Session 2
Linear Regression I

PSYC 560
Heungsun Hwang

Linear Regression

- Linear regression is a very simple approach for supervised learning.
- It is a good starting point for more complex supervised learning tools.

Linear Regression

- It is useful for explaining or predicting a quantitative response (DV).

  - Two main goals:
    - To investigate the relationship between a DV and predictors

    - To find the best prediction equation for a DV regardless of the meaning of predictors in the equation

Linear Regression

- Linear regression assumes that there is a "linear" relationship between predictors and a response.

$$Y = f(X) + e$$

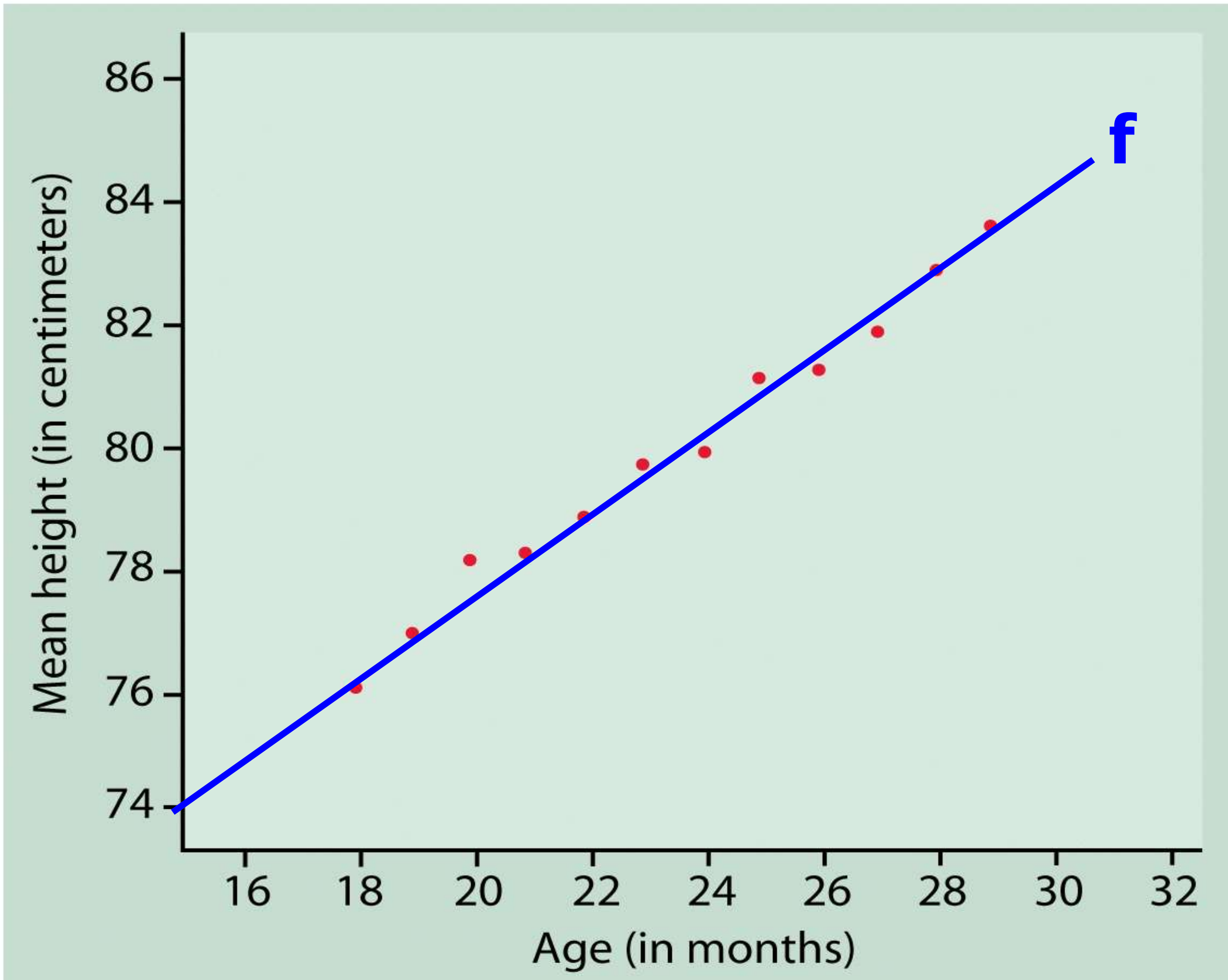
Simple Linear Regression

- When there is a linear relationship between the DV (Y) and a single predictor (X), we can represent this relationship by a straight line, called a regression line.

- A regression line gives a compact description of the dependency of Y on X.

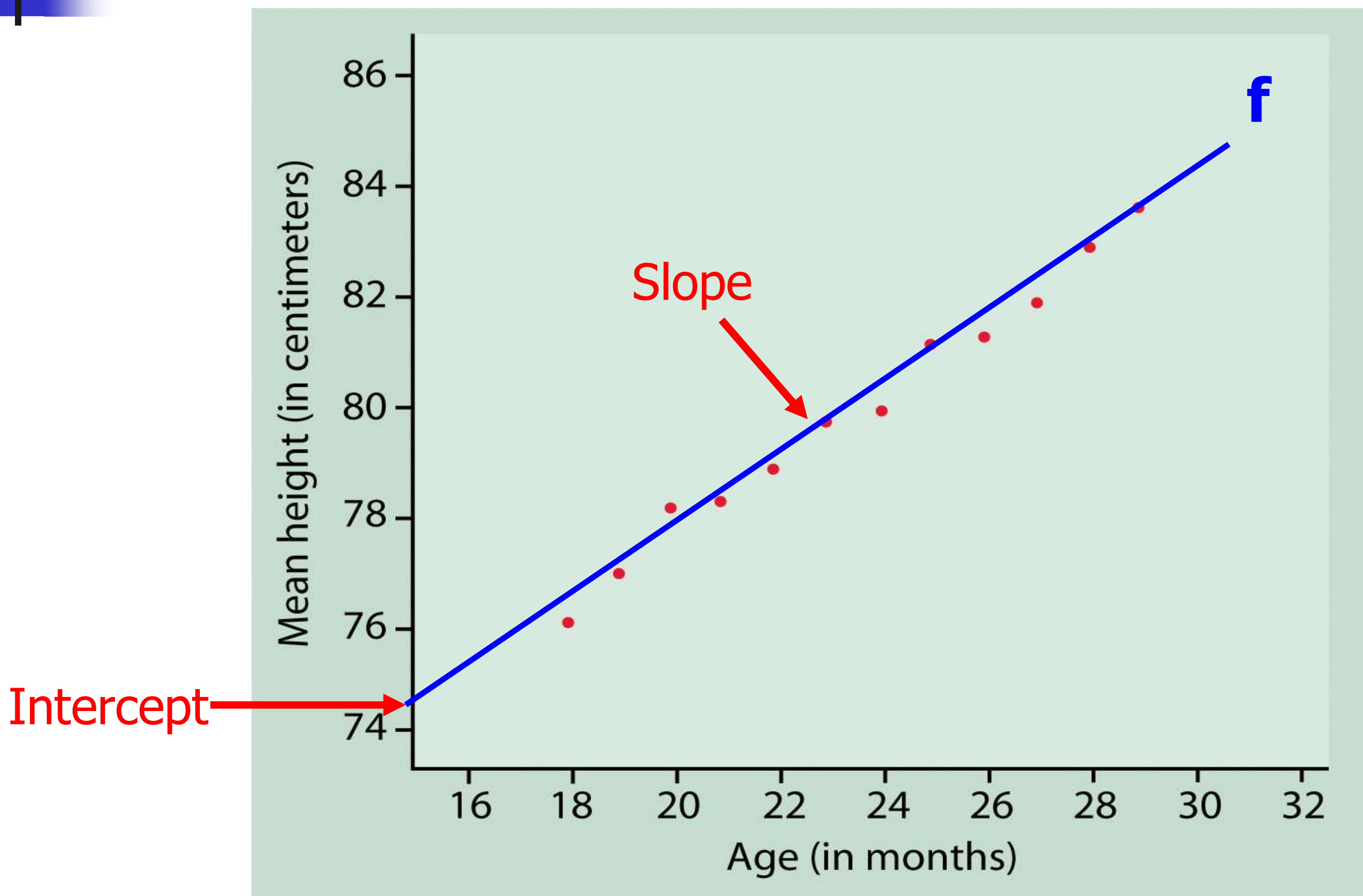
- The equation of a line is a mathematical model for the linear relationship – simple linear regression model.

Simple Linear Regression



Moore & McCabe (1998, p.136)

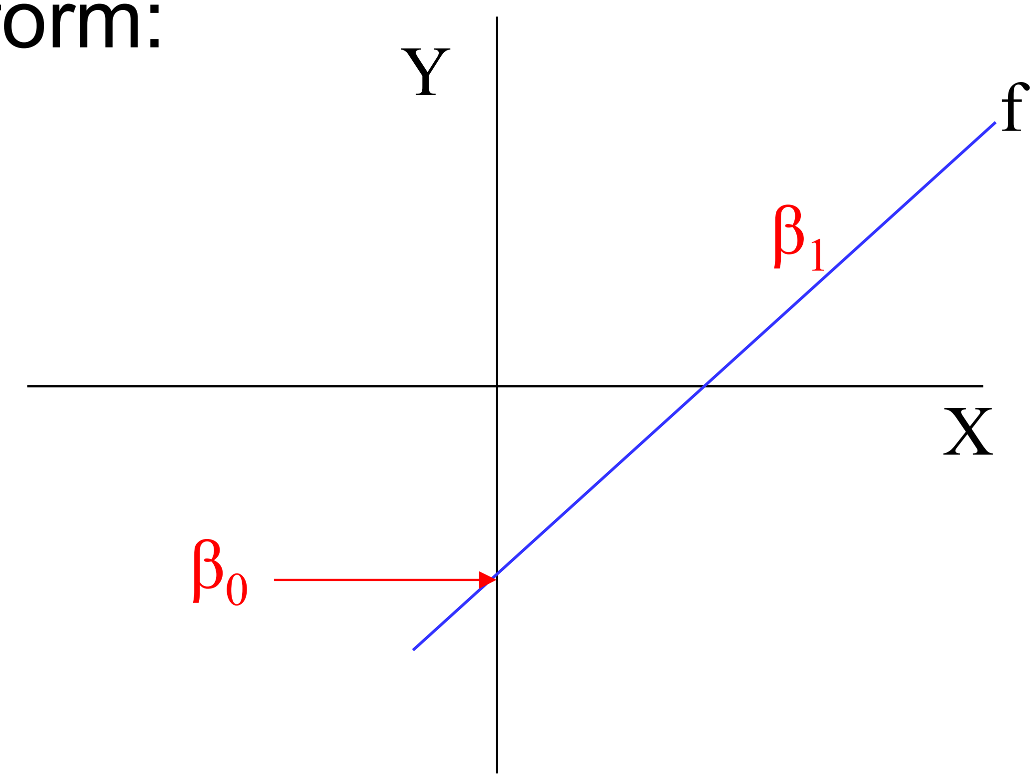
What is the “equation” of a line?



Moore & McCabe (1998, p.136)

Simple Linear Regression Model

- A straight-line relating Y to X has the following form:
  - $f(X) = \beta_0 + \beta_1 X$
    - β_0 = intercept
    - β_1 = slope



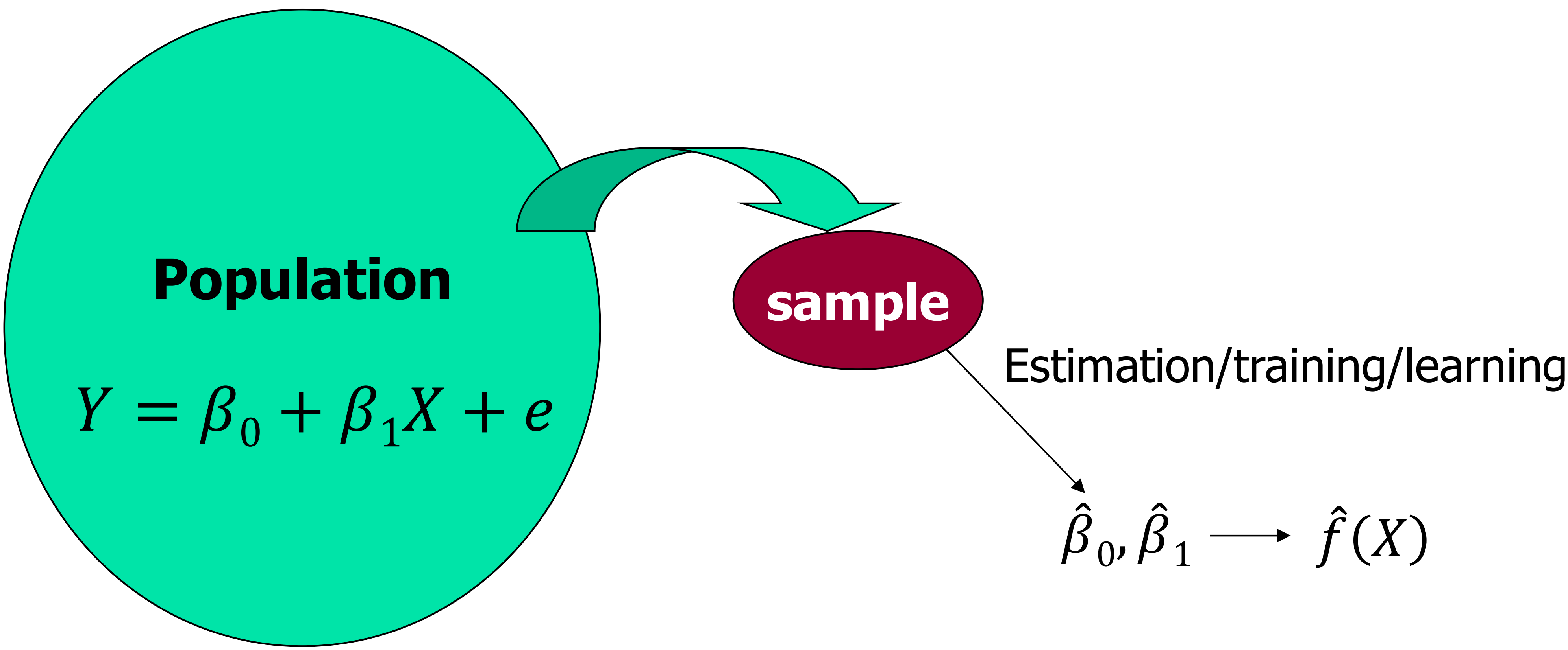
$$Y = f(X) + e$$
$$= \beta_0 + \beta_1 X + e$$

Simple Linear Regression Model

$$
\begin{aligned}
Y &= f(X) + e \\
&= \beta_0 + \beta_1 X + e
\end{aligned}
$$

- β_0 = Average value of Y when X = 0.
- β_1 = Amount by which Y changes on average when X changes by one unit.

Parameter Estimation - How to determine the best regression line?



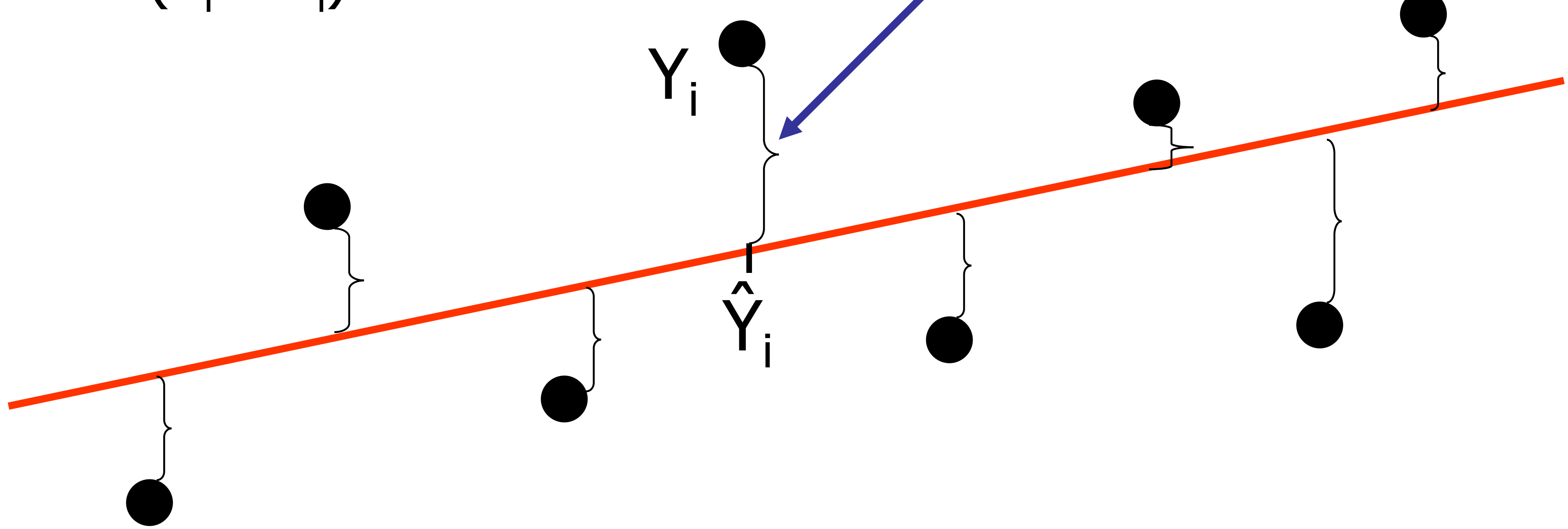
Training data/sample

Parameter Estimation - How to determine the best regression line?

- The "best" regression line is the one that comes the closest to the data points in the vertical direction. There are many ways to make this distance "as small as possible."
  - **The method of least squares** – the most common method. The least-squares regression line of Y on X is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.

Concept of Least Squares

vertical distance between a data point and a line (residual) $= (Y_i - \hat{Y}_i)$



Sum of the **S**quares of all residuals = SS(Residual) = $\sum_{i=1}^{N} \left(Y_i - \hat{Y}_i\right)^2$

Concept of Least Squares

- The least-squares regression line of Y on X is determined in such a way that it makes SS(Residual) as small as possible.

$$\text{SS(Residual)} = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \left(Y_i - \beta_0 - \beta_1 X_i \right)^2$$

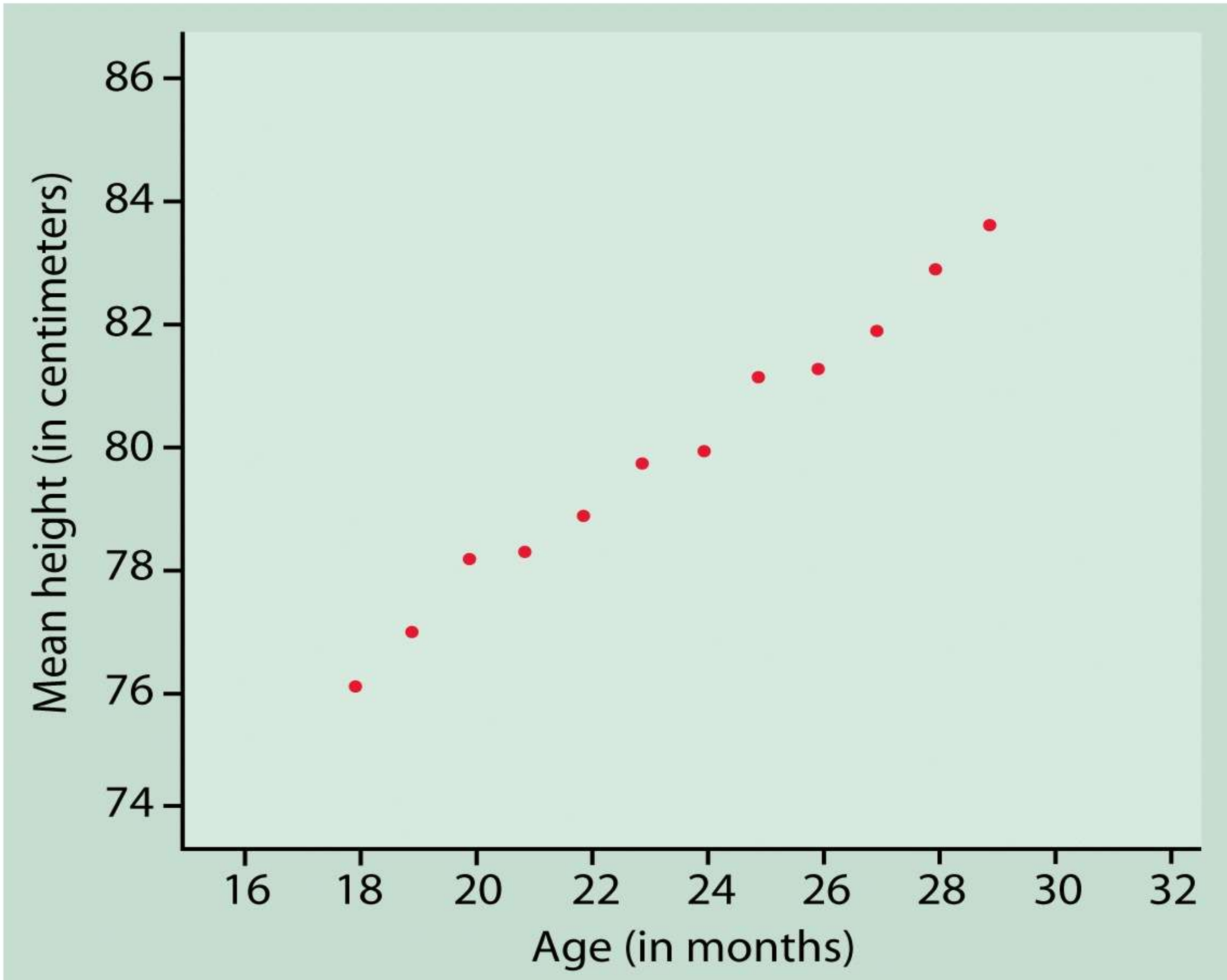
- β_0 and β_1 are estimated to minimize SS(Residual).
- In other words, this method aims to minimize the unexplained portion of Y by the regression line.

Least Squares Coefficient Estimates

$$\hat{\boldsymbol{\beta}}_1 = \frac{\sum_{i=1}^{N}\left(\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)\right)}{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2}$$

$$\hat{\boldsymbol{\beta}}_0 = \overline{Y} - \hat{\boldsymbol{\beta}}_1 \overline{X}$$

Example: LS Coefficient Estimates



Moore & McCabe (1998, p.136)

Example: LS Coefficient Estimates

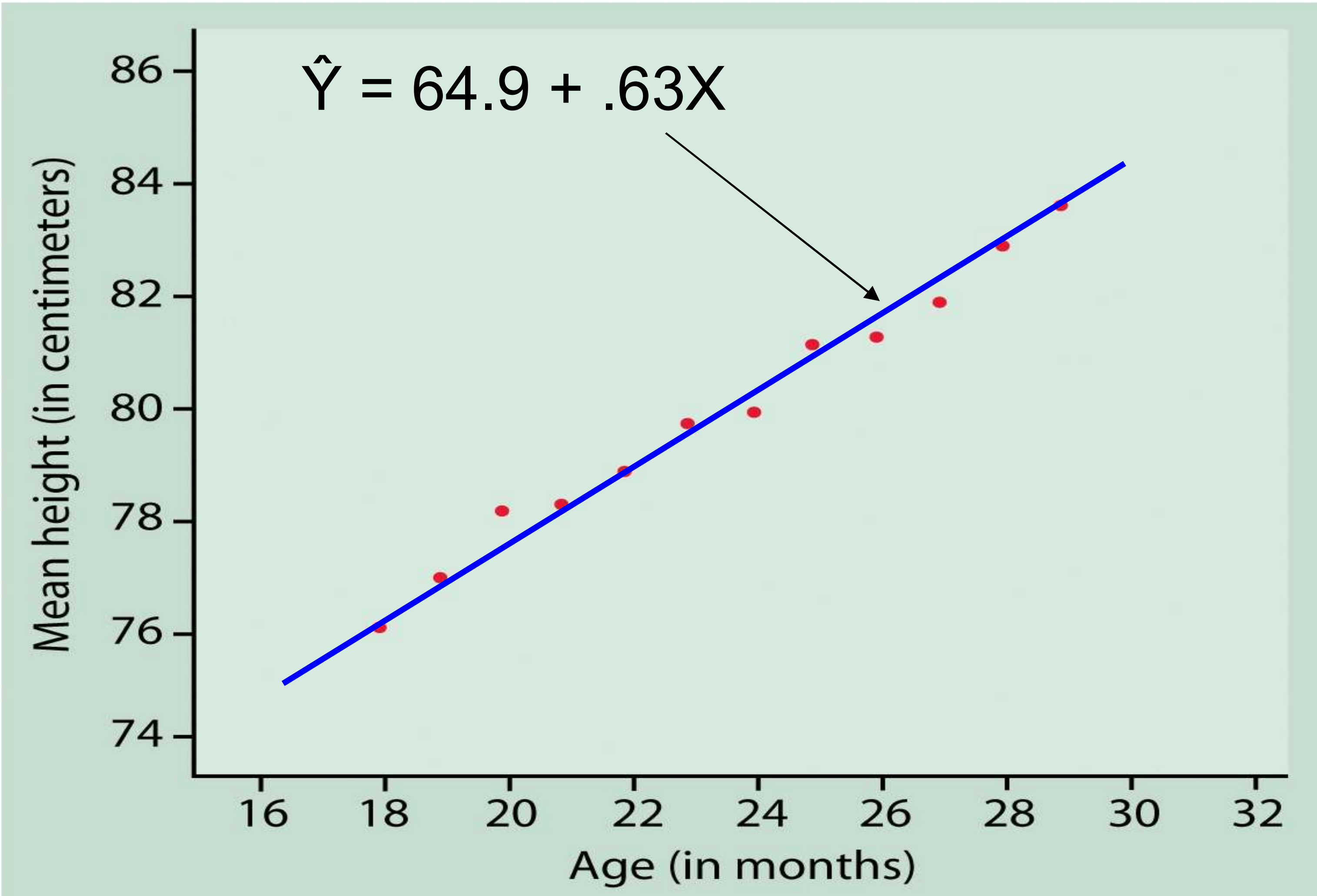
Age X (month)	Height Y (cm)
18	76.1
19	77.0
20	78.1
21	78.2
22	78.8
23	79.7
24	79.9
25	81.1
26	81.2
27	81.8
28	82.8
29	83.5

Moore & McCabe (1998, p.136)

- N = 12
- $\bar{X}$ = 23.5
- $\bar{Y}$ = 79.85

- $\hat{\beta}_1$=.6348
- $\hat{\beta}_0$= 64.93

Example: LS Coefficient Estimates



Moore & McCabe (1998, p.136)

Statistical test for the significance of slope

- We can perform a hypothesis test on the relationship between X and Y in simple linear regression (the effect of X on Y)
  - $H_0 : \beta_1 = 0$
    - There is no linear relationship between X and Y (no effect of X on Y)
  - $H_1 : \beta_1 \neq 0$
    - There is a linear relationship (an effect of X on Y)

Statistical test for the significance of slope

- We compute a t-statistic, given by

$$t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)}$$

where $SE(\hat{\beta}_1)$ is the standard error of the estimate.

Statistical test for the significance of slope

- $SE(\hat{\beta}_1)$ is computed as follows:

$$\mathrm{SE}(\widehat{\boldsymbol{\beta}}_1) = \sqrt{\frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{N-2}} \Bigg/ \sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

Statistical test for the significance of slope

- If the p-value of the t statistic is small enough, e.g., $p < \alpha = .05$ (or .01), we may reject the null hypothesis.

- This indicates that the slope is different from zero, suggesting a statistically significant effect of X on Y.

Statistical test for the significance of slope

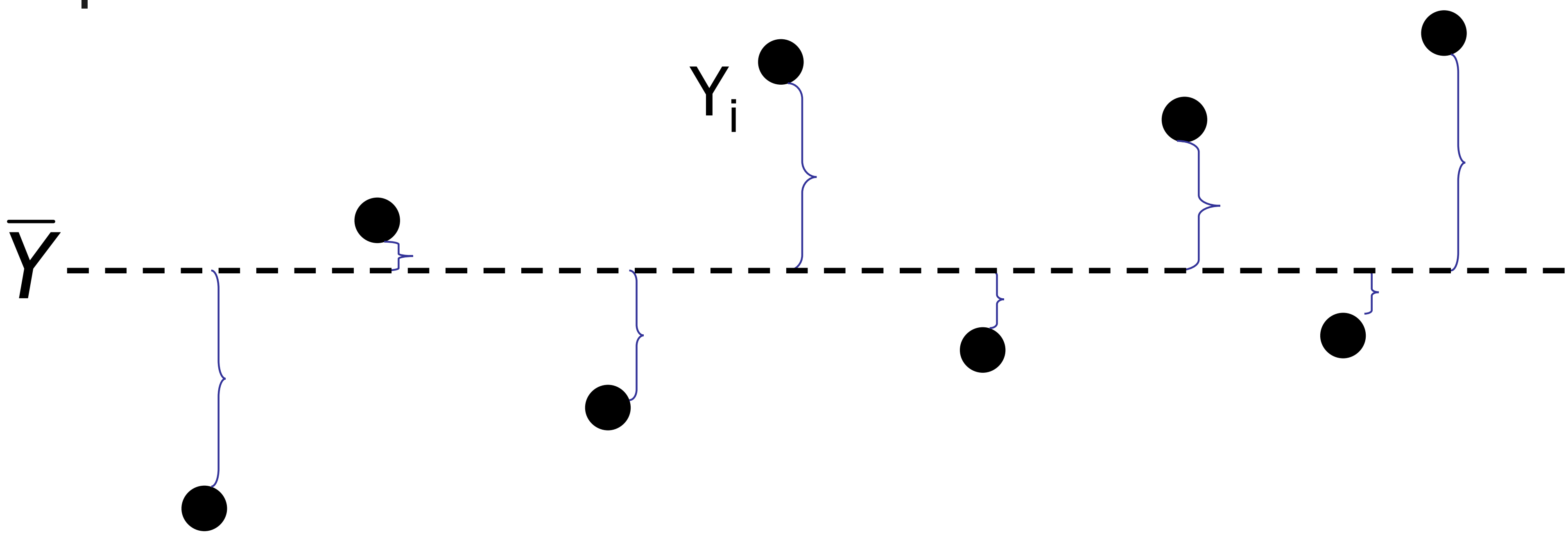
- To apply the *t* test for the slope, the following assumptions are required:
  - Normal distribution
  - Independent observations

Partitioning Variation: Simple Linear Regression

- As in ANOVA, we can also divide the variance/variation in the DV (Y) into different parts resulting from different sources.

- In regression analysis, the total variation in Y is partitioned into:
  - **SS(Regression):** The variation in Y explained by the regression line.
  - **SS(Residual):** The variation in Y unexplained by the regression line (residuals).

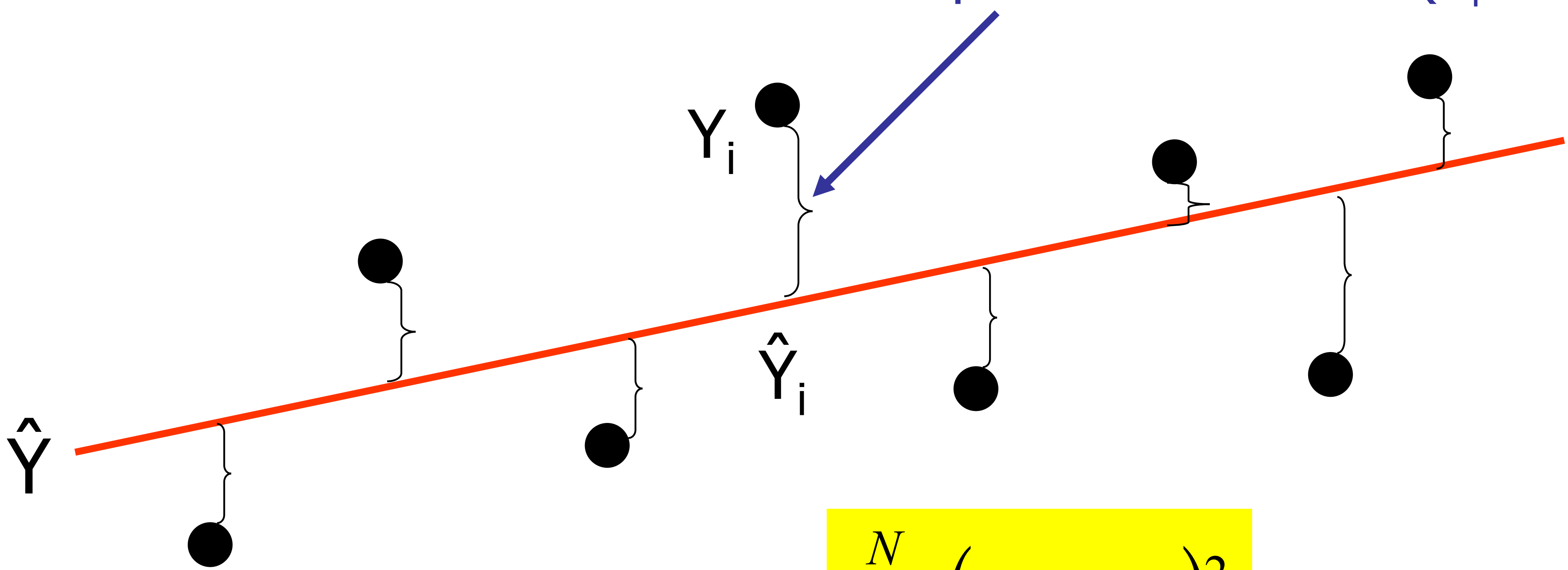
Partitioning Variation: SS(T)



$$\mathbf{SS(T)} = \sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2$$

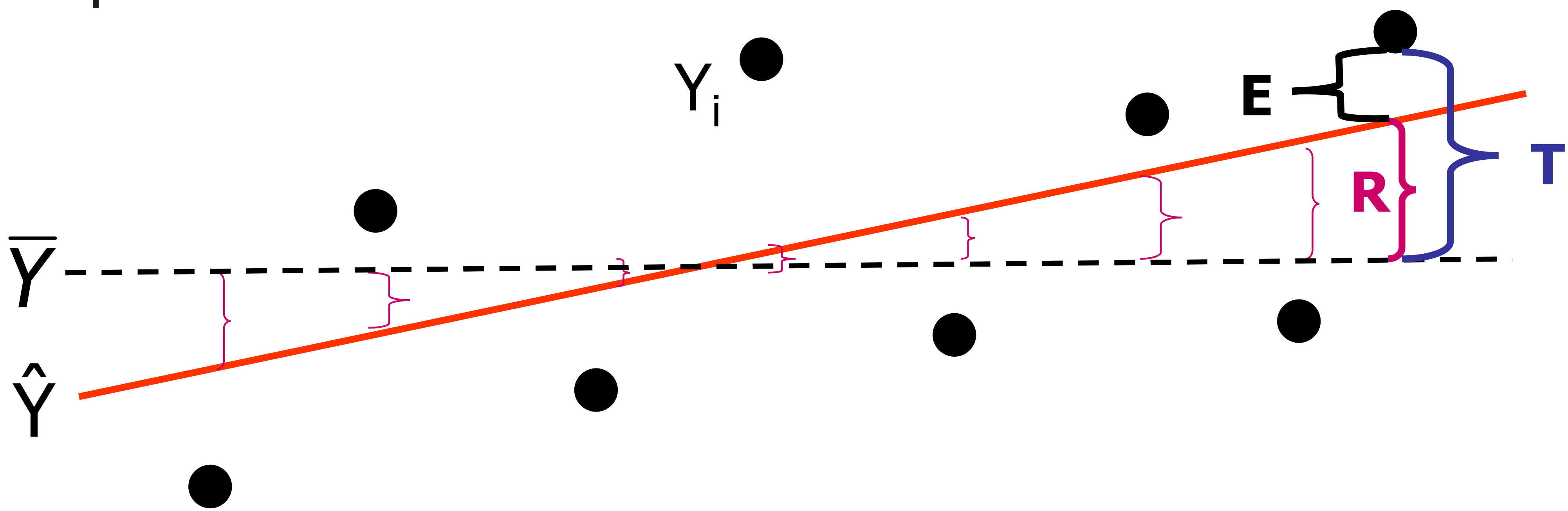
Partitioning Variation: SS(Residual)

Vertical distance between a data point and a line = $(Y_i - \hat{Y}_i)$



SS(Residual) = $\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2$

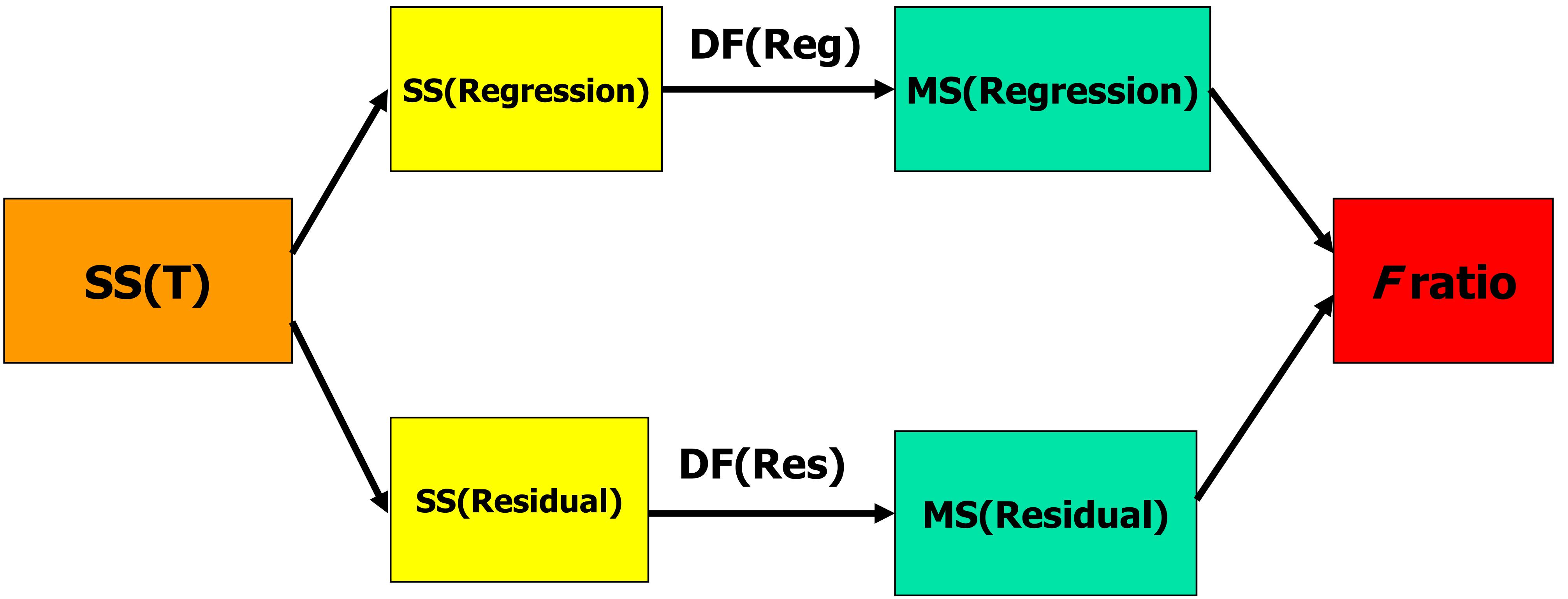
Partitioning Variation: SS(Regression)



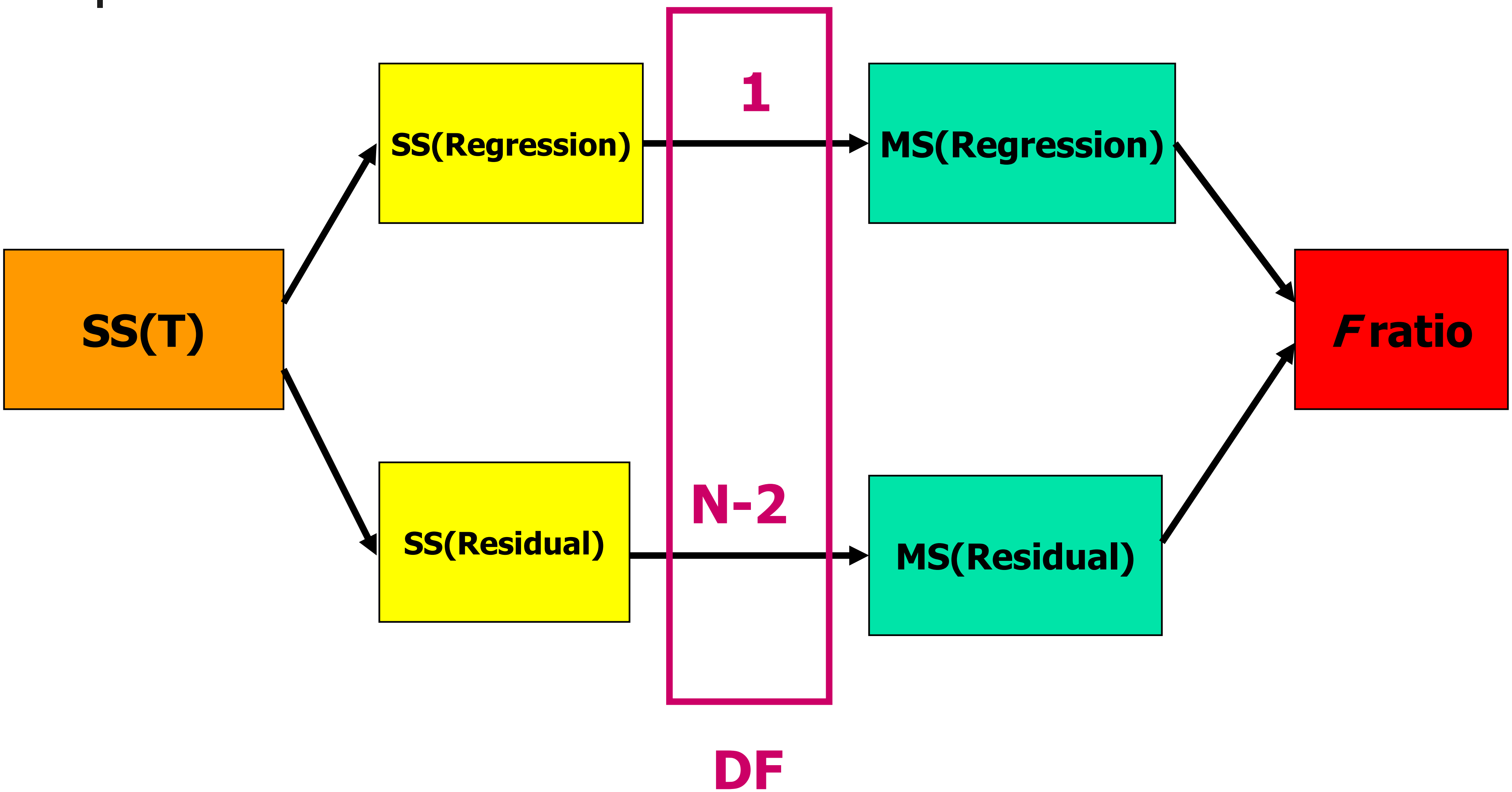
SS(Regression) = SS(T) – SS(Residual)

$$= \sum_{i=1}^{N} \left(\hat{Y}_i - \bar{Y} \right)^2$$

Simple Linear Regression – ANOVA Table



Simple Linear Regression – ANOVA Table



Simple Linear Regression

- The F statistic is used for testing
  - $H_0 : \beta_1 = 0$
  - $H_1 : \beta_1 \neq 0$

- If the observed F value is greater than a critical value of F with DF(Reg) and DF(Res) at $\alpha = .05$, we may reject H_0.

- This is the same as using the t test for β_1.
  - In fact, $t^2 = F$

Assessing the goodness-of-fit of the regression model: Coefficient of Determination (R^2)

- Proportion of the total variation in Y accounted for by the regression model.

$$R^2 = 1 - \frac{\text{SS(Residual)}}{\text{SS(Total)}}$$

- Ranges from 0 to 1.
  - The larger R^2, the more variance of Y explained
  - 0 = No explanation at all
  - 1 = Perfect explanation

- In simple linear regression, $r = \sqrt{R^2}$

Assessing the goodness-of-fit of the model: Residual Standard Error (RSE)

- The residual standard error (RSE) is an estimate of the standard deviation of e.

$$\mathrm{RSE} = \sqrt{\frac{1}{N-P-1} SS(\mathrm{Residual})}$$

- Roughly speaking, it is the average amount that the response deviates from the true regression line.
  - RSE is considered a measure of the lack of fit of the model.

Assessing the goodness-of-fit of the model: Mean Squared Error (MSE)

- The mean squared error (MSE) is the mean of the sum of squared residuals, i.e., it measures the average of the squares of the errors.

$$\mathrm{MSE} = \frac{1}{N} SS(\mathrm{Residual})$$

- Root mean squared error (RMSE) is the square root of MSE:

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}}$$

Example: Simple Linear Regression

- Wine consumption and heart attacks (Data from M. H. Criqui, reported in *New York Times*, Dec. 12., 1994) – wineheartattach.csv

TABLE 2.2 Wine consumption and heart attacks

Country	Alcohol from wine	Heart disease deaths	Country	Alcohol from wine	Heart disease deaths
Australia	2.5	211	Netherlands	1.8	167
Austria	3.9	167	New Zealand	1.9	266
Belgium	2.9	131	Norway	0.8	227
Canada	2.4	191	Spain	6.5	86
Denmark	2.9	220	Sweden	1.6	207
Finland	0.8	297	Switzerland	5.8	115
France	9.1	71	United Kingdom	1.3	285
Iceland	0.8	211	United States	1.2	199
Ireland	0.7	300	West Germany	2.7	172
Italy	7.9	107			

Moore & McCabe (1998, p.119)

Example: Simple Linear Regression

```
Call:
lm(formula = heartattack ~ wine, data = mydata)

Residuals:
   Min     1Q Median     3Q    Max 
-62.95 -25.91 -12.35  26.97  55.52 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  260.563     13.835  18.833 7.97e-13 ***
wine         -22.969      3.557  -6.457 5.91e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 37.88 on 17 degrees of freedom
Multiple R-squared:  0.7103,	Adjusted R-squared:  0.6933 
F-statistic: 41.69 on 1 and 17 DF,  p-value: 5.913e-06
```

Example: Simple Linear Regression

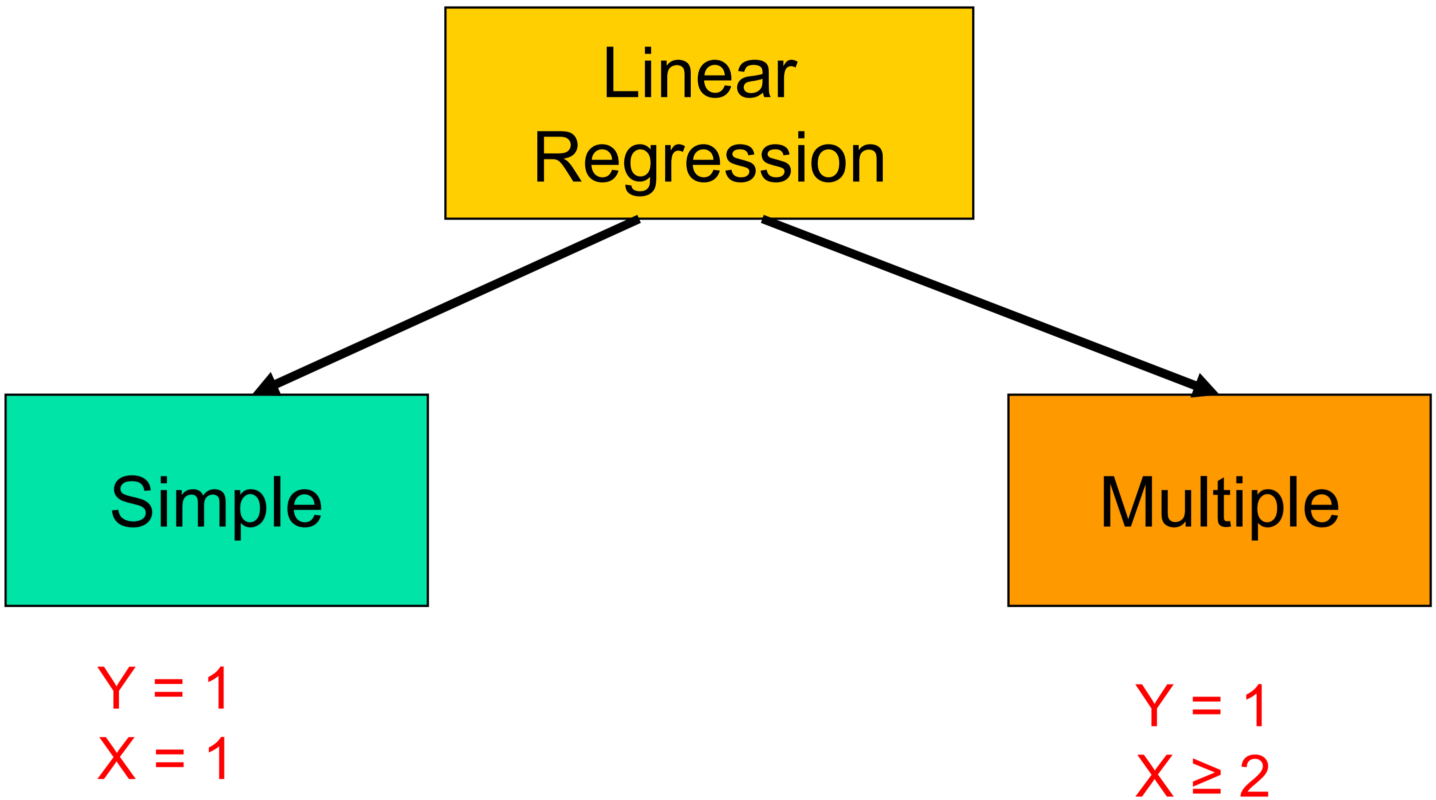
- This model summary table shows the goodness of fit of the fitted regression model/line. It indicates that about 71% of the total variance of the DV (heart disease deaths) is explained by the predictor (wine) ($R^2 = .71$).
- We may reject the null hypothesis that the population slope is zero because the p-value of the *t* statistic is less than .05 ($t = -6.457$, $p < .00$).
  - Note that in simple regression analysis, both F and t tests are used for testing the significance of the single slope, resulting in the same conclusion.

Example: Simple Linear Regression

- This indicates that the slope is statistically significantly different from zero, suggesting a statistically significant and negative effect of wine on heart disease deaths.

- More specifically, the estimated slope (wine = -22.97) indicates that as wine consumption increases by one unit, heart disease deaths decrease by 22.97 units on average.

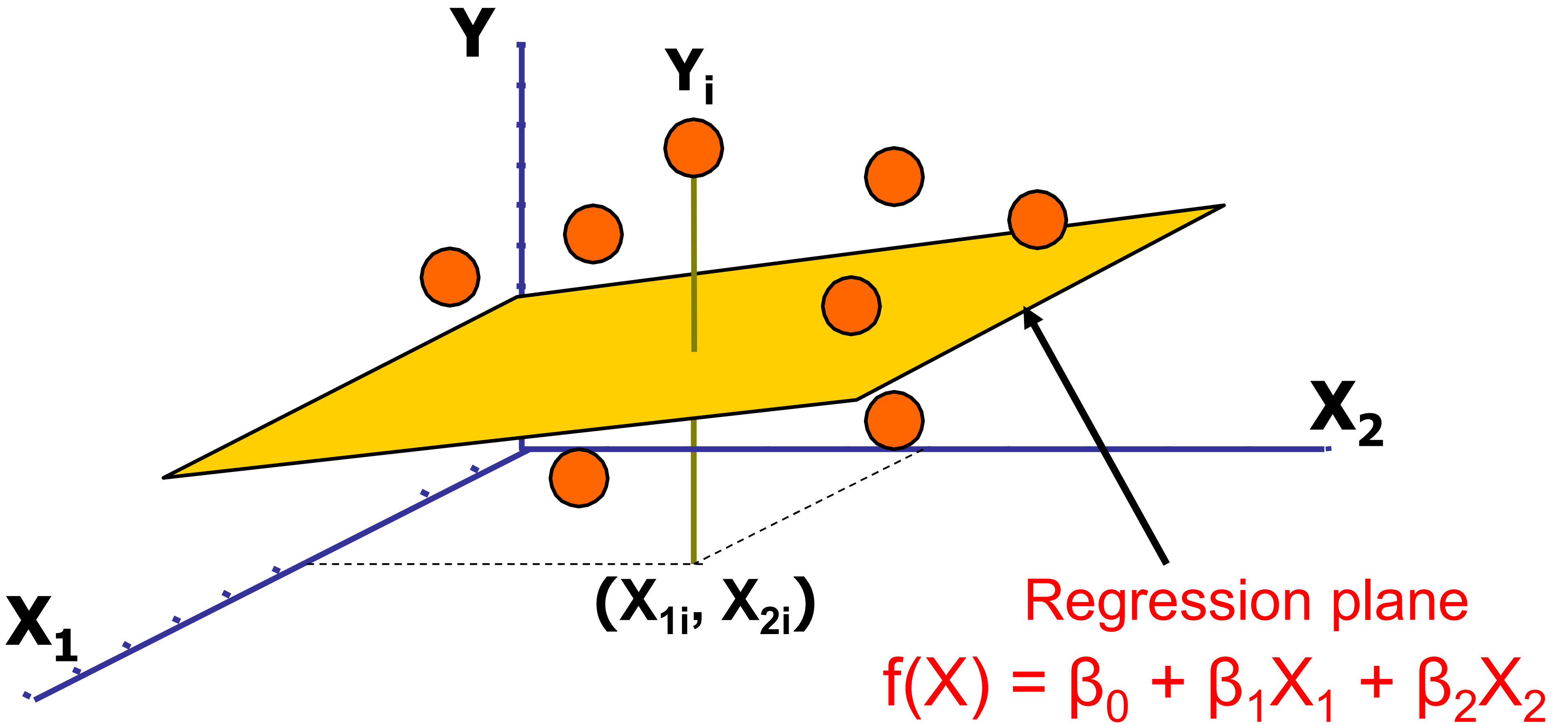
Multiple Linear Regression



Multiple Linear Regression

- Describes how the DV (Y) changes as multiple predictors (X_P) ($P \geq 2$) change.

Multiple Linear Regression: P = 2



Multiple Linear Regression Model

- Linear relation between a continuous DV and P predictors ($P \geq 2$):

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_P X_P$$

- Predictors can be continuous or discrete.

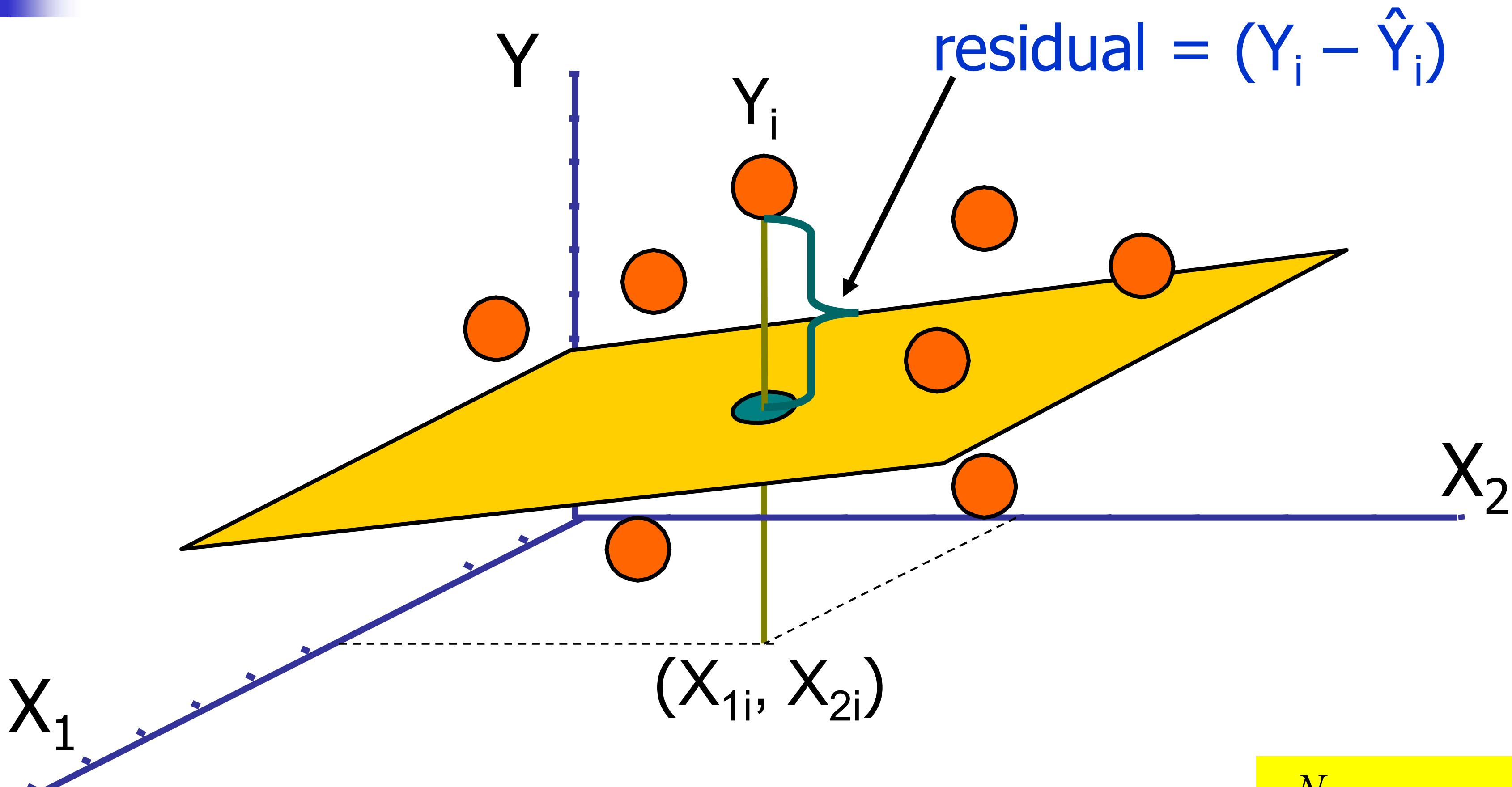
Multiple Linear Regression Model

- Intercept (β_0):
  - Average value of *Y* when all predictors are zero.
- Slope (β_p):
  - Amount by which Y changes on average when X_p changes by one unit, holding the other predictors constant (or controlling for the other predictors).
    - Example: If $\beta_1 = 2$, on average, Y is expected to increase by 2 for each 1 unit increase in X_1 with the other predictors constant.

How to estimate coefficients?

- As in simple linear regression, the method of least squares can be used for estimating the coefficients in the regression model.
  - This method chooses the values of the intercept and slopes that make the sum of the squared residuals as small as possible.

Multiple Linear Regression: Least Squares



Sum of the Squares of all residuals $= SS(\text{Residual}) =$

$$\sum_{i=1}^{N} \left(Y_i - \hat{Y}_i\right)^2$$

Multiple Linear Regression: Least Squares

- In other words, the coefficients are estimated to minimize SS(Residual).

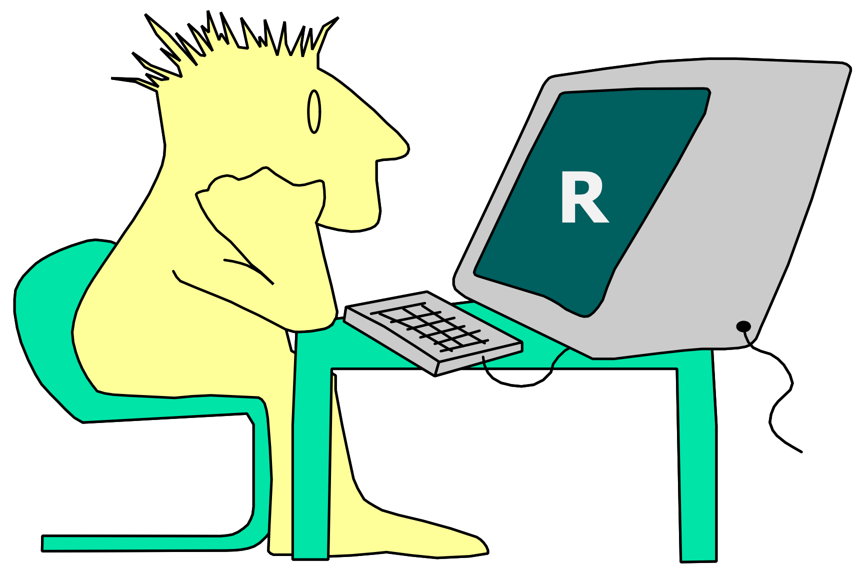
SS(Residual)

$$= \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{N} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_1 - \cdots - \hat{\beta}_P X_P)^2$$

- The computations for obtaining the least squares estimates are complicated…

Multiple Linear Regression: Least Squares

Just let software do the computations!



Partitioning Variation in Multiple Linear Regression

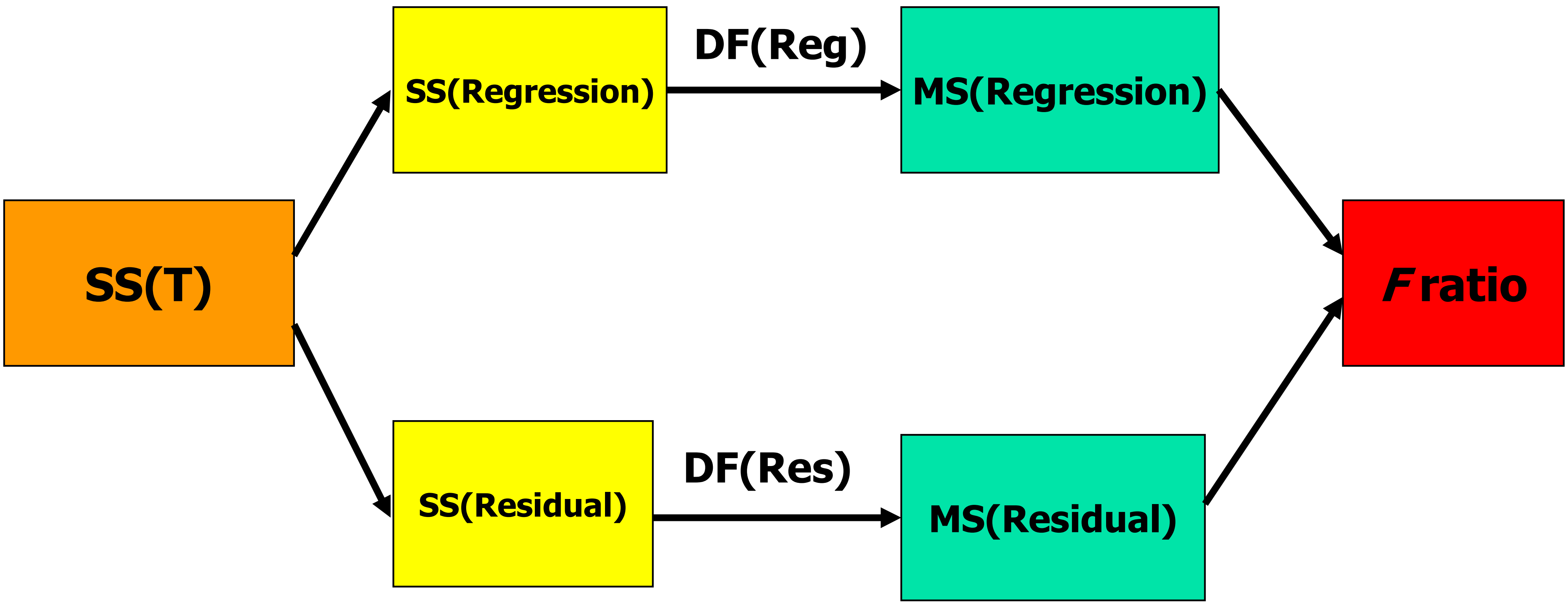
- As in simple regression analysis, the total variation in Y (SS(T)) is partitioned into:
  - SS(Regression): The variation in Y explained by the regression model.
  - SS(Residual): The variation in Y unexplained by the regression model.

$$SS(T) = \sum_{i=1}^{N} \left(Y_i - \bar{Y}\right)^2$$

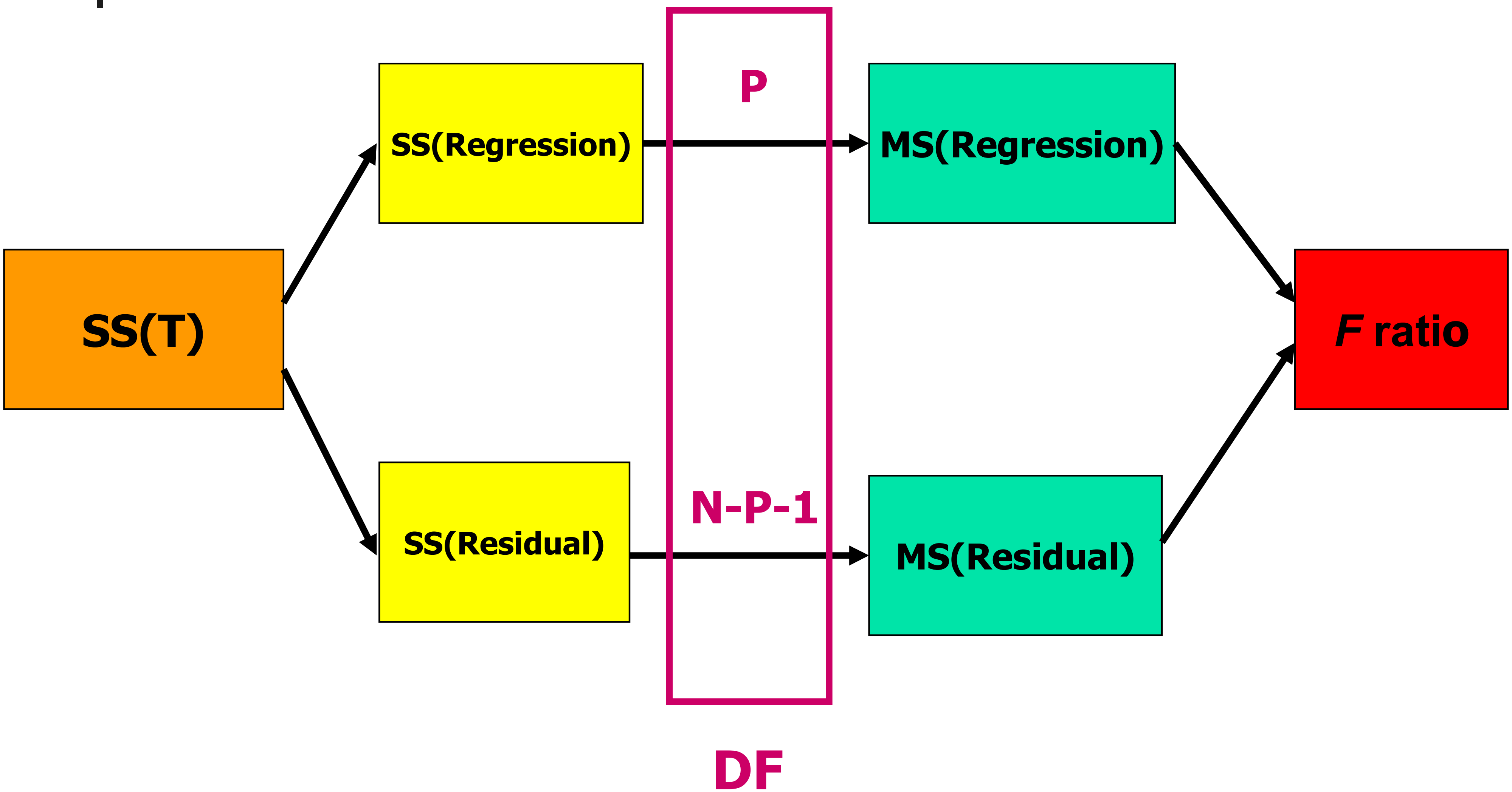
$$SS(Res) = \sum_{i=1}^{N} \left(Y_i - \hat{Y}_i\right)^2$$

$$SS(Reg) = \sum_{i=1}^{N} \left(\hat{Y}_i - \bar{Y}\right)^2$$

Multiple Linear Regression – ANOVA Table



Multiple Linear Regression – ANOVA Table



Testing Overall Significance

- The F statistic is used to examine if there is a linear relationship between all X variables together and Y.
- Hypotheses
  - $H_0: \beta_1 = \beta_2 = \ldots = \beta_P = 0$
    - None of the Xs are linearly related to Y.
  - H_1: At least one coefficient is not 0
    - At least one X is linearly related to Y

Testing Overall Significance

- If the observed value of F is greater than the critical value of $F(DF(Reg), DF(Res))$ at $\alpha = .05$, we may reject the null hypothesis.

- This indicates that at least one regression coefficient is statistically significantly different from zero.

Testing Individual coefficients

- $H_0 : \beta_p = 0$
  - X_p is not linearly related to Y (no linear relationship)
- $H_1 : \beta_p \neq 0$

- We can apply a t test for testing the significance of an individual coefficient.

Assessing the goodness-of-fit of the regression model: Coefficient of Determination (R^2)

- Proportion of the total variation in Y accounted for by the regression model.

$$R^2 = 1 - \frac{\text{SS(Residual)}}{\text{SS(Total)}}$$

- Also called the Squared Multiple Correlation (SMC).
  - Ranges from 0 to 1.
  - The larger R^2, the more variance of Y explained
  - 0 = No explanation at all
  - 1 = Perfect explanation

Coefficient of Determination (R^2)

- If a new predictor is added to the model, how will it affect the R^2 value for the model?
  - SS(Residual) will always be smaller by adding more predictors (although the amount of decrease may be negligible).
  - In other words, SS(Regression) will become always larger.
  - Thus, R^2 never decreases when a new predictor is added to the model.

Adjusted R^2

- R^2 never decreases when a new predictor is added to model
  - Not attractive when deciding a reasonable set of predictors (or model comparisons)
- Solution = Adjusted R^2

$$\text{Adjusted } R^2 = 1 - (1 - R^2)\left[\frac{N-1}{N-P-1}\right] \le R^2$$

  - If no substantial increase in R^2 is obtained by adding a new predictor, adjusted R^2 tends to decrease.
    - This index prefers a more parsimonious model when different models provide a similar explanation power (R^2).

Assessing the goodness-of-fit of the model: Residual Standard Error (RSE)

- The residual standard error (RSE) is an estimate of the standard deviation of e.

$$\mathrm{RSE} = \sqrt{\frac{1}{N - P - 1} SS(\mathrm{Residual})}$$

- Roughly speaking, it is the average amount that the response deviates from the true regression line.
  - RSE is considered a measure of the lack of fit of the model.

Assessing the goodness-of-fit of the model: Mean Squared Error (MSE)

- The mean squared error (MSE) is the mean of the sum of squared residuals, i.e., it measures the average of the squares of the errors.

$$\mathrm{MSE} = \frac{1}{N}\mathit{SS}(\mathrm{Residual})$$

- Root mean squared error (RMSE) is the square root of MSE:

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}}$$

Example: Multiple Linear Regression

- The children's antisocial behaviour data: Part of the National Longitudinal Survey of Youth (NLSY) reported in Curran (1998) (curran_training.csv).
  - In the NLSY, a large sample of children and their mothers were administered a set of assessment instruments every other year starting from 1986 to 1992.
  - From the original NLSY sample, Curran (1998) selected 221 pairs of children and mothers based on three selection criteria. First, children must have aged between 6 and 8 years at the first time point of assessment. Second, they had to complete interviews at all four time points. Finally, only one biological child was considered from each mother.
  - The average child's age was 6.9 years (SD = .62) and the average mother's age was 25.5 years (SD = 1.87) at the first time point.
  - We used 186 pairs as a training sample and 35 pairs as a test sample.

Example: Multiple Linear Regression

- DV = The antisocial behaviour of children measured at the first time point (0-12)

- Predictors (measured at the first time point):
  - Gender (female = 0 and male = 1)
  - Cognitive stimulation for children at home (0-14)
  - Emotional support for children at home (0-13)

- N = 186

Example: Multiple Linear Regression

```
Call:
lm(formula = anti1 ~ gender + cogstm + emotsup, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5089 -1.1187 -0.3290  0.9261  5.4064 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.619633   0.560243   4.676 5.69e-06 ***
gender       0.646018   0.229772   2.812  0.00547 ** 
cogstm       0.008342   0.049073   0.170  0.86521    
emotsup     -0.159691   0.053881  -2.964  0.00345 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 1.557 on 182 degrees of freedom
Multiple R-squared:  0.08062, Adjusted R-squared:  0.06546 
F-statistic:  5.32 on 3 and 182 DF,  p-value: 0.001549
```

Example: Multiple Linear Regression

```
Call:
lm(formula = anti1 ~ gender + emotsup, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5291 -1.1019 -0.3367  0.8981  5.4285 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.66742    0.48330   5.519 1.15e-07 ***
gender       0.64441    0.22897   2.814  0.00542 ** 
emotsup     -0.15656    0.05049  -3.101  0.00224 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.553 on 183 degrees of freedom
Multiple R-squared:  0.08047, Adjusted R-squared:  0.07042 
F-statistic: 8.007 on 2 and 183 DF,  p-value: 0.0004637
```

Linear Regression

- Two main goals:
  - To investigate the relationship between a DV and multiple predictors.
  - To find the best prediction equation for a DV regardless of the meaning of predictors in the equation.

Linear Regression: Prediction

- Once we have fit the regression model, it is straightforward to use $\hat{Y}$ to predict the response Y based on a set of values for the predictors.

$$Y = \hat{Y} + e$$

Example - Linear Regression: Prediction

- The original antisocial behaviour data were divided into training (curran_training.csv, N = 186) and test (curran_test.csv, N = 35) samples.
- We applied the same linear regression model to the training data to estimate the regression coefficients.
- Then, we used the coefficient estimates to predict new observations of the DV in the test sample and calculated RSE, MSE, and RMSE for the test sample.

Example - Linear Regression: Prediction

```
Call:
lm(formula = anti1 ~ gender + cogstm + emotsup, data =
mydata_training)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5089 -1.1187 -0.3290  0.9261  5.4064

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.619633   0.560243   4.676 5.69e-06 ***
gender       0.646018   0.229772   2.812  0.00547 **
cogstm       0.008342   0.049073   0.170  0.86521
emotsup     -0.159691   0.053881  -2.964  0.00345 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.557 on 182 degrees of freedom
Multiple R-squared:  0.08062, Adjusted R-squared:  0.06546
F-statistic:  5.32 on 3 and 182 DF,  p-value: 0.0015494
```

Example - Linear Regression: Prediction

```
mydata_test$anti1
[1] 3 0 0 0 1 0 1 2 3 0 1 1 0 0 0 2 2 1 2 2 1 2 0 3 0 3 1 0 1 1 1 2 2 2 0

pred_y_test
        1         2         3         4         5         6         7         8
9        10        11
1.9035081 1.4338636 0.6437511 0.9631327 2.0559438 1.2491482 2.2228897 1.6091511
2.0965651 0.9047419 2.4922221
       12        13        14        15        16        17        18        19
20        21        22
1.7688419 1.2741728 1.2408066 0.7784172 1.7438172 2.0559438 0.7784172 1.1061404
1.5674433 1.4244356 1.7271342
       23        24        25        26        27        28        29        30
31        32        33
1.7521588 0.8034419 1.4244356 1.6008095 0.9381081 1.6008095 0.9130834 1.5757849
1.4077525 1.7605003 2.0298327
       34        35
0.9214250 1.92019

RSE
[1] 1.112484
MSE
[1] 1.096179
RMSE
[1] 1.046986
```

Nominal Predictors in Linear Regression

- We can use a nominal variable (with more than two categories/levels) as a predictor in linear regression.
- However, unlike a binary or continuous variable, this variable cannot be added to the regression model just as it is.
- Instead, a nominal variable needs to be recoded into a set of variables (e.g., dummy variables) which can then be added to the regression model.
  - A nominal variable is **dummy-coded**.

Dummy Coding

- Dummy coding involves the assignment of binary values (0 or 1) to represent membership in each level of a nominal variable.
  - Expresses group memberships of observations using only zeros and ones.
  - The number of dummy variables is 1 less than that of levels of the nominal variable.

Dummy Coding

- For example, if a nominal variable has three groups/categories (1, 2, & 3), this variable is represented by two dummy variables.

Dummy variables

subjects	X	D_1	D_2
	1	1	0
	2	0	1
	3	0	0
	1	1	0

Original nominal variable

Steps of Dummy Coding

- Step 1: Create K – 1 new variables as dummy variables, where K = number of groups/categories.
- Step 2: Choose one of K groups as a baseline (a group against which all other groups will be compared). Usually this is a control group.
- Step 3: Assign the baseline group values of 0 for all dummy variables.
- Step 4: For the *k*th dummy variable (k = 1, …,K-1), assign the value 1 to the *k*th group. Assign all other groups 0 for this variable.

Dummy Coding

Y	SES
18	H
19	H
15	H
23	H
11	H
11	M
16	M
13	M
12	M
10	M
4	L
8	L
10	L
9	L
6	L

	Y	D1	D2
G1	18	1	0
	19	1	0
	15	1	0
	23	1	0
	11	1	0
G2	11	0	1
	16	0	1
	13	0	1
	12	0	1
	10	0	1
G3	4	0	0
	8	0	0
	10	0	0
	9	0	0
	6	0	0

Dummy Coding

- In this example, we can have a linear regression equation for each observation (i = 1,…,N), as follows.

$$Y_i = \beta_0 + \beta_1 D1_i + \beta_2 D2_i + e_i$$

	Y	D1	D2
G1	18	1	0
	19	1	0
	15	1	0
	23	1	0
	11	1	0
G2	11	0	1
	16	0	1
	13	0	1
	12	0	1
	10	0	1
G3	4	0	0
	8	0	0
	10	0	0
	9	0	0
	6	0	0

Dummy Coding

- In this regression equation,
  - β_0 = Mean on Y of G3 (baseline group)
  - β_1 = difference between the means of G1 and G3 on Y
  - β_2 = difference between the means of G2 and G3 on Y

	Y	D1	D2
G1	18	1	0
	19	1	0
	15	1	0
	23	1	0
	11	1	0
G2	11	0	1
	16	0	1
	13	0	1
	12	0	1
	10	0	1
G3	4	0	0
	8	0	0
	10	0	0
	9	0	0
	6	0	0

Example: Nominal Predictors in Linear Regression

- Field (2005) (GlastonburyDummy.csv): A biologist is interested in the potential health effects of music festivals. She went to a music festival and measured the hygiene of concert goers over the three days, which ranges from 0 (you smell like you’ve bathed in sewage) to 5 (you smell of freshly baked bread).

Example: Nominal Predictors in Linear Regression

- DV (change): the change in hygiene over the three days

- Predictor (music affiliation):
  - 1 = alternative (indie kid)
  - 2 = heavy metal (metaller)
  - 3 = hippy/folky/ambient (crusty)
  - 4 = others (no musical affiliation)

Example: Nominal Predictors in Linear Regression

```
Call:
lm(formula = change ~ music, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-1.81596 -0.49385  0.07615  0.42615  1.60404

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.56404    0.09144  -6.169 1.02e-08 ***
musicindiekid -0.40025    0.20592  -1.944   0.0543 .
musicmetaller  0.01788    0.16337   0.109   0.9130
musiccrusty   -0.40180    0.16798  -2.392   0.0184 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6903 on 117 degrees of freedom
Multiple R-squared:  0.07191, Adjusted R-squared:  0.04811
F-statistic: 3.022 on 3 and 117 DF,  p-value: 0.03254
```

Lab: Linear Regression I

1. Simple Linear Regression – wineheartattach.csv
2. Multiple Linear Regression – curran_training.csv
3. Multiple Linear Regression with Dummy Variables – GlastonburyDummy.csv