Simple Linear Regression in SPSS **STAT 314**

1. Ten Corvettes between 1 and 6 years old were randomly selected from last year's sales records in Virginia Beach, Virginia. The following data were obtained, where x denotes age, in years, and y denotes sales price, in hundreds of dollars.

x	6	6	6	4	2	5	4	5	1	2
y	125	115	130	160	219	150	190	163	260	260

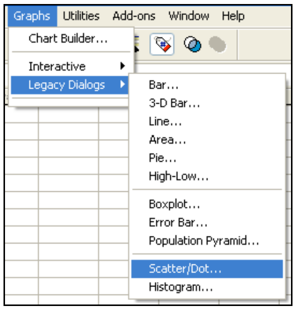
a. Graph the data in a scatterplot to determine if there is a possible linear relationship.
b. Compute and interpret the linear correlation coefficient, r.
c. Determine the regression equation for the data.
d. Graph the regression equation and the data points.
e. Identify outliers and potential influential observations.
f. Compute and interpret the coefficient of determination, r^2.
g. Obtain the residuals and create a residual plot. Decide whether it is reasonable to consider that the assumptions for regression analysis are met by the variables in questions.
h. At the 5% significance level, do the data provide sufficient evidence to conclude that the slope of the population regression line is not 0 and, hence, that age is useful as a predictor of sales price for Corvettes?
i. Obtain and interpret a 95% confidence interval for the slope, β, of the population regression line that relates age to sales price for Corvettes.
j. Obtain a point estimate for the mean sales price of all 4-year-old Corvettes.
k. Determine a 95% confidence interval for the mean sales price of all 4-year-old Corvettes.
l. Find the predicted sales price of Jack Smith's 4-year-old Corvette.
m. Determine a 95% prediction interval for the sales price of Jack Smith's 4-year-old Corvette.

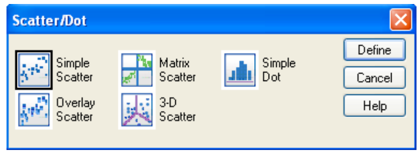
Note that the following steps are not required for all analyses…only perform the necessary steps to complete your problem. Use the above steps as a guide to the correct SPSS steps.

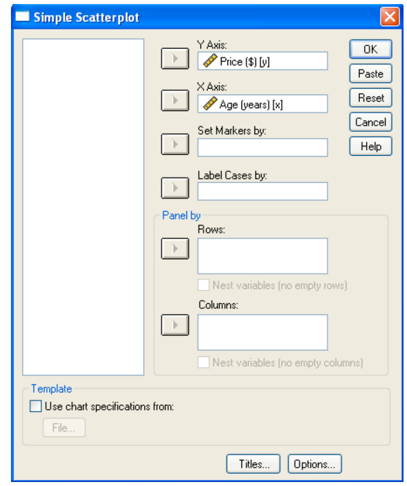
1. Enter the age values into one variable and the corresponding sales price values into another variable *(see figure, below)*.

	x	y
1	6	12500
2	6	11500
3	6	13000
4	4	16000
5	2	21900
6	5	15000
7	4	19000
8	5	16300
9	1	26000
10	2	26000

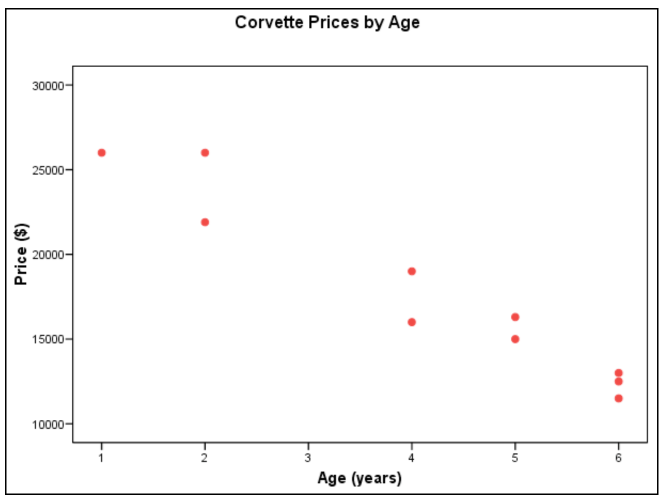
2. Select Graphs → Legacy Dialogs → Scatter/Dot… (select Simple then click the Define button) with the Y Axis variable (Price) and the X Axis variable (Age) entered *(see figures, below)*. Click "Titles…" to enter a descriptive title for your graph, and click "Continue". Click "OK".







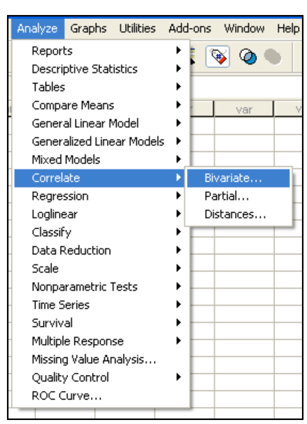
Your output should look similar to the figure below.



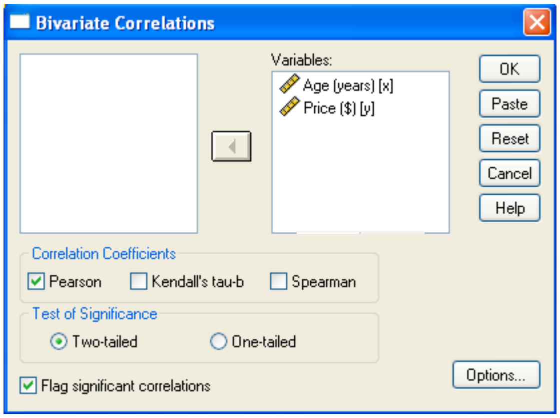
a. Graph the data in a scatterplot to determine if there is a possible linear relationship.

The points seem to follow a somewhat linear pattern with a negative slope.

3. Select Analyze → Correlate → Bivariate… *(see figure, below)*.



4. Select “Age” and “Price” as the variables, select “Pearson” as the correlation coefficient, and click “ “OK” (*see the left figure, below*).



Correlations

		Age (years)	Price ($)
Age (years)	Pearson Correlation	1	-.9679**
	Sig. (2-tailed)		.00000448
	N	10	10
Price ($)	Pearson Correlation	-.9679**	1
	Sig. (2-tailed)	.00000448	
	N	10	10

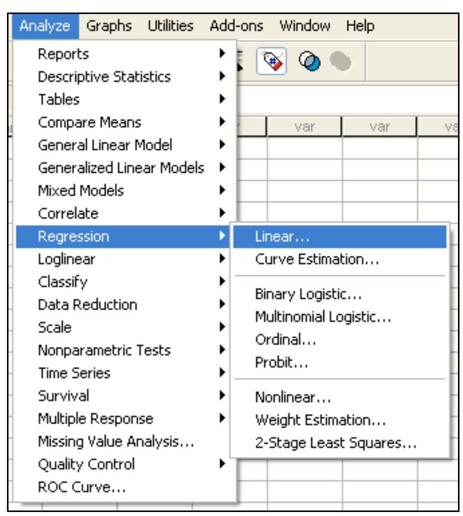
**. Correlation is significant at the 0.01 level (2-tailed).

b. Compute and interpret the linear correlation coefficient, *r*.

The correlation coefficient is –0.9679 *(see the right figure, above)*. This value of *r* suggests a strong negative linear correlation since the value is negative and close to –1. Since the above value of *r* suggests a strong negative linear correlation, the data points should be clustered closely about a negatively sloping regression line. This is consistent with the graph obtained above. Therefore, since we see a strong negative linear relationship between Age and Price, linear regression analysis can continue.

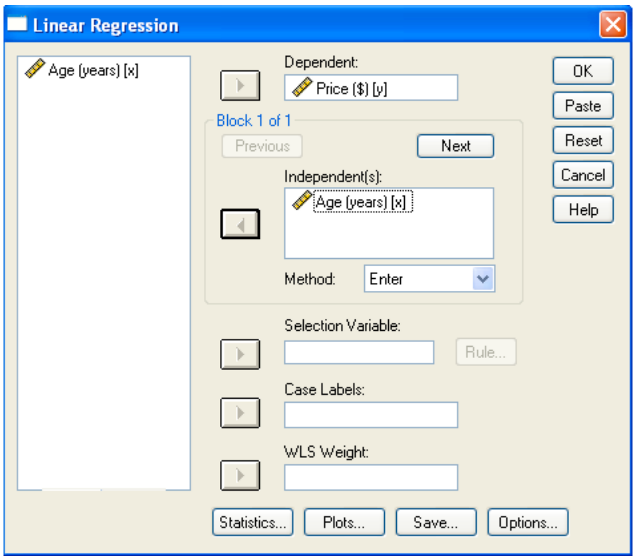
5. Since we eventually want to predict the price of 4-year-old Corvettes (parts j–m), enter the number “4” in the “Age” variable column of the data window after the last row. Enter a “.” for the corresponding “Price” variable value (this lets SPSS know that we want a prediction for this value and not to include the value in any other computations) *(see left figure, below)*.

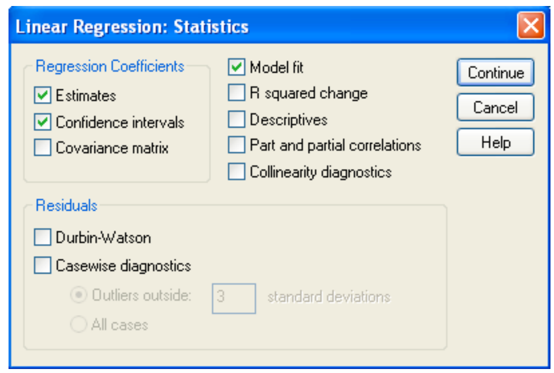
	x	y
1	6	12500
2	6	11500
3	6	13000
4	4	16000
5	2	21900
6	5	15000
7	4	19000
8	5	16300
9	1	26000
10	2	26000
11	4	.

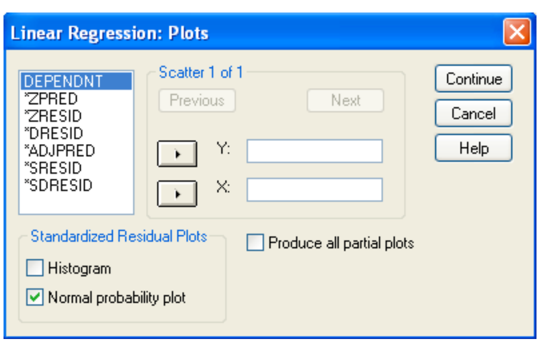


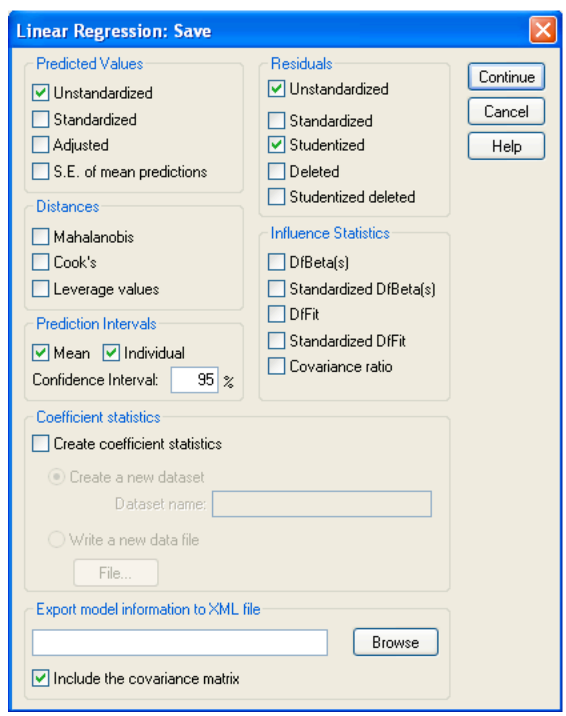
6. Select Analyze → Regression → Linear… *(see right figure, above)*.

7. Select “Price” as the dependent variable and “Age” as the independent variable *(see upper-left figure, below)*. Click “Statistics”, select “Estimates” and “Confidence Intervals” for the regression coefficients, select “Model fit” to obtain r^2, and click “Continue” *(see upper-right figure, below)*. Click “Plots…”, select “Normal Probability Plot” of the residuals, and click “Continue” *(see lower-left figure, below)*. Click “Save…”, select “Unstandardized” predicted values, select “Unstandardized” and “Studentized” residuals, select “Mean” (to obtain a confidence interval…output in the Data Window) and “Individual” (to obtain a prediction interval…output in the Data Window) at the 95% level (or whatever level the problem requires), and click “Continue” *(see lower-right figure, below)*. Click “OK”.









The output from this procedure is extensive and will be shown in parts in the following answers.

c. Determine the regression equation for the data.

Coefficients[a]

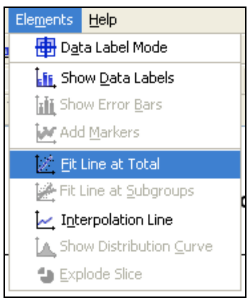
Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	29160.1942	1143.2899		25.5055	.00000001	26523.7629	31796.6254
	Age (years)	-2790.2913	256.2889	-.9679	-10.8873	.00000448	-3381.2946	-2199.2880

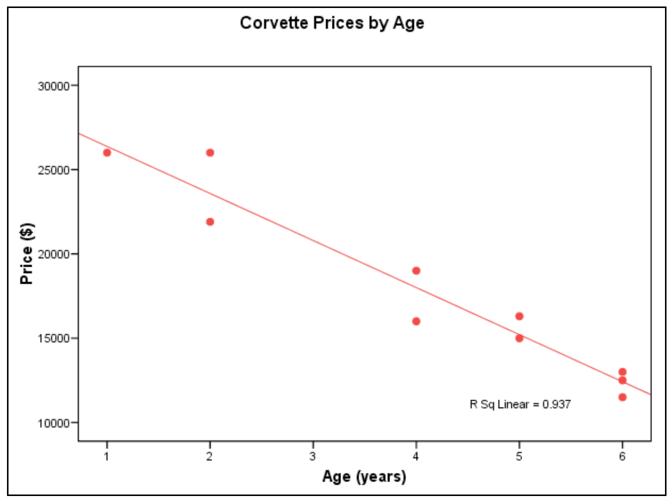
a. Dependent Variable: Price ($)

From above, the regression equation is: Price = 29160.1942 – (2790.2913)(Age).

8. From within the output window, double-click on the scatterplot to enter Chart Editor mode. From the "Elements" menu, select "Fit Line at Total". Click the close box. Now your scatterplot displays the linear regression line computed above.

d. Graph the regression equation and the data points.





e. Identify outliers and potential influential observations.

There do not appear to be any points that lie far from the cluster of data points or far from the regression line; thus there are no possible outliers or influential observations.

f. Compute and interpret the coefficient of determination, r^2.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.9679[a]	.9368	.9289	1424.6529

a. Predictors: (Constant), Age (years)

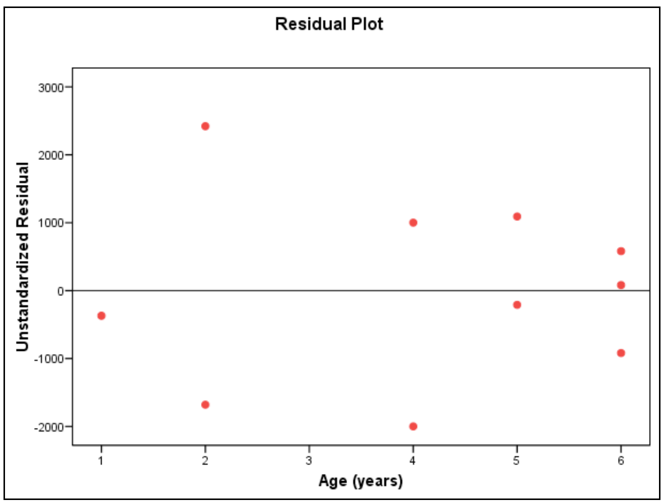
b. Dependent Variable: Price ($)

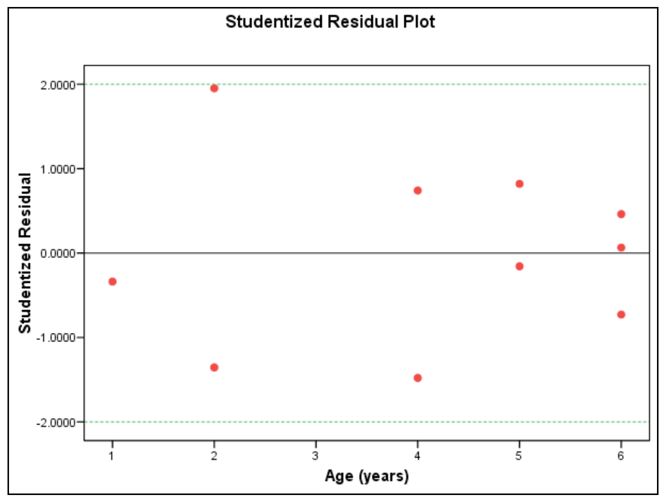
The coefficient of determination is 0.9368; therefore, about 93.68% of the variation in the price data is explained by age. The regression equation appears to be very useful for making predictions since the value of r^2 is close to 1.

9. The residuals and standardized values (as well as the predicted values, the confidence interval endpoints, and the prediction interval endpoints) can be found in the data window.

	x	y	PRE_1	RES_1	SRE_1	LMCI_1	UMCI_1	LICI_1	UICI_1
1	6	12500	12418.4466	81.5534	.0647	10888.6725	13948.2207	8794.4829	16042.4103
2	6	11500	12418.4466	-918.4466	-.7285	10888.6725	13948.2207	8794.4829	16042.4103
3	6	13000	12418.4466	581.5534	.4613	10888.6725	13948.2207	8794.4829	16042.4103
4	4	16000	17999.0291	-1999.0291	-1.4793	16958.4604	19039.5978	14552.9173	21445.1410
5	2	21900	23579.6117	-1679.6117	-1.3548	21961.0824	25198.1409	19917.2981	27241.9252
6	5	15000	15208.7379	-208.7379	-.1567	14041.5998	16375.8759	11722.3192	18695.1565
7	4	19000	17999.0291	1000.9709	.7407	16958.4604	19039.5978	14552.9173	21445.1410
8	5	16300	15208.7379	1091.2621	.8194	14041.5998	16375.8759	11722.3192	18695.1565
9	1	26000	26369.9029	-369.9029	-.3383	24263.7409	28476.0649	22467.4906	30272.3153
10	2	26000	23579.6117	2420.3883	1.9523	21961.0824	25198.1409	19917.2981	27241.9252
11	4	.	17999.0291	.	.	16958.4604	19039.5978	14552.9173	21445.1410

10. To create a residual plot, select Graphs → Legacy Dialogs → Scatter/Dot… (Simple) with the residuals (RES_1) as the Y Axis variable and Age as the X Axis variable. Click "Titles…" to enter "Residual Plot" as the title for your graph, and click "Continue". Click "OK". Double-click the resulting graph in the output window, select "Options" → "Y Axis Reference Line", select the "Reference Line" tab in the properties window, add position of line "0", and click "Apply". Click the close box to exit the chart editor (see left plot, below).



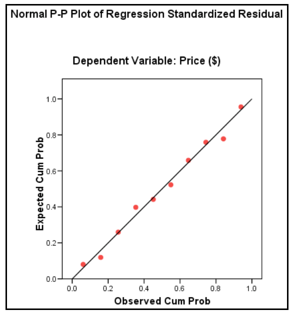


11. To create a studentized residual plot (what the textbook calls a standardized residual plot), select Graphs → Legacy Dialogs → Scatter/Dot… (Simple) with the studentized residuals (SRES_1) as the Y Axis variable and Age as the X Axis variable. Click "Titles…" to enter "Studentized Residual Plot" as the title for your graph, and click "Continue". Click "OK". Double-click the resulting graph in the output window, select "Options" → "Y Axis Reference Line", select the "Reference Line" tab in the properties window, add position of line "0", and click "Apply".

If 2 and/or -2 are in the range covered by the y-axis, repeat the last steps to add a reference line at "2" and "-2" (see right plot, above); any points that are not between these lines are considered potential outliers.

If 3 and/or -3 are in the range covered by the y-axis, repeat the last steps to add a reference line at "3" and "-3"; any points that are beyond these lines are considered outliers.

12. To assess the normality of the residuals, consult the P-P Plot from the regression output.



g. Obtain the residuals and create a residual plot. Decide whether it is reasonable to consider that the assumptions for regression analysis are met by the variables in questions.

The residual plot shows a random scatter of the points (independence) with a constant spread (constant variance). The studentized residual plot shows a random scatter of the points (independence) with a constant spread (constant variance) with no values beyond the ±2 standard deviation reference lines (no outliers). The normal probability plot of the residuals shows the points close to a diagonal line; therefore, the residuals appear to be approximately normally distributed. Thus, the assumptions for regression analysis appear to be met.

h. At the 10% significance level, do the data provide sufficient evidence to conclude that the slope of the population regression line is not 0 and, hence, that age is useful as a predictor of sales price for Corvettes?

Step 1: **Hypotheses**

$H_0: \beta = 0$ (Age is not a useful predictor of price.)

$H_a: \beta \neq 0$ (Age is a useful predictor of price.)

Step 2: **Significance Level**

$\alpha = 0.05$

Step 3: **Critical Value(s) and Rejection Region(s)**

Reject the null hypothesis if p-value ≤ 0.05.

Step 4: **Test Statistic** *(choose either the T-test method or the F-test method…not both)*

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B Upper Bound
1	(Constant)	29160.1942	1143.2899		25.5055	.00000001	26523.7629	31796.6254
	Age (years)	-2790.2913	256.2889	-.9679	-10.8873	.00000448	-3381.2946	-2199.2880

a. Dependent Variable: Price ($)

$T = -10.8873$, and p-value = 0.00000448

ANOVA[b]

Model		df	Sum of Squares	Mean Square	F	Sig.
1	Regression	1	240578912.6214	240578912.6214	118.5330	.00000448[a]
	Residual	8	16237087.3787	2029635.9223		
	Total	9	256816000.0000			

a. Predictors: (Constant), Age (years)

b. Dependent Variable: Price ($)

$F = 118.5330$, and p-value = 0.00000448

Step 5: **Conclusion**

Since p-value = 0.00000448 ≤ 0.05, we shall reject the null hypothesis.

Step 6: **State conclusion in words**

At the $\alpha = 0.05$ level of significance, there exists enough evidence to conclude that the slope of the population regression line is not zero and, hence, that age is useful as a predictor of price for Corvettes.

i. Obtain and interpret a 95% confidence interval for the slope, β, of the population regression line that relates age to sales price for Corvettes.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	29160.1942	1143.2899		25.5055	.00000001	26523.7629	31796.6254
	Age (years)	-2790.2913	256.2889	-.9679	-10.8873	.00000448	-3381.2946	-2199.2880

a. Dependent Variable: Price ($)

We are 95% confident that the slope of the true regression line is somewhere between –3381.2946 and –2199.2880. In other words, we are 95% confident that for every year older Corvettes get, their average price decreases somewhere between $3,381.2946 and $2,199.2880.

j. Obtain a point estimate for the mean sales price of all 4-year-old Corvettes.

	x	y	PRE_1	RES_1	SRE_1	LMCI_1	UMCI_1	LICI_1	UICI_1
11	4	.	17999.0291	.	.	16958.4604	19039.5978	14552.9173	21445.1410

The point estimate (PRE_1) is 17999.0291 dollars ($17,999.0291).

k. Determine a 95% confidence interval for the mean sales price of all 4-year-old Corvettes.

	x	y	PRE_1	RES_1	SRE_1	LMCI_1	UMCI_1	LICI_1	UICI_1
11	4	.	17999.0291	.	.	16958.4604	19039.5978	14552.9173	21445.1410

We are 95% confident that the mean sales price of all four-year-old Corvettes is somewhere between $16,958.4604 (LMCI_1) and $19,039.5978 (UMCI_1).

l. Find the predicted sales price of Jack Smith's selected 4-year-old Corvette.

	x	y	PRE_1	RES_1	SRE_1	LMCI_1	UMCI_1	LICI_1	UICI_1
11	4	.	17999.0291	.	.	16958.4604	19039.5978	14552.9173	21445.1410

The predicted sales price is 17999.0291 dollars ($17,999.0291).

m. Determine a 95% prediction interval for the sales price of Jack Smith's 4-year-old Corvette.

	x	y	PRE_1	RES_1	SRE_1	LMCI_1	UMCI_1	LICI_1	UICI_1
11	4	.	17999.0291	.	.	16958.4604	19039.5978	14552.9173	21445.1410

We are 95% certain that the individual sales price of Jack Smith's Corvette will be somewhere between $14,552.9173 (LICI_1) and $21,445.1410 (UICI_1).