Multiple Linear Regression: Theory, Geometry, Inference, Diagnostics, and Fully Solved MSc-Level Assignment

Regression Techniques Course

Abstract

This handout presents a rigorous and conceptually clear account of multiple linear regression for MSc students. The purpose is not merely to teach the formula $\widehat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \boldsymbol{y}$, but to explain why the formula arises, what it means geometrically, when it is valid, how inference is performed, and how regression should be interpreted in practice. Topics include matrix formulation, least squares derivation, projection geometry, Gauss–Markov theorem, sampling distribution under normality, confidence intervals, prediction intervals, ANOVA decomposition, partial F-tests, multicollinearity, diagnostics, leverage, residuals, and model interpretation. A detailed home assignment with fully worked solutions is included.

Contents

1 Learning Objectives

After studying this handout, a student should be able to:

1. write the multiple linear regression model in scalar and matrix form;
2. derive the ordinary least squares estimator rigorously;
3. interpret regression coefficients as partial effects;
4. understand fitted values as projections;
5. derive the variance of the OLS estimator;
6. explain the Gauss–Markov theorem;
7. perform t-tests, F-tests, confidence intervals, and prediction intervals;
8. understand ANOVA decomposition in regression;
9. diagnose multicollinearity, leverage, and residual problems;
10. solve standard MSc-level theoretical and numerical problems in multiple regression.

2 From Simple to Multiple Linear Regression

In simple linear regression, we study the relationship between one response variable Y and one predictor X:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

In multiple linear regression, we allow several predictors:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \qquad i = 1, \ldots, n.$$

Here:

$$Y_i = \text{response for unit } i,$$

$$x_{ij} = \text{value of predictor } j \text{ for unit } i,$$

$$\beta_0 = \text{intercept},$$

$$\beta_j = \text{partial regression coefficient for predictor } j,$$

$$\varepsilon_i = \text{random error}.$$

Key conceptual shift

In simple regression, the slope measures the association between Y and one predictor. In multiple regression, β_j measures the effect of x_j on Y after adjusting for all other predictors in the model.

Thus, β_j is not merely a marginal association. It is a conditional or partial effect.

3 Matrix Form of the Model

Let

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

The design matrix is

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}.$$

Then the model is compactly written as

$$\boxed{\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}.}$$

The matrix X has n rows and $p+1$ columns. The first column is usually the intercept column $\mathbf{1}_n$.

4 Classical Assumptions

The usual fixed-design multiple linear regression assumptions are:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0},$$
$$\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n,$$

and X is treated as fixed.

Often, for exact finite-sample inference, we additionally assume

$$\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 I_n).$$

Meaning of the assumptions

1. $\mathbb{E}(\varepsilon_i) = 0$: the model is correctly centered.
2. $\mathrm{Var}(\varepsilon_i) = \sigma^2$: errors have common variance.
3. $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$: errors are uncorrelated.
4. Normality is not required for least squares estimation, but it is used for exact t- and F-tests.

5 Ordinary Least Squares Estimation

The ordinary least squares estimator minimizes the residual sum of squares:

$$\mathrm{RSS}(\boldsymbol{\beta}) = \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2.$$

Expanding,

$$\mathrm{RSS}(\boldsymbol{\beta}) = (\boldsymbol{y} - X\boldsymbol{\beta})^\top (\boldsymbol{y} - X\boldsymbol{\beta}).$$

Therefore,

$$\mathrm{RSS}(\boldsymbol{\beta}) = \boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{\beta}^\top X^\top \boldsymbol{y} + \boldsymbol{\beta}^\top X^\top X\boldsymbol{\beta}.$$

Differentiate with respect to $\boldsymbol{\beta}$:

$$\frac{\partial \,\mathrm{RSS}}{\partial \boldsymbol{\beta}} = -2X^\top \boldsymbol{y} + 2X^\top X\boldsymbol{\beta}.$$

Setting the derivative equal to zero gives the normal equations:

$$\boxed{X^\top X\widehat{\boldsymbol{\beta}} = X^\top \boldsymbol{y}.}$$

If $X^\top X$ is nonsingular, then

$$\boxed{\widehat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \boldsymbol{y}.}$$

Theorem 1 (Existence and uniqueness of OLS)**.** *If* $\mathrm{rank}(X) = p+1$*, then* $X^\top X$ *is nonsingular and the least squares estimator is unique:*

$$\widehat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \boldsymbol{y}.$$

Proof. For any nonzero $\boldsymbol{a} \in \mathbb{R}^{p+1}$,

$$\boldsymbol{a}^\top X^\top X\boldsymbol{a} = (X\boldsymbol{a})^\top (X\boldsymbol{a}) = \|X\boldsymbol{a}\|_2^2.$$

If $\mathrm{rank}(X) = p+1$, then $X\boldsymbol{a} \neq 0$ for all $\boldsymbol{a} \neq 0$. Hence

$$\|X\boldsymbol{a}\|_2^2 > 0.$$

Therefore, $X^\top X$ is positive definite and hence nonsingular. The RSS is a strictly convex quadratic function of $\boldsymbol{\beta}$, so the minimizer is unique. □

6 Fitted Values, Residuals, and the Hat Matrix

The fitted values are

$$\widehat{\boldsymbol{y}} = X\widehat{\boldsymbol{\beta}}.$$

Substituting the OLS estimator,

$$\widehat{\boldsymbol{y}} = X(X^\top X)^{-1} X^\top \boldsymbol{y}.$$

Define the hat matrix:

$$\boxed{H = X(X^\top X)^{-1} X^\top.}$$

Then

$$\boxed{\widehat{\boldsymbol{y}} = H\boldsymbol{y}.}$$

The residual vector is

$$\boldsymbol{e} = \boldsymbol{y} - \widehat{\boldsymbol{y}}.$$

Since $\widehat{\boldsymbol{y}} = H\boldsymbol{y}$,

$$\boxed{\boldsymbol{e} = (I_n - H)\boldsymbol{y}.}$$

Theorem 2 (Properties of the hat matrix)**.** *The matrix $H = X(X^\top X)^{-1}X^\top$ is symmetric and idempotent:*

$$H^\top = H, \qquad H^2 = H.$$

Proof. First,

$$H^\top = \left\{X(X^\top X)^{-1}X^\top\right\}^\top = X\left\{(X^\top X)^{-1}\right\}^\top X^\top.$$

Since $X^\top X$ is symmetric, its inverse is symmetric. Hence

$$H^\top = X(X^\top X)^{-1}X^\top = H.$$

Next,

$$H^2 = X(X^\top X)^{-1}X^\top X(X^\top X)^{-1}X^\top.$$

Since $X^\top X$ appears in the middle,

$$H^2 = X(X^\top X)^{-1}(X^\top X)(X^\top X)^{-1}X^\top.$$

Therefore,

$$H^2 = X(X^\top X)^{-1}X^\top = H.$$

□

7 Geometric Interpretation of Least Squares

The column space of X is

$$\mathcal{C}(X) = \{X\boldsymbol{b} : \boldsymbol{b} \in \mathbb{R}^{p+1}\}.$$

This is the set of all fitted vectors possible under the model.

OLS chooses $\widehat{\boldsymbol{y}} \in \mathcal{C}(X)$ closest to $\boldsymbol{y}$ in Euclidean distance:

$$\widehat{\boldsymbol{y}} = \underset{\boldsymbol{z}\in\mathcal{C}(X)}{\arg\min} \|\boldsymbol{y} - \boldsymbol{z}\|_2^2.$$

The residual vector is orthogonal to the column space of X:

$$X^\top \boldsymbol{e} = \boldsymbol{0}.$$

Indeed,

$$X^\top \boldsymbol{e} = X^\top(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}) = X^\top\boldsymbol{y} - X^\top X\widehat{\boldsymbol{\beta}} = \boldsymbol{0},$$

by the normal equations.

Geometry in one sentence

OLS projects the response vector $\boldsymbol{y}$ orthogonally onto the column space of the design matrix X.

8 Unbiasedness and Variance of OLS

Assume

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}, \qquad \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n.$$

The OLS estimator is

$$\widehat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \boldsymbol{y}.$$

Substitute the model:

$$\widehat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top (X\boldsymbol{\beta} + \boldsymbol{\varepsilon}).$$

Therefore,

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (X^\top X)^{-1} X^\top \boldsymbol{\varepsilon}.$$

Taking expectation,

$$\boxed{\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}.}$$

Thus OLS is unbiased.

The variance is

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \mathrm{Var}\left\{(X^\top X)^{-1} X^\top \boldsymbol{\varepsilon}\right\}.$$

Therefore,

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = (X^\top X)^{-1} X^\top \,\mathrm{Var}(\boldsymbol{\varepsilon}) X (X^\top X)^{-1}.$$

Since $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$,

$$\boxed{\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2 (X^\top X)^{-1}.}$$

9 Gauss–Markov Theorem

Theorem 3 (Gauss–Markov theorem)**.** *Under the assumptions*

$$\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}, \qquad \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n, \qquad \mathrm{rank}(X) = p + 1,$$

the OLS estimator $\widehat{\boldsymbol{\beta}}$ is the Best Linear Unbiased Estimator, abbreviated BLUE, of $\boldsymbol{\beta}$.

Meaning of BLUE

- Linear: the estimator is a linear function of $\boldsymbol{y}$.
- Unbiased: its expectation equals $\boldsymbol{\beta}$.
- Best: it has the smallest variance among all linear unbiased estimators.

Proof. Consider any linear unbiased estimator of $\boldsymbol{\beta}$ of the form

$$\widetilde{\boldsymbol{\beta}} = A\boldsymbol{y},$$

where A is a $(p+1) \times n$ matrix.

Unbiasedness means

$$\mathbb{E}(\widetilde{\boldsymbol{\beta}}) = A\mathbb{E}(\boldsymbol{y}) = AX\boldsymbol{\beta} = \boldsymbol{\beta}$$

for all $\boldsymbol{\beta}$. Hence

$$AX = I_{p+1}.$$

The OLS estimator can be written as

$$\widehat{\boldsymbol{\beta}} = C\boldsymbol{y}, \qquad C = (X^\top X)^{-1}X^\top.$$

Since

$$CX = (X^\top X)^{-1}X^\top X = I_{p+1},$$

OLS is unbiased.

Now write

$$A = C + D,$$

where

$$D = A - C.$$

Then

$$DX = AX - CX = I_{p+1} - I_{p+1} = 0.$$

The variance of $\widetilde{\boldsymbol{\beta}}$ is

$$\mathrm{Var}(\widetilde{\boldsymbol{\beta}}) = \sigma^2 AA^\top.$$

Now

$$AA^\top = (C+D)(C+D)^\top = CC^\top + CD^\top + DC^\top + DD^\top.$$

But

$$CD^\top = (X^\top X)^{-1}X^\top D^\top = (X^\top X)^{-1}(DX)^\top = 0.$$

Similarly,

$$DC^\top = 0.$$

Therefore,

$$AA^\top = CC^\top + DD^\top.$$

Since $DD^\top$ is positive semidefinite,

$$AA^\top - CC^\top = DD^\top \geq 0.$$

Hence

$$\mathrm{Var}(\widetilde{\boldsymbol{\beta}}) - \mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2 DD^\top \geq 0.$$

Thus OLS has minimum variance among all linear unbiased estimators. □

10 Estimating the Error Variance

The residual sum of squares is

$$\mathrm{RSS} = \boldsymbol{e}^\top \boldsymbol{e} = (\boldsymbol{y} - \widehat{\boldsymbol{y}})^\top (\boldsymbol{y} - \widehat{\boldsymbol{y}}).$$

Since

$$\boldsymbol{e} = (I - H)\boldsymbol{y},$$

we have

$$\mathrm{RSS} = \boldsymbol{y}^\top (I - H)\boldsymbol{y}.$$

An unbiased estimator of σ^2 is

$$\boxed{\widehat{\sigma}^2 = \frac{\mathrm{RSS}}{n - p - 1}.}$$

Why $n-p-1$?

The model estimates $p+1$ regression coefficients, including the intercept. Therefore, the residual degrees of freedom are

$$n-(p+1)=n-p-1.$$

11 Sampling Distribution under Normal Errors

If

$$\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 I_n),$$

then

$$\widehat{\boldsymbol{\beta}} \sim N_{p+1}\left(\boldsymbol{\beta}, \sigma^2 (X^\top X)^{-1}\right).$$

Let

$$C = (X^\top X)^{-1}.$$

Then the variance of $\widehat{\beta}_j$ is

$$\mathrm{Var}(\widehat{\beta}_j) = \sigma^2 c_{jj}.$$

Hence

$$\frac{\widehat{\beta}_j - \beta_j}{\sigma\sqrt{c_{jj}}} \sim N(0,1).$$

Since σ^2 is unknown, replace it by $\widehat{\sigma}^2$. Then

$$\boxed{\frac{\widehat{\beta}_j - \beta_j}{\widehat{\sigma}\sqrt{c_{jj}}} \sim t_{n-p-1}.}$$

12 Confidence Interval for a Regression Coefficient

A $100(1-\alpha)\%$ confidence interval for β_j is

$$\boxed{\widehat{\beta}_j \pm t_{\alpha/2,n-p-1}\widehat{\sigma}\sqrt{c_{jj}},}$$

where c_{jj} is the j-th diagonal element of $(X^\top X)^{-1}$.

13 Testing a Single Regression Coefficient

To test

$$H_0 : \beta_j = 0 \qquad \text{against} \qquad H_1 : \beta_j \neq 0,$$

use

$$\boxed{t_j = \frac{\widehat{\beta}_j}{\widehat{\sigma}\sqrt{c_{jj}}} \sim t_{n-p-1} \quad \text{under } H_0.}$$

A large absolute value of t_j provides evidence against H_0.

Interpretation warning

A significant t-test for β_j means that predictor x_j contributes after adjusting for all other predictors in the model. It does not necessarily mean that x_j has a strong marginal association with y.

14 Linear Contrasts

A linear combination of regression coefficients is

$$\theta = \boldsymbol{a}^\top \boldsymbol{\beta},$$

where $\boldsymbol{a} \in \mathbb{R}^{p+1}$ is fixed.

Its estimator is

$$\widehat{\theta} = \boldsymbol{a}^\top \widehat{\boldsymbol{\beta}}.$$

Since

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2 (X^\top X)^{-1},$$

we have

$$\mathrm{Var}(\widehat{\theta}) = \sigma^2 \boldsymbol{a}^\top (X^\top X)^{-1} \boldsymbol{a}.$$

Therefore,

$$\boxed{\frac{\boldsymbol{a}^\top \widehat{\boldsymbol{\beta}} - \boldsymbol{a}^\top \boldsymbol{\beta}}{\widehat{\sigma}\sqrt{\boldsymbol{a}^\top (X^\top X)^{-1} \boldsymbol{a}}} \sim t_{n-p-1}.}$$

15 Prediction at a New Point

Let

$$\boldsymbol{x}_0 = \begin{pmatrix} 1 \\ x_{01} \\ \vdots \\ x_{0p} \end{pmatrix}$$

be a new predictor vector.

The mean response at $\boldsymbol{x}_0$ is

$$\mu_0 = \boldsymbol{x}_0^\top \boldsymbol{\beta}.$$

Its estimator is

$$\widehat{\mu}_0 = \boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}}.$$

The variance of $\widehat{\mu}_0$ is

$$\mathrm{Var}(\widehat{\mu}_0) = \sigma^2 \boldsymbol{x}_0^\top (X^\top X)^{-1} \boldsymbol{x}_0.$$

A confidence interval for the mean response is

$$\boxed{\widehat{\mu}_0 \pm t_{\alpha/2, n-p-1} \widehat{\sigma} \sqrt{\boldsymbol{x}_0^\top (X^\top X)^{-1} \boldsymbol{x}_0}.}$$

For predicting a new observation

$$Y_0 = \boldsymbol{x}_0^\top \boldsymbol{\beta} + \varepsilon_0,$$

the prediction variance is

$$\mathrm{Var}(Y_0 - \widehat{\mu}_0) = \sigma^2 \left\{ 1 + \boldsymbol{x}_0^\top (X^\top X)^{-1} \boldsymbol{x}_0 \right\}.$$

Thus the prediction interval is

$$\boxed{\widehat{\mu}_0 \pm t_{\alpha/2, n-p-1} \widehat{\sigma} \sqrt{1 + \boldsymbol{x}_0^\top (X^\top X)^{-1} \boldsymbol{x}_0}.}$$

Confidence interval versus prediction interval

A confidence interval estimates the average response at $\boldsymbol{x}_0$. A prediction interval predicts a future individual response. The prediction interval is wider because it includes both estimation uncertainty and new observation noise.

16 ANOVA Decomposition in Multiple Regression

Assume the model includes an intercept. Define

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

The total sum of squares is

$$\mathrm{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

The regression sum of squares is

$$\mathrm{SSR} = \sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2.$$

The residual sum of squares is

$$\mathrm{SSE} = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2.$$

Then

$$\boxed{\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}\,.}$$

The coefficient of determination is

$$\boxed{R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}.}$$

The adjusted coefficient of determination is

$$\boxed{R^2_{\mathrm{adj}} = 1 - \frac{\mathrm{SSE}\,/(n-p-1)}{\mathrm{SST}\,/(n-1)}.}$$

Meaning of R^2

R^2 measures the proportion of total variability in the response explained by the fitted regression model. However, R^2 never decreases when new predictors are added. Adjusted R^2 penalizes unnecessary predictors.

17 Overall F-Test

To test whether the predictors jointly explain the response, we test

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

against

$$H_1 : \text{at least one } \beta_j \neq 0.$$

The test statistic is

$$\boxed{F = \frac{\text{SSR}/p}{\text{SSE}/(n-p-1)}.}$$

Under H_0,

$$F \sim F_{p,n-p-1}.$$

The ANOVA table is:

Source	df	Sum of Squares	Mean Square
Regression	p	SSR	$\text{MSR} = \text{SSR}/p$
Error	$n-p-1$	SSE	$\text{MSE} = \text{SSE}/(n-p-1)$
Total	$n-1$	SST	

18 Partial F-Test

Suppose a full model contains all predictors and a reduced model contains a subset of predictors.

Let

$$\text{SSE}_R = \text{residual sum of squares of reduced model},$$

$$\text{SSE}_F = \text{residual sum of squares of full model}.$$

Suppose the reduced model has q fewer predictors than the full model. Then

$$\boxed{F = \frac{(\text{SSE}_R - \text{SSE}_F)/q}{\text{SSE}_F/(n-p_F-1)},}$$

where p_F is the number of predictors in the full model.

Under the null hypothesis that the excluded variables have zero coefficients,

$$F \sim F_{q,n-p_F-1}.$$

Meaning of partial F-test

The partial F-test asks whether a group of variables adds significant explanatory power after the variables in the reduced model have already been included.

19 Partial Regression Interpretation

Consider the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon.$$

The coefficient β_1 is the effect of X_1 on Y after adjusting for X_2.

This can be understood by the Frisch–Waugh–Lovell idea:

1. Regress Y on X_2, and keep the residuals.
2. Regress X_1 on X_2, and keep the residuals.
3. Regress the first residuals on the second residuals.

The slope in step 3 equals the coefficient of X_1 in the full multiple regression.

Deep interpretation

Multiple regression removes from Y and X_1 the part already explained by other predictors, then measures the remaining association.

20 Multicollinearity

Multicollinearity occurs when predictors are highly linearly related.

In matrix terms, multicollinearity means $X^\top X$ is close to singular. Then $(X^\top X)^{-1}$ has large diagonal elements, causing large standard errors.

The variance of $\widehat{\beta}_j$ is

$$\mathrm{Var}(\widehat{\beta}_j) = \sigma^2 c_{jj},$$

where

$$C = (X^\top X)^{-1}.$$

Thus large c_{jj} gives unstable estimates.

A common diagnostic is the variance inflation factor:

$$\boxed{\mathrm{VIF}_j = \frac{1}{1 - R_j^2},}$$

where R_j^2 is obtained by regressing predictor X_j on all other predictors.

Interpretation of VIF

If R_j^2 is close to 1, then X_j is almost explained by the other predictors. Then VIF_j is large, and the standard error of $\widehat{\beta}_j$ is inflated.

21 Leverage and Residual Diagnostics

The leverage of observation i is the i-th diagonal element of the hat matrix:

$$\boxed{h_{ii} = \boldsymbol{x}_i^\top (X^\top X)^{-1} \boldsymbol{x}_i.}$$

Leverage measures how far the predictor vector $\boldsymbol{x}_i$ is from the center of the predictor cloud.

Properties:

$$0 \le h_{ii} \le 1,$$

$$\sum_{i=1}^{n} h_{ii} = p + 1.$$

Thus the average leverage is

$$\bar{h} = \frac{p+1}{n}.$$

A common rule is that observations with

$$h_{ii} > \frac{2(p+1)}{n} \quad \text{or} \quad h_{ii} > \frac{3(p+1)}{n}$$

may be considered high-leverage points.

The residual variance is

$$\mathrm{Var}(e_i) = \sigma^2 (1 - h_{ii}).$$

The standardized residual is

$$r_i = \frac{e_i}{\widehat{\sigma}\sqrt{1 - h_{ii}}}.$$

Leverage versus outlier

A response outlier has an unusual y-value. A leverage point has unusual predictor values. A point can have high leverage without being an outlier, but a high-leverage outlier can strongly influence the regression fit.

22 Worked Numerical Example

A small multiple regression example

Suppose

$$X^\top X = \begin{pmatrix} 5 & 10 & 6 \\ 10 & 30 & 14 \\ 6 & 14 & 10 \end{pmatrix}, \qquad X^\top \boldsymbol{y} = \begin{pmatrix} 20 \\ 56 \\ 34 \end{pmatrix}.$$

The model has an intercept and two predictors:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i.$$

The normal equations are

$$\begin{pmatrix} 5 & 10 & 6 \\ 10 & 30 & 14 \\ 6 & 14 & 10 \end{pmatrix} \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 20 \\ 56 \\ 34 \end{pmatrix}.$$

Solving this system gives

$$\widehat{\beta}_0 = 1, \qquad \widehat{\beta}_1 = 1.4, \qquad \widehat{\beta}_2 = 0.8.$$

Thus the fitted model is

$$\widehat{y} = 1 + 1.4x_1 + 0.8x_2.$$

Interpretation:

- Holding x_2 fixed, a one-unit increase in x_1 increases the fitted mean response by 1.4.
- Holding x_1 fixed, a one-unit increase in x_2 increases the fitted mean response by 0.8.

23 Home Assignment: Multiple Linear Regression

Course: Regression Techniques

Level: MSc Statistics

Instruction: Answer all questions. Each solution must include mathematical derivation and interpretation.

Problem 1: Derivation of the OLS Estimator

Let

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where X is an $n \times (p+1)$ matrix of full column rank. Derive the OLS estimator by minimizing

$$\mathrm{RSS}(\boldsymbol{\beta}) = \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2.$$

Solution to Problem 1

We minimize

$$\mathrm{RSS}(\boldsymbol{\beta}) = (\boldsymbol{y} - X\boldsymbol{\beta})^\top (\boldsymbol{y} - X\boldsymbol{\beta}).$$

Expanding,

$$\mathrm{RSS}(\boldsymbol{\beta}) = \boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{\beta}^\top X^\top \boldsymbol{y} + \boldsymbol{\beta}^\top X^\top X\boldsymbol{\beta}.$$

Differentiating with respect to $\boldsymbol{\beta}$,

$$\frac{\partial \mathrm{RSS}}{\partial \boldsymbol{\beta}} = -2X^\top \boldsymbol{y} + 2X^\top X\boldsymbol{\beta}.$$

At the minimum,

$$-2X^\top \boldsymbol{y} + 2X^\top X\widehat{\boldsymbol{\beta}} = 0.$$

Therefore,

$$X^\top X\widehat{\boldsymbol{\beta}} = X^\top \boldsymbol{y}.$$

Since X has full column rank, $X^\top X$ is nonsingular. Hence

$$\boxed{\widehat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \boldsymbol{y}.}$$

The Hessian is $2X^\top X$, which is positive definite. Therefore, the solution is the unique global minimizer.

Problem 2: Orthogonality of Residuals

Show that the residual vector

$$\boldsymbol{e} = \boldsymbol{y} - X\widehat{\boldsymbol{\beta}}$$

is orthogonal to every column of X.

Solution to Problem 2

The normal equations are

$$X^\top X\widehat{\boldsymbol{\beta}} = X^\top \boldsymbol{y}.$$

Rearranging,

$$X^\top \boldsymbol{y} - X^\top X\widehat{\boldsymbol{\beta}} = 0.$$

Thus

$$X^\top(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}) = 0.$$

Since

$$\boldsymbol{e} = \boldsymbol{y} - X\widehat{\boldsymbol{\beta}},$$

we have

$$\boxed{X^\top \boldsymbol{e} = 0.}$$

This means that the residual vector is orthogonal to every column of X. Geometrically, the fitted vector $X\widehat{\boldsymbol{\beta}}$ is the orthogonal projection of $\boldsymbol{y}$ onto the column space of X.

Problem 3: Hat Matrix

Define

$$H = X(X^\top X)^{-1}X^\top.$$

Show that H is symmetric and idempotent. Also show that

$$\operatorname{tr}(H) = p + 1.$$

Solution to Problem 3

First,

$$H^\top = \left\{X(X^\top X)^{-1}X^\top\right\}^\top = X\left\{(X^\top X)^{-1}\right\}^\top X^\top.$$

Since $X^\top X$ is symmetric, $(X^\top X)^{-1}$ is also symmetric. Therefore,

$$H^\top = H.$$

Next,

$$H^2 = X(X^\top X)^{-1}X^\top X(X^\top X)^{-1}X^\top.$$

Using $X^\top X$ in the middle,

$$H^2 = X(X^\top X)^{-1}(X^\top X)(X^\top X)^{-1}X^\top.$$

Thus

$$H^2 = H.$$

Finally,

$$\operatorname{tr}(H) = \operatorname{tr}\left\{X(X^\top X)^{-1}X^\top\right\}.$$

Using cyclic invariance of trace,

$$\operatorname{tr}(H) = \operatorname{tr}\left\{(X^\top X)^{-1}X^\top X\right\}.$$

Hence

$$\operatorname{tr}(H) = \operatorname{tr}(I_{p+1}) = p + 1.$$

Problem 4: Unbiasedness and Variance of OLS

Assume

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}, \qquad \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n.$$

Show that

$$\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

and

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2 (X^\top X)^{-1}.$$

Solution to Problem 4

The OLS estimator is

$$\widehat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \boldsymbol{y}.$$

Substitute

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Then

$$\widehat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top (X\boldsymbol{\beta} + \boldsymbol{\varepsilon}).$$

Hence

$$\widehat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top X\boldsymbol{\beta} + (X^\top X)^{-1} X^\top \boldsymbol{\varepsilon}.$$

Therefore,

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (X^\top X)^{-1} X^\top \boldsymbol{\varepsilon}.$$

Taking expectation,

$$\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + (X^\top X)^{-1} X^\top \mathbb{E}(\boldsymbol{\varepsilon}).$$

Since $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$,

$$\boxed{\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}.}$$

For variance,

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \mathrm{Var}\left\{ (X^\top X)^{-1} X^\top \boldsymbol{\varepsilon} \right\}.$$

Thus

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = (X^\top X)^{-1} X^\top \,\mathrm{Var}(\boldsymbol{\varepsilon}) X (X^\top X)^{-1}.$$

Since $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$,

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2 (X^\top X)^{-1} X^\top X (X^\top X)^{-1}.$$

Therefore,

$$\boxed{\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2 (X^\top X)^{-1}.}$$

Problem 5: Estimator of σ^2

Show that

$$\widehat{\sigma}^2 = \frac{\mathrm{RSS}}{n - p - 1}$$

is an unbiased estimator of σ^2.

Solution to Problem 5

The residual vector is

$$\boldsymbol{e} = (I - H)\boldsymbol{y}.$$

Since

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

and

$$(I - H)X\boldsymbol{\beta} = 0,$$

we get

$$\boldsymbol{e} = (I - H)\boldsymbol{\varepsilon}.$$

Therefore,

$$\text{RSS} = \boldsymbol{e}^\top \boldsymbol{e} = \boldsymbol{\varepsilon}^\top (I - H)^\top (I - H)\boldsymbol{\varepsilon}.$$

Because $I - H$ is symmetric and idempotent,

$$\text{RSS} = \boldsymbol{\varepsilon}^\top (I - H)\boldsymbol{\varepsilon}.$$

For a random vector $\boldsymbol{\varepsilon}$ with mean zero and variance $\sigma^2 I_n$,

$$\mathbb{E}(\boldsymbol{\varepsilon}^\top A \boldsymbol{\varepsilon}) = \sigma^2 \operatorname{tr}(A)$$

for any fixed symmetric matrix A. Hence

$$\mathbb{E}(\text{RSS}) = \sigma^2 \operatorname{tr}(I - H).$$

Now

$$\operatorname{tr}(I - H) = \operatorname{tr}(I) - \operatorname{tr}(H) = n - (p + 1) = n - p - 1.$$

Thus

$$\mathbb{E}(\text{RSS}) = \sigma^2 (n - p - 1).$$

Therefore,

$$\mathbb{E}\left(\frac{\text{RSS}}{n - p - 1}\right) = \sigma^2.$$

Hence

$$\boxed{\widehat{\sigma}^2 = \frac{\text{RSS}}{n - p - 1}}$$

is unbiased.

Problem 6: Confidence Interval for a Coefficient

Suppose

$$\widehat{\beta}_j = 2.40, \qquad \widehat{\sigma} = 1.20, \qquad c_{jj} = 0.25, \qquad n - p - 1 = 20.$$

Find a 95% confidence interval for β_j. Use $t_{0.025,20} = 2.086$.

Solution to Problem 6

The standard error of $\widehat{\beta}_j$ is

$$\mathrm{SE}(\widehat{\beta}_j) = \widehat{\sigma}\sqrt{c_{jj}}.$$

Substituting,

$$\mathrm{SE}(\widehat{\beta}_j) = 1.20\sqrt{0.25} = 1.20(0.5) = 0.60.$$

The 95% confidence interval is

$$\widehat{\beta}_j \pm t_{0.025,20}\,\mathrm{SE}(\widehat{\beta}_j).$$

Thus

$$2.40 \pm 2.086(0.60).$$

Now

$$2.086(0.60) = 1.2516.$$

Therefore, the interval is

$$(2.40 - 1.2516,\ 2.40 + 1.2516).$$

Hence

$$\boxed{(1.1484,\ 3.6516).}$$

Approximately,

$$\boxed{(1.15,\ 3.65).}$$

Since the interval does not contain zero, the coefficient is statistically significant at the 5% level.

Problem 7: Testing a Single Coefficient

For a regression coefficient, suppose

$$\widehat{\beta}_j = 1.8, \qquad \mathrm{SE}(\widehat{\beta}_j) = 0.6, \qquad n - p - 1 = 25.$$

Test

$$H_0 : \beta_j = 0$$

against

$$H_1 : \beta_j \neq 0.$$

Solution to Problem 7

The test statistic is

$$t = \frac{\widehat{\beta}_j - 0}{\mathrm{SE}(\widehat{\beta}_j)}.$$

Substitute the values:

$$t = \frac{1.8}{0.6} = 3.$$

Under H_0,

$$t \sim t_{25}.$$

For a two-sided test at 5%, the critical value is approximately

$$t_{0.025,25} \approx 2.06.$$

Since

$$|3| > 2.06,$$

we reject H_0.
Thus there is significant evidence that $\beta_j \neq 0$, after adjusting for other predictors in the model.

Problem 8: ANOVA Decomposition

Suppose a multiple regression model with $n = 30$ observations and $p = 3$ predictors gives

$$\mathrm{SST} = 500, \qquad \mathrm{SSE} = 120.$$

Find SSR, R^2, adjusted R^2, and the overall F-statistic.

Solution to Problem 8

Since

$$\text{SST} = \text{SSR} + \text{SSE},$$

we get

$$\text{SSR} = \text{SST} - \text{SSE} = 500 - 120 = 380.$$

The coefficient of determination is

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{380}{500} = 0.76.$$

The adjusted R^2 is

$$R^2_{\text{adj}} = 1 - \frac{\text{SSE}/(n-p-1)}{\text{SST}/(n-1)}.$$

Here

$$n = 30, \qquad p = 3, \qquad n - p - 1 = 26.$$

Therefore,

$$R^2_{\text{adj}} = 1 - \frac{120/26}{500/29}.$$

Compute

$$120/26 = 4.6154, \qquad 500/29 = 17.2414.$$

Thus

$$R^2_{\text{adj}} = 1 - \frac{4.6154}{17.2414} = 1 - 0.2677 = 0.7323.$$

So

$$\boxed{R^2_{\text{adj}} \approx 0.732.}$$

The overall F-statistic is

$$F = \frac{\text{SSR}/p}{\text{SSE}/(n-p-1)}.$$

Substitute:

$$F = \frac{380/3}{120/26}.$$

Now

$$380/3 = 126.6667, \qquad 120/26 = 4.6154.$$

Therefore,

$$F = \frac{126.6667}{4.6154} = 27.444.$$

Thus

$$\boxed{F \approx 27.44.}$$

This is a large value, suggesting that the predictors jointly explain a significant part of the response variation.

Problem 9: Partial F-Test

A reduced model gives

$$\text{SSE}_R = 180.$$

A full model gives

$$\mathrm{SSE}_F = 120.$$

The full model has $p_F = 5$ predictors and $n = 40$ observations. The reduced model excludes $q = 2$ predictors. Test whether the excluded predictors are jointly significant.

Solution to Problem 9

The partial F-statistic is

$$F = \frac{(\mathrm{SSE}_R - \mathrm{SSE}_F)/q}{\mathrm{SSE}_F/(n - p_F - 1)}.$$

Substitute:

$$F = \frac{(180 - 120)/2}{120/(40 - 5 - 1)}.$$

Compute:

$$(180 - 120)/2 = 60/2 = 30.$$

The denominator degrees of freedom are

$$40 - 5 - 1 = 34.$$

Thus

$$120/34 = 3.5294.$$

Therefore,

$$F = \frac{30}{3.5294} = 8.50.$$

So

$$\boxed{F = 8.50.}$$

Under the null hypothesis,

$$F \sim F_{2,34}.$$

Since 8.50 is large, we reject the null hypothesis at conventional significance levels. Thus the two excluded predictors jointly add significant explanatory power to the model.

Problem 10: Prediction Interval

Suppose for a new point $\boldsymbol{x}_0$,

$$\widehat{\mu}_0 = 25, \qquad \widehat{\sigma} = 3, \qquad \boldsymbol{x}_0^\top (X^\top X)^{-1} \boldsymbol{x}_0 = 0.10, \qquad n - p - 1 = 20.$$

Find a 95% confidence interval for the mean response and a 95% prediction interval for a new observation. Use $t_{0.025,20} = 2.086$.

Solution to Problem 10

The confidence interval for the mean response is

$$\widehat{\mu}_0 \pm t_{0.025,20}\widehat{\sigma}\sqrt{\boldsymbol{x}_0^\top (X^\top X)^{-1}\boldsymbol{x}_0}.$$

Substitute:

$$25 \pm 2.086(3)\sqrt{0.10}.$$

Since

$$\sqrt{0.10} = 0.3162,$$

we get

$$2.086(3)(0.3162) = 1.979.$$

Thus the confidence interval is

$$\boxed{(23.021,\ 26.979).}$$

The prediction interval is

$$\widehat{\mu}_0 \pm t_{0.025,20}\widehat{\sigma}\sqrt{1 + \boldsymbol{x}_0^\top (X^\top X)^{-1}\boldsymbol{x}_0}.$$

Substitute:

$$25 \pm 2.086(3)\sqrt{1.10}.$$

Since

$$\sqrt{1.10} = 1.0488,$$

we get

$$2.086(3)(1.0488) = 6.565.$$

Thus the prediction interval is

$$\boxed{(18.435,\ 31.565).}$$

The prediction interval is wider because it includes the random variation of a new observation.

Problem 11: Variance Inflation Factor

Suppose predictor X_j is regressed on all other predictors and the resulting coefficient of determination is

$$R_j^2 = 0.90.$$

Compute the VIF and interpret it.

Solution to Problem 11

The variance inflation factor is

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2}.$$

Substitute:

$$\mathrm{VIF}_j = \frac{1}{1 - 0.90} = \frac{1}{0.10} = 10.$$

Thus

$$\boxed{\mathrm{VIF}_j = 10.}$$

This means that the variance of $\widehat{\beta}_j$ is inflated by a factor of 10 compared with the situation where X_j is uncorrelated with the other predictors. Equivalently, the standard error is inflated by a factor of

$$\sqrt{10} \approx 3.16.$$

This indicates serious multicollinearity involving X_j.

Problem 12: Leverage

Suppose a model has $n = 50$ observations and $p = 4$ predictors. What is the average leverage? If an observation has $h_{ii} = 0.30$, would you consider it high leverage?

Solution to Problem 12

The average leverage is

$$\bar{h} = \frac{p+1}{n}.$$

Here

$$p = 4, \qquad n = 50.$$

Therefore,

$$\bar{h} = \frac{5}{50} = 0.10.$$

Common high-leverage cutoffs are

$$\frac{2(p+1)}{n} = \frac{10}{50} = 0.20$$

and

$$\frac{3(p+1)}{n} = \frac{15}{50} = 0.30.$$

The given observation has

$$h_{ii} = 0.30.$$

It equals the stricter cutoff $3(p+1)/n$. Therefore, it should be treated as a potentially high-leverage observation and examined carefully.

Problem 13: Conceptual Interpretation of Coefficients

Consider the fitted model

$$\widehat{y} = 12 + 3.5x_1 - 2.1x_2 + 0.8x_3.$$

Interpret each coefficient carefully.

Solution to Problem 13

The intercept is

$$12.$$

It represents the fitted mean response when

$$x_1 = x_2 = x_3 = 0.$$

This interpretation is meaningful only if the zero values of the predictors are meaningful and within the range of the data.
The coefficient of x_1 is

$$3.5.$$

Holding x_2 and x_3 fixed, a one-unit increase in x_1 is associated with an increase of 3.5 units in the fitted mean response.
The coefficient of x_2 is

$$-2.1.$$

Holding x_1 and x_3 fixed, a one-unit increase in x_2 is associated with a decrease of 2.1 units in the fitted mean response.
The coefficient of x_3 is

$$0.8.$$

Holding x_1 and x_2 fixed, a one-unit increase in x_3 is associated with an increase of 0.8 units in the fitted mean response.
The phrase "holding other variables fixed" is essential in multiple regression.

Problem 14: Gauss–Markov Theorem

State the Gauss–Markov theorem and explain why normality is not required for it.

Solution to Problem 14

The Gauss–Markov theorem states that, under the assumptions

$$\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}, \qquad \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n, \qquad \mathrm{rank}(X) = p + 1,$$

the OLS estimator

$$\widehat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \boldsymbol{y}$$

is the Best Linear Unbiased Estimator of $\boldsymbol{\beta}$.
This means that among all estimators that are:

1. linear functions of $\boldsymbol{y}$,

2. unbiased for $\boldsymbol{\beta}$,

OLS has the smallest variance matrix in the positive semidefinite ordering.
Normality is not required because the theorem uses only the first two moments:

$$\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}, \qquad \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n.$$

Normality is needed for exact finite-sample t-tests and F-tests, but not for the BLUE property.

Problem 15: Short Conceptual Questions

Answer briefly but rigorously.

(a) Why does adding more predictors never increase SSE?

(b) Why can R^2 be misleading?

(c) What is the difference between residual and error?

(d) Why is multicollinearity a problem?

(e) What is the difference between a confidence interval and a prediction interval?

Solution to Problem 15

(a) Adding predictors and SSE.
Adding predictors enlarges the column space of the design matrix. OLS minimizes RSS over this column space. A larger space cannot give a worse minimum. Therefore, SSE cannot increase when predictors are added.

(b) Why R^2 can be misleading.
R^2 never decreases when new predictors are added, even if those predictors are irrelevant. Thus a large R^2 may reflect overfitting rather than genuine explanatory power. Adjusted R^2, cross-validation, and residual diagnostics are needed for better assessment.

(c) Residual versus error.
The error is

$$\varepsilon_i = y_i - \mathbb{E}(Y_i),$$

which is unobserved. The residual is

$$e_i = y_i - \widehat{y}_i,$$

which is observed after fitting the model. Residuals estimate errors but are not identical to them.

(d) Multicollinearity.
Multicollinearity makes $X^\top X$ close to singular. This inflates the diagonal elements of $(X^\top X)^{-1}$, leading to large standard errors and unstable coefficient estimates. As a result, individual t-tests may be insignificant even when the overall model is useful.

(e) Confidence interval versus prediction interval.
A confidence interval estimates the mean response at a given predictor value. A prediction interval predicts a future individual observation. The prediction interval is wider because it includes both uncertainty in estimating the mean and the random error of a new observation.

24 Additional Instructor Notes for Teaching

How to teach multiple regression in a job talk

Do not begin with software output. Begin with the question:

> What does it mean to isolate the effect of one variable while adjusting for others?

Then introduce the matrix form, least squares geometry, and inference.

Three messages students must remember

1. Multiple regression coefficients are partial effects, not marginal effects.
2. OLS is an orthogonal projection of $\boldsymbol{y}$ onto the column space of X.
3. Inference depends on uncertainty quantification through $\widehat{\sigma}^2$, standard errors, and degrees of freedom.

Common student mistakes

1. Interpreting β_j without saying "holding other predictors fixed."
2. Thinking that high R^2 always means a good model.
3. Ignoring multicollinearity when standard errors are large.
4. Confusing residuals with true errors.
5. Confusing confidence intervals with prediction intervals.

Suggested References

1. Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis*. Wiley.
2. Montgomery, D. C., Peck, E. A. and Vining, G. G. (2012). *Introduction to Linear Regression Analysis*. Wiley.
3. Weisberg, S. (2005). *Applied Linear Regression*. Wiley.
4. Seber, G. A. F. and Lee, A. J. (2003). *Linear Regression Analysis*. Wiley.
5. Kutner, M. H., Nachtsheim, C. J., Neter, J. and Li, W. (2005). *Applied Linear Statistical Models*. McGraw–Hill.