the form of interval estimation for the model parameters and for values of the dependent variable was considered in Sections 4.5 and 4.6. This development will continue in Chapter 5 where we will consider hypothesis testing and model selection.

Finally, we considered some practical problems that arise when data are less than perfect for the estimation and analysis of the regression model, including multicollinearity, missing observations, measurement error, and outliers.

Key Terms and Concepts

- Assumptions
- Asymptotic covariance matrix
- Asymptotic distribution
- Asymptotic efficiency
- Asymptotic normality
- Asymptotic properties
- Attrition
- Bootstrap
- Condition number
- Confidence interval
- Consistency
- Consistent estimator
- Data imputation
- Efficient scale
- Estimator
- Ex ante forecast
- Ex post forecast
- Finite sample properties
- Gauss–Markov theorem
- Grenander conditions
- Highest posterior density interval
- Identification
- Ignorable case
- Inclusion of superfluous (irrelevant) variables
- Indicator
- Interval estimation
- Least squares attenuation
- Lindeberg–Feller Central Limit Theorem
- Linear estimator
- Linear unbiased estimator
- Maximum likelihood estimator
- Mean absolute error
- Mean square convergence
- Mean squared error
- Measurement error
- Method of moments
- Minimum mean squared error
- Minimum variance linear unbiased estimator
- Missing at random
- Missing completely at random
- Missing observations
- Modified zero-order regression
- Monte Carlo study
- Multicollinearity
- Not missing at random
- Oaxaca's and Blinder's decomposition
- Omission of relevant variables
- Optimal linear predictor
- Orthogonal random variables
- Panel data
- Pivotal statistic
- Point estimation
- Prediction error
- Prediction interval
- Prediction variance
- Pretest estimator
- Principal components
- Probability limit
- Root mean squared error
- Sample selection
- Sampling distribution
- Sampling variance
- Semiparametric
- Smearing estimator
- Specification errors
- Standard error
- Standard error of the regression
- Stationary process
- Statistical properties
- Stochastic regressors
- Theil U statistic
- t ratio
- Variance inflation factor
- Zero-order method

Exercises

1. Suppose that you have two independent unbiased estimators of the same parameter θ, say $\hat{\theta}_1$ and $\hat{\theta}_2$, with different variances v_1 and v_2. What linear combination $\hat{\theta} = c_1\hat{\theta}_1 + c_2\hat{\theta}_2$ is the minimum variance unbiased estimator of θ?
2. Consider the simple regression $y_i = \beta x_i + \varepsilon_i$ where $E[\varepsilon \mid x] = 0$ and $E[\varepsilon^2 \mid x] = \sigma^2$
   a. What is the minimum mean squared error linear estimator of β? [*Hint:* Let the estimator be $(\hat{\beta} = \mathbf{c}'\mathbf{y})$. Choose $\mathbf{c}$ to minimize $\text{Var}(\hat{\beta}) + (E(\hat{\beta} - \beta))^2$. The answer is a function of the unknown parameters.]

b. For the estimator in part a, show that ratio of the mean squared error of $\hat{\beta}$ to that of the ordinary least squares estimator b is

$$\frac{\text{MSE}[\hat{\beta}]}{\text{MSE}[b]} = \frac{\tau^2}{(1+\tau^2)}, \quad \text{where } \tau^2 = \frac{\beta^2}{[\sigma^2/\mathbf{x}'\mathbf{x}]}.$$

Note that τ is the square of the population analog to the "t ratio" for testing the hypothesis that $\beta = 0$, which is given in (5-11). How do you interpret the behavior of this ratio as $\tau \to \infty$?

3. Suppose that the classical regression model applies but that the true value of the constant is zero. Compare the variance of the least squares slope estimator computed without a constant term with that of the estimator computed with an unnecessary constant term.
4. Suppose that the regression model is $y_i = \alpha + \beta x_i + \varepsilon_i$, where the disturbances ε_i have $f(\varepsilon_i) = (1/\lambda)\exp(-\varepsilon_i/\lambda), \varepsilon_i \geq 0$. This model is rather peculiar in that all the disturbances are assumed to be nonnegative. Note that the disturbances have $E[\varepsilon_i \mid x_i] = \lambda$ and $\text{Var}[\varepsilon_i \mid x_i] = \lambda^2$. Show that the least squares slope is unbiased but that the intercept is biased.
5. Prove that the least squares intercept estimator in the classical regression model is the minimum variance linear unbiased estimator.
6. As a profit-maximizing monopolist, you face the demand curve $Q = \alpha + \beta P + \varepsilon$. In the past, you have set the following prices and sold the accompanying quantities:

Q	3	3	7	6	10	15	16	13	9	15	9	15	12	18	21
P	18	16	17	12	15	15	4	13	11	6	8	10	7	7	7

Suppose that your marginal cost is 10. Based on the least squares regression, compute a 95 percent confidence interval for the expected value of the profit-maximizing output.

7. The following sample moments for $x = [1, x_1, x_2, x_3]$ were computed from 100 observations produced using a random number generator:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 100 & 123 & 96 & 109 \\ 123 & 252 & 125 & 189 \\ 96 & 125 & 167 & 146 \\ 109 & 189 & 146 & 168 \end{bmatrix}, \quad \mathbf{X}'\mathbf{y} = \begin{bmatrix} 460 \\ 810 \\ 615 \\ 712 \end{bmatrix}, \quad \mathbf{y}'\mathbf{y} = 3924.$$

The true model underlying these data is $y = x_1 + x_2 + x_3 + \varepsilon$.

a. Compute the simple correlations among the regressors.
b. Compute the ordinary least squares coefficients in the regression of y on a constant x_1, x_2, and x_3.
c. Compute the ordinary least squares coefficients in the regression of y on a constant, x_1 and x_2, on a constant, x_1 and x_3, and on a constant, x_2 and x_3.
d. Compute the variance inflation factor associated with each variable.
e. The regressors are obviously collinear. Which is the problem variable?

8. Consider the multiple regression of $\mathbf{y}$ on K variables $\mathbf{X}$ and an additional variable $\mathbf{z}$. Prove that under the assumptions A1 through A6 of the classical regression model, the true variance of the least squares estimator of the slopes on $\mathbf{X}$ is larger when $\mathbf{z}$

is included in the regression than when it is not. Does the same hold for the sample estimate of this covariance matrix? Why or why not? Assume that $\mathbf{X}$ and $\mathbf{z}$ are nonstochastic and that the coefficient on $\mathbf{z}$ is nonzero.

9. For the classical normal regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with no constant term and K regressors, assuming that the true value of $\boldsymbol{\beta}$ is zero, what is the exact expected value of $F[K, n-K] = (R^2/K)/[(1-R^2)/(n-K)]$?
10. Prove that $E[\mathbf{b}'\mathbf{b}] = \boldsymbol{\beta}'\boldsymbol{\beta} + \sigma^2 \sum_{k=1}^{K}(1/\lambda_k)$ where $\mathbf{b}$ is the ordinary least squares estimator and λ_k is a characteristic root of $\mathbf{X}'\mathbf{X}$.
11. For the classical normal regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with no constant term and K regressors, what is plim $F[K, n-K] = \text{plim } \frac{R^2/K}{(1-R^2)/(n-K)}$, assuming that the true value of $\boldsymbol{\beta}$ is zero?
12. Let e_i be the ith residual in the ordinary least squares regression of $\mathbf{y}$ on $\mathbf{X}$ in the classical regression model, and let ε_i be the corresponding true disturbance. Prove that $\text{plim}(e_i - \varepsilon_i) = 0$.
13. For the simple regression model $y_i = \mu + \varepsilon_i$, $\varepsilon_i \sim N[0, \sigma^2]$, prove that the sample mean is consistent and asymptotically normally distributed. Now consider the alternative estimator $\hat{\mu} = \sum_i w_i y_i$, $w_i = \frac{i}{(n(n+1)/2)} = \frac{i}{\sum_i i}$. Note that $\sum_i w_i = 1$. Prove that this is a consistent estimator of μ and obtain its asymptotic variance. [*Hint:* $\sum_i i^2 = n(n+1)(2n+1)/6$.]
14. Consider a data set consisting of n observations, n_c complete and n_m incomplete, for which the dependent variable, y_i, is missing. Data on the independent variables, $\mathbf{x}_i$, are complete for all n observations, $\mathbf{X}_c$ and $\mathbf{X}_m$. We wish to use the data to estimate the parameters of the linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Consider the following the imputation strategy: Step 1: Linearly regress $\mathbf{y}_c$ on $\mathbf{X}_c$ and compute $\mathbf{b}_c$. Step 2: Use $\mathbf{X}_m$ to predict the missing $\mathbf{y}_m$ with $\mathbf{X}_m\mathbf{b}_c$. Then regress the full sample of observations, $(\mathbf{y}_c, \mathbf{X}_m\mathbf{b}_c)$, on the full sample of regressors, $(\mathbf{X}_c, \mathbf{X}_m)$.
    a. Show that the first and second step least squares coefficient vectors are identical.
    b. Is the second step coefficient estimator unbiased?
    c. Show that the sum of squared residuals is the same at both steps.
    d. Show that the second step estimator of σ^2 is biased downward.
15. In (4-13), we find that when superfluous variables $\mathbf{X}_2$ are added to the regression of $\mathbf{y}$ on $\mathbf{X}_1$ the least squares coefficient estimator is an unbiased estimator of the true parameter vector, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \mathbf{0}')'$. Show that in this long regression, $\mathbf{e}'\mathbf{e}/(n - K_1 - K_2)$ is also unbiased as estimator of σ^2.
16. In Section 4.7.3, we consider regressing $\mathbf{y}$ on a set of principal components, rather than the original data. For simplicity, assume that $\mathbf{X}$ does not contain a constant term, and that the K variables are measured in deviations from the means and are "standardized" by dividing by the respective standard deviations. We consider regression of $\mathbf{y}$ on L principal components, $\mathbf{Z} = \mathbf{X}\mathbf{C}_L$, where $L < K$. Let $\mathbf{d}$ denote the coefficient vector. The regression model is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. In the discussion, it is claimed that $E[\mathbf{d}] = \mathbf{C}_L'\boldsymbol{\beta}$. Prove the claim.
17. Example 4.10 presents a regression model that is used to predict the auction prices of Monet paintings. The most expensive painting in the sample sold for \$33.0135M (log = 17.3124). The height and width of this painting were 35" and 39.4", respectively. Use these data and the model to form prediction intervals for the log of the price and then the price for this painting.