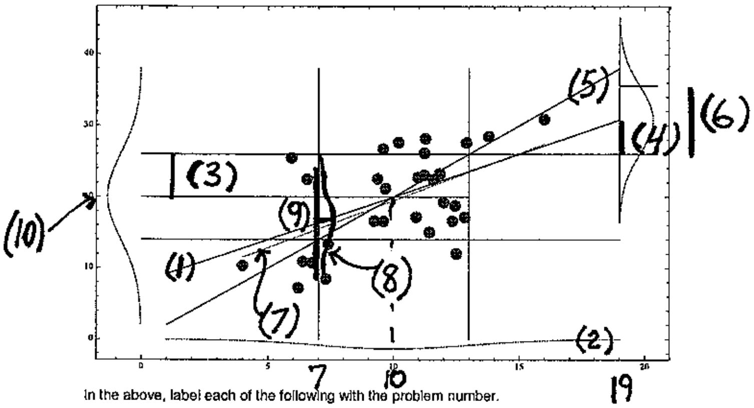
Exam 4 Prep

1-10. **Normal Population.** The scatter plot show below is a random sample from a 2D normal population. The bell curves and dark lines refer to the population. The sample Least Squares Line (shorter) is in red.

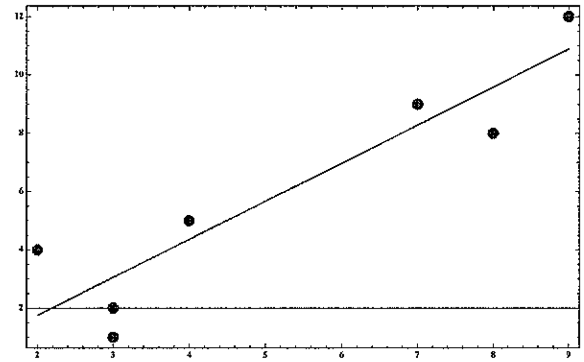


In the above, label each of the following with the problem number.

1. The population Least Squares line.
2. The population distribution of x.
3. The population sd of y (and indicate by a thick line segment).
4. The sd of all population y whose x-score is 19.

5. The population SD line.
6. A 68% y-interval for population points with x = 19.
7. The sample Least Squares Line.
8. Sketch in place the distribution of population y with x = 7.
9. The 95% prediction interval for population y with x = 7.
10. Use the sample Least Squares Line to read-off the regression-based estimator of μ_y **if it is known that** $\mu_x = 10$.

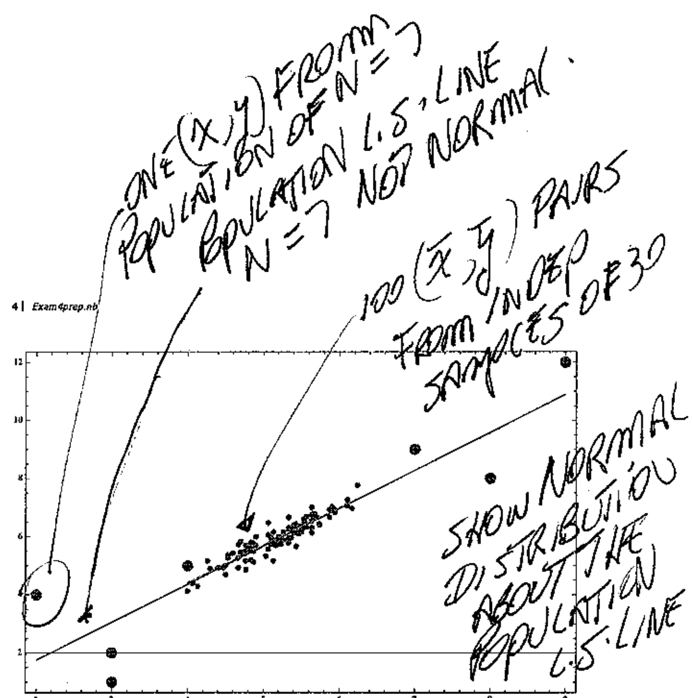
11-20. **Non-normal Population.** A population need not be normal in order for Least Squares to be useful. The plot below shows a decidedly non-normal population of N = 7 points together with its population Least Squares Line.

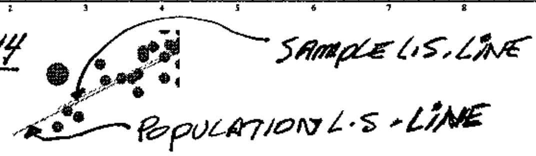


x	y	x^2	y^2	xy
2	4	4	16	8
4	5	16	25	20
3	2	9	4	6
7	9	49	81	63
8	8	64	64	64
3	1	9	1	3
9	12	81	144	108
—	—	—	—	—
5.14286	5.85714	33.1429	47.8571	38.8571

From the table determine the following. You should check the results using your calculator's built-in statistical routines.

11. μ_x 5.14286
12. μ_y 5.85714
13. σ_x $\sqrt{33.1429-5.14286^2}$
14. σ_y $\sqrt{47.8571-5.85714^2}$
15. ρ $\dfrac{38.8571-5.14286\ 5.85714}{\sigma_x\ \sigma_y}$
16. Use these to re-plot the Least Squares Line in the above.
17. σ_{3x-2} (properties of population sd in relation to location or scale changes).
    $3\sigma_x$
18. $\rho_{3x-2,9y+4}$ (properties of population correlation in relation to location or scale changes).
    ρ
19. The plot of the population is not at all normal, and the strength of regression is not great. Nor would the plot of any sample from this population look normal. However, the plot of 100 pairs $(\bar{x}, \bar{y})$, each from a with-replacement sample of $n = 30$, **is nearly normal**. The importance of this is that various issues surrounding the joint variation of $(\bar{x}, \bar{y})$ are subject to normal theory even though the population is not normal.



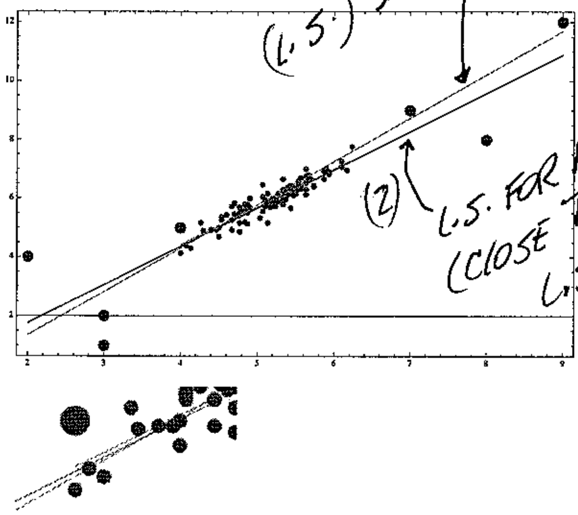


The sample least squares line for 100 pairs $(\bar{x}, \bar{y})$ is shown in red (see enlargement).

Moreover, almost any sample of $n = 30$ from the population can give us a good idea of the population Least Squares Line. Below, I've overlaid the above plot with one such sample of 30 (there are only 7 points in the population so the sample of 30 falls entirely on these 7 but with unequal numbers because of random sampling). Owing to the unequal representation of the 7 population points in the sample of 30 the sample Least Squares Line does not fall so perfectly on the population Least Squares Line as has the sample L.S. for 100 pairs $(\bar{x}, \bar{y})$.

ALSO, THE SAMPLE L.S. LINE HAS APPROXIMATED THE L.S. LINE OF THE "POPULATION" OF ALL $(\bar{x}, \bar{y})$ FROM SAMPLES OF $n = 30$ (ONLY 100 $(\bar{x}, \bar{y})$ SHOWN).

RANDOM WITH REPLACEMENT SAMPLE OF $n = 30$ FROM A DECIDEDLY NON-NORMAL POPULATION OF N = 7 FINDS SAMPLE L.S. LINE ~ POPULATION L.S.



(1, 5) SAMPLE (1)

(2) L.S. FOR 100 $(\bar{x}, \bar{y})$ (CLOSE TO POPULATION L.S.)

Identify (1) the sample L.S. line from n = 30 pairs (x, y) in plot and enlargement above.

Identify (2) the sample L.S. line from n = 100 pairs $(\bar{x}, \bar{y})$ in plot and enlargement above.

20. Which are correct?

A random sample of large n from a population is likely to produce a least squares line close to the population least squares line.

The variation of sample values (x, y) from a population is universally going to exhibit normal looking plots of the points (x, y) provided n is large.

A random sample of large n of random sample pairs $(\bar{x}, \bar{y})$ is likely to show plot consistent with normal variation even if the underlying population is not normal.

21-22. Sampling Distribution slope b_1 of sample regression line. This is keyed to #14 and #16 of page 748.

$$s_x = \sqrt{\sum (x-\bar{x})^2/(n-1)} = \frac{\sqrt{n}}{\sqrt{n-1}}\sqrt{\overline{x^2} - (\bar{x})^2}$$

$$s_e = \sqrt{\sum (y-\hat{y})^2/(n-2)} = \frac{\sqrt{n}}{\sqrt{n-2}}\sqrt{1-r^2}\,\sqrt{\overline{y^2} - (\bar{y})^2}$$

$$\text{SE}(b_1) = \frac{s_e}{\sqrt{n-1}\; s_x} \text{ estimate of SD of } b_1$$

$$\text{SO } \text{SE}(b_1) = \frac{\sqrt{1-r^2}\; s_y}{\sqrt{n-2}\; s_x} = \frac{\sqrt{1-r^2}}{r\sqrt{n-2}} b_1$$

SE(b_1) EXPRESSED IN TERMS OF r, n, b_1

This estimate of the standard deviation of b_1 is useful for CI and tests about the population slope β_1.

If the population is 2D normal then we have

CI for β_1: $b_1 \pm t_{df=n-2}\,\text{SE}(b_1)$ (exact)

Same as: $b_1 \pm t_{df=n-2} \frac{\sqrt{1-r^2}}{r\sqrt{n-2}} b_1$ (exact)

test statistic $\frac{b_1-\beta_1}{\text{SE}(b_1)}$ has exactly $t_{df=n-2}$ distribution

and if $H_0: \beta_1 = 0$ the test statistic is $\frac{b_1-\beta_1}{\text{SE}(b_1)} = \frac{\sqrt{1-r^2}}{r\sqrt{n-2}}$

TEST STATISTIC FOR $H_0: \beta_1 = 0$ DEPENDS ONLY ON r, n.

For large n, **even if the population is not normal**

CI for β_1: $b_1 \pm z\,\text{SE}(b_1)$ (approximate)

test statistic $\frac{b_1-\beta_1}{\text{SE}(b_1)}$ has ~ Z distribution.

As applied to #14 and #16 pg. 748

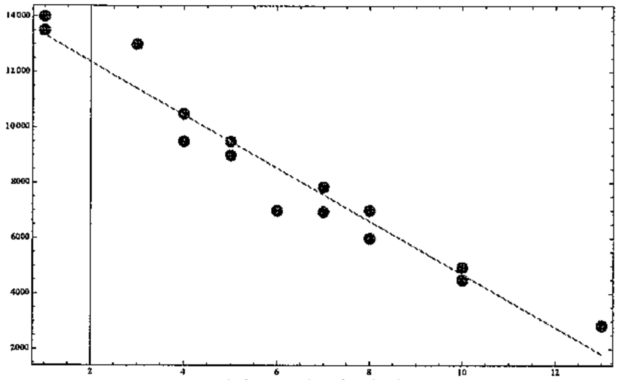
x	y	x^2	y^2	xy
1	13 990	1	195 720 100	13 990
1	13 495	1	182 115 025	13 495
3	12 999	9	168 974 001	38 997
4	9500	16	90 250 000	38 000
4	10 495	16	110 145 025	41 980
5	8995	25	80 910 025	44 975
5	9494	25	90 136 036	47 470
6	6999	36	48 986 001	41 994
7	6950	49	48 302 500	48 650
7	7850	49	61 622 500	54 950
8	6999	64	48 986 001	55 992
8	5995	64	35 940 025	47 960
10	4950	100	24 502 500	49 500
10	4495	100	20 205 025	44 950
13	2850	169	8 122 500	37 050
–	–	–	–	–
6.13333	8403.73	48.2667	8.09945×10^7	41 330.2

$\bar{x} = 6.1333$ $\bar{y} = 8403.73$

$\hat{\sigma}_x = \sqrt{48.2667 - 6.1333^2}$

$\hat{\sigma}_y = \sqrt{80994500 - 8403.73^2}$

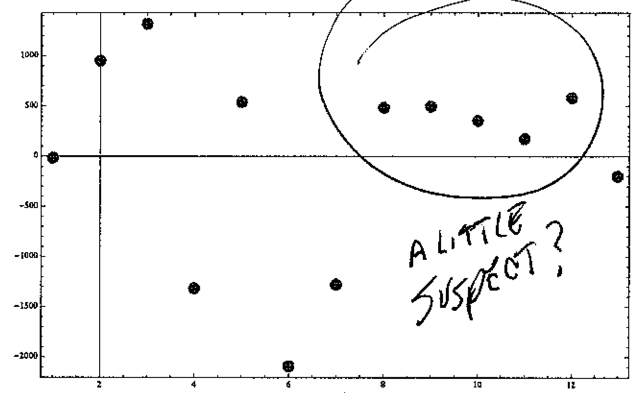
$r = \dfrac{41330.2 - 6.1333 \; 8403.73}{\hat{\sigma}_x \; \hat{\sigma}_y}$



x = age (yr), y = advertised price (dollars)

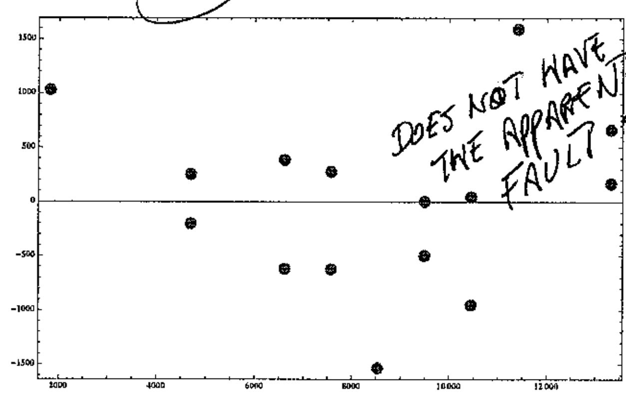
n = 15 pairs (x, y)
means (6.1333, 8403.73)
s_x = 3.3778 s_y = 3333.56
r = -0.971767 t_{13} = 2.16 for 95%
b_1 = -959.039
SE(b_1) = 64.5816 (applicable df = 15-2 = 13)
95%CI = -959.039 + {-1, 1} (2.16) (64.5816)
= {-1098.54, -819.543}

Plot of **residuals vs. x = year**:



A LITTLE SUSPECT?

Plot of **residuals vs. predicted** values as favored by the textbook:

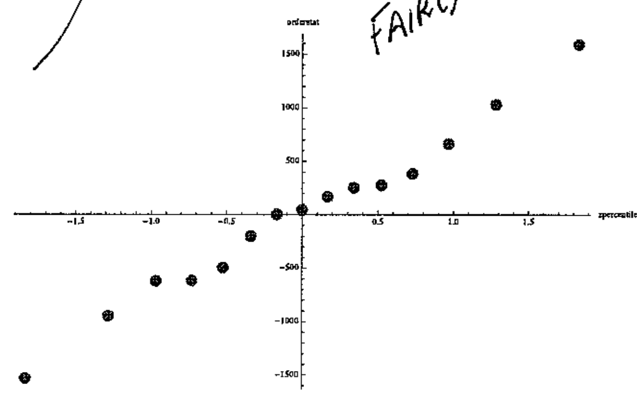


DOES NOT HAVE THE APPARENT FAULT SHOWN ABOVE

PREDICTED

Normal Probability Plot of residuals.

FAIRLY LINEAR



? USE THE t-CI FOR β_1
WITH CAUTION.