UNIT 4

Regression Analysis I

Linear Regression Model

Relationship between variables is a linear function

Population y-intercept

Population Slope

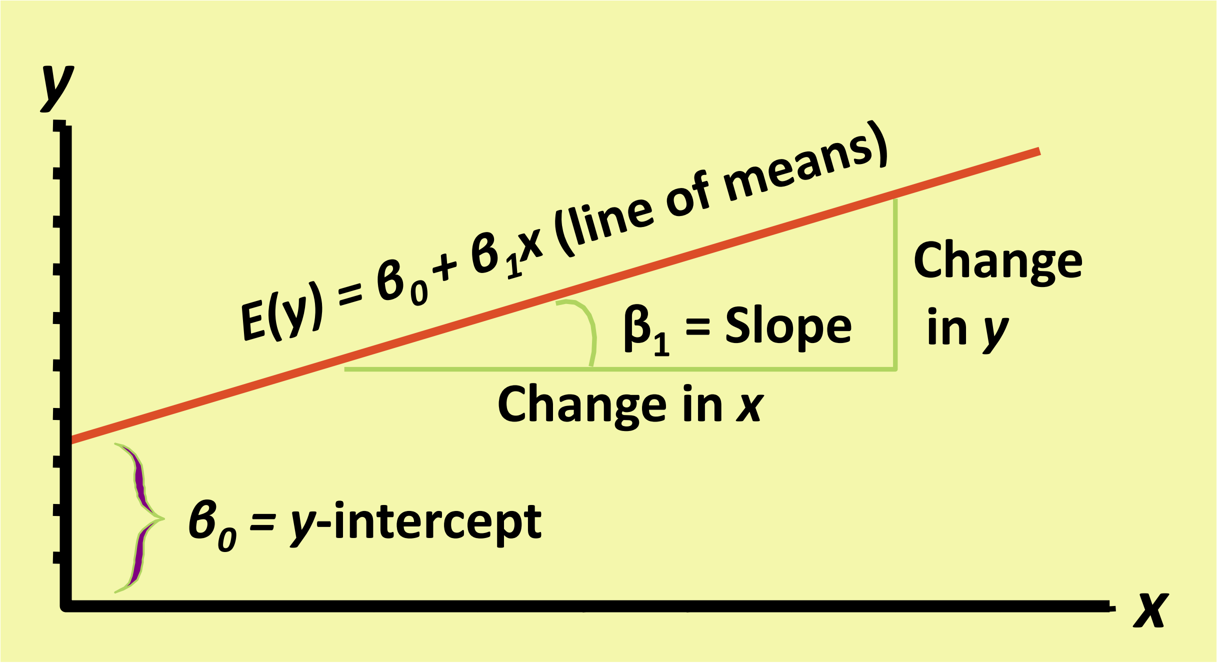
Random Error

$$y = \beta_0 + \beta_1 x + \varepsilon$$

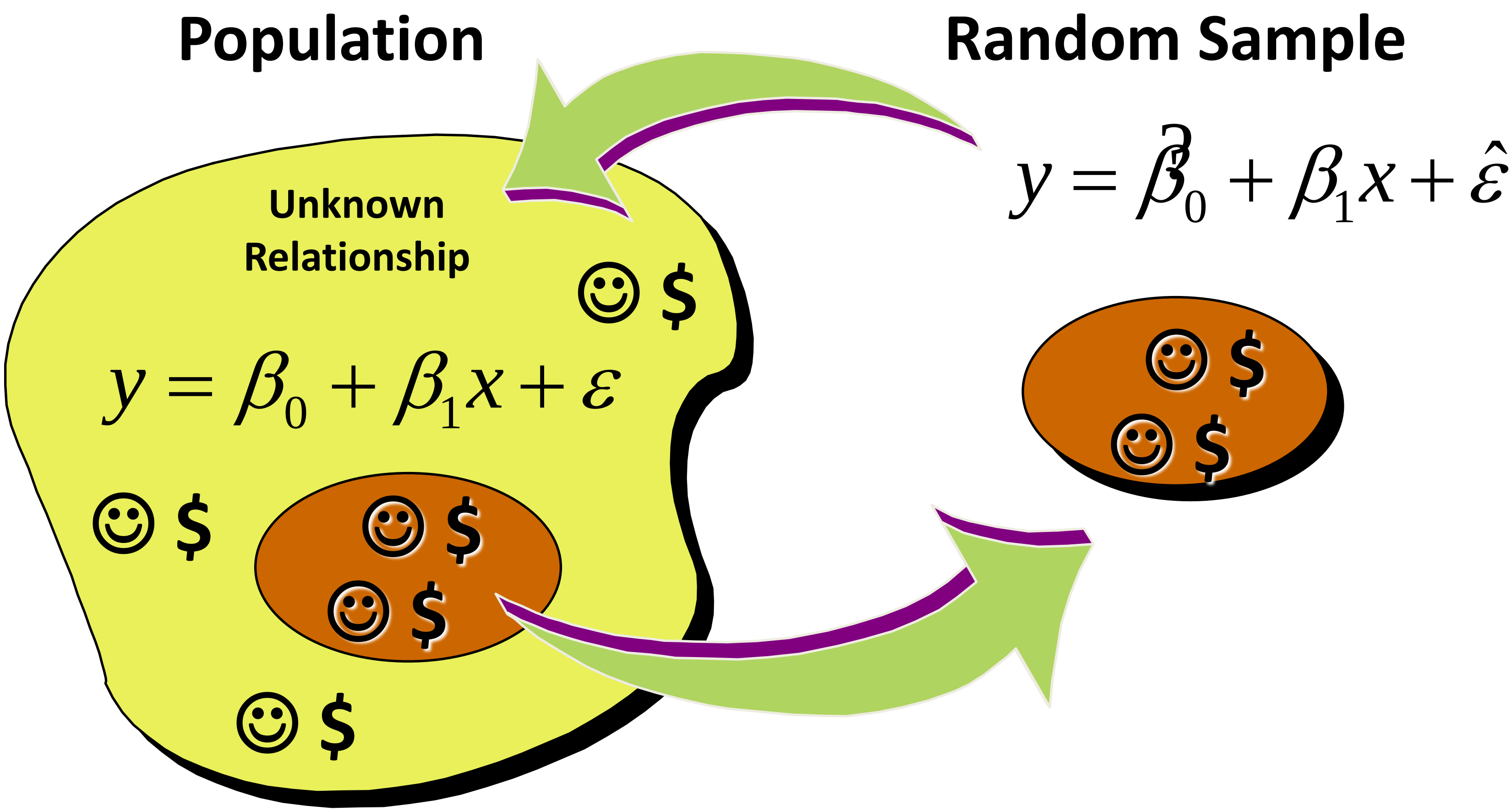
Dependent (Response) Variable

Independent (Explanatory) Variable

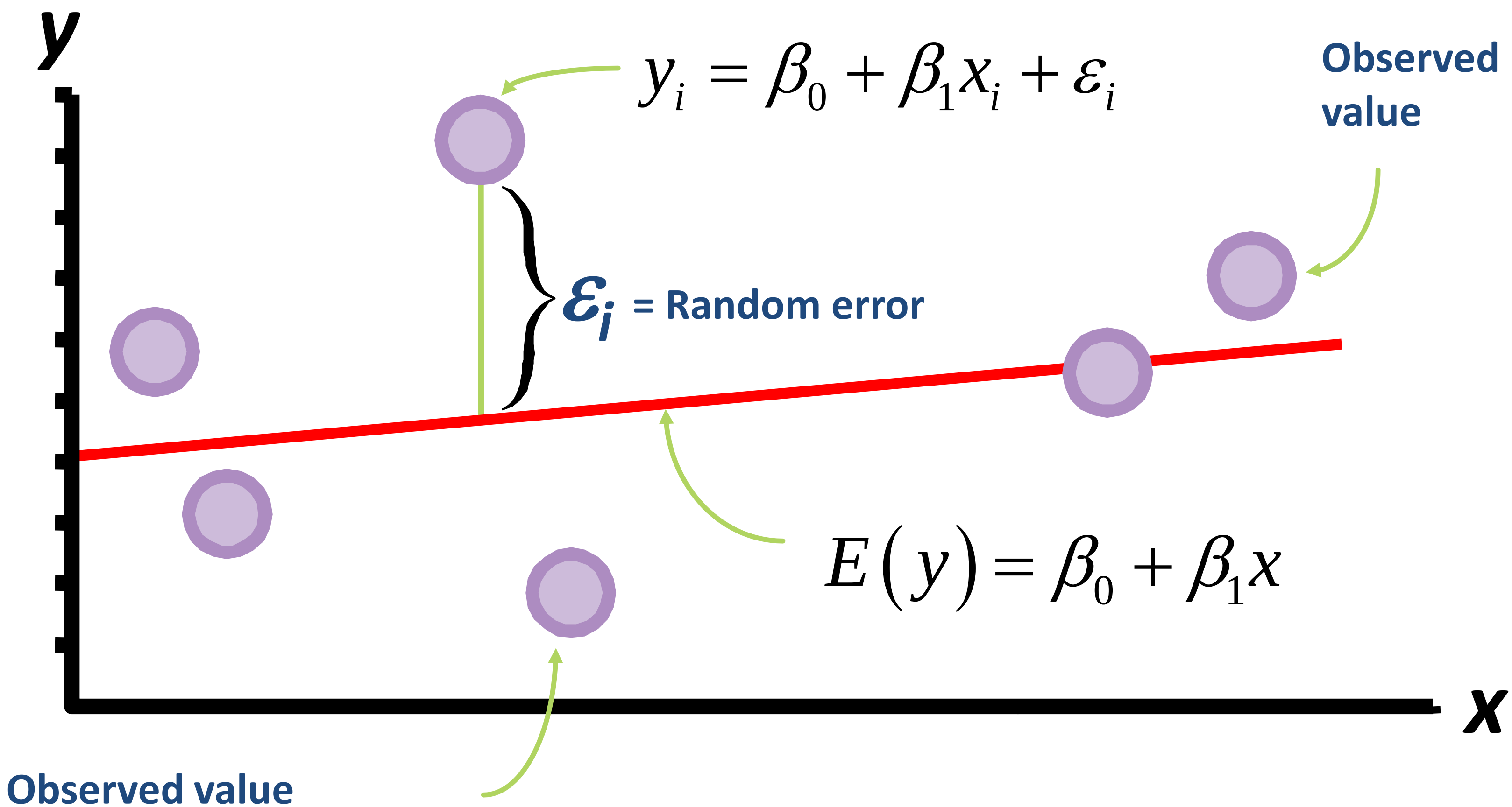
Line of Means



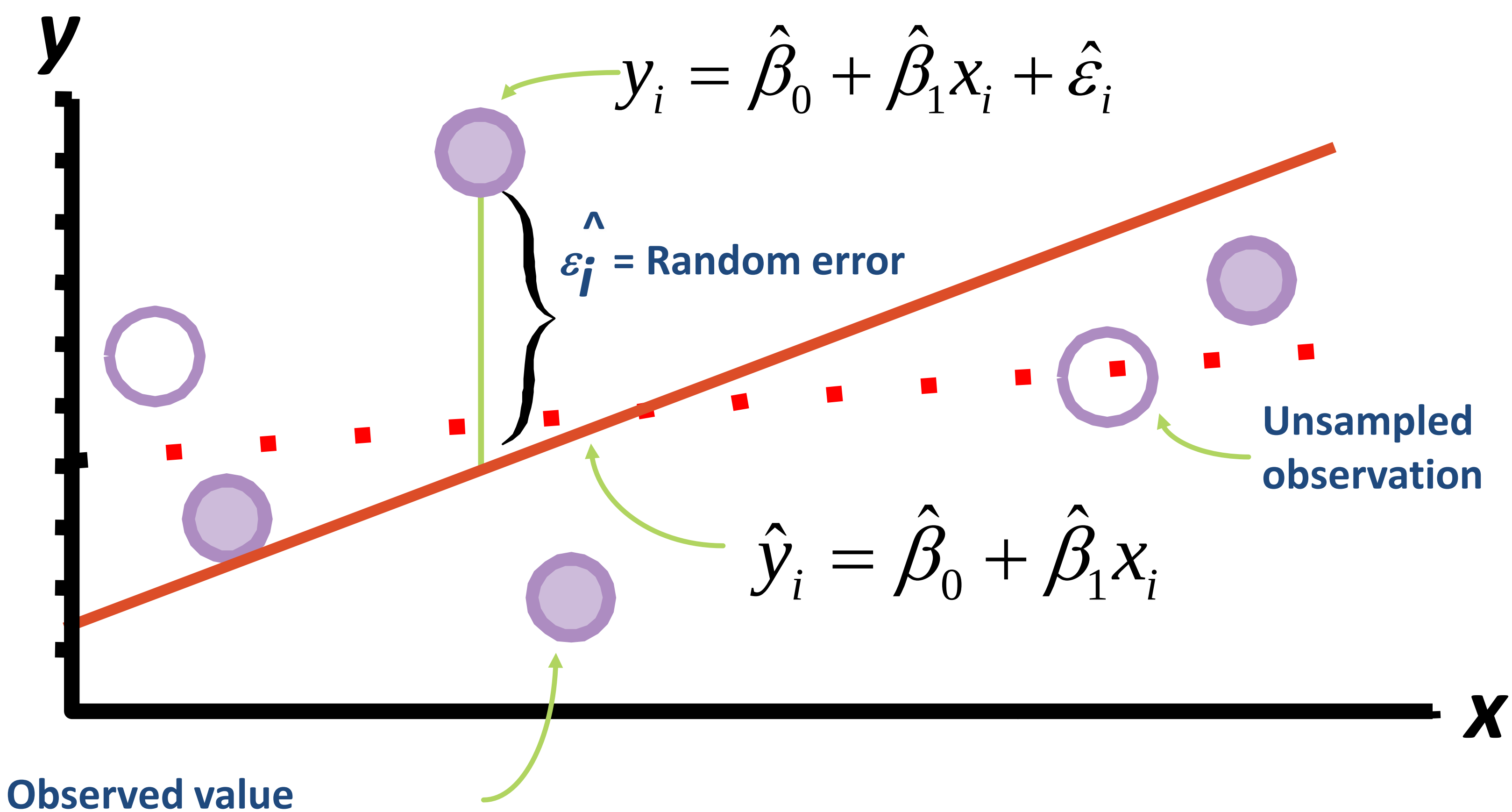
Population & Sample Regression Models



Population Linear Regression Model



Sample Linear Regression Model



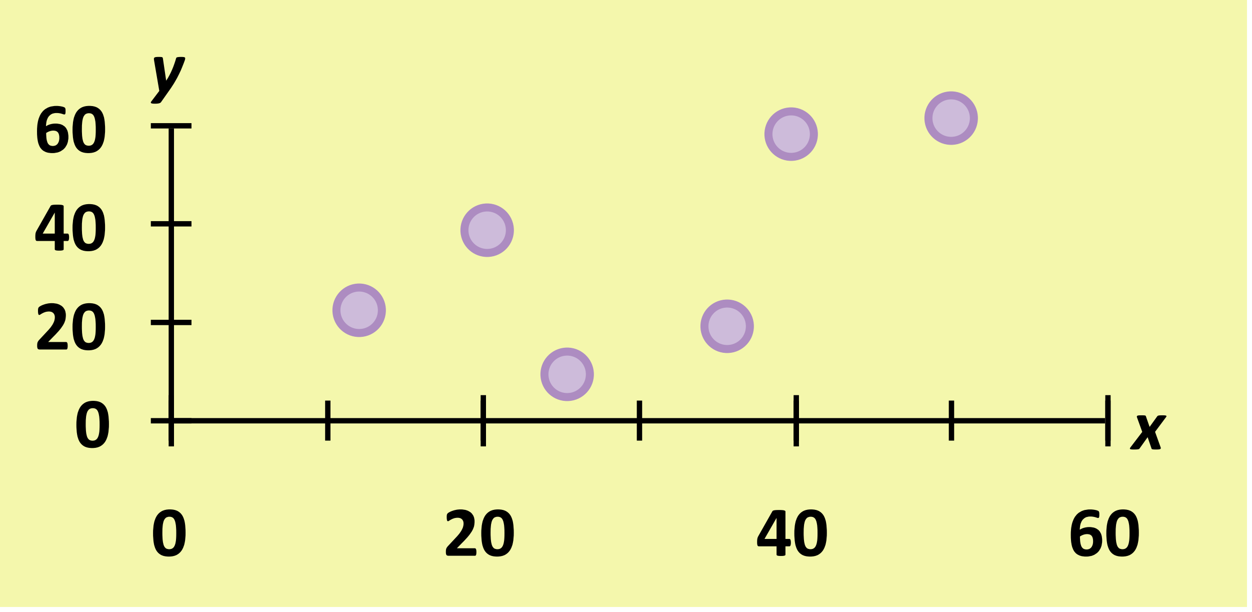
Estimating Parameters: Least Squares Method

Regression Modeling Steps

1. Hypothesize deterministic component
2. **Estimate unknown model parameters**
3. Specify probability distribution of random error term
   - Estimate standard deviation of error
4. Evaluate model
5. Use model for prediction and estimation

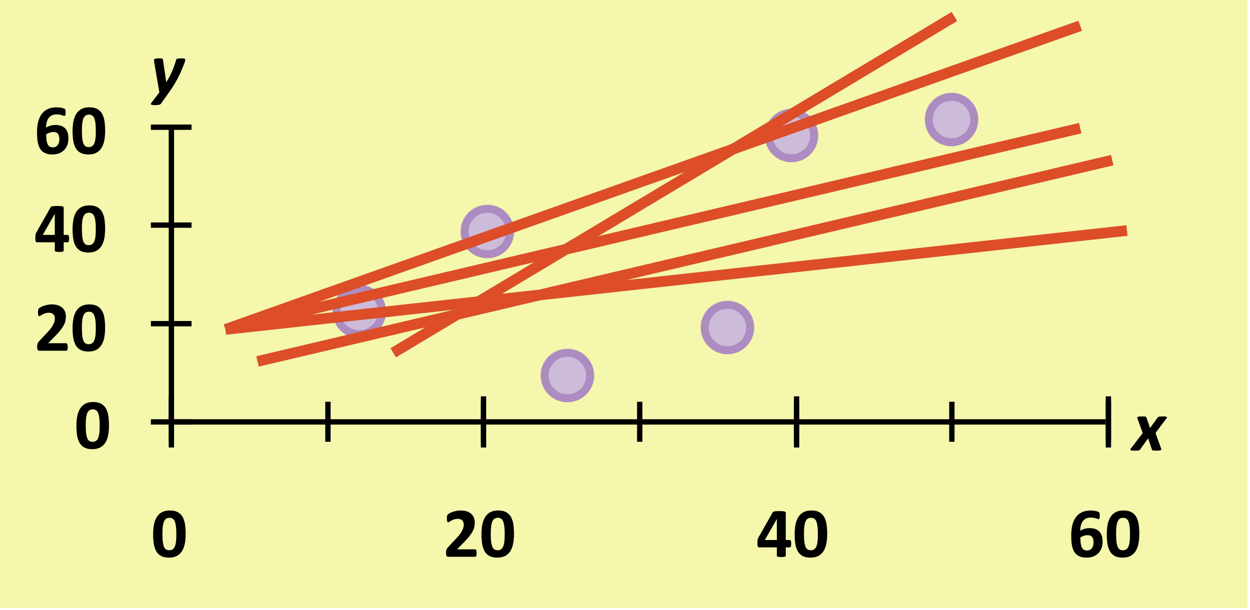
Scattergram

1. Plot of all (x_i, y_i) pairs
2. 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’?



Least Squares

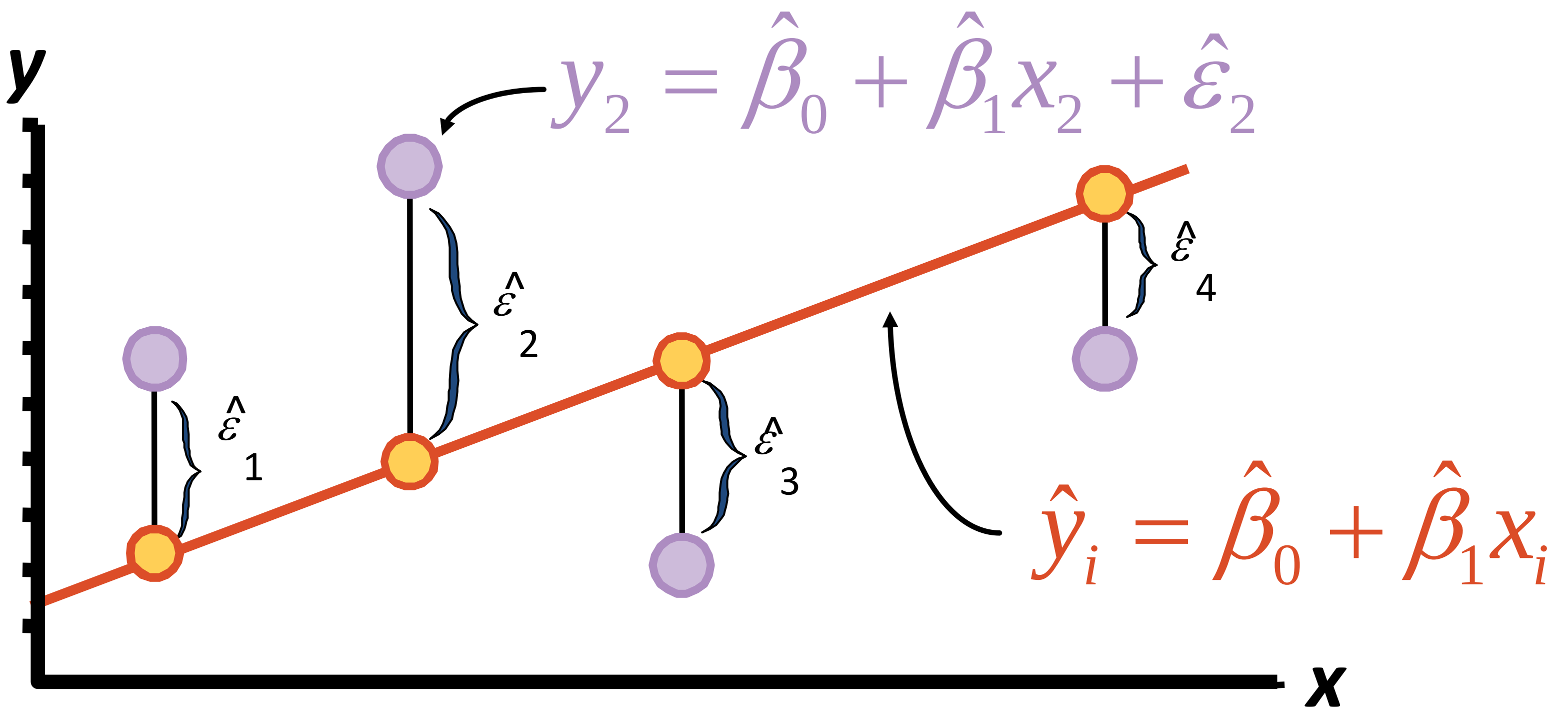
- 'Best fit' means difference between actual *y* values and predicted *y* values are a minimum
  - *But* positive differences off-set negative

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

- Least Squares minimizes the Sum of the Squared Differences (SSE)

Least Squares Graphically

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



Coefficient Equations

Prediction Equation $$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Slope $$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

***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

1. Slope ($\hat{\beta}_1$)
   - Estimated *y* changes by $\hat{\beta}_1$ for each 1unit increase in *x*
     - If $\hat{\beta}_1 = 2$, then Sales (*y*) is expected to increase by 2 for each 1 unit increase in Advertising (*x*)

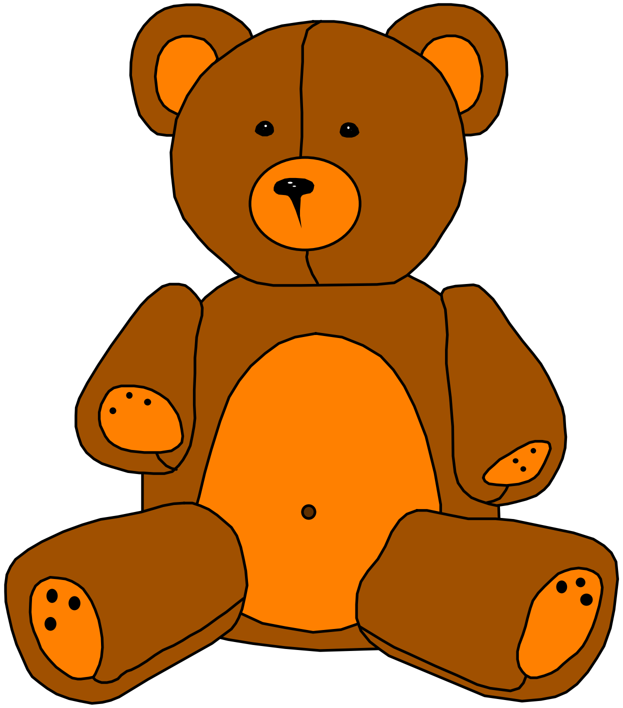
2. *Y*-Intercept ($\hat{\beta}_0$)
   - Average value of *y* when *x* = 0
     - If $\hat{\beta}_0 = 4$, then Average Sales (*y*) is expected to be 4 when Advertising (*x*) is 0

Least Squares Example

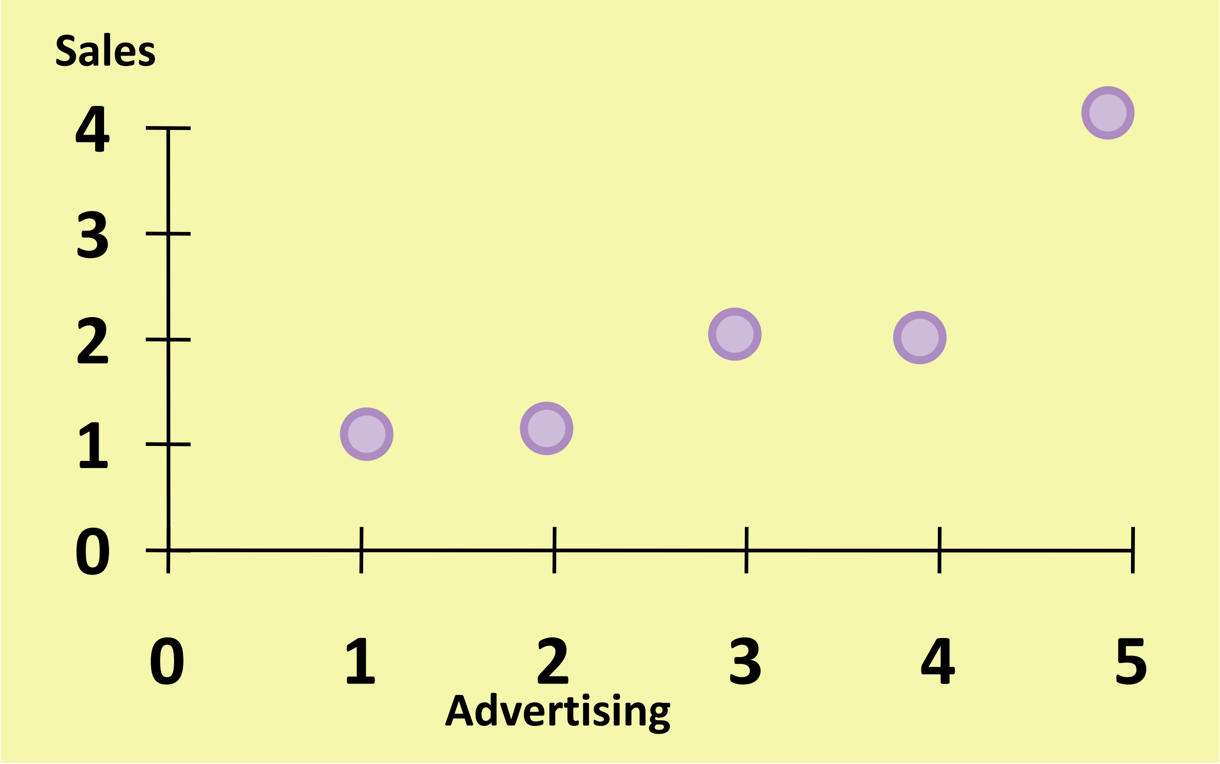
You're a marketing analyst for Hasbro Toys. You gather the following data:

Ad $	**Sales (Units)**
1	**1**
2	**1**
3	**2**
4	**2**
5	**4**

Find the **least squares line** relating sales and advertising.



Scattergram
Sales vs. Advertising



Parameter Estimation Solution Table

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	1	1	1	1
2	1	4	1	2
3	2	9	4	6
4	2	16	4	8
5	4	25	16	20
15	10	55	26	37

Parameter Estimation Solution

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{37 - \frac{(15)(10)}{5}}{55 - \frac{(15)^2}{5}} = .70$$

$$\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x} = 2 - (.70)(3) = -.10$$

$$\hat{y} = -.1 + .7x$$

Parameter Estimation Computer Output

Parameter Estimates

$\hat{\beta}_0$

Variable	DF	Parameter Estimate	Standard Error	T for H0: Param=0	Prob>|T|
INTERCEP	1	-0.1000	0.6350	-0.157	0.8849
ADVERT	1	0.7000	0.1914	3.656	0.0354

$\hat{\beta}_1$

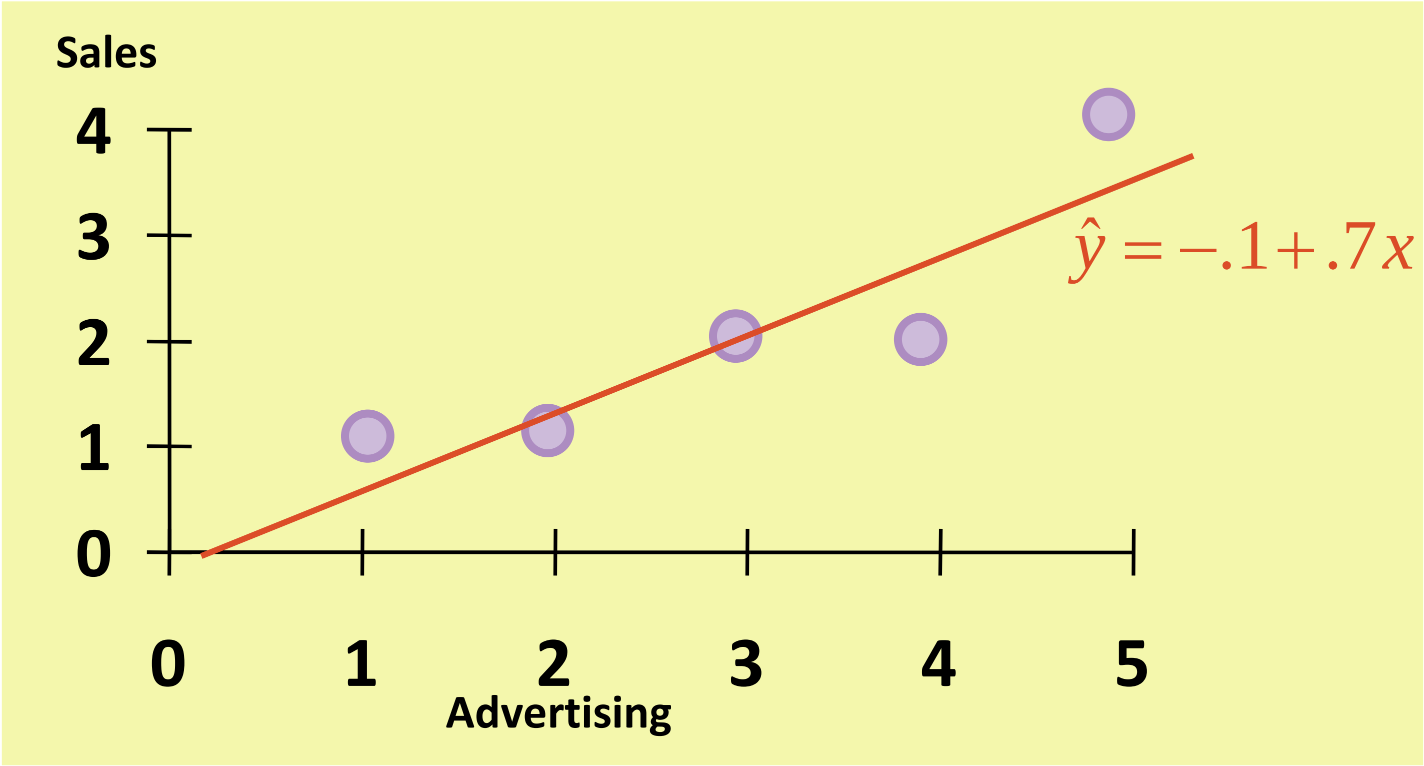
$$\hat{y} = -.1 + .7x$$

Coefficient Interpretation Solution

1. Slope ($\hat{\beta}_1$)
    - Sales Volume (*y*) is expected to increase by .7 units for each $1 increase in Advertising (*x*)

2. *Y*-Intercept ($\hat{\beta}_0$)
    - Average value of Sales Volume (*y*) is -.10 units when Advertising (*x*) is 0
        - Difficult to explain to marketing manager
        - Expect some sales without advertising

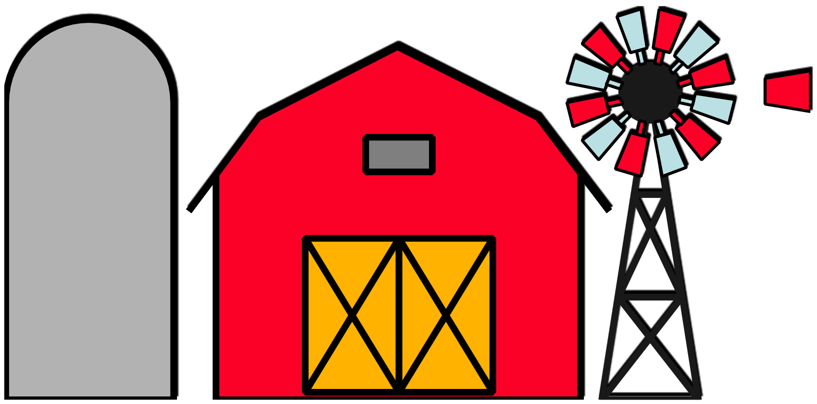
Regression Line Fitted to the Data



Least Squares Thinking Challenge

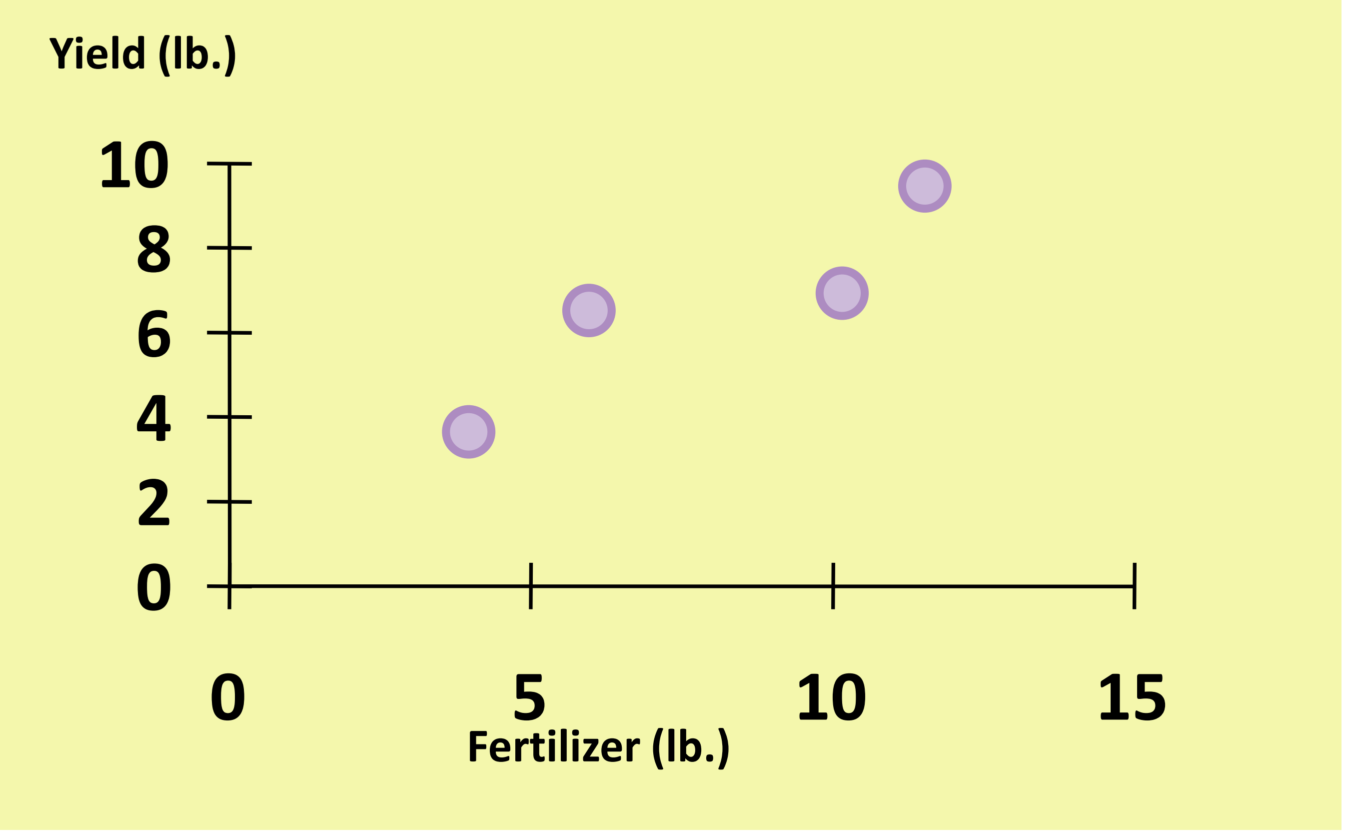
You're an economist for the county cooperative. You gather the following data:

Fertilizer (lb.)	Yield (lb.)
4	3.0
6	5.5
10	6.5
12	9.0



Find the **least squares line** relating crop yield and fertilizer.

Scattergram
Crop Yield vs. Fertilizer*



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*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{218 - \frac{(32)(24)}{4}}{296 - \frac{(32)^2}{4}} = .65$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 6 - (.65)(8) = .80$$

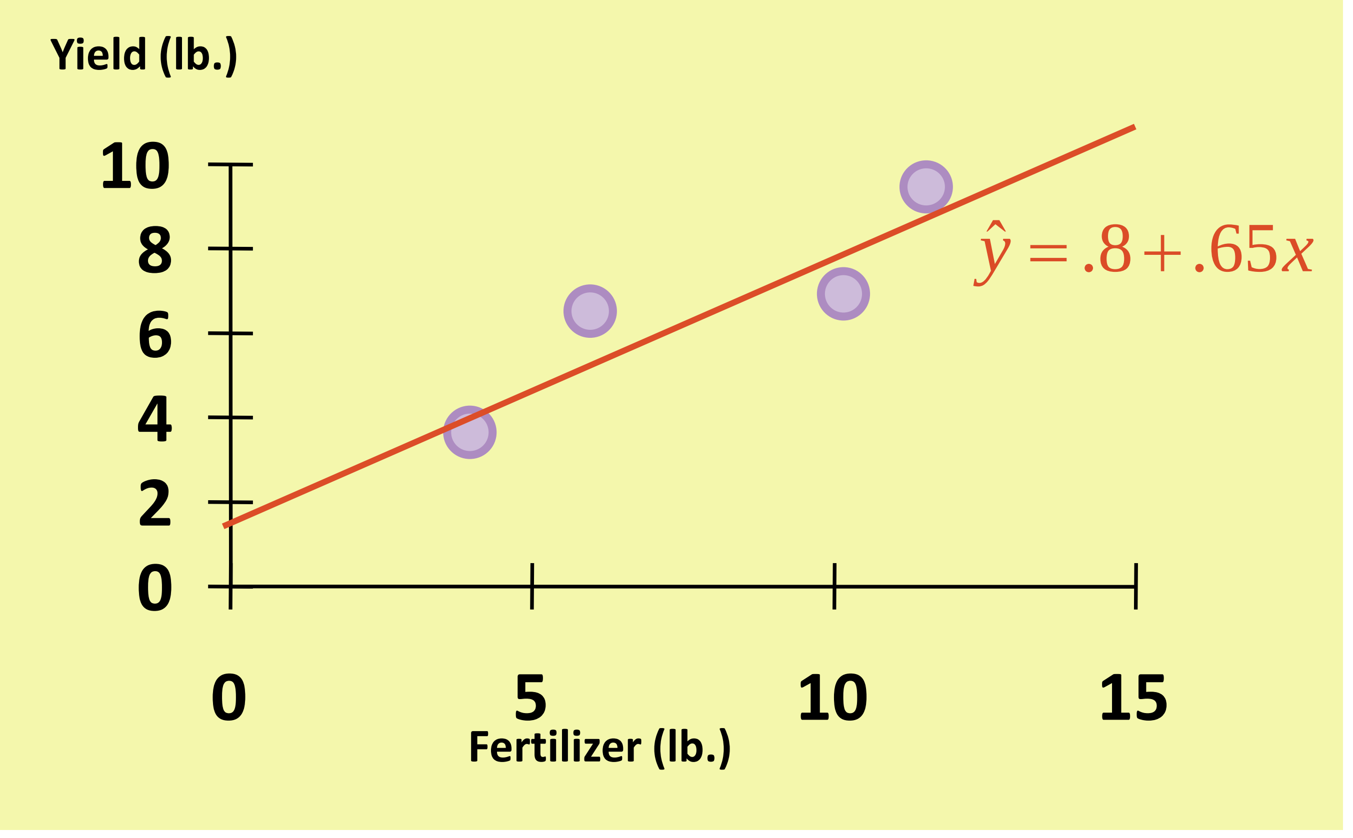
$$\hat{y} = .8 + .65x$$

Coefficient Interpretation Solution*

1. Slope ($\hat{\beta}_1$)
   - Crop Yield (y) is expected to increase by .65 lb. for each 1 lb. increase in Fertilizer (x)

2. Y-Intercept ($\hat{\beta}_0$)
   - Average Crop Yield (y) is expected to be 0.8 lb. when no Fertilizer (x) is used

Regression Line Fitted to the Data*



Probability Distribution of Random Error

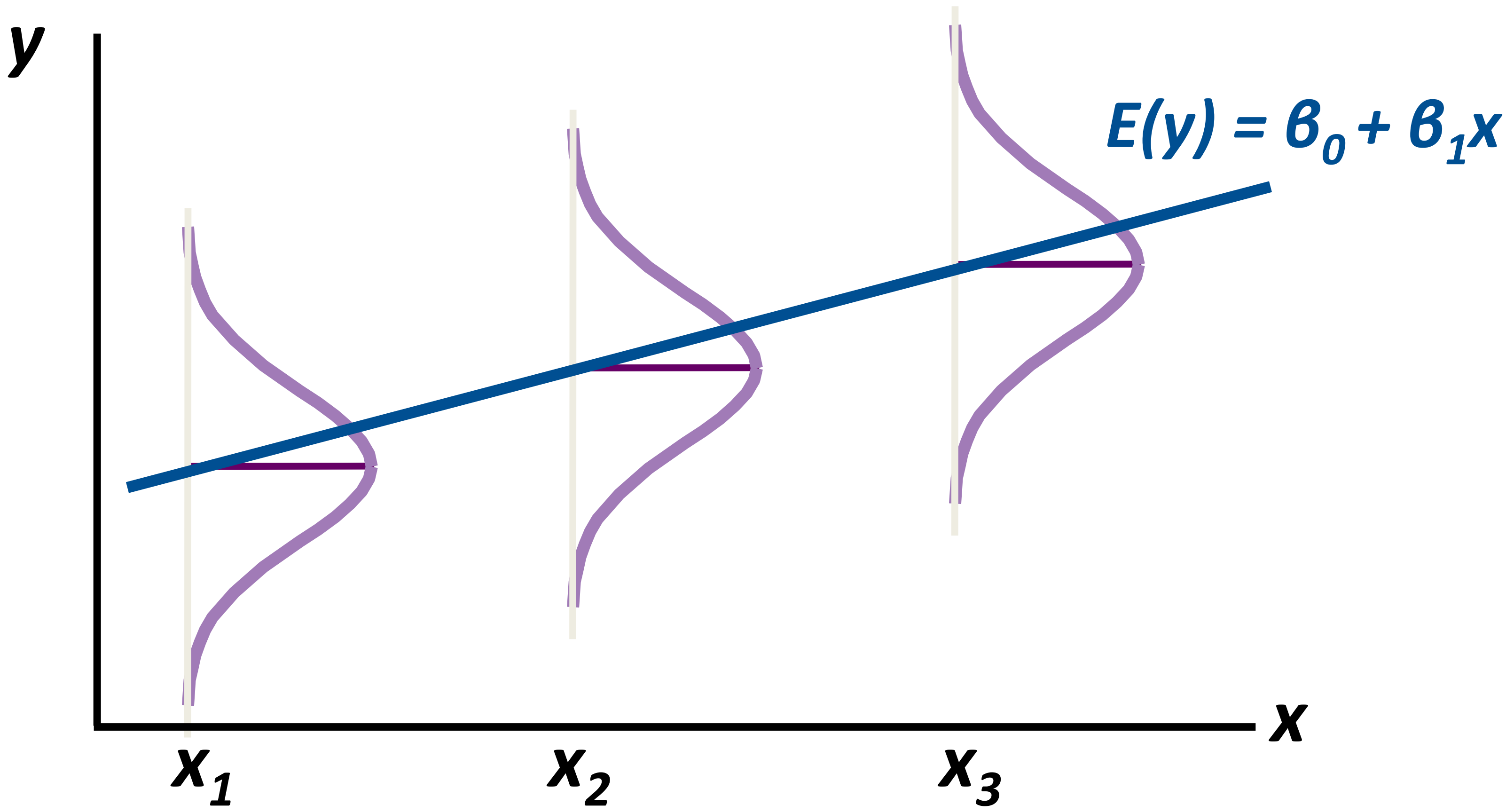
Regression Modeling Steps

1. Hypothesize deterministic component
2. Estimate unknown model parameters
3. **Specify probability distribution of random error term**
    - Estimate standard deviation of error
4. Evaluate model
5. Use model for prediction and estimation

Linear Regression Assumptions

1. Mean of probability distribution of error, ε, is 0
2. Probability distribution of error has constant variance
3. Probability distribution of error, ε, is normal
4. Errors are independent

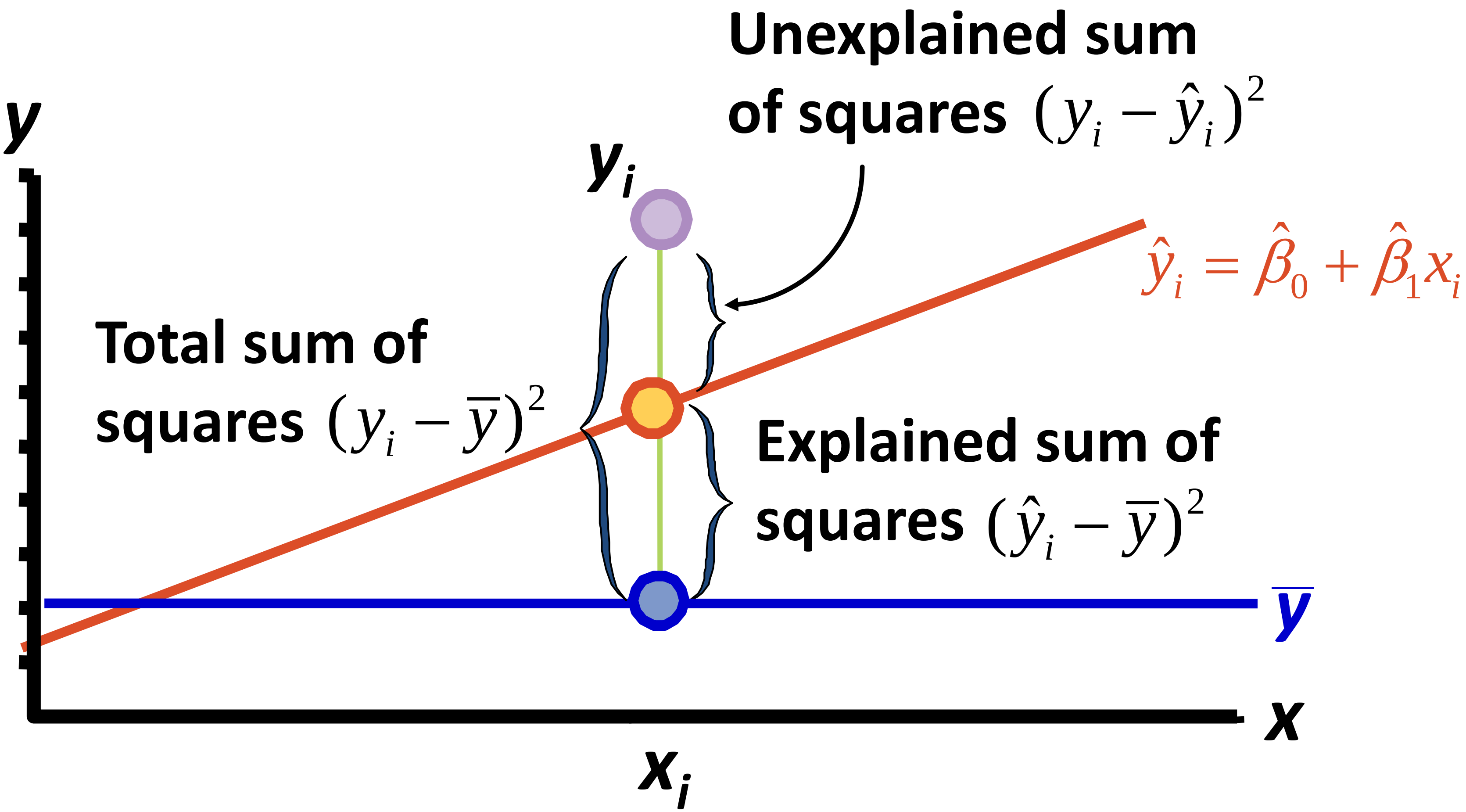
Error
Probability Distribution



Random Error Variation

- Variation of actual *y* from predicted *y*, $\hat{y}$
- Measured by standard error of regression model
  - Sample standard deviation of $\hat{\varepsilon}$: s
- Affects several factors
  - Parameter significance
  - Prediction accuracy

Variation Measures



Estimation of σ^2

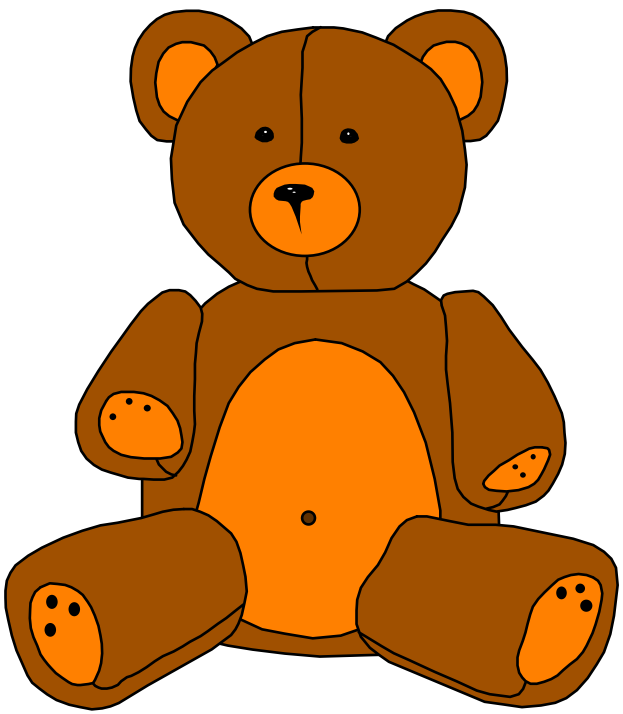
$$s^2 = \frac{SSE}{n-2} \quad where \quad SSE = \sum \left(y_i - \hat{y}_i \right)^2$$

$$s = \sqrt{s^2} = \sqrt{\frac{SSE}{n-2}}$$

Calculating SSE, s^2, s Example

You're a marketing analyst for Hasbro Toys.
You gather the following data:

Ad $	Sales (Units)
1	1
2	1
3	2
4	2
5	4



Find **SSE**, **s^2**, and **s**.

Calculating SSE Solution

x_i	y_i	$\hat{y} = -.1 + .7x$	$y - \hat{y}$	$(y - \hat{y})^2$
1	1	.6	.4	.16
2	1	1.3	-.3	.09
3	2	2	0	0
4	2	2.7	-.7	.49
5	4	3.4	.6	.36
				SSE=1.1

Calculating s^2 and s Solution

$$s^2 = \frac{SSE}{n-2} = \frac{1.1}{5-2} = .36667$$

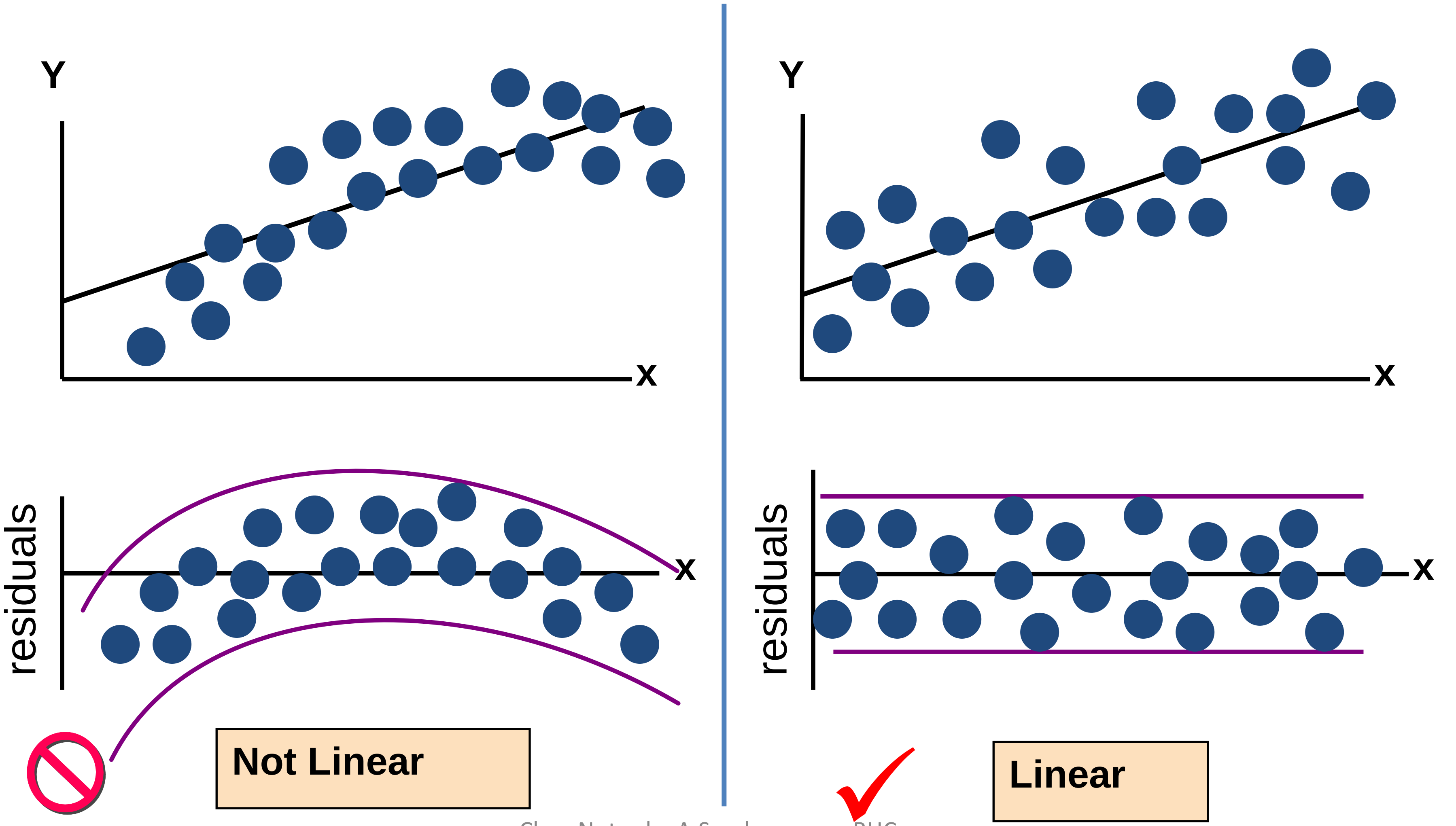
$$s = \sqrt{.36667} = .6055$$

Residual Analysis

$$\boxed{e_i = Y_i - \hat{Y}_i}$$

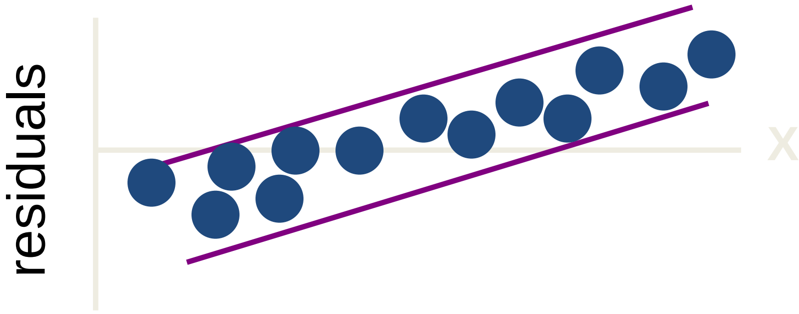
- The residual for observation i, e_i, is the difference between its observed and predicted value
- Check the assumptions of regression by examining the residuals
  - Examine for linearity assumption
  - Evaluate independence assumption
  - Evaluate normal distribution assumption
  - Examine for constant variance for all levels of X (homoscedasticity)

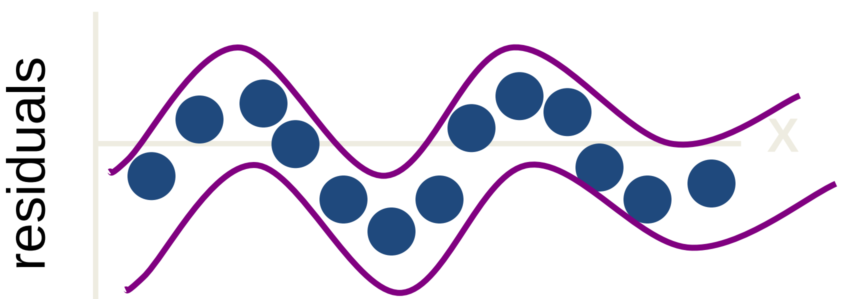
Residual Analysis for Linearity



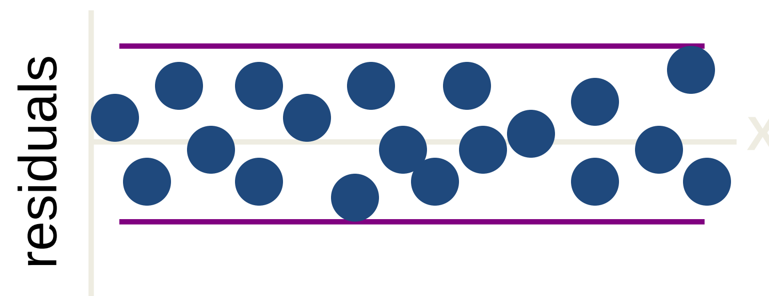
Residual Analysis for Independence

Not Independent





Independent

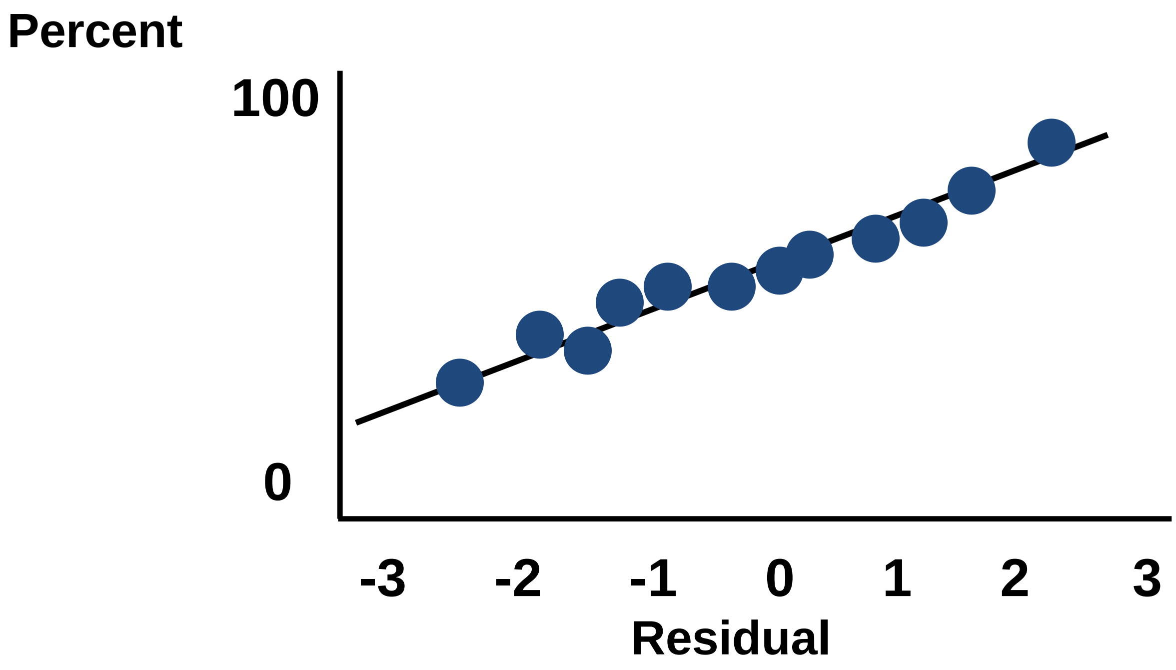


Checking for Normality

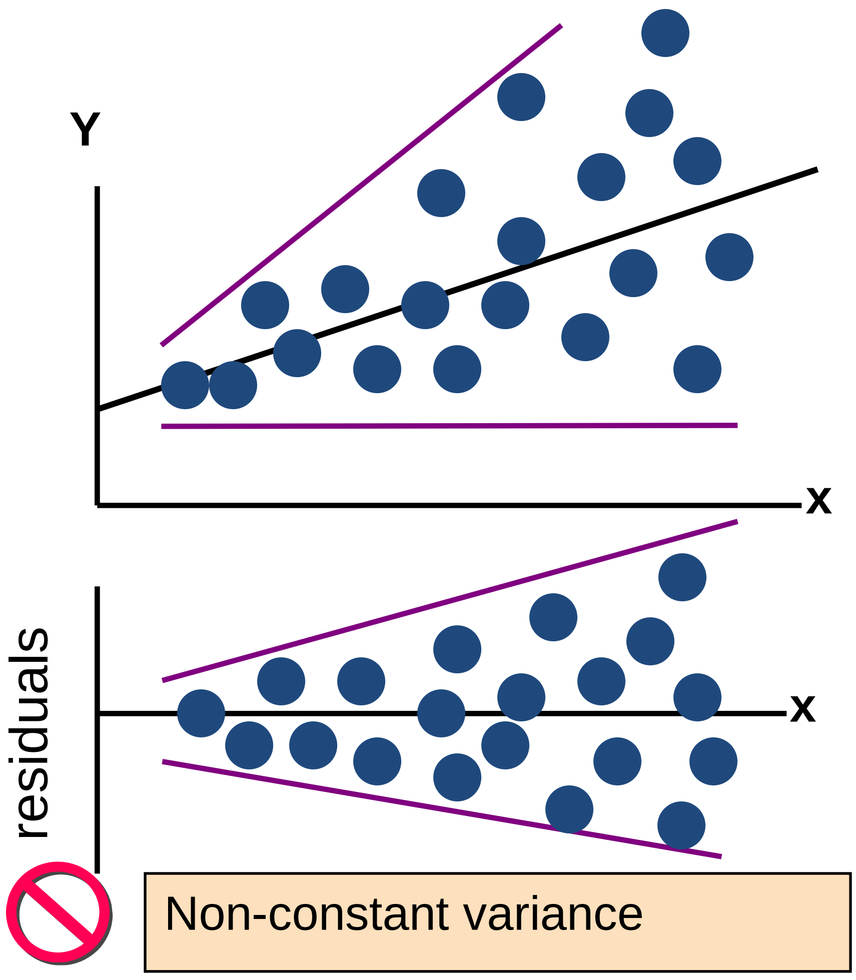
- Examine the Stem-and-Leaf Display of the Residuals
- Examine the Boxplot of the Residuals
- Examine the Histogram of the Residuals
- Construct a Normal Probability Plot of the Residuals

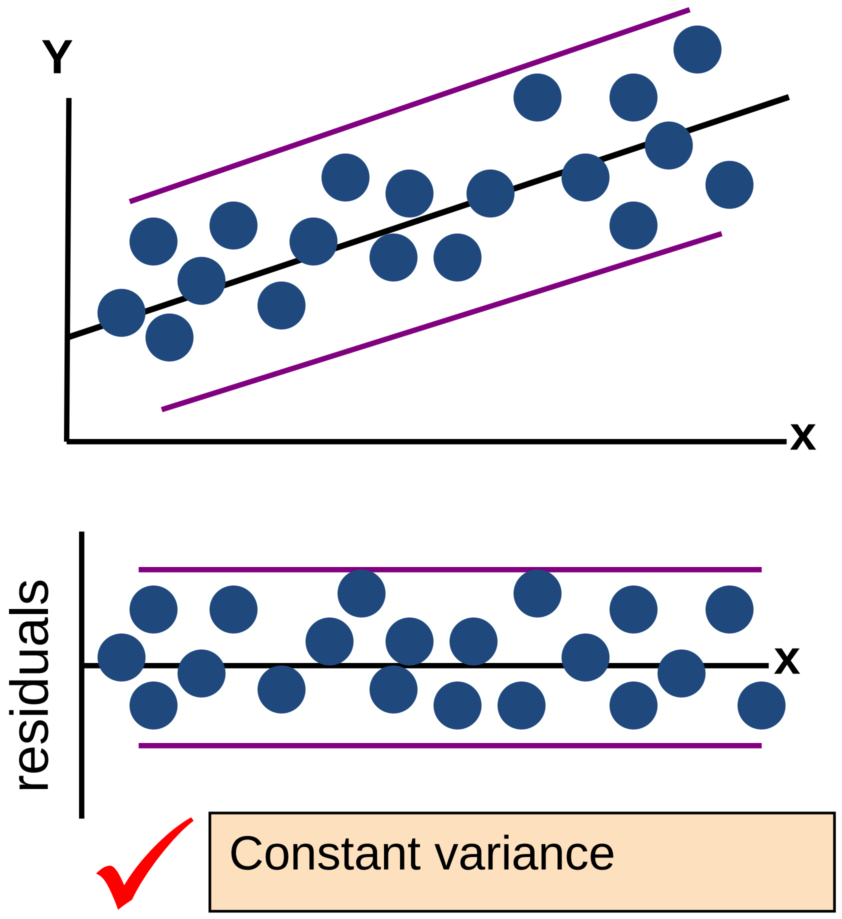
Residual Analysis for Normality

When using a normal probability plot, normal errors will approximately display in a straight line



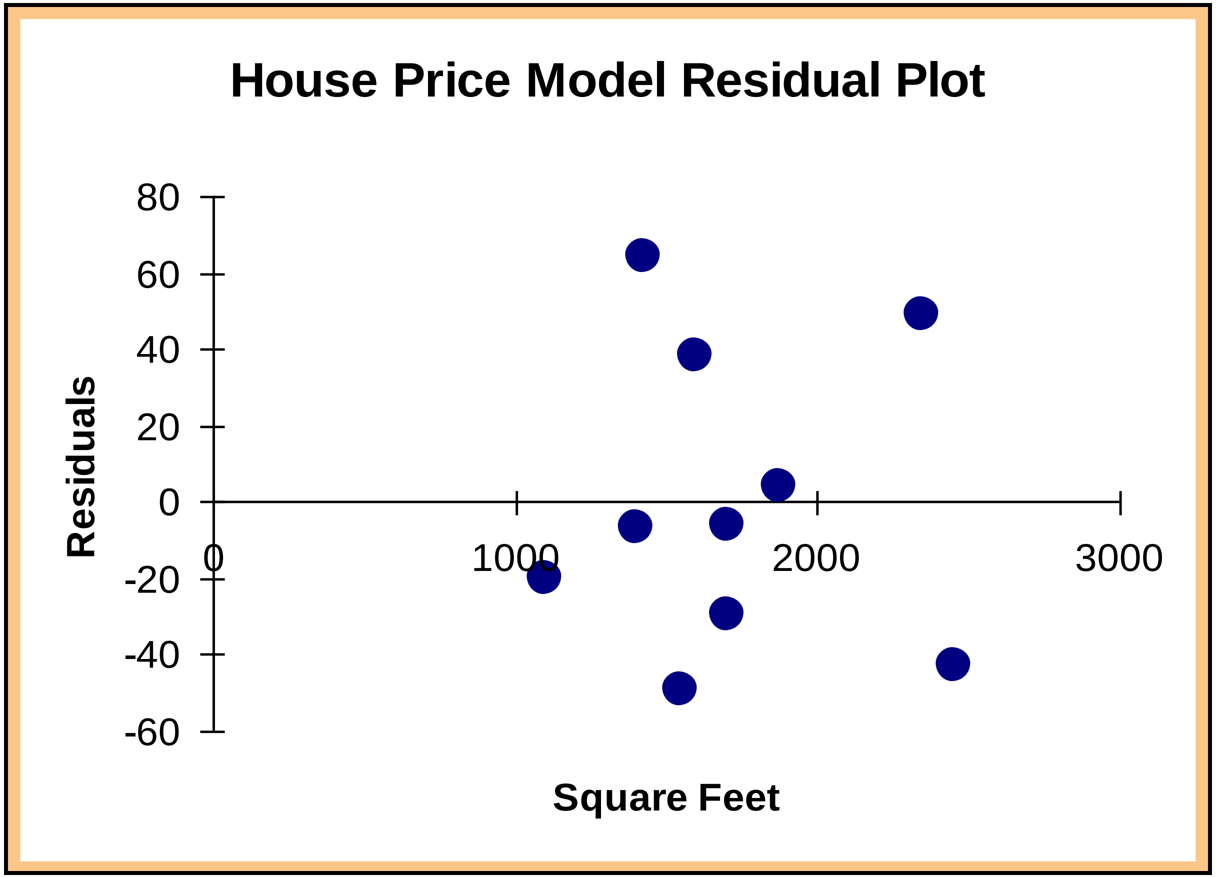
Residual Analysis for Equal Variance





Simple Linear Regression Example: Excel Residual Output

RESIDUAL OUTPUT		
	Predicted House Price	*Residuals*
1	251.92316	-6.923162
2	273.87671	38.12329
3	284.85348	-5.853484
4	304.06284	3.937162
5	218.99284	-19.99284
6	268.38832	-49.38832
7	356.20251	48.79749
8	367.17929	-43.17929
9	254.6674	64.33264
10	284.85348	-29.85348



Does not appear to violate any regression assumptions

Evaluating the Model

Testing for Significance

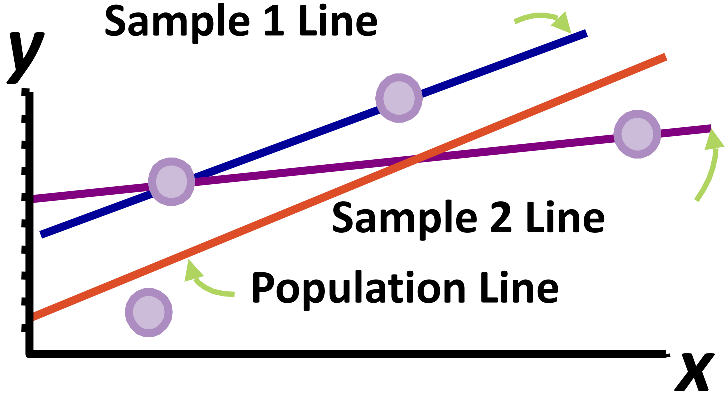
Regression Modeling Steps

1. Hypothesize deterministic component
2. Estimate unknown model parameters
3. Specify probability distribution of random error term
   - Estimate standard deviation of error
4. **Evaluate model**
5. Use model for prediction and estimation

Test of Slope Coefficient

- Shows if there is a linear relationship between *x* and *y*
- Involves population slope β_1
- Hypotheses
  - H_0: $\beta_1 = 0$ (No Linear Relationship)
  - H_a: $\beta_1 \neq 0$ (Linear Relationship)
- Theoretical basis is sampling distribution of slope

Sampling Distribution of Sample Slopes

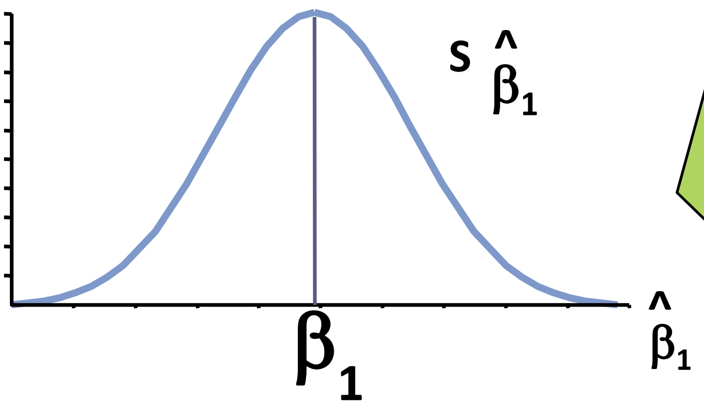


All Possible Sample Slopes

Sample 1:	2.5
Sample 2:	1.6
Sample 3:	1.8
Sample 4:	2.1
:	:

Very large number of sample slopes

Sampling Distribution



Slope Coefficient Test Statistic

$$t = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{s \big/ \sqrt{SS_{xx}}} \qquad df = n - 2$$

where

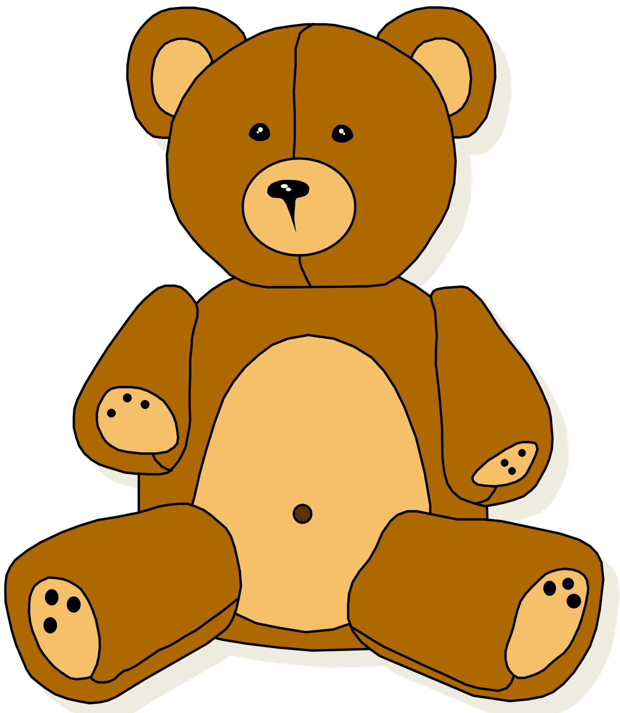
$$SS_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

Test of Slope Coefficient Example

You're a marketing analyst for Hasbro Toys.

You find $\hat{\beta}_0 = -.1$, $\hat{\beta}_1 = .7$ and $s = .6055$.

Ad $	Sales (Units)
1	1
2	1
3	2
4	2
5	4



Is the relationship **significant** at the **.05** level of significance?

Solution Table

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	1	1	1	1
2	1	4	1	2
3	2	9	4	6
4	2	16	4	8
5	4	25	16	20
15	10	55	26	37

Multiple Linear Regression Models

Introduction

- Many applications of regression analysis involve situations in which there are more than one regressor variable.
- A regression model that contains more than one regressor variable is called a **multiple regression model**.

The Multiple Linear Regression Model

- Regression applications in which there are several independent variables, $x_1, x_2, \ldots, x_k$. A multiple linear regression model with p independent variables has the equation

$$\mu_y = \beta_o + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$$

  – β_i is the intercept and β_i determines the contribution of the independent variable x_i

- The ε is a random variable with mean 0 and variance σ^2.

The Prediction Equation

- The equation for this model fitted to data is

$$\hat{y} = b_o + b_1 x_1 + \ldots + b_p x_p$$

- Where $\hat{y}$ denotes the "predicted" value computed from the equation, and b_i denotes an estimate of β_i.

- As with Simple Linear Regression, they're obtained by the method of **least squares**
  - Among the set of all possible values for the parameter estimates, I find the ones which *minimize* the sum of squared residuals.

Multiple Linear Regression Models

Introduction

In general, the **dependent variable** or **response** Y may be related to k **independent** or **regressor variables.** The model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \qquad (12\text{-}2)$$

is called a multiple linear regression model with k regressor variables. The parameters β_j, $j = 0, 1, \ldots, k$, are called the regression coefficients. This model describes a hyperplane in the k-dimensional space of the regressor variables $\{x_j\}$. The parameter β_j represents the expected change in response Y per unit change in x_j when all the remaining regressors x_i $(i \neq j)$ are held constant.

Multiple Linear Regression Models

Least Squares Estimation of the Parameters

The **method of least squares** may be used to estimate the regression coefficients in the multiple regression model, Equation 12-2. Suppose that $n > k$ observations are available, and let x_{ij} denote the ith observation or level of variable x_j. The observations are

$$(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i), \qquad i = 1, 2, \ldots, n \quad \text{and} \quad n > k$$

It is customary to present the data for multiple regression in a table such as Table 12-1.

Table 12-1 Data for Multiple Linear Regression

y	x_1	x_2	...	x_k
y_1	x_{11}	x_{12}	...	x_{1k}
y_2	x_{21}	x_{22}	...	x_{2k}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
y_n	x_{n1}	x_{n2}	...	x_{nk}

Multiple Linear Regression Models

Least Squares Estimation of the Parameters

- The least squares function is given by

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2$$

- The least squares estimates must satisfy

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) = 0$$

and

$$\left.\frac{\partial L}{\partial \beta_j}\right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) x_{ij} = 0 \quad j = 1, 2, \ldots, k$$

Multiple Linear Regression Models

Matrix Approach to Multiple Linear Regression

Suppose the model relating the regressors to the response is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i \qquad i = 1, 2, \ldots, n$$

In matrix notation this model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Multiple Linear Regression Models

Matrix Approach to Multiple Linear Regression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Multiple Linear Regression Models

Matrix Approach to Multiple Linear Regression

We wish to find the vector of least squares estimators that minimizes:

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

The resulting least squares estimate is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{y} \tag{12-13}$$

Multiple Linear Regression Models

Matrix Approach to Multiple Linear Regression

The fitted regression model is

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \qquad i = 1, 2, \ldots, n \tag{12-14}$$

In matrix notation, the fitted model is

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

The difference between the observation y_i and the fitted value $\hat{y}_i$ is a **residual,** say, $e_i = y_i - \hat{y}_i$. The $(n \times 1)$ vector of residuals is denoted by

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} \tag{12-15}$$

Multiple Linear Regression Models

Table 12-2 Wire Bond Data for Example 12-1

Observation Number	Pull Strength y	Wire Length x_1	Die Height x_2	Observation Number	Pull Strength y	Wire Length x_1	Die Height x_2
1	9.95	2	50	14	11.66	2	360
2	24.45	8	110	15	21.65	4	205
3	31.75	11	120	16	17.89	4	400
4	35.00	10	550	17	69.00	20	600
5	25.02	8	295	18	10.30	1	585
6	16.86	4	200	19	34.93	10	540
7	14.38	2	375	20	46.59	15	250
8	9.60	2	52	21	44.88	15	290
9	24.35	9	100	22	54.12	16	510
10	27.50	8	300	23	56.63	17	590
11	17.08	4	412	24	22.13	6	100
12	37.00	11	400	25	21.15	5	400
13	41.95	12	500				

Multiple Linear Regression Models

Example

In Example 12-1, we illustrated fitting the multiple regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

where y is the observed pull strength for a wire bond, x_1 is the wire length, and x_2 is the die height. The 25 observations are in Table 12-2. We will now use the matrix approach to fit the regression model above to these data. The model matrix $\mathbf{X}$ and $\mathbf{y}$ vector for this model are

Example

$$\mathbf{X} = \begin{bmatrix} 1 & 2 & 50 \\ 1 & 8 & 110 \\ 1 & 11 & 120 \\ 1 & 10 & 550 \\ 1 & 8 & 295 \\ 1 & 4 & 200 \\ 1 & 2 & 375 \\ 1 & 2 & 52 \\ 1 & 9 & 100 \\ 1 & 8 & 300 \\ 1 & 4 & 412 \\ 1 & 11 & 400 \\ 1 & 12 & 500 \\ 1 & 2 & 360 \\ 1 & 4 & 205 \\ 1 & 4 & 400 \\ 1 & 20 & 600 \\ 1 & 1 & 585 \\ 1 & 10 & 540 \\ 1 & 15 & 250 \\ 1 & 15 & 290 \\ 1 & 16 & 510 \\ 1 & 17 & 590 \\ 1 & 6 & 100 \\ 1 & 5 & 400 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 9.95 \\ 24.45 \\ 31.75 \\ 35.00 \\ 25.02 \\ 16.86 \\ 14.38 \\ 9.60 \\ 24.35 \\ 27.50 \\ 17.08 \\ 37.00 \\ 41.95 \\ 11.66 \\ 21.65 \\ 17.89 \\ 69.00 \\ 10.30 \\ 34.93 \\ 46.59 \\ 44.88 \\ 54.12 \\ 56.63 \\ 22.13 \\ 21.15 \end{bmatrix}$$

Multiple Linear Regression Models

The $\mathbf{X}'\mathbf{X}$ matrix is

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 2 & 8 & \cdots & 5 \\ 50 & 110 & \cdots & 400 \end{bmatrix} \begin{bmatrix} 1 & 2 & 50 \\ 1 & 8 & 110 \\ \vdots & \vdots & \vdots \\ 1 & 5 & 400 \end{bmatrix}$$

$$= \begin{bmatrix} 25 & 206 & 8{,}294 \\ 206 & 2{,}396 & 77{,}177 \\ 8{,}294 & 77{,}177 & 3{,}531{,}848 \end{bmatrix}$$

and the $\mathbf{X}'\mathbf{y}$ vector is

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 2 & 8 & \cdots & 5 \\ 50 & 110 & \cdots & 400 \end{bmatrix} \begin{bmatrix} 9.95 \\ 24.45 \\ \vdots \\ 21.15 \end{bmatrix} = \begin{bmatrix} 725.82 \\ 8{,}008.47 \\ 274{,}816.71 \end{bmatrix}$$

The least squares estimates are found from Equation 12-13 as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Multiple Linear Regression Models

or

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 25 & 206 & 8{,}294 \\ 206 & 2{,}396 & 77{,}177 \\ 8{,}294 & 77{,}177 & 3{,}531{,}848 \end{bmatrix}^{-1} \begin{bmatrix} 725.82 \\ 8{,}008.37 \\ 274{,}811.31 \end{bmatrix}$$

$$= \begin{bmatrix} 0.214653 & -0.007491 & -0.000340 \\ -0.007491 & 0.001671 & -0.000019 \\ -0.000340 & -0.000019 & +0.0000015 \end{bmatrix} \begin{bmatrix} 725.82 \\ 8{,}008.47 \\ 274{,}811.31 \end{bmatrix}$$

$$= \begin{bmatrix} 2.26379143 \\ 2.74426964 \\ 0.01252781 \end{bmatrix}$$

Therefore, the fitted regression model with the regression coefficients rounded to five decimal places is

$$\hat{y} = 2.26379 + 2.74427x_1 + 0.01253x_2$$

This is identical to the results obtained in Example 12-1.

Multiple Linear Regression Models

This regression model can be used to predict values of pull strength for various values of wire length (x_1) and die height (x_2). We can also obtain the **fitted values** $\hat{y}_i$ by substituting each observation (x_{i1}, x_{i2}), $i = 1, 2, \ldots, n$, into the equation. For example, the first observation has $x_{11} = 2$ and $x_{12} = 50$, and the fitted value is

$$\begin{aligned}\hat{y}_1 &= 2.26379 + 2.74427x_{11} + 0.01253x_{12}\\ &= 2.26379 + 2.74427(2) + 0.01253(50)\\ &= 8.38\end{aligned}$$

The corresponding observed value is $y_1 = 9.95$. The *residual* corresponding to the first observation is

$$\begin{aligned}e_1 &= y_1 - \hat{y}_1\\ &= 9.95 - 8.38\\ &= 1.57\end{aligned}$$

Table 12-3 displays all 25 fitted values $\hat{y}_i$ and the corresponding residuals. The fitted values and residuals are calculated to the same accuracy as the original data.

Multiple Linear Regression Models

Table 12-3 Observations, Fitted Values, and Residuals for Example 12-2

Observation Number	y_i	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$	Observation Number	y_i	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$
1	9.95	8.38	1.57	14	11.66	12.26	−0.60
2	24.45	25.60	−1.15	15	21.65	15.81	5.84
3	31.75	33.95	−2.20	16	17.89	18.25	−0.36
4	35.00	36.60	−1.60	17	69.00	64.67	4.33
5	25.02	27.91	−2.89	18	10.30	12.34	−2.04
6	16.86	15.75	1.11	19	34.93	36.47	−1.54
7	14.38	12.45	1.93	20	46.59	46.56	0.03
8	9.60	8.40	1.20	21	44.88	47.06	−2.18
9	24.35	28.21	−3.86	22	54.12	52.56	1.56
10	27.50	27.98	−0.48	23	56.63	56.31	0.32
11	17.08	18.40	−1.32	24	22.13	19.98	2.15
12	37.00	37.46	−0.46	25	21.15	21.00	0.15
13	41.95	41.46	0.49				

Multiple Linear Regression Models

Estimating σ^2

An unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p} = \frac{SS_E}{n-p} \qquad (12\text{-}16)$$

Multiple Linear Regression Models

Properties of the Least Squares Estimators

Unbiased estimators:

$$
\begin{aligned}
E(\hat{\boldsymbol{\beta}}) &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\
&= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})] \\
&= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}] \\
&= \boldsymbol{\beta}
\end{aligned}
$$

Covariance Matrix:

$$
\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} C_{00} & C_{01} & C_{02} \\ C_{10} & C_{11} & C_{12} \\ C_{20} & C_{21} & C_{22} \end{bmatrix}
$$

Multiple Linear Regression Models

Properties of the Least Squares Estimators

Individual variances and covariances:

$$V(\hat{\beta}_j) = \sigma^2 C_{jj}, \qquad j = 0, 1, 2$$

$$\text{cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 C_{ij}, \qquad i \neq j$$

In general,

$$\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 \mathbf{C}$$

Logistic Regression

- Regression used to fit a curve to data in which the dependent variable is binary, or dichotomous
- Typical application: Medicine
  - We might want to predict response to treatment, where we might code survivors as 1 and those who don't survive as 0

Example

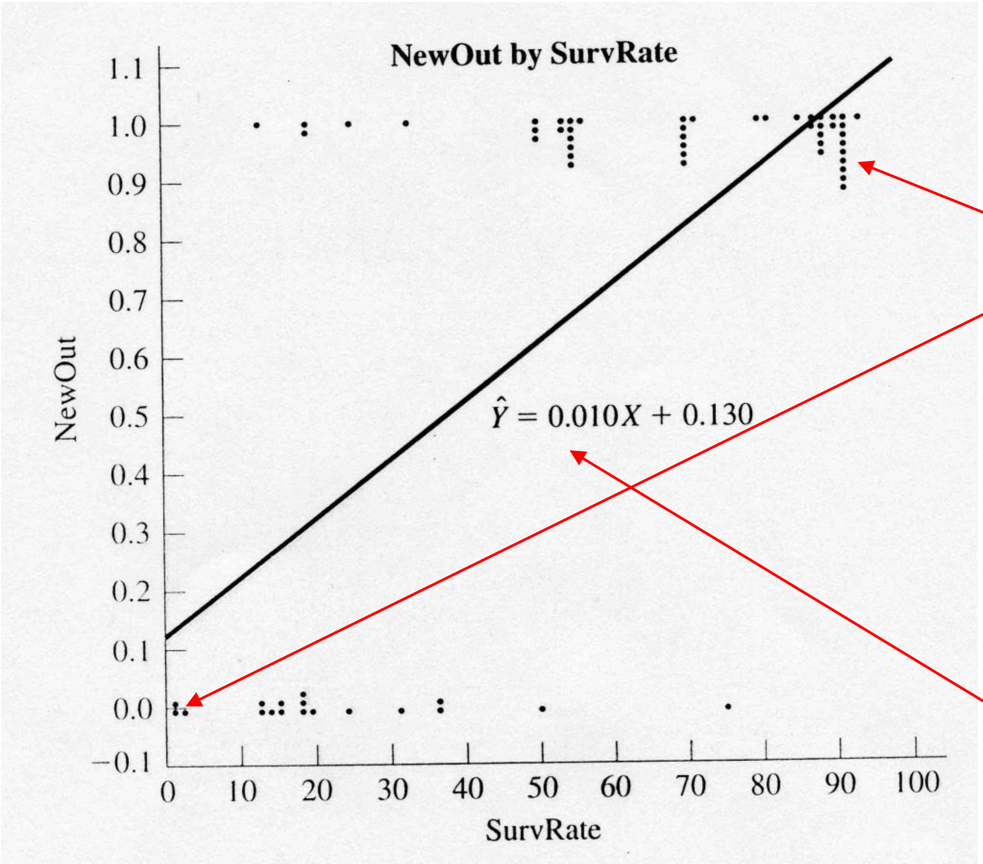


FIGURE 15·7 *Outcome as a function of SurvRate*

Observations:
For each value of SurvRate, the number of dots is the number of patients with that value of NewOut

Regression:
Standard linear regression

Problem: extending the regression line a few units left or right along the X axis produces predicted probabilities that fall outside of [0,1]

A Better Solution

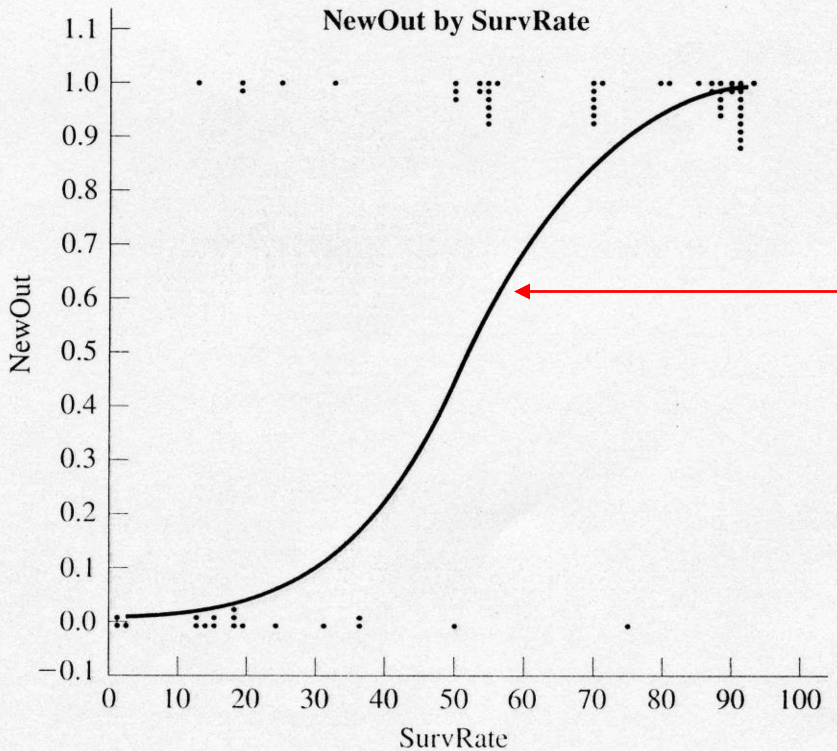


FIGURE 15.8 *More appropriate regression line for predicting outcome*

<u>Regression Curve</u>: Sigmoid function!

(bounded by asymptotes *y*=0 and *y*=1)

Odds

- Given some event with probability *p* of being 1, the odds of that event are given by:

$$\text{odds} = p\ /\ (1 - p)$$

- Consider the following data

Delinquent

Testosterone		**Yes**	**No**	**Total**
	Normal	402	3614	4016
	High	101	345	446
		503	3959	4462

- The odds of being delinquent if you are in the Normal group are:

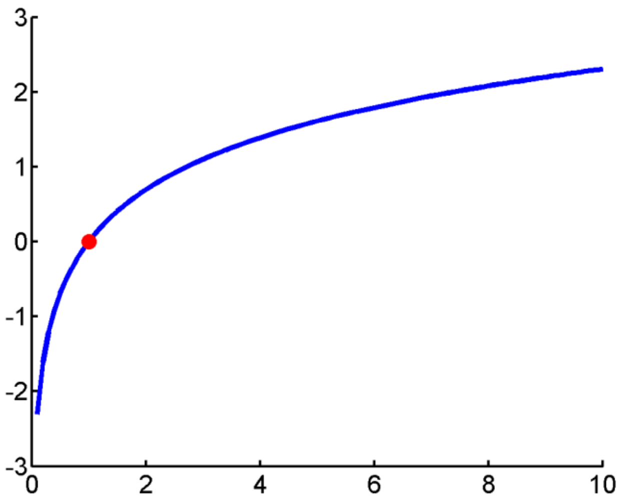
pdelinquent/(1–pdelinquent) = (402/4016) / (1 - (402/4016)) = 0.1001 / 0.8889 = 0.111

Odds Ratio

- The odds of being not delinquent in the Normal group is the reciprocal of this:
  - 0.8999/0.1001 = 8.99
- Now, for the High testosterone group
  - odds(delinquent) = 101/345 = 0.293
  - odds(not delinquent) = 345/101 = 3.416
- When we go from Normal to High, the odds of being delinquent nearly triple:
  - Odds ratio: 0.293/0.111 = 2.64
  - 2.64 times more likely to be delinquent with high testosterone levels

Logit Transform

- The logit is the natural log of the odds



- logit(p) = ln(odds) = ln (p/(1-p))

Logistic Regression

- In logistic regression, we seek a model:

$$\text{logit}(p) = b_0 + b_1 X$$

- That is, the log odds (logit) is assumed to be linearly related to the independent variable X
- So, now we can focus on solving an ordinary (linear) regression!

Recovering Probabilities

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

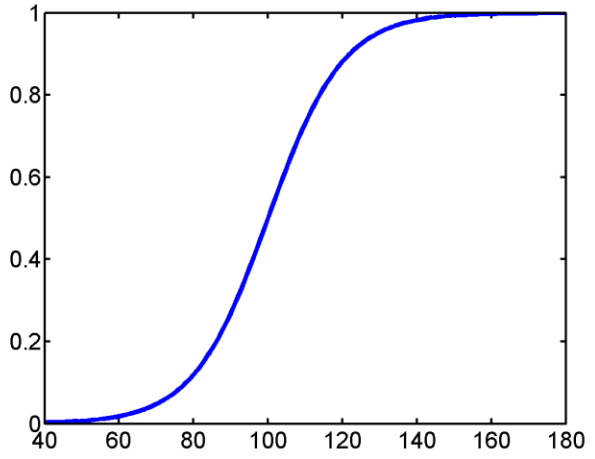
$$\Leftrightarrow \frac{p}{1-p} = e^{\beta_0 + \beta_1 X}$$

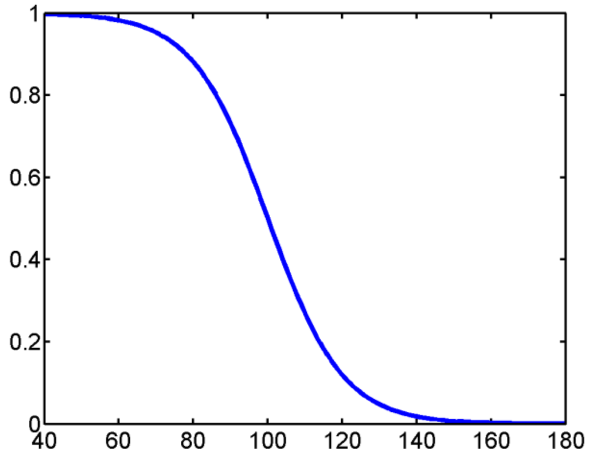
$$\Leftrightarrow p = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

which gives *p* as a sigmoid function!

Logistic Response Function

- When the response variable is binary, the shape of the response function is often sigmoidal:





Interpretation of β1

- Let:
  - odds1 = odds for value X (p/(1–p))
  - odds2 = odds for value X + 1 unit
- Then:

$$\frac{odds2}{odds1} = \frac{e^{b_0 + b_1(X+1)}}{e^{b_0 + b_1 X}}$$

$$= \frac{e^{(b_0 + b_1 X) + b_1}}{e^{b_0 + b_1 X}} = \frac{e^{(b_0 + b_1 X)} e^{b_1}}{e^{b_0 + b_1 X}} = e^{b_1}$$

- Hence, the exponent of the slope describes the proportionate rate at which the predicted odds ratio changes with each successive unit of X

Sample Calculations

- Suppose a cancer study yields:
  - log odds = –2.6837 + 0.0812 SurvRate
- Consider a patient with SurvRate = 40
  - log odds = –2.6837 + 0.0812(40) = 0.5643
  - odds = $e^{0.5643}$ = 1.758
  - patient is 1.758 times more likely to be improved than not
- Consider another patient with SurvRate = 41
  - log odds = –2.6837 + 0.0812(41) = 0.6455
  - odds = $e^{0.6455}$ = 1.907
  - patient's odds are 1.907/1.758 = 1.0846 times (or 8.5%) better than those of the previous patient
- Using probabilities
  - p40 = 0.6374 and p41 = 0.6560
  - Improvements appear different with odds and with *p*

Poisson Regression

Review of Regression

You may have come across:

Dependent Variable	Regression Model
Continuous	Linear
Binary	Logistic
Multicategory (unordered) (nominal variable)	Multinomial Logit
Multicategory (ordered) (ordinal variable)	Cumulative Logit

Dependent Variable	Regression Model
Continuous	Linear
Binary	Logistic
Multicategory (unordered) (nominal variable)	Multinomial Logit
Multicategory (ordered) (ordinal variable)	Cumulative Logit
Count variable	**Poisson Regression (Log-linear model)**

Poisson Regression

- In many cases the dependent variable is of the count type, such as:
  - ➢The number of phone calls made in a year
  - ➢The number of visits to a national park in a year
  - ➢The number of accidents occurred in a year
- The underlying variable is discrete, taking only a finite non-negative number of values.
- In many cases the count is 0 for several observations.
- Each count example is measured over a certain finite time period.

Data

Data for this session are assumed to be:

- A **count** variable Y (e.g. number of accidents, number of suicides)
- One **categorical** variable (X) with C possible categories (e.g. days of week, months)
- Hence Y has C possible outcomes $y_1, y_2, \ldots, y_C$

Probability Distributions Used For Count Data

- Poisson Probability Distribution: Regression models based on this probability distribution are known as Poisson Regression Models (PRM).

- Negative Binomial Probability Distribution: An alternative to PRM is the Negative Binomial Regression Model (NBRM), used to remedy some of the deficiencies of the PRM.

Poisson Regression Models

- If a discrete random variable Y follows the Poisson distribution, its probability density function (PDF) is given by:

$$f(Y = y_i) = \Pr(Y = y_i) = \frac{e^{-\lambda_i} \lambda_i^{\,y_i}}{y_i!}, \;\; y_i = 0,1,2\ldots$$

  where $f(Y|y_i)$ denotes the probability that the discrete random variable Y takes non-negative integer value y_i, and λ is the parameter of the Poisson distribution.

- **Equidispersion**: A unique feature of the Poisson distribution is that the mean and the variance of a Poisson-distributed variable are the same
  - If variance > mean, there is **overdispersion**

Poisson Regression Models

- The Poisson regression model can be written as:

$$y_i = E(y_i) + u_i = \lambda_i + u_i$$

- where the *y*s are independently distributed as Poisson random variables with mean λ for each individual expressed as:

- $\lambda_i = E(y_i|X_i) = \exp[B_1 + B_2X_{2i} + \ldots + B_kX_{ki}] = \exp(BX)$

- Taking the exponential of BX will guarantee that the mean value of the count variable, λ, will be positive.

Poisson model

Poisson model

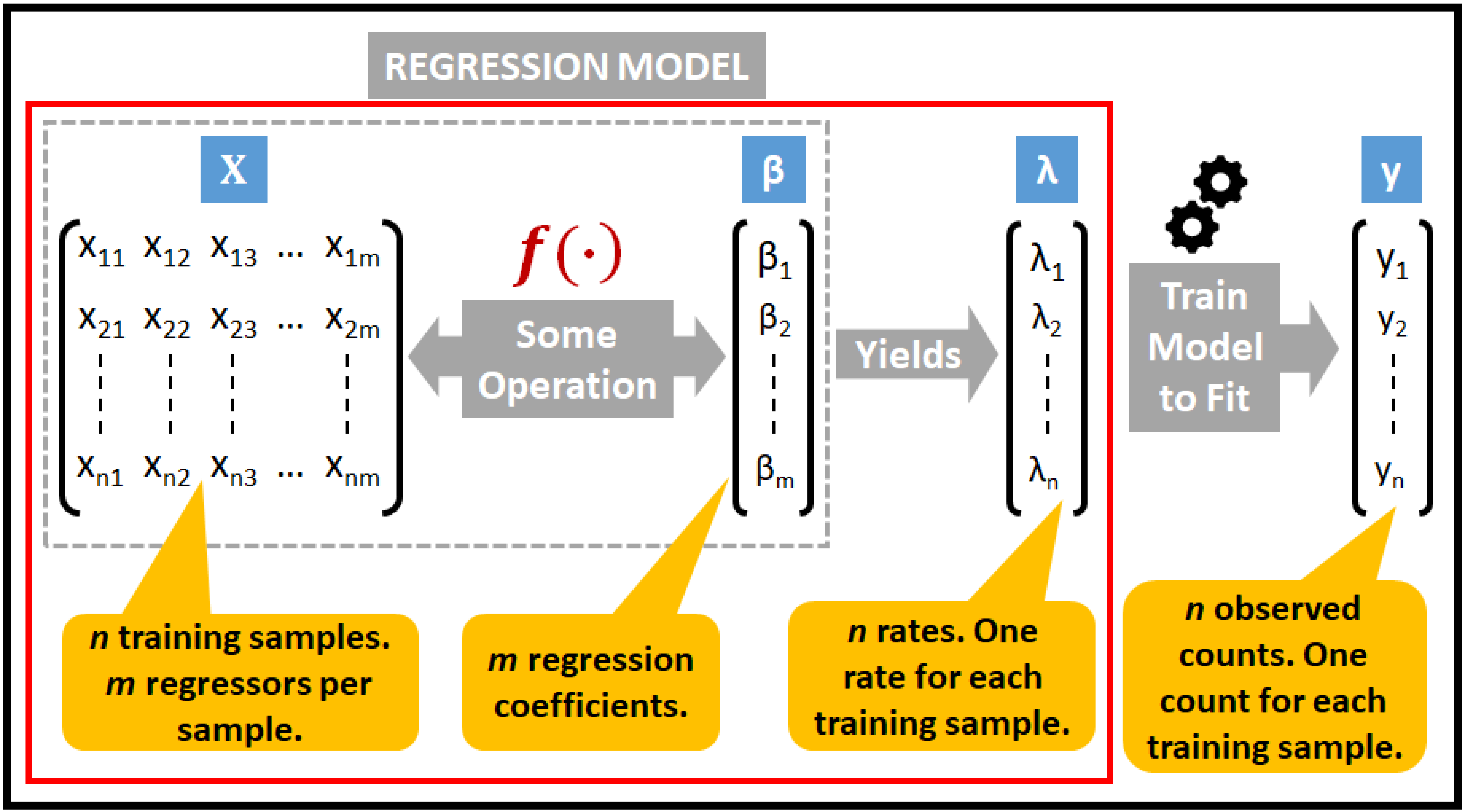
- The Poisson model predicts the number of occurrences of an event.
- The Poisson model states that the probability that the dependent variable Y will be equal to a certain number y is:

$$p(Y = y) = \frac{e^{-\mu}\mu^{y}}{y!}$$

- For the Poisson model, μ is the intensity or rate parameter.

$$\mu = \exp\left(\mathbf{x}_i'\beta\right)$$

- Interpretation of the coefficients: one unit increase in x will increase/decrease the average number of the dependent variable by the coefficient expressed as a percentage.


REGRESSION MODEL
X
β
λ
y
x_{11} x_{12} x_{13} … x_{1m}
x_{21} x_{22} x_{23} … x_{2m}
x_{n1} x_{n2} x_{n3} … x_{nm}
$f(\cdot)$
Some Operation
β_1 β_2 β_m
Yields
λ_1 λ_2 λ_n
Train Model to Fit
y_1 y_2 y_n
n training samples. *m* regressors per sample.
m regression coefficients.
n rates. One rate for each training sample.
n observed counts. One count for each training sample.

Event rate for the i^{th} sample

$$PMF(y_i|\boldsymbol{x}_i) = \frac{e^{-\lambda_i} * \lambda_i^{y_i}}{y_i!}$$

Probability of seeing count y_i given the regression vector x_i

Regression coefficients.
An (m x 1) matrix.

$$\lambda = e^{X\beta}$$

Event rates.
An (n x 1) matrix.

The training set.
An (n x m) matrix.

Training the Poisson regression model

- The technique for identifying the coefficients **β** is called **M**aximum **L**ikelihood **E**stimation (MLE).

Date	Day	High Temp (°F)	Low Temp (°F)	Precipitation	Brooklyn Bridge
6/1	Thursday	78.1	62.1	0.00	3,468
6/2	Friday	73.9	60.1	0.01	3,271
6/3	Saturday	72.0	55	0.01	2,589
6/4	Sunday	68.0	60.1	0.09	1,805
6/5	Monday	66.9	60.1	0.02	2,171
6/6	Tuesday	55.9	53.1	0.06	1,193
6/7	Wednesday	[illegible]	[illegible]	[illegible]	[illegible]

$$P(3468|\boldsymbol{x_1}) = \frac{e^{-\lambda_1} * \lambda_1^{3468}}{3468!}$$

$$P(3271|\boldsymbol{x_2}) = \frac{e^{-\lambda_2} * \lambda_2^{3271}}{3271!}$$

$$P(2589|\boldsymbol{x_3}) = \frac{e^{-\lambda_3} * \lambda_3^{2589}}{2589!}$$

$$P(1805|\boldsymbol{x_4}) = \frac{e^{-\lambda_4} * \lambda_4^{1805}}{1805!}$$

$$\lambda_1 = e^{x_1\boldsymbol{\beta}}$$

$$\lambda_2 = e^{x_2\boldsymbol{\beta}}$$

$$\lambda_3 = e^{x_3\boldsymbol{\beta}}$$

$$\lambda_4 = e^{x_4\boldsymbol{\beta}}$$

Joint probability of observing all *n* counts

$$P(\boldsymbol{y}|\boldsymbol{X}) = P(3468|\boldsymbol{x_1}) * P(3271|\boldsymbol{x_2}) * P(2589|\boldsymbol{x_3}) * \ldots * P(2727|\boldsymbol{x_n})$$

$$\therefore L(\boldsymbol{\beta}) = P(\boldsymbol{y}|\boldsymbol{X}) = \frac{e^{-\lambda_1} * \lambda_1^{3468}}{3468!} * \frac{e^{-\lambda_2} * \lambda_2^{3271}}{3271!} * \frac{e^{-\lambda_3} * \lambda_3^{2589}}{2589!} * \ldots * \frac{e^{-\lambda_n} * \lambda_n^{2727}}{2727!}$$

Likelihood Function for β

Joint probability after plugging in the individual count probabilities

Unit 5

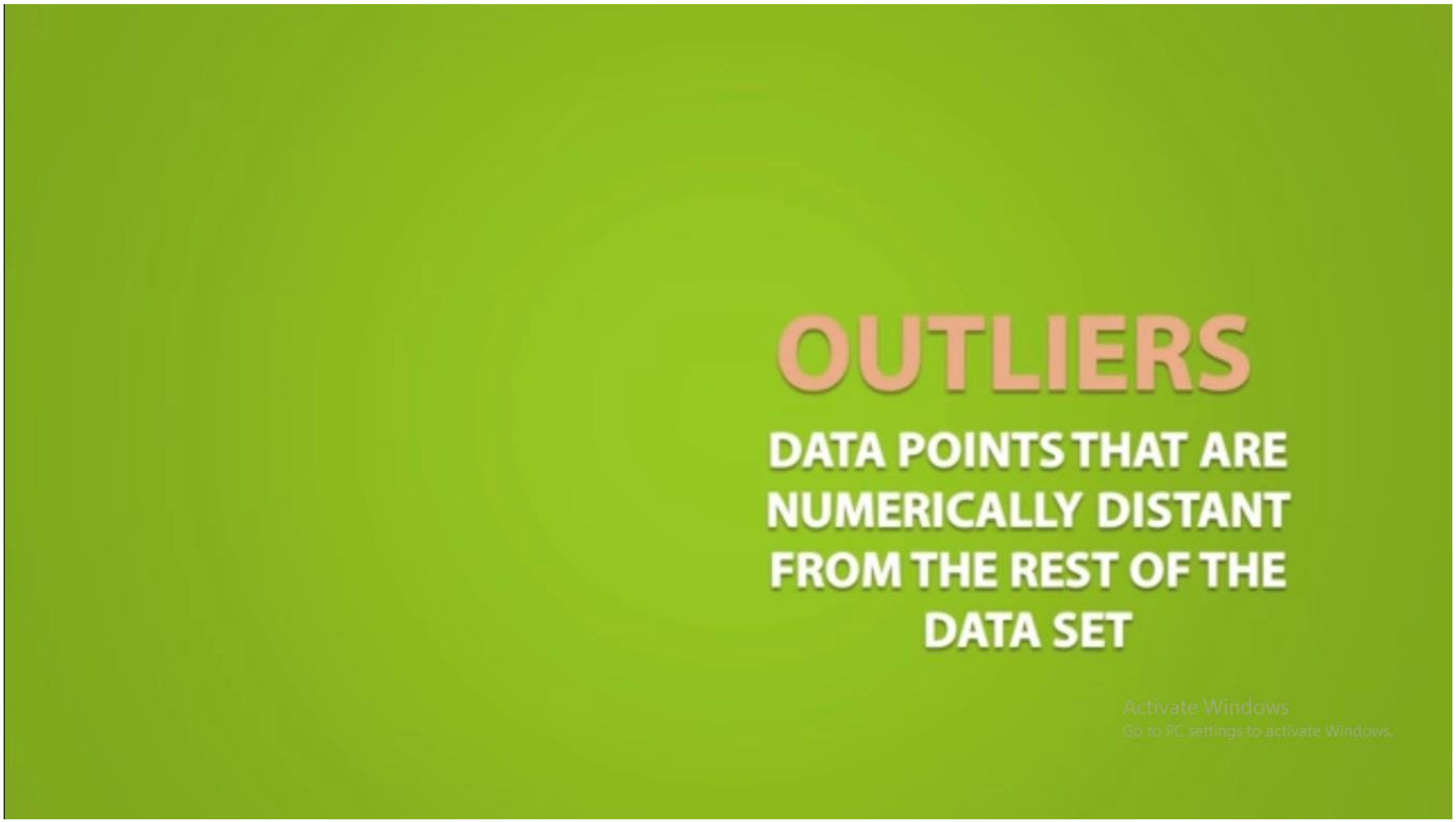
OUTLIER and INFLUENTIAL OBSERVATION

Definition

outlier: A point in a scatterplot that has an extreme x value, an extreme y value, or both. A point can also be an outlier if it is well away from the main trend of points.

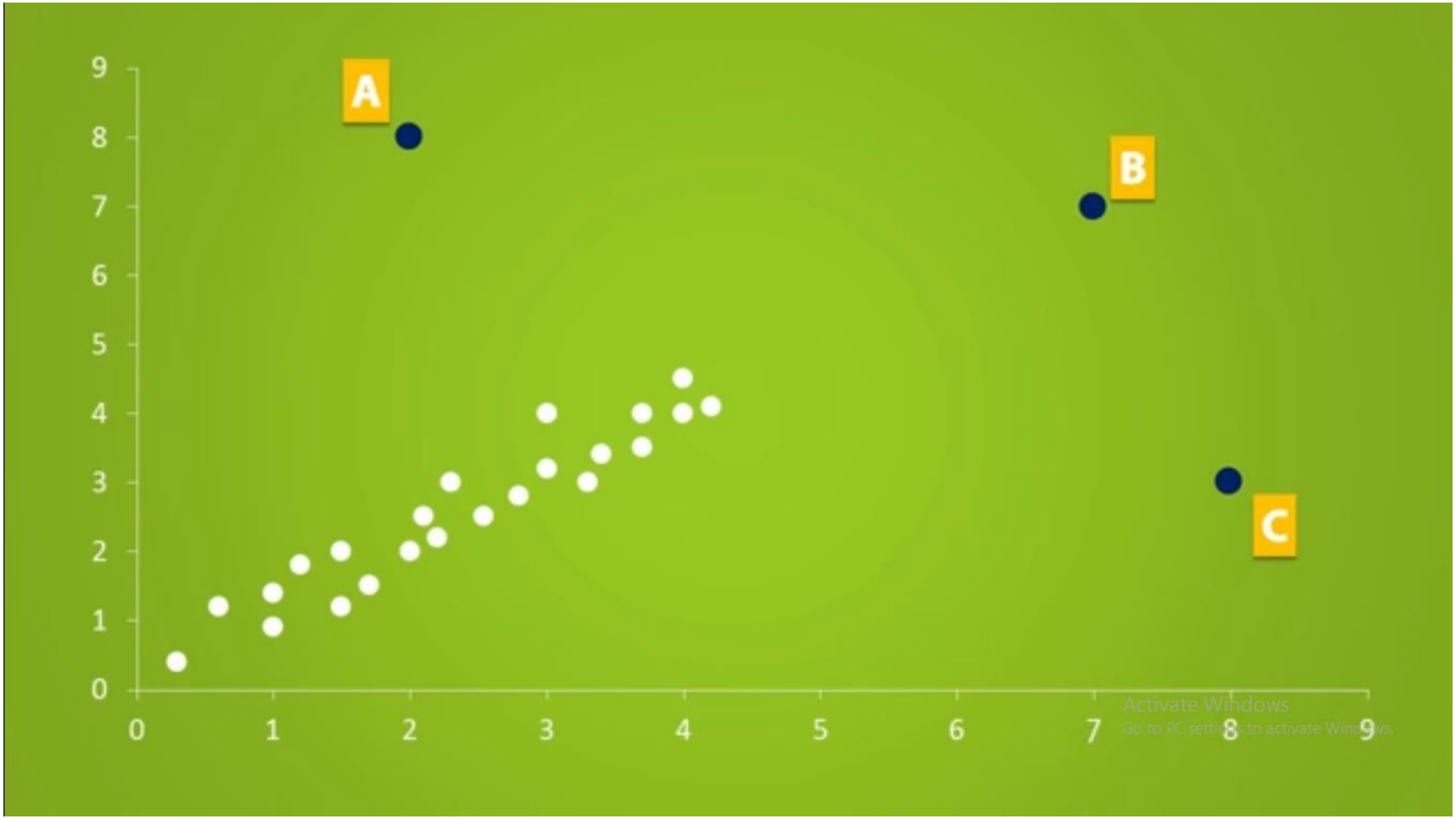
WHAT IS AN OUTLIER?

- Always do scatter plots! The simple act of looking at the shape and pattern of your data can tell you a lot.
- Sometimes a point(s) just don't "look right" and seem out of place.
- For at least one of your variables, a value can appear out of the norm or extreme.
- A point(s) may have an extremely large residual value even though it is within the range of values for that variable.
- A point(s) can influence the regression line pulling it in the direction of the outlier.
- A point(s) fall outside the general pattern of the data.

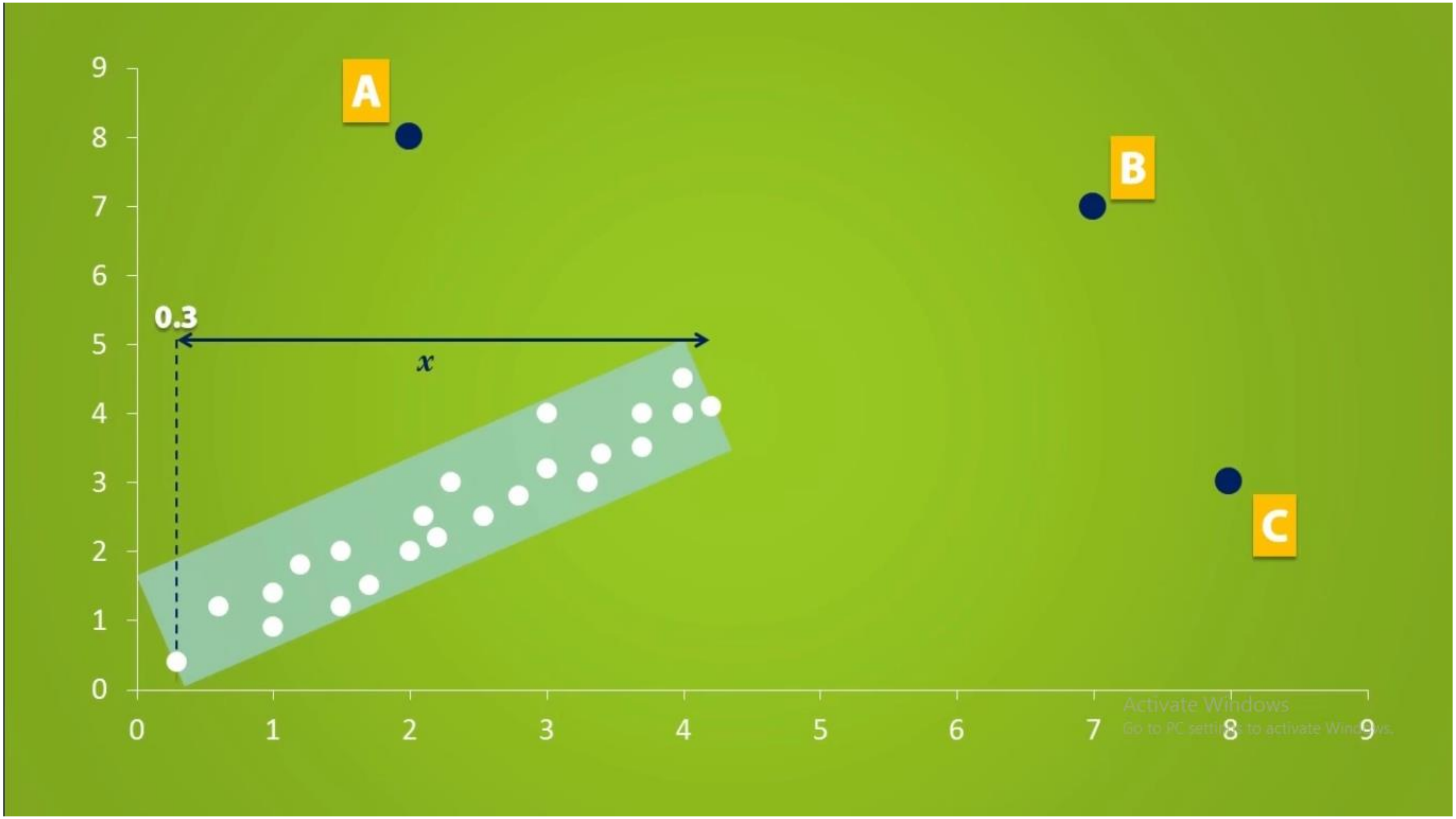

OUTLIERS
DATA POINTS THAT ARE NUMERICALLY DISTANT FROM THE REST OF THE DATA SET

OUTLIERS

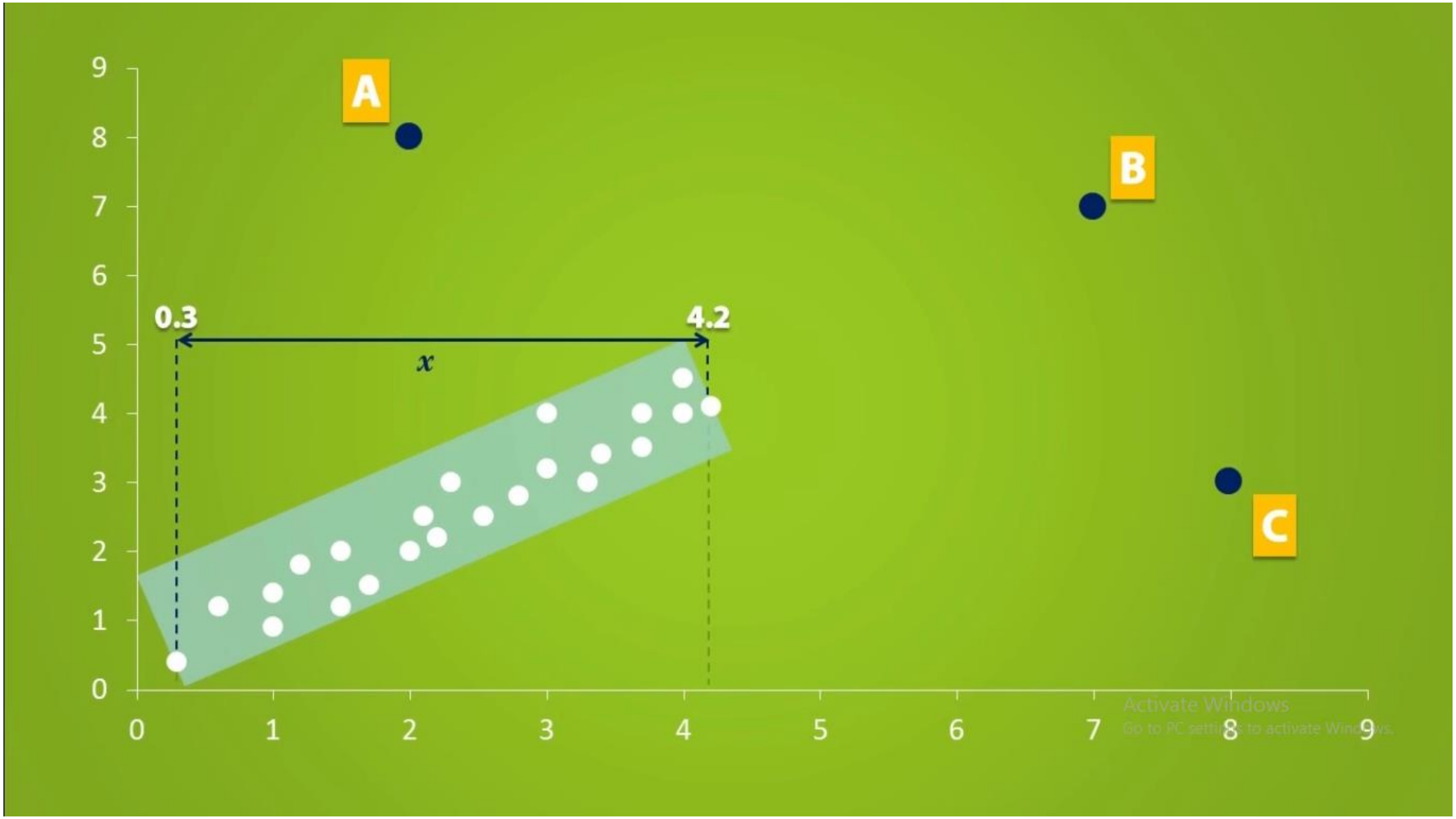
DATA POINTS THAT ARE
NUMERICALLY DISTANT
IN THE "Y" DIRECTION
AND/OR IN THE "X" DIRECTION

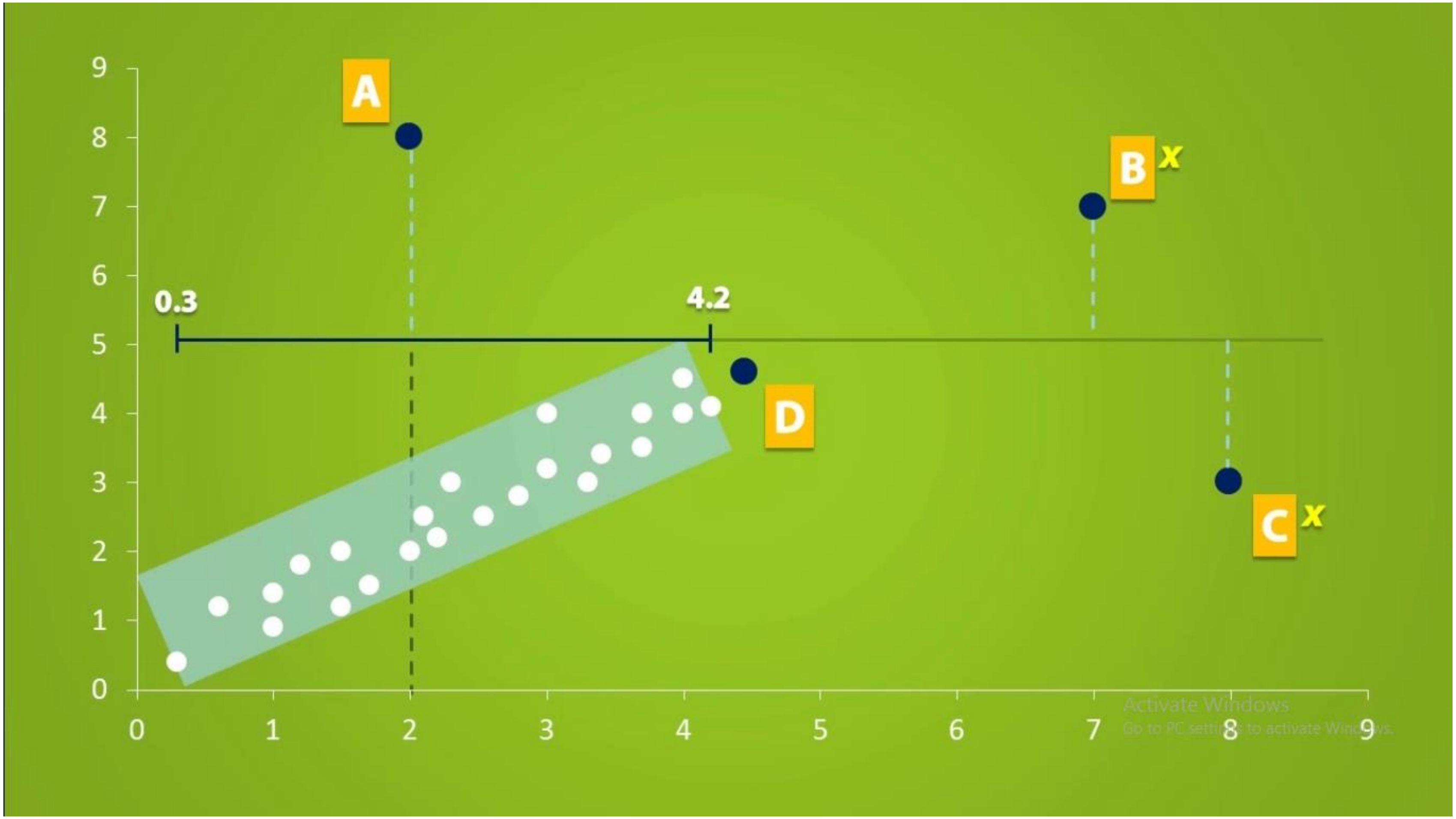


A
B
C
0
1
2
3
4
5
6
7
8
9

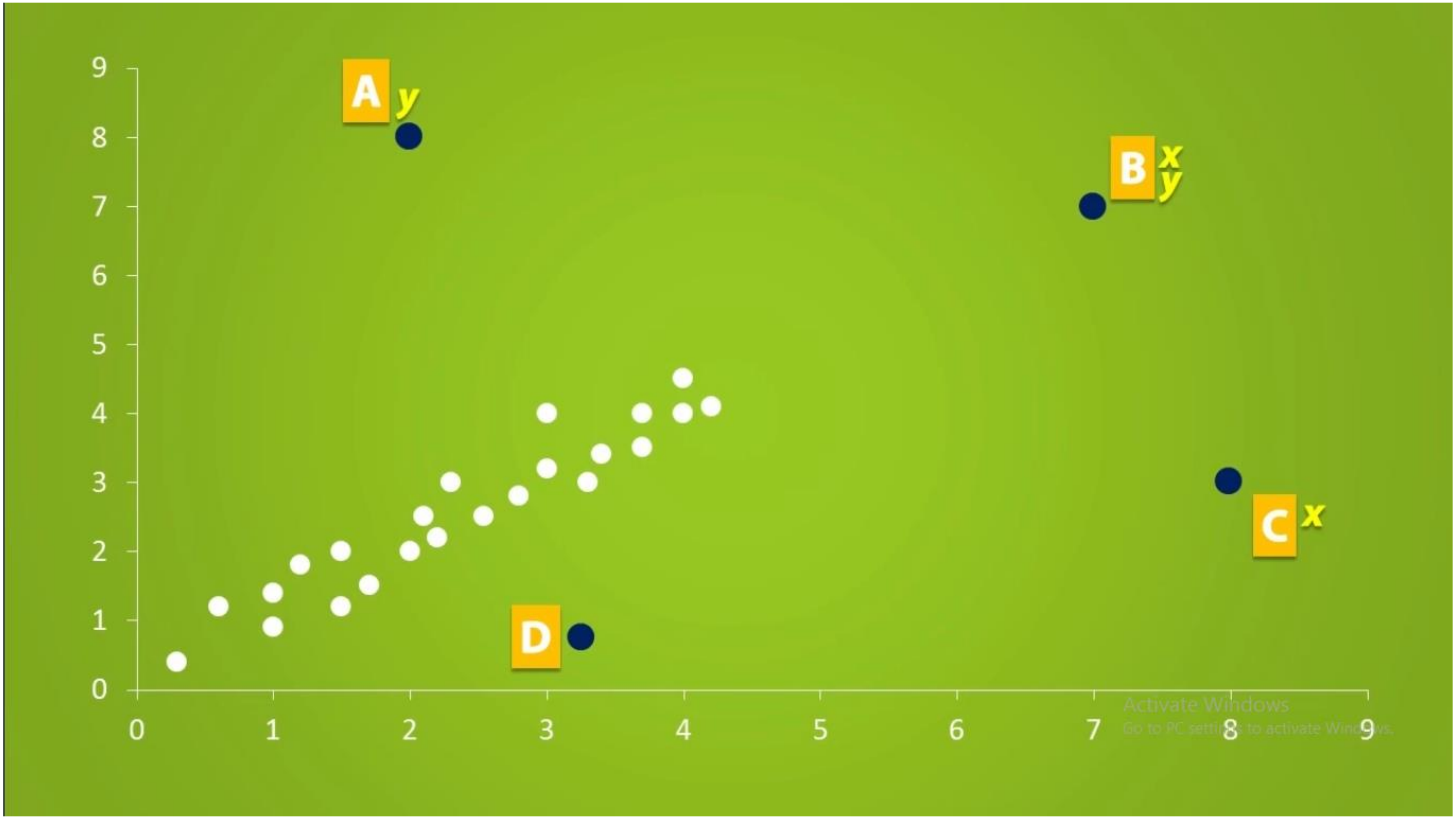


A
B
C
0.3
x
0 1 2 3 4 5 6 7 8 9
Activate Windows

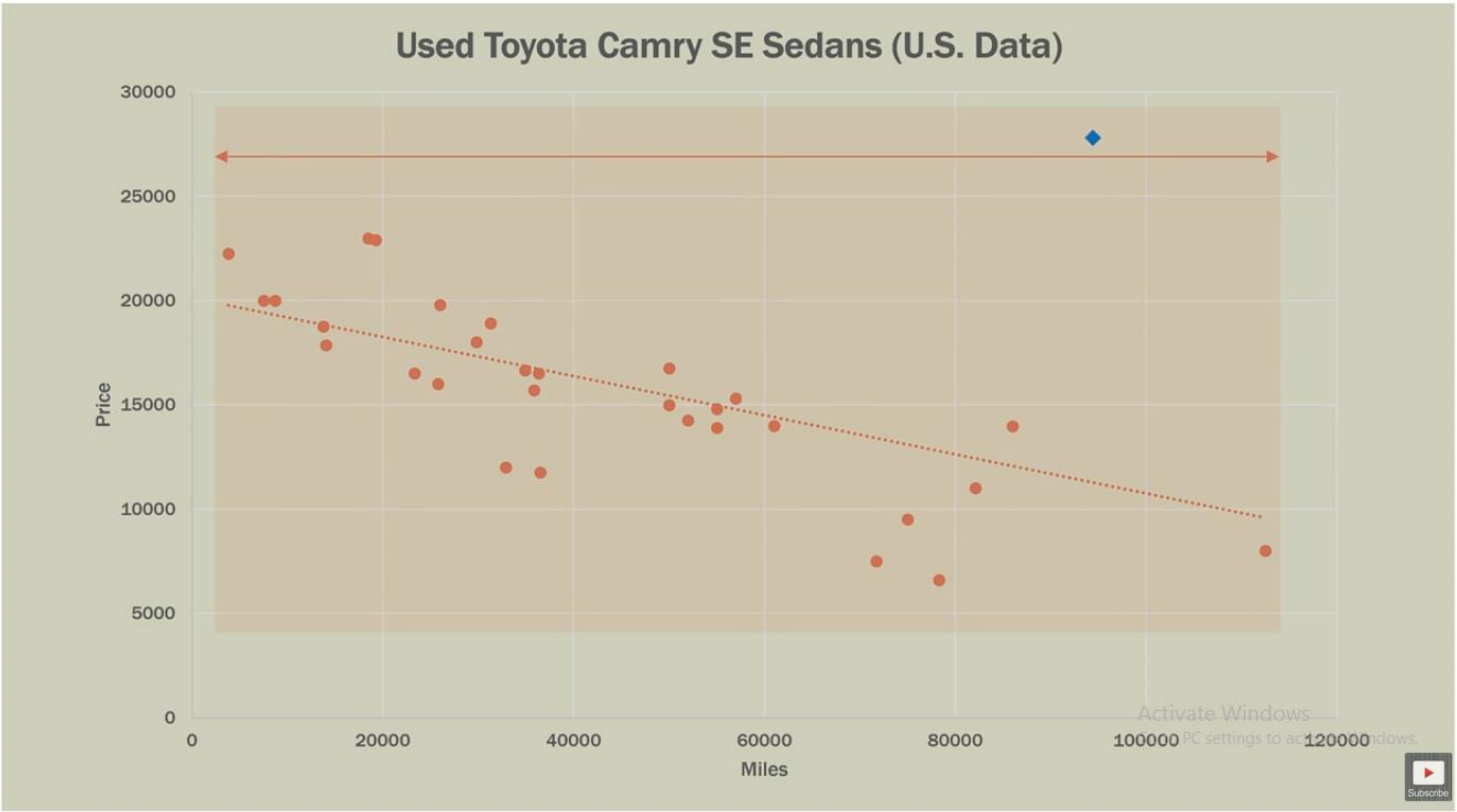

A
B
C
0.3
4.2
x
9
8
7
6
5
4
3
2
1
0
0
1
2
3
4
5
6
7
8
9



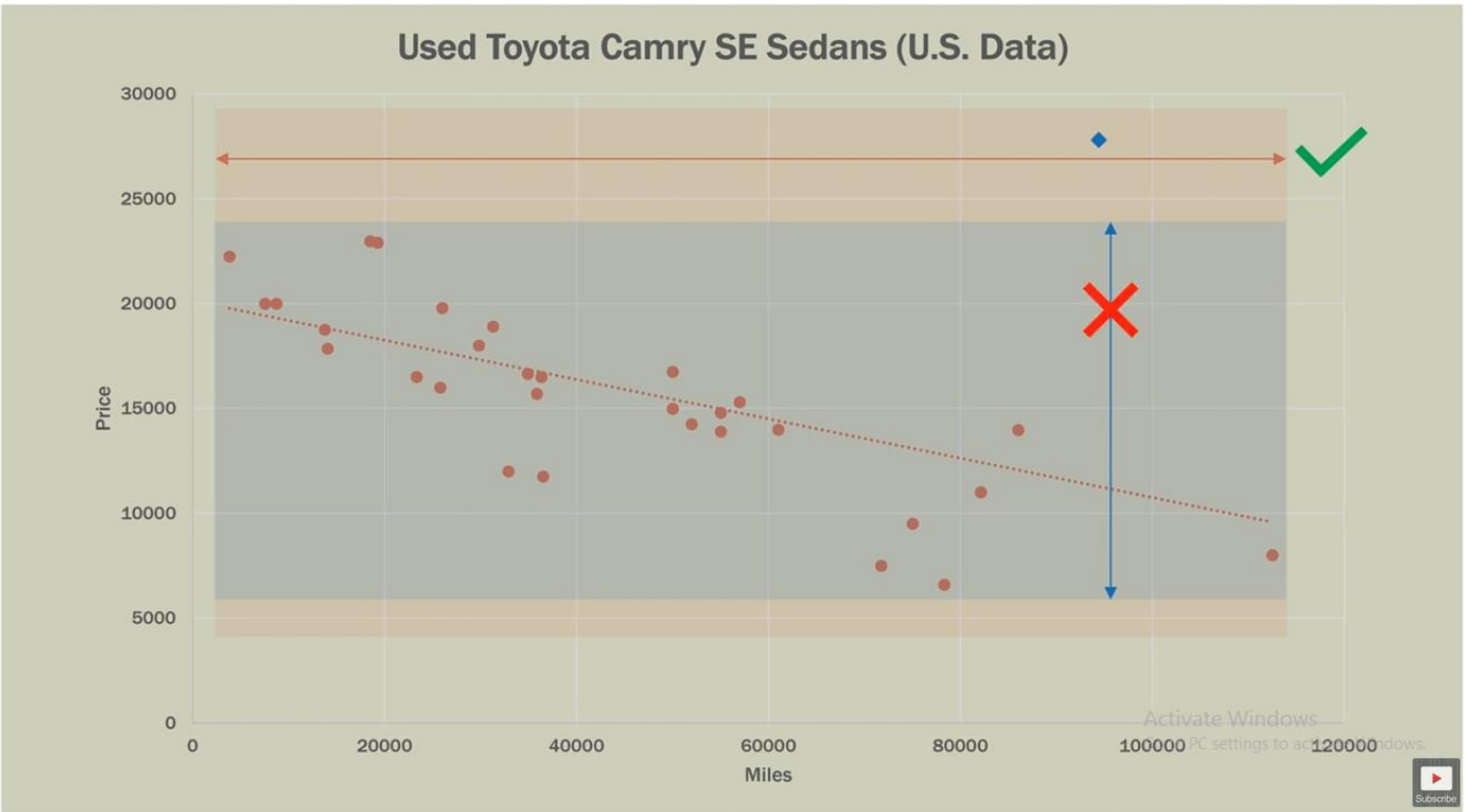
A
B x
C x
D
0.3
4.2
0
1
2
3
4
5
6
7
8
9



A y
B x y
C x
D
0
1
2
3
4
5
6
7
8
9
0
1
2
3
4
5
6
7
8
9

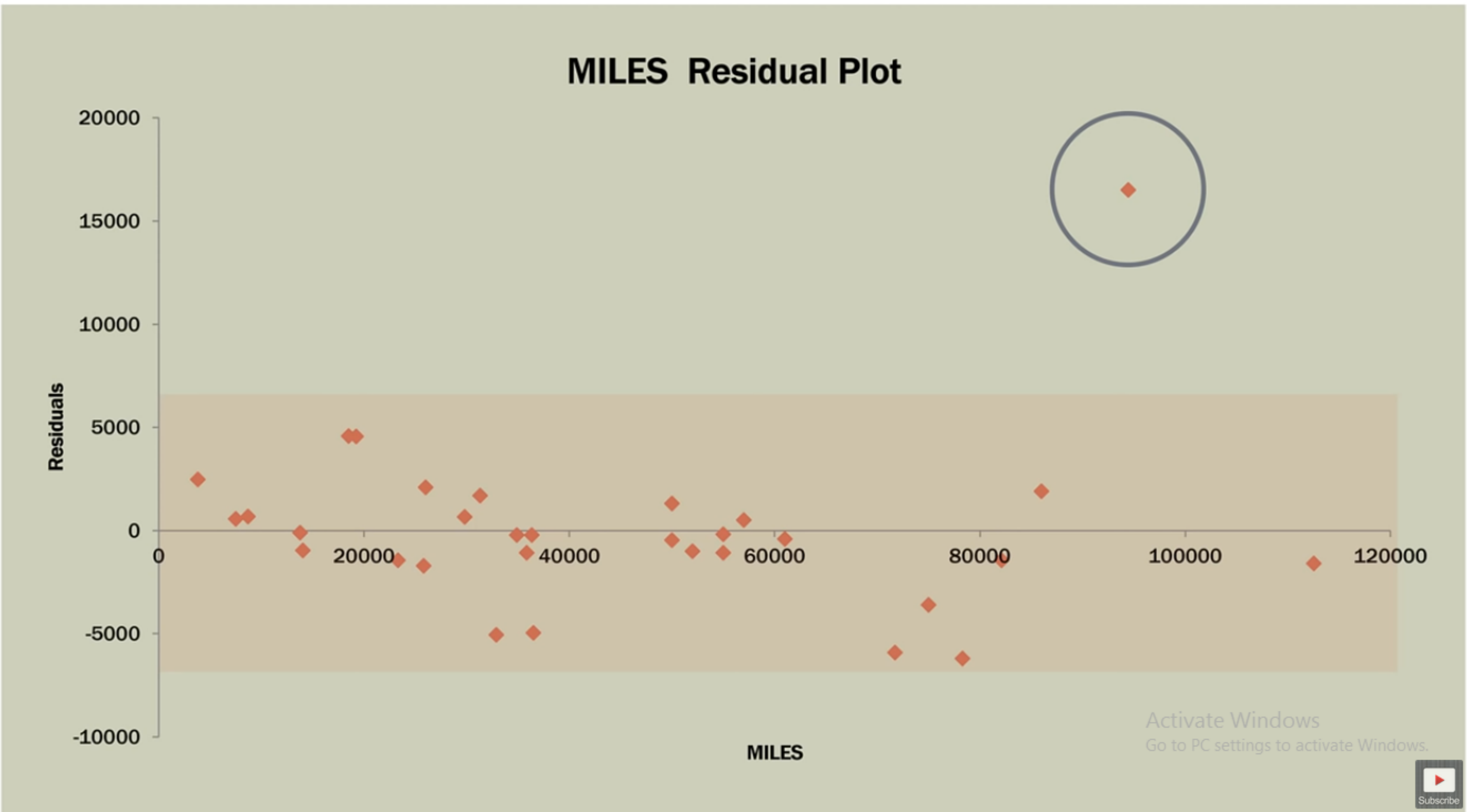


Used Toyota Camry SE Sedans (U.S. Data)
Price
Miles
30000
25000
20000
15000
10000
5000
0
0
20000
40000
60000
80000
100000
120000


Used Toyota Camry SE Sedans (U.S. Data)
Price
Miles
30000
25000
20000
15000
10000
5000
0
0
20000
40000
60000
80000
100000
120000

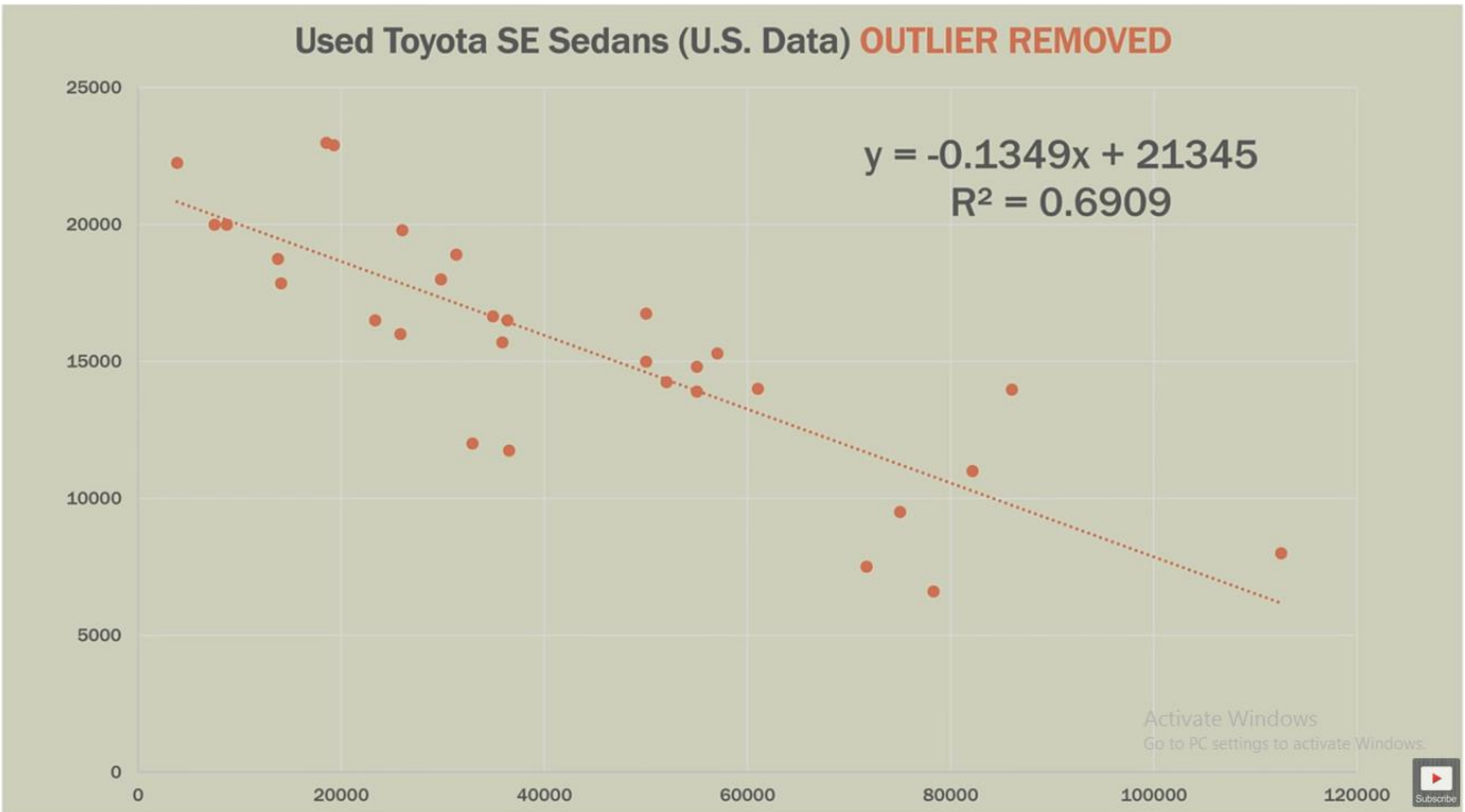
QUESTIONS TO ASK

- Was there an error when recording, entering, or coding the data? A "fat finger" error?
  - Example: A dataset indicates that student received a 40 on a test when it should have been entered as a 70.
- Do the outliers or influential observations suggest a different model should be used? Curvilinear, etc.?
  - Example: Financial yield curves, if modeled using linear regression, can create outliers where the curve bends.



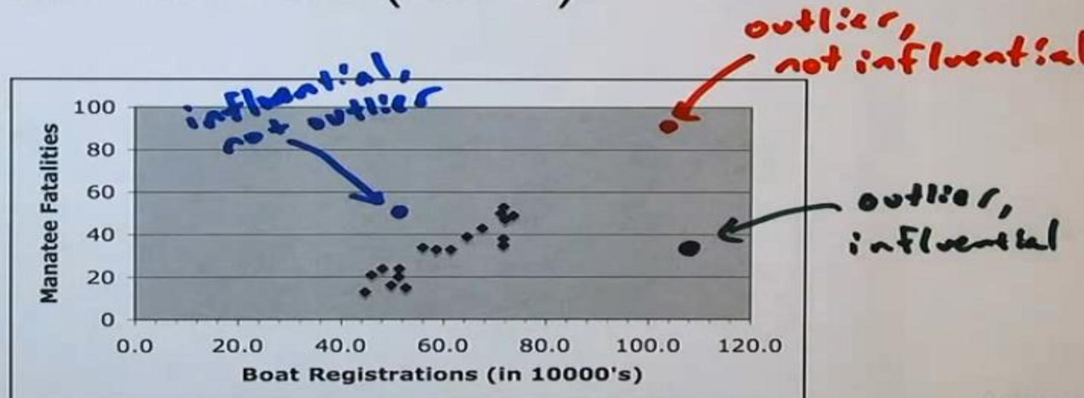
MILES Residual Plot
20000
15000
10000
5000
0
-5000
-10000
Residuals
0
20000
40000
60000
80000
100000
120000
MILES

Observation	Predicted PRICE	Residuals	Standard Residuals	OUTLIER
11	16860.61276	-210.6127571	-0.052330674	FALSE
12	12435.40088	-1435.40088	-0.356652163	FALSE
13	18397.53514	4591.464863	1.140835218	FALSE
14	18328.09249	4570.907514	1.13572736	FALSE
15	12793.29762	-6193.297622	-1.53884049	FALSE
16	17713.97954	-1714.979537	-0.426118703	FALSE
17	9587.877312	-1588.877312	-0.39478625	FALSE
18	15445.51959	-455.5195901	-0.113182352	FALSE
19	16724.72633	-224.7263272	-0.055837454	FALSE
20	16771.58372	-1071.583717	-0.266254993	FALSE
21	13410.97173	-5910.971732	-1.468691349	FALSE
22	15258.09003	-1008.090032	-0.2504788	FALSE
23	14976.94569	-1076.945694	-0.267587276	FALSE
24	13102.65011	-3602.650108	-0.89514572	FALSE
25	17946.11105	-1446.111045	-0.359313304	FALSE
26	11278.39821	16520.60179	4.104852133	TRUE
27	17338.55813	660.4418686	0.16409912	FALSE
28	18842.68034	-92.68033875	-0.023028162	FALSE
29	14976.94569	-176.9456937	-0.043965463	FALSE
30	14414.65702	-414.6570181	-0.103029282	FALSE
31	17048.04232	-5048.042316	-1.254280415	FALSE


Used Toyota SE Sedans (U.S. Data) OUTLIER REMOVED
y = -0.1349x + 21345
R² = 0.6909
25000
20000
15000
10000
5000
0
0
20000
40000
60000
80000
100000
120000

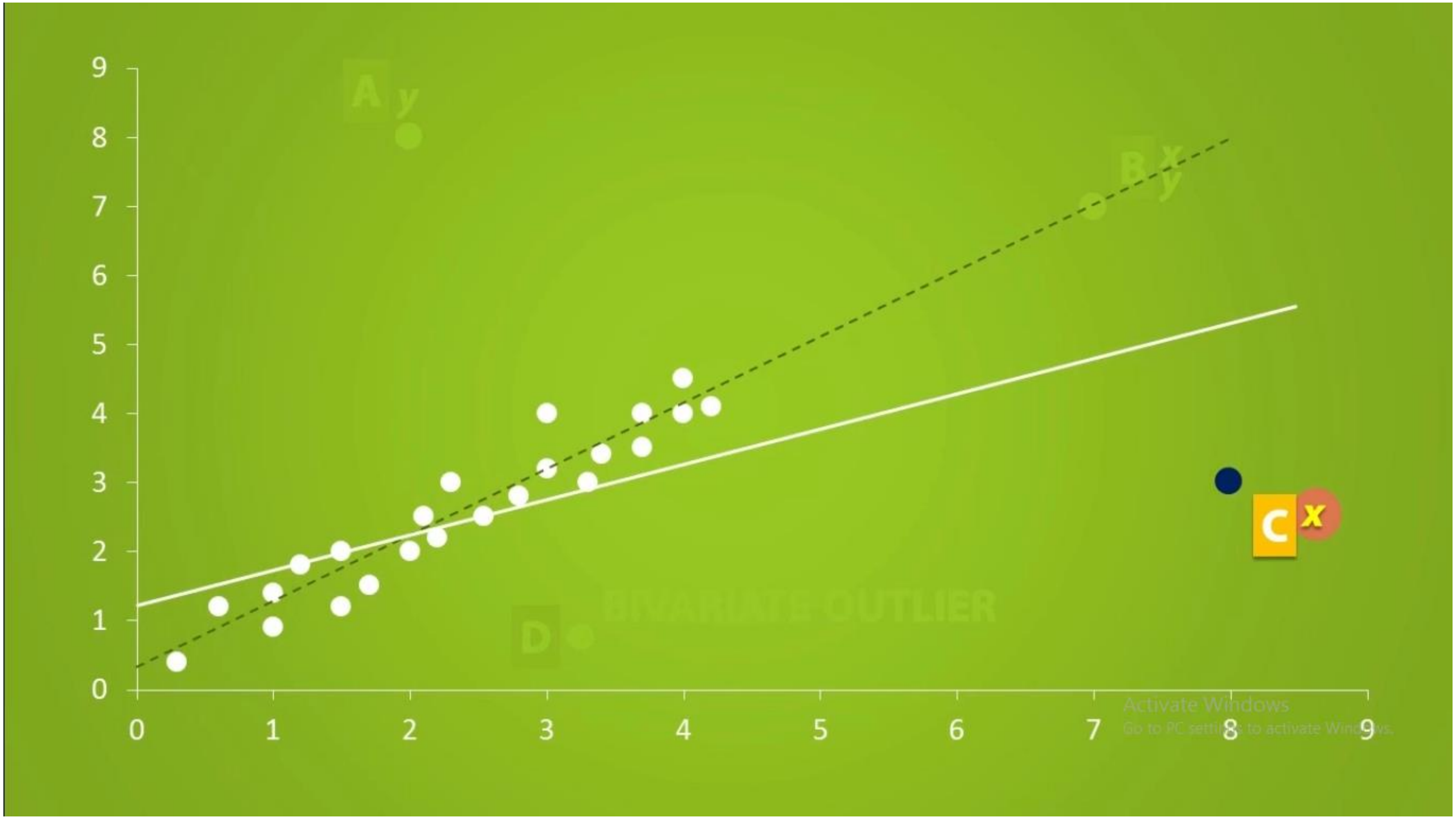
Definition

influential point: An observation that, if removed, significantly changes a statistical measure (like r).



INFLUENTIAL OBSERVATION

- May be an outlier, but not necessarily
- It could be a value of the independent variable that is far outside the rest of the values for the data
- It could be a value of the dependent variable that is far outside the rest of the values for the data
- Or a combination of the two factors above
- Influential observations can dramatically change the regression output and even change the slope of the regression line from positive to negative or negative to positive, model significance, etc.



9
8
7
6
5
4
3
2
1
0
0
1
2
3
4
5
6
7
8
9
C
x

Detection of Influential Observations in Multiple Linear Regression

- An influential observation is the data point that causes a significant change in the regression parameter estimates if it is deleted from the whole dataset.
- Based on this idea we remove one observation at a time to fit the same regression model then calculate the fitted value

- The **prediction sum of squares** (or **PRESS**) is a model validation method used to assess a model's predictive ability that can also be used to compare regression models
- For a data set of size *n*, PRESS is calculated by omitting each observation individually and then the remaining *n* - 1 observations are used to calculate a regression equation which is used to predict the value of the omitted response value.

- To measure the difference between response y_i and $\hat{y}_{i,-i}$
- *we introduce the following PRESS residual.*

The PRESS residual is defined as

$$e_{i,-i} = y_i - \hat{y}_{i,-i}$$

The PRESS statistic is defined as

$$PRESS = \sum_{i=1}^{n} (y_i - \hat{y}_{i,-i})^2 = \sum_{i=1}^{n} (e_{i,-i})^2$$

Therefore, a regression model with a smaller value of the PRESS statistic should be a preferred model.

Graphical Display of Regression Diagnosis

- *Partial Residual Plot*

- Partial residual plots attempt to show the relationship between a given independent variable and the response variable given that other independent variables are also in the model.

we rearrange the X as (x_j, X_{-j}), where x_j is the jth column in the X and X_{-j} is the remaining X after deleting the jth column.

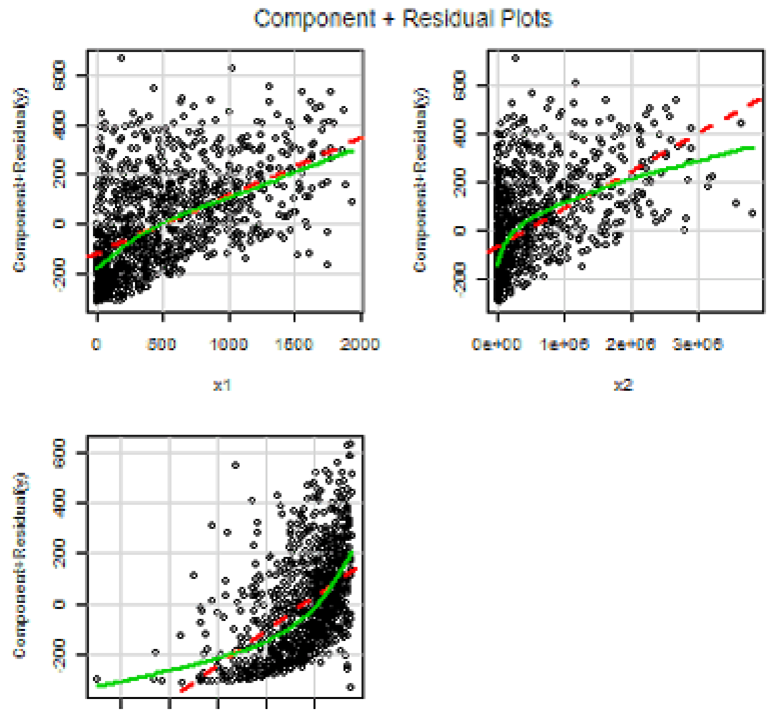
$$\boxed{e_{y|X_{-j}} \text{ against } e_{x_j|X_{-j}}}$$

Component-plus-residual Plot

The component-plus-residual (CPR) plot is one of regression diagnosis plots. It is useful for assessing nonlinearity in independent variables in the model. The CPR plot is the scatter plot of

$$\boxed{e_{y|X} + x_j b_j \text{ against } x_j}$$

- A component residual plot adds a line indicating where the line of best fit lies. A significant difference between the residual line and the component line indicates that the predictor does not have a linear relationship with the dependent variable.



Component + Residual Plots
Component+Residual(y)
x1
x2
0
500
1000
1500
2000
0e+00
1e+06
2e+06
3e+06
-200
0
200
400
600

Augmented Partial Residual Plot

The augmented partial residual (APR) plot is another graphical display of regression diagnosis. The APR plot is the plot of

$$\boxed{e_{y|Xx_j^2} + x_j b_j + x_j^2 b_{jj} \ \text{ against } \ x_j}$$

Test for influential

Various methods have been proposed for measuring influence. Assume an estimated regression $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$, where $\mathbf{y}$ is an n×1 column vector for the response variable, $\mathbf{X}$ is the n×k design matrix of explanatory variables (including a constant), $\mathbf{e}$ is the n×1 residual vector, and $\mathbf{b}$ is a k×1 vector of estimates of some population parameter $\beta \in \mathbb{R}^k$. Also define $\mathbf{H} \equiv \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top$, the projection matrix of $\mathbf{X}$. Then we have the following measures of influence:

1. $\mathrm{DFBETA}_i \equiv \mathbf{b} - \mathbf{b}_{(-i)} = \dfrac{\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{x}_i^\top e_i}{1 - h_{i\cdot}}$, where $\mathbf{b}_{(-i)}$ denotes the coefficients estimated with the i-th row $\mathbf{x}_i$ of $\mathbf{X}$ deleted, $h_{i\cdot} = \mathbf{x}_i\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{x}_i^\top$ denotes the i-th row of $\mathbf{H}$.

DFBETA measures the difference in each parameter estimate with and without the influential point.

Cook's D: The Cook's D is the distance between the least squares estimates of regression coefficients with $\boldsymbol{x}_i$ included and excluded. The composite measurement Cook's D is defined as

$$D_i = \frac{(\boldsymbol{b} - \boldsymbol{b}_{-i})^{'}(\boldsymbol{X}^{'}\boldsymbol{X})(\boldsymbol{b} - \boldsymbol{b}_{-i})}{ps^2},$$

Cook’ D

- it is used to identify influential data points. It depends on both the residual and leverage i.e it takes it account both the **x** value and **y** value of the observation.
- **Steps to compute Cook’s distance:**
- delete observations one at a time.
- refit the regression model on remaining (n−1)
- observations examine how much all of the fitted values change when the ith observation is deleted.

Model Selection Criteria – All Possible Regressions

$P-1$ predictors $\Rightarrow$ 2^{P-1} potential models (each variable can be in or out of model)

R_p^2 or SSE_p criterion (Goal: find p so that $\max\left(R_p^2\right)$ or $\min\left(SSE_p\right)$ "flattens out"):

$$R_p^2 = \frac{SSR_p}{SSTO} = 1 - \frac{SSE_p}{SSTO} \qquad p = \text{\# of parameters in current model}$$

$R_{a,p}^2$ or MSE_p criterion (Goal: find model that maximizes (or close to) $R_{a,p}^2$ and minimizes MSE_p):

$$R_{a,p}^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE_p}{SSTO} = 1 - \frac{\left(SSE_p/(n-p)\right)}{\left(SSTO/(n-1)\right)} = 1 - \frac{MSE_p}{\left(SSTO/(n-1)\right)}$$

Mallow's C_p criterion (Goal: find model with smallest p so that $C_p \le p$):

$$C_p = \frac{SSE_p}{MSE\left(X_1,\ldots,X_{P-1}\right)} - \left(n-2p\right)$$

AIC_p and SBC_p criteria (Goal: choose model that minimize these values):

$$AIC_p = n\ln\left(SSE_p\right) - n\ln(n) + 2p \qquad SBC_p = n\ln\left(SSE_p\right) - n\ln(n) + \left[\ln(n)\right]p$$

$PRESS_p$ criterion (Goal: Small values):

$$PRESS_p = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_{i(i)}\right)^2 \qquad \hat{Y}_{i(i)} \equiv \text{fitted value for } i^{th} \text{ case when it was not used in fitting model}$$

Regression Model Building

- Setting: Possibly a large set of predictor variables (including interactions).
- Goal: Fit a parsimonious model that explains variation in Y with a small set of predictors
- Automated Procedures and all possible regressions:
  - Backward Elimination (Top down approach)
  - Forward Selection (Bottom up approach)
  - Stepwise Regression (Combines Forward/Backward)

Backward Elimination Traditional Approach

- Select a significance level to stay in the model (e.g. SLS=0.20, generally .05 is too low, causing too many variables to be removed)
- Fit the full model with all possible predictors
- Consider the predictor with lowest *t*-statistic (highest *P*-value).
  - If $P >$ SLS, remove the predictor and fit model without this variable (must re-fit model here because partial regression coefficients change)
  - If $P \leq$ SLS, stop and keep current model
- Continue until all predictors have *P*-values below SLS
- Note: R uses model based criteria: AIC, SBC instead

Forward Selection – Traditional Approach

- Choose a significance level to enter the model (e.g. SLE=0.20, generally .05 is too low, causing too few variables to be entered)
- Fit all simple regression models.
- Consider the predictor with the highest *t*-statistic (lowest *P*-value)
  – If $P \leq$ SLE, keep this variable and fit all two variable models that include this predictor
  – If $P >$ SLE, stop and keep previous model
- Continue until no new predictors have $P \leq$ SLE
- Note: R uses model based criteria: AIC, SBC instead

Stepwise Regression – Traditional Approach

- Select SLS and SLE (SLE<SLS)
- Starts like Forward Selection (Bottom up process)
- New variables must have $P \leq$ SLE to enter
- Re-tests all "old variables" that have already been entered, must have $P \leq$ SLS to stay in model
- Continues until no new variables can be entered and no old variables need to be removed
- Note: R uses model based criteria: AIC, SBC instead

Model Validation

- When we have a lot of data, we would like to see how well a model fit on one set of data (training sample) compares to one fit on a new set of data (validation sample), and how the training model fits the new data.
- Want data sets to be similar wrt levels of the predictors
- Training set should have at least 6-10 times as many observations than potential predictors
- Models should give similar model fits based on SSE_p, $PRESS_p$, C_p, and MSE_p and regression coefficients
- Mean Square Prediction Error when training model is applied to validation sample:

$$MSPR = \frac{\sum_{i=1}^{n^*}\left(Y_i - \hat{Y}_i\right)^2}{n^*} \qquad \hat{Y}_i = b_0^T + b_1^T X_{i1}^V + \ldots + b_{p-1}^T X_{i,p-1}^V$$