

## CHAPTER ONE

# Review of Ordinary Linear Regression and Its Assumptions

### 1.1 THE ORDINARY LINEAR REGRESSION EQUATION AND ITS ASSUMPTIONS

A linear regression equation can be alternatively specified as

$$\begin{aligned}y_i &= \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{or} \\ \mu_{y|x} &= \beta_0 + \beta_1 x \quad \text{or} \\ E(y|x) &= \beta_0 + \beta_1 x\end{aligned}\tag{1.1}$$

to describe the quantitative relationship between a single predictor  $x$  and an outcome  $y$ . In the population health research projects described in the Introduction,  $y$  may be a measured GHb or the score of a very-low-birth-weight (VLBW) child on a test, or the systolic blood pressure (SBP) at a visit to the sleep clinic. In the first equation  $\epsilon_i$  is a random regression error describing the deviation of a given value  $y_i$  from its mean. It can be viewed as capturing unmeasured influence on the outcome. In order to make both the first and the second equations of (1.1) correct, it is assumed that  $E(\epsilon_i|x_i) = 0$ . In other words, if the second equation is to describe the relationship of the mean  $y$  to  $x$  correctly, the random errors in the first equation must average to 0 for all  $x$ . This also implies that  $\epsilon_i$  does not depend on  $x_i$ . The last two equations are just saying the same thing in different notation because the “expected value”  $E(\cdot)$  of a variable is by definition the mean of that variable.

We assume that the reader is familiar with the “conditional on” notation implied by the “|”. Conditioning on a variable means that the variable is (at that moment) considered a constant, so the parameters of the distribution of  $y$  may depend on  $x$ . In other words, when conditioning systolic blood pressure on a given age  $x$ , we are interested in the parameters of the distribution of blood pressure at that age. Estimation of the parameters of equations (1.1) usually proceeds by the method

of least squares. In dealing with the regression equation, forming estimators, and drawing inference, we commonly make a number of assumptions:

### 1.1.1 Straight-Line Relationship

Equation (1.1) implies that  $x$  and the mean of  $y$  are related in a straight-line fashion. This assumption can be alternatively stated as a constant difference in mean  $y$  between every pair of  $x$ 's that are separated by the same number of steps. For example, if  $y$  is systolic blood pressure from visit 1 in the Sleep Cohort Study and  $x$  is age, linearity implies that the difference in mean blood pressure between a 50-year-old and a 40-year-old is the same as that between a 40-year-old and a 30-year-old. Regardless of the level of  $x$ ,  $\mu_{y|x+1} - \mu_{y|x} = \beta_1$ , so that the regression coefficient is the difference in mean with one step increase in  $x$ . Again, if  $y$  is systolic blood pressure and age  $x$  is recorded in years,  $\beta_1$  is the increase in mean blood pressure every year. The linearity assumption is an inherent structural assumption, the validity of which is driven by the biological, sociological, and so on, mechanisms that relate  $y$  to  $x$ . When the linearity assumption holds, we are ahead statistically, because we need to estimate only two parameters  $\beta_0$  and  $\beta_1$  instead of a separate  $\mu_{y|x}$  for every  $x$ .

Only in the situation that  $x$  is binary (e.g., designating two treatment groups) is the linearity assumption moot, or automatically satisfied. If  $y$  is systolic blood pressure and  $x$  is a 0–1 indicator of gender where 1 indicates male, then  $\beta_1$  is the difference in mean blood pressure between males and females, and  $\beta_0$  is the mean for females. In this situation,  $\mu_{y|x}$  is simply a notation for representing the means of two groups (females and one-step difference involved). Since no assumptions are made on the mean structure, equations (1.1) estimate two parameters either way.

In other situations, the original  $x$  may just serve as a label for different groups, such as ethnic categories or treatments. The linearity assumption then makes little sense. However, we can expand (1.1) through the device of binary indicator variables, which bypass the linearity assumption, but again do not save us parameters as compared to estimating  $\mu_{y|x}$  separately for each group. In the Wisconsin Sleep Cohort Study, we may wish to compare mean blood pressure between the four state agencies surveyed, by using three indicator variables. In SAS, indicator variables are created in many procedures by the CLASS statement [1].

In the simple cases presented in this chapter, we emphasize linearity of  $\mu_{y|x}$  versus a single predictor. We can easily generalize equation (1.1) to more complicated cases by transforming  $y$  or  $x$  or by adding squared, cubic, and so on, terms in  $x$ . Note, however, that even when  $x$  or  $y$  is transformed or when polynomial terms are added, ordinary regression remains a linear expression of the regression parameters. This simplifies estimation. In Chapter 12, we will consider some situations when the regression equation for the mean is not linear in the parameters.

#### 1.1.1.1 Example

OUTPUT PACKET I shows regression equations, plots of residuals versus predicted values, and mean plots for some variables from the data sets of interest. Later, we

will analyze some of these data sets longitudinally. However, for now we chose only one data point for each individual. Systolic blood pressure is analyzed from visit 1 to the sleep lab, and we selected GHb measured around 4 years diabetes duration. These variables are both regressed on age. To discern nonlinearity in the regression analysis, we look for any trend in the mean residual. Recall that the mean of the residual  $\epsilon_i$  should be 0 at all levels of the predicted value  $\mu_{y|x}$  and at all levels of  $x$ . Even a linear trend in the residual plot would indicate nonlinearity in the regression. (In contrast, we would look for curvature in a scatter plot, or a plot of means versus levels of  $x$ .) We see in the plot of GHb residuals on age that the residuals appear higher in the middle of the plot, at ages corresponding to the teenage years. It is known that adolescence is associated with poor glycemic control [2]. In the Sleep Cohort data set there is a hint of downturn of blood pressure residuals at the highest predicted values. We will see later that this may be due to some individuals with the highest levels of predicted systolic blood pressure taking blood-pressure-lowering medications. However, we see that all three residual plots display a great deal of variability in the data. This is typical of many epidemiologic and health services studies and can make it difficult to discern nonlinearity from such plots.

The regression analysis and residual plots can be generated by statements such as

```
PROC REG; MODEL SBP=AGE; PLOT RESIDUAL. *PREDICTED.;
```

However, to have residuals and predicted values available for further analysis, and especially for producing a histogram, we used statements

```
PROC REG; MODEL SBP=AGE; OUTPUT OUT=dataset R=RESID P=PRED;  
PROC PLOT; PLOT RESID*PRED;  
PROC UNIVARIATE PLOT: VAR RESID;
```

Here, residuals and predicted values are stored in the data set “dataset” together with the original variables. PROC UNIVARIATE provides a histogram of the residuals. PROC REG differs from most other SAS regression programs in its lack of ability to automatically create indicator variables and interactions. It has the advantage of being easy to run, and of accepting multiple model statements.

The plots of mean  $y_i$  were obtained by grouping  $x_i$  into intervals, so that  $\bar{y}_{group}$  can be plotted against  $\bar{x}_{group}$ . To do so (except for the duration of diabetes, which was already an integer), we used statements

```
AGEGP=5*INT(AGE/5);  
PROC SORT; BY AGEGP;  
PROC MEANS NOPRINT; BY AGEGP; VAR SBP;  
OUTPUT OUT=MM MEAN=SBPMEAN;  
PROC PLOT; PLOT SBPMEAN*AGEGP;
```

Note how the integer function was applied to efficiently create 5-year age groups (e.g.,  $5*INT(57/5) = 5*INT(11.2) = 5*11 = 55$ , so that subjects age 55–59 are in age group 55). The over-65 group was pooled with the 60–64 group. There is

a fairly linear relationship between age and systolic blood pressure at the time of the first visit to the sleep laboratory. On the other hand, the mean plot confirms that the relationship of mean GHb to age is far from linear. It appears that GHb rises rapidly until about age 14 or 15 and then declines.

### 1.1.1.2 *Comment on Bias*

If the linearity assumption does not hold in (1.1), the mean of  $\epsilon_i$  is not 0 for all  $x_i$ , and (1.1) is not a good representation of  $\mu_{y|x}$ . The linearity assumption can also be phrased as lack of bias in representing, and later estimating, the mean  $\mu_{y|x}$  by the regression equation. Lack of bias is often seen as the most important attribute of an estimator, making the linearity assumption of paramount importance in (1.1).

Technically, unbiasedness in an estimator is defined as the property of being “correct on average.” As long as the linearity assumption holds, it can be shown (as we do later) that least-squares estimators of  $\beta_0$  and  $\beta_1$  are unbiased. This means that if studies producing estimators for (1.1) were done many, many times over, and the estimators  $\hat{\beta}_0$  and  $\hat{\beta}_1$  averaged across all these studies, the result would be the true  $\beta_0$  and  $\beta_1$ . When  $\hat{\beta}_0$  and  $\hat{\beta}_1$  average to their correct values and (1.1) is a correct formulation,  $\hat{\mu}_{y|x}$  is an unbiased estimator of the mean  $y$  at a given  $x$ .

As we will discuss in Chapter 7, bias can also be created by unequal probability sampling from the population. We will demonstrate there how to correct for such bias. A special case of selection bias occurs when subjects are chosen based on screening high on some risk factor [3].

### 1.1.1.3 *Comment on Causal Interpretation*

Even when there is technically no bias in equation (1.1) as stated conditionally on  $x$ , it is important to remember that  $\beta_1$  may not have a causal interpretation. Consider a situation where the causal model conditional on both  $x$  and all confounders can be formulated:

$$y_i = \beta'_0 + \beta'_1 x_i + \beta_2 w_i + \epsilon_i \quad (1.2)$$

where  $\beta'_1$  measures the causal effect of  $x$  on  $y$  and where  $w$  is a confounder. (Recall the definition of a confounder as a variable associated with both the outcome and risk factors.) If  $x$  and  $w$  are both normally distributed with correlation  $\rho_{xw}$  and variances  $\sigma_x^2$  and  $\sigma_w^2$ , respectively, it can be shown that the formulation of (1.1), conditionally on only  $x$ , becomes

$$y_i = \beta_0 + \left( \beta'_1 + \beta_2 \rho_{xw} \frac{\sigma_w}{\sigma_x} \right) x_i + \epsilon_i \quad (1.3)$$

with all assumptions of (1.1) satisfied. However,  $\beta_1$  is confounded as

$$\beta_1 = \beta'_1 + \beta_2 \rho_{xw} \frac{\sigma_w}{\sigma_x}$$

The above situation is known as confounding in epidemiology and as endogeneity in econometrics. Another way to describe endogeneity is that the unknown or

error part  $\beta_2 w_i + e_i$  in (1.2) does not have mean 0 for all  $x$ . However, we will not know that when we are fitting (1.1), unaware of the presence of  $w$ .

### 1.1.2 Equal Variance Assumption

The equal variance assumption can be written  $\text{Var}(\epsilon|x) = \sigma_{y|x}^2$ , which implies that the variability of  $y$  around its mean is the same at every  $x$ . For example, the variability in systolic blood pressure is assumed to be the same at every age. The assumption enters in the formulation of the least-squares equations. Recall that in applying the least-squares principle, the expression  $\sum (y_i - \hat{\mu}_{y_i|x_i})^2$  is minimized, where  $\hat{\mu}_{y_i|x_i} = \hat{\beta}_0 + \hat{\beta}_1 x_i$ . This expression treats all observation points equally. But, if  $\text{Var}(\epsilon_i|x_i)$  is not equal at all points, it would be sensible to give less weight to the points where this variability is greater. Where variability is wider, observations  $y_i$  tell us less about where the mean is. For example, at ages where GHb variability is the greatest, measurements are less informative about the location of the curve describing the mean. With biological measurements the variance often tends to increase with the size of the measurement, so if  $y$  is positively related to  $x$ ,  $\text{Var}(\epsilon_i|x_i)$  is often higher when  $x_i$  is higher.

In addition to the above consideration, the estimate of the residual variance is difficult to interpret and cannot be assumed to yield valid significance tests when the true residual variance is not constant. We will see, in later chapters, how significance tests can be corrected to take unequal variance into account.

#### 1.1.2.1 Example

In a plot of residuals versus predicted values, we assess the equal variance assumption by looking for whether the plot tends to fan out, usually to the right side. Some caution is in order, as there will be more spread among points in areas with many observations. In our residual plots, we see fanning out toward the high side of the predicted GHb. This corresponds to greater variability during the adolescent years when GHb peaks.

#### 1.1.2.2 Comment on Efficiency

It is a general principle in statistics that weighting an observation by the inverse of its variance yields estimators with the smallest standard errors. Hence we foresee that similarly weighting observations in the least-squares estimator may be beneficial. However, we will need to provide mathematical justification for exactly how to do this weighting.

The property of smallest possible standard errors is referred to as *efficiency*. The word efficiency is used similarly when referring to an electrical device or engine: how well the input (i.e., the electricity/fuel or data) is utilized in producing the desired product (e.g., refrigeration/mileage or regression estimators). One familiar situation where the equal variance assumption is clearly violated is when  $y_i$  is a binary variable. Then  $\text{Var}(y_i)$  is  $\pi_i(1 - \pi_i)$ , which depends on the proportion of success  $\pi_i$ . When  $y_i$  is coded 0, 1 we have  $\pi_i = \mu_{y_i|x_i}$ , so the variance depends on the mean of  $y_i$ . We will see later how such relationships are taken into account. A

common solution for binary outcome is to apply logistic regression, abandon the idea of least-squares estimation, and apply maximum likelihood. We will emphasize the link between the two approaches.

### 1.1.3 Normality Assumption

We usually assume that the distribution of the error term  $\epsilon_i$  is normal. This assumption enters in forming inferences for the estimators. The normality assumption allows us use of the  $t$ -distribution to obtain tests and confidence intervals for individual coefficients and the validity of  $F$ -tests for the model as a whole. However, it can be shown that in large samples  $\hat{\beta}_0$ ,  $\hat{\beta}_1$ , and their test statistics have approximately normal sampling distributions regardless of whether  $\epsilon_i$  is normally distributed. The beauty of the  $t$ - and  $F$ -tests is that they are applicable even with small sample sizes as long as the normality assumption holds for the residuals. The reader is probably familiar with how, when we abandon the normality assumption, in logistic regression, we are “stuck with”  $\chi^2$ -tests that yield correct inference only in large samples. It may be noted that for large sample sizes when the degrees of freedom for the estimator  $s_{y|x}^2$  is large, so its random error vanishes, the  $t$ -distribution approaches the normal,  $t^2$  approaches  $\chi^2(1)$ , and the  $F$  distribution with  $m$  numerator degrees of freedom approaches  $\chi^2(m)/m$ .

Using maximum likelihood requires that a distributional assumption be made on  $\epsilon_i$ . We will see later that the normality assumption leads to equality of least-squares and maximum likelihood estimators for model (1.1). However, even later we will see that the normal distribution is one example in a broader framework and that it is convenient to draw connections between maximum likelihood and least-squares estimation procedures for many common distributions.

#### 1.1.3.1 Examples

The normality assumption can be assessed informally, but adequately, by looking at a histogram of the residuals. Our graph for the residuals of systolic blood pressure on age shows very slight skewness. This is rather typical. With a large data set, it is not of great concern, unless skewness is extreme. With a small data set, on the other hand, it may be difficult to assess normality.

Note that it is normality of the residuals that is required for  $t$ - and  $F$ -tests of regression coefficients to be valid in small samples, not normality of the outcome before taking  $x$  into account. It is not uncommon to see that normality improves when conditioning on  $x$ . The final example in OUTPUT PACKET I shows a regression analysis of number of days VLBW infants spend in the NICU on birth weight. We see that while the number of days has a rather skewed distribution, the skewness disappears when conditioning on the infant’s birth weight.

#### 1.1.3.2 Comment on Normality

Sometimes, a variable’s distribution is fairly well known from previous studies to be normally distributed in populations that are unselected for the variable. Systolic blood pressure is an example of a measure that has been widely studied. It is

usually found to be very close to normally distributed, although some investigators have applied the transformation  $\log(\text{SBP}-50)$  [4]. In our case this transformation reduced skewness only slightly, and the distribution became skewed in the opposite direction instead. It is interesting that designed variables such as IQ are typically scaled to have normal distributions in the population. Hence, the normality of such variables is “man-made,” while the normality of biological variables is considered to be the result of many factors being added up, so that by the central limit theorem the end result is normally distributed. Recall that the central limit theorem states that means and sums of many independent variables tend toward normal distributions.

#### 1.1.4 Independence Assumption

We usually assume that  $\epsilon_i$  and  $\epsilon_{i'}$  are independent for  $i \neq i'$ —that is, that the residuals for two different observations on  $y$  do not “travel together” once their corresponding  $x$ ’s are taken into account. We will see that the assumption enters when deriving the standard errors of regression coefficients. Intuitively, one can see that if  $\epsilon_i$  and  $\epsilon_{i'}$  are positively correlated, we really have less information than we presume, when we base our inference on thinking that all  $y$ ’s contribute a given piece. In other words, lack of independence implies that only part of the information about  $\beta_0$  and  $\beta_1$  imparted by  $y_{i'}$  is new; the rest has already been gained from  $y_i$ .

Consider the data from the Wisconsin Diabetes Registry. Each individual provided a number of measurements on glycosylated hemoglobin across several years. Say we wish to examine how glycosylated hemoglobin (GHb) relates to the duration of diabetes. Obviously, much variability in GHb exists at each duration and is reflected in the  $\epsilon$ . While it may be fairly reasonably assumed that  $\epsilon$  from different individuals are independent, GHb of the same individual may tend to be uniformly on the high or low side, depending on the person’s diet, diabetes care regimen, exercise level, and so on. We can usually not begin to hope that we have captured all these influences in the  $x$ ’s we have in the study. Hence  $\epsilon$  on the same individual are not independent. Similarly, in the sleep study, having blood pressure data on 1251 measurements from three visits on 520 individuals does not convey the same information on, for example, gender differences as having 1251 measurements on 1251 different individuals.

Because there is less information in the data when residuals are positively correlated than when they are independent, standard errors can be underestimated. We will see that lack of independence can be dealt with either by maximum likelihood estimation where the joint distribution of the measurements is correctly modeled, or a modified least-squares approach.

##### 1.1.4.1 Example

It is not easy to discern most cases of dependence in an overall scatter plot, or residual plot. One may target special cases of dependency that are expected in a given study. In OUTPUT PACKET I, we included only one GHb observation

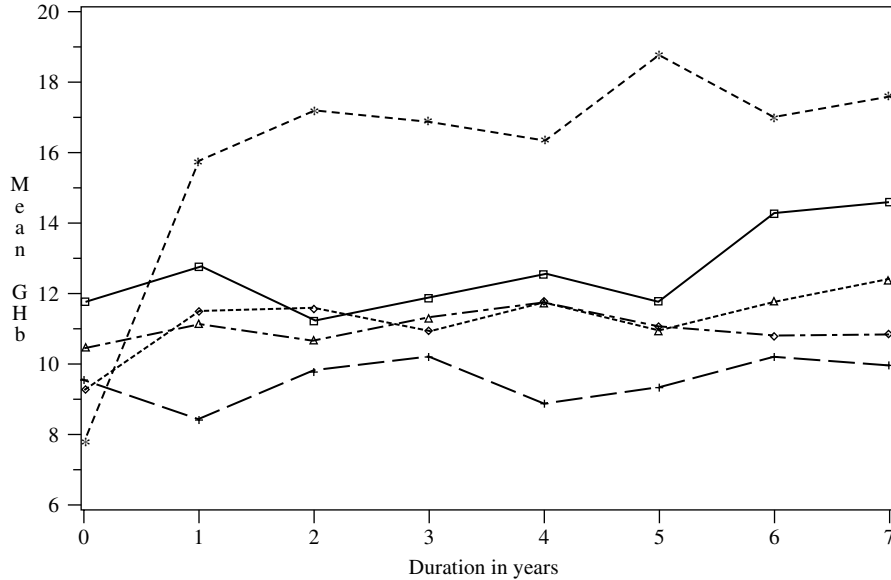


Fig. 1.1 Wisconsin Diabetes Registry Study GHb for the first five subjects

per individual, to improve independence of measurements. (Of course this is very wasteful, and one of our goals is to enable inclusion of all the data in the analysis). In Figure 1.1, GHb versus duration is shown for the first five individuals in the data set. We see that the values of a given individual tend to be on the high or low side. In fact, the average correlation among residuals from the same individual is 0.51. Systolic blood pressure residuals across visits in the sleep study correlate at 0.35 within person.

## 1.2 A NOTE ON HOW THE LEAST-SQUARES ESTIMATORS ARE OBTAINED

We need calculus to minimize the expression  $\sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$  with respect to the estimators  $\hat{\beta}_0, \hat{\beta}_1$ . In this case, the two estimators constitute two “variables” in the calculus sense, and we take the derivatives of the least-squares expression, invoking the chain rule. Noting that

$$LS = \sum LS_i = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum \hat{\epsilon}_i^2$$

we have that

$$\frac{dLS_i}{d\hat{\epsilon}_i} = 2\hat{\epsilon}_i$$

so

$$\frac{\partial LS_i}{\partial \hat{\beta}_0} = 2\hat{\epsilon}_i \frac{\partial \hat{\epsilon}_i}{\partial \hat{\beta}_0} = 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-1)$$

and

$$\frac{\partial LS_i}{\partial \hat{\beta}_1} = 2\hat{\epsilon}_i \frac{\partial \hat{\epsilon}_i}{\partial \hat{\beta}_1} = 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i)$$

We apply the rule that the derivative of a sum is the sum of derivatives. Then setting the final derivatives to 0 we obtain:

$$\begin{aligned} \frac{\partial LS}{\partial \hat{\beta}_0} &= -2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \frac{\partial LS}{\partial \hat{\beta}_1} &= -2 \sum x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \end{aligned} \tag{1.4}$$

Simultaneously solving these equations results in the usual estimators

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned}$$



**OUTPUT PACKET I: EXAMPLES OF ORDINARY REGRESSION ANALYSES**

**I.1. Analysis of SBP Versus Age: Wisconsin Sleep Cohort Study**

*The REG Procedure*

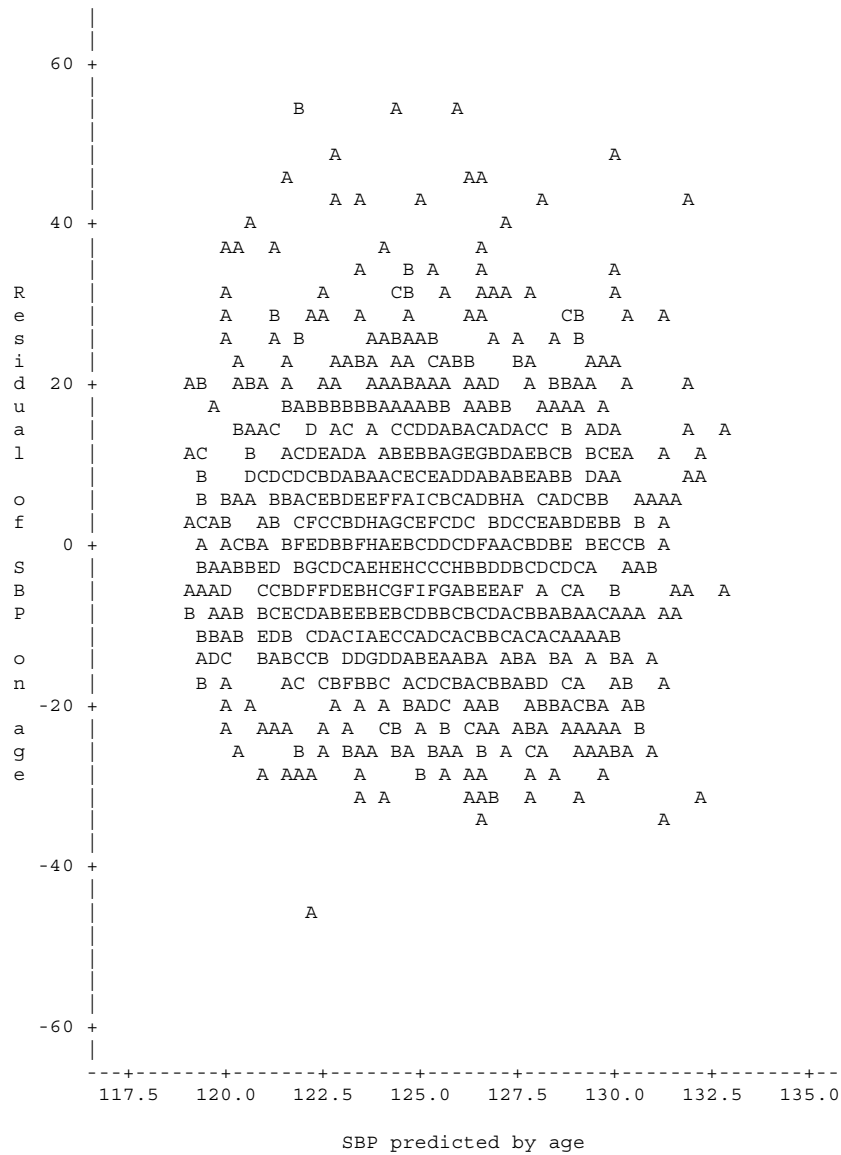
Model: MODEL1  
Dependent Variable: SBP

**Analysis of Variance**

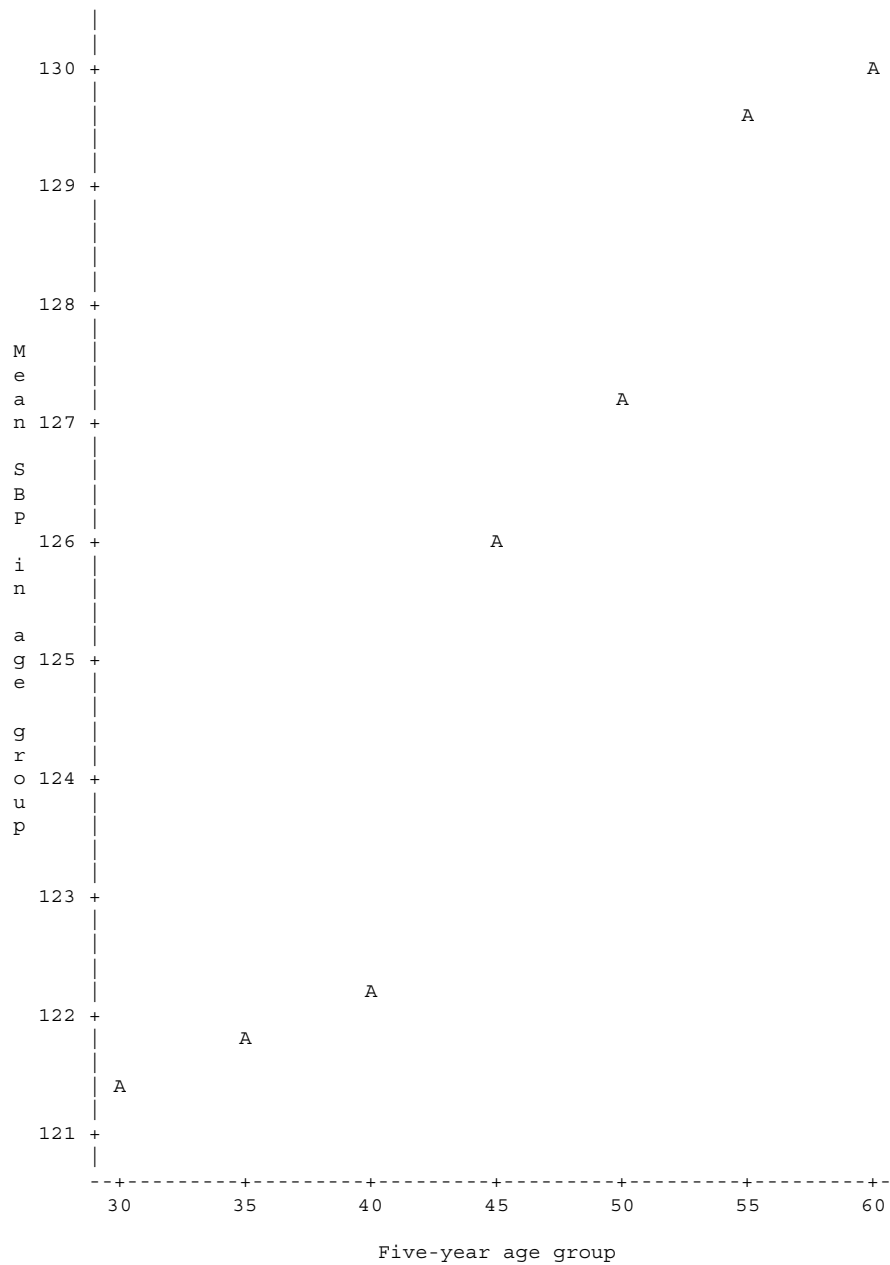
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	11604	11604	56.43	<0.0001
Error	1363	280294	205.64520		
Corrected total	1364	291898			
Root MSE		14.34033	R-square	0.0398	
Dependent mean		125.09145	Adjusted R-square	0.0390	
Coefficient of variation		11.46388			

**Parameter Estimates**

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	107.98092	2.31066	46.73	<0.0001
Age	1	0.36605	0.04873	7.51	<0.0001



Plot of resid\*pred. Legend: A = 1 obs, B = 2 obs, and so on. Note: 13 obs had missing values.



*The UNIVARIATE Procedure*

Variable: resid (residual of SBP versus age)

**Moments**

<i>N</i>	1365	Sum weights	1365
Mean	0	Sum observations	0
Standard deviation	14.3350769	Variance	205.49443
Skewness	0.41763912	Kurtosis	0.61294189
Uncorrected SS	280294.403	Corrected SS	280294.403
Coefficient of variation		Standard error of mean	0.38800177

**Basic Statistical Measures**

Location		Variability	
Mean	0.00000	Standard deviation	14.33508
Median	-0.64707	Variance	205.49443
Mode	13.66973	Range	99.99362
		Interquartile range	18.47016

**Tests for Location:  $\mu_0 = 0$** 

Test	Statistic	<i>p</i> Value
Student's <i>t</i>	t      0	Pr >   <i>t</i>      1.0000
Sign	M      -23.5	Pr ≥   <i>M</i>      0.2131
Signed rank	S     -15323.5	Pr ≥   <i>S</i>      0.2930

**Quantiles (Definition 5)**

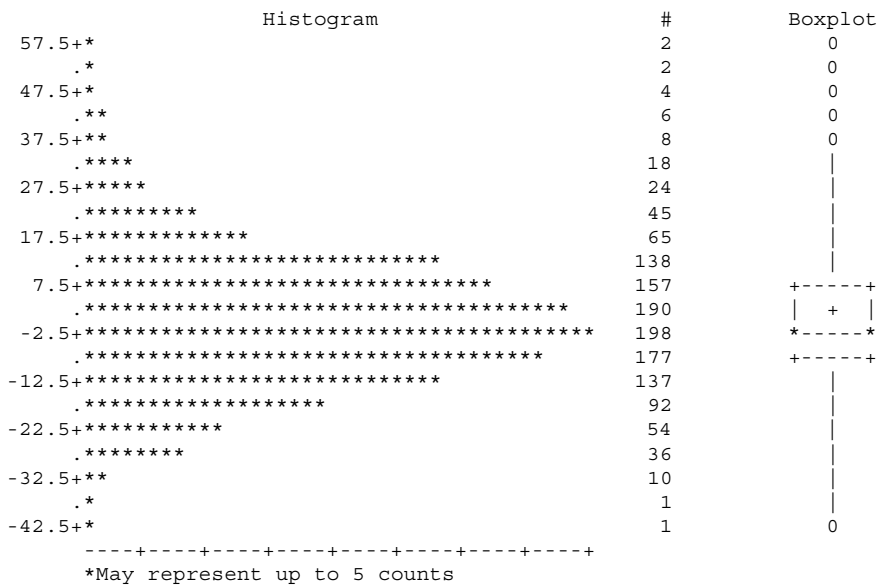
Quantile	Estimate
100% Max	55.669359
99%	41.669569
95%	24.621515
90%	17.771212
75% Q3	8.770605
50% Median	-0.647072
25% Q1	-9.699556
10%	-17.683597
5%	-22.715642
1%	-29.571613
0% Min	-44.324258

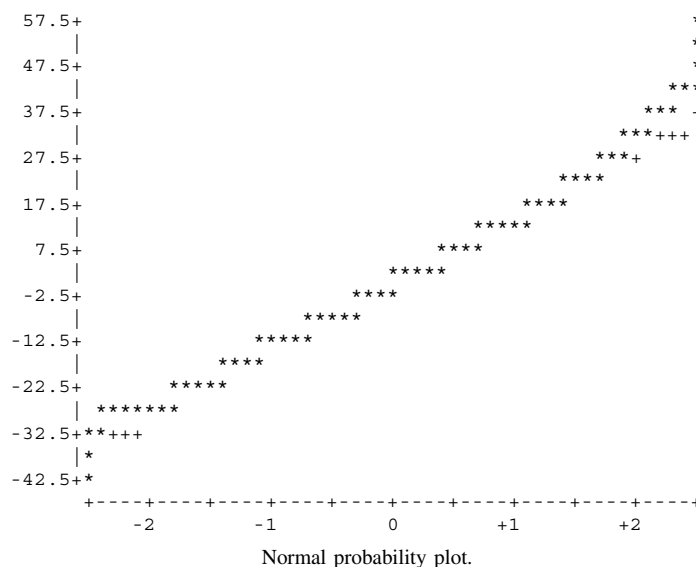
**Extreme Observations**

Lowest		Highest	
Value	Obs	Value	Obs
-44.3243	1167	47.8313	85
-35.4823	1080	53.4092	802
-33.3630	1190	54.6771	367
-32.3541	896	55.0596	1208
-32.0275	1194	55.6694	1322

**Missing Values**

Missing Value	Count	Percent of	
		All Observations	Missing Obs
	13	0.94	100.00





**I.2. Analysis of GHb Versus Age—Wisconsin Diabetes Registry**

*GHb Versus Age—Wisconsin Diabetes Registry*

*The REG Procedure*

Model: MODEL1

Dependent Variable: GHb

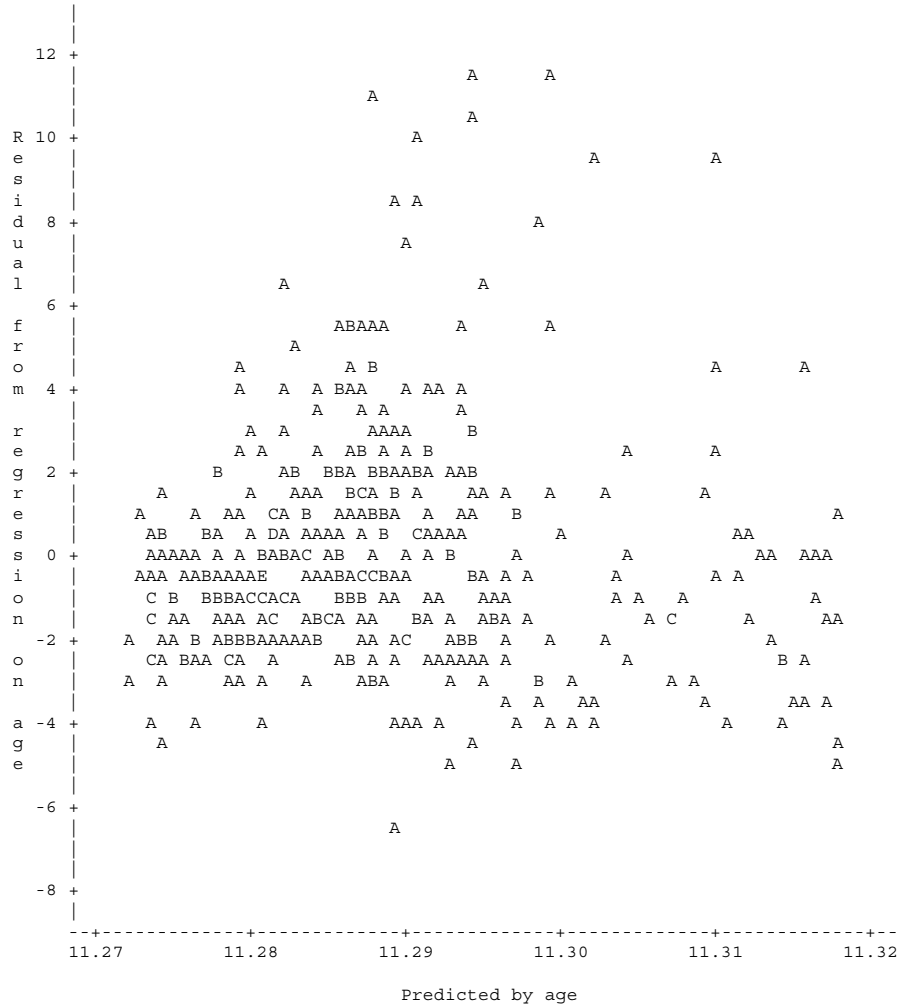
**Analysis of Variance**

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.04566	0.04566	0.01	0.9381
Error	413	3119.52820	7.55334		
Corrected total	414	3119.57387			
Root MSE		2.74833	R-square	0.0000	
Dependent mean		11.28859	Adjusted R-square	-0.0024	
Coefficient of variation		24.34613			

**Parameter Estimates**

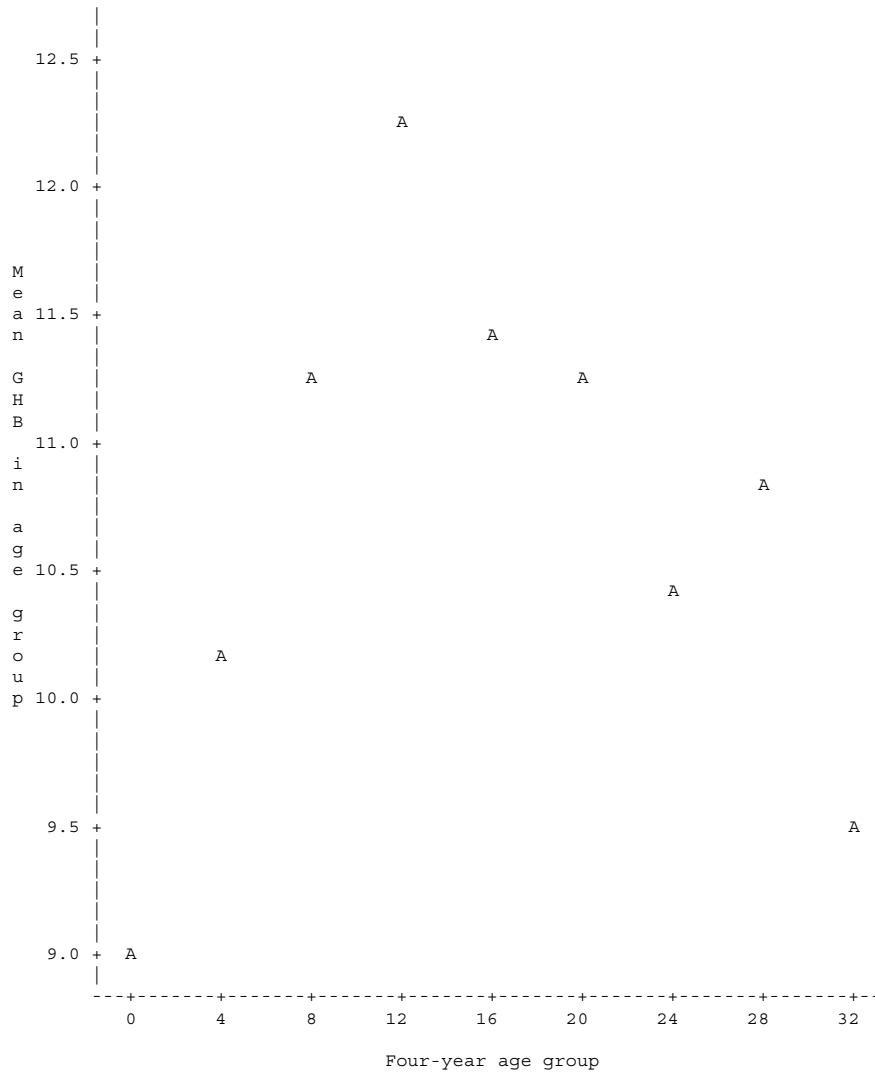
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	Intercept	1	11.26650	0.31444	35.83	<0.0001
Age	Age	1	0.00155	0.02000	0.08	0.9381

*Residual Plot Versus Predicted Value*



Plot of resid\*pred. Legend: A = 1 obs, B = 2 obs, and so on.

*Mean GHb by Four-Year Age Groups*



Plot of ghmean\*iage. Legend: A = 1 obs, B = 2 obs, and so on.

*Histogram of GHb Residuals*  
*The UNIVARIATE Procedure*

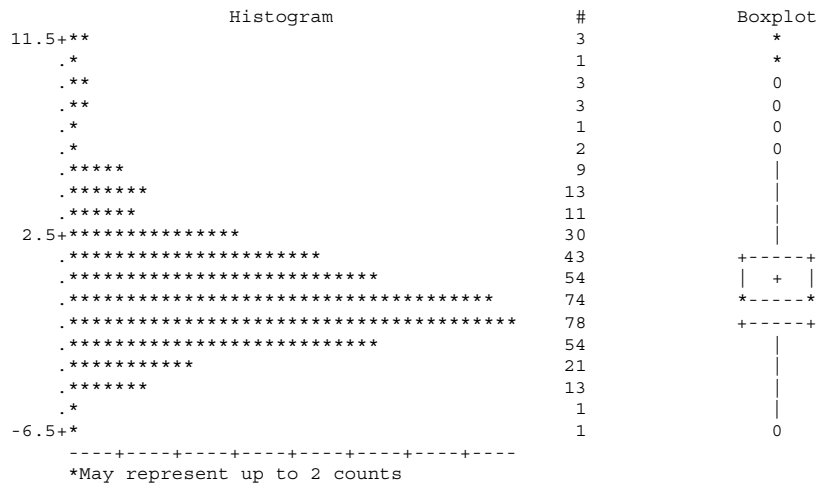
Variable: resid (residual from regression versus age)

**Moments**

<i>N</i>	415	Sum of weights	415
Mean	0	Sum of observations	0
Standard deviation	2.74501226	Variance	7.53509228
Skewness	1.28441431	Kurtosis	2.835934
Uncorrected SS	3119.5282	Corrected SS	3119.5282
Coefficient of variation		Standard error of mean	0.13474735

**Basic Statistical Measures**

Location		Variability	
Mean	0.00000	Standard deviation	2.74501
Median	-0.43380	Variance	7.53509
Mode	-1.10334	Range	17.89502
		Interquartile range	3.02800





*Regression of Number of Days in NICU*  
*The REG Procedure*

Model: MODEL1  
 Dependent Variable: len (days in NICU)

**Analysis of Variance**

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	280517	280517	572.28	<0.0001
Error	765	374986	490.17803		
Corrected total	766	655503			
	Root MSE	22.13996	R-square	0.4279	
	Dependent mean	60.71186	Adjusted R-square	0.4272	
	Coefficient of variation	36.46728			

**Parameter Estimates**

Variable	Label	DF	Parameter Estimate	Standard Error
Intercept	Intercept	1	145.15576	3.61932
Birth weight	Birth weight in grams	1	-7.49741	0.31341

**Parameter Estimates**

Variable	Label	DF	t Value	Pr >  t
Intercept	Intercept	1	40.11	<0.0001
Birth weight	Birth weight in grams	1	-23.92	<0.0001

