CHILD HEALTH AND DEVELOPMENT STUDY

15. Brief Version of the Case Study

15.1 Problem Formulation
15.2 Study Design
15.3 Displaying and Describing the Data
15.4 Multiple Linear Regression Model
15.5 Basic Diagnostics
15.6 Summary

15.1 Problem Formulation

The data in the case study were collected for 680 live-born, white, male infants born to members of the Kaiser Foundation Health Plan who reside in the San Francisco-East Bay area. We will use the parental observations as independent variables, and assess their relationship to the characteristics of the infants (principally birth weight) with multiple linear regression. In particular, we will determine the influence of parental observations on infant birth weight and estimate the relative impact of maternal and paternal variables on birth weight.

The data from the study are available in the SPSS file *child.sav* located in the STAT 252 directory on the FTP server.

There are 12 variables in the data file describing each infant-mother-father set. They are organized in three blocks: *Infant Measurements*, *Maternal Measurements*, and *Paternal Measurements*. The following is a description of the variables in the data file:

	COLUMN	VARIABLE CODE	VARIABLE DESCRIPTION
INFANT MEASUREMENTS	1	ID	ID Number
	2	HEADCIR	Head circumference (inches)
	3	LENGTH	Length (inches)
	4	BWT	Birth Weight (pounds)
MATERNAL MEASUREMENTS	5	GESTWKS	Gestation (weeks)
	6	MAGE	Maternal Age (years)
	7	MNOCIG	Cigarettes (number smoked/day)
	8	MHEIGHT	Maternal Height (inches)
	9	MPPWT	Pre-pregnancy Weight (pounds)
PATERNAL MEASUREMENTS	10	FAGE	Father's Age
	11	FEDYRS	Father's education (years)
	12	FNOCIG	Cigarettes (number smoked/day)
	13	FHEIGHT	Father's Height (inches)

In this module we will consider a set of nine predictor variables: GESTWKS, MNOCIG, MAGE, MHEIGHT, MPPWT, FAGE, FEDYRS, FNOCIG, FHEIGHT. We will apply a multiple linear regression model to examine the relationship between infant birth weight and the predictors in the following form:

$$BWT = \beta_0 + \beta_1 * GESTWKS + \beta_2 * MAGE + \beta_3 * MNOCIG + \beta_4 * MHEIGHT +$$
$$\beta_5 * MPPWT + \beta_6 * FAGE + \beta_7 * FEDYRS + \beta_8 * FNOCIG + \beta_9 * FHEIGHT + ERROR.$$

The child health and development data were already examined in *Simple Regression* module in *STAT 252 Laboratories* Web site. We used simple linear regression to develop two alternative models describing the relationship between birth weight of male infants and each of the two independent variables: length of gestation period and amount of maternal smoking. We found that the interrelationship between gestation time and maternal smoking makes interpretation of the two separate simple linear regression analyses unsatisfactory because neither analysis accounts for the possible influence of the other variable. With multiple regression, the obstacle will be removed.

The results of the study are described by J. Yerushalmy in the report "The California Child Health and Development Studies-Study Design, and Some Illustrative Findings on Congenital Heart Disease" published in *Congenital Malformations, Proceedings of The Third International Conference, pp. 299-306, International Congress Series No.204, New York, 1970.*

15.2 Study Design

Live-born, white, male infants make up a study sample of 680 observations. The 680 infants were not selected randomly from any well-defined population. Therefore, the observed pattern cannot be inferred to hold in some general population, for example the population of white, male infants born in California, unless we assume that the infants are representative of the population.

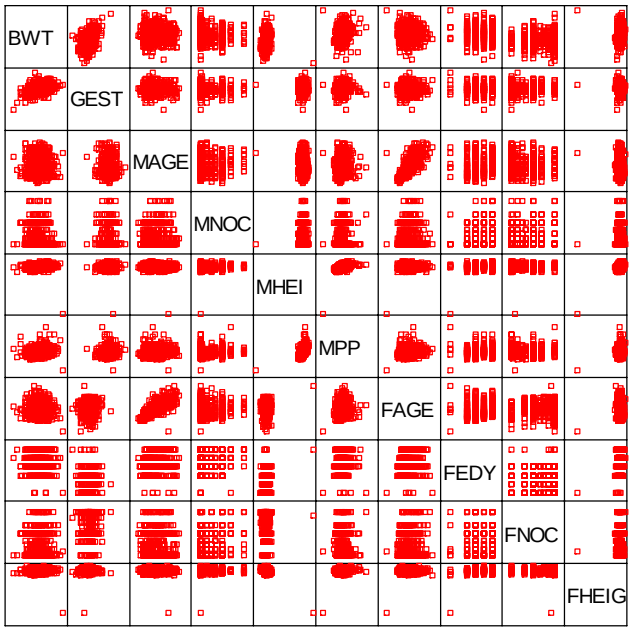
What can then be inferred? As there is no random sample, the statistical results apply only to the infants actually measured. Any extrapolation of the pattern to other children comes from the assumption that the relationship between birth weight and gestation time (amount of maternal smoking) is similar for others. This is not necessarily a bad assumption. The point is that extending the inference to other infants is surely open to question.

The status of each infant (born to whom) is established beyond the control of the investigator. Thus the study is an observational study. That means that we cannot draw any causal conclusions from the statistical analysis alone. One cannot rule out the possibility that confounding variables are responsible for the differences in birth weights. For example, the result - that the infants born to smoking mothers tended to have lower birth weight relative to non-smoking mothers -does not prove that being born to a smoker is responsible for the difference. Diets during pregnancy, for example, may differ in the two groups and may be responsible for the different weights.

Nevertheless the study may be very useful to examine the factors affecting birth weight and to establish cause-and-effect relationship based on theory.

15.3 Displaying and Describing the Data

The following scatterplot matrix allows you to visualize the relationships between any two variables in the data:



In the scatterplot matrix, for each plot, the names of variables are indicated in the corresponding row and the column of the matrix. For example, the middle plot in the first column of the matrix is the scatterplot of birth weight (BWT) versus gestation time (GESTWKS).

The first row of the matrix displays the relationship between the dependent variable (birth weight) and each of the nine independent variables. The degree of linearity of the relationship between birth weight and each of the nine independent predictors varies from very strong for GESTWKS (gestation time) to very weak for mother's age or father's age. The matrix also reveals high positive correlation between father's age and mother's age.

The correlation matrix for the variables in the study is displayed below:

	BW	GES	MNO	MAG	MHE	MPP	FNO	FAG	FHEI	FED
BWT	**1**	**.426**	**-.179**	**.0013**	**.2025**	**.2216**	**-.023**	**.017**	**.154**	**.033**
GES	**.426**	**1**	**-.071**	**.003**	**.048**	**.052**	**-.003**	**.042**	**.024**	**.035**
MNO	**-.179**	**-.071**	**1**	**.045**	**.026**	**-.026**	**.262**	**.028**	**.011**	**.024**
MAG	**.001**	**.003**	**.045**	**1**	**.018**	**.116**	**.017**	**.817**	**-.071**	**.241**
MHE	**.202**	**.048**	**.026**	**.018**	**1**	**.494**	**-.015**	**.018**	**.303**	**.108**
MPP	**.222**	**.052**	**-.026**	**.116**	**.494**	**1**	**-.028**	**.124**	**.166**	**.001**
FNO	**-.023**	**-.003**	**.262**	**.017**	**-.015**	**-.028**	**1**	**.040**	**.014**	**-.182**
FAG	**.017**	**.042**	**.028**	**.817**	**.018**	**.124**	**.040**	**1**	**-.134**	**.220**
FHEI	**.154**	**.0240**	**.0108**	**-.071**	**.303**	**.166**	**.014**	**-.134**	**1**	**.108**
FED	**.033**	**.035**	**.024**	**.241**	**.108**	**.001**	**-.182**	**.220**	**.108**	**1**

The heighest correlation (r=0.817) is, not surprisingly, between age of mother (MAGE) and age of father (FAGE). The first row of the table shows that the maternal variables are stronger correlated to infant birth weight (BWT) than the paternal variables. The second column of the table indicates that length of gestation (GESTWKS) is essentially uncorrelated with the parental variables (the correlation $| r | < 0.071$) and is strongly correlated with birth weight (r=0.426).

There is a negative moderate correlation between number of cigarettes smoked per day by the mother (MNOCIG) and birth weight (r=-.179). However, there is a very weak negative correlation between infant birth weight and the number of cigarettes smoked by father (FNOCIG, r=-.023). There is a very weak negative correlation between number of cigarettes smoked and gestation time (r=-.071).

A number of variables show rather expected associations-variables related to adult size. Mother's pre-pregnancy weight (MPPWT), mother's height (MHEIGHT), and father's height (FHEIGHT) are all moderately correlated with each other (the correlation coefficient $r > 0.166$. Also maternal (MNOCIG) and paternal (FNOCIG) smoking habits are correlated (r=0.262).

It is important to keep in mind that correlation coefficients only indicate pairwise linear associations when typically more complicated relationships exist.

15.4 Multiple Linear Regression Model

In this section we will examine the relationship between infant birth weight (BWT) and the nine independent variables with the following multiple regression model:

$$BWT = \beta_0 + \beta_1 * GESTWKS + \beta_2 * MNOCIG + \beta_3 * MAGE + \beta_4 * MHEIGHT + \beta_5 * MPPWT + \beta_6 * FNOCIG + \beta_7 * FAGE + \beta_8 * FHEIGHT + \beta_9 * FEDYRS + ERROR.$$

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 nine independent variables *GESTWKS*, *MNOCIG*, *MAGE*, *MHEIGHT*, *MPPWT*, *FNOCIG*, *FAGE*, *FHEIGHT*, and *FEDYRS*.

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

$$\mu\{BWT\} = \beta_0 + \beta_1 * GESTWKS + \beta_2 * MNOCIG + \beta_3 * MAGE + \beta_4 * MHEIGHT + \beta_5 * MPPWT + \beta_6 * FNOCIG + \beta_7 * FAGE + \beta_8 * FHEIGHT + \beta_9 * FEDYRS.$$

The following table displays the initial regression results for this data set.

MULTIPLE LINEAR REGRESSION

Multiple R	.51180
R Square	.26194
Adjusted R Square	.25203
Standard Error	.94472

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	9	212.22479	23.58053
Residual	670	597.97074	.89249

F = 26.42095 Signif F = .0000

The squared multiple correlation coefficient R^2 (0.26194) says that a significant portion (over 26 %) of the variation in infant birth weights is explained by these nine predictors. The adjusted squared multiple correlation coefficient (0.25203) is not very different from the unadjusted value of 0.26194 because the number of independent variables (k=9) is much less than the number of observations (n=680).

We analyze the ANOVA table associated with the multiple linear regression. The sum of squares due to the regression model is reported as 212.22479, and the sum of squares due to error (residual sum of squares) is 597.97074. The residual mean square is an estimate of the variance σ^2 and is equal to 0.89249.

The value of the F statistic is equal to 26.42095 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	Beta	VIF	T	Sig T
GESTWKS	.233447	.019471	.400802	1.014	11.989	.0000
MNOCIG	-.015221	.003357	-.157066	1.089	-4.534	.0000
MAGE	-.001308	.011634	-.006544	3.073	-.112	.9105
MHEIGHT	.039756	.017511	.090378	1.439	2.270	.0235
MPPWT	.008421	.002374	.137819	1.370	3.548	.0004
FNOCIG	.001864	.002718	.024178	1.129	.686	.4931
FAGE	.000066	.010447	.000373	3.123	.006	.9949
FHEIGHT	.039018	.014742	.094239	1.151	2.647	.0083
FEDYRS	.004157	.017668	.008383	1.152	.235	.8140
(Constant)	-8.090991	1.43988			-5.619	.0000

According to the output, the estimated regression line of birth weights on the nine predictors is

$$\mu\{BWT\} = .233 * GESTWKS - .015 * MNOCIG - .001 * MAGE + .040 * MHEIGHT +$$
$$.008 * MPPWT + .002 * FNOCIG + .00007 * FAGE + .039 * FHEIGHT + .004 * FEDYRS - 8.091.$$

The comparison of the nine independent variables by means of individual t-statistics indicates the relative magnitude of the unique contribution of each variable to the overall variability in birth weight. According to the above SPSS output, gestation time is the largest contributor to the explained variation in birth weight. The regression coefficient associated with gestation time is .233447 with a corresponding t ratio of 11.989, indicating a very strong effect of gestation time on infant birth weight after accounting for the effect of maternal and paternal variables.

Maternal smoking, maternal pre-pregnancy weight, paternal height, and maternal height are the next four most important contributors. The remaining four variables (MAGE, FAGE, FNOCIG, and FEDYRS) have significance probabilities greater than .49, and therefore can be considered nonsignificant contributors.

The three variable selection techniques (forward selection, backward elimination, and stepwise regression) available in SPSS applied to the data produced the following regression model with five independent variables GESTWKS, MNOCIG, MPPWT, FHEIGHT, and MHEIGHT:

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

The SPSS output for the model is displayed below:

MULTIPLE LINEAR REGRESSION

Multiple R	.51127
R Square	.26140
Adjusted R Square	.25592
Standard Error	.94226

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	5	211.78250	42.35650
Residual	674	598.41303	.88785

F = 47.70665 Signif F = .0000

The squared multiple correlation coefficient R^2 (0.26140) is now slightly smaller than the value obtained for the full model (0.26194). The linear regression of birth weight on the five predictors explains almost the same portion (over 26 %) of the variation in infant birth weights as the model with the nine variables. The value of the F statistic equal to 47.70665 is now much larger than the value of 26.42095 obtained for the full model.

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

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

Variable	B	SE B	Beta	VIF	T	Sig T
GESTWKS	.233851	.019363	.401496	1.008	12.077	.0000
MNOCIG	-.014619	.003220	-.150857	1.008	-4.540	.0000
MHEIGHT	.040042	.017358	.091028	1.421	2.307	.0214
MPPWT	.008308	.002330	.135979	1.327	3.565	.0004
FHEIGHT	.039711	.014387	.095913	1.102	2.760	.0059
(Constant)	-8.114024	1.40914			-5.758	.0000

According to the above output, the estimated regression line of birth weights on the five predictors is

$$\mu\{BWT\} = .234 * GESTWKS - .015 * MNOCIG + .040 * MHEIGHT + .008 * MPPWT + .040 * FHEIGHT - 8.114.$$

All independent variables in the model are significant with the p-value <= .0214. The regression diagnostics for the model is discussed in **Section 15.5.**

15.5 Basic Diagnostics

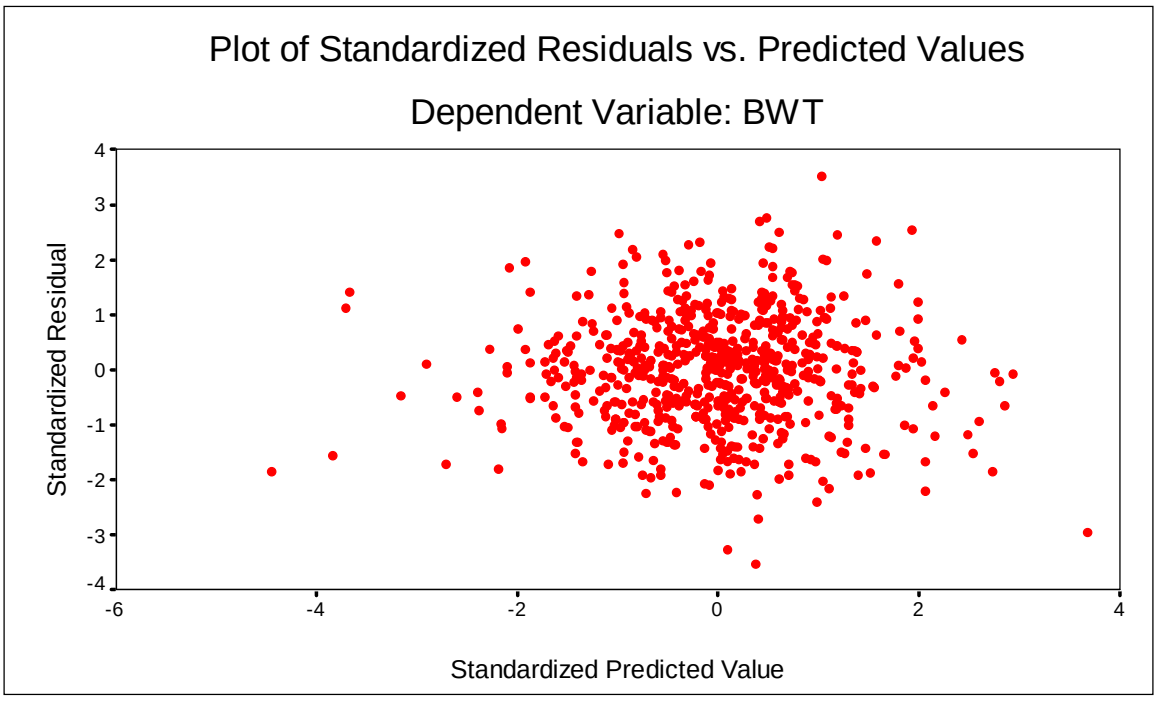
In the previous section we have described the relationship between the response variable BWT and the predictors GESTWKS, MNOCIG, MHEIGHT, MPPWT, and FHEIGHT by the following multiple linear regression model:

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

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*.

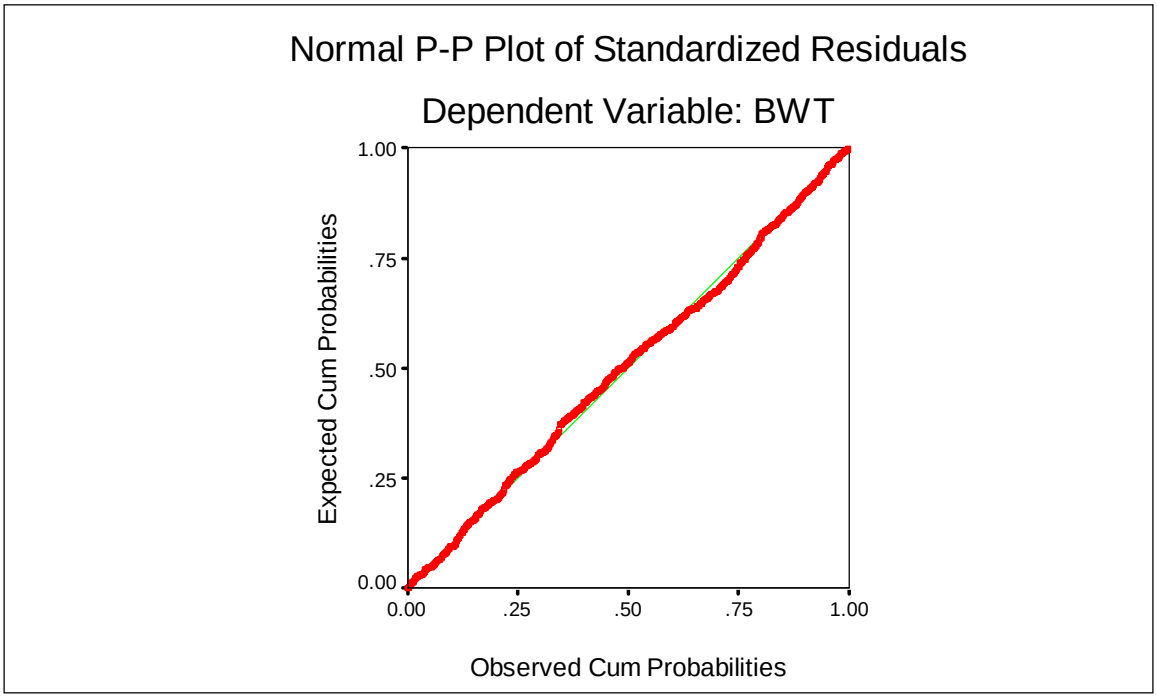
In order to see whether the assumption of linearity and constant variance are not violated, we plot residuals (standardized residuals) against the fitted and also against each predictor variable. If the assumptions of linearity and constant variance appear to be met, then these residual plots should 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 plots of standardized residuals against each of the nine independent variables do not provide any evidence to question the linearity and constant variance assumptions.

In order to assess whether the normality 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.

The multicollinearity and diagnostics for outliers and influential cases for the data is studied in **Section 9.4** and **Section 9.5**, respectively.

15.6 Summary

The above model showed a very strong effect of gestation time on infant birth weight after accounting for the effect of maternal and paternal variables. Maternal smoking (MNOCIG), maternal pre-pregnancy weight (MPPWT), paternal height (FHEIGHT), and maternal height (MHEIGHT) are the next four most important contributors. However, the remaining four variables: maternal age (MAGE), paternal age (FAGE), paternal smoking (FNOCIG), and paternal education (FEDYRS) were found to be nonsignificant contributors.

In **Section 10** we compared the relative influence on birth weight from maternal variables compared to the paternal variables. We found considerably stronger influence of the maternal variables on an infant's birth weight.