Kevin Scharnhorst
Med Inf 409
Spring 2011

Homework 4

DIRECTIONS

Problems should be done using the computer. For each problem, be sure to:

- Discuss assumptions of the analysis for each problem
- Display all appropriate data and discuss the results as you would for an article

Problem 1

Consider the following data:

X	Y
2	3
3	6
4	8
6	4
7	10
8	14
9	8
10	12
11	14
12	12
13	16

a.) Regress Y on X. Obtain and interpret the prediction equation. Predict the value of Y when X=5.
b.) Plot the standardized residuals against the standardized predicted values. Do you see any pattern? What does this suggest?
c.) Overall, how well does the model fit?

Hypothesis

H_0 – There is no significance relationship between the independent variable (X) and dependent variable (Y)

H_A – There is a significant relationship between our independent and dependent variables.

Assumptions

- Our independent variable (X) is continuous and they appear to be.

- Our samples are normally distributed

- Independence between variables
- Simple random sampling
- Constant variance exists
- Linearity

Results

Part A:

Before we continue further with analysis we evaluate the R^2 value from our model summary and see that we have a value of .707 or 70.7%. That percentage represents the degree of variability which can be described by our model. 70.7% is considered moderate coorelation so we continue

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.841[a]	.707	.675	2.448

a. Predictors: (Constant), X

b. Dependent Variable: Y

To determine our prediction equation, we use the Coefficients. The Coefficients[a] model can be interpreted by pulling out elements from the model. Our equation will be

Predicted variable (dependent variable or Y) = slope * independent variable + intercept

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	2.170	1.781		1.218	.254	-1.859	6.199
	X	.978	.210	.841	4.662	.001	.503	1.453

a. Dependent Variable: Y

The slope tells us how steep the line regression is. A slope of 0 is a horizontal line, a slope of 1 (the case in this example) is a diagonal line from the lower left to the upper right, and a vertical line has an infinite slope. The intercept is where the regression line strikes the Y axis when the independent variable has a value of 0. So evaluating the model above, our equation is:

Y = .978 X + 2.170

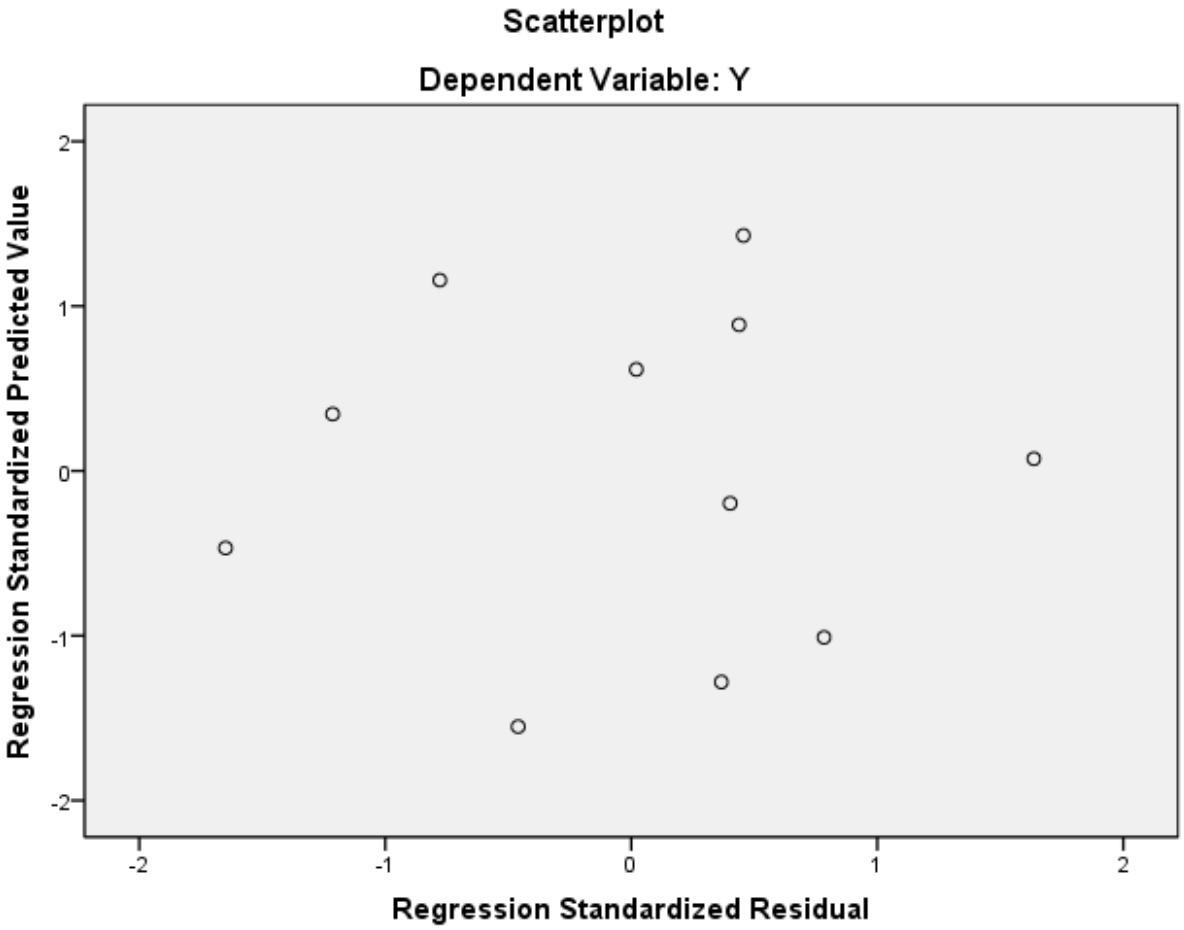
So substituting a value of 5 for X, we can predict the value of Y as

Y = (.978)(5) + 2.170 or
Y = 4.89 + 2.170 or

Y = 7.06 when X = 5

Part B

A scatter plot with standardized residuals against the standardized predicted values helps us to to determine if we have any abnormal data points. In evaluating below, we can see that the majority of our data centers around zero, so we can assume that it is safe to proceed forward.



Part C

The below ANOVA shows that our significance factor is .001 or .1%. Since .1% is less than 5%, We would therefore **reject the null hypothesis and accept the alternative hypothesis.**

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	130.248	1	130.248	21.735	.001[a]
	Residual	53.934	9	5.993		
	Total	184.182	10			

a. Predictors: (Constant), X

b. Dependent Variable: Y

Problem 2

An experimenter was interested in the possible linear relation between the time spent per day in practicing a foreign language and the ability of the person to speak the language at the end of a 6-week period. Some 50 students were assigned at random among five experimental conditions ranged from 15 minutes practice daily to 3 hours practice per day. Then, at the end of 6 weeks, each student was scored for proficiency in the language. The data follow:

Proficiency Scores, By Daily Practice Time
(x = Practice, in hours)

.25	.50	1	2	3
117	106	86	140	105
85	81	98	128	149
112	74	125	108	110
81	79	123	104	144
105	118	118	132	137
109	110	94	133	151
80	82	93	96	117
73	86	91	101	113
110	111	122	103	142
78	113	130	135	112

Find the linear regression equation for predicting Y, the proficiency of a student, from X, the practice time per day.

Hypothesis

H_0 – There is no significant relationship between the practice time variable(s) and the resultant proficiency of a student to speak a foreign language.

H_A – Significant relationship exists between practice time and the resultant proficiency to speak a foreign language.

Assumptions

- X is the independent variable, practice time (in hours)
- Y is the dependent variable, proficiency of a student
- Our independent variable (X) is continuous and they appear to be.
- Our samples are normally distributed
- Independence between variables
- Simple random sampling
- Constant variance exists
- Linearity

Results and Inference

To determine predictive equation for Y, we run a linear regression test on our variables. We have five different groups of variables (x1, x2, x3, x4, x5) which represent the different times of practicing a foreign language. Our dependent variable is the resulting proficiency scores and is represented by Y. To compare X and Y, we setup X by assigning a value to the associated group practice time where 1=.25 hrs, 2 = .5 hrs, 3 = 1 hr, 4 = 2 hr and 5 = 3 hr. The Y variable then are the corresponding proficiency scores for each 10 students for each group.

To determine the predictive equation, we evaluate the Coefficient equation and consider the below formula to replace it with pertinent values.

Predicted variable (dependent variable or Y) = slope * independent variable + intercept

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.602[a]	.362	.349	16.861

a. Predictors: (Constant), Practice Time in Hours

b. Dependent Variable: Proficiency Score

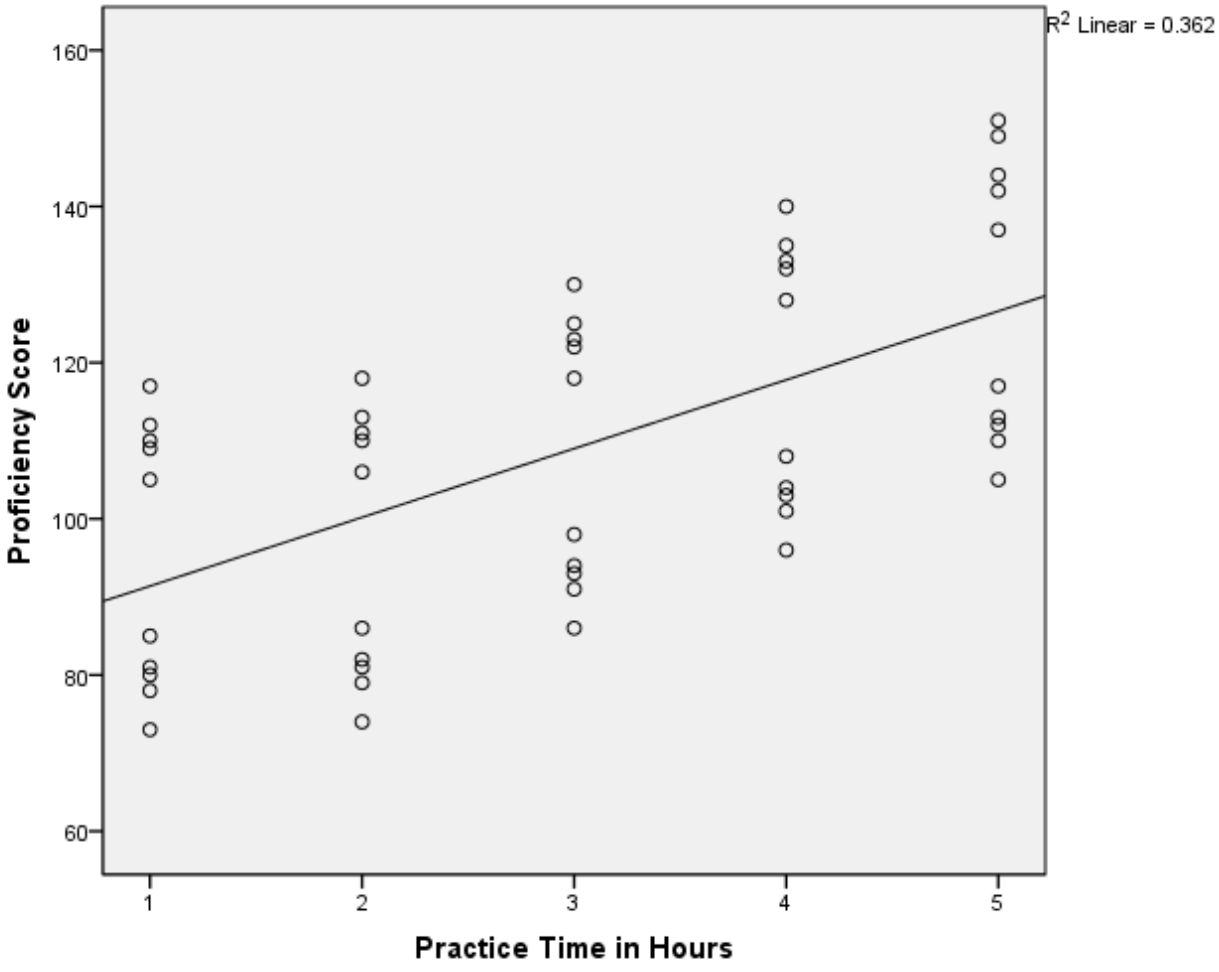
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	82.600	5.592		14.771	.000
	Practice Time in Hours	8.800	1.686	.602	5.219	.000

a. Dependent Variable: Proficiency Score

So, our predictive formula is then:
Y (Proficiency in speaking a foreign language) = 8.80 X + 82.600

We can also see from the scatter graph that there is a significant relationship between practice time and proficiency score. As more time is spent practicing the proficiency score increases. The biggest improvement shown with the more hours spent practicing a foreign language.



Further from the Coefficients, we see that both are significant. In considering the ANOVA, we see that the significance is .000 which is less than .005. We would therefore conclude that we would **reject the null hypothesis and accept the alternative hypothesis**.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7744.000	1	7744.000	27.240	.000[a]
	Residual	13646.000	48	284.292		
	Total	21390.000	49			

a. Predictors: (Constant), Practice Time in Hours

b. Dependent Variable: Proficiency Score

Problem 3

An experimenter was interested in the possible linear relationship between the measure of finger dexterity *X* and another measure representing general muscular coordination *Y*. A random sample of 25 persons showed the following scores:

Person	*X* Value	*Y* Value
1	75	84
2	77	94
3	75	90
4	76	90
5	75	91
6	76	86
7	73	87
8	75	95
9	74	83
10	75	85
11	76	88
12	74	91
13	72	80
14	75	85
15	73	87
16	75	82
17	78	86
18	76	83
19	74	85
20	74	88
21	77	100
22	75	98
23	76	89
24	74	91
25	75	99

Compute the correlation coefficient, and test its significance (α=.05)

Hypothesis

H_0 – There is no significant relationship between finger dexterity and general muscular coordination.

H_A – There is a significant relationship between finger dexterity and general muscular coordination.

Assumptions

- X is the independent variable, finger dexterity
- Y is the dependent variable, muscular coordination
- Our independent variable (X) is continuous and they appear to be.

- Our samples are normally distributed
- Independence between variables
- Simple random sampling

Results and Inference

In looking at the Model summary below we see that the R^2 value is .105 or 10.5% demonstrates hardly any correlation. The Correlation Coefficent (R) is .324. Since there is no correlation, we would not continue any further with analysis.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.324[a]	.105	.066	5.134

a. Predictors: (Constant), Finger Dexterity

Further, we can evaluate using an ANOVA analysis and see below that the significance factor of P is 11.4% which is greater than 5%. We can therefore conclude that **we fail to reject the null hypothesis** and conclude there is not significant relationship between finger dexterity and muscular coordination. We can further see below from the scatter chart below that there appears to be no relationship shown.

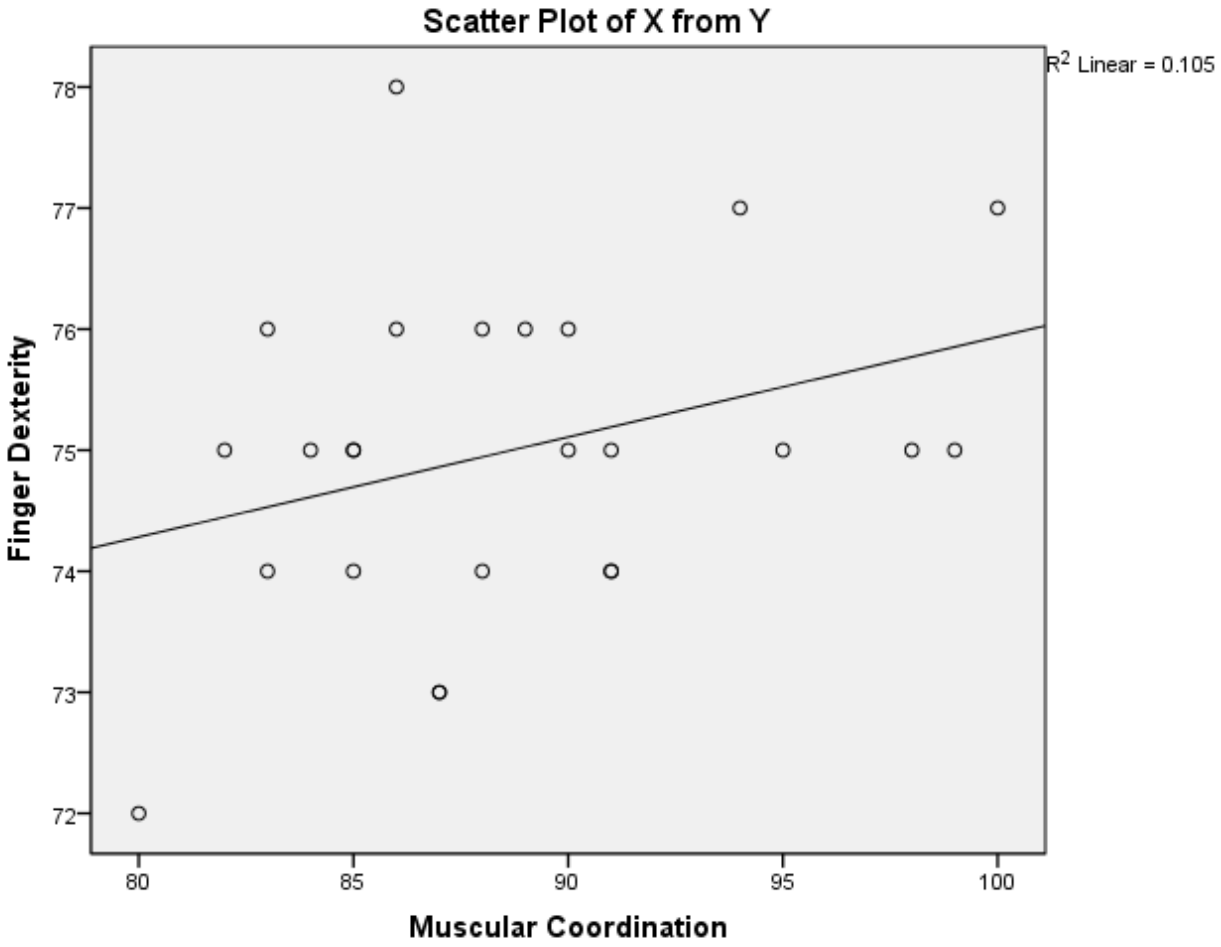
ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	71.273	1	71.273	2.704	.114[a]
	Residual	606.167	23	26.355		
	Total	677.440	24			

Problem 4

Based on the data in Exercise 3, find the regression equation for predicting *X* from *Y*. Plot this regression equation along with the raw data. What is the appropriate measure of the "scatter" or horizontal deviations of the obtained points in this plot about the regression line?

We reverse the axis from the previous problem to predict X (our independent variable) from Y (our dependent variable). Below we see that the slope is diagonal.



To form our predictive equation we reverse our dependent and independent variables in SPSS and consider the resulting coefficients.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	67.669	4.465		15.154	.000	58.432	76.907
	Muscular Coordination	.083	.050	.324	1.644	.114	-.021	.187

a. Dependent Variable: Finger Dexterity

Our equation to predict our independent variable (X) then is $X = .083Y + 67.669$.

Problem 5

A developmental psychologist believes that the age at which a normal child begins to speak words clearly is closely related to the age at which the child first begins to use complete sentences. A random sample of 33 normal children was taken, and careful records were kept for each. Let *X* be the age at which words are first clearly used, and let *Y* be the age at which complete sentences are used. The following data gives the values of *X* and *Y* in months. Find the correlation between *X* and *Y*, and compute and interpret an appropriate regression equation.

Child	*X*	*Y*	Child	*X*	*Y*	Child	*X*	*Y*
1	15.1	25.2	12	14.3	25.7	23	13.6	24.3
2	12.7	24.3	13	11.5	23.4	24	15.2	26.3
3	11.7	22.1	14	13.4	25.7	25	12.1	23.4
4	13.1	23.3	15	13.7	24.5	26	12.6	24.5
5	13.0	24.1	16	13.5	26.0	27	14.1	26.2
6	11.2	23.6	17	12.8	24.6	28	11.2	23.0
7	13.3	25.5	18	13.2	25.4	29	14.0	24.3
8	12.3	24.3	19	14.7	26.3	30	13.1	25.3
9	13.7	25.5	20	12.2	25.2	31	11.5	24.2
10	12.2	23.2	21	14.7	26.4	32	14.9	27.2
11	13.3	27.1	22	14.6	25.8	33	13.8	26.3

Hypothesis

H_0 – No significant relationship exists between the age at which words are first spoken clearly and the age at which a complete sentence is used.

H_A – There is a significant relationship between the age at which words are first spoken clearly and the age at which a complete sentence is used.

Assumptions

- X is the independent variable, age at which words are first clearly used
- Y is the dependent variable, age at which complete sentences are used
- Our independent variable (X) is continuous and it appear to be.
- Our samples are normally distributed
- Independence between variables
- Simple random sampling

Results

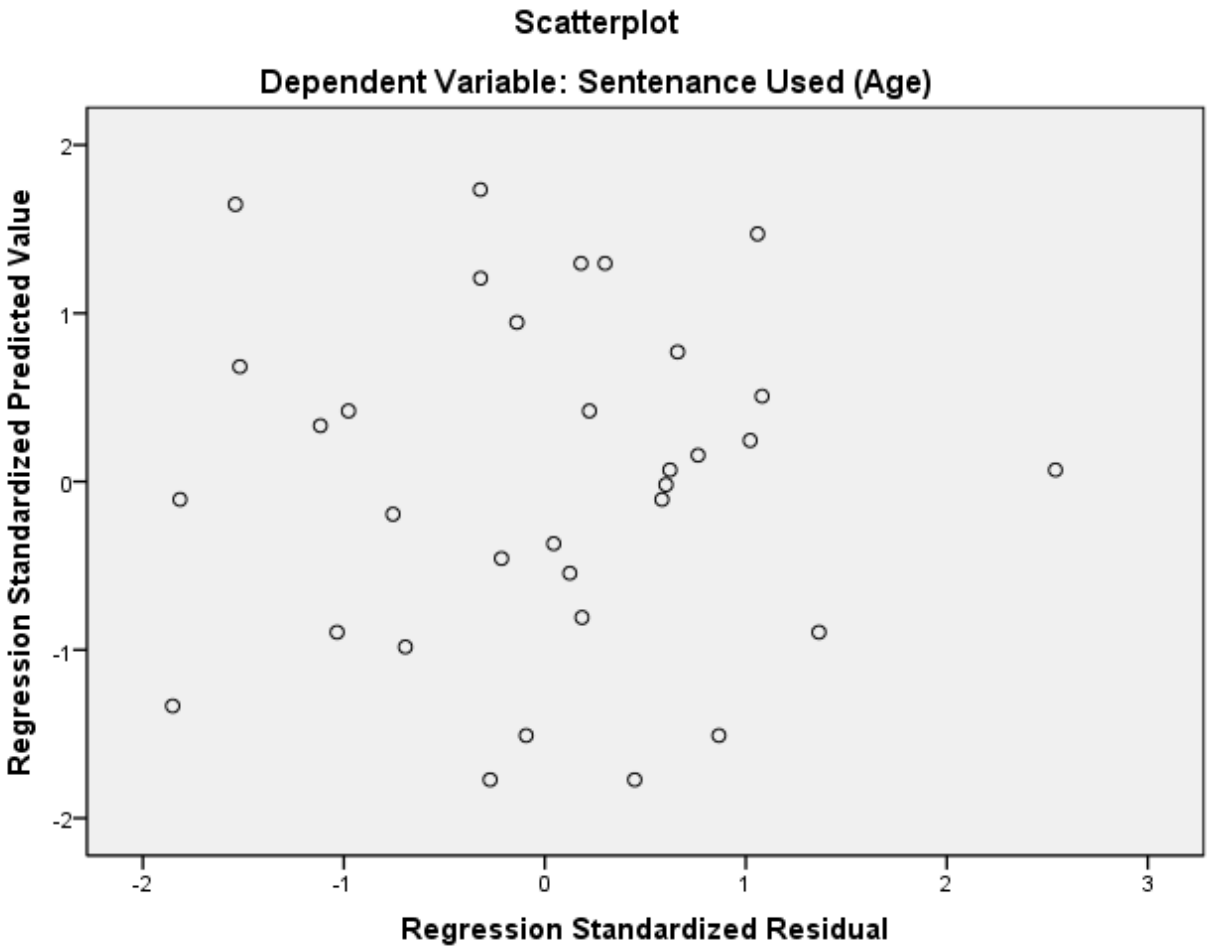
From the model summary from a linear regression analysis we see that R^2 is .574 or 57.4%. R2 represents the correlation between our dependent and independent variables. Because it is between 30% and 70%, this is considered a moderate correlation and we would continue with further analysis.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.758[a]	.574	.560	.8341

a. Predictors: (Constant), Words Spoken (Age)

To further interpret the correlation we can run a scatter plot to show the standardized residuals against the standardized predicted values. Below are the results of that graph. Seeing that all the values group around zero, we can gain greater confidence that there is high correlation between our variables.



We can calculate the regression equation with the coefficents. The equation would be

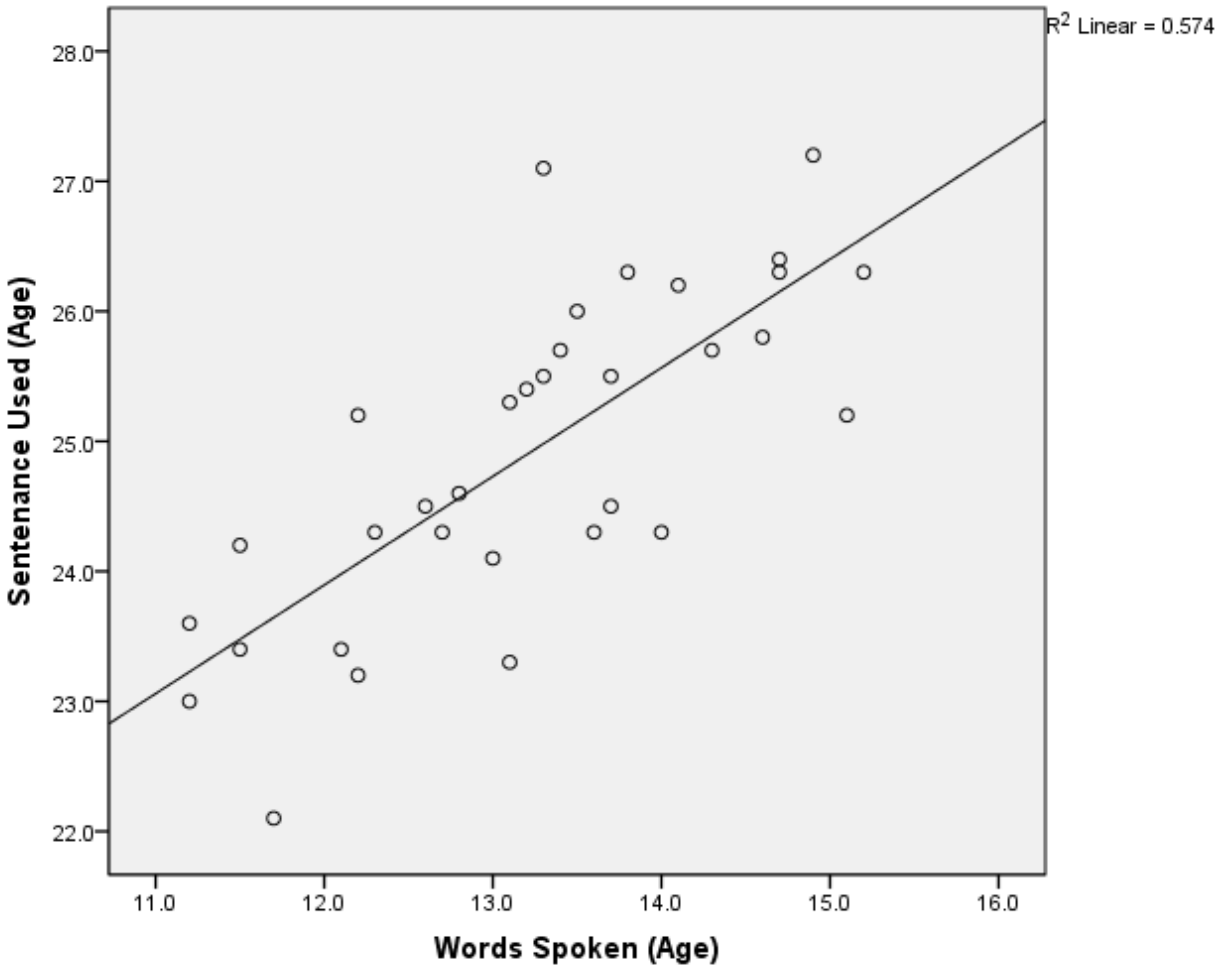
$Y = .835X + 13.873$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	13.873	1.715		8.090	.000
	Words Spoken (Age)	.835	.129	.758	6.462	.000

a. Dependent Variable: Sentenance Used (Age)

In interpreting the values and equation, we can see that the slope is .835 and would expect to see a diagonal slope increasing from bottom left to upper right. The scatter plot below shows us this.



To determine the model fit, we would then look at our ANOVA and from below we see that the P measure or significance factor is .000 which is less than 5%. We **therefore reject the null hypothesis** and accept the alternative hypothesis and assume there is a significant relationship between when children first speak words clearly and then use sentences clearly.