# 6 *Simple Linear Regression*

Everything so far in this course asked questions about parameters of one or two populations: ''is the mean equal to 50?'', ''what is a plausible range for $p$?'' This chapter asks a different kind of question: *how does one variable relate to another?* Does more verification testing reduce an AI system's failure rate? Does a heavier shipment burn more fuel? Does more ad spending bring more signups? The tool for answering such questions from data is *regression analysis*, and the simplest and most used version of it, one response and one predictor connected by a straight line, is the subject of this chapter. The chapter is organized in the following order: empirical models and the scatter diagram; the simple linear regression model and its assumptions; the method of least squares, derived from first principles; estimating the error variance; the statistical properties of the estimators; hand computation on real-sized data sets; hypothesis tests and the analysis of variance for regression; the coefficient of determination and the sample correlation; confidence intervals for the coefficients, for the mean response, and prediction intervals for new observations; residual analysis for checking the model; regression on transformed variables; and, finally, the two classic abuses of regression, extrapolation and the causation trap.

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

**Q1.** A logistics analyst fits the model $\hat{y} = 20.4 + 1.48x$ relating fuel use $y$ (liters) to shipment weight $x$ (tonnes). What fuel use does the model predict for a 20-tonne shipment, and what does the number 1.48 mean?

**Q2.** An AI safety team has fitted a regression of failure rate on testing hours. They want (a) a range for the *average* failure rate at $x_0 = 30$ hours, and (b) a range for the failure rate of the *next single* system tested for 30 hours. Which range is wider, and why?

**Q3.** A startup regresses weekly signups on weekly ad spend and reports $R^2 = 0.81$. State what 0.81 means, and state one thing it does *not* mean.

**Answers.**

**Q1.** $\hat{y} = 20.4 + 1.48(20) = 50.0$ liters. The slope 1.48 means that each additional tonne of shipment weight is associated with 1.48 more liters of fuel, *on average*, over the range of weights in the data.

**Q2.** Range (b), the prediction interval, is wider. The average in (a) is uncertain only because the fitted line is uncertain. A single new observation carries that same uncertainty *plus* its own random deviation $\varepsilon$ from the line, so its interval must be wider.

**Q3.** It means 81% of the sample variability in signups is explained by the linear relationship with ad spend. It does *not* mean that ad spend *causes* 81% of signups; regression on observational data measures association, not causation, and it also does not confirm that the model assumptions hold.

**Lecture 15 starts here** — *Simple linear regression, part 1*

**Lecture 15.** This section is covered by Lecture 15.

**Reference.** Montgomery & Runger, 6th ed., Section 11-1.

**Reference.** Lecture 16 is the software session for this chapter. Spreadsheet and Python recipes for fitting and checking every model in this chapter are collected in Chapter 10, *Software and Tools*.

**Empirical model.** A model built from observed data rather than derived from physical first principles.

**Regressor.** The predictor or independent variable $x$, treated as fixed or measured without error.

**Response.** The dependent variable $Y$, the quantity to be modeled or predicted. It is a random variable.

## 6.1 Empirical Models and the Scatter Diagram

Many relationships in engineering cannot be written down from theory. No physical law states exactly how a perception model's failure rate falls with verification effort, or how signups respond to ad spend. But data on such pairs of variables can be collected, and a model can be built *from the data*. A model of this kind is called an *empirical model*. It does not claim to explain the mechanism; it summarizes the observed relationship well enough to interpolate, predict, and quantify uncertainty.

The data for this chapter are $n$ pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. The variable $x$ is called the *regressor* (also predictor, or independent variable); the variable $y$ is the *response* (or dependent variable). The roles are not interchangeable: we model $y$ as depending on $x$, and the choice of which variable plays which role comes from the engineering question, not from the mathematics. If the question

is "how much fuel will this shipment need?", then fuel is the response and weight is the regressor.

The first step of every regression analysis, before any formula, is the *scatter diagram*: plot each observation as a point $(x_i, y_i)$ and look. Three questions should be answered by eye. Does $y$ tend to move with $x$ at all? If it does, is the trend reasonably straight, or curved? And are there isolated points far from the rest? Figure 6.1 shows the three basic patterns. Only the first one, an approximately straight trend, is the subject of this chapter; the second one is handled by the transformations of Section 6.13, and the third pattern means $x$ carries little information about $y$.

**Scatter diagram.** A plot of the pairs $(x_i, y_i)$ as points. The first step of every regression analysis.

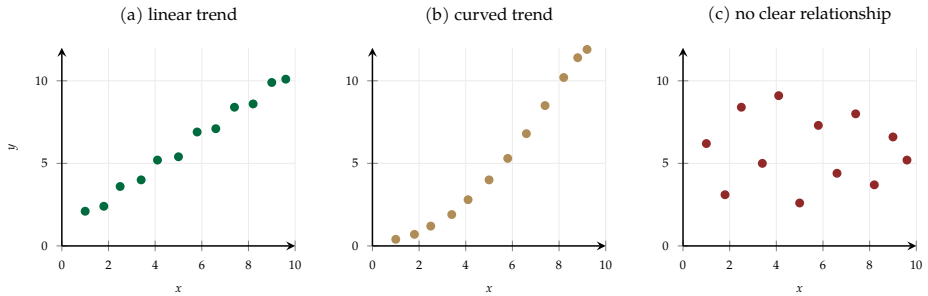


Figure 6.1. Three scatter-diagram patterns. Simple linear regression is appropriate for pattern (a). Pattern (b) calls for a transformation (Section 6.13); in pattern (c), $x$ has little predictive value for $y$.

**Startup lens.** Every startup dashboard is full of candidate regressions: spend vs. signups, price vs. conversion, onboarding steps vs. retention. The scatter diagram is the cheapest analysis you will ever run, and it prevents the most expensive mistake: fitting a straight line to a relationship that is not straight.

**Reference.** Montgomery & Runger, 6th ed., Section 11-2.

## 6.2 The Simple Linear Regression Model

Suppose the scatter diagram shows an approximately straight trend. We model the *expected value* of the response at each level of the regressor as a straight line,

$$E(Y \mid x) = \beta_0 + \beta_1 x, \tag{6.1}$$

and each individual observation as that line plus a random error:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, 2, \ldots, n. \tag{6.2}$$

**Intercept $\beta_0$.** The value of $E(Y \mid x)$ at $x = 0$. Meaningful only when $x = 0$ lies in or near the range of the data.

**Slope $\beta_1$.** The change in $E(Y \mid x)$ produced by a one-unit increase in $x$.

**Random error $\varepsilon$.** The deviation of an individual observation from the population regression line.

Equation (6.2) is the *simple linear regression model*: "simple" because there is one regressor, "linear" because the parameters enter linearly. Its three ingredients are:

- $\beta_0$, the *intercept*: the height of the line at $x = 0$;

- $\beta_1$, the *slope*: the change in the mean response per one-unit increase in $x$;
- $\varepsilon_i$, the *random error*: everything that moves observation $i$ off the line—measurement noise, omitted variables, natural unit-to-unit variation.

The parameters $\beta_0$, $\beta_1$, and the error variance are population quantities: fixed, unknown, and to be estimated from data, exactly like $\mu$ and $\sigma^2$ in earlier chapters.

The model comes with assumptions on the errors, and the whole chapter's inference machinery rests on them. Please pay attention here, because checking these assumptions (Section 6.12) is part of every serious regression analysis, not an optional extra:

- $E(\varepsilon_i) = 0$: the line is the true mean of $Y$ at every $x$;
- $\mathrm{Var}(\varepsilon_i) = \sigma^2$ for all $i$: the scatter around the line has the *same* spread at every $x$;
- the $\varepsilon_i$ are uncorrelated with one another;
- for tests and intervals, the $\varepsilon_i$ are normally distributed.

Under these assumptions, at each fixed $x$ the response $Y$ is a normal random variable with mean $\beta_0 + \beta_1 x$ and variance $\sigma^2$. Figure 6.2 draws this picture: the model is not one line of points but a whole *family of distributions*, one at each $x$, whose means line up on the population regression line and whose spreads are all equal.

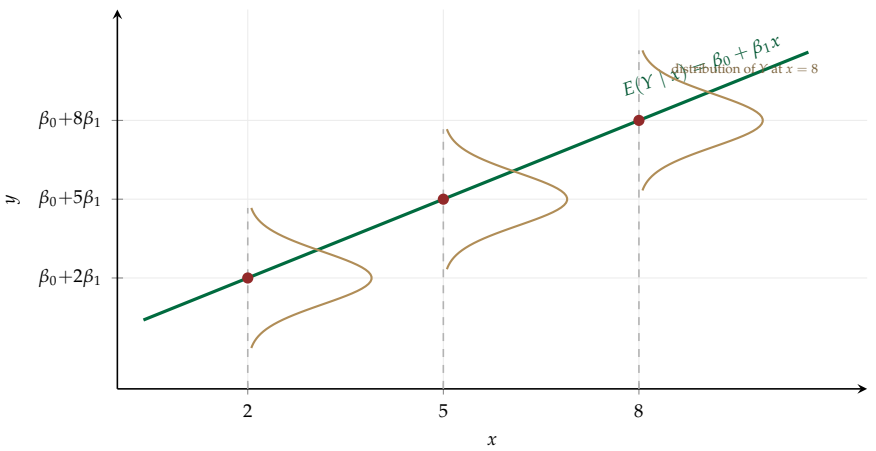

Figure 6.2. The simple linear regression model (6.2). At each value of $x$, the response $Y$ has a normal distribution (gold curves, drawn sideways) centered on the population regression line (green) with the *same* variance $\sigma^2$ everywhere.

## 6.3 The Method of Least Squares

The parameters $\beta_0$ and $\beta_1$ are unknown. Once we have estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, the *fitted regression line* is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x, \tag{6.3}$$

and the *residual* of observation $i$ is the vertical gap between the observed point and the fitted line,

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i. \tag{6.4}$$

How should the line be chosen? Many candidate criteria exist, but the one used almost everywhere, in statistics and in machine learning alike, is the *method of least squares*: choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the sum of squared vertical deviations,

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \big(y_i - \beta_0 - \beta_1 x_i\big)^2. \tag{6.5}$$

Squaring does two jobs: it makes positive and negative deviations count equally, and it makes $S$ a smooth function that calculus can minimize.

### 6.3.1 Deriving the normal equations

$S(\beta_0, \beta_1)$ is a function of two variables, so we set both partial derivatives to zero. Differentiating (6.5) with respect to $\beta_0$ and to $\beta_1$ and evaluating at the minimizers $\hat{\beta}_0, \hat{\beta}_1$:

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} \big(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\big) = 0,$$
$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} \big(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\big) x_i = 0.$$

Dividing by $-2$ and rearranging gives the *normal equations*:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \qquad \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i. \tag{6.6}$$

**Engineering judgment.** Slopes carry units, and the units carry the engineering meaning. A slope of 1.48 in a fuel model is not a bare number: it is 1.48 liters *per tonne*. Writing the units forces you to check that the magnitude is physically plausible before you trust the fit.

**Reference.** Montgomery & Runger, 6th ed., Section 11-2.

**Residual.** The vertical distance $e_i = y_i - \hat{y}_i$ from an observed point to the fitted line. The observable stand-in for the unobservable error $\varepsilon_i$.

**Method of least squares.** Choosing the line that minimizes the sum of squared vertical deviations of the data from the line.

**Normal equations.** The two linear equations obtained by setting the partial derivatives of the least squares criterion to zero.

These are two linear equations in the two unknowns $\hat{\beta}_0$ and $\hat{\beta}_1$. Solving them is routine algebra, and the solution is cleanest when written in terms of two summary quantities that will appear in every formula for the rest of this chapter: the *corrected sum of squares of* $x$,

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}, \tag{6.7}$$

and the *corrected sum of cross products*,

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}. \tag{6.8}$$

The second form of each is the computational shortcut: it needs only the raw totals $\sum x_i$, $\sum y_i$, $\sum x_i^2$, $\sum x_i y_i$, with no subtraction of means observation by observation. In terms of these, the *least squares estimators* are

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \tag{6.9}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{6.10}$$

The slope estimator (6.9) is the co-movement of $x$ and $y$ divided by the spread of $x$: how much $y$ changes with $x$, per unit of variation in $x$. The intercept estimator (6.10) is whatever value makes the line pass through the point $(\bar{x}, \bar{y})$.

**Insight.** Two facts fall straight out of the normal equations (6.6), and both are useful checks on any fit. The first normal equation says $\sum e_i = 0$: least squares residuals always sum to zero, so a residual list that does not (up to rounding) contains an arithmetic error. Dividing that same equation by $n$ gives $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$: the fitted line always passes exactly through the center point $(\bar{x}, \bar{y})$ of the data. Every least squares line is anchored at the data's center of gravity.

The complete fitting recipe, which we will use many times, is recorded as a procedure.

**Procedure 6.1 — Fitting the least squares line.**

1. **Plot the data.** Draw the scatter diagram and confirm an approximately straight trend with no wild outliers. If the trend is curved, stop and consider a transformation (Section 6.13).
2. **Accumulate the totals.** Compute $n$, $\sum x_i$, $\sum y_i$, $\sum x_i^2$, $\sum x_i y_i$ (and $\sum y_i^2$, which the inference steps of this chapter will need). Then $\bar{x} = \sum x_i / n$ and $\bar{y} = \sum y_i / n$.
3. **Compute the building blocks.** $S_{xx}$ from (6.7) and $S_{xy}$ from (6.8).
4. **Compute the estimates.** $\hat{\beta}_1 = S_{xy}/S_{xx}$ by (6.9), then $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ by (6.10).
5. **Write and check the model.** State $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ with units. Check that the sign of $\hat{\beta}_1$ matches the scatter diagram and that the line passes through $(\bar{x}, \bar{y})$.

**Example 6.1 — Does more verification testing reduce failures?**

An ISE engineer evaluates an autonomous inspection system at a refinery. She subjects eight copies of the system to different amounts of verification testing $x$ (hours) and records each one's deployed failure rate $y$ (failures per 1000 inspections):

| $x$ (test hours) | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 |
|---|---|---|---|---|---|---|---|---|
| $y$ (failures/1000) | 48 | 42 | 38 | 32 | 29 | 24 | 22 | 18 |

Fit the least squares line and interpret the slope.

***Solution.***

***Step 1 — Plot.***

Figure 6.3 shows the scatter: a clear, nearly straight, decreasing trend. Simple linear regression is appropriate.

***Step 2 — Totals.***
With $n = 8$: $\sum x_i = 220$, $\sum y_i = 253$, $\sum x_i^2 = 7100$, $\sum x_i y_i = 6070$, and $\sum y_i^2 = 8761$ (kept for later). Then $\bar{x} = 220/8 = 27.5$ and $\bar{y} = 253/8 = 31.625$.

***Step 3 — Building blocks.***
By (6.7) and (6.8),

$$S_{xx} = 7100 - \frac{220^2}{8} = 7100 - 6050 = 1050, \qquad S_{xy} = 6070 - \frac{(220)(253)}{8} = 6070 - 6957.5 =$$

***Step 4 — Estimates.***
By (6.9) and (6.10),

$$\hat{\beta}_1 = \frac{-887.5}{1050} = -0.8452, \qquad \hat{\beta}_0 = 31.625 - (-0.8452)(27.5) = 31.625 + 23.24 = 54.87.$$

***Step 5 — Model and check.***
The fitted model is $\boxed{\hat{y} = 54.87 - 0.8452\,x}$. The slope is negative, matching the scatter. Interpretation: over the tested range of 10 to 45 hours, each additional hour of verification testing is associated with about 0.85 *fewer* failures per 1000 inspections, on average. The intercept 54.87 would be the failure rate at zero testing hours; $x = 0$ sits below the smallest observed $x = 10$, so it is a mild extrapolation and should be quoted with caution.

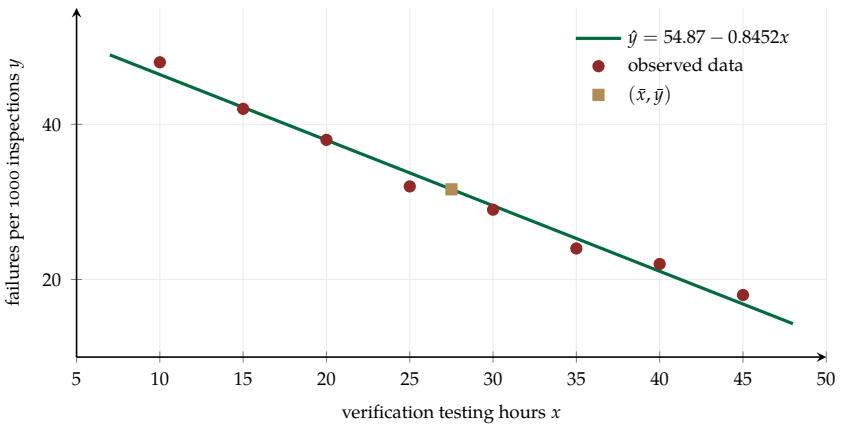


Figure 6.3. The verification-testing data of Example 6.1 with the fitted least squares line. The line passes exactly through $(\bar{x}, \bar{y}) = (27.5,\ 31.625)$, as every least squares line must.

**Safety lens.** The fitted line in Example 6.1 summarizes eight systems tested between 10 and 45 hours. It says nothing about 200 hours: substituting $x = 200$ gives $\hat{y} = 54.87 - 0.8452(200) = -114.2$ failures per 1000, a negative rate, which is meaningless. A safety argument built on an extrapolated regression is not evidence; Section 6.14 returns to this.

## 6.4 Estimating the Error Variance

**Reference.** Montgomery & Runger, 6th ed., Section 11-2.

**Error sum of squares.** $SS_E = \sum e_i^2$, the squared scatter remaining after the line is fitted.

The model (6.2) has a third parameter, the error variance $\sigma^2$, and every test and interval in this chapter needs an estimate of it. The natural raw material is the residuals: the *error sum of squares*

$$SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 \tag{6.11}$$

measures how much squared scatter the fitted line fails to explain. Computing $SS_E$ residual by residual is tedious; a shortcut does it from the summary quantities. Define the corrected sum of squares of $y$,

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}, \tag{6.12}$$

which is also written $SS_T$ (the *total* sum of squares, a name explained in Section 6.8). Then

$$SS_E = S_{yy} - \hat{\beta}_1 S_{xy}. \tag{6.13}$$

The unbiased estimator of $\sigma^2$ divides by $n - 2$:

$$\hat{\sigma}^2 = \frac{SS_E}{n-2} = MS_E. \tag{6.14}$$

The quantity $MS_E$ (“error mean square”) is the regression analogue of the sample variance, and $\hat{\sigma} = \sqrt{MS_E}$ estimates the typical size of the scatter around the line, in the units of $y$.

**Insight.** Why $n - 2$ and not $n - 1$? Degrees of freedom count observations minus estimated parameters. The sample variance $S^2$ divides by $n - 1$ because one parameter, $\bar{x}$, was estimated from the data. Here the residuals are measured from a line with *two* estimated parameters, $\hat{\beta}_0$ and $\hat{\beta}_1$, so two degrees of freedom are spent and $n - 2$ remain. The pattern continues: multiple regression with $p$ estimated coefficients divides by $n - p$.

**Example 6.2 — Error variance for the verification data**

For the data of Example 6.1, estimate $\sigma^2$ and $\sigma$.

***Solution.***

***Step 1 — Total sum of squares.***

By (6.12), using $\sum y_i^2 = 8761$ and $\sum y_i = 253$:

$$S_{yy} = 8761 - \frac{253^2}{8} = 8761 - 8001.125 = 759.875.$$

***Step 2 — Error sum of squares.***

By (6.13), with $\hat{\beta}_1 = -0.8452$ and $S_{xy} = -887.5$:

$$SS_E = 759.875 - (-0.8452)(-887.5) = 759.875 - 750.1 = 9.726.$$

Both factors are negative, so the product $\hat{\beta}_1 S_{xy}$ is positive; a negative $SS_E$ at this step always signals a sign error.

***Step 3 — Estimate.***

By (6.14), with $n - 2 = 6$ degrees of freedom,

$$\hat{\sigma}^2 = MS_E = \frac{9.726}{6} = 1.621, \qquad \hat{\sigma} = \sqrt{1.621} = 1.273.$$

The answer is $\boxed{\hat{\sigma}^2 = 1.621}$: after the linear effect of testing hours is removed, individual systems scatter about the line by roughly $\pm 1.3$ failures per 1000 inspections.

**Reference.** Montgomery & Runger, 6th ed., Section 11-3.

## 6.5 *Properties of the Least Squares Estimators*

The estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are statistics: they are built from the random responses $Y_i$, so they have sampling distributions, exactly as $\bar{X}$ did in Chapter 2. Under the model assumptions, both estimators are *unbiased*,

$$E(\hat{\beta}_1) = \beta_1, \qquad E(\hat{\beta}_0) = \beta_0,$$

and their variances are $\mathrm{Var}(\hat{\beta}_1) = \sigma^2 / S_{xx}$ and $\mathrm{Var}(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$. Replacing $\sigma^2$ by its estimate $MS_E$ from (6.14) gives the *estimated standard errors* that all the

inference formulas use:

$$\operatorname{se}(\hat{\beta}_1) = \sqrt{\frac{MS_E}{S_{xx}}}, \tag{6.15}$$

$$\operatorname{se}(\hat{\beta}_0) = \sqrt{MS_E\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}. \tag{6.16}$$

Three remarks connect these formulas to engineering practice. First, both standard errors shrink when $MS_E$ shrinks: cleaner data give sharper estimates. Second, $\operatorname{se}(\hat{\beta}_1)$ shrinks when $S_{xx}$ grows: spreading the $x$ values over a wider range pins down the slope more firmly. An experimenter who can choose the test conditions should spread them out rather than cluster them. Third, the intercept's standard error (6.16) contains the extra term $\bar{x}^2/S_{xx}$: when the data sit far from $x = 0$, the intercept is a long lever-arm extrapolation and is estimated poorly even when the slope is estimated well.

The structure of this chapter is now exactly the structure of Chapters 2 through 5, with new symbols. Table 6.1 makes the correspondence explicit; if the left column is familiar, the right column is the same routine.

| Concept | One-sample inference | Regression |
|---|---|---|
| Parameters | $\mu, \sigma^2$ | $\beta_0, \beta_1, \sigma^2$ |
| Estimators | $\bar{X}, S^2$ | $\hat{\beta}_1, \hat{\beta}_0, MS_E$ |
| Unbiased? | yes | yes |
| Standard error | $s/\sqrt{n}$ | $\sqrt{MS_E/S_{xx}}$ (slope) |
| Degrees of freedom | $n-1$ | $n-2$ |
| Test statistic | $T_0 = \dfrac{\bar{X}-\mu_0}{s/\sqrt{n}}$ | $T_0 = \dfrac{\hat{\beta}_1}{\sqrt{MS_E/S_{xx}}}$ |
| Interval form | estimate $\pm\, t\times$ se | estimate $\pm\, t\times$ se |

Table 6.1. Regression inference reuses the estimate–standard error–$t$ structure of one-sample inference. Only the formulas for the estimate and its standard error change.

**Lecture 17 starts here** — *Simple linear regression, continued*

**Lecture 17.** This section is covered by Lecture 17.

**Reference.** Montgomery & Runger, 6th ed., Sections 11-2 and 11-4 examples.

## 6.6 Fitting the Line in Practice

This section does what a practice lecture does: it works the fitting routine of Procedure 6.1 on full data sets, slowly, exactly as you will do it in an exam with a calculator. The first example carries a ten-observation data set through the complete hand computation, including the running-totals table you should reproduce on paper. The same data set returns in Sections 6.7 and 6.10 for inference, so the work invested here pays twice.

**Example 6.3 — Shipment weight and fuel use: the full hand computation**

A logistics company suspects that the fuel a delivery truck uses on its standard route depends linearly on the shipment weight it carries. Ten runs of the same route give the data in Table 6.2: shipment weight $x$ (tonnes) and fuel used $y$ (liters). Fit the least squares line, estimate $\sigma^2$, and interpret both coefficients.

***Solution.***

***Step 1 — Plot.***

The scatter (Figure 6.4) is strongly linear and increasing; proceed.

***Step 2 — Totals.***

Build the computation table column by column and total each column. Table 6.2 shows the result: $n = 10$, $\sum x_i = 170$, $\sum y_i = 455$, $\sum x_i^2 = 3220$, $\sum y_i^2 = 21429$, $\sum x_i y_i = 8222$. Hence $\bar{x} = 17.00$ tonnes and $\bar{y} = 45.50$ liters.

***Step 3 — Building blocks.***

By (6.7), (6.8), and (6.12),

$$S_{xx} = 3220 - \frac{170^2}{10} = 3220 - 2890 = 330.0,$$
$$S_{xy} = 8222 - \frac{(170)(455)}{10} = 8222 - 7735 = 487.0,$$
$$S_{yy} = 21429 - \frac{455^2}{10} = 21429 - 20702.5 = 726.5.$$

***Step 4 — Estimates.***
By (6.9) and (6.10),

$$\hat{\beta}_1 = \frac{487.0}{330.0} = 1.476\,\text{L/tonne}, \qquad \hat{\beta}_0 = 45.50 - (1.476)(17.00) = 45.50 - 25.09 = 20.41\,\text{L}.$$

***Step 5 — Error variance.***
By (6.13) and (6.14),

$$SS_E = 726.5 - (1.476)(487.0) = 726.5 - 718.7 = 7.806, \qquad \hat{\sigma}^2 = \frac{7.806}{8} = 0.9758.$$

***Step 6 — Model and interpretation.***
The fitted model is $\boxed{\hat{y} = 20.41 + 1.476\,x}$ with $\hat{\sigma} = \sqrt{0.9758} = 0.9878$ L. Each additional tonne of load is associated with about 1.48 more liters of fuel on this route, on average. The intercept, 20.41 L, estimates the fuel the route itself costs with an empty truck; since the observed weights start at 8 tonnes, $x = 0$ is a modest extrapolation, but here it has a physical meaning (the vehicle's own weight and the fixed route) that makes it credible. Individual runs scatter about the line by roughly $\pm 1$ liter.

| $i$ | $x_i$ (t) | $y_i$ (L) | $x_i^2$ | $y_i^2$ | $x_i y_i$ |
|---|---|---|---|---|---|
| 1 | 8 | 32 | 64 | 1024 | 256 |
| 2 | 10 | 36 | 100 | 1296 | 360 |
| 3 | 12 | 37 | 144 | 1369 | 444 |
| 4 | 14 | 42 | 196 | 1764 | 588 |
| 5 | 16 | 43 | 256 | 1849 | 688 |
| 6 | 18 | 48 | 324 | 2304 | 864 |
| 7 | 20 | 49 | 400 | 2401 | 980 |
| 8 | 22 | 53 | 484 | 2809 | 1166 |
| 9 | 24 | 57 | 576 | 3249 | 1368 |
| 10 | 26 | 58 | 676 | 3364 | 1508 |
| $\sum$ | 170 | 455 | 3220 | 21429 | 8222 |

Table 6.2. Computation table for Example 6.3. Building the five columns and totaling them is Step 2 of Procedure 6.1; every later formula in the analysis reads from the bottom row.

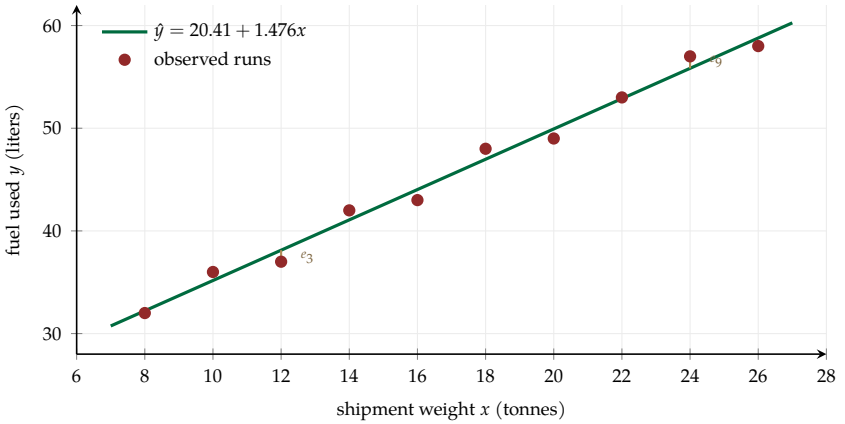


Figure 6.4. The fuel data of Example 6.3 with the fitted line. The gold segments show two residuals: least squares chooses the line that minimizes the sum of the squares of all ten such vertical gaps.

**Example 6.4 — Ad spend and signups: fit, predict, residual, interpret**

Your startup varies its weekly advertising budget and records the resulting signups. Five weeks of data, with $x$ = ad spend (thousand SAR) and $y$ = signups (hundreds):

| Week | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $x$ (thousand SAR) | 2 | 4 | 6 | 8 | 10 |
| $y$ (hundred signups) | 2 | 5 | 4 | 7 | 7 |

The totals are $\sum x_i = 30$, $\sum y_i = 25$, $\sum x_i^2 = 220$, $\sum y_i^2 = 143$, $\sum x_i y_i = 174$. (a) Fit the least squares line. (b) Estimate signups for a week with $x = 8$. (c) Find the residual for the observed week with $x = 8$; does the model over- or underestimate that week? (d) What signup change does the model associate with spending 3 thousand SAR more per week?

***Solution.***

***(a) Fit (Procedure 6.1).***

$\bar{x} = 30/5 = 6$, $\bar{y} = 25/5 = 5$. By (6.7) and (6.8),

$$S_{xx} = 220 - \frac{30^2}{5} = 40, \qquad S_{xy} = 174 - \frac{(30)(25)}{5} = 24.$$

Then by (6.9) and (6.10), $\hat{\beta}_1 = 24/40 = 0.60$ and $\hat{\beta}_0 = 5 - 0.60(6) = 1.40$, so $\boxed{\hat{y} = 1.40 + 0.60\,x}$.

**(b) *Prediction at* $x = 8$.**

$\hat{y} = 1.40 + 0.60(8) = 6.20$, that is, about 620 signups.

**(c) *Residual.***

The observed value at $x = 8$ is $y = 7$. By (6.4), $e = 7 - 6.20 = +0.80$. The residual is positive: the actual signups exceeded the fitted value, so the model *underestimates* that week.

**(d) *Slope interpretation.***

The slope is the change in mean response per unit of $x$, so a 3-unit increase changes the estimate by $\Delta\hat{y} = 0.60 \times 3 = 1.80$ hundred, about 180 additional signups per week. The answer is $\boxed{\Delta\hat{y} = 1.80}$ hundred signups. Note the phrasing: the model *associates* the extra spend with extra signups; whether spending *causes* them is a separate question (Section 6.14).

**Startup lens.** Example 6.4(d) is the number a founder actually uses: 180 signups per extra 3,000 SAR is about 16.7 SAR per signup at the margin. Regression turns a scatter of weekly noise into a unit cost you can compare against customer lifetime value.

**Lecture 18 starts here** — *Linear regression, part 2*

**Lecture 18.** This section is covered by Lecture 18.

**Reference.** Montgomery & Runger, 6th ed., Section 11-4.1.

## 6.7 Testing for Significance of Regression

A least squares line can *always* be computed, even from pure noise. Before using a fitted model we must therefore ask: is the relationship real, or could a slope of this size arise by chance when the true slope is zero? This is a hypothesis test, and it has exactly the structure of Chapter 4. The hypotheses are

$$H_0 : \beta_1 = 0 \qquad \text{versus} \qquad H_1 : \beta_1 \neq 0, \tag{6.17}$$

where $H_0$ says the mean of $Y$ does not change with $x$ at all: the population regression line is flat, and $x$ has no linear predictive value. Figure 6.5 shows the two situations.

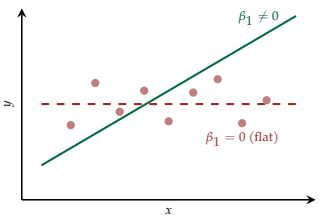


Figure 6.5. If $\beta_1 = 0$, knowing $x$ tells you nothing about $Y$. The test asks whether the observed slope is too far from zero to be noise.

The test statistic standardizes the estimated slope by its standard error (6.15):

$$T_0 = \frac{\hat{\beta}_1}{\operatorname{se}(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\sqrt{MS_E / S_{xx}}}, \tag{6.18}$$

and under $H_0$ it follows the $t$ distribution with $n - 2$ degrees of freedom. At significance level $\alpha$, reject $H_0$ when $|T_0| > t_{\alpha/2,\,n-2}$. The same statistic with

$\hat{\beta}_1 - \beta_{1,0}$ in the numerator tests $H_0\colon \beta_1 = \beta_{1,0}$ for any hypothesized slope, and one-sided versions work exactly as in Chapter 4.

**Procedure 6.2 — Testing the slope.**

1. **State the hypotheses** (6.17) (or a one-sided pair, if the engineering question is directional) and choose $\alpha$.
2. **Fit and estimate.** Obtain $\hat{\beta}_1$ by (6.9) and $MS_E$ by (6.14).
3. **Compute the standard error** of the slope by (6.15): $\operatorname{se}(\hat{\beta}_1) = \sqrt{MS_E / S_{xx}}$.
4. **Compute the statistic** $T_0$ by (6.18).
5. **Find the critical value** $t_{\alpha/2,\,n-2}$ from the $t$ table (or bracket the P-value between table columns).
6. **Decide and conclude in context.** Reject $H_0$ if $|T_0| > t_{\alpha/2,\,n-2}$; state the conclusion in the problem's own units and language.

**Example 6.5 — Is the ad-spend effect real?**

For the ad-spend data of Example 6.4, test $H_0\colon \beta_1 = 0$ against $H_1\colon \beta_1 \neq 0$ at $\alpha = 0.05$, and bracket the P-value.

***Solution.***

***Step 1 — Hypotheses.***

$H_0\colon \beta_1 = 0$, $H_1\colon \beta_1 \neq 0$, $\alpha = 0.05$, $n - 2 = 3$ degrees of freedom.

***Step 2 — Estimates.***

From Example 6.4, $\hat{\beta}_1 = 0.60$, $S_{xx} = 40$. We also need $S_{yy}$ by (6.12): $S_{yy} = 143 - 25^2/5 = 18.00$. Then by (6.13) and (6.14), $SS_E = 18.00 - (0.60)(24) = 3.600$ and $MS_E = 3.600/3 = 1.200$.

***Step 3 — Standard error.***

By (6.15), $\operatorname{se}(\hat{\beta}_1) = \sqrt{1.200/40} = \sqrt{0.03000} = 0.1732$.

***Step 4 — Statistic.***
By (6.18),

$$T_0 = \frac{0.60}{0.1732} = 3.464.$$

***Step 5 — Critical value.***
Table 6.3 gives $t_{0.025,3} = 3.182$.

| $\nu$ | $t_{0.05}$ | $t_{0.025}$ | $t_{0.01}$ |
|---|---|---|---|
| 2 | 2.920 | 4.303 | 6.965 |
| 3 | 2.353 | 3.182 | 4.541 |
| 4 | 2.132 | 2.776 | 3.747 |

Table 6.3. $t$ table excerpt: $t_{0.025,3} = 3.182$.

***Step 6 — Decide.***
$|T_0| = 3.464 > 3.182$, so reject $H_0$ at $\alpha = 0.05$: the data give significant evidence of a linear relationship between ad spend and signups. For the P-value, $t_{0.025,3} = 3.182 < 3.464 < t_{0.01,3} = 4.541$, so the one-tail area is between 0.01 and 0.025 and the two-sided P-value satisfies $0.02 < P < 0.05$. The evidence is significant but not overwhelming; five weeks of data leave only three degrees of freedom for estimating the noise.

**Reference.** Montgomery & Runger, 6th ed., Section 11-4.2.

**Analysis of variance.** Splitting the total variability of $y$ into a part explained by the model and a residual part, and comparing the two.

## 6.8 The Analysis of Variance for Regression

There is a second, equivalent way to test the significance of regression, and it introduces the bookkeeping that Chapters 8 and 9 will use constantly: the *analysis of variance* (ANOVA). Start from the total variability of the responses about their mean, $SS_T = S_{yy}$, and split each deviation as $y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$. Squaring and summing (the cross term vanishes, by the normal equations) gives the *ANOVA identity*

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{SS_T} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{SS_R} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{SS_E}. \tag{6.19}$$

$SS_R$ is the *regression sum of squares*, the variability the fitted line explains; $SS_E$ is the error sum of squares, the variability it does not. The computing forms follow

from (6.13):

$$SS_T = S_{yy}, \qquad SS_R = \hat{\beta}_1 S_{xy}, \qquad SS_E = S_{yy} - \hat{\beta}_1 S_{xy}.$$

If the model is useful, $SS_R$ should be large relative to $SS_E$. The formal comparison divides each sum of squares by its degrees of freedom (1 for regression, since one slope is fitted; $n-2$ for error) to form mean squares, and takes their ratio:

$$F_0 = \frac{MS_R}{MS_E} = \frac{SS_R/1}{SS_E/(n-2)}, \tag{6.20}$$

which under $H_0\colon \beta_1 = 0$ follows the $F$ distribution with 1 and $n-2$ degrees of freedom. Reject $H_0$ when $F_0 > f_{\alpha,1,n-2}$. The computation is organized in the standard ANOVA table, Table 6.4; learn this layout, because exams routinely ask you to complete a partially filled version of it.

| Source | df | SS | MS | $F_0$ |
|---|---|---|---|---|
| Regression | 1 | $SS_R = \hat{\beta}_1 S_{xy}$ | $MS_R = SS_R/1$ | $F_0 = MS_R/MS_E$ |
| Error | $n-2$ | $SS_E = S_{yy} - \hat{\beta}_1 S_{xy}$ | $MS_E = SS_E/(n-2)$ | |
| Total | $n-1$ | $SS_T = S_{yy}$ | | |

Table 6.4. The ANOVA table for simple linear regression. The df and SS columns each add up: $1 + (n-2) = n-1$ and $SS_R + SS_E = SS_T$ by (6.19).

Please pay attention to one exact fact, because it saves work and catches errors: for simple linear regression,

$$F_0 = T_0^2,$$

where $T_0$ is the slope statistic (6.18). The $F$ test and the two-sided $t$ test are the *same* test, and they always reach the same conclusion. If your computed $F_0$ is not the square of your computed $T_0$ (up to rounding), one of them is wrong.

**Example 6.6 — ANOVA for the ad-spend regression**

Build the complete ANOVA table for the ad-spend data of Example 6.4, carry out the $F$ test at $\alpha = 0.05$, and verify the $F_0 = T_0^2$ identity against Example 6.5.

***Solution.***

***Step 1 — Sums of squares.***
From earlier work, $S_{yy} = 18.00$, $\hat{\beta}_1 = 0.60$, $S_{xy} = 24$. Then

$$SS_T = 18.00, \qquad SS_R = (0.60)(24) = 14.40, \qquad SS_E = 18.00 - 14.40 = 3.600.$$

***Step 2 — Mean squares and $F_0$.***
$MS_R = 14.40/1 = 14.40$ and $MS_E = 3.600/3 = 1.200$, so by (6.20), $F_0 = 14.40/1.200 = 12.00$. Table 6.5 assembles the results.

| Source | df | SS | MS | $F_0$ |
|---|---|---|---|---|
| Regression | 1 | 14.40 | 14.40 | 12.00 |
| Error | 3 | 3.600 | 1.200 | |
| Total | 4 | 18.00 | | |

Table 6.5. Completed ANOVA table for the ad-spend regression (Example 6.6).

***Step 3 — Test.***
Table 6.6 gives $f_{0.05,1,3} = 10.13$. Since $F_0 = 12.00 > 10.13$, reject $H_0\colon \beta_1 = 0$ at $\alpha = 0.05$.

| $\nu_2$ | $\nu_1 = 1$ | $\nu_1 = 2$ |
|---|---|---|
| 2 | 18.51 | 19.00 |
| 3 | 10.13 | 9.55 |
| 4 | 7.71 | 6.94 |

Table 6.6. $F$ table excerpt ($\alpha = 0.05$): $f_{0.05,1,3} = 10.13$.

***Step 4 — Consistency check.***
From Example 6.5, $T_0 = 3.464$ and $T_0^2 = (3.464)^2 = 12.00 = F_0$. The two tests agree exactly, as they must.

**Reference.** Montgomery & Runger, 6th ed., Sections 11-7.2 and 11-8.

## 6.9 $R^2$ and the Sample Correlation Coefficient

The ANOVA identity (6.19) suggests a natural one-number summary of how well the line fits: the fraction of the total variability that the regression explains. This is the *coefficient of determination*,

**Coefficient of determination.** $R^2 = SS_R/SS_T$, the proportion of the total variability of $y$ explained by the regression.

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}, \qquad 0 \le R^2 \le 1. \tag{6.21}$$

$R^2 = 1$ means every point lies exactly on the line ($SS_E = 0$); $R^2 = 0$ means the line explains nothing ($\hat{\beta}_1 = 0$). For the ad-spend regression, $R^2 = 14.40/18.00 = 0.80$: 80% of the week-to-week variability in signups is explained by the linear relationship with ad spend, and 20% remains unexplained. Figure 6.6 shows what high and low $R^2$ look like on a scatter diagram.

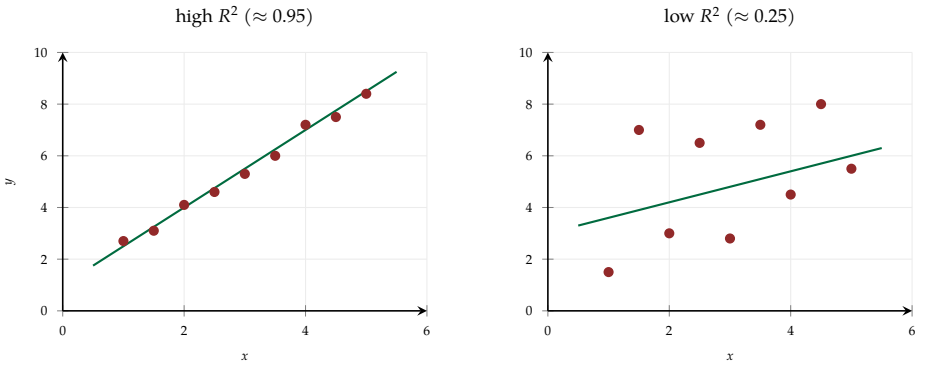


Figure 6.6. Two data sets with the same kind of fitted line. Left: the points hug the line and the model explains most of the variability. Right: the trend is present but weak, and most of the variability remains in the residuals.

**Sample correlation coefficient.** $r = S_{xy}/\sqrt{S_{xx}S_{yy}}$, a unitless measure of linear association between $-1$ and $+1$.

Closely related is the *sample correlation coefficient*,

$$r = \frac{S_{xy}}{\sqrt{S_{xx}\,S_{yy}}}, \qquad -1 \le r \le +1, \tag{6.22}$$

a unitless measure of the strength and direction of linear association. For simple linear regression the two summaries are tied together exactly: $R^2 = r^2$, and the sign of $r$ equals the sign of $\hat{\beta}_1$ (both share the numerator $S_{xy}$). For the ad-spend data, $r = 24/\sqrt{40 \times 18} = 24/26.83 = 0.8944$, and $r^2 = 0.7999 \approx 0.80 = R^2$, with $r > 0$ matching the positive slope. The correlation and the slope answer different questions: $r$ says how *tightly* the points follow a line, in unitless terms; $\hat{\beta}_1$ says how *steep* that line is, in the units of $y$ per unit of $x$. A relationship can be tight but shallow, or steep but noisy.

Please pay attention here, because $R^2$ is the most over-interpreted number in applied statistics. Three things a large $R^2$ does *not* mean:

- It does not mean the model assumptions hold. A curved relationship can produce $R^2 = 0.92$ while the residuals show a blatant pattern (Section 6.12); the linear model is still wrong.

- It does not mean the model predicts well outside the observed range of $x$. $R^2$ is computed entirely inside the data.
- It does not mean $x$ causes $y$. Association carried by a third variable produces high $R^2$ just as easily as causation does (Section 6.14).

**Safety lens.** A vendor showing $R^2 = 0.97$ for "predicted vs. actual" performance of a safety component has shown you a fit, not a validation. Ask for the residual plots, the range of conditions covered, and the holdout data. High $R^2$ on the training range is the beginning of the argument, not the end.

**Lecture 19.** This section is covered by Lecture 19.

**Reference.** Montgomery & Runger, 6th ed., Sections 11-3 and 11-5.1.

---

**Lecture 19 starts here** — *Linear regression, part 3*

## 6.10 Confidence Intervals for the Slope and Intercept

Chapter 3's pattern, estimate $\pm$ ($t$ critical value) $\times$ (standard error), applies verbatim to the regression coefficients. A $100(1-\alpha)\%$ confidence interval for the slope is

$$\hat{\beta}_1 \;\pm\; t_{\alpha/2,\,n-2}\sqrt{\frac{MS_E}{S_{xx}}}, \tag{6.23}$$

and for the intercept,

$$\hat{\beta}_0 \;\pm\; t_{\alpha/2,\,n-2}\sqrt{MS_E\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}. \tag{6.24}$$

The duality from Chapter 4 carries over: the interval (6.23) excludes zero exactly when the $t$ test of Procedure 6.2 rejects $H_0\colon \beta_1 = 0$ at the same $\alpha$. An hypothesis test on the intercept uses the matching statistic

$$T_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{MS_E\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}} \;\sim\; t_{n-2} \text{ under } H_0\colon \beta_0 = \beta_{0,0}. \tag{6.25}$$

Testing or interpreting the intercept is meaningful only when $x = 0$ lies inside or near the range of the data, or when it has a physical meaning (as the empty-truck fuel did in Example 6.3); otherwise the intercept is a geometric anchor, not an engineering quantity.

**Example 6.7 — Interval estimates for the fuel model**

For the fuel regression of Example 6.3 ($n = 10$, $\hat{\beta}_1 = 1.476$, $\hat{\beta}_0 = 20.41$, $MS_E = 0.9758$, $S_{xx} = 330.0$, $\bar{x} = 17.00$), find 95% confidence intervals for $\beta_1$

and $\beta_0$, and state what each interval tells the fleet manager.

***Solution.***

***Step 1 — Critical value.***

Degrees of freedom $n-2=8$; Table 6.7 gives $t_{0.025,8}=2.306$.

| $\nu$ | $t_{0.05}$ | $t_{0.025}$ | $t_{0.01}$ |
|---|---|---|---|
| 7 | 1.895 | 2.365 | 2.998 |
| 8 | 1.860 | 2.306 | 2.896 |
| 9 | 1.833 | 2.262 | 2.821 |

Table 6.7. $t$ table excerpt: $t_{0.025,8}=2.306$.

***Step 2 — Slope interval.***

By (6.15), $\operatorname{se}(\hat\beta_1)=\sqrt{0.9758/330.0}=\sqrt{0.002957}=0.05438$. By (6.23),

$$1.476 \pm (2.306)(0.05438) = 1.476 \pm 0.1254 \quad\Longrightarrow\quad \boxed{1.350 \le \beta_1 \le 1.601}\ \text{L/tonne}.$$

***Step 3 — Intercept interval.***

By (6.16),

$$\operatorname{se}(\hat\beta_0)=\sqrt{0.9758\left(\frac{1}{10}+\frac{17.00^2}{330.0}\right)}=\sqrt{0.9758\,(0.1000+0.8758)}=\sqrt{0.9521}=0.9758,$$

so by (6.24), $20.41 \pm (2.306)(0.9758) = 20.41 \pm 2.250$, giving $18.16 \le \beta_0 \le 22.66$ liters.

***Step 4 — Interpretation.***

The manager can state with 95% confidence that each tonne of payload costs between 1.35 and 1.60 liters of fuel on this route; the interval excludes zero decisively, so the weight effect is real. The base fuel cost of the route (empty truck) is between about 18.2 and 22.7 liters.

**Reference.** Montgomery & Runger, 6th ed., Sections 11-5.2 and 11-6.

**Mean response.** $E(Y \mid x_0)=\beta_0+\beta_1 x_0$, the height of the population regression line at $x_0$.

## 6.11 *Mean Response and Prediction of New Observations*

The most common use of a fitted model is to say something about the response at a chosen setting $x_0$. Two different questions hide behind "what is $y$ at $x_0$?", and they get two different intervals. Please read every problem statement for which

one it asks, because confusing them is the classic regression exam error.

The first question is about the *mean response*: where is the population line at $x_0$? The point estimate is $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$, and a $100(1-\alpha)\%$ *confidence interval on the mean response* is

$$\hat{y}_0 \pm t_{\alpha/2,\,n-2}\sqrt{MS_E\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}. \tag{6.26}$$

**Prediction interval.** An interval for a single future observation $Y_{\text{new}}$ at $x_0$; always wider than the confidence interval for the mean response.

The second question is about a *single new observation*: if one more run is made at $x_0$, what will its $y$ be? A $100(1-\alpha)\%$ *prediction interval* for $Y_{\text{new}}$ at $x_0$ is

$$\hat{y}_0 \pm t_{\alpha/2,\,n-2}\sqrt{MS_E\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}. \tag{6.27}$$

The only difference is the extra "1+" inside the square root, and it is not a detail: it is the variance of the new observation's own error $\varepsilon$, which no amount of data about the line can remove.

**Insight.** Where the "1+" comes from. The prediction error is $Y_{\text{new}} - \hat{y}_0 = \varepsilon_{\text{new}} + \big(E(Y \mid x_0) - \hat{y}_0\big)$: the new point's own scatter, plus our error in locating the line. The two pieces are independent, so their variances add: $\sigma^2 + \sigma^2\big(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\big)$. Collecting more data shrinks the second piece toward zero, but the first piece, the irreducible unit-to-unit variation, stays. A prediction interval can never be narrower than the scatter of individual observations.

Both intervals share a second important feature, visible in Figure 6.7: the term $(x_0-\bar{x})^2/S_{xx}$ makes them *narrowest at* $x_0 = \bar{x}$ and steadily wider as $x_0$ moves toward the edge of the data. The fitted line is best known at the center of the data and least known at its ends, which is one more reason extrapolation beyond the data is dangerous: the formulas themselves warn you by blowing up the width.

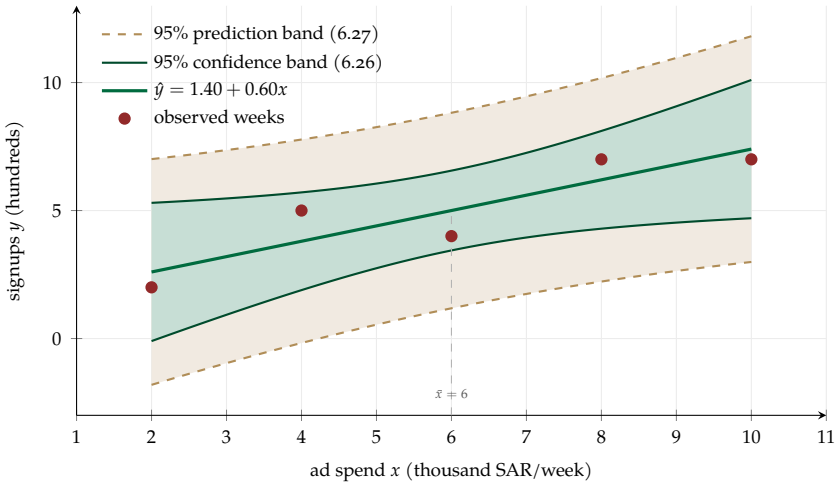


Figure 6.7. Confidence band for the mean response (inner, green) and prediction band for a new observation (outer, gold) for the ad-spend regression. Both bands are narrowest at $\bar{x} = 6$ and widen toward the edges of the data; the prediction band is wider everywhere because of the extra "1+" in (6.27).

| | Confidence interval (6.26) | Prediction interval (6.27) |
|---|---|---|
| Target | $E(Y \mid x_0)$, the mean | $Y_{\text{new}}$, one observation |
| Extra "1+"? | no | yes |
| Width | narrower | always wider |
| As $n$ grows | width shrinks toward 0 | width cannot shrink below $2t\hat{\sigma}$ |
| Question answered | "where is the true line?" | "what will the next run give?" |

Table 6.8. Confidence interval on the mean response versus prediction interval for a new observation. Exam wording: "*mean* response" or "*average*" calls for (6.26); "*predict*", "the *next*", or "a *new*" observation calls for (6.27).

**Procedure 6.3 — Interval estimation at a new setting $x_0$.**

1. **Check the range.** Confirm that $x_0$ lies inside (or very near) the observed range of $x$. If it does not, stop: the model has no evidence there.
2. **Compute the point estimate** $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

3. **Identify the question.** Mean response at $x_0 \rightarrow$ confidence interval (6.26). A single future observation at $x_0 \rightarrow$ prediction interval (6.27).
4. **Compute the standard error**, without the "1+" for the mean response, with it for prediction.
5. **Apply the $t$ multiplier** $t_{\alpha/2,\,n-2}$ and form $\hat{y}_0 \pm t \times \text{se}$.
6. **Interpret in context**, naming what the interval covers (the average, or the next single outcome) with units.

**Example 6.8 — Planning next week's campaign: CI versus PI**

For the ad-spend regression ($\hat{y} = 1.40 + 0.60x$, $n = 5$, $\bar{x} = 6$, $S_{xx} = 40$, $MS_E = 1.200$, $t_{0.025,3} = 3.182$), the growth team plans to spend $x_0 = 8$ thousand SAR next week. Find (a) a 95% confidence interval for the *mean* weekly signups at this spend level, and (b) a 95% prediction interval for *next week's* signups. Compare the two.

***Solution.***

***Step 1 — Range check and point estimate.***

$x_0 = 8$ lies inside the observed range $[2, 10]$. The point estimate is $\hat{y}_0 = 1.40 + 0.60(8) = 6.20$ hundred signups.

***Step 2 — Confidence interval for the mean response.***

By (6.26),

$$\text{se} = \sqrt{1.200\left(\frac{1}{5} + \frac{(8-6)^2}{40}\right)} = \sqrt{1.200\,(0.2000 + 0.1000)} = \sqrt{0.3600} = 0.6000,$$

so $6.20 \pm (3.182)(0.6000) = 6.20 \pm 1.909$, giving $\boxed{4.29 \le E(Y \mid 8) \le 8.11}$ hundred signups.

***Step 3 — Prediction interval for a new week.***

By (6.27),

$$\text{se} = \sqrt{1.200\,(1 + 0.2000 + 0.1000)} = \sqrt{1.560} = 1.249,$$

so $6.20 \pm (3.182)(1.249) = 6.20 \pm 3.974$, giving $\boxed{2.23 \le Y_{\text{new}} \le 10.17}$ hundred signups.

***Step 4 — Compare.***

The CI has width 3.82; the PI has width 7.95, about twice as wide. If the team repeated 8-thousand-SAR weeks many times, the long-run *average* is pinned between 429 and 811 signups; but any *single* week could plausibly land anywhere from 223 to 1017. Promising a client next week's number from the CI would badly overstate the model's precision.

**Engineering judgment.** Choosing between (6.26) and (6.27) is an engineering decision, not a statistical one. Sizing a fuel budget for a whole quarter of runs is a question about the mean: use the CI. Deciding whether one specific truck can finish its route on one tank is a question about a single run: use the PI.

**Reference.** Montgomery & Runger, 6th ed., Section 11-7.1.

## 6.12 *Residual Analysis: Checking the Model*

A significant $F$ test and a high $R^2$ say the line explains variability. They do *not* say the model assumptions of Section 6.2 hold, and every interval and P-value in this chapter leans on those assumptions. The assumptions are statements about the unobservable errors $\varepsilon_i$; the observable stand-ins are the residuals $e_i = y_i - \hat{y}_i$. If the model is adequate, the residuals should look like structureless noise: no trend, no curve, no fan, no extreme points. Residual analysis is the discipline of actually looking.

Two plots do most of the work. The plot of $e_i$ against the fitted values $\hat{y}_i$ (or against $x_i$) reveals curvature and unequal variance. The normal probability plot of the residuals (or, for small samples, a simple dot diagram) checks the normality assumption. To flag outliers, scale each residual by the estimated error standard deviation to get the *standardized residual*,

**Standardized residual.** $d_i = e_i / \sqrt{MS_E}$; under a correct normal model, about 95% of the $d_i$ fall in $(-2, 2)$.

$$d_i = \frac{e_i}{\sqrt{MS_E}} = \frac{e_i}{\hat{\sigma}}. \tag{6.28}$$

If the model and normality are correct, about 95% of the $d_i$ should fall between $-2$ and $+2$; a value beyond $\pm 2$ deserves investigation, and a value beyond $\pm 3$ almost certainly marks a data error or a special cause. Figure 6.8 shows the four patterns you must be able to recognize on sight.

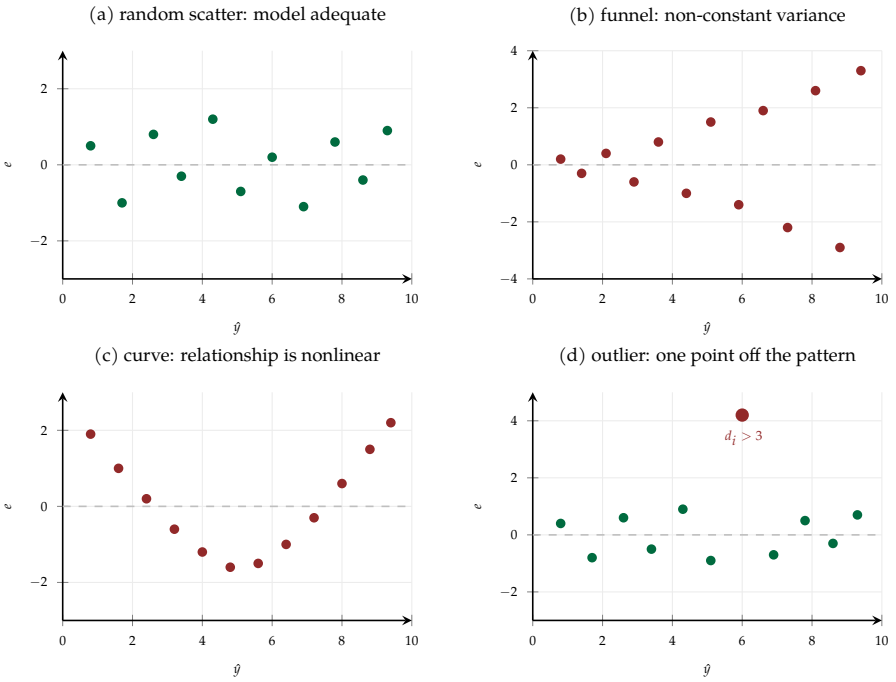


Figure 6.8. Four residual patterns. (a) is what an adequate model produces. (b) violates constant variance: consider transforming $y$ (Section 6.13). (c) means the true relationship is not a straight line: add curvature or transform. (d) shows an outlier: check the data before anything else.

**Procedure 6.4 — Residual analysis.**

1. **Compute the residuals** $e_i = y_i - \hat{y}_i$ for all $i$, and verify $\sum e_i \approx 0$ (an exact property of least squares; a violation is an arithmetic error).
2. **Plot** $e_i$ **versus** $\hat{y}_i$ (or versus $x_i$). Look for the patterns of Figure 6.8: a curve (nonlinearity), a funnel (non-constant variance), or any systematic structure.
3. **If the data have a time order, plot** $e_i$ **in that order.** Long runs of same-sign residuals suggest correlated errors, which invalidate the standard errors even when the fit looks good.

4. **Check normality** with a normal probability plot or, for small $n$, a dot diagram of the residuals.
5. **Compute standardized residuals** $d_i$ by (6.28); investigate any $|d_i| > 2$ and treat any $|d_i| > 3$ as a likely data error or special cause. Never delete a point without an identified reason.
6. **Act on what you find.** Curvature or a funnel: transform a variable (Section 6.13) and refit. Outlier with a confirmed cause: correct or remove, then refit. Clean plots: the analysis stands.

**Example 6.9 — Residual check for the ad-spend model**

Carry out steps 1 and 5 of Procedure 6.4 for the ad-spend regression ($\hat{y} = 1.40 + 0.60x$, $MS_E = 1.200$): compute all residuals and standardized residuals, and decide whether any week is an outlier.

***Solution.***

***Step 1 — Residuals.***

With $\hat{\sigma} = \sqrt{1.200} = 1.095$, Table 6.9 lists $\hat{y}_i$, $e_i$, and $d_i = e_i/1.095$ for the five weeks.

| $i$ | $x_i$ | $y_i$ | $\hat{y}_i$ | $e_i$ | $d_i = e_i/\hat{\sigma}$ |
|---|---|---|---|---|---|
| 1 | 2 | 2 | 2.60 | $-0.60$ | $-0.55$ |
| 2 | 4 | 5 | 3.80 | $+1.20$ | $+1.10$ |
| 3 | 6 | 4 | 5.00 | $-1.00$ | $-0.91$ |
| 4 | 8 | 7 | 6.20 | $+0.80$ | $+0.73$ |
| 5 | 10 | 7 | 7.40 | $-0.40$ | $-0.37$ |
| $\sum$ | | | | 0.00 | |

Table 6.9. Residuals and standardized residuals for Example 6.9.

***Step 2 — Checks.***

The residuals sum to $-0.60 + 1.20 - 1.00 + 0.80 - 0.40 = 0.00$, as least squares guarantees. Every standardized residual satisfies $|d_i| < 2$; the largest in magnitude is $d_2 = +1.10$. The answer is $\boxed{\text{no outliers}}$: with only five

points no strong pattern can be read, but nothing in the residuals contradicts the model.

**Reference.** Montgomery & Runger, 6th ed., Section 11-9.

**Transformation.** Replacing $y$, $x$, or both by functions of themselves (such as $\ln y$ or $1/x$) so that the relationship becomes linear in the new variables.

## 6.13 *Regression on Transformed Variables*

When the scatter diagram or the residual plot shows a curve, the straight line model is wrong for the data as recorded, but it may be exactly right for a *transformation* of the data. The strategy: transform $y$, $x$, or both so that the relationship in the new scale is linear, then apply every formula of this chapter unchanged to the transformed variables. No new machinery is needed; that is the whole point.

Two transformations cover most engineering cases. If the true relationship is *exponential*, $y = \beta_0 e^{\beta_1 x}$, taking natural logs of both sides gives

$$\ln y = \ln \beta_0 + \beta_1 x,$$

a straight line of $y' = \ln y$ against $x$. If the relationship is a *power law*, $y = \beta_0 x^{\beta_1}$, then

$$\ln y = \ln \beta_0 + \beta_1 \ln x,$$

a straight line of $y' = \ln y$ against $x' = \ln x$. In both cases the fitted intercept lives in the log scale, and the original-scale constant is recovered as $e^{\hat{\beta}'_0}$. Table 6.10 summarizes which pattern calls for which transformation; a funnel-shaped residual plot (variance growing with the mean) is also commonly fixed by $\ln y$ or $\sqrt{y}$, because these transformations compress the large values where the scatter is widest.

| Pattern in the data | Underlying form | Transformation |
|---|---|---|
| Exponential growth or decay | $y = \beta_0 e^{\beta_1 x}$ | $y' = \ln y$ |
| Power-law relationship | $y = \beta_0 x^{\beta_1}$ | $y' = \ln y,\ x' = \ln x$ |
| Variance grows with mean (funnel) | — | $y' = \ln y$ or $y' = \sqrt{y}$ |
| Inverse relationship | $y = \beta_0 + \beta_1 / x$ | $x' = 1/x$ |

Table 6.10. Common transformations. After transforming, every formula of this chapter applies to the primed variables unchanged.

**Example 6.10 — Container backlog growing exponentially**

A port disruption leaves a container backlog that appears to grow exponentially with time. For days $x = 1, \dots, 5$, the backlog $y$ (hundreds of TEU) gives a strongly curved scatter (Figure 6.9, left), so the analyst transforms to $y' = \ln y$ and computes the summary statistics of the transformed data:

$$n = 5, \quad \bar{x} = 3.0, \quad \bar{y}' = 1.80, \quad S_{xx} = 10.0, \quad S_{xy'} = 6.30, \quad S_{y'y'} = 3.98.$$

(a) Fit the linear model in the transformed scale. (b) Compute $R^2$. (c) Predict the backlog on day 6, in TEU. (d) Interpret the slope in the original scale.

***Solution.***

***(a) Fit.***

By (6.9) and (6.10) applied to $(x, y')$:

$$\hat{\beta}_1 = \frac{6.30}{10.0} = 0.6300, \qquad \hat{\beta}_0' = 1.80 - (0.6300)(3.0) = -0.0900,$$

so $\boxed{\widehat{\ln y} = -0.090 + 0.630\,x}$.

***(b) Fit quality.***

$SS_R = \hat{\beta}_1 S_{xy'} = (0.6300)(6.30) = 3.969$, and by (6.21), $R^2 = 3.969/3.98 = 0.9972$: the log-linear model explains 99.7% of the variability of $\ln y$.

***(c) Prediction on day 6.***

In the log scale, $\hat{y}' = -0.090 + 0.630(6) = 3.690$. Back-transforming, $\hat{y} = e^{3.690} = 40.04$ hundreds of TEU, so about $\boxed{4000 \text{ TEU}}$. Note that day 6 is one step beyond the observed range; a short-horizon extrapolation of a mechanism known to persist (the disruption is ongoing) is defensible, but it should be flagged as an extrapolation all the same.

***(d) Slope interpretation.***

In the original scale the model is $y = e^{-0.090} e^{0.630x}$: each day multiplies the backlog by $e^{0.630} = 1.878$, an 88% daily growth rate. In log-scale models the slope is a *growth rate*, not an additive change.

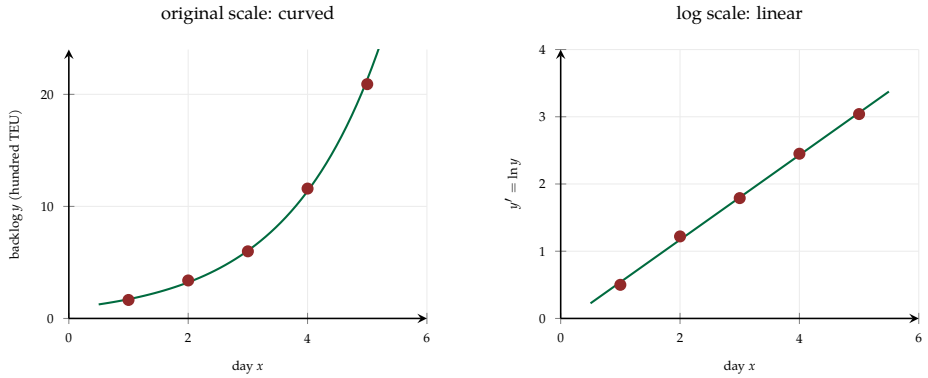


Figure 6.9. The backlog data of Example 6.10. Left: in the original scale the growth is exponential and a straight line would fit badly. Right: after the transformation $y' = \ln y$, the trend is linear and ordinary least squares applies.

**Safety lens.** Log transformations are the working language of AI scaling laws: model loss versus training compute is close to a power law, so researchers fit straight lines in log-log scale, exactly the second row of Table 6.10. The regression mechanics are this chapter's; only the axes are logarithmic.

**Reference.** Montgomery & Runger, 6th ed., Section 11-4 (comments) and 11-8.

## 6.14 *Abuses of Regression*

Two misuses of regression cause more real damage than every computational error combined. Both have appeared in passing; this section states them plainly.

**Extrapolation.** The fitted line summarizes the data over the observed range of $x$, and only there. Outside that range the data contain no information: the true relationship may bend, saturate, or break entirely, and the fit gives no warning (Figure 6.10). The verification model of Example 6.1 predicted $-114$ failures per 1000 at $x = 200$ hours: an absurd output produced by perfectly correct arithmetic. Two safeguards are built into this chapter: step 1 of Procedure 6.3 refuses out-of-range $x_0$, and the interval widths (6.26) and (6.27) grow as $x_0$ leaves the center of the data. When a prediction outside the range is genuinely needed, the honest options are to collect data there or to bring in subject-matter theory; a longer line is not evidence.

**Extrapolation.** Using a fitted model at $x$ values outside the range of the data that produced it.

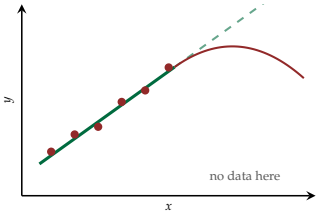


Figure 6.10. Inside the data (solid) the line is supported by evidence. Beyond it, the fitted line (dashed) and the truth (maroon) are free to part ways, and nothing in the data can warn you.

**Association is not causation.** A strong regression on observational data shows that $x$ and $y$ move together; it does not show that changing $x$ would change $y$. A *lurking variable* can drive both. Warehouses with more overtime hours also record more safety incidents, with a beautiful $R^2$, because shipment volume drives both overtime and incidents; cutting overtime would not deliver the incident reduction the regression seems to promise. The designed experiments of Chapters 8 and 9, where the engineer *sets* $x$ and randomizes everything else, are what license

**Lurking variable.** An unmeasured variable that drives both $x$ and $y$, creating an association between them without any causal link.

**Startup lens.** Growth dashboards are lurking-variable factories. Signups and ad spend both rise before Ramadan and both fall in summer; a regression of one on the other partly measures the calendar, not the ads. Before scaling a budget on the strength of a slope, run the experiment: randomize spend across comparable weeks or regions, then Chapter 5's two-sample tools apply to something you actually controlled.

causal conclusions. For observational regressions, the correct language is the one used throughout this chapter: ''associated with,'' not ''causes.''

## *Chapter Summary*

This chapter built the complete workflow for modeling one response against one regressor. The simple linear regression model (6.2) places a normal distribution of responses at each $x$, centered on the line $\beta_0 + \beta_1 x$ with constant variance $\sigma^2$. The *method of least squares* minimizes the sum of squared residuals; solving the normal equations gives the estimators $\hat{\beta}_1 = S_{xy}/S_{xx}$ (6.9) and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ (6.10), computed by the routine of Procedure 6.1. The error variance is estimated by $\hat{\sigma}^2 = SS_E/(n-2)$ (6.14). Both estimators are unbiased, with standard errors (6.15) and (6.16), so the familiar estimate-$\pm$-$t\times$se machinery applies: the slope $t$ test (6.18) (Procedure 6.2), the equivalent regression $F$ test (6.20) built on the ANOVA identity $SS_T = SS_R + SS_E$ (6.19), and confidence intervals (6.23) and (6.24). Fit quality is summarized by $R^2 = SS_R/SS_T$ (6.21) and the correlation $r$ (6.22), with $R^2 = r^2$. At a new setting $x_0$, the confidence interval for the mean response (6.26) and the prediction interval for a single new observation (6.27) differ by the ''1+'' term, and both widen away from $\bar{x}$ (Procedure 6.3). Residual analysis (Procedure 6.4) checks the assumptions the inference rests on; curved or funnel-shaped residuals call for the transformations of Section 6.13. And two abuses must be refused every time: extrapolating beyond the data, and reading causation into association.

| Quantity | Formula | Equation |
|---|---|---|
| Model | $Y = \beta_0 + \beta_1 x + \varepsilon$ | (6.2) |
| Building blocks | $S_{xx} = \sum x_i^2 - (\sum x_i)^2/n$ | (6.7) |
| | $S_{xy} = \sum x_i y_i - (\sum x_i)(\sum y_i)/n$ | (6.8) |
| Slope estimator | $\hat{\beta}_1 = S_{xy}/S_{xx}$ | (6.9) |
| Intercept estimator | $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ | (6.10) |
| Error variance | $\hat{\sigma}^2 = MS_E = SS_E/(n-2)$ | (6.14) |
| SE of slope | $\text{se}(\hat{\beta}_1) = \sqrt{MS_E/S_{xx}}$ | (6.15) |
| SE of intercept | $\text{se}(\hat{\beta}_0) = \sqrt{MS_E(1/n + \bar{x}^2/S_{xx})}$ | (6.16) |
| Slope $t$ test | $T_0 = \hat{\beta}_1/\sqrt{MS_E/S_{xx}} \sim t_{n-2}$ | (6.18) |
| Slope CI | $\hat{\beta}_1 \pm t_{\alpha/2,n-2}\sqrt{MS_E/S_{xx}}$ | (6.23) |
| ANOVA identity | $SS_T = SS_R + SS_E$ | (6.19) |
| Regression $F$ | $F_0 = MS_R/MS_E = T_0^2$ | (6.20) |
| CI, mean response | $\hat{y}_0 \pm t_{\alpha/2,n-2}\sqrt{MS_E\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$ | (6.26) |
| PI, new observation | $\hat{y}_0 \pm t_{\alpha/2,n-2}\sqrt{MS_E\left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$ | (6.27) |
| Coefficient of determination | $R^2 = SS_R/SS_T = 1 - SS_E/SS_T$ | (6.21) |
| Sample correlation | $r = S_{xy}/\sqrt{S_{xx}S_{yy}},\ r^2 = R^2$ | (6.22) |
| Standardized residual | $d_i = e_i/\sqrt{MS_E}$ | (6.28) |

Table 6.11. Key formulas of Chapter 6 at a glance.

## *Exercises*

### *Fitting the line*

**Exercise 6.1.** The method of least squares chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize which quantity?

(A) the sum of the residuals, $\sum e_i$

(B) the sum of the absolute residuals, $\sum |e_i|$

(C) the sum of the squared vertical deviations, $\sum (y_i - \beta_0 - \beta_1 x_i)^2$

(D) the sum of the perpendicular distances from the points to the line

Also state two properties that every least squares fit automatically has.

**Exercise 6.2.** A red team probes an LLM guardrail in successive hardening rounds. After each round $x$ (1 to 5), the number of jailbreak prompts (out of a fixed suite) that still get through is recorded:

| round $x$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| jailbreaks $y$ | 10 | 8 | 7 | 5 | 4 |

Fit the least squares line and interpret the slope in context.

**Exercise 6.3.** A startup's fitted model relating support tickets $y$ (per day) to active users $x$ (thousands) is $\hat{y} = 3.2 + 0.45x$. On a day with $x = 10$ thousand users, 8.5 tickets arrived. Compute the fitted value and the residual for that day, and state whether the model over- or underestimated the ticket load.

### *Inference on the slope*

**Exercise 6.4.** A rail operator regresses unloading time $y$ (minutes) on freight load $x$ (tonnes) for $n = 12$ trains and obtains $\hat{\beta}_1 = 0.350$, $MS_E = 1.960$, $S_{xx} = 800.0$. (a) Test $H_0 \colon \beta_1 = 0$ against $H_1 \colon \beta_1 \neq 0$ at $\alpha = 0.05$, given $t_{0.025,10} = 2.228$. (b) Give a 95% confidence interval for $\beta_1$ and connect it to the test result.

**Exercise 6.5.** An ML team regresses a model's evaluation score $y$ on fine-tuning epochs $x$ over $n = 8$ training runs and gets $\hat{\beta}_1 = 0.90$, $MS_E = 4.80$, $S_{xx} = 30.0$. Using $t_{0.025,6} = 2.447$, test whether the epoch effect is significant at $\alpha = 0.05$, give the 95% CI for $\beta_1$, and explain what the team should conclude.

### *Prediction and intervals*

**Exercise 6.6.** A warehouse regresses picking hours $y$ on orders received $x$ (tens) over $n = 10$ days: $\hat{y} = 5.0 + 1.2x$, with $\bar{x} = 20$, $S_{xx} = 250$, $MS_E = 2.50$, and $t_{0.025,8} = 2.306$. For a day with $x_0 = 25$: (a) find a 95% confidence interval for the *mean* picking hours at this order level; (b) find a 95% prediction interval for a *single* such day; (c) explain why (b) is wider, and at which $x_0$ both intervals are narrowest.

**Exercise 6.7.** The verification model of Example 6.1, $\hat{y} = 54.87 - 0.8452x$, was fitted on testing budgets between 10 and 45 hours. A planner asks for the predicted failure rate at $x = 120$ hours. Compute what the formula gives, explain why the number should not be reported, and state what Procedure 6.3 says to do instead.

### *$R^2$ and correlation*

**Exercise 6.8.** A regression of daily signups on ad impressions over $n = 22$ days produced this partial ANOVA table:

| Source | df | SS | MS | $F_0$ |
|---|---|---|---|---|
| Regression | 1 | 120 | 120 | ? |
| Error | ? | ? | 4.0 | |
| Total | ? | ? | | |

Complete the table, then find $R^2$ and test the significance of regression at $\alpha = 0.05$, given $f_{0.05,1,20} = 4.35$.

**Exercise 6.9.** A port analyst reports a sample correlation of $r = -0.95$ between the number of cranes down for maintenance and the containers moved per shift. (a) What is $R^2$ for the corresponding regression, and what does it mean? (b) What is the sign of the fitted slope, and why? (c) A colleague says "$r = -0.95$ means each crane down costs us 0.95 containers." Correct the misunderstanding.

### *Residuals and model adequacy*

**Exercise 6.10.** After fitting delivery cost against route distance for a courier fleet, an analyst plots residuals against fitted values and sees the scatter widening steadily from left to right, like a funnel opening. Which model assumption is violated, what does the violation do to the chapter's tests and intervals, and what remedy does Procedure 6.4 suggest?

**Exercise 6.11.** A dwell-time regression at a container terminal has $MS_E = 1.44$ (hours$^2$). Two observations have residuals $e_7 = +2.4$ hours and $e_{13} = -4.2$ hours. Compute the standardized residuals and classify each observation according to Procedure 6.4.

*Exam-style problems*

**Exercise 6.12.** A distribution center studies how the number of forklifts assigned to a shift ($x$) affects the pallets moved per hour ($y$, tens). Six shifts give:

| $x$ (forklifts) | 2 | 4 | 6 | 8 | 10 | 12 |
|---|---|---|---|---|---|---|
| $y$ (tens of pallets/hr) | 3 | 5 | 4 | 7 | 8 | 9 |

Summary values: $\sum x = 42$, $\sum y = 36$, $\sum x^2 = 364$, $\sum y^2 = 244$, $\sum xy = 294$. Critical values: $f_{0.05,1,4} = 7.71$, $t_{0.025,4} = 2.776$.

1. **[3 pt]** Compute $S_{xx}$, $S_{xy}$, $S_{yy}$ and fit the least squares line.
2. **[2 pt]** Build the complete ANOVA table.
3. **[1 pt]** Test the significance of regression at $\alpha = 0.05$ using the $F$ statistic.
4. **[2 pt]** Give a 95% confidence interval for $\beta_1$ and connect it to part (c).
5. **[2 pt]** Predict $y$ at $x = 8$, find the residual of the observed shift with $x = 8$, and compute $R^2$.

**Exercise 6.13.** A subscription startup believes monthly churn rate $y$ (percent) follows a power law in account age $x$ (weeks): $y = \beta_0 x^{\beta_1}$. Taking $y' = \ln y$ and $x' = \ln x$ for $n = 6$ customer cohorts gives

$$\bar{x}' = 4.00, \quad \bar{y}' = 1.20, \quad S_{x'x'} = 8.00, \quad S_{x'y'} = -4.00, \quad S_{y'y'} = 2.05.$$

Critical value: $t_{0.025,4} = 2.776$.

1. **[2 pt]** Fit the linear model in the transformed variables.
2. **[1 pt]** Compute $R^2$ for the transformed model.
3. **[2 pt]** Test $H_0 \colon \beta_1 = 0$ at $\alpha = 0.05$.
4. **[2 pt]** Estimate the churn rate of a cohort with $x' = 5$ (about 148 weeks old), in the *original* scale.
5. **[1 pt]** Interpret $\hat{\beta}_1$: what happens to churn when account age doubles?

*Assessing outside recommendations*

**Exercise 6.14.** An AI assistant analyzes the ad-spend data of Example 6.4 ($x$ between 2 and 10 thousand SAR, $\hat{y} = 1.40 + 0.60x$, $R^2 = 0.80$) and reports: ''The model is strong ($R^2 = 0.80$). Recommendation: raise weekly spend to 60,000 SAR, which the model predicts will deliver $\hat{y} = 1.40 + 0.60(60) = 37.4$ hundred, i.e. about 3,740 signups per week.'' Identify every flaw in this recommendation, explain the consequences, and describe what a correct analysis would do.

**Exercise 6.15.** A consultant regresses monthly safety incidents $y$ on overtime hours $x$ across your company's 15 warehouses and reports $R^2 = 0.93$ with a highly significant slope. The report concludes: ''Overtime *causes* incidents; the high $R^2$ proves it, and it also confirms the regression assumptions are satisfied. Cap overtime and incidents will fall proportionally.'' Identify the two distinct errors, give a plausible alternative explanation for the finding, and state what evidence would actually support the recommendation.

## In-Class Activity 6.1: Draw the Line Before You Trust It

Week 8 — Lecture 15

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

---

Your team is evaluating a warehouse-robot perception module. Each build of the module received a different amount of adversarial stress testing, and its undetected-failure count in a fixed acceptance suite was recorded:

| stress-test batches $x$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| undetected failures $y$ | 9 | 7 | 6 | 4 | 4 |

### *Part 1 — Before any arithmetic*

Identify the regressor and the response, sketch the scatter by hand, and state (with one sentence of reasoning) what sign you expect for $\hat{\beta}_1$.

### *Part 2 — Fit the line*

Compute $\sum x$, $\sum y$, $\sum x^2$, $\sum xy$, then $S_{xx}$, $S_{xy}$, $\hat{\beta}_1$, and $\hat{\beta}_0$. Write the fitted model and check that your slope sign matches Part 1.

*t, carefully*

4 and compute the residual of the observed build with $x = 4$. Then: a teammate wants the e count at $x = 20$ batches. In two sentences, explain whether your team should provide it.

rough which point does every least squares line pass, and why is that a useful arithmetic check?

## In-Class Activity 6.2: One Line Through the Port

Week 9 — Lecture 17

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

---

A terminal supervisor suspects that a container ship's unloading time grows linearly with the number of ships already queued when it arrives. Five arrivals give:

| ships queued $x$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| unloading time $y$ (hours) | 4 | 6 | 5 | 8 | 7 |

### *Part 1 — Fit*

Compute the totals, then $S_{xx}$, $S_{xy}$, $\hat{\beta}_1$, $\hat{\beta}_0$, and write the fitted model with units.

### *Part 2 — Residual audit*

Compute all five fitted values and residuals, and verify that the residuals sum to zero. Which arrival does the model miss by the most?

*pret for the supervisor*

nce for the supervisor interpreting $\hat{\beta}_1$ in context, and one sentence explaining why the model sed to quote unloading times for a 20-ship queue.

what does a positive residual say about the model's estimate for that observation?

## In-Class Activity 6.3: Is the Trend Real or Noise?

Week 9 — Lecture 18

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

---

Your startup regressed weekly revenue on weekly active users over $n = 12$ weeks. The fit produced a positive slope with $SS_T = 50.0$ and $SS_R = 40.0$. Critical value: $f_{0.05,1,10} = 4.96$.

### *Part 1 — Complete the ANOVA table*

Fill in every entry: sources, degrees of freedom, sums of squares, mean squares, and $F_0$.

### *Part 2 — Test and translate*

Carry out the $F$ test at $\alpha = 0.05$, then write the conclusion as one sentence a non-statistician investor would understand.

*summaries, one story*

l the value of $|T_0|$ that the equivalent $t$ test would give (no $t$ table needed). How do you know re?

th one reason: a significant $F$ test confirms that the regression assumptions are satisfied.

## In-Class Activity 6.4: Two Intervals, One Decision

Week 10 — Lecture 19

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

---

An AV team models the distance between disengagements $y$ (km) against the volume of curated training scenarios $x$ (thousands). From $n = 10$ software builds: $\hat{y} = 120 + 8x$, $\bar{x} = 5$, $S_{xx} = 60$, $MS_E = 25$, and $t_{0.025,8} = 2.306$.

### *Part 1 — Which interval does each question need?*

For each question below, state CI on the mean response or PI for a new observation, with one clause of justification:
(i) ''What average disengagement distance should we expect across the whole fleet if all builds train on $x_0 = 5$?''
(ii) ''The next build trains on $x_0 = 5$; what range should its road-test result fall in?''

### *Part 2 — Compute both at $x_0 = 5$*

Find the point estimate, then the 95% CI and the 95% PI. Show the standard error computation for each.

*the residuals*

lots from rival models are described: (a) a fan opening to the right; (b) a clear U-shape; (c) itive residuals followed by long runs of negative ones when plotted in build order. Name the tion in each case and one remedy or follow-up.

hy can a prediction interval never shrink to zero width, even with unlimited data?