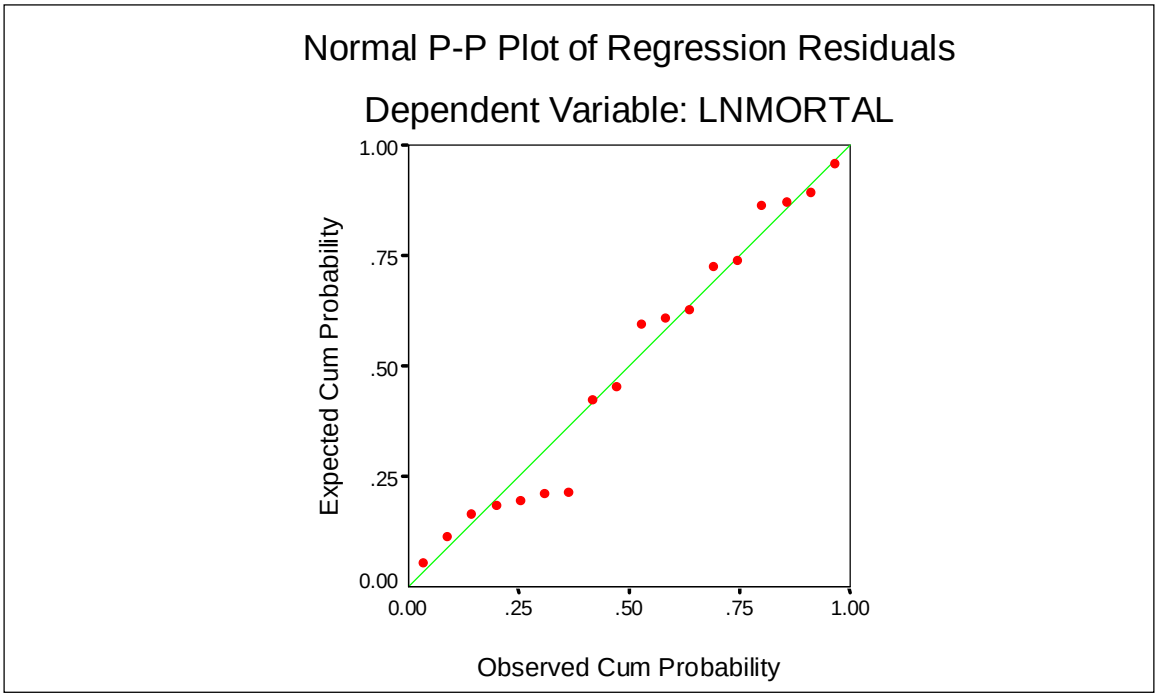
6. Checking the Assumptions of Simple Linear Regression Model

In order to describe the relationship between heart disease mortality and wine consumption, we have used the simple regression model for the log-transformed data. However, the conclusions based on the model are valid only if the underlying assumptions are satisfied. In this section we will check the assumptions.

6.1 Checking the Normality Assumption

Estimates of the coefficients and their standard errors are robust to nonnormal distributions. The consequences of violating this assumption are usually minor for the tests and confidence intervals. However, if prediction intervals are used, departures from normality become important. This is because the prediction intervals are based directly on the normality of the population distributions whereas tests and confidence intervals are based on the sampling distributions of the estimates which may be approximately normal even when the population distributions are not.

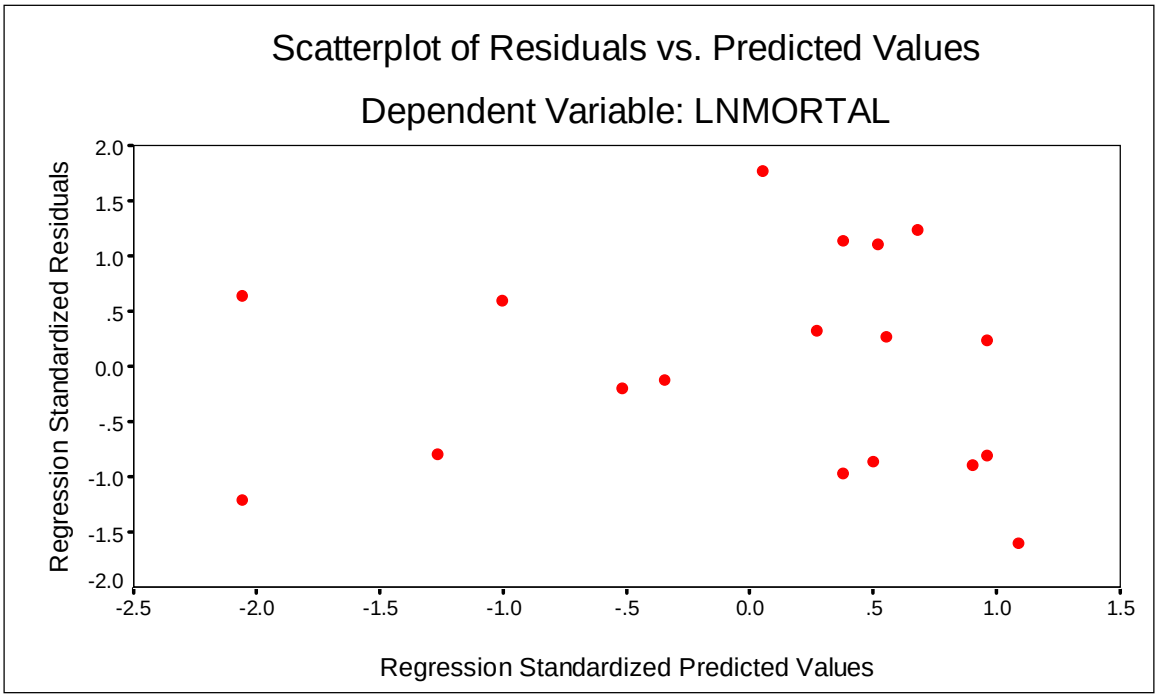
In order to assess whether the assumption is not violated with SPSS, the normal P-P plot of regression standardized residuals is obtained. The plot plots the cumulative proportions of standardized residuals against the cumulative proportions of the normal distribution. If the normality assumption is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



As you can see, the above plot supports the normality assumption. There is some slight departure from a straight-line pattern, but overall the plot is close enough to a straight line.

6.2 Checking the Constant Variance Assumption

It is assumed that mortality (log scale) is normally distributed with equal variance at each value of the independent variable (wine consumption on log scale). One method of checking whether the assumption of constant variance is not violated is to plot the residuals against the predicted values. We then look for a change in the spread or dispersion of the plotted points.



The relatively small number of observations (18) makes the validation of the assumption difficult. Nevertheless, the spread of residuals appears to be approximately the same over the whole range of standardized predicted values. It is not very likely that the assumption of constant variance is violated in this case. The plot also indicates a reasonable degree of randomness about the horizontal line at 0.

6.3 Checking the Independence Assumption

Of the all assumptions, independence is the most crucial. Lack of independence causes no bias in estimates of the coefficients, but standard errors are seriously violated. As a consequence the tests and confidence intervals can be effected.

The heart disease mortality rate for each country is obtained independently of the readings for other countries. There is no indication that the assumption of independence might be violated.

6.4 Summary

In the simple linear regression models discussed above, there was a need to transform the original data in order to fit the data to the straight-line regression model. The normal probability plot shows that the assumption of normality is not violated. The plot of standardized residuals versus standardized predicted values supports the assumption of equal spread. There is no sufficient information to indicate that the assumption of independence might be violated. Summarizing, the simple linear regression model fits the transformed data well.