Types of Sums of Squares

TESTING DIFFERENT HYPOTHESES

Types of Sums of Squares

There are four different types of sums of squares

The two most important (that is, most common) are Type I and Type III

In many cases, the types will be equivalent and hence the distinction isn't important

Goal: Explore when the types are the same, when they are different, and which one to use if they are different

Extra Sums of Squares Tests

There are many different hypotheses that we can test

Different types of (differences between) sums of squares have been developed to test different types of hypotheses

These are known as:

- TYPE I SUMS OF SQUARES: (Known as sequential sum of squares)
- TYPE II SUMS OF SQUARES:
- TYPE III SUMS OF SQUARES: (Known as adjusted/partial sums of squares)
- TYPE IV SUMS OF SQUARES: (We won't go into this one)

(These are not to be confused with Type I Error nor Type II Error)

We will focus in this class on Type I and Type III, but we will briefly discuss Type II in these notes as well.

Type III and IV are the same unless there are terms that have no observations

(e.g. $\mu\{Y|A,B\} = \mu + A + B + AB$ w/ $J = K = 2$ and treatment combination A = 2 and B = 2 do not occur together)

General Notation

We can test different hypotheses by comparing the sums of squares explained by the different models

To facilitate these tests , we need to define appropriate notation

Suppose we are considering two models M_1 and M_2

Example:

M_1: $\mu\{Y|X\} = \beta_0 + \beta_1 x_1$

M_2: $\mu\{Y|X\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

We want to test between these two models…

General Notation

Example (continued):

M_1: $\mu\{Y|X\} = \beta_0 + \beta_1 x_1$

M_2: $\mu\{Y|X\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

We want to test between these two models:

One possible test: Look at t-test for β_1

Another possible test: Look at extra-sums-of squares of M_2 vs. M_1

(In this case, these are the same tests)

For a general model M, write SS(M) to be the sums of squares explained by M

(sometimes called the regression or model SS)

Definition: $\text{SS}(M_2|M_1) = \text{SS}(M_1, M_2) - \text{SS}(M_1)$

Using this Notation

Back to example (continued):

M_1: $\mu\{Y|X\} = \beta_0 + \beta_1 x_1$

M_2: $\mu\{Y|X\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

We want to test between these two models:

The extra-sums-of squares of M_2 vs. M_1 is $\text{SS}(M_2|M_1) = \text{SS}(M_1, M_2) - \text{SS}(M_1)$,

Where the notation $\text{SS}(M_1, M_2)$ means the model that combines M_1 & M_2

Back to example (conclusion):

So, the extra-sums-of-squares F-test compares these two things (via division):

- $SS(M_2|M_1)$ divided by the additional degrees of freedom (DF) of M_2 vs. M_1
- An estimate of the variance, commonly $\text{SS}(M_2)$ divided by the (*n*-DF) of M_2

Type I Sums of Squares

Type I Sums of Squares

When fitting an ANOVA or multiple regression model, the results we have discussed thus far would be the same no matter the order the terms are written:

Example:

$$\mu\{income|education, gender\} = \beta_0 + \beta_1 education + \beta_2 gender$$

$$\mu\{income|education, gender\} = \beta_0 + \beta_1 gender + \beta_2 education$$

However, for Type I sums of squares, the written order matters

Type I Sums of Squares (SS1)

The Type I Sums of Squares (SS1) looks at the incremental improvement via SS for each term in the model, reading left to right

Let's get the SS1 from an old example:

$$\mu\{Y|X\} = sqFootage + zipcode + nBedrooms + nBathrooms$$

(here, I'm using reductive notation by suppressing the "betas")

Define the following sequence of models:

1. $\mu\{Y|X\} = \beta_0$
2. $\mu\{Y|X\} = sqFootage$
3. $\mu\{Y|X\} = sqFootage + zipcode$
4. $\mu\{Y|X\} = sqFootage + zipcode + nBedrooms$
5. $\mu\{Y|X\} = sqFootage + zipcode + nBedrooms + nBathrooms$

Type I Sums of Squares (SS1): Sequential

1. $\mu\{Y|X\} = \beta_0$
2. $\mu\{Y|X\} = sqFootage$
3. $\mu\{Y|X\} = sqFootage + zipcode$
4. $\mu\{Y|X\} = sqFootage + zipcode + nBedrooms$
5. $\mu\{Y|X\} = sqFootage + zipcode + nBedrooms + nBathrooms$

Then SS1 tests the following sequence of extra sums of squares:

SS(2.|1.)

SS(3.|2.)

SS(4.|3.)

SS(5.|4.)

Type I Sums of Squares (SS1): Sequential

```
TITLE 'SS1: SqFt first';
PROC GLM DATA = housing;
    CLASS zipCode (ref = '75224');
    MODEL salePrice = sqFootage zipCode nBedrooms nBathrooms / SS1;
RUN;
```

Source	DF	Type I SS	Mean Square	F Value	Pr > F
sqFootage	1	3.7511734E13	3.7511734E13	334.57	<.0001
zipcode	1	297924155925	297924155925	2.66	0.1161
nBedrooms	1	163226851347	163226851347	1.46	0.2394
nBathrooms	1	1.9143457E12	1.9143457E12	17.07	0.0004

```
TITLE 'SS1: Baths first';
PROC GLM DATA = housing;
    CLASS zipCode (ref = '75224');
    MODEL salePrice = nBathrooms sqFootage zipCode nBedrooms / SS1;
RUN;
```

Source	DF	Type I SS	Mean Square	F Value	Pr > F
nBathrooms	1	3.9210804E13	3.9210804E13	349.73	<.0001
sqFootage	1	411977926265	411977926265	3.67	0.0672
zipcode	1	19545685233	19545685233	0.17	0.6800
nBedrooms	1	244903183640	244903183640	2.18	0.1524

```
TITLE 'SS1: Zip first';
PROC GLM DATA = housing;
    CLASS zipCode (ref = '75224');
    MODEL salePrice = zipCode nBathrooms sqFootage nBedrooms / SS1;
RUN;
```

Source	DF	Type I SS	Mean Square	F Value	Pr > F
zipcode	1	2.6982E13	2.6982E13	240.66	<.0001
nBathrooms	1	1.2288638E13	1.2288638E13	109.60	<.0001
sqFootage	1	371689482968	371689482968	3.32	0.0811
nBedrooms	1	244903183640	244903183640	2.18	0.1524

(these are with respect to the order in each MODEL statement)

SS(2.|1.)
SS(3.|2.)
SS(4.|3.)
SS(5.|4.)

SS(2.|1.)
SS(3.|2.)
SS(4.|3.)
SS(5.|4.)

SS(2.|1.)
SS(3.|2.)
SS(4.|3.)
SS(5.|4.)

Type I Sums of Squares: Sequential

Reasons for using SS1:

Primarily, if we have an interest in a specific nesting of explanatory variables

Example: Suppose we are looking at the Zillow data and we want to make sure all comparisons control for location via including zipcode

Also, we have the following research questions:

Is there an estimated effect of square footage given zipcode is fixed?

Is there an estimated effect of bedrooms given zipcode and square footage are fixed?

Is there an estimated effect of bathrooms given zipcode, square footage, and bedrooms are fixed?

Then SS1, alternatively known as sequential sums of squares, is the appropriate analysis

Type I Sums of Squares: Sequential

The most commonly used job for SS1 is with polynomial models

$$\mu\{Y|X\} = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_2 X^3 + \cdots + \beta_p X^{p-1}$$

In this case, the usual imposed model restriction is to include all lower level polynomial transformations

(just like with interactions and main effects)

Hence, we want to know:

- Do we need the linear effect vs. the intercept?
- Do we need the quadratic effect vs the linear effect and the intercept?
- …
- Do we need the X^{p-1} effect vs X^{p-2}, … , X effects and the intercept?

This is exactly what SS1 tests

Type I Sums of Squares: Sequential

Properties of SS1:

(These are very related concepts)

- Under the independent, $N(0, \sigma^2)$ assumption, SS1 tests are independent of each other

(general idea, orthogonal Gaussians are independent)

- SS1 depends on the order that the effects are specified in the MODEL statement
- SS1 for all effects add up to the "Model" SS.

(None of the other SS types have this property, except in special cases)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	3.988723E13	9.9718076E12	88.94	<.0001
Error	24	2.6908501E12	112118754091		
Corrected Total	28	4.257808E13			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
sqFootage	1	3.7511734E13	3.7511734E13	334.57	<.0001
zipcode	1	297924155925	297924155925	2.66	0.1161
nBedrooms	1	163226851347	163226851347	1.46	0.2394
nBathrooms	1	1.9143457E12	1.9143457E12	17.07	0.0004

```
> 3.7511734e13+297924155925+163226851347+1.9143457e12
[1] 3.988723e+13
```

Type II Sums of Squares

Type II Sums of Squares (SS2)

A model effect j is **CONTAINED** in model effect k if effect j can be formed by removing other explanatory variables from effect k

Example: the main effect x_1 is contained in the interaction x_1x_2 but the interaction x_1x_3 is not contained in x_1x_2

Definition: The Type II Sums of Squares (SS2) examines the reduction in sums of squares of adding an effect *x* to a model after all the other effects except for all the effects that contain *x* are added to the model

Example: If we are looking at a two-way crossed, interactive ANOVA scenario with factors A and B: $\mu\{Y|A,B\} = \mu + A + B + AB$

Test for A: SS(A$|\mu, B$)

Test for B: SS(B$|\mu, A$)

Test for AB: SS(AB$|\mu, A, B$)

Type III Sums of Squares

Type III Sums of Squares (SS3)

The type III Sums of Squares are the most familiar and commonly used

The test is similar to SS2, but eliminates the “containment” condition

Definition: The Type II Sums of Squares (SS2) examines the reduction in sums of squares of adding an effect *x* to a model after all the other effects are added to the model

Example: If we are looking at a two-way crossed, interactive ANOVA scenario with factors A and B: $\mu\{Y|A,B\} = \mu + A + B + AB$

Test for A: SS(A$|\mu, B, AB$)

Test for B: SS(B$|\mu, A, AB$)

Test for AB: SS(AB$|\mu, A, B$)

Type III Sums of Squares (SS3)

It is fundamentally a test of an effect vs. all other effects in the model

Compare this to the regression interpretation:

"We estimate the an association between Y and *x, holding all other terms constant*"

This is the idea behind writing SS($A|\mu, B, AB$), which in words would be sums of squares of the model with A, B, AB vs. the model with only B and AB (and an intercept term)

Type III Sums of Squares (SS3)

```
TITLE 'SS2';
PROC GLM DATA = housing;
    CLASS zipCode (ref = '75224');
    MODEL salePrice = zipCode nBathrooms sqFootage nBedrooms sqFootage*zipCode / SS2;
RUN;
```

Source	DF	Type II SS	Mean Square	F Value	Pr > F
zipcode	1	390187942918	390187942918	4.19	0.0522
nBathrooms	1	1.1488842E12	1.1488842E12	12.35	0.0019
sqFootage	1	611415468145	611415468145	6.57	0.0174
nBedrooms	1	5376784691.4	5376784691.4	0.06	0.8122
sqFootage*zipcode	1	550906165856	550906165856	5.92	0.0231

This tests the ESS for model

- zipcode, nBaths, nBeds, sqFt

vs.

- nBaths, nBeds, sqFt

```
TITLE 'SS3';
PROC GLM DATA = housing;
    CLASS zipCode (ref = '75224');
    MODEL salePrice = zipCode nBathrooms sqFootage nBedrooms sqFootage*zipCode / SOLUTION SS3;
RUN;
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
zipcode	1	390187942918	390187942918	4.19	0.0522
nBathrooms	1	1.1488842E12	1.1488842E12	12.35	0.0019
sqFootage	1	105440671390	105440671390	1.13	0.2981
nBedrooms	1	5376784691.4	5376784691.4	0.06	0.8122
sqFootage*zipcode	1	550906165856	550906165856	5.92	0.0231

This tests the ESS for model

- zipcode, nBaths, nBeds, sqFt, zipcode*sqFt

vs.

- nBaths, nBeds, sqFt, zipcode*sqFt

Type III Sums of Squares (SS3)

```
TITLE 'SS3';
PROC GLM DATA = housing;
    CLASS zipCode (ref = '75224');
    MODEL salePrice = zipCode nBathrooms sqFootage nBedrooms sqFootage*zipCode / SOLUTION SS3;
RUN;
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
zipcode	1	390187942918	390187942918	4.19	0.0522
nBathrooms	1	1.1488842E12	1.1488842E12	12.35	0.0019
sqFootage	1	105440671390	105440671390	1.13	0.2981
nBedrooms	1	5376784691.4	5376784691.4	0.06	0.8122
sqFootage*zipcode	1	550906165856	550906165856	5.92	0.0231

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-242417.9238	B	296441.9769	-0.82	0.4219
zipcode 75225	-674673.2583	B	329453.2933	-2.05	0.0522
zipcode 75224	0.0000	B	.	.	.
nBathrooms	311379.2010		88611.2275	3.51	0.0019
sqFootage	-47.0805	B	157.8412	-0.30	0.7682
nBedrooms	-30730.1490		127832.3057	-0.24	0.8122
sqFootage*zipcode 75225	334.1428	B	137.3190	2.43	0.0231
sqFootage*zipcode 75224	0.0000	B	.	.	.

When the factor explanatory variables have two levels, SS3 will exactly mimic the regression table

However, when the factors have more than 2 levels, the SS3 will test all of the level combinations at the same time, while the regression table will test them all individually

Let's look at an example..
(Note: We will just look at SS3 but all 4 SS types test factor levels as a group)

Type III Sums of Squares (SS3): More than 2 Factor Levels

```
PROC GLM DATA = housing;
    CLASS zipCode (ref = '75224');
    MODEL salePrice = zipCode nBathrooms sqFootage nBedrooms sqFootage*zipCode / SOLUTION SS3;
RUN;
```

Class Level Information		
Class	Levels	Values
zipcode	4	75222 75223 75225 75224

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-1015095.904	B	379904.3335	-2.67	0.0151
zipcode 75222	980271.176	B	383337.9106	2.56	0.0193
zipcode 75223	353497.717	B	584597.1377	0.60	0.5525
zipcode 75225	538045.201	B	440382.8041	1.22	0.2367
zipcode 75224	0.000	B	.	.	.
nBathrooms	365206.725		105554.2721	3.46	0.0026
sqFootage	284.295	B	112.2856	2.53	0.0203
nBedrooms	-75180.864		140792.2770	-0.53	0.5995
sqFootage*zipcode 75222	-381.804	B	153.7336	-2.48	0.0225
sqFootage*zipcode 75223	-52.996	B	350.5349	-0.15	0.8814
sqFootage*zipcode 75225	-98.265	B	84.9012	-1.16	0.2614
sqFootage*zipcode 75224	0.000	B	.	.	.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
zipcode	3	669689023572	223229674524	2.23	0.1179
nBathrooms	1	1.1984609E12	1.1984609E12	11.97	0.0026
sqFootage	1	117198304237	117198304237	1.17	0.2928
nBedrooms	1	28546634041	28546634041	0.29	0.5995
sqFootage*zipcode	3	713133807441	237711269147	2.37	0.1022

ANOVA Table: Testing levels as group

Regression Table: Testing levels individually

Comparing SS2 to SS3

Type II Sums of Squares (SS2)

A claim about SS2:

If all of the interaction terms are insignificant, then SS2 is a more powerful test than SS3 (which we will discuss in a moment)

Back to an Example: If we are looking at a two-way crossed, interactive ANOVA scenario with factors A and B: $\mu\{Y|A,B\} = \mu + A + B + AB$

Suppose we test this model using SS2 and SS3

Also, suppose that the test for AB (which is the same test in both SS2 and SS3) fails to reject the null hypothesis

Now, we interpret the tests for the main effects without refitting the model after removing the interaction

Then, SS2 is a more powerful test than SS3

However, it is much more common practice to use SS3 and refit w/o interaction