

Estimation & Inference

Aayush Bajaj

August 31, 2026

fitting a regression model produces numbers; inference is what licenses saying anything about them. this page collects the estimation machinery every model on this branch shares — likelihood, score, information, and the iterative algorithms that maximise them — and the two inferential regimes it feeds: *exact* small-sample t and F results in the linear gaussian model, and *asymptotic* wald, score and likelihood-ratio results everywhere else.

*distilled from my math5806
(applied regression analysis)
weeks 2 and 3, which follow
dobson and barnett chapters 4-6*

Contents

1	likelihood, score, information	1
1.1	the score function	1
1.2	newton–raphson and the method of scoring	2
1.3	worked: a weibull scale parameter	2
2	three tests of one hypothesis	3
2.1	worked: one proportion, three verdicts	3
2.2	the deviance connection	4
3	the linear gaussian model: exact inference	4
3.1	least squares is maximum likelihood	4
3.2	reading a coefficient table	4
3.3	the f-test and nested models	5
3.4	anova and ancova are regressions	5
3.5	residual diagnostics	5
4	the two regimes, side by side	6
5	results	6
6	see also	6

1 likelihood, score, information

1.1 the score function

given independent observations with joint log-likelihood $\ell(\theta; \mathbf{y})$, the **score** is its gradient,

$$U(\theta) = \frac{\partial \ell(\theta; \mathbf{y})}{\partial \theta}, \quad (1)$$

and the maximum likelihood estimator solves $U(\hat{\theta}) = 0$. the score is itself a random variable — a fresh sample gives a fresh score curve — and its two moments carry the whole asymptotic theory:

$$\mathbb{E}[U(\theta)] = 0, \quad \text{Var}[U(\theta)] = \mathbb{E}[U(\theta)^2] = -\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \theta^2}\right] = \mathcal{I}(\theta), \quad (2)$$

the **fisher information**. the second equality is the information identity; the geometry is that the information is the expected curvature of the log-likelihood at its peak. a sharply curved likelihood pins the parameter down; a flat one leaves it loose — fatter likelihood, more uncertainty.

*the cramér-rao bound makes
this precise: no unbiased
estimator beats variance
1/I(theta)*

by a central limit argument the score is asymptotically $N(0, \mathcal{I})$, and inverting through the usual taylor expansion gives the master result:

$$\hat{\theta} \text{ is asymptotically } N(\theta, \mathcal{I}(\theta)^{-1}), \quad (3)$$

so standard errors are square roots of the inverse information's diagonal — the numbers every `summary()` prints.

1.2 newton–raphson and the method of scoring

outside charmed cases the equation $U(\theta) = 0$ has no closed form. newton–raphson linearises the score at the current iterate and solves:

$$\theta^{(m)} = \theta^{(m-1)} - \frac{U(\theta^{(m-1)})}{U'(\theta^{(m-1)})}, \quad (4)$$

and the **method of scoring** swaps the observed second derivative $U' = \ell''$ for its expectation $-\mathcal{I}$, which is smoother, often cheaper, and hands back the estimator's covariance as a by-product. for a glm with canonical link the two coincide exactly; in general they differ but share the fixed point. the multivariate versions replace division by matrix solves, and fisher scoring on a glm rearranges into iteratively reweighted least squares — the derivation lives on the glm page.

1.3 worked: a weibull scale parameter

the course's week-2 exercise estimates the scale θ of a weibull with known shape $\lambda = 2$ from $n = 49$ failure lifetimes. the score and its derivative are

$$U(\theta) = -\frac{n\lambda}{\theta} + \frac{\lambda \sum_i y_i^\lambda}{\theta^{\lambda+1}}, \quad U'(\theta) = \frac{n\lambda}{\theta^2} - \frac{\lambda(\lambda+1) \sum_i y_i^\lambda}{\theta^{\lambda+2}}, \quad (5)$$

and this model happens to admit a closed form to check against: $U(\hat{\theta}) = 0$ gives $\hat{\theta} = (\frac{1}{n} \sum_i y_i^\lambda)^{1/\lambda}$, the root-mean-square of the data when $\lambda = 2$.

```

1 newton.weibull <- function(y, theta0, lambda = 2, tol = 1e-10, maxit = 20) {
2   n <- length(y); theta <- theta0
3   for (i in 1:maxit) {
4     s <- -n * lambda / theta + lambda * sum(y^lambda) / theta^(lambda + 1)
5     h <- n * lambda / theta^2 - lambda * (lambda + 1) * sum(y^lambda) /
6       theta^(lambda + 2)
7     theta.new <- theta - s / h
8     if (abs(theta.new - theta) < tol) return(list(theta = theta.new, iter = i))
9     theta <- theta.new
10  }
```

```

10 list(theta = theta, iter = maxit)
11 }

```

```

starting from theta0 = mean(y):
Iter 1 : theta = 9633.7774
Iter 2 : theta = 9875.8983
Iter 3 : theta = 9892.1100
Iter 4 : theta = 9892.1768
Iter 5 : theta = 9892.1768      <- converged
closed form sqrt(mean(y^2)):    9892.1768 (difference 3.6e-12)
maximised log-likelihood:      -480.85

```

five iterations from the sample mean, agreeing with the closed form to machine precision — the quadratic convergence near the root is characteristic of newton methods.

2 three tests of one hypothesis

to test $H_0 : \theta = \theta_0$, three statistics interrogate the same log-likelihood curve at different landmarks:

- **wald**: how far is $\hat{\theta}$ from θ_0 , in standard-error units? $W = (\hat{\theta} - \theta_0)^2 \mathcal{I}(\hat{\theta})$;
- **score**: how steep is the log-likelihood at θ_0 ? (a true value should sit near a stationary point) $S = U(\theta_0)^2 / \mathcal{I}(\theta_0)$;
- **likelihood ratio**: how much height is lost by descending from $\hat{\theta}$ to θ_0 ? $C = 2[\ell(\hat{\theta}) - \ell(\theta_0)]$.

all three are asymptotically χ_1^2 under H_0 (with the obvious degrees-of-freedom generalisation for vector hypotheses), and they are first-order equivalent — yet they can disagree in finite samples, and the disagreement is informative.

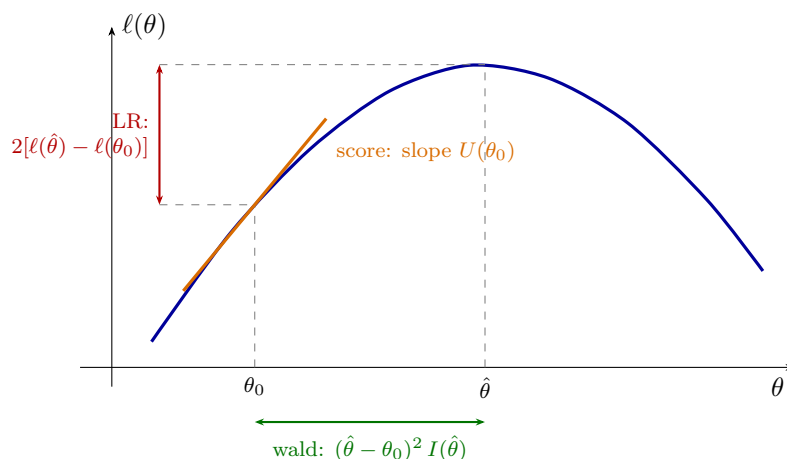


Figure 1: one curve, three tests. the likelihood ratio reads the vertical drop, the wald statistic the horizontal gap (scaled by curvature at the peak), the score statistic the slope at the null. only the score test never needs the mle.

2.1 worked: one proportion, three verdicts

the course's binomial exercise: $y = 460$ successes in $n = 828$ trials, test $H_0 : \pi = 0.5$. the mle is $\hat{\pi} = 0.5556$, and for a binomial proportion the three statistics specialise

to

$$S = \frac{(\hat{\pi} - \pi_0)^2}{\pi_0(1 - \pi_0)/n} = 10.22, \quad W = \frac{(\hat{\pi} - \pi_0)^2}{\hat{\pi}(1 - \hat{\pi})/n} = 10.35, \quad C = 2[\ell(\hat{\pi}) - \ell(\pi_0)] = 10.17, \quad (6)$$

with p-values 0.00139, 0.00129 and 0.00137 against χ_1^2 — the same verdict three ways, differing only in whether the variance is evaluated at the null (score), at the mle (wald), or integrated along the path (lr).

note the score statistic is exactly the square of the classical one-proportion z-test

when they *do* disagree, trust order is usually lr, then score, then wald: the wald statistic is computed solely from local curvature at the mle and degrades badly when the likelihood is asymmetric — pathologically so near separation in logistic regression, where an enormous effect can shrink its own wald statistic to non-significance (the hauck–donner effect). the likelihood ratio costs a second fit but is invariant to reparametrisation.

2.2 the deviance connection

for nested regression models the likelihood-ratio statistic wears a different name: the difference of deviances. $\Delta D = D_0 - D_1 = 2[\ell_1 - \ell_0] \sim \chi^2(p_1 - p_0)$ — every “does this covariate earn its keep” test on this branch is this statistic.

3 the linear gaussian model: exact inference

3.1 least squares is maximum likelihood

for $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ with $\varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$, the log-likelihood is a decreasing function of $\|\mathbf{y} - \mathbf{X}\beta\|^2$, so the mle *is* the least-squares solution

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}, \quad (7)$$

one shot of linear algebra, no iteration. this is the only member of the glm family where the score equations are linear; fisher scoring run on it converges in a single lap (the working response is $\mathbf{y}$ itself), which is why `glm()` with the default gaussian family reproduces `lm()` to the digit.

in practice lm() uses a qr decomposition of X rather than forming and inverting X-transpose-X — same answer, better conditioning

what makes this model special is that inference here is **exact at every sample size**, not asymptotic:

$$\hat{\beta} \sim N\left(\beta, \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}\right), \quad \frac{(N-p)s^2}{\sigma^2} \sim \chi_{N-p}^2, \quad \frac{\hat{\beta}_j - \beta_j}{\text{se}(\hat{\beta}_j)} \sim t_{N-p}, \quad (8)$$

with $s^2 = \|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2 / (N-p)$, and $\hat{\beta}$ independent of s^2 . the gauss–markov theorem adds a distribution-free guarantee: among linear unbiased estimators, least squares has minimal variance — normality buys exactness, not optimality.

3.2 reading a coefficient table

the course’s tutorial regression (points-per-minute of esports players on height, assists-per-minute, minutes-per-game and age) is a drill in reading `summary(lm(...))`

honestly. each row's $t = \hat{\beta}_j / \text{se}(\hat{\beta}_j)$ tests $H_0 : \beta_j = 0$ given everything else in the model:

term	estimate	std. error	t	p	verdict
intercept	0.2207	0.3160	0.70	0.487	not significant
height	0.000531	0.001566	0.34	0.735	not significant
apm	-0.5491	0.1847	-2.97	0.0038	significant at 1%
mpg	0.007778	0.001051	7.40	6.6e-11	highly significant
age	-0.000522	0.002626	-0.20	0.843	not significant

the discipline: a fat p-value says the coefficient is *indistinguishable from zero at this sample size*, not that it is zero; an intercept extrapolated to covariate values no player possesses (height zero) is not meaningfully interpretable however small its p-value; and a sensible sign (playmakers trade assists against scoring) is corroboration, not proof.

3.3 the f-test and nested models

dropping q coefficients at once is a nested-model comparison, and the exact-sampling analogue of the deviance difference:

$$F = \frac{(\text{RSS}_0 - \text{RSS}_1)/q}{\text{RSS}_1/(N - p)} \sim F_{q, N-p} \quad \text{under } H_0, \quad (9)$$

which for a single coefficient squares the t -statistic. the **F-statistic** line at the foot of a summary is this test against the intercept-only model.

3.4 anova and ancova are regressions

categorical covariates enter through dummy coding, at which point classical analysis of variance is a linear gaussian model whose F -tests compare factor levels. the course's ancova example — achievement scores on training method (three levels) plus initial aptitude — fits $\mathbf{y} \sim \text{method} + \mathbf{x}$ and reads off both stories at once: aptitude matters ($\hat{\beta}_x = 0.74$, $t = 5.23$), and both alternative methods beat the baseline (+2.19 and +1.86 points) with the aptitude effect held common across groups; overall $F = 29.43$ on (3, 17) degrees of freedom. running the same formula through `glm()` prints identical estimates with the bookkeeping renamed: residual deviance 10.30 where `lm()` said residual sum of squares, a dispersion estimate 0.606 where it said residual variance.

3.5 residual diagnostics

the assumptions bought all that exactness, so check them: plot **standardised residuals** $r_i = (y_i - \hat{y}_i) / (s\sqrt{1 - h_{ii}})$ against fitted values (curvature betrays a wrong mean structure, a funnel betrays heteroscedasticity), against each covariate, and as a normal q-q plot. the leverages h_{ii} come from the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$; points with large leverage *and* large residual are the ones that move the fit.

confounding warning from the tutorial: unbalanced factors make the anova table order-dependent — fit both orders or use type-ii sums of squares before pronouncing on a factor

4 the two regimes, side by side

question	linear gaussian model	any other glm
estimator	closed form, one solve	irls, iterated to convergence
coefficient test	exact t_{N-p}	asymptotic wald z
nested comparison	exact F	asymptotic $\Delta D \sim \chi^2$
fit summary	R^2 , residual variance	deviance, aic
failure mode	exact collinearity only	separation, non-convergence

5 results

definition 5.1 (score). $U(\theta) = \partial \ell(\theta; \mathbf{y}) / \partial \theta$; the mle solves $U(\hat{\theta}) = 0$.

definition 5.2 (fisher information). $\mathcal{I}(\theta) = \text{Var}[U(\theta)] = \mathbb{E}[U(\theta)^2]$ — the expected curvature of the log-likelihood.

theorem 5.3 (moments of the score). *under regularity, $\mathbb{E}[U(\theta)] = 0$ and $\text{Var}[U(\theta)] = -\mathbb{E}[\ell''(\theta)] = \mathcal{I}(\theta)$.*

theorem 5.4 (asymptotic normality of the mle). *$\hat{\theta}$ is consistent and asymptotically $N(\theta, \mathcal{I}(\theta)^{-1})$; equivalently $(\hat{\theta} - \theta)^2 \mathcal{I}(\theta) \rightarrow \chi_1^2$ in distribution.*

definition 5.5 (wald, score and likelihood-ratio statistics). for $H_0 : \theta = \theta_0$: $W = (\hat{\theta} - \theta_0)^2 \mathcal{I}(\hat{\theta})$, $S = U(\theta_0)^2 / \mathcal{I}(\theta_0)$, $C = 2[\ell(\hat{\theta}) - \ell(\theta_0)]$; each is asymptotically χ_1^2 under H_0 , and they are first-order equivalent.

theorem 5.6 (wilks). *for nested models with $p_0 < p_1$ parameters, $2[\ell_1(\hat{\theta}_1) - \ell_0(\hat{\theta}_0)] \rightarrow \chi^2(p_1 - p_0)$ under the smaller model.*

theorem 5.7 (distribution theory of least squares). *with gaussian errors: $\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1})$, $(N - p)s^2 / \sigma^2 \sim \chi_{N-p}^2$, the two are independent, and $(\hat{\beta}_j - \beta_j) / \text{se}(\hat{\beta}_j) \sim t_{N-p}$ exactly.*

theorem 5.8 (gauss–markov). *if $\mathbb{E}(\epsilon) = \mathbf{0}$ and $\text{Var}(\epsilon) = \sigma^2 \mathbf{I}$ (no normality needed), the least-squares estimator has minimal variance among linear unbiased estimators of β .*

theorem 5.9 (nested f-test). *dropping q coefficients: $F = \frac{(\text{RSS}_0 - \text{RSS}_1)/q}{\text{RSS}_1/(N-p)} \sim F_{q, N-p}$ exactly under H_0 and normal errors; for $q = 1$, $F = t^2$.*

remark 5.10 (cramér–rao). any unbiased estimator T of θ satisfies $\text{Var}(T) \geq \mathcal{I}(\theta)^{-1}$; the mle attains the bound asymptotically.

6 see also

- generalised linear models — where fisher scoring becomes irls, and deviance is defined
- ordinary least squares — the estimator this page’s exact theory attaches to
- logistic regression — the asymptotic regime in daily use, wald traps included
- dobson & barnett, *an introduction to generalized linear models* — chapters 4-6 are this page’s spine