Prediction

A 95% confidence interval for $\mu_{Y|X=x^*}$ is given by

$$\hat{y} \pm t_{0.025} s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$$

A 95% prediction interval for $Y|X = x^*$ is given by

$$\hat{y} \pm t_{0.025} s_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$$

Critical values for intervals are determined from Student's t distribution with $df = n - 2$.

Supplementary Exercises 12.S.1–12.S.30

12.S.1 In a study of the Mormon cricket (*Anabrus simplex*), the correlation between female body weight and ovary weight was found to be $r = 0.836$. The standard deviation of the ovary weights of the crickets was 0.429 g. Assuming that the linear model is applicable, estimate the standard deviation of ovary weights of crickets whose body weight is 4 g.[33]

12.S.2 In a study of crop losses due to air pollution, plots of Blue Lake snap beans were grown in open-top field chambers, which were fumigated with various concentrations of sulfur dioxide. After a month of fumigation, the plants were harvested, and the total yield of bean pods was recorded for each chamber. The results are shown in the table.[34]

	X = Sulfur dioxide concentration (ppm)			
	0	0.06	0.12	0.30
Y = yield (kg)	1.15	1.19	1.21	0.65
	1.30	1.64	1.00	0.76
	1.57	1.13	1.11	0.69
Mean	1.34	1.32	1.11	0.70

Preliminary calculations yield the following results.

$$\bar{x} = 0.12 \qquad \bar{y} = 1.117$$
$$s_X = 0.11724 \qquad s_Y = 0.31175$$
$$r = -0.8506 \qquad \text{SS(resid)} = 0.2955$$

(a) Calculate the linear regression of Y on X.

(b) Plot the data and draw the regression line on your graph.

(c) Calculate s_e. What are the units of s_e?

12.S.3 Refer to Exercise 12.S.2.

(a) Assuming that the linear model is applicable, find estimates of the mean and the standard deviation of yields of beans exposed to 0.24 ppm of sulfur dioxide.

(b) Is the estimate in part (a) an interpolation or extrapolation? How can you tell?

(c) Which condition of the linear model appears doubtful for the snap bean data?

12.S.4 Refer to Exercise 12.S.2. Consider the null hypothesis that sulfur dioxide concentration has no effect on yield.

(a) Assuming that the linear model holds, formulate this as a hypothesis about the true regression line.

(b) Write a directional alternative, in symbols, that says increasing sulfur dioxide tends to decrease yield.

(c) How many degrees of freedom are there for the test that compares the hypotheses in (a) and (b)?

(d) The sample slope, b_1, is -2.262 and the standard error of the slope is 0.4421. Compute the value of the test statistic.

(e) The P-value for the test is 0.0002. If $\alpha = 0.05$, state your conclusion regarding the null hypothesis in the context of this setting.

12.S.5 Another way to analyze the data of Exercise 12.S.2 is to take each treatment mean as the observation Y; then the data would be summarized as in the accompanying table.

	Sulfur dioxide X (ppm)	Mean yield Y (kg)
	0.00	1.34
	0.06	1.32
	0.12	1.11
	0.30	0.70
Mean	0.1200	1.1175
SD	0.12961	0.29714
	$r = -0.98666$	
	SS(resid) = 0.007018	

(a) For the regression of mean yield on X, calculate the regression line and the residual standard deviation, and compare with the results of Exercise 12.S.2. Explain why the discrepancy is not surprising.

(b) What proportion of the variability in mean yield is explained by the linear relationship between mean yield and sulfur dioxide? Using the data in Exercise 12.S.2, what proportion of the variability in individual chamber yield is explained by the linear relationship between individual chamber yield and sulfur dioxide? Explain why the discrepancy is not surprising.

12.S.6 In a study of the tufted titmouse (*Parus bicolor*), an ecologist captured seven male birds, measured their wing lengths and other characteristics, and then marked and released them. During the ensuing winter, he repeatedly observed the marked birds as they foraged for insects and seeds on tree branches. He noted the branch diameter on each occasion, and calculated (from 50 observations) the average branch diameter for each bird. The results are shown in the table.[35]

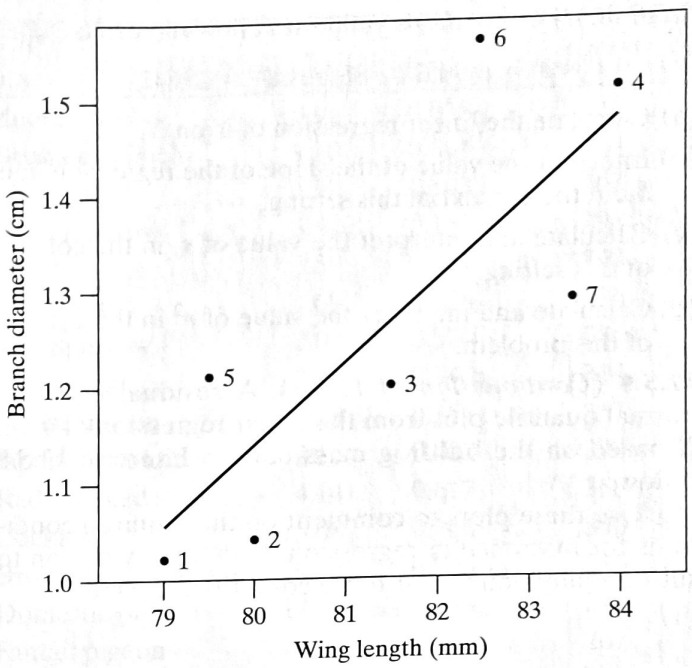
Bird	Wing length X (mm)	Branch diameter Y (cm)
1	79.0	1.02
2	80.0	1.04
3	81.5	1.20
4	84.0	1.51
5	79.5	1.21
6	82.5	1.56
7	83.5	1.29
Mean	81.429	1.2614
SD	1.98806	0.21035
	$r = 0.80335$	
	SS(resid) = 0.09415	

(a) Calculate s_e and specify the units. Verify the approximate relationship between s_Y and s_e, and r.

(b) Do the data provide sufficient evidence to conclude that the diameter of the forage branches chosen by male titmice is correlated with their wing length? Test an appropriate hypothesis against a nondirectional alternative. Let $\alpha = 0.05$.

(c) The test in part (a) was based on 7 observations, but each branch diameter value was the mean of 50 observations. If we were to test the hypothesis of part (a) using the raw numbers, we would have 350 observations rather than only 7. Why would this approach not be valid?

12.S.7 (*Continuation of 12.S.6*) A scatterplot and fitted regression line of the data from Exercise 12.S.6 follow. The individual birds are labeled in the plot.



(a) Which bird/point has the largest regression residual?

(b) Which bird(s)/points(s) have the most leverage?

(c) Are there any birds/points that are influential?

(d) Invent your own bird observation of x = wing length and y = branch diameter that would be an example of a regression outlier, but not an influential observation.

(e) Invent your own bird observation of x = wing length and y = branch diameter that would be an example of a leverage point that is not influential.

(f) Invent your own bird observation of x = wing length and y = branch diameter that would an example of an influential point.

12.S.8 Exercise 12.3.7 deals with data on the relationship between body length and jumping distance of bullfrogs. A third variable that was measured in that study was the mass of each bullfrog. The following table shows these data.[16]

Bullfrog	Length X (mm)	Mass Y (g)
1	155	404
2	127	240
3	136	296
4	135	303
5	158	422
6	145	308
7	136	252
8	172	533.8
9	158	470
10	162	522.9
11	162	356
Mean	149.636	373.427
SD	14.4725	104.2922

Preliminary calculations yield the following results:

$$r = 0.90521 \quad \text{SS(resid)} = 19642$$

(a) Calculate the linear regression of Y on X.

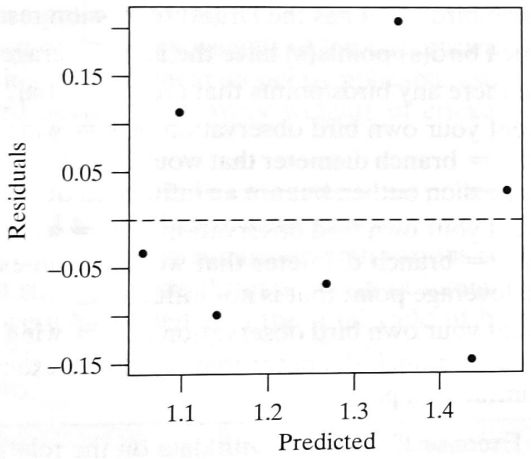
(b) Interpret the value of the slope of the regression line, b_1, in the context of this setting.

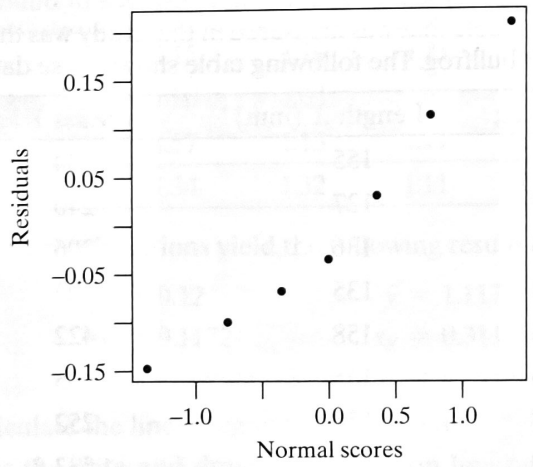
(c) Calculate and interpret the value of s_e in the context of this setting.

(d) Calculate and interpret the value of r^2 in the context of this problem.

12.S.9 (*Continuation of 12.S.8*). A residual plot and normal quantile plot from the linear regression of Y on X based on the bullfrog mass data in Exercise 12.S.8 follow.

Use these plots to comment on the required conditions for inference in regression. Is there any reason to substantially doubt that these conditions are met?



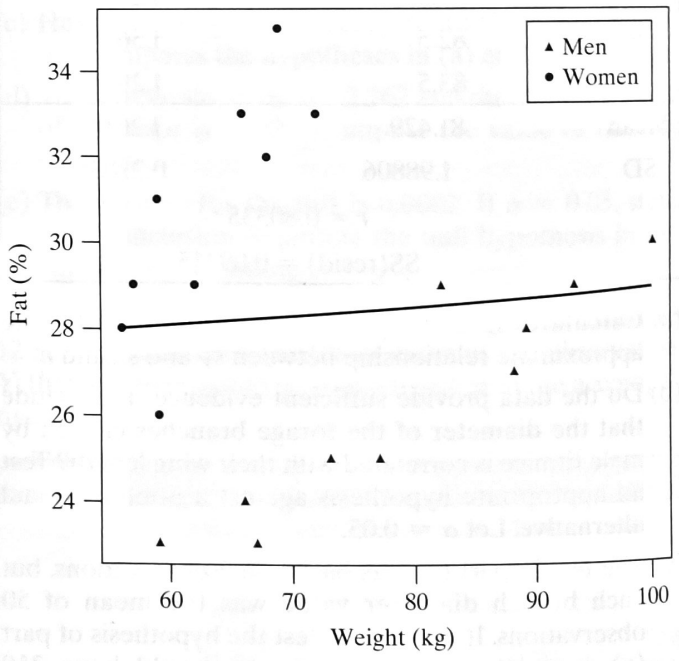


12.S.10 An exercise physiologist used skinfold measurements to estimate the total body fat, expressed as a percentage of body weight, for 19 participants in a physical fitness program. The body fat percentages and the body weights are shown in the table.[36]

Participant	Weight X (kg)	Fat Y (%)	Participant	Weight X (kg)	Fat Y (%)
1	89	28	11	57	29
2	88	27	12	68	32
3	66	24	13	69	35
4	59	23	14	59	31
5	93	29	15	62	29
6	73	25	16	59	26
7	82	29	17	56	28
8	77	25	18	66	33
9	100	30	19	72	33
10	67	23			

Actually, participants 1 to 10 are men, and participants 11 to 19 are women. A summary and graph of the data for men, women, and both sexes combined into a single sample follow.

Men ($n = 10$)	Women ($n = 9$)	Both sexes ($n = 19$)
$\bar{x} = 79.40$	$\bar{x} = 63.1$	$\bar{x} = 71.68$
$\bar{y} = 26.30$	$\bar{y} = 30.67$	$\bar{y} = 28.37$
$s_X = 13.2430$	$s_X = 5.7975$	$s_X = 13.1320$
$s_Y = 2.6269$	$s_Y = 2.8723$	$s_Y = 3.4835$
$r = 0.9352$	$r = 0.8132$	$r = 0.0780$



(a) Compute the regression equations for the males and females separately.

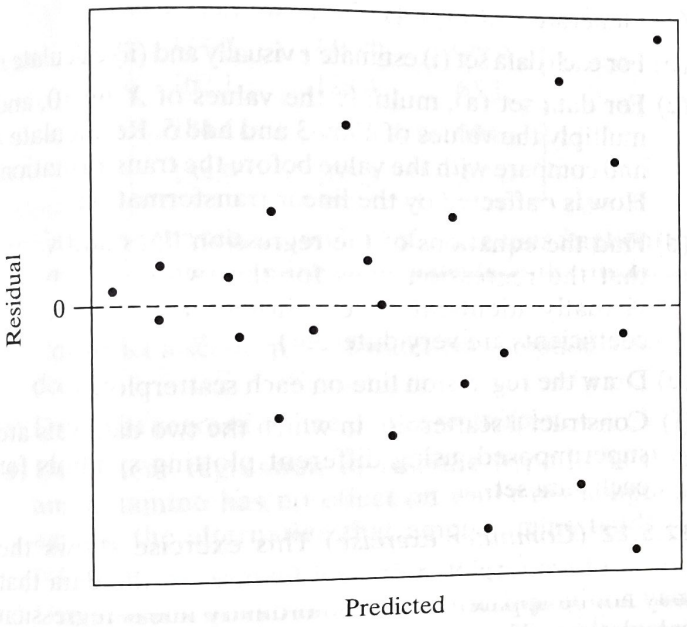
(b) The equation to the fitted regression line for both sexes combined, which is shown on the plot, is $\hat{y} = 26.88 + 0.021x$. How does the slope of this line

compare to the slopes you computed in part (a)? Can you explain the discrepancy?

(c) Examine the correlation coefficients for (i) the males, (ii) the females, and (iii) both sexes combined. Do these values agree with your reasoning provided in part (b)?

12.S.11 Refer to the respiration rate data of Exercise 12.3.7. Construct a 95% confidence interval for β_1.

12.S.12 The following plot is a residual plot from fitting a regression model to some data. Make a sketch of the scatterplot of the data that led to this residual plot. (*Note*: There are two possible scatterplots—one in which b_1 is positive and one in which b_1 is negative.)



12.S.13 Biologists studied the relationship between embryonic heart rate and egg mass for 20 species of birds. They found that heart rate, Y, has a linear relationship with the logarithm of egg mass, X. The data are given in the following table.[37]

Species	Egg mass (g)	Log-(egg mass) X	Heart rate Y (beats/min)
Zebra finch	0.96	−0.018	335
Bengalese finch	1.10	0.041	404
Marsh tit	1.39	0.143	363
Bank swallow	1.42	0.152	298
Great tit	1.59	0.201	348
Varied tit	1.69	0.228	356
Tree sparrow	2.09	0.320	335
Budgerigar	2.19	0.340	314
House martin	2.25	0.352	357
Japanese bunting	2.56	0.408	370
Red-cheeked starling	4.14	0.617	358
Cockatiel	5.08	0.706	300
Brown-eared bulbul	6.40	0.806	333
Domestic pigeon	17.10	1.233	247
Fantail pigeon	19.70	1.294	267
Homing pigeon	19.80	1.297	230
Barn owl	20.10	1.303	219
Crow	20.50	1.312	297
Cattle egret	27.50	1.439	251
Lanner falcon	41.20	1.615	242
Mean	9.94	0.690	311

For these data the fitted regression equation is

$$\hat{y} = 368.06 - 82.452x$$

and

$$\text{SS(resid)} = 15748.6$$

(a) Interpret the value of the intercept of the regression line, b_0, in the context of this setting.

(b) Interpret the value of the slope of the regression line, b_1, in the context of this setting.

(c) Calculate s_e and specify the units.

(d) Interpret the value of s_e in the context of this setting.

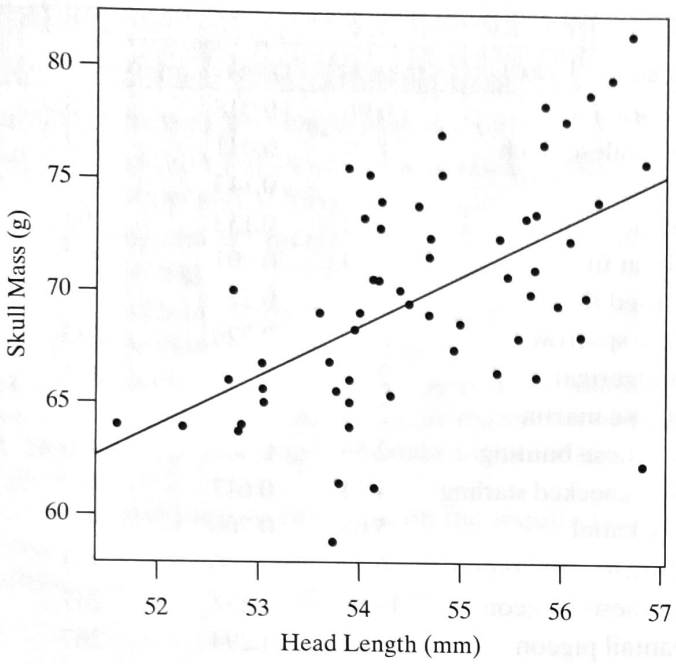
12.S.14 An ornithologist measured the mass (g) and head length (the distance from the tip of the bill to the back of the skull, in mm) for a sample of 60 female blue jays. Here is a plot of the data and computer output:

```
Coefficients:
             Estimate  Std. Error  t value     Pr(>|t|)
(Intercept) -50.6960     24.1843   -2.096       0.0404 *
Head          2.2052      0.4425    4.984 5.95e-06 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1 Residual standard error: 4.23 on 58 degrees of
freedom
Multiple R-squared: 0.2999,Adjusted R-squared: 0.2878
F-statistic: 24.84 on 1 and 58 DF, p-value: 5.954e-06
```

(a) Interpret the fitted slope of the regression model, in the context of this setting.

(b) The computer output says that the residual standard error (which we call the *residual standard deviation*) is 4.23. In the context of this setting, what does that mean? (Be sure to state the units for 4.23 as part of your answer.)

(c) The computer output says that R^2 is 0.2999. In the context of this setting, what does that mean?



12.S.15 Consider the study and regression output in Exercise 12.S.14. The *P*-value given on the "Head" line is 0.00000595.

(a) What hypothesis is being tested using this *P*-value? State your answer symbolically and in plain English.

(b) What conditions are necessary for the *P*-value to be trustworthy?

12.S.16 Consider the study and regression output in Exercise 12.S.14. Construct a 95% confidence interval for the population value of the slope.

12.S.17 Consider the study and regression output in Exercise 12.S.14. Use the regression model to predict the mass of a blue jay with head measurement of 56 mm.

12.S.18 Consider the study and regression output in Exercise 12.S.14. Use the regression model to predict the mean and SD of the masses for blue jays with heads that are 53 mm in length.

12.S.19 Consider the study and regression output in Exercise 12.S.14. Sadly, an ornithologist's cat brought in just the head of a blue jay. The head length was 47 mm. What would you predict the mass of the bird to have been? Is your prediction trustworthy? Explain.

12.S.20 (*Challenge question*) Consider the study and regression output in Exercise 12.S.14. Using only the numeric output to support your answer, would it be unusual for a female blue jay with a head length of 52 mm to weigh less than 54 g?

12.S.21 (*Computer exercise*) The accompanying table gives two data sets: (A) and (B). The values of X are the same for both data sets and are given only once.

	(A)	(B)		(A)	(B)
X	Y	Y	X	Y	Y
0.61	0.88	0.96	2.56	1.97	1.20
0.93	1.02	0.97	2.74	2.02	3.59
1.02	1.12	0.07	3.04	2.26	3.09
1.27	1.10	2.54	3.13	2.27	1.55
1.47	1.44	1.41	3.45	2.43	0.71
1.71	1.45	0.84	3.48	2.57	3.05
1.91	1.41	0.32	3.79	2.53	2.54
2.00	1.59	1.46	3.96	2.73	3.33
2.27	1.58	2.29	4.12	2.92	2.38
2.33	1.66	2.51	4.21	2.96	3.08

(a) Generate scatterplots of the two data sets.

(b) For each data set (i) estimate r visually and (ii) calculate r.

(c) For data set (a), multiply the values of X by 10, and multiply the values of Y by 3 and add 5. Recalculate r and compare with the value before the transformation. How is r affected by the linear transformation?

(d) Find the equations of the regression lines and verify that the regression lines for the two data sets are virtually identical (even though the correlation coefficients are very different).

(e) Draw the regression line on each scatterplot.

(f) Construct a scatterplot in which the two data sets are superimposed, using different plotting symbols for each data set.

12.S.22 (*Computer exercise*) This exercise shows the power of scatterplots to reveal features of the data that may not be apparent from the ordinary linear regression calculations. The accompanying table gives three fictitious data sets, A, B, and C. The values of X are the same for each data set, but the values of Y are different.[38]

Data set:	A	B	C
X	Y	Y	Y
10	8.04	9.14	7.46
8	6.95	8.14	6.77
13	7.58	8.74	12.74
9	8.81	8.77	7.11
11	8.33	9.26	7.81
14	9.96	8.10	8.84
6	7.24	6.13	6.08
4	4.26	3.10	5.39
12	10.84	9.13	8.15
7	4.82	7.26	6.42
5	5.68	4.74	5.73

(a) Verify that the fitted regression line is almost exactly the same for all three data sets. Are the residual standard deviations the same? Are the values of r the same?

(b) Construct a scatterplot for each of the data sets. What does each plot tell you about the appropriateness of linear regression for the data set?

(c) Plot the fitted regression line on each of the scatterplots.

12.S.23 (*Computer exercise*) In a pharmacological study, 12 rats were randomly allocated to receive an injection of amphetamine at one of two dosage levels or an injection of saline. Shown in the table is the water consumption of each animal (ml water per kg body weight) during the 24 hours following injection.[39]

Dose of amphetamine (ml/kg)		
0	1.25	2.5
122.9	118.4	134.5
162.1	124.4	65.1
184.1	169.4	99.6
154.9	105.3	89.0

(a) Calculate the regression line of water consumption on dose of amphetamine, and calculate the residual standard deviation.

(b) Construct a scatterplot of water consumption against dose.

(c) Draw the regression line on the scatterplot.

(d) Use linear regression to test the hypothesis that amphetamine has no effect on water consumption against the alternative that amphetamine tends to reduce water consumption. (Use $\alpha = 0.05$.)

(e) Use analysis of variance to test the hypothesis that amphetamine has no effect on water consumption. (Use $\alpha = 0.05$.) Compare with the result of part (d).

(f) What conditions are necessary for the validity of the test in part (d) but not for the test in part (e)?

(g) Calculate the pooled standard deviation from the ANOVA, and compare it with the residual standard deviation calculated in part (a).

12.S.24 (*Computer exercise*) Consider the Amazon tree data from Exercise 12.6.9. The researchers in this study were interested in how age, Y, is related to X = "growth rate," where growth rate is defined as diameter/age (i.e., cm of growth per year).

(a) Create the variable "growth rate" by dividing each diameter by the corresponding tree age.

(b) Make a scatterplot of Y = age versus X = growth rate and fit a regression line to the data.

(c) Make a residual plot from the regression in part (b). Then make a normal quantile plot of the residuals. How do these plots call into question the use of a linear model and regression inference procedures?

(d) Take the logarithm of each value of age and of each value of growth rate. Make a scatterplot of $Y = \log(\text{age})$ versus $X = \log(\text{growth rate})$ and fit a regression line to the data.

(e) Make a residual plot from the regression in part (d). Then make a normal quantile plot of the residuals. Based on these plots, does a regression model in log scale, from part (d), seem appropriate?

12.S.25 (*Computer exercise*) Researchers measured the blood pressures of 22 students in two situations: when the students were relaxed and when the students were taking an important examination. The table lists the systolic and diastolic pressures for each student in each situation.[40]

During exam		Relaxed	
Systolic pressure (mm Hg)	Diastolic pressure (mm Hg)	Systolic pressure (mm Hg)	Diastolic pressure (mm Hg)
132	75	110	70
124	170	90	75
110	65	90	65
110	65	110	80
125	65	100	55
105	70	90	60
120	70	120	80
125	80	110	60
135	80	110	70
105	80	110	70
110	70	85	65
110	70	100	60
110	70	120	80
130	75	105	75
130	70	110	70
130	70	120	80
120	75	95	60
130	70	110	65
120	70	100	65
120	80	95	65
120	70	90	60
130	80	120	70

(a) Compute the change in systolic pressure by subtracting systolic pressure when relaxed from systolic pressure during the exam; call this variable X.

(b) Repeat part (a) for diastolic pressure. Call the resulting variable Y.

(c) Make a scatterplot of Y versus X and fit a regression line to the data.

(d) Make a residual plot from the regression in part (c).

(e) Note the outlier in the residual plot [and on the scatterplot from part (c)]. Delete the outlier from the data set. Then repeat parts (c) and (d).

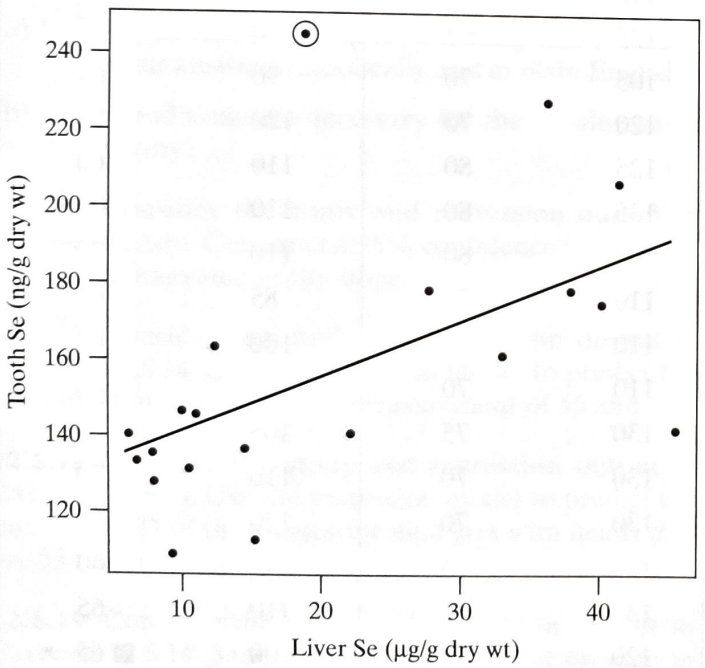
(f) What is the fitted regression model (after the outlier has been removed)?

12.S.26 (*Continuation of 12.S.25*) Consider the data from Exercise 12.S.25, part (f).

(a) Construct a 95% confidence interval for β_1.

(b) Interpret the confidence interval from part (a) in the context of this setting.

12.S.27 Selenium (Se) is an essential element that has been shown to play an important role in protecting marine mammals against the toxic effects of mercury (Hg) and other metals. It has been suggested that metal concentrations in marine mammal teeth can potentially be used as bioindicators for body burden. Twenty Belugas (*Delphinapterus leucas*) were harvested from the Mackenzie Delta, Northwest Territories, in 1996 and 2002, as part of an annual traditional Inuit hunt. Tooth and liver Se concentrations are reported in the table, summarized, and graphed.[41]



(a) Can we regard the sample correlation between Tooth (Y) and Liver (X) selenium, $r = 0.53726$, as an estimate of the population correlation coefficient? Briefly explain.

(b) If the circled point were removed from the data set, would the sample correlation listed in part (a) increase, decrease, or stay about the same?

Whale	Liver Se (µg/g)	Tooth Se (ng/g)	Whale	Liver Se (µg/g)	Tooth Se (ng/g)
1	6.23	140.16	11	15.28	112.63
2	6.79	133.32	12	18.68	245.07
3	7.92	135.34	13	22.08	140.48
4	8.02	127.82	14	27.55	177.93
5	9.34	108.67	15	32.83	160.73
6	10.00	146.22	16	36.04	227.60
7	10.57	131.18	17	37.74	177.69
8	11.04	145.51	18	40.00	174.23
9	12.36	163.24	19	41.23	206.30
10	14.53	136.55	20	45.47	141.31

(c) If the roles of X and Y were reversed (i.e., Y = Liver and X = Tooth selenium), would the sample correlation listed in part (a) increase, decrease, or stay about the same?

(d) Is the circled point on the plot a leverage and/or influential point? Explain briefly.

(e) Is the circled point on the plot an outlier?

12.S.28 (*Continuation of 12.S.27*) The following are summary statistics for the selenium data in Exercise 12.S.27.

$$\bar{x} = 20.685 \qquad \bar{y} = 156.599$$
$$s_X = 13.4491 \qquad s_Y = 36.0595$$
$$r = 0.53729 \qquad \text{SS(resid)} = 17{,}573.4$$

(a) Calculate the regression line of Tooth selenium on Liver selenium.

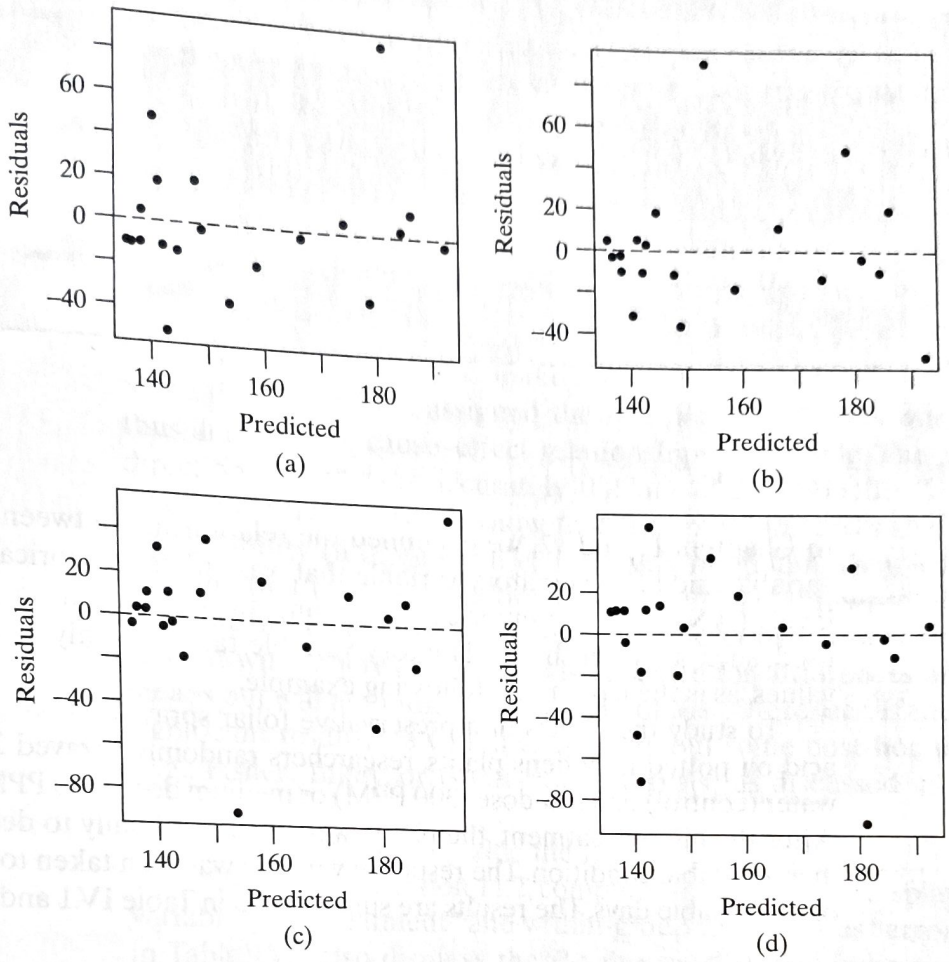
(b) Compute a 95% confidence interval for the slope of the regression line.

(c) Interpret the interval computed in part (b) in the context of the problem.

(d) Using the interval computed in part (b), is it reasonable to believe that the slope is as small as 0.25 (ng/g)/(µg/g)?

12.S.29 (*Continuation of 12.S.27 and 12.S.28*) Referring to the data plotted in Exercise 12.S.27, which of the following is a residual plot resulting from fitting the regression line in Exercise 12.S.28, part (a)? Justify your choice.

12.S.30 (*Continuation of 12.S.27*) The whales observed in this study were harvested during a traditional Inuit hunt in two particular years. What are we assuming about the captured whales to justify our analyses of these data in the preceding problems?


Residuals
60
20
0
−40
140
160
180
Predicted
(a)
Residuals
60
20
0
−40
140
160
180
Predicted
(b)
Residuals
20
0
−40
−80
140
160
180
Predicted
(c)
Residuals
20
0
−40
−80
140
160
180
Predicted
(d)