Statistics 21c_SPSS.pdf	Michael Hallstone, Ph.D. hallston@hawaii.edu

Lecture 21c: Using SPSS for Regression and Correlation

The purpose of this lecture is to illustrate the how to create SPSS output for correlation and regression.

You will notice that this document follows the order of the test questions for regression and correlation on the take home exam.

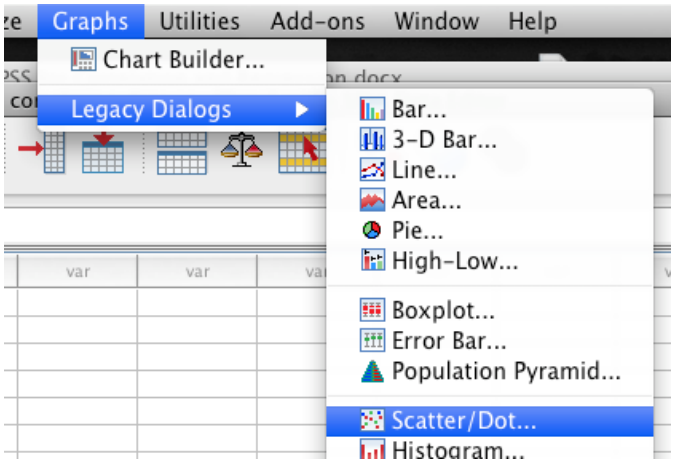
Regression

Recall simple regression/correlation require two interval/ratio level variables. Thus you will need two of these

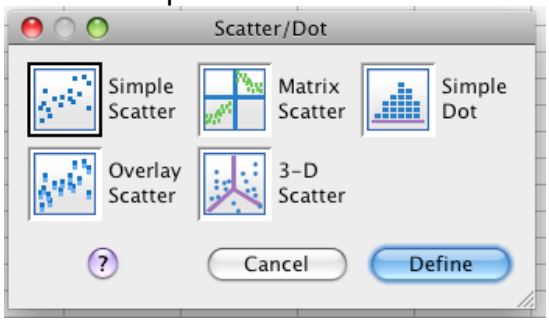
We will use the data from lecture 21. I have two ratio-level variables: **waveht** = *wave ht. in feet at Waimea Bay* & **surfers** = *# of surfers in the water at Waimea Bay.* Pretend the data comes from randomly selecting observations from the lifeguards for 7 days in January and February of this winter (n=7).

Wave Ht.	# of surfers
2	0
5	2
8	10
12	30
15	40
18	45
20	60

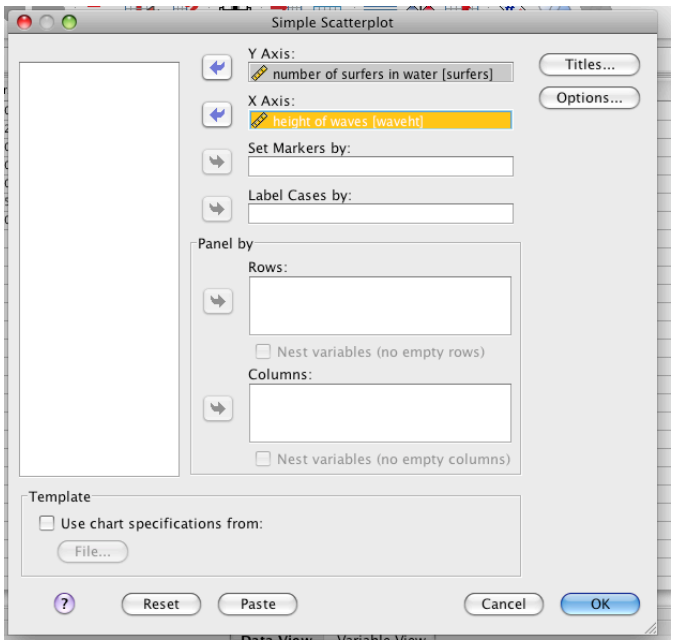
How to have SPSS do a scatter plot



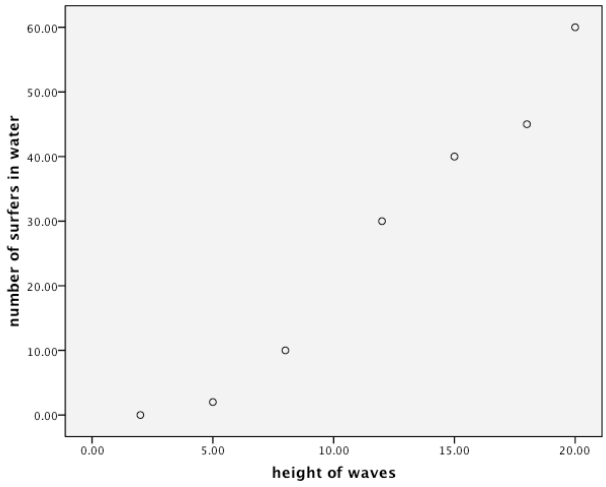
Choose Simple Scatter and select **Define.**



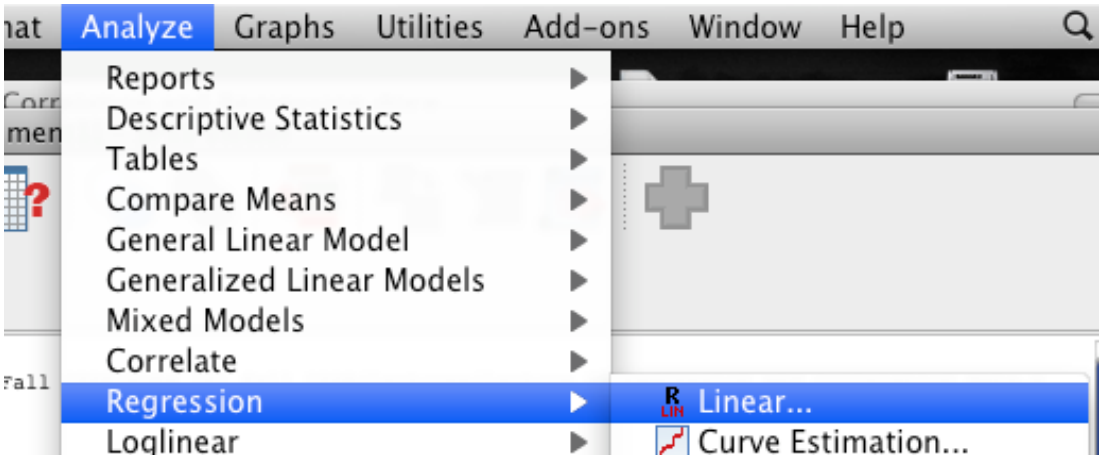
It is VERY important that you do not "mix" up your two variables in this screen! Recall that the dependent variable (y) is the variable you wish to predict and you will predict it using the independent variable (x). Use the arrows to put your dependent variable (y) [ours is number of surfers in the water] into the y axis box and the independent variable (x) [ours is height of waves] into the x axis box and push **OK.**



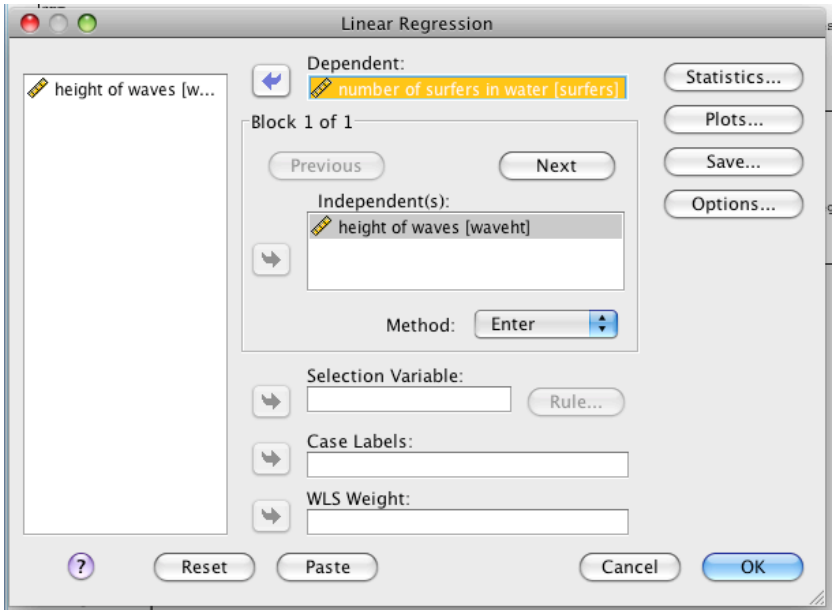
Below is the output of the scatter plot



How to have SPSS produce simple linear regression output



Again is VERY important that you do not "mix" up your two variables in this screen! Move your dependent variable (y) into the **Dependent** box and your independent variable (x) into the **Independent** box and push **OK.** Recall our dependent variable is *number of surfers* and our independent variable is *height of waves*.



Below is the full output for regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	height of waves[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: number of surfers in water

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.985[a]	.969	.963	4.45003

a. Predictors: (Constant), height of waves

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3134.415	1	3134.415	158.282	.000[a]
	Residual	99.014	5	19.803		
	Total	3233.429	6			

a. Predictors: (Constant), height of waves

b. Dependent Variable: number of surfers in water

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-12.102	3.514		-3.444	.018
	height of waves	3.396	.270	.985	12.581	.000

a. Dependent Variable: number of surfers in water

TR value for step 6 and 7

The TR value is lower number under the "T" column: 12.581. The slope (b)= 3.396 and is found as described above. S_b is the lower number under the "Std. Error" column: .270

•

TR= $\underline{b\text{-}B_{Ho}}$ =	$\underline{3.396 - O}$=	$\underline{3.396}$=	**12.58**
S_b	0.270	0.270	

p-value for step 7

We find the p-value as the lower number under the "Sig." column. If it is less than our level of significance (α) it is significant. If it is greater than our α it IS NOT significant. (Our p-value is not =.000! SPSS ran out of decimal points!) Our p value <.001 which is less than .05 and is SIGNIFICANT!!

Regression line or "y-hat line" equation

The values for regression line are also found in the very bottom box of the regression output. The value next to "constant" is the y-intercept and the slope is directly below that. I have highlighted each in blue like this. So for

y-hat=mx+b where m = slope and b = y intercept

y-hat= 3.396X +(-12.102)

now because "plus a minus" is actually a minus the correct equation for the line is

y-hat=3.396X -12.102

Finding r^2

Look at the second box "Model Summary." For this example our r^2 or "coefficient of determination" is under the R square column and r^2 = .969 (0.969x100=96.9 or 96.9%). *Finally we can explain the r^2 value in English -- relating it to our variables!!!* In this case we would say that *"96.9% of the total variation in # of surfers in the water at Waimea Bay is explained by variations in wave height."*

By the way you can also find the "r" from the SPSS output above; it is the *r* or "coefficient of correlation." In this case *r* = .985 -- a very strong positive relationship.

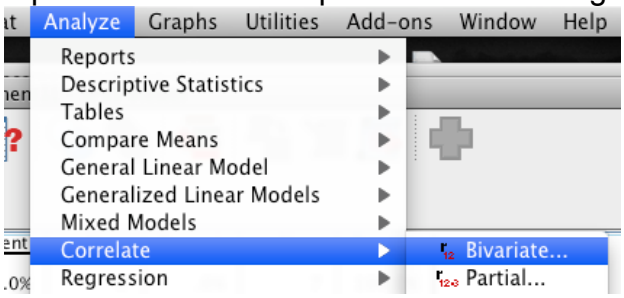
What a correlation tests in plain English

Recall that correlations test for "association" between two variables. As noted on the first page of the main lecture, "In correlation we see how closely the two variables under examination are associated. All "association" means is that changes in one variable are associated with changes in another. We might expect that "ability to drive a car" is associated with the number of drinks one has. A change in one variable is associated with a change in another variable."

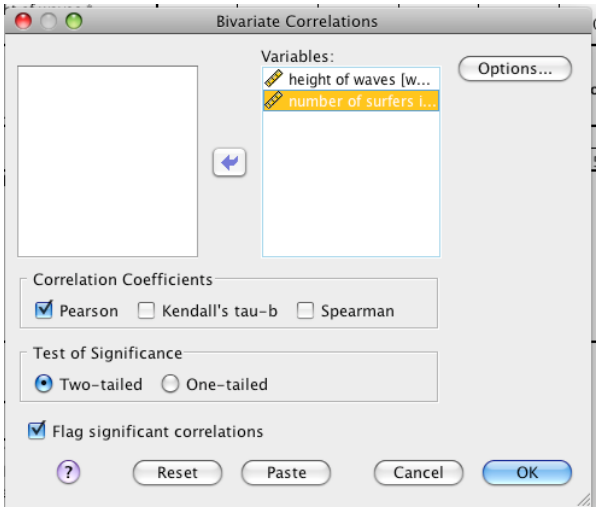
So in this case we would be seeing if changes in the height of waves at Waimea Bay are associated with changes in the number of surfers in the water at Waimea Bay

To have SPSS create correlation output

While you can find all of this information in the SPSS regression output above, SPSS also has separate correlation output. Here is how to get it all in one place:



Move both of your variables over into the Variables box the push OK.



Below is the full SPSS output for a regression analysis:

Correlations

		height of waves	number of surfers in water
height of waves	Pearson Correlation	1	.985**
	Sig. (2-tailed)		.000
	N	7	7
number of surfers in water	Pearson Correlation	.985**	1
	Sig. (2-tailed)	.000	
	N	7	7

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

The "strength" of r actually depends upon previous research, so I will give you a contrived or "pretend" rule that we will use for this class only. ***But remember for the r or "coefficient of correlation" to be a meaningful number we must have a "significant correlation" (p<.05 or p< your alpha from step 2)!***

Recall the "sign" of r matters!

- 0.00< r <0.33 = "weak relationship"(the mathematical sign (+ or -) determines whether or not it's a positive or negative relationship)
- 0.34< r <0.66 = "moderate relationship"(the mathematical sign (+ or -) determines whether or not it's a positive or negative relationship)
- 0.67< r <1.0 = "strong relationship"(the mathematical sign (+ or -) determines whether or not it's a positive or negative relationship)

The "R" from the SPSS output above is the *r* or "coefficient of correlation." In this case *r* = .985. We can see that the R= .985 again and most importantly we can see the p-value for the correlation test in the second "Sig. (2-tailed)" box. (Our p-value is not =.000! SPSS ran out of decimal points!) Our p value <.001 which is less than .05 and is SIGNIFICANT!! and .98457 is very close to a perfect correlation of +1 we can say represents "a very strong positive relationship."