214B: Lab 1

Correlation and Regression

TA: Melissa Gordon Wolf

Research Question: A researcher at a school district wants to know about the relationship between student test scores and the percentage of English language learners within the district.

Testable Hypothesis: Does the percentage of English language learners within a district predict test scores?

Step 1: Label the data.

Open the SPSS file from Gauchospace and use the codebook below to add the variable labels. *Hint: Do this in the Variable View tab.*

Variable	Label	Type
ID	School ID indicator	Nominal
testscr	Combined average test scores of math and English	Continuous
meal_pct	Percent qualifying for free and reduced priced lunch	Continuous
comp_stu	Computers per students	Continuous
expn_stu	EXPENTITURES PER STUDENT ($'S)	Continuous
str	Student teach ratio (ENRL_TOT/TEACHERS)	Continuous
avginc	District average income (in units of thousands)	Continuous
el_pct	Percent of English Language Learners	Continuous
completeA_G	Percentage of Students who complete A-G UC requirement	Continuous

In R:

1. First, import the dataset using `haven`.
2. Use the `var_label` command from the `labelled` package to add variable labels. *Hint below*

```
labelled::var_label(lab1data)<-list
(ID="School ID indicator",
testscr="Combined average test scores of math and English" ...)
```

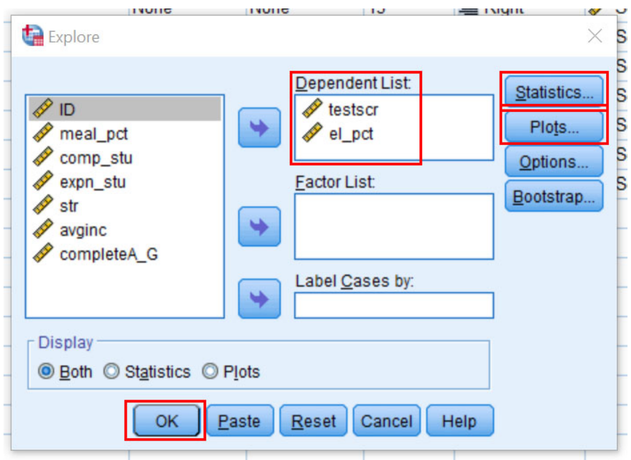
Step 2: Univariate Statistics

Answer questions 1 and 2 on the Gauchospace Quiz.

Let's generate the univariate statistics for the two variables of interest: **testscr** (*Combined average test scores of math and English*) and **el_pct** (*Percent of English Language Learners*). Both of these variables are continuous, so we'll get descriptive statistics such as the mean and skew of each variable.

1. Select **Analyze** > **Descriptive Statistics** > **Explore**
2. Move **testscr** and **el_pct** into the **Dependent List**

3. Under **Statistics**, select **Descriptives** and **Percentiles**
4. Under **Plots**, only select **Histogram**
5. Press **OK**



In R

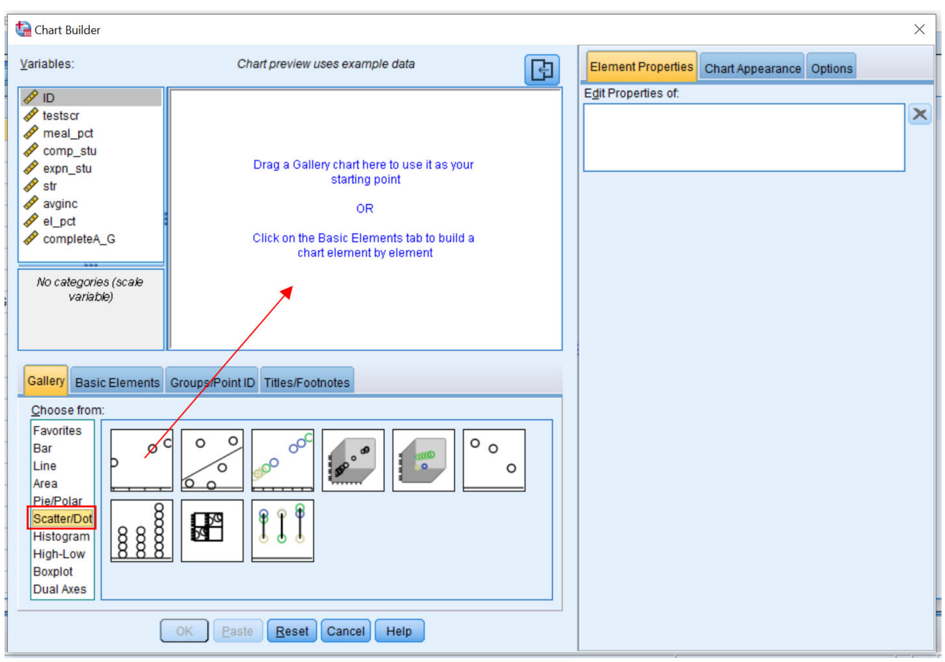
1. Use the `describe` command from the `psych` package. *Hint below*

```
psych::describe(lab1data$testscr)
```

Step 3a: Bivariate Relationships -> Graphing

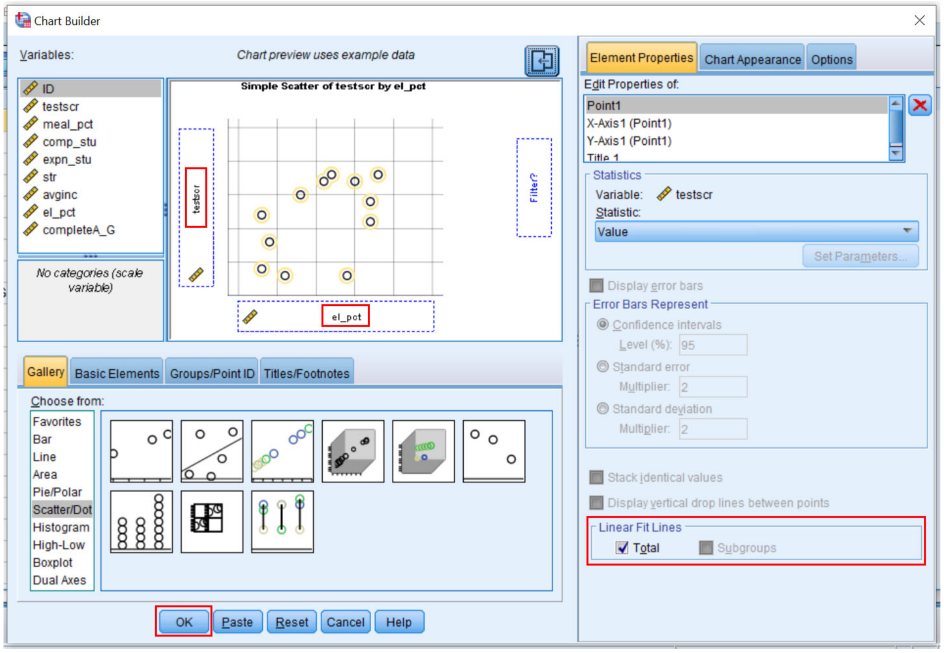
Let's graph the relationship between **testscr** and **el_pct** using a scatterplot.

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



3. Drag **testscr** onto the y-axis and **el_pct** onto the x-axis.

4. At the bottom of the **Element Properties** tab, select **Total** under **Linear Fit Lines**
5. Press **OK**



In R

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

```
ggplot(lab1data,aes(el_pct,testscr))+
  geom_point()+
  theme_minimal()
```

Step 3b: Bivariate Statistics -> Modeling

To evaluate the relationship between two continuous variables, we can use a correlation and a regression model. However, both of these models assume that there is a linear relationship between the two variables. Is there?

Correlation

Answer questions 3 and 4 on the Gauchospace Quiz.

1. Select **Analyze** > **Correlate** > **Bivariate**
2. Drag **testscr** and **el_pct** into the **Variables** box
3. Press **OK**

Correlations

		testscr	el_pct
testscr	Pearson Correlation	1	-.644**
	Sig. (2-tailed)		.000
	N	420	420
el_pct	Pearson Correlation	-.644**	1
	Sig. (2-tailed)	.000	
	N	420	420

**. Correlation is significant at the 0.01 level (2-tailed).

Interpreting Correlations

Rubric for Basic Correlation Analysis Write-up	
Conclusion Sentence	-negative or positive correlation? -between which variables?
Report Numerical Evidence	-correlation coefficient +/- .2 to +/- .29 = weak +/- .3 to +/- .39 = moderate +/- .4 to +/- .69 = strong +/- .7 to +/- 1 = very strong -*p*-value – statistically significant?
Interpret	-statements that either… …interpret your numerical evidence into words …interpret your conclusion without statistical language – for example, what does it mean that two variables are negatively correlated?

For more information: http://www.psychwiki.com/wiki/How_do_I_write_a_Results_section_for_Correlation%3F

In R

1. Use corr from base R. *Figure this out yourself*
2. Use rcorr from Hmisc package.

```
Hmisc::rcorr(lab1data$el_pct,lab1data$testscr)
```

Regression

Test of independence:

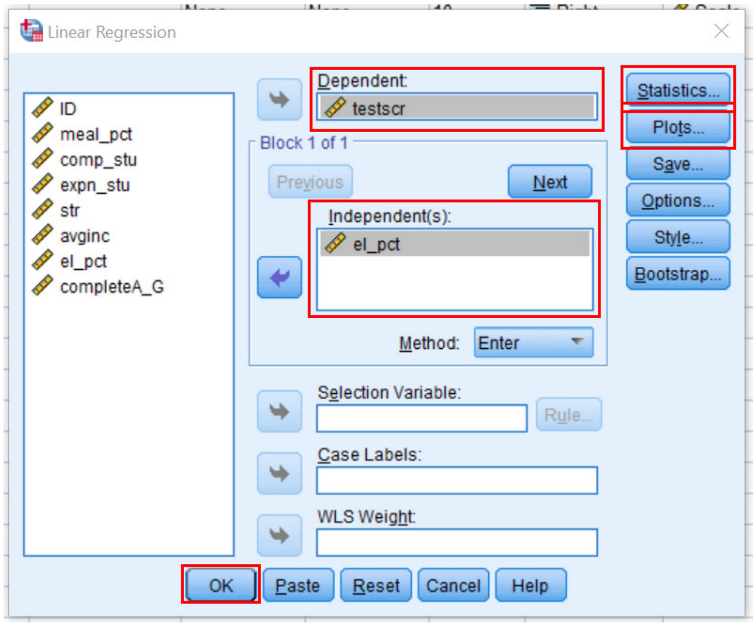
$H_0 : \beta = 0$

$H_A : \beta \neq 0$

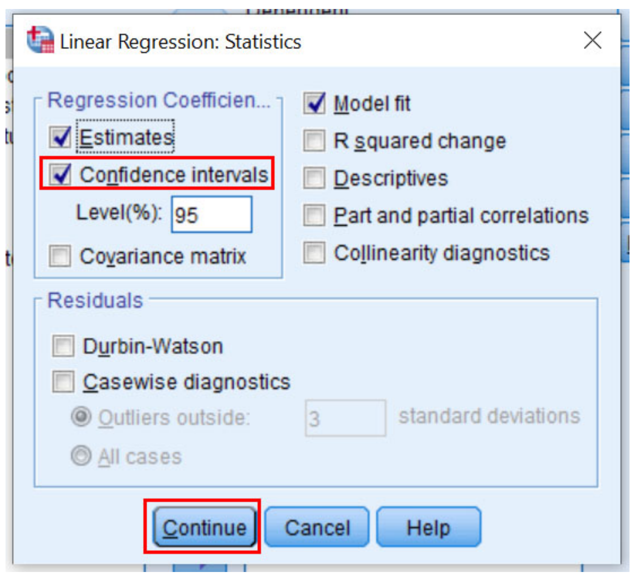
We want to know if there is a relationship between **el_pct** and **testscr**. We assume there is no relationship (H_0 or the null hypothesis). Note that β is the slope.

Answer questions 5, 6 and 7 on the Gauchospace Quiz.

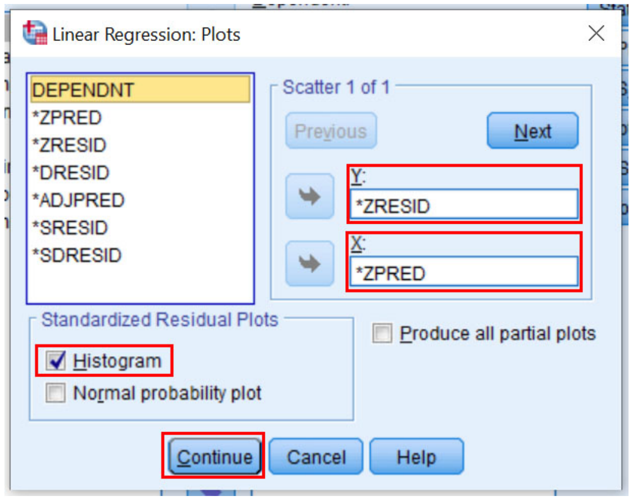
1. Select **Analyze** > **Regression** > **Linear**
2. Drag **testscr** into the **Dependent** box and **el_pct** into the **Independent(s)** box.



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

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.644[a]	.415	.413	14.59173

a. Predictors: (Constant), el_pct

b. Dependent Variable: testscr

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	63109.569	1	63109.569	296.402	.000[b]
	Residual	89000.025	418	212.919		
	Total	152109.59	419			

a. Dependent Variable: testscr

b. Predictors: (Constant), el_pct

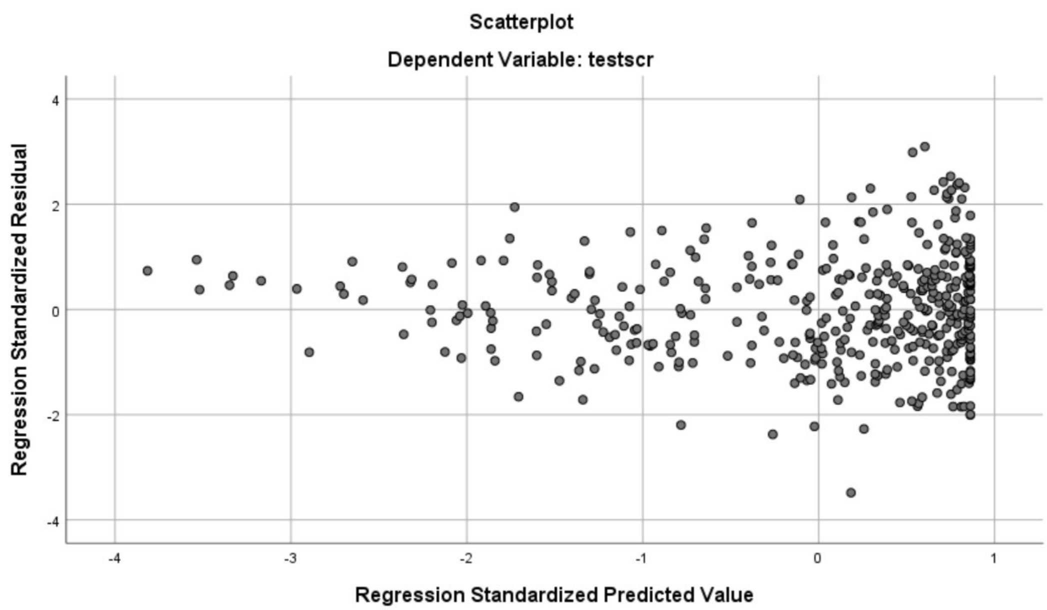
- **Coefficients table**
  - Note on the output that the slope (β) is -.671
  - Note on the output that the intercept is 664.74
  - In a simple linear regression model with one predictor, the *standardized* (β) is the same as the correlation coefficient.
    - A one standard deviation increase in el_pct is associated with a .644 *standard deviation* decrease in testscr.
    - Standardizing puts the two variables on the same scale
  - Note the 95% confidence intervals for the coefficients. Do they contain 0?
    - If the confidence interval contains 0, the effect is non-significant

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	664.739	.941		706.687	.000	662.890	666.588
	el_pct	-.671	.039	-.644	-17.216	.000	-.748	-.595

a. Dependent Variable: testscr

- **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

$$\hat{y} = \beta_0 + \beta_1 * x$$

$$\hat{y} = 664.74 - .671 * el - pct$$

In R

1. Use base R to run the model

```
model<-lm(testscr~el_pct,data=lab1data)
summary(model)
anova(model)
```

Advanced: Pretty output created using the stargazer package

```r
stargazer(model)
```

Table 1:

	Dependent variable:
	testscr
el_pct	-0.671^{***}
	(0.039)
Constant	664.739^{***}
	(0.941)
Observations	420
R^2	0.415
Adjusted R^2	0.413
Residual Std. Error	14.592 (df = 418)
F Statistic	296.402^{***} (df = 1; 418)
Note:	*p<0.1; **p<0.05; ***p<0.01

2. Plot the residuals on a scatterplot using base R

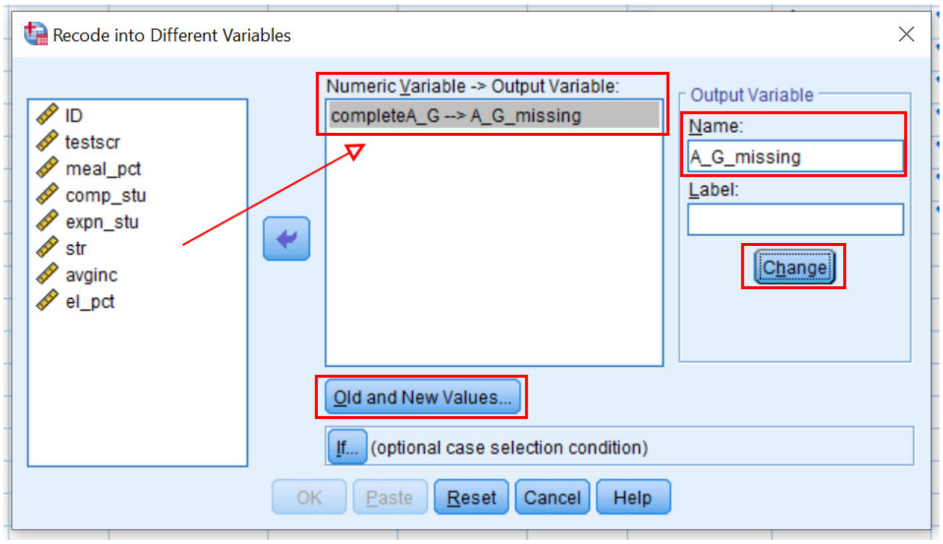
Graph will be reversed from SPSS but will contain same info

```r
plot(model)
```

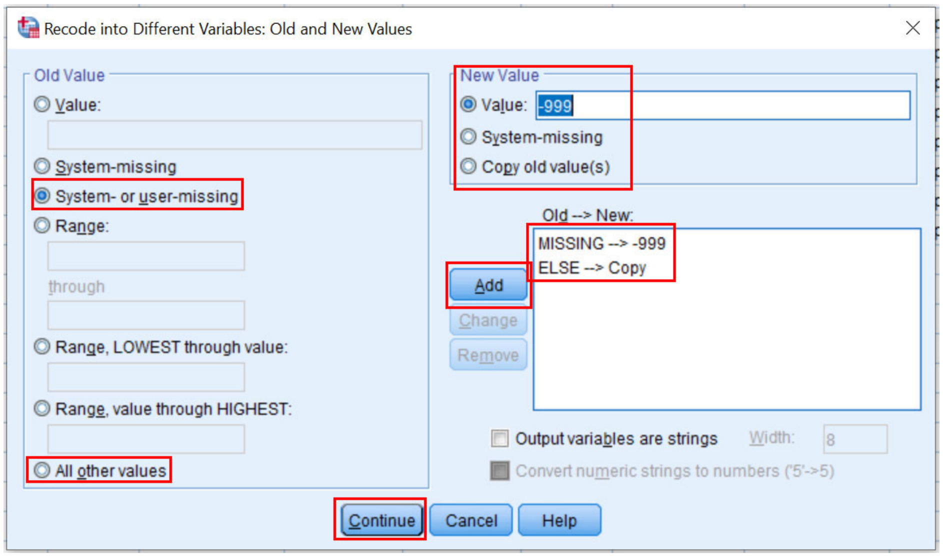
Variable recode (HW)

Notice that the variable **completeA_G** has some missing values. Let's recode those missing values to -999.

1. Select **Transform > Recode into different variables**
2. Drag **completeA_G** into the **Variable** box.
3. Under **Output Variable**, type a new name. We'll use **A_G_missing** but you can use whatever you want.
4. Press **CHANGE!**
5. Select **Old and New Values**



6. Under **Old Value**, select **System or user missing**
7. Under **New Value**, type **-999**
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 a new variable called **A_G_missing** where all missing values are coded as -999.
15. Select the **Variable View** tab.
16. Go down to the row that contains the variable **A_G_missing** and select the cell in the **Missing** column. Click on the "…" on the right side of the cell.
17. Enter **-999** under **Discrete Missing Values** and select **OK**. You have now told SPSS that all -999 values are missing values.

HW HINTS:

- **Important**: Make sure to **USE** the new variable you created in subsequent analyses.
- **Important**: In the HW, we are recoding numeric values into missing (opposite of above). Also, the missing value in the HW is not -999. Look at the **Missing** column of the **Variable View** to figure out what the value is.

In R

1. Use the `sjmisc` package

- *Make sure to tell R to copy the other values using else=copy*
- *Note: This is just an exercise to copy what we're doing in SPSS. We would never do this in R. Always tell R that missing values are NA. For example, if you read data in from SPSS that had all missing values coded as -999, we'd use the recode command and write -999=NA*

```
lab1data$A_G_missing<-sjmisc::rec(lab1data$completeA_G,rec="NA=-999;else=copy")
```

General R hints

- To install packages:

```
install.packages("packagename")
```

- To load libraries:

```
library(packagename)
```

- To call a command from a specific package:

```
packagename::command
```

- Revisit "Intro to R" lab from Week 1 on gauchospace
- If you just run the code above, you won't really learn much. Force yourself to retype it and mess around with the code to learn more. Remember, you're learning a language!
- Google is your friend!
    - *Seriously, all I do is Google everything*
- We have a stat software support person (Adam Garber). Email him to schedule a meeting if you have more questions about R.