Linear & Generalized Linear Models/ Advanced Regression for Independent Data

Andrew J. Spieker, Ph.D.

Associate Professor of Biostatistics
Vanderbilt University

Set 3: Hypothesis testing; ANOVA

Version: 07/03/2026

Table of Contents

Hypothesis testing:

- We seek to evaluate statistical strength of evidence. Example null hypotheses that I suspect you'll be familiar with (examples provided in parentheses):
  - $H_0 : \beta_1 = 0$ (association).
  - $H_0 : \beta_1 = \beta_2 = 0$ (overall/omnibus association).
  - $H_0 : \beta_2 - \beta_1 = 0$ (between-strata comparison).
  - $H_0 : \beta_0 + \beta_1 = 0$ (stratum-specific mean).
- How do we perform these tests?
- More generally, can we test hypotheses of the form $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$?
- Even more generally, can we test hypotheses of the form $\mathbf{C}\boldsymbol{\beta} = \mathbf{c}$?

Hypothesis testing:

- To begin with, suppose we want to test $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$, where $\boldsymbol{\beta}$ is of length K and $\mathbf{C}$ is a $Q \times K$ matrix (i.e., Q linear hypotheses).
- Intuitively, we want a test that rejects H_0 if $\mathbf{C}\widehat{\boldsymbol{\beta}}$ is "very different" from $\mathbf{0}$, and otherwise fails to reject.
- A Wald test statistic takes the form: $(\mathbf{C}\widehat{\boldsymbol{\beta}})^\intercal \left[\widehat{\text{Cov}}[\mathbf{C}\widehat{\boldsymbol{\beta}}]\right]^{-1} \mathbf{C}\widehat{\boldsymbol{\beta}}$.
- If $\mathbf{X}$ is of full rank and we use OLS under the assumption that $\text{Cov}[\boldsymbol{\epsilon}] = \sigma^2\mathbf{I}$, then a Wald test would reject H_0 for large values of:

$$(\mathbf{C}\widehat{\boldsymbol{\beta}})^\intercal \left[\mathbf{C}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{C}^\intercal\right]^{-1} \mathbf{C}\widehat{\boldsymbol{\beta}}/S^2,$$

where $S^2 = \text{RSS}/(N - K)$.

Assumptions and road map:

- We will derive a sensible test statistic similar to the Wald test described on the previous slide.
- We will then—under certain parametric assumptions—derive its exact distribution under the null and alternative hypotheses.
- We will be sure to talk a bit more about justification for the Wald test when parametric assumptions are not met.

Table of Contents

Setup:

- As has generally been the case thus far, the model is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.
- Assume $\text{Cov}[\boldsymbol{\epsilon}] = \sigma^2\mathbf{I}$ until further notice.
- We will present the results assuming $\mathbf{X}$ is of full rank.
  - Results generalize to the cases of $\text{rank}(\mathbf{X}) = R < K$ (you would see more pseudo-inverses, and "K" would generally be replaced with "R" throughout the notes).

Testable hypotheses:

- Suppose we want to test the hypothesis:

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0} \quad \text{or} \quad H_0 : \mathbf{c}_1^\intercal\boldsymbol{\beta} = \cdots = \mathbf{c}_Q^\intercal\boldsymbol{\beta} = 0.$$

- We say the hypothesis is *testable* if $\mathbf{c}_i^\intercal = \mathbf{d}_i^\intercal\mathbf{X}$ for $i = 1, \ldots, Q$.
  - This means $\mathbf{c}_1^\intercal\boldsymbol{\beta}$, ..., and $\mathbf{c}_Q^\intercal\boldsymbol{\beta}$ are estimable functions.
  - This occurs precisely when the rows of $\mathbf{C}$ are linear combinations of the rows of $\mathbf{X}$.
  - In the case of full rank, any row of $\mathbf{C}$ will be a linear combination of the rows of $\mathbf{X}$ (all linear hypotheses are testable). When $\mathbf{X}$ is not of full rank, not all hypotheses are testable, but we are not concerning ourselves with this case.
- In these notes, we will assume $Q \leq K$ and $\text{rank}(\mathbf{C}) = Q$.
  - That is, no redundancy.

Unrestricted and restricted models:

- Hypothesis: $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$.
- The premise of the hypothesis test that we're about to develop is based on a fit to an unrestricted model and a fit to the restricted model (under H_0). Let's talk about a couple of examples of what this means, since it's so central to the concepts.

Example 3.1: Testing an overall effect of a categorical variable

- Unrestricted model:

$$Y = \beta_0 + \beta_1 1(X = 1) + \beta_2 1(X = 2) + \beta_3 1(X = 3) + \epsilon.$$

- Hypothesis: $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$, or:

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \mathbf{0}.$$

- Restricted model (under H_0):

$$Y = \beta_0^* + \epsilon^*.$$

Example 3.2: Testing an overall effect in a two-way interaction model

- Unrestricted model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon.$$

- Hypothesis: $H_0 : \beta_1 = \beta_3 = 0$, or:

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \mathbf{0}.$$

- Restricted model (under H_0):

$$Y = \beta_0^* + \beta_2^* X_2 + \epsilon^*.$$

Example 3.3: Testing a stratum-specific effect an interaction model

- Unrestricted model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon.$$

- Hypothesis: $H_0 : \beta_1 + 2\beta_3 = 0$, or:

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \begin{bmatrix} 0 & 1 & 0 & 2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = 0.$$

- Restricted model (under H_0):

$$Y = \beta_0^* + \beta_1^*(X_1 - X_1 X_2/2) + \beta_2^* X_2 + \epsilon^*.$$

Testable hypotheses and key lemmas

Notation and intuition:

- RSS_H denotes the residual sum of squares under the null, H_0.
- RSS_H never smaller than RSS (i.e., under the unrestricted model).
  - I'll need to supply justification for this. Coming up!
- If H_0 is true, we expect RSS_H to be "only slightly greater" than RSS.
- If H_0 is false, we expect RSS_H to be "much greater" than RSS.

Helping to visualize the hypothesis test:

- Under the full model, $\mathbf{X}\boldsymbol{\beta} \in \mathcal{C}(\mathbf{X}) =: \Omega$.
- Since $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ is testable, we may write $\mathbf{C} = \mathbf{D}\mathbf{X}$ for some $\mathbf{D}$.
    - The rows of $\mathbf{C}$ are linear combinations of the rows of $\mathbf{X}$.
    - The columns of $\mathbf{C}$ are linear combinations of the columns of $\mathbf{D}$.
- Under H_0, $\mathbf{C}\boldsymbol{\beta} = \mathbf{D}\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$ such that $\mathbf{X}\boldsymbol{\beta} \in \mathcal{N}(\mathbf{D})$.
- Therefore, under the restricted model, $\mathbf{X}\boldsymbol{\beta} \in \mathcal{C}(\mathbf{X}) \cap \mathcal{N}(\mathbf{D}) =: \boldsymbol{\omega} \subset \Omega$.
- Let $\widehat{\mathbf{y}} = \mathbf{P}_{\Omega}\mathbf{y}$, denote the fitted value under the unrestricted model.
- Let $\widehat{\mathbf{y}}_H = \mathbf{P}_{\omega}\mathbf{y}$ denote the fitted value under the restricted model.

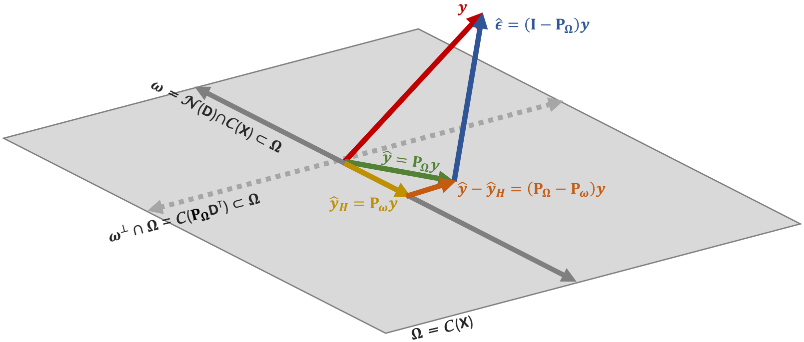

Figure 3.1: Visualizing hypothesis testing.

Understanding RSS:

- For the full model, we have:

$$\text{RSS} = (\mathbf{y} - \widehat{\mathbf{y}})^\intercal (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{y}^\intercal (\mathbf{I} - \mathbf{P}_\Omega)\mathbf{y}.$$

- For the restricted model, we have:

$$\text{RSS}_H = (\mathbf{y} - \widehat{\mathbf{y}}_H)^\intercal (\mathbf{y} - \widehat{\mathbf{y}}_H) = \mathbf{y}^\intercal (\mathbf{I} - \mathbf{P}_\omega)\mathbf{y}.$$

- It follows, then, that:

$$\text{RSS}_H - \text{RSS} = \mathbf{y}^\intercal (\mathbf{P}_\Omega - \mathbf{P}_\omega)\mathbf{y}.$$

- What have we learned about nested linear models?

Lemma 3.1: Difference between $\mathbf{P}_\Omega$ and $\mathbf{P}_\omega$

$\mathbf{P}_\Omega - \mathbf{P}_\omega = \mathbf{P}_{\omega^\perp \cap \Omega}$, and $\omega^\perp \cap \Omega = \mathcal{C}(\mathbf{P}_\Omega \mathbf{D}^\intercal)$.

- See an advanced regression textbook (e.g., Seber & Lee) for proofs.
- This is important; it implies that $\text{RSS}_H - \text{RSS} \geq 0$ (specifically, $\mathbf{P}_\Omega - \mathbf{P}_\omega \succcurlyeq 0$ because it is a projection matrix as we previously learned from Lemma 1.4).

Understanding RSS:

- We already know that $\mathbf{X}\boldsymbol{\beta} \in \Omega$, such that $(\mathbf{I} - \mathbf{P}_\Omega)\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$. Therefore,

$$\text{RSS} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\intercal(\mathbf{I} - \mathbf{P}_\Omega)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

- Under the restricted model, $\mathbf{X}\boldsymbol{\beta} \in \omega$, such that $(\mathbf{I} - \mathbf{P}_\omega)\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$. Therefore,

$$\text{RSS}_H = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\intercal(\mathbf{I} - \mathbf{P}_\omega)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

- It follows, then, that **under the restricted model**, we have:

$$\text{RSS}_H - \text{RSS} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\intercal(\mathbf{P}_\Omega - \mathbf{P}_\omega)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

- It may not be clear yet why I want to write RSS and RSS_H in terms of the (unknown) parameter $\boldsymbol{\beta}$; I will later invoke theorems/lemmas from Set 1 that demand I do so. This slide is extremely useful.

Lemma 3.2: Difference between $\mathbf{P}_\Omega$ and $\mathbf{P}_\omega$ (more explicitly)

If $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ is a testable hypothesis, then:

$$\mathbf{P}_\Omega - \mathbf{P}_\omega = \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{C}^\intercal(\mathbf{C}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{C}^\intercal)^{-1}\mathbf{C}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal.$$

- To prove this, note that $\mathbf{P}_\Omega - \mathbf{P}_\omega$ is a projection matrix that, by Lemma 3.1, projects onto the columns of $\mathbf{P}_\Omega\mathbf{D}^\intercal$. Expand out the following quantity:

$$\mathbf{P}_\Omega\mathbf{D}^\intercal((\mathbf{P}_\Omega\mathbf{D}^\intercal)^\intercal(\mathbf{P}_\Omega\mathbf{D}^\intercal))^{-1}(\mathbf{P}_\Omega\mathbf{D}^\intercal)^\intercal,$$

  and recognize that $\mathbf{P}_\Omega = \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal$ and $\mathbf{C} = \mathbf{D}\mathbf{X}$ (by the fact that $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ is testable).

Theorem 3.1: Expression for $\mathrm{RSS}_H - \mathrm{RSS}$

If $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ is a testable hypothesis, then:

$$\mathrm{RSS}_H - \mathrm{RSS} = (\mathbf{C}\widehat{\boldsymbol{\beta}})^\intercal(\mathbf{C}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{C}^\intercal)^{-1}\mathbf{C}\widehat{\boldsymbol{\beta}},$$

where $\widehat{\boldsymbol{\beta}}$ is the OLS estimate from the unrestricted model.

- Using the fact that $\mathrm{RSS}_H - \mathrm{RSS} = \mathbf{y}^\intercal(\mathbf{P}_\Omega - \mathbf{P}_\omega)\mathbf{y}$, this follows directly from Lemma 3.2 (use the fact that $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}$).

Lemma 3.3: Properties involving full-rank C

If $\mathbf{C}$ is of full (row) rank and corresponds to a testable hypothesis $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$, then $\mathbf{C}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{C}^\top \succ 0$.

- To prove this, note that $\mathbf{C} = \mathbf{D}\mathbf{X}$; therefore, $\mathbf{C}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{C}^\top = \mathbf{D}\mathbf{P}_\mathbf{X}\mathbf{D}^\top$, which is symmetric.
- Because $\mathbf{P}_\mathbf{X}$ is positive semi-definite, so too is $\mathbf{D}\mathbf{P}_\mathbf{X}\mathbf{D}^\top$.
- Letting $\mathbf{d}$ denote a column of $\mathbf{D}^\top$, note that $\mathbf{d}^\top\mathbf{P}_\mathbf{X}\mathbf{d} = 0$ if and only if $\mathbf{d} \in \mathcal{N}(\mathbf{X}^\top)$, but since $\mathbf{C} = \mathbf{D}\mathbf{X}$ is of full rank, this implies $\mathbf{d} = \mathbf{0}$.
- Hence, $\mathbf{D}\mathbf{P}_\mathbf{X}\mathbf{D}^\top$ is a $Q \times Q$ matrix of full rank (Q), and so we can strengthen positive semi-definiteness to positive-definiteness.

Testable hypotheses and key lemmas

Theorem 3.2: Expectation of $\text{RSS}_H - \text{RSS}$

If $\text{rank}(\mathbf{C}) = Q$, then:

$$\text{E}[\text{RSS}_H - \text{RSS}] = \sigma^2 Q + \boldsymbol{\beta}^\intercal \mathbf{C}^\intercal (\mathbf{C}(\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{C}^\intercal)^{-1} \mathbf{C}\boldsymbol{\beta}.$$

- To prove this, recall that $\text{RSS}_H = \text{RSS} = \mathbf{y}^\intercal (\mathbf{P}_\Omega - \mathbf{P}_\omega)\mathbf{y}$, employ the formula for the expectation of a quadratic form, and use Lemma 3.2.
- By Lemma 3.3, $\boldsymbol{\beta}^\intercal \mathbf{C}^\intercal (\mathbf{C}(\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{C}^\intercal)^{-1} \mathbf{C}\boldsymbol{\beta} = 0$ only when $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ is true, and is otherwise positive (which is *exactly* the property that we would want).

Motivating a reasonable test statistic:

- We've argued that $\text{RSS}_H - \text{RSS}$ will be (in expectation) equal to $Q\sigma^2$ under H_0, and will be larger when H_0 is false.
- This implies that the following may be reasonable to test H_0:

$$F = \frac{(\text{RSS}_H - \text{RSS})/Q}{\text{RSS}/(N-K)}.$$

- If F is large, we want to reject H_0.
- Thus far, we have only assumed $\text{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\text{Cov}[\epsilon] = \sigma^2 \mathbf{I}$.
- To obtain an exact form for the distribution under the null and under the alternative, we have to make a distributional assumption on $\boldsymbol{\epsilon}$.
- Certainly, the fact that we're calling it the F-statistic is suggestive of its distribution! :)

Table of Contents

Theorem 3.3: Distribution of of $\text{RSS}_H - \text{RSS}$ under H_0

Suppose $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $\text{rank}(\mathbf{X}) = K$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. If $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ is a testable hypothesis with $\text{rank}(\mathbf{C}_{Q\times K}) = Q$, then under the null hypothesis:

$$F = \frac{(\text{RSS}_H - \text{RSS})/Q}{\text{RSS}/(N-K)} \sim \mathcal{F}_{Q,N-K}.$$

Theorem 3.3: Proof

- Note that $\text{rank}(\mathbf{P}_\Omega) = K$ and $\text{rank}(\mathbf{P}_\omega) = K - Q$.
- Recall that $\mathbf{y}^\intercal(\mathbf{I} - \mathbf{P}_\Omega)\mathbf{y} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\intercal(\mathbf{I} - \mathbf{P}_\Omega)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$.
- Therefore, applying Theorem 1.3:
  - $\text{RSS}/\sigma^2 = \mathbf{y}^\intercal(\mathbf{I} - \mathbf{P}_\Omega)\mathbf{y}/\sigma^2 \sim \chi^2_{N-K}$.
- By Theorem 1.5 (together with Lemma 1.4):
  - $\text{RSS}_H - \text{RSS}$ is independent of RSS.
- By Theorem 1.6:
  - $(\text{RSS}_H - \text{RSS})/\sigma^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\intercal(\mathbf{P}_\Omega - \mathbf{P}_\omega)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})/\sigma^2 \sim \chi^2_Q$ under H_0.
- The result therefore follows immediately by the definition of the F-distribution.

Table of Contents

The distribution of F under the alternative

Motivation:

- Knowing the distribution of F under the null is what allows us to conduct the hypothesis test of interest having a specific level.
- However, finding a test that achieves the nominal level is not the hard part (if you don't believe me, construct a 20-sided fair die and reject the null hypothesis if and only if it comes up as a 17).
- We also want to know that the distribution of F is meaningfully different from its null distribution when in fact the null is false.
- Ideally, the "more different" $\mathbf{C}\boldsymbol{\beta}$ is from $\mathbf{0}$, the "more different" we would want the distribution of F to be from its null distribution. Given Theorem 3.2, we should feel pretty optimistic that this will be the case!

The distribution of F under the alternative

Theorem 3.4: Distribution of of $\text{RSS}_H - \text{RSS}$

Suppose $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $\text{rank}(\mathbf{X}) = K$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. If $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ is a testable hypothesis with $\text{rank}(\mathbf{C}_{Q\times K}) = Q$, then:

$$F = \frac{(\text{RSS}_H - \text{RSS})/Q}{\text{RSS}/(N-K)} \sim \mathcal{F}_{Q,N-K}(\lambda), \quad \text{where}$$

$$\lambda = \boldsymbol{\beta}^\intercal\mathbf{X}^\intercal(\mathbf{P}_\Omega - \mathbf{P}_\omega)\mathbf{X}\boldsymbol{\beta}/(2\sigma^2) = (\mathbf{C}\boldsymbol{\beta})^\intercal(\mathbf{C}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{C}^\intercal)^{-1}\mathbf{C}\boldsymbol{\beta}/(2\sigma^2).$$

Note: You can use this information to determine, for instance:

1. Sample size needed to achieve a certain power for a specified effect.
2. The power to detect a specified effect under a specified sample size.
3. The minimum effect at which a certain power will be achieved under a specified sample size.

R functions `pf` and `qf()` are useful!

Theorem 3.4: Proof

- As we've seen, $\text{rank}(\mathbf{P}_\Omega) = K$ and $\text{rank}(\mathbf{P}_\omega) = K - Q$.
- Recall that $\mathbf{y}^\intercal(\mathbf{I} - \mathbf{P}_\Omega)\mathbf{y} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\intercal(\mathbf{I} - \mathbf{P}_\Omega)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$.
- Therefore, applying Theorem 1.3:
  - $\text{RSS}/\sigma^2 = \mathbf{y}^\intercal(\mathbf{I} - \mathbf{P}_\Omega)\mathbf{y} \sim \chi^2_{N-K}$.
- By Theorem 1.5 (together with Lemma 1.4):
  - $\text{RSS}_H - \text{RSS}$ is independent of RSS.
- By Theorem 1.7:
  - $(\text{RSS}_H - \text{RSS})/\sigma^2 = \mathbf{y}^\intercal(\mathbf{P}_\Omega - \mathbf{P}_\omega)\mathbf{y}/\sigma^2 \sim \chi^2_Q((\mathbf{X}\boldsymbol{\beta})^\intercal(\mathbf{P}_\Omega - \mathbf{P}_\omega)\mathbf{X}\boldsymbol{\beta}/2\sigma^2)$.
- The result therefore follows immediately by the definition of the non-central F-distribution.

More about the non-centrality parameter:

- Recall that Lemma 3.2 presented an expression for $\mathbf{P}_\Omega - \mathbf{P}_\omega$, which we can leverage to write down the non-centrality parameter more specifically as:

$$\lambda = \frac{1}{2\sigma^2}\boldsymbol{\beta}^\intercal \mathbf{C}^\intercal (\mathbf{C}(\mathbf{X}^\intercal \mathbf{X})^{-1}\mathbf{C}^\intercal)^{-1}\mathbf{C}\boldsymbol{\beta}.$$

- It is straightforward to show that:

$$\mathrm{E}[F] = \left(1 + \frac{2\lambda}{Q}\right)\frac{N-K}{N-K-2}.$$

Table of Contents

Example: One-way ANOVA

Setup:

- Consider the following model:

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1N_1} \\ Y_{21} \\ \vdots \\ Y_{2N_2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1N_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2N_2} \end{bmatrix}.$$

- To-do:
  - Argue that $H_0 : \mu_1 = \mu_2$ is testable.
  - Write out H_0 in the form $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$.
  - Determine the F statistic and its distribution under H_0 assuming the errors are normally distributed.

Arguing testability:

- Model:

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1N_1} \\ Y_{21} \\ \vdots \\ Y_{2N_2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1N_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2N_2} \end{bmatrix}.$$

- $H_0 : \mu_1 = \mu_2 \Rightarrow \mu_1 - \mu_2 = 0$ can be written in the form $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$, with

$$\mathbf{C} = \begin{bmatrix} 1 & -1 \end{bmatrix},$$

which is testable because the (single) row of $\mathbf{C}$ is a linear combination of the rows of $\mathbf{X}$.

Deriving the $\boldsymbol{F}$-statistic:

- By Theorem 3.1 (expression for $\text{RSS}_H - \text{RSS}$) and Lemma 3.3 ($\mathbf{C}$ is of full row-rank), $\text{RSS}_H - \text{RSS} = (\mathbf{C}\widehat{\boldsymbol{\beta}})^\top(\mathbf{C}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{C}^\top)^{-1}\mathbf{C}\widehat{\boldsymbol{\beta}}$.
- Easiest to take this in in pieces:

$$\begin{aligned}
\mathbf{X}^\top\mathbf{X} &= \begin{bmatrix} N_1 & 0 \\ 0 & N_2 \end{bmatrix} \Rightarrow (\mathbf{X}^\top\mathbf{X})^{-1} = \begin{bmatrix} 1/N_1 & 0 \\ 0 & 1/N_2 \end{bmatrix} \\
\Rightarrow \mathbf{C}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{C}^\top &= \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} 1/N_1 & 0 \\ 0 & 1/N_2 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \\
&= 1/N_1 + 1/N_2. \\
\widehat{\boldsymbol{\beta}} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = \begin{bmatrix} \overline{Y}_1 \\ \overline{Y}_2 \end{bmatrix} \\
\Rightarrow (\mathbf{C}\widehat{\boldsymbol{\beta}})^\top(\mathbf{C}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{C}^\top)^{-1}(\mathbf{C}\widehat{\boldsymbol{\beta}}) &= \frac{(\overline{Y}_1 - \overline{Y}_2)^2}{1/N_1 + 1/N_2}.
\end{aligned}$$

Deriving the $\boldsymbol{F}$-statistic:

- Additionally, RSS $= \sum_{i=1}^{N_1}(Y_{1i} - \overline{Y}_1)^2 + \sum_{i=1}^{N_2}(Y_{2i} - \overline{Y}_2)^2$.
- In this example, $N = N_1 + N_2$, $Q = 1$ and $K = 2$. It follows, then, that

$$F = \frac{(\mathrm{RSS}_H - \mathrm{RSS})/Q}{\mathrm{RSS}/(N_1 + N_2 - 2)} = \frac{(\overline{Y}_1 - \overline{Y}_2)^2}{s_p^2(1/N_1 + 1/N_2)}.$$

- You may recognize this as the square of the t-statistic from a two-sample t-test that assumes equal variances between groups.
- Under H_0, Theorem 3.3 says that $F \sim \mathcal{F}_{1,N_1+N_2-2}$.
  - Note that if $T \sim t_{df}$, then $T^2 \sim \mathcal{F}_{1,df}$, which agrees with our result.

Example: One-way ANOVA

Simulation: Ideas

- It would be nice to see that this theoretical result holds up in a simulation setting.
- To check that the theory holds up, we want to generate data under the null, compute the F-statistic (repeatedly), and compare to, say, randomly generated data from F distribution with the degrees of freedom that the statistic should have ("simulation studies 101").
  - In this example, we want to check to see that the "whole distribution" matches. How do we do this?
    - Histograms.
    - QQ-plot (plot quantiles).
    - Descriptive statistics (e.g., mean, standard deviation, quantiles).

Simulation: Setup under H_0

```
## Set seed for reproducibility
set.seed(7345)

## Number of simulations
nsim <- 50000

## Space to store results
res <- matrix(0, nrow = nsim, ncol = 1)

## Set sample size
n1 <- 25
n2 <- 35

## Set common standard deviation
sigma <- 2

## Set mean difference (H0 : mean difference = 0)
delta <- 0
```

Simulation: Run simulation under H_0

```r
## Run simulations
for (j in 1:nsim)
{
  ## Generate data
  Y0 <- rnorm(n1, 0, sigma)
  Y1 <- rnorm(n2, delta, sigma)

  ## Compute RSS
  RSS <- sum((Y0 - mean(Y0))^2) + sum((Y1 - mean(Y1))^2)

  ## Compute F statistic
  Fstat <- ((mean(Y1) - mean(Y0))^2)/((1/n1 + 1/n2)*RSS/(n1 + n2 - 2))
  res[j,1] <- Fstat
}

## Generate data from theoretical distribution
f.compare <- rf(20000, df1 = 1, df2 = n1 + n2 - 2)
```

Simulation: Present results under H_0

```
## Histograms overlaid (harder to compare)
hist(res, freq = FALSE, col = rgb(1, 0, 0, 0.2), xlab = "F statistic",
     main = "Histogram (Null distribution)", ylim = c(0,0.8),
     xlim = c(0,35), breaks = seq(0,30,1))
par(new = TRUE)
hist(f.compare, freq = FALSE, col = rgb(0, 0, 1, 0.2),
     xlab = "", ylab = "", main = "", ylim = c(0,0.8),
     xlim = c(0,35), breaks = seq(0,35,1))
legend(10, 0.7, pch = 15, col = c(rgb(1, 0, 0, 0.2),
       rgb(0, 0, 1, 0.2)), lwd = 1, lty = 0,
       legend = c("Observed", "Theoretical"))

## Quantile-quantile plot (easier to compare, but be cautious)
qqplot(res, f.compare, frame.plot = FALSE, cex = 0.8, pch = 20,
       col = "darkgray", xlim = c(0,30), ylim = c(0,30),
       xlab = "F statistic quantile",
       ylab = "F(1, 58) - Random/Theoretical",
       main = "QQ-plot (Null distribution)")
abline(0,1, col = "blue")
```

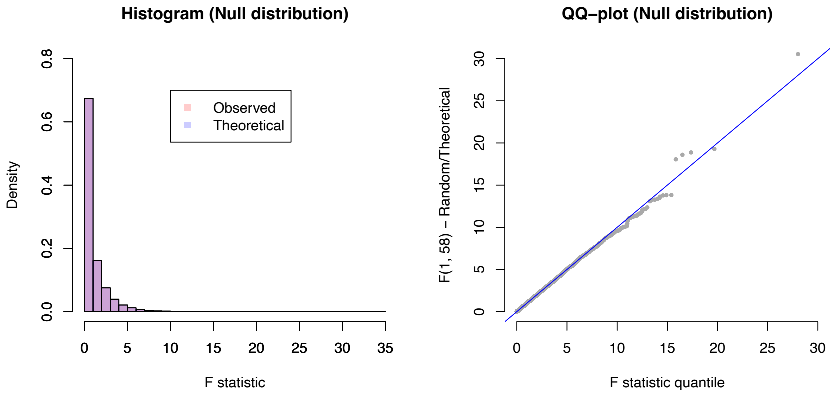

Figure 3.2: Simulation results under H_0.

Simulation: Setup and run simulation under alternative

```r
## Set seed for reproducibility
set.seed(7345)

## Set mean difference (mean difference = 2)
delta <- 2

## Space to store results
res <- matrix(0, nrow = nsim, ncol = 1)

## Run simulations
for (j in 1:nsim)
{
  ## Generate data
  Y0 <- rnorm(n1, 0, sigma)
  Y1 <- rnorm(n2, delta, sigma)

  ## Compute RSS
  RSS <- sum((Y0 - mean(Y0))^2) + sum((Y1 - mean(Y1))^2)

  ## Compute F statistic
  Fstat <- ((mean(Y1) - mean(Y0))^2)/((1/n1 + 1/n2)*RSS/(n1 + n2 - 2))
  res[j,1] <- Fstat
}

## Non-centrality parameter
lambda <- (1/(2*sigma^2))*((2^2)/(1/n1 + 1/n2))

## R parameterizes the non-centrality parameter differently
f.compare <- rf(20000, df1 = 1, df2 = n1 + n2 - 2, ncp = 2*lambda)
```

Example: One-way ANOVA

Simulation: Present results under alternative

```r
## Histograms overlaid (harder to compare)
hist(res, freq = FALSE, col = rgb(1, 0, 0, 0.2), xlab = "F statistic",
     main = "Histogram (Alternative distribution)", ylim = c(0,0.06),
     xlim = c(0,100), breaks = seq(0,110,2))
par(new = TRUE)
hist(f.compare, freq = FALSE, col = rgb(0, 0, 1, 0.2),
     xlab = "", ylab = "", main = "", ylim = c(0,0.06), xlim = c(0,100),
     breaks = seq(0,110,2))
legend(40, 0.05, pch = 15, col = c(rgb(1, 0, 0, 0.2),
       rgb(0, 0, 1, 0.2)), lwd = 1, lty = 0,
       legend = c("Observed", "Theoretical"))

## Quantile-quantile plot (easier to compare, but be cautious)
qqplot(res, f.compare, frame.plot = FALSE, cex = 0.8, pch = 20,
       col = "darkgray", xlim = c(0,100), ylim = c(0,100),
       xlab = "F statistic quantile",
       ylab = "F(1, 58; lambda) - Random/Theoretical",
       main = "QQ-plot (Altnernative distribution)")
abline(0,1, col = "blue")
```

Example: One-way ANOVA

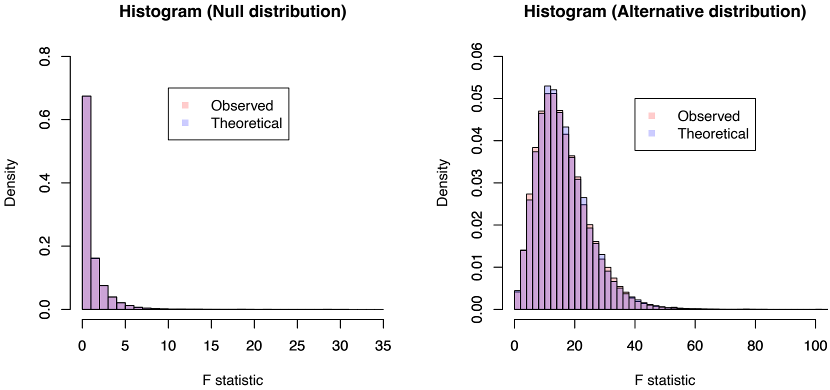

Figure 3.3: Simulation results under H_1.

Table of Contents

Ideas behind asymptotic justification:

- Building off of our previous example, note that key results tell us that the t-test is sensible in large samples even if the errors are not normally distributed.
- The natural question is: can we justify an F-test asymptotically even if our errors are not normally distributed?
  - Short answer: Yes!
- Without going into much technical detail, the F-test has theoretical justification if the errors are not normally distributed, provided the sample size is sufficiently large.
  - You will illustrate (but not prove) this on an upcoming problem set!

Note: Connection to χ^2 test

- If $X \sim \mathcal{F}_{d_1,d_2}$, then $Y = d_1 X \dot{\sim} \chi^2_{d_1}$ for large values of d_2.
- It stands to reason that, under H_0,

$$Q \times F \dot{\sim} \chi^2_Q,$$

provided $N - K$ is sufficiently large.

Note: Connection to χ^2 test

- If $X \sim \mathcal{F}_{d_1,d_2}(\lambda)$, then $Y = d_1 X \mathrel{\dot\sim} \chi^2_{d_1}(\lambda)$ for large values of d_2.
- It stands to reason that, under alternative hypotheses,

$$Q \times F \mathrel{\dot\sim} \chi^2_Q\left(\frac{1}{2\sigma^2}\boldsymbol{\beta}^\intercal \mathbf{C}^\intercal (\mathbf{C}(\mathbf{X}^\intercal \mathbf{X})^{-1}\mathbf{C}^\intercal)^{-1}\mathbf{C}\boldsymbol{\beta}\right),$$

provided $N - K$ is sufficiently large.

Table of Contents

TESTS OF $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{c}$

Expanding our results:

- Suppose you're interested in a set of the hypothesis $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{c}$.
- Another way to express this is that you're interested in the hypothesis $\mathbf{C}\boldsymbol{\beta} - \mathbf{c} = \mathbf{0}$.
- Results generalize in the expected way. For instance:
  - $\text{RSS}_H - \text{RSS} = (\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{c})^{\intercal}(\mathbf{C}(\mathbf{X}^{\intercal}\mathbf{X})^{-1}\mathbf{C}^{\intercal})^{-1}(\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{c})$.

Table of Contents

ANOVA: A little more formally

Ideas:

- ANOVA is an abbreviation for **an**alysis **o**f **va**riance. We will see now why it is so termed.
- Historically, ANOVA has been the approach of choice in the social sciences, laboratory research, etc., allowing one to compare means across discrete strata.

ANOVA: A little more formally

Ideas:

- Regression: design matrix is front and center:
  - $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.
- ANOVA: design matrix is an afterthought. Instead, we express the outcome in terms of group-specific means or between-group contrasts.
- Example: $i = 1, \ldots, n_j$ observations strata $j = 0, 1$.
  - Possible model: $Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$.
  - Possible model: $Y_{ij} = \alpha_j + \epsilon_{ij}$.
- Example: $i = 1, \ldots, n_{kj}$ observations in strata defined by $j = 1, \ldots, J$ and $k = 1, \ldots, K$.
  - Possible model: $Y_{ijk} = \mu_{jk} + \epsilon_{ijk}$.
  - Possible model: $Y_{ijk} = \alpha_j + \beta_k + \epsilon_{ijk}$.

Table of Contents

Example 3.4: One-way ANOVA with two groups

- Suppose we have $i = 1, \ldots, n_j$ independent observations in strata $j = 1, 2$ (shared outcome variance, σ^2). Let $N = n_1 + n_2$.

	Observations	Mean
Population 1	$y_{11}, \ldots, y_{1n_1}$	$\overline{y}_{1\cdot}$
Population 2	$y_{21}, \ldots, y_{2n_2}$	$\overline{y}_{2\cdot}$
Overall		$\overline{y}_{\cdot\cdot}$

- Model: $Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$.
- Suppose we seek to test $H_0 : \alpha_1 = \alpha_2$. We can use the F-statistic:

$$\begin{aligned} F &= \frac{(\text{RSS}_H - \text{RSS})/(J-1)}{\text{RSS}/(N-2)} \\ &= \frac{(n_1(\overline{y}_{1\cdot} - \overline{y}_{\cdot\cdot})^2 + n_2(\overline{y}_{2\cdot} - \overline{y}_{\cdot\cdot})^2)/(2-1)}{(\sum_{i=1}^{n_1}(y_{1i} - \overline{y}_{1\cdot})^2 + \sum_{i=1}^{n_2}(y_{2i} - \overline{y}_{2\cdot})^2)/(N-2)}. \end{aligned}$$

Example 3.5: One-way ANOVA with J groups

- Suppose we have $i = 1, \ldots, n_j$ independent observations in strata $j = 1, \ldots, J$ (shared outcome variance, σ^2). Let $N = \sum_{j=1}^{J} n_j$.

	Observations	Mean
Population 1	$y_{11}, \ldots, y_{1n_1}$	$\overline{y}_{1\cdot}$
Population 2	$y_{21}, \ldots, y_{2n_2}$	$\overline{y}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$
Population J	$y_{J1}, \ldots, y_{Jn_J}$	$\overline{y}_{J\cdot}$
Overall		$\overline{y}_{\cdot\cdot}$

- Model: $Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$.
- To test $H_0 : \alpha_1 = \cdots = \alpha_J$, we can use the F-statistic:

$$F = \frac{(\text{RSS}_H - \text{RSS})/(J-1)}{\text{RSS}/(N-J)} = \frac{(\sum_{j=1}^{J} n_i(\overline{y}_{j\cdot} - \overline{y}_{\cdot\cdot})^2)/(J-1)}{(\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ji} - \overline{y}_{j\cdot})^2)/(N-J)}.$$

Example 3.5: One-way ANOVA with J groups

- The F-statistic has an exact F-distribution under the null hypothesis *if* the errors are normally distributed.
- If normality does not hold, the F-distribution is approximate for large samples.

Example 3.5: One-way ANOVA with J groups

- Why can the quantity $\text{RSS}_H - \text{RSS}$ be written in this format?
  - Note first that $\text{RSS}_H = \sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{ji} - \overline{y}_{..})^2$.
  - Then, take note of the following very useful decomposition:

$$\begin{aligned} \sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{ji} - \overline{y}_{..})^2 &= \sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{ji} - \overline{y}_{j.})^2 + \sum_{j=1}^{J} n_j (\overline{y}_{j.} - \overline{y}_{..})^2 \\ SS_T &= SS_W + SS_B. \end{aligned}$$

- The total sum of squares can be written as the within-group sum of squares plus the between-group sum of squares.
- The residual sum of squares is none other than the within-group sum of squares.

Example 3.5: One-way ANOVA with J groups

- Suppose $n_1 = \cdots = n_J =: n$, so that $N = n \times J$.
- An ANOVA table typically would have the following information:

Source of variation	df	Sum of squares	Mean sum of squares
Between-group	$J-1$	SS_B	$SS_B/(J-1)$
Within-group	$N-J=J(n-1)$	SS_W	$SS_W/(N-J)$
Total	$N-1=nJ-1$	SS_T	$SS_T/(N-1)$

- With this, we can see more readily why this is called ANOVA!
- You can think of "between-group" as marking variability explained by the predictor of interest and "within group" as marking everything else (residual).

Table of Contents

Factorial design:

- Suppose we want to compare the effects of J chemotherapy agents together with K different dosage levels. In total, there are $J \times K$ combinations of levels. A total of n_{jk} patients are assigned to each group, with $N = \sum_{1 \leq j \leq J; 1 \leq k \leq K} n_{jk}$.
- ANOVA model: $Y_{ijk} = \mu + \alpha_{jk} + \epsilon_{ijk}$.
  - What does the design matrix look like?
  - What is the interpretation of each of the following hypotheses?
    - $H_0 : \alpha_{jk} = 0$ for all $1 \leq j \leq J$, $1 \leq k \leq K$.
    - $H_0 : \alpha_{11} = \alpha_{12} = \cdots = \alpha_{1K}$.
    - $H_0 : \alpha_{11} = \alpha_{12} = \cdots = \alpha_{1K}; \cdots ; \alpha_{J1} = \alpha_{J2} = \cdots = \alpha_{JK}$.
    - $H_0 : \alpha_{jk} = \alpha_j + \gamma_k$.

Two-way ANOVA:

- Suppose $n_{jk} = n$, so that $N = n \times J \times K$.
- A two-way ANOVA table typically would have the following information:

Source of variation	df	Sum of squares	Mean sum of squares
Factor X	$J-1$	SS_B^X	$SS_B^X/(J-1)$
Factor Z	$K-1$	SS_B^Z	$SS_B^Z/(K-1)$
Interaction XZ	$(J-1)(K-1)$	SS_B^{XZ}	$SS_B^{XZ}/((J-1)(K-1))$
Error	$JK(n-1)$	SS_W	$SS_W/(JK(n-1))$
Total	$N-1$	SS_T	$SS_T/(N-1)$

Example 3.6: Doxorubicin and serum

- Chemotherapy agents are generally preliminarily evaluated in a laboratory setting for activity.
- A sample drawn from a liquid culture of some cell line is exposed to a new drug or combinations of new drugs at varying concentrations.
- The cells are then put in a solid culture medium and incubated. The resulting colonies of cells can be counted using an optical scanner.
- The number of colonies visible at the end of incubation are a fair representation of the number of cells not killed by the treatment.
- As an example, consider a laboratory experiment of doxorubicin (concentrations of 0.05, 0.10, 0.50, and 1.00 μmol/L). Further, half the samples are exposed to 10% calf serum.
- Each of the eight defined groups has a sample size of $n = 24$.

Example: One-way ANOVA

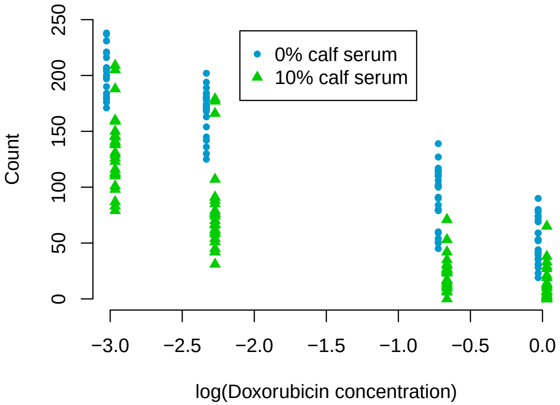

Figure 3.4: Doxorubicin, serum type, and count.

Example 3.6: Doxorubicin and serum

```
## Run ANOVA
zz <- aov(count ~ factor(doxconc) * factor(serum), data = dat)

> summary(zz)
                                Df Sum Sq Mean Sq F value   Pr(>F)
factor(doxconc)                  3 558897  186299 285.870  < 2e-16 ***
factor(serum)                    1 214067  214067 328.480  < 2e-16 ***
factor(doxconc):factor(serum)    3  16039    5346   8.204 3.74e-05 ***
Residuals                      184 119911     652
---
```

- How do we interpret each p-value?

Concluding thoughts on ANOVA:

- There is a lot more that can be said about the ANOVA framework.
- The "balanced" case results in particularly nice properties.
- Special considerations are necessary for unbalanced cases.
- The assumption of equal variances is dreadful, and ANOVA doesn't provide a way out of it as easily as regression does.
- Mixed models/random effects models fall under the category of "ANCOVA" (**an**anlysis of **cova**riance). This falls under the umbrella of BIOS 7346.
- I am biased toward the more flexible regression framework. However, ANOVA remains the approach of choice in some fields and you should familiarize yourself with the essential ideas.

Table of Contents

Decomposing variability:

- Another topic that has been notably absent so far is the coefficient of determination (and the related correlation coefficient).
- It's easiest to motivate some of these measures of association once we've wrapped our minds around sources of variation.

Important decomposition:

- Consider the regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $\text{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\text{Cov}[\boldsymbol{\epsilon}] = \sigma^2\mathbf{I}$. Assume that $\mathbf{X}$ is of full rank.
- Let $\widehat{y}$ denote the fitted value based on $\widehat{\boldsymbol{\beta}}$, the OLS estimate of $\boldsymbol{\beta}$.
- Then,

$$\begin{aligned} SS_T &= \sum_{i=1}^{N}(y_i - \overline{y})^2 \overset{\text{math}}{=} \cdots = \sum_{i=1}^{N}(y_i - \widehat{y}_i)^2 + \sum_{i=1}^{N}(\widehat{y}_i - \overline{y})^2 \\ &= SS_R + SS_M. \end{aligned}$$

- SS_T: total sum of squares.
  - Marks variability of Y about $\overline{Y}$.
- SS_R: residual sum of squares (akin to "within").
  - Marks variability of Y about $\widehat{Y}$.
- SS_M: model sum of squares (akin to "between").
  - Marks variability of $\widehat{Y}$ about $\overline{Y}$.

Additional decomposition:

- Note: $SS_R = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^\intercal(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \mathbf{y}^\intercal\mathbf{y} - \widehat{\boldsymbol{\beta}}^\intercal\mathbf{X}^\intercal\mathbf{y}$. Why?
- With algebra, $SS_R = \sum_{i=1}^{N}(y_i - \overline{y})^2 - \widehat{\boldsymbol{\beta}}_1^\intercal(\mathbf{X}^C)^\intercal\mathbf{y}$, where $\mathbf{X}^C$ denotes the design matrix with predictors centered about respective means (omit column of ones for the intercept) and $\widehat{\boldsymbol{\beta}}_1$ denotes coefficients of $\widehat{\boldsymbol{\beta}}$ without the intercept.
    - Hint 1: $SS_T = \sum_{i=1}^{N}(y_i - \overline{y})^2 = \mathbf{y}^T\mathbf{y} - N\overline{y}^2$.
    - Hint 2: Centering X's alters $\widehat{\beta}_0$ (updated to $\overline{y}$) and the remaining estimates are unaffected.
    - Hint 3: $\widehat{\boldsymbol{\beta}}_1^\intercal(\mathbf{X}^C)^\intercal\mathbf{y} = \widehat{\boldsymbol{\beta}}^\intercal\mathbf{X}^\intercal\mathbf{y} - N\overline{y}^2$.
- Taken together, this implies that $SS_M = \widehat{\boldsymbol{\beta}}_1^\intercal(\mathbf{X}^C)^\intercal\mathbf{y}$.
- Now, since $\widehat{\boldsymbol{\beta}}$ solves the appropriate normal equations, we further find that $SS_M = \widehat{\boldsymbol{\beta}}_1^\intercal[(\mathbf{X}^C)^\intercal\mathbf{X}^C]\widehat{\boldsymbol{\beta}}_1$.

Interpretation:

- The coefficient of determination is given as:

$$R^2 = \frac{SS_T - SS_R}{SS_T} = \frac{SS_M}{SS_T} = \frac{\widehat{\boldsymbol{\beta}}_1^\intercal [(\mathbf{X}^C)^\intercal \mathbf{X}^C] \widehat{\boldsymbol{\beta}}_1}{\sum_{i=1}^N (y_i - \overline{y})^2}.$$

- Marks the proportion of "total variation" in the outcome explained by the overall model.
- Sometimes described as a measure of model fit, although I find this ambiguous. I want to avoid conflating concepts of correct model specification and a model's predictive capacity.
- Now, SS_T can also be partitioned as:

$$\begin{aligned} SS_T &= \sum_{i=1}^N (y_i - \overline{y})^2 = \mathbf{y}^\intercal \mathbf{y} - N\overline{y}^2 = (\mathbf{y}^\intercal \mathbf{y} - \widehat{\boldsymbol{\beta}} \mathbf{X}^\intercal \mathbf{y}) + (\widehat{\boldsymbol{\beta}} \mathbf{X}^\intercal \mathbf{y} - N\overline{y}^2) \\ &= SS_R + SS_M. \end{aligned}$$

Coefficient of determination

Properties:

1. $0 \leq R^2 \leq 1$.
2. $R = \sqrt{R^2}$ marks the correlation between Y and $\widehat{Y}$.
3. Adding variables to a model cannot decrease the value of R^2.

Distribution under the null:

- Consider the hypothesis $H_0 : \beta_1 = \cdots = \beta_{K-1} = 0$ (i.e., the predictors are not linearly associated with the outcome).
- The F-test associated with this hypothesis can be written as:

$$F = \frac{R^2/(K-1)}{(1-R^2)/(N-K)}.$$

- With a little algebra, we can rearrange, finding:

$$R^2 = \frac{(K-1)F}{(N-K)+(K-1)F} = \frac{\left(\frac{K-1}{N-K}\right)F}{1+\left(\frac{K-1}{N-K}\right)F}.$$

Distribution under the null: $H_0 : \beta_1 = \cdots = \beta_{K-1} = 0$

- If we are willing to assume that $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$, then we know that $F \sim F_{K-1,N-K}$ under H_0, from which we learn that:

$$R^2 = \frac{\left(\frac{K-1}{N-K}\right) F}{1 + \left(\frac{K-1}{N-K}\right) F} \sim \text{Beta}\left(\frac{K-1}{2}, \frac{N-K}{2}\right).$$

- From this, we also learn that (under H_0):

$$\text{E}[R^2] = \frac{K-1}{N-1}.$$

- This is the motivation for the adjusted R^2, which will have expectation zero under the null.

Coefficient of determination

Adjusted $\mathbf{R^2}$:

- The adjusted R^2 penalizes the number of parameters in the model:

$$R^2_{\text{adj}} = 1 - (1 - R^2)\frac{N-1}{N-K}.$$

- In so doing, we have that (under H_0),

$$\mathrm{E}[R^2_{\text{adj}}] = 0.$$

- Trade-off: The adjusted R^2 can take on a negative value.

Coefficient of partial determination: Partial R^2

- Suppose we seek to understand the proportion of variation not explained in a reduced model that is explained by predictors in a full(er) model.
- The partial R^2 is given by:

$$R^2_{\text{Full}|\text{Reduced}} = \frac{SS_R^{\text{Reduced}} - SS_R^{\text{Full}}}{SS_R^{\text{Reduced}}}.$$

- Here, SS_R^{Reduced} is taking the place of SS_T, which you can think of as being based on a model that has been reduced to the maximal extent.

Relationship to more general *F*-test:

- Similarly, suppose we seek to test (intercept *not* involved):

$$H_0 : \begin{bmatrix} 0 & \mathbf{C}_{Q\times(K-1)} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \boldsymbol{\beta}_{-0} \end{bmatrix} = \mathbf{0}.$$

- Then, the F-statistic can be expressed as:

$$F = \frac{(R^2 - R^2_H)/Q}{(1 - R^2)/(N - K)},$$

  where R^2_H is the coefficient of determination under the restricted model. You should be able to verify that this is the more general version of what we just saw.

- Keep in mind that we are assuming $\text{rank}(\mathbf{C}) = Q$ and $\mathbf{X}$ is of full rank.

Summary

This unit:

- Testable hypotheses.
- The F-statistic.
- ANOVA.
- The coefficient of determination.

Summary: So far

- Random vectors and matrices; multivariate normal theory.
- Ordinary least squares.
- Hypothesis testing and ANOVA.

Summary: Coming up

- Weighted least squares.
- Confidence regions and prediction.
- Diagnostics.
- Exponential families.
- Generalized linear models.
- Sandwich and bootstrap.
- Quasi-likelihood.
- Hypothesis testing for GLMs.
- Diagnostics for GLMs.
- Further considerations for binary outcomes.
- Nonlinear least squares.