**2.1.1: SAS: Simple Linear Regression**

Dr. Bean – Stat 5100

**Example:** The Toluca Company makes replacement parts for refrigeration equipment. For a certain part, it takes some time to set up the production process, and then the production of a given lot size can begin. As part of a cost improvement program, the company wished to better understand the relationship between the lot size (X) and the total work hours (Y). Data were reported for 25 representative lots of varying size.

**/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/**

**/\* Input data \*/**

**data toluca; input lotsize workhours @@; cards;**

**80 399 30 121 50 221 90 376 70 361 60 224**

**120 546 80 352 100 353 50 157 40 160 70 252**

**90 389 20 113 110 435 100 420 30 212 50 268**

**90 377 110 421 30 273 90 468 40 244 80 342**

**70 323**

**;**

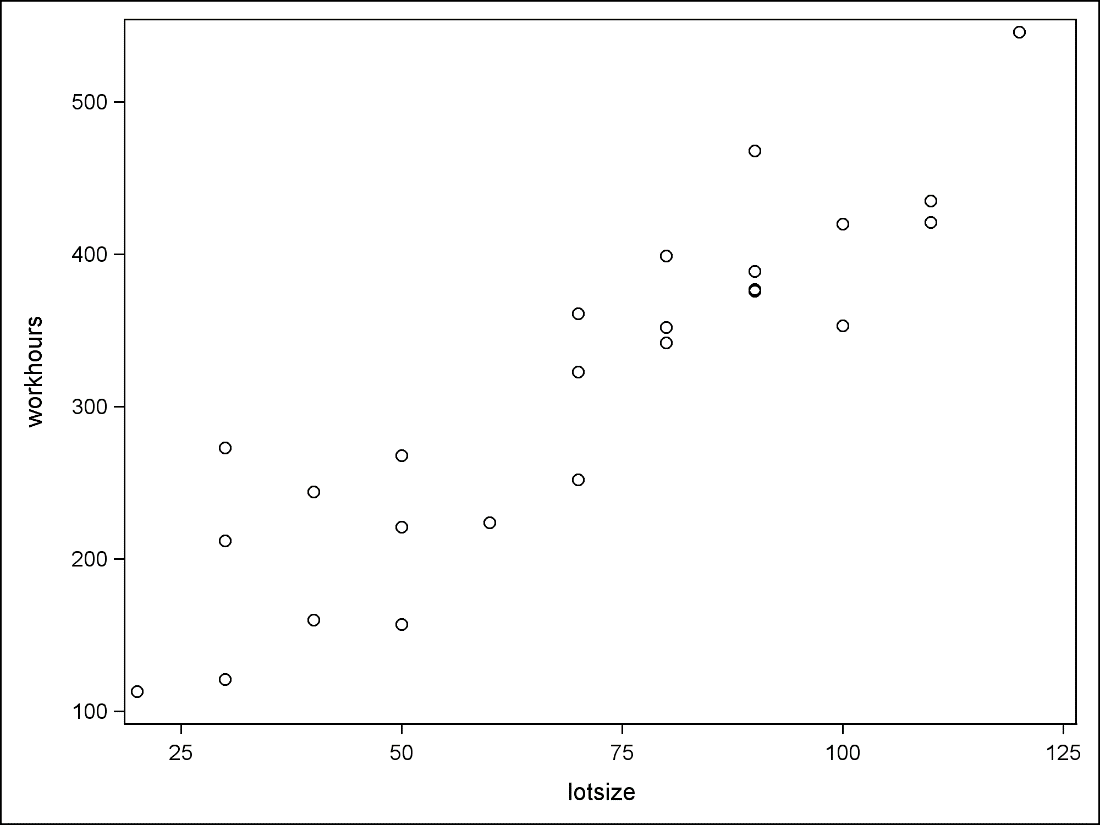
**run;**

**/\* Make a scatterplot of Y=workhours and X=lotsize \*/**

**proc sgplot data=toluca;**

**scatter x=lotsize y=workhours ;**

**run;**



**/\* Be professional -- make it look nice \*/**

**proc sgplot data=toluca;**

**scatter x=lotsize y=workhours /**

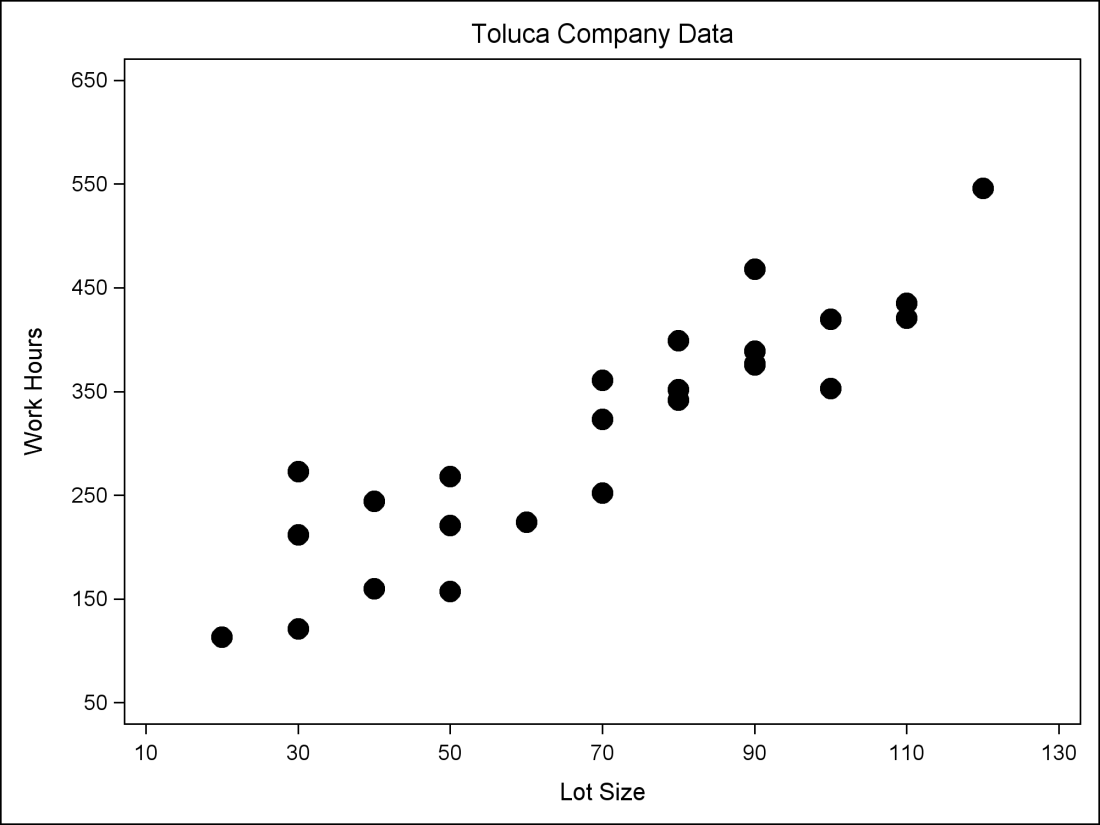
**markerattrs=(symbol=CIRCLEFILLED size=3pt);**

**title1 'Toluca Company Data';**

**xaxis label='Lot Size' values=(10 to 130 by 20);**

**yaxis label='Work Hours' values=(50 to 650 by 100);**

**run;**



**/\* Look at correlation between these variables \*/**

**proc corr data=toluca;**

**var workhours lotsize;**

**run;**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | ***Toluca Company Data*** |   The CORR Procedure   |  |  | | --- | --- | | **2 Variables:** | workhours lotsize |      | **Simple Statistics** | | | | | | | | --- | --- | --- | --- | --- | --- | --- | | **Variable** | **N** | **Mean** | **Std Dev** | **Sum** | **Minimum** | **Maximum** | | **workhours** | 25 | 312.28000 | 113.13764 | 7807 | 113.00000 | 546.00000 | | **lotsize** | 25 | 70.00000 | 28.72281 | 1750 | 20.00000 | 120.00000 |      | **Pearson Correlation Coefficients, N = 25  Prob > |r| under H0: Rho=0** | | | | --- | --- | --- | |  | **workhours** | **lotsize** | | **workhours** | |  | | --- | | 1.00000 | |  | | |  | | --- | | 0.90638 | | <.0001 | | | **lotsize** | |  | | --- | | 0.90638 | | <.0001 | | |  | | --- | | 1.00000 | |  | | |

**/\* Now fit simple linear model with Y=workhours and X=lotsize \*/**

**proc reg data=toluca;**

**model workhours = lotsize;**

**title1 'Simple linear model';**

**run;**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | ***Simple linear model*** |   The REG Procedure  Model: MODEL1  Dependent Variable: workhours   |  |  | | --- | --- | | **Number of Observations Read** | 25 | | **Number of Observations Used** | 25 |      | **Analysis of Variance** | | | | | | | --- | --- | --- | --- | --- | --- | | **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** | | **Model** | 1 | 252378 | 252378 | 105.88 | <.0001 | | **Error** | 23 | 54825 | 2383.71562 |  |  | | **Corrected Total** | 24 | 307203 |  |  |  |      |  |  |  |  | | --- | --- | --- | --- | | **Root MSE** | 48.82331 | **R-Square** | 0.8215 | | **Dependent Mean** | 312.28000 | **Adj R-Sq** | 0.8138 | | **Coeff Var** | 15.63447 |  |  |      | **Parameter Estimates** | | | | | | | --- | --- | --- | --- | --- | --- | | **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | | **Intercept** | **1** | 62.36586 | 26.17743 | 2.38 | 0.0259 | | **lotsize** | **1** | 3.57020 | 0.34697 | 10.29 | <.0001 |   Panel of fit diagnostics for workhours.  Scatterplot of workhours by lotsize overlaid with the fit line, a 95% confidence band and lower and upper 95% prediction limits. |

**/\* See predicted values \*/**

**proc reg data=toluca noprint;**

**model workhours = lotsize;**

**output out=PredictedValues p=Predict;**

**proc print data=PredictedValues;**

**title1 'Predicted Values';**

**run;**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | ***Predicted Values*** |  | **Obs** | **lotsize** | **workhours** | **Predict** | | --- | --- | --- | --- | | **1** | 80 | 399 | 347.982 | | **2** | 30 | 121 | 169.472 | | **3** | 50 | 221 | 240.876 | | **…** |  |  |  | | **24** | 80 | 342 | 347.982 | | **25** | 70 | 323 | 312.280 | |