Lecture 11: Hypothesis Testing

Module 3: part 4

Spring 2025

Logistics

- Continue hypothesis testing for linear models

Hypothesis Testing: A Systematic Approach

1. **Formulate Hypotheses**
   - H_0 (Null hypothesis): The "status quo" or "no effect" statement.
   - H_1 or H_A (Alternative hypothesis): What we seek evidence for.
   - Example: $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$ (two-sided)
2. **Choose an Appropriate Test Statistic**
   - Select based on data type and hypothesis structure.
   - Common tests: t-test, z-test, chi-square, F-test, etc.
   - Verify assumptions required for the selected test.
3. **Determine Significance Level (α)**
   - Set *before* collecting data (typically $\alpha = 0.05$ or 0.01).
   - Represents the probability of Type I error (rejecting a true H_0).
4. **Collect Data & Calculate Test Statistic**
   - Ensure proper sampling techniques.
   - Apply the selected statistical test formula.
5. **Determine p-value**
   - p-value = P(observing data at least as extreme as ours — H_0 is true).
   - Lower p-values indicate stronger evidence against H_0.
6. **Draw Conclusions**
   - If p-value $\leq \alpha$: Reject H_0 (statistically significant result).
   - If p-value $> \alpha$: Fail to reject H_0 (insufficient evidence).
   - Interpret in context of the original research question.

Common Misconceptions About p-values

- **What p-values are NOT:**
  - NOT the probability that H_0 is true
  - NOT the probability that results occurred by chance
  - NOT an indicator of effect size or practical significance
- **Statistical vs. Practical Significance**
  - Statistical significance: Evidence against H_0 (p-value $\leq \alpha$)
  - Practical significance: Meaningful real-world impact
  - Large samples can detect tiny, practically irrelevant effects
- **Types of Errors**
  - Type I Error: Rejecting H_0 when it's true (false positive)
  - Type II Error: Failing to reject H_0 when it's false (false negative)
  - Power = 1 - P(Type II Error) = Probability of correctly rejecting false H_0

Testing a Single Coefficient in Multiple Regression

Model Specification

Assume the following linear model with normally distributed errors:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$

Research Question

Does X_1 have a significant association with Y after controlling for $X_2, X_3, \ldots, X_p$?

Hypothesis Formulation

$$H_0 : \beta_1 = \beta_{1,0} \quad \text{(typically } \beta_{1,0} = 0\text{)}$$
$$H_A : \beta_1 \neq \beta_{1,0}$$

Note: The values of other coefficients ($\beta_0, \beta_2, \ldots, \beta_p$) are not specified in these hypotheses.

Testing a Single Regression Coefficient

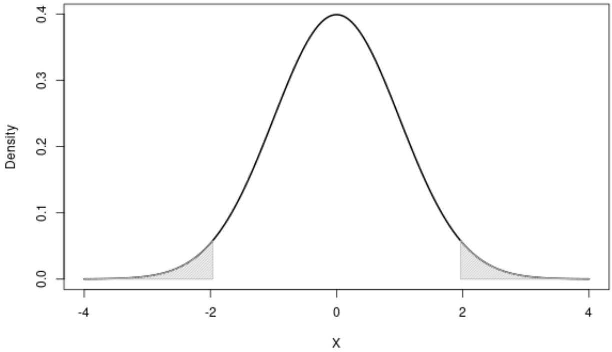
Test Statistic

$$t = \frac{\hat{\beta}_j - \beta_{j,0}}{\sqrt{\widehat{\text{var}}(\hat{\beta}_j)}}$$

- Testing $H_0 : \beta_j = \beta_{j,0}$ vs. $H_1 : \beta_j \neq \beta_{j,0}$ (or one-sided alternatives)
- Under H_0, $t \sim T_{n-p-1}$ (Student's t-distribution with $n - p - 1$ degrees of freedom)
  - $n =$ number of observations
  - $p =$ number of predictors (excluding intercept)
- **Critical region**: Values of t that lead to rejecting H_0
  - Two-sided test: Reject if $|t| > t_{\alpha/2,n-p-1}$
  - One-sided test: Reject if $t > t_{\alpha,n-p-1}$ or $t < -t_{\alpha,n-p-1}$
- Significance level $\alpha =$ probability of Type I error (rejecting H_0 when true)
- Most common case: testing $H_0 : \beta_j = 0$ (no effect of predictor j)

Testing a single coefficient

Find the appropriate cut-off for a T_{n-p-1} so that each shaded region has area $\alpha/2$



Understanding P-values

Definition

The **p-value** is the probability of observing a test statistic as extreme or more extreme than the one we calculated, *assuming the null hypothesis is true*.

- Equivalent to the rejection region approach:
  - Reject H_0 if p-value $< \alpha$
  - Fail to reject H_0 if p-value $\geq \alpha$
- For two-sided tests: p-value $= 2 \times P(|T| > |t_{\text{observed}}|)$
- For upper one-sided tests: p-value $= P(T > t_{\text{observed}})$
- For lower one-sided tests: p-value $= P(T < t_{\text{observed}})$

Common Misconceptions

The p-value is the probability that the null hypothesis is true.

Correct Interpretation

A p-value of 0.03 means: If the null hypothesis were true, we would observe a test statistic at least as extreme as ours in only 3% of repeated experiments.

Hypothesis Testing for a Single Regression Coefficient

1. **Formulate Hypotheses**
   - $H_0 : \beta_j = \beta_{j,0}$ (typically $\beta_{j,0} = 0$)
   - $H_A : \beta_j \neq \beta_{j,0}$ (or one-sided)
2. **Choose Significance Level**
   - Set α (typically 0.05)
3. **Fit Regression Model**
   - Obtain $\hat{\beta}_j$ and $\text{SE}(\hat{\beta}_j)$
4. **Calculate Test Statistic**
   - $t = \frac{\hat{\beta}_j - \beta_{j,0}}{\text{SE}(\hat{\beta}_j)}$
5. **Determine p-value**
   - $p = 2 \times P(|T_{n-p-1}| > |t|)$ for two-sided test
6. **Draw Conclusion**
   - Reject H_0 if $p < \alpha$
   - Fail to reject H_0 if $p \geq \alpha$

The Duality of Hypothesis Tests and Confidence Intervals

Confidence Interval (CI) for β_1 at level $1-\alpha$

$$\hat{\beta}_1 \pm t_{\alpha/2,n-p-1} \cdot \mathrm{SE}(\hat{\beta}_1) = \hat{\beta}_1 \pm t_{\alpha/2,n-p-1} \cdot \sqrt{\widehat{\mathrm{var}}(\hat{\beta}_1)} \tag{1}$$

Or written as an interval:

$$\left[\hat{\beta}_1 - t_{\alpha/2,n-p-1} \cdot \mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + t_{\alpha/2,n-p-1} \cdot \mathrm{SE}(\hat{\beta}_1)\right] \tag{2}$$

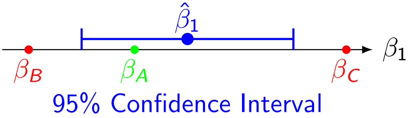
Key Insight

If a value $\beta_{1,0}$ is included in the $(1-\alpha)$ confidence interval, then:

$$\left|\frac{\hat{\beta}_1 - \beta_{1,0}}{\mathrm{SE}(\hat{\beta}_1)}\right| < t_{\alpha/2,n-p-1} \tag{3}$$

This means we would **fail to reject** the null hypothesis $H_0 : \beta_1 = \beta_{1,0}$ at significance level α.

Hypothesis Test vs Confidence Interval



- β_A: Inside CI $\Rightarrow$ Fail to reject $H_0 : \beta_1 = \beta_A$
- β_B, β_C: Outside CI $\Rightarrow$ Reject $H_0 : \beta_1 = \beta_B$ or $H_0 : \beta_1 = \beta_C$

Equivalence Relationships

- CI contains zero $\Leftrightarrow$ Fail to reject $H_0 : \beta_1 = 0$
- CI excludes zero $\Leftrightarrow$ Reject $H_0 : \beta_1 = 0$
- CI width $\propto$ Standard error of $\hat{\beta}_1$

Testing Multiple Coefficients

Testing Multiple Coefficients Simultaneously

Model Setup

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \beta_4 X_{i,4} + \ldots + \beta_p X_{i,p} + \varepsilon_i$$

Joint Hypothesis

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \quad \text{(other coefficients unrestricted)}$$

$$H_A : \text{At least one of } \beta_1, \beta_2, \beta_3 \text{ is non-zero}$$

Testing Multiple Coefficients Simultaneously

Multiple Testing Problem

Individual tests at $\alpha = 5\%$ level lead to inflated family-wise error rate:

# of tests p	P(false rej.)
1	0.05
2	0.975
5	0.2262
10	0.4013
20	0.6415

$P(FR) = 1 - (1 - \alpha)^p$ (ind.)

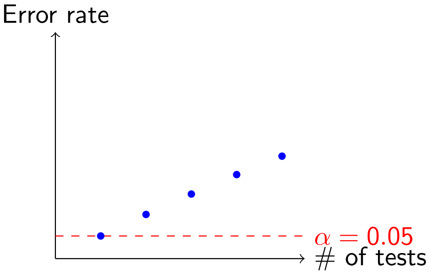


Figure: Family-wise error rate increases with number of tests

Solutions

- **F-test**: Tests joint significance of multiple coefficients (appropriate for our hypothesis).
- **Bonferroni correction**: Adjust individual test α levels by dividing by the number of tests.

F-test: Testing Multiple Coefficients Simultaneously

Joint Hypothesis

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \quad \text{(other coefficients unrestricted)}$$

$$H_A : \text{At least one of } \beta_1, \beta_2, \beta_3 \text{ is non-zero}$$

F-test Approach

1. Fit two models:
   - **Full model:** All predictors included
   - **Restricted model:** Set $\beta_1 = \beta_2 = \beta_3 = 0$
2. Calculate the F-statistic:

$$F = \frac{(RSS_{restricted} - RSS_{full})/q}{RSS_{full}/(n-p-1)}$$

where:

- RSS = Residual Sum of Squares
- q = Number of restrictions (here $q = 3$)
- n = Sample size
- p = Number of predictors in full model

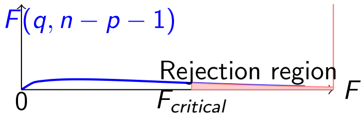


Figure: F distribution with rejection region

Decision Rule

- Reject H_0 if $F > F_{critical}$
- $F_{critical} = F_{1-\alpha,q,n-p-1}$
- Typical $\alpha = 0.05$

F-test for Nested Models

Hypothesis Setup:

- H_0: The excluded variables have no effect on the response.
- H_1: At least one of the excluded variables contributes significantly.

Models:

- **Alternative model (full model):**

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \ldots + \beta_p x_{i,p} + \varepsilon_i$$

- **Null model (reduced model):**

$$Y_i = \beta_0 + \underbrace{0 \cdot x_{i,1} + 0 \cdot x_{i,2} + 0 \cdot x_{i,3}}_{\text{excluded variables}} + \beta_4 x_{i,4} + \ldots + \beta_p x_{i,p} + \varepsilon_i$$

F-test for Nested Models

Hypothesis Setup:

- H_0: The excluded variables have no effect on the response.
- H_1: At least one of the excluded variables contributes significantly.

Models:

- **Alternative model (full model):**

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \ldots + \beta_p x_{i,p} + \varepsilon_i$$

- **Null model (reduced model):**

$$Y_i = \beta_0 + \underbrace{0 \cdot x_{i,1} + 0 \cdot x_{i,2} + 0 \cdot x_{i,3}}_{\text{excluded variables}} + \beta_4 x_{i,4} + \ldots + \beta_p x_{i,p} + \varepsilon_i$$

F-statistic:

$$F = \frac{(\text{RSS}_{\text{null}} - \text{RSS}_{\text{alt}})/d_f}{\text{RSS}_{\text{alt}}/d_f}$$

where RSS is the residual sum of squares and d_f denotes degrees of freedom.

F-test for Nested Models

Hypothesis Setup:

- H_0: The excluded variables have no effect on the response.
- H_1: At least one of the excluded variables contributes significantly.

Models:

- **Alternative model (full model):**

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \ldots + \beta_p x_{i,p} + \varepsilon_i$$

- **Null model (reduced model):**

$$Y_i = \beta_0 + \underbrace{0 \cdot x_{i,1} + 0 \cdot x_{i,2} + 0 \cdot x_{i,3}}_{\text{excluded variables}} + \beta_4 x_{i,4} + \ldots + \beta_p x_{i,p} + \varepsilon_i$$

F-statistic:

$$F = \frac{(\text{RSS}_{\text{null}} - \text{RSS}_{\text{alt}})/d_f}{\text{RSS}_{\text{alt}}/d_f}$$

where RSS is the residual sum of squares and d_f denotes degrees of freedom.

Interpretation:

- If F is large, the additional variables improve the model significantly.
- The F-test follows an F-distribution under H_0.
- If p-value $< \alpha$, reject H_0 and conclude that the excluded variables matter.

F-test

Under the alternative hypothesis, we can fit an **alternative model**: If the null hypothesis is false, we would expect

$$\underbrace{\sum_i (y_i - \tilde{y}_i)^2}_{RSS(\tilde{b}_0, \tilde{b}_1, \tilde{b}_2 \ldots \tilde{b}_p)} < \underbrace{\sum_i (y_i - \hat{y}_i)^2}_{RSS(\hat{b}_0, \hat{b}_1 = 0, \hat{b}_2 = 0, \hat{b}_3 = 0, \hat{b}_4, \ldots \hat{b}_p)}$$

F-test

Under the alternative hypothesis, we can fit an **alternative model**: If the null hypothesis is false, we would expect

$$\underbrace{\sum_i (y_i - \tilde{y}_i)^2}_{RSS(\tilde{b}_0, \tilde{b}_1, \tilde{b}_2 \ldots \tilde{b}_p)} < \underbrace{\sum_i (y_i - \hat{y}_i)^2}_{RSS(\hat{b}_0, \hat{b}_1 = 0, \hat{b}_2 = 0, \hat{b}_3 = 0, \hat{b}_4, \ldots \hat{b}_p)}$$

But even if the null hypothesis is true, we would expect

$$RSS(\tilde{b}_0, \tilde{b}_1, \tilde{b}_2 \ldots \tilde{b}_p) \leq RSS(\hat{b}_0, \hat{b}_1 = 0, \hat{b}_2 = 0, \hat{b}_3 = 0, \hat{b}_4, \ldots \hat{b}_p)$$

- Null model is always a choice under the alternative model
- When given more options, the minimizer cannot be worse
- If all the covariates of model A are included in model B and the dependent variables are the same, then we say model A is **nested** in model B

F-test

We want

- A statistic which is *extreme* when the null hypothesis is false.
- A statistic whose distribution we can describe under the null hypothesis.

$$RSS(Null) - RSS(Alt)$$

.

F-test

We want

- A statistic which is *extreme* when the null hypothesis is false.
- A statistic whose distribution we can describe under the null hypothesis.

$$\frac{RSS(H_0) - RSS(H_1)}{p_0 - p_1}$$

where p_0 and p_1 are the number of parameters (not including the intercept) which are being estimated in the null and alternative models respectively.

F-test

We want

- A statistic which is *extreme* when the null hypothesis is false.
- A statistic whose distribution we can describe under the null hypothesis.

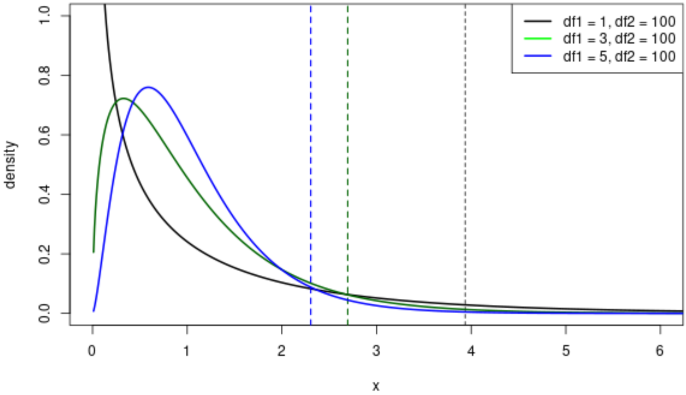
.

$$F = \frac{[RSS(H_0) - RSS(H_1)]/(p_1 - p_0)}{RSS(H_1)/(n - p_1 - 1)}$$

where p_{alt} and p_{null} are the number of parameters (not including the intercept) which are being estimated in the null and alternative models

F-distributions

F distributions have two parameters: df_1 and df_2



- For our statistic, $\text{df}_1 = p_{\text{alt}} - p_{\text{null}}$
- For our statistic, $\text{df}_2 = n - p_{\text{alt}} - 1$

Wrapup

- Hypothesis testing in Linear models
- Can test a single coefficient using T-test
- Can test several coefficients at once using F-test