CHAPTER 10

Regression and Correlation

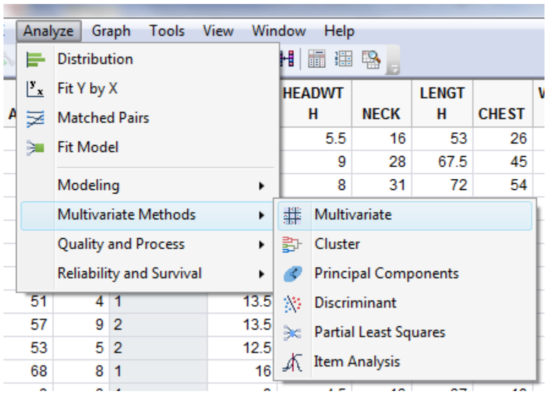
In this Chapter we assess the strength of the linear relationship between two continuous variables. If a significant linear relationship is found, the next step would be to set up a linear equation that describes the relationship between these variables for the purposes of explanation and prediction. Then, we will compute a prediction interval for the dependent variable, given a certain value for the independent variable. This chapter discuss simple linear regression for the purpose of this course.

Class Example 1:

Assess the strength of the relationship between the variables “CHEST” and “WEIGHT”, separate male bears from female bears using the variable “SEX” (Coding: males =1, females = 2). Find the linear equation that describes this relationship.

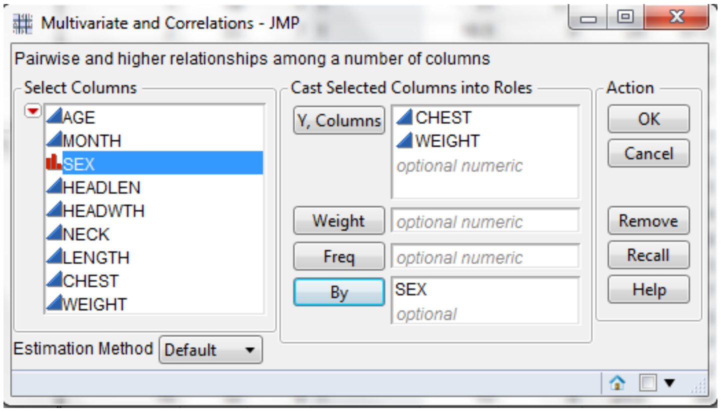
Open the file “Bears” as in Figures 8.1 to 8.2 (or ignore these steps if the file is already open), and change the variable “SEX” to character, as in Figures 8.3 and 8.4. Then, let’s compute the correlation coefficient by selecting “Multivariate Methods” from the “Analyze” menu.

Figure 10.1



You can select many variables, and you will get all pairwise correlations between these variables. Let’s select the variables “CHEST” and “WEIGHT”, and click over “Y, Columns”, then select “SEX” and click over “BY” as follows,

Figure 10.2



Multivariate and Correlations - JMP

Pairwise and higher relationships among a number of columns

Select Columns: AGE, MONTH, SEX, HEADLEN, HEADWTH, NECK, LENGTH, CHEST, WEIGHT

Cast Selected Columns into Roles

Y, Columns: CHEST, WEIGHT, optional numeric

Weight: optional numeric

Freq: optional numeric

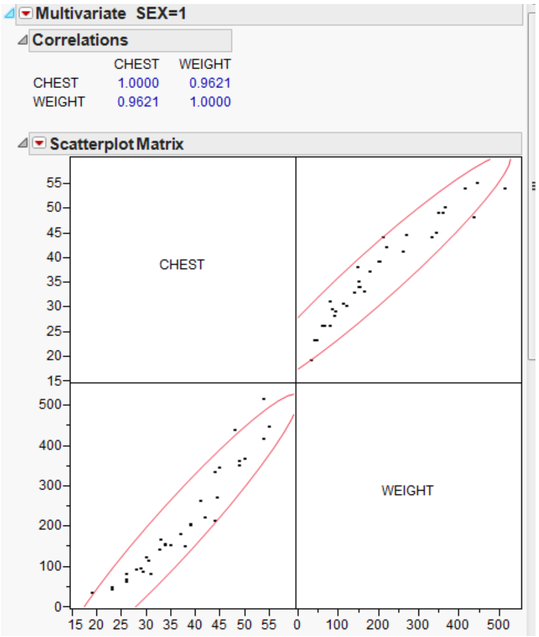
By: SEX, optional

Estimation Method: Default

click over “OK” and you will get the following results

For males:

Figure 10.3



Multivariate SEX=1

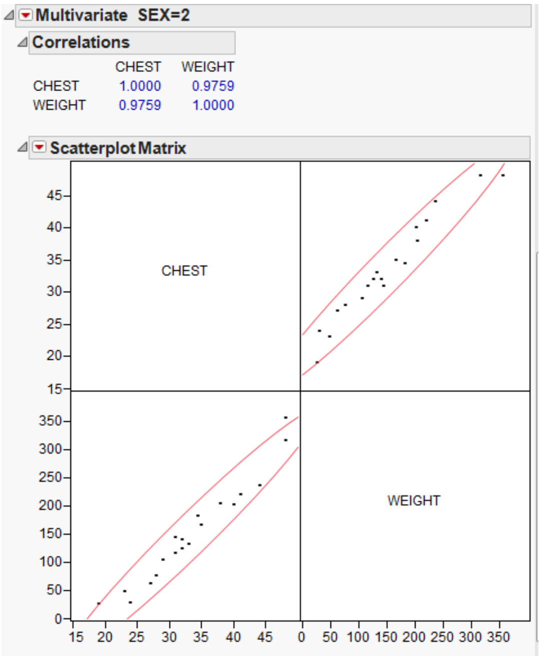
Correlations

	CHEST	WEIGHT
CHEST	1.0000	0.9621
WEIGHT	0.9621	1.0000

Scatterplot Matrix

the scatter plot for these two variables is shown, and the linear correlation coefficient for male bears is *r* = 0.9621. You can repeat the procedure for females and the results are as follows:

Figure 10.4



Multivariate SEX=2

Correlations

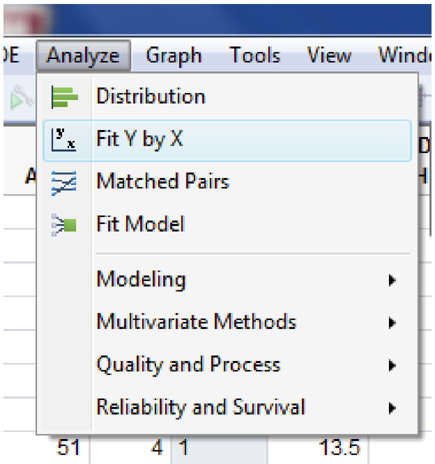
	CHEST	WEIGHT
CHEST	1.0000	0.9759
WEIGHT	0.9759	1.0000

For females, the linear correlation coefficient is 0.9759 as can be seen on Figure 10.4. This procedure can be extended to include several variables where you can get all correlation coefficients for every pair of variables. The arrangement of correlation coefficients in a matrix form, is called the correlation matrix.

Next, we are interested to compute the linear regression equation to describe this relationship. The procedure is as follows:

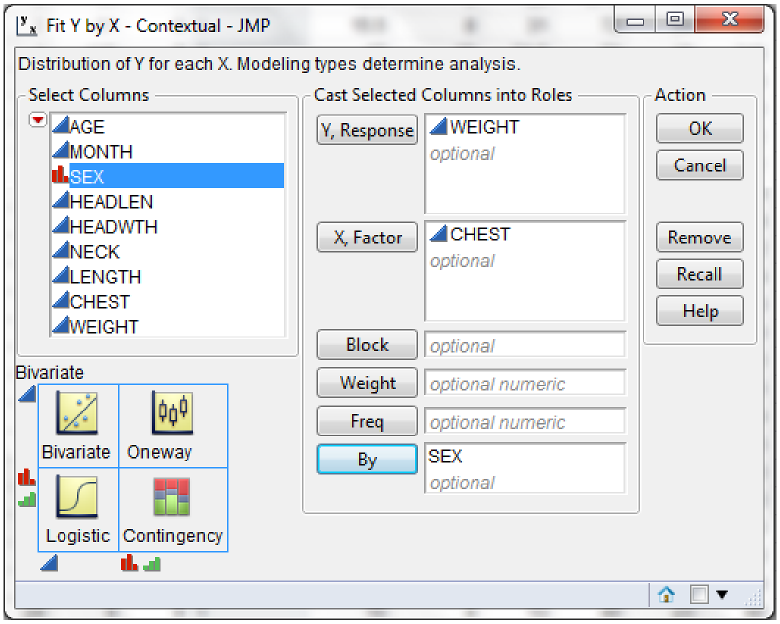
select “Fit Y by X” from the “Analyze” menu as shown below,

Figure 10.5



select the variable “CHEST” and click over “X, Factor”, select the variable “WEIGHT” and click over “Y, Response” , select variable “SEX” and click over “BY” as shown below

Figure 10.6



Fit Y by X - Contextual - JMP

Distribution of Y for each X. Modeling types determine analysis.

Select Columns: AGE, MONTH, SEX, HEADLEN, HEADWTH, NECK, LENGTH, CHEST, WEIGHT

Cast Selected Columns into Roles

Y, Response: WEIGHT

X, Factor: CHEST

Block: optional

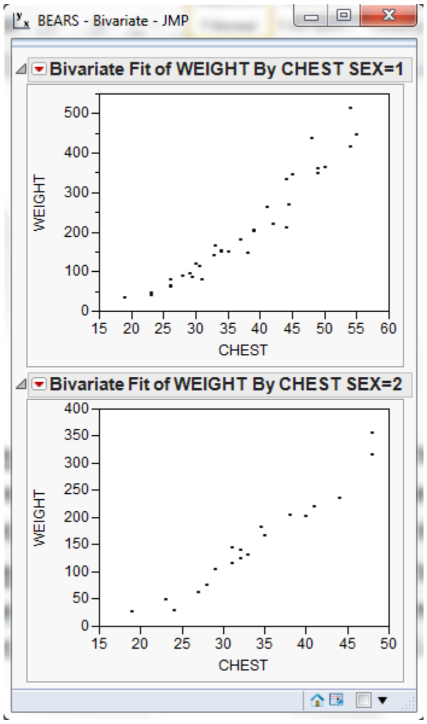
Weight: optional numeric

Freq: optional numeric

By: SEX

click over “OK”, and you get the following graphs,

Figure 10.7



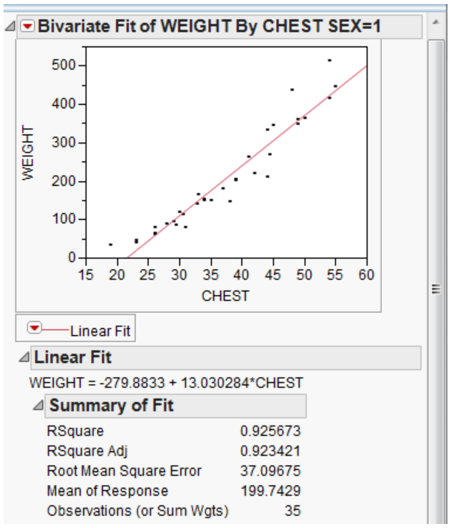
BEARS - Bivariate - JMP

Bivariate Fit of WEIGHT By CHEST SEX=1

Bivariate Fit of WEIGHT By CHEST SEX=2

As you can see on previous scatter plots, there is a linear trend among these variables, then click over the red triangle, and select “Fit Line”, you will get the following results for “SEX” = 1 (males)

Figure 10.8



Linear Fit

WEIGHT = -279.8833 + 13.030284*CHEST

Summary of Fit

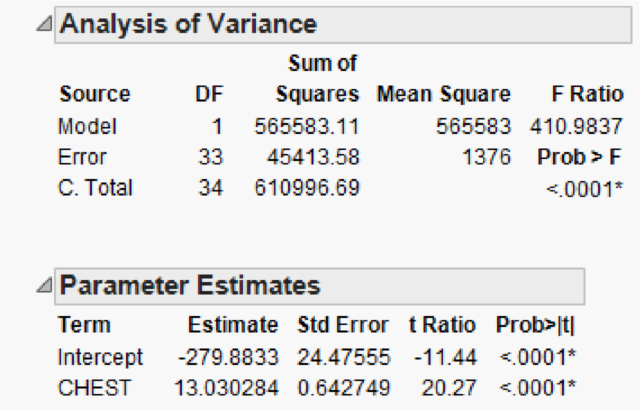
RSquare	0.925673
RSquare Adj	0.923421
Root Mean Square Error	37.09675
Mean of Response	199.7429
Observations (or Sum Wgts)	35

you get the regression line and also, you can see the regression equation expressed as

$$\text{WEIGHT} = -279.8833 + 13.030284*\text{CHEST} \qquad (10.1)$$

The R^2,or coefficient of determination is 0.925673, this number can also be expressed as a percentage as 92.57%. This number shows the proportion of variability on the dependent variable that can be explained using the independent variable. In this case, most of the variability (92.57%) on WEIGHT can be explained using the variable CHEST. The “Root Mean Square Error” is a quantity that measures the deviations around the regression line in a similar way that the standard deviation measures the deviation around the mean. The next part of the computer output is as follows:

Figure 10.9



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	565583.11	565583	410.9837
Error	33	45413.58	1376	Prob > F
C. Total	34	610996.69		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-279.8833	24.47555	-11.44	<.0001*
CHEST	13.030284	0.642749	20.27	<.0001*

The table entitled “Analysis of Variance” compares the explained variance (Model) with the unexplained variance (Error). Please notice that if you take the square root of the “Mean Square Error” term (shown at the ANOVA table as the number 1376), it is equal to the term labeled as “Root Mean Square Error” on Figure 10.8 (√(1376) = 37.09675). Also, if you divide the model sum of squares by the total sum of squares you get:

$$\frac{565583.11}{610996.69} = 0.9257$$

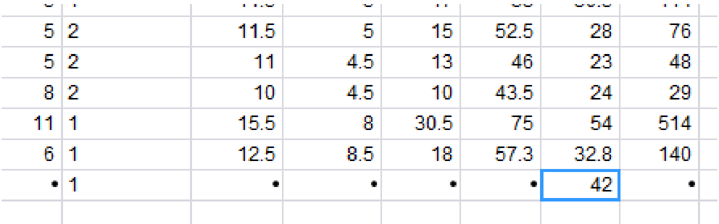
This is R^2, as shown before. This computation fits the definition of R^2, which is the ratio of the explained variability to the total variability. The last part of the computer output lists the parameter estimates for the linear regression equation, the same numbers are shown at equation 10.1, however in this table we have some additional information. In addition to the parameter estimates (point estimates), the standard error is shown in order to assess the variability associated with the parameter estimates. The test statistic and the corresponding *p*-values are also shown at the same row. This *p*-value allows you to test the null hypothesis that the parameter is equal to zero versus the alternative hypothesis that the parameter is not equal to zero. A small *p*-value (less than α) indicates that the null hypothesis should be rejected. In this case both *p*-values indicate that the intercept and the CHEST coefficient are statistically significant, and therefore, that the null hypotheses that the coefficient is equal to zero should be rejected. Thus, these parameters should be included in the regression equation, and they can be used to explain the linear relationship between the variables CHEST and WEIGHT. The same procedure can be repeated for female bears but this activity is left as an exercise for the student.

PREDICTION

Frequently we want to predict the value of the dependent variable using a particular value for the independent variable. In order to obtain a point estimate, you need to substitute the value of the independent variable in the linear equation and compute the corresponding value for the dependent variable. However, we also usually need a prediction interval to assess the variability of our prediction, in this case we have to use the formulas studied during class time and listed in the textbook. Also, the prediction intervals can be computed in JMP using the following procedure:

Let’s assume that we want to predict the weight for a male bear that has a chest measurement of 42”. First, we have to add the new value of 42” to the dataset, let’s add this value and the code of “1” for the variable “SEX” at the bottom of the list as shown,

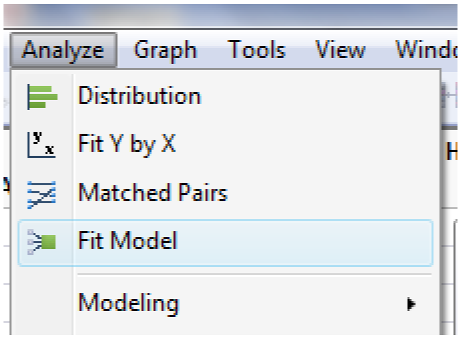
Figure 10.10



5	2		11.5	5	15	52.5	28	76
5	2		11	4.5	13	46	23	48
8	2		10	4.5	10	43.5	24	29
11	1		15.5	8	30.5	75	54	514
6	1		12.5	8.5	18	57.3	32.8	140
•	1		•	•	•	•	42	•

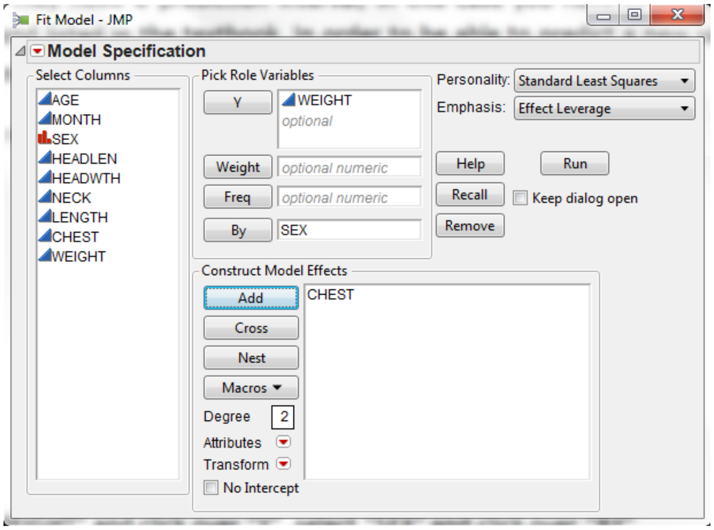
Then , you have to select “Fit Model” from the “Analyze” menu.

Figure 10.11



next, we select “WEIGHT” and click over “Y”, then select “SEX” and click over “BY”, and select “CHEST” and click over “ADD”, then click over “RUN”

Figure 10.12



Y: WEIGHT
optional

Weight: optional numeric

Freq: optional numeric

By: SEX

Personality: Standard Least Squares

Emphasis: Effect Leverage

☐ Keep dialog open

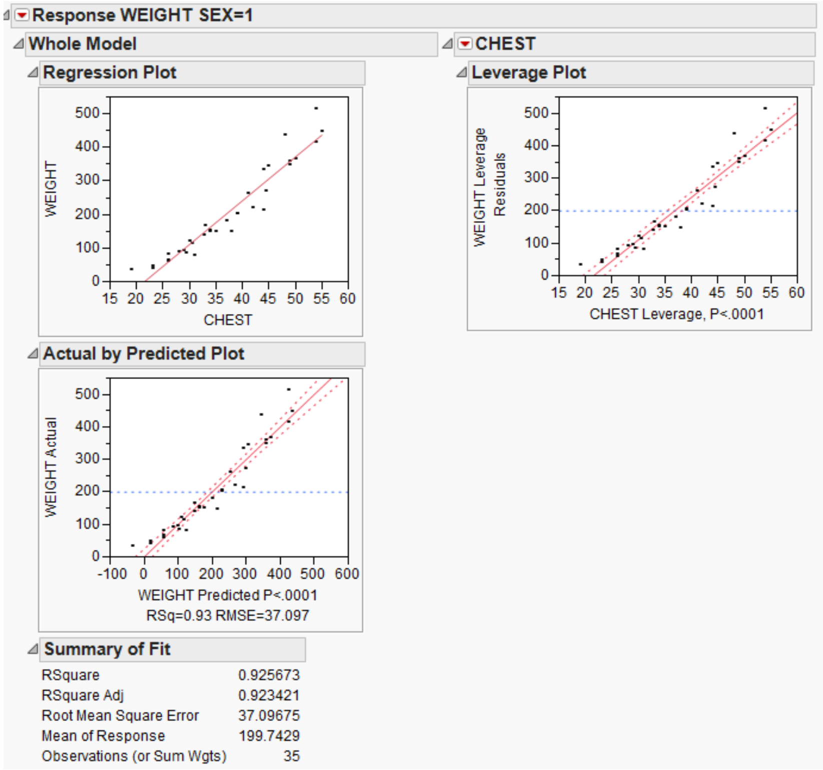
Construct Model Effects: CHEST

Degree: 2

☐ No Intercept

you can see the computer output that matches previous results obtained on Figures 10.8 and 10.9

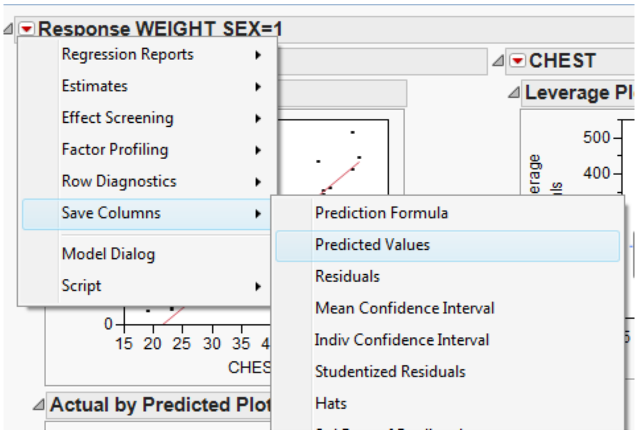
Figure 10.13



RSquare	0.925673
RSquare Adj	0.923421
Root Mean Square Error	37.09675
Mean of Response	199.7429
Observations (or Sum Wgts)	35

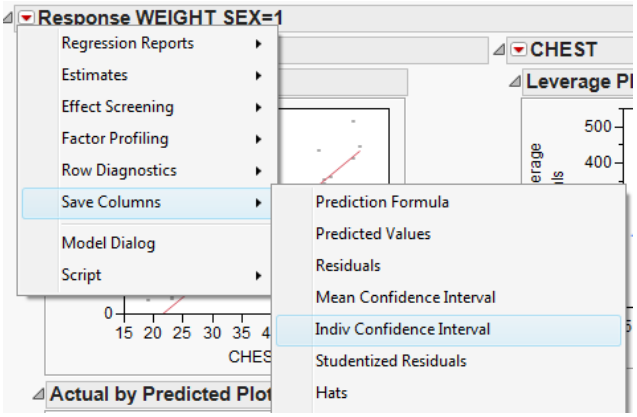
next, click over the red triangle on the upper left corner of the screen and select "Save Columns", then select "Predicted Values" as shown below,

Figure 10.14



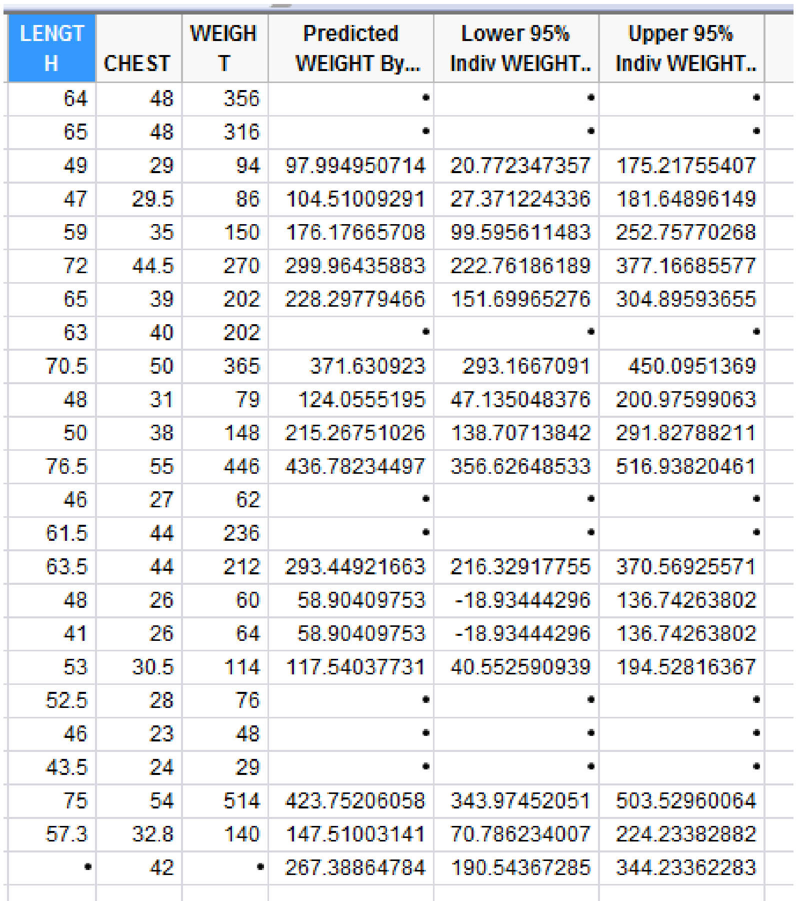
next, we need to obtain the prediction intervals by selecting "Indiv Confidence Interval" from the same menu (Save Columns)

Figure 10.15



Then, you can minimize or close the results window and in the data window you are going to find three additional columns with the information shown below:

Figure 10.16



LENGTH	CHEST	WEIGHT	Predicted WEIGHT By...	Lower 95% Indiv WEIGHT..	Upper 95% Indiv WEIGHT..
64	48	356	•	•	•
65	48	316	•	•	•
49	29	94	97.994950714	20.772347357	175.21755407
47	29.5	86	104.51009291	27.371224336	181.64896149
59	35	150	176.17665708	99.595611483	252.75770268
72	44.5	270	299.96435883	222.76186189	377.16685577
65	39	202	228.29779466	151.69965276	304.89593655
63	40	202	•	•	•
70.5	50	365	371.630923	293.1667091	450.0951369
48	31	79	124.0555195	47.135048376	200.97599063
50	38	148	215.26751026	138.70713842	291.82788211
76.5	55	446	436.78234497	356.62648533	516.93820461
46	27	62	•	•	•
61.5	44	236	•	•	•
63.5	44	212	293.44921663	216.32917755	370.56925571
48	26	60	58.90409753	-18.93444296	136.74263802
41	26	64	58.90409753	-18.93444296	136.74263802
53	30.5	114	117.54037731	40.552590939	194.52816367
52.5	28	76	•	•	•
46	23	48	•	•	•
43.5	24	29	•	•	•
75	54	514	423.75206058	343.97452051	503.52960064
57.3	32.8	140	147.51003141	70.786234007	224.23382882
•	42	•	267.38864784	190.54367285	344.23362283

As you can see at the table above, you can find a prediction interval and a points estimate for every observation listed in the dataset for male bears. Notice there are no predictions for female bears because this results window only included male bears (SEX=1), the prediction intervals for female bears can be done using the same procedure on the next results window.

The weight point estimate prediction for a male bear with a chest measurement of 42” is 267.39 pounds. This computation can also be verified using expression (10.1) as follows:

$$\text{WEIGHT} = -279.8833 + 13.030284*(42) = 267.39$$

The prediction interval for a male bear with a chest measurement of 42” is

$$190.54 \text{ lbs} < \text{Weight} < 344.23 \quad 95\% \text{ PI}$$

This result can also be verified using the prediction formulas seen during class and listed at the textbook. These results complete the procedure to obtain prediction estimates for simple linear regression.

Class Exercises:

1- Regression and Correlation: Open the file “Bears.JMP”, find the correlation coefficient and regression equation for the variables LENGTH and WEIGHT separating males and females. Use WEIGHT as the dependent variable and LENGTH as the independent variable and show the corresponding scatter plots. Predict the weight and obtain a 95% PI for a male bear that measures 62” and for a female bear that measures 57” in length.

2- Hypothesis test for the means: Open the file “Bears.JMP”, find the correlation coefficient and regression equation for the variables HEADLEN and LENGTH separating males and females. Use LENGTH as the dependent variable and show the corresponding scatter plots. Predict the length and obtain a 95% PI for a male bear that has a head length of 14” and for a female bear that has a head length of 12”.

Team Assignment: Regression and Correlation:

Use the random sample that you obtained at the beginning of the semester from the file “Small Town.xls” (do not use the whole dataset, if for some reason you do not have the sample with you, obtain a new sample and save it) and do the following:

1- Find the answers to the following questions:

   a. Find the correlation coefficient and the regression equation for the variables WEIGHT and HEIGHT. Separate the regression models for males and females. Use height as the independent variable. If the correlation is significant, obtain a point estimate and PI for the weight of a male that has a height of 70” and for a female that has a height of 62”.

   b. Find the correlation coefficient and the regression equation for the variables SYS_BP and BMI. Separate the regression models for males and females. Use BMI as the independent variable. If the correlation is significant, obtain a point estimate and PI for the SYS_BP of a male that has a BMI of 24 and for a female that has a BMI of 23.

   c. Find the correlation coefficient and the regression equation for the variables SALARY and CHARITY. Use SALARY as the independent variable. If the correlation is significant, obtain a point estimate and PI for the CHARITY amount that a person contributes if we know that he has a salary of $42,000.

2- Write a report that summarizes your findings.