Lab 2: Regression Continued

214B Winter 2020

TA: Melissa Gordon Wolf

Simple Linear Regression

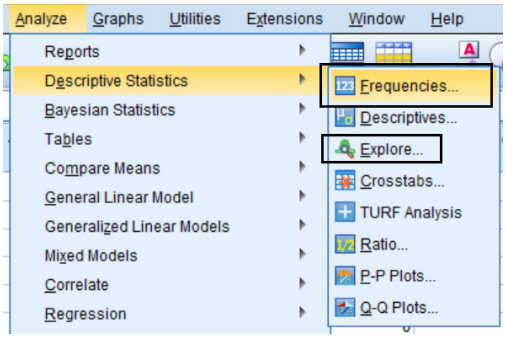
Research question: Do people that enjoy reading tend to do better in school?

Testable hypothesis: Does reading enjoyment (*enjoyread*) predict verbal test scores (*verbalscr*)?

Always begin by checking the descriptive statistics

(Answer quiz questions 1 and 2)

Run the appropriate descriptive statistics for each variable given it's scale type. Select **Analyze** > **Descriptive Statistics**, and then select **Frequencies** or **Explore**.



If you run **Frequencies**, select the **Statistics** menu and the **Charts** menu.

Under **Statistics**, select:

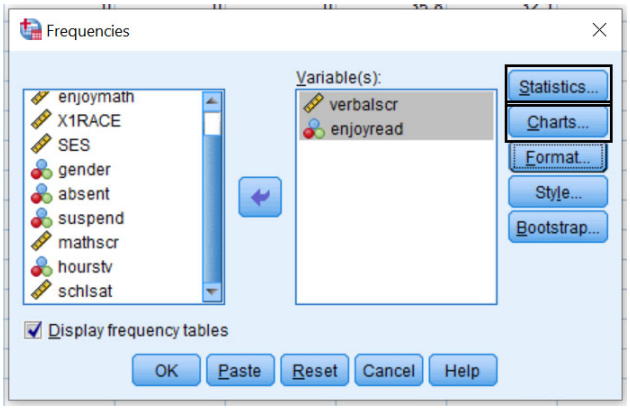
1. Mean
2. Median
3. Std. deviation
4. Minimum
5. Maximum
6. Skewness

Note: Generating these statistics is only appropriate if the variable is continuous.

Under **Charts**, select:

- **Bar charts** if the variable is categorical or ordinal
- **Histograms** if the variable is ordinal or continuous

Hint: We can use either a bar chart or a histogram for an ordinal variable



In R

Frequencies:

```
summarytools::freq(df$variable)

df %>%
  ggplot(aes(x=variable))+
  geom_bar()
```

Descriptives:

```
psych::describe(df$variable)

df %>%
  ggplot(aes(x=variable))+
  geom_histogram()
```

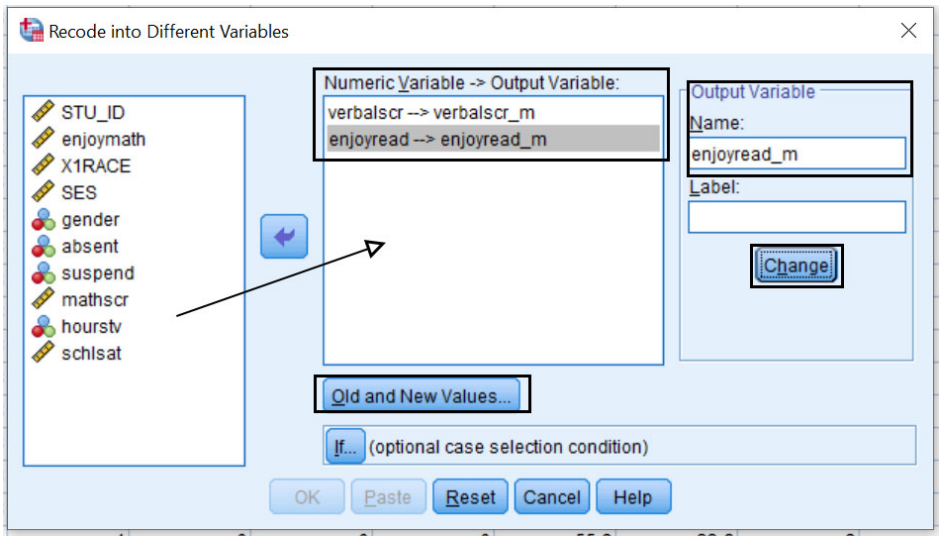
Why do you think these graphs aren't working?

Recode the unusual values as missing

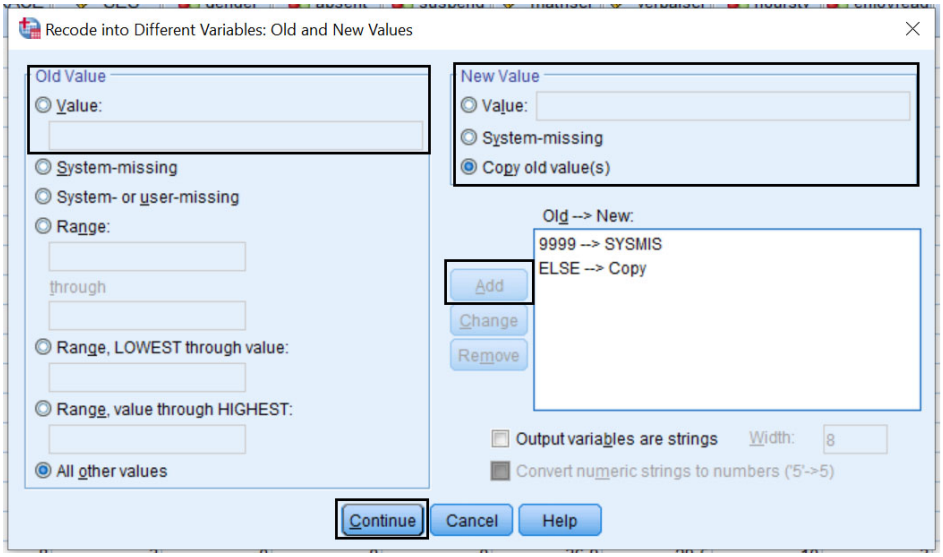
(Answer quiz question 3)

You will often get datasets where the coding scheme isn't transparent, so it's important to check the data to make sure there aren't any unusual values. In this case, we see a bunch of "9999's" in each variable. We should immediately be suspicious about these values. At this point, you would contact the person who gave you the data or check the reference manual. As your point of contact, I will tell you that these 9999 values are missing data.

1. Select **Transform** > **Recode into different variables**
2. Drag **enjoyread** and **verbalscr** into the **Variable** box.
3. Under **Output Variable**, type a new name for each variable. We'll use **enjoyread_m** and **verbalscr_m** but you can use whatever you want.
4. Press **CHANGE** each time you update the variable name!
5. Select **Old and New Values**



6. Under **Old Value**, select **9999**
7. Under **New Value**, select **System-missing**
8. Press **ADD!**
9. Under **Old Value**, select **All other values**
10. Under **New Value**, select **Copy old value(s)**
11. Press **ADD!**
12. Select **Continue**
13. Press **OK**



14. Look at your **Data View** tab in your dataset and scroll to the right. You'll notice that you have two new variables called **enjoyread_m** and **verbalscr_m** where all of the 9999 values are now coded as missing.

Important: Make sure to USE these new variables in your analyses!

In R

1. Use the `sjmisc` package

```
df$enjoyread_m<-sjmisc::rec(df$enjoyread,rec="9999=NA;else=copy")
```

2. Use `dplyr` (from `tidyverse`)

```
df$enjoyread_m<-na_if(df$enjoyread,9999)
```

Do the same for verbalscr

Check the descriptives again of the new variable

Make sure that you aren't missing anything and that the variables look okay

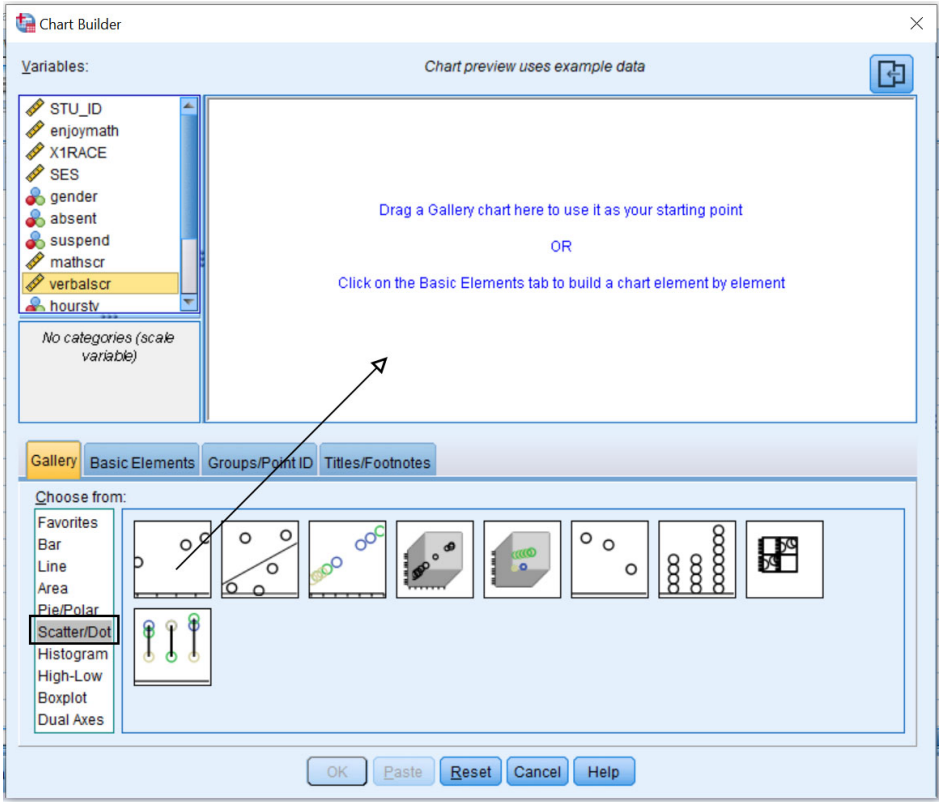
Finally... Simple Linear Regression!

Scatterplot

(Answer quiz question 4)

First, let's plot the relationship between the two variables to make sure the relationship is linear, and therefore appropriate to run a simple linear regression model on.

1. Select **Graphs** > **Chart Builder**
2. Under **Gallery**, select **Scatter/Dot** and drag **Simple Scatter** onto the **Chart Preview**



3. Drag **enjoyread_m** onto the x-axis and **verbalscr_m** onto the y-axis.
4. Press **OK**

In R

1. Use the plot command from base R. *Figure this out yourself*
2. Use ggplot.

```
ggplot(df,aes(enjoyread_m,verbalscr_m))+
  geom_point()+
  theme_minimal()
```

Regression Model

(Answer quiz questions 5, 6 and 7)

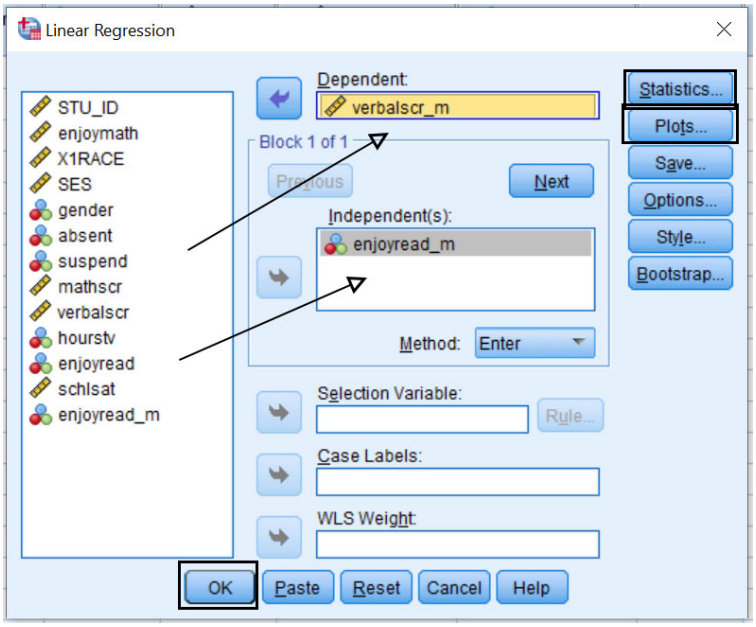
Test of independence:

$H_0 : \beta = 0$

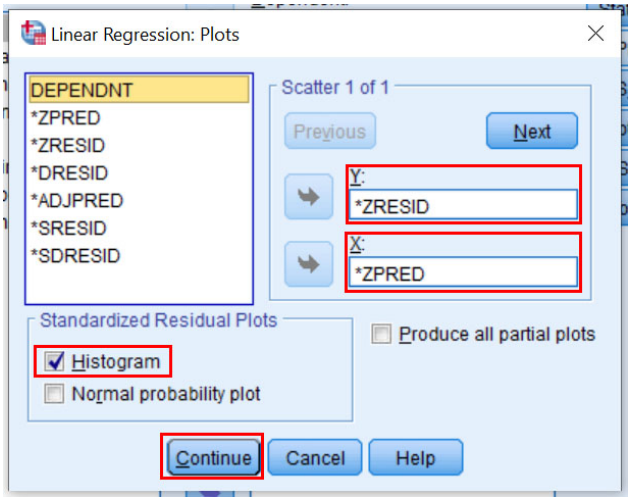
$H_A : \beta \neq 0$

We want to know if there is a relationship between **enjoyread_m** and **verbalscr_m**. We assume there is no relationship (H_0 or the null hypothesis). Note that β is the slope.

1. Select **Analyze** > **Regression** > **Linear**
2. Drag **verbalscr_m** into the **Dependent** box and **enjoyread_m** into the **Independent(s)** box.
3. Under **Statistics** select **Confidence Intervals** and press **Continue**



4. Under **Plot**, drag **ZRESID** (standardized residuals) into the **Y** box, and **ZPRED** (standardized predicted values) into the **X** box. Select the **Histogram** under the **Standardized Residual Plots**.



You should see the following output:

- **Model Summary**
  - Gives us the "omnibus" or "overall" model results
  - Look for the correlation and R^2
- **ANOVA Table**
  - Gives us the "omnibus" or "overall" model results
  - Regression Sum of Squares + Residual Sum of Squares = Total Sum of Squares
    - These are used to determine if the F-test is significant
- **Coefficients table**
  - Slope and intercept
  - A one unit increase in enjoyread_m is associated with a 2.77 increase in verbalscr_m.
  - Note the 95% confidence intervals for the coefficients. Do they contain 0?
- **Residual Scatterplot**
  - Plots the residuals against the predicted values
    - AKA: Is there a relationship between the model predicted test scores and the residuals (errors)
  - Used to evaluate the homoscedasticity assumption
    - Ideally, we'd like to see a completely random pattern
  - Because the residuals and predicted values are standardized, this tells us how "spread out" they are

Regression Equation

$$\hat{y} = \beta_0 + \beta_1 * x$$
$$\hat{y} = 20.76 + 2.77 * enjoyread - m$$

In R

1. Use base R to run the model

```
model<-lm(y~x,data=df)
summary(model)
anova(model)
```

Advanced: Pretty output created using the stargazer package

```
stargazer(model,ci=TRUE)
```

2. Plot the residuals on a scatterplot using base R

Table 1:

	Dependent variable:
	verbalscr_m
enjoyread_m	2.772***
	(2.688, 2.856)
Constant	20.760***
	(20.486, 21.034)
Observations	3,056
R^2	0.579
Adjusted R^2	0.579
Residual Std. Error	2.091 (df = 3054)
F Statistic	4,195.663*** (df = 1; 3054)
Note:	*p<0.1; **p<0.05; ***p<0.01

Scatterplot will be reversed from SPSS but will contain same info

```
plot(model)
hist(model$residuals)
```

Multiple Linear Regression

(**Answer quiz questions 8 and 9**)

Research question: Do people that enjoy reading and enjoy school tend to do better in school?

Testable hypothesis: Does reading enjoyment (*enjoyread*) and school satisfation (*schsat*) predict verbal test scores (*verbalscr*)?

Always begin by checking the descriptive statistics

Same as above

Recode any missing values

Same as above

Check the descriptives of the new variables

Same as above

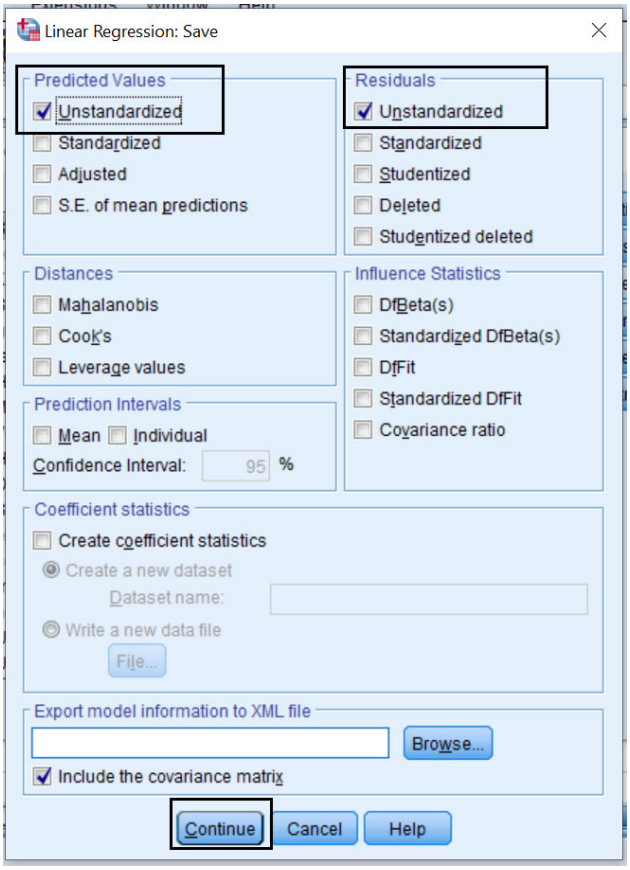
Plot the relationship between all the DV and the IV

Same as above

Run the regression equation!

1. Select **Analyze** > **Regression** > **Linear**
2. Drag **verbalscr_m** into the **Dependent** box and **enjoyread_m** and **schlsat** into the **Independent(s)** box.
3. Under **Statistics** select **Confidence Intervals** and press **Continue**

4. Under **Plot**, drag **ZRESID** (standardized residuals) into the **Y** box, and **ZPRED** (standardized predicted values) into the **X** box. Select the **Histogram** under the **Standardized Residual Plots**.
5. Under **Save**, select **Understandized** under both **Predicted Values** and **Residuals**.



In R

1. Add another variable in the model by simply adding it!
2. Add the residuals and predicted values to the dataset

```
model2<-lm(y~x+z, data=df)
df$resid<-model2$residuals
df$pred<-mode2$fitted.values
View(df)
```

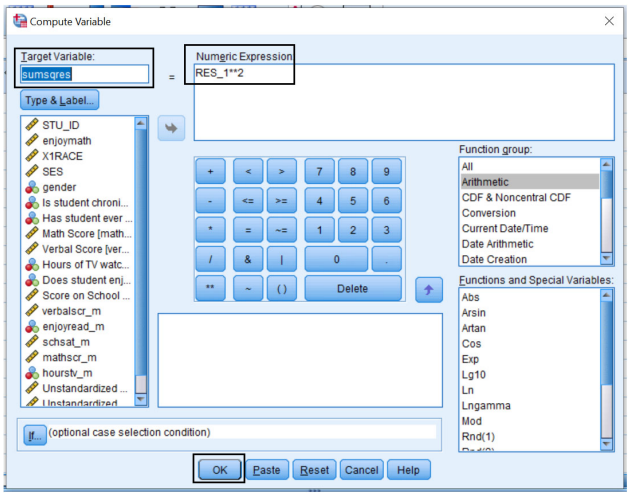
Computing the residual sum of squares

(**Answer quiz questions 10 and 11**)

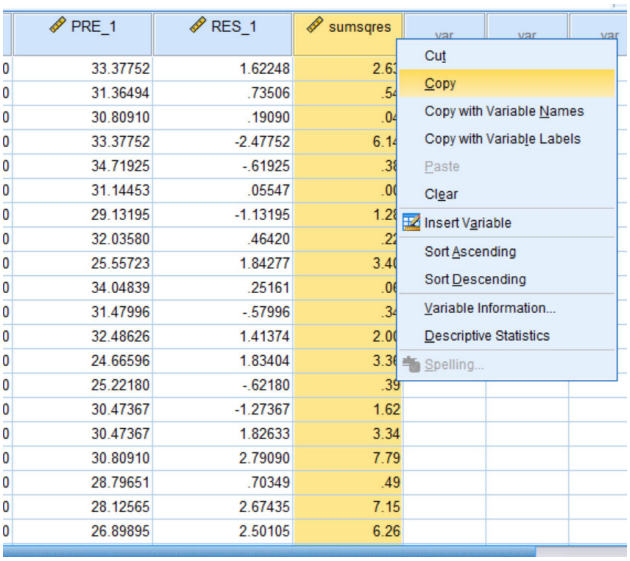
We have to do half of this in SPSS (1) and half in Excel (2), unfortunately.

1. Compute the square of the residuals

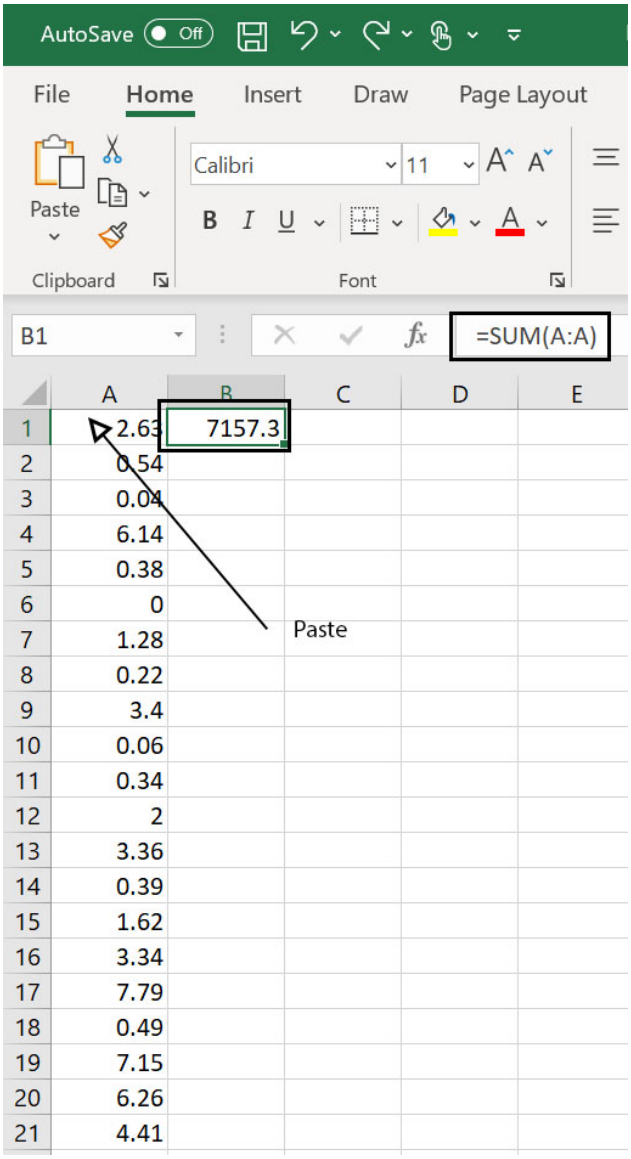
- Select **Transform** > **Compute Variable**
- Under **Target Variable** enter the name of the new variable you want to compute. We are going to compute the square of the residuals from the model, so we will call it **sumsqres** but you can call it whatever you want
- Under **Numeric expression** enter the equation for the square of the residuals: RES_1**2
- RES_1 is the residual variable we just created _ **2 is how we tell SPSS to square it



2. Copy the new variable you just created (**sumsqres**) by right clicking on the column and selecting **copy**.



3. Open Excel.
4. Right click on cell **A1** and press **Paste**.
5. In cell **B1**, type **=SUM(A:A)** to sum the entire column.



In R

Much simpler!

```
residsq<-(model2$residuals)^2
sum(residsq)
```