1 Means and predictors

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

\[
\bar{y} = \arg \min_{a} \sum_{i=1}^{n} (y_i - a)^2.
\]

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 \( \hat{y}_i = \hat{a} + \hat{b} x_i \) where

\[
\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \arg \min_{a,b} \sum_{i=1}^{n} (y_i - a - bx_i)^2.
\]

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

\[
\hat{\beta} = \arg \min_{\beta} \sum_{i=1}^{n} (y_i - x_i' \beta)^2.
\]

The algebra of linear predictors  

First order conditions of (1) are

\[
\sum_{i=1}^{n} x_i \left( y_i - x_i' \hat{\beta} \right) = 0.
\]

(2)

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

\[
\hat{\beta} = \left( \sum_{i=1}^{n} x_i x_i' \right)^{-1} \sum_{i=1}^{n} x_i y_i.
\]

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

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

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 \( \hat{y}_i \) is a linear combination of \( x_i \), the covariance between \( \hat{u}_i \) and \( \hat{y}_i \) is also zero. We conclude that a linear regression decomposes \( y_i \) into two orthogonal components:

\[
y_i = \hat{y}_i + \hat{u}_i,
\]

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

\[
R^2 = \frac{\text{Var}(\hat{y}_i)}{\text{Var}(y_i)}.
\]
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 \( \hat{\beta} \) as an estimator of the corresponding population quantity:

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

where we require that \( E (x_i x'_i) \) has full rank.

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

\[
\hat{\beta} - \beta = \left( \frac{1}{n} \sum_{i=1}^n x_i x'_i \right)^{-1} \frac{1}{n} \sum_{i=1}^n x_i u_i.
\]

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

\[
\text{plim}_{n \to \infty} \left( \hat{\beta} - \beta \right) = \left( \text{plim}_{n \to \infty} \frac{1}{n} \sum_{i=1}^n x_i x'_i \right)^{-1} \frac{1}{n} \sum_{i=1}^n x_i u_i = \left[ E (x_i x'_i) \right]^{-1} E (x_i u_i) = 0.
\]

Therefore, \( \hat{\beta} \) is a consistent estimator of \( \beta \).

Moreover, because of the central limit theorem

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

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

\[
\sqrt{n} \left( \hat{\beta} - \beta \right) \xrightarrow{d} \mathcal{N} (0, W)
\]

where

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

and also for individual coefficients:

\[
\sqrt{n} \left( \hat{\beta}_j - \beta_j \right) \xrightarrow{d} \mathcal{N} (0, w_{jj})
\]

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

**Asymptotic standard errors and confidence intervals** A consistent estimator of \( W \) is:

\[
\hat{W} = \left( \frac{1}{n} \sum_{i=1}^n x_i x'_i \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n u_i^2 x_i x'_i \right) \left( \frac{1}{n} \sum_{i=1}^n x_i x'_i \right)^{-1}.
\]

The quantity \( \sqrt{w_{jj}/n} \) is called an asymptotic standard error of \( \hat{\beta}_j \), or simply a standard error. It is an approximate standard deviation of \( \hat{\beta}_j \) in a large sample, and it is used as a measure of the precision of an estimate.
Due to Cramér’s theorem:

\[
\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{w}_{jj}/n}} \xrightarrow{d} N(0, 1).
\] (10)

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

\[
\left(\hat{\beta}_j - 1.96\sqrt{\hat{w}_{jj}/n}, \; \hat{\beta}_j + 1.96\sqrt{\hat{w}_{jj}/n}\right).
\] (11)

3 Classical regression model

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

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

If \( E \left\{ \left[ E(y_i | x_i) - x_i' \beta \right]^2 \right\} \) is not zero or close to zero, \( x_i' \beta \) 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:

\[
E(y | X) = X\beta \quad \text{(A1)}
\]

\[
Var(y | X) = \sigma^2 I_n. \quad \text{(A2)}
\]

The first assumption (A1) asserts that \( E(y_i | x_1, ..., x_n) = x_i' \beta \) for all \( i \). This assumption contains two parts. The first one is that \( E(y_i | x_1, ..., x_n) = E(y_i | x_i) \); 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 \( E(y_i | x_i) = x_i' \beta \). Under A1 \( \hat{\beta} \) is an unbiased estimator:

\[
E(\hat{\beta} | X) = (X'X)^{-1} X' E(y | X) = \beta
\] (13)

and therefore also \( E(\beta) = \beta \) by the law of iterated expectations.

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

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

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

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

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

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

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

$$\hat{\sigma}^2 = \frac{\hat{u}' \hat{u}}{n - k}. \quad (17)$$

**Sampling distributions under conditional normality** Consider as a third assumption:

$$y \mid X \sim N \left( X\beta, \sigma^2 I_n \right). \quad (A3)$$

Under A3:

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

so that also

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

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

$$z_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 a_{jj}}} \sim N \left( 0, 1 \right). \quad (20)$$

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 $\sqrt{w_{jj}/n}$ 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
\[
\frac{\hat{u}'\hat{u}}{\sigma^2} \sim \chi^2_{n-k} \text{ independent of } z_j
\] (21)
and
\[
\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 a_{jj}}} \sim t_{n-k}.
\] (22)
In addition, letting now \( \hat{\beta}_j \) denote a subset of \( r \) coefficients and \( A_{jj} \) the corresponding submatrix of \((X'X)^{-1}\), we have
\[
\frac{(\hat{\beta}_j - \beta_j)' A_{jj}^{-1} (\hat{\beta}_j - \beta_j)}{\sigma^2} \sim \chi^2_r
\] (23)
and
\[
\frac{(\hat{\beta}_j - \beta_j)' A_{jj}^{-1} (\hat{\beta}_j - \beta_j) / r}{\sigma^2} \sim F_{r,(n-k)}.\] (24)

4 Weighted least squares

The ordinary least squares (OLS) statistic \( \hat{\beta} \) 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
\[
\tilde{\beta} = \left(\sum_{i=1}^n w_i x_i x_i' \right)^{-1} \sum_{i=1}^n w_i x_i y_i \tag{25}
\]
where \( w_i \) are some weights. OLS is the special case in which \( w_i = 1 \) for all \( i \).

Under appropriate regularity conditions
\[
\text{plim} \left( \tilde{\beta} - \beta \right) = \left[ E \left( w_i x_i x_i' \right) \right]^{-1} E \left( w_i x_i u_i \right).
\] (26)

Thus, in general to ensure consistency of \( \tilde{\beta} \) 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:
\[
E \left( w_i x_i u_i \right) = E \left( w_i x_i E \left( u_i \mid x_i \right) \right) = 0,
\]
but more generally \( \tilde{\beta} \) is not a consistent estimator of the population linear projection coefficient \( \beta \) when \( E(y_i \mid x_i) \neq x_i' \beta \).\(^1\)

Subject to consistency, the asymptotic normality result is
\[
\sqrt{n} \left( \tilde{\beta} - \beta \right) \xrightarrow{d} N \left( 0, \left[ E \left( w_i x_i x_i' \right) \right]^{-1} E \left( w_i^2 x_i^2 x_i x_i' \right) \left[ E \left( w_i x_i x_i' \right) \right]^{-1} \right).
\] (27)
\(^1\) Actually, if \( x_i \) has density \( f(x) \), \( \tilde{\beta} \) 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 $\sigma_i^2 = E(u_i^2 \mid x_i)$, the asymptotic variance in (27) becomes

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

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

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

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

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

where

$$H = \begin{pmatrix} I \\ - E(\sigma_i^2 w_i^2 x_i x_i')^{-1} E(w_i x_i x_i') \end{pmatrix}, \quad m_i = \begin{pmatrix} x_i \\ \sigma_i w_i x_i \end{pmatrix}. \quad \text{Also note that for any } q \text{ we have } q' [H'E(m_i m_i') H] q \geq 0.$$

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

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

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

In matrix notation:

$$\tilde{\beta}_{GLS} = (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} y \quad \text{(32)}$$

where $\Omega = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$.

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

The asymptotic normality result is

$$\sqrt{n} \left( \tilde{\beta}_{GLS} - \beta \right) \overset{d}{\to} N \left( 0, \left[ E \left( \frac{x_i x_i'}{\sigma_i^2} \right) \right]^{-1} \right). \quad \text{(33)}$$

Usually $\tilde{\beta}_{GLS}$ is an infeasible estimator because $\sigma_i^2$ is an unknown function of $x_i$. In a feasible GLS estimation $\sigma_i^2$ 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

\[
y_h = X_h\beta + u_h
\]

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

\[
\hat{\beta} = (X'X)^{-1} X'y = \left( \sum_{h=1}^{H} X_h'X_h \right)^{-1} \sum_{h=1}^{H} X_h'y_h.
\]

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

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

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

\[
\left( \frac{X'X}{H} \right)^{-1} \left( \frac{1}{H} \sum_{h=1}^{H} X_h'\hat{u}_h\hat{u}_h'X_h \right) \left( \frac{X'X}{H} \right)^{-1},
\]

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

\[
\text{Var} \left( \hat{\beta} \right) = (X'X)^{-1} \left( \sum_{h=1}^{H} X_h'\hat{u}_h\hat{u}_h'X_h \right) (X'X)^{-1}.
\]

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.