

Regression as statistical modeling

Hicham Janati

Fall 2026

For two chapters we modeled the *joint* distribution of the row: one Gaussian when the data form one cluster, a mixture of Gaussians when they form several. That is the most complete thing a statistician can do with a table, and it is not always the right thing. It can be *too complicated*: the joint model's number of parameters grows quadratically with the dimension, and a mixture multiplies that count by the number of components. And it can be *unnecessary*: often one component of the vector has a special status (the quantity to be predicted, priced, or acted on), and the question is not how the whole row is distributed but what that one component is worth when the others are known. Call it Y , and call the rest $\mathbf{X}$. The prologue already named the way out: the joint distribution factors into the law of $\mathbf{X}$ times the conditional law of Y given $\mathbf{X}$, and prediction needs only the second factor. This chapter models that factor alone, regardless of what the distribution of $\mathbf{X}$ looks like, and it does so in the spirit of Part 1: a parametric model, its assumptions examined one at a time, and exact inference. The model is one every reader has met before, the linear model of first year; the point of the chapter is to re-read it as a *statistical model*: what each assumption buys, what the tests actually say, and what happens to twenty-eight of them at once.

Throughout, the course conventions hold: bold letters denote vectors and matrices, plain italic letters denote scalars.

1 From the joint to a conditional model

Return to the store of chapter 2, and zoom in on the segment the mixture called the *regulars*: $n = 400$ customers. For each of them the database records the monthly spend, and three more columns: visits per month, household size, and distance to the store in kilometers,

	spend (Dh)	visits / month	household	distance (km)
customer 1	362	13.0	3	7.8
customer 2	222	10.7	2	6.6
customer 3	977	11.7	5	4.9
customer 4	753	13.8	4	4.2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
customer 400	334	11.4	2	6.2

The store's question is not symmetric in these four quantities. Spend is the one it acts on (revenue to forecast, offers to target), and the other three are context: how much does an extra visit bring? does distance cost money? what should the store expect from a new customer whose three covariates are known? Write Y for the spend and $\mathbf{x} = (x_1, \dots, x_p)^\top$ for the covariates ($p = 3$ here).

A full chapter-1 treatment would model the joint distribution of the vector $(\mathbf{X}, Y)$ on $\mathbb{R}^{p+1}$. A Gaussian model of the four-dimensional row needs $(p+1) + (p+1)(p+2)/2$ parameters, asserts joint Gaussianity of covariates that may be nothing of the sort (household size takes whole values; distance is skewed), and spends most of its parameters describing how the covariates co-vary among themselves, which the store's question never asks. The conditional route keeps only what the question needs. When a vector is jointly Gaussian, the conditional law of one coordinate given the others is exactly Gaussian, with a mean *affine* in the observed values and a variance

that does not depend on them. The conditional model takes that conclusion and promotes it to an assumption, while dropping every claim about $\mathbf{X}$:

Definition 1.1 (Gaussian linear model)

The *Gaussian linear model* for a response Y given covariates $\mathbf{x} \in \mathbb{R}^p$ is the conditional model

$$Y \mid \mathbf{X} = \mathbf{x} \sim \mathcal{N}(\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}, \sigma^2), \quad \text{equivalently} \quad Y = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x} + \varepsilon, \quad (1)$$

with parameters $\beta_0 \in \mathbb{R}$, $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\sigma^2 > 0$, and with $\varepsilon := Y - \beta_0 - \boldsymbol{\beta}^\top \mathbf{x}$ the deviation from the affine mean. The assumptions, stated separately:

- (A1) *correct mean (linearity and exogeneity)*: $\mathbb{E}(Y \mid \mathbf{X} = \mathbf{x}) = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}$ for every $\mathbf{x}$, equivalently $\mathbb{E}(\varepsilon \mid \mathbf{X}) = 0$;
- (A2) *independent rows*: the customers are sampled independently (given the design, the deviations $\varepsilon_1, \dots, \varepsilon_n$ of distinct rows are independent);
- (A3) *homoscedasticity*: $\mathbb{V}(\varepsilon \mid \mathbf{X}) = \sigma^2$, the same constant at every $\mathbf{x}$;
- (A4) *conditional Gaussianity*: given $\mathbf{X} = \mathbf{x}$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

A word on (A1), which some textbooks split into two: “linearity” of the mean and “exogeneity” of the noise. With ε defined as the deviation from the affine mean, the two are one and the same statement, and we state it once. The content that matters in practice is this: (A1) claims that the conditional mean is correctly specified *with the covariates included in the model*. Its typical failure is not a curve where a line was drawn but an *absent* covariate: when a variable that influences Y is left out and is correlated with an included one, the included coefficient absorbs part of the missing effect. Section 5 will show why this particular failure is the hardest to detect.

Nothing is assumed about the distribution of $\mathbf{X}$: the covariates may be discrete, skewed or dependent. This is where the economy comes from: instead of a distribution on $\mathbb{R}^{p+1}$, the model has $p + 2$ parameters. Figure 1 draws the model’s claim on the regulars’ data, for the single covariate visits: one straight line for the conditional mean, and around it the same Gaussian at every visit frequency (same shape, same width, only the center moving along the line). Whether the data agree with that claim is a question the chapter takes up in section 5; until then, they are working assumptions.

Reality check

(A1)–(A4) are not hypotheses to be verified and ticked off: no dataset can prove that a conditional mean is exactly affine or that errors are exactly Gaussian. They are working assumptions. Each one, as section 3 maps out, buys a specific inferential product, and each one, when wrong, silently invalidates what it bought. The honest attitude is the one this course keeps returning to: put the assumptions to work, then gather evidence of how far off they are (section 5), knowing the evidence can refute them but never certify them.

2 Least squares is maximum likelihood

Stack the n observed rows: let $\mathbf{y} = (y_1, \dots, y_n)^\top$ collect the responses and let the *design matrix* $\mathbb{X} \in \mathbb{R}^{n \times (p+1)}$ have rows $(1, \mathbf{x}_i^\top)$ (a column of ones for the intercept, then the covariates). Absorbing β_0 into $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^\top$, the n equations (1) become

$$\mathbf{y} = \mathbb{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n), \quad (2)$$

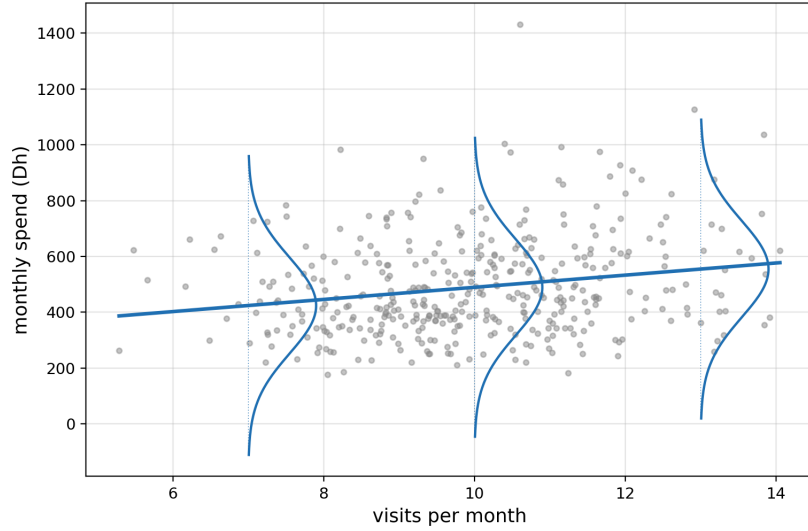


Figure 1: The model’s claim, drawn on the regulars’ spend against visits: a straight conditional mean and the same Gaussian conditional distribution at every visit frequency (three slices shown). Whether the data agree is section 5’s question.

the independence across rows being assumption (A2) at work. Under the model, the likelihood of the observed responses given the design is

$$L(\beta, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i^\top \beta)^2}{2\sigma^2}\right) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\|\mathbf{y} - \mathbb{X}\beta\|^2}{2\sigma^2}\right),$$

writing $\mathbf{x}_i$ for the i -th row of the design. Whatever σ , maximizing L over β means minimizing the exponent: maximum likelihood in the Gaussian linear model *is* least squares, $\min_{\beta} \|\mathbf{y} - \mathbb{X}\beta\|^2$, the first-year criterion, now obtained as a consequence of the model.

Theorem 2.1 (OLS estimator)

If $\mathbb{X}^\top \mathbb{X}$ is invertible, the likelihood is maximized at

$$\hat{\beta} = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \mathbf{y}, \quad \hat{\sigma}_{\text{MLE}}^2 = \frac{\text{RSS}}{n}, \quad \text{RSS} = \|\mathbf{y} - \mathbb{X}\hat{\beta}\|^2. \quad (3)$$

Proof. The map $g(\beta) = \|\mathbf{y} - \mathbb{X}\beta\|^2 = \mathbf{y}^\top \mathbf{y} - 2\beta^\top \mathbb{X}^\top \mathbf{y} + \beta^\top \mathbb{X}^\top \mathbb{X} \beta$ has gradient $\nabla g = -2\mathbb{X}^\top \mathbf{y} + 2\mathbb{X}^\top \mathbb{X} \beta$. Setting it to zero gives the *normal equations* $\mathbb{X}^\top \mathbb{X} \beta = \mathbb{X}^\top \mathbf{y}$, whose unique solution is $\hat{\beta}$ when $\mathbb{X}^\top \mathbb{X}$ is invertible; the critical point is the global minimum because the Hessian $2\mathbb{X}^\top \mathbb{X}$ is positive definite (for any $\mathbf{u} \neq 0$, $\mathbf{u}^\top \mathbb{X}^\top \mathbb{X} \mathbf{u} = \|\mathbb{X}\mathbf{u}\|^2 > 0$ by invertibility). With $\hat{\beta}$ fixed, the likelihood as a function of σ^2 is $(\sigma^2)^{-n/2} e^{-\text{RSS}/(2\sigma^2)}$, maximized at $\sigma^2 = \text{RSS}/n$, the same one-dimensional maximization as chapter 1’s Gaussian MLE. $\square$

As in chapter 1, the MLE of the variance is biased; the unbiased version divides by the residual degrees of freedom,

$$\hat{\sigma}^2 = \frac{\text{RSS}}{n - p - 1}, \quad (4)$$

$n - p - 1$ playing the role that $n - 1$ played for the sample variance: one degree of freedom is spent per fitted coefficient.

The estimator fluctuates, with a known law. Substituting (2) into (3),

$$\hat{\beta} = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top (\mathbb{X} \beta + \varepsilon) = \beta + (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \varepsilon :$$

conditionally on the design, $\hat{\beta}$ is an *affine map of a Gaussian vector*, precisely the situation chapter 1’s stability property handles. Its mean is β and its covariance is $\mathbf{A} \sigma^2 \mathbf{I}_n \mathbf{A}^\top$ with $\mathbf{A} = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top$, which collapses to:

$$\hat{\beta} | \mathbb{X} \sim \mathcal{N}(\beta, \sigma^2 (\mathbb{X}^\top \mathbb{X})^{-1}). \quad (5)$$

Moreover $\text{RSS}/\sigma^2 \sim \chi^2(n - p - 1)$ and RSS is independent of $\hat{\beta}$ (the Cochran theorem again, stated without proof exactly as in chapter 1). These two facts are the basis of all the inference in this chapter.

3 What each assumption buys

The definition listed four assumptions. They are not a package: each one buys a specific product, and the products stack. Adding them one at a time, starting from none, shows what each contributes.

No assumptions: a fit. With no model at all, OLS is arithmetic: the line (or hyperplane) closest to the cloud in squared distance. It always exists and always produces numbers, but nothing can be said about those numbers: they estimate no defined target and carry no error bar. Inference requires a model.

A correct mean buys unbiasedness. Under (A1), the estimator has a target and hits it on average: taking conditional expectations in the displayed decomposition above,

$$\mathbb{E}(\hat{\beta} | \mathbb{X}) = \beta + (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \mathbb{E}(\varepsilon | \mathbb{X}) = \beta,$$

the last step being exactly (A1)’s exogeneity form. This is the assumption that most often fails in practice: if a covariate that influences Y is missing from the model and correlated with an included one, the noise absorbs it, (A1) fails, and the included coefficient is biased, partly reporting the absent variable’s effect. No amount of data repairs the bias.

Independence buys the error accounting. Assumption (A2) is easy to forget because the customers of a loyalty file satisfy it so naturally, and it is the one that makes the covariance computation of section 2 possible at all: independent rows mean the error covariance matrix is diagonal, so information adds up row by row. When rows are dependent (several customers from the same household, or repeated measurements of the same person), the file contains fewer effective observations than lines, and the standard errors computed as if n rows were n pieces of information come out wrong, usually too small. This failure is in one respect worse than the funnel of section 5: neither residual plot shows it, because each residual taken alone looks perfectly ordinary.

Homoscedasticity buys the standard errors. Under (A1)–(A3), the covariance formula in (5) is correct, so the reported standard error of each coefficient, $\text{SE}(\hat{\beta}_j) = \hat{\sigma} \sqrt{[(\mathbb{X}^\top \mathbb{X})^{-1}]_{jj}}$, estimates the real sampling fluctuation. If the noise variance instead changes with $\mathbf{x}$, OLS remains unbiased but the formula is wrong, and every quantity computed from a standard error (tests, intervals) is wrong with it. The same assumptions also carry a classical optimality: among all linear unbiased estimators, OLS has the smallest variance (the Gauss–Markov theorem, a classical result we use without proof).

Gaussianity buys exact finite- n inference. Under the full model (A1)–(A4), the laws of section 2 are exact for every n , and they deliver the t and F tests and confidence intervals of the next section at their stated levels, for $n = 30$ as for $n = 400$, with no large-sample approximation. Without (A4), those levels are only approximations, good when n is large (the CLT at work), unreliable when it is not.

Each of these assumptions can be wrong, and its inferential product is then lost. The next section puts the products to work; section 5 then confronts the assumptions with the data, and the chapter’s closing table lists each assumption, its product, and whether the diagnostics can detect its failure.

4 Testing a coefficient

Does distance to the store really cost money, once visits and household size are accounted for? The question is about β_{dist} , and the model makes it answerable with a number.

4.1 Testing one coefficient

From (5), the j -th coefficient satisfies $\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2 v_j)$ with $v_j = [(\mathbb{X}^\top \mathbb{X})^{-1}]_{jj}$. Standardizing and replacing the unknown σ by $\hat{\sigma}$ produces a statistic whose law is exactly known under the model; the replacement is legitimate because RSS is independent of $\hat{\beta}$:

$$T_j = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma} \sqrt{v_j}} \sim t(n - p - 1), \quad (6)$$

Student’s distribution with $n - p - 1$ degrees of freedom. Two products follow immediately. A *confidence interval* at level $1 - \alpha$: $\hat{\beta}_j \pm t_{1-\alpha/2}(n - p - 1) \text{SE}(\hat{\beta}_j)$. And a *test* of $H_0 : \beta_j = 0$: reject when the observed $|t_j| = |\hat{\beta}_j|/\text{SE}(\hat{\beta}_j)$ is beyond the quantile. The *p-value* attached to the test is the probability, *computed under H_0* , of drawing a statistic at least as extreme as the one observed: $p_j = \mathbb{P}_{H_0}(|T_j| \geq |t_j|)$.

On the regulars ($n = 400$, $p = 3$, hence 396 degrees of freedom), the fitted spend equation gives $\hat{\beta}_{\text{visits}} = 24.6$ Dh per visit with $\text{SE} = 3.8$, hence $t = 6.5$ and $p = 2 \cdot 10^{-10}$; and $\hat{\beta}_{\text{dist}} = -16.6$ Dh per kilometer with $\text{SE} = 3.4$, $t = -4.9$, $p = 1.6 \cdot 10^{-6}$. Under the model, distance costs money: about seventeen dirhams of monthly spend per kilometer, with an uncertainty the interval makes precise. (These numbers hold under the model; section 5 examines whether the model itself holds.)

What the numbers do not say. The t -test is widely used and often misread. The following misconceptions are common.

- *A large p-value does not mean H_0 is true.* Failing to reject is absence of evidence, not evidence of absence: the same $p = 0.4$ arises for a null effect and for a real effect measured with too little data. One never “accepts” H_0 .
- *The p-value is not $\mathbb{P}(H_0 \mid \text{data})$.* It is computed assuming H_0 ; it quantifies the surprise of the data under that assumption, not the probability of the assumption.
- *Significance is not importance.* The t -statistic scales with $\sqrt{n}$: with enough rows, a coefficient of one centime per kilometer becomes “highly significant”. Read the estimate and its confidence interval (quantities in dirhams), not the stars next to the p-value.
- *A large R^2 is neither necessary nor sufficient.* Valid inference on coefficients does not require a large R^2 , and a large R^2 certifies nothing about the assumptions the inference rests on. The paragraph after this list unpacks both directions.

- *Coefficients are ceteris paribus, and local.* $\hat{\beta}_{\text{visits}}$ is the effect of one more visit *holding household size and distance fixed*, estimated over the observed range of the data; the fitted equation says nothing about a customer visiting forty times a month. Extrapolation is silent: the formula produces a number without warning that it has left the data.

What R^2 does and does not measure. R^2 measures one thing: the share of the variance of Y that the fitted mean absorbs, or equivalently, how small the noise σ^2 is relative to the spread of the response. The validity of inference does not depend on that share, in either direction. *Not necessary:* nothing in the construction of (6) involves R^2 ; the exact laws behind the t -statistic require (A1)–(A4), not a small σ . Under a valid model with a noisy response, a low R^2 means individual predictions are imprecise, and the intervals say so honestly by being wide; but their coverage, and the level of every test, remain exact. A small real effect in a noisy response is estimated *correctly*, just less precisely. *Not sufficient:* R^2 is computed from the fit alone and never looks at the assumptions, so it cannot certify them. The chapter’s own fits make this concrete: the raw-spend regression of section 5 reports a respectable $R^2 = 0.49$ while the funnel beneath it invalidates every standard error; and adding the redundant items column of the multicollinearity paragraph below leaves that same R^2 unchanged (0.4908 to 0.4909) while every t -test on the pair dissolves.

Reality check

This section has said “the effect of one more visit” and “distance costs money”. Under (A1)–(A3) such phrases describe an *association* in the population as it is: among the store’s customers, those who live one kilometer farther spend less on average, the other included covariates held fixed. Nothing in the model says what would happen *if* the store induced an extra visit or relocated a customer: a causal claim, which requires that no unrecorded variable drives both the covariate and the response. No assumption of this chapter provides that, and an omitted variable can manufacture a convincing “effect” out of nothing. Reading coefficients as causal effects is a decision that must be defended outside the model.

4.2 Predicting one customer

The same machinery gives the uncertainty of an individual prediction. For a new customer with covariates $\mathbf{x}_0$ (as a design row, with the leading 1), the point prediction is $\hat{y}_0 = \mathbf{x}_0^\top \hat{\beta}$, and its error has two independent sources: the estimation error of $\hat{\beta}$, with variance $\sigma^2 \mathbf{x}_0^\top (\mathbb{X}^\top \mathbb{X})^{-1} \mathbf{x}_0$ by (5), and the customer’s own noise ε_0 , with variance σ^2 . Adding them and studentizing as in (6) gives:

Property 4.1 (Prediction interval)

Under (A1)–(A4), a level- $(1 - \alpha)$ prediction interval for the response of a new, independent row $\mathbf{x}_0$ is

$$\hat{y}_0 \pm t_{1-\alpha/2}(n-p-1) \hat{\sigma} \sqrt{1 + \mathbf{x}_0^\top (\mathbb{X}^\top \mathbb{X})^{-1} \mathbf{x}_0}, \quad \hat{y}_0 = \mathbf{x}_0^\top \hat{\beta}. \quad (7)$$

The two terms under the root continue the R^2 discussion. The design term $\mathbf{x}_0^\top (\mathbb{X}^\top \mathbb{X})^{-1} \mathbf{x}_0$ shrinks with n (for a typical customer of the log-spend model it is 0.003); the constant 1 does not. For any reasonable sample size the interval’s width is essentially $2 t_{1-\alpha/2} \hat{\sigma}$, governed by the noise level, which is exactly what R^2 measures the size of. On the regulars, a customer with 10 visits, a household of 3 and 5 km gets a median spend prediction of 435 Dh with a 95% interval of [268, 707] Dh, obtained by exponentiating the log-scale interval 6.08 ± 0.49 . (Monotone transformations carry quantiles, so the interval survives the return to dirhams.) The interval is honest and wide: with $R^2 = 0.51$, half the variance of log spend is noise, and no amount of data narrows an individual prediction below it. The R^2 that was irrelevant to coefficient inference is decisive for individual prediction.

4.3 Several coefficients at once

The F test. To test a group of coefficients jointly (do the store’s three covariates explain anything at all?), one compares the residual sums of squares of the model with and without them: with q coefficients set to zero under H_0 ,

$$F = \frac{(\text{RSS}_0 - \text{RSS})/q}{\text{RSS}/(n - p - 1)} \sim F(q, n - p - 1) \quad \text{under } H_0.$$

A large F means the q extra coefficients reduce the residual sum of squares by more than chance would explain: one joint answer, where the t -tests give q marginal ones. The distinction matters in practice; the following situation, common in any database with redundant columns, shows why.

When covariates overlap: multicollinearity. Suppose the store also records the number of *items* bought per month. Regulars buy roughly the same basket per trip, so items is very nearly a multiple of visits: in our data the correlation between the two columns is $\rho = 0.988$. Adding items to the regression of the worked example above degrades the inference on visits sharply: its standard error jumps from 3.8 to 24.3 (more than six times larger), its t -statistic drops from 6.5 to 1.3 ($p = 0.20$), and items itself comes out at $p = 0.79$. Read one t -test at a time, $n = 400$ customers cannot confirm that either of the two matters. The F test of the pair, however, shows clearly that together they do: $F = 21.2$, $p = 2 \cdot 10^{-9}$.

There is no contradiction: the two tests answer different questions. Each t -test asks whether its covariate still matters *once the other is in the model*; and given items, visits adds almost nothing (nor the reverse). The F test asks whether the pair matters at all, and it plainly does. The mechanism sits in the formula the chapter has used throughout: two nearly proportional columns make $\mathbb{X}^\top \mathbb{X}$ nearly singular, and the diagonal entries of its inverse, the v_j of (6), blow up. This is the same near-singularity chapter 1 met in covariance estimation. For two standardized covariates with correlation ρ , the standard errors inflate by the factor $1/\sqrt{1-\rho^2}$, which at the observed $\rho = 0.988$ equals 6.4, matching the jump from 3.8 to 24.3 (the residual $\hat{\sigma}$ barely moves when items is added, so the factor carries through). This is *multicollinearity*, and it dictates how the output may be read: the fitted predictions are unharmed (the model knows what spend to expect, since either column carries the information), but the *attribution* of the effect between near-twin covariates is undetermined by the data, and t -tests must not be read as a ranking of variable importance. What to do about it, beyond the obvious housekeeping of dropping a redundant column, is a question the next chapter takes up.

5 Diagnostics: evidence, not proof

5.1 Residual plots and the log repair

Everything so far holds *under the model*. The reality check of section 1 demands the confrontation: the assumptions cannot be verified, but the fitted model leaves traces (the residuals), and residuals can contradict it. Two plots do most of the work.

- *Residuals against fitted values.* Under (A1)–(A3) the residuals should hover around zero in a band of constant width, at every fitted level. Curvature indicates a conditional mean that is not affine in the included covariates, the visible part of (A1); a funnel (spread growing with the fitted value) indicates heteroscedasticity (A3 violated).
- *QQ-plot.* Sorted standardized residuals against Gaussian quantiles: systematic departure from the diagonal (curved tails, an S shape) indicates non-Gaussian errors (A4 violated).

Figure 2 (top row) runs both on the regulars’ regression of section 4, and both plots contradict the model. The residual plot funnels clearly: the spread at a fitted spend of 700 Dh is more than

half again the spread at 350 Dh, so (A3) fails. With it fail the standard errors, t -statistics and p -values just computed, all of which rest on a constant σ . The QQ-plot bends away from the diagonal in the right tail: big spenders are more extreme than a Gaussian allows.

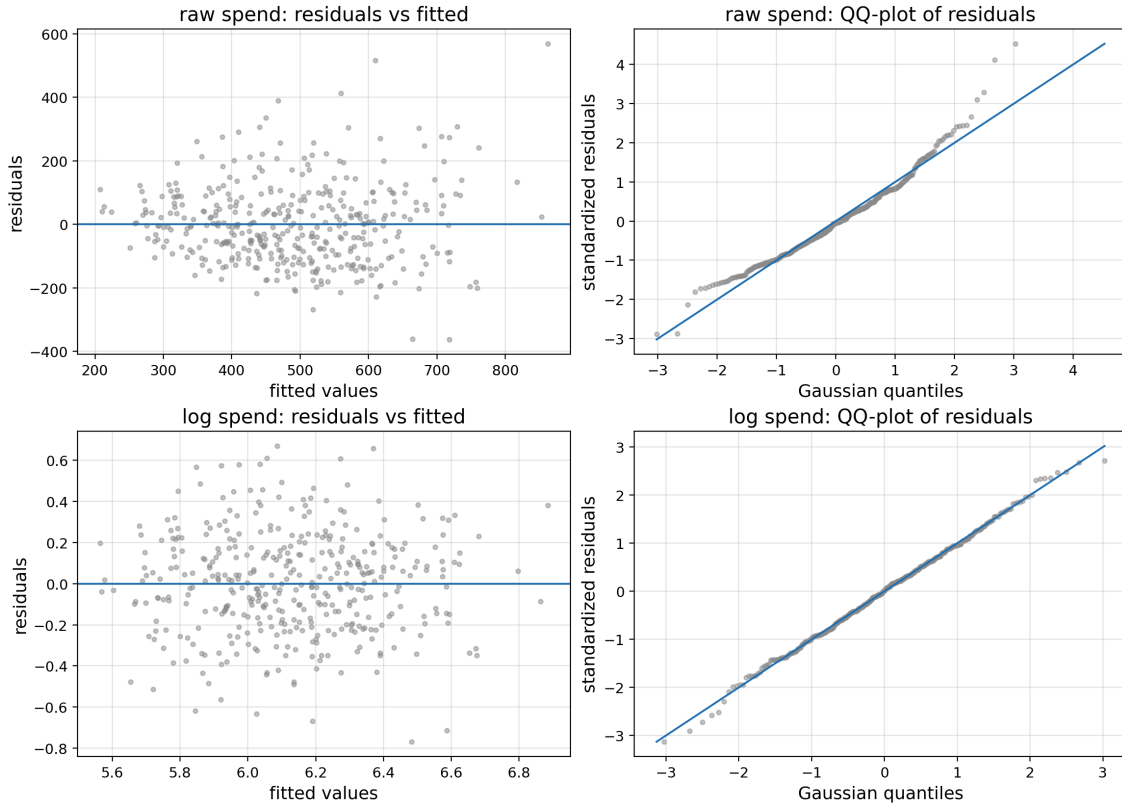


Figure 2: Diagnostics for the regulars' regression. Top: raw spend; the residuals funnel outward (A3 fails) and the QQ-plot bends in the right tail (A4 strained). Bottom: the same customers, response $\log(\text{spend})$; both plots are clean. The remedy was a modeling decision, not a test.

The remedy here is a modeling decision rather than an adjustment to the tests. Nothing in the chapter requires the modeled response to be the raw recorded quantity: for any strictly increasing transformation g , one may posit the Gaussian linear model of definition 1.1 for the *transformed* response $g(Y)$,

$$g(Y) \mid \mathbf{X} = \mathbf{x} \sim \mathcal{N}(\beta_0 + \beta^\top \mathbf{x}, \sigma^2), \quad (8)$$

and every result of sections 2–4 applies verbatim with y_i replaced by $g(y_i)$. The choice of g is part of the model specification, on the same footing as (A1)–(A4), and it is judged by the same diagnostics; the identity $g(y) = y$ (modeling raw spend) was itself such a choice, and the funnel is its refutation. Here the nature of the response points to $g = \log$: spend is positive, and effects on it plausibly act multiplicatively (an extra household member scales a basket rather than adding a fixed amount). The adopted model is therefore (8) with $g = \log$: $\log Y \mid \mathbf{X} = \mathbf{x} \sim \mathcal{N}(\beta_0 + \beta^\top \mathbf{x}, \sigma^2)$. The bottom row of figure 2 refits the same three covariates on the log response: the funnel is gone, the QQ-plot straightens, and the model's assumptions are no longer contradicted by its own residuals. The coefficients change meaning accordingly: on the log scale each unit of a covariate multiplies expected spend by a fixed factor: one more monthly visit by $e^{0.048} \approx 1.05$ (five percent), one more household member by $e^{0.17} \approx 1.19$, one more kilometer by $e^{-0.033} \approx 0.97$. These are readings at least as useful to the store as dirham amounts. One caution when mapping predictions back to dirhams: $e^{\beta_0 + \beta^\top \mathbf{x}}$ is the conditional *median* of spend, not its mean; for a log-Gaussian response the mean is larger, by the factor $e^{\sigma^2/2}$.

5.2 When the variance will not be repaired

A transformation fixed (A3) here because the effects were themselves multiplicative; nothing guarantees that a scale exists on which the mean is affine *and* the variance is constant. When heteroscedasticity remains, a standard partial remedy keeps the OLS estimates, still unbiased under (A1), and replaces the variance formula, which is what the funnel falsifies, by an estimate built from the squared residuals themselves, valid without a common σ^2 : the *heteroscedasticity-consistent* (robust, or White) standard errors, which we name as a tool and use without proof. They repair the standard errors, not the estimator: OLS is no longer the most efficient fit under unequal variances, and the exact t laws become large-sample approximations. Section 6 reaches the same numbers by a different route, one that needs no variance formula at all.

5.3 Leverage and influence

The two plots read the residuals; two further diagnostics read each row's *weight* in the fit. Both are stated here as tools, without derivation.

Definition 5.1 (Leverage and Cook's distance)

Let $\mathbf{H} = \mathbb{X}(\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top$ be the *hat matrix*, which maps observed to fitted responses ($\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$). The *leverage* of row i is the diagonal entry h_{ii} , and its *Cook's distance* is

$$D_i = \frac{r_i^2 h_{ii}}{(p+1) \hat{\sigma}^2 (1 - h_{ii})^2}, \quad r_i = y_i - \hat{y}_i. \quad (9)$$

The leverage measures how strongly observation i 's own response pulls its own fitted value. Leverages lie in $[0, 1]$ and sum to $p+1$, so their average is $(p+1)/n$; a row with a leverage several times that average sits in an unusual corner of covariate space and can tilt the fit by itself. Leverage is potential rather than harm: Cook's distance combines an unusual position with a poorly fitted response, and measures how far the whole coefficient vector moves when row i is deleted. Rows with D_i near or above 1 change the fitted equation single-handedly and deserve individual inspection: a data-entry error, an exceptional customer, or a reason to doubt the model. Figure 3 shows both tools catching exactly such a row, planted in the regulars' file: one customer whose 14 visits were entered as 41, the two digits transposed. Its leverage is 0.46 against an average of 0.005, it single-handedly halves the fitted slope, and its Cook's distance of 4.0 towers over every honest row's, none of which exceeds 0.03.

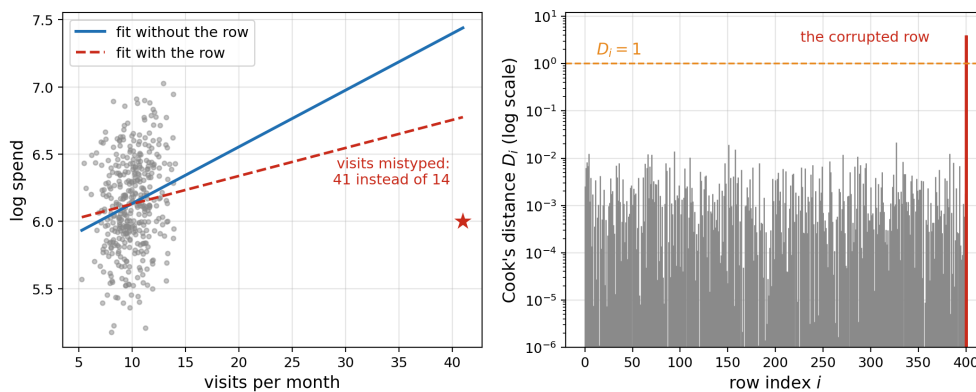


Figure 3: Influence diagnostics on the regulars' file with one corrupted row: 14 visits entered as 41, digits transposed. Left: the single red customer drags the fitted slope from 0.042 down to 0.021. Right: Cook's distance for all 401 rows (log scale); the honest rows stay below 0.03 while the corrupted one reaches $D = 4.0$, far past the $D_i = 1$ alarm line.

5.4 The assumption no plot can check

One violation is conspicuously absent from what the diagnostics can reveal: the omitted-variable failure of (A1). The residuals are orthogonal to every included column by construction; the normal equations state exactly $\mathbb{X}^\top(\mathbf{y} - \mathbb{X}\hat{\boldsymbol{\beta}}) = 0$. No plot of residuals against fitted values, against a covariate, or against anything computed from the design can therefore show that an absent variable is correlated with an included one. A fit biased by an omitted variable produces residuals exactly as healthy-looking as an unbiased one's. The most dangerous assumption is therefore the one the diagnostics cannot see: clean plots rule out curvature, heteroscedasticity and heavy tails, and say nothing at all about confounding, which is also why the causal warning of section 4 cannot be discharged by any amount of plotting. Independence (A2) shares this invisibility: correlated rows leave each residual individually ordinary, so neither plot betrays them either. The closing table of the chapter records which violations the diagnostics can and cannot see.

Diagnostics refute but do not certify: clean plots mean the model survived this confrontation, not that it is true. And the choice of response variable is part of the model: here the effective repair of a broken assumption was not a different estimator but a better description of the quantity being modeled. Linearity (A1) is now an assumption about the log scale, and a transformation moves the assumptions rather than removing them.

6 The bootstrap: inference by resampling

Every inferential product of this chapter was derived from one object: the sampling distribution of the estimator, the law (5) that says how $\hat{\boldsymbol{\beta}}$ would fluctuate across repetitions of the sampling. That law rests on (A1)–(A4), and the diagnostics have just shown that these can fail: under the funnel, the covariance formula is wrong, and with it every standard error, test and interval built on it. The repairs so far were bespoke: a transformation when one exists, a robust formula quoted without proof when it does not, each fixing one failure. This section adds the general-purpose tool: a way to *see* the sampling distribution when no formula delivers it.

Resampling the sample. Ask what the sampling distribution of $\hat{\beta}_{\text{household}}$ concretely means: if the store could recruit 400 fresh customers from the same population and we refitted, the coefficient would come out different; recruit again, different again. The histogram of those hypothetical refits *is* the sampling distribution, and its standard deviation the standard error we are after. The population cannot be redrawn from, but the sample can. The *bootstrap* [1] treats the 400 observed rows as if they were the population, and redraws samples from them. Concretely, a *bootstrap replicate* draws n complete rows $(\mathbf{x}_i, y_i)$ from the file *with replacement* (some customers drawn twice or more, others not at all) and refits the regression on the redrawn file. Repeating B times gives $\hat{\boldsymbol{\beta}}^{*1}, \dots, \hat{\boldsymbol{\beta}}^{*B}$, and the spread of the replicated coefficients estimates the sampling fluctuation:

$$\widehat{\text{SE}}_{\text{boot}}(\hat{\beta}_j) = \sqrt{\frac{1}{B-1} \sum_{b=1}^B (\hat{\beta}_j^{*b} - \bar{\beta}_j^*)^2}, \quad \bar{\beta}_j^* = \frac{1}{B} \sum_{b=1}^B \hat{\beta}_j^{*b}, \quad (10)$$

while the empirical 2.5% and 97.5% quantiles of the replicates give a *percentile confidence interval* that reads the distribution's shape directly, symmetric or not. Because entire rows are resampled (the *pairs* bootstrap), each customer's response stays attached to their own covariates: no model of the noise enters anywhere. Nothing is assumed about the variance (A3) or the distribution (A4) of the errors; what the substitution does need is (A2), independent rows drawn from one population, so that redrawing rows independently imitates how the file itself arose.

Checking the raw-spend model. Run this on the raw-spend regression, the one whose funnel falsified the printed standard errors, with $B = 2000$ refits (a few seconds of computation). The

household coefficient comes out at $\widehat{SE}_{boot} = 5.1$ against the nominal 4.65: the formula understates the fluctuation by ten percent. Visits moves from a nominal 3.8 to a bootstrap 4.0. The robust standard errors of section 5.2 report 5.2 and 4.0 for the same two coefficients: the resampling recovers them almost exactly, without anyone deriving a formula. Figure 4 (left) shows the disagreement as distributions: the histogram of the 2000 replicates against the Gaussian law the nominal formula asserts. On the adopted log-spend model, whose residuals contradict nothing, the same check passes: nominal, robust and bootstrap standard errors land within a few percent of one another (figure 4, right). The raw-scale discrepancy is ten percent, not an order of magnitude. Nothing in the printed regression output signals an error of this size, and B refits are enough to check the formula against the data.

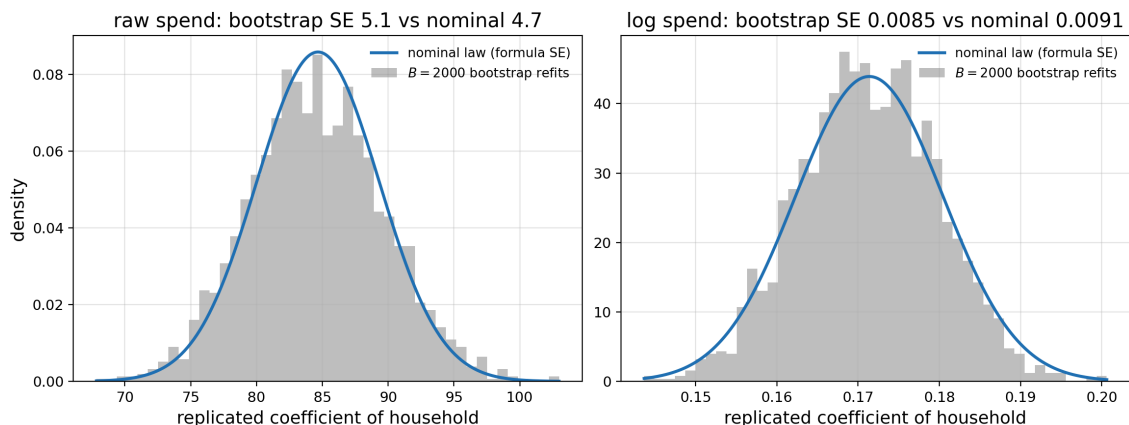


Figure 4: The pairs bootstrap on the regulars’ regression: histogram of $B = 2000$ replicated household coefficients against the Gaussian law asserted by the nominal formula. Raw spend (left): the replicates spread wider than the formula claims, the funnel’s ten-percent understatement made visible. Log spend (right): formula and replicates agree.

What the bootstrap does and does not buy. The cost is computation (B refits, trivial for least squares) plus one genuine assumption: (A2). Resampling rows independently *manufactures* independence, so when the file contains dependent rows (several customers of one household, repeated measurements), bootstrap standard errors are wrong in the same direction as the nominal ones, and the failure that no residual plot shows is one no naive resampling shows either. The other limit is subtler: the bootstrap estimates the *fluctuation* of an estimator, never its *bias*. The omitted-variable failure of (A1) resamples into a reassuringly tight histogram centered on the wrong value. Finally, the substitution of sample for population is itself an approximation that only n improves; B controls only the simulation error around it. Within those limits, the tool is remarkably general: any statistic (a coefficient, a quantile, a difference of medians) can be resampled the same way. And the device of refitting on resampled replicas returns twice in Part 2: as a sibling of chapter 5’s cross-validation, and in chapter 7, where the refits on the replicas are averaged into a predictor in their own right.

7 Multiple tests

Encouraged, the store’s data team dumps into the (log-scale) regression every column the database holds: beyond visits, household size and distance, twenty-five further recorded attributes (tenure, age, the day of the week of the first visit, and so on down the schema). Twenty-eight coefficients, one t -test each, at the customary 5%. The data being simulated, we know the truth the team does not: exactly three of the twenty-eight covariates actually influence spend.

The problem. Each single test is calibrated: if $\beta_j = 0$, the probability of flagging covariate j is 5%. But the team reads all twenty-eight at once, and this is exactly the many-tests situation of the previous chapter: among m true nulls, the family-wise error rate $1 - (1 - \alpha)^m$ compounds (46% for $m = 12$, and 72% for the $m = 25$ junk covariates of our regression). With many true nulls, at least one false rejection is therefore likely. On our data the naive screen flags four covariates: the three real ones, and the twentieth attribute, a pure noise column, at $p = 0.023$ (figure 5).

The corrections at work. The previous chapter built the two remedies; here they meet a regression. Bonferroni tests each coefficient at level α/m , a guarantee that needs no independence: with $m = 28$ the working threshold drops to $0.05/28 \approx 0.0018$, and real effects with mid-sized p-values are lost. Family-wise control costs power. Holm’s step-down procedure spends the same budget more efficiently and rejects everything Bonferroni rejects, sometimes more. For an exploratory screen like this one, the Benjamini–Hochberg procedure controls the false discovery rate instead: sort the p-values and flag up to the last one under the rising line $k\alpha/m$.

One qualification matters more here than it did in the previous chapter. The Benjamini–Hochberg guarantee covers independent tests and certain positively dependent families, and regression coefficient tests are dependent in a specific, visible way: all twenty-eight t -statistics share the same $\hat{\sigma}$, and the components of $\hat{\beta}$ correlate through the off-diagonal entries of $(\mathbb{X}^\top \mathbb{X})^{-1}$, the multicollinearity mechanism of section 4.3. Arbitrary dependence requires a modified, more conservative threshold; Bonferroni and Holm, resting on the union bound alone, are immune to the issue.

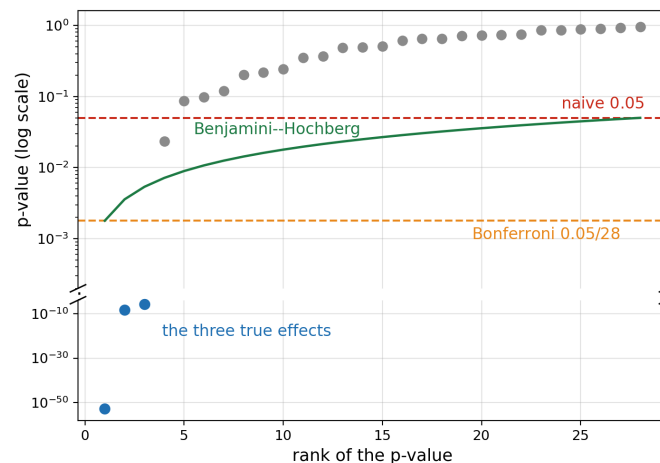


Figure 5: The twenty-eight sorted p-values of the full regression. The y-axis is broken: the top panel shows the threshold region, the bottom panel the three truly active covariates (blue) at their actual p-values, down to 10^{-53} . The naive 5% line admits a false discovery (the fourth point, $p = 0.023$); Bonferroni and the Benjamini–Hochberg line both flag exactly the three real effects.

On the store’s regression, Bonferroni and Benjamini–Hochberg each flag exactly the three real covariates. In this example the signal is strong, so the conservative and the liberal correction happen to agree; with weaker effects Bonferroni would typically reject fewer. The previous chapter’s division of labor applies unchanged (confirmatory analyses control the FWER, exploratory screens like this one the FDR), and so does its standing rule: the correction is chosen before the p-values are examined, or it controls nothing.

8 Conclusion

The table collects the assumptions and products of section 3, together with what the diagnostics of section 5 can and cannot see.

assumption	buys	breaks without it	in the plots?
(A1) affine mean	the fitted shape	curved fit	yes: curvature
(A1) no omitted confounder	unbiased coefficients	OVB	no
(A2) independent rows	error accounting	SEs too small	no
(A3) homoscedasticity	SE formula; Gauss–Markov	SEs wrong	yes: funnel
(A4) cond. Gaussianity	exact t/F , PIs	approximate levels	yes: QQ-plot

This chapter examined a conditional model (nothing assumed about the covariates, everything about the response given them) assumption by assumption: exogeneity bought unbiasedness, independence and homoscedasticity together bought the standard errors, Gaussianity bought exact tests and honest intervals from four hundred rows. Each product is lost when its assumption fails. This is why the diagnostics that caught the funnel, the log transform that repaired it, the resampling that estimated what the formulas could not, and the corrections for twenty-eight simultaneous tests are part of using the model, not accessories to it.

Part 1 ends here. The next chapter takes the same linear model and asks it the prologue’s other question: not whether its assumptions deserve belief, but whether its predictions hold up on rows it has never seen. That question is different enough to deserve a part of its own.

References

- [1] B. Efron. Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7(1):1–26, 1979.