Math 3080 § 1. **Final Exam** Name: Example
Treibergs April 19, 2010

Problems mostly about material since 2nd midterm.

Problems taken from Prof. Roberts' Math 3080 Exams given Spring 2004

(1.) *Data was collected from an experiment to investigate the effects of 3 herbicides and 4 levels of nitrogen on the yield of wheat. For each combination of nitrogen level and herbicide type, two observations were made. We have the partial results from the Anova table below.*

ANALYSIS OF VARIANCE *Yield of Wheat*

SOURCE	*DF*	*SS*	*MS*	*F*
Herbicide		*12.00*		
Nitrogen Level		*240.00*		
INTERACTION		*36.00*		
ERROR				
TOTAL		*431.33*		

i) Complete the ANOVA table. Perform the appropriate hypothesis tests to examine the effects of herbicide and nitrogen on the yield of wheat. For each test you perform be brief, stating only the null hypothesis, the value of the test statistic you use, and your statistical conclusion as compared to the correct critical value. Use level of significance $\alpha = .05$.

ii) Summarize your conclusions briefly in the practical context of the problem.

iii) In the above problem, suppose that the observations in the ij-th cell are X_{ijk}. *What assumptions about* X_{ijk} *do the above tests require?*

This is is a two factor fixed effects analysis of variance. The observations

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

where μ, α_i, β_j and γ_{ij} are constants so that $\sum_i \alpha_i = 0$, $\sum_j \beta_j = 0$, $\sum_j \gamma_{ij} = 0$ for each i, $\sum_i \gamma_{ij} = 0$ for each j and $\epsilon_{ijk} \sim N(0, \sigma^2)$ are independent, identically distributed normal random variables.

There are $I = 3$ herbicide levels, $J = 4$ nitrogen levels and $K = 2$ replications, all together $n = IJK = 24$ observations. There are $I-1 = 2$, $J-1 = 3$, $(I-1)(J-1) = 6$, $IJ(k-1) = 12$ and $n-1 = 23$ degrees of freedom for herbicide, nitrogen, interaction, error and total. By the sum of squares identity, $SSE = SST - SSA - SSB - SSAB = 431.33 - 12.00 - 240.00 - 36.00 = 143.33$. $MS = SS/DF$ so $MSA = 12.00/2 = 6.00$, $MSB = 240.00/3 = 120.00$, $MSAB = 36.00/6 = 6.00$ and $MSE = 143.33/12 = 11.94417$. Finally, $F = MS/MSE$ in fixed effects model, so $FA = 6.00/MSE = .5023373$, $FB = 10.04675$ and $FAB = 6.00/MSE = .5023373$. Thus we are able to complete the ANOVA table.

Source of Variation	DF	SS	MS	F
Herbicide	2	12.00	6.00	.5023373
Nitrogen	3	240.00	120.00	10.04675
Interaction	6	36.00	6.00	.5023373
Error	12	143.33	11.94	
Total	23	431.33		

The first test of hypothesis is whether interactions are significant

$$\begin{array}{lll} \mathcal{H}_{AB\ 0}: & \gamma_{ij} = 0 & \text{for all } i, j; \quad \text{v.} \\ \mathcal{H}_{AB\ 1}: & \gamma_{ij} \neq 0 & \text{for some } i, j. \end{array}$$

Under $\mathcal{H}_{AB\ 0}$, $FAB \sim f_{(I-1)(J-1),IJ(K-1)}$ so we reject $\mathcal{H}_{AB\ 0}$ if $FAB > f_{6,12}(.05) = 3.00$. Since FAB is less, we are unable to reject $\mathcal{H}_{AB\ 0}$: the interactions are plausibly negligible.

The second test of hypothesis is whether the herbicide effect is significant

$$\begin{array}{lll} \mathcal{H}_{A\ 0}: & \alpha_i = 0 & \text{for all } i; \quad \text{v.} \\ \mathcal{H}_{A\ 1}: & \alpha_i \neq 0 & \text{for some } i. \end{array}$$

Under $\mathcal{H}_{A\ 0}$, $FA \sim f_{I-1,IJ(K-1)}$ so we reject $\mathcal{H}_{A\ 0}$ if $FA > f_{2,12}(.05) = 3.89$. Since FA is less, we are unable to reject $\mathcal{H}_{A\ 0}$: the herbicide effect is plausibly negligible.

The final test of hypothesis is whether the nitrogen effect is significant

$$\begin{array}{lll} \mathcal{H}_{B\ 0}: & \beta_j = 0 & \text{for all } j; \quad \text{v.} \\ \mathcal{H}_{B\ 1}: & \beta_j \neq 0 & \text{for some } j. \end{array}$$

Under $\mathcal{H}_{B\ 0}$, $FAB \sim f_{J-1,IJ(K-1)}$ so we reject $\mathcal{H}_{AB\ 0}$ if $FAB > f_{2,12}(.05) = 3.49$. Since FB is larger, we reject $\mathcal{H}_{AB\ 0}$: the nitrogen effect is significant at the 0.05 level.

(2.) *A study of the hours of relief provided by 5 different headache tablets was conducted on 25 patients assigned at random, 5 to each type of tablet. The mean and standard deviation of the sample for each tablet is reported in the table below. Assume that the hours of relief for each tablet was approximately normally distributed with the same variance for each tablet. Use this information to answer the following questions.*

	A	B	C	D	E
Mean	5.44	7.90	4.30	2.98	6.96
Stdev	1.89	1.34	2.11	1.30	1.85

The grand mean is $X_{\cdot\cdot} = 5.516$ *and* $\sum\sum X_{ij}^2 = 929.974$.

i) Set up the Anova table for this experiment. Do the sample results indicate that there is a difference in the mean hours of relief offered by the different tablets? Use $\alpha = .05$.

ii) Using Tukey's method to determine which pairs of tablets differ significantly in the mean hours of relief provided by the tablets. Use $\alpha = .05$.

Let us construct the ANOVA table. This is a one-way fixed effects model. That means that we assume the observations

$$X_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

where μ, α_i are constants so that $\sum_i \alpha_i = 0$ and $\epsilon_{ij} \sim N(0, \sigma^2)$ are independent, identically distributed normal random variables. In this case there are $I = 5$ levels of the factor and $J = 5$ replicates and $n = IJ = 25$ observations. Hence the degrees of freedom is a $I-1 = 4$ for the tablet A, $I(J-1) = 20$ for the error and $IJ-1 = 24$ for the total. $SST = \sum_{ij} X_{ij}^2 - n(\overline{X_{cdot}\,\cdot})^2 = 190.134$. $SSA = J(\sum_i (\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2) = 78.4216$. Thus $SSE = SST - SSA = 169.3176 - 78.4216 = 90.896$. $MS = SS/DF$ so $MSA = 78.4216/4 = 19.6054$ and $MSE = 90.896/20 = 4.5448$. Finally $F = MSA/MSE = 19.6054/4.5448 = 4.313809$. The ANOVA table is

Source of Variation	DF	SS	MS	F
A	4	78.4216	19.6054	4.313809
Error	20	90.896	4.5448	
Total	24	169.3176		

To test the hypothesis whether the effect is significant

$$\begin{array}{lll} \mathcal{H}_0: & \alpha_i = 0 & \text{for all } i; \quad \text{v.} \\ \mathcal{H}_1: & \alpha_i \neq 0 & \text{for some } i. \end{array}$$

Under $\mathcal{H}_0$, $F \sim f_{I-1, I(J-1)}$ so we reject $\mathcal{H}_0$ if $F > f_{4,20}(.05) = 2.87$. Since F is larger, we reject $\mathcal{H}_0$: the hours of relief are significantly different at the 0.05 level.

To determine which differences are significant, we compute the Tukey's HSD. In this case the critical value of the Studentized Range is $q(.05, 5, 20) = 4.23$ so

$$HSD = q(\alpha, I, I(J-1))\sqrt{\frac{MSE}{J}} = 4.23\sqrt{\frac{4.5448}{5}} = 4.03.$$

Only $|\bar{X}_D - \bar{X}_B| > HSD$ and the other differences are smaller than HSD. Sorting the means, we put Tukey Bars under subsets whose pairs that are not significantly different.

	D	C	A	E	B
Mean	2.98	4.30	5.44	6.96	7.90

(3.) *A random sample of 445 college students was classified according to frequency of marijuana use and parental use of alcohol and psychoactive drugs. Does the data suggest that parental usage and student usage are independent in the population sampled?*

		Student Usage			
		Never	Occasional	Regular	Total
	Neither	141	54	40	235
Parental Usage	One	68	44	51	163
	Both	17	11	19	47
	Total	226	109	110	445

What is the null hypothesis here? Assuming the null hypothesis is true, what is the probability and the expected number for the "Neither $*$ Never" cell? The total of the contributions to the test statistic from all cells except "Neither $*$ Never" cell is: 18.5. *Find the contribution from that cell and state the value of this test statistic result. give the approximate p-value of your test statistic. State what the p-value of your test statistic means in the practical terms of the problem. Finally, what conclusion would you recommend and why?*

This is a χ^2 test for independence. The null hypothesis is

$$\begin{array}{lll} & \mathcal{H}_0: & p_{ij} = p_{i\cdot}p_{\cdot j} & \text{for all } i, j; \\ \text{v.} & \mathcal{H}_1: & p_{ij} \neq p_{i\cdot}p_{\cdot j} & \text{for some } i, j. \end{array}$$

Under $\mathcal{H}_0$, the expected probability of parental neither and student never is $\hat{p}_{11} = \hat{p}_{1\cdot}\hat{p}_{\cdot 1} = (235/445)(226/445) = 0.2681985$. Thus the expected number is $\hat{e}_{11} = \hat{p}_{11}n = 445 \times 0.2681985 = 119.3483$. The test statistic is, according to the given data

$$\chi^2 = \sum_{i,j} \frac{(y_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} = \frac{(y_{11} - \hat{e}_{11})^2}{\hat{e}_{11}} + \sum_{(i,j)\neq(1,1)}^{J} \frac{(y_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} = \frac{(141 - 119.3483)^2}{119.3483} + 18.5 = 22.4.$$

Under $\mathcal{H}_0$, the statistic is distributed according to $\chi^2_{(I-1)(J-1)}$. The P-value is thus $P = P(\chi^2_4 > 22.4) < .005$ from the table. (In fact, $P = 0.0001689$.) Thus we reject $\mathcal{H}_0$: the data shows that the rows and columns are not independent.

Problems taken from my Math 3080 Exams given Spring 2005

(1.) *Complete the ANOVA table, given that there were five samples taken from each of the three populations.*

Source of Variation	df	SS	MS	f	P-value	F-crit
Treatment			9.8			
Error						
Total		31.6				

(2.) *The yield of wheat in bushels per acre were compared for five different vaieties A, B, C, D, E and at six different locations. Each variety was randomly assigned to a plot at each location. The ANOVA table for the data, generated by Excel, is shown below, with $\alpha = .05$. Use the table to answer the questions.*

Anova: Two-Factor Without Replication

SUMMARY	Count	Sum	Average	Variance
A	6	206.2	34.36667	5.97467
B	6	193.7	32.23333	2.23767
C	6	217.8	36.3	4.912
D	6	220.7	36.73333	4.22167
E	6	187.2	31.2	3.816
Loc 1	5	171.5	34.3	8.285
Loc 2	5	161.7	32.34	7.163
Loc 3	5	162.1	32.42	5.152
Loc 4	5	174.6	34.92	11.972
Loc 5	5	183.6	36.72	5.897
Loc 6	5	172.1	34.42	6.717

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Rows	142.70467	4	35.67617	18.75751	0.000001516	2.86608
Columns	67.77067	5	13.55413	7.12638	0.000565452	2.71089
Error	38.03933	20	1.90197			
Total	248.51467	29				

a. *State the assumptions that are being made in order to do the ANOVA test.*

Random blocks design which is the same as the two-factor without replication design. $X_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$ where $i = 1, \ldots, I$, $j = 1, \ldots, J$, μ, α_i and β_j are constants so that $\sum_i \alpha_i = 0$, $\sum_j \beta_j = 0$, $\epsilon_{ij} \sim N(0, \sigma)$ is a random sample taken from normal RV.

b. *Do the data provide sufficient evidence to indicate (at the $\alpha = .05$ level) a difference in mean yield for the five varieties of wheat?*

Yes. The P-value for f_{Variety} is $0.00000152 < .05$ so is significant.

c. *If another plot at location 2 were planted variety C, what is your best guess of the yield? Why?*

Note $\bar{x}_{\cdot\cdot} = \frac{1}{5}(34.37 + 32.23 + 36.30 + 36.73 + 31.20) = 34.4533$. Since $E(X_{32}) = \mu + \alpha_3 + \beta_2 = E(\bar{x}_{3\cdot}) + E(\bar{x}_{\cdot 2}) - E(\bar{x}_{\cdot\cdot})$ we take the estimator $\bar{x}_{3\cdot} + \bar{x}_{\cdot 2} - \bar{x}_{\cdot\cdot} = 36.30 + 32.34 - 34.45 = 34.19$.

(3.) *From the data for problem (2), use Tukey's procedure (at the $\alpha = 0.5$ level) to identify significant differences among the different varieties. Order the means and indicate the relationship between the means using underscoring.*

Tukey's test is to compare all differences simultaneously with $w = q\sqrt{\frac{MSE}{J}}$ where the one tailed, level α studentized range statistic for $I = 5$ levels is $q = q_{\alpha, I, (I-1)(J-1)} = q_{.05,5,20} = 4.23$. So $w = 4.23 \cdot \sqrt{\frac{1.90197}{6}} = 2.38$.

Since $\bar{x}_{4,\cdot} - \bar{x}_{1\cdot} = 2.36 < w < \bar{x}_{4,\cdot} - \bar{x}_{2\cdot} = 4.50$, $\bar{x}_{1,\cdot} - \bar{x}_{2\cdot} = 2.13 < w < \bar{x}_{3,\cdot} - \bar{x}_{2\cdot} = 4.07$ and $\bar{x}_{2,\cdot} - \bar{x}_{5\cdot} = 1.03 < w < \bar{x}_{1,\cdot} - \bar{x}_{5\cdot} = 3.17$, evidence allows us to find conclude that means E and B do not differ; means B and A do not differ; means A, C and D do not differ; although C and D differ from B, neither differs from A; although A and E differ from each other, neither differs from B.

The underscoring pattern is:

```
  E         B         A         C         D
31.20     32.23     34.37     36.30     36.73
==============
          ++++++++++++++
                    -----------------------
```

(3.) *Consider the following experiment to test whether or not three different college entrance exams are interchangeable. A list of all colleges in the United States with 100 or more students was obtained, and 20 schools were randomly selected from the list. Then 30 students were randomly sampled from each of the twenty schools and randomly divided into three groups of 10 students each. The first group took Exam A, the second took Exam B, and the third took Exam C. Partial results of the two way ANOVA with "Exam" as one of the factors and "School" as the other is listed. State the model. What assumptions are made about the data in order to make the ANOVA analysis? Complete the ANOVA table. Carry out the tests of the three hypotheses. What is your conclusion?*

Two factor mixed fixed/random effects model with replication. $X_{ijk} = \mu + \alpha_i + B_j + G_{ij} + \varepsilon_{ijk}$ for $i = 1, \ldots, I$, $j = 1, \ldots J$ and $k = 1, \ldots, K$ where μ and α_i are constant, such that $\sum_i \alpha_i = 0$. The B_i, G_{ij} and ε_{ijk} are independent normal RV's with mean zero and $V(B_j) = \sigma_B^2$, $V(G_{ij}) = \sigma_G^2$ and $V(\varepsilon_{ijk}) = \sigma^2$, constant, but unknown variances. A is the fixed effect (test) and B is the random effect (School.) State the three hypotheses of interest, and their corresponding tests.

ANOVA

Source of Variation	df	SS	MS	F	F crit
Exam	2	589	294.5	1.950	3.25
School	19	2833	151.7	1.005	1.84
Exam*School	38	5738	151.0	1.756	1.42
Error	540	46440	86.0		
Total	599	56650			

$\mathcal{H}_{\mathcal{AB}'}$: $\sigma_G^2 = 0$. Compute $f_{AB} = \frac{MSAB}{MSE}$ to reject $\mathcal{H}_{\mathcal{AB}'}$ if $f_{AB} > f_{.05,38,540}$. $\mathcal{H}_{\mathcal{A}'}$: $\alpha_1 = \cdots = \alpha_I = 0$. Compute $f_A = \frac{MSA}{MSAB}$ to reject $\mathcal{H}_{\mathcal{A}'}$ if $f_A > f_{.05,2,38}$. $\mathcal{H}_{\mathcal{B}'}$: $\sigma_B^2 = 0$. Compute $f_B = \frac{MSB}{MSAB}$ to reject $\mathcal{H}_{\mathcal{B}'}$ if $f_B > f_{.05,19,38}$.

The interaction term is significant. Normally this ends the analysis. It turns out that neither other effect is significant.

(1.) Answer:

ANOVA

Source of Variation	df	SS	MS	F	F crit
Treatment	2	19.6	9.8	9.8	3.89
Error	12	12.0	1.0		
Total	14	31.6			

(4.)*In a study about ball bearing life, Box (1992) considered four fixed factors in the manufacturing process: XX=type of ball, YY=cage design, ZZ=type of grease and WW=amount of grease and $life$=relative time. Each of the factor had two levels: $1 =$ "standard" or $2 =$ "modified." Test at the 95% confidence level whether any of the main effects are significant. State your hypotheses. Here is the printout of the data file. Also, here is a partial printout of the ANOVA table made requesting all interactions of the four variables in the model. The rest of the entries in the table were* "`undefined.`"

XX	YY	ZZ	WW	life
1	1	1	1	0.31
2	2	1	1	2.17
2	1	2	1	1.37
2	1	1	2	1.38
1	2	2	1	0.92
1	2	1	2	0.73
1	1	2	2	0.95
2	2	2	2	2.57

	DF	SS
CONSTANT	1	13.52
XX	1	2.622
YY	1	0.70805
XX.YY	1	0.32
ZZ	1	0.18605
XX.ZZ	1	0.0242
YY.ZZ	1	0.0002
XX.YY.ZZ	1	0.09245

This is a 2^4 fixed effects experiment with half replication. From the eight data points and factor levels, it is evident that $XYZW$ was the defining contrast and the experimental treatments were the eight combinations with an even number of effects in common with $XYZW$. In particular XYZ is confounded with W. There is one degree of freedom for each of the sum of squares in the table. To test whether any of the main effects are significant, we assume that there are no interactions, and use the remaining sum of squares for a proxy for SSE. Thus we get $SSE1 = SSXY + SSXZ + SSYZ = 0.32 + 0.0242 + 0.0002 = 0.34440$ and $SST = SSX + SSY + SSZ + SSW + SSE1 = 2.622 + 0.70805 + 0.18605 + 0.09245 + 0.34440 = 3.95295$. The resulting ANOVA table is thus

Source Var.	DF	Sum of Sq.	Mean Square	F
XX	1	2.622	2.622	22.8397213
YY	1	0.70805	0.70805	6.1676829
ZZ	1	0.18605	0.18605	1.6206446
WW	1	0.09245	0.09245	0.8053136
Error 1	3	0.34440	0.11480	
Total	7	3.95295		

To test $\mathcal{H}_{A0} : \alpha_1 = 0$ vs $\mathcal{H}_{A\infty} : \alpha_\infty \neq \prime$ we accept $\mathcal{H}_{A1}$ if $f > F_{\alpha,1,3}$. The other factors are tested similarly. For example, if $\alpha = 0.10$ then $F_{.10,1,3} = 5.54$ so there is no strong evidence that the ZZ and WW factors is nonzero whereas XX and YY are significantly nonzero. If $\alpha = .05$ then $F_{.05,1,3} = 10.13$ so XX is significant at this level too.

More Problems

(1.) *A 2001 study by Y. Li & J. Chen investigated the effect of levels of several factors on glucose consumption (in g/L.) A single measurement is provided for each combination of factors.*

A	B	C	Glucose Consumption
-1	-1	-1	68.0
1	-1	-1	77.5
-1	1	-1	98.0
1	1	-1	98.0
-1	-1	1	74.0
1	-1	1	77.0
-1	1	1	97.0
1	1	1	98.0

i) *Compute estimates for the main effects and the interactions. Is it possible to compute an error sum of squares? Explain.*

Using Yates's method (not in Rosenkrantz), where columns 1, 2, EC are sums in pairs followed by differences in pairs of the previous columns, we get

Treatment Condition	x_{ijk}	1	2	Effects Contrast	Effect	$SS = (\text{contrast})^2/8$
1	68	145.5	341.5	687.5	$\hat{\mu} = 85.94$	59082.03
a	77.5	196	346	13.5	$\hat{\alpha}_1 = 1.69$	22.78
b	98	151	9.5	94.5	$\hat{\beta}_1 = 11.81$	1116.28
ab	98	195	4	-11.5	$\hat{\gamma}_{11}^{AB} = -1.44$	16.53
c	74	9.5	50.5	4.5	$\hat{\delta}_1 = 0.56$	2.53
ac	77	0	44	-5.5	$\hat{\gamma}_{11}^{AC} = -.69$	3.78
bc	97	3	-9.5	-6.5	$\hat{\gamma}_{11}^{BC} = -.81$	2.64
abc	98	1	-2	7.5	$\hat{\gamma}_{111}^{ABC} = .94$	7.03

Thus the effects are gotten by, e.g., $\alpha_1 = (\text{effects contrast})/8 = 13.5/8 = 1.69$. It is not possible to compute an error sum of squares, as there are no replications, there is no error sum of squares.

The alternative to Yates' Method is to form the sums with the treatment basis signs. $e_1 = (1,1,1,1,1,1,1,1)$, $e_a = (-1,1,-1,1,-1,1,-1,1)$, $e_b = (-1,-1,1,1,-1,-1,1,1)$ and $e_c = (-1,-1,-1,-1,1,1,1,1)$ so $L_1 = \langle e_1, x_{ijk}\rangle = 68 + 77.5 + \cdots + 98 = 687.5$, $L_a = \langle e_a, x_{ijk}\rangle = -68 + 77.5 - 98 + 98 - \cdots + 98 = 13.5$, $L_b = \langle e_b, x_{ijk}\rangle = -68 - 77.5 + 98 + 98 - \cdots + 98 = 94.5$, $L_c = \langle e_1, x_{ijk}\rangle = -68 - \cdots - 98 + 74 + \cdots + 98 = 4.5$. The others follow the same pattern. But to get ANOVA table, these contrasts suffice.

ii) *Are there any interactions among the larger effects? If so, which ones?*

No, none of the interaction terms are nearly as large as the effect of factor B.

iii) *Assume that it is known from past experience that the additive model holds. Add the sums of squares for the interactions and use the result to in place of an error sum of squares to test the hypothesis that the main effects are equal to zero.*

If the additive model holds, then we may add the higher order interactions to get an estimate for the σ^2 to replace the error sum of squares. We have $\sum x_{ijk}^2 = 60256.25$ and $x_{\dots}^2/8 = 59082.03$ so that $SST = 60256.25 - 59082.03 = 1174.22$. Thus $SSE = SST - (SSA + SSB + SSC) = 1174.22 - (22.82 + 1116.28 + 2.53) = 32.63$ so we construct the ANOVA table. There is strong evidence that the main effect B is nonzero whereas not enough evidence to rule out that the other effects are zero.

Source	df	Sum of Squares	Mean Square	f	P-value
A	1	22.78	22.78	2.78	.1699
B	1	116.28	1116.28	136.86	.00031
C	1	2.53	2.53	.31	.6072
Error	4	32.63	32.63	8.16	
Total	7	1174.22			

(2.) *In a 2^5 experiment with half replication, suppose it is determined to confound E with $ABCD$. Which treatment combinations should be used to carry out the experiment? List the resulting alias pairs.*

Since $ABCD$ and E are complementary, the word $ABCDE$ is being used to determine the halves. Thus to list the set of experimental treatment combinations, we list all words which have either an odd number of letters in common with the defining word $ABCDE$ or an even number. Thus, using the odd ones, the experimental treatments are

$$\{A, B, C, D, E, ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE, BDE, CDE, ABCDE\}$$

The aliased pairs are the words in the odd group and their generalized inteaction with the defining word $ABCDE$, namely,

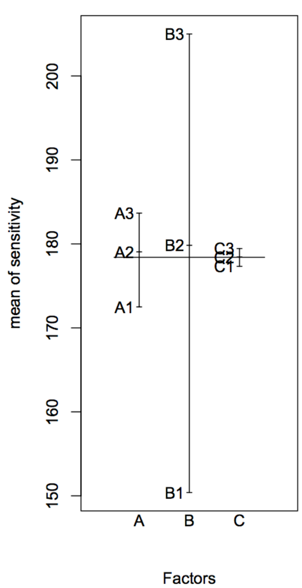
$$\begin{array}{cccc} \{1, ABCDE\} & \{A, BCDE\} & \{B, ACDE\} & \{C, ABDE\} \\ \{D, ABCE\} & \{E, ABCD\} & \{ABC, DE\} & \{ABD, CE\} \\ \{ABE, CD\} & \{ACD, BE\} & \{ACE, BD\} & \{ADE, AD\} \\ \{BCD, AE\} & \{BCE, AD\} & \{BDE, AC\} & \{CDE, AB\} \end{array}$$

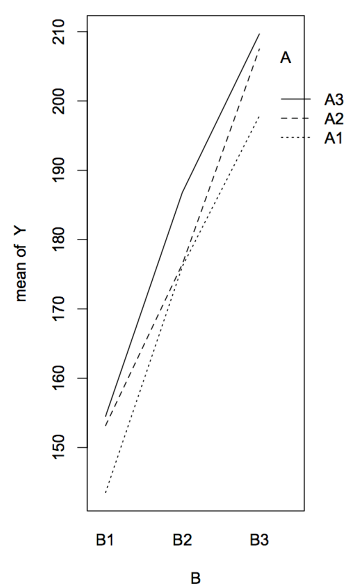
(3.) *A study "SN-Radio for the Quality Evaluation," (Japanese Standards Association, 1988) considered three mechanical and electrical factors that affect the measured sensitivity of a control valve: A relative position of control bolt (center−0.5, center, cemter+0.5), B control bolt range (2, 4, 5 and 7 mm) and C input voltage (100, 120, 150 V).*

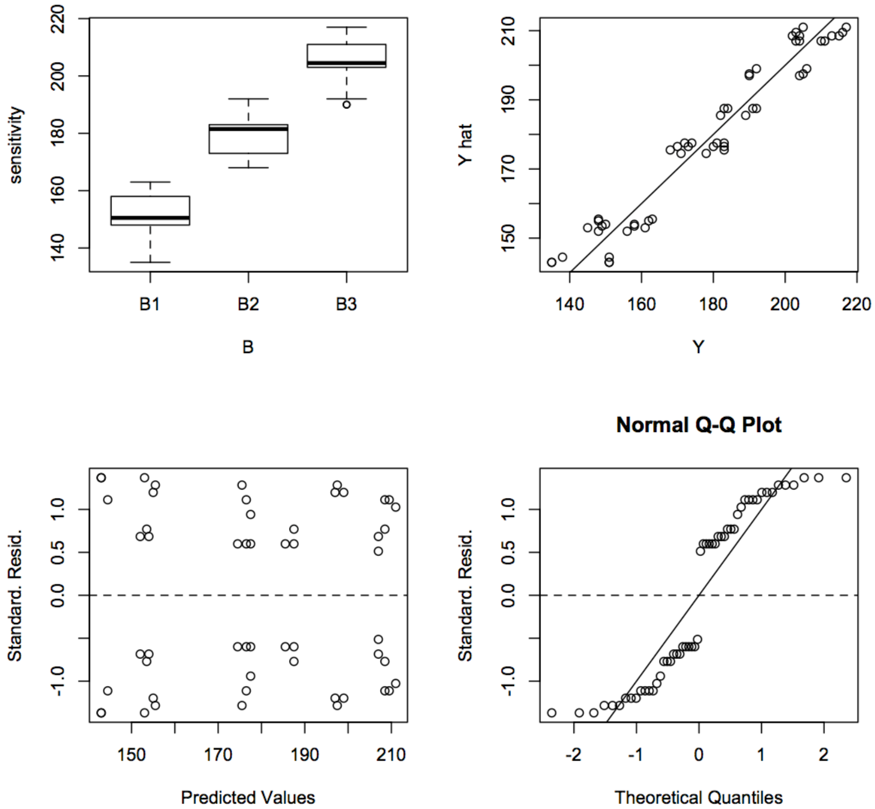
		C1		C2		C3	
	B1	151	135	151	135	151	138
A1	B2	178	171	180	173	181	174
	B3	204	190	205	190	206	192
	B1	156	148	158	149	158	150
A2	B2	183	168	183	170	183	172
	B3	210	204	211	203	213	204
	B1	161	145	162	148	163	148
A3	B2	189	182	191	184	192	183
	B3	215	202	216	203	217	205

Perform an analysis of variance to test for significant main and interaction effects. Find the coefficient of determination, R^2. State the model, the test statistics, the rejection regions, and your conclusions. The partial **R** *output, interaction plots and diagnostic plots are attached. Comment about what each plot says about the appropriateness of the model.*

```
\Response: sensitivity
          Df  Sum Sq Mean Sq  F value    Pr(>F)
A          2  1133.6   566.8   8.2991  0.001551
B          2 26896.3 13448.1 196.9086 < 2.2e-16
C          2    40.1    20.1   0.2939  0.747686
A:B        4   216.5    54.1   0.7926  0.540317
A:C        4     1.6     0.4   0.0060  0.999924
B:C        4     2.3     0.6   0.0084  0.999850
A:B:C      8     2.6     0.3   0.0047  1.000000
Residuals 27  1844.0    68.3
```







This is a three-way fixed effects anova with replicates. The model is

$$y_{ijk\ell} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijk\ell}$$

where μ, α_i, β_j, γ_k, $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$, $(\beta\gamma)_{jk}$, $(\alpha\beta\gamma)_{ijk}$ are constants such that $\sum_i \alpha_i = 0$, $\sum_j \beta_j = 0$, $\sum_k \gamma_k = 0$, $\sum_i (\alpha\beta)_{ij} = 0$ for all j, $\sum_j (\alpha\beta)_{ij} = 0$ for all i, $\sum_i (\alpha\gamma)_{ik}$ for all k, $\sum_k (\alpha\gamma)_{ik}$ for all i, $\sum_j (\beta\gamma)_{jk}$ for all k, $\sum_k (\beta\gamma)_{jk}$ for all j, $\sum_i (\alpha\beta\gamma)_{ijk} = 0$ for all (j,k), $\sum_j (\alpha\beta\gamma)_{ijk} = 0$ for all (i,k), $\sum_k (\alpha\beta\gamma)_{ijk} = 0$ for all (i,j) and $\epsilon_{ijk\ell} \sim N(0, \sigma^2)$ are IID normal variables. In our case, $I = J = K = 3, L = 2$.

The first test of hypothesis is whether three-way interactions are significant

$$\begin{array}{lll} \mathcal{H}_{ABC\ 0}: & (\alpha\beta\gamma)_{ijk} = 0 & \text{for all } i,j,k; \qquad \text{v.} \\ \mathcal{H}_{ABC\ 1}: & (\alpha\beta\gamma)_{ijk} \neq 0 & \text{for some } i,j,k. \end{array}$$

Under $\mathcal{H}_{ABC\ 0}$, $FABC \sim f_{(I-1)(J-1)(K-1), IJK(L-1)}$ so we reject $\mathcal{H}_{ABC\ 0}$ if $FABC > f_{8,27}(.05)$. Since the P value is $P = 1.000$, we are unable to reject $\mathcal{H}_{ABC\ 0}$: the 3-way interactions are plausibly negligible.

The second tests of hypothesis is whether two-way interactions are significant. BC and AC are similar to AB

$$\begin{array}{lll} \mathcal{H}_{AB\ 0}: & (\alpha\beta)_{ij} = 0 & \text{for all } i,j; \qquad \text{v.} \\ \mathcal{H}_{AB\ 1}: & (\alpha\beta)_{ij} \neq 0 & \text{for some } i,j. \end{array}$$

Under $\mathcal{H}_{AB\ 0}$, $FAB \sim f_{(I-1)(J-1), IJK(L-1)}$ so we reject $\mathcal{H}_{AB\ 0}$ if $FAB > f_{4,27}(.05)$. Since the P value is $P = .540$, we are unable to reject $\mathcal{H}_{AB\ 0}$. Similarly, the P-values for $\mathcal{H}_{AC0}$ and $\mathcal{H}_{BC0}$ are both .999, we fail to reject those also: all 2-way interactions are plausibly negligible.

The last tests of hypothesis is whether the main effects are significant. The B and C effects are tested similarly.

$$\begin{aligned} &\mathcal{H}_{A\ 0}: && \alpha_i = 0 && \text{for all } i; && \text{v.} \\ &\mathcal{H}_{A\ 1}: && \alpha_i \neq 0 && \text{for some } i. \end{aligned}$$

Under $\mathcal{H}_{A\ 0}$, $FA \sim f_{I-1,IJK(L-1)}$ so we reject $\mathcal{H}_{A\ 0}$ if $FA > f_{2,27}(.05)$. Since the P value is $P = .0016$, we reject $\mathcal{H}_{A\ 0}$. Similarly, the P-values for $\mathcal{H}_{B0}$ is $P = 2.2 \times 10^{-16}$ so we reject $\mathcal{H}_{B0}$ also. The P value for $\mathcal{H}_{C0}$ is .748, we fail to reject it: the main A and B effects are significant, the C effect is not. We could perform Tukey's HSD family comparisons on the A and B factors at this point.

The design plot shows that factor B seems to have a large influence on the sensitivity. The interaction plot shows that the lines are quite parallel, so we expect no AB interaction.

The box plot shows that the B factor accounts for widely different means. The size of the boxes are similar, indicating the same variance for each level of the factor B are approximately equal as we require of the model.

The plot of standardized residuals v. fitted values again shows that the data is "tubular," meaning that the spread is constant for different fitted values, as required of the model. This data has a curious almost perfect symmetry suggesting that the B factor has symmetric influence.

The plot of fitted values $\hat{y}_{ijk\ell}$ versus observed values $y_{ijk\ell}$ shows good alignment with the 45° line, again showing that the variance is uniform across the data, as we expect. The spread is not large indicating that the model explains a lot of the variation.

Finally, the QQ-plot is significantly S-shaped. In other words, the large observed values of the residuals are not as large as the corresponding normal quantiles: the distribution of the residuals light tailed and is not very normal. In this case, a transformation of the data is in order. In fact the transformation $z = ((y - \hat{\mu})/s)^3$ spreads the points and straightens our the S-shape.

To compute the coefficient of determination, $SST = SSA + SSB + \cdots + SSABC = 1133.6 + 26896.3 + 40.1 + 216.5 + 1.6 + 2.3 + 2.6 + 1844.0 = 30137.0$ so

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{1844.0}{30137.0} = .9388.$$

(4.) *Consider a balanced one-way random effects analysis of variance model.*

a. State the model and its assumptions.

b. Prove the sum of squares identity $SST = SSA + SSE$.

c. Find $\mathbb{E}(SSA)$ and $\mathbb{E}(SSE)$ for this model.

a. The balanced single factor random-effects model assumes that the observations are random variables with I factors and the same number J observations per level

$$Y_{ij} = \mu + A_i + \epsilon_{ij}$$

for all $i = 1, \ldots, I$ and $j = 1, \ldots, J$ where μ is constant, and the A_i's and ϵ_{ij}'s are mutually independent random variables such that $A_i \sim N(0, \sigma_A^2)$ for each i and $\epsilon_{ij} \sim N(0, \sigma^2)$ for each i, j are normal, where σ_A and σ are two constants.

b. Using the usual dot notation to indicate summing over a subscript or bar-dot for averaging, we have to relate three sums of squares

$$SST = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(Y_{ij} - \bar{Y}_{\cdot\cdot}\right)^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ij}^2 - IJ\bar{Y}_{\cdot\cdot}^2,$$

$$SSA = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\right)^2 = J\sum_{i=1}^{I} Y_{i\cdot}^2 - IJ\bar{Y}_{\cdot\cdot}^2,$$

$$SSE = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ij}^2 - J\sum_{i=1}^{I}\bar{Y}_{i\cdot}^2.$$

The vanishing of the cross term happens for any model because

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\left(Y_{ij} - \bar{Y}_{i\cdot}\right)\left(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\right) = \sum_{i=1}^{I}\left(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\right)\sum_{j=1}^{J}\left(Y_{ij} - \bar{Y}_{i\cdot}\right) = \sum_{i=1}^{I}\left(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\right)\left(J\bar{Y}_{i\cdot} - J\bar{Y}_{i\cdot}\right) = 0.$$

Hence

$$\begin{aligned}
SST &= \sum_{i=1}^{I}\sum_{j=1}^{J}\left(Y_{ij} - \bar{Y}_{\cdot\cdot}\right)^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\left(Y_{ij} - \bar{Y}_{i\cdot} + \bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\right)^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\left\{\left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2 + 2\left(Y_{ij} - \bar{Y}_{i\cdot}\right)\left(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\right) + \left(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\right)^2\right\} \\
&= SSE + 0 + SSA.
\end{aligned}$$

c. The expected values of the sum of squares uses the formula for variance of a random variable X, namely $\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X)$. Thus, using independence,

$$\mathbb{E}(Y_{ij}) = \mathbb{E}\left(\mu + A_i + \epsilon_{ij}\right) = \mu,$$

$$\mathbb{E}(Y_{ij}^2) = \mathbb{V}(Y_{ij}) + \mathbb{E}^2(Y_{ij}) = \mathbb{V}\left(\mu + A_i + \epsilon_{ij}\right) + \mu^2 = \sigma_A^2 + \sigma^2 + \mu^2,$$

$$\bar{Y}_{i\cdot} = \frac{1}{J}\sum_{j=1}^{J} Y_{ij} = \frac{1}{J}\sum_{j=1}^{J}\left(\mu + A_i + \epsilon_{ij}\right) = \mu + A_i + \frac{1}{J}\sum_{j=1}^{J}\epsilon_{ij}$$

$$\mathbb{E}(\bar{Y}_{i\cdot}) = \mathbb{E}\left(\mu + A_i + \frac{1}{J}\sum_{j=1}^{J}\epsilon_{ij}\right) = \mu,$$

$$\mathbb{E}(\bar{Y}_{i\cdot}^2) = \mathbb{V}\left(\mu + A_i + \frac{1}{J}\sum_{j=1}^{J}\epsilon_{ij}\right) + \mathbb{E}^2(\bar{Y}_{i\cdot}) = \sigma_A^2 + \frac{1}{J}\sigma^2 + \mu^2,$$

$$\bar{Y}_{\cdot\cdot} = \frac{1}{I}\sum_{i=1}^{I}\bar{Y}_{i\cdot} = \frac{1}{I}\sum_{i=1}^{I}\left(\mu + A_i + \frac{1}{J}\sum_{j=1}^{J}\epsilon_{ij}\right) = \mu + \frac{1}{I}\sum_{i=1}^{I}A_i + \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\epsilon_{ij}$$

$$\mathbb{E}(\bar{Y}_{\cdot\cdot}) = \mathbb{E}\left(\mu + \frac{1}{I}\sum_{i=1}^{I}A_i + \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\epsilon_{ij}\right) = \mu$$

$$\mathbb{E}(\bar{Y}_{\cdot\cdot}^2) = \mathbb{V}(\bar{Y}_{\cdot\cdot}) + \mathbb{E}^2(\bar{Y}_{\cdot\cdot}) = \mathbb{V}\left(\mu + \frac{1}{I}\sum_{i=1}^{I}A_i + \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\epsilon_{ij}\right) + \mu^2 = \frac{1}{I}\sigma_A^2 + \frac{1}{IJ}\sigma^2 + \mu^2$$

Putting these together we have the formulæ for the expectations of the sums of squares

$$\begin{aligned}\mathbb{E}(SSA) &= \mathbb{E}\left(J\sum_{i=1}^{I}\bar{Y}_{i\cdot}^2 - IJ\bar{Y}_{\cdot\cdot}^2\right) = J\sum_{i=1}^{I}\mathbb{E}\left(\bar{Y}_{i\cdot}^2\right) - IJ\,\mathbb{E}\left(\bar{Y}_{\cdot\cdot}^2\right)\\ &= J\sum_{i=1}^{I}\left(\sigma_A^2 + \frac{1}{J}\sigma^2 + \mu^2\right) - IJ\left(\frac{1}{I}\sigma_A^2 + \frac{1}{IJ}\sigma^2 + \mu^2\right)\\ &= J(I-1)\sigma_A^2 + (I-1)\sigma^2\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}(SSE) &= \mathbb{E}\left(\sum_{i=1}^{I}\sum_{j=1}^{J}Y_{ij}^2 - J\sum_{i=1}^{I}\bar{Y}_{i\cdot}^2\right)\\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}\mathbb{E}\left(\bar{Y}_{ij}^2\right) - J\sum_{i=1}^{I}\mathbb{E}\left(\bar{Y}_{i\cdot}^2\right)\\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}\left(\sigma_A^2 + \sigma^2 + \mu^2\right) - J\sum_{i=1}^{I}\left(\sigma_A^2 + \frac{1}{J}\sigma^2 + \mu^2\right)\\ &= I(J-1)\sigma^2.\end{aligned}$$

These agree with equations (12.22 – 12.23) of text (WAY TO GO UTES!)
(5.) *Consider a* 2^3 *fixed effects analysis of variance model with* n *replicates per cell. Show that* $SSAB = L_{ab}^2/(8n)$*, where* L_{ab} *is the contrast corresponding to the experimental condition ab.*

The contrast basis vectors are $e_1 = (1,1,1,1,1,1,1,1)$, $e_a = (-1,1,-1,1,-1,1,-1,1)$, $e_b = (-1,-1,1,1,-1,-1,1,1)$ so their componentwise product is $e_{ab} = (1,-1,-1,1,1,-1,-1,1)$. The cell sums are denoted (1), (a) (b), (ab) and so on. Let the vector of cell sums be $V = ((1),(a),(b),(ab),(c),(ac),(bc),(abc))$. Then the contrast is the inner product

$$L_{ab} = \langle e_{ab}, V\rangle = (1) - (a) - (b) + (ab) + (c) - (ac) - (bc) + (abc).$$

On the other hand, in the model

$$y_{ijk\ell} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijk\ell}$$

the side conditions (see problem 3.) imply

$$(\alpha\beta)_{00} + (\alpha\beta)_{01} = (\alpha\beta)_{01} + (\alpha\beta)_{11} = (\alpha\beta)_{10} + (\alpha\beta)_{11} = 0$$

so

$$(\alpha\beta)_{11} = -(\alpha\beta)_{01} = -(\alpha\beta)_{10} = (\alpha\beta)_{00}.$$

The estimators $\widehat{(\alpha\beta)_{ij}}$ satisfy the same equations. Hence their squares are all equal and

$$SSAB = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}\sum_{\ell=1}^{n}\widehat{(\alpha\beta)}_{ij}^{\,2} = 8n\widehat{(\alpha\beta)_{11}}^{\,2}.$$

But the estimator for the effect is given by

$$
\begin{aligned}
\widehat{(\alpha\beta)_{11}} &= \bar{Y}_{11\cdot\cdot} - \bar{Y}_{1\cdot\cdot\cdot} - \bar{Y}_{\cdot1\cdot\cdot} + \bar{Y}_{\cdot\cdot\cdot\cdot} \\
&= \frac{4Y_{11\cdot\cdot} - 2Y_{1\cdot\cdot\cdot} - 2Y_{\cdot1\cdot\cdot} + Y_{\cdot\cdot\cdot\cdot}}{8n} \\
&= \frac{4Y_{110\cdot} + 4Y_{111\cdot} - 2Y_{100\cdot} - 2Y_{110\cdot} - 2Y_{101\cdot} - 2Y_{111\cdot} - 2Y_{010\cdot} - 2Y_{110\cdot} - 2Y_{011\cdot} - 2Y_{111\cdot}}{8n} \\
&\quad + \frac{Y_{000\cdot} + Y_{100\cdot} + Y_{010\cdot} + Y_{110\cdot} + Y_{001\cdot} + Y_{101\cdot} + Y_{011\cdot} + Y_{111\cdot}}{8n} \\
&= \frac{Y_{000\cdot} - Y_{100\cdot} - Y_{010\cdot} + Y_{110\cdot} + Y_{001\cdot} - Y_{101\cdot} - Y_{011\cdot} + Y_{111\cdot}}{8n} \\
&= \frac{(1) - (a) - (b) + (ab) + (c) - (ac) - (bc) + (abc)}{8n} = \frac{L_{ab}}{8n}.
\end{aligned}
$$

Hence we complete the computation

$$SSAB = 8n\widehat{(\alpha\beta)_{11}}^{\,2} = 8n\left(\frac{L_{ab}}{8n}\right)^2 = \frac{L_{ab}^2}{8n}.$$

(5.) *Consider a 2^3 fixed effects analysis of variance model with n replicates per cell. Show that the sum of squares identity holds $SST = SSE + SSA + SSB + SSAB + SSC + SSAC + SSBC + SSABC$.*

The fact that $SST - SSE = \sum_i \sum_j \sum_k \sum_\ell \bar{Y}_{ijk\cdot}^2 - 8n\bar{Y}_{\cdot\cdot\cdot\cdot}^2$ follows just like in problem 4a. Using the fact that $SSfactor = L_{factor}^2/(8n)$ for all factors, the result follows from a basic fact of linear algebra. Observe that all basis vectors $e_1 = (1,1,1,1,1,1,1,1)$, $e_a = (-1,1,-1,1,-1,1,-1,1)$, $e_b = (-1,-1,1,1,-1,-1,1,1)$, $e_{ab} = (1,-1,-1,1,1,-1,-1,1)$ and so on. All have length $\sqrt{8}$ and are orthogonal. Thus the vectors $e_1/\sqrt{8}, e_a/\sqrt{8}, \ldots$ are orthonormal. Thus the orthogonal decomposition of the vector $V = ((1),(a),(b),(ab),(c),(ac),(bc),(abc))$ into the orthonormal basis is

$$V = c_1\,\frac{e_1}{\sqrt{8}} + c_a\,\frac{e_a}{\sqrt{8}} + \cdots c_{abc}\,\frac{e_{abc}}{\sqrt{8}}$$

where c_{factor} is gotten by orthogonal projection

$$c_{factor} = \langle V, \frac{e_{factor}}{\sqrt{8}}\rangle = \frac{L_{factor}}{\sqrt{8}}.$$

It follows that the square length of the vector $V/\sqrt{n}$ in $\mathbb{R}^8$ is

$$
\begin{aligned}
\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{k=0}^{1}\sum_{\ell=1}^{n} \bar{Y}_{ijk\cdot}^2 = \frac{|V|^2}{n} &= \frac{c_1^2 + c_a^2 + \cdots + c_{abc}^2}{n} \\
&= \frac{L_1^2 + L_a^2 + \cdots L_{abc}^2}{8n} = 8n\bar{Y}_{\cdot\cdot\cdot\cdot}^2 + SSA + \cdots + SSABC,
\end{aligned}
$$

since $SSfactor = L_{factor}^2/(8n)$ and $Y_{\cdot\cdot\cdot\cdot} = L_1$. Subtracting $8n\bar{Y}_{\cdot\cdot\cdot\cdot}^2$ completes the computation.