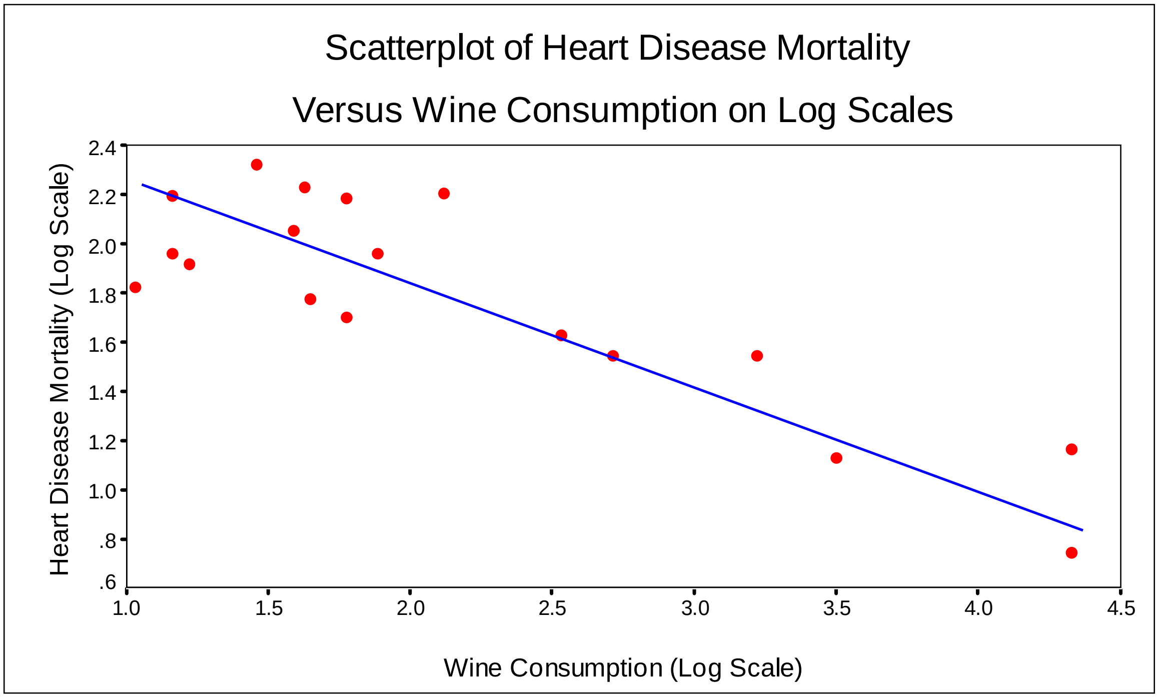
5. Simple Linear Regression Model

In the previous section we obtained a scatterplot of heart disease mortality on the log scale versus wine consumption on the log scale. The plot revealed a linear pattern between the two variables.



Thus the following simple regression models is suitable:

$$LNMORTAL = \beta_0 + \beta_1 * LNWINE + ERROR.$$

Here LNMORTAL is the natural logarithm of MORTAL, and LNWINE is the natural logarithm of WINE. 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 *LNWINE*. The variable *ERROR* follows a normal distribution at each level of LNWINE.

The simple linear regression model can be stated equivalently as follows:

$$\mu\{LNMORTAL \mid LNWINE)\} = \beta_0 + \beta_1 * LNWINE.$$

The above model with *LNWINE* as a predictor is useful only if the slope β_1 is different from zero. The hypothesis that $\beta_1 = 0$ (the model is useful) can be tested using either t or F tests. The F-statistic is the square of the t-statistic and the corresponding p-values of the two tests are identical.

From the above equation, we can obtain

$$MEDIAN\{LNMORTAL \mid LNWINE\} = \exp(\beta_0)(WINE)^{\beta_1}.$$

The SPSS simple linear regression model output for the problem has the following form:

LINEAR REGRESSION

Multiple R	.85932
R Square	.73843
Adjusted R Square	.72209
Standard Error	.22854

Analysis of Variance

	DF	**Sum of Squares**	**Mean Square**
Regression	1	2.35918	2.35918
Residual	16	.83566	0.05223

F = 45.16980 Signif F = .0000

According to the output, the value of the correlation coefficient between the logarithm of heart disease mortality and the logarithm of wine consumption is 0.85932. The value of R^2 (0.73843) says that 73.843% of the variation in the log-mortality was explained by the linear regression on log-wine consumption. The remaining variation was due to some other variables.

We analyze the ANOVA table associated with the simple regression. The sum of squares due to the regression model is reported as 2.35918, and the sum of squares due to error (residual sum of squares) is 0.83566. The residual mean square is an estimate of the variance σ^2 and is equal to 0.05223. Thus an estimate of the residual standard deviation is $\sqrt{0.05223} = 0.228539$.

The value of the F statistic is equal to 45.16980 with the corresponding p-value of 0 provides very strong evidence of the utility of the model.

Now we analyze the part of the output providing the estimates of the regression parameters.

---------------------- **Variables in the Equation** ------------------------------------

Variable	**B**	**SE B**	**95% Confidence Interval B**		**Beta**
LNWINE	-.355596	.052909	-0.467759	-0.243433	-0.859321
(Constant)	2.555552	.126897	2.286542	2.824561	

Variable	**T**	**Sig T**
LNWINE	-6.721	.0000
(Constant)	20.139	.0000

According to the output, the estimated regression line of heart disease mortality (log scale) on wine consumption (log scale) is

$$\mu\{LNMORTAL \mid LNWINE)\} = -0.355596 * LNWINE + 2.555552$$

The association between mortality (log scale) and wine consumption (log scale) is negative and significant (estimate of the slope is -0.355596 with reported p-value of zero).

The estimated regression line was obtained when both the response (*MORTAL*) and explanatory variable (*WINE*) were logged. The log transformation was necessary to fit the data to a straight line model and to make the assumptions of the simple linear regression model satisfied. The validity of the simple linear regression assumptions for the data is discussed in **Section 6**.

Based on the estimated regression line equation, we have

$$MEDIAN\{LNMORTAL \mid LNWINE\} = \exp(2.555552)(WINE)^{-0.355596},$$

or

$$MEDIAN\{LNMORTAL \mid LNWINE\} = 12.87841 * (WINE)^{-0.355596},$$

Thus, a doubling of wine consumption is associated with a $2^{-0.355596} = 0.781568$ fold change in the median of heart disease mortality (22% drop in heart disease mortality).