Week 3

MA5821- Advanced Statistical Methods for Data Scientists

Kazi Arif Hossain
kazi.hossain@jcu.edu.au

AGENDA

- Week 2 recap
- Week 2 Workbook
- Week 3 Content

DIFFERENT TYPES OF REGRESSION

- **Stepwise Regression**: is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure.
  - Forward Selection
  - Backward Selection
  - Bidirectional

- **Ridge regression** is a way to create a parsimonious model (use L2 regularization) when the number of predictor variables in a set exceeds the number of observations, or when a data set has multicollinearity (correlations between predictor variables)

- **Lasso regression** analysis is a shrinkage and variable selection method for linear regression models. The goal of lasso regression is to obtain the subset of predictors that minimizes prediction error for a quantitative response variable. The lasso does this by imposing a constraint on the model parameters that causes regression coefficients for some variables to shrink toward zero

- **Ecological regression** is a statistical technique used especially in political science and history to estimate group voting behaviour from aggregate data

- **Bayesian linear regression** allows a fairly natural mechanism to survive insufficient data, or poor distributed data. It allows to put a prior on the coefficients and on the noise so that in the absence of data, the priors can take over.

- **Quantile regression** is the extension of linear regression and we use it when the conditions of linear regression are not applicable (eg, fails the test of normality)

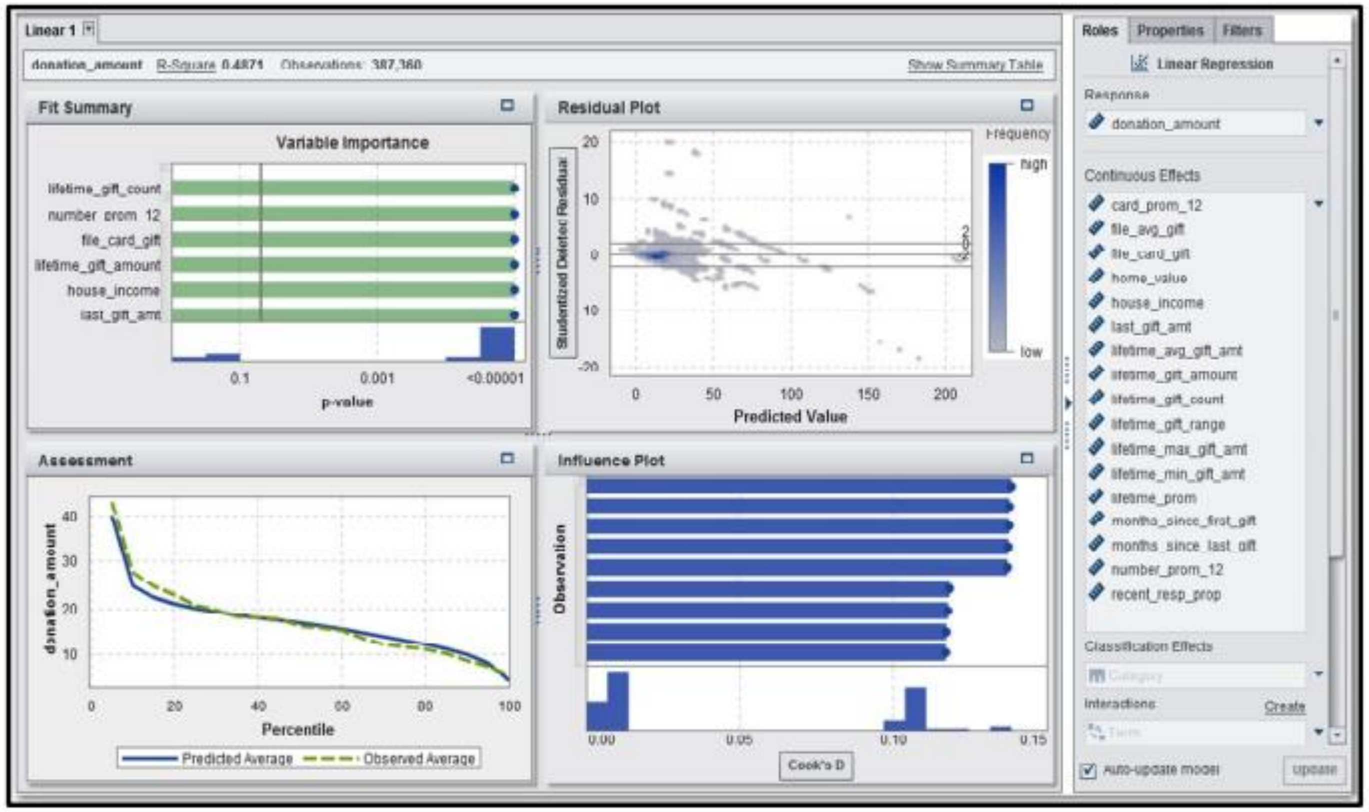
WEEK 2 RECAP: ASSUMPTIONS FOR LINEAR REGRESSION

- Linearity
- Multivariate normality
- No or little multicollinearity
- Homoscedasticity (same variance)
- Independent of errors / No auto-correlation

The value of R-square is always between 0 and 1, where 0 means that the model does not model explain any variability in the target variable (Y) and 1 meaning it explains full variability in the target variable.

In shrinkage methods we don't actually select variables explicitly but rather we fit a model containing **all p** predictors using a technique that constrains or regularizes the coefficient estimates that shrinks the coefficient estimates towards zero relative to the least squares estimates.

SAS REGRESSION ANALYSIS



- **Fit Summary** - displays how significant the effect variables are to the response variable.
- **Residual Plot** - displays the difference between the predicted and the actual data.
- **Assessment** - displays the values for the observed response along with the model's predicted response.
- **Influence Plot** - displays the observations that might influence the overall analysis.

SAS REGRESSION ANALYSIS

Linear Regression: Summary Table

- Overall ANOVA
- Dimensions
- Fit Statistics
- Model ANOVA
- Type III Test
- Parameter Estimates

Overall ANOVA | Dimensions | Fit Statistics | Model ANOVA | Type III Test | Parameter Estimates

Source	Deg Freedom	Sum of Squares	Mean Square	F Value	Pr > F	R-Square
Model	16	28946917	1809182	22992.56	<0.0001	0.487115
Error	387343	30478299	78.68555	.	.	.
Corrected Total	387359	59425216	.	.	.	.

SAS REGRESSION ANALYSIS

Linear Regression: Parameter Estimates

The *Parameter Estimates* tab displays the parameter estimates (coefficients) of each model effect and their associated statistics.

Overall ANOVA | Dimensions | Fit Statistics | Model ANOVA | Type III Test | Parameter Estimates

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	11.32107	0.142154	79.63963	<0.0001
card_prom_12	0.01774	0.012788	1.387177	0.1654
file_avg_gift	0.348892	0.297737	1.171811	0.2413
file_card_gift	0.103976	0.008109	12.8225	<0.0001
home_value	0.000081	0.00002	4.123035	<0.0001
house_income	0.002275	0.000117	19.38853	<0.0001
last_gift_amt	0.527597	0.002207	239.033	<0.0001
lifetime_avg_gif...	-0.23445	0.297723	-0.78749	0.431
lifetime_gift_a...	0.028689	0.000269	106.4795	<0.0001
lifetime_gift_co...	-0.26844	0.004924	-54.5134	<0.0001
lifetime_gift_ra...	0.332109	0.004325	76.78409	<0.0001
lifetime_max_g...	-0.35113	0.004526	-77.5837	<0.0001
lifetime_min_gi...	0	.	.	.
lifetime_prom	-0.05038	0.00231	-21.8071	<0.0001
months_since_...	-0.02383	0.001032	-23.0945	<0.0001
months_since_...	0.089265	0.004405	20.26551	<0.0001
number_prom_...	0.059703	0.005615	10.63353	<0.0001
recent_resp_pr...	-10.8921	0.162732	-66.9325	<0.0001

predicted **donation_amount** = 11.32107 + 0.01774(**card_prom_12**) + 0.348892(**file_avg_gift**) + 0.103976(**file_card_gift**) + …

Linear Regression: Overall ANOVA

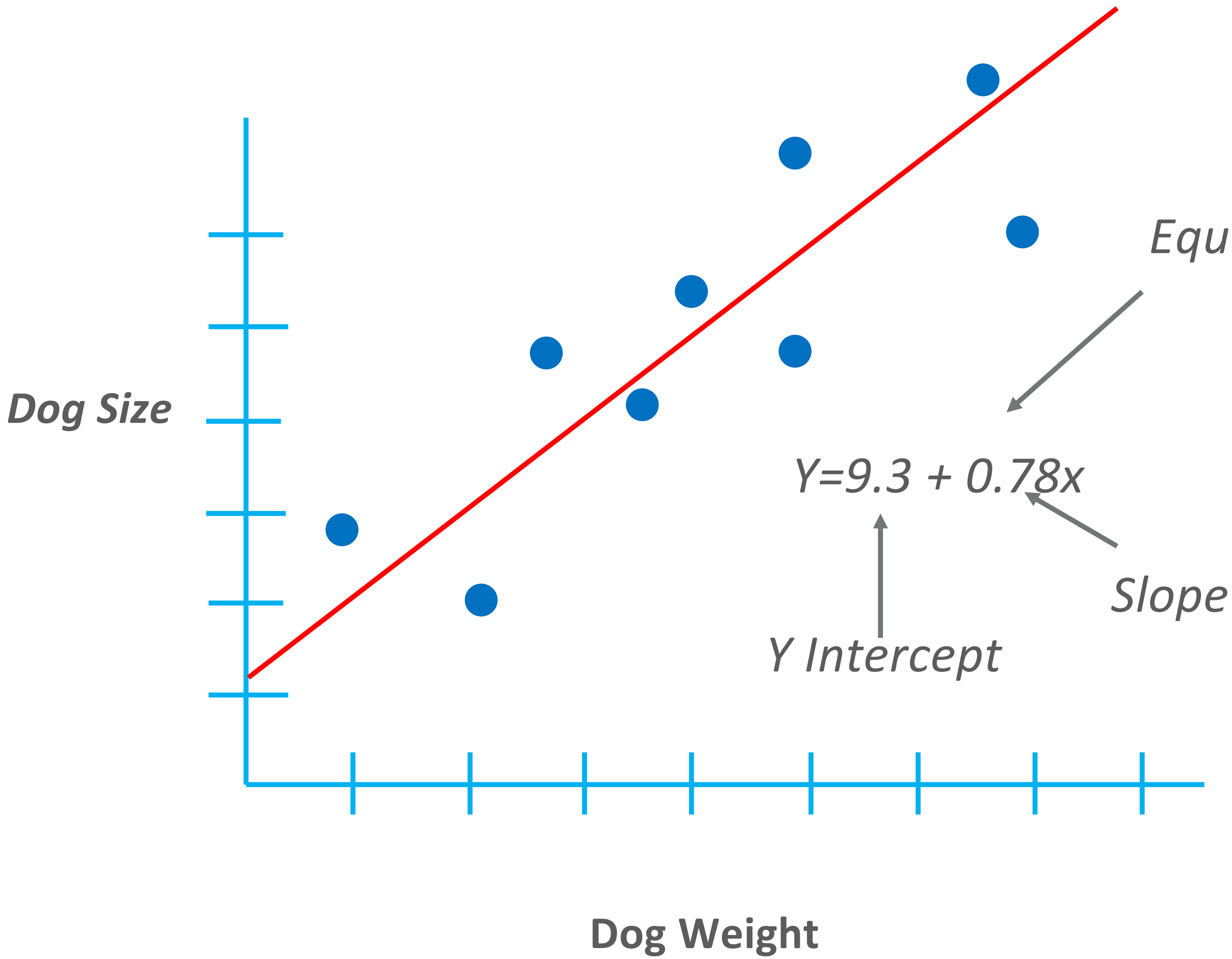
The *Overall ANOVA* tab provides information about how well the model fits the data.

- Source – source of the variation in the data
- Deg Freedom – degrees of freedom associated with the source
- Sum of Squares – sum of squared errors of prediction
- Mean Square = Sum of Squares ÷ Deg Freedom
- F Value = Model Mean Square ÷ Error Mean Square
- Pr > F – *p*-value
- R-Square – the proportion of variation in the response variable explained by the factors in the model

Overall ANOVA | Dimensions | Fit Statistics | Model ANOVA | Type III Test | Parameter Estimates

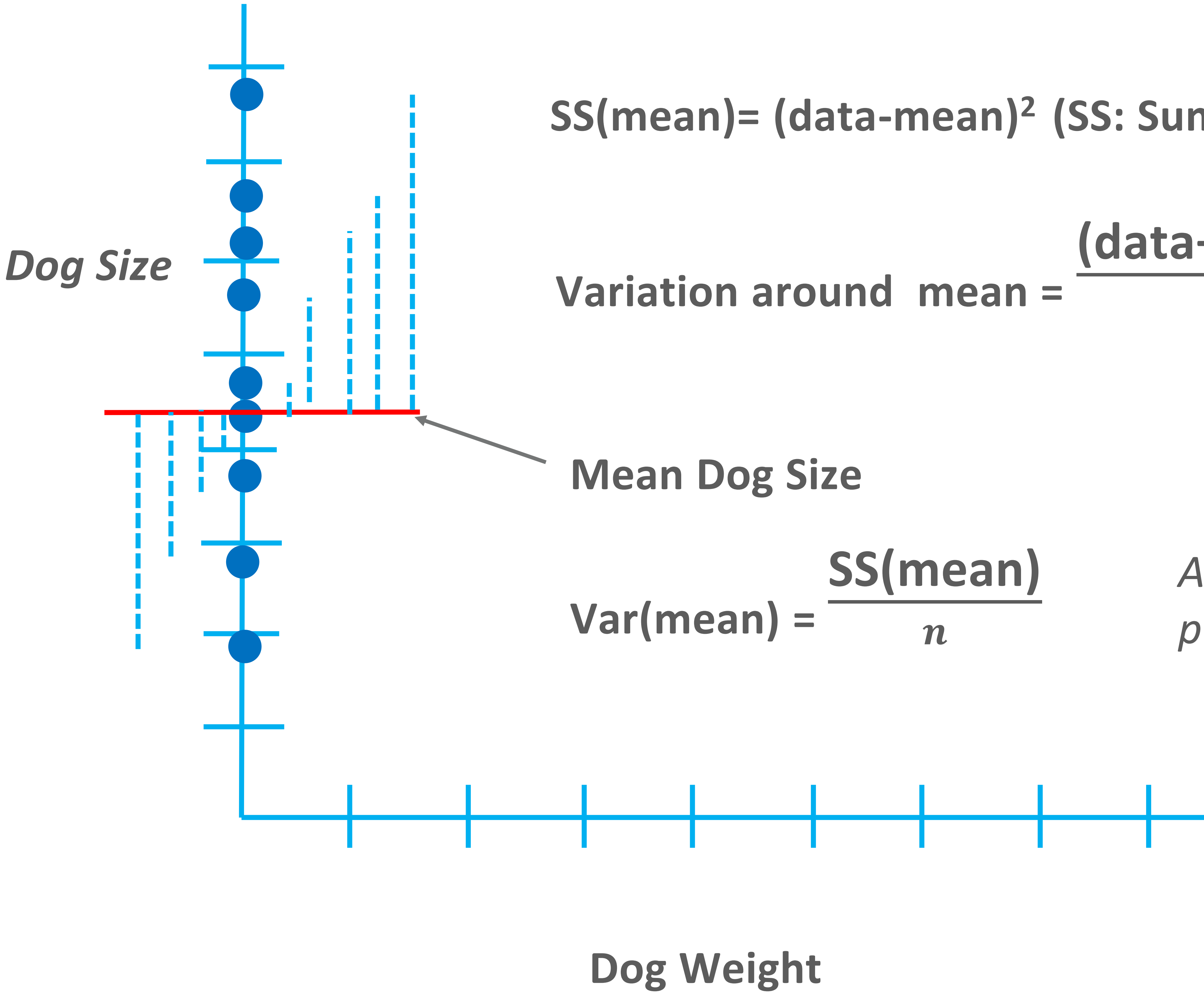
Source	Deg Freedom	Sum of Squares	Mean Square	F Value	Pr > F	R-Square
Model	16	28946917	1809182	22992.56	<0.0001	0.487115
Error	387343	30478299	78.68555	.	.	.
Corrected Total	387359	59425216	.	.	.	.

WEEK 2 RECAP: LINEAR REGRESSION



Since the slope is not 0, it means knowing a dog's weight we can guess about that dog's size

How good is that guess ?? **R^2** will help us



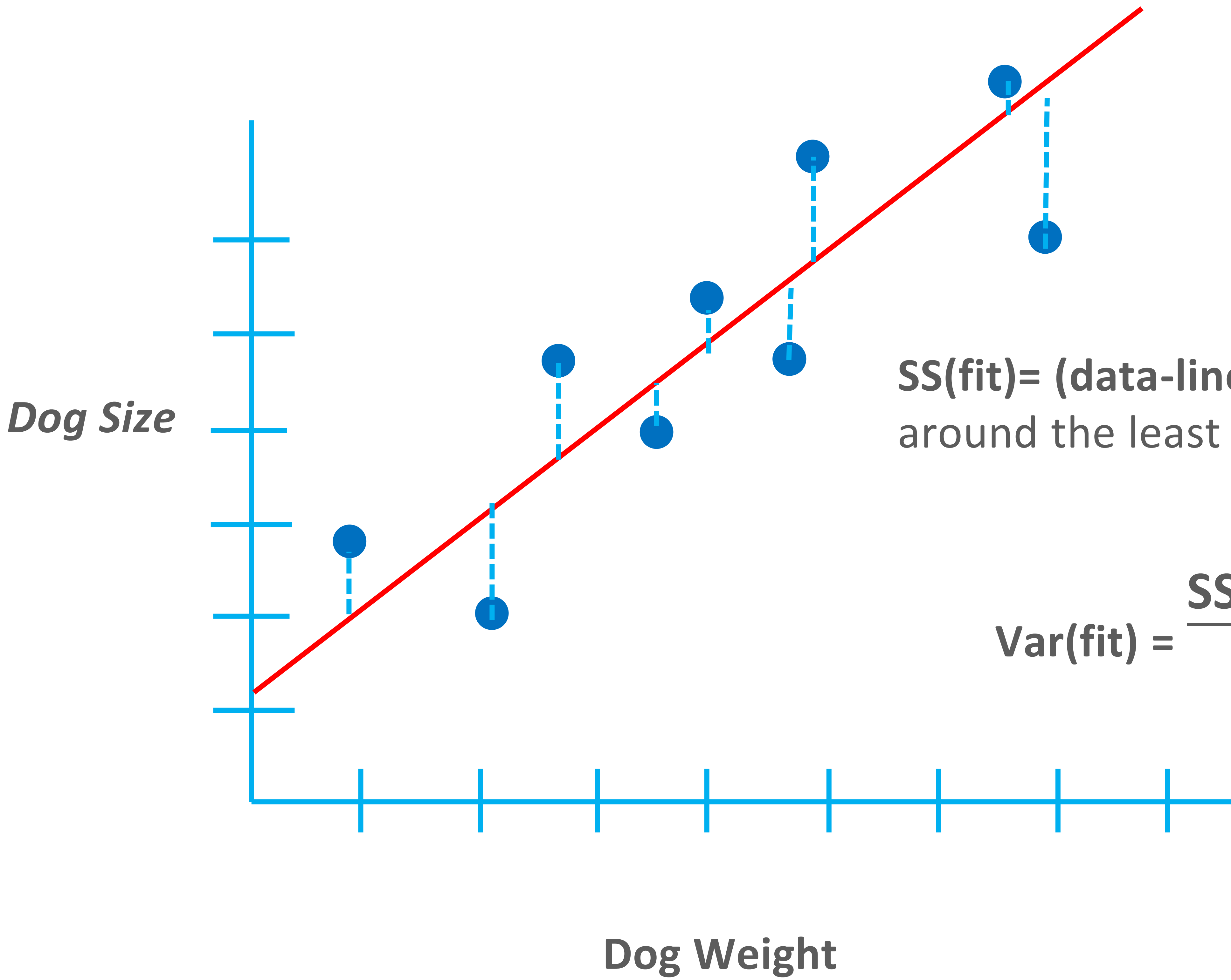
SS(mean)= (data-mean)^2 (SS: Sum of square)

Variation around mean = $\dfrac{\textbf{(data−mean)2}}{n}$

Var(mean) = $\dfrac{\textbf{SS(mean)}}{n}$

Avg Sum of squares per dog

WEEK 2 RECAP: LINEAR REGRESSION



SS(fit)= (data-line)² (sum of squares around the least squares fit)

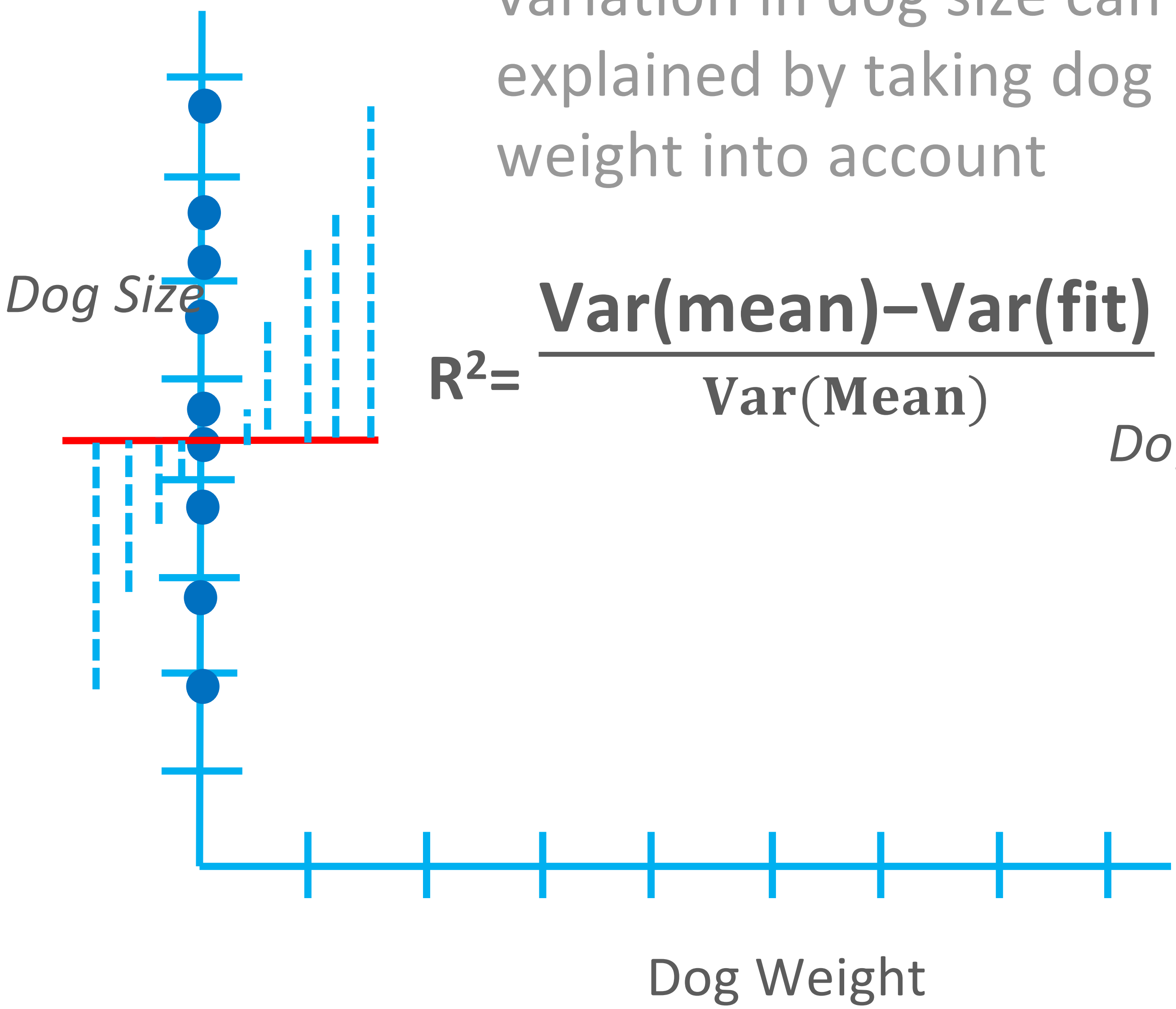
$$\textbf{Var(fit)} = \frac{\textbf{SS(fit)}}{n}$$

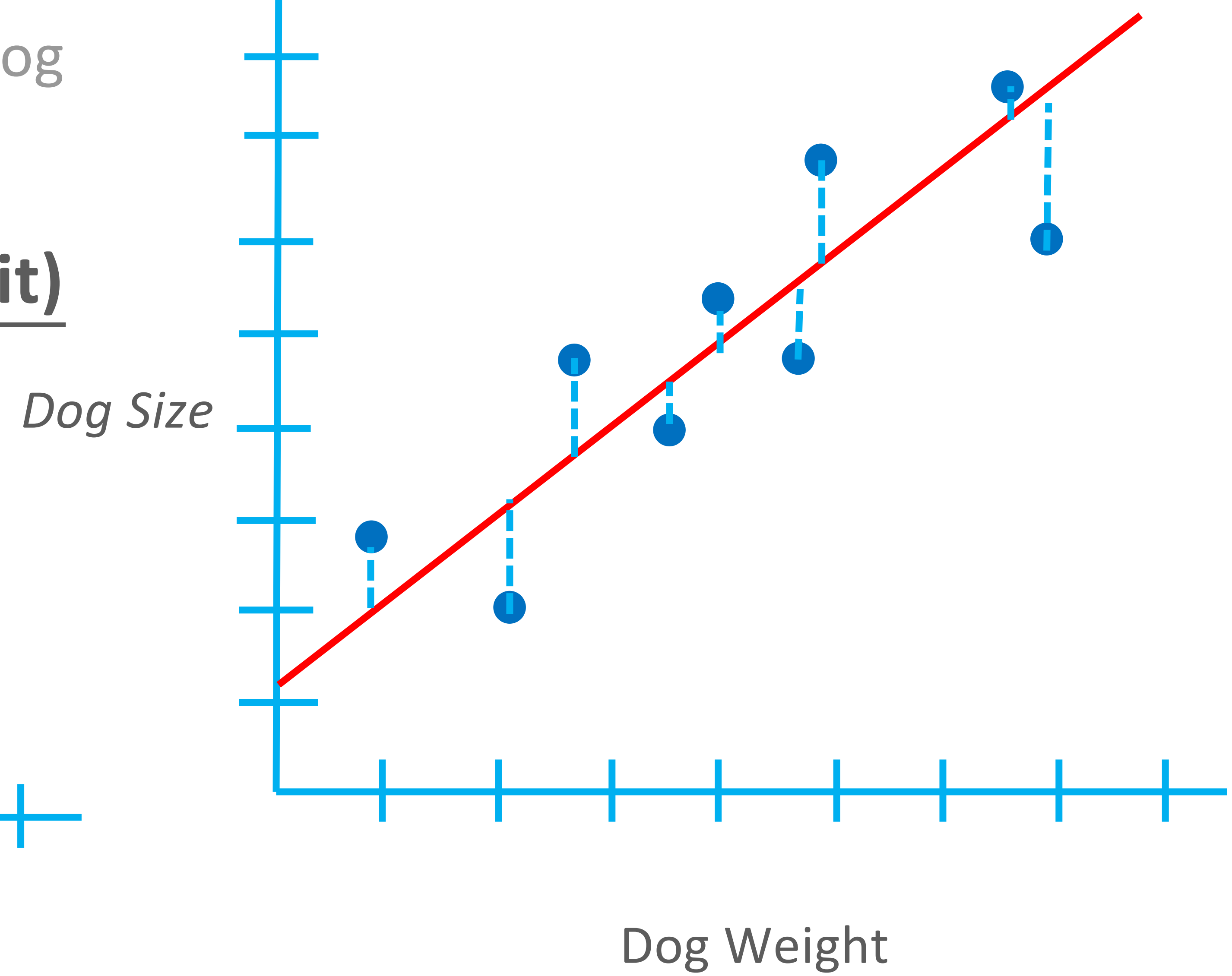
WEEK 2 RECAP: LINEAR REGRESSION

JAMES COOK UNIVERSITY AUSTRALIA

R^2 tells us how much of the variation in dog size can be explained by taking dog weight into account

$$R^2 = \frac{\text{Var(mean)} - \text{Var(fit)}}{\text{Var(Mean)}}$$

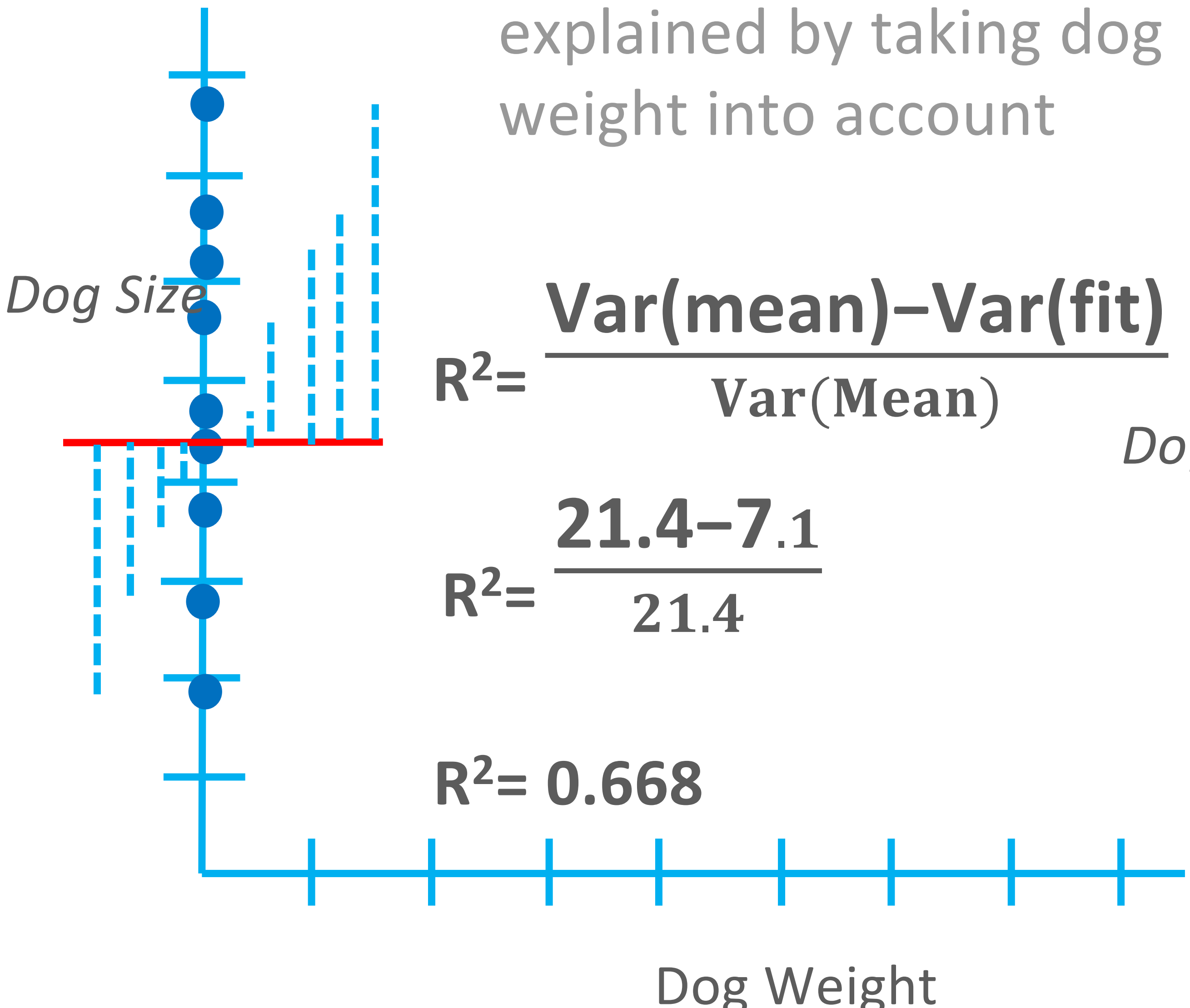




JAMES COOK UNIVERSITY AUSTRALIA

WEEK 2 RECAP: LINEAR REGRESSION

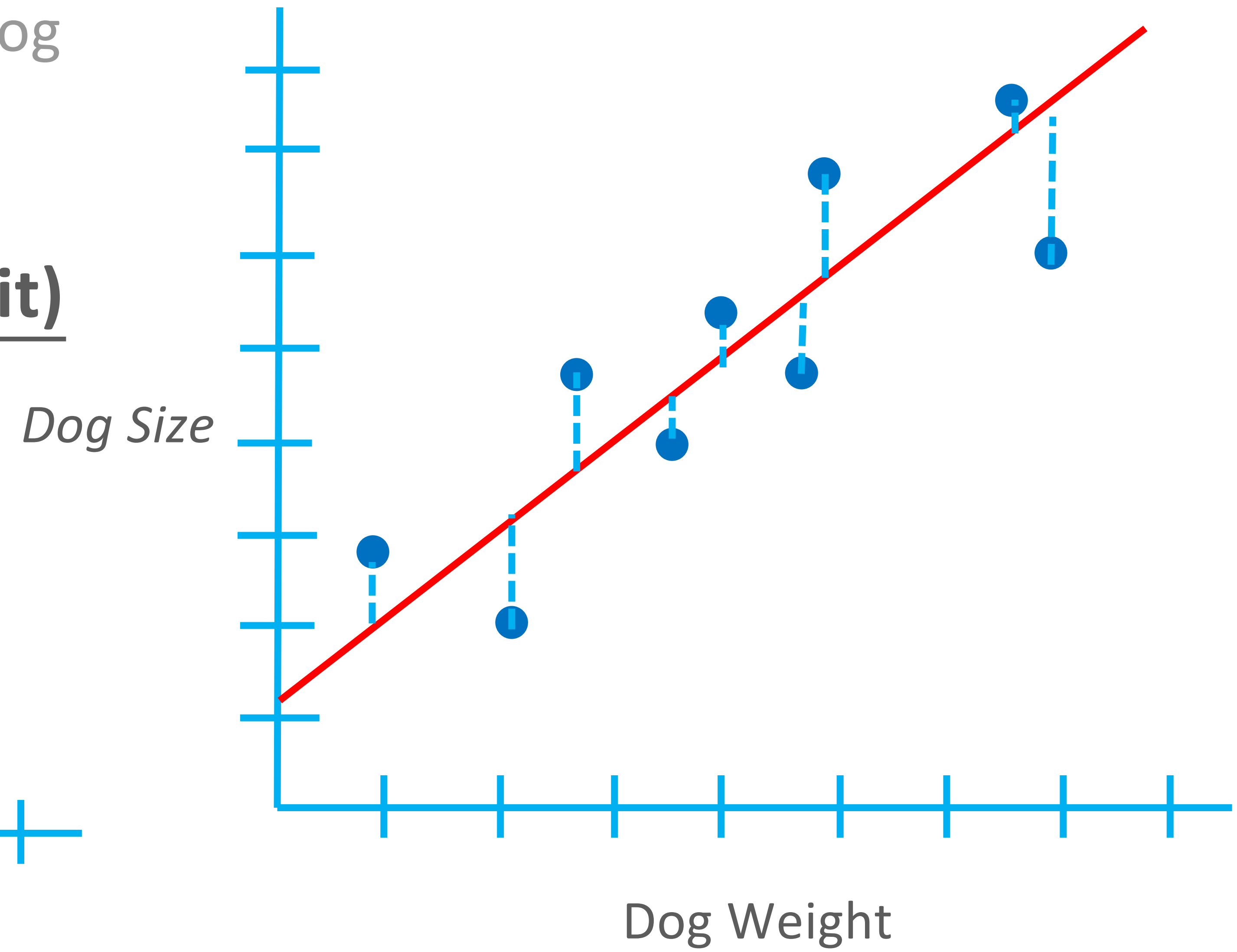
R^2 tells us how much of the variation in dog size can be explained by taking dog weight into account



$$R^2 = \frac{\text{Var(mean)} - \text{Var(fit)}}{\text{Var(Mean)}}$$

$$R^2 = \frac{21.4 - 7.1}{21.4}$$

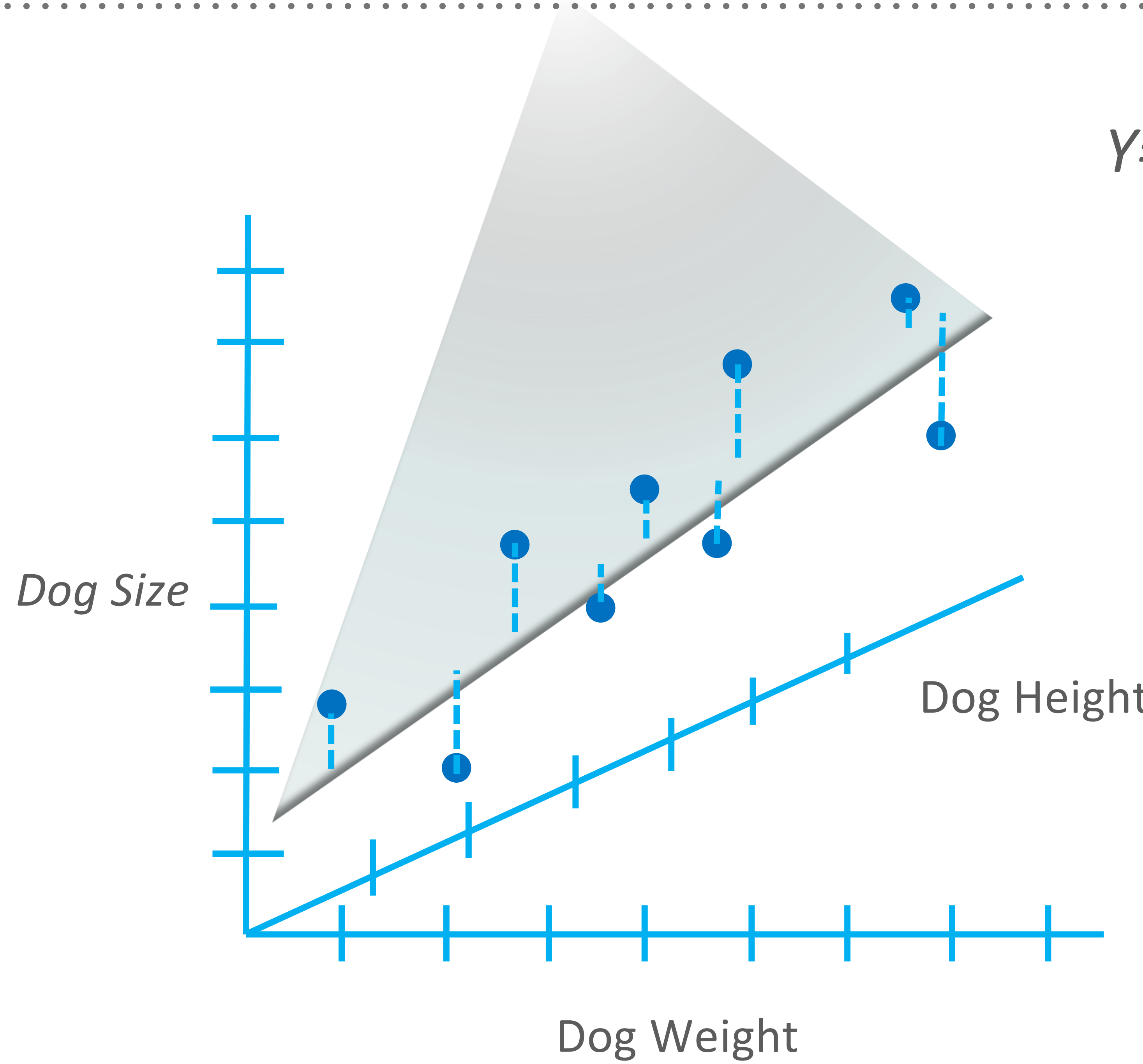
$$R^2 = 0.668$$



There is a 66% reduction in variance when we take dog weight into account

Alternatively, dog weight explains 66% of the variation in dog size

WEEK 2 RECAP: LINEAR REGRESSION



Y=9.3 + 0.78x +0z

If height (z-axis) is useless and doesn't make ss(fit) any smaller, then **least squares** will ignore it by making that parameter =0

WEEK 2 RECAP: F-VALUE

$$R^2 = \frac{\textbf{The variation in dog size explained by weight}}{\textbf{The variation in dog size without taking weight into account}}$$

$$F = \frac{\textbf{The variation in dog size explained by weight}}{\textbf{The variation in dog size not explained by weight}}$$

$$F = \frac{SS(mean) - SS(fit)/(P_{fit} - P_{mean})}{SS(fit)/(n - P_{fit})}$$

***(P- value** comes from F)*

$P_{fit} = number\ of parameters\ in\ the\ fit\ line = 3$ (*Y=9.3 + 0.78x +.98z*)

$P_{mean} = number\ of parameters\ in\ the\ mean\ line = 1(only\ y\ intercept\ or$ *Y=9.3*)

JAMES COOK UNIVERSITY AUSTRALIA

If Fit is good then

$$F = \frac{\textbf{The variation explained by the extra parameter in the fit}}{\textbf{The variation not explained by the extra parameter in the fit}}$$

$$F = \frac{\textbf{Large Number}}{\textbf{\textit{Small Number}}}$$

$$F = \frac{SS(mean) - SS(fit)/(P_{fit} - P_{mean})}{SS(fit)/(n - P_{fit})}$$

*(**P- value** comes from F)*

Y=y-intercept + slope x + slope z

size= 0.9034 – 0.3523 weight + 1.2347 height

Coefficients:

This is the p-value

	Estimate	*std. Error*	*t value*	*Pr(>|t|)*
Intercept	0.9034	0.2372	1.034	0.546
weight	-0.3523	0.4556	0.7565	0.3523
height	1.2347	0.5455	2.7342	0.0315

..it means using **weight** and **height** isn't significantly better than using **height** alone to predict **size**

$Y = y\text{-}intercept + slope_1 \ x \ weight + slope_2 \ x \ height$

$Y = y\text{-}intercept + \cancel{slope_1 \ x \ weight} + slope_2 \ x \ height$

WEEK 2 RECAP: REGRESSION ANALYSIS

Y= y-intercept + slope x + slope z

size= 0.9034 – 0.3523 weight + 1.2347 height

Coefficients:

This is the p-value

	Estimate	*std. Error*	*t value*	*Pr(>|t|)*
Intercept	0.9034	0.2372	1.034	0.546
weight	-0.3523	0.4556	0.7565	0.3523
height	1.2347	0.5455	2.7342	0.0315

..it means using **weight** and **height** is significantly better than using **weight** alone to predict **size**

$$Y = y\text{-}intercept + slope_1 \times weight + slope_2 \times height$$

$$Y = y\text{-}intercept + slope_1 \times weight \cancel{+ slope_2 \times height}$$

REGULARISATION

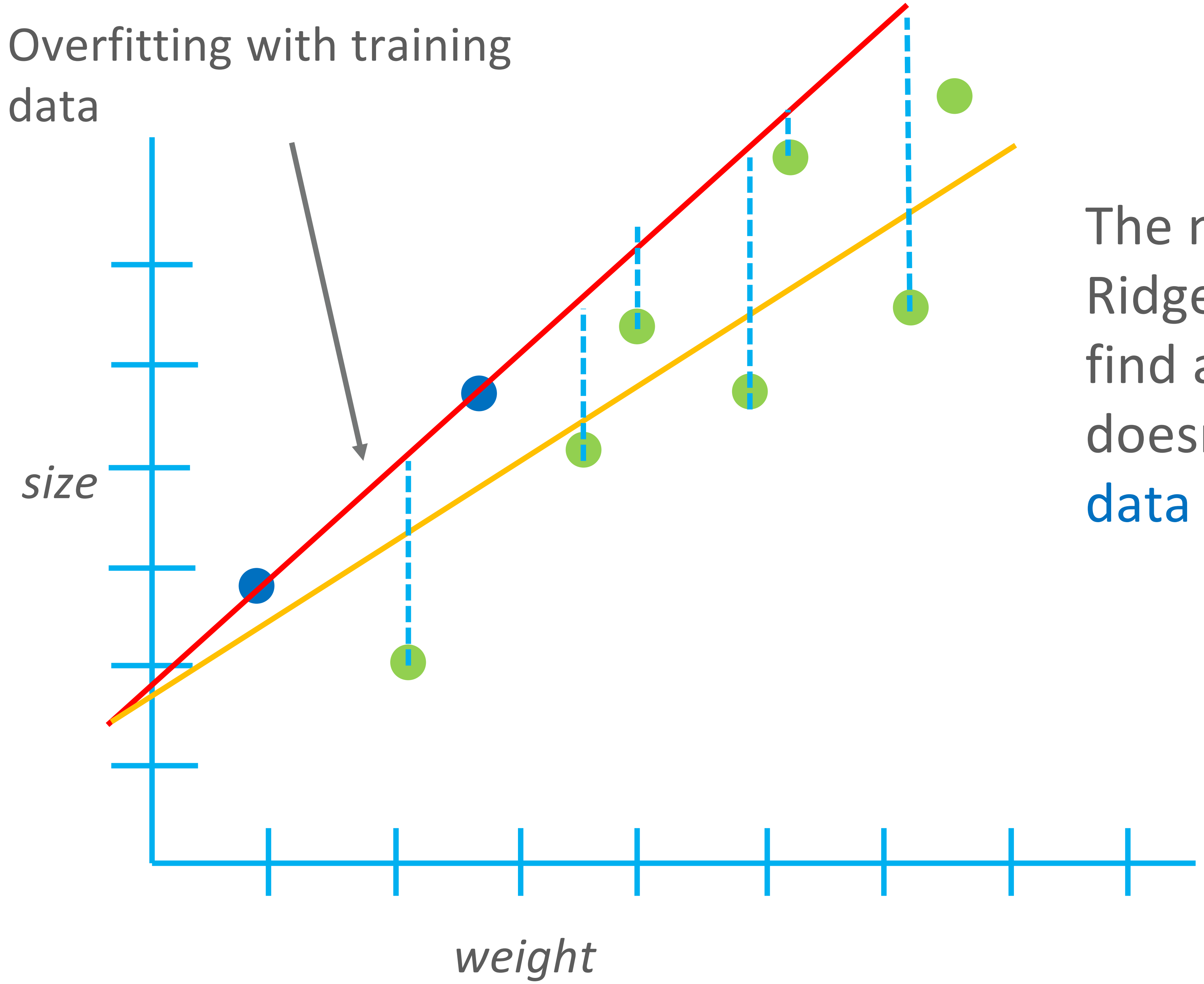
Regularization is a way to give a penalty to certain models (usually overly complex ones).

Two commonly used types of regularized regression methods are ridge regression and lasso regression.

Ridge regression is a way to create a parsimonious model when the number of predictor variables in a set exceeds the number of observations, or when a data set has multicollinearity (correlations between predictor variables).

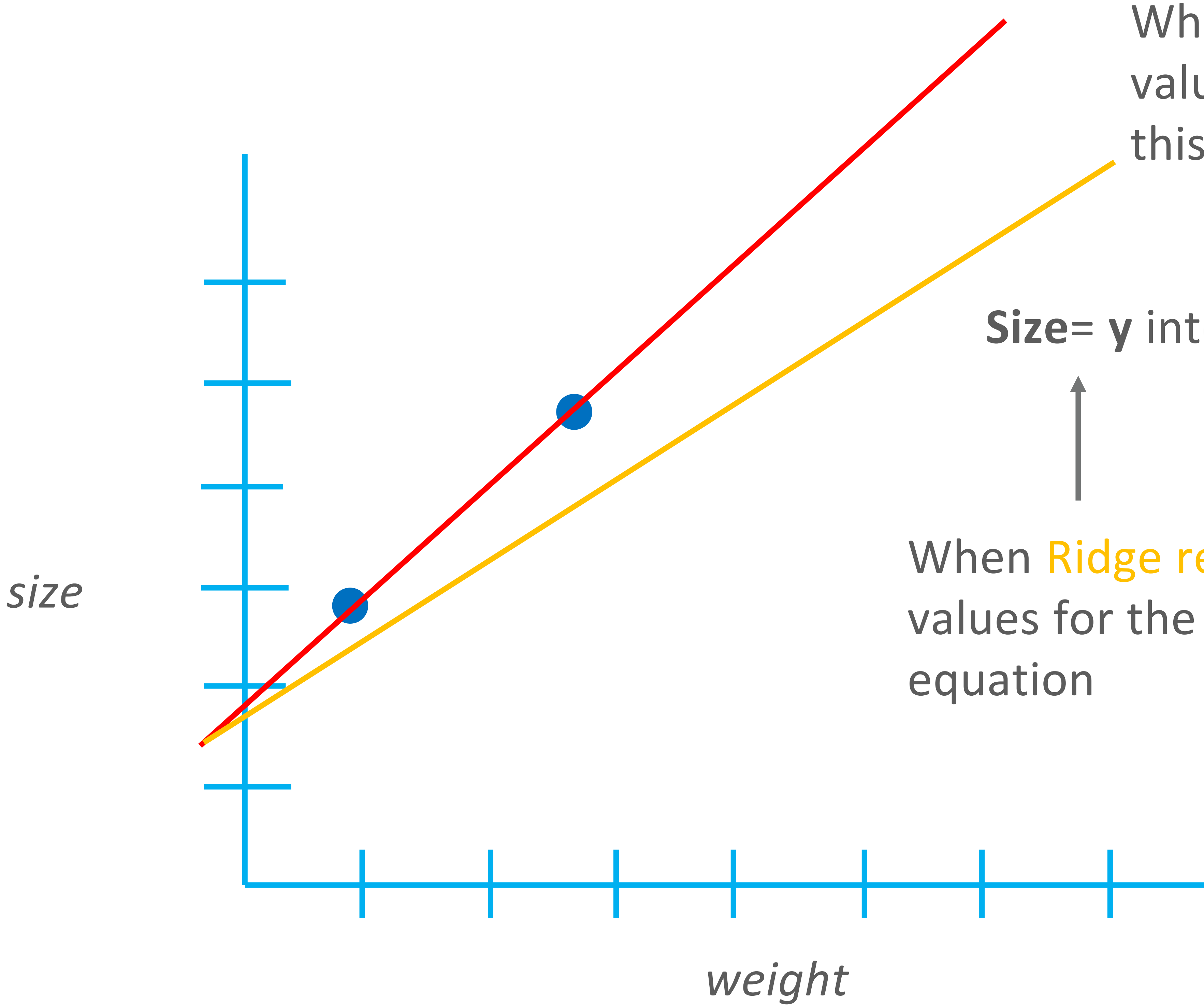
Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. This type is very useful when you have high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

RIDGE REGRESSION



The main idea behind Ridge regression is to find a new line that doesn't fit the training data as well

RIDGE REGRESSION



When least square determines values for the parameters in this equation

Size= **y** intercept + **slope** X **weight**

..It minimizes the **SSR**

When Ridge regression determines values for the parameters in this equation

..It minimizes the (sum of squared residuals)**SSR**

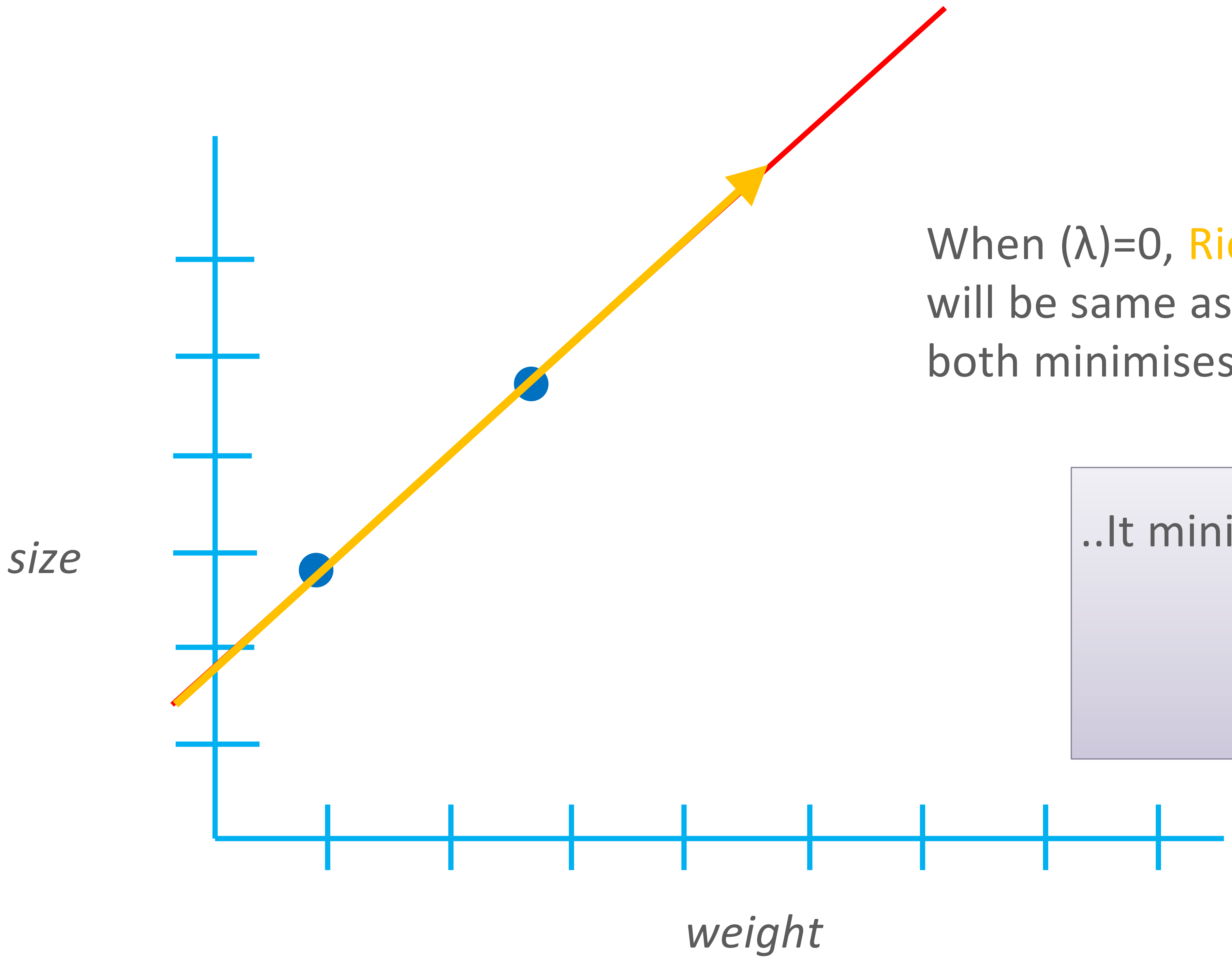
+

(λ) x **Slope²**

..this is called **Penalty**

(**λ** can be any value from 0 to +ve infinity)

RIDGE REGRESSION



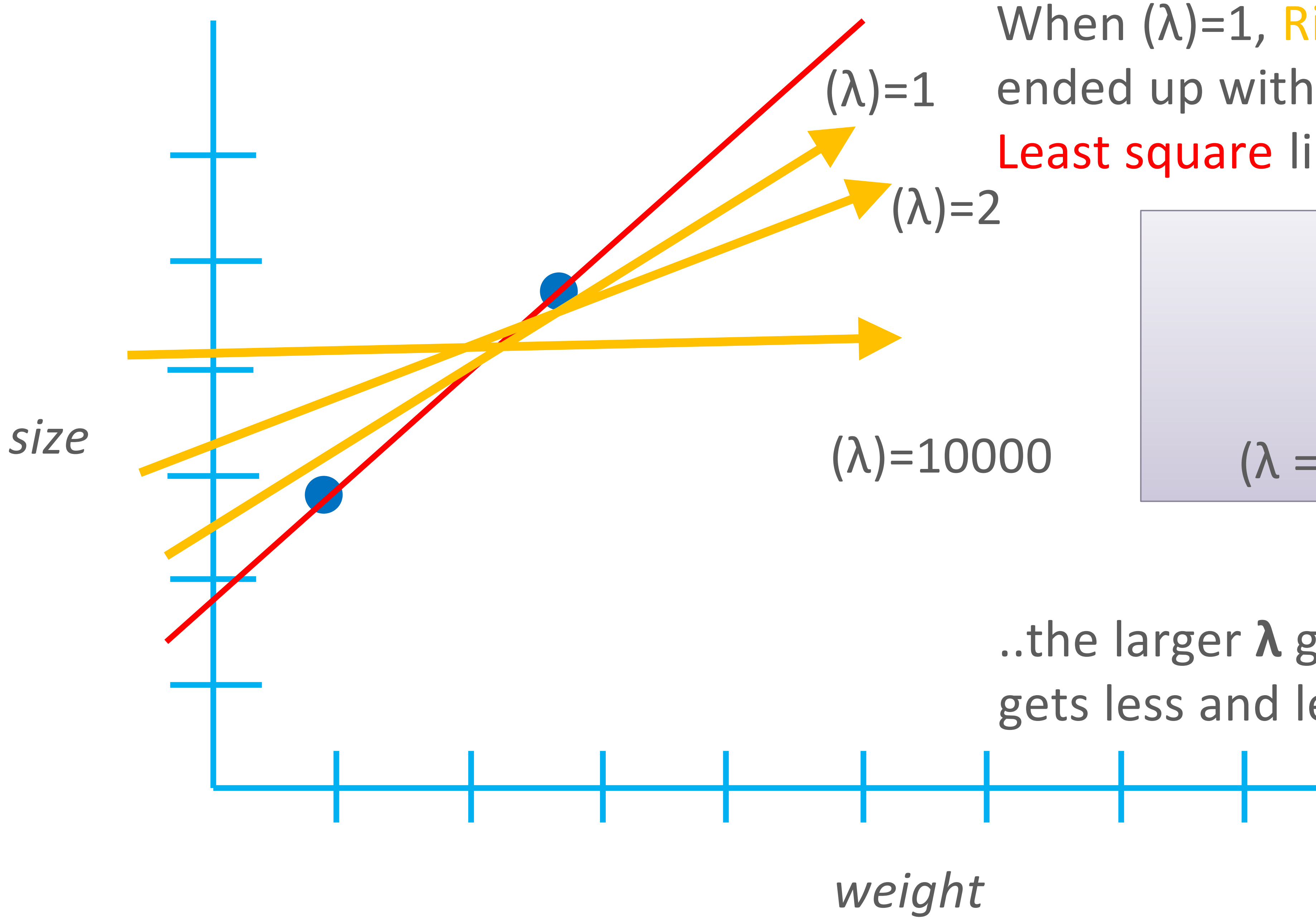
When (λ)=0, Ridge regression line will be same as Least square line as both minimises same thing

..It minimizes the **SSR**

+

0

RIDGE REGRESSION



When (λ)=1, Ridge regression line ended up with a smaller slope than Least square line

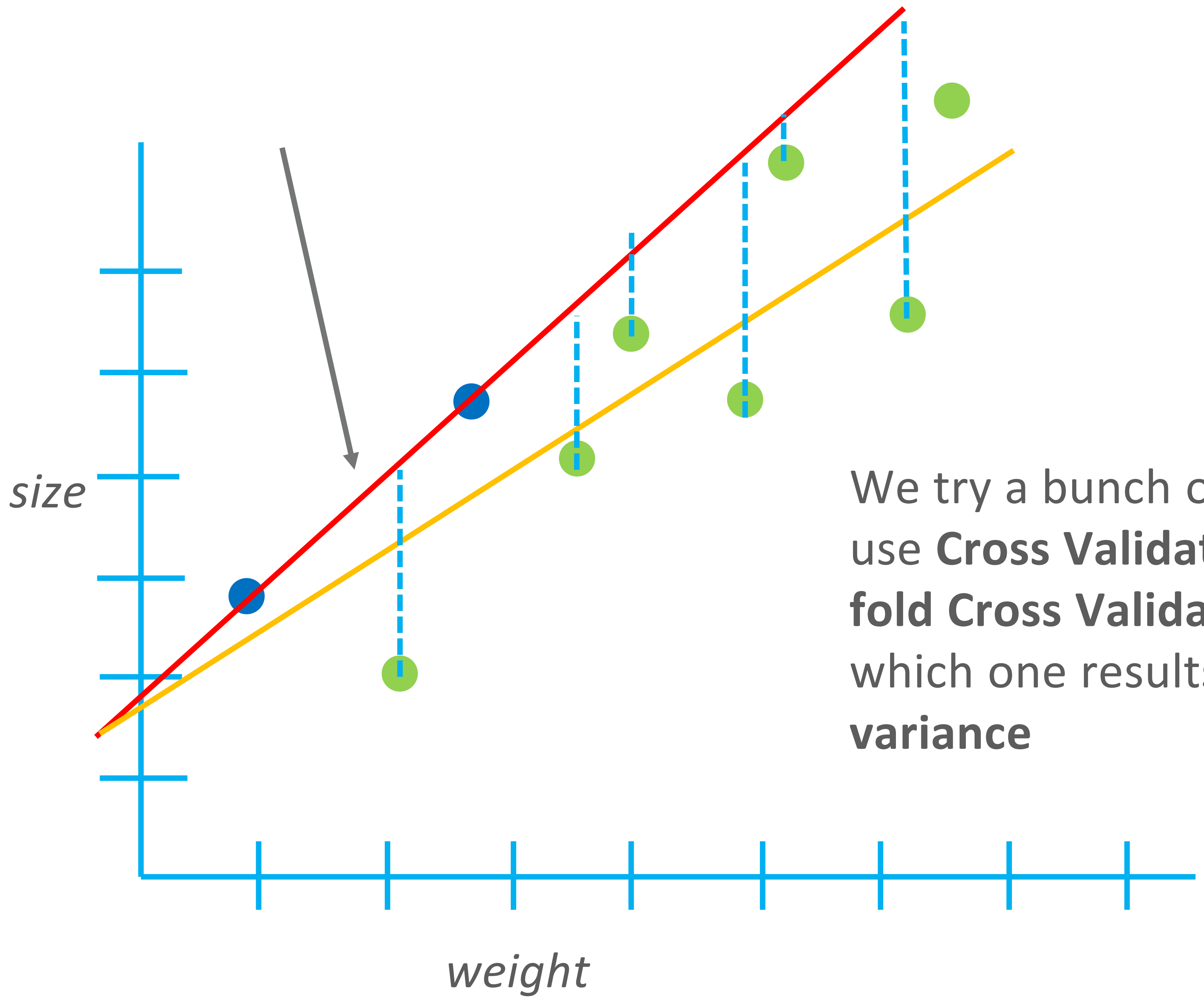
SSR

+

(λ =**1**) x **Slope²**

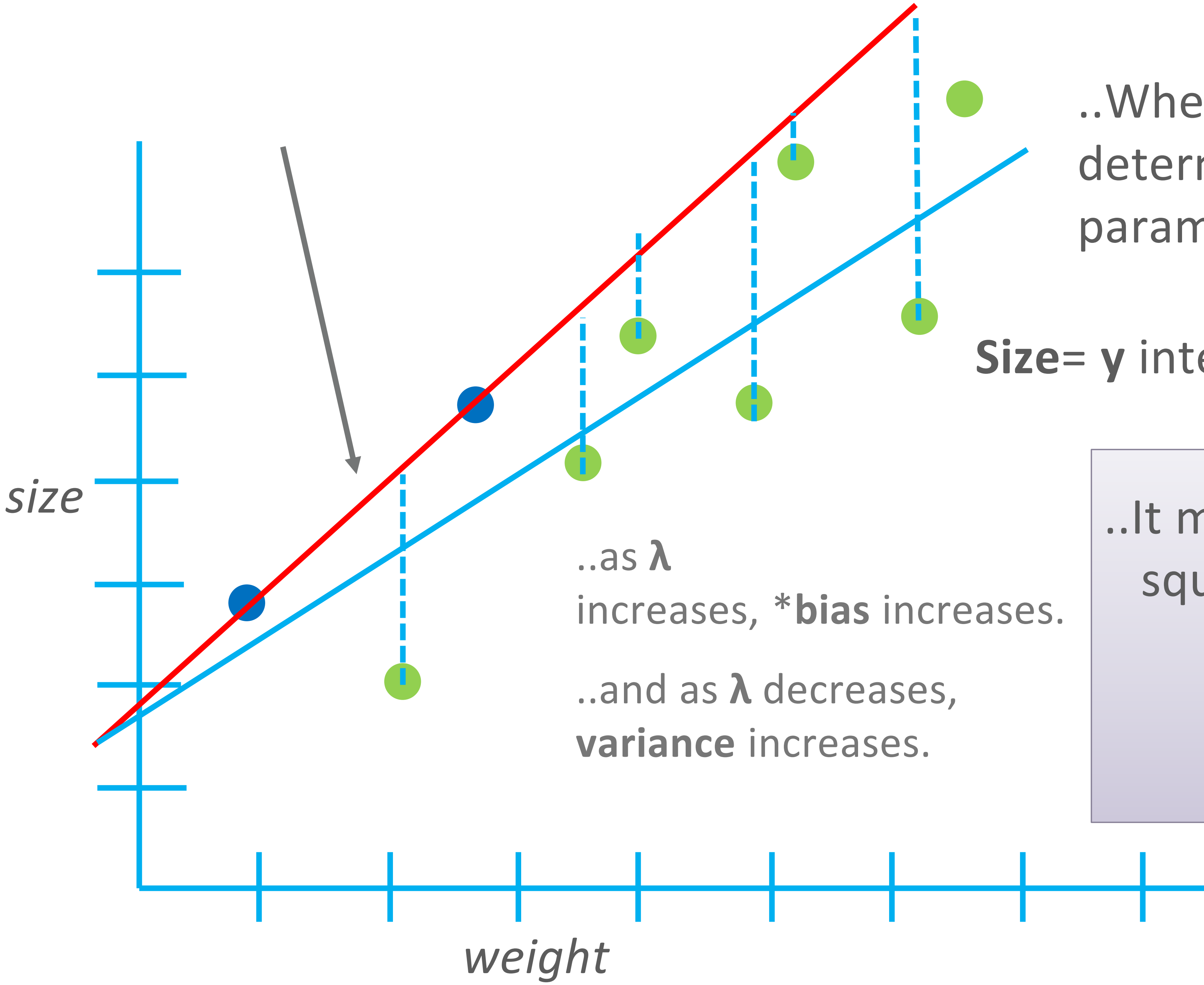
..the larger **λ** gets, prediction for **size** gets less and less sensitive to **weight**

RIDGE REGRESSION



We try a bunch of values for **λ**, and use **Cross Validation** (typically **10-fold Cross Validation**) to determine which one results in the lowest **variance**

LASSO REGRESSION



..When LASSO regression determines values for the parameters in this equation

Size= **y** intercept + **slope** X **weight**

..It minimizes the (sum of squared residuals)**SSR**

+

(λ) x **|Slope|**

..as **λ** increases, ***bias** increases.

..and as **λ** decreases, **variance** increases.

* **bias** is the tendency of a statistic to overestimate or underestimate a parameter. Bias can seep into your results for a slew of reasons including **sampling** or **measurement** errors, or **unrepresentative** samples

DIFF BETWEEN RIDGE & LASSO REGRESSION

..When RIDGE regression determines values for the parameters in this equation

..When LASSO regression determines values for the parameters in this equation

Size= **y** intercept + **slope$_1$** X **weight** + **slope$_2$** X **length +…**

..It minimizes the (sum of squared residuals)**SSR**

+

(λ) x **($Slope_1^2$ + $Slope_2^2$+..)**

..It minimizes the (sum of squared residuals)**SSR**

+

(λ) x **(|$Slope_1$|+| $Slope_2$ +.. |**

The big difference between RIDGE and LASSO regression is, RIDGE regression can only shrink the slope asymptotically close to 0, while LASSO regression can shrink the slope all the way to 0

DIFF BETWEEN RIDGE & LASSO REGRESSION

Since LASSO regression can exclude useless variables from equations, it is a little better than RIDGE regression at reducing **Variance** in models that contain a lot of useless variables

In contrast RIDGE regression tends to a little better when most variables are useful.

MISSING VALUES AND REGRESSION MODELLING

Some common reason or missing values

- There was no data to be captured.
- The information is not applicable, so no values were entered
- In some software applications there may be no mandatory requirement that data be entered.
- System integration problems as data is passed from one platform to another.

Synthetic distribution methods use a ‘one size fits all’ approach to handle missing values. Any case with a missing input measurement has the missing value replaced with a fixed number. The net effect is to modify an input’s distribution to include a point mass at the selected fixed number. many modelling methods, this can be achieved by locating the point mass at the input’s mean value.

Estimation methods provide tailored imputations for each case with missing values. This is done by viewing the missing value problem as a prediction problem. You can train a model to predict an input’s value from other inputs. Then, when an input’s value is unknown, you can use this model to predict or estimate the unknown missing value. This approach is best suited for missing values that result from a lack of knowledge about values that have no match or are not disclosed.

GENERAL LINEAR MODELS (GLMS)

The **GLM** generalizes linear regression by allowing the linear model to be related to the response variable via a *link function* and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.

GLM provide a higher-level framework for specifying models that can deal with many types of data. GLM extend the theory and methods of linear models to data that is not **normally** distributed or where there are multiple predictor **variables** that come from a **different** data type (categorical, ordinal, positive real, positive integer).

In SAS terminology, GLM variables are as follows:

- There is only one continuous response variable (**Response**)
- Multiple **effects** (independent or predictor) variables, which can be any of the following types:
    - Continuous (**Continuous Effects**)
    - Categorical (**Classification Effects**)
    - Interaction (**Interaction Effects**).