EPHE 591

Analysis of Variance

Field, Chapter 10

Analysis of Variance

Basic Logic

The logic is simple, we will conduct a test to disprove the null hypothesis:

$$\mu_1 = \mu_2 = \mu_3$$

We will test the null hypothesis by analyzing the VARIANCES within samples and between the means of the samples.

Omnibus Tests versus Planned Comparisons

Unequally Sized Groups

The variance estimates need complex adjustments to weight information from different groups.

Use a computer to do this!

BUT… using unequal groups makes the analysis of variance much more sensitive to violations of homogeneity of variance

Unequally Sized Groups

So, what is to be done…

1) Make group sizes equal

2) Interpolate “missing” data… group means, bootstrapping, etc

Estimating Population Variance From Variation Within Each Sample

- we do not know the true population variance

- however, we can estimate the population variance from the sample

- note, this estimate comes out the same whether or not the null hypothesis is true, because it is based entirely on variance within each sample.

Estimating Population Variance From Variation Within Each Sample

$$S^2{}_{within} = \frac{S_1^2 + S_2^2 + S_3^2}{N_{groups}}$$

- this is also called $\mathrm{MS}_{\mathrm{within}}$, or mean squares within, or $\mathrm{MS}_{\mathrm{error}}$

Why is this “error”

Simply put, because this reflects variance within a group and we are interested in differences between groups…

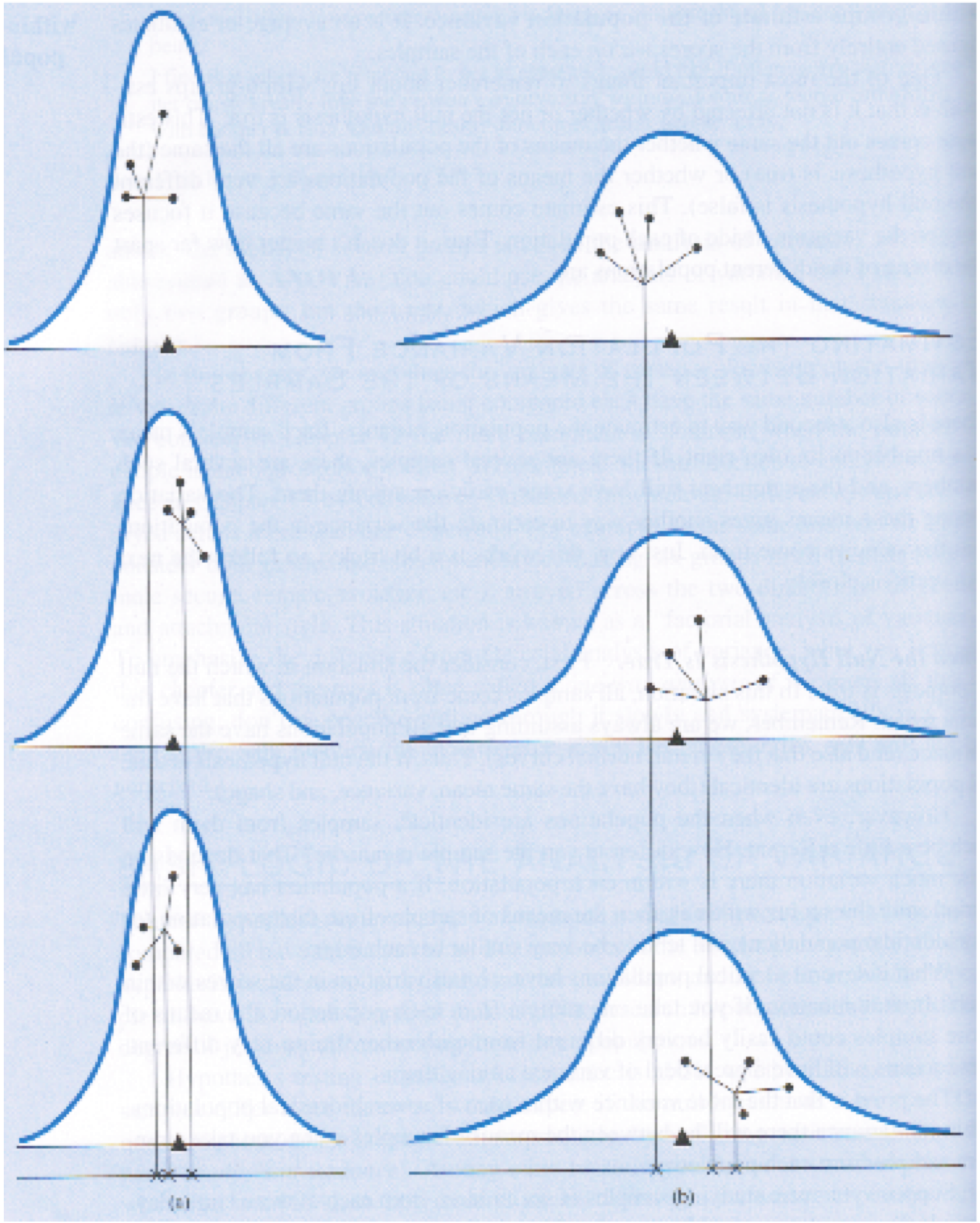
In an ideal world…

Group 1	Group 2	Group 3
8	4	4
8	4	4
8	4	4
8	4	4

Estimating Population Variance From Variation Within Each Sample

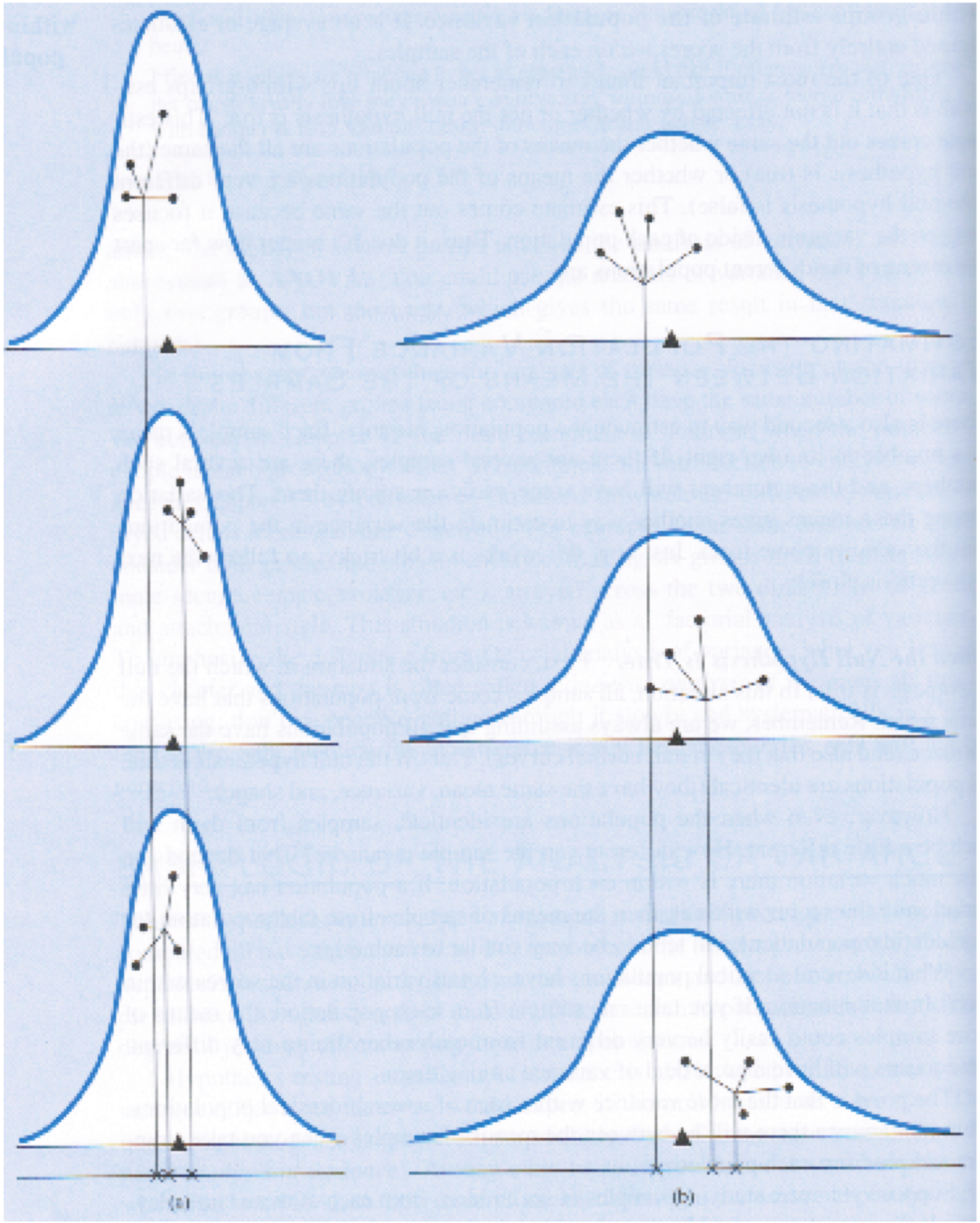
So, MS_{within} or MS_{error} reflects the variability among scores in a sample

Estimating Population Variance From Variation Between the Means of Each Sample



If a population has more variance, then the means of samples taken from that population will be more variable.

This allows us to estimate the variance within each population by examining the variation among the means of our samples.



This would be a <u>between-groups estimate of the population variance</u>.

Estimating Population Variance From Variation Between the Means of Each Sample

if the null hypothesis is true:

between estimate = within estimate

if the null hypothesis is not true:

between estimate > within estimate

Sources of Variation in Within-Group and Between-Group Variance Estimates

	Variation Within Populations	Variation Between Populations
Null Hypothesis is True		
Within-Groups Estimate	X	
Between-Groups Estimate	X	
Null Hypothesis is False		
Within-Groups Estimate	X	
Between-Groups Estimate	X	X

Estimating Population Variance From Variation Between the Means of Each Sample

$$S_M^2 = \frac{\sum (M - GM)^2}{df_{between}}$$

The Estimated Variance of the Distribution of Means

Estimating Population Variance From Variation Between the Means of Each Sample

$$S_{between}^{2} = (S_{M}^{2})(n)$$

Where n is the number of scores in each group

Note, this is referred to as $MS_{between}$

Finally, because n is in the equation, there is an assumption here that group sizes are equal.

Estimating Population Variance From Variation Between the Means of Each Sample

So, $MS_{between}$ reflects the variability among group means

Recall our rules for hypothesis testing…

1) <u>Check assumptions</u>

2) Calculate a statistic

3) Compare that statistic to the sampling distribution of the means for that statistic

ANOVA Assumptions

Continuous Data

No Outliers

Normality

(recall this goes away with a decent sample size)

Homogeneity of Variance

(rule of thumb, Levene, Bartlett)

Recall our rules for hypothesis testing…

1) Check assumptions

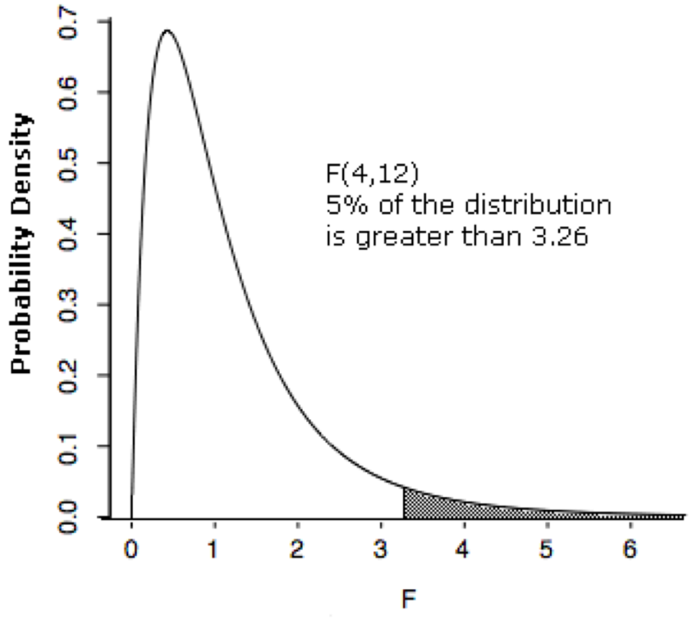
2) Calculate a statistic

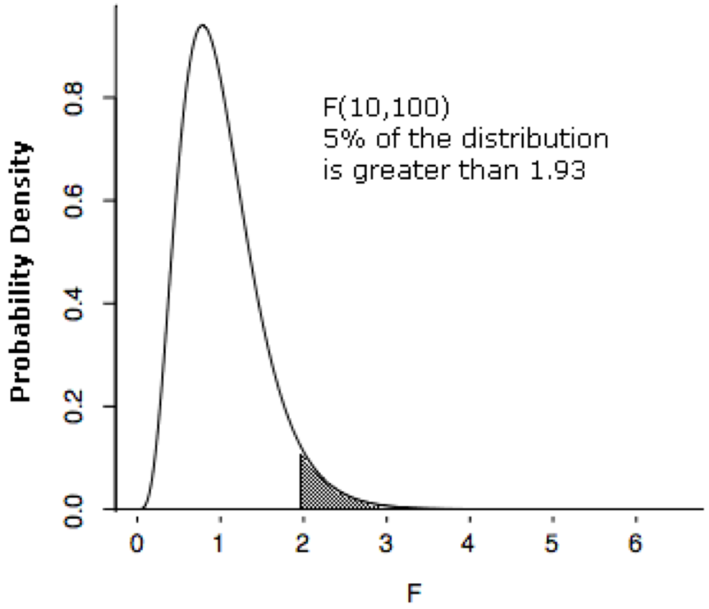
3) Compare that statistic to the sampling distribution of the means for that statistic

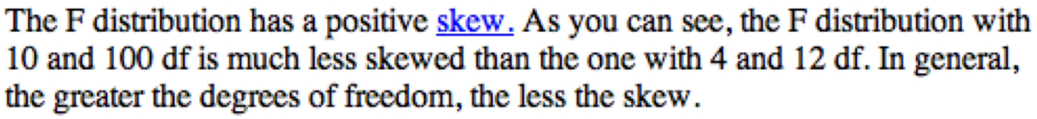
The F Statistic

$$F = \frac{MS_{between}}{MS_{within}}$$

The F Distribution







The F distribution has a positive skew. As you can see, the F distribution with 10 and 100 df is much less skewed than the one with 4 and 12 df. In general, the greater the degrees of freedom, the less the skew.

Recall our rules for hypothesis testing…

1) Check assumptions

2) Calculate a statistic

3) Compare that statistic to the sampling distribution of the means for that statistic

The ANOVA Summary Table

Source	df	SS	MS	F
Between	#	#	#	#
Within	#	#	#	
Total	#	#		

$df_{between}$ = a – 1, where a is the total number of groups

df_{within} = N – a, where N is the total number of scores and a is the total number of groups

df_{total} = N - 1

Score Deviations

$$\begin{tikzcd}
& X \arrow[dl, leftrightarrow] \arrow[dr, leftrightarrow] & \\
\overline{X} \arrow[rr, leftrightarrow] & & \overline{X}_{GM}
\end{tikzcd}$$

Using Sums of Squares to obtain MS$_{within}$ and MS$_{between}$

$$SS_{total} = \sum (X - \overline{X_{GM}})^2$$

Using Sums of Squares to obtain MS_{within} and $MS_{between}$

$$SS_{within} = \sum (X - \overline{X})^2$$

Using Sums of Squares to obtain MS_{within} and $MS_{between}$

$$SS_{between} = n\sum(\overline{X} - \overline{X_{GM}})^2$$

$$SS_{\text{total}} = SS_{\text{between}} + SS_{\text{within}}$$

Effect Size

$$R^2 = \frac{SS_{between}}{SS_{total}}$$

This reflects the proportion of variance accounted for by the measure.

Also called eta squared (η^2), or the correlation ratio.

Effect Size: Eta Squared

$$R^2 = \frac{SS_{between}}{SS_{total}}$$

An effect size of 0.01 is small, 0.06 is medium, 0.14 is large (Cohen, 1988)

Effect Size: Partial Eta Squared

$$R^2 = \frac{SS_{between}}{SS_{between} + SS_{error}}$$

More typically reported than eta squared (because it is the SPSS default???)

A Note…

If you are trying to compare SPSS and R results they may differ.

SPSS uses Type III Sums of Squares by default, R uses Type I (which can be changed).

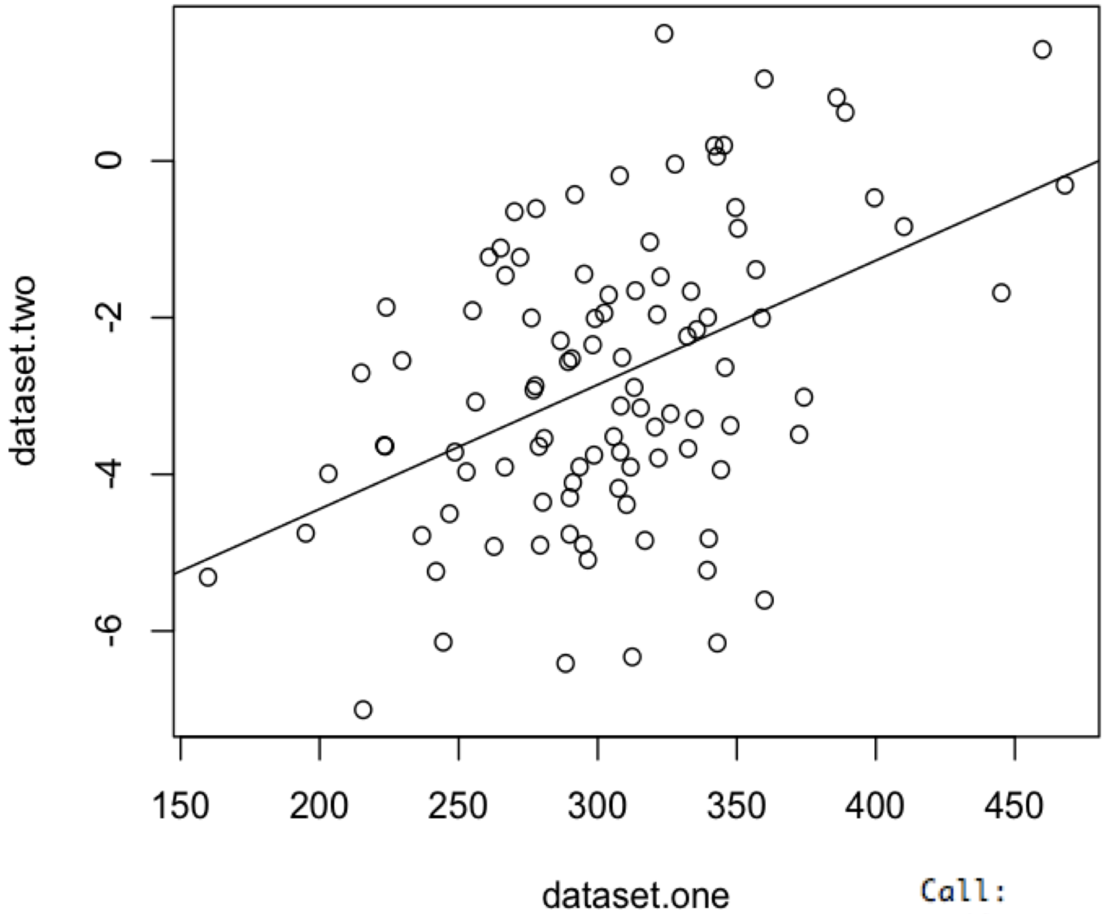
So, the results should be very similar – but a bit different. If they differ greatly, you have made a mistake!

Theory

Recall the multiple regression linear model:

$$Y = X_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + \ldots$$

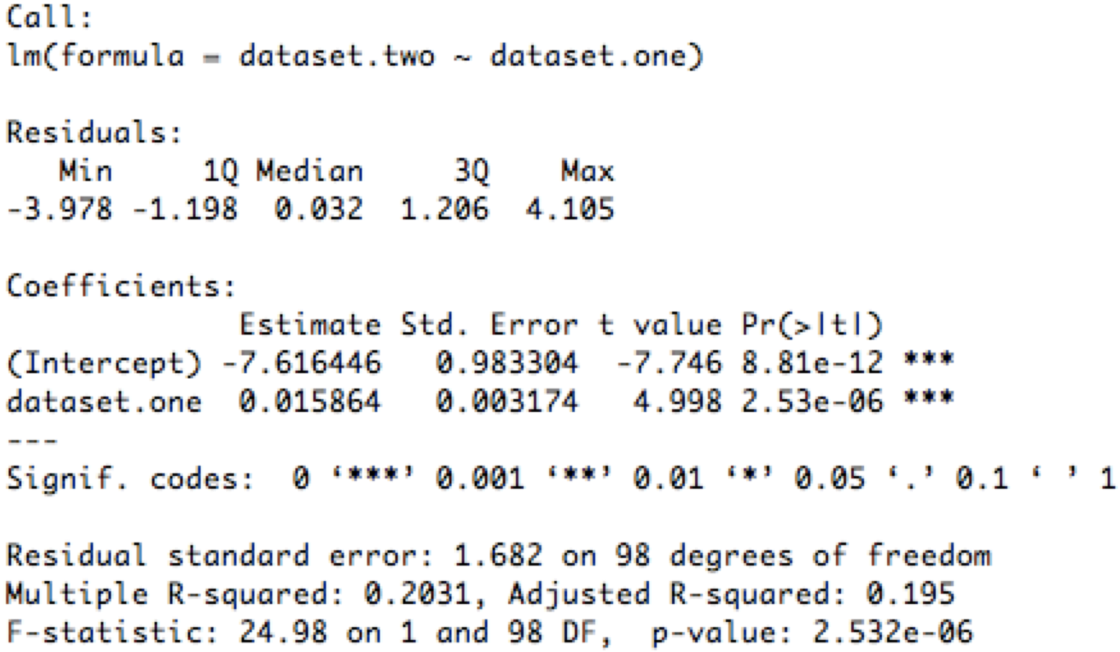
ANOVA is just a special case of this model



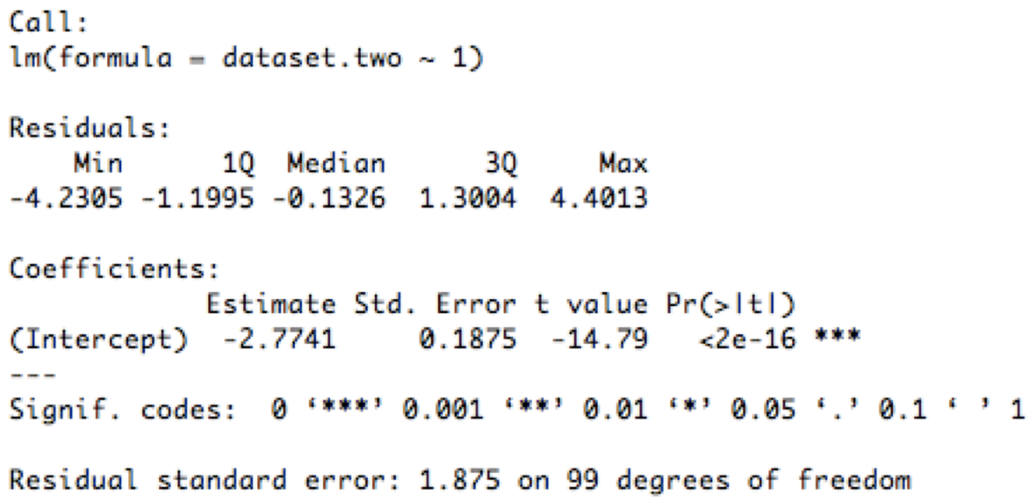
$y = mx + b$

or, more appropriately…

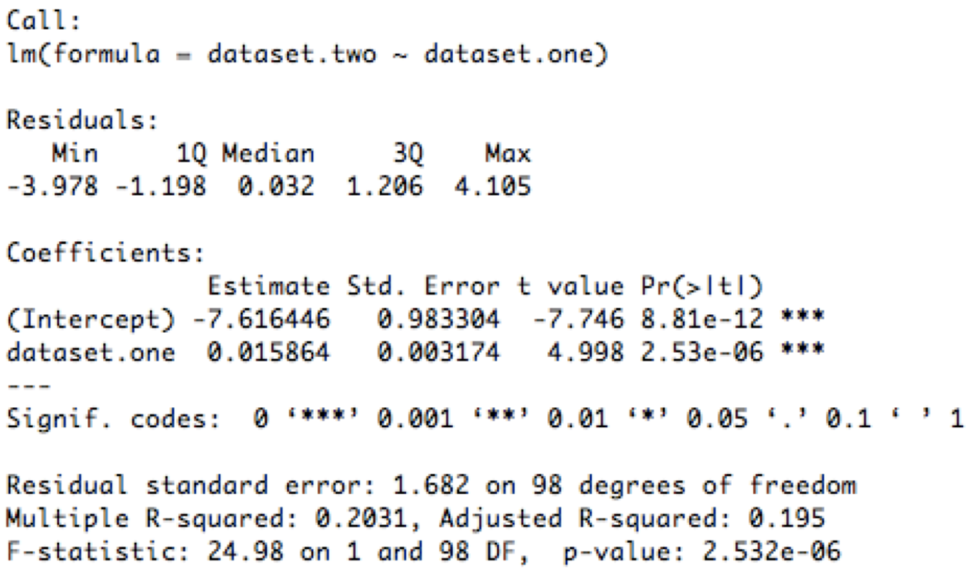
$y = \beta_0 + \beta_1 x + \varepsilon$



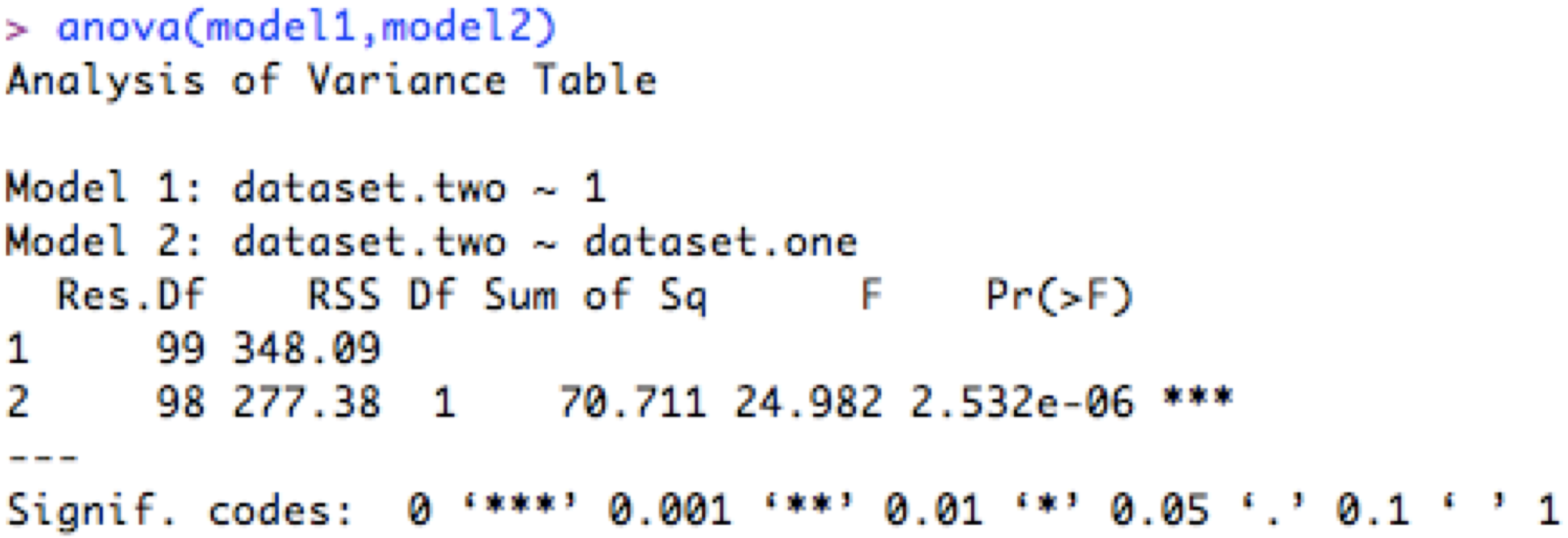
```
Call:
lm(formula = dataset.two ~ dataset.one)

Residuals:
   Min     1Q Median     3Q    Max 
-3.978 -1.198  0.032  1.206  4.105 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -7.616446   0.983304  -7.746 8.81e-12 ***
dataset.one  0.015864   0.003174   4.998 2.53e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.682 on 98 degrees of freedom
Multiple R-squared: 0.2031, Adjusted R-squared: 0.195 
F-statistic: 24.98 on 1 and 98 DF,  p-value: 2.532e-06
```



```
Call:
lm(formula = dataset.two ~ 1)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.2305 -1.1995 -0.1326  1.3004  4.4013 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -2.7741     0.1875  -14.79   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.875 on 99 degrees of freedom
```



```
Call:
lm(formula = dataset.two ~ dataset.one)

Residuals:
   Min     1Q Median     3Q    Max 
-3.978 -1.198  0.032  1.206  4.105 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -7.616446   0.983304  -7.746 8.81e-12 ***
dataset.one  0.015864   0.003174   4.998 2.53e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.682 on 98 degrees of freedom
Multiple R-squared: 0.2031, Adjusted R-squared: 0.195 
F-statistic: 24.98 on 1 and 98 DF,  p-value: 2.532e-06
```



```
> anova(model1,model2)
Analysis of Variance Table

Model 1: dataset.two ~ 1
Model 2: dataset.two ~ dataset.one
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1     99 348.09                                  
2     98 277.38  1    70.711 24.982 2.532e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The General Linear Model

A mathematical relationship between a criterion variable and one or more predictor variables.

The General Linear Model

A GLM has three components:

1) Some fixed influence that is the same for all data points

(expressed as the y intercept)

$y = \boldsymbol{\beta_0} + \beta_1 x + \varepsilon$

The General Linear Model

A GLM has three components:

2) Influence of other variables

(expressed as the slope)

$y = \beta_0 + \boldsymbol{\beta_1} x + \varepsilon$

The General Linear Model

A GLM has three components:

3) Other unmeasured influence, i.e., error

$$y = \beta_0 + \beta_1 x + \boldsymbol{\varepsilon}$$

The General Linear Model

Why is it a linear model?

Because the graphed relationship is a straight line, there are no power terms in the equation.

$$y = \beta_0 + \beta_1 x + \varepsilon$$

vs

$$y = \beta_0 + \beta_1 x^{2} + \varepsilon$$

The General Linear Model

Note, you could use a transformed variable, i.e., instead of using x^2 in the equation square x in the data, then a form of

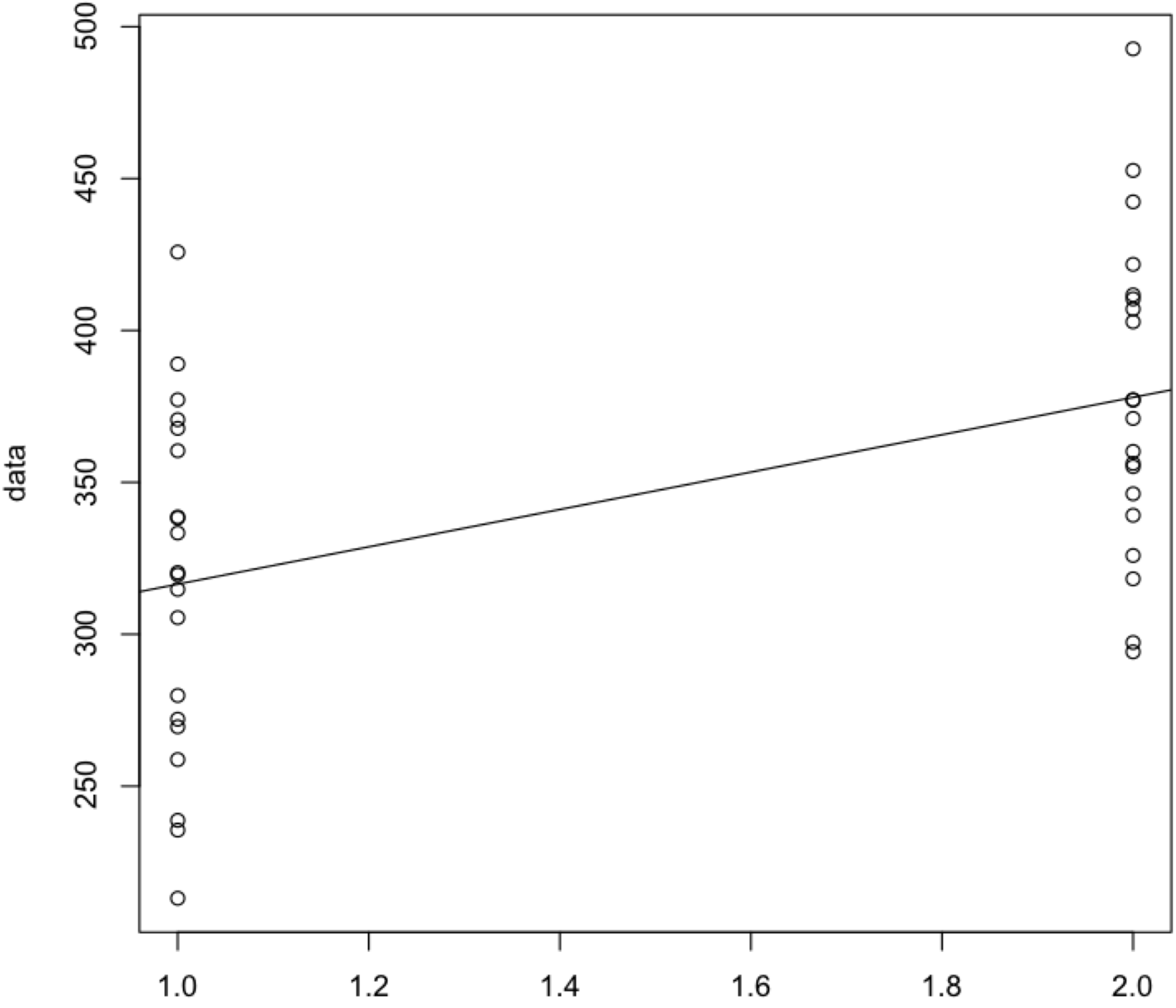
$$y = \beta_0 + \beta_1 x + \varepsilon$$

would be valid for the transformed data, and preserve the linearity of the model.

ANOVA as a special case of Multiple Regression

- ANOVA tests whether there is a difference on a measured variable between groups

- MR tests whether there is a relationship between a criterion variable and a predicted variable


data
500
450
400
350
300
250
1.0
1.2
1.4
1.6
1.8
2.0
group

Thus…

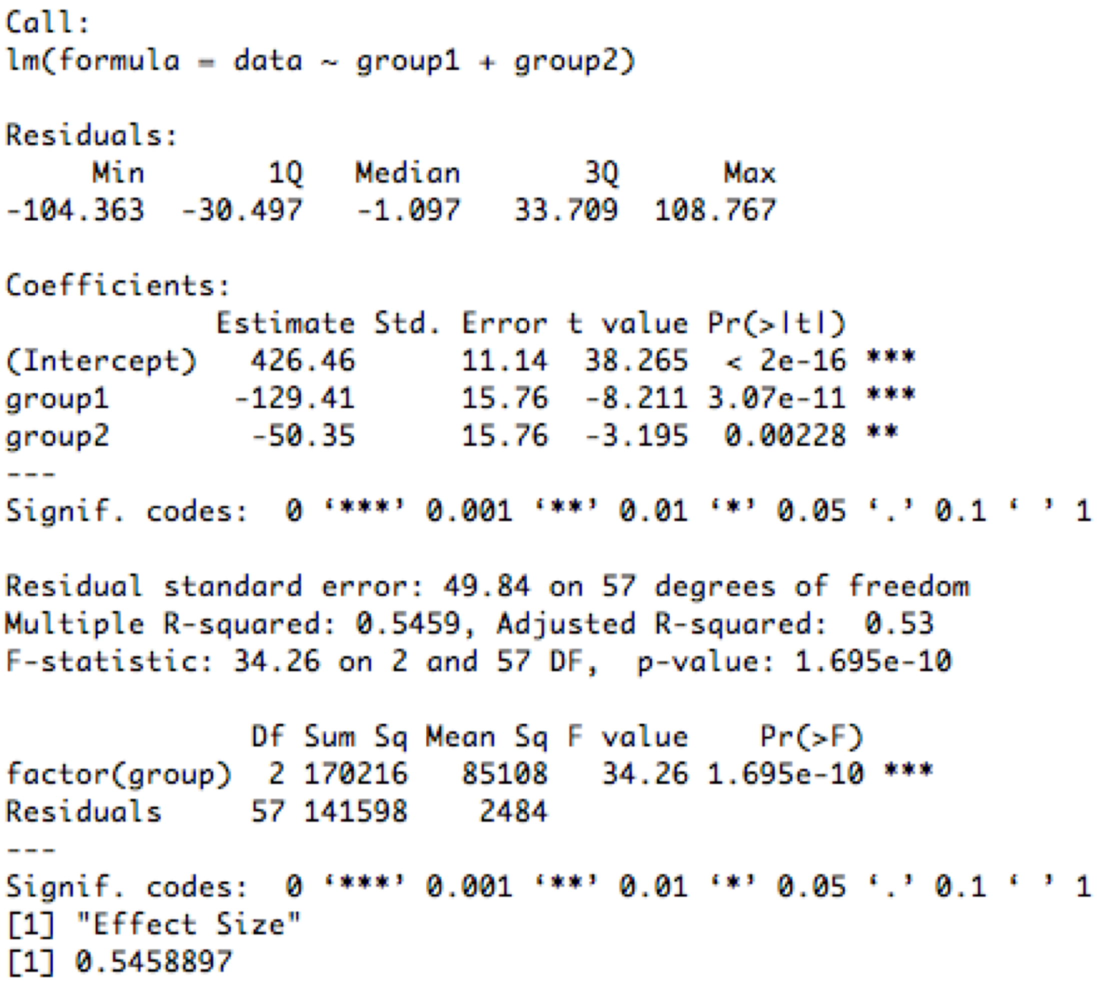
A single factor ANOVA has a linear model of the form:

$y = \beta_0 + \beta_1 x + \varepsilon$

But is typically expressed as:

$y = \upsilon + \alpha + \varepsilon$

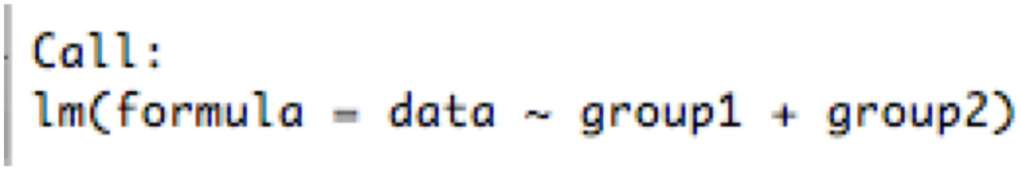
Now to implement this in R for a three level factor (i.e., groups 1,2,3) you need to use nominal coding



```
Call:
lm(formula = data ~ group1 + group2)

Residuals:
     Min       1Q   Median       3Q      Max 
-104.363  -30.497   -1.097   33.709  108.767 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   426.46      11.14  38.265  < 2e-16 ***
group1       -129.41      15.76  -8.211 3.07e-11 ***
group2        -50.35      15.76  -3.195  0.00228 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.84 on 57 degrees of freedom
Multiple R-squared: 0.5459,	Adjusted R-squared:  0.53 
F-statistic: 34.26 on 2 and 57 DF,  p-value: 1.695e-10

              Df Sum Sq Mean Sq F value    Pr(>F)    
factor(group)  2 170216   85108   34.26 1.695e-10 ***
Residuals     57 141598    2484                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] "Effect Size"
[1] 0.5458897
```

Now to implement this in R for a three level factor (i.e., groups 1,2,3) you need to use nominal coding

ANOVA Group	Regression Factor 1	Regression Factor 2
1	1	0
1	1	0
2	0	1
2	0	1
3	0	0
3	0	0



```
Call:
lm(formula = data ~ group1 + group2)
```

Note the difference between the model and the actual regression equation:

MODEL: $y = \upsilon + \alpha + \varepsilon$

REGRESSION: $y = \beta_0 + \beta_1 x + \beta_2 x + \varepsilon$