Chapter 11

REGRESSION AND CORRELATION METHODS

REVIEW OF KEY CONCEPTS

SECTION 11.1 The Linear-Regression Model

A study was performed relating vegetable and fruit consumption to systolic blood pressure (SBP) in a group of 60 lactovegetarians (vegetarians who eat dairy products but not meat, poultry, or fish). The descriptive statistics for these variables are given as follows:

	Mean	sd	*n*	Range	Unit
SBP	109.3	11.9	60	82–144	mm Hg
Vegetables	3.13	2.06	60	0–7.95	Number of servings/day
Fruit	3.41	2.44	60	0–10.63	Number of servings/day

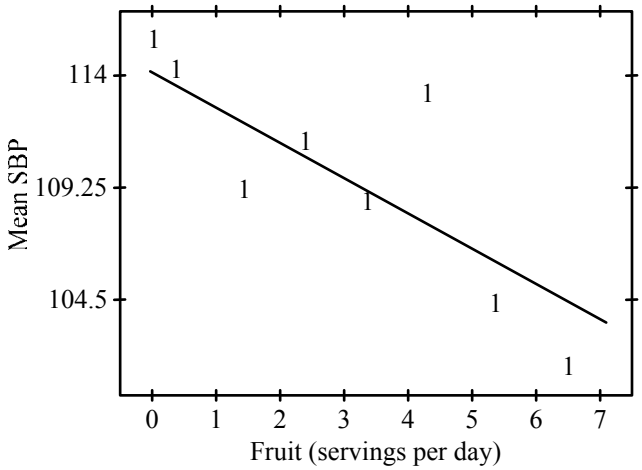
We focus on fruit consumption, subdividing fruit consumption into "one serving per day" groups and computing mean SBP in each fruit-consumption group. These are presented numerically in Table 11.1 and graphically in Figure 11.1.

Table 11.1 Mean systolic blood pressure (SBP) (mm Hg) by number of servings per day of fruit

Fruit consumption (servings per day)	*n*	Mean SBP	Predicted SBP
0.0	8	115.9	114.5
0.01–0.99	4	114.4	113.8
1.00–1.99	6	109.4	112.2
2.00–2.99	8	110.7	110.7
3.00–3.99	12	109.1	109.2
4.00–4.99	6	112.8	107.7
5.00–5.99	7	104.2	106.1
6.0+[a]	9	101.9	104.6

[a] It is assumed that the average fruit consumption in this group is 6.5 servings per day.

FIGURE 11.1 Plot of systolic blood pressure versus fruit consumption (servings per day) in a group of 60 lactovegetarians



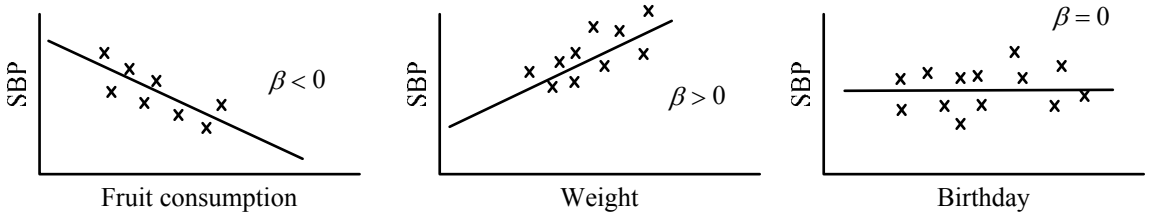
It appears that mean SBP decreases as fruit consumption increases, with the decline being roughly linear. This suggests the following *linear-regression model*:

General Linear-Regression Model

$$y_i = \alpha + \beta x_i + e_i \qquad y_i = \text{SBP } i\text{th person},\ x_i = \text{fruit consumption } i\text{th person}$$
$$e_i \sim N\left(0,\ \sigma^2\right) \text{ error term}$$

The parameter σ^2 is a measure of the variability of the points about the line. If all points fall exactly on a line, then $\sigma^2 = 0$. In general, $\sigma^2 > 0$. The parameters α and β are the intercept and slope of the regression line. The slope is likely to be negative $(\beta < 0)$ for SBP versus fruit consumption, positive $(\beta > 0)$ for SBP versus body weight and zero $(\beta = 0)$ for SBP versus birthday as depicted in Figure 11.2.

FIGURE 11.2 Relationship of SBP to fruit consumption, body weight, and birthday, respectively



SECTION 11.2 Fitting Regression Lines—The Method of Least Squares

We use the *method of least squares* to estimate the parameters of a regression line. The coefficients *a* and *b* are estimated parameters, corresponding to the true parameters α and β. We find *a* and *b* so as to minimize *S*, where

$$S = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2 = \sum_{i=1}^{n} d_i^2$$

In words, S is the sum of squares of the vertical deviations of the points from the line. The least-squares estimates are

$$b = \frac{L_{xy}}{L_{xx}},\ a = \bar{y} - b\bar{x} = \frac{\sum y_i - b\sum x_i}{n}$$

where L_{xy} = corrected sum of cross-products given by

$$L_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$

L_{xx} = corrected sum of squares for x given by

$$L_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

L_{yy} = corrected sum of squares for y given by

$$L_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

The line $y = a + bx$ is called the *least-squares or regression line*. In this example,

$$\sum y_i = 6559 \qquad \sum y_i^2 = 725{,}343 \quad \sum x_i = 204.83$$
$$\sum x_i^2 = 1051.31 \quad \sum x_i y_i = 21{,}853.86$$
$$L_{yy} = 725{,}343 - \frac{6559^2}{60} = 8334.98$$
$$L_{xx} = 1051.31 - \frac{204.83^2}{60} = 352.05$$
$$L_{xy} = 21{,}853.86 - \frac{204.83(6559)}{60} = -537.47$$
$$b = \frac{-537.47}{352.05} = -1.53 \qquad a = \frac{6559 + 1.53 \times 204.83}{60} = 114.53$$
$$y = 114.53 - 1.53x$$

This is the least-squares or regression line.

SECTION 11.3 Testing for the Statistical Significance of a Regression Line

We wish to test the hypothesis $H_0: \beta = 0$ versus $H_1: \beta \neq 0$. We can subdivide the Total Sum of Squares (SS) $= \sum_{i=1}^{n}(y_i - \bar{y})^2$ into two components called the Regression SS and the Residual SS as shown in the equation:

$$\underset{\text{Total SS}}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = \underset{\text{Residual SS}}{\sum_{i=1}^{n}(y_i-\hat{y})^2} + \underset{\text{Regression SS}}{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}$$

where $\hat{y}_i$ is the predicted value of y for a specific value of x, given by $a+bx_i$. The Residual SS (or S in Section 11.2, study guide) is a measure of how close the individual sample points are from the regression line. The Regression SS is a measure of how far the slope of the regression line is from zero (in absolute value). For a good-fitting regression line, we want a high Regression SS and a low Residual SS. We use $\text{Regression SS}/\text{Residual SS}$ as a test criterion to test the hypothesis that $\beta=0$ and reject H_0 for large values of Regression SS/Residual SS. For convenience, the test statistic is usually formulated in terms of the Regression Mean Square (MS) and the Residual Mean Square (MS) defined by

$$\text{Regression MS} = \frac{\text{Regression SS}}{1}$$

$$\text{Residual MS} = \frac{\text{Residual SS}}{n-2}$$

The test statistic is given by $F = \text{Regression MS}/\text{Residual MS} \sim F_{1,\,n-2}$ under H_0. The p-value is

$$p\text{-value} = Pr\left(F_{1,\,n-2} > F\right).$$

This is called the *F test for simple linear regression*.

We can arrange the results in the form of an ANOVA table as follows:

	SS	*df*	MS	*F*-stat	*p*-value
Regression					
Residual					
Total					

11.3.1 Short Computational Form for the *F* Test

The Regression SS, Residual SS and test statistic F can also be written in the more computationally convenient forms

$$\text{Regression SS} = b^2L_{xx} = bL_{xy} = \frac{L_{xy}^2}{L_{xx}}$$

$$\text{Residual SS} = \text{Total SS} - \text{Regression SS} = L_{yy} - \frac{L_{xy}^2}{L_{xx}}$$

$$F = \frac{\dfrac{L_{xy}^2}{L_{xx}}}{\dfrac{\left(L_{yy}-L_{xy}^2/L_{xx}\right)}{n-2}} \sim F_{1,\,n-2} \text{ under } H_0$$

In our example, $L_{xy} = -537.47$, $L_{xx} = 352.05$, $L_{yy} = 8334.98$. Thus,

$$\text{Regression SS} = \frac{(-537.47)^2}{352.05} = 820.55$$

$$\text{Residual SS} = 8334.98 - 820.55 = 7514.44$$

$$\text{Residual MS} = \frac{7514.44}{58} = 129.56$$

$$F = \frac{\text{Regression MS}}{\text{Residual MS}} = \frac{820.55}{129.56} = 6.33 \sim F_{1,\,58}$$

Because

$$F_{1,\,40,\,.975} = 5.42,\ F_{1,\,40,\,.99} = 7.31,\ 5.42 < 6.33 < 7.31$$

it follows that if we had 40 *df* for the Residual MS, then $1 - .99 < p < 1 - .975$ or $.01 < p < .025$. Also, because

$$F_{1,\,60,\,.975} = 5.29,\ F_{1,\,60,\,.99} = 7.08,\ 5.29 < 6.33 < 7.08$$

it follows that if we had 60 *df* for the Residual MS, then $1 - .99 < p < 1 - .975$ or $.01 < p < .025$. Therefore, since we have 58 Residual *df* and $40 < 58 < 60$, it follows that $.01 < p < .025$, which implies that there is a significant relationship between fruit consumption and systolic blood pressure. The results can be displayed in an ANOVA table as follows:

	SS	*df*	MS	*F*-stat	*p*-value
Regression	820.55	1	820.55	6.33	$.01 < p < .025$
Residual	7514.44	58	129.56		
Total	8334.98	59			

SECTION 11.4 The *t* Test Approach to Significance Testing for Linear Regression

Another approach to hypothesis testing is based on the estimated slope *b*. It can be shown that

$$se(b) = \sqrt{\frac{\text{Residual MS}}{L_{xx}}}$$

We also denote Residual MS by $s^2_{y \cdot x}$ = variance of *y* for a given *x* as opposed to the unconditional variance of $y = s^2_y$. We use the test statistic

$$t = \frac{b}{se(b)} \sim t_{n-2} \text{ under } H_0$$

The *p*-value $= 2 \times Pr(t_{n-2} > t)$ if $t > 0$, or $2 \times Pr(t_{n-2} < t)$ if $t \le 0$. This test is called the *t test for simple linear regression*. In our case,

$$b = -1.53,\ se(b) = \sqrt{\frac{129.56}{352.05}} = 0.61$$

$$t = \frac{b}{se(b)} = \frac{-1.53}{0.61} = -2.517 \sim t_{58} \text{ under } H_0$$

The *p*-value $= 2 \times Pr(t_{58} < -2.517) = .015$.

The F test for simple linear regression and the t test for simple linear regression are equivalent in the sense that $t^2 = F$. In this case, $t^2 = (-2.517)^2 = 6.33 = F$. Also, the p-values obtained from either test are the same.

SECTION 11.5 Interval Estimation for Linear Regression

We can obtain interval estimates for the parameters of a regression line. A 95% CI for β is given by

$$\begin{aligned} b \pm t_{n-2,\ .975} se(b) &= -1.53 \pm t_{58,\ .975}(0.61) \\ &= -1.53 \pm 2.001(0.61) \\ &= -1.53 \pm 1.21 = (-2.74,\ -0.31) \end{aligned}$$

A 95% CI for α is given by $a \pm t_{n-2,\ 0.975} se(a)$, where

$$se(a) = \sqrt{s_{y \cdot x}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{L_{xx}} \right)}$$

The intercept is usually more of interest in the comparison of two regression lines.

We can also obtain interval estimates for predictions made from regression lines. Suppose

$x_1 = x$-value for a new individual not used in constructing the regression line

$\bar{x}$ = mean value for the subjects used in the regression

Then

$\hat{y}$ = average value of y for individuals with x-value $= x_1 = a + bx_1$

There are two types of *se*'s used in predictions from regression lines:

$$se_1(\hat{y}) = \sqrt{s_{y \cdot x}^2 \left[1 + \frac{1}{n} + \frac{(x_1 - \bar{x})^2}{L_{xx}} \right]}$$

= standard error of $\hat{y}$ for a given value of $x(x_1)$ for one individual

$$se_2(\hat{y}) = \sqrt{s_{y \cdot x}^2 \left[\frac{1}{n} + \frac{(x_2 - \bar{x})^2}{L_{xx}} \right]}$$

= standard error of $\hat{y}$ for a given value of $x(x_2)$ where $\hat{y}$ is an average value over many individuals with the value x_2

$se_1(\hat{y})$ is used in making predictions for one specific individual; $se_2(\hat{y})$ is used in making predictions for the average value of y over many individuals with the same value of x.

SECTION 11.6 R^2

A commonly used measure of goodness of fit for a regression line is R^2, which is defined by

$R^2 = \dfrac{\text{Regression SS}}{\text{Total SS}}$ and is interpreted as the proportion of the variance of y that is explained by x

To see this, note that it can be shown algebraically that

$$R^2 \approx 1 - \frac{s^2_{y \cdot x}}{s^2_y}$$

or $R^2 \approx 1 -$ proportion of variance of y that is not explained by $x =$ proportion of variance of y that is explained by x.

If the fit is perfect, then $R^2 = 1$. If x has no relationship to y, then $R^2 = 0$. In our case,

$$R^2 = \frac{820.55}{8334.98} = .10.$$

Only 10% of the variance in y is explained by x. *Caution*: We can have a highly significant regression with a small R^2 due to a large n. Conversely, we can also have a marginally significant regression with a large R^2 if n is small. Thus, a high R^2 and a low p-value are not synonymous.

SECTION 11.7 Assessing Goodness of Fit of Regression Lines

Assumptions Made In Linear Regression

1. **Linearity** Variables must be in the proper scale so that the average value of y for a given value of x is a linear function of x.
2. **Homoscedasticity** Residual variance (i.e., the variance of y for a given x) is the same for all data points, and in particular, is independent of x.
3. **Independence** All residuals are assumed to be independent; e.g., usually, you can't enter multiple data points for the same person because they are not independent. Special methods which account for the dependence must be used in this case.

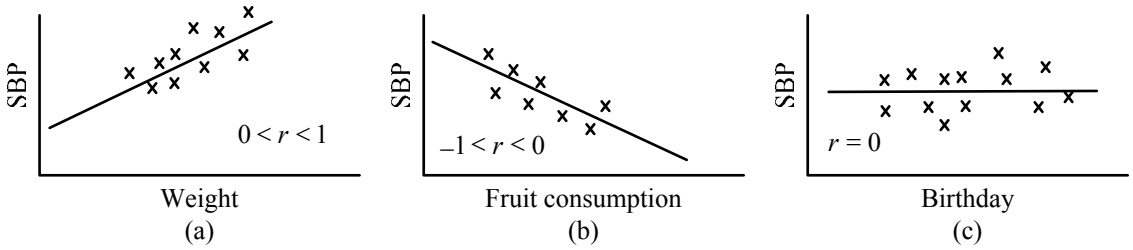
To check the validity of the first two assumptions, we can use *residual analysis*. The *residual* for an individual point (x_i, y_i) is $y_i - \hat{y}_i$, where $\hat{y}_i = a + bx_i =$ fitted value of y for the value $x = x_i$. With computer packages such as SAS or MINITAB, we can plot the residuals versus either x_i or $\hat{y}_i$. If assumptions 1 and 2 hold, then no pattern should be evident in the residual plot. For example, violations of assumption 1 will often result in clumps of positive and/or negative residuals for specific ranges of x. Violations of assumption 2 will often result in greater spread of the residuals for either large or small values of x. Residual analysis is also helpful in identifying *outliers* (points that are very far from the regression line), and *influential points* (points that have an extreme influence on the regression parameter estimates). A point is influential if the regression slope and/or intercept estimates change a lot in magnitude when that point is removed. We may want to remove outliers or influential points before reporting the final analysis (or at least perform the analysis with and without including these points to assess their impact on the results.) It is almost always worthwhile to perform some residual analyses to check for the validity of regression models.

In Figure 11.1, we have plotted the mean SBP versus fruit consumption, where the latter is grouped in one-serving-per-day groups, and have superimposed the fitted regression line. The observed mean values appear to be randomly scattered about the regression line, indicating no particular pattern in the residuals. The actual observed mean and predicted SBP for each fruit-consumption group are also provided in Table 11.1. To compute the predicted value for a group, the midpoint of the group interval was used. For example, for the 1.00–1.99-servings-per-day group, the midpoint $= 1.5$ and the predicted SBP was $114.53 - 1.53(1.5) = 112.2$ mm Hg. It appears that the linearity and homoscedasticity assumptions are reasonable for these data.

SECTION 11.8 The Correlation Coefficient

The sample correlation coefficient r is defined as $L_{xy}/\sqrt{L_{xx}L_{yy}}$. It has the advantage over the regression coefficient of being a dimensionless quantity ($-1 \le r \le 1$), and is not affected by the units of measurement. If $r > 0$, then y tends to increase as x increases [e.g., SBP versus weight in Figure 11.3(a)]; if $r < 0$, then y tends to decrease as x increases [e.g., SBP versus fruit consumption in Figure 11.3(b)]; if $r = 0$, then y is unrelated to x [e.g., SBP versus birthday in Figure 11.3(c)]. In general, r is a more appropriate measure than b if one is interested in whether there is an association between two variables, rather than in predicting one variable from another. Conversely, if prediction is the main interest, then b is the more appropriate measure. The sample correlation coefficient r is the sample analog to the population correlation coefficient ρ introduced in Chapter 5.

FIGURE 11.3 Scatter plot of SBP versus weight, fruit consumption, and birthday, respectively



The relationship between b and r is given by $b = rs_y/s_x$. Notice that b is a rescaled version of r in the units of y divided by x, while r is a dimensionless quantity.

11.8.1 One-Sample Inference

We may wish to test the hypothesis $H_0: \rho = 0$ versus $H_1: \rho \neq 0$, where ρ is the population correlation coefficient. For this purpose, we use the test statistic

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2} \text{ under } H_0$$

The p-value $= 2 \times Pr(t_{n-2} > t)$ if $t > 0$, or $= 2 \times Pr(t_{n-2} < t)$ if $t \le 0$. It is mathematically equivalent to either the F test or the t test for simple linear regression discussed in Sections 11.3 and 11.4 (it always yields the same p-value).

In some instances, we expect a nonzero correlation and are interested in the magnitude (or strength) of the correlation. For example, suppose we know from large studies that the familial correlation of SBP between two sibs in the same family is 0.3. We measure the correlation between SBPs of twin pairs and find a correlation of 0.5 based on 50 twin pairs. Is this a significant difference?

We wish to test the hypothesis $H_0: \rho = \rho_0 \neq 0$. versus $H_1: \rho \neq \rho_0$. The appropriate test is called *Fisher's one-sample z-test for correlation*. The test statistic for this test is

$$\lambda = (z - z_0)\sqrt{n-3} \sim N(0, 1) \text{ under } H_0$$

where

$$\begin{aligned} z &= \text{Fisher's } z \text{ transformation of } r \\ &= 0.5 \ln\left(\frac{1+r}{1-r}\right) \text{ (given in Table 13, in the Appendix of the text)} \\ z_0 &= \text{Fisher's z transformation of } \rho_0 \end{aligned}$$

The p-value $= 2 \times \Phi(\lambda)$ if $\lambda < 0$, or $= 2 \times [1 - \Phi(\lambda)]$ if $\lambda \geq 0$.

In this case, a one-sided alternative is more appropriate, because twins can only plausibly have a stronger correlation than sib pairs. Thus, we test the hypothesis $H_0: \rho = 0.3$ versus $H_1: \rho > 0.3$.

We have

$$\begin{aligned} z_0 &= \frac{1}{2}\ln\left(\frac{1.3}{0.7}\right) = 0.310 \\ z &= \frac{1}{2}\ln\left(\frac{1.5}{0.5}\right) = 0.549 \\ \lambda &= \sqrt{47}(0.549 - 0.310) = 1.644 \sim N(0, 1) \text{ under } H_0. \\ \text{One-tailed } p &= Pr[N(0, 1) > 1.644] = 1 - \Phi(1.644) = .050. \end{aligned}$$

11.8.2 Two-Sample Inference

The tracking correlation is an index of the relationship between the same physiologic measure at different times. Suppose we measure systolic blood pressure at two points in time three years apart. We find that the correlation between the two repeated measures (the tracking correlation) $= 0.7$ for 50 people age 30–49 at baseline and 0.5 for 40 people age 50–69 at baseline. Suppose we wish to compare the underlying two tracking correlations. We wish to test the hypothesis

$$H_0: \rho_1 = \rho_2 \text{ versus } H_1: \rho_1 \neq \rho_2$$

We use the test statistic

$$\lambda = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} \sim N(0, 1)$$

where z_1, z_2 are Fisher's z transformation of r_1 and r_2, respectively. The p-value $= 2 \times \Phi(\lambda)$ if $\lambda < 0$, or $2 \times [1 - \Phi(\lambda)]$ if $\lambda \geq 0$.

In this case,

$$\begin{aligned} z_1 &= 0.867 \qquad z_2 = 0.549 \\ \lambda &= \frac{0.318}{\sqrt{\frac{1}{47} + \frac{1}{37}}} = \frac{0.318}{0.220} = 1.447 \sim N(0, 1) \end{aligned}$$

The p-value $= 2 \times [1 - \Phi(1.447)] = 2(1 - .926) = .148$. Thus, there is no significant difference between the tracking correlation coefficients for the two age groups.

11.8.3 Sample Size Estimation for Correlation Coefficients

In some, instances, the primary goal of an investigation is to test hypotheses concerning a correlation coefficient. For example, a number of specific genetic mutations have been found to be associated with the ocular disease retinitis pigmentosa, a condition which often results in loss of visual field and in some cases leads to legal blindness. One issue is whether there are additional environmental factors that are associated with the course of the disease. To test this hypothesis, two subjects in the same family with a specific genetic mutation of about the same age are identified and the correlation coefficient between their visual fields is computed. The correlation coefficient is computed over all pairs of affected people for families with the genetic mutation. We wish to test if this correlation is significantly different from zero which would indicate either that there are other genes involved or that there are other common environmental factors. It is expected that the correlation coefficient will be 0.2.

In general, if the expected correlation is ρ_0 and a 1-sided test is performed with significance level α and power $= 1-\beta$, then we need to study

$$n = \left[\frac{\left(z_{1-\alpha} + z_{1-\beta}\right)^2}{z_0^2} + 3 \right] \text{ pairs of subjects.}$$

If $\rho_0 = .2$, $\alpha = .05$, and $1-\beta = .80$, then $z_0 = 0.203$ and

$$n = \frac{\left(z_{.95} + z_{.80}\right)^2}{(0.203)^2} + 3 = \frac{(1.645 + 0.84)^2}{(0.203)^2} + 3 = 153$$

Thus, we need to study 153 pairs of subjects to achieve 80% power.

If alternatively, we only have *n* pairs of subjects, then the power is given by

$$\text{Power} = \Phi\left(\sqrt{n-3}\, z_0 - z_{1-\alpha}\right)$$

Therefore, if we are only able to enroll 100 families with two affected members each, then the power is given by

$$\Phi\left[\sqrt{97}(0.203) - 1.645\right] = \Phi(0.354) = .64.$$

SECTION 11.9 Multiple Regression

The purpose of multiple regression is to relate a single normlly distributed outcome variable *y* to several $(k > 1)$ predictor variables $x_1, \ldots, x_k$.

11.9.1 Multiple-Regression Model

$$y_i = \alpha + \sum_{j=1}^{k} \beta_j x_{ij} + e_i \text{ where } x_{ij} = j \text{ th variable for the } i\text{th person,}$$

$$y_i = \text{outcome variable for the } i\text{th person, } e_i \sim N\left(0,\ \sigma^2\right)$$

We use the method of least squares to fit the model.

11.9.2 Interpretation of Regression Coefficients

The regression coefficients β_j in a multiple-regression model are referred to as *partial-regression coefficients.* β_j can be interpreted as the average increase in y per unit increase in x_j, *holding all other variables constant.* This is different from the regression coefficient β in a simple linear-regression model, where no other variables are controlled for.

11.9.3 Hypothesis Testing

There are two types of hypotheses that can be tested with multiple regression. The first type of hypothesis is

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0 \text{ versus } H_1: \text{at least one } \beta_j \neq 0$$

This is a test of the overall hypothesis that at least some of the β_j are different from zero without specifying which one is different. An F test can be employed to test this hypothesis. The second type of hypothesis is $H_0: \beta_j = 0$, all other $\beta_l \neq 0$ versus H_1: all $\beta_j \neq 0$. This is a test of the hypothesis that the *j*th variable has a significant effect on y after controlling for the other covariates in the model. For example, in the blood pressure data in Section 11.2, the variables age, sex and weight were also available for all subjects. If we wish to study the relationship between SBP vs. age, sex, weight and fruit consumption when considered simultaneously, then we have

y_i = SBP for the *i*th subject
x_{i1} = age = age (yrs) for the *i*th subject
x_{i2} = sex for the *i*th subject = 1 if male, 2 if female
x_{i3} = weight (lb) for the *i*th subject
x_{i4} = fruit = number of servings of fruit per day for the *i*th subject

We fit the multiple-regression model

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + e_i$$

The results are as follows, using SAS PROC REG:

STATISTICAL ANALYSIS SYSTEM

Model: Model 02		SSE	5096.259	F Ratio	8.74	
		DFE	55	Prob>F	0.0001	
Dep Var: MSYS		MSE	92.659246	R-square	0.3885	
Variable	DF	Parameter estimate	Standard error	T Ratio	Prob>|T|	Variable label
Intercept	1	96.541534	12.464154	7.7455	0.0001	
Age	1	0.013926	0.176773	0.0788	0.9375	
Sex	1	-6.738063	2.782668	-2.4214	0.0188	
Weight	1	0.196530	0.065980	2.9786	0.0043	
Fruit	1	-1.137050	0.538329	-2.1122	0.0392	

We can test the hypothesis $H_0: \beta_4 = 0$, $\beta_1, \beta_2, \beta_3 \neq 0$ versus H_1: all $\beta_j \neq 0$; this is a test of the hypothesis that fruit consumption has an effect on SBP after controlling for age, sex, and weight. We have

$$b_4 = -1.14 \pm 0.54, \; p = .039$$

This indicates that fruit consumption has a significant effect on SBP even after controlling for age, sex, and weight. However, the magnitude of the effect is much smaller than in the simple linear-regression model fit in Section 11.2, where $b = -1.53$. Roughly 1/4 of the effect of fruit consumption can be explained by the other variables in the model.

We also see that sex and weight have significant effects on SBP, indicating that after holding all other variables constant, males have higher SBP than females (by 6.7 mm Hg on average), and heavier people have higher SBP than lighter people (by 0.197 mm Hg per 1 pound of weight). At the top of the SAS output is a test of the first hypothesis H_0: all $\beta_j = 0$ versus H_1: at least one $\beta_j \neq 0$. The *p*-value (listed under Prob>F) is 0.0001, indicating that at least some of the variables are having a significant effect on SBP. Finally, the $R^2 = 0.39$, indicating that 39% of the variation in SBP can be explained by age, sex, weight, and fruit consumption.

11.9.4 Several Types of Correlation Associated with Multiple Regression

1. **Partial Correlation** The correlation between a normally distributed outcome variable *y* (e.g., SBP) and a single predictor variable (e.g., fruit consumption) after controlling both the outcome and the predictor variable for one or more other predictors in a multiple-regression model (e.g., age, sex, weight).
2. **Multiple Correlation** The correlation between a normally distributed outcome variable *y* (e.g., SBP) and the linear predictor $\beta_1 x_1 + \ldots + \beta_k x_k$, where $\beta_1, \ldots, \beta_k$ are the coefficients in a multiple regression model. The multiple correlation represents the highest possible correlation between *y* and a linear function of the *x*'s. For example, in the SAS multiple-regression model shown in Section 11.9.3, the multiple correlation $= \sqrt{.3885} = .62$.

SECTION 11.10 Rank Correlation

In the Data Set VALID.DAT (in Chapter 2, text), we provide data from a validation study based on 173 women in the Nurses' Health Study of two methods for assessing dietary intake: a food-frequency questionnaire (FFQ) and a diet record (DR). The DR is considered the gold standard, but is very expensive to administer; the FFQ is inexpensive, but has substantial measurement error. Data are presented on four nutrients. Here we focus on the correlation between alcohol intake as measured by the FFQ versus the DR. Since alcohol intake is very skewed and nonnormal, with a substantial fraction of women reporting no intake, we are reluctant to use the ordinary Pearson correlation coefficient. Instead, we will use a nonparametric analogue to the Pearson correlation known as the *Spearman rank-correlation coefficient*, denoted by r_s. To compute the rank correlation, we first rank each of the 173 FFQ values and the 173 DR values and then compute an ordinary Pearson correlation based on the ranked data. Once the correlation is obtained we can test the null hypothesis of no association by using the test statistic

$$t_s = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} \sim t_{n-2} \text{ under } H_0$$

The *p*-value $= 2 \times Pr(t_{n-2} < t_s)$ if $t_s < 0$, or $= 2 \times Pr(t_{n-2} > t_s)$ if $t_s \geq 0$. This test should only be used if $n \geq 10$.

The computations were performed using MINITAB. A sample of the ranked data for DR (labeled ALCD_R) and FFQ (labeled ALCF_R) are shown on the following page, together with the corresponding raw data (labeled ALCD and ALCF).

The calculation of the rank correlation is given as follows:

CORRELATION OF ALCD_R AND ALCF_R = 0.899

The rank correlation $= .899$, which is quite high. To perform the significance test, we compute the test statistic

$$t_s = \frac{.899\sqrt{171}}{\sqrt{1-(.899)^2}} = 26.8 \sim t_{171} \text{ under } H_0$$

The p-value $= 2 \times Pr(t_{171} > 26.8) < .001$. Clearly, there is a strong correlation between alcohol intake as assessed by DR and FFQ.

For smaller n ($n \le 9$), a small-sample table should be used (see Table 14, Appendix, text) to assess significance. To use this table, we reject H_0 if $|r_s| \ge c$, where c = the critical value for a two-sided α level (e.g., $\alpha = .05$).

Suppose $n = 6$ and $r_s = .850$. Since the critical value $= .829$ for $\alpha = .10$ and .886 for $\alpha = .05$, it follows that $.05 \le p < .10$.

	ALCD	ALCF	ALCD_R	ALCF_R
155	8.18	6.63	107.0	102.0
156	0.51	0.00	25.0	17.5
157	7.82	12.08	102.0	134.5
158	6.60	11.14	90.0	124.5
159	28.79	52.07	165.0	171.0
160	3.82	1.81	69.0	61.0
161	0.00	0.00	9.0	17.5
162	8.31	2.91	110.0	79.0
163	8.22	13.76	108.0	142.5
164	0.96	0.76	38.0	42.5
165	0.63	0.00	29.0	17.5
166	15.72	16.81	140.0	151.0
167	23.97	12.43	158.0	136.0
168	7.28	6.63	98.0	102.0
169	4.60	7.25	75.5	108.5
170	2.56	1.81	56.0	61.0
171	13.91	8.64	134.0	112.5
172	5.39	7.19	86.0	107.0
173	0.60	3.49	27.5	80.0

PROBLEMS

Cardiology
The data in Table 11.2 are given for 27 patients with acute dilated cardiomyopathy [1].

11.1 Fit a regression line relating age (x) to LVEF (y).

11.2 What is the averge LVEF for a 45-year-old patient with this condition?

11.3 Test for the significance of the regression line in Problem 11.1 using the F test.

11.4 What is the R^2 for the regression line in Problem 11.1?

11.5 What does R^2 mean in Problem 11.4?

11.6 Test for the significance of the regression line in Problem 11.1 using the t test.

11.7 What are the standard errors of the slope and intercept for the regression line in Problem 11.1?

11.8 Provide 95% Cl's for the slope and intercept.

11.9 What is the z transformation of 0.435?

Table 11.2 Data for patients with acute dilated cardiomyopathy

Patient number	Age, x	Left ventricular ejection fraction (LVEF), y
1	35	0.19
2	28	0.24
3	25	0.17
4	75	0.40
5	42	0.40
6	19	0.23
7	54	0.20
8	35	0.20
9	30	0.30
10	65	0.19
11	26	0.24
12	56	0.32
13	60	0.32
14	47	0.28
15	50	0.24
16	43	0.18
17	30	0.22
18	56	0.23
19	23	0.14
20	26	0.14
21	58	0.30
22	65	0.07
23	34	0.12
24	63	0.13
25	23	0.17
26	23	0.24
27	46	0.19

Note: $\sum x_i = 1137$, $\sum x_i^2 = 54{,}749$, $\sum y_i = 6.05$, $\sum y_i^2 = 1.5217$, $\sum x_i y_i = 262.93$

Source: Reprinted with permission of the *New England Journal of Medicine*, **312**(14), 885–890, 1985.

Hypertension

Many hypertension studies involve more than one observer, and it is important to verify that the different observers are in fact comparable. A new observer and an experienced observer simultaneously take blood pressure readings on 50 people (x_i, y_i), $i = 1, \ldots, 50$, and the sample correlation coefficient is .75. Suppose the minimum acceptable population correlation is .9.

11.10 Are the sample data compatible with this population correlation?

Pathology

11.11 Assuming normality, use the data in Table 6.1 (in Chapter 6 of this Study Guide) to test if there is any relation between total heart weight and body weight within each of the two groups.

Hypertension

Suppose that 20 married couples, each in the age group 25–34, have their systolic blood pressures taken, with the data listed in Table 11.3.

Table 11.3 SBP measurements from 20 married couples

Couple	Male	Female	Couple	Male	Female
1	136	110	11	156	135
2	121	112	12	98	115
3	128	128	13	132	125
4	100	106	14	142	130
5	110	127	15	138	132
6	116	100	16	126	146
7	127	98	17	124	127
8	150	142	18	137	128
9	180	143	19	160	135
10	172	150	20	125	110

11.12 Test the hypothesis that there is no correlation between the male and female scores.

11.13 Suppose a study in the literature has found $\rho = 0.1$ based on a sample of size 100. Are the data in Table 11.3 consistent with this finding?

Hypertension

The level of a vasoactive material called *kallikrein* in the urine is a variable that has been associated in some studies with level of blood pressure in adults. Generally, people with low levels of kallikrein tend to have high levels of blood pressure. A study was undertaken in a group of 18 infants to see if this relationship persisted. The data in Table 11.4 were obtained.

Suppose we are given the following statistics:

$$\sum_{i=1}^{18} x_i = 67.765,\ \sum_{i=1}^{18} x_i^2 = 267.217,\ \sum_{i=1}^{18} y_i = 4.104,$$

$$\sum_{i=1}^{18} y_i^2 = 28.767,\ \sum_{i=1}^{18} x_i y_i = 17.249$$

11.14 Assuming that a linear relationship exists between SBP z score and ln kallikrein, derive the best-fitting linear relationship between these two variables.

11.15 Is there a significant relationship between these two variables based on your answer to Problem 11.14? Report a p-value.

11.16 In words, what does a p-value mean in the context of Problem 11.15?

Table 11.4 The relationship of kallikrein to blood pressure

Observation	ln[a] (kallikrein) x	SBP z score[b] y
1	2.773	1.929
2	5.545	–1.372
3	3.434	–0.620
4	3.434	1.738
5	2,639	0.302
6	3.091	0.679
7	4.836	0.999
8	3.611	0.656
9	4.554	0.027
10	3.807	–0.057
11	4.500	1.083
12	2.639	–2.265
13	3.555	0.963
14	3.258	–1.062
15	4.605	2.771
16	3.296	–0.160
17	4.787	–0.217
18	3.401	–1.290

[a] The ln transformation was used to better normalize the underlying distribution.

[b] Instead of raw blood pressure, a z score was used, which is defined as the raw blood pressure standardized for body weight and expressed in standard deviation units, with positive scores indicating high blood pressure and negative scores indicating low blood pressure.

11.17 Assess the goodness of fit of the regression line fitted in Problem 11.14.

Pediatrics

It is often mentioned anecdotally that very young children have a higher metabolism that gives them more energy than older children and adults. We decide to test this hypothesis by measuring the pulse rates on a selected group of children. The children are randomly chosen from a community census so that two male children are selected from each 2-year age group starting with age 0 and ending with age 21 (i.e., 0–1/2–3/4–5/.../20–21). The data are given in Table 11.5.

We are given the following basic statistics:

$$n = 22 \qquad \sum_{i=1}^{22} y_i = 1725 \qquad \sum_{i=1}^{22} y_i^2 = 140{,}933$$

$$\sum_{i=1}^{22} x_i = 233 \qquad \sum_{i=1}^{22} x_i^2 = 3345 \qquad \sum_{i=1}^{22} x_i y_i = 16{,}748$$

11.18 Suppose we hypothesize that there is a linear-regression model relating pulse rate (y) and age (x). What are the assumptions for such a model?

11.19 Fit the parameters for the model in Problem 11.18.

11.20 Test the model fitted in Problem 11.19 for statistical significance.

11.21 What is the predicted pulse for an average 12 year-old child?

Table 11.5 Pulse rate and age in children ages 0–21

Age group	Age, x_i	Pulse rate, y_i
0–1	1 0	103 125
2–3	3 3	102 86
4–5	5 5	88 78
6–7	6 6	77 68
8–9	9 8	90 75
10–11	11 11	78 66
12–13	12 13	76 82
14–15	14 14	58 56
16–17	16 17	72 70
18–19	19 18	56 64
20–21	21 21	81 74

11.22 What is the standard error of the estimate in Problem 11.21?

11.23 Suppose John Smith is 12 years old. What is his estimated pulse rate from the regression line?

11.24 What is the standard error of the estimate in Problem 11.23?

11.25 How does John compare with other children in his age group if his actual pulse rate is 75?

11.26 What is the difference between the two predictions and standard errors in Problems 11.21–11.24?

11.27 Plot the data and assess whether the assumptions of the model are satisfied. In particular, does the linearity assumption appear to hold?

Hypertension, Genetics

Much research has been devoted to the etiology of hypertension. One general problem is to determine to what extent hypertension is a genetic phenomenon. This issue ws examined in 20 families by measuring the SBP of the mother, father, and first-born child in the family. The data are given in Table 11.6.

Table 11.6 Familial systolic blood-pressure (SBP) relationships

Family	SBP mother (mm Hg), y	SBP father (mm Hg), x	SBP child (mm Hg), t
1	130	140	90
2	125	120	85
3	140	180	120
4	110	150	100
5	145	175	105
6	160	120	100
7	120	145	110
8	180	160	140
9	120	190	115
10	130	135	105
11	125	150	100
12	110	125	80
13	90	140	70
14	120	170	115
15	150	150	90
16	145	155	90
17	130	160	115
18	155	125	110
19	110	140	90
20	125	150	100

$$\sum_{i=1}^{20} x_i = 2980,\ \sum_{i=1}^{20} x_i^2 = 451{,}350,\ \sum_{i=1}^{20} y_i = 2620,$$
$$\sum_{i=1}^{20} y_i^2 = 351{,}350,\ \sum_{i=1}^{20} x_i y_i = 390{,}825,\ \sum_{i=1}^{20} t_i = 2030,$$
$$\sum_{i=1}^{20} t_i^2 = 210{,}850,\ \sum_{i=1}^{20} x_i t_i = 305{,}700,\ \sum_{i=1}^{20} y_i t_i = 269{,}550$$

11.28 Test the assumption that the mother's SBP and father's SBP are uncorrelated. (We would expect a lack of correlation, since the mother and father are genetically unrelated.)

11.29 We would expect from genetic principles that the correlation between the mother's SBP and the child's SBP is .5. Can this hypothesis be tested?

11.30 Suppose a nonzero correlation has been found in Problem 11.29. Is there some explanation other than a genetic one for the existence of this correlation?

We would like to be able to predict the child's SBP on the basis of the parents' SBP.

11.31 Find the best-fitting linear relationship between the child's SBP and the mother's SBP.

11.32 Test for the significance of this relationship.

11.33 What would be the average child's SBP if the mother's SBP is 130 mm Hg?

11.34 What would be the average child's SBP if the mother's SBP is 150 mm Hg?

11.35 What would be the average child's SBP if the mother's SBP is 170 mm Hg?

11.36 Find the standard errors for the estimates in Problems 11.33-11.35.

11.37 Why are the three standard errors calculated in Problem 11.36 not the same?

The two-variable regression model with the father's SBP and the mother's SBP as independent variables and the child's SBP as the dependent variable is fitted in Table 11.7. Suppose we wish to assess the independent effects of the mother's SBP and father's SBP as predictors of the child's SBP.

Table 11.7 Two-variable regression model predicting child's SBP as a function of the parental SBP's.

Predictor	COEF	STDEV	T-Ratio	P
Constant	–15.69	23.65	–0.66	0.516
SBP-Moth	0.4225	0.1185	3.57	0.002
SBP-Fath	0.4150	0.1248	3.32	0.004

S = 10.67 R-SQ= 59.7%

Analysis of variance

Source	DF	SS	MS	F	P
Regression	2	2870.1	1435.0	12.61	0.000
Error	17	1934.9	113.8		
Total	19	4805.0			

11.38 What hypotheses should be tested for this assessment?

11.39 Test the hypotheses proposed in Problem 11.38.

11.40 What are the standardized regression coefficients in Table 11.7 and what do they mean?

11.41 Assess the goodness of fit of the multiple regression model in Table 11.7.

Nutrition
A study was performed looking at the reproducibility of dietary intake among a group of 181 children ages 45–60 months at baseline [2]. The children provided dietary recall data in year 1 and also in year 3. Suppose the correlation coefficient com-puted for polyunsaturated fat between these two assessments and .27.

11.42 What test can be used to assess if the true correlation coefficient is different from zero?

11.43 Implement the test in Problem 11.42 and report a *p*-value.

11.44 Provide a 95% Cl for the true correlation (ρ).

11.45 Suppose in the same study that the correlation coefficient for saturated fat is .43, and we wish to compare the correlation coefficients for saturated and polyunsaturated fat. Can the two-sample test for correlations (Equation 11.27, in Chapter 11, text) be employed to compare these correlations? Why or why not? (Do not perform the test.)

Hypertension, Nutrition
A study was performed to investigate the relationship between alcohol intake and systolic blood pressure (SBP). The results are given as follows:

Alcohol Intake (g/day)	Mean SBP	sd	*n*
0	130.8	18.2	473
≥ 30	128.5	20.9	185

11.46 What test can be performed to compare mean SBP between the high-intake, ≥ 30 (g/day), and no-intake group, if we assume that the underlying variances of the two groups are the same? (Do not implement this test.)

11.47 It is well known that age is related to SBP, and it is suspected that age may be related to alcohol intake as well. Can an analysis be performed to look at differences in mean SBP between the high-intake and no-intake group, while controlling for age? Write down the type of model and the hypotheses to be tested under H_0 and H_1. (Do not implement this analysis.)

11.48 The authors also studied a low alcohol intake group (alcohol intake > 0, but < 30 g/day). Suppose we want to consider all three groups and wish to test the hypothesis that mean SBP is the same in all three groups versus the alternative hypothesis that mean SBP is different between at least two of the groups while controlling for age. What type of analysis could be performed to answer this question? (Do not implement this analysis.)

Hypertension
The variability of blood pressure is important in planning screening programs for detecting people with high blood pressure. There are two schools of thought concerning this variability: Some researchers feel that a subgroup of people have extremely variable blood pressures, whereas most people have relatively stable blood pressures; other researchers feel that this variability is common to all people. A study was set up to answer this question. A group of 15 people had their blood pressures measured on three separate visits in each of 2 years, and the between-visit variance of blood pressure was measured each year. The results for systolic blood pressure (SBP) are given in Table 11.8. We wish to assess if there is any relationship between the year 1 and year 2 variances.

Table 11.8 Blood-pressure variability measured at two points in time

Person	Between-visit variance, year 1	Between-visit variance, year 2
1	8.8	19.0
2	10.0	25.7
3	6.7	17.9
4	13.3	95.7
5	10.0	16.7
6	6.2	15.2
7	35.0	3.9
8	51.2	11.9
9	30.9	14.6
10	61.0	21.6
11	5.4	31.3
12	46.6	21.4
13	37.0	10.1
14	2.6	18.7
15	2.0	22.7

11.49 Why might a rank correlation be a useful method for expressing such a relationship?

11.50 Test for the significance of the rank correlation based on the previous data.

11.51 What are your conclusions based on these data?

Cardiovascular Disease
An emerging field is the use of non-invasive markers of atherosclerosis as early warning signs of coronary disease. One marker that has been frequently studied is the carotid artery intima-media thickness (IMT) which is a marker of blockage of the carotid artery. An investigator wants to set up a study relating systolic blood pressure to IMT. In other studies, the correlation between these measures is about .3.

11.52 How many subjects should be enrolled in the study to test the hypothesis that there is an association between these 2 measures if a one-sided test is to be used with $\alpha = .05$ and 80% power is desired?

11.53 Suppose that 100 subjects are enrolled in the study. How much power will the study have if a one-sided test is used with $\alpha = .05$?

SOLUTIONS

11.1
$$L_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$
$$= 54{,}749 - \frac{1137^2}{27}$$
$$= 54{,}749 - 47{,}880.33 = 6868.67$$
$$L_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$
$$= 262.93 - \frac{1137(6.05)}{27}$$
$$= 262.93 - 254.77 = 8.16$$
$$b = \frac{L_{xy}}{L_{xx}} = \frac{8.16}{6868.67} = 0.0012$$
$$a = \frac{\sum y_i - b\sum x_i}{n} = \frac{6.05 - 0.0012(1137)}{27}$$
$$= \frac{6.05 - 1.350}{27} = \frac{4.700}{27} = 0.174$$

Thus, the regression line is given by

$$y = 0.174 + 0.0012x\,.$$

11.2 The expected LVEF is

$$0.174 + 0.0012(45) = 0.174 + 0.053 = 0.228\,.$$

11.3 First compute L_{yy}.

$$L_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = 1.5217 - \frac{(6.05)^2}{27}$$
$$= 1.522 - 1.356 = 0.166$$

Now compute the Regression and Residual Sum of Squares and Mean Square:

$$\text{Reg SS} = \frac{L_{xy}^2}{L_{xx}} = \frac{8.16^2}{6868.67} = 0.0097 = \text{Reg MS}$$
$$\text{Res SS} = 0.166 - 0.0097 = 0.156$$
$$\text{Res MS} = \frac{0.156}{25} = 0.0063$$

Finally, the F statistic is given by

$$F = \frac{\text{Reg MS}}{\text{Res MS}} = \frac{0.0097}{0.0063} = 1.55 \sim F_{1,\,25} \text{ under } H_0\,.$$

Since

$$F_{1,\,25,\,.95} > F_{1,\,30,\,.95} = 4.17 > 1.55,$$

it follows that $p > .05$. Therefore, H_0 is accepted and we conclude that there is no significant relationship between LVEF and age.

11.4 $R^2 = \text{Reg SS}/\text{Total SS} = 0.0097/0.166 = 0.058$

11.5 It means that only 5.8% of the variance in LVEF is explained by age.

11.6 Compute the standard error of the estimated slope as follows:

$$se(b) = \sqrt{\frac{\text{Res MS}}{L_{xx}}} = \sqrt{\frac{0.0063}{6868.67}} = 0.00095$$

Thus, the t statistic is given by

$$t = \frac{b}{se(b)} = \frac{0.0012}{0.00095} = 1.24 \sim t_{25} \text{ under } H_0$$

Since $t_{25,\,.975} = 2.060 > t$, it follows that $p > .05$, and H_0 is accepted at the 5% level.

11.7 $se(b) = 0.00095$ as given in Problem 11.6. The standard error of the intercept is given by

$$se(a) = \sqrt{\text{Res MS}\left(\frac{1}{n} + \frac{\bar{x}^2}{L_{xx}}\right)}$$
$$= \sqrt{0.0063\left[\frac{1}{27} + \frac{(\frac{1137}{27})^2}{6868.67}\right]}$$
$$= \sqrt{0.0063(0.0370 + 0.2582)}$$
$$= \sqrt{0.0063(0.2952)} = 0.043$$

11.8 A 95% CI for the slope is given by

$$b \pm t_{25,\ .975} se(b) = 0.0012 \pm 2.060(0.00095)$$
$$= 0.0012 \pm 0.0020$$
$$= (-0.0008,\ 0.0032)$$

A 95% CI for the intercept is given by

$$a \pm t_{25,\ .975} se(a) = 0.174 \pm 2.060(0.043)$$
$$= 0.174 \pm 0.089 = (0.086,\ 0.263)$$

11.9 We have that

$$z = 0.5 \times \ln\left(\frac{1+0.435}{1-0.435}\right) = 0.5 \times \ln\left(\frac{1.435}{0.565}\right)$$
$$= 0.5 \times \ln(2.540) = 0.5(0.932) = 0.466$$

11.10 We wish to test the hypothesis $H_0: \rho = .9$ versus $H_1: \rho < .9$. We use the one-sample z test with test statistic

$$\lambda = \sqrt{n-3}(z - z_0) \sim N(0,\ 1) \text{ under } H_0$$

where from Table 13 (in the Appendix, text)

$$z = 0.5 \times \ln\left(\frac{1+.75}{1-.75}\right) = 0.973$$
$$z = 0.5 \times \ln\left(\frac{1+.9}{1-.9}\right) = 1.472$$

We have

$$\lambda = \sqrt{47}(0.973 - 1.472) = 6.856(-0.499)$$
$$= -3.423 \sim N(0,\ 1) \text{ under } H_0.$$

The p-value $= \Phi(-3.423) < .001$. Thus, the observed correlation is significantly lower than .9, indicating that the new observer does *not* meet the standard.

11.11 Compute the correlation coefficient between THW and BW in each of the two groups. Use the CORRELATION function of Excel and obtain

Left-heart disease	Normal
$r = 0.296$	$r = 0.657$

We now compute the t statistic for the one sample t test for correlation and use the TDIST function of Excel to obtain the p-value (see Appendix).

Left-heart disease	Normal
$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.296(3)}{\sqrt{1-0.296^2}}$	$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.657\sqrt{8}}{\sqrt{1-0.657^2}}$
$= 0.93 \sim t_9,\ p = .38.$	$= 2.465 \sim t_8,\ p = .039.$

Thus, there is a significant association between total heart weight and body weight in the normal group, but not in the left-heart disease group.

11.12 If $x =$ male, $y =$ female, then we have that

$$r = \frac{L_{xy}}{\sqrt{L_{xx} \times L_{yy}}},$$

where

$$L_{xx} = 8923.8,\ L_{yy} = 4412.95,\ L_{xy} = 4271.9.$$

Therefore,

$$r = \frac{4271.9}{\sqrt{8923.8 \times 4412.95}} = .681$$

We use the one sample t test for correlation with the following test statistic:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{.681\sqrt{18}}{\sqrt{1-.681^2}} = \frac{2.888}{0.733} = 3.94 \sim t_{18} \text{ under } H_0$$

Since $t_{18,\ .9995} = 3.922 < 3.94$, we have that $p < .001$. Thus, there is a significant correlation between spouse blood pressures. Alternatively, we could use the CORRELATION and TDIST functions of Excel to solve this problem (see Appendix).

11.13 Let ρ_1 = our population correlation coefficient, ρ_2 = literature population correlation coefficient. We wish to test the hypothesis $H_0: \rho_1 = \rho_2$ versus $H_1: \rho_1 \neq \rho_2$. We have the test statistic

$$\lambda = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} \sim N(0,\ 1) \text{ under } H_0$$

We have

$$z_1 = 0.5 \times \ln\left(\frac{1+0.681}{1-0.681}\right) = 0.830$$
$$z_2 = 0.5 \times \ln\left(\frac{1+0.10}{1-0.10}\right) = 0.100$$

Thus,

$$\lambda = \frac{0.830 - 0.100}{\sqrt{\frac{1}{17} + \frac{1}{97}}}$$
$$= \frac{0.730}{0.263} = 2.777 \sim N(0,\ 1) \text{ under } H_0$$

The p-value

$$2 \times [1 - \Phi(\lambda)] = 2 \times [1 - \Phi(2.777)] = 2 \times (1 - .9973) = .005.$$

Thus, we can say that our correlation is significantly different from the correlation in the literature.

11.14 We fit a least-squares line of the form $y = a + bx$ where $b = L_{xy}/L_{xx}$, $a = \bar{y} - b\bar{x}$. We have that

$$L_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{18} = 267.217 - \frac{67.765^2}{18} = 12.101$$

$$L_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{18} = 28.767 - \frac{4.104^2}{18} = 27.831$$

$$L_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{18} = 17.249 - \frac{67.765(4.104)}{18} = 1.799$$

Thus,

$$b = \frac{1.799}{12.101} = 0.149$$

$$a = \frac{\sum y_i - b\sum x_i}{18} = \frac{4.104 - 0.149 \times 67.765}{18} = -0.332$$

The best-fitting linear relationship is $y = -0.332 + 0.149x$.

11.15 We wish to test the hypothesis $H_0: \beta = 0$ versus $H_1: \beta \neq 0$. We use the test statistic

$$F = \frac{\text{Reg MS}}{\text{Res MS}} \sim F_{1,\, n-2} \text{ under } H_0.$$

We have

$$\text{Reg SS} = \frac{L_{xy}^2}{L_{xx}} = \frac{1.799^2}{12.101} = 0.267 = \text{Reg MS}$$

$$\text{Total SS} = L_{yy} = 27.831$$

$$\text{Res SS} = \text{Tot SS} - \text{Reg SS} = 27.831 - 0.267 = 27.564$$

$$\text{Res MS} = \frac{\text{Res SS}}{n-2} = \frac{27.564}{16} = 1.723$$

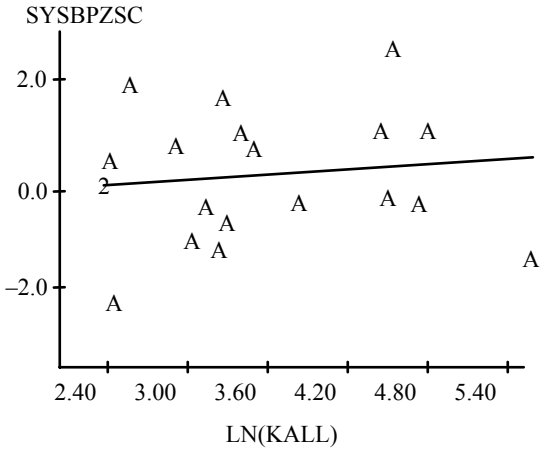
Thus,

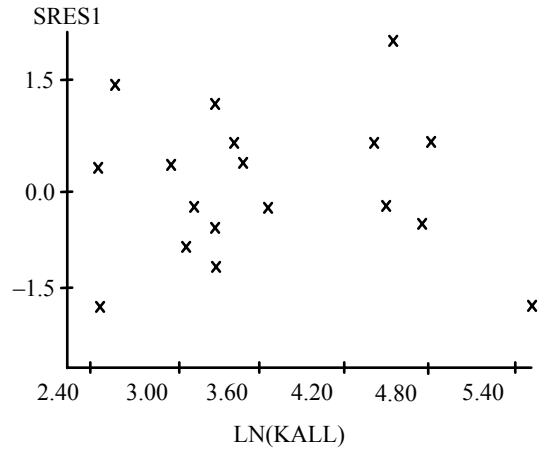
$$F = \frac{\text{Reg MS}}{\text{Res MS}} = \frac{0.267}{1.723} = 0.155 \sim F_{1,\,16} \text{ under } H_0.$$

We see from the F table that $F_{1,\,16,\,.90} = 3.05 > F$. Thus, $p > .10$ and there is no significant relationship between these two variables in this data set.

11.16 A p-value in this context is the probability of obtaining an estimated slope at least as large (in absolute value) as the one obtained between blood pressure and kallikrein, when in fact no relationship exists.

11.17 We use MINITAB to run the regression of SBP z score on ln(kallikrein). A scatter plot of the raw data (labeled A) and the fitted regression line are given below. They do not reveal any outliers, influential points, or systematic departures from the linear-regression assumptions. This is reinforced by the plot of the Studentized residuals (SRES 1) versus ln(kallikrein). (see next page).





11.18 The underlying model here is $y_i = \alpha + \beta x_i + e_i$ where $e_i \sim N(0, \sigma^2)$, the e_i are independent random variables, y_i = pulse rate for the ith child, and x_i = age for the ith child.

11.19 We first compute

$$L_{xy} = 16{,}748 - \frac{233(1725)}{22} = -1521.32$$

$$L_{xx} = 3345 - \frac{233^2}{22} = 877.32$$

$$L_{yy} = 140{,}933 - \frac{1725^2}{22} = 5677.32$$

We then have

$$b = \frac{L_{xy}}{L_{xx}} = \frac{-1521.32}{877.32} = -1.73$$
$$a = \frac{\sum y_i - b\sum x_i}{22} = \frac{1725 + 1.73(233)}{22} = 96.8$$

The least-squares line is thus $y = 96.8 - 1.73x$.

11.20 We test the hypothesis $H_0: \beta = 0$ versus $H_1: \beta \neq 0$. We will use the F test for statistical significance. We have that

$$\text{Reg SS} = \frac{L_{xy}^2}{L_{xx}} = \frac{(-1521.32)^2}{877.32} = 2638.05$$
$$\text{Total SS} = L_{yy} = 5677.32$$
$$\text{Res SS} = 5677.32 - 2638.05 = 3039.27$$
$$\text{Res MS} = \frac{3039.27}{20} = 151.96$$

We thus have the following ANOVA table:

	SS	*df*	MS	*F*-stat
Regression	2638.05	1	2638.05	17.36
Residual	3039.27	20	151.96	
Total	5677.32	21		

We compute

$$F = \frac{\text{Reg MS}}{\text{Res MS}} = 17.36 \sim F_{1,\,20} \text{ under } H_0.$$

Since $F > F_{1,\,20,\,.999} = 14.82$, we have that $p < .001$ and we conclude that there is a highly significant correlation between pulse rate and age.

11.21 From the regression equation, we have that

$$\hat{y} = 96.8 - 1.73(12) = 76.0 \text{ beats per minute.}$$

11.22 The standard error is given by

$$\sqrt{s_{y\cdot x}^2\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{L_{xx}}\right]} = \sqrt{151.96\left[\frac{1}{22} + \frac{(12 - \frac{233}{22})^2}{877.32}\right]}$$
$$= \sqrt{151.96(0.048)} = 2.69$$

11.23 The estimated pulse is the same as for an average 12-year-old = 76.0 beats per minute (as given in the solution to Problem 11.21).

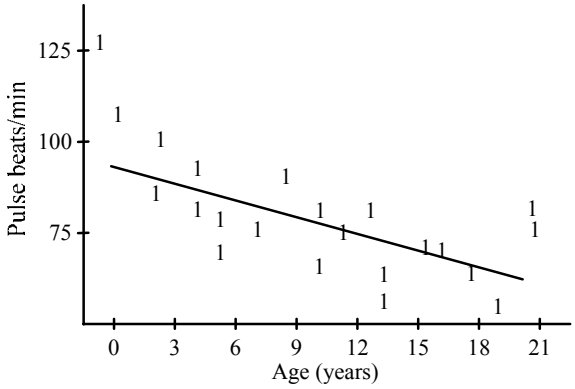
11.24 The standard error is given by

$$se = \sqrt{s_{y\cdot x}^2\left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{L_{xx}}\right]}$$
$$= \sqrt{151.96\left[1 + \frac{1}{22} + \frac{(12 - \frac{233}{22})^2}{877.32}\right]}$$
$$= \sqrt{151.96(1.048)} = 12.62$$

11.25 John is about average for children in his age group because his pulse is 75 while the average pulse is 76.0, with a standard error of 12.62.

11.26 The predicted pulse is the same for a particular 12-year-old and for an average 12-year-old (76.0). However, the standard error for a particular 12-year-old (12.62) is much larger than for an average 12-year-old (2.69).

11.27 A plot of the data in Table 11.5 is provided as follows. The fit of the regression line appears adequate. There is one mild outlier corresponding to a pulse of 125 at age 0. It may be that pulse rates for children under 1 year of age need to be modelled separately. However, more data are needed in this age range before being certain of this point.



11.28 Test the hypothesis $H_0: \rho = 0$ versus $H_1: \rho \neq 0$. Use the test statistic

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2} \text{ under } H_0$$

In this case

$$r = \frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}$$

We have

$$L_{xx} = 451{,}350 - \frac{2980^2}{20} = 7330$$

$$L_{yy} = 351{,}350 - \frac{2620^2}{20} = 8130$$

$$L_{xy} = 390{,}825 - \frac{2980(2620)}{20} = 445$$

Thus,

$$r = \frac{445}{\sqrt{7330 \times 8130}} = \frac{445}{7719.64} = .058$$

The test statistic is

$$t = \frac{.058\sqrt{18}}{\sqrt{1 - .058^2}} = \frac{0.245}{0.998} = 0.245 \sim t_{18}$$

Since $t_{18,\ .975} = 2.101$, this is clearly not statistically significant, and there is no significant correlation between the mother's and father's SBP, which is what would be expected on purely genetic grounds.

11.29 Compute

$$r_{yt} = \frac{L_{yt}}{\sqrt{L_{yy} \times L_{tt}}}$$

We have

$$L_{tt} = 210{,}850 - \frac{2030^2}{20} = 4805$$

$$L_{yt} = 269{,}550 - \frac{2620(2030)}{20} = 3620$$

and

$$r = \frac{3620}{\sqrt{8130 \times 4805}} = \frac{3620}{6250.17} = .579$$

Test the null hypothesis $H_0: \rho = .5$ versus the alternative $H_1: \rho \neq .5$. Use the Fisher z statistic, which yields

$$z = \frac{1}{2}[\ln(1 + r) - \ln(1 - r)]$$

$$= \frac{1}{2}[\ln(1.579) - \ln(0.421)]$$

$$= \frac{1}{2}(0.4569 + 0.8656) = 0.661$$

Also, from Table 13 (Appendix, text), $z_0 = 0.549$. Use the test statistic

$$\lambda = (z - z_0)\sqrt{n - 3} \sim N(0,\ 1) \text{ under } H_0$$

Thus,

$$\lambda = \sqrt{17}(0.661 - 0.549) = 0.461 \sim N(0,\ 1) \text{ under } H_0.$$

The p-value $= 2 \times [1 - \Phi(0.461)] = .64$. Thus, we accept H_0 that $\rho = .5$.

11.30 The mother and first-born child were both living in the same environment, which might explain all or part of the observed correlation of blood pressure.

11.31 Use the model $t_i = \alpha + \beta y_i + e_i$. Test the null hypothesis that $\beta = 0$ versus the alternative hypothesis that $\beta \neq 0$.

$$b = \frac{L_{ty}}{L_{yy}}$$

where

$$L_{ty} = 3620$$

$$L_{yy} = 8130$$

from Problems 11.28 and 11.29. Thus,

$$b = \frac{3620}{8130} = 0.445$$

Furthermore,

$$a = \frac{\sum_{i=1}^{20} t_i - b\sum_{i=1}^{20} y_i}{20} = \frac{2030 - 0.445(2620)}{20} = 43.2$$

We thus have the linear relation $t = 43.2 + 0.445y$.

11.32 Test the hypothesis $H_0: \beta = 0$ versus $H_1: \beta \neq 0$. Use the test statistic

$$F = \frac{\text{Reg MS}}{\text{Res MS}} \sim F_{1,\ n-2} \text{ under } H_0$$

We have

$$\text{Reg MS} = \frac{L_{ty}^2}{L_{yy}} = \frac{3620^2}{8130} = 1611.86 = \text{Reg SS}$$

$$\text{Res MS} = \frac{L_{tt} - \text{Reg SS}}{18} = \frac{L_{tt} - 1611.86}{18}$$

Thus,

$$\text{Res MS} = \frac{4805 - 1611.86}{18} = 177.40 = s_{t \cdot y}^2$$

and

$$F = \frac{1611.86}{177.40} = 9.09 \sim F_{1,\,18} \text{ under } H_0$$

Since $F_{1,\,18,\,.99} = 8.29$, $F_{1,\,18,\,.995} = 10.22$, and

$$8.29 < 9.09 < 10.22,\ .005 < p < .01,$$

and there is a significant relationship between the mother's and child's SBP.

11.33 The child's SBP can be predicted from the mother's SBP from the relationship in Problem 11.31;

$$E(t) = 43.2 + 0.445 \times 130 = 101.1 \text{ mm Hg.}$$

11.34 In this case $E(t) = 43.2 + 0.445 \times 150 = 110.0$ mm Hg.

11.35 In this case $E(t) = 43.2 + 0.445 \times 170 = 118.9$ mm Hg.

11.36 The standard errors are given by the formula

$$se = \sqrt{s^2_{t\cdot y}\left[\frac{1}{n} + \frac{(y-\bar{y})^2}{L_{yy}}\right]} = \sqrt{177.40\left[\frac{1}{20} + \frac{(y-\bar{y})^2}{8130}\right]}$$

We have

$$\bar{y} = \frac{2620}{20} = 131 \text{ mm Hg}$$

Thus, for $y = 130$, 150, and 170 mm Hg, we have, respectively,

$$se(130) = \sqrt{177.40\left[\frac{1}{20} + \frac{(130-131)^2}{8130}\right]} = 2.98$$

$$se(150) = \sqrt{177.40\left[\frac{1}{20} + \frac{(150-131)^2}{8130}\right]} = 4.09$$

$$se(170) = \sqrt{177.40\left[\frac{1}{20} + \frac{(170-131)^2}{8130}\right]} = 6.49$$

11.37 The three standard errors are not the same because the standard error increases the further the value of the mother's SBP is from the mean SBP for all mothers (131 mm Hg). Thus, the *se* corresponding to 130 mm Hg is smallest, whereas the *se* corresponding to 170 mm Hg is largest.

11.38 To test for the independent effect of the father's SBP on the child's SBP after controlling for the effect of the mother's SBP, test the hypothesis $H_0: \beta_1 = 0,\ \beta_2 \neq 0$ versus $H_1: \beta_1 \neq 0,\ \beta_2 \neq 0$. Similarly, to test for the independent effect of the mother's SBP on the child's SBP after controlling for the effect of the father's SBP, test the hypothesis $H_0: \beta_2 = 0,\ \beta_1 \neq 0$ versus $H_1: \beta_2 \neq 0,\ \beta_1 \neq 0$.

11.39 Compute the following test statistics:

$$\text{Father: } t_1 = \frac{b_1}{se(b_1)} = \frac{0.4150}{0.1248} = 3.32 \sim t_{17},\ .001 < p < .01$$

$$\text{Mother: } t_2 = \frac{b_2}{se(b_2)} = \frac{0.4225}{0.1185} = 3.57 \sim t_{17},\ .001 < p < .01$$

Thus, each parent's SBP is significantly associated with the child's SBP after controlling for the other spouse's SBP.

11.40 From Problems 11.28 and 11.29,

$$s_t^2 = \frac{L_{tt}}{n-1} = \frac{4805}{19} = 252.89$$

$$s_x^2 = \frac{L_{xx}}{n-1} = \frac{7330}{19} = 385.79$$

$$s_y^2 = \frac{L_{yy}}{n-1} = \frac{8130}{19} = 427.89$$

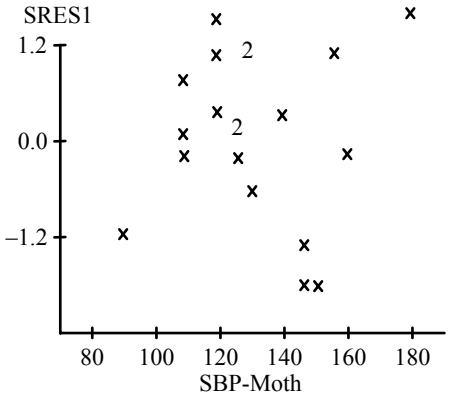
The standardized regression coefficients are given by

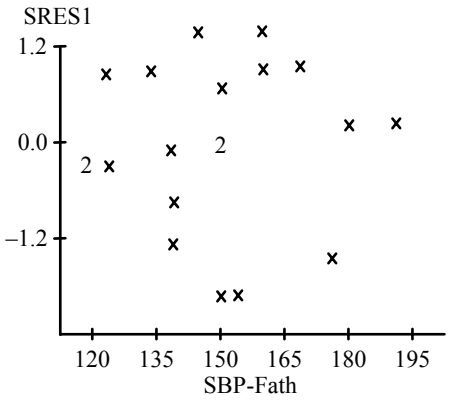
$$\text{Father: } b_{s_1} = b_1 \times \frac{s_x}{s_t} = 0.4150 \times \sqrt{\frac{385.79}{252.89}} = 0.513$$

$$\text{Mother: } b_{s_2} = b_2 \times \frac{s_y}{s_t} = 0.4225 \times \sqrt{\frac{427.89}{252.89}} = 0.550$$

The standardized coefficients represent the increase in the number of standard deviations in the child's SBP that would be expected per standard-deviation increase in the parent's SBP after controlling for the other spouse's SBP.

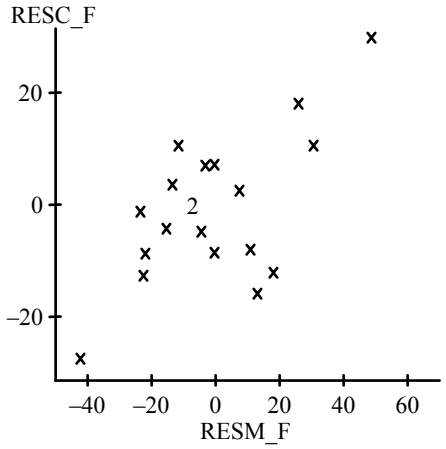
11.41 To assess goodness of fit, we first plot the Studentized residuals from the multiple-regression model (SRES1) versus the mother's and father's SBP, respectively (see the following figures). The plots do not reveal any violation of regression assumptions.



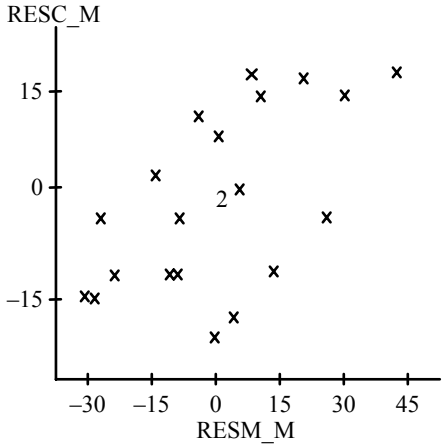


We also obtain partial-residual plots to assess the relationship between the child's SBP and one parent's SBP, after controlling each for the effects of the other parent's SBP. Since the Student version of MINITAB does not support partial-residual plots, we construct them ourselves. Specifically, to obtain the partial-residual plot of child's SBP on mother's SBP, we (1) run a regression of child's SBP on father's SBP and obtain residuals (labeled RESC_F), (2) run a regression of mother's SBP on father's SBP and obtain residuals (labeled RESM_F), and (3) generate a scatter plot of RESC_F on RESM_F. In a similar manner, we obtain a partial-residual plot of child's SBP on father's SBP after controlling for mother's SBP. Both plots appear approximately linear with no outliers or violation of regression assumptions.

Partial-residual Plot of child's SBP on mother's SBP after controlling each for father's SBP



Partial-residual plot of child's SBP on father's SBP after controlling each for mother's SBP



11.42 The one-sample *t* test for correlation.

11.43 We test the hypothesis $H_0: \rho = 0$ versus $H_1: \rho \neq 0$. We have the test statistic

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{.27\sqrt{179}}{\sqrt{1-.27^2}}$$
$$= \frac{3.612}{0.963} = 3.75 \sim t_{179}$$

Since $t > t_{120,\ .9995} = 3.373 > t_{179,\ .9995}$, it follows that $p < .001$. Therefore, we reject H_0 and conclude that the correlation coefficient between polyunsaturated fat intake at years 1 and 3 is significantly greater than zero.

11.44 A 95% CI for $\rho = (\rho_1,\ \rho_2)$, where

$$\rho_1 = \frac{\exp(2z_1) - 1}{\exp(2z_1) + 1}$$

$$\rho_2 = \frac{\exp(2z_2) - 1}{\exp(2z_2) + 1}$$

$$z_1 = z - 1.96\sqrt{\frac{1}{n-3}}$$

$$z_2 = z + 1.96\sqrt{\frac{1}{n-3}}$$

From Table 13 (Appendix, text), the z transform corresponding to $r = .27$ is 0.277. Thus,

$$z_1 = 0.277 - 1.96\sqrt{\frac{1}{178}} = 0.277 - 0.147 = 0.130$$

$$z_2 = 0.277 + 0.147 = 0.424$$

Thus,

$$\rho_1 = \frac{\exp(0.260) - 1}{\exp(0.260) + 1} = \frac{0.297}{2.297} = .129$$

$$\rho_2 = \frac{\exp(0.848) - 1}{\exp(0.848) + 1} = \frac{1.334}{3.334} = .400$$

Thus, the 95% CI for $\rho = (.129,\ .400)$

11.45 No. The two-sample z test can only be used for independent samples. Since in this case the samples are derived from the same subjects, they are not independent samples. Instead, the two-sample test for comparing dependent correlation coefficients must be used (see Section 11.8.6, text).

11.46 The two-sample *t* test with equal variances.

11.47 We can use a multiple-regression model of the form

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$$

where

y_i = SBP for the *i*th subject

x_{1i} = age for the *i*th subject

x_{2i} = alcohol-intake group for the *i* th subject

= 1 if alcohol intake $\geq$ 30 g/day

= 0 if alcohol intake = 0 g/day

$e_i \sim N\left(0,\ \sigma^2\right)$

We wish to test the hypothesis $H_0:\ \beta_2 = 0,\ \beta_1 \neq 0$ versus $H_1:\ \beta_2 \neq 0,\ \beta_1 \neq 0$. β_2 is the partial-regression coefficient for alcohol intake and represents the difference in mean SBP between the high-intake and no-intake groups for subjects of the same age.

11.48 We can use a multiple-regression model with two indicator variables for alcohol intake of the form

$$y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e_i$$

where

y_i = SBP for the *i*th subject

x_{1i} = age for the *i*th subject

x_{2i} = 1 if alcohol intake $\geq$ 30 g / day for the *i*th subject

= 0 otherwise

x_{3i} = 1 if alcohol intake > 0 and < 30 g / day for the *i*th subject

= 0 otherwise

$e_i \sim N(0,\ \sigma^2)$

We wish to test the hypothesis $H_0:\ \beta_2 = \beta_3 = 0,\ \beta_1 \neq 0$ versus H_1: at least one of $(\beta_2,\ \beta_3) \neq 0,\ \beta_1 \neq 0$.

β_2 = the mean difference in SBP between the high-intake and the no-intake group for subjects of the same age. β_3 = the mean difference in SBP between the low-intake and the no-intake group for subjects of the same age (see Section 12.5.3, in Chapter 12, text) for more details on this type of modeling).

11.49 The rank correlation is a useful measure here, because the distribution of the variances is quite skewed and very far from being normal.

11.50 We first rank the two sets of scores as follows:

Between-visit variance (year 1)	Frequency	Rank range	Average rank
2.0	1	1	1.0
2.6	1	2	2.0
5.4	1	3	3.0
6.2	1	4	4.0
6.7	1	5	5.0
8.8	1	6	6.0
10.0	2	7–8	7.5
13.3	1	9	9.0
30.9	1	10	10.0
35.0	1	11	11.0
37.0	1	12	12.0
46.6	1	13	13.0
51.2	1	14	14.0
61.0	1	15	15.0

Between-visit variance (year 2)	Frequency	Rank range	Average rank
3.9	1	1	1.0
10.1	1	2	2.0
11.9	1	3	3.0
14.6	1	4	4.0
15.2	1	5	5.0
16.7	1	6	6.0
17.9	1	7	7.0
18.7	1	8	8.0
19.0	1	9	9.0
21.4	1	10	10.0
21.6	1	11	11.0
22.7	1	12	12.0
25.7	1	13	13.0
31.3	1	14	14.0
95.7	1	15	15.0

The ranks are then displayed for each person as follows:

Person number	Between-visit variance (year 1 rank)	Between-visit variance (year 2 rank)
1	6.0	9.0
2	7.5	13.0
3	5.0	7.0
4	9.0	15.0
5	7.5	6.0
6	4.0	5.0
7	11.0	1.0
8	14.0	3.0
9	10.0	4.0
10	15.0	11.0
11	3.0	14.0
12	13.0	10.0
13	12.0	2.0
14	2.0	8.0
15	1.0	12.0

We now compute the Spearman rank-correlation coefficient as follows:

$$r_s = \frac{L_{xy}}{\left(L_{xx}L_{yy}\right)^{1/2}} = \frac{-91.5}{[279.5(280)]^{1/2}}$$
$$= \frac{-91.5}{279.75} = -.327$$

We then have the following test statistic:

$$t_s = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} = \frac{-.327\sqrt{13}}{\sqrt{1-(-0.327)^2}} = \frac{-1.179}{0.945}$$
$$= -1.248 \sim t_{13} \text{ under } H_0$$

Since $t_{13,\ .975} = 2.160 > |t_s|$, it follows that the results are not statistically significant.

11.51 The results are consistent with the theory that blood-pressure variability is not a characteristic of specific people, but rather is common among all people. Otherwise, there should have been a strong positive rank correlation. Indeed, the estimated correlation in this small group of people is negative.

11.52 We use the sample size formula in Equation 11.26 (in Chapter 11, text):

$$n = \frac{\left(z_{1-\alpha} + z_{1-\beta}\right)^2}{z_0^2} + 3$$

We have $\alpha = .05$, $\beta = .20$ and $z_0 = 0.310$ (from Table 13, Appendix, text). Thus,

$$n = \frac{(z_{.95} + z_{.80})^2}{(0.310)^2} + 3 = \frac{(1.645 + 0.84)^2}{(0.310)^2} + 3 = 67.5.$$

Therefore, 68 subjects need to be studied to achieve 80% power.

11.53 We use the power formula in Equation 11.26 (in Chapter 11, text):

$$\text{Power} = \Phi\left(\sqrt{n-3}\, z_0 - z_{1-\alpha}\right)$$

We have $n = 100$, $z = 0.310$, $z_{1-\alpha} = z_{.95} = 1.645$. Thus,

$$\text{Power} = \Phi\left[\sqrt{97}(0.310) - 1.645\right] = \Phi(1.403) = .920.$$

Thus, the study would have 92% power.

REFERENCES

[1] Dec, G. W., Palacios, I. F., Fallon, J. T, Aretz, H. T, Mills, J., Lee, D. C. S., & Johnson, R. A. (1985). Active myocarditis in the spectrum of acute dilated cardiomyopathies: Clinical features, histologic correlates and clinical outcome. *New England Journal of Medicine*, **312**(14), 885–890.

[2] Stein, A. D., Shea, S., Basch, C. E., Contento, I. R., & Zybert, P. (1991, December). Variability and tracking of nutrient intakes of preschool children based on multiple administrations of the 24-hour dietary recall. *American Journal of Epidemiology*, **134**(12), 1427–1437.