Homework Assignment 3

Stat 502 WD

Please submit ONE document to the Dropbox for Homework 3. This can be a Word document (preferred) or a pdf file.

\*\*\* **Question 1 and 3 are to be done in SAS**

\*\*\* **Question 2 is to be done in Mintab**.

Your submission can include ‘copy and paste’ output results from the software, but be sure to indicate important results in your text.

**1)** Consider an experiment wherein four methods of tablet compaction are being tested, and the measurement of interest is the elastic modulus (em). The researchers think that the method of compaction will affect the em. Four methods (m1, m2, m3, and m4) were used, and there were 5 replications of the test for each method. The data were as follows:

m1 m2 m3 m4

5.95 10.61 17.00 7.51

4.78 13.13 20.88 6.82

4.28 10.51 20.06 6.23

5.04 9.26 15.25 6.49

4.56 11.04 23.74 5.91

1. (10 pts) State their Null and Alternative Hypothesis.

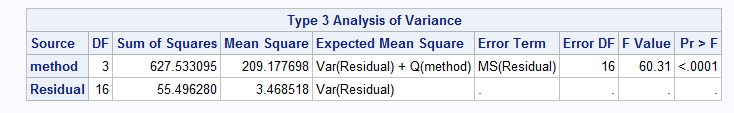
The null hypothesis (the opposite of the alternative) states that there are no differences (or that they are all equal) among the group means.

H0:μ1=μ2= μ3=μ4

Ha: One of the means is different

1. (10 pts) Run an ANOVA in SAS to test the Null Hypotheses, and refer to specific output obtained to draw a conclusion about the Null Hypothesis (note that you will need to re-organize the data into a ‘stacked format’ for SAS input).

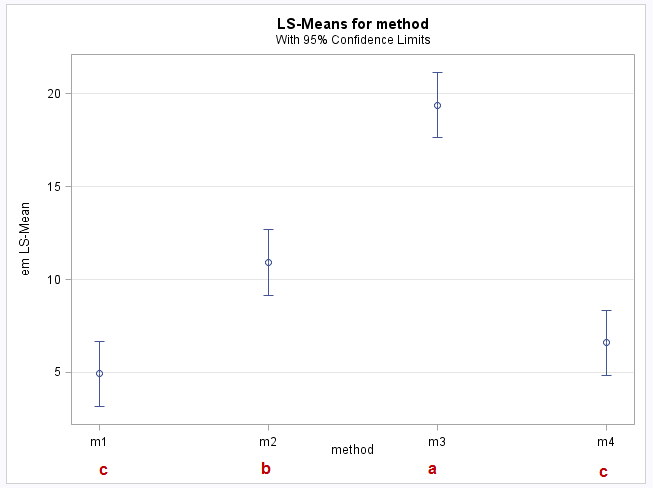
We see from the anova output:

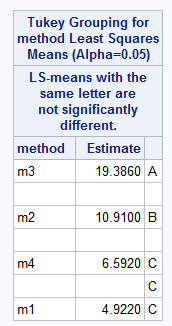


Since the P-value < alpha (assume 0.05) and therefore we reject the null hypothesis. We have sufficient evidence to reject the null hypothesis that the means are the same and conclude the alternative.

1. (10 pts) Perform a mean comparison using the Tukey method.

**SAS Anova output:**





**Manual computation:**

In the Tukey procedure we compute a ‘yardstick’ value based on the MSError and the number of means being compared. If any two means differ by more than the Tukeyw value, then they are significantly different.

Tukey’s w value w = qα(p,dfError)⋅sY¯

We have

qα(p,dfError) = q.05(4, 16) = 4.05

r= number of replications = 5

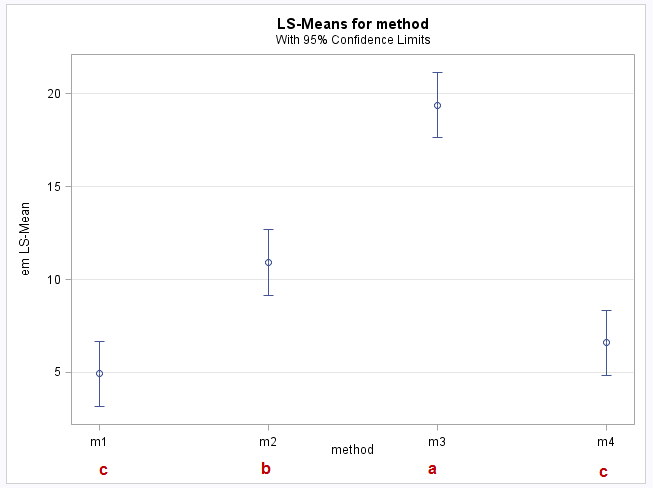
sY¯ =standard error of a treatment mean =  = = 0.833

Therefore w = 4.05 \* 0.833 = 3.374

For the example the tukey analysis gives:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Tukey w | 3.374 |  |  |
| Brand | m3 | m2 | m4 | m1 |
| Means | 19.386 | 10.91 | 6.592 | 4.922 |
|  | a | b | c | c |
|  |  |  |  |  |
|  |  | Is < w | Label |  |
| m3-m2 | 8.476 | No | a, b to m3, m2 |  |
| m2-m4 | 4.318 | No | c to m4 |  |
| m4-m1 | 1.67 | Yes | c to m1 |  |

d) (10pts) Create a means plot in SAS and label the means with the lettering that you obtain from the mean comparisons.



**2)** a) (10 pts) Run and ANOVA for the same data as in Problem I, produce a plots of residuals, use the Box-Cox method in MINITAB to find the ‘best’ λ.





**One-way ANOVA: em versus method**

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

method 3 627.53 209.178 60.31 0.000

Error 16 55.50 3.469

Total 19 683.03

Model Summary

S R-sq R-sq(adj) R-sq(pred)

1.86240 91.87% 90.35% 87.30%

Means

method N Mean StDev 95% CI

m1 5 4.922 0.639 (3.156, 6.688)

m2 5 10.910 1.407 (9.144, 12.676)

m3 5 19.39 3.33 (17.62, 21.15)

m4 5 6.592 0.613 (4.826, 8.358)

Pooled StDev = 1.86240

**Tukey Pairwise Comparisons**

Grouping Information Using the Tukey Method and 95% Confidence

method N Mean Grouping

m3 5 19.39 A

m2 5 10.910 B

m4 5 6.592 C

m1 5 4.922 C

Means that do not share a letter are significantly different.



Best λ is -0.5

b) (10 pts) Create a new response variable using the transformation suggested in part (a), and re-run the ANOVA producing a new residual plot. (A new variable can be created in the Minitab Worksheet by using the ‘Calc’ tab in the toolbar).





**One-way ANOVA: 1/sqrt(em) versus method**

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

method 3 0.143902 0.047967 103.92 0.000

Error 16 0.007385 0.000462

Total 19 0.151287

Model Summary

S R-sq R-sq(adj) R-sq(pred)

0.0214842 95.12% 94.20% 92.37%

**Tukey Pairwise Comparisons**

Grouping Information Using the Tukey Method and 95% Confidence

method N Mean Grouping

m1 5 0.4529 A

m4 5 0.39047 B

m2 5 0.30420 C

m3 5 0.22919 D

Means that do not share a letter are significantly different.

1. (10 pts) Find the Tukey mean comparisons for the transformed variable and compare the results with the original (un-transformed data). How do the residual plots compare with the residual plots obtained before transformation?

**Tukey Pairwise Comparisons**

Grouping Information Using the Tukey Method and 95% Confidence

method N Mean Grouping

m1 5 0.4529 A

m4 5 0.39047 B

m2 5 0.30420 C

m3 5 0.22919 D

Means that do not share a letter are significantly different.

The mean comparisons lead to all methods having significantly different means. While in the case with un-transformed variable m1 and m4 had the same letter (C) – not significantly different means.

**Comparison of residual plots:**

The following points are noteworthy:

* The residual plot for the untransformed variable confirms that the "equal variance" assumption is violated: the residuals vs. fits plot exhibits fanning. However the transformation helped with that and indicates that the equal variance condition is met after the transformation.
* While we need more tests, visually it appears that the normality of the residuals is improved after the transformation.
* Both the variables seemingly meet the linearity condition. One caveat here is that since the transformation is improving the normality and equal variance condition, it will have a positive impact on the linearity.
* Both the conditions appear to indicate independence.
* The R-squared increases from 91.87 to 95.12

In summary the y-transformation has increased the R-squared and improved the situation with normality and equal variance.

d) (10 pts) What power did we have when running this ANOVA?

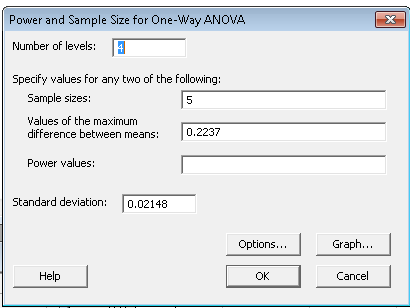
~~When using em: Number of levels: 4, Sample sizes: 5, max difference: (19.39-4.922) = 14.468, SD: sqrt(MSE)=sqrt(3.469)= 1.863~~

~~~~

~~When using 1/sqrt(em): Number of levels: 4, Sample sizes: 5, max difference: 0.22371, SD: sqrt(MSE)=sqrt(0.000462)= 0.0215~~

~~~~

Correct Answer



*This yields*:

**Power and Sample Size**

One-way ANOVA

Alpha = 0.05 Assumed standard deviation = 0.02148

Factors: 1 Number of levels: 4

Maximum Sample

Difference Size Power

0.2237 5 1

Power is 1

**3)** In this set of questions (2pts each) we will be working through some important material that we see in Chapter 16. We will be using SAS because it allows total control of model specification and direct output of the estimates of the model parameters.

Objectives: First, we want to take a look at what is going on with the ‘Cell Means’ model discussed in 16.3 compared to the ‘Effects Model’ discussed in 16.7.

Second, we want to explore the General Linear F test as shown at the very end of 16.6 as a Comment.

Lastly, we will explore the relationship of ANOVA to Regression, and compare two types of coding: Indicator coding and Effects Coding.

Instructions: To complete the homework assignment, cut and paste the SAS code provided into the SAS Program Editor window for each objective, and then use the output to compute the requested quantities. You should include the Output tables from SAS in the document you submit.

Experimental Setting: Conductivity (the response variable) of electrical components are measured after application of different coating materials. (Data from Hicks, 1973) The interest is focused on evaluating differences among 4 types of coatings, labelled C1, C2, C3, and C4. The different coatings (the treatment levels) were assigned to components in a completely randomized design, with 5 replications of each coating.

Steps

Start a SAS program, which includes columns for the indicator (or ‘dummy’) coded variables (ic1 ic2 ic3) and the effect coded variables (ec1 ec2 ec3). The SAS code fragments below can be copied and pasted directly to your SAS program editor and run.

data hicks;

input trt $ resp ic1 ic2 ic3 ec1 ec2 ec3;

/\* ic=indicator coding ec=effect coding \*/

datalines;

c1 56 1 0 0 1 0 0

c1 55 1 0 0 1 0 0

c1 62 1 0 0 1 0 0

c1 59 1 0 0 1 0 0

c1 60 1 0 0 1 0 0

c2 64 0 1 0 0 1 0

c2 61 0 1 0 0 1 0

c2 50 0 1 0 0 1 0

c2 55 0 1 0 0 1 0

c2 56 0 1 0 0 1 0

c3 45 0 0 1 0 0 1

c3 46 0 0 1 0 0 1

c3 45 0 0 1 0 0 1

c3 39 0 0 1 0 0 1

c3 43 0 0 1 0 0 1

c4 42 0 0 0 -1 -1 -1

c4 39 0 0 0 -1 -1 -1

c4 45 0 0 0 -1 -1 -1

c4 43 0 0 0 -1 -1 -1

c4 41 0 0 0 -1 -1 -1

;

run;

The means are seen in the results of the Summary procedure:

Run 1:

proc summary data=hicks;

class trt; var resp; output out=a mean=; run;

proc print; title 'Summary Output'; run;

In the output, the \_Type\_0 mean is the overall or grand mean, and the ‘cell’ means (the treatment means) are listed as \_Type\_1.



Run 2: Fitting the Full means model.

proc mixed data=hicks method=type3;

class trt;

model resp = trt / noint solution;

ods select Type3 SolutionF;

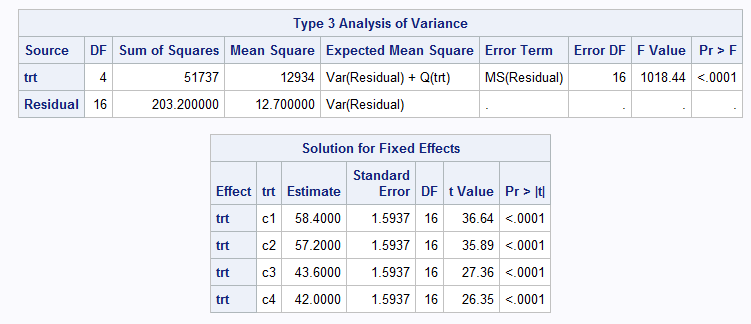
title 'Full (Means) Model';

run;

Notice that the model statement has the option “ / noint solution”. The model statement automatically creates an intercept, and here we don’t want this to happen. We simply want to estimate parameters for the 4 treatment level means. The parameter estimates are shown in the output by specifying the option “solution”.

The “ods select …” statement is restricting the output to only include the tables of interest to be output.

The output should look like:

****

1. How do the Estimates in the Solutions table compare to the \_Type1\_ means in the Summary procedure output?

\_\_They are same\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. What is the SSE (or SSResidual) for this ‘**Full**’ means model?

203.2\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Run 3: Now we can fit a ‘**Reduced**’ model, in which we fit only a single, overall mean.

proc mixed data=hicks method=type3;

class;

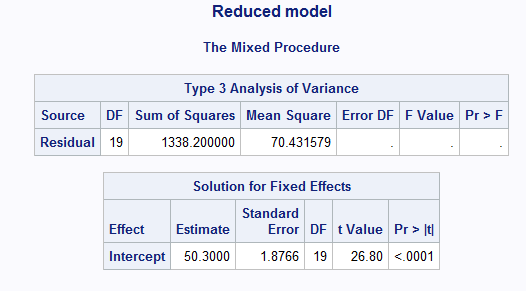
model resp = / solution;

ods select Type3 SolutionF;

title 'Reduced model';

run;

Here you can see that we didn’t specify anything in the class statement and that there no variables specified on the right hand side of the equation in the model statement. Because the model statement automatically generates an intercept, this code will produce the estimate of the overall mean and it will be labelled ‘intercept’.



1. Find the F statistic for testing for the Treatment effect by constructing a General Linear F test. Use the SSE(Full) and SSE(Reduced) from the output above.

F= (1338.2-203.2)/(3) / 203.2/16 = 378.33/12.7 = 29.79

The General Linear F test is computed as:

(see textbook section 16.6, under *Comments*).

Run 4: Now run the ‘regular’ ANOVA, specifying the model as we have done before:

proc mixed data=hicks method=type3;

class trt;

model resp=trt / solution;

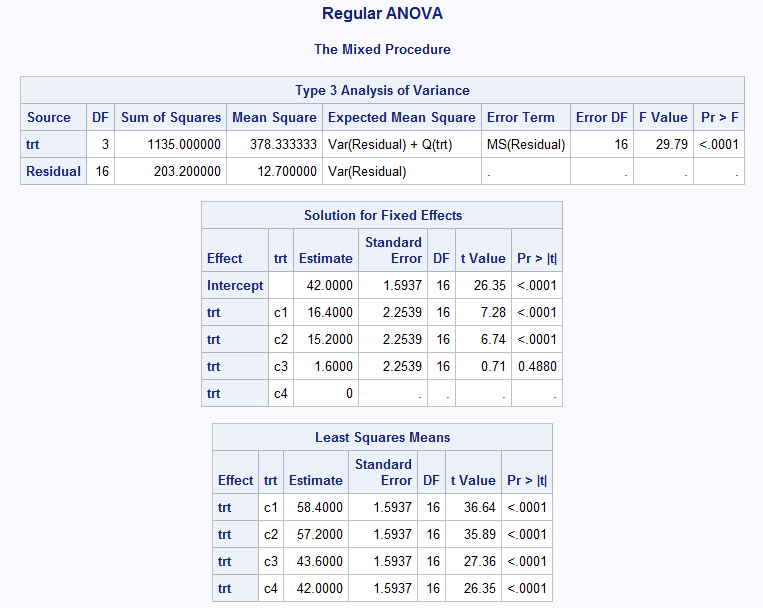
lsmeans trt;

ods select Type3 SolutionF LSmeans;

title 'Regular ANOVA';

run;

Copy the output to your submission.



**d)** From the ANOVA you produced in Run 4 you can see that the difference in residual sums of squares (SSE Reduced – SSE Full) calculated above in question (3) is the same as ~~\_\_Mean Square of Treatment – 378.33\_\_\_\_\_\_\_\_\_~~**~~.~~  SSTrt**

In addition to the ANOVA table we get a listing of estimates for the ANOVA model:

| **Solution for Fixed Effects** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Effect** | **trt** | **Estimate** | **Standard Error** | **DF** | **t Value** | **Pr > |t|** |
| **Intercept** |  | 42.0000 | 1.5937 | 16 | 26.35 | <.0001 |
| **trt** | **c1** | 16.4000 | 2.2539 | 16 | 7.28 | <.0001 |
| **trt** | **c2** | 15.2000 | 2.2539 | 16 | 6.74 | <.0001 |
| **trt** | **c3** | 1.6000 | 2.2539 | 16 | 0.71 | 0.4880 |
| **trt** | **c4** | 0 | . | . | . | . |

The form of the output is determined by the type of coding that is being used by SAS to find the estimates. With Indicator coding, the estimate shown for ‘Intercept’ is in fact the mean for the C4 treatment level. The estimates for the means for the remaining treatment levels are obtained by adding the value shown for a treatment level to the ‘Intercept’.

**e)** Show below a confirmation of this

C1 = \_\_\_58.4\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

C2 = \_\_\_57.2\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

C3 = \_\_\_43.6\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

C4 = \_\_\_42\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Run 5: We can confirm that SAS is using indicator coding for ANOVA by running the ANOVA ourselves as a Multiple Linear Regression. In SAS we can conveniently use Proc Reg to specify the dependent variable as a function of the three indicator variables:

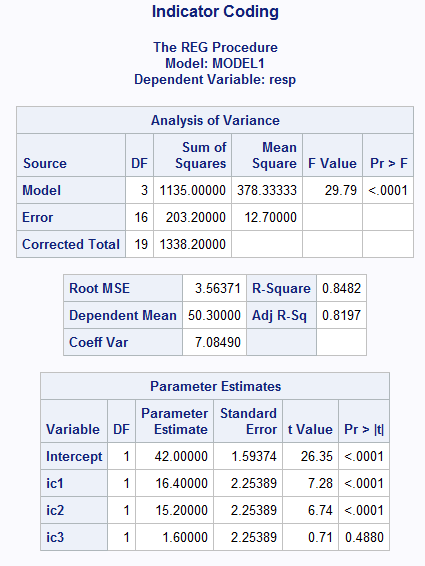
proc reg data=hicks;

model resp=ic1 ic2 ic3;

ods select ANOVA FitStatistics ParameterEstimates;

title 'Indicator Coding';

run;



**f)** How do the Model SS, ErrorSS, and F statistic from the Regression output compare to the ANOVA SSTrt, SSResidual, and F results we obtained above using the ‘regular ANOVA' in Proc Mixed?

\_\_We get the same values for Model SS, Error SS and F stat\_\_\_\_\_\_\_\_

**g)** How do the parameter estimates in the regression output compare to the solutions obtained from the ‘regular ANOVA’?

\_\_Same\_\_\_\_\_\_\_

Run 6: Finally, we can use the regression procedure to run the model with the effect coded variables.

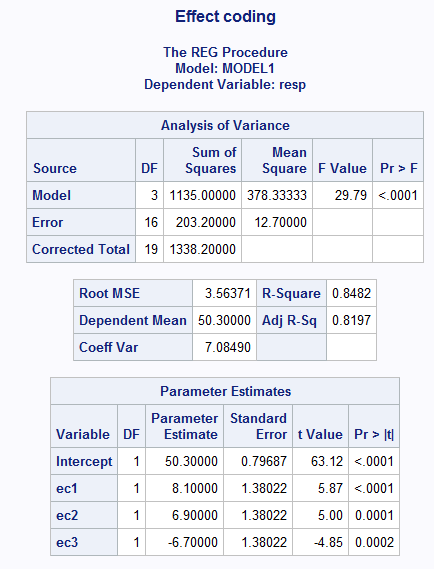
proc reg data=hicks;

model resp=ec1 ec2 ec3;

ods select ANOVA FitStatistics ParameterEstimates;

title 'Effect coding';

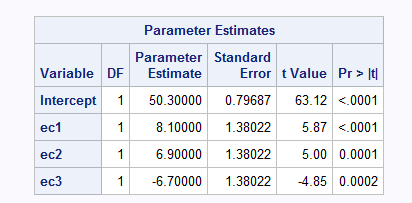
run;



**h)** How do the Model SS, ErrorSS, and F compare with the regression results using the indicator coding?

\_\_\_\_\_Same\_\_\_\_\_\_\_\_

| **Parameter Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | **1** |  |  |  |  |
| **ec1** | **1** |  |  |  |  |
| **ec2** | **1** |  |  |  |  |
| **ec3** | **1** |  |  |  |  |
|  |  |  |  |  |  |

****

**i)** The output will contain the list of estimates of the parameters in the ‘Effects’ Model (Section 16.7). Compare the intercept here with the results of the Summary procedure in Step 1. \_Same as grand mean\_\_Type\_0\_\_\_\_\_\_\_\_\_\_\_\_

The estimate shown in the table above for ec1 would be added to the intercept to represent this treatment level in the effects model. In other words, the b1 estimate shown above for in the regression model is the value of for treatment level C1 in the Effects Model representation of the ANOVA (Equation 16.62 in the text). The estimates for C2 and C3 are also shown, but to get the estimate for C4 we have to use the relationship .

**j)** Show these estimates of below:

~~C1 \_\_50.3+8.1=58.4\_\_\_\_\_\_\_\_~~

~~C2 \_\_\_50.3+6.9=57.2\_\_\_\_\_\_\_~~

~~C3 \_\_\_50.3-6.7=43.6\_\_\_\_\_\_\_~~

~~C4 \_\_\_50.3-8.1-6.9+6.7= 42\_\_\_\_\_\_\_~~

C1 \_\_\_\_\_8.1\_\_\_\_\_

C2 \_\_\_\_\_6.9\_\_\_\_\_

C3 \_\_\_\_\_-6.7\_\_\_\_\_

C4 \_\_\_\_\_-8.3\_\_\_\_\_