7 *Multiple Linear Regression*

The previous chapter fitted a straight line through data with one regressor: $Y = \beta_0 + \beta_1 x + \varepsilon$. In practice, one regressor is rarely enough. The delivery time of an order depends on the distance, the number of stops, and the load. The churn rate of a subscription product depends on the price, the onboarding time, and the support load. The evaluation score of an AI model depends on the model size and the quality of its training data. This chapter extends the regression machinery to several regressors at once. The chapter is organized in the following order: the multiple linear regression model and how to read each coefficient; the matrix form of the model and the least squares solution, with a fully worked hand computation; the estimate of σ^2 and the standard errors of the coefficients; the analysis of variance and the overall F test; R^2 and the adjusted R^2; how to read a regression summary table from software; tests and confidence intervals on individual coefficients; the partial F test for a group of coefficients; intervals for the mean response and for a new observation; multicollinearity and the variance inflation factor; indicator, interaction, and polynomial terms; a step-by-step model-building procedure; and residual diagnostics.

Almost everything in this chapter is a direct generalization of the previous one. The sums of squares, the ANOVA table, the t statistics, and the residual plots all keep the same structure. What changes is the bookkeeping: with k regressors there are $p = k + 1$ parameters, the residual degrees of freedom become $n - p$, and the convenient scalar formulas $\hat{\beta}_1 = S_{xy}/S_{xx}$ are replaced by one matrix formula that covers every case. Please pay attention to the degrees of freedom throughout this chapter, because using $n - 2$ where $n - p$ is required is the most common error in multiple regression problems.

Pre-class quiz. *Try these before lecture; the answers follow at the end of the quiz.*

Q1. A logistics analyst fits $\hat{y} = 2.5 + 0.09\,x_1 + 0.60\,x_2$, where y is delivery time in hours, x_1 is distance in km, and x_2 is the number of stops. What exactly does the coefficient 0.60 mean?

Q2. An AI evaluation team fits a multiple regression with $k = 4$ regressors on $n = 26$ test runs. How many degrees of freedom does the residual (error) term have, and which distribution does the overall F statistic follow under H_0?

Q3. A startup analyst adds a fifth regressor to a churn model and reports that R^2 rose from 0.810 to 0.815. Does this prove the new regressor is useful?

Answers.

Q1. Each additional stop increases the *predicted* delivery time by 0.60 hours, *for deliveries at the same distance*. A coefficient in multiple regression always describes the effect of its own regressor with the other regressors held fixed.

Q2. The model estimates $p = k + 1 = 5$ parameters, so the residual degrees of freedom are $n - p = 26 - 5 = 21$. Under H_0 the overall statistic follows $F_{4,21}$: numerator degrees of freedom $k = 4$, denominator $n - p = 21$.

Q3. *No.* R^2 never decreases when a regressor is added, even a completely irrelevant one, so a small increase proves nothing. To judge the new regressor, look at its t statistic or at the *adjusted* R^2, which can decrease when a useless regressor is added.

Lecture 20 starts here — *Multiple linear regression, part 1*

Lecture 20. This section is covered by Lecture 20.

Reference. Montgomery & Runger, 6th ed., Section 12-1.1.

7.1 From One Regressor to Several

Suppose an e-commerce operation wants to model the delivery time y of an order. Using only the distance to the customer as a regressor gives a fitted line with a modest R^2: distance matters, but it is not the whole story. The number of items in the order and the experience of the delivery staff also move the response. A model that includes all of these inputs can explain much more of the variability in y. The extension of simple linear regression to k regressors is the *multiple linear regression* (MLR) model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon. \tag{7.1}$$

Regressor. An input variable x_j used to explain or predict the response y. Also called a predictor or an independent variable.

Multiple linear regression model. A regression model with two or more regressors, $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$.

For a data set of n observations, the model says that each observed response is a linear combination of that observation's regressor values plus a random error:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, \qquad i = 1, 2, \ldots, n,$$

Partial regression coefficient. The coefficient β_j: the change in $E(Y)$ per unit change in x_j when all other regressors are held fixed.

where x_{ij} is the value of regressor j for observation i. The error assumptions are the same as in simple linear regression: $E(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$ for every i (constant variance), and the errors are uncorrelated. For confidence intervals and hypothesis tests we add $\varepsilon_i \sim N(0, \sigma^2)$.

There are two things to notice about the model before any computation.

First, how to read a coefficient. The parameter β_j is the change in the *mean* response per unit change in x_j, *with all the other regressors held fixed*. For this reason the β_j are called *partial* regression coefficients: each one isolates the contribution of its own regressor after the others have been accounted for. This is the single most frequently tested interpretation in this chapter, so please state it in full whenever you interpret a fitted model: ''holding the other regressors fixed, a one-unit increase in x_j changes the predicted response by $\hat{\beta}_j$ units.'' A coefficient read without the ''holding the others fixed'' clause can be badly misleading, because the same regressor can have a different coefficient, or even a different sign, when the set of other regressors in the model changes.

Safety lens. In AI evaluation work, a model score depends jointly on model scale, data quality, and test difficulty. Fitting the score on scale alone attributes the data-quality effect to scale, which misdirects the training budget. MLR separates the contributions.

Second, what ''linear'' means. The model is linear *in the parameters* β_j, not necessarily in the x's. The model $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$ contains a squared regressor, but it is still a linear regression model, because Y is a linear combination of the unknown β's. The model $Y = \beta_0 e^{\beta_1 x} \varepsilon$ is not, because β_1 appears in an exponent. Section 7.12 uses this fact to fit curves and interactions with the same machinery developed here.

Figure 7.1 shows the geometric picture for $k = 2$. With one regressor, least squares fits a line through a scatter of points in the plane. With two regressors, the fitted equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$ is a *plane* floating over the (x_1, x_2) region, and the residuals are the vertical distances from the data points to that plane. With $k > 2$ regressors the picture becomes a hyperplane that we can no longer draw, but the algebra, which is the subject of the next section, does not care about the dimension.

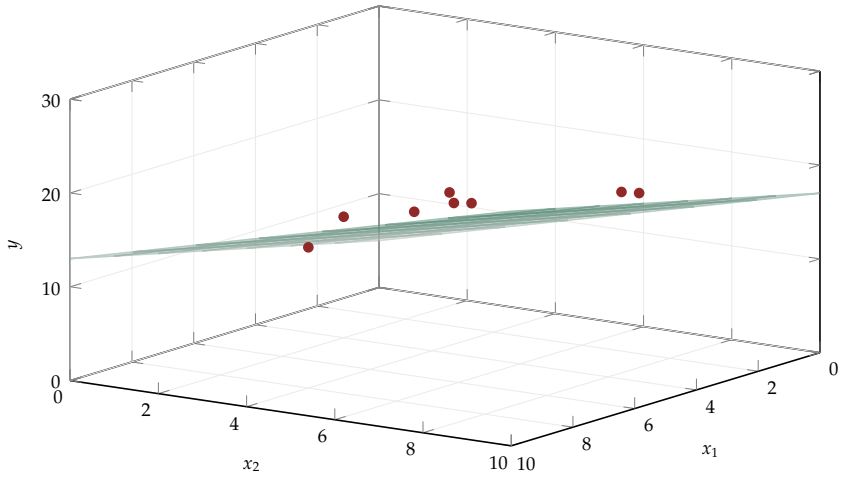

Figure 7.1. The fitted MLR surface for $k = 2$ regressors is a plane, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$. Each residual is the vertical distance from a data point (dots) to the plane. Least squares chooses the plane that minimizes the sum of the squared vertical distances.

With k regressors the model has $p = k + 1$ unknown parameters: $\beta_0, \beta_1, \ldots, \beta_k$, plus σ^2. We therefore need $n > p$ observations, and the residual degrees of freedom will be $n - p$. Write this identity down now and keep it visible; it drives every degrees-of-freedom entry in the chapter:

$$p = k + 1, \qquad \text{residual df} = n - p = n - k - 1.$$

Reference. Montgomery & Runger, 6th ed., Sections 12-1.2 and 12-1.3.

7.2 *The Matrix Form and Least Squares*

In simple linear regression the least squares estimates had closed scalar formulas: $\hat{\beta}_1 = S_{xy}/S_{xx}$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. With k regressors, writing one scalar formula per coefficient becomes hopeless. Instead we stack all n model equations into a single matrix equation:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{7.2}$$

where

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Design matrix. The $n \times p$ matrix $\boldsymbol{X}$ whose first column is all 1's (for the intercept) and whose remaining columns hold the regressor values.

Here $\boldsymbol{y}$ is $n \times 1$, $\boldsymbol{X}$ is $n \times p$, $\boldsymbol{\beta}$ is $p \times 1$, and $\boldsymbol{\varepsilon}$ is $n \times 1$.

The matrix $\boldsymbol{X}$ is called the *design matrix*. Its first column is all 1's; that column multiplies β_0 and produces the intercept in every row. Each remaining column holds one regressor. Building $\boldsymbol{X}$ correctly, with the column of 1's first, is the first step of every matrix regression computation, and forgetting the column of 1's is a common trap.

Least squares chooses $\boldsymbol{\beta}$ to minimize the sum of squared residuals,

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Differentiating with respect to each β_j and setting the derivatives to zero gives the *normal equations*:

Normal equations. The system $\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$ whose solution is the least squares estimator.

$$\boldsymbol{X}'\boldsymbol{X}\,\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}. \tag{7.3}$$

This is a system of p linear equations in the p unknowns $\hat{\beta}_0, \ldots, \hat{\beta}_k$. If $\boldsymbol{X}'\boldsymbol{X}$ is invertible, which requires that no regressor column is an exact linear combination of the others, the solution is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{7.4}$$

Equation (7.4) is the one formula that replaces both SLR formulas: applied with $k = 1$, it reproduces $\hat{\beta}_1 = S_{xy}/S_{xx}$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ exactly. Once $\hat{\boldsymbol{\beta}}$ is available, the fitted values and the residuals are

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}, \qquad e_i = y_i - \hat{y}_i.$$

Insight. Why the residuals sum to zero. The normal equations (7.3) can be rewritten as $\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{0}$: the residual vector is orthogonal to every column of $\boldsymbol{X}$. The first column of $\boldsymbol{X}$ is all 1's, and orthogonality to that column says exactly $\sum_i e_i = 0$. Orthogonality to each regressor column says $\sum_i x_{ij} e_i = 0$: the residuals carry no leftover linear information about any regressor in the model. Least squares has squeezed out everything the columns of $\boldsymbol{X}$ can explain.

On an exam, inverting a 3×3 matrix by hand is not the point of the question, so $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ will be given. Your job is the rest: build $\boldsymbol{X}$ and $\boldsymbol{y}$, compute $\boldsymbol{X}'\boldsymbol{y}$, and multiply. Procedure 7.1 lists the steps, and Example 7.1 carries them out in full.

Procedure 7.1 — Fitting an MLR model by matrix least squares.

1. Build the design matrix $\boldsymbol{X}$ ($n \times p$): a first column of 1's, then one column per regressor. Build the response vector $\boldsymbol{y}$ ($n \times 1$).
2. Compute $\boldsymbol{X}'\boldsymbol{y}$ (a $p \times 1$ vector) and, if it is not given, $\boldsymbol{X}'\boldsymbol{X}$ (a $p \times p$ matrix).
3. Obtain $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ (usually given) and multiply: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ by (7.4).
4. Write the fitted equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$ and interpret each coefficient holding the others fixed.
5. If residuals are needed, compute $\hat{y}_i$ for each row and $e_i = y_i - \hat{y}_i$; check that $\sum_i e_i = 0$.

Example 7.1 — Evaluation score from model size and data quality

An AI safety team evaluates $n = 8$ candidate perception models on a fixed benchmark. Each candidate is described by two coded regressors: $x_1 = +1$ for a large network and $x_1 = -1$ for a small one, and $x_2 = +1$ for high-quality (curated) training data and $x_2 = -1$ for uncurated data. The response y is the benchmark score (points out of 100). The data are shown in Table 7.1. Fit the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ by least squares.

Run i	x_{i1} (size)	x_{i2} (data quality)	y_i (score)
1	-1	-1	56
2	$+1$	-1	68
3	-1	$+1$	74
4	$+1$	$+1$	82
5	-1	-1	58
6	$+1$	-1	66
7	-1	$+1$	72
8	$+1$	$+1$	84

Table 7.1. Benchmark scores for eight evaluation runs with coded model size and data quality.

Solution.

Step 1 — Build X and y.

With $k = 2$ regressors and $n = 8$ runs, $\boldsymbol{X}$ is 8×3:

$$\boldsymbol{X} = \begin{bmatrix} 1 & -1 & -1 \\ 1 & +1 & -1 \\ 1 & -1 & +1 \\ 1 & +1 & +1 \\ 1 & -1 & -1 \\ 1 & +1 & -1 \\ 1 & -1 & +1 \\ 1 & +1 & +1 \end{bmatrix}, \qquad \boldsymbol{y} = \begin{bmatrix} 56 \\ 68 \\ 74 \\ 82 \\ 58 \\ 66 \\ 72 \\ 84 \end{bmatrix}.$$

Step 2 — Compute $X'X$.

The entries of $\boldsymbol{X}'\boldsymbol{X}$ are sums over the eight rows: the $(1,1)$ entry is $n = 8$; the $(1,2)$ entry is $\sum x_{i1} = 0$; the $(2,3)$ entry is $\sum x_{i1}x_{i2} = 0$ (each $(\pm 1, \pm 1)$ combination appears twice, so the products cancel); and the diagonal entries are $\sum x_{i1}^2 = \sum x_{i2}^2 = 8$. Therefore

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} 8 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 8 \end{bmatrix}, \qquad (\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} 0.125 & 0 & 0 \\ 0 & 0.125 & 0 \\ 0 & 0 & 0.125 \end{bmatrix}.$$

The inverse is easy here because the coded design makes $\boldsymbol{X}'\boldsymbol{X}$ diagonal; inverting a diagonal matrix just inverts each diagonal entry. Real data rarely give a diagonal $\boldsymbol{X}'\boldsymbol{X}$, which is why the inverse is normally supplied.

Step 3 — Compute $X'y$.

Each entry is a sum of the y_i weighted by one column of $\boldsymbol{X}$:

$$\begin{aligned} \textstyle\sum_i y_i &= 56 + 68 + 74 + 82 + 58 + 66 + 72 + 84 = 560, \\ \textstyle\sum_i x_{i1} y_i &= -56 + 68 - 74 + 82 - 58 + 66 - 72 + 84 = 40, \\ \textstyle\sum_i x_{i2} y_i &= -56 - 68 + 74 + 82 - 58 - 66 + 72 + 84 = 64, \end{aligned}$$

so $\boldsymbol{X}'\boldsymbol{y} = (560,\ 40,\ 64)'$.

Step 4 — Multiply to get $\hat{\boldsymbol{\beta}}$.

By (7.4),

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 0.125 & 0 & 0 \\ 0 & 0.125 & 0 \\ 0 & 0 & 0.125 \end{bmatrix} \begin{bmatrix} 560 \\ 40 \\ 64 \end{bmatrix} = \begin{bmatrix} 70 \\ 5 \\ 8 \end{bmatrix}.$$

The fitted model is $\boxed{\hat{y} = 70 + 5x_1 + 8x_2}$.

Step 5 — Interpret and check the residuals.

Holding data quality fixed, moving from a small to a large network ($x_1 = -1$ to $+1$, a two-unit change) raises the predicted score by $2 \times 5 = 10$ points. Holding model size fixed, moving from uncurated to curated data raises it by $2 \times 8 = 16$ points. On this benchmark, data quality moves the score more than model size does. The fitted values and residuals are:

i	1	2	3	4	5	6	7	8
$\hat{y}_i$	57	67	73	83	57	67	73	83
e_i	-1	$+1$	$+1$	-1	$+1$	-1	-1	$+1$

The residuals sum to zero, as they always must when the model contains an intercept.

Startup lens. The same coded ± 1 design works for a pricing experiment: $x_1 = \pm 1$ for two price points, $x_2 = \pm 1$ for two onboarding flows. Balanced coding makes $\boldsymbol{X}'\boldsymbol{X}$ diagonal, so each effect is estimated independently of the others—the cleanest experiment your data can buy.

Reference. Montgomery & Runger, 6th ed., Section 12-1.4.

7.3 *Estimating σ^2 and the Standard Errors*

The error variance σ^2 measures the scatter of the observations around the true regression surface. As in simple linear regression, it is estimated from the sum of squared residuals, but the divisor changes: fitting p parameters uses up p degrees of freedom, so

$$\hat{\sigma}^2 = \frac{SS_E}{n-p} = MS_E, \qquad SS_E = \sum_{i=1}^{n} e_i^2 = \boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}. \tag{7.5}$$

The matrix expression for SS_E is worth memorizing, because it avoids computing the residuals one by one: $\boldsymbol{y}'\boldsymbol{y} = \sum y_i^2$ is one number, and $\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}$ is a short dot product of two vectors you have already computed.

The estimator $\hat{\boldsymbol{\beta}}$ is a linear function of the responses, so its sampling behavior

follows directly from the model assumptions. It is unbiased, $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, and its variances and covariances are collected in one matrix:

$$\operatorname{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2\boldsymbol{C}, \qquad \boldsymbol{C} = (\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{7.6}$$

The diagonal entries of $\boldsymbol{C}$ carry the information we use most. Writing C_{jj} for the diagonal entry belonging to $\hat{\beta}_j$ (with C_{00} belonging to the intercept), the variance of one coefficient is $\operatorname{Var}(\hat{\beta}_j) = \sigma^2 C_{jj}$, and replacing σ^2 by $\hat{\sigma}^2$ gives the estimated standard error

$$\operatorname{se}(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2\, C_{jj}} = \sqrt{MS_E \cdot C_{jj}}. \tag{7.7}$$

Standard error of a coefficient. $\operatorname{se}(\hat{\beta}_j) = \sqrt{MS_E\, C_{jj}}$, where C_{jj} is the diagonal entry of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ belonging to $\hat{\beta}_j$.

Every t statistic and every confidence interval for a coefficient in this chapter is built from (7.7). Note where each ingredient comes from: MS_E comes from the residuals through (7.5), and C_{jj} comes from the design through $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Noisy data inflate the first factor; a poorly spread or nearly collinear design inflates the second (Section 7.10 returns to this point).

Example 7.2 — Standard errors for the evaluation model

Continue Example 7.1. Estimate σ^2 and compute the standard error of each coefficient.

Solution.

Step 1 — SS_E by the matrix shortcut.

First, $\boldsymbol{y}'\boldsymbol{y} = 56^2 + 68^2 + 74^2 + 82^2 + 58^2 + 66^2 + 72^2 + 84^2 = 39{,}920$. Next,

$$\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} = \begin{bmatrix} 70 & 5 & 8 \end{bmatrix} \begin{bmatrix} 560 \\ 40 \\ 64 \end{bmatrix} = 39{,}200 + 200 + 512 = 39{,}912.$$

By (7.5), $SS_E = 39{,}920 - 39{,}912 = 8$. (Check against the residuals of Example 7.1: eight residuals of ± 1 give $\sum e_i^2 = 8$.)

Step 2 — $\hat{\sigma}^2$.

Here $n = 8$ and $p = k + 1 = 3$, so

$$\hat{\sigma}^2 = \frac{SS_E}{n - p} = \frac{8}{5} = 1.600.$$

Step 3 — Standard errors.

From Example 7.1, $\boldsymbol{C} = \operatorname{diag}(0.125,\ 0.125,\ 0.125)$. By (7.7), all three coefficients share the same standard error:

$$\operatorname{se}(\hat{\beta}_j) = \sqrt{1.600 \times 0.125} = \sqrt{0.2000} = 0.4472.$$

So the fitted model with standard errors is $\hat{y} = 70\,(0.447) + 5\,(0.447)\,x_1 + 8\,(0.447)\,x_2$, and $\boxed{\hat{\sigma}^2 = 1.60,\ \operatorname{se}(\hat{\beta}_j) = 0.447}$. Both slope estimates are many standard errors away from zero, which previews the formal tests of Section 7.7.

Reference. Montgomery & Runger, 6th ed., Section 12-2.1.

7.4 *ANOVA for MLR and the Overall F Test*

The analysis-of-variance identity from simple linear regression carries over unchanged: the total variability of the response splits into a part explained by the regression and a leftover residual part,

$$\underbrace{SS_T}_{n-1 \text{ df}} = \underbrace{SS_R}_{k \text{ df}} + \underbrace{SS_E}_{n-k-1 \text{ df}}, \tag{7.8}$$

with the matrix computing formulas

$$SS_T = \boldsymbol{y}'\boldsymbol{y} - n\bar{y}^2, \qquad SS_R = \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} - n\bar{y}^2, \qquad SS_E = \boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}.$$

What changes from SLR is only the degrees of freedom: the regression now has k of them (one per regressor) and the residual has $n - k - 1 = n - p$. Always verify that the degrees of freedom add up, $k + (n - k - 1) = n - 1$; if they do not, an entry of your table is wrong. The mean squares divide each sum of squares by its degrees of freedom: $MS_R = SS_R/k$ and $MS_E = SS_E/(n-k-1) = \hat{\sigma}^2$. Table 7.2 shows the standard layout.

Source	df	SS	MS	F_0
Regression	k	SS_R	$MS_R = SS_R/k$	MS_R/MS_E
Residual (error)	$n-k-1$	SS_E	$MS_E = SS_E/(n-k-1)$	
Total	$n-1$	SS_T		

Table 7.2. The ANOVA table for multiple linear regression. In SLR, $k = 1$ and the residual df reduce to $n - 2$; the structure is otherwise identical.

The first inference question about a fitted MLR model is whether the model as a whole is worth anything: does at least one regressor have a linear relationship with the response? The hypotheses are

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \qquad \text{vs.} \qquad H_1: \beta_j \neq 0 \text{ for at least one } j.$$

Please note that β_0 is not part of the null hypothesis; the test asks whether the regressors help beyond a constant, not whether the constant is zero. The test statistic compares the explained mean square to the residual mean square:

$$F_0 = \frac{MS_R}{MS_E} = \frac{SS_R/k}{SS_E/(n-k-1)} \sim F_{k,n-k-1} \quad \text{under } H_0, \tag{7.9}$$

and we reject H_0 at level α when $F_0 > f_{\alpha,k,n-k-1}$. Rejecting H_0 means *at least one* regressor matters. It does not mean all of them matter, and it does not say which; that is the job of the individual t tests in Section 7.7.

Procedure 7.2 — The overall F test for significance of regression.

1. State the hypotheses: $H_0: \beta_1 = \cdots = \beta_k = 0$ vs. $H_1:$ at least one $\beta_j \neq 0$. Fix α.
2. Compute (or read off) SS_R, SS_E, and the degrees of freedom k and $n-k-1$; check they sum to $n-1$.
3. Form $MS_R = SS_R/k$, $MS_E = SS_E/(n-k-1)$, and $F_0 = MS_R/MS_E$ by (7.9).
4. Find the critical value $f_{\alpha,k,n-k-1}$ from the F table (or the p-value $P(F_{k,n-k-1} > F_0)$).
5. Reject H_0 if $F_0 > f_{\alpha,k,n-k-1}$. Conclude in context: "the model explains a significant amount of the variability in y" — not "every regressor is significant."

Engineering judgment. An overall F test that fails is an engineering signal, not just a statistical one: the inputs you chose to measure do not move the output. The fix is usually a better list of candidate regressors from process knowledge, not a fancier model on the same columns.

Example 7.3 — ANOVA for the evaluation model

Build the ANOVA table for the model of Example 7.1 and test the significance of regression at $\alpha = 0.05$.

Solution.

Step 1 — Sums of squares.

With $\bar{y} = 560/8 = 70$ and the quantities from Example 7.2:

$$SS_T = \boldsymbol{y}'\boldsymbol{y} - n\bar{y}^2 = 39{,}920 - 8(70)^2 = 39{,}920 - 39{,}200 = 720,$$
$$SS_R = \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} - n\bar{y}^2 = 39{,}912 - 39{,}200 = 712,$$
$$SS_E = SS_T - SS_R = 720 - 712 = 8.$$

Step 2 — The table.

Here $k = 2$ and $n - k - 1 = 5$ (check: $2 + 5 = 7 = n - 1$).

Source	df	SS	MS	F_0
Regression	2	712	356.0	222.5
Residual	5	8	1.6	
Total	7	720		

Table 7.3. ANOVA table for the evaluation-score model of Example 7.1.

Step 3 — Test.

By (7.9), $F_0 = 356.0/1.6 = 222.5$. The critical value with 2 and 5 degrees of freedom is $f_{0.05,2,5} = 5.79$ (Table 7.4). Since $222.5 > 5.79$, we reject H_0. $\boxed{F_0 = 222.5 > 5.79;\ \text{the regression is significant}}$. At least one of model size and data quality has a real linear relationship with the benchmark score.

	ν_1 (numerator df)		
ν_2	1	2	3
4	7.71	6.94	6.59
5	6.61	5.79	5.41
6	5.99	5.14	4.76

Table 7.4. Excerpt of the F table, $f_{0.05,\nu_1,\nu_2}$. The highlighted cell is $f_{0.05,2,5} = 5.79$ used in Example 7.3.

Reference. Montgomery & Runger, 6th ed., Section 12-2.1.

7.5 R^2 and the Adjusted R^2

The coefficient of determination keeps its SLR definition:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}, \tag{7.10}$$

the proportion of the variability in y explained by the model. For the evaluation model, $R^2 = 712/720 = 0.9889$: the two coded regressors explain 98.9% of the score variability.

R^2 has one serious defect in multiple regression. Adding *any* regressor to a model never decreases R^2, even when the regressor is pure noise. So a modeler who keeps adding columns will watch R^2 climb toward 1 regardless of whether the model is getting better or just more cluttered.

Insight. Why R^2 can never decrease. The larger model can always imitate the smaller one: set the new coefficient to zero and keep the old coefficients, and SS_E is unchanged. Least squares then minimizes SS_E over a strictly larger set of candidate models, so the minimized SS_E can only stay the same or fall—and $R^2 = 1 - SS_E/SS_T$ can only stay the same or rise. The increase says nothing about whether the new regressor carries real information.

Adjusted R^2. A version of R^2 that charges the model for each parameter it uses; it can decrease when a useless regressor is added.

The repair is to charge the model for the degrees of freedom it spends. The *adjusted* R^2 replaces each sum of squares by its mean square:

$$R^2_{\text{adj}} = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)} = 1 - \frac{MS_E}{SS_T/(n-1)}. \tag{7.11}$$

The denominator $SS_T/(n-1)$ is fixed for a given data set, so R^2_{adj} moves opposite to MS_E. Adding a useless regressor lowers SS_E a little but lowers the residual degrees of freedom by a whole unit, so MS_E typically *rises* and R^2_{adj} *falls*. That is exactly the behavior we want from a model-comparison score: use R^2 to report how well a given model fits, and use R^2_{adj} to compare models with different numbers of regressors. Figure 7.2 shows the typical pattern as regressors are added one at a time.

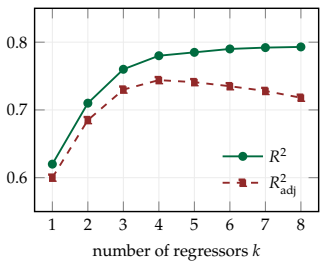

Figure 7.2. As regressors are added one at a time, R^2 never falls, but R^2_{adj} peaks (here at $k = 4$) and then declines: the later regressors do not earn the degrees of freedom they cost.

For the evaluation model of Example 7.3,

$$R^2_{\text{adj}} = 1 - \frac{8/5}{720/7} = 1 - \frac{1.600}{102.86} = 1 - 0.01556 = 0.9844,$$

slightly below $R^2 = 0.9889$, as it always is. A large gap between the two is a warning sign that the model carries regressors it does not need, or that the residual degrees of freedom are very few.

Startup lens. Investor dashboards reward a high R^2, and it is tempting to stuff a churn model with every metric the product logs. R^2_{adj} is the discipline: if adding "daily active minutes" drops it, the metric is decoration, not signal.

Reference. Montgomery & Runger, 6th ed., Section 12-1.4.

7.6 Reading a Regression Summary from Software

In practice the fitting is done by software, and exam problems frequently hand you the software's two standard tables: the ANOVA table and the coefficient summary. You must be able to read both, fill in deliberately blanked entries, and act on them. The coefficient summary has one row per parameter and four standard columns:

- **Coef**: the least squares estimate $\hat{\beta}_j$ from (7.4);
- **SE**: the standard error $\operatorname{se}(\hat{\beta}_j) = \sqrt{MS_E\, C_{jj}}$ from (7.7);
- ***t*-Stat**: the ratio $\hat{\beta}_j / \operatorname{se}(\hat{\beta}_j)$, which tests $H_0 : \beta_j = 0$ (Section 7.7);
- ***p*-value**: the probability, under H_0, of a t_{n-p} value at least as extreme as the observed t-Stat.

Every entry is recomputable from the others: given any two of Coef, SE, and t-Stat, you can recover the third. Exam problems exploit this constantly.

Example 7.4 — Reading output: delivery time from distance, stops, and load

A regional carrier models the delivery time y (hours) of $n = 25$ routes using x_1 = distance (km), x_2 = number of stops, and x_3 = load (tonnes). The software reports the fitted equation $\hat{y} = 3.147 + 0.093\,x_1 + 0.633\,x_2 - 0.302\,x_3$ together with Tables 7.5 and 7.6.

Source	df	SS	MS	F_0
Regression	3	83.343	27.781	134.3
Residual	21	4.343	0.2068	
Total	24	87.686		

Table 7.5. ANOVA table for the delivery-time model, $n = 25$ routes.

Variable	Coef	SE	t-Stat	p-value
Intercept	3.147	0.448	7.03	< 0.001
Distance (x_1)	0.093	0.0146	6.37	< 0.001
Stops (x_2)	0.633	0.132	4.80	< 0.001
Load (x_3)	-0.302	0.146	-2.07	0.051

Table 7.6. Coefficient summary for the delivery-time model. Each t-Stat is Coef divided by SE; each p-value refers to t_{21}.

(a) Confirm the sample size and the parameter count from the table alone. (b) Test the significance of regression at $\alpha = 0.05$. (c) Compute R^2 and R^2_{adj}. (d) Predict the delivery time of a route with distance 30 km, 3 stops, and a 2-tonne load.

Solution.

Step 1 — Read the structure.

Total df $= n - 1 = 24$, so $n = 25$. Regression df $= k = 3$, so $p = k + 1 = 4$ parameters, and residual df $= 25 - 4 = 21$; indeed $3 + 21 = 24$.

Step 2 — Overall F test (Procedure 7.2).

$F_0 = MS_R / MS_E = 27.781 / 0.2068 = 134.3$. The critical value is $f_{0.05,3,21} = 3.07$. Since $134.3 > 3.07$, reject H_0: the model is highly significant — at least one of distance, stops, and load predicts delivery time.

Step 3 — Fit statistics.

By (7.10) and (7.11),

$$R^2 = \frac{83.343}{87.686} = 0.9505, \qquad R^2_{\text{adj}} = 1 - \frac{4.343/21}{87.686/24} = 1 - \frac{0.2068}{3.654} = 0.9434.$$

The three regressors explain 95.0% of the variability in delivery time.

Step 4 — Predict.

$$\hat{y}_0 = 3.147 + 0.093(30) + 0.633(3) - 0.302(2) = 3.147 + 2.790 + 1.899 - 0.604 = 7.232.$$

The predicted delivery time is $\boxed{\hat{y}_0 = 7.23 \text{ hours}}$. If the route actually took 7.0 hours, the residual would be $e = 7.0 - 7.23 = -0.23$ hours: the model overestimated slightly.

Engineering judgment. The negative load coefficient looks wrong at first—heavier trucks should be slower. Operations knowledge resolves it: heavily loaded routes are assigned to senior drivers on highway corridors. The coefficient is the effect of load *holding distance and stops fixed*, and within that slice the driver-assignment policy dominates. Always reconcile signs with how the system actually operates.

Lecture 21. This section is covered by Lecture 21.

Reference. Montgomery & Runger, 6th ed., Sections 12-2.2 and 12-3.1.

Lecture 21 starts here — *Multiple linear regression, part 2*

7.7 Tests and Intervals for Individual Coefficients

The overall F test answers ''is the model as a whole useful?'' The next question is finer: ''does *this particular* regressor contribute, given that the other regressors are already in the model?'' For regressor x_j the hypotheses are $H_0 : \beta_j = \beta_{j,0}$ vs. $H_1 : \beta_j \neq \beta_{j,0}$, where the hypothesized value $\beta_{j,0}$ is 0 in the most common case (''is x_j needed at all?'') but can be any claimed value. The test statistic is the familiar standardized distance:

$$T_0 = \frac{\hat{\beta}_j - \beta_{j,0}}{\mathrm{se}(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta_{j,0}}{\sqrt{MS_E\, C_{jj}}} \;\sim\; t_{n-p} \quad \text{under } H_0, \tag{7.12}$$

and we reject H_0 at level α when $|T_0| > t_{\alpha/2,\,n-p}$ (or use the one-sided variant when H_1 is directional). With $\beta_{j,0} = 0$ this is exactly the t-Stat column of the software output. The matching $100(1-\alpha)\%$ confidence interval is

$$\hat{\beta}_j - t_{\alpha/2,\,n-p}\,\mathrm{se}(\hat{\beta}_j) \;\le\; \beta_j \;\le\; \hat{\beta}_j + t_{\alpha/2,\,n-p}\,\mathrm{se}(\hat{\beta}_j), \tag{7.13}$$

with the usual duality: the CI contains exactly the values $\beta_{j,0}$ that a two-sided level-α test would fail to reject. If 0 is outside the interval, x_j is significant; if a vendor's claimed value is outside the interval, the claim is rejected.

Please pay attention here, because this is a common trap: the degrees of freedom are $n - p = n - k - 1$, not $n - 2$. The value $n - 2$ is the special case $k = 1$. A second trap is the subscript: if the question asks about ''number of stops'' and stops is x_2 in the model, the test is on β_2; match the subscript to the regressor's position, not to the order the question mentions it.

The test in (7.12) is a *marginal* (or partial) test: it measures the contribution of x_j *in the presence of the other regressors in the model*. The same x_j can be significant in one model and insignificant in another, because other regressors may already carry its information. This is not a contradiction; it is what ''holding the others fixed'' means for inference.

Procedure 7.3 — Testing one coefficient and building its CI.

1. Identify which β_j the question is about (match the variable name to its position in the model) and the hypothesized value $\beta_{j,0}$ (0 unless a specific value is claimed).
2. Get $\hat{\beta}_j$ and $\text{se}(\hat{\beta}_j)$: from the output table, or from (7.7) using MS_E and C_{jj}.
3. Compute T_0 by (7.12) and find $t_{\alpha/2,\,n-p}$ with $n-p = n-k-1$ degrees of freedom.
4. Reject H_0 if $|T_0| > t_{\alpha/2,\,n-p}$; or build the CI by (7.13) and check whether $\beta_{j,0}$ lies inside.
5. State the conclusion in context, always with the qualifier "given the other regressors in the model."

df	$t_{0.05}$	$t_{0.025}$	$t_{0.01}$
5	2.015	2.571	3.365
21	1.721	2.080	2.518
26	1.706	2.056	2.479

Table 7.7. Excerpt of the t table. The highlighted cell, $t_{0.025,5} = 2.571$, serves the evaluation model ($n-p=5$); the rows for 21 and 26 df serve the delivery and churn models.

Example 7.5 — Which knobs move churn?

A subscription startup models monthly churn y (% of users lost) on $n = 30$ customer cohorts using $x_1 =$ price (SAR/month), $x_2 =$ median onboarding time (minutes), and $x_3 =$ support tickets per user. Table 7.8 shows the coefficient summary (MS_E-based standard errors, $n-p = 26$ df).

Variable	Coef	SE	t-Stat	p-value
Intercept	4.800	1.520	3.16	0.004
Price (x_1)	0.052	0.021	2.48	0.020
Onboarding (x_2)	0.310	0.078	3.97	< 0.001
Tickets (x_3)	0.145	0.290	0.50	0.621

Table 7.8. Coefficient summary for the churn model, $n = 30$ cohorts, 26 residual df.

(a) At $\alpha = 0.05$, which regressors contribute significantly? (b) Build a 95% CI for the onboarding coefficient β_2. (c) A growth consultant claims "every extra minute of onboarding adds a full percentage point of churn" ($\beta_2 = 1$). Test the claim at $\alpha = 0.05$.

Solution.

Step 1 — Individual tests.

The critical value is $t_{0.025,26} = 2.056$ (Table 7.7). Compare each $|t|$-Stat (equivalently, each p-value to 0.05): price has $|2.48| > 2.056$ (significant), onboarding has $|3.97| > 2.056$ (significant), tickets has $|0.50| < 2.056$ (not significant). Given price and onboarding are in the model, the ticket rate adds no detectable explanatory power.

Step 2 — CI for β_2.

By (7.13),

$$0.310 \pm 2.056 \times 0.078 = 0.310 \pm 0.1604,$$

so $\boxed{0.150 \le \beta_2 \le 0.470}$ (% churn per minute). Zero is outside the interval, in agreement with the t test.

Step 3 — Test the consultant's claim.

By (7.12) with $\beta_{2,0} = 1$:

$$T_0 = \frac{0.310 - 1}{0.078} = \frac{-0.690}{0.078} = -8.846.$$

Since $|-8.846| > 2.056$, reject $H_0\colon \beta_2 = 1$. Equivalently, $1 \notin (0.150,\ 0.470)$. The data flatly contradict the claim: onboarding friction matters, but at roughly a third of a point per minute, not a full point. $\boxed{\text{Reject } \beta_2 = 1}$

Startup lens. Notice what Example 7.5 does *not* say: that support tickets are irrelevant to churn. It says tickets add nothing *once price and onboarding are known*—perhaps because slow onboarding generates the tickets. Marginal tests rank knobs given the model, not in isolation.

Table 7.9 collects the question-to-test mapping. On an exam, identify the question type first, then pick the statistic.

If the question says...	**Use...**
"Is the model significant?"	overall F (7.9): $H_0\colon \beta_1 = \cdots = \beta_k = 0$
"Is x_j significant?"	t test (7.12): $H_0\colon \beta_j = 0$
"Is the effect of x_j equal to c?"	t test (7.12): $H_0\colon \beta_j = c$
"Is the relationship with x_j inverse?"	one-sided t: $H_0\colon \beta_j = 0$, $H_1\colon \beta_j < 0$
"Do $x_{j_1}, \ldots, x_{j_r}$ (a group) contribute?"	partial F test (7.14)

Table 7.9. Choosing the right hypothesis test in multiple regression.

7.8 The Partial F Test: Extra Sum of Squares

Reference. Montgomery & Runger, 6th ed., Section 12-2.2.

Partial F test. A test of whether a *group* of regressors contributes to the model, given that the remaining regressors are already included. Also called the extra sum of squares method.

Extra sum of squares. $SS_R(\text{full}) - SS_R(\text{reduced})$: the additional variability explained by the tested group of regressors.

Between the overall F (all slopes at once) and the individual t (one slope) sits a frequent practical question: does a *group* of regressors contribute, beyond what the rest of the model already explains? For example, with five regressors, are x_1, x_2, and x_4 needed given that x_3 and x_5 stay? The hypotheses for a group of r coefficients are

$$H_0 : \beta_{j_1} = \beta_{j_2} = \cdots = \beta_{j_r} = 0 \qquad \text{vs.} \qquad H_1 : \text{at least one of them} \neq 0.$$

The method compares two fitted models. The *full model* B_1 contains all k regressors and has regression sum of squares $SS_R(B_1)$. The *reduced model* B_2 drops the r regressors under test, keeping only the ones not being questioned, and has $SS_R(B_2)$. Because the full model can always do at least as well, $SS_R(B_1) \geq SS_R(B_2)$, and the difference is the *extra sum of squares*: the variability explained by the tested group over and above the reduced model. Scaled by its degrees of freedom and by the full model's MS_E, it becomes the test statistic

$$F_0 = \frac{\left[SS_R(B_1) - SS_R(B_2)\right]/r}{MS_E} \sim F_{r,n-p} \quad \text{under } H_0, \tag{7.14}$$

where r is the number of coefficients being tested, $p = k + 1$ counts the parameters of the *full* model, and MS_E is the *full* model's residual mean square. Reject H_0 when $F_0 > f_{\alpha,r,n-p}$. Two bookkeeping rules prevent most errors: B_1 is always the bigger model, and the numerator degrees of freedom count the coefficients *tested*, not the ones retained.

Procedure 7.4 — The partial F test (extra sum of squares method).

1. Write the hypotheses: H_0 sets the r tested coefficients to zero; the remaining regressors are not questioned.
2. Identify the two models: full (B_1) with all k regressors, reduced (B_2) with only the untested ones. To compute SS_R for each model, use only that model's regressor columns plus the response column.
3. Compute the extra sum of squares $SS_R(B_1) - SS_R(B_2)$ and the full model's

MS_E.

4. Form F_0 by (7.14); degrees of freedom are r (numerator) and $n - p$ from the full model (denominator).
5. Reject H_0 if $F_0 > f_{\alpha, r, n-p}$: the group adds significant explanatory power beyond the reduced model.

Example 7.6 — Do load and traffic earn their place?

The carrier of Example 7.4 expands its study: for $n = 25$ routes it fits a full model with $k = 4$ regressors, $x_1 =$ distance, $x_2 =$ stops, $x_3 =$ load, and $x_4 =$ a traffic congestion index. The full-model ANOVA gives $SS_R(B_1) = 320.0$ and $SS_E = 80.0$ with $SS_T = 400.0$. A reduced model using only distance and stops gives $SS_R(B_2) = 280.0$. At $\alpha = 0.05$, do load and traffic together contribute significantly, given that distance and stops are already in the model?

Solution.

Step 1 — Hypotheses and models (Procedure 7.4).

H_0: $\beta_3 = \beta_4 = 0$ vs. H_1: at least one of $\beta_3, \beta_4 \neq 0$. Full model B_1: x_1, x_2, x_3, x_4. Reduced model B_2: x_1, x_2 only. Here $r = 2$ coefficients are tested.

Step 2 — Full-model MS_E.

The full model has $p = 5$ parameters, so its residual df are $n - p = 25 - 5 = 20$ and

$$MS_E = \frac{SS_E}{n - p} = \frac{80.0}{20} = 4.000.$$

Step 3 — Extra sum of squares and F_0.

The extra sum of squares is $320.0 - 280.0 = 40.0$, so by (7.14)

$$F_0 = \frac{40.0/2}{4.000} = \frac{20.00}{4.000} = 5.000.$$

Step 4 — Decision.

The critical value is $f_{0.05,2,20} = 3.49$. Since $5.000 > 3.49$, reject H_0. $F_0 = 5.00 > 3.49$; keep
Together, load and the traffic index explain a significant amount of delivery-

time variability beyond distance and stops. Note that this is a statement about the *pair*; whether each one individually earns its place is a question for the t tests.

Safety lens. Partial F tests are the statistical backbone of ablation studies. ''Do the three red-team features add anything once the standard telemetry is in the model?'' is exactly $H_0\colon \beta_{j_1} = \beta_{j_2} = \beta_{j_3} = 0$ tested against the full pipeline.

Reference. Montgomery & Runger, 6th ed., Sections 12-3.2 and 12-4.

7.9 *Intervals for the Mean Response and for a New Observation*

Once the model is fitted and vetted, we use it at a specific operating point. Collect the point's regressor values, with a leading 1 for the intercept, into the vector $\boldsymbol{x}_0 = (1, x_{01}, x_{02}, \ldots, x_{0k})'$. The point estimate is $\hat{y}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_k x_{0k}$, and, exactly as in simple linear regression, there are two different interval questions with two different answers.

For the *mean response* at $\boldsymbol{x}_0$, the $100(1-\alpha)\%$ confidence interval is

$$\hat{y}_0 \pm t_{\alpha/2,\,n-p}\sqrt{\hat{\sigma}^2\,\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}, \tag{7.15}$$

and for a *single new observation* at $\boldsymbol{x}_0$, the $100(1-\alpha)\%$ prediction interval is

$$\hat{y}_0 \pm t_{\alpha/2,\,n-p}\sqrt{\hat{\sigma}^2\left(1 + \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0\right)}. \tag{7.16}$$

The scalar $\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0$ generalizes the SLR quantity $1/n + (x_0 - \bar{x})^2/S_{xx}$: it is small near the center of the observed design and grows toward the edges, so both intervals widen as $\boldsymbol{x}_0$ moves away from the data.

Insight. Why the prediction interval carries the extra ''1.'' The CI answers ''where is the *average* response at this operating point?'' and its width reflects only the uncertainty in the fitted surface. The PI answers ''where will *one* future observation land?'' and a single observation scatters around the surface with variance σ^2 even if the surface were known exactly. The two variances add: estimation variance plus σ^2, giving the $1+$ term in (7.16). The PI is therefore always wider, and unlike the CI it does not shrink to zero as n grows.

On an exam you will either be handed the standard error $\mathrm{se}(\hat{Y}_0)$ directly, or be given $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ and asked to compute the quadratic form yourself. Read the notation carefully: if the given SE ''already includes the $+1$,'' it is a prediction standard error, and multiplying by another $+1$ double-counts.

Example 7.7 — CI and PI for the evaluation model

For the model of Examples 7.1–7.3 ($\hat{y} = 70 + 5x_1 + 8x_2$, $\hat{\sigma}^2 = 1.600$, $n - p = 5$, $\boldsymbol{C} = \operatorname{diag}(0.125, 0.125, 0.125)$), the team plans a run with a large model on curated data: $\boldsymbol{x}_0 = (1,\ +1,\ +1)'$. Find the 95% CI for the mean score at this configuration and the 95% PI for the score of one future run.

Solution.

Step 1 — Point estimate and quadratic form.

$\hat{y}_0 = 70 + 5(1) + 8(1) = 83$. Because $\boldsymbol{C}$ is diagonal,

$$\boldsymbol{x}_0' \boldsymbol{C}\, \boldsymbol{x}_0 = (1)^2(0.125) + (1)^2(0.125) + (1)^2(0.125) = 0.375.$$

Step 2 — CI for the mean response.

With $t_{0.025,\,5} = 2.571$ (Table 7.7), by (7.15):

$$83 \pm 2.571\sqrt{1.600 \times 0.375} = 83 \pm 2.571\sqrt{0.6000} = 83 \pm 2.571(0.7746) = 83 \pm 1.991.$$

The 95% CI for the mean score is $\boxed{(81.01,\ 84.99)}$.

Step 3 — PI for one new run.

By (7.16):

$$83 \pm 2.571\sqrt{1.600\,(1 + 0.375)} = 83 \pm 2.571\sqrt{2.200} = 83 \pm 2.571(1.483) = 83 \pm 3.813.$$

The 95% PI is $\boxed{(79.19,\ 86.81)}$. The PI is roughly twice as wide as the CI: the mean score at this configuration is pinned down to about ± 2 points, but any single run can still land anywhere in a ± 3.8-point band.

Reference. Montgomery & Runger, 6th ed., Section 12-6.

Multicollinearity. Strong linear relationships among the regressors themselves, which make individual coefficients unstable and their standard errors large.

7.10 *Multicollinearity and the Variance Inflation Factor*

The interpretation ''the effect of x_j holding the others fixed'' quietly assumes that the data contain variation in x_j *while the others are fixed*. When two regressors move together in the data — distance and travel time, model size and training cost, price and plan tier — the data contain almost no such variation, and the model cannot tell their effects apart. This condition is called *multicollinearity*. It does not bias the fitted surface as a whole, and predictions inside the data region

remain fine; what it destroys is the stability of the *individual* coefficients. Small changes in the data can swing $\hat{\beta}_j$ wildly, standard errors balloon, coefficients take signs that contradict engineering sense, and it becomes possible for the overall F to be highly significant while every individual t is insignificant.

Insight. Where the instability comes from. $\mathrm{Var}(\hat{\beta}_j) = \sigma^2 C_{jj}$, and C_{jj} is a diagonal entry of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. When one regressor column is nearly a linear combination of the others, $\boldsymbol{X}'\boldsymbol{X}$ is nearly singular, and inverting a nearly singular matrix produces huge entries—exactly in the positions of the entangled regressors. The design, not the noise, is what inflates the variance.

Variance inflation factor. $\mathrm{VIF}_j = 1/(1 - R_j^2)$: the factor by which multicollinearity multiplies $\mathrm{Var}(\hat{\beta}_j)$, where R_j^2 comes from regressing x_j on the other regressors.

The standard diagnostic is the *variance inflation factor*. Regress regressor x_j on all the *other* regressors and record the resulting coefficient of determination R_j^2; then

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2}. \tag{7.17}$$

VIF_j is exactly the factor by which $\mathrm{Var}(\hat{\beta}_j)$ is inflated relative to a design where x_j is unrelated to the other regressors. If x_j is orthogonal to the rest, $R_j^2 = 0$ and $\mathrm{VIF}_j = 1$ (no inflation); as $R_j^2 \to 1$ the factor explodes. Common working rules: $\mathrm{VIF}_j > 5$ deserves attention, and $\mathrm{VIF}_j > 10$ signals serious multicollinearity. Figure 7.3 shows how sharply the factor rises.

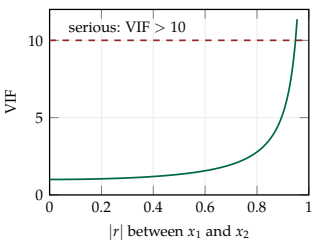

Figure 7.3. With two regressors, $R_1^2 = r^2$, so $\mathrm{VIF} = 1/(1 - r^2)$. The inflation is mild until $|r| \approx 0.8$ and then rises steeply: $|r| = 0.95$ already gives $\mathrm{VIF} = 10.3$.

For two regressors, R_1^2 is just the squared sample correlation between x_1 and x_2, so the computation is quick. Suppose the churn model of Example 7.5 had used both "price" and "plan tier," with sample correlation $r = 0.95$ between them. Then $R_1^2 = 0.95^2 = 0.9025$ and

$$\mathrm{VIF} = \frac{1}{1 - 0.9025} = \frac{1}{0.0975} = 10.26,$$

so each coefficient's variance is about ten times what it would be in a clean design, and each standard error is inflated by $\sqrt{10.26} = 3.203$. Effects that would be clearly significant become statistically invisible. Figure 7.4 shows the practical consequence: with a collinear design, refitting the same model on repeated samples from the same process produces wildly different coefficient estimates.

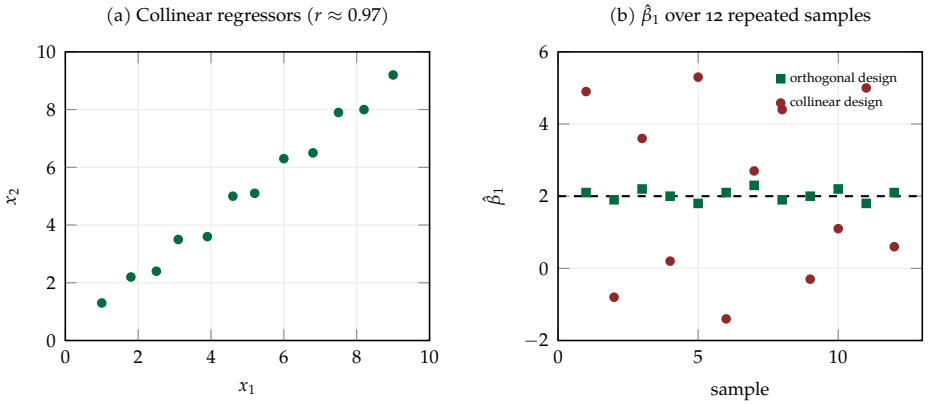

Figure 7.4. The cost of multicollinearity. Panel (a): two regressors that track each other almost perfectly. Panel (b): the estimate $\hat{\beta}_1$ (true value 2, dashed) refitted on twelve independent samples of the same size. With an orthogonal design the estimates cluster tightly; with the collinear design of panel (a) they swing between -1.4 and 5.3, sometimes even changing sign.

What to do about multicollinearity. The choices, in rough order of preference: collect data where the entangled regressors vary independently (a designed experiment, as in Example 7.1, has VIF $= 1$ by construction); drop one of the redundant regressors; combine them into a single meaningful index (cost per km rather than cost and km separately); or accept the entanglement and refuse to interpret the individual coefficients, using the model only for prediction inside the data region.

Safety lens. Evaluation suites often log dozens of correlated capability metrics. A safety report that ranks "which capability drives the risk score" from a collinear regression is quoting noise; check the VIFs before any coefficient is allowed to guide policy.

Reference. Montgomery & Runger, 6th ed., Section 12-6.

Indicator variable. A regressor coded 0 or 1 to represent a category, letting one regression model hold both groups at once.

7.11 *Indicator Variables*

Regression is not restricted to measured quantities. A categorical input with two levels — annual vs. monthly plan, curated vs. uncurated data, highway vs. city route — enters the model as an *indicator variable*, coded $z = 1$ for one level and $z = 0$ for the other:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 z + \varepsilon. \tag{7.18}$$

The single model (7.18) contains two parallel regression lines. For the $z = 0$ group the mean response is $\beta_0 + \beta_1 x_1$; for the $z = 1$ group it is $(\beta_0 + \beta_2) + \beta_1 x_1$. The coefficient β_2 is the vertical gap between the groups at any fixed x_1, and the standard t test of $H_0\colon \beta_2 = 0$ from (7.12) asks whether the groups genuinely differ once x_1 is accounted for. A category with m levels needs $m - 1$ indicators,

one level serving as the baseline; using m indicators together with an intercept makes the columns of $\boldsymbol{X}$ linearly dependent and $\boldsymbol{X}'\boldsymbol{X}$ singular.

Example 7.8 — Churn for monthly vs. annual plans

The startup of Example 7.5 compares plan types. For $n = 24$ cohorts it fits churn (%) on price x_1 (SAR/month) and an indicator z (1 for annual plans, 0 for monthly), obtaining

$$\hat{y} = 6.20 + 0.050\,x_1 - 3.10\,z, \qquad \operatorname{se}(\hat{\beta}_2) = 0.85, \qquad n - p = 21.$$

(a) Write the fitted line for each plan type and interpret $\hat{\beta}_2$. (b) At $\alpha = 0.05$, do annual plans churn differently from monthly plans at the same price?

Solution.

Step 1 — Two parallel lines.

Monthly ($z = 0$): $\hat{y} = 6.20 + 0.050\,x_1$. Annual ($z = 1$): $\hat{y} = 3.10 + 0.050\,x_1$. At *any* given price, annual-plan cohorts are predicted to churn 3.10 percentage points less than monthly cohorts; the price sensitivity (0.050 per SAR) is assumed common to both (Figure 7.5).

Step 2 — Test the gap.

By (7.12) with $\beta_{2,0} = 0$:

$$T_0 = \frac{-3.10}{0.85} = -3.647,$$

and $t_{0.025,\,21} = 2.080$. Since $|-3.647| > 2.080$, reject H_0. Annual plans churn significantly less, by 3.10 points

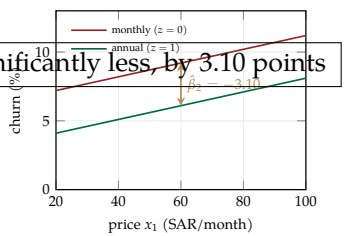

Figure 7.5. One indicator variable, two parallel lines. The gap between the lines is the indicator coefficient, constant across all prices.

Reference. Montgomery & Runger, 6th ed., Section 12-6.

7.12 *Interaction and Polynomial Terms*

Because ''linear'' means linear in the β's, the MLR machinery fits curved and interacting relationships without any new theory. Two constructions cover most engineering needs.

Interaction terms. If the effect of x_1 on the response depends on the level of x_2,

Interaction. A product term $x_1 x_2$ in the model; the effect of one regressor then depends on the level of the other.

add their product as a new regressor:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12}\, x_1 x_2 + \varepsilon. \tag{7.19}$$

Setting $x_3 = x_1 x_2$ turns (7.19) into an ordinary three-regressor linear model, fitted by (7.4) as usual. To read it, group the terms in x_1: the slope of the response with respect to x_1 is $\beta_1 + \beta_{12} x_2$, a slope that changes with x_2. When $x_2 = z$ is an indicator, the model draws two lines with different intercepts *and* different slopes (Figure 7.6), relaxing the parallel-lines assumption of Section 7.11. The t test on β_{12} asks whether the interaction is real.

Polynomial terms. If a residual plot shows curvature, add powers of the regressor:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon. \tag{7.20}$$

Setting $x_1 = x$ and $x_2 = x^2$ again gives a standard two-regressor linear model (Figure 7.7). The same idea handles transformed responses: fitting $\ln y = \beta_0 + \beta_1 x + \varepsilon$ is a linear regression with response $\ln y$. One caution on transformed models: predictions must be back-transformed as a whole. From $\widehat{\ln y} = 2.0 + 0.50x$ the prediction at $x = 3$ is $\hat{y} = e^{2.0+0.50(3)} = e^{3.5} = 33.12$, *not* $e^{2.0} + e^{0.50}x$; the exponential of a sum is a product, not a sum of exponentials.

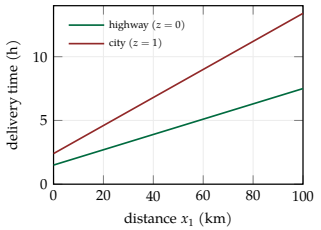

Figure 7.6. An interaction between distance and a route-type indicator: the lines have different slopes, so each extra kilometer costs more time in the city than on the highway.

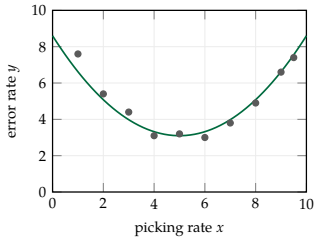

Figure 7.7. A quadratic model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$ fitted by ordinary linear least squares: the model is linear in the β's even though the curve is not a line.

Reference. Montgomery & Runger, 6th ed., Sections 12-6.1 and 12-6.3.

7.13 *Building a Model Step by Step*

The tools are now all on the table; what remains is the order in which to use them. There is no automatic recipe that replaces judgment, but the following sequence keeps the analysis honest and is the one we will use in this course.

Procedure 7.5 — Building and vetting an MLR model.

1. **Choose candidates from knowledge, not from the data.** List the regressors that the physics, the process, or the business logic says should matter, including any indicators, interactions, or polynomial terms suggested by how the system works.
2. **Fit the full candidate model** by (7.4) and run the overall F test (Proce-

dure 7.2). If the model as a whole is not significant, stop and rethink the candidate list.

3. **Check for multicollinearity**: compute the VIFs by (7.17). Resolve any VIF > 10 (drop, combine, or redesign) before interpreting individual coefficients.
4. **Assess terms one at a time**: examine each t statistic (Procedure 7.3); test doubtful *groups* with the partial F test (Procedure 7.4). Remove regressors that do not contribute, one at a time, refitting after each removal — coefficients and t values change every time the model changes.
5. **Compare competing models with R^2_{adj}** by (7.11), never with plain R^2.
6. **Check the residuals** of the surviving model (Section 7.14). Patterns send you back to step 1 with a new candidate term.
7. **Report** the fitted equation with standard errors, the ANOVA table, R^2_{adj}, and the checks performed — not just the coefficients.

The refit-after-each-removal rule in step 4 deserves emphasis. Because the t tests are marginal, removing one regressor changes every other regressor's estimate, standard error, and t value. Deleting all the insignificant regressors in one sweep is wrong: two regressors can be individually insignificant only because each is standing in the other's shadow, and removing both at once throws away real signal. Remove the weakest term, refit, and look again.

As an illustration, apply steps 4 and 5 to the churn model of Example 7.5. The full model ($k = 3$) has $R^2 = 0.7420$, so $R^2_{\text{adj}} = 1 - (1 - 0.7420)\frac{29}{26} = 0.7122$. The tickets coefficient was the clear non-contributor ($T_0 = 0.50$). Refitting without it ($k = 2$) gives $R^2 = 0.7395$ — lower, as it must be — but $R^2_{\text{adj}} = 1 - (1 - 0.7395)\frac{29}{27} = 0.7202$, which is *higher*. The smaller model explains essentially the same variability while spending one fewer degree of freedom, so it is the better model. Plain R^2 would have voted, wrongly, for the larger one.

Engineering judgment. Step 1 is where engineering earns its keep. A regression can only choose among the columns it is given; it cannot invent the missing regressor. If dwell time obviously depends on crane availability and no one logged crane availability, no amount of model search will fix the data set.

Reference. Montgomery & Runger, 6th ed., Section 12-5.

7.14 Residual Diagnostics for MLR

Standardized residual. $d_i = e_i/\sqrt{MS_E}$; values beyond ± 2 flag potential outliers, beyond ± 3 almost certain ones.

Everything inferential in this chapter — the F and t distributions, the CIs and PIs — leans on the error assumptions, so the fitted model must be checked before its intervals are trusted. The diagnostic toolkit carries over from simple linear regression with one addition.

Compute the residuals $e_i = y_i - \hat{y}_i$ and the standardized residuals $d_i = e_i/\sqrt{MS_E}$. Then plot: (i) a normal probability plot of the residuals, to check normality; (ii) e_i against the fitted values $\hat{y}_i$; and (iii) — the MLR addition — e_i against *each regressor* x_j separately. A healthy model shows structureless horizontal bands in all of them (Figure 7.8, left). The two classic pathologies are the *funnel*, a spread that grows with the fitted value and signals nonconstant variance (a transformation of y, such as $\ln y$ or $\sqrt{y}$, usually helps), and the *curve*, a U- or arch-shaped pattern signaling a missing quadratic or interaction term in whichever plot shows it. A pattern in the plot of e_i vs. x_3, for example, points specifically at a missing x_3^2 or $x_3 x_m$ term.

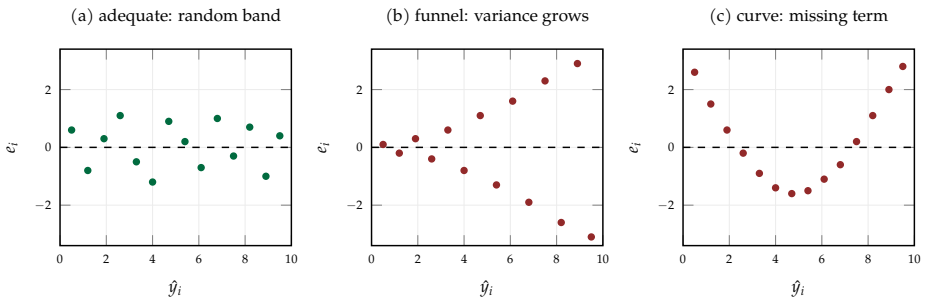

Figure 7.8. Reading residual-vs-fitted plots in MLR. (a) A random horizontal band: the model is adequate. (b) A funnel: the error variance is not constant; transform the response. (c) A systematic curve: the mean model is missing a quadratic or interaction term. In MLR, make the same plot against each regressor x_j as well; the plot that shows the pattern tells you which term is missing.

Influential observation. A data point whose removal would substantially change the fitted coefficients; the dangerous case combines a large residual with high leverage.

Finally, look for unusual points, which come in two distinct kinds. An *outlier* has a large residual ($|d_i| > 2$): the point sits far from the fitted surface. A *high-leverage* point has extreme regressor values: it sits far from the other points in the x-space and pulls the fitted surface toward itself, so its own residual can look deceptively small. A point that is both — extreme in x *and* poorly fitted — is an *influential observation*, the most dangerous case: removing it substantially changes $\hat{\boldsymbol{\beta}}$. When a point is flagged, investigate it as an engineer before touching it as a statistician: a transcription error is corrected, a legitimate extreme condition

is kept and reported, and in either case the model should be refitted with and without the point to show how much the conclusions depend on it.

Safety lens. In disengagement data, the influential points are often the most informative ones: a single extreme-weather run can dominate the fit. Deleting it makes the dashboard prettier and the safety case weaker. Report both fits.

Chapter Summary

Multiple linear regression extends the fitted line to k regressors, $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$, with each coefficient read as the effect of its regressor *holding the others fixed*. Stacking the n observations gives the matrix form $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, and one formula, $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, replaces all the scalar SLR estimators. The ANOVA identity $SS_T = SS_R + SS_E$ survives with degrees of freedom k and $n - p$, where $p = k + 1$; the overall F tests all slopes at once, the t statistic $\hat{\beta}_j / \operatorname{se}(\hat{\beta}_j)$ tests one coefficient given the rest, and the partial F tests a group through the extra sum of squares. R^2 never falls when regressors are added, so competing models are compared with R^2_{adj}. Multicollinearity inflates coefficient variances by the factor $\text{VIF}_j = 1/(1 - R_j^2)$; indicators, interactions, and polynomial terms extend the model while keeping it linear in the parameters; and residual plots — against the fitted values and against each regressor — must be checked before any interval is trusted. Table 7.10 collects the formulas.

Quantity	Formula	Eq.
MLR model	$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$	(7.1)
Matrix form	$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$	(7.2)
Least squares	$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$	(7.4)
Error variance	$\hat{\sigma}^2 = SS_E/(n-p) = MS_E,\ p = k+1$	(7.5)
Coefficient SE	$\operatorname{se}(\hat{\beta}_j) = \sqrt{MS_E\, C_{jj}},\ \boldsymbol{C} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$	(7.7)
Overall F	$F_0 = MS_R/MS_E \sim F_{k,n-k-1}$	(7.9)
Adjusted R^2	$R^2_{\text{adj}} = 1 - \dfrac{SS_E/(n-p)}{SS_T/(n-1)}$	(7.11)
Coefficient t	$T_0 = (\hat{\beta}_j - \beta_{j,0})/\operatorname{se}(\hat{\beta}_j) \sim t_{n-p}$	(7.12)
Coefficient CI	$\hat{\beta}_j \pm t_{\alpha/2,n-p}\operatorname{se}(\hat{\beta}_j)$	(7.13)
Partial F	$F_0 = \dfrac{[SS_R(B_1) - SS_R(B_2)]/r}{MS_E} \sim F_{r,n-p}$	(7.14)
CI mean response	$\hat{y}_0 \pm t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2\, \boldsymbol{x}_0'\boldsymbol{C}\boldsymbol{x}_0}$	(7.15)

Table 7.10. Key formulas of Chapter 7 at a glance.

Exercises

Model and interpretation

Exercise 7.1. A carrier fits $\hat{y} = 2.1 + 0.08\,x_1 + 0.55\,x_2$, where y is delivery time (hours), x_1 is distance (km), and x_2 is the number of stops. The correct reading of the coefficient 0.55 is:

(A) every delivery with one more stop takes exactly 0.55 hours longer.

(B) the predicted delivery time rises by 0.55 hours per additional stop, for routes of the same distance.

(C) stops explain 55% of the variability in delivery time.

(D) the predicted time of a one-stop route is 0.55 hours.

Exercise 7.2. For each model, state whether it can be fitted by linear least squares (7.4), possibly after a transformation or a relabeling of regressors, and name the trick used: (a) $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon$; (b) $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon$; (c) $Y = \beta_0\, e^{\beta_1 x}\, \varepsilon$; (d) $Y = \beta_0 + e^{\beta_1} x + \varepsilon$.

Matrix least squares

Exercise 7.3. A red-team study records, for $n = 5$ testing campaigns, $x_1 =$ prompts issued (hundreds), $x_2 =$ attack categories covered, and $y =$ distinct defects found:

$$(x_1, x_2, y) \in \{(2, 10, 15),\ (4, 15, 22),\ (6, 12, 20),\ (8, 18, 28),\ (10, 20, 35)\}.$$

Write the design matrix $\boldsymbol{X}$ and the response vector $\boldsymbol{y}$ for the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, state the matrix equation for $\hat{\boldsymbol{\beta}}$, and give the dimensions of $\boldsymbol{X}'\boldsymbol{X}$ and $\boldsymbol{X}'\boldsymbol{y}$.

Exercise 7.4. A warehouse studies picking time y (seconds above standard) using coded shelf height x_1 and coded load x_2. Four picks give $(x_1, x_2, y) =$

$(1,0,12)$, $(0,1,9)$, $(-1,0,6)$, $(0,-1,3)$, and

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} 0.25 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.5 \end{bmatrix}.$$

(a) Compute $\hat{\boldsymbol{\beta}}$ and write the fitted equation. (b) Compute all four residuals, verify they sum to zero, and estimate σ^2.

Inference on coefficients and the model

Exercise 7.5. A warehouse automation team fits a picking-time model with $k = 3$ regressors on $n = 30$ shifts. The software prints an incomplete ANOVA table: Regression — df 3, $SS = 234.0$, $MS = 78.0$; Error — $MS = 3.0$; Total — df 29. (a) Complete the table (error df, SS_E, SS_T, F_0) and test the significance of regression at $\alpha = 0.05$, given $f_{0.05,3,26} = 2.98$. (b) Compute $\hat{\sigma}^2$, R^2, and R^2_{adj}.

Exercise 7.6. An LLM evaluation model uses $k = 3$ regressors on $n = 28$ runs. The row for $x_3 =$ context length (thousands of tokens) reads: Coef $= -0.045$, SE $= 0.019$. (a) At $\alpha = 0.05$, does context length contribute to the model? (b) Build a 95% CI for β_3. (c) A blog post claims long context degrades the score at $\beta_3 = -0.10$ per thousand tokens. Test the claim at $\alpha = 0.05$.

Exercise 7.7. An AI lab models benchmark score on $n = 22$ training runs with $x_1 =$ model size, $x_2 =$ data quality index, and $x_3 =$ fine-tuning epochs. The full model gives $SS_R(B_1) = 150.0$ and $SS_E = 38.0$. A reduced model with x_1 alone gives $SS_R(B_2) = 120.0$. At $\alpha = 0.05$ ($f_{0.05,2,18} = 3.55$), do data quality and fine-tuning epochs together add explanatory power beyond model size?

Model building and multicollinearity

Exercise 7.8. A startup compares two demand-forecast models on the same $n = 28$ weeks of data. Model A uses $k = 3$ regressors and has $R^2 = 0.78$; Model B uses $k = 6$ and has $R^2 = 0.80$. Which model should be preferred? Support the choice with R^2_{adj}, and explain why comparing the plain R^2 values is not a fair contest.

Exercise 7.9. A delivery-time model includes both route distance (x_1, km) and expected travel time from the map service (x_2, min). Their sample correlation is $r = 0.98$. (a) Compute the VIF for these regressors and the factor by which each standard error is inflated. (b) The overall F is highly significant, yet neither t statistic is. Explain how this happens and propose two remedies.

Exam-style problems

Exercise 7.10. A rail terminal models wagon-set unloading time y (hours) on $n = 18$ train arrivals using $x_1 =$ number of wagons and $x_2 =$ total load (tonnes). The fit gives $SS_T = 250.0$ and $SS_R = 210.0$; the coefficient row for load reads Coef $= 1.85$, SE $= 0.60$. Use $\alpha = 0.05$ ($f_{0.05,2,15} = 3.68$, $t_{0.025,15} = 2.131$).

1. **[4 pt]** Complete the ANOVA table: all df, SS_E, both mean squares, and F_0.
2. **[4 pt]** Test the significance of regression and state the conclusion in context.
3. **[4 pt]** Compute R^2 and R^2_{adj}.
4. **[5 pt]** Test whether load contributes to the model, and give a 95% CI for its coefficient.

Exercise 7.11. A guardrail team studies the false-positive rate y (%) of a content filter as a function of the coded decision threshold x_1 and the coded prompt length x_2. Six calibration runs give

$$(x_1, x_2, y):\ (-1,-1,10),\ (1,-1,12),\ (-1,1,5),\ (1,1,9),\ (0,0,8.5),\ (0,0,9.5),$$

and the design yields $(\boldsymbol{X}'\boldsymbol{X})^{-1} = \operatorname{diag}(1/6,\ 1/4,\ 1/4)$. Use $\alpha = 0.05$ ($t_{0.025,3} = 3.182$).

1. **[5 pt]** Compute $\boldsymbol{X}'\boldsymbol{y}$ and $\hat{\boldsymbol{\beta}}$, and write the fitted equation.
2. **[3 pt]** Compute SS_E and $\hat{\sigma}^2$.
3. **[4 pt]** Test $H_0\colon \beta_1 = 0$ vs. $H_1\colon \beta_1 \neq 0$.
4. **[5 pt]** At $\boldsymbol{x}_0 = (1,\ 1,\ -1)'$, give the 95% CI for the mean false-positive rate and the 95% PI for one future run.

Assessing outside recommendations

Exercise 7.12. An AI assistant reviews the delivery model of Example 7.4 and recommends: ''Drop distance from the model. Its coefficient is only 0.093, tiny next to the stops coefficient of 0.633, so distance barely matters.'' Route distances in the data run from 5 to 400 km, and $\text{se}(\hat{\beta}_1) = 0.0146$. Find the two errors in this reasoning, explain them, and give the correct assessment of the distance regressor.

Exercise 7.13. A consultant upgrades a startup's churn model, fitted on $n = 20$ cohorts, from $k = 2$ regressors ($R^2 = 0.71$) to $k = 7$ regressors ($R^2 = 0.74$) and reports: ''Explained variance is up three points, so the new model is better. All seven coefficients were tested against $t_{0.025,\,18} = 2.101$ and five passed.'' Identify the two statistical errors, correct the numbers, and state which model you would keep.

In-Class Activity 7.1: Two Regressors, One Port

Week 10 — Lecture 20

Work in teams of 3–4. Write your reasoning, not just the final number. One submission per team, all names on it.

Your team models container dwell time at a port terminal. For each sampled container, y is the dwell time (hours above the tariff-free window), x_1 is the coded vessel size (-1 feeder, $+1$ deep-sea), and x_2 is the coded customs channel (-1 green, $+1$ red). Four sampled containers give

$$(x_1, x_2, y):\ (-1, -1,\ 8),\ (+1, -1,\ 14),\ (-1, +1,\ 18),\ (+1, +1,\ 24).$$

Part 1 — Build the pieces

Write the design matrix $\boldsymbol{X}$ and the response vector $\boldsymbol{y}$ for the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$. State the dimensions of $\boldsymbol{X}$, $\boldsymbol{X}'\boldsymbol{X}$, and $\boldsymbol{X}'\boldsymbol{y}$, and say in one sentence why the column of 1's must be there.

Part 2 — Estimate by hand

Compute $\boldsymbol{X}'\boldsymbol{y}$. Using $(\boldsymbol{X}'\boldsymbol{X})^{-1} = \operatorname{diag}(0.25,\ 0.25,\ 0.25)$, compute $\hat{\boldsymbol{\beta}}$ and write the fitted equation. Then interpret $\hat{\beta}_2$ in one sentence that includes the words "holding ... fixed."

ees-of-freedom detective

pposes adding two more regressors (crane assignment and shipping line) to this same four-
et and running the overall F test. Using only degrees-of-freedom bookkeeping ($p = k + 1$,
$- p$), explain what goes wrong, and state the minimum n needed to have at least 3 residual
om with $k = 4$ regressors.

what does the overall F test conclude that an individual t test does not, and vice versa?

In-Class Activity 7.2: Which Knob Moves Churn?

Week 11 — Lecture 21

Work in teams of 3–4. Write your reasoning, not just the final number. One submission per team, all names on it.

Your startup's churn model, fitted on $n = 30$ cohorts (residual df 26, $t_{0.025,26} = 2.056$), produced this coefficient summary:

Variable	Coef	SE	t-Stat	p-value
Intercept	4.800	1.520	3.16	0.004
Price (SAR/month)	0.052	0.021	2.48	0.020
Onboarding (min)	0.310	0.078	3.97	< 0.001
Tickets per user	0.145	0.290	0.50	0.621

Part 1 — Read the table

At $\alpha = 0.05$, which regressors contribute significantly, given the others? For each decision, write the comparison you used (either $|t|$ vs. 2.056 or p vs. 0.05).

Part 2 — Judge a claim

Marketing claims that cutting onboarding from 12 to 7 minutes will reduce churn by at least 3 percentage points. Using the fitted coefficient, compute the predicted reduction for that 5-minute cut, then build the 95% CI for β_2 and state whether the data can support the ''at least 3 points'' promise.

back on a shortcut

s: ''Tickets has the third-largest coefficient (0.145), bigger than price (0.052), so if anything we ce, not tickets.'' Write two or three sentences correcting this reasoning, using both the units of nd their standard errors.

hy can a regressor be insignificant in this model yet still be strongly related to churn on its own?