LECTURE 6: MORE ON THE K-VARIABLE LINEAR MODEL

COMPUTATION AND DISTRIBUTION OF CONSTRAINED ESTIMATORS:

Consider the null hypothesis $H_0 : R\beta = r$ where $R$ is $q \times k$ and $r$ is $q \times 1$. We suppose there are genuinely $q$ restrictions under $H_0$, so rank $(R) = q$.

Let $\hat{\beta}$ be the unconstrained estimator, i.e. $\hat{\beta} = (X'X)^{-1}X'y$, and $b$ be the constrained estimator satisfying $Rb = r$. (Typically, $R\hat{\beta} \neq r$.)

**Proposition:** $b = \hat{\beta} + (X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}(r - R\hat{\beta})$

**Proof:** Let $S(\tilde{b}) = (y - X\tilde{b})'(y - X\tilde{b}) - 2\lambda(R\tilde{b} - r)$. The constrained estimator $b$ satisfies the first order conditions:

1. $-X'y + X'Xb - R'\lambda = 0$
2. $Rb - r = 0$

Thus $b = \hat{\beta} + (X'X)^{-1}R'\lambda$ (on premultiplying (1) by $(X'X)^{-1}$)

Let's eliminate $\lambda$.

$Rb = R\hat{\beta} + R(X'X)^{-1}R'\lambda$

Since $Rb = r$,

$[R(X'X)^{-1}R']^{-1}r = [R(X'X)^{-1}R']^{-1}R\hat{\beta} + \lambda$.

Thus, $\lambda = [R(X'X)^{-1}R']^{-1}(r - R\hat{\beta})$.

Substitute out $\lambda$ in the definition of $b$:

$b = \hat{\beta} + (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(r - R\hat{\beta})$
Now we can get the sampling distribution of $b$ by using $\hat{\beta} = \beta + (X'X)^{-1}X'\varepsilon$. For this, we first calculate the mean and the variance of $b$.

**Proposition:** $Eb = \beta.$ (Under $H_0$)

**Proof:** Substitute $\hat{\beta}$ in the definition of $b$:

$$b = \beta + (X'X)^{-1}X'\varepsilon + (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}[r - R\beta - R(X'X)^{-1}X'\varepsilon]$$

$$= \beta + [I - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R]X'X^{-1}X'\varepsilon,$$ using $r = R\beta$.

From this we see that $Eb = \beta$. ■

**Proposition:** $V(b) \leq V(\hat{\beta})$.

**Proof:** Let $A = R(X'X)^{-1}R'$. Note that $b - \beta = [I - (X'X)^{-1}R'A^{-1}R](X'X)^{-1}X'\varepsilon$.

$$V(b) = E(b - \beta)(b - \beta)'$$

$$= \sigma^2 [I - (X'X)^{-1}R'A^{-1}R][I - (X'X)^{-1}R'A^{-1}R]' \text{ since } E\varepsilon\varepsilon' = \sigma^2 I$$

$$= \sigma^2 [(X'X)^{-1} - 2(X'X)^{-1}R'A^{-1}R(X'X)^{-1} + (X'X)^{-1}R'A^{-1}R(X'X)^{-1}R'A^{-1}R(X'X)^{-1}]$$

Using the definition of $A$, this becomes

$$V(b) = \sigma^2 [(X'X)^{-1} - (X'X)^{-1}R'A^{-1}R(X'X)^{-1}]$$

$$\leq V(\hat{\beta}) = \sigma^2 (X'X)^{-1} \quad (why?) \quad ■$$

What is the relation to the Gauss-Markov theorem? Why doesn't this expression depend on $r$?

**Proposition:** Under normality, we have the complete distribution of $b$ with the mean and the variance calculated above.

**Estimation of $\sigma^2$**:
What is the unbiased estimator under restriction? What is the ML estimator?

**F TESTS:**

Let \( e \) and \( e^* \) be the vector of restricted and unrestricted residuals respectively.

**Proposition:**

\[
    e'e - e^*e^* = (r - R\hat{\beta})'[R(X'X)^{-1}R']^{-1}(r - R\hat{\beta})
\]

**Proof:**

\[
e = y - Xb = y - X\hat{\beta} - X(b - \hat{\beta}) = e^* - X(b - \hat{\beta})
\]

\[
    \Rightarrow e'e = e^*e^* + (b - \hat{\beta})'X'(b - \hat{\beta})
\]

\[
    \Rightarrow e'e - e^*e^* = (r - R\hat{\beta})'[R(X'X)^{-1}R']^{-1}(r - R\hat{\beta}) \quad \blacksquare
\]

This shows that the numerator of F can be calculated from the unrestricted estimates. However, LS is so easy to compute that we typically use restricted and unrestricted sums of squared residuals from separate regressions. Consider the following example.

**Example:** Consider \( y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \) with the restriction \( \beta_1 + \beta_2 = 2 \). If we substitute for \( \beta_1 \), we get

\[
y = \beta_0 + (2 - \beta_2)x_1 + \beta_2 x_2 + \varepsilon
\]

\[
y = \beta_0 + 2x_1 - \beta_2 x_1 + \beta_2 x_2 + \varepsilon
\]

\[
    \Rightarrow y - 2x_1 = \beta_0 + \beta_2(x_2 - x_1) + \varepsilon
\]

Regress \( (y - 2x_1) \) on a constant term and \( (x_2 - x_1) \), and get the sum of squared residuals from this restricted regression (\( e'e \)). Regress \( y \) on a constant term, \( x_1 \) and \( x_2 \), and get the sum of squared residuals from this unrestricted regression (\( e'e \)). Compare the sums of squared residuals from these regressions.
Structural Change:

This is covered in our combining samples example (i.e. example 2) and also subsets of coefficients example (i.e. example 3) in Lecture 5.

DUMMY VARIABLES:

Here we define a new variable equal to 0 or 1 indicating absence or presence of a characteristic. This allows the intercept to differ.

See figure 6.1

Example: homeowners/renters, male/female, regulation applies/regulation doesn't apply, etc.

Dummy variable trap:

Suppose \( X_2 = 1 \) if the characteristic is present

\[ = 0 \text{ if the characteristic is not present} \]

and \( X_3 = 1 \) if the characteristic is not present

\[ = 0 \text{ if the characteristic is present} \]

Then \( X_2 + X_3 = 1 = X_1 \) if the regression contains the constant term \( X_1 = 1 \in \mathbb{R}^N \). .... And?

Interactions between dummies for different characteristics:

Suppose \( X_2 \) is the dummy variable for characteristic 1 and \( X_3 \) is the dummy variable for characteristic 2. Let \( X_4 = X_2 \times X_3 \) (elementwise). That is,

\[ X_4 = 1 \text{ if both characteristics are present} \]

\[ = 0 \text{ if only one or none of the characteristics is present.} \]

Then the (marginal) effect of characteristic 1 (i.e. partial derivative of \( y \) with respect to characteristic 1) is \( \beta_1 \); effect of characteristic 2 is \( \beta_3 \); effect of both is \( \beta_2 + \beta_3 + \beta_4 \).
This could be set up differently. There are lots of alternatives. For example:

\[ X_2 = 1 \text{ if characteristic 1 is present but characteristic 2 is not} \]

\[ X_3 = 1 \text{ if characteristic 2 is present but characteristic 1 is not} \]

\[ X_4 = 1 \text{ if both characteristics are present} \]

Although different set ups will give different coefficients, correct interpretation of these coefficients will give the same estimated effects.

*Interactions with continuous regressors:*

See figures 6.2a and 6.2b

Suppose education is reported in grouped form: 0 - 8 years; 9 - 12 years; 12+ years

How should we set up the dummy variables?

One temptation is to code

\[ d = 0 \text{ if 0 - 8 years of education} \]

\[ = 1 \text{ if 9 - 12 years of education} \]

\[ = 2 \text{ if 12+ years of education} \]

This is very restrictive and probably unsound.

A better set up would be to use 2 dummies:

\[ d_1 = 1 \text{ if 0 - 8 years of education} \]

\[ = 0 \text{ else} \]

\[ d_2 = 1 \text{ if 9 - 12 years of education} \]

\[ = 0 \text{ else} \]

The first set up imposes that the effect of having 12+ years of education is twice the effect of having 9 - 12 years of education. In general, class variables with several classes require many
MULTICOLLINEARITY:

The problem is lack of data information when \( X'X \) is singular (recall picture) or "nearly" singular. If some X's move together, it is difficult to sort their separate effects on y. More data does help.

Other sources of information are useful. Purely "technical" remedies for collinearity work by imposing arbitrary and sometimes hidden "information". Never use ridge regression in an economic application.

The problem of multicollinearity in K-variate regression is equivalent to the problem of small sample size in estimating a mean. The problem of small sample size is referred to as "micronumerosity" by A.S.Goldberger. See the following amusing discussion.
NOTE ON THE GAUSS-MARKOV THEOREM:

Consider estimation of a mean $\mu$ based on an observation $X$.

Assume: $X \sim F$ and

$$\int x dF = \mu ; \int x^2 dF = \mu^2 + \sigma^2. \quad (*)$$

(Note that $\int x dF$ takes into account both continuous and discrete random variables. In the continuous random variable case $\int x F = \int x f(x) dx$ where $f(x)$ is the probability density function of $X$. In the discrete random variable case $\int x dF = \sum_{k=1}^{m} a_k \cdot \Pr(X = a_k)$ where $X$ takes on values $a_1, ..., a_m$.)

Suppose that the estimator for $\mu$ is $h(x)$. Unbiasedness implies that

$$\int h(x) dF = \mu.$$

*Theorem:* The only function $h$ unbiased for all $F$ and $\mu$ satisfying (*) is $h(x) = x$. 
**Proof:** Let \( h(x) = x + \phi(x) \). Then

\[
\int \phi(x)dF = 0 \quad \text{for all } F.
\]

Suppose, on the contrary, that \( \phi(x) > 0 \) on some set \( A \). Let \( \mu \in A \) and \( F = \delta_\mu \) (the distribution assigning point mass to \( x = \mu \)). Then (\*) is satisfied and

\[
\int h(x)d\delta_\mu = \mu + \phi(\mu) \neq \mu,
\]

which is a contradiction.

Argument is the same if \( \phi(x) < 0 \). This shows that \( \phi(x) = 0 \). ■

The logic is that if the estimator is nonlinear, we can choose a distribution so that it is biased.

**Application to regression:**

\[ y = X\beta + \varepsilon \Rightarrow (X'X)^{-1}X'y = \beta + (X'X)^{-1}X'\varepsilon. \]

This is just the problem of estimating a \( K \)-variate mean with one observation.

Details?