6. Linear Regression with Two Predictors

Two simple linear regression models were used to analyze the *Child Health and Development* data in *Simple Regression* module. The dependent variable in each case was birth weight (BWT, in pounds) and the independent variables were length of gestation period (GESTWKS, in weeks) and amount of maternal smoking (MNOCIG, reported number of cigarettes smoked per day).

The percentage of variation in birth weight that can be explained by the simple regression on gestation time is 18.1%. On the other hand, the percentage of total variation explained by postulating a simple linear relationship between mother's smoking exposure and her infant's birth weight is only 3.2%. We found that the interrelationship between these two independent variables makes interpretation of the two separate simple linear regression analyses unsatisfactory because neither analysis accounts for the possible influence of the other variable.

Now we will reconsider the problem by using a multiple regression model with the two independent variables. We will demonstrate that multiple regression is able to take into account and assess simultaneous influences of the two independent variables on birth weight.

More precisely, we will examine the relationship between infant birth weight and the two independent variables with the following multiple regression model:

$$BWT = \beta_0 + \beta_1 * GESTWKS + \beta_2 * MNOCIG + ERROR.$$

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*.

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

$$\mu\{BWT\} = \beta_0 + \beta_1 * GESTWKS + \beta_2 * MNOCIG,$$

where the regression coefficients β_1, β_2 are to be estimated with the sample data. The above model with *GESTWKS* and *MNOCIG* as predictors is useful only if at least one slope β_i is different from zero. The hypothesis that the model is useful can be tested using F test.

The analysis of variance table contains the components necessary to assess statistically the worth of the regression model as a tool for describing the data set. The SPSS output is displayed below:

MULTIPLE LINEAR REGRESSION

Multiple R	.45137
R Square	.20374
Adjusted R Square	.20138
Standard Error	.97618

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	2	165.06694	82.53347
Residual	677	645.12859	.95292

F = 86.61089 Signif F = .0000

Thus, 20.4% (R^2=0.20374) of the total variation in birth weight (BWT) can be explained by a linear model with gestational age (GESTWKS) and maternal smoking (MNOCIG) simultaneously as independent variables. That is, the total variation of 810.1955 separates into two pieces with 20.4% (165.06694) associated with the regression sum of squares and the remaining 79.6% (645.12859) associated with the residual sum of squares.

The squared multiple correlation coefficient increases with each variable added to the regression equation. The value of R^2 was slightly smaller (0.181) for a simple linear regression of BWT on GESTWKS (see *Child Development Study* in ***Simple Regression*** module).

The residual mean square is an estimate of the variance σ^2 and is equal to 0.95292.

The value of the F statistic is equal to 86.61089 with the corresponding p-value < 0.0001 indicates the strength of the joint influence of the two independent variables on an infant's birth weight.

The separate effects of the independent variables are inferred from the regression coefficients. 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
GESTWKS	.241850	.020026	.202531	.281170	.415230
MNOCIG	-.014535	.003332	-.021077	-.007994	-.149993
(Constant)	-1.994051	.799441	-3.563733	-.424369	

Variable	Tolerance	VIF	T	Sig T
GESTWKS	.994982	1.005	12.077	.0000
MNOCIG	.994982	1.005	-4.363	.0000
(Constant)			-2.494	.0000

According to the output, the estimated regression line of log beginning salaries on the two predictors is

$$\mu\{BWT \mid GESTWKS, MNOCIG\} = -1.994051 + .241850 * GESTWKS - .014535 * MNOCIG.$$

As we expected gestational time (GESTWKS) shows a positive influence on infant birth weight. Each additional week of gestational age is expected to increase an infant's birth weight an estimated 0.242 pounds or 3.8 ounces. This increase represents the influence on birth weight from gestational time adjusted for the effects of maternal smoking. The unadjusted value obtained with a simple linear regression model (see *Child Development Study* in ***Simple Regression*** module) is slightly larger (0.248). Obviously, the influence of gestation is not adjusted for other associated variables. The effects of the independent variables are adjusted by a regression model for only those variables included in the linear equation.

The association between gestational age and infant birth weight is significant with the value of the t-statistic of 12.077 and the corresponding p-value < 0.0001.

Maternal smoking (MNOCIG) shows a negative influence on infant birth weight, which is again inferred from the regression coefficient. For a constant gestational age, smoking a pack of cigarettes per day (20 cigarettes) reduces an infant's birth weight, on the average, an estimated 20(0.014535)=0.300 lbs. or 4.7 ounces.

The t-statistic serves as a useful way to compare the magnitude of the relative contributions to the variability in birth weight from the two independent variables free from the fact that they are measured in different units (weeks and cigarettes/day). Comparison of the two t ratios (GESTWKS: t= 12.1 and MNOCIG: t= -4.4) shows that maternal smoking exposure has considerably less influence on birth weight than gestational age.

Now we analyze the part of the output displaying collinearity statistics. No VIF value exceeds 10.0, and the tolerance values show that collinearity does not explain more than 1 percent of any independent variable's variance. Thus, we can find no support for the existence of multicollinearity based on the values of VIF and tolerance provided above.

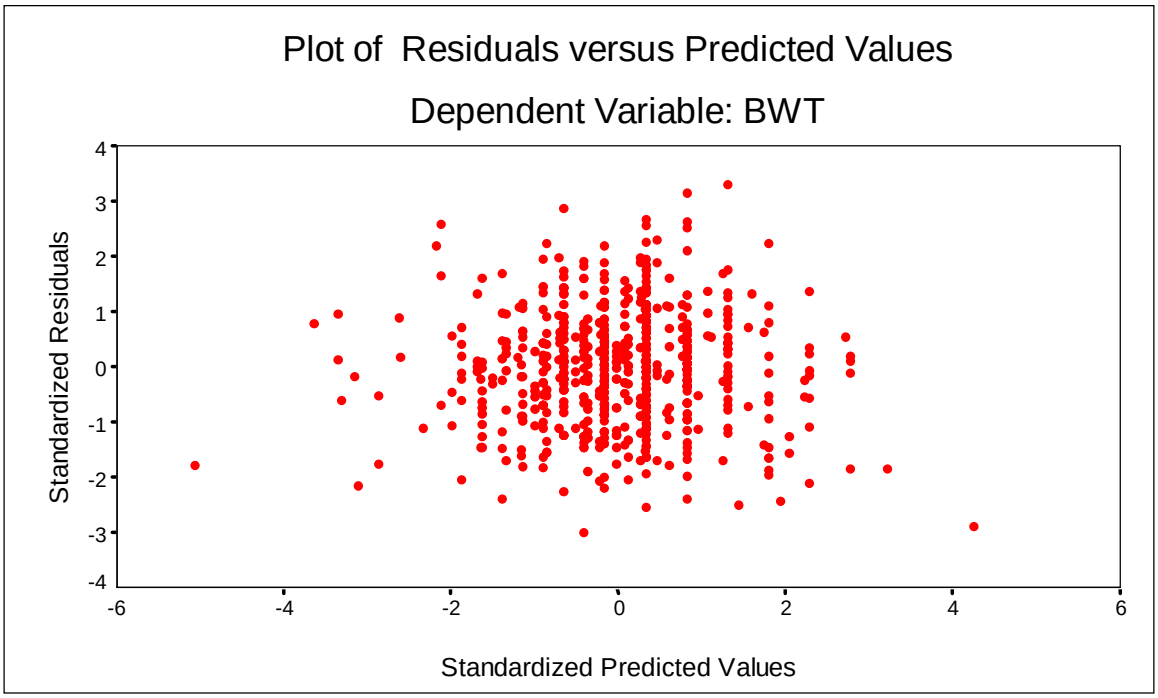
The collinearity statistics for the problem is also provided by the coefficient variance decomposition analysis with condition indices:

Collinearity Diagnostics

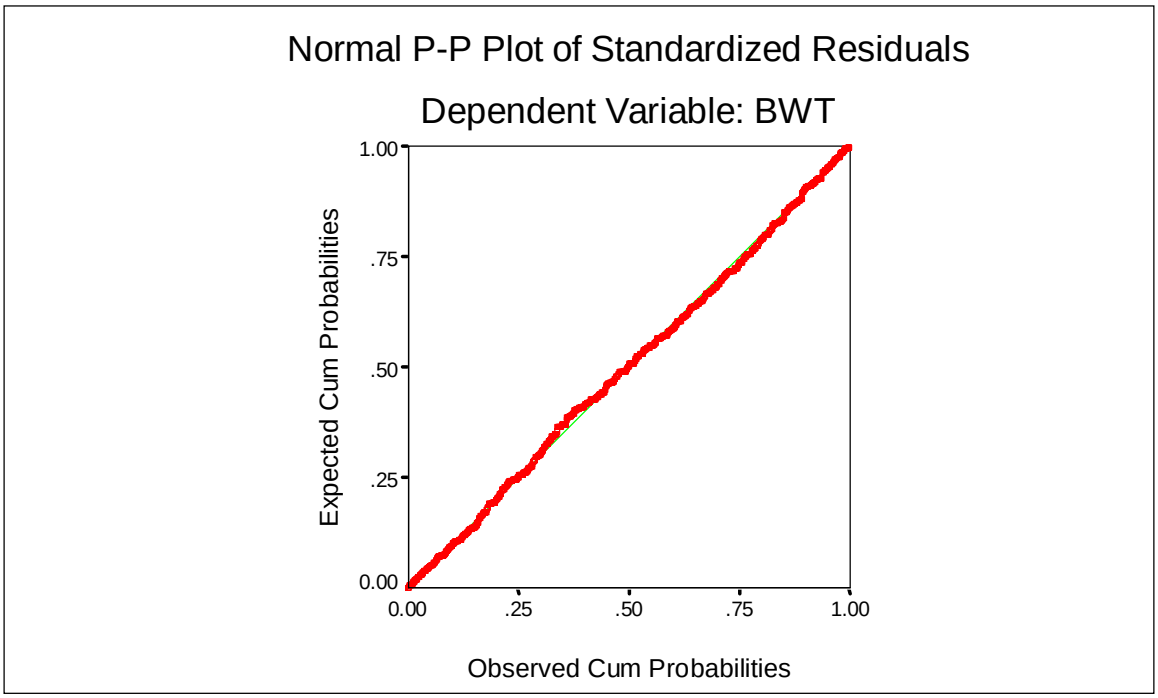
Number	**Eigenval**	**Cond Index**	**Variance Proportions** Constant	GESTWKS	MNOCIG
1	2.42260	1.000	.00035	.00035	.06567
2	.57630	2.050	.00043	.00045	.92686
3	.00110	46.927	**.99922**	**.99920**	.00746

The condition index for the last component exceeds 30 and the last component accounts for over 99.9% of the constant, over 99.9% of GESTWKS, and less than 1% of the variance of MNOCIG. Thus, for a more stable model, it might be wise to explore a model without MNOCIG.

Now we analyze the residuals. 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, and the linearity assumption is not violated. There are no cases with significantly higher residuals that can be classified as outliers.



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 almost lying on a straight line, the normality assumption is strongly supported.