Correlation and Regression

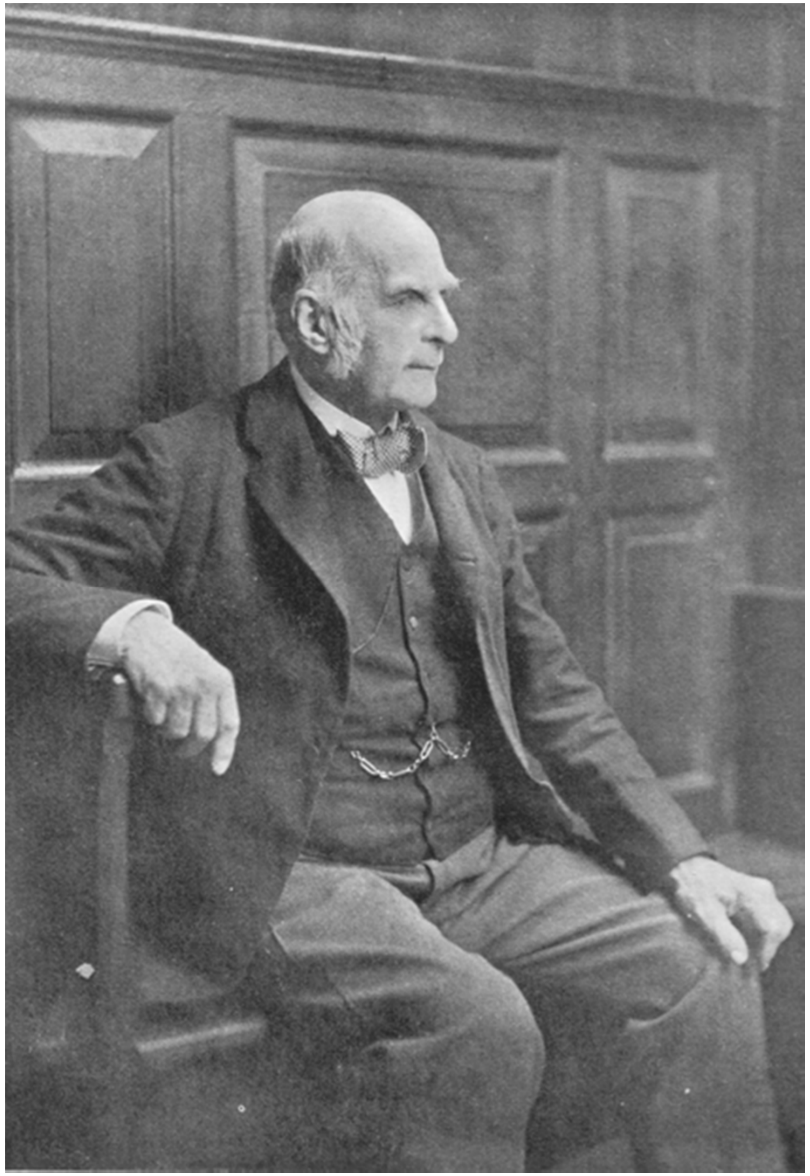
Sir Francis Galton

1822-1911

Geographer, meteorologist, tropical explorer, inventor of fingerprint identification, eugenicist, half-cousin of Charles Darwin and best-selling author

Galton

- Obsessed with measurement
- Tried to measure everything from the weather to female beauty
- Invented correlation and regression



Sweet Peas

- Galton’s experiment with sweet peas (1875) led to the development of initial concepts of linear regression.
- Sweet peas could self-fertilize: “daughter plants express genetic variations from mother plants without contribution from a second parent.”

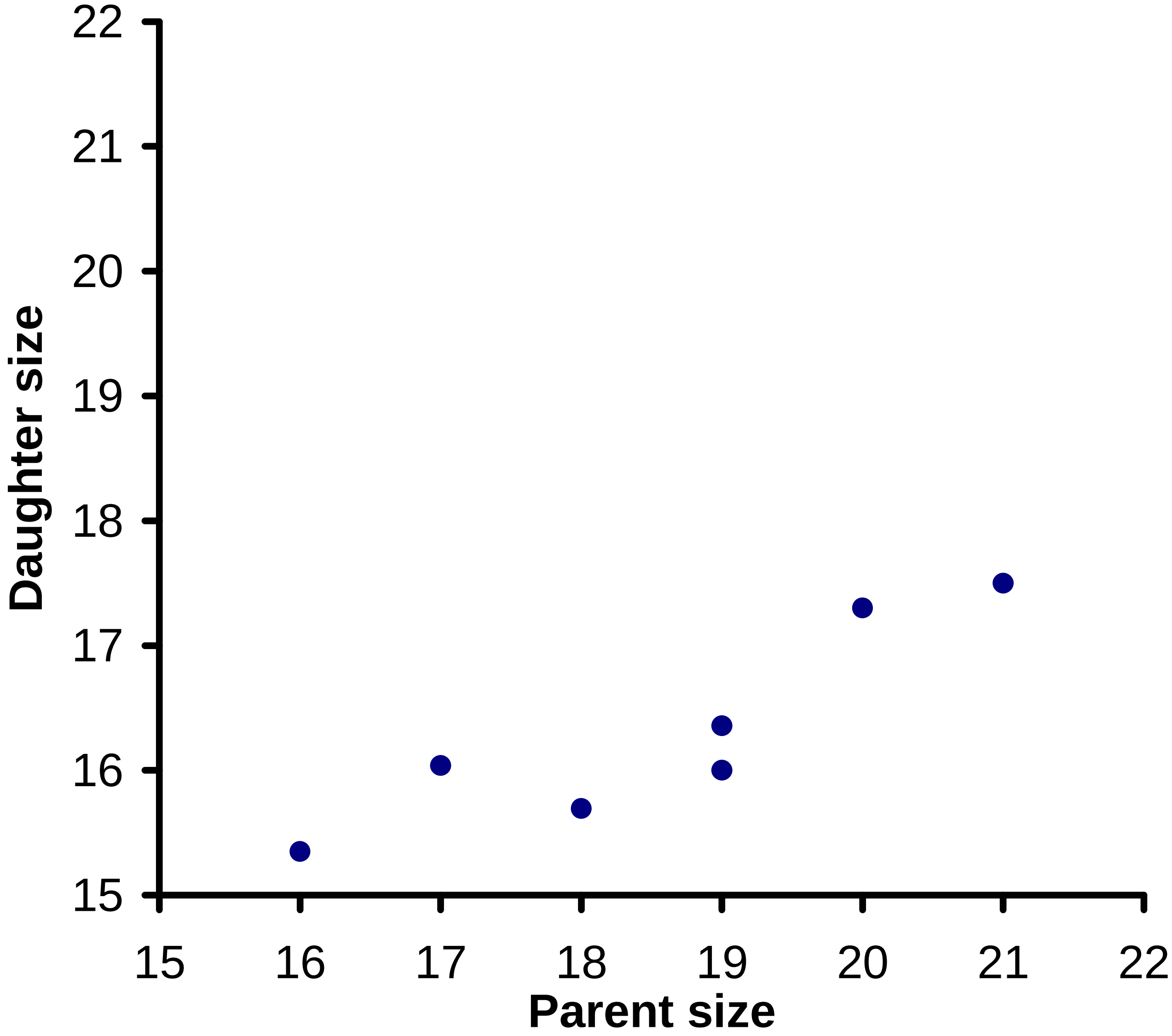
Sweet Peas (2)

- Distributed packets of seeds to 7 friends
- Uniformly distributed sizes, split into 7 size groups with 10 seeds per size.
- There was substantial variation among packets.
- 7 sizes 10 seeds per size 7 friends = 490 seeds
- Friends were to harvest seeds from the new generation of friends and return them to Galton.

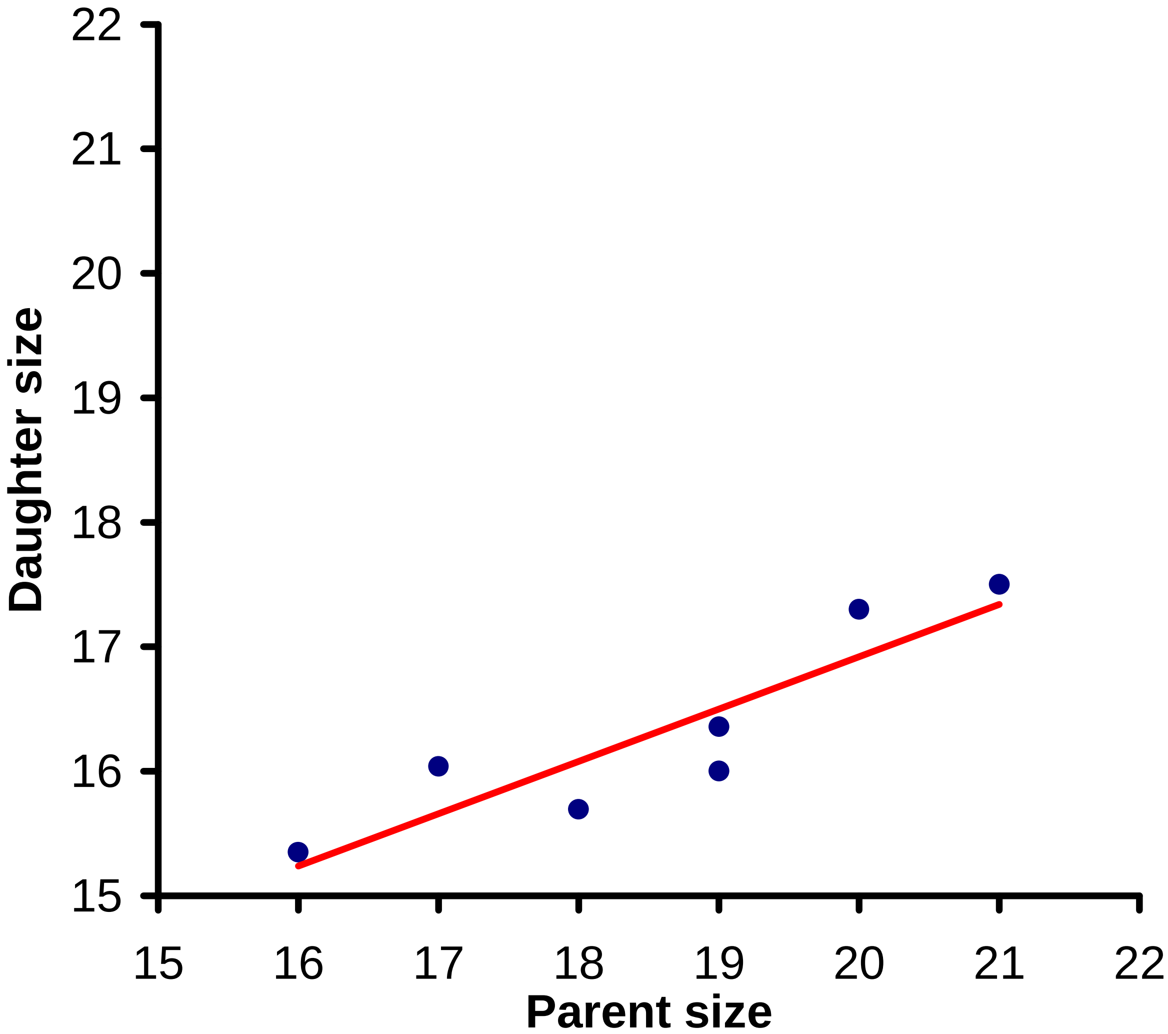
Sweet Peas (3)

- Plotted the weights of daughter seeds against weights of mother seeds.
- Hand fitted a line to the data
- Slope of the line connecting the means of different columns is equivalent to regression slope.

Sweet Pea Scatterplot



Sweet Pea Scatterplot with Regression Line



Slope indicates strength of association

Galton drew his line by hand, and estimated that for every thousandth of an inch of increased size of parents, the daughters size was affected by 0.33 thousandths

Things Galton Noticed

- Mother plants of a given seed size tended to have daughter seeds of pretty similar sizes
- Extremely large mother seeds grew into plants whose daughter seeds were **generally not so large**
- Extremely small mother seeds grew into plants whose daughter seeds were **generally not so small**
- Galton put a name on the loss of extremity:

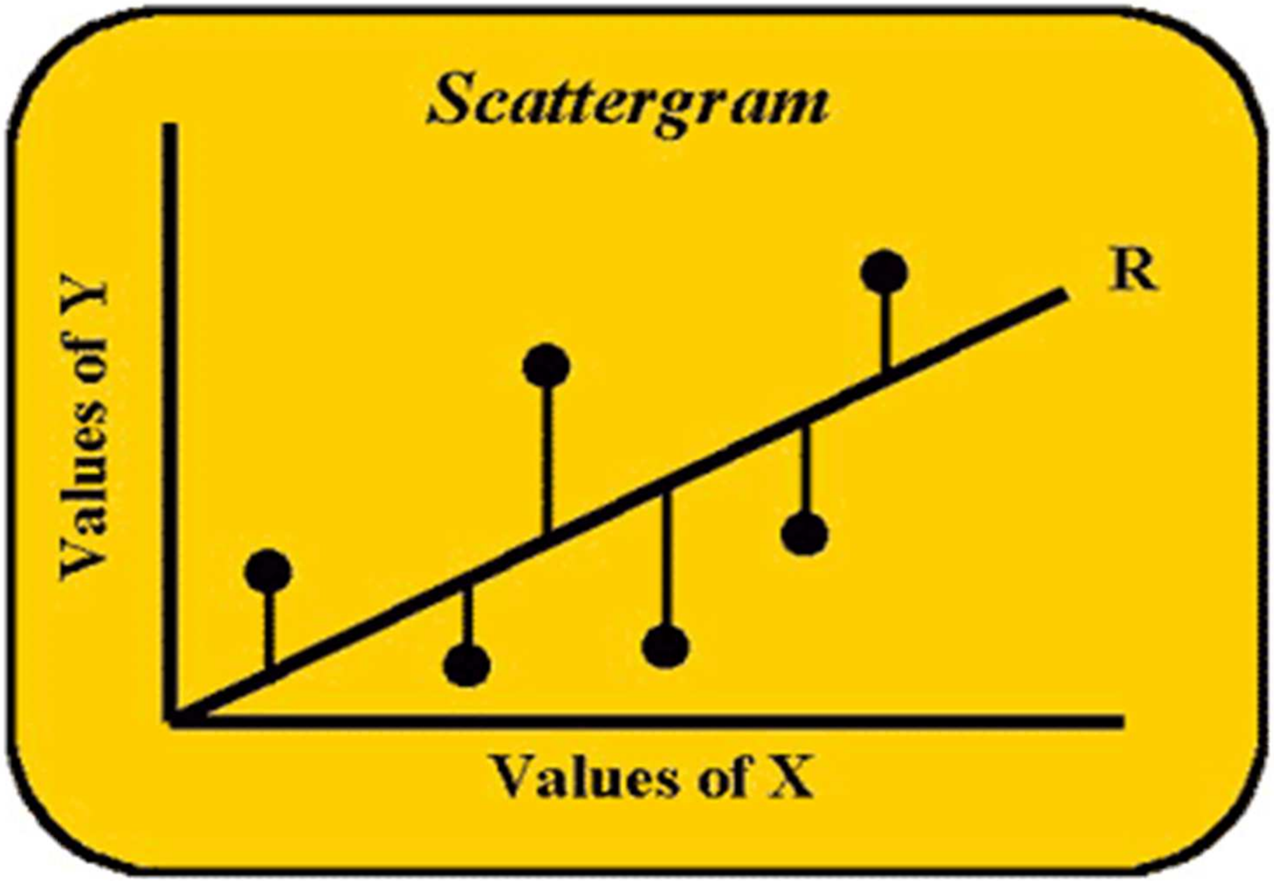
  "Regression to the mean"

Karl Pearson (1857-1936)

- Formalized Galton's method
- Invented least squares method of determining regression line

Pearson with Galton, c. 1900

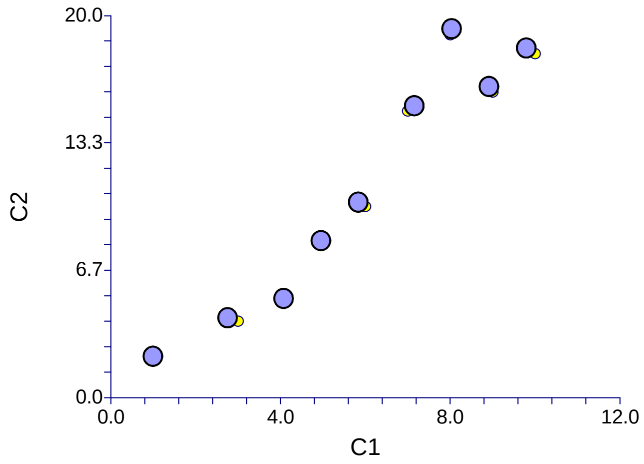
Least squares idea: Choose the line that minimizes the sum of the squares of the deviations of each observation from the regression line



Scatterplots

Correlations:

Positive

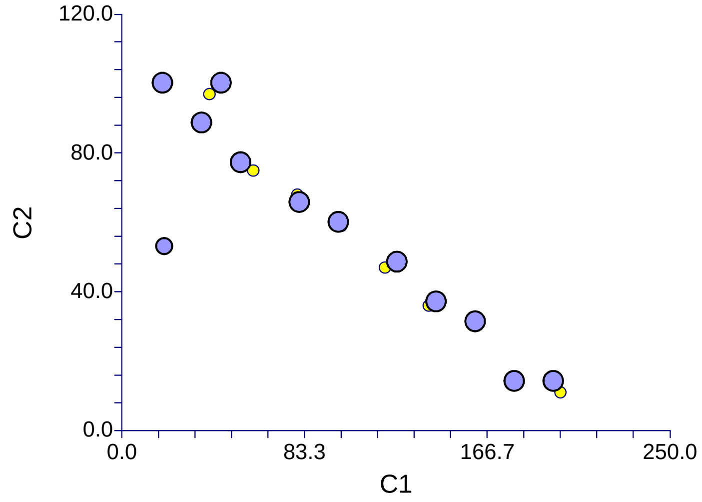


Large values of X associated with large values of Y,

small values of X associated with small values of Y

e.g. IQ and SAT

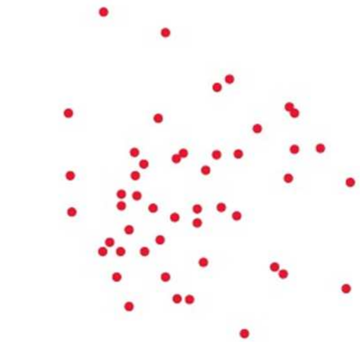
Negative



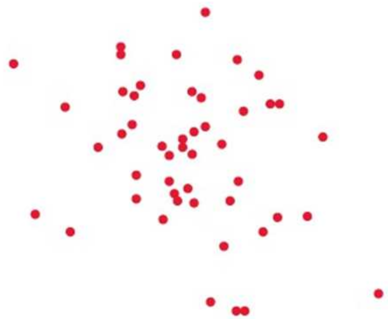
Large values of X associated with small values of Y

& vice versa

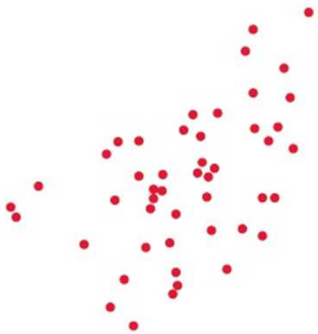
e.g. SPEED and ACCURACY



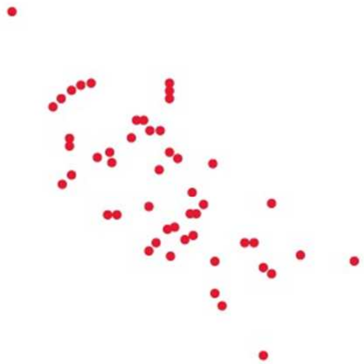
Correlation $r = 0$



Correlation $r = -0.3$

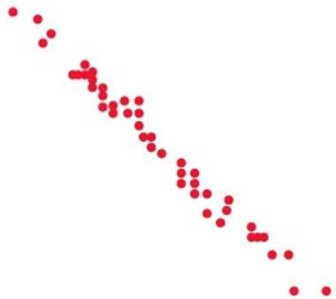


Correlation $r = 0.5$



Correlation $r = -0.7$

Correlation $r = 0.9$



Correlation $r = -0.99$

Correlation does not imply causality

- Two variables might be associated because they share a common cause
- For example, SAT scores and College Grade are highly associated, but probably not because scoring well on the SAT causes a student to get high grades in college
- Being a good student, etc., would be the common cause of the SATs and the grades

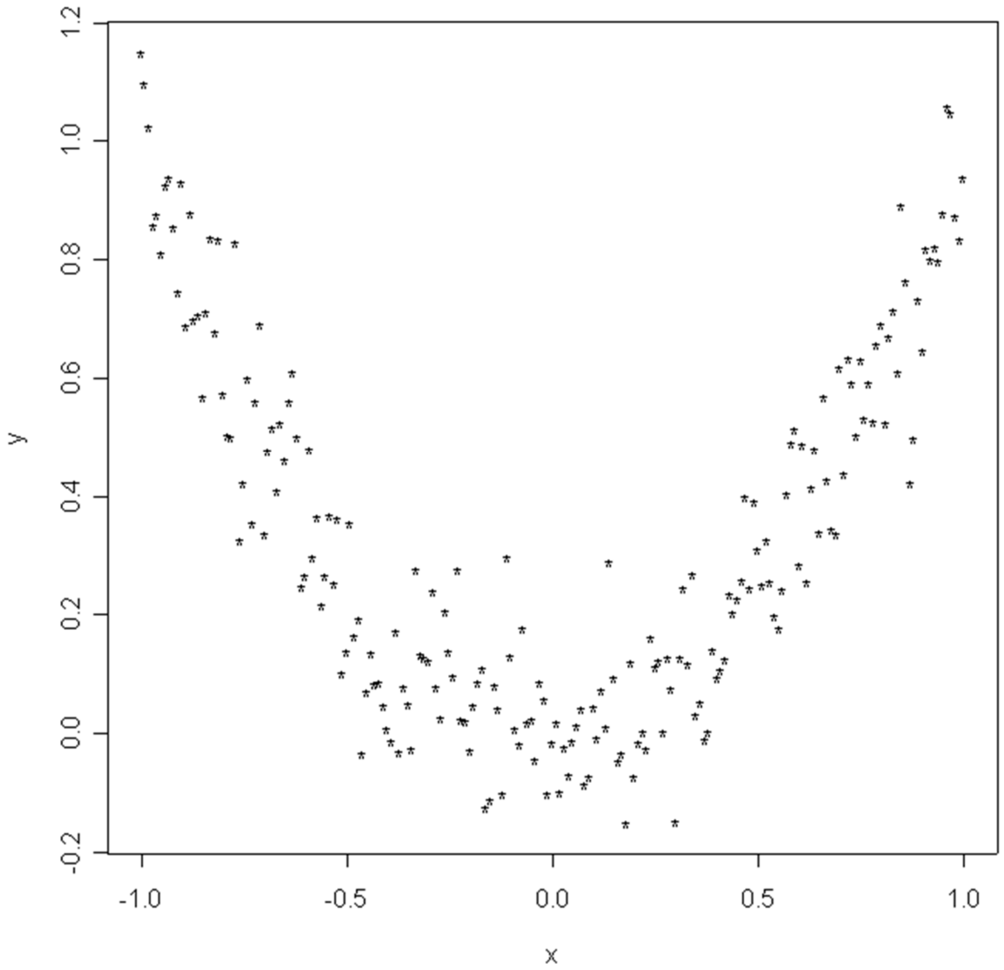
Intervening and confounding factors

There is a positive correlation between ice cream sales and drownings

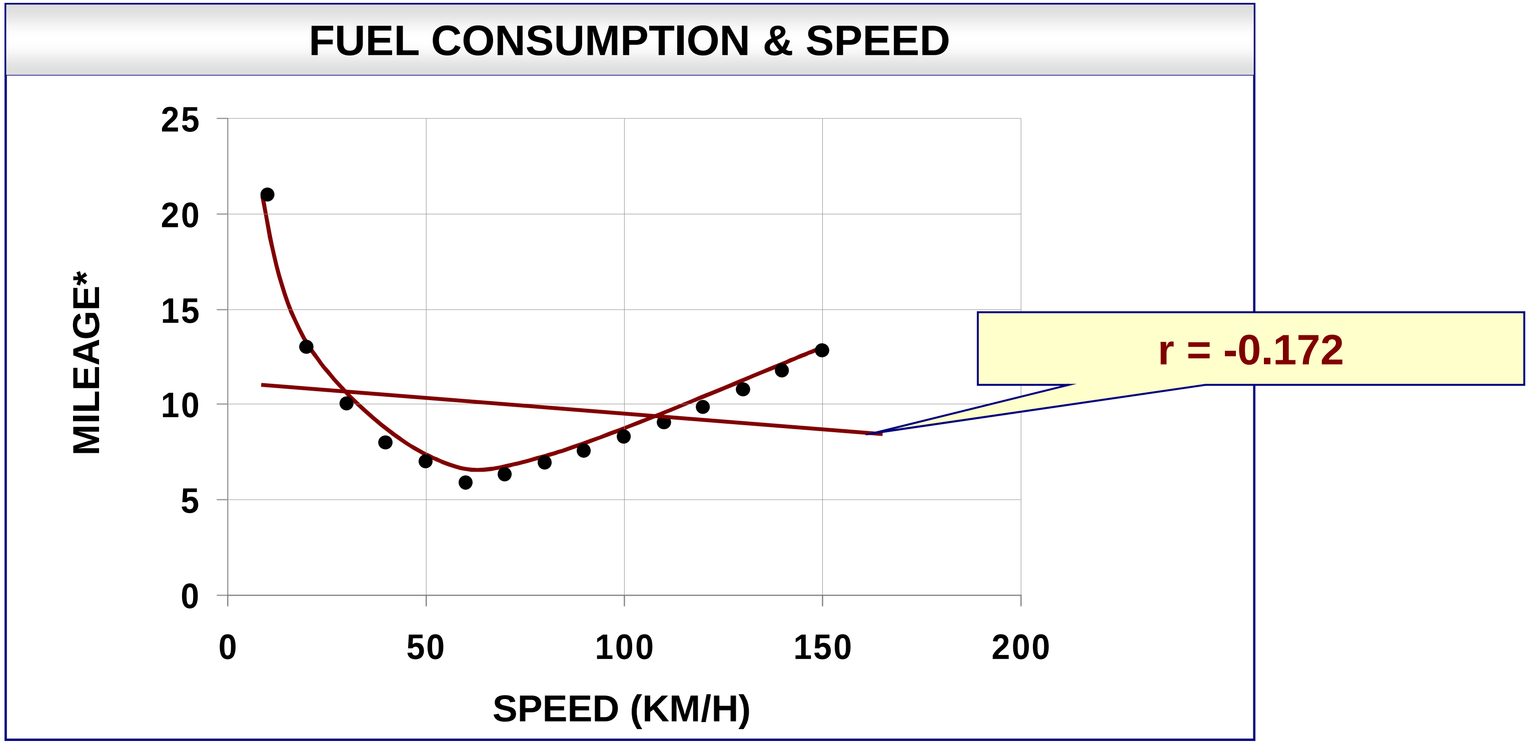
There is a strong positive association between Number of Years of Education and Annual Income

- In part, getting more education allows people to get better, higher-paying jobs.
- But these variables are *confounded* with others, such as socio-economic status

Regression and correlation will not capture nonlinear relationships



Speed And Mileage Have A Curved Relationship. Using r would be inappropriate



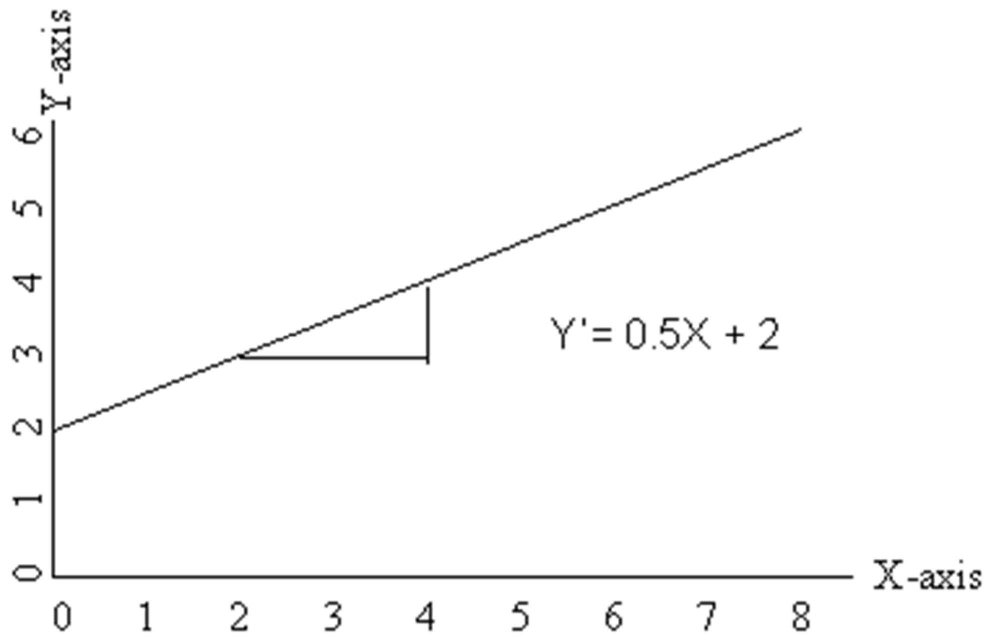
** Mileage: liters / 100 km*

Bivariate regression equation

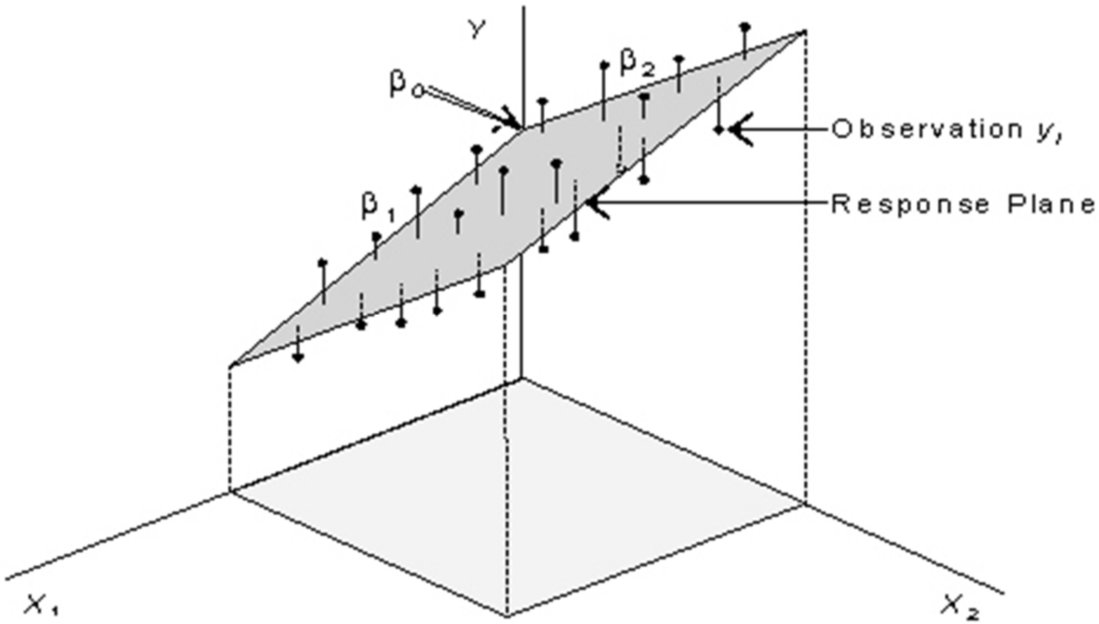
Y=bx+c

b=slope

c=intercept



Adding another variable



Multiple regression equation

$Y = b_1 x_1 + b_2 x_2 + c$

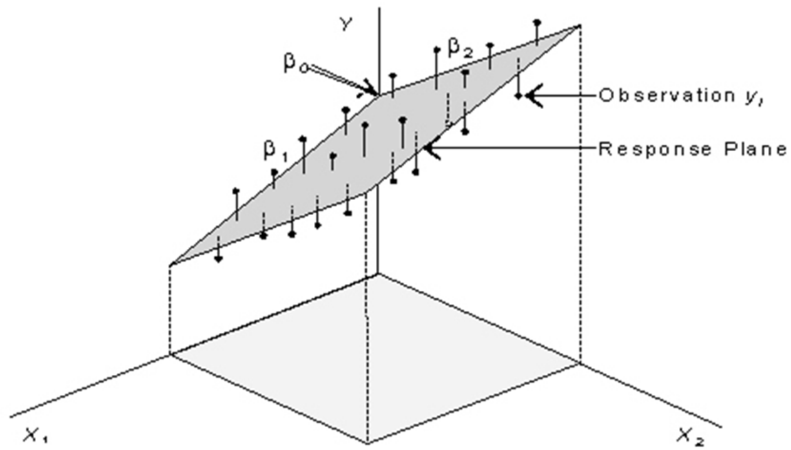
b_1=slope of 1^{st} variable

$X_1 = 1^{st}$ variable

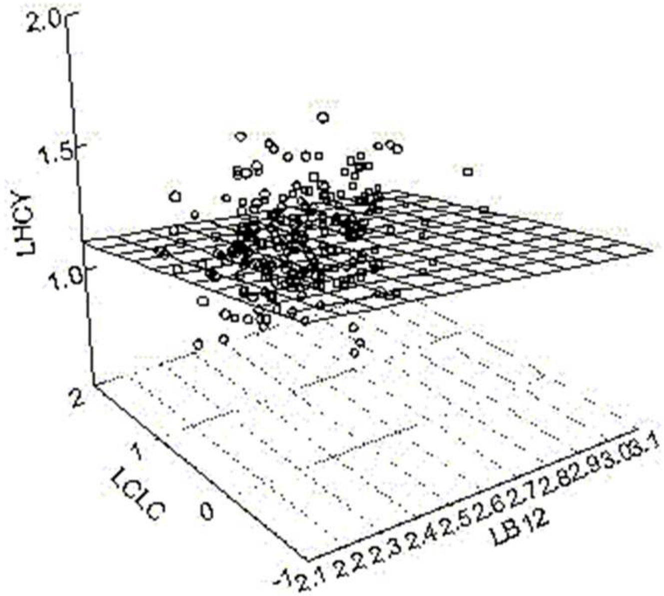
B_2=slope of 2^{nd} variable

X_2=second variable

c=intercept

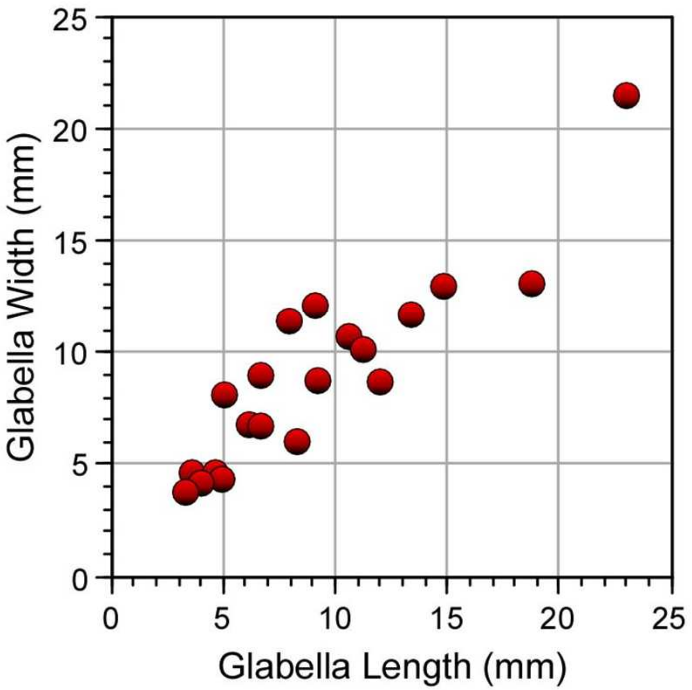


Another view

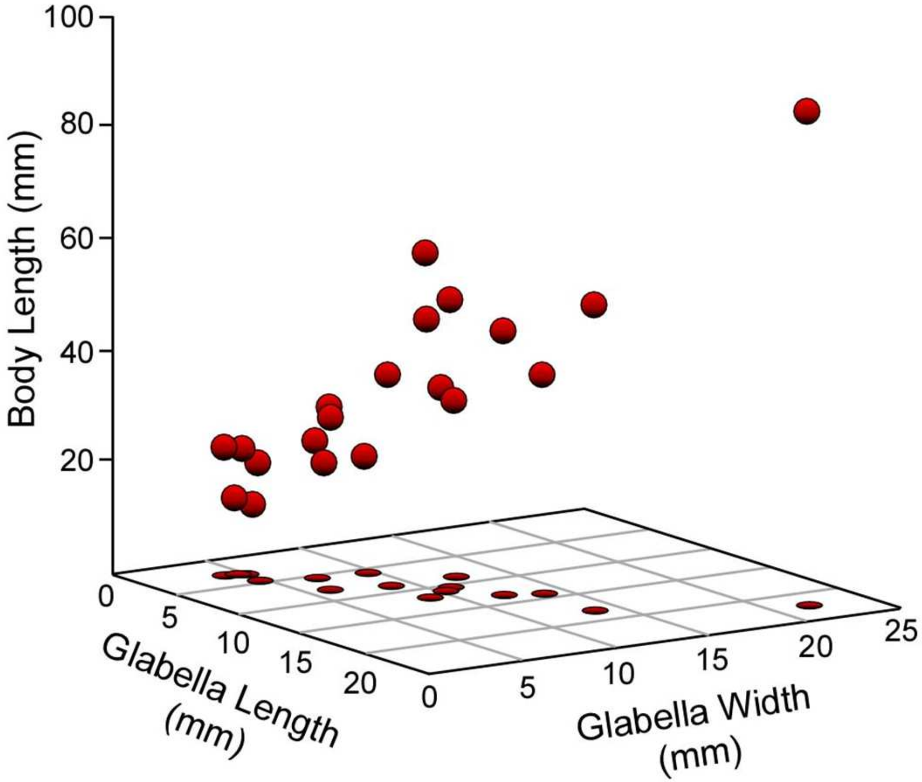


Another view

A.



B.

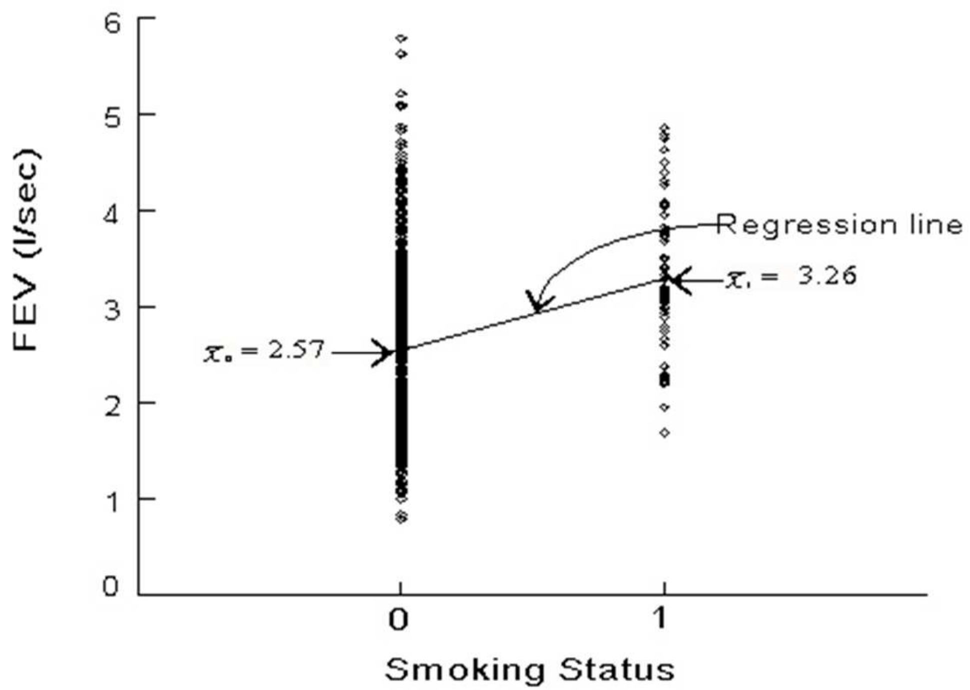


Dummy variables

- All examples have been continuous variables, but most historical data is categorical
- We can recode categorical variables into dummy variables
  - Race becomes codes
    - white (0,1)
    - black (0,1)
    - Asian (0,1)
  - One category must be omitted

Dummy variable scatter plot

(FEV is forced expiratory volume)



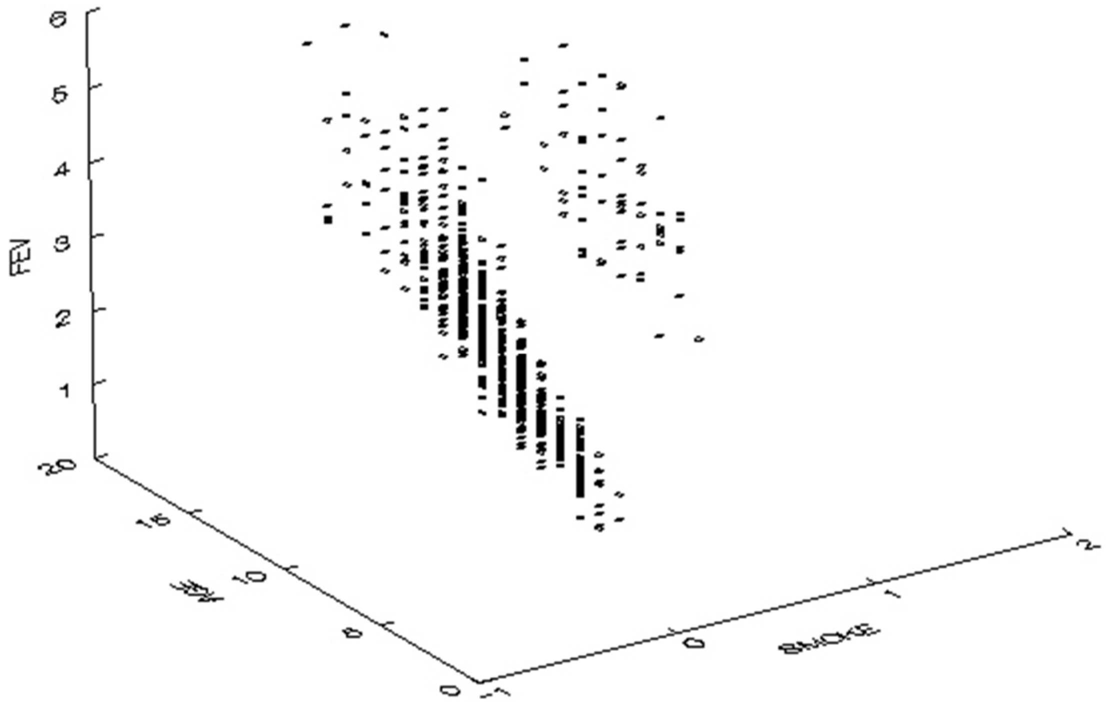
Bivariate regression results

Variable	ß (coefficient	F-test
SMOKE	0.71	41.7892

Y-Intercept 2.5661426

- The slope coefficient of 0.71 for SMOKE and FEV suggests hat FEV *increases* 0.71 units (on the average) as we go from SMOKE = 0 (non-smoker) to SMOKE = 1 (smoker). This is surprisingly given what we know about smoking. How can a positive relation between SMOKE and FEV exist given what we know about the physiological effects of smoking? The answer lies in understanding the confounding effects of AGE on SMOKE and FEV. In this child and adolescent population, nonsmokers are younger than nonsmokers (mean ages: 9.5 years vs. 13.5 years, respectively). It is therefore *not* surprising that AGE confounds the relation between SMOKE and FEV. So what are we to do?

Dummy variable scatterplot (3-D)



Multiple regression results

Variable	b coefficient	F-test
SMOKE	-0.2089949	6.6994
AGE	0.2306046	793.8988
Y-Intercept	0.3673730	

Intergenerational coresidence of the aged, 1950-2000: Independent variables

Table 1. State-level measures of intergenerational coresidence, earnings, and education, 1950-2000

	1950	1960	1970	1980	1990	2000	All years
Mean of states:							
Percent of persons aged 65+ residing with their children	34.1	24.4	17.4	13.9	13.5	14.5	19.6
Percent of persons aged 30-39 with low incomes	55.8	44.8	39.0	32.3	29.7	28.5	38.7
Percent of persons aged 65+ with low incomes	75.4	61.6	49.1	32.1	27.5	23.7	44.9
Percent of persons aged 30-39 completed high school	44.0	54.6	66.7	81.9	89.3	89.8	71.1
Percent of persons aged 65+ completed high school	17.1	19.3	27.0	38.8	56.5	69.2	38.0
Standard deviation:							
Percent of persons aged 65+ residing with their children	6.6	5.9	4.3	3.8	3.6	3.8	8.9
Percent of persons aged 30-39 with low incomes	7.1	5.1	3.7	3.4	4.5	4.3	11.0
Percent of persons aged 65+ with low incomes	6.0	7.9	8.2	6.5	5.3	3.3	20.0
Percent of persons aged 30-39 completed high school	10.9	9.1	8.0	5.9	3.9	3.4	18.9
Percent of persons aged 65+ completed high school	5.0	5.0	6.5	9.0	8.9	7.1	20.5
Number of cases:	46	46	46	46	46	46	276

Note: Alaska, Delaware, Hawaii, Nevada, and Wyoming excluded because of insufficient cases; the District of Colombia is treated as a state. Low income is defined as half the median income for each age group in the 2000 census (under $12,046 for persons 30-39, and under $6,998 for persons aged 65 or over in 2000 dollars).

Source: Ruggles et al. (2004)

State-level regression results

Table 2. State-level models of education and income on percent of elders residing with adult
Ordinary Least Squares (OLS) Models with Pooled data, 1950-2000

	Model 1 OLS		Model 3 OLS		Model 5 OLS	
	B	t	B	t	B	t
Census Year						
1950	19.61	36.00 ***	4.29	1.64	0.97	0.25
1960	9.87	18.13 ***	-0.34	-0.15	-3.99	-1.22
1970	2.91	5.34 ***	-3.41	-2.26 *	-5.88	-2.30 *
1980	-0.61	-1.11	-2.79	-4.07 ***	-3.42	-2.12 *
1990	-1.03	-1.89	-1.87	-3.48 **	-1.14	-1.58
2000	(reference)		(reference)		(reference)	
Income and education						
Percent of persons aged 30-39 with low incomes			0.30	6.42 ***	0.12	2.85 ***
Percent of persons aged 65+ with low incomes			0.12	2.31 *	-0.01	-0.30
Percent of persons aged 30-39 completing high school					-0.37	-8.60 ***
Percent of persons aged 65+ completing high school					0.01	0.30
State effects	Yes		Yes		Yes	
Constant	14.42	1.12	4.29	2.59	43.84	6.77 ***
rho / lamda						
Adjusted R Square/pseudo F	0.91		0.93		0.95	
Log likelihood						
N	276		276		276	

* p < .05 ** p < .01 *** p < .001

Source: Ruggles et al. (2004)

Note: Omitted state is New Hampshire