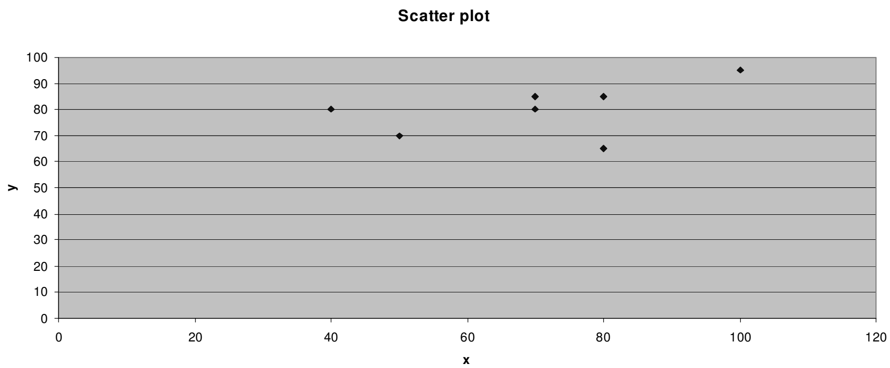
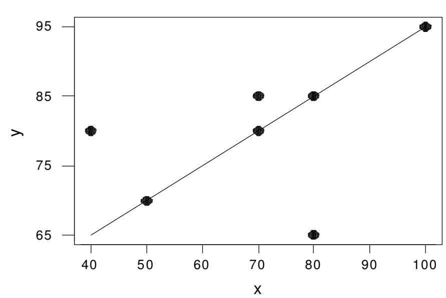
Chapter 15
Correlation and Regression

15.1 **a.**



b. The major axis has slope s_y / s_x and goes through the point $(\bar{x}, \bar{y})$. Here $\bar{x}$ = 70, s_x = 20, $\bar{y}$ = 80, s_y = 10, and n = 7 so an equation for the line along the major axis is
y = 80 + (10/20)·(x – 70) or y = 45 + 0.5x.



c.

Student ID	1	2	3	4	5	6	7	Sum
z_x	–1.5	–1.0	0.0	0.5	0.5	1.5	0.0	0
z_y	0.0	–1.0	0.5	–1.5	0.5	1.5	0.0	0
z_xz_y	0.0	1.0	0.0	–0.75	0.25	2.25	0.0	2.75

15.3

obs	x	y	$(x-\bar{x})^2$	$(y-\bar{y})^2$	z_x	z_y	$z_x z_y$
1	0	2	4	0	–1	0	**0**
2	2	1	0	1	0	–1	**0**
3	4	3	4	1	1	1	**1**
Total	**6**	**6**	**8**	**2**	**0**	**0**	**1**

Here n = 3, $\bar{x}$ = 6/3 = 2, s_x = 2, $\bar{y}$ = 6/3 = 2, s_y = 1, and r = 1/2

The correlation coefficient is 0.5.

15.5

obs	x	y	$(x-\bar{x})^2$	$(y-\bar{y})^2$	z_x	z_y	$z_x z_y$
1	10	7	9	–1.5	–1.5	+1.5	**–2.25**
2	20	5	25	1	–0.5	+0.5	**–0.25**
3	25	4	0	0	0	0	**0.00**
4	25	3	0	1	0	–0.5	**0.00**
5	30	4	25	0	0.5	0	**0.00**
6	40	1	225	9	1.5	–1.5	**–2.25**
Total	**150**	**24**	**500**	**20**	**0**	**0**	**–4.75**

So the 5 number summary is $\bar{x}$ = 150/6 = 25, s_x = $\sqrt{500/5}$ = 10, $\bar{y}$ = 24/6 = 4, s_y = $\sqrt{20/5} = 2$, and r =

$$\frac{\sum z_x z_y}{n-1} = \frac{-4.75}{5} = \frac{-19}{20} = -0.95$$

15.9 From Exercise 15.1, the point of averages is $(\bar{x}, \bar{y})$ = (70,80) and the slope of the regression line is

$$r\frac{s_y}{s_x} = \frac{2.75}{6} \times \frac{10}{20} = \frac{11}{48} \approx 0.22917$$

so the regression line is $\hat{y} = 0.22917x + 63.958\overline{3}$ or $\hat{y} = \dfrac{11x + 3070}{48}$.

15.11 **a.** From Exercise 15.3 the point of averages is $(\bar{x}, \bar{y})$ = (2, 2) and the slope of the regression line is rs_y/s_x = (0.5)(0.5) = 0.25. Use the point-slope formula to find the regression line

$$\hat{y} = 2 + \frac{1}{4}(x - 2) \quad \text{or} \quad \hat{y} = \frac{x+6}{4}.$$

b. SSE = 1.50 as seen from the following table:

OBS	x	y	$\hat{y}$	$e = y - \hat{y}$	e^2
1	0	2	1.50	0.50	**0.25**
2	2	1	2.00	–1.00	**1.00**
3	4	3	2.50	0.50	**0.25**
Totals	**6**	**6**	**6.00**	**0.00**	**1.50 = SSE**

15.13 **a.** An equation for the regression line is $\hat{y} = \bar{y} + r\frac{s_y}{s_x}(x - \bar{x})$.

Using values from Exercise 15.5, we have

$$\hat{y} = 4 - 0.95 \cdot \tfrac{2}{10}\,(x - 25)$$

or $\hat{y}$ = 4 – 0.19 (x – 25) or $\hat{y}$ = –0.19x + 8.75 .

b. SSE = 1.95 as seen from the last entry in the following table.

OBS	x	y	$\hat{y}$	$(y - \hat{y})^2$
1	10	7	6.85	**0.0225**
2	20	5	4.95	**0.0025**
3	25	4	4.00	**0.0000**
4	25	3	4.00	**1.0000**
5	30	4	3.05	**0.9025**
6	40	1	1.15	**0.0225**
Totals	**150**	**24**	**24.00**	**1.9500 = SSE**

$$s_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{1.95}{4}} = 0.6982$$

15.15 **a.** Since x is unknown, estimate y by $\bar{y}$ = 18.

b. Since x = 69 = $\bar{x}$, use the regression line and estimate y by $\bar{y}$ = 18.

c. x = 73 is (73 – $\bar{x}$)/s_x = 1.6 standard units. Estimate y to be 1.6r = 1.28 standard units. So predict the forearm length to be $\bar{y}$ + 1.28s_y = 19.28 inches.

d. x = 64 is (64 – $\bar{x}$)/s_x = –2 standard units. Now estimate the forearm length to be –2r = –1.6 standard units. Predict the forearm length to be $\bar{y}$ – 1.6s_y = 16.40 inches.

15.17 **a.** 19 inches is (19 – 18)/1 = 1 standard unit. The area under the standard normal curve to the right of 1 is 0.15866 (use Table 3). So 15.87% of the men have forearm length over 19 inches.

b. Use the result from Exercise 15.15c. Men who are 73 inches tall have an average forearm length of 19.28 inches with standard deviation $\sqrt{1-r^2}\,s_y$ = 0.60 inches. So 19 inches is (19 – 19.28)/0.60 = –0.4667 standard units. The area under the standard normal curve to the right of –0.4667 is 0.67963. So 67.96% of the men who are 73 inches tall have forearm lengths over 19 inches.

15.19 Use the appropriate regression line to compute. The average height of men whose wives are 68 inches tall is 68 + 0.35 $\left(\frac{68-64}{2.5}\right)$2.7 = 69.512 inches. The standard deviation of these heights is $\sqrt{1-r^2}\,s_x$ = 2.5292 inches. So estimate the height of the man to be 69.51 inches. Assuming their heights follow a normal curve, about 95% of these men will be within 1.96 standard deviations of average. This gives a prediction interval of 69.51 ± 1.96×2.53 or 69.51 ± 4.96 inches. So a 95% prediction interval for the heights of these men is 64.55 inches to 74.47 inches.

15.21 The husbands IQ is (115 – 100)/15 = 1 standard unit. The average IQ of the wives is 1 × r = 0.5 standard units. This is 100 + 0.5 × 15 = 107.5 IQ units. It is not equal to 130 because there are two regression lines, one for predicting IQ's of wives from the IQ of the husbands, and the other for predicting IQ's of husbands from the IQ of their wife.

15.25 **a.** A 95% confidence interval for β is

$b \pm t_{5,.975}\, SE_b$ or 0.04 ± 0.146. So, 95%C.I.: –0.106 < β < 0.186. Here

$$b = r\frac{s_y}{s_x} = \frac{.3\cdot 2}{15} = -0.04$$

$t_{.975,\,5}$ = 2.5706,

$$SE_b = \frac{b}{r}\sqrt{\frac{1-r^2}{n-2}} = \frac{0.04}{0.30}\sqrt{\frac{1-0.09}{5}} = 0.05688$$

.

b. Since the 95% confidence for β includes both positive and negative values, the hypothesis that β (or ρ) is positive is not supported at the 5% level.
Test H_o : β = 0 versus the alternative hypothesis Ha: β > 0. The P-value is

$$\Pr(b \geq 0.04) = \Pr(T_5 \leq \frac{0.04-0}{0.05688}) = \Pr(T_5 \geq 0.70323) \approx 0.257.$$

Since the P-value is not small, there is insufficient evidence to conclude β > 0.

15.27 **a.** From Exercise 15.2, n = 6, r = 2/5, and $s_y/s_x = 0.5$ so

$$b = r\frac{s_y}{s_x} = \frac{1}{5} \quad \text{and} \quad SE_b = \frac{b}{r}\sqrt{\frac{1-r^2}{n-2}} = \frac{1}{2}\sqrt{0.21} \approx 0.22913$$

Also $t_{4,.95} = 2.1318$ (Using Table 4). A 90% confidence interval for β has limits b ± $t_{4,.95}$ SE_b or 0.2000 ± 0.4885 so 90% C.I.: –0.29 < β < 0.69.

b. The two-sided 90% confidence interval includes negative values for β, so there is insufficient evidence to conclude ρ > 0 (or β > 0) at the 5% significance level. To test H_o: β = 0 versus the alternative hypothesis H_a: β > 0, the P-value is

$$\Pr(b \geq \tfrac{1}{5}) = \Pr(T_4 \leq \frac{0.2}{0.22913}) = \Pr(T_4 \geq 0.87287) \approx 0.22$$

15.33 **a.** The regression line goes through $(\bar{x}, \bar{y})$ = (75,85) with slope $rs_y/s_x = 0.60$. An equation for the regression line is $\hat{y}$ = 0.6x + 40.

b. If x = 63, predict y to be on the regression line. Thus, predict y to be $\hat{y}$ = 77.8.

c. A 95% confidence interval for β has bounds b ± $t_{6,.075}$ SE_b with b = 0.60,

$$SE_b = \frac{b}{r}\sqrt{\frac{1-r^2}{n-2}} = 0.4861$$

and $t_{6,0.975} = 2.4469$ (using Table 4).
So the limits are β = 0.6± 1.189 and the interval is –0.59 < β < 1.79.

d. Since the 95% confidence interval for β includes both negative and positive values, the alternative hypothesis β > 0 (or ρ > 0) is not accepted at the 2.5% significance level. The P-value for H_o: β = 0 versus H_a: β > 0 is

$$\Pr(b > .6 \mid H_o) = \Pr(T_6 > \frac{.6-0}{0.4861}) = \Pr(T_6 > 1.2343) \approx 0.13$$

The P-value is large, so the null hypothesis ρ = 0 (equivalent to β = 0) is retained at the 5% significance level (because P-value > α).

15.35

Specimen	Humerus (x)	Femur (y)	$(x-\bar{x})^2$	$(y-\bar{y})^2$	z_x	z_y	$z_x z_y$
1	312	430	4	1.96	0.1139	–0.073	–0.0083
2	335	458	625	707.56	1.4234	1.3889	1.9769
3	286	407	576	595.36	–1.3664	–1.2740	1.7408
4	312	440	4	73.96	0.1139	0.4490	0.0511
5	305	422	25	88.36	–0.2847	–.4908	0.1397
Totals	1550	2157	1234	1467.20	0	0	3.9002
Average	310	431.4					

$$s_x = \sqrt{\frac{1234}{4}} \qquad s_y = \sqrt{\frac{1467.2}{4}}$$

a.

r = $$\bar{z}_{xy} = \frac{3.9002}{4} = 0.97505$$

b. The major axis goes through the point $(\bar{x}, \bar{y})$ = (310, 431.4) and has slope s_y/s_x = 1.0904. Thus, y = 431.4 + 1.0904 (x – 310).

c. The regression line goes through the point $(\bar{x}, \bar{y})$ = (310, 431.4) and has slope rs_y/s_x = 1.0632. Thus $\hat{y}$ = 431.4 + 1.0632 (x – 310).

d. A 95% confidence interval for β is b ± $t_{3,.975}$ SE_b = 1.0632 ± 0.4448.
Here b = 1.0632,

$$SE_b = \frac{b}{r}\sqrt{\frac{1-r^2}{n-2}} = 0.13975$$

and $t_{3,0.975}$ = 3.1824. So the 95% confidence interval is 0.62 < β < 1.51.