# Linear Models Lecture 12: Hypothesis Testing

Robert Gulotty

University of Chicago

February 25, 2026

# Where We Are

**Last lecture:** We derived the Wald statistic

$$W = (\boldsymbol{R}\hat{\beta} - \boldsymbol{r})' \left[\boldsymbol{R}\hat{\boldsymbol{V}}\boldsymbol{R}'\right]^{-1} (\boldsymbol{R}\hat{\beta} - \boldsymbol{r}) \xrightarrow{d} \chi^2_q$$

and showed how to construct robust tests and confidence intervals.

**Today:** We go deeper into the *practice* of hypothesis testing.

- The F test: a criterion-based alternative
- Score (LM) tests: testing from the restricted model
- The trinity of classical tests
- Test inversion for confidence regions
- Multiple testing and Bonferroni corrections
- Power: what determines whether you can detect an effect?
- Hansen's practical advice for applied work

# Constrained Least Squares: Setup

Many tests today compare an **unrestricted** model to a **restricted** one. We need to know how to estimate under restrictions.

**Problem:** Minimize the sum of squared errors subject to $q$ linear constraints:

$$\tilde{\beta}_{\text{CLS}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (Y_i - \boldsymbol{X}_i'\beta)^2 \quad \text{subject to} \quad \boldsymbol{R}'\beta = \boldsymbol{r}$$

where $\boldsymbol{R}$ is $k \times q$ and $\boldsymbol{r}$ is $q \times 1$.

**Examples:**

- $\beta_3 = 0$: set $\boldsymbol{R}' = (0, 0, 1, 0, \ldots)$, $\boldsymbol{r} = 0$ (exclusion restriction)
- $\beta_2 = \beta_3$: set $\boldsymbol{R}' = (0, 1, -1, 0, \ldots)$, $\boldsymbol{r} = 0$ (equality restriction)
- $\beta_2 + \beta_3 = 1$: set $\boldsymbol{R}' = (0, 1, 1, 0, \ldots)$, $\boldsymbol{r} = 1$ (adding-up constraint)

# The CLS Estimator

Solve via Lagrange multipliers. The solution is:

$$\boxed{\tilde{\beta}_{\text{CLS}} = \hat{\beta}_{\text{OLS}} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}\left[\boldsymbol{R}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}\right]^{-1}(\boldsymbol{R}'\hat{\beta}_{\text{OLS}} - \boldsymbol{r})}$$

**Reading the formula:**

- Start from unrestricted OLS $\hat{\beta}_{\text{OLS}}$, subtract a correction toward the constraint
- Correction is proportional to $(\boldsymbol{R}'\hat{\beta}_{\text{OLS}} - \boldsymbol{r})$: how far OLS violates $H_0$
- If OLS already satisfies the restriction, $\tilde{\beta}_{\text{CLS}} = \hat{\beta}_{\text{OLS}}$

The restricted residuals and variance estimate are:

$$\tilde{e}_i = Y_i - \boldsymbol{X}_i'\tilde{\beta}_{\text{CLS}}, \qquad \tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\tilde{e}_i^2$$

Since the restriction constrains the parameter space: $\tilde{\sigma}^2 \geq \hat{\sigma}^2$ always.

## CLS in Practice

**Simple case:** If $H_0$ sets some coefficients to zero, CLS just means dropping those regressors.

```
# Unrestricted
mod_U <- lm(lwage ~ education + exper + I(exper^2) +
              female + union, data = wages)
# Restricted (H0: beta_female = beta_union = 0)
mod_R <- lm(lwage ~ education + exper + I(exper^2),
              data = wages)
```

**General case:** For arbitrary linear restrictions (e.g., $\beta_2 = \beta_3$), reparameterize or use the closed-form formula.

The key outputs from CLS are $\text{SSE}_R = \sum \tilde{e}_i^2$ and $\tilde{\sigma}^2$, which we use to build the F and Score statistics.

# The F Test: Idea

The F test asks: **how much does the fit worsen when we impose the null?**

- Run the **unrestricted** regression: get $\text{SSE}_U = \sum(Y_i - \boldsymbol{X}_i'\hat{\beta}_{\text{OLS}})^2$
- Run the **restricted** regression (imposing $H_0$): get $\text{SSE}_R = \sum(Y_i - \boldsymbol{X}_i'\tilde{\beta}_{\text{CLS}})^2$

- Since $H_0$ constrains the parameter space, $\text{SSE}_R \geq \text{SSE}_U$ always.
- If the fit barely worsens $\Rightarrow$ the restriction is consistent with the data.
- If the fit worsens a lot $\Rightarrow$ the restriction is costly $\Rightarrow$ evidence against $H_0$.

# The F Statistic

Testing $H_0 \colon \boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{r}$ with $q$ restrictions:

$$\boxed{F = \frac{(\text{SSE}_R - \text{SSE}_U)/q}{\text{SSE}_U/(n-k)} = \frac{(\tilde{\sigma}^2 - \hat{\sigma}^2)/q}{\hat{\sigma}^2/(n-k)}}$$

- **Numerator**: average increase in residual variance per restriction
- **Denominator**: unrestricted estimate of error variance ($s^2$)
- Reject $H_0$ when $F$ is large

**Key relationship:** For linear hypotheses, $F = W^0/q$ where $W^0$ is the homoskedastic Wald statistic. Under homoskedasticity and normality, $F \sim F_{q,\,n-k}$ exactly. Asymptotically, $F \xrightarrow{d} \chi^2_q/q$.

# F Test: When and Why

**Advantages:**

- Directly computable from standard output (just need SSE from two regressions)
- Exact distribution under normality + homoskedasticity
- Slightly more conservative than $\chi^2$ critical values (good in small samples)

**Limitations:**

- Requires homoskedasticity for the $F_{q,n-k}$ distribution to be valid
- Under heteroskedasticity, use the robust Wald test from Lecture 11 instead

**Hansen's warning:** Many packages automatically report an "F-statistic" testing that all slopes are zero. With modern sample sizes this is nearly always significant. **There is no reason to report this F statistic.**

# F Test in R

```r
mod_U <- lm(lwage ~ education + exper + I(exper^2) +
              female + union, data = wages)
mod_R <- lm(lwage ~ education + exper + I(exper^2),
              data = wages)

# Manual F test
SSE_U <- sum(resid(mod_U)^2)
SSE_R <- sum(resid(mod_R)^2)
q <- 2; n <- nobs(mod_U); k <- length(coef(mod_U))
F_stat <- ((SSE_R - SSE_U)/q) / (SSE_U/(n - k))
p_val  <- 1 - pf(F_stat, q, n - k)

# Or simply:
anova(mod_R, mod_U)
```

# Example: Joint Significance

Wage regression with "Male Union Member" and "Female Union Member" indicators.

**Test:** $H_0$: Union membership has no effect on wages (both coefficients $= 0$).

- $W = 23$ (Wald), so $F = 23/2 = 11.5$, $p < 0.001$. **Reject.**

**Interpretation:** Rejecting the joint null means *at least one* coefficient is nonzero. It does not mean both are. Always examine both the joint test and individual t-statistics for a complete picture.

# Score Tests: The Idea

The Wald test starts from the **unrestricted** estimate and asks: is $\hat{\boldsymbol{\theta}}$ far from $\boldsymbol{\theta}_0$?

The **Score test** (also called **Lagrange Multiplier test**) starts from the **restricted** estimate and asks: does the objective function want to move away from the restriction?

- Compute the restricted estimator $\tilde{\boldsymbol{\beta}}$ (constrained least squares)
- Evaluate the *gradient* (score) of the objective function at $\tilde{\boldsymbol{\beta}}$
- If $H_0$ is true, the score should be near zero (we're near the optimum)
- If $H_0$ is false, the score should be large (the restriction is pulling us away)

# Score Statistic for Linear Restrictions

For $H_0: \boldsymbol{R}'\beta = \boldsymbol{c}$ in the normal regression model:

$$S = \frac{(\boldsymbol{R}'\hat{\beta} - \boldsymbol{c})'[\boldsymbol{R}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}]^{-1}(\boldsymbol{R}'\hat{\beta} - \boldsymbol{c})}{\tilde{\sigma}^2}$$

- Identical to $W^0$ except $s^2$ replaced by $\tilde{\sigma}^2$ (restricted variance).
- Under $H_0$: $S \xrightarrow{d} \chi^2_q$

**Key result:** $S$ is a monotone transformation of $F$: $S = n\left(1 - \frac{1}{1+qF/(n-k)}\right)$, so the Score test and the F test always give the same accept/reject decision.

# When Are Score Tests Useful?

For linear regression with linear restrictions, the Score test offers nothing new over $F$ or Wald.

**But** in more complex settings, Score tests have a major advantage:

- They only require estimation under $H_0$ (the restricted model)
- The "unrestricted" model may not even be a simple OLS regression

**Example:** Testing for heteroskedasticity (Breusch–Pagan). The null is homoskedastic OLS—easy. The unrestricted model allows $\text{Var}(e_i \mid X_i)$ to vary with $X$, which requires modeling the variance function. The Score test only needs the OLS residuals.

**Same idea applies to:**

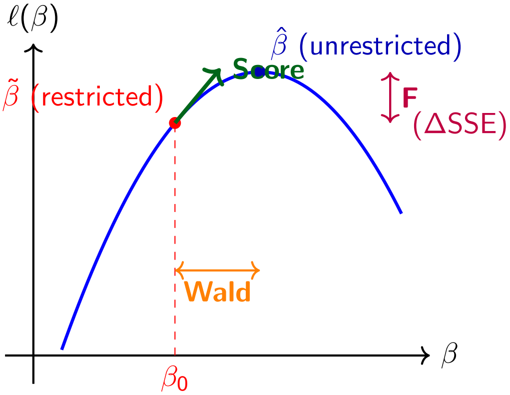
- Testing for serial correlation (Breusch–Godfrey): null is OLS with iid errors
- Testing for omitted nonlinearities: null is a simple linear specification
- Any setting where $H_0$ gives you a clean model but $H_1$ is messy

# Three Ways to Test the Same Hypothesis

| | Wald | Score (LM) | F / LR-like |
|---|---|---|---|
| **Estimates from** | Unrestricted | Restricted | Both |
| **Measures** | Distance of $\hat{\boldsymbol{\theta}}$ from $\boldsymbol{\theta}_0$ | Gradient at restriction | Change in fit (SSE) |
| **Null dist.** | $\chi^2_q$ | $\chi^2_q$ | $\chi^2_q/q$ or $F_{q,n-k}$ |
| **Robust?** | Yes (sandwich) | Score-like | No |

For linear restrictions in the normal model, all three are monotone transformations of each other and give identical decisions. They differ for nonlinear restrictions or non-normal models.

# Visualizing the Trinity




- **Wald**: horizontal distance between $\hat{\beta}$ and $\beta_0$ **Score**: slope at restricted estimate **F**: vertical distance ($\Delta$SSE)

# Confidence Intervals *Are* Inverted Tests

Recall the standard 95% CI:

$$\hat{C} = \left[\hat{\theta} - 1.96 \cdot s(\hat{\theta}),\ \ \hat{\theta} + 1.96 \cdot s(\hat{\theta})\right]$$

This is exactly the set of values $\theta$ that are *not rejected* by a two-sided $t$-test:

$$\hat{C} = \{\theta \ : \ |T(\theta)| \leq 1.96\}$$

**General principle:** Inverting a test with good Type I error control produces a confidence set with good coverage.

$$P[\theta \in \hat{C}] = P[\text{Accept} \mid \theta] = 1 - P[\text{Type I error}]$$

# Why Test Inversion Matters: Nonlinear Parameters

Consider $\theta = \beta_1/\beta_2$ (e.g., peak experience in a wage equation).

**Approach 1:** Delta method CI

$$\hat{\theta} \pm 1.96 \cdot s(\hat{\theta})$$

From Lecture 11, this can be inaccurate for ratios (Fieller's problem).

**Approach 2:** Rewrite as a *linear* restriction $\beta_1 - \theta\beta_2 = 0$ and invert:

$$\hat{C} = \left\{ \theta \ : \ \frac{(\hat{\beta}_1 - \theta\hat{\beta}_2)^2}{\boldsymbol{R}'(\theta)\hat{\boldsymbol{V}}\boldsymbol{R}(\theta)} \leq 1.96^2 \right\}$$

where $\boldsymbol{R}(\theta) = (1, -\theta)'$. This requires a grid search over $\theta$.

**Hansen's example:** Peak experience $= -50\beta_1/\beta_2$.
Delta method CI: [29.8, 29.9]. Inverted linear test CI: [29.1, 30.6].
The delta method interval is **far too narrow**.

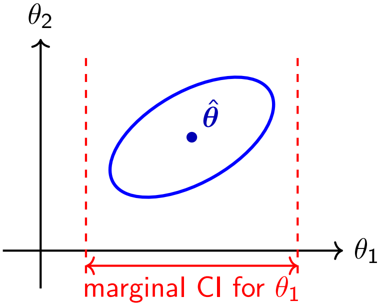
# Test Inversion in R

```r
# Wage equation: log(wage) ~ exper + exper^2/100 + ...
mod <- lm(lwage ~ exper + I(exper^2/100) + education,
           data = wages)
b <- coef(mod); V <- vcovHC(mod)

# Grid search: invert the t-test for theta = -50*b1/b2
theta_grid <- seq(20, 40, by = 0.01)
in_CI <- sapply(theta_grid, function(th) {
  R <- c(0, 1, -th/50, 0)  # gradient of b[2] - th*b[3]/50
  num <- (b[2] - th * b[3] / 50)^2
  den <- t(R) %*% V %*% R
  num / den <= qchisq(0.95, 1)
})
CI <- range(theta_grid[in_CI])
```

# Confidence Regions for Multiple Parameters

For $q > 1$ parameters, invert the Wald test: $\hat{C} = \left\{\boldsymbol{\theta} : W(\boldsymbol{\theta}) \leq \chi^2_{q,\,1-\alpha}\right\}$ — an **ellipsoid** in $\mathbb{R}^q$ centered at $\hat{\boldsymbol{\theta}}$, shaped by $\hat{\boldsymbol{V}}_{\theta}^{-1}$.




Marginal CIs (projections) are always wider than the ellipsoid in each dimension.

## The Multiple Testing Problem

Suppose you test $k$ hypotheses, each at level $\alpha = 0.05$. Under the **global null** (all $k$ true), what is $P$[at least one false rejection]?

By **Boole's inequality**:

$$P\left[\min_{j \le k} p_j < \alpha\right] \le \sum_{j=1}^{k} P[p_j < \alpha] \rightarrow k\alpha$$

| $k$ (tests) | $\alpha$ per test | Familywise error $\le$ |
|---|---|---|
| 5 | 0.05 | 0.25 |
| 10 | 0.05 | 0.50 |
| 20 | 0.05 | 1.00 |

With 20 tests, you are *virtually certain* to get a false rejection!

# The Bonferroni Correction

**Goal:** Control the *familywise error rate* (FWER) at $\alpha$.

**Rule:** Reject the $j$th hypothesis only if $p_j < \alpha/k$. Equivalently: $p_{\text{Bonf}} = k \cdot \min_{j \leq k} \; p_j$.

**Proof:**

$$P\left[\min_{j \leq k} p_j < \frac{\alpha}{k}\right] \leq \sum_{j=1}^{k} P\left[p_j < \frac{\alpha}{k}\right] \rightarrow k \cdot \frac{\alpha}{k} = \alpha$$

**Simple and conservative.** Controls FWER for *any* dependence structure among the tests. The cost: reduced power when $k$ is large.

# Bonferroni: Worked Example

Two coefficients tested at $\alpha = 0.05$:

| | Individual $p$ | Bonferroni $p$ $(= 2 \times p)$ |
|---|---|---|
| Union membership | 0.04 | 0.08 |
| Married status | 0.15 | 0.30 |

- **Without correction:** Union membership "significant" at 5%.
- **With Bonferroni:** Neither significant at FWER $= 0.05$ (need $p < 0.025$).

**When to worry:** exploring many specifications/subgroups, examining many coefficients, or reporting the "most significant" result.

# Bonferroni in R

```
# Get p-values from a regression
mod <- lm(lwage ~ education + exper + I(exper^2) +
              female + union + married, data = wages)
pvals <- summary(mod)$coefficients[-1, 4]  # drop intercept

# Bonferroni correction
p_bonf <- p.adjust(pvals, method = "bonferroni")
cbind(raw = round(pvals, 4), bonferroni = round(p_bonf, 4))

# Other options: Holm (less conservative, still controls FWER)
p_holm <- p.adjust(pvals, method = "holm")
```

**Holm's method** is uniformly more powerful than Bonferroni while still controlling FWER. Use it when available.

# Type I and Type II Errors

| | $H_0$ **true** | $H_0$ **false** |
|---|---|---|
| **Reject** $H_0$ | Type I error ($\alpha$) | Correct (Power $= 1 - \beta$) |
| **Accept** $H_0$ | Correct ($1 - \alpha$) | Type II error ($\beta$) |

- **Size** ($\alpha$): probability of falsely rejecting a true null
- **Power** ($\pi$): probability of correctly rejecting a false null: $\pi(\theta) = P[\text{Reject } H_0 \mid \theta \neq \theta_0]$
- Power depends on the *true value* of $\theta$, on $n$, and on $\alpha$.

**Key trade-off:** Making $\alpha$ smaller reduces Type I errors but also reduces power. A test with $\alpha = 0$ never rejects and has zero power.

# Power of the $t$-Test

For $H_0 : \theta = \theta_0$ vs. $H_1 : \theta > \theta_0$:

$$T = \frac{\hat{\theta} - \theta_0}{s(\hat{\theta})} \approx Z + \delta, \quad Z \sim N(0,1)$$

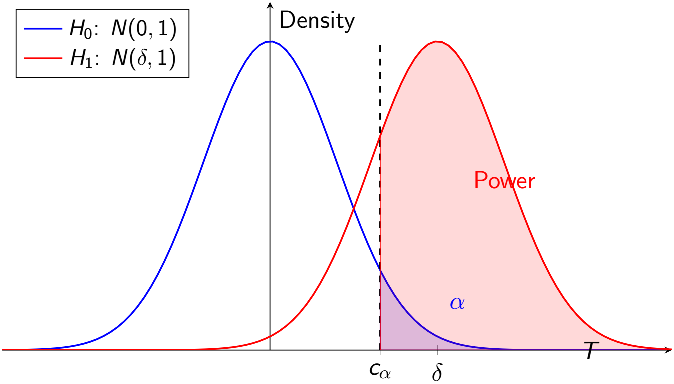
where $\delta = (\theta - \theta_0)/s(\hat{\theta})$ is the **signal-to-noise ratio**.

Power of a one-sided test at level $\alpha$: $\quad \pi(\delta) = \Phi(\delta - z_\alpha)$

**Power increases when:**

- The true effect $|\theta - \theta_0|$ is larger
- The standard error $s(\hat{\theta})$ is smaller (more precise estimation)
- The significance level $\alpha$ is larger (less stringent threshold)
- The sample size $n$ is larger (since $s(\hat{\theta}) \propto 1/\sqrt{n}$)

# Visualizing Power



# Power in OLS: What You Can Control

In OLS, $s(\hat{\theta}) \approx \sigma_e/(\mathrm{sd}(X) \cdot \sqrt{n})$ for a single regressor. So:

$$\delta \approx \frac{(\theta - \theta_0) \cdot \mathrm{sd}(X) \cdot \sqrt{n}}{\sigma_e}$$

**Levers for increasing power:**

1. **Increase** $n$: power grows with $\sqrt{n}$
2. **Reduce** $\sigma_e$: add controls that explain $Y$ (reduce residual variance)
3. **Increase sd**$(X)$: more variation in the regressor of interest
4. **Increase** $\alpha$: use 5% instead of 1% (but more Type I errors)

**Controls help power!** Relevant covariates reduce $\sigma_e$ without reducing $\mathrm{sd}(X)$—a free power boost. But irrelevant covariates cost degrees of freedom.

# Power and Test Dimension

For joint tests ($q > 1$), the Wald statistic under local alternatives: $W \xrightarrow{d} \chi^2_q(\lambda)$, where $\lambda = \boldsymbol{h}' \boldsymbol{V}_\theta^{-1} \boldsymbol{h}$ is the **non-centrality parameter**.

**Critical fact:** For a fixed $\lambda$, power *decreases* as $q$ increases.

| $q$ (restrictions) | $\lambda$ for 50% power | Sample size increase |
|---|---|---|
| 1 | 3.85 | baseline |
| 2 | 4.96 | +28% |
| 3 | 5.77 | +50% |

**Takeaway:** Testing more restrictions simultaneously dilutes power. A single-coefficient $t$-test is more powerful than a joint $F$-test that includes other restrictions.

# The 50% Power Benchmark

How far must the true parameter be from $\theta_0$ for 50% power?

**One-sided test:**

- At $\alpha = 0.05$: need $\delta \geq 1.65$ standard errors
- At $\alpha = 0.01$: need $\delta \geq 2.33$ standard errors

**Implication for sample size:** $(2.33/1.65)^2 \approx 2$.

A test at $\alpha = 0.01$ requires roughly **twice the sample size** as $\alpha = 0.05$ to achieve the same power.

**Two-sided** ($\alpha = 0.05$): need $|\delta| \geq 1.96$ for 50% power. Since se $\propto 1/\sqrt{n}$, you need $n \propto (\sigma_e/(\theta - \theta_0))^2$.

# Power Calculation in R

```
# Approximate power for a two-sided t-test in OLS
# Inputs: effect size, residual SD, regressor SD, n, alpha
power_ols <- function(effect, sigma_e, sd_x, n, alpha=0.05) {
  se <- sigma_e / (sd_x * sqrt(n))
  delta <- effect / se
  z <- qnorm(1 - alpha/2)
  power <- pnorm(delta - z) + pnorm(-delta - z)
  return(power)
}

# Example: detect beta = 0.1 with sigma_e = 1, sd(x) = 2
sapply(c(50, 100, 200, 500, 1000), function(n)
  round(power_ols(0.1, 1, 2, n), 3))
# [1] 0.080 0.117 0.198 0.463 0.803
```

With $\beta = 0.1$, $\sigma_e = 1$, $\mathrm{sd}(X) = 2$: you need $n \approx 1000$ for 80% power!

# Test Consistency

## Definition

*A test is* **consistent against fixed alternatives** *if for any true* $\theta \neq \theta_0$*:* $P[\text{Reject } H_0 \mid \theta] \to 1$ *as* $n \to \infty$.

**Good news:** The $t$-test and Wald test are consistent. As $n \to \infty$, $s(\hat{\theta}) \to 0$, so $|T| \to \infty$ whenever $\theta \neq \theta_0$.

**Caution:** Consistency means you will eventually reject any false null—including *economically trivial* deviations from $\theta_0$.

In very large samples, statistical significance $\neq$ economic significance. A statistically significant but tiny coefficient may not be meaningful.

# Hansen's Rules for Applied Work

1. **Report standard errors, not $t$-ratios.**
   - Standard errors focus attention on precision and confidence intervals.
2. **Report $p$-values, not asterisks.**
   - $p$-values contain more information than $*/{*}{*}/{*}{*}{*}$ ("an inferior practice").
3. **Focus on economically motivated hypotheses.**
   - Don't mechanically test every coefficient against zero.
   - Report the $t$-test for $\beta_j = 0$ when this is a scientifically interesting question.

# More Practical Advice

4. **"Do Not Reject" $\neq$ "Accept."**
   - Failing to reject means insufficient evidence, *not* that $H_0$ is true.
   - Never write: "the regression finds that $X$ has no effect on $Y$."
5. **Statistical significance $\neq$ economic significance.**
   - With large $n$, even tiny effects become "significant."
   - Always discuss the *magnitude* and substantive meaning.
6. **For nonlinear hypotheses, use minimum distance or test inversion.**
   - The Wald statistic is *not invariant* to the algebraic formulation of $H_0$.
   - Different equivalent formulations can give different results!

# The "What to Report" Checklist

For any coefficient of interest $\theta$:

1. **Point estimate** $\hat{\theta}$: the "best guess"
2. **Standard error** $s(\hat{\theta})$: measure of precision
3. **Confidence interval**: range of values consistent with the data
4. **Economic interpretation**: what the magnitude means in context

When testing:

5. State the **hypothesis** clearly (what question does this answer?)
6. Report the *p*-**value** (not just "significant" / "not significant")
7. Discuss **power** if you fail to reject (could you have detected a meaningful effect?)
8. Account for **multiple testing** if examining many hypotheses

# Summary

| Topic | Key Takeaway |
|---|---|
| F test | Measures fit loss from imposing $H_0$; $F = W^0/q$; needs homoskedasticity |
| Score test | Tests from restricted model; equivalent to $F$ for linear restrictions |
| Trinity | Wald, Score, F agree for linear $H_0$ in normal model |
| Test inversion | Confidence sets = non-rejected values; essential for nonlinear parameters |
| Bonferroni | With $k$ tests, use $\alpha/k$ to control familywise error |
| Power | $\delta =$ effect/se; grows with $\sqrt{n}$; 50% power needs $\delta \approx 2$ |
| Practical advice | Report SEs, CIs, $p$-values; focus on magnitudes; consider power |

# Next Time

**Lecture 13: Instrumental Variables**

- What happens when $E[\boldsymbol{Xe}] \neq \mathbf{0}$?
- Omitted variable bias, simultaneity, measurement error
- The IV solution: find $\boldsymbol{Z}$ such that $E[\boldsymbol{Ze}] = \mathbf{0}$ but $E[\boldsymbol{ZX}'] \neq \mathbf{0}$