Regression Analysis

Objective of Regression analysis is to explain variability in dependent variable by means of one or more of independent or control variables.

Applications

There are four broad classes of applications of regression analysis.

- Descriptiveorexplanatory:interestmaybeondescribing“Whatfactorsinfluencevari-ability

  in dependent variable?” For example, factor contributing to higher sales among company’s sales force.

- Predictive,forexamplesettingnormalquotaorbaselinesales.Wecanalsouseestimated equation to determine “normal” and “abnormal” or outlier observations.

- ComparingAlternativetheoreticalexplanations,

  – Consumers use reference price in comparing alternatives,

  – Consumers use specific price points in comparing alternatives.

- Decisionpurpose,

  – Estimating variable and fixed costs having calibrated cost function.

  – Estimating sales, revenues and profits having calibrated demand function.

  – Setting optimal values of marketing mix variables.

  – Using estimated equation for “What if” analysis.

Data Requirement

- Measurementontwoormorevariablesoneofwhichmustbedependent.

- Dependentvariablemusthaveintervalorratioscalemeasurement.

- Ifindependentvariablesarenominalscaled(e.g.brandchoice),thenappropriatecautionmust

  be maintained so that results from analysis can be interpreted. For example, it may be necessary to create variables that take values 0 and 1 or dummy variables.

Steps in Regression Analysis

1. Decide on purpose of model and appropriate dependent variable to meet that purpose.

2. Decide on independent variables.

3. Estimate parameters of regression equation.
4. Interpret estimated parameters, goodness of fit and qualitative and quantitative assess-ment of parameters.
5. Assess appropriateness of assumptions.
6. If some assumptions are not satisfied, modify and revise estimated equation.
7. Validate estimated regression equation.

We will examine these steps with the assumption that purpose of model is already been decided and we need to perform remaining steps.

Decision about Independent Variables

Here are some suggestion for variable(s) to be included in regression analysis as independent variables.

- Basedontheory.
  - Economic, sales is a function of price,
  - Psychological, behavioural intention and attitude toward a product,
  - Biological, fertilizer usage, generally increase plant growth.
- Priorresearch,
  - Replicate findings for earlier efforts.
  - Extend results for alternative product category.
  - Bring new insights to earlier efforts.
- Educated“Guesses”,goodideaorcommonsense.
- Statisticalapproaches.
  - Stepwise Forward, add a variable that contributes most to explaining dependent variable, continue this, until either no variables are left to add or none of remaining variables contribute in explaining variation in dependent variable.
  - Stepwise Backward, add all variables to the model and remove one variable at a time, starting with one that explains least amount of variation in dependent variable.
  - All Subset, estimate all combinations containing two variables at a time, then three variables at a time etc. Then, choose a subset that has most stable set of independent variables.

- Allvariablescontainedindataset.

Estimating Parameters

- Methodofleastsquares,or
- Methodofmaximumlikelihood,or
- Weightedleastsquares,or
- Methodofleastabsolutedeviations.

We will examine several alternative approaches to estimate parameters including situation where we have only two observations.
A Simple Regression Model can be written as

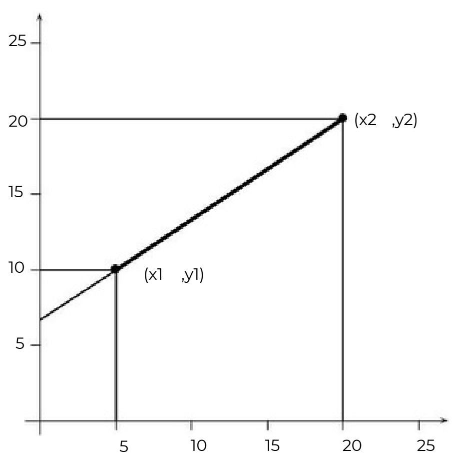
ValueofDependentvariable = Constant+
Slope × Value of Indep. variable + Error

$$y = a + b \times x + E$$

- Constant(a),Slope(b)andError(E)areunknown.
- You observe N pair of values of dependent and independent variables.
- Regressionanalysisprovidesreasonable(statisticallyunbiased)valuesforslope(s)and intercept.

An Illustrative Example - Two observations only.

Suppose we have two observation (x1, y1) and (x2, y2) or (5,10) and (20,20). These observations graphically can be shown as follows.



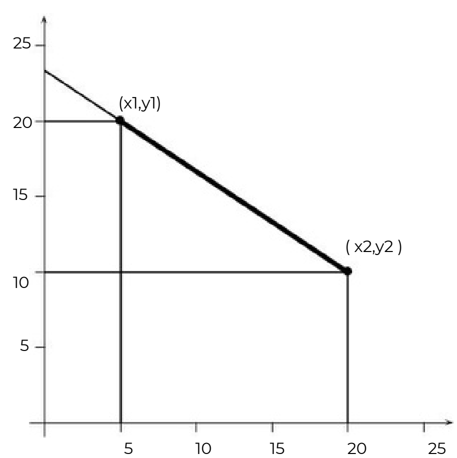
$$\text{Slope} = \frac{y2-y1}{x2-x1} = \frac{20-10}{20-5} = 0.66$$

Slope is positive because $y2 \geq y1$ and $x2 \geq x1$

Intercept is $y1 - b \times x1 = 6.67$

The resulting equation would be y = 6.67 + .66 × x.

Now, suppose we have two observation (x_1, y_1) and (x_2, y_2) or (5,20) and (20,10). These observations graphically can be shown as follows.



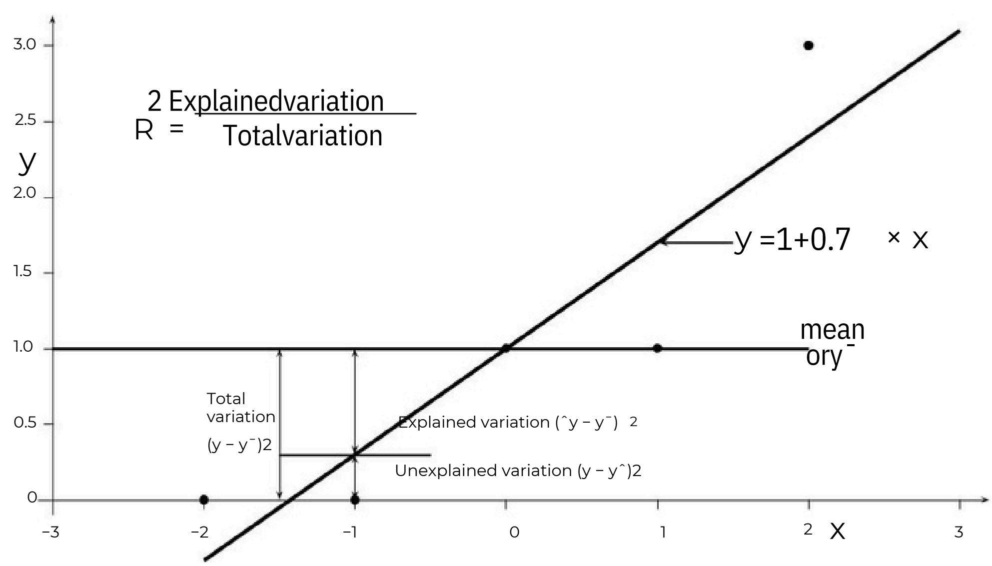
$$\text{Slope} = \frac{y_2 - y_1}{x_2 - x_1} = \frac{10-20}{20-5} = -0.66$$

Slope is negative because $y_2 < y_1$ and $x_2 \geq x_1$

Intercept is $y_1 - b \times x_1 = 23.33$

The resulting equation would be y = 23.33 − .66 × x.

Now suppose we observe five pairs of x and y observations as follows: (−2, 0), (−1, 0), (0, 1), (1, 1) and (2, 3). These are displayed below along with regression line which is shown in dashed for-mat.



As you can see from above examples, estimating parameters is nothing more than assigning appropriate values to parameters. Let us re-write our observations again, in somewhat different format and see another alternative approach to obtain parameter estimates.

$$y_i = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 3 \end{pmatrix} \qquad x_i = \begin{pmatrix} -2 \\ -1 \\ 0 \\ 1 \\ 2 \end{pmatrix}$$

Our regression equation can be written as

$$y_i = a + b \times x_i + E_i \qquad i = 1, \cdots, 5.$$

Suppose we added both sides (over all observations) of above equation, the we could write

$$\sum_{i=1}^{5} y_i = \sum_{i=1}^{5} a + \sum_{i=1}^{5} b x_i + \sum_{i=1}^{5} E_i.$$

Further let us divide both sides by 5 or number of observations, we would get,

$$\frac{\sum_{i=1}^{5} y_i}{5} = \frac{\sum_{i=1}^{5} a}{5} + \frac{\sum_{i=1}^{5} b x_i}{5} + \frac{\sum_{i=1}^{5} E_i}{5}.$$

This is equal to

$$\bar{y} = a + b\bar{x} + \bar{E}.$$

Let us assume that $\bar{E}$ is zero, which simply says that positive differences and negative differences cancel each other and on an average random noise is zero. Now subtract the average equation from our original equation. That is,

$$y_i - \bar{y} = b(x_i - \bar{x}) + E_i.$$

Suppose now we multiply both sides by $(x_i - \bar{x})$, then we would get a complicated expression like

$$(x_i - \bar{x})(y_i - \bar{y}) = b(x_i - \bar{x})(x_i - \bar{x}) + E_i(x_i - \bar{x}).$$

Let us now take average of both sides and divide by $(5 - 1)$ or $(N - 1)$ where N is number of observations. This would lead to

$$\frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{N - 1} = b \frac{\sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})}{N - 1} + \frac{\sum_{i=1}^{N} E_i(x_i - \bar{x})}{N - 1}.$$

We now have to make our second assumption which states that independent variable and error term are not correlated. That is, $\sum_{i=1}^{N} E_i(x_i - \bar{x}) = 0$. This is one of the difficult assumption to test but one that is required, to derive value of b. With this assumption, we are in position to write estimate of b or $\hat{b}$. That is,

$$\hat{b} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})}.$$

We are also assuming that $x_i - \bar{x}$ is not equal to zero. That is, there is some variation in independent variable, one that is useful to explain variation in dependent variable. Once we know estimate of b, we can go back to $\bar{y} = a + b\bar{x}$ and solve for a. This we will call as $\hat{a}$ and it can be obtained by $\hat{a} = \bar{y} - \hat{b}\bar{x}$. Implicit in our effort to compute various averages, we assumed that each observation is equally weighted. This assumption is satisfied if error variability across observation is about the same. That is, $(y_i - \hat{y}_i)^2$ is similar over all the observations.

Let us see applicability of above work to our example. First note that $\bar{y} = 1$ and $\bar{x} = 0$. Then, $y_i - \bar{y}$ and $x_i - \bar{x}$ is

$$y - \bar{y} = \begin{pmatrix} 0-1 \\ 0-1 \\ 1-1 \\ 1-1 \\ 3-1 \end{pmatrix} \quad x - \bar{x} = \begin{pmatrix} -2-0 \\ -1-0 \\ 0-0 \\ 1-0 \\ 2-0 \end{pmatrix}.$$

This simplifies to

$$y_i - \bar{y} = \begin{pmatrix} -1 \\ -1 \\ 0 \\ 0 \\ 2 \end{pmatrix} \quad x_i - \bar{x} = \begin{pmatrix} -2 \\ -1 \\ 0 \\ 1 \\ 2 \end{pmatrix}.$$

This would result in

$$(y - \bar{y})(x - \bar{x}) = \begin{pmatrix} 2 \\ 1 \\ 0 \\ 0 \\ 4 \end{pmatrix} \text{ and }$$

$$(x - \bar{x})^2 = \begin{pmatrix} 4 \\ 1 \\ 0 \\ 1 \\ 4 \end{pmatrix}$$

This would mean that $\hat{b} = \frac{7}{10}$ and $\hat{a} = 1$. Note that our equation in this case would be $y_i = 1 + 0.7 \times x_i$. This is exactly same equation written on our graph as well. Note that we could also estimate proportion of variability explained by independent variable by computing R^2 and set of other summary measures.

Multiple independent variables

Nothing much changes, if we had multiple variables. We, however, need to worry about joint variability of independent variables. Consider a situation with two independent variables(x_{1i}

and x_{2i}). That is,

$$y_i = a + b_1 \times x_{1i} + b_2 \times x_{2i} + E_i.$$

Here our interest lies with finding best values of a, b1 and b2. To derive these, we could follow above steps. That is, first averaging of both sides, then subtracting the averages and finally multiplying by $(x_{1i} - \bar{x}_1)$ and $(x_{2i} - \bar{x}_2)$. This will give us two equations with two unknowns. That is,

$$(y_i - \bar{y}) = b_1(x_{1i} - \bar{x}_1) + b_2(x_{2i} - \bar{x}_2)$$

Multiply first by $(x_{1i} - \bar{x}_1)$ and then by $(x_{2i} - \bar{x}_2)$. This will result in,

$$(y_i - \bar{y})(x_{1i} - \bar{x}_1) = b_1(x_{1i} - \bar{x}_1)(x_{1i} - \bar{x}_1) + b_2(x_{2i} - \bar{x}_2)(x_{1i} - \bar{x}_1)$$
$$(y_i - \bar{y})(x_{2i} - \bar{x}_2) = b_1(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) + b_2(x_{2i} - \bar{x}_2)(x_{2i} - \bar{x}_2)$$

We would sum both sides of both equations and divide by N − 1. Moreover for simplicity, we could make following substitutions.

$$S_{yx1} = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(x_{1i} - \bar{x}_1)}{N-1}$$
$$S_{yx2} = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(x_{2i} - \bar{x}_2)}{N-1}$$
$$S_{x2x1} = S_{x1x2} = \frac{\sum_{i=1}^{N}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{N-1}$$
$$S_{x1x1} = \frac{\sum_{i=1}^{N}(x_{1i} - \bar{x}_1)(x_{1i} - \bar{x}_1)}{N-1}$$
$$S_{x2x2} = \frac{\sum_{i=1}^{N}(x_{2i} - \bar{x}_2)(x_{2i} - \bar{x}_2)}{N-1}.$$

These terms are called averages of sums of squared values of cross products (SSCP). These are very useful quantities in various multivariate analysis procedures. After substituting these terms, we may write our earlier equation as

$$S_{yx1} = \hat{b}_1 S_{x1x1} + \hat{b}_2 S_{x1x2}$$
$$S_{yx2} = \hat{b}_1 S_{x1x2} + \hat{b}_2 S_{x2x2}$$

Suppose we assumed that $S_{x1x2} = 0$, then we could at once write estimates for b_1 and b_2. That is,

$$\hat{b}_1 = \frac{S_{yx1}}{S_{x1x1}}$$
$$\hat{b}_2 = \frac{S_{yx2}}{S_{x2x2}}$$

If $S_{x_1x_2} \neq 0$, then we need to solve these two equations simultaneously and obtain estimates. There is also a possibility that $S_{x_1x_2} = S_{x_1x_1}$ which would also imply that $S_{x_1x_2} = S_{x_2x_2}$. This would result in collapse of two unknown to just one, that is, $(b_1 + b_2)$. This condition is called perfect multicollinearity. Not that

$$S_{yx_1} = \hat{b}_1 S_{x_1x_1} + \hat{b}_2 S_{x_1x_2}$$
$$S_{yx_2} = \hat{b}_1 S_{x_1x_2} + \hat{b}_2 S_{x_2x_2},$$

can be written in matrix form as follows:

$$\begin{pmatrix} S_{yx_1} \\ S_{yx_2} \end{pmatrix} = \begin{pmatrix} S_{x_1x_1} & S_{x_1x_2} \\ S_{x_1x_2} & S_{x_2x_2} \end{pmatrix} \begin{pmatrix} \hat{b}_1 \\ \hat{b}_2 \end{pmatrix}.$$

The solution to such matrix equations could be written as

$$\begin{pmatrix} \hat{b}_1 \\ \hat{b}_2 \end{pmatrix} = \begin{pmatrix} S_{x_1x_1} & S_{x_1x_2} \\ S_{x_1x_2} & S_{x_2x_2} \end{pmatrix}^{-1} \begin{pmatrix} S_{yx_1} \\ S_{yx_2} \end{pmatrix}.$$

Let us summarize assumptions that were made up to this point.

Assumptions of Regression Equation

- Onanaveragedifferencebetweentheobservedvalue(y_i)andthepredictedvalue($\hat{y}_i$) is zero.
- On an average the estimated values of errors and values of independent variables are not related to each other.
- Thesquareddifferencesbetweentheobservedvalueandthepredictedvaluearesimilar.
- Thereissomevariationinindependentvariable.Iftherearemorethanonevariableinthe equation, then two variables should not be perfectly correlated.

We could also make following observations about slope and intercept.

Intercept or Constant

- Interceptisthepointatwhichtheregressioninterceptsy-axis.
- Intercept provides a measure about the mean of dependent variable when slope(s) are zero.
- Ifslope(s)arenotzerotheninterceptisequaltothemeanofdependentvariableminusslope× mean of independent variable.

Slope

- Changeisdependentvariableaswechangeindependentvariable.
- Zeroslopemeansthatindependentvariabledoesnothaveanyinfluenceondependent variable.
- For a linear model, slope is not equal to elasticity. That is because, elasticity is percent change in dependent variable, as a result one percent change in independent variable.

Interpretation and Assessment

In this step, I envision explaining obtained results and providing insights about set of vari-ables. This should be both from conceptual point of view as well as statistical perspective. Fur-thermore, statistical measures could either be qualitative[1] such as r-square (R^2) or quantitative measure like F-statistic. When computing R^2, we do not make any additional assumptions. On the other hand, application of F-statistics we need additional assumptions. F-statistics is used to test whether set of regressors significantly explain variations in the dependent variable. To use F-statistic or t-statistic, we require two additional assumptions. First, which is our fourth assumption, require that error values be normally and identically distributed. Finally, we also need to decide on appropriate probability level to reject or accept our null hypothesis. I will usually follow prob. of 0.05 to reject null hypothesis. This in common language says that I will accept the null hypothesis 19 times out of 20 and reject it once out of 20. Here is a summary of steps that one could follow in testing hypothesis.

1. Decide on null hypothesis. Most computer programs, unless we specify, test using the F-statistic whether all regressor slopes are equal to zero. The t-statistic test whether a particular regressor is equal to zero.
2. Decide on probability level at which to reject the null hypothesis. You may recall this as alpha (α) level associated with Type I error. Although the most scientific research tradi-tions use probability level of 0.05, you might be risk-taker and willing to use something else like 0.25.
3. Computeteststatistic[2].

[1] Consider a measure like R^2. We know that it is bounded between zero and one. But actual magnitude that might be acceptable varies from applications to applications as well as quality of data. Hence indicators like R^2, I consider them to be qualitative measures of goodness-of-fit. On the other hand, F-statistic require that we make assumptions about distribution of errors, probability level to reject or accept null hypothesis and specifies whether null or alternative hypothesis is true or false. Hence, indicators like F-statistic I will call them as quantitative measures.

[2] The F-statistic is ratio of two mean squared errors, the average squared deviations explained to the average squared deviations not explained. Since we assume that errors are normally distributed, squared values of such errors are chi-squared (χ^2) distributed. The F-statistic then is a ratio of two χ^2 distributed variables. The t-statistic is ratio of the estimated parameter value to the standard error of parameter estimate.

4. Decide whether to reject or accept null hypothesis. At a particular probability level, if the tabled[3] value is less than the computed statistic, then we should reject the null hypothesis and vice versa. There is an alternative for this step. Most computer programs, print statistic as well as probability of the computed statistic. In such a situation, if probability is less than or equal to 0.05, then we reject the null hypothesis.

Following table summarizes above discussion about interpretation of parameters.

Interpretational Measures

Specific Aspect	Descriptive	DecisionOriented
Goodness-of-fit	R^2 or adjusted R^2, indicates percentvariationindependent variable explained by a set of independent variables.	F-statistic,largernumbermeansreject thenullhypothesisthatallparametersare zero
Individual parameters	Sign,Magnitudeandelasticity	t-statisticindicateswhetherspecificparameterisdifferentfromzero.Incomparing,t-statisticfortwoparameters,alarger t-statistic indicates that the independent variable is more important than other.

Let us apply all this to our small problem. First the SAS input.

```
options nocenter nodate ps = 70 ls =80 nonumber formchar=|---- |+|----- |;
data toy;

 input y x;
datalines;
 0  -2
 0  -1
 1   0
 1   1
 3   2
;;;;
proc reg;model y = x; run;
```

[3] I am here referring to table of t- or F-statistics.

SAS output produced following:

```
Dependent Variable: Y

Analysis of Variance

                            Sumof        Mean
Source          DF        Squares       Square     FValue       Prob>F

Model            1     4.90000     4.90000      13.364      0.0354
Error            3     1.10000     0.36667
CTotal           4     6.00000

   RootMSE       0.60553   R-square      0.8167
   DepMean       1.00000   AdjR-sq       0.7556
   C.V.         60.55301

Parameter Estimates

              Parameter     Standard    TforH0:
Variable DF    Estimate        Error  Parameter=0  Prob>|T|

INTERCEP  1    1.000000  0.27080128        3.693     0.0345
X         1    0.700000  0.19148542        3.656     0.0354
```

Our null hypothesis for this example would state “variable x does not explain statistically significant variations in y”. Our computed F-statistic is 13.4 with prob. of 0.035 which suggest that we should reject the null hypothesis. Moreover, R2 is 0.817 which indicates that substantial proportion of variation in y is accounted by variable x. Since there is only one variable in our equation, many of conclusions in F-statistic also will be matched by t-statistic. That is, reject null hypothesis that b = 0.

Evaluating Assumptions

Of the various assumptions in our analysis, following assumption lend to some form test procedure.

1. The squared differences between the observed dependent variable value and the predicted value are similar for all observations.
2. Each observation has equal influence on estimated parameters.
3. Independent variables are not correlated, or correlation among them is low.
4. If dependent variable is sorted in ascending or descending order, then the estimated residuals (yi – yˆi) are not related to each other.

5. The estimated residuals ($y_i - \hat{y}_i$) are normally distributed.

We will examine each of them below.

Assumptions and and Tests

Assumption	Descriptive	DecisionOriented
Similar variation	Visualinspectionorplotobservationnumberandparticularmeasure	Studentresiduals,normalizedresidual. Checkforobservationswiththeabsolutevalue ofnormalizedresiduals≥2. Rstudent, value of residual when a particular observation is deleted. Check observation with the absolute value of Rstudent ≥2. Cook's D, same as above and check observation with Cook's D ≥ 8/[N − 2(k + 1)].
Equal Weight orinfluence	Visualinspectionorplotobservationnumberandparticularmeasure	COVRATIO,ratioofcovariationamongindependentvariablesbasedonparticularobservationexcludedtoonebasedontotalsample. If the absolute value of COVARATIO − 1 is ≥ 3(k + 1)/[N − k − 1], then examine particular observation. DFFITS indicate change in parameter estimates taken all together when a particular observation is excluded. The absolute value of DFFITS ≥ $2\sqrt{(k-1)/N}$ consideredextreme observation. DFBETAS indicate change individual parameter estimate, when particular observation is excluded. The absolute value of DFBETAS ≥ $2/\sqrt{N}$ shouldbeconsideredextremeobservation.
Independent variables uncorrelated or collinearity	Visualinspectionofcorrelations andproportionofvarianceshared acrossvariables.	Variance inflation factor (VIF) greater than 10 is considered a case of multicollinear- ity. Condition Index, more than 15 to 20 is considered a case for multicollinearity.
Successive errorterms related or auto-correlation	Visualinspectionorplotobservationnumberandresiduals.	autocorrelation should be equal to zero and statistically not significant. Durbin-Watson's Statistic farther away from 2 is considered a situation with autocorrelation.
Normality of residuals	Q-Qorprobabilityplot,fora normallydistributedvariable,plot would be straight line passing through origin.	Tests of skewness, kurtosis and / or other test procedure to detect departure from normality.

Let us see how all these things apply to our simple example along with some of statistical derivations. Suppose our regression equation can be written as

$$y_i = a + b \times x_i + \epsilon_i \qquad i = 1, \cdots, 5.$$

For the first observation, then the predicted value is

$$\hat{y}_1 = \hat{a} + \hat{b} x_1$$

where $\hat{a}$ and $\hat{b}$ are used to denote the estimated intercept and slope respectively. It follows that the estimated residual for observation i is $\hat{\epsilon}_i = y_i - (\hat{a} + \hat{b} x_i)$ and sum of squared residuals is $\sum_{i=1}^{n} \hat{\epsilon}_i^2$ and the standard deviation, often denoted by s is

$$s = \sqrt{\sum_{i=1}^{n} \hat{\epsilon}_i^2}.$$

Note that under the assumptions of linear regression, it can be shown that

$$E(\hat{a}) = a$$
$$E(\hat{b}) = b$$
$$\text{var}(\hat{a}) = \frac{s^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2}$$
$$\text{var}(\hat{b}) = \frac{s^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$
$$\text{cov}(\hat{a}, \hat{b}) = \frac{-s^2 \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

where $\bar{x}$ is the average of x_i, $i = 1, \cdots, 5$.

Suppose we want to know the standard error of the predicted value for the first observations, $\hat{y}_1$, then we determine the variance of $\hat{y}_1$ and from that we compute the standard error. Note that variance of $\hat{y}_1$ is

$$\text{var}(\hat{y}_1) = \text{var}(\hat{a}) + \text{var}(\hat{b}) x_1^2 + 2 x_1 \text{cov}(\hat{a}, \hat{b})$$

It can be shown that

$$\text{var}(\hat{y}_1) = s^2 \left[\frac{1}{n} + \frac{(x_1 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right],$$

and square root of $\text{var}(\hat{y}_1)$ is usually reported as the standard error of prediction. Note that quantity inside square bracket is called diagonal elements of hat matrix and indicates distance between independent variable values for specific observation and the mean values. Similarly it can be shown that

$$\text{var}(\hat{\epsilon}_1) = s^2 \left[1 - \frac{1}{n} - \frac{(x_1 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right],$$

and square root of $\text{var}(\hat{\epsilon}_1)$ is usually reported as the standard error of residual. Following output indicates that SAS generates numbers as we would expect.

Obs	Dep Var Y	Predict Value	Std Err Predict	Residual	Std Err Residual	Student Residual
1	0	-0.4000	0.469	0.4000	0.383	1.044
2	0	0.3000	0.332	-0.3000	0.507	-0.592
3	1.0000	1.0000	0.271	0	0.542	0.000
4	1.0000	1.7000	0.332	-0.7000	0.507	-1.382
5	3.0000	2.4000	0.469	0.6000	0.383	1.567

Obs	-2-1-0 1 2	Cook's D	Rstudent	Hat Diag H	Cov Ratio	Dffits
1	| |** |	0.818	1.0690	0.6000	2.2779	1.3093
2	| *| |	0.075	-0.5145	0.3000	2.5068	-0.3368
3	| | |	0.000	0.0000	0.2000	2.8125	0.0000
4	| **| |	0.409	-1.8708	0.3000	0.4250	-1.2247
5	| |*** |	1.841	3.0000	0.6000	0.1860	3.6742

Obs	INTERCEP Dfbetas	X Dfbetas
1	0.7559	-1.0690
2	-0.2750	0.1945
3	0.0000	0.0000
4	-1.0000	-0.7071
5	2.1213	3.0000

```
Sum of Residuals                        0
Sum of Squared Residuals           1.1000
Predicted Resid SS (Press)         4.4337
```

Note that in this example,

$$var(\hat{y}_1) = s^2\left[\frac{1}{n} + \frac{(x_1-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right],$$

where s^2 is sum of squared residuals divided by $(n-2)$ or 3 in this case. Furthermore, $\bar{x}=0$ and $\sum_{i=1}^{5}(x_i-\bar{x})^2 = 10$. This results in

$$var(\hat{y}_1) = \frac{1.1}{3}\left[\frac{1}{5} + \frac{4}{10}\right] = \frac{1.1}{5}$$

and square root of 0.22 results in the standard error of prediction of 0.469 for this observation. Similarly,

$$\widehat{\mathrm{var}}(E_1) = s^2\left[1 - \frac{1}{n} - \frac{(x_1-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right],$$

$$= \frac{1.1}{3}\left(1 - \frac{1}{5} - \frac{4}{10}\right)$$

$$= 0.14667,$$

and square root of this is 0.383. Note that column Student Residual is ratio of column Residual to Std Err Residual. Note that all other remaining measures reported above (Cook's D, Rstudent etc.) require estimate based on particular observation being deleted. For example, estimating a and b when first observation is deleted, denoted by $\hat{a}_{(1)}$ and $\hat{b}_{(1)}$. It is possible to obtain these estimate without actually conducting separate regression analyses. Thus,

$$\hat{a}_{(1)} = \hat{a} - \frac{\hat{E}_1}{n(1-h_{11})}$$

$$\hat{b}_{(1)} = \hat{b} - \frac{x_i\hat{E}_1}{(1-h_{11})\sum_{i=1}^{n}(x_i-\bar{x})^2},$$

where h_{11} is diagonal elements of hat matrix or H (see notes above). For the first observation, $\hat{a}_{(1)}$ and $b_{(1)}$ is equal to 0.8 and 0.9 respectively. Similarly, RSTUDENT is normalized residual when ith observation is excluded from analysis. For the first observation,

$$\mathrm{RSTUDENT}(1) = \frac{E_1}{s_{(1)}\sqrt{1-h_{11}}},$$

where $s_{(1)}$ is estimated standard error when the first observation is excluded and that can be estimated by

$$s^2_{(1)} = \frac{1}{n-p-1}\left[(n-p)s^2 - \frac{E_1^2}{1-h_{11}}\right]$$

$$= \frac{1}{4-1-1}\left[(4-1)\frac{1.1}{5-1-1} - \frac{0.4\times0.4}{1-0.6}\right]$$

$$= 0.5 \times (1.1 - 0.4) = 0.35.$$

Then substituting square root of 0.35 in expression of RSTUDENT to obtain

$$\mathrm{RSTUDENT}(1) = \frac{0.4}{0.5916\sqrt{0.4}} = 1.069,$$

which is reported for the first observation.

A Realistic Example

As you might be aware that computer system vary dramatically in prices. My interest in following example is to use regression analysis to predict likely prices that may be charged by retailers. Using variety of sources including retailer websites and local Pennysaver, in December 2001, I compiled information about 40 Desktop systems. Although each computer can be characterized by number of features, I focused on four attributes; central processing unit (CPU) speed in MHz, amount of random access memory in megabytes (RAM), Size of hard disk in gigabytes (HARDDISK) and size of monitor in inches (smallest screen that one can buy is 15inches). My SAS input follows:

```
options nocenter nodate ps=80 ls=80; data pc; input price cpu ram harddisk
monitor retail $ cpu_type $;


 cards;
828.00 1000 128 20 17Selltek EZ Celeron
949.00 1400 128 20 17 Pctek Pentium 4
969.98 1000 256 40 17 Datamatrix Celeron
 978.00  800  256  20  17  Selltek  Power  800Mhz
 Celeron  1009.99  900  128  60  17  FutureShop
eMachines Celeron
1068.00 1000 256 20 17 Selltek Power 1000Mhz Celeron
1128.00 1300 256 20 17 Selltek Power 1300Mhz Pentium 4
1149.99 1400 256 20 17 TCC System #1 Pentium 4
1169.99 1200 256 40 17 TCC System #2 AMD K7
1176.53 1100 128 20 15 Gateway 300Cb Celeron
1199.00 1100 128 40 17 Business Depot HP 7917 / Pavilion Celeron
1229.99 1100 256 20 17FutureShop Compaq 5310 Celeron
1238.53 1000 128 20 15 Gateway E1800 Celeron
1249.00 1100 256 40 17 RadioShack Compaq Presario 5310CA Celeron
1249.98 1500 256 40 17 Datamatrix Pentium 4
1249.99 1000 192 60 17 FutureShop HP XT858 Pentium 3
1249.99 1200 256 40 17 FutureShop Cicero SC2511 Celeron
1269.98 1600 256 40 17 Datamatrix AMD K7
1299.99 1300 128 40 17 FutureShop HP 7935 AMD Athlon
1329.99 1200 256 40 17FutureShop Compaq 5320 Celeron
1349.00 1200 256 40 17 RadioShack Compaq Presario 5320CA Celeron
1378.00 1200 256 40 17 Selltek Ultimate 1200Mhz Pentium 3
1399.00 1100 128 20 17 Dell Dimension 2100 Celeron
1478.00 1600 256 40 17 Selltek Ultimate 1600Mhz Pentium 4
1549.00 1600 256 20 17 Dell Dimension 4300S Pentium 4
1549.99 1200 256 60 17 FutureShop Sony PC540 Celeron
1628.00 1800 256 40 17 Selltek Ultimate 1800Mhz Pentium 4
1649.99 1500 256 60 17 FutureShop eMachines Pentium 4
1749.00 1500 256 60 17 RadioShack Compaq Presario 5330CA Pentium 4
1749.00 1700 256 40 17 Pctek Pentium 4
1749.00 1000 256 40 17 Business Depot Compaq Presario 5330CA Celeron
1849.99 1700 256 60 19 TCC System #3 Pentium 4
1899.00 1500 256 40 17 RadioShack HP 7955/MX70 Pentium 4
```

```
1949.97 1700 256 60 17 FutureShop Cicero SC6411 Pentium 4
2010.48 1500 256 20 17 Gateway 500Sb Pentium 4
2019.99 1700 256 40 17 FutureShop Compaq 5340 Pentium 4
2149.00 1700 512 80 17 Business Depot HP 7965 / Pavilion Pentium 4
2509.00 1900 256 20 19 Dell Dimension 8200 Pentium 4
2649.00 1800 512 60 17 Business Depot Compaq Presario 5350CA Pentium 4
2649.99 2000 512 80 17 FutureShop HP 7975 Pentium 4
;;;;
proc reg;
model price = cpu ram harddisk monitor ; run;
```

SAS output is as follows:

```
Model: MODEL1
Dependent Variable: PRICE

Analysis of Variance

                       Sumof       Mean
Source        DF     Squares      Square     F Value      Prob>F

Model         45896914.7629 1474228.69      22.511     0.0001
Error         352292100.542107
CTotal        398189015.305 65488.5869
                             18
   RootMSE      255.90738  R-square      0.7201
   DepMean    1497.70800   AdjR-sq       0.6881
    C.V.        17.08660

Parameter Estimates

               Parameter    Standard   TforH0:
Variable DF     Estimate       Error  Parameter=0 Prob>|T|

INTERCEP  1  -526.647108 1120.2966356     -0.470     0.6412
CPU       1     0.833024  0.17123187      4.865     0.0001
RAM       1     1.524821  0.61138600      2.494     0.0175
HARDDISK  1     2.781324  2.87411046      0.968     0.3398
MONITOR   1    24.098373 69.51531811      0.347     0.7309
```

Here are my observations in point form.

- ThenullhypothesisstatesthatvariationinpricecannotbeexplainedbyCPUspeed,amount RAM, size of hard disk and size of monitor. We reject this hypothesis, because probability of F-statistic is less than or equal to 0.05.

- Weareexplainingabout72%ofvariationinpricebythesefourvariables.

- Regression equation can be written as

$$\text{Price} = -526.65 + 0.833 \times \text{CPU} + 1.525 \times \text{RAM} + 2.781 \times \text{HARDDISK} + 24.098 \times \text{MONITOR}.$$

- NotethattheparameterassociatedwithvariablesCPUandRAMhavecorrectsigns[4]and statistically significant (probability of t-statistic is less than 0.05).

- TheparametersassociatedwithvariablesHARDDISKandMONITORhavecorrectsignbut statistically not significant. That means, these parameters could be equal to zero.

- Consideradesktopwith1Ghz,with256MegabytesofRAM,about40gigabytesharddrive

and 17 inches MONITOR. For such machine, I should be expected to pay about $1,218. This is concluded as follows:

$$\begin{aligned} \text{Price} &= -526.65 + 0.833 \times 1000 + 1.525 \times 256 + 2.781 \times 40 + 24.098 \times 17. \\ &= -526.65 + 833 + 390.4 + 111.24 + 409.67 \\ &= 1217.66 \end{aligned}$$

Note that holding everything else same, if we decide to purchase desktop computer with 1.5 Ghz CPU, price of computer would go up by $416.5. A constructed equation like this would be useful tool to understand competitive market behaviour. Let us turn our attention to evaluating assumptions. First SAS input and then followed by relevant output.

```
proc reg data=pc;
 model price = cpu ram harddisk cdrom / collinoint r dw influence vif ;
 id brand;
 output out=predpc
 residual=resprc student=stprc dffits=dfprc covratio = covprc;
run;
```

Obs	RETAIL	Dep Var PRICE (1)	Predict Value (2)	Std Err Predict (3)	Residual (4)	Std Err Residual (5)	Student Residual (6)
1	Selltek	828.	966.9	75.171	-138.	244.618	-0.568
2	Pctek	0	1300.1	84.236	9	241.646	-1.453
3	Datamat	949.	1217.7	74.704	-351.	244.761	-1.012
		0			1		
		970.			-247.		
		0			7		

[4] I would think that desktops with faster CPUs should be more expensive than slower CPUs.

4	978.0	995.4	117.797 -17.4251	227.184	-0.077
Selltek	1010.0	994.8	125.129 15.1867	223.229	0.068
5	1068.0	116 2.	092.814 -94.0298	238.483	-0.394
FutureS	1128.0	141 1. 9	71.552 -283.9	245.701	-1.156
6	1150.0	149 5. 2	71.607 -345.2	245.685	-1.405
Selltek	1170.0	138 4. 3	49.600 -214.3	251.055	-0.853
7 10 Gateway	1176.5	100 2. 0	142.267 174.6	212.718	0.821
Selltek 11 Busines	1199.0	110 5. 8	77.940 93.2184	243.750	0.382
8TCC 12 FutureS	1230.0	124 5. 3	82.844 -15.3422	242.127	-0.063
9TCC 13 Gateway	1238.5	918.7	138.660 319.9	215.086	1.487
14 RadioSh	1249.0	130 1. 0	61.051 -51.9587	248.518	-0.209
15 Datamat	1250.0	163 4. 2		251.617	-1.527
16 FutureS	1250.0	117 5. 7	46.662 -384.2	235.091	0.316
17 FutureS	1250.0	138 4. 3	101.098 74.2958	251.055	-0.535
18 Datamat	1270.0	171 7. 5	49.600 -134.3	249.465	-1.794
19 FutureS	1300.0	127 2. 4	57.060 -447.5	242.685	0.114
20 FutureS	1330.0	138 4. 3	81.196 27.6037	251.055	-0.216
21 RadioSh	1349.0	138 4. 3	49.600 -54.2711	251.055	-0.140
22 Selltek	1378.0	138 4. 3	49.600 -35.2611	251.055	-0.025
23 Dell	1399.0	105 0. 2	49.600 -6.2611	245.675	1.420
24 Selltek	1478.0	171 7.	571.640 348.8	249.465	-0.960
25 Dell	1549.0	166 1.	857.060 -239.5	242.045	-0.466
26 FutureS	1550.0	143 9.	983.082 -112.8	244.250	0.451
27 Selltek	1628.0	188 4.	176.358 110.1	241.488	-1.060
28 FutureS	1650.0	168 9.	884.689 -256.1	245.454	-0.162
29 RadioSh	1749.0	168 9.	872.395 -39.8047	245.454	0.241
30 Pctek	1749.0	180 0.	872.395 59.2053	246.105	-0.210
31 Busines	1749.0	121 7.	770.148 -51.7730	244.761	2.171
32 TCC	1850.0	190 4. 6	74.704 531.3	211.118	-0.259
33 RadioSh	1899.0	163 4. 2	144.629 -54.6063	251.617	1.053
34 FutureS	1950.0	185 6. 4	46.662 264.8	240.226	0.390
35 Gateway	2010.5	157 8. 5	88.206 93.5705	244.472	1.767
36 FutureS	2020.0	180 0. 8	75.643 431.9	246.105	0.891
37 Busines	2149.0	230 2. 4	70.148 219.2	217.482	-0.705
38 Dell	2509.0	195 9. 9	134.871 -153.4	199.216	2.756
39 Busines	2649.0	233 0. 1	160.628 549.1	221.193	1.442
40 FutureS	2650.0	255 2. 3	128.694 318.9	216.323	0.452
			136.722 97.7026		

① column is values of dependent variable (yi). This variable is sorted in ascending order to help us interpret other statistical measures.

② column is predicted values for dependent variable (ˆyi). For the first observation,

$$\hat{y}_1 = -526.65 + 0.833 \times 1000 + 1.525 \times 128 + 2.781 \times 20 + 24.098 \times 17 = 966.9.$$

③ column is the standard error associated with predicted values, a larger number indicates that values of independent variables are farther away from the “average” observation. For the first observation independent variable vector, x1 is [110001282017]. Then Var(y1) = s2x01(X0X)−1x.

④ column is residual or error values, (yi − yˆi).

⑤ column is the standard error associated with error, and again a larger number indicates that values of independent variables are farther away from the “average” observation.

⑥ column the Student residuals are also called normalized (generally normalized means divided by the standard error) residuals. If residuals are normally distributed then normalized residuals more than 2 should be considered extreme observations.

Obs	RETAIL	⑦ -2-1-0 1 2	⑧ Cook's D	⑨ Rstudent	⑩ HatDiag H	⑪ Cov Ratio
1	Selltek	| *| |	0.006	-0.5621	0.0863	1.2080
2	Pctek	| **| |	0.051	-1.4771	0.1083	0.9499
3	Datamat	| **| |	0.019	-1.0123	0.0852	1.0893
4	Selltek	| | |	0.000	-0.0756	0.2119	1.4655
5	FutureS	| | |	0.000	0.0671	0.2391	1.5182
6	Selltek	| | |	0.005	-0.3895	0.1315	1.3018
7	Selltek	| **| |	0.023	-1.1614	0.0782	1.0323
8	TCC	| **| |	0.034	-1.4258	0.0783	0.9381
9	TCC	| *| |	0.006	-0.8501	0.0376	1.0812
10	Gateway	| |* |	0.060	0.8168	0.3091	1.5181
11	Busines	| | |	0.003	0.3777	0.0928	1.2478
12	FutureS	| | |	0.000	-0.0625	0.1048	1.2906
13	Gateway	| |** |	0.184	1.5144	0.2936	1.1807
14	RadioSh	| | |	0.001	-0.2062	0.0569	1.2181
15	Datamat	| ***| |	0.016	-1.5577	0.0332	0.8471
16	FutureS	| | |	0.004	0.3119	0.1561	1.3503
17	FutureS	| *| |	0.002	-0.5293	0.0376	1.1528
18	Datamat	| ***| |	0.034	-1.8553	0.0497	0.7511
19	FutureS	| | |	0.000	0.1121	0.1007	1.2830
20	FutureS	| | |	0.000	-0.2132	0.0376	1.1931
21	RadioSh	| | |	0.000	-0.1385	0.0376	1.1977
22	Selltek	| | |	0.000	-0.0246	0.0376	1.2010
23	Dell	| |** |	0.034	1.4417	0.0784	0.9323
24	Selltek	| *| |	0.010	-0.9588	0.0497	1.0645
25	Dell	| | |	0.005	-0.4609	0.1054	1.2525
26	FutureS	| | |	0.004	0.4456	0.0890	1.2325
27	Selltek	| **| |	0.028	-1.0624	0.1095	1.1026
28	FutureS	| | |	0.000	-0.1599	0.0800	1.2518
29	RadioSh	| | |	0.001	0.2379	0.0800	1.2461
30	Pctek	| | |	0.001	-0.2075	0.0751	1.2420
31	Busines	| |**** |	0.088	2.3001	0.0852	0.6132
32	TCC	| | |	0.006	-0.2552	0.3194	1.6823
33	RadioSh	| |** |	0.008	1.0542	0.0332	1.0181
34	FutureS	| | |	0.004	0.3847	0.1188	1.2836
35	Gateway	| |*** |	0.060	1.8247	0.0874	0.7939
36	FutureS	| |* |	0.013	0.8881	0.0751	1.1145
37	Busines	| *| |	0.038	-0.7001	0.2778	1.4900
38	Dell	| |***** |	0.988	3.0699	0.3940	0.5613
39	Busines	| |** |	0.141	1.4654	0.2529	1.1392
40	FutureS	| | |	0.016	0.4465	0.2854	1.5711

⑦ column is a plot of normalized residuals and these numbers generally vary between −2 and 2.

⑧ column Cook's D is a summary measure of the influence of a single observation on the total changes in all other residuals when observation is excluded from the estimation. In our case, Cook's $D \geq \frac{8}{N-2(k+1)}$ is $\frac{8}{40-10}$ or 0.267 would be considered influential observation (see observation number 38).

⑨ column Rstudent is similar to Cook's D with the exception that error variances are estimated using without the ith observation.

⑩ column Hat Diag H (Diagonal of Hat matrix H, also sometimes denoted as hii) is a ratio of variability for an observation to the sample variability in independent variables. If each observation has equal influence on regression equation, then the average influence would be k/N and observation with hii ≥ 2k/N (2 ×4/40 or 0.2 for our example) would be considered an influential observation. There are number of observations with such problem, especially towards the end of dataset or higher priced desktop systems.

(11) column Cov ratio (Covariance ratio) is a ratio covariances when ith observation is excluded to the sample covariances. A value of COVRATIO close to 1 indicates the "average" influence by an observation while the absolute value of (COVRATIO - 1) $\geq \frac{3(k+1)}{N-k-1}$ is considered significant influential observation. For our case, COVRATIO ≥ $1 + (335\times5)$ or 1.429 would be observations with higher than the normal influence.

Obs	RETAIL	(12) Dffits	(13) INTERCEP Dfbetas	(14) CPU Dfbetas	RAM Dfbetas	HARDDISK Dfbetas	MONITOR Dfbetas
1	Selltek	-0.1727	0.0247	0.0545	0.0464	0.0452	-0.0475
2	Pctek	-0.5149	-0.0181	-0.2738	0.3160	0.1419	0.0082
3	Datamat	-0.3090	0.0441	0.2592	-0.1324	-0.0097	-0.0805
4	Selltek	-0.0392	0.0064	0.0313	-0.0246	0.0178	-0.0112
5	FutureS	0.0376	-0.0033	-0.0141	-0.0194	0.0289	0.0060
6	Selltek	-0.1516	0.0202	0.0975	-0.0948	0.0900	-0.0370
7	Selltek	-0.3382	0.0097	0.0394	-0.1628	0.2732	-0.0289
8	TCC	-0.4156	-0.0079	-0.0510	-0.1541	0.3406	-0.0035
9	TCC	-0.1679	0.0129	0.0963	-0.0550	-0.0019	-0.0286
10	Gateway	0.5463	0.4983	0.1465	-0.1210	-0.0494	-0.4755
11	Busines	0.1208	-0.0088	-0.0148	-0.0839	0.0572	0.0186
12	FutureS	-0.0214	0.0023	0.0110	-0.0129	0.0144	-0.0044
13	Gateway	0.9763	0.8897	0.1480	-0.1664	-0.0844	-0.8332
14	RadioSh	-0.0507	0.0060	0.0378	-0.0200	-0.0012	-0.0116
15	Datamat	-0.2889	-0.0400	-0.1420	0.0457	0.0129	0.0500
16	FutureS	0.1341	-0.0109	-0.0632	-0.0440	0.1029	0.0203
17	FutureS	-0.1046	0.0081	0.0600	-0.0343	-0.0012	-0.0178
18	Datamat	-0.4244	-0.0735	-0.2979	0.1135	0.0222	0.1011
19	FutureS	0.0375	0.0006	0.0114	-0.0323	0.0162	0.0005
20	FutureS	-0.0421	0.0032	0.0242	-0.0138	-0.0005	-0.0072
21	RadioSh	-0.0274	0.0021	0.0157	-0.0090	-0.0003	-0.0047
22	Selltek	-0.0049	0.0004	0.0028	-0.0016	-0.0001	-0.0008
23	Dell	0.4204	-0.0429	-0.0386	-0.1647	-0.1206	0.0891
24	Selltek	-0.2193	-0.0380			0.0115	0.0522
25	Dell	-0.1582	-0.0157	-0.1540	0.0586	0.1151	0.0198
26	FutureS	0.1393	-0.0039	-0.0820	-0.0206	0.1059	0.0096
27	Selltek	-0.3726	-0.0736	-0.0573	-0.0179	0.0209	0.1082
28	FutureS	-0.0472	-0.0053	-0.3269	0.1364	-0.0361	0.0073
29	RadioSh	0.0702	0.0079	-0.0130	0.0218	0.0537	-0.0109
30	Pctek	-0.0591	-0.0112	0.0193	-0.0324	0.0033	0.0161
31	Busines	0.7020	-0.1003	-0.0482	0.0195	0.0221	0.1828
32	TCC	-0.1748	0.1476	-0.5891	0.3008	-0.0555	-0.1411
33	RadioSh	0.1955	0.0271	-0.0115	0.0553	-0.0088	-0.0339
34	FutureS	0.1413	0.0240	0.0961	-0.0309	0.0859	-0.0357
35	Gateway	0.5646	0.0358	0.0868	-0.0788	-0.4446	-0.0366
36	FutureS	0.2531	0.0480			-0.0140	-0.0689
37	Busines	-0.4342	-0.0433	0.1934	0.1394	-0.1292	0.0694
38	Dell	2.4752	-1.8277	0.2063	-0.0835	-1.0910	1.7273
39	Busines	0.8526	0.1007	0.0669	-0.2616	-0.1207	-0.1584
40	FutureS	0.2822	0.0490	0.7657	-0.1447	0.0774	-0.0786
				-0.6588			
				-0.0045			
				0.1189			
				0.0631			

(12) column Dffits indicates influence of an observation on the overall fit of model. DFFITS outside of range $\pm 2\sqrt{(k-1)/N}$ is considered influential observation. In our case, $\pm 2\sqrt{3/40}$ or ±0.548 would be an influential observations.

13 14 columns DFBETAs indicate influence of particular observation on a specific parameter estimate. Observations outside $\pm 2/\sqrt{N}$ would influencing particular observations. In our case, the appropriate range is $\pm 2/\sqrt{40}$ or ±0.316. There will one DFBETA for each parameter estimated. In our case there are five such measures.

```
               15
               Variance
Variable  DF   Inflation

INTERCEPT 1   0.00000000
CPU       1   1.67434954
RAM       1   1.87908442
HARDDISK  1   1.46192377
MONITOR   1   1.18063429


Collinearity Diagnostics(intercept adjusted)
           16            17          18
                  Condition Var Prop Var Prop Var PropVar Prop
Number Eigenvalue      Index CPU      RAM      HARDDISK MONITOR

    1    2.18504   1.00000   0.0823  0.0783   0.0767  0.0513
    2    0.89578   1.56181   0.0180  0.0542   0.1508  0.6471
    3    0.57072   1.95667   0.3952  0.0459   0.5049  0.2253
    4    0.34845   2.50413   0.5045  0.8216   0.2676  0.0762

Durbin-WatsonD           1.634 19
(For Numberof Obs.)         40
1stOrderAutocorrelation  0.177 20

SumofResiduals
Sum of Squared Residuals 2292100.5421
Predicted Resid SS (Press) 3350277.8100
```

15 column variance inflation is a measure of collinearity among independent variables and a larger number indicates that variables highly correlated. This does not appear to be a problem in our illustration.

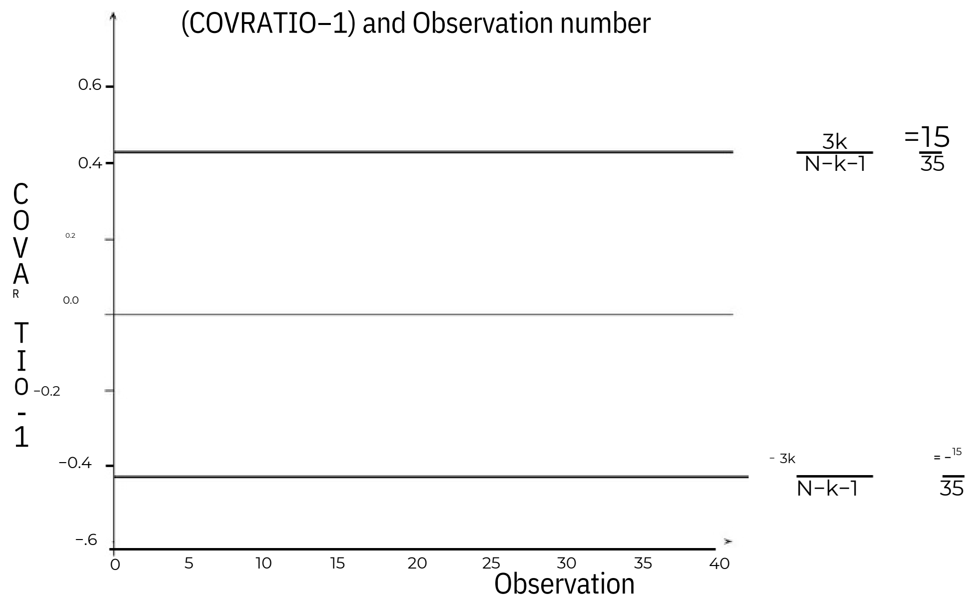
16 column eigenvalue is another measure of degree to which independent variables are correlated. (see the next item for interpreting these).

17 column condition index is square root of the ratio of largest eigenvalue to a particular eigenvalue.

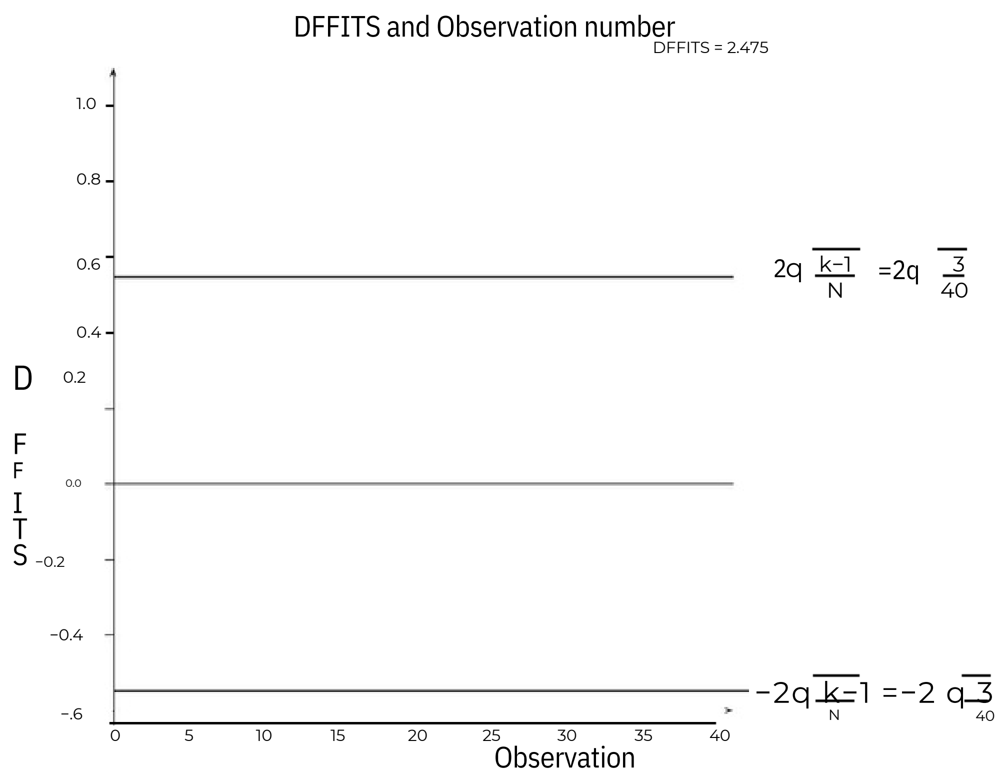
18 columns var prop (proportion of variance shared) is degree to which two or more variables have common variability.

19, 20 are measures of whether successive error terms are correlated.

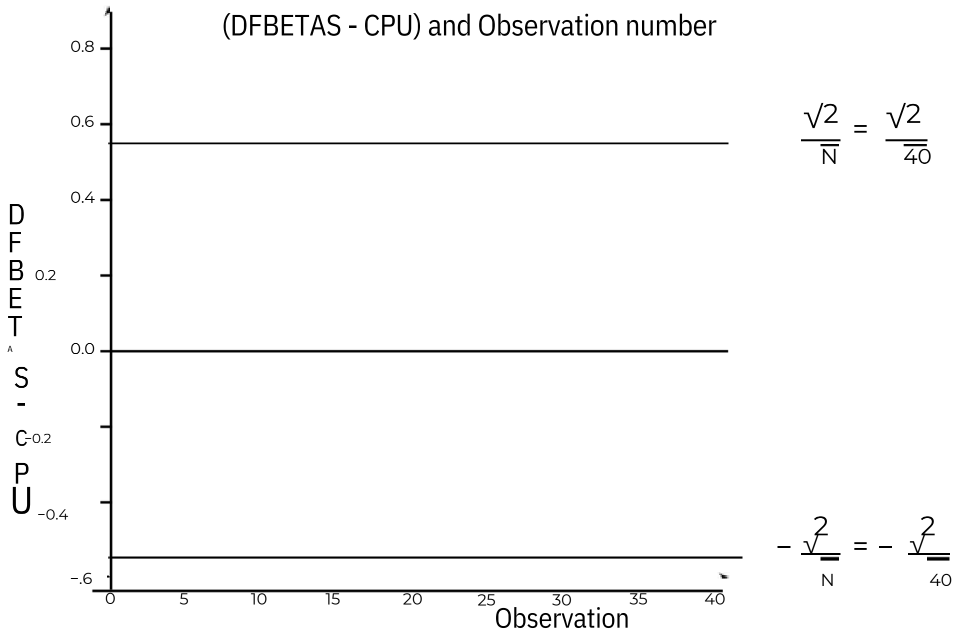
There is a graphical alternative to visualizing various diagonistics discussed above. Consider measure COVRATIO. If observations are sorted in ascending or descending order, then plot of COVRATIO and observation number could be used to visually understand nature of violations related to this measure. Several of such graphs are provided for illustrative purposes.



Note that there are seven observations outside the limits.



Note that there are four observations outside the limits and observation number 38 is particularly noteworthy.



Note that there are two observations outside the limits.

Testing Normality

The purpose of this material is to provide procedures that can be used to evaluate the univariate normality. If tests reveal problems, then it is advisable to turn to the alternative approaches to analysis, including transformation or weighted least squares.

The moments around the mean of a distribution reveal departures from normality. Suppose we have a random variable y with a population mean of μ_1, then the rth moment about the mean is defined as

$$\mu_r = E(y - \mu_1)^r, \qquad \text{for } r > 1,$$

where E is used to denote the expected value or the average. If we know mean (μ_1) and its variance (μ_2), then it is possible to describe the univariate normal distribution. This is because its higher-order moments are either zero or can be written as functions of mean or variance. Consequently, if we examine and test higher order moments, it should be possible to detect departures from normality. We will look at the second, third and fourth moments for a sample and population below.

The Sample Variance

The population variance (μ_2) is the expected value of the squared difference of the values from the population mean:

$$\mu_2 = E(y - \mu_1)^2.$$

The sample variance (s^2) is usually computed as

$$s^2 = \frac{1}{(N-1)} \sum_{i=1}^{N} (y_i - \bar{y})^2.$$

Test of Skewness to Detect Non-normality

Skewness is a measure of the tendency of the deviations to be larger in one direction than in the other.

A population skewness is defined as

$$\frac{E(y-\mu_1)^3}{\mu_2^{3/2}}.$$

The sample third moment (g_1) is computed as[5]

$$g_1=\frac{N}{(N-1)(N-2)}\frac{\sum_{i=1}^{N}(y_i-\bar{y})^3}{s^3}.$$

The coefficient of skewness (CS) or $\sqrt{b_1}$ is

$$CS=\sqrt{b_1}=\frac{\sqrt{N(N-1)}}{N-2}g_1.$$

For a normally distributed variable, CS is 0. Moreover, if CS is negative and statistically significant, then skew is to the left. On the other hand, if CS is positive, then skew is to the right. In the large samples, hypothesis test for CS can be performed by converting CS as a unit normal deviate. That is,

$$z_{\sqrt{b_1}}=\pm CS\sqrt{\frac{(N+1)(N+3)}{6(N-2)}}$$

where the undetermined sign is the same as that of the third moment and this quantity is approximately normally distributed under the null hypothesis of population normality.

Tests of Kurtosis to Detect Non-Normality

The heaviness of the tails is measured by kurtosis or the coefficient of kurtosis (b2). The population kurtosis is defined as

$$\frac{\mu_4}{\mu_2^2}=\frac{E(y-\mu_1)^4}{\mu_2^2}-3.$$

The sample fourth moment is calculated as

$$g_2=\frac{N(N+1)}{(N-1)(N-2)(N-3)}\frac{\sum_{i=1}^{N}(y_i-\bar{y})^4}{s^4}-\frac{3(N-1)^2}{(N-2)(N-3)}.$$

To convert fourth moment to kurtosis (b2) we need to compute

$$b_2=3\frac{N-1}{N+1}+\frac{(N-2)(N-3)}{(N+1)(N-1)}g_2.$$

For a normally distributed variable, b2 is equal to 3. In large samples, hypothesis test for b2 can be performed by converting b2 as a unit normal deviate. That is,

$$z_{b_2}=\left[b_2+\frac{6}{(N+1)}\right]\sqrt{\frac{(N+1)^2(N+3)(N+5)}{24N(N-2)(N-3)}}.$$

[5] PROC UNIVARIATE in SAS reports the third and fourth moments but not coefficent of skewness and kurtosis as indicated below.

and this estimate is approximately normally distributed under the null hypothesis of population nor-mality. Note that values less than zero indicate that the distribution is more peaked with longer tails than the normal distribution; values greater than zero indicate flatter distribution in the centre and with shorter tails than the normal distribution.

Omnibus Tests of Normality

It is possible to combine test of skewness and kurtosis into one test that detects departure from normality due to either of these measures. Such tests are called *omnibus*. The test statistic

$$K^2 = z^2\sqrt{b_1} + z^2 b_2$$

where the K^2 statistic has approximately a chi-square (χ^2) distribution, with 2 degrees of freedom when the population is normally distributed.

There are many other tests to determine departure of a variable from normality. The program NORMTEST also prints statistic called Shapiro-Wilk test[6]. It is based on assumption that ordered observations of normally distributed variable will have equal and similar weights. Thus, if weight assigned to the first observation (the lowest value of y_i, let us call it $y_{(1)}$) is 1/N and the second observation (one that is more than or equal to $y_{(1)}$, let us call it $y_{(2)}$) will have weight of 2/N and so on[7]. The test statistic of Shapiro-Wilk (W) is

$$W = \frac{\left(\sum_{i=1}^{N} a_i y_{(i)}\right)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}$$

where a_i is weight associated with i observation and variable y is ordered such that $y_{(1)} \le y_{(2)} \le \cdots \le y_{(N)}$.

Small values of W correspond to departure from normality.

We will examine below SAS input and output to conduct these tests. As you have seen above, numerical calculations involved in above are extensive. To assist you with these calculations, I have a SAS macro[8] To access this macro, I would use following SAS input.

```
%include "c:\sas6_12\normtest.sas";
%normtest(stprc,predpc);
```

In this instance predpc is name of SAS dataset and stprc is a variable whose normality is being tested. SAS will produce two sorts of outputs; one graphical and another textual. These follow here. First SAS and then graphical output.

```
Normality Test for variable stprc N=40
```

[6] Shapiro, S. S. and Wilk M. B. (1965) "A analysis of variance test for Normality", *Biometrika*, vol. 52, 591–611.

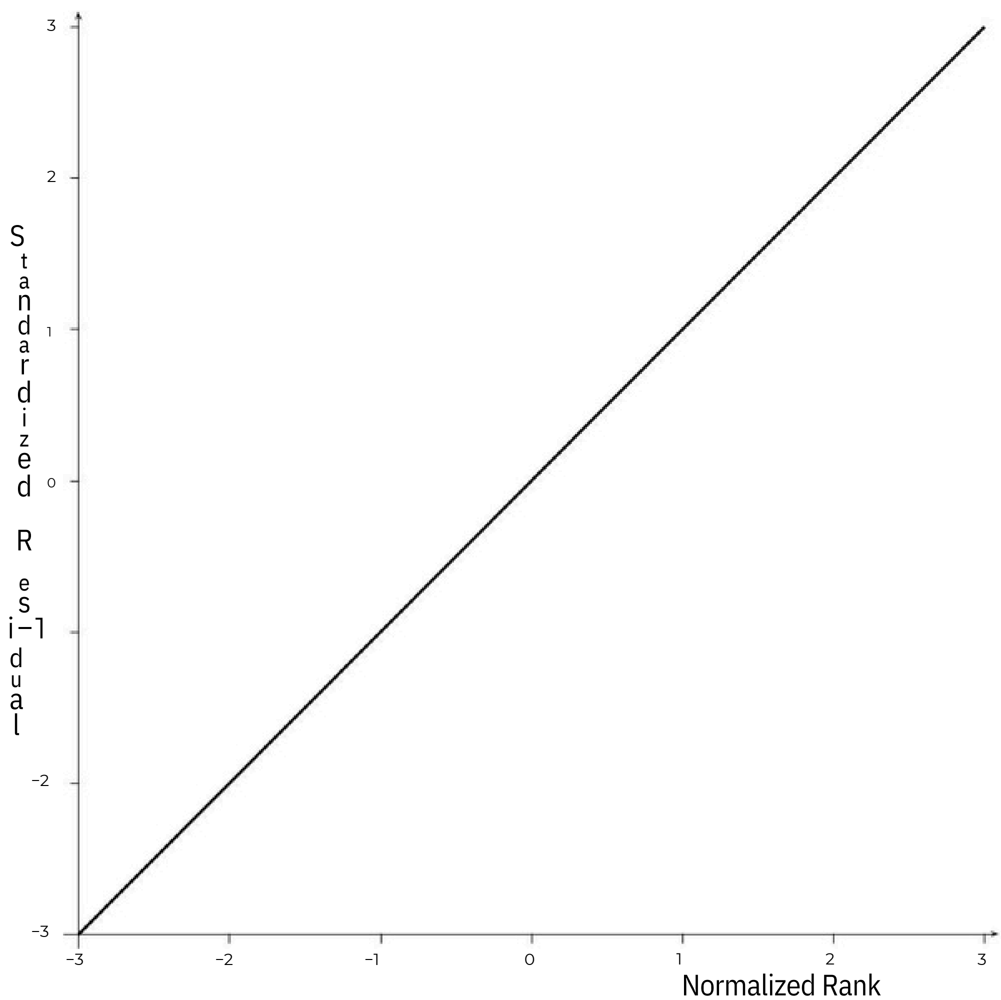
[7] This is intuitive description of the statistic and not the exact method.

[8] This macro is modified version of as it appeared in *American Statistician* and it was originally written by D'Agostino Ralph B., Albert Belanger and Ralph B. D'Agostino Jr. (1990) "A Suggestion for Using Pow-erful and informative Tests of Normality", Vol. 44, pp. 316–321. The macro for your usage is kept in file
G:\courses\COST6060\NORMTEST.SAS

```
G1=0.592      SQRTB1=0.569      z=    1.598prob=0.1101
G2=0.239      B2=3.064          z=    0.554prob=0.5798

          K**2 = Chisquare (2 df) = 2.860 prob = 0.2393
          Shapiro-Wilk Test = 0.966Prob = 0.3704
```

These numbers indicate that residuals have slight skew to the right (since $\sqrt{b_1}$ is 0.569) and we would accept the null hypothesis that residuals are normally distributed. We also conclude that the coefficient of kurtosis is close to normally distributed variable. Both K2 and Shapiro-Wilk test indicate that we may accept null hypothesis that residuals are normally distributed.



Revise Model to meet Assumptions

- Failure of Similar variation or equal influence
    1. Transformdependentvariable,orindependentvariableorboth.
    2. Excludeobservationswithmoreinfluence.
    3. Apply weighting scheme that are managerially or statistically meaningful.
    4. Estimatemodelwithweightedleastsquaresortheleastabsolutedeviation.
- Presence of Collinearity
    1. Createnewindexvariablesthatmaycapturecorrelationsamongindependentvariableseither conceptually (for example SES, instead income, occupation and education etc.)
    2. Determinestabilityofparametersbyexcludingoneormorevariables.
    3. Usestatisticalproceduresfordealingwiththisproblem,forexample,transformation,alternative criterion to minimize.
- Lack of Independence of successive error values
    1. Maybecausedbymissingvariables,competitivevariablesorcustomerloyalty,theninclude missing variables.
    2. Re-estimatemodelwithautocorrelatederrors.
- Error values not normally distributed
    1. Usenon-normaldistributiontoestimateparameters.
    2. Usetransformationconvertdependentvariablesothatnewvariableisnormallydistributed.
    3. Break sample in subsegments and estimate parameters for each subsegment.

Validate Revised Model

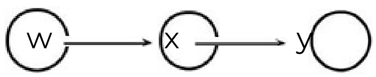
1. Use limited number of explanatory variables. Avoid using all variables to be included in your regression model. If there are large number of variables, then create indices, groupings with conceptual idea. Then, use selected such variables to estimate models.

2. Usealargesample,40-50observationspervariableincludedwillhavebetterstabilitytoestimates than 5 - 10 observations.

3. Validateyourmodelwithhold-outorsplit-halfsampleornewsample.

Limitations of Regression Analysis

- Moderatingeffectsofvariables
  1. Bygroupdifferences,
  2. Interactioneffects,
  3. Effectoccuronlyatcertainlevel.
- Mediating effects of variables. I will indicate first by picture that variable x affects y and variable w affects x. If you include, say variable w and regressed on y, we may get unexpected results.



Alternatively, this could be written in form of equations as follows.

$$y = a+bx+ey$$
$$x = c+dw+ex$$

- Not-linear effects.
- Effects associated with data collection.
  1. Measurementerrors,
  2. Responseeffects,
  3. Truncationofvariables.

Estimating Regression Model using Matrix Algebra

A simple model that you may be familiar with, viz.,

$$y = X\beta + u \quad (1)$$

where y is $(y_1, y_2, \cdots, y_N)0$, u is $(u_1, u_2, \cdots, u_N)0$, and β is $(\beta_1, \beta_2, \cdots, \beta_k)0$ are vectors, and X $(X_1, X_2, \cdots, X_N)0$ is matrix and 0 is used to denote transpose of a matrix or a vector. In this model, vector y of size $N \times 1$ is called dependent variable and matrix X of size $N \times k$ is a set of independent variables. In estimation of parameter vector β, I am interested in the "best" possible estimate. In the following discussion I want to demonstrate to you that two of the commonly used estimators, least squares and maximum likelihood, are the same for the above model.

Least Squares Estimator

In the least squares method, I want to find $\hat{\beta}$ of the regression parameter β so as to minimize the sum of squared residuals. Mathematically I may write

$$\text{Minimize} f(\beta) = (y - X\beta)'(y - X\beta) \\ = y'y - 2y'X\beta + \beta'X'X\beta \tag{2}$$

To minimize this function, I obtain the first derivative of $f(\beta)$ with respect to β and set equal to zero. Thus, I may write

$$\frac{\partial f}{\partial \beta} = -2X'y + 2X'X\beta = 0 \text{ or}$$

$$\hat{\beta} = (X'X)^{-1}X'y \tag{3}$$

It is can be shown that that $E(\hat{\beta}) = \beta$ and $V(\hat{\beta}) = \sigma^2(X'X)^{-1}$ where E and V denote statistical expectation and variance respectively.

I made four important assumptions in deriving these estimates. First, it is assumed that $E(u) = 0$ and implies that the mean of random noise is zero. Second, it is also assumed that $E(X'u) = 0$ and implies that random noise values and independent variable values are not correlated. Third assumption requires that $E(uu') = \sigma^2 I_N$ where I_N denotes an identity matrix of size $N \times N$. In words, this assumption requires that each element of random noise vector u be independent and identically distributed. This assumption is clearly violeted if the observed dependent variable takes either 0 or 1 values. (As an excercise you may show this). Similarly, if sucessive values of dependent variable are related, as in case of time series data, then this assumption is also violeted. Finally, matrix $(X'X)$ is nonsingular, which is equivalent to stating rank of matrix X is k. Note that a mere presence of high correlation among the set of independent variables does not violet this assumption.

It is also possible to show (with lot of algebraic manipulation) that the estimated value of σ^2 is $(\hat{u}'\hat{u})/(N - k)$. Note also that second derivatives of $f(\beta)$ with respect to β are positive. This assures me that I have actually minimized the function.

Maximum Likelihood Estimation

Suppose I assume further that u vector is normally distributed. This is an extension to the third assumption that I have written above. Then, likelihood of observing u_1 is given by

$$f(u_1) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{u_1^2}{2\sigma^2}\right) \tag{4}$$

If there are N independent observations, then the joint likelihood of observing $f(u_1), f(u_2), \cdots, f(u_N)$ will be denoted by L and may be written as

$$
\begin{aligned}
L &= f(u_1) \times f(u_2) \times \cdots \times f(u_N) \\
&= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{u_i^2}{2\sigma^2}\right) \\
&= (2\pi\sigma^2)^{-N/2} \exp\left(-\frac{\sum_{i=1}^{N} u_i^2}{2\sigma^2}\right)
\end{aligned} \tag{5}
$$

Instead of using likelihood, it is customary in the literature to use logarithm of likelihood. Thus taking the logarithm of equation (5), I may obtain

$$\log L = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N} u_i^2.$$

Above equation in matrix form can be written as

$$
\begin{aligned}
\log L &= -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}u'u \\
&= -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)
\end{aligned} \tag{6}
$$

The maximum likelihood estimator of the regression parameter vector is an estimator that maximizes likelihood function (or log of likelihood function). To maximize log L, I would take the derivatives of log L with respect to β and σ^2 and set equal to zero. Thus, I may write

$$
\begin{aligned}
\frac{\partial \log L}{\partial \beta} &= -\frac{1}{2\sigma^2}(-2X'y + 2X'X\beta) = 0 \\
\frac{\partial \log L}{\partial \sigma^2} &= -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}(y - X\beta)'(y - X\beta) = 0
\end{aligned}
$$

Solving for $\hat{\beta}$ and $\hat{\sigma}^2$ I may obtain

$$
\begin{aligned}
\hat{\beta} &= (X'X)^{-1}X'y \quad \text{and} \\
\hat{\sigma}^2 &= \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{N}
\end{aligned}
$$

Although the estimate of vector β using the least squares and maximum likelihood method is same, the estimate of σ^2 is not equal. In fact, $\hat{\sigma}^2$ estimate based on the maximum likelihood method is biased and the estimate based on the least squares method is unbiased. Finally, note also that second derivatives of log L with respect to β and σ^2 are negative. This assures me that I have actually maximized the function.

Finally, it is possible to obtain logL value if $u'u$ is known from the least squares estimation procedure. To obtain this, substitute unbiased value of $\hat{\sigma}^2$ in the expression of logL. Thus, we may write

$$
\begin{aligned}
\log L &= -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\frac{(u'u)}{N-k} - \frac{(u'u)}{2u'u / (N-k)} \\
&= -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\frac{(u'u)}{N-k} - \frac{N-k}{2}
\end{aligned} \tag{7}
$$

In expression (7) $u'u$ is sums of squares of residuals and remaining terms contain known constants. Thus, it is possible to obtain logarithm of likelihood, if one knows sums of squares, criterion used in the least squares method.

Formulae for Various Quantities Reported in Regression Analysis

I am now in position to summarize various formulae that one normally encounters when using a regression program. Here N refers to length of vector y and k refers to length of β vector excluding constant term. So k = 4 means that there are four independent variables and N = 24 means that I had 24 observed values of dependent variable.

Mean of Dependent Variable is also called expected value of random variable;

$$E(y) = y'1/N \text{ or } \frac{\sum_{i=1}^{N} y_i}{N}$$

Standard Deviation of Dependent Variable is

$$\left(\frac{y'y}{N-1} - \frac{y'1}{N-1}\right)^{1/2} = \left[E(y^2) - [E(y)]^2\right]^{1/2}$$

Sum of Squared Residuals is $\hat{u}'\hat{u} = \sum_{i=1}^{N} \hat{u}_i^2 = y'y - \hat{\beta}'X'y$. This is the quantity minimized in the least squares method. You may prove that the equality is true.

Standard Error of Regression is

$$\left[\frac{\hat{u}'\hat{u}}{N-k}\right]^{1/2} = \left[\frac{\sum_{i=1}^{N} \hat{u}_i^2}{N-k}\right]^{1/2}$$

R^2 is always between zero and one and is computed

$$1 - \frac{\hat{u}'\hat{u}/N}{y'y/N} \text{ or } 1 - \frac{\sum_{i=1}^{N} \hat{u}_i^2/N}{\sum_{i=1}^{N} y_i^2/N}$$

It is known that R^2 is an increasing function of number of independent variables in the model.

R^2_{adj} is an improvement over R^2 so as to adjust for the number of variables in the model. It is computed as

$$1 - \frac{\hat{u}'\hat{u}/(N-k-1)}{y'y/(N-1)} \text{ or } 1 - \frac{\sum_{i=1}^{N} \hat{u}_i^2/(N-k-1)}{\sum_{i=1}^{N} y_i^2/(N-1)}.$$

Durbin-Watson Statistics is commonly used statistics to test whether successive values of random noise are related to each other. It is estimated by

$$dw = \frac{\sum_{i=2}^{N} (\hat{u}_i - \hat{u}_{i-1})^2}{\sum_{i=1}^{N} \hat{u}_i^2},$$

and expected value of this statistics for a normally distributed random variable is 2.

Estimated Autocorrelation or correlation among successive observation is

$$\rho = \frac{\sum_{i=2}^{N} \hat{u}_i \hat{u}_{i-1}/(N-1)}{\sum_{i=1}^{N} \hat{u}_i^2/N},$$

and expected value of this statistics for a normally distributed random variable is 0.

F-statistics is used to test whether β vector is significantly different from zero and it is the ratio of mean sums of squares due regression to the error mean sums of squares, i.e.

$$\frac{\hat{\beta}' X' y/k}{\hat{u}'\hat{u}/(N-k)}.$$

This statistics is distributed according to F-distribution with k and $(N-k)$ degrees of freedom.

Standard Error of Coefficient is

$$SEC_i = \sqrt{\frac{\sum_{i=1}^{N} \hat{u}_i^2}{(N-k)}} \sqrt{a_{ii}}$$

where a_{ii} are diagonal elements of $(X'X)^{-1}$ matrix.

t-statistics is $\frac{\hat{\beta}_i - \beta_i}{SEC_i}$ and this is distributed according to t-distribution with $(N-1)$ degrees of freedom. Note that expected value of β_i in above expression is zero.

Cook's Distance (CDi) is a measure of the change in the regression coefficents that would occur if a ith case is omitted. The measure reveals observations that are most influential in affecting estimated regression equation. It is affected by both the case being an outlier on dependent variable and on the set of predictors. It is computed as

$$CD_i = \frac{(\hat{\beta} - \hat{\beta}_{(-i)})'(X'X)(\hat{\beta} - \hat{\beta}_{(-i)})}{(k+1)\, MS_{res}},$$

where $\hat{\beta}_{(-i)}$ is the vector of estimated regression coefficients with the ith observation deleted, and MS_{res} is the residual variance for all the observations. It is easier to compute Cook's D by

$$CD_i = \frac{1}{k+1} r_i^2 \frac{h_{ii}}{1-h_{ii}},$$

where r_i is standardized residual when ith observation is excluded and h_{ii} is diagonal of $X_i(X'X)^{-1}X_i'$

Standard Error of Prediction If x_0 is vector associated with independent variable values and y_0 is value of dependent variable, then the standard error of prediction is given by

$$\sqrt{var(\hat{y}_0)} = \sqrt{x_0'(X'X)^{-1}x_0 s^2},$$

where s^2 is error variance associated with all observations.

Standard Error of Residuals is

$$\sqrt{var(\hat{y} - x_0'\beta)} = \sqrt{s^2[1 + x_0'(X'X)^{-1}x_0]}.$$

Rstudent Residuals are normalized residuals with ith observation excluded and it is computed as

$$RSTUDENT = \frac{r_i}{s_i\sqrt{1 - h_{ii}}},$$

where r_i is normalized residual, s_i is standard error when ith observation is excluded from analysis and h_{ii} is diagonal of $X_i(X'X)^{-1}X_i'$. Observations with RSTUDENT larger than 2 in absolute value may be considered extreme observation.

COVRATIO is ratio of determinants of covariances when the ith observation is deleted (denoted by $s^2_{(-i)}(X_{(i)}'X_{(i)})^{-1}$ to covariance using all the data, $s^2(X'X)^{-1}$. That is,

$$COVRATIO = \frac{\det\left[s^2_{(-i)}(X_{(i)}'X_{(i)})^{-1}\right]}{\det[s^2(X'X)^{-1}]}.$$

HAT matrix H is

$$H = X(X'X)^{-1}X'$$

or covariation within an observation to the average covariation. The diagonal entries of this matrix (h_{ii}) often are used for detecting influential observations.

DFFITS measures change in fit when ith observation is deleted, or $DFFITS = x_i[\beta - \beta_{(-1)}]$.

DFBETA is change in estimated coefficients when ith observations is deleted. $DFBETA_i = \beta - \beta_{(-1)}$.

VIF If R_i^2 is the multiple correlation coefficient of X_i regressed on the remaining explanatory variables, $VIF_i = \frac{1}{1 - R_i^2}$.

Condition Index If $\lambda_{max}, \lambda_2 \cdots \lambda_k$ denotes eigenvalues associated with matrix $(X'X)$, then

$$\text{Condition Index} = \sqrt{\frac{\lambda_{max}}{\lambda_i}}.$$

Proportions of variance of the kth regression coefficient shared with jth components. If eigenvectors are represented by v_{kj} and jth eigenvalue as λ_j, then shared variance kth variable is given by

$$var(\beta_k) = s^2 \sum_{j=1} \frac{v_{kj}}{\lambda_j}.$$