# Lab 2: Regression Continued

### $214\mathrm{B}$ 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**

Recode into Different Variables: Old and New Valu	Jes ×
Old Value	New Value
© <u>V</u> alue:	© Va <u>l</u> ue:
	© System-missing
© <u>S</u> ystem-missing	Opy old value(s)
◎ System- or <u>u</u> ser-missing	
© Ra <u>ng</u> e:	
through	ELSE -> Copy
© Range, LOWEST through value:	Remove
© Range, value through HIGHEST:	
	Output variables are strings Width: 8
All other values	Convert nu <u>m</u> eric strings to numbers ('5'->5)
	Tue Cancel Help

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!

### ${\rm In}\ {\rm R}$

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  $\mathbf{OK}$

### 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  $\beta$  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**.

tinear Regression: Plots	×				
DEPENDNT *ZPRED *ZRESID *DRESID *ADJPRED *SRESID *SDRESID	Scatter 1 of 1 Previous Next  Y: *ZRESID  X: *ZPRED				
Standardized Residual Plots					
Cancel Help					

#### You should see the following output:

- Model Summary
  - Gives us the "omnibus" or "overall" model results
  - Look for the correlation and  $\mathbb{R}^2$
- 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
    - $\ast\,$  I deally, 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**

$$\label{eq:prod} \begin{split} \hat{y} &= \beta_0 + \beta_1 * x \\ \hat{y} &= 20.76 + 2.77 * enjoyread - m \end{split}$$

#### ${\rm In}~{\rm R}$

1. Use base R to run the model

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

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		
$\mathbb{R}^2$	0.579		
Adjusted $\mathbb{R}^2$	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

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.
 Under Save, select Understandized under both Predicted Values and Residuals.

Linear Regression: Save	X				
Predicted Values Unstandardized Standardized Adjusted S.E. of mean gredictions	Residuals         ✓ Unstandardized         Standardized         Studentized         Deleted         Studentized deleted				
Distances Maḫalanobis Cook's Leverage values Prediction Intervals Mean Individual Confidence Interval: 95 %	Influence Statistics DfBeta(s) Standardized DfBeta(s) DfFit Standardized DfFit Covariance ratio				
Coefficient statistics Create coefficient statistics Create a new dataset Dataset name: Write a new data file File					
Export model information to XML file Browse Include the covariance matrix					
Continue Cancel Help					

#### 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)</pre>
```

#### 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

Compute Variable		×
Target Variable: Sumscrees Type & LabeL STU_ID enjoymath X XTRACE	= Numgric Expression RES_1**2	Function group:
SES gender ls student chroni Has student ever Vath Score (math Verbal Score (ver Hours of TV watc	+ < > 7 8 9 - <= >= 4 5 6 + = ~= 1 2 3 / &   0 .	All Arithmetic CDF & Noncentral CDF Conversion Current Date/Time Date Arithmetic Date Creation
Does student enj     Score on School     Verbalsor_m     enjoyread_m     schsal_m     mathsor_m     hourstv_m     Unstandardized     Unstandardized	••• •• () Delete 🗲	Eunctions and Special Variables: Abs Arsin Artan Cos Exp Lg10 Ln Lnamma
(optional case selectio	n condition)	Mod Rnd(1)

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

	PRE_1	RES_1	🖋 sumsqres	var	var	var
0	33.37752	1.62248	2.63	Cuţ		
0	31.36494	.73506	.54	<u>C</u> opy		
0	30.80910	.19090	.04	Copy with Variable <u>N</u> ames		nes
0	33.37752	-2.47752	6.14	Copy with Variable Labels		
0	34.71925	61925	.38	Paste		
0	31.14453	.05547	.00	Clear		
0	29.13195	-1.13195	1.28	- Insert Vari	iahle	
0	32.03580	.46420	.22	Sort Ascending		
0	25.55723	1.84277	3.4(			
0	34.04839	.25161	.00	Sort Descending		
0	31.47996	57996	.34	Variable Information		
0	32.48626	1.41374	2.00	Descriptive Statistics		
0	24.66596	1.83404	3.30	Spelling		
0	25.22180	62180	.39			
0	30.47367	-1.27367	1.62			
0	30.47367	1.82633	3.34			
0	30.80910	2.79090	7.79			
0	28.79651	.70349	.49			
0	28.12565	2.67435	7.15			
0	26.89895	2.50105	6.26			

- 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.

AutoSave 💽 🕞 🤟 🗸 🤆 👻 🗸 🗸					
Fi	le <b>Hon</b>	ne Inse	ert Dra	w Page	Layout
Pa	ste 💞	Calibri B I	U v III Font	√ <u>11</u> ~ A^ ~ ∕⁄ ~ <u>A</u>	A <sup>*</sup> ≡
B1 $\cdot$ : $\times$ $f_x$ =SUM(A:A)					IM(A:A)
	A	B	с	D	Е
1	₽2.63	7157.3			
2	0,54				
3	0.04				
4	6.14				
5	0.38				
6	0				
7	1.28	`	Paste		
8	0.22				
9	3.4				
10	0.06				
11	0.34			_	
12	2				
13	3.36				
14	0.39				
15	1.62				
16	3.34				
17	7.79				
18	0.49				
19	7.15				
20	6.26				
21	4.41				

# In R

Much simpler!

residsq<-(model2\$residuals)^2
sum(residsq)</pre>