30 Lecture 30: April 23

Last time

- Analysis of Variance (JF chapter 8)
  - two-way anova
  - higher-way anova

Today

- analysis of covariance (ANCOVA)
- Final exam (take-home) will be posted April 30th and due midnight May 7th.
- Course evaluation.

Analysis of Covariance

Analysis of covariance (ANCOVA) is a term used to describe linear models that contain both qualitative and quantitative explanatory variables. The method is, therefore, equivalent to dummy-variable regression, discussed in the previous lectures, although the ANCOVA model is parametrized differently from the dummy-regression model.

Covariate is a variable known to affect the response that

1. differs among EUs
2. reflects differences that exist independently of experimental treatment.

A nutrition example

A nutrition scientist conducted an experiment to evaluate the effects of four vitamin supplements on the weight gain of laboratory animals. The experiment was conducted in a completely randomized design with $N = 20$ animals randomized to $a = 4$ supplement groups, each with sample size $n \equiv 5$. The response variable of interest is weight gain, but calorie intake z was measured simultaneously.

Diet	$y(g)$	Diet	y	Diet	y	Diet	y
1	48	2	65	3	79	4	59
1	67	2	49	3	52	4	50
1	78	2	37	3	63	4	59
1	69	2	75	3	65	4	42
1	53	2	63	3	67	4	34
1	$\bar{y}_{1+} = 63$	2	$\bar{y}_{2+} = 57.8$	3	$\bar{y}_{3+} = 65.2$	4	$\bar{y}_{4+} = 48.8$
1	$s_1 = 12.3$	2	$s_2 = 14.9$	3	$s_3 = 9.7$	4	$s_4 = 10.9$

Question: Is there evidence of a vitamin supplement effect?

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Diet	3	797.8	265.9	1.823	0.184
Residuals	16	2334.4	145.9		

But calorie intake z was measured simultaneously:

Diet	$y(g)$	z	Diet	y	z	Diet	y	z	Diet	y	z
1	48	350	2	65	400	3	79	510	4	59	530
1	67	440	2	49	450	3	52	410	4	50	520
1	78	440	2	37	370	3	63	470	4	59	520
1	69	510	2	75	530	3	65	470	4	42	510
1	53	470	2	63	420	3	67	480	4	34	430

Question: How and why could these new data be incorporated into analysis?
Answer: ANCOVA can be used to reduce unexplained variation.

ANCOVA model,

$$y_{ij} = \mu + \alpha_i + \beta z_{ij} + \epsilon_{ij}$$

where μ is the reference level, α_i is the main effect of treatment, β is the partial regression coefficient, and $\epsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. The model is equivalent as the dummy-variable regression model,

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_z z_i + \epsilon_i \quad \text{for } i = 1, \ldots, 20$$

Finish the table below

Source	df
Diet	
Covariate	1
Residual	
Total	

Answer:

To test for difference among treatments. The null hypothesis in terms of α_i is
$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_4 = 0$ v.s. $H_a :$ at least one $\alpha_i \neq 0$
And the null hypothesis in terms of β_i is
$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ v.s. $H_a :$ at least one $\beta_i \neq 0$

Question: which two models do we compare when testing the above null hypothesis?
Answer:

Linear contrasts of means

With ANOVA (or ANCOVA) models, we do not generally test hypotheses about individual coefficients (but we can do so if we wish). For dummy-coded regressors in one-way ANOVA, a t-test or F-test of $H_0 : \alpha_1 = 0$, for example, is equivalent to testing for the difference in means between the first group and the baseline group, $H_0 : \mu_1 = \mu_m$.

Consider the one-way ANOVA model:

$$Y_{ij} = \mu_i + \epsilon_{ij}, i = 1, 2, \ldots, t, \text{ and } j = 1, 2, \ldots, n_i$$

with $\epsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.

A linear function of the group means of the form

$$\theta = c_1\mu_1 + c_2\mu_2 + \cdots + c_t\mu_t$$

is called a linear combination of the treatment means. And the c_i's are the coefficients of the linear combination. If

$$c_1 + c_2 + \cdots + c_t = \sum_{j=1}^{t} c_j = 0,$$

the linear combination is called a contrast. Contrasts with more than two non-zero coefficients are called complex contrasts.

Let two contrasts θ_1 and θ_2 be given by

$$\theta_1 = c_1\mu_1 + \cdots + c_t\mu_t = \sum_{j=1}^{t} c_j\mu_j$$

$$\theta_2 = d_1\mu_1 + \cdots + d_t\mu_t = \sum_{j=1}^{t} d_j\mu_j,$$

then the two contrasts θ_1 and θ_2 are mutually orthogonal if the products of their coefficients sum to zero:

$$c_1d_1 + \cdots + c_td_t = \sum_{j=1}^{t} c_jd_j = 0$$

θ_i and θ_j are orthogonal $\implies$ $\hat{\theta}_i$ and $\hat{\theta}_j$ are statistically independent.

Types of effects

Consider the following two-way ANOVA model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$
$$i = 1, 2 = a \text{ and } j = 1, 2 = b \text{ and } k = 1, 2, \ldots, 7 = n.$$

$\epsilon_{ijk} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. Parameter constraints: $\sum_i \alpha_i = \sum_j \beta_j = 0$ and $\sum_i (\alpha\beta)_{ij} = 0$ for each j and $\sum_j (\alpha\beta)_{ij} = 0$ for each i.

- Factor A: AGE has $a = 2$ levels - A_1 : younger and A_2 : older
- Factor B: GENDER has $b = 2$ levels - B_1 : female and B_2 : male

Three kinds of effects in this 2×2 design:

1. Simple effects are simple contrasts.
   - $\mu(A_1B) = \mu_{12} - \mu_{11}$ - simple effect of gender for young folks.
   - $\mu(AB_1) = \mu_{21} - \mu_{11}$ - simple effect of age for women.
2. Interaction effects are differences of simple effects: $\mu(AB) = \mu(AB_2) - \mu(AB_1) = (\mu_{22} - \mu_{12}) - (\mu_{21} - \mu_{11})$
   - difference between simple age effects for men and women
   - difference between simple gender effects for old and young folks
   - interaction effect of AGE and GENDER.
3. Main effects are averages or sums of simple effects

$$\mu(A) = \frac{1}{2}(\mu(AB_1) + \mu(AB_2))$$

$$\mu(B) = \frac{1}{2}(\mu(A_1B) + \mu(A_2B))$$

Sampling distribution of linear contrast estimates

For a linear contrast

$$\theta = c_1\mu_1 + \cdots + c_t\mu_t$$

The *best* estimator for a contrast of interest can be obtained by substituting treatment group sample means $\bar{y}_{i+}$ for treatment population means μ_i in the contrast θ:

$$\hat{\theta} = c_1\bar{Y}_{1+} + c_2\bar{Y}_{2+} + \cdots + c_t\bar{Y}_{t+}$$

Example

Recall the binding fraction data that investigate binding fraction for several antibiotics using $n = 20$ bovine serum samples:

Antibiotic	Binding Percentage	Sample mean
Penicillin G	29.6 24.3 28.5 32.0	28.6
Tetracyclin	27.3 32.6 30.8 34.8	31.4
Streptomycin	5.8 6.2 11.0 8.3	7.8
Erythromycin	21.6 17.4 18.3 19	19.1
Chloramphenicol	29.2 32.8 25.0 24.2	27.8

Consider the pairwise contrast comparing penicillin (population) mean to Tetracyclin mean:

$$\theta = \mu_1 - \mu_2 = (1)\mu_1 + (-1)\mu_2 + (0)\mu_3 + (0)\mu_4 + (0)\mu_5$$

Obtain a point estimator of θ.
Answers:

Question: How good is this estimate? In other words, how much uncertainty associated with the estimate?

We want to characterize the sampling distribution of $\hat{\theta}$. According to our model setup, Y_{ij} follow normal distributions. $\hat{\theta}$ is a linear function of Y_{ij}, so that $\hat{\theta}$ follows a normal distribution. We want to derive the mean and variance to characterize the normal distribution that $\hat{\theta}$ follows:

$$\hat{\theta} \sim \mathcal{N}(\theta, Var(\hat{\theta}))$$

Derive expressions for the mean and the variance:

Therefore, the standard error:

$$SE(\hat{\theta}) = \sqrt{Var(\hat{\theta})} = \sqrt{\sigma^2 \sum_{j=1}^{t} \frac{c_j^2}{n_j}}$$

which is estimated by

$$\widehat{SE}(\hat{\theta}) = \sqrt{MS[E] \sum_{j=1}^{t} \frac{c_j^2}{n_j}}$$

To test $H_0 : \theta = \theta_0$ (often 0) versus $H_1 : \theta \neq \theta_0$, use t-test:

$$t = \frac{\hat{\theta} - \theta_0}{\widehat{SE}(\hat{\theta})} \overset{H_0}{\sim} t_{N-t}$$

At level α, the critical value for this test is $t(N-t, \alpha/2)$ and $100(1-\alpha)\%$ confidence interval for a contrast $\theta = \sum c_j \mu_j$ is given by

$$\sum c_j \bar{Y}_{j+} \pm t(N-t, \alpha/2) \sqrt{MS[E] \sum \frac{c_j^2}{n_j}}$$

Multiple Comparisons

Let's first review type I and type II errors.

	H_0 is True	H_0 is False
Don't reject H_0	Probability $1-\alpha$	Probability β
Reject H_0	Probability α	Probability $1-\beta$

- Type I error: rejection of a true null hypothesis (false positive).
- Type II error: failure to reject a false null hypothesis (false negative).
- Type I error rate or significance level (α): the probability of rejecting the null hypothesis given the null hypothesis is true.
- Type II error rate (β): the probability of failure to reject the null hypothesis given the null hypothesis is false. $1-\beta$ gives the power of a test.

Now, let's consider all simple (pairwise) contrasts for the binding fraction data with $t = 5$ antibiotic treatments of the form $\theta = \mu_i - \mu_j$.

- We have $\begin{pmatrix} 5 \\ 2 \end{pmatrix} = 10$ tests for significance each at level $\alpha = 0.05$
- what is the probability of committing at least one type I error?

We need to consider the familywise error rate (fwe) when testing k contrasts:

$$fwe = \Pr(\text{at least one type I error})$$

Methods for simultaneous inference for multiple contrasts include

- Bonferroni
- Scheffé
- Tukey

When the number of comparisons is in the hundreds or thousands (e.g. genome-wide association studies), and FWE control is hopeless, more manageable type I error rate is the False Discovery Rate (FDR):

$$FDR = E(\frac{\text{Falsely rejected null hypotheses}}{\text{Number of rejected null hypotheses}})$$

Bonferroni correction

Suppose interest lies in exactly k contrasts. The Bonferroni adjustment to α controls fwe is

$$\alpha_{bonferroni} = \frac{\alpha}{k}$$

and simultaneous 95% confidence intervals for the k contrasts are given by

$$a_1\bar{Y}_{1+} + \cdots + a_t\bar{Y}_{t+} \pm t(\frac{\alpha_{bonferroni}}{2}, \nu)\sqrt{MS[E]\sum\frac{a_j^2}{n_j}}$$

$$b_1\bar{Y}_{1+} + \cdots + b_t\bar{Y}_{t+} \pm t(\frac{\alpha_{bonferroni}}{2}, \nu)\sqrt{MS[E]\sum\frac{b_j^2}{n_j}}$$

$$\cdots$$

$$k_1\bar{Y}_{1+} + \cdots + k_t\bar{Y}_{t+} \pm t(\frac{\alpha_{bonferroni}}{2}, \nu)\sqrt{MS[E]\sum\frac{k_j^2}{n_j}}$$

where ν denotes df for error.

Example: for the binding fraction example, consider only pairwise comparisons with Penicillin:

$$\theta_1 = \mu_1 - \mu_2, \theta_2 = \mu_1 - \mu_3, \theta_3 = \mu_1 - \mu_4, \theta_4 = \mu_1 - \mu_5$$

We have $k = 4, \alpha_{bonferroni} = 0.05/k = 0.0125$ and $t(\frac{\alpha_{bonferroni}}{2}, 15) = 2.84$. Substitution leads to

$$t(\frac{\alpha_{bonferroni}}{2}, 15)\sqrt{MS[E]\left(\frac{1^2}{4} + \frac{(-1)^2}{4} + \frac{0^2}{4} + \cdots + \frac{0^2}{4}\right)}$$

$$=2.84\sqrt{(9.05)\frac{2}{4}} = 6.0$$

so that **simultaneous** 95% confidence intervals for θ_1, θ_2, θ_3 and θ_4 take the form

$$\bar{y}_{1+} - \bar{y}_{i+} \pm 6.0$$

Scheffé

Another method to construct **simultaneous** 95% confidence intervals for **ALL** contrasts, use

$$\sum_{j=1}^{t} c_j \bar{y}_{j+} \pm \sqrt{(t-1)(F^*)MS[E]\sum_{j=1}^{t}\frac{c_j^2}{n_j}}$$

where $F^* = F(\alpha, t-1, N-t)$. For a pairwise comparisons of means, μ_j and μ_k, this yields

$$\bar{y}_{j+} - \bar{y}_{k+} \pm \sqrt{(t-1)(F^*)MS[E](1/n_j + 1/n_k)}$$

Using $\alpha = 0.05$, need to specify

- t (from the design)
- F^* (same critical value as for $H_0 : \alpha_i \equiv 0$).
- $MS[E]$ (from the data)
- $\bar{y}_{j+}$, $\bar{y}_{k+}$
- n_j, n_k (from the data)

For binding fraction data,

$$\sqrt{(t-1)(F^*)MS[E](\frac{1}{n_j}+\frac{1}{n_k})} = \sqrt{(5-1)(3.06)9.05(\frac{1}{4}+\frac{1}{4})} = 7.44$$

If any two sample means differ by more than 7.44, they differ significantly.

Tukey

Tukey's method is better than Scheffé's method when making **all pairwise** comparisons in balanced designs ($n = n_1 = n_2 = \cdots = n_t$). It is conservative, controlling the experimentwise error rate, and has a lower type II error rate in these cases than Scheffé. (It is more powerful.)

For simple contrasts of the form

$$\theta = \mu_j - \mu_k$$

to test

$$H_0 : \theta = 0 \text{ vs } H_1 : \theta \neq 0$$

reject H_0 at level α if

$$|\hat{\theta}| > q(t, N-t, \alpha)\sqrt{\frac{MS[E]}{n}}$$

where $q(t, N-t, \alpha)$ denotes α level studentized range for t means and $N-t$ degrees of freedom, the quantity $q(t, N-t, \alpha)\sqrt{\frac{MS[E]}{n}}$ is referred to as Tukey's honestly significant difference (HSD). The studentized ranges can be calculated using R function `qtukey(1 − α, t, N − t)`.

Sample size computations for one-way ANOVA

Now consider the null hypothesis in a balanced experiment using one-way ANOVA to compare t treatment means and $\alpha = 0.05$:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_t = \mu$$

versus the alternative

$$H_a : \mu_i \neq \mu_j \text{ for some } i \neq j$$

Suppose that we intend to use a balanced design. How big does our sample size $n_1 = n_2 = \cdots = n_t = n$ need to be?

The answer depends on lots of things, namely, σ^2 and how many treatment groups t and how much of a difference among the means we hope to be able to detect, and with how big a probability.

Given α, $\mu_1, \ldots, \mu_t$ and σ^2, we can choose n to ensure a power of at least $1 - \beta$ (i.e., an upper bound on type II error rate) using the noncentral F distribution.

Recall that the critical region for the statistic $F = MS[Trt]/MS[E]$ is everything bigger than $F(\alpha, t-1, N-t) = F^*$.

The power of the F-test conducted using $\alpha = 0.05$ to reject H_0 under this alternative is given by

$$1 - \beta = \Pr(MS[Trt]/MS[E] > F^*; H_1 \text{ is true}). \tag{1}$$

Let $\tau_i = \mu_i - \mu$ for each treatment i so that

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_t = 0$$

When some H_1 is true and the sample size n is used in each group, it can be shown that the F ratio has the noncentral F distribution with noncentrality parameter

$$\gamma = \sum_{j=1}^{t} n_j \left(\frac{\tau_j}{\sigma}\right)^2 = n \sum_{j=1}^{t} \left(\frac{\tau_j}{\sigma}\right)^2$$

This is the parameterization for the F distribution used in both SAS and R.

One way to obtain an adequate sample size is trial and error. Software packages can be used to get probabilities of the form 1 for various values of n.

Example

Suppose we want to test equal mean binding fractions among antibiotics against the alternative

$$H_1 : \mu_P = \mu + 3, \mu_T = \mu + 3, \mu_S = \mu - 6, \mu_E = \mu, \mu_C = \mu$$

so that

$$\tau_1 = \tau_2 = 3, \tau_3 = -6, \tau_4 = \tau_5 = 0.$$

Assume $\sigma = 3$ (is it arbitrary? any idea of how to guess?) and we need to use $\alpha = 0.05$. The noncentrality parameter is given by

$$\gamma = n[(\frac{3}{3})^2 + (\frac{3}{3})^2 + (\frac{-6}{3})^2]$$

The $\alpha = 0.05$ critical value for H_0 is given by

$$F^* = F(5-1, 5n-5, 0.05).$$

We need the area to the right of F^* for the noncentral F distribution with degrees of freedom 4 and $5(n-1)$ and noncentrality parameter $\gamma = 6n$ to be greater or equal to the desired power level of $1 - \beta = 0.8$.

We will revisit this example in the lab session on Friday.

Multivariate normal distribution

- The standard multivariate normal is a vector of independent standard normals, denoted $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$. The joint density is

  $$f_{\mathbf{Z}}(\mathbf{z}) = \frac{1}{(2\pi)^{p/2}} e^{-\sum_{i=1}^p z_i^2/2}.$$

  The mgf is

  $$m_{\mathbf{Z}}(\mathbf{t}) = \prod_{i=1}^p m_{Z_i}(t_1) = \prod_{i=1}^p e^{t_i^2/2} = e^{\mathbf{t}^T\mathbf{t}/2}.$$

- Consider the affine transformation $\mathbf{X} = \boldsymbol{\mu} + \mathbf{A}\mathbf{Z}$ where $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$. $\mathbf{X}$ has mean and variance

  $$\mathrm{E}(\mathbf{X}) = \boldsymbol{\mu}, \quad \mathrm{Var}(\mathbf{X}) = \mathbf{A}\mathbf{A}^T$$

  and the moment generating function is

  $$m_{\mathbf{X}}(\mathbf{t}) = \mathrm{E}(e^{\mathbf{t}^T(\boldsymbol{\mu}+\mathbf{A}\mathbf{Z})}) = e^{\mathbf{t}^T\boldsymbol{\mu}}\mathrm{E}(e^{\mathbf{t}^T\mathbf{A}\mathbf{Z}}) = e^{\mathbf{t}^T\boldsymbol{\mu}+\mathbf{t}^T\mathbf{A}\mathbf{A}^T\mathbf{t}/2}.$$

- $\mathbf{X} \in \mathbb{R}^p$ has a multivariate normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance $\mathbf{V} \in \mathbb{R}^{p\times p}, \mathbf{V} \succeq_{p.s.d.} \mathbf{0}$, denoted $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$, if its mgf takes the form

  $$m_{\mathbf{X}}(\mathbf{t}) = e^{\mathbf{t}^T\boldsymbol{\mu}+\mathbf{t}^T\mathbf{V}^T\mathbf{t}/2}, \quad \mathbf{t} \in \mathbb{R}^p$$

- if $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$ and $\mathbf{V}$ is non-singular, then
  - $\mathbf{V} = \mathbf{A}\mathbf{A}^T$ for some non-singular $\mathbf{A}$
  - $\mathbf{A}^{-1}(\mathbf{X} - \boldsymbol{\mu}) \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$
  - The density of $\mathbf{X}$ is

    $$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{V}|^{1/2}} e^{-(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu})/2}.$$

- (Any affine transform of normal is normal) If $\mathbf{X} \in \mathbb{R}^p, \mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$ and $\mathbf{Y} = \mathbf{a} + \mathbf{B}\mathbf{X}$, where $\mathbf{a} \in \mathbb{R}^q$ and $\mathbf{B} \in \mathbb{R}^{q \times p}$, then $\mathbf{Y} \sim \mathcal{N}(\mathbf{a} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B}\mathbf{V}\mathbf{B}^T)$.
- (Marginal of normal is normal) If $\mathbf{X} \in \mathbb{R}^p, \mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$, then any subvector of $\mathbf{X}$ is normal too.
- A convenient fact about normal random variables/vectors is that zero correlation/covariance implies independence.
  If $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$ and is partitioned as
  $$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_m \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \vdots \\ \boldsymbol{\mu}_m \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} \mathbf{V}_{11} & \dots & \mathbf{V}_{1m} \\ \vdots & & \vdots \\ \mathbf{V}_{m1} & \dots & \mathbf{V}_{mm} \end{bmatrix}$$
  then $\mathbf{X}_1, \dots, \mathbf{X}_m$ are jointly independent if and only if $\mathbf{V}_{ij} = \mathbf{0}$ for all $i \neq j$.
  Proof:

Independence and Cochran's theorem

- (Independence between two linear forms of a multivariate normal) Let $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$, $\mathbf{Y}_1 = \mathbf{a}_1 + \mathbf{B}_1\mathbf{X}$ and $\mathbf{Y}_2 = \mathbf{a}_2 + \mathbf{B}_2\mathbf{X}$. Then $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are independent if and only if $\mathbf{B}_1\mathbf{V}\mathbf{B}_2^T = 0$.
  Proof:
- Consider the normal linear model $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\mathbf{b}, \sigma^2\mathbf{I}_n)$
  - Using $\mathbf{A} = (1/\sigma^2)(\mathbf{I} - \mathbf{P}_\mathbf{X})$, we have
    $$SSE/\sigma^2 = ||\hat{\boldsymbol{\epsilon}}||_2^2/\sigma^2 = \mathbf{y}^T\mathbf{A}\mathbf{y} \sim \chi^2_{n-r},$$
    where $r = \text{rank}(\mathbf{X})$. Note the noncentrality parameter is
    $$\phi = \frac{1}{2}(\mathbf{X}\mathbf{b})^T(1/\sigma^2)(\mathbf{I} - \mathbf{P}_\mathbf{X})(\mathbf{X}\mathbf{b}) = 0 \text{ for all } \mathbf{b}.$$
  - Using $\mathbf{A} = (1/\sigma^2)\mathbf{P}_\mathbf{X}$, we have
    $$SSR/\sigma^2 = ||\hat{\mathbf{y}}||_2^2/\sigma^2 = \mathbf{y}^T\mathbf{A}\mathbf{y} \sim \chi^2_r(\phi),$$
    with the noncentrality parameter
    $$\phi = \frac{1}{2}(\mathbf{X}\mathbf{b})^T(1/\sigma^2)\mathbf{P}_\mathbf{X}(\mathbf{X}\mathbf{b}) = \frac{1}{2\sigma^2}||\mathbf{X}\mathbf{b}||_2^2.$$
  - The joint distribution of $\hat{\mathbf{y}}$ and $\hat{\boldsymbol{\epsilon}}$ is
    $$\begin{bmatrix} \hat{\mathbf{y}} \\ \hat{\boldsymbol{\epsilon}} \end{bmatrix} = \begin{bmatrix} \mathbf{P}_\mathbf{X} \\ \mathbf{I}_n - \mathbf{P}_\mathbf{X} \end{bmatrix} \mathbf{y} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{X}\mathbf{b} \\ \mathbf{0}_n \end{bmatrix}, \begin{bmatrix} \sigma^2\mathbf{P}_\mathbf{X} & \mathbf{0} \\ \mathbf{0} & \sigma^2(\mathbf{I} - \mathbf{P}_\mathbf{X}) \end{bmatrix} \right).$$
    So $\hat{\mathbf{y}}$ is independent of $\hat{\boldsymbol{\epsilon}}$. Thus $||\hat{\mathbf{y}}||_2^2/\sigma^2$ is independent of $||\hat{\boldsymbol{\epsilon}}||_2^2/\sigma^2$ and
    $$F = \frac{||\hat{\mathbf{y}}||_2^2/\sigma^2/r}{||\hat{\boldsymbol{\epsilon}}||_2^2/\sigma^2/(n-r)} \sim F_{r,n-r}(\frac{1}{2\sigma^2}||\mathbf{X}\mathbf{b}||_\mathbf{2}^\mathbf{2}).$$

- (Independence between linear and quadratic forms of a multivariate normal) Let $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$. Let $\mathbf{A}$ be symmetric with rank s. Then $\mathbf{BX}$ and $\mathbf{X}^T\mathbf{AX}$ are independent if $\mathbf{BVA} = \mathbf{0}$.
Proof:

- (Independence between two quadratic forms of a multivariate normal) Let $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$, $\mathbf{A}$ be symmetric with rank r, and $\mathbf{B}$ be symmetric with rank s. If $\mathbf{BVA} = \mathbf{0}$, then $\mathbf{X}^T\mathbf{AX}$ and $\mathbf{X}^T\mathbf{BX}$ are independent.
Proof:

- (Cochran's theorem) Let $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I}_n)$ and $\mathbf{A}_i$, $i = 1, \ldots, k$ be symmetric idempotent matrix with rank s_i. If $\sum_{i=1}^k \mathbf{A}_i = \mathbf{I}_n$, then $(1/\sigma^2)\mathbf{y}^T\mathbf{A}_i\mathbf{y}$ are independent $\chi^2_{s_i}(\phi_i)$, with $\phi_i = \frac{1}{2\sigma^2}\boldsymbol{\mu}^T\mathbf{A}_i\boldsymbol{\mu}$ and $\sum_{i=1}^k s_i = n$.
Proof:

- Application to the one-way ANOVA: $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$. We have the classical ANOVA table

Source	df	Projection	SS	Noncentrality
Mean	1	$\mathbf{P_1}$	$SSM = n\bar{y}^2$	$\frac{1}{2\sigma^2}n(\mu + \bar{\alpha})^2$
Group	$a-1$	$\mathbf{P_X} - \mathbf{P_1}$	$SSA = \sum_{i=1}^a n_i\bar{y}_i^2 - n\bar{y}^2$	$\frac{1}{2\sigma^2}\sum_{i=1}^a n_i(\alpha_i - \bar{\alpha})^2$
Error	$n-a$	$\mathbf{I} - \mathbf{P_X}$	$SSE = \sum_{i=1}^a \sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$	0
Total	n	$\mathbf{I}$	$SST = \sum_i \sum_j y_{ij}^2$	$\frac{1}{\sigma^2}\sum_{i=1}^a n_i(\mu + \alpha_i)^2$

Bootstrap

We follow JF Chapter 21 to discuss the version of nonparametric bootstrap here. The term *bootstrapping*, coined by Efron (1979), refers to using the sample to learn about the sampling distribution of a statistic without reference to external assumptions – as in "pulling oneself up by one's bootstraps."

Bootstrapping offers a number of advantages:

- The bootstrap is quite general, although there are some cases in which it fails.
- Because it does not require distributional assumptions (such as normally distributed errors), the bootstrap can provide more accurate inferences when the data are not well behaved or when the sample size is small.
- It is possible to apply the bootstrap to statistics with sampling distributions that are difficult to derive, even asymptotically.
- It is relatively simple to apply the bootstrap to complex data collection plans.

Bootstrap standard errors

For simplicity, we start with an iid sample $Y_1, \ldots, Y_n$ with each Y_i having distribution function F, and a real parameter θ is estimated by $\hat{\theta}$. When necessary, we think of $\hat{\theta}$ as a function of the sample, $\hat{\theta}(Y_1, \ldots, Y_n)$. The variance of $\hat{\theta}$ is then

$$\mathrm{Var}_F(\hat{\theta}) = \int \left\{ \hat{\theta}(y_1, \ldots, y_n) - \mathrm{E}_F(\hat{\theta}) \right\}^2 dF(y_1) \ldots dF(y_n),$$

where

$$\mathrm{E}_F(\hat{\theta}) = \int \hat{\theta}(y_1, \ldots, y_n) dF(y_1) \ldots dF(y_n).$$

The nonparametric bootstrap estimate of $\mathrm{Var}(\hat{\theta})$ is just to replace F by the empirical distribution function $F_n(y) = n^{-1} \sum_{i=1}^{n} I(Y_i \leqslant y)$:

$$\mathrm{Var}_{F_n}(\hat{\theta}) = \int \left\{ \hat{\theta}(y_1, \ldots, y_n) - \mathrm{E}_{F_n}(\hat{\theta}) \right\}^2 dF_n(y_1) \ldots dF_n(y_n),$$

Please refer to Chapter 11 of Boos and Stefanski for a complete discussion.

A practical bootstrapping procedure follows:

1. Create r number of bootstrap replications or pseudo-replicates – that is, for each bootstrap sample (replicate) $b = 1, \ldots, r$, we randomly draw n observations $\{Y_{b_1}^*, Y_{b_2}^*, \ldots, Y_{b_n}^*\}$ **with replacement** from the original sample $\{Y_1, Y_2, \ldots, Y_n\}$.
2. Obtain an estimate $\hat{\theta}_b^*$ of each bootstrap sample.
3. Use the distribution of $\hat{\theta}_b^*$ to estimate properties of the sampling distribution of $\hat{\theta}$.For example, the sample standard deviation of $\hat{\theta}_b^*$ gives the bootstrap standard error estimates of $\widehat{SE}^*(\hat{\theta})$.

Bootstrap example

We use the example in JF 21.1 for illustration. Imagine that we sample (fake) ten working, married couples, determining in each case the husband's and wife's income, as recorded in the table (JF table 21.3) below.

Observation	husband's Income	Wife's Income	Difference Y_i
1	34	28	6
2	24	27	-3
3	50	45	5
4	54	51	3
5	34	28	6
6	29	19	10
7	31	20	11
8	32	40	-8
9	40	33	7
10	34	25	9

A point estimate of this population mean difference μ is the sample mean,

$$\bar{Y} = \frac{\sum Y_i}{n} = 4.6$$

Elementary statistical theory tells us that the standard deviation of the sampling distribution of sample means is $\mathrm{SD}(\bar{Y}) = \sigma/\sqrt{n}$, where σ is the population standard deviation of Y. Because we do not know σ in most real applications, the usual estimator of σ is the sample standard deviation

$$\hat{S} = \sqrt{\frac{\sum(Y_i - \bar{Y})^2}{n-1}}$$

and we obtain the 95% confidence interval by

$$\bar{Y} \pm t_{n-1,0.025}\frac{\hat{S}}{\sqrt{n}}$$

In the present case, $\hat{S} = 5.948$, $\widehat{SE}(\bar{Y}) = 5.948/\sqrt{10} = 1.881$, and $t_{9,0.025} = 2.262$. The 95% confidence interval for the population mean μ is therefore

$$4.6 \pm 2.262 \times 1.881 = 4.6 \pm 4.255$$

or equivalently,

$$0.345 < \mu < 8.855$$

To illustrate the bootstrap procedure,

1. We can draw $r = 2000$ bootstrap samples (using a computer), each of size $n = 10$, from the original data given in table 21.3.

2. We then calculate the mean $\bar{Y}_b^*$, with $b = 1, \ldots, r$ for each bootstrap sample.

3. The bootstrap estimate of the standard error is then given by $\widehat{SE}^*(\bar{Y}^*) = \sqrt{\frac{\sum_{b=1}^{r}\left(\bar{Y}_b^* - \bar{\bar{Y}}^*\right)^2}{r-1}}$

From the 2000 replicates that Dr. Fox drew, he obtained $\bar{\bar{Y}}^* = 4.693$ and $\widehat{SE}(\bar{Y}^*) = 1.750$. Both are quite close to the theoretical values (read JF 21.1 for a discussion over $\sqrt{n/n-1}$ for the differences in calculating the standard errors, which is often negligible, especially when n is large).

Now, we can get a bootstrap estimate for the $100(1-\alpha)\%$ confidence interval by using the $\alpha/2$ and $(1-\alpha/2)$ quantiles of the bootstrap sampling distribution of $\hat{\theta}_b^*$ which means

1. We order $\hat{\theta}_b^*$ such that $\hat{\theta}_{(1)}^* \leqslant \hat{\theta}_{(2)}^* \leqslant \cdots \leqslant \hat{\theta}_{(r)}^*$.

2. Find the two quantiles $\hat{\theta}_{(lower)}^* = \hat{\theta}_{(\alpha/2\times r)}^*$ and $\hat{\theta}_{(upper)}^* = \hat{\theta}_{((1-\alpha/2)\times r)}^*$

3. Construct the confidence interval by $(\hat{\theta}_{(lower)}^*, \hat{\theta}_{(upper)}^*)$.

In this case,

$$\begin{aligned} \text{lower} &= 2000(0.05/2) = 50 \\ \text{upper} &= 2000(1 - 0.05/2) = 1950 \\ \bar{Y}_{(50)}^* &= 0.7 \\ \bar{Y}_{(1950)}^* &= 7.8 \\ 0.7 < \mu &< 7.8 \end{aligned}$$

Bias-corrected bootstrap intervals

We introduce the bias-corrected version of the above bootstrap intervals through two "correction factors" Z and A defined below:.

1. Calculate

$$Z \equiv \Phi^{-1}\left[\frac{\sum_{b=1}^{r} I(\hat{\theta}_b^* < \hat{\theta})}{r}\right]$$

where $\Phi^{-1}(\cdot)$ is the inverse of the standard-normal distribution and $\sum_{b=1}^{r} I(\hat{\theta}_b^* < \hat{\theta})/r$ is the proportion of bootstrap replicates below the estimate $\hat{\theta}$. If the bootstrap sampling distribution is symmetric and if $\hat{\theta}$ is unbiased, then this proportion will be close to 0.5, and the "correction factor" Z will be close to 0.

2. Let $\hat{\theta}_{(-i)}$ represent the value of $\hat{\theta}$ produced when ith observation is deleted from the sample (known as the jackknife values of $\hat{\theta}$). There are n of these quantities. Let $\bar{\theta} = \sum \hat{\theta}_{(-i)}/n$. Then calculate

$$A \equiv \frac{\sum_{i=1}^{n}(\bar{\theta} - \hat{\theta}_{(-i)})^3}{6\left[\sum_{i=1}^{n}(\bar{\theta} - \hat{\theta}_{(-i)})^2\right]^{3/2}}$$

With the correction factors Z and A, compute

$$A_1 \equiv \Phi\left[Z + \frac{Z - z_{\alpha/2}}{1 - A(Z - z_{\alpha/2})}\right]$$
$$A_2 \equiv \Phi\left[Z + \frac{Z + z_{\alpha/2}}{1 - A(Z + z_{\alpha/2})}\right]$$

And the corrected interval is

$$\hat{\theta}^*_{(lower)} < \theta < \hat{\theta}^*_{(upper)}$$

where lower* $= rA_1$ and upper* $= rA_2$ (rounding or interpolating as required).

When the correction factors Z and A are both 0, $A_1 = \Phi(-z_{\alpha/2}) = \alpha/2$ and $A_2 = \Phi(z_{\alpha/2}) = 1 - \alpha/2$.

For the 2000 bootstrap samples that Dr. Fox drew, there are 926 bootstrapped means below $\bar{Y} = 4.6$, and so $Z = \Phi^{-1}(926/2000) = -0.09288$. The $\bar{Y}_{(-i)}$ are $4.444, 5.444, \ldots, 4.111$. And $A = -0.05630$. Using the correction factors Z and A,

$$\begin{aligned} A_1 &= \Phi\left[-0.09288 + \frac{-0.09288 - 1.96}{1 - [-0.05630(-0.09288 - 1.96)]}\right] \\ &= \Phi(-2.414) = 0.007889 \\ A_2 &= \Phi\left[-0.09288 + \frac{-0.09288 + 1.96}{1 - [-0.05630(-0.09288 + 1.96)]}\right] \\ &= \Phi(1.597) = 0.9449 \end{aligned}$$

Multiplying by r, we have $2000 \times 0.007889 \approx 16$ and $2000 \times 0.9449 \approx 1890$, from which

$$\bar{Y}^*_{(16)} < \mu < \bar{Y}^*_{(1890)}$$
$$-0.4 < \mu < 7.3$$

Logistic regression

So far, we only considered cases where the response variable is continuous. Logistic regression belongs in the family of Generalized Linear Model that can be used for analyzing binary responses.

Motivation Let p be the probability of a specific outcome. We are interested in how this probability is affected by the explanatory variables. A naive approach could be:

$$p = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

Problem p must be between 0 and 1.

Solution Model log odds of p (i.e. logit of p) which are defined as

$$\text{odds} = \frac{p}{1-p} \in [0, \infty)$$
$$\text{logit} = \log(\frac{p}{1-p}) \in (-\infty, \infty)$$

This forms the logistic regression

$$\text{logit}(p) = \log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Note that

1. Increase in log odds $\iff$ increase in p.
   Decrease in log odds $\iff$ decrease in p.
2. No ϵ in logistic regression because we observe a binary outcome y_i, not p itself.

The density

$$\begin{aligned} f(y_i|p_i) &= p_i^{y_i}(1-p_i)^{1-y_i} \\ &= e^{y_i \log(p_i) + (1-y_i)\log(1-p_i)} \\ &= e^{y_i \log(\frac{p_i}{1-p_i}) + \log(1-p_i)} \end{aligned}$$

where

$$\text{E}(y_i) = p_i = \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \quad \text{(mean function, inverse link function)}$$
$$\mathbf{x}_i^T \boldsymbol{\beta} = \log(\frac{p_i}{1-p_i}) \quad \text{(logit link function)}$$

We obtain parameter estimates by maximum likelihood. Read page 131 - page 133 of Dr. Hua Zhou's Computational Statistics notes (link) for algorithms to find these MLE (maximum likelihood estimates).