Statistics: Continuous Methods

STAT452/652, Spring 2013

Computer Lab 5

Tuesday, April 9, 2013
DMS, 106
1:00-2:15PM

SIMPLE LINEAR REGRESSION

with

Minitab®

Instructor: Ilya Zaliapin

Topic: Simple Linear Regression

Goal: Learn how to perform simple linear regression analysis and interpret its results.

Assignments:

1) Download the data set **Bears_lab5.MTW** from the course web site to your Minitab session; it contains selected columns from the data file **Bears.MTW.** Use the Minitab help to learn more about the measurements in this data set.
2) Use scatterplot to analyze the association between the bears' lengths and weights. Choose a data transformation that will make the association approximately linear and homoskedastic.
3) Perform linear regression analysis with the bear's weight as response (dependent variable) and the bear's length as predictor (independent variable) using the data transformation found in 2):
   a. Find the fitted regression line;
   b. Check the residuals' mean, homoskedasticity, and normality;
   c. Find the coefficient of determination, regression sum of squares, error sum of squares, total sum of squares, and error sample standard deviation; discuss and interpret the values;
   d. Find the standard deviations and P-values for the fitted regression coefficients, decide whether the coefficients are significant;
   e. Discuss unusual observations.
4) Find the CI and PI for the bear's weight at the mean bear's length;
5) Perform correlation analysis of the regression residuals with the bears' head, neck, and chest measurements; discuss and interpret the results.

Report:

A printed report for this Lab is due on Thursday, April 18 in class. BW printouts are OK. Reports will not be accepted by mail.

1. Introduction

In applications, it is often important to forecast the value of a random variable *Y* (which can be difficult or impossible to measure directly) using the observations of another random variable *X* (which can be more accessible for direct measurements). An important result of the probability theory is that when the two random variables are normal the forecast of *Y* in terms of *X* that minimizes the mean-square error is linear:

$$\hat{Y} = a + bX \,.$$

Often, even for non-Normal random variables, a practically useful forecast can be obtained using a linear function of the predictor *X*. This explains the importance of linear regression analysis, one of the most common tools of applied statistics.

2. Simple Linear Regression

[This section is not intended to provide all the necessary theoretical background for regression analysis; refer to the text or lecture notes for the missing details.]

Recall that a linear regression (LR) model with one dependent variable Y and one independent variable X (also called *simple linear regression*, SLR) takes the form

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \,, \tag{1}$$

where (Y_i, X_i) are *observations*, β_0 and β_1 are *regression coefficients*, and ε_i are *model errors*. We assume that the errors have zero expected value. The regression analysis is focused on estimating the regression coefficients, making statistical inference about these estimations, and constructing forecasts for the new values of Y, given the new value of X. The estimation is done using the mean-square error minimization; that is the fitted regression coefficients are chosen to minimize the sum of the squared residuals:

$$\sum_{i=1}^{n} e_i^{\,2} = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_1 \right)^2 \to \min \,. \tag{2}$$

Remark: The residuals

$$e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$$

are *not* equal to the errors ε_i of the model (1), since they depend on the estimated, not true, values of the regression coefficients.

The estimated (fitted) regression coefficients in (1) can be found using the standard formulas (see the textbook or lecture notes) using the criterion (2) without any additional assumptions. However, the standard methods of regression *inference* (deriving the distribution of the estimations, testing their significance, finding the distribution for the forecast of Y) are based on additional assumption about the model errors:

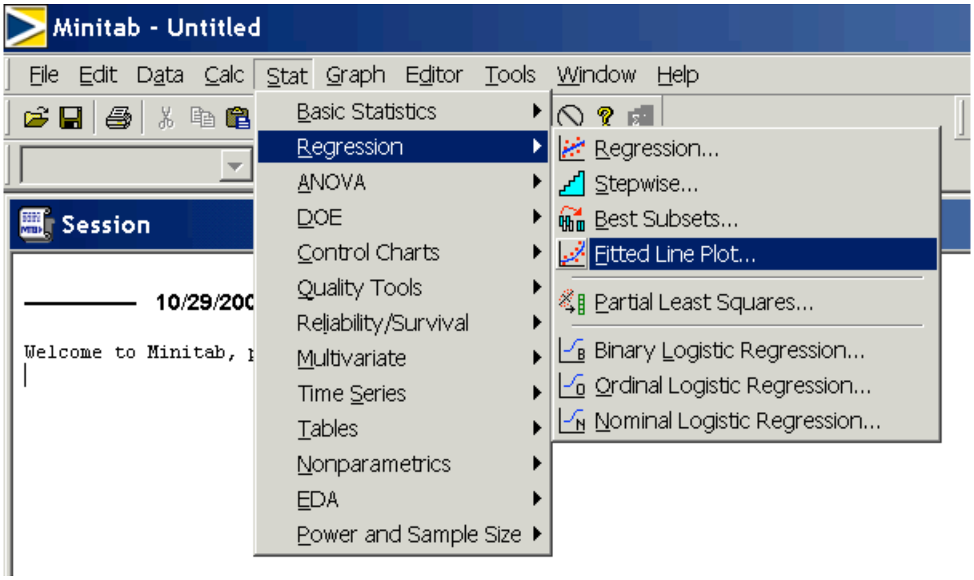
$$\varepsilon_i \text{ are iid } N(0,\sigma^2)\,. \tag{3}$$

In particular, a model with errors of the same variance (not necessarily independent) is called *homoskedastic*; otherwise it is *heteroskedastic*.

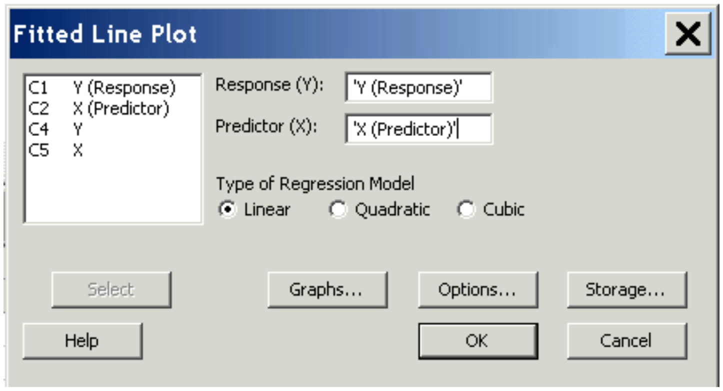
We will consider two ways of performing simple linear regression analysis. The first method is good for presentation purposes (obtaining a quick visual summary of the results), while the second one focuses on technical details.

2.1 SLR: Presentation of Results

A useful summary of regression results can be obtained using the menu **Stat/Regression/Fitted Line Plot…**



… which brings up the following submenu


Fitted Line Plot

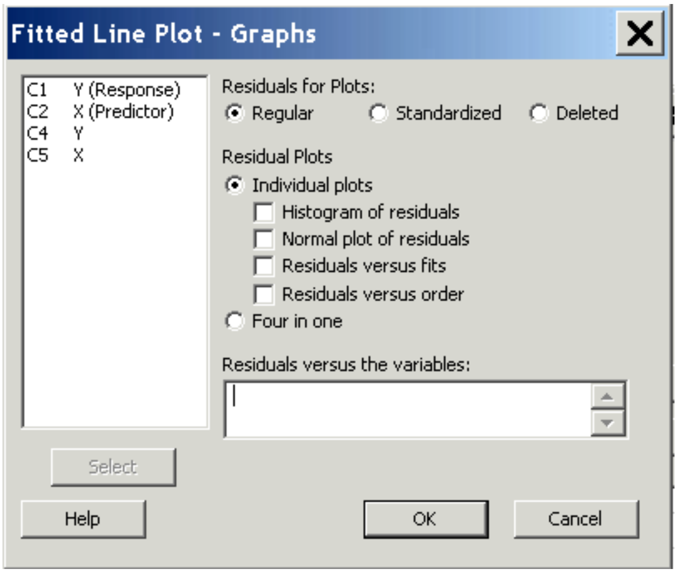
C1 Y (Response)
C2 X (Predictor)
C4 Y
C5 X

Response (Y): 'Y (Response)'

Predictor (X): 'X (Predictor)'

Type of Regression Model
● Linear ○ Quadratic ○ Cubic

For now, we only consider **Linear** type of regression model. The other three sub-submenus to use are **Graphs, Options, and Storage:**



Fitted Line Plot - Graphs

C1 Y (Response)
C2 X (Predictor)
C4 Y
C5 X

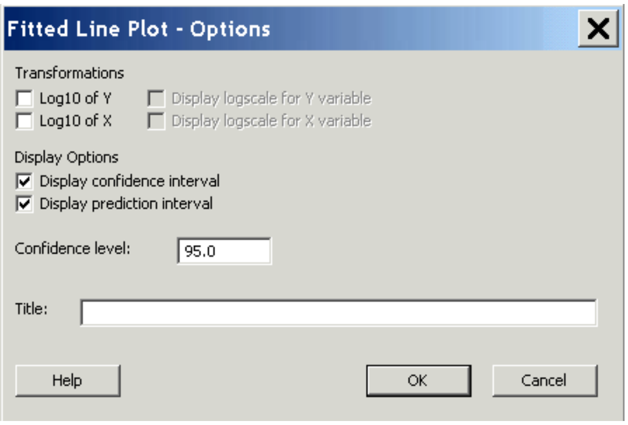
Residuals for Plots:
● Regular ○ Standardized ○ Deleted

Residual Plots
● Individual plots
□ Histogram of residuals
□ Normal plot of residuals
□ Residuals versus fits
□ Residuals versus order
○ Four in one

Residuals versus the variables:

Minitab can plot the regression residuals, which is useful for checking the model assumptions. In **Graphs** you choose what type of residuals to plot (details will be discussed in class).

In **Options** you specify whether to apply a log-transformation to the data and whether to display confidence and prediction intervals for the regression.



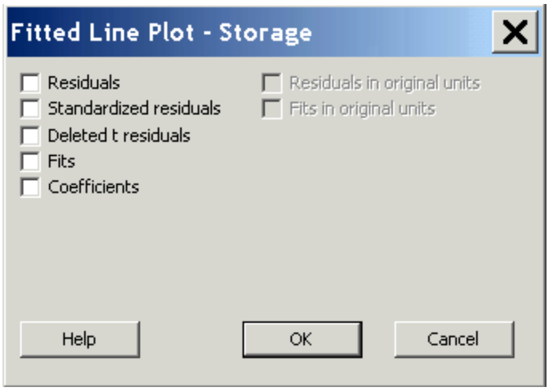
Fitted Line Plot - Options

Transformations
□ Log10 of Y □ Display logscale for Y variable
□ Log10 of X □ Display logscale for X variable

Display Options
☑ Display confidence interval
☑ Display prediction interval

Confidence level: 95.0

Title:

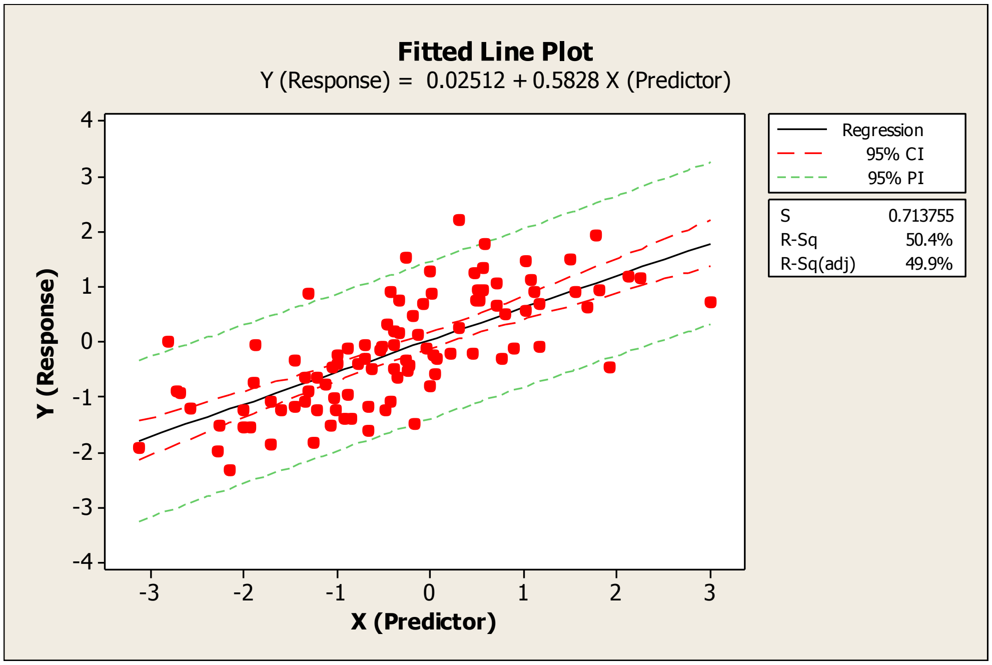


Fitted Line Plot - Storage

□ Residuals □ Residuals in original units
□ Standardized residuals □ Fits in original units
□ Deleted t residuals
□ Fits
□ Coefficients

In **Storage** you choose the outputs to store in the current worksheet.

The regression results are summarized in a graph like this:

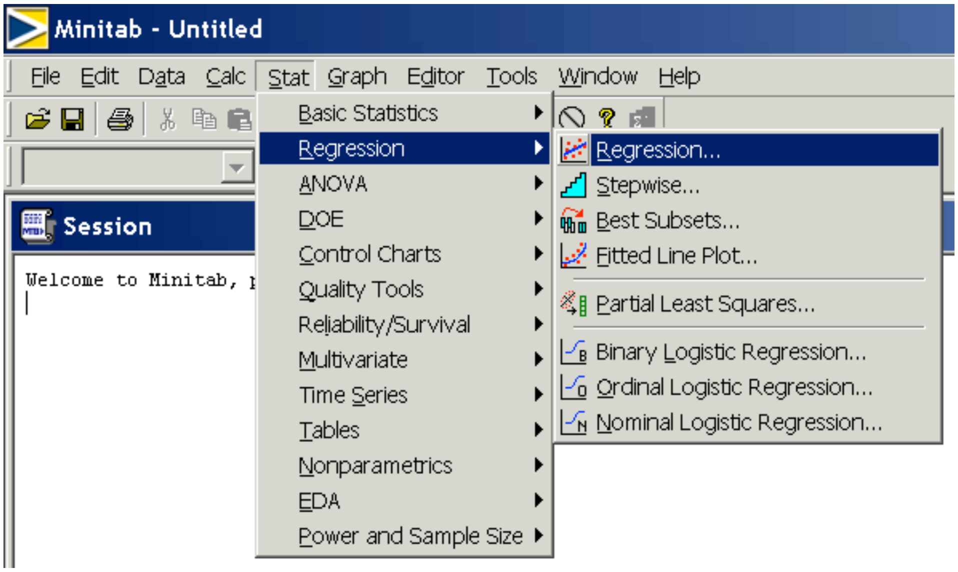


The graph contains the following info:

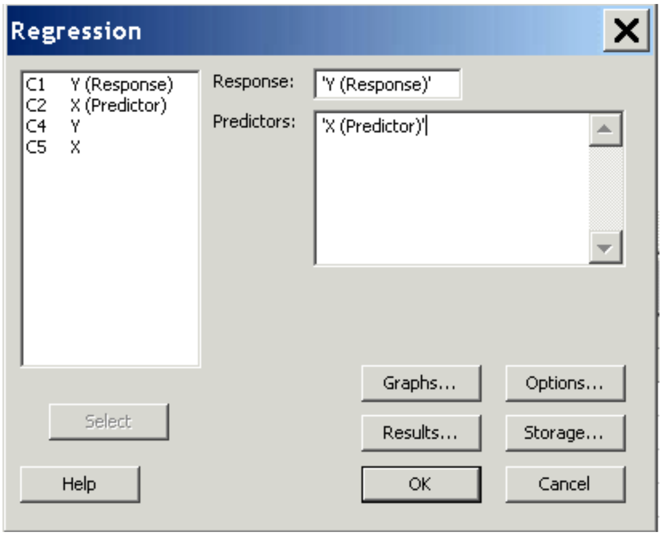
- The fitted regression equation is shown in the title
- Confidence interval (**CI**) is for the fitted values of the regression line. (The position of the regression line is random.)
- Prediction interval (**PI**) is for the actual values of Y: it shows where the new values of Y may fall.
- **S** is the sample standard deviation of residuals
- **R-Sq** is the coefficient of determination (r^2)
- **R-Sq(adj)** is the adjusted coefficient of determination (we are not paying much attention to it in a simple regression).

2.2 SLR: Technical Details

An access to all the technical details of the regression analysis is provided by the menu **Stat/Regression/Regression…**



…which leads to the following submenu

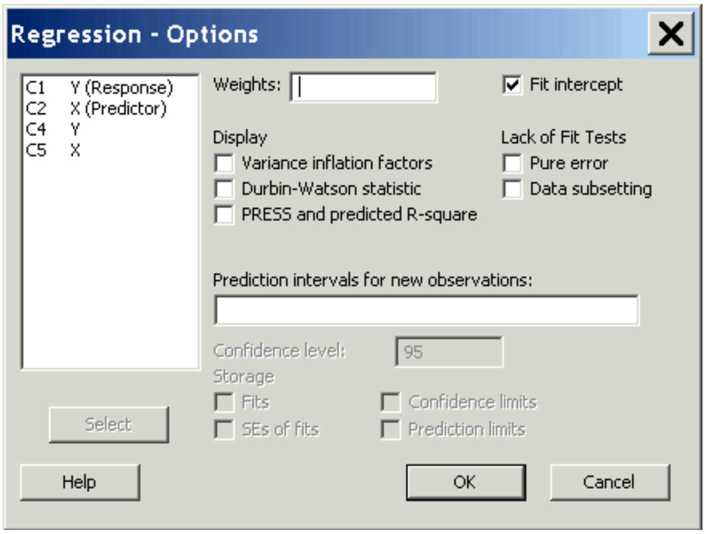


C1	Y (Response)
C2	X (Predictor)
C4	Y
C5	X

Response: 'Y (Response)'

Predictors: 'X (Predictor)'

The sub-submenu **Graphs** is the same as in **Section 2.1.**


Regression - Options

C1 Y (Response)
C2 X (Predictor)
C4 Y
C5 X

Weights:

☑ Fit intercept

Display
□ Variance inflation factors
□ Durbin-Watson statistic
□ PRESS and predicted R-square

Lack of Fit Tests
□ Pure error
□ Data subsetting

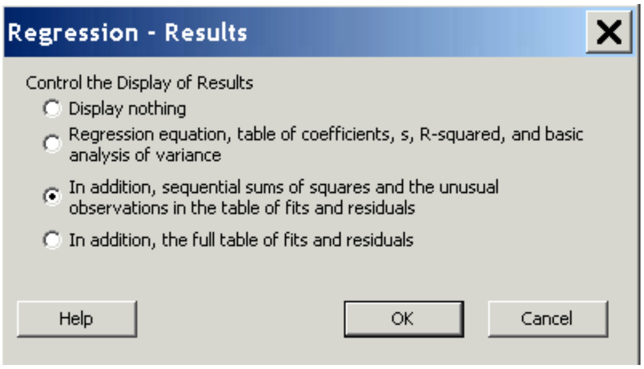
Prediction intervals for new observations:

Confidence level: 95

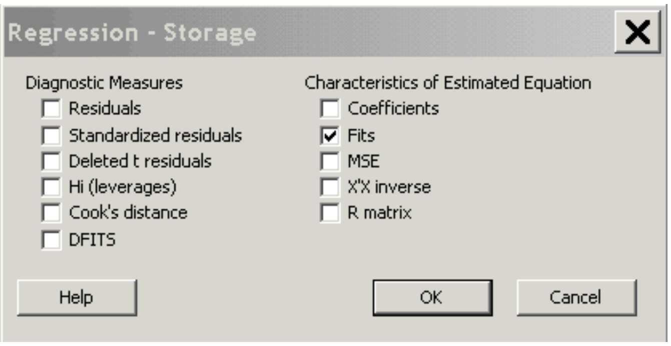
Storage
□ Fits
□ SEs of fits
□ Confidence limits
□ Prediction limits

In **Options**, we are primarily interested in the option "Prediction intervals for new observations". Here, you give the values of X for which the forecast will be constructed.

In **Results,** you choose what results will be displayed in the Session window.


Regression - Results

Control the Display of Results
○ Display nothing
○ Regression equation, table of coefficients, s, R-squared, and basic analysis of variance
● In addition, sequential sums of squares and the unusual observations in the table of fits and residuals
○ In addition, the full table of fits and residuals

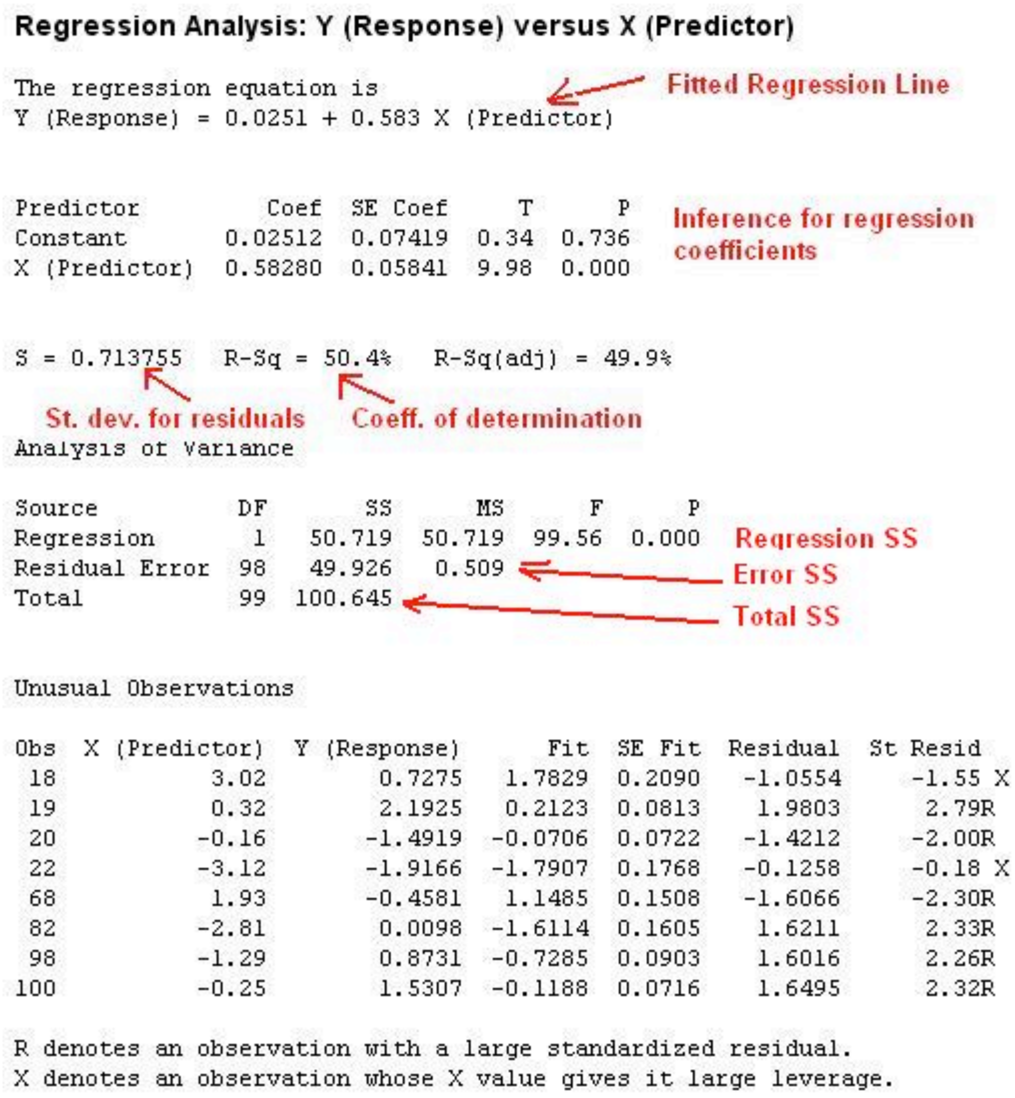

Regression - Storage

Diagnostic Measures
□ Residuals
□ Standardized residuals
□ Deleted t residuals
□ Hi (leverages)
□ Cook's distance
□ DFITS

Characteristics of Estimated Equation
□ Coefficients
☑ Fits
□ MSE
□ X'X inverse
□ R matrix

In **Storage,** you choose which results will be stored in the current worksheet and/or in **Data** storage.

The results of the analysis are displayed in the **Session** window and look like this (the details will be discussed in class):



Regression Analysis: Y (Response) versus X (Predictor)

```
The regression equation is
Y (Response) = 0.0251 + 0.583 X (Predictor)


Predictor          Coef  SE Coef     T      P
Constant        0.02512  0.07419  0.34  0.736
X (Predictor)   0.58280  0.05841  9.98  0.000


S = 0.713755   R-Sq = 50.4%   R-Sq(adj) = 49.9%

Analysis of Variance

Source          DF       SS      MS      F      P
Regression       1   50.719  50.719  99.56  0.000
Residual Error  98   49.926   0.509
Total           99  100.645


Unusual Observations

Obs  X (Predictor)  Y (Response)      Fit  SE Fit  Residual  St Resid
 18           3.02        0.7275   1.7829  0.2090   -1.0554    -1.55 X
 19           0.32        2.1925   0.2123  0.0813    1.9803     2.79R
 20          -0.16       -1.4919  -0.0706  0.0722   -1.4212    -2.00R
 22          -3.12       -1.9166  -1.7907  0.1768   -0.1258    -0.18 X
 68           1.93       -0.4581   1.1485  0.1508   -1.6066    -2.30R
 82          -2.81        0.0098  -1.6114  0.1605    1.6211     2.33R
 98          -1.29        0.8731  -0.7285  0.0903    1.6016     2.26R
100          -0.25        1.5307  -0.1188  0.0716    1.6495     2.32R

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.
```

Fitted Regression Line

Inference for regression coefficients

St. dev. for residuals

Coeff. of determination

Regression SS

Error SS

Total SS