REGRESSION ANALYSIS

What is regression analysis?

Regression analysis is a very popular and very useful statistical method for investigating functional relationships among variables. The relationship is expressed in the form of an equation or a model connecting the *response* or *dependent* variable and one or more *explanatory* or *predictor* variables. We may wish to examine whether cigarette consumption is related to various socioeconomic and demographic variables such as age, education, income, and price of cigarettes. In this example, the response variable is cigarette consumption (measured by the number of packs of cigarette sold) and the explanatory or predictor variables are the various socioeconomic and demographic variables.

We denote the response variable by Y and the set of predictor variables by X_1, X_2,…, X_p, where p denotes the number of predictor variables. The true relationship between Y and the predictors X_1, X_2,…, X_p can be approximated by the regression model

$$Y = f\left(X_1, X_2, ..., X_p\right) + \varepsilon$$

where ε is assumed to be a random error representing the discrepancy in the approximation. It accounts for the failure of the model to fit the data exactly. The function $f\left(X_1, X_2, ..., X_p\right)$ describes the relation between Y and X_1, X_2,…, X_p. An example of is the linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \qquad \textbf{(1)}$$

where $\beta_0, \beta_1, \cdots, \beta_p$, called the regression parameters or coefficients, are unknown constants to be determined (estimated) from the data.

Regression analysis is applied in various fields: economics, business, meteorology, medicine, biology, engineering, physics, education, agriculture, and psychology. Typically, a regression analysis is used for one (or more) of three purposes:

1. modeling the relationship between the predictors (X_i) and the response (Y);
2. prediction of the target variable (forecasting);
3. and testing of hypotheses.

The regression equation given in (1) is known as the multiple linear regression model, precisely because there are more than one predictor variables in the equation. It is linear in the sense that the regression coefficients are of degree (exponent) one. When there is only one predictor in the equation, then the model is called a simple linear regression model. The simple regression model is given by

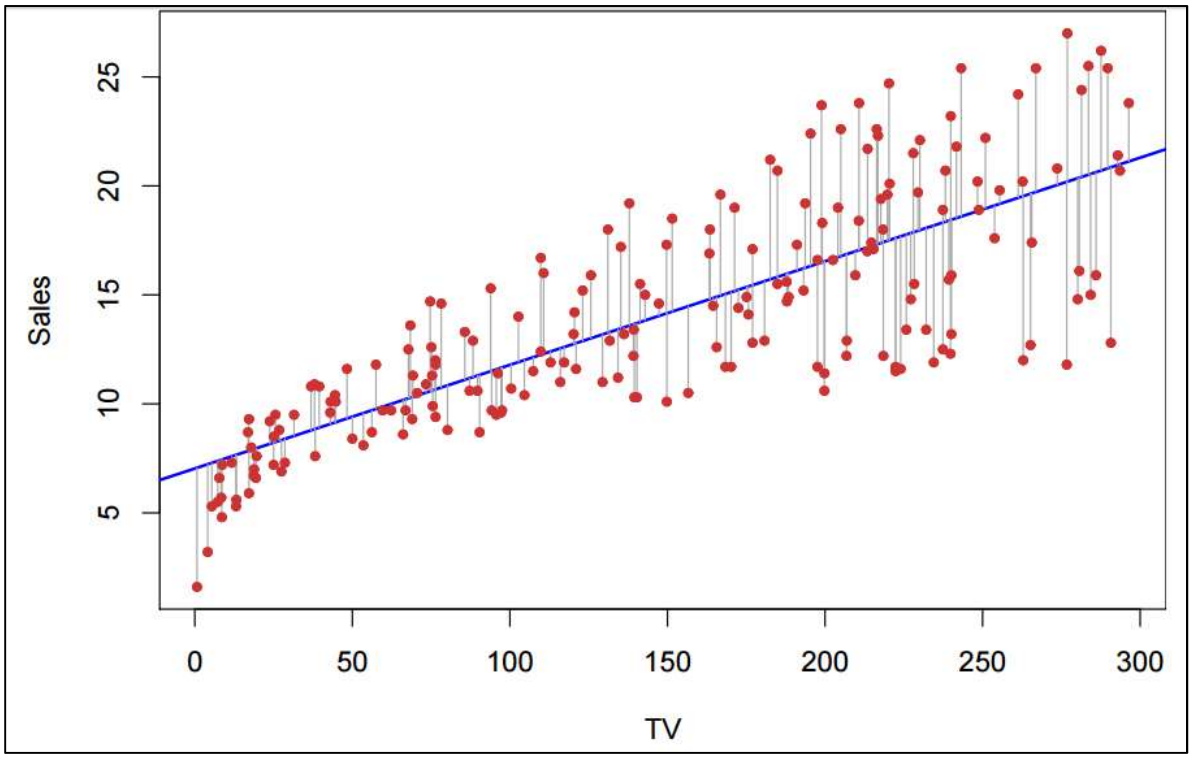
$$Y = \beta_0 + \beta_1 X_1 + \varepsilon \qquad \textbf{(2)}$$

Estimation of Parameters using Method of Least Squares

The true regression function represents the expected relationship between the response variable and the predictor variables, which is unknown. A primary goal of a regression analysis is to estimate this relationship, or equivalently, to estimate the unknown parameters (β_i). The standard approach is the least squares method, where the estimates are chosen to minimize

$$\sum_{i=1}^{n}\left[Y_i - \left(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\right)\right]^2 \quad \textbf{(3)}$$

Graphically, this simply means finding that straight line which is the closest to all the data points, on the average.



Standard Assumptions of Ordinary Least Squares Regression

The least squares method will not necessarily yield sensible results unless certain assumptions hold. These are:

1. **Assumption about the form of the model**: The model that relates the response Y to the predictors X_1, X_2,…, X_p is assumed to be linear in the regression parameters $\beta_0, \beta_1, \cdots, \beta_p$. This is commonly referred to as the *linearity* assumption.

2. **Assumptions about the errors (ε)**:
   a) The errors have a normal distribution (normality assumption)
   b) The errors have a mean of zero.
   c) The errors have the same variance (homogeneity of variance assumption)
   d) The errors are independent, that is, an error from one observation (ε_i) is independent of the error from another observation (ε_j).

3. **Assumptions on the predictor variables**:
   a) The predictor variables are assumed fixed or selected in advance.
   b) The values of the predictors are assumed to be measured without error.
   c) The predictor variables are assumed to be linearly independent of each other.

4. **Assumption about the observations**: All observations are equally important in determining the results and in influencing conclusions (absence of outliers and influential observations).

A feature of the method of least squares is that small or minor violations of the underlying assumptions do not invalidate the inferences or conclusions drawn from the analysis in a major way. Gross violations of the model assumptions can, however, seriously distort conclusions. Hence, a fundamental part of any regression analysis is to check them using various plots, tests, and diagnostics.

Interpreting the Regression Coefficients

The least squares regression coefficients have very specific meanings. They are often misinterpreted, so it is important to be clear on what they mean (and do not mean). Consider first the intercept, $\hat{\beta}_0$. It is the estimated expected value of the response variable when the predictors are all equal to zero. Note that this might not have any physical interpretation, since a zero value for all predictors might be impossible or might not be logically and practically possible. In that case, it is pointless to try to interpret this value.

The estimated coefficient for the *j*th predictor (*j=1, 2, …, p*) is interpreted as follows:

$\hat{\beta}_j$: The estimated expected change in the response variable associated with one unit change in the *j*th predictor variable, holding all else in the model fixed or equal.

Suppose a random sample of college students at a particular university is taken in order to understand the relationship between college GPA and other variables. A linear regression model is built with college GPA (CGPA) as a function of the of high school GPA (HSGPA) and the standardized Scholastic Aptitude Test (SAT), with the resulting estimated regression equation given below.

$$CGPA = 1.3 + 0.7 \times HSGPA - 0.0001 \times SAT .$$

Here:

$\hat{\beta}_0 = 1.3$: This is the estimated CGPA if HSGPA=0 and SAT=0. *Is there such a student with HSGP=0 and SAT=0?*

$\hat{\beta}_1 = 0.7$: This is the estimated increase in CGPA for every unit increase in HSGPA, holding SAT fixed or constant. *What does holding SAT fixed or constant?*

$\hat{\beta}_2 = -0.0001$: This is the estimated decrease in CGPA for every unit increase in SAT, holding HSGPA fixed or constant.

Measures of Goodness-of-fit of the Regression Model

A well-fitting regression model results in predicted values close to the observed data values. Three statistics are used in Ordinary Least Squares (OLS) regression to evaluate model fit: R-squared, the overall F-test, and the Root Mean Square Error (RMSE).

1. **Coefficient of Determination (R^2)**. It estimates the proportion of variability in *the response variable* accounted for by the best linear combination of the predictors. Values closer to 1 indicate a good deal of predictive power of the predictors for the response variable, while values closer to 0 indicate little predictive power. It can be shown that R^2 is affected by the number of predictors in the model. That is, for fixed sample size, R^2 is higher for models with more predictors and lesser for models with fewer predictors.

   The adjusted R^2 corrects this bias of the R^2. Therefore, for multiple regression models it is always suggested to look at the adjusted R^2 to get an estimate of the proportion of the variability in the response variable accounted for the predictor variables. For simple linear regression models, the R^2 is sufficient.

2. **Root Mean Square Error (RMSE)**. The RMSE is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data–how close the observed data points are to the model's predicted values. Whereas R-squared is a relative measure of fit, RMSE is an absolute measure of fit. As the square root of a variance, RMSE can be interpreted as the standard deviation of the unexplained variance, and has the useful property of being in the same units as the response variable. Lower values of RMSE indicate better fit. RMSE is a good measure of how accurately the model predicts the response, and is the most important criterion for fit if the main purpose of the model is prediction.

3. **F test**. It evaluates the null hypothesis that all regression coefficients are equal to zero versus the alternative that at least one does not. An equivalent null hypothesis is that R-squared equals zero. A significant F-test indicates that the observed R-squared is reliable, and is not a spurious result of oddities in the data set. Thus, the F-test determines whether the proposed relationship between the response variable and the set of predictors is statistically reliable, and can be useful when the research objective is either prediction or explanation.

Hypothesis Test

There are two types of hypothesis tests related to the regression coefficients of immediate interest.

1. Test of the overall significance of the regression

   Ho: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$

   Ha: some $\beta_j \neq 0$, *j=1, 2, …, p*

   Test stat.: F test

2. Test of the significance of an individual coefficient

   Ho: $\beta_j = 0$, *j=1, 2, …, p*

   Ha: $\beta_j \neq 0$

   Test stat.: T test

Example:

Determining the appropriate sale price for a home is clearly of great interest to both buyers and sellers. For illustration purposes, sale prices of a set of homes in a particular area are regressed on important characteristics of the home such as the number of bedrooms, the living area, the lot size, and so on. A public data on sales of 85 one-family homes in the Levittown, NY area from June 2010 through May 2011 will be used here. For each of the 85 houses in the sample, the number of bedrooms, number of bathrooms, living area (in square feet), lot size (in square feet), the year the house was built, and the property taxes are used as potential predictors of the sale price.

This data set is contained in Sheet1 of the Excel file named "*Data for regression analysis*".

Regression Analysis using Stata

1. Import the Excel file into Stata.
2. Type the following command,

   ***regress* saleprice bedrooms bathrooms livingarea lotsize yearbuilt propertytax**
3. Hit Enter.

OUTPUT

```
. regress saleprice bedrooms bathrooms livingarea lotsize yearbuilt propertytax

      Source |       SS       df       MS              Number of obs =      85
-------------+------------------------------           F(  6,    78) =   13.34
       Model |  1.7970e+11     6  2.9951e+10           Prob > F      =  0.0000
    Residual |  1.7511e+11    78  2.2450e+09           R-squared     =  0.5065
-------------+------------------------------           Adj R-squared =  0.4685
       Total |  3.5481e+11    84  4.2240e+09           Root MSE      =   47381

------------------------------------------------------------------------------
   saleprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    bedrooms |  -12291.01   9346.727    -1.32   0.192    -30898.92    6316.893
   bathrooms |   51699.24   13094.17     3.95   0.000     25630.75    77767.73
  livingarea |   65.90302   15.97905     4.12   0.000     34.09118    97.71486
     lotsize |   -.897068   4.194203    -0.21   0.831    -9.247084    7.452948
   yearbuilt |   3760.898   1962.504     1.92   0.059    -146.1484    7667.944
 propertytax |    1.47611   2.832372     0.52   0.604     -4.16271     7.11493
       _cons |   -7148819    3820094    -1.87   0.065    -1.48e+07    456403.4
------------------------------------------------------------------------------
```

Interpretation:

1. *Goodness-of-fit measures*
   a) The adjusted R^2 is 0.4685. The linear combination of the predictors (characteristics of a house) explains only 46.85% of the variation observed in the response variable (sale price).
   b) There is (highly) significant relationship between the response and the predictors taken altogether (F=13.34, p=0.0000).
   c) RMSE=47381. This value is very high implying a poor fit.
   d) Only two of the predictors provide significant explanation to house price: *number of bathrooms* (t=3.95, p=0.000) and *living area* (t=4.12, p=0.000).

2. Regression coefficients

   Let X_1=no. of bedrooms, X_2=no. of bathrooms, X_3=living area, X_4=lot size, X_5=year built, and X_6=property tax.

   Estimated regression equation:

   $$\hat{Y} = -7148819 - 12291.01X_1 + 51699.24X_2 + 65.903X_3 - 0.897X_4 + 3760.898X_5 + 1.476X_6$$

   a) $\hat{\beta}_2 = 51699.24$: Given all else in the model is held fixed, one additional bathroom in a house is associated with an estimated expected price that is $51,699.24 higher.

b) $\hat{\beta}_3 = 65.90$: One additional square foot of living area is associated with an estimated expected price that is $65.90, assuming all other variables are fixed.

Remarks:

1. The above example simply illustrates how to perform regression analysis using Stata and how to look at model fit and test significance of individual predictors.
2. The model does not provide a good fit to the data. There may be some problems with the data.
3. The interpretations of the significant coefficients are valid only if the assumptions are satisfied.

Regression Diagnostics: Detection of Model Violations

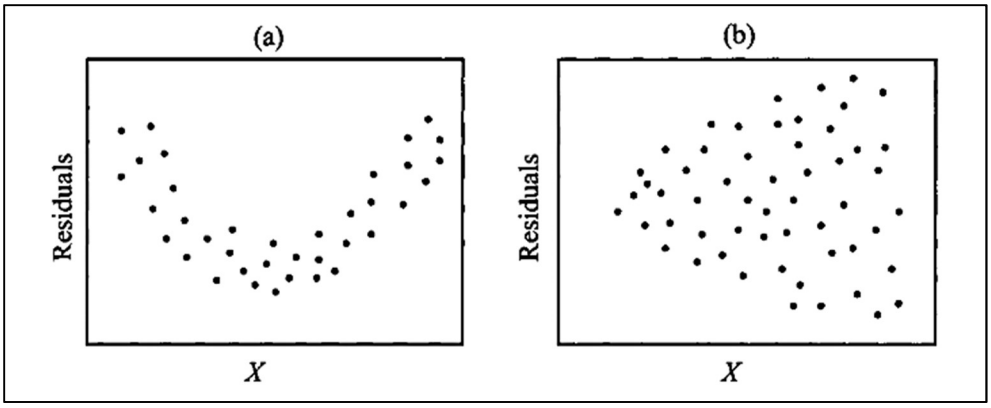
As is true of all statistical methods, linear regression analysis can be a very effective way to model data as long as the assumptions being made are true, but if they are violated least squares can potentially lead to misleading results.

An ***outlier*** is an observation with a response value y_i that is unusual relative to its expected value. Since the fitted value $\hat{y}_i$ is the best available estimate of the expected response for the i^{th} observation, it is natural that the ***residual*** $e_i = y_i - \hat{y}_i$ should be the key statistic used to evaluate if an observation is an outlier. An outlier is an observation with a large absolute residual (note that residuals can be positive or negative). The issue is then to determine what is meant by "large." Most often, the raw residuals are standardized and it is the ***standardized residuals*** which are used to gauge if an observation is an outlier. Standardized residuals outside ±2.5 can be flagged as potentially outlying and examined further.

Residual Plots to Check Model Assumptions

Residual plots can be used to assess the quality of a regression. Below are some of these plots.

1. Normal probability plot: If the residuals are normally distributed, this plot should resemble a nearly straight line with intercept of zero and a slope of one.
2. Histogram of the standardized residuals: It should resemble a bell and symmetrical centered at zero.
3. Scatter plot of the standardized residual versus the fitted values: This plot should show a random scatter of points. Any discernible pattern in this plot may indicate violation of some assumptions.
4. Scatter plot of the standardized residual against each of the predictor variables: This plot should also show a random scatter of points.



5. Residual lag plot: It is constructed by plotting i^{th} residual against $(i-1)^{th}$ residual and is useful for examining the independence of the error terms. Any non-random pattern in a lag plot suggests that the variance is not random or the errors are not independent.

Generating predicted values, residuals, and residual plots using Stata

1. To generate predicted or fitted values, type the command: ***predict* pred, *xb.***

2. To generate residuals, type the command: ***predict* resid, *residuals.***

3. To generate standard residuals, type the command: ***predict* zresid, *rstandard.***

4. To generate the lagged(1) standardized residuals, type the command: ***generate* lag1=zresid[_*n-1*].**

5. To generate the normal probability plot of the standardized residuals, type the command: ***pnorm* zresid.**

6. To generate the histogram of the standardized residuals, type the command: ***hist* zresid.**

7. To generate the scatter plot of the standardized residual versus the fitted values, type the command: ***scatter* zresid pred.**

8. To generate the scatter plot of the standardized residual versus each of the predictors, type the command: ***scatter* zresid livingarea.**

9. To generate the lag plot of the standardized residual, type the command: ***scatter* zresid lag1.**

Formal Tests for Checking Regression Assumptions

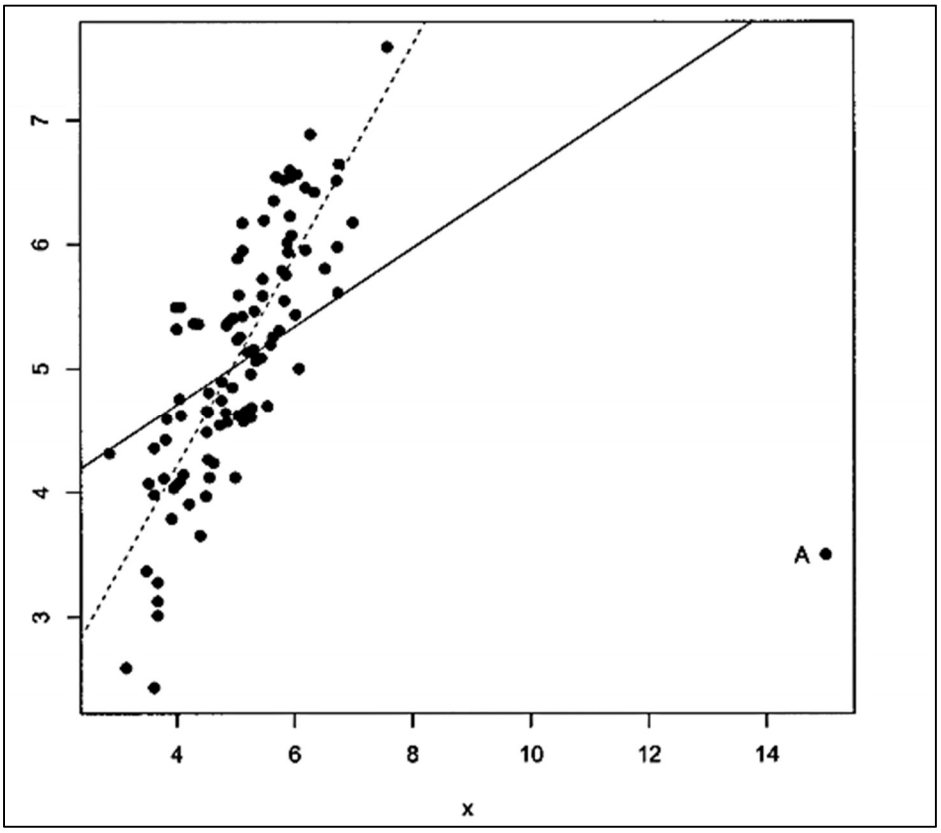
1. *Normality*: Wilk-Shapiro or Shapiro-Francia test
   Command: ***swilk* zresid**

2. *Homogeneity or homoscedasticity*: White's test or Breusch-Pagan / Cook-Weisberg test
   Command: ***estat imtest, white***
   Command: ***estat hettest***

Leverage and Influential Points

An observation with an extreme value on a predictor variable is called a point with high **leverage**. Leverage is a measure of how far an observation deviates from the mean of that variable. These points with high leverage values can have an effect on the estimate of regression coefficients. Observations with leverage values greater than $\frac{2.5(p+1)}{n}$ should be investigated as possible leverage points.

Influential points are those observations whose deletion, singly or in combination with others, causes substantial changes in the fitted model (estimated coefficients, fitted values, t-tests, etc.). Consider the plot below. Point A is an influential point. If it is deleted, the fitted regression line is the dashed line; the solid line represents the fitted regression line it A is not deleted.



Measures of Influence

The influence of an observation may be measured by the effects it produces on the fit when it is omitted from the data in the fitting process. There are numerous measures of influence in the literature.

1. Cook's distance. It measures the difference between the regression coefficients obtained from the full data and the regression coefficients obtained by deleting the i^{th} observation, or equivalently, the difference between the fitted values obtained from the full data and the fitted values obtained by deleting the i^{th} observation. Observations with Cook's distance greater than 1 is considered influential and has to be investigated.

2. DFITS. A measure similar to Cook's distance has been proposed by Welsch and Kuh (1977) and named DFITS. It represents the difference between the i^{th} fitted value from the full data and the i^{th} fitted value by deleting the i^{th} observation. Observations with $|DFITS| > 2\sqrt{\frac{p+1}{n-p-1}}$ are usually classified as influential points.

3. Hadi's Influence Measure. *Hadi (*1992) proposed a measure of the influence of the i^{th} observation based on the fact that influential observations are outliers in either the response variable or in the predictors, or both.

Generating leverage values and measures of influence using Stata

1. To generate leverage values, type the command: ***predict*** **lev,** ***leverage***.

2. To generate Cook's distance, type the command: ***predict*** **d,** ***cooksd***.

3. To generate DFITS measure, type the command: ***predict*** **dfit,** ***dfits***.

4. To generate DFITS measure, type the command: ***hinflu6*** **h**.
   [NOTE: Hadi's influence measure is computed using the hinflu6 package which needs to be installed into Stata by typing ***findit hinflu6*** in the command window.]

Qualitative variables as predictors

Qualitative or categorical variables can be very useful as predictor variables in regression analysis. Qualitative variables such as gender, marital status, or political affiliation can be represented by *indicator* or *dummy* variables. These variables take on only two values, usually 0 and 1. The two values signify that the observation belongs to one of two possible categories. Numerical values of indicator variables are not intended to reflect a quantitative ordering of the categories, but only serve to identify category or class membership.

Example:

A survey on the salary of computer professionals in a large company was conducted. The objective of the survey was to identify and quantify those variables that determine salary differentials. The response variable is salary (**S**) and the predictors are: experience (**X**)**,** measured in years; education (**E**), coded as 1 for completion of a high school (H.S.) diploma, 2 for completion of a bachelor degree (B.S.), and 3 for the completion of an advanced degree; and management (**M**), which is coded as 1 for a person with management responsibility and 0 otherwise. We shall try to measure the effects of these three variables on salary using regression analysis. This data set is contained in Sheet2 of the Excel file named "*Data for regression analysis*".

Note that when using indicator variables to represent a set of categories, the number of these variables required is one less than the number of categories. For example, in case of education categories we create two indicator variables $\mathbf{E_{i1}}$ and $\mathbf{E_{i2}}$, defined as follows:

$$E_{i1} = \begin{cases} 1, \text{if ith person is in the HS category} \\ 0, \text{otherwise} \end{cases}$$

and

$$E_{i2} = \begin{cases} 1, \text{if ith person is in the BS category} \\ 0, \text{otherwise} \end{cases}$$

The category that is not represented by an indicator variable is referred to as the *base category* or the *control group* because the regression coefficients of the indicator variables are interpreted relative to the control group.

Regression analysis with categorical predictors using Stata

1. We need to create indicators variables for Education (**E**). We do it in Stata using the following command: ***tabulate* e**, ***gen***(**ed**). Notice that this command creates three indicator variables (**ed1**, **ed2**, **ed3**). We will use only two of these indicator variables in the regression analysis. The indicator variable being left out becomes the reference group or control group.
2. Type the following command: ***regress* s x ed1 ed2 m** to generate the following output.

```
. regress s x ed1 ed2 m

      Source |       SS       df       MS              Number of obs =      46
-------------+------------------------------           F(  4,    41) =  226.84
       Model |  957816858      4  239454214            Prob > F      =  0.0000
    Residual |  43280719.5     41  1055627.3            R-squared     =  0.9568
-------------+------------------------------           Adj R-squared =  0.9525
       Total |  1.0011e+09     45  22246612.8           Root MSE      =  1027.4

------------------------------------------------------------------------------
           s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |    546.184   30.51919    17.90   0.000     484.5493    607.8188
         ed1 |   -2996.21   411.7527    -7.28   0.000    -3827.762   -2164.659
         ed2 |   147.8249   387.6593     0.38   0.705    -635.0689    930.7188
           m |   6883.531    313.919    21.93   0.000     6249.559    7517.503
       _cons |   11031.81   383.2171    28.79   0.000     10257.89    11805.73
------------------------------------------------------------------------------
```

Interpretation: Assuming all assumptions are satisfied!

1. Taken together, the predictor variables (experience, education and management responsibility) have significant relationship with salary (F=226.84, p=0.0000).
2. 95.25% of the variability in salary can be explained by the linear combination of experience, education and management responsibility.
3. The estimated regression equation is given by:

$$\hat{S} = 11031.81 + 546.18X - 2996.21Ed1 + 147.82Ed2 + 6883.53M$$

$\hat{\beta}_1 = 546.18$: An additional year of experience is estimated to be worth a salary increment of $546.18, assuming all other predictors are held fixed/constant.

$\hat{\beta}_2 = -2996.21$: The estimated salary of an employee who has a HS diploma is $2996.21 lower than another employee who has an advanced degree, assuming all else are constant.

$\hat{\beta}_4 = 6883.53$: The estimated salary of an employee who has a managerial position is $6883.53 higher than another employee who has none, assuming all else are constant.

Other issues with linear regression analysis

A. Multicollinearity

Multicollinearity exists when two or more of the predictors in a regression model are moderately or highly correlated. From a practical point of view, multicollinearity leads to two problems. First, it can happen that the overall F-statistic is significant, yet each of the individual t-statistics is not. Second, if the data are changed only slightly, the fitted regression coefficients can change dramatically. Hence, the usual way to interpret a regression coefficient as measuring the change in the response variable when the corresponding predictor variable is increased by one unit and all other predictor
variables are held constant is no longer valid. Note, however, that while multicollinearity can have a large effect on regression coefficients and associated t-statistics, it does not have a large effect on overall measures of fit like the overall F-test or R^2.

Example:

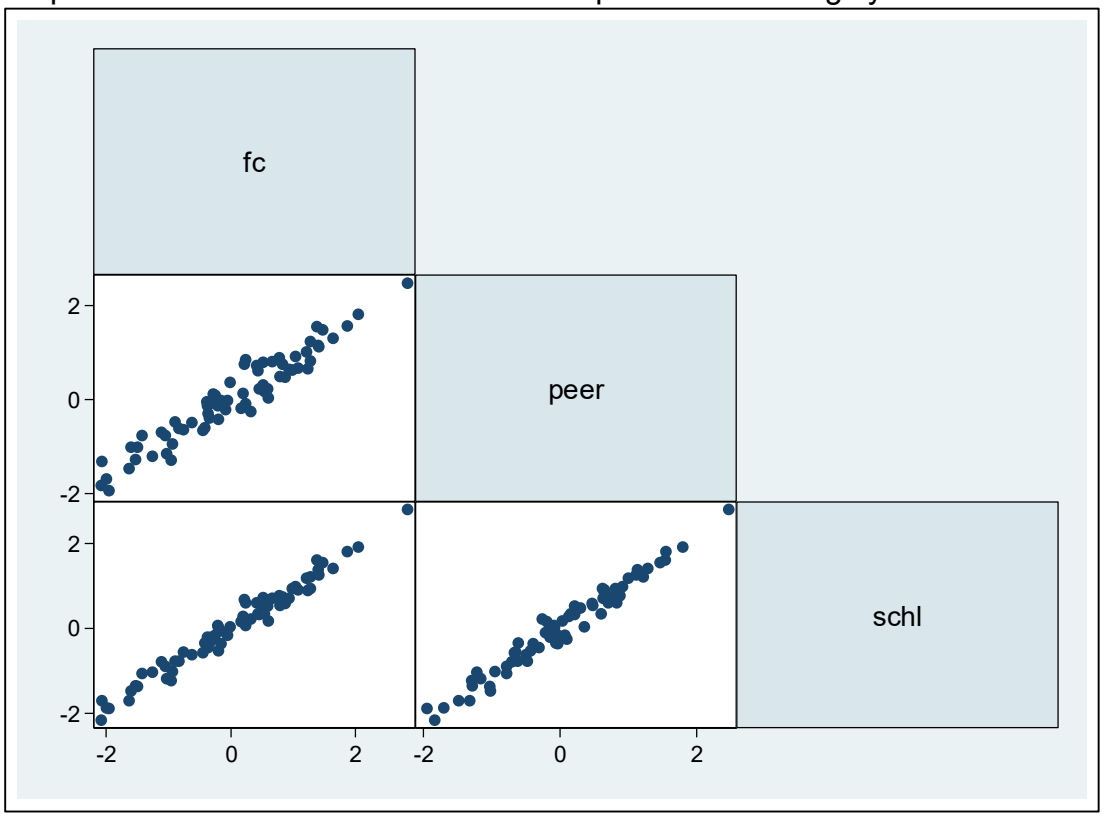
Consider a data set consisting of measurements taken in 1965 for 70 schools selected at random. The data consist of variables that measure student achievements (ACHV), faculty credentials (FC), the influence of their peer group in the school (PEER), and school facilities (SCHL). The objective is to evaluate the effect of school inputs on achievement. The output of regression analysis using Stata is given on the next page.

Source	SS	df	MS	Number of obs =	70
				F(3, 66) =	5.72
Model	73.5062325	3	24.5020775	Prob > F =	0.0015
Residual	282.873224	66	4.28595794	R-squared =	0.2063
				Adj R-squared =	0.1702
Total	356.379456	69	5.16491966	Root MSE =	2.0703

achv	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
fc	1.10126	1.410562	0.78	0.438	-1.715017	3.917537
peer	2.322057	1.481287	1.57	0.122	-.6354284	5.279542
schl	-2.280996	2.220448	-1.03	0.308	-6.714263	2.152272
_cons	-.0699591	.2506421	-0.28	0.781	-.5703821	.430464

Comments:

1. The F statistic is significant indicating that all together the variables FC, PEER, and SCHL are important in explaining variation in ACHV. However, the adjusted R^2 is very low (17.02%) and none of the predictors is significant.
2. This is a typical scenario when there is multicollinearity among the predictors. The scatterplot matrix below shows that the three predictors are highly correlated.



3. Note that the algebraic sign of SCHL is negative, which is opposite than what we expected.

Example:

Another data set considered here to illustrate the effect of multicollinearity is based on aggregate data concerning import activity in the French economy. The data have been analyzed by *Malinvaud* (1968). The variables are imports (IMPORT), domestic production (DOPROD), stock formation (STOCK), and domestic consumption (CONSUM), all measured in billions of French francs for the years 1949-1966. The output of regression analysis using Stata is given below.

Source	SS	df	MS		Number of obs =	18
					F(3, 14) =	168.45
Model	2576.92062	3	858.97354		Prob > F =	0.0000
Residual	71.3903825	14	5.09931304		R-squared =	0.9730
					Adj R-squared =	0.9673
Total	2648.311	17	155.783		Root MSE =	2.2582

import	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
doprod	.0322051	.1868844	0.17	0.866	-.3686221	.4330324
stock	.4141991	.3222598	1.29	0.220	-.2769794	1.105378
consum	.242746	.2853608	0.85	0.409	-.3692921	.8547841
_cons	-19.7251	4.125254	-4.78	0.000	-28.57289	-10.87731

Comment:

The F statistic is significant and adjusted R^2 is very high (96.73%) but none of the predictors is significant. This is another sign of multicollinearity.

Measure to detect multicollinearity

The relationship between the predictor variables can be judged by examining a quantity called the *variance inflation factor*. As the name suggests, a variance inflation factor (*VIF*) quantifies how much the variance is inflated. Thus, the value of VIF measures the amount by which the variance of the regression coefficient of a predictor is increased due to the linear association of that predictor with other predictor variables relative to the variance that would result if it were not related to them linearly. Values of variance inflation factors greater than 10 is often taken as a signal that the data have collinearity problem.

To generate the VIF using Stata, simply type the command: ***vif***. The VIFS of the three predictors in the Achievement example are:

Variable	VIF	1/VIF
schl	83.16	0.012026
fc	37.58	0.026609
peer	30.21	0.033100
Mean VIF	50.32	

While for the Import example, the VIF values are:

Variable	VIF	1/VIF
doprod	469.74	0.002129
consum	469.37	0.002131
stock	1.05	0.952492
Mean VIF	313.39	

B. Autocorrelation

Recall that one of the assumptions when building a linear regression model is that the errors are independent. This section discusses methods for dealing with dependent errors. In particular, the dependency usually appears because of a temporal component. Error terms correlated over time are said to be **autocorrelated** or **serially correlated**. When error terms are autocorrelated, some issues arise when using ordinary least squares. These problems are:

- Estimated regression coefficients are still unbiased, but they no longer have the minimum variance property.
- The MSE may seriously underestimate the true variance of the errors.
- The standard error of the regression coefficients may seriously underestimate the true standard deviation of the estimated regression coefficients.
- Statistical intervals and inference procedures are no longer strictly applicable.

C. Model misspecification

Specification error occurs when an independent variable is correlated with the error term. There are several different causes of specification error:

- incorrect functional form could be employed;
- a variable omitted from the model may have a relationship with both the dependent variable and one or more of the independent variables (omitted-variable bias);
- an irrelevant variable may be included in the model;
- the dependent variable may be part of a system of simultaneous equations (simultaneity bias);
- measurement errors may affect the independent variables.

Variable selection procedures (Reading assignment)

1. Forward selection
2. Backward elimination
3. Stepwise method