214B: Lab 4

ANOVA: Categorical Predictors

TA: Melissa G Wolf

HW3 Recap

Transforming variables

In the last homework assignment, you were asked to interpret the slope for Income in the MLR model. The correct interpretation was: "Controlling for math identity, a one-unit increase in income is associated with a .001 unit increase in math scores ($\beta = .001$, $p < .001$)."

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	13.526	.578		23.420	.000	12.394	14.658
	Income	.001	.000	.399	61.933	.000	.000	.001
	Does teen see self as math person?	5.357	.128	.269	41.779	.000	5.105	5.608

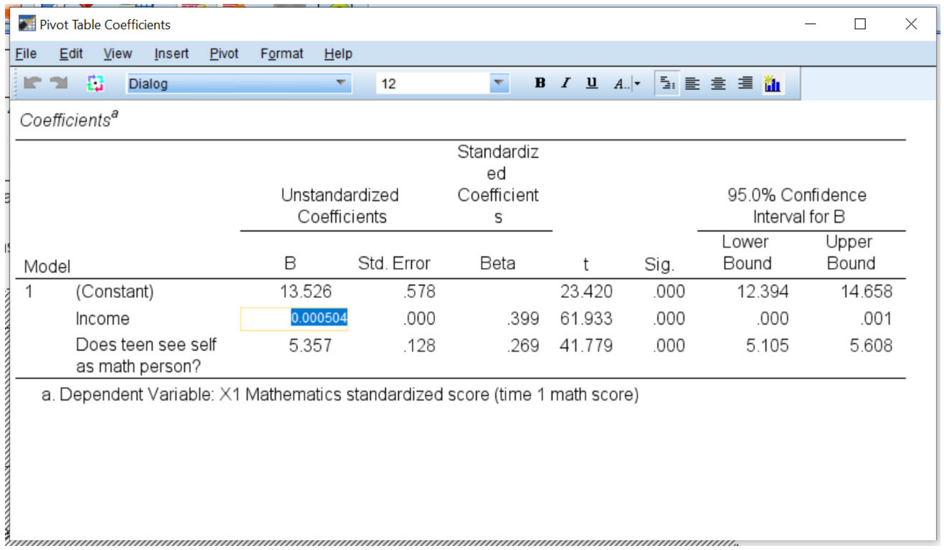
a. Dependent Variable: X1 Mathematics standardized score (time 1 math score)

Recall that in the last lab, we discussed how **math scores** and **income** were on very different scales. For example, income was reported in USD, with a minimum of \$50,698 and a maximum of \$98,807, while math scores ranged from 24.02 to 82.19. Thus, an increase in annual income of \$1 USD was associated a .001 unit increase in math scores.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Income	21444	50698.00	98807.00	70540.942	7803.0420
X1 Mathematics standardized score (time 1 math score)	21444	24.02	82.19	51.1096	10.07767
Valid N (listwise)	21444				

It's clear that we need to change the scale of the variables to make our coefficients more interpretable. However, before we do that, we need to examine the income coefficient more closely. What happens when we double click on it?



Pivot Table Coefficients

File Edit View Insert Pivot Format Help

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	13.526	.578		23.420	.000	12.394	14.658
	Income	0.000504	.000	.399	61.933	.000	.000	.001
	Does teen see self as math person?	5.357	.128	.269	41.779	.000	5.105	5.608

a. Dependent Variable: X1 Mathematics standardized score (time 1 math score)

We see that the slope coefficient is actually .000504 (rounded to the 6th decimal place). Thus, our updated interpretation of the slope coefficient is: "Controlling for math identity, a one-unit increase in income is associated with a .000504 unit increase in math scores ($\beta = .000504$, $p < .001$)."

Answer quiz question 1

Now, let's make our regresion coefficients more interpretable. What happens if we multiply each of the numbers by 1,000?

Income unit:

```
1*1000
```

```
## [1] 1000
```

Income slope coefficient:

```
.000504*1000
```

```
## [1] 0.504
```

What is our new interpretation of this regression coefficient?

Answer quiz question 2

Dummy coding

To use a categorical predictor with more than 2 levels in a regression model, we need to recode the variable into multiple variables to create a reference category. It is most common (and easiest) to use a method called "dummy coding". We can think of dummy coding as creating a bunch of dichotomous variables, where 0 is always the same reference category throughout. *We always need k-1 dummy variables, where k is the number of categories.*

Imagine we have a predictor variable called **Couch Color** with 3 categories: Brown, Dark Grey, and Light Grey. Let's make Brown the reference category, Dark Grey the first dummy variable, and Light Grey the second dummy variable. Our outcome variable is **Couch Cost**.

Let's say we observe the following equation:

$$\hat{y} = 800 + 300 * d_1 + 500 * d_2 + e_i$$

We'll use this matrix below to plug values into an equation.

	Dummy 1 Coefficient (d1)	Dummy 2 Coefficient (d2)
Reference Category (Brown)	0	0
Dark Grey	1	0
Light Grey	0	1

What is the expected cost a brown couch?

$$\hat{y} = 800 + (300 * 0) + (500 * 0)$$
$$\hat{y} = 800$$

What is the expected cost a dark grey couch?

$$\hat{y} = 800 + (300 * 1) + (500 * 0)$$
$$\hat{y} = 1100$$

Answer quiz question 3

Regression with one categorical predictor with 3+ levels

Open the Week 4 dataset from the lab folder. Our predictor variable will be Political Party (**Party**) and our outcome variable will be Voter Likelihood (**Likelihood**). Let's see which party is the most likely to vote in this election!

Creating the dummy variables

The first thing we need to do is create dummy variables. Let's begin by running the frequencies for this variable to see how many dummy variables we need to create (and make sure we have no missing data).

Select **Analyze** > **Descriptive Statistics** > **Frequencies** and move **Party** into the **Variables(s)** box. Press **OK**.

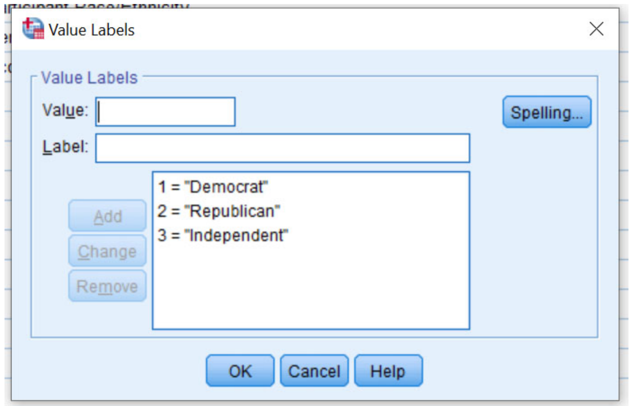
Political Party

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Democrat	175	35.0	35.0	35.0
	Republican	182	36.4	36.4	71.4
	Independent	143	28.6	28.6	100.0
	Total	500	100.0	100.0	

In R:

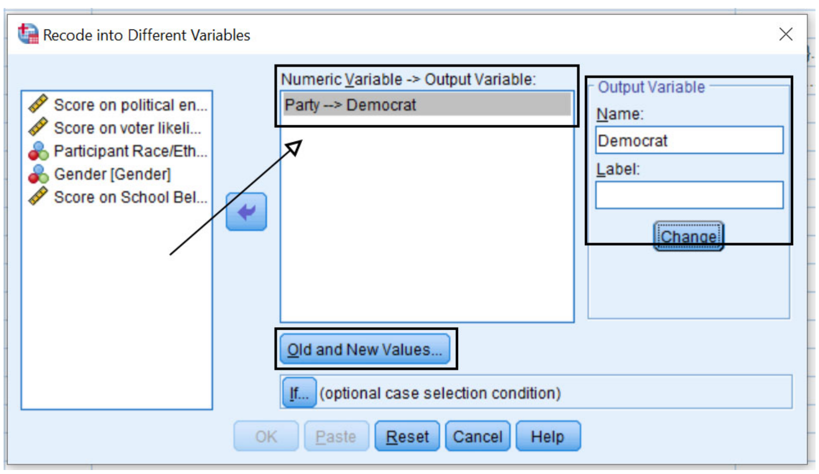
```
sjmisc::frq(Week4$Party)
```

There are three categories, so we need two dummy variables. Let's make **Independent** our reference category. This means we need to create two dummy variables: one for Democrat, and one for Republican. A quick glance at **Variable View** shows us what each party is coded as:

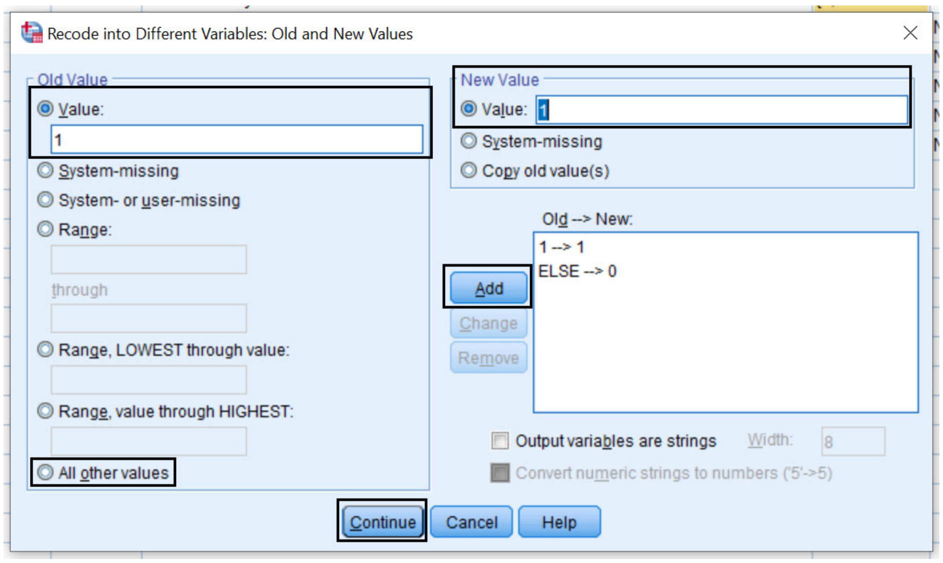


Recoding. . . we've done this before!

1. Select **Transform** > **Recode into Different Variables**.
2. Drag **Party** into the **Variable** box.
3. Under **Output Variable**, type **Democrat** under **Name** and press **Change**.
4. Select **Old and New Values**

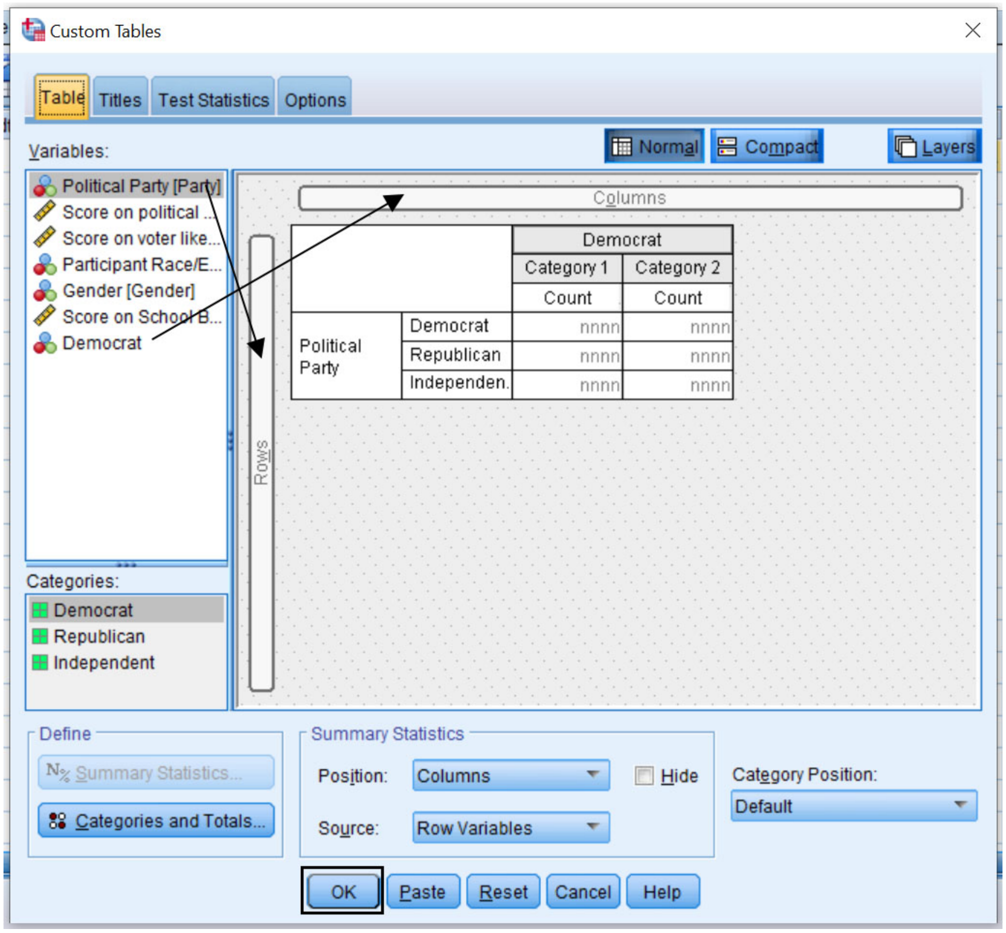


5. Under Variable View, we saw that **Democrat** is coded as **1**. Under **Old Value**, type in **1**. Under **New Value**, type in **1**. Press **Add**.
6. As we saw in the matrix above, all other variables get a value of **0** to drop out of the model when this variable is activated. Under **Old Value**, select **All other values**. Under **New Value**, type in **0**. Press **Add**.
7. Press **Continue** and then select **OK**.



How can we double check that we created this dummy variable correctly?

1. Select **Analyze** > **Tables** > **Custom Tables**
2. Drag **Party** onto **Rows** and the new variable **Democrat** onto **Columns**
3. Press **OK**

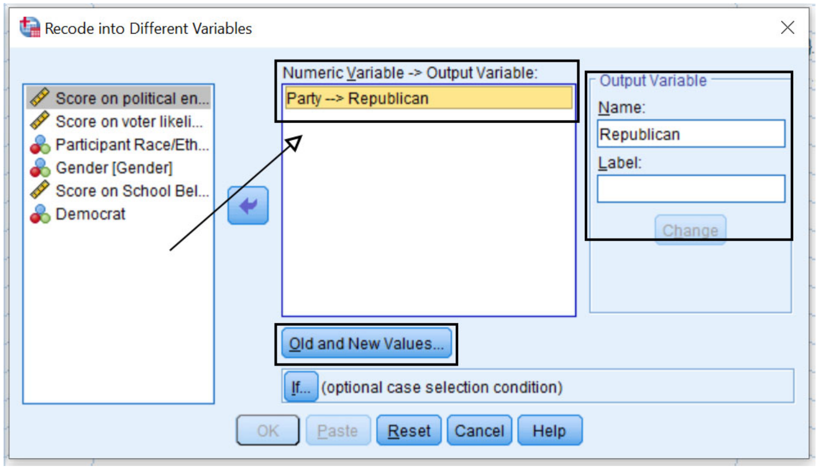


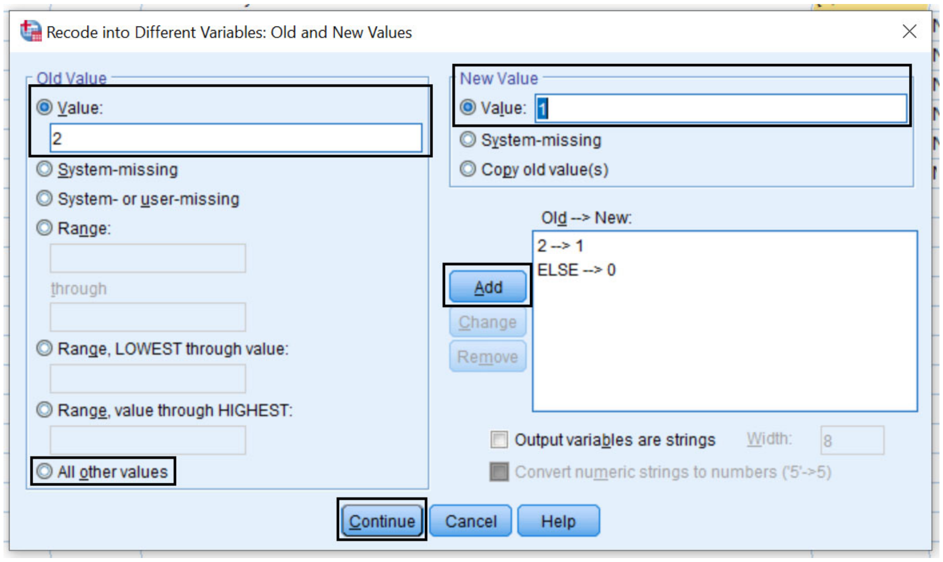
As we see in the table below, all values of Democrat are correctly coded as 1, and all other values are correctly coded as 0.

		Democrat	
		.00	1.00
		Count	Count
Political Party	Democrat	0	175
	Republican	182	0
	Independent	143	0

Repeat the same process to create the Republican dummy variable

Important: Make sure to the value of 2 into 1, and all other values into 0. The dummy variable always gets a value of 1.





As we see in the table below, we've successfully created two dummy variables:

- Democrat: 1's for Democrats, 0's for Independents and Republicans
- Republican: 1's for Republicans, 0's for Independents and Democrats

Custom Tables

		Democrat		Republican	
		.00	1.00	.00	1.00
		Count	Count	Count	Count
Political Party	Democrat	0	175	175	0
	Republican	182	0	0	182
	Independent	143	0	143	0

In R

```
Week4$Democrat <- sjmisc::rec(Week4$Party, rec="1=1;else=0")
Week4$Republican <- sjmisc::rec(Week4$Party, rec="2=1;else=0")

xtabs(~Party+Democrat,data=Week4)
xtabs(~Party+Republican,data=Week4)
```

Running the regression model

1. Select **Analyze** > **Regression** > **Linear**.
2. Drag **Likelihood** into the **Dependent** box.
3. Drag **Democrat** and **Republican** into the **Independent(s)** box.
4. Under **Statistics** select **Confidence Intervals**

Your output should look like this:

Coefficients[a]

	Independent	Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
Model		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	51.124	1.125		45.457	.000	48.914	53.333
	Democrat	-11.547	1.516	-.389	-7.617	.000	-14.526	-8.569
	Republican	-7.251	1.503	-.246	-4.825	.000	-10.204	-4.298

a. Dependent Variable: Score on voter likelihood survey (highest = most likely to vote)

In R:

```
summary(lm(Likelihood~Democrat+Republican,data=Week4))
```

$$\hat{y} = 51.124 - 11.547 * d_{dem} - 7.251 * d_{rep} + e_i$$

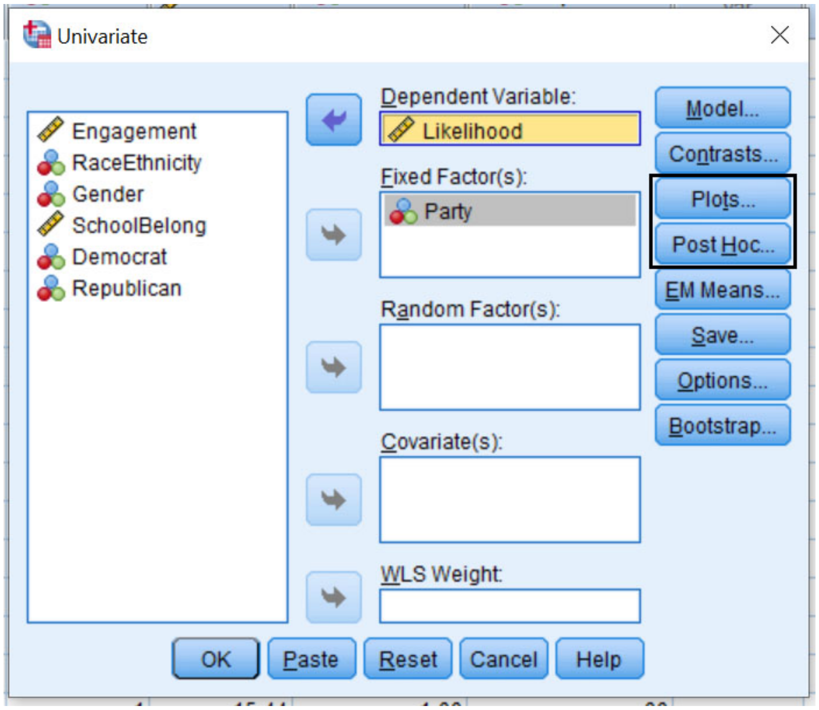
Answer quiz questions 4, 5 and 6

ANOVA model

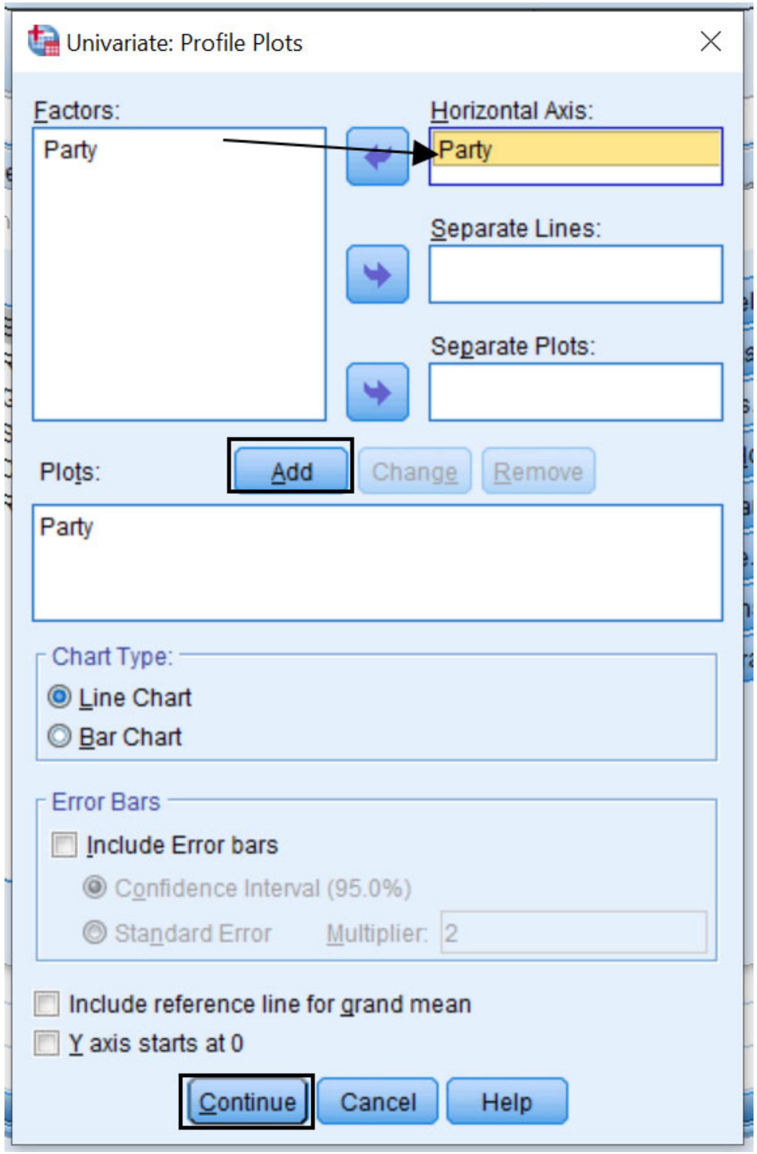
ANOVA models are helpful because they use a type of coding called effects coding, which allows us to compare all of the groups with each other. Let's replicate this problem using an ANOVA model so that we can compare all of the groups. The ANOVA model gives us the omnibus F-test ("Is there a significant difference anywhere in the model?") and we can then use post hoc tests to compare the group means. We'll use a Type I error correction so that our Type I error rate does not exceed .05.

1. Select **Analyze** > **General Linear Model** > **Univariate**.

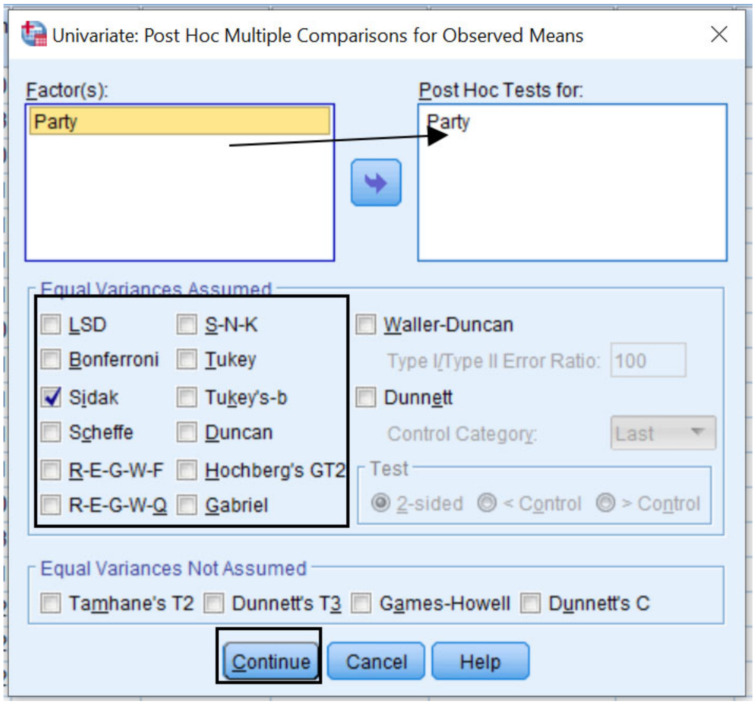
2. Drag **Likelihood** into the **Dependent Variable** box.
3. Drag **Party** into the **Fixed Factor(s)** box.



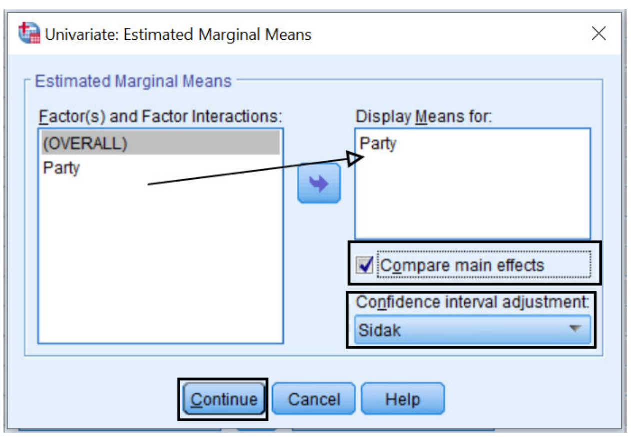
4. Select **Plots**.
5. Move **Party** from **Factors** to **Horizontal Axis** and press **Add**. Select **Continue**.



6. Select **Post Hoc**.
7. Move **Party** from **Factors** to the **Post Hoc Tests for** box.
8. Select **Sidak** and press **Continue**.



9. Select **EM Means**.
10. Move **Party** from **Factors** to the **Display Means for** box.
11. Select **Compare Main Effects**.
12. Under Confidence Interval Adjustment, select **Sidak**.
13. Press **Continue**.
14. Press **OK**



You should see the following output:

Between-Subjects Factors

		Value Label	N
Political Party	1	Democrat	175
	2	Republica n	182
	3	Independe nt	143

Tests of Between-Subjects Effects

Dependent Variable: Score on voter likelihood survey (highest = most likely to vote)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	10586.49[a]	2	5293.245	29.265	.000
Intercept	994941.71	1	994941.71	5500.781	.000
Party	10586.491	2	5293.245	29.265	.000
Error	89893.792	497	180.873		
Total	1088054.4	500			
Corrected Total	100480.28	499			

a. R Squared = .105 (Adjusted R Squared = .102)

Post Hoc Tests

Political Party

Multiple Comparisons

Dependent Variable: Score on voter likelihood survey (highest = most likely to vote)
Sidak

(I) Political Party	(J) Political Party	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Democrat	Republican	-4.2964*	1.42386	.008	-7.7077	-.8851
	Independent	-11.5475*	1.51605	.000	-15.1796	-7.9153
Republican	Democrat	4.2964*	1.42386	.008	.8851	7.7077
	Independent	-7.2511*	1.50288	.000	-10.8517	-3.6504
Independent	Democrat	11.5475*	1.51605	.000	7.9153	15.1796
	Republican	7.2511*	1.50288	.000	3.6504	10.8517

Based on observed means.
The error term is Mean Square(Error) = 180.873.
*. The mean difference is significant at the .05 level.

Estimated Marginal Means

Political Party

Estimates

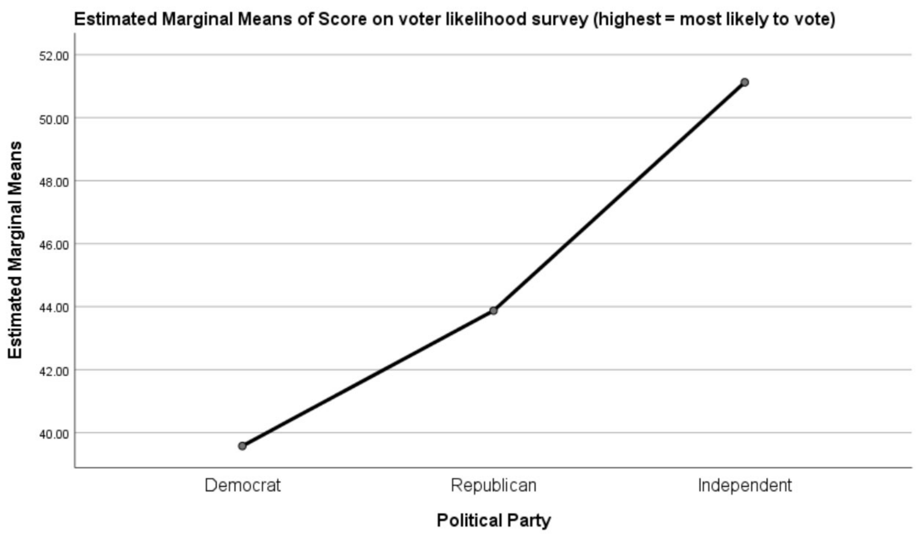
Dependent Variable: Score on voter likelihood survey (highest = most likely to vote)

Political Party	Mean	Std. Error	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Democrat	39.576	1.017	37.579	41.574
Republican	43.873	.997	41.914	45.831
Independent	51.124	1.125	48.914	53.333

Pairwise Comparisons

Dependent Variable: Score on voter likelihood survey (highest = most likely to vote)

(I) Political Party	(J) Political Party	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
Democrat	Republican	-4.296*	1.424	.008	-7.708	-.885
	Independent	-11.547*	1.516	.000	-15.180	-7.915
Republican	Democrat	4.296*	1.424	.008	.885	7.708
	Independent	-7.251*	1.503	.000	-10.852	-3.650
Independent	Democrat	11.547*	1.516	.000	7.915	15.180
	Republican	7.251*	1.503	.000	3.650	10.852



In R:

```
Week4$Party<-factor(Week4$Party,
                    levels=c(1,2,3),
                    labels=c("Democrat","Republican","Other"))
anova.model <- aov(Likelihood~Party,data=Week4)
summary(anova.model)
DescTools::PostHocTest(anova.model,method="hsd")
emmeans::emmeans(anova.model, ~Party)
emmeans::emmip(anova.model, ~Party)
```

How do we interpret this in APA format?

A one-way ANOVA was conducted to compare the effect of political party membership (Independent/Republican/Democrat) on voter likelihood. Voter likelihood was calculated using responses to a survey about voter likelihood (M = 44.44, SD = 14.19, min = 4.07, max = 92.88; see Figure 1 below). There was a significant effect of political party membership on voter likelihood $[F(2, 497) = 29.265, p < .01]$. Post hoc comparisons using a Dunn-Sidak correction revealed significant differences between all three political parties at the .01 alpha level. Independents were the most likely to vote (M = 51.124) followed by Republicans (M = 43.873) and Democrats (M = 39.576; see Figure 2 below).

Figure 1

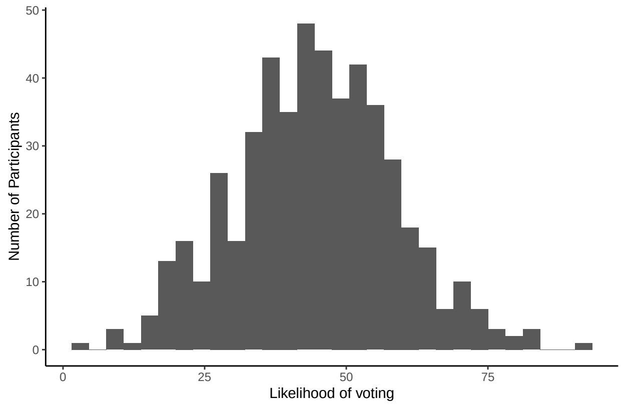
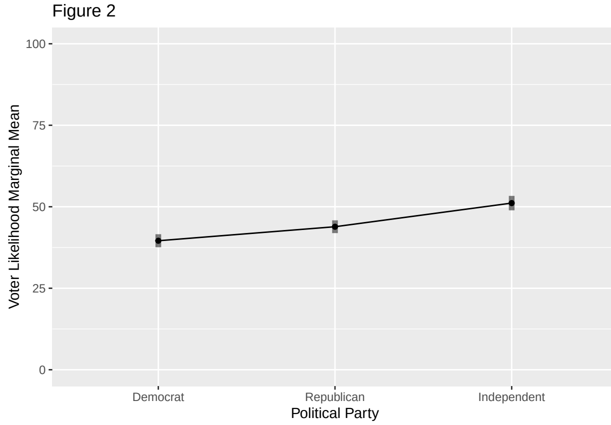


Figure 2



What are Marginal Means?: Marginal means are the model predicted means for each group, controlling for other variables in the model. In this case, there are no other variables in the model so the marginal means are the actual means. We want to use the marginal means because we are creating a "model" for a reason - we are trying to estimate the population parameters and we do not want to report sample dependent estimates.

How did we get the overall mean/sd/min/max of the outcome variable?: Simple descriptive statistics!

Note: We get the F-statistic from the "Test of Between Subjects Effects" table in SPSS, or the ANOVA table in R. We get the means from the estimated marginal means table. We get the statistical significance from the Post Hoc Tests table.

Note: Figure 2 is the same as the marginal means plot from SPSS but with a rescaled y-axis.

Answer quiz questions 7 and 8.