CHILD HEALTH AND DEVELOPMENT STUDY

9. Diagnostics

In this section various diagnostic tools will be used to evaluate the adequacy of the regression model with the five independent variables developed in Section 8. These tools include residual plots to investigate whether the assumptions of linearity of the response variable, and normality and constant variance of the error appear to be met. They also provide insight into how the predictor variables are related to one another and how they influence the model.

9.1 Checking the Linearity of Birth Weights

In the previous section we have described the relationship between the response variable BWT and the predictors GESTWKS, MNOCIG, MHEIGHT, MPPWT, and FHEIGHT by the following multiple linear regression model:

$$BWT = \beta_0 + \beta_1 * GESTWKS + \beta_2 * MNOCIG + \beta_3 * MHEIGHT + \beta_4 * MPPWT +$$
$$+ \beta_5 * FHEIGHT + ERROR.$$

The random variable *ERROR* is assumed to follow a normal distribution with the mean of zero and an unknown standard deviation σ. The standard deviation is constant at all levels of the response variable *BWT* under a range of settings of the independent variables *GESTWKS, MNOCIG, MHEIGHT, MPPWT, and FHEIGHT*.

We have found the estimates of the model coefficients and discussed the significance of the five independent variables in the model.

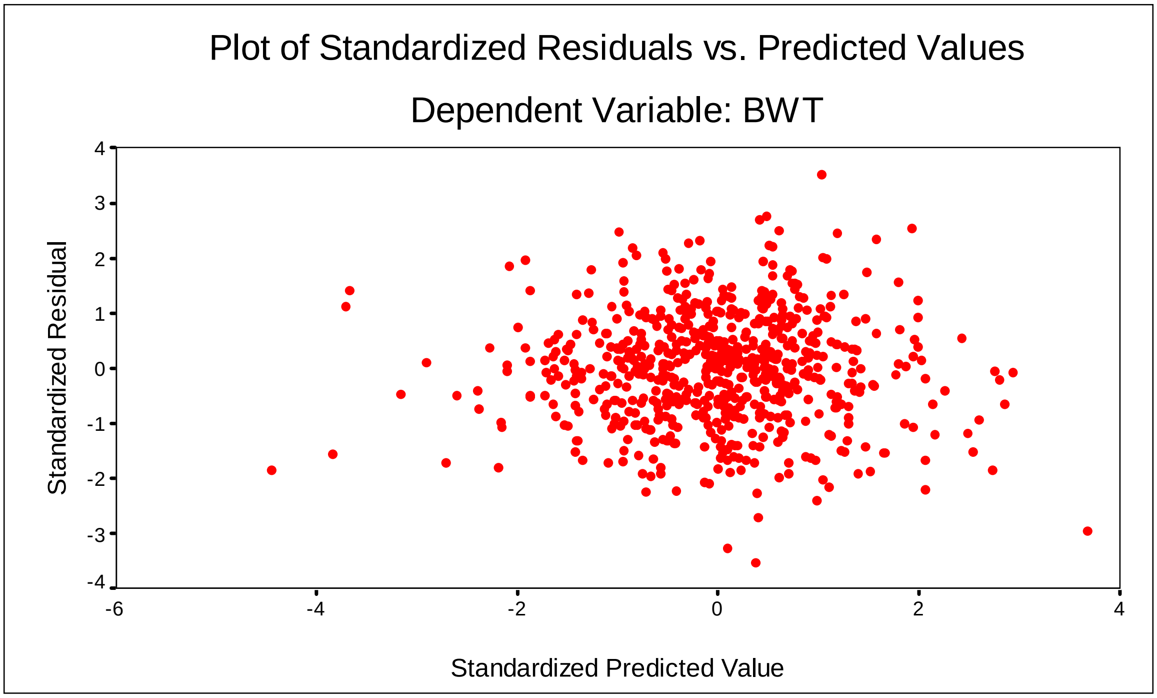
The above model is based on the assumption of a linear relationship between birth weight and the predictor variables. The assumption of linearity is easily examined through plots of residuals (standardized residuals) against predicted values and also against each independent variable. The standardized residuals are ordinary residuals divided by the sample standard deviation of the residuals and have mean 0 and standard deviation 1. If the linearity assumption appears to be met, then these residual plots should exhibit a random scatter of points about zero. Any consistent curvilinear pattern in the residuals indicates a nonlinear relationship between the dependent variable and independent variables and calls for a nonlinear regression model.

In order to check the assumption for the above multiple linear regression model with SPSS, we will obtain the plots of the standardized residuals against the predicted values and against each independent variable. As the plots will also be used to assess the constant variance of the error assumption, we will discuss the two assumptions together in the next section.

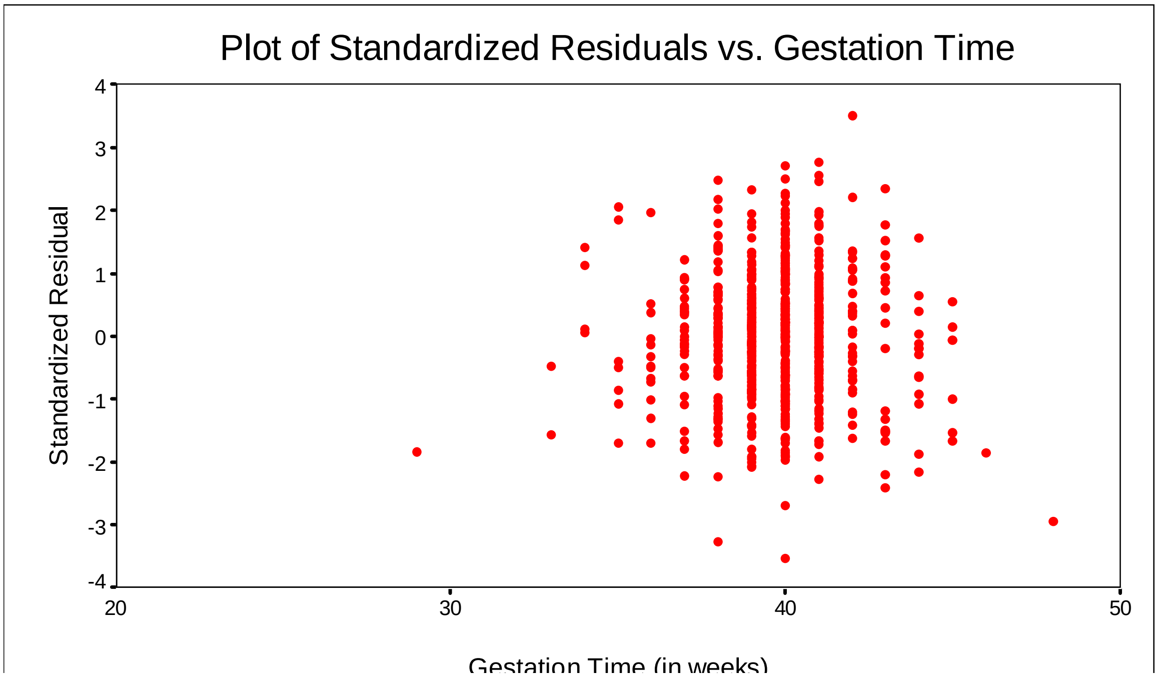
9.2 Checking Constant Variance Assumption

In order to see whether the assumption of constant variance is not violated, we plot residuals (standardized residuals) against the fitted and also against each predictor variable. If the assumptions of linearity and constant variance appear to be met, then these residual plots should exhibit a random scatter of points with similar spread across all levels of fitted and independent variable values.

The plot displayed below shows the scatterplot of standardized residuals against the corresponding fitted values. No obvious difficulties are revealed in this display. With the exception of the smallest fitted values, the variability appears to be quite similar across all levels of fitted values. A random pattern is apparent in the plot, the linearity assumption is not violated.

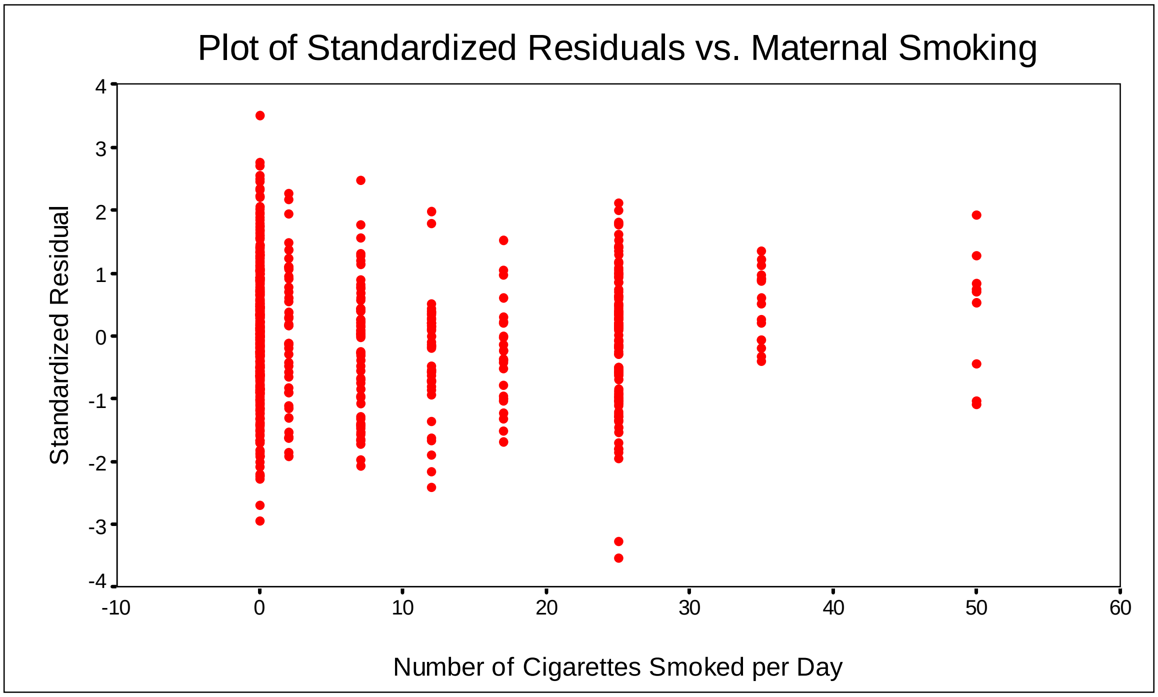


The next plot gives a similar display of standardized residuals plotted against gestation time.



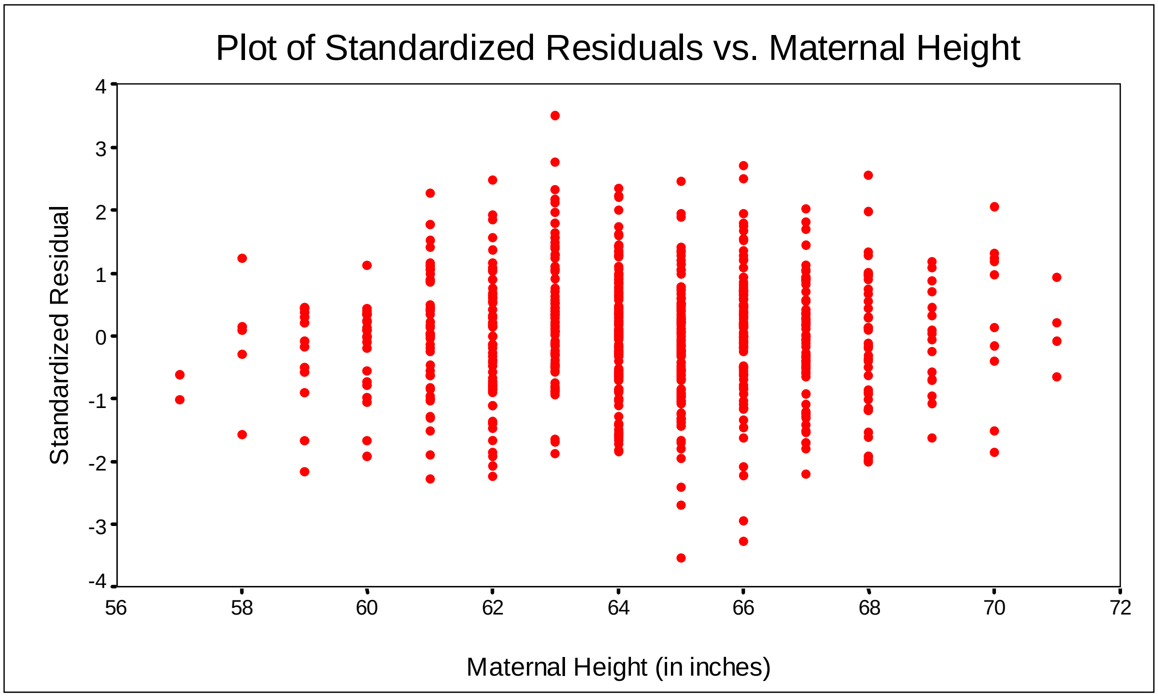
The points in the plot are randomly scattered and constant variability across all values for the independent variable is supported. The random pattern in the plot does not provide any evidence to question the linearity assumption.

The next plot gives a similar plot for the amount of maternal smoking variable.



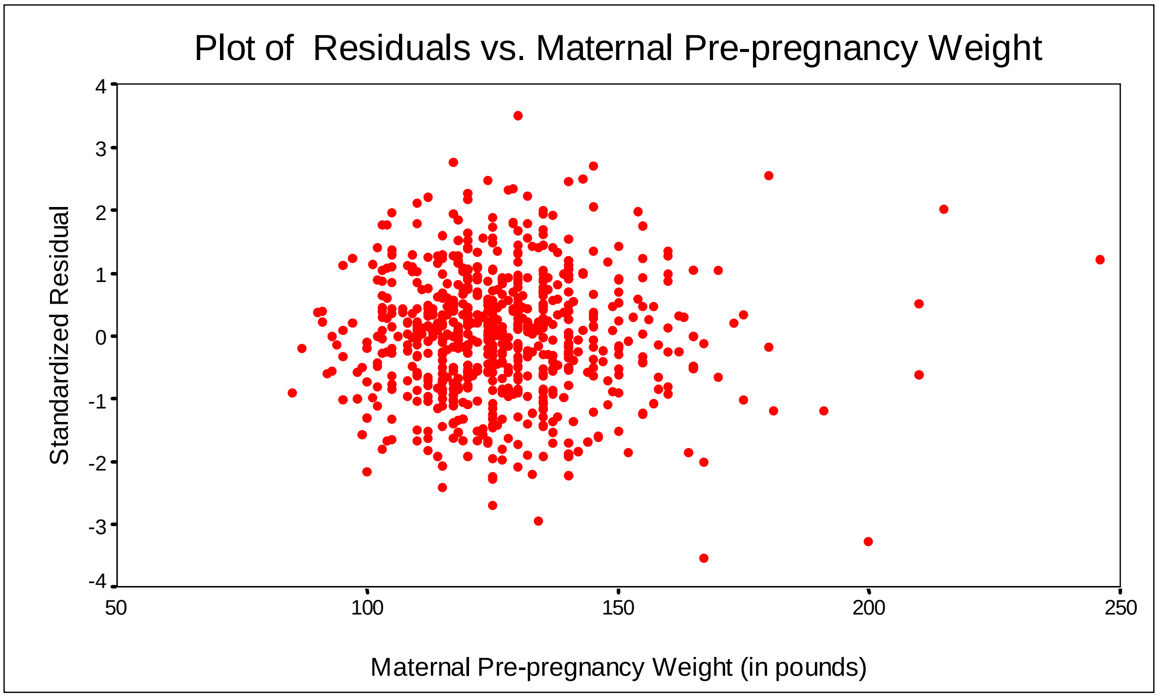
At the first glance, it seems that the variability decreases as the number of cigarettes smoked per day increases. As the change in spread refers to a relatively small fraction of observations, this plot does not suggest any weakness of the model with respect to nonconstant variability.

The next plot shows the standardized residuals versus maternal height.



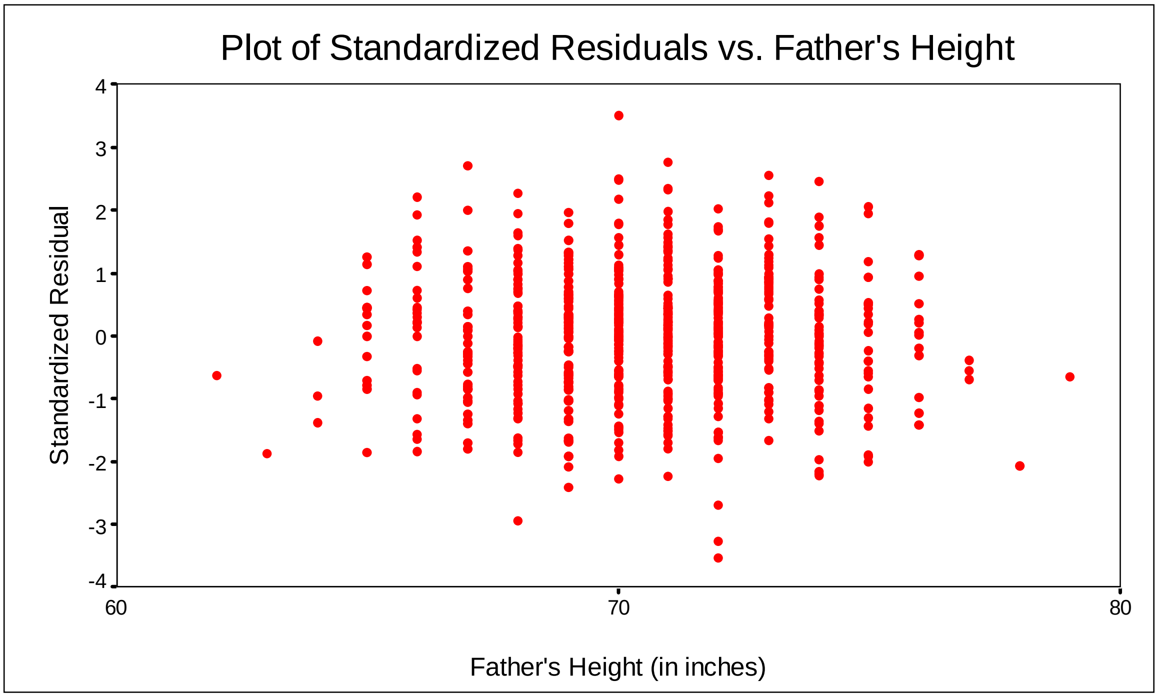
There appear to be some differences of variability in residuals for different levels of education, but the effect is not severe. There is not strong enough evidence to question the assumption of equal variance.

Now we plot the standardized residuals against maternal pre-pregnancy weight.



The plot shows residuals falling randomly, with relatively equal spread about zero and no strong tendency to be either greater or less than zero.

Finally we examine the plot of standardized residuals against father's height.

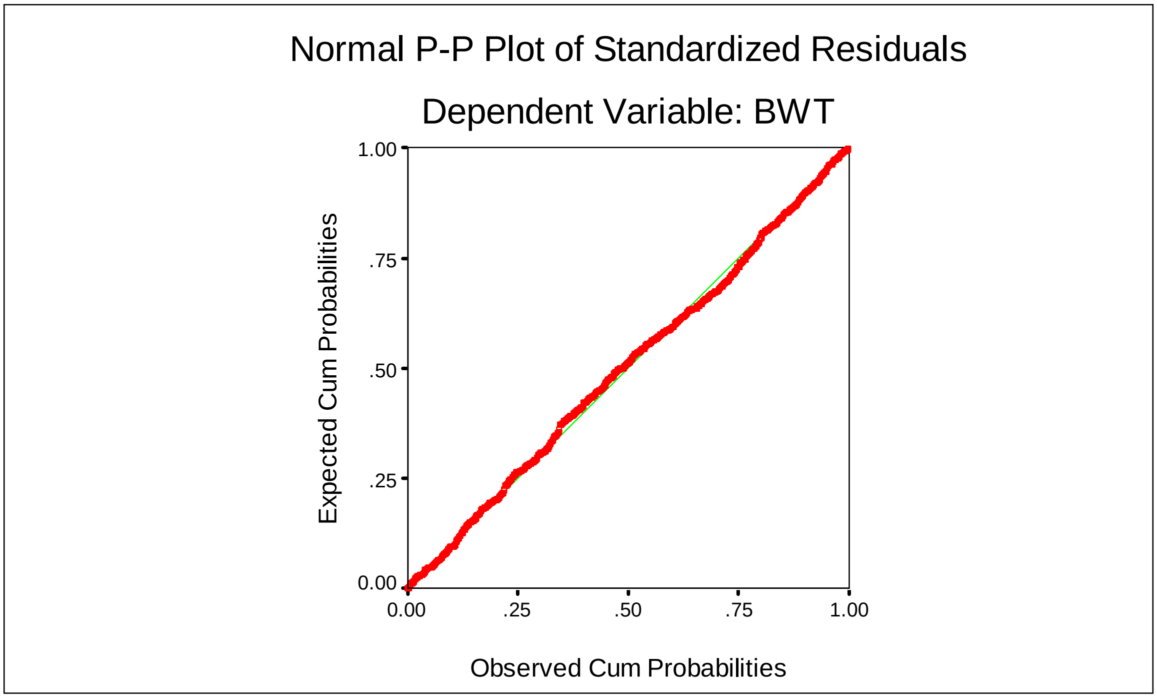


There appear to be some differences of variability in residuals for different levels of father's height, but the effect is not severe. There is not strong enough evidence to question the assumption of linearity and equal variance.

In summary, the assumptions of linearity and constant variance of the error seem very reasonable for this choice of regression model.

9.3 Checking Normality of the Error Assumption

In order to assess whether the assumption is not violated, the normal P-P plot of regression standardized residuals is obtained. The plot plots the cumulative proportions of standardized residuals against the cumulative proportions of the normal distribution. If the normality assumption is not violated, points will cluster around a straight line.



As the points in the plot are lying along a straight line, the normality assumption is strongly supported.

9.4 Multicollinearity of the Independent Variables

It is well known that collinearity and multicollinearity can have harmful effects on multiple regression, both in the interpretation of the results and in how they are obtained. In particular, collinearity affects parameter estimates and their standard errors, and consequently t ratios. Inflated standard errors mean wider confidence intervals for the regression coefficients and a diminished ability of tests to find significant results.

The use of several variables as predictors in the child health and development regression model makes the assessment of multiple correlation necessary to identify multicollinearity. But this is not possible by examining only the correlation matrix, which shows only simple correlations between two variables.

The simplest means of identifying collinearity is an examination of the correlation matrix for the predictor variables. The presence of high correlations (generally .90 and above) is the first indication of substantial collinearity. Lack of any high correlation values, however, does not ensure a lack of collinearity. Collinearity may be due to the combined effect of two or more other independent variables. The correlation matrix shows only simple correlations between two variables.

The correlation matrix for the child health data does not reveal any high correlation between two predictor variables. The highest correlation of 0.817 is between age of

Two measures for assessing both pairwise and multiple variable collinearity available in SPSS are the tolerance and the variance inflation factor (VIF). **Tolerance** is the amount of variability of the selected independent variable not explained by the other independent variables. It is obtained by making each independent variable a dependent variable and regressing it against the remaining independent variables. Tolerance values approaching zero indicate that the variable is highly collinear with the other predictor variables. The **variance inflation factor** (VIF) is inversely related to the tolerance value: $VIF = 1/TOLERANCE.$ Large VIF values (a usual threshold is 10.0, which corresponds to a tolerance of .10) indicate a high degree of collinearity or multicollinearity among the independent variables.

The following output displays the values of tolerance and VIF for the predictor variables in the case study.

```
--------------------- Variables in the Equation ----------------------------------

Variable          B        SE B       Beta      Tolerance   VIF     T
GESTWKS      .233851  .019363     .401496    .991583    1.008  12.077
MNOCIG      -.014619  .003220    -.150857    .992331    1.008  -4.540
MHEIGHT      .040042  .017358     .091028    .703816    1.421   2.307
MPPWT        .008308  .002330     .135979    .753340    1.327   3.565
FHEIGHT      .039711  .014387     .095913    .907532    1.102   2.760
(Constant)  -8.114024  1.40914                                   -5.758
```

No VIF value exceeds 10.0, and the tolerance values show that collinearity does not explain more than 10 percent of any independent variable's variance. There is no evidence of a significant collinearity in the problem.

SPSS regression collinearity diagnostics includes also the condition indices and the regression coefficient variance-decomposition matrix. A large condition index (over 30) indicates a high degree of collinearity. The regression coefficient variance-decomposition matrix shows the proportion of variance for each regression coefficient (and its associated variable) attributable to each condition index.

In order to examine collinearity, we first identify all condition indices above the threshold value of 30. Then for all condition indices exceeding the threshold, we identify variables with variance proportions above 0.90. A collinearity problem is indicated when a condition index identified as above the threshold value accounts for a substantial proportion of variance (.90 or above) for *two or more coefficients*. Thus each row in the matrix with the proportions exceeding 0.90 for at least two coefficients indicates significant correlations among the corresponding variables.

The collinearity diagnostics table for the child health and development data is displayed below:

```
Collinearity Diagnostics

Num  Eigen     Cond                    Variance Proportions
               Index  Constant   GEST   MNOCIG   MHEIG    MPPWT   FHEIGHT
   1  5.32943   1.000   .00002  .00008   .00964  .00004   .00050    .00004
   2   .65271   2.857   .00001  .00005   .98056  .00002   .00036    .00003
   3   .01455  19.137   .00300  .01201   .00173  .00096   .79216    .00463
   4   .00193  52.585   .00366  .76198   .00675  .05693   .03514    .13819
   5   .00090  76.825   .00571  .01197   .00114  .61305   .09391    .62993
   6   .00048  105.79   .98760  .21391   .00019  .32899   .07792    .22718
```

As you can see, there are three condition indices exceeding the threshold value of 30. However, none of the three condition indices accounts for a substantial proportion of variance (0.90 or above) for two or more coefficients. Thus, we can find no support for the existence of multicollinearity.

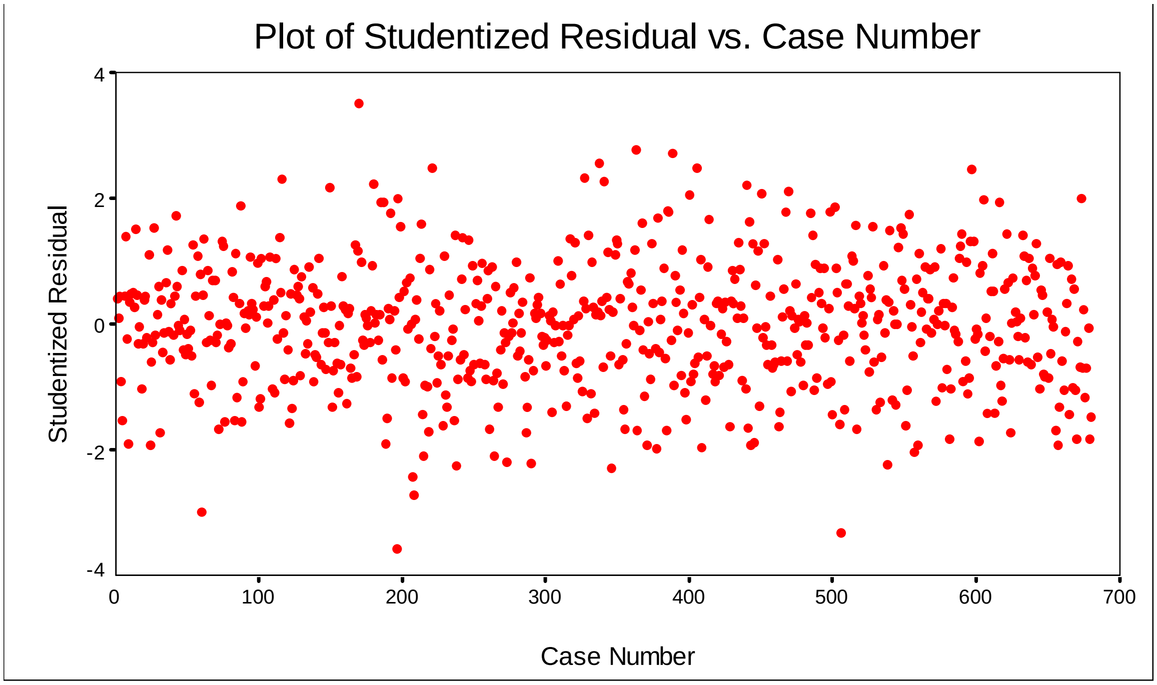
9.5 Diagnostics for Outliers and Influential Cases

Now we consider diagnostics for outliers and influential cases. An outlier is not necessarily an influential point, nor do all influential points have to be outliers. Thus, different statistical tools are used to identify outliers and influential observations. Studentized residuals are used for flagging outliers, and leverages and Cook's distances for flagging influential cases.

A studentized residual is a residual divided by its estimated standard deviation. The standardization makes the residuals directly comparable (larger predicted values have larger residuals). The studentized residual is the primary indicator of an observation that is an outlier on the dependent variable. With a fairly large sample size (50 or above), we may use a rule of thumb that studentized residuals smaller than -2 or larger than 2 are substantial. Observations falling outside the range can be considered *potential* outliers.

Instead of using cutoffs based on distributional assumptions, many researchers plot the studentized residuals, looking for points that stand apart from the others.

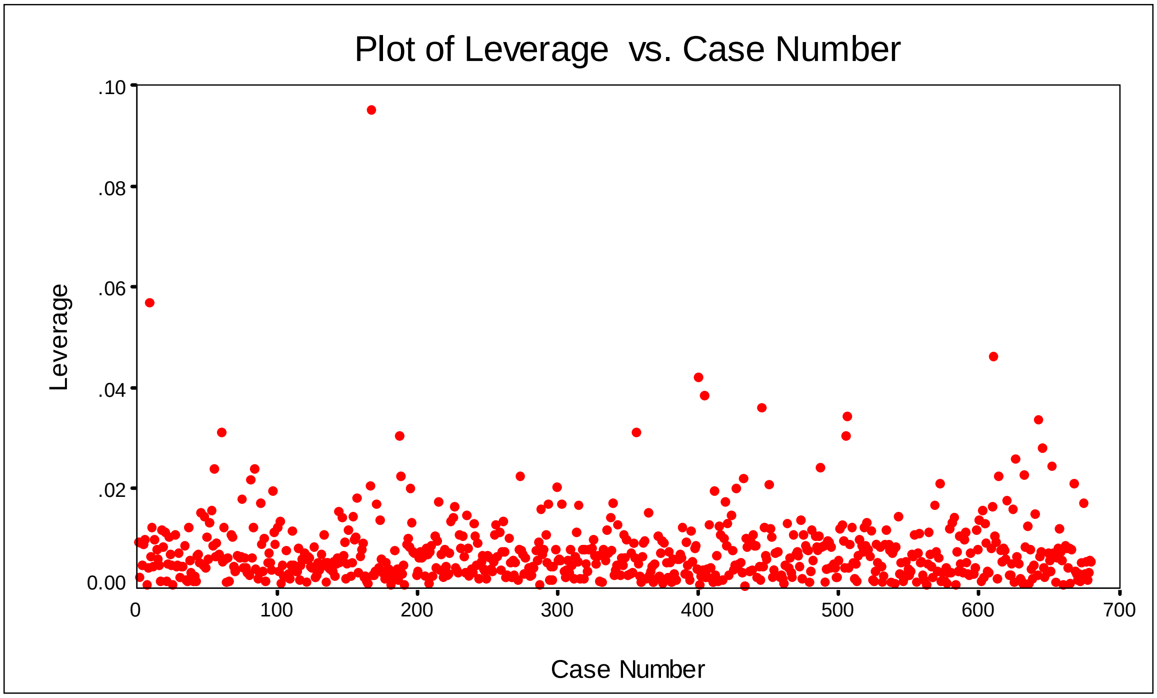
The following plot is a plot of studentized residuals versus case number for the child health and development data.



There are several cases with the studentized residuals above 2 or below -2 threshold. However, there are no points that stand apart from the others.

The leverage of a case is a measure of the distance between its explanatory variable values and the average of the explanatory variable values in the entire data set. This observation has substantial impact on the regression results due to its differences from other observations. Leverages are greater than 1/n and less than 1, and the average of all leverages in a data set is always p/n, where p is the number of regression variables. While a large leverage does not necessarily indicate that the case is influential, it does imply that the case has a high potential for influence. Statisticians use (2*p)/n as a lower cutoff point for flagging potential influential cases (if p>10 and n>50), (3*p)/n otherwise. Instead of using cutoffs, many researchers are looking for points that stand apart from the others.

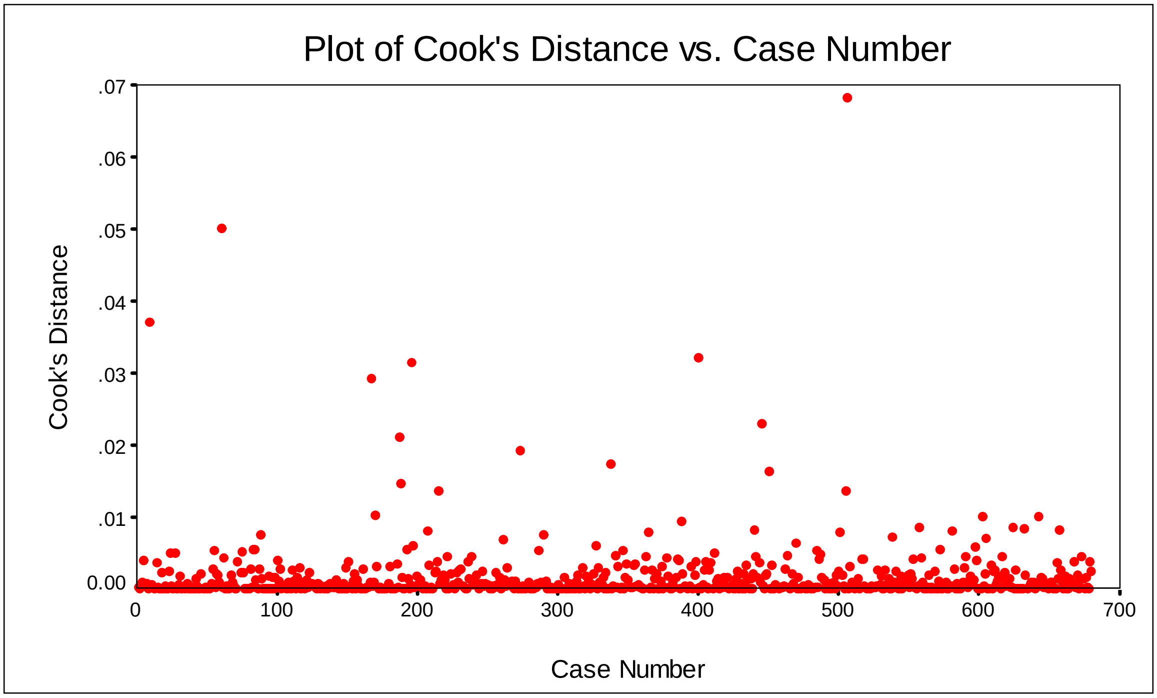
In the child health problem, the threshold leverage value is (3*6)/680=0.026. There are several cases with the leverage exceeding the threshold value, but two of them are obviously substantially larger than the rest: .09537 (case 167) and .05693 (case 9).



In order to get some overall assessment of influence and see whether the above cases 167 and 9 are indeed influential, we will look at some other case influence statistics. One of these statistics is the Cook's distance.

Cook's Distance measures overall influence of a single case on the estimated regression coefficients when the case is deleted from the estimation process. Large values (usually greater than 1) indicate substantial influence by the case in affecting the estimated regression coefficients. However, even if no observations exceed this threshold, additional attention is dictated if a small set of observations has substantially higher values than the rest.

The following plot is the plot of Cook's distances against case number for the data.



Although the value of Cook's Distance for case 506 is only equal to 0.06830, which is considerably smaller than 1, it is obviously substantially larger than the rest. Therefore, it is worthy to rerun the regression without the case to see its influence on the regression results.

The case 506 is a case of a heavy smoker with the pre-pregnancy weight of 200 pounds who gave birth to a child with extremely low birth weight of 4.5 pounds. Running multiple regression without the case changes slightly the coefficient of determination (it is equal to .26140 without the case and .26520 for all observations), the residual sum of squares (588.63924 without the case and 598.41303 for all observations), and the coefficients of the regression equation. The regression equation without the case is

$$\mu\{BWT\} = .236 * GESTWKS - .013 * MNOCIG + .036 * MHEIGHT + .01 * MPPWT +$$
$$.040 * FHEIGHT - 7.979.$$