1 Correlation of Variables

So far we've analyzed variables in a vacuum. We've computed summary statistics on a single variable with no regard to the effect one variable may have on another. Let's change that now. A simple example of two variables that are correlated are height and the ability to play in the NBA. The taller a person is the more likely they are to play basketball. This is not a perfect correlation! There are tall people who do not play basketball. There are (relatively speaking) short people who play professional basketball.

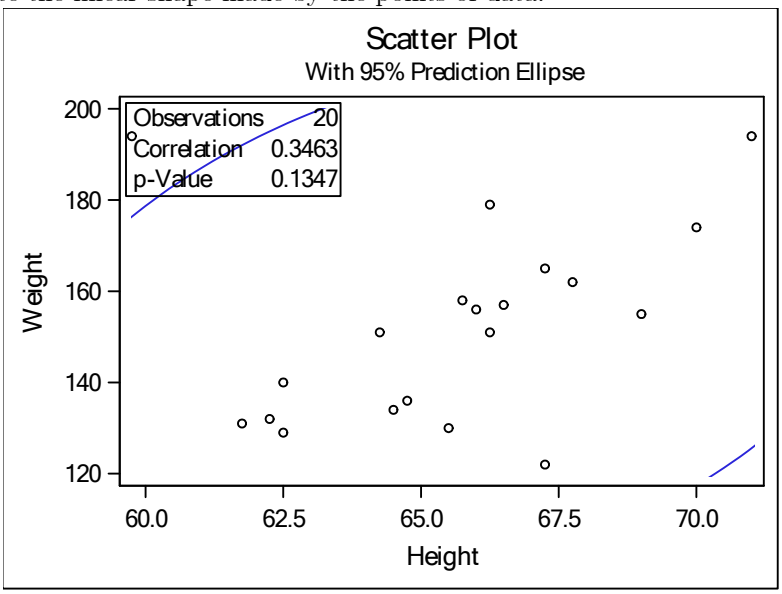
The 7'7" Manute Bol and 5'3" Muggsy Bogues

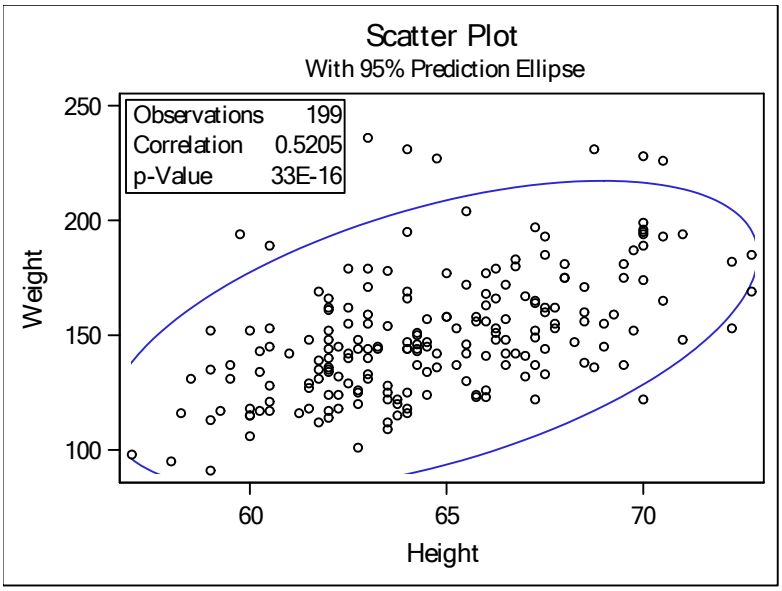
2 Scatterplots

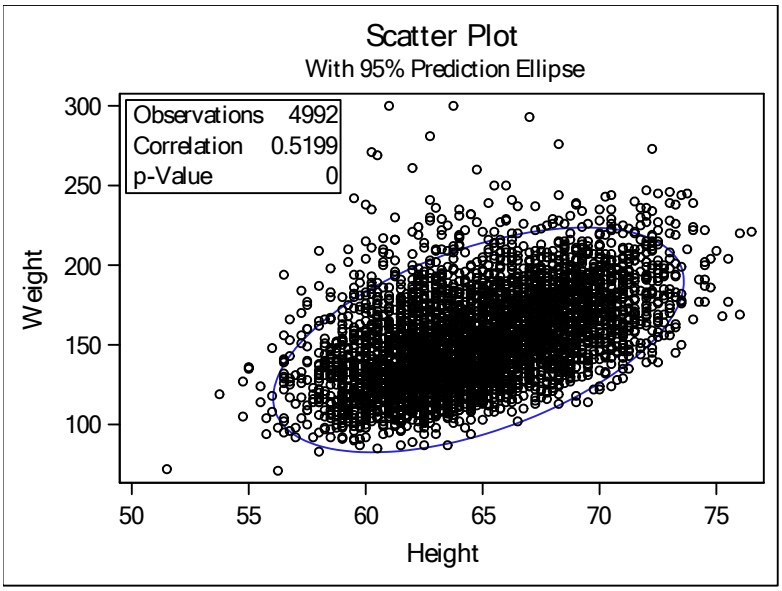
How do we look at data and determine if a correlation exists between a pair of variables? Recall out health data set.

Obs	Status	DeathCause	AgeCHDdiag	Sex	AgeAtStart	Height	Weight	Diastolic	Systolic	MRW
1	Dead	Other	.	Female	29	62.50	140	78	124	121
2	Dead	Cancer	.	Female	41	59.75	194	92	144	183
3	Alive		.	Female	57	62.25	132	90	170	114
4	Alive		.	Female	39	65.75	158	80	128	123
5	Alive		.	Male	42	66.00	156	76	110	116
6	Alive		.	Female	58	61.75	131	92	176	117
7	Alive		.	Female	36	64.75	136	80	112	110
8	Dead	Other	.	Male	53	65.50	130	80	114	99
9	Alive		.	Male	35	71.00	194	68	132	124
10	Dead	Cerebral Vascular Disease	.	Male	52	62.50	129	78	124	106

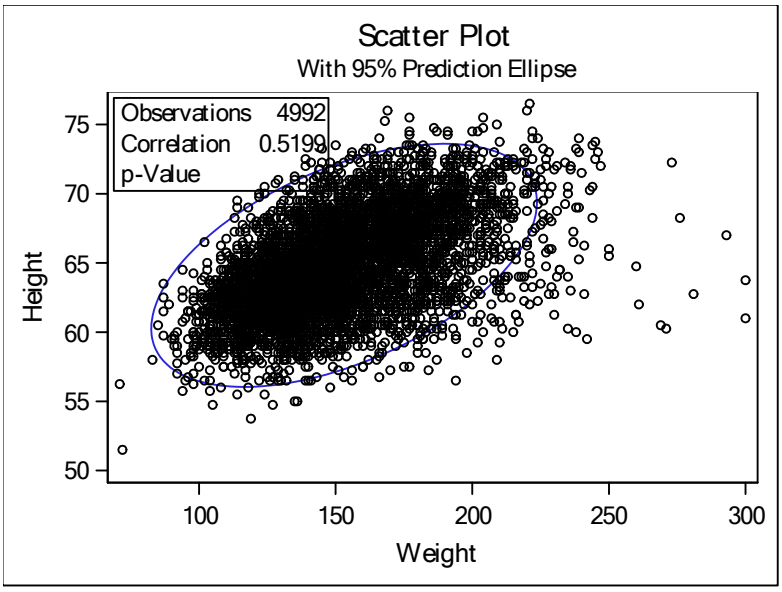
Does any correlation exist between the height and weight of a person in this data set? We begin to answer this question by creating a scatterplot of the data. Note the linear shape made by the points of data.



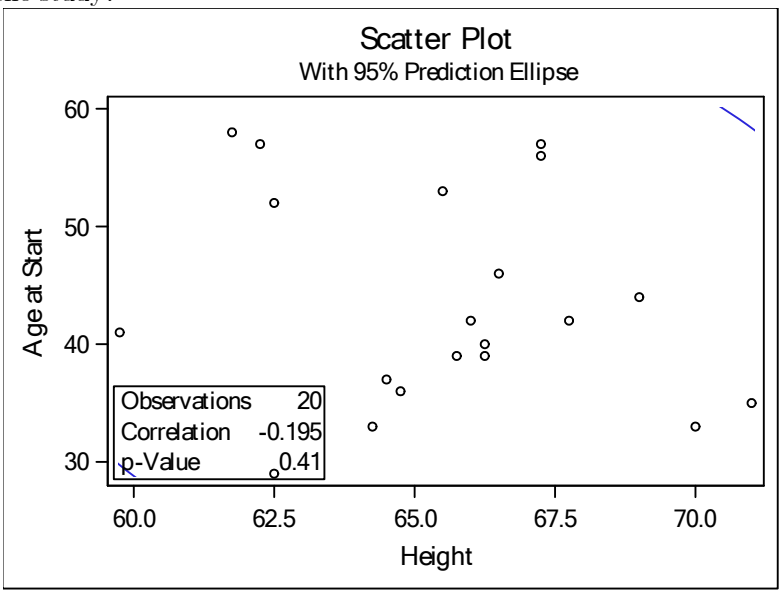


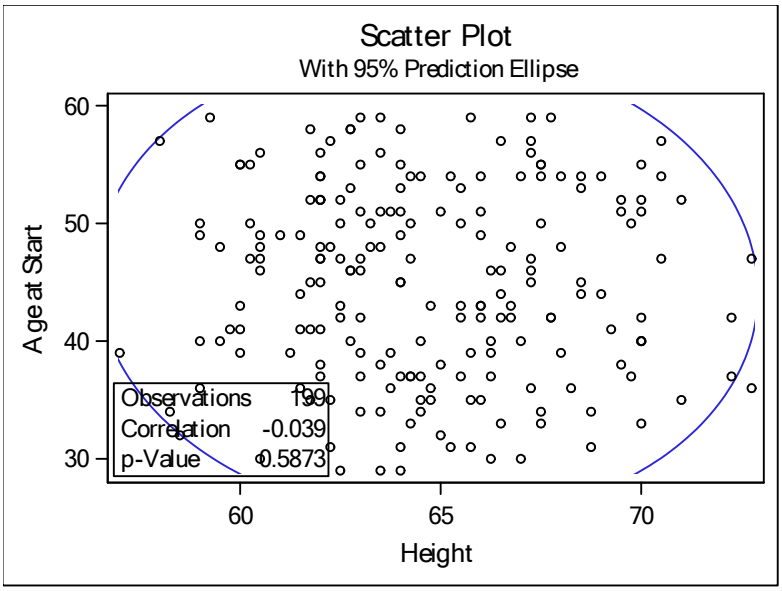


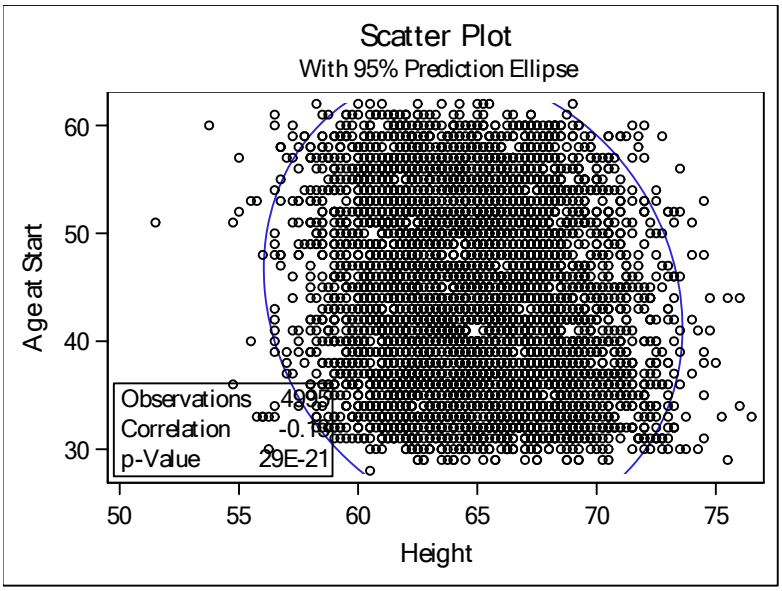
Also note that the order of the columns of data makes a difference.



Do we think a correlation should exist between height and age at the start of the study?



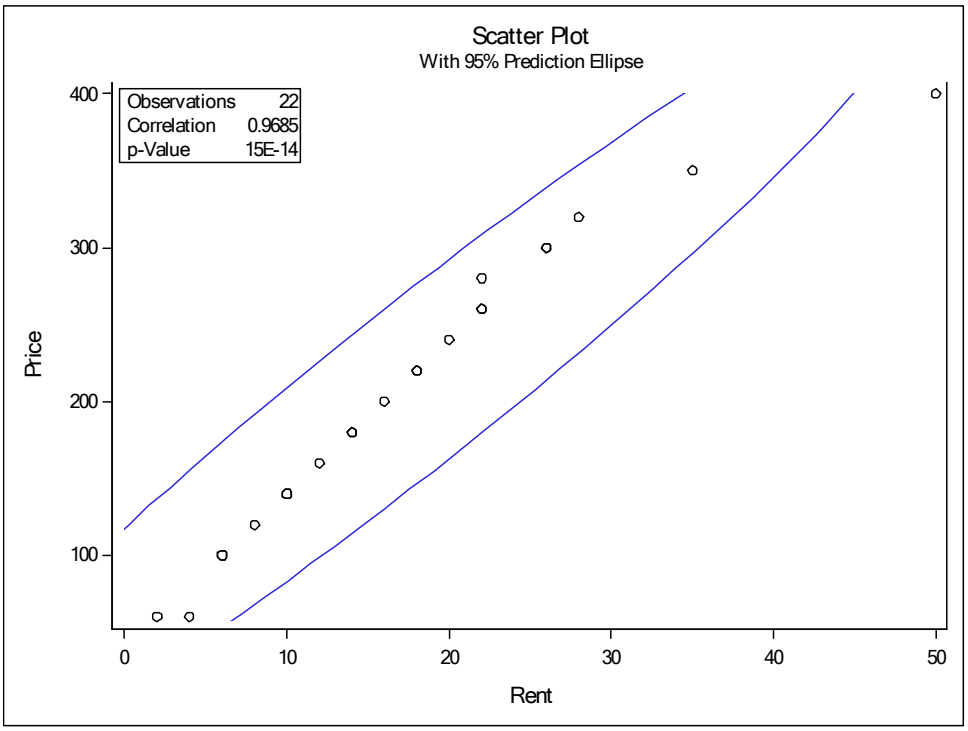




How do we create a scatterplot in the TI-84? Place the x-values (**explanatory**) in L_1 and the y-values (**response**) in L_2. Press **2nd**, **Y=** and select Plot 1. Select **On** and then the Scatterplot icon. Press **ZOOM** and **9**.

Exercise 1 *Use your TI-83/84 to create a scatterplot for property rent and property price in Monopoly.*

Obs	Name	Price	Rent	Group	E	F	G
1	Mediterranean Ave.	60	2	Purple			
2	Baltic Ave.	60	4	Purple			
3	Oriental Ave.	100	6	Light-Green			
4	Vermont Ave.	100	6	Light-Green			
5	Connecticut Ave.	120	8	Light-Green			
6	St. Charles Place	140	10	Violet			
7	States Ave.	140	10	Violet			
8	Virginia Ave.	160	12	Violet			
9	St. James Place	180	14	Orange			
10	Tennessee Ave.	180	14	Orange			
11	New York Ave.	200	16	Orange			
12	Kentucky Ave.	220	18	Red			
13	Indiana Ave.	220	18	Red			
14	Illinois Ave.	240	20	Red			
15	Atlantic Ave.	260	22	Yellow			
16	Ventnor Ave.	260	22	Yellow			
17	Marvin Gardens	280	22	Yellow			
18	Pacific Ave.	300	26	Dark-Green			
19	North Carolina Ave.	300	26	Dark-Green			
20	Pennsylvania Ave.	320	28	Dark-Green			
21	Park Place	350	35	Dark-Blue			
22	Boardwalk	400	50	Dark-Blue			

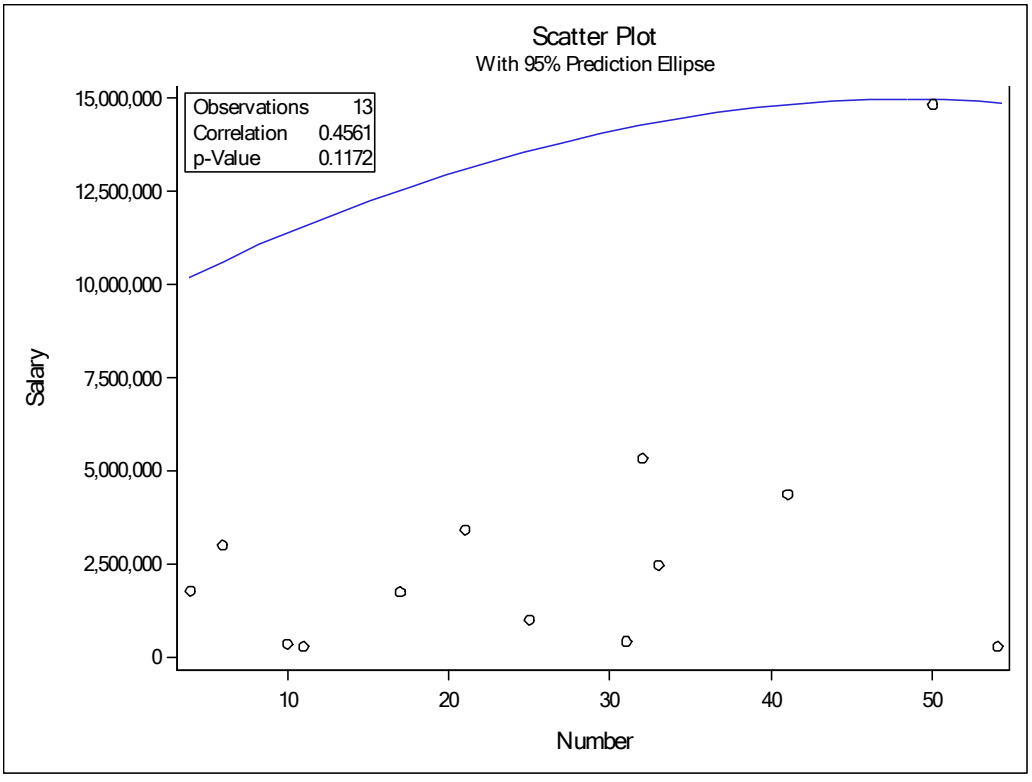


3 Correlation Coefficient

It would be nice if we had a numerical measure of the strength of the correlation between two variables. We do and it is called the **correlation coefficient,** denoted r. The correlation coefficient r is a number between -1 and 1 which indicates the strength and direction of a linear relation between two variables. The closer r is to 0 the less correlation there is between variables. The further away from 0 the stronger the correlation is. The sign of r indicates if the correlation is positive or negative. The correlation between height and weight is 0.5199 which is a medium positive correlation. The correlation between height and age at the start of the study is -0.129 which indicates little to no correlation.

Be careful with small data sets. They might indicate a correlation where none should exist. Odd things can happen with small data sets. Do we think the relationship between jersey number and salary for the '98-'99 San Antonio Spurs is real or a coincidence permitted by a small population?

Obs	Number	Player	Pos	Ht	Wt	Birth Date	G	Exp	College	Salary
1	33	Antonio Daniels	PG	04JUN	195	19MAR1975	us	1	Bowling Green State University	2,472,000
2	21	Tim Duncan	PF	11JUN	250	25APR1976	vi	1	Wake Forest University	3,413,000
3	17	Mario Elie	SG	05JUN	210	26NOV1963	us	8	American International College	1,750,000
4	32	Sean Elliott	SF	08JUN	205	02FEB1968	us	9	University of Arizona	5,333,000
5	10	Andrew Gaze	SG	07JUN	205	24JUL1965	au	1	Seton Hall University	350,000
6	6	Avery Johnson	PG	10MAY	175	25MAR1965	us	10	Southern University and A&M College	3,000,000
7	4	Steve Kerr	PG	03JUN	175	27SEP1965	lb	10	University of Arizona	1,774,000
8	25	Jerome Kersey	SF	07JUN	215	26JUN1962	us	14	Longwood University	1,000,000
9	54	Gerard King	SF	09JUN	230	25NOV1972	us	R	Nicholls State University	287,500
10	41	Will Perdue	C	01JUL	240	29AUG1965	us	10	Vanderbilt University	4,373,000
11	50	David Robinson	C	01JUL	235	06AUG1965	us	9	United States Naval Academy	14,841,000
12	31	Malik Rose	PF	07JUN	250	23NOV1974	us	2	Drexel University	425,000
13	11	Brandon Williams	SG	06JUN	215	27FEB1975	us	1	Davidson College	287,500



How do we compute the correlation coefficient r on the TI-84? Again, place the x-values (**explanatory**) in L_1 and the y-values (**response**) in L_2. Press **STAT** and select the **CALC** menu. Now select **8: LinReg(a+bx)**.

Exercise 2 *Use the TI-84 to find the correlation coefficient for property rent and cost in Monopoly.*

4 Homework

1. Navidi/Monk: Section 11.1: 15-18, 23-25