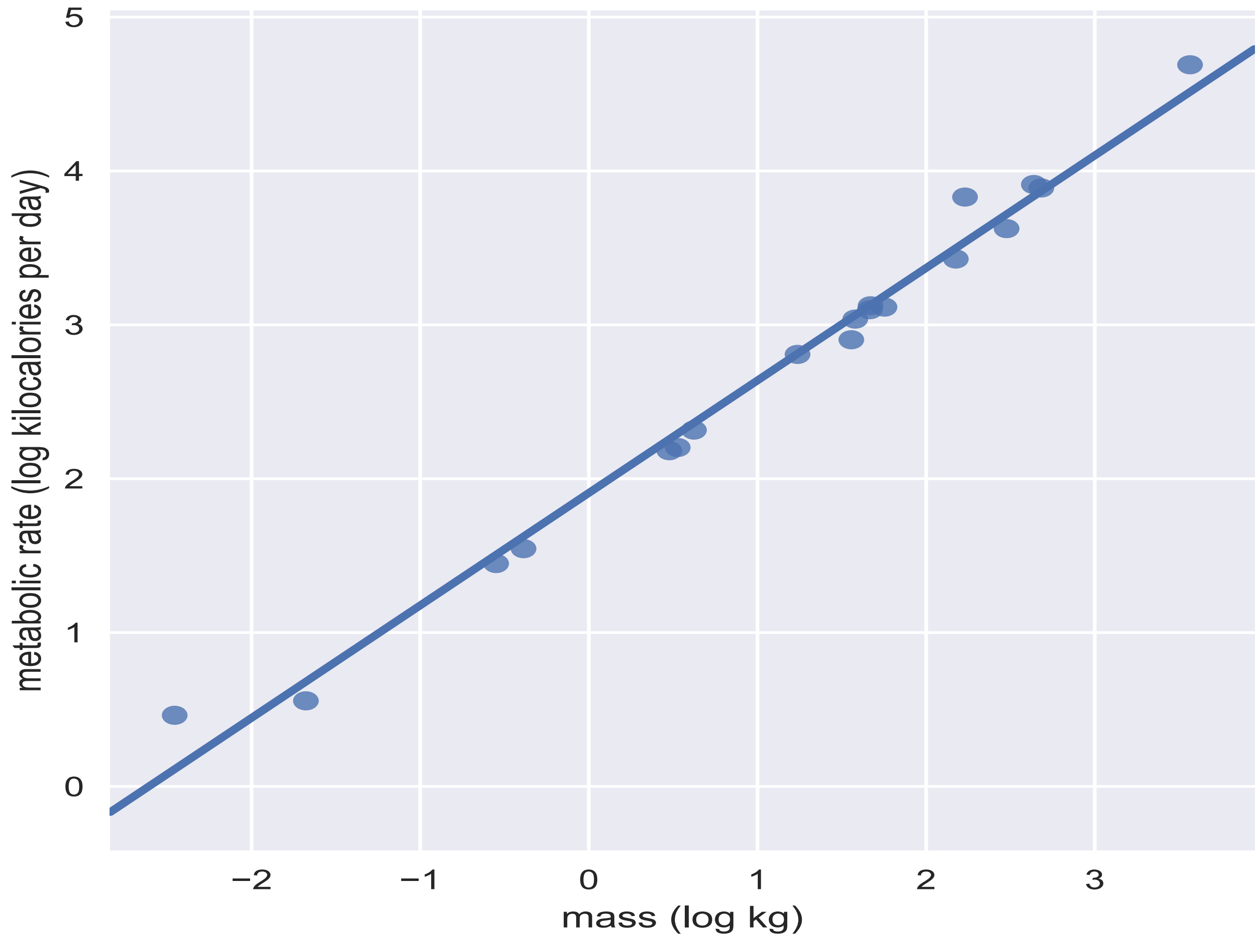
$$\frac{1}{n}\sum(\underbrace{\log(y_i)}_{\tilde{y}_i} - [a + b\underbrace{\log(x_i)}_{\tilde{x}_i}])^2$$

The model is linear in the transformed data

Here we minimize with respect to *a* and *b*.

$$\sum_{i=1}^{n}[\tilde{y}_i - (a+b\tilde{x}_i)]^2$$

Return to our Fitted line



Line has slope 0.75

5 kg

Question: Do 720 cats produce more heat than 1 elephant?

$$5 \times 720 = 3600$$

3600 kg

What does the slope of the line tell us?

$$\log(rate) = a + 0.75\log(mass)$$

Or

$$rate \propto mass^{0.75}$$

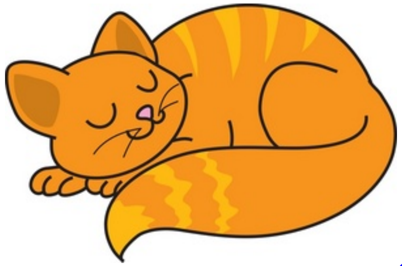
If body mass of elephant is 720 times that of a cat, then metabolic rate is $720^{0.75}$ = 140-fold greater than a cat's

5 kg

Question: Do 720 cats produce more heat than 1 elephant?

YES!
140 cats have the same metabolic rate as 1 elephant

3600 kg



5 kg

Question: Why not just use the values for cat and elephant, rather than fitting a line?

If this relationship holds for mammals in general then we gain in accuracy by using a line fitted to all of the data

3600 kg

5 kg

Question: If we feed our cat enough to gain 3595 kg, will it produce the same heat as an elephant?

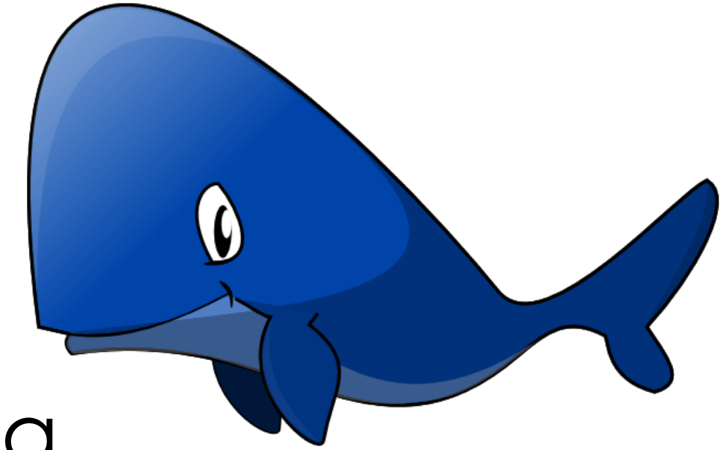
NO!
That's silly!
This is an observational study.
We have observed a relationship between mass and metabolic rate.
It is not a causal relationship.

3600 kg

5 kg

Question: Can we estimate the metabolic rate for a 135,00 kg blue whale using our regression line?

Best not – It would mean extrapolating well beyond the range of the original data and we don't know if the same linear relationship still holds.



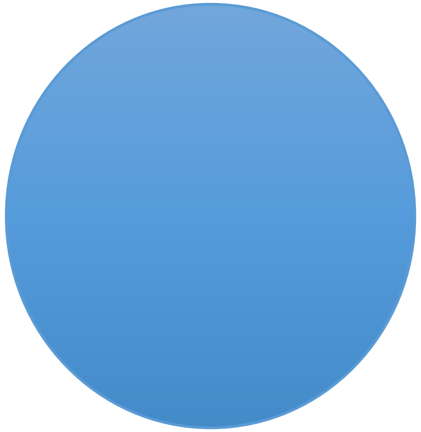
135000 kg

Inference & Bootstrapping

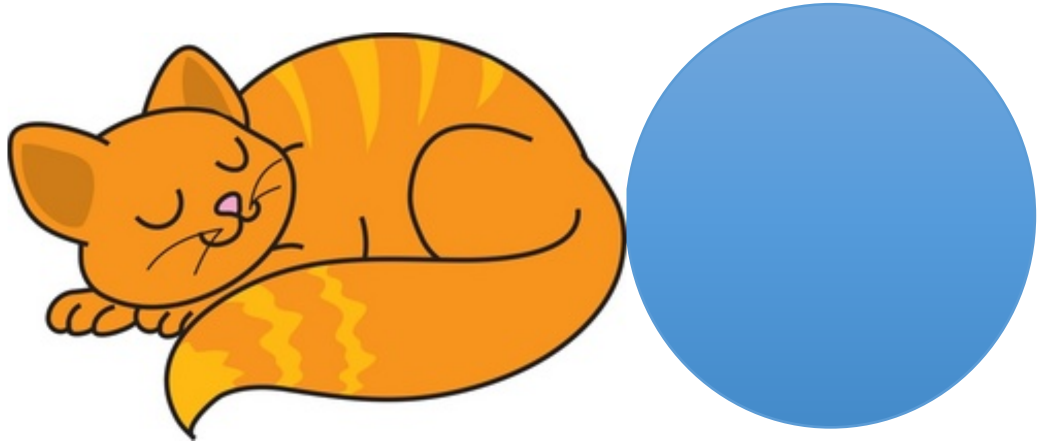
Why is the slope $\frac{3}{4}$?

Why is the slope ¾?

- An alternative theory is that the exponent should be 2/3 because of the relationship between mass and surface area.
- The **spherical cat**:



Explain 2/3



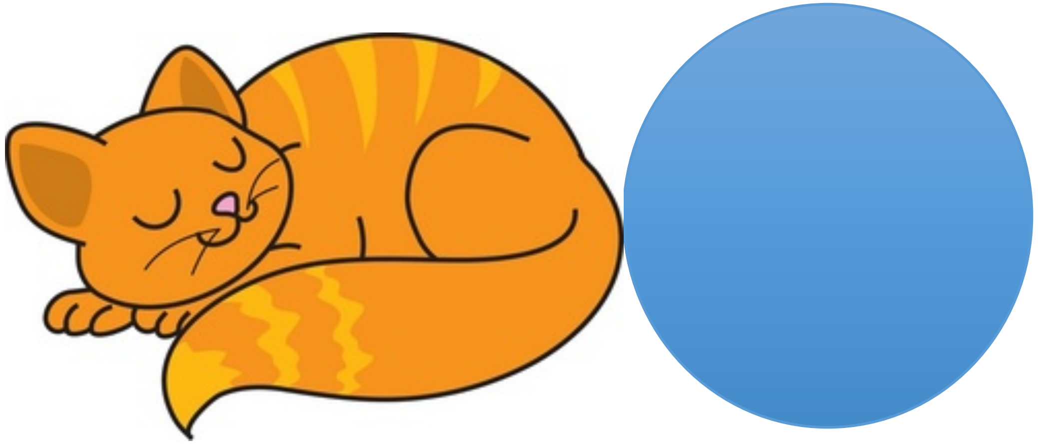
Explain 2/3

$$mass \propto volume \propto diameter^3$$

$$rate \propto surface\ area \propto diameter^2$$

$$rate \propto (diameter^3)^{2/3}$$

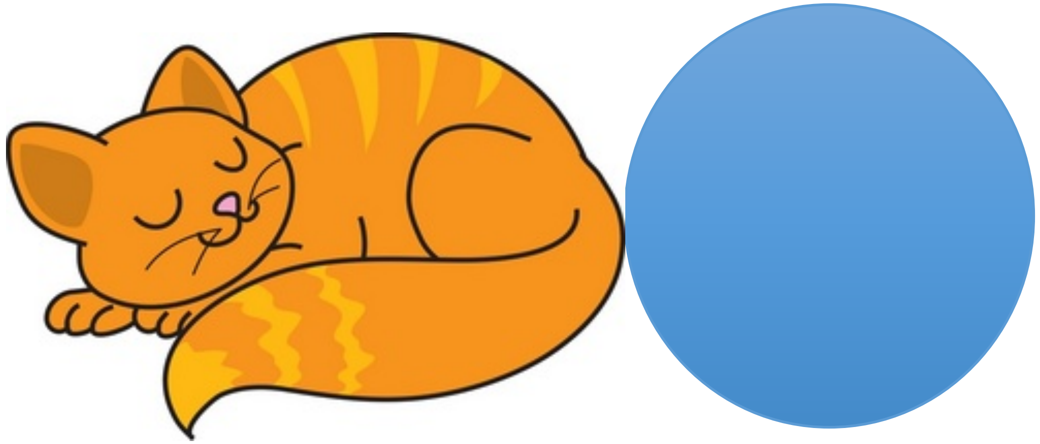
$$rate \propto mass^{2/3}$$



Why isn't the slope 2/3?

Statistical Models are not the same as physical models.

Statistical models can be used to infer
Statistical models can be used to predict

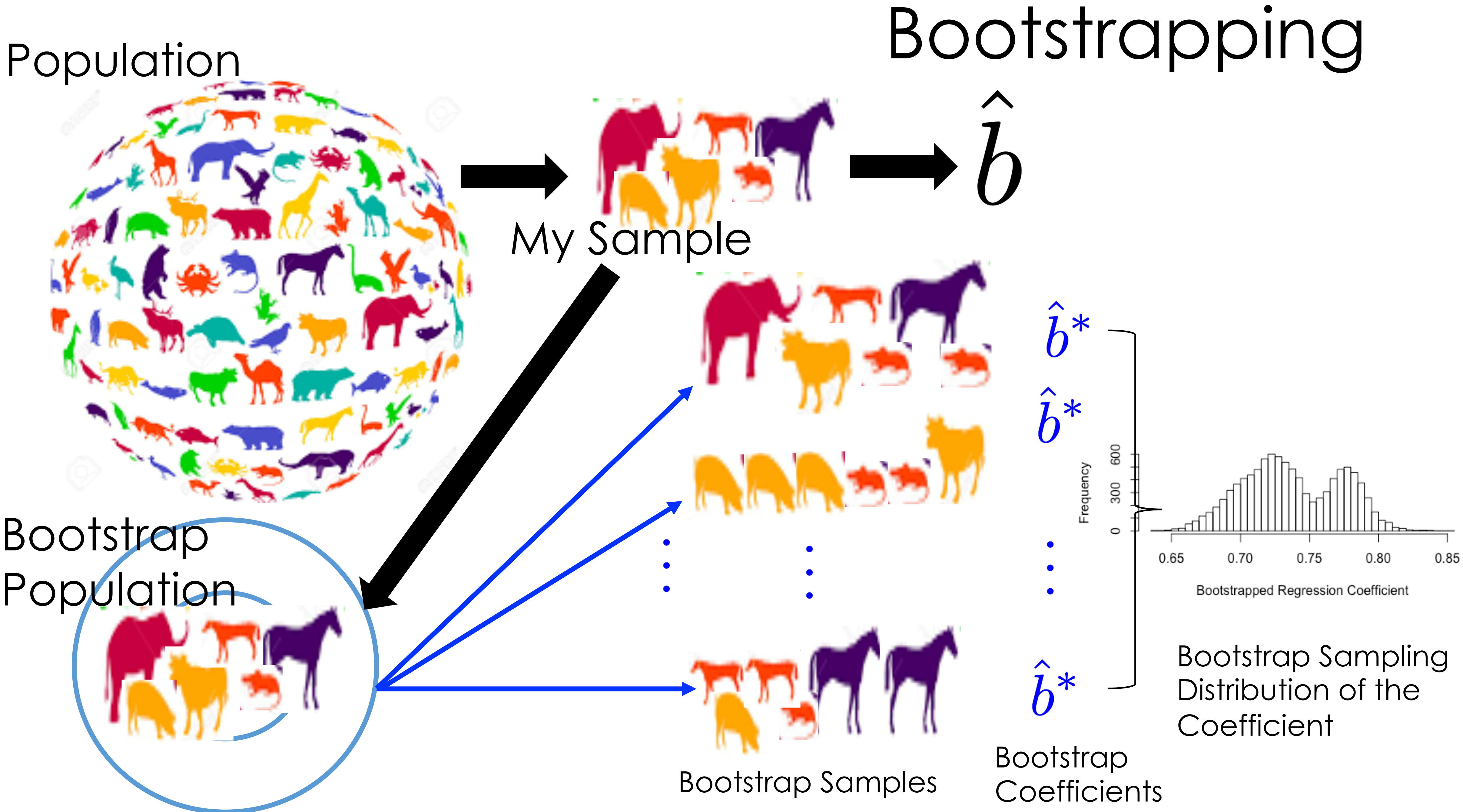


Test the hypothesis: slope = 2/3

Null Hypothesis: true slope is 2/3 AND

the observed difference between fitted coefficient and the true coefficient of 2/3 is *due to chance* in the sampling of the mammals

How to get a sense of this chance?



Bootstrapping
Population
My Sample
$\hat{b}$
Bootstrap Population
$\hat{b}^*$
$\hat{b}^*$
$\hat{b}^*$
Frequency
0
300
600
0.65
0.70
0.75
0.80
0.85
Bootstrapped Regression Coefficient
Bootstrap Sampling Distribution of the Coefficient
Bootstrap Samples
Bootstrap Coefficients

Bootstrapping - Ideas

The sample of mammals should look like the population of mammals

Substitute our sample for the "population"; call it the bootstrap population

Imitate the data generation process by sampling from the bootstrap population; call it the bootstrap sample .

Fit a linear model to the bootstrap sample.

Repeat many times and examine the variability in the bootstrapped coefficient

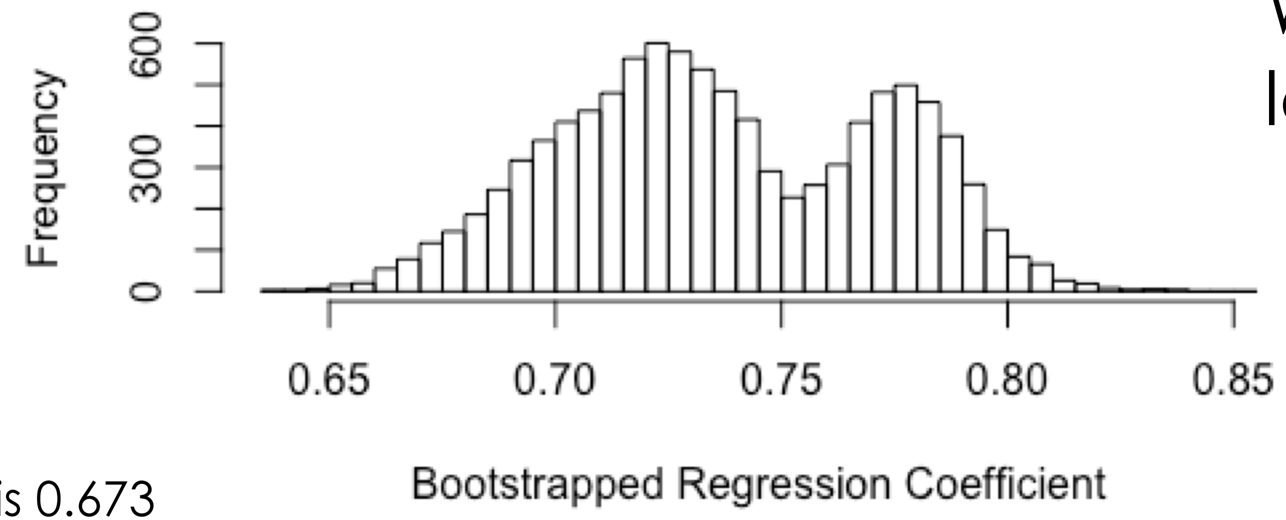
Bootstrap the coefficient

- Bootstrap population: 19 (x,y) pairs of mass and metabolic rate
- Bootstrap sample gives us a bootstrap statistic - the slope of the regression line
- Take 10,000 bootstrap samples from the bootstrap population
- Examine the distribution of bootstrapped coefficients.
- If 2/3 is not within the (0.025, 0.975) percentiles of the bootstrapped distribution of the coefficient, then reject the hypothesis

Bootstrap Sampling Distribution

Based on these percentiles we would reject the hypothesis that the slope is 2/3. But…

Why does it look like that?



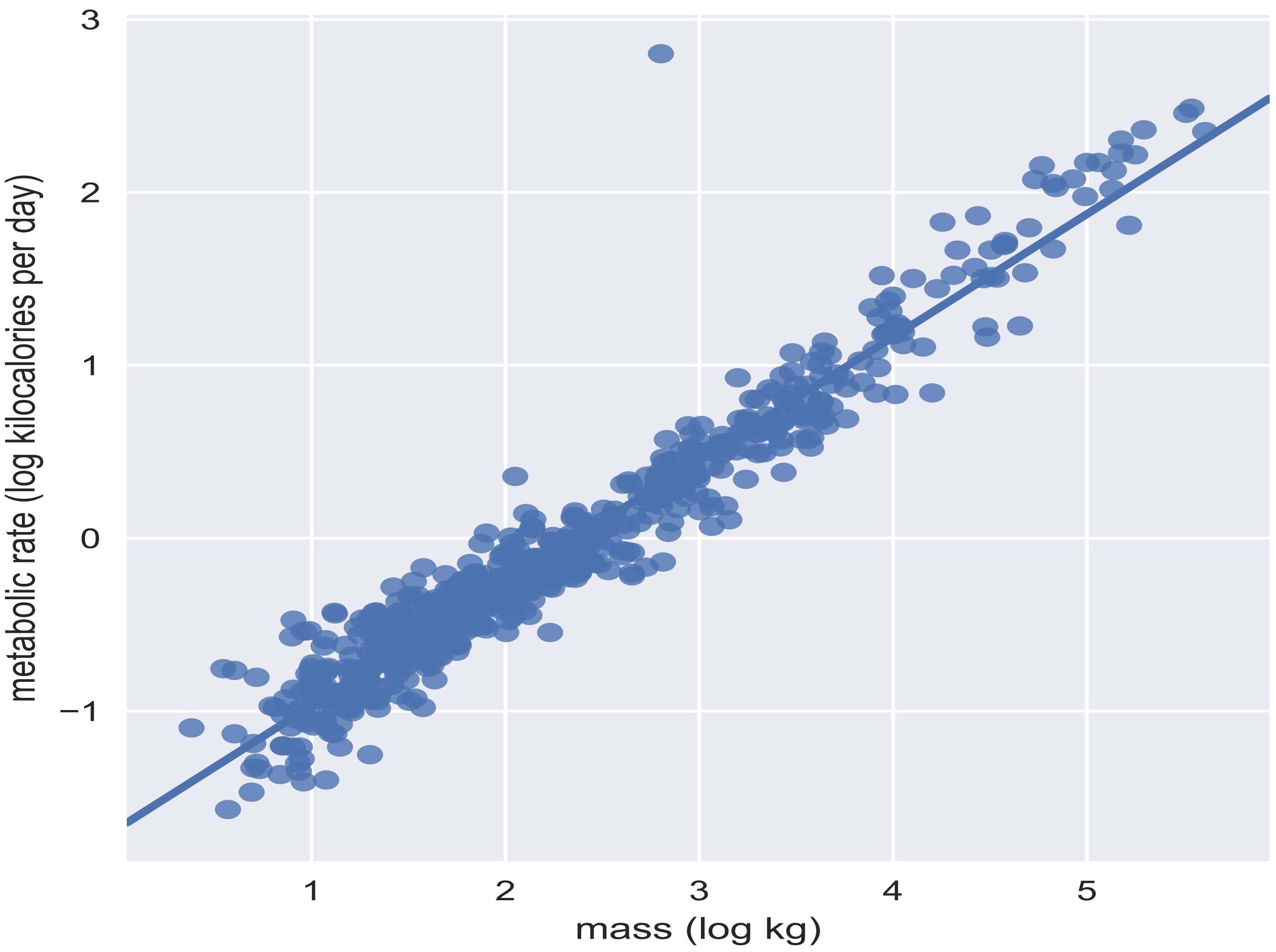
Does this mean that we shouldn't be doing the bootstrap?

Percentiles:
0.025 percentile is 0.673
0.975 percentile is 0.799

Since 0.667 isn't in [0.673, 0.799] then we reject the hypothesis that slope is $\frac{2}{3}$

With only 19 obs the bootstrap may not perform well.

Kleiber's rule studied by Clarke (2010)



Slope remains 3/4

Statistical Models

- To be useful must be an accurate description of data
- Can assist in discovery of physical facts or social phenomena
- Physical models may suggest a particular relationship, which we can fit and test.
- Wish to generalize beyond the subjects studied (even when an entire population is studied)

Summary Points

- With observational studies we cannot make causal claims such as increasing mass by 1 kg leads to a predicted increase in metabolic rate.
- It's not a good idea to extrapolate beyond the range of values observed.

Summary Points

- Even a high correlation, need not mean the relationship is linear.
- Residual plots help us determine the adequacy of the fit.
- Depending on the situation, we may be satisfied with a less complex model that does not fit the data as well, if the size of the errors are tolerable.

Extensions to Simple Linear Regression

- Multiple regression
  - Linear algebra
  - Geometric interpretation
- Qualitative variables
  - explanatory (x)
  - response (y)
- Prediction & Inference
  - Probability Model
  - Bias-Variance tradeoff

Extensions to Simple Linear Regression

- Variable Selection
  - Feature engineering
  - Test-train split
  - Cross-validation
  - Regularization

- Loss – L_2, L_1, and Huber
  - Minimization – L_2
  - Gradient Descent