Statistics: Continuous Methods
STAT452/652, Spring 2013

Computer Lab 6
Tuesday, April 23, 2013
DMS, 106
1:00-2:15PM

MULTIPLE LINEAR REGRESSION
with

Minitab®

Instructor: Ilya Zaliapin

Topic: Multiple Linear Regression

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

Assignments:

1) Download the data set **Lab6_data.MTW** from the course web site to your Minitab session; it contains a data set for multiple linear regression analysis.
2) Perform simple linear regression of Y on each of X_i, discuss the results.
3) Perform correlation analysis among the predictors, discuss the results.
4) Find the best multiple linear regression model for predicting Y using X_i.
5) Identify the variables with (i) collinearity, (ii) confounding, and (iii) predictors uncorrelated with response but useful for regression. We do have each case in this data set.

Report:

A printed report for this Lab is due on Tuesday, April 30 in class. BW printouts are OK. Reports will not be accepted by mail.

1. Introduction

Multiple Linear Regression (MLR) analysis is used to predict the values of a variable of interest (response) using *several* other variables (predictors). Hence, MLR is a natural generalization of Simple Liner Regression (SLR) discussed in Lab 5. Despite the fact that many technical details of MLR are very similar to those of SLR, there exist important differences, which are mainly connected to the (possibly) complicated correlation structure of the response and predictors. Two important topics to discuss are *collinearity* and *confounding*. Another noteworthy issue deals with predictors that are uncorrelated with the response although might be useful for prediction.

2. Model, parameter estimation

Multiple Linear Regression (MLR) model is used to predict the values of a *response variable Y* using a linear combination of the values of *predictors* X_i:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \varepsilon_i . \qquad (1)$$

Notably, the predictors can be nonlinear transformations of the *observations,* like in a polynomial model, where we observe only two variables, *Y* and *X*:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^{\,2} + \ldots + \beta_p X_i^{\,p} + \varepsilon_i$$

The estimation of the *regression parameters* β_i is done using the mean-square error minimization:

$$\sum_{i=1}^{n} e_i^{\,2} = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \ldots - \hat{\beta}_p X_{pi} \right)^2 \to \min \qquad (2)$$

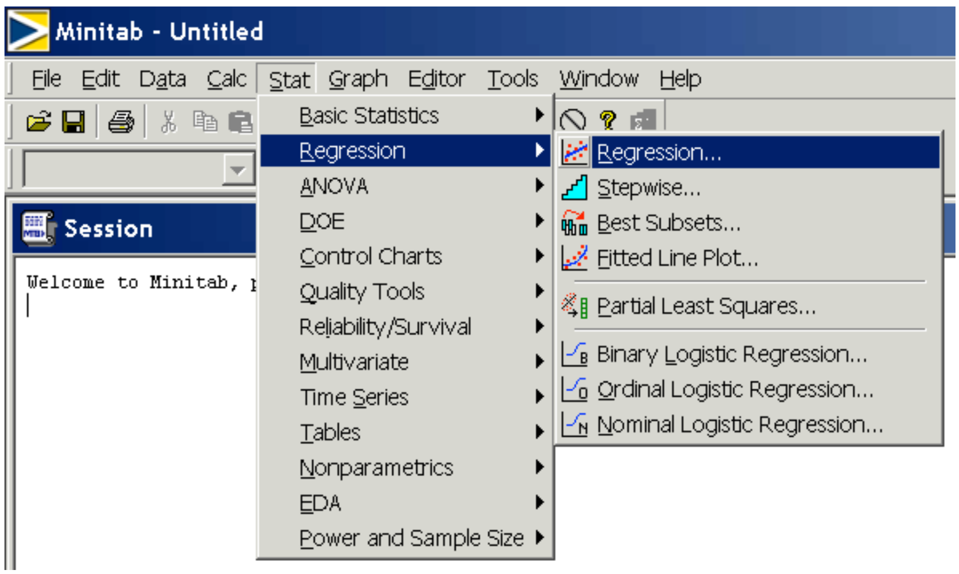
The estimated (fitted) regression coefficients can be found by a statistical package using the criterion (2) without any additional assumptions. However, the standard methods of regression *inference* (deriving the distribution of the estimations, testing their significance, finding the distribution for the forecast of Y, *etc.*) are based on an additional assumption about the model errors:

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

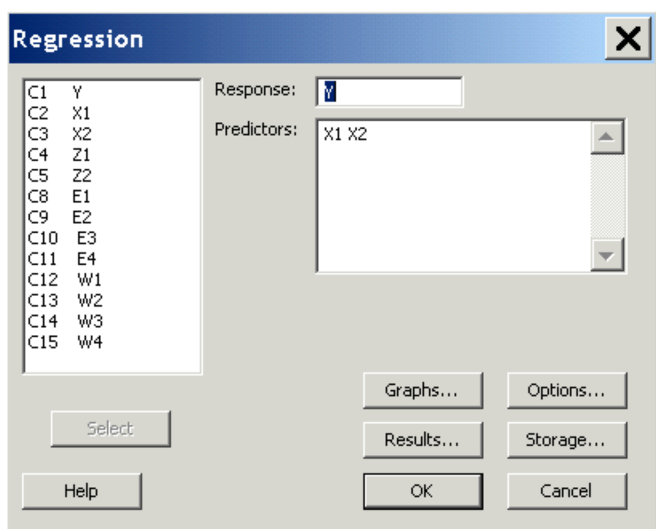
A model with errors of the same variance (not necessarily independent) is called *homoskedastic*; otherwise it is *heteroskedastic*.

3. Data analysis and inference

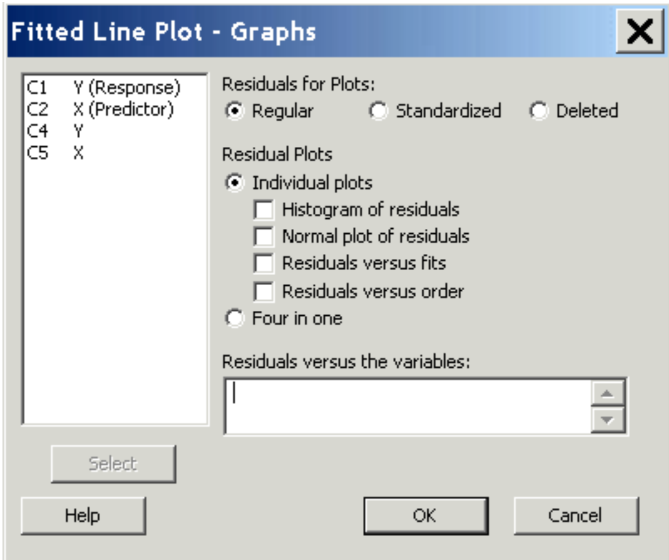
The standard methods of multiple regression analysis are implemented in the menu **Stat/Regression/Regression…**



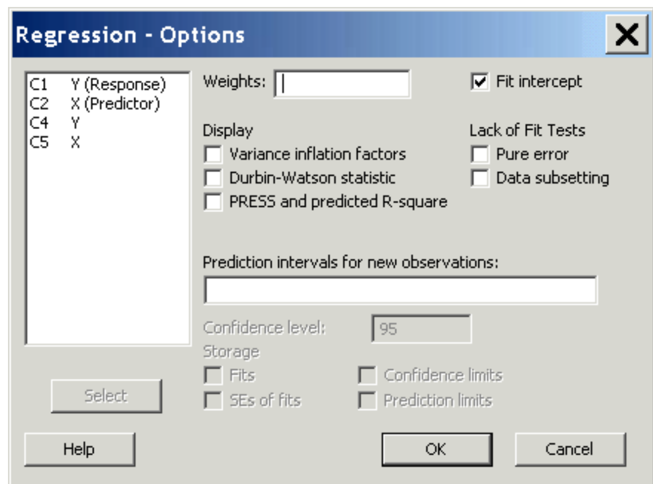
…which leads to the following submenu



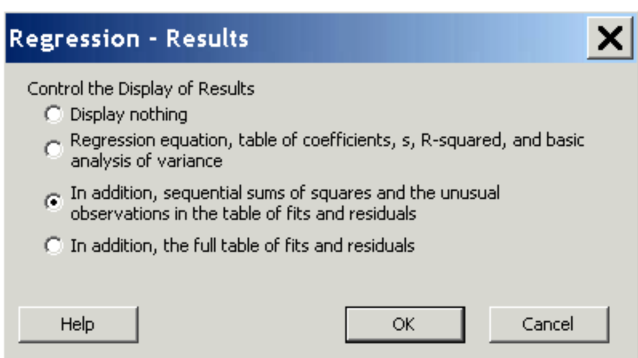
The sub-submenus **Graphs, Options, Results** and **Storage** are the same as for simple linear regression:



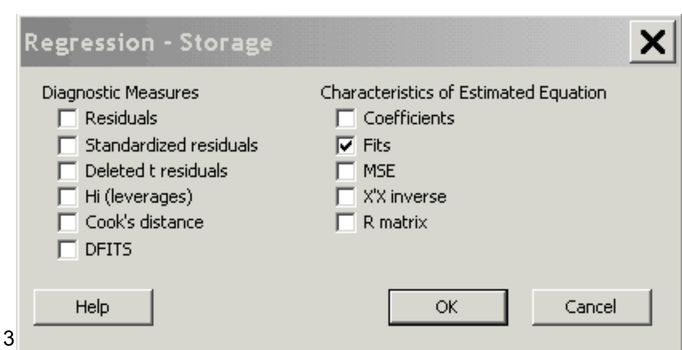
Minitab can plot the regression residuals, which is useful for checking the model assumptions. In **Graphs** you choose what type of residuals to plot (details will be discussed in class).



In **Options**, we are primarily interested in the option "Prediction intervals for new observations". Here, you give the values of X_i for which the forecast will be constructed (the number of values should match the number of predictors)



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



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 versus X1, X2

```
The regression equation is
Y = - 0.028 + 0.936 X1 - 0.0400 X2
```

Fitted Regression Equation

Predictor	Coef	SE Coef	T	P	VIF
Constant	-0.0285	0.1086	-0.26	0.794	
X1	0.9356	0.1507	6.21	0.000	1.745
X2	-0.04003	0.09974	-0.40	0.689	1.745

Estimated parameters

St. dev. of estimated parameters

Standardized value, T-statistics

P-value, the probability to have this estimation in a model where the true value is 0

Variance Inflation Factors, large values indicate collinearity

```
S = 1.07567   R-Sq = 38.9%   R-Sq(adj) = 37.7%
```

St. dev. of residuals

Coefficient of determination, R^2

Adjusted value of coefficient of determination, the proportion of the variance of response explained by predictors

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	71.580	35.790	30.93	0.000
Residual Error	97	112.236	1.157		
Total	99	183.816			

Regression SS

Error SS

Total SS

F statistics and P-value for testing the hypothesis that all the regression coefficients except intercept are zeros

4. Confounding

Confounding is a situation when a significant correlation between *Y* and *X* is explained not by actual physical association between the variables, but by a third variable *Z*, which is actually related to both *Y* and *X.* Recall an example where *Y* is a size of a child vocabulary, *X* is her shoe size, and *Z* is her age. Indeed, Y and X are positively associated (correlated), although there is no relationship between the shoe size and learning new words. The correlation is due to the age.

5. Collinearity

Collinearity (*multicollinearity*) is a situation when several predictors are so highly correlated that we can't decide which one is important in explaining the association among the predictors and response. In this case, the regression coefficients for all collinear predictors might be insignificant, despite the fact that each of them might have a significant correlation with the response.

To detect collinearity we might

a) Perform correlation analysis of the predictors to detect the highly correlated ones;

b) Compute the Variance Inflation Factors (VIF), which estimate by how much the variance of the estimated coefficients is increased due to the correlation among them. Usually, VIF >5 signals a dangerous collinearity.

When collinearity is detected we can

a) Leave for analysis only the predictors that are not highly correlated among each other using the knowledge of the process.

b) Use Stepwise Regression, or Best Subsets Regression Minitab options to select the predictors without collinearity.

c) Use Partial Least Square Regression, which will transform the correlated variables in order to extract the part that causes the correlation.

As always, other approaches are also available to detect and deal with collinearity.

6. Predictors can be uncorrelated with the response

Surprisingly enough, you might encounter a situation when random variables *Y* and *X* are uncorrelated, although including X in the multiple regression model to predict *Y* significantly improves its performance. Consider the following example, where E_i are iid standard Normal rvs:

$$W_1 = E_1 + E_2;\ W_2 = E_2 + E_3;\ W_3 = E_3 + E_4.$$

Clearly, W_1 and W_3 are uncorrelated. However, when we predict W_1 using W_2 and W_3, the latter helps explaining the variance introduced in the model by W_2 and not related to W_1 (the one caused by E_3).