Module 19: Simple Linear Regression

This module focuses on simple linear regression and thus begins the process of exploring one of the more used and powerful statistical tools.

Goldman-Tono-Pen Example

An ophthalmologist who is assessing intraocular pressures as a part of a community program for the prevention of glaucoma is interested in using a portable device (Tono-Pen) for making these measurements. An important question is how well the measurements made with this device compare to those made with a more standard device (Goldman) used in clinical settings. To address this question, the ophthalmologist compared the two devices by using each on n = 40 eyes. For this comparison, each eye was measured once with each device.

Goldman-Tono-Pen Example Data

ID	Goldman	T-Pen	ID	Goldman	T-Pen
1	17	22	21	26	24
2	19	19	22	13	12
3	20	14	23	22	19
4	27	20	24	19	18
5	19	15	25	21	23
6	17	20	26	23	24
7	22	29	27	19	16
8	17	22	28	21	20
9	19	19	29	17	18
10	23	16	30	20	14
11	29	17	31	15	17
12	19	20	32	20	18
13	13	12	33	12	14
14	18	14	34	20	18
15	22	20	35	22	20
16	23	17	36	20	21
17	18	14	37	23	20
18	20	24	38	30	30
19	19	20	39	27	27
20	21	21	40	17	18

Comparing the two Devices

One approach to comparing the two devices would be to do a paired t-test, which would be appropriate since the measurements made by the two devices on the same eyes could not be considered independent and since the differences between the two measurements are of interest.

Goldman-Tono-Pen Worksheet

ID	Goldman x = G	Tono-Pen y = T	d	d^2	ID	Goldman x = G	Tono-Pen y = T	d	d^2
1	17	22	-5	25	**21**	26	24	2	4
2	19	19	0	0	**22**	13	12	1	1
3	20	14	6	36	**23**	22	19	3	9
4	27	20	7	49	**24**	19	18	1	1
5	19	15	4	16	**25**	21	23	-2	4
6	17	20	-3	9	**26**	23	24	-1	1
7	22	29	-7	49	**27**	19	16	3	9
8	17	22	-5	25	**28**	21	20	1	1
9	19	19	0	0	**29**	17	18	-1	1
10	23	16	7	49	**30**	20	14	6	36
11	19	17	2	4	**31**	15	17	-2	4
12	19	20	-1	1	**32**	20	18	2	4
13	13	12	1	1	**33**	12	14	-2	4
14	18	14	4	16	**34**	20	18	2	4
15	22	20	2	4	**35**	22	20	2	4
16	23	17	6	36	**36**	20	21	-1	1
17	18	14	4	16	**37**	23	20	3	9
18	20	24	-4	16	**38**	30	30	0	0
19	19	20	-1	1	**39**	27	27	0	0
20	21	21	0	0	**40**	17	18	-1	1
					Sum	799	766	33	451
					Mean	19.975	19.15	0.825	

	Goldman	Tono-Pen		
ID	X=G	Y=T	d=G-T	d^2
N	40	40	40	
Sum	799	766	33	451
Mean	19.975	19.15	0.825	
SD	3.7106	4.185	3.296	
Sum^2/n	15,960.03	14,668.90	27.23	
$Sum(x^2)$	16,497	15,352	451	
SS	536.98	683.10	423.78	
s^2	13.77	17.52	10.87	
SE	0.587	0.662	0.521	
t = mean(d)/SE(d)		1.58		
df = n-1		39		
$t_{0.975}(39)$		2.02		

1. **Hypothesis:** $H_0: \Delta = \mu_G - \mu_T = 0$ vs. $H_1: \Delta \neq 0$,

2. **Assumptions:** Differences are a random sample with normal distribution,

3. **The α level:** $\alpha = 0.05$,

4. **Test statistic:** $$t = \frac{\bar{d}}{s_d / \sqrt{n}} = \frac{\bar{d}}{s_{\bar{d}}}$$

5. **The Rejection Region:** Reject if t is not between $\pm\, t_{0.975}(39) = 2.02$

6. **The Result:** $$n = 40,\ \bar{d} = 0.8,\ s_{\bar{d}} = 0.52$$
$$t = \frac{0.8}{0.52} = 1.58$$

7. **The conclusion:** Accept $H_0: \Delta = \mu_G - \mu_T = 0$, since t is between $\pm$ 2.02.

Hence, from this standpoint, we do not have compelling evidence that the two devices are measuring intra-ocular pressures differently. Is this a sufficient assessment of the situation, or should we look further?

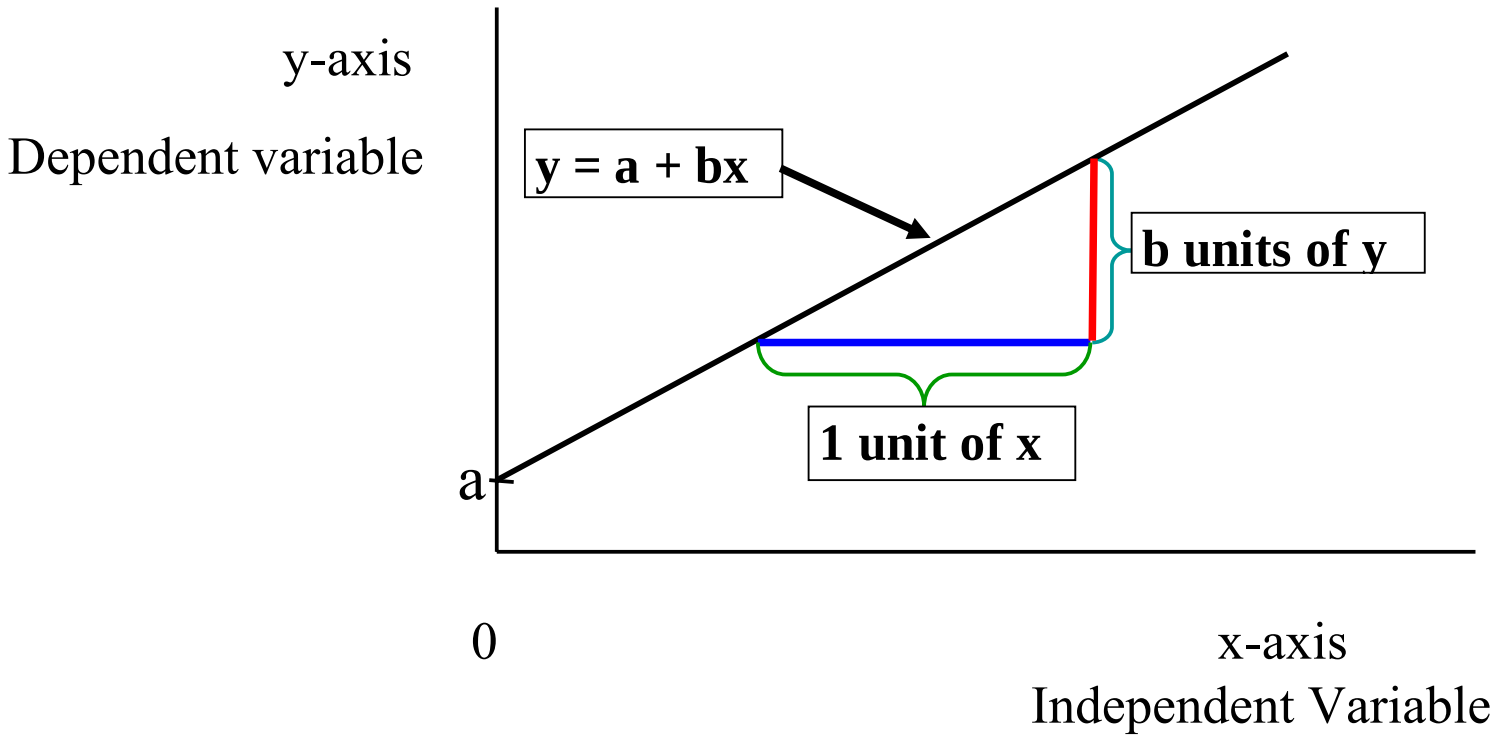
Looking Further

One way to look further at this situation is to think about the relationship between the measurements made by the two machines in terms of simple linear regression. In this context, we would wonder if higher values on one machine more directly imply higher values on the other.

Simple linear regression focuses on a possible straight line relationship between the measurements made by the two machines.

Simple Linear Regression Concepts

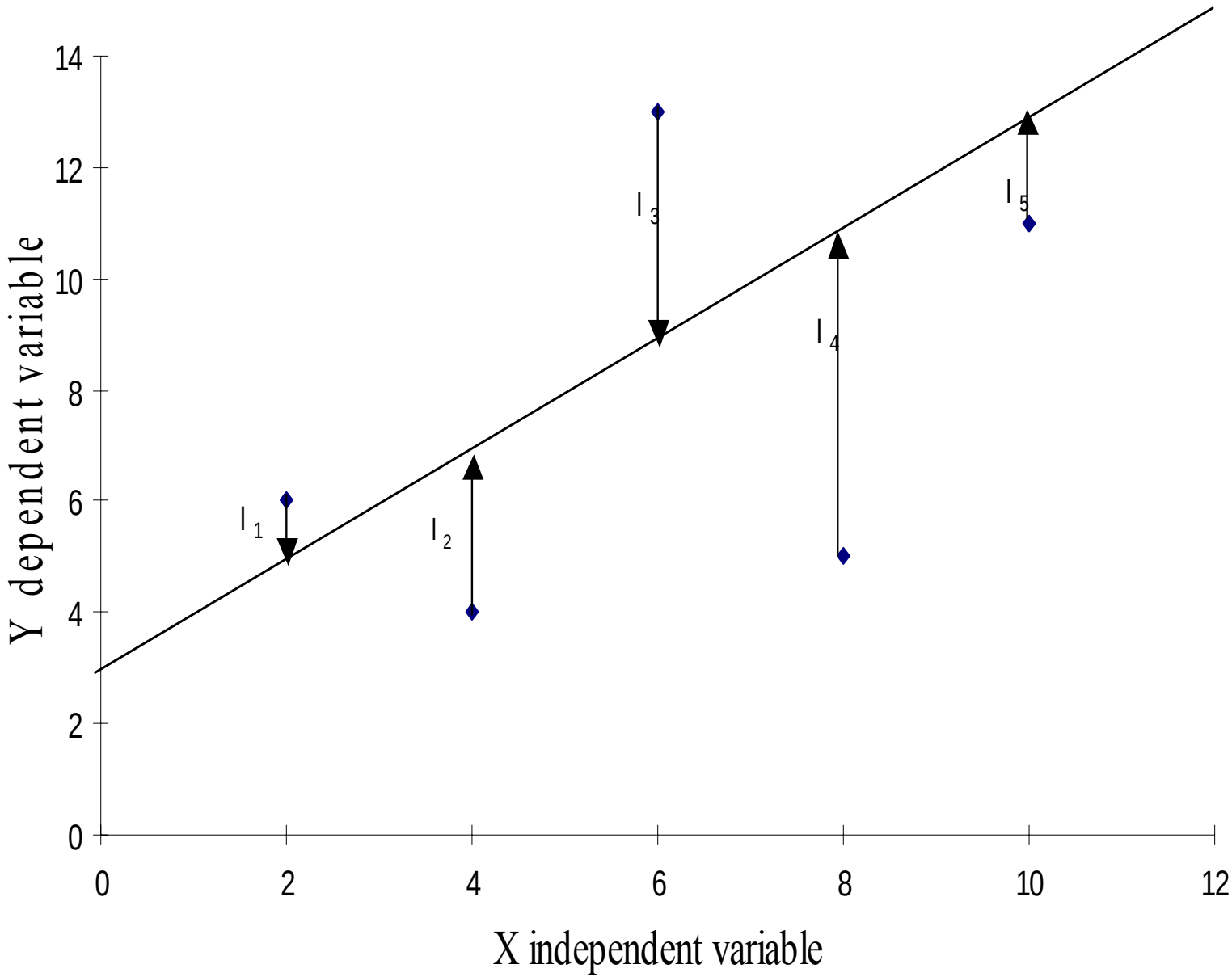
In general, simple linear regression finds the best straight line for describing the relationship between two variables. In its simplest form, which is what we consider here, it does not do a very good job of assessing how well the line describes the data, but nevertheless provides useful information.

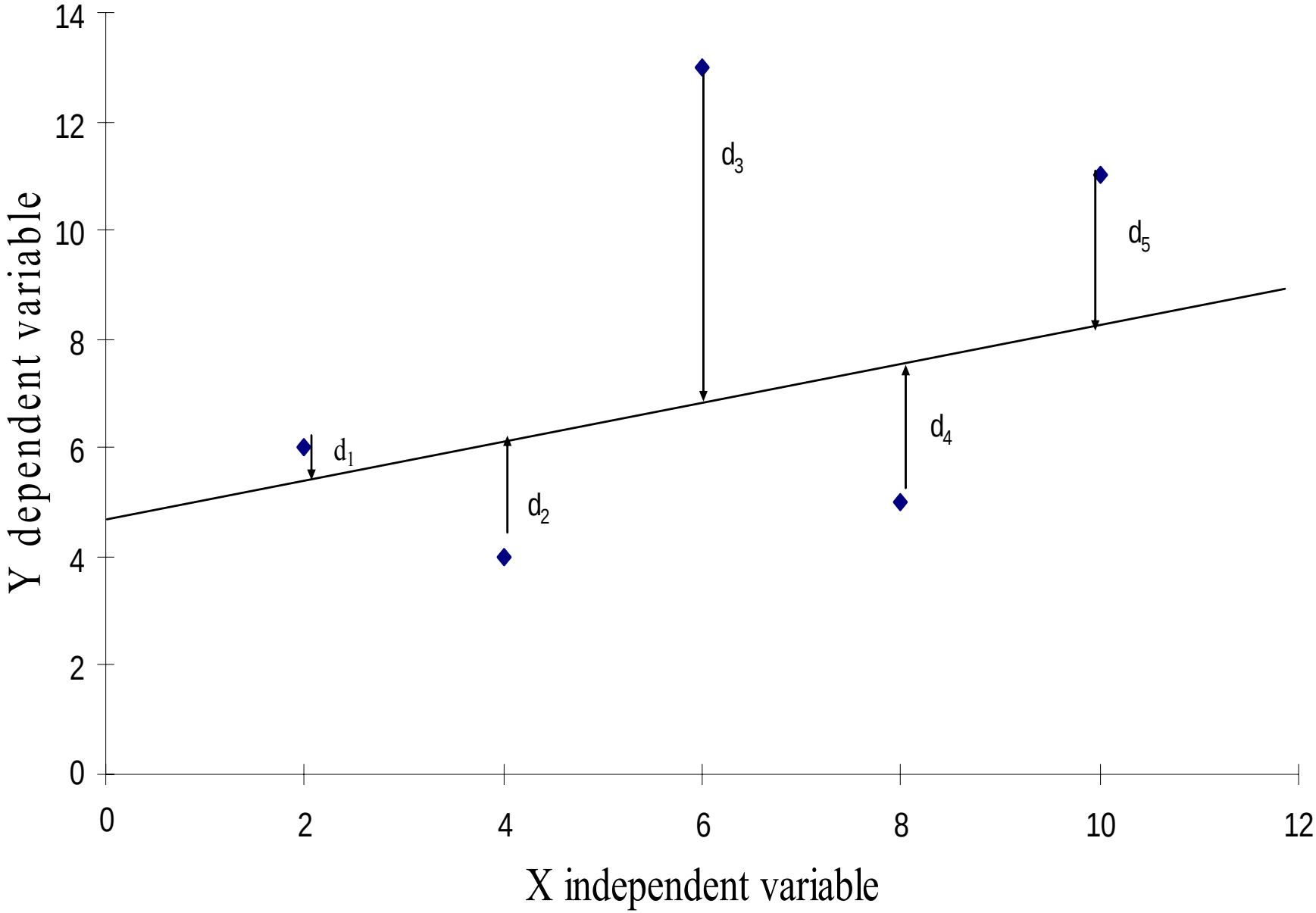


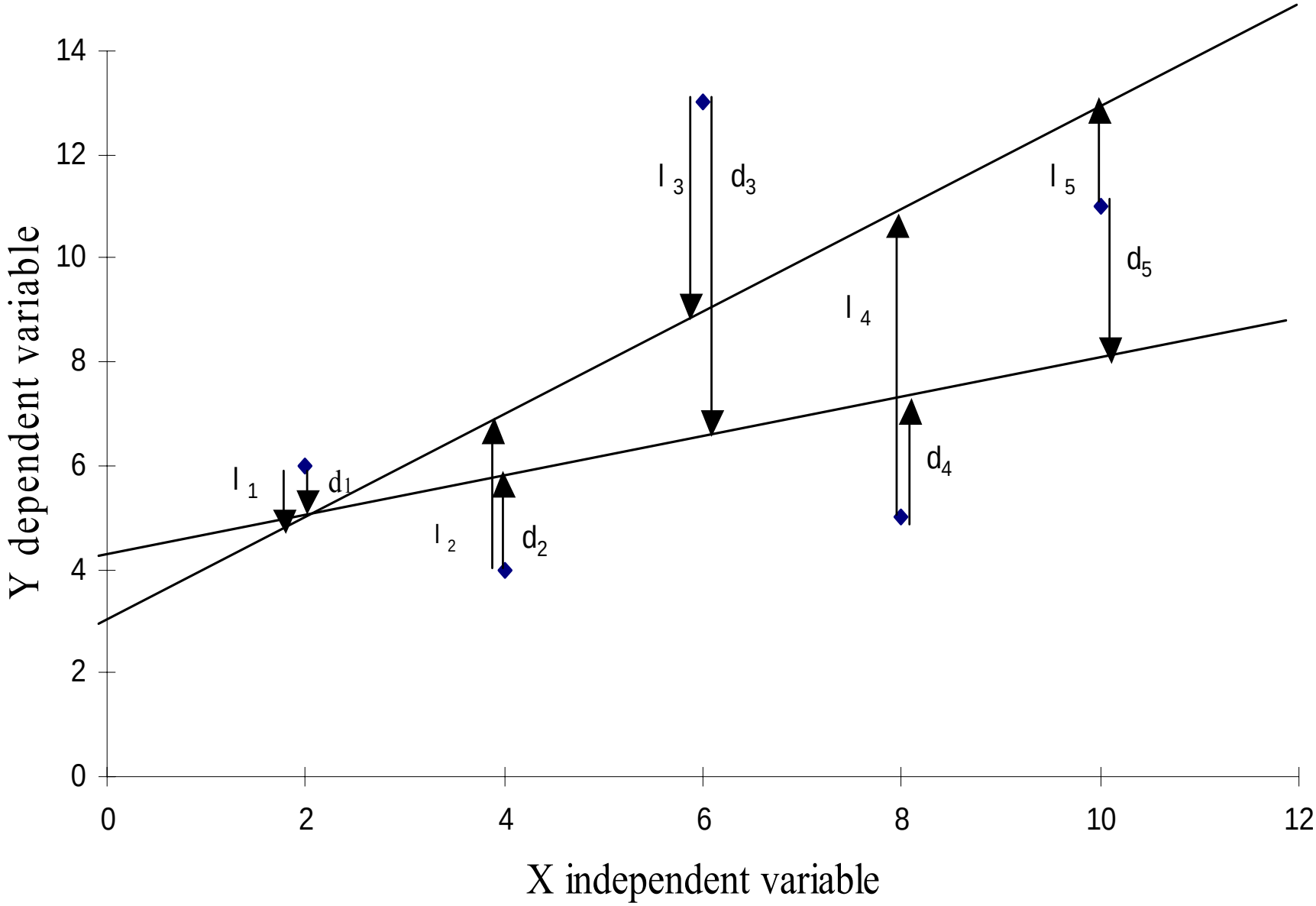
a = Intercept, that is, the point where the line crosses the y-axis, which is the value of y at **x** = 0.

b = Slope of the regression line, that is, the number of units of increase (positive slope) or decrease (negative slope) in **y** for each unit increase in **x**.

The Regression Line




Y dependent variable
X independent variable
0
2
4
6
8
10
12
14
d_1
d_2
d_3
d_4
d_5


14
12
10
8
6
4
2
0
0
2
4
6
8
10
12
Y dependent variable
X independent variable
l 1
d1
l 2
d2
l 3
d3
l 4
d4
l 5
d5

The context for simple linear regression is that we have a random sample of persons from a set of well-defined populations, each defined by a specific value for x-variable. We have measurements of another variable, the y-variable so that we have two variables for each person. For simple linear regression, we focus on a straight line that depicts the relationship between these two variables. The best straight line is the one for which the sum of the squared vertical distances of each point from the line is the least. This "least squares" line has ***slope***

$$b = \frac{\sum xy - \sum x \sum y / n}{\sum x^2 - (\sum x)^2 / n} = \frac{SS(xy)}{SS(x)},$$

and ***intercept***

$$a = \overline{y} - b\overline{x}.$$

For this situation, the sample line

$$y = a + b\,x$$

is an estimate of the population line

$$Y = \alpha + \beta x,$$

and a and b are estimates of α and β respectively. For a specific value of x, such as x = 10, the value for y calculated from the regression equation is

$$\hat{y} = a + b(x = 10),$$

which is called the regression estimate of Y at the value x = 10.

Simple Regression Example

The following data are diastolic blood pressure (DBP) measurements taken at different times after an intervention for n = 5 persons. For each person, the data available include the time of the measurement and the DBP level. Of interest is the relationship between these two variables.

	Time		DPB		
Patient	x	x^2	y	y^2	xy
1	0	0	72	5,184	0
2	5	25	66	4,356	330
3	10	100	70	4,900	700
4	15	225	64	4,096	960
5	20	400	66	4,356	1,320
Sum	50	750	338	22,892	3,310
Mean	10		67.6		
n	5		5		

For the blood pressure data,

$$\bar{x} = 50/5 = 10,$$

$$\bar{y} = 338/5 = 67.6,$$

the slope is

$$b = \frac{\sum xy - \sum x \sum y / n}{\sum x^2 - (\sum x)^2 / n} = \frac{SS(xy)}{SS(x)},$$

$$b = \frac{3,310 - (50)(338)/5}{750 - (50)^2/5} = -0.28$$

and the intercept is

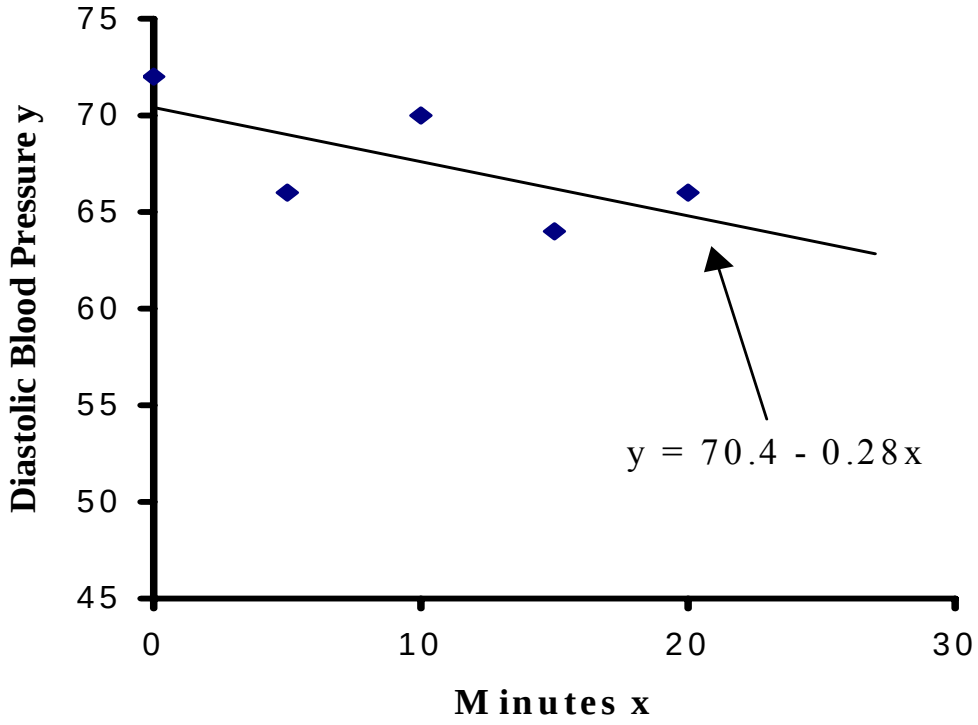
$$a = \bar{y} - b\bar{x},$$

$$a = 67.6 - (-0.28)10 = 70.4$$

The best line is

$$y = a + bx = 70.4 - 0.28x$$

Patient	Time x	DBP y
1	0	72
2	5	66
3	10	70
4	15	64
5	20	66



Example: *AJPH, Dec. 2003; 93: 2099-2104*

Secular Trends in Adolescent Never Smoking From 1990 to 1999 in California: An Age–Period–Cohort Analysis

Xinguang Chen, MD, PhD, Guohua Li, MD, DrPh, Jennifer B. Unger, PhD, Xiaowei Liu, MS, and C. Anderson Johnson, PhD

Objectives. We analyzed age, time period, and cohort effects on trends in adolescent cigarette smoking in California from 1990 to 1999.

Methods. Data from subjects aged 12 to 17 years (n = 26 536; 50.4% male) from the California Tobacco Survey and the California Youth Tobacco Survey were analyzed, and never smokers were used as the outcome measure.

Results. The proportion of never smokers increased from 60% for males and 66% for females in 1990 to around 70% for both sexes in 1999. Respondents were more likely to be never smokers if born in 1978 or later (i.e., aged 12 years or younger in 1990, when most tobacco control programs started in California).

Conclusions. The statewide antitobacco programs prevented adolescents from starting to smoke, primarily through a cohort effect. (*Am J Public Health.* 2003;93: 2099–2104)

TABLE 1—Data Sources and Sample Characteristics: Adolescent Respondents to the California Tobacco Survey and the California Youth Tobacco Survey

Survey and Year	Total No. Subjects	Age (y) at Survey (%)			Male (%)	Race/Ethnicity (%)			
		12–13	14–15	16–17		White	Hispanic/Latino	African American	Asian American
California Tobacco Survey									
1990–91	5015	34.02	33.84	32.14	50.49	57.51	26.08	5.84	8.09
1992	1782	35.30	33.84	30.86	49.38	52.24	30.64	6.51	8.25
1993	5495	34.69	34.09	31.23	51.21	55.65	27.84	5.79	8.39
Subtotal	12 292	34.50	33.95	31.55	50.65	55.91	27.53	5.91	8.21
California Youth Tobacco Survey									
1994	1738	36.54	34.52	28.94	49.37	48.68	32.51	7.31	8.80
1995	2153	35.21	34.84	29.96	50.49	46.73	34.42	6.36	9.38
1996	2504	34.03	34.58	31.39	50.60	46.73	36.30	6.39	8.07
1997	2691	34.86	33.85	31.29	50.76	48.42	36.08	6.35	7.88
1998	2460	34.88	33.17	31.95	51.30	44.96	39.88	7.56	5.45
1999	2698	34.62	34.66	30.73	48.85	43.81	40.44	7.49	6.12
Subtotal	14 244	34.93	34.25	30.83	50.25	46.43	36.91	6.90	7.50
Total	26 536	36.87	36.34	33.26	50.44	50.82	32.57	6.44	7.83

Note. Statistical analysis indicated no significant differences in age, sex, or ethnicity compositions between the California Tobacco Survey and the California Youth Tobacco Survey data. A total of 1179 subjects from the Los Angeles County Minority Health Survey of 1990–1991 were excluded from the analysis owing to increased proportions of ethnic minority respondents in its data, which may affect the estimates of the overall trends.

Never Smoking Regression Worksheet

	Year (x)	Female (y_1)	Male (y_2)	x^2	xy_1	xy_2	y_1^2	y_2^2
	1990.89	66.25	60.05	3963643	131896.5	119552.9445	4389.0625	3606.0025
	1992.4	67.125	64.6	3969657.8	133739.9	128709.04	4505.765625	4173.16
	1993.35	66.55	60.95	3973444.2	132657.4	121494.6825	4428.9025	3714.9025
	1994.35	65.85	62.65	3977431.9	131327.9	124946.0275	4336.2225	3925.0225
	1995.55	66.425	66.125	3982219.8	132554.4	131955.7438	4412.280625	4372.5156
	1996.65	67.65	64.55	3986611.2	135073.4	128883.7575	4576.5225	4166.7025
	1997.465	66.02	64.845	3989866.4	131872.6	129525.6179	4358.6404	4204.874
	1998.69	68.275	67.315	3994761.7	136460.6	134541.8174	4661.475625	4531.3092
	1999.55	69.775	69.425	3998200.2	139518.6	138818.7588	4868.550625	4819.8306
Total	17958.895	603.92	580.51	35835836	1205101	1158428.39	40537.4229	37514.32
Mean	1995.432778	67.10222222	64.5011111					
Sum^2	322521909.6							
Num b		19.52089444	59.7079472					
Denum b		68.53125555	68.5312556					
b		0.28484659	0.87125133					
a		-501.29	-1674.0223					

For the never smoking data

$$\overline{x} = 17958.895/9 = 1995.433$$

$$\overline{y}_{female} = 603.92/9 = 67.102\ ,$$

$$\overline{y}_{male} = 580.51/9 = 64.501$$

The slopes are $$b = \frac{\sum xy - \sum x \sum y / n}{\sum x^2 - (\sum x)^2 / n} = \frac{SS(xy)}{SS(x)},$$

$$b_{female} = \frac{1205101.284 - ((17958.895)(603.92)/9)}{35835836.27 - ((17958.895)^2/9)} = 0.285$$

$$b_{male} = \frac{1158428.39 - ((17958.895)(580.51)/9)}{35835836.27 - ((17958.895)^2/9)} = 0.871$$

The intercepts are

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

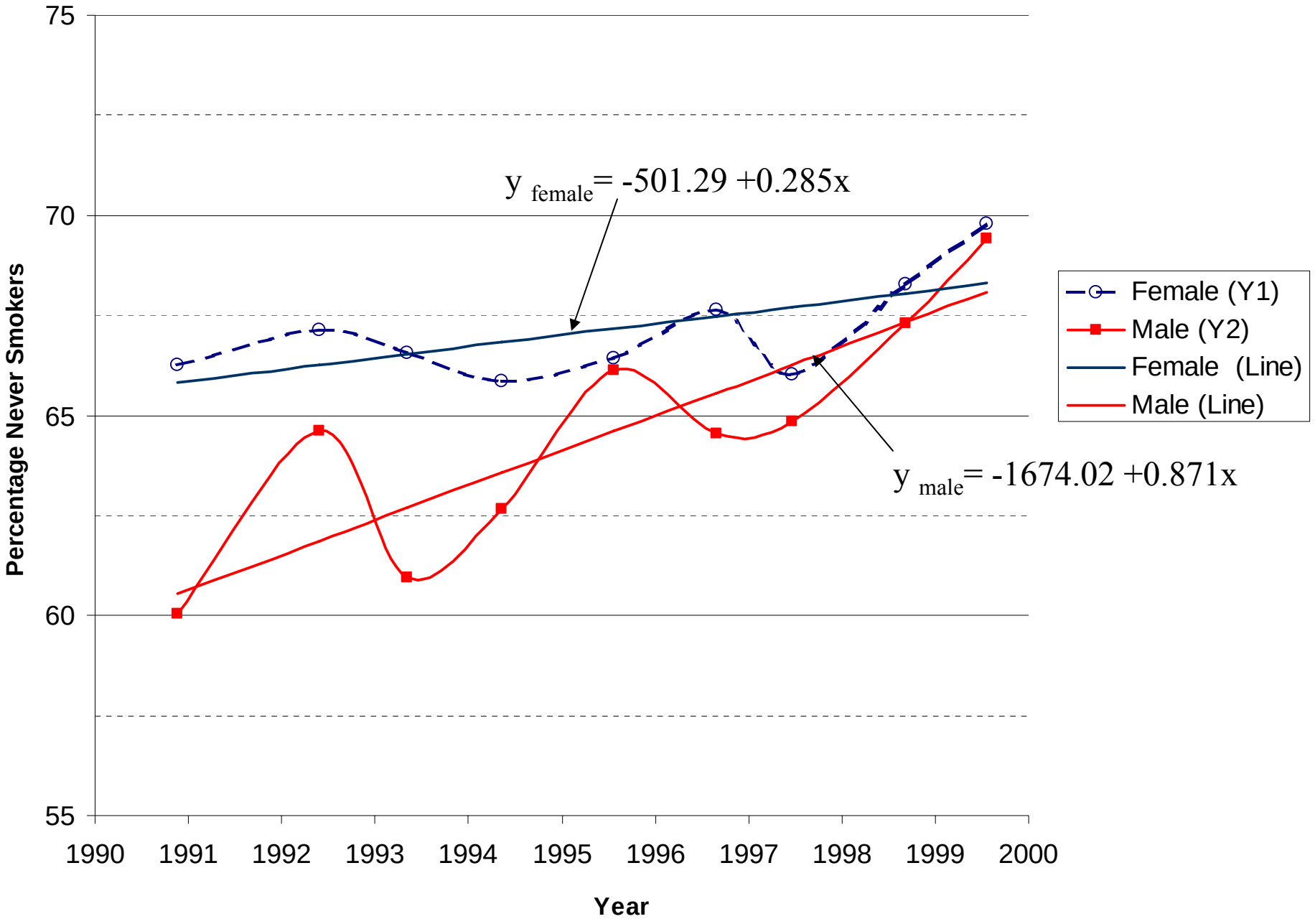
$$a_{female} = 67.102 - (0.285 * 1995.433) = \text{-}501.290$$

$$a_{male} = 64.501 - (0.871 * 1995.433) = \text{-}1674.022$$

The best lines are:

$$y_{female} = a_{female} + b_{female}x = -501.290 + 0.285x$$

$$y_{male} = a_{male} + b_{male}x = -1674.022 + 0.871x$$


75
70
65
60
55
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
Percentage Never Smokers
Year
y female= -501.29 +0.285x
y male= -1674.02 +0.871x
Female (Y1)
Male (Y2)
Female (Line)
Male (Line)

Regression ANOVA

If the regression line is flat in the sense that the regression estimate of Y, being $\hat{y}$, is the same for all values of x, then there is no gain from considering the x variable as it is having no impact on $\hat{y}$. This situation occurs when the estimated slope b = 0. An important question is whether or not the population parameter $\beta = 0$, that is, whether the truth is that there is no linear relationship between y and x. To test this situation, we can proceed with a formal test.

1. The Hypothesis: $H_0: \beta = 0$ vs $H_1: \beta \neq 0$

2. The α level: $\alpha = 0.05$

3. The assumptions: Random normal samples for y-variable from populations defined by x-variable

4. The test statistic:

ANOVA				
Source	df	SS	MS	F
Regression	1	SS(*Reg*)	SS(*Reg*)/1	MS(*Reg*)/MS(*Res*)
Residual	n-2	SS(*Res*)	SS(*Res*)/(n-2)	
Total	n-1	SS(y)		

5. The rejection region : Reject $H_0: \beta = 0$ if the value calculated for F is greater than $F_{0.95}(1, n-2)$

$$R^2 = SS(Reg) / SS(Total)$$

R^2 is the total amount of variation in the dependent variable y explained by its regression relationship with x .

Blood Pressure Example

$$SS(Total) = SS(y) = \sum (y - \bar{y})^2$$

$$= 22,892 - \frac{(338)^2}{5} = 43.2$$

$$SS(Regression) = bSS(xy)$$

$$= b\left\{\sum xy - \sum x \sum y / n\right\}$$

$$= -0.28\{3310 - (50)(338)/5\} = 19.6$$

$$SS(Residual) = SS(Total) - SS(Regression)$$

$$= 43.2 - 19.6 = 23.6$$

ANOVA

Source	df	SS	MS	F
Regression	1	19.6	19.6	2.49
Residual	3	23.6	7.89	
Total	4	43.2		

$H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$

For $\alpha = 0.05$ $F_{0.95(1,3)} = 10.1$, Hence accept $H_0 : \beta = 0$

$$R^2 = \frac{SS(Regression)}{SS(Total)} = \frac{19.6}{43.2} = 0.4537 \quad \text{or} \quad 45.37\%$$

Note: The above hypothesis test does not asses how well the straight line fits the data.

Goldman-Tono-Pen Example

We can apply these tools to the Goldman-Tono-Pen example. Note that while we test the null hypothesis H_0: $\beta = 0$, it is of little interest as it is not a very meaningful hypothesis.

Goldman Tono-Pen Example

ID	Goldman x = G	T-Pen y = T	d	d^2	G^2	T^2	GxT
1	17	22	-5	25	289	484	374
2	19	19	0	0	361	361	361
3	20	14	6	36	400	196	280
4	27	20	7	49	729	400	540
5	19	15	4	16	361	225	285
6	17	20	-3	9	289	400	340
7	22	29	-7	49	484	841	638
8	17	22	-5	25	289	484	374
9	19	19	0	0	361	361	361
10	23	16	7	49	529	256	368
11	19	17	2	4	361	289	323
12	19	20	-1	1	361	400	380
13	13	12	1	1	169	144	156
14	18	14	4	16	324	196	252
15	22	20	2	4	484	400	440
16	23	17	6	36	529	289	391
17	18	14	4	16	324	196	252
18	20	24	-4	16	400	576	480
19	19	20	-1	1	361	400	380
20	21	21	0	0	441	441	441
21	26	24	2	4	676	576	624
22	13	12	1	1	169	144	156
23	22	19	3	9	484	361	418
24	19	18	1	1	361	324	342
25	21	23	-2	4	441	529	483
26	23	24	-1	1	529	576	552
27	19	16	3	9	361	256	304
28	21	20	1	1	441	400	420
29	17	18	-1	1	289	324	306
30	20	14	6	36	400	196	280
31	15	17	-2	4	225	289	255
32	20	18	2	4	400	324	360
33	12	14	-2	4	144	196	168
34	20	18	2	4	400	324	360
35	22	20	2	4	484	400	440
36	20	21	-1	1	400	441	420
37	23	20	3	9	529	400	460
38	30	30	0	0	900	900	900
39	27	27	0	0	729	729	729
40	17	18	-1	1	289	324	306
Sum	799	766	33	1089	16,497	15,352	15,699

	Goldman X = G	Tono-Pen Y = T	d	d^2	G^2	T^2	GxT GT
Sum	799	766	33	451	16,497	15,352	15,699
Mean	20.0	19.2	0.8				
Sum^2/n	15,960.03	14,668.90	27.23				15,300.85
SS	536.98	683.10	423.78				398.15
s^2	13.77	17.52	10.87				
s = SD	3.7	4.2	3.3				
SE	0.59	0.66	0.52				
s^2_p (G,T)	15.64						
s_p (G,T)	4.0						
b	0.74						
a	4.34						

$$\hat{y} = a + bx$$

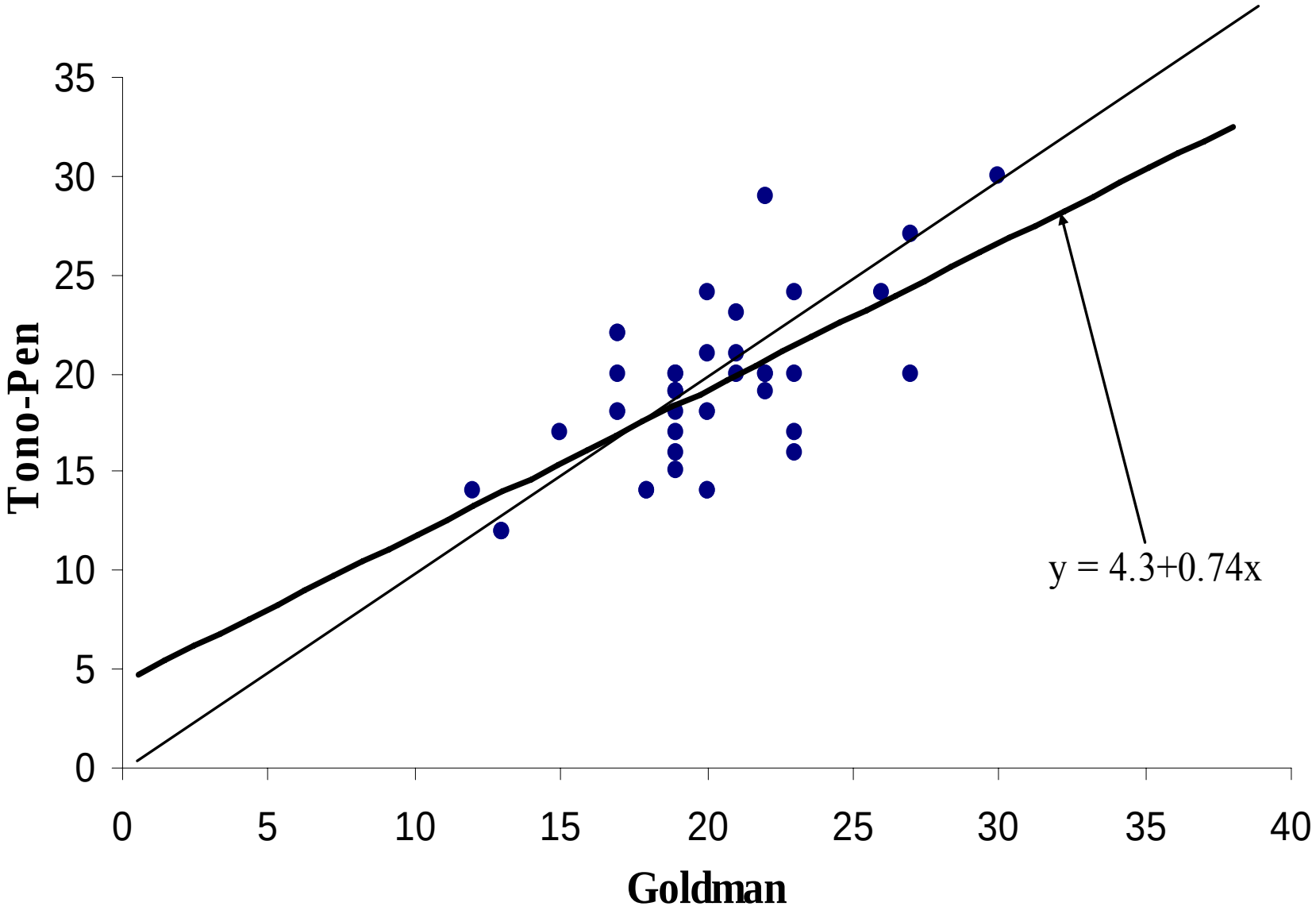
$$\hat{y} = 4.34 + 0.74x$$

SS(Reg)	295.22	R^2	0.4322
SS(Residual)	387.88		
SS(Total)	683.10	r	0.6574

Estimated Tono-Pen measurements for specific Goldman results

Goldman	Estimated Tono-Pen	Difference
5	8.0	3.0
10	11.8	1.8
15	15.5	0.5
20	19.2	-0.8
25	22.9	-2.1
30	26.6	-3.4
35	30.3	-4.7

Create a new table


Goldman-Tono-Pen Example
Tono-Pen
Goldman
y = 4.3+0.74x
0
5
10
15
20
25
30
35
40

Regression ANOVA – Goldman Tono-Pen Example

1. The Hypothesis: $H_0: \beta = 0$ vs $H_1: \beta \neq 0$

2. The Assumptions: Random samples, x measured without error, y normal distributed for each level of x

3. **The α-level:** $\alpha = 0.05$

4. **The test statistic:** ANOVA

5. **The rejection region:** Reject $H_0: \beta = 0$, if

$$F = \frac{MS(\text{Re}\,gression)}{MS(\text{Re}\,sidual)} > F_{0.95(1,38)} \approx 4.08$$

6. The result: n = 40, SS(*Regression*) = 295.22
SS(*Residual*) = 387.88
SS(*Total*) = 683.10

$F_{0.95(1,38)} \approx 4.08$

ANOVA

Source	DF	SS	MS	F
Regression	1	295.22	295.22	28.91
Residual	38	387.88	10.21	
Total	39	683.10		

7. The conclusion: Reject H_0: $\beta = 0$ since $28.91 > 4.08$

Example: *AJPH, Aug. 1999; 89: 1187-1193*

Social Capital and Self-Rated Health: A Contextual Analysis

Ichiro Kawachi, MD, PhD, Bruce P. Kennedy, EdD, and Roberta Glass, MS

ABSTRACT

Objectives. Social capital consists of features of social organization—such as trust between citizens, norms of reciprocity, and group membership—that facilitate collective action. This article reports a contextual analysis of social capital and individual self-rated health, with adjustment for individual household income, health behaviors, and other covariates.

Methods. Self-rated health ("Is your overall health excellent, very good, good, fair, or poor?") was assessed among 167 259 individuals residing in 39 US states, sampled by the Behavioral Risk Factor Surveillance System. Social capital indicators, aggregated to the state level, were obtained from the General Social Surveys.

Results. Individual-level factors (e.g., low income, low education, smoking) were strongly associated with self-rated poor health. However, even after adjustment for these proximal variables, a contextual effect of low social capital on risk of self-rated poor health was found. For example, the odds ratio for fair or poor health associated with living in areas with the lowest levels of social trust was 1.41 (95% confidence interval = 1.33, 1.50) compared with living in high-trust states.

Conclusions. These results extend previous findings on the health advantages stemming from social capital. (*Am J Public Health.* 1999;89:1187–1193)

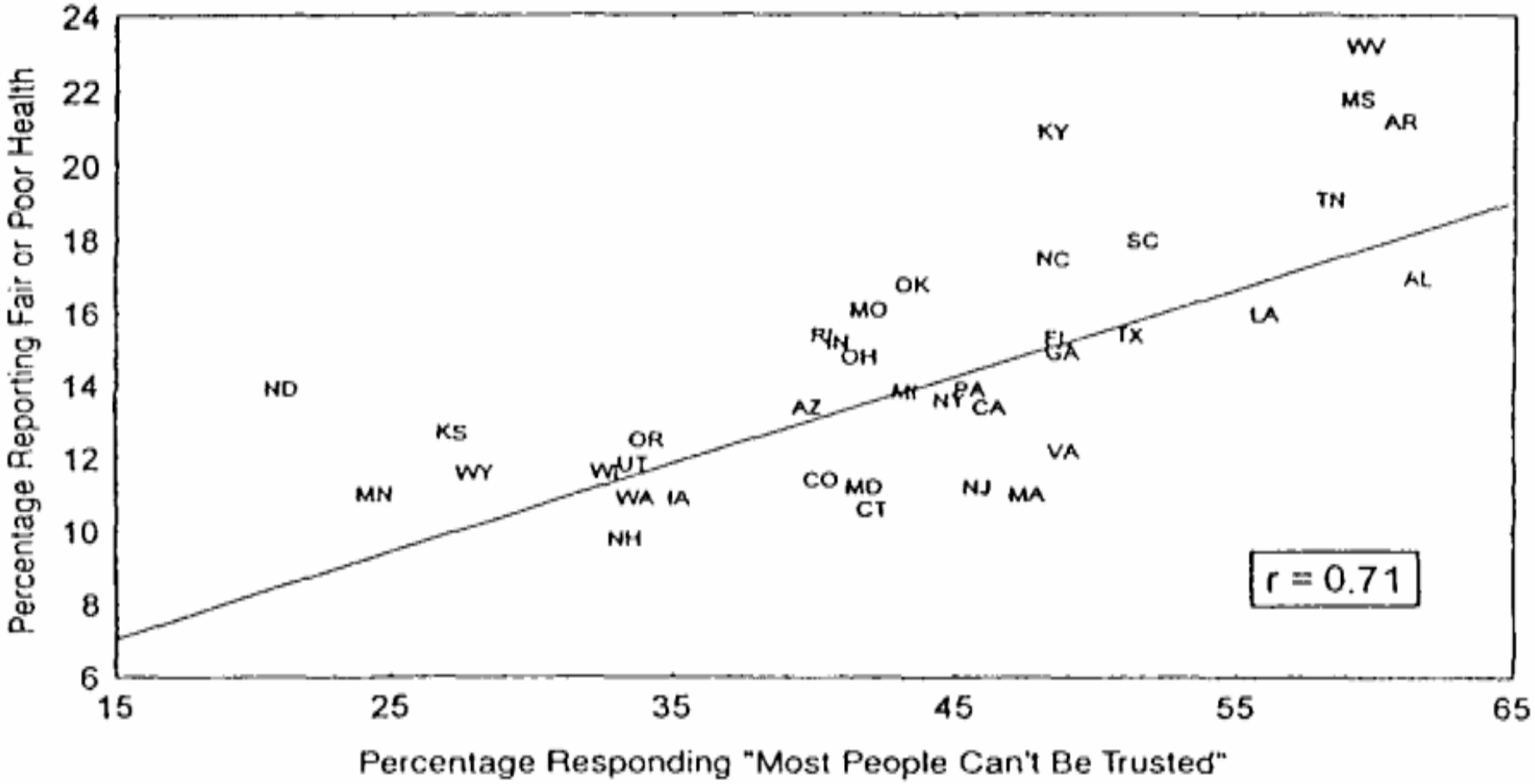
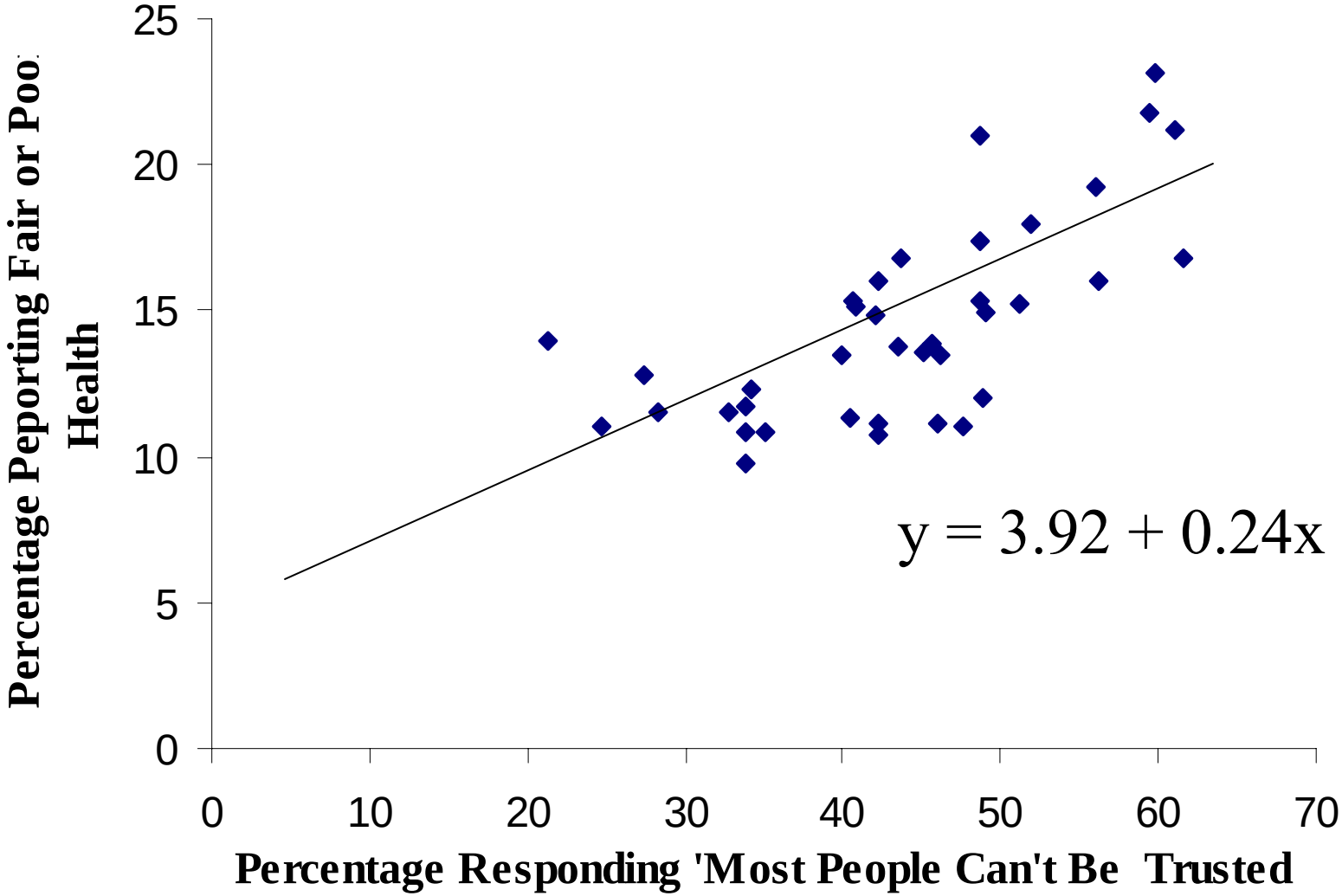


FIGURE 1—Scatterplot of levels of interpersonal trust and percentage of residents in each state reporting fair or poor health, Behavioral Risk Factor Surveillance System, 1993–1994.

	State	Y	X	Y^2	X^2	XY
1	AL	16.80	61.63	282.24	3797.64	1035.30
2	AR	21.20	61.13	449.44	3736.27	1295.85
3	AZ	13.50	40.00	182.25	1600.00	540.00
4	CA	13.50	46.25	182.25	2139.06	624.38
5	CO	11.30	40.38	127.69	1630.14	456.24
6	CT	10.75	42.25	115.56	1785.06	454.19
7	FL	15.30	48.63	234.09	2364.39	743.96
8	GA	14.90	49.00	222.01	2401.00	730.10
9	IA	10.80	35.13	116.64	1233.77	379.35
10	IN	15.10	40.88	228.01	1670.77	617.21
11	KS	12.75	27.38	162.56	749.39	349.03
12	KY	21.00	48.63	441.00	2364.39	1021.13
13	LA	16.00	56.13	256.00	3150.02	898.00
14	MA	11.00	47.63	121.00	2268.14	523.88
15	MD	11.10	42.25	123.21	1785.06	468.98
16	MI	13.80	43.50	190.44	1892.25	600.30
17	MN	11.00	24.69	121.00	609.47	271.56
18	MO	16.00	42.25	256.00	1785.06	676.00
19	MS	21.75	59.50	473.06	3540.25	1294.13
20	NC	17.40	48.63	302.76	2364.39	846.08
21	ND	14.00	21.25	196.00	451.56	297.50
22	NH	9.80	33.88	96.04	1147.52	331.98
23	NJ	11.10	46.00	123.21	2116.00	510.60
24	NY	13.60	45.13	184.96	2036.27	613.70
25	OH	14.80	42.00	219.04	1764.00	621.60
26	OK	16.80	43.63	282.24	1903.14	732.90
27	OR	12.30	34.13	151.29	1164.52	419.74
28	PA	13.90	45.63	193.21	2081.64	634.19
29	RI	15.30	40.63	234.09	1650.39	621.56
30	SC	18.00	52.00	324.00	2704.00	936.00
31	TN	19.25	56.00	370.56	3136.00	1078.00
32	TX	15.25	51.25	232.56	2626.56	781.56
33	UT	11.75	33.75	138.06	1,139.06	396.56
34	VA	12.00	48.88	144.00	2388.77	586.50
35	WA	10.80	33.88	116.64	1147.52	365.85
36	WI	11.50	32.75	132.25	1072.56	376.63
37	WV	23.10	59.88	533.61	3585.02	1383.11
38	WY	11.50	28.25	132.25	798.06	324.88
	Total	549.70	1654.69	8391.24	75779.10	24838.49
	Mean	14.47	43.54			
	SD	3.75	23.6			
	r	0.7				
	slope	0.24				
	intercept	3.92				
	Value at 15	7.56				



At x = 45, y = 14.72

r = 0.70

Regression ANOVA
Social Capital and Self-Rated Health Example

1. The Hypothesis: $H_0: \beta = 0$ vs $H_1: \beta \neq 0$

2. The Assumptions: Random samples, x measured without error, y normal distributed for each level of x

3. **The α-level:** $\alpha = 0.05$

4. **The test statistic:** ANOVA

5. **The rejection region:** Reject $H_0: \beta = 0$, if

$$F = \frac{MS(Regression)}{MS(Residual)} > F_{0.95(1,36)} \approx 4.11$$

6. The result: n = 38, SS(*Regression*) = 218.37
SS(*Residual*) = 221.03
SS(*Total*) = 439.40

$$F_{0.95(1,36)} \approx 4.11$$

ANOVA

Source	DF	SS	MS	F
Regression	1	218.37	218.37	35.57
Residual	36	221.03	6.14	
Total	37	439.40		

7. The conclusion: Reject H_0: $\beta = 0$ since $35.57 > 4.11$

Example: *AJPH, July 1999; 89: 1059 -1065*

The Duration and Timing of Exposure: Effects of Socioeconomic Environment on Adult Health

Chris Power, PhD, Orly Manor, PhD, and Sharon Matthews, MSc

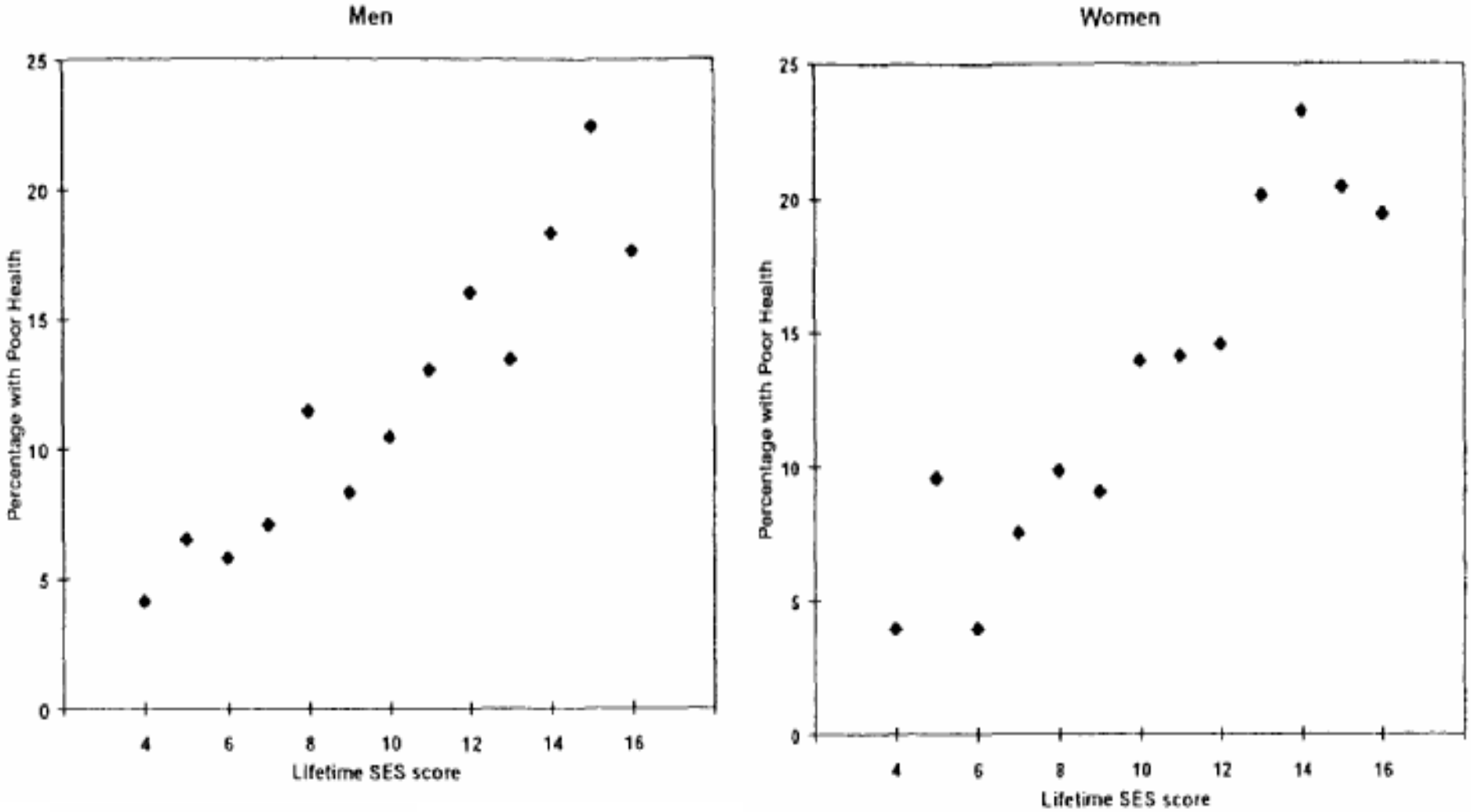
ABSTRACT

Objectives. This study investigated timing and duration effects of socioeconomic status (SES) on self-rated health at 33 years of age and established whether health risks are modified by changing SES and whether cumulative SES operates through education.

Methods. Data were from the 1958 British birth cohort. Occupational class at birth and at 16, 23, and 33 years of age was used to generate a lifetime SES score.

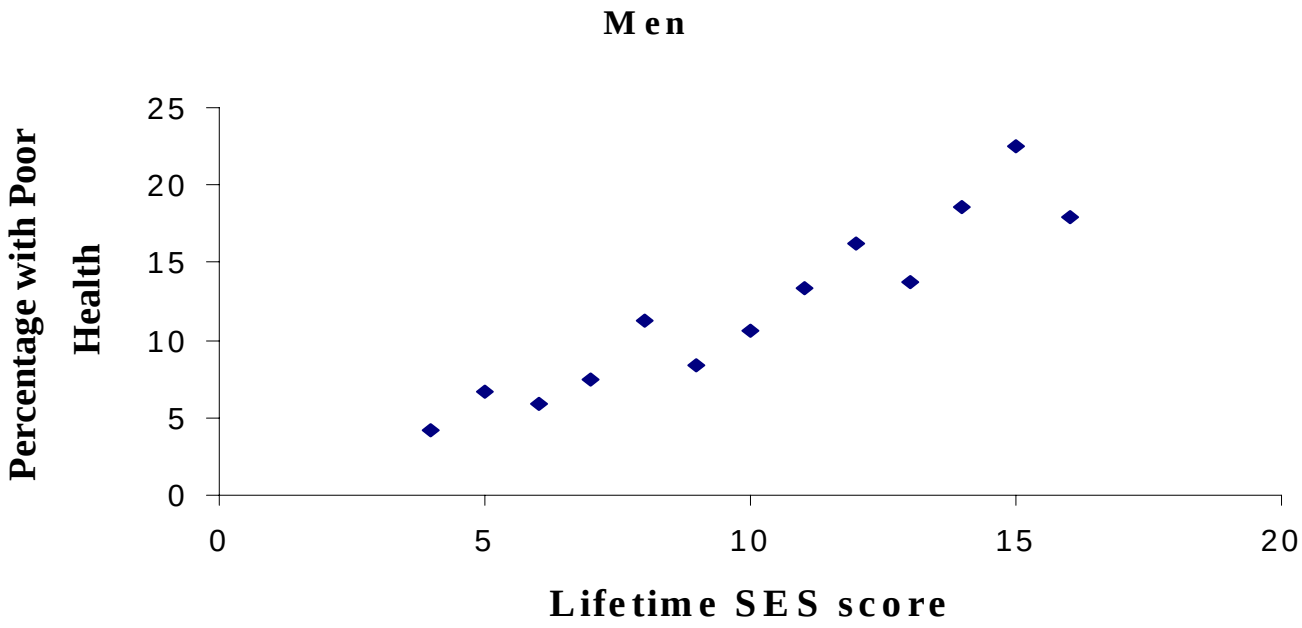
Results. At 33 years of age, 12% of men and women reported poor health. SES at birth and at 16, 23, and 33 years of age was significantly associated with poor health: all ages except 16 years in men made an additional contribution to the prediction of poor health. No large differences in effect sizes emerged, suggesting that timing was not a major factor. Odds of poor health increased by 15% (men) and 18% (women) with a 1-unit increase in the lifetime SES score. Strong effects of lifetime SES persisted after adjustment for education level.

Conclusions. SES from birth to 33 years of age had a cumulative effect on poor health in early adulthood. This highlights the importance of duration of exposure to socioeconomic conditions for adult health. (*Am J Public Health.* 1999;89:1059–1065)



Note. SES = socioeconomic status.

FIGURE 1—Poor health (including subjects who rated their health as fair) at 33 years of age and cumulative socioeconomic circumstances (birth to 33 years of age).


Men
Percentage with Poor Health
25
20
15
10
5
0
0
5
10
15
20
Lifetime SES score


Women
Percentage with Poor Health
25
20
15
10
5
0
0
5
10
15
20
Lifetime SES score

Socioeconomic Environment and Adult Health Example

	Men				Women				
X	X^2	Y	Y^2	XY	X	X^2	Y	Y^2	XY
4.0	15.9	4.1	17.1	16.5	3.9	15.4	4.3	18.1	16.7
5.0	25.0	6.7	45.2	33.6	5.0	25.0	9.4	88.5	47.1
6.0	36.1	5.9	34.3	35.2	6.0	35.4	4.3	18.1	25.3
7.0	49.4	7.2	52.4	50.9	7.0	48.4	7.4	54.0	51.2
8.1	65.6	11.2	125.7	90.8	8.0	63.5	9.9	97.0	78.5
9.0	80.8	8.5	71.4	76.0	9.0	80.8	9.1	83.2	82.0
10.0	100.0	10.5	110.7	105.2	10.0	100.0	14.3	203.3	142.6
11.0	119.9	13.5	180.9	147.3	10.9	118.6	14.3	203.3	155.3
12.0	144.7	16.2	262.8	195.0	12.0	143.0	14.7	216.4	175.9
13.0	168.2	13.8	190.2	178.9	12.9	166.7	20.3	411.7	261.9
13.9	193.8	18.6	346.7	259.2	13.9	193.8	23.7	560.7	329.6
14.9	223.2	22.6	510.3	337.5	14.9	221.1	20.9	436.0	310.5
15.9	252.5	18.1	327.6	287.6	15.8	250.3	19.6	382.6	309.4
$\sum$ 129.8	1475.2	156.9	2275.2	1813.6	129.2	1462.0	171.9	2773.1	1986.1
n 13		13			13		13		
$\bar{X}$ 10.0		12.1			9.9		13.2		
SD 3.9		5.6			3.9		6.5		

X: Lifetime socioeconomic status (SES) score

Y : Percentage with Poor Health

Socioeconomic Environment and Adult Health Example

Men	Women
SS(x) = 179.20	SS(x) = 177.95
SS(y) = 381.54	SS(y) = 500.05
SS(xy) = 247.01	SS(xy) = 277.68
b = 1.38	b = 1.56
a = -1.57	a = -2.25
r = 0.9447	r = 0.9309
SS(*Reg*) = 340.50	SS(*Reg*) = 433.30
SS(*Res*) = 41.04	SS(*Res*) = 66.75
SS(*Total*) = 381.54	SS(*Total*) = 500.05
$\hat{y}_M = -1.57 + 1.38x$	$\hat{y}_W = -2.25 + 1.56x$

Socioeconomic Environment and Adult Health Example

	Men	Women
1. **The hypothesis:**	H_0: $\beta = 0$ vs H_1: $\beta \neq 0$	H_0: $\beta = 0$ vs H_1: $\beta \neq 0$
2. **The assumptions:**	Random samples x measured without error y normal distributed for each level of x	The same as that of men
3. **The α-level :**	$\alpha = 0.05$	$\alpha = 0.05$
4. **The test statistic:**	ANOVA	ANOVA
5. **The rejection region:**	Reject H_0: $\beta = 0$, if	The same as that of men

$$F = \frac{MS(Regression)}{MS(Residual)} > F_{0.95(1,n-2)} \approx 4.08$$

Regression ANOVA
Socioeconomic Environment and Adult Health Example

6. The result:

ANOVA	Men					Women			
Source	df	SS	MS	F		df	SS	MS	F
Regression	1	340.50	340.50	91.29		1	433.30	433.30	70.38
Residual	11	41.04	3.73			11	66.75	6.07	
Total	12	381.54				12	500.05		

7. The conclusion: Reject H_0: $\beta = 0$ since $F > F_{0.95(1,11)} = 4.08$