STAT 425 Note 2

Wenxiao Yang*

*Department of Mathematics, University of Illinois at Urbana-Champaign

2021

目录

1 Generalized Least Squares

What do we do if the errors are **correlated** or **heteroscedastic**?
Suppose $\varepsilon \sim \mathcal{N}_n(0, \Sigma)$, where Σ is the variance-covariance matrix.

1.1 GLS, Σ known ($\hat{\beta} = \left(\mathbf{X}^\top \Sigma^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^\top \Sigma^{-1} \mathbf{y}$, $RSS = (\mathbf{y} - \mathbf{X}\beta)^\top \Sigma^{-1} (\mathbf{y} - \mathbf{X}\beta)$)

1.1.1 Method 1: Transform back to OLS

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

where $\varepsilon \sim \mathcal{N}_n(0, \Sigma)$ and Σ is a known, symmetric, positive definite covariance matrix.
When the errors are heteroscedastic or correlated:
Transform this problem back to Ordinary Least-Squares (OLS):
1. Assume S^{-1} exists and write

$$\Sigma = SS^\top$$

(We could use, for example, the Cholesky decomposition from linear algebra to obtain S.)

2. Multiply the model equation by S^{-1} on both sides:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\beta + \varepsilon \\ S^{-1}\mathbf{y} &= S^{-1}(\mathbf{X}\beta + \varepsilon) \\ \underbrace{S^{-1}\mathbf{y}}_{:=\mathbf{y}^*} &= \underbrace{S^{-1}\mathbf{X}}_{:=\mathbf{x}^*}\beta + \underbrace{S^{-1}\varepsilon}_{:=\varepsilon^*} \\ \mathbf{y}^* &= \mathbf{X}^*\beta + \varepsilon^* \end{aligned}$$

This implies that

$$\varepsilon^* \sim \mathcal{N}(S^{-1}\mathbf{0}, \underbrace{S^{-1}\Sigma \left(S^{-1}\right)^\top}_{=\text{Identity}}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$$

since $S^{-1}\Sigma \left(S^{-1}\right)^\top = S^{-1}SS^\top \left(S^{-1}\right)^\top = I$

3. For the transformed model, we can solve for β using OLS:

$$\mathbf{y}^* = \mathbf{X}^*\beta + \varepsilon^*$$

where $\mathbf{y}^* = S^{-1}\mathbf{y}, \mathbf{X}^* = S^{-1}\mathbf{X}$
So, the estimator for β computes as

$$\begin{aligned} \hat{\beta} &= \left(\mathbf{X}^{*\top}\mathbf{X}^*\right)^{-1} \mathbf{X}^{*\top}\mathbf{y}^* \\ &= (\mathbf{X}^\top \underbrace{\left(S^{-1}\right)^\top S^{-1}}_{=\Sigma^{-1}} \mathbf{X})^{-1} \mathbf{X}^\top \underbrace{\left(S^{-1}\right)^\top S^{-1}}_{=\Sigma^{-1}} \mathbf{y} \\ &= \left(\mathbf{X}^\top \Sigma^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^\top \Sigma^{-1} \mathbf{y} \end{aligned}$$

Note that the solution we obtained minimizes:

$$\|\mathbf{y}^* - \mathbf{X}^*\beta\|^2 = (\mathbf{y} - \mathbf{X}\beta)^\top \Sigma^{-1}(\mathbf{y} - \mathbf{X}\beta)$$

1.1.2 Weighted Least Squares (WLS)

Suppose that Σ is a diagonal matrix of unequal error variances:

$$\Sigma = \text{diag}\left(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2\right)$$

The GLS estimate of β minimizes:

$$(\mathbf{y} - \mathbf{X}\beta)^\top \Sigma^{-1}(\mathbf{y} - \mathbf{X}\beta) = \sum_{i=1}^{n} \frac{\left(y_i - \mathbf{x}_i^\top \beta\right)^2}{\sigma_i^2}$$

This problem is known as the Weighted Least-Squares (WLS).
Note that the errors are weighted by

$$w_i = \frac{1}{\sigma_i^2}$$

smaller weights for samples with larger variances.

1.1.3 WLS special example : Replicated Observations

Suppose we collected multiple observations for each $\mathbf{x}_i$. We use double subscripts to indicate the replicate observations:

$$(\mathbf{x}_i, y_{i1}, y_{i2}, \ldots, y_{in_i})$$

Let y_i denote the average of the n_i observations sharing $\mathbf{x}_i$. Then the residual sum of squares for β equals

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(y_{ij} - \mathbf{x}_i^\top \beta\right)^2 = \sum_{i=1}^{n} n_i\left(y_i - \mathbf{x}_i^\top \beta\right)^2 + \sum_{i=1}^{n}\sum_{j=1}^{n_i}(y_{ij} - y_i)^2$$

Minimizing the RSS to solve for β is the same as minimizing the first term on the right only.

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} n_i(y_i - \mathbf{x}_i^T\beta)^2$$

1.1.4 Method 2: Likelihood Estimation

Model: $\mathbf{y} \sim N_n(\mathbf{X}\beta, \Sigma)$
Log-likelihood:

$$\begin{aligned}
&\log(p(\mathbf{y} \mid \beta, \Sigma)) \\
&= \log\left\{\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{y} - \mathbf{X}\beta)^\top \Sigma^{-1}(\mathbf{y} - \mathbf{X}\beta)\right]\right\} \\
&= -\frac{1}{2}(\mathbf{y} - \mathbf{X}\beta)^\top \mathbf{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\beta) + \text{ Constant.}
\end{aligned}$$

Therefore the MLE is given by

$$\hat{\beta}_{mle} = \arg\min_{\beta}(\mathbf{y} - \mathbf{X}\beta)^\top \Sigma^{-1}(\mathbf{y} - \mathbf{X}\beta)$$

1.2 GLS, Σ unknown

1.2.1 Method 1: Estimation of Variance (r_i^2)/Standard Deviation Function ($|r_i|$)

$$\sigma_i^2 = \mathbb{E}(\varepsilon_i^2) - (\mathbb{E}(\varepsilon_i))^2$$

Since we assume $E(\varepsilon_i) = 0$, we have

$$\sigma_i^2 = \mathbb{E}(\varepsilon_i^2)$$

Which implies r_i^2 is an estimator of σ_i^2; $|r_i|$ is an estimator of the standard deviation σ_i.

Estimate Variance Function $\hat{v}_i(x)$

1. Fit a regression model using OLS, and obtain the residuals r_i.
2. Regress the squared residuals r_i^2 against the appropriate predictor variables.

Denote $\hat{v}_i$ be the fitted value from variance function

$$w_i = \frac{1}{\hat{v}_i}$$

Estimate Standard Deviation Function $\hat{s}_i(x)$

1. Fit a regression model using OLS, and obtain the residuals r_i.
2. Regress the absolute residuals $|r_i|$ against the appropriate predictor variables.

Denote $\hat{s}_i$ be the fitted value from standard deviation function

$$w_i = \frac{1}{(\hat{s}_i)^2}$$

The estimated variances are then placed in the variance-covariance matrix Σ and the regression coefficients are estimated using the WLS (Weighted Least Squares method).

1.2.2 Method 2: iterative approach

1. Start with some initial guess of Σ
2. Use Σ to estimate β
3. Use residuals (since we have known β) to estimate Σ
4. Iterate until convergence.

It looks like a good idea; however the methods will not work if we do not assume some structure about Σ (too many parameters to be estimated).

Based on the application, we can assume a particular structure for Σ that does not involve too many parameters.

Then, we can model β and Σ simultaneously.

For example , for AR(1) times series (auto-regressive model of order 1), the structure of Σ would be:

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots \\ \rho & 1 & \rho & \rho^2 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho^{n-1} & \rho^{n-2} & \cdots & \cdots & 1 \end{pmatrix}$$

Σ as a function of ρ and σ^2.

2 Lack of Fit Tests

2.1 Gaussian Assumption

Gaussian Assumption, which can be summarized concisely as:

$$y \sim \mathcal{N}_n(\mathbf{X}\beta, \sigma^2\mathbf{I})$$

Under these assumptions:

$$\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T y \sim \mathcal{N}_p(\beta, \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}),$$
$$\hat{y} = \mathbf{X}\hat{\beta} \sim \mathcal{N}_n(\mathbf{X}\beta, \sigma^2\mathbf{H})$$

independently,

$$\hat{\sigma}^2 = \frac{RSS}{n-p} = \frac{||y-\hat{y}||^2}{n-p} \sim \sigma^2\frac{\chi^2_{n-p}}{n-p}$$

2.2 When σ^2 is known

Intuition:
If the model is correct, then $\hat{\sigma}^2$ is an unbiased estimate of σ^2.
If we know σ^2, we could construct a test based on the ratio $\frac{\hat{\sigma}^2}{\sigma^2}$, a measure of *lack-of-fit*.
In this case we want to test the hypothesis:

$$\begin{cases} H_0: \text{ There is no lack of fit.} \\ H_\alpha: \text{ There is lack of fit.} \end{cases}$$

We use the test statistic:

$$\frac{\hat{\sigma}^2}{\sigma^2} = \frac{RSS/(n-p)}{\sigma^2} \sim \frac{\chi^2_{n-p}}{n-p}$$

Lack of fit means the error variance is large related to the value of σ^2, i.e., the test statistic is large.
Conclude that there is lack of fit (i.e. Reject H_0), if:

$$(n-p)\frac{\hat{\sigma}^2}{\sigma^2} \geq \chi^2_{n-p}(1-\alpha)$$

2.3 When σ^2 is unknown

2.3.1 Hypothesis

If σ^2 is unknown, a general approach is to compare an estimate of σ^2 based on a much bigger/general model. If we can derive the distribution (under H_0) of $\hat{\sigma}^2_{\text{LinearModel}} / \hat{\sigma}^2_{\text{BigModel}}$, then we reduce this problem to a two model comparison test problem.
The null hypothesis is the current model:
$H_0 : \mathbb{E}(y_i) = \mathbf{x}_i^\top \beta, \quad i = 1, 2, \ldots, n, \quad$ for some vector β
The more general model is assumed under the alternative hypothesis:
$H_\alpha : \mathbb{E}(y_i) = f(\mathbf{x}_i), \quad i = 1, 2, \ldots, n, \quad$ for some function f

2.3.2 Under the null hypothesis H_0

$y_{ij} = \mathbf{x}_i^\top \beta + \varepsilon_{ij}$, some $\beta, \varepsilon_{ij} \sim$ iid $\mathcal{N}\left(0, \sigma^2\right)$
RSS_0 with $df = n - p$

2.3.3 Under the alternative big-model hypothesis H_α :

$y_{ij} = f(\mathbf{x}_i) + \varepsilon_{ij}$, some function $f, \varepsilon_{ij} \sim$ iid $\mathcal{N}\left(0, \sigma^2\right)$
Can we estimate σ^2 for the big model in H_α ?
- The answer is yes, if there is some replication in the data, i.e., there are multiple observations (replicates) for some (at least) of the same $\mathbf{x}_i$ values.
- Schematically we can represent these replicates as:

$$(\mathbf{x}_i, y_{i1}, y_{i2}, \ldots, y_{in_i}), \quad i = 1 : m, \quad n = \sum_i n_i$$

RSS_a with $df = n - m = \sum_i (n_i - 1)$, where

$$RSS_a = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2$$

2.3.4 $F-$Test

All of the degrees of freedom for RSS_a come from the replications. Therefore, with replication we can do an F test for lack of fit:

$$F = \frac{(RSS_0 - RSS_a)/(m-p)}{RSS_a/(n-m)} \sim F_{m-p,n-m}$$

3 Polynomials Regression

3.1 Basic Function

$$y_i = f(x_i) + \varepsilon_i$$
$$y_i = \beta_0 + \sum_{j=1}^{d} b_j(x_i)\beta_j + \varepsilon_i$$
$$y_i = \beta_0 + \sum_{j=1}^{d} \beta_j x_i^j + \varepsilon_i$$

d is the **degree of the polynomial component**.

3.2 Choose Order d

1.**Forward Approach**: Keep adding terms until the last added term is not significant.
2.**Backward Approach**: Start with a large d, and keep eliminating the terms that are not statistically significant, starting with the highest order term.

Once we pick up a d, we do not test the significance of the lower-order terms. (include all the lower-order terms in our model by default)
Reasoning: we do not want our results to be affected by a change of location/scale of the data. $(z_i - 2)^2 = z_i^2 - 4z_i + 4$.
Exception: particular polynomial function (physics law).

3.3 Orthogonal Polynomials

Successive predictors x^j are **highly correlated** introducing multicollinearity problems.

$$y_i = \beta_0 + \beta_1 \mathbf{z}_i + ... + \beta_d \mathbf{z}_d + \varepsilon_i$$

where $\mathbf{z}_j = a_1 + b_2 x + ... + \kappa_j x^j$ is a polynomial of order j with coefficients chosen such that $\mathbf{z}_j^T \mathbf{z}_j = 0$

3.4 Piece-wise Polynomials

If the true mean of $E(Y|X = x) = f(x)$ is too wiggly, we might need to fit a higher order polynomial, which is not always a good idea.

Instead we will consider **piece-wise polynomials**:
1.we divide the range of x into several intervals, and
2.within each interval $f(x)$ is a low-order polynomial, e.g., cubic or quadratic, but the polynomial coefficients change from interval to interval;

3.in addition we require the overall $f(x)$ to be continuous up to certain derivatives.

3.5 Cubic Splines

3.5.1 Why Spline

Polynomials: smooth, but each point affects the fit globally.
Piece-wise Polynomials: localizes the influence of each data point, but are not smooth enough.
Splines: combines the beneficail aspects of both approaches.

3.5.2 Settings

A **Cubic Spline** is a curve constructed from sections of cubic polynomials, joined together so that the curve is *continuous up to second derivative*.
The points at which the sections join are called the **knots** of the spline.

We want to define a cubic spline function in the interval $[a, b]$
- Define m knots such that: $a < \xi_1 < \xi_2 < \ldots < \xi_m < b$
- A function g defined on $[a, b]$ is a **cubic spline with respect to knots** $\{\xi_i\}_{i=1}^m$ if:
1. g is a cubic polynomial in each of the $m + 1$ intervals,

$$g(x) = d_i x^3 + c_i x^2 + b_i x + a_i, \quad x \in [\xi_i, \xi_{i+1}]$$

where $i = 0, \ldots, m, \xi_0 = a$ and $\xi_{m+1} = b$
2. g is continuous up to the 2 nd derivative: since g is continuous up to the 2nd derivative for any point inside an interval, it suffices to check the following conditions:

$$g^{(0,1,2)}\left(\xi_i^+\right) = g^{(0,1,2)}\left(\xi_i^-\right), \quad i = 1 : m$$

This expression indicates that the function and the first and second order derivatives are continuous at the knots.

3.5.3 Number of free parameters: $m + 4$

How many free parameters do we need to represent a cubic spline?
(i) 4 parameters (d_i, c_i, b_i, a_i) for each of the $(m + 1)$ intervals.
(ii) 3 constraints at each of the m knots (continuity constraints).
The **total number of free parameters** (similar to the number of degrees of freedom) is:

$$4(m + 1) - 3m = m + 4$$

3.5.4 Properties: linear combination also cubic spline

Given knots $\{\xi_i\}_{i=1}^m$, the linear combinations of cubic splines are also cubic splines.
That is, for a set of given knots, the corresponding cubic splines form a linear space (of functions) with $dim(m+4)$.

3.5.5 Cubic Splines Basis

A set of basis functions for cubic splines (w.r.t knots $\{\xi_i\}_{i=1}^m$) is given by:

$$\begin{aligned} h_0(x) &= 1 \\ h_1(x) &= x \\ h_2(x) &= x^2 \\ h_3(x) &= x^3 \\ h_{i+3}(x) &= (x-\xi_i)_+^3, \quad i = 1, 2, \ldots, m \end{aligned}$$

That is, any cubic spline can be uniquely expressed as:

$$\beta_0 + \sum_{j=1}^{m+3} \beta_j h_j(x)$$

Given knot locations, there are many alternative, but equivalent ways of writing down a basis for cubic splines.

Example 1 (Other Basis)**.** *For example, another basis for cubic splines can be the following:*

$$\begin{aligned} h_0(x) &= 1 \\ h_1(x) &= x \\ h_{i+1}(x) &= R\left(x, \xi_i^*\right), i = 1, \ldots, q-1 \end{aligned}$$

where

$$\begin{aligned} R(x,z) = &\left[(z-1/2)^2 - 1/12\right]\left[(x-1/2)^2 - 1/12\right]/4 \\ &- \left[(|x-z|-1/2)^4 - 1/2(|x-z|-1/2)^2 + 7/240\right]/24 \end{aligned}$$

3.6 Natural Cubic Splines (NCS)

A cubic spline on $[a, b]$ is a **Natural Cubic Spline** if its *second and third derivatives are zero at a and b.*

3.6.1 Degree of Freedom (Number of free parameters): m

This condition implies that NCS is a linear function in the two extreme intervals $[a, \xi_1]$ and $[\xi_m, b]$. The linear functions in the two extreme intervals are completely determined by their neighboring

intervals.
The degree of freedom of NCS's with m knots is:

$$4(m+1) - 3m - 4 = m$$

(We have 4 additional constraints.)

3.6.2 NCS Basis

A Natural Cubic Spline with m knots is represented by m basis functions, for example, one such basis is given by

$$N_1(x) = 1$$
$$N_2(x) = x$$
$$N_{k+2}(x) = d_k(x) - d_{k-1}(x)$$

where

$$d_k(x) = \frac{(x-\xi_k)_+^3 - (x-\xi_m)_+^3}{\xi_m - \xi_k}$$

Each of these derivatives can be seen to have zero second and third derivative for $x \geq \xi_m$.

3.6.3 Note: Waste of Data points

Recall that the **linear functions** in the two extreme intervals are completely determined by the other cubic splines. So data points in the two extreme intervals (i.e., outside the two boundary knots) are wasted since they do not affect the fitting.

3.7 Regression Splines

We can represent the model on the observed n data points using matrix notation:

$$\begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}_{n\times 1} = \begin{pmatrix} h_0(x_1) & h_1(x_1) & \dots & h_{p-1}(x_1) \\ h_0(x_2) & h_1(x_2) & \dots & h_{p-1}(x_2) \\ & & \dots & \\ h_0(x_n) & h_1(x_n) & \dots & h_{p-1}(x_n) \end{pmatrix}_{n\times p} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_p \end{pmatrix}_{p\times 1}$$

where our design matrix is the matrix $\mathbf{F}$ of basis functions.
We can find β by solving the problem:

$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{F}\beta\|^2$$

3.8 $K-$Fold Cross-Validation

How to select the optimal number of knots (or df)?

K-Fold Cross-Validation Steps:

1. Set a fixed number of knots (or df).
2. Divide the set of observations into k groups (or folds).
3. Leave the first fold as a validation set (not used to fit the model). Fit the Regression Spline with a fixed number of knots using the remaining k $-$ 1 folds.
4. Calculate the *Mean Square Error* for fold 1: MSE_1.
5. Repeat the previous steps k times. Each time a new validation set is used to calculate MSE_i.
6. Calculate the average k-fold *Cross-Validation error*:

$$CV(k) = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

7. Repeat 2 to 6 with a new number of knots (or df).
8. Select the number of knots that **minimizes** the k-fold CV error or $CV(k)$.

4 ANOVA: ANalysis of COVAriance

These are regression problems where some predictors are quantitative (i.e. numerical) and some are qualitative (i.e. categorical).

4.1 Two level example

For simplicity, we will focus on examples with just two predictors: X (numerical) and D (categorical). D has two levels: 0 or 1.

General Model

$$y = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 (x \cdot d) + \varepsilon$$

Model 1: Coincident regression lines

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Model 1'

$$y = \beta_0 + \beta_2 d + \varepsilon$$

Model 2: Parallel regression lines

$$y = \beta_0 + \beta_1 x + \beta_2 d + \varepsilon$$

Model 3: Regression lines with equal intercepts but different slopes

$$y = \beta_0 + \beta_1 x + \beta_3 (x \cdot d) + \varepsilon$$

Model 4: Unrelated regression lines

$$y = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 (x \cdot d) + \varepsilon$$

Hierarchical Rule for interactions: *an interaction term will be included in a model only if all its main effects have been included.*
Due to this rule, we would include both β_1 and β_2, once β_3 is significant. So, we don't consider model 3 this place.

4.2 Two-level Test: t-test

Which model to pick?
1. test whether the interaction term is significant:

$$H_0 : \text{model 2} \quad H_\alpha : \text{model 4}$$

2. if don't reject H_0, test whether you can further reduce model 2 to model 1

$$H_0 : \text{model 1} \quad H_\alpha : \text{model 2}$$

4.3 Multi-level example

Model the response Y by two predictors X and D, where X is a numerical variable and D is categorical with k levels.
We need to generate $k-1$ dummy variables: $D_2, \ldots, D_k$ where:

$$D_i = \begin{cases} 0, & \text{if not level i} \\ 1, & \text{if level i} \end{cases}$$

Level 1 is the reference level.
Model 0: $Y \sim 1$
Model 1: $Y \sim X$
Model 1': $Y \sim D$
Model 2: $Y \sim D + X$
Model 4: $Y \sim D + X + D : X$

4.4 Multi-level Test: F-test

Note that when D has more than two levels, the difference between model parameter number may not be one, so t-test is no longer appropriate.
1) Compare models:

$$H_0 : Y \sim X + D \quad \text{vs.} \quad H_\alpha : Y \sim D + X + D : X$$

If the interaction $D:X$ is significant, stop.

2) If X is significant, keep X.

2′) If D is significant, keep D.

3) If neither X nor D are significant, report the intercept model $Y \sim 1$

Sequential ANOVA We can use the anova function to get sequential F-tests. The sequence of F -tests given by anova ($\text{lm}(Y \sim X + D + X:D)$)

H_0	H_α
$Y \sim 1$	$Y \sim X$
$Y \sim X$	$Y \sim X + D$
$Y \sim X + D$	$Y \sim X + D + X:D$

The sequence of F -tests given by anova ($\text{lm}(Y \sim D + X + X:D)$)

H_0	H_α
$Y \sim 1$	$Y \sim D$
$Y \sim D$	$Y \sim D + X$
$Y \sim D + X$	$Y \sim D + X + X:D$

Note: Some of the F-stats and p-values from the sequential ANOVA table are **different** from the ones we calculated based on usual F-test (we learned) for comparing two nested models.
Suppose we want to compare:

$$H_0 : Y \sim X \quad \text{vs} \quad H_\alpha : Y \sim X + D$$

The usual F-stat is given by:

$$\frac{(RSS_0 - RSS_a)/(k-1)}{RSS_a/(n-p_a)} = \frac{(RSS_0 - RSS_a)/(k-1)}{\hat{\sigma}_a^2}$$

which follows $F_{k-1,n-p_a}$ under the null hypothesis. k is the total number of categories of variable D
The F-stat from the sequential ANOVA table:

$$\frac{(RSS_0 - RSS_a)/(k-1)}{RSS_A/(n-p_A)} = \frac{(RSS_0 - RSS_a)/(k-1)}{\hat{\sigma}_A^2}$$

which follows $F_{k-1,n-p_A}$ under the null hypothesis, where RSS_A denotes the RSS from the biggest model $Y \sim X + D + X:D$ and $p_A = 2k$

5 Variable Selection

5.1 Training and Test Errors

Training data: $(\mathbf{x}_i, y_i)_{i=1}^n$
Test data: $(\mathbf{x}_i, y_i^*)_{i=1}^n$ is an independent (imaginary) data set collected at the same location $\mathbf{x}_i$'s (also

known as in-sample prediction)

Assume the data comes from a linear model:

$\mathbf{y}_{n\times 1}, \mathbf{y}^*_{n\times 1}$ are i.i.d $\sim N_n\left(\mu, \sigma^2\mathbf{I}_n\right)$ and $\mu = \mathbf{X}\beta$

We can also write:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$
$$\mathbf{y}^* = \mathbf{X}\beta + \varepsilon^*$$

with $\varepsilon_{n\times 1}, \varepsilon^*_{n\times 1}$ i.i.d $\sim \mathcal{N}_n\left(\mathbf{0}, \sigma^2\mathbf{I}_n\right)$ are independent.

$$\begin{aligned}
\mathbb{E}(\text{ Test Error })^2 &= \mathbb{E}\left\|\mathbf{y}^* - \mathbf{X}\hat{\beta}\right\|^2 \\
&= \mathbb{E}\left\|(\mathbf{y}^* - \mathbf{X}\beta) + (\mathbf{X}\beta - \mathbf{X}\hat{\beta})\right\|^2 \\
&= \mathbb{E}\left\|\mathbf{y}^* - \mu\right\|^2 + \mathbb{E}\|\mathbf{X}\beta - \mathbf{X}\hat{\beta}\|^2 \\
&= \mathbb{E}\left\|\varepsilon^*\right\|^2 + \operatorname{Tr}\left(\mathbf{X}\operatorname{Cov}(\hat{\beta})\mathbf{X}^\top\right) \\
&= n\cdot\sigma^2 + \sigma^2\operatorname{Tr}\mathbf{H} = n\cdot\sigma^2 + p.\sigma^2 \\
\mathbb{E}(\text{ Train Error })^2 &= \mathbb{E}||\mathbf{y} - \hat{\mathbf{y}}\,||^2 = \mathbb{E}||(\mathbf{I}-\mathbf{H})\mathbf{y}\|^2 \\
&= \operatorname{Tr}\left((\mathbf{I}-\mathbf{H})\operatorname{Cov}(\mathbf{y})(\mathbf{I}-\mathbf{H})^\top\right) \\
&= \sigma^2\operatorname{Tr}((\mathbf{I}-\mathbf{H})) = (n-p)\cdot\sigma^2
\end{aligned}$$

Index each model (i.e., each subset of the p variables) by a p-dimensional binary vector γ :

$$\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_p), \quad \gamma_j = 0/1$$

where $\gamma_j = 1$ indicates that X_j is included in the model, and $\gamma_j = 0$ otherwise.

So there are a total of 2^p possible subsets or sub-models. In particular

$$\gamma = (1, 1, \ldots, 1)$$

refers to the full model including all p variables (largest dim), and

$$\gamma = (0, 0, \ldots, 0)$$

refers to the intercept-only model (smallest dim).

Suppose that $\mu = \mathbf{X}\beta$ where μ is the mean of $\mathbf{y}$. If we fit the data $\mathbf{y}$ with respect to model γ, i.e., we fit a linear model with a sub-design matrix X_γ where X_γ contains only columns from X such that $\gamma_j = 1$

We can show that the Testing Error and the Training error for model γ are:

$$\mathbb{E}(\text{ Test Error }) = n\sigma^2 + p\sigma^2 + \text{Bias}_\gamma$$
$$\mathbb{E}(\text{ Training Error }) = n\sigma^2 - p\sigma^2 + \text{ Bias}_\gamma$$

Bigger model (i.e., p large) $\rightarrow$ small Bias, but large variance $\left(p\sigma^2\right)$;

Smaller model (i.e., p small) $\rightarrow$ large Bias, but small variance $\left(p\sigma^2\right)$.

So to reduce the test error (i.e., prediction error), the key is to find the best trade-off between Bias and Variance.

5.2 Model selection procedures

5.2.1 Testing-based procedures

Testing-based procedures: Select best model based on statistical tests for model comparison.

Backward elimination

- Start with all the predictors in the model.

1. Remove the predictor with highest $p-$value $> \alpha_0$ (most insignificant).
2. Refit the model, and repeat the above process.
3. Stop when all $p-$values $\leq \alpha_0$.

(α_0 is often set to 15% or 20% which is higher than usual)

Forward elimination

1. Start with the intercept-only model.
2. For all predictors not in the model, check their p-value if being added to the model. Add the one with the lowest $p-$ value $\leq \alpha_0$ (most significant).
3. Refit the model, and repeat the above process.
4. Stop when no more predictors can be added.

Pros and Cons of Testing-based procedures

- Main advantage: Computation cost is low.
- Due to the "one-at-a-time" nature of adding/dropping variables, this type of procedures does not compare all possible models. So it's possible to miss the "optimal" model.
- It's not clear how to choose α_0, the cut-off for p-values.

5.2.2 Criterion-based procedures

Criterion-based procedures: Select best model based on an information criteria (combining model fit and model complexity) for model comparison.

1. Score each model according to an information criteria
2. Use a searching algorithm to find the optimal model

Model selection criteria/scores often takes the following form:

$$\text{Training error} + \text{Complexity-penalty}$$

Model Selection Criteria:

AIC/BIC

$$AIC : -2 \times \log lik_\gamma + 2p_\gamma$$
$$BIC : -2 \times \log lik_\gamma + \log(n)p_\gamma$$

where p_γ is the number of predictors included in model γ

For the linear regression model:

$$-2 \times \log lik_\gamma = n \log \frac{RSS_\gamma}{n}$$

The lower the AIC/BIC the better. Note that when n is large, adding an additional predictor costs a lot more in BIC than AIC. So AIC tends to pick a bigger model than BIC.

Adjusted $-R^2$ for model γ

$$\begin{aligned} R_a^2 &= 1 - \frac{RSS/\left(n - p_\gamma - 1\right)}{TSS/(n-1)} \\ &= 1 - \left(1 - R^2\right)\left(\frac{n-1}{n - p_\gamma - 1}\right) \\ &= 1 - \frac{\hat{\sigma}_\gamma^2}{\hat{\sigma}_0^2} \end{aligned}$$

The higher the R_a^2 the better.

Mallow's C_p

$$C_p = \frac{RSS_\gamma}{\hat{\sigma}^2} + 2p_\gamma - n$$

where $\hat{\sigma}^2$ is the estimate of the error variance from the full model. Mallow's C_p behaves very similar to AIC.

Searching Algorithms:

Leap and Bounds:

return *the global optimal solution* among all possible models, but *only feasible for less than* 50 *variables*.

–Find the p models with the smallest RSS amongst all models of the same size.

–Then evaluate the score on the p models and report the optimal one.

Greedy algorithms:

fast, but *only return a local optimal solution* (which might be good enough in practice).

- Backward: start with the full model and sequentially delete predictors until the score does not improve.

- Forward: start with the null model and sequentially add predictors until the score does not improve.

- Stepwise: consider both deleting and adding one predictor at each stage.

6 Shrinkage Methods

Find a *trade-off* between *model bias* and *prediction error*.

6.1 Principal Components Regression (PCR)

When we have too many predictors, we need dimensionality reduction in the predictors space. Predictors might be highly correlated.

1. Take matrix X of predictors and center the columns of X to have zero mean. Consider X with no intercept column. (In order to focus on the vaiation).
2. Find directions of greater variation in the data. (找到最能表示 X 的向量)

6.1.1 Principal Component Analysis (PCA)

The steps to find directions of greater variation in matrix $\mathbf{X}$:
- Find $\mathbf{u}_1$ to maximize variance of $\mathbf{u}_1^\top \mathbf{X}$ subject to $\mathbf{u}_1^\top \mathbf{u}_1 = 1$.
- Find $\mathbf{u}_2$ to maximize variance of $\mathbf{u}_2^\top \mathbf{X}$ subject to $\mathbf{u}_1^\top \mathbf{u}_2 = 0$ and $\mathbf{u}_2^\top \mathbf{u}_2 = 1$
- Continue looking for directions of greatest variation in the data which are orthogonal to the previous ones.
- Continue until the total number of dimensions is exhausted. The principal components are given by the columns of matrix $\mathbf{Z}$, where

$$\mathbf{Z} = \mathbf{XU}$$

$$[\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_m] = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m][\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_m]$$

$\mathbf{z}_i$ and $\mathbf{u}_i$ are the columns of $\mathbf{Z}$ and $\mathbf{U}$ respectively. $\mathbf{U}$ is called the **rotation matrix**. $\mathbf{Z}$ is a version of the data rotated in such a way that the resulting principal components are orthogonal.
- Each Principal Component is a linear combination of the original variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ with weights given by each column $\mathbf{u}_i$ of matrix $\mathbf{U}$:

$$\mathbf{z}_i = u_{1i}\mathbf{x}_1 + u_{2i}\mathbf{x}_2 + \ldots + u_{mi}\mathbf{x}_m$$

- Principal components are very sensitive to outliers.
- The **Mahalanobis distance** can be used to measure the distance of a point to the data mean, after adjusting for correlation in the data.
- Under the multivariate normality assumption in m dimensions, the **Mahalanobis distance** $d_i = \sqrt{(\mathbf{x}_i - \mu)^\top \Sigma^{-1} (\mathbf{x}_i - \mu)}$ can be estimated using the sample estimators for μ and Σ and the quantity d_i^2 follows a χ_m^2 distribution. This can be used to detect outliers in higher dimensions.

6.1.2 After PCA

- Replace model $\mathbf{Y} \sim \mathbf{X}$ by the model $\mathbf{Y} \sim \mathbf{Z}$
- Only need to use the first few columns of $\mathbf{Z}$ as predictors
- Interpretation of the PCAs as predictors might be challenging. We need to use the values of $\mathbf{u}_i$ in the rotation matrix (also called the loadings) for interpretation.
- Sometimes we can make better predictions with a small number of PCs in $\mathbf{Z}$ than with a large number of predictors in $\mathbf{X}$

6.1.3 Use How many Principal Components?

- The trace of the sample variance-covariance S of Z (total sample variance) is equal to the sum of its eigenvalues:

$$\text{trace}(S) = s_1^2 + s_2^2 + \ldots + s_m^2 = \lambda_1 + \lambda_2 + \ldots + \lambda_m$$

Since, sample variance-covariance matrix is symmetric, the equation must hold.

- Most of the total variance of a data set is concentrated in the first principal components.

PCs 解释能力逐次递减，我们只需要用前几个就行了：

1. Make a plot of the PCs standard deviations $\left(\sqrt{\lambda_i}\right)$ vs. the PC index i. This is called the scree plot.

- Look for the PC index i where there is a big change in slope (the elbow) in the scree plot.

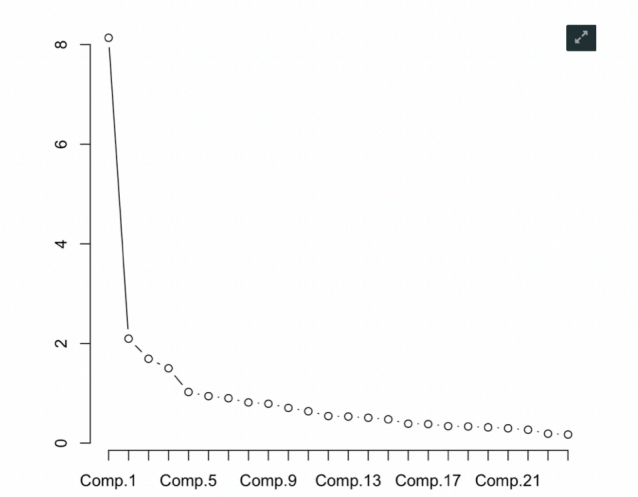

图 1:

Example 2.

第一个 comp 使第一个点到第二个点,..., 第 4 个 comp 使第四个点到第五个点，所以这里我们只需要前 4 个 comp.

2. Another way is to calculate the cumulative variance explained by the first PCs, and retain the number of PC s explaining between 70% to 90% of the total variation.

3. An alternative way is to discard PCs such that $\lambda_i < \bar{\lambda}$

6.2 Ridge Regression

- Although the aim of PCR is to reduce dimensionality in the number of predictors, you still have to measure all the predictors since each PC is a linear combination of all predictors.

- Ridge regression assumes that after normalization, some of the regression coefficients should not be

very large.
- Ridge regression is very useful when you have collinearity and the LS regression coefficients are unstable.
- The method uses a **penalized regression** since the LS minimization problem has a penalty term:

$$\text{minimize}(y - X\boldsymbol{\beta})^\top (y - X\boldsymbol{\beta}) + \lambda \sum_j \boldsymbol{\beta}_j^2$$

for some $\lambda \geq 0$. The penalty term is $\sum_j \beta_j^2$
– Usually predictors are standardized first (centered by their means and scaled by their standard deviations) and the response y is centered.
–The ridge regression estimates are:

$$\hat{\beta} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T y$$

Or, the β minimize

$$(y - X\beta)^T(y - X\beta) \text{ subject to } \sum_j \boldsymbol{\beta}_j^2 \leq t^2$$

- The parameter λ (or t) should be chosen to have stable estimates of β.

- Note that when $\lambda = 0$ the ridge regression estimation problem reduces to the standard LS problem, while when $\lambda \to \infty, \hat{\boldsymbol{\beta}} \to 0$
- It is useful to plot the values of $\hat{\beta}_j$ as a function of λ.
- The value of λ can be also chosen using automated methods as **Generalized Cross-Validation (GCV)** (similar to Cross-Validation).
- Ridge regression coefficient estimates are **biased**.

6.3 Lasso Regression

- In this case the estimated $\hat{\boldsymbol{\beta}}$ minimizes:

$$\text{minimize}(y - X\boldsymbol{\beta})^\top (y - X\boldsymbol{\beta}) + \lambda \sum_j |\beta_j|$$

for some $\lambda \geq 0$. The penalty term is $\sum_j |\beta_j|$ (L_1 constraint)
Or, the β minimize

$$(y - X\beta)^T(y - X\beta) \text{ subject to } \sum_j |\beta_j| \leq t$$

- In two-dimensions the constraint defines a square. In higher dimensions it defines a polytope.
- Lasso is useful when the response can be explained by few predictors with zero effect on the remaining predictors (Lasso is similar to a variable selection method).
- When $\beta_j = 0$ the corresponding predictor is eliminated. This is not the case for ridge regression.

- Use Lasso when the effect of predictors is **sparse**. This means that only few predictors will have an effect on the response (e.g. gene expression data) or when number of predictors is large ($p > n$)
- Use the lars R package for Lasso
- Select t in the constraint $\sum_{j=1}^{p} |\beta|_j \leq t$ by using **Cross-Validation (CV)**
- As t increases, the number of predictors increases.

6.4 Comparing Ridge Regression and Lasso

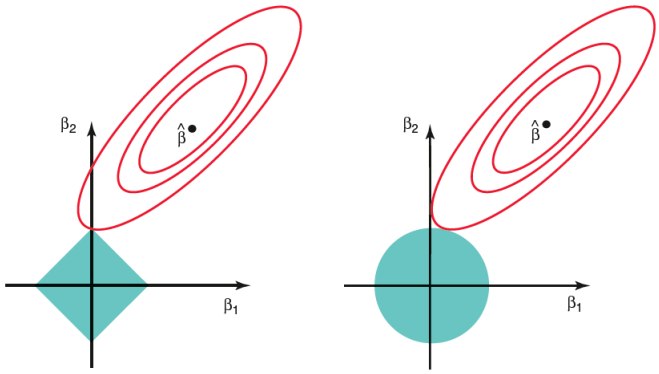

图 2:

2 维下，Lasso 是正方形，Ridge 是圆形。
- Lasso selects a sub-set of predictors (some coefficients equal to zero).
- Ridge regression performs better when the response is a function of many predictors with coefficients around the same size.
- Lasso will perform better when a relatively small number of predictors have large coefficients and the rest are very small or equal to zero.
- Since the number of predictors is never known a priori, cross-validation can be used to decide which approach is better for a particular data set.

7 ANOVA: Comparative Experiments

7.1 Terminology

- **Factor**: an Independent variable. They can be experimental or observational. In our example: Diet
- **Level**: A particular form of the factor. In our example: Levels of the Diet: A, B, C, D
- **Treatments**: Factor levels or factor level combinations (if the study contains more than one factors). They provide insights into mechanisms causing the variation being studied. Control treatments?

- **Complete Randomized Design**: Experimental units are randomly split into r groups, and r treatments are assigned, one per group.

7.2 Data

$$\begin{array}{llllll} \text{group } 1 & y_{11}, & y_{12} & \cdots & y_{1n_1} \\ \text{group } 2 & y_{21}, & y_{22} & \cdots & y_{2n_2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \text{group } r & y_{r1}, & y_{r2} & \cdots & y_{rn_r} \end{array}$$

- r is the number of groups
- n_i denotes the number of obs in the ith group
- $n = \sum_{i=1}^{r} n_i$ is the total sample size
- y_{ij} = observation j for the ith factor.

7.3 ANOVA Model

7.3.1 ANOVA Means Model (Cell Means Model)

$$y_{ij} = \mu_i + \varepsilon_{ij}, i = 1, \ldots, r; \quad j = 1, \ldots, n_i$$

- y_{ij} : the value of the response in the j th trial for the ith factor.
- μ_i : the population mean for the ith factor level (treatment).
- $\varepsilon_{ij} \sim^{iid} \mathcal{N}\left(0, \sigma^2\right)$

7.3.2 Factor Effects Model

Define the effect of factor level i on the response, i.e. the treatment effect as

$$\alpha_i = \mu_i - \mu$$

where μ is the overall mean.
Factor Effects Model:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, i = 1, \ldots, r; j = 1, \ldots, n_i$$
$$\varepsilon_{ij} \sim^{iid} \mathcal{N}\left(0, \sigma^2\right)$$

- The factor effects model has $r + 1$ model parameters, i.e.

$$(\mu, \alpha_1, \ldots, \alpha_r)$$

- In order for the α 's to be (uniquely) estimated, we need to impose restrictions.
- The restrictions on the α 's depend on how μ is defined.

Model	μ Definition	α 's Restriction
Reference Cell	$\mu = \mu_1 \quad \alpha_1 = 0$	
Sum-to-Zero	$\mu = \frac{1}{r}\sum_i \mu_i$	$\sum_i \alpha_i = 0$
Weighted Sum-to-Zero	$\mu = \frac{1}{n}\sum_i n_i \mu_i$	$\sum_i n_i \alpha_i = 0$

- The default in R is the Reference Cell model.

7.4 Model Properties

$- E(y_{ij}) = \mu_i$

$- \text{Var}(y_{ij}) = \text{Var}(\varepsilon_{ij}) = \sigma^2$

Thus, all observations have the same variance, regardless of factor level.

- $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ and independent
- $y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$ and independent.

We can re-state the model as

$$y_{ij} \text{ are independent } \mathcal{N}(\mu_i, \sigma^2)$$

7.5 Model Estimation

Minimize the sum of squared deviations of the observations around their expected values with respect to the parameters:

$$Q = \sum_{i=1}^{r}\sum_{j=1}^{n_i} (y_{ij} - \mathbb{E}(y_{ij}))^2$$

If we re-write Q we have

$$Q = \sum_j (y_{1j} - \mu_1)^2 + \sum_j (y_{2j} - \mu_2)^2 + \ldots + \sum_j (y_{rj} - \mu_r)^2$$

So the **least squares estimator** of μ_i, denoted by $\hat{\mu}_i$ is

$$\hat{\mu}_i = \bar{y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$$

Using the appropriate constraints, we can easily extract the estimators for μ and α_i.

- The LS fit for y_{ij} is the corresponding group mean

$$\hat{y}_{ij} = \bar{y}_i$$

- Residuals

$$r_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_i$$

- RSS

$$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2$$

i.e. the within-group variation.

Source of Variation	SS	df	MS
Between Groups	$FSS=\sum n_i\left(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot}\right)^2$	$r-1$	$\frac{FSS}{r-1}$
Error (within Groups)	$RSS=\sum\sum\left(y_{ij}-\bar{y}_{i\cdot}\right)^2$	$n-r$	$\frac{RSS}{n-r}$
Total	$TSS=\sum\sum\left(y_{ij}-\bar{y}_{\cdot\cdot}\right)^2$	$n-1$	

7.6 F-test

- We want to test whether the means of the groups are really different. We can express this as

$$\begin{cases} H_0:\mu_1=\mu_2=\ldots=\mu_r \\ H_\alpha:\ \text{not all } \mu_i, i=1,\ldots,r \text{ are equal} \end{cases}$$

- or in terms of models

$$\begin{cases} H_0: y_{ij}=\mu+\varepsilon_{ij} \\ H_\alpha: y_{ij}=\mu+\alpha_i+\varepsilon_{ij} \end{cases}$$

- They are two nested models, so we can use the F-test

$$\frac{\left(RSS_0-RSS_\alpha\right)/(r-1)}{RSS_\alpha/(n-r)}\sim F_{r-1,n-r}$$

under H_0.

- The test statistic can also be written as

$$\frac{FSS/(r-1)}{RSS/(n-r)}=\frac{\text{Between-group Variation }/(r-1)}{\text{Within-group Variation }/(n-r)}$$

where FSS, RSS are defined in the ANOVA table.

7.7 Diagnostics for ANOVA Models

–Check for outliers/ unusual observations.

–Check the residuals vs. fitted values plot for departures from the constant variance assumption.

–Check the Q-Q plot for departures from the normality assumption.

Levene's Test for Equality of Variances:

–Run Regression abs(residuals) X, i.e. use abs(residuals) as the response in a new one-way ANOVA.

–If the p-value for the F-test is **greater** than 1% level, then we conclude that there is no evidence of a non-constant variance.

7.8 Inference for Factor Level Means (function about the μ_is)

7.8.1 A single factor level mean

- Estimation of

$$\mu_i : \hat{\mu}_i = \bar{y}_i$$

- Distribution of

$$\hat{\mu}_i : E(\hat{\mu}_i) = \mu_i, \quad \text{Var}(\hat{\mu}_i) = \frac{\sigma^2}{n_i}$$

- The estimated variance of $\bar{y}_i.$ is

$$s^2_{\bar{y}_i} = \frac{1}{n_i} \cdot \frac{RSS}{n-r}$$

- Under the ANOVA model assumptions $\frac{\bar{y}_i. - \mu_i}{s_{\bar{y}_i}}$ is distributed as T_{n-r}
- Confidence Interval for μ_i :

$$\mu_i \in \bar{y}_i. \pm T_{n-r}(\alpha/2) s_{\bar{y}_i}$$

7.8.2 A difference between two factor level means

The difference between two factor level means (pairwise comparison) is defined as

$$D = \mu_i - \mu_{i'}$$

- Estimation of D:

$$\hat{D} = \bar{y}_i - \bar{y}_{i'}$$

- Distribution of $\hat{D}$:

$$E(\hat{D}) = \mu_i - \mu_{i'}, \text{Var}(\hat{D}) = \sigma^2 \left(\frac{1}{n_i} + \frac{1}{n_{i'}} \right)$$

The estimated variance of $\hat{D}$ is

$$s^2_{\hat{D}} = \frac{RSS}{n-r} \cdot \left(\frac{1}{n_i} + \frac{1}{n_{i'}} \right)$$

- Under the ANOVA model assumptions

$$\frac{\hat{D} - D}{s_{\hat{D}}} \text{ is distributed as } T_{n-r}$$

- Confidence Interval for D:

$$D \in \hat{D} \pm T_{n-r}(\alpha/2) s_{\hat{D}}$$

- Hypothesis Test for D :

$$\begin{cases} H_0 : \mu_i = \mu_{i'} \\ H_\alpha : \mu_i \neq \mu_{i'} \end{cases} \Leftrightarrow \begin{cases} H_0 : \mu_i - \mu_{i'} = D = 0 \\ H_\alpha : \mu_i - \mu_{i'} \neq 0 \end{cases}$$

The test statistic is

$$t = \frac{\hat{D}}{s_{\hat{D}}} \sim T_{n-r}$$

7.8.3 A contrast among factor level means

A contrast is a comparison involving two or more level means:

$$L = \sum_{i=1}^{r} c_i \mu_i, \quad \text{where } \sum_{i=1}^{r} c_i = 0$$

- Estimation of L:

$$\hat{L} = \sum_{i=1}^{r} c_i \bar{y}_i$$

- Distribution of $\hat{L}$:

$$E(\hat{L}) = \sum_{i=1}^{r} c_i \mu_i, \mathrm{Var}(\hat{L}) = \sigma^2 \sum_{i=1}^{r} \frac{c_i^2}{n_i}$$

The estimated variance of $\hat{L}$ is

$$s^2_{\hat{L}} = \frac{RSS}{n-r} \cdot \sum_{i=1}^{r} \frac{c_i^2}{n_i}$$

- Under the ANOVA model assumptions $\frac{\hat{L}-L}{s_{\hat{L}}}$ is distributed as T_{n-r}
- Confidence Interval for L:

$$L \in \hat{L} \pm T_{n-r}(\alpha/2) s_{\hat{L}}$$

- Hypothesis Testing for L :

$$\begin{cases} H_0 : L = 0 \\ H_\alpha : L \neq 0 \end{cases}$$

The test statistic is

$$t = \frac{\hat{L}}{s_{\hat{L}}} \sim T_{n-r}$$

7.8.4 A linear combination of factor level means

$$L = \sum_{i=1}^{r} c_i \mu_i, \quad \text{no restrictions on } c_i' s$$

- Point estimator and estimated variance same as before.
- Single Degree of Freedom Tests

$$\begin{cases} H_0 : L = c \\ H_\alpha : L \neq c \end{cases}$$

The test statistic here is

$$F = t^2 = \left(\frac{\hat{L} - c}{s_{\hat{L}}} \right)^2 \sim F_{1,n-r}$$

7.9 Limitations of Inference Procedures

The confidence coefficient $1 - \alpha$ for the estimation procedures described is a statement confidence coefficient and applies **only to a particular estimate**, **not to a series of estimates**.
Similarly the specified Type I error rate α applies only to a particular test and not to a series of tests.

7.10 Bonferroni Correction $\frac{\alpha}{m}$

Example 3. *If the confidence coefficents are* 95% *for all individual* μ_i*, the confidence coefficient will be* $(95\%)^n < 95\%$ *for family* $f(\mu_1, ..., \mu_n)$*.*
So the 95% *confidence interval of family will be* ***wider*** *than individual.*

When? The family of interest is a particular set of pairwise comparisons, contrasts, or linear combinations that is specified by the user.
- Suppose m is the number of statements in the family.
- In order to control the family wise error rate to be α, we need to reduce the error rate for each individual comparison to be α/m.
- That is we need to increase the significance level from $(1-\alpha)$ to $(1-\alpha/m)$
- Not applicable when m is large, since the Cls would be too wide due to the increase of the significant level.

i.e. 我们用 $1 - \frac{5\%}{n}$ for all μ_i 组成 95% 的 $f(\mu_1, ..., \mu_n)$

7.11 Tukey's Paired Comparison Procedures

When? the family of interest is a set of all pairwise comparisons of factor level means, i.e. it consists of estimates of all pairs $D = \mu_i - \mu_{i'}$
A confidence interval is given by

$$D \in \hat{D} + \frac{q(\alpha/2; r, n-r)}{\sqrt{2}} s(\hat{D})$$

where $q(\alpha/2; r, n-r)$ refers to the $\alpha/2$ upper quantile of the studentized range for r means and $n-r$ degrees of freedom.
The coverage probability is exact when the sample sizes in each group are identical and is approximate otherwise.
Remark: The studentized range refers to the distribution of

$$\max_{i \neq j} \sqrt{n} \left(\bar{y}_i - \bar{y}_j\right) / \hat{\sigma}$$

where $\bar{y}_i$ and $\bar{y}_j$ are sample means from independent samples of size n from normal distributions with common means and variance σ^2.
Note: Tukey is always better than Bonferroni and Scheffe in pairwise comparisons.

7.12 Scheffe's Method for Contrasts

When? The family of interest is the set of contrasts among the factor level means:

$$L = \sum c_i \mu_i, \text{ where } \sum c_i = 0$$

An confidence interval is given by

$$L \in \hat{L} + (r-1) F_{r-1, n-r}(\alpha) s_{\hat{L}}$$

8 Two Way ANOVA

Single-factor:

1. Do not explore the entire space of treatment combinations.
2. Interactions cannot be estimated.
3. Full randomization is not possible.
4. Multiple stages increase complexity of the analysis.

MultiFactor:

1. Efficient replication.
2. Assessment of Interactions.
3. Validity of Findings.

8.1 Factor Effects Model for Two Factors

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

	Sum	Average
Cell (i,j)	$y_{ij.} = \sum_{k=1}^{n} y_{ijk}$	$\bar{y}_{ij.} = \frac{y_{ij}}{n}$
Row i	$y_{i..} = \sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}$	$\bar{y}_{i..} = \frac{y_{i..}}{bn}$
Column j	$y_{j.j} = \sum_{i=1}^{a}\sum_{k=1}^{n} y_{ijk}$	$\bar{y}_{j.} = \frac{y_{j.j}}{an}$
Overall	$y_{...} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}$	$\bar{y}_{...} = \frac{y_{...}}{nab}$

- Using least squares method, the estimated treatment means are:

$$\hat{\mu}_{ij} = \bar{y}_{ij}$$

- The factor effects estimators depend on the constraints that we impose. For example, under the sum-constraints we have

$$\hat{\alpha}_i = \bar{y}_{i..} - \bar{y}\ldots, \quad \hat{\beta}_j = \bar{y}_{.j.} - \bar{y}\ldots$$
$$\hat{(\alpha\beta)}_{ij} = \bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}..$$

- The fitted values and residuals compute as usual as

$$\hat{y}_{ijk} = \bar{y}_{ij}, \quad r_{ij} = y_{ijk} - \hat{y}_{ijk}$$

8.2 Interaction Plots

If the lines are not parallel, interaction is presented.

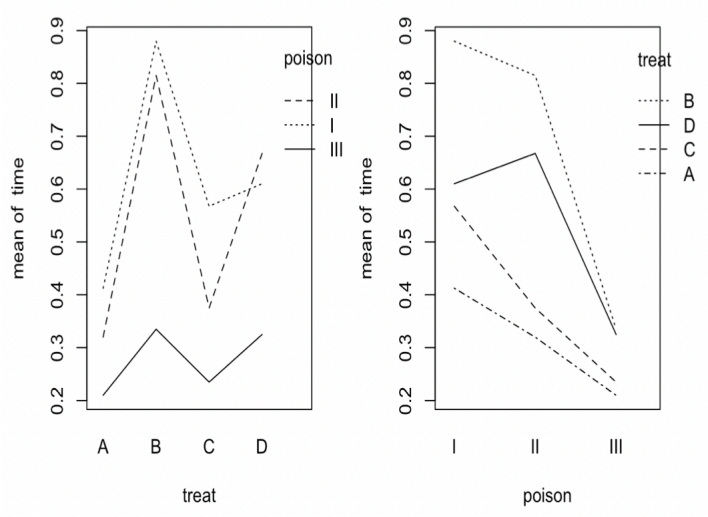

图 3:

8.3 Partitioning of Total Sum of Squares

$$\underbrace{y_{ijk} - \bar{y}_{...}}_{\text{Total Deviation}} = \underbrace{\bar{y}_{ij} - \bar{y}_{...}}_{\substack{\text{Deviation of estimated} \\ \text{treatment mean around} \\ \text{overall mean}}} + \underbrace{y_{ijk} - \bar{y}_{ij.}}_{\substack{\text{Deviation} \\ \text{around estimated} \\ \text{treatment mean}}}$$

$$TSS = FSS + RSS$$

where

$$TSS = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{...})^2$$

$$FSS = n \sum_i \sum_j (\bar{y}_{ij\cdot} - \bar{y}_{...})^2$$

$$RSS = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij.})^2 = \sum_i \sum_j \sum_k e_{ijk}^2$$

8.4 Partitioning of Treatment Sum of Squares

$$\underbrace{\bar{y}_{ij.} - \bar{y}_{...}}_{\substack{\text{Deviation of estimated} \\ \text{treatment mean around} \\ \text{overall mean}}} = \underbrace{\bar{y}_{i..} - \bar{y}_{...}}_{\substack{\text{A main} \\ \text{effect}}} + \underbrace{\bar{y}_{.j.} - \bar{y}_{...}}_{\substack{\text{B main} \\ \text{effect}}} + \underbrace{\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}}_{\substack{\text{A B interaction} \\ \text{effect}}}$$

$$FSS = SSA + SSB + SSAB \text{ (Orthogonal Decomposition)}$$

where

$$SSA = nb \sum_i (\bar{y}_{i..} - \bar{y}\ldots)^2$$

$$SSB = na \sum_j (\bar{y}_{.j.} - \bar{y}\ldots)^2$$

$$SSAB = n \sum_i \sum_j (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}\ldots)^2$$

8.5 ANOVA Table

Source of Variation	SS	df	MS
Factor A	SSA	$a-1$	MSA $= \frac{SSA}{a-1}$
Factor B	SSB	$b-1$	MSB $= \frac{SSB}{b-1}$
AB Interactions	SSAB	$(a-1)(b-1)$	MSAB $= \frac{SSAB}{(a-1)(b-1)}$
Error	RSS	$ab(n-1)$	MSE $= \frac{RSS}{ab(n-1)}$
Total	TSS	nab -1	

8.6 F-test

- In order to test for the statistical significance of the interaction terms, we use partial F-tests. So, we fit a main effects model (i.e. no interactions)

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$$

- Then, we compare the two nested models:

$$\begin{cases} H_0: \text{ smaller model with } p_0 \text{ coefficients} \\ H_\alpha: \text{ larger model with } p_\alpha \text{ coefficients} \end{cases}$$

The F-test is formulated as

$$F = \frac{(RSS_0 - RSS_\alpha) \,/\, (p_\alpha - p_0)}{MSE_\alpha} \sim F_{p_\alpha - p_0, n - p_\alpha} \text{ under the } H_0$$

We can also perform F-tests directly using the ANOVA table, where for the interaction term we have:

$$F_{AB} = \frac{MSAB}{MSE} \sim F_{(a-1)(b-1), nab-1}$$

Hierarchy principle: we test for main effects only if the interaction term is not statistically significant.

8.7 Estimation of Factor Level Means

When interactions are not statistically significant, we analyze the factor level means:
- Factor Level Means:

$$\hat{\mu}_{i.} = \bar{y}_{i..}, s^2_{\hat{\mu}_{i.}} = \frac{MSE}{bn}$$

- Differences of Factor Level Means:

$$\hat{\mu}_{i\cdot} - \hat{\mu}_{i'\cdot} = \bar{y}_{i\cdot\cdot} - \bar{y}_{i'\cdot\cdot}, s^2_{\hat{D}} = \frac{2MSE}{bn}$$

- Contrasts of Factor Level Means:

$$\hat{L} = \sum c_i \hat{\mu}_{i\cdot} = \sum c_i \bar{y}_{i\cdot\cdot\cdot}, s^2_{\hat{L}_{(i)}} = \frac{MSE}{bn} \sum c_i^2$$

where $\sum c_i = 0$

For individual hypothesis test and Cls, the multiplier is $T_{(n-1)ab}(\alpha/2)$.

For family hypothesis tests/intervals, we select the desired family multiplier:

- Tukey Multiplier:

$$\frac{1}{\sqrt{2}} q_{a,(n-1)ab}(1-\alpha)$$

- Bonferroni Multiplier: $B = T_{(n-1)ab}(1-\alpha/2m)$, where m refers to the number of multiple comparisons.

- Scheffé Multiplier:

- $S^2 = (b-1)F_{b-1,(n-1)ab}(1-\alpha)$, if the contrasts involve the $\mu_{i\cdot}$ and

- $S^2 = (a-1)F_{a-1,(n-1)ab}(1-\alpha)$, if the contrasts involve the $\mu_{\cdot j}$.

8.8 Estimation of Treatment Means

When interactions are statistically significant, we analyze the treatment means:

- Treatment Means:

$$\hat{\mu}_{ij} = \bar{y}_{ij\cdot}, \quad s^2_{\hat{\mu}_{ij}} = \frac{MSE}{n}$$

- Differences of Treatment Means:

$$\hat{D} = \hat{\mu}_{ij} - \hat{\mu}_{i'j'} = \bar{y}_{ij\cdot} - \bar{y}_{i'j'\cdot}, \quad i,j \neq i',j' \text{ and } s^2_{\hat{D}} = \frac{2MSE}{n}$$

- Contrasts of Treatment Means:

$$\hat{L} = \sum\sum c_{ij}\hat{\mu}_{ij} = \sum\sum c_{ij}\bar{y}_{ij}, \text{ where } \sum\sum c_{ij} = 0$$

with variance $s^2_{\hat{L}} = \frac{MSE}{n}\sum c_{ij}^2$

For individual hypothesis test and Cls, the multiplier is $T_{(n-1)ab}(\alpha/2)$.

For family hypothesis tests/intervals, we choose the desired family multiplier:

- Tukey Multiplier: $\frac{1}{\sqrt{2}} q_{ab,(n-1)ab}(1-\alpha)$

- Bonferroni Multiplier: $B = T_{(n-1)ab}(1-\alpha/2m)$, where m refers to the number of multiple comparisons.

- Scheffé Multiplier: $S^2 = (ab-1)F_{ab-1,(n-1)ab}(1-\alpha)$, if the contrasts involve the $\mu_{i\cdot\cdot}$

9 Two Way ANOVA: Special Cases

9.1 Unbalanced ANOVA (Use Partial $F-$Test or ANOVA type III in R)

– **When the treatment sample sizes are unequal**, the analysis of variance for two-factor studies becomes more complex.
– The least-squares equations are no longer of a simple structure and the regular analysis of variance formulas are now inappropriate.
– Furthermore, the factor effect component sum of squares are no longer orthogonal; that is, they **do not sum up to TSS**.

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

– Due to the lack of orthogonality, the **ANOVA F-tests are not applicable**.

9.1.1 Use Partial $F-$Test

– We will express the ANOVA model as a regression model with **indicator (dummy) variables**.
– We need $a-1$ indicator variables for factor A main effects and $b-1$ indicator variables for factor B main effects. The interactions correspond to the cross products of the indicator variables for A and B.
Use **Partial $F-$Test**.

9.1.2 ANOVA type III in R

This type tests for the presence of an effect given that both the other effects are in the model.

```
Anova(lm(1/time ~ treat*poison, data=newrats), type="III")
```

```
## Anova Table (Type III tests)
##
## Response: 1/time
##               Sum Sq Df F value    Pr(>F)
## (Intercept)  15.0605  1 66.5967 1.298e-09 ***
## treat         2.1340  3  3.1455   0.03723 *
## poison       11.7375  2 25.9514 1.225e-07 ***
## treat:poison  1.9800  6  1.4592   0.22073
## Residuals     7.9151 35
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

图 4:

Example 4. *Compute SSTotal*

$$
\begin{aligned}
RSS_0 - RSS_\alpha = &(SSA + SSR) - SSR \quad &= SSA = 2.1340 \\
&(SSB + SSR) - SSR \quad &= SSB = 11.7375 \\
&(SSAB + SSR) - SSR \quad &= SSAB = 1.9800 \\
& & SST = 2.1340 + 11.7375 + 1.9800 + 7.9151 = 23.7666
\end{aligned}
$$

9.2 Balanced ANOVA with $n = 1$ (Tukey's Test)

- Only one observation in each cell, so we cannot fit the interaction model.
- There are no degrees of freedom left for estimating the error.
- RSS = 0 when the model includes main effects and interaction term. (error is 0)
- All F-tests are valid, but the interaction model is not a candidate model.

9.2.1 Tukey's Test for Additivity

Consider the following model that includes interactions:

$$y_{ij} = \mu + \alpha_i + \beta_j + \theta\alpha_i\beta_j + \varepsilon_{ij}$$

Here, we assume that the interactions are of *multiplicative* nature, i.e.

$$(\alpha\beta)_{ij} = \theta\alpha_i\beta_j$$

- Consider the SSA, SSB as before and:

$$SSAB^* = \frac{\left(\sum_i \sum_j (\bar{y}_i - \bar{y}_{..})(\bar{y}_{.j} - \bar{y}_{..})\, y_{ij}\right)^2}{\sum_i (\bar{y}_{i.} - \bar{y}_{..})^2 \sum_j (\bar{y}_{.j} - \bar{y}_{..})^2}$$

- The TSS is computed as usual and is decomposed as

$$TSS = SSA + SSB + SSAB^* + SSRem^*$$

where the remainder is

$$SSRem^* = TSS - SSA - SSB - SSAB^*$$

- We want to test the following hypothesis

$$\begin{cases} H_0 : \theta = 0 & \text{(no interactions)} \\ H_\alpha : \theta \neq 0 & \text{(interactions)} \end{cases}$$

which is essentially a test for model additivity. - The test statistic computes as

$$F^* = \frac{SSAB^*/1}{SSRem^*/(ab - a - b)}$$

10 Introduction to Experimental Designs

10.1 Experimental vs. Observational Study

Experimental Study is a scienti c procedure undertaken to make a discovery, test a hypothesis or verify a claim.
Observational Study is one in which the experimenter observes the effect of a factor on the response, or measures an outcome without an attempt to a ect the outcome by intervention.

10.2 Principles of Experimental Design

Randomization
- Random allocation of treatment and order
- Ensures that collected data are IID random variables
- Averages-out the effects of exogenous factors.

Replication
- Estimate of experimental error
- Higher Precision

Blocking
- Higher precision when comparisons of factors are made
- Reduced variability transmitted from nuisance factors.

10.3 Randomization Test

- A test based directly on re-randomizing - with the same kind of randomization originally used to assign the treatments - is called a **randomization test**.

- *Advantage*: No need for any distributional assumptions (independence, normality, etc.) - just need to assume that the treatment randomization was performed properly.
- *Disadvantage*: Requires more computation, and you must implement for yourself or use specialized software.

11 Blocking in Experimental Designs

11.1 Randomized Complete Block Design (RCBD) Model

Example:

- Treatment Factor ‘variety‘: 8 levels
- Block Factor ‘block‘: 5 levels
- Observe that we have *only one* observation per treatment-block combination.

```
xtabs(yield~variety + block, oatvar)
```

```
##         block
## variety   I  II III  IV   V
##       1 296 357 340 331 348
##       2 402 390 431 340 320
##       3 437 334 426 320 296
##       4 303 319 310 260 242
##       5 469 405 442 487 394
##       6 345 342 358 300 308
##       7 324 339 357 352 220
##       8 488 374 401 338 320
```

图 5:

Suppose that there are r treatments (factor levels) and n_b blocks.

$$y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$$

where
- μ. is a constant
- τ_i are the treatment effects
- β_j are the block effects
- ε_{ij} are independent $\mathcal{N}\left(0, \sigma^2\right)$
- $i = 1, \ldots, n_b$ (total number of blocks), $j = 1, \ldots, r$ (total number of treatments)

Remarks
- y_{ij} is the response for the j th treatment in the ith block.
- There is a single observation per block. This implies that we have a limited ability to detect an interaction between treatment and block. So, we are working with the additive model.
- We can check for treatment and block main effects, but blocking is a feature of the design which means that if insignificant, we cannot gain the degrees of freedom.

ANOVA Display for the RCBD (Two factors: Treatments and Blocks)

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Treatments	$SS_{\text{Treatments}}$	$r-1$	$\frac{SS_{\text{Treatments}}}{r-1}$	$\frac{MS_{\text{Treatments}}}{MS_g}$
Blocks	SS_{Blocks}	n_b-1	$\frac{SS_{\text{Blocks}}}{n_b-1}$	
Error	SS_E	$(r-1)(n_b-1)$	$\frac{SS_g}{(r-1)(n_b-1)}$	
Total	SS_T	rn_b-1		

11.2 Latin Squares

11.2.1 Example

- A, B, C: 3 treatments
- Day of week (Monday, Wednesday, Friday): Blocking Variable
- Operator ID: 1, 2, 3: Blocking Variable

	Operator		
Day	1	2	3
Monday	B	A	C
Wednesday	A	C	B
Friday	C	B	A

* Each operator runs each treatment, and all treatments are run on each day

11.2.2 Features of a Latin Square Design

- There are r treatments.
- There are 2 blocking variables, each containing r classes.
- Each row and each column in the design square contains all treatments.
- Each treatment is assigned to each block only once.

Advantages
- Reduces more experimental error than with 1 blocking factor.
- Small scale studies can isolate important treatment features.
- Repeated measures designs can remove order effects.

Disadvantages
- Each blocking factor must have r levels.
- No interactions among factors.
- With small r, we have very few error degrees of freedom.
- Complex Randomization.

11.2.3 Randomization in Latin Square Designs

- Determine r, the number of treatments, row blocks, and column blocks.
- Select a Standard Latin Square (from tables or with software).
- Use Capital Letters to represent treatments $(A, B, C, \ldots)$ and randomly assign treatments to labels.
- Randomly assign Row Block levels to Square Rows.
- Randomly assign Column Block levels to Square Columns.

11.2.4 Latin Square Model

$$y_{ijk} = \mu + \tau_i + \beta_j + \gamma_k + e_{ijk}$$

where
- μ is a constant
- τ_i treatment effect (latin letter)
- β_j (column) blocking effect
- γ_k (row) blocking effect
- e_{ijk} are independent $\mathcal{N}\left(0, \sigma^2\right)$
- $i, k, j = 1, \ldots, r$

NOVA Display for the Latin Square Model (Three factors: Treatments, Rows and Cols)

AOV	*df*
Rows (blocks)	$r-1$
Cols (blocks)	$r-1$
Treatments	$r-1$
Error	$(r-1)(r-2)$
Total	$\left(r^2-1\right)$

11.3 Balanced Incomplete Block Design (BIBD)

- [→] Why Balanced?

Each pair of treatments occur together λ times.

- [→] Why Incomplete?

Cannot fit all treatments in each block.

Notation
- t treatments
- b blocks
- k treatments per block (block size)

- r times each treatment occurs
- $N = t \cdot r = b \cdot k$ observations in total

- Treatment i occurs in r blocks.
- To have balance, each other treatment is equally likely to be treatment i in a block.
- Since there are $k-1$ other units in a block and $t-1$ other treatments, the number of times each pair of treatments appears in the same block is

$$\begin{aligned}\lambda &= \text{一个 block 中剩余的 units 可能存在 treatment j 的概率} \times \text{treatment i 总共存在的 block 数量} \\ &= \frac{k-1}{t-1} r = \frac{r(k-1)}{t-1}\end{aligned}$$

where λ is an integer.

Examples
- $t = 3, b = 3, k = 2 \to r = 2, \lambda = 1$.
$t =\#$ treatments $\{A, B, C\} = 3$
The form of square: $k \times b = 2 \times 3$.
$r =\#$ A occurs$=\#$ B occurs$=\#$ C occurs$= 2$
$\lambda =\#$ A and B in one block$=\#$ A and C in one block$=\#$ B and C in one block$= 1$

Block		
1	2	3
A	B	A
B	C	C

- $t = 4, k = 2, b = 6 \to r = 3, \lambda = 1$

Block					
1	2	3	4	5	6
A	A	A	B	B	C
B	C	D	C	D	D

11.4 BIBD Remarks

Advantages:
- A BIBD enables us to run an experiment when the size of the available blocks of experimental units is smaller than the number of treatments.
- Estimates of treatment effects have equal precision and expressions for the variances of the estimated cell means and of contrasts of treatment means or effects are relatively simple.
- The presence of balance permits the use of Scheffé and Tukey procedures for the analysis of treatment effects.

Disadvantages:
- BIBD exist only for certain combinations of numbers of treatments, block sizes, and numbers of blocks.
- The assumption that there are no interactions between the blocking variable and the treatments is restrictive.
- The analysis of a BIBD is more complex than that of a RCBD.

$$y_{ij} = \mu + \tau_i + \beta_j + e_{ij}$$

- μ constant
- τ_i treatment effects
- β_j the block effects
- e_{ij} independent $N\left(0, \sigma^2\right)$

Remarks
- Not all y_{ij} exist, because of incompleteness.
- Non-orthogonality of treatments and blocks.

12 Linear Models with Random Effects

12.1 Random Effects

One-way ANOVA model

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Previously, we assumed the α_i s were parameters: fixed, unknown values. (We also gave a restriction, to make them identifiable.) These are fixed effects, corresponding to a fixed factor.
Now suppose that the α_i s are unobserved random variables:

$$\alpha_i \sim N\left(0, \sigma_\alpha^2\right)$$

and assume they are independent of each other and of the ε_{ij} s.
These α_i s are called random effects, and the corresponding factor variable is a random factor.

- **Fixed effects** are appropriate when the levels of the factor are individually important or meaningful (e.g. treatments in a designed experiment, levels of education).
- **Random effects** are appropriate when the levels of the factor are meaningful only as representatives of a more general collection (e.g. as if sampled from a population, or representative of some hypothetical population).

- For the one-way ANOVA model with a random factor, the random effects satisfy

$$\mathrm{E}(\alpha_i) = 0 \quad \mathrm{Var}(\alpha_i) = \sigma_\alpha^2$$

so random effects contribute only to the variance structure of the model, not to the mean structure.
- The parameter σ_α^2 is generally unknown, and we usually seek to estimate it (or σ_α) and test the null hypothesis $\sigma_\alpha^2 = 0$.
Parameters like σ_α^2 (and σ^2) are called **variance components**.

12.2 Intraclass Correlation

- Under this model, the responses can be correlated:

$$\mathrm{Cov}\left(y_{ij}, y_{ij'}\right) = \sigma_\alpha^2 \quad \text{for } j \neq j'$$

So different observations from the same "class" (same level i of the random factor) may have a nonzero correlation.
- The **intraclass correlation coefficient (ICC)** is the correlation between y_{ij} and $y_{ij'}$ (for any i and $j \neq j'$):

$$\rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma^2}$$

12.3 Mixed Models

- A **fixed effects model** has only fixed factors.
- A **random effects model** has only random factors.
- A **mixed (effects) model** has both fixed and random factors.
The general form (matrix-vector):

$$\mathrm{y} = \mathrm{X}\beta + \mathrm{Z}_\gamma + \varepsilon$$

where **X** and **Z** are known design matrices, β contains the fixed effect (mean-related) parameters, and

$$\gamma \sim N\left(0, \sigma^2 D\right) \quad \text{independent of} \quad \varepsilon \sim N\left(0, \sigma^2 I\right)$$

are the random effects and the errors, with D containing the (unknown) random effect parameters.

The variance components (in D) are typically estimated via one of three different methods:
- **ANOVA estimation**, based on quantities in an ANOVA table; complicated for general models
- **maximum likelihood**
- **restricted maximum likelihood (REML)**, generally less biased than maximum likelihood
For balanced data, REML and ANOVA estimation tend to coincide.

R package lme4 with mixed model function lmer

```
library(lme4)

## Loading required package:  Matrix
```

REML estimation

```
milk.reml = lmer(bac ~ (1|shipment), data=milk)
```

图 6:

12.4 REML: restricted maximum likelihood

```
summary(milk.reml)

## Linear mixed model fit by REML ['lmerMod']
## Formula: bac ~ (1 | shipment)
##    Data: milk
##
## REML criterion at convergence: 184.5
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.66731 -0.60005  0.06269  0.50602  2.14554
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  shipment (Intercept) 29.74    5.454
##  Residual             22.29    4.721
## Number of obs: 30, groups:  shipment, 5
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   15.167      2.587   5.863
```

图 7:

- The formula bac (1|shipment) specifies the one-way random effects ANOVA model. As usual, there is an automatically-added intercept (representing μ), and the term (1|shipment) represents the random effect term α_i.
- Function Imer uses the REML method, by default. We see that the REML estimates of the variance components are

$$\widehat{\sigma}_{\alpha}^2 \approx 29.74 \quad \widehat{\sigma}^2 \approx 22.29$$

(The Std.Dev. column simply gives the square roots of these: $\widehat{\sigma}_{\alpha}$ and $\widehat{\sigma}$.)
- The only fixed effect is the intercept, μ.

```r
# ML estimation

milk.ml = lmer(bac ~ (1|shipment), data=milk, REML=FALSE)
```

图 8:

12.5 ML Estimation

```r
summary(milk.ml)

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: bac ~ (1 | shipment)
##    Data: milk
##
##      AIC      BIC   logLik deviance df.resid
##    194.1    198.3    -94.1    188.1       27
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.61636 -0.63533  0.08278  0.48918  2.19649
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  shipment (Intercept) 23.05    4.801
##  Residual             22.29    4.721
## Number of obs: 30, groups:  shipment, 5
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   15.167      2.314   6.555
```

图 9:

- So the MLE for σ_α^2 is

$$\widehat{\sigma}_\alpha^2 \approx 23.05$$

smaller than the REML estimate 29.74. MLEs for variance components are often biased low.
- The estimate listed for σ^2 is apparently not the MLE, but is still the REML estimate.
- In this case, the estimate of μ has remained unchanged, but its standard error has changed.

12.6 Testing and Confidence Intervals

- For the fixed effects (in β), likelihood ratio tests are available. (For this to work, the variance components should be estimated with MLE, not REML.)
(These LRTs are sometimes unreliable, so a parametric bootstrap approach can be used - later.)
- The methods of generalized least squares (F-tests and t-tests) could alternatively be used (though

this would ignore the additional uncertainty of replacing $\mathbf{D}$ with $\widehat{\mathbf{D}}$).
- There are also confidence intervals for fixed effect parameters based on the Wald approach or (perhaps more reliably) on profile likelihood.

12.7 Testing the Random Effect Variance

- For the random effects, the null hypothesis is usually that a variance component equals zero.
- For technical reasons, the usual chi-square approximation in the LRT often fails to be adequate (most often leading to a test that is too conservative).
- An improvement is to use the parametric bootstrap to perform the LRT (see example later).
- Methods based on ANOVA are also available, and can be useful in single-factor or balanced cases.
- Profile likelihood confidence intervals for variance components can be computed (but may have problems, as the LRT does).

12.8 Parametric Bootstrap

The parametric bootstrap may be more accurate in small samples. Here are the steps:
1. Compute the LR statistics for the null and alternative models
2. Generate data under the null hypothesis model
3. Fit the null and alternative model for the generated data
4. Compute the LR statistic
5. Repeat steps 2 to 4 many times
6. Find the Bootstrap probability of exceeding the observed LR value