Correlation (Pearson & Spearman) & Linear Regression

Azmi Mohd Tamil

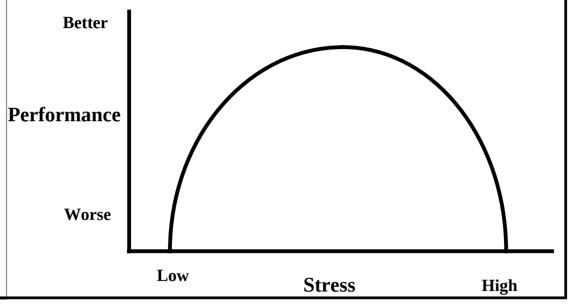
Key Concepts

- Correlation as a statistic
- Positive and Negative Bivariate Correlation
- Range Effects
- Outliers
- Regression & Prediction
- Directionality Problem (& cross-lagged panel)
- Third Variable Problem (& partial correlation)

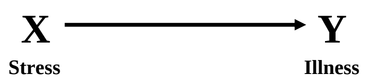
Assumptions

- Related pairs
- Scale of measurement. For Pearson, data should be interval or ratio in nature.
- Normality
- Linearity
- Homocedasticity

Example of Non-Linear Relationship Yerkes-Dodson Law – not for correlation



Correlation



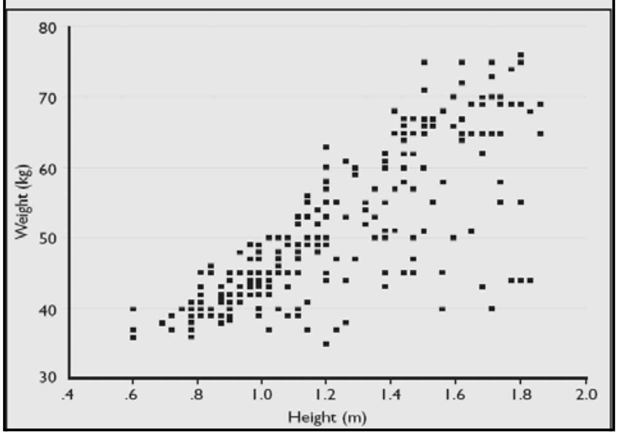
Correlation – parametric & non-para

- **2 Continuous Variables - Pearson**
  - **linear relationship**
  - **e.g., association between height and weight**
- **1 Continuous, 1 Categorical Variable (Ordinal) Spearman/Kendall**
  - **–e.g., association between Likert Scale on work satisfaction and work output**
  - **–pain intensity (no, mild, moderate, severe) and dosage of pethidine**

Pearson Correlation

- **2 Continuous Variables**
  - **linear relationship**
  - **e.g., association between height and weight, +**
- measures the degree of linear association between two interval scaled variables
- analysis of the relationship between two quantitative outcomes, e.g., height and weight,

Fig. 1 Relationship between height and weight.



How to calculate r?

$$r = \frac{\Sigma XY - \frac{\Sigma X \Sigma Y}{N}}{\sqrt{(\Sigma X^2 - \frac{(\Sigma X)^2}{N})\ (\Sigma Y^2 - \frac{(\Sigma Y)^2}{N})}}$$

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

Example

- $\sum$ x = 4631 $\sum$ x2 = 688837
- $\sum$ y = 2863 $\sum$ y2 = 264527
- $\sum$ xy = 424780 n = 32

- a=424780-(4631*2863/32)
  =10450.22
- b=688837-46312/32=18644.47
- c=264527-28632/32=8377.969
- r=a/(b*c)$^{0.5}$
  =10450.22/(18644.47*8377.969)$^{0.5}$
  =0.836144

- t= 0.836144*((32-2)/(1-0.8361442))$^{0.5}$
  t = 8.349436 & d.k. = 30,
  p < 0.001

nores	bps1	bpd1	x2	y2	xy
234	118	67	13924	4489	7906
235	126	76	15876	5776	9576
238	105	68	11025	4624	7140
240	112	71	12544	5041	7952
243	99	55	9801	3025	5445
244	99	66	9801	4356	6534
245	110	75	12100	5625	8250
274	133	85	17689	7225	11305
248	134	88	17956	7744	11792
253	129	83	16641	6889	10707
255	140	80	19600	6400	11200
256	117	72	13689	5184	8424
259	137	86	18769	7396	11782
231	164	95	26896	9025	15580
232	164	94	26896	8836	15416
233	164	89	26896	7921	14596
236	156	87	24336	7569	13572
237	147	103	21609	10609	15141
239	186	108	34596	11664	20088
241	170	102	28900	10404	17340
242	170	99	28900	9801	16830
246	176	121	30976	14641	21296
247	186	116	34596	13456	21576
249	157	107	24649	11449	16799
250	142	91	20164	8281	12922
251	159	85	25281	7225	13515
252	144	97	20736	9409	13968
254	155	113	24025	12769	17515
257	162	72	26244	5184	11664
258	151	98	22801	9604	14798
260	164	109	26896	11881	17876
261	155	105	24025	11025	16275
	4631	2863	688837	264527	424780

Interpreting Correlations

- Statistics
- Problems with causal interpretation

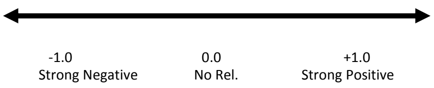
Correlation

Two pieces of information:

- The strength of the relationship
- The direction of the relationship

Strength of relationship

- r lies between -1 and 1. Values near 0 means no (linear) correlation and values near ± 1 means very strong correlation.



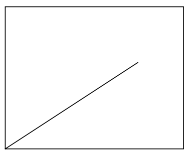
How to interpret the value of r?

Table II. Strength of linear relationship.

Correlation Coefficient value	Strength of linear relationship
At least 0.8	Very strong
0.6 up to 0.8	Moderately strong
0.3 to 0.5	Fair
Less than 0.3	Poor

Correlation (+ direction)

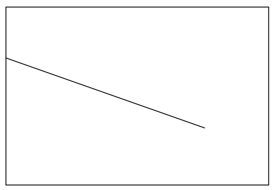
- Positive correlation: high values of one variable associated with high values of the other
- Example: Higher course entrance exam scores are associated with better course grades during the final exam.



Positive and Linear

Correlation (- direction)

- Negative correlation: The negative sign means that the two variables are inversely related, that is, as one variable increases the other variable decreases.
- Example: Increase in body mass index is associated with reduced effort tolerance.



Negative and Linear

Pearson's *r*

- A .9 is a very strong positive association (as one variable rises, so does the other)
- A -.9 is a very strong negative association (as one variable rises, the other falls)

r=.9 has nothing to do with 90%

r=correlation coefficient

Coefficient of Determination Defined

- Pearson's *r* can be squared , r^2, to derive a coefficient of determination.
- Coefficient of determination – the portion of variability in one of the variables that can be accounted for by variability in the second variable

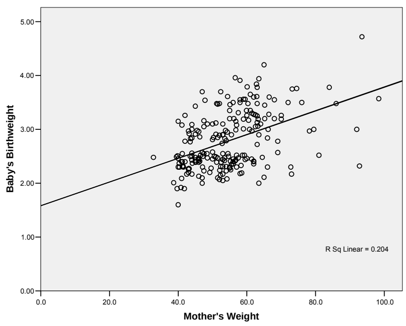
Coefficient of Determination

- Pearson's r can be squared , r^2, to derive a coefficient of determination.
- Example of depression and CGPA
  - Pearson's r shows negative correlation, r=-.5
  - r^2=.25
  - In this example we can say that 1/4 or .25 of the variability in CGPA scores can be accounted for by depression (remaining 75% of variability is other factors, habits, ability, motivation, courses studied, etc)

Coefficient of Determination and Pearson's *r*

- Pearson's r can be squared , r^2
- If r=.5, then r^2=.25
- If r=.7 then r^2=.49
- Thus while r=.5 versus .7 might not look so different in terms of strength, r^2 tells us that r=.7 accounts for about twice the variability relative to r=.5

A study was done to find the association between the mothers' weight and their babies' birth weight. The following is the scatter diagram showing the relationship between the two variables.



The coefficient of correlation (r) is 0.452

The coefficient of determination (r^2) is 0.204

Twenty percent of the variability of the babies' birth weight is determined by the variability of the mothers' weight.

Causal Silence: Correlation Does Not Imply Causality

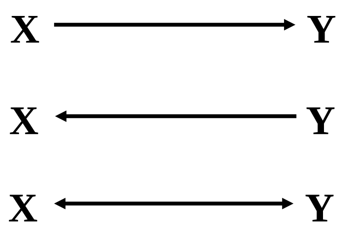
Causality – must demonstrate that variance in one variable can only be due to influence of the other variable

- **Directionality of Effect Problem**
- **Third Variable Problem**

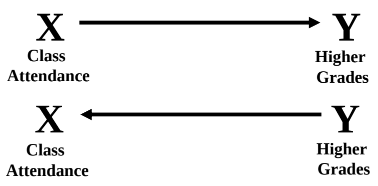
CORRELATION DOES NOT MEAN CAUSATION

- A high correlation **does not** give us the evidence to make a cause-and-effect statement.
- A common example given is the high correlation between the cost of damage in a fire and the number of firemen helping to put out the fire.
- Does it mean that to cut down the cost of damage, the fire department should dispatch less firemen for a fire rescue!
- The intensity of the fire that is highly correlated with the cost of damage and the number of firemen dispatched.
- The high correlation between smoking and lung cancer. However, one may argue that both could be caused by stress; and smoking does not cause lung cancer.
- In this case, a correlation between lung cancer and smoking may be a result of a cause-and-effect relationship (by clinical experience + common sense?). To establish this cause-and-effect relationship, controlled experiments should be performed.

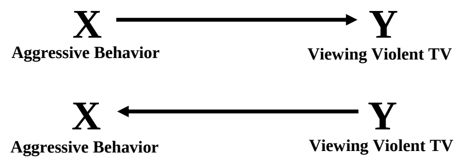
Directionality of Effect Problem



Directionality of Effect Problem



Directionality of Effect Problem



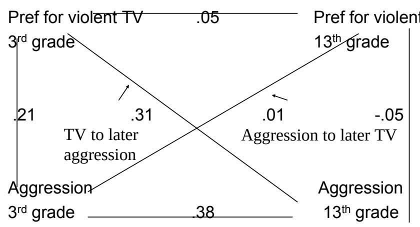
Aggressive children may prefer violent programs or
Violent programs may promote aggressive behavior

Methods for Dealing with Directionality

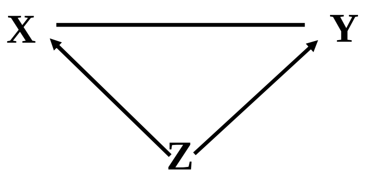
- Cross-Lagged Panel design
  - A type of longitudinal design
  - Investigate correlations at several points in time
  - STILL NOT CAUSAL

Example next page

Cross-Lagged Panel



Third Variable Problem



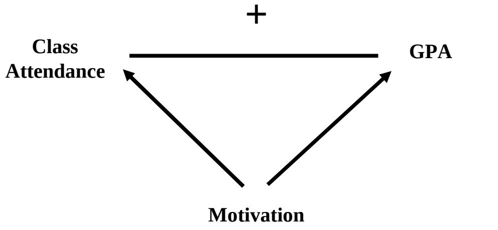
Class Exercise

Identify the

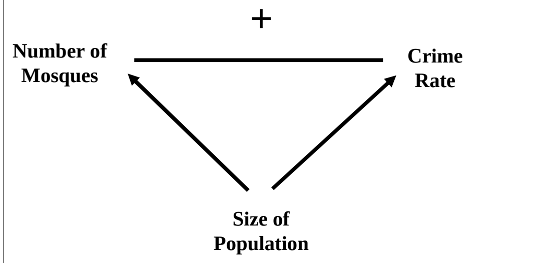
third variable

that influences both X and Y

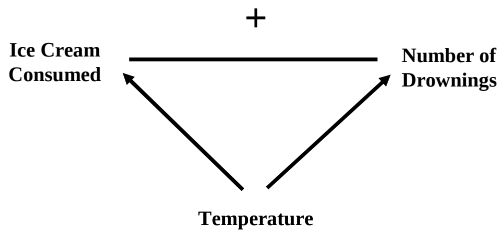
Third Variable Problem



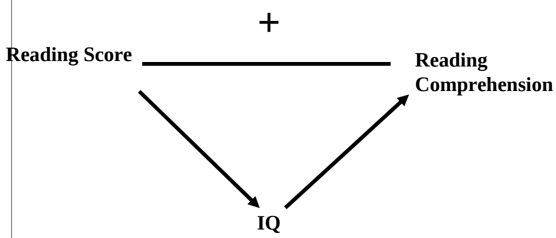
Third Variable Problem



Third Variable Problem



Third Variable Problem

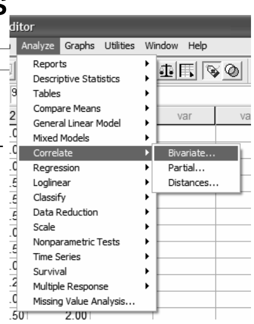


Data Preparation - Correlation

- Screen data for outliers and ensure that there is evidence of linear relationship, since correlation is a measure of linear relationship.
- Assumption is that each pair is bivariate normal.
- If not normal, then use Spearman.

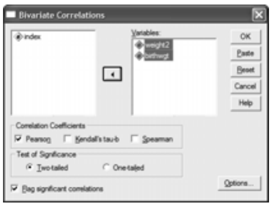
Correlation In SPSS

- For this exercise, we will be using the data from the CD, under Chapter 8, korelasi.sav
- This data is a subset of a case-control study on factors affecting SGA in Kelantan.
- Open the data & select - >Analyse
  >Correlate
  >Bivariate…



Correlation in SPSS

- We want to see whether there is any association between the mothers' weight and the babies'weight. So select the variables (weight2 & birthwgt) into 'Variables'.
- Select 'Pearson' Correlation Coefficients.
- Click the 'OK' button.



Correlation Results

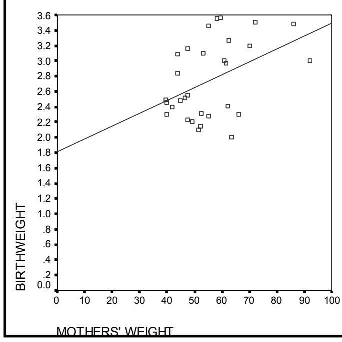
Correlations

		WEIGHT2	BIRTHWGT
WEIGHT2	Pearson Correlation	1	.431*
	Sig. (2-tailed)	.	.017
	N	30	30
BIRTHWGT	Pearson Correlation	.431*	1
	Sig. (2-tailed)	.017	.
	N	30	30

*. Correlation is significant at the 0.05 level (2-tailed).

- The r = 0.431 and the p value is significant at 0.017.
- The r value indicates a fair and positive linear relationship.

Scatter Diagram



Rsq = 0.1861

- If the correlation is significant, it is best to include the scatter diagram.
- The r square indicated mothers' weight contribute 19% of the variability of the babies' weight.

Spearman/Kendall Correlation

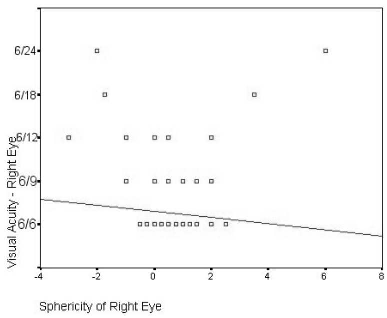
- To find correlation between a related pair of continuous data (not normally distributed); or
- **Between 1 Continuous, 1 Categorical Variable (Ordinal)**
  - **e.g., association between Likert Scale on work satisfaction and work output.**

Spearman's rank correlation coefficient

- In statistics, **Spearman's rank correlation coefficient**, named for Charles Spearman and often denoted by the Greek letter ρ (rho), is a non-parametric measure of correlation – that is, it assesses how well an arbitrary monotonic function could describe the relationship between two variables, without making any assumptions about the frequency distribution of the variables. Unlike the Pearson product-moment correlation coefficient, it does not require the assumption that the relationship between the variables is linear, nor does it require the variables to be measured on interval scales; it can be used for variables measured at the ordinal level.

Example

•Correlation between sphericity and visual acuity.
•Sphericity of the eyeball is continuous data while visual acuity is ordinal data (6/6, 6/9, 6/12, 6/18, 6/24), therefore Spearman correlation is the most suitable.
•The Spearman rho correlation coefficient is -0.108 and p is 0.117. P is larger than 0.05, therefore there is no significant association between sphericity and visual acuity.



Correlations

			Visual Acuity - Right Eye	Sphericity of Right Eye
Spearman's rho	Visual Acuity - Right Eye	Correlation Coefficient	1.000	-.108
		Sig. (2-tailed)	.	.117
		N	215	211
	Sphericity of Right Eye	Correlation Coefficient	-.108	1.000
		Sig. (2-tailed)	.117	.
		N	211	211

Example 2

•- Correlation between glucose level and systolic blood pressure.
•Based on the data given, prepare the following table;
•For every variable, sort the data by rank. For ties, take the average.
•Calculate the difference of rank, d for every pair and square it. Take the total.
•Include the value into the following formula;

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

•$\sum d^2 = 4921.5$ n = 32
•Therefore $r_s = 1-((6*4921.5)/(32*(32^2-1)))$
= 0.097966.
This is the value of Spearman correlation coefficient (or ϒ).
•Compare the value against the Spearman table;
•p is larger than 0.05.
•Therefore there is no association between systolic BP and blood glucose level.

nores	glu	rank x	bps1	rank y	d	d2
231	123	23	164	25.5	-2.5	6.25
232	97	9	164	25.5	-16.5	272.25
233	325	32	164	25.5	6.5	42.25
234	124	24	118	7	17	289
235	107	12.5	126	8	4.5	20.25
236	95.7	8	156	20	-12	144
237	122	22	147	16	6	36
238	112	17	105	3	14	196
239	119	20	186	31.5	-11.5	132.25
240	132	25	112	5	20	400
241	105	11	170	28.5	-17.5	306.25
242	219	30	170	28.5	1.5	2.25
243	141	26	99	1.5	24.5	600.25
244	93.6	4	99	1.5	2.5	6.25
245	206	29	110	4	25	625
246	113	18.5	176	30	-11.5	132.25
247	167	28	186	31.5	-3.5	12.25
248	95.6	7	134	11	-4	16
249	108	14.5	157	21	-6.5	42.25
250	297	31	142	14	17	289
251	109	16	159	22	-6	36
252	100	10	144	15	-5	25
253	83.3	2	129	9	-7	49
254	145	27	155	18.5	8.5	72.25
255	90.2	3	140	13	-10	100
256	113	18.5	117	6	12.5	156.25
257	108	14.5	162	23	-8.5	72.25
258	121	21	151	17	4	16
259	94.5	6	137	12	-6	36
260	69.4	1	164	25.5	-24.5	600.25
261	94.2	5	155	18.5	-13.5	182.25
274	107	12.5	133	10	2.5	6.25
						4921.5

Spearman's table

•0.097966 is the value of Spearman correlation coefficient (or ρ).
•Compare the value against the Spearman table;
•0.098 < 0.364 (p=0.05)
•p is larger than 0.05.
•Therefore there is no association between systolic BP and blood glucose level.

N (the number of pairs of scores):	*0.05*	*0.02*	*0.01*
5	1	1	
6	0.886	0.943	1
7	0.786	0.893	0.929
8	0.738	0.833	0.881
9	0.683	0.783	0.833
10	0.648	0.746	0.794
12	0.591	0.712	0.777
14	0.544	0.645	0.715
16	0.506	0.601	0.665
18	0.475	0.564	0.625
20	0.45	0.534	0.591
22	0.428	0.508	0.562
24	0.409	0.485	0.537
26	0.392	0.465	0.515
28	0.377	0.448	0.496
30	0.364	0.432	0.478

SPSS Output

Correlations

			GLU	BPS1
Spearman's rho	GLU	Correlation Coefficient	1.000	.097
		Sig. (2-tailed)	.	.599
		N	32	32
	BPS1	Correlation Coefficient	.097	1.000
		Sig. (2-tailed)	.599	.
		N	32	32

Linear Regression

Regression

X ⟶ **Y**

Predictor Variable **Criterion Variable**

Predicting Y based on a given value of X

Regression and Prediction

X ⟶ **Y**

Height **Weight**

REGRESSION

- Regression
  - one variable is a direct cause of the other
  - or if the value of one variable is changed, then as a direct consequence the other variable also change
  - or if the main purpose of the analysis is prediction of one variable from the other

REGRESSION

- Regression: looking for a dependence of one variable, the dependent variable on another, the independant variable
- relationship is summarised by a <u>regression equation.</u>
- y = a + bx

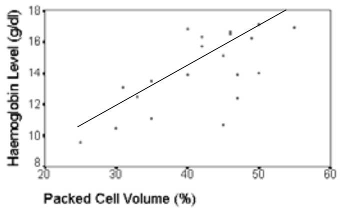
REGRESSION

- Regression
  - The regression line
    - x - independent variable - horizontal axis
    - y - dependent variable - vertical axis
  - Regression equation

    y = a + bx
    - a = intercept at y axis
    - b = regression coefficient
  - Test of significance - b is not equal to zero

REGRESSION

- Regression



Linear Regression

- Come up with a **Linear Regression Model** to predict a continous outcome with a continous risk factor, i.e. predict BP with age. Usually the next step after correlation is found to be strongly significant.
- y = a + bx
  - BP = regression estimate (b) * age + constant (a) + error term (å)
- b=

$$\frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}$$

Testing for significance

test whether the slope is significantly different from zero by:

t = b/SE(b)

$$SE_{(b)} = \frac{S_{res}}{\sqrt{\Sigma(x-\bar{x})^2}} \qquad S_{res} = \sqrt{\frac{\Sigma(y - y_{fit})^2}{n-2}}$$

$$\sqrt{((SD((y)^2(1-r^2)(n-1)/(n-2)}$$

Regression Line

- In a scatterplot showing the association between 2 variables, the regression line is the “best-fit” line and has the formula

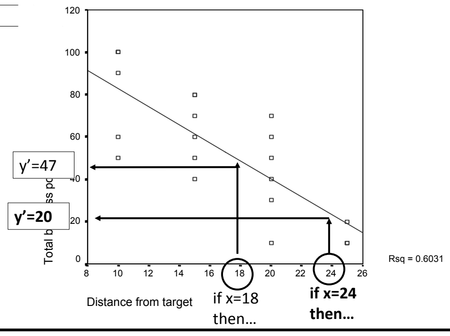
y=a + bX

a=place where line crosses Y axis

b=slope of line (rise/run)

Thus, given a value of X, we can predict a value of Y

Regression Graphic – Regression Line

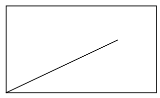


Regression Equation

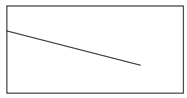
- y’= a + bx
  - y’ = predicted value of y
  - b = slope of the line
  - x = value of x that you enter
  - a = y-intercept (where line crosses y axis)
- In this case….
  - y’ = 125.401 - 4.263(x)
- So if the distance is **20** feet
  - y’ = -4.263(**20**) + 125.401
  - y’ = -85.26 + 125.401
  - **y’ = 40.141**

Regression Line (Defined)

Regression line is the line where absolute values of vertical distances between points on scatterplot and a line form a minimum sum (relative to other possible lines)



Positive and Linear



Negative and Linear

Example

∑x = 6426 ∑ x2 = 1338088
∑ y = 4631 ∑ xy = 929701
n = 32

b = (929701-(6426*4631/32))/(1338088-(64262/32)) = -0.00549

Mean x = 6426/32=200.8125
mean y = 4631/32=144.71875

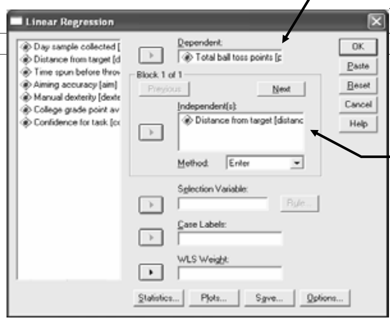
a = 144.71875+(0.00549*200.8125) = 145.8212106

Systolic BP = 144.71875 - 0.00549.chol

nores	chol	bps1	x2	y2	xy
234	162	118	26244	13924	19116
235	210	126	44100	15876	26460
238	239	105	57121	11025	25095
240	187	112	34969	12544	20944
243	181	99	32761	9801	17919
244	180	99	32400	9801	17820
245	156	110	24336	12100	17160
274	191	133	36481	17689	25403
248	203	134	41209	17956	27202
253	169	129	28561	16641	21801
255	221	140	48841	19600	30940
256	223	117	49729	13689	26091
259	269	137	72361	18769	36853
231	151	164	22801	26896	24764
232	151	164	22801	26896	24764
233	249	164	62001	26896	40836
236	206	156	42436	24336	32136
237	252	147	63504	21609	37044
239	219	186	47961	34596	40734
241	129	170	16641	28900	21930
242	150	170	22500	28900	25500
246	194	176	37636	30976	34144
247	164	186	26896	34596	30504
249	223	157	49729	24649	35011
250	264	142	69696	20164	37488
251	232	159	53824	25281	36888
252	165	144	27225	20736	23760
254	232	155	53824	24025	35960
257	286	162	81796	26244	46332
258	180	151	32400	22801	27180
260	198	164	39204	26896	32472
261	190	155	36100	24025	29450
	6426	4631	1338088	688837	929701

SPSS Regression Set-up

- “**Criterion**,”
- y-axis variable,
- what you’re trying to predict



- “**Predictor**,”
- x-axis variable,
- what you’re basing the prediction on

Getting Regression Info from SPSS

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.777[a]	.603	.581	18.476

a. Predictors: (Constant), Distance from target

y' = a + b (**x**)
y' = 125.401 - 4.263(**20**)

a

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	125.401	14.265		8.791	.000
	Distance from targ	-4.263	.815	-.777	-5.230	.000

a. Dependent Variable: Total ball toss points

b