Generalized Added-Variable Plots

John Luke Gallup *

January 4, 2024

Abstract

An added-variable plot shows the correlation between an outcome variable and an explanatory variable *conditional on* other explanatory variables. Although usually applied to OLS regression, this paper extends the method to a broad class of linear and nonlinear estimators including generalized least squares, instrumental variables, maximum likelihood and generalized methods of moments estimators. Added-variable plots show the contribution of each data observation to the estimated correlation, providing an intuitive presentation of complex estimation results whose meaning is often opaque to the uninitiated.

1 Introduction

An added-variable plot displays the impact of each observation of a particular explanatory variable on the predicted line of fit on an outcome variable. If there were only one explanatory variable, the correlation could be shown by a simple scatter plot with a prediction line of the outcome versus the explanatory variable. With multiple explanatory variables, though, the scatter plot is an unreliable representation of this relationship, because it does not condition on the other explanatory variables. The added-variable plot properly shows the conditional correlation. The added-variable plot is sometimes called a partial regression plot (Belsley et al., 1980).

The added-variable plot present the results of multivariate estimation in a more intuitive way than tables of coefficient estimates. The plots are particularly helpful for presenting the results of complex estimation procedures, making them more intelligible to those who haven't studied them before. The plot shows how closely the observations hew to the predicted fit and serves as a visual diagnostic for the influence of outlier observations. A confidence interval shows the statistical significance of the correlation. If the confidence interval includes a zero slope, the coefficient is not significantly different from zero.

*Department of Economics, Portland State University (email: jlgallup@pdx.edu). I gratefully acknowledge helpful comments from Mead Over and Stephen Portnoy.

This paper extends the theory of added-variable plots to most commonly used estimation methods: generalized least squares (GLS) , instrumental variables (IV), linear and nonlinear maximum likelihood (MLE), nonlinear least squares (NLS) and generalized method of moments (GMM).

Although the origin of the added-variable plot is somewhat unclear,[1] it follows from the logic of the Frisch–Waugh–Lovell theorem (Frisch & Waugh, 1933; Lovell, 1963). Developed in the context of detrending time series variables, the Frisch–Waugh–Lovell theorem shows that the OLS estimate of the coefficient on one variable is equal to regressing that variable and the outcome variable on all the other explanatory variables first, and then regressing the residuals from those regressions on each other (Davidson & MacKinnon, 1993, Section 1.4).

Added-variable plots have previously been extended to the generalized linear model, a limited class of maximum likelihood estimators, by Wang (1985), Pregibon (1985), and Hines and Carter (1993).

Figure 1 is an example of an added-variable plot displaying the results of a GMM dynamic panel estimation. It shows the partial correlation of one explanatory variable, the average early childhood health conditions of workers, with the

[1] Chien (2011) attributes added-variable plots to Mosteller and Tukey (1977, 12C. Graphical Fitting by Stages), proposed as an aid in deciding whether to include additional variables in a regression. Cook and Weisberg (1982) notes that added-variable plots for the analysis of residuals to diagnose the influence of outlier observations are discussed in Draper and Smith (1966), Belsley et al. (1980), and Weisberg (1980).

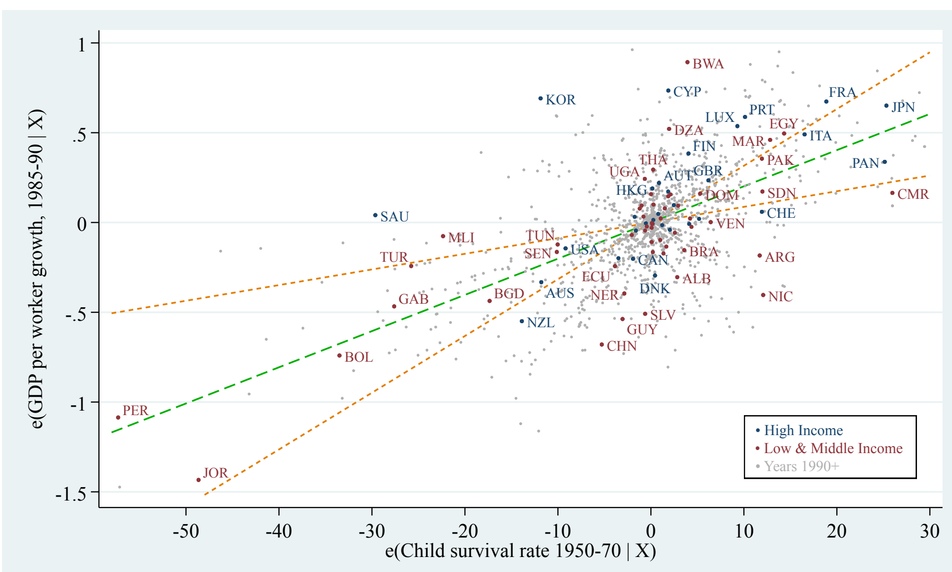

Figure 1: An added-variable plot of GMM dynamic panel estimation:
Economic growth versus child survival

output variable, economic growth per worker, from Gallup (2023). The estimation model includes an instrumented lagged dependent variable, fixed effects, time effects and four other explanatory variables.

The added-variable method transforms the displayed variables to remove the influence of additional explanatory variables from both the explanatory and outcome variables (indicated by "e(· | X)" in the axis titles, where X refers to the other explanatory variables). The green dashed line has a slope equal to the estimated coefficient on the child survival rate. The two orange shorter-dashed lines are the 95% confidence limits. Because the confidence interval in this graph does not include a zero slope, the coefficient estimate is statistically significant. The unfamiliar shape of the confidence interval is discussed in the next section.

The added-variable plot produces a scatter plot and fitted line of the relationship between an explanatory variable and the outcome variable, after accounting for the influence of the other explanatory variables. The graph of transformed added-variable observations conveys the strength of the partial correlation between the explanatory and outcome variables which would be harder to grasp in a table of coefficients.

2 Derivation of Added-Variable Plots

2.1 Ordinary Least Squares

Ordinary least squares (OLS) regression estimates the parameters of a linear relationship between an outcome variable and multiple explanatory variables. Let the outcome variable $\mathbf{y}$ be an $n \times 1$ vector, the explanatory variables $\mathbf{X}$ be an $n \times k$ matrix. Let the unknown parameters $\boldsymbol{\beta}$ be an $k \times 1$ vector and a random error $\boldsymbol{\varepsilon}$ be an $n \times 1$ vector. The estimation equation is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

The OLS estimator $\mathbf{b}$ of $\boldsymbol{\beta}$ is

$$\mathbf{b} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}. \tag{1}$$

We can partition the $\mathbf{X}$ matrix into $\mathbf{X} = [\mathbf{x}_1 \mathbf{X}_2]$ where $\mathbf{X}_2 = [\mathbf{x}_2 \cdots \mathbf{x}_k]$, partition $\mathbf{b}$ into $\mathbf{b} = \begin{bmatrix} b_1 \\ \mathbf{b}_2 \end{bmatrix}$ where $\mathbf{b}_2 = \begin{bmatrix} b_2 \\ \vdots \\ b_k \end{bmatrix}$. After partitioning, we rewrite Equation (1) as

$$\begin{bmatrix} b_1 \\ \mathbf{b}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^\top \mathbf{x}_1 & \mathbf{x}_1^\top \mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{x}_1 & \mathbf{X}_2^\top \mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{x}_1^\top \mathbf{y} \\ \mathbf{X}_2^\top \mathbf{y} \end{bmatrix}.$$

After some manipulation, $b_1 = (\mathbf{x}_1^\top \mathbf{M}_2 \mathbf{x}_1)^{-1} \mathbf{x}_1^\top \mathbf{M}_2 \mathbf{y}$ where $\mathbf{M}_2 = (\mathbf{I} - \mathbf{X}_2 (\mathbf{X}_2^\top \mathbf{X}_2)^{-1} \mathbf{X}_2^\top)$. Since $\mathbf{M}_2$ is symmetric and idempotent, we can rewrite b_1 as

$$b_1 = (\mathbf{x}_1^\top \mathbf{M}_2^\top \mathbf{M}_2 \mathbf{x}_1)^{-1} \mathbf{x}_1^\top \mathbf{M}_2^\top \mathbf{M}_2 \mathbf{y} = (\mathbf{e}_{\mathbf{x}_1}^\top \mathbf{e}_{\mathbf{x}_1})^{-1} \mathbf{e}_{\mathbf{x}_1}^\top \mathbf{e}_{\mathbf{y}} \tag{2}$$

where $\mathbf{e_{x_1}} = \mathbf{M}_2\mathbf{x}_1$ and $\mathbf{e_y} = \mathbf{M}_2\mathbf{y}$.

Inspection of the equation for "residual-maker" $\mathbf{M}_2$ shows that $\mathbf{e_y} = \mathbf{M}_2\mathbf{y}$ is the vector of residuals from the regression of $\mathbf{y}$ on $\mathbf{X}_2$ and likewise $\mathbf{e_{x_1}} = \mathbf{M}_2\mathbf{x}_1$ is the vector of residuals from the regression of $\mathbf{x}_1$ on $\mathbf{X}_2$. Equation (2) is the conclusion of the Frisch–Waugh–Lovell theorem (Frisch & Waugh, 1933; Lovell, 1963), that regressing the residuals of $\mathbf{y}$ on $\mathbf{X}_2$ on the residuals of $\mathbf{x}_1$ on $\mathbf{X}_2$ produces the OLS coefficient estimate for b_1.

$\mathbf{e_y}$ and $\mathbf{e_{x_1}}$ can be interpreted as $\mathbf{y}$ and $\mathbf{x}_1$ purged of the influence of the $\mathbf{X}_2$ variables. $\mathbf{e_y} = \mathbf{y} - \hat{\mathbf{y}}_{\mathbf{x}_2}$, where $\hat{\mathbf{y}}_{\mathbf{x}_2}$ is the predicted value of $\mathbf{y}$ from the regression of $\mathbf{y}$ on $\mathbf{X}_2$. That is, $\mathbf{e_y}$ is what is left over when all the variation in $\mathbf{y}$ that can be predicted by $\mathbf{X}_2$ has been removed. Similarly, $\mathbf{e_{x_1}}$ is the residual of $\mathbf{x}_1$ remaining when the variation of $\mathbf{x}_1$ that can be predicted by $\mathbf{X}_2$ has been removed. So the correlation of $\mathbf{e_y}$ and $\mathbf{e_{x_1}}$ is the partial correlation of $\mathbf{y}$ and $\mathbf{x}_1$, conditional on $\mathbf{X}_2$.

The decomposition of b_1 into $\mathbf{e_{x_1}}$ and $\mathbf{e_y}$ gives rise to the added-variable plot. A scatterplot of $\mathbf{e_{x_1}}$ versus $\mathbf{e_y}$ shows the correlation of the x_1 variable with the y variable, *controlling for the influence of the other explanatory variables* in the regression. From equation (2), we can see that the OLS estimator b_1 of β_1 is the result of regressing $\mathbf{e_y}$ on $\mathbf{e_{x_1}}$ (with no intercept term).

The predicted fit in the added-variable plot is a line connecting $\mathbf{e_{x_1}}$ and the predicted values of $\hat{\mathbf{e}}_\mathbf{y}$: $\hat{\mathbf{e}}_\mathbf{y} = b_1\mathbf{e_{x_1}}$. Thus the slope of the OLS linear fit of the data in the scatterplot of $\mathbf{e_{x_1}}$ versus $\mathbf{e_y}$ is equal to b_1, the estimated partial effect of x_1 on y.

2.2 The Confidence Interval and its Unfamiliar Shape

The variance of an observation $\hat{e}_{yi}$ of predicted $\mathbf{e_y}$ conditional on the explanatory variables $\mathbf{X}$ is $V[\hat{e}_{y_i}|\mathbf{X}] = V[b_1 e_{x_{1i}}|\mathbf{X}] = e^2_{x_{1i}} V[b_1|\mathbf{X}]$, since $\mathbf{e_{x_1}} = \mathbf{M}_2\mathbf{x}_1$ is entirely constructed from elements of $\mathbf{X}$. A confidence interval at the $1-\alpha$ level around the prediction line has boundaries at $\hat{\mathbf{e}}_\mathbf{y} \pm t_{[n-k,1-\alpha/2]}\hat{\sigma}_{b_1}|\mathbf{e_{x_1}}|$, where $t_{[n-k,1-\alpha/2]}$ is the $1-\alpha/2$ percentile of the t distribution with $n-k$ degrees of freedom and $\hat{\sigma}_{b_1}$ is the standard error of b_1.

Whether the added-variable plot confidence interval includes a slope of zero indicates whether the coefficient b_1 is statistically different from zero. The usual test for statistical significance of coefficient b_1=0 at level α has a boundary at t test statistic $|\frac{b_1-0}{\hat{\sigma}_{b_1}}| = t_{[n-k,1-\alpha/2]}$. Thus b_1 is judged not to be significantly different from zero when $|b_1| = t_{[n-k,1-\alpha/2]}\hat{\sigma}_{b_1}$ or less. When $|b_1|$ is equal to this value, the confidence interval $\hat{\mathbf{e}}_\mathbf{y} \pm t_{[n-k,1-\alpha/2]}\hat{\sigma}_{b_1}|\mathbf{e_{x_1}}| = \hat{\mathbf{e}}_\mathbf{y} \pm |b_1||\mathbf{e_{x_1}}| = \hat{\mathbf{e}}_\mathbf{y} \pm |\hat{\mathbf{e}}_\mathbf{y}|$, which includes $\hat{\mathbf{e}}_\mathbf{y} = 0$. Thus if the confidence interval does not include $\mathbf{e_y} = 0$ for $\mathbf{e_{x_1}} \neq 0$, the estimate b_1 is statistically significant at the level α. The slopes of all the straight lines in the confidence interval are the range of likely values for b_1, so if that range excludes a zero-sloped b_1, that corresponds to rejecting $\beta_1 = 0$.

The confidence interval in the added-variable plot has an unfamiliar appearance with a zero width at $e_{x_1} = 0$ and straight lines emanating from that point.

This shape is entirely due to there not being an intercept term in the regression of $\mathbf{e_y}$ on $\mathbf{e_{x_1}}$.

A regression of variable y on a single variable x *with an intercept term* has quadratic confidence interval boundaries with a confidence band always greater than zero, due the intercept. A regression without an intercept, say of $y_i - \bar{y}$ on $x_i - \bar{x}$, will have a confidence interval with the same shape as in the added-variable plot, and a confidence band of zero at $x_i - \bar{x} = 0$. At $x_i - \bar{x} = 0$, $\widehat{y_i - \bar{y}} = b_1(x_i - \bar{x}) = 0$, with a confidence interval $\widehat{y_i - \bar{y}} \pm t_{[n-1,1-\alpha/2]}\hat{\sigma}_{b_1}|x_i - \bar{x}| = 0$.

The substantive reason why the added-variable plot has a zero-width confidence interval at $e_{x_{1i}} = 0$ is that, at that point, the $\mathbf{X}_2$ variables account for *all* the variation in $\mathbf{x}_1$, resulting in the residual $e_{x_{1i}}$ equalling 0. Since x_{1i} is contributing nothing to predicted $\hat{y}_i$ at this point, the variation in $\hat{y}_i$ not accounted for by $\mathbf{X}_2$, namely $\hat{e}_{y_i}$, is known to be unambiguously zero.

2.3 Generalized Added-Variable Plots for Non-OLS Estimators

None of the non-OLS estimators in this paper have the exact form of Equation (1). However, each of the subsequent estimators can be transformed to have a parallel form of

$$\hat{\boldsymbol{\beta}} = (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\top \tilde{\mathbf{y}}. \tag{3}$$

where $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{y}}$ are transformations of the original data $\mathbf{X}$ and $\mathbf{y}$, respectively. When $\tilde{\mathbf{X}}$ is partitioned into $[\tilde{\mathbf{x}}_1 \tilde{\mathbf{X}}_2]$, the Frisch–Waugh–Lovell theorem ensures that

$$\hat{\beta}_1 = (\mathbf{e}_{\tilde{\mathbf{x}}_1}^\top \mathbf{e}_{\tilde{\mathbf{x}}_1})^{-1} \mathbf{e}_{\tilde{\mathbf{x}}_1}^\top \mathbf{e}_{\tilde{\mathbf{y}}}$$

where $\mathbf{e}_{\tilde{\mathbf{x}}_1} = \tilde{\mathbf{M}}_2 \tilde{\mathbf{x}}_1$, $\mathbf{e}_{\tilde{\mathbf{y}}} = \tilde{\mathbf{M}}_2 \mathbf{y}$ and $\tilde{\mathbf{M}}_2 = \mathbf{I} - \tilde{\mathbf{X}}_2(\tilde{\mathbf{X}}_2^\top \tilde{\mathbf{X}}_2)^{-1} \tilde{\mathbf{X}}_2^\top$.

The transformed observations in $\mathbf{e}_{\tilde{\mathbf{x}}_1}$ graphed against the observations in $\mathbf{e}_{\tilde{\mathbf{y}}}$ form the added-variable plot. A linear fit of the added-variable observations has a slope equal to the coefficient $\hat{\beta}_1$.

The fitted line in the added-variable plot and its confidence interval have formulas similar to the OLS regression case. The fitted line connects $\mathbf{e}_{\tilde{\mathbf{x}}_1}$ and the predicted values of $\hat{\mathbf{e}}_{\tilde{\mathbf{y}}} = \mathbf{e}_{\tilde{\mathbf{x}}_1} \hat{\beta}_1$.

The variance of an observation of predicted $\hat{e}_{\tilde{y}i}$ conditional on the explanatory variables $\tilde{\mathbf{X}}$ is $V[\hat{e}_{\tilde{y}_i}|\tilde{\mathbf{X}}] = V[e_{\tilde{x}_{1i}} \hat{\beta}_1|\tilde{\mathbf{X}}] = e^2_{\tilde{x}_{1i}} V[\hat{\beta}_1|\tilde{\mathbf{X}}]$, since $\mathbf{e}_{\tilde{\mathbf{x}}_1} = \tilde{\mathbf{M}}_2 \tilde{\mathbf{x}}_1$ is a linear combination of the columns of $\tilde{\mathbf{X}}$.

A confidence interval around the prediction line has boundaries at $\hat{\mathbf{e}}_{\tilde{\mathbf{y}}} \pm z_{[1-\alpha/2]} \hat{\sigma}_{\hat{\beta}_1} |\mathbf{e}_{\tilde{\mathbf{x}}_1}|$, where $z_{[1-\alpha/2]}$ is the $1-\alpha/2$ percentile of the cumulative standard normal distribution and $\hat{\sigma}_{\hat{\beta}_1}$ is the asymptotic standard error of $\hat{\beta}_1$.

3 Linear Estimators

Added-variable plots can be constructed for linear estimators other than OLS: generalized least squares (GLS), and two-stage least squares (2SLS) and three-stage least squares (3SLS) instrumental variables estimators. Linear generalized method of moments (GMM) estimators are addressed as a special case of non-linear GMM in Section 4.3 below.

3.1 Generalized Least Squares

The OLS estimator is efficient when the error term is homoscedastic and non-autocorrelated, with a variance of $V[\boldsymbol{\varepsilon}] = \sigma^2 \mathbf{I}$. For a general heteroscedastic and autocorrelated error process with variance $V[\boldsymbol{\varepsilon}] = \boldsymbol{\Sigma}$, generalized least squares is the efficient linear estimator:

$$\hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y}.$$

Because $\boldsymbol{\Sigma}$ is a variance matrix, it is symmetric and positive definite with a symmetric positive definite square root $\boldsymbol{\Sigma}^{-1/2}$. Defining $\tilde{\mathbf{X}} = \hat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{X}$ and $\tilde{\mathbf{y}} = \hat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{y}$,, where $\hat{\boldsymbol{\Sigma}}$ is a consistent estimate of $\boldsymbol{\Sigma}$, then the Feasible Generalized Least Squares (FGLS) estimator is

$$\hat{\boldsymbol{\beta}}_{FGLS} = (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\top \tilde{\mathbf{y}}.$$

An added variable plot can be constructed graphing $\mathbf{e}_{\tilde{\mathbf{x}}_1}$ versus $\mathbf{e}_{\tilde{\mathbf{y}}}$ for any variable in $\mathbf{X}$, as in Section 2.3.

3.2 Two-stage and Three-stage Least Squares

3.2.1 Two-stage least squares

Instrumental variables (IV) is a frequently employed estimation method in econometrics to address endogenous explanatory variables where $E[\boldsymbol{\varepsilon}|\mathbf{X}] \neq \mathbf{0}$, violating OLS assumptions (*e.g.* Davidson & MacKinnon, 1993, chapter 7). If "instrumental" variables $\mathbf{Z}$ can be found which are correlated with the endogenous explanatory variables $\mathbf{X}$ but uncorrelated with the errors $\boldsymbol{\varepsilon}$, they can be employed to make consistent estimates of the parameters. Assume there is an $n \times k$ matrix of explanatory variables $\mathbf{X}$, an $n \times l$ matrix of instrumental variables $\mathbf{Z}$ where $l \geq k$,

$$E[\boldsymbol{\varepsilon}|\mathbf{Z}] = \mathbf{0},$$
$$\text{plim}(\tfrac{1}{n}\mathbf{Z}'\mathbf{Z}) \text{ is finite and positive definite, and}$$
$$\text{plim}(\tfrac{1}{n}\mathbf{Z}'\mathbf{X}) \text{ is finite and has rank } k.$$

The first stage regression of $\mathbf{X}$ on $\mathbf{Z}$ produces predicted $\tilde{\mathbf{X}}$,

$$\tilde{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{X}$$

which is asymptotically uncorrelated with the errors because it is a projection of $\mathbf{X}$ onto the span of $\mathbf{Z}$. This enables a second stage consistent estimator

$$\hat{\boldsymbol{\beta}}_{2SLS} = (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\top \mathbf{y}.$$

This two-stage least squares (2SLS) equation can be used to produce added-variable plots for any of the explanatory variables as in Section 2.3.

3.2.2 Three-stage least squares

In a system of equations where endogenous variables in one equation are outcome variables in a different equation, some of the explanatory variables in the latter equation may serve as instruments for the endogenous variables in the former equation. When these equations are estimated separately, 2SLS can be employed. However, it is more efficient to proceed to a third stage, estimating the whole system of equations by applying feasible generalized least squares if there is cross-equation correlation of the errors. This is a three-stage least squares (3SLS) estimator (Davidson & MacKinnon, 1993, chapter 18).

For a system of M estimation equations

$$\begin{aligned} \mathbf{y}_1 &= \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1 \\ &\vdots \\ \mathbf{y}_M &= \mathbf{X}_M\boldsymbol{\beta}_M + \boldsymbol{\varepsilon}_M, \end{aligned}$$

the 2SLS estimator of each equation $m \in \{1, \ldots, M\}$ is

$$\hat{\boldsymbol{\beta}}_m = (\hat{\mathbf{X}}_m^\top \hat{\mathbf{X}}_m)^{-1}\hat{\mathbf{X}}_m^\top \mathbf{y}_m$$

where

$$\hat{\mathbf{X}}_m = \mathbf{Z}_m(\mathbf{Z}_m^\top \mathbf{Z}_m)^{-1}\mathbf{Z}_m^\top \mathbf{X}_m,$$

$\mathbf{X}_m$ are potentially endogenous explanatory variables in equation m, and $\mathbf{Z}_m$ are valid instruments, typically found in the equations other than m. To apply cross-equation FGLS, construct

$$\hat{\mathbf{X}} = \begin{bmatrix} \hat{\mathbf{X}}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \hat{\mathbf{X}}_M \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_M \end{bmatrix} \quad \text{and} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_M \end{bmatrix}.$$

A consistent estimate $\hat{\boldsymbol{\Sigma}}$ of $V[\boldsymbol{\varepsilon}]$ can be calculated from the 2SLS residuals. Then the efficient 3SLS estimator is

$$\hat{\boldsymbol{\beta}}_{3SLS} = (\hat{\mathbf{X}}^\top \hat{\boldsymbol{\Sigma}}^{-1}\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}^\top \hat{\boldsymbol{\Sigma}}^{-1}\mathbf{y}.$$

Defining

$$\tilde{\mathbf{X}} = \hat{\boldsymbol{\Sigma}}^{-1/2}\hat{\mathbf{X}} \quad \text{and} \quad \tilde{\mathbf{y}} = \hat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{y}$$

makes it possible to write

$$\hat{\boldsymbol{\beta}}_{3SLS} = (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top \tilde{\mathbf{y}}.$$

This formula for $\hat{\boldsymbol{\beta}}_{3SLS}$ can produce an added-variable plot of any explanatory variable in any of the M equations.

4 Nonlinear Estimators

Added-variable plots can be applied to extremum estimators, a wide class which includes maximum likelihood, nonlinear least squares and generalized method of moments estimators (Amemiya, 1985; Newey & McFadden, 1994). Each of these nonlinear estimators can be expressed in linear form in terms of transformed variables $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{y}}$ of Section 2.3.

Extremum estimators maximize a criterion function $q(\cdot)$ which depends on the parameters conditioned on the observed data:

$$\max_{\boldsymbol{\beta}} q(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X})$$

In general, the extremum estimator criterion function $q(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X})$ can depend on other parameters besides $\boldsymbol{\beta}$, the parameters for the conditional mean of $\mathbf{y}$, but they are not the focus of added-variable plots. Since added-variable plots are post-estimation commands at which point we have consistent estimates of all parameters, we treat any other parameters as known constants and replace their true values with the consistent estimates in the formulas.

The first order conditions for maximization implicitly define the estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$:

$$\mathbf{s}(\hat{\boldsymbol{\beta}}) \equiv \frac{\partial q(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X})}{\partial \boldsymbol{\beta}} = \mathbf{0}$$

Taking a first-order Taylor series expansion of $\mathbf{s}(\hat{\boldsymbol{\beta}})$ around the true parameter values $\boldsymbol{\beta}_0$ and applying the Mean Value Theorem,

$$\mathbf{s}(\hat{\boldsymbol{\beta}}) = \mathbf{s}(\boldsymbol{\beta}_0) + \mathbf{H}(\boldsymbol{\beta}^*) \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)$$

where $\mathbf{H}(\boldsymbol{\beta}^*) \equiv \frac{\partial \mathbf{s}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \mid_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}$ for some $\boldsymbol{\beta}^* \in [\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0]$. Since $\mathbf{s}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$,

$$\boldsymbol{\beta}_0 = \mathbf{H}(\boldsymbol{\beta}^*)^{-1}\left(\mathbf{H}(\boldsymbol{\beta}^*)\hat{\boldsymbol{\beta}} + \mathbf{s}(\boldsymbol{\beta}_0)\right) \tag{4}$$

In the context of creating an added-variable plot, we already have the estimated $\hat{\boldsymbol{\beta}}$ in hand, but we would like to linearize its formula. Equation 4 still holds when $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^* = \boldsymbol{\beta}_0$. When this occurs, it becomes

$$\hat{\boldsymbol{\beta}} = \hat{\mathbf{H}}^{-1}(\hat{\mathbf{H}}\hat{\boldsymbol{\beta}} + \hat{\mathbf{s}}) \tag{5}$$

where $\hat{\mathbf{H}} \equiv \mathbf{H}(\hat{\boldsymbol{\beta}})$ and $\hat{\mathbf{s}} \equiv \mathbf{s}(\hat{\boldsymbol{\beta}})$. For this equation to be graphed as an added-variable plot, it needs to have the linear form

$$\hat{\boldsymbol{\beta}} = (\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top\tilde{\mathbf{y}}. \tag{6}$$

Equation 5 can have this form if $\hat{\mathbf{H}}$ can be factored such that $\hat{\mathbf{H}} = \tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}$ and $\hat{\mathbf{H}}\hat{\boldsymbol{\beta}} + \hat{\mathbf{s}}$ can be factored such that $\hat{\mathbf{H}}\hat{\boldsymbol{\beta}} + \hat{\mathbf{s}} = \tilde{\mathbf{X}}^\top\tilde{\mathbf{y}}$. Note the difference in dimensions between equations (5) and (6). $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{s}}$ are $k \times 1$ and $\hat{\mathbf{H}}$ is $k \times k$, but $\tilde{\mathbf{X}}$ is $n \times k$ and $\tilde{\mathbf{y}}$ is $n \times 1$.

Once $\hat{\boldsymbol{\beta}}$ is decomposed into a linear combination of $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{y}}$, added-variable plots can be created as in Section 2.3.

4.1 Maximum Likelihood Estimation

To derive added-variable plots for maximum likelihood estimation, we need an explicit functional form for the random distribution assumed in the statistical model. With little loss of generality, we limit consideration to the exponential family of distributions (*e.g.* Dasgupta, 2011; "Exponential Family", 2022), which includes all of the distributions commonly used in maximum likelihood estimation.[2]

A distribution in the exponential family has a probability density function (or a probability mass function in the case of a discrete distrbution) in the form of

$$f_Y(\mathbf{y}|\boldsymbol{\theta}) = \exp\Big(\boldsymbol{\theta}^\top \mathbf{t}(\mathbf{y}) - a(\boldsymbol{\theta}) + b(\mathbf{y})\Big)$$

for a random variable $\mathbf{y}$ with parameters $\boldsymbol{\theta}$ and known functions $\mathbf{t}(\cdot), a(\cdot),$ and $b(\cdot)$. $\mathbf{t} : \mathbb{R}^n \to \mathbb{R}^n$ ($\mathbf{t}$ has the same $n \times 1$ dimension as $\mathbf{y}$) and $a, b : \mathbb{R}^n \to \mathbb{R}$. The corresponding log likelihood function is

$$\ell(\boldsymbol{\theta}|\mathbf{y}) = \boldsymbol{\theta}^\top \mathbf{t}(\mathbf{y}) - a(\boldsymbol{\theta}) + b(\mathbf{y}).$$

with derivatives

$$\frac{\partial \ell}{\partial \boldsymbol{\theta}} = \mathbf{t}(\mathbf{y}) - \frac{\partial a(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

and

$$\frac{\partial^2 \ell}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = \frac{\partial^2 a(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}.$$

It can be shown that

$$\begin{aligned} \boldsymbol{\mu} &\equiv E[\mathbf{t}(\mathbf{y})] = \frac{\partial a(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ \boldsymbol{\Sigma} &\equiv Var[\mathbf{t}(\mathbf{y})] = \frac{\partial^2 a(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \end{aligned}$$

(Dasgupta, 2011, p. 597, Theorem 18.7).

Hence,

$$\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}^\top} = \boldsymbol{\Sigma} \quad \text{and} \quad \frac{\partial \boldsymbol{\theta}^\top}{\partial \boldsymbol{\mu}} = \boldsymbol{\Sigma}^{-1}.$$

Let $\boldsymbol{\mu}$, the mean of $\mathbf{t}(\mathbf{y})$, be a function of parameters $\boldsymbol{\beta}$ and explanatory variables $\mathbf{X}$:

$$\boldsymbol{\mu} = \boldsymbol{\mu}(\mathbf{X}, \boldsymbol{\beta})$$

[2] The exponential family includes the normal, binomial, multinomial, chi-squared, Gamma, Beta, exponential, Wishart, Poisson, negative binomial and Weibull distributions, among others. The familiar distributions it excludes are the uniform, t, F and Cauchy as well as non-parametric distributions. Under certain conditions, the Darmois-Koopman-Pitman theorem proves that the exponential family is made up of the distributions having a sufficient statistic (Andersen, 1970), denoted by $\mathbf{t}(\mathbf{y})$ in the probability density function.

where $\boldsymbol{\mu}(\cdot)$ is a known function, $\boldsymbol{\beta}$ is a $k \times 1$ vector and $\mathbf{X}$ is an $n \times k$ matrix. Then the log likelihood $\ell(\boldsymbol{\theta}|\mathbf{y})$ conditional on the parameters in $\boldsymbol{\theta}$ corresponds to the M-estimator criterion function $q(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta})$ in the previous section.

Denoting $\boldsymbol{\mathcal{M}} \equiv \frac{\partial \boldsymbol{\mu}(\mathbf{X},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top}$ ($\boldsymbol{\mathcal{M}}$ is upper case $\boldsymbol{\mu}$), the log likelihood score with respect to $\boldsymbol{\beta}$ is

$$\begin{aligned}\mathbf{s}(\boldsymbol{\beta}) \equiv \frac{\partial \ell}{\partial \boldsymbol{\beta}} &= \frac{\partial \boldsymbol{\mu}^\top}{\partial \boldsymbol{\beta}} \frac{\partial \boldsymbol{\theta}^\top}{\partial \boldsymbol{\mu}} \frac{\partial \ell}{\partial \boldsymbol{\theta}} \\ &= \boldsymbol{\mathcal{M}}^\top \boldsymbol{\Sigma}^{-1}(\mathbf{t}(\mathbf{y}) - \boldsymbol{\mu}).\end{aligned}$$

Noting that

$$\frac{\partial \boldsymbol{\theta}^\top}{\partial \boldsymbol{\beta}} = \boldsymbol{\mathcal{M}}^\top \boldsymbol{\Sigma}^{-1},$$

the Hessian is

$$\begin{aligned}\mathbf{H}(\boldsymbol{\beta}) \equiv \frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} &= \frac{\partial \boldsymbol{\theta}^\top}{\partial \boldsymbol{\beta}} \frac{\partial^2 \ell}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\beta}^\top} \\ &= \boldsymbol{\mathcal{M}}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mathcal{M}} \\ &= \boldsymbol{\mathcal{M}}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mathcal{M}}.\end{aligned}$$

For a consistent estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ let $\hat{\boldsymbol{\mathcal{M}}} = \boldsymbol{\mathcal{M}}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$. Then with consistent estimate $\hat{\boldsymbol{\Sigma}}$ of $\boldsymbol{\Sigma}$,

$$\hat{\mathbf{H}} = \hat{\boldsymbol{\mathcal{M}}}^\top \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mathcal{M}}} \text{ and } \hat{\mathbf{s}} = \hat{\boldsymbol{\mathcal{M}}}^\top \hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{t}(\mathbf{y}) - \hat{\boldsymbol{\mu}}).$$

By Equation 5,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{MLE} &= \hat{\mathbf{H}}^{-1}(\hat{\mathbf{H}}\hat{\boldsymbol{\beta}} + \hat{\mathbf{s}}) \\ &= \Big(\hat{\boldsymbol{\mathcal{M}}}^\top \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mathcal{M}}})\Big)^{-1} \hat{\boldsymbol{\mathcal{M}}}^\top \hat{\boldsymbol{\Sigma}}^{-1} \Big(\hat{\boldsymbol{\mathcal{M}}}\hat{\boldsymbol{\beta}} + \mathbf{t}(\mathbf{y}) - \hat{\boldsymbol{\mu}}\Big) \\ &= (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\top \tilde{\mathbf{y}}\end{aligned}$$

where

$$\tilde{\mathbf{X}} = \hat{\boldsymbol{\Sigma}}^{-1/2} \hat{\boldsymbol{\mathcal{M}}} \text{ and } \tilde{\mathbf{y}} = \hat{\boldsymbol{\Sigma}}^{-1/2} \Big(\mathbf{t}(\mathbf{y}) + \hat{\boldsymbol{\mathcal{M}}}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\mu}}\Big). \tag{7}$$

The following subsections apply this linearization of maximum likelihood estimators to two examples of exponential family distributions: multivariate normal and binary.

4.1.1 Multivariate Normal Maximum Likelihood

For a nonlinear maximum likelihood estimation model $\boldsymbol{\mu} = \boldsymbol{\mu}(\mathbf{X}, \boldsymbol{\beta})$ with normally distributed outcomes $\mathbf{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the log likelihood function is

$$\ell(\boldsymbol{\mu}|\mathbf{y}, \boldsymbol{\Sigma}) = -\tfrac{1}{2}[n \log 2\pi + \log |\boldsymbol{\Sigma}| + (\mathbf{y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})]$$

which can be reorganized as

$$\ell(\boldsymbol{\mu}|\mathbf{y},\boldsymbol{\Sigma}) = \boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\mathbf{y} - \tfrac{1}{2}\left[n\log 2\pi - \log|\boldsymbol{\Sigma}^{-1}| + \boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right] - \tfrac{1}{2}\mathbf{y}^\top\boldsymbol{\Sigma}^{-1}\mathbf{y}.$$

This fits the log likelihood of exponential family distributions

$$\ell(\boldsymbol{\theta}|\mathbf{y}) = \boldsymbol{\theta}^\top\mathbf{t}(\mathbf{y}) - a(\boldsymbol{\theta}) + b(\mathbf{y})$$

where $\boldsymbol{\theta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu},\ \mathbf{t}(\mathbf{y}) = \mathbf{y},\ a(\boldsymbol{\theta}) = \frac{1}{2}\left[n\log 2\pi - \log|\boldsymbol{\Sigma}^{-1}| + \boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right]$ and $b(\mathbf{y}) = -\frac{1}{2}\mathbf{y}^\top\boldsymbol{\Sigma}^{-1}\mathbf{y}$.[3]

Applying Equation 7,

$$\hat{\boldsymbol{\beta}}_{MLE} = (\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top\tilde{\mathbf{y}}$$

$$\text{where } \tilde{\mathbf{X}} = \hat{\boldsymbol{\Sigma}}^{-1/2}\hat{\boldsymbol{\mathcal{M}}} \text{ and } \tilde{\mathbf{y}} = \hat{\boldsymbol{\Sigma}}^{-1/2}\left(\mathbf{y} + \hat{\boldsymbol{\mathcal{M}}}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\mu}}\right).$$

Note that the estimation model for nonlinear maximum likelihood estimation with a normal distribution can also be written as

$$\mathbf{y} = \boldsymbol{\mu}(\mathbf{X},\boldsymbol{\beta}) + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\varepsilon}|\mathbf{X} \sim N(\mathbf{0},\boldsymbol{\Sigma})$.

When $\boldsymbol{\mu}$ is a linear function of $\boldsymbol{\beta}$ such that $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, the estimation model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\varepsilon}|\mathbf{X} \sim N(\mathbf{0},\boldsymbol{\Sigma})$. In this case,

$$\tilde{\mathbf{X}} = \hat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{X} \text{ and } \tilde{\mathbf{y}} = \hat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{y}, \text{ so}$$

$$\hat{\boldsymbol{\beta}}_{MLE} = (\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top\tilde{\mathbf{y}} = (\mathbf{X}^\top\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{y}.$$

The linear maximum likelihood estimator $\hat{\boldsymbol{\beta}}_{MLE}$ has the same functional form as the Feasible Generalized Least Squares (FGLS) estimator, but the variance $\hat{\boldsymbol{\Sigma}}$ is estimated by maximum likelihood which is typically different from the FGLS estimate of $\hat{\boldsymbol{\Sigma}}$.

4.1.2 Binary Outcome Maximum Likelihood: probit, logit, etc.

Let the outcome variable $\mathbf{y}$ have independent elements with only two outcomes designated by $y_i \in \{0,1\}$. $y_i \sim Bernoulli(p_i)$ where p_i is the probability that $y_i = 1$. Each outcome y_i potentially has a different probability p_i. Let $\mathbf{p} = [p_1 \cdots p_n]^\top$ and $\mathbf{1}$ be an $n \times 1$ vector of ones.

[3] This likelihood function is conditional on the value of $\boldsymbol{\Sigma}$, which is convenient for decomposing $\boldsymbol{\beta}$. For the unconditional multivariate normal likelihood function, $\boldsymbol{\theta} = \begin{bmatrix}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ \text{vec}(-\frac{1}{2}\boldsymbol{\Sigma}^{-1})\end{bmatrix}$, $\mathbf{t}(\mathbf{y}) = \begin{bmatrix}\mathbf{y} \\ \text{vec}(\mathbf{y}\mathbf{y}^\top)\end{bmatrix}$, $a(\boldsymbol{\theta}) = -\frac{1}{2}[n\log 2\pi + \log|\boldsymbol{\Sigma}| + \boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}]$ and $b(\mathbf{y}) = 0$ (Camara Escudero, 2020).

Then the log likelihood function is

$$\ell(\mathbf{p}|\mathbf{y}) = \mathbf{y}^\top \log(\mathbf{p}) + (\mathbf{1} - \mathbf{y})^\top \log(\mathbf{1} - \mathbf{p})$$

which can be reorganized as

$$\ell(\mathbf{p}|\mathbf{y}) = \left[\log(\mathbf{p}) - \log(\mathbf{1} - \mathbf{p})\right]^\top \mathbf{y} + \log(\mathbf{1} - \mathbf{p})^\top \mathbf{1}.$$

The binary log likelihood belongs in the exponential family

$$\ell(\boldsymbol{\theta}|\mathbf{y}) = \boldsymbol{\theta}^\top \mathbf{t}(\mathbf{y}) - a(\boldsymbol{\theta}) + b(\mathbf{y})$$

when $\boldsymbol{\theta} = \log(\mathbf{p}) - \log(\mathbf{1} - \mathbf{p})$, $\mathbf{t}(\mathbf{y}) = \mathbf{y}$, $a(\boldsymbol{\theta}) = -\log(\mathbf{1} - \mathbf{p})^\top \mathbf{1}$, $b(\mathbf{y}) = 0$. From the properties of the Bernoulli distribution, $\boldsymbol{\mu} \equiv E[\mathbf{y}] = \mathbf{p}$ and $\boldsymbol{\Sigma} \equiv V[\mathbf{y}] = \text{vec}^{-1}[\mathbf{p}(\mathbf{1} - \mathbf{p})]$.

For typical binary outcome models, the expected value of $\mathbf{y}$ in $\boldsymbol{\mu}$ depends on the explanatory variables $\mathbf{X}$ and parameters $\boldsymbol{\beta}$ via a known cumulative distribution function $F(\cdot)$:

$$\boldsymbol{\mu} = \mathbf{p} = F(\mathbf{X}\boldsymbol{\beta})$$

so that $\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}^\top}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \equiv \hat{\boldsymbol{\mathcal{M}}} = \text{vec}^{-1}[f(\mathbf{X}\hat{\boldsymbol{\beta}})]\mathbf{X}$ where $f(\cdot)$ is the probability density function corresponding to $F(\cdot)$, and $\hat{\boldsymbol{\Sigma}} = \text{vec}^{-1}[F(\mathbf{X}\hat{\boldsymbol{\beta}})]\text{vec}^{-1}[\mathbf{1} - F(\mathbf{X}\hat{\boldsymbol{\beta}})]$.

As in Equation 7,

$$\tilde{\mathbf{X}} = \hat{\boldsymbol{\Sigma}}^{-1/2}\text{vec}^{-1}[f(\mathbf{X}\hat{\boldsymbol{\beta}})]\mathbf{X}, \quad \tilde{\mathbf{y}} = \hat{\boldsymbol{\Sigma}}^{-1/2}(\mathbf{y} + \text{vec}^{-1}[f(\mathbf{X}\hat{\boldsymbol{\beta}})]\mathbf{X}\hat{\boldsymbol{\beta}} - F(\mathbf{X}\hat{\boldsymbol{\beta}})), \text{ and}$$

$$\hat{\boldsymbol{\beta}}_{MLE} = (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\top \tilde{\mathbf{y}}.$$

In perhaps more intuitive notation, each row of $\tilde{\mathbf{X}}$ is $\tilde{\mathbf{x}}_i = \frac{f(\mathbf{x}_i\hat{\boldsymbol{\beta}})\mathbf{x}_i}{\sqrt{F(\mathbf{x}_i\hat{\boldsymbol{\beta}})(\mathbf{1} - F(\mathbf{x}_i\hat{\boldsymbol{\beta}}))}}$ and each element of $\tilde{\mathbf{y}}$ is $\tilde{y}_i = \frac{y_i + f(\mathbf{x}_i\hat{\boldsymbol{\beta}})\mathbf{x}_i\hat{\boldsymbol{\beta}} - F(\mathbf{x}_i\hat{\boldsymbol{\beta}})}{\sqrt{F(\mathbf{x}_i\hat{\boldsymbol{\beta}})(\mathbf{1} - F(\mathbf{x}_i\hat{\boldsymbol{\beta}}))}}$, where $\tilde{\mathbf{x}}_i$ and $\mathbf{x}_i$ are row vectors.

The most commonly employed cumulative density functions (CDF) $F(\cdot)$ for binary outcome models are the normal CDF for the probit model and the logistic CDF for the logit model. The $\hat{\boldsymbol{\beta}}_{MLE}$ for binary distributions provides the basis for added-variable plots for probit and logistic coefficient estimates.

4.2 Nonlinear Least Squares

The nonlinear least squares estimator follows the assumptions of ordinary least squares with the exception that the parameters $\boldsymbol{\beta}$ affect the outcome variable according to a known nonlinear function $\boldsymbol{\mu}(\mathbf{X}, \boldsymbol{\beta})$:

$$\mathbf{y} = \boldsymbol{\mu}(\mathbf{X}, \boldsymbol{\beta}) + \boldsymbol{\varepsilon}$$

where $\mathbf{y}$ is an $n \times 1$ vector of outcome variable observations, $\boldsymbol{\mu}(\cdot)$ is a known $n \times 1$ function, $\mathbf{X}$ is an $n \times k$ matrix of explanatory variable observations, $\boldsymbol{\beta}$ is a $k \times 1$ vector of unknown parameters and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of unobserved random errors.

The nonlinear least squares estimator minimizes the sum of squared errors, solving

$$\min_{\beta} \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}$$

Defining $\boldsymbol{\mathcal{M}} \equiv \frac{\partial \boldsymbol{\mu}(\mathbf{X},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top}$, the score is

$$\mathbf{s}(\boldsymbol{\beta}) = \frac{\partial \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}}{\partial \boldsymbol{\beta}} = 2\boldsymbol{\mathcal{M}}^\top \boldsymbol{\varepsilon}$$

and the Hessian is

$$\mathbf{H}(\boldsymbol{\beta}) = \frac{\partial^2 \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = 2 \frac{\partial^2 \boldsymbol{\mu}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \boldsymbol{\varepsilon} + 2\boldsymbol{\mathcal{M}}^\top \boldsymbol{\mathcal{M}}.$$

Davidson and MacKinnon (1993, 156, Equation 5.38) show that the first term in the Hessian is asymptotically negligible because $\frac{\partial^2 \boldsymbol{\mu}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \boldsymbol{\varepsilon}$ is a weighted sum of n errors.

Hence, $\hat{\mathbf{H}}$, a consistent estimate of the Hessian, is

$$\hat{\mathbf{H}} = 2\hat{\boldsymbol{\mathcal{M}}}^\top \hat{\boldsymbol{\mathcal{M}}}$$

where consistent estimates of the parameters replace the unknown parameters.

By Equation 5,

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{NLS} &= \hat{\mathbf{H}}^{-1} \left(\hat{\mathbf{H}} \hat{\boldsymbol{\beta}} + \hat{\mathbf{s}} \right) \\
&= (2\hat{\boldsymbol{\mathcal{M}}}^\top \hat{\boldsymbol{\mathcal{M}}})^{-1} \left(2\hat{\boldsymbol{\mathcal{M}}}^\top \hat{\boldsymbol{\mathcal{M}}} \hat{\boldsymbol{\beta}} + 2\hat{\boldsymbol{\mathcal{M}}}^\top (\mathbf{y} - \hat{\boldsymbol{\mu}}) \right) \\
&= (\hat{\boldsymbol{\mathcal{M}}}^\top \hat{\boldsymbol{\mathcal{M}}})^{-1} \hat{\boldsymbol{\mathcal{M}}}^\top (\mathbf{y} + \hat{\boldsymbol{\mathcal{M}}} \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\mu}}) \\
&= (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\top \tilde{\mathbf{y}}
\end{aligned}$$

when $\tilde{\mathbf{X}} = \hat{\boldsymbol{\mathcal{M}}}$ and $\tilde{\mathbf{y}} = \mathbf{y} + \hat{\boldsymbol{\mathcal{M}}} \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\mu}}$.

If $\boldsymbol{\mu}$ were linear so that $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, then $\tilde{\mathbf{X}} = \hat{\boldsymbol{\mathcal{M}}} = \mathbf{X}$ and $\tilde{\mathbf{y}} = \mathbf{y} + \hat{\boldsymbol{\mathcal{M}}} \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\mu}} = \mathbf{y}$. Hence one can interpret the NLS $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{y}}$ as being adjusted for the nonlinearity in $\boldsymbol{\mu}$ in order to display the $\hat{\boldsymbol{\beta}}$ coefficients as a linear trend of the transformed variables in the added-variable plot. The plot shows the contribution of each data observation to the nonlinear regression fit.

4.3 Nonlinear (and Linear) Generalized Method of Moments Estimation

Nonlinear generalized method of moments (GMM) estimators encompass a wide class of estimators. All the estimators discussed previously (OLS, GLS, 2SLS, 3SLS, MLE and NLS) are special cases of GMM, which also covers nonlinear instrumental variables estimation.

GMM estimators are the solution to a set of sample moments which correspond to hypothesized population moments. Let

$$\mathbf{g}(\boldsymbol{\beta}, \mathbf{y}, \mathbf{X})$$

be a known $n \times 1$ dimensional "elementary zero function" (Davidson & MacKinnon, 2004, p. 367) where, as before, $\mathbf{y}$ is an $n \times 1$ vector of outcome variable observations, $\mathbf{X}$ is an $n \times k$ matrix of explanatory variable observations, $\boldsymbol{\beta}$ is a $k \times 1$ vector of unknown parameters. $\mathbf{g}(\boldsymbol{\beta}, \mathbf{y}, \mathbf{X})$ has the property that for the true value $\boldsymbol{\beta}_0$ of $\boldsymbol{\beta}$

$$E[\mathbf{g}(\boldsymbol{\beta}_0, \mathbf{y}, \mathbf{X})] = \mathbf{0} \text{ and } E[\mathbf{g}(\boldsymbol{\beta}, \mathbf{y}, \mathbf{X})] \neq \mathbf{0} \;\forall\; \boldsymbol{\beta} \neq \boldsymbol{\beta}_0.$$

If the elementary zero function $\mathbf{g}(\boldsymbol{\beta}, \mathbf{y}, \mathbf{X})$ were linear, it would be the residual $\mathbf{e} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$.

For possibly endogenous explanatory variables $\mathbf{X}$, let $\mathbf{Z}$ be a $n \times l$ matrix of l linearly independent instruments where $l \geq k$. When $\mathbf{X}$ is not endogenous, $\mathbf{Z} = \mathbf{X}$. The assumed population moment for GMM estimation is

$$E[\mathbf{Z}^\top \mathbf{g}(\boldsymbol{\beta_0}, \mathbf{y}, \mathbf{X})] = \mathbf{0}.$$

Let the variance of $\mathbf{g}(\boldsymbol{\beta}, \mathbf{y}, \mathbf{X})$ be

$$E\big[\mathbf{g}(\boldsymbol{\beta})\mathbf{g}(\boldsymbol{\beta})^\top\big] = \boldsymbol{\Sigma}.$$

The GMM estimator of $\boldsymbol{\beta}$ is derived by solving the sample moment corresponding to the assumed population moment:

$$\mathbf{Z}^\top \mathbf{g}(\boldsymbol{\beta}, \mathbf{y}, \mathbf{X}) = \mathbf{0}.$$

When the number of instruments in $\mathbf{Z}$ is greater than the number of explanatory variables in $\mathbf{X}$ (*i.e.* $l > k$), however, there is no unique solution to the sample moments. Hansen (1982) famously showed that the most efficient linear combination of sample moments using a quadratic metric is weighted by the inverse of the variance of the sample moment, so that the moment conditions which better fit the data are given greater weight.

The criterion function to be maximized for efficient GMM estimation is

$$q(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) = -\frac{1}{n}\mathbf{g}(\boldsymbol{\beta})^\top \mathbf{Z}(\mathbf{Z}^\top \hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^\top \mathbf{g}(\boldsymbol{\beta})$$

(Davidson & MacKinnon, 2004, p. 376), where $\hat{\boldsymbol{\Sigma}}$ is a consistent estimate of $\boldsymbol{\Sigma}$.

Denoting $\mathbf{G} \equiv \frac{\partial \mathbf{g}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top}$, the score of $q(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta})$ is

$$\mathbf{s}(\boldsymbol{\beta}) \equiv \frac{\partial q(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\frac{2}{n}\mathbf{G}^\top \mathbf{Z}\big(\mathbf{Z}^\top \hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Z}\big)^{-1}\mathbf{Z}^\top \mathbf{g}(\boldsymbol{\beta}).$$

The Hessian is

$$\mathbf{H}(\boldsymbol{\beta}) \equiv \frac{\partial^2 q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = -\frac{2}{n}\left(\frac{\partial \mathbf{G}^\top}{\partial \boldsymbol{\beta}}\mathbf{Z}\big(\mathbf{Z}^\top \hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Z}\big)^{-1}\mathbf{Z}^\top \mathbf{g}(\boldsymbol{\beta}) + \mathbf{G}^\top \mathbf{Z}\big(\mathbf{Z}^\top \hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Z}\big)^{-1}\mathbf{Z}^\top \mathbf{G}\right)$$

The first term in the Hessian is a weighted average of the elementary zero function $\mathbf{g}(\boldsymbol{\beta})$, similar to the Hessian for the nonlinear least squares estimator. When $\mathbf{g}(\boldsymbol{\beta})$ is evaluated at a consistent estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ and $\mathbf{W}$ is an $k \times n$ matrix of weights, plim $\frac{1}{n}\mathbf{W}\mathbf{g}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$. Hence the first term has a probability limit of zero and can be dropped from the predicted Hessian. The predicted Hessian is

$$\hat{\mathbf{H}} = -\frac{2}{n}\hat{\mathbf{G}}^\top \mathbf{Z}\hat{\boldsymbol{\Omega}}^{-1}\mathbf{Z}^\top \hat{\mathbf{G}},$$

and the predicted score is

$$\hat{\mathbf{s}} = -\frac{2}{n}\hat{\mathbf{G}}^\top \mathbf{Z}\hat{\boldsymbol{\Omega}}^{-1}\mathbf{Z}^\top \hat{\mathbf{g}},$$

where $\hat{\boldsymbol{\Omega}} \equiv \mathbf{Z}^\top \hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Z}$, $\hat{\mathbf{G}} \equiv \mathbf{G}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$ and $\hat{\mathbf{g}} \equiv \mathbf{g}(\hat{\boldsymbol{\beta}}, \mathbf{y}, \mathbf{X})$.

By Equation 5,

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{GMM} &= \hat{\mathbf{H}}^{-1}(\hat{\mathbf{H}}\hat{\boldsymbol{\beta}} + \hat{\mathbf{s}}) \\
&= \Big(\hat{\mathbf{G}}^\top \mathbf{Z}\hat{\boldsymbol{\Omega}}^{-1}\mathbf{Z}^\top \hat{\mathbf{G}}\Big)^{-1}\Big(\hat{\mathbf{G}}^\top \mathbf{Z}\hat{\boldsymbol{\Omega}}^{-1}\mathbf{Z}^\top(\hat{\mathbf{G}}\hat{\boldsymbol{\beta}} + \hat{\mathbf{g}})\Big) \\
&= (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top \tilde{\mathbf{y}}
\end{aligned}$$

where

$$\begin{aligned}
\tilde{\mathbf{X}} &= \hat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{Z}\Big(\mathbf{Z}^\top \hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Z}\Big)^{-1}\mathbf{Z}^\top \hat{\mathbf{G}}, \text{ and} \\
\tilde{\mathbf{y}} &= \hat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{Z}\Big(\mathbf{Z}^\top \hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Z}\Big)^{-1}\mathbf{Z}^\top(\hat{\mathbf{G}}\hat{\boldsymbol{\beta}} + \hat{\mathbf{g}}).
\end{aligned}$$

All of the estimators discussed previously in this paper can be represented as special cases of the nonlinear GMM estimator. To represent MLE estimators as a GMM estimator, the population moment assumption is $E[\frac{\partial \ell}{\partial \boldsymbol{\beta}}] = 0$.

Table 1 shows the form of the elementary zero function $\mathbf{g}(\boldsymbol{\beta}, \mathbf{y}, \mathbf{X})$, the instruments $\mathbf{Z}$, the variance of $\mathbf{g}(\boldsymbol{\beta}, \mathbf{y}, \mathbf{X})$ and the $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{y}}$ values used to form the residuals $\mathbf{e}_{\tilde{\mathbf{x}}_1}$ and $\mathbf{e}_{\tilde{\mathbf{y}}}$ which can be graphed in an added-variable plot.

5 Conclusion

Added-variable plots are an intuitive and underutilized method of presenting the results of multivariate estimation to non-specialists and specialists alike. A scatter plot with a regression fit is a compelling way to display the relationship between data for an outcome variable and one explanatory variable. The added-variable plot shows a similar graph when there are more than one explanatory variable included in an estimated relationship. The added-variable plot shows the influence of each observation in the data for a chosen explanatory variable, while accounting for the effect of all the other explanatory variables. The plots show how well the data fit the model as well as which observations are outliers.

Table 1: Special cases of GMM estimators

Estimator	$\mathbf{g}(\boldsymbol{\beta})$	$\mathbf{Z}$	$V[\mathbf{g}]$	$\tilde{\mathbf{X}}$	$\tilde{\mathbf{y}}$
OLS	$\mathbf{y}-\mathbf{X}\boldsymbol{\beta}$	$\mathbf{X}$	$\sigma^2\mathbf{I}$	$\mathbf{X}$	$\mathbf{y}$
GLS	$\mathbf{y}-\mathbf{X}\boldsymbol{\beta}$	$\mathbf{X}$	$\boldsymbol{\Sigma}$	$\hat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{X}$	$\hat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{y}$
2SLS	$\mathbf{y}-\mathbf{X}\boldsymbol{\beta}$	$\mathbf{Z}$	$\sigma^2\mathbf{I}$	$\mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{X}$	$\mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{y}$
3SLS	$\mathbf{y}-\mathbf{X}\boldsymbol{\beta}$	$\mathbf{Z}$	$\boldsymbol{\Sigma}$	$\hat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{Z}(\mathbf{Z}^\top\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{X}$	$\hat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{Z}(\mathbf{Z}^\top\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{y}$
MLE exp	$\mathbf{t}(\mathbf{y})-\boldsymbol{\mu}(\mathbf{X},\boldsymbol{\beta})$	$\hat{\boldsymbol{\mathcal{M}}}$	$\boldsymbol{\Sigma}$	$\hat{\boldsymbol{\Sigma}}^{-1/2}\hat{\boldsymbol{\mathcal{M}}}$	$\hat{\boldsymbol{\Sigma}}^{-1/2}(\mathbf{t}(\mathbf{y})+\hat{\boldsymbol{\mathcal{M}}}\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\mu}})$
NLS	$\mathbf{y}-\boldsymbol{\mu}(\mathbf{X},\boldsymbol{\beta})$	$\hat{\boldsymbol{\mathcal{M}}}$	$\sigma^2\mathbf{I}$	$\hat{\boldsymbol{\mathcal{M}}}$	$\mathbf{y}+\hat{\boldsymbol{\mathcal{M}}}\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\mu}}$
NLS $\boldsymbol{\Sigma}$	$\mathbf{y}-\boldsymbol{\mu}(\mathbf{X},\boldsymbol{\beta})$	$\hat{\boldsymbol{\mathcal{M}}}$	$\boldsymbol{\Sigma}$	$\hat{\boldsymbol{\Sigma}}^{-1/2}\hat{\boldsymbol{\mathcal{M}}}$	$\hat{\boldsymbol{\Sigma}}^{-1/2}(\mathbf{y}+\hat{\boldsymbol{\mathcal{M}}}\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\mu}})$
lin GMM	$\mathbf{y}-\mathbf{X}\boldsymbol{\beta}$	$\mathbf{Z}$	$\boldsymbol{\Sigma}$	$\hat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{Z}(\mathbf{Z}^\top\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{X}$	$\hat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{Z}(\mathbf{Z}^\top\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{y}$
NL GMM	$\mathbf{g}(\boldsymbol{\beta},\mathbf{y},\mathbf{X})$	$\mathbf{Z}$	$\boldsymbol{\Sigma}$	$\hat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{Z}(\mathbf{Z}^\top\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^\top\hat{\mathbf{G}}$	$\hat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{Z}\hat{\boldsymbol{\Omega}}^{-1}\mathbf{Z}^\top\left(\hat{\mathbf{G}}\hat{\boldsymbol{\beta}}+\hat{\mathbf{g}}\right)$

OLS - ordinary least squares, GLS - generalized least squares, 2SLS - two stage least squares, 3SLS - three stage least squares, MLE exp - maximum likelihood estimator of an exponential family distribution, NLS - nonlinear least squares, NLS $\boldsymbol{\Sigma}$ - nonlinear least squares with heteroskedastic or autocorrelated errors, lin GMM - linear generalized method of moments, NL GMM - nonlinear generalized method of moments.

Previously, added-variable plots have only been derived for ordinary least squares and General Linear Model (GLM) estimators. Here the method is extended to most common estimation methods with the exception of Bayesian and nonparametric estimators. Derivations are provided for constructing added-variable plots of generalized least squares and instrumental variables estimators. A method is presented for constructing plots for extremum estimators with solutions for maximum likelihood using the exponential family of distributions, nonlinear least squares and generalized method of moments estimators. The results apply to most time series, limited dependent variable and survival estimators (maximum likelihood) as well as static and dynamic panel data estimators (generalized method of moments).

References

Amemiya, T. (1985). Asymptotic properties of extremum estimators. *Advanced Econometrics* (pp. 105–158). Harvard University Press.

Andersen, E. B. (1970). Sufficiency and Exponential Families for Discrete Sample Spaces. *Journal of the American Statistical Association*, *65*(331), 1248–1255. https://doi.org/10.1080/01621459.1970.10481160

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. Wiley.

Camara Escudero, M. (2020). *Multivariate Normal as an Exponential Family Distribution*. Retrieved February 3, 2023, from https://maurocamaraescudero.

netlify.app/post/multivariate-normal-as-an-exponential-family-distribution/

Chien, L.-C. (2011). A robust diagnostic plot for explanatory variables under model mis-specification. *Journal of Applied Statistics*, *38*(1), 113–126. https://doi.org/10.1080/02664760903271940

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. Chapman and Hall.

Dasgupta, A. (2011). The Exponential Family and Statistical Applications. *Probability for Statistics and Machine Learning:Fundamentals and Advanced Topics* (pp. 583–612). Springer.

Davidson, R., & MacKinnon, J. G. (1993). *Estimation and inference in econometrics*. Oxford.

Davidson, R., & MacKinnon, J. G. (2004). *Econometric Theory and Methods*. Oxford.

Draper, N. R., & Smith, H. (1966). *Applied Regression Analysis*. Wiley.

Exponential family. (2022). In *Wikipedia*. Retrieved January 22, 2023, from https://en.wikipedia.org/wiki/Exponential_family

Frisch, R., & Waugh, F. V. (1933). Partial Time Regressions as Compared with Individual Trends. *Econometrica*, *1*(4), 387. https://doi.org/10.2307/1907330

Gallup, J. L. (2023). Cognitive and Economic Development. *Portland State Economics Working Papers 165*. https://archives.pdx.edu/ds/psu/40716

Hansen, L. P. (1982). Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*, *50*(4), 1029. https://doi.org/10.2307/1912775

Hines, R. J. O., & Carter, E. M. (1993). Improved Added Variable and Partial Residual Plots for the Detection of Influential Observations in Generalized Linear Models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *42*(1), 3–20. https://doi.org/10.2307/2347405

Lovell, M. C. (1963). Seasonal Adjustment of Economic Time Series and Multiple Regression Analysis. *Journal of the American Statistical Association*, *58*(304), 993–1010.

Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression*. Addison-Wesley.

Newey, W. K., & McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics* (pp. 2111–2245). Elsevier. https://doi.org/10.1016/S1573-4412(05)80005-4

Pregibon, D. (1985). Link tests. In S. Kotz & N. Johnson (Eds.), *Encyclopedia of Statistical Sciences* (pp. 82–85). Wiley.

Wang, P. C. (1985). Adding a Variable in Generalized Linear Models. *Technometrics*, *27*(3), 273–276. https://doi.org/10.2307/1269708

Weisberg, S. (1980). *Applied linear regression*. Wiley-Interscience.