TABLE 9.3 Nonlinear Least Squares and Instrumental Variable Estimates

	Instrumental Variables		*Least Squares*	
Parameter	***Estimate***	***Standard Error***	***Estimate***	***Standard Error***
α	627.031	26.6063	468.215	22.788
β	0.040291	0.006050	0.0971598	0.01064
γ	1.34738	0.016816	1.24892	0.1220
σ	57.1681	—	49.87998	—
$\mathbf{e'e}$	650,369.805	—	495,114.490	—

The model is

$$y = h(\mathbf{x}, \boldsymbol{\beta}, \mathbf{w}, \boldsymbol{\gamma}) + \varepsilon.$$

We consider cases in which the auxiliary parameter $\boldsymbol{\gamma}$ is estimated separately in a model that depends on an additional set of variables $\mathbf{w}$. This first step might be a least squares regression, a nonlinear regression, or a maximum likelihood estimation. The parameters $\boldsymbol{\gamma}$ will usually enter $h(.)$ through some function of $\boldsymbol{\gamma}$ and $\mathbf{w}$, such as an expectation. The second step then consists of a nonlinear regression of y on $h(\mathbf{x}, \boldsymbol{\beta}, \mathbf{w}, \mathbf{c})$ in which $\mathbf{c}$ is the first-round estimate of $\boldsymbol{\gamma}$. To put this in context, we will develop an example.

The estimation procedure is as follows.

1. Estimate $\boldsymbol{\gamma}$ by least squares, nonlinear least squares, or maximum likelihood. We assume that this estimator, however obtained, denoted $\mathbf{c}$, is consistent and asymptotically normally distributed with asymptotic covariance matrix $\mathbf{V}_c$. Let $\hat{\mathbf{V}}_c$ be any appropriate estimator of $\mathbf{V}_c$.

2. Estimate $\boldsymbol{\beta}$ by nonlinear least squares regression of y on $h(\mathbf{x}, \boldsymbol{\beta}, \mathbf{w}, \mathbf{c})$. Let $\sigma^2\mathbf{V}_b$ be the asymptotic covariance matrix of this estimator of $\boldsymbol{\beta}$, assuming $\boldsymbol{\gamma}$ is known and let $s^2\hat{\mathbf{V}}_b$ be any appropriate estimator of $\sigma^2\mathbf{V}_b = \sigma^2(\mathbf{X}^{0\prime}\mathbf{X}^0)^{-1}$, where $\mathbf{X}^0$ is the matrix of pseudoregressors evaluated at the true parameter values $\mathbf{x}_i^0 = \partial h(\mathbf{x}_i, \boldsymbol{\beta}, \mathbf{w}_i, \boldsymbol{\gamma})/\partial\boldsymbol{\beta}$.

The argument for consistency of $\mathbf{b}$ is based on the Slutsky Theorem, D.12 as we treat $\mathbf{b}$ as a function of $\mathbf{c}$ and the data. We require, as usual, well-behaved pseudoregressors. As long as $\mathbf{c}$ is consistent for $\boldsymbol{\gamma}$, the large-sample behavior of the estimator of $\boldsymbol{\beta}$ conditioned on $\mathbf{c}$ is the same as that conditioned on $\boldsymbol{\gamma}$, that is, as if $\boldsymbol{\gamma}$ were known. Asymptotic normality is obtained along similar lines (albeit with greater difficulty). The asymptotic covariance matrix for the two-step estimator is provided by the following theorem.

THEOREM 9.4 Asymptotic Distribution of the Two-Step Nonlinear Least Squares Estimator [Murphy and Topel (1985)]

Under the standard conditions assumed for the nonlinear least squares estimator, the second-step estimator of $\boldsymbol{\beta}$ is consistent and asymptotically normally distributed with asymptotic covariance matrix

$$\mathbf{V}_b^* = \sigma^2\mathbf{V}_b + \mathbf{V}_b[\mathbf{C}\mathbf{V}_c\mathbf{C}' - \mathbf{C}\mathbf{V}_c\mathbf{R}' - \mathbf{R}\mathbf{V}_c\mathbf{C}']\mathbf{V}_b,$$

THEOREM 9.4 (Continued)
where

$$\mathbf{C} = n \operatorname{plim} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i^0 \hat{\boldsymbol{\varepsilon}}_i^2 \left(\frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta}, \mathbf{w}_i, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}'} \right)$$

and

$$\mathbf{R} = n \operatorname{plim} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i^0 \hat{\boldsymbol{\varepsilon}}_i \left(\frac{\partial g(\mathbf{w}_i, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}'} \right).$$

The function $\partial g(.)/\partial \boldsymbol{\gamma}$ in the definition of **R** *is the gradient of the ith term in the log-likelihood function if $\boldsymbol{\gamma}$ is estimated by maximum likelihood. (The precise form is shown below.) If $\boldsymbol{\gamma}$ appears as the parameter vector in a regression model,*

$$z_i = f(\mathbf{w}_i, \boldsymbol{\gamma}) + u_i, \tag{9-32}$$

then $\partial g(.)/\partial \boldsymbol{\gamma}$ will be a derivative of the sum of squared deviations function,

$$\frac{\partial g(.)}{\partial \boldsymbol{\gamma}} = u_i \frac{\partial f(\mathbf{w}_i, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}.$$

If this is a linear regression, then the derivative vector is just $\mathbf{w}_i$.

Implementation of the theorem requires that the asymptotic covariance matrix computed as usual for the second-step estimator based on **c** instead of the true $\boldsymbol{\gamma}$ must be corrected for the presence of the estimator **c** in **b**.

Before developing the application, we note how some important special cases are handled. If $\boldsymbol{\gamma}$ enters $h(.)$ as the coefficient vector in a prediction of another variable in a regression model, then we have the following useful results.

Case 1 Linear regression models. If $h(.) = \mathbf{x}_i'\boldsymbol{\beta} + \delta E[z_i \mid \mathbf{w}_i] + \varepsilon_i$, where $E[z_i \mid \mathbf{w}_i] = \mathbf{w}_i'\boldsymbol{\gamma}$, then the two models are just fit by linear least squares as usual. The regression for y includes an additional variable, $\mathbf{w}_i'\mathbf{c}$. Let d be the coefficient on this new variable. Then

$$\hat{\mathbf{C}} = d \sum_{i=1}^{n} e_i^2 \mathbf{x}_i \mathbf{w}_i'$$

and

$$\hat{\mathbf{R}} = \sum_{i=1}^{n} (e_i u_i) \mathbf{x}_i \mathbf{w}_i'.$$

Case 2 Uncorrelated linear regression models. In Case 1, if the two regression disturbances are uncorrelated, then $\mathbf{R} = \mathbf{0}$.

Case 2 is general. The terms in **R** vanish asymptotically if the regressions have uncorrelated disturbances, whether either or both of them are linear. This situation will be quite common.

Case 3 Prediction from a nonlinear model. In Cases 1 and 2, if $E[z_i \mid \mathbf{w}_i]$ is a nonlinear function rather than a linear function, then it is only necessary to change $\mathbf{w}_i$ to $\mathbf{w}_i^0 = \partial E[z_i \mid \mathbf{w}_i]/\partial \boldsymbol{\gamma}$—a vector of pseudoregressors—in the definitions of **C** and **R**.

Case 4 Subset of regressors. In case 2 (but not in case 1), if **w** contains all the variables that are in **x**, then the appropriate estimator is simply

$$\mathbf{V}_b^* = s_e^2 \left(1 + \frac{c^2 s_u^2}{s_e^2}\right) (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1},$$

where $\mathbf{X}^*$ includes all the variables in **x** as well as the prediction for z.

All these cases carry over to the case of a nonlinear regression function for y. It is only necessary to replace $\mathbf{x}_i$, the actual regressors in the linear model, with $\mathbf{x}_i^0$, the pseudoregressors.

9.5.3 TWO-STEP ESTIMATION OF A CREDIT SCORING MODEL

Greene (1995c) estimates a model of consumer behavior in which the dependent variable of interest is the number of major derogatory reports recorded in the credit history of a sample of applicants for a type of credit card. In fact, this particular variable is one of the most significant determinants of whether an application for a loan or a credit card will be accepted. This dependent variable y is a discrete variable that at any time, for most consumers, will equal zero, but for a significant fraction who have missed several revolving credit payments, it will take a positive value. The typical values are zero, one, or two, but values up to, say, 10 are not unusual. This count variable is modeled using a Poisson regression model. This model appears in Sections B.4.8, 22.2.1, 22.3.7, and 21.9. The probability density function for this discrete random variable is

$$\text{Prob}[y_i = j] = \frac{e^{-\lambda_i}\lambda_i^j}{j!}.$$

The expected value of y_i is λ_i, so depending on how λ_i is specified and despite the unusual nature of the dependent variable, this model is a linear or nonlinear regression model. We will consider both cases, the linear model $E[y_i \mid \mathbf{x}_i] = \mathbf{x}_i'\boldsymbol{\beta}$ and the more common loglinear model $E[y_i \mid \mathbf{x}_i] = e^{\mathbf{x}_i'\boldsymbol{\beta}}$, where $\mathbf{x}_i$ might include such covariates as age, income, and typical monthly credit account expenditure. This model is usually estimated by maximum likelihood. But since it is a bona fide regression model, least squares, either linear or nonlinear, is a consistent, if inefficient, estimator.

In Greene's study, a secondary model is fit for the outcome of the credit card application. Let z_i denote this outcome, coded 1 if the application is accepted, 0 if not. For purposes of this example, we will model this outcome using a **logit** model (see the extensive development in Chapter 21, esp. Section 21.3). Thus

$$\text{Prob}[z_i = 1] = P(\mathbf{w}_i, \boldsymbol{\gamma}) = \frac{e^{\mathbf{w}_i'\boldsymbol{\gamma}}}{1 + e^{\mathbf{w}_i'\boldsymbol{\gamma}}},$$

where $\mathbf{w}_i$ might include age, income, whether the applicants own their own homes, and whether they are self-employed; these are the sorts of variables that "credit scoring" agencies examine.

Finally, we suppose that the probability of acceptance enters the regression model as an additional explanatory variable. (We concede that the power of the underlying theory wanes a bit here.) Thus, our nonlinear regression model is

$$E[y_i \mid \mathbf{x}_i] = \mathbf{x}_i'\boldsymbol{\beta} + \delta P(\mathbf{w}_i, \boldsymbol{\gamma}) \quad \text{(linear)}$$

or

$$E[y_i \mid \mathbf{x}_i] = e^{\mathbf{x}_i'\boldsymbol{\beta} + \delta P(\mathbf{w}_i, \boldsymbol{\gamma})} \quad \text{(loglinear, nonlinear)}.$$

The two-step estimation procedure consists of estimation of $\boldsymbol{\gamma}$ by maximum likelihood, then computing $\hat{P}_i = P(\mathbf{w}_i, \mathbf{c})$, and finally estimating by either linear or nonlinear least squares $[\boldsymbol{\beta}, \delta]$ using $\hat{P}_i$ as a constructed regressor. We will develop the theoretical background for the estimator and then continue with implementation of the estimator.

For the Poisson regression model, when the conditional mean function is linear, $\mathbf{x}_i^0 = \mathbf{x}_i$. If it is loglinear, then

$$\mathbf{x}_i^0 = \partial\lambda_i/\partial\boldsymbol{\beta} = \partial \exp(\mathbf{x}_i'\boldsymbol{\beta})/\partial\boldsymbol{\beta} = \lambda_i \mathbf{x}_i,$$

which is simple to compute. When $P(\mathbf{w}_i, \boldsymbol{\gamma})$ is included in the model, the pseudoregressor vector $\mathbf{x}_i^0$ includes this variable and the coefficient vector is $[\boldsymbol{\beta}, \delta]$. Then

$$\hat{\mathbf{V}}_b = \frac{1}{n}\sum_{i=1}^{n}[y_i - h(\mathbf{x}_i, \mathbf{w}_i, \mathbf{b}, \mathbf{c})]^2 \times (\mathbf{X}^{0\prime}\mathbf{X}^0)^{-1},$$

where $\mathbf{X}^0$ is computed at $[\mathbf{b}, d, \mathbf{c}]$, the final estimates.

For the logit model, the gradient of the log-likelihood and the estimator of $\mathbf{V}_c$ are given in Section 21.3.1. They are

$$\partial \ln f(z_i \mid \mathbf{w}_i, \boldsymbol{\gamma})/\partial\boldsymbol{\gamma} = [z_i - P(\mathbf{w}_i, \boldsymbol{\gamma})]\mathbf{w}_i$$

and

$$\hat{\mathbf{V}}_c = \left[\sum_{i=1}^{n}[z_i - P(\mathbf{w}_i, \hat{\boldsymbol{\gamma}})]^2 \mathbf{w}_i \mathbf{w}_i'\right]^{-1}.$$

Note that for this model, we are actually inserting a prediction from a regression model of sorts, since $E[z_i \mid \mathbf{w}_i] = P(\mathbf{w}_i, \boldsymbol{\gamma})$. To compute $\mathbf{C}$, we will require

$$\partial h(.)/\partial\boldsymbol{\gamma} = \lambda_i \delta\, \partial P_i/\partial\boldsymbol{\gamma} = \lambda_i \delta P_i(1 - P_i)\mathbf{w}_i.$$

The remaining parts of the corrected covariance matrix are computed using

$$\hat{\mathbf{C}} = \sum_{i=1}^{n}\left(\hat{\lambda}_i \hat{\mathbf{x}}_i^0 \hat{\varepsilon}_i^2\right)[\hat{\lambda}_i d \hat{P}_i(1 - \hat{P}_i)]\mathbf{w}_i'$$

and

$$\hat{\mathbf{R}} = \sum_{i=1}^{n}\left(\hat{\lambda}_i \hat{\mathbf{x}}_i^0 \hat{\varepsilon}_i\right)(z_i - \hat{P}_i)\mathbf{w}_i'.$$

(If the regression model is linear, then the three occurrences of λ_i are omitted.)

TABLE 9.4 Two-Step Estimates of a Credit Scoring Model

	Step 1. $P(w_i, \gamma)$		*Step 2.* $\mathrm{E}[y_i \mid x_i] = x_i'\beta + \delta P_i$			*Step 2.* $\mathrm{E}[y_i \mid x_i] = e^{x_i'\beta + \delta P_i}$		
Variable	*Est.*	*St.Er.*	*Est.*	*St.Er.**	*St.Er.**	*Est.*	*St.Er.*	*Se.Er.**
Constant	2.7236	1.0970	−1.0628	1.1907	1.2681	−7.1969	6.2708	49.3854
Age	−0.7328	0.02961	0.021661	0.018756	0.020089	0.079984	0.08135	0.61183
Income	0.21919	0.14296	0.03473	0.07266	0.082079	−0.1328007	0.21380	1.8687
Self-empl	−1.9439	1.01270						
Own Rent	0.18937	0.49817						
Expend			−0.000787	0.000368	0.000413	−0.28008	0.96429	0.96969
$\boldsymbol{P(w_i, \gamma)}$			1.0408	1.0653	1.177299	6.99098	5.7978	49.34414
ln L	−53.925							
$\mathbf{e'e}$			95.5506			80.31265		
s			0.977496			0.89617		
R^2			0.05433			0.20514		
Mean	0.73		0.36			0.36		

Data used in the application are listed in Appendix Table F9.1. We use the following model:

$$\text{Prob}[z_i = 1] = P(\text{age, income, own rent, self-employed}),$$

$$E[y_i] = h(\text{age, income, expend}).$$

We have used 100 of the 1,319 observations used in the original study. Table 9.4 reports the results of the various regressions and computations. The column denoted St.Er.* contains the corrected standard error. The column marked St.Er. contains the standard errors that would be computed ignoring the two-step nature of the computations. For the linear model, we used $\mathbf{e'e}/n$ to estimate σ^2.

As expected, accounting for the variability in $\mathbf{c}$ increases the standard errors of the second-step estimator. The linear model appears to give quite different results from the nonlinear model. But this can be deceiving. In the linear model, $\partial E[y_i \mid \mathbf{x}_i, P_i]/\partial \mathbf{x}_i = \boldsymbol{\beta}$ whereas in the nonlinear model, the counterpart is not $\boldsymbol{\beta}$ but $\lambda_i \boldsymbol{\beta}$. The value of λ_i at the mean values of all the variables in the second-step model is roughly 0.36 (the mean of the dependent variable), so the marginal effects in the nonlinear model are [0.0224, −0.0372, −0.07847, 1.9587], respectively, including P_i but not the constant, which are reasonably similar to those for the linear model. To compute an asymptotic covariance matrix for the estimated marginal effects, we would use the delta method from Sections D.2.7 and D.3.1. For convenience, let $\mathbf{b}_p = [\mathbf{b}', d]'$, and let $\mathbf{v}_i = [\mathbf{x}_i', \hat{P}_i]'$, which just adds P_i to the regressor vector so we need not treat it separately. Then the vector of marginal effects is

$$\mathbf{m} = \exp(\mathbf{v}_i'\mathbf{b}_p) \times \mathbf{b}_p = \lambda_i \mathbf{b}_p.$$

The matrix of derivatives is

$$\mathbf{G} = \partial \mathbf{m}/\partial \mathbf{b}_p = \lambda_i(\mathbf{I} + \mathbf{b}_p \mathbf{v}_i'),$$

so the estimator of the asymptotic covariance matrix for $\mathbf{m}$ is

$$\text{Est.Asy. Var}[\mathbf{m}] = \mathbf{G}\mathbf{V}_b^*\mathbf{G}'.$$

TABLE 9.5 Maximum Likelihood Estimates of Second-Step Regression Model

	Constant	*Age*	*Income*	*Expend*	*P*
Estimate	−6.3200	0.073106	0.045236	−0.00689	4.6324
Std.Error	3.9308	0.054246	0.17411	0.00202	3.6618
Corr.Std.Error	9.0321	0.102867	0.402368	0.003985	9.918233

One might be tempted to treat λ_i as a constant, in which case only the first term in the quadratic form would appear and the computation would amount simply to multiplying the asymptotic standard errors for $\mathbf{b}_p$ by λ_i. This approximation would leave the asymptotic t ratios unchanged, whereas making the full correction will change the entire covariance matrix. The approximation will generally lead to an understatement of the correct standard errors.

Finally, although this treatment is not discussed in detail until Chapter 18, we note at this point that nonlinear least squares is an inefficient estimator in the Poisson regression model; maximum likelihood is the preferred, efficient estimator. Table 9.5 presents the maximum likelihood estimates with both corrected and uncorrected estimates of the asymptotic standard errors of the parameter estimates. (The full discussion of the model is given in Section 21.9.) The corrected standard errors are computed using the methods shown in Section 17.7. A comparison of these estimates with those in the third set of Table 9.4 suggests the clear superiority of the maximum likelihood estimator.

9.6 SUMMARY AND CONCLUSIONS

In this chapter, we extended the regression model to a form which allows nonlinearity in the parameters in the regression function. The results for interpretation, estimation, and hypothesis testing are quite similar to those for the linear model. The two crucial differences between the two models are, first, the more involved estimation procedures needed for the nonlinear model and, second, the ambiguity of the interpretation of the coefficients in the nonlinear model (since the derivatives of the regression are often nonconstant, in contrast to those in the linear model.) Finally, we added two additional levels of generality to the model. A nonlinear instrumental variables estimator is suggested to accommodate the possibility that the disturbances in the model are correlated with the included variables. In the second application, two-step nonlinear least squares is suggested as a method of allowing a model to be fit while including functions of previously estimated parameters.

Key Terms and Concepts

- Box–Cox transformation
- Consistency
- Delta method
- GMM estimator
- Identification
- Instrumental variables estimator
- Iteration
- Linearized regression model
- LM test
- Logit
- Multicollinearity
- Nonlinear model
- Normalization
- Orthogonality condition
- Overidentifying restrictions
- P_E test
- Pseudoregressors
- Semiparametric
- Starting values
- Translog
- Two-step estimation
- Wald test