Chapter 7

Linear and generalised linear models

7.1 Linear model

7.1.1 Model

So far y_i has been iid. Here we extend this to the case where the observations depend linearly on some known (fixed) $p \times 1$ vector of regressors x_i. The basic structure will be that

$$y_i = x_i'\beta + \varepsilon_i, \quad E(\varepsilon_i) = 0, \quad Var(\varepsilon_i) = \sigma^2,$$

and the ε_i are iid. This model is indexed by the parameters β and σ^2. We can write the model in matrix form as $y = X\beta + \varepsilon$, where $y = (y_1, ..., y_n)'$, $X = (x_1, ..., x_n)'$ and $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)'$.

7.2 Definition

$$\widehat{\beta} = (X'X)^{-1} X'y, \quad \text{assuming } X \text{ has full rank.}$$

7.2.1 Interpretation

(1) The most common association with the above estimator is with the so called least squares principle. This write

$$\begin{aligned} \widehat{\beta} &= \arg\min_{\beta} \varepsilon'\varepsilon = \arg\min_{\beta} \sum_{i=1}^{n} \varepsilon_i^2 \\ &= \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 . \end{aligned}$$

That is it determines $\widehat{\beta}$ by finding the value of β which minimises the squared vertical distances between y_i and $x_i'\beta$. This construction leaves many questions unanswered for the choice of squared distances looks completely ad hoc. Justifications usually involve the resulting estimator has good properties. Alternatives include

$$\widetilde{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} |y_i - x_i'\beta| ,$$

which is called an L_1 (or least absolute deviation) regression estimator, or

$$\widetilde{\widetilde{\beta}} = \arg\min_{\beta} medium \{|y_1 - x_1'\beta|, ..., |y_n - x_n'\beta|\} ,$$

the least medium absolute deviation (high breakdown) estimator.

(2) Suppose we regard X as fixed (or we condition on it for purposes of inference) and $y_i \sim N(x_i'\beta, \sigma^2)$, with the y_i being independent over i. Then the ML estimator is

$$\begin{aligned} \widehat{\beta} &= \arg\max_{\beta} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i'\beta)^2\right\} \\ &= \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 , \end{aligned}$$

the same as the least squares estimator. This implies least squares estimation is the ML estimator if the model is Gaussian and has many optimality properties if the truth is a Gaussian regression.

(3) Suppose $y_i \sim Laplace(x_i'\beta, \sigma^2)$, that is

$$f(y_i; \beta, \sigma^2) = \frac{1}{2\sigma} \exp\left\{-\frac{1}{\sigma}|y_i - x_i'\beta|\right\},$$

and that the data are independent over i. For this distribution the usual mean and variance of the Gaussian regression carry over. Then the model has much fatter tails in $y_i - x_i'\beta$ than the normal as the Laplace regression model involves only exponentiating minus the absolute value, while the Gaussian model squares these distances. The ML estimator for β is

$$\begin{aligned} \widetilde{\beta} &= \arg\max_{\beta} \left(\frac{1}{2\sigma}\right)^n \exp\left\{-\frac{1}{\sigma}\sum_{i=1}^{n}|y_i - x_i'\beta|\right\} \\ &= \arg\min_{\beta} \sum_{i=1}^{n} |y_i - x_i'\beta| . \end{aligned}$$

That is the ML estimator for this distribution is an L_1 regression. Such a regression is typically less sensitive to unusual observations than the least squares regression — this can be a good thing or a bad thing in practice. NOTE: the Laplace log-likelihood function is not continuously differentiable with respect to β and so many of the typical properties of MLE do not go through even if the data is Laplace.

example

Think about the dataset $x = (1, 2, 3, 5)'$ and $y = (1, 2, 3, 1)'$, so 3 observations are in a straight-line with gradient one and a single observation is an outlier to that. The LS estimator is $(1 + 4 + 9 + 5) / (1 + 4 + 9 + 25) = 0.487$ and is deeply effected by the outlier. The L_1 estimator minimises sum of the absolute values. These decrease with a gradient of -4 until $\beta = 1/5$ when it slows to -2 which continues until $\beta = 1$ when the gradient switches to 4. So the L_1 regressor gives an estimator of 1, so ignores the outlier completely.

7.3 Properties of least squares

For $y = X\beta + \varepsilon$ and $E(\varepsilon) = 0$ while $Var(\varepsilon) = \sigma^2 I$, then

$$\widehat{\beta} = (X'X)^{-1} X'y = \beta + (X'X)^{-1} X'\varepsilon,$$

so $\widehat{\beta}$ is unbiased while $Var\left(\widehat{\beta}\right) = \sigma^2 (X'X)^{-1}$.

If ε is Gaussian, then $\widehat{\beta}$ is a linear combination of Gaussian and so if Gaussian. Further,

$$\frac{\partial \log L}{\partial \beta} = \frac{1}{\sigma^2} X'(y - X\beta) \quad \text{implying} \quad Var\frac{\partial \log L}{\partial \beta} = \frac{1}{\sigma^2} X'X,$$

so $\widehat{\beta}$ achieves the Cramer-Rao lower bound and is the minimum variance unbiased estimator.

If the parametric model is not Gaussian $\widehat{\beta}$ has more limited optimality properties. In particular it is possible to show that in the class of estimators which is linear in y, $\widehat{\beta}$ is the minimum variance unbiased estimator (Gauss-Markov theorem).

This result is less helpful than it at first seems for when the data is non-Gaussian it seems clear that we should be using a non-linear estimator! See, for example, the discussion above about the Laplace regression model.

A more thorough discussion of these issues is given in Chapter **??**.

7.4 Generalized linear models*

7.4.1 Models

Instead of working with a particular density function it is sometimes helpful to construct a class of densities. One such class is the exponential family, which puts

$$f(y;\theta,\psi) = \exp\left[\frac{1}{\psi}\{y\theta - b(\theta)\} - c(y,\psi)\right].$$

We are free to chose the functions $b(.)$ and $c(.,.)$ to place familiar distributions in this framework.. Simple examples of this includes

- $N(\mu,\sigma^2)$, $\theta=\mu$, $\psi=\sigma^2$, $b(\theta)=\theta^2/2$.
- Poisson with mean μ, puts $\theta=\log\mu$, $b(\theta)=\exp\theta$ and $\psi=1$.
- Bernoulli with probability of success of p. Then we put $\theta=\log\{p/(1-p)\}$, $b(\theta)=\log(1+\exp\theta)$ and $\psi=1$.

Other common distributions which go into this framework includes the gamma and some extreme value distributions. An excellent review of this material is given in Azzalini (1996, Ch. 6) and Garthwaite, Jolliffe and Jones (1995, Ch. 10).

Having a class of densities is useful as it allows us to derive some generic results. First, the log-likelihood is

$$\log L(\theta,\psi;y) = const + \frac{\theta}{\psi}\sum_{i=1}^{n} y_i - \frac{n\,b(\theta)}{\psi} - \sum c(y_i,\psi),$$

which implies

$$\frac{\partial \log L(\theta,\psi;y)}{\partial\theta} = \frac{1}{\psi}\sum_{i=1}^{n} y_i - \frac{n\,b'(\theta)}{\psi}, \quad \text{where} \quad b'(\theta) = \frac{\partial b(\theta)}{\partial\theta}.$$

As the expectation of the score is zero when evaluated at the true parameter point this implies

$$E(y_i) = b'(\theta),$$

while the information equality implies

$$Var(y_i) = \psi b''(\theta).$$

7.4.2 Regression models

The Gaussian distribution $Y_i \sim N(\mu,\sigma^2)$ extends to the regression problem where $Y_i \sim N(\beta' x_i,\sigma^2)$. An interesting question is how this argument extends to non-Gaussian densities. Clearly we cannot just write $\mu_i = E(y_i) = x_i'\beta$ as some non-Gaussian densities require the mean to be positive! Instead we will work with a function of the mean

$$g(\mu_i) = x_i'\beta = \eta_i,$$

which is called a link function. Throughout I will assume $g(.)$ is a monotonic function. An example of this is the Poisson case where $\mu_i > 0$ and so it makes some sense to write $\log\mu_i = x_i'\beta$. Economic examples of the use of Poisson type regressions includes models of patents by Hausman, Hall and Griliches (1984). Likewise in the Bernoulli case we have $\mu_i \in (0,1)$ and has been used by, for example, Micklewright (1989) in labour economics. Hence it might make sense to use a logistic transformation $\log\{\mu_i/(1-\mu_i)\} = x_i'\beta$.

If we combine an exponential family model with a regression model then we call the result a generalized linear model. It has the generic form

$$\log f(y;\theta_i,\psi) = \frac{1}{\psi}\{y\theta_i - b(\theta_i)\} - c(y,\psi),$$

where

$$\mu_i = E(y_i) = b'(\theta_i) \quad \text{and} \quad g(\mu_i) = x_i'\beta = \eta_i.$$

Hence the likelihood depends now only on the parameters ψ and β as the θ_i are solely determined by the regressors and the regression coefficients.

7.4.3 Likelihood analysis

We think of β as the parameters of interest, then

$$\log L(\beta,\psi;y) = \frac{1}{\psi}\sum_{i=1}^{n}\{y_i\theta_i - b(\theta_i)\} - \sum_{i=1}^{n} c(y_i,\psi),$$

and so

$$\frac{\partial \log L(\beta,\psi;y)}{\partial\beta} = \frac{1}{\psi}\sum_{i=1}^{n}\left[\frac{\partial\eta_i}{\partial\beta}\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\theta_i}{\partial\mu_i}\{y_i - b'(\theta_i)\}\right].$$

Although this looks complicated it simplifies to the important result

$$\frac{\partial \log L(\beta,\psi;y)}{\partial\beta} = \sum_{i=1}^{n}\left[x_i\frac{\{y_i-\mu_i\}}{Var(y_i)}\frac{\partial\mu_i}{\partial\eta_i}\right],$$

as

$$\frac{\partial\eta_i}{\partial\beta} = x_i, \frac{\partial\mu_i}{\partial\theta_i} = b''(\theta_i) = \frac{Var(y_i)}{\psi}.$$

This score cannot be immediately solved to deliver the ML estimator of β as in general $\frac{\partial\mu_i}{\partial\eta_i}, Var(y_i)$ and μ_i all depend on β. Hence we will have to use a numerical maximisation procedure to iterate to the maximum. Here we will look at a Fisher scoring method[1], which uses the expected information matrix.

The expected information for the sample can be found by looking at

$$E\left\{\frac{\partial \log L(\beta,\psi;Y_i)}{\partial\beta_j}\frac{\partial \log L(\beta,\psi;Y_i)}{\partial\beta_k}\right\} = E\left[x_{ij}\frac{\{y_i-\mu_i\}}{Var(y_i)}\frac{\partial\mu_i}{\partial\eta_i}x_{ik}\frac{\{y_i-\mu_i\}}{Var(y_i)}\frac{\partial\mu_i}{\partial\eta_i}\right],$$

which simplifies to

$$x_{ij}x_{ik}\frac{1}{Var(y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2.$$

This implies, using independence across the Y_i that

$$-E\left(\frac{\partial^2\log L}{\partial\beta\partial\beta'}\right) = E\left\{\frac{\partial\log L}{\partial\beta}\frac{\partial\log L}{\partial\beta'}\right\} = \sum_{i=1}^{n} x_i x_i'\frac{1}{Var(y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 = X'WX,$$

where W is diagonal with $ii-th$ element

$$w_i = \frac{1}{Var(y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2,$$

and $X_{ij} = x_{ij}$. Likewise the score can be written as

$$\frac{\partial \log L(\beta,\psi;y)}{\partial\beta} = X'Wu, \quad \text{where} \quad u_i = \{y_i-\mu_i\}\,\partial\eta_i/\partial\mu_i.$$

The implication of this is that the Fisher scoring algorithm actually carries out iteratively weighted least squares, computing the update of β by

$$\beta^{(k+1)} = \beta^{(k)} + (X'WX)^{-1}X'Wu = (X'WX)^{-1}X'W\left(X\beta^{(k)} + u\right) = (X'WX)^{-1}X'Wy^{(k)},$$

say.

[1] Fisher scoring generically iterates

$$\theta^{(k+1)} = \theta^{(k)} + \left(-E\frac{\partial^2\log L}{\partial\theta\partial\theta'}\bigg|_{\theta=\theta^{(k)}}\right)^{-1}\frac{\partial\log L(\theta^{(k)};y)}{\partial\theta}$$

Hence it is a particular case of the quasi-Newton method.

exercise

Suppose we wish to model durations of employment, y_i, as a function of some (assumed fixed) explanatory variable, x_i. We suppose

$$f(y_i;\beta) = \lambda_i \exp(-y_i\lambda_i), \quad y_i > 0, \quad \lambda_i = \exp(x_i\beta),$$

and $E(y_i) = \lambda_i^{-1}$. Assume the y_i are independent over i. Write down the model's likelihood and find the score function for β. Show that the log-likelihood is concave and hence suggest how you might numerically compute the ML estimator?

7.4.4 Asymptotic distribution

In our discussion of the asymptotic distribution of the ML estimator we have always assumed that the observations are iid. This model does not have identical distributions, although the independence across individuals remains.

The average expected information per observation is

$$\frac{1}{n}E\left(\frac{\partial \log L}{\partial \beta}\frac{\partial \log L}{\partial \beta'}\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(\frac{\partial \log L(\theta;Y_i)}{\partial \beta}\frac{\partial \log L(\theta;Y_i)}{\partial \beta'}\right) = \frac{1}{n}X'WX.$$

Now suppose that

$$\lim_{n\to\infty}\frac{1}{n}E\left(\frac{\partial \log L}{\partial \beta}\frac{\partial \log L}{\partial \beta'}\right) = M > 0,$$

then we have that

$$\sqrt{n}\left(\widehat{\beta} - \beta\right) \xrightarrow{d} N\left(0, M^{-1}\right).$$

This result will hold as long as M is finite and positive semi-definite. This will happen as long as the expected information per observation

$$E\left\{\frac{\partial \log L(\theta;Y_i)}{\partial \beta}\frac{\partial \log L(\theta;Y_i)}{\partial \beta'}\right\}$$

is always strictly positive and bounded.

7.5 Generalized Least Squares and instrumental variables

7.5.1 Generalized Least Squares, GLS

Since $\boldsymbol{\Omega}$ is p.d, the inverse exists and is also p.d. Therefore, there exists a nonsingular matrix $\boldsymbol{P}$ so that

$$\boldsymbol{\Omega}^{-1} = \boldsymbol{P}'\boldsymbol{P}.$$

This gives

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{P}'\boldsymbol{P}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{P}'\boldsymbol{P}\boldsymbol{y} = \{(\boldsymbol{P}\boldsymbol{X})'(\boldsymbol{P}\boldsymbol{X})\}^{-1}(\boldsymbol{P}\boldsymbol{X})'(\boldsymbol{P}\boldsymbol{y}).$$

This would have been obtained from the regression of $\boldsymbol{P}\boldsymbol{y}$ on $\boldsymbol{P}\boldsymbol{X}$. To deal with the nonspherical model, pre-multiply $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$ by $\boldsymbol{P}$ so that

$$\boldsymbol{y}_* = \boldsymbol{X}_*\boldsymbol{\beta} + \boldsymbol{u}_*$$

where $\boldsymbol{y}_* = \boldsymbol{P}\boldsymbol{y}$, $\boldsymbol{X}_* = \boldsymbol{P}\boldsymbol{X}$ and $\boldsymbol{u}_* = \boldsymbol{P}\boldsymbol{u}$. Since, $\boldsymbol{\Omega} = \boldsymbol{P}^{-1}(\boldsymbol{P}')^{-1}$,

$$\text{var}(\boldsymbol{u}_*) = \text{E}(\boldsymbol{P}\boldsymbol{u}\boldsymbol{u}'\boldsymbol{P}') = \sigma^2\boldsymbol{P}\boldsymbol{\Omega}\boldsymbol{P}' = \sigma^2\boldsymbol{I},$$

satisfying the conditions under which OLS is BLUE. The *GLS* estimator is defined as

$$\bar{\boldsymbol{\beta}_{\text{GLS}}} = (\boldsymbol{X}_*'\boldsymbol{X}_*)^{-1}\boldsymbol{X}_*'\boldsymbol{y}_* = (\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{y})$$

so that

$$\text{var}(\bar{\boldsymbol{\beta}_{\text{GLS}}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}$$

which is also the asymptotic variance matrix had we adopted the ML approach. An unbiased estimate of σ^2 can be obtained from OLS applied to the transformed model

$$\begin{aligned} s^2 &= (\boldsymbol{y}_* - \boldsymbol{X}_*\bar{\boldsymbol{\beta}_{\text{GLS}}})'(\boldsymbol{y}_* - \boldsymbol{X}_*\bar{\boldsymbol{\beta}_{\text{GLS}}})/(n-k) \\ &= \{\boldsymbol{P}(\boldsymbol{y} - \boldsymbol{X}\bar{\boldsymbol{\beta}_{\text{GLS}}})\}'\{\boldsymbol{P}(\boldsymbol{y} - \boldsymbol{X}\bar{\boldsymbol{\beta}_{\text{GLS}}})\}/(n-k) \\ &= (\boldsymbol{y} - \boldsymbol{X}\bar{\boldsymbol{\beta}_{\text{GLS}}})'\boldsymbol{\Omega}^{-1}(\boldsymbol{y} - \boldsymbol{X}\bar{\boldsymbol{\beta}_{\text{GLS}}})/(n-k), \end{aligned}$$

which differs from the biased MLE $\hat{\sigma}^2$ by the factor $n/(n-k)$. To test the finite sample restriction $H_0 : \boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r}$, the test can be based on

$$F = \frac{(\boldsymbol{r} - \boldsymbol{R}\bar{\boldsymbol{\beta}_{\text{GLS}}})'\{\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{R}'\}^{-1}(\boldsymbol{r} - \boldsymbol{R}\bar{\boldsymbol{\beta}_{\text{GLS}}})/q}{s^2} \sim F_{q,n-k}.$$

Note that we could have also written $\boldsymbol{u}$ as $\boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{V})$ (where $\boldsymbol{V} = \sigma^2\boldsymbol{\Omega}$), so that,

$$\begin{aligned} \bar{\boldsymbol{\beta}_{\text{GLS}}} &= (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y} \\ \text{var}(\bar{\boldsymbol{\beta}_{\text{GLS}}}) &= (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}. \end{aligned}$$

7.5.2 Instrumental Variable (IV) estimators

If the condition stating that the regressors and the disturbance are independent does not hold, then the OLS estimators are biased and inconsistent. As an example, consider the *errors in variables* problem,

$$y = x\beta + u$$

without a constant term, but that the variables are measured with error, that is,

$$\begin{aligned} x &= \bar{x} + v \\ \text{so that } y &= \beta\bar{x} + u \end{aligned}$$

where u, $\bar{x}$ and v are mutually independent. This scenario is classic under models with *measurement error*. The OLS equations give

$$\begin{aligned} \hat{\beta} &= \frac{\sum y_i x_i}{\sum x_i^2} = \frac{\sum x_i(\beta\bar{x}_i + u_i)}{\sum x_i^2} \\ &= \beta\frac{\sum x_i\bar{x}_i}{\sum x_i^2} + \frac{\sum x_i u_i}{\sum x_i^2}. \end{aligned}$$

If the second moments and their plims's exist, then,

$$\begin{aligned} \text{plim}\left(\frac{1}{n}\sum x_i^2\right) &= \sigma_{\bar{x}}^2 + \sigma_v^2 \\ \text{plim}\left(\frac{1}{n}\sum x_i\bar{x}_i\right) &= \sigma_{\bar{x}}^2 \\ \text{plim}\left(\frac{1}{n}\sum x_i u_i\right) &= 0 \end{aligned}$$

giving

$$\text{plim}\hat{\beta} = \beta\left(\frac{\sigma_{\bar{x}}^2}{\sigma_{\bar{x}}^2 + \sigma_v^2}\right).$$

Hence OLS is biased and inconsistent. Suppose we now re-write the model as

$$y = \beta x + (u - \beta v)$$

so that

$$\hat{\beta} = \beta + \frac{\sum x_i(u_i - \beta v_i)}{\sum x_i^2},$$

then, $\text{plim}(1/n)\sum x_i(u_i-\beta v_i) = \text{plim}(1/n)\sum x_i u_i - \beta\,\text{plim}(1/n)\sum x_i v_i = -\beta\sigma_v^2$, giving the same results for plim$\hat{\beta}$ as before. For the general linear model,

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}$$

so that,

$$\text{plim}\,\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \text{plim}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \cdot \text{plim}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{u}\right).$$

If we assume that plim $(\boldsymbol{X}'\boldsymbol{X}/n) = \boldsymbol{\Sigma}_{\boldsymbol{XX}}$, a p.d. matrix of full rank, and plim $(\boldsymbol{X}'\boldsymbol{u}/n) = \boldsymbol{\Sigma}_{\boldsymbol{Xu}} \neq \boldsymbol{0}$, then,

$$\text{plim}\,\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \boldsymbol{\Sigma}_{\boldsymbol{XX}}{}^{-1} \cdot \boldsymbol{\Sigma}_{\boldsymbol{Xu}}$$

so that correlation of $\boldsymbol{u}$ with one or more of the disturbance terms makes OLS inconsistent. A consistent estimator can be obtained by using *instrumental variables*.

Suppose the model is still $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$, with var$(\boldsymbol{u}) = \sigma^2\boldsymbol{I}$, but plim $(\boldsymbol{X}'\boldsymbol{u}/n) \neq \boldsymbol{0}$. If we can find a matrix $\boldsymbol{Z}$ of order $n \times l$, $(l \geq k)$, such that

- the variables in $\boldsymbol{Z}$ are correlated with those in $\boldsymbol{X}$ and in the limit, plim$(\boldsymbol{Z}'\boldsymbol{X}/n) = \boldsymbol{\Sigma}_{\boldsymbol{ZX}}$, a finite matrix of full rank.
- and plim $(\boldsymbol{Z}'\boldsymbol{u}/n) = \boldsymbol{0}$.

Then

$$\boldsymbol{Z}'\boldsymbol{y} = \boldsymbol{Z}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}'\boldsymbol{u} \text{ with } \text{var}(\boldsymbol{Z}'\boldsymbol{u}) = \sigma^2(\boldsymbol{Z}'\boldsymbol{Z})$$

which suggests the use of GLS. The estimator is thus

$$\hat{\boldsymbol{\beta}_{\text{GLS}}} = \hat{\boldsymbol{\beta}_{\text{IV}}} = \{\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}\}^{-1}\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y}$$

so that $\hat{\boldsymbol{\beta}_{\text{GLS}}} = \hat{\boldsymbol{\beta}_{\text{IV}}} = (\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{y}$, where $\boldsymbol{P}_{\boldsymbol{Z}} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$. The variance-covariance matrix is

$$\text{var}(\hat{\boldsymbol{\beta}_{\text{IV}}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X})^{-1}$$

and σ^2 can be consistently estimated by

$$\hat{\sigma}^2 = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}_{\text{IV}}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}_{\text{IV}}})/n.$$

To check that the IV estimator is consistent,

$$\begin{aligned}\hat{\boldsymbol{\beta}_{\text{IV}}} &= \boldsymbol{\beta} + \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{u}\right)\\ \text{and } \frac{1}{n}\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X} &= \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right),\end{aligned}$$

so that

$$\text{plim}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}\right) = \boldsymbol{\Sigma}_{\boldsymbol{XZ}}\boldsymbol{\Sigma}_{\boldsymbol{ZZ}}{}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{ZX}}$$

which is finite and nonsingular. Also,

$$\text{plim}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{u}\right) = \boldsymbol{\Sigma}_{\boldsymbol{XZ}}\boldsymbol{\Sigma}_{\boldsymbol{ZZ}}{}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Zu}} = \boldsymbol{0}$$

(since the instruments are assumed to be uncorrelated in the limit with the disturbance), so that the IV estimator is consistent.

Example - special case

If $l = k$, then $\boldsymbol{X}'\boldsymbol{Z}$ is a $k \times k$ nonsingular matrix. The estimator now simplifies to

$$\begin{aligned}\hat{\boldsymbol{\beta}_{\text{IV}}} &= (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{y}\\ \text{with } \text{var}(\hat{\boldsymbol{\beta}_{\text{IV}}}) &= \sigma^2(\boldsymbol{Z}'\boldsymbol{X})^{-1}(\boldsymbol{Z}'\boldsymbol{Z})(\boldsymbol{X}'\boldsymbol{Z})^{-1}.\end{aligned}$$

7.5.2.1 Two-Stage Least Squares (2SLS)

The IV estimator can also be seen as the result of applying least squares twice.

(1) Regress each of the variables in the $\boldsymbol{X}$ matrix on $\boldsymbol{Z}$ to obtain a matrix of fitted values, $\hat{\boldsymbol{X}}$,

$$\hat{\boldsymbol{X}} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X} = \boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}.$$

(2) Regress $\boldsymbol{y}$ on $\hat{\boldsymbol{X}}$ to obtain the estimated $\boldsymbol{\beta}$ vector

$$\begin{aligned}\hat{\boldsymbol{\beta}_{\text{2SLS}}} &= (\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}(\hat{\boldsymbol{X}}'\boldsymbol{y})\\ &= (\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{y})\\ &= \hat{\boldsymbol{\beta}_{\text{IV}}}.\end{aligned}$$