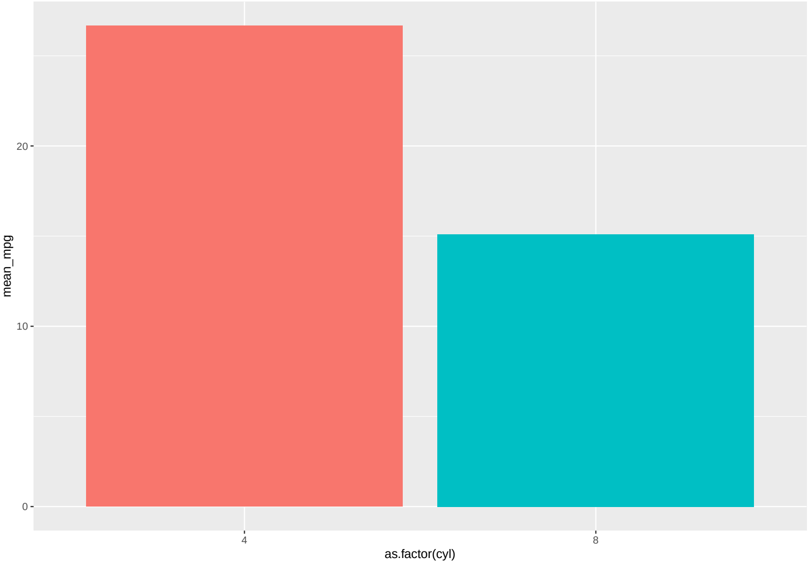
One-way ANOVA

Let's Add Another Group

- **ANOVA** or **Analysis of Variance** is a statistical model taht is used when we want to compare more than two independent means
  - What test have we covered that examined mean differences for different groups?
- really, its just an extension to the linear model that we have been covering from the beginning
- one major difference is the inclusion of the F statistic and therefore, the F table

Technically Categorical Predictors in the Linear Model



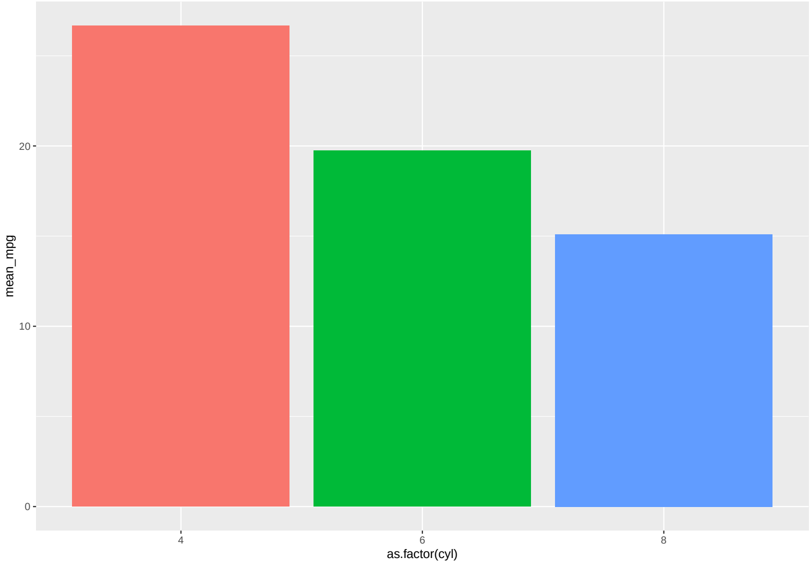
Linear Model to Compare Several Means

- the only difference now is that since we have multiple groups/samples to compare, we now have to incorporate dummy coding
  - binary variables can be handled by SPSS
- dummy coded variables will now represent the differences between means between the **reference group** and the other groups, this will be seen with our *b* values
  - Our example, we will be comparing 4 cylinder cars to the other cylinder cars(6- and 8-cyl cars)
- a one-way ANOVA with only two groups will give you the same answer as an independent-samples t-test

Linear Model to Compare Several Means

- our ANOVA will take two steps though
  - this is our first real instance of testing a linear model and the main points of what an ANOVA does
    - we get an F statistic that tells us there is a difference between our groups **generally**
    - then we make comparisons between the means of all of the groups to see which groups **specifically** differ from one another
- ANOVA is the same thing as Linear Regression
  - both are linear models
  - both can accept categorical IVs
  - both have continuous DVs
  - linear regression can also include continuous IVs
- linear regression can be useful for more complex issues, such as multiple predictors and unequal group sizes

Example



Example

- Hypotheses
  - H0: There will be no differences between the cylinder sizes in miles per gallon (MPG)
  - H1: There will be differences between the cylinder sizes in MPG
  - H1: 4-cylinder cars will differ in MPG from 6-cylinder and 8-cylinder cars
  - H1: 4-cylinder cars will have better MPG than 6-cylinder and 8-cylinder cars

Example

```
     8 4 6
[1,] 0 0 1
[2,] 0 0 1
[3,] 0 1 0
[4,] 0 0 1
[5,] 1 0 0
[6,] 0 0 1
```

```
   vars  n  mean   sd median trimmed  mad  min  max range s
X1    1 32 20.09 6.03   19.2    19.7 5.41 10.4 33.9  23.5 0
```

Example

```
 Descriptive statistics by group
group: 4
   vars  n  mean   sd median trimmed  mad  min  max range s
X1    1 11 26.66 4.51     26   26.44 6.52 21.4 33.9  12.5 0
------------------------------------------------------------
group: 6
   vars n  mean   sd median trimmed  mad  min  max range  s
X1    1 7 19.74 1.45   19.7   19.74 1.93 17.8 21.4   3.6 -0
------------------------------------------------------------
group: 8
   vars  n mean   sd median trimmed  mad  min  max range  s
X1    1 14 15.1 2.56   15.2   15.15 1.56 10.4 19.2   8.8 -0
```

Example

$$outcome_i = (model) + error_i$$

```
                Df Sum Sq Mean Sq F value   Pr(>F)    
as.factor(cyl)  2  824.8   412.4    39.7 4.98e-09 ***
Residuals      29  301.3    10.4                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Example

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = mpg ~ as.factor(cyl), data = mtcars)

$`as.factor(cyl)`
          diff        lwr        upr     p adj
6-4  -6.920779 -10.769350 -3.0722086 0.0003424
8-4 -11.563636 -14.770779 -8.3564942 0.0000000
8-6  -4.642857  -8.327583 -0.9581313 0.0112287
```

Example

```
lm(formula = mpg ~ as.factor(cyl), data = mtcars)
                coef.est coef.se t value Pr(>|t|)
(Intercept)      26.66     0.97   27.44    0.00
as.factor(cyl)6  -6.92     1.56   -4.44    0.00
as.factor(cyl)8 -11.56     1.30   -8.90    0.00
---
n = 32, k = 3
residual sd = 3.22, R-Squared = 0.73
```

Linear Model to Compare Several Means

- previously we used a dummy variable comparing 4-cylinder and 8-cylinder cars with one of the dummy variables included in the model

$$outcome_i = (model) + error_i$$

- now because we have multiple groups, we will be including two dummy variables into our model
  - we will compare two groups to our reference group (which can be thought of as a control group)

$$MPG_i = b_0 + b1(6cyl_i) + b2(8cyl_i) + \epsilon_i$$

Linear Model to Compare Several Means

- so we can first look at the value for our reference group to get the intercept
  - because we dummy coded these variables, since we are only focused on the 4-cylinder group, we will include zeros for the other two groups

$$MPG_i = b_0 + b1(0) + b2(0)$$

$$MPG_i = b_0$$

$$X_{4cyl} = b_0$$

Linear Model to Compare Several Means

- now if we look at the 6-cylinder group, we can then change the dummy coding to reflect that group

$$MPG_i = b_0 + b1(1) + b2(0)$$

$$MPG_i = b_0 + b_1$$

Linear Model to Compare Several Means

- we can then get the expected value for a 6-cylinder car with the information we already know
  - we know that the intercept is now equal to average MPG for our reference group (4-cylinder)

$$MPG_i = b_0 + b_1$$

$$X_{6cyl} = X_{4cyl} + b_1$$

$$X_{6cyl} - X_{4cyl} = b_1$$

Linear Model to Compare Several Means

- now if we look at the 8-cylinder group, we can then change the dummy coding to reflect that group

$$MPG_i = b_0 + b1(0) + b2(1)$$

$$MPG_i = b_0 + b_2$$

Linear Model to Compare Several Means

$$MPG_i = b_0 + b_2$$

$$X_{8cyl} = X_{4cyl} + b_2$$

$$X_{8cyl} - X_{4cyl} = b_2$$

Linear Model to Compare Several Means

- by utilizing dummy coding, we can now have the differences in means between our three groups
  - you can do this with as many groups as you'd like but after so many comparisons, they begin to get meaningless
  - Ex: if you were to compare all 50 states in violent crime rates
    - what state would be your reference group
    - does it matter if you compare one state to the other 49
- we'll also cover contrast coding, which uses the dummy variables and the b values to represent differences between groups before collecting data and go along with your hypotheses
  - this is different from the common approach of using post-hoc analyses, which compares every single possible comparison, even if you did not hypothesize about a specific comparison

Linear Model to Compare Several Means

- from the example above, we will get a F statistic
  - within that, we will have the model fit
  - then the residual/error, which is the unknown from our tested model
- Additionally, we will have coefficients (bs) that are once again the differences between the reference group and the other group we are comparing to the reference group

Logic of the F-statistic

- the F statistic or F ratio is the overall fit of the linear model
- some guidelines for the F statistic
  - the model that represents “no effect/relationship” is a model where the predicted value of the outcome is always the grand mean (mean of the outcome variable)
  - a different model that is fit represents our alternative hypothesis
    - we compare the fits of the two models using the grand mean
  - intercept and additional parameters describe the model

Logic of the F-statistic

- parameters determine the shape of the model fit
  - bigger coefficients, larger deviation between model and the null model (grand mean)
- parameters (b) represent differences between group means
- if differences between group means are large enough, the model will fit better than the null model (grand mean)
- if this is the case, then our model of comparing group means is better than the null model (grand mean) and the group means are significantly different from the null