3.4b, which plots the residuals versus predicted salaries, shows a clear violation of the constant variance assumption. For lower predicted salaries there is little variability about 0, but for the high salaries there is considerable variability of the residuals. The implication of this is that the model will predict lower salaries quite accurately, but not so for the higher salaries.

Figure 3.4c plots the residuals versus number of years in the major leagues. This plot shows a clear curvilinear clustering, that is, quadratic. The curved lines encompass the vast majority of points to make this trend even more evident. The implication of this curvilinear trend is that the regression model will tend to overestimate the salaries of players who have been in the major years only a few years or over 15 years, and it will underestimate the salaries of players who have been in the majors about 5 to 9 years.

In concluding this section, note that if nonlinearity or nonconstant variance is found, there are various remedies. For nonlinearity, perhaps a polynomial model is needed. Or sometimes a transformation of the data will enable a nonlinear model to be approximated by a linear one. For nonconstant variance, weighted least squares is one possibility, or more commonly, a variance-stabilizing transformation (such as square root or log) may be used. I refer the reader to Weisberg (1985, chapter 6) for an excellent discussion of remedies for regression model violations.

3.11 MODEL VALIDATION

We indicated earlier that it was crucial for the researcher to obtain some measure of how well the regression equation will predict on an independent sample(s) of data. That is, it was important to determine whether the equation had generalizability. We discuss here three forms of model validation, two being empirical and the other involving an *estimate* of average predictive power on other samples. First I give a brief description of each form, and then elaborate on each form of validation.

1. *Data splitting*. Here the sample is randomly split in half. It does not have to be split evenly, but we use this for illustration. The regression equation is found on the so-called derivation sample (also called the screening sample, or the sample that "gave birth" to the prediction equation by Tukey). This prediction equation is then applied to the other sample (called validation or calibration) to see how well it predicts the y scores there.
2. *Compute an adjusted* R^2. There are various adjusted R^2 measures, or measures of shrinkage in predictive power, but they do not all estimate the same thing. The one most commonly used, and that which is printed out by both major statistical packages, is due to Wherry. It is very important to note here that the Wherry formula estimates how much variance on y would be accounted for if we had derived the prediction equation in the

population from which the sample was drawn. The Wherry formula does *not* indicate how well the derived equation will predict on other samples from the same population. A formula due to Stein (1960) does estimate average cross-validation predictive power. As of this writing it is not printed out by any of the three major packages. The formulas due to Wherry and Stein are presented shortly.

3. *Use the PRESS statistic.* As pointed out by several authors, in many instances one does not have enough data to be randomly splitting it. One can obtain a good measure of *external* predictive power by use of the PRESS statistic. In this approach the y value for *each* subject is set aside and a prediction equation derived on the remaining data. Thus, n prediction equations are derived and n true prediction errors are found. To be very specific, the prediction error for subject 1 is computed from the equation derived on the remaining $(n - 1)$ data points, the prediction error for subject 2 is computed from the equation derived on the other $(n - 1)$ data points, and so on. As Myers (1990) put it, "PRESS is important in that one has information in the form of n validations in which the fitting sample for each is of size $n - 1$" (p. 171).

Data Splitting

Recall that the sample is randomly split. The regression equation is found on the derivation sample and then is applied to the other sample (validation) to determine how well it will predict y there. Next we give a hypothetical example, randomly splitting 100 subjects.

Derivation Sample
$n = 50$
Prediction Equation $\hat{y}_i = 4 + .3x_1 + .7x_2$

Validation Sample
$n = 50$

y	x_1	x_2
6	1	.5
4.5	2	.3
	...	
7	5	.2

Now, using this prediction equation we predict the y scores in the validation sample:

$$\hat{y}_1 = 4 + .3\,(1) + .7\,(.5) = 4.65$$

$$\hat{y}_2 = 4 + .3\,(2) + .7\,(.3) = 4.81$$

$$\cdots$$

$$\hat{y}_{50} = 4 + .3\,(5) + .7\,(.2) = 5.64$$

The cross-validated R then is the correlation for the following set of scores:

y	$\hat{y}_i$
6	4.65
4.5	4.81
...	
7	5.64

Random splitting and cross validation can be easily done using SPSS and the filter case function.

Cross Validation with SPSS

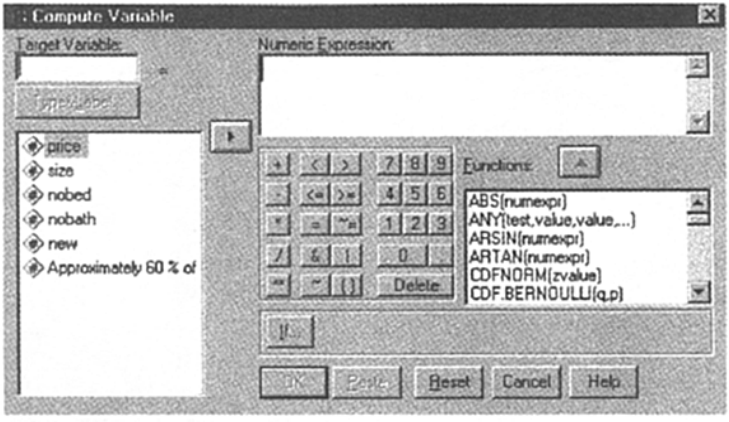
To illustrate cross validation with SPSS for Windows 10.0 we use the Agresti data. Recall that the sample size here was 93. First we randomly select a sample and do a stepwise regression on this random sample. We have selected an approximate random sample of 60%. It turns out there is an n = 60 in our sample. This is done by clicking on DATA, choosing SELECT CASES from the drop-down menu, then choosing RANDOM SAMPLE and finally selecting a random sample of approximately 60%. When this is done a FILTER_$ variable is created, with value = 1 for those cases included in the sample and value = 0 for those cases *not* included in the sample. When the stepwise regression was done the variables SIZE, NOBATH and NEW were included as predictors and the coefficients, etc are given below for that run:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-28.948	8.209		-3.526	.001
	SIZE	78.353	4.692	.910	16.700	.000
2	(Constant)	-62.848	10.939		-5.745	.000
	SIZE	62.156	5.701	.722	10.902	.000
	NOBATH	30.334	7.322	.274	4.143	.000
3	(Constant)	-62.519	9.976		-6.267	.000
	SIZE	59.931	5.237	.696	11.444	.000
	NOBATH	29.436	6.682	.266	4.405	.000
	NEW	17.146	4.842	.159	3.541	.001

a. Dependent Variable: PRICE

The next step in the cross validation is to use the COMPUTE statement to compute the predicted values for the dependent variable. This COMPUTE statement is obtained by clicking on TRANSFORM and then selecting COMPUTE from the dropdown menu. When this is done the following screen appears:



Using the coefficients obtained from the above regression we have:

PRED = −62.519 + 59.931*SIZE + 29.436*NOBATH + 17.146*NEW

We wish to correlate the predicted values in the other part of the sample with the y values there to obtain the cross validated value. We click on DATA again, and use SELECT IF FILTER_$ = 0. That is, we select those cases in the *other* part of the sample. There are 33 cases in the other part of the random sample. When this is done all the cases with FILTER_$ = 1 are selected, and a partial listing of the data appears as follows:

	price	size	nobed	nobath	new	filter_$	pred
1	48.50	1.10	3.00	1.00	.00	0	32.84
2	55.00	1.01	3.00	2.00	.00	0	56.88
3	68.00	1.45	3.00	2.00	.00	1	83.25
4	137.00	2.40	3.00	3.00	.00	0	169.62
5	309.40	3.30	4.00	3.00	1.00	0	240.71
6	17.50	.40	1.00	1.00	.00	1	-9.11
7	19.60	1.28	3.00	1.00	.00	0	43.63
8	24.50	.74	3.00	1.00	.00	0	11.27

Finally, we use the CORRELATION program to obtain the bivariate correlation between PRED and PRICE (the dependent variable) in this sample of 33. That correlation is .878, which is a drop from the maximized correlation of .944 in the derivation sample.

Adjusted R^2

Herzberg (1969) presented a discussion of various formulas that have been used to estimate the amount of shrinkage found in R^2. As mentioned earlier, the one most commonly used, and due to Wherry, is given by

$$\hat{\rho}^2 = 1 - \frac{(n-1)}{(n-k-1)}(1 - R^2) \tag{8}$$

where $\hat{\rho}$ is the estimate of ρ, the population multiple correlation coefficient. This is the adjusted R^2 printed out by SAS and SPSS. Draper and Smith (1981) commented on Equation 8: "A related statistic . . . is the so called adjusted $r(R_a^2)$, the idea being that the statistic R_a^2 can be used to compare equations fitted not only to a specific set of data but also to two or more entirely different sets of data. The value of this statistic for the latter purpose is, in our opinion, not high" (p. 92).

Herzberg noted:

> In applications, the population regression function can never be known and one is more interested in how effective the *sample* regression function is in *other* samples. A measure of this effectiveness is r_c, the sample cross-validity. For any given regression function r_c will vary from validation sample to validation sample. The average value of r_c will be approximately equal to the correlation, in the *population*, of the sample regression function with the criterion. This correlation is the population cross-validity, ρ_c. Wherry's formula estimates ρ rather than ρ_c. (p. 4)

There are two possible models for the predictors: (a) regression—the values of the predictors are fixed, that is, we study y only for certain values of x, and (b) correlation—the predictors are random variables—this is a much more reasonable model for social science research. Herzberg presented the following formula for estimating ρ_c^2 under the correlation model:

$$\hat{\rho}_c^2 = 1 - \left(\frac{n-1}{n-k-1}\right)\left(\frac{n-2}{n-k-2}\right)\left(\frac{n+1}{n}\right)(1 - R^2) \tag{9}$$

where n is sample size and k is the number of predictors. It can be shown that $\rho_c < \rho$.

If you are interested in cross validity predictive power, then the Stein formula (Equation 9) should be used. As an example, suppose $n = 50$, $k = 10$ and $R^2 = .50$. If you used the Wherry formula (Equation 8), then your estimate is

$$\hat{\rho}^2 = 1 - 49/39\,(.50) = .372$$

whereas with the proper Stein formula you would obtain

$$\hat{\rho}_c^2 = 1 - (49/39)(48/38)(51/50)(.50) = .191$$

In other words, use of the Wherry formula would give a misleadingly positive impression of the cross validity predictive power of the equation.

Table 3.9 shows how the estimated predictive power drops off using the Stein formula (Equation 9) for small to fairly large subject/variable ratios when $R^2 = .50$.

PRESS Statistic

The PRESS approach is important in that one has n validations, each based on $(n - 1)$ observations. Thus, each validation is based on essentially the entire sample. This is very important when one does not have large n, for in this situation

TABLE 3.8
Estimated Predictive Power Using the Stein Formula for Small to Fairly Large Subject/Variable Ratios

Subject/Variable Ratio	Stein Estimate $1 - \left(\frac{n-1}{n-k-1}\right)\left(\frac{n-2}{n-k-2}\right)\left(\frac{n+1}{n}\right)(1 - R^2)$	Comment
Small (5:1) $n = 50, k = 10$ $R^2 = .50$ ②	$1 - (49/39)\,(48/38)\,(51/50)\,(.5)$ $= .191$ ①	The estimated amount of shrinkage is great, i.e., on the average we expect the predictive power to be reduced by about 60%.
Moderate (10:1) $n = 100, k = 10$ $R^2 = .50$	$1 - (99/98)\,(98/88)\,(101/100)\,(.5)$ $= .374$	The shrinkage is still fairly substantial.
Fairly Large (15:1) $n = 150, k = 10$ $R^2 = .50$	$1 - (149/139)\,(148/138)\,(151/150)\,(.5)$ $= .421$	We finally reach a point where the expected amount of shrinkage is fairly small, about 16%.

① If we were to apply the prediction equation to many other samples from the same population, then on the *average* we would account for 19.1% of the variance on y.

② We have chosen this value to illustrate because the typical R^2 values found in social science are often around .50.

data splitting is really not practical. For example, if $n = 60$ and we have 6 predictors, randomly splitting the sample involves obtaining a prediction equation on only 30 subjects.

Recall that in deriving the prediction (via the least squares approach), the sum of the squared errors is *minimized*. The PRESS residuals, on the other hand, are true prediction errors, because the y value for each subject was not simultaneously used for fit and model assessment. Let us denote the predicted value for subject i, where that subject was *not* used in developing the prediction equation, by $\hat{y}_{(-i)}$. Then the PRESS residual for each subject is given by

$$\hat{e}_{(-i)} = y_i - \hat{y}_{(-i)}$$

and the PRESS sum of squared residuals is given by

$$\text{PRESS} = \Sigma \hat{e}_{(-i)}^2 \tag{10}$$

Therefore, one might prefer the model with the smallest PRESS value. The preceding PRESS value can be used to calculate an R^2-like statistic that more accurately reflects the generalizability of the model. It is given by

$$R_{\text{Press}}^2 = 1 - (\text{PRESS})/\Sigma (y_i - \bar{y})^2 \tag{11}$$

Importantly, the SAS REG program does routinely print out PRESS, although it is called PREDICTED RESID SS (PRESS). Given this value, it is a simple matter to calculate the R^2 PRESS statistic, because $s_y^2 = \Sigma(y_i - \bar{y})^2 / (n - 1)$.

3.12 IMPORTANCE OF THE ORDER OF THE PREDICTORS IN REGRESSION ANALYSIS

The order in which the predictors enter a regression equation can make a great deal of difference with respect to how much variance on y they account for, especially for moderate or highly correlated predictors. Only for uncorrelated predictors (which would rarely occur in practice) does the order not make a difference. We give two examples to illustrate.

Example 7

A dissertation by Crowder (1975) attempted to predict ratings of trainably mentally (TMs) retarded individuals using IQ (x_2) and scores from a Test of Social Inference (TSI). He was especially interested in showing that the TSI had incremental predictive validity. The criterion was the average ratings by two individuals in charge of the TMs. The intercorrelations among the variables were:

$$r_{x_1x_2} = .59,\ r_{yx_2} = .54,\ r_{yx_1} = .566$$