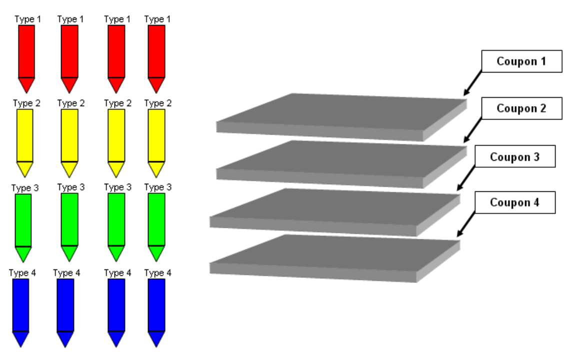
MODULE 5: RANDOMIZED COMPLETE BLOCK DESIGN (RCBD)

Learning Objectives

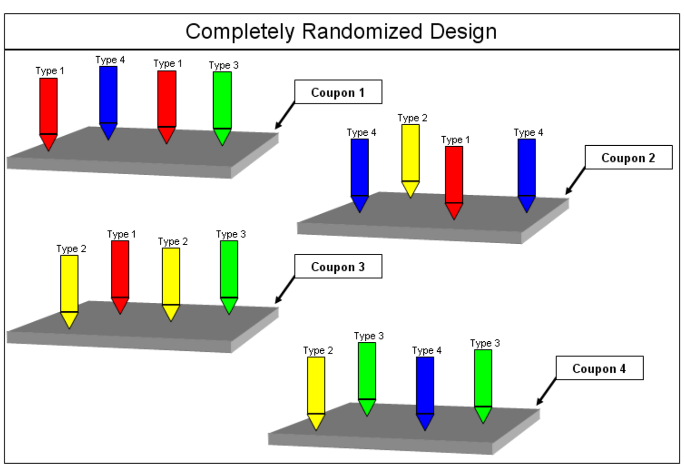
- Explain the purpose of blocking and when to use a RCBD.
- Design and implement a RCBD, including appropriate blocking variables.
- Compare the statistical models and ANOVA tables for CRD and RCBD.
- Analyze RCBD data using statistical software and interpret results.
- Evaluate the effectiveness of blocking in reducing experimental error.

Example 5.1: Indentation Tips
A study was conducted to compare four different tips used to measure hardness on a testing machine. Hardness testing is a crucial process in materials science and engineering that determines a metal's resistance to deformation, particularly permanent indentation, scratching, or wear. The metal sheets used to compare tips vary in density due to extraneous variables. It is thought that one particular sheet is fairly consistent, but consistency between sheets is not likely. These metal sheets are referred to as coupons.

Suppose the experimenter has four different coupons available to test.



Suppose there was no restriction on treatment randomization. If this is the case, the following randomization of a completely randomized design is possible.



1. Suppose that coupon 1 is atypically soft. How will this affect the results?

2. Suppose that coupon 2 is atypically hard. How will this affect the results?

Up to this point, we have considered only one experimental design: the Completely Randomized Design. This design is appropriate when the experimental units are *homogeneous* If this is not the case, then the experimental units should be grouped into *blocks* of homogenous units to reduce the experimental error variance. This type of design is known as a **Randomized Complete Block Design (RCBD)**.

Elements of RCBDs

- **Block:** This is a ___________________ group of experimental units. A RCBD consists of first sorting the experimental units into blocks.

- **Complete:** Each __________ consists of one complete replication of the set of _______________. Therefore, each treatment will show up __________ within each block.

- **Randomized:** The treatments are ____________ assigned to experimental units separately ______________ each block.

When blocking the experimental units, keep the following objectives in mind:

- Within blocks, make the experimental units as *homogeneous* as possible with respect to the response variable.
- Make the different blocks as *heterogeneous* as possible with respect to the response variable.

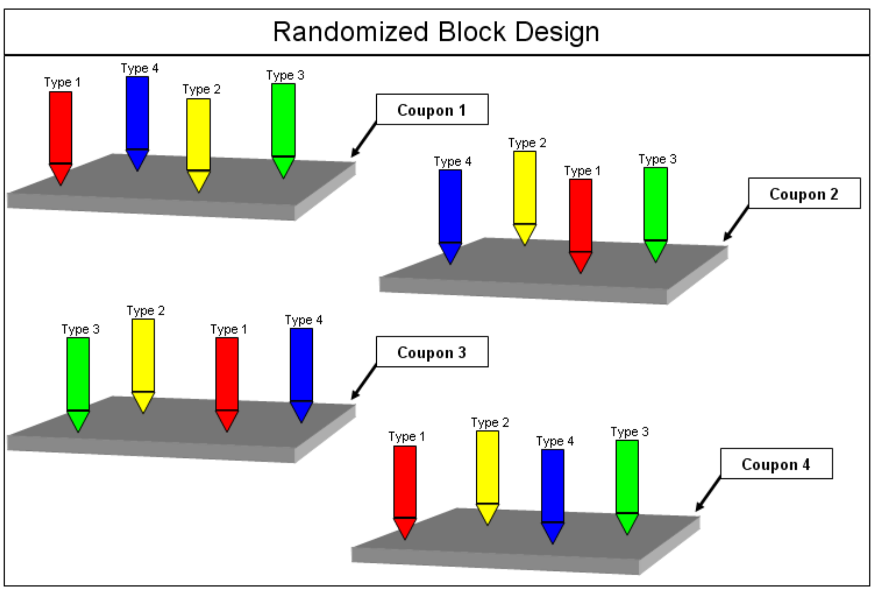
Advantages

- Effective blocking can lead to *reduced* experimental error and more precise estimates of treatment effects.
- The RCBD can accommodate *any* number of treatments and replications.
- The statistical analysis is relatively simple.

Disadvantages

- The degrees of freedom for experimental error are not as large as for a CRD.
- *More* assumptions are required than for a CRD.

Recall that each coupon in our study is fairly consistent, but consistency between coupons is not likely. Therefore, we will set up the experiment so that **each coupon represents a block**, and each tip will be tested on each block. One possible randomization scheme is given as follows:



> Blocking is used to reduce the effect of one or more extraneous variables. If you can control the variable or you have a genuine interest in the effect of a variable on the response, then it is **NOT** a blocking variable.

Example 5.2: Advertising
In an experiment on the affects of four levels of newspaper advertising saturation (Levels A, B, C, and D) on sales volume, the experimental unit is a city, and 16 cities are available for the study. The size of the city is usually highly correlated with sales volume.

1. How would you create blocks for this experiment?

2. Identify the treatment design.

3. Identify the experimental design.

Next, we randomly assign treatments to the experimental units. To do so, we randomly permute **WITHIN EACH BLOCK** to assign treatments to experimental units. We have four blocks of size four:

See "RCBD Randomization" tutorials on Canvas

4. Using JMP/R, create a possible random assignment of advertising saturation to cities (e.u.) in a RCBD with 4 replications (blocks) where the blocking factor is city size. Indicate the advertising saturation level for each experimental unit in the table below.

	Block -- City Size			
e.u.	**Block 1 (smallest cities)**	**Block 2**	**Block 3**	**Block 4 (largest cities)**
1				
2				
3				
4				

Treatment Effects Model for a RCBD

$$y_{ij} + \tau_i + \rho_j + \epsilon_{ij} \quad \text{with} \quad \epsilon_{ij} \sim \text{independent } N(0, \sigma^2)$$

$$\text{for } i = 1,2, \ldots, t \qquad j = 1,2, \ldots, r$$

where

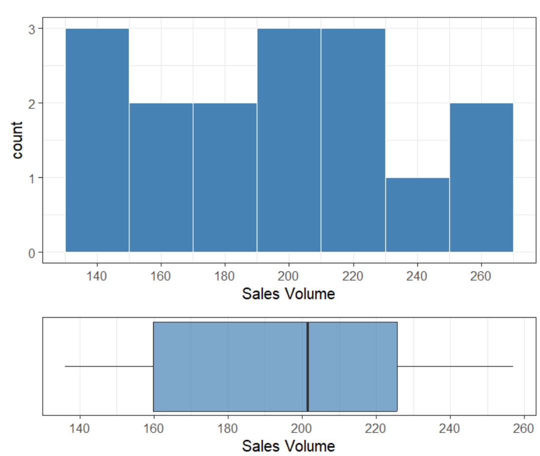
- ________ is the response from the city in the j^{th} city size block receiving the i^{th} advertising saturation
- ________is the overall mean sales volume
- ________ is the effect of the i^{th} advertising saturation
- ________ is the effect of the j^{th} city size block (this represents the average deviation of the units in block j from the overall mean)
- ________ is the random error term associated with the city in the j^{th} city size block receiving the i^{th} advertising saturation.

The treatment and blocks are assumed to be *additive* (there is no interaction between treatments and blocks).

Analysis of Variance

Using the values for sales below, let's further investigate what is meant by controlling for variation with a "block effect" (ρ_j).

The histogram, box plot, and summary statistics provide the distribution and variation of the sales volume regardless of advertisement saturation.



Variable	N	Mean	Std Dev
Sales Volume	16	196.06	38.46

5. Using the output above, calculate the SSTotal. *Note $s^2 = \frac{\sum_i \sum_j (y_{ij} - \bar{y}_{..})^2}{N-1}$ (you can check your answer from the output below).*

$$\text{SST} = \sum_i \sum_j (y_{ij} - \bar{y}_{..})^2 \to \text{SST} = (N-1)s^2 =$$

⚠ Incorrect Output -- **CRD ANOVA (ignoring city size block effect)**

To understand the variation in sales volumes explained by advertisement saturation (SSTrt), we can run a one-way ANOVA.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	6249.188	2083.06	1.5686
Error	12	15935.750	1327.98	**Prob > F**
C. Total	15	22184.938		0.2482

Summary of Fit

RSquare	0.281686
RSquare Adj	0.102108
Root Mean Square Error	36.44145
Mean of Response	196.0625
Observations (or Sum Wgts)	16

Expanded Estimates

Nominal factors expanded to all levels

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	196.0625	9.110362	21.52	<.0001*
advertising_saturation[A]	-23.0625	15.77961	-1.46	0.1696
advertising_saturation[B]	-12.3125	15.77961	-0.78	0.4503
advertising_saturation[C]	6.4375	15.77961	0.41	0.6905
advertising_saturation[D]	28.9375	15.77961	1.83	0.0916

Level	Least Sq Mean	Std Error	Mean
A	173.000	18.22	173.00
B	183.750	18.22	183.75
C	202.500	18.22	202.50
D	225.000	18.22	225.00

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
advertising_saturation	3	3	6249.1875	1.5686	0.2482

6. From the *incorrect* output above, identify:

- $R^2 =$
- $\text{SSTrt} =$

- $\tau_A =$
- $\tau_B =$
- $\tau_C =$
- $\tau_D =$

7. What is SSError in the *incorrect* output above? What are some potential sources of unexplained error in this study?

- $\text{SSE} =$

Since the study was run as a block design, we can compute the amount of variation in sales explained by the city size (blocks) independent of the variation explained by the advertisement saturation.

8. Using the table below, fill in the Mean Sales (see *incorrect* output above) and calculate the block effect for each city size block. Then calculate SSBlk.

Original Data

Advertising Saturation	Block -- City Size Block 1 (smallest cities)	 Block 2	 Block 3	 Block 4 (largest cities)	Mean Sales ($\bar{y}_{i\cdot}$)
A	136	153	203	200	
B	147	146	217	225	
C	162	189	231	228	
D	184	208	251	257	
Block Effects $\boldsymbol{\rho_{\cdot j} = \bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}}$					$\bar{y}_{\cdot\cdot} =$

$$\mathrm{SSBlk} = t\sum_j \left(\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}\right)^2 =$$

Recall, the purpose of the block design is to "remove" the variation explained by the blocking variable from the MSError, thus reducing the experimental error variation (MSE).

9. Using the block effects from the table above, adjust each sales volume value by the associated block effect. Then calculate the block adjusted mean sales.

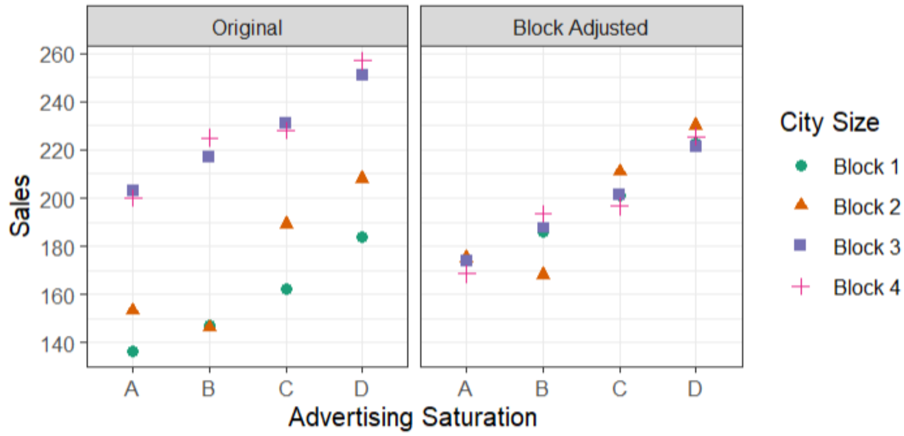
Block adjusted – $\boldsymbol{y_{ij} - \rho_{\cdot j}}$

Advertising Saturation	Block -- City Size Block 1 (smallest cities)	 Block 2	 Block 3	 Block 4 (largest cities)	Block-adjusted Mean Sales
A					
B					
C					
D					

- What do you notice about the block-adjusted mean sales compared to the original mean sales?

- Would the SSTrt change when using the block-adjusted mean sales? Recall $\mathrm{SSTrt} = r\sum_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$.

10. What do you notice about the variability of the block-adjusted sales within each advertising saturation compared to the original variability of sales (without block-adjusted sales)?



To sum all this up, in a RCBD, the least squares estimators of the parameters in are given by:

Parameter	Least Squares Estimator
μ	
τ_i	
ρ_j	

The **fitted values** are given by:

$$\hat{y}_{ij} = \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{..}) = \bar{y}_{i.} + \bar{y}_{.j} - \bar{y}_{..}$$

Therefore, the **residuals** are given by:

$$\epsilon_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - (\bar{y}_{i.} + \bar{y}_{.j} - \bar{y}_{..}) = y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}$$

Then, the **total sum of squares** (**SST**) can then be partitioned into the sum of squares for blocks, treatments, and error:

$$\text{SSTrt} = r\sum_i (\bar{y}_{i.} - \bar{y}_{..})^2$$

$$\text{SSBlk} = t\sum_j (\bar{y}_{.j} - \bar{y}_{..})^2$$

$$\text{SSE} = \sum_i \sum_j (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$$

Degrees of Freedom, Mean Squares and the ANOVA Table

The degrees of freedom are listed in the following table, and the mean squares are obtained the usual way. The ANOVA table for a RCBD is presented below:

Source of Variation	**df**	**SS**	**MS**	**F**
Blocks	$r - 1$	SSBlk	MSBlk	MSBlk/MSE
Treatment	$t - 1$	SSTrt	MSTrt	MSTrt/MSE
Block x Treatment – error	$(r - 1)(t - 1)$	SSE	MSE	
Total ($N = rt$)	$N - 1$			

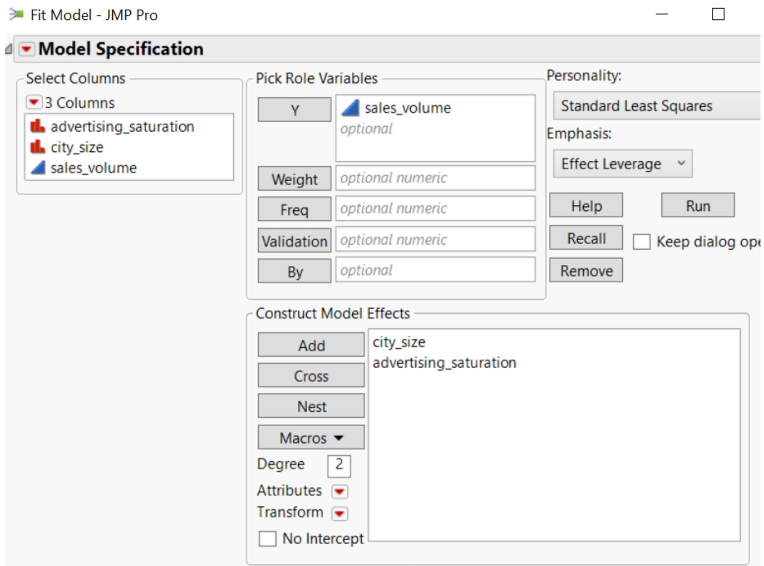
11. Using what you learned from the previous questions (you should not have to do any more hefty calculations; you have all of the sums of squares), fill out the ANOVA table below.

Source of Variation	**df**	**SS**	**MS**	**F**
City Size				
Advertisement Saturation				
City Size x Advertisement – error				
Total ($N = rt$)		22184.93		

Analyzing a RCBD in JMP/R (Correct Analysis)

See "RCBD Analysis" tutorials on Canvas

Analyze > Fit Model



Summary of Fit

RSquare	0.975499
RSquare Adj	0.959164
Root Mean Square Error	7.771476
Mean of Response	196.0625
Observations (or Sum Wgts)	16

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	21641.375	3606.90	59.7209
Error	9	543.563	60.40	Prob > F
C. Total	15	22184.938		<.0001*

Parameter Estimates

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
city_size	3	3	15392.188	84.9517	<.0001*
advertising_saturation	3	3	6249.188	34.4902	<.0001*

Expanded Estimates

Nominal factors expanded to all levels

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	196.0625	1.942869	100.91	<.0001*
city_size[Block 1]	-38.8125	3.365148	-11.53	<.0001*
city_size[Block 2]	-22.0625	3.365148	-6.56	0.0001*
city_size[Block 3]	29.4375	3.365148	8.75	<.0001*
city_size[Block 4]	31.4375	3.365148	9.34	<.0001*
advertising_saturation[A]	-23.0625	3.365148	-6.85	<.0001*
advertising_saturation[B]	-12.3125	3.365148	-3.66	0.0052*
advertising_saturation[C]	6.4375	3.365148	1.91	0.0880
advertising_saturation[D]	28.9375	3.365148	8.60	<.0001*

```
> sales_rcbdmod <- lm(sales_volume ~ advertising_saturation + city_size,
+                     data = sales_data)
> anova(sales_rcbdmod)
Analysis of Variance Table

Response: sales_volume
                       Df  Sum Sq Mean Sq F value    Pr(>F)
advertising_saturation  3  6249.2  2083.1  34.490 2.901e-05 ***
city_size               3 15392.2  5130.7  84.952 6.376e-07 ***
Residuals               9   543.6    60.4
```

```
> summary(sales_rcbdmod)

Call:
lm(formula = sales_volume ~ advertising_saturation + city_size,
    data = sales_data)

Residuals:
     Min       1Q   Median       3Q      Max
-15.6875  -2.5000   0.5625   2.5000   9.8125

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              196.062      1.943 100.914 4.67e-15 ***
advertising_saturation1  -23.062      3.365  -6.853 7.45e-05 ***
advertising_saturation2  -12.312      3.365  -3.659 0.005245 **
advertising_saturation3    6.437      3.365   1.913 0.088039 .
city_size1               -38.812      3.365 -11.534 1.08e-06 ***
city_size2               -22.062      3.365  -6.556 0.000104 ***
city_size3                29.438      3.365   8.748 1.08e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.771 on 9 degrees of freedom
Multiple R-squared:  0.9755,    Adjusted R-squared:  0.9592
F-statistic: 59.72 on 6 and 9 DF,  p-value: 9.683e-07
```

Recall the statistical effects model to predict the sales response for each city in the data set is:

$$\hat{y}_{ij} = \hat{\mu} + \hat{\tau}_i + \hat{\rho}_j$$

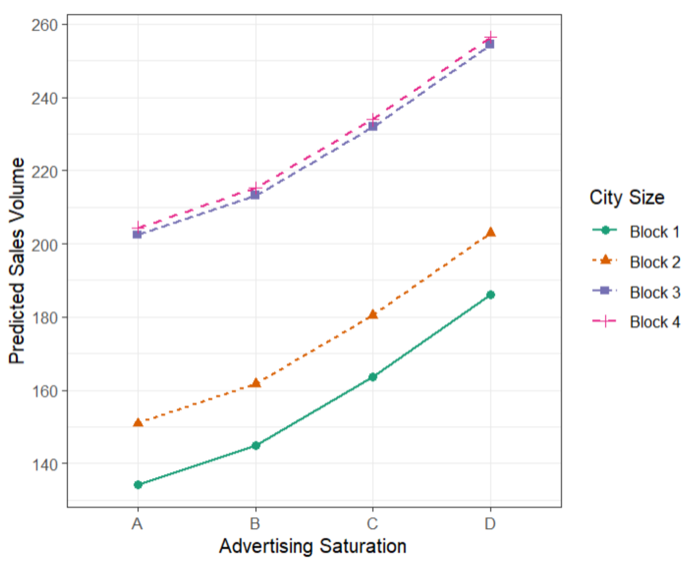
Where $\hat{\tau}_i$ represent the effects of advertising saturation and the $\hat{\rho}_j$ represent the city size block effects.

12. Using the statistical effects model and expanded estimates shown in the output above, fill out the table below with the predicted sales for each block and advertisement saturation.

Predicted sales – $\hat{y}_{ij}$

Advertising Saturation	Block -- City Size			
	Block 1 (smallest cities)	**Block 2**	**Block 3**	**Block 4 (largest cities)**
A				
B				
C				
D				

Notice that the predictions (shown in the plot) show that the model assumes no interaction between advertising saturation and city size in the randomized complete block design (RCBD). This means the effect of advertising saturation on sales is consistent across cities – while overall sales levels vary by city, the differences between advertising saturation levels remain unchanged.



13. Looking at the output below, what happens if we fit a model which contains the interaction (*incorrect analysis*)? Why does this happen?

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	MSE used	F Ratio
				DFE used	Prob > F
Error	0	0.000	.		
C. Total	15	22184.938		.	.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	196.0625	.	.	.
city_size[Block 1]	-38.8125	.	.	.
city_size[Block 2]	-22.0625	.	.	.
city_size[Block 3]	29.4375	.	.	.
advertising_saturation[A]	-23.0625	.	.	.
advertising_saturation[B]	-12.3125	.	.	.
advertising_saturation[C]	6.4375	.	.	.
advertising_saturation[A]*city_size[Block 1]	1.8125	.	.	.
advertising_saturation[A]*city_size[Block 2]	2.0625	.	.	.
advertising_saturation[A]*city_size[Block 3]	0.5625	.	.	.
advertising_saturation[B]*city_size[Block 1]	2.0625	.	.	.
advertising_saturation[B]*city_size[Block 2]	-15.6875	.	.	.
advertising_saturation[B]*city_size[Block 3]	3.8125	.	.	.
advertising_saturation[C]*city_size[Block 1]	-1.6875	.	.	.
advertising_saturation[C]*city_size[Block 2]	8.5625	.	.	.
advertising_saturation[C]*city_size[Block 3]	-0.9375	.	.	.

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
city_size	3	3	15392.188	.	.
advertising_saturation	3	3	6249.188	.	.
advertising_saturation*city_size	9	9	543.563	.	.

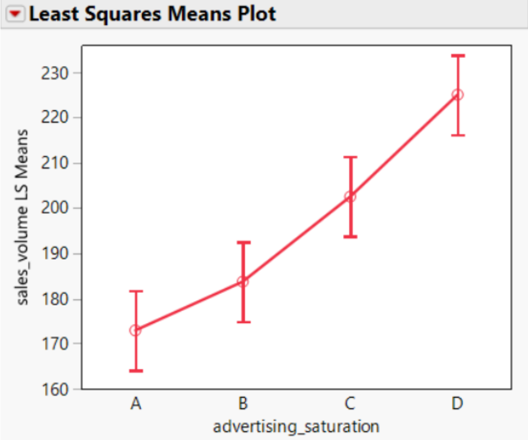
Testing Hypotheses

- The usual F-Statistic is used to test the null hypothesis of no difference among the treatment means.
- Little interest exists in formal inference about block effects, so we typically ignore the F-Statistic given for blocks in the ANOVA table.
- If a significant difference in treatments exists, you should proceed in the usual manner by figuring out WHERE the differences exist.

In our analysis above, we found evidence of an effect of advertisement saturation on the sales volume ($F = 34.49$; $df = 3, 15$; $p < 0.0001$). Therefore, we can look into the pairwise comparisons to determine which advertisement saturation results in the highest mean sales.

advertising_saturation > LSMeans Plot + Tukey HSD + Ordered Differences Report

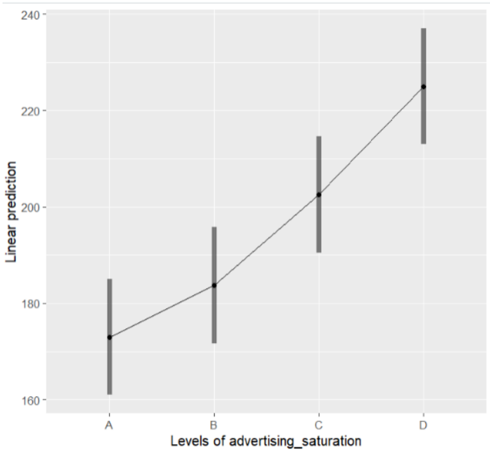
Least Squares Means Plot



Level				Least Sq Mean	Std Error
D	A			225.00	3.8857
C		B		202.50	3.8857
B			C	183.75	3.8857
A			C	173.00	3.8857

Levels not connected by same letter are significantly different.

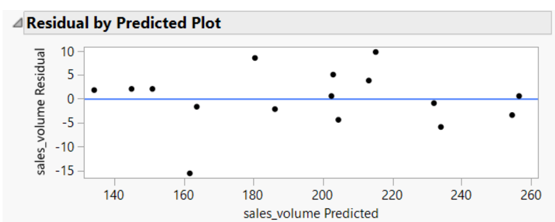
Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
D	A	52.00000	5.495263	34.8449	69.15514	<.0001*
D	B	41.25000	5.495263	24.0949	58.40514	0.0002*
C	A	29.50000	5.495263	12.3449	46.65514	0.0021*
D	C	22.50000	5.495263	5.3449	39.65514	0.0118*
C	B	18.75000	5.495263	1.5949	35.90514	0.0323*
B	A	10.75000	5.495263	-6.4051	27.90514	0.2721

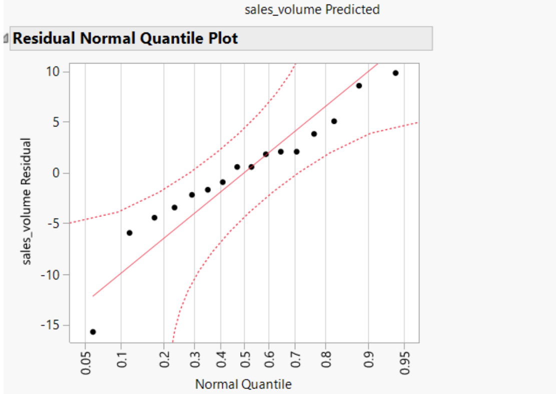



```
> sales_emmeans <- emmeans(sales_rcbdmod, specs = ~ advertising_saturation)
> emmip(sales_emmeans, ~ advertising_saturation, CIs = T, adjust = "tukey")
> cld(sales_emmeans, Letters = LETTERS, decreasing = T, adjust = "tukey")
Note: adjust = "tukey" was changed to "sidak"
because "tukey" is only appropriate for one set of pairwise comparisons
 advertising_saturation emmean   SE df lower.CL upper.CL .group
 D                         225 3.89  9      213      237  A
 C                         202 3.89  9      190      215   B
 B                         184 3.89  9      172      196    C
 A                         173 3.89  9      161      185    C

Results are averaged over the levels of: city_size
> pairs(sales_emmeans)
 contrast estimate  SE df t.ratio p.value
 A - B       -10.8 5.5  9  -1.956  0.2721
 A - C       -29.5 5.5  9  -5.368  0.0021
 A - D       -52.0 5.5  9  -9.463  <.0001
 B - C       -18.8 5.5  9  -3.412  0.0323
 B - D       -41.2 5.5  9  -7.506  0.0002
 C - D       -22.5 5.5  9  -4.094  0.0118

Results are averaged over the levels of: city_size
```

14. Which advertisement saturation results in the largest mean sales? On average, how many more sales is this than the advertisement saturation with the second largest mean sales?

15. Does our model appear to violate constant variance and normality of residuals?





Why go to the trouble to block?

If not blocked, then variation in sales caused by the city size (block) is contained within the error. This 'larger' MSError could make it difficult to find statistical differences among the advertisement saturation means.

CRD Analysis (No blocking)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	6249.188	2083.06	1.5686
Error	12	15935.750	1327.98	**Prob > F**
C. Total	15	22184.938		0.2482

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
advertising_saturation	3	3	6249.1875	1.5686	0.2482

RCBD Analysis (Block by City Size)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	21641.375	3606.90	59.7209
Error	9	543.563	60.40	**Prob > F**
C. Total	15	22184.938		<.0001*

Parameter Estimates

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
city_size	3	3	15392.188	84.9517	<.0001*
advertising_saturation	3	3	6249.188	34.4902	<.0001*

16. Recall from Module 1 that Inclusion criteria, covariates, and blocking all seek to reduce unexplained variation in the response. We could use an inclusion criteria (e.g., only large cities) to remove the variation caused by city size. What is the limitation with doing this?

Key Statistical Concept

We block on an inherent characteristic of the experimental unit (e.g., city size). If we know that we will be using experimental units that will respond very differently based on this inherent characteristic (responsible for a large amount of extraneous variation in the response) AND that characteristic is easily identifiable and thus, makes it easy to form homogeneous groups ➔ BLOCK on it

Expanding the scope of inference may result in increased variation in the response. Use blocking to expand the scope, but then remove the extra variation from the Error.

Practice Problems

1. **Coronary Heart Disease**. A researcher studied the effects of three experimental diets with varying fat contents on the total lipid (fat) level in plasma (Extremely Low = 1, Fairly Low = 2, and Moderately Low = 3). Total lipid level is a widely used predictor of coronary heart disease. Fifteen male subjects were grouped into five blocks according to age. Within each block, the three experimental diets were randomly assigned to the three subjects. Data on reduction in lipid level (grams per liter) after the subjects were on the diet for a fixed period of time can be found on Canvas in the file **fat_diets.csv.**

   a. Identify the treatment design:

   b. Identify the experimental design:

   c. Sketch out the Skeleton ANOVA table for this study.

   d. Write the statistical model for this experiment, making sure to fully explain each component.

e. Check the appropriate model assumptions.

f. Is there a significant difference in the reduction lipid level across the 3 diets? Cite all evidence.

g. Does this mean all the diets differ from one another? Explain.

h. Conduct the Tukey letter groupings and all pairwise comparisons. Which diet results in the largest reduction in lipid level?

2. **Remotivation.** A remotivation team in a psychiatric hospital conducted an experiment to compare five methods for remotivating patients. Patients were grouped according to their level of initial motivation. Patients in each group were randomly assigned to the five remotivation methods. At the end of the experiment patients were evaluated by a team composed of a psychiatrist, a psychologist, a nurse, and a social worker, none of whom was aware of the method to which patients had been assigned. The team assigned each patient a composite score as a measure of his or her level of motivation. The data can be found in the file **remotivation_data.csv** on Canvas.

   a. Identify the treatment design.

   b. Identify the experimental design.

c. Sketch out the Skeleton ANOVA table for this study.

d. Write the statistical model for this experiment.

e. Determine whether any of the model assumptions have been violated.

f. Do the data provide significant evidence of a difference in method of remotivation? If so, where?