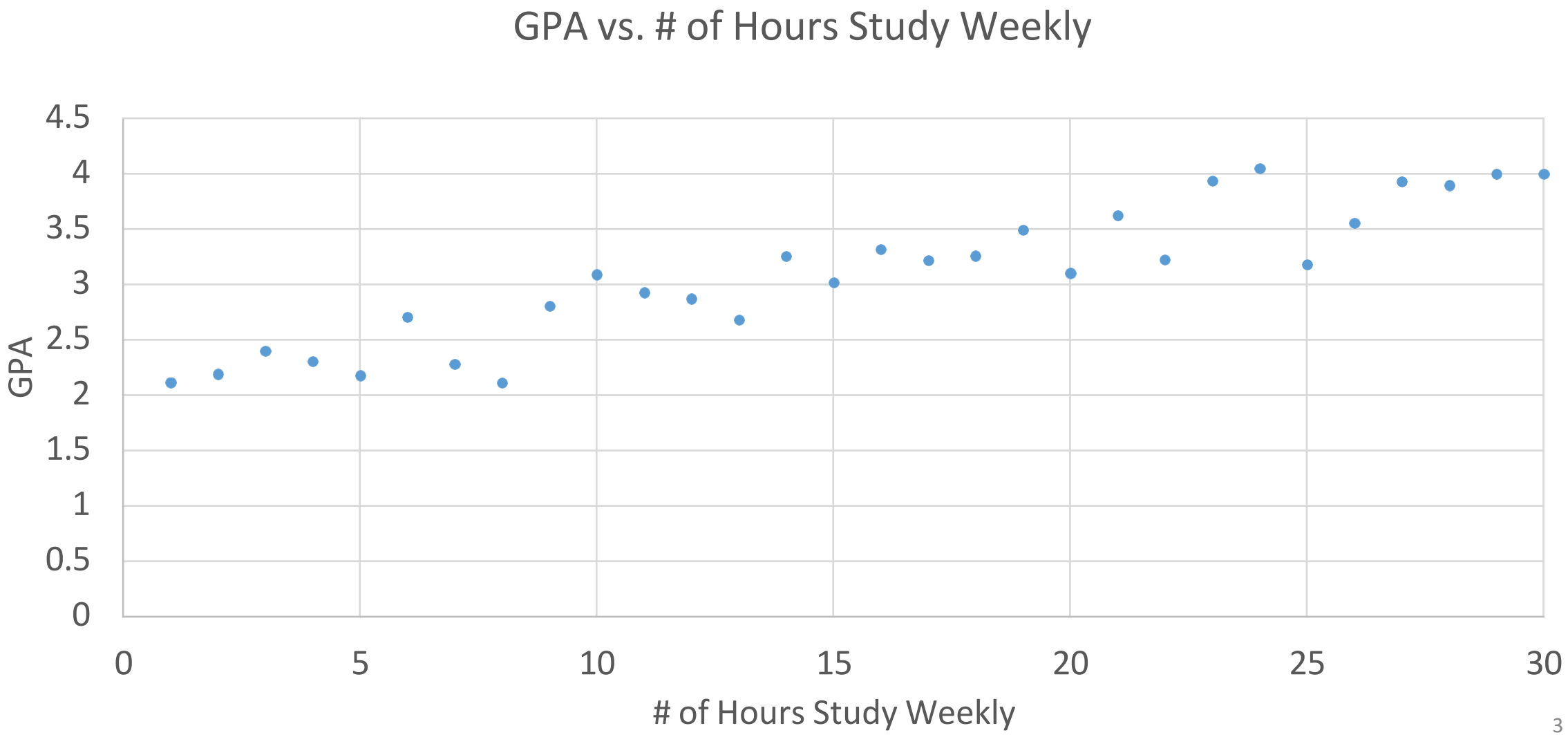
Introduction to Linear Regression

Kazi Aminul Islam
Department of Computer Science
Kennesaw State University

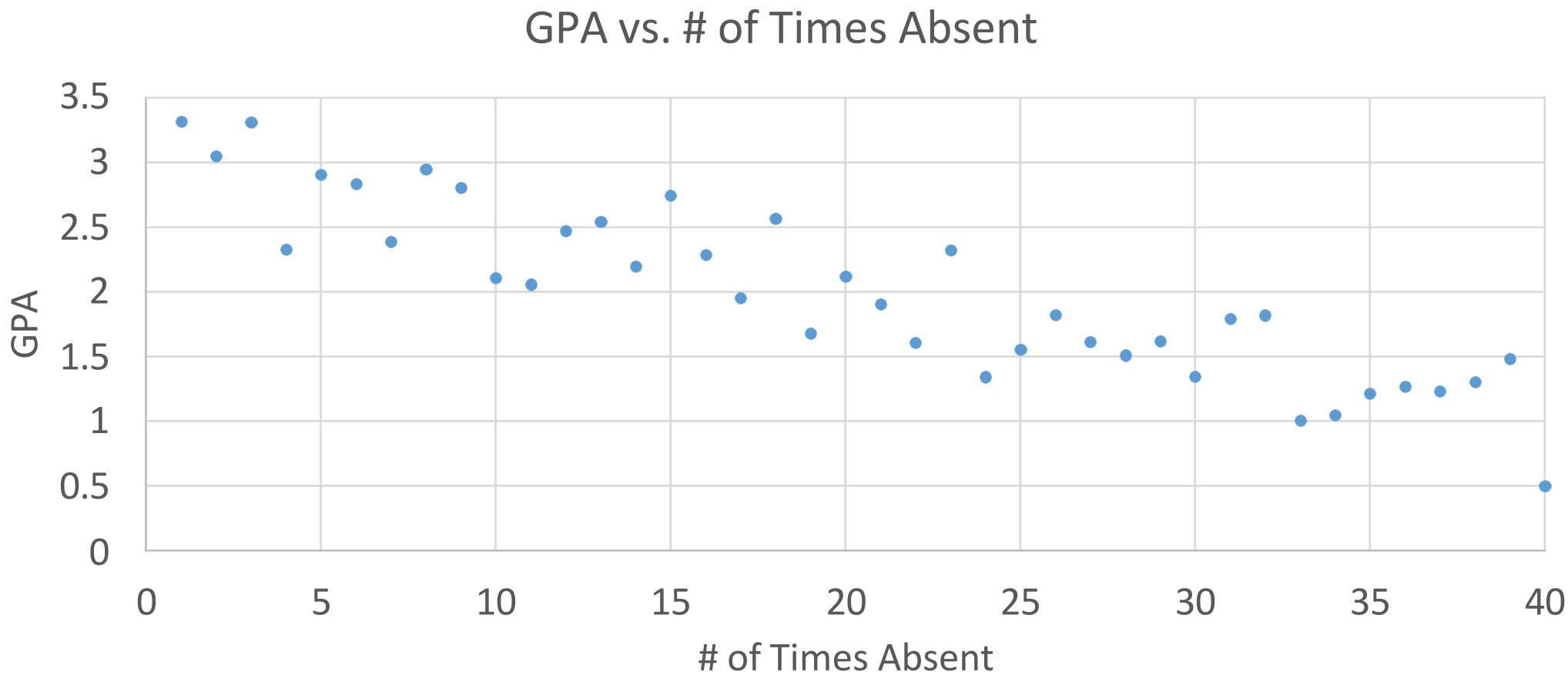
Learning Objectives

1. Scatterplots
2. Describe the Linear Regression Model
3. Strength and Significance of Linear Relations
4. State the Regression Modeling Steps
5. Explain Ordinary Least Squares
6. Compute Regression Coefficients
7. Goodness of Fit

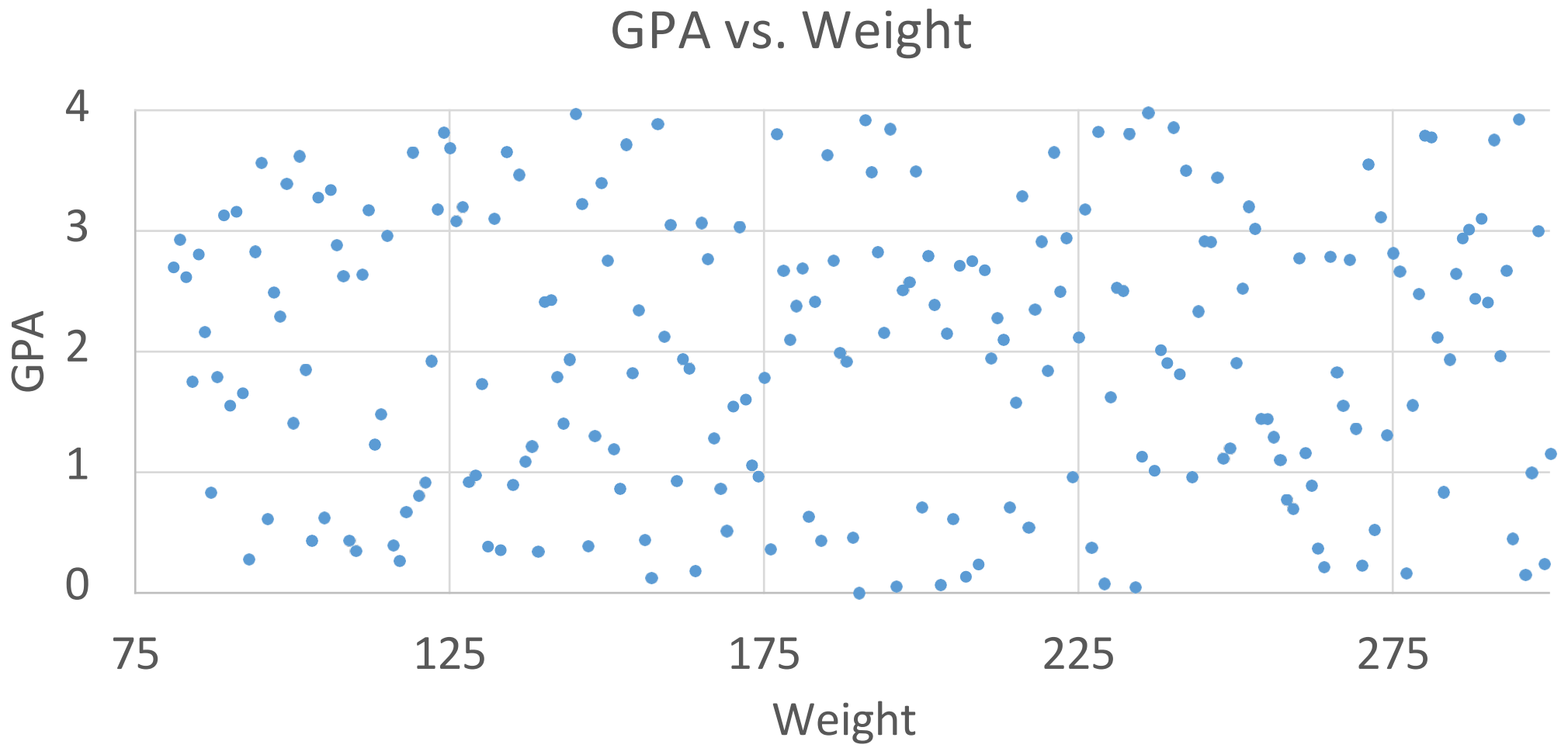
Observations



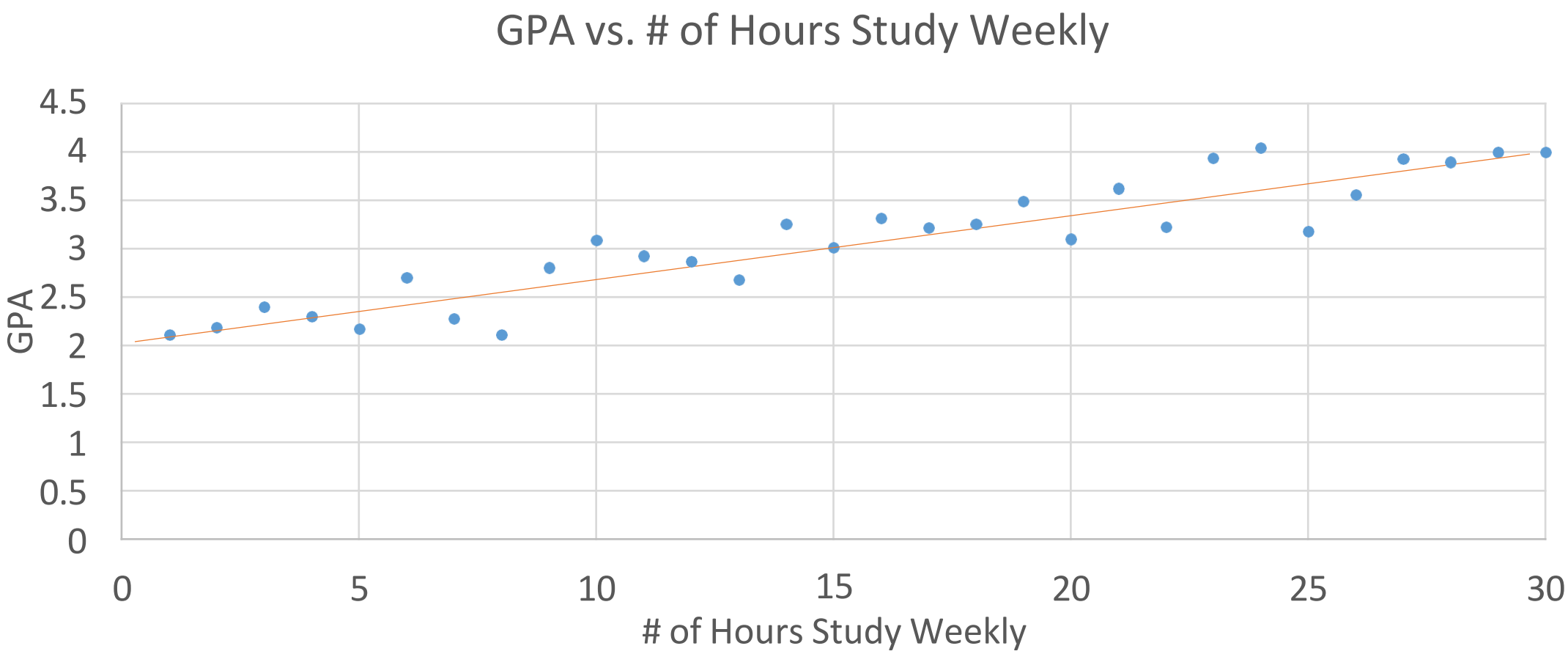
Observations (cont.)



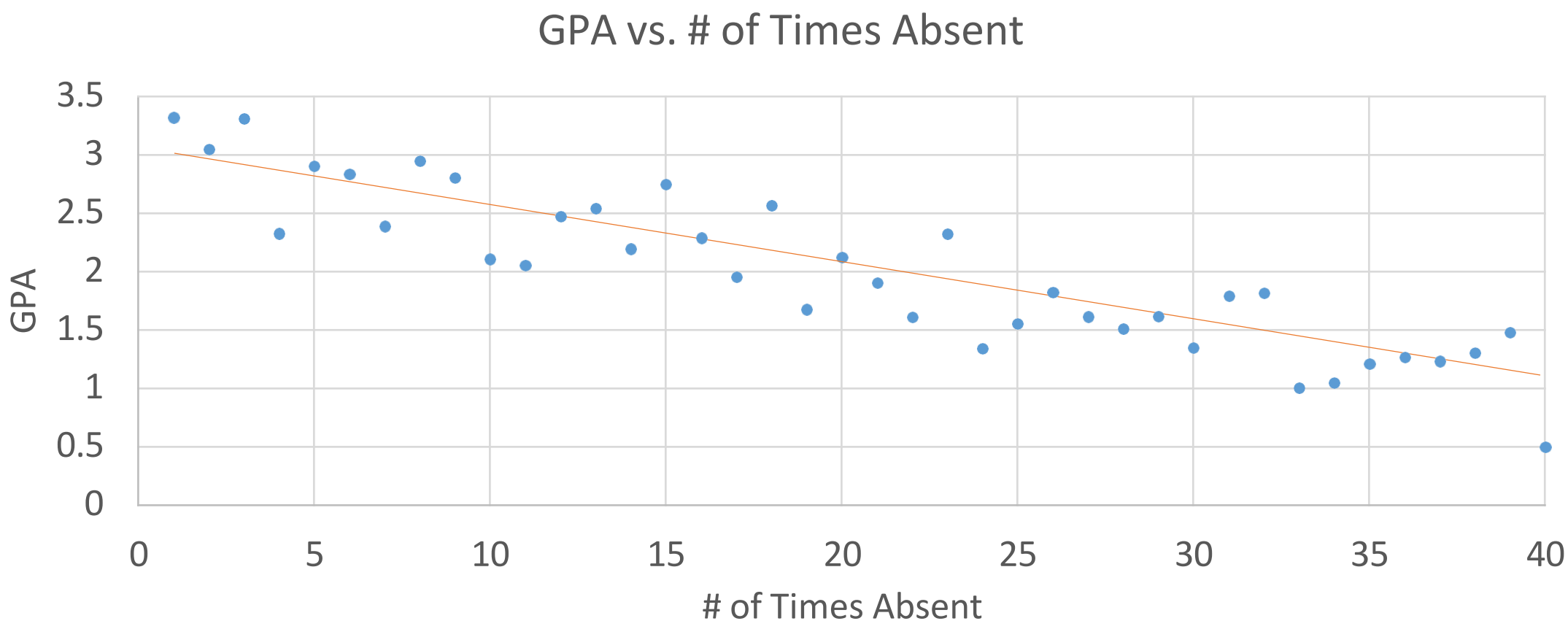
Observations (Cont.)



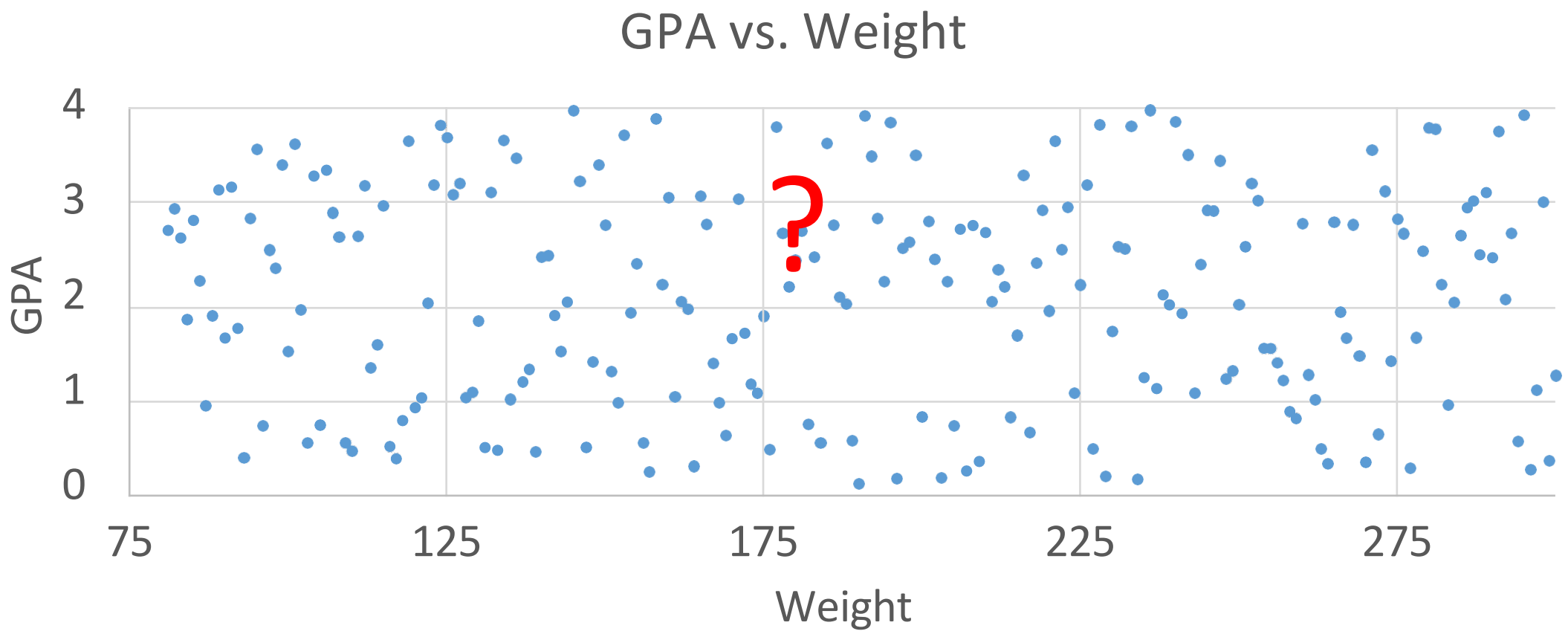
Find a Line that Fits the Data



Find a Line that Fits the Data (Cont.)



Find a Line that Fits the Data (Cont.)

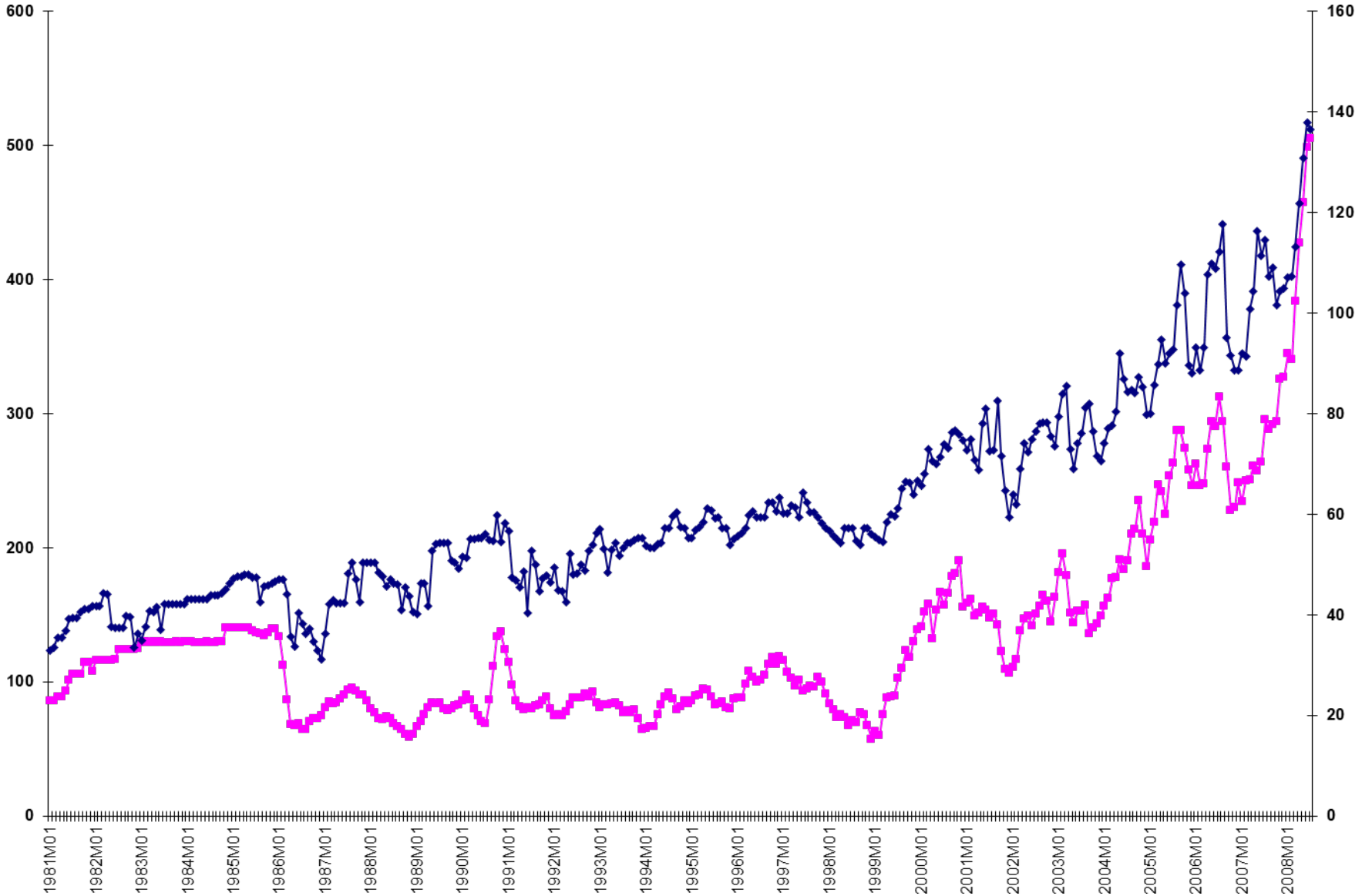


Regression Model

- Relation between variables where changes in some variables may "explain" or possibly "cause" changes in other variables.
- Explanatory variables are termed the **independent** variables and the variables to be explained are termed the **dependent** variables.
- Regression model estimates the nature of the relationship between the independent and dependent variables.
  - Change in dependent variables that results from changes in independent variables, i.e. # of hours study weekly of the relationship.
- If we assume the relation can be described by a line, it is linear regression.

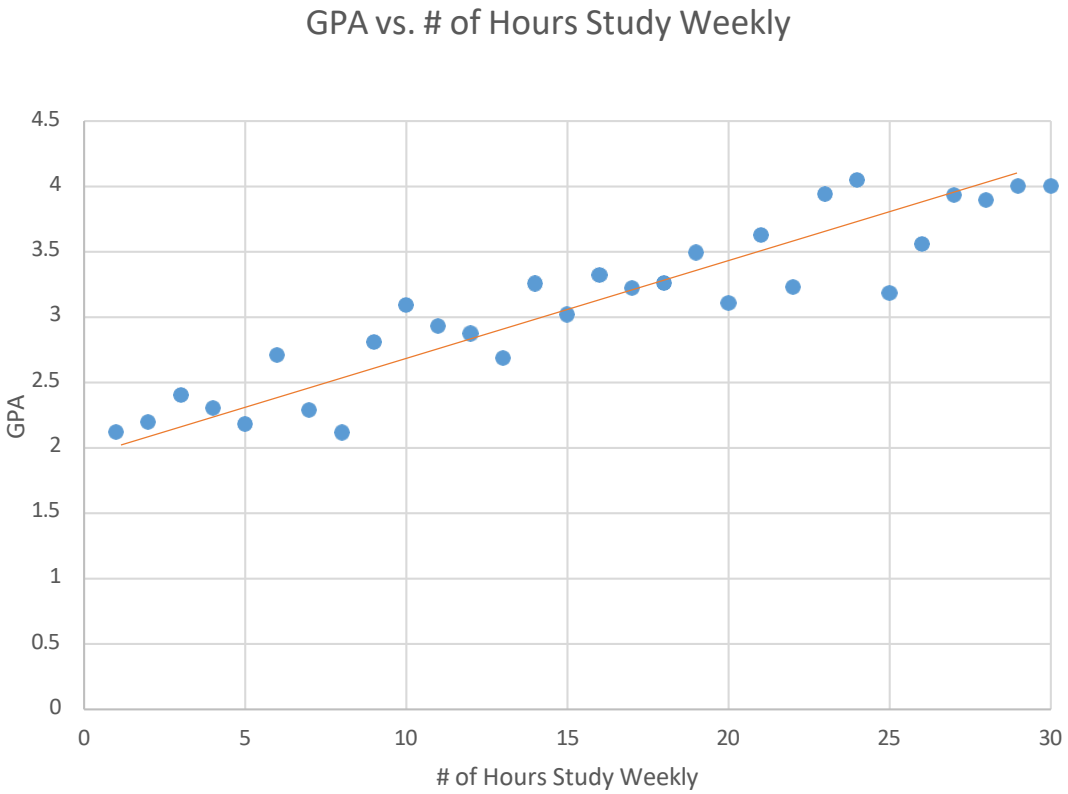
Examples of Dep/Independent Variables

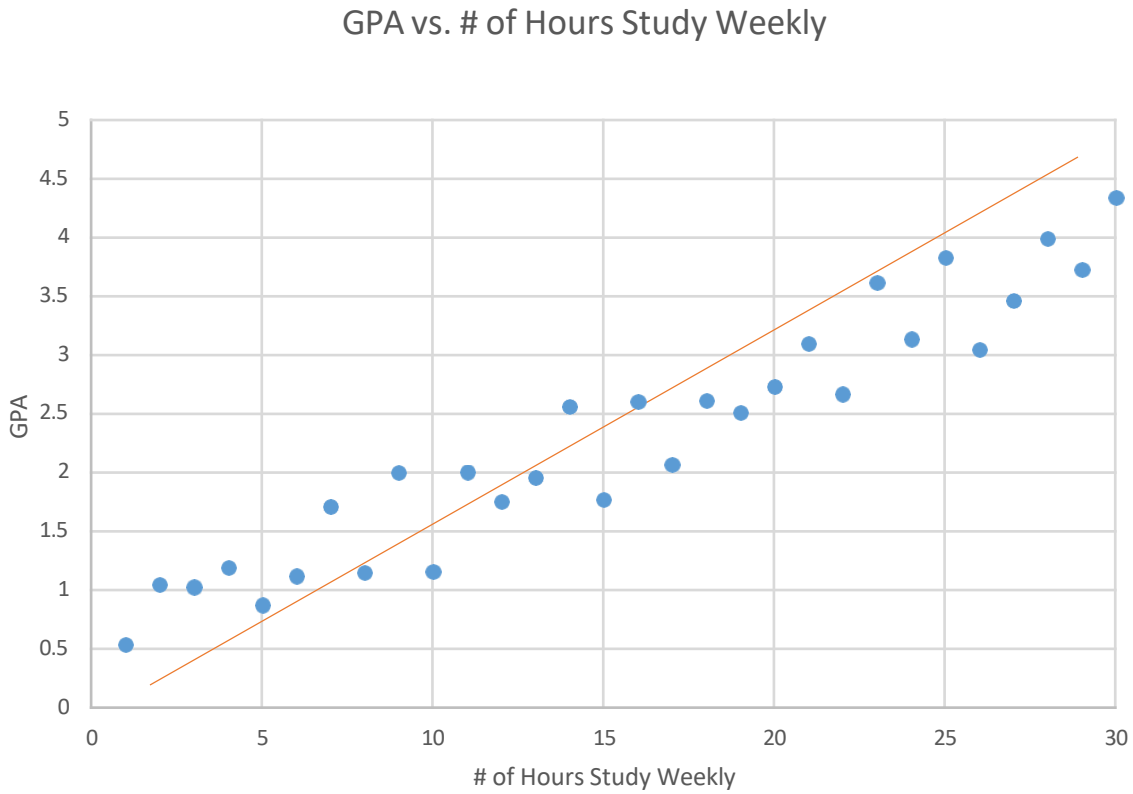
- Dependent variable is retail price of gasoline in Georgia whereas independent variable is the price of crude oil.
- Dependent variable is employment income whereas independent variables might be hours of work, education, occupation, sex, age, region, years of experience, unionization status, etc.
- Price of a product and quantity produced or sold:
  - Quantity sold affected by price. Dependent variable is quantity of product sold – independent variable is price.
  - Price affected by quantity offered for sale. Dependent variable is price – independent variable is quantity sold.



Pink: crude oil price index, 100@1997; Blue: regular gasoline prices, cents per liter

Strength of a Relation





The slope gives the information on the strength of the relation.

Measuring the Strength of a Linear Relation

- The correlation r measures the strength of a linear relation between two quantities.

$$r = \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{S_x}\right)\left(\frac{y_i-\bar{y}}{S_y}\right)$$

$$S_x = \sqrt{\sum(x_i-\bar{x})^2/(n-1)} \text{ and } S_y = \sqrt{\sum(y_i-\bar{y})^2/(n-1)}$$

- *r* is always a number between -1 and 1.
- $r > 0$ indicates a positive association.
- $r < 0$ indicates a negative association.

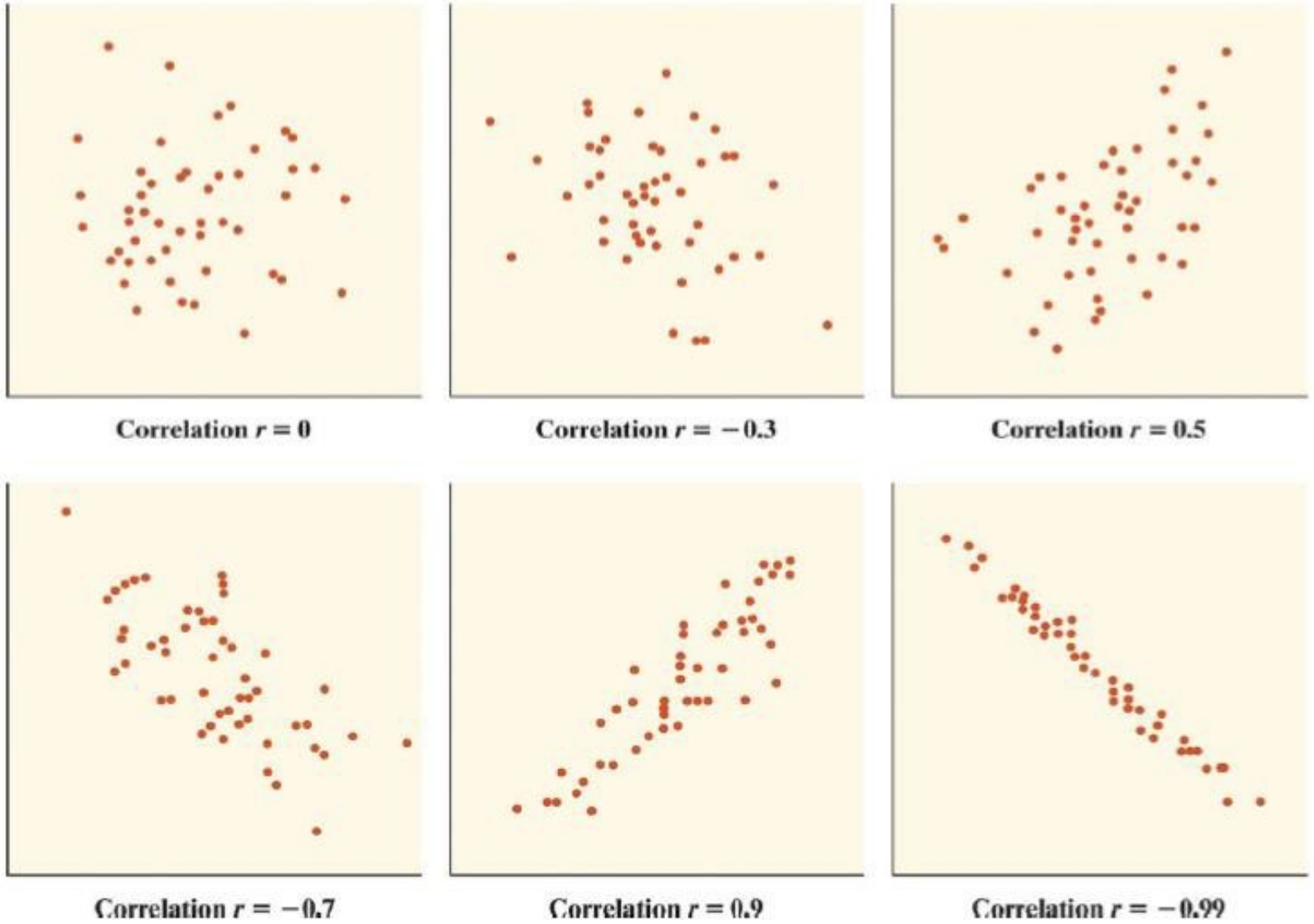
$$r = \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{S_x}\right)\left(\frac{y_i-\bar{y}}{S_y}\right)$$

$$S_x = \sqrt{\sum(x_i-\bar{x})^2/(n-1)} \text{ and } S_y = \sqrt{\sum(y_i-\bar{y})^2/(n-1)}$$

Measuring the Strength of a Linear Relation (cont.)

- Values of *r* near 0 indicate a very weak linear relationship.
- The strength of the linear relationship increases as *r* moves away from 0 toward -1 or 1.
- The extreme values *r* = -1 and r = 1 occur only in the case of a perfect linear relationship.
- Absolute value of r determines the strength of a relation:
  - $|r| < 0.3$ indicates none or very weak relation
  - $0.3 \leq \left|r\right| < 0.5$ indicates weak relation
  - $0.5 \leq \left|r\right| < 0.7$ indicates moderate relation
  - $0.7 \leq \left|r\right|$ indicates strong relation

Examples of Scatterplots with Different Correlation Coefficients



Statistical significance of the relationship

- Correlation coefficients have a probability (p-value), which shows the probability that the relationship between the two variables is equal to zero (null hypotheses; no relationship).
- Strong correlations have low p-values because the probability that they have no relationship is very low.
- Correlations are typically considered statistically significant if the p-value is lower than 0.05 in the social sciences, but the researcher has the liberty to decide the p-value for which he or she will consider the relationship to be significant.

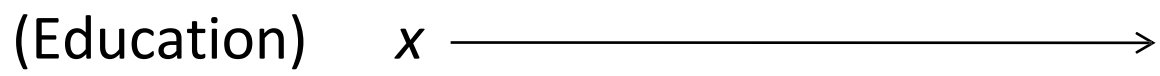
Hypothesis Testing

- We can test correlation equal to 0 using t test:
- $H_0: r = 0$

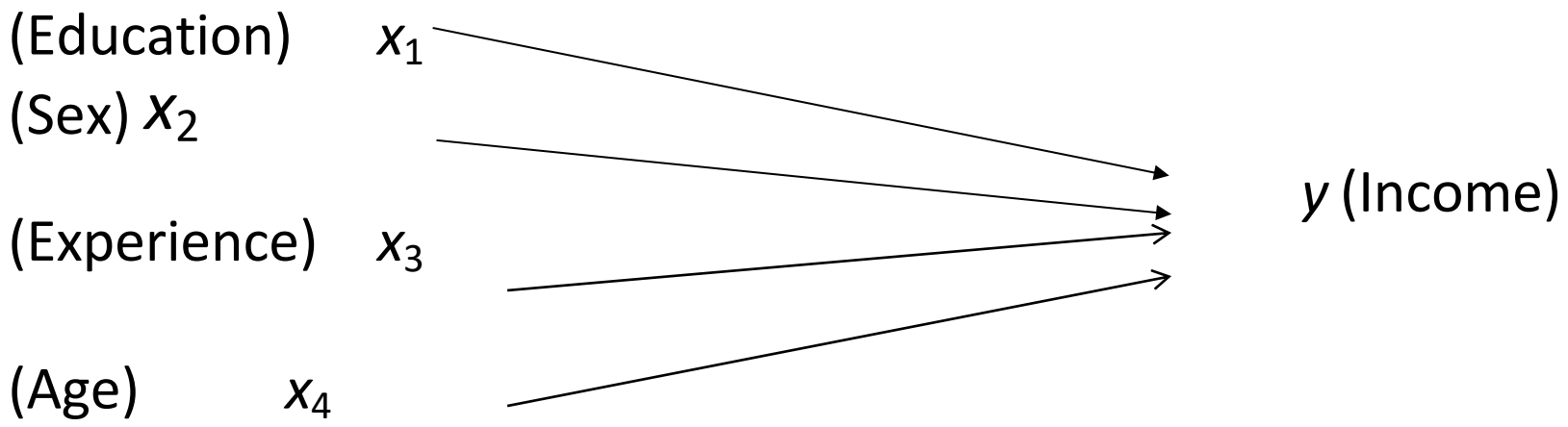
$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Bivariate and multivariate models

Bivariate or simple regression model



Multivariate or multiple regression model



Bivariate or simple linear regression

- *x* is the independent variable
- *y* is the dependent variable
- The regression model is $$y = \beta_0 + \beta_1 x + \varepsilon$$

- The model has two variables, the independent or explanatory variable, *x,* and the dependent variable *y*, the variable whose variation is to be explained.
- The relationship between *x* and *y* is a linear or straight line relationship.
- Two parameters to estimate – the slope of the line β_1 and the *y*-intercept β_0 (where the line crosses the vertical axis).
- ε is the unexplained, random, or error component. Much more on this later.

Regression line

- The regression model is $y = \beta_0 + \beta_1 x + \varepsilon$
- Data about *x* and *y* are obtained from a sample.
- From the sample of values of *x* and *y*, estimates b_0 of β_0 and b_1 of β_1 are obtained using the least squares or another method.
- The resulting estimate of the model is

$$\hat{y} = b_0 + b_1 x$$

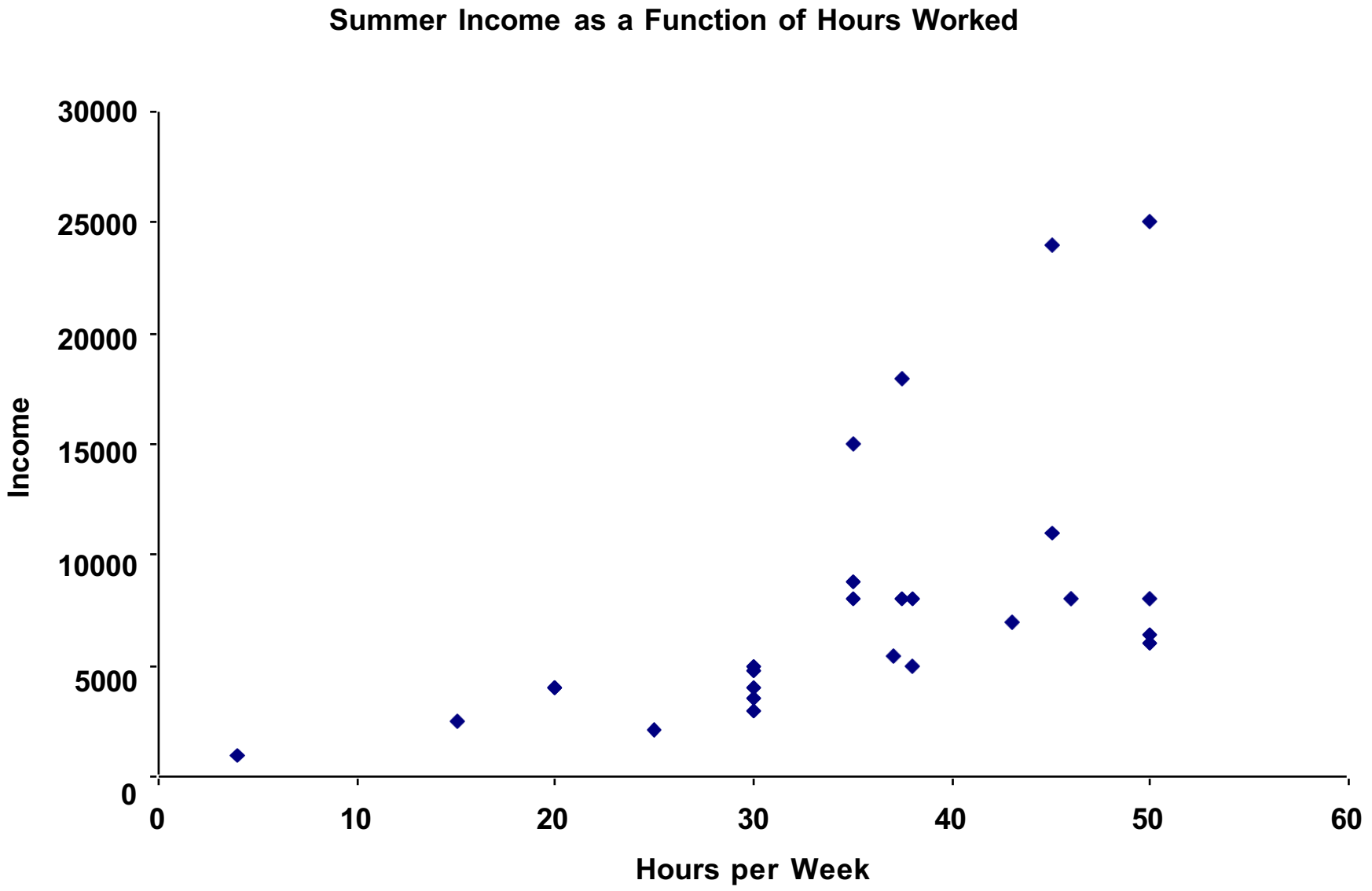
- The symbol is $\hat{y}$ termed "*y* hat" and refers to the predicted values of the dependent variable *y* that are associated with values of *x*, given the linear model.

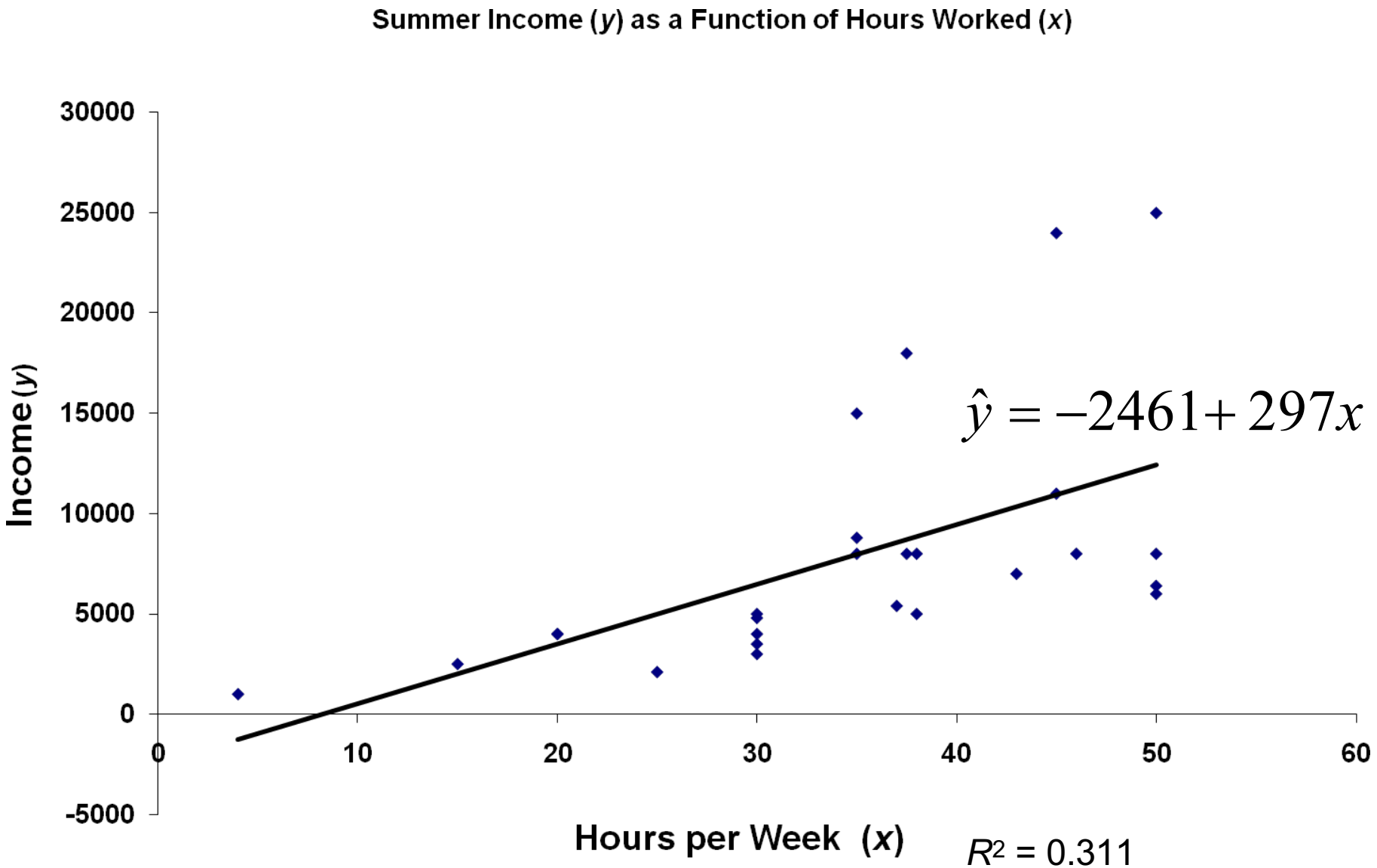
Uses of regression

- Amount of change in a dependent variable that results from changes in the independent variable(s) – can be used to estimate elasticities, returns on investment in human capital, etc.
- Attempt to determine causes of phenomena.
- Prediction and forecasting of sales, economic growth, etc.
- Support or negate theoretical model.
- Modify and improve theoretical models and explanations of phenomena.

Income	hrs/week
8000	38
6400	50
2500	15
3000	30
6000	50
5000	38
8000	50
4000	20
11000	45
25000	50
4000	20
8800	35
5000	30
7000	43

Income	hrs/week
8000	35
18000	37.5
5400	37
15000	35
3500	30
24000	45
1000	4
8000	37.5
2100	25
8000	46
4000	30
1000	200
2000	200
4800	30

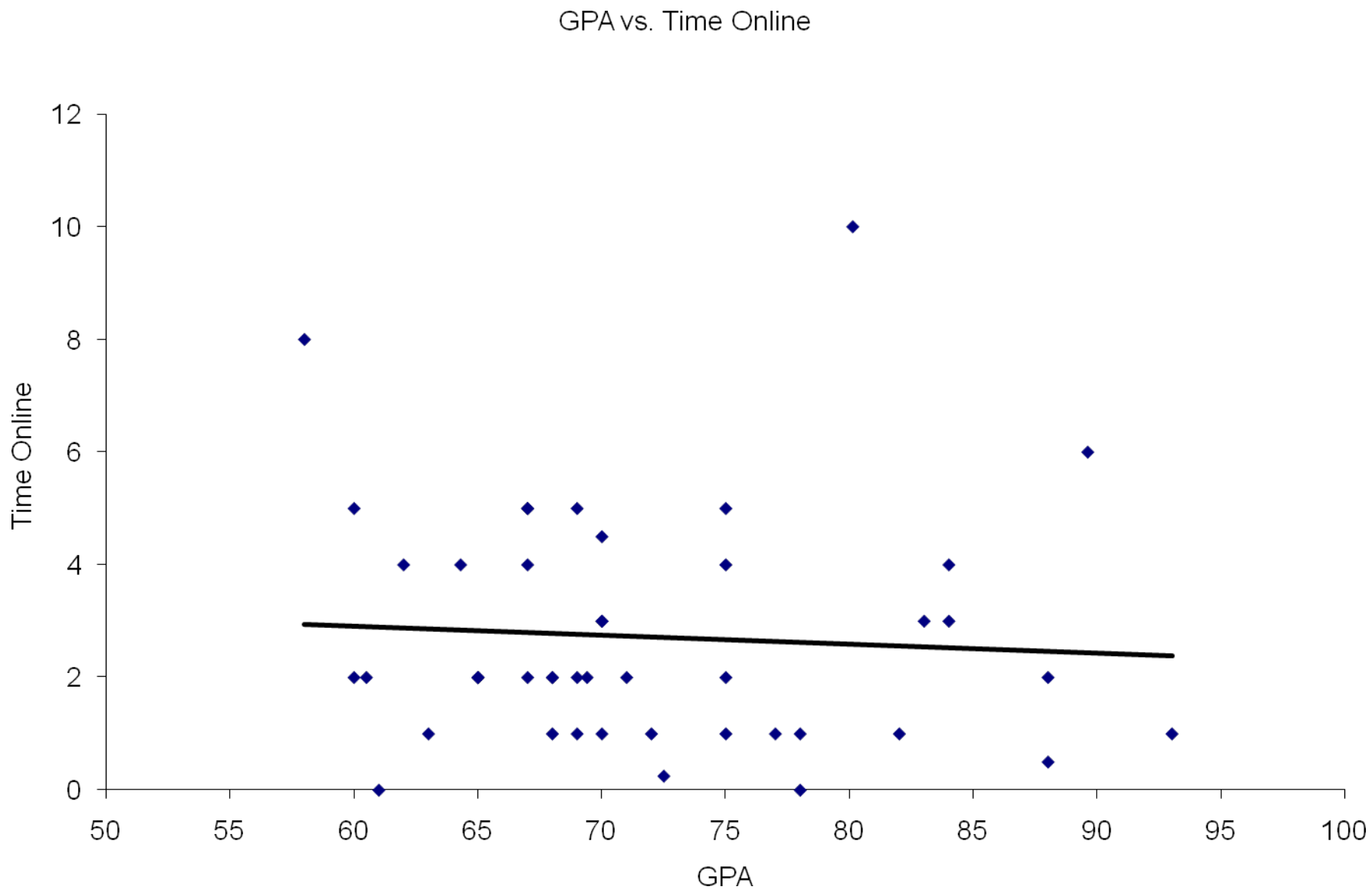

Summer Income as a Function of Hours Worked
Income
30000
25000
20000
15000
10000
5000
0
0
10
20
30
40
50
60
Hours per Week


Summer Income (y) as a Function of Hours Worked (x)
Income (y)
30000
25000
20000
15000
10000
5000
0
-5000
0
10
20
30
40
50
60
Hours per Week (x)
$\hat{y} = -2461 + 297x$
R^2 = 0.311
Significance = 0.0031

GPA	Online
75	5
63	1
70	3
68	2
60.5	2
58	8
67	5
80.12	10
64.3	4
84	4
62	4
72.5	0.25
70	4.5
75	2
77	1

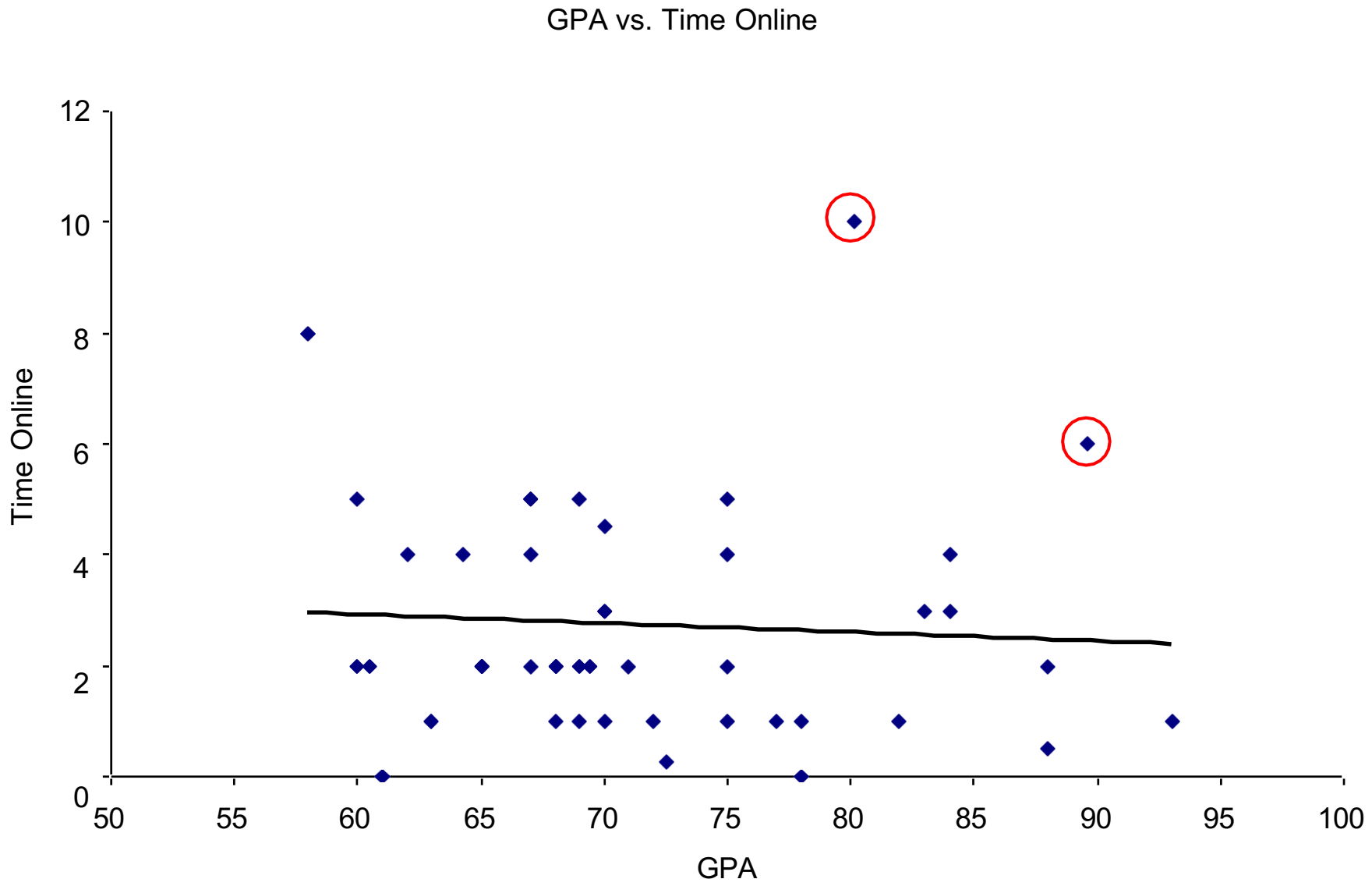
GPA	Online
61	0
69	1
84	3
71	2
78	0
75	4
70	3
82	1
60	5
89.6	6
93	1
83	3
65	2
67	4
78	1

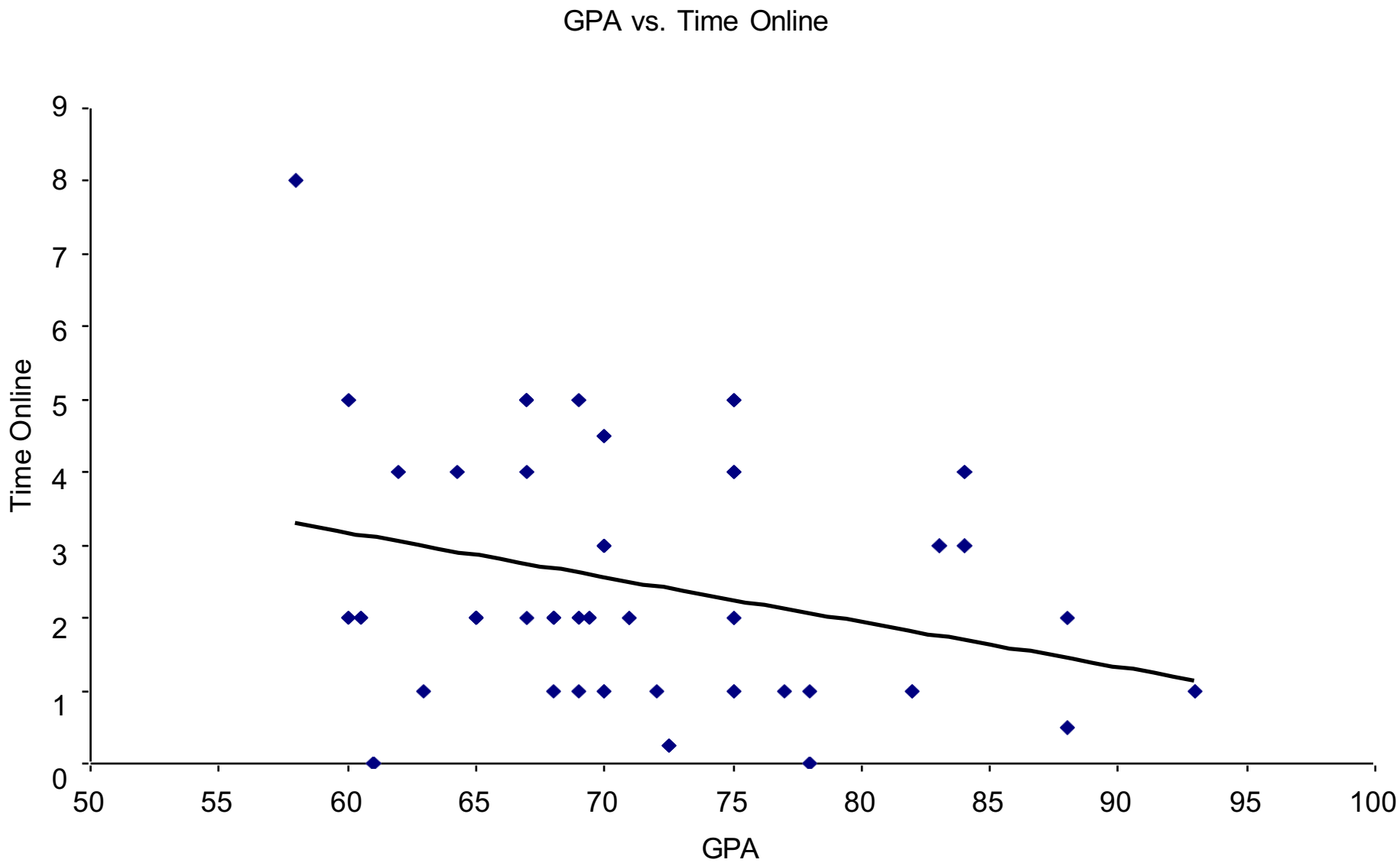
GPA	Online
88	0.5
72	1
67	5
69	2
69	5
68	1
65	2
65	2
88	2
68	2
67	2
60	2
70	1
75	1
69.4	2

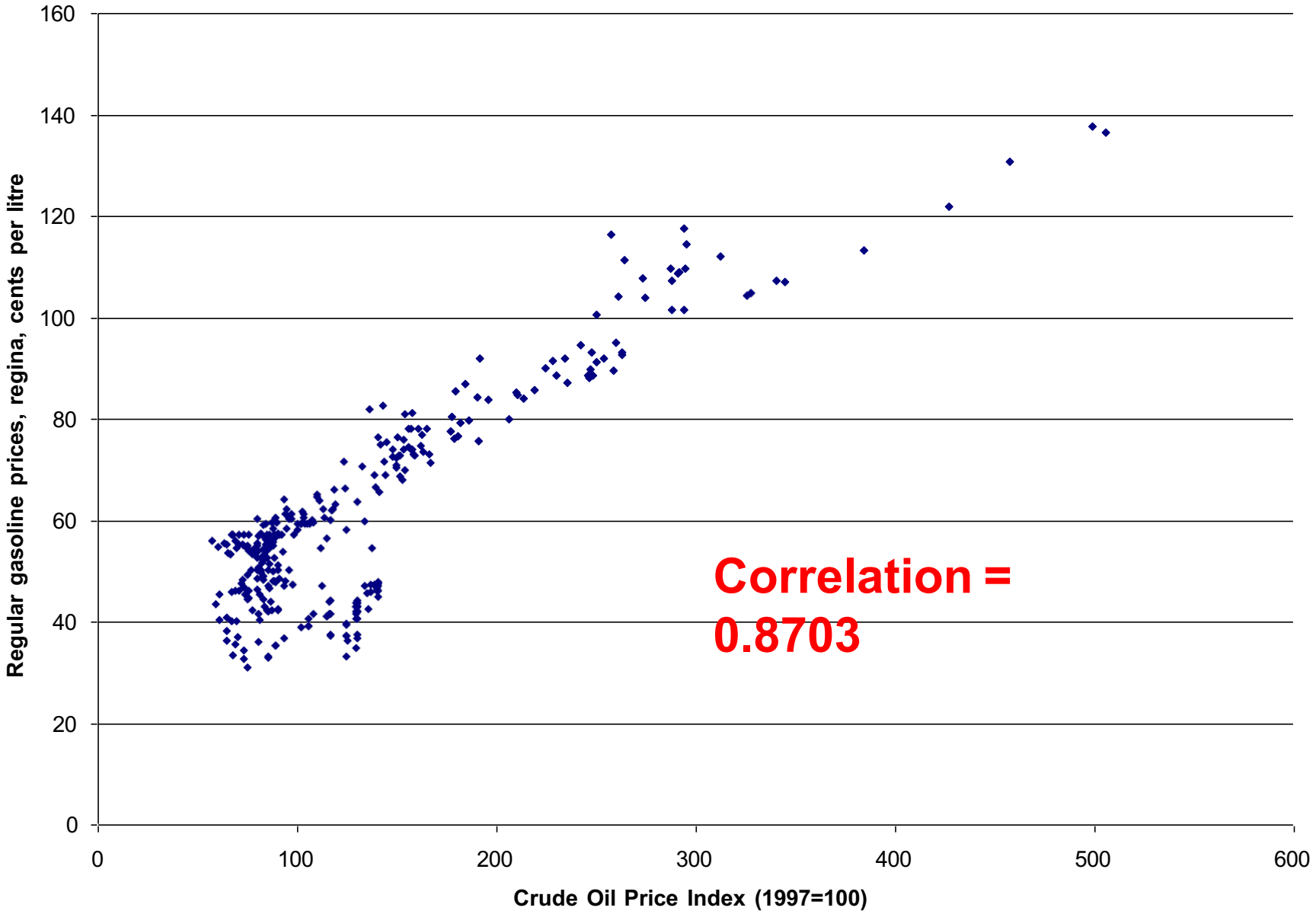

GPA vs. Time Online
Time Online
GPA
0
2
4
6
8
10
12
50
55
60
65
70
75
80
85
90
95
100

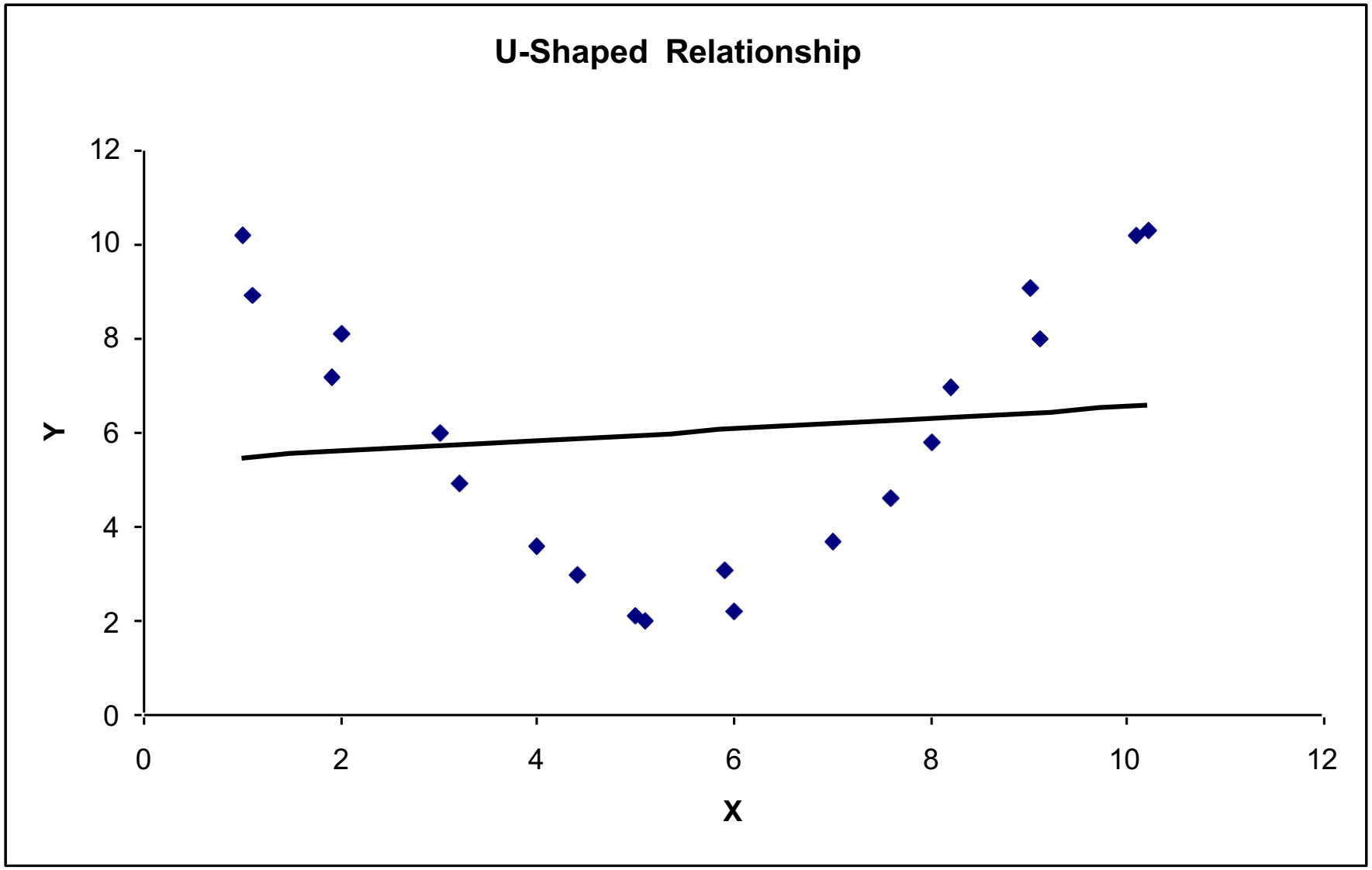
Outliers

- Rare, extreme values may distort the outcome.
  - Could be an error.
  - Could be a very important observation.
- Outlier: more than 3 standard deviations from the mean.


GPA vs. Time Online
Time Online
GPA
0
2
4
6
8
10
12
50
55
60
65
70
75
80
85
90
95
100


GPA vs. Time Online
Time Online
GPA
0
1
2
3
4
5
6
7
8
9
50
55
60
65
70
75
80
85
90
95
100


160
140
120
100
80
60
40
20
0
Regular gasoline prices, regina, cents per litre
0
100
200
300
400
500
600
Crude Oil Price Index (1997=100)
Correlation =
0.8703



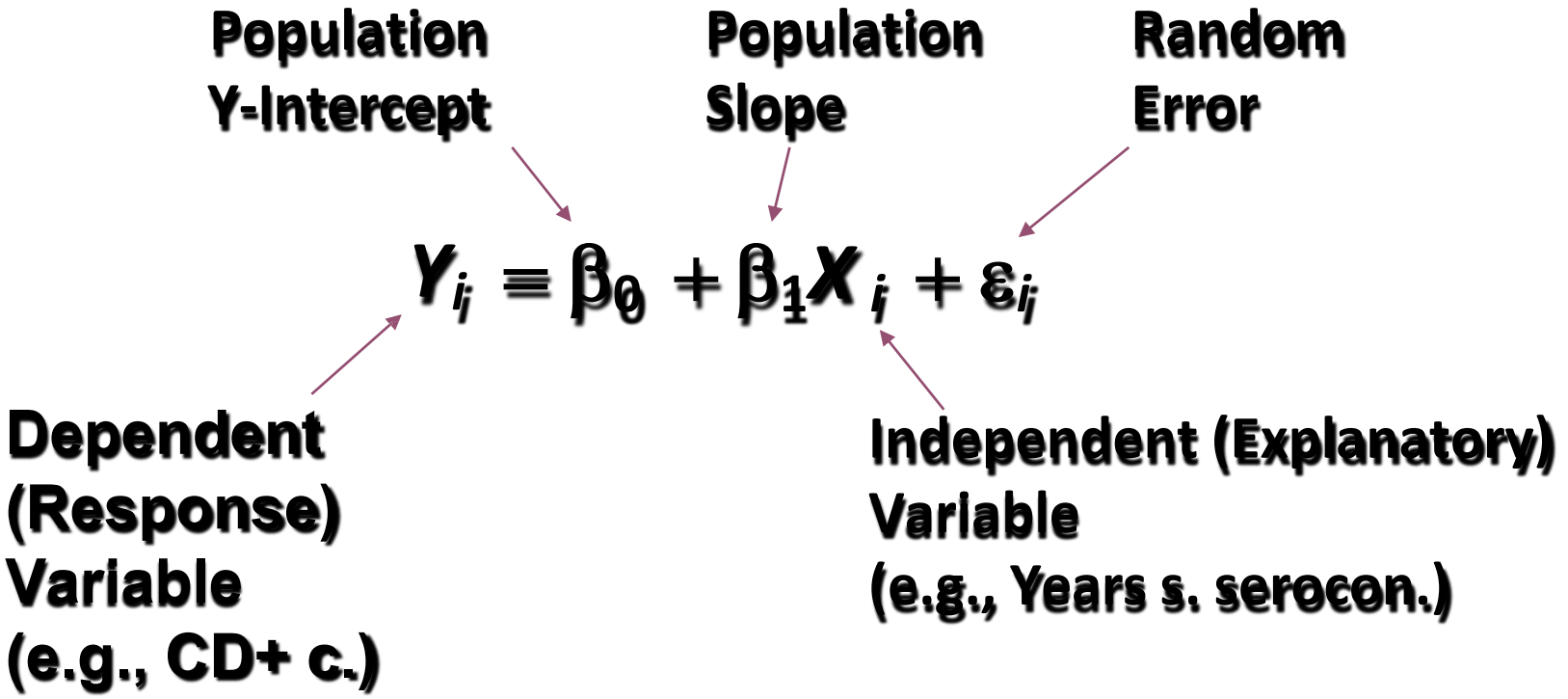
Correlation = +0.12.

Regression Modeling Steps

1. Hypothesize Deterministic Component
    - Estimate Unknown Parameters
2. Specify Probability Distribution of Random Error Term
    - Estimate Standard Deviation of Error
3. Evaluate the fitted Model
4. Use Model for Prediction & Estimation

Linear Regression Model

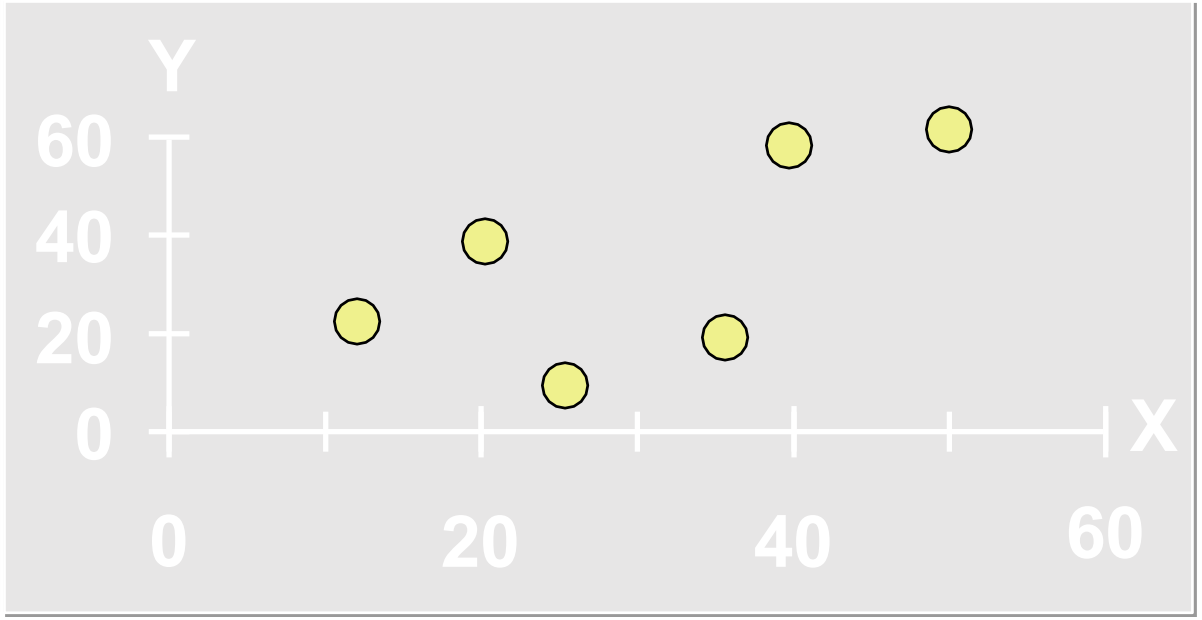
- 1. Relationship Between Variables Is a Linear Function



Estimating Parameters: Least Squares Method

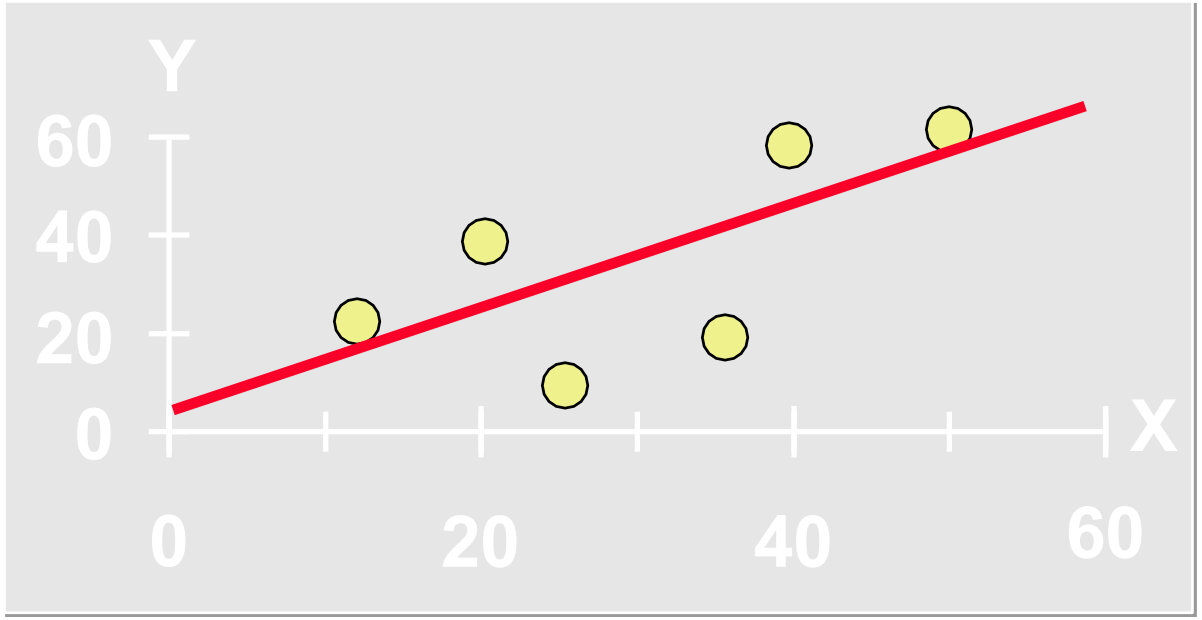
Scatter plot

- Plot of All (X_i, Y_i) Pairs
- Suggests How Well Model Will Fit



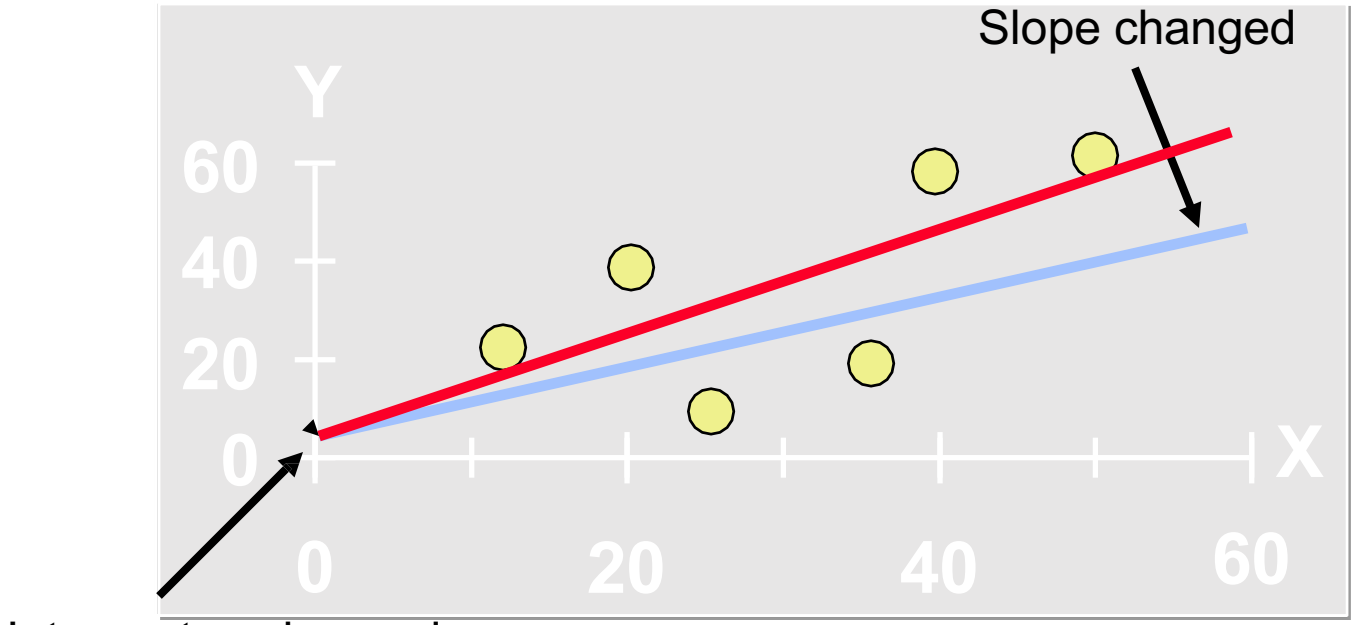
Thinking Challenge

- **How would you draw a line through the points?**
- **How do you determine which line 'fits best'?**



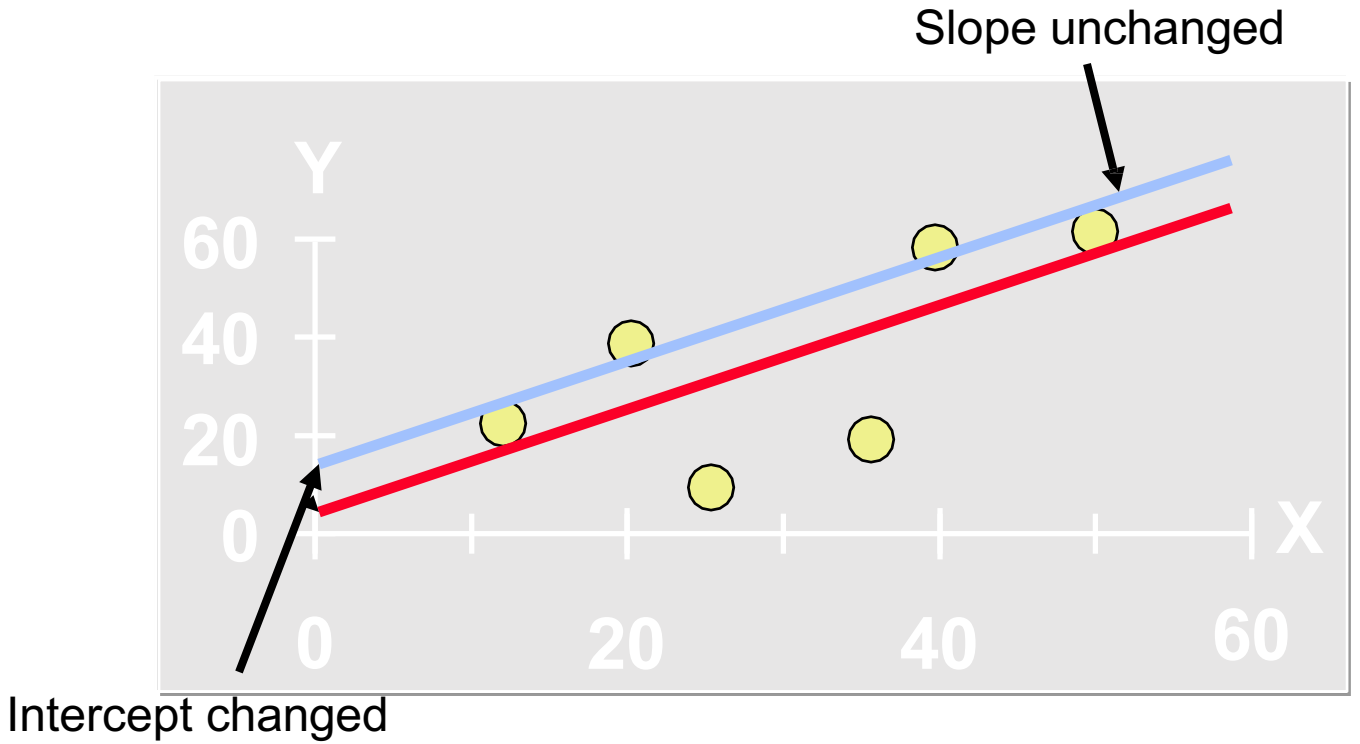
Thinking Challenge

How would you draw a line through the points?
How do you determine which line 'fits best'?



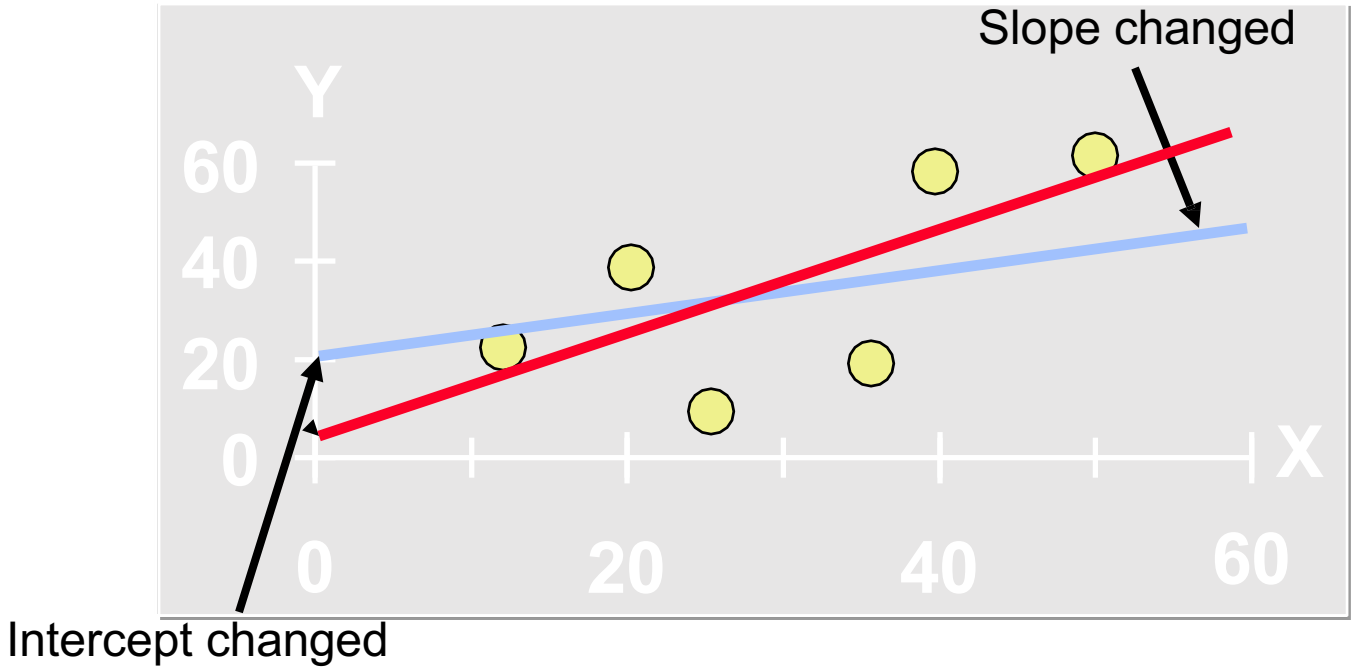
Thinking Challenge

How would you draw a line through the points?
How do you determine which line 'fits best'?



Thinking Challenge

How would you draw a line through the points?
How do you determine which line "fits best"?



Least Squares

- ‘Best Fit’ Means Difference Between Actual Y Values & Predicted Y Values Are a Minimum.
- *But* Positive Differences Off-Set Negative ones

Least Squares

- 'Best Fit' Means Difference Between Actual Y Values & Predicted Y Values is a Minimum. *But* Positive Differences Off-Set Negative ones. So square errors!

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

Least Squares

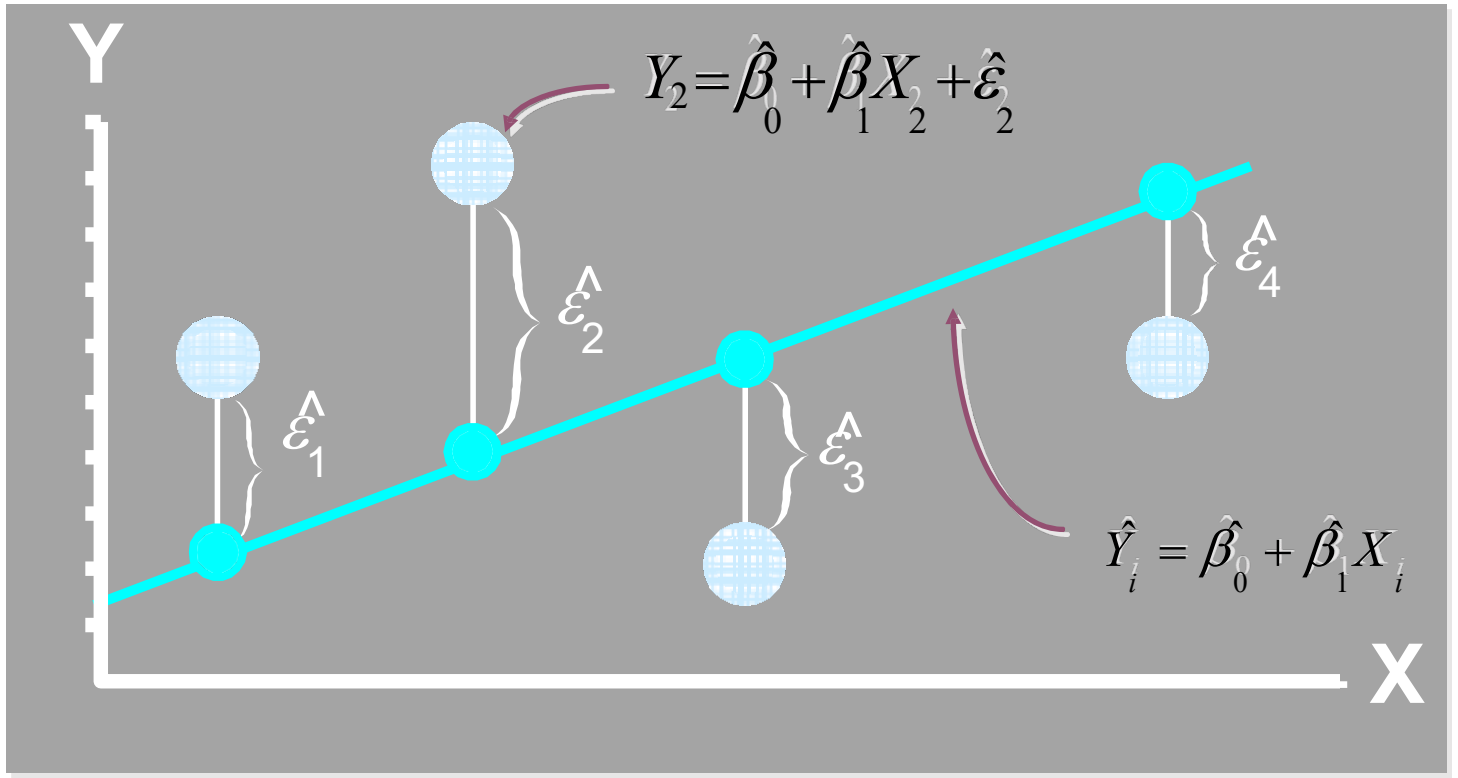
- 'Best Fit' Means Difference Between Actual Y Values & Predicted Y Values Are a Minimum. *But* Positive Differences Off-Set Negative. So square errors!

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

- LS Minimizes the Sum of the Squared Differences (errors) (SSE)

Least Squares Graphically

LS minimizes $\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \hat{\varepsilon}_1^2 + \hat{\varepsilon}_2^2 + \hat{\varepsilon}_3^2 + \hat{\varepsilon}_4^2$



Derivation of Parameters: β_0

- Least Squares (L-S):

  Minimize squared error

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$0 = \frac{\partial \sum \varepsilon_i^2}{\partial \beta_0} = \frac{\partial \sum (y_i - \beta_0 - \beta_1 x_i)^2}{\partial \beta_0}$$

$$= -2(n\bar{y} - n\beta_0 - n\beta_1 \bar{x})$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Derivation of Parameters: β_1

- Least Squares (L-S):

  Minimize squared error

$$0 = \frac{\partial \sum \varepsilon_i^2}{\partial \beta_1} = \frac{\partial \sum (y_i - \beta_0 - \beta_1 x_i)^2}{\partial \beta_1}$$
$$= -2 \sum x_i (y_i - \beta_0 - \beta_1 x_i)$$
$$= -2 \sum x_i (y_i - \bar{y} + \beta_1 \bar{x} - \beta_1 x_i)$$

$$\beta_1 = \frac{\Sigma xy - \frac{1}{n}(\Sigma x)(\Sigma y)}{\Sigma x^2 - \frac{1}{n}(\Sigma x)^2}$$

Coefficient Equations

- Prediction equation

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- Sample slope

$$\hat{\beta}_1 = \frac{\Sigma xy - \frac{1}{n}(\Sigma x)(\Sigma y)}{\Sigma x^2 - \frac{1}{n}(\Sigma x)^2}$$

- Sample Y - intercept

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Computation Table

X_i	Y_i	X_i^2	Y_i^2	$X_i Y_i$
X_1	Y_1	X_1^2	Y_1^2	$X_1 Y_1$
X_2	Y_2	X_2^2	Y_2^2	$X_2 Y_2$
:	:	:	:	:
X_n	Y_n	X_n^2	Y_n^2	$X_n Y_n$
ΣX_i	ΣY_i	ΣX_i^2	ΣY_i^2	$\Sigma X_i Y_i$

Interpretation of Coefficients

- Slope (β_1)
  - Estimated *Y* Changes by β_1 for Each 1 Unit Increase in *X*
  - If $\beta_1 = 2$, then *Y* Is Expected to Increase by 2 for Each 1 Unit Increase in *X*
- Y-Intercept (β_0)
  - Average Value of *Y* When $X = 0$
  - If $\beta_0 = 4$, then Average *Y* Is Expected to Be 4 When *X* Is 0

Parameter Estimation Solution Table*

X_i	Y_i	X_i^2	Y_i^2	$X_i Y_i$
4	3.0	16	9.00	12
6	5.5	36	30.25	33
10	6.5	100	42.25	65
12	9.0	144	81.00	108
32	24.0	296	162.50	218

Parameter Estimation Solution*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}} = \frac{218 - \frac{(32)(24)}{4}}{296 - \frac{(32)^2}{4}} = 0.65$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} = 6 - (0.65)(8) = 0.80$$

Coefficient Interpretation Solution

- Slope
  - Milk Yield (*Y*) Is Expected to Increase by .65 lb. for Each 1 lb. Increase in Food intake (*X*)

- Y-Intercept
  - Average Milk yield (*Y*) Is Expected to Be 0.8 lb. When Food intake (*X*) Is 0

Goodness of Fit

- *y* is the dependent variable, or the variable to be explained.
- How much of *y* is explained statistically from the regression model, in this case the line?
- Total variation in *y* is termed the total sum of squares, or SST.

$$\mathrm{SST} = \sum (y_i - \underline{y})_2$$

- The common measure of goodness of fit of the line is the **coefficient of determination**, the proportion of the variation or SST that is "explained" by the line.

SST or Total Variation of *y*

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

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

Difference from mean

"Error" of prediction

Value of *y* "explained" by the line

Difference of any observed value of *y* from the mean is the difference between the observed and predicted value plus the difference of the predicted value from the mean of *y*. From this, it can be proved that:

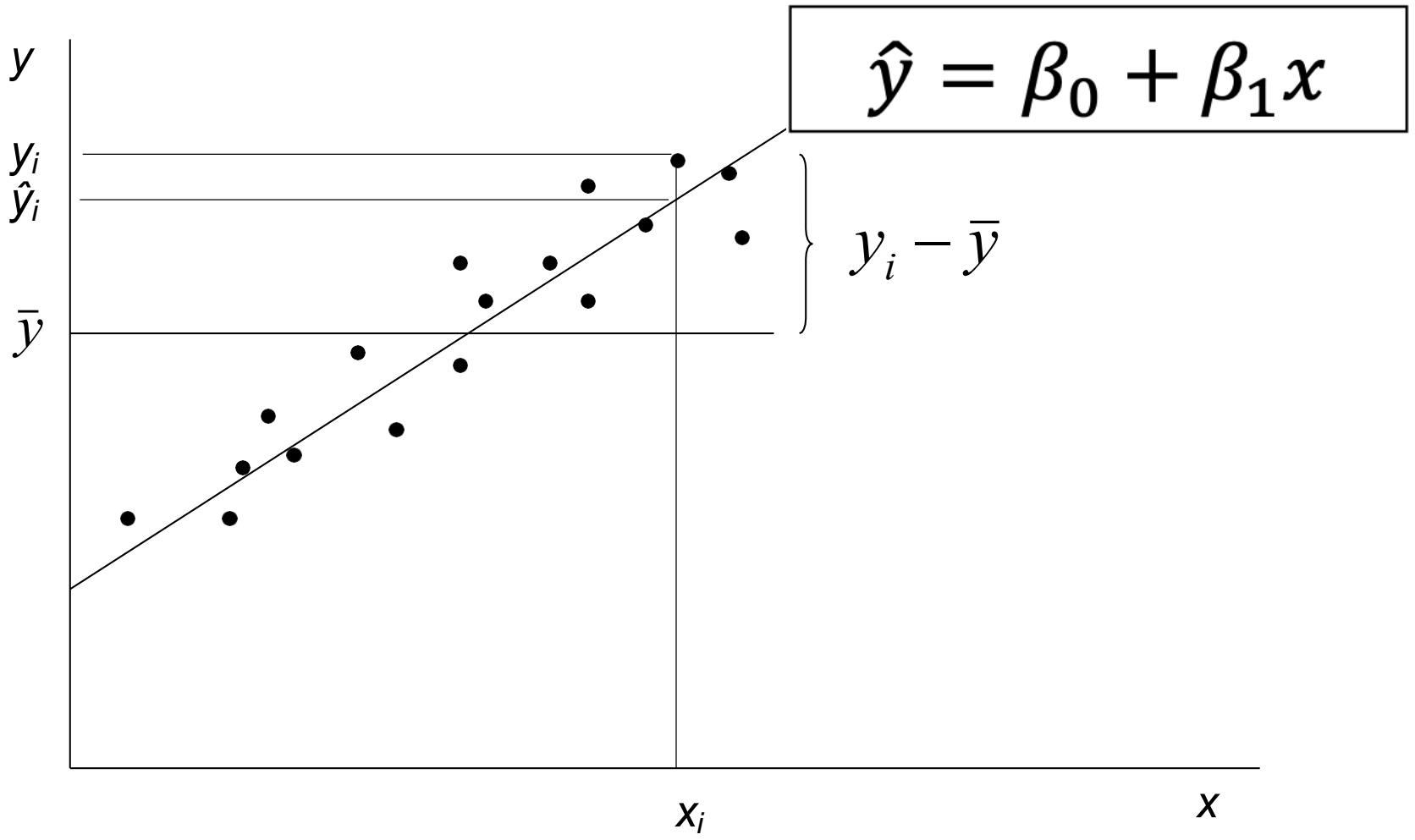
$$\sum (y_i - \underline{y})_2 = \sum (y_i - \hat{y}_i)_2 + \sum (\hat{y}_i - \underline{y})_2$$

SST= Total variation of *y*

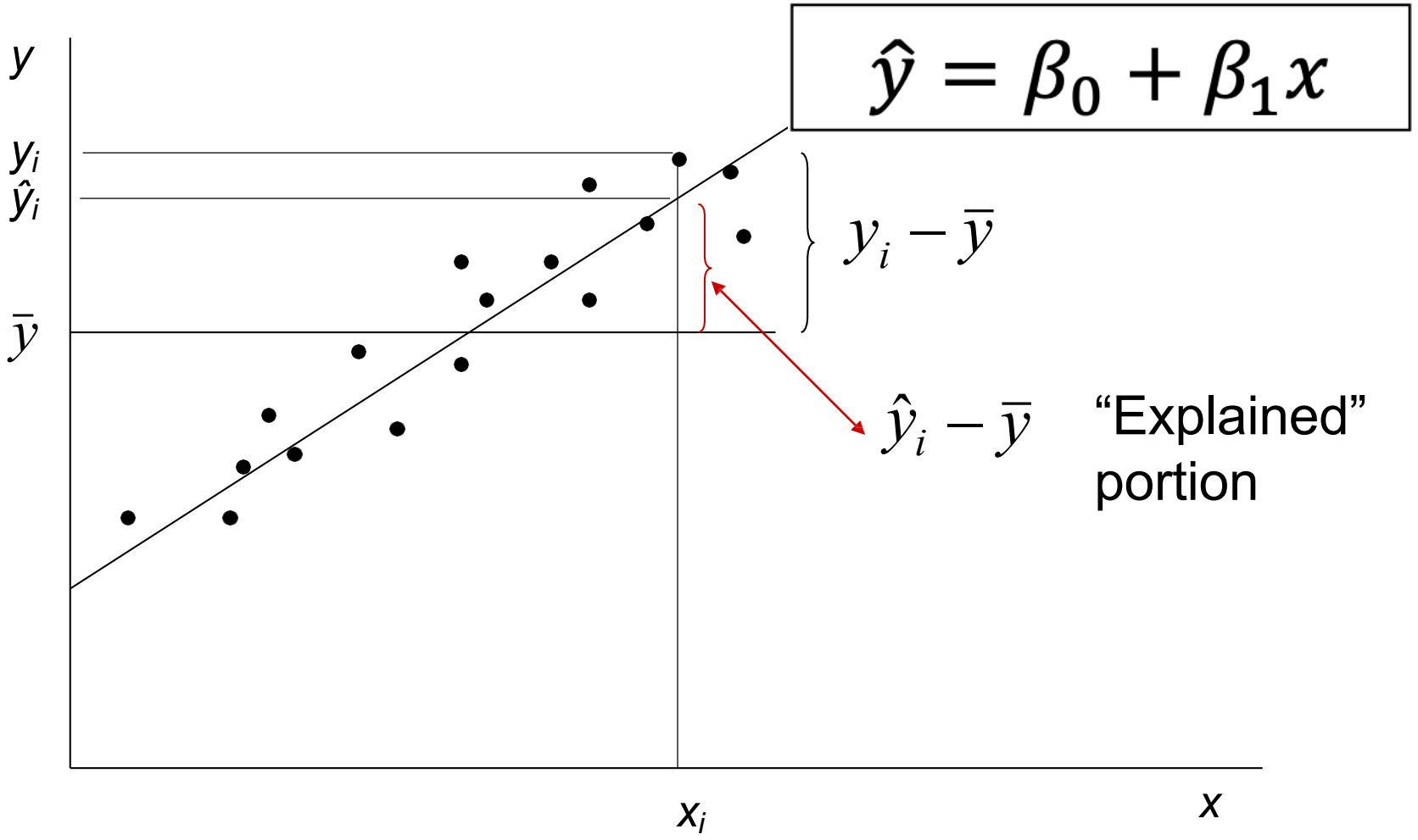
SSE = "Unexplained" or "error" variation of *y*

SSR = "Explained" variation of *y*

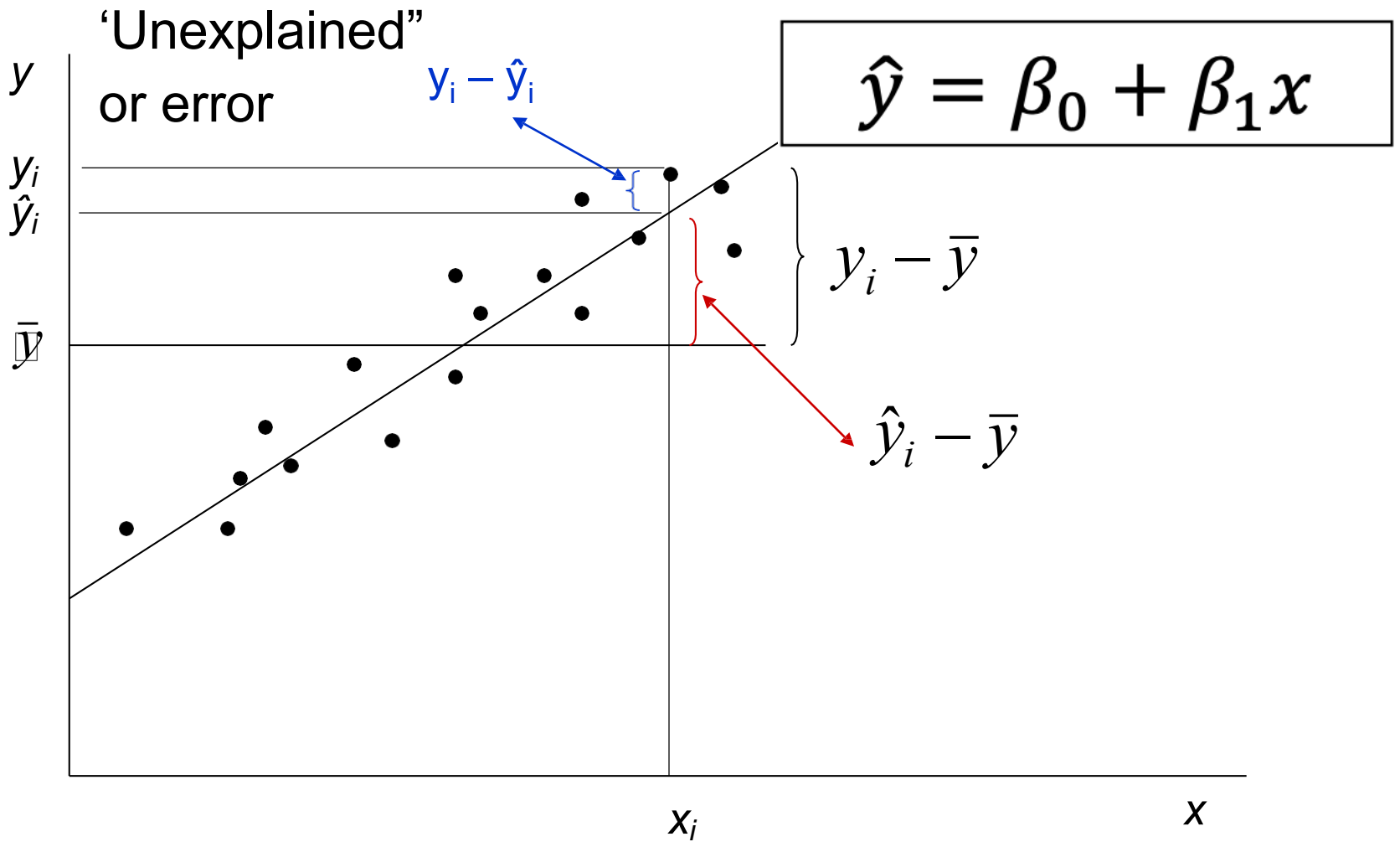
Variation in *y*



Variation in *y* "explained" by the line



Variation in *y* that is "unexplained" or error



SST = SSR + SSE

CONNECTION?

Total variability = Explained variability + Unxplained variability

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

Coefficient of Determination

The **coefficient of determination**, r^2 or R^2 (the notation used in many texts), is defined as the ratio of the "explained" or regression sum of squares, SSR, to the total variation or sum of squares, SST.

$$r^2 = R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

The coefficient of determination is the square of the correlation coefficient *r.* The correlation coefficient, *r*, is the square root of the coefficient of determination, but with the same sign (positive or negative) as β_1.

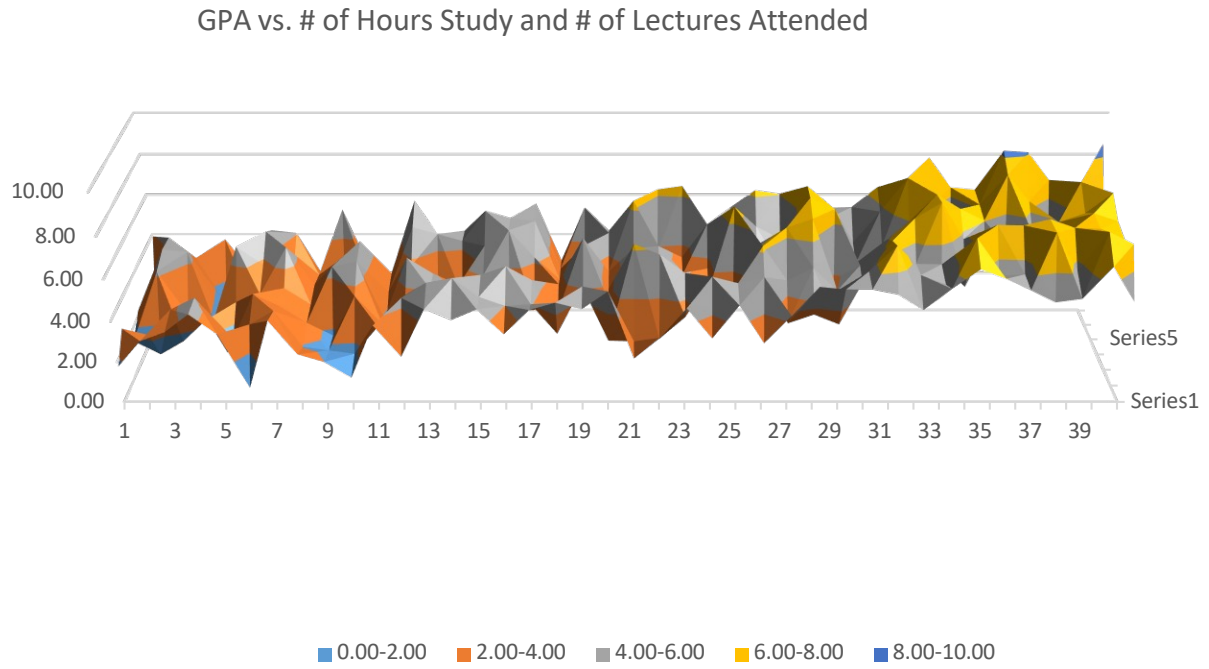
Multiple Linear Regression

- Multiple linear regression (MLR), also known simply as multiple regression, is a linear regression that uses multiple explanatory variables (features) to predict the outcome of a response variable.
- The goal of multiple linear regression (MLR) is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable.
- Consider two features: $x_1, x_2 \Rightarrow y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i$ We basically extend the bivariate model to have one extra feature value.
- So the sample data point looks like: $(y_i, x_{i,1}, x_{i,2})$

Example

- Assume we want to model GPA versus # of hours study and # of lectures attended.
- We need to collect data points (GPA, # of hours study, # of lectures attended).

# of Hours Study\ # of Lectures Attended	0	1	2	3	4	5	6
1	1.78	2.87	1.46	2.23	1.06	0.83	3.88
2	3.03	2.55	0.78	4.21	4.60	4.51	2.59
3	3.47	3.81	1.46	2.04	4.48	3.48	1.60
4	4.33	4.39	2.52	1.76	1.43	1.02	3.72
5	3.37	1.67	1.58	2.68	4.74	1.49	1.93
6	0.71	3.08	3.83	3.07	3.84	4.81	4.20
7	4.03	3.74	3.23	3.38	4.71	3.07	4.06
8	2.38	1.96	2.39	2.57	2.83	2.61	1.68
9	2.00	1.78	1.75	4.46	3.98	1.53	5.25
10	1.22	1.61	3.73	4.71	5.02	2.32	2.01



Matrix Representation

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_i, p + \epsilon_i$$

$$Y^{n\times 1} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} \mathrm{X}^{n\times(p+1)} = \begin{bmatrix} 1 x_{1,1} \cdots x_{1,p} \\ \vdots \quad \vdots \quad \ldots \quad \vdots \\ 1 x_{n,1} \ldots x_{n,p} \end{bmatrix} \beta^{(p+1)\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} \epsilon^{n\times 1} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$Y = X\beta + \epsilon$$

$$\hat{Y} = X\hat{\beta}$$

LSE Estimate

$$SSE = \sum(y_i - \widehat{y_i})^2 = \sum \epsilon_i^2 = \epsilon^T \epsilon$$
$$SSE = \left(Y - \hat{Y}\right)^T \left(Y - \hat{Y}\right)$$

Plug in $\hat{Y} = X\hat{\beta} \Rightarrow SSE = \left(Y - X\hat{\beta}\right)^T \left(Y - X\hat{\beta}\right)$

Apply transpose in the terms, we have

$$SSE = \left(Y^T - \hat{\beta}^T X^T\right) \left(Y - X\hat{\beta}\right)$$
$$SSE = Y^T Y - Y^T X\hat{\beta} - \hat{\beta}^T X^T Y + \hat{\beta}^T X^T X\hat{\beta}$$

Thus, SSE is a function of $\hat{\beta}$. We want to minimize SSE w.r.t $\hat{\beta}$.

Multiple Regression Process

- Preprocess data points and labels to build $X\ and\ Y$ such that
- $$Y^{n\times 1} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} \mathrm{X}^{n\times(p+1)} = \begin{bmatrix} 1 x_{1,1} \cdots x_{1,p} \\ \vdots \quad \vdots \quad \ldots \quad \vdots \\ 1 x_{n,1} \ldots x_{n,p} \end{bmatrix}$$

Compute $\hat{\beta} = (X^T X)^{-1} X^T Y$

Predict new samples given test data X after prefixed one and compute $\hat{Y} = X\hat{\beta}$.