Regression

Class Notes

Manuel Arellano

January 20, 2017

1 Means and predictors

Given some data $\{y_1, ..., y_n\}$ we could calculate a mean $\overline{y} = (1/n) \sum_{i=1}^{n} y_i$ as a single quantity that summarizes the n data points. $\overline{y}$ is an optimal predictor that minimizes mean squared error:

$$\overline{y} = \arg\min_a \sum_{i=1}^{n} (y_i - a)^2 .$$

Now if we have data on two variables for the same units $\{y_i, x_i\}_{i=1}^{n}$, we can get a better predictor of y using the additional information in x calculating the regression line $\widehat{y}_i = \widehat{a} + \widehat{b} x_i$ where

$$\left(\widehat{a}, \widehat{b}\right) = \arg\min_{a,b} \sum_{i=1}^{n} (y_i - a - b x_i)^2 .$$

More generally, if x_i is a vector $x_i = (1, x_{2i}, ..., x_{ki})'$, we calculate the linear predictor $\widehat{y}_i = x_i'\widehat{\beta}$ where

$$\widehat{\beta} = \arg\min_b \sum_{i=1}^{n} \left(y_i - x_i' b\right)^2 . \tag{1}$$

The algebra of linear predictors First order conditions of (1) are

$$\sum_{i=1}^{n} x_i \left(y_i - x_i'\widehat{\beta}\right) = 0. \tag{2}$$

If $\sum_{i=1}^{n} x_i x_i'$ is full rank (which requires $n \geq k$) there is a unique solution:

$$\widehat{\beta} = \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1} \sum_{i=1}^{n} x_i y_i. \tag{3}$$

We may use the compact notation $X'X = \sum_{i=1}^{n} x_i x_i'$ and $X'y = \sum_{i=1}^{n} x_i y_i$ where $y = (y_1, ..., y_n)'$ and $X = (x_1, ..., x_n)'$.

Denoting residuals as $\widehat{u}_i = y_i - x_i'\widehat{\beta}$, from the first order conditions (2) we can immediately say that as long as a constant term is included in x_i:

$$\frac{1}{n} \sum_{i=1}^{n} \widehat{u}_i = 0, \quad \frac{1}{n} \sum_{i=1}^{n} x_{ji} \widehat{u}_i = 0 \text{ for } j = 2, ..., k.$$

Therefore, the mean of the residuals is zero and the covariance between the residuals and each of the x variables is also zero. Moreover, since $\widehat{y}_i$ is a linear combination of x_i, the covariance between $\widehat{u}_i$ and $\widehat{y}_i$ is also zero. We conclude that a linear regression decomposes y_i into two orthogonal components:

$$y_i = \widehat{y}_i + \widehat{u}_i,$$

so that $\widehat{Var}(y_i) = \widehat{Var}(\widehat{y}_i) + \widehat{Var}(\widehat{u}_i)$. An R^2 measures the fraction of the variance of y_i that is accounted by $\widehat{y}_i$:

$$R^2 = \frac{\widehat{Var}(\widehat{y}_i)}{\widehat{Var}(y_i)}.$$

2 Consistency and asymptotic normality of linear predictors

If our data $\{y_i, x_i\}_{i=1}^n$ are a random sample from some population we can study the properties of $\widehat{\beta}$ as an estimator of the corresponding population quantity:

$$\beta = \left[E\left(x_i x_i'\right)\right]^{-1} E\left(x_i y_i\right), \tag{4}$$

where we require that $E\left(x_i x_i'\right)$ has full rank.

Letting the population linear predictor error be $u_i = (y_i - x_i'\beta)$, the estimation error is

$$\widehat{\beta} - \beta = \left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^n x_i u_i.$$

Clearly, $E\left(x_i u_i\right) = 0$, since β solves the first-order conditions $E\left[x_i\left(y_i - x_i'\beta\right)\right] = 0$. By Slutsky's theorem and the law of large numbers:

$$\operatorname*{plim}_{n\to\infty}\left(\widehat{\beta} - \beta\right) = \left(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1} \operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^n x_i u_i = \left[E\left(x_i x_i'\right)\right]^{-1} E\left(x_i u_i\right) = 0. \tag{5}$$

Therefore, $\widehat{\beta}$ is a consistent estimator of β.

Moreover, because of the central limit theorem

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n x_i u_i \xrightarrow{d} \mathcal{N}\left(0, V\right)$$

where $V = E\left(u_i^2 x_i x_i'\right)$. In addition, using Cramér's theorem we can assert that

$$\sqrt{n}\left(\widehat{\beta} - \beta\right) \xrightarrow{d} \mathcal{N}\left(0, W\right) \tag{6}$$

where

$$W = \left[E\left(x_i x_i'\right)\right]^{-1} E\left(u_i^2 x_i x_i'\right)\left[E\left(x_i x_i'\right)\right]^{-1}, \tag{7}$$

and also for individual coefficients:

$$\sqrt{n}\left(\widehat{\beta}_j - \beta_j\right) \xrightarrow{d} \mathcal{N}\left(0, w_{jj}\right) \tag{8}$$

where w_{jj} is the j-th diagonal element of W.

Asymptotic standard errors and confidence intervals A consistent estimator of W is:

$$\widehat{W} = \left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \widehat{u}_i^2 x_i x_i'\right)\left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1}. \tag{9}$$

The quantity $\sqrt{\widehat{w}_{jj}/n}$ is called an asymptotic standard error of $\widehat{\beta}_j$, or simply a standard error. It is an approximate standard deviation of $\widehat{\beta}_j$ in a large sample, and it is used as a measure of the precision of an estimate.

Due to Cramér's theorem:

$$\frac{\widehat{\beta}_j - \beta_j}{\sqrt{\widehat{w}_{jj}/n}} \xrightarrow{d} \mathcal{N}(0,1). \tag{10}$$

The use of this statement is in calculating approximate confidence intervals. A 95% large sample confidence interval is:

$$\left(\widehat{\beta}_j - 1.96\sqrt{\widehat{w}_{jj}/n},\ \widehat{\beta}_j + 1.96\sqrt{\widehat{w}_{jj}/n}\right). \tag{11}$$

3 Classical regression model

A linear predictor is the best linear approximation to the conditional mean of y given x in the sense:

$$\beta = \arg\min_b E\left\{\left[E(y_i \mid x_i) - x_i'b\right]^2\right\}. \tag{12}$$

That is, $x_i'\beta$ minimizes the mean squared approximation errors where the mean is taken with respect to the distribution of x. Therefore, changing the distribution of x will change the linear predictor unless the conditional mean is linear, in which case $E(y_i \mid x_i) = x_i'\beta$.

If $E\left\{\left[E(y_i \mid x_i) - x_i'\beta\right]^2\right\}$ is not zero or close to zero, $x_i'\widehat{\beta}$ will not be a very informative summary of the dependence in mean between y and x. In general, the use of a linear predictor is hard to motivate if the conditional mean is notoriously nonlinear.

The classical regression model is a linear model that makes the following two assumptions:

$$E(y \mid X) = X\beta \tag{A1}$$

$$Var(y \mid X) = \sigma^2 I_n. \tag{A2}$$

The first assumption (A1) asserts that $E(y_i \mid x_1, ..., x_n) = x_i'\beta$ for all i. This assumption contains two parts. The first one is that $E(y_i \mid x_1, ..., x_n) = E(y_i \mid x_i)$; this part of the assumption will always hold if $\{y_i, x_i\}_{i=1}^n$ is a random sample and is sometimes called strict exogeneity. The second part is the linearity assumption $E(y_i \mid x_i) = x_i'\beta$. Under A1 $\widehat{\beta}$ is an unbiased estimator:

$$E\left(\widehat{\beta} \mid X\right) = \left(X'X\right)^{-1} X'E(y \mid X) = \beta \tag{13}$$

and therefore also $E\left(\widehat{\beta}\right) = \beta$ by the law of iterated expectations.

The second assumption (A2) says that $Var(y_i \mid x_1, ..., x_n) = \sigma^2$ and $Cov(y_i, y_j \mid x_1, ..., x_n) = 0$ for all i and j. Under random sampling $Var(y_i \mid x_1, ..., x_n) = Var(y_i \mid x_i)$ and $Cov(y_i, y_j \mid x_1, ..., x_n) = 0$ always hold. Assumption A2 also requires that $Var(y_i \mid x_i)$ is constant for all x_i and this situation is called homoskedasticity. The alternative situation when $Var(y_i \mid x_i)$ may vary with x_i is called heteroskedasticity. When the data are time series the zero covariance condition $Cov(y_i, y_j \mid x_1, ..., x_n) = 0$ is called lack of autocorrelation.

Under A2 the variance matrix of $\widehat{\beta}$ given X is

$$Var\left(\widehat{\beta} \mid X\right) = \sigma^2 \left(X'X\right)^{-1}. \tag{14}$$

Moreover, under A2 since $E\left(u_i^2 x_i x_i'\right) = \sigma^2 E\left(x_i x_i'\right)$ the sandwich formula (7) becomes

$$W = \sigma^2 \left[E\left(x_i x_i'\right)\right]^{-1}. \tag{15}$$

To obtain an unbiased estimator of σ^2 note that under A2, letting $M = I_n - X\left(X'X\right)^{-1} X'$, we have

$$E\left(\widehat{u}'\widehat{u}\right) = E\left[E\left(u'Mu \mid X\right)\right] = E\left(tr\left[ME\left(uu' \mid X\right)\right]\right) = \sigma^2 tr\left(M\right) = \sigma^2\left(n-k\right), \tag{16}$$

so that an unbiased estimator of σ^2 is given by the degrees of freedom corrected residual variance:

$$\widehat{\sigma}^2 = \frac{\widehat{u}'\widehat{u}}{n-k}. \tag{17}$$

Sampling distributions under conditional normality Consider as a third assumption:

$$y \mid X \sim \mathcal{N}\left(X\beta, \sigma^2 I_n\right). \tag{A3}$$

Under A3:

$$\widehat{\beta} \mid X \sim \mathcal{N}\left(\beta, \sigma^2\left(X'X\right)^{-1}\right), \tag{18}$$

so that also

$$\widehat{\beta}_j \mid X \sim \mathcal{N}\left(\beta_j, \sigma^2 a_{jj}\right) \tag{19}$$

where a_{jj} is the j-th diagonal element of $(X'X)^{-1}$. Moreover, conditionally and unconditionally we have

$$z_j \equiv \frac{\widehat{\beta}_j - \beta_j}{\sqrt{\sigma^2 a_{jj}}} \sim \mathcal{N}\left(0,1\right). \tag{20}$$

This result, which holds exactly for the normal classical regression model, also holds under homoskedasticity as a large-sample approximation for linear predictors and non-normal populations, in light of (8), (15), and Cramér's theorem.

Heteroskedasticity-consistent standard errors Note that the validity of the large sample results in (9), (10) and (11) does not require homoskedasticity. This is why the asymptotic standard errors $\sqrt{\widehat{w}_{jj}/n}$ calculated from (9) are usually called heteroskedasticity-consistent or White standard errors, after the work of Halbert White.

Other distributional results The other key exact distributional results in this context are

$$\frac{\widehat{u}'\widehat{u}}{\sigma^2} \sim \chi^2_{n-k} \quad \text{independent of } z_j \tag{21}$$

and

$$\frac{\widehat{\beta}_j - \beta_j}{\sqrt{\widehat{\sigma}^2 a_{jj}}} \sim t_{n-k}. \tag{22}$$

In addition, letting now $\widehat{\beta}_j$ denote a subset of r coefficients and A_{jj} the corresponding submatrix of $(X'X)^{-1}$, we have

$$\frac{\left(\widehat{\beta}_j - \beta_j\right)' A_{jj}^{-1} \left(\widehat{\beta}_j - \beta_j\right)}{\sigma^2} \sim \chi^2_r \tag{23}$$

and

$$\frac{\left(\widehat{\beta}_j - \beta_j\right)' A_{jj}^{-1} \left(\widehat{\beta}_j - \beta_j\right)/r}{\widehat{\sigma}^2} \sim F_{r,(n-k)}. \tag{24}$$

4 Weighted least squares

The ordinary least squares (OLS) statistic $\widehat{\beta}$ is a function of simple means of $x_i x_i'$ and $x_i y_i$. Under heteroskedasticity it may make sense to consider weighted means in which observations with a smaller variance receive a larger weight. Let us consider estimators of the form

$$\widetilde{\beta} = \left(\sum_{i=1}^n w_i x_i x_i'\right)^{-1} \sum_{i=1}^n w_i x_i y_i \tag{25}$$

where w_i are some weights. OLS is the special case in which $w_i = 1$ for all i.

Under appropriate regularity conditions

$$\text{plim}\left(\widetilde{\beta} - \beta\right) = \left[E\left(w_i x_i x_i'\right)\right]^{-1} E\left(w_i x_i u_i\right). \tag{26}$$

Thus, in general to ensure consistency of $\widetilde{\beta}$ we need that $E(w_i x_i u_i) = 0$. This result will hold if $E(u_i \mid x_i) = 0$ and $w_i = w(x_i)$ is a function of x_i only:

$$E(w_i x_i u_i) = E(w_i x_i E(u_i \mid x_i)) = 0,$$

but more generally $\widetilde{\beta}$ is not a consistent estimator of the population linear projection coefficient β when $E(y_i \mid x_i) \neq x_i'\beta$.[1]

Subject to consistency, the asymptotic normality result is

$$\sqrt{n}\left(\widetilde{\beta} - \beta\right) \xrightarrow{d} \mathcal{N}\left(0, \left[E\left(w_i x_i x_i'\right)\right]^{-1} E\left(u_i^2 w_i^2 x_i x_i'\right) \left[E\left(w_i x_i x_i'\right)\right]^{-1}\right). \tag{27}$$

[1] Actually, if x_i has density $f(x)$, $\widetilde{\beta}$ is consistent for the optimal linear predictor under an alternative probability distribution of x_i given by $g(x) \propto f(x) w(x)$.

Asymptotic efficiency When weights are chosen to be proportional to the reciprocal of $\sigma_i^2 = E\left(u_i^2 \mid x_i\right)$, the asymptotic variance in (27) becomes

$$\left[E\left(\frac{x_i x_i'}{\sigma_i^2}\right)\right]^{-1}. \tag{28}$$

Moreover, it can be shown that for any (conformable) vector q:

$$q'\left[E\left(w_i x_i x_i'\right)\right]^{-1} E\left(\sigma_i^2 w_i^2 x_i x_i'\right)\left[E\left(w_i x_i x_i'\right)\right]^{-1} q \geq q'\left[E\left(\frac{x_i x_i'}{\sigma_i^2}\right)\right]^{-1} q. \tag{29}$$

Statement (29) says that the asymptotic variance of any linear combination of weighted LS estimates $q'\widetilde{\beta}$ is the smallest when the weights are $w_i \propto 1/\sigma_i^2$. To prove (29) note that[2]

$$E\left(\frac{x_i x_i'}{\sigma_i^2}\right) - E\left(w_i x_i x_i'\right)\left[E\left(\sigma_i^2 w_i^2 x_i x_i'\right)\right]^{-1} E\left(w_i x_i x_i'\right) = H' E\left(m_i m_i'\right) H \tag{30}$$

where

$$H = \begin{pmatrix} I \\ -\left[E\left(\sigma_i^2 w_i^2 x_i x_i'\right)\right]^{-1} E\left(w_i x_i x_i'\right) \end{pmatrix}, \quad m_i = \begin{pmatrix} \frac{x_i}{\sigma_i} \\ \sigma_i w_i x_i \end{pmatrix}.$$

Also note that for any q we have $q'\left[H' E\left(m_i m_i'\right) H\right] q \geq 0$.

Generalized least squares In view of (29) we can say that the estimator

$$\widetilde{\beta}_{GLS} = \left(\sum_{i=1}^n \frac{x_i x_i'}{\sigma_i^2}\right)^{-1} \sum_{i=1}^n \frac{x_i y_i}{\sigma_i^2} \tag{31}$$

is asymptotically efficient in the sense of having the smallest asymptotic variance among the class of consistent weighted least squares estimators. $\widetilde{\beta}_{GLS}$ is a generalized least squares estimator (GLS).

In matrix notation:

$$\widetilde{\beta}_{GLS} = \left(X'\Omega^{-1} X\right)^{-1} X'\Omega^{-1} y \tag{32}$$

where $\Omega = diag\left(\sigma_1^2, ..., \sigma_n^2\right)$.

In a generalized classical regression model we have $E\left(y \mid X\right) = X\beta$ and $Var\left(y \mid X\right) = \Omega$.

The asymptotic normality result is

$$\sqrt{n}\left(\widetilde{\beta}_{GLS} - \beta\right) \xrightarrow{d} \mathcal{N}\left(0, \left[E\left(\frac{x_i x_i'}{\sigma_i^2}\right)\right]^{-1}\right). \tag{33}$$

Usually $\widetilde{\beta}_{GLS}$ is an infeasible estimator because σ_i^2 is an unknown function of x_i. In a feasible GLS estimation σ_i^2 is replaced by a (parametric or nonparametric) estimated quantity. The large-sample properties of the resulting estimator may or may not coincide with those of the infeasible GLS.

---

[2] We are using the fact that if A and B are positive definite matrices, then $A - B$ is positive definite if and only if $B^{-1} - A^{-1}$ is positive definite.

5 Cluster-robust standard errors

Suppose the sample $\{y_i, x_i\}_{i=1}^n$ consists of H groups or clusters of M_h observations each ($n = M_1 + ... + M_H$), such that observations are independent across groups but dependent within groups, H is large and M_h is small (fixed) for all h. For convenience let us order observations by groups and use a double-index notation (y_{hm}, x_{hm}) for $h = 1, ..., H$(group index) and $m = 1, ..., M_h$ (within group index).

The compact notation for linear regression was $y = X\beta + u$. A similar notation for the observations in cluster h is

$$y_h = X_h\beta + u_h \tag{34}$$

where $y_h = (y_{h1}, ..., y_{hM_h})'$, etc. Using this notation the OLS estimator is

$$\widehat{\beta} = \left(X'X\right)^{-1} X'y = \left(\sum_{h=1}^{H} X_h' X_h\right)^{-1} \sum_{h=1}^{H} X_h' y_h. \tag{35}$$

Note that in terms of individual observations we can write $X'y = \sum_{h=1}^{H}\sum_{m=1}^{M_h} x_{hm}y_{hm}$, etc.

The scaled estimation error is

$$\sqrt{H}\left(\widehat{\beta} - \beta\right) = \left(\frac{X'X}{H}\right)^{-1} \frac{1}{\sqrt{H}} \sum_{h=1}^{H} X_h' u_h.$$

Applying the central limit theorem at cluster level, a consistent estimate of the variance of $\sqrt{H}\left(\widehat{\beta} - \beta\right)$ is given by

$$\left(\frac{X'X}{H}\right)^{-1} \frac{1}{H} \sum_{h=1}^{H} X_h' \widehat{u}_h \widehat{u}_h' X_h \left(\frac{X'X}{H}\right)^{-1}, \tag{36}$$

so that cluster-robust standard errors can be obtained as the square roots of the diagonal elements of the covariance matrix

$$\widehat{Var}\left(\widehat{\beta}\right) = \left(X'X\right)^{-1} \left(\sum_{h=1}^{H} X_h' \widehat{u}_h \widehat{u}_h' X_h\right) \left(X'X\right)^{-1}. \tag{37}$$

This is the sandwich formula associated with clustering. Its rationale is as a large H approximation. There are many applications of this tool, both with actual cluster survey designs and with other data sets with potential group-level dependence.