Chapter 21: Omission of Relevant Variables, Inclusion of Irrelevant Variables, and Model Selection

This chapter discusses the consequences of misspecifying a regression model by either omitting relevant variables or including irrelevant ones. It also covers model selection techniques.

21.1 Omission of Relevant Variables

Suppose the true model is:

$$y = X_1\beta_1 + X_2\beta_2 + \epsilon,$$

where Assumption A1 holds. However, we regress y on X_1 only, perhaps because we did not observe X_2.

The OLS estimator for β_1 in this misspecified model is:

$$\tilde{\beta_1} = (X_1^T X_1)^{-1} X_1^T y.$$

Substituting the true model for y:

$$\begin{aligned}\tilde{\beta_1} &= (X_1^T X_1)^{-1} X_1^T (X_1\beta_1 + X_2\beta_2 + \epsilon)\\ &= (X_1^T X_1)^{-1} X_1^T X_1\beta_1 + (X_1^T X_1)^{-1} X_1^T X_2\beta_2 + (X_1^T X_1)^{-1} X_1^T \epsilon\\ &= \beta_1 + (X_1^T X_1)^{-1} X_1^T X_2\beta_2 + (X_1^T X_1)^{-1} X_1^T \epsilon.\end{aligned}$$

Taking expectations conditional on X:

$$\begin{aligned}E(\tilde{\beta_1}|X) &= \beta_1 + (X_1^T X_1)^{-1} X_1^T X_2\beta_2\\ &= \beta_1 + \beta_{12},\end{aligned}$$

where $\beta_{12} = (X_1^T X_1)^{-1} X_1^T X_2\beta_2$.

Intuition: The OLS estimator $\tilde{\beta_1}$ is biased. The bias, β_{12}, arises from the correlation between the included regressor X_1 and the omitted variable X_2. If X_1 and X_2 are uncorrelated, then $X_1^T X_2 = 0$, and there is no bias. If they are correlated, the OLS estimator of the coefficient on X_1 picks up some of the effect of X_2 on y.

In general, $\tilde{\beta_1}$ is biased and inconsistent. The direction and magnitude of the bias depend on β_2 and on $X_1^T X_2$.

Example 21.1

Some common examples of omitted variables:

1. **Seasonality:** In time series data, failing to include seasonal dummy variables can bias coefficients if other variables have seasonal patterns.
2. **Dynamics:** Ignoring lagged variables in a dynamic model can lead to bias.
3. **Nonlinearity:** If the true relationship is nonlinear but a linear model is fitted, the coefficients may be biased.

4. **Endogeneity:** Omitting variables that are correlated with both the dependent variable and included independent variables leads to omitted variable bias, a specific case of which is endogeneity.

For example: consider a wage equation. Wages depend on education and ability. If you omit ability, and ability is positively correlated with both wages and education, then omitting ability results in upward bias of the effect of education. Similarly, if studying the effect of race or gender on wages, you might omit experience or education which could bias the estimated effect of race or gender.

Variance of the estimator:

The conditional variance of the biased estimator $\tilde{\beta_1}$ given X is:

$$\text{var}(\tilde{\beta_1}|X) = (X_1^T X_1)^{-1} X_1^T \Sigma(X) X_1 (X_1^T X_1)^{-1},$$

where $\Sigma(X)$ is the variance-covariance matrix of the error term.

This should be compared with the variance of the estimator from the correct model that includes X_2:

$$\text{var}(\hat{\beta_1}|X) = (X_1^T M_2 X_1)^{-1} X_1^T M_2 \Sigma(X) M_2 X_1 (X_1^T M_2 X_1)^{-1},$$

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

In the homoskedastic case, $\Sigma(X) = \sigma^2 I$, and the variances simplify to:

$$\text{var}(\tilde{\beta_1}|X) = \sigma^2 (X_1^T X_1)^{-1}$$

and

$$\text{var}(\hat{\beta_1}|X) = \sigma^2 (X_1^T M_2 X_1)^{-1}.$$

Since $X_1^T X_1$ is “larger” than $X_1^T M_2 X_1$, including the relevant variable (X_2) reduces the variance of the estimator of the coefficients on X_1. The comparison between the MSE of the two estimators is:

$$(X_1^T X_1)^{-1} X_1^T X_2 \beta_2 \beta_2^T X_2^T X_1 (X_1^T X_1)^{-1} + \sigma^2 (X_1^T X_1)^{-1} \text{ versus } \sigma^2 (X_1^T M_2 X_1)^{-1}$$

There’s no ranking between these two performance measures in general. However, with large samples, the variance components go to zero. So you prefer the estimator that controls for X_2.

Effect on Inference:

The standard estimated variance of $\tilde{\beta_1}$ is $s_u^2 (X_1^T X_1)^{-1}$, where

$$s_u^2 = \frac{y^T M_1 y}{n-K_1} = \frac{(X_2\beta_2+\epsilon)^T M_1 (X_2\beta_2+\epsilon)}{n-K_1}.$$

Expanding the numerator and taking the expectation (under homoskedasticity):

$$E[s_u^2] = \sigma^2 + \frac{\beta_2^T X_2^T M_1 X_2 \beta_2}{n-K_1} \geq \sigma^2.$$

The inequality follows because M_1 is a positive semi-definite matrix. Therefore, the estimated variance of $\tilde{\beta_1}$ is upwardly biased. The t-statistic is not distributed as t anymore.

If $X_1^T X_2 = 0$, then $\tilde{\beta}$ is unbiased, but the OLS standard errors are still biased. In this case, the t-ratio is downward biased, and so t-tests are less likely to reject the null hypothesis.

In general, the direction of the bias in the t-ratio is ambiguous, depending on the bias of $\tilde{\beta_1}$.

Nonlinear Example:

Suppose the true regression function is nonlinear:

$$y_i = \alpha + \beta x_i + \gamma x_i^2 + \epsilon_i.$$

If we omit the quadratic term x_i^2, the OLS estimators for the intercept and slope coefficients are:

$$E[\tilde{\beta}] = \beta + \gamma \frac{\sum_{i=1}^n (x_i - \bar{x}) x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$E[\tilde{\alpha}] = \alpha + \gamma \left[\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x} \frac{\sum_{i=1}^n (x_i - \bar{x}) x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$. If $\bar{x} = 0$, the bias of the slope depends on the skewness of the regressors and γ. The bias of the intercept depends on γ and variance.

21.2 Inclusion of Irrelevant Variables/Knowledge of Parameters

Suppose the true model is:

$$y = X_1 \beta_1 + \epsilon,$$

but we regress y on both X_1 and X_2. The estimator of β_1 is:

$$\tilde{\beta_1} = (X_1^T M_2 X_1)^{-1} X_1^T M_2 y = \beta_1 + (X_1^T M_2 X_1)^{-1} X_1^T M_2 \epsilon.$$

The expected value and conditional variance are:

$$E(\tilde{\beta_1} | X) = \beta_1$$

$$\text{var}(\tilde{\beta_1} | X) = \sigma^2 (X_1^T M_2 X_1)^{-1}.$$

Compare this to the correct model's estimator:

$$\text{var}(\hat{\beta_1} | X) = \sigma^2 (X_1^T X_1)^{-1}.$$

Since $X_1^T X_1 - X_1^T M_2 X_1 = X_1^T P_2 X_1 \geq 0$,

$$(X_1^T X_1)^{-1} - (X_1^T M_2 X_1)^{-1} \leq 0,$$

the variance is smaller when using only the relevant regressors (X_1). Including irrelevant variables inflates the variance.

Restricted least squares You have linear restrictions: $R\beta = r$. Using restricted least squares gives smaller variances.

Bias Variance Tradeoff: There is a trade-off between bias and variance. - Big Model: low bias, high variance. - Small Model: small variance, large bias.

Known parameters: If we know β_2 and use $y - X_2 \beta_2$ on X_1, this gives better results than using y on X_1 and X_2. Knowledge of parameters is valuable.

21.3 Model Selection

Let M be a collection of linear regression models obtained from a given set of K regressors $X = (X_1, \dots, X_K)$. There are $2^K - 1$ different subsets of X, and hence different regression models that can be counted as submodels. Suppose the true model lies in M:

$$y = X\beta + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2 I_n)$. Some of the β_i could be zero.

We want to select among different submodels when $K < n$, in which case all submodels can be estimated.

Let K_j be the number of explanatory variables in a given regression (with some selection of K_j regressors), and let $\hat{\epsilon_j}$ denote the vector of residuals. The unbiased estimator of the error variance is:

$$s^2_{*j} = \frac{\hat{\epsilon_j}^T \hat{\epsilon_j}}{n-K_j}.$$

This estimator makes a trade-off between goodness of fit and parsimony. Expanding around the average sum of squared residuals:

$$s^2_{*j} = \frac{\hat{\epsilon_j}^T \hat{\epsilon_j}}{n-K_j} = \frac{\hat{\epsilon_j}^T \hat{\epsilon_j}}{n} \cdot \frac{1}{1-\frac{K_j}{n}} = \frac{\hat{\epsilon_j}^T \hat{\epsilon_j}}{n} \left(1 + \frac{K_j}{n} + \dots\right).$$

The first term $\frac{\hat{\epsilon_j}^T \hat{\epsilon_j}}{n}$ measures goodness of fit. Minimizing s^2_{*j} is equivalent to maximizing adjusted R^2:

$$\bar{R_j}^2 = 1 - \frac{n-1}{n-K_j}(1 - R_j^2) = 1 - \frac{n-1}{n-K_j} \frac{\hat{\epsilon_j}^T \hat{\epsilon_j}}{\hat{\epsilon}^T \hat{\epsilon}}.$$

Other model selection criteria: - **Prediction Criterion:** $PC_j = \frac{\hat{\epsilon_j}^T \hat{\epsilon_j}}{n-K_j}(1 + \frac{K_j}{n})$ - **Akaike Information Criterion:** $AIC_j = \ln(\frac{\hat{\epsilon_j}^T \hat{\epsilon_j}}{n}) + \frac{2K_j}{n}$ - **Bayesian Information Criterion:** $BIC_j = \ln(\frac{\hat{\epsilon_j}^T \hat{\epsilon_j}}{n}) + \frac{K_j \ln(n)}{n}$

These criteria allow model comparisons with different covariates. - All criteria have the property that the selected model is bigger or equal than true model with probability going to one. - Only BIC_j correctly selects the true model with probability going to one (because it has the largest penalty).

Theorem 21.1

$\bar{R}^2$ falls (rises) when the deleted variable has $t > (<)1$.

21.3.1 Problems and Issues

1. **Computational Issues:** Searching over all 2^K regressions can be infeasible. You can start small and add variables, or start large and remove them.
2. **True model not in M:** Even if the true model is not in M, the search is guaranteed to find the best in the model set M.
3. **Other criteria:** Consistency with economic theory, consistency with the data (dependent variable in reasonable range, random residuals, out of sample performance)
4. **Data mining:** White's (2000) Reality Check Paper Consider regressions with 2 variables from a large set K. Total number of regressions is $K(K-1)/2$. The best fitting regression will have high in-sample R^2 even if variables are all mutually independent.

21.4 LASSO

The Lasso method (Tibshirani, 1996) is an alternative model selection procedure.

Suppose we have n observations on candidate covariates $X_1, X_2, \ldots, X_K$ and an outcome variable y, where $K > n$.

Ridge Regression Estimator:

$$\hat{\beta_R} = (X^T X + \lambda I_K)^{-1} X^T y,$$

where $\lambda > 0$. This solution always exists because $X^T X \geq 0$ implies $X^T X + \lambda I_K > 0$.

The ridge regression is the solution to the penalized least squares criterion:

$$\min_{b_1, b_2, \ldots, b_K} \sum_{i=1}^{n} (y_i - b_1 - b_2 x_{2i} - b_3 x_{3i} - \cdots - b_K x_{Ki})^2 + \lambda \sum_{j=1}^{K} b_j^2.$$

Sparsity Assumption:

Assume that the true model is of much smaller dimension.

$$E(Y|X_1, \ldots, X_K) = \sum_{k=1}^{K_0} \beta_{jk} X_{jk},$$

for some $K_0 \leq K$. Denote the active set by

$$A = \{j : \beta_j \neq 0\},$$

and the inactive set by its complement in $\{1, \ldots, K\}$. The assumption that K_0 is much smaller than n is called **sparsity**.

The Lasso method solves the following constrained minimization problem:

$$\min_{b_1, b_2, \ldots, b_K} \sum_{i=1}^{n} (y_i - b_1 - b_2 x_{2i} - \cdots - b_K x_{Ki})^2 + \lambda \sum_{j=1}^{K} |b_j|.$$

This is called L_1 penalty. The problem can be equivalently written as

$$\min_{b_1, \ldots, b_K} \sum_{i=1}^{n} (y_i - b_1 - b_2 x_{2i} - \cdots - b_K x_{Ki})^2 \text{ s.t. } \sum_{j=1}^{K} |b_j| \leq s.$$

- s is a turning parameter
- If s is large enough the constraint does not matter, and it’s OLS.
- If s is small, we get shrunken versions of the least squares estimates with several coefficients set to zero.

Define the selected set: $A = \{j : \beta_j \neq 0\}$

It’s been shown under some conditions that: $P(A = \hat{A}) \to 1$ That is the Lasso selects the correct model with probability going to one.

Exercises

Exercise 1

Solution 1

Consider the model $y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$, where ϵ_i are i.i.d. with mean 0 and variance σ^2. Suppose you estimate the model by regressing y_i on x_{1i} only. Derive the expression for the bias of the OLS estimator of β_1.

Under what condition is the estimator unbiased?

Exercise 2

Solution 2

Consider the model in Exercise 1. Assuming homoskedasticity, derive the variance of the OLS estimator of β_1 when you regress y_i only on x_{1i}. Compare this variance to the variance of the OLS estimator of β_1 when you regress y_i on both x_{1i} and x_{2i}. Which variance is smaller? Explain the intuition.

Exercise 3

Solution 3

Consider the model in Exercise 1. Suppose you regress y_i on x_{1i} only. Show that the usual OLS estimator of the error variance is biased upwards.

Exercise 4

Solution 4

True or False: When relevant variables are omitted, the t-ratio for the included variables is always downward biased. Explain.

Exercise 5

Solution 5

Consider the true model $y_i = \alpha + \beta x_i + \gamma x_i^2 + \epsilon_i$. Suppose you estimate the model by regressing y_i on a constant and x_i, ignoring the quadratic term. Show that if $\bar{x} = 0$, the bias of the OLS estimator of the slope coefficient depends on the skewness of x.

Exercise 6

Solution 6

Consider the true model $y = X_1\beta_1 + \epsilon$. You estimate the model by regressing y on X_1 and X_2, where X_2 are irrelevant variables. Show that the OLS estimator of β_1 is unbiased.

Exercise 7

Solution 7

Consider the scenario in Exercise 6. Show that the variance of the OLS estimator of β_1 is larger when X_2 is included compared to when it is excluded.

Exercise 8

Solution 8

Explain the bias-variance trade-off in the context of model selection.

Exercise 9

Solution 9

Explain how the adjusted R^2 is related to the unbiased estimator of the error variance in a regression model.

Exercise 10

Solution 10

What is the key difference between the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) in terms of model selection?

Exercise 11

Solution 11

Among the model selection criteria discussed in the text (adjusted R^2, PC, AIC, BIC), which one is guaranteed to select the true model with probability tending to one as the sample size grows?

Exercise 12

Solution 12

What does Theorem 21.1 in the text state about the relationship between the adjusted R^2 and the t-statistic of a deleted variable?

Exercise 13

Solution 13

What does “sparsity” mean in the context of the LASSO method?

Exercise 14

Solution 14

Write down the objective function that the LASSO method minimizes. Explain the role of the tuning parameter λ.

Exercise 15

Solution 15

How does the LASSO method achieve variable selection?

Exercise 16

Solution 16

What is the main difference between the penalty terms in Ridge regression and LASSO?

Exercise 17

Solution 17

Explain the concept of “active set” in the context of the LASSO method.

Exercise 18

Solution 18

Under what condition does the LASSO select the correct model with probability tending to one?

Exercise 19

Solution 19

If the bound s in the alternative formulation of the LASSO problem is very large, what does the LASSO solution converge to?

Exercise 20

Solution 20 How is cross-validation used in the context of the LASSO?

Solutions

Solution 1

Exercise 1

The true model is:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i.$$

The misspecified model regresses y_i on x_{1i} only:

$$y_i = \beta_1 x_{1i} + v_i,$$

where $v_i = \beta_2 x_{2i} + \epsilon_i$.

The OLS estimator of β_1 in the misspecified model is:

$$\tilde{\beta}_1 = (X_1^T X_1)^{-1} X_1^T y.$$

Substituting the true model for y:

$$\begin{aligned}\tilde{\beta}_1 &= (X_1^T X_1)^{-1} X_1^T (\beta_1 X_1 + \beta_2 X_2 + \epsilon)\\ &= (X_1^T X_1)^{-1} X_1^T X_1 \beta_1 + (X_1^T X_1)^{-1} X_1^T X_2 \beta_2 + (X_1^T X_1)^{-1} X_1^T \epsilon\\ &= \beta_1 + (X_1^T X_1)^{-1} X_1^T X_2 \beta_2 + (X_1^T X_1)^{-1} X_1^T \epsilon.\end{aligned}$$

Taking expectations:

$$E(\tilde{\beta}_1) = \beta_1 + (X_1^T X_1)^{-1} X_1^T X_2 \beta_2.$$

The **bias** is $E(\tilde{\beta_1}) - \beta_1 = (X_1^T X_1)^{-1} X_1^T X_2 \beta_2$.

The estimator is unbiased if $(X_1^T X_1)^{-1} X_1^T X_2 \beta_2 = 0$. This occurs if $X_1^T X_2 = 0$ (i.e., X_1 and X_2 are orthogonal) or if $\beta_2 = 0$ (i.e., X_2 does not belong in the true model).

Solution 2

Exercise 2

Misspecified model (regressing y_i on x_{1i} only):

$$\tilde{\beta_1} = (X_1^T X_1)^{-1} X_1^T y.$$

Under homoskedasticity, $Var(\epsilon|X) = \sigma^2 I$. Since $v_i = \beta_2 x_{2i} + \epsilon_i$, $Var(v|X) = \sigma^2 I$ assuming X_2 is non-stochastic or conditioning on X_2

$$\begin{aligned} Var(\tilde{\beta_1}|X) &= Var((X_1^T X_1)^{-1} X_1^T y|X) \\ &= (X_1^T X_1)^{-1} X_1^T Var(y|X) X_1 (X_1^T X_1)^{-1} \\ &= (X_1^T X_1)^{-1} X_1^T Var(\beta_2 X_2 + \epsilon|X) X_1 (X_1^T X_1)^{-1} \\ &= \sigma^2 (X_1^T X_1)^{-1}. \end{aligned}$$

Correct model (regressing y_i on x_{1i} and x_{2i}):

$\hat{\beta} = (X^T X)^{-1} X^T y$, where $X = [X_1, X_2]$.

$Var(\hat{\beta}|X) = \sigma^2 (X^T X)^{-1}$. Partitioning: $Var(\hat{\beta}_1|X) = \sigma^2 (X_1^T M_2 X_1)^{-1}$

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

Comparison:

We need to compare $\sigma^2 (X_1^T X_1)^{-1}$ and $\sigma^2 (X_1^T M_2 X_1)^{-1}$.

Since M_2 is idempotent and symmetric, we can write: $X_1^T X_1 = X_1^T M_2 X_1 + X_1^T P_2 X_1$, where $P_2 = X_2 (X_2^T X_2) X_2^T$

Because $X_1^T P_2 X_1$ is positive semi-definite, $(X_1^T X_1)^{-1} - (X_1^T M_2 X_1)^{-1} \leq 0$ by properties of positive definite matrices.

Therefore, $Var(\tilde{\beta_1}|X) \geq Var(\hat{\beta}_1|X)$. The variance is smaller when the correct model (including X_2) is used.

Intuition: Including the relevant variable X_2 reduces the noise in the estimation of β_1, leading to a more precise (lower variance) estimator.

Solution 3

Exercise 3

The misspecified model is $y_i = \beta_1 x_{1i} + v_i$, where $v_i = \beta_2 x_{2i} + \epsilon_i$. The OLS estimator of the error variance is:

$s_u^2 = \frac{\tilde{\epsilon}^T\tilde{\epsilon}}{n-1} = \frac{y^TM_1y}{n-1}$, where $M_1 = I - X_1(X_1^TX_1)^{-1}X_1^T$ Substituting the true model for y: $s_u^2 = \frac{(\beta_2X_2+\epsilon)^TM_1(\beta_2X_2+\epsilon)}{n-1}$

Expanding the numerator and using $M_1X_1 = 0$:

$s_u^2 = \dfrac{\beta_2^TX_2^TM_1X_2\beta_2 + 2\beta_2^TX_2^TM_1\epsilon + \epsilon^TM_1\epsilon}{n-1}$. Taking expectations, and noting that $E[\epsilon|X] = 0$ and $E[\epsilon^TM_1\epsilon|X] = \sigma^2 tr(M_1) = \sigma^2(n-1)$

$$E[s_u^2|X] = \frac{\beta_2^TX_2^TM_1X_2\beta_2 + \sigma^2(n-1)}{n-1} = \sigma^2 + \frac{\beta_2^TX_2^TM_1X_2\beta_2}{n-1}.$$

Since M_1 is positive semi-definite, $\beta_2^TX_2^TM_1X_2\beta_2 \geq 0$. Thus, $E[s_u^2|X] \geq \sigma^2$.

The usual OLS estimator of the error variance is biased upwards.

Solution 4

Exercise 4

False. The t-ratio can be either upward or downward biased. The direction of the bias in the t-ratio depends on the direction of the bias in the coefficient estimate and the bias in the estimated standard error. As shown in the text and in Solution 3, the estimated variance of the coefficient is upward biased. However, the coefficient estimate itself can be either upward or downward biased, depending on the correlation between the included and omitted variables, and the sign of the coefficient on the omitted variable.

Solution 5

Exercise 5

The true model is:

$$y_i = \alpha + \beta x_i + \gamma x_i^2 + \epsilon_i.$$

The misspecified model is:

$y_i = \alpha + \beta x_i + v_i$, where $v_i = \gamma x_i^2 + \epsilon_i$.

The OLS estimator for β is:

$$\tilde{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Substituting the true model for y_i and simplifying:

$$\tilde{\beta} = \beta + \gamma\frac{\sum_{i=1}^n (x_i - \bar{x})x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{\sum_{i=1}^n (x_i - \bar{x})\epsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Taking expectations:

$$E[\tilde{\beta}] = \beta + \gamma\frac{\sum_{i=1}^n (x_i - \bar{x})x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

If $\bar{x} = 0$, then

$$E[\tilde{\beta}] = \beta + \gamma \frac{\sum_{i=1}^{n} x_i^3}{\sum_{i=1}^{n} x_i^2}.$$

The term $\frac{\sum_{i=1}^{n} x_i^3}{n}$ is a measure of the skewness of x. Thus, the bias depends on the skewness of x and the value of γ.

Solution 6

Exercise 6

True model: $y = X_1\beta_1 + \epsilon$.

Misspecified model: $y = X_1\beta_1 + X_2\beta_2 + \epsilon$. Since X_2 are irrelevant, $\beta_2 = 0$ in population.

The OLS estimator of $\beta = [\beta_1, \beta_2]$ is: $\tilde{\beta} = (X^T X)^{-1} X^T y$ with $X = [X_1 \vdots X_2]$

We can partition the estimator as follows:

$$\begin{bmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_2 \end{bmatrix} = \begin{bmatrix} X_1^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_2^T X_2 \end{bmatrix}^{-1} \begin{bmatrix} X_1^T \\ X_2^T \end{bmatrix} y$$

Using the formula for the inverse of partitioned matrix and simplifying:

$\tilde{\beta}_1 = (X_1^T M_2 X_1)^{-1} X_1^T M_2 y$ where $M_2 = I - X_2(X_2^T X_2)^{-1} X_2^T$ Substitute for $y = X_1\beta_1 + \epsilon$
$\tilde{\beta}_1 = (X_1^T M_2 X_1)^{-1} X_1^T M_2 (X_1\beta_1 + \epsilon) = \beta_1 + (X_1^T M_2 X_1)^{-1} X_1^T M_2 \epsilon$

Taking expectations:

$$E[\tilde{\beta}_1 | X] = \beta_1.$$

The OLS estimator of β_1 is unbiased, even when irrelevant variables are included.

Solution 7

Exercise 7

From Solution 6, the estimator of β_1 when X_2 is included is:

$$\tilde{\beta}_1 = (X_1^T M_2 X_1)^{-1} X_1^T M_2 y,$$

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

The variance is:

$$\begin{aligned} Var(\tilde{\beta}_1 | X) &= (X_1^T M_2 X_1)^{-1} X_1^T M_2 Var(y|X) M_2 X_1 (X_1^T M_2 X_1)^{-1} \\ &= \sigma^2 (X_1^T M_2 X_1)^{-1}. \end{aligned}$$

When X_2 is excluded, the estimator of β_1 is:

$$\hat{\beta}_1 = (X_1^T X_1)^{-1} X_1^T y.$$

The variance is:

$$Var(\hat{\beta}_1|X) = \sigma^2(X_1^T X_1)^{-1}.$$

Since $X_1^T X_1 - X_1^T M_2 X_1 = X_1^T P_2 X_1 \geq 0$,

$$(X_1^T X_1)^{-1} \leq (X_1^T M_2 X_1)^{-1}.$$

Therefore, $Var(\hat{\beta}_1|X) \leq Var(\tilde{\beta}_1|X)$. The variance is larger when irrelevant variables are included.

Solution 8

Exercise 8

The **bias-variance trade-off** refers to the relationship between the bias and variance of an estimator in different model specifications.

- **Large Model (more variables):** Tends to have *low bias* because it is more likely to include all relevant variables. However, it has *high variance* because it includes irrelevant variables, adding noise to the estimation.
- **Small Model (fewer variables):** Tends to have *low variance* because it excludes irrelevant variables. However, it has *high bias* because it may omit relevant variables.

Model selection involves finding a balance between these two extremes.

Solution 9

Exercise 9

The unbiased estimator of the error variance is:

$$s_{*j}^2 = \frac{\hat{\epsilon_j}^T \hat{\epsilon_j}}{n-K_j}.$$

The adjusted R^2 is:

$$\bar{R_j}^2 = 1 - \frac{n-1}{n-K_j}(1 - R_j^2) = 1 - \frac{n-1}{n-K_j}\frac{ESS}{TSS} = 1 - \frac{\frac{\hat{\epsilon_j}^T \hat{\epsilon_j}}{n-K_j}}{\frac{\sum(y_i-\bar{y})^2}{n-1}}.$$

Where ESS is the error sum of squares, TSS is the total sum of squares, and K_j is the number of regressors in model j. The adjusted R^2 can be written in terms of the unbiased error of variance (s_{*j}^2):

$$\bar{R_j}^2 = 1 - \frac{s_{*j}^2}{\frac{\sum(y_i-\bar{y})^2}{n-1}}.$$

Maximizing $\bar{R_j}^2$ is equivalent to minimizing s_{*j}^2. The adjusted R^2 penalizes the inclusion of additional variables by dividing the sum of squared residuals by $(n - K_j)$ instead of n. This penalty reflects the loss of degrees of freedom.

Solution 10

Exercise 10

AIC and BIC both penalize model complexity, but BIC imposes a larger penalty for additional parameters.

- **AIC:** $\ln(\frac{\hat{\epsilon_j}^T\hat{\epsilon_j}}{n}) + \frac{2K_j}{n}$. Penalty term: $\frac{2K_j}{n}$.
- **BIC:** $\ln(\frac{\hat{\epsilon_j}^T\hat{\epsilon_j}}{n}) + \frac{K_j\ln(n)}{n}$. Penalty term: $\frac{K_j\ln(n)}{n}$.

Since $\ln(n) > 2$ for $n > 7$, BIC penalizes additional parameters more heavily than AIC. This means that BIC tends to favor simpler models compared to AIC.

Solution 11

Exercise 11

The **Bayesian Information Criterion (BIC)** is guaranteed to select the true model with probability tending to one as the sample size grows. This is because it has the largest penalty term among the mentioned criteria.

Solution 12

Exercise 12

Theorem 21.1 states that the adjusted R^2 ($\bar{R}^2$) *falls* when the deleted variable has a t-statistic with absolute value *greater* than 1, and $\bar{R}^2$ *rises* when the deleted variable has a t-statistic with absolute value *less* than 1. In other words, $\bar{R}^2$ increases if and only if the t-statistic of the added (or deleted) variable is less than 1 in absolute value.

Solution 13

Exercise 13

Sparsity means that the true model has a relatively small number of non-zero coefficients compared to the total number of potential predictors. In other words, only a few of the available covariates actually affect the dependent variable; most have coefficients of zero.

Solution 14

Exercise 14

The LASSO minimizes the following objective function:

$$\min_{b_1,b_2,...,b_K} \sum_{i=1}^n (y_i - b_1 - b_2x_{2i} - \cdots - b_Kx_{Ki})^2 + \lambda\sum_{j=1}^K |b_j|.$$

The first part is the residual sum of squares, measuring the goodness of fit. The second part is the L_1 penalty, which is the sum of the absolute values of the coefficients, multiplied by the tuning parameter λ.

- λ controls the strength of the penalty.
  - If $\lambda = 0$, there is no penalty, and LASSO reduces to OLS.
  - As λ increases, the penalty becomes stronger, forcing more coefficients to be exactly zero. This leads to variable selection.

Solution 15

Exercise 15

The LASSO achieves variable selection through the L_1 penalty, $\lambda\sum_{j=1}^K |b_j|$. Because of the absolute value in the penalty, the optimization problem has a non-differentiable point at zero for each coefficient. This encourages some coefficients to be *exactly* zero, effectively removing the corresponding variables from the model. This is in contrast to Ridge regression, which uses an L_2 penalty and shrinks coefficients towards zero but rarely sets them exactly to zero.

Solution 16

Exercise 16

- **Ridge Regression:** Uses an L_2 penalty: $\lambda \sum_{j=1}^{K} b_j^2$. This penalty shrinks coefficients towards zero but generally does not set them exactly to zero.
- **LASSO:** Uses an L_1 penalty: $\lambda \sum_{j=1}^{K} |b_j|$. This penalty encourages sparsity, setting some coefficients to exactly zero.

The key difference is the use of the absolute value (L_1) versus the square (L_2) of the coefficients in the penalty term.

Solution 17

Exercise 17

The **active set** in the context of the LASSO is the set of variables with non-zero coefficients in the LASSO solution. It represents the variables that the LASSO has selected as being important for predicting the outcome variable. Formally, $A = \{j : \beta_j \neq 0\}$.

Solution 18

Exercise 18

Under certain conditions (which are beyond the scope of the provided text but generally involve restrictions on the correlation between predictors and the signal strength of the true coefficients), the LASSO selects the correct model with probability tending to one as the sample size grows. That is, $\Pr(\hat{A} = A) \to 1$, where $\hat{A}$ is the estimated active set, and A is the true active set.

Solution 19

Exercise 19

If the bound s in the alternative formulation of the LASSO problem is very large, the constraint $\sum_{j=1}^{K} |b_j| \leq s$ becomes non-binding. In this case, the LASSO solution converges to the ordinary least squares (OLS) solution. This is because the constraint is no longer restricting the magnitude of the coefficients.

Solution 20

Exercise 20

Cross-validation is used in LASSO to select the optimal value of the tuning parameter λ (or equivalently, the bound s). The data is split into multiple folds. For each value of λ in a grid, the LASSO is trained on a subset of the folds and validated on the remaining fold. The value of λ that minimizes the average cross-validation error (e.g., mean squared error) is chosen as the optimal value. This helps prevent overfitting and provides a more reliable estimate of the model's performance on unseen data.

R Scripts

R Script 1: Omitted Variable Bias

```
# 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
```

```
library(MASS) # For multivariate normal simulation
```

```

Attaching package: 'MASS'

The following object is masked from 'package:dplyr':

    select
```

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

# Simulation parameters
n <- 100  # Sample size
beta1 <- 2 # True coefficient for x1
beta2 <- 3 # True coefficient for x2
sigma <- 1 # Error standard deviation

# Generate correlated x1 and x2
mu <- c(0, 0)
Sigma <- matrix(c(1, 0.5, 0.5, 1), nrow = 2) # Covariance matrix with correlation 0.5
X <- mvrnorm(n, mu, Sigma)
x1 <- X[, 1]
x2 <- X[, 2]

# Generate the dependent variable y
epsilon <- rnorm(n, 0, sigma)
y <- beta1 * x1 + beta2 * x2 + epsilon

# --- Model 1: Correctly specified model (includes x1 and x2) ---
df_correct <- data.frame(y, x1, x2)
model_correct <- lm(y ~ x1 + x2, data = df_correct)
summary(model_correct) # Observe estimates close to beta1 and beta2
```

```

Call:
lm(formula = y ~ x1 + x2, data = df_correct)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8730 -0.6607 -0.1245  0.6214  2.0798
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.13507    0.09614   1.405    0.163
x1           1.89930    0.11346  16.741   <2e-16 ***
x2           2.94692    0.11857  24.853   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9513 on 97 degrees of freedom
Multiple R-squared:  0.9433,    Adjusted R-squared:  0.9421
F-statistic: 806.3 on 2 and 97 DF,  p-value: < 2.2e-16
```

```
# --- Model 2: Misspecified model (omits x2) ---
df_misspecified <- data.frame(y, x1)
model_misspecified <- lm(y ~ x1, data = df_misspecified)
summary(model_misspecified) # Observe biased estimate of beta1
```

```
Call:
lm(formula = y ~ x1, data = df_misspecified)

Residuals:
    Min      1Q  Median      3Q     Max
-6.1108 -1.6878 -0.2221  1.6733  6.5244

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.03752    0.25941   0.145    0.885
x1           3.18502    0.27268  11.680   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.569 on 98 degrees of freedom
Multiple R-squared:  0.582, Adjusted R-squared:  0.5777
F-statistic: 136.4 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
# --- Visualization ---
# Plot y vs x1, coloring by x2 to illustrate the omitted variable effect
ggplot(df_correct, aes(x = x1, y = y, color = x2)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, linetype = "dashed") +
  labs(title = "Omitted Variable Bias Illustration",
       x = "x1",
       y = "y",
       color = "x2") +
  theme_bw()
```

```
`geom_smooth()` using formula = 'y ~ x'
```

```
Warning: The following aesthetics were dropped during statistical transformation:
colour.
ℹ This can happen when ggplot fails to infer the correct grouping structure in
  the data.
ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
  variable into a factor?
```

```
# --- Calculate the theoretical bias ---
# bias = (X1'X1)^(-1) X1'X2 * beta2
X1 <- matrix(x1, ncol = 1)
X2 <- matrix(x2, ncol = 1)
```

```
theoretical_bias <- solve(t(X1) %*% X1) %*% t(X1) %*% X2 * beta2
print(paste("Theoretical Bias:", theoretical_bias))
```

```
[1] "Theoretical Bias: 1.29438426797482"
```

```
print(paste("Estimated Coefficient of x1 (misspecified model):", coef(model_misspecified)[2]))
```

```
[1] "Estimated Coefficient of x1 (misspecified model): 3.18502020668086"
```

```
print(paste("Estimated Coefficient of x1 (correct model) + Theoretical Bias:" ,coef(model_correct)[2]
         + theoretical_bias ))
```

```
[1] "Estimated Coefficient of x1 (correct model) + Theoretical Bias: 3.19368636237153"
```

Explanation:

1. **Setup:**
   - Loads the `tidyverse` for data manipulation and visualization and `MASS` for generating multivariate normal data.
   - Sets a seed for reproducibility.
   - Defines simulation parameters: sample size (`n`), true coefficients (`beta1`, `beta2`), and error standard deviation (`sigma`).
2. **Data Generation:**
   - Creates correlated regressors `x1` and `x2` using `mvrnorm` from the `MASS` package. A correlation of 0.5 is introduced between x_1 and x_2.
   - Generates the dependent variable `y` according to the true model: $y = \beta_1 x_1 + \beta_2 x_2 + \epsilon$.
3. **Model Estimation:**
   - **Correct Model:** Creates a data frame `df_correct` and estimates the correct model using `lm(y ~ x1 + x2, data = df_correct)`. The `summary()` function provides estimates of β_1 and β_2, which should be close to their true values.
   - **Misspecified Model:** Creates a data frame `df_misspecified` (including only x_1) and estimates the misspecified model using `lm(y ~ x1, data = df_misspecified)`. The `summary()` function now provides a *biased* estimate of β_1. This illustrates **omitted variable bias**, as discussed in Section 21.1 of the text.
4. **Visualization:**
   - Creates a scatter plot of `y` versus `x1`, with points colored by the value of `x2`. This visually demonstrates how the omitted variable `x2` affects the relationship between `y` and `x1`. The dashed line shows the fitted relationship from the misspecified model.
5. **Theoretical Bias Calculation:**
   - Calculates the theoretical bias using the formula: $\text{bias} = (X_1^T X_1)^{-1} X_1^T X_2 \beta_2$.
   - Prints the theoretical bias.
   - Prints the estimated coefficient of x1 in misspecified model.
   - Prints the sum of the estimated coefficient for x_1 in the correct model and the theoretical bias, it should match closely the coefficient from the misspecified model. This concretely demonstrates the formula for omitted variable bias presented in the text.

R Script 2: Inclusion of Irrelevant Variables

```
# Load necessary libraries
library(tidyverse)

# Set seed for reproducibility
set.seed(456)

# Simulation parameters
```

```
n <- 100
beta1 <- 2
sigma <- 1

# Generate x1 (relevant) and x2 (irrelevant)
x1 <- rnorm(n)
x2 <- rnorm(n)  # x2 is independent of y

# Generate y
epsilon <- rnorm(n, 0, sigma)
y <- beta1 * x1 + epsilon

# --- Model 1: Correct model (includes only x1) ---
df_correct <- data.frame(y, x1)
model_correct <- lm(y ~ x1, data = df_correct)
summary(model_correct)
```

```

Call:
lm(formula = y ~ x1, data = df_correct)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.44544 -0.70333 -0.01888  0.59612  2.53030 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.08417    0.09601   0.877    0.383    
x1           2.11989    0.09565  22.163   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9532 on 98 degrees of freedom
Multiple R-squared:  0.8337,	Adjusted R-squared:  0.832 
F-statistic: 491.2 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
# --- Model 2: Includes irrelevant variable x2 ---
df_irrelevant <- data.frame(y, x1, x2)
model_irrelevant <- lm(y ~ x1 + x2, data = df_irrelevant)
summary(model_irrelevant)
```

```

Call:
lm(formula = y ~ x1 + x2, data = df_irrelevant)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.43186 -0.70864 -0.00242  0.59762  2.50212 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.08686    0.09715   0.894    0.373    
x1           2.12028    0.09613  22.057   <2e-16 ***
x2           0.02346    0.09910   0.237    0.813    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9578 on 97 degrees of freedom
Multiple R-squared:  0.8338,	Adjusted R-squared:  0.8303 
F-statistic: 243.3 on 2 and 97 DF,  p-value: < 2.2e-16
```

```
# --- Compare variances of beta1 estimates ---
var_beta1_correct <- vcov(model_correct)[2,2]  # Variance of beta1 in correct model
```

```
var_beta1_irrelevant <- vcov(model_irrelevant)[2,2] # Variance of beta1 with irrelevant variable

print(paste("Variance of beta1 (correct model):", var_beta1_correct))
```

```
[1] "Variance of beta1 (correct model): 0.00914900135959905"
```

```
print(paste("Variance of beta1 (irrelevant variable included):", var_beta1_irrelevant))
```

```
[1] "Variance of beta1 (irrelevant variable included): 0.00924066938296475"
```

```
# --- Visualization: Coefficient estimates and confidence intervals ---
coef_df <- data.frame(
  Model = c("Correct", "Irrelevant"),
  Estimate = c(coef(model_correct)[2], coef(model_irrelevant)[2]),
  Lower = c(confint(model_correct)[2,1], confint(model_irrelevant)[2,1]),
  Upper = c(confint(model_correct)[2,2], confint(model_irrelevant)[2,2])
)

ggplot(coef_df, aes(x = Model, y = Estimate, ymin = Lower, ymax = Upper)) +
  geom_pointrange() +
  labs(title = "Comparison of Coefficient Estimates and Confidence Intervals",
       y = "Coefficient Estimate (beta1)") +
  theme_bw()
```

Explanation:

1. **Setup:** Loads the `tidyverse` package, sets a seed, and defines simulation parameters.

2. **Data Generation:** Generates a relevant variable `x1` and an *irrelevant* variable `x2` (independent of `y`). `y` is generated based only on `x1` and the error term.

3. **Model Estimation:**

   - **Correct Model:** Estimates the model with only the relevant variable `x1`.
   - **Model with Irrelevant Variable:** Estimates the model including both `x1` and the irrelevant variable `x2`.

4. **Variance Comparison:**

   - Extracts the variance of the $\hat{\beta}_1$ estimates from both models using `vcov()`.
   - Prints the variances. The variance of $\hat{\beta}_1$ will be *larger* when the irrelevant variable `x2` is included, demonstrating the concept from Section 21.2 of the text that including irrelevant variables inflates the variance of the estimators.

5. **Visualization:** Creates a plot showing the point estimates and confidence intervals for β_1 from both models. The confidence interval will be wider for the model including the irrelevant variable, visually confirming the increased variance.

R Script 3: Model Selection with Adjusted R-squared

```
# Load necessary libraries
library(tidyverse)

# Set seed
set.seed(789)
```

```r
# Generate data with multiple potential predictors
n <- 100
k <- 5  # Number of potential predictors
X <- matrix(rnorm(n * k), ncol = k)
beta <- c(2, 0, 1.5, 0, 0)  # True coefficients: x1 and x3 are relevant, others are not
epsilon <- rnorm(n)
y <- X %*% beta + epsilon
df <- data.frame(y, X)
colnames(df) <- c("y", paste0("x", 1:k))

# --- Function to calculate adjusted R-squared ---
calculate_adj_r_squared <- function(model, data) {
  n <- nrow(data)
  k <- length(coef(model)) - 1 # Number of predictors (excluding intercept)
  r_squared <- summary(model)$r.squared
  adj_r_squared <- 1 - (1 - r_squared) * (n - 1) / (n - k - 1)
  return(adj_r_squared)
}

# --- Model selection using adjusted R-squared ---
# Consider all possible combinations of predictors
all_combinations <- unlist(lapply(1:k, function(i) combn(paste0("x", 1:k), i, simplify = FALSE)),
         recursive = FALSE)

results <- data.frame(Model = character(), Adj_R_squared = numeric())

for (combination in all_combinations) {
  formula_str <- paste("y ~", paste(combination, collapse = " + "))
  model <- lm(as.formula(formula_str), data = df)
  adj_r2 <- calculate_adj_r_squared(model, df)
  results <- rbind(results, data.frame(Model = formula_str, Adj_R_squared = adj_r2))
}

# Find the model with the highest adjusted R-squared
best_model <- results[which.max(results$Adj_R_squared), ]
print(best_model)
```

```
              Model Adj_R_squared
20 y ~ x1 + x3 + x5     0.8980251
```

```r
# --- Visualization: Adjusted R-squared for different models ---
ggplot(results, aes(x = reorder(Model, Adj_R_squared), y = Adj_R_squared)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Adjusted R-squared for Different Models",
       x = "Model",
       y = "Adjusted R-squared") +
  theme_bw()
```

Explanation:

1. **Setup:** Loads `tidyverse`, sets a seed, and generates data with multiple potential predictors (`x1` through `x5`). The true coefficients are set so that only `x1` and `x3` are truly relevant.

2. **`calculate_adj_r_squared` Function:** Defines a function to calculate the adjusted R^2 given a model and data. This implements the formula from the text (Equation 21.1).

3. **Model Selection:**
    - `all_combinations`: Generates all possible combinations of predictors (from single predictors up to all five predictors). `combn` function helps to get combinations, `lapply` applies this for each possible number of variables and `unlist(,recursive = FALSE)` makes a flat list from a list of lists.
    - Iterates through each combination:
        - Constructs the model formula string (e.g., “y ~ x1 + x3”).
        - Fits the linear model using `lm()`.
        - Calculates the adjusted R^2 using the defined function.
        - Stores the model formula and adjusted R^2 in the `results` data frame.
    - Finds the model with the maximum adjusted R^2 using `which.max()`.
    - Prints the best model. This demonstrates the model selection process based on adjusted R^2, as described in Section 21.3.
4. **Visualization:** Creates a bar plot showing the adjusted R^2 for each model, sorted in ascending order. This makes it easy to visually identify the best model.

R Script 4: Model Selection with AIC and BIC

```
# Load necessary libraries
library(tidyverse)

# Use the same data generation as in Script 3
# (Re-run the data generation part of Script 3 here if not already in your environment)
# Set seed
set.seed(789)

# Generate data with multiple potential predictors
n <- 100
k <- 5  # Number of potential predictors
X <- matrix(rnorm(n * k), ncol = k)
beta <- c(2, 0, 1.5, 0, 0)  # True coefficients: x1 and x3 are relevant, others are not
epsilon <- rnorm(n)
y <- X %*% beta + epsilon
df <- data.frame(y, X)
colnames(df) <- c("y", paste0("x", 1:k))

# --- Model selection using AIC and BIC ---
all_combinations <- unlist(lapply(1:k, function(i) combn(paste0("x", 1:k), i, simplify = FALSE)),
         recursive = FALSE)

results <- data.frame(Model = character(), AIC = numeric(), BIC = numeric())

for (combination in all_combinations) {
  formula_str <- paste("y ~", paste(combination, collapse = " + "))
  model <- lm(as.formula(formula_str), data = df)
  aic_val <- AIC(model)
  bic_val <- BIC(model)
  results <- rbind(results, data.frame(Model = formula_str, AIC = aic_val, BIC = bic_val))
}

# Find the best model according to AIC
best_model_aic <- results[which.min(results$AIC), ]
print("Best model according to AIC:")
```

```
[1] "Best model according to AIC:"
```

```
print(best_model_aic)
```

```
              Model      AIC      BIC
20 y ~ x1 + x3 + x5 258.1546 271.1805
```

```
# Find the best model according to BIC
best_model_bic <- results[which.min(results$BIC), ]
print("Best model according to BIC:")
```

```
[1] "Best model according to BIC:"
```

```
print(best_model_bic)
```

```
        Model      AIC      BIC
7 y ~ x1 + x3 258.4832 268.9039
```

```
# --- Visualization: AIC and BIC for different models ---
results_long <- results %>%
  pivot_longer(cols = c(AIC, BIC), names_to = "Criterion", values_to = "Value")

ggplot(results_long, aes(x = reorder(Model, Value), y = Value, fill = Criterion)) +
  geom_bar(stat = "identity", position = "dodge") +
  coord_flip() +
  labs(title = "AIC and BIC for Different Models",
       x = "Model",
       y = "Criterion Value") +
  theme_bw()
```

Explanation:

1. **Data:** Uses the same data generation process as R Script 3.

2. **Model Selection:**

    - Similar to Script 3, generates all possible combinations of predictors.
    - Iterates through each combination:
        - Constructs the model formula.
        - Fits the linear model.
        - Calculates AIC and BIC using the built-in `AIC()` and `BIC()` functions. These functions implement the formulas from the text (Equations 21.3 and 21.4).
        - Stores the model formula, AIC, and BIC in the `results` data frame.
    - Finds the best model according to AIC (minimum AIC).
    - Finds the best model according to BIC (minimum BIC).
    - Prints the best models. Note that AIC and BIC may select different models. BIC will tend to favor simpler models.

3. **Visualization:** Uses `pivot_longer` from `tidyr` to reshape the data for easier plotting. Creates a bar plot showing both AIC and BIC for each model. This allows for a visual comparison of the models based on both criteria.

R Script 5: LASSO Regression

```
# Load necessary libraries
library(tidyverse)
library(glmnet)  # For LASSO regression
```

```
Loading required package: Matrix
```

```

Attaching package: 'Matrix'

The following objects are masked from 'package:tidyr':

    expand, pack, unpack

Loaded glmnet 4.1-8
```

```
# Set seed
set.seed(1011)

# Generate data with many predictors, but only a few are relevant
n <- 100
k <- 20 # Number of predictors
X <- matrix(rnorm(n * k), ncol = k)
beta <- c(rep(2, 5), rep(0, k - 5))  # First 5 predictors are relevant, rest are irrelevant
epsilon <- rnorm(n)
y <- X %*% beta + epsilon

# --- LASSO Regression with Cross-Validation ---

# Create a data frame for glmnet
df <- data.frame(y,X)

# glmnet requires a matrix for X and a vector for y
x_matrix <- model.matrix(y ~ ., df)[,-1] # Remove intercept column
y_vector <- df$y
# Perform cross-validation to find the optimal lambda
cv_model <- cv.glmnet(x_matrix, y_vector, alpha = 1)  # alpha = 1 for LASSO

# Find the best lambda
best_lambda <- cv_model$lambda.min
print(paste("Best lambda:", best_lambda))
```

```
[1] "Best lambda: 0.103203794548873"
```

```
# Fit the LASSO model with the best lambda
lasso_model <- glmnet(x_matrix, y_vector, alpha = 1, lambda = best_lambda)

# Extract the coefficients
coefficients <- coef(lasso_model)
print(coefficients)
```

```
21 x 1 sparse Matrix of class "dgCMatrix"
                     s0
(Intercept)  0.05261701
X1           2.12012238
X2           1.94425413
X3           2.00064168
X4           1.91416014
X5           1.82219096
X6           0.08016068
X7           0.01767611
X8          -0.12686034
X9           .
X10          .
X11          .
X12          .
X13          .
X14          .
X15          .
X16          .
```

```
X17          0.01608481
X18          .
X19          .
X20          .

# --- Visualization: Coefficient Path ---

# Fit LASSO for a range of lambda values (without cross-validation, for visualization)
lasso_path <- glmnet(x_matrix, y_vector, alpha = 1)

# Plot the coefficient path
plot(lasso_path, xvar = "lambda", label = TRUE)
title("LASSO Coefficient Path", line = 2.5)
```

```
# --- Visualization: Cross-validation Error ---
plot(cv_model)
title("Cross-Validation Error for LASSO", line=2.5)
```

Explanation:

1. **Setup:** Loads `tidyverse` and `glmnet`, sets a seed, and generates data with many predictors (`k = 20`), but only the first five have non-zero coefficients (illustrating **sparsity**).

2. **Data Preparation for `glmnet`:**

   - `glmnet` requires the input data to be in a specific format: a matrix for the predictors (`x_matrix`) and a vector for the response (`y_vector`). `model.matrix` is used to create dummy variable and remove intercept.

3. **LASSO with Cross-Validation:**

   - `cv.glmnet(x_matrix, y_vector, alpha = 1)` performs cross-validation to find the optimal value of the regularization parameter λ. `alpha = 1` specifies LASSO regression (as opposed to Ridge regression or elastic net).
   - `cv_model$lambda.min` extracts the value of λ that minimizes the cross-validation error.
   - `glmnet(x_matrix, y_vector, alpha = 1, lambda = best_lambda)` fits the final LASSO model using the best λ.

4. **Coefficient Extraction:** `coef(lasso_model)` extracts the estimated coefficients from the fitted LASSO model. Many of these coefficients will be exactly zero, demonstrating LASSO’s **variable selection** property.

5. **Visualization:**

   - **Coefficient Path:** `glmnet(x_matrix, y_vector, alpha = 1)` (without specifying `lambda`) fits the LASSO model for a *range* of λ values. `plot(lasso_path, xvar = "lambda", label = TRUE)` creates a “coefficient path” plot. This plot shows how the estimated coefficients change as λ changes. As λ increases (moving from right to left on the plot), more coefficients are shrunk to zero.
   - **Cross-Validation Error:** `plot(cv_model)` plots the cross-validation error as a function of λ. This helps to visualize how the choice of λ affects the model’s performance. The minimum point on this curve corresponds to `cv_model$lambda.min`.

These scripts illustrate the key concepts from Chapter 21 of the text, using simulations, model fitting, and visualizations to provide a practical understanding of omitted variable bias, inclusion of irrelevant variables, model selection techniques, and LASSO regression.

YouTube Videos on Econometrics Concepts

Here are some YouTube videos that explain the concepts discussed in Chapter 21, along with explanations of how they relate to the text:

Omitted Variable Bias

1. **Title:** "Omitted Variable Bias Explained" **Channel:** Ben Lambert **Link:** https://www.youtube.com/watch?v=FW-v5HXYv_M **Relevance:** This video provides a clear and intuitive explanation of **omitted variable bias (OVB)**. It covers the core concepts presented in Section 21.1 of the text, including:
   - The conditions under which OVB occurs (correlation between the omitted variable and both the dependent and included independent variables).
   - The direction and magnitude of the bias.
   - The mathematical derivation of the bias, similar to the derivation presented in the text ($\beta_{12} = (X_1^T X_1)^{-1} X_1^T X_2 \beta_2$).
   - The use of directed acyclic graphs (DAGs) to visualize the relationships between variables, which is helpful, although not directly in the text.
2. **Title:** "Omitted Variable Bias: An Introduction" **Channel:** A Crash Course in Causality **Link:** https://www.youtube.com/watch?v=oPXeWfIuJs0 **Relevance:** This video provides another excellent explanation of OVB, covering similar ground to the Ben Lambert video but with slightly different examples. It reinforces the core concepts of Section 21.1. It does a good job in connecting the bias in the coefficient with a bias in estimating causal effects.

Inclusion of Irrelevant Variables

1. **Title:** "Irrelevant variables in OLS" **Channel:** Economics and Guitars **Link:** https://www.youtube.com/watch?v=G3YD7-b5jO8 **Relevance:** This short video directly addresses the consequences of including irrelevant variables in a regression model. It aligns with Section 21.2 of the text, particularly the points that:
   - The OLS estimator remains *unbiased*.
   - The *variance* of the estimator *increases* when irrelevant variables are included.
   - It does the demonstration of the matrix algebra proof in the case where a single extra variable is added.

Model Selection (AIC, BIC)

1. **Title:** "AIC, BIC, and Model Selection in R" **Channel:** Quant Psych **Link:** https://www.youtube.com/watch?v=i-e6g_SzJ90 **Relevance:** This video explains and demonstrates the use of the **Akaike Information Criterion (AIC)** and the **Bayesian Information Criterion (BIC)** for model selection, specifically in the context of R. This directly relates to Section 21.3 of the text, where AIC and BIC are introduced as model selection criteria (Equations 21.3 and 21.4). The video covers:
   - The formulas for AIC and BIC.
   - The interpretation of AIC and BIC values (lower is better).
   - How to use AIC and BIC to compare different models.
   - Practical implementation in R.
2. **Title**: StatQuest: AIC and BIC **Channel**: StatQuest with Josh Starmer **Link**: https://www.youtube.com/watch?v=5-OM5yFAGaA **Relevance**: Very clear and conceptual explanation of the use of AIC and BIC for model selection. Discusses the underlying likelihood considerations.

LASSO Regression

1. **Title:** "StatQuest: Lasso Regression, Clearly Explained!!!" **Channel:** StatQuest with Josh Starmer **Link:** https://www.youtube.com/watch?v=NGf0voTMlcs **Relevance:** This video provides a very clear, intuitive explanation of **LASSO regression**. It aligns perfectly with Section 21.4 of the text, covering:
   - The motivation for LASSO (dealing with a large number of predictors).
   - The L_1 penalty and how it leads to variable selection (setting coefficients to exactly zero).
   - The concept of the tuning parameter λ.
   - A visual explanation of how LASSO works geometrically.
   - The connection between LASSO and other methods (like best subset selection).
2. **Title:** "Lasso Regression - a Lasso in a Haystack | SciPy 2019 Tutorial | Gaël Varoquaux" **Channel:** Enthought **Link:** https://www.youtube.com/watch?v=74-SNkbp_6c **Relevance**: Although this video uses Python instead of R, it is very valuable to understand LASSO. It covers:
   - The motivation of the Lasso method.
   - The concept of sparsity.
   - The meaning of the parameter s that appears in the alternative formulation of the LASSO method.
3. **Title**: 17.1 Lasso method - an introduction **Channel**: mathematicalmonk **Link**: https://www.youtube.com/watch?v=2P-e-AGv4DU **Relevance**: Introduction to the lasso method and motivation, using mathematical notation, similar to the text.

These videos provide a mix of theoretical explanations, visual aids, and practical demonstrations (using R for AIC/BIC) that complement the material presented in Chapter 21. They are all currently available on YouTube at the provided links (verified on October 26, 2023).

Multiple Choice Exercises

MC Exercise 1

MC Solution 1

Omitted variable bias occurs when:

a. An irrelevant variable is included in the regression.
b. A relevant variable is excluded from the regression, and it is correlated with an included regressor.
c. The error term is heteroskedastic.
d. The sample size is too small.

MC Exercise 2

MC Solution 2

The direction of omitted variable bias depends on:

a. The sample size.
b. The variance of the error term.
c. The sign of the coefficient of the omitted variable and the correlation between the omitted and included variables.
d. Whether the model includes an intercept.

MC Exercise 3

MC Solution 3

If you omit a relevant variable from a regression, the OLS estimator of the coefficient of an included variable is generally:

a. Unbiased.
b. Biased and inconsistent.
c. Biased but consistent.
d. Efficient.

MC Exercise 4

MC Solution 4

Including irrelevant variables in a regression model will:

a. Decrease the variance of the OLS estimators.
b. Increase the variance of the OLS estimators.
c. Bias the OLS estimators.
d. Improve the adjusted R^2.

MC Exercise 5

MC Solution 5

When irrelevant variables are included in a regression, the OLS estimators of the coefficients of the *relevant* variables are:

a. Biased.
b. Unbiased.
c. Inconsistent.
d. Biased and Inconsistent.

MC Exercise 6

MC Solution 6

The adjusted R^2:

a. Always increases when a new variable is added to the model.
b. Can be negative.
c. Penalizes the inclusion of additional variables.
d. Is always larger than the unadjusted R^2.

MC Exercise 7

MC Solution 7

Which of the following model selection criteria imposes the largest penalty for adding more variables?

a. R^2
b. Adjusted R^2
c. AIC
d. BIC

MC Exercise 8

MC Solution 8

The Akaike Information Criterion (AIC) is defined as:

a. $\ln(\frac{\hat{\epsilon}_j{}^T \hat{\epsilon}_j}{n}) + \frac{K_j}{n}$
b. $\ln(\frac{\hat{\epsilon}_j{}^T \hat{\epsilon}_j}{n}) + \frac{2K_j}{n}$
c. $\frac{\hat{\epsilon}_j{}^T \hat{\epsilon}_j}{n-K_j}$
d. $\frac{\hat{\epsilon}_j{}^T \hat{\epsilon}_j}{n}$

MC Exercise 9

MC Solution 9

The Bayesian Information Criterion (BIC) is defined as:

a. $\ln(\frac{\hat{\epsilon}_j{}^T \hat{\epsilon}_j}{n}) + \frac{K_j}{n}$
b. $\ln(\frac{\hat{\epsilon}_j{}^T \hat{\epsilon}_j}{n}) + \frac{2K_j}{n}$
c. $\ln(\frac{\hat{\epsilon}_j{}^T \hat{\epsilon}_j}{n}) + \frac{K_j \ln(n)}{n}$
d. $\frac{\hat{\epsilon}_j{}^T \hat{\epsilon}_j}{n}$

MC Exercise 10

MC Solution 10

When using AIC or BIC for model selection, you should choose the model with the:

a. Highest AIC and highest BIC.
b. Lowest AIC and lowest BIC.
c. Highest AIC and lowest BIC.
d. Lowest AIC and highest BIC.

MC Exercise 11

MC Solution 11

The LASSO method is used for:

a. Estimating coefficients in a linear regression.
b. Variable selection.
c. Dealing with multicollinearity.
d. All of the above.

MC Exercise 12

MC Solution 12

The LASSO uses which type of penalty?

a. L_0 penalty
b. L_1 penalty
c. L_2 penalty
d. No penalty

MC Exercise 13

MC Solution 13

Ridge regression uses which type of penalty?

a. L_0 penalty
b. L_1 penalty
c. L_2 penalty
d. No penalty

MC Exercise 14

MC Solution 14

As the tuning parameter λ in LASSO increases, the number of non-zero coefficients in the model generally:

a. Increases.
b. Decreases.
c. Stays the same.
d. Becomes equal to the sample size.

MC Exercise 15

MC Solution 15

The term “sparsity” in the context of LASSO refers to:

a. A large number of non-zero coefficients.
b. A small number of non-zero coefficients.
c. A large sample size.
d. A small sample size.

MC Exercise 16

MC Solution 16

In the alternative formulation of the LASSO problem, what happens as the bound ‘s’ increases?

a. More variables enter the model.
b. Fewer variables enter the model.
c. The solution approaches the ridge regression solution.
d. The solution approaches the ordinary least squares solution.

MC Exercise 17

MC Solution 17

Cross-validation in the context of LASSO is used to:

a. Estimate the coefficients.
b. Choose the optimal value of the tuning parameter.
c. Calculate the standard errors.
d. Test for heteroskedasticity.

MC Exercise 18

MC Solution 18

Theorem 21.1 states that the adjusted R-squared falls when a variable is deleted if its t-statistic is:

a. Greater than 1 in absolute value
b. Less than 1 in absolute value
c. Greater than 2 in absolute value
d. Less than 2 in absolute value.

MC Exercise 19

MC Solution 19

When omitting a relevant variable, the usual OLS estimator of the error variance is:

a. Unbiased
b. Biased downwards
c. Biased upwards
d. Consistent

MC Exercise 20

MC Solution 20

The active set in the LASSO method is defined as:

a. The set of all variables in the model.
b. The set of variables with non-zero coefficients.
c. The set of variables with zero coefficients.
d. The set of variables used for cross-validation.

Multiple Choice Solutions

MC Solution 1

MC Exercise 1

Correct Answer: b)

Explanation: Omitted variable bias arises when a variable that is both *relevant* (affects the dependent variable) and *correlated with an included regressor* is left out of the regression. This is the core concept of Section 21.1.

MC Solution 2

MC Exercise 2

Correct Answer: c)

Explanation: The direction (positive or negative) of the omitted variable bias depends on two factors: (1) the sign of the coefficient (β_2) of the omitted variable in the true model, and (2) the sign of the correlation between the omitted variable and the included regressor(s). This is discussed in Section 21.1.

MC Solution 3

MC Exercise 3

Correct Answer: b)

Explanation: Omitting a relevant variable generally makes the OLS estimators of the included variables both biased (not centered on the true value) and inconsistent (does not converge to the true value as the sample size increases). This is a key result from Section 21.1.

MC Solution 4

MC Exercise 4

Correct Answer: b)

Explanation: Including irrelevant variables *increases* the variance of the OLS estimators. While the estimators remain unbiased, they become less precise. This is discussed in Section 21.2.

MC Solution 5

MC Exercise 5

Correct Answer: b)

Explanation: A crucial result from Section 21.2 is that including irrelevant variables does *not* bias the OLS estimators of the coefficients of the *relevant* variables. The estimators remain unbiased.

MC Solution 6

MC Exercise 6

Correct Answer: c)

Explanation: The adjusted R^2 is a modified version of R^2 that penalizes the inclusion of additional variables. It accounts for the degrees of freedom used by the model. This is explained in Section 21.3.

MC Solution 7

MC Exercise 7

Correct Answer: d)

Explanation: The Bayesian Information Criterion (BIC) imposes a larger penalty for additional variables than AIC or adjusted R^2 (especially for larger sample sizes). This is because the BIC's penalty term includes $\ln(n)$, where n is the sample size. This is discussed in Section 21.3.

MC Solution 8

MC Exercise 8

Correct Answer: b)

Explanation: The AIC is defined as $\ln(\frac{\hat{\epsilon_j}^T\hat{\epsilon_j}}{n}) + \frac{2K_j}{n}$, where K_j is the number of parameters in model j. This is Equation 21.3 in the text.

MC Solution 9

MC Exercise 9

Correct Answer: c)

Explanation: The BIC is defined as $\ln(\frac{\hat{\epsilon_j}^T\hat{\epsilon_j}}{n}) + \frac{K_j \ln(n)}{n}$. This is Equation 21.4 in the text.

MC Solution 10

MC Exercise 10

Correct Answer: b)

Explanation: For both AIC and BIC, lower values indicate a better model fit, taking into account model complexity. So you want the model with the lowest AIC and the lowest BIC.

MC Solution 11

MC Exercise 11

Correct Answer: d)

Explanation: The LASSO (Least Absolute Shrinkage and Selection Operator) is a method that does all of these: estimates coefficients, performs variable selection by shrinking some coefficients to exactly zero, and helps with multicollinearity by reducing the impact of highly correlated predictors. This corresponds to Section 21.4.

MC Solution 12

MC Exercise 12

Correct Answer: b)

Explanation: LASSO uses an L_1 penalty, which is the sum of the absolute values of the coefficients: $\lambda \sum_{j=1}^{K} |b_j|$. This is the defining characteristic of LASSO, as explained in Section 21.4.

MC Solution 13

MC Exercise 13

Correct Answer: c)

Explanation: Ridge regression, mentioned as a precursor to LASSO, uses an L_2 penalty which involves the sum of *squared* coefficients: $\lambda \sum_{j=1}^{K} b_j^2$.

MC Solution 14

MC Exercise 14

Correct Answer: b)

Explanation: As the tuning parameter λ in LASSO increases, the penalty on the magnitude of the coefficients becomes stronger. This forces more coefficients to be shrunk to zero, thus *decreasing* the number of non-zero coefficients. This is how LASSO performs variable selection.

MC Solution 15

MC Exercise 15

Correct Answer: b)

Explanation: Sparsity, in the context of LASSO and other regularization methods, means that only a *small number* of the potential predictors have non-zero coefficients in the true model (or in the estimated model).

MC Solution 16

MC Exercise 16

Correct Answer: d)

Explanation: The alternative formulation of the LASSO problem minimizes the residual sum of squares subject to the constraint $\sum_{j=1}^{K} |b_j| \leq s$. As ‘s’ increases, the constraint becomes less restrictive. When ‘s’ is large enough, the constraint is no longer binding, and the LASSO solution becomes equivalent to the ordinary least squares (OLS) solution.

MC Solution 17

MC Exercise 17

Correct Answer: b)

Explanation: Cross-validation is a technique used to estimate the prediction error of a model and to choose the optimal value of the tuning parameter (λ in LASSO) that minimizes this error.

MC Solution 18

MC Exercise 18

Correct Answer: b)

Explanation: Theorem 21.1 states that the adjusted R^2 will *fall* if a variable is deleted and its t-statistic is *greater* than 1 in absolute value. Conversely, the adjusted R^2 will rise if a variable is added and its t-statistic is less than one in absolute value.

MC Solution 19

MC Exercise 19

Correct Answer: c)

Explanation: When a relevant variable is omitted, the usual OLS estimator of the error variance (s_u^2) is biased *upwards*. This is because the variation explained by the omitted variable is incorrectly attributed to the error term. This is shown mathematically in Section 21.1.

MC Solution 20

MC Exercise 20

Correct Answer: b)

Explanation: The active set in LASSO is the set of variables that have *non-zero* coefficients in the LASSO solution. These are the variables that LASSO has deemed important for predicting the outcome.

Author: Peter Fuleky

This book was built with Quarto