Chapter 16

Multiple Linear Regression

Detailed Solutions

16.1 Basic Concept

16.1 The "Ceteris Paribus" Concept

Problem: Interpret b_1 in $\hat{Y} = b_0 + b_1X_1 + b_2X_2$.
Solution:

(a) **Interpretation:** The coefficient b_1 represents the estimated change in the mean response $\hat{Y}$ for a one-unit increase in X_1, **holding X_2 constant** (Ceteris Paribus). It isolates the effect of X_1 from X_2.

(b) **Difference from Simple Regression:** In simple regression (Y vs X_1), if X_1 is correlated with X_2, the coefficient b_1 absorbs some of the effect of X_2 (Omitted Variable Bias). In Multiple Regression, we explicitly control for X_2, so b_1 reflects the "pure" effect of X_1 (assuming no other omitted variables).

(a) Change in Y per unit X1, holding X2 constant. (b) Multiple regression removes bias from X2.

16.2 R-squared vs. Adjusted R-squared

Problem: Comparing metrics.
Solution:

(a) **Why Adjusted?** Standard R^2 never decreases when you add variables, even junk ones. It rewards complexity. Adjusted R^2 penalizes the model for adding useless variables, providing a fairer comparison between models of different sizes.

(b) **Adding "Shoe Size" (R^2):** The standard R^2 will **increase slightly** (or stay exactly the same), simply because the model can fit the random noise slightly better. It will never decrease.

(c) **Adding "Shoe Size" (Adj R^2):** The Adjusted R^2 will likely **decrease**. The penalty for adding a variable (loss of degree of freedom) outweighs the tiny, non-existent improvement in fit.

(b) R^2 Increases/Stays same.
(c) Adj R^2 Decreases.

16.2 Intermediate

16.3 Real Estate Price Prediction

Problem: Y vs Size(X_1), Age(X_2), Dist(X_3).
Solution:

(a) **Model Equation:** From Coefficients table: Intercept=50, Size=0.15, Age=-2.00, Dist=-1.50.

$$\hat{Y} = 50 + 0.15X_1 - 2.00X_2 - 1.50X_3$$

(b) **Global F-Test:** Hypothesis: $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$. Look at "Significance F" = 0.0000. Since $0.0000 < 0.05$, we **Reject** H_0. At least one variable is significant; the model is useful.

(c) **Individual T-Tests:** Compare P-values with 0.05:

- Size ($P = 0.000$): Significant.
- Age ($P = 0.0004$): Significant.
- Distance ($P = 0.145$): **Not Significant**.

(d) **Interpretation of Age (X_2):** Coefficient is -2.00. For every additional year of age, the house price decreases by $2,000 (assuming unit is thousands), **holding Size and Distance constant**.

(e) **Prediction:** $X_1 = 2000, X_2 = 10, X_3 = 5$.

$$\begin{aligned}\hat{Y} &= 50 + 0.15(2000) - 2.00(10) - 1.50(5)\\ &= 50 + 300 - 20 - 7.5\\ &= 350 - 27.5\\ &= 322.5\end{aligned}$$

Predicted Price: 322.5 (unit).

(a) $\hat{Y} = 50 + 0.15X_1 - 2X_2 - 1.5X_3$
(b) Significant Model.
(c) Distance is not significant.
(e) 322.5

16.3 Challenge

16.4 Interaction Effects (Visualized)

Problem: $\hat{Y} = 30 + 2X_1 + 10X_2 + 1.5(X_1X_2)$. X_1 (Exp), X_2 (Edu: 0=HS, 1=PhD).
Solution:

(a) **Visual Check:** The lines diverge (spread apart) as Experience (X_1) increases.

- **Parallel lines** = No Interaction.
- **Diverging lines** = Significant Interaction.

This implies the return on experience is **higher** for PhDs than for High School grads. The gap widens over time.

(b) **Slope Calculation:**

- **High School ($X_2 = 0$):**

$$\begin{aligned}\hat{Y} &= 30 + 2X_1 + 10(0) + 1.5(X_1 \cdot 0)\\ &= 30 + \mathbf{2}X_1\end{aligned}$$

Slope = 2.

- **PhD ($X_2 = 1$):**

$$\begin{aligned}\hat{Y} &= 30 + 2X_1 + 10(1) + 1.5(X_1 \cdot 1)\\ &= 30 + 10 + (2 + 1.5)X_1\\ &= 40 + \mathbf{3.5}X_1\end{aligned}$$

Slope = 3.5.

(c) **Interpretation:** No, the value is not the same. For HS, 1 year exp adds 2 units of salary. For PhD, 1 year exp adds 3.5 units of salary. The interaction coefficient (+1.5) is the **extra boost** in slope gained by having a PhD.

(a) Diverging lines $\implies$ Interaction.
(b) HS Slope = 2, PhD Slope = 3.5.
(c) Value of experience depends on education.

16.5 Multicollinearity

Problem: X_1 (kg) and X_2 (lbs).
Solution:

(a) **Correlation:** 1 kg $\approx$ 2.2 lbs. $X_2 = 2.2X_1$. The correlation is exactly **1.0** (Perfect Multicollinearity).

(b) **Calculation Problem:** Regression coefficients are calculated using $(X^T X)^{-1}$. If variables are perfectly correlated, the matrix $(X^T X)$ becomes **Singular** (Determinant = 0) and cannot be inverted. The coefficients are undefined or unstable.

(c) **P-values Effect:** Even if correlation is high but not perfect (e.g., 0.99), the standard errors of the coefficients inflate massively. $t = b/SE$. If SE is huge, t becomes small, and P-value becomes large (Insignificant). **Result:** The F-test says the *whole model* is great (Significant), but individual T-tests say *neither variable* is significant. They "steal" each other's significance.

(a) $r = 1.0$
(b) Matrix non-invertible.
(c) Inflated SE $\rightarrow$ Insignificant P-values.