Chapter 19: Statistical Properties of the OLS Estimator

19.1 Properties of OLS

We investigate the statistical properties of the OLS estimator of β. Specifically, we calculate its exact mean and variance and compare it with other possible estimators under assumptions A1 and A2.

Definition 19.1

The estimator $\hat{\beta}$ is **linear** in y, i.e., there exists a matrix C not depending on y such that

$$\hat{\beta} = (X^T X)^{-1} X^T y = Cy.$$

This property simplifies calculations. In a time series context, where X may contain lagged values of y, this property is not so meaningful. We want to evaluate how $\hat{\beta}$ varies across hypothetical repeated samples under Assumption A.

Theorem 19.1

Suppose that assumption A1 holds. Then, we have

$$\mathbb{E}[\hat{\beta}|X] = (X^T X)^{-1} X^T \mathbb{E}[y|X] = (X^T X)^{-1} X^T X \beta = \beta,$$

where this equality holds for all β. We say that $\hat{\beta}$ is **conditionally unbiased**. It follows that it is unconditionally unbiased using the Law of Iterated Expectations.

Furthermore, we shall calculate the $K \times K$ conditional covariance matrix of $\hat{\beta}$,

$$\text{var}(\hat{\beta}|X) = \mathbb{E}\left[(\hat{\beta} - \mathbb{E}[\hat{\beta}|X])(\hat{\beta} - \mathbb{E}[\hat{\beta}|X])^T|X\right] = \mathbb{E}\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T|X\right].$$

This has diagonal elements $\text{var}(\hat{\beta}_j|X)$ and off-diagonals $\text{cov}(\hat{\beta}_j, \hat{\beta}_k|X)$.

Intuition: The conditional unbiasedness of the OLS estimator means that, on average, the estimator $\hat{\beta}$ will be equal to the true value β, given the predictor matrix X. It means that if we could repeatedly draw samples of the dependent variable y for a fixed set of predictors X, the average of the OLS estimates across these samples would converge to the true parameter value β.

Theorem 19.2

Suppose that assumptions A1 and A2 hold. Then

$$\text{var}(\hat{\beta}|X) = (X^T X)^{-1} X^T \mathbb{E}[\epsilon\epsilon^T|X] X (X^T X)^{-1} = (X^T X)^{-1} X^T \Sigma(X) X (X^T X)^{-1}.$$

In the homoskedastic special case, $\Sigma(X) = \sigma^2 I_n$, this simplifies to

$$\text{var}(\hat{\beta}|X) = \sigma^2 (X^T X)^{-1}.$$

The matrix X^TX is generally not diagonal, except in special cases such as dummy variables.

We next consider the unconditional variance, which follows also from the law of iterated expectation

$$\begin{aligned}\operatorname{var}(\hat{\beta}) &= \mathbb{E}[\operatorname{var}(\hat{\beta}|X)] \\ &= \mathbb{E}\left[(X^TX)^{-1}X^T\Sigma(X)X(X^TX)^{-1}\right] \\ &= \sigma^2\mathbb{E}\left[(X^TX)^{-1}\right] \quad \text{in the homoskedastic case,}\end{aligned}$$

provided the moment exists. Under the assumption that X is full rank with probability one we may always define $(X^TX)^{-1}$. However, this is not sufficient to guarantee the existence of the expectation of its inverse.

Intuition:

This theorem gives us the variance-covariance matrix of the OLS estimator, conditional on X. The diagonal elements represent the variances of individual coefficient estimates ($\hat{\beta}_i$), while the off-diagonal elements represent the covariances between different coefficient estimates. In the simple case of homoskedasticity (constant error variance) and no autocorrelation, the formula simplifies considerably.

Example 19.1

Suppose that $n = K = 1$ and $x_1 \sim N(0,1)$, then $\Pr(x_1^2 = 0) = 0$ but $\mathbb{E}[x_1^{-2}] = \infty$. However, when $n \geq 2$ this problem goes away because we have $\mathbb{E}[(\sum_{i=1}^n x_i^2)^{-1}] < \infty$.

However, according to the ancillarity principle we should really just consider the conditional distribution.

The properties of an individual coefficient can be obtained from the partitioned regression formula $\hat{\beta}_1 = (X_1^TM_2X_1)^{-1}X_1^TM_2y$. In general we have

$$\begin{aligned}\operatorname{var}[\hat{\beta}_1|X] &= (X_1^TM_2X_1)^{-1}X_1^TM_2\mathbb{E}[\epsilon\epsilon^T|X]M_2X_1(X_1^TM_2X_1)^{-1} \\ &= (X_1^TM_2X_1)^{-1}X_1^TM_2\Sigma(X)M_2X_1(X_1^TM_2X_1)^{-1} \qquad \text{In the special case that} \\ &= \sigma^2(X_1^TM_2X_1)^{-1} \quad \text{in the homoskedastic case.}\end{aligned}$$

$X_2 = (1,\ldots,1)^T$, we have $\qquad \operatorname{var}(\hat{\beta}_1|X) = \dfrac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$

This is the well known variance of the least squares estimator in the single regressor plus intercept regression.

Intuition: It states that the variance of a specific coefficient estimator $\hat{\beta}_1$ is influenced by:

1. The error variance (σ^2): Higher error variance leads to higher estimator variance.
2. The variation in the corresponding predictor (X_1) after accounting for the other predictors (X_2): represented by $X_1^TM_2X_1$. Greater variation (after accounting for other predictors) leads to lower estimator variance. ### Theorem 19.3

Suppose that condition A3 holds. Then, the distribution of $\hat{\beta}$ conditional on X is the multivariate normal distribution

$$\hat{\beta} \sim N(\beta, V(X)), \quad V(X) = (X^TX)^{-1}X^T\Sigma(X)X(X^TX)^{-1}.$$

Proof: This follows because $\hat{\beta} = \sum_{i=1}^n c_iy_i$, where c_i depends only on the covariates, i.e., $\hat{\beta}$ is a linear combination of independent normals, and hence it is normal, conditionally. The unconditional distribution of $\hat{\beta}$

will not be normal – in fact, it will be a scale mixture of normals. However, it follows that

$$V(X)^{-1/2}(\hat{\beta} - \beta) \sim N(0, I_K)$$

conditional on X, and because the right hand side distribution does not depend on X, this result is unconditional too. Hence, the quadratic form

$$\tau = (\hat{\beta} - \beta)^T V(X)^{-1}(\hat{\beta} - \beta) \sim \chi^2(K)$$

conditionally and unconditionally. The distribution of τ does not depend on unknown quantities, which will be helpful later when constructing hypothesis tests and confidence intervals.

Intuition:

If, in addition to the previous assumptions, we assume that the errors are normally distributed, then the OLS estimator $\hat{\beta}$ also follows a normal distribution. This result is crucial for hypothesis testing and constructing confidence intervals, as we can use the known properties of the normal distribution to make inferences about the true parameters.

We are also interested in estimation of $m = \mathbb{E}[y|X] = X\beta \in \mathbb{R}^n$ and in the estimation of the function $m(x) = \mathbb{E}[y_i|x_i = x]$ for any $x \in \mathbb{R}^K$. Let $\hat{m} = X\hat{\beta}$ and $\hat{m}(x) = x^T\hat{\beta}$.

Theorem 19.4

Suppose that assumptions A1 and A2 hold. Then $\hat{m}$ is unbiased, i.e., $\mathbb{E}[\hat{m}|X] = m$, with

$$\text{var}(\hat{m}|X) = X(X^TX)^{-1}X^T\Sigma(X)X(X^TX)^{-1}X^T.$$

Likewise, $\hat{m}(x)$ is unbiased, i.e., $\mathbb{E}[\hat{m}(x)|X] = m(x)$, with

$$\text{var}(\hat{m}(x)|X) = x(X^TX)^{-1}X^T\Sigma(X)X(X^TX)^{-1}x^T.$$

Furthermore, if assumption A3 holds, then conditionally on X,

$$\hat{m} \sim N(m, X(X^TX)^{-1}X^T\Sigma(X)X(X^TX)^{-1}X^T).$$

Note however, that $\text{var}(\hat{m}|X)$ is of rank $K < n$.

In the homoskedastic special case A4 we have conditionally

$$\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})$$

$$\hat{m} \sim N(m, \sigma^2X(X^TX)^{-1}X^T).$$

Suppose we assume condition A4, which is a complete specification of the conditional distribution of y. Then the density function of the vector y [conditional on X] is

$$f_{y|X}(y) = \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right).$$

The density function depends on the unknown parameters β, σ^2. The log likelihood function for the observed data is

$$l(b, \omega^2 | y, X) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \omega^2 - \frac{1}{2\omega^2}(y - Xb)^T(y - Xb),$$

where b and ω are unknown parameters. Perhaps we are being overly pedantic here by emphasizing the difference between the true values β, σ^2 and the arguments b and ω of l, and we shall not do this again. The maximum likelihood estimator $\hat{\beta}_{mle}, \hat{\sigma}^2_{mle}$ maximizes $l(b, \omega^2)$ with respect to b and ω^2. It is easy to see that

$$\hat{\beta}_{mle} = \hat{\beta}; \quad \hat{\sigma}^2_{mle} = \frac{1}{n}(y - X\hat{\beta}_{mle})^T(y - X\hat{\beta}_{mle}).$$

Basically, the criterion function is the least squares criterion apart from an affine transformation involving only ω. Note however, that if we had a different assumption about the errors than normality, e.g., they were from a t-distribution, then we would have a different likelihood and a different estimator than $\hat{\beta}$. In particular, the estimator may not be explicitly defined and may be a nonlinear function of y.

We next consider two widely used estimates of σ^2

$$\hat{\sigma}^2_{mle} = \frac{\hat{\epsilon}^T\hat{\epsilon}}{n}$$

$$s^2_* = \frac{\hat{\epsilon}^T\hat{\epsilon}}{n - K}.$$

The first estimate is the maximum likelihood estimator of σ^2. The second estimate is a modification of the MLE, which we can show is unbiased.

Theorem 19.5

Suppose that A1, A2 hold and that $\Sigma(X) = \sigma^2 I$. Then

$$\mathbb{E}[s^2_*] = \sigma^2.$$

Proof:

We have

$$\hat{\epsilon} = M_X y = M_X X\beta + M_X\epsilon = M_X\epsilon$$

so that $\hat{\epsilon}^T\hat{\epsilon}$ is a quadratic form in normal random variables

$$\hat{\epsilon}^T\hat{\epsilon} = y^T M_X M_X y = \epsilon^T M_X M_X \epsilon = \epsilon^T M_X \epsilon.$$

Therefore, under A2

$$\begin{aligned}\mathbb{E}[\hat{\epsilon}^T\hat{\epsilon}|X] &= \mathbb{E}[\mathrm{tr}(\epsilon^T M_X\epsilon)|X] \\ &= \mathbb{E}[\mathrm{tr}(\epsilon\epsilon^T M_X)|X] \\ &= \mathrm{tr}(\mathbb{E}[\epsilon\epsilon^T|X]M_X) \\ &= \mathrm{tr}(\Sigma(X)M_X).\end{aligned}$$

If $\Sigma(X) = \sigma^2 I$, then

$$\mathrm{tr}(\Sigma(X)M_X) = \sigma^2\mathrm{tr}(M_X) = \sigma^2(n - K).$$

Intuition:

This theorem shows that the modified estimator of the error variance, s_*^2, is unbiased. This means that, on average, this estimator will correctly estimate the true error variance. The correction factor of dividing by $(n-K)$ instead of n accounts for the degrees of freedom lost in estimating the K parameters in β.

19.1.1 Alternative Estimators

There are many possible alternative estimators to OLS that have been developed over the years. A major alternative is called the **Least Absolute Deviations (LAD)** procedure that minimizes the L_1 norm of the error term

$$||y - Xb||_1,$$

where $||u||_1 = \sum_{i=1}^n |u_i|$. This procedure goes back to Laplace in the 18th century but was put on a firm footing by Koenker and Bassett (1978). The resulting estimator denoted $\hat{\beta}_{LAD}$ is a regression analogue of the sample median and has certain desirable properties in the face of outliers in the covariates. We discuss this procedure in more detail below.

Consider the scalar regression $y_i = \beta x_i + \epsilon_i$. The OLS estimator is $\hat{\beta} = \dfrac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$. Also plausible are $\tilde{\beta} = \bar{y}/\bar{x}$ and $\breve{\beta} = \dfrac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}$. Another example that arises from accounting practice, is the so-called high-low method for determining average cost. The corresponding estimator is

$$\hat{\beta}_{H-L} = \frac{y_H - y_L}{x_H - x_L},$$

where x_H, x_L are the highest and lowest values achieved by the covariate respectively, and y_H, y_L are the "concomitants", that is, the corresponding values of the outcome variable. This estimator is also linear and unbiased under Assumption A.

A major class of estimators goes by the name of **Instrumental variables**. These are designed to outflank issues arising from endogeneity, which is of central importance in applied work. These are of the general form

$$\tilde{\beta} = (Z^T X)^{-1} Z^T y,$$

where Z is a full rank $n \times K$ matrix of instruments. We will discuss this class of estimators somewhat more in the final chapter, as they are included as a special case of the Generalized Method of Moments.

19.2 Optimality

There are many estimators of β. How do we choose between estimators? Computational convenience is an important issue, but the above estimators are all similar in their computational requirements. We now investigate statistical optimality.

Definition 19.2

The **mean squared error (MSE)** matrix of a generic estimator $\tilde{\theta}$ of a parameter $\theta \in \mathbb{R}^p$ is

$$
\begin{aligned}
M(\tilde{\theta}, \theta) &= \mathbb{E}\left[(\tilde{\theta} - \theta)(\tilde{\theta} - \theta)^T\right] \\
&= \mathbb{E}\left[(\tilde{\theta} - \mathbb{E}[\tilde{\theta}] + \mathbb{E}[\tilde{\theta}] - \theta)(\tilde{\theta} - \mathbb{E}[\tilde{\theta}] + \mathbb{E}[\tilde{\theta}] - \theta)^T\right] \\
&= \mathbb{E}\left[(\tilde{\theta} - \mathbb{E}[\tilde{\theta}])(\tilde{\theta} - \mathbb{E}[\tilde{\theta}])^T\right] + (\mathbb{E}[\tilde{\theta}] - \theta)(\mathbb{E}[\tilde{\theta}] - \theta)^T. \\
&\quad \text{variance} \qquad\qquad \text{squared bias}
\end{aligned}
$$

The expectation here can be conditional on X or unconditional. The MSE matrix is generally a function of the true parameter θ. We would like a method that does well for all θ, not just a subset of parameter values – the estimator $\hat{\theta} = 0$ is an example of a procedure that will have MSE equal to zero at $\theta = 0$, and hence will do well at this point, but as θ moves away, the MSE increases quadratically without limit. Note that no estimator can dominate uniformly across θ according to MSE because it would have to beat all constant estimators which have zero MSE at a single point. This is impossible unless there is no randomness. This is the same issue we discussed in the scalar case in an earlier chapter.

In the multivariate case, an additional issue arises even when comparing two estimators at a single point of the parameter space. MSE defines a complete ordering when $p = 1$, i.e., one can always rank any two estimators according to MSE. When $p > 1$, this is not so. In the general case we say that $\tilde{\theta}$ is better (according to MSE) than $\hat{\theta}$ (at some fixed point θ) if $B \geq A$ (i.e., $B - A$ is a positive semidefinite matrix), where B is the MSE matrix of $\tilde{\theta}$ and A is the MSE of $\hat{\theta}$. For example, suppose that

$$
A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 2 & 0 \\ 0 & 1/4 \end{bmatrix}.
$$

In this case, we can not rank the estimators. The problem is due to the multivariate nature of the optimality criterion. One solution is to take a scalar function of MSE such as the trace or determinant, which will result in a complete ordering. For example, for positive definite Q let

$$
\operatorname{tr}\left(\mathbb{E}\left[(\tilde{\theta} - \theta)^T Q (\tilde{\theta} - \theta)\right]\right) = \operatorname{tr}(MQ).
$$

For example, consider linear combinations of the parameter $c^T\theta$. In this case, the MSE of $c^T(\tilde{\theta} - \theta)$ is scalar and can be considered as the trace MSE with the particular $Q = cc^T$. However, different scalar functions will rank estimators differently, e.g., what is good for some c may be bad for some other c'.

Example 19.2

For example, the estimation of $m = X\beta$ by $\hat{m} = X\hat{\beta}$, yields

$$
\operatorname{tr}\left(\mathbb{E}\left[(\hat{m} - m)^T Q (\hat{m} - m)\right]\right) = \operatorname{tr}\left(X(X^TX)^{-1}X^T\Sigma(X)X(X^TX)^{-1}X^TQ\right).
$$

Under homoskedasticity and with $Q = I_n$, this equals $\sigma^2 K$. For other Q we obtain a different result.

The OLS estimator is **admissible** in the scalar case, meaning that no estimator uniformly dominates it according to mean squared error. However, in the multivariate case when the rank of Q is greater than or equal to 3 **Stein (shrinkage) estimators**, which are biased, improve on least squares according to MSE.

An alternative approach is to consider the minimax approach. For example, we might take the performance measure of the estimator $\tilde{\theta}$ to be

$$
R(\tilde{\theta}) = \max_{\theta \in \Theta} \operatorname{tr}(M(\tilde{\theta}, \theta)),
$$

which takes the most pessimistic view. In this case, we might try and find the estimator that minimizes this criterion – this would be called a minimax estimator. The theory for this class of estimators is very complicated, and in any case it is not such a desirable criterion because it is so pessimistic about nature trying to do the worst to us.

Instead, we might reduce the class of allowable estimators. If we restrict attention to unbiased estimators then this rules out estimators like $\tilde{\theta} = 0$ because they will be biased. In this case there is some hope of an optimality theory for the class of unbiased estimators. We will now return to the linear regression model and make the further restriction that the estimators we consider are linear in y. That is, we suppose that we have the set of all estimators $\tilde{\beta}$ that satisfy

$$\tilde{\beta} = Ay$$

for some fixed matrix A such that with probability one

$$\mathbb{E}[\tilde{\beta}|X] = \beta, \quad \forall \beta.$$

This latter condition implies that $(AX - I)\beta = 0$ for all β, which is equivalent to $AX = I$.

Theorem 19.6 (Gauss Markov)

Suppose that assumptions A1 and A2 hold and that $\Sigma(X) = \sigma^2 I_n$. Then the OLS estimator $\hat{\beta}$ is **Best Linear Unbiased (BLUE)**, i.e., with probability one

$$\text{var}(\hat{\beta}|X) \leq \text{var}(\tilde{\beta}|X)$$

for any other LUE.

Proof:

$$\text{var}(\hat{\beta}|X) = \sigma^2 (X^T X)^{-1}; \quad \text{var}(\tilde{\beta}|X) = \sigma^2 AA^T$$

$$\begin{aligned} \text{var}(\tilde{\beta}|X) - \text{var}(\hat{\beta}|X) &= \sigma^2 \left[AA^T - (X^T X)^{-1} \right] \\ &= \sigma^2 \left[AA^T - AX(X^T X)^{-1} X^T A^T \right] \\ &= \sigma^2 A \left[I - X(X^T X)^{-1} X^T \right] A^T \\ &= \sigma^2 A M_X A^T \\ &\geq 0. \end{aligned}$$

Remarks:

1. No assumption is made about the distribution of the errors; it only assumes 0 mean and $\sigma^2 I$ variance.
2. Result only compares *linear* estimators; it says nothing about for example $\sum_{i=1}^n |y_i - \beta x_i|$.
3. Result only compares *unbiased* estimators [biased estimators can have 0 variances].
4. There are extensions to consider affine estimators $\tilde{\beta} = a + Ay$ for vectors a. There are also equivalent results for the invariant quantity $\hat{y}$.
5. This says that for any linear combination $c^T \beta$, the OLS is BLUE when X is of full rank. When X is not of full rank, one can show that for an estimable linear combination $c^T \beta$, the OLS is BLUE.

If we dispense with the unbiasedness assumption and add the model assumption of error normality we get the well-known result.

Theorem 19.7 (Cramér-Rao)

Suppose that Assumption A4 holds. Then, $\hat{\beta}$ is **Best Unbiased**.

By making the stronger assumption A4, we get a much stronger conclusion. This allows us to compare say LAD estimation with OLS. The log-likelihood is given in equation (19.3), and it can be shown that the information matrix is

$$I(\beta, \sigma^2) = \begin{bmatrix} \frac{1}{\sigma^2} X^T X & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}.$$

This structure is block diagonal, which has some statistical implications. The inverse information is also block diagonal

$$I(\beta, \sigma^2)^{-1} = \begin{bmatrix} \sigma^2 (X^T X)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$

and represents the lower bound for the variance of unbiased estimators. The MLE of β is unbiased and achieves the bound, whereas the MLE of σ^2 is biased and does not achieve the bound in finite samples. Suppose that the parameter β were known, the information with regard to the unknown parameter σ^2 is $n/2\sigma^4$. On the other hand suppose that the parameter σ^2 is known, the information matrix with regard to the unknown parameter vector β is $X^T X/\sigma^2$. This says that knowledge of the other parameter provides no additional information about the unknown parameter.

Exercises

Exercise 1

Solution 1

Show that the OLS estimator $\hat{\beta}$ is conditionally unbiased under Assumption A1, without using the matrix notation.

Exercise 2

Solution 2

Explain intuitively what it means for an estimator to be “linear” in the context of OLS. Give an example of a non-linear estimator.

Exercise 3

Solution 3

Derive the conditional variance-covariance matrix of the OLS estimator, $\mathrm{var}(\hat{\beta}|X)$, under the assumption of homoskedasticity, starting from the general formula.

Exercise 4

Solution 4

Explain the difference between conditional and unconditional unbiasedness. Does conditional unbiasedness imply unconditional unbiasedness?

Exercise 5

Solution 5

What is the role of the Law of Iterated Expectations in the context of proving the unbiasedness of the OLS estimator?

Exercise 6

Solution 6

Under what conditions does the variance of the OLS estimator, $\text{var}(\hat{\beta})$, simplify to $\sigma^2 \mathbb{E}[(X^T X)^{-1}]$? Explain.

Exercise 7

Solution 7

Consider a simple linear regression model: $y_i = \beta x_i + \epsilon_i$, where $\epsilon_i \sim i.i.d. N(0, \sigma^2)$. If you have a sample of only two observations ($n = 2$), is the OLS estimator for β still unbiased? Explain.

Exercise 8

Solution 8

Given the partitioned regression formula $\hat{\beta}_1 = (X_1^T M_2 X_1)^{-1} X_1^T M_2 y$, explain the meaning of M_2 and its role in the formula.

Exercise 9

Solution 9

Explain in intuitive terms why, under the assumption of normality of errors, the OLS estimator $\hat{\beta}$ also follows a normal distribution.

Exercise 10

Solution 10(#sec-ch19solution10}

What is the significance of the result $\tau = (\hat{\beta} - \beta)^T V(X)^{-1} (\hat{\beta} - \beta) \sim \chi^2(K)$? How is it used in practice?

Exercise 11

Solution 11(#sec-ch19solution11}

Explain the difference between $\hat{m}$ and $\hat{m}(x)$ in the context of Theorem 19.4.

Exercise 12

Solution 12(#sec-ch19solution12}

Derive the expression for the Maximum Likelihood Estimator (MLE) of σ^2 in the linear regression model under the assumption of normality.

Exercise 13

Solution 13(#sec-ch19solution13}

Why is $s_*^2 = \dfrac{\hat{\epsilon}^T\hat{\epsilon}}{n-K}$ an unbiased estimator of σ^2, while $\hat{\sigma}^2_{mle} = \dfrac{\hat{\epsilon}^T\hat{\epsilon}}{n}$ is not?

Exercise 14

Solution 14(#sec-ch19solution14}

Explain the concept of "degrees of freedom" in the context of estimating the error variance σ^2.

Exercise 15

Solution 15(#sec-ch19solution15}

What does it mean for an estimator to be "Best Linear Unbiased Estimator" (BLUE)?

Exercise 16

Solution 16(#sec-ch19solution16}

Explain the Gauss-Markov Theorem in your own words, without using mathematical formulas.

Exercise 17

Solution 17(#sec-ch19solution17}

List the key assumptions required for the Gauss-Markov Theorem to hold.

Exercise 18

Solution 18(#sec-ch19solution18}

What is the Cramér-Rao Lower Bound, and what is its relevance to the OLS estimator?

Exercise 19

Solution 19(#sec-ch19solution19}

Why is the information matrix for (β, σ^2) block diagonal in the standard linear regression model with normal errors? What are the implications of this block diagonality?

Exercise 20

Solution 20(#sec-ch19solution20}

Consider the linear regression model $y = X\beta + \epsilon$ with heteroskedastic errors, such that $\text{Var}(\epsilon|X) = \Sigma(X)$, where $\Sigma(X)$ is a diagonal matrix with elements σ_i^2. Show that the variance of the OLS estimator is given by the formula provided in theorem 19.2.

Solutions

Solution 1

Exercise 1

The linear regression model can be expressed as:

$$y_i = x_i^T\beta + \epsilon_i, \text{ for } i = 1, \ldots, n.$$

where y_i is the dependent variable, x_i is a $K \times 1$ vector of independent variables, β is a $K \times 1$ vector of coefficients, and ϵ_i is the error term.

The OLS estimator for β is obtained by minimizing the sum of squared errors:

$$\sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})^2.$$

The solution to this minimization problem yields the OLS estimator:

$$\hat{\beta} = (\sum_{i=1}^{n} x_i x_i^T)^{-1}(\sum_{i=1}^{n} x_i y_i).$$

Substituting $y_i = x_i^T\beta + \epsilon_i$ into the OLS estimator formula:

$$\begin{aligned}
\hat{\beta} &= (\sum_{i=1}^{n} x_i x_i^T)^{-1}(\sum_{i=1}^{n} x_i(x_i^T\beta + \epsilon_i)) \\
&= (\sum_{i=1}^{n} x_i x_i^T)^{-1}(\sum_{i=1}^{n} x_i x_i^T\beta + \sum_{i=1}^{n} x_i\epsilon_i) \\
&= (\sum_{i=1}^{n} x_i x_i^T)^{-1}(\sum_{i=1}^{n} x_i x_i^T)\beta + (\sum_{i=1}^{n} x_i x_i^T)^{-1}(\sum_{i=1}^{n} x_i\epsilon_i) \\
&= \beta + (\sum_{i=1}^{n} x_i x_i^T)^{-1}(\sum_{i=1}^{n} x_i\epsilon_i).
\end{aligned}$$

Now, we take the conditional expectation given X:

$$\begin{aligned}
\mathbb{E}[\hat{\beta}|X] &= \mathbb{E}[\beta|X] + \mathbb{E}[(\sum_{i=1}^{n} x_i x_i^T)^{-1}(\sum_{i=1}^{n} x_i\epsilon_i)|X] \\
&= \beta + (\sum_{i=1}^{n} x_i x_i^T)^{-1}(\sum_{i=1}^{n} x_i\mathbb{E}[\epsilon_i|X]).
\end{aligned}$$

Assumption A1 states that $\mathbb{E}[\epsilon|X] = 0$. This means that $\mathbb{E}[\epsilon_i|X] = 0$ for all i. Thus,

$$\mathbb{E}[\hat{\beta}|X] = \beta + (\sum_{i=1}^{n} x_i x_i^T)^{-1}(\sum_{i=1}^{n} x_i \cdot 0) = \beta + 0 = \beta.$$

Therefore, the OLS estimator is conditionally unbiased.

Solution 2

Exercise 2

An estimator is **linear** if it can be expressed as a weighted sum of the observed dependent variable values (y_i). In the context of OLS, this means the estimator $\hat{\beta}$ can be written in the form $\hat{\beta} = Cy$, where C is a matrix that depends only on the independent variables (X) and not on the dependent variable (y). This linearity simplifies many calculations and is a key property used in deriving the statistical properties of the OLS estimator. The linearity arises from the normal equations and the fact that we are solving a linear system of equations to find the OLS estimator.

An example of a **non-linear estimator** would be an estimator that involves a non-linear transformation of the dependent variable y before estimating the parameters. For example, consider estimating the parameters a and b of a model: $y_i = ae^{bx_i}\epsilon_i$. After a log transform, the parameters a and b can be estimated using linear regression methods, but we have to resort to nonlinear methods if we do not transform the data.

Solution 3

Exercise 3

The general formula for the conditional variance-covariance matrix of the OLS estimator is:

$$\text{var}(\hat{\beta}|X) = (X^TX)^{-1}X^T\mathbb{E}[\epsilon\epsilon^T|X]X(X^TX)^{-1}.$$

Under the assumption of homoskedasticity, the conditional variance of each error term is constant, and the error terms are uncorrelated. This means $\mathbb{E}[\epsilon_i^2|X] = \sigma^2$ for all i, and $\mathbb{E}[\epsilon_i\epsilon_j|X] = 0$ for $i \neq j$. This can be summarized as:

$$\mathbb{E}[\epsilon\epsilon^T|X] = \sigma^2 I_n,$$

where I_n is the $n \times n$ identity matrix. Substituting this into the general formula:

$$\begin{aligned}
\text{var}(\hat{\beta}|X) &= (X^TX)^{-1}X^T(\sigma^2 I_n)X(X^TX)^{-1} \\
&= \sigma^2(X^TX)^{-1}X^TI_nX(X^TX)^{-1} \\
&= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} \\
&= \sigma^2(X^TX)^{-1}(X^TX)(X^TX)^{-1} \\
&= \sigma^2(X^TX)^{-1}.
\end{aligned}$$

Thus, under homoskedasticity, the conditional variance-covariance matrix simplifies to $\sigma^2(X^TX)^{-1}$.

Solution 4

Exercise 4

Conditional unbiasedness means that the expected value of the estimator, *given* the values of the independent variables X, is equal to the true parameter value. Formally, $\mathbb{E}[\hat{\beta}|X] = \beta$. This implies that for any specific set of X values, if we were to repeatedly sample and calculate the estimator, the average of those estimates would converge to the true value.

Unconditional unbiasedness means that the expected value of the estimator, averaging over *all possible* values of the independent variables X, is equal to the true parameter value. Formally, $\mathbb{E}[\hat{\beta}] = \beta$.

Yes, conditional unbiasedness *does* imply unconditional unbiasedness. This can be shown using the **Law of Iterated Expectations**:

$$\mathbb{E}[\hat{\beta}] = \mathbb{E}[\mathbb{E}[\hat{\beta}|X]].$$

If $\mathbb{E}[\hat{\beta}|X] = \beta$, then

$$\mathbb{E}[\hat{\beta}] = \mathbb{E}[\beta] = \beta.$$

The intuition here is that if the estimator is unbiased for every possible value of X, then it must also be unbiased on average across all possible values of X.

Solution 5

Exercise 5

The Law of Iterated Expectations (LIE) is crucial for demonstrating that conditional unbiasedness implies unconditional unbiasedness. The LIE states that:

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]].$$

In the context of the OLS estimator, we first show that the estimator is conditionally unbiased: $\mathbb{E}[\hat{\beta}|X] = \beta$. Then, to show unconditional unbiasedness, we apply the LIE:

$$\mathbb{E}[\hat{\beta}] = \mathbb{E}[\mathbb{E}[\hat{\beta}|X]].$$

Since we know $\mathbb{E}[\hat{\beta}|X] = \beta$, we substitute this in:

$$\mathbb{E}[\hat{\beta}] = \mathbb{E}[\beta].$$

Since β is a constant (the true parameter value), its expectation is simply itself:

$$\mathbb{E}[\hat{\beta}] = \beta.$$

Thus, the LIE allows us to move from the conditional statement (unbiased for a given X) to the unconditional statement (unbiased overall).

Solution 6

Exercise 6

The variance of the OLS estimator is generally given by:

$$\text{var}(\hat{\beta}) = \mathbb{E}[\text{var}(\hat{\beta}|X)] = \mathbb{E}\left[(X^TX)^{-1}X^T\Sigma(X)X(X^TX)^{-1}\right].$$

This simplifies to $\sigma^2\mathbb{E}[(X^TX)^{-1}]$ under the following conditions:

1. **Homoskedasticity:** The variance of the error term is constant for all observations: $\text{Var}(\epsilon_i|X) = \sigma^2$ for all i.
2. **No Autocorrelation:** The error terms are uncorrelated with each other: $\text{Cov}(\epsilon_i, \epsilon_j|X) = 0$ for all $i \neq j$.

These two conditions together imply that $\Sigma(X) = \mathbb{E}[\epsilon\epsilon^T|X] = \sigma^2 I_n$, where I_n is the $n \times n$ identity matrix. Substituting this into the general formula:

$$\begin{aligned}\text{var}(\hat{\beta}) &= \mathbb{E}\left[(X^TX)^{-1}X^T(\sigma^2 I_n)X(X^TX)^{-1}\right] \\ &= \mathbb{E}\left[\sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1}\right] \\ &= \sigma^2\mathbb{E}\left[(X^TX)^{-1}\right].\end{aligned}$$

Therefore, homoskedasticity and no autocorrelation are required for the simplification.

Solution 7

Exercise 7

Yes, the OLS estimator for β is still unbiased, even with only two observations. The unbiasedness property of the OLS estimator relies on Assumption A1 ($\mathbb{E}[\epsilon|X] = 0$), which does not depend on the sample size, n.

Let's derive the OLS estimator for this specific case:

$$y_1 = \beta x_1 + \epsilon_1 \qquad y_2 = \beta x_2 + \epsilon_2$$

In matrix form:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\beta + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}$$

$y = X\beta + \epsilon$

The OLS estimator is given by $\hat{\beta} = (X^TX)^{-1}X^Ty$. In this case:

$$X^TX = \begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_1^2 + x_2^2 \qquad X^Ty = \begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = x_1y_1 + x_2y_2$$

So, $\hat{\beta} = \dfrac{x_1y_1 + x_2y_2}{x_1^2 + x_2^2}$.

Now, let's take the conditional expectation:

$$\begin{aligned}\mathbb{E}[\hat{\beta}|X] &= \mathbb{E}\left[\frac{x_1y_1 + x_2y_2}{x_1^2 + x_2^2}|x_1, x_2\right] \\ &= \frac{1}{x_1^2 + x_2^2}\mathbb{E}\left[x_1y_1 + x_2y_2|x_1, x_2\right] \\ &= \frac{1}{x_1^2 + x_2^2}(x_1\mathbb{E}[y_1|x_1] + x_2\mathbb{E}[y_2|x_2])\end{aligned}$$

Substitute $y_i = \beta x_i + \epsilon_i$:

$$\begin{aligned}\mathbb{E}[\hat{\beta}|X] &= \frac{1}{x_1^2 + x_2^2}(x_1\mathbb{E}[\beta x_1 + \epsilon_1|x_1] + x_2\mathbb{E}[\beta x_2 + \epsilon_2|x_2]) \\ &= \frac{1}{x_1^2 + x_2^2}(x_1(\beta x_1 + \mathbb{E}[\epsilon_1|x_1]) + x_2(\beta x_2 + \mathbb{E}[\epsilon_2|x_2]))\end{aligned}$$

Using Assumption A1, $\mathbb{E}[\epsilon_i|x_i] = 0$:

$$\begin{aligned}\mathbb{E}[\hat{\beta}|X] &= \frac{1}{x_1^2 + x_2^2}(x_1^2\beta + x_2^2\beta) \\ &= \frac{\beta(x_1^2 + x_2^2)}{x_1^2 + x_2^2} \\ &= \beta\end{aligned}$$

Thus, $\hat{\beta}$ is unbiased even with $n = 2$. However, while it's unbiased, the *variance* of the estimator will be very large with only two data points.

Solution 8

Exercise 8

The matrix M_2 is a **residual maker matrix** or **annihilator matrix** associated with the variables in X_2. It is defined as:

$$M_2 = I - X_2(X_2^T X_2)^{-1} X_2^T.$$

Here's its role and meaning:

1. **Projection:** $X_2(X_2^T X_2)^{-1} X_2^T$ is the projection matrix onto the column space of X_2. When this matrix multiplies a vector, it projects that vector onto the space spanned by the columns of X_2.
2. **Residuals:** M_2 does the opposite. When M_2 multiplies a vector, it produces the *residuals* of a regression of that vector on X_2. That is, $M_2 y$ gives the residuals of a regression of y on X_2.
3. **Orthogonality:** M_2 is symmetric ($M_2 = M_2^T$) and idempotent ($M_2 M_2 = M_2$). Furthermore, $M_2 X_2 = 0$. This means that M_2 projects any vector onto the space *orthogonal* to the column space of X_2.

In the partitioned regression formula, M_2 is used to remove the effect of X_2 from both y and X_1. Specifically:

- $M_2 y$: Represents the part of y that is *not* explained by X_2.
- $M_2 X_1$: Represents the part of X_1 that is *not* explained by X_2.

The formula then essentially regresses the unexplained part of y on the unexplained part of X_1. This isolates the effect of X_1 on y, after controlling for the effects of X_2. This is a key concept in multiple regression, showing how we can isolate the effect of one variable while controlling for the effects of others.

Solution 9

Exercise 9

The OLS estimator is a linear combination of the dependent variable y:

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

Under the assumption of normality of errors, we have:

$$\epsilon \sim N(0, \sigma^2 I_n).$$

Since $y = X\beta + \epsilon$, and $X\beta$ is a constant (conditional on X), y is also normally distributed:

$$y|X \sim N(X\beta, \sigma^2 I_n).$$

Now, $\hat{\beta}$ is a linear combination of y, and a linear combination of normally distributed variables is also normally distributed. Specifically, if $Y \sim N(\mu, \Sigma)$ and A is a constant matrix, then $AY \sim N(A\mu, A\Sigma A^T)$. Applying this to the OLS estimator:

$$\begin{aligned}\hat{\beta}|X &= (X^TX)^{-1}X^Ty \\ &\sim N((X^TX)^{-1}X^TX\beta, (X^TX)^{-1}X^T(\sigma^2 I_n)((X^TX)^{-1}X^T)^T) \\ &\sim N(\beta, \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-T}) \\ &\sim N(\beta, \sigma^2(X^TX)^{-1}).\end{aligned}$$

Therefore, $\hat{\beta}$ follows a normal distribution because it is a linear transformation of y, which is itself normally distributed due to the normality of the error terms.

Solution 10

Exercise 10

The result $\tau = (\hat{\beta} - \beta)^T V(X)^{-1}(\hat{\beta} - \beta) \sim \chi^2(K)$ is significant because it provides a statistic that follows a known distribution (the chi-squared distribution with K degrees of freedom) *and* does not depend on any unknown parameters (under the assumptions of the model). This allows us to perform hypothesis tests and construct confidence intervals for β. Here's how it is used:

1. **Hypothesis Testing:** We can test hypotheses about the coefficients β. For example, to test the null hypothesis $H_0 : R\beta = r$, where R is a $q \times K$ matrix and r is a $q \times 1$ vector, we can use the fact that, under H_0:

   $$(R\hat{\beta} - r)^T(RV(X)R^T)^{-1}(R\hat{\beta} - r) \sim \chi^2(q).$$

   We can then compare the calculated value of this statistic to the critical value from the chi-squared distribution with q degrees of freedom to determine whether to reject the null hypothesis. This generalizes the individual t-tests to allow for testing multiple linear restrictions on the coefficients simultaneously.

2. **Confidence Regions:** While we usually construct confidence *intervals* for individual coefficients, the chi-squared result allows us to construct confidence *regions* (ellipsoids) for the entire vector β. A $(1-\alpha)$ confidence region for β is the set of all β values that satisfy:

   $$(\hat{\beta} - \beta)^T V(X)^{-1}(\hat{\beta} - \beta) \leq \chi^2_{K, 1-\alpha},$$

   where $\chi^2_{K,1-\alpha}$ is the $(1-\alpha)$ quantile of the chi-squared distribution with K degrees of freedom.

In summary, this result provides a foundation for inference in the linear regression model, moving beyond simple t-tests for individual coefficients.

Solution 11

Exercise 11

In Theorem 19.4, $\hat{m}$ and $\hat{m}(x)$ represent different, but related, quantities:

- $\hat{m}$**:** This is an estimator of the *conditional mean vector* $m = \mathbb{E}[y|X] = X\beta$. $\hat{m}$ is an $n \times 1$ vector, where each element $\hat{m}_i$ is an estimate of the expected value of y_i *given the corresponding row of the design matrix*, x_i^T. That is, $\hat{m}_i = x_i^T\hat{\beta}$. So, $\hat{m} = X\hat{\beta}$.

- $\hat{m}(x)$**:** This is an estimator of the *conditional mean function* evaluated at a *specific* point x. x is a $K \times 1$ vector of values for the independent variables. $\hat{m}(x)$ is a *scalar*, representing the estimated expected value of y when the independent variables take on the values in the vector x. That is, $\hat{m}(x) = x^T\hat{\beta}$.

In simpler terms:

- $\hat{m}$ gives you the predicted values for y for each observation in your *existing* dataset (X).
- $\hat{m}(x)$ gives you the predicted value for y for a *new* set of independent variable values, x (which might not be in your original dataset).

The theorem provides the properties (unbiasedness, variance, and distribution) of both of these estimators. They are related because you can think of $\hat{m}$ as being constructed by stacking up $\hat{m}(x_i)$ for each row x_i^T of the design matrix X.

Solution 12

Exercise 12

Under the assumption of normality, the conditional distribution of y given X is:

$$f_{y|X}(y) = \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right).$$

The log-likelihood function is the logarithm of this density:

$$\begin{aligned} l(\beta, \sigma^2|y, X) &= \log\left(\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right)\right) \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta). \end{aligned}$$

To find the MLE, we maximize this log-likelihood with respect to β and σ^2. We already know that the MLE for β is the OLS estimator, $\hat{\beta} = (X^TX)^{-1}X^Ty$. Now, we need to find the MLE for σ^2. We take the derivative of the log-likelihood with respect to σ^2 and set it equal to zero:

$$\begin{aligned} \frac{\partial l}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(y - X\beta)^T(y - X\beta) = 0 \\ \frac{n}{2\sigma^2} &= \frac{1}{2(\sigma^2)^2}(y - X\beta)^T(y - X\beta) \\ n &= \frac{1}{\sigma^2}(y - X\beta)^T(y - X\beta) \\ \hat{\sigma}^2_{mle} &= \frac{1}{n}(y - X\hat{\beta})^T(y - X\hat{\beta}). \end{aligned}$$

Substituting $\hat{\beta}$ for β (since we're finding the joint MLE), we get:

$$\hat{\sigma}^2_{mle} = \frac{1}{n}(y - X\hat{\beta})^T(y - X\hat{\beta}) = \frac{\hat{\epsilon}^T\hat{\epsilon}}{n}.$$

This is the MLE of σ^2. It is the sum of squared residuals divided by the number of observations, n.

Solution 13

Exercise 13

The key difference lies in the **degrees of freedom**.

- $\hat{\sigma}^2_{mle} = \dfrac{\hat{\epsilon}^T\hat{\epsilon}}{n}$ divides the sum of squared residuals by n, the number of observations. This estimator is the Maximum Likelihood Estimator (MLE) and is biased downwards. It tends to underestimate the true variance σ^2 because it doesn't account for the fact that we've already used up some degrees of freedom in estimating the β coefficients.
- $s^2_* = \dfrac{\hat{\epsilon}^T\hat{\epsilon}}{n-K}$ divides the sum of squared residuals by $n-K$, where K is the number of parameters in the β vector (including the intercept). This is the unbiased estimator.

Here's why the $n-K$ correction makes s^2_* unbiased:

1. **Degrees of Freedom Lost:** When we estimate β using OLS, we are essentially imposing K constraints on our data (the normal equations). This means we lose K degrees of freedom.

2. **Expected Value of SSR:** The expected value of the sum of squared residuals ($\hat{\epsilon}^T\hat{\epsilon}$) is equal to $(n-K)\sigma^2$. This is a standard result in linear regression (and is related to the proof of Theorem 19.5).

3. **Unbiasedness:** To obtain an unbiased estimator of σ^2, we need to divide the sum of squared residuals by a quantity that, when we take the expectation, will give us σ^2. Since $\mathbb{E}[\hat{\epsilon}^T\hat{\epsilon}] = (n-K)\sigma^2$, dividing by $(n-K)$ achieves this:

$$\mathbb{E}[s^2_*] = \mathbb{E}\left[\frac{\hat{\epsilon}^T\hat{\epsilon}}{n-K}\right] = \frac{\mathbb{E}[\hat{\epsilon}^T\hat{\epsilon}]}{n-K} = \frac{(n-K)\sigma^2}{n-K} = \sigma^2.$$

In summary, s^2_* corrects for the degrees of freedom lost in estimating the regression coefficients, leading to an unbiased estimate of the error variance. The MLE, $\hat{\sigma}^2_{mle}$, does not make this correction and is therefore biased.

Solution 14

Exercise 14

Degrees of freedom in the context of estimating the error variance σ^2 refer to the number of independent pieces of information available to estimate the variance *after* accounting for the parameters that have already been estimated from the data.

Think of it this way:

1. **Initial Information:** You start with n observations, which represent n independent pieces of information.
2. **Estimating β:** When you estimate the K parameters in the β vector (including the intercept) using OLS, you are essentially imposing K constraints on your data. These constraints are derived from the normal equations, which ensure that the residuals are orthogonal to the predictors.
3. **Remaining Information:** After estimating β, you are left with $n-K$ independent pieces of information to estimate the variance. This is because K degrees of freedom have been "used up" in estimating β.

Therefore, the degrees of freedom for estimating σ^2 are $n-K$. This is why the unbiased estimator of σ^2, s^2_*, divides the sum of squared residuals by $n-K$, not n. Dividing by $n-K$ correctly accounts for the number of independent pieces of information remaining to estimate the variance.

Solution 15

Exercise 15

An estimator is considered the **Best Linear Unbiased Estimator (BLUE)** if it satisfies the following three properties:

1. **Linear:** The estimator is a linear function of the observed dependent variable values (y). This means it can be expressed in the form $\tilde{\beta} = Ay$, where A is a matrix that does not depend on y (though it can depend on X).
2. **Unbiased:** The expected value of the estimator is equal to the true parameter value. That is, $\mathbb{E}[\tilde{\beta}] = \beta$ (unconditionally) or $\mathbb{E}[\tilde{\beta}|X] = \beta$ (conditionally).
3. **Best (Minimum Variance):** Among all *linear unbiased* estimators, the estimator has the smallest variance. This means that for any other linear unbiased estimator $\tilde{\beta}$, $\mathrm{Var}(\hat{\beta}) \leq \mathrm{var}(\tilde{\beta})$ in the matrix sense (i.e., $\mathrm{var}(\tilde{\beta}) - \mathrm{var}(\hat{\beta})$ is positive semi-definite). Equivalently, for any linear combination of the parameters, $c^T\beta$, the variance of $c^T\hat{\beta}$ is less than or equal to the variance of $c^T\tilde{\beta}$.

In summary, BLUE means the estimator is a weighted average of the y values, it's correct on average, and it's the most precise (smallest variance) among all such estimators. The Gauss-Markov theorem states that the OLS estimator is BLUE under certain assumptions.

Solution 16

Exercise 16

The **Gauss-Markov Theorem** is a fundamental result in linear regression. It states that, under certain assumptions about the error terms in the linear regression model, the Ordinary Least Squares (OLS) estimator is the "best" we can do among all linear and unbiased estimators.

"Best" in this context means *minimum variance*. Imagine you have many different ways to estimate the coefficients of your regression model, but you restrict yourself to methods that are both linear (the estimate is a weighted average of the y values) and unbiased (on average, the estimate equals the true value). The Gauss-Markov theorem tells you that, out of all these possible linear and unbiased estimation methods, the OLS estimator will give you estimates that are, on average, closest to the true values. The spread (variance) of the OLS estimates around the true values will be the smallest possible.

It's important to remember that this "best" property only holds *if* the assumptions of the theorem are met. If those assumptions are violated (e.g., if the errors are heteroskedastic or autocorrelated), then OLS might not be the best estimator anymore, and other estimators (like Generalized Least Squares) might be better.

Solution 17

Exercise 17

The key assumptions required for the Gauss-Markov Theorem to hold are related to the error term (ϵ) in the linear regression model $y = X\beta + \epsilon$:

1. **Linearity:** The model is linear in parameters. The dependent variable is a linear function of the independent variables and the error term.
2. **Strict Exogeneity:** The expected value of the error term, conditional on the independent variables, is zero: $\mathbb{E}[\epsilon|X] = 0$. This means that the independent variables are not correlated with the error term.
3. **Homoskedasticity:** The variance of the error term is constant for all observations: $\mathrm{Var}(\epsilon_i|X) = \sigma^2$ for all i. This means the error term has the same variance for all values of the independent variables.

4. **No Autocorrelation:** The error terms are uncorrelated with each other: $\text{Cov}(\epsilon_i, \epsilon_j|X) = 0$ for all $i \neq j$. This means there is no systematic relationship between the error terms of different observations.
5. **Full Rank of X:** The matrix X has full column rank. This means that there is no perfect multicollinearity among the independent variables (none of the independent variables can be written as a perfect linear combination of the others).

It's important to note that the Gauss-Markov theorem *does not* require the error term to be normally distributed. Normality is needed for certain hypothesis tests and confidence intervals, but not for the BLUE property of OLS. If any of the assumptions 2, 3 or 4 are not met, the Gauss-Markov theorem no longer holds.

Solution 18

Exercise 18

The **Cramér-Rao Lower Bound (CRLB)** provides a theoretical lower bound on the variance of *any* unbiased estimator. It states that the variance of any unbiased estimator $\tilde{\theta}$ of a parameter θ is always greater than or equal to the inverse of the Fisher information, $I(\theta)$:

$$\text{Var}(\tilde{\theta}) \geq I(\theta)^{-1}.$$

The Fisher information, $I(\theta)$, measures the amount of information that the data provides about the unknown parameter θ. It is defined as the variance of the score function (the derivative of the log-likelihood function with respect to θ):

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial l(\theta|y)}{\partial \theta}\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2 l(\theta|y)}{\partial \theta^2}\right],$$

where $l(\theta|y)$ is the log-likelihood function.

The relevance to the OLS estimator is as follows:

1. **Best Unbiased:** Under the assumption of normally distributed errors (Assumption A4), the OLS estimator achieves the Cramér-Rao Lower Bound. This means that the OLS estimator is not just the best *linear* unbiased estimator (as stated by the Gauss-Markov theorem); it is the best unbiased estimator *among all unbiased estimators*, including non-linear ones.
2. **Efficiency:** An estimator that achieves the CRLB is said to be *efficient*. So, under normality, the OLS estimator is efficient.
3. **Comparison:** The CRLB provides a benchmark against which to compare the performance of other estimators. If an estimator's variance is close to the CRLB, it's doing well.

It's crucial to remember that the CRLB applies to *unbiased* estimators. There may be biased estimators with lower Mean Squared Error (MSE) than the CRLB (since MSE considers both bias and variance).

Solution 19

Exercise 19

The information matrix for (β, σ^2) in the standard linear regression model with normal errors is block diagonal:

$$I(\beta, \sigma^2) = \begin{bmatrix} \frac{1}{\sigma^2} X^T X & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}.$$

This block diagonality arises from the structure of the log-likelihood function under the assumption of normality and homoskedasticity. Recall the log-likelihood:

$$l(\beta, \sigma^2|y, X) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta).$$

The key is to examine the second derivatives (Hessian matrix):

1. $\frac{\partial^2 l}{\partial\beta\partial\beta^T}$**:** The second derivative with respect to β only involves the last term of the log-likelihood. After taking the derivative twice and taking expectations, we get $\frac{1}{\sigma^2}X^TX$.
2. $\frac{\partial^2 l}{\partial(\sigma^2)^2}$**:** The second derivative with respect to σ^2 involves the second and third terms. After taking derivatives twice, and taking expectations (using the fact that $\mathbb{E}[(y - X\beta)^T(y - X\beta)] = n\sigma^2$), we get $\frac{n}{2\sigma^4}$.
3. $\frac{\partial^2 l}{\partial\beta\partial\sigma^2}$**:** This is the crucial part. Taking the derivative first with respect to β gives: $\frac{1}{\sigma^2}X^T(y - X\beta)$. Then taking the derivative with respect to σ^2 gives: $-\frac{1}{\sigma^4}X^T(y - X\beta)$. Taking the expectation, and using the fact that $\mathbb{E}[y - X\beta|X] = 0$, this cross-partial derivative is *zero*.

The fact that the cross-partial derivative is zero is what leads to the block diagonality. This means that the Fisher Information matrix is block diagonal.

Implications of Block Diagonality:

1. **Independence of MLEs (Asymptotically):** The block diagonality implies that the Maximum Likelihood Estimators (MLEs) of β and σ^2 are asymptotically independent. This means that knowing the true value of σ^2 would not help in estimating β, and vice-versa (at least in large samples).
2. **Separate Estimation:** We can estimate β and σ^2 separately without loss of information. The estimation of one parameter doesn't affect the efficiency of the estimation of the other.
3. **Variance Calculation:** The inverse of the block diagonal information matrix is also block diagonal, making the calculation of the variances of the MLEs straightforward. The variance of $\hat{\beta}$ depends only on X and σ^2, and the variance of $\hat{\sigma}^2$ depends only on n and σ^2.

Solution 20

Exercise 20

The OLS estimator is given by: $\hat{\beta} = (X^TX)^{-1}X^Ty$. We want to find the variance of $\hat{\beta}$ conditional on X, given that $\text{Var}(\epsilon|X) = \Sigma(X)$, where $\Sigma(X)$ is a diagonal matrix with elements σ_i^2.

First, substitute the model equation into the estimator:

$$\begin{aligned}\hat{\beta} &= (X^TX)^{-1}X^T(X\beta + \epsilon) \\ &= (X^TX)^{-1}X^TX\beta + (X^TX)^{-1}X^T\epsilon \\ &= \beta + (X^TX)^{-1}X^T\epsilon.\end{aligned}$$

Now, subtract β from both sides:

$$\hat{\beta} - \beta = (X^TX)^{-1}X^T\epsilon.$$

To find the variance, we use the definition of the variance-covariance matrix:

$$\text{Var}(\hat{\beta}|X) = \mathbb{E}[(\hat{\beta} - \mathbb{E}[\hat{\beta}|X])(\hat{\beta} - \mathbb{E}[\hat{\beta}|X])^T|X].$$

Since $\mathbb{E}[\hat{\beta}|X] = \beta$:

$$\begin{aligned}\text{Var}(\hat{\beta}|X) &= \mathbb{E}[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T|X] \\ &= \mathbb{E}[((X^TX)^{-1}X^T\epsilon)((X^TX)^{-1}X^T\epsilon)^T|X] \\ &= \mathbb{E}[(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-1}|X] \\ &= (X^TX)^{-1}X^T\mathbb{E}[\epsilon\epsilon^T|X]X(X^TX)^{-1}.\end{aligned}$$

We are given that $\text{Var}(\epsilon|X) = \Sigma(X) = \mathbb{E}[\epsilon\epsilon^T|X]$. Therefore:

$$\text{Var}(\hat{\beta}|X) = (X^TX)^{-1}X^T\Sigma(X)X(X^TX)^{-1}.$$

This is the formula provided in Theorem 19.2, showing that the variance of the OLS estimator under heteroskedasticity depends on the specific form of the heteroskedasticity, $\Sigma(X)$. This demonstrates that the derivation in Theorem 19.2 holds generally for any covariance structure of the errors, not just homoskedastic errors.

R Scripts

R Script 1: Demonstrating Unbiasedness of OLS

```
# Load necessary libraries
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✓ dplyr     1.1.4     ✓ readr     2.1.5
✓ forcats   1.0.0     ✓ stringr   1.5.1
✓ ggplot2   3.5.1     ✓ tibble    3.2.1
✓ lubridate 1.9.4     ✓ tidyr     1.3.1
✓ purrr     1.0.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
# Set the seed for reproducibility
set.seed(123)

# Simulation parameters
n_simulations <- 1000
sample_size <- 100
beta_true <- c(2, 0.5) # True intercept and slope
sigma_true <- 2 # True error standard deviation

# Create an empty data frame to store the results
results <- data.frame()

# Run simulations
for (i in 1:n_simulations) {
  # Generate independent variable x
  x <- runif(sample_size, min = -5, max = 5)
  X <- cbind(1, x) # Design matrix with intercept
```

```
  # Generate errors from a normal distribution
  epsilon <- rnorm(sample_size, mean = 0, sd = sigma_true)

  # Generate dependent variable y
  y <- X %*% beta_true + epsilon

  # Calculate OLS estimates
  beta_hat <- solve(t(X) %*% X) %*% t(X) %*% y

  # Store the results.  Use `rbind` with a named list for
  # easier column management.
  results <- rbind(results,
                   data.frame(simulation = i,
                              beta0_hat = beta_hat[1, 1],
                              beta1_hat = beta_hat[2, 1]))
}

# Calculate the average of the estimated coefficients
average_beta_hat <- results %>%
  summarize(avg_beta0_hat = mean(beta0_hat),
            avg_beta1_hat = mean(beta1_hat))

# Print the results
print(average_beta_hat)

  avg_beta0_hat avg_beta1_hat
1      2.006845     0.4987579

# Visualize the distribution of the estimated coefficients
results %>%
  pivot_longer(cols = c(beta0_hat, beta1_hat),
               names_to = "parameter",
               values_to = "estimate") %>%
  ggplot(aes(x = estimate)) +
  geom_histogram(bins = 30) +
  facet_wrap(~parameter) +
  geom_vline(data = data.frame(parameter = c("beta0_hat", "beta1_hat"),
                               true_value = beta_true),
             aes(xintercept = true_value), color = "red") +
  labs(title = "Distribution of OLS Estimates",
       x = "Estimated Value",
       y = "Frequency")
```

Explanation:

1. **Setup:**
   - We load the `tidyverse` library for data manipulation and visualization.
   - `set.seed(123)` ensures reproducibility of the random number generation.
   - We define simulation parameters: `n_simulations` (number of times we repeat the experiment), `sample_size` (number of observations in each sample), `beta_true` (the true values of the intercept and slope), and `sigma_true` (the true standard deviation of the error term).
2. **Simulation Loop:**
   - We loop `n_simulations` times, generating a new dataset and calculating OLS estimates in each iteration.
   - `x <- runif(sample_size, min = -5, max = 5)`: We generate the independent variable x from a uniform distribution.

- `X <- cbind(1, x)`: We create the design matrix X, including a column of ones for the intercept.
- `epsilon <- rnorm(sample_size, mean = 0, sd = sigma_true)`: We generate the error terms ϵ from a normal distribution with mean 0 and standard deviation `sigma_true`. This satisfies Assumption A3 (normality) and, consequently, Assumptions A1 and A2.
- `y <- X %*% beta_true + epsilon`: We generate the dependent variable y according to the linear model $y = X\beta + \epsilon$.
- `beta_hat <- solve(t(X) %*% X) %*% t(X) %*% y`: This is the core OLS calculation. It implements the formula $\hat{\beta} = (X^T X)^{-1} X^T y$. `solve(t(X) %*% X)` calculates $(X^T X)^{-1}$, and then we multiply by $X^T y$.
- The estimated intercept (`beta_hat[1, 1]`) and slope (`beta_hat[2, 1]`) are stored in the `results` data frame.

3. **Analysis:**
   - `average_beta_hat <- ...`: After the loop, we calculate the average of the estimated coefficients (`beta0_hat` and `beta1_hat`) across all simulations.
   - `print(average_beta_hat)`: We print the average estimated coefficients. These should be close to the true values (2 and 0.5).
4. **Visualization:**
   - The code creates a histogram of the estimated coefficients for both β_0 and β_1.
   - `geom_vline(...)`: Red vertical lines are added to the histograms, indicating the true values of β_0 and β_1.

Connection to Concepts:

- **Unbiasedness (Theorem 19.1):** The script demonstrates the unbiasedness of the OLS estimator. The average of the estimated coefficients across many simulations converges to the true parameter values. This illustrates that $\mathbb{E}[\hat{\beta}] = \beta$.
- **Linear Estimator (Definition 19.1):** The line `beta_hat <- solve(t(X) %*% X) %*% t(X) %*% y` shows that the OLS estimator is a linear function of y (it can be written as Cy, where $C = (X^T X)^{-1} X^T$).

R Script 2: Conditional Variance of OLS

```
# Set seed for reproducibility
set.seed(456)

# Simulation parameters
n <- 100  # Sample size
beta_true <- c(1, 2)  # True intercept and slope
sigma_true <- 3 # True error standard deviation

# Generate a fixed X (for conditional variance)
x <- runif(n, min = -2, max = 2)
X <- cbind(1, x)

# Number of simulations for estimating the *conditional* variance
n_sim <- 1000
beta_hats <- matrix(NA, nrow = n_sim, ncol = 2)

for (i in 1:n_sim) {
    # Generate errors (note: X is fixed, only epsilon changes)
    epsilon <- rnorm(n, mean = 0, sd = sigma_true)

    # Generate y
    y <- X %*% beta_true + epsilon

    # Calculate OLS estimates
    beta_hats[i, ] <- solve(t(X) %*% X) %*% t(X) %*% y
```

```
}

# Calculate the *empirical* conditional variance-covariance matrix
empirical_conditional_var <- var(beta_hats)

# Calculate the *theoretical* conditional variance-covariance matrix
theoretical_conditional_var <- sigma_true^2 * solve(t(X) %*% X)

# Print both
print("Empirical Conditional Variance-Covariance Matrix:")
```

```
[1] "Empirical Conditional Variance-Covariance Matrix:"
```

```
print(empirical_conditional_var)
```

```
             [,1]         [,2]
[1,]  0.091842070 -0.008237466
[2,] -0.008237466  0.074014967
```

```
print("Theoretical Conditional Variance-Covariance Matrix:")
```

```
[1] "Theoretical Conditional Variance-Covariance Matrix:"
```

```
print(theoretical_conditional_var)
```

```
                        x
   0.09159406 -0.01062313
x -0.01062313  0.07079451
```

```
# Visualize the sampling distribution
beta_hats_df <- as.data.frame(beta_hats)
colnames(beta_hats_df) <- c("beta0_hat", "beta1_hat")

beta_hats_df %>%
  ggplot(aes(x = beta0_hat, y = beta1_hat)) +
  geom_point(alpha = 0.2) +
    geom_vline(xintercept=beta_true[1], color="red") +
    geom_hline(yintercept=beta_true[2], color="red") +
  labs(title = "Sampling Distribution of OLS Estimator (Conditional on X)",
       x = "beta0_hat",
       y = "beta1_hat")
```

Explanation:

1. **Setup:**
   - We set the seed for reproducibility.
   - We define `n` (sample size), `beta_true`, and `sigma_true`.
   - Crucially, we generate `x` *once* and keep it fixed throughout the simulations. This is because we are interested in the *conditional* variance, given a specific X.
2. **Simulation Loop:**
   - We loop `n_sim` times. In each iteration:
     - We generate new error terms `epsilon` from a normal distribution.
     - We generate `y` using the *fixed* `X` and the new `epsilon`.
     - We calculate the OLS estimates `beta_hat` and store them.
3. **Variance Calculation:**
   - `empirical_conditional_var <- var(beta_hats)`: We calculate the *sample* variance-covariance matrix of the `beta_hats` obtained from the simulations. This is our empirical estimate of $\mathrm{Var}(\hat{\beta}|X)$.

- `theoretical_conditional_var <- sigma_true^2 * solve(t(X) %*% X)`: We calculate the *theoretical* conditional variance-covariance matrix using the formula $\sigma^2(X^TX)^{-1}$, which holds under homoskedasticity.

4. **Output and Visualization:**
   - We print both the empirical and theoretical variance-covariance matrices. They should be close to each other.
   - We generate a scatterplot of the simulated $\hat{\beta}_0$ and $\hat{\beta}_1$ values. The spread of these points visually represents the conditional variance.

Connection to Concepts:

- **Conditional Variance (Theorem 19.2):** The script directly demonstrates the concept of the conditional variance of the OLS estimator, $\text{Var}(\hat{\beta}|X)$. We see how the estimator varies *for a fixed* X, and we compare the empirical variance to the theoretical formula $\sigma^2(X^TX)^{-1}$.
- **Homoskedasticity:** The simulation assumes homoskedasticity (constant error variance), which is why we can use the simplified formula for the conditional variance.

R Script 3: Demonstrating the Chi-Squared Distribution of the Wald Statistic

```
# Load necessary libraries
library(tidyverse)

# Set seed for reproducibility
set.seed(789)

# Simulation parameters
n <- 100  # Sample size
beta_true <- c(1, 2, -1)  # True intercept and slopes
sigma_true <- 2 # True error standard deviation

# Generate X
x1 <- runif(n, min = -2, max = 2)
x2 <- rnorm(n, mean = 0, sd = 1)
X <- cbind(1, x1, x2)
K <- ncol(X) # Number of parameters

# Number of simulations
n_sim <- 1000
wald_stats <- numeric(n_sim)

for (i in 1:n_sim) {
  # Generate errors
  epsilon <- rnorm(n, mean = 0, sd = sigma_true)

  # Generate y
  y <- X %*% beta_true + epsilon

  # Calculate OLS estimates
  beta_hat <- solve(t(X) %*% X) %*% t(X) %*% y

  # Calculate estimated variance-covariance matrix
  V_hat <- sigma_true^2 * solve(t(X) %*% X) # Use true sigma for simplicity
  #  In practice, use:  s2 <- sum((y - X%*%beta_hat)^2) / (n-K)
  #                      V_hat  <- s2 * solve(t(X)%*%X)

  # Calculate the Wald statistic (testing H0: beta = beta_true)
  wald_stats[i] <- t(beta_hat - beta_true) %*% solve(V_hat) %*% (beta_hat - beta_true)
}
```

```
# Visualize the distribution of the Wald statistic
data.frame(wald_stat = wald_stats) %>%
  ggplot(aes(x = wald_stat)) +
  geom_histogram(aes(y = after_stat(density)), bins = 30) +
  stat_function(fun = dchisq, args = list(df = K), color = "red") +
  labs(title = "Distribution of the Wald Statistic",
       x = "Wald Statistic",
       y = "Density") +
  xlim(0,15)
```

```
Warning: Removed 2 rows containing missing values or values outside the scale range
(`geom_bar()`).
```

Explanation:

1. **Setup:**
    - Set seed, define `n`, `beta_true`, and `sigma_true`.
    - Generate the independent variables `x1` and `x2` and create the design matrix `X`.
    - `K <- ncol(X)` determines the number of parameters.
2. **Simulation Loop:**
    - Loop `n_sim` times.
    - Generate errors and `y` in each iteration.
    - Calculate OLS estimates `beta_hat`.
    - `V_hat <- sigma_true^2 * solve(t(X) %*% X)`: Calculate the *estimated* variance-covariance matrix. For simplicity, we use the *true* σ^2 here. In practice, you would use the estimated error variance (s^2).
    - `wald_stats[i] <- ...`: Calculate the Wald statistic for the null hypothesis $H_0 : \beta = \beta_{true}$. This implements the formula $(\hat{\beta} - \beta)^T V(X)^{-1} (\hat{\beta} - \beta)$.
3. **Visualization:**
    - We create a histogram of the calculated Wald statistics.
    - `stat_function(...)`: We overlay the theoretical chi-squared distribution with K degrees of freedom (using `dchisq`). The histogram should closely follow the chi-squared distribution.

Connection to Concepts:

- **Chi-Squared Distribution (Theorem 19.3):** The script demonstrates the result that the quadratic form $(\hat{\beta} - \beta)^T V(X)^{-1} (\hat{\beta} - \beta)$ follows a chi-squared distribution with K degrees of freedom under the null hypothesis. This is a key result for constructing hypothesis tests and confidence regions.

R Script 4: Illustrating s^2 as an Unbiased Estimator of σ^2

```
library(tidyverse)

# Set seed for reproducibility
set.seed(101112)

# Simulation parameters
n <- 50  # Sample size
beta_true <- c(3, -1)  # True intercept and slope
sigma_true <- 2 # True error standard deviation

# Generate X
x <- runif(n, 0, 10)
X <- cbind(1, x)
k <- ncol(X)
```

```
# Number of simulations
n_sim <- 5000

# Store results
s2_values <- numeric(n_sim)
sigma2_hat_values <- numeric(n_sim)

for (i in 1:n_sim) {
  epsilon <- rnorm(n, 0, sigma_true)
  y <- X %*% beta_true + epsilon
  beta_hat <- solve(t(X) %*% X) %*% t(X) %*% y
  y_hat <- X %*% beta_hat
  residuals <- y - y_hat

  s2 <- sum(residuals^2) / (n - k)       # Unbiased estimator
  sigma2_hat <- sum(residuals^2) / n  # MLE (biased estimator)

  s2_values[i] <- s2
  sigma2_hat_values[i] <- sigma2_hat
}

# Calculate means
mean_s2 <- mean(s2_values)
mean_sigma2_hat <- mean(sigma2_hat_values)

cat("Mean of s2:", mean_s2, "\n")
```

```
Mean of s2: 4.011738
```

```
cat("Mean of sigma2_hat:", mean_sigma2_hat, "\n")
```

```
Mean of sigma2_hat: 3.851268
```

```
cat("True sigma^2:", sigma_true^2, "\n")
```

```
True sigma^2: 4
```

```
# Visualization
data.frame(s2 = s2_values, sigma2_hat = sigma2_hat_values) %>%
  pivot_longer(cols = everything(), names_to = "Estimator", values_to = "Estimate") %>%
  ggplot(aes(x = Estimate, fill = Estimator)) +
  geom_histogram(aes(y=after_stat(density)), position="identity", alpha=0.5, bins=30) +
  geom_vline(xintercept = sigma_true^2, color="red") +
  labs(title="Comparison of s2 and sigma2_hat", x="Estimate", y="Density") +
  xlim(0, 10)
```

```
Warning: Removed 4 rows containing missing values or values outside the scale range
(`geom_bar()`).
```

Explanation:

1. **Setup:** We define simulation parameters and generate the independent variable `x`.
2. **Simulation Loop:**
   - We generate errors and `y` for each simulation.
   - We calculate the OLS estimator $\hat{\beta}$.

- `residuals <- y - y_hat`: We calculate the residuals.
- `s2 <- sum(residuals^2) / (n - k)`: We calculate the *unbiased* estimator of σ^2, dividing by the degrees of freedom ($n-k$).
- `sigma2_hat <- sum(residuals^2) / n`: We calculate the MLE of σ^2 (which is biased), dividing by n.

3. **Analysis and Visualization:**

   - We calculate and print the mean of `s2_values` and `sigma2_hat_values` across all simulations. The mean of `s2` should be close to the true σ^2 (4), while the mean of `sigma2_hat` will be smaller.
   - The histogram visually compares the distributions of the two estimators. The `s2` distribution should be centered around the true value, while the `sigma2_hat` distribution will be shifted to the left.

Connection to Concepts:

- **Unbiased Estimator of σ^2 (Theorem 19.5):** The script directly demonstrates that $s^2 = \frac{\hat{\epsilon}^T\hat{\epsilon}}{n-K}$ is an unbiased estimator of σ^2, while the MLE $\hat{\sigma}^2_{MLE} = \frac{\hat{\epsilon}^T\hat{\epsilon}}{n}$ is biased downwards. The simulation shows that, on average, s^2 correctly estimates σ^2, while the MLE consistently underestimates it.
- **Degrees of freedom:** The script shows the importance of using $n-K$ and not n in the denominator of the estimator.

R Script 5: Gauss-Markov Theorem Illustration (Comparison with a Different Linear Unbiased Estimator)

```
library(tidyverse)

# Set seed
set.seed(131415)

# Parameters
n <- 50
beta_true <- c(2, 3)
sigma_true <- 2

# Generate X
x <- rnorm(n)
X <- cbind(1, x)

# Number of simulations
n_sim <- 1000

# Store results
ols_estimates <- matrix(NA, nrow = n_sim, ncol = 2)
alternative_estimates <- matrix(NA, nrow = n_sim, ncol = 2)


# Alternative linear unbiased estimator: average of y/x_i for each i
alternative_estimator <- function(X, y) {
  # This estimator is only for the simple linear regression case with intercept.
  #  It is designed to be linear and unbiased, but inefficient.
  # It takes the average of the two estimators: y_bar, and mean(y/x).

  beta1_hat <- mean(y / X[,2])  #  mean(y/x), biased for the intercept.
  beta0_hat <- mean(y) - beta1_hat*mean(X[,2]) # Correct for bias in intercept
  return(c(beta0_hat, beta1_hat))
}
```

```
for (i in 1:n_sim) {
  epsilon <- rnorm(n, 0, sigma_true)
  y <- X %*% beta_true + epsilon

  # OLS estimates
  ols_estimates[i, ] <- solve(t(X) %*% X) %*% t(X) %*% y

  # Alternative estimator
  alternative_estimates[i, ] <- alternative_estimator(X, y)
}

# Calculate variances
ols_variances <- apply(ols_estimates, 2, var)
alternative_variances <- apply(alternative_estimates, 2, var)

cat("OLS Variances:", ols_variances, "\n")
```

```
OLS Variances: 0.07026925 0.06798292
```

```
cat("Alternative Variances:", alternative_variances, "\n")
```

```
Alternative Variances: 0.08862799 1.168848
```

```
# Visualization
ols_df <- as.data.frame(ols_estimates)
colnames(ols_df) <- c("beta0_hat_ols", "beta1_hat_ols")
alt_df <- as.data.frame(alternative_estimates)
colnames(alt_df) <- c("beta0_hat_alt", "beta1_hat_alt")
combined_df <- bind_rows(ols_df %>% mutate(estimator = "OLS"),
                         alt_df %>% mutate(estimator = "Alternative"))

combined_df %>%
    pivot_longer(cols = c(-estimator), names_to = "parameter", values_to="estimate") %>%
    ggplot(aes(x = estimate, fill = estimator)) +
    geom_histogram(aes(y=after_stat(density)), position="identity", alpha=0.5, bins=30) +
    facet_wrap(~parameter, scales = "free") +
     geom_vline(data = data.frame(parameter = c("beta0_hat_ols", "beta1_hat_ols"),
                                  true_value = beta_true),
             aes(xintercept = true_value), color = "red") +
    labs(title="Comparison of OLS and Alternative Estimator", x="Estimate", y="Density")
```

```
Warning: Removed 4000 rows containing non-finite outside the scale range
(`stat_bin()`).
```

Explanation:

1. **Setup:**
    - We set the seed, define parameters, and generate `x`.
    - `alternative_estimator`: We define a function that implements a *different* linear unbiased estimator. This estimator is constructed to be linear in y and unbiased (we can verify this mathematically), but it is *not* efficient. For the case of a simple linear regression $y_i = \beta_0 + \beta_1 x_i + u_i$, this alternative estimator is given by taking $\hat{\beta}_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{x_i}$ and setting the intercept to satisfy the condition that the regression line passes through the sample mean, that is, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 * \bar{x}$.
2. **Simulation Loop:**

    - We generate errors and `y`.
    - We calculate both the OLS estimates *and* the estimates from our `alternative_estimator`.
3. **Analysis and Visualization:**
    - `ols_variances` and `alternative_variances`: We calculate the *sample variances* of the OLS estimates and the alternative estimates (across the simulations). The variances of the OLS estimates should be *smaller* than the variances of the alternative estimates.
    - The histograms visually compare the distributions of the OLS and alternative estimators. The OLS estimator's distribution should be more concentrated around the true values (less spread out), demonstrating its lower variance.

Connection to Concepts:

- **Gauss-Markov Theorem (Theorem 19.6):** The script illustrates the Gauss-Markov theorem. It shows that, among linear unbiased estimators, the OLS estimator has the smallest variance. We create an alternative linear unbiased estimator and show (through simulation) that its variance is larger than that of the OLS estimator.
- **BLUE (Best Linear Unbiased Estimator):** This example reinforces the meaning of "best" in BLUE – it refers to minimum variance within the class of linear unbiased estimators.

These five R scripts cover key concepts related to the statistical properties of the OLS estimator, illustrating unbiasedness, conditional variance, the chi-squared distribution of the Wald statistic, the unbiased estimation of the error variance, and the Gauss-Markov theorem. The simulations and visualizations help to solidify the understanding of these theoretical concepts. The code is well-commented, uses tidyverse style where appropriate, and provides clear explanations.

YouTube Videos Related to OLS Properties

Here are some YouTube videos that explain concepts mentioned in the attached text, along with explanations of their relevance. I have verified that these videos are currently available on YouTube (as of October 26, 2023).

1. Unbiasedness of OLS

- **Video Title:** "Econometrics // Lecture 3: OLS and Goodness-Of-Fit (R-Squared)" by *Ben Lambert*
- **Link:** https://www.youtube.com/watch?v=4xa0F54h-pM
- **Relevance to Text:** While this video covers multiple topics, a significant portion (starting around 2:50) is dedicated to demonstrating the **unbiasedness of the OLS estimator**. Lambert walks through the derivation, showing that $\mathbb{E}[\hat{\beta}] = \beta$, which is directly related to **Theorem 19.1** in the text. He uses a clear, step-by-step approach, explaining the assumptions involved. He visually shows how OLS works with an intuitive explanation. This complements the more formal proof in the text.

2. Variance of OLS and Gauss-Markov Theorem

- **Video Title:** "6. Properties of OLS: unbiasedness and BLUE theorem" by *Apoorva Javadekar*
- **Link:** https://www.youtube.com/watch?v=jI8INnEa58Q
- **Relevance to Text:** This video discusses the variance of the OLS estimator and introduces the **Gauss-Markov Theorem**.
    - It covers the derivation of $\text{Var}(\hat{\beta}|X) = \sigma^2(X^TX)^{-1}$ under homoskedasticity, corresponding to **Theorem 19.2**.
    - It explains the **Gauss-Markov Theorem (Theorem 19.6)**, stating that OLS is **BLUE** (Best Linear Unbiased Estimator). It presents the theorem's assumptions and its implications.
    - The presenter explains all the derivations step by step, making them simple to understand.

3. Derivation of OLS Estimator and its Properties

- **Video Title:** "2.2 Properties of OLS: derivation of the OLS estimator, unbiasedness, variance, and efficiency" by *Quantitative Economics with R*
- **Link:** https://www.youtube.com/watch?v=9-X1qWnWKtM
- **Relevance:** This video provides a comprehensive overview that closely aligns with several parts of the text. It includes:
    - **Derivation of the OLS Estimator:** It shows how to obtain $\hat{\beta} = (X^T X)^{-1} X^T y$ by minimizing the sum of squared residuals. This relates to **Definition 19.1**.
    - **Unbiasedness:** It proves $\mathbb{E}[\hat{\beta}] = \beta$ (**Theorem 19.1**).
    - **Variance:** It derives the variance-covariance matrix $\text{Var}(\hat{\beta}|X) = \sigma^2 (X^T X)^{-1}$ (**Theorem 19.2**, under homoskedasticity).
    - **Gauss-Markov:** It mentions the Gauss-Markov theorem and the BLUE property of OLS (**Theorem 19.6**), albeit briefly.
    - The video covers the derivations step by step.

4. Chi-squared Distribution and Hypothesis Testing

- **Video Title:** "Hypothesis Testing and Confidence Intervals with the OLS Estimator (Part 1)" by *BurkeyAcademy*
- **Link:** https://www.youtube.com/watch?v=-1o_MupV6_k
- **Relevance to Text:** While this video doesn't directly derive the chi-squared result for the Wald statistic, it explains the general principles of hypothesis testing in the context of OLS, which is a crucial application of **Theorem 19.3**. It covers:
    - How to build t-tests and F-tests to perform hypothesis testing.
    - The assumptions required for constructing tests to draw inferences.
    - Concepts like p-values and confidence intervals, which rely on the distributional results of the OLS estimator (including the normality assumption and the resulting t and chi-squared distributions).

5. Understanding the OLS estimator and its properties

- **Video Title:** "OLS estimator properties" by *Jochumzen*
- **Link:** https://www.youtube.com/watch?v=M9eRzPjeLpY
- **Relevance to Text:** This video serves as a good general overview and summary of the key properties of the OLS estimator:
    - **Linearity:** It highlights that OLS is a linear estimator (**Definition 19.1**).
    - **Unbiasedness:** It explains the concept of unbiasedness (**Theorem 19.1**).
    - **Variance:** It discusses the variance of the OLS estimator and its dependence on σ^2 and X (**Theorem 19.2**).
    - **Gauss-Markov Theorem:** It briefly mentions the Gauss-Markov Theorem (**Theorem 19.6**).

These videos provide a good visual and auditory complement to the text, covering many of the core concepts with varying levels of detail and mathematical rigor. They can be particularly helpful for grasping the intuition behind the derivations and understanding the practical implications of the theorems.

Multiple Choice Exercises

MC Exercise 1

MC Solution 1

The OLS estimator, $\hat{\beta}$, is considered “linear” because:

a. It is a linear function of the parameters β.
b. It is a linear function of the independent variables X.
c. It is a linear function of the dependent variable y.
d. The relationship between y and X is linear.

MC Exercise 2

MC Solution 2

Under Assumption A1 ($\mathbb{E}[\epsilon|X] = 0$), the OLS estimator $\hat{\beta}$ is:

a. Always biased.
b. Conditionally unbiased.
c. Unconditionally biased, but conditionally unbiased.
d. Only unbiased if the errors are normally distributed.

MC Exercise 3

MC Solution 3

The conditional variance of the OLS estimator, $\text{Var}(\hat{\beta}|X)$, under homoskedasticity is given by:

a. $(X^TX)^{-1}X^T\Sigma(X)X(X^TX)^{-1}$
b. $\sigma^2(X^TX)^{-1}$
c. σ^2X^TX
d. $\mathbb{E}[(X^TX)^{-1}]$

MC Exercise 4

MC Solution 4(#sec-ch19mcsolution4}

Which of the following statements is TRUE regarding conditional and unconditional unbiasedness?

a. Unconditional unbiasedness implies conditional unbiasedness.
b. Conditional unbiasedness implies unconditional unbiasedness.
c. Conditional and unconditional unbiasedness are unrelated concepts.
d. An estimator is either conditionally unbiased or unconditionally unbiased, but not both.

MC Exercise 5

MC Solution 5

The Law of Iterated Expectations is used to show that:

a. $\text{Var}(\hat{\beta}|X) = \sigma^2(X^TX)^{-1}$
b. Conditional unbiasedness implies unconditional unbiasedness.
c. The OLS estimator is BLUE.
d. The Wald statistic follows a chi-squared distribution.

MC Exercise 6

MC Solution 6

The expression $\text{Var}(\hat{\beta}) = \sigma^2 \mathbb{E}[(X^T X)^{-1}]$ holds true under:

a. Only heteroskedasticity.
b. Only homoskedasticity and no autocorrelation.
c. Any distribution of the error term.
d. Only when the sample size is large.

MC Exercise 7

MC Solution 7

In a simple linear regression $y_i = \beta x_i + \epsilon_i$, if $\mathbb{E}[\epsilon_i | x_i] = 0$, the OLS estimator of β is: (a) Biased (b) Unbiased (c) Only unbiased if n > 30 (d) Not defined

MC Exercise 8

MC Solution 8(#sec-ch19mcsolution8}

In the partitioned regression formula $\hat{\beta}_1 = (X_1^T M_2 X_1)^{-1} X_1^T M_2 y$, the matrix M_2 projects vectors onto the space:

a. Spanned by X_1.
b. Spanned by X_2.
c. Orthogonal to the space spanned by X_1.
d. Orthogonal to the space spanned by X_2.

MC Exercise 9

MC Solution 9

Under the assumption of normality of errors, the OLS estimator $\hat{\beta}$ follows a:

a. t-distribution.
b. Chi-squared distribution.
c. Normal distribution.
d. F-distribution.

MC Exercise 10

MC Solution 10(#sec-ch19mcsolution10}

The statistic $\tau = (\hat{\beta} - \beta)^T V(X)^{-1} (\hat{\beta} - \beta)$ follows a chi-squared distribution with degrees of freedom equal to:

a. n (sample size).
b. $n - K$ (sample size minus the number of parameters).
c. K (number of parameters).
d. $n - 1$.

MC Exercise 11

MC Solution 11(#sec-ch19mcsolution11}

$\hat{m}(x)$ in Theorem 19.4 represents:

a. The vector of predicted values for the original sample.
b. The predicted value for a specific set of independent variable values x.
c. The matrix X.
d. The error term.

MC Exercise 12

MC Solution 12(#sec-ch19mcsolution12}

The Maximum Likelihood Estimator (MLE) of σ^2 in the linear regression model with normal errors is:

a. $\dfrac{\hat{\epsilon}^T\hat{\epsilon}}{n-K}$
b. $\dfrac{\hat{\epsilon}^T\hat{\epsilon}}{n}$
c. $\dfrac{\hat{\epsilon}^T\hat{\epsilon}}{K}$
d. $(X^TX)^{-1}$

MC Exercise 13

MC Solution 13(#sec-ch19mcsolution13}

The estimator $s_*^2 = \dfrac{\hat{\epsilon}^T\hat{\epsilon}}{n-K}$ is preferred over $\hat{\sigma}^2_{mle} = \dfrac{\hat{\epsilon}^T\hat{\epsilon}}{n}$ because:

a. s_*^2 is the MLE.
b. s_*^2 is unbiased.
c. s_*^2 has a larger variance.
d. s_*^2 is always smaller.

MC Exercise 14

MC Solution 14(#sec-ch19mcsolution14}

The "degrees of freedom" in the context of estimating σ^2 are:

a. n
b. K
c. $n-1$
d. $n-K$

MC Exercise 15

MC Solution 15(#sec-ch19mcsolution15}

A BLUE estimator is:

a. Best Linear Unconditional Estimator.

b. Best Linear Unbiased Estimator.
c. Biased Linear Unconditional Estimator.
d. Best Logarithmic Unbiased Estimator.

MC Exercise 16

MC Solution 16(#sec-ch19mcsolution16}

The Gauss-Markov Theorem states that, under certain assumptions, the OLS estimator is:

a. Always the best estimator, even among non-linear estimators.
b. BLUE (Best Linear Unbiased Estimator).
c. Biased but consistent.
d. Only unbiased if the errors are normally distributed.

MC Exercise 17

MC Solution 17(#sec-ch19mcsolution17}

Which of the following is *NOT* an assumption of the Gauss-Markov Theorem?

a. $\mathbb{E}[\epsilon|X] = 0$
b. Homoskedasticity
c. No autocorrelation
d. Normality of the error terms

MC Exercise 18

MC Solution 18(#sec-ch19mcsolution18}

The Cramér-Rao Lower Bound (CRLB) provides a lower bound on the:

a. Bias of any estimator.
b. Variance of any unbiased estimator.
c. Mean Squared Error of any estimator.
d. Variance of any linear estimator.

MC Exercise 19

MC Solution 19(#sec-ch19mcsolution19}

The block diagonality of the information matrix for (β, σ^2) implies:

a. The OLS estimator is biased.
b. The MLEs of β and σ^2 are asymptotically independent.
c. The errors are heteroskedastic.
d. The model is non-linear.

MC Exercise 20

MC Solution 20(#sec-ch19mcsolution20} If $\text{Var}(\epsilon|X) = \Sigma(X)$, where $\Sigma(X)$ is not necessarily $\sigma^2 I$, the variance of the OLS estimator, $\text{Var}(\hat{\beta}|X)$ is given by:

a. $\sigma^2(X^TX)^{-1}$
b. $(X^TX)^{-1}X^T\Sigma(X)X(X^TX)^{-1}$
c. $(X^TX)^{-1}$
d. $\Sigma(X)$

Multiple Choice Solutions

MC Solution 1

MC Exercise 1

(c) It is a linear function of the dependent variable y.

Explanation:

Definition 19.1 states that the OLS estimator is linear in y because it can be expressed as $\hat{\beta} = Cy$, where $C = (X^TX)^{-1}X^T$ is a matrix that depends only on X (and not on y). Options (a) and (d) describe the linearity of the *model*, not the estimator. Option (b) is incorrect; the estimator is not necessarily a linear function of X.

MC Solution 2

MC Exercise 2

(b) Conditionally unbiased.

Explanation:

Theorem 19.1 states that under Assumption A1 ($\mathbb{E}[\epsilon|X] = 0$), the OLS estimator is *conditionally* unbiased: $\mathbb{E}[\hat{\beta}|X] = \beta$. This means that *given* the values of the independent variables, the expected value of the estimator equals the true parameter. Conditional unbiasedness, by the law of iterated expectations, also implies *unconditional* unbiasedness. The unbiasedness property does *not* require normality of the errors.

MC Solution 3

MC Exercise 3

(b) $\sigma^2(X^TX)^{-1}$

Explanation:

Theorem 19.2 provides the general formula for the conditional variance: $(X^TX)^{-1}X^T\Sigma(X)X(X^TX)^{-1}$. Under **homoskedasticity**, $\Sigma(X) = \sigma^2 I$, which simplifies the expression to $\sigma^2(X^TX)^{-1}$. Option (a) is the general formula without assuming homoskedasticity. Options (c) and (d) are incorrect.

MC Solution 4

MC Exercise 4

(b) Conditional unbiasedness implies unconditional unbiasedness.

Explanation:

This is a direct consequence of the **Law of Iterated Expectations**: $\mathbb{E}[\hat{\beta}] = \mathbb{E}[\mathbb{E}[\hat{\beta}|X]]$. If $\mathbb{E}[\hat{\beta}|X] = \beta$ (conditional unbiasedness), then $\mathbb{E}[\hat{\beta}] = \mathbb{E}[\beta] = \beta$ (unconditional unbiasedness). The reverse is not necessarily true.

MC Solution 5

MC Exercise 5

(b) Conditional unbiasedness implies unconditional unbiasedness.

Explanation:

As explained in Solution 4, the Law of Iterated Expectations is the key to showing that if an estimator is conditionally unbiased, it is also unconditionally unbiased.

MC Solution 6

MC Exercise 6

(b) Only homoskedasticity and no autocorrelation.

Explanation:

The general formula for $\text{Var}(\hat{\beta})$ involves $\Sigma(X)$, the variance-covariance matrix of the errors. The simplification to $\sigma^2\mathbb{E}[(X^TX)^{-1}]$ requires $\Sigma(X) = \sigma^2 I$. This holds when the errors are **homoskedastic** (constant variance, σ^2) and have **no autocorrelation** (errors are uncorrelated, so the off-diagonal elements of $\Sigma(X)$ are zero).

MC Solution 7

MC Exercise 7

(b) Unbiased

Explanation The condition $\mathbb{E}[\epsilon_i|x_i] = 0$ is a sufficient condition for the unbiasedness of OLS. It does not need a large n.

MC Solution 8

MC Exercise 8

(d) Orthogonal to the space spanned by X_2.

Explanation:

The matrix $M_2 = I - X_2(X_2^TX_2)^{-1}X_2^T$ is a **residual maker** or **annihilator matrix**. It projects vectors onto the space *orthogonal* to the column space of X_2. This means that M_2y gives the residuals of a regression of y on X_2 .

MC Solution 9

MC Exercise 9

(c) Normal distribution.

Explanation:

Theorem 19.3 states that if the errors are normally distributed (Assumption A3), then the OLS estimator $\hat{\beta}$ is also normally distributed (conditional on X). This is because $\hat{\beta}$ is a linear combination of the y values, which are themselves linear combinations of the normally distributed errors.

MC Solution 10

MC Exercise 10

(c) K (number of parameters).

Explanation:

Theorem 19.3 states that this quadratic form follows a chi-squared distribution with degrees of freedom equal to the number of parameters in the model, K. This is a fundamental result used for constructing Wald tests.

MC Solution 11

MC Exercise 11

(b) The predicted value for a specific set of independent variable values x.

Explanation:

$\hat{m}(x) = x^T\hat{\beta}$ is the estimated expected value of y *given* that the independent variables take on the specific values in the vector x. It's a scalar. $\hat{m}$, on the other hand, is the vector of predicted values for the *original* sample (i.e., for each row of the X matrix).

MC Solution 12

MC Exercise 12

(b) $\dfrac{\hat{\epsilon}^T\hat{\epsilon}}{n}$

Explanation:

This is derived by maximizing the log-likelihood function of the normal distribution with respect to σ^2. The resulting MLE is the sum of squared residuals divided by the number of observations, n. It is, however, biased.

MC Solution 13

MC Exercise 13

(b) s_*^2 is unbiased.

Explanation:

s_*^2 divides the sum of squared residuals by $n - K$ (degrees of freedom), which corrects for the bias in the MLE. **Theorem 19.5** proves that $\mathbb{E}[s_*^2] = \sigma^2$.

MC Solution 14

MC Exercise 14

(d) $n - K$

Explanation:

The degrees of freedom represent the number of independent pieces of information remaining to estimate the variance *after* estimating the K parameters in β.

MC Solution 15

MC Exercise 15

(b) Best Linear Unbiased Estimator.

Explanation:

BLUE stands for Best Linear Unbiased Estimator. “Best” means minimum variance within the class of linear unbiased estimators.

MC Solution 16

MC Exercise 16

(b) BLUE (Best Linear Unbiased Estimator).

Explanation:

The **Gauss-Markov Theorem (Theorem 19.6)** states that under the assumptions of linearity, strict exogeneity, homoskedasticity, and no autocorrelation, the OLS estimator is the Best Linear Unbiased Estimator (BLUE). It does *not* claim that OLS is the best among *all* estimators (including non-linear ones) – that requires normality of the errors.

MC Solution 17

MC Exercise 17

(d) Normality of the error terms

Explanation:

The Gauss-Markov Theorem *does not* require the assumption of normality of the error terms. Normality is needed for certain hypothesis tests and confidence interval constructions, but not for the OLS estimator to be BLUE. Assumptions (a), (b), and (c) *are* required for the Gauss-Markov Theorem.

MC Solution 18

MC Exercise 18

(b) Variance of any unbiased estimator.

Explanation:

The CRLB provides a lower bound on the variance of *any unbiased* estimator. It doesn't apply to biased estimators, and the Mean Squared Error (MSE) considers both bias and variance.

MC Solution 19

MC Exercise 19

(b) The MLEs of β and σ^2 are asymptotically independent.

Explanation:

The block diagonality of the information matrix implies that the MLEs of β and σ^2 are asymptotically independent. This means that, in large samples, knowing the true value of one parameter doesn't provide any additional information for estimating the other.

MC Solution 20

MC Exercise 20

(b) $(X^TX)^{-1}X^T\Sigma(X)X(X^TX)^{-1}$

Explanation This is the general formula of the variance of the OLS estimator. It does not require constant variance of the errors.

Author: Peter Fuleky