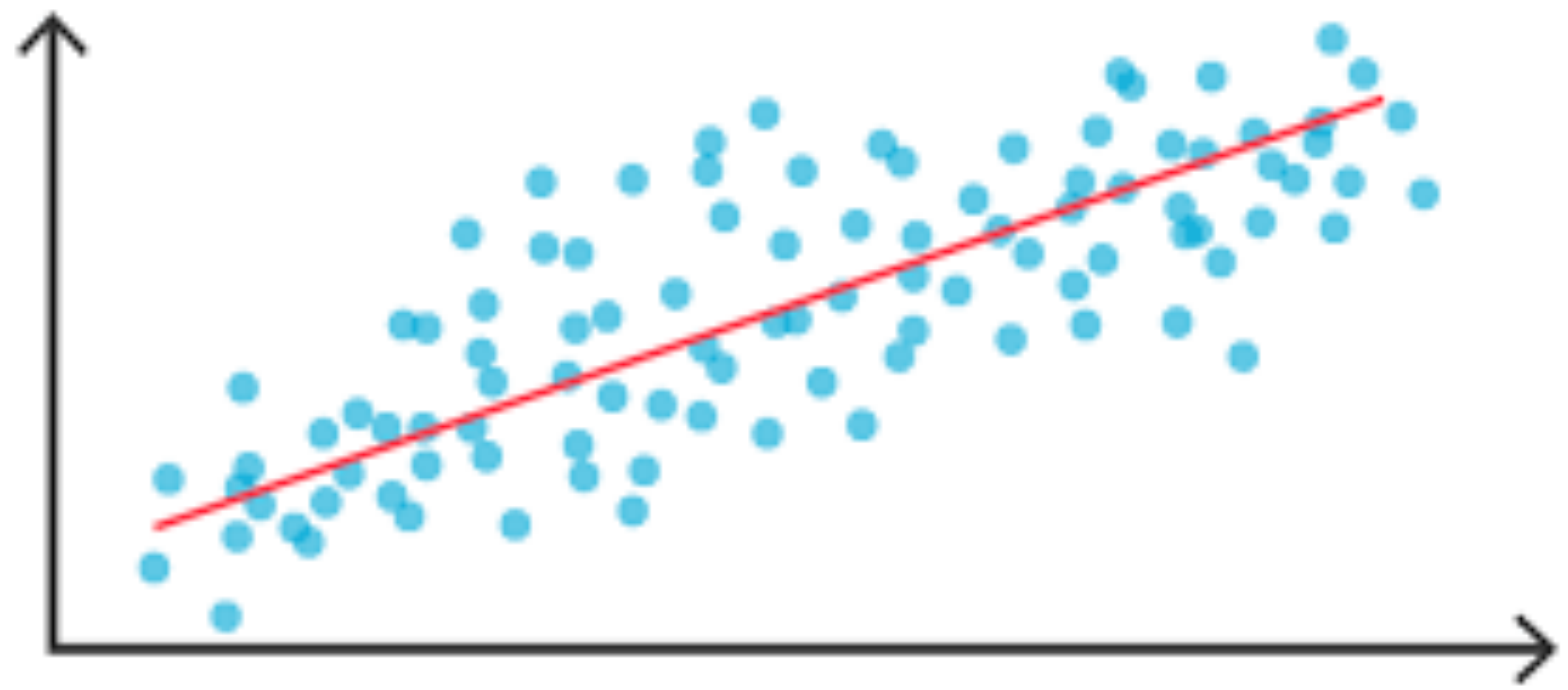
Foundational Statistics

Simple Linear Regression



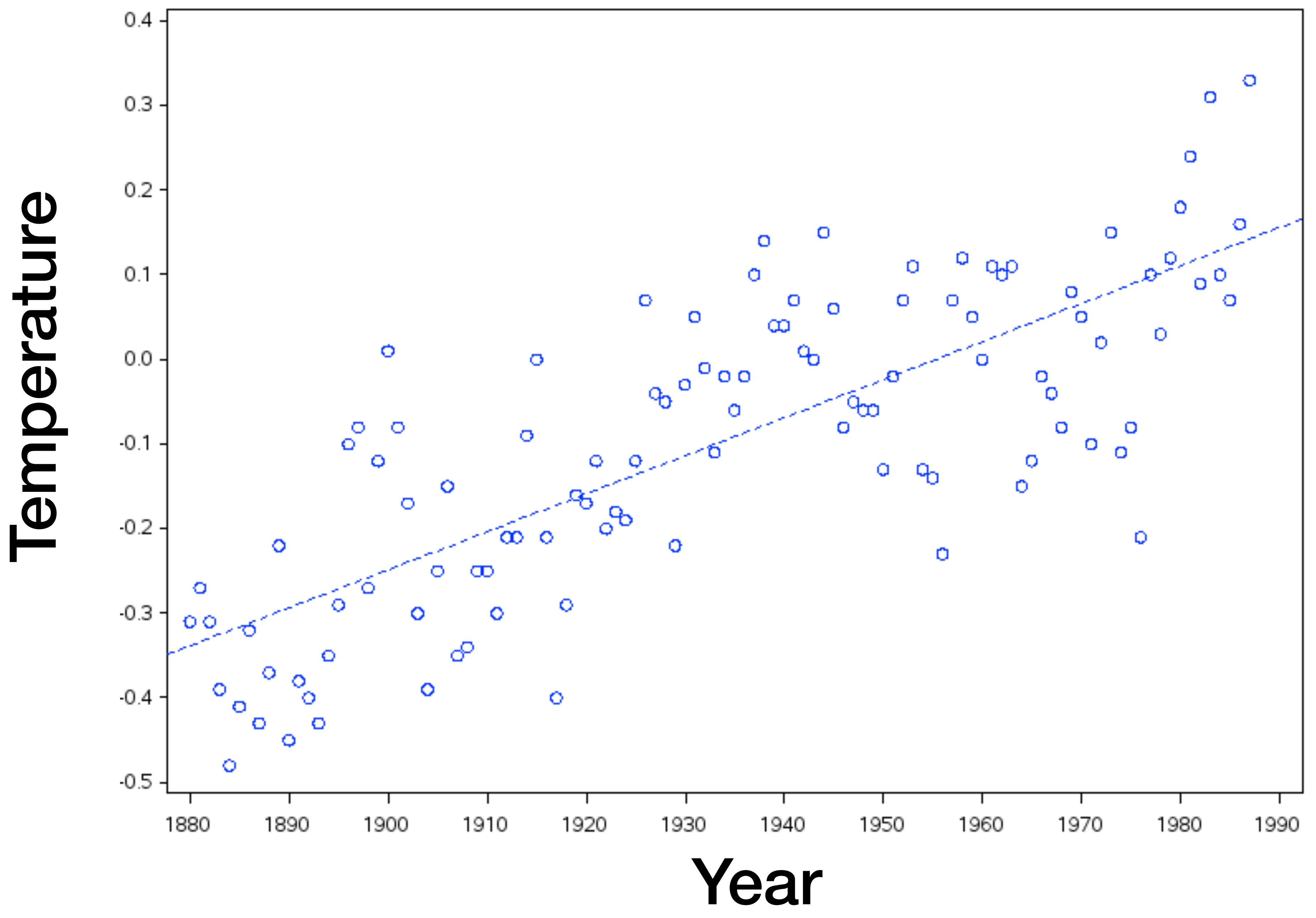
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

A linear model for two numeric variables

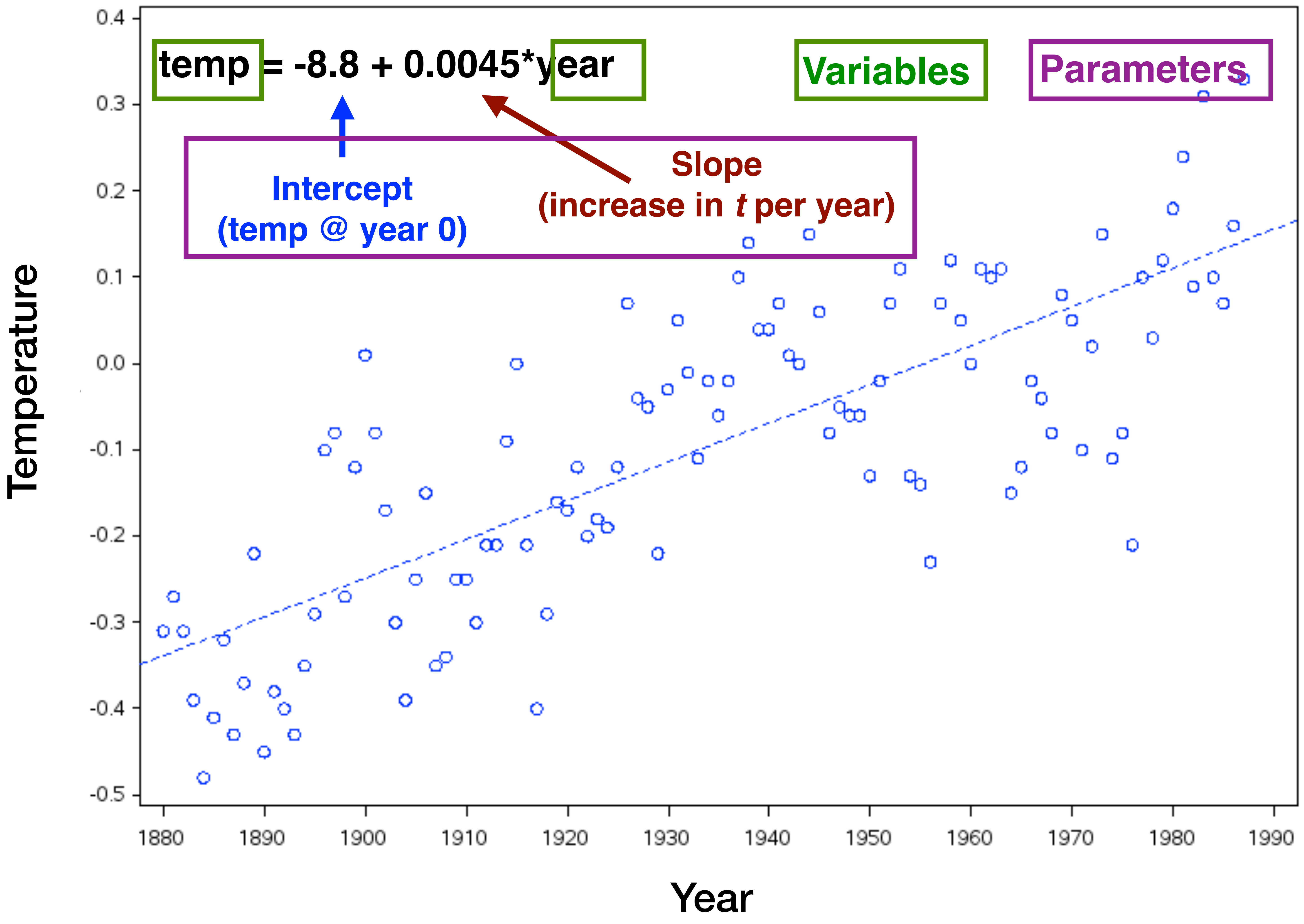
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$y_i = \beta_0 + \beta_1 \times x_i + \varepsilon_i$$

$$\text{response variable} = \underbrace{\underbrace{\text{population intercept}}_{\text{intercept term}} + \underbrace{\text{population slope} \times \text{predictor variable}}_{\text{slope term}}}_{\text{model}} + \text{error}$$



A generic linear model for two numeric variables



Ordinary Least Squares (OLS): one method for estimating β_0 and β_1 from a random sample.

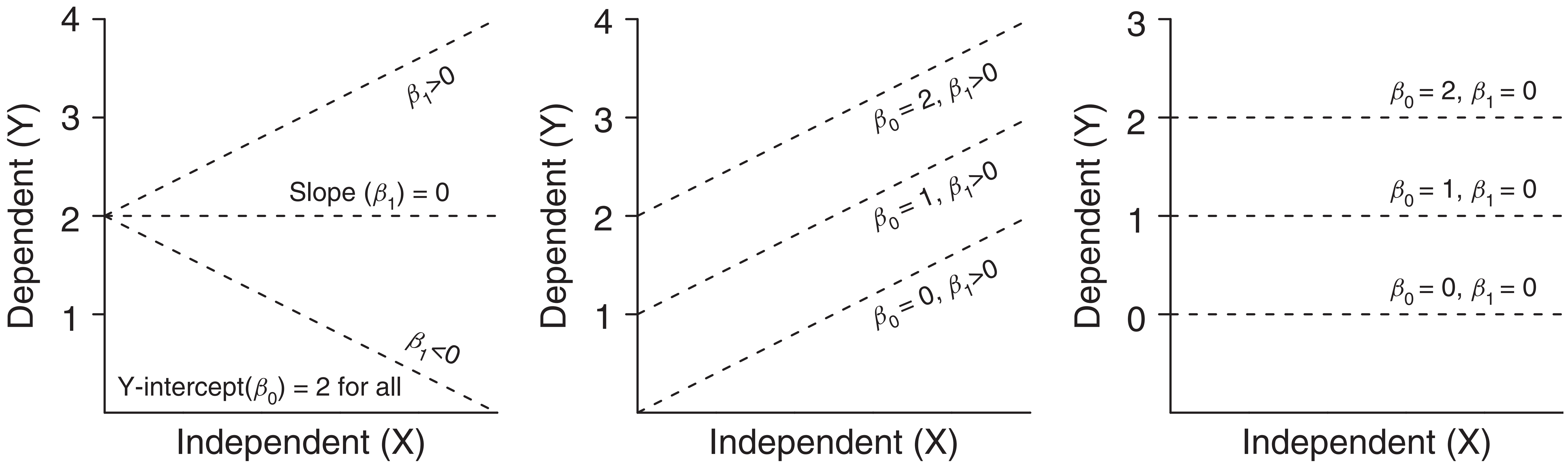
- **Construct a "best fit" linear function to model variation in *y***
- **Function is derived such that the total vertical distance between observed *y*-values and the line are minimized**
- **The *y*-intercept and slope of the line are our sample-based estimates for the population β_0 and β_1**
- **This approach (also called "Model I regression") assumes that the *x*-variable is measured without error**

Ordinary Least Squares (OLS): How it works

Hypothesis tests in linear regression

$$H_0 : \beta_1 = 0$$ (the population slope equals zero)

$$H_0 : \beta_0 = 0$$ (the population y-intercept equals zero)



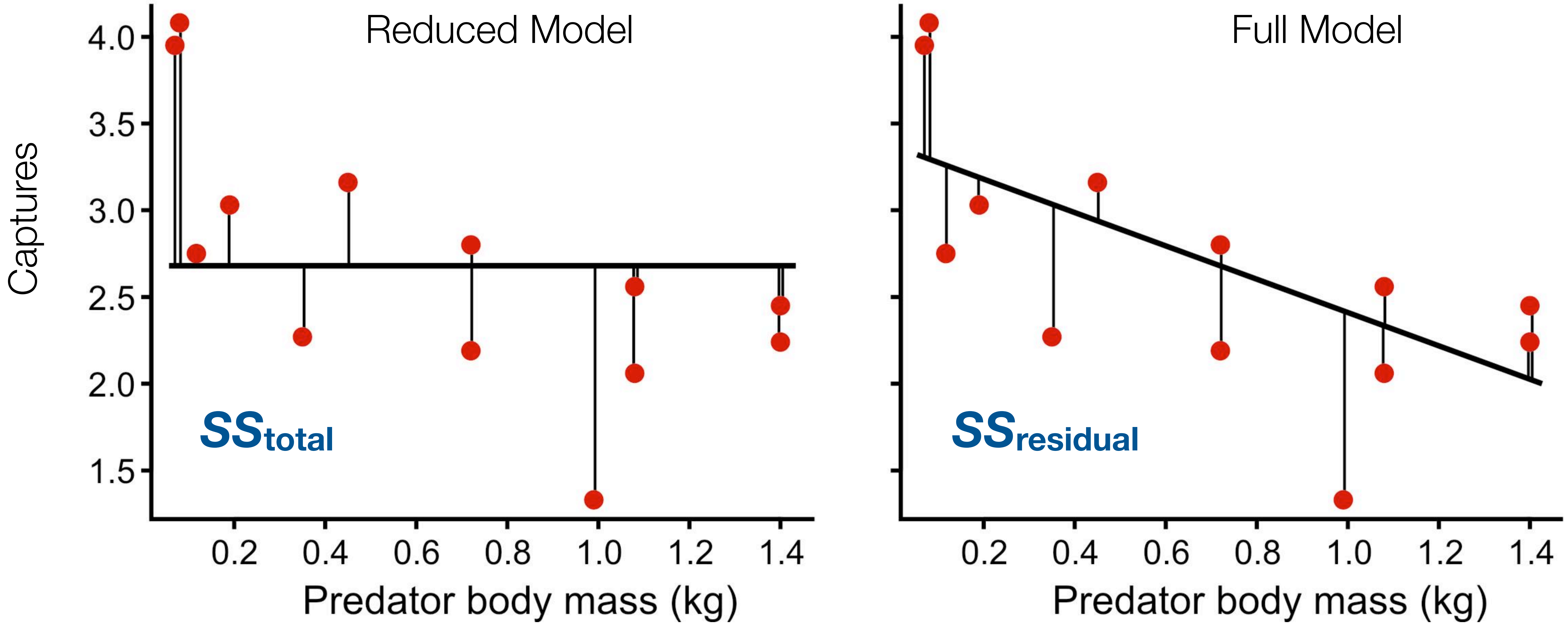
Comparing full and reduced models to test the Null Hypothesis of $\beta_1 = 0$

$$\textit{full model}\ (H_A)\ \text{-} \quad y_i = \beta_0 + \beta_1 x_i + error_i$$

$$\textit{reduced model}\ (H_0)\ \text{-} \quad y_i = \beta_0 + 0x_i + error_i$$
$$= \boxed{\beta_0 + error_i}$$

Model with no slope:
No linear effect of *x* on *y*

Comparing full and reduced models to test the Null Hypothesis of $\beta_1 = 0$



To test null hypothesis:

1) fit a "reduced" model without slope term (fit under H_0)

2) fit the "full" model with slope term added back (fit under H_A)

3) use the full and reduced models to calculate a test statistic that reflects the ratio of explained to unexplained variation by the full model

Comparing full and reduced models to test the Null Hypothesis of $\beta_1 = 0$

(i) the variation that is explained by the model (SS_{Model})

$$SS_{Model} = SS_{Total}\ (reduced\ model) - SS_{Residual}\ (full\ model)$$

The bigger this difference, the better our full model is, relative to the null model

(ii) the variation that is unexplained by the model ($SS_{Residual}$)

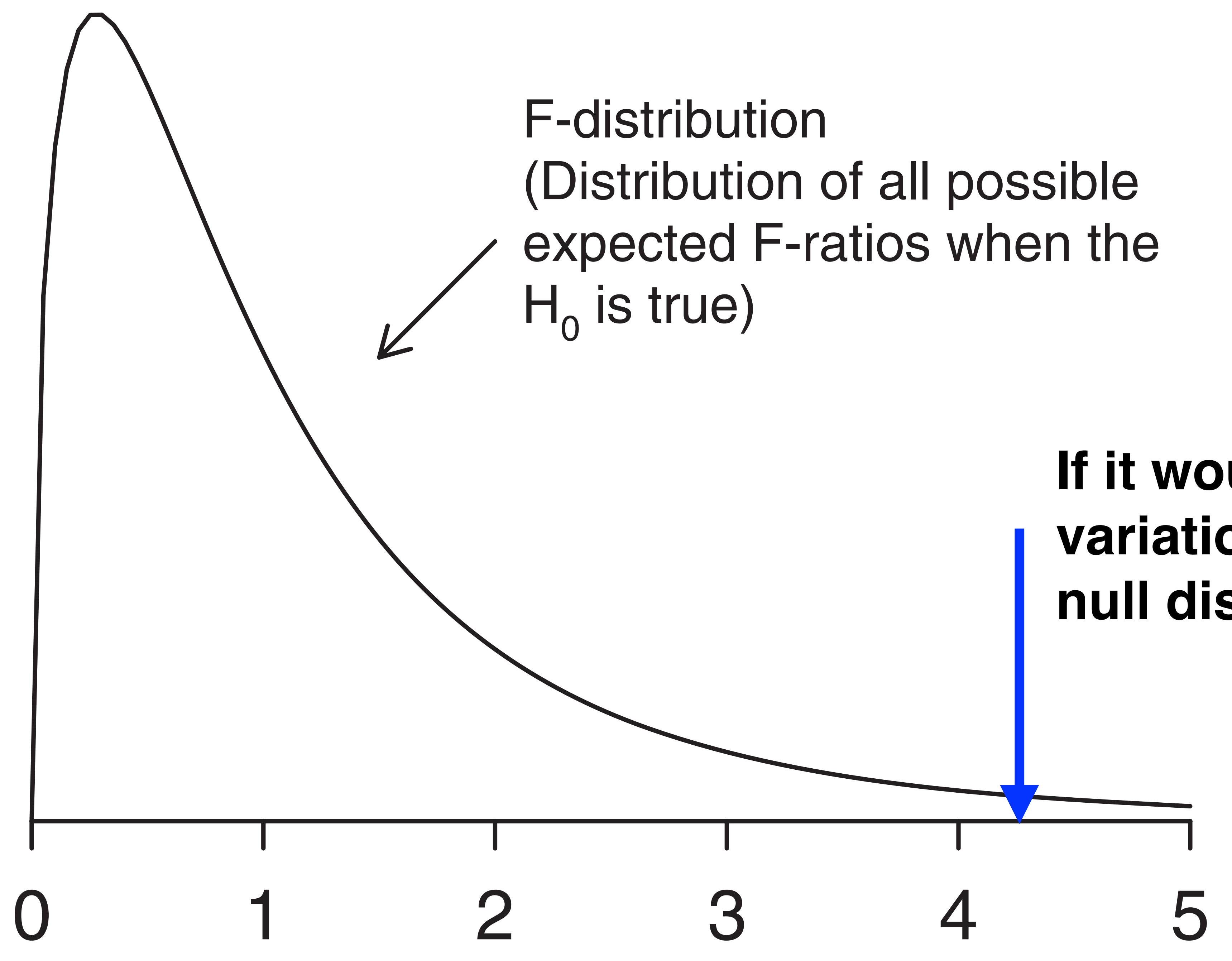
$$SS_{Residual}\ (full\ model)$$

If this value is small, our model explains much of the variation in *y*

So, If our model's ratio of explained to unexplained variation is high,
we reject the null hypothesis of our predictor variables having zero explanatory power

Comparing full and reduced models to test the Null Hypothesis of $\beta_1 = 0$

***F*-ratio (our test statistic) =** $\dfrac{SS_{reduced} - SS_{full}}{SS_{full}} = \dfrac{Var_{explained}}{Var_{unexplained}}$



If it would be rare to observe the level of variation explained by our model, relative to the null distribution, we reject the null hypothesis

Assumptions of the *F*-ratio test for $\beta_1 = 0$

1. Linear relationship between *y* and *x*, under H_A

(Check using scatter plot)

2. Bivariate normally distributed

(Check using histograms, boxplots, etc.)

3. Variance of residuals is homogeneous across all values of *x*

(Check using residuals vs. predicted-*y* plot)

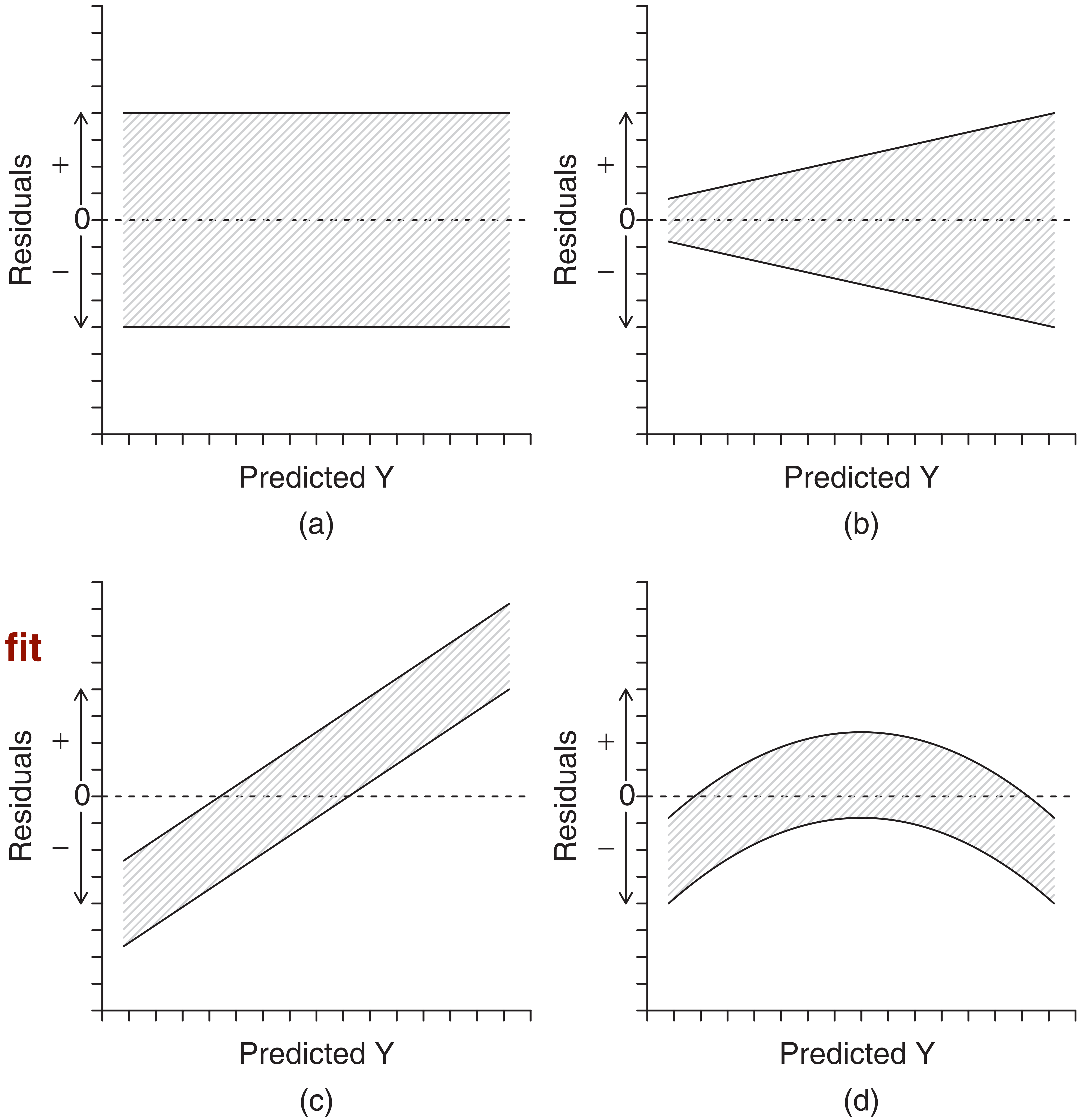
Assumptions of the *F*-ratio test for $\beta_1 = 0$ (Using residual plots)

uniform variance across pred. *Y*

increasing variance as pred. *Y* increases

model is systematically poor fit (may be an outlier)

original relationship is curvilinear, not linear



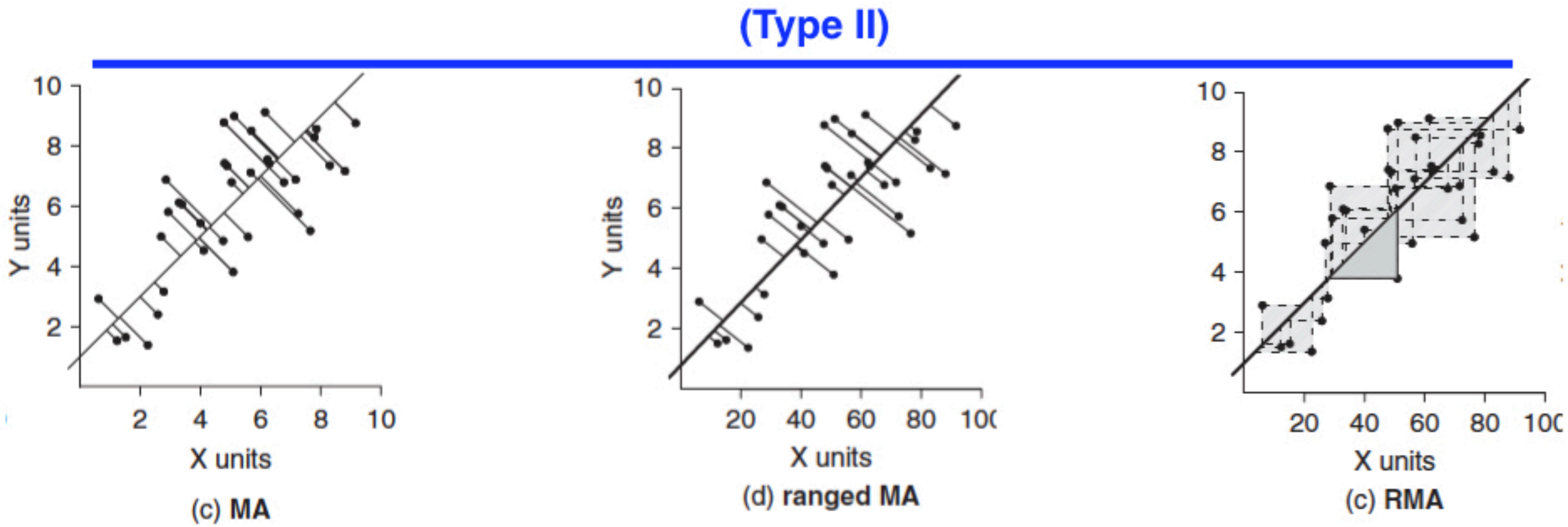
The coefficient of determination (r^2)

$$r^2 = \frac{SS_{regression}}{SS_{total}} = 1 - \frac{SS_{residual}}{SS_{total}}$$

$$r^2 = 1 - \frac{SS_{residual(full)}}{SS_{total(reduced)}}$$

- proportion of variance in Y that is explained by X

Model II regression: when *x* and *y* are both measured with error



MA (Major Axis): *x* and *y* have similar error and have same units

Ranged MA: *x* and *y* in different units or on different scales.

Assumes no outliers

RMA or SMA (Reduced Major Axis): *x* and *y* in different units or on different scales. Robust to outliers