One-Way ANOVA
Factor effect model

One-way Analysis of Variance (ANOVA)

- Y is a continuous variable (just like regular regression)
- X is a categorical variable with $r \geq 2$ distinct values

- In ANOVA terminology, X is a *factor* with r *levels*
- Typically, the levels represent different groups, subpopulations, or treatments
- Because X is no longer continuous, our model no longer expresses $\hat{Y}$ as a smooth function of X. We are now interested in *differences* among the mean responses for the various factor levels.

Relation between Regression and ANOVA

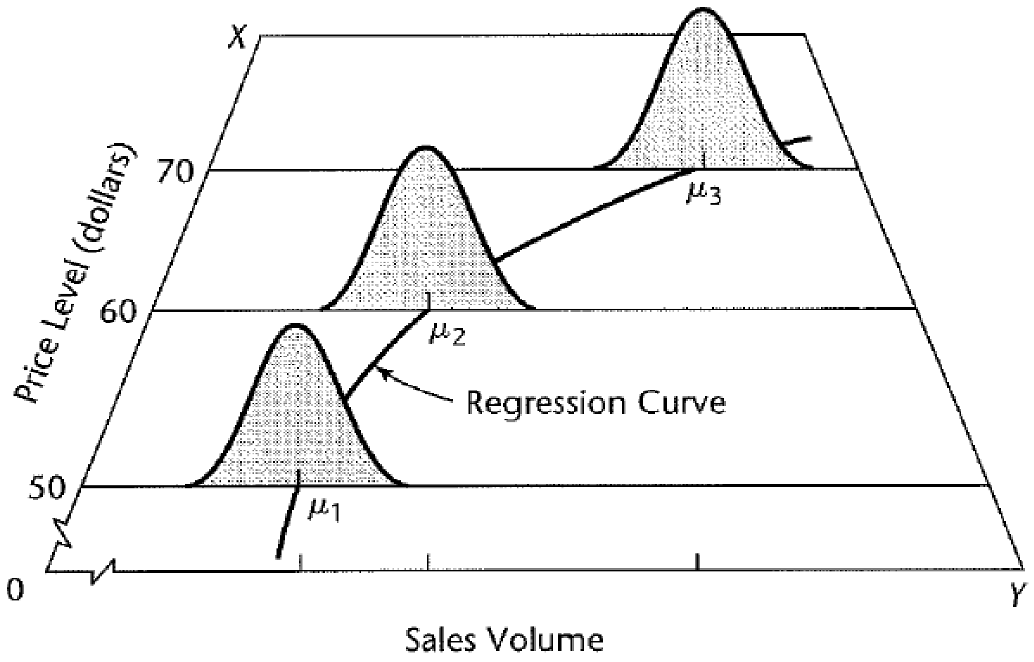
In regression, we aimed to estimate the parameters of a deterministic equation that expressed the conditional expectation of Y as a function of X.

In ANOVA, our common objectives are slightly different:

1. Determine whether any differences in $\mathrm{E}(Y)$ exist among the factor levels
2. Determine which specific factor levels differ from each other
3. Estimate the differences in $\mathrm{E}(Y)$ among various levels

   (or equivalently, estimate the population means for Y within different levels)

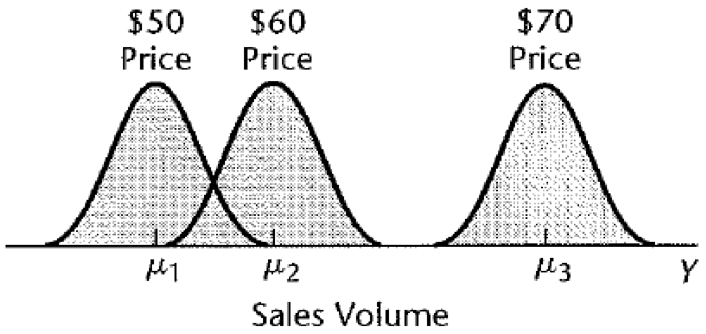
(a) Regression Model

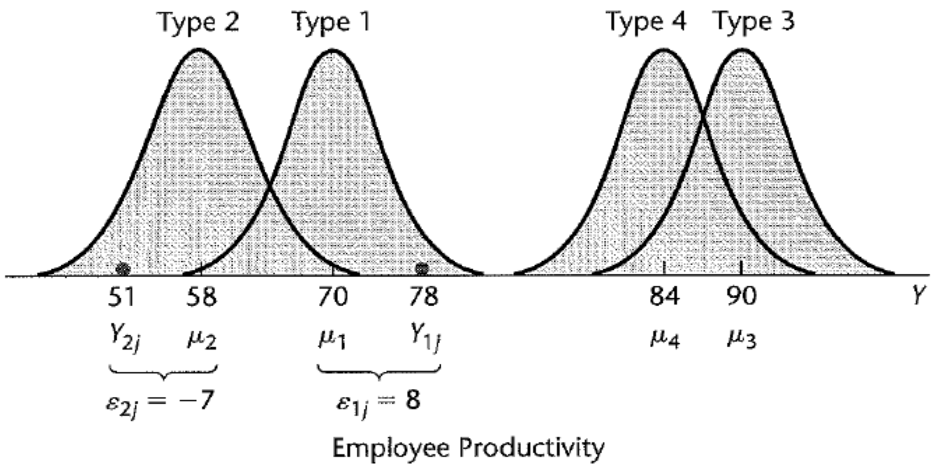


$$\varepsilon = Y - \hat{Y} = \mathrm{Y} - \mathrm{X}\beta$$

(b) Analysis of Variance Model

No assumptions is made about the nature of regression function





The Cell means model

- Until now, we have used the index i to represent individual cases in the data.
- For ANOVA, use i to represent a factor level, $i = 1, \ldots r$
- Individual cases within each level are represented by the index j, $j = 1,2, \ldots, n_i$
- Y_{ij} is the value of the response for the j-th individual in factor level i.
- We will also (eventually) transition away from representing parameters as β to representing them as μ or τ

The Cell means model

$$Y_{ij} = \mu_i + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim Normal(0, \sigma^2)$$

$$E(Y_{ij}) = \mu_i \qquad \sigma(Y_{ij}) = \sigma^2 \qquad Y_{ij} \sim Normal(\mu_i, \sigma^2)$$

$\bar{Y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{i,j}$ is the sample mean for observations from level i.

$\bar{Y}_{\cdot\cdot} = \frac{1}{n_T} \sum_{i=1}^{r} \sum_{j=1}^{n_i} Y_{i,j}$ is the mean over all of the observations.

$n_T = \sum_{i=1}^{r} n_i$ is the total sample size.

The Cell means model is an essentially a linear model, $Y = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

For example, if $r = 3, n_1 = n_2 = n_3 = 2, n = 6$

$$\mathbf{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{bmatrix}$$

$$\mathbf{E}\{\mathbf{Y}\} = \begin{bmatrix} E\{Y_{11}\} \\ E\{Y_{12}\} \\ E\{Y_{21}\} \\ E\{Y_{22}\} \\ E\{Y_{31}\} \\ E\{Y_{32}\} \end{bmatrix} = \mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_1 \\ \mu_2 \\ \mu_2 \\ \mu_3 \\ \mu_3 \end{bmatrix}$$

$$\boldsymbol{\sigma}^2\{\mathbf{Y}\} = \boldsymbol{\sigma}^2\{\boldsymbol{\varepsilon}\} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}$$

The Cell means model is a linear model, $Y = X\mu + \varepsilon$, *with* ***no intercepts***

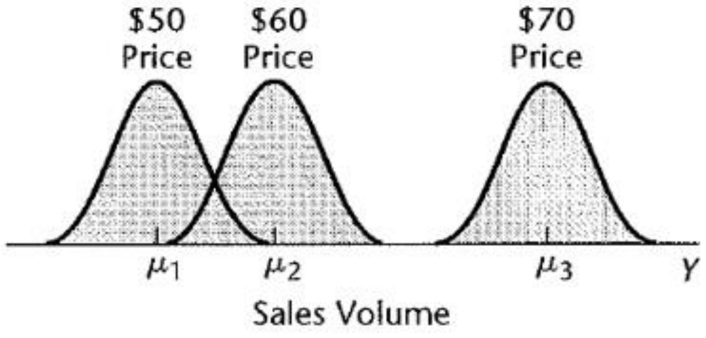
For example, if $r = 3, n_1 = n_2 = n_3 = 2, n = 6$

$$\begin{bmatrix} Y_{1,1} \\ Y_{1,2} \\ Y_{2,1} \\ Y_{2,2} \\ \vdots \\ Y_{r,n_r} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} \cdot \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_r \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,1} \\ \varepsilon_{1,2} \\ \varepsilon_{2,1} \\ \varepsilon_{2,2} \\ \vdots \\ \varepsilon_{r,n_r} \end{bmatrix}$$

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\mu} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I})$$

The model assumes that,

- The errors (and therefore the observations) are independent
- The errors are normally distributed (but the CLT still applies)
- The errors have constant variance
- Subpopulations associated with different levels of the factor *might* have different mean responses



Estimation for the cell means model

Because X is discrete, estimates for the ANOVA model can be computed without linear algebra by using the standard equations for the sample mean and sample variance.

For example, minimize $Q_i = \Sigma(Y_{ij} - \mu_i)^2$ with respect to μ_i

For each level i, the true *within-group mean,* μ_i, is estimated as,

$$\hat{\mu}_i = \bar{Y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{i,j}$$

and the *within-group sample variance* is

$$s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(Y_{i,j} - \bar{Y}_{i\cdot})^2$$

The within-group sample variances are treated as “data” about the value of the true error variance, σ^2, which is estimated by taking a weighted average (“pooling” the variances):

$$
\begin{aligned}
s^2 &= \frac{\sum_{i=1}^{r}(n_i - 1)s_i^2}{\sum_{i=1} r(n_i - 1)} \\
&= \frac{1}{n_T - r}\sum_{i=1}^{r}(n_i - 1)s_i^2 \\
&= \frac{1}{n_t - r}\sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{i,j} - \bar{Y}_{i\cdot})^2 \\
&= MSE
\end{aligned}
$$

Note: If $n_i = n$ for all i, this equation reduces to a simple mean, $s^2 = \frac{1}{r}\Sigma s_i^2$

This is also known as the “balanced design”. If $n_i \neq n$, s^2 will be weighted by group size.

ANOVA table

Source of Variation	SS	df	MS	E{MS}
ANOVA model fixed effect	$SSR = \Sigma n_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$	$r-1$	MSR= $\frac{SSR}{r-1}$	$\sigma^2 + \frac{n\Sigma(\mu_i - \mu_{\cdot})^2}{r-1}$
Error	$SSE = \Sigma\left(Y_{ij} - \bar{Y}_{i.}\right)^2$	$n-r$	MSE= $\frac{SSE}{n-r}$	σ^2
Total	$SSTO = \Sigma\left(Y_{ij} - \bar{Y}_{\cdot\cdot}\right)^2$	$n-1$		

Under $\mathrm{H}_0 : (\mu_1 = \mu_2 = \ldots = \mu_r)$,

$$F^* = \frac{MSM}{MSE} \sim F_{r-1, n_t - r}$$

If $p = P\left(F_{r-1, n_t - r} \geq F^*\right) \leq \alpha \rightarrow$ reject H_0

If we reject H_0, we conclude that *at least one* of the factor levels has a group mean that is different from the others.

Factor Effects Model

Factor effects simply reparameterize the cell means model so that the parameters now represent differences (i.e., “effects”) relative to a selected baseline reference.

Advantages:

- easier to interpret null hypotheses
- an effect of 0 for a particular level indicates that the level is not different from the **reference**
- positive and negative effects have similarly natural interpretations Disadvantage:

somewhat more convoluted notation. Choice of reference matters.

Factor Effects Model

$$\mu_i = \mu. + (\mu_i - \mu.)$$
$$Let\ \ \tau_i = \mu_i - \mu.$$
$$\mu_i = \mu. + \tau_i$$
$$Then\ Y_{ij} = \mu. + \tau_i + \varepsilon_{ij}$$

- $\mu.$ is the (unknown) population mean for the **baseline reference**, common to all observations
- τ_i is the *$i\ th$ factor level effect* or the *$i\ th$ treatment effect.*
- ε_{ij} are independent $N(0, \sigma^2)$ $\quad i = 1, \ldots, r\ \ and\ j = 1, \ldots, n_i$
- Factor effects model and cell means model are equivalent for modeling data.

Factor effects model $Y_{ij} = \mu. + \tau_i + \varepsilon_{ij}$

Cell means model $Y_{ij} = \mu_i + \varepsilon_{ij}$

- Factor effect model uses intercept ($\beta_0\ or\ \mu.$) to represent the baseline level. Other levels are compared to the baseline ($\beta_i = \tau_i = \mu_i - \mu.$) $i = 1, ..r$
- Cell mean model doesn't use intercept. All levels are estimated with β_i, $i = 1, ..r$

Some basic choices of the reference and the response function

Example: suppose r=3

Unweighted mean

$$\mu. = \frac{(\Sigma_{i=1}^{r}\mu_i)}{r}$$

Subject to restriction that $(\Sigma_{i=1}^{r}\tau_i = 0)$

- The parameter vector is $(\mu., \tau_1, \tau_2)$
- For level 1: $\mathrm{E(Y)} = \mu_1 = \mu. + \tau_1$
- For level 2: $\mathrm{E(Y)} = \mu_2 = \mu. + \tau_2$
- For level 3: $\mathrm{E(Y)} = \mu_3 = \mu. + \tau_3 = \mu. - \tau_1 - \tau_2$

The first factor mean

$$\mu. = \mu_1$$

- The parameter vector is $(\mu., \tau_2, \tau_3)$
- For level 1: $\mathrm{E(Y)} = \mu_1 = \mu. + \tau_1 = \mu_1 + 0$
- For level 2: $\mathrm{E(Y)} = \mu_2 = \mu. + \tau_2 = \mu_1 + (\mu_2 - \mu_1)$
- For level 3: $\mathrm{E(Y)} = \mu_3 = \mu. + \tau_3 = \mu_1 + (\mu_3 - \mu_1)$

The second factor mean

$$\mu. = \mu_2$$

- The parameter vector is $(\mu., \tau_1, \tau_3)$
- For level 1: $\mathrm{E(Y)} = \mu_1 = \mu. + \tau_1 = \mu_2 + (\mu_1 - \mu_2)$
- For level 2: $\mathrm{E(Y)} = \mu_2 = \mu. + \tau_2 = \mu_2 + 0$
- For level 3: $\mathrm{E(Y)} = \mu_3 = \mu. + \tau_3 = \mu_2 + (\mu_3 - \mu_2)$

Factor Effects Model with Unweighted Mean

$$\mu. = \frac{(\Sigma_{i=1}^{r}\mu_i)}{r} \qquad \text{Subject to restriction that} (\Sigma_{i=1}^{r}\tau_i = 0)$$

We shall use only the parameters $\mu., \tau_1, \ldots \tau_{r-1}$ for the linear model, since $\tau_r = -\tau_1 - \tau_2 - \cdots - \tau_{r-1}$

Consider a single factor study with $r = 3$ factor levels when $n_1 = n_2 = n_3 = 2$.

The matrix form $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ can be specified as

$$Y = \begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{bmatrix} \quad X = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{bmatrix} \quad \beta = \begin{bmatrix} \mu. \\ \tau_1 \\ \tau_2 \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{bmatrix} \quad \boldsymbol{E}\{\boldsymbol{Y}\} = \boldsymbol{X\beta} = \begin{bmatrix} \mu. + \tau_1 \\ \mu. + \tau_1 \\ \mu. + \tau_2 \\ \mu. + \tau_2 \\ \mu. - \tau_1 - \tau_2 \\ \mu. - \tau_1 - \tau_2 \end{bmatrix}$$

The intercept is back for the reference mean

Factor Effects Model with Unweighted Mean

Ho: $\tau_1 = \tau_2 = \cdots = \tau_{r-1} = 0$

Ho: $not\ all\ \tau_i = 0$

	(Intercept)	design1 (x_1)	design2 (x_2)	design3 (x_3)
1	1	1	0	0
2	1	1	0	0
3	1	1	0	0
4	1	1	0	0
5	1	1	0	0
6	1	0	1	0
7	1	0	1	0
8	1	0	1	0
9	1	0	1	0
10	1	0	1	0
11	1	0	0	1
12	1	0	0	1
13	1	0	0	1
14	1	0	0	1
15	1	-1	-1	-1
16	1	-1	-1	-1
17	1	-1	-1	-1
18	1	-1	-1	-1
19	1	-1	-1	-1

```
Analysis of Variance Table

Response: y
            Df Sum Sq Mean Sq F value    Pr(>F)    
Design_uw    3 588.22 196.074  18.591 2.585e-05 ***
Residuals 15 158.20  10.547
```

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
μ.  (Intercept)      18.6750     0.7485  24.949 1.25e-13 ***
τ1  Design_uwdesign1 -4.0750     1.2708  -3.207 0.005884 ** 
τ2  Design_uwdesign2 -5.2750     1.2708  -4.151 0.000854 ***
τ3  Design_uwdesign3  0.8250     1.3706   0.602 0.556221    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.248 on 15 degrees of freedom
Multiple R-squared:  0.7881,	Adjusted R-squared:  0.7457 
F-statistic: 18.59 on 3 and 15 DF,  p-value: 2.585e-05
```

$E\{Y_{1.}\} = \underline{\mu. + \tau_1} = 18.675 - 4.075 = 14.6$

$E\{Y_{2.}\} = \underline{\mu. + \tau_2} = 18.675 - 5.275 = 13.4$

$E\{Y_{3.}\} = \underline{\mu. + \tau_3} = 18.675 + 0.825 = 19.5$

$E\{Y_4\} = \underline{\mu. - \tau_1 - \tau_2 - \tau_3} = 18.675 + 4.075 + 5.275 - 0.825 = 27.2$

Note: the t value and the p value are for testing the significance of the corresponding coefficients of the same row.

Factor Effects Model with the first group 1 as reference mean (default)

$\mu. = \mu_1 \qquad \tau_1 = 0$

	(Intercept)	design2	design3	design4
1	1	0	0	0
2	1	0	0	0
3	1	0	0	0
4	1	0	0	0
5	1	0	0	0
6	1	1	0	0
7	1	1	0	0
8	1	1	0	0
9	1	1	0	0
10	1	1	0	0
11	1	0	1	0
12	1	0	1	0
13	1	0	1	0
14	1	0	1	0
15	1	0	0	1
16	1	0	0	1
17	1	0	0	1
18	1	0	0	1
19	1	0	0	1

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)    
Design_uw  3 588.22 196.074  18.591 2.585e-05 ***
Residuals 15 158.20  10.547                      
```

Coefficients:

		Estimate	Std. Error	t value	Pr(>|t|)	
$\mu.$	(Intercept)	14.600	1.452	10.053	4.66e-08	***
τ_2	Designdesign2	-1.200	2.054	-0.584	0.5677	
τ_3	Designdesign3	4.900	2.179	2.249	0.0399	*
τ_4	Designdesign4	12.600	2.054	6.135	1.91e-05	***

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.248 on 15 degrees of freedom
Multiple R-squared:  0.7881,	Adjusted R-squared:  0.7457 
F-statistic: 18.59 on 3 and 15 DF,  p-value: 2.585e-05
```

$E\{Y_{1.}\} = \underline{\mu. + \tau_1} = 14.6 + 0 = 14.6$ $\qquad E\{Y_{2.}\} = \underline{\mu. + \tau_2} = 14.6 - 1.2 = 13.4$

$E\{Y_{3.}\} = \underline{\mu. + \tau_3} = 14.6 + 4.9 = 19.5$ $\qquad E\{Y_{4.}\} = \underline{\mu. + \tau_4} = 14.6 + 12.6 = 27.2$

Note: the t value and the p value are for testing the significance of the corresponding coefficients of the same row.

Factor Effects Model with the first group 2 as reference mean (Relevel)

$\mu. = \mu_2 \qquad \tau_2 = 0$

	(Intercept)	design1	design3	design4
1	1	1	0	0
2	1	1	0	0
3	1	1	0	0
4	1	1	0	0
5	1	1	0	0
6	1	0	0	0
7	1	0	0	0
8	1	0	0	0
9	1	0	0	0
10	1	0	0	0
11	1	0	1	0
12	1	0	1	0
13	1	0	1	0
14	1	0	1	0
15	1	0	0	1
16	1	0	0	1
17	1	0	0	1
18	1	0	0	1
19	1	0	0	1

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)    
Design_uw  3 588.22 196.074  18.591 2.585e-05 ***
Residuals 15 158.20  10.547                      
```

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
μ.  (Intercept)    13.400      1.452   9.226 1.43e-07 ***
τ1  Design2design1  1.200      2.054   0.584   0.5677    
τ3  Design2design3  6.100      2.179   2.800   0.0135 *  
τ4  Design2design4 13.800      2.054   6.719 6.88e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.248 on 15 degrees of freedom
Multiple R-squared:  0.7881,    Adjusted R-squared:  0.7457 
F-statistic: 18.59 on 3 and 15 DF,  p-value: 2.585e-05
```

$E\{Y_{1.}\} = \underline{\mu. + \tau_1} = 13.4 + 1.2 = 14.6$

$E\{Y_{2.}\} = \underline{\mu. + \tau_2} = 13.4$

$E\{Y_{3.}\} = \underline{\mu. + \tau_3} = 13.4 + 6.1 = 19.5$

$E\{Y_{4.}\} = \underline{\mu. + \tau_4} = 13.4 + 13.8 = 27.2$

Note: the t value and the p value are for testing the significance of the corresponding coefficients of the same row.

Estimation and hypotheses on the following effects

- A single factor level mean μ_i
- A difference between two factor level means
- A contrast among factor level means
- A linear combination of factor level means.
- Multiple and simultaneous comparison

A single factor level and difference between two factor levels

$Ho: \mu_i = c \quad Ha: \mu_i \neq c$

$$ts = \frac{\bar{Y}_i - c}{s\{\bar{Y}_i\}} \qquad s^2\{\bar{Y}_i\} = \frac{MSE}{n_i} \quad \leftarrow \sigma^2(\bar{Y}) = \frac{\sigma^2}{n}$$

$$CI: \bar{Y}_i \pm t\left(1 - \frac{\alpha}{2}; n_T - r\right) s\{\bar{Y}_i\}$$

$Ho: \mu_2 = 0 \quad Ha: \mu_2 \neq 0$

$\bar{Y}_2 = 13.4 \qquad s^2\{\bar{Y}_2\} = \frac{10.55}{5} = 2.11$, so $s\{\bar{Y}_2\} = 1.453$

$ts = \frac{\bar{Y}_i}{s\{\bar{Y}_i\}} = 9.22$ Reject if $t_s > t(0.975; 15) = 2.131$

$CI: \bar{Y}_i \pm t\left(1 - \frac{\alpha}{2}; n_T - r\right) s\{\bar{Y}_i\} = 13.4 \pm 2.131(1.453)$
$= 13.4 \pm 3.096 = 10.3,\ 16.6$

$Ho: \mu_i - \mu_j = 0 \quad Ha: \mu_i - \mu_j \neq 0$

$$ts = \frac{\bar{Y}_i - \bar{Y}_j}{s\{\bar{Y}_i - \bar{Y}_j\}} \quad s^2\{\bar{Y}_i - \bar{Y}_j\} = MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right) \leftarrow \sigma^2(\bar{Y}_1 \pm \bar{Y}_2) = \sigma^2(\bar{Y}_1) + \sigma^2(\bar{Y}_2)$$

$$CI: \bar{Y}_i - \bar{Y}_j \pm t\left(1 - \frac{\alpha}{2}; n_T - r\right) s\{\bar{Y}_i - \bar{Y}_j\}$$

$Ho: \mu_2 - \mu_1 = 0 \quad Ha: \mu_2 - \mu_1 \neq 0$

$\bar{Y}_2 - \bar{Y}_1 = 13.4 - 14.6 = -1.2$

$s^2\{\bar{Y}_i - \bar{Y}_j\} = MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right) = 4.22$

$ts = \frac{-1.2}{2.054} = -0.584$ Reject if $|t_s| > t(0.975; 15) = 2.131$

$CI: \bar{Y}_i - \bar{Y}_j \pm t\left(1 - \frac{\alpha}{2}; n_T - r\right) s\{\bar{Y}_i - \bar{Y}_j\} = -1.2 \pm 2.131(2.054)$
$= -1.2 \pm 4.377 = -5.58,\ 3.18$

Contrast of factor level means (not simultaneous comparison)

A ***contrast*** is a comparison involving two or more factor level means. A contrast will be denoted by $\boldsymbol{L}$, and is defined as

$L = \Sigma_{i=1}^{r} c_i \mu_i$ Where $\Sigma_{i=1}^{r} c_i = 0$

For example:

1. $L = \mu_1 - \mu_2$ $c_1 = 1, c_2 = -1, c_3 = 0, c_4 = 0$

2. $L = \frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4}{2}$ $c_1 = \frac{1}{2}, c_2 = \frac{1}{2}, c_3 = -\frac{1}{2}, c_4 = -\frac{1}{2}$

3. $L = \mu_1 - \frac{\mu_1 + \mu_2 + \mu_3 + \mu_4}{4}$ $c_1 = \frac{3}{4}, c_2 = -\frac{1}{4}, c_3 = -\frac{1}{4}, c_4 = -\frac{1}{4}$

$\hat{L} = \Sigma_{i=1}^{r} c_i \bar{Y}_i$ $s^2\{L\} = \text{MSE}\, \Sigma_{i=1}^{r} c_i^2 / n_i$ $\frac{\hat{L}-L}{s\{L\}} \sim t(n_T - r)$ for ANOVA

For example: Ho: $\frac{\mu_1+\mu_2}{2} - \frac{\mu_3+\mu_4}{2} = 0$ and Ha: $\frac{\mu_1+\mu_2}{2} - \frac{\mu_3+\mu_4}{2} \neq 0$

$\hat{L} = \Sigma_{i=1}^{r} c_i \bar{Y}_i$=-9.35 $s^2\{L\} = \text{MSE}\, \Sigma_{i=1}^{r} c_i^2 / n_i$=2.242 $t_s = \frac{\hat{L} - L}{s\{L\}} = -6.23 \sim t(15)$ The CI for L: (-12.54, -6.16)

```
oneway(cereal$y, cereal$design,mc=matrix(c(0.5,0.5,-0.5,-0.5),1,4))$Contrast.NOT.simultaneous
```

```
$Contrast.NOT.simultaneous
         L      lower     upper         t   p-value
 -9.350000 -12.540892 -6.159108 -6.245605  0.000016
```

Bonferroni multiple comparison

We want compare g linear combination Ls'. $L = \Sigma_{\mathrm{i=1}}^{\mathrm{r}} \mathrm{c_i} \mu_i$ where $\Sigma_{\mathrm{i=1}}^{\mathrm{r}} \mathrm{c_i} = 0$

$$\hat{L} \pm Bs\{\hat{L}\}, where\ B = t(1 - \frac{\alpha}{2g}; n_T - r)$$

For example:

1. $L_1 = \frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4}{2}$

$c_1 = \frac{1}{2}, c_2 = \frac{1}{2}, c_3 = -\frac{1}{2}, c_4 = -\frac{1}{2}$

2. $L_2 = \frac{\mu_1 + \mu_3}{2} - \frac{\mu_2 + \mu_4}{2}$

$c_1 = \frac{1}{2}, c_2 = -\frac{1}{2}, c_3 = \frac{1}{2}, c_4 = -\frac{1}{2}$

$\widehat{L_1} = \Sigma_{\mathrm{i=1}}^{\mathrm{r}} \mathrm{c_i} \bar{Y}_i$=-9.35

$s^2\{L_1\} = 2.242$

$\widehat{L_2} = \Sigma_{\mathrm{i=1}}^{\mathrm{r}} \mathrm{c_i} \bar{Y}_i$=-3.25

$s^2\{L_2\} = 2.242$

$$B = t\left(1 - \frac{\alpha}{2g}; n_T - r\right) = \mathbf{2.84}$$

The simultaneous CI for L_1 : $(-\mathbf{13.6}, -\mathbf{5.1})$ L_2: $(-\mathbf{7.5}, \quad \mathbf{1})$

```
mc2<-matrix(c(0.5,0.5,-0.5,-0.5, 0.5, -0.5, 0.5, -0.5),2,4, byrow=TRUE)
oneway(cereal$y, cereal$design, mc=mc2)
```

The procedure of diagnostic and remedial measures in ANOVA is like regular regression model

- Non-constancy of error variance
- Non-independence of error terms
- Outliers
- Omission of important predictors
- Non-normality of error terms

The oneway() function in ALSM package serves multiple purpose for single factor ANOVA

- Fitting of ANOVA model
- ANOVA table
- Test and confidence interval for single factor level mean
- Inferences for difference between two factor level means
- Contrast of factor level means
- ANOVA diagnostic
- Nonparametric Rank F test
- Plots for exploration and residuals

Usage $oneway(y, group, alpha = 0.05,\ c.vallue = 0, mc = NULL)$

Arguments
$y: vector$
$group: vector, factor$
$alpha: 0.05\ by\ default$
$c.value:\ \ single\ factor\ test: Ho: \mu_i = c, 0\ by\ defult$
$mc: matric\ \ contrast$

Example

1. Find the test statistic and p value for a hypothesis test $Ho: \mu_1 = \mu_3,\ Ha: \mu_1 \neq \mu_3$

2. Find the test statistic and p value for a hypothesis test $Ho: \mu_2 = \mu_3,\ Ha: \mu_2 \neq \mu_3$

3. Find the test statistic and p value for a hypothesis test $Ho: L = 0, Ha: L \neq 0$ $where\ L = \mu_1 - \frac{\mu_1 + \mu_2 + \mu_3 + \mu_4}{4}$

4. Find the simultaneous confidence interval for $(\mu_1 - \mu_3)$, $(\mu_2 - \mu_3)$

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)     14.600      1.452  10.053 4.66e-08 ***
Designdesign2   -1.200      2.054  -0.584   0.5677    
Designdesign3    4.900      2.179   2.249   0.0399 *  
Designdesign4   12.600      2.054   6.135 1.91e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.248 on 15 degrees of freedom
Multiple R-squared:  0.7881,    Adjusted R-squared:  0.7457 
F-statistic: 18.59 on 3 and 15 DF,  p-value: 2.585e-05
```