Jonathan Tran
CSC 21700

Regression

The study of linear regression is a method of analyzing a linear relationship between a dependent variable “y” and a independent variable “x”, if there is a relationship. The regression model will generated using a MD5 hash of the name “Jonathan”. The MD5 hash will be split into four different strings, each having its own purpose. The regression model population will be randomly generated by a seed. The approach of creating this regression model and its randomness leads to believe that we will be seeing little to no correlation between variables, meaning there is no significant relationship between the two variables. In addition, depending on the sample we obtained, it may be difficult to make a statistical inference.

The population of this simulation will consist of randomly generated “y” values created by randomly generated “x” values. The “x” values will be randomly generated using a seed, and ranges from 0 to 1. The population was generated using C++ to obtain the “x” values. In addition to the “x” values, there will be an error value added to it. The error term is due to regression models not being entirely close to reality, so this must be accounted for. The error values will be normally distributed. The population regression model parameters are generated below.

Hash: **C1F80EDDEA77F14650A2062DDA3EB15C**
-Divide hash into 4 strings.
C1F80EDD EA77F146 50A2062D DA3EB15C

Using a hex calculator divide the first and second hash by 7FFFFFFF. The third hash is the seed for the random number generator. The fourth hash will be divided by 7FFFFFFF, which will give us the standard deviation.

βo = 1.5153826319218533
$\beta 1$ =1.8317853546849383
seed = 1352795693
σ_e = 1.7050382335228091

All of this gives us the formula:
Y = 1.5153826319218533 + 1.8317853546849383x + ε, ε~N(0, 1.7050382335228091)

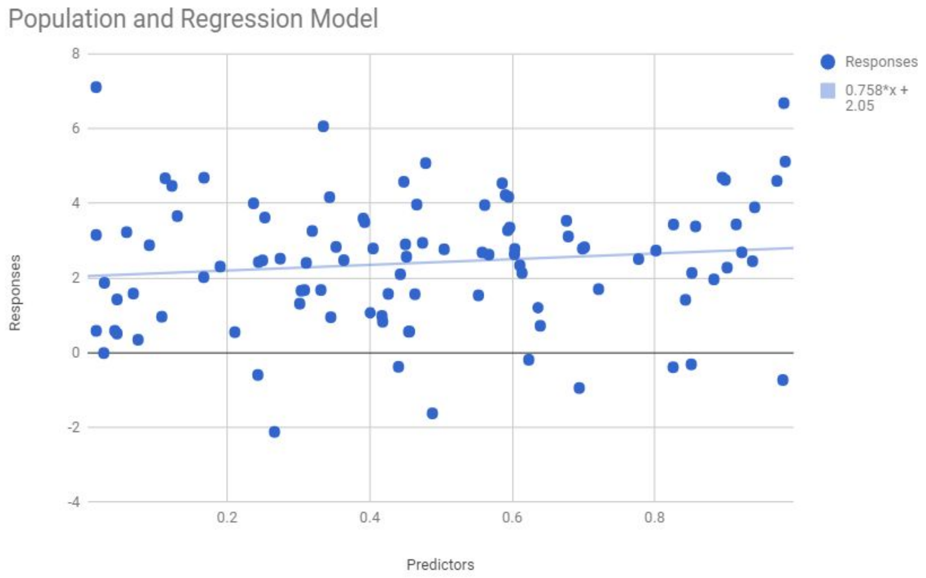
Population

A population size of a 100 will be generated. The predictor x will be generated using the seed along with normally distributed values of ε, giving us our responses. Below is a table summary of our population.

N=100	Mean	Standard Deviation
Predictor X	0.475246149	0.280706573916607
Responses Y	2.411304998	1.7214543499611019

N=100	Min	1st Quartile	Median	3rd Quartile	Max
Predictor X	0.0159307	0.2628865	0.4530165	0.676687	0.983032
Responses Y	-2.12033	1.3919025	2.514685	3.4536925	7.11557

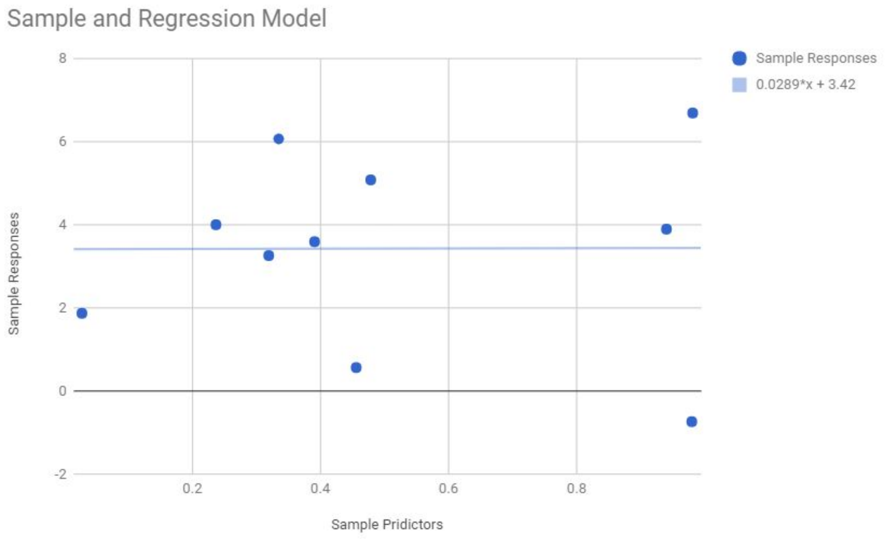
Below is a Regression Model of our population.

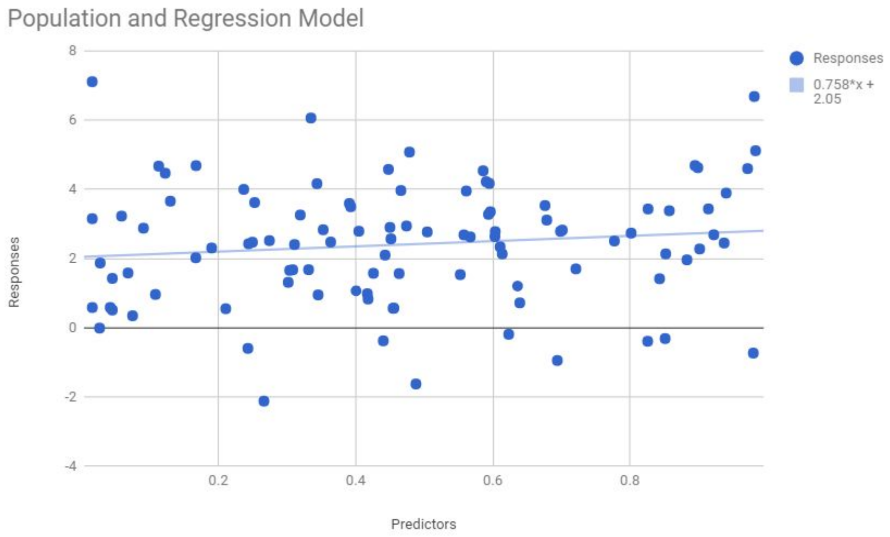


<u>**Sample of Regression Model**</u>

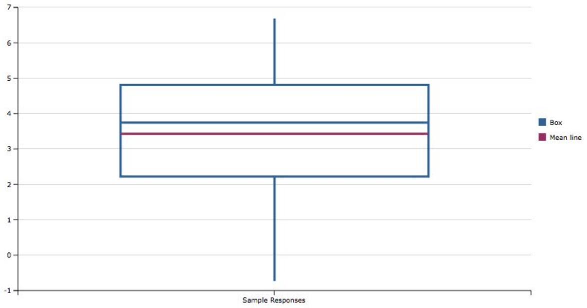
Usually we don't look at the population because its too expensive to analyze all the data. For this project, the sample regression model is what will be analyzed. The sample will be every 10th predictor and its associated response, because its every 10th value the sample size will a 10.

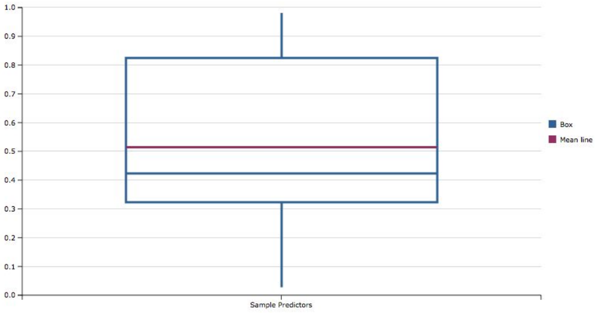
On the next page is the Regression model of the sample and below that is the population model. Notice the difference; because of the sample size, the regression line of the sample is flatter almost as if the slope of the line is 0. The points on the graph are also more spread out and it's difficult to see if there is a relationship between the variables.





Below is the distribution of the predictors and the responses.





Below is a table summary of our sample.

N=10	Mean	Standard Deviation
Predictor X	0.51445672	0.336667046501435
Responses Y	3.4299051	2.3284586790407083

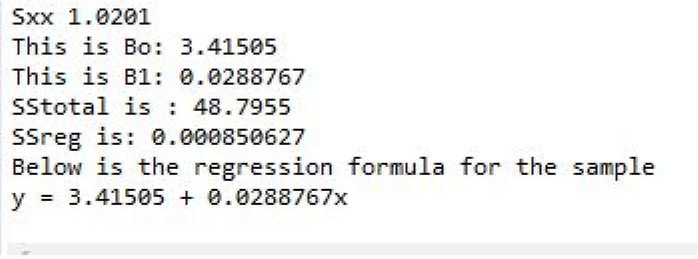
N=10	Min	1st Quartile	Median	3rd Quartile	Max
Predictor X	0.02756	0.32312	0.42328	0.82476	0.98117
Responses Y	-0.73062	2.22045	3.74515	4.81061	6.68726

Construction of the regression model will require these formulas below.

Regression estimates

$$
\begin{aligned}
b_0 &= \widehat{\beta}_0 = \bar{y} - b_1\,\bar{x} \\
b_1 &= \widehat{\beta}_1 = S_{xy}/S_{xx} \\
&\text{where} \\
S_{xx} &= \sum_{i=1}^{n}(x_i - \bar{x})^2 \\
S_{xy} &= \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})
\end{aligned}
$$

Using C++, the regression model has been obtained, code will be provided.



```
Sxx 1.0201
This is Bo: 3.41505
This is B1: 0.0288767
SStotal is : 48.7955
SSreg is: 0.000850627
Below is the regression formula for the sample
y = 3.41505 + 0.0288767x
```

The program obtained the regression model formula y = 3.41505 + 0.0288767x. With this equation we can make some kind of prediction on what the response will be base on our x value. If we were to plug in 0.5 into the equation, we will get a y value.

Y = 3.41505 + 0.0288767(0.5)

Y = 3.42948835

With the given regression formula we can now work on analyzing the equation. First we will start with finding the confidence interval. A significant level of 95% will be used for the confidence interval to find the range of which the true slope is located.

Find the regression variance:
Regression Variance = SS_{ERR} /(n-2)
SS_{REG} = β1*β1*Sxx;
SS_{TOT} = $\sum$(yi- $\bar{y}$)^2
SS_{ERR} = SS_{TOT} - SS_{REG}
S_{xx} = Xn i=1 (xi − x¯)^2
S_{xx}= 1.0201 (given in code)

SS_{REG} = 0.0288767*0.0288767 *1.0201

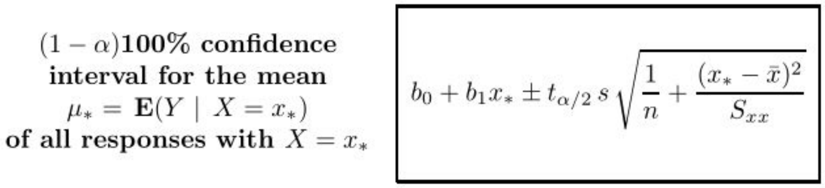
SS_{REG} = 0.000850624465328089
SS_{TOT} = 48.7955
SS_{ERR} = 48.794649375534671911

Regression Variance s^2 = SSerr / n-2
Regression Variance = 48.794649375534671911 / 8

Regression Variance = 6.0993
Regression standard deviation = 2.470

Degrees of Freedom = 8
Using this equation Standard Deviation of b Std (β1) = Standard Deviation / Square Root(Sxx) = **2.445**
Using the student’s T-distribution with the degrees of freedom 8, we found the value = **2.306.**
β1 +||- t * s/square root(Sxx)
0.0288767 +||- 2.306*2.445
The confidence interval is shown below.
(-5.6092933,5.6670467)

With the information we have we can calculate the mean of all responses.



$(1-\alpha)$100% confidence interval for the mean $\mu_* = \mathbf{E}(Y \mid X = x_*)$ of all responses with $X = x_*$

$$b_0 + b_1 x_* \pm t_{\alpha/2}\, s \sqrt{\frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}}}$$

x* = 0.487899, a value from our population.
Plugging in x* into regression model.
Y = 3.4291389130533
3.42 +||- 2.306*2.470*sqrt(1/10+(0.487899- 0.51445672)/1.0201)
The mean of all responses = (1.87093,4.96907)

What the confidence interval about the slope tells is that the we are 95% confident that the true slope lies between these two values. With this we can see that the confidence interval contains the value 0. If the true slope was actually 0, then we should see a straight line going across the graph with no change in height. Since 0 is within the confidence interval, there is a evidence that there is no relationship. We can use the information we have so far to build an anova table. P-value is calculated via software.

ANOVA TABLE

	d.f	SS	MS	F	p-value
Regression	1	0.00085	0.00085	0.00014	0.99087
Residual	8	48.79464	6.09933		
Total	9	48.7955			

Calculating the T statistic is 0.02888/(sqrt(6.09933)/1.0201) = 0.01193. Using the F Table Numerator 1 and denominator 8, the closest value is 1.54, which is why we need software to calculate the p-value. Without it, we would have to estimate. We still need to conduct a hypothesis test on the slope. We calculated the 95% confidence interval, but we still need evidence due to the range (-5.6092933,5.6670467) .

<u>Hypothesis Testing</u>

Ho: β1 = 0
Ha: β1 ≠ 0.

Assume Ho is true. We will use a significant level of 0.05 for this test. The p-value = 0.99087. If the p-value is less than 0.05, reject null, if its greater accept null. Seeing as how the p-value is far beyond 0.05, there is strong significant evidence to accept the null hypothesis. Accepting the null hypothesis means there is evidence that the slope is 0.

The correlation coefficient measures the strength of the linear relationship between the two variables

$$R^2 = \frac{SS_{\text{REG}}}{SS_{\text{TOT}}}$$

R^2 = 0.00085/48.79464 = 0.000017

R = 0.00417

The correlation close to 0 represent no linear relationship between the variables. 1 represents a position correlation and -1 a negative correlation. Seeing as how close 0.00417 is to 0. There is basically no relationship.

The information gathered so far is enough to make a statistical analysis of the regression model. In the introduction, I explained that due to the randomness of the variables, the regression model will have no relationship. The gathered data supports my claim. The confidence interval calculated with a 95% confidence contains the true value of the slope within a range (-5.6092933,5.6670467). The value 0 is contained within this range. For hypothesis testing, we assumed that the null hypothesis is true (slope = 0). The p-value is 0.99087, and because this p-value is far greater than 0.05, there is strong significant evidence that the slope is 0. In Addition, the r value is close to 0. All of this supports my claim that the regression model has very strong evidence that there is no relationship, but we overlooked one problem with our sample.

Looking back at our population and how we obtained our sample; with such a small sample size it's difficult to make a statistical inference on it. Looking back at our population, it had some kind of relationship; however, due to the method of sampling, the relationship was difficult to see. The population had a weak relationship, but it had a relationship nonetheless. The slope of the population model was also slightly changing and increasing, whereas our sample had a slope close to 0. In conclusion, the sample size was too small, and it didn’t give us enough information for the population model. We need a different method of sampling or a larger sample of the population. Statistic requires multiply sampling and a good sample size. There is always a change for false positives when rejecting and accepting a null hypothesis.