Regression Analysis

To call in the statistician after the experiment is done may be no more than asking him to perform a postmortem examination: he may be able to say what the experiment died of-R.A.Fisher

Section 4.0 Introduction to regression analysis

The moniker *regression analysis* derives from its application in measuring receding relationships between certain hereditary traits and their passage to succeeding generations. This was formally studied by Sir Francis Galton, who was a British Geneticist. As traits to antecedents become fainter when generation gap increases, the relation was coined regression analysis. Since its introduction, the ubiquity of regression analysis has been astounding. It finds application in a wide swathe of disciplines from commerce to medicine. In this approach, an approximate relationship, $y \cong f(x)$, is posited, where $f(x, \beta_0, \beta_1)$ may be linear in *x* and the coefficients β_0, β_1 , or polynomial function of *x*. By adding an error component, the approximation $y \cong f(x)$, becomes an equality, i.e., $y = f(x, \beta_0, \beta_1) + \varepsilon$. The goal is to find a $f\left(x, \overset{\wedge}{\beta_0}, \overset{\wedge}{\beta_1}\right)$ such that the error ε is minimized. Note that $\overset{\wedge}{\beta} 0, and \overset{\wedge}{\beta} 1$ are computed from the sample data. The estimated equation $y = f\left(x, \overset{\wedge}{\beta_0}, \overset{\wedge}{\beta_1}\right)$ is then utilized to predict the response *y* for a given value of *x*.

Section 4.1 Simple linear regression and the least squares method

The simple linear regression model is a method to model a single dependent variable *y* as a function of single independent variable *x*. Depending on the discipline in which it finds application, the dependent variable *y* is variously alluded to as response variable, output variable, or endogenous variable. The independent variable aliases include predictor variable, attribute variable, exogenous variable, or input variable among others.

Mathematically, the relationship between *y* and *x* may be expressed as;

Observed output = function + noise (4.1)

It is assumed the function "*f*" is linear in β and *x* i.e.,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1,2,\ldots\ldots,n \qquad (4.2)$$

Whatever cannot be explained by our chosen function "*f*" will be collapsed into the error term ε. The goal here is to find those values of β_0 and β_1 such that the sum-of-squares of error is minimized.

Notes:

The sum-of-squares of error is minimized by minimizing the objective function; $\phi = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$ (4.3)

Solving the system of linear equations, by setting the first derivatives of ϕ with respect to β_0 and β_1 zero, we have

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \qquad (4.4)$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (4.5)$$

Section 4.2 Regression model with normal errors

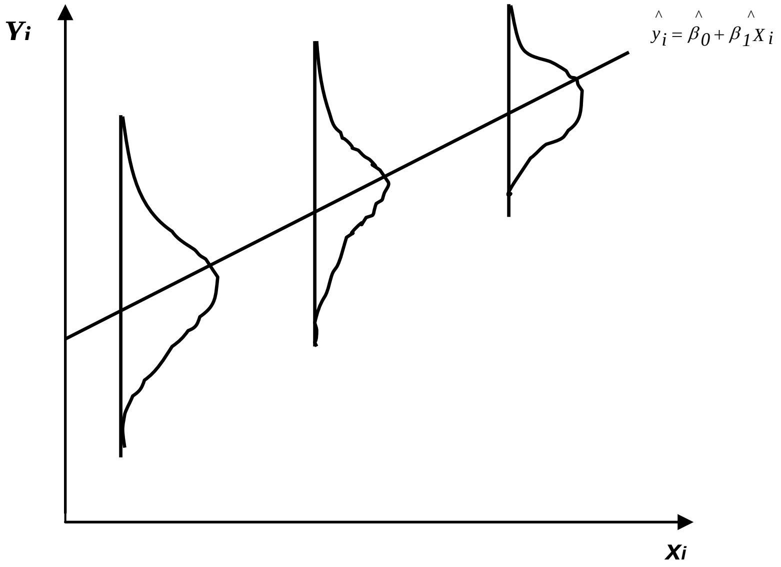


Figure 4.1 Linear regression model with normal errors

Example 4.1

Consider the following example, where *y* is the response variable, and *x* is the predictor variable. Table 4.1 contains 20 observations. We will fit a simple linear regression (SLR) model, given by $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1,2,....,n$

y	*x*	*y*	*x*
1.63	0.97	1.65	1.23
2.48	2.16	1.37	1.16
1.01	0.74	1.64	1.28
1.44	1.22	1.36	0.72
1.42	0.56	1.36	0.75
1.46	0.78	1.18	0.44
2.00	1.40	1.96	1.95
1.70	1.64	2.12	1.86
0.97	0.36	1.36	0.99
1.75	1.34	1.00	0.31
2.00	1.34	1.52	0.93
1.95	1.68	1.80	1.30
		1.00	0.39

Table 4.1 Data for regression model fitting, where y and x are the dependent and independent variables respectively.

The plot of *y* versus *x* and the overlaid regression function is given in Figure 4.2

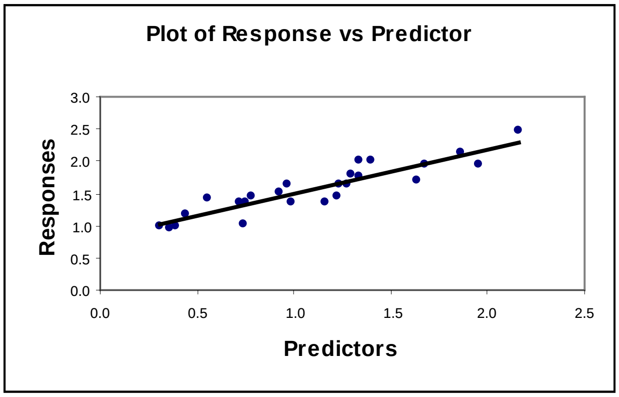


Figure 4.2. Plot of y versus x with the least squares fitted line

Here, the estimated values of the regression coefficients $\beta_0, and\ \beta_1$ are:

$\hat{\beta}_1 = 0.69$ and $\hat{\beta}_0 = 0.81$ respectively.

The regression problem at its core is a decomposition of the variances problem. The goal is to break the overall variation in the data into its component pieces known as *analysis of variance (ANOVA)*. Figure 4.3 below demonstrates graphically the component parts of total variance.

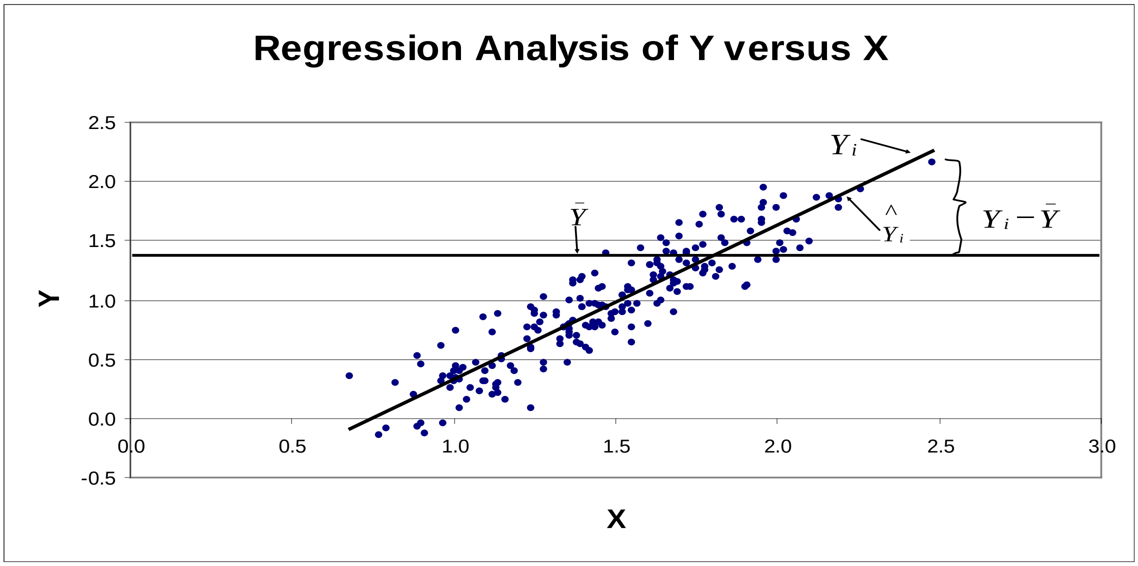


Figure 4.3. Decomposition of total variance into component variances

Notes:

The total variation given by the total sum of squares and the component sums of squares is given by the equation:

$$\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2=\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2+\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2 \qquad (4.6)$$

Where, $\bar{y}$ is the overall average of the responses irregardless of the predictor variable *x*, $\hat{y}_i$ is the predicted value of the response at $x = x_i$, and y_i is the value of the response variable at $x = x_i$.

$\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2$ is called the total sum of squares shortened to SST, $\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2$ is called the error sum of squares shortened to SSE, and $\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2$ is known as the regression sum of squares rewritten as SSR in short.

To determine the fitness of the function estimated, we first set up the Analysis of variance (ANOVA) table, and determine the statistical significance of the parameters estimated. Table 4.2 shows a decomposition of the total variation in the data for a sample size equal to *n* and number of parameters equal to *p* (including the intercept term).

Sources of Variation	*Degrees of Freedom*	*Sum of Squares*	*Mean Square*	*F-Statistic*	*Pr>F*
Regression	*p-1*	*ssr*	*msr = ssr/p-1*		
Error	*n-p*	*sse*	*mse = sse/n-p*		
Total	*n-1*	*sst*			

Table 4.2: Analysis of variance (ANOVA)

As shown in the Table, the F-statistic is computed which follows an F-distribution with *p-1, n-p* degrees of freedom. The last column Pr > F, is the *p*-value, which is the area under the F-distribution curve to the right of the F-value (percentile). Also, a descriptive measure of goodness of fit is determined using the R^2-statistic. $R^2 = 1 - \frac{SSE}{SST}$ is simply square of the correlation coefficient $r\,(-1 \le r \le 1)$ - a measure of degree of linear relationship between the dependent and the independent variable. A value of R^2 close to 1 suggests a strong linear relationship. The R^2-statistic $\left(0 \le R^2 \le 1\right)$ together with the F-statistic may be used to determine predictive ability of the model.

Proper implementation of simple linear regression requires the following steps, namely, *model selection, model fitting and model validity.* We will discuss each step in meticulous detail.

Section 4.3 Model Selection

In this first step, graph the dependent variable *Y* versus the independent variable *X* as a scatter plot for visual display to determine the relationship. An approximate

Notes:

linear trend would suggest fitting a straight line, and any non-linear trend may require polynomial fitting. While grasping the relationship is straight forward in many instances, certain amount of art is involved in determining the approximate relationship. Upon determining the presumptive model for fitting, ordinary least squares is applied to estimate model parameters. The following example is used throughout this chapter to demonstrate ideas and techniques.

Example 4.2 Aerobic fitness data

Consider the aerobic fitness data set, we introduced in chapter 2. We will use simple linear regression to develop an equation to predict fitness based on exercise tests rather than on expensive and cumbersome oxygen consumption measurements

As an illustration, we will model oxygen(*y*) as a function of rum time(*x*). Figure 4.4 is a plot of *y* versus *x*,

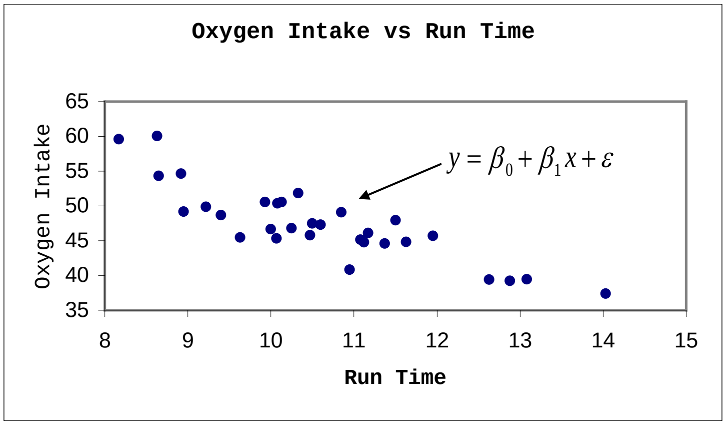


Figure 4.4 There appears to be a definite linear and negative relationship between oxygen intake and run time.

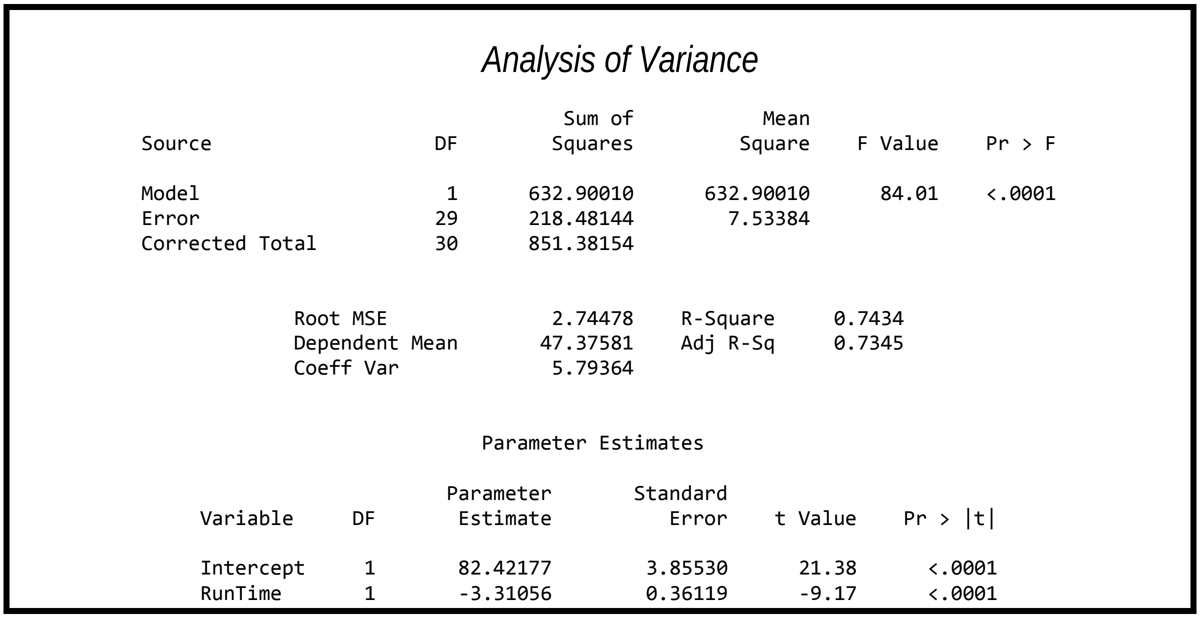
Since the relationship appears to be linear, we posit a linear model of the type, $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

The SAS code for fitting the proposed model is

```
Data slr ;
Input runtime oxygen;
Datalines;
11.37      44.609
10.07      45.313
 8.65      54.297
 8.17      59.571
 9.22      49.874
  .           .
  .           .
  .           .
  .           .
  .           .
 9.93      50.545
  9.4      48.673
 11.5       47.92
 10.5      47.467

  ;
  proc reg data = slr ;
  model oxygen = runtime ;
  run ;
```

Notes:

The SAS output of the program is given below.



Analysis of Variance

```
                                  Sum of           Mean
Source                  DF       Squares         Square    F Value    Pr > F

Model                    1     632.90010      632.90010      84.01    <.0001
Error                   29     218.48144        7.53384
Corrected Total         30     851.38154


             Root MSE              2.74478    R-Square     0.7434
             Dependent Mean       47.37581    Adj R-Sq     0.7345
             Coeff Var             5.79364


                             Parameter Estimates

                         Parameter       Standard
     Variable     DF      Estimate          Error    t Value    Pr > |t|

     Intercept     1      82.42177        3.85530      21.38      <.0001
     RunTime       1      -3.31056        0.36119      -9.17      <.0001
```

Table 4.3. Analysis of variance output using proc reg in SAS®

The first part is the familiar analysis of variance results. The model, and error components under the “source” column corresponds to “regression” and “error” in Table 4.2. Notice that the fitted model is statistically significant at the 99.99% confidence level. The column “`Pr > F`” equal to 0.0001. This is saying that the area to the right of 84.01(F value) is 0.0001. The R-square (R^2) = 0.7434 is measure of

Notes:

linear relationship between oxygen intake and run time. The F-statistic also known as the signal to noise ratio is merely, the ratio of model mean square and error mean square. A large value of the F-statistic implies a significant relationship.

In the “parameter estimates” portion of the output, the estimated values of the intercept and slope coefficients are $\hat{\beta}_0 = 82.42177$ and $\hat{\beta}_1 = -3.31056$ respectively. A negative value of the slope coefficient indicates negative relationship between oxygen intake and run time. The graph in Figure 4.4 confirms this finding. As we pointed out earlier, model significance relates to testing the hypothesis $H_0 : \beta_1 = 0$ using the *t*-statistic. From the output, $t = \dfrac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} = \dfrac{-3.31056 - 0}{0.36119} = -9.17$.

Since -9.17 less than $t_{0.025,29} = -2.04523$, the slope coefficient is statistically significant at 95% confidence level. Notice that the square of *t* value equal to -9.17=84.01. The fitted model is given by $\hat{Y} = 82.42177 - 3.31056X$.

The estimated slope coefficient means that for every unit increase in the

Notes:

independent variable *x*, the change in the dependent variable is $\overset{\wedge}{\beta}_1$ units.

In the next section we will discuss goodness of fit of the fitted model.

Section 4.4. Determination of goodness of fit

The validity of the simple linear model is based on the following assumptions

1. Constant variance of the error term, i.e., $\text{var}(\varepsilon_i) = \sigma^2$
2. The error term is normally distributed, i.e., $\varepsilon_i \sim N(0, \sigma^2)$ (4.7)
3. The errors are independent and identically distributed

Upon fitting a model, the three assumptions in 4.7 should be verified for model validity.

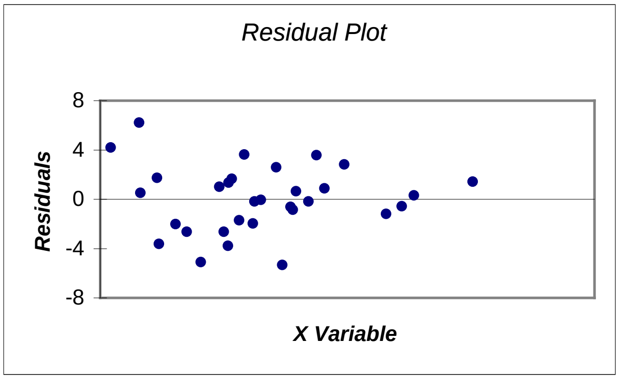


Figure 4.5. Residuals appear to be randomly scattered around zero satisfying assumptions one and three.

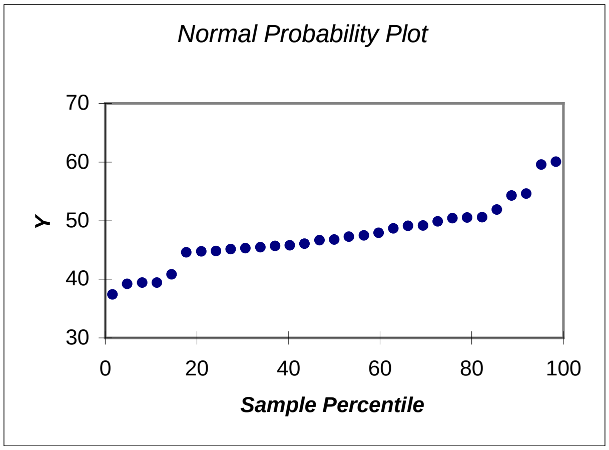


Figure 4.6 Normal quantile-quantile graph which is approximately linear confirms assumption three of normality of the residuals.

Section 4.5 Some useful hints

The fitted model should be used for predictions within the range of the values of the independent variable. In the oxygen intake example, the run time range is (8.17, 14.03).

Even against minor violations of the assumption of normality, the estimated model may be used for predictive purposes.

Section 4.6 Properties of the least squares estimator

The ordinary least squares estimator is an unbiased estimator, i.e.,

Notes:

$$E\left(\overset{\wedge}{\beta_1}\right) = \beta_1 \qquad (4.8)$$

This simply means that the average of estimated of values of $\overset{\wedge}{\beta_1}$ equals β_1.

The least squares estimator is a minimum variance estimator.

Section 4.7 Multiple Regression Analysis

It is quite common in applications to find relationships between a quantitative variable dependent *Y* versus several independent variables. Notice that this is an obvious extension of the single dependent versus the single independent simple linear regression. For example expenditure (y) may be related to education and gender in addition to earnings. In this scenario, an extension of simple linear regression, known as *multiple linear regression* is employed for model fitting.

All the assumptions of the simple linear model hold for the multiple regression models as well and implementationally it follows along the lines of the single variable model. Returning to our oxygen intake study, we will fit a multiple regression model with all the independent variables included. The next few lines are the SAS code.

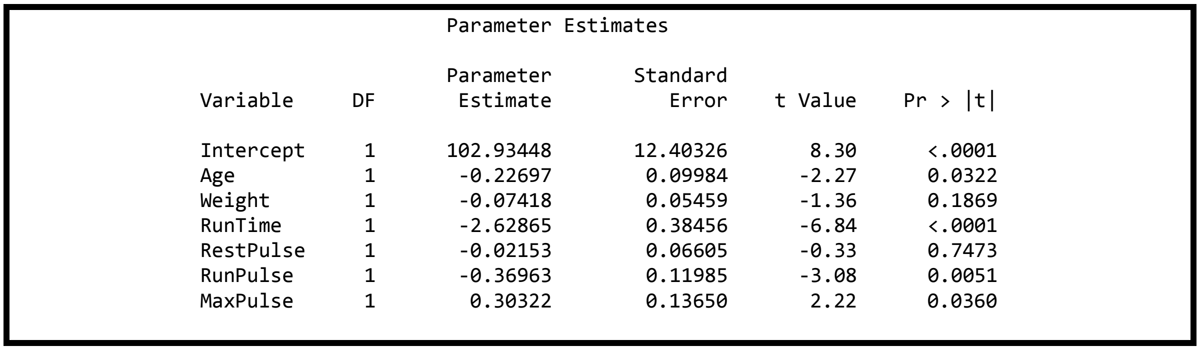
Notes:

Example 4.3 Oxygen intake study

```
Data mlr ;

Input Age Weight Oxygen RunTime RestPulse RunPulse MaxPulse;

Cards;
    44      89.47    44.609    11.37     62      178      182
    40      75.07    45.313    10.07     62      185      185
    44      85.84    54.297     8.65     45      156      168
    42      68.15    59.571     8.17     40      166      172
    38      89.02    49.874     9.22     55      178      180
     .        .         .        .        .        .        .
     .        .         .        .        .        .        .
     .        .         .        .        .        .        .
    51      73.71     45.79    10.47     59      186      188
    57      59.08    50.545     9.93     49      148      155
    49      76.32    48.673      9.4     56      186      188
    48      61.24     47.92     11.5     52      170      176
    52      82.78    47.467     10.5     53      170      172
    ;
    proc reg datal = mlr ;
    model oxygen = Age Weight RunTime RestPulse RunPulse MaxPulse ;
    run ;
```

The output is in the following.

```
                              The REG Procedure
                                Model: MODEL1
                          Dependent Variable: Oxygen

                   Number of Observations Read          31
                   Number of Observations Used          31


                             Analysis of Variance

                                      Sum of           Mean
 Source                   DF         Squares         Square    F Value    Pr > F

 Model                     6       722.54361      120.42393      22.43    <.0001
 Error                    24       128.83794        5.36825
 Corrected Total          30       851.38154


              Root MSE             2.31695    R-Square     0.8487
              Dependent Mean      47.37581    Adj R-Sq     0.8108
              Coeff Var            4.89057
```



```
                             Parameter Estimates

                             Parameter       Standard
Variable      DF              Estimate          Error    t Value    Pr > |t|

Intercept      1             102.93448       12.40326       8.30      <.0001
Age            1              -0.22697        0.09984      -2.27      0.0322
Weight         1              -0.07418        0.05459      -1.36      0.1869
RunTime        1              -2.62865        0.38456      -6.84      <.0001
RestPulse      1              -0.02153        0.06605      -0.33      0.7473
RunPulse       1              -0.36963        0.11985      -3.08      0.0051
MaxPulse       1               0.30322        0.13650       2.22      0.0360
```

Table 4.4. Anova table and parameter estimates of oxygen intake study data

The "*p*-value" corresponding to the F-value of 22.43 is equal to 0.0001, which means that the overall model with the six input variables is statistically significant at the (1-0.0001)*100%=99.99 percent confidence level. In multiple regression modeling, it is more appropriate to use the adjusted R^2 statistic as a measure of linear relationship. It adjusts for the number of independent variables included in the model. It is well known that as more independent variables are included, the R^2 statistic increases. Notice that adjusted value of R^2 is equal to 0.81 as opposed to 0.85. The parameter estimates portion of the output shows the individual effect of the independent variables. Note that variables weight and rest pulse are statistically insignificant. This simply means that neither variable contributes to explaining the behavior of

Notes:

oxygen intake. These variables may be dropped from the independent variable list and the model may be refit. In situations where the experimenter is not sure about which variables to be included for analysis, an exploratory variable selection methodology may be undertaken. These procedures are popularly called *variable selection* methods. Variable selection methods entail fitting many regression models in a methodical fashion to derive the *best fitting model. Some of the popular ones are the forward selection, backward* selection, and *max r*. We will explain each method briefly and demonstrate the application of the max *r* using our oxygen intake data.

Section 4.8 Model selection methods

- *Forward Selection*

The forward selection procedure at the outset fits all possible one variable regression models and selects the best fitting one variable model. In the second round of fitting, all possible 2 variable models are fit and likewise, selects the best fitting two variable model. The procedure iterates until all possible variables are considered. It may be

Notes:

that in each round, significant variables of the preceding round may be dropped as their effect may be diminished, when considered in conjunction with variables entering the model in the current iteration of model fitting.

- *Backward selection*

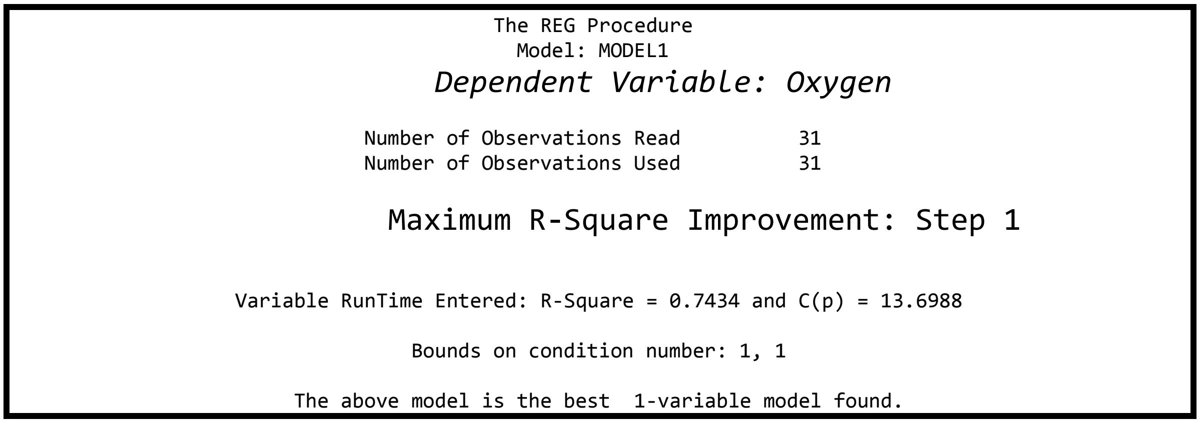
In this approach, model selection begins by considering all the independent variables. Variables that do not contribute to explaining variation in the dependent variable are systematically eliminated until only those variables that are significant are found and retained for further analysis.

- *Max r Selection*

The max r option utilizes the R^2 statistic to determine the variables to be utilized for modeling. The procedure evolves by adding variables sequentially. A log consisting of contribution of each entering variable is maintained. At the point where no further improvement in R^2 is observed, the procedure is terminated. For visual appreciation, a graph of the R^2 statistic versus the number of independent variables is plotted. The point at which the curve reaches a plateau, is a guide to selecting the number of variables to be included for further analysis.

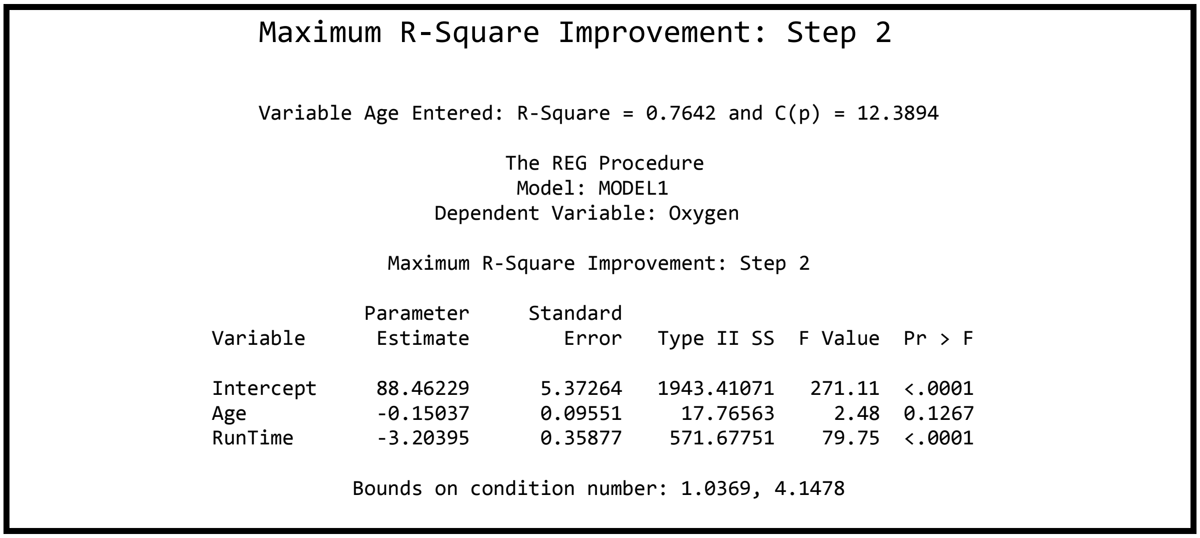
Example 4.3 continued Oxygen intake example

For illustration of the max r procedure for model selection, we only include partial output from SAS.



```
                          The REG Procedure
                            Model: MODEL1
                    Dependent Variable: Oxygen

               Number of Observations Read          31
               Number of Observations Used          31

             Maximum R-Square Improvement: Step 1


     Variable RunTime Entered: R-Square = 0.7434 and C(p) = 13.6988

                  Bounds on condition number: 1, 1

          The above model is the best  1-variable model found.
```

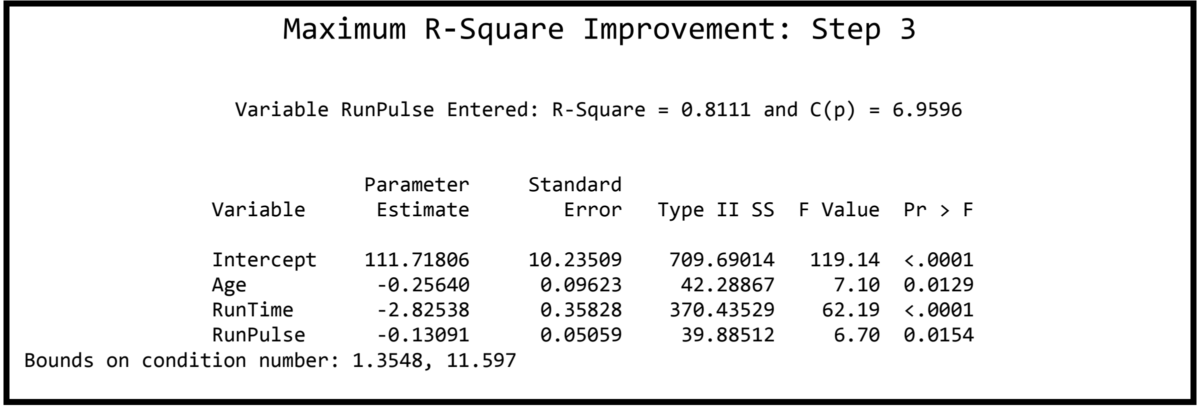
Table 4.5(a)Variable run time entered into the model



```
        Maximum R-Square Improvement: Step 2


        Variable Age Entered: R-Square = 0.7642 and C(p) = 12.3894

                           The REG Procedure
                             Model: MODEL1
                       Dependent Variable: Oxygen

                   Maximum R-Square Improvement: Step 2

                  Parameter     Standard
      Variable      Estimate        Error   Type II SS  F Value  Pr > F

      Intercept     88.46229      5.37264   1943.41071   271.11  <.0001
      Age           -0.15037      0.09551     17.76563     2.48  0.1267
      RunTime       -3.20395      0.35877    571.67751    79.75  <.0001

                 Bounds on condition number: 1.0369, 4.1478
```

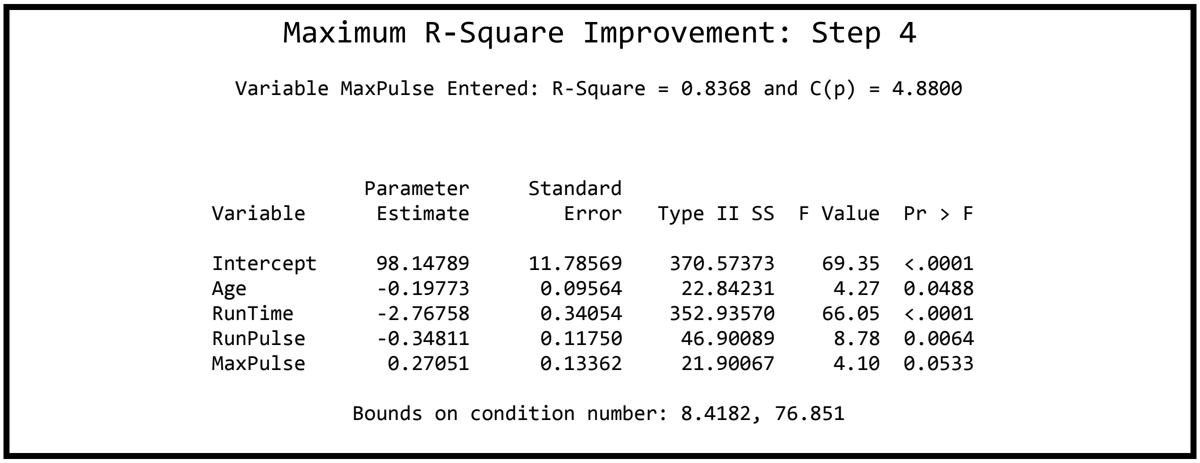
Notes:

Table 4.5(b)Variables run time and age entered into the model



```
                  Maximum R-Square Improvement: Step 3


       Variable RunPulse Entered: R-Square = 0.8111 and C(p) = 6.9596


                Parameter     Standard
  Variable       Estimate        Error   Type II SS  F Value  Pr > F

  Intercept     111.71806     10.23509    709.69014   119.14  <.0001
  Age            -0.25640      0.09623     42.28867     7.10  0.0129
  RunTime        -2.82538      0.35828    370.43529    62.19  <.0001
  RunPulse       -0.13091      0.05059     39.88512     6.70  0.0154
Bounds on condition number: 1.3548, 11.597
```

Table 4.5(c)Variables run time, age and run pulse entered into the model



```
                  Maximum R-Square Improvement: Step 4

       Variable MaxPulse Entered: R-Square = 0.8368 and C(p) = 4.8800



                Parameter     Standard
  Variable       Estimate        Error   Type II SS  F Value  Pr > F

  Intercept      98.14789     11.78569    370.57373    69.35  <.0001
  Age            -0.19773      0.09564     22.84231     4.27  0.0488
  RunTime        -2.76758      0.34054    352.93570    66.05  <.0001
  RunPulse       -0.34811      0.11750     46.90089     8.78  0.0064
  MaxPulse        0.27051      0.13362     21.90067     4.10  0.0533

             Bounds on condition number: 8.4182, 76.851
```

Table 4.5(d)Variables run time, age, run pulse and maximum pulse entered into the model

Continuing, the final six variable model yields an R^2 equal to ~0.85. The last two variables weight and rest pulse only add a meager 1% to R^2, suggesting that their contribution is negligible. The max r output produces an ANOVA table and other information, which we excised for reasons of space.

Figure 4.7 is a plot of R^2 versus the independent variables coded 1,2,3,4,5, and 6.

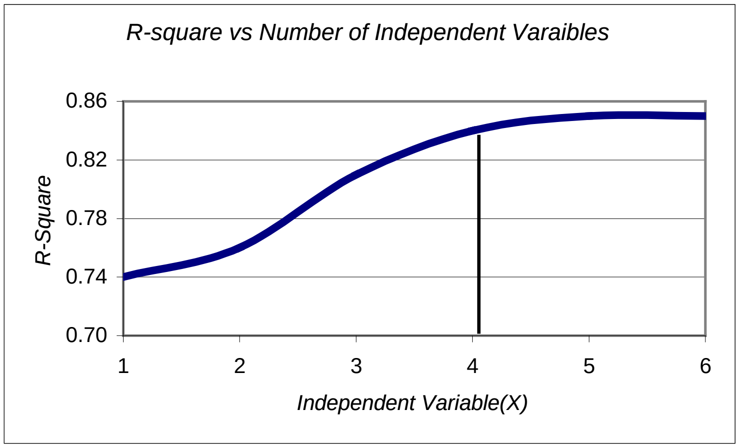


Figure 4.7. R-square appears to not increase significantly beyond X equal to 4. The recommended number of variables to be included in the model is four. Choose variables from the fourth step of model improvement from the SAS output.

From the graph, the four variable model appears to be suitable. Further study is conducted to finalize the model for predictions. The steps include evaluation of effects of multicollinearity, and outliers, and finally verification of model assumptions.

Section 4.9 Multicollinearity and diagnostics

Correlations among independent variables cause regression estimators to be erratic. This problem is termed as *multicollinearity*. Another way to state the problem is; when an independent variable is expressed approximately as a linear combination of other variables in the model, the estimated values of parameters are unstable. This simply means that they have large standard errors. This situation is quite undesirable. It is a good idea to find out which variables are nearly collinear. Multicollinearity is a common problem with observational data, i.e., when the investigator does not have control over the values that independent variables can assume. This is especially a common problem in economic data and was recognized by economists. PROC REG procedure provides several methods for detecting collinearity with the COLLIN, COLLINOINT, TOL, and VIF options.

The material below is optional

The COLLIN option in the MODEL statement requests that a collinearity analysis be performed. First, *X′X* is standardized to have ones on the diagonal. If the COLLINOINT option is specified, the intercept variable is adjusted out first. Then the eigen values and eigen vectors are computed. The PROC REG outputs the eigen values of *X′X* rather than singular values of *X*. The eigen values of *X′X* are the squares of the singular values of *X*.

The condition indices are the square roots of the ratio of the largest eigen value to each individual eigen value. The largest condition index is the condition number of the scaled *X* matrix. Belsey, et al (1980) suggest that, when this number is around 10, weak dependencies may be starting to affect the regression estimates. When

Notes:

this number is larger than 100, the estimates may have a fair amount of numerical error (although the statistical standard error almost always is much greater than the numerical error).

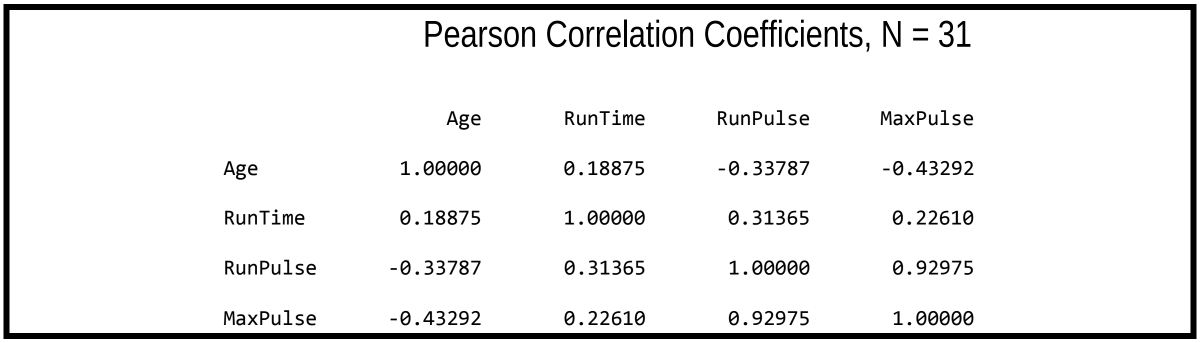
For each variable, PROC REG produces the proportion of the variance of the estimate accounted for by each principal component. A collinearity problem occurs when a component associated with a high condition index contributes strongly (variance proportion greater than about 0.5) to the variance of two or more variables.

The VIF option in the MODEL statement provides the Variance Inflation Factors (VIF). These factors measure the inflation in the variances of the parameter estimates due to collinearities that exist among the independent variables. There are no formal criteria for deciding if a VIF is large enough to affect the predicted values. The TOL option requests the tolerance values for the parameter estimates. The tolerance is defined as 1/*VIF*.

Using the COLLIN option in the model statement, and also computing the pair-wise

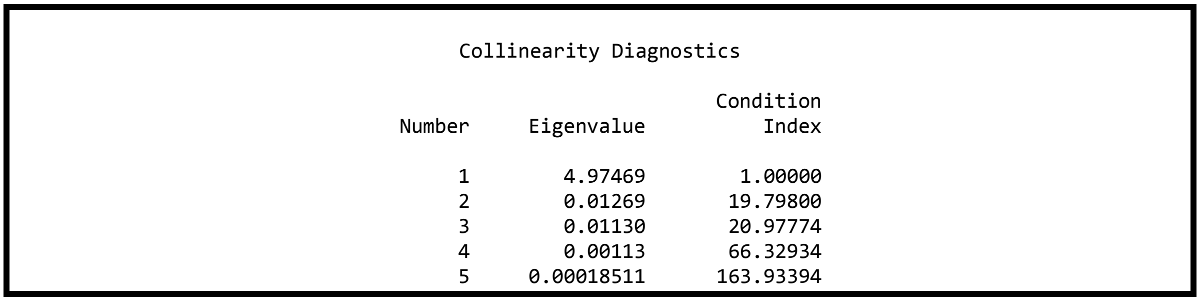
Notes:

correlation matrix using PROC CORR we find that strong collinearities do exist among the independent variables. The correlation matrix and the collinearity diagnostics obtained from SAS are reproduced.



```
                    Pearson Correlation Coefficients, N = 31

                         Age      RunTime     RunPulse     MaxPulse

Age                  1.00000      0.18875     -0.33787     -0.43292

RunTime              0.18875      1.00000      0.31365      0.22610

RunPulse            -0.33787      0.31365      1.00000      0.92975

MaxPulse            -0.43292      0.22610      0.92975      1.00000
```

Table 4.6(a). Pair-wise correlation matrix of the four independent variables age, runtime, run pulse, and maximum pulse



```
              Collinearity Diagnostics

                                     Condition
       Number      Eigenvalue            Index

            1         4.97469          1.00000
            2         0.01269         19.79800
            3         0.01130         20.97774
            4         0.00113         66.32934
            5      0.00018511        163.93394
```

Table 4.6(b). The Eigen values and condition number to determine multicollinearity

As there is a strong correlation between run pulse and max pulse, one of them may be chosen to be removed from the model. This is a very crude approach, and more sophisticated approaches such as *ridge regression* should be considered. Ridge regression is a technique which is able to handle the variance of ordinary least squares estimates by perturbing the *on diagonal* elements of the covariance matrix of independent variables.

Section 4.10 Effect of outliers in regression

Outliers have a tremendous impact on curve fitting. They have to be treated carefully. Models fitted with outliers included may lead to spurious results. The cartoon in Figure 4.8 conveys the serious effect of outliers, albeit humorously.

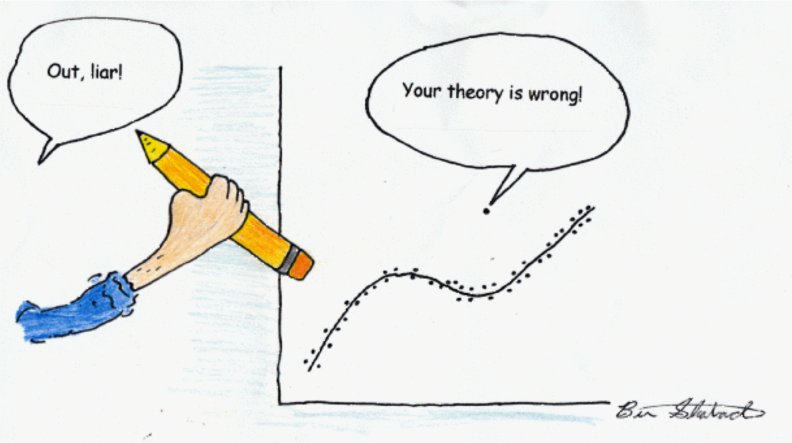


Figure 4.8. Outliers may have a significant effect on curve fitting.

Notes:

Outliers are anomalous observations. In a sense they are extreme observations. They can be identified by such tools as *box and whisker* plots, residual plots of standardized residuals versus *X* or the predicted values $\hat{Y}_i, i = 1,2,.....,n$. In a residual plot, outlier points are those that lie far beyond the main body of the scatter of residuals. In general they are beyond 3 standard deviations from zero. Outliers can have serious consequences in model fitting. In the presence of outliers, the fitted line or plane may be shifted disproportionately towards the aberrant observation. Because using method of least squares, entails minimizing sum of squares of error, squared sum of far out observations contribute significantly to the error term. If an outlier observation is suspected to be due to a typographical error, or measurement error, it should be discarded.

Notes:

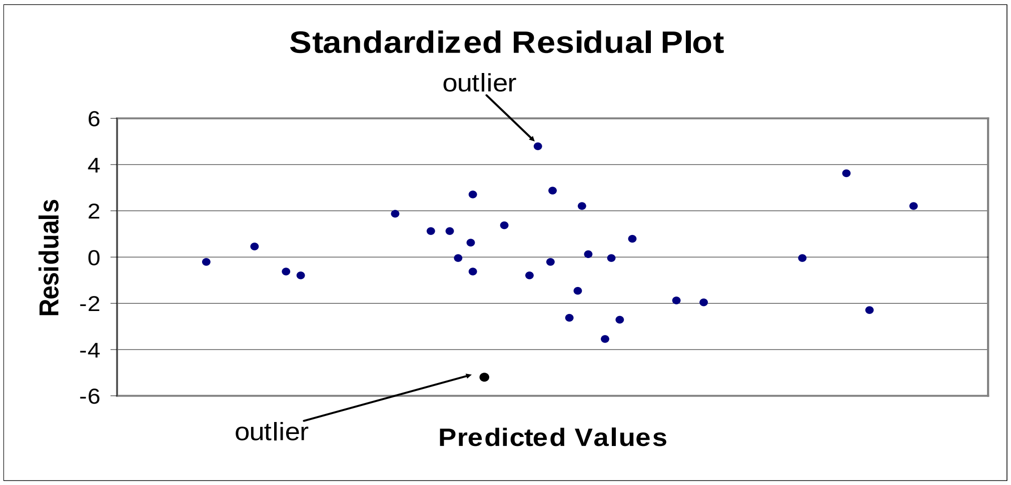


Figure 4.9. Residual plot of standardized residuals versus predicted values yhat. Observations with arrows pointing are outliers greater than four standard deviations away from zero. They may be discarded from future fitting and analysis.

Needless to say, outlier effects should be examined carefully. While discarding obvious ones is appropriate, outliers should be evaluated to determine if they should be eliminated or retained. If retained, we must somehow mitigate their influence on the fitted model. Figure 4.9 is an illustration of outliers in the exercise example data. There are other formal methods for outlier detection and elimination which we shall undertake in the following.

Section 4.11 A simple ad-hoc test for outliers

Notes:

After deleting the "i^{th}"observation denoted by $Y_i^{(d)}$, we a fit a regression function based on *n-1* observations. Let us assume that the *n-1* observations $Y_i, (i = 1,2,,....,n) \sim N\left(\hat{\beta}_0 + \hat{\beta}_1 X_i, \hat{\sigma}^2\right)$. If the probability $P\left(Y_i^{(d)} > 3\left(\hat{\sigma}\right)^2\right)$ is very small, it is evidence that the deleted observation does not belong to the distribution of *Y* values. The test requires fitting *n* different regression functions, with $Y_i^{(d)}, i = 1,2,,....,n$ deleted at each fitting epoch.

Regression analysis another example from Econometrics

We use the following example from Zaman, Rousseeuw, and Orhan (2001) to apply ordinary least squares (OLS) to study the growth in the gross domestic product. Towards that effort, De Long and Summers (1991) studied the national growth of 61 countries from 1960 to 1985 using OLS. Here the response and predictor variables are given as follows: The response variable is the growth in gross domestic product per worker (*GDP*) and the predictors are labor force growth (***LFG***), relative GDP gap (*GAP*), equipment investment (*EQP*), and non-equipment investment (*NEQ*). The data set is given in the Table below. Each column in the Table corresponds to, country, GDP, LFG, EQP, NEQ, and GAP

```
   Argentin  0.0089 0.0118 0.0214 0.2286 0.6079
   Austria   0.0332 0.0014 0.0991 0.1349 0.5809
   Belgium   0.0256 0.0061 0.0684 0.1653 0.4109
   Bolivia   0.0124 0.0209 0.0167 0.1133 0.8634
   Botswana  0.0676 0.0239 0.1310 0.1490 0.9474
```

Brazil	0.0437	0.0306	0.0646	0.1588	0.8498
Cameroon	0.0458	0.0169	0.0415	0.0885	0.9333
Canada	0.0169	0.0261	0.0771	0.1529	0.1783
Chile	0.0021	0.0216	0.0154	0.2846	0.5402
Colombia	0.0239	0.0266	0.0229	0.1553	0.7695
CostaRic	0.0121	0.0354	0.0433	0.1067	0.7043
Denmark	0.0187	0.0115	0.0688	0.1834	0.4079
Dominica	0.0199	0.0280	0.0321	0.1379	0.8293
Ecuador	0.0283	0.0274	0.0303	0.2097	0.8205
ElSalvad	0.0046	0.0316	0.0223	0.0577	0.8414
Ethiopia	0.0094	0.0206	0.0212	0.0288	0.9805
Finland	0.0301	0.0083	0.1206	0.2494	0.5589
France	0.0292	0.0089	0.0879	0.1767	0.4708
Germany	0.0259	0.0047	0.0890	0.1885	0.4585
Greece	0.0446	0.0044	0.0655	0.2245	0.7924
Guatemal	0.0149	0.0242	0.0384	0.0516	0.7885
Honduras	0.0148	0.0303	0.0446	0.0954	0.8850
HongKong	0.0484	0.0359	0.0767	0.1233	0.7471
India	0.0115	0.0170	0.0278	0.1448	0.9356
Indonesi	0.0345	0.0213	0.0221	0.1179	0.9243
Ireland	0.0288	0.0081	0.0814	0.1879	0.6457
Israel	0.0452	0.0305	0.1112	0.1788	0.6816
Italy	0.0362	0.0038	0.0683	0.1790	0.5441
IvoryCoa	0.0278	0.0274	0.0243	0.0957	0.9207
Jamaica	0.0055	0.0201	0.0609	0.1455	0.8229
Japan	0.0535	0.0117	0.1223	0.2464	0.7484
Kenya	0.0146	0.0346	0.0462	0.1268	0.9415
Korea	0.0479	0.0282	0.0557	0.1842	0.8807
Luxembou	0.0236	0.0064	0.0711	0.1944	0.2863
Madagasc	-0.0102	0.0203	0.0219	0.0481	0.9217
Malawi	0.0153	0.0226	0.0361	0.0935	0.9628
Malaysia	0.0332	0.0316	0.0446	0.1878	0.7853
Mali	0.0044	0.0184	0.0433	0.0267	0.9478
Mexico	0.0198	0.0349	0.0273	0.1687	0.5921
Morocco	0.0243	0.0281	0.0260	0.0540	0.8405
Netherla	0.0231	0.0146	0.0778	0.1781	0.3605
Nigeria	-0.0047	0.0283	0.0358	0.0842	0.8579
Norway	0.0260	0.0150	0.0701	0.2199	0.3755
Pakistan	0.0295	0.0258	0.0263	0.0880	0.9180
Panama	0.0295	0.0279	0.0388	0.2212	0.8015
Paraguay	0.0261	0.0299	0.0189	0.1011	0.8458
Peru	0.0107	0.0271	0.0267	0.0933	0.7406
Philippi	0.0179	0.0253	0.0445	0.0974	0.8747
Portugal	0.0318	0.0118	0.0729	0.1571	0.8033
Senegal	-0.0011	0.0274	0.0193	0.0807	0.8884
Spain	0.0373	0.0069	0.0397	0.1305	0.6613
SriLanka	0.0137	0.0207	0.0138	0.1352	0.8555
Tanzania	0.0184	0.0276	0.0860	0.0940	0.9762
Thailand	0.0341	0.0278	0.0395	0.1412	0.9174
Tunisia	0.0279	0.0256	0.0428	0.0972	0.7838
U.K.	0.0189	0.0048	0.0694	0.1132	0.4307
U.S.	0.0133	0.0189	0.0762	0.1356	0.0000
Uruguay	0.0041	0.0052	0.0155	0.1154	0.5782
Venezuel	0.0120	0.0378	0.0340	0.0760	0.4974
Zambia	-0.0110	0.0275	0.0702	0.2012	0.8695

```
   Zimbabwe  0.0110 0.0309 0.0843 0.1257 0.8875
```

Robust regression

The main purpose of robust regression is to detect outliers and provide resistant (stable) results in the presence of outliers. In order to achieve this stability, robust regression limits the influence of outliers. Historically, three classes of problems have been addressed with robust regression techniques:

- problems with outliers in the *y*-direction (response direction)
- problems with multivariate outliers in the *x*-space (i.e., outliers in the covariate space, which are also referred to as leverage points)
- problems with outliers in both the *y*-direction and the *x*-space

Many methods have been developed in response to these problems. However, in statistical applications of outlier detection and robust regression, the methods most commonly used today are Huber M estimation, high breakdown value estimation, and combinations of these two methods. The new ROBUSTREG procedure in this version provides four such methods: M estimation, LTS estimation, S estimation, and MM estimation.

1. M estimation was introduced by Huber (1973), and it is the simplest approach both computationally and theoretically. Although it is not robust with respect to leverage points, it is still used extensively in analyzing data for which it can be assumed that the contamination is mainly in the response direction.

2. Least Trimmed Squares (LTS) estimation is a high breakdown value method introduced by Rousseeuw (1984). The breakdown value is a measure of the proportion of contamination that an estimation method can withstand and still maintain its robustness. The performance of this method was improved by the FAST-LTS algorithm of Rousseeuw and Van Driessen (2000).
3. S estimation is a high breakdown value method introduced by Rousseeuw and Yohai (1984). With the same breakdown value, it has a higher statistical efficiency than LTS estimation.
4. MM estimation, introduced by Yohai (1987), combines high breakdown value estimation and M estimation. It has both the high breakdown property and a higher statistical efficiency than S estimation.

Section 4.12 Analysis of variance

In many testing situations, more than two populations are compared to study if the differences between the populations are due to random fluctuations in the data or

Notes:

rather due to arising from different populations. The two-sample *t*-test in its most general form is a tool to study differences between two populations. *Analysis of variance* is an extension of the two-sample t-test for comparing more than two populations. In analysis of variance, a continuous response variable, known as a *dependent variable*, is measured under experimental conditions identified by group variables, known as *independent variables or factors*. The variation in the response is assumed to be due to the sum of the effects of the group variable, and the random error term included to account for natural variations in measurement processes.

A one-way analysis of variance considers one treatment factor with two or more treatment levels. The goal of the analysis is to test for differences among the means of the levels and to quantify these differences. If there are two treatment levels, analysis of variance degenerates to the standard two-sample *t*-test used to compare means of two normal populations. The assumptions of analysis of variance modeling are that the treatment effects are additive and the random error terms follow normal distribution with mean equal to zero and a constant variance σ^2 and are

Notes:

independent.

The analysis of variance (ANOVA) model for modeling a single response variable versus a single independent variable measured at *k*-levels is mathematically expressed as

$$y_{ij} = \mu + T_i + \varepsilon_{ij}, \begin{cases} i = 1,2,...,n \\ j = 1,2,....,k \end{cases} \qquad (4.9)$$

and $\varepsilon_{ij} \sim N\left(0, \sigma^2\right)$, T_i is the treatment effect.

The formulation of the analysis of variance model argues that an observed response value fluctuates around the overall mean μ. The fluctuation is induced due to the treatment effect and the Gaussian random error term.

The data structure for an ANOVA model with *k*-levels of the independent variable is given in Figure 4.10.

T_1	T_2	.	.	.	T_k
y_{11}	y_{12}	.	.	.	y_{1k}
y_{21}	y_{22}	.	.	.	y_{2k}
.	.	.	.	.	.
.	.	.	.	.	.
y_{i1}	y_{i2}	.	y_{ij}	.	y_{ik}
.	.	.	.	.	.
.	.	.	.	.	.
y_{n1}	y_{n2}	.	.	.	y_{nk}

Figure 4.10. Data structure of a one-way analysis of variance layout.

The independent variable is often referred to in the analysis of variance jargon as the treatment variable. The origin of the word "treatment" is due to its applications in agricultural experiments. ANOVA was invented by the famous mathematical statistician Dr. Ronald Aylmer Fisher who is considered the father of modern statistics.

Suppose we want to test the hypothesis

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k \quad versus \quad \mu_i \neq \mu_j \text{ for at least one pair } \mu_i , \mu_j \qquad (4.10)$$

The analysis of variance technique decomposes the overall variation into its component parts, *within group variation* and *between groups variation*. Each level of the treatment variable may be viewed as a group. If the between groups variation dominates the within group variation, it may be evidence that there are significant differences between the treatment levels. The total variation and the within group, and between groups are simply the total sum of squares, within group sum of

Notes:

squares, and between group sum of squares. Mathematically,

$$\underbrace{\sum_{j=1}^{k}\sum_{i=1}^{n}\left(y_{ij}-\bar{y}\right)^2}_{SST} = \sum_{j=1}^{k}\sum_{i=1}^{n}\left(y_{ij}-\bar{y}_{.j}+\bar{y}_{.j}+\bar{y}\right)^2 = \underbrace{\sum_{j=1}^{k}\sum_{i=1}^{n}\left(y_{ij}-\bar{y}_{.j}\right)^2}_{SSW} + \underbrace{\sum_{j=1}^{k}\sum_{i=1}^{n}\left(\bar{y}_{.j}-\bar{y}\right)^2}_{SSB} \quad (4.11)$$

The notation, $\bar{y}_{.j}$ means that the average is computed by holding the variable indexed by *j* as a constant, and adding over the cells in the ANOVA data structure indexed by $"i"$. The terms SST, SSW, and SSB are abbreviations for total sum of squares, within group sum of squares, and between groups sum of squares. The ANOVA decomposition is represented in a tabular form known as the ANOVA table akin to the decomposition in regression modeling.

Sources of Variation	*Degrees of Freedom*	*Sum of Squares*	*Mean Square*	*F-Statistic*	*Pr>F*
Between Gropus	*k-1*	*ssb*	*msr = ssb/k-1*		
Within Group	*k(n-1)*	*ssw*	*msw = sse/k(n-1)*		
Total	*nk-1*	*sst*			

Figure 4.11. Sources of variation due to the decomposition of total variation using a one-way analysis of variance model

Notes:

If the F-statistic which is a ratio involving between group sum of squares and within group sum of squares is much greater then one, it suggests that between group variation dominates within group variation, pointing to differences among the *k* classes of the treatment factor. An *F-statistic* value greater than the $100(1-\alpha)\%$ percentile point of the *F-distribution* given by $F_{\alpha,k-1,k(n-1)}$ is indicative of significant differences among the *k*-classes. Note that *k-1*, and *k(n-1)* are the numerator and denominator degrees of freedom. The ANOVA table is useful only to the extent of establishing presence or lack of the treatment effect on the response variable. It does not tell which two groups are statistically significantly different. We typically use the multiple comparisons test to find levels of the treatment factor that are significantly different from each other. The *least significant distance* or the *Tukey* test is a popular choice for multiple comparisons.

Example 4.4 Effect of bacteria on nitrogen content of red clover plants

The following example studies the effect of bacteria on the nitrogen content of red

Notes:

clover plants. The treatment factor is bacteria strain, and it has six levels. Five of the six levels consist of five different *Rhizobium trifolii* bacteria cultures combined with a composite of five *Rhizobium meliloti* strains. The sixth level is a composite of the five *Rhizobium trifolii* strains with the composite of the *Rhizobium meliloti*. Red clover plants are inoculated with the treatments, and nitrogen content is later measured in milligrams. The data are derived from an experiment by Erdman (1946). The data is given Table 4.7 below. Each pair of observations corresponds to a particular strain and the observed nitrogen content on each red clover plant.

```
3DOK1  19.4 3DOK1  32.6 3DOK1  27.0 3DOK1  32.1 3DOK1  33.0
3DOK5  17.7 3DOK5  24.8 3DOK5  27.9 3DOK5  25.2 3DOK5  24.3
3DOK4  17.0 3DOK4  19.4 3DOK4   9.1 3DOK4  11.9 3DOK4  15.8
3DOK7  20.7 3DOK7  21.0 3DOK7  20.5 3DOK7  18.8 3DOK7  18.6
3DOK13 14.3 3DOK13 14.4 3DOK13 11.8 3DOK13 11.6 3DOK13 14.2
COMPOS 17.3 COMPOS 19.4 COMPOS 19.1 COMPOS 16.9 COMPOS 20.8
```

Table 4.7 Six strains of bacteria and nitrogen content of thirty red clover plants. Experimental data (courtesy, Erdman, 1946)

Notes: