MONASH University

Information Technology

FIT1006
Business Information Analysis

Lecture 8
Linear Regression

Topics covered:

- Estimating the regression equation by eye.
- Fitting a regression using Excel and SYSTAT.
- Measuring the goodness of fit.
- Modelling with the regression equation.

Linear Regression

- Regression is the practice of describing the (linear) relationship between 2 or more quantitative variables.
- Thus if we know the value of one variable, we can estimate the value of the related variable of interest.
- Origin: The 19th century scientist Francis Galton collected data on the heights of fathers and their sons. He found that tall fathers had slightly shorter sons and that short fathers had slightly taller sons. Thus in each case there was a regression (reversion) to the mean. Over time the details of the investigation have been forgotten but the name has stuck to this method of modelling.

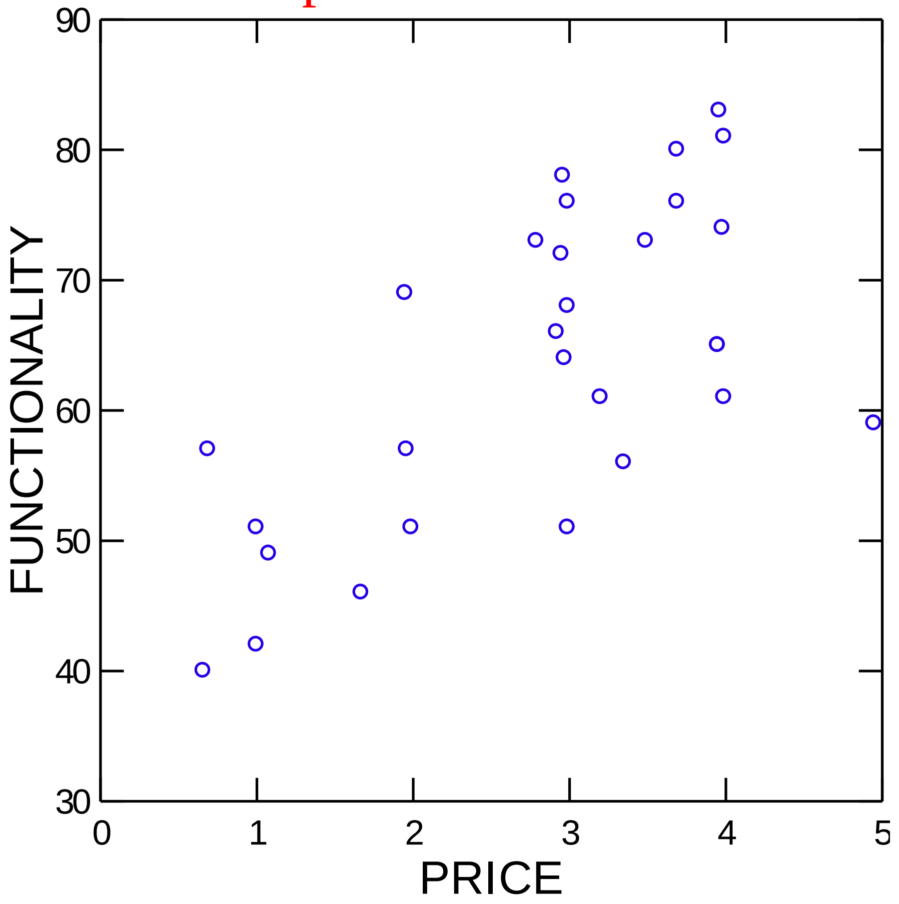
Motivating Question

- In 1998, *Choice* magazine tested 1500 toothbrushes.
- A summary of price and functionality score is on the right.
- What is the relationship between price and functionality?
- How reliable is the model? (Selvanathan 4th Ed p 679)
- Answers later…

Price	Functionality
3.96	83
3.99	81
3.69	80
2.96	78
3.69	76
2.99	76
3.98	74
2.79	73
3.49	73
2.95	72
1.95	69
2.99	68
2.92	66
3.95	65
3.95	65
2.97	64
3.99	61
3.20	61
4.95	59
0.69	57
1.96	57
3.35	56
1.00	51
2.99	51
1.99	51
1.08	49
1.67	46
1.00	42
0.66	40

Motivating Question

Scatterplot



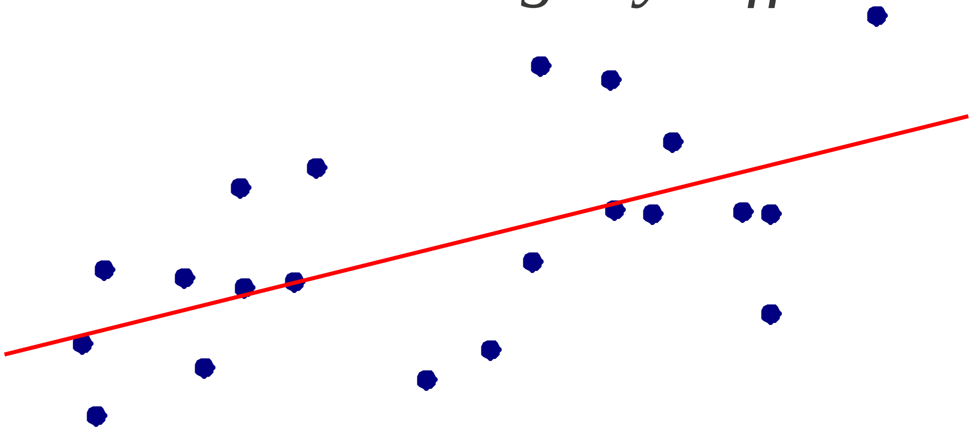
Price	Functionality
3.96	83
3.99	81
3.69	80
2.96	78
3.69	76
2.99	76
3.98	74
2.79	73
3.49	73
2.95	72
1.95	69
2.99	68
2.92	66
3.95	65
3.95	65
2.97	64
3.99	61
3.20	61
4.95	59
0.69	57
1.96	57
3.35	56
1.00	51
2.99	51
1.99	51
1.08	49
1.67	46
1.00	42
0.66	40

The underlying assumption

When we calculate the regression of y on x, we are assuming that the relationship between x and y is linear and thus we can say that $y = ax + b + e$, where e are random, normally distributed errors.

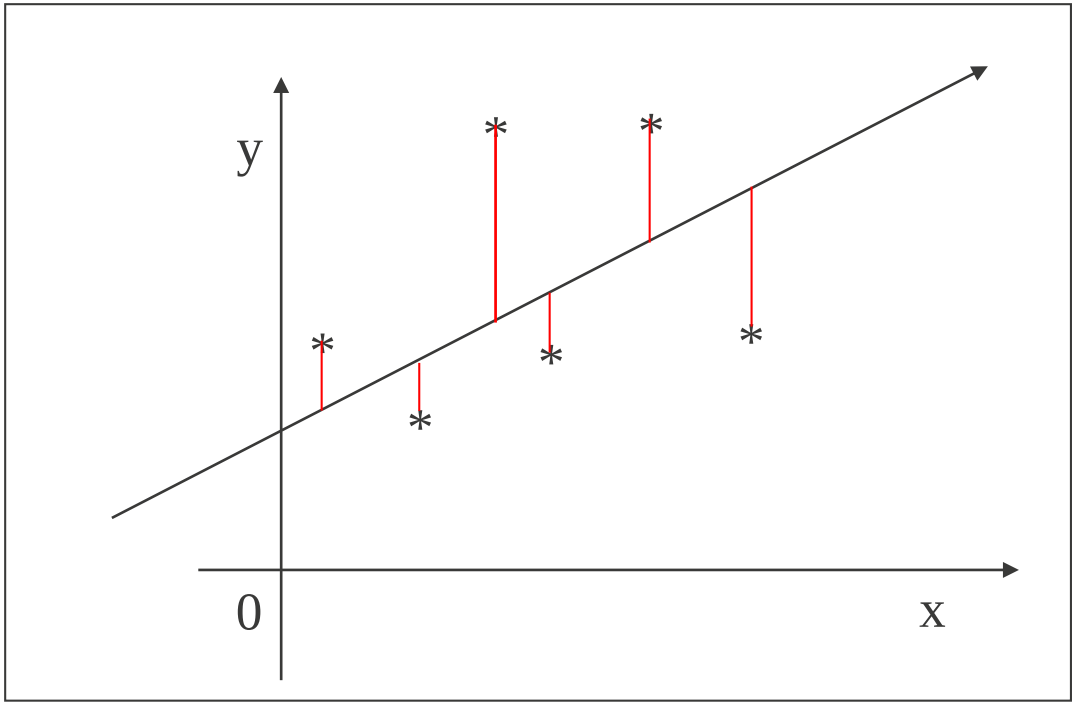
We want to find the value of a and b.

(*Note: the textbook uses slightly different notation*)



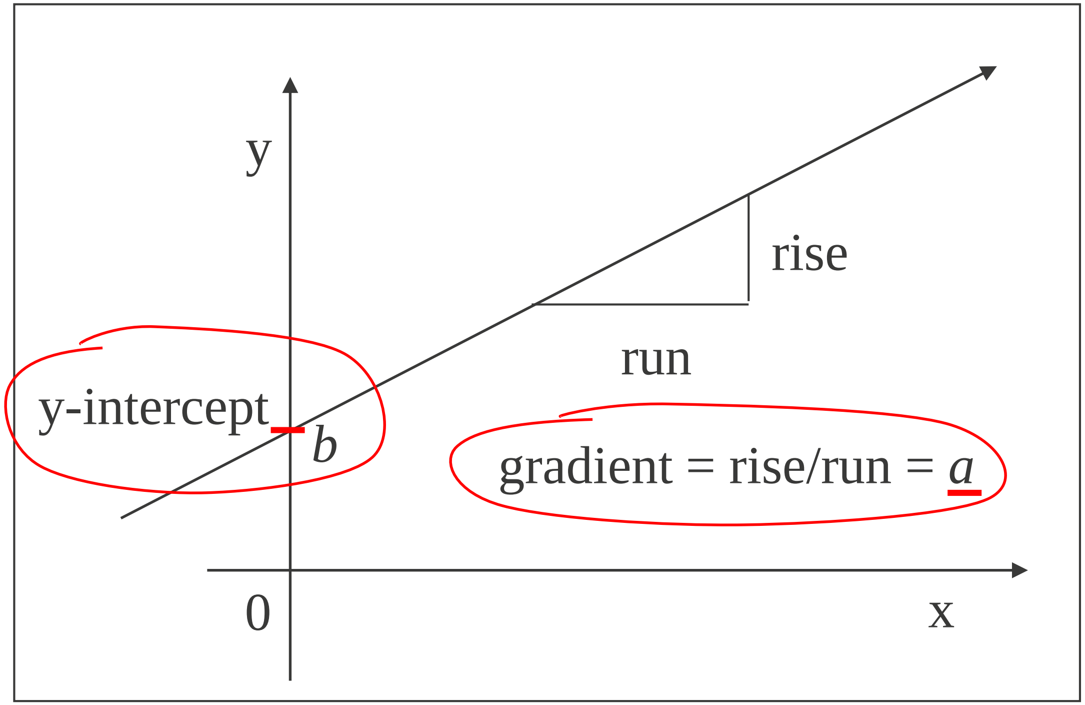
The basic idea

We want to fit a line through the data that minimises the sum of the squared errors, or differences, between the fitted model (line) and the data.



The equation of a straight line

We can use the basic equation of a straight line as the model for our regression equation. A line with gradient ‘*a*’ and y-intercept ‘*b*’ has equation: *y = ax + b.*



Least Squares Regression

- Ordinary Least Squares Regression minimises the sum of squared errors in the data.
- The regression of *y* on *x* as *y* = *ax* + *b* is:

$$a = \frac{s_{xy}}{s_x{}^2} = \frac{\Sigma xy - \dfrac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \dfrac{\left(\Sigma x\right)^2}{n}} \text{ and } b = \bar{y} - a\bar{x}$$

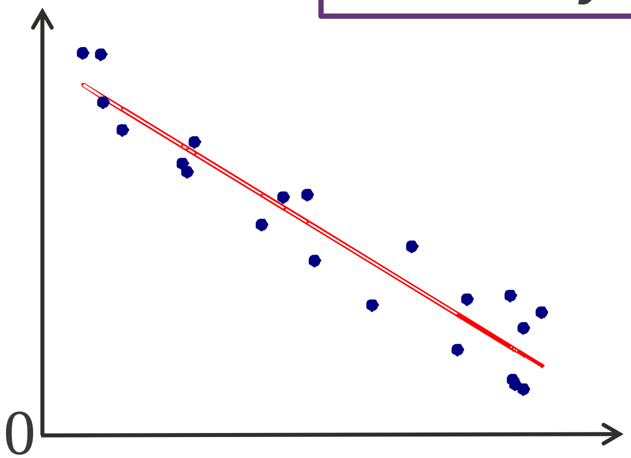
$$\boxed{note : \bar{y} = \frac{\Sigma y}{n} etc.}$$

https://flux.qa (Feed code: SJ6KGV)

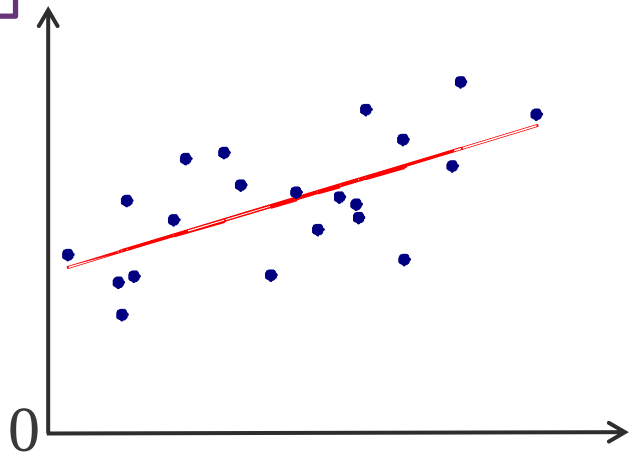
Question 1: In which plot is ‘*b*’ greatest?

Assume: $y = ax + b$

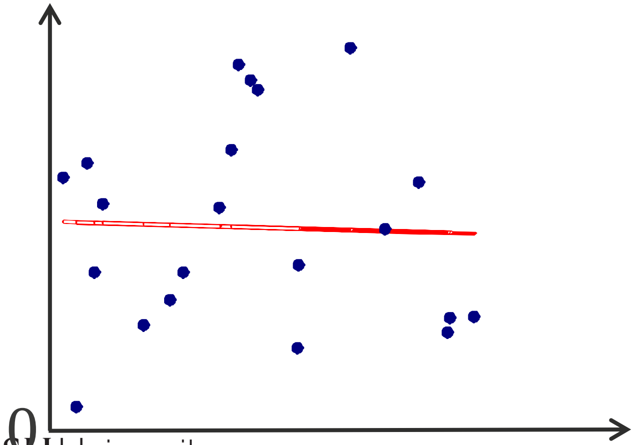
✓ A.



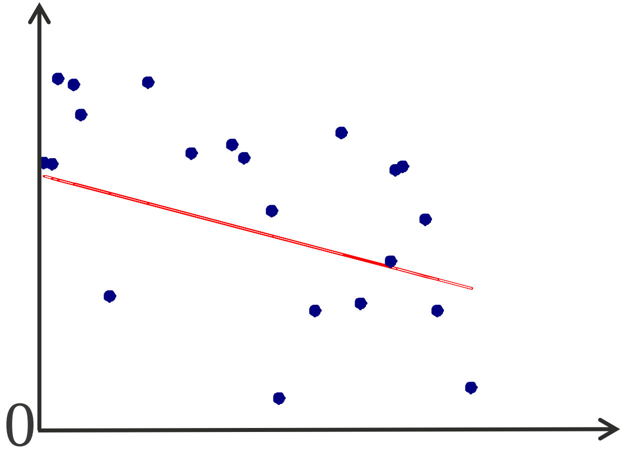
B.



C.



D.

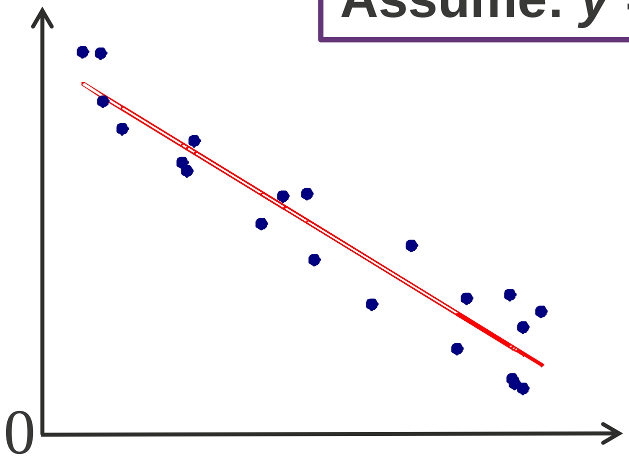


https://flux.qa (Feed code: SJ6KGV)

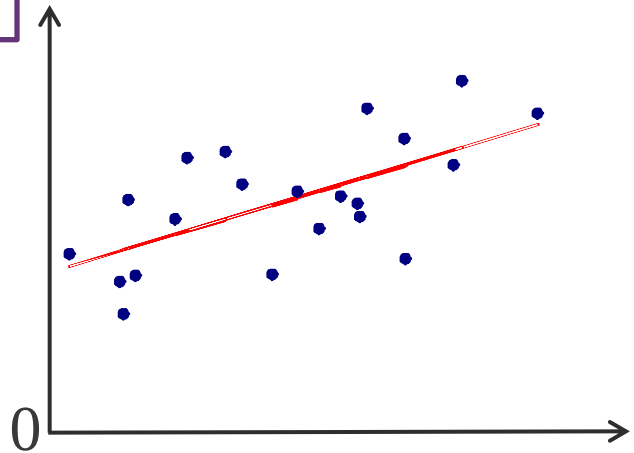
Question 2: In which plot is ‘a’ greatest

Assume: *y* = *ax* + *b*

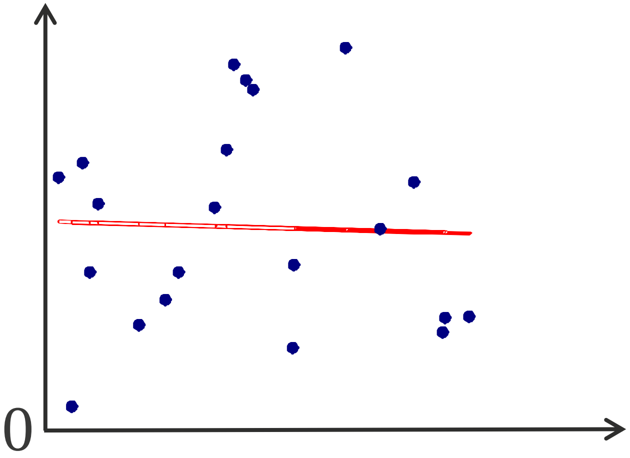
A.



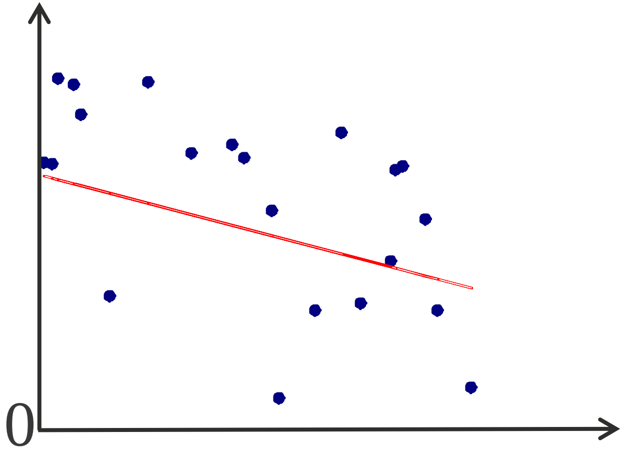
✓ B.



C.

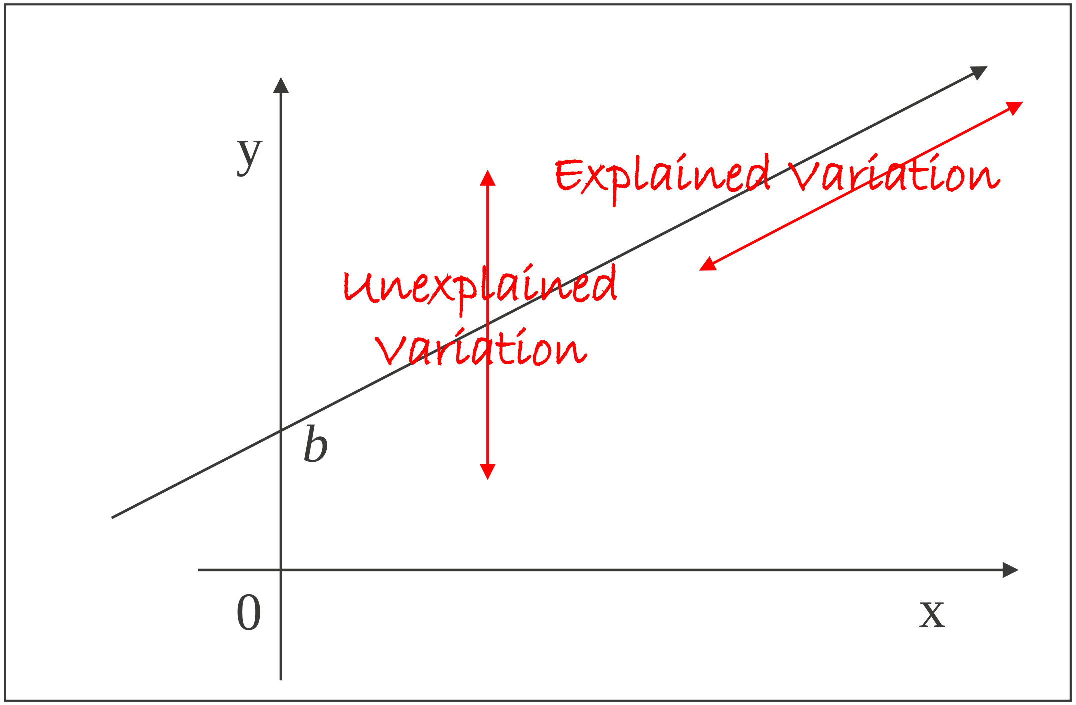


D.



How good is the fit?

- One measure of how well the regression model is the proportion of variation in *y* that is explained by the regression equation.



Coefficient of Determination

- The coefficient of determination is the proportion of variation in *y* that is explained by variation in *x* through the regression equation.

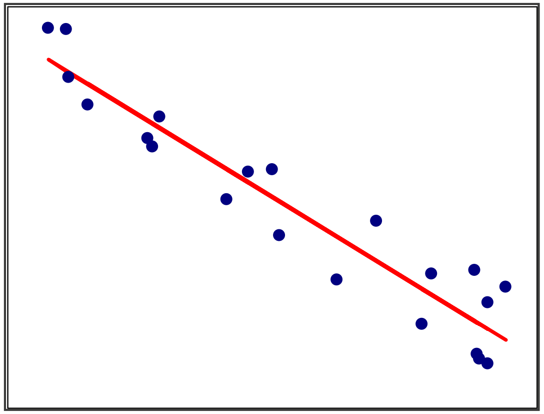
- The coefficient of determination is r^2 – the square of Pearson’s correlation coefficient *r*.

- Often 100 r^2 is calculated and the result expressed as a percentage.

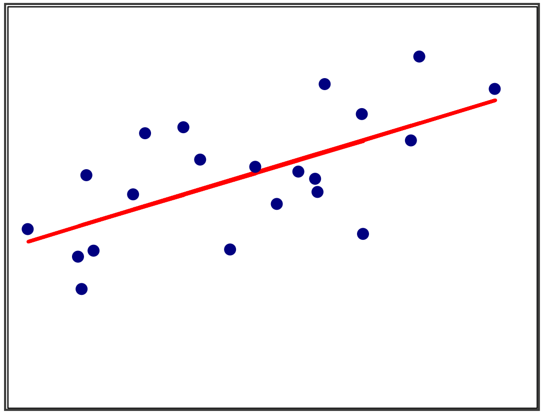
https://flux.qa (Feed code: SJ6KGV)

Question 3: In which plot has r^2 closest to 1?

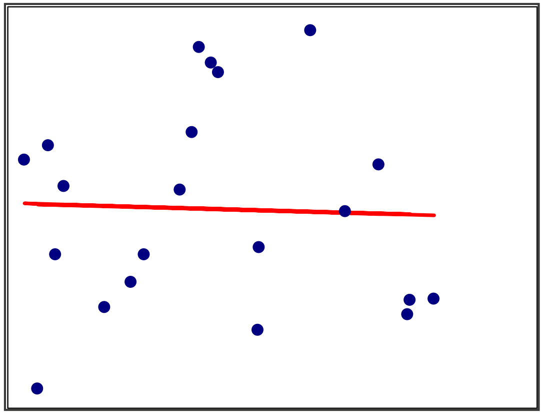
✓ A.



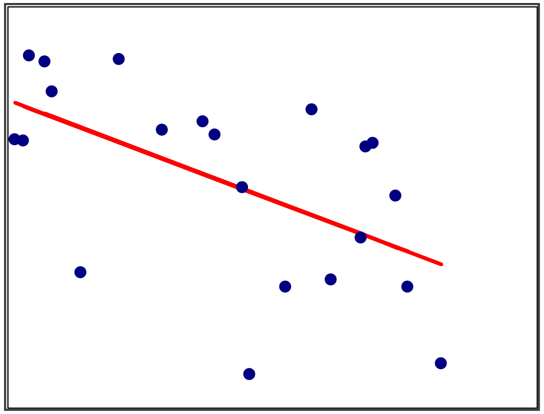
B.



C.



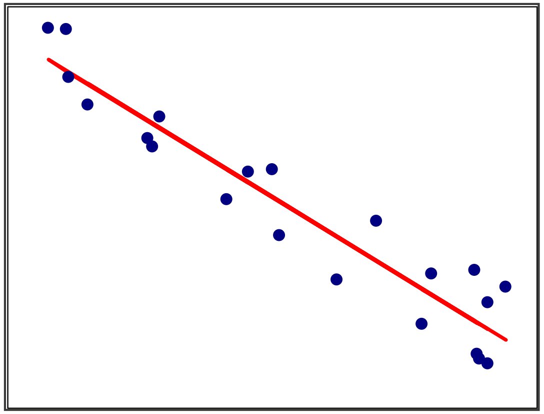
D.



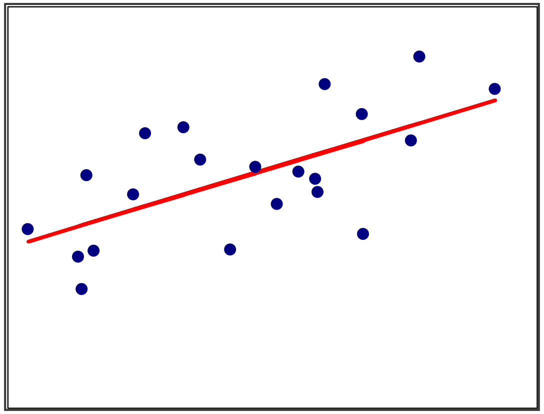
https://flux.qa (Feed code: SJ6KGV)

Question 4: In which plot has r^2 closest to 0?

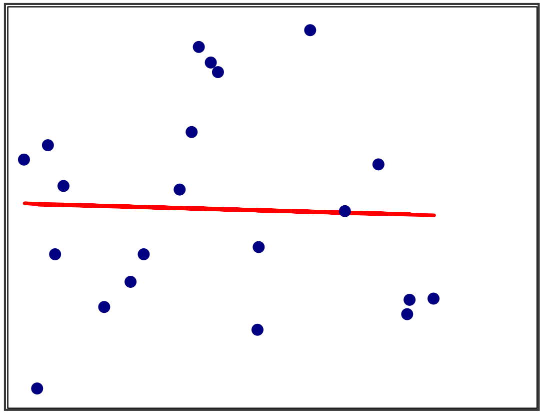
A.



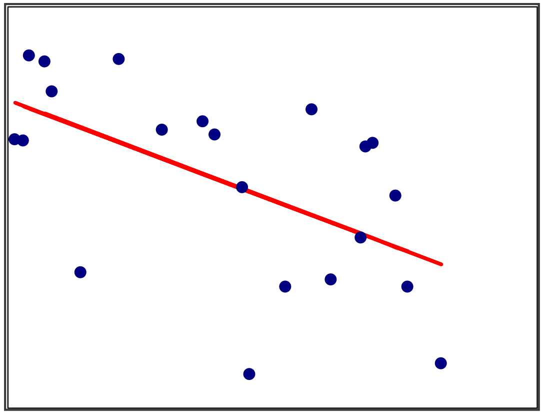
B.



✓ C.



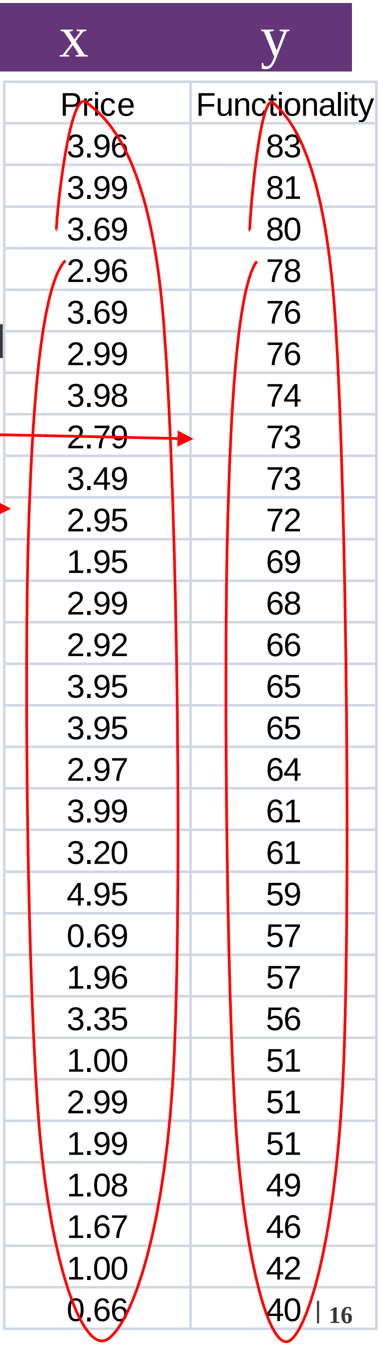
D.



MONASH University

Regression in EXCEL

- Regression is a built in analysis function, or you can also calculate the formulas manually with:
  - *a* = SLOPE(y values, x values)
  - *b* = INTERCEPT(y values, x values)
- Also,
  - *r* = CORREL(y values, x values)
  - r^2 = CORREL(y values, x values)^2
- Regression is also a ‘Chart Tool’ if you first draw a scatter plot and then choose this option.



x	y
Price	Functionality
3.96	83
3.99	81
3.69	80
2.96	78
3.69	76
2.99	76
3.98	74
2.79	73
3.49	73
2.95	72
1.95	69
2.99	68
2.92	66
3.95	65
3.95	65
2.97	64
3.99	61
3.20	61
4.95	59
0.69	57
1.96	57
3.35	56
1.00	51
2.99	51
1.99	51
1.08	49
1.67	46
1.00	42
0.66	40

Regression in SYSTAT

SYSTAT calculates regression and gives a diagnostic output of the fitted model.

- Select: Analyze > Regression > Linear > Least Squares
- The dependent variable is the one we’re trying to predict.
- The independent variable is the one that is free to change.
- The model and residuals can be saved to a data file.

Regression by hand

We stopped here … will continue the rest next week

- Use the same terms you calculated for the least squares correlation: Σ*x*, Σ*y*, Σ*x*², Σ*y*², Σ*xy*, and *n*.

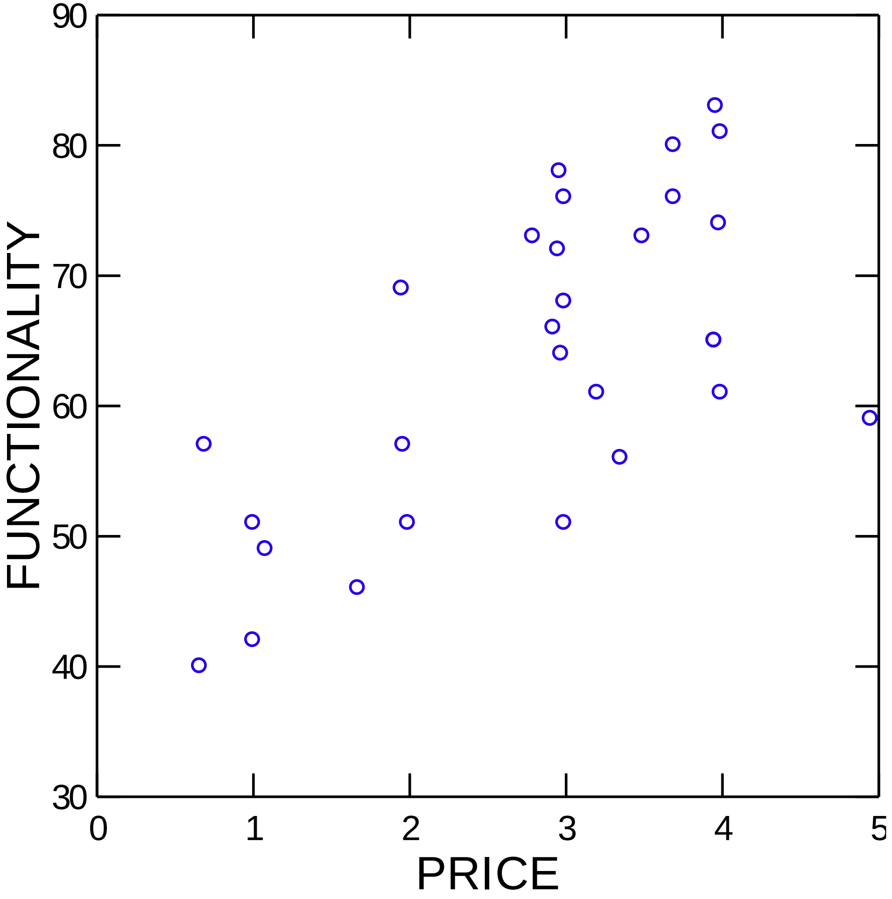
$$a = \frac{s_{xy}}{s_x^{2}} = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma x^{2} - \frac{\left(\Sigma x\right)^{2}}{n}} \text{ and } b = \bar{y} - a\bar{x}$$

$$\bar{y} = \frac{\Sigma y}{n}$$

- Know how to calculate the regression equation using your calculator.

Motivating Question

- Let’s try and fit the Line of Best Fit by eye:



https://flux.qa (Feed code: SJ6KGV)

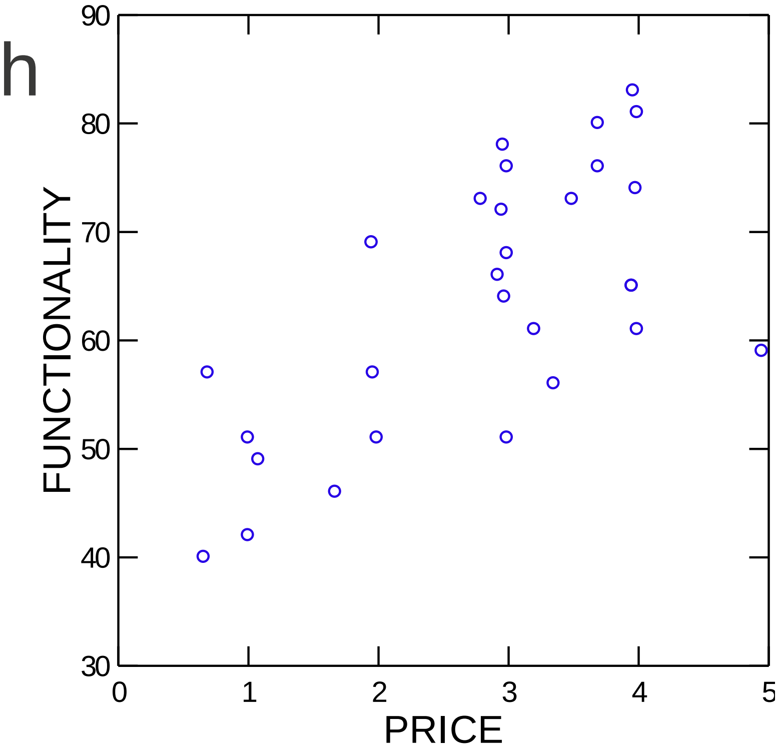
Question 5: For the toothbrush problem which assumption is true:

A. Price and function are both independent.

B. Function is independent

C. Price is dependent.

D. Price is independent.



SYSTAT Output (a) report

```
Dependent Variable            ¦ FUNCTIONALITY
N                             ¦ 29
Multiple R                    ¦ 0.664
Squared Multiple R            ¦ 0.441
Adjusted Squared Multiple R   ¦ 0.421
Standard Error of Estimate    ¦ 9.187


Regression Coefficients B = (X'X)-1X'Y

         ¦                                   Std.
Effect   ¦ Coefficient   Standard Error   Coefficient   Tolerance       t   p-Value
---------+-----------------------------------------------------------------------
CONSTANT ¦      44.025            4.567         0.000           .   9.640     0.000
PRICE    ¦       6.939            1.503         0.664       1.000   4.618     0.000
```

Pearson's r

Coeff of determination: r^2

Y-intercept: b = 44.025

Gradient: a = 6.939

Least square regression line is: Y = 6.639 x + 44.025

SYSTAT Output (a) report

```
Analysis of Variance

Source      ¦          SS    df    Mean Squares    F-Ratio    p-Value
------------+-----------------------------------------------------------
Regression  ¦ 1,800.032     1        1,800.032     21.325      0.000
Residual    ¦ 2,279.003    27           84.408


Durbin-Watson D-Statistic    ¦ 0.946
First Order Autocorrelation ¦ 0.482


Information Criteria

AIC              ¦ 214.860
AIC (Corrected)  ¦ 215.820
Schwarz's BIC    ¦ 218.962
```

SYSTAT Output (a) report

- Tweaking the data by changing data point 19 from (4.95, 59) to (4.95, 39) results in a warning:

```
Analysis of Variance

Source       ¦          SS    df    Mean Squares    F-Ratio    p-Value
-------------+------------------------------------------------------
Regression   ¦ 1,257.113     1       1,257.113     10.008      0.004
Residual     ¦ 3,391.577    27         125.614

*** WARNING *** :

Case 19 is an Outlier (Studentized Residual : -4.698)
```

- Other warnings are for ‘leverage’ and large residuals.

https://flux.qa (Feed code: SJ6KGV)

Question 6

If regression equation is: Function = 44 + Price × 7, a toothbrush having a Price of $3 would have a Function of:

A. 44
B. 51
C. 54
D. 65
E. None of the above.

The Challenger Disaster

- The Space Shuttle Challenger disaster occurred on January 28, 1986, when Space Shuttle Challenger broke apart 73 seconds into its flight, leading to the deaths of its seven crew members… *(Text and images: Wikipedia)*

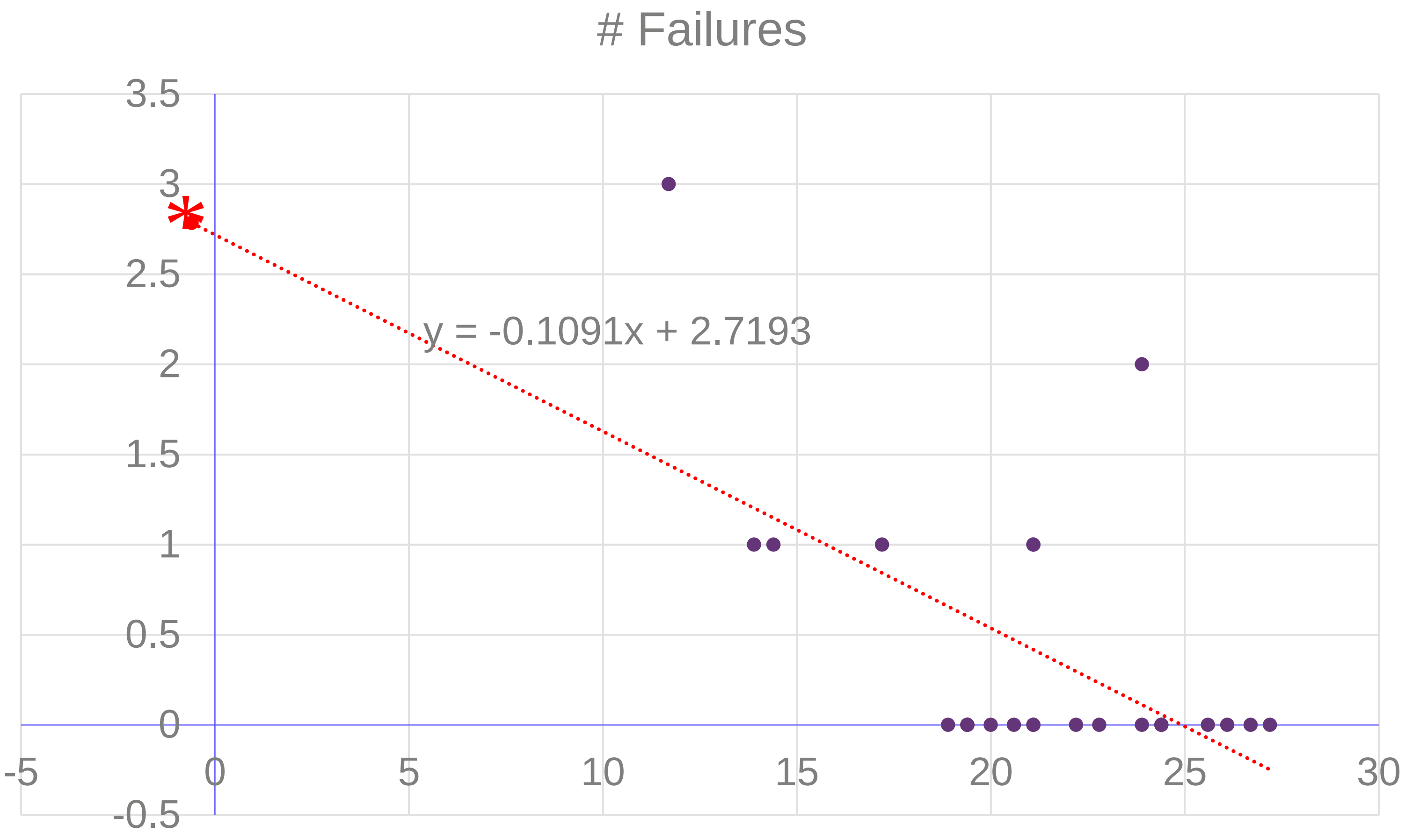
The Challenger Disaster

Data

Flight No	Date	Temp F	Temp C	# Failures
1	04-12-81	66	18.9	0
2	11-12-81	70	21.1	1
3	03-22-82	69	20.6	0
4	06-27-82	80	26.7	*
5	11-11-82	68	20.0	0
6	04-04-83	67	19.4	0
7	06-18-83	72	22.2	0
8	08-30-83	73	22.8	0
9	11-28-83	70	21.1	0
10	02-03-84	57	13.9	1
11	04-06-84	63	17.2	1
12	08-30-84	70	21.1	1
13	10-05-84	78	25.6	0
14	11-08-84	67	19.4	0
15	01-24-85	53	11.7	3
16	04-12-85	67	19.4	0
17	04-29-85	75	23.9	0
18	06-17-85	70	21.1	0
19	07-29-85	81	27.2	0
20	08-27-85	76	24.4	0
21	10-03-85	79	26.1	0
22	10-30-85	75	23.9	2
23	11-26-85	76	24.4	0
24	01-12-86	58	14.4	1
Temperature on launch		31	-0.6	
From: http://wps.aw.com/wps/media/objects/15/15719/projects/ch5_challenger/index.html				

The Challenger Disaster

Temperature at launch *



Group activity

$Y_f = 0.535x_f + 138.56$

$Y_m = 0.425x_m + 150.47$

- The Height v Weight of 102 elite male and 100 elite female athletes at the AIS.
- See: FIT1006 Lecture 08 Worksheet.
- In groups work out the target height for a 100kg male and an 80kg female athlete.
- The most accurate prediction can be made for which gender? Why?
- Comment on the differences between genders of their weight profile.
- Data Source: http://www.statsci.org/data/oz/ais.html

One More Thing

- If we interchange *x* and *y* of our model we get a different regression equation, not just the inverse equation.
- Why?

Necessary Skills

- Calculate the least squares regression by hand (using your calculator) for a small data set.
- Interpret the basic SYSTAT output and comment on any data points that have a significant affect on the regression model.
- Draw a scatterplot and superimpose the line of best fit.
- Calculate Pearson’s *r*, r^2 and Comment on the goodness of fit of the regression.

Reading/Questions

- **Reading:**
  - 7th Ed Sections 15.1 - 15.4, 15.7, 16.1*, 16.2*.
  - *Additional reading on multiple regression.
- **Questions:**
  - 7th Ed Questions 15.6, 15.7, 15.8, 15.10, 15.12, 15.14, 15.19, 15.21, 15.17, 15.63, 15.64.