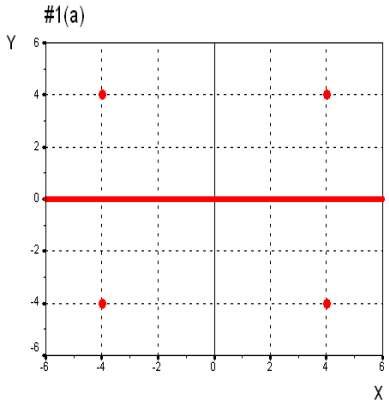
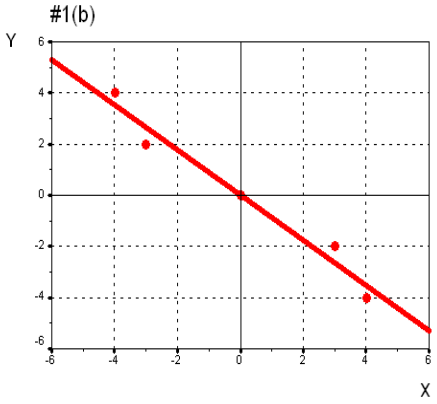
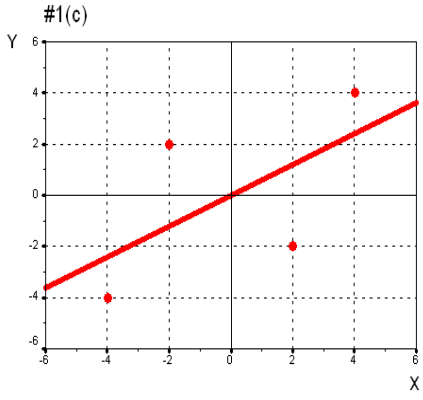
REGRESSION AND CORRELATION: ANSWERS & DISCUSSION

1. Here are scattergrams (drawn by SPSS: Graphs => Legacy Dialogs => Scattergram => Simple):

(a)







	(a)	(b)	(c)
Mean (X)	0	0	0
Mean (Y)	0	0	0
SD (X)	4	$\sqrt{10}$	$\sqrt{10}$
SD (Y)	4	$\sqrt{8}$	$\sqrt{10}$
r	.000	−.984	+.600
r^2	.000	.968	.360
b	.000	−0.880	+0.600
a	0	0	0

Here are calculations (on a worksheet identical to that at the end of the Correlation and Regression handout) for (b) only:

Case #	X	Y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
1	4	−4	+4	−4	16	16	−16
2	3	−2	+3	−2	9	4	−6
3	0	0	0	0	0	0	0
4	−3	2	−3	+2	9	4	−6
5	−4	4	−4	4	16	16	−16
Sum	0	0	0	0	50	40	−44
Av.	0	0			10	8	−8.8
	$(\bar{x})$	$(\bar{y})$			Var(X)	Var(Y)	Cov(X,Y)
√av.					3.16	2.83	
					SD(X)	SD(Y)	

$$\text{Correlation coefficient} = r = \frac{\text{Cov}(X,Y)}{\text{SD}(X)\text{SD}(Y)} = \frac{-8.8}{(3.16)(2.83)} = -0.9845$$

$$r^2 = 0.968$$

$$\text{Regression coefficient (Y dep.)} = b = \frac{\text{Cov}(X,Y)}{\text{Var}(X)} = \frac{-8.8}{10} = -0.88$$

$$\text{Intercept (constant)} = a = \bar{y} - b\,\bar{x} = 0 - (-0.88) \times 0 = 0$$

$$\text{Regression equation: } \hat{y} = a + bx = (-.88)x$$

2. (d) There's no dispersion in the dependent variable, so all points lie on a horizontal line. The covariance of X and Y is zero because the all y-deviations are zero. The regression slope is zero divided by ⅔ (the variance of X), so $b = 0$, which makes sense. However, the correlation coefficient is zero divided by zero [$SD(x) \times 0 = 0$], so the correlation is undefined. ***Lesson: correlation can exist only if the dependent variable actually varies***.

(e) There's no dispersion in the independent variable, so all points lie on a vertical line. The covariance of X and Y is zero, because all x-deviations are zero. Both the regression slope and the correlation coefficient are zero divided by zero [$Var(X) = 0$ and $0 \times SD(y) = 0$, respectively). ***Lesson: regression slope and correlation can exist only if the independent variable actually varies***.

(f) $b = 1.5$, $a = 3$, and $r = +1$. You should see that, if there are only two cases (two plotted points), the calculated correlation must be perfect (i.e., either +1 or −1), provided both variables vary (i.e., provided we avoid the problems entailed by (d) and (e) above). Put graphically, we can always draw a straight line that perfectly fits (by the least squares or any other criterion) two points. ***Lesson: correlation is meaningful only given three or more cases***. We can generalize this consideration.

A statistic based on a sample of size $n = 1$ provides some information for estimating the population *average* with respect to some variable. For example, if the value of the one sampled case is "big" (or "small"), that's a bit of evidence that the average value in the population is "big" (or "small"). However, a sample of size $n = 1$ provides *no information whatsoever for estimating the population dispersion* with respect to any variable, because the sample SD is zero no matter what case constitutes the sample of $n = 1$.

A statistic based on a sample of size $n = 2$ provides some information for estimating the population *dispersion* with respect to some variable. For example, if the values of the two sampled case are "far apart" (or "close together"), that's a bit of evidence that the dispersion in the population is "big" (or "small"). However, a sample of size $n = 2$ provides *no information whatsoever for estimating the magnitude of the population correlation coefficient* between two variables (beyond whether the correlation is positive or negative), because the sample correlation is +1 or −1 (or possibly undefined) no matter what two cases constitutes a sample of $n = 2$.

3. MPH is the independent variable X; MPG is the dependent variable Y. Here is the worksheet:

Case #	X	Y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
1	20	24	−20	−2.8	400	7.84	+56
2	30	28	−10	+1.2	100	1.44	−12
3	40	30	0	+3.2	0	10.24	0
4	50	28	+10	+1.2	100	1.44	+12
5	60	24	+20	−2.8	400	7.84	−56
Total	200	134	0	0	1000	28.8	0
Average	40	26.8			200	5.76	0

Since Cov $(X,Y) = 0$,

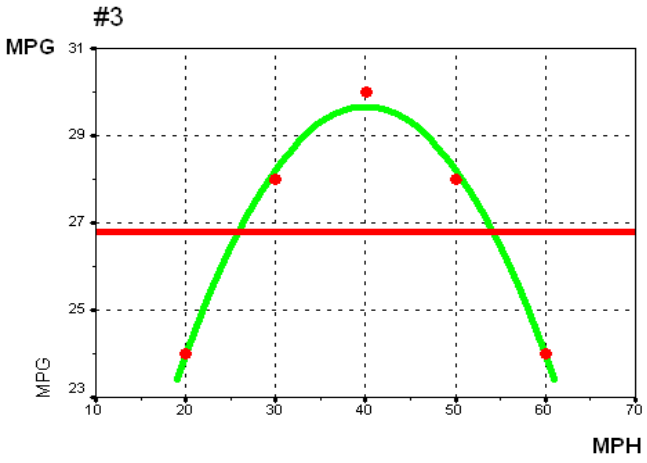
$r = 0$

$r^2 = 0$

$b = 0$

$a = \bar{y} = 26.8$

Here is the scattergram: =>



The scattergram shows a very nice relationship between *MPH* and *MPG*, but it is not a *linear* (straight line) relationship; rather it is a *curvilinear* (curved line) relationship. The correlation coefficient measures the closeness of points to the best fitting *straight* line, and the best fitting straight line here is horizontal ($y = \bar{y} = 26.8$), indicating no (linear) relationship. ***Lesson: don't be satisfied with just calculating regression and correlation coefficients; also to look at scattergrams to see obvious patterns in or features of the data that may not show up in the standard summary statistics.***

4. (a) years
   (b) seconds
   (c) none [correlation is a pure number, which would not change even if we changed the age units and/or reaction time units]
   (d) years

(e) years squared

(f) seconds (of reaction time) per year (of age), i.e., the regression coefficient answers this question: for each additional year of age of the subjects, how much on average does their reaction time (in seconds) increase [the coefficient presumably is negative — that is, reaction time presumably decreases with age]

5. The student newspaper said, in effect, that the correlation between teaching ability and research productivity is *negative*, whereas Professor McDaniel said merely that the correlation is *about zero*. In plain language, productive researchers are neither more nor less likely to be effective teachers than unproductive researchers and vice versa.

6. (a) The term *correlation* usually applies to association (measured by the correlation coefficient r) between two *interval* variables. Since the variable GENDER is dichotomous and the nominal/ordinal/interval distinction applies only to non-dichotomous variables, typically some measure of association other than r would be used. In any event, simply reporting an association between GENDER and INCOME would tell your audience that men and women have (on average) different levels of income but would not indicate which group has the higher (average) income. But trying to specify the direction of the association as positive (as is done here) or negative is a blunder, because GENDER is not measured on a LO-HI scale like INCOME. Probably the best way to report your finding would be to report the mean income of male workers and of female workers; the magnitude of this "gender gap" with respect to income is essentially the regression coefficient and answers this question: how big a difference does gender (independent variable) make for what people can expect to earn (dependent variable)? The magnitude of the association (or correlation) between these two variables answers (in effect) this question: how big is this "between-the-two- genders" gap compared with the "within-each-gender" gaps, i.e., how big is the income gap between men and women compared with the income dispersion *among men* and also *among women*.

(b) No correctly calculated correlation coefficient (or other measure of association) can be greater than +1 (or less than –1).

(c) The value of a correlation coefficient is a pure number. The correlation between age and income is just $r = +0.53$, not +0.53 years (or dollars or any other units — and it would be the same if we measured people's ages in months, their income in Euros, etc.).

7. (a) a substantial negative correlation (the older a car, the less its value)

(b) a substantial negative correlation (the heavier car, the lower its MPG)

(c) a modest to substantial positive correlation (per our favorite scattergram; we can actually estimate this correlation coefficient quite precisely; see #8 (c) below)

(d) a modest positive correlation (the taller a man, the greater his weight [on average and obviously with many exceptions, so the positive r would probably be closer to 0 than to 1]

(e) presumably (almost) zero correlation

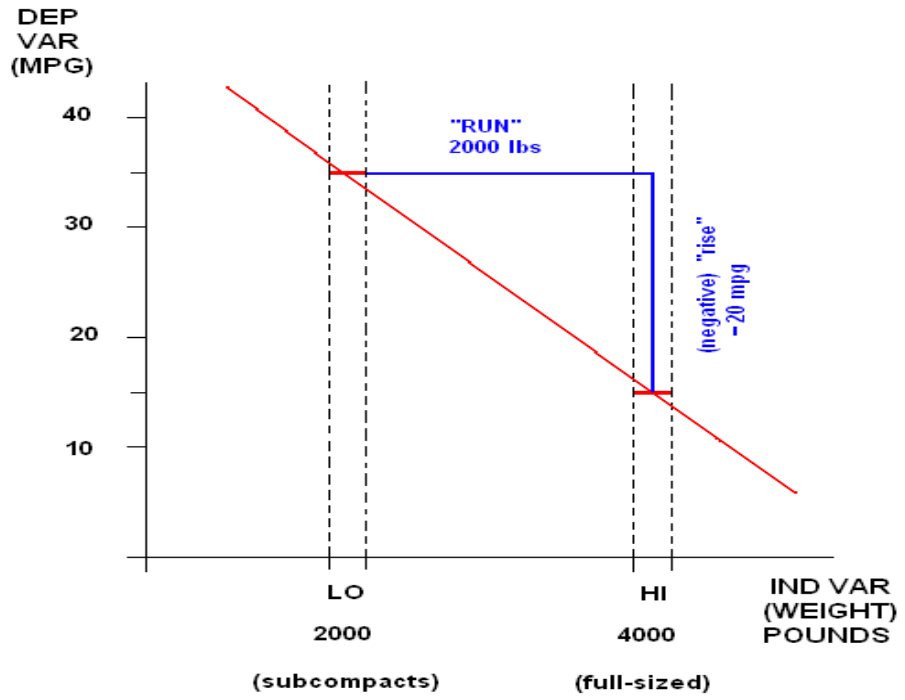
8. I find that students can typically make reasonable estimates of the magnitude and (especially) the direction of association between variables (such as in the previous question) but find it difficult to make a reasonable guess at the numerical magnitude of a regression coefficient to be a difficult task. However, if you understand the question that the regression coefficient is answering and have a little contextual knowledge (e.g., about cars), it possible to make good "ball-park" estimates. In particular, remember that the regression coefficient answers this question: ***how much, on average, does the dependent variable increase when the independent variable increases by one unit***. Clearly, *the magnitude of the regression coefficient depends on (i) which variable is considered to be independent and which dependent and also on (i) the units in which each variable is measured.* Then (1) think in terms of the two "vertical strips" (such as we have used in examining scattergrams, in particular SON'S HEIGHT by FATHER'S HEIGHT) corresponding to some LOW value and some HIGH value of the independent variable, (2) make a reasonable guess at the average value of the dependent variable for cases within each of these strips, (3) connect these two estimated averages to form a line of averages (effectively a regression line), and finally (4) determine its slope from the "rise over run" triangle. See the hand-drawn figure on the next page (that pertains to MPG by WEIGHT).

   (a) Clearly AGE is the independent variable, influencing the dependent variable PRICE. As noted above, the negative correlation means the older a car, the less its value. The regression coefficient is also negative and answers this further question: "*How much* (on average) does PRICE decline with AGE?" To give any numerical answer to this question, we must specify the *units* in which we are expressing the two variables. It is natural to measure AGE in *years* and PRICE in *dollars*. Then the question becomes: *on the average, how many dollars do cars lose in value for each year they get older*? A plausible, though *very* approximate, answer might something on the order of \$3000, in which case $b \approx -3000$ dollars (dependent variable units) per year (independent variable units). One way in which this answer is approximate is that new cars depreciate more rapidly and older cars less rapidly. (Furthermore, price is a ratio variable, and a car's value can never fall below zero. In addition, cars old enough to qualify as "classics" or antiques begin to appreciate in value.) Thus, the relationship between the age and values of cars is truly a *curvilinear*, not linear, relationship. But the relationship is approximately linear (and quite strongly negative) for the first half dozen years or so of age.

   (b) Clearly WEIGHT influences MPG. As noted above, the negative correlation means the heavier car, the lower its MPG. The regression coefficient answers this question: "*How much* (on average) does additional weight 'cost' in terms of MPG?" To answer this question we must specify in what units we are expressing the two variables. Let us (initially) measure WEIGHT in pounds; the units for MPG are already specified. The question then becomes: *what is the average "penalty," in terms of MPG, paid for each additional pound of vehicle weight*? Here is the basis for a rough guess: a really big car weighs about 4000 pounds and gets about 15 MPG, a subcompact car weighs about 2000 pounds and gets about 35 MPG; so b is something like $(15 - 35) = -20$ MPG (negative "rise") divided by $(4000 - 2000) = 2000$ pounds ("run") or about -0.01 MPG per pound. A probably more comprehensible way of stating this is to express WEIGHT in terms of *hundreds of pounds*, which makes b is about -1. This translates into the very

comprehensible and useful (if you're in the market for a new car) conclusion that *each additional hundred pounds of vehicle weight imposes an average penalty of about 1 MPG.*

Consider the question in the framework of a scattergram without plotted points:



$$\text{regression slope} = b = \frac{-20 \text{ mpg}}{2000 \text{ lbs}} = -.01 \text{ mpg per lb} = -1 \text{ mpg per 100 lbs}$$

Substantive conclusion: each addition 100 lbs in vehicle weight costs you about 1 MPG

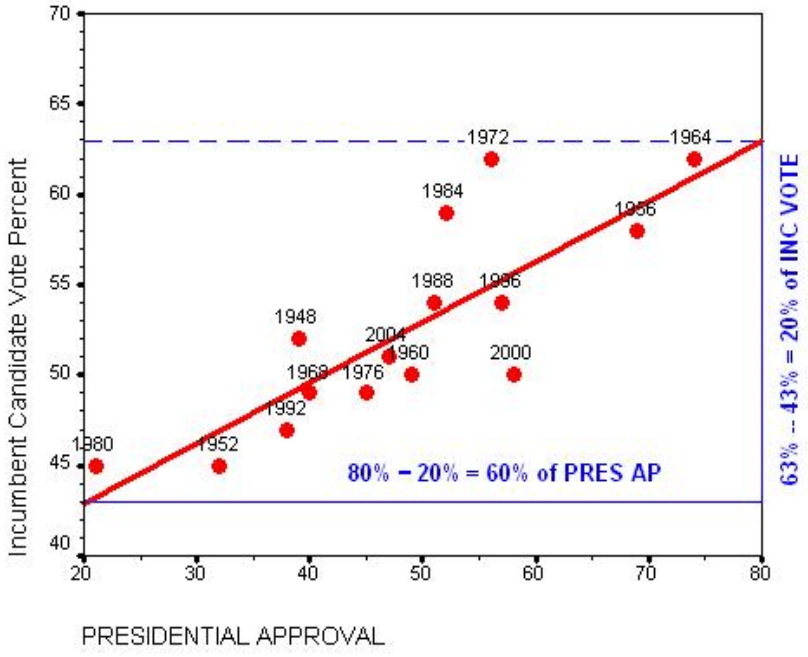
(c) In the xeroxed scattergram of fathers' and sons' heights (where both heights are measured in the same units, i.e., inches), we saw (by estimating the mean height of sons in several vertical strips, "connecting the dots" to form the line of averages and measuring its slope) that b is about $+0.5$. Note that the independent and dependent variables in this data have just about the same SD, which implies that b and r are just about equal (a point that you can confirm by checking their respective formulas), so we can guess rather precisely that r is about +0.5.

(d) Suppose we consider HEIGHT to be independent and WEIGHT to be dependent (though this is rather arbitrary). As noted, the positive correlation means that, on average, the taller people are, the greater their weight. The regression coefficient answers this question: "*How much* (on average) more do people weight as height increases?" Let us express height in inches and weight in pounds. *On average* (around which there is a lot of dispersion, for clearly people of the same height may vary greatly in weight), we might

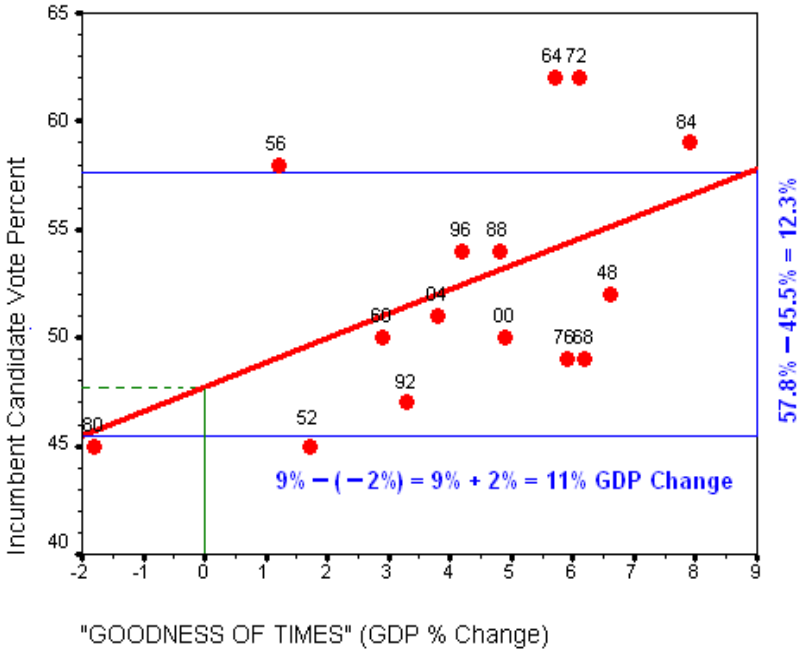
guess that people about 63" tall (5'3") probably weigh *on average* something like 130 lbs., and people around 75" tall (6'3") probably weigh *on average* about 210 lbs. So *b* is something like (210 – 130) = 70 pounds ("rise") divided by (75 – 63) = 12 inches ("run") or about +6 pounds per inch, i.e., on average, an increase of one inch in height is associated with an increase of about 6 pounds in weight.

(e) Since the correlation is zero (or almost zero), the covariance must be zero (or almost zero), so the regression coefficient is also zero (or almost zero), i.e., average IQ does not vary at all with height.

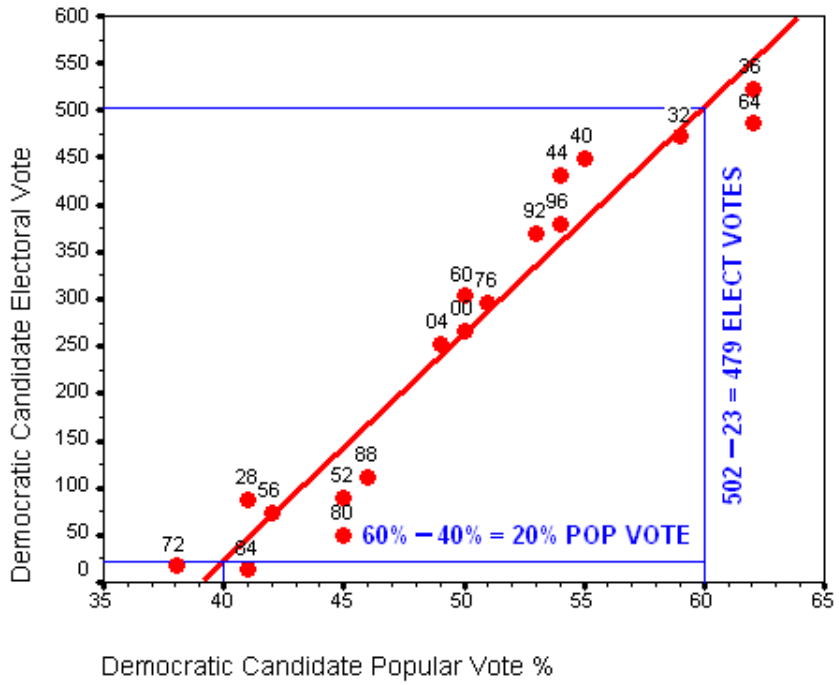
9. The regression lines in these scattergrams were calculated exactly by SPSS. You should be able to draw in approximately similar lines by "eyeball" methods and compute "rise over run" slopes that are approximately the same as those given here:



The regression coefficient *b* = "rise" (20% of INC VOTE) over "run" (60% of PRES AP) = +0.33. The substantive conclusion is that each additional 1% of PRES APPROVAL translates on average (around which there is modest dispersion, i.e., correlation is quite strong) into a gain of one-third of a percentage point in the popular vote for the incumbent candidate.



The regression coefficient b = "rise" (12.3% of INC VOTE) over "run" (11% of GDP%) = +1.12. The substantive conclusion is that each additional 1% of GDP% translates on average (around which there is considerable dispersion, i.e., the correlation is modest) into a gain of 1.12% in the popular vote for the incumbent candidate.



The regression coefficient b = "rise" (479 ELECT VOTES) over "run" (20% of POP VOTE) = +24. The substantive conclusion is that each additional 1% of popular votes translates on average (around which there is very little dispersion, i.e., the correlation is very high) into 24 additional electoral votes. (This conclusion applies to Republican candidates as well as Democratic, because everything has been on a two-party basis. If the dependent variable were *percent* of electoral votes, instead of *actual number* of electoral votes, then we would have $b = +4.5$, since 24/538 $\simeq$ 4.5%.)

The intercept for INC VOTE by GDP can be read directly off the graph and it answers a perfectly sensible question: what percent of the popular vote do we expect a Presidential candidate of the incumbent party to get when GPD growth is 0%?

Remember that the intercept for INC VOTE by PRES AP is not 43%, since the vertical axis shown in the graph does not correspond to PRES AP = 0% but rather PRES AP = 20%. Since INC VOTE falls by .33% for each 1% fall in PRES AP, the intercept is 43% $-0.33\times20\%$ = 43% – 6.67% = 36.33%. It answers the perhaps not very sensible question: what percent of the popular vote do we expect a candidate of the incumbent party to get when the President's approval rating is 0%.

In like manner, we can calculate the intercept for ELECT VOTE by POP VOTE to be −938. Substantively, this is saying that we would expect that a candidate who wins 0% of the popular votes would get (on average) −938 electoral votes. Obviously, this conclusion is substantively absurd. It comes about because we are treating as *linear* a relationship which cannot be linear over the full range of the independent variable (because a party's electoral vote is a ratio variable and cannot fall below zero). We have data for a only narrow range of the independent variable (from about 38% to 62%), within which the relationship *is* very close to linear (as the scattergram shows). When we *blindly* project that (steeply sloping) straight line back to popular vote = 0%, we get the ridiculous answer of −938 votes.

Under *proportional representation*, each 1% increase in popular votes necessarily results in an increase in electoral votes of (just about) 1% (i.e., 5.38 electoral votes [which would in practice mean 5 or 6 electoral votes, electoral votes are discrete) . Thus we would have $b \approx +5.38$. (If the dependent variable were *percent* of electoral votes won, not *actual number* of electoral votes won, we would have $b \approx 1$. In either event, the relationship would be essentially perfect, i.e., $r \approx +1$, with intercept $a \approx 0$.)

10. On the scattergram, draw in the line DUKAKIS VOTE = MONDALE VOTE. (If the two axes had been drawn on the same scale, this would be a 45° line. But the SPSS Scattergram you were given compresses the vertical scale a bit relative to the horizontal scale, so the DV88 = DV84 line is not actually 45°. However, both scales run from 25 to 60, so you can draw the DV88 = DV84 line simply by putting a ruler on the southwest and northeast corners of the scattergram and drawing a line from one corner to the other.) For this line, $a = 0$ and $b = 1$. Any point on this line represents a state in which Mondale and Dukakis got the same percent of the vote; any point "southeast" of this line represents a state in which Mondale did better than Dukakis; and any point "northwest" of this line represents a state in which Dukakis did better than Mondale.

(a) It may be checked that almost every point in the scattergram lies "northwest" of the DUKAKIS = MONDALE line. Thus Dukakis did better than Mondale in almost every state.

(b) While no points lies clearly "southeast" of the DUKAKIS = MONDALE line, two appear to lie just about on the line. Thus there was no state in which Dukakis did clearly worse than Mondale and only two states in which they got (just about) the same vote.

For the record, these are the two points that appear to lie on the DUKAKIS = MONDALE:

	Mondale 1984	*Dukakis 1988*
Georgia	39.8022%	39.7479%
Tennessee	41.8182%	41.7814%

For the further record, Maryland is the state on the northwest side in which Dukakis improved the least over Mondale:

Maryland	47.2425%	48.5356%

(c) From the information given, the regression equation is:

DUKAKIS VOTE = 12.3 + 0.84 × MONDALE VOTE

Evaluate this expression for a LO and HI value of MONDALE VOTE, say 24 and = 50.

DUKAKIS VOTE = 12.3 + 0.84×24 = 12.3 + 20.2 = 32.5.
DUKAKIS VOTE = 12.3 + 0.84×50 = 12.3 + 45.4 = 54.3.

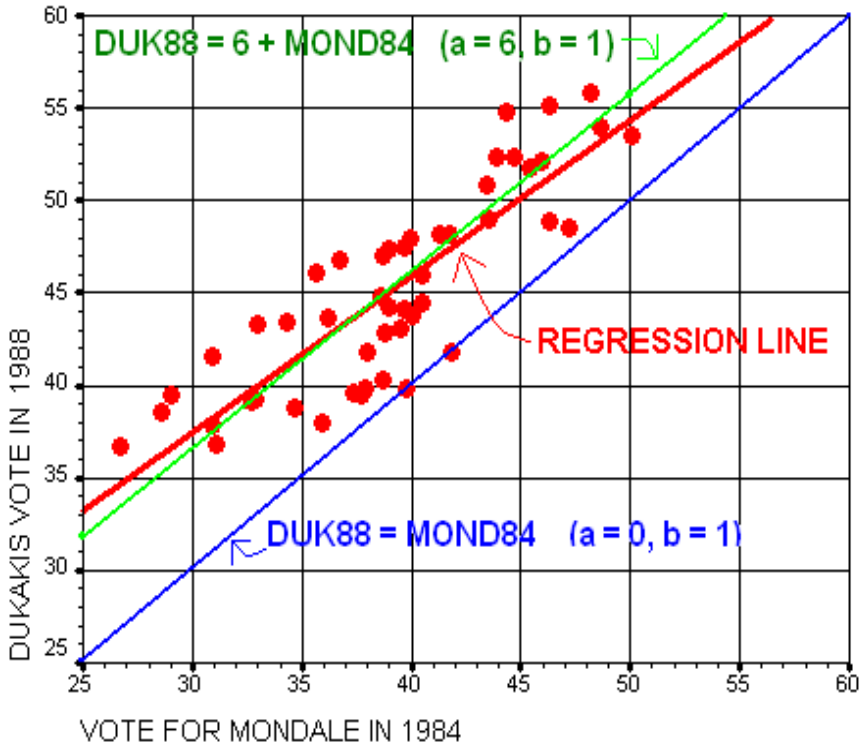
Now plot the two points (24%,32.5%) and (50%,54.3%). The straight line through these points is the regression line.

(c) (1) Given a uniform national swing, for every state in which Mondale got X % of the vote in 1984, Dukakis would have received $X + 6$% of the vote in 1988. All points would all lie exactly on the line with this equation:

DUKAKIS VOTE = 6 + MONDALE VOTE.

(2) From the equation above, $a = 6$ and $b = +1$. From the fact that all points would lie exactly on this line, it follows that $r = +1$.

(3) We know the swing was not a uniform 6% from the fact that the actual regression line is not DUKAKIS = 6 + MONDALE, and we know the swing was not uniform by any magnitude from the fact that r, though very high, is clearly not +1 as there is some "scatter" in the scattergram. (The actual correlation is $r = +0.87$.)



DUK88 = 12.3 + 0.84 x MOND84
DUK88 = 12.3 + 0.84 x 25 = 33.3
DUK88 = 12.3 + 0.84 x 55 = 58.5