

SEX DISCRIMINATION PROBLEM

9. Diagnostics

In this section various diagnostic tools will be used to evaluate the adequacy of the regression model 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 the Log-Transformed Salaries

9.2 Checking Constant Variance of the Error Assumption

9.3 Checking the Normality of the Error Assumption

9.4 Multicollinearity of the Independent Variables

9.5 Diagnostics for Outliers and Influential Cases

9.1 Checking the Linearity of the Log-Transformed Salaries

In the previous section we have described the relationship between the response variable *LNBSAL* and the predictors *EDUC*, *SENIOR*, *TREXP*, and *FSEX* by the following multiple linear regression model:

$$LNBSAL = \beta_0 + \beta_1 * EDUC + \beta_2 * SENIOR + \beta_3 * TREXP + \beta_4 * FSEX + ERROR.$$

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 *LNBSAL* under a range of settings of the independent variables *EDUC*, *SENIOR*, *TREXP*, and *FSEX*.

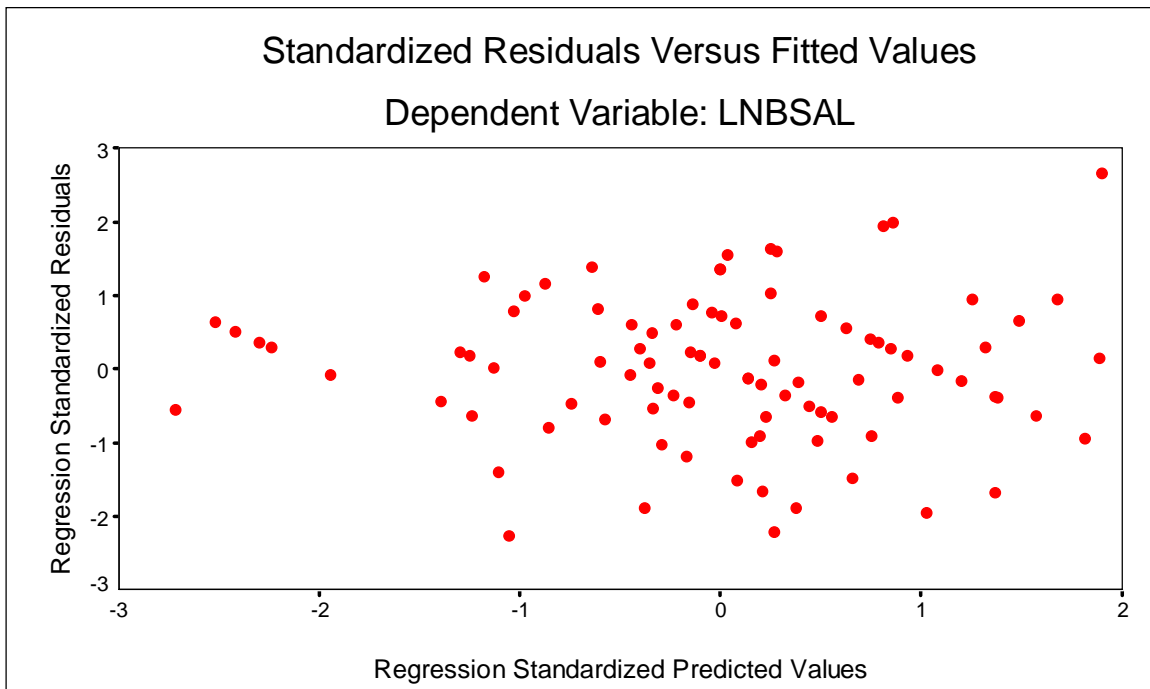
The above model is based on the assumption of a linear relationship between the log-transformed salaries and the predictor variables. The assumption of linearity is easily examined through plots of residuals (standardized residuals) against the predictor values. If the linearity assumption appears to be met, then these residual plots should exhibit a random scatter of points. Any consistent curvilinear pattern in the residuals indicates a nonlinear relationship between the response variable and the predictor 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 predictor variables. 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.

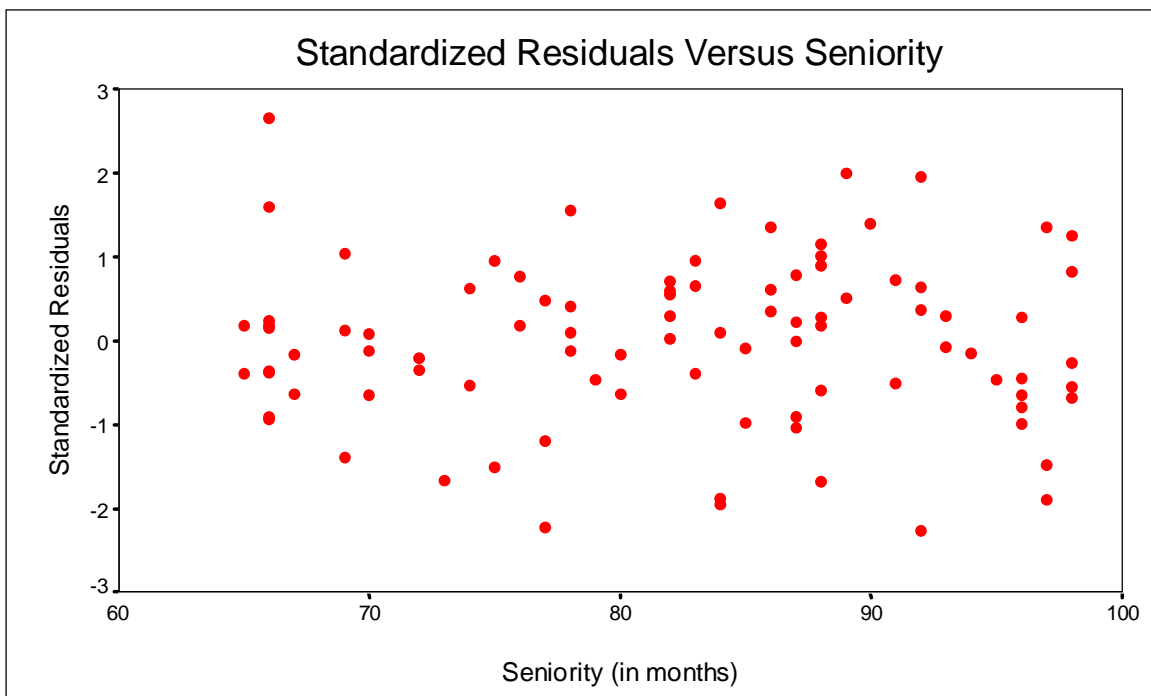
9.2 Checking Constant Variance Assumption

In order to see whether the assumption of constant variance is not violated, we plot residuals (standardized residuals) against the fitted and predictor values. 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 predictor 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. There are no cases with significantly higher residuals that can be classified as outliers.

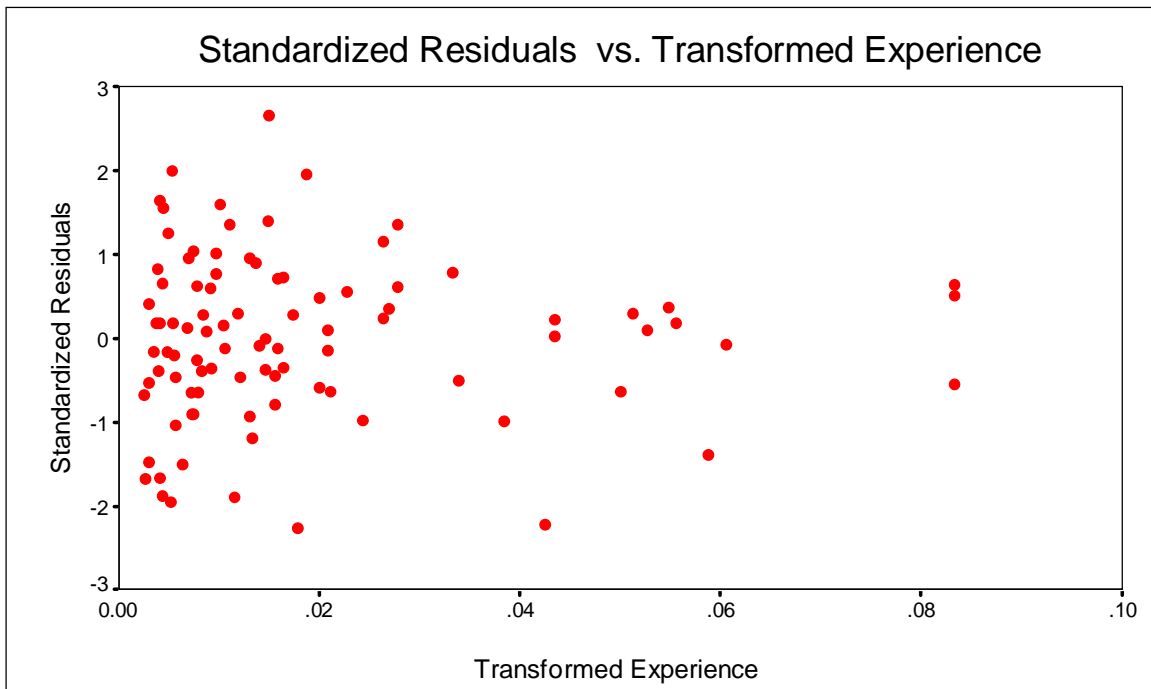


The next plot gives a similar display of standardized residuals plotted against seniority.



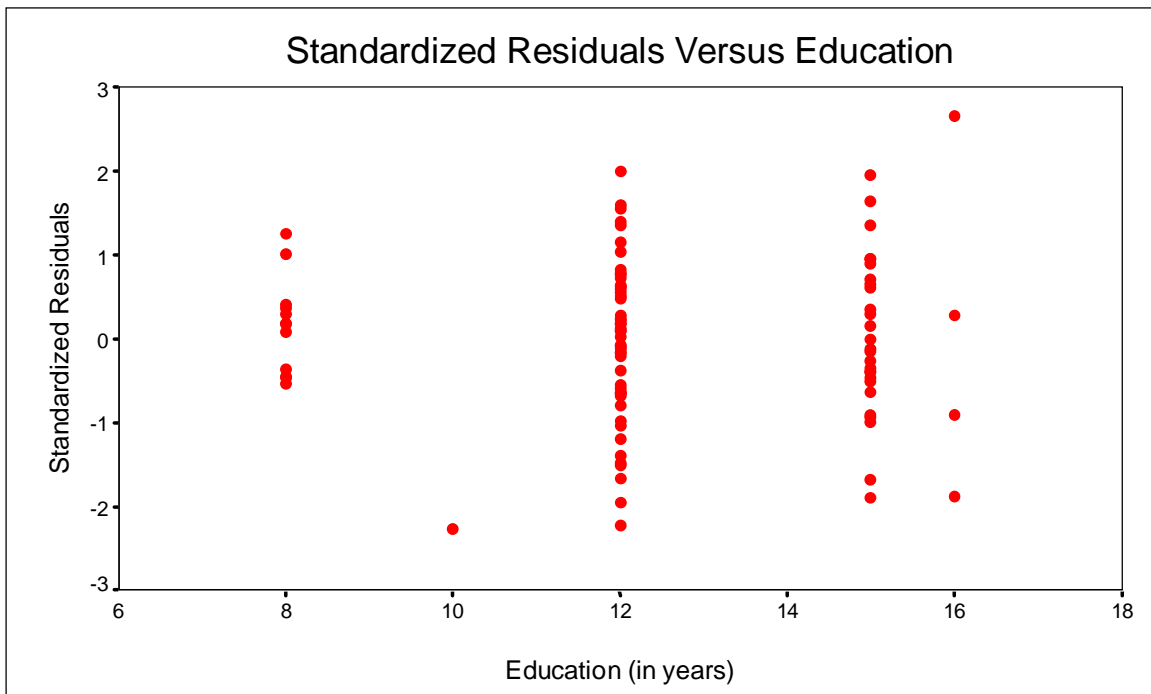
The points in the plot are randomly scattered and constant variability across all values for the seniority predictor is supported. The random pattern in the plot does not provide any evidence to question the linearity assumption.

The next plot gives a similar plot for the transformed experience predictor.



At the first glance, it seems that the variability decreases as the transformed experience 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 education level.

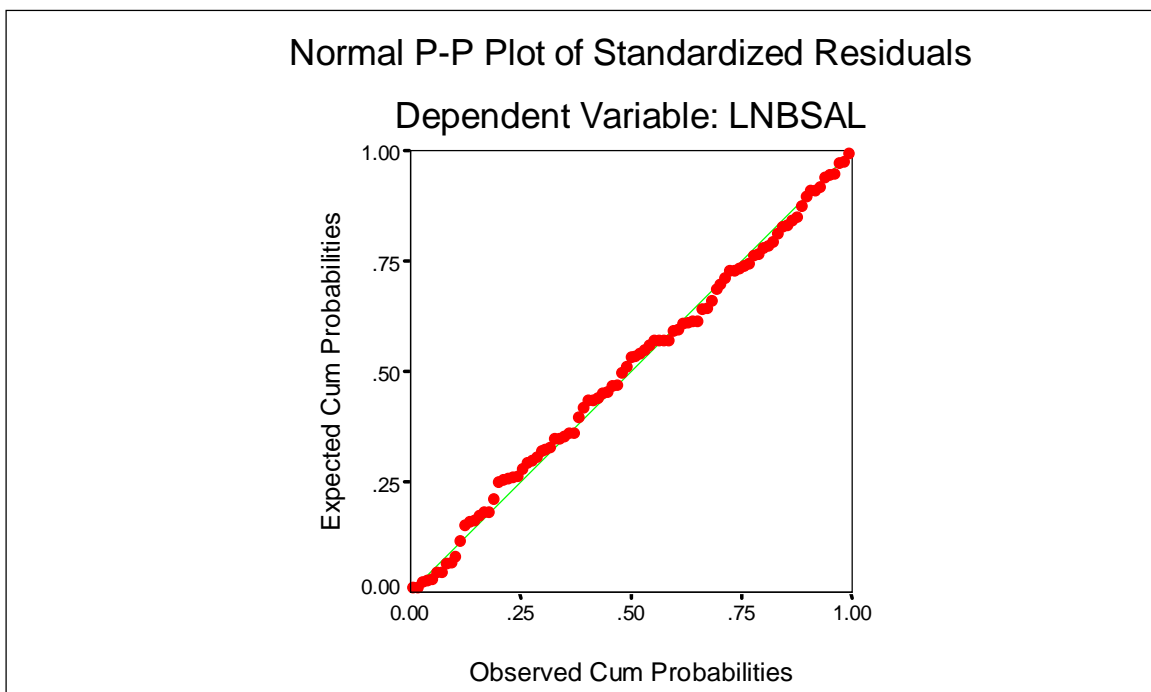


There appear to be some differences of variability in residuals for different levels of education, but the effect is not severe. There is not strong enough evidence to question the assumption of equal variance.

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 close to a straight line, the normality assumption is 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 sex discrimination 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 sex discrimination data does not reveal any high correlations between two predictor variables.

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 sex discrimination problem.

Variables in the Equation							
Variable	B	SE B	Beta	Tolerance	VIF	T	Sig T
EDUC	.013759	.003920	.243035	.888564	1.125	3.510	.0007
SENIOR	-.003373	.000850	-.267687	.936504	1.068	-3.969	.0001
TRANEXP	-2.705475	.465684	-.391774	.937007	1.067	-5.810	.0000
FSEX	-.123657	.018817	-.457104	.880709	1.135	-6.572	.0000
(Constant)	8.826894	.087202				101.224	.0000

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 sex discrimination 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 sex discrimination problem is displayed below:

Collinearity Diagnostics							
Number	Eigenval	Cond Index	Variance Proportions				
			Constant	EDUC	SENIOR	TREXP	FSEX
1	4.24658	1.000	.00050	.00147	.00076	.01651	.01251
2	.42607	3.157	.00069	.00239	.00063	.91359	.05433
3	.30055	3.759	.00159	.01255	.00303	.01610	.74678
4	.02074	14.309	.02319	.77460	.25563	.03915	.07976
5	.00605	26.487	.97403	.20899	.73996	.01466	.10662

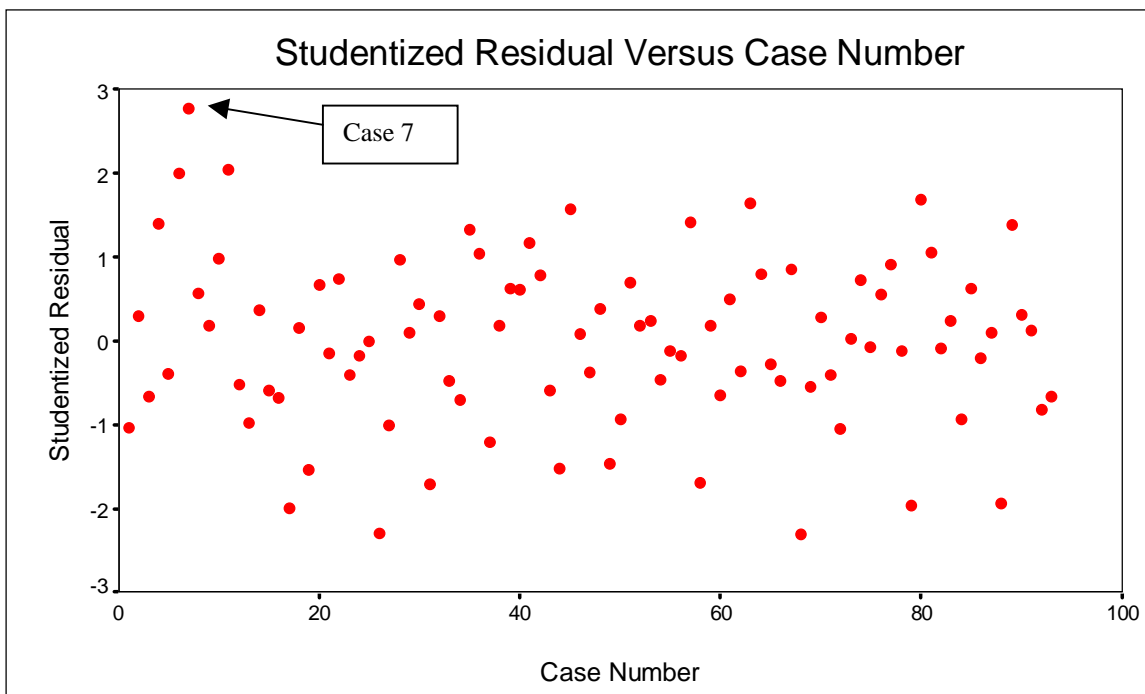
As there are no condition indices exceeding 30, there is no evidence of a collinearity problem for the data.

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.

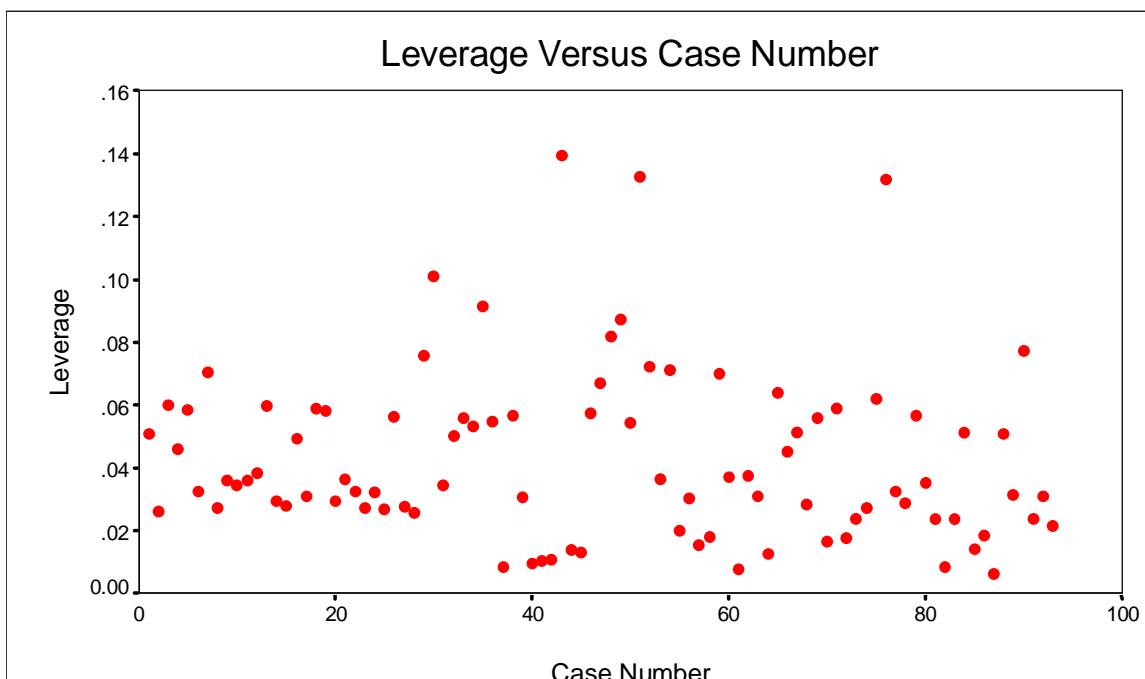
The following plot is a plot of studentized residuals versus case number for the sex discrimination data. As you can see, there are four cases with the standardized residual smaller than -2 or larger than 2: cases 7, 11, 26, and 68. Case 7 produces the highest studentized residual with its value close to 3



Nevertheless, the plot of the standardized residuals versus the fitted values for the sex discrimination problems displayed in **Section 9.2** does not reveal any cases with significantly higher residuals that can be classified as outliers.

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 coefficients. 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.

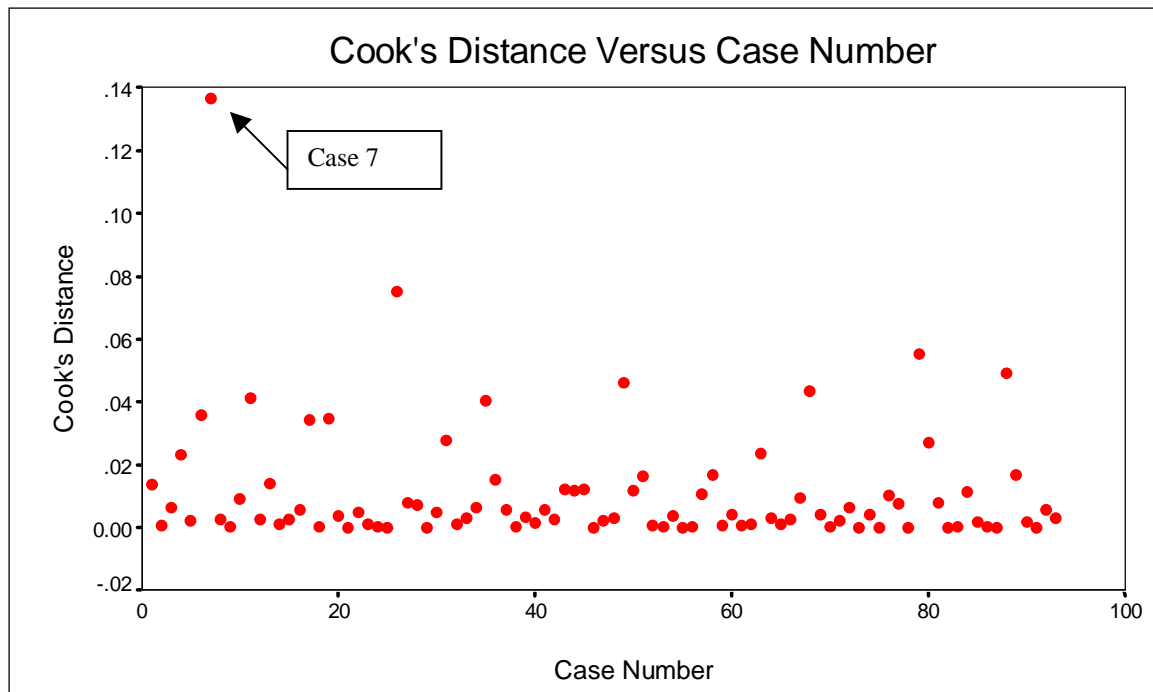
In the sex discrimination problem, the threshold leverage value is $(3*5)/93=0.16129$. The plot of leverage versus case number displayed below does not reveal any cases exceeding the threshold value.



The leverage measure is used in conjunction with the other case influence statistics to get some overall assessment of influence. 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 all of the remaining observations.

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



Although the value of Cook's Distance for case 7 is only equal to 0.137, 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.