STAT 210
Applied Statistics and Data Analysis
Experimental Design I

Joaquin Ortega

Design of Experiments

Design of Experiments

Based partly on

Chapter 11, Ugarte, Militino and Arnholt, *Probability and Statistics with R*, Chapman & Hall 2008.

and

Chapter 2, J. Lawson, *Design and Analysis of Experiments with R*, Chapman and Hall 2015.

Design of Experiments

Experiments are usually carried out to determine the effect of a factor or factors in some response of interest.

We will assume that the response is a continuous variable.

The factors are independent variables whose levels are set by the experimenter. They may be categorical variables or continuous variables that have been set to a fixed number of discrete levels.

These different configurations of the factors are known as treatments, and they are applied to experimental units.

The response variable is measured for each treatment, and the objective is to compare the observed responses.

When two or more factors are involved, the experiment is known as a factorial design.

Design of Experiments

For example, a researcher may be interested in the effect of adding a certain amount of carbon on steel hardness, and the experiment is designed so that three different amounts are considered.

In this case, the treatments are the three levels of carbon that the experimenter is interested in comparing, carbon is the factor being considered, and strength is the response.

In the event a second factor is of interest, such as temperature with two levels, the experiment will consist of $2 \times 3 = 6$ treatment combinations and is known as a factorial design.

Design of Experiments

Suppose first that the experimenter has one furnace to carry out the experiment.

There may be variations in the performance of the furnace at different times.

To minimize their effect, the researcher should randomly assign the order in which the experiments are run.

By randomizing the assignment of treatments, the possibility of confounding differences due to levels of carbon with differences due to the performance of the furnace is minimized.

When the assignment of treatments is done in a completely random way, we speak of a *completely randomized design* (CRD).

Design of Experiments

Suppose now that the experimenter has two furnaces to run the experiment.

Even though the furnaces may have the same characteristics, there may be operational differences that may affect the result of the experiment.

In this case, experiments run on the same furnace may be more homogeneous than experiments run on different furnaces.

Experiments run on the same furnace are grouped into **blocks**

In this case, treatments are randomly assigned to experimental units within a block, i.e., the order in which experiments are performed is assigned randomly *for each furnace*.

Such a design is known as a randomized complete block design (RCBD).

Design of Experiments

- **Treatments** are levels of a factor or combinations of factor levels the experimenter wants to compare.
- **Experimental units** are anything to which treatments are applied, for example, animals, plots, plants, or people.
- **Responses** are outcomes observed after the application of a treatment to an experimental unit.
- **Experimental error** is random variation present in the experiment, not under the control of the experimenter. Experimental error may be due to many things, including but not limited to: measurement error, different responses from measuring the same quantity in separate trials, and different responses from experimental units given the same treatment.
- **Treatment structure** specifies the set of factors the experimenter has selected to study or compare.

Design of Experiments

- **Design structure** defines how experimental units are assigned to treatment groups.
- **Randomization** is the use of some well-defined probabilistic mechanism to assign treatments to experimental units. Randomization reduces the possibility of bias and confounding. Randomization should also be used, if possible, with any variable not under the direct control of the experimenter that may influence the measured response.
- **Replication** is the independent assignment of several experimental units to each treatment (factor combination), resulting in independent observations. Replication shows the results are reproducible and allows the experimenter to estimate the experimental error. When the number of experimental units is the same for all treatments, the design is referred to as a balanced design. Unbalanced designs do not have an equal number of experimental units for all treatments.

Design of Experiments

When all treatment levels are fixed by the experimenter, we speak of a **fixed effects** model or experiment.

If treatment levels are random, we talk of **random effects**.

Experiments which include both kinds are **mixed-effects** experiments.

The fixed effects models assume:

1. The measured responses are independent of one another.
2. The model errors are independent of one another and follow a normal distribution.
3. The variance is homogeneous across treatments.

Design of Experiments

The experimenter seeks a model that explains how the response varies in terms of the independent variables in the experiment, the predictors.

There may be more than one model that adequately describes this relation. In this case, we are guided by the principle of parsimony (Occam's razor), and we should choose the simplest model that reasonably describes the experimental results.

Models are expressed in `R` with the syntax

```r
response ~ predictors
```

Design of Experiments

Finding an adequate model is an iterative process that starts by

1. Identifying an appropriate model based on the treatment and design structure of the experiment.
2. Validating the model's assumptions using diagnostic plots.
3. Selecting a different model or transforming the response variable when the model's assumptions are not satisfied until a plausible model is found.

Once a model has been validated, formal inference to test for no treatment effects (equality of treatment means) and estimation of the model's parameters can be undertaken.

If the analysis shows that not all treatment means are equal, multiple comparisons may be used to determine which treatment produces the 'best' result.

A General Model for One-way Anova.

A General Model for One-way Anova.

Suppose we have an experiment where one factor or treatment T has k levels and there are r_i replications for configuration $i, i = 1, \ldots, k$.

By this, we mean that for each configuration, the experiment is repeated r_i times, and the results obtained are the replications.

It does not mean that the experiment is performed once, and r_i measurements are taken.

A General Model for One-way Anova.

Let y denote the response variable, then a model for this experiment would be

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}; \qquad j = 1, \ldots, r_i, i = 1, \ldots k, \tag{1}$$

where

- y_{ij} represents the outcome of the j-th replication of level i,
- μ is the overall mean,
- τ_i is the **effect** of treatment i and
- ϵ_{ij} represents the error for the j-th replication of level i.

A General Model for One-way Anova.

The errors are assumed to be independent and normally distributed with mean 0 and equal variance σ^2.

In this model, μ is the global mean for the experiment, and $\mu + \tau_i$ represents the average response for the *i*-th group.

(1) is usually known as the *effects model*.

The total number of data points is $n = \sum_{i=1}^{k} r_i$.

A General Model for One-way Anova.

An alternative model for this experiment would be

$$y_{ij} = \mu_i + \epsilon_{ij}; \qquad j = 1, \ldots, r_i, i = 1, \ldots k, \tag{2}$$

where μ_i represents the average response for level i of the treatment factor and the same assumptions are made for the errors ϵ_{ij}.

Comparing μ_s with μ_t is equivalent to comparing τ_s with τ_t:

$$\mu_t - \mu_s = (\mu + \tau_t) - (\mu + \tau_s) = \tau_t - \tau_s$$

Model (2) is sometimes known as the *cell means model*. The two models are equivalent.

A General Model for One-way Anova.

These models can be used in two different scenarios.

When the experimenter specifically chooses the treatments, and there is no desire to extend the results to other treatments, the model is referred to as a **fixed effects model**.

When the treatments are selected at random from a larger population of possible treatments and the experimenter would like to extend the conclusions of the experiment to all treatments in the population, the model is called a **random effects model**.

We will only consider the fixed effects model.

Least Squares Estimation

Least Squares Estimation

Least square estimators for the one-way Anova model are values for the parameters $\hat{\mu}, \hat{\tau}_1, \ldots, \hat{\tau}_k$ that minimize the error sum of squares

$$\sum_{i=1}^{k} \sum_{j=1}^{r_i} \epsilon_{ij}^2 = \sum_{i=1}^{k} \sum_{j=1}^{r_i} (y_{ij} - \mu - \tau_i)^2. \tag{3}$$

The resulting model $y_{ij} = \hat{\mu} + \hat{\tau}_i$ is the best-fitting model in the sense of minimizing (3).

Least Squares Estimation

The procedure for minimizing this expression is the usual. The expression in (3) is differentiated with respect to the parameters $\mu, \tau_1, \ldots, \tau_k$ in turn and each of the resulting expressions is set equal to zero, yielding a set of $k+1$ equations. These are known as the *normal equations*.

It is an exercise in calculus to verify that these equations are

$$y_{\bullet\bullet} - n\hat{\mu} - \sum_{i=1}^{k} r_i \hat{\tau}_i = 0, \tag{4}$$

$$y_{i\bullet} - r_i(\hat{\mu} + \hat{\tau}_i) = 0, \quad i = 1, \ldots, k, \tag{5}$$

where the hat notation indicates that these are the values that minimize (3).

Least Squares Estimation

From (5) we get that

$$\hat{\mu} + \hat{\tau}_i = \frac{1}{r_i} y_{i\bullet} = \bar{y}_{i\bullet}$$

for $i = 1, \ldots, k$, so the least squares estimate for the i-th treatment mean is the corresponding sample mean $\bar{y}_{i\bullet}$.

However, there is a problem with the normal equations. If we add up the equations in (5) we get (4). The k equations in (5) are linearly independent, but if we add (4) we get an undetermined system of equations that does not have a unique solution.

This means that the $k + 1$ parameters in model (1) are not all estimable.

Least Squares in Matrix Notation

Least Squares in Matrix Notation

Consider model (1) with $k = 3$ factor levels and $r = 2$ replicates for each level. We can write the effects model using matrix notation as

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\epsilon} \qquad (6)$$

where

$$\mathbf{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{31} \\ \epsilon_{32} \end{pmatrix}.$$

Least Squares in Matrix Notation

The least squares estimators are the solution to the normal equations $\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{y}$. The problem is that the matrix $\mathbf{X}'\mathbf{X}$ is singular and cannot be inverted.

The R function `lm` makes the matrix $\mathbf{X}$ full rank by dropping the column that corresponds to the first level of the factor:

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}.$$

Least Squares in Matrix Notation

This coding makes the first level of the treatment the standard, and all other levels are compared to it.

For example, with $k = 3$ levels the solution to the normal equations is

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\mu} + \hat{\tau}_1 \\ \hat{\tau}_2 - \hat{\tau}_1 \\ \hat{\tau}_3 - \hat{\tau}_1 \end{pmatrix}. \tag{7}$$

Least Squares in Matrix Notation

This is equivalent to adding the equation $\hat{\tau}_1 = 0$ to the equations in (5).

There are other alternatives. For practical purposes, any one of the infinite number of solutions will be satisfactory, since they lead to identical solutions for the estimable parameters.

In fact, any extra equation can be added, provided that it is not a linear combination of the equations already present.

The trick is to add whichever equation will aid most in solving the entire set of equations.

Least Squares in Matrix Notation

A common solution is obtained by adding the extra equation $\sum_i r_i \hat{\tau}_i = 0$. In this case the normal equations become

$$\begin{aligned} \sum_i r_i \hat{\tau}_i &= 0 \\ y_{\bullet\bullet} - n\hat{\mu} &= 0 \\ y_{i\bullet} - r_i(\hat{\mu} + \hat{\tau}_i) &= 0, \quad i = 1, \dots, k \end{aligned}$$

from which we get the least squares solutions

$$\hat{\mu} = \bar{y}_{\bullet\bullet}, \qquad \hat{\tau}_i = \bar{y}_{i\bullet} - y_{\bullet\bullet}.$$

Least Squares in Matrix Notation

Parameter	Estimator
μ	$\overline{Y}_{\bullet\bullet}$
μ_i	$\overline{Y}_{i\bullet}$
τ_i	$\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}$
ϵ_{ij}	$Y_{ij} - \overline{Y}_{i\bullet}$
σ^2	$\frac{1}{N-k}\sum_{i=1}^{k}\sum_{j=1}^{r_i}(Y_{ij} - \overline{Y}_{i\bullet})^2$

Example

Example

A tire manufacturer is interested in investigating the braking performance for different types of tread patterns.

There are four different tread patterns identified with the letters `A`, `B`, `C`, and `D`. Six measurements were taken with each one.

Measurements (`StopDist`) correspond to the braking distance in feet of a medium sized car from a speed of 60 miles per hour.

The same driver and car were used for all the experiments.

The order of the treatments was assigned at random.

Example

The data can be found in the `Tire` file in the `PASWR` package.

```
library(PASWR)
str(Tire)
```

```
## 'data.frame':    24 obs. of  2 variables:
##  $ StopDist: int  391 374 416 363 353 381 394 413 398 39
##  $ tire    : Factor w/ 4 levels "A","B","C","D": 1 1 1 1
```

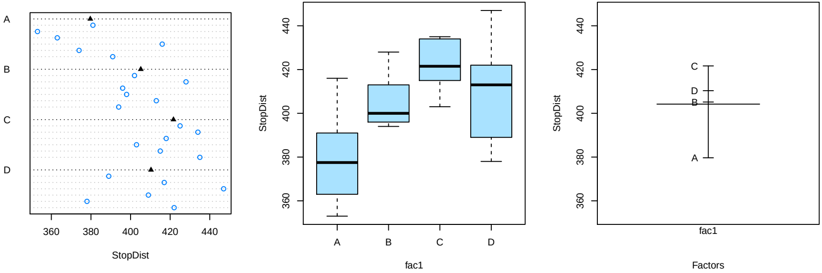
```
head(Tire, n=4)
```

```
##   StopDist tire
## 1      391    A
## 2      374    A
## 3      416    A
## 4      363    A
```

Example

For an initial graphical exploration of the results we use the function `oneway.plots()` from the `PASWR` package.

```
with(Tire, oneway.plots(StopDist, tire))
```



Example

We now use the `lm` function to fit the linear model. The results are stored in the file `mod0`.

```r
mod0 <- lm(StopDist ~ tire, data = Tire)
```

Example

```r
summary(mod0)
```

```
##
## Call:
## lm(formula = StopDist ~ tire, data = Tire)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.333  -9.667  -2.250  11.417  36.667
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  379.667      7.691  49.363  < 2e-16 ***
## tireB         25.500     10.877   2.344 0.029497 *
## tireC         42.000     10.877   3.861 0.000973 ***
## tireD         30.667     10.877   2.819 0.010594 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.84 on 20 degrees of freedom
## Multiple R-squared:  0.4442, Adjusted R-squared:  0.3608
## F-statistic: 5.328 on 3 and 20 DF,  p-value: 0.007316
```

Example

To interpret these results recall from (7) that the default coding for treatment in `R` means that the first value (`Intercept`) corresponds to the average value for the first treatment level, $\hat{\mu} + \hat{\tau}_1$ while `tireB`$= \hat{\tau}_2 - \hat{\tau}_1$, `tireC`$= \hat{\tau}_3 - \hat{\tau}_1$, and `tireD`$= \hat{\tau}_4 - \hat{\tau}_1$. Thus

$$\begin{aligned}
\hat{\mu} + \hat{\tau}_1 &= 379.7; \\
\hat{\mu} + \hat{\tau}_2 &= 379.7 + 25.5 = 405.2 \\
\hat{\mu} + \hat{\tau}_3 &= 379.7 + 42.0 = 421.7 \\
\hat{\mu} + \hat{\tau}_4 &= 379.7 + 30.7 = 410.4
\end{aligned}$$

.

Variance Estimation

Variance Estimation

The **residuals** $\hat{\epsilon}_{ij}, j = 1 \ldots, r_i, i = 1, \ldots, k$ are defined as

$$\hat{\epsilon}_{ij} = y_{ij} - (\hat{\mu} + \hat{\tau}_i)$$

and represent the difference between the j-th replication of the i-th treatment and the estimated treatment mean $\hat{\mu} + \hat{\tau}_i = \bar{y}_{i\bullet}$.

The sum of squares for error or error sum of squares is

$$\begin{aligned} SSE &= \sum_i \sum_j \hat{\epsilon}_{ij}^2 = \sum_i \sum_j (y_{ij} - (\hat{\mu} + \hat{\tau}_i))^2 \\ &= \sum_i \sum_j (y_{ij} - \bar{y}_{i\bullet})^2 \qquad (8) \\ &= \sum_i \sum_j y_{ij}^2 - \sum_i r_i \bar{y}_{i\bullet}^2. \end{aligned}$$

Variance Estimation

The sample variance for the *i*-th treatment is given by

$$\hat{\sigma}_i^2 = \frac{1}{r_i - 1} \sum_{j=1}^{r_i} (y_{ij} - \bar{y}_{i\bullet})^2 \tag{9}$$

which is an unbiased estimator for the common variance of the errors: $E(\hat{\sigma}_i^2) = \sigma^2$. From (9) we get that

$$\sum_{j=1}^{r_i} (y_{ij} - \bar{y}_{i\bullet})^2 = (r_i - 1)\hat{\sigma}_i^2,$$

and using this in (8)

$$SSE = \sum_{i=1}^{k} (r_i - 1)\hat{\sigma}_i^2$$

Variance Estimation

Taking expectations we get

$$E(SSE) = \sum_{i=1}^{k}(r_i - 1)E(\hat{\sigma}_i^2) = \sigma^2 \sum_{i=1}^{k}(r_i - 1) = \sigma^2(n - k)$$

and we conclude that

$$\hat{\sigma}^2 = \frac{SSE}{n - k} = MSE \tag{10}$$

is an unbiased estimator for the variance σ^2.