Inference in Linear Models

Statistics · Module 8 — Multivariate Linear Models

Node 8.5

Position in Knowledge Graph

System: Statistics. Structural Domain: Module 8 — Multivariate and Linear Models. Node 8.5 — Inference in Linear Models.

This node develops the exact finite-sample distributional theory for inference in the normal linear model: t-tests, F-tests, confidence intervals, and prediction intervals. These results depend on the geometry of OLS (Node 8.3), Gauss-Markov (Node 8.4), the MVN (Node 8.2), and the general testing and estimation frameworks (Nodes 5.1, 4.1). This node enables the extension to GLMs (Node 8.6).

Formal Definition

Let $Y = X\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ and $\operatorname{rank}(X) = p$. Under normality:

$$\hat\beta \sim \mathcal{N}(\beta,\ \sigma^2(X^\top X)^{-1}), \qquad \frac{\text{RSS}}{\sigma^2} \sim \chi^2_{n-p}, \qquad \hat\beta \perp\!\!\!\perp \hat\sigma^2.$$

Intuition

Under Gaussian errors, the OLS estimator is exactly normal (not just approximately). The residual sum of squares follows a chi-squared distribution because it is a quadratic form in normal variables via the idempotent matrix $I - H$. The independence of $\hat\beta$ and $\hat\sigma^2$ — which follows from the orthogonality of H and $I - H$ — allows us to form exact pivotal quantities. The ratio of a normal to a chi-squared produces a t-distribution; the ratio of two chi-squareds produces an F-distribution. These are exact for any sample size, not asymptotic approximations.

Structural Role

Under normality, the linear model admits exact finite-sample inference. The t-test for single coefficients and the F-test for general linear restrictions are both instances of likelihood ratio testing (Node 5.4) specialized to the Gaussian linear model, where the chi-squared approximation happens to be exact. Cochran's theorem (Node 8.2) provides the distributional foundation, connecting the rank of projection matrices to chi-squared degrees of freedom.

Mathematical Development

Distribution of $\hat\beta$. Since $\hat\beta = (X^\top X)^{-1}X^\top Y$ is a linear function of $Y \sim \mathcal{N}(X\beta, \sigma^2 I)$, it is exactly normal: $\hat\beta \sim \mathcal{N}(\beta, \sigma^2(X^\top X)^{-1})$.

Distribution of RSS. Write $\hat{\varepsilon} = (I-H)Y = (I-H)\varepsilon$ (since $(I-H)X\beta = 0$). Then $\hat{\varepsilon}/\sigma \sim \mathcal{N}(0, I-H)$ and

$$\frac{\text{RSS}}{\sigma^2} = \frac{\varepsilon^\top (I-H)\varepsilon}{\sigma^2}.$$

Since $I-H$ is symmetric idempotent with rank $n-p$, this is a chi-squared with $n-p$ degrees of freedom: $\text{RSS}/\sigma^2 \sim \chi^2_{n-p}$.

Independence of $\hat{\beta}$ and RSS. $\hat{\beta}$ depends on Y through HY, and RSS depends on Y through $(I-H)Y$. Since $H(I-H) = 0$ and Y is normal, $HY \perp\!\!\!\perp (I-H)Y$. Therefore $\hat{\beta} \perp\!\!\!\perp \hat{\sigma}^2$.

Cochran's theorem application. The decomposition $\varepsilon = H\varepsilon + (I-H)\varepsilon$ corresponds to projection onto $\mathcal{C}(X)$ and $\mathcal{C}(X)^\perp$. By Cochran's theorem, $\varepsilon^\top H\varepsilon/\sigma^2 \sim \chi^2_p$ and $\varepsilon^\top(I-H)\varepsilon/\sigma^2 \sim \chi^2_{n-p}$, independently. Under the alternative $\mathbb{E}[Y] = X\beta \neq 0$, the first quadratic form has a noncentral chi-squared distribution, while RSS remains $\sigma^2\chi^2_{n-p}$.

t-test for a single coefficient. To test $H_0 : \beta_j = \beta_j^0$:

$$T_j = \frac{\hat{\beta}_j - \beta_j^0}{\hat{\sigma}\sqrt{(X^\top X)^{-1}_{jj}}} \sim t_{n-p} \quad \text{under } H_0.$$

Derivation. $\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2 (X^\top X)^{-1}_{jj})$, so $(\hat{\beta}_j - \beta_j)/(\sigma\sqrt{(X^\top X)^{-1}_{jj}}) \sim \mathcal{N}(0,1)$. Dividing by $\sqrt{\hat{\sigma}^2/\sigma^2}$, where $(n-p)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}$ independently, gives a t_{n-p} distribution.

Confidence interval for β_j:

$$\hat{\beta}_j \pm t_{n-p,\,\alpha/2} \cdot \hat{\sigma}\sqrt{(X^\top X)^{-1}_{jj}}.$$

F-test for general linear hypotheses. For $H_0 : R\beta = r$ with $R \in \mathbb{R}^{q\times p}$ of rank q:

$$F = \frac{(R\hat{\beta} - r)^\top [R(X^\top X)^{-1}R^\top]^{-1}(R\hat{\beta} - r)/q}{\hat{\sigma}^2} \sim F_{q,\,n-p} \quad \text{under } H_0.$$

Derivation. Under H_0, $R\hat{\beta} - r \sim \mathcal{N}(0, \sigma^2 R(X^\top X)^{-1}R^\top)$. The numerator quadratic form divided by σ^2 is χ^2_q. Dividing by the independent $\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}/(n-p)$ gives $F_{q,n-p}$.

Nested model comparison. To compare the full model $Y = X\beta + \varepsilon$ with a reduced model $Y = X_0\beta_0 + \varepsilon$ where $\mathcal{C}(X_0) \subseteq \mathcal{C}(X)$:

$$F = \frac{(\text{RSS}_0 - \text{RSS})/(p - p_0)}{\text{RSS}/(n-p)} \sim F_{p-p_0,\,n-p} \quad \text{under the reduced model.}$$

This is the ANOVA F-test.

Confidence ellipsoid for β:

$$\left\{ b \in \mathbb{R}^p : (b - \hat{\beta})^\top (X^\top X)(b - \hat{\beta}) \le p\hat{\sigma}^2 F_{p,\,n-p,\,\alpha} \right\}.$$

This is the inversion of the F-test for $H_0 : \beta = b$.

Prediction interval. For a new observation $Y_{\text{new}} = x_0^\top \beta + \varepsilon_{\text{new}}$:

$$\hat{Y}_{\text{new}} = x_0^\top \hat{\beta}, \quad \mathrm{Var}(Y_{\text{new}} - \hat{Y}_{\text{new}}) = \sigma^2(1 + x_0^\top (X^\top X)^{-1} x_0).$$

The $(1-\alpha)$ prediction interval is

$$x_0^\top \hat{\beta} \pm t_{n-p,\,\alpha/2} \cdot \hat{\sigma}\sqrt{1 + x_0^\top (X^\top X)^{-1} x_0}.$$

The extra σ^2 accounts for the irreducible noise in the new observation.

Coefficient of determination. $R^2 = 1 - \mathrm{RSS}/\mathrm{TSS}$ where $\mathrm{TSS} = \|Y - \bar{Y}\mathbf{1}\|^2$. The adjusted $R^2 = 1 - (1 - R^2)(n-1)/(n-p)$ penalizes for additional predictors. Under the null model (no predictors beyond the intercept), $R^2 \cdot (n-p)/((1-R^2)(p-1)) \sim F_{p-1,n-p}$.

Simultaneous confidence intervals. The Bonferroni correction produces simultaneous $(1-\alpha)$ intervals for all p coefficients by using $t_{n-p,\alpha/(2p)}$ in place of $t_{n-p,\alpha/2}$. The Scheffe method uses the F-distribution: $\hat{\beta}_j \pm \sqrt{pF_{p,n-p,\alpha}} \cdot \hat{\sigma}\sqrt{(X^\top X)^{-1}_{jj}}$, which provides coverage for all linear combinations $a^\top \beta$ simultaneously.

Power of the F-test. Under the alternative $R\beta \neq r$, the F-statistic has a noncentral F distribution $F_{q,n-p,\delta}$ with noncentrality parameter

$$\delta = \frac{(R\beta - r)^\top [R(X^\top X)^{-1} R^\top]^{-1} (R\beta - r)}{\sigma^2}.$$

The power increases with δ, which depends on the effect size, the sample size (through $X^\top X$), and the error variance. Power analysis uses this noncentrality parameter to determine the required sample size.

Residual diagnostics. Under the model, the standardized residuals $r_i = \hat{\varepsilon}_i/(\hat{\sigma}\sqrt{1 - h_{ii}})$ are approximately t_{n-p-1}. The externally studentized residuals $t_i = \hat{\varepsilon}_i/(\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}})$, where $\hat{\sigma}_{(i)}$ is estimated from the data excluding observation i, follow t_{n-p-1} exactly. Cook's distance $D_i = r_i^2 h_{ii}/(p(1 - h_{ii}))$ combines leverage and residual magnitude to measure influence.

Analysis of variance (ANOVA) decomposition. The total sum of squares decomposes as

$$\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS}, \quad \text{where } \mathrm{ESS} = \|\hat{Y} - \bar{Y}\mathbf{1}\|^2,\ \mathrm{RSS} = \|Y - \hat{Y}\|^2.$$

Under $H_0 : \beta_2 = \cdots = \beta_p = 0$ (only intercept), $\mathrm{ESS}/\sigma^2 \sim \chi^2_{p-1}$ and is independent of $\mathrm{RSS}/\sigma^2 \sim \chi^2_{n-p}$, yielding the overall F-test $F = (\mathrm{ESS}/(p-1))/(\mathrm{RSS}/(n-p)) \sim F_{p-1,n-p}$.

Connection to likelihood ratio. Under normality, the LR statistic for $H_0 : R\beta = r$ is

$$\Lambda = n \log(\mathrm{RSS}_0/\mathrm{RSS}),$$

where RSS_0 is the residual sum of squares under the restricted model. The F-statistic is a monotone transformation: $F = ((e^{\Lambda/n} - 1)(n-p))/q$. Thus the F-test is equivalent to the likelihood ratio test, with the F-distribution providing the exact (not asymptotic) null distribution.

Multiple testing. When testing many coefficients simultaneously, the individual t-test p-values require adjustment. The Bonferroni correction controls the family-wise error rate by using α/m as the per-test level. The Benjamini-Hochberg procedure controls the false discovery rate, which is less conservative and more appropriate for exploratory analysis with many predictors.

Worked Examples

Example 1. In simple linear regression with $n = 30$ observations, $\hat{\beta}_1 = 2.5$, $\mathrm{SE}(\hat{\beta}_1) = 0.8$. The t-statistic is $T = 2.5/0.8 = 3.125$ on 28 df. The p-value is $2\mathbb{P}(t_{28} > 3.125) \approx 0.004$. The 95% CI is $2.5 \pm 2.048 \times 0.8 = (0.86, 4.14)$.

Example 2. Test $H_0 : \beta_2 = \beta_3 = 0$ in a model with $p = 5$ predictors and $n = 50$. Here $q = 2$, $R = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$, $r = 0$. If $\mathrm{RSS}_{\text{full}} = 120$ and $\mathrm{RSS}_{\text{reduced}} = 160$, then $F = ((160 - 120)/2)/(120/45) = 20/2.667 = 7.5$ on $(2, 45)$ df. The critical value $F_{2,45,0.05} \approx 3.20$, so we reject H_0.

Cross-Links

AI. The F-statistic for nested model comparison is the basis for feature selection in linear models. In deep learning, the degrees-of-freedom concept generalizes via effective degrees of freedom $\mathrm{tr}(H_\lambda)$ for regularized estimators, connecting to generalization bounds.

Economics. The t-test and F-test are the primary tools in empirical economics. The Chow test for structural breaks is an F-test comparing pooled and split regressions. Robust inference under heteroskedasticity replaces the exact F with asymptotic chi-squared/F using the sandwich covariance (Node 7.5).

Linear Algebra. The distributional results rest on the spectral properties of idempotent matrices: a symmetric idempotent matrix of rank r has r eigenvalues equal to 1 and the rest zero. Quadratic forms in standard normals with such matrices give chi-squared distributions.

Quantum Mechanics. Hypothesis testing in quantum state discrimination uses analogous chi-squared statistics for goodness-of-fit testing of quantum states, with degrees of freedom determined by the dimension of the state space minus the number of estimated parameters.

Structural Compression

Invariant. Under Gaussian errors, all finite-sample inference in the linear model is determined by $\hat{\beta}$, $\hat{\sigma}^2$, and the hat matrix H; the t and F distributions arise from the chi-squared distribution of RSS/σ^2 and the independence of $\hat{\beta}$ and $\hat{\sigma}^2$.

Mechanism. Orthogonal decomposition $Y = HY + (I - H)Y$ separates signal from noise; independence follows from orthogonality under normality (Cochran's theorem); quadratic forms in standard normals yield chi-squared distributions; ratios of independent chi-squareds yield F; a standard normal divided by the square root of an independent chi-squared yields t.

Cross-System Echo. The F-statistic is the likelihood ratio statistic for the Gaussian linear model; the exact F-distribution replaces the asymptotic χ^2 approximation that holds in general parametric models. In ANOVA, the same F-test compares group means, decomposing total variation into between-group and within-group components.

Exercises

1. Derive the distribution of the t-statistic T_j from the joint distribution of $\hat{\beta}_j$ and $\hat{\sigma}^2$, verifying the t_{n-p} result.
2. Show that the F-test for $H_0 : \beta_j = 0$ is equivalent to T_j^2 by verifying $F_{1,n-p} = t_{n-p}^2$.
3. Prove that $\hat{\beta} \perp\!\!\!\perp \hat{\sigma}^2$ using the fact that HY and $(I - H)Y$ are functions of Y through orthogonal projections and Y is normal.

4. For the prediction interval, explain why the variance $\sigma^2(1 + x_0^\top (X^\top X)^{-1} x_0)$ has two components and identify which component is reducible with more data.
5. In a one-way ANOVA with $k = 3$ groups of size $n_j = 10$, derive the F-statistic as a ratio of between-group and within-group mean squares, and state the null distribution.
6. Show that $R^2 = 0$ if and only if $\hat{\beta}_j = 0$ for all $j \geq 2$ (assuming an intercept is included), and explain why R^2 always increases when adding predictors.
7. Argue, structurally, why the exact distributional results of this node (t-tests, F-tests) require Gaussian errors while the Gauss-Markov theorem (Node 8.4) does not, identifying exactly which properties of the normal distribution are needed and which are not.

Statistics · Module 8 — Multivariate Linear Models · Node 8.5
Concepts: hypothesis-testing, variance-and-covariance, vector-space
Prerequisites: 8.3, 8.4, 8.2, 5.1, 4.1
Enables: 8.6

hbar.university — own.your.cognition