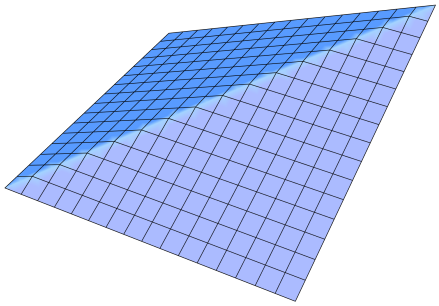


STA 2101H: Methods of Applied Statistics I

Prof. Jerry Brunner • Fall 2018 • University of Toronto
F 2:00–5:00PM in MP 202

Transcribed, LaTeXed, and edited by Rob Zimmerman

Note: These are personal lecture notes, not official course materials. They may contain errors (which by default you should attribute to me, Rob Zimmerman, rather than the course instructor). The permanent home of these — and all of my other lecture notes — is https://rob-zimmerman.github.io/coursenotes.

Contents

Lecture 1 — Friday, September 07, 2018

1. Statistical Foundations and Large Sample Inference

Statistical inference begins with a probability model and an explicit inferential target. The first topic develops estimators, tests, identifiability, convergence, and the multivariate approximations used throughout the course.

Models, Estimators, and Tests

A **statistical model** is a set of probability distributions proposed as possible descriptions of an observable random process. It separates the observable data from the unknown features that generate them. If the observable random variable is Y, a parametric model has the form

$$\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\},$$

where θ is the parameter and Θ is the parameter space. The model is an approximation: its value lies in whether it captures the aspects of the process needed for the scientific question.

A **statistic** is a function of the observed data that does not depend on an unknown parameter. An **estimator** is a statistic used to estimate a parameter, and its realized value is an **estimate**. Statistical inference uses the sampling distribution of an estimator to quantify what the data reveal about the parameter.

Example 1.1 Suppose $Y_i = 1$ when customer i prefers a new coffee blend and $Y_i = 0$ otherwise. Model $Y_1, \ldots, Y_n$ as independent Bernoulli random variables with common success probability θ. The likelihood is

$$L(\theta) = \prod_{i=1}^{n} \theta^{Y_i}(1-\theta)^{1-Y_i} = \theta^{\sum_{i=1}^n Y_i}(1-\theta)^{n-\sum_{i=1}^n Y_i}.$$

Differentiating the log likelihood shows that the maximum likelihood estimator is

$$\widehat{\theta} = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

In a sample of 100 customers, 60 prefer the new blend, so $\widehat{\theta} = 0.60$.

Consider the null hypothesis $H_0 : \theta = 0.5$ against the two sided alternative $H_1 : \theta \neq 0.5$. Under H_0,

$$Z_0 = \frac{\overline{Y} - 0.5}{\sqrt{0.5(1-0.5)/n}}$$

is approximately standard normal for large n. For the coffee data, $Z_0 = 2$, giving the two sided p-value

$$2\mathbb{P}(Z \geq 2) = 0.0455.$$

Replacing the null standard error by the estimated standard error gives the Wald statistic

$$Z_W = \frac{\overline{Y} - 0.5}{\sqrt{\overline{Y}(1-\overline{Y})/n}} = 2.041,$$

whose two sided p-value is 0.0412. These procedures are asymptotically equivalent, but they need not agree exactly in a finite sample.

Definition 1.2 For a test of H_0 against H_1, a **Type I error** occurs when H_0 is rejected even though it is true, and a **Type II error** occurs when H_0 is not rejected even though H_1 is true. The **significance level** is the maximum Type I error probability allowed by the test. The **power function** is

$$\pi(\theta) = \mathbb{P}_\theta(\text{reject } H_0).$$

A two sided level-α test may be viewed as two one sided tests, each at level $\alpha/2$. Consequently, a significant positive statistic supports the positive directional alternative, and a significant negative statistic supports the negative directional alternative. Failure to reject H_0 is not evidence that H_0 is true; it says only that the data did not provide sufficient evidence against it at the chosen level.

The approximate $100(1-\alpha)\%$ Wald confidence interval for a Bernoulli probability is

$$\overline{Y} \pm z_{1-\alpha/2}\sqrt{\frac{\overline{Y}(1-\overline{Y})}{n}}. \tag{1.1}$$

Example 1.3 For the coffee data, the 95% interval is $(0.504, 0.696)$. Its repeated sampling interpretation is that, under repeated sampling from a fixed population parameter, approximately 95% of intervals constructed by the same procedure cover that parameter. It is not a probability statement about the fixed parameter after the interval has been observed.

Confidence intervals and two sided tests are dual: at level α, $H_0 : \theta = \theta_0$ is rejected precisely when the corresponding $100(1-\alpha)\%$ confidence interval excludes θ_0, provided both procedures use the same standard error and approximation.

For planning a test of $H_0 : \theta = \theta_0$ against a two sided alternative, a useful normal approximation to the power at $\theta = \theta_1$ is

$$\pi(\theta_1) \approx \mathbb{P}_{\theta_1}\left(\frac{\overline{Y} - \theta_0}{\sqrt{\theta_0(1-\theta_0)/n}} > z_{1-\alpha/2}\right) + \mathbb{P}_{\theta_1}\left(\frac{\overline{Y} - \theta_0}{\sqrt{\theta_0(1-\theta_0)/n}} < -z_{1-\alpha/2}\right).$$

For $\theta_0 = 0.5$, $\theta_1 = 0.6$, and $\alpha = 0.05$, a sample size of 194 gives approximate power 0.8003. The following compact computation keeps the planning assumptions visible.

```
binomial_power <- function(n, theta0, theta1, alpha = 0.05) {
  cutoff <- qnorm(1 - alpha / 2)
  se0 <- sqrt(theta0 * (1 - theta0) / n)
  se1 <- sqrt(theta1 * (1 - theta1) / n)
  upper <- 1 - pnorm(theta0 + cutoff * se0, theta1, se1)
  lower <- pnorm(theta0 - cutoff * se0, theta1, se1)
  upper + lower
}
binomial_power(189, 0.5, 0.6)
```

Power calculations expose a central design principle: a large p-value may result from a small effect or from insufficient information. Sample size should therefore be chosen before data collection by specifying scientifically meaningful alternatives, rather than by asking how many observations happen to be convenient.

Identifiability and the Zipper Example

Suppose children are each given one minute to close a zipper. A first model takes Y_i to be an independent Bernoulli random variable with a common success probability. This model ignores obvious heterogeneity: some children have more dexterity or experience than others.

A richer two stage model assigns child i a personal probability θ_i, with

$$Y_i \mid \theta_i \sim \text{Bernoulli}(\theta_i) \qquad \text{and} \qquad \theta_i \sim \text{Beta}(\alpha, \beta),$$

independently across children, where $\alpha, \beta > 0$. The beta density is

$$f(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad 0 < \theta < 1. \tag{1.2}$$

The two shape parameters allow a wide range of population heterogeneity, as illustrated in Figure 1.

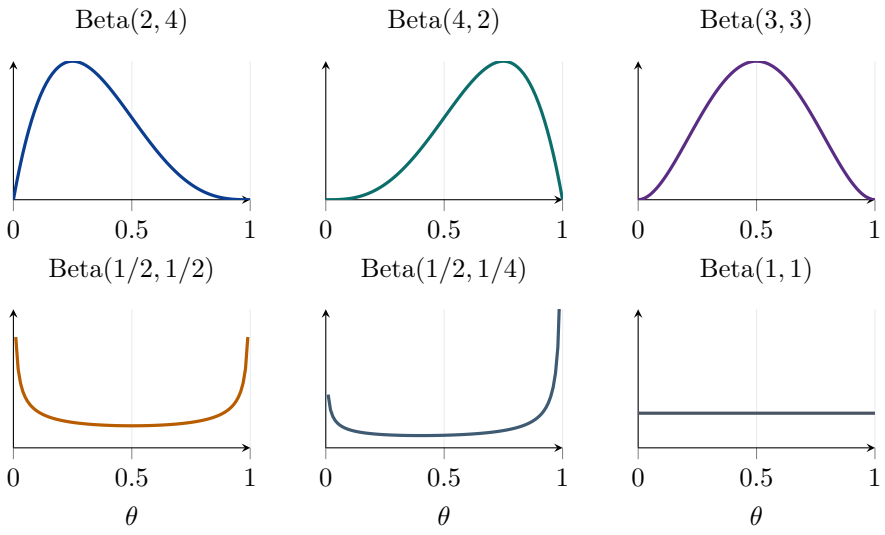


Figure 1

By the law of total probability,

$$\mathbb{P}(Y_i = 1) = \int_0^1 \mathbb{P}(Y_i = 1 \mid \theta_i) f(\theta_i)\, \mathrm{d}\theta_i = \int_0^1 \theta_i f(\theta_i)\, \mathrm{d}\theta_i = \mathbb{E}[\theta_i] = \frac{\alpha}{\alpha+\beta}.$$

Therefore, after integrating out the latent probabilities,

$$\mathbb{P}(Y_1 = y_1, \ldots, Y_n = y_n) = \prod_{i=1}^{n} \left(\frac{\alpha}{\alpha+\beta}\right)^{y_i} \left(\frac{\beta}{\alpha+\beta}\right)^{1-y_i}. \tag{1.3}$$

The observable distribution depends on α and β only through their ratio $\alpha/(\alpha+\beta)$.

Definition 1.4 A parametric model $\{\mathbb{P}_\theta : \theta \in \Theta\}$ is **identifiable** if

$$\mathbb{P}_{\theta_1} = \mathbb{P}_{\theta_2} \Longrightarrow \theta_1 = \theta_2.$$

Equivalently, distinct parameter values must imply distinct distributions of the observable data.

The zipper model is not identifiable because every pair satisfying

$$\frac{\alpha}{\alpha+\beta} = \overline{y}$$

gives the same observable distribution and the same maximized likelihood. Solving for β produces the likelihood ridge

$$\beta = \frac{1-\overline{y}}{\overline{y}}\alpha. \tag{1.4}$$

The geometry of this nonidentifiability is shown in Figure 2: every point on the ridge has the same likelihood.

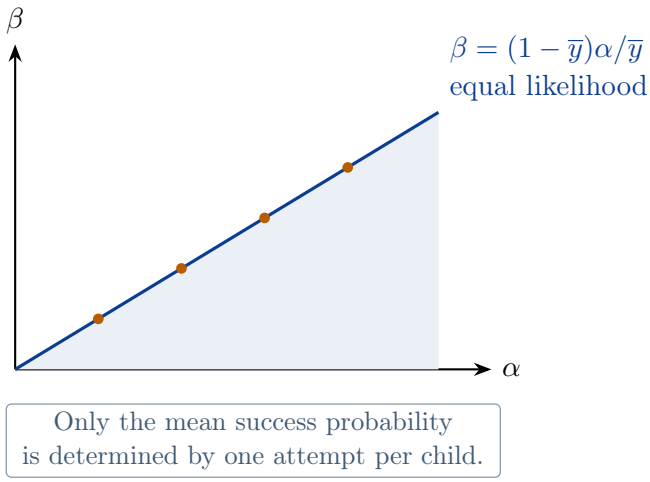


Figure 2

Theorem 1.5 If a parameter is not identifiable, no estimator can be consistent at every point of the parameter space.

Proof. Suppose $\theta_1 \neq \theta_2$ but $\mathbb{P}_{\theta_1} = \mathbb{P}_{\theta_2}$, and suppose T_n were consistent at both values. Choose disjoint open neighbourhoods B_1 and B_2 containing θ_1 and θ_2, respectively. Consistency would imply

$$\mathbb{P}_{\theta_1}(T_n \in B_1) \longrightarrow 1 \qquad \text{and} \qquad \mathbb{P}_{\theta_2}(T_n \in B_2) \longrightarrow 1.$$

Because the two parameter values induce the same sampling distribution, the first convergence also gives $\mathbb{P}_{\theta_2}(T_n \in B_1) \to 1$. This is impossible because $B_1 \cap B_2 = \varnothing$. □

The model is not useless. The identifiable quantity $\mathbb{E}[\theta_i] = \alpha/(\alpha + \beta)$ can still be estimated consistently by $\overline{Y}$. What fails is the attempt to estimate the entire population distribution of personal success probabilities from only one binary response per child. Repeating the task for each child would reveal within child information and could identify both shape parameters. Thus, when a scientifically meaningful parameter is not identifiable, the remedy is often a better research design rather than a more elaborate optimizer.

Modes of Convergence and Laws of Large Numbers

Let $X_1, X_2, \ldots$ be random variables and let X be another random variable. The principal modes of convergence are as follows.

Definition 1.6 The following statements define the three principal modes of convergence.

1. The sequence converges **almost surely**, written $X_n \xrightarrow{\text{a.s.}} X$, if
$$\mathbb{P}\big(\{\omega : X_n(\omega) \to X(\omega)\}\big) = 1.$$
2. The sequence converges **in probability**, written $X_n \xrightarrow{P} X$, if, for every $\varepsilon > 0$,
$$\mathbb{P}(|X_n - X| > \varepsilon) \longrightarrow 0.$$
3. The sequence converges **in distribution**, written $X_n \xrightarrow{d} X$, if
$$F_{X_n}(x) \longrightarrow F_X(x)$$

at every continuity point x of F_X.

Almost sure convergence implies convergence in probability, which in turn implies convergence in distribution. Neither converse holds in general. When the limit is a constant, however, convergence in distribution and convergence in probability are equivalent.

Theorem 1.7 (Laws of large numbers) Let $X_1, X_2, \ldots$ be independent random variables with a common distribution, mean μ, and finite first absolute moment. Then

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{a.s.}} \mu.$$

In particular, $\overline{X}_n \xrightarrow{P} \mu$. The latter conclusion also follows under the finite variance assumptions of the weak law of large numbers.

For Bernoulli observations, Theorem 1.7 says that the sample proportion converges to the population probability. It also justifies simulation: if $h(X)$ is integrable and $X_1, \ldots, X_m$ are simulated from a density f, then

$$\frac{1}{m}\sum_{j=1}^{m} h(X_j) \longrightarrow \mathbb{E}_f[h(X)] = \int h(x)f(x)\,\mathrm{d}x.$$

If direct simulation from the target density g is difficult but simulation from f is easy, the same idea applies when h is integrable under g and yields the importance sampling identity

$$\int h(x)g(x)\,\mathrm{d}x = \mathbb{E}_f\left[h(X)\frac{g(X)}{f(X)}\right],$$

provided $f(x) > 0$ whenever $h(x)g(x) \neq 0$.

Proposition 1.8 Let $\mathbb{1}_1, \ldots, \mathbb{1}_m$ be independent indicators of whether a simulated data set leads to rejection. If the true rejection probability is π, then

$$\widehat{\pi}_m = \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}_j$$

is unbiased for π and

$$\sqrt{m}(\widehat{\pi}_m - \pi) \xrightarrow{d} \mathcal{N}(0, \pi(1-\pi)).$$

Consequently, an approximate confidence interval for the simulation target is

$$\widehat{\pi}_m \pm z_{1-\alpha/2}\sqrt{\frac{\widehat{\pi}_m(1-\widehat{\pi}_m)}{m}}.$$

This interval measures *Monte Carlo error*: uncertainty caused by using only finitely many simulated data sets. It is distinct from the sampling error inside any one simulated data set. A simulation result without a Monte Carlo standard error can look more precise than the computation warrants.

Example 1.9 Suppose $Y_1, \ldots, Y_{100}$ are Bernoulli observations and the goal is to test $H_0 : \theta = 0.50$ when the true value is $\theta = 0.60$. For each of $m = 10{,}000$ simulated samples, form

$$Z = \frac{\sqrt{n}(\overline{Y} - \theta_0)}{\sqrt{\overline{Y}(1-\overline{Y})}}$$

and reject when $|Z| > 1.96$. The estimated power was 0.5394. A 99% Monte Carlo confidence interval is

$$0.5394 \pm 2.576\sqrt{\frac{0.5394(1-0.5394)}{10{,}000}} = [0.5266, 0.5522].$$

The calculation supports a power near 0.54, not the assertion that it is exactly 0.5394.

The computational core of Example 1.9 is short.

```
set.seed(9999)
ybar <- rbinom(10000, size = 100, prob = 0.60) / 100
z <- sqrt(100) * (ybar - 0.50) / sqrt(ybar * (1 - ybar))
power_hat <- mean(abs(z) > qnorm(0.975))
mc_se <- sqrt(power_hat * (1 - power_hat) / length(z))
```

Simulation design should mirror the complete inferential procedure. If a standard error is estimated from each generated sample, it must be re-estimated inside the simulation loop. Reusing a population standard deviation, selecting a model only once, or conditioning on successful fits can estimate the performance of a different procedure from the one that will actually be used.

Importance sampling introduces an additional diagnostic. With weights $w_j = g(X_j)/f(X_j)$, the

estimate

$$\widehat{I}_m = \frac{1}{m}\sum_{j=1}^{m} h(X_j)w_j$$

can be unbiased yet unstable when a few weights dominate. A convenient summary is the effective sample size based on the weights,

$$m_{\text{eff}} = \frac{\left(\sum_{j=1}^{m} w_j\right)^2}{\sum_{j=1}^{m} w_j^2},$$

which is well defined when the weights are not all zero and is then at most m. A very small value signals that the proposal density f does not adequately cover the important region of the target integrand.

Definition 1.10 An estimator T_n of θ is **consistent** if $T_n \xrightarrow{P} \theta$, and it is **strongly consistent** if $T_n \xrightarrow{\text{a.s.}} \theta$.

For example, the sample variance

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

is consistent for $\sigma^2 = \text{Var}(X_1)$ when the second moment is finite. Indeed,

$$\frac{n-1}{n}S_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2 \xrightarrow{P} \mathbb{E}[X_1^2] - \mu^2 = \sigma^2,$$

by the law of large numbers and continuity of $x \mapsto x^2$.

Central Limit Theorems and the Delta Method

Consistency identifies an estimator's large sample target. Central limit theorems and the delta method describe the estimator's fluctuations around that target and thereby support approximate inference.

Theorem 1.11 (Univariate central limit theorem) Let $X_1, X_2, \ldots$ be independent random variables with a common distribution, mean μ, and variance $0 < \sigma^2 < \infty$. Then

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0,1).$$

Equivalently, for large n,

$$\overline{X}_n \dot{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

The approximation symbol emphasizes that the finite sample distribution need not be normal. Theorem 1.11 concerns the shape of the standardized sampling distribution as the sample size grows.

For a random vector $\boldsymbol{X} \in \mathbb{R}^p$, convergence almost surely and in probability are defined using any norm, and convergence in distribution may be characterized by convergence of all scalar projections. This yields the multivariate form of the central limit theorem.

Theorem 1.12 (Multivariate central limit theorem) Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ be independent random vectors with a common distribution, mean $\boldsymbol{\mu}$, and finite covariance matrix $\boldsymbol{\Sigma}$. Then

$$\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{d} \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}).$$

Two closure principles make limiting results useful in applications.

Theorem 1.13 (Continuous mapping and Slutsky's theorem) Suppose $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{P} c$, where c is constant.

1. If g is continuous, then $g(X_n) \xrightarrow{d} g(X)$.
2. The sum satisfies $X_n + Y_n \xrightarrow{d} X + c$.
3. The product satisfies $X_n Y_n \xrightarrow{d} cX$.
4. If $c \neq 0$, then the ratio satisfies $X_n / Y_n \xrightarrow{d} X/c$.

For example, replacing σ by a consistent estimator S_n in Theorem 1.11 and applying Part (4) of Theorem 1.13 gives

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} \xrightarrow{d} \mathcal{N}(0, 1).$$

Theorem 1.14 (Multivariate delta method) Suppose

$$\sqrt{n}(\boldsymbol{T}_n - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}),$$

and $g : \mathbb{R}^p \to \mathbb{R}^q$ is continuously differentiable in a neighbourhood of $\boldsymbol{\theta}$. Let $\mathbf{D}g(\boldsymbol{\theta})$ denote the $q \times p$ derivative matrix. Then

$$\sqrt{n}\{g(\boldsymbol{T}_n) - g(\boldsymbol{\theta})\} \xrightarrow{d} \mathcal{N}_q\left(\mathbf{0}, \mathbf{D}g(\boldsymbol{\theta})\boldsymbol{\Sigma}\mathbf{D}g(\boldsymbol{\theta})^\top\right).$$

Proof. The first order Taylor expansion is

$$g(\boldsymbol{T}_n) = g(\boldsymbol{\theta}) + \mathbf{D}g(\boldsymbol{\theta})(\boldsymbol{T}_n - \boldsymbol{\theta}) + \boldsymbol{r}_n,$$

where $\|\boldsymbol{r}_n\|/\|\boldsymbol{T}_n - \boldsymbol{\theta}\| \to 0$ in probability. Multiplication by $\sqrt{n}$, followed by Theorem 1.13, gives the result because the linear transformation of a multivariate normal vector is multivariate normal. □

Example 1.15 Let $X_1, \ldots, X_n$ have the geometric distribution

$$\mathbb{P}(X_i = x) = \theta(1-\theta)^{x-1}, \quad x = 1, 2, \ldots,$$

where $0 < \theta < 1$. Since $\mathbb{E}[X_i] = 1/\theta$ and $\mathrm{Var}(X_i) = (1-\theta)/\theta^2$, the maximum likelihood estimator is $\widehat{\theta} = 1/\overline{X}_n$. Apply Theorem 1.14 to $g(x) = 1/x$, whose derivative at $1/\theta$ is $-\theta^2$, to obtain

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\big(0, \theta^2(1-\theta)\big).$$

Thus $\widehat{\theta}$ is approximately normal with variance $\theta^2(1-\theta)/n$.

The first order delta method becomes degenerate when the derivative vanishes. Suppose $\sqrt{n}\,\overline{X}_n \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ and $g(x) = x^2$. Because $g'(0) = 0$, the first order limit is zero. A second order expansion instead gives

$$n\overline{X}_n^2 = (\sqrt{n}\,\overline{X}_n)^2 \xrightarrow{d} \sigma^2\chi^2_1.$$

This example is a warning: a zero first derivative does not imply that the transformed estimator has no sampling variation; it means that a different rate and a higher order expansion are required.

The delta method also suggests variance stabilizing transformations. If T_n has approximate vari-

ance $v(\theta)/n$, seek g satisfying

$$\{g'(\theta)\}^2 v(\theta) = c$$

for a constant c. For a Poisson mean, $v(\theta) = \theta$, so $g(\theta) = 2\sqrt{\theta}$ stabilizes variance. For a Bernoulli proportion, $v(\theta) = \theta(1-\theta)$, and one convenient choice is $g(\theta) = 2\arcsin\sqrt{\theta}$.

Lecture 2 — Friday, September 14, 2018

Random Vectors and the Multivariate Normal Distribution

A random matrix is a matrix whose entries are random variables. For a random matrix with integrable entries, expectation is taken entrywise, and, for fixed matrices $\mathbf{A}$ and $\mathbf{B}$ of compatible dimensions,

$$\mathbb{E}[\mathbf{AXB}] = \mathbf{A}\mathbb{E}[\mathbf{X}]\mathbf{B}. \tag{2.1}$$

For random vectors $\boldsymbol{X} \in \mathbb{R}^p$ and $\boldsymbol{Y} \in \mathbb{R}^q$ with finite second moments, define

$$\begin{aligned}\operatorname{Cov}(\boldsymbol{X}) &= \mathbb{E}\left[(\boldsymbol{X} - \mathbb{E}\boldsymbol{X})(\boldsymbol{X} - \mathbb{E}\boldsymbol{X})^\mathsf{T}\right],\\ \operatorname{Cov}(\boldsymbol{X}, \boldsymbol{Y}) &= \mathbb{E}\left[(\boldsymbol{X} - \mathbb{E}\boldsymbol{X})(\boldsymbol{Y} - \mathbb{E}\boldsymbol{Y})^\mathsf{T}\right].\end{aligned}$$

Consequently,

$$\operatorname{Cov}(\mathbf{A}\boldsymbol{X}) = \mathbf{A}\operatorname{Cov}(\boldsymbol{X})\mathbf{A}^\mathsf{T} \qquad \text{and} \qquad \operatorname{Cov}(\mathbf{A}\boldsymbol{X}, \mathbf{B}\boldsymbol{Y}) = \mathbf{A}\operatorname{Cov}(\boldsymbol{X}, \boldsymbol{Y})\mathbf{B}^\mathsf{T}. \tag{2.2}$$

Definition 2.1 A random vector $\boldsymbol{X} \in \mathbb{R}^p$ has a multivariate normal distribution with mean $\boldsymbol{\mu}$ and positive semidefinite covariance matrix $\boldsymbol{\Sigma}$, written $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if every linear combination $\boldsymbol{a}^\mathsf{T}\boldsymbol{X}$ is univariate normal. The distribution may be degenerate. If $\boldsymbol{\Sigma}$ is positive definite, it has density

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right). \tag{2.3}$$

Every linear combination of a multivariate normal vector is normal. More generally, if $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

$$\mathbf{A}\boldsymbol{X} + \boldsymbol{b} \sim \mathcal{N}_q(\mathbf{A}\boldsymbol{\mu} + \boldsymbol{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\mathsf{T}).$$

For jointly multivariate normal subvectors, zero cross covariance is equivalent to independence.

This equivalence is special to the multivariate normal family.

Proposition 2.2 If $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ positive definite, then

$$(\boldsymbol{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) \sim \chi^2_p.$$

Proof. Choose a nonsingular square root $\mathbf{C}$ satisfying $\mathbf{C}\mathbf{C}^\top = \boldsymbol{\Sigma}$. Then $\boldsymbol{Z} = \mathbf{C}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I}_p)$, and

$$(\boldsymbol{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) = \boldsymbol{Z}^\top \boldsymbol{Z} = \sum_{j=1}^{p} Z_j^2 \sim \chi^2_p.$$

□

Example 2.3 Suppose

$$\boldsymbol{X} \sim \mathcal{N}_3\left(\begin{bmatrix}1\\0\\6\end{bmatrix}, \begin{bmatrix}2 & 1 & 0\\1 & 4 & 0\\0 & 0 & 2\end{bmatrix}\right),$$

and define $\boldsymbol{Y} = (X_1 + X_2, X_2 + X_3)^\top$. By (2.2),

$$\boldsymbol{Y} \sim \mathcal{N}_2\left(\begin{bmatrix}1\\6\end{bmatrix}, \begin{bmatrix}8 & 5\\5 & 6\end{bmatrix}\right).$$

The normal linear model is

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \text{and} \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n),$$

where $\mathbf{X}$ is a fixed full rank design matrix. The least squares estimator is

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{Y},$$

so the transformation rule gives

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_p\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\right). \tag{2.4}$$

The same geometry proves the familiar independence of $\overline{X}$ and S^2 in a normal sample. The sample mean is a linear projection of the data vector onto the span of the all ones vector, while the residual vector is its orthogonal projection onto the complementary subspace. Their covariance is zero, and

joint normality therefore makes them independent. It follows that

$$\frac{\sqrt{n}(\overline{X}-\mu)}{S} \sim t_{n-1}.$$

Nonlinear Hypotheses

Suppose an estimator satisfies

$$\widehat{\boldsymbol{\beta}} \dot{\sim} \mathcal{N}_p(\boldsymbol{\beta}, \widehat{\mathbf{V}}).$$

For a differentiable scalar function g, Theorem 1.14 gives

$$g(\widehat{\boldsymbol{\beta}}) \dot{\sim} \mathcal{N}\left(g(\boldsymbol{\beta}), \boldsymbol{\nabla} g(\boldsymbol{\beta})^{\mathsf{T}} \mathbf{V} \boldsymbol{\nabla} g(\boldsymbol{\beta})\right).$$

Under $H_0 : g(\boldsymbol{\beta}) = 0$, the Wald statistic is therefore

$$Z = \frac{g(\widehat{\boldsymbol{\beta}})}{\sqrt{\boldsymbol{\nabla} g(\widehat{\boldsymbol{\beta}})^{\mathsf{T}} \widehat{\mathbf{V}} \boldsymbol{\nabla} g(\widehat{\boldsymbol{\beta}})}} \dot{\sim} \mathcal{N}(0,1). \tag{2.5}$$

Example 2.4 In a regression of grade point average on verbal and mathematics scores, suppose the hypothesis of interest is

$$H_0 : \beta_1 \beta_2 = 0.$$

Here $g(\boldsymbol{\beta}) = \beta_1 \beta_2$ and

$$\boldsymbol{\nabla} g(\boldsymbol{\beta}) = \begin{bmatrix} 0 & \beta_2 & \beta_1 \end{bmatrix}^{\mathsf{T}}.$$

The fitted coefficients were

$$\widehat{\beta}_0 = 0.608075, \qquad \widehat{\beta}_1 = 0.002307, \qquad \text{and} \qquad \widehat{\beta}_2 = 0.000997.$$

Using the fitted covariance matrix in (2.5) gives $Z = 1.690$ and a two sided p-value of 0.0911. The data therefore do not provide evidence against the product restriction at level 0.05.

The calculation requires only the estimate, its covariance matrix, and the gradient:

```r
b <- coef(fit)
V <- vcov(fit)
g <- b["VERBAL"] * b["MATH"]
gradient <- c(0, b["MATH"], b["VERBAL"])
se_g <- sqrt(drop(t(gradient) %*% V %*% gradient))
z <- g / se_g
2 * pnorm(-abs(z))
```

Nonlinear restrictions require care near singular points. Under $H_0 : \beta_1\beta_2 = 0$, the gradient vanishes at $\beta_1 = \beta_2 = 0$, so the usual first order Wald approximation is unreliable there. A likelihood ratio test based directly on the restricted parameter space may behave better.

2. Regression, Likelihood, and Bayesian Methods

Regression models express conditional relationships, while likelihood and Bayesian methods provide two broad frameworks for learning unknown parameters. Scientific interpretation requires separating mathematical association from the design assumptions needed for causal conclusions.

Linear Regression and Scientific Interpretation

The multiple linear regression model is

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.6}$$

with independent errors satisfying $\mathbb{E}[\varepsilon_i] = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2$. Under normality, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. In matrix form,

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

The response variable is the outcome to be explained or predicted. The explanatory variables describe the conditions under which the response is observed.

When the explanatory variables are quantitative, β_j is the change in the conditional mean response associated with a one unit increase in x_j, while the other explanatory variables are held fixed. The phrase *held fixed* describes a conditional comparison within the model. It does not, by itself, establish a causal effect.

In a randomized experiment, assignment makes treatment independent of pretreatment characteristics, so a conditional treatment comparison can support causal interpretation. In an observational

study, unmeasured common causes may create confounding. Regression can adjust for measured covariates, but it cannot automatically remove bias due to unmeasured or poorly measured variables.

A categorical explanatory variable with k levels requires $k-1$ indicator variables when the model includes an intercept. For a drug trial with levels A, B, and placebo, define

$$x_{i1} = \mathbb{1}\{\text{drug A}\} \qquad \text{and} \qquad x_{i2} = \mathbb{1}\{\text{drug B}\}.$$

Then

$$\mathbb{E}[Y_i \mid \boldsymbol{x}_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2},$$

and the three conditional means are

$$\mu_{\text{placebo}} = \beta_0, \qquad \mu_{\text{A}} = \beta_0 + \beta_1, \qquad \text{and} \qquad \mu_{\text{B}} = \beta_0 + \beta_2.$$

Thus $H_0 : \beta_1 = \beta_2 = 0$ tests equality of all three means, while $H_0 : \beta_1 - \beta_2 = 0$ compares drugs A and B.

Adding a quantitative covariate z_i gives

$$\mathbb{E}[Y_i \mid \boldsymbol{x}_i, z_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 z_i.$$

This formulation assumes parallel regression lines across the treatment groups. Product terms such as $x_{i1} z_i$ allow the treatment comparison to vary with z_i; such terms are interactions and are developed systematically later.

For a full rank design matrix, least squares minimizes

$$\text{SSE}(\boldsymbol{\beta}) = (\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})^{\mathsf{T}}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}),$$

giving the estimator and residual sum of squares

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\boldsymbol{Y} \qquad \text{and} \qquad \text{SSE} = \|\boldsymbol{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2.$$

The mean squared error is $\text{MSE} = \text{SSE}/(n-p)$.

Many regression questions are linear hypotheses of the form

$$H_0 : \mathbf{L}\boldsymbol{\beta} = \boldsymbol{h}, \tag{2.7}$$

where $\mathbf{L}$ has r linearly independent rows. Let the reduced model impose (2.7) and let the full

model leave $\boldsymbol{\beta}$ unrestricted. The general linear test statistic is

$$F = \frac{\{\mathrm{SSE}(R) - \mathrm{SSE}(F)\}/r}{\mathrm{SSE}(F)/(n-p)} \sim F_{r,n-p} \tag{2.8}$$

under the null hypothesis and the normal model. A test of one coefficient is the square of the corresponding t-test, while the overall regression test compares the fitted model with an intercept only reduced model.

If the rows of $\mathbf{X}$ are random, the usual fixed design calculations remain valid conditionally on $\mathbf{X}$. For example,

$$\mathbb{E}[\widehat{\boldsymbol{\beta}} \mid \mathbf{X}] = \boldsymbol{\beta}$$

implies unconditional unbiasedness by the law of total expectation. Similarly, if a conditional test has size α for every admissible $\mathbf{X}$, then

$$\mathbb{P}_{H_0}(\text{reject}) = \mathbb{E}[\mathbb{P}_{H_0}(\text{reject} \mid \mathbf{X})] = \alpha.$$

Randomness of the explanatory variables is therefore not itself a problem. The essential requirement is that the conditional mean model be correct, especially that the error have conditional mean zero.

Omitted Variables and Instrumental Variables

Suppose the data-generating equation is

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad \text{and} \qquad \mathrm{Cov}(X, \varepsilon) = c. \tag{2.9}$$

The population least squares slope obtained by regressing Y on X is

$$\beta_1^\star = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} = \beta_1 + \frac{c}{\sigma_X^2}. \tag{2.10}$$

Unless $c = 0$, the least squares estimator converges to $\beta_1^\star$ rather than β_1. Increasing the sample size then makes the biased target more precise; it does not remove the bias. If $\beta_1 = 0$ but $c \neq 0$, the rejection probability of the usual test for $H_0 : \beta_1 = 0$ can approach one.

In multiple regression, omitting a relevant predictor places its contribution inside the error term. If the omitted predictor is correlated with an included predictor, the included predictor becomes correlated with the new error. This mechanism explains why coefficients may change sign or significance when covariates are added.

Example 2.5 Within both chimpanzees and humans, strength decreases with age. Chimpanzees are also stronger on average and, in the observed sample, younger on average. If species is omitted, the pooled association between age and strength can be positive even though both slopes within species are negative.

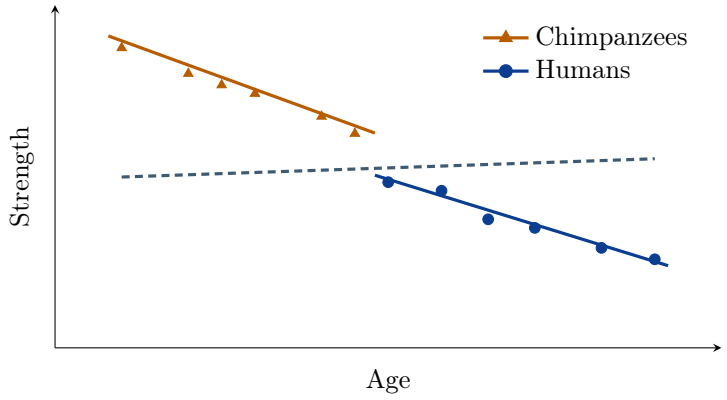


Figure 3

Figure 3 is schematic, but the geometry is general: pooled association combines within group association and between group composition. Including the grouping variable recovers the scientifically relevant conditional comparison only when the resulting model is otherwise adequate.

Under a jointly normal model for (X, Y), the observable bivariate distribution provides five moments: two means, two variances, and one covariance. Model (2.9) contains six structural quantities $(\beta_0, \beta_1, \mu_X, \sigma_X^2, \sigma_\varepsilon^2, c)$. Without an additional restriction or source of information, β_1 is not identifiable.

An **instrumental variable** W is associated with the explanatory variable X but has no direct effect on Y and is uncorrelated with the structural error. In the simple model

$$X = \alpha_0 + \alpha_1 W + \varepsilon_1,$$
$$Y = \alpha_2 + \beta X + \varepsilon_2,$$

assume $\text{Cov}(W, \varepsilon_2) = 0$ and $\text{Cov}(W, X) \neq 0$. The causal structure is shown in Figure 4.

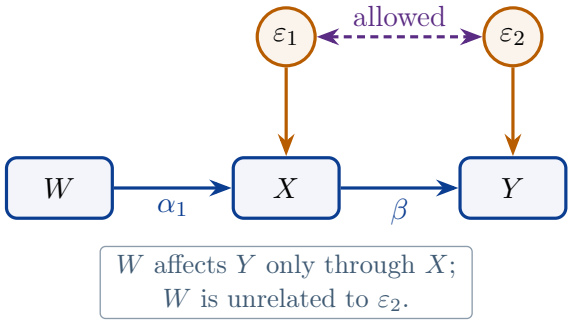


Figure 4

Taking covariances with W gives

$$\operatorname{Cov}(W, Y) = \beta \operatorname{Cov}(W, X),$$

so the structural slope is identified by

$$\beta = \frac{\operatorname{Cov}(W, Y)}{\operatorname{Cov}(W, X)}. \tag{2.11}$$

The corresponding sample covariance ratio is consistent.

Example 2.6 In a real estate application, metropolitan house price may act as an instrument for household income when studying credit card debt, provided house price affects debt only through income and is unrelated to the debt equation's unmeasured causes.

The assumptions behind (2.11) are substantive and cannot usually be proved from the observed data alone. A weak association between W and X also makes the ratio unstable. Instruments are valuable precisely because they bring information from outside the confounded relationship, but credible instruments are difficult to find.

Lecture 3 — Friday, September 21, 2018

Bayesian Inference and Decision Theory

Bayesian inference treats uncertainty about a parameter through a probability distribution. If $\pi(\theta)$ is a prior density and $p(\boldsymbol{y} \mid \theta)$ is the sampling density, Bayes' theorem gives

$$\pi(\theta \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \theta)\pi(\theta)}{\int p(\boldsymbol{y} \mid t)\pi(t)\,\mathrm{d}t} \propto p(\boldsymbol{y} \mid \theta)\pi(\theta). \tag{3.1}$$

The posterior distribution combines the information represented by the prior with the information in the likelihood. With increasing sample size, the likelihood commonly becomes more concentrated, although the actual influence of the prior depends on the model and on whether the prior puts adequate mass near the data supported region.

The frequentist parameter is fixed and probability describes repeated samples. The Bayesian parameter is assigned a distribution that expresses uncertainty conditional on the adopted model and prior. These are different probability statements, even when the resulting numerical intervals are similar.

Example 3.1 Let $Y_1, \ldots, Y_n$ be conditionally independent Bernoulli observations with success probability θ, and take $\theta \sim \mathrm{Beta}(\alpha, \beta)$. If $s = \sum_{i=1}^{n} Y_i$, then (3.1) gives

$$\begin{aligned}\pi(\theta \mid \boldsymbol{y}) &\propto \theta^{s}(1-\theta)^{n-s}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{\alpha+s-1}(1-\theta)^{\beta+n-s-1}.\end{aligned}$$

Therefore,

$$\theta \mid \boldsymbol{y} \sim \mathrm{Beta}(\alpha + s, \beta + n - s). \tag{3.2}$$

The beta family is a conjugate prior family for the Bernoulli model.

With a uniform prior, $\alpha = \beta = 1$. For the coffee data, $s = 60$ and $n = 100$, so

$$\theta \mid \boldsymbol{y} \sim \mathrm{Beta}(61, 41) \qquad \text{and} \qquad \mathbb{E}[\theta \mid \boldsymbol{y}] = \frac{61}{102} = 0.5980.$$

The posterior in Figure 5 concentrates around the observed proportion while retaining uncertainty about the population probability.

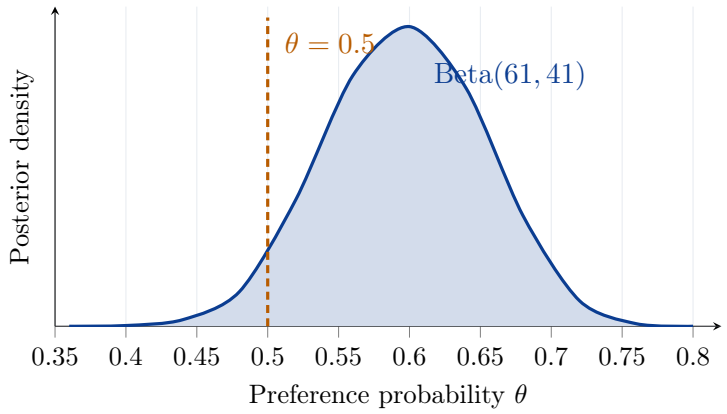


Figure 5

Bayesian estimation can be expressed as a decision problem. Let d be a decision, let θ be the unknown state, and let $L(d, \theta)$ be the loss. After observing $\boldsymbol{y}$, the posterior expected loss is

$$\rho(d \mid \boldsymbol{y}) = \mathbb{E}[L(d, \theta) \mid \boldsymbol{y}].$$

A Bayes decision minimizes this quantity. Under squared error loss $L(d, \theta) = (d - \theta)^2$, expand around the posterior mean $m = \mathbb{E}[\theta \mid \boldsymbol{y}]$:

$$\mathbb{E}[(d - \theta)^2 \mid \boldsymbol{y}] = (d - m)^2 + \mathrm{Var}(\theta \mid \boldsymbol{y}).$$

The unique minimizer is $d = m$, so the Bayes estimator under squared error loss is the posterior mean.

For testing $H_0 : \theta \leq 0.5$ against $H_1 : \theta > 0.5$ under a symmetric loss that assigns zero to a correct decision and one to an incorrect decision, choose the hypothesis with the larger posterior probability.

Example 3.2 Under the symmetric decision rule, the coffee posterior gives

$$\mathbb{P}(\theta > 0.5 \mid \boldsymbol{y}) = 0.97698,$$

so the decision favours the new blend. If a false positive costs k times as much as a false negative, reject only when the posterior odds in favour of H_1 exceed k. Decision theory makes such asymmetry explicit rather than hiding it inside a conventional significance level.

Maximum Likelihood and Likelihood Ratio Tests

For independent observations with density $f(y;\boldsymbol{\theta})$, the likelihood and log likelihood are

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i;\boldsymbol{\theta}) \qquad \text{and} \qquad \ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(y_i;\boldsymbol{\theta}).$$

The maximum likelihood estimator is any point

$$\widehat{\boldsymbol{\theta}} \in \arg\max_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}\in\Theta} \ell(\boldsymbol{\theta}).$$

Numerical software often minimizes the negative log likelihood. Starting values should be chosen in the admissible parameter space and, when practical, from method of moments estimates.

Example 3.3 Use the shape–scale parameterization

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}}, \quad x>0 \qquad \text{and} \qquad \alpha,\beta>0.$$

Then $\mathbb{E}[X]=\alpha\beta$ and $\mathrm{Var}(X)=\alpha\beta^2$, which suggest the starting values

$$\widetilde{\alpha} = \frac{\overline{X}^2}{S^2} \qquad \text{and} \qquad \widetilde{\beta} = \frac{S^2}{\overline{X}}.$$

For the sample under consideration, numerical maximization produced

$$\widehat{\alpha} = 1.80593, \qquad \widehat{\beta} = 3.80867, \qquad \text{and} \qquad -\ell(\widehat{\alpha},\widehat{\beta}) = 142.0316.$$

The essential computation is short; the optimizer's trace and repetitive console output do not contribute to the statistical argument.

```
minus_loglik <- function(par, x) {
  alpha <- par[1]
  beta <- par[2]
  -sum(dgamma(x, shape = alpha, scale = beta, log = TRUE))
}
start <- c(mean(x)^2 / var(x), var(x) / mean(x))
fit <- optim(start, minus_loglik, x = x, hessian = TRUE)
theta_hat <- fit$par
```

```
cov_hat <- solve(fit$hessian)
```

Example 3.4 At the estimate in Example 3.3, the Hessian of the negative log likelihood was

$$\mathbf{H}(\widehat{\alpha}, \widehat{\beta}) = \begin{bmatrix} 36.68932 & 13.12727 \\ 13.12727 & 6.22228 \end{bmatrix}.$$

Its eigenvalues are 41.56514 and 1.34647, both positive, so the fitted point is a strict local minimum of the negative log likelihood. The estimated covariance matrix is

$$\mathbf{H}(\widehat{\alpha}, \widehat{\beta})^{-1} = \begin{bmatrix} 0.11118 & -0.23456 \\ -0.23456 & 0.65556 \end{bmatrix}.$$

The large negative covariance reflects the elongated likelihood contours: increasing the shape while decreasing the scale can preserve a similar mean $\alpha\beta$. The nested contours in Figure 6 visualize this negative association.

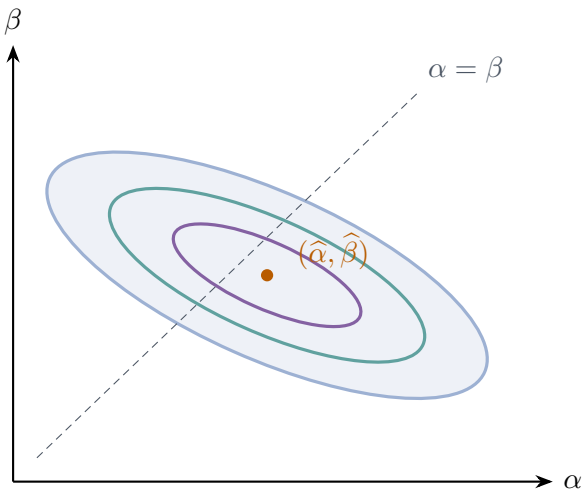


Figure 6

A candidate interior maximizer should have a zero gradient and a negative definite Hessian for ℓ, equivalently a positive definite Hessian for $-\ell$. These checks are local; they do not rule out a larger value elsewhere or a boundary maximum.

Algorithm 3.5 (Numerical likelihood audit) For a likelihood fitted by numerical optimization:

1. Encode parameter constraints directly or transform to an unconstrained scale, such as $\eta_\alpha = \log \alpha$ for a positive shape parameter.
2. Start the search from more than one admissible point.
3. Verify that the reported gradient is close to zero.
4. Inspect the Hessian eigenvalues and the condition number.
5. Compare the fitted value across starting points and, in low dimensions, inspect a profile or contour plot.
6. Check that the reported estimate is not an unintended boundary point.

A transformation can make an optimizer more stable, but the covariance must then be returned to the original scale. If $\boldsymbol{\eta} = g(\boldsymbol{\theta})$ is fitted and $\widehat{\mathbf{V}}_\eta$ is its estimated covariance, the delta method gives

$$\widehat{\mathbf{V}}_\theta = \mathbf{D}g^{-1}(\widehat{\boldsymbol{\eta}})\widehat{\mathbf{V}}_\eta \mathbf{D}g^{-1}(\widehat{\boldsymbol{\eta}})^\mathsf{T}.$$

For $\eta_\alpha = \log \alpha$, this implies $\widehat{\mathrm{Var}}(\widehat{\alpha}) \doteq \widehat{\alpha}^2 \widehat{\mathrm{Var}}(\widehat{\eta}_\alpha)$.

Suppose the null parameter space Θ_0 is contained in the full parameter space Θ. The likelihood ratio statistic is

$$G^2 = -2\log\left(\frac{\max_{\boldsymbol{\theta}\in\Theta_0} L(\boldsymbol{\theta})}{\max_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta})}\right) = 2\{\ell(\widehat{\boldsymbol{\theta}}) - \ell(\widetilde{\boldsymbol{\theta}})\}, \tag{3.3}$$

where $\widehat{\boldsymbol{\theta}}$ is unrestricted and $\widetilde{\boldsymbol{\theta}}$ is restricted. Under regularity conditions and r independent restrictions,

$$G^2 \xrightarrow{d} \chi^2_r$$

under the null hypothesis. This is **Wilks' theorem**.

The number of restrictions is a geometric dimension difference, not merely the number of equations written down; redundant constraints do not add degrees of freedom. In the three category multinomial model, the unrestricted parameter space is the interior of a triangle:

$$\theta_1 > 0, \qquad \theta_2 > 0, \qquad \text{and} \qquad \theta_1 + \theta_2 < 1.$$

The restriction $\theta_1 = 2\theta_2$ is a one dimensional line segment, as shown in Figure 7.

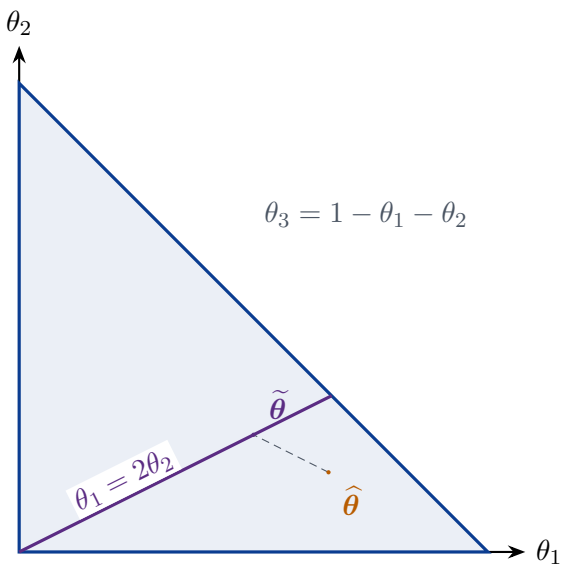


Figure 7

In Example 3.3, testing $H_0 : \alpha = \beta$ gives the restricted estimate $\widetilde{\alpha} = \widetilde{\beta} = 2.56237$ and $-\ell(\widetilde{\alpha}, \widetilde{\beta}) = 144.1704$. Hence

$$G^2 = 2(144.1704 - 142.0316) = 4.2776,$$

with one degree of freedom and $p = 0.0386$. The restriction can be imposed directly or through a reparameterization. Maximum likelihood and the likelihood ratio statistic are invariant to a one-to-one change of parameters, which makes reparameterization especially useful for nonlinear null hypotheses.

More generally, a collection of linear restrictions can be written

$$H_0 : \mathbf{L}\boldsymbol{\theta} = \boldsymbol{h}, \tag{3.4}$$

where $\mathbf{L}$ has one row per written contrast. The number of independent restrictions is $\text{rank}(\mathbf{L})$, so redundant rows must be removed before assigning degrees of freedom. For example, the statements

$$\theta_1 = \theta_2, \qquad \theta_2 = \theta_3, \qquad \text{and} \qquad \theta_1 = \theta_3$$

contain only two independent restrictions. The third equality follows from the first two.

The same geometry accommodates a smooth nonlinear null hypothesis. If $H_0 : q(\boldsymbol{\theta}) = \mathbf{0}$ and the derivative $\mathbf{D}q(\boldsymbol{\theta})$ has rank r on the null set, then the null space is locally r dimensions smaller

than the full space. A one-to-one reparameterization can make $q(\boldsymbol{\theta})$ the first block of the new parameter. For instance, testing $\mu = \sigma^2$ in a normal model can use

$$\eta_1 = \mu - \sigma^2 \qquad \text{and} \qquad \eta_2 = \mu,$$

so that the null hypothesis becomes $\eta_1 = 0$. Regularity matters: if a parameter lies on a boundary or the derivative loses rank, the usual chi-squared limit need not hold.

Information, Wald Tests, and Score Tests

Let the log likelihood contribution from one observation be $\ell_1(\boldsymbol{\theta})$. The score and Fisher information are

$$\boldsymbol{U}(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \tag{3.5}$$

$$\mathbf{I}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\left[\boldsymbol{U}_1(\boldsymbol{\theta})\boldsymbol{U}_1(\boldsymbol{\theta})^\mathsf{T}\right] = -\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial^2 \ell_1(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\mathsf{T}}\right]. \tag{3.6}$$

Under regularity conditions, the maximum likelihood estimator is consistent and

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}_p(\mathbf{0}, \mathbf{I}(\boldsymbol{\theta})^{-1}). \tag{3.7}$$

An estimated covariance matrix is obtained by inverting the observed information,

$$\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\theta}}) = \left\{-\frac{\partial^2 \ell(\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\mathsf{T}}\right\}^{-1}.$$

For the linear restriction $H_0 : \mathbf{L}\boldsymbol{\theta} = \boldsymbol{h}$, the Wald statistic is

$$W = (\mathbf{L}\widehat{\boldsymbol{\theta}} - \boldsymbol{h})^\mathsf{T}\{\mathbf{L}\widehat{\mathbf{V}}\mathbf{L}^\mathsf{T}\}^{-1}(\mathbf{L}\widehat{\boldsymbol{\theta}} - \boldsymbol{h}) \xrightarrow{d} \chi^2_r. \tag{3.8}$$

The score test evaluates the gradient of the unrestricted likelihood at the restricted estimate. The likelihood ratio test compares the restricted and unrestricted likelihood values. Thus:

1. The score test requires fitting only the restricted model.
2. The Wald test requires fitting only the unrestricted model.
3. The likelihood ratio test requires both fits.

All three are asymptotically equivalent under standard regularity conditions. Their finite sample behaviour can differ, and the likelihood ratio statistic is often less sensitive than the Wald statistic

to nonlinear parameterization.

Lecture 4 — Friday, September 28, 2018

For the gamma restriction $H_0 : \alpha = \beta$, the Wald calculation gave $W = 3.2952$ and $p = 0.0695$, compared with the likelihood ratio value $p = 0.0386$. The discrepancy is not a contradiction: both conclusions rely on large sample approximations and the Wald statistic measures distance in a local quadratic geometry.

Example 4.1 (Chick weights) In an independent groups analysis of chick weights, a heteroscedastic Wald test of equality among four group means gave

$$W = 14.8909 \qquad \text{and} \qquad p = 0.00191.$$

For comparison, the ordinary equal variance analysis of variance gave $p = 0.00686$, and the Kruskal–Wallis test gave $p = 0.0142$. The procedures answer closely related but not identical questions under different assumptions. Agreement strengthens the qualitative conclusion, but it does not make the assumptions interchangeable.

Bayesian Computation

Conjugate models such as Example 3.1 are analytically convenient, but most posterior expectations have the form

$$\mathbb{E}[h(\theta) \mid \boldsymbol{y}] = \int h(\theta)\pi(\theta \mid \boldsymbol{y})\,\mathrm{d}\theta$$

and cannot be evaluated in closed form. If $\theta^{(1)}, \ldots, \theta^{(M)}$ are draws from the posterior, then

$$\frac{1}{M}\sum_{m=1}^{M} h(\theta^{(m)})$$

approximates the posterior expectation.

A Markov chain Monte Carlo method constructs a dependent sequence whose stationary distribution is the posterior. The dependence does not invalidate Monte Carlo averaging, but it reduces the effective information relative to the same number of independent draws.

Algorithm 4.2 (Gibbs sampler) Partition the parameter as $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. Starting from an admissible state, repeat the following sweep:

1. Draw θ_1 from its full conditional distribution given the current values of $\theta_2, \ldots, \theta_k$ and the data.
2. Draw θ_2 from its full conditional distribution using the newly updated θ_1 and the current remaining coordinates.
3. Continue in this manner through θ_k.

After an initial warmup period, retain draws for posterior summaries.

The warmup period removes the portion most affected by the arbitrary starting state. Trace plots reveal poor mixing, trends, or isolated excursions; autocorrelation plots quantify serial dependence. Several dispersed chains are preferable to a single chain because agreement across them is evidence, though not proof, of convergence.

Example 4.3 For the coffee model, a compact probabilistic program is sufficient:

```
model {
  for (i in 1:n) {
    y[i] ~ dbern(theta)
  }
  theta ~ dbeta(1, 1)
}
```

The simulation estimate of the posterior mean and variance should agree with the exact Beta$(61, 41)$ values. This comparison is a valuable check of both the code and the sampler.

Example 4.4 As a nonconjugate example, suppose

$$Y_i \mid \mu, \tau \sim \mathcal{N}(\mu, \tau^{-1}), \qquad \mu \sim \mathcal{N}(0, 100), \qquad \text{and} \qquad \tau \sim \text{Lognormal}(0, 1).$$

For a sample of size 50 generated near mean 100 and standard deviation 15, posterior simu-

lation produced the summaries

$$\mathbb{E}[\mu \mid \boldsymbol{y}] \approx 100.63 \qquad \text{and} \qquad 95\% \text{ credible interval } (97.07, 103.91),$$

and

$$\mathbb{E}[\sigma \mid \boldsymbol{y}] \approx 13.52, \qquad 95\% \text{ credible interval } (11.35, 16.36), \qquad \text{and} \qquad \sigma = \tau^{-1/2}.$$

Prior distributions are part of the model, not harmless defaults. In a larger simulated sample, incompatible choices such as an exponential prior for a mean known by the data to be near 100 produced a pathological chain with extreme excursions and distorted summaries. A large sample cannot rescue an algorithm that fails to explore the intended posterior. Prior predictive checks, multiple chain diagnostics, and sensitivity analysis are therefore essential parts of Bayesian computation.

Lecture 5 — Friday, October 05, 2018

3. Generalized Linear Models

Generalized linear models extend regression to responses whose distributions and mean–variance relationships are not well represented by a normal model. Binary, count, and unordered categorical outcomes illustrate the common structure and its distinct interpretations.

Binary Logistic Regression

A generalized linear model has three ingredients:

1. The response has a distribution from the exponential family.
2. The model has a linear predictor $\eta_i = \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta}$.
3. A link function g relates the conditional mean to the linear predictor according to

$$g(\mathbb{E}[Y_i \mid \boldsymbol{x}_i]) = \eta_i.$$

Ordinary normal linear regression uses the identity link. Binary and count outcomes require links that respect the range of their conditional means.

For a binary response, write

$$\pi_i = \mathbb{P}(Y_i = 1 \mid \boldsymbol{x}_i) = \mathbb{E}[Y_i \mid \boldsymbol{x}_i].$$

The logistic regression model is

$$\operatorname{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}. \tag{5.1}$$

Solving for the probability gives

$$\pi_i = \frac{\exp\left(\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}\right)}{1+\exp\left(\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}\right)}. \tag{5.2}$$

The inverse logit maps the entire real line into $(0, 1)$, so fitted probabilities are automatically admissible. Other links are possible. The probit model, for example, takes $\pi_i = \Phi(\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta})$.

Exponentiating (5.1) gives

$$\frac{\pi_i}{1-\pi_i} = \exp(\beta_0)\prod_{j=1}^{p-1}\exp(\beta_j x_{ij}).$$

Thus $\exp(\beta_j)$ is an odds ratio: increasing x_j by one unit, while holding the other explanatory variables fixed, multiplies the conditional odds of $Y = 1$ by $\exp(\beta_j)$. The coefficient itself is a change in log odds, not a change in probability. Because (5.2) is nonlinear, the corresponding probability change depends on the starting probability and the other covariates.

Example 5.1 Let $Y = 1$ indicate survival, let x measure disease severity, and compare chemotherapy, radiation, and the combination of both. With chemotherapy as the reference group, define indicators d_1 for radiation and d_2 for combined therapy. A parallel curves model is

$$\operatorname{logit}\{\mathbb{P}(Y = 1 \mid d_1, d_2, x)\} = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 x.$$

At any fixed severity, the survival odds under combined therapy are $\exp(\beta_2)$ times the chemotherapy odds. The common coefficient β_3 assumes that severity has the same log odds effect for all therapies.

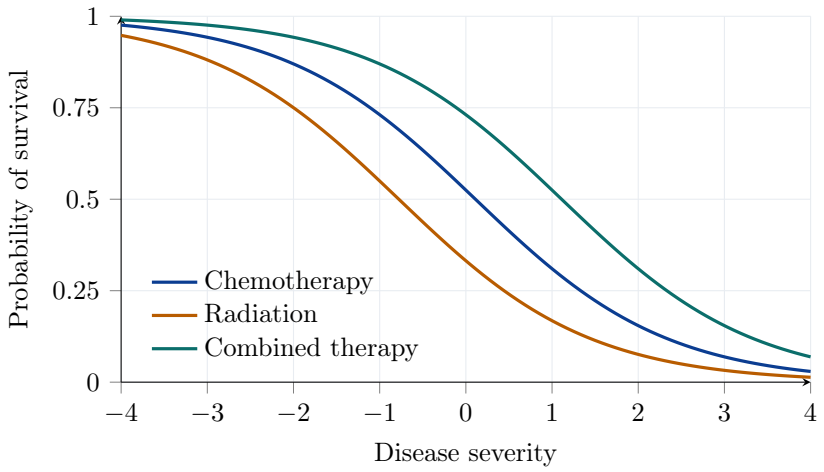


Figure 8

Figure 8 translates the parallel log odds model into probability curves. Equal slopes on the logit scale do not produce parallel curves on the probability scale. Adding products $d_1 x$ and $d_2 x$ would permit therapy by severity interaction.

Conditional independence gives the binary likelihood

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{1-y_i}, \tag{5.3}$$

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \log \pi_i + (1-y_i) \log(1-\pi_i) \right].$$

There is generally no closed form maximum likelihood estimator. Iteratively reweighted least squares is one standard numerical method. Likelihood ratio and Wald tests use the large sample theory developed earlier; a likelihood ratio test compares deviances, because the deviance difference equals twice the fitted log likelihood difference.

For a fitted covariate vector $\boldsymbol{x}_\star$, let $\widehat{\eta}_\star = \boldsymbol{x}_\star^{\mathsf{T}} \widehat{\boldsymbol{\beta}}$. Its estimated standard error is

$$\operatorname{se}(\widehat{\eta}_\star) = \sqrt{\boldsymbol{x}_\star^{\mathsf{T}} \widehat{\mathbf{V}} \boldsymbol{x}_\star}.$$

A confidence interval should first be formed on the unbounded link scale and then transformed by the inverse logit. Applying a normal interval directly to the probability estimate may produce

endpoints outside $[0, 1]$.

Logistic Regression in Practice

Example 5.2 (Student performance and logistic regression) The mathematics performance data contain 394 students and a binary response indicating whether each student passed. A model using high school grade point average and high school English score gave

$$\widehat{\text{logit}}(\pi) = -14.69568 + 0.22982\,\text{GPA} - 0.04020\,\text{English}.$$

Holding English score fixed, a one point increase in grade point average multiplies the passing odds by

$$e^{0.22982} = 1.258.$$

The Wald test for English score gave $W = 5.53$ and $p = 0.0187$, while the likelihood ratio test of both predictors gave $G^2 = 92.97$.

For a student with grade point average 80 and English score 75, the fitted probability was

$$\widehat{\pi} = 0.66265 \qquad \text{and} \qquad \text{se}(\widehat{\pi}) = 0.02859,$$

where the probability scale standard error was obtained by the delta method.

Course stream has three levels: catch-up, elite, and mainstream. The observed pass counts were

Course stream	Passed	Total	Observed proportion
Catch-up	8	35	0.229
Elite	24	31	0.774
Mainstream	204	328	0.622

The catch up to mainstream sample odds ratio is 0.180. Such an unadjusted comparison mixes stream differences with differences in prior preparation.

A model that includes English score, grade point average, and course stream is

$$\begin{aligned}\widehat{\text{logit}}(\pi) = &-14.18265 - 0.03534\,\text{English} + 0.21939\,\text{GPA} \\ &- 1.29137\,\mathbb{1}\{\text{catch-up}\} + 0.75847\,\mathbb{1}\{\text{elite}\},\end{aligned}$$

with mainstream as the reference category. The course stream likelihood ratio test gave

$p = 0.00156$, and the corresponding Wald test gave $p = 0.00347$. Controlling for the other variables reverses the apparent direction of the English score association. The reversal is evidence of confounding among the preparation measures, not a reason to interpret the negative coefficient causally.

The adjusted odds ratio comparing the elite and catch-up streams is

$$\exp(0.75847 - (-1.29137)) = 7.767.$$

Predictions far outside the observed covariate range can nevertheless be absurd. For example, extrapolating to an English score of -40 uses the algebraic model correctly but has no empirical support. A fitted formula does not enlarge the population represented by the data.

The reusable fitting workflow is compact:

```
fit <- glm(passed ~ hsengl + hsgpa + course,
           family = binomial, data = math)
drop1(fit, test = "LRT")

x_star <- data.frame(hsengl = 75, hsgpa = 80,
                     course = "Mainstream")
predict(fit, x_star, type = "link", se.fit = TRUE)
predict(fit, x_star, type = "response")
```

Model formulas, tests, and prediction commands matter; pages of printed coefficient correlations and residual summaries do not need to be reproduced.

Poisson Regression

Given their covariates, let Y_i be conditionally independent counts with

$$Y_i \mid \boldsymbol{x}_i \sim \text{Poisson}(\lambda_i) \qquad \text{and} \qquad \mathbb{E}[Y_i \mid \boldsymbol{x}_i] = \text{Var}(Y_i \mid \boldsymbol{x}_i) = \lambda_i.$$

The log link model is

$$\log \lambda_i = \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta} \qquad \text{and} \qquad \lambda_i = \exp\left(\boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta}\right). \tag{5.4}$$

Consequently, $\exp(\beta_j)$ is a multiplicative effect on the conditional mean count. If x_j increases by one while the other variables remain fixed, the expected count is multiplied by $\exp(\beta_j)$.

When observations have unequal exposure t_i, model the rate through

$$\log \lambda_i = \log t_i + \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta}.$$

The known term $\log t_i$ is an offset, not an estimated coefficient. It converts count comparisons into rate comparisons.

Example 5.3 (Training programs and support calls) In a training program study, 300 workers were assigned among programs A, B, and C, and the response was the number of information technology support calls. The raw mean counts were 4.07, 3.47, and 4.05. After adjustment for test score and experience, the fitted model was

$$\begin{aligned}\widehat{\log \lambda} = 1.9927 &- 0.009205\,\text{Score} - 0.028014\,\text{Experience} \\ &- 0.170519\,\mathbb{1}\{\text{B}\} - 0.007833\,\mathbb{1}\{\text{C}\},\end{aligned}$$

with program A as the reference. Holding the other variables fixed, each additional unit of experience multiplies the expected call count by

$$e^{-0.028014} = 0.9724,$$

a decrease of approximately 2.76%. The program likelihood ratio test gave $G^2 = 6.8767$ with $p = 0.0321$, and the Wald test gave $W = 6.7335$ with $p = 0.0345$.

The equality of the Poisson mean and variance is a model restriction. Extra Poisson variation can arise from unobserved heterogeneity or dependence. It tends to make model based standard errors too small. Residual diagnostics, a dispersion assessment, a negative binomial model, or a random effects model may then be appropriate.

Multinomial Logit Models

Let Y take one of K unordered categories. Choose category K as the reference and define $K - 1$ baseline category logits

$$\log\left(\frac{\pi_j(\boldsymbol{x})}{\pi_K(\boldsymbol{x})}\right) = \boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta}_j, \quad j = 1, \ldots, K-1. \tag{5.5}$$

Writing $\eta_j = \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}_j$, the probabilities are

$$\pi_j(\boldsymbol{x}) = \frac{e^{\eta_j}}{1 + \sum_{h=1}^{K-1} e^{\eta_h}}, \quad j = 1, \ldots, K-1, \tag{5.6}$$
$$\pi_K(\boldsymbol{x}) = \frac{1}{1 + \sum_{h=1}^{K-1} e^{\eta_h}} = 1 - \sum_{j=1}^{K-1} \pi_j(\boldsymbol{x}).$$

These expressions are positive and sum to one. For $K = 2$, the construction reduces exactly to ordinary binary logistic regression.

For three categories, suppose the logits relative to category 3 are

$$L_1 = \log\left(\frac{\pi_1}{\pi_3}\right) \qquad \text{and} \qquad L_2 = \log\left(\frac{\pi_2}{\pi_3}\right).$$

Then

$$\pi_1 = \frac{e^{L_1}}{1 + e^{L_1} + e^{L_2}}, \qquad \pi_2 = \frac{e^{L_2}}{1 + e^{L_1} + e^{L_2}}, \qquad \text{and} \qquad \pi_3 = \frac{1}{1 + e^{L_1} + e^{L_2}}.$$

A positive coefficient in logit j means that the explanatory variable raises the odds of category j relative to the reference category, holding the other variables fixed. It need not increase the marginal probability of category j over every covariate range because all category probabilities share the same denominator.

Example 5.4 (Student performance and multinomial regression) The mathematics outcome was also recorded in three categories: failed, gone, and passed, with counts 61, 97, and 236. Using failed as the reference and grade point average as the only predictor gave

$$\log\left(\frac{\pi_{\text{gone}}}{\pi_{\text{failed}}}\right) = 1.904 - 0.0188\,\text{GPA},$$
$$\log\left(\frac{\pi_{\text{passed}}}{\pi_{\text{failed}}}\right) = -13.393 + 0.18644\,\text{GPA}.$$

At a grade point average of 90, the fitted probabilities were

$$\widehat{\pi}_{\text{gone}} = 0.03884, \qquad \widehat{\pi}_{\text{passed}} = 0.92971, \qquad \text{and} \qquad \widehat{\pi}_{\text{failed}} = 0.03146.$$

In the fuller model, a likelihood ratio test for course stream gave $p = 0.0184$. Jointly removing English and calculus scores gave $p = 0.0762$, so a reduced model retained grade point average and course stream. Model simplification should be driven by the scientific

question and prespecified comparisons when possible; repeated data dependent deletion can make nominal p-values optimistic.

The essential software pattern is the same as in binary regression:

```
library(nnet)
fit_multi <- multinom(outcome ~ hsgpa + course,
                      data = math, trace = FALSE)
eta <- predict(fit_multi, newdata = x_star, type = "probs")

fit_reduced <- update(fit_multi, . ~ . - course)
anova(fit_reduced, fit_multi, test = "Chisq")
```

The response probabilities, coefficient contrasts, and likelihood comparisons carry the statistical content. Raw optimization traces and repeated printouts do not.

Lecture 6 — Friday, October 12, 2018

4. Prediction, Missing Data, and Machine Learning

Prediction quality depends on the population target of a fitted procedure and on whether its validation resembles future use. Measurement error, missing values, and data dependent model selection can change that target even when the fitting computation is formally correct.

Least Squares Targets Under Misspecification

Consider a regression with an intercept. It is often convenient to centre every variable at its sample mean:

$$Y_i - \overline{Y} = \beta_1(X_{i1} - \overline{X}_1) + \cdots + \beta_p(X_{ip} - \overline{X}_p) + \varepsilon_i.$$

Centring changes the intercept but leaves the fitted slopes and residuals unchanged. Let

$$\widehat{\boldsymbol{\Sigma}}_{XX} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^{\mathsf{T}}$$

and

$$\widehat{\boldsymbol{\Sigma}}_{XY} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \overline{\boldsymbol{X}})(Y_i - \overline{Y}).$$

The least squares slope vector is

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\Sigma}}_{XX}^{-1}\widehat{\boldsymbol{\Sigma}}_{XY}. \tag{6.1}$$

If the relevant second moments exist and $\boldsymbol{\Sigma}_{XX}$ is nonsingular, the law of large numbers gives

$$\widehat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}^{\star} = \boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}. \tag{6.2}$$

This probability limit exists even when the conditional mean is not linear. It is the coefficient vector of the best population linear predictor, not automatically a structural or causal parameter.

Under the correctly specified model

$$Y = \beta_0 + \boldsymbol{X}^{\mathsf{T}}\boldsymbol{\beta} + \varepsilon \qquad \text{and} \qquad \mathbb{E}[\varepsilon \mid \boldsymbol{X}] = 0,$$

we have

$$\boldsymbol{\Sigma}_{XY} = \mathrm{Cov}(\boldsymbol{X}, Y) = \boldsymbol{\Sigma}_{XX}\boldsymbol{\beta},$$

so (6.2) reduces to $\boldsymbol{\beta}^{\star} = \boldsymbol{\beta}$. Under misspecification, the same calculation identifies the parameter that least squares actually estimates and therefore makes the bias mechanism explicit.

Measurement Error in Explanatory Variables

Suppose the structural model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \tag{6.3}$$

but the latent explanatory variables are observed with additive error:

$$W_1 = X_1 + e_1 \qquad \text{and} \qquad W_2 = X_2 + e_2.$$

Assume that $\mathbb{E}[\varepsilon \mid X_1, X_2] = 0$ and that the measurement errors have mean zero, are mutually independent, and are independent of (X_1, X_2, ε). Write

$$\phi_{jk} = \mathrm{Cov}(X_j, X_k) \qquad \text{and} \qquad \omega_j = \mathrm{Var}(e_j).$$

The reliability of W_j is

$$\mathrm{Rel}(W_j) = \frac{\mathrm{Var}(X_j)}{\mathrm{Var}(W_j)} = \frac{\phi_{jj}}{\phi_{jj} + \omega_j}. \tag{6.4}$$

If Y is naively regressed on (W_1, W_2), then

$$\boldsymbol{\Sigma}_{WW} = \begin{bmatrix} \phi_{11} + \omega_1 & \phi_{12} \\ \phi_{12} & \phi_{22} + \omega_2 \end{bmatrix}$$

and

$$\boldsymbol{\Sigma}_{WY} = \begin{bmatrix} \beta_1\phi_{11} + \beta_2\phi_{12} \\ \beta_1\phi_{12} + \beta_2\phi_{22} \end{bmatrix}.$$

Applying (6.2) gives the naive probability limit. Under the null hypothesis $\beta_2 = 0$, its second component simplifies to

$$\beta_2^\star = \frac{\beta_1\omega_1\phi_{12}}{(\phi_{11} + \omega_1)(\phi_{22} + \omega_2) - \phi_{12}^2}. \tag{6.5}$$

This target is nonzero whenever all of the following hold:

1. The genuinely active predictor has $\beta_1 \neq 0$.
2. The predictor X_1 is measured with error, so $\omega_1 > 0$.
3. The latent predictors are correlated, so $\phi_{12} \neq 0$.

Measurement error in the variable being controlled for leaves residual confounding, which is then attributed to the correlated predictor W_2. The path diagram in Figure 9 makes the two sources of measurement noise and the latent predictor correlation explicit.

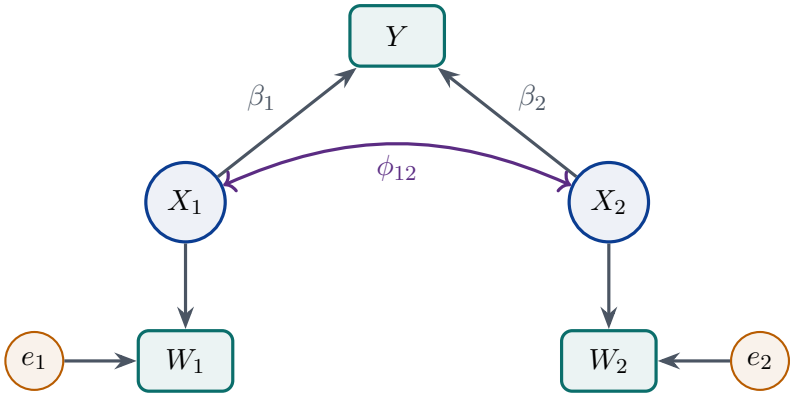


Figure 9

Several special cases clarify (6.5). If $\phi_{12} = 0$, the numerator vanishes. If W_1 is measured without error, the numerator also vanishes. Error in W_2 enlarges the denominator and can attenuate the

spurious coefficient, but it does not repair the underlying misspecification. When X_1 is randomly assigned, it is independent of pretreatment X_2, so the correlation mechanism disappears.

Example 6.1 A simulation study crossed sample size, predictor correlation, signal strength, two reliabilities, and four distributional families. It included 7,500 conditions and 10,000 generated data sets per condition, for 75 million fits. Every data set satisfied $\beta_2 = 0$, and the naive test was conducted at level 0.05.

The following representative results use normally distributed variables and reliability 0.90 for both measurements. Entries are empirical Type I error probabilities when X_1 explains 50% of the response variance.

	Corr(X_1, X_2)				
n	0.00	0.20	0.40	0.60	0.80
50	0.0460	0.0520	0.0963	0.1106	0.1633
100	0.0535	0.0569	0.1461	0.1857	0.2837
250	0.0483	0.0625	0.3068	0.3731	0.5864
500	0.0515	0.0780	0.5323	0.6488	0.8837
1000	0.0481	0.1185	0.8273	0.9088	0.9907

Larger samples make the test more certain about the wrong nonzero target. This is why Type I error can approach one rather than 0.05.

The signal strength controls how quickly this breakdown becomes visible. The same experiment with X_1 explaining 25% of the response variance gave the following results.

	Corr(X_1, X_2)				
n	0.00	0.20	0.40	0.60	0.80
50	0.0476	0.0505	0.0636	0.0715	0.0913
100	0.0504	0.0521	0.0834	0.0940	0.1294
250	0.0467	0.0533	0.1402	0.1624	0.2544
500	0.0468	0.0595	0.2300	0.2892	0.4649
1000	0.0505	0.0734	0.4094	0.5057	0.7431

Table 1: Empirical Type I error with a weak signal.

When X_1 explained 75% of the response variance, even moderate correlation caused severe

inflation.

	$\text{Corr}(X_1, X_2)$				
n	0.00	0.20	0.40	0.60	0.80
50	0.0485	0.0579	0.1727	0.2089	0.3442
100	0.0541	0.0679	0.3101	0.3785	0.6031
250	0.0479	0.0856	0.6450	0.7523	0.9434
500	0.0445	0.1323	0.9109	0.9635	0.9992
1000	0.0522	0.2179	0.9959	0.9998	1.0000

Table 2: Empirical Type I error with a strong signal.

The three tables isolate a systematic pattern:

1. Increasing the latent predictor correlation increases the magnitude of the spurious target in (6.5).
2. Increasing the relationship between X_1 and Y increases $|\beta_1|$ and therefore increases the spurious target.
3. Decreasing the reliability of W_1 increases ω_1 and leaves more residual confounding.
4. Decreasing the reliability of W_2 can attenuate the spurious coefficient by increasing the denominator.
5. Increasing n reduces the standard error around a nonzero probability limit, so the false rejection probability tends toward one.

The fifth pattern is the most counterintuitive. More data improve precision around the estimand chosen by the fitted model; they do not force that estimand to equal the intended structural coefficient. The divergence across predictor correlations is plotted in Figure 10.

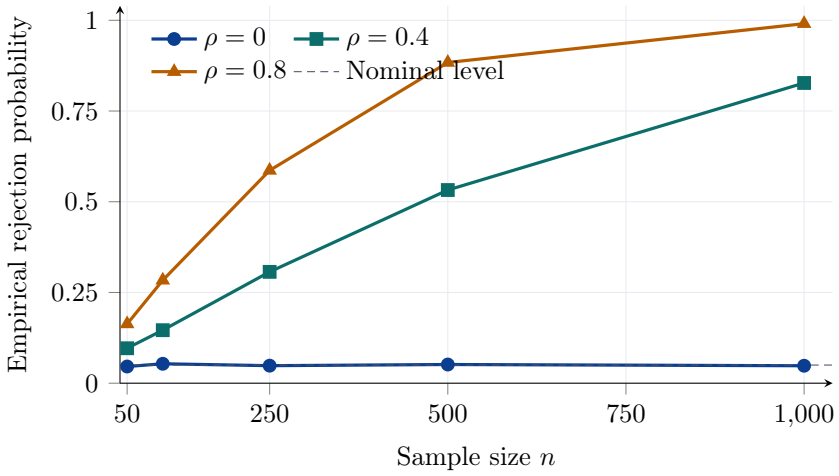


Figure 10

When the signal was stronger, the effect was even more severe. With X_1 explaining 75% of the response variance, correlation 0.8, and $n = 250$, the rejection probability was 0.9434; at $n = 1000$, it was essentially one. The distributional family had comparatively little influence.

Measurement error is not restricted to normal linear regression. The same principle affects logistic regression, survival models, log linear models, and coarsened predictors. Converting a noisy quantitative variable into categories or ranks does not recover the lost information. Possible remedies include validation data, replicate measurements, structural measurement error models, and credible instrumental variables. Their success depends on adding information or assumptions sufficient for identifiability.

Lecture 7 — Friday, October 26, 2018

Missing Data and Imputation

Missing value imputation creates a measured surrogate for an unobserved value. It can therefore reproduce the measurement error mechanism in (6.5). Consider

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \quad \beta_2 = 0,$$

with $\mathrm{Corr}(X_1, X_2) = 0.8$. Values of X_1 are missing completely at random, and the fitted regression uses an imputed version W_1 while testing $H_0 : \beta_2 = 0$.

The phrase **missing completely at random** means that the probability of missingness does not depend on observed or unobserved data. This favourable mechanism does not make an arbitrary imputation method valid for inference. The imputed value still has error, and treating it as observed ignores imputation uncertainty.

Definition 7.1 Let $\boldsymbol{R}$ denote the missingness indicators, $\boldsymbol{Y}_{\text{obs}}$ the observed components, and $\boldsymbol{Y}_{\text{mis}}$ the missing components.

1. Data are **missing completely at random** when
$$\mathbb{P}(\boldsymbol{R} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{mis}}) = \mathbb{P}(\boldsymbol{R}).$$
2. Data are **missing at random** when
$$\mathbb{P}(\boldsymbol{R} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{mis}}) = \mathbb{P}(\boldsymbol{R} \mid \boldsymbol{Y}_{\text{obs}}).$$
3. Missingness is **not at random** when the missingness probability still depends on $\boldsymbol{Y}_{\text{mis}}$ after conditioning on the observed data.

The names concern conditional distributions, not visual randomness in a data table. A missingness mechanism is an assumption about how the unseen values relate to the act of being missing. In particular, a variable can be missing more often in one observed treatment group and still be missing at random after conditioning on treatment.

Complete case analysis is unbiased for many regression targets under missing completely at random, but it discards information. If each of p predictors is independently missing with probability q, the expected complete case fraction is
$$(1 - q)^p. \tag{7.1}$$
With $p = 5$ and $q = 0.20$, only 32.768% of cases are expected to be complete. Correlated missingness or different missingness rates change (7.1), but the exponential loss with the number of variables remains an important warning.

Single deterministic imputation alters two parts of the analysis at once. It changes the empirical covariance matrix and then treats the altered values as known. To see the first effect, suppose a fraction q of a centred variable X is replaced by its mean, zero. Ignoring the negligible distinction

between the sample and population mean,

$$\operatorname{Var}(W) \approx (1-q)\operatorname{Var}(X).$$

If missingness is independent and a centred companion variable Z is retained, then

$$\operatorname{Cov}(W, Z) \approx (1-q)\operatorname{Cov}(X, Z).$$

The corresponding correlation is attenuated by approximately $\sqrt{1-q}$. In a multiple regression, these changes propagate through the inverse covariance matrix in (6.2); the result is not generally a simple attenuation of one coefficient.

Example 7.2 The simulation study repeatedly generated data under the null hypothesis and then reproduced the complete imputation and testing procedure, as summarized in Figure 11.

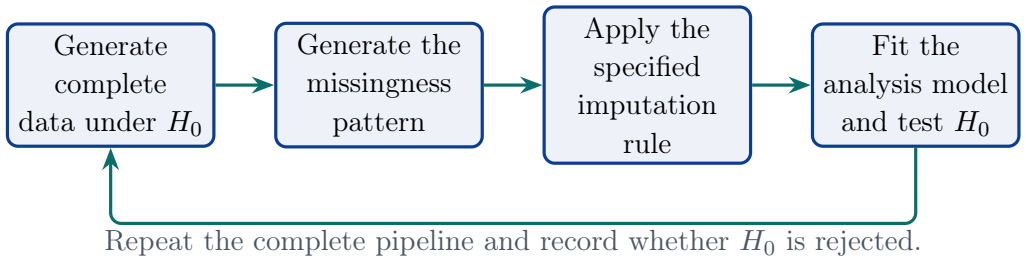


Figure 11

The simulation results in Table 3 show how strongly the choice of method affects a nominal level-0.05 test.

Imputation strategy	Main setting	Rejection probability
Mean or median substitution	$n = 200$	0.5881
Random observed value substitution	$n = 200$	0.8921
One nearest neighbour	25% missing	0.5390
Regression prediction of X_1 from X_2	25% missing	0.0340
Nearest neighbours using (X_1, X_2, Y)	response included	0.2730
Ten neighbour imputation, three predictors	$n = 200$	0.1960
Ten neighbour imputation, three predictors	$n = 500$	0.2340
Same, with an informative third predictor	$n = 500$	0.1860

Table 3

Mean substitution reduces variance and weakens the relationship between the imputed covariate and other variables. Random substitution preserves a marginal distribution but destroys joint relationships. Nearest neighbour methods preserve local patterns only to the extent that the distance variables are informative and properly scaled.

Regression imputation performed well in the simple simulation because it used the correct linear relationship between the two predictors. Its apparent success is not a general guarantee. Single imputation treats predicted values as known, understates uncertainty, and may fail when the imputation model is wrong or the number of complete cases is small.

Using the response to impute an explanatory variable changes the joint distribution in a way that can contaminate the subsequent test. Omitting the response may also be improper when it contains information about the missing value. A principled imputation model must be compatible with the eventual analysis and propagate uncertainty, as multiple imputation does.

Multiple imputation replaces each missing value by several draws rather than one fixed prediction. Let $\widehat{Q}_\ell$ and U_ℓ be the estimate and its complete data variance from imputed data set ℓ, for $\ell = 1, \ldots, M$. The pooled estimate is

$$\overline{Q} = \frac{1}{M} \sum_{\ell=1}^{M} \widehat{Q}_\ell.$$

Define the average within imputation variance and the between imputation variance by

$$\overline{U} = \frac{1}{M}\sum_{\ell=1}^{M} U_\ell \qquad \text{and} \qquad B = \frac{1}{M-1}\sum_{\ell=1}^{M}(\widehat{Q}_\ell - \overline{Q})^2.$$

The total variance estimate

$$T = \overline{U} + \left(1 + \frac{1}{M}\right) B \tag{7.2}$$

adds uncertainty about the missing values to the usual complete data uncertainty. Formula (7.2) does not rescue a poor imputation model; it addresses the additional defect of pretending that imputed values were observed without error. The contrast among the principal imputation methods is summarized graphically in Figure 12.

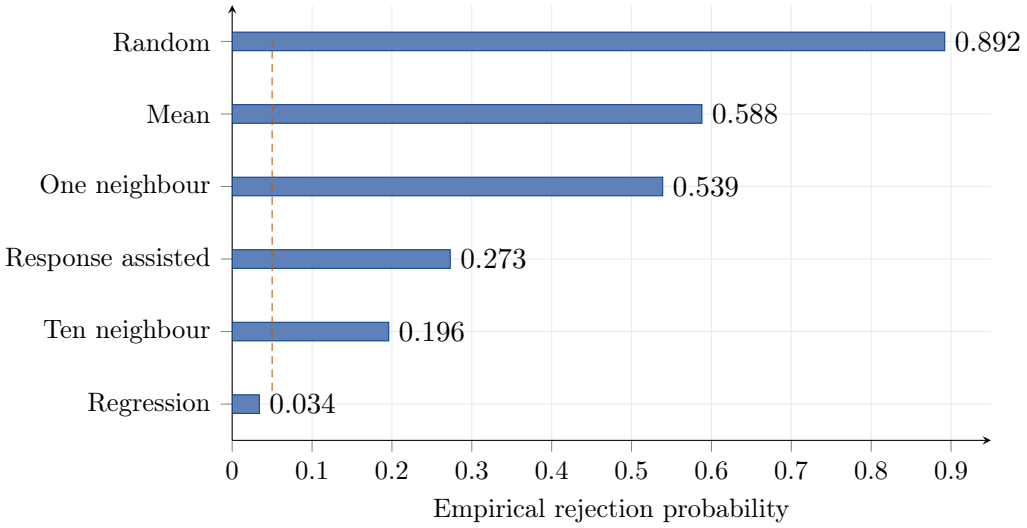


Figure 12

The five predictor experiment exposed a further danger. With 20% missingness in each predictor, the probability that a case was complete was only

$$(1 - 0.20)^5 = 0.32768.$$

The nearest neighbour procedure sometimes failed for particular missingness patterns. Retaining only successful imputations produced rejection proportions near 0.061–0.072, but this apparent improvement came with strong selection: in one condition, 22,810 attempted

data sets were needed to obtain 1,000 successful fits. Conditioning on algorithmic success changes the simulation population and can hide severe practical instability.

The compact simulation skeleton in the following code makes the inferential target and the missingness mechanism visible without reproducing every experimental variation.

```
for (sim in seq_len(M)) {
  X <- MASS::mvrnorm(n, mu = c(0, 0), Sigma = Sigma)
  Y <- beta0 + beta1 * X[, 1] + rnorm(n, sd = sigma)

  missing <- rbinom(n, 1, p_missing) == 1
  W1 <- X[, 1]
  W1[missing] <- NA

  x2 <- X[, 2]
  imp_fit <- lm(W1 ~ x2)
  W1[missing] <- predict(imp_fit,
                         newdata = data.frame(x2 = x2[missing]))
  p_value[sim] <- coef(summary(lm(Y ~ W1 + X[, 2])))[3, 4]
}
mean(p_value < 0.05)
```

The notation required by a software formula can make the displayed code less elegant than the underlying idea. The methodological point is to simulate under a true null, reproduce the complete analysis pipeline, and estimate its actual rejection probability.

Confidence and Prediction Intervals

Return to the normal linear model

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \text{and} \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

For a fixed vector $\boldsymbol{l} \in \mathbb{R}^p$, the target $\boldsymbol{l}^\mathsf{T}\boldsymbol{\beta}$ is a linear combination of regression coefficients, and

$$\boldsymbol{l}^\mathsf{T}\widehat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{l}^\mathsf{T}\boldsymbol{\beta}, \sigma^2 \boldsymbol{l}^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\boldsymbol{l}\right).$$

Because

$$\frac{(n-p)\,\mathrm{MSE}}{\sigma^2} \sim \chi^2_{n-p}$$

independently of $\widehat{\boldsymbol{\beta}}$, an exact $100(1-\alpha)\%$ confidence interval is

$$\boldsymbol{l}^\mathsf{T}\widehat{\boldsymbol{\beta}} \pm t_{n-p,1-\alpha/2}\sqrt{\mathrm{MSE}\;\boldsymbol{l}^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\boldsymbol{l}}. \tag{7.3}$$

For a covariate vector $\boldsymbol{x}_\star$, choosing $\boldsymbol{l} = \boldsymbol{x}_\star$ gives a confidence interval for the conditional mean

$$\mu_\star = \mathbb{E}[Y \mid \boldsymbol{x}_\star] = \boldsymbol{x}_\star^\mathsf{T}\boldsymbol{\beta}.$$

This interval quantifies uncertainty in the fitted regression surface at $\boldsymbol{x}_\star$.

Now let $Y_\star$ be a future response satisfying

$$Y_\star = \boldsymbol{x}_\star^\mathsf{T}\boldsymbol{\beta} + \varepsilon_\star \qquad \text{and} \qquad \varepsilon_\star \sim \mathcal{N}(0,\sigma^2),$$

independently of the fitted sample. The prediction error is

$$Y_\star - \boldsymbol{x}_\star^\mathsf{T}\widehat{\boldsymbol{\beta}} = \varepsilon_\star - \boldsymbol{x}_\star^\mathsf{T}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

The two terms are independent, so

$$\mathrm{Var}\left(Y_\star - \boldsymbol{x}_\star^\mathsf{T}\widehat{\boldsymbol{\beta}}\right) = \sigma^2\left\{1 + \boldsymbol{x}_\star^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\boldsymbol{x}_\star\right\}.$$

Therefore, an exact $100(1-\alpha)\%$ prediction interval is

$$\boldsymbol{x}_\star^\mathsf{T}\widehat{\boldsymbol{\beta}} \pm t_{n-p,1-\alpha/2}\sqrt{\mathrm{MSE}\,\{1 + \boldsymbol{x}_\star^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\boldsymbol{x}_\star\}}. \tag{7.4}$$

The extra 1 in (7.4) is irreducible future case variation. A mean confidence interval can become narrow as the sample grows, but an individual prediction interval retains the random spread of a new response. The mean confidence band and the wider prediction band are compared in Figure 13.

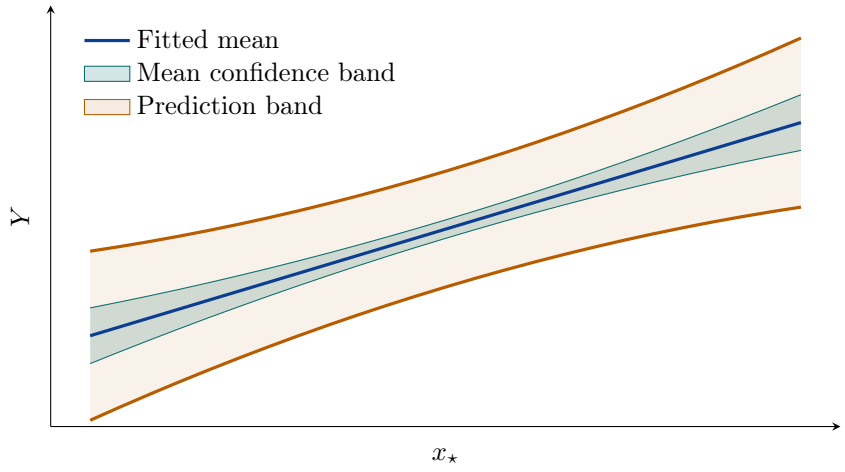


Figure 13

Both intervals widen away from the centre of the observed design because $\boldsymbol{x}_\star^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\boldsymbol{x}_\star$ increases with leverage. This algebraic widening does not make extreme extrapolation safe; outside the observed region, model form error can dominate the nominal uncertainty calculation.

Example 7.3 (Fuel consumption intervals) In a fuel consumption regression based on 95 cars,

$$\text{litres}/100\,\text{km} \sim \text{weight} + \text{length} + \text{country},$$

a Japanese Toyota Camry with weight 1295 kg and length 4.52 m had fitted consumption 12.387. The 95% interval for its mean response was $(11.371, 13.403)$, while the 95% prediction interval for a new comparable car was $(8.857, 15.918)$. The difference in widths illustrates the extra future case variance in (7.4).

The software interface distinguishes the two targets:

```
fit <- lm(lper100k ~ weight + length + country, data = cars)
camry <- data.frame(weight = 1295, length = 4.52,
                    country = "Japan")
predict(fit, camry, interval = "confidence", level = 0.95)
predict(fit, camry, interval = "prediction", level = 0.95)
```

Lecture 8 — Friday, November 02, 2018

Regression and External Prediction

Example 8.1 (External prediction of mathematics performance) The mathematics performance data initially contained 579 records. Several numeric fields used special values such as 0, 998, and 999 for missing or inapplicable measurements. These values must be recoded before modelling; otherwise they become extreme observations rather than missing values. After cleaning, 287 complete cases remained for the main analysis.

An initial model included 19 predictors. Joint tests supported removing nation, course, and sex effects. The resulting model was

$$\begin{aligned}\widehat{\text{mark}} = &-71.4517 + 0.9721\,\text{diagnostic} + 1.5892\,\text{GPA} \\ &+ 0.2176\,\text{calculus} - 0.3002\,\text{English} + 4.6966\,\mathbb{1}\{\text{other language}\}.\end{aligned}$$

Its adjusted coefficient of determination was 0.454. Forward selection returned the same set of predictors, but this agreement does not constitute independent confirmation: both procedures used the same data.

The model was evaluated on a separate replication sample. English score and language did not replicate at a Bonferroni adjusted level of 0.01, while diagnostic test and grade point average remained useful. This distinction between exploration and confirmation is fundamental. A variable selected because it looked promising in one sample requires new data for an honest test of the same claim.

Prediction in the replication sample had to accommodate different missing predictor patterns. For each case, a model using the available predictors was fit to the training data, and a prediction interval was then formed for that case. Across the replication cases,

$$\text{MAE} = 11.4345, \qquad R^2_{\text{prediction}} = 0.3392, \qquad \text{and} \qquad \text{interval coverage} = 0.95.$$

The **mean absolute error** is

$$\text{MAE} = \frac{1}{m}\sum_{i=1}^{m} |y_i - \widehat{y}_i|.$$

It is expressed in response units and is less sensitive to a few large errors than mean squared error.

Prediction performance and coefficient significance are different goals. A predictor can make a statistically detectable but practically negligible contribution, and a model with individually nonsignificant correlated predictors can predict well. The external sample evaluates the entire fitting and prediction procedure, including data cleaning and any selection steps.

Lecture 9 — Friday, November 16, 2018

Learning, Generalization, and Regularization

Machine learning includes classification, regression, transcription, translation, anomaly detection, missing data completion, denoising, and density estimation. The common objective is to use experience from observed data to perform a task according to a chosen performance measure.

In **supervised learning**, the training data contain inputs and target outputs. In **unsupervised learning**, the target structure is not directly labelled. The distinction concerns the information supplied during training, not whether the final algorithm uses an optimization problem.

The training error measures fit to the data used to choose the model. The test error estimates performance on new cases from the target population. **Generalization** is the ability to maintain good performance beyond the training sample.

Model capacity describes the range of functions a learning procedure can represent. Low capacity can cause **underfitting**: the model cannot capture important structure even in the training data. Excessive capacity can cause **overfitting**: the model follows accidental sample variation and generalizes poorly. The resulting tradeoff between training and test error is shown in Figure 14.

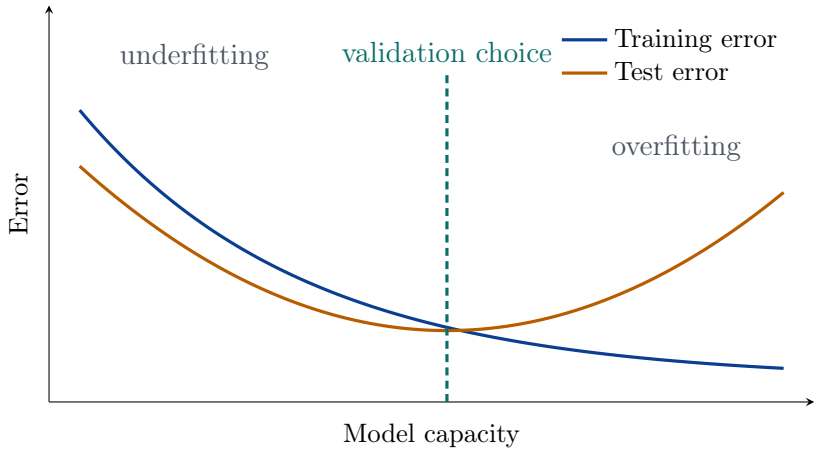


Figure 14

The irreducible lower bound for prediction under the true data-generating distribution is sometimes called the Bayes error. The no free lunch principle says that no learning algorithm is uniformly best over all possible data-generating mechanisms. Inductive assumptions are unavoidable; they encode which patterns the learner expects to generalize.

Regularization controls capacity by adding a preference to the training criterion. Ridge regression, for example, chooses

$$\widehat{\boldsymbol{\beta}}_\lambda = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta})^2 + \lambda \boldsymbol{\beta}^\mathsf{T}\boldsymbol{\beta} \right\}. \tag{9.1}$$

The tuning parameter λ is not estimated by ordinary least squares. It is a **hyperparameter** chosen through a validation procedure. Larger λ shrinks coefficients more strongly and reduces variance at the cost of additional bias.

If a data set is divided into training, validation, and test portions, the training portion estimates parameters, the validation portion chooses hyperparameters and model variants, and the test portion is used once for final assessment. Repeatedly inspecting test performance turns the test set into another validation set and invalidates its role as an independent assessment.

When data are limited, K-fold cross-validation divides the sample into K approximately equal folds. For each fold, fit the procedure on the other $K - 1$ folds and score it on the held out fold.

Average the held out losses:

$$\widehat{R}_{\mathrm{CV}} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{|I_k|}\sum_{i\in I_k} L(y_i, \widehat{f}^{(-k)}(\boldsymbol{x}_i)). \tag{9.2}$$

Every response is evaluated out of sample, although the fold scores are dependent because their training sets overlap. The rotating validation fold is visualized in Figure 15.

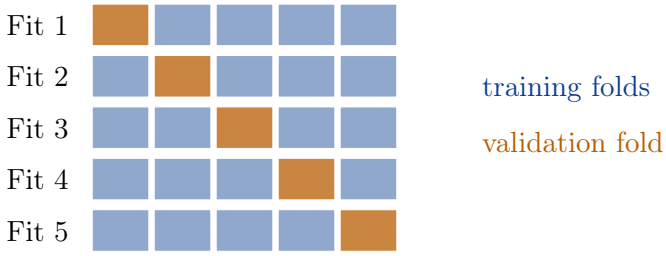


Figure 15

Bayesian prediction averages over posterior parameter uncertainty:

$$p(y_\star \mid \boldsymbol{x}_\star, \boldsymbol{y}) = \int p(y_\star \mid \boldsymbol{x}_\star, \boldsymbol{\theta})\pi(\boldsymbol{\theta} \mid \boldsymbol{y})\,\mathrm{d}\boldsymbol{\theta}.$$

Maximum a posteriori estimation instead maximizes the posterior density and can often be interpreted as penalized likelihood. Support vector machines seek a large margin separating surface, while stochastic gradient descent updates parameters from small random subsets of the observations. These methods differ in construction, but all require a clearly defined target population, loss function, validation protocol, and final test.

Nearest Neighbour Regression

For a query point $\boldsymbol{x}_\star$, k nearest neighbour regression finds the k training inputs closest to $\boldsymbol{x}_\star$ and predicts

$$\widehat{f}_k(\boldsymbol{x}_\star) = \frac{1}{k}\sum_{i\in N_k(\boldsymbol{x}_\star)} y_i. \tag{9.3}$$

This is a local averaging method. Small k produces a flexible, low bias but high variance predictor. Large k smooths more heavily, increasing bias and reducing variance.

Distance depends on units. If one explanatory variable ranges in thousands and another ranges

in tenths, Euclidean distance will be dominated by the first unless the inputs are standardized. Irrelevant variables also degrade neighbourhoods, especially in high dimensions, where points become uniformly far apart. This is one manifestation of the curse of dimensionality. The local neighbourhood used for a query point is illustrated in Figure 16.

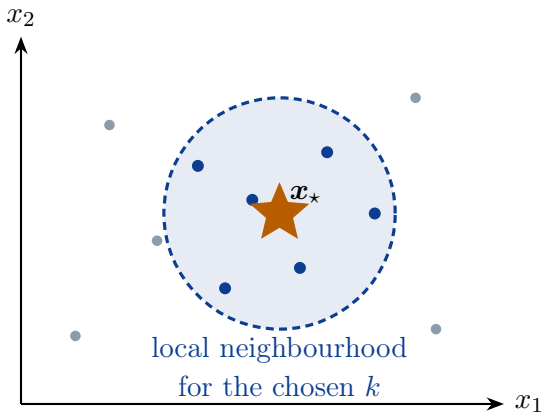


Figure 16

Example 9.1 (Nearest neighbour prediction of mathematics performance) The mathematics data were divided into fitting and validation portions. Predicting every case by the training mean gave validation mean absolute error 14.61. A full nearest neighbour model improved this to about 13.21. Forward validation based variable selection favoured grade point average, diagnostic test, and calculus score, with validation mean absolute error about 12.56.

The neighbour count was then tuned on validation data; values near $k = 25$ produced mean absolute error around 11.87 in one split. This number is itself variable because the training and validation split is random. The same validation data should not be used to report the final performance.

On the independent replication sample, nearest neighbour regression achieved

$$R^2_{\text{prediction}} = 0.2204 \qquad \text{and} \qquad \text{MAE} = 12.938.$$

The corresponding least squares procedure achieved $R^2_{\text{prediction}} = 0.3392$ and MAE $= 11.4345$. Thus the more flexible method did not generalize better in this problem. Flexibility is a resource to tune, not a guarantee of accuracy.

The central workflow fits in a few lines:

```
scale_fit <- preProcess(train[predictors],
                        method = c("center", "scale"))
x_train <- predict(scale_fit, train[predictors])
x_valid <- predict(scale_fit, valid[predictors])

mae <- sapply(k_grid, function(k) {
  pred <- FNN::knn.reg(x_train, x_valid,
                       y = train$mark, k = k)$pred
  mean(abs(valid$mark - pred))
})
k_best <- k_grid[which.min(mae)]
```

Scaling parameters are learned from the training data and then applied to validation or test inputs. Computing them from the entire data set would leak information from held out cases into the fitted procedure.

Lecture 10 — Friday, November 23, 2018

5. Interactions, Factorial Designs, and Within Case Data

Interactions describe effects that vary across conditions. Factorial designs organize these comparisons, and mixed models extend them to data in which the same case contributes several dependent responses.

Interactions in Regression

An **interaction** is present when the relationship between one explanatory variable and the conditional mean response depends on the value of another explanatory variable. This statement is symmetric: if the effect of X_1 depends on X_2, then the effect of X_2 depends on X_1.

For two quantitative variables, the model

$$\mathbb{E}[Y \mid x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \tag{10.1}$$

has partial slopes

$$\frac{\partial}{\partial x_1}\mathbb{E}[Y \mid x_1, x_2] = \beta_1 + \beta_3 x_2 \qquad \text{and} \qquad \frac{\partial}{\partial x_2}\mathbb{E}[Y \mid x_1, x_2] = \beta_2 + \beta_3 x_1.$$

Thus $H_0 : \beta_3 = 0$ is the hypothesis of no interaction. When $\beta_3 \neq 0$, neither β_1 nor β_2 is an overall main effect: β_1 is the x_1 slope specifically at $x_2 = 0$, and β_2 is the x_2 slope specifically at $x_1 = 0$.

Centring quantitative variables makes lower order coefficients easier to interpret. If $z_j = x_j - c_j$, then a model containing both main effects and their product represents the same fitted surface after reparameterization. Choosing c_j to be a meaningful reference value makes a lower order slope describe the effect at that reference rather than at an arbitrary zero.

For a quantitative variable x and a binary group indicator d, consider

$$\mathbb{E}[Y \mid x, d] = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 x d. \tag{10.2}$$

The two group regressions are

$$\begin{aligned} d = 0 : &\qquad \mathbb{E}[Y \mid x, d = 0] = \beta_0 + \beta_1 x, \\ d = 1 : &\quad \mathbb{E}[Y \mid x, d = 1] = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x. \end{aligned}$$

Here β_2 compares groups at $x = 0$, and β_3 compares their slopes. If x is centred at $\overline{x}$, the coefficient of d compares the groups at the sample mean of x.

With three groups, two indicators and two product terms allow three separate lines. Let group 3 be the reference:

$$\mathbb{E}[Y \mid x, d_1, d_2] = \beta_0 + \beta_1 x + \beta_2 d_1 + \beta_3 d_2 + \beta_4 x d_1 + \beta_5 x d_2.$$

The equal slopes hypothesis is

$$H_0 : \beta_4 = \beta_5 = 0.$$

If it is retained, $H_0 : \beta_2 = \beta_3 = 0$ tests equality of the parallel regressions. If the interaction is present, a group comparison must be made at a stated value of x. The contrast between equal slopes and group dependent slopes is shown in Figure 17.

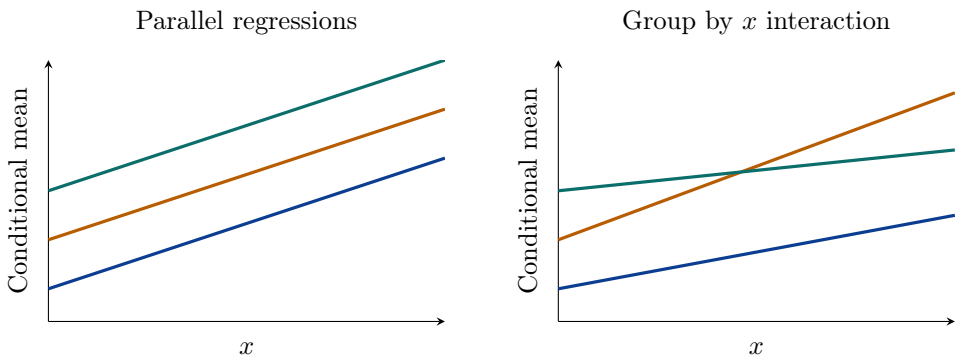


Figure 17

Example 10.1 (Weight and fuel consumption by country) In the car data, the model

$$\text{fuel} \sim \text{weight} * \text{country}$$

allowed the weight slope to vary across United States, Japanese, and European cars. The interaction test gave

$$F = 4.3573 \qquad \text{and} \qquad p = 0.0155.$$

The fitted weight slopes were steeper for Japanese and European cars than for United States cars. An overall country comparison without specifying weight would therefore be incomplete.

The software formula is brief:

```
fit_parallel <- lm(fuel ~ weight + country, data = cars)
fit_interact <- lm(fuel ~ weight * country, data = cars)
anova(fit_parallel, fit_interact)

weight_ref <- mean(cars$weight)
cars$weight_c <- cars$weight - weight_ref
fit_centered <- lm(fuel ~ weight_c * country, data = cars)
```

Centring changes the reported country coefficients but not the fitted values, residuals, interaction test, or comparisons evaluated at the same physical weight.

Factorial Designs and Cell Means

A **factor** is a categorical explanatory variable whose possible values are called levels. A study with observations at every combination of factor levels has a **complete factorial design**. Random assignment of experimental units to the treatment combinations permits causal interpretation of the treatment contrasts.

Example 10.2 (Potato rot experiment) In the potato study, potato pieces were inoculated with one of three bacterial types and stored at either a cool or warm temperature. The response was the percentage of surface area with visible rot. There are $3 \times 2 = 6$ treatment conditions. Let

$$\mu_{ij} = \mathbb{E}[Y \mid \text{temperature } i, \text{bacteria } j], \quad i = 1, 2 \qquad \text{and} \qquad j = 1, 2, 3.$$

The six cell means completely describe the conditional mean structure.

For a balanced design, define the temperature marginal means

$$\mu_{i\cdot} = \frac{1}{3}\sum_{j=1}^{3} \mu_{ij},$$

the bacterial marginal means

$$\mu_{\cdot j} = \frac{1}{2}\sum_{i=1}^{2} \mu_{ij},$$

and the grand mean

$$\mu_{\cdot\cdot} = \frac{1}{6}\sum_{i=1}^{2}\sum_{j=1}^{3} \mu_{ij}.$$

The temperature main effect concerns differences among the $\mu_{i\cdot}$; the bacterial main effect concerns differences among the $\mu_{\cdot j}$.

An interaction is a difference of differences. For bacterial levels j and k, the temperature contrast is

$$(\mu_{2j} - \mu_{1j}) - (\mu_{2k} - \mu_{1k}).$$

There is no interaction precisely when all such contrasts are zero. Equivalently,

$$\mu_{ij} = \mu_{\cdot\cdot} + (\mu_{i\cdot} - \mu_{\cdot\cdot}) + (\mu_{\cdot j} - \mu_{\cdot\cdot}). \tag{10.3}$$

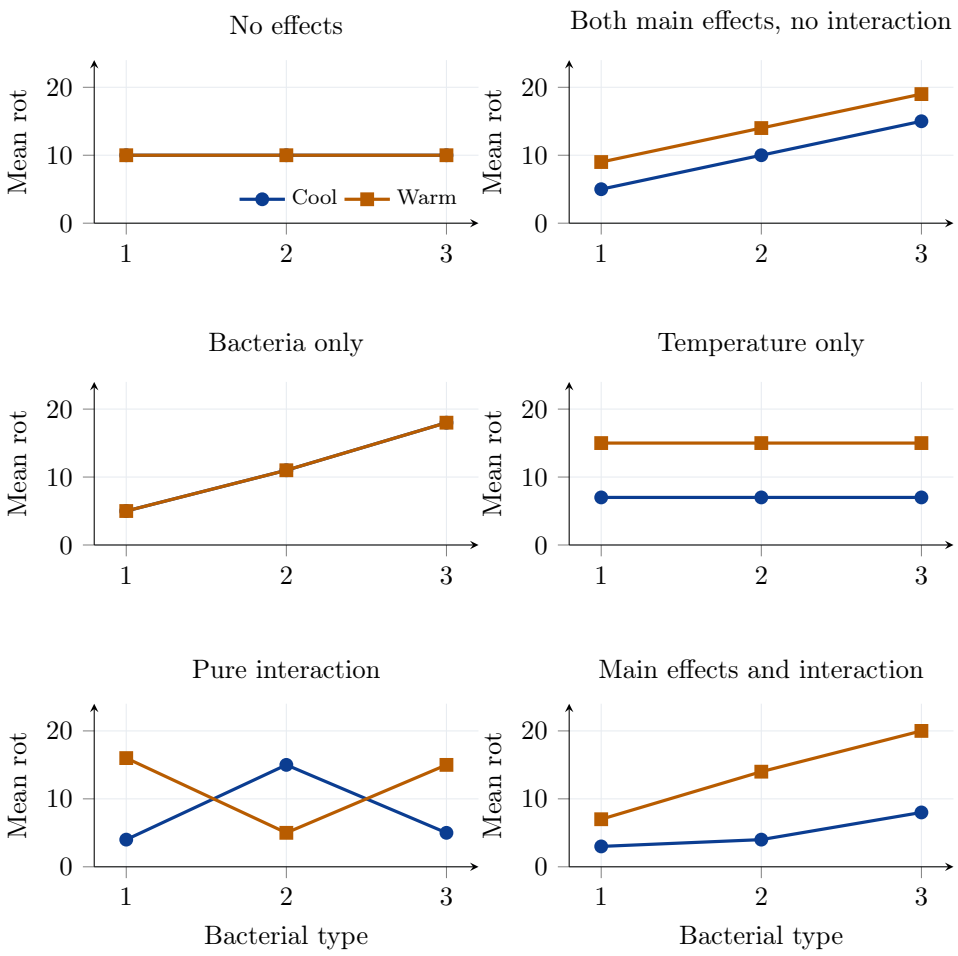


Figure 18

Parallel mean profiles in Figure 18 correspond to no interaction. Crossing or nonparallel profiles indicate interaction, but visual size must be judged relative to sampling variability. With interaction, a marginal main effect averages across factor levels that may have very different simple effects. Such an average can be mathematically defined yet scientifically misleading.

Cell means coding uses one indicator for each treatment combination and omits the intercept:

$$\mathbb{E}[Y \mid \boldsymbol{x}] = \sum_{i=1}^{2} \sum_{j=1}^{3} \mu_{ij} x_{ij}, \tag{10.4}$$

where exactly one x_{ij} equals one for each observation. Every regression coefficient is then a cell

mean. This coding is especially convenient for planned contrasts.

For example, the average contrast between warm and cool conditions is

$$C_T = \frac{1}{3}\{(\mu_{21} - \mu_{11}) + (\mu_{22} - \mu_{12}) + (\mu_{23} - \mu_{13})\}.$$

The bacterial main effect has two degrees of freedom and may be expressed by any two independent contrasts among the three bacterial marginal means. The interaction has

$$(2 - 1)(3 - 1) = 2$$

degrees of freedom.

Example 10.3 (Potato rot analysis) In the balanced potato experiment, there were nine observations per cell. The sample cell means were

	Bacteria 1	Bacteria 2	Bacteria 3
Cool	3.556	4.778	8.000
Warm	7.000	13.556	19.556

The profile is shown in Figure 19.

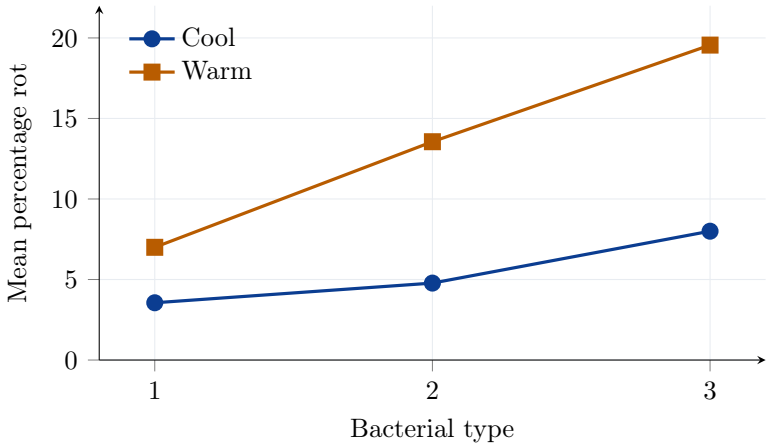


Figure 19

The overall test among the six cell means gave $F = 15.0509$. The temperature main effect

test gave $F = 38.6138$, the bacterial main effect test gave $F = 14.8390$, and the interaction test gave

$$F = 3.4815 \qquad \text{and} \qquad p = 0.0387.$$

Because the interaction is detectable, simple effects are informative. At cool temperature, the bacterial comparison gave $F = 2.160$ and $p = 0.126$, whereas at warm temperature it gave $F = 16.160$ with a very small p-value. A contrast comparing the warm minus cool effect for bacteria 1 with that for bacteria 3 gave $F = 6.74$ and $p = 0.0125$.

For three or more factors, terminology follows the same principle. A two way interaction says that the relation between two factors depends on their levels. A three way interaction says that a two way interaction depends on the level of a third factor. A four way interaction compares three way interactions across a fourth factor. Interpretation should proceed from the highest order supported interaction downward.

Effect Coding and Planned Contrasts

Reference cell coding compares each nonreference level with a selected baseline. Effect coding instead assigns the omitted level the value -1 on every indicator. For a factor with three levels, one convenient coding is

Level	x_1	x_2
1	1	0
2	0	1
3	-1	-1

in the model

$$\mathbb{E}[Y \mid x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

The resulting means are

$$\mu_1 = \beta_0 + \beta_1, \qquad \mu_2 = \beta_0 + \beta_2, \qquad \text{and} \qquad \mu_3 = \beta_0 - \beta_1 - \beta_2.$$

Consequently,

$$\beta_0 = \frac{\mu_1 + \mu_2 + \mu_3}{3},$$

and β_1, β_2 are deviations from this unweighted grand mean. The coding changes coefficient interpretation, not the fitted cell means or any test of a coding invariant hypothesis.

For a 3×2 design, effect coding for the bacterial factor and the temperature factor is

Condition	b_1	b_2	t
Bacteria 1, cool	1	0	1
Bacteria 2, cool	0	1	1
Bacteria 3, cool	-1	-1	1
Bacteria 1, warm	1	0	-1
Bacteria 2, warm	0	1	-1
Bacteria 3, warm	-1	-1	-1

Product columns $b_1 t$ and $b_2 t$ encode the interaction. The saturated mean model is

$$\mathbb{E}[Y \mid b_1, b_2, t] = \beta_0 + \beta_1 b_1 + \beta_2 b_2 + \beta_3 t + \beta_4 b_1 t + \beta_5 b_2 t. \tag{10.5}$$

In a balanced design, β_0 is the grand mean, β_3 is half the difference in marginal means between the cool and warm conditions, and β_4, β_5 describe interaction deviations.

A **contrast** among p population means is

$$C = \sum_{j=1}^{p} a_j \mu_j \qquad \text{and} \qquad \sum_{j=1}^{p} a_j = 0. \tag{10.6}$$

The restriction that the coefficients sum to zero makes the comparison invariant to adding a common constant to every mean. Its sample estimator is

$$\widehat{C} = \sum_{j=1}^{p} a_j \overline{Y}_j.$$

Under independent groups with common variance,

$$\mathrm{se}(\widehat{C}) = \sqrt{\mathrm{MSE} \sum_{j=1}^{p} \frac{a_j^2}{n_j}}.$$

Any collection of r contrasts can be tested through the general linear hypothesis (2.7).

Contrasts should be chosen to answer scientific questions. In a dose study, one contrast might compare placebo with the average of all active doses, another might represent a linear dose trend, and a third might measure curvature. Orthogonality can simplify interpretation in balanced designs, but scientific relevance is more important than algebraic convenience.

When covariates are present, raw marginal means may average over different covariate distributions in different treatment groups. Model based adjusted means evaluate treatment means at a common covariate distribution or reference value. These are sometimes called least squares means. Their definition must state the averaging population because, in an unbalanced design with interactions, different weighting rules answer different questions.

Sequential sums of squares allocate shared variation according to the order in which terms enter a model. In a balanced orthogonal design, this order does not matter. In an unbalanced design, it can matter substantially. A safer approach is to formulate the scientific null hypothesis as a full versus reduced model comparison or as an explicit contrast matrix.

Warning Software may silently choose sequential, marginal, or coding dependent hypotheses. Inspect the model matrix and state the null hypothesis before interpreting a test table.

Example 10.4 (Potato rot analysis in R) For the potato data, the main analyses can be expressed directly:

```
potato$bacteria <- C(potato$bacteria, contr.sum)
potato$temp <- C(potato$temp, contr.sum)

fit <- lm(rot ~ bacteria * temp, data = potato)
car::Anova(fit, type = 3)

cell_fit <- lm(rot ~ 0 + interaction(temp, bacteria),
               data = potato)
car::linearHypothesis(cell_fit, L)
```

The first fit gives effect coded main and interaction tests. The cell means fit is convenient for custom contrasts through the matrix **L**.

Fractional Factorial Designs

A complete factorial design can be expensive because the number of treatment combinations multiplies across factors. A **fractional factorial design** observes only a selected subset of combinations. The reduction in cost is paid for by assumptions: some effects or interactions become inseparable.

The simplest example is a 2×2 design with one missing cell. Write the cell means as

	$B = 1$	$B = 2$
$A = 1$	μ_{11}	μ_{12}
$A = 2$	μ_{21}	μ_{22}

and suppose μ_{22} is unobserved. The no interaction restriction is

$$\mu_{11} - \mu_{12} - \mu_{21} + \mu_{22} = 0.$$

It identifies the missing cell mean as

$$\mu_{22} = \mu_{12} + \mu_{21} - \mu_{11}. \tag{10.7}$$

The observed data alone do not verify this equality. It is the assumption that replaces the missing treatment condition.

More generally, let $\boldsymbol{\mu}$ be the vector of complete design cell means and let $\boldsymbol{\beta}$ be a vector of effect coded parameters. There is a nonsingular design matrix $\mathbf{A}$ satisfying

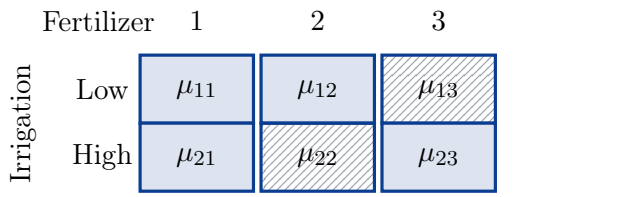
$$\boldsymbol{\mu} = \mathbf{A}\boldsymbol{\beta} \qquad \text{and} \qquad \boldsymbol{\beta} = \mathbf{A}^{-1}\boldsymbol{\mu}.$$

Removing treatment combinations deletes rows of $\mathbf{A}$. To retain identifiability, remove an equal number of columns corresponding to effects assumed absent. The remaining square submatrix must be nonsingular.

Algorithm 10.5 (Constructing an identifiable fraction) For a proposed fractional factorial design:

1. Write the complete cell mean model and its effect coded design matrix.
2. Identify the treatment combinations that can be observed.
3. State which effects or interactions will be assumed absent.
4. Delete the corresponding rows and columns.
5. Verify that the remaining matrix has full rank.
6. Record which estimable contrasts the retained coefficients represent.

Example 10.6 As an illustration, consider three fertilizers and two irrigation levels. The complete 3×2 design has six cell means and six saturated parameters: one grand mean, two fertilizer components, one irrigation component, and two interaction components. If only four cells can be observed, two independent components must be sacrificed. Assuming the entire interaction absent leaves the additive four parameter model identifiable for a suitable connected set of four cells. The observed and omitted treatment combinations are displayed in Figure 20.



Four observed cells identify an additive model;
the two interaction components are assumed absent.

Figure 20

A disconnected fraction cannot compare all factor levels, even under additivity. For example, observing only diagonal cells can confound fertilizer with irrigation. Full rank verification in Algorithm 10.5 detects this problem.

Fractional designs are most compelling when subject matter knowledge makes high order interactions negligible and when design generators deliberately control aliasing. In observational data, empty cells may arise accidentally; then an additive model can mathematically fill them, but the required extrapolation should be stated explicitly.

Normal Within Case Data

Repeated measures and longitudinal studies observe several responses from the same case. Responses from one case are usually dependent. Treating them as independent gives the wrong sampling variance and can reverse a conclusion.

The matched pairs setting is the simplest example. Let

$$Y_{i1} = \mu_1 + b_i + \varepsilon_{i1} \qquad \text{and} \qquad Y_{i2} = \mu_2 + b_i + \varepsilon_{i2},$$

where b_i is a case specific random effect. The within case difference is

$$D_i = Y_{i2} - Y_{i1} = (\mu_2 - \mu_1) + (\varepsilon_{i2} - \varepsilon_{i1}),$$

so the shared case effect cancels. A paired t-test is a one sample test on the D_i and uses this cancellation. An independent samples test discards the pairing and includes between case variation that is irrelevant to the within case contrast.

The general normal linear mixed model is

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{b} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}), \qquad \text{and} \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{R}), \tag{10.8}$$

with $\boldsymbol{b} \perp\!\!\!\perp \boldsymbol{\varepsilon}$. The vector $\boldsymbol{\beta}$ contains fixed effects shared by the population. The vector $\boldsymbol{b}$ contains random deviations associated with sampled cases or other random units.

Marginalizing over the random effects gives

$$\boldsymbol{Y} \sim \mathcal{N}\left(\mathbf{X}\boldsymbol{\beta},\ \mathbf{Z}\mathbf{G}\mathbf{Z}^{\mathsf{T}} + \mathbf{R}\right). \tag{10.9}$$

The covariance structure, rather than a change in the fixed mean formula, is what accommodates within case dependence.

Example 10.7 Suppose treatment i is applied to several plots on farm j:

$$Y_{ij} = \mu + \tau_i + b_j + \varepsilon_{ij}, \qquad b_j \sim \mathcal{N}(0, \sigma_b^2), \qquad \text{and} \qquad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2).$$

Two observations from the same farm have covariance σ_b^2, while observations from different farms are independent. Within a farm,

$$\mathrm{Var}(Y_{ij}) = \sigma_b^2 + \sigma^2 \qquad \text{and} \qquad \mathrm{Corr}(Y_{ij}, Y_{i'j}) = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2}.$$

For m observations from one case, the random intercept covariance matrix is

$$\mathbf{V}_i = \sigma_b^2 \mathbf{1}_m \mathbf{1}_m^{\mathsf{T}} + \sigma^2 \mathbf{I}_m = \begin{bmatrix} \sigma_b^2 + \sigma^2 & \sigma_b^2 & \cdots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \sigma^2 & \cdots & \sigma_b^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_b^2 & \sigma_b^2 & \cdots & \sigma_b^2 + \sigma^2 \end{bmatrix}. \tag{10.10}$$

Equal marginal variances and equal pairwise correlations constitute compound symmetry. This can be reasonable for exchangeable repeated conditions, but it is often unrealistic for longitudinal

observations whose correlation decreases with time separation. The repeated diagonal and off diagonal entries are visualized in Figure 21.

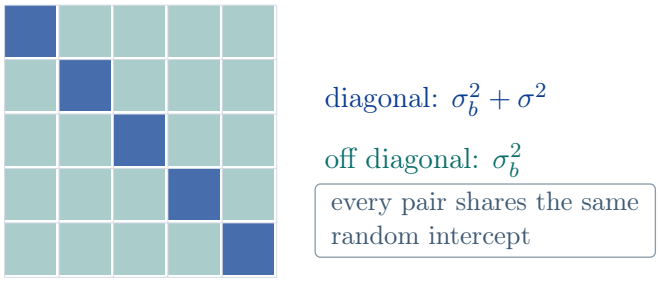


Figure 21

Remark 10.8 (Fixed and random effects) The distinction concerns the inferential population, not whether a variable is numerically coded. A treatment with deliberately selected levels is usually represented by fixed effects because those level contrasts are the objects of interest. A grouping factor is represented by random effects when its observed levels are treated as a sample from a larger population and the variation among levels is itself an inferential target. Thus salt concentration can be a fixed factor while restaurant location is random; a treatment condition can be fixed while participant is random.

For the one factor farm model with q farms and m observations per farm, $\mathbf{Z}$ has one column per farm. Its row for an observation from farm j contains a one in column j and zeros elsewhere. With three farms and two observations per farm,

$$\mathbf{Z} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \qquad \text{and} \qquad \mathbf{G} = \sigma_b^2 \mathbf{I}_3.$$

Therefore $\mathbf{Z}\mathbf{G}\mathbf{Z}^\top$ places σ_b^2 in every position corresponding to a pair from the same farm and zero elsewhere. Adding $\mathbf{R} = \sigma^2 \mathbf{I}_6$ adds the residual variance only to the diagonal. This construction explains the block structure in (10.10) directly from the design matrix.

Theorem 10.9 Suppose

$$\boldsymbol{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}),$$

where $\mathbf{V}$ is known and positive definite and $\mathbf{X}$ has full column rank. Then the maximum likelihood estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}_{\mathbf{V}} = (\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{V}^{-1} \boldsymbol{Y}, \tag{10.11}$$

with covariance

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}_{\mathbf{V}}) = (\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X})^{-1}.$$

Proof. Apart from constants, the negative log likelihood is

$$\frac{1}{2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{V}^{-1} (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}).$$

Its derivative with respect to $\boldsymbol{\beta}$ is $-\mathbf{X}^\top \mathbf{V}^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})$. Setting the derivative equal to zero gives (10.11). The covariance follows by applying (2.2) to the resulting linear transformation of $\boldsymbol{Y}$. □

In a mixed model, $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^\top + \mathbf{R}$ contains unknown covariance parameters. The estimation is therefore iterative: propose covariance parameters, compute the generalized least squares fixed effects, update the covariance fit, and continue until the likelihood stabilizes.

Once the parameters are estimated, the conditional mean of the random effects is

$$\widehat{\boldsymbol{b}} = \mathbf{G}\mathbf{Z}^\top \mathbf{V}^{-1} (\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathbf{V}}). \tag{10.12}$$

This is a prediction, not an estimate of a new population level parameter. It shrinks group specific deviations toward zero. A group with few or noisy observations is shrunk more strongly than a group with abundant precise information.

Example 10.10 Suppose each group has m observations and let $\overline{r}_i$ be group i's mean residual after removing the fixed effects. Then (10.12) reduces to

$$\widehat{b}_i = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2/m}\, \overline{r}_i.$$

The multiplier lies between zero and one. It approaches one when the between group variation dominates or when m is large, and it approaches zero when the group mean is mostly residual

noise.

Classical repeated measures procedures rely on restrictive covariance assumptions and handle imbalance and covariates poorly. Mixed models allow unequal numbers and timings of observations, subject to assumptions about the missing data process and covariance model.

Example 10.11 For a study with sex as a between case factor and three noise conditions measured within each case, a random intercept model is

$$\begin{aligned} Y_{ijk} = \beta_0 &+ \beta_1 \,\mathrm{sex}_i + \beta_2 \,\mathrm{noise}_{2k} + \beta_3 \,\mathrm{noise}_{3k} \\ &+ \beta_4 \,\mathrm{sex}_i \mathrm{noise}_{2k} + \beta_5 \,\mathrm{sex}_i \mathrm{noise}_{3k} + b_i + \varepsilon_{ik}. \end{aligned}$$

The fixed interaction asks whether the within case noise effect differs by sex. The random intercept accounts for stable differences among cases.

A random slope extension for longitudinal time t_{ij}, with independent residuals of variance σ^2 that are independent of the random effects, is

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{0i} + b_{1i} t_{ij} + \varepsilon_{ij} \qquad \text{and} \qquad \begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} \sim \mathcal{N}_2 \left(\mathbf{0}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{bmatrix} \right). \tag{10.13}$$

Cases have individual intercepts and slopes. The induced covariance depends on time:

$$\mathrm{Cov}(Y_{ij}, Y_{ik}) = \sigma_0^2 + \sigma_{01}(t_{ij} + t_{ik}) + \sigma_1^2 t_{ij} t_{ik} + \sigma^2 \mathbb{1}_{\{j=k\}}.$$

Unlike compound symmetry, the covariance need not be constant across pairs.

Maximum likelihood jointly estimates fixed effects and covariance parameters by maximizing the marginal likelihood from (10.9). Ordinary maximum likelihood covariance estimates account for the fixed effect fit only through the maximized likelihood and can be biased downward in small samples.

Restricted maximum likelihood instead forms linear combinations $\mathbf{K}^\top \boldsymbol{Y}$ satisfying

$$\mathbf{K}^\top \mathbf{X} = 0.$$

These error contrasts remove the fixed mean, so their likelihood depends only on covariance parameters. Restricted maximum likelihood commonly reduces small sample variance component bias. Models with different fixed effect spaces should be compared by ordinary maximum likelihood, whereas restricted maximum likelihood is appropriate for comparing covariance structures with

the same fixed effects.

The marginal log likelihood for covariance parameter $\boldsymbol{\psi}$ is, apart from an additive constant,

$$\ell(\boldsymbol{\beta}, \boldsymbol{\psi}) = -\frac{1}{2}\left[\log|\mathbf{V}_{\boldsymbol{\psi}}| + (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^{\mathsf{T}}\mathbf{V}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})\right]. \tag{10.14}$$

Substituting the generalized least squares estimate into (10.14) profiles out the fixed effects. Restricted maximum likelihood includes an additional adjustment,

$$-\frac{1}{2}\log|\mathbf{X}^{\mathsf{T}}\mathbf{V}_{\boldsymbol{\psi}}^{-1}\mathbf{X}|,$$

which accounts for degrees of freedom spent estimating $\boldsymbol{\beta}$. This is why restricted maximum likelihood often reduces the downward bias of variance component estimates.

Warning Variance components lie on a boundary because they cannot be negative. A test of $H_0 : \sigma_b^2 = 0$ therefore violates the interior point assumptions of the ordinary likelihood ratio theorem. A naive chi-squared reference distribution can be wrong; an appropriate boundary result, parametric bootstrap, or specialized small sample procedure is needed.

Lecture 11 — Friday, November 30, 2018

Mixed Model Examples

Example 11.1 (Grapefruit prices) In a grapefruit price experiment, eight stores were observed at three price levels. The mean sales were

Price condition	1	2	3
Mean sales	55.4375	53.6000	51.3375

The random intercept estimates were

$$\widehat{\sigma}_b^2 = 35.257 \qquad \text{and} \qquad \widehat{\sigma}^2 = 0.6837.$$

Most response variation was between stores, so repeated measurements from a store were highly correlated.

A naive independent groups analysis gave $F = 0.9388$ and $p = 0.4069$. The random intercept analysis used the precise within store comparison and gave

$$F = 49.346 \qquad \text{and} \qquad p = 4.57 \times 10^{-7}.$$

Dependence does not always reduce significance. Positive pairing can remove large irrelevant between case variation and make a within case effect much clearer. The individual store trajectories and their mean profile appear in Figure 22.

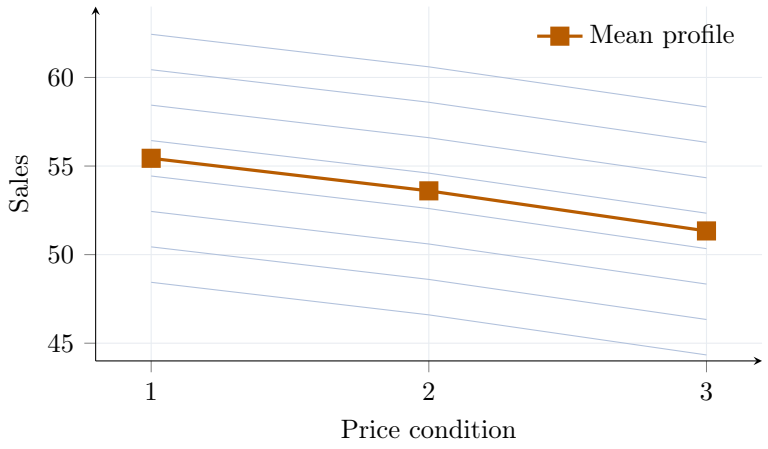


Figure 22

Example 11.2 (Dichotic listening) In a dichotic listening study, 40 participants were classified as left- or right-handed and tested in three ear conditions. The model included handedness as a between case factor, ear condition as a within case factor, their interaction, and a participant random intercept. The tests gave

Effect	p-value
Handedness	0.331
Ear condition	0.0121
Handedness by ear condition	0.329

There was evidence of an ear condition effect but not of a differential effect by handedness. Bonferroni adjusted pairwise comparisons gave $p = 0.0157$ when comparing both ears with the left, $p = 0.00654$ when comparing both ears with the right, and $p = 0.746$ when com-

paring the left with the right. The two nearly parallel handedness profiles are shown in Figure 23.

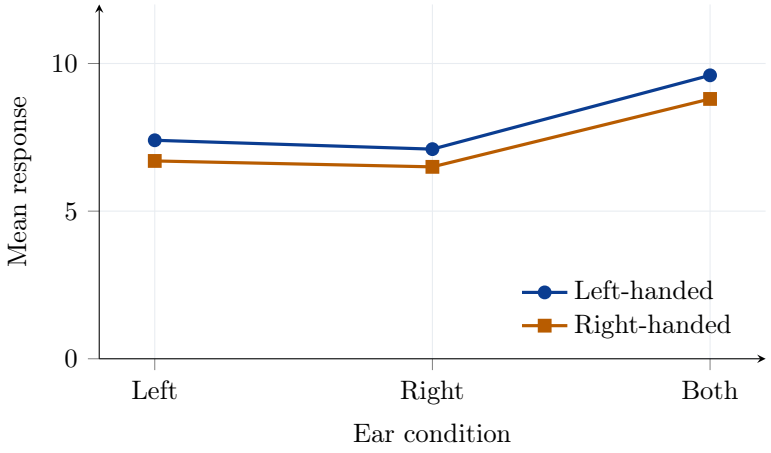


Figure 23

Example 11.3 (Reaction times) An unbalanced reaction time study crossed language as a between case factor with sentence type and context as within case factors. Effect coding made the main effect tests correspond to unweighted marginal comparisons. The three factor mixed model detected a context effect ($p = 2.35 \times 10^{-7}$), a sentence by context interaction ($p = 0.00690$), and a language by sentence by context interaction ($p = 0.0347$).

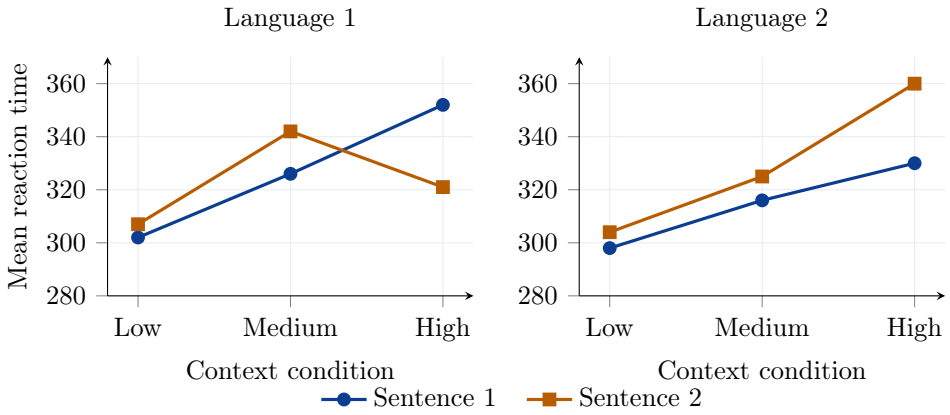


Figure 24

The nonparallel changes in Figure 24 illustrate the meaning of a three way interaction: the sentence by context interaction has a different form across languages. The plotted means are schematic, while the reported tests come from the fitted mixed model.

The core model syntax mirrors the hierarchy:

```
library(lme4)
library(lmerTest)

fit <- lmer(reaction ~ language * sentence * context +
              (1 | participant),
            data = reaction_data, REML = TRUE)
anova(fit, type = 3)

fit_slope <- lmer(reaction ~ time * language +
                    (1 + time | participant),
                  data = longitudinal_data, REML = TRUE)
```

Random effects identify the sampling units that generate dependence. A random intercept for participant is not interchangeable with a fixed participant indicator when population level variance components and prediction for new participants are of interest.

Binary Within Case Data

Suppose case i produces binary responses Y_{ij} under repeated conditions. A random intercept logistic model is

$$Y_{ij} \mid b_i \sim \text{Bernoulli}(\pi_{ij}), \qquad \text{logit}(\pi_{ij}) = \boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\beta} + b_i, \qquad \text{and} \qquad b_i \sim \mathcal{N}(0, \sigma_b^2). \tag{11.1}$$

Responses are conditionally independent given b_i, but they are marginally dependent because responses from the same case share the latent intercept. We also assume that the random intercepts are independent across cases, so the case-specific likelihood contributions are independent.

The random intercept represents persistent heterogeneity in response propensity. A case with $b_i > 0$ has higher conditional odds in every condition. For two distinct observations from the same

case,

$$\operatorname{Cov}(Y_{ij}, Y_{ik}) = \operatorname{Cov}(\mathbb{E}[Y_{ij} \mid b_i], \mathbb{E}[Y_{ik} \mid b_i]) \geq 0.$$

The inequality is strict when $\sigma_b^2 > 0$, because both condition specific probabilities increase strictly with b_i. Unlike the normal random intercept model, this covariance has no simple constant closed form and depends on the fixed predictors.

Conditional on b_i, increasing x_j by one multiplies the odds by e^{β_j}. After integrating over the random effect, the marginal population odds ratio is generally not e^{β_j}. Logistic regression is noncollapsible: conditional and marginal effects can differ even without confounding.

For case i, the marginal likelihood contribution is

$$L_i(\boldsymbol{\beta}, \sigma_b^2) = \int_{-\infty}^{\infty} \left\{ \prod_{j=1}^{m_i} \pi_{ij}(b)^{y_{ij}} [1 - \pi_{ij}(b)]^{1-y_{ij}} \right\} \phi(b; 0, \sigma_b^2) \, \mathrm{d}b. \tag{11.2}$$

The full likelihood is $\prod_{i=1}^{n} L_i$. The integral in (11.2) normally requires numerical approximation. Laplace approximation and adaptive Gaussian quadrature are common choices.

The integration is performed once per independent case. This factorization is why the data must identify the true independent units: treating each binary response as its own case would remove the shared integral and falsely impose marginal independence.

To expose the numerical problem, write the conditional log likelihood for case i as

$$h_i(b) = \sum_{j=1}^{m_i} [y_{ij} \log \pi_{ij}(b) + (1 - y_{ij}) \log(1 - \pi_{ij}(b))] + \log \phi(b; 0, \sigma_b^2).$$

Then

$$L_i = \int_{\mathbb{R}} \exp(h_i(b)) \, \mathrm{d}b.$$

Directly multiplying many small Bernoulli probabilities can underflow to zero. A stable computation centres the integrand at its mode $\widetilde{b}_i$:

$$\log L_i = h_i(\widetilde{b}_i) + \log \int_{\mathbb{R}} \exp\Big(h_i(b) - h_i(\widetilde{b}_i) \Big) \, \mathrm{d}b. \tag{11.3}$$

The exponent in (11.3) is never positive, which greatly improves numerical stability.

Laplace approximation replaces h_i by its quadratic Taylor expansion at $\widetilde{b}_i$:

$$\log L_i \doteq h_i(\widetilde{b}_i) + \frac{1}{2}\log(2\pi) - \frac{1}{2}\log\Big(-h_i''(\widetilde{b}_i)\Big).$$

Adaptive Gaussian quadrature instead evaluates the integrand at several nodes centred and scaled around the same mode. More nodes usually improve accuracy but increase computation. Agreement across reasonable quadrature settings is a useful sensitivity check; disagreement is evidence that the likelihood approximation materially affects the conclusion.

Proposition 11.4 Under (11.1), suppose $j \neq k$ and the two condition specific probabilities $\pi_{ij}(b)$ and $\pi_{ik}(b)$ are increasing functions of b. Then

$$\operatorname{Cov}(Y_{ij}, Y_{ik}) = \operatorname{Cov}(\pi_{ij}(b_i), \pi_{ik}(b_i)) \geq 0.$$

Proof. Conditional independence makes the expected conditional covariance zero. The law of total covariance therefore gives

$$\operatorname{Cov}(Y_{ij}, Y_{ik}) = \operatorname{Cov}(\mathbb{E}[Y_{ij} \mid b_i], \mathbb{E}[Y_{ik} \mid b_i]).$$

Two increasing functions of the same scalar random variable have nonnegative covariance. One way to see this is to let B' be an independent copy of B and use

$$2\operatorname{Cov}(f(B), g(B)) = \mathbb{E}\big[(f(B) - f(B'))(g(B) - g(B'))\big] \geq 0.$$

□

Binary mixed model likelihoods can be numerically delicate. Warning signs include a large gradient at the reported optimum, a nearly singular Hessian, variance estimates on the boundary, or material changes as the quadrature accuracy changes. Zero responses in a treatment by item cell can also create separation or near separation.

Algorithm 11.5 (Binary mixed model validation) After fitting a binary mixed model:

1. Confirm that the grouping variables match the independent sampling units and that repeated observations have not been treated as independent.
2. Tabulate outcomes in each important factor combination to detect zero or all one cells.

3. Inspect convergence, gradients, Hessian diagnostics, and covariance parameters near zero or one.
4. Refit with a more accurate likelihood approximation and compare the estimates and target tests.
5. Simplify an unsupported random effects structure only with a stated scientific and numerical reason.
6. Distinguish conditional odds ratios from marginal population comparisons in the interpretation.

Effect coding is useful for factorial fixed effects because it makes lower order tests correspond to average comparisons when the design and model support that interpretation. A marginal or Type III test of a main effect in the presence of an interaction should still be interpreted as an average over the interacting factor, not as a universal effect.

Example 11.6 (Recursion responses) Participants responded to several linguistic items under conditions A, B, C, and E. The binary response recorded whether all target constructions were produced. There were 2,100 item level records, with 325 from adults and 1,775 from children. Eight binary outcomes were missing, leaving 323 observed adult responses and 1,769 observed child responses. Items were nested within condition, and conditions were repeated within participant.

The adult and child success counts were

Group	No	Yes	Success proportion
Adult	152	171	0.5294
Child	1433	336	0.1899

The large raw difference does not account for condition, item, or repeated responses.

Condition specific success percentages and adult to child odds ratios were

	A	B	C	E
Adult percentage	72.22	71.79	39.74	24.68
Child percentage	26.42	29.81	11.74	6.82
Odds ratio	7.24	5.99	4.96	4.47

The raw group by condition pattern is displayed in Figure 25.

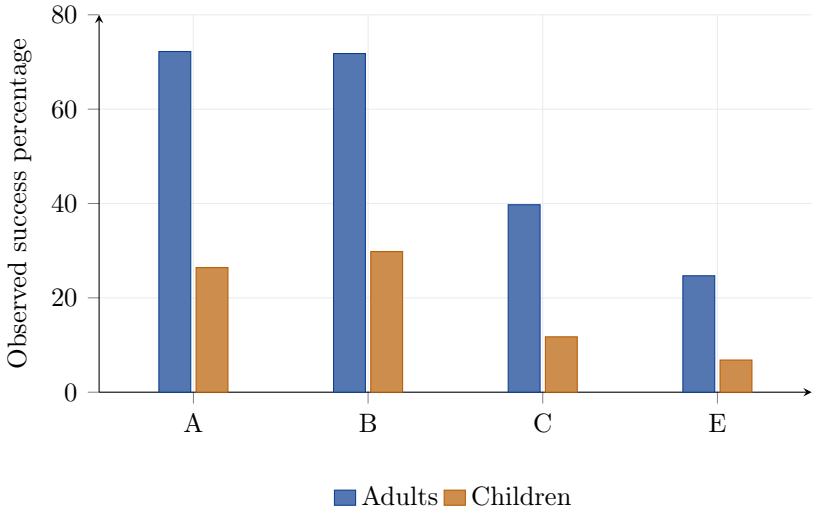


Figure 25

A first model included fixed effects for group, condition, and their interaction, with a participant random intercept:

$$\text{logit}(\pi_{ij}) = \beta_0 + \text{group}_i + \text{condition}_{ij} + (\text{group} \times \text{condition})_{ij} + b_i.$$

Adaptive quadrature with two points gave

Effect	Wald statistic	degrees of freedom	p-value
Group	32.9325	1	9.54×10^{-9}
Condition	132.5334	3	$< 2.2 \times 10^{-16}$
Group by condition	1.2293	3	0.746

The adult and child difference and condition differences were clear, while the data did not support a varying group difference across conditions.

The default fit produced a small convergence warning. Refitting with a faster zero point approximation gave interaction $p = 0.778$, while two point adaptive quadrature gave $p = 0.746$ and satisfied the convergence criterion. Agreement of the substantive conclusion across approximations is reassuring, but the better converged fit is the appropriate basis for reporting.

Items e1 and e6 had no affirmative responses in some groups and were removed from an item level comparison. With item treated as a fixed effect nested within condition, the no intercept model was

$$\text{logit}(\pi_{ij}) = \mu_{\text{item}(ij)} + b_i.$$

A Wald test with 19 degrees of freedom for equality among items within their conditions gave

$$W = 153.88 \qquad \text{and} \qquad p < 2.2 \times 10^{-16}.$$

The corresponding likelihood ratio comparison of condition means with item specific means gave

$$G^2 = 193.0 \qquad \text{and} \qquad p < 2.2 \times 10^{-16}.$$

Both tests show substantial item heterogeneity.

Treating item as fixed estimates a separate mean for every observed item and restricts conclusions to those items. Treating item as random models the observed items as a sample from a larger population. A model with participant and item random intercepts gave the variance estimates

$$\widehat{\sigma}^2_{\text{participant}} = 2.200 \qquad \text{and} \qquad \widehat{\sigma}^2_{\text{item}} = 0.928.$$

The choice between fixed and random item effects is therefore a choice about the target of inference, not merely a device for obtaining a preferred p-value.

For children only, a model with age group by condition fixed effects and random intercepts for participant and item gave

Effect	Wald statistic	degrees of freedom	p-value
Age group	31.6554	2	1.34×10^{-7}
Condition	12.1541	3	0.00687
Age group by condition	3.9003	6	0.690

The estimated marginal means on the conditional log odds scale, averaged over the other fixed factor and evaluated at zero random effects, were

Age group	4-year-olds	5-year-olds	6-year-olds
Mean log odds	−2.803	−2.594	−0.772

and

Condition	A	B	C	E
Mean log odds	-1.600	-1.300	-2.792	-3.284

The fitted age and condition margins are shown in Figure 26.

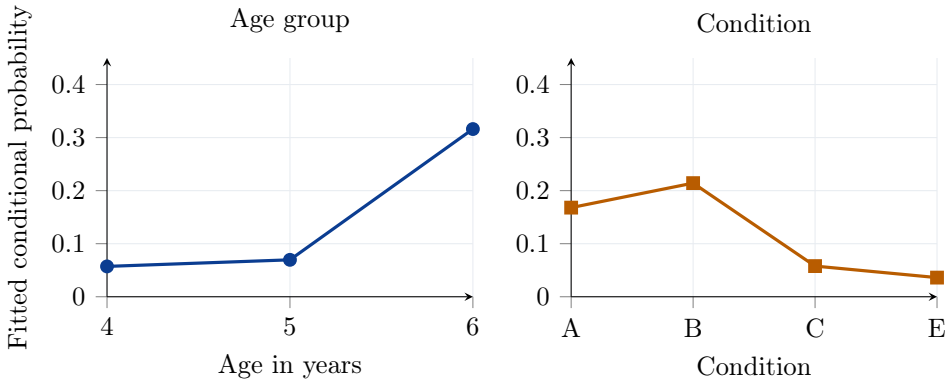


Figure 26

Pairwise comparisons protected across all nine age and condition contrasts used the Bonferroni threshold $0.05/9 = 0.00556$. Six-year-olds responded more often than four- and five-year-olds, while conditions A and B elicited more target responses than C and E. The four- and five-year groups did not differ, nor did A and B or C and E.

The essential model progression is:

```
library(lme4)
library(car)

contrasts(recursion$group) <- contr.sum
contrasts(recursion$condition) <- contr.sum

fit_group <- glmer(all_targets ~ group * condition +
                     (1 | participant),
                   family = binomial, data = recursion,
                   nAGQ = 2)
```

```
Anova(fit_group, type = 3)

fit_items <- glmer(all_targets ~ condition +
                     (1 | participant) + (1 | item),
                   family = binomial, data = children)
fit_age <- glmer(all_targets ~ age_group * condition +
                   (1 | participant) + (1 | item),
                 family = binomial, data = children)
```

The data hierarchy appears directly in the two random intercept terms. Item is nested within condition by design, while participant crosses conditions and items. A model formula must reflect this actual sampling structure.

Remark 11.7 In a generalized linear mixed model, increasing numerical precision is not a substitute for checking identifiability, separation, boundary estimates, or an overly ambitious random effects structure. Computation can approximate a well defined likelihood; it cannot create information absent from the design.

Appendix: Lecture and Tutorial Index