

Should the Number of Clinical Trials be Limited at Medical Centers?

By
Student 5

A. Summary

I wanted to determine if the number of clinical trials, or studies, (explanatory variable) a medical center participates in affects the number of protocol violations (response variable) reported for that center. The two variables were collected for the year 2001 from a central (proprietary) database of a national clinical trials program. The data appears to have a strong, positive linear relationship. The relationship's equation is:

$$violations = -0.515078 + 0.509984(studies)$$

B. Data Gathering

Data was collected from a central database for a national clinical trials program. This database is proprietary. The first of three variables to be collected is medical center number. This data is ordinal and is only used to identify the medical centers. The second variable (number of studies) and the third variable (number of protocol violations) are both ratio. A scatter plot (Figure 1) of the data indicates a strong linear relationship.

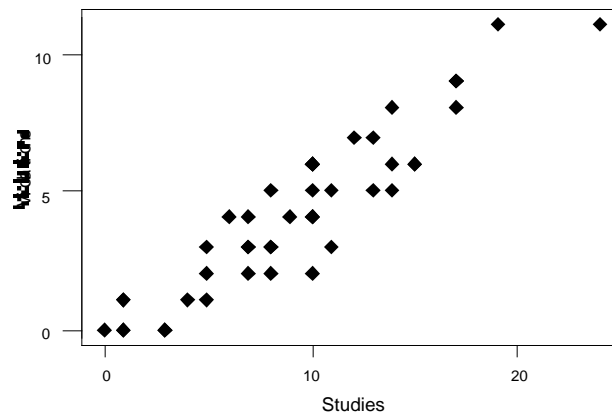


Figure 1. Scatter plot of Data.

C. Population and Sample

The population was limited to the medical facilities in a specific healthcare system. The data was further limited to those facilities authorized to perform clinical research. Consequently, the data is not from a random sample. A couple of possible biases exist. The first bias is that a center with a larger number of research projects has a greater chance of having more protocol violations. A possible second bias is that larger centers would have more resources available to manage their research portfolio and oversee their program thereby reducing their risk of protocol violations.

D. Box Plots

A scatter plot (Figure 1) of the data shows the possibility of an outlier. For the variable *studies*, any value below -7.5 or above 26.5 would be an outlier, and for the variable *violations*, any value below -4 or above 12 would be an outlier. The coordinates of the possible outlier are (24, 11), therefore it is not an outlier. However, if it had been an outlier I would have kept the data point since it would identify a site that is possibly overworked or has an excessive number of protocol violations and is consequently putting the research patients at an increased risk.

The boxplot for *studies* (explanatory variable) (Figure 2) displays a skew to the right (positive). The median for this variable is slightly off-center indicating that the mid 50% of this data is slightly skewed left (negatively).

Boxplot of Studies

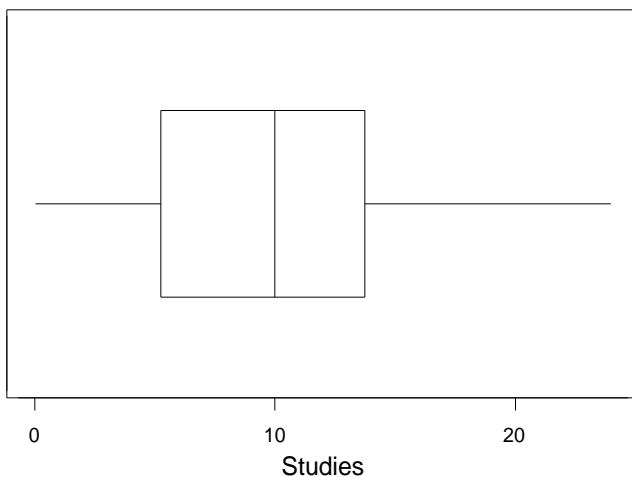
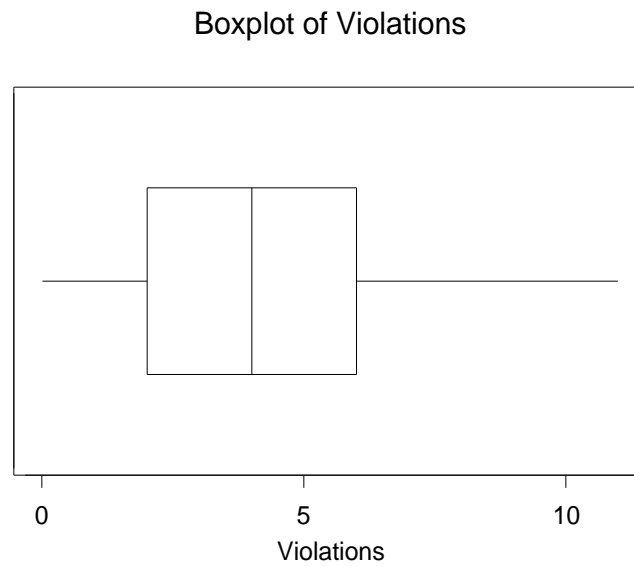


Figure 2. Box plot of Explanatory Variable (*studies*).



The boxplot for *violations* (response variable) (Figure 3) also shows a positive (right) skew. The middle 50% of data for this variable appears symmetrical since the median is centered.

Figure 3. Box plot of Response Variable (*violations*).

E. Regression Analysis

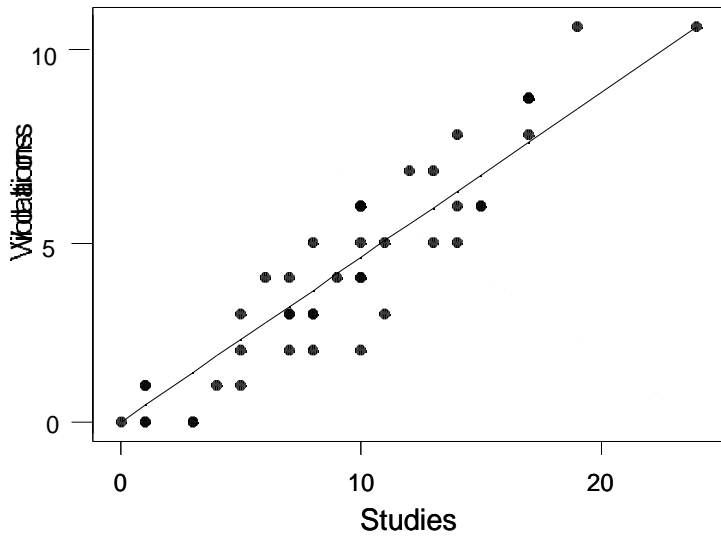
The data table (Table 1) is given below, and as discussed earlier, there were no outliers and therefore, no adjustments were made to the data (original data = final data).

Site#	Studies	Violations		Site#	Studies	Violations
59	5	2		157	12	7
63	1	0		159	14	6
65	10	4		166	7	3
68	3	0		169	0	0
75	8	3		173	8	3
80	13	5		177	14	8
82	14	5		185	10	6
85	5	3		189	10	4
87	7	4		194	7	3
90	1	1		199	19	11
96	7	2		205	17	8
99	1	0		207	15	6
108	17	9		209	10	5
120	11	3		221	13	7
132	10	2		222	15	6
137	24	11		226	4	1
139	17	9		229	5	1
142	10	6		230	10	6
143	6	4		232	11	5
148	9	4		233	3	0
149	1	1		248	8	2
596	8	5		250	17	9

Table 1. Data Table

Below is a scatter plot with regression line (Figure 4). Following this plot are the printed regression statistics from Minitab (Figure 5) and the fitted line plot (Figure 6). R^2 (coefficient of determination) is 0.867 meaning that 86.7% of the variation of *violations* (response variable) is explained by the equation. The correlation coefficient (r) is 0.931 and confirms a strong, positive relationship.

Figure 4. Scatter plot with regression line.



S

R-Sq(adj) = 86.4%

MS	F	P
334.202	273.384	0
1.222		

Figure 5. Regression

Statistics from Minitab.

Regression Plot

Violations = -0.515078 + 0.509984 Studies

S = 1.10565 R-Sq = 86.7 % R-Sq(adj) = 86.4 %

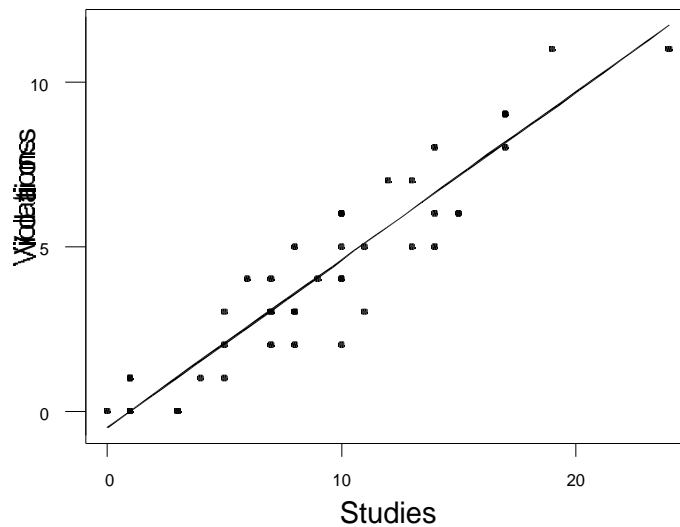
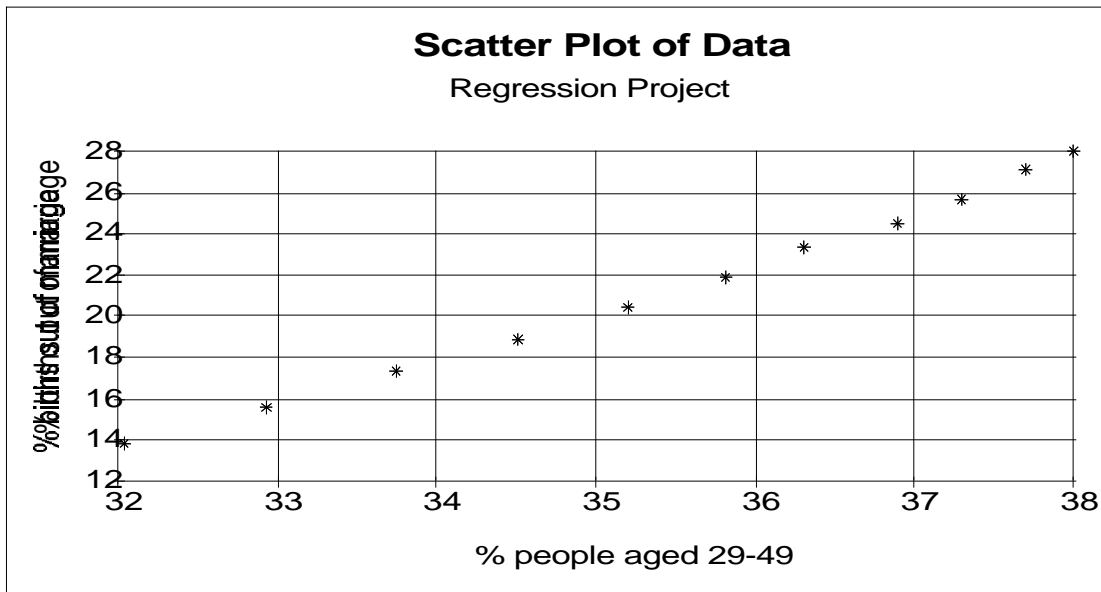


Figure 6. Fitted Line Plot

Regression Project - Written Report

A. The purpose of this project is to show a relationship between the percent of live births outside of marriage out of all live births (response variable) and the percent of people between the ages of twenty-five and forty-nine out of the total population of the United States (explanatory variable). The data for this project was retrieved from a website. In the end, there is a relationship between the two variables. It is a linear relationship, and the equation for its regression line is $y = 2.35756x - 62.1700$. The line's correlation coefficient (r) of **.9970648398** shows that the line has a strong, positive linear relationship.

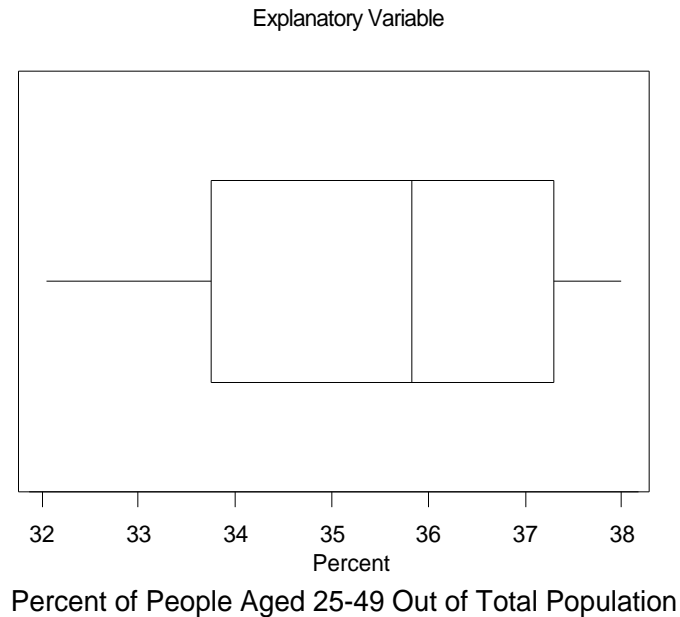
B. The response variable in this project is the percent of live births outside of marriage out of all live births; it is a ratio measurement. The explanatory variable is the percent of people between the ages of twenty-five and forty-nine out of the total population of the United States; it is also a ratio measurement. This data came from a website, <http://www.worldinfigures.org>. I found this website by typing "United States statistics" into a search engine. Once I was there, I tried to find two variables that could be logically related. A copy of the webpage from which the data was gathered is included in the appendix. The data is from the years 1980 through 1990, and is from the United States.



C. The population of the explanatory variable is all citizens of the United States between 1980 and 1990 who were aged twenty-five through forty-nine. The population of the response variable is all the babies born in the United States between 1980 and 1990 whose mother was not married. Because it would have been nearly impossible for researchers to survey everyone in both populations, the data must be a sample. The sample was most likely not random, because not everyone who was aged twenty-five through forty-nine had an equal chance of being included in this data, nor did every

single mother. In order for a sample to be random, all parts of the population must have an equal chance of being selected. This data could be bias because it does not contain information about babies born secretly out of hospitals who were abandoned by their mothers.

D.



The box plot for the explanatory variable is nearly symmetrical, with only a slight skew to the right. The skew is due to the fact that the median (Quartile 2) is not directly in the center of Quartile 1 and Quartile 3. There are no visible outliers.

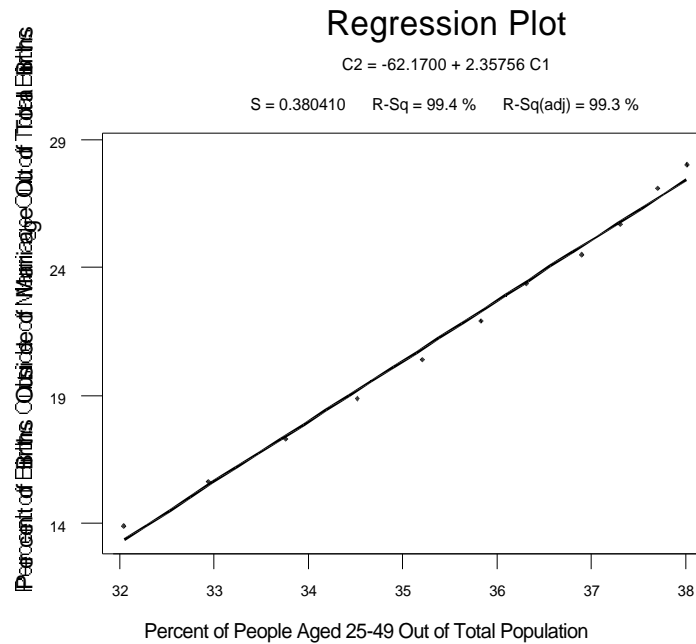


The box plot for the response variable is even more symmetrical than that of the explanatory variable; that is, the median is almost directly in the center of Quartile 1 and Quartile 3. Also, the whisker lines are nearly equal. Once again, there are no visible outliers.

None of the points appear to be outliers. To be sure, I used the formula for finding outliers to calculate any possible outliers for both sets of data. For the explanatory data (the percent of people aged 25-49 as compared to the total population) I found that Quartile One equaled 33.75 and Quartile Three equaled 37.3. By finding the difference between the two quartiles and multiplying the answer by 1.5, I found that the Interquartile Range (IQR) equaled 5.325. To find outliers, I subtracted the IQR from Quartile One and added the IQR to Quartile Three. Any data that was below or above the resulting number respectively was an outlier. None of the data were outliers.

For the response data (percent of live births outside of marriage as compared out of all live births), I followed the same process. Quartile One equaled 17.29 and Quartile Three equaled 25.71. The IQR turned out to equal 12.63. By subtracting the IQR from Quartile One and adding it to Quartile Three, I found that, once again, none of the data qualified as being an outlier.

E.



Regression Analysis: C2 versus C1

The regression equation is

$$C2 = -62.1700 + 2.35756 C1$$

$$S = 0.380410 \quad R\text{-Sq} = 99.4 \% \quad R\text{-Sq}(\text{adj}) = 99.3 \%$$

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	220.886	220.886	1526.39	0.000
Error	9	1.302	0.145		
Total	10	222.188			

*C1 = Explanatory Variable (population) C2 = Response Variable (births)

Because I removed no outliers from my data, the original and final data are the same.



Explanatory Variable (C1)	Response Variable(C2)
32.04	13.88
32.93	15.62
33.75	17.29
34.51	18.88
35.20	20.41
35.82	21.89
36.30	23.39
36.90	24.49
37.30	25.71
37.70	27.08
38.00	28.00