Stats for Engineers Lecture 9

Summary From Last Time

Confidence Intervals for the mean

If σ^2 is ***known***, confidence interval for μ is $\bar{X} - z\sqrt{\frac{\sigma^2}{n}}$ to $\bar{X} + z\sqrt{\frac{\sigma^2}{n}}$, where z is obtained from Normal tables.

If σ^2 is ***unknown*** (only know sample variance s^2):

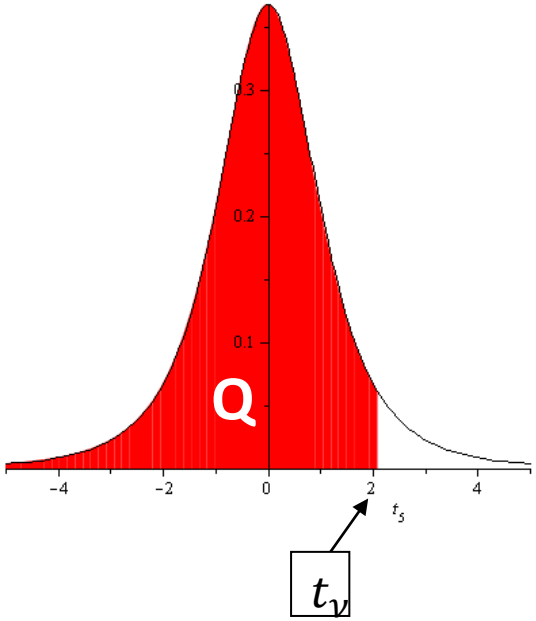
Assuming independent Normal data, the confidence interval for μ is:

$\bar{X} - t_{n-1}\sqrt{\frac{s^2}{n}}$ to $\bar{X} + t_{n-1}\sqrt{\frac{s^2}{n}}$.

Student t-distribution

t-tables $Q(t_\nu) = \int_{-\infty}^{t_\nu} f_\nu(t)dt$

$\nu = n - 1$



Sample size

How many random samples do you need to reach desired level of precision?

Suppose we want to estimate μ to within $\pm\delta$, where δ (and the degree of confidence) is given.

Want $$\delta = t_{n-1}\sqrt{\frac{s^2}{n}}$$

$$\Rightarrow n = \frac{t_{n-1}^2 s^2}{\delta^2}$$

Need:

- Estimate of s^2 (e.g. previous experiments)

- Estimate of t_{n-1}. This depends on n, but not very strongly.

e.g. take $t_{n-1} = 2.1$ for 95% confidence.

Rule of thumb: for 95% confidence, choose $n = \frac{2.1^2 \times \text{Estimate of variance}}{\delta^2}$

Student t distribution: values of *x*

DF	Q					
ν	0.95	0.975	0.99	0.995	0.999	0.9995
1	6.3138	12.7062	31.8205	63.6567	318.309	636.619
2	2.9200	4.3027	6.9646	9.9248	22.327	31.599
3	2.3534	3.1824	4.5407	5.8409	10.215	12.924
4	2.1318	2.7764	3.7469	4.6041	7.173	8.610
5	2.0150	2.5706	3.3649	4.0321	5.893	6.869
6	1.9432	2.4469	3.1427	3.7074	5.208	5.959
7	1.8946	2.3646	2.9980	3.4995	4.785	5.408
8	1.8595	2.3060	2.8965	3.3554	4.501	5.041
9	1.8331	2.2622	2.8214	3.2498	4.297	4.781
10	1.8125	2.2281	2.7638	3.1693	4.144	4.587
11	1.7959	2.2010	2.7181	3.1058	4.025	4.437
12	1.7823	2.1788	2.6810	3.0545	3.930	4.318
13	1.7709	2.1604	2.6503	3.0123	3.852	4.221
14	1.7613	2.1448	2.6245	2.9768	3.787	4.140
15	1.7531	2.1314	2.6025	2.9467	3.733	4.073
16	1.7459	2.1199	2.5835	2.9208	3.686	4.015
17	1.7396	2.1098	2.5669	2.8982	3.646	3.965
18	1.7341	2.1009	2.5524	2.8784	3.610	3.922
19	1.7291	2.0930	2.5395	2.8609	3.579	3.883
20	1.7247	2.0860	2.5280	2.8453	3.552	3.850
21	1.7207	2.0796	2.5176	2.8314	3.527	3.819
22	1.717	2.0739	2.5083	2.8188	3.505	3.792
23	1.7139	2.0687	2.4999	2.8073	3.485	3.768
24	1.7109	2.0639	2.4922	2.7969	3.467	3.745
25	1.7081	2.0595	2.4851	2.7874	3.450	3.725
26	1.7056	2.0555	2.4786	2.7787	3.435	3.707
27	1.7033	2.0518	2.4727	2.7707	3.421	3.690
28	1.7011	2.0484	2.4671	2.7633	3.408	3.674
29	1.6991	2.0452	2.4620	2.7564	3.396	3.659
30	1.6973	2.0423	2.4573	2.7500	3.385	3.646
31	1.6955	2.0395	2.4528	2.7440	3.375	3.633
32	1.6939	2.0369	2.4487	2.7385	3.365	3.622
33	1.6924	2.0345	2.4448	2.7333	3.356	3.611
34	1.6909	2.0322	2.4411	2.7284	3.348	3.601
35	1.6896	2.0301	2.4377	2.7238	3.340	3.591
36	1.6883	2.0281	2.4345	2.7195	3.333	3.582
37	1.6871	2.0262	2.4314	2.7154	3.326	3.574
38	1.6860	2.0244	2.4286	2.7116	3.319	3.566
39	1.6849	2.0227	2.4258	2.7079	3.313	3.558
40	1.6839	2.0211	2.4233	2.7045	3.307	3.551
45	1.6794	2.0141	2.4121	2.6896	3.281	3.520
50	1.6759	2.0086	2.4033	2.6778	3.261	3.496
60	1.6706	2.0003	2.3901	2.6603	3.232	3.460
70	1.6669	1.9944	2.3808	2.6479	3.211	3.435
80	1.6641	1.9901	2.3739	2.6387	3.195	3.416
90	1.6620	1.9867	2.3685	2.6316	3.183	3.402
100	1.6602	1.9840	2.3642	2.6259	3.174	3.390
∞	1.6449	1.9600	2.3263	2.5758	3.090	3.291

Example

A large number of steel plates will be used to build a ship. Ten are tested and found to have sample mean weight $\bar{X} = 2.13\text{kg}$ and sample variance $s^2 = (0.25\ \text{kg})^2$. How many need to be tested to determine the mean weight with 95% confidence to within $\pm 0.1\ \text{kg}$?

Answer:

Assuming plates have independent weights with a Normal distribution $\delta = 0.1\text{kg} = t_{n-1}\sqrt{\frac{s^2}{n}}$

Take $t_{n-1} \approx 2.1$ for 95% confidence.

$$\Rightarrow n = \frac{t_{n-1}^2 s^2}{\delta^2} = \frac{2.1^2\ 0.25^2}{0.1^2} = 27.6$$

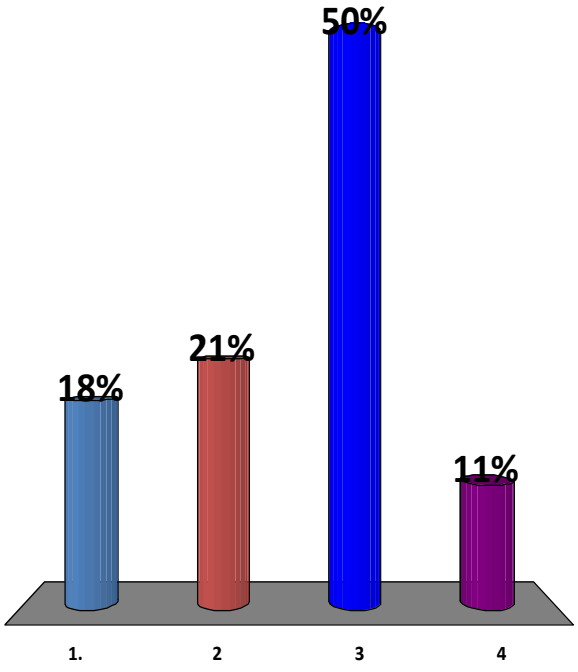
i.e. need to test about 28

Number of samples

If you need 28 samples for the confidence interval to be ± 0.1 kg, approximately how many samples would you need to get a more accurate answer with confidence interval ± 0.01 kg?

$$\delta = t_{n-1}\sqrt{\frac{s^2}{n}}$$

1. 88.5
2. 280
3. ✓ 2800
4. 28000



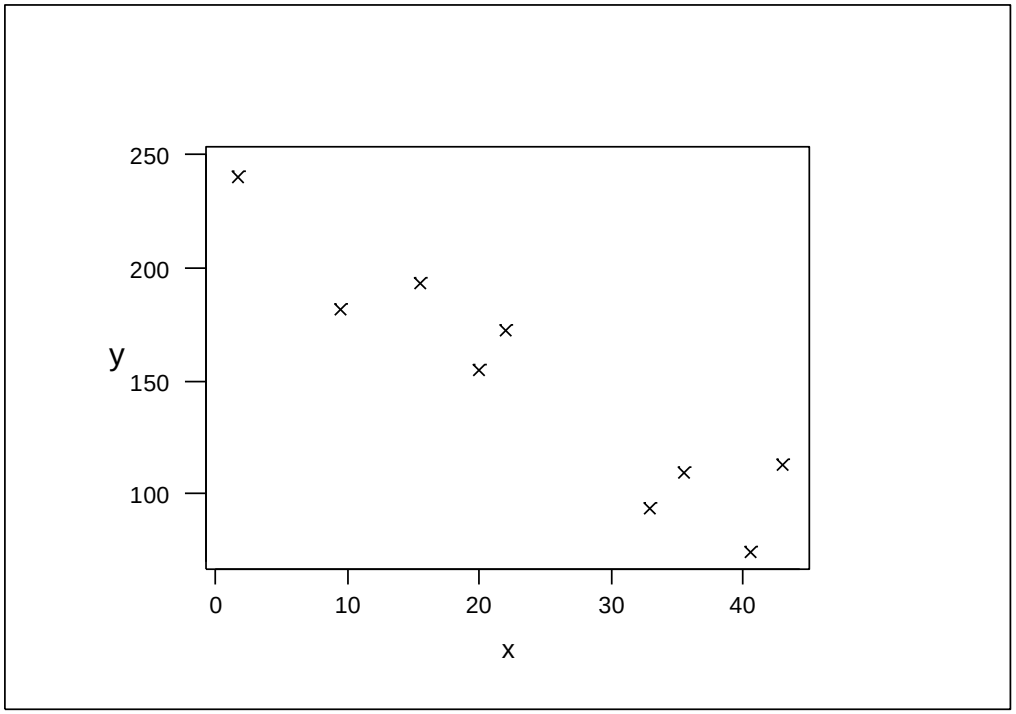
$$\delta = t_{n-1}\sqrt{\frac{s^2}{n}}$$

$$\Rightarrow \frac{\delta}{10} = t_{n-1}\sqrt{\frac{s^2}{100n}}$$ so need 100 × more. i.e. 2800

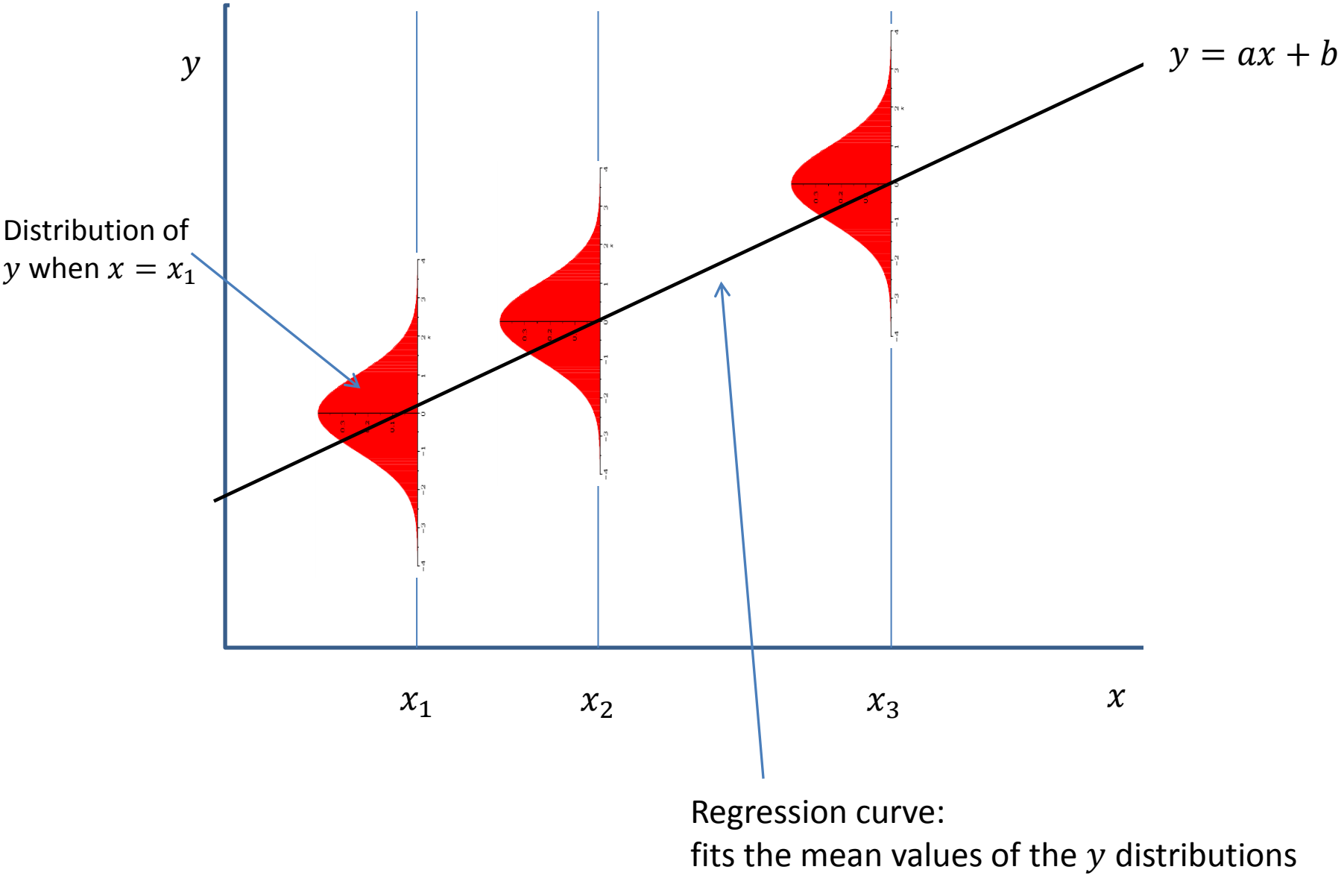
Linear regression

We measure a response variable y at various values of a controlled variable x

e.g. measure fuel efficiency y at various values of an experimentally controlled external temperature x

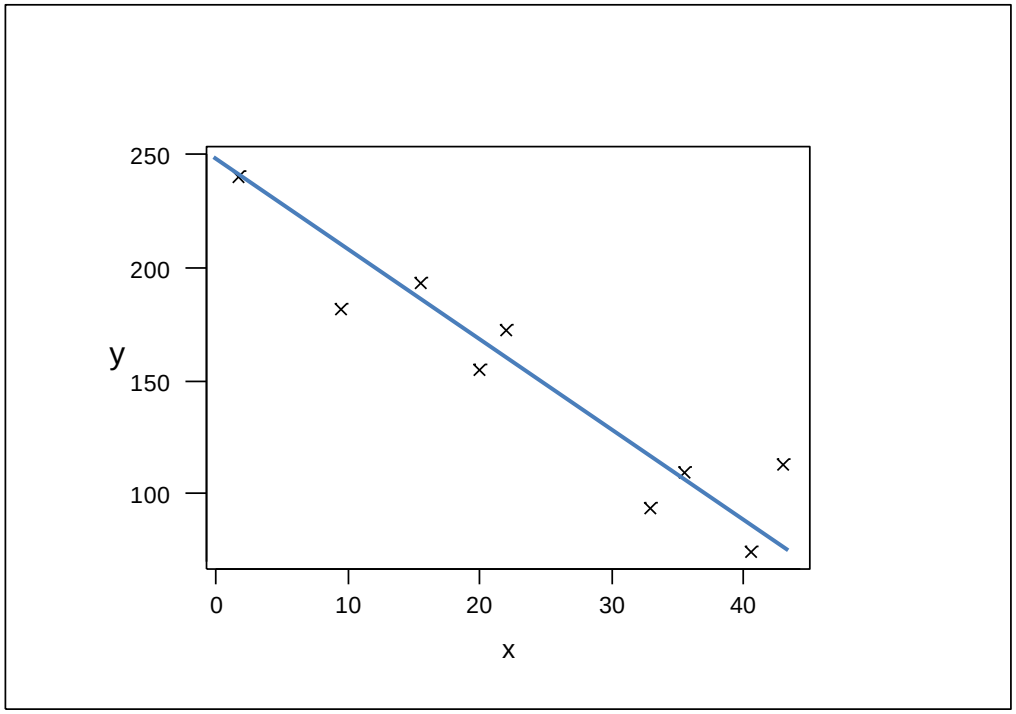


Linear regression: fitting a straight line to the *mean* value of y as a function of x


y
$y = ax + b$
Distribution of y when $x = x_1$
x_1
x_2
x_3
x
Regression curve:
fits the mean values of the y distributions

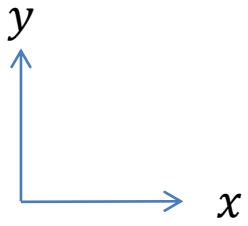
From a sample of y values at various x, we want to fit the regression curve.

e.g.

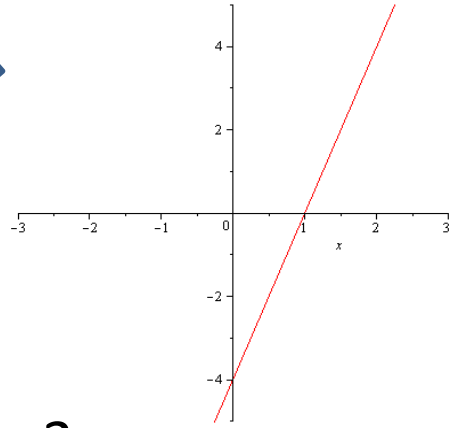


Straight line plots

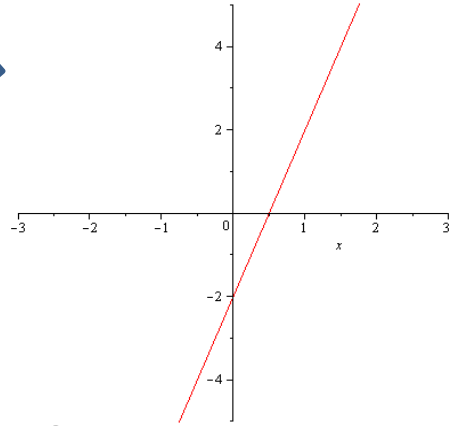
Which graph is of the line $y = 2x - 4$?



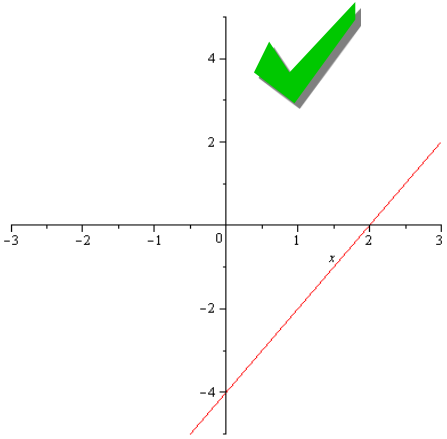
1.



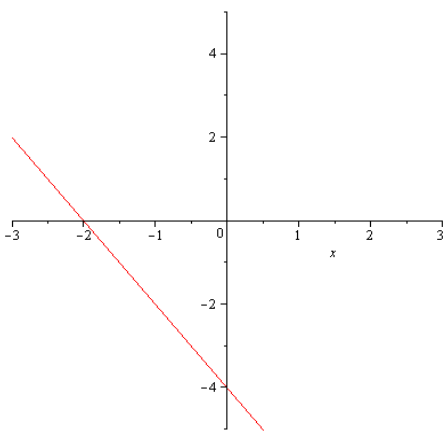
2.

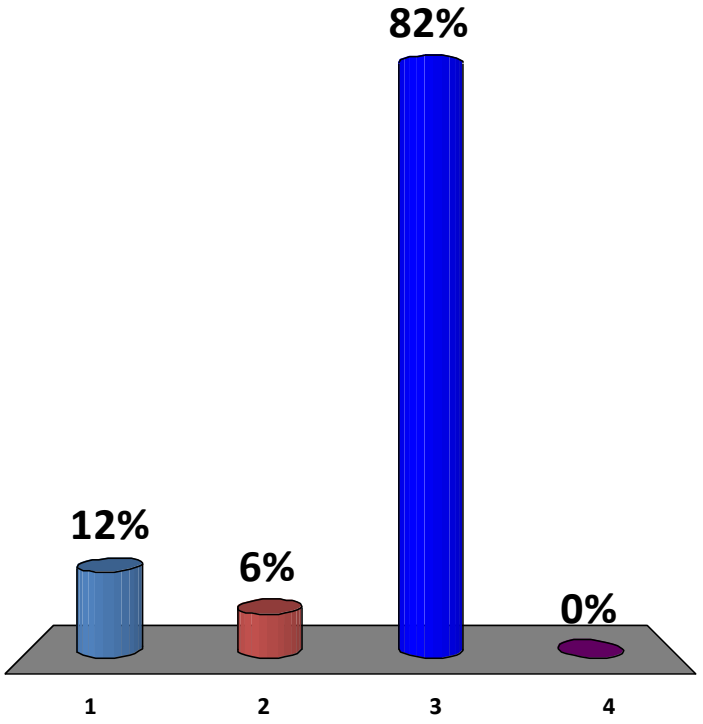


3.



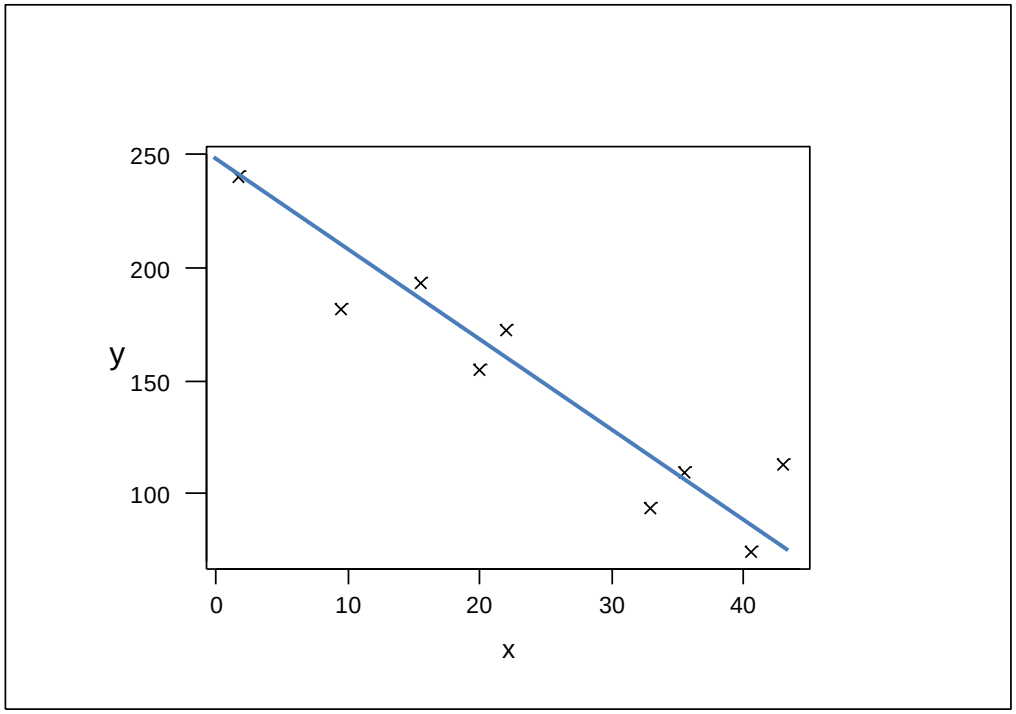
4.



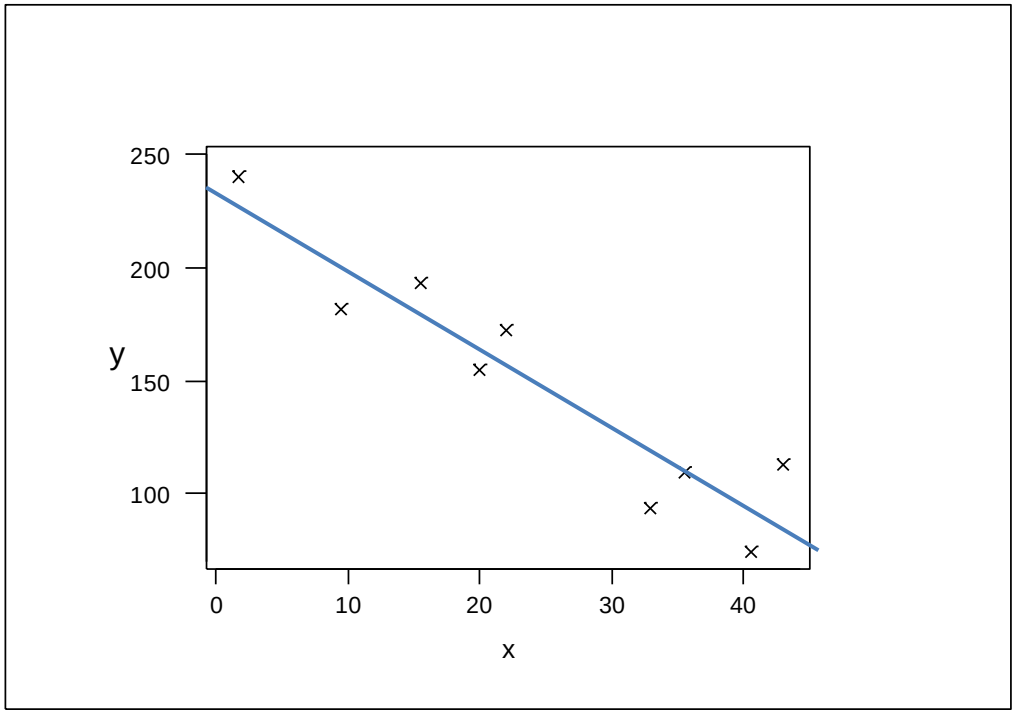


From a sample of y values at various x, we want to fit the regression curve.

e.g.



Or is it



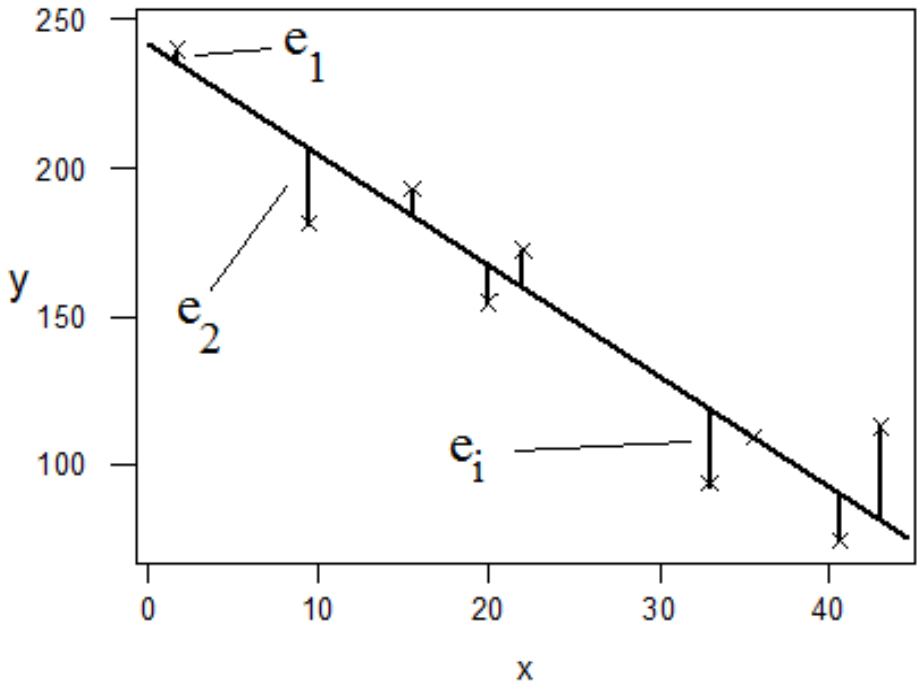
What do we mean by a line being a ‘good fit’?

Equation of straight line is $y = a + bx$

Simple model for data:

$$y_i = a + b\,x_i + e_i$$

Straight line Random error



Simplest assumption: $e_i \sim N(0, \sigma^2)$ for all i, and e_i's are independent

- Linear regression model

Model is $y_i = a + b\,x_i + e_i$

Want to estimate parameters *a* and *b*, using the data.

e.g. - choose a and b to minimize the errors

Maximum likelihood estimate = least -squares estimate

Minimize $$E = \sum_i e_i^2 = \sum_i (y_i - \hat{y}_i) = \sum_i (y_i - a - bx_i)^2$$

Data point

Straight-line prediction

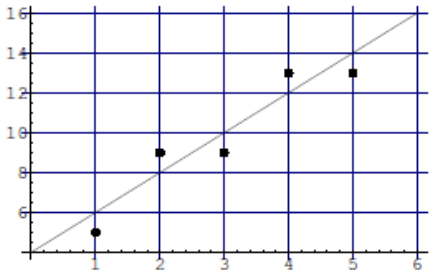
E is defined and can be minimized even when errors not Normal
– least-squares is simple general prescription for fitting a straight line

(but statistical interpretation in general less clear)

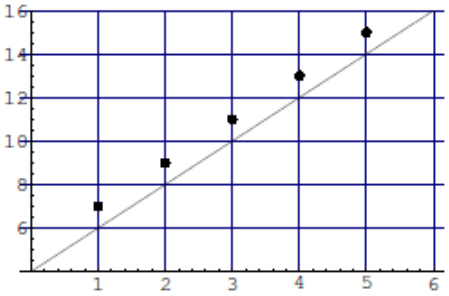
The line $y = 4 + 2x$ has been proposed as a line of best fit for the following four sets of data. For which data set is this line the best fit (minimum $E = \sum_i e_i^2$)?

Question from Derek Bruff

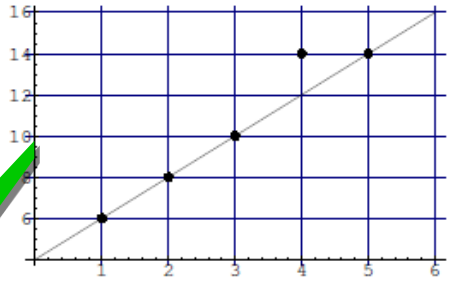
1.



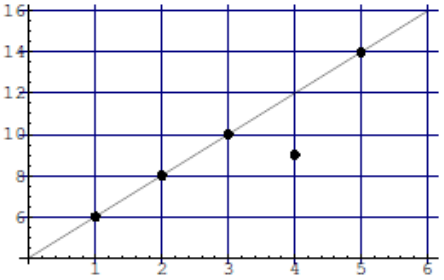
2.

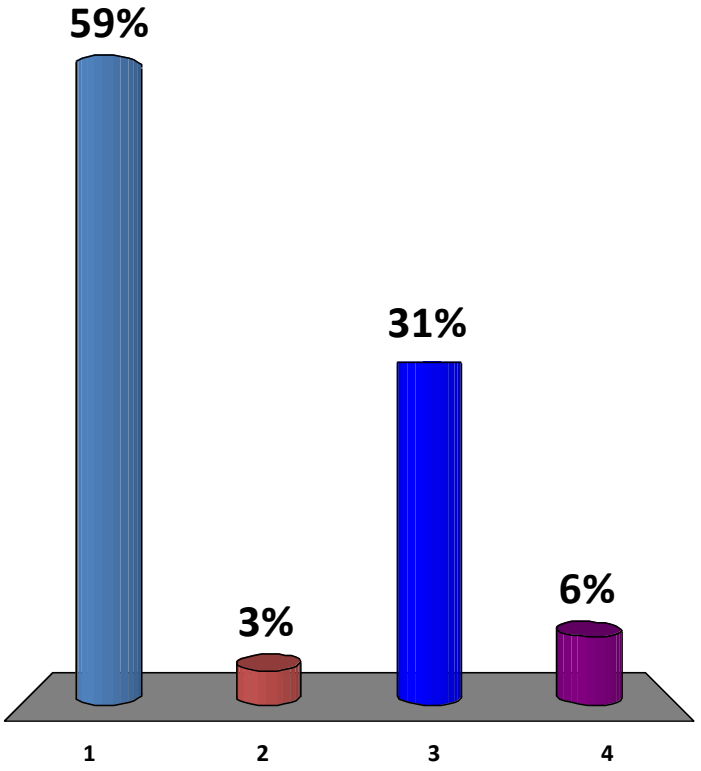


3.



4.





How to find a and b that minimize $E = \sum_i e_i^2 = \sum_i (y_i - a - bx_i)^2$?

For minimum want $\frac{\partial E}{\partial a} = 0$ and $\frac{\partial E}{\partial b} = 0$, see notes for derivation

Solution is the least-squares estimates $\hat{a}$ and $\hat{b}$:

$$\hat{b} = \frac{S_{xy}}{S_{xx}} \text{ and } \hat{a} = \bar{y} - \hat{b}\,\bar{x}$$

Sample means

Where

$$S_{xx} = \sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n} = \sum_i (x_i - \bar{x})^2$$

$$S_{xy} = \sum_i x_i y_i - \frac{\sum_i x_i \sum_i y_i}{n} = \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

Equation of the fitted line is $\hat{y} = \hat{a} + \hat{b}x$

Most of the things you need to use are on the formula sheet

STATISTICAL FORMULAE

1. For X having the Binomial distribution, B(n, p): $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$,
for $k = 0, 1, 2, \ldots, n$ and X has mean np and variance $np(1\text{-}p)$.

2. For Y having the Poisson distribution, parameter λ: $P(Y = k) = \frac{e^{-\lambda}\lambda^k}{k!}$,
for $k = 0, 1, 2, \ldots$ and Y has mean λ and variance λ.

3. For the sample $x_1, x_2, \ldots, x_n$, the sample mean is $\bar{x} = \sum_{i=1}^{n} x_i / n$
and the sample variance is $s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 / (n-1) = \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right) / (n-1)$.

4. Observations $x_1, x_2, \ldots, x_m$ occur with frequencies $f_1, f_2, \ldots, f_m$, the sample mean is $\bar{x} = \frac{\sum_{i=1}^{m} f_i x_i}{n}$ and the sample variance is $s^2 = \frac{\sum_{i=1}^{m} f_i (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{m} f_i x_i^2 - n\bar{x}^2}{n-1}$.

5. Random sample $X_1, X_2, \ldots, X_n$ from a N(μ, σ^2) distribution, then $\frac{\bar{X} - \mu}{\sqrt{\sigma^2 / n}}$ has a N(0, 1) distribution and $\frac{\bar{X} - \mu}{\sqrt{s^2 / n}}$ has a t_{n-1} distribution.

6. B(n, p) is approximated by N(np, $np(1\text{-}p)$), when n is large and np is not too close to 0 or n. Poisson, λ, is approximated by N(λ, λ), when λ is large.

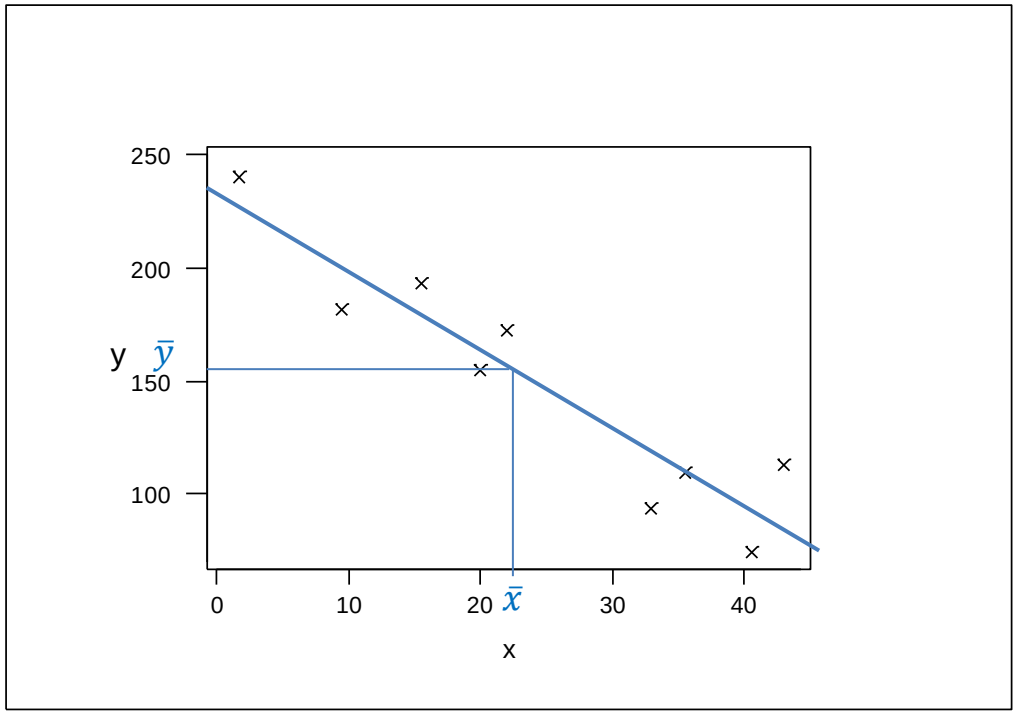
7. The linear regression line is estimated by $y = \hat{a} + \hat{b}x$ where $\hat{b} = S_{xy} / S_{xx}$,
$\hat{a} = \bar{y} - \hat{b}\bar{x}$, $\hat{\sigma}^2 = \frac{S_{yy} - \hat{b} S_{xy}}{n-2}$, $S_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$, $S_{yy} = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}$ and
$S_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$. $\frac{\hat{b} - b}{\sqrt{\hat{\sigma}^2 / S_{xx}}}$ has a t_{n-2} distribution. The mean value of y at x_0, $a + bx_0$, has confidence interval $\hat{a} + \hat{b}x_0 \pm t_{n-2} \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}$.
A confidence interval for a single response at x_0 is
$\hat{a} + \hat{b}x_0 \pm t_{n-2} \sqrt{\hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}$.
The (Pearson product-moment) sample correlation is $r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$.

Note that since $\hat{a} = \bar{y} - \hat{b}\,\bar{x}$

$$\begin{aligned}\hat{y} &= \hat{a} + \hat{b}x \\ &= \bar{y} - \hat{b}\bar{x} + \hat{b}x\end{aligned}$$

$$\Rightarrow \hat{y} - \bar{y} = \hat{b}(x - \bar{x})$$

i.e. $(\bar{x}, \bar{y})$ is on the line



***Example*:**

The data *y* has been observed for various values of *x*, as follows:

y	240	181	193	155	172	110	113	75	94
x	1.6	9.4	15.5	20.0	22.0	35.5	43.0	40.5	33.0

Fit the simple linear regression model using least squares.

$$\hat{b} = \frac{S_{xy}}{S_{xx}} \text{ and } \hat{a} = \bar{y} - \hat{b}\,\bar{x}$$

$$S_{xx} = \sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n}$$

$$S_{xy} = \sum_i x_i y_i - \frac{\sum_i x_i \sum_i y_i}{n}$$

***Answer*:**

Want to fit $\hat{y} = \hat{a} + \hat{b}x$

$n = 9$

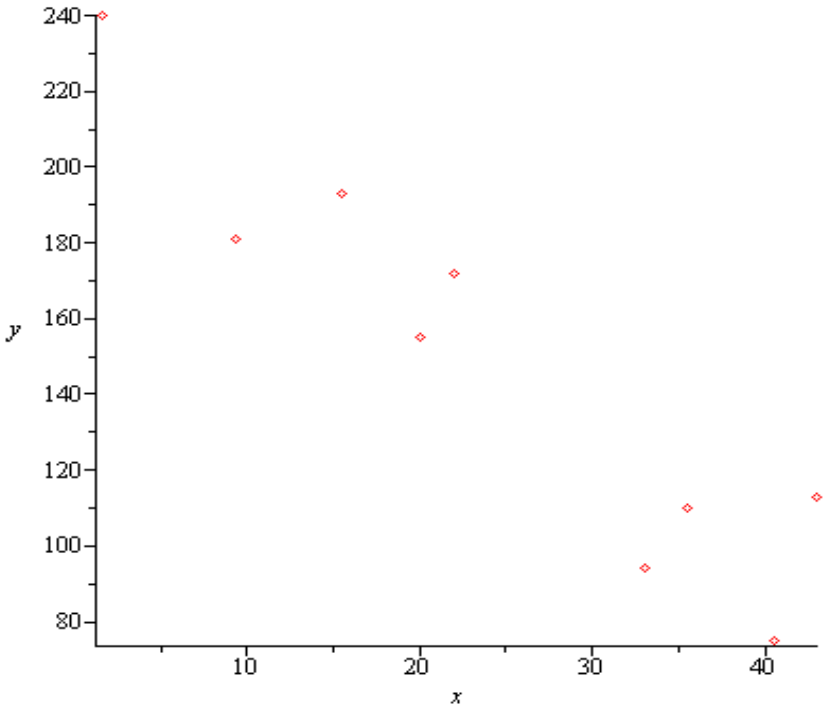
$\sum_i x_i = 220.5\,,\ \sum_i y_i = 1333.0$

$\sum_i x_i^2 = 7053.7, \quad \sum_i y_i^2 = 220549, \quad \sum_i x_i y_i = 26864$

$$S_{xx} = 7053.7 - \frac{220.5^2}{9} = 1651.42$$

$$S_{xy} = 26864 - \frac{220.50 \times 1333.0}{9} = -5794.1$$

$$\Rightarrow \hat{b} = \frac{S_{xy}}{S_{xx}} = -\frac{5794.5}{1651.45} = -3.5086$$



***Answer*:**

$$\hat{b} = \frac{S_{xy}}{S_{xx}} \text{ and } \hat{a} = \bar{y} - \hat{b}\,\bar{x}$$

$$S_{xx} = \sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n}$$

$$S_{xy} = \sum_i x_i y_i - \frac{\sum_i x_i \sum_i y_i}{n}$$

Want to fit $\hat{y} = \hat{a} + \hat{b}x$

$n = 9$

$\sum_i x_i = 220.5\ ,\ \sum_i y_i = 1333.0$

$\sum_i x_i^2 = 7053.7, \quad \sum_i y_i^2 = 220549, \quad \sum_i x_i y_i = 26864$

$$S_{xx} = 7053.7 - \frac{220.5^2}{9} = 1651.42$$

$$S_{xy} = 26864 - \frac{220.50 \times 1333.0}{9} = -5794.1$$

$$\Rightarrow \hat{b} = \frac{S_{xy}}{S_{xx}} = -\frac{5794.5}{1651.45} = -3.5086$$

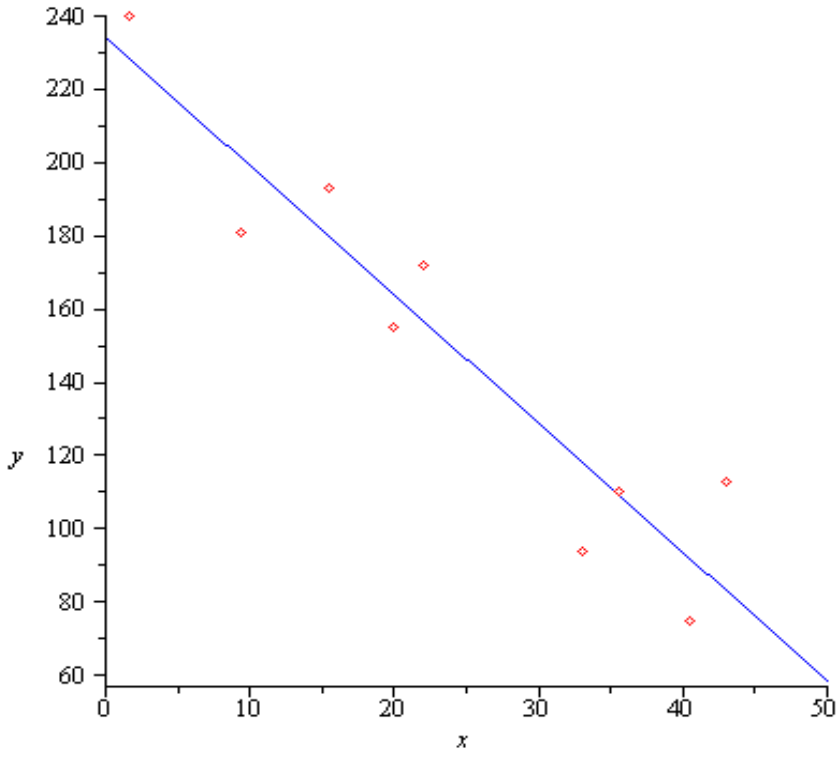
Now just need $\hat{a}$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

$$= \frac{1333.0}{9} - (-3.5086) \times \frac{(220.50)}{9} = 234.1$$

So the fit is approximately

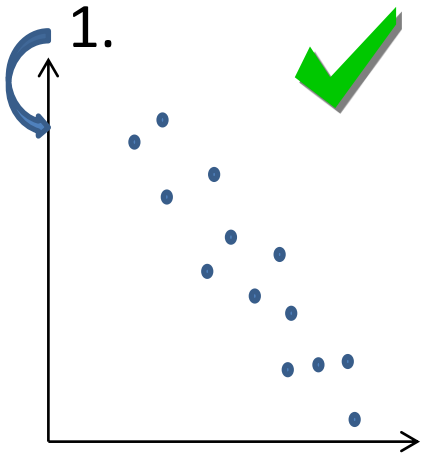
$$y = 234.1 - 3.509x$$



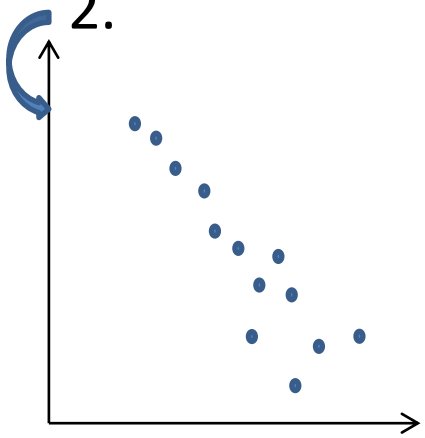
Q

Which of the following data are likely to be most appropriately modelled using a linear regression model?

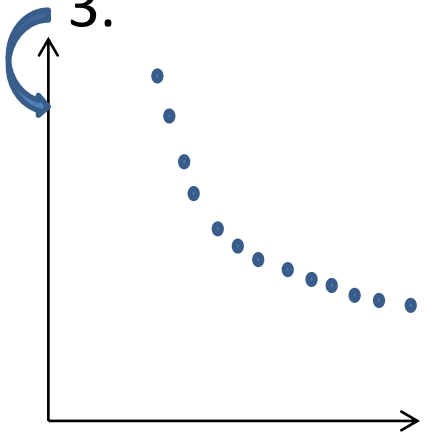
1.

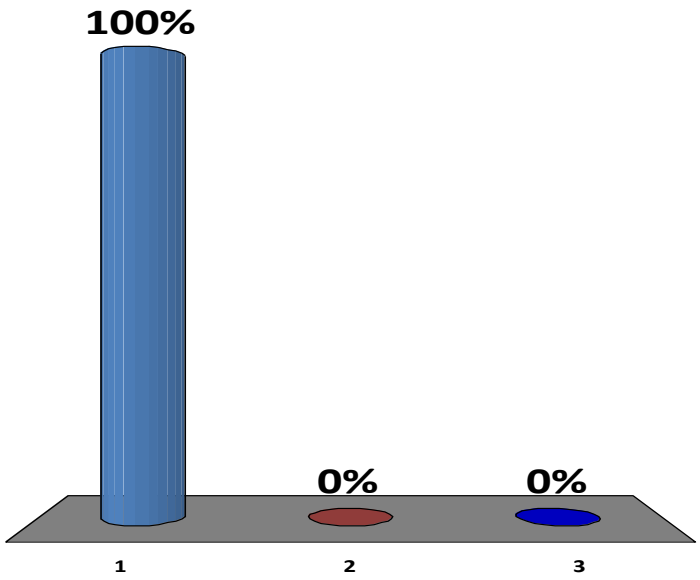


2.



3.

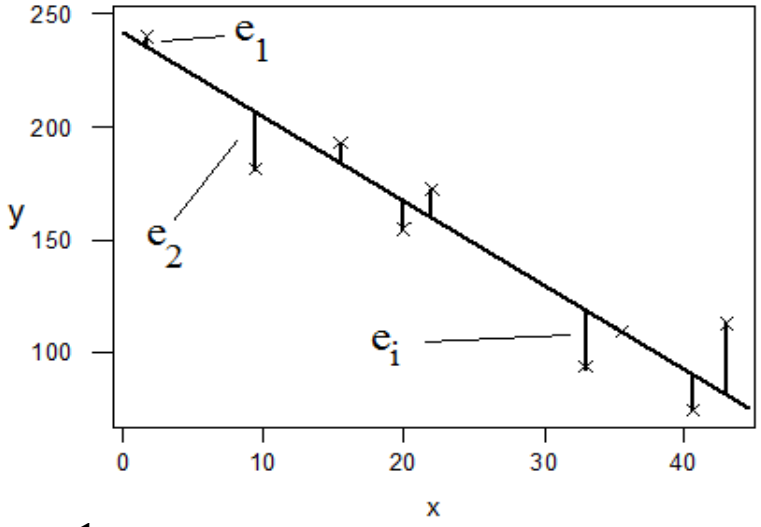




Quantifying the goodness of the fit

Estimating σ^2: variance of *y* about the fitted line

Estimated error is: $\hat{e}_i = y_i - \hat{y}_i$



$\mu_e = 0$, so the ordinary sample variance of the e_i's is $\sim \frac{1}{n-1}\sum_i \hat{e}_i^2$

In fact, this is biased since two parameters, *a* and *b* have been estimated. The unbiased estimate is:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum \hat{e}_i^2 = \frac{1}{n-2}\sum (y_i - \hat{y}_i)^2$$

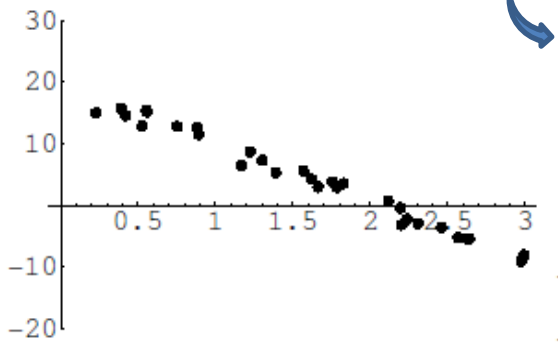
$$= \frac{S_{yy} - \hat{b}S_{xy}}{n-2}$$ *[derivation in notes]*

Residual sum of squares

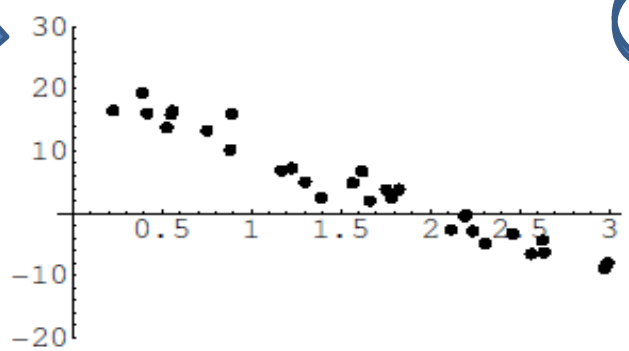
Which of the following plots would have the greatest residual sum of squares [variance of y about the fitted line]?

Question from Derek Bruff

1.



2.



3.