Appendix

Population and Sampling Distributions

Statisticians view sample data as representative of some real or imaginary population. A randomly selected sample may represent a real population, but even if the sample is nonrandom or the population is fictitious, the concept of representative samples remains analytically useful.

This appendix introduces expectation algebra, which offers tools for describing theoretical distributions, like the population distribution of a variable or the distribution of a statistic across all possible random samples (*sampling distribution*). Expectation algebra defines basic statistical properties of estimators, such as ordinary least squares (OLS). We conclude with a look at some important theoretical sampling distributions.

Expected Values

A sample mean equals the sum of values for each case, divided by the number of cases. This definition does not work well for populations, which may have an

infinite number of cases. To define population means and related parameters, we turn to the algebra of expectations.

The *expected value* of variable Y, written $E[Y]$, is its population mean:

$$E[Y] = \mu_Y$$

A *discrete* variable can take on only certain distinct values. For discrete distributions, the expected value may be defined as the sum of each possible Y value times its probability:

$$E[Y] = \sum_{i=1}^{I} Y_i f(Y_i) = \mu_Y \qquad [A1.1]$$

where I is the number of discrete Y values. The probability (or population proportion) of Y value Y_i is denoted by $f(Y_i)$. $f(Y)$ is the *probability density* function of Y.

Continuous variables are not limited to discrete values (although they may be restricted to a certain range). Between any two continuous-variable values, infinitely many other values exist. We need calculus to define the expected value of a continuous variable:

$$E[Y] = \int_{-\infty}^{\infty} Yf(Y)\,dY = \mu_Y \qquad [A1.2]$$

Again, $f(Y)$ refers to the probability density function of Y.

We can define population means, but we often cannot calculate them because the probability density function is unknown. The sample mean

$$\bar{Y} = \frac{\Sigma Y_i}{n} \qquad [A1.3]$$

is an *unbiased estimator* of the population mean, $E[Y]$. (A later section defines "unbiased estimator.")

Expectation has the following properties. For any constants a and b, and any variables X and W:

$$E[a] = a \qquad [A1.4]$$

$$E[bX] = bE[X] \qquad [A1.5]$$

$$E[X + W] = E[X] + E[W] \qquad [A1.6]$$

Sample means have similar properties. The mean of a constant is the constant ([A1.4]); the mean of a constant times a variable equals the constant times the variable's mean ([A1.5]); and the mean of the sum of two variables equals the sum of the two variables' means ([A1.6]).

The basic properties of expectation lead to further deductions. For example, the expectation of any linear function of X equals the same linear function of $E[X]$:

$$\begin{aligned} E[a + bX] &= E[a] + E[bX] \\ &= a + bE[X] \end{aligned}$$

The expectation of a linear function of X and W is

$$\begin{aligned} E[bX + cW] &= E[bX] + E[cW] \\ &= bE[X] + cE[W] \end{aligned}$$

The expectation operator, $E[\ ]$, can be applied in this manner to any linear equation.

Covariance

Covariance, a basic measure of the relationship between X and Y, equals the expected value of the cross product [X minus the mean of X, times Y minus the mean of Y]:

$$\begin{aligned} \text{Cov}[X, Y] &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY] - E[X]E[Y] \end{aligned} \qquad \text{[A1.7]}$$

An unbiased estimator of the population covariance is the sample covariance, s_{XY}:

$$s_{XY} = \frac{\Sigma\{(X_i - \bar{X})(\mathrm{Y}_i - \bar{Y})\}}{n - 1} \qquad \text{[A1.8]}$$

Algebraic properties of covariance follow from those of expectation. For any variables X, Y, and W, and any constants a and b:

$$\text{Cov}[a, Y] = 0 \qquad \text{[A1.9]}$$

$$\text{Cov}[bX, Y] = b\,\text{Cov}[X, Y] \qquad \text{[A1.10]}$$

$$\text{Cov}[X + W, Y] = \text{Cov}[X, Y] + \text{Cov}[W, Y] \qquad \text{[A1.11]}$$

Sample covariances have similar properties.

The covariance of a variable with itself equals the *variance* (discussed in the next section):

$$\text{Cov}[X, X] = \text{Var}[X] \qquad \text{[A1.12]}$$

Covariance is unaffected by the addition of a constant to either variable:

$$\begin{aligned}\text{Cov}[a + X, Y] &= \text{Cov}[a, Y] + \text{Cov}[X, Y] \\ &= \text{Cov}[X, Y]\end{aligned}$$

Covariances between sums of variables reduce to sums of covariances between their components:

$$\text{Cov}[X + W, Y] = \text{Cov}[X, Y] + \text{Cov}[W, Y]$$

or

$$\begin{aligned}\text{Cov}[X, Y - X] &= \text{Cov}[X, Y] - \text{Cov}[X, X] \\ &= \text{Cov}[X, Y] - \text{Var}[X]\end{aligned}$$

Variance

Expectation locates a distribution's center. Variance measures variation around that center. The variance of X is the expected value of squared deviations from the mean:

$$\begin{aligned}\text{Var}[X] &= \text{Cov}[X, X] \\ &= E[X - E[X]]^2\end{aligned} \qquad \text{[A1.13]}$$

The sample variance s_X^2 provides an unbiased estimator of $\text{Var}[X]$:

$$s_X^2 = \frac{\Sigma(X_i - \bar{X})^2}{n - 1} \qquad \text{[A1.14]}$$

Basic algebraic properties include the following:

$$\text{Var}[a] = 0 \qquad \text{[A1.15]}$$

$$\text{Var}[bX] = b^2 \text{Var}[X] \qquad \text{[A1.16]}$$

$$\text{Var}[X + W] = \text{Var}[X] + \text{Var}[W] + 2\,\text{Cov}[X, W] \qquad \text{[A1.17]}$$

These properties follow from the definitions of expectation and covariance. We can combine them for further applications, such as the variance of a linear function of X:

$$\begin{aligned}\text{Var}[a + bX] &= \text{Var}[a] + \text{Var}[bX] \\ &= b^2 \text{Var}[X]\end{aligned}$$

or the variance of a linear function of X and W:

$$\text{Var}[bX + cW] = b^2 \text{Var}[X] + c^2 \text{Var}[W] + 2bc \text{Cov}[X, W]$$

Sample variances obey similar rules.

Further Definitions

A *covariance matrix* (also called a *variance-covariance matrix*) displays the variances and covariances of many variables at once. With three variables, a covariance matrix has the form

$$\begin{matrix} \text{Var}[Y] & \text{Cov}[Y, X] & \text{Cov}[Y, W] \\ \text{Cov}[X, Y] & \text{Var}[X] & \text{Cov}[X, W] \\ \text{Cov}[W, Y] & \text{Cov}[W, X] & \text{Var}[W] \end{matrix}$$

Covariance is symmetrical: $\text{Cov}[X, Y] = \text{Cov}[Y, X]$. Consequently, the upper triangular part of a covariance matrix repeats the information of the lower triangle. Variances make up the *major diagonal* (upper left to lower right).

Many univariate, bivariate, and multivariate analyses require only the information in the covariance matrix. The raw data are not needed to find standard deviations, correlations, or multiple regression coefficients, for instance. Expectation, covariance, and variance are statistical building blocks from which other statistics can be defined. For example:

1. The *standard deviation* (σ) equals the square root of the variance:

$$\sigma_X = \sqrt{\text{Var}[X]} \qquad \text{[A1.18]}$$

2. *Correlation* (ρ_{XY}) is a standardized covariance:[1]

$$\rho_{XY} = \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X] \text{Var}[Y]}} \qquad \text{[A1.19]}$$

Correlation equals covariance if both variables are standard scores, with variances of 1 ([3.16]). A correlation matrix is laid out like a covariance matrix:

$$\begin{matrix} 1 & \rho_{YX} & \rho_{YW} \\ \rho_{XY} & 1 & \rho_{XW} \\ \rho_{WY} & \rho_{WX} & 1 \end{matrix}$$

Like covariance, correlation is symmetrical, so cells above the diagonal are redundant. A correlation matrix contains less information than a covariance matrix. It gives no indication of the variables' original scales.

3. In a two-variable regression the OLS *slope* or *regression coefficient* is

$$\beta_1 = \frac{\text{Cov}[X, Y]}{\text{Var}[X]} \qquad \text{[A1.20]}$$

The *Y-intercept* is

$$\beta_0 = E[Y] - \beta_1 E[X] \qquad \text{[A1.21]}$$

Equations [A1.20]–[A1.21] are for the *regression of Y on X*. A different slope and intercept result if we regress X on Y; unlike covariance and correlation, regression is not symmetrical.[2]

4. Multiple regression slopes (*partial regression coefficients*) can likewise be defined from variances and covariances. In the regression of Y on X and W, the OLS coefficient on X is

$$\beta_{YX.W} = \frac{\text{Var}[W]\,\text{Cov}[X, Y] - \text{Cov}[X, W]\,\text{Cov}[W, Y]}{\text{Var}[X]\,\text{Var}[W] - (\text{Cov}[X, W])^2} \qquad \text{[A1.22]}$$

Here $\beta_{YX.W}$ denotes "the regression of Y on X, controlling for W." (Simpler subscripting is employed elsewhere in this book.) The coefficient on W in the same regression will be

$$\beta_{YW.X} = \frac{\text{Var}[X]\,\text{Cov}[W, Y] - \text{Cov}[X, W]\,\text{Cov}[X, Y]}{\text{Var}[X]\,\text{Var}[W] - (\text{Cov}[X, W])^2} \qquad \text{[A1.23]}$$

We obtain unbiased estimates of these population parameters by substituting sample covariance and variance ([A1.8] and [A1.14]) into [A1.18]–[A1.23].

Continuing in this manner, using variance and covariance to define higher-order regression leads to unwieldy equations. Matrix algebra (Appendix 3) provides a better way to work with the large amounts of information needed for multivariate calculations.

Properties of Sampling Distributions

The *sampling distribution* of a statistic is its theoretical distribution over all possible random samples of a given size. Researchers typically have just one sample, but theories about sampling distributions guide what inferences they draw from their sample.

Sample statistics provide estimates of population parameters. A statistic b is an *unbiased* estimator of parameter β if

$$E[b] = \beta \qquad \text{[A1.24]}$$

This means that, over all possible random samples of size n, the average value of b equals the parameter we are trying to estimate. *Bias* equals the expected difference between statistic and parameter:

$$\text{bias} = E[b - \beta] \qquad \text{[A1.25]}$$

A biased estimator ($E[b] \neq \beta$, or $E[b - \beta] \neq 0$) tends on average to be too high or too low, relative to the population parameters. Other things being equal, we prefer unbiased estimators.

The *variance* of an estimator equals average squared deviation around its mean:

$$\text{Var}[b] = E[b - E[b]]^2 \qquad \text{[A1.26]}$$

The theoretical *standard error* equals the square root of the variance:

$$\sigma_b = \sqrt{\text{Var}[b]} \qquad \text{[A1.27]}$$

Other things being equal, we prefer an estimator with a low variance or standard error.

The *mean squared error* is the expected value of squared deviations from the parameter:

$$\begin{aligned} \text{MSE} &= E[b - \beta]^2 \\ &= \text{Var}[b] + \text{bias}^2 \end{aligned} \qquad \text{[A1.28]}$$

Note that, if b is an unbiased estimator,

$$\text{MSE} = \text{Var}[b]$$

Among unbiased estimators, the one with the lowest MSE or variance is the most efficient.

These definitions of bias, variance, and mean squared error apply to sampling distributions based on samples of any fixed, finite size. They are called *small-sample* properties. Sometimes the small-sample properties of an estimator are unknown, but we do know its *large-sample* or *asymptotic* properties: theoretical behavior as sample size approaches infinity. For example, b is an *asymptotically unbiased* estimator of β if

$$\lim_{n \to \infty} E[b] = \beta \qquad \text{[A1.29]}$$

This means that, as n (sample size) approaches infinity, the expected value of b approaches β. β is the mean of the *limiting distribution* of b.

For many estimators the sampling distribution would collapse to a single point, with zero variance, if sample size *reached* infinity. *Asymptotic variance* is the variance as sample size *approaches* infinity. Among asymptotically unbiased

estimators of any parameter, that with the lowest asymptotic variance is called *asymptotically efficient*. *Asymptotic normality* means that, as $n \to \infty$, the distribution of b approaches normality.

We call b a *consistent* estimator of β if the probability that b is very close to β approaches 1 as n approaches infinity. Formally, for any small constant c:

$$\lim_{n \to \infty} P[|b - \beta| < c] = 1 \qquad \text{[A1.30]}$$

Consistency implies that, as sample size increases, we can be increasingly confident that b is close to β.

Ordinary Least Squares

We begin with a linear model for expected Y_i:

$$E[Y_i] = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{K-1} X_{i,K-1} \qquad \text{[A1.31]}$$

and view X values as (hypothetically) fixed in repeated sampling. ε_i represents errors, which cause individual Y_i to differ from $E[Y_i]$:

$$Y_i = E[Y_i] + \varepsilon_i$$

If errors have zero mean:

$$E[\varepsilon_i] = 0 \qquad \text{for all } i \qquad \text{[A1.32]}$$

then ordinary least squares (OLS) estimates of the β parameters in [A1.31] should be unbiased:

$$E[b_k] = \beta_k \qquad \text{for } k = 0, 1, 2, \ldots, K-1$$

To assess the likelihood of being close to the parameter, we want to know the estimator's variance. OLS variance and standard error estimates rest on two further assumptions:

Homoscedasticity: Errors have the same variance for every case:

$$\text{Var}[\varepsilon_i] = \text{Var}[\varepsilon] \qquad \text{for all } i \qquad \text{[A1.33]}$$

No autocorrelation: Errors for case i are independent of errors for case j:

$$\text{Cov}[\varepsilon_i, \varepsilon_j] = 0 \qquad \text{for } i \neq j \qquad \text{[A1.34]}$$

If [A1.33] is false (*heteroscedasiticity*) or if [A1.34] is false (*autocorrelation*), estimated standard errors may be biased, invalidating the usual hypothesis tests and confidence intervals.

Assuming fixed X and [A1.32]–[A1.34], the *Gauss-Markov Theorem* demonstrates that OLS is the most efficient linear unbiased estimator. Note the restriction to *linear unbiased estimators*; nonlinear or biased estimators sometimes perform better, even assuming [A1.31]–[A1.34].

Some Theoretical Distributions

Confidence intervals and hypothesis tests build upon theoretical sampling distributions. The normal (Gaussian) and three related distributions (χ^2, t, and F) are widely used.

The normal or Gaussian distribution is a theoretical probability distribution for variables that represent sums of many independent, identically distributed random variables. Certain sample statistics, such as means or regression coefficients, can be viewed in this way and hence should have approximately normal sampling distributions—given large enough samples.

Normal distributions are defined by the probability density function $f(Y)$:

$$f(Y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(Y-\mu)^2/2\sigma^2} \qquad \text{[A1.35]}$$

Graphing $f(Y)$ against Y produces the familiar bell shape of the normal curve. There exists a unique normal distribution for every possible combination of mean (μ) and standard deviation (σ). The mean is the distribution's center, and standard deviation measures spread around this center.

A *standard normal distribution* has a mean of 0 and a standard deviation of 1. If Y is distributed normally with mean μ and standard deviation σ, then

$$Z = \frac{Y - \mu}{\sigma} \qquad \text{[A1.36]}$$

follows a standard normal distribution (Figure A1.1). All normal distributions range from negative to positive infinity; only the range $\mu \pm 5\sigma$ is shown in Figure A1.1.

Sums of independent squared standard normal variables follow a *chi-square* (χ^2) distribution. If each Z_d is an independent standard normal variable ($d = 1, 2, \ldots \text{df}$), then

$$\chi^2_{\text{df}} = \sum_{d=1}^{\text{df}} Z_d^2 \qquad \text{[A1.37]}$$

follows a χ^2 distribution with df degrees of freedom. There is a different χ^2 distribution for every possible value of df. Unlike normal variables, χ^2 variables can

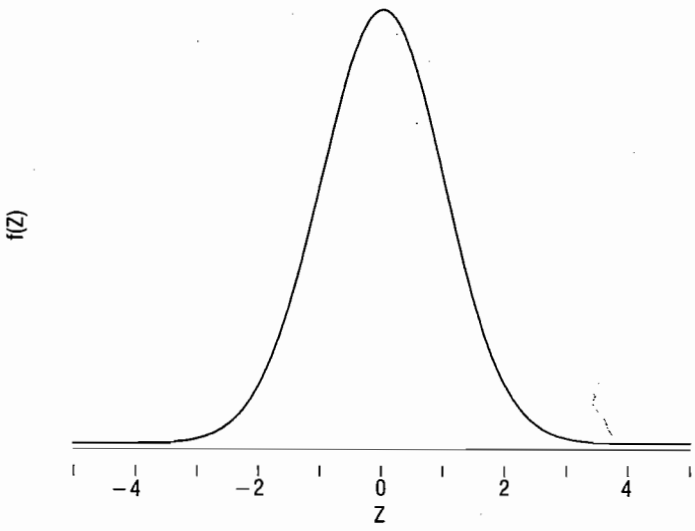

Figure A1.1 Standard normal distribution.

only be positive, and χ^2 distributions are positively skewed (Figure A1.2]. Their skew is most pronounced at low degrees of freedom.

If Z is a standard normal variable and s^2 is a chi-square variable, independent of Z and with df degrees of freedom, then the ratio

$$t = \frac{Z}{s/\sqrt{\text{df}}} \qquad \text{[A1.38]}$$

follows a t-distribution with df degrees of freedom. We construct t-statistics for hypothesis tests by expressing the distance between statistic and hypothesized parameter in estimated standard errors. t-distributions resemble the standard normal distribution but have heavier tails, especially with few degrees of freedom (Figure A1.3). With many degrees of freedom, t-distributions become approximately standard normal.

The ratio of two independent chi-square variables, each divided by its degrees of freedom, follows an *F-distribution.* If $s_1^2 \sim \chi^2_{\text{df}1}$ and $s_2^2 \sim \chi^2_{\text{df}2}$:

$$F^{\text{df}1}_{\text{df}2} = \frac{s_1^2/\text{df}1}{s_2^2/\text{df}2} \qquad \text{[A1.39]}$$

Two parameters, the numerator and denominator degrees of freedom, characterize an F-distribution. Like χ^2, an F-variable cannot be negative. F-tests can compare two estimated variances, such as the explained and residual variance in OLS.

Appendix 4 includes tables of critical values for the theoretical t- (Table A4.1), F- (Table A4.2), and χ^2- (Table A4.3) distributions. Standard normal distribution (Z) critical values equal the t_∞ critical values of Table A4.1.

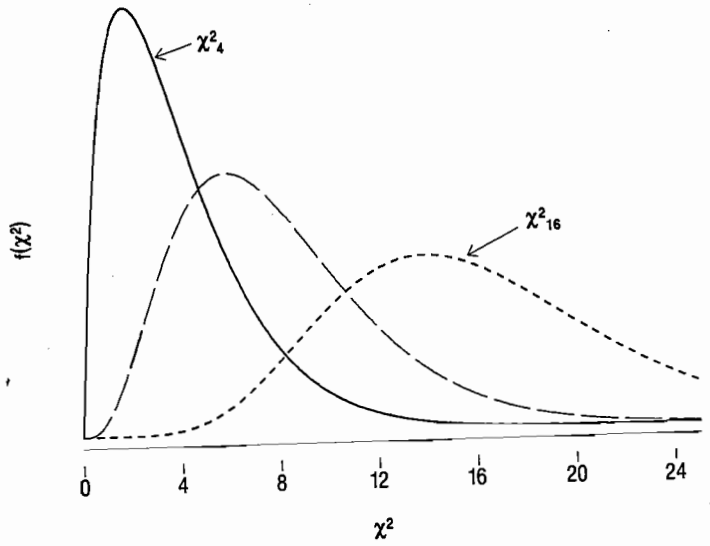

Figure A1.2 Chi-square (χ^2) distributions with 4, 8, and 16 df.

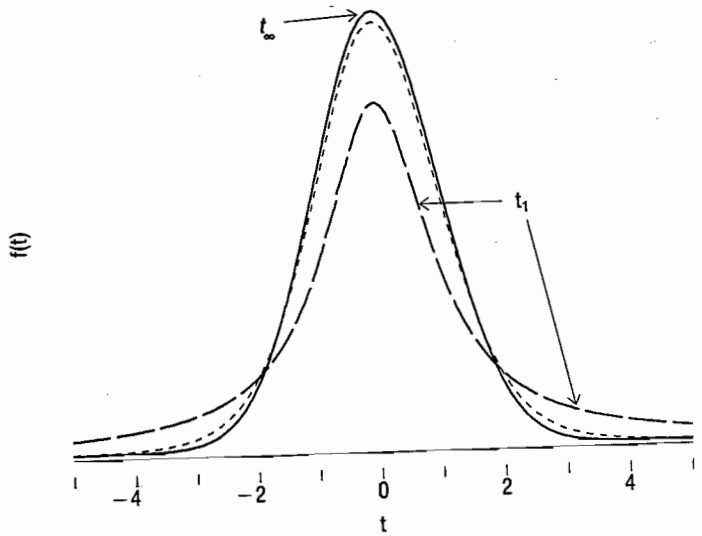

Figure A1.3 *t*-distributions with 1, 10, and infinite df.

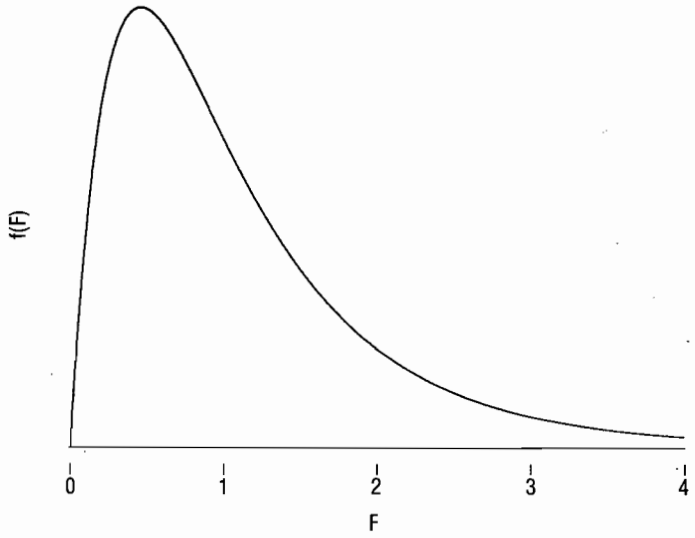

Figure A1.4 F-distribution with $df_1 = 4$ and $df_2 = 16$.

Exercises

1. Apply rules of expectation and covariance to simplify:
   a. $E[1201 + 47.5X]$
   b. $E[-11.4 + .176X + 3.9W]$
   c. $\text{Cov}[Y, 1201 + 47.5X]$
   d. $\text{Cov}[Y, -11.4 + .176X + 3.9W]$
   e. $\text{Var}[1201 + 47.5X]$
2. Use the properties of expectations ([A1.4]–[A1.6]) and the definition of covariance ([A1.7]) to derive Equations [A1.9]–[A1.11].
3. Use properties of expectation and covariance to derive Equations [A1.15]–[A1.17].
4. We wish to know the true values of a variable X, but our measurements contain some error (as all measurements do). Let X represent the true values, $\tilde{X}$ our *measured* values, and ε the errors. (For example, X might be individuals' true incomes, $\tilde{X}$ the incomes they claim in a survey, and ε whatever difference there is between reported ($\tilde{X}$) and true (X) incomes.) Measured values equal the true values plus error:

   $$\tilde{X} = X + \varepsilon$$

   where $\tilde{X}$, X, and ε are all variables.
   a. What is the expected value of $\tilde{X}$? What therefore is required for $\tilde{X}$ to be an "unbiased estimator" of X, in the sense that both have the same expectation?

b. What is the variance $\tilde{X}$?
c. Assume that errors are *random*, meaning that $\text{Cov}[X, \varepsilon] = 0$. What does this imply about the relative variation in measured ($\tilde{X}$) and true (X) values?
d. If errors are *not* random and $\text{Cov}[X, \varepsilon] \neq 0$, does your generalization of part c still hold?

Notes

1. The Greek letter ρ (rho) traditionally denotes population correlations (sample correlations are denoted by r). In robustness literature (Chapter 6), ρ has a different, unrelated meaning.
2. If X, rather than Y, is the left-hand-side variable (*regression of X on Y*), the slope becomes

$$\beta_1 = \text{Cov}[X, Y]/\text{Var}[Y]$$

instead of [A1.20]. Similarly, the intercept becomes

$$\beta_0 = E[X] - \beta_1 E[Y]$$

instead of [A1.21].