KTH Engineering Sciences
Harald Lang 24/2–09

Comments on Bruce Hansen's course notes.

Some formulas can be a bit hard to get hold of, since one has to combine formulas from different places. Therefore I summarise some results here. (My notation may differ from Hansen's in some places.)

Consider the regression equation

$$y = x'\beta + e \quad \mathrm{E}[\, e \mid x] = 0$$

In sample form:

$$Y = X\beta + e \quad (e \text{ is now an } n \times 1 \text{ matrix})$$
$$\mathrm{E}[\, e \mid X] = 0, \quad \text{which implies } \mathrm{E}[X'e] = 0.$$

The OLS estimate $\hat{\beta}$ of β is defined by

$$Y = X\hat{\beta} + \hat{e} \quad X'\hat{e} = 0$$

which yields $\hat{\beta} = (X'X)^{-1}X'Y$. This is equivalent to minimising the sum of squares of the residuals. The assumption that observations are inependent implies

$$\mathrm{E}[\, ee' \mid X] = \text{diagonal-matrix}(\sigma_1^2, \ldots, \sigma_n^2)$$

- *White's estimate of the covariance matrix* of $\hat{\beta}$, conditional on X, is

$$\hat{\Omega} = (X'X)^{-1}[X'D(\hat{e}^2)X](X'X)^{-1}$$

where

$$D(\hat{e}^2) = \text{diagonal-matrix}(\hat{e}_1^2, \ldots, \hat{e}_n^2)$$

i.e.,

$$X'D(\hat{e}^2)X = \sum_1^n x_i x_i' \hat{e}_i^2$$

- *The homoskedastic covariance matrix* is

$$\hat{\Omega} = \hat{\sigma}^2(X'X)^{-1} = \hat{\sigma}^2 \begin{pmatrix} a & b' \\ b & n^{-1}C(X_*, X_*)^{-1} \end{pmatrix}$$

where $C(X_*, X_*)$ is the sample covariance matrix of the non-constant regressors, i.e., X_* is the matrix X with the first column of ones deleted. Here a is a 1×1-matrix, b is a column matrix and consequently b' is a row matrix. The covariances of the coefficients for the non-constant regressors are thus determined by their sample covariances and the estimated variance of the residual. (This fact does not appear in Hansen.)

- *Wald's test* is a follows: We want to test the null hypothesis $\mathrm{H}_0 : \ R\beta = \mu$. The test statistic is the variable

$$(R\hat{\beta} - \mu)'(R\hat{\Omega}R')^{-1}(R\hat{\beta} - \mu)$$

which is $\chi^2(r)$ if H_0 is true, where r is the number of rows in R. (Note that Hansen has R' for R.)

- *The Instrumental Variable Method and 2SLS.* Assume that we have the same number of instruments as explanatory variables. Let X be the matrix of observations on the explanatory variables x and Z the observations on the instrumental variables (recall that the instrumental variables comprise the exogenous x-variables.) **Note that Hansen reverses the notation:** for some strange reason he calls the explanatory variables (the variables entering the equation under study) Z and the instruments X. I think this is very confusing, and at odds with convention and the notation in previous chapters.

  The estimate of β is then obtained from $Z'\hat{e} = 0$, i.e.,

$$\hat{\beta} = (Z'X)^{-1}Z'Y$$

  (compare with the formula given first in section 9.5 in Hansen. Note that X and Z are swapped.) The estimated MSE (Mean Squared Error) matrix of $\hat{\beta}$ is given by

$$\hat{\Omega} = (Z'X)^{-1}[Z'D(\hat{e}^2)Z](X'Z)^{-1}$$

  This is Whites heteroskedasticity-consistent estimator. I can't find this in Hansen! The instrumental estimator is not unbiased, the MSE matrix measures the deviation between the estimated value and the *true* value rather than the *expected* value.

  When we have more instruments (z-variables) than regressors (x-variables) we do Two Stage Least Squares (2SLS), which is the following: regress each regressor on the instruments

$$X = Z\hat{\Gamma} + \hat{u} \quad \text{where} \quad Z'\hat{u} = 0$$

  and define $\hat{Z}$ by

$$\hat{Z} = Z\hat{\Gamma}$$

  Note that the columns of $\hat{Z}$ are the same as those in X except for the endogeneous one(s). $\hat{Z}$ will now have as many columns (instruments) as the explanatory variables. Use $\hat{Z}$ as instruments, and proceed as above.

- In order for the parameters in an IV or 2SLS regression to be *identified,* a necessary (but not sufficient) condition is that there are *at least* as many instruemnts as there are endogeneous variables (the *order condition.*) A further test is the following: for each endogeneous variable, regress it on *all* exogeneous regressors (exogeneous regressors plus instruments). Then test if the coefficients for the instruments are all equal to zero (a Wald test, if more than one instrument.) If this null can not be rejected, there is a problem with identification.

  If there is only one endogeneous variable, this test is also a *sufficient* test for identification (if there is at least one instrument, of course.)

Lemma

Here is a useful lemma that does not appear in Hansen: *Let Ω be a symmetric, positive definite $k \times k$ matrix, and M a $k \times j$-matrix. If A is any $j \times k$-matrix such that $AM = I$ (the identity-matrix), then*

$$A\Omega A' \geq A_0 \Omega A_0'$$

where

$$A_0 = (M'\Omega^{-1}M)^{-1}M'\Omega^{-1}.$$

Note that also A_0 satisfies $A_0 M = I$, and that $A_0 \Omega A_0' = (M'\Omega^{-1}M)^{-1}$.

Proof of the lemma:

Note first that

$$A\Omega A_0' = A\Omega\Omega^{-1}M(M'\Omega^{-1}M)^{-1} = (M'\Omega^{-1}M)^{-1} = A_0\Omega A_0'$$

hence also (transpose)

$$A_0\Omega A' = A_0\Omega A_0'$$

Employing these two relations we get

$$\begin{aligned} 0 \leq (A - A_0)\Omega(A - A_0)' &= A\Omega A' - A\Omega A_0' - A_0\Omega A' + A_0\Omega A_0' \\ &= A\Omega A' - A_0\Omega A_0' \end{aligned}$$

Q.E.D.

- *The GLS estimator* (Hansen ch. 5.1) can be derived as follows. We assume that the covariance matrix of e is Ω and estimate β by $Z'\hat{e} = 0$ for some suitable $n \times (k+1)$-matrix Z. We seek the optimal choice of Z. We get

$$\hat{\beta} = (Z'X)^{-1}Z'Y = \beta + (Z'X)^{-1}Z'e$$

so the covariance matrix of $\hat{\beta}$ is $(Z'X)^{-1}Z'\Omega Z(X'Z)^{-1}$.

Now define

$$A = (Z'X)^{-1}Z'$$

so that the covariance matrix is

$$A\Omega A'.$$

Note that $AX = I$. Hence, by the lemma above, in order to minimise the covariance matrix, we should choose

$$A = A_0 = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}$$

i.e., we should choose $Z' = X'\Omega^{-1}$. The GLS estimate is thus derived from

$$X'\Omega^{-1}\hat{e} = 0.$$

yielding the covariance matrix of $\hat{\beta}$

$$(X'\Omega^{-1}X)^{-1}$$

In summary: The GLS estimator is

$$\hat{\beta} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y$$

with estimated covariance matrix

$$(X'\Omega^{-1}X)^{-1}$$

The problem in practise is of course that Ω is typically unknown. In some cases it can be estimated, and the method is then called FGLS ("Feasible Generalised Least Squares".)

- *Goodness of fit.* (Hansen ch. 3.3.) In the model

$$y = x'\beta + e$$

the part $\tilde{y} = x'\beta$ is the part of the variation of y that is "explained" by x. In the usual regression model, where $\mathrm{E}[\, e \mid x] = 0$, x and e are uncorrelated, hence also $\tilde{y}$ and e are uncorrelated, so we can decompose the variance of y:

$$\mathrm{Var}(y) = \mathrm{Var}(\tilde{y}) + \mathrm{Var}(e)$$

and a measure of goodness of fit is the ratio $\dfrac{\mathrm{Var}(\tilde{y})}{\mathrm{Var}(y)}$. This ratio is in fact the square of the correlation coefficient between y and $\tilde{y}$. Indeed, it follows from $y = \tilde{y} + e$ and the fact that $\tilde{y}$ and e are uncorrelated that

$$\mathrm{Cov}(\tilde{y}, y) = \mathrm{Var}(\tilde{y})$$

hence, denoting the correlation coefficient between y and $\tilde{y}$ by ρ,

$$\rho^2 = \frac{\mathrm{Cov}(\tilde{y}, y)^2}{\mathrm{Var}(\tilde{y})\,\mathrm{Var}(y)} = \frac{\mathrm{Var}(\tilde{y})}{\mathrm{Var}(y)}$$

The estimate of ρ^2, when degrees of freedom have been appropriately taken into account, is called "adjuster R-square" and denoted $\bar{R}^2$. It is reported by every OLS software. It can be computed as

$$\bar{R}^2 = 1 - \frac{\frac{1}{n-k-1}\sum_1^n \hat{e}_j^2}{\frac{1}{n-1}\sum_1^n (y_j - \bar{y})^2} \quad \text{where} \quad \bar{y} = \tfrac{1}{n}\textstyle\sum_1^n y_j$$

Note that when x and e are correlated, as when instrumental variables are employed, $\bar{R}^2$ has little meaning. If it is reported at all, it has presumably been estimated as $\dfrac{\mathrm{Cov}(\tilde{y}, y)^2}{\mathrm{Var}(\tilde{y})\,\mathrm{Var}(y)}$, i.e., as the square of the correlation coefficient mentioned above.

- *Prediction.* Assume that we want to predict the value of y given values c on $x = (x_1 \dots x_k)'$. The best prediction $\tilde{y}$, in the sense of least MSE (Mean Squared Error) $\mathrm{E}[(y - \tilde{y})^2]$ conditional on x is the conditional mean: $\tilde{y} = \mathrm{E}[\, y \mid x]$. We specify a functional form $g(x; \beta)$ for $\tilde{y} = \mathrm{E}[\, y \mid x]$:

$$y = g(x; \beta) + e, \quad \mathrm{E}[\, e \mid x] = 0 \tag{1}$$

where β is a vector of unknown parameters. If $g(x, \beta)$ is linear in β this is a usual linear regression equation:

$$y = x'\beta + e, \quad \mathrm{E}[\, e \mid x] = 0 \tag{2}$$

and the estimation method is OLS or possibly GLS. We consider the linear case (2). The estimate produces estimated values $\hat{\beta}$ with estimated covariance matrix $\hat{\Omega}$ of β. For given $x = c$ the estimated predictor is then

$$\hat{y} = c'\hat{\beta} \quad \text{with MSE (or variance)} \quad c'\hat{\Omega}c$$

The MSE of the predict*ion* (as opposed to the predict*or*) is an estimate of $\mathrm{E}[(y - \hat{y})^2]$ and is the sum of the error in the estimate of $\hat{y}$ and the residual in (2):

$$\text{prediction MSE} = c'\hat{\Omega}c + \hat{\sigma}^2(c)$$

where $\hat{\sigma}^2(c)$ is an estimate of the residual in (2) when $x = c$. In the homoskedastic case, an unbiased estimate is

$$\hat{\sigma}^2 = \tfrac{1}{n-k-1}\sum_1^n \hat{e}^2$$

The RMSE of the prediction (Root Mean Squared Error, equivalent to the standard error) is thus

$$\text{SD} = \text{RMSE} = \sqrt{c'\hat{\Omega}c + \hat{\sigma}^2(c)}$$

Note that it is never appropriate to use instrumental variables for estimating a prediction equation! The reason is that in this case we seek the values of $\hat{\beta}$ that produces the highest $\bar{R}^2$; there is no other interpretation involved.

- *Non Linear Least Squares* (NLLS). Assume that we want to estimate the model

$$y = g(x;\beta) + e, \quad \text{E}[\, e \mid x] = 0 \tag{1}$$

where $g(x;\beta)$ is some function of the random variables $x = (x_1, \ldots, x_j)$ and $\beta = (\beta_0 \ \ldots \ \beta_k)'$ is a column vector of unknown parameters that we want to estimate. In the linear case, $g(x,\beta) = x'\beta$, i.e., $g(x,\beta)$ is linear *in the parameters* β.

I will use notation different from Hansen's: $g_\beta(x,\beta) = (\frac{\partial g(x;\beta)}{\partial \beta_0} \ \ldots \ \frac{\partial g(x;\beta)}{\partial \beta_k})$, i.e., $g_\beta(x,\beta)$ is a row vector (not a column vector, as in Hansen.) A first order Taylor expansion of β is thus denoted

$$g(x;\beta) - g(x;\hat{\beta}) = g_\beta(x;\hat{\beta})(\beta - \hat{\beta})$$

Let X and Y be the matrices of observations on n values $x_1, \ldots, x_n$ of x and $y_1, \ldots, y_n$ of y as before. I denote

$$G(X;\beta) = \begin{pmatrix} g(x_1;\beta) \\ \vdots \\ g(x_n;\beta) \end{pmatrix}$$

and

$$G_\beta(X;\beta) = \begin{pmatrix} \frac{\partial g(x_1;\beta)}{\partial \beta_0} & \cdots & \frac{\partial g(x_1;\beta)}{\partial \beta_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial g(x_n;\beta)}{\partial \beta_0} & \cdots & \frac{\partial g(x_n;\beta)}{\partial \beta_k} \end{pmatrix}$$

We now turn to the estimation of (1). Let Z be some $n \times (k+1)$-matrix. An MME is obtained by solving

$$0 = Z'\hat{e} = Z'\big(Y - G(X,\hat{\beta})\big)$$

for $\hat{\beta}$. We employ a first order Taylor approximation:

$$\begin{aligned} 0 = Z'\big(Y - G(X,\hat{\beta})\big) &= Z'\big(G(X;\beta) + e - G(X,\hat{\beta})\big) \\ &\approx Z'\big(G_\beta(X;\hat{\beta})(\beta - \hat{\beta}) + e\big) \end{aligned}$$

Ignoring the approximation, we get

$$\hat{\beta} - \beta = \big(Z' G_\beta(X;\hat{\beta})\big)^{-1} Z' e$$

and as usual we want to find the matrix Z which minimises the MSE. A calculation similar to what we have done before shows that $Z = G_\beta(X;\hat{\beta})$ is optimal if we have homoskedasticity. The NLLS estimator is thus obtained from

$$G_\beta(X;\hat{\beta})'\big(Y - G(X,\hat{\beta})\big) = 0 \tag{2}$$

and the MSE-matrix (covariance matrix, $\mathrm{E}[(\hat{\beta}-\beta)(\hat{\beta}-\beta)']$) can be estimated by

$$\hat{\Omega} = \big(G_\beta' G_\beta\big)^{-1} G_\beta'\, D(\hat{e}^2)\, G_\beta \big(G_\beta' G_\beta\big)^{-1}$$

where $$G_\beta = G_\beta(X;\hat{\beta})$$

which heteroskedasticity consistent.

Note that (2) is also the solution to the minimisation problem

$$\min_\beta \big(Y - G(X,\beta)'\big(Y - G(X,\beta)$$

which is of course why the estimator is called Non Linear Least Squares. In this non linear case, the minimisation problem might be simpler to solve than (2), since there are no derivatives involved.

In summary: The NLLS estimator of (1) is

$$\hat{\beta} = \operatorname*{argmin}_\beta \big(Y - G(X,\beta)'\big(Y - G(X,\beta)\big)$$

with estimated MSE-matrix

$$\hat{\Omega} = \big(G_\beta' G_\beta\big)^{-1} G_\beta'\, D(\hat{e}^2)\, G_\beta \big(G_\beta' G_\beta\big)^{-1}$$

and prediction MSE for an observation c on x

$$g_\beta(c;\hat{\beta})\, \hat{\Omega}\, g_\beta(c;\hat{\beta})' + \sigma^2(c)$$

- *NLLS with Instrumental Variables.* This doesn't appear in Hansen. But you should know that instrumental variables in conjunction with NLLS is feasible.

  Assume first that we have as many instrumental variables as parameters β. Denote the matrix of instruments Z as in the linear case. In this case the estimate $\hat{\beta}$ is obtained from

$$Z'\big(Y - G(X,\hat{\beta})\big) = 0$$

and $$\hat{\Omega} = \big(Z' G_\beta\big)^{-1} Z'\, D(\hat{e}^2)\, Z \big(G_\beta' Z\big)^{-1}$$

  The problem is to figure out what good instruments are. The instrument $z = (z_0 \ldots z_k)$ should be uncorrelated with e but well correlated with $g_\beta(x,\beta)$.

  If we have more instruments than parameters, then the formulas are

$$G_\beta(X,\hat{\beta})' Z (Z'Z)^{-1} Z'\big(Y - G(X,\hat{\beta})\big) = 0 \tag{1}$$

and $$\hat{\Omega} = \big(\hat{Z}' G_\beta\big)^{-1} \hat{Z}'\, D(\hat{e}^2)\, \hat{Z} \big(G_\beta' \hat{Z}\big)^{-1}$$

where $$\hat{Z} = Z(Z'Z)^{-1} Z' G_\beta$$

  The equation (1) can also be formulated as a minimisation problem, not involving derivatives:

$$\hat{\beta} = \operatorname*{argmin}_{\beta} \big(Y - G(X,\beta)\big)' Z (Z'Z)^{-1} Z' \big(Y - G(X,\beta)\big) \tag{1'}$$

The prediction error has the same formula as in the case without instruments.

The procedures and formulae I have presented in this section are not standard—I suppose different software programmes use different estimation methods for instrumental variable NLLS. The purpose of this section is just to make you aware of the possibility of using instrumental variables also in the non linear case.

- *Generalised NLLS.* Consider again the non linear equation

$$y = g(x;\beta) + e, \quad \mathrm{E}[\, e \mid x] = 0 \tag{1}$$

Just like in the linear case, if the covariance matrix of e is known, say V, then a more efficient estimation can be obtained. The calculations are analogous to those in the linear case, so I don't repeat them here. The end results are as follows:

To compute $\hat{\beta}$: $$G'_\beta V^{-1} \big(Y - G(X,\hat{\beta})\big) = 0$$

or equivalently: $$\hat{\beta} = \operatorname*{argmin}_{\beta} \big(Y - G(X,\beta)\big)' V^{-1} \big(Y - G(X,\beta)\big)$$

Covariance of $\hat{\beta}$: $$\big(G'_\beta V^{-1} G_\beta\big)^{-1}$$

- *Logit and Probit.* We want to estimate a model

$$p = p(x;\beta)$$

where p is the probability of an event occurring, depending on the values of x. Typically

$$p(x;\beta) = \begin{cases} \dfrac{\exp(x'\beta)}{1+\exp(x'\beta)} & \text{(Logit), or} \\ \Phi(x'\beta) & \text{(Probit)} \end{cases}$$

where Φ is the cumulative normal distribution. We estimate the parameters β with generalised NLLS in the regression model

$$y_i = p(x_i;\beta) + e_i \tag{1}$$
$$\mathrm{E}[\, e_i \mid x_i] = 0 \tag{2}$$
$$\mathrm{E}[\, e_i^2 \mid x_i] = p(x_i;\beta)(1 - p(x_i;\beta)) \tag{3}$$

where $y_i = 1$ if the event occurred and $y_i = 0$ if it did not occur. It is easy to verify (2) and (3). In this case we do not know V in advance, but once we have estimates $\hat{\beta}$ of β, we can estimate V:

$$\widehat{V} = \mathrm{diag.}\big(p(x_1;\hat{\beta})(1 - p(x_1;\hat{\beta}))\ \ldots\ p(x_n;\hat{\beta})(1 - p(x_n;\hat{\beta}))\big)$$

The equation for $\hat{\beta}$ thus becomes

$$P_\beta(X,\hat{\beta})' \widehat{V}^{-1} \big(Y - P(X,\hat{\beta})\big) = 0 \tag{4}$$

with covariance matrix

$$\Big(P_\beta(X,\hat{\beta})' \widehat{V}^{-1} P_\beta(X,\hat{\beta})\Big)^{-1} \tag{5}$$

The orthodox way of estimating (1) is by Maximum Likelihood but, as is easy to show, this is equivalent to the estimator given above.

With the logit specification these fomulas become particularly simple. Indeed, a simple calculation shows that in this case $p_\beta(x;\beta) = p(x;\beta)\big(1-p(x;\beta)\big)x'$, so (4) and (5) simplifies to

$$X'\big(Y - P(X,\hat\beta)\big) = 0 \tag{4'}$$

with covariance matrix

$$\Big(\sum_i p(x_i;\hat\beta)\big(1-p(x_i;\hat\beta)\big)x_i x_i'\Big)^{-1} \tag{5'}$$

- *Bootstrap.* It is a good idea to use bootstrap to test hypotheses and estimate confidence intervals in case of IV-estimation, and in particular for NLLS, LAD (Least Absolute Deviation) and Quantile Regression. Hansen shows how to test $\mathrm{H}_0 : \theta = \theta_0$ where θ is a vector of parameters in the section "Symmetric Percentile-t Intervals." I suggest using a similar method in the context of "Percentile intervals". The procedure is as follows: Let $\hat\theta(x)$ be our point estimate, and $\hat\theta(x_j^*),\ j = 1,\dots,b$ be the corresponding estimates from the b bootstrap resamples. Now estimate a covariance matrix for $\hat\theta$ (the accuracy of this estimate is not crucial):

$$\widehat V = \frac{1}{b}\sum_{j=1}^{b}\big(\hat\theta(x_j^*) - \hat\theta(x)\big)\big(\hat\theta(x_j^*) - \hat\theta(x)\big)'$$

Choose a risk level α and determine the constant c such that

$$\big(\hat\theta(x_j^*) - \hat\theta(x)\big)'\widehat V^{-1}\big(\hat\theta(x_j^*) - \hat\theta(x)\big) > c \quad \text{for } \alpha b \text{ of the } j\text{:s.}$$

Now an approximate confidence region for the true value θ at the level $1-\alpha$ is given by

$$\big(\hat\theta(x) - \theta\big)'\widehat V^{-1}\big(\hat\theta(x) - \theta\big) \le c$$

and the hypothesis $\mathrm{H}_0 : \theta = \theta_0$ is rejected with risk level α if

$$\big(\hat\theta(x) - \theta_0\big)'\widehat V^{-1}\big(\hat\theta(x) - \theta_0\big) > c$$

- *Model Selection.* Assume that we are interested in the coefficients β_1 in the regression equation

$$y = \beta_0 + x_1'\beta_1 + x_2'\beta_2 + e \tag{1}$$

Here i assume that the means of x_1 and x_2 are all equal to zero; their means are incorporated in the intercept β_0.

Is it then correct, a good idea or a bad idea, to estimate β_1 from the shorter equation

$$y = \beta_0 + x_1'\beta_1 + u\ ? \tag{2}$$

It is *incorrect* (i.e., gives the wrong estimate asymptotically) unless one (or both) of the following conditions are satisfied:

i) x_1 and x_2 are uncorrelated,

ii) $\beta_2 = 0$.

In case *i)* the estimate of (1) more efficient. The reason is that the covariance matrix for β_1 is (in the homoskedastic case) $(X_1'X_1)^{-1}\sigma^2(e)$ if estimated by (1), and $(X_1'X_1)^{-1}\sigma^2(u)$ if estimated by (2). But $\sigma^2(u) = \mathrm{Var}(x_2'\beta_2) + \sigma^2(e) > \sigma^2(e)$, so (1) is more efficient.

The case *ii)* is different. Since $\beta_2 = 0$ employing (1) rather than (2) means that we impose the restriction $\beta_2 = 0$ on (2), which of course enhances the efficiency—more information improves the estimate.

- *Self Selection Bias.* A not so uncommon problem is that we want to estimate an equation

$$y = x'\beta + e$$

but one of the regressors is endogenous through *self selection.* For example, assume I want to figure out if my teaching in this course does any good, or if I could just as well give you the course literature, and you all read for yourselves. In order to measure the impact on performance on the final test of attendance to class, I run an regression like

$$(\text{performance in final test}) = \beta_0 + (\text{attendance in class})\beta_1 + e \tag{1}$$

Assume that I, to my great disappointment, find that β_1 is negative. It appears that my teaching actually is detrimental rather than helpful. But it might be that it is the *very clever students who choose* not to attend class. I.e., those students who are very clever, and thus have a large residual, have a lower than average attendance. In that case, there is a negative correlation between "attendance" and the residual. This causes the OLS estimate of β_1 to be biased downwards.

I find it easiest to think about this as follows: If the equation (1) is estimated with OLS, it can be interpreted as a *prediction* equation. What performance do I predict for a student that do not attend class? If I believe that those who do not attend are more clever than average, then I might expect a good performance by those with a low attendance, i.e., a low—or even negative—value of β_1. Its value, however, would not fully reflect the impact of teaching on performance.

The selection mechanism causes a bias since the dummy variable "attendance in class" is correlated with the "unobserved heterogeneity" contained in the error term. A remedy is to find an instrument variable (or several) and employ IV or 2SLS.

Here is another type of self selection. Assume that we want to estimate a wage equation:

$$\ln(w) = x'\beta + e \tag{2}$$

(w = wage) but we only observe wages on those who work, obviously. We want to interpret (2) as the wage a person would get *if* she works, whether or not she does. A person chooses to work if the wage is above her reservation wage, and we have another equation for the reservation wage rw :

$$\ln(rw) = z'\gamma + u \tag{3}$$

We only observe w if $w > rw$, so in the sample used for estimating (2) we have

$$\begin{aligned} \mathrm{E}[\, e \mid x] &= \mathrm{E}[\, e \mid w > rw] \\ &= \mathrm{E}[\, e \mid x'\beta + e > z'\gamma + u] \\ &= \mathrm{E}[\, e \mid e > z'\gamma - x'\beta + u] > 0 \end{aligned}$$

so we have an endogeneity bias. In this case some of the observations with small (e.g., large negative) e have a tendency not to appear in the sample. If we include these observations in the sample, we have data for these dependent variables (the covariates,) but not for the corresponding dependent variable (the wage, if the person had chosen to work.) The data sample is *censored.* In cases like this, a popular estimation method is "Heckman's Lambda".

Hansen writes "The problem of sample selection arises when the sample is a non-random selection of potential observations. This occurs when the observed data is (sic! should be 'are') systematically different from the population of interest." *This is not quite true.* The problem arises if the selection is based on some other criteria than the covariates! For example, assume that I want to see if there is some difference in, say, risk of accidents for firemen, depending on their gender. "Gender" will then be one of my covariates in a (logit) regression, and it is a *good* strategy to select data so that about as many females as males are selected, even though the proportion of females in the entire population of firemen is much less than one half. The sample selection bias arises if the selection procedure favours some *unobserved* heterogeneity that influences the risk for accidents, e.g., females who are stronger than average among female firemen (and "strength" is not one of the covariates.)

- *Heckman's Lambda.* An orthodox way to estimate an equation which suffers from censoring by self selection is by a "Heckit", AKA "Heckman's lambda". I will not go into details, but the procedure is as follows. We want to estimate

$$\begin{gathered} y = x'\beta + e \\ \text{where } \mathrm{E}[\, e \mid x] \neq 0 \end{gathered} \tag{1}$$

The idea is that we should define a second equation, a selection equation, from which we can estimate $\mathrm{E}[\, e \mid x]$, say by some expression $\mathrm{E}[\, e \mid x] = \rho\lambda(x)$. Then $e = \rho\lambda(x) + u$, where $\mathrm{E}[\, u \mid x] = 0$ and we can write our equation (1)

$$y = x'\beta + \rho\lambda(x) + u, \quad \mathrm{E}[\, u \mid x] = 0 \tag{2}$$

which can be estimated by OLS. Here $\lambda(x)$ is to be estimated from the selection equation (see below), but ρ is estimated in (2). We can immediately see a problem with this approach: if λ is linear in x, then we have multicolinearity. The only way this can work is if λ is highly non-linear *and* the specification (1) is indeed perfectly correct, i.e., that there is no non-linearity in the true specification that we haven't taken into account. The situation improves if there are exogenous variables z that appear in the selection equation, so that $\lambda = \lambda(x, z)$. In this way ρ can be identified essentially via z.

The selection equation is typically defined as follows: let D be a dummy for selection, i.e., D=1 for all data in the sample. Then the model says that $D = 1$ if and only if $x'\gamma + v > 0$ where v is a $N(0, 1)$-variable; γ is to be estimated. The probability that $D = 1$ is thus $\mathrm{Prob}(v > -x'\gamma) = \Phi(x'\gamma)$ where Φ is the (cumulative) normal distribution function. Hence γ can be estimated by a Probit:

$$D = \Phi(x'\gamma) + \mathit{error}$$

The final assumption is then that v and e (the residual in (1)) have a joint normal distribution with covariance ρ. After some computations, one comes up with the specification (2) where λ has a known functional form.

The "Heckit" model is common knowledge in econometrics, so you must know about it. If you ever are going to use it, you can look up the details, but personally I would be very reluctant to employ it because of the heroic strong distribution assumptions. It is important, though, that you can recognise the problem with self selection. In any case, Heckman's two step procedure is inferior to a MLE (Maximum Likelihood Estimator,) which thus is a preferred alternative when available.

- *Tobit Equations.* Sometimes we have to work with data that are either *truncated* or *censored.* Data are *censored* when in some cases we don't know the true value of the dependent variable, but only that it is above (or below) some threshold. For example, assume that we want to estimate the demand for hotel rooms in some area. This demand varies with the current prices for rooms, with the season, and so on. In some cases, perhaps under the most popular holiday season, *all* rooms are booked, and then we don't know the real demand, only that it exceeds the actual capacity.

  Data are *truncated* if in the case the dependent variable is above (or below) some threshold we don't even know the corresponding explanatory variables. For example, in the wage equation (2) in the discussion on "Self Selection Bias", if we only have data on persons who work, we have truncated data.

  When data are censored, the *Tobit Model* assumes that the residual has a normal distribution, and a rather straightforward MLE can be employed; see ch. 12.3 in Hansen. There are several problems with this model, though, which are well explained by Peter Kennedy in his book "A Guide to Econometrics".

- *Duration Models.* Unfortunately there is nothing in Hansen about duration models, but I feel that the case with censored data is not fully treated if we leave them out. So here goes.

  A duration model is one where we want to estimate the duration of something, depending on some characteristics x. How long does it take for a patient to recover from some disease depending on some treatment; how long does a laid off person stay unemployed, given her characteristics, how long does it take for a KTH student to finish her studies, given her characteristics and the tuition offered, how long does it take for a certain product to break, given some production characteristics, and so on. Assume that the duration has an exponential distribution with intensity $\mu(x'\beta)$ for some function μ. Since μ must be positive, a possible choice is $\mu(x'\beta) = \ln\big(\exp(x'\beta) + 1\big)$. The probability density function for $y = \textit{duration}$ is then

$$f(y) = \mu \exp(-\mu y) \quad \text{and distribution function} \quad F(y) = 1 - \exp(-\mu y)$$

  Assume that we have a number of observations, some of which are censored, i.e., the duration still goes on, so we only know that the duration will exceed some value y^*. Denote the non-censored durations y_i, $i = 1, \ldots, n$, and the censored ones y_i^*, $i = n+1, \ldots, m$. The log likelihood is then (essentially)

$$L(\beta) = \sum_{i=1}^{n} \Big(\ln\big(\mu(x_i'\beta)\big) - \mu(x_i'\beta)\, y_i \Big) - \sum_{i=n+1}^{m} \mu(x_i'\beta)\, y_i^*$$

  and the MLE of β is thus the β that maximises this expression.

  A problem with this is that it does not take care of *duration dependence.* The exponential distribution has the property that the "hazard" is independent of the length of the duration spell so far. Indeed, assume that a spell has lasted for a time of t days, and consider the conditional probability that it will end within the the next Δt days. This probability is

$$\begin{aligned} \text{Prob}(y \le t + \Delta t \mid y > t) &= \frac{F(t+\Delta t) - F(t)}{1 - F(t)} \\ &= \frac{\exp(\mu t) - \exp\big(\mu(t+\Delta t)\big)}{\exp(\mu t)} \\ &= 1 - \exp(-\mu \Delta t) \end{aligned}$$

  which is independent of t.

An alternative approach, which also makes it easy to model duration dependence, is to measure the duration in discrete time steps. Let us choose the unit of time such that $\Delta t = 1$ is a suitable time step; a day or a week, or whatever, and denote $p(x'\beta) = 1 - \exp(-\mu(x'\beta))$. With our specification of μ this becomes

$$p(x'\beta) = \frac{\exp(x'\beta)}{1 + \exp(x'\beta)},$$

i.e., a logit specification. This is thus the probability that the spell will end during the next time period. Assume we are studying unemployment spells. We view "each individual as contributing not one but several observations to a giant logit likelihood function. In the first time period, each individual either stays or leaves unemployment, so a logit likelihood could be structured, with appropriate explanatory variables, to capture this. Now consider all the individuals who have not yet left the unemployment state and who have not become censored, namely all the individuals for whom it is possible to leave the unemployment state during the second time period. In the second time period each of these individuals either stays or leaves the state of unemployment, so a second logit likelihood, with the same explanatory variables (whose values could be different if they vary with time,) can be structured to capture this. Similar logit likelihoods can be formulated for each of the remaining time periods, with the number of observations contributing to these likelihoods diminishing as individuals are censored or leave the unemployment state. A giant likelihood can then be formed by multiplying together all these separate-period likelihoods. Each individual contributes several terms to this giant likelihood, one term for each time period for which that individual was at risk of leaving the unemployment state.

A baseline hazard can be built into this specification by including a function of time among the explanatory variables. Alternatively, we could allow the intercept in each of the separate-period logit formulations to be different. If there are a a total of k time periods, k dummy variables, one for each period, (taking the value one for that period and zero for all other periods,) are entered as additional explanatory variables in the logit specification in place of the intercept. These dummy variables allow each duration length to contribute to the intercept to the logit specification separately, thereby modeling a completely unrestricted baseline hazard." [Citation from P. Kennedy's book *"A Guide to Econometrics,"* Blackwell Publishing.]

Note that we still have a problem with *unobserved heterogeneity.* Those with a "bad" unobserved heterogeneity, i.e., those who have a small hazard, are more likely to appear many times in the likelihood, thus creating a bias in the estimate.

- *Transformation of Dependent Variable.* Consider a model

$$y = x'\beta + \varepsilon \tag{1}$$

where the dependent variable y can only take positive values. It is then often advisable to consider an alternative model where y is transformed by a logarithm:

$$\ln y = x'\gamma + u \tag{2}$$

For instance, wage equations, where y is wage, are often formulated in this way. Three things to consider:

1. The interpretation of β and γ are quite different. If x_1 increases by one unit in equation (1), then y increases by β_1 units, so the dimension of β_1 is "y-units per x_1-unit". However, an increase by x_1 by one unit in equation (2) means that $\ln y$ increases by γ_1, i.e., that y is multiplied by e^{γ_1}. Hence, γ_1 is dimension-less.

2. Note that in equation (2), it is *not true* that

   $\mathrm{E}[y] = e^{x'\gamma}$. Indeed, if u has a Normal distribution, the true relation is that

   $$\mathrm{E}[y] = e^{\frac{1}{2}\sigma^2 + x'\gamma}, \quad \text{where } \sigma^2 = \mathrm{E}[\, u^2 \mid x].$$

3. The specification (2) is particularly suitable when we believe that (1) suffers from heteroskedasticity, and that the standard error of ε is proportional to y. Indeed, in this case we could write (1) as

   $$y = x'\beta(1 + v)$$

   where the variance of v is approximately independent of x. Taking logarithms gives

   $$\ln y = \ln(x'\beta) + \ln(1 + v) = \ln(x'\beta) + u$$

   where $u = \ln(1 + v)$, and the variance of u is independent of x. This is thus a homoskedastic equation, and we might consider replacing $\ln(x'\beta)$ by the linear specification $x'\gamma$ to get (2).

A common situation is that y can only take values between zero and one. This is the case for instance if y is a fraction. The model (1) can then easily suffer from the problem that $x'\beta$ for some reasonable values of x takes values outside of this range. A common transformation for y is then the logistic:

$$\ln\left(\frac{y}{1-y}\right) = x'\gamma + u$$

Note that as y varies from zero to one, the transformed value of y varies from $-\infty$ to $+\infty$.