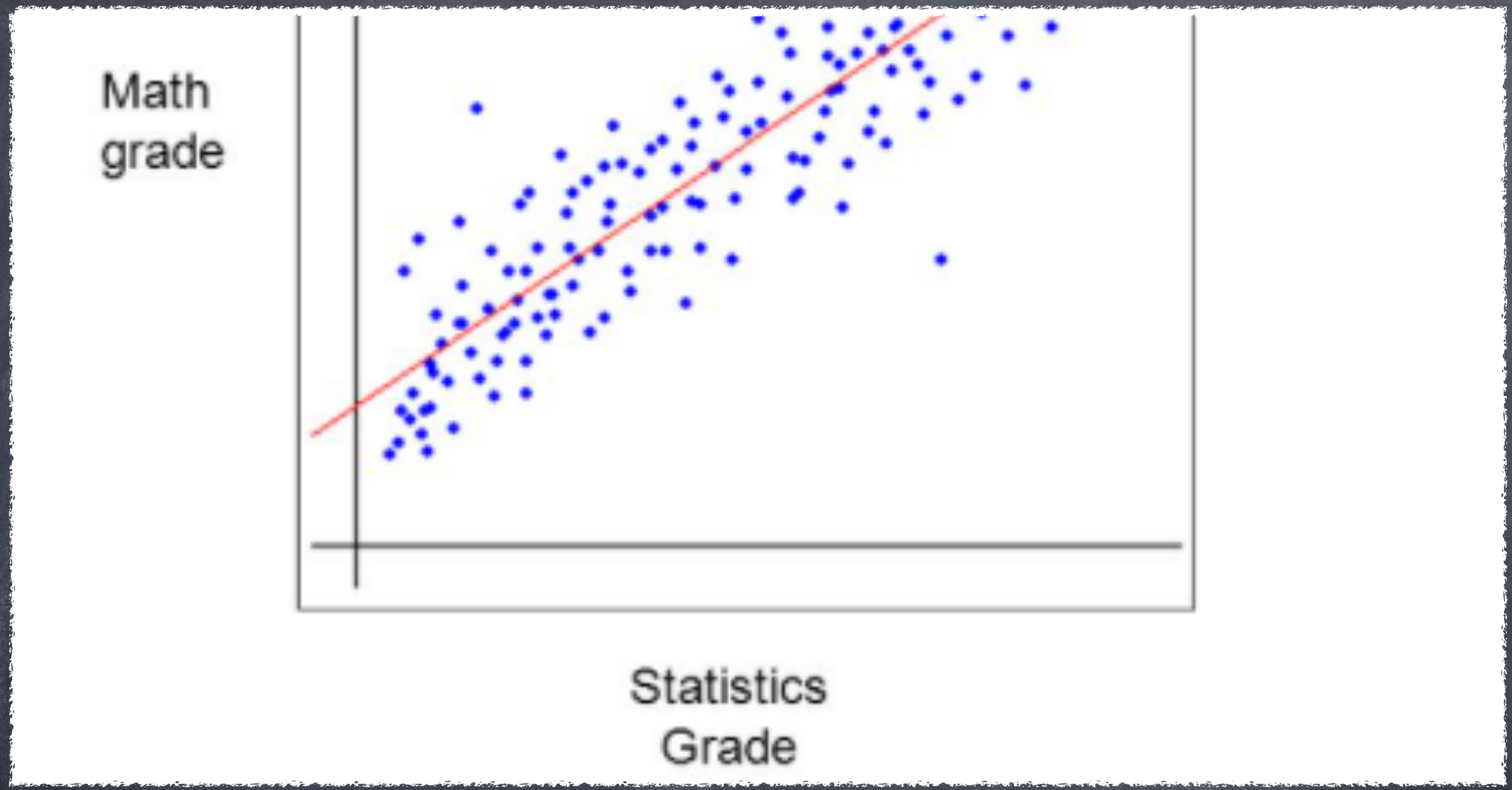
KBTU | KAZAKH-BRITISH TECHNICAL UNIVERSITY

DATA ANALYSIS

Christian Capone
Kazakh-British Technical University
Almaty, Kazakhstan



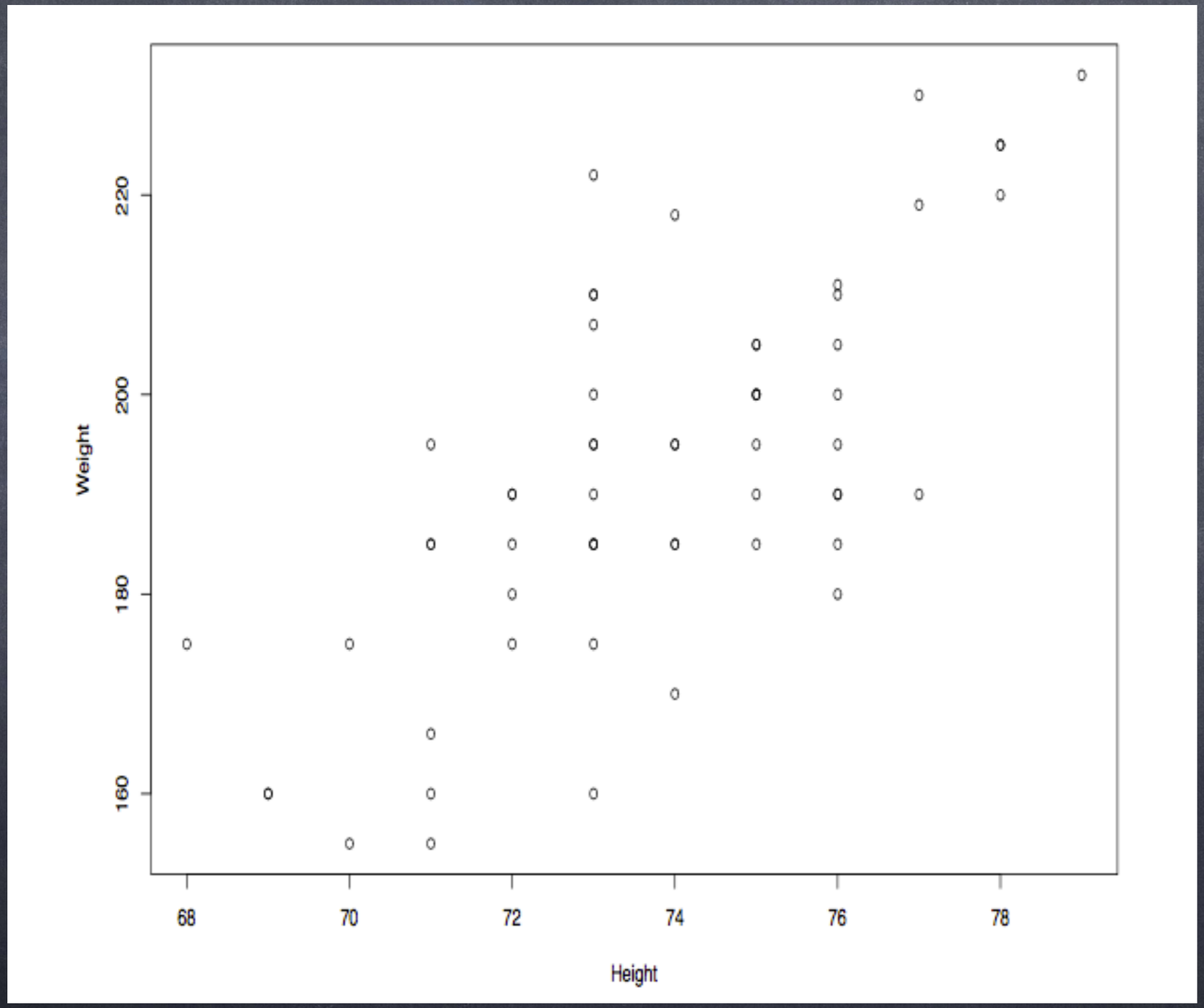
Relationships between Variables

the starting point of inferring statistics

Linear Models

- X is the grade of a student on first attestation and Y is his/her grade on second attestation
- X is the height of a person and Y is his/her weight
- We are interested in the relationship between X and Y and we might be further interested in predicting one variable in terms of the other

Linear Models



Linear Model

- $Y = a + bX + e$
  - a = intercept
  - b = slope
  - e = random error
- How can we set this model, i.e. obtain a good estimation of "a" and "b"?
  - Eyeball Fit: we take two points and we calculate the function the line passing by them, taking care to divide the quadrant in a way that same quantity of measures are above and below the line
  - Least Squares Fit (LS): minimise the averaged squared deviation from the chosen line; it is not robust
  - Wilcoxon Fit: it is less sensitive against outliers at least in Y direction; without outliers it is very close to LS

Least Square Fit

$$Y' = b_{yx} X + a_{yx}$$

$$b_{yx} = r_{xy} \frac{\sigma_y}{\sigma_x} = \frac{N\sum XY - (\sum X \sum Y)}{N\sum X^2 - (\sum X)^2}$$

$$a_{yx} = \overline{Y} - b_{yx} \overline{X}$$

Exercise

1. *Let X be the length (cm) of a laboratory mouse and let Y be its weight (gm). Consider the data for X and Y given below. Obtain a scatterplot of the data and comment on the plot.*

X	Y
16	32
15	26
20	40
13	27
15	30
17	38
16	34
21	43
22	64
23	45
24	46
18	39

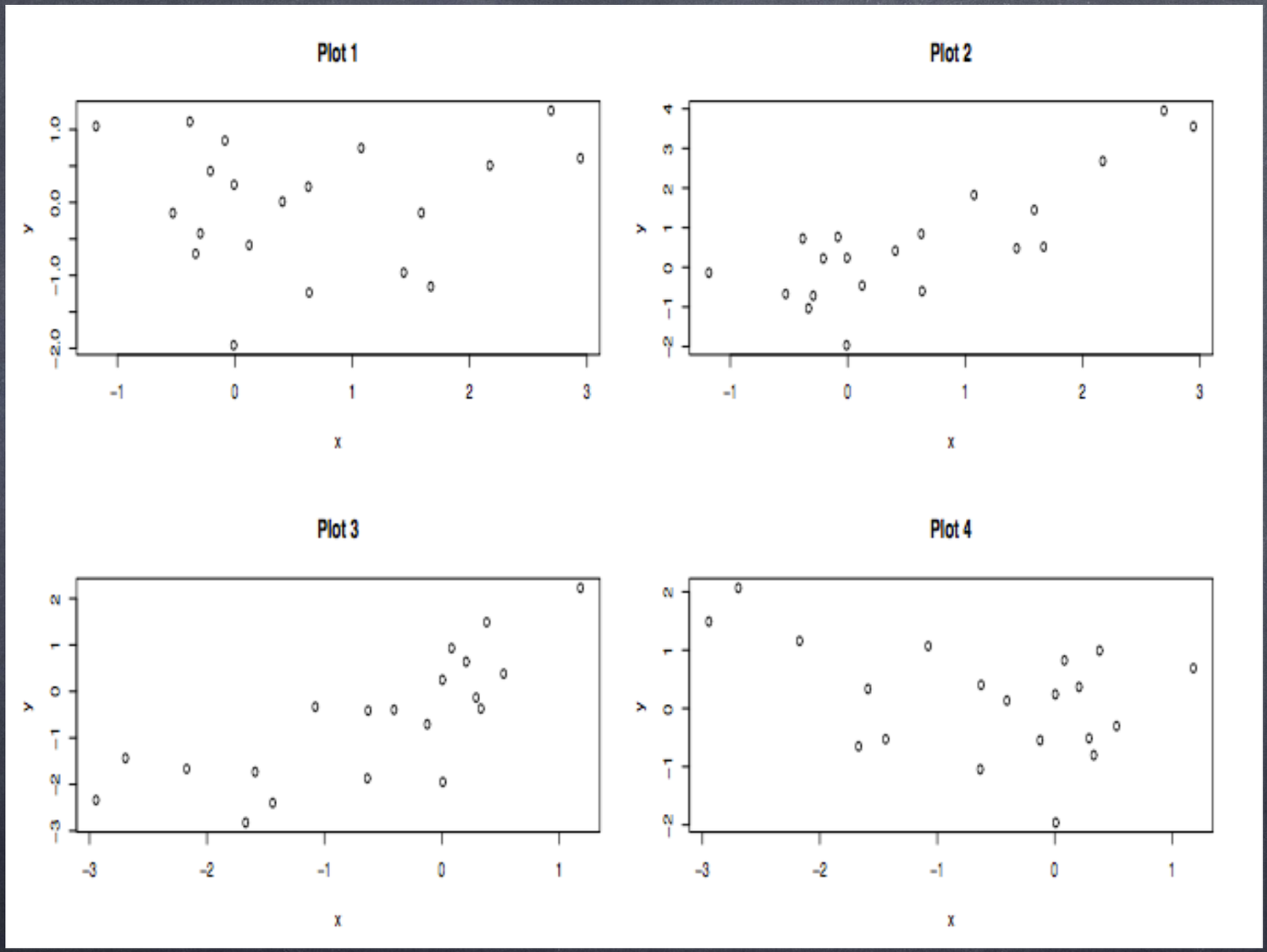
2. *For the data set in Problem #1, eyeball a linear fit obtaining an estimate of the slope and the intercept.*
   (a) *Plot your fit.*
   (b) *Use your plotted fit, to predict the weight of a mouse that is 20 cm long.*
   (c) *Use your predicition equation to predict the weight of a mouse that is 25 cm long.*
   (d) *What does the estimate of slope mean in terms of the problem?*
   (e) *What does the estimate of intercept mean in terms of the problem?*
3. *Use the formulas given in class to determine the LS fit for the data given in Problem #1. (ANS: LS slope is: 2.405).*
4. *Plot your fit.*
5. *Compare the LS fit with your eyeball fit? Which is a better fit? Why?*

Residual Analysis

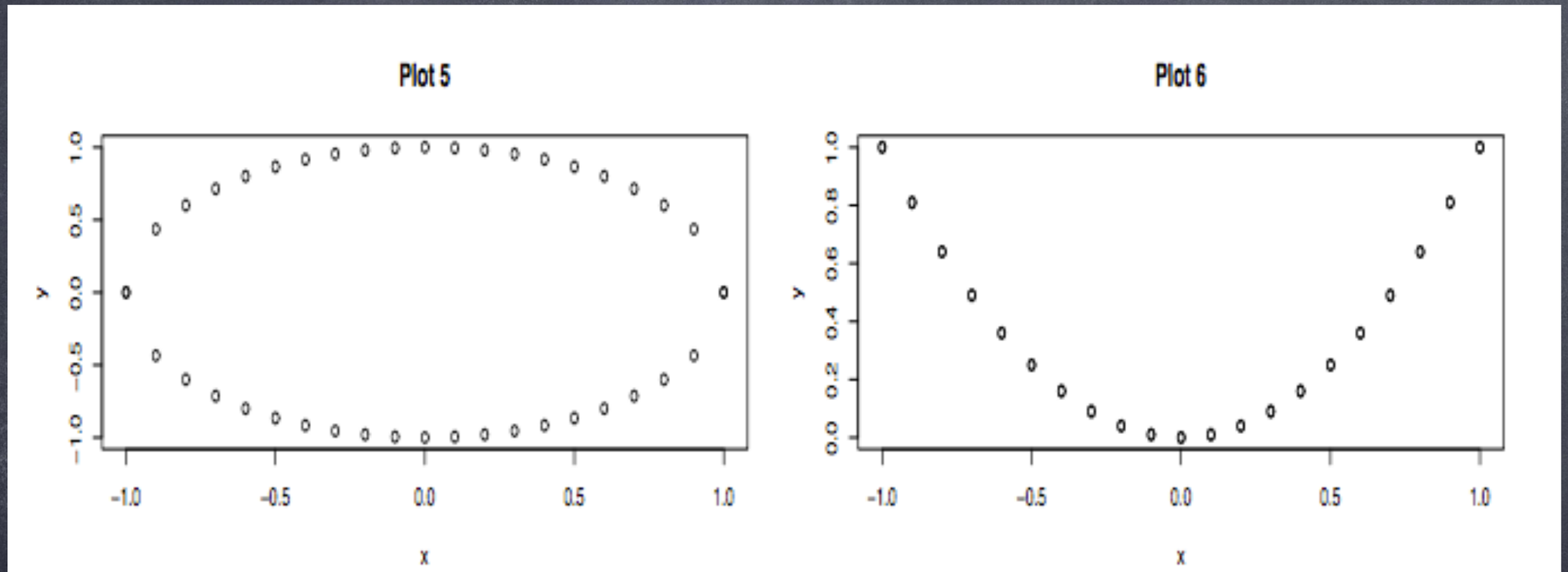
- Picking a model for a problem is a major undertaking: if the model fits well then it can be used to increase understanding of the problem and/or for prediction
- When a specific model is good? When there is no error? But we considered all dataset, till now, where there is an error
- We define a good model when the random error "e" is not a function of "X". In order to check that we can perform a "residual plot": random scatter is good

$$\hat{e} = Y - (\hat{a} + \hat{b}X)$$ versus $\hat{Y}$

Relationships

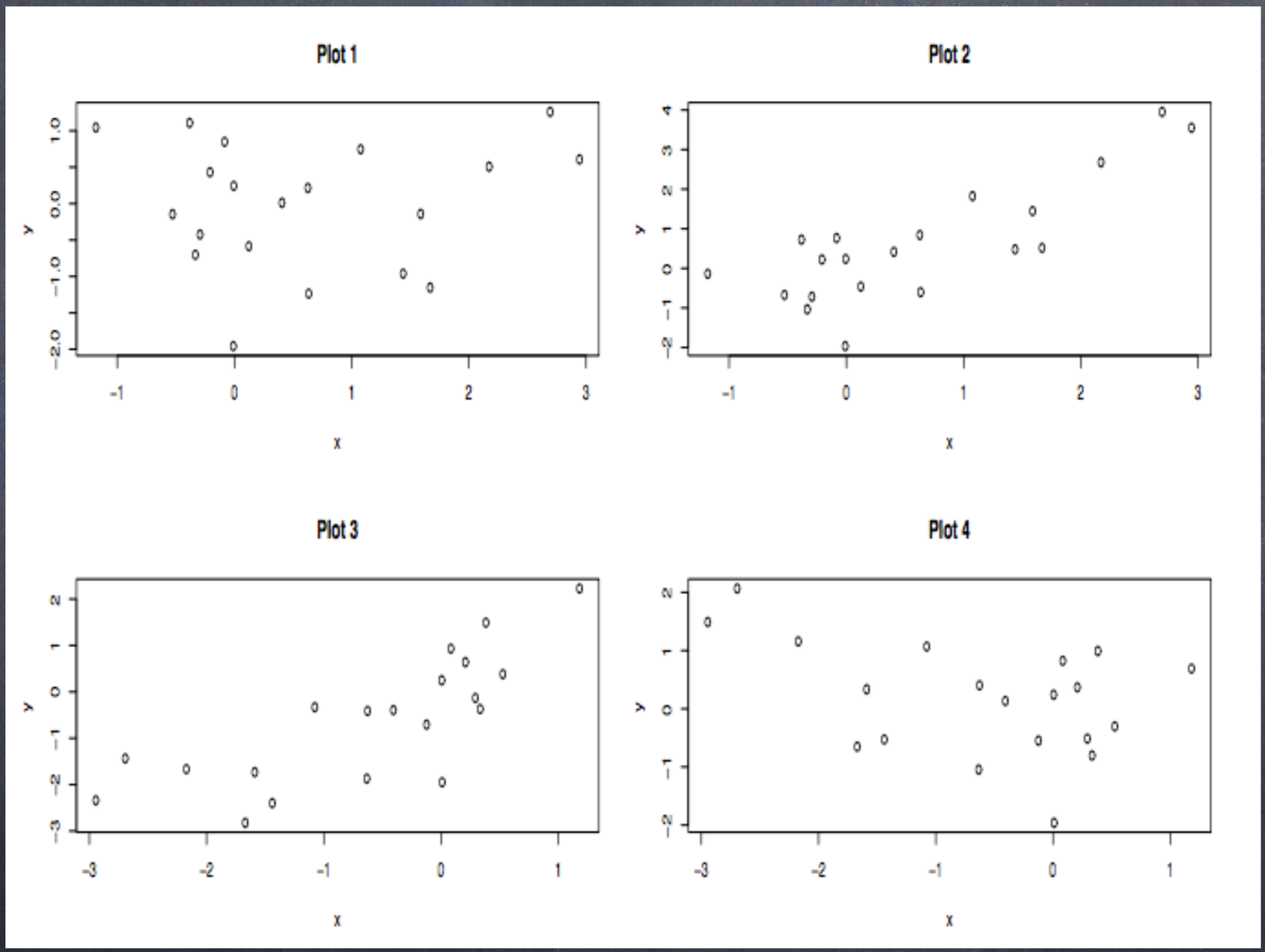


Relationships



Correlations with R^2

.086



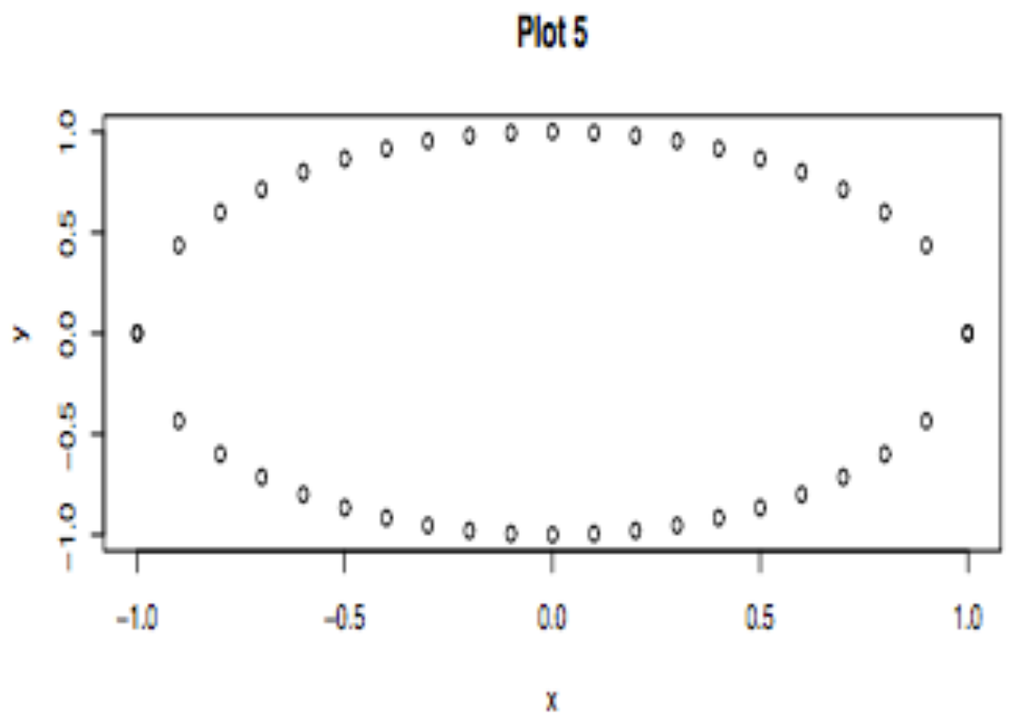
.810

.770

.430

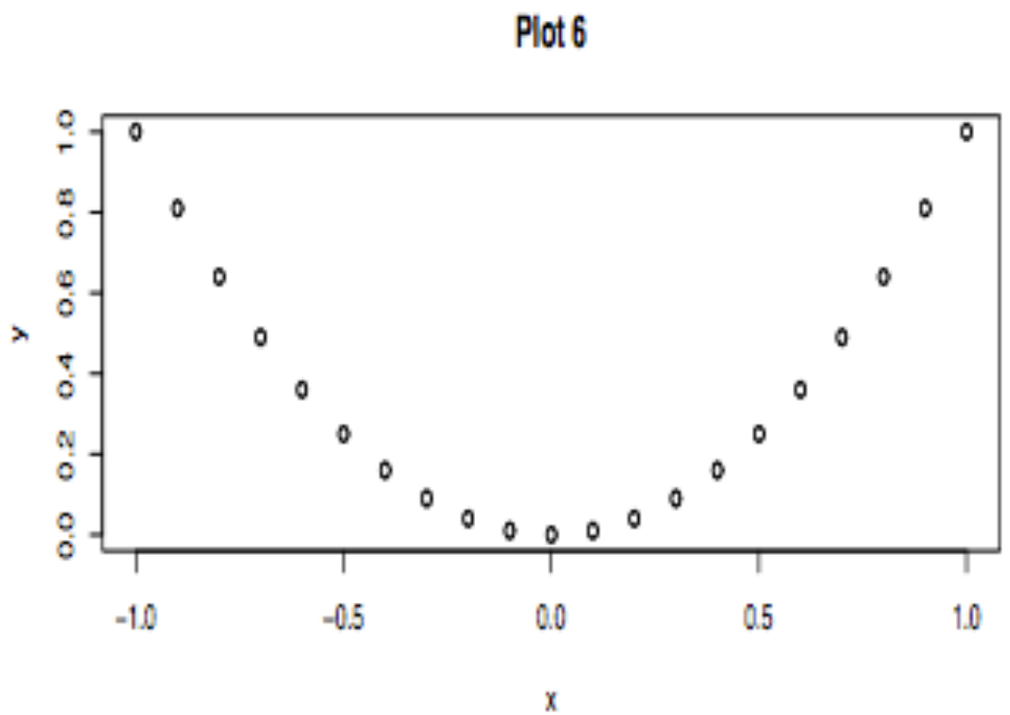
Correlations with R^2

Circular



0.00

Quadratic



0.00

How to find a measure of linear relationships?

1. on X and Y axes locate the mean
2. Count the number of points in quadrants 1 and 3 and subtract the number of points in quadrants 2 and 4
3. an high positive number indicate a positive relationship, an high negative number a negative relationship

Sample Covariance and Correlation Coefficient

$$s_{XY} = \frac{\text{Sum}((X - \bar{X})(Y - \bar{Y}))}{n}$$

Covariance

Correlation Coefficient

$$r = \frac{s_{XY}}{\sqrt{\frac{\text{Sum}(X-\bar{X})^2 \text{Sum}(Y-\bar{Y})^2}{n^2}}}$$

- Covariance is not robust against unit measures
- Correlation Coefficient (r) is robust (value range: -1 to +1) against unit measures
- Value of "r" close to "0" indicate little or no linear relationship
- r^2 (often R^2) is a measure of LS fit and it is call COEFFICIENT OF DETERMINATION

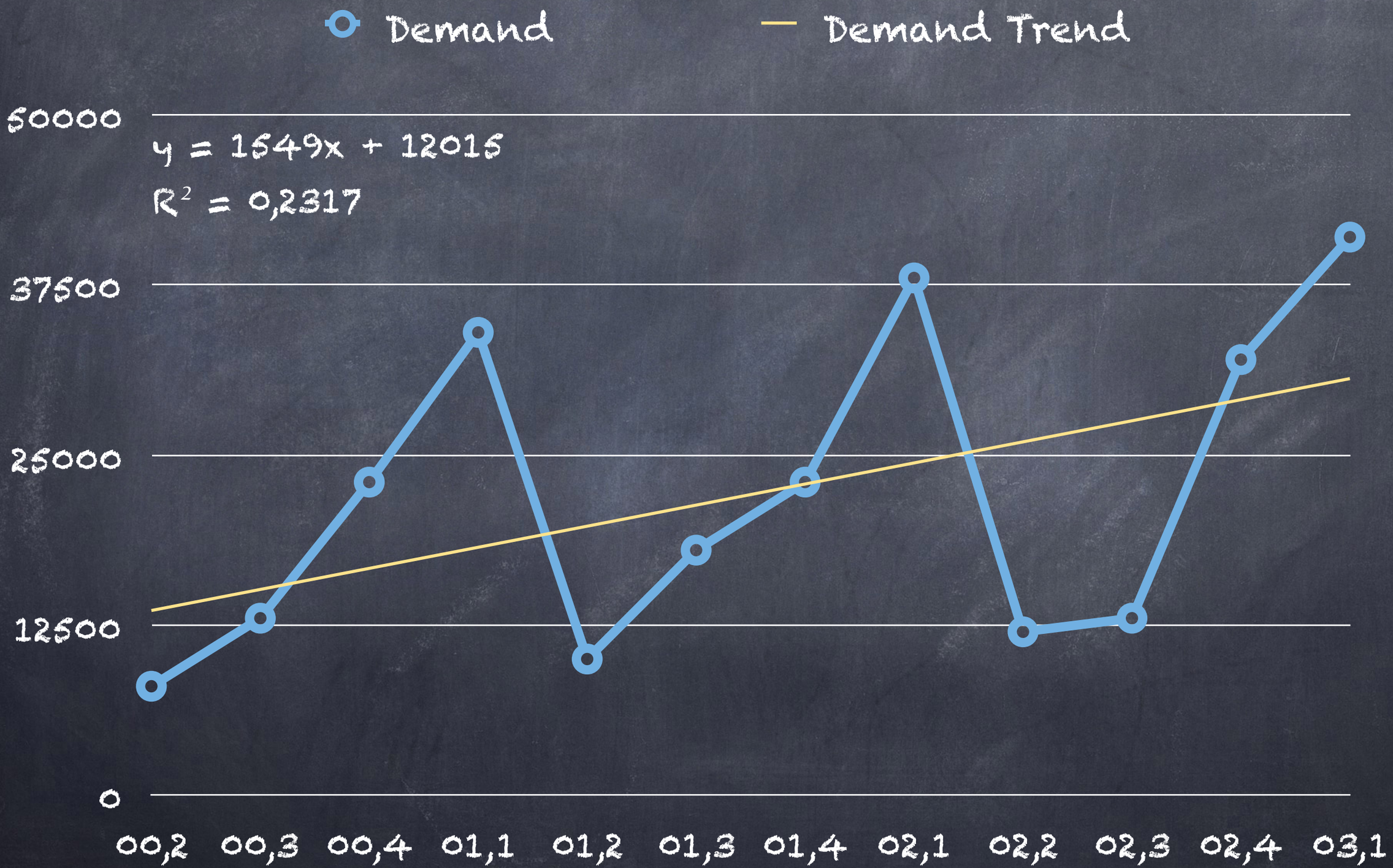
Demand Forecast

Static Method

Quarterly Demand Example

Year, Qtr	Period t	Demand Dt
00,2	1	8.000
00,3	2	13.000
00,4	3	23.000
01,1	4	34.000
01,2	5	10.000
01,3	6	18.000
01,4	7	23.000
02,1	8	38.000
02,2	9	12.000
02,3	10	13.000
02,4	11	32.000
03,1	12	41.000

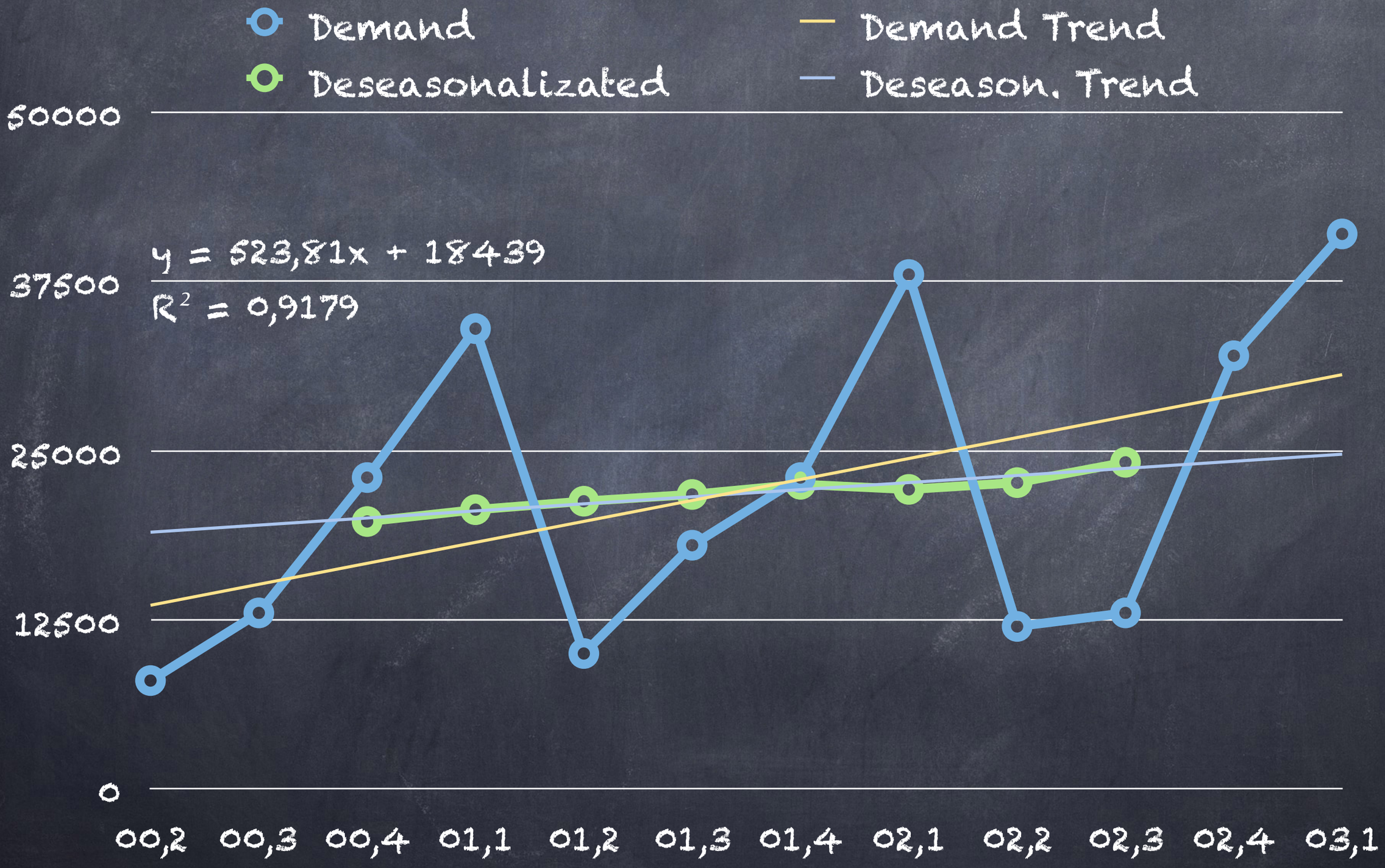
Quarterly Demand Example



Deseasonalized Demand

Year,	Period t	Demand Dt	Deseasonsalized
00,2	1	8.000	
00,3	2	13.000	
00,4	3	23.000	19.750
01,1	4	34.000	20.625
01,2	5	10.000	21.250
01,3	6	18.000	21.750
01,4	7	23.000	22.500
02,1	8	38.000	22.125
02,2	9	12.000	22.625
02,3	10	13.000	24.125
02,4	11	32.000	
03,1	12	41.000	

Deseasonalized Demand



Deseasonalized Demand

Year,	Period t	Demand Dt	Deseas. Demand	Deseas. Eq	Deseas. F.
00,2	1	8.000		18.963	0,42
00,3	2	13.000		19.487	0,67
00,4	3	23.000	19.750	20.010	1,15
01,1	4	34.000	20.625	20.534	1,66
01,2	5	10.000	21.250	21.058	0,47
01,3	6	18.000	21.750	21.582	0,83
01,4	7	23.000	22.500	22.106	1,04
02,1	8	38.000	22.125	22.629	1,68
02,2	9	12.000	22.625	23.153	0,52
02,3	10	13.000	24.125	23.677	0,55
02,4	11	32.000		24.201	1,32
03,1	12	41.000		24.725	1,66

Deseasonalized Demand

Year,	Period t	Demand Dt	Deseas. Demand	Deseas. Eq	Deseas. F.
00,2	1	8.000		18.963	0,42
00,3	2	13.000		19.487	0,67
00,4	3	23.000	19.750	20.010	1,15
01,1	4	34.000	20.625	20.534	1,66
01,2	5	10.000	21.250	21.058	0,47
01,3	6	18.000	21.750	21.582	0,83
01,4	7	23.000	22.500	22.106	1,04
02,1	8	38.000	22.125	22.629	1,68
02,2	9	12.000	22.625	23.153	0,52
02,3	10	13.000	24.125	23.677	0,55
02,4	11	32.000		24.201	1,32
03,1	12	41.000		24.725	1,66
F03,2	13	11.909		25.249	0,47
F03,3	14	17.613		25.772	0,68
F03,4	15	30.785		26.296	1,17
F04,1	16	44.640		26.820	1,66

P	D.F.
1	0,47
2	0,68
3	1,17
4	1,66

Demand Forecast

Linear Regression

Year, Qtr	X	Y	X*Y	X*X
00,4	3	19.750	59.250	9
01,1	4	20.625	82.500	16
01,2	5	21.250	106.250	25
01,3	6	21.750	130.500	36
01,4	7	22.500	157.500	49
02,1	8	22.125	177.000	64
02,2	9	22.625	203.625	81
02,3	10	24.125	241.250	100
8	52	174.750	1.157.875	380

Regression Equation $y = a + bx$

Slope $b = (N\Sigma XY - (\Sigma X)(\Sigma Y)) / (N\Sigma X^2 - (\Sigma X)^2)$

Intercept $a = (\Sigma Y - b(\Sigma X)) / N$

Slope	523,81
Intercept	18439

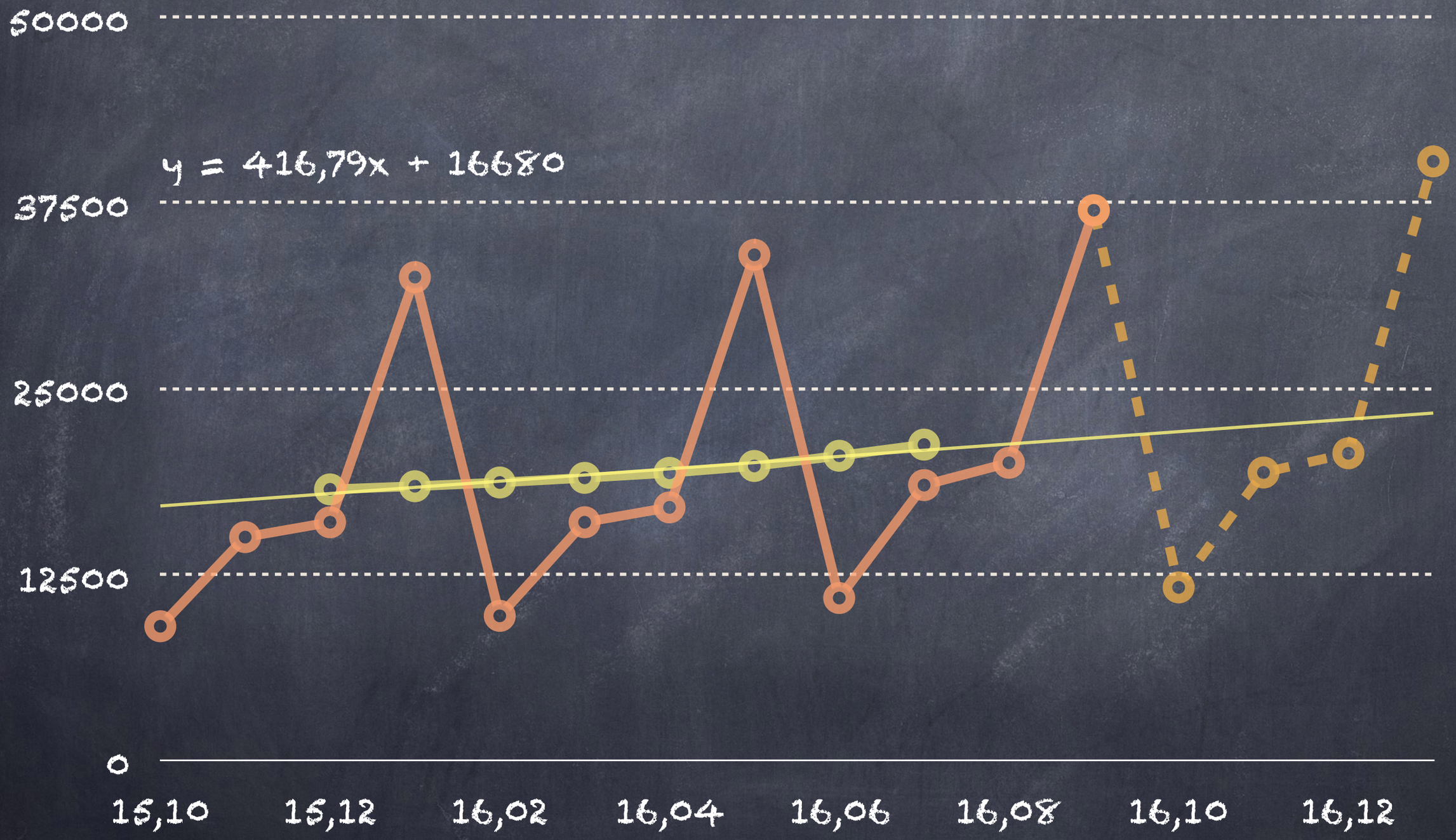
Demand Example #1

Year, Month	Period t	Demand Dt
15,10	1	9.000
15,11	2	15.000
15,12	3	16.000
16,01	4	32.500
16,02	5	9.700
16,03	6	16.000
16,04	7	17.000
16,05	8	34.000
16,06	9	10.900
16,07	10	18.500
16,08	11	20.000
16,09	12	37.000

Demand Example

Y, M	Period t	Demand Dt	Deseas. Demand	Linear Regr	S.F.
15,10	1	9.000		17.097	0,53
15,11	2	15.000		17.514	0,86
15,12	3	16.000	18.213	17.930	0,89
16,01	4	32.500	18.425	18.347	1,77
16,02	5	9.700	18.675	18.764	0,52
16,03	6	16.000	18.988	19.181	0,83
16,04	7	17.000	19.325	19.598	0,87
16,05	8	34.000	19.788	20.014	1,70
16,06	9	10.900	20.475	20.431	0,53
16,07	10	18.500	21.225	20.848	0,89
16,08	11	20.000		21.265	0,94
16,09	12	37.000		21.681	1,71
F16,10	13	11.615		22.098	0,53
F16,11	14	19.348		22.515	0,86
F16,12	15	20.641		22.932	0,90
F17,01	16	40.290		23.349	1,73

Graph Example #1



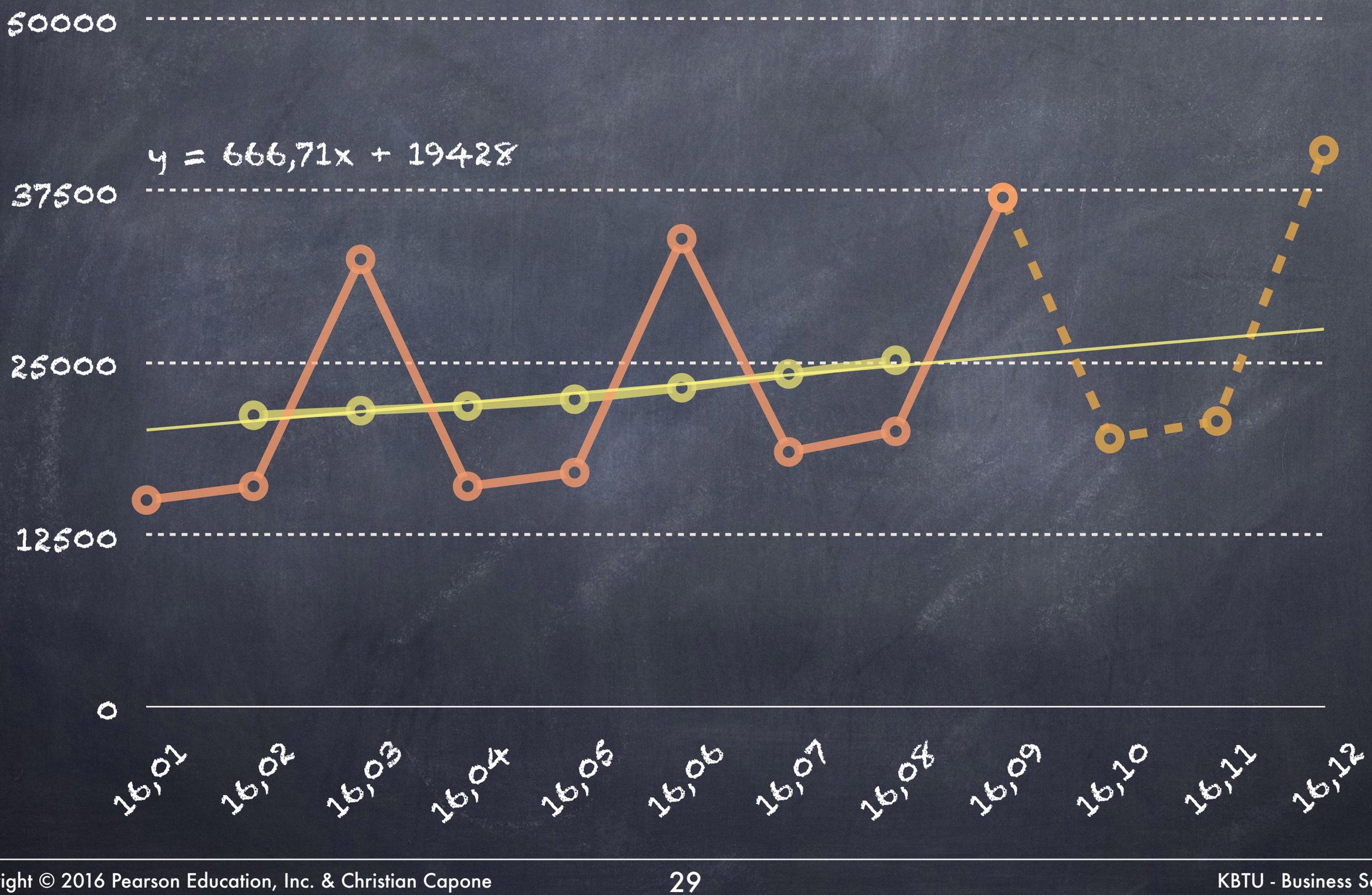
Demand Example #2

Year, Month	Period t	Demand Dt
16,01	1	15.000
16,02	2	16.000
16,03	3	32.500
16,04	4	16.000
16,05	5	17.000
16,06	6	34.000
16,07	7	18.500
16,08	8	20.000
16,09	9	37.000

Demand Example #2

Year,	Period	Demand Dt	Deseas. Demand	Linear Regr	S.F.
16,01	1	15.000		20.095	0,75
16,02	2	16.000	21.167	20.761	0,77
16,03	3	32.500	21.500	21.428	1,52
16,04	4	16.000	21.833	22.095	0,72
16,05	5	17.000	22.333	22.762	0,75
16,06	6	34.000	23.167	23.428	1,45
16,07	7	18.500	24.167	24.095	0,77
16,08	8	20.000	25.167	24.762	0,81
16,09	9	37.000		25.428	1,46
F16,10	10	19.471		26.095	0,75
F16,11	11	20.742		26.762	0,78
F16,12	12	40.439		27.429	1,47

Graph Example #2




Q&A