Regession 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.

- Descriptive or explanatory: interest may be on describing “What factors influence variability in dependent variable?” For example, factor contributing to higher sales among company’s sales force.
- Predictive, for example setting normal quota or baseline sales. We can also use estimated equation to determine “normal” and “abnormal” or outlier observations.
- Comparing Alternative theoretical explanations,
  - Consumers use reference price in comparing alternatives,
  - Consumers use specific price points in comparing alternatives.
- Decision purpose,
  - 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

- Measurement on two or more variables one of which must be dependent.
- Dependent variable must have interval or ratio scale measurement.
- If independent variables are nominal scaled (e.g. brand choice), then appropriate caution must 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 assessment 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 suggestions for variable(s) to be included in regression analysis as independent variables.

- Based on theory.
  - Economic, sales is a function of price,
  - Psychological, behavioural intention and attitude toward a product,
  - Biological, fertilizer usage, generally increase plant growth.

- Prior research,
  - Replicate findings for earlier efforts.
  - Extend results for alternative product category.
  - Bring new insights to earlier efforts.

- Educated “Guesses”, good idea or common sense.

- Statistical approaches.
  - **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.
Regression Analysis

- All variables contained in dataset.

Estimating Parameters

- Method of least squares, or
- Method of maximum likelihood, or
- Weighted least squares, or
- Method of least absolute deviations.

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

\[
\text{Value of Dependent variable} = \text{Constant} + \text{Slope} \times \text{Value of Indep. variable} + \text{Error}
\]

\[
y = a + b \times x + E
\]

- Constant (a), Slope (b) and Error (E) are unknown.
- You observe \(N\) pair of values of dependent and independent variables.
- Regression analysis provides reasonable (statistically unbiased) values for slope(s) and intercept.

An Illustrative Example - Two observations only.
Suppose we have two observations \((x_1, y_1)\) and \((x_2, y_2)\) or \((5,10)\) and \((20,20)\). These observations graphically can be shown as follows.

\[
\text{Slope} = \frac{y_2 - y_1}{x_2 - x_1}
\]

\[
= \frac{20 - 10}{20 - 5}
\]

\[
= \frac{10}{15}
\]

\[
= 0.66
\]

Slope is positive because

\[
y_2 \geq y_1
\]

and\n
\[
x_2 \geq x_1
\]

Intercept is \(y_1 - b \times x_1 = 6.67\)
The resulting equation would be \( y = 6.67 + 0.66 \times x \).

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

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 \( y = 23.33 - 0.66 \times x \).

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 format.

\[ R^2 = \frac{\text{Explained variation}}{\text{Total variation}} \]

\[ y = 1 + 0.7 \times x \]

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.
\[
y_i = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 3 \end{pmatrix} \quad x_i = \begin{pmatrix} -2 \\ -1 \\ 0 \\ 1 \\ 2 \end{pmatrix}
\]

Our regression equation can be written as
\[y_i = a + b \times x_i + E_i \quad i = 1, \ldots, 5.
\]

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 $(N-1)$ 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

\[
y_i - \bar{y} = \begin{pmatrix} 0 & -1 \\ 0 & -1 \\ 1 & -1 \\ 1 & -1 \\ 3 & -1 \end{pmatrix}, \quad x_i - \bar{x} = \begin{pmatrix} -2 & -0 \\ -1 & -0 \\ 0 & 0 \\ 1 & 0 \\ 2 & 0 \end{pmatrix}.
\]

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

\[
(y_i - \bar{y})(x_i - \bar{x}) = \begin{pmatrix} 2 \\ 1 \\ 0 \\ 0 \end{pmatrix} \quad \text{and} \quad (x_i - \bar{x})^2 = \begin{pmatrix} 4 \\ 1 \\ 0 \\ 1 \end{pmatrix}
\]

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 \((x_{1i}, x_{2i})\).
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 $a$, $b_1$ and $b_2$. 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,

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

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

$$S_{yx1} = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(x_{1i} - \bar{x}_1)}{N - 1}$$

$$S_{yx2} = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(x_{2i} - \bar{x}_2)}{N - 1}$$

$$S_{x1x1} = S_{x1x2} = \frac{\sum_{i=1}^{N}(x_{1i} - \bar{x}_1)(x_{1i} - \bar{x}_1)}{N - 1}$$

$$S_{x2x1} = S_{x2x2} = \frac{\sum_{i=1}^{N}(x_{2i} - \bar{x}_2)(x_{2i} - \bar{x}_2)}{N - 1}.$$ 

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**

- On an average difference between the observed value \((y_i)\) and the predicted value \((\hat{y}_i)\) is zero.
- On an average the estimated values of errors and values of independent variables are not related to each other.
- The squared differences between the observed value and the predicted value are similar.
- There is some variation in independent variable. If there are more than one variable in the equation, then two variables should not be perfectly correlated.

We could also make following observations about slope and intercept.

**Intercept or Constant**

- Intercept is the point at which the regression intercepts y-axis.
- Intercept provides a measure about the mean of dependent variable when slope(s) are zero.
- If slope(s) are not zero then intercept is equal to the mean of dependent variable minus slope× mean of independent variable.
Slope

- Change is dependent variable as we change independent variable.
- Zero slope means that independent variable does not have any influence on dependent 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 variables. This should be both from conceptual point of view as well as statistical perspective. Furthermore, 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 \((\alpha)\) level associated with Type I error. Although the most scientific research traditions use probability level of 0.05, you might be risk-taker and willing to use something else like 0.25.

3. Compute test statistic\(^2\).

\(^1\)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 \((\chi^2)\) distributed. The F-statistic then is a ratio of two \(\chi^2\) 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 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**

<table>
<thead>
<tr>
<th>Specific Aspect</th>
<th>Descriptive</th>
<th>Decision Oriented</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goodness-of-fit</td>
<td>$R^2$ or adjusted $R^2$, indicates percent variation in dependent variable explained by a set of independent variables.</td>
<td><strong>F-statistic</strong>, larger number means reject the null hypothesis that all parameters are zero</td>
</tr>
<tr>
<td>Individual parameters</td>
<td>Sign, Magnitude and elasticity</td>
<td><strong>t-statistic</strong> indicates whether specific parameter is different from zero. In comparing, t-statistic for two parameters, a larger t-statistic indicates that the independent variable is more important than other.</td>
</tr>
</tbody>
</table>

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

```sas
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;
```

---

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

---

**Regression Analysis**

---

Multivariate Research Methods

Course: COST*6060
Regression Analysis

SAS output produced following:

Dependent Variable: Y

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Prob&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>4.90000</td>
<td>4.90000</td>
<td>13.364</td>
<td>0.0354</td>
</tr>
<tr>
<td>Error</td>
<td>3</td>
<td>1.10000</td>
<td>0.36667</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C Total</td>
<td>4</td>
<td>6.00000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 0.60553  
R-square 0.8167
Dep Mean 1.00000  
Adj R-sq 0.7556

Parameter Estimates

| Variable | Parameter  | Standard          | T for H0: Parameter=0 | Prob > |T|
|----------|------------|-------------------|-----------------------|--------|
| INTERCEP | 1.000000   | 0.27080128        | 3.693                 | 0.0345 |
| X        | 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, $R^2$ 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 $(y_i - \hat{y}_i)$ 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 Tests

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Descriptive</th>
<th>Decision Oriented</th>
</tr>
</thead>
</table>
| Similar variation                 | Visual inspection or plot observation number and particular measure | **Student residuals**, normalized residual. Check for observations with the absolute value of normalized residuals \(\geq 2\).**  
|                                   |                                                   | **Rstudent**, value of residual when a particular observation is deleted. Check observation with the absolute value of Rstudent \(\geq 2\).**  
|                                   |                                                   | **Cook’s D**, same as above and check observation with Cook’s D \(\geq 8/[N - 2(k + 1)]\).** |
| Equal Weight or influence         | Visual inspection or plot observation number and particular measure | **COVRATIO**, ratio of covariation among independent variables based on particular observation excluded to one based on total sample. If the absolute value of COVARATIO – 1 is \(\geq 3(k + 1)/[N - k - 1]\), then examine particular observation.**  
|                                   |                                                   | **DFFITS** indicate change in parameter estimates taken all together when a particular observation is excluded. The absolute value of DFFITS \(\geq 2\sqrt{(k - 1)/N}\) considered extreme observation.**  
|                                   |                                                   | **DFBETAS** indicate change individual parameter estimate, when particular observation is excluded. The absolute value of DFBETAS \(\geq 2\sqrt{N}\) should be considered extreme observation.** |
| Independent variables uncorrelated or collinearity | Visual inspection of correlations and proportion of variance shared across variables. | **Variance inflation factor** (VIF) greater than 10 is considered a case of multicollinearity.**  
|                                   |                                                   | **Condition Index**, more than 15 to 20 is considered a case for multicollinearity.** |
| Successive error terms related or autocorrelation | Visual inspection or plot observation number and residuals. | **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-Q or probability plot, for a normally distributed variable, plot would be straight line passing through origin. | Tests of skewness, kurtosis and / or other test procedure to detect departure from normality.** |

Multivariate Research Methods                      Course: COST*6060
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

\[ y_i = a + b \times x_i + E_i \quad i = 1, \ldots, 5. \]

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 \( \hat{E}_i = y_i - (\hat{a} + \hat{bx}_i) \) and sum of squared residuals is \( \sum_{i=1}^{n} \hat{E}_i^2 \) and the standard deviation, often denoted by \( s \) is

\[ s = \sqrt{\frac{\sum_{i=1}^{n} \hat{E}_i^2}{n - 2}}. \]

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

\[
\begin{align*}
\mathcal{E}(\hat{a}) &= a \\
\mathcal{E}(\hat{b}) &= b \\
\text{var}(\hat{a}) &= \frac{s^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2} \\
\text{var}(\hat{b}) &= \frac{s^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\
\text{cov}(\hat{a}, \hat{b}) &= -\frac{s^2 \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}
\end{align*}
\]

where \( \bar{x} \) is the average of \( x_i, \quad i = 1, \ldots, 5. \)

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

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

It can be shown that

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

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

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

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

<table>
<thead>
<tr>
<th>Obs</th>
<th>Dep Var</th>
<th>Predict Value</th>
<th>Std Err Predict</th>
<th>Residual</th>
<th>Std Err Residual</th>
<th>Student Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>-0.4000</td>
<td>0.469</td>
<td>0.4000</td>
<td>0.383</td>
<td>1.044</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0.3000</td>
<td>0.332</td>
<td>-0.3000</td>
<td>0.507</td>
<td>-0.592</td>
</tr>
<tr>
<td>3</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.271</td>
<td>0</td>
<td>0.542</td>
<td>0.000</td>
</tr>
<tr>
<td>4</td>
<td>1.0000</td>
<td>1.7000</td>
<td>0.332</td>
<td>-0.7000</td>
<td>0.507</td>
<td>-1.382</td>
</tr>
<tr>
<td>5</td>
<td>3.0000</td>
<td>2.4000</td>
<td>0.469</td>
<td>0.6000</td>
<td>0.383</td>
<td>1.567</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Obs</th>
<th>Cook’s D</th>
<th>Hat Diag Rstudent</th>
<th>H Ratio</th>
<th>Dffits</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>**</td>
<td>0.818 1.0690 0.6000 2.2779 1.3093</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>*</td>
<td></td>
<td>0.075 -0.5145 0.3000 2.5068 -0.3368</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>0.000 0.0000 0.2000 2.8125 0.0000</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>**</td>
<td></td>
<td>0.409 -1.8708 0.3000 0.4250 -1.2247</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>***</td>
<td>1.841 3.0000 0.6000 0.1860 3.6742</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Obs</th>
<th>INTERCEP D</th>
<th>X</th>
<th>Dfbetas</th>
<th>Dfbetas</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.7559</td>
<td>-1.0690</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-0.2750</td>
<td>0.1945</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-1.0000</td>
<td>-0.7071</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2.1213</td>
<td>3.0000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

Note that in this example,

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

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

$$\text{var}(\hat{y}_1) = \frac{1.1}{3} \left[ \frac{1}{5} + \frac{4}{10} \right] = \frac{1.1}{5}$$
and square root of 0.22 results in the standard error of prediction of 0.469 for this observation. Similarly,

\[
\text{var}(\hat{E}_1) = s^2 \left[ 1 - \frac{1}{n} - \frac{(x_1 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right],
\]
\[
= \frac{1.1}{3} \left[ 1 - \frac{1}{5} - \frac{4}{10} \right]
\]
\[
= 0.1467,
\]

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 \(\hat{b}_{(1)}\) is equal to 0.8 and 0.9 respectively. Similarly, RSTUDENT is normalized residual when \(i\)th observation is excluded from analysis. For the first observation,

\[
\text{RSTUDENT}_{(1)} = \frac{E_1}{s_{(1)}\sqrt{1 - h_{11}}},
\]

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

\[
s_{(1)}^2 = \frac{1}{n - p - 1} \left[ (n - p)s^2 - \frac{E_1^2}{1 - h_{11}} \right]
\]
\[
= \frac{1}{4 - 1 - 1} \left[ (4 - 1) \frac{1.1}{5 - 1} - \frac{0.4 \times 0.4}{1 - 0.6} \right]
\]
\[
= 0.5 \times (1.1 - 0.4) = 0.35.
\]

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

\[
\text{RSTUDENT}_{(1)} = \frac{0.4}{0.5916\sqrt{0.4}} = 1.069,
\]

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:

```sas
options nocenter nodate ps=80 ls=80;
data pc;
input price cpu ram harddisk monitor retail $ cpu_type $;
cards;
  828.00 1000 128 20 17 Selltek 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 17 FutureShop 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 17 FutureShop 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
  1594.99 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
```

Multivariate Research Methods

Course: COST*6060
SAS output is as follows:

Model: MODEL1
Dependent Variable: PRICE

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Prob&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4</td>
<td>5896914.7629</td>
<td>1474228.6907</td>
<td>22.511</td>
<td>0.0001</td>
</tr>
<tr>
<td>Error</td>
<td>35</td>
<td>2292100.5421</td>
<td>65488.586918</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C Total</td>
<td>39</td>
<td>8189015.305</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 255.90738  R-square 0.7201  Adj R-sq 0.6881

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | T for H0: 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.
The null hypothesis states that variation in price can not be explained by CPU speed, 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.

We are explaining about 72% of variation in price by these four variables.

Regression equation can be written as

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

Note that the parameter associated with variables CPU and RAM have correct signs\(^4\) and statistically significant (probability of t-statistic is less than 0.05).

The parameters associated with variables HARDDISK and MONITOR have correct sign but statistically not significant. That means, these parameters could be equal to zero.

Consider a desktop with 1 Ghz, with 256 Megabytes of RAM, about 40 gigabytes hard drive and 17 inches MONITOR. For such machine, I should be expected to pay about $1,218. This is concluded as follows:

\[
\text{Price} = -526.65 + 0.833 \times 1000 + 1.525 \times 256 + 2.781 \times 40 + 24.098 \times 17.
\]

\[
= -526.65 + 833 + 390.4 + 111.24 + 409.67
\]

\[
= 1217.66
\]

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.

```sas
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;
```

<table>
<thead>
<tr>
<th>Obs</th>
<th>RETAIL</th>
<th>Dep Var Predict</th>
<th>Std Err Predict</th>
<th>Std Err Residual</th>
<th>Student Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Selltek 828.0</td>
<td>966.9</td>
<td>75.171</td>
<td>-138.9</td>
<td>244.618</td>
</tr>
<tr>
<td>2</td>
<td>Pctek 949.0</td>
<td>1300.1</td>
<td>84.236</td>
<td>-351.1</td>
<td>241.646</td>
</tr>
<tr>
<td>3</td>
<td>Datamat 970.0</td>
<td>1217.7</td>
<td>74.704</td>
<td>-247.7</td>
<td>244.761</td>
</tr>
</tbody>
</table>

\(^4\)I would think that desktops with faster CPUs should be more expensive than slower CPUs.
<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Selltek</td>
<td>978.0</td>
<td>995.4</td>
<td>117.797</td>
<td>-17.4251</td>
</tr>
<tr>
<td>5</td>
<td>FutureS</td>
<td>1010.0</td>
<td>994.8</td>
<td>125.129</td>
<td>15.1867</td>
</tr>
<tr>
<td>6</td>
<td>Selltek</td>
<td>1068.0</td>
<td>1162.0</td>
<td>92.814</td>
<td>-94.0298</td>
</tr>
<tr>
<td>7</td>
<td>Selltek</td>
<td>1128.0</td>
<td>1411.9</td>
<td>71.552</td>
<td>-283.90</td>
</tr>
<tr>
<td>8</td>
<td>TCC</td>
<td>1150.0</td>
<td>1495.2</td>
<td>71.607</td>
<td>-345.20</td>
</tr>
<tr>
<td>9</td>
<td>TCC</td>
<td>1170.0</td>
<td>1384.3</td>
<td>49.600</td>
<td>-214.30</td>
</tr>
<tr>
<td>10</td>
<td>Gateway</td>
<td>1176.5</td>
<td>1002.0</td>
<td>142.267</td>
<td>174.60</td>
</tr>
<tr>
<td>11</td>
<td>Business</td>
<td>1199.0</td>
<td>1105.8</td>
<td>77.940</td>
<td>93.218</td>
</tr>
<tr>
<td>12</td>
<td>FutureS</td>
<td>1230.0</td>
<td>1245.3</td>
<td>82.844</td>
<td>-15.342</td>
</tr>
<tr>
<td>13</td>
<td>Gateway</td>
<td>1238.5</td>
<td>918.7</td>
<td>138.660</td>
<td>319.90</td>
</tr>
<tr>
<td>14</td>
<td>RadioSh</td>
<td>1249.0</td>
<td>1301.0</td>
<td>61.051</td>
<td>-51.96</td>
</tr>
<tr>
<td>15</td>
<td>Datamat</td>
<td>1250.0</td>
<td>1634.2</td>
<td>46.662</td>
<td>-384.20</td>
</tr>
<tr>
<td>16</td>
<td>FutureS</td>
<td>1250.0</td>
<td>1175.7</td>
<td>101.098</td>
<td>74.295</td>
</tr>
<tr>
<td>17</td>
<td>FutureS</td>
<td>1250.0</td>
<td>1384.3</td>
<td>49.600</td>
<td>-134.30</td>
</tr>
<tr>
<td>18</td>
<td>RadioSh</td>
<td>1270.0</td>
<td>1717.5</td>
<td>57.060</td>
<td>-447.5</td>
</tr>
<tr>
<td>19</td>
<td>Dell</td>
<td>1399.0</td>
<td>1050.2</td>
<td>71.640</td>
<td>348.80</td>
</tr>
<tr>
<td>20</td>
<td>Selltek</td>
<td>1478.0</td>
<td>1717.5</td>
<td>57.060</td>
<td>-239.5</td>
</tr>
<tr>
<td>21</td>
<td>Dell</td>
<td>1549.0</td>
<td>1661.8</td>
<td>83.082</td>
<td>-112.8</td>
</tr>
<tr>
<td>22</td>
<td>FutureS</td>
<td>1550.0</td>
<td>1439.9</td>
<td>76.358</td>
<td>110.10</td>
</tr>
<tr>
<td>23</td>
<td>Busines</td>
<td>1749.0</td>
<td>1217.7</td>
<td>134.871</td>
<td>-153.4</td>
</tr>
<tr>
<td>24</td>
<td>TCC</td>
<td>1850.0</td>
<td>1904.6</td>
<td>144.629</td>
<td>-54.606</td>
</tr>
<tr>
<td>25</td>
<td>RadioSh</td>
<td>1899.0</td>
<td>1634.2</td>
<td>46.662</td>
<td>264.80</td>
</tr>
<tr>
<td>26</td>
<td>FutureS</td>
<td>1950.0</td>
<td>1856.4</td>
<td>88.206</td>
<td>93.571</td>
</tr>
<tr>
<td>27</td>
<td>Gateway</td>
<td>2010.5</td>
<td>1578.5</td>
<td>75.643</td>
<td>431.90</td>
</tr>
<tr>
<td>28</td>
<td>FutureS</td>
<td>2020.0</td>
<td>1800.8</td>
<td>70.148</td>
<td>549.10</td>
</tr>
<tr>
<td>29</td>
<td>Busines</td>
<td>2149.0</td>
<td>2302.4</td>
<td>134.871</td>
<td>-153.4</td>
</tr>
<tr>
<td>30</td>
<td>Dell</td>
<td>2509.0</td>
<td>1959.9</td>
<td>160.628</td>
<td>549.1</td>
</tr>
<tr>
<td>31</td>
<td>FutureS</td>
<td>2650.0</td>
<td>2552.3</td>
<td>136.722</td>
<td>97.702</td>
</tr>
</tbody>
</table>

- Column 1 is values of dependent variable (\(y_i\)). This variable is sorted in ascending order to help us interpret other statistical measures.
- Column 2 is predicted values for dependent variable (\(\hat{y}_i\)). For the first observation, \(\hat{y}_1 = -526.65 + 0.833 \times 1000 + 1.525 \times 128 + 2.781 \times 20 + 24.098 \times 17 = 966.9\).
- Column 3 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, \(x_1 = [110001282017]\). Then \(\text{Var}(y_1) = s^2 x_1' (X'X)^{-1} x_1\).
- Column 4 is residual or error values, (\(y_i - \hat{y}_i\)).
- Column 5 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.
- Column 6 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.

Multivariate Research Methods Course: COST*6060
<table>
<thead>
<tr>
<th>Obs</th>
<th>RETAIL</th>
<th>-2-1-0</th>
<th>1</th>
<th>2</th>
<th>Cook’s D</th>
<th>Rstudent</th>
<th>Hat Diag</th>
<th>Cov</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Selltek</td>
<td>*</td>
<td>0.006</td>
<td>-0.5621</td>
<td>0.0863</td>
<td>1.2080</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Pctek</td>
<td>**</td>
<td>0.051</td>
<td>-1.4771</td>
<td>0.1083</td>
<td>0.9499</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Datamat</td>
<td>**</td>
<td>0.019</td>
<td>-1.0123</td>
<td>0.0852</td>
<td>1.0893</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Selltek</td>
<td></td>
<td>0.000</td>
<td>-0.0756</td>
<td>0.2119</td>
<td>1.4655</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>FutureS</td>
<td></td>
<td>0.000</td>
<td>0.0671</td>
<td>0.2391</td>
<td>1.5182</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Selltek</td>
<td></td>
<td>0.005</td>
<td>-0.3895</td>
<td>0.1315</td>
<td>1.3018</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Selltek</td>
<td>**</td>
<td>0.023</td>
<td>-1.1614</td>
<td>0.0782</td>
<td>1.0323</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>TCC</td>
<td>**</td>
<td>0.034</td>
<td>-1.4258</td>
<td>0.0783</td>
<td>0.9381</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>TCC</td>
<td>*</td>
<td>0.006</td>
<td>-0.8501</td>
<td>0.0376</td>
<td>1.0812</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Gateway</td>
<td></td>
<td>0.060</td>
<td>0.8168</td>
<td>0.3091</td>
<td>1.5181</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Busines</td>
<td></td>
<td>0.003</td>
<td>0.3777</td>
<td>0.0928</td>
<td>1.2478</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>FutureS</td>
<td></td>
<td>0.000</td>
<td>-0.0625</td>
<td>0.1048</td>
<td>1.2906</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Gateway</td>
<td></td>
<td>0.184</td>
<td>1.5144</td>
<td>0.2936</td>
<td>1.1807</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>RadioSh</td>
<td></td>
<td>0.001</td>
<td>-0.2062</td>
<td>0.0569</td>
<td>1.2181</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Datamat</td>
<td>***</td>
<td>0.016</td>
<td>-1.6577</td>
<td>0.0332</td>
<td>0.8471</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>FutureS</td>
<td></td>
<td>0.004</td>
<td>0.3119</td>
<td>0.1561</td>
<td>1.3503</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>FutureS</td>
<td>**</td>
<td>0.002</td>
<td>-0.5293</td>
<td>0.0376</td>
<td>1.1528</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Datamat</td>
<td>***</td>
<td>0.034</td>
<td>-1.8553</td>
<td>0.0497</td>
<td>0.7511</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>FutureS</td>
<td></td>
<td>0.000</td>
<td>0.1121</td>
<td>0.1007</td>
<td>1.2830</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>FutureS</td>
<td></td>
<td>0.000</td>
<td>-0.2132</td>
<td>0.0376</td>
<td>1.1931</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>RadioSh</td>
<td></td>
<td>0.000</td>
<td>-0.1385</td>
<td>0.0376</td>
<td>1.1977</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Selltek</td>
<td></td>
<td>0.000</td>
<td>-0.0246</td>
<td>0.0376</td>
<td>1.2010</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Dell</td>
<td>***</td>
<td>0.034</td>
<td>1.4417</td>
<td>0.0784</td>
<td>0.9323</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Selltek</td>
<td>*</td>
<td>0.010</td>
<td>-0.9588</td>
<td>0.0497</td>
<td>1.0645</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Dell</td>
<td></td>
<td>0.005</td>
<td>-0.4609</td>
<td>0.1054</td>
<td>1.2525</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>FutureS</td>
<td></td>
<td>0.004</td>
<td>0.4456</td>
<td>0.0890</td>
<td>1.2325</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Selltek</td>
<td>**</td>
<td>0.028</td>
<td>-1.0624</td>
<td>0.1095</td>
<td>1.1026</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>FutureS</td>
<td></td>
<td>0.000</td>
<td>-0.1599</td>
<td>0.0800</td>
<td>1.2518</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>RadioSh</td>
<td></td>
<td>0.001</td>
<td>0.2379</td>
<td>0.0800</td>
<td>1.2461</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>Pctek</td>
<td></td>
<td>0.001</td>
<td>-0.2075</td>
<td>0.0751</td>
<td>1.2420</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>Busines</td>
<td>*****</td>
<td>0.088</td>
<td>2.3001</td>
<td>0.0852</td>
<td>0.6132</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>TCC</td>
<td></td>
<td>0.006</td>
<td>-0.2552</td>
<td>0.3194</td>
<td>1.6823</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>RadioSh</td>
<td>**</td>
<td>0.008</td>
<td>1.0542</td>
<td>0.0332</td>
<td>1.0181</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>FutureS</td>
<td></td>
<td>0.004</td>
<td>0.3847</td>
<td>0.1188</td>
<td>1.2836</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>Gateway</td>
<td>*****</td>
<td>0.060</td>
<td>1.8247</td>
<td>0.0874</td>
<td>0.7939</td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>FutureS</td>
<td>*</td>
<td>0.013</td>
<td>0.8881</td>
<td>0.0751</td>
<td>1.1145</td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>Busines</td>
<td></td>
<td>0.038</td>
<td>-0.7001</td>
<td>0.2778</td>
<td>1.4900</td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>Dell</td>
<td>*****</td>
<td>0.988</td>
<td>3.0699</td>
<td>0.3940</td>
<td>0.5613</td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>Busines</td>
<td></td>
<td>0.141</td>
<td>1.4654</td>
<td>0.2629</td>
<td>1.1392</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>FutureS</td>
<td>**</td>
<td>0.016</td>
<td>0.4465</td>
<td>0.2854</td>
<td>1.5711</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- column 7 is a plot of normalized residuals and these numbers generally vary between −2 and 2.
- column 8 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 ≥ \( \frac{8}{N-2(k+1)} \) is 0.267 would be considered influential observation (see observation number 38).
- column 9 is similar to Cook’s D with the exception that error variances are estimated without the ith observation.
- column 10 Hat Diag H (Diagonal of Hat matrix H, also sometimes denoted as \( h_{ii} \)) 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
Regression Analysis

$k/N$ and observation with $h_{ij} \geq 2k/N$ (2 x 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.

The column Cov ratio (Covariance ratio) is a ratio covariances when $i$th 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 $(\text{COVRATIO} - 1) \geq \frac{3(k+1)}{N-k-1}$ is considered significant influential observation. For our case, $\text{COVRATIO} \geq 1 + \frac{(3 \times 5)}{30} = 1.429$ would be observations with higher than the normal influence.

<table>
<thead>
<tr>
<th>Obs</th>
<th>RETAIL</th>
<th>INTERC</th>
<th>CPU</th>
<th>RAM</th>
<th>HARDDISK</th>
<th>MONITOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Selltek</td>
<td>-0.1727</td>
<td>0.0247</td>
<td>0.0545</td>
<td>0.0464</td>
<td>0.0452</td>
</tr>
<tr>
<td>2</td>
<td>Pctek</td>
<td>-0.5149</td>
<td>-0.0181</td>
<td>-0.2738</td>
<td>0.3160</td>
<td>0.1419</td>
</tr>
<tr>
<td>3</td>
<td>Datamat</td>
<td>-0.3090</td>
<td>0.0441</td>
<td>0.2692</td>
<td>-0.1324</td>
<td>-0.0997</td>
</tr>
<tr>
<td>4</td>
<td>Selltek</td>
<td>-0.0392</td>
<td>0.0064</td>
<td>0.0313</td>
<td>-0.0246</td>
<td>0.0178</td>
</tr>
<tr>
<td>5</td>
<td>FutureS</td>
<td>0.0376</td>
<td>-0.0033</td>
<td>-0.0141</td>
<td>-0.0194</td>
<td>0.0289</td>
</tr>
<tr>
<td>6</td>
<td>Selltek</td>
<td>-0.1516</td>
<td>0.0202</td>
<td>0.0975</td>
<td>-0.0948</td>
<td>0.0900</td>
</tr>
<tr>
<td>7</td>
<td>Selltek</td>
<td>-0.3382</td>
<td>0.0097</td>
<td>0.0394</td>
<td>-0.1628</td>
<td>0.2732</td>
</tr>
<tr>
<td>8</td>
<td>TCC</td>
<td>-0.4156</td>
<td>-0.0079</td>
<td>-0.0510</td>
<td>-0.1541</td>
<td>0.3406</td>
</tr>
<tr>
<td>9</td>
<td>TCC</td>
<td>-0.1679</td>
<td>0.0129</td>
<td>0.0963</td>
<td>-0.0550</td>
<td>-0.0019</td>
</tr>
<tr>
<td>10</td>
<td>Gateway</td>
<td>0.5463</td>
<td>0.4983</td>
<td>0.1465</td>
<td>-0.1210</td>
<td>-0.0494</td>
</tr>
<tr>
<td>11</td>
<td>Business</td>
<td>0.1208</td>
<td>-0.0088</td>
<td>-0.0148</td>
<td>-0.0839</td>
<td>0.0572</td>
</tr>
<tr>
<td>12</td>
<td>FutureS</td>
<td>-0.0214</td>
<td>0.0023</td>
<td>0.0110</td>
<td>-0.0129</td>
<td>0.0144</td>
</tr>
<tr>
<td>13</td>
<td>Gateway</td>
<td>0.9763</td>
<td>0.8897</td>
<td>0.1480</td>
<td>-0.1664</td>
<td>-0.0844</td>
</tr>
<tr>
<td>14</td>
<td>RadioSh</td>
<td>-0.0507</td>
<td>0.0060</td>
<td>0.0378</td>
<td>-0.0200</td>
<td>-0.0012</td>
</tr>
<tr>
<td>15</td>
<td>Datamat</td>
<td>-0.2889</td>
<td>-0.0400</td>
<td>-0.1420</td>
<td>0.0457</td>
<td>0.0129</td>
</tr>
<tr>
<td>16</td>
<td>FutureS</td>
<td>0.1341</td>
<td>-0.0109</td>
<td>-0.0632</td>
<td>-0.0440</td>
<td>0.1029</td>
</tr>
<tr>
<td>17</td>
<td>FutureS</td>
<td>-0.1046</td>
<td>0.0081</td>
<td>0.0600</td>
<td>-0.0343</td>
<td>-0.0012</td>
</tr>
<tr>
<td>18</td>
<td>Datamat</td>
<td>-0.4244</td>
<td>-0.0735</td>
<td>-0.2979</td>
<td>0.1135</td>
<td>0.0222</td>
</tr>
<tr>
<td>19</td>
<td>FutureS</td>
<td>0.0375</td>
<td>0.0006</td>
<td>0.0114</td>
<td>-0.0323</td>
<td>0.0162</td>
</tr>
<tr>
<td>20</td>
<td>FutureS</td>
<td>-0.0421</td>
<td>0.0032</td>
<td>0.0242</td>
<td>-0.0138</td>
<td>-0.0005</td>
</tr>
<tr>
<td>21</td>
<td>RadioSh</td>
<td>-0.0274</td>
<td>0.0021</td>
<td>0.0157</td>
<td>-0.0090</td>
<td>-0.0003</td>
</tr>
<tr>
<td>22</td>
<td>Selltek</td>
<td>-0.0049</td>
<td>0.0004</td>
<td>0.0028</td>
<td>-0.0016</td>
<td>-0.0001</td>
</tr>
<tr>
<td>23</td>
<td>Dell</td>
<td>0.4204</td>
<td>-0.0429</td>
<td>-0.0386</td>
<td>-0.1647</td>
<td>-0.1206</td>
</tr>
<tr>
<td>24</td>
<td>Selltek</td>
<td>-0.2193</td>
<td>-0.0380</td>
<td>-0.1540</td>
<td>0.0586</td>
<td>0.0115</td>
</tr>
<tr>
<td>25</td>
<td>Dell</td>
<td>-0.1822</td>
<td>-0.0157</td>
<td>-0.0820</td>
<td>-0.0206</td>
<td>0.1151</td>
</tr>
<tr>
<td>26</td>
<td>FutureS</td>
<td>0.1393</td>
<td>-0.0039</td>
<td>-0.0573</td>
<td>-0.0179</td>
<td>0.1059</td>
</tr>
<tr>
<td>27</td>
<td>Selltek</td>
<td>-0.3726</td>
<td>-0.0736</td>
<td>-0.3269</td>
<td>0.1364</td>
<td>0.0209</td>
</tr>
<tr>
<td>28</td>
<td>FutureS</td>
<td>-0.0472</td>
<td>-0.0053</td>
<td>-0.0130</td>
<td>0.0218</td>
<td>-0.0361</td>
</tr>
<tr>
<td>29</td>
<td>RadioSh</td>
<td>0.0702</td>
<td>0.0079</td>
<td>0.0193</td>
<td>-0.0324</td>
<td>0.0537</td>
</tr>
<tr>
<td>30</td>
<td>Pctek</td>
<td>-0.0591</td>
<td>-0.0112</td>
<td>-0.0482</td>
<td>0.0195</td>
<td>0.0033</td>
</tr>
<tr>
<td>31</td>
<td>Business</td>
<td>0.7020</td>
<td>-0.1003</td>
<td>-0.5891</td>
<td>0.3008</td>
<td>0.0221</td>
</tr>
<tr>
<td>32</td>
<td>TCC</td>
<td>-0.1748</td>
<td>0.1476</td>
<td>-0.0115</td>
<td>0.0553</td>
<td>-0.0555</td>
</tr>
<tr>
<td>33</td>
<td>RadioSh</td>
<td>0.1955</td>
<td>0.0271</td>
<td>0.0961</td>
<td>-0.0309</td>
<td>-0.0088</td>
</tr>
<tr>
<td>34</td>
<td>FutureS</td>
<td>0.1413</td>
<td>0.0240</td>
<td>0.0868</td>
<td>-0.0788</td>
<td>0.0859</td>
</tr>
<tr>
<td>35</td>
<td>Gateway</td>
<td>0.5646</td>
<td>0.0358</td>
<td>0.1934</td>
<td>0.1394</td>
<td>-0.4446</td>
</tr>
<tr>
<td>36</td>
<td>FutureS</td>
<td>0.2531</td>
<td>0.0480</td>
<td>0.2063</td>
<td>-0.0835</td>
<td>-0.0140</td>
</tr>
<tr>
<td>37</td>
<td>Business</td>
<td>-0.4342</td>
<td>-0.0433</td>
<td>0.0669</td>
<td>-0.2616</td>
<td>-0.1292</td>
</tr>
<tr>
<td>38</td>
<td>Dell</td>
<td>2.4752</td>
<td>-1.8277</td>
<td>0.7657</td>
<td>-0.1447</td>
<td>-1.0910</td>
</tr>
<tr>
<td>39</td>
<td>Business</td>
<td>0.8526</td>
<td>0.1007</td>
<td>-0.0045</td>
<td>0.6588</td>
<td>-0.1207</td>
</tr>
<tr>
<td>40</td>
<td>FutureS</td>
<td>0.2822</td>
<td>0.0490</td>
<td>0.0631</td>
<td>0.1189</td>
<td>0.0774</td>
</tr>
</tbody>
</table>

The 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 $\pm 0.548$ would be an influential observations.


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 \( \pm 0.316 \). There will one DFBETA for each parameter estimated. In our case there are five such measures.

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>Variance Inflation</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERCEP</td>
<td>1</td>
<td>0.00000000</td>
</tr>
<tr>
<td>CPU</td>
<td>1</td>
<td>1.67434954</td>
</tr>
<tr>
<td>RAM</td>
<td>1</td>
<td>1.87908442</td>
</tr>
<tr>
<td>HARDDISK</td>
<td>1</td>
<td>1.46192377</td>
</tr>
<tr>
<td>MONITOR</td>
<td>1</td>
<td>1.18063429</td>
</tr>
</tbody>
</table>

Collinearity Diagnostics (intercept adjusted)

<table>
<thead>
<tr>
<th>Number</th>
<th>Eigenvalue</th>
<th>Condition Index</th>
<th>Var Prop CPU</th>
<th>Var Prop RAM</th>
<th>Var Prop HARDDISK</th>
<th>Var Prop MONITOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.18504</td>
<td>1.00000</td>
<td>0.0823</td>
<td>0.0783</td>
<td>0.0767</td>
<td>0.0513</td>
</tr>
<tr>
<td>2</td>
<td>0.89578</td>
<td>1.56181</td>
<td>0.0180</td>
<td>0.0542</td>
<td>0.1508</td>
<td>0.6471</td>
</tr>
<tr>
<td>3</td>
<td>0.57072</td>
<td>1.95667</td>
<td>0.3952</td>
<td>0.0459</td>
<td>0.5049</td>
<td>0.2253</td>
</tr>
<tr>
<td>4</td>
<td>0.34845</td>
<td>2.50413</td>
<td>0.5045</td>
<td>0.8216</td>
<td>0.2676</td>
<td>0.0762</td>
</tr>
</tbody>
</table>

Durbin-Watson D 1.634
(For Number of Obs.) 40
1st Order Autocorrelation 0.177

Sum of Residuals 0
Sum of Squared Residuals 2292100.5421
Predicted Resid SS (Press) 3350277.8100


[12] 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.

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

[14] column condition index is square root of the ratio of largest eigenvalue to a particular eigenvalue.

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

[16] , 20 are measures of whether successive error terms are correlated.

There is a graphical alternative to visualizing various diagnostics 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 \( \mu_1 \), then the \( r \)th moment about the mean is defined as

\[
\mu_r = \mathbb{E}(y - \mu_1)^r, \quad \text{for } r > 1,
\]

where \( \mathbb{E} \) is used to denote the expected value or the average. If we know mean (\( \mu_1 \)) and its variance (\( \mu_2 \)), 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 (\( \mu_2 \)) is the expected value of the squared difference of the values from the population mean:

\[
\mu_2 = \mathbb{E}(y - \mu_1)^2.
\]

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

\[ \mathcal{E}(y - \mu_1)^3 / \mu_2^{3/2} \]

The sample third moment \((g_1)\) is computed as

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

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

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

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,

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

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 \((b_2)\). The population kurtosis is defined as

\[ \mu_4 = \mathcal{E}(y - \mu_1)^4 / \mu_2^2 - 3. \]

The sample fourth moment is calculated as

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

To convert fourth moment to kurtosis \((b_2)\) 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, \(b_2\) is equal to 3. In large samples, hypothesis test for \(b_2\) can be performed by converting \(b_2\) 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 coefficient of skewness and kurtosis as indicated below.
and this estimate is approximately normally distributed under the null hypothesis of population normality. 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 (\(\chi^2\)) 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

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

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

---


\(^7\)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 Powerful and informative Tests of Normality”, Vol. 44, pp. 316–321. The macro for your usage is kept in file G:\courses\COST6060\NORMTEST.SAS.
Regression Analysis

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

K**2 = Chisquare (2 df) = 2.860 prob = 0.2393
Shapiro-Wilk Test = 0.966 Prob = 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 $K^2$ 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. Transform dependent variable, or independent variable or both.
  2. Exclude observations with more influence.
  3. Apply weighting scheme that are managerially or statistically meaningful.
  4. Estimate model with weighted least squares or the least absolute deviation.

• Presence of Collinearity
  1. Create new index variables that may capture correlations among independent variables either conceptually (for example SES, instead income, occupation and education etc.)
  2. Determine stability of parameters by excluding one or more variables.
  3. Use statistical procedures for dealing with this problem, for example, transformation, alternative criterion to minimize.

• Lack of Independence of successive error values
  1. May be caused by missing variables, competitive variables or customer loyalty, then include missing variables.
  2. Re-estimate model with autocorrelated errors.

• Error values not normally distributed
  1. Use non-normal distribution to estimate parameters.
  2. Use transformation convert dependent variable so that new variable is normally distributed.
  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. Use a large sample, 40 - 50 observations per variable included will have better stability to estimates than 5 - 10 observations.

3. Validate your model with hold-out or split-half sample or new sample.
Limitations of Regression Analysis

- Moderating effects of variables
  1. By group differences,
  2. Interaction effects,
  3. Effect occur only at certain level.

- 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.

\[ w \rightarrow x \rightarrow y \]

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

\[
\begin{align*}
y & = a + bx + e_y \\
x & = c + dw + e_x
\end{align*}
\]

- Not-linear effects.

- Effects associated with data collection.
  1. Measurement errors,
  2. Response effects,
  3. Truncation of variables.

Estimating Regression Model using Matrix Algebra

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

\[
y = X\beta + u
\]

where $y$ is $(y_1, y_2, \ldots, y_N)'$, $u$ is $(u_1, u_2, \ldots, u_N)'$, and $\beta$ is $(\beta_1, \beta_2, \ldots, \beta_k)'$ are vectors, and $X$ $(X_1, X_2, \ldots, X_N)'$ is matrix and $'$ 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 $\beta$, 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 $\beta$ so as to minimize the sum of squared residuals. Mathematically I may write

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

$$= y'y - 2y'X\beta + \beta'X'X\beta$$

(2)

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

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

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

(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 $E(uu') = \sigma^2I_N$ 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 violated if the observed dependent variable takes either 0 or 1 values. (As an exercise you may show this). Similarly, if successive values of dependent variable are related, as in case of time series data, then this assumption is also violated. 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 violate this assumption.

It is also possible to show (with lot of algebraic manipulation) that the estimated value of $\sigma^2$ is $(\hat{u}'\hat{u})/(N - k)$. Note also that second derivatives of $f(\beta)$ with respect to $\beta$ 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

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

(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 $\mathcal{L}$ and may be written as
\[ L = f(u_1) \times f(u_2) \times \cdots \times f(u_N) \]
\[ = \left( \frac{1}{\sqrt{2\pi \sigma^2}} \right)^N \exp\left( -\frac{\sum_{i=1}^{N} u_i^2}{2\sigma^2} \right) \]
\[ = (2\pi \sigma^2)^{-N/2} \exp\left( -\frac{\sum_{i=1}^{N} u_i^2}{2\sigma^2} \right) \]  
(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

\[
\log L = -\frac{N}{2} \log(2\pi) - \frac{N}{2} \log \sigma^2 - \frac{1}{2\sigma^2} u^T u
\]
\[= -\frac{N}{2} \log(2\pi) - \frac{N}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (y - X\beta)^T (y - X\beta)
\]  
(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 \(\beta\) and \(\sigma^2\) and set equal to zero. Thus, I may write

\[
\frac{\partial \log L}{\partial \beta} = -\frac{1}{2\sigma^2} (2X' y + 2X' X \beta) = 0
\]
\[
\frac{\partial \log L}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} (y - X\beta)^T (y - X\beta) = 0
\]

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

\[
\hat{\beta} = (X' X)^{-1} X' y \quad \text{and} \quad \hat{\sigma}^2 = \frac{(y - X\beta)^T (y - X\beta)}{N}
\]

Although the estimate of vector \(\beta\) using the least squares and maximum likelihood method is same, the estimate of \(\sigma^2\) is not equal. In fact, \(\sigma^2\) 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 \(\beta\) and \(\sigma^2\) are negative. This assures me that I have actually maximized the function.

Finally, it is possible to obtain \(\log L\) 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 \(\log L\). Thus, we may write

\[
\log L = -\frac{N}{2} \log(2\pi) - \frac{N}{2} \log \left( \frac{u'u}{N-k} \right) - \frac{N-k}{2} (u'u)
\]
\[= -\frac{N}{2} \log(2\pi) - \frac{N}{2} \log \left( \frac{u'u}{N-k} \right) - \frac{N-k}{2}
\]  
(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 $\beta$ 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 observations 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 statistic for a normally distributed random variable is 0.

F-statistics is used to test whether \( \beta \) vector is significantly different from zero and it is the ratio of mean sums of squares due to 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 \( \beta_i \) in above expression is zero.

Cook’s Distance (\( CD_i \)) is a measure of the change in the regression coefficients that would occur if a \( i \)th 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 \( i \)th observation deleted, and \( MS_{res} \) is the residual variance for all the observations. It is easier to compute Cook’s D by

\[
CD_i = \frac{1}{k + 1} r_i^2 h_{ii} / (1 - h_{ii}),
\]

where \( r_i \) is standardized residual when \( i \)th 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

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

where \( s^2 \) is error variance associated with all observations.
Standard Error of Residuals is
\[ \sqrt{\text{var}(\hat{y}_0 - x'_0\beta)} = \sqrt{s^2 \left[ 1 + x'_0(X'X)^{-1}x_0 \right]} \]

**Rstudent Residuals** are normalized residuals with ith observation excluded and it is computed as
\[ \text{RSTUDENT} = \frac{r_i}{s_i \sqrt{1 - h_{ii}}} \]
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 \( s^2_{(i)}(X'_iX_i)^{-1} \)) to covariance using all the data, \( s^2(X'X)^{-1} \). That is,
\[ \text{COVRATIO} = \frac{\det \left[ s^2_{(i)}(X'_iX_i)^{-1} \right]}{\det \left[ s^2(X'X)^{-1} \right]} \]

**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 = \( x_i[\beta - \beta_{(i-1)}] \).

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

**VIF** If \( R^2_i \) is the multiple correlation coefficient of \( X_i \) regressed on the remaining explanatory variables, \( \text{VIF}_i = \frac{1}{1 - R^2_i} \).

**Condition Index** If \( \lambda_1, \lambda_2, \ldots, \lambda_k \) denotes eigenvalues associated with matrix \( (X'X) \), then
\[ \text{Condition Index} = \sqrt{\frac{\lambda_{\text{max}}}{\lambda_i}} \]

**Proportions of variance** of the kth regression coefficient shared with jth components. If eigenvectors are represented by \( v_{kj} \) and jth eigenvalue as \( \lambda_j \), then shared variance kth variable is given by
\[ \text{var}(\beta_k) = s^2 \sum_{j=1}^{k} \frac{v_{kj}}{\lambda_j} \]