Lecture 18

Statistics---the linear model, multivariate style

Let's consider a more general linear model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i,$$
$$i = 1, \ldots, n$$

This is a job for matrix notation!

Statistics---the linear model, multivariate style

Let's consider a more general linear model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i,$$
$$i = 1, \ldots, n$$

This is a job for matrix notation!

Let $X_i = (X_{0i}, \ldots, X_{ki})$ $1\mathrm{x}(k+1)$ (row) vector (X_{0i}==1)

Let $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^T$ $(k+1)\mathrm{x}1$ (column) vector

Statistics---the linear model, multivariate style

So we have:

$$Y_i = X_i\beta + \varepsilon_i, \qquad i = 1, \ldots, n$$

But we can go further:

Let $Y = (Y_1, \ldots, Y_n)^T$ nx1 (column) vector

Let $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ nx1 (column) vector

Let $X = \begin{bmatrix} X_{01} & \ldots & X_{k1} \\ X_{02} & \ldots & X_{k2} \\ & & \\ X_{0n} & \ldots & X_{kn} \end{bmatrix}$ nx(k+1) matrix (X_{0i}==1)

Statistics---the linear model, multivariate style

So we have:

$$\underset{n\times 1}{Y} = \underset{(n\times(k+1))((k+1)\times 1)}{X\beta} + \underset{n\times 1}{\varepsilon}$$

Assumptions:

i) identification: $n > k+1$, X has full column rank $k+1$ (i.e., regressors are linearly independent, i.e., X^TX is invertible)

ii) error behavior: $E(\varepsilon) = 0$, $E(\varepsilon\varepsilon^T)$ ($= Cov(\varepsilon)$) $= \sigma^2 I_n$ (stronger version $\varepsilon \sim N(0, \sigma^2 I_n)$)

Statistics---the linear model, multivariate style

So we have:

$$\underset{n\times 1}{Y} = \underset{(n\times(k+1))((k+1)\times 1)}{X\beta} + \underset{n\times 1}{\varepsilon}$$

Assumptions:

i) identification: $n > k+1$, X has full column rank $k+1$ (i.e., regressors are linearly independent, i.e., X^TX is invertible)

ii) error behavior: $E(\varepsilon) = 0$, $E(\varepsilon\varepsilon^T)$ ($= Cov(\varepsilon)$) $= \sigma^2 I_n$ (stronger version $\varepsilon \sim N(0, \sigma^2 I_n)$)

nxn identity matrix

Statistics---the linear model, multivariate style

So we have:

$$\underset{n\times 1}{Y} = \underset{(n\times(k+1))((k+1)\times 1)}{X\beta} + \underset{n\times 1}{\varepsilon}$$

Assumptions:

i) identification: $n > k+1$, X has full column rank $k+1$ (i.e., regressors are linearly independent, i.e., X^TX is invertible)

ii) error behavior: $E(\varepsilon) = 0$, $E(\varepsilon\varepsilon^T)$ ($= Cov(\varepsilon)$) $= \sigma^2 I_n$ (stronger version $\varepsilon \sim N(0, \sigma^2 I_n)$)

Let's take a closer look at these assumptions.

Statistics---the linear model, multivariate style

i) $n > k+1$, X has full column rank $k+1$ (i.e., regressors are linearly independent, i.e., X^TX is invertible)---what does this mean?

---need to have more observations than regressors

---can't have any regressors that do not have positive sample variation

---can't have any regressors that are linear functions of one or more other regressors

Statistics---the linear model, multivariate style

---can't have any regressors that are linear functions of one or more other regressors

Imagine a case where we want to estimate the effect of schooling, work experience, and age, on salary, so we estimate the following model on individual-level data:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

Y_i salary

X_{1i} years of schooling

X_{2i} years of work experience

X_{3i} age

could be that $X_{1i} + X_{2i} + 6 = X_{3i}$ for all i in our sample

Statistics---the linear model, multivariate style

---can't have any regressors that are linear functions of one or more other regressors

Actually, researchers most often run afoul of this assumption when using dummy variables to indicate, say, observations falling into an exhaustive and mutually exclusive set of classes. Suppose each observation in your data set of pets is either a dog, cat, or fish. You cannot create and include a dummy variable for each type of pet because together they add up to a column of 1s, which is perfectly collinear with the first column of the X matrix. You need to omit one of them.

R will tell you if you make this mistake.

Statistics---the linear model, multivariate style

ii) $E(\varepsilon) = 0$, $E(\varepsilon\varepsilon^T) = \sigma^2 I_n$---what does this mean?

$\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is $n \times 1$, so $\varepsilon\varepsilon^T$ $n \times n$

$$E(\varepsilon\varepsilon^T) = \begin{bmatrix} E(\varepsilon_1\varepsilon_1) & \cdots & E(\varepsilon_1\varepsilon_n) \\ \\ E(\varepsilon_n\varepsilon_1) & \cdots & E(\varepsilon_n\varepsilon_n) \end{bmatrix} = \begin{bmatrix} Var(\varepsilon_1) & \cdots & Cov(\varepsilon_1, \varepsilon_n) \\ \\ Cov(\varepsilon_1, \varepsilon_n) & \cdots & Var(\varepsilon_n) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^2 & 0 & \ldots & 0 \\ 0 & \sigma^2 & \ldots & 0 \\ \\ 0 & 0 & \ldots & \sigma^2 \end{bmatrix} = \sigma^2 I_n$$

Statistics---the linear model, multivariate style

$E(\varepsilon) = 0$, $E(\varepsilon\varepsilon^T) = \sigma^2 I_n$---what does this mean?

This is because the $E(\varepsilon) = 0$.

$\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is $n \times 1$, so $\varepsilon\varepsilon^T$ $n \times n$

$$E(\varepsilon\varepsilon^T) = \begin{bmatrix} E(\varepsilon_1\varepsilon_1) & \ldots & E(\varepsilon_1\varepsilon_n) \\ \\ E(\varepsilon_n\varepsilon_1) & \ldots & E(\varepsilon_n\varepsilon_n) \end{bmatrix} = \begin{bmatrix} Var(\varepsilon_1) & \ldots & Cov(\varepsilon_1, \varepsilon_n) \\ \\ Cov(\varepsilon_1, \varepsilon_n) & \ldots & Var(\varepsilon_n) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^2 & 0 & \ldots & 0 \\ 0 & \sigma^2 & \ldots & 0 \\ \\ 0 & 0 & \ldots & \sigma^2 \end{bmatrix} = \sigma^2 I_n$$

Statistics---the linear model, multivariate style

$E(\varepsilon) = 0$, $E(\varepsilon\varepsilon^T) = \sigma^2 I_n$---what does this mean?

$\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is $n \times 1$, so $\varepsilon\varepsilon^T$ $n \times n$

$$E(\varepsilon\varepsilon^T) = \begin{bmatrix} E(\varepsilon_1\varepsilon_1) & \ldots & E(\varepsilon_1\varepsilon_n) \\ \\ E(\varepsilon_n\varepsilon_1) & \ldots & E(\varepsilon_n\varepsilon_n) \end{bmatrix} = \begin{bmatrix} Var(\varepsilon_1) & \ldots & Cov(\varepsilon_1, \varepsilon_n) \\ \\ Cov(\varepsilon_1, \varepsilon_n) & \ldots & Var(\varepsilon_n) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^2 & 0 & \ldots & 0 \\ 0 & \sigma^2 & \ldots & 0 \\ \\ 0 & 0 & \ldots & \sigma^2 \end{bmatrix} = \sigma^2 I_n$$

This is called the variance-covariance matrix of ε---we denote it $Cov(\varepsilon)$

Statistics---the linear model, multivariate style

Linear model:

$$\underset{n\times 1}{Y} = \underset{(n\times(k+1))((k+1)\times 1)}{X\beta} + \underset{n\times 1}{\varepsilon}$$

What is $\hat{\beta}$? Well, it is the vector that minimizes the sum of squared errors, i.e., $\hat{\varepsilon}^T\hat{\varepsilon} = (Y - X\hat{\beta})^T(Y - X\hat{\beta})$

Statistics---the linear model, multivariate style

Linear model:

$$\underset{n\times 1}{Y} = \underset{(n\times(k+1))((k+1)\times 1)}{X\beta} + \underset{n\times 1}{\varepsilon}$$

What is $\hat{\beta}$? Well, it is the vector that minimizes the sum of squared errors, i.e., $\hat{\varepsilon}^T\hat{\varepsilon} = (Y - X\hat{\beta})^T(Y - X\hat{\beta})$

So, take the derivative w.r.t. β and set equal to zero to obtain $-2X^T(Y - X\hat{\beta}) = 0$. Then solve for $\hat{\beta}$.

Statistics---the linear model, multivariate style

Linear model:

$$\underset{n\times 1}{Y} = \underset{(n\times(k+1))((k+1)\times 1}{X\beta} + \underset{n\times 1}{\varepsilon}$$

What is $\hat{\beta}$? Well, it is the vector that minimizes the sum of squared errors, i.e., $\hat{\varepsilon}^T\hat{\varepsilon} = (Y - X\hat{\beta})^T(Y - X\hat{\beta})$

So, take the derivative w.r.t. β and set equal to zero to obtain $-2X^T(Y - X\hat{\beta}) = 0$. Then solve for $\hat{\beta}$.

$X^TY = X^TX\hat{\beta}$

$\hat{\beta} = (X^TX)^{-1}X^TY$ if (X^TX) is invertible.

Statistics---the linear model, multivariate style

Linear model:

$$\underset{n\times 1}{Y} = \underset{(n\times(k+1))((k+1)\times 1)}{X\beta} + \underset{n\times 1}{\varepsilon}$$

What is $\hat{\beta}$? Well, it is the vector that minimizes the sum of squared errors, i.e., $\hat{\varepsilon}^T\hat{\varepsilon} = (Y-X\hat{\beta})^T(Y-X\hat{\beta})$

So, take the derivative w.r.t. β and set equal to zero to obtain $-2X^T(Y-X\hat{\beta}) = 0$. Then solve for $\hat{\beta}$.

$$X^TY = X^TX\hat{\beta}$$

$\hat{\beta} = (X^TX)^{-1}X^TY$ if (X^TX) is invertible.

Wow, beautiful.

Statistics---the linear model, multivariate style

$\hat{\beta}$

Linear model:

$$Y = X\beta + \varepsilon$$

$$n\times 1 \quad (n\times(k+1))((k+1)\times 1) \quad n\times 1$$

What do we want to know about $\hat{\beta}$? Its distribution!

Statistics---the linear model, multivariate style

$\hat{\beta}$

Linear model:

$$\underset{n\times 1}{Y} = \underset{(n\times(k+1))((k+1)\times 1)}{X\beta} + \underset{n\times 1}{\varepsilon}$$

What do we want to know about $\hat{\beta}$? Its distribution!

$E(\hat{\beta}) = \beta$ (Treat Xs as fixed, they come outside of the expectation operator, and it's easy to show.)

Statistics---the linear model, multivariate style

Linear model:

$$\underset{n\times 1}{Y} = \underset{(n\times(k+1))((k+1)\times 1)}{X\beta} + \underset{n\times 1}{\varepsilon}$$

What do we want to know about $\hat{\beta}$? Its distribution!

$E(\hat{\beta}) = \beta$ (Treat Xs as fixed, they come outside of the expectation operator, and it's easy to show.)

$Cov(\hat{\beta}) = \sigma^2(X^TX)^{-1}$ (Again, not too hard to show if you treat the Xs as fixed---details in notes on website.)

Statistics---the linear model, multivariate style

Linear model:

$$\underset{n\times 1}{Y} = \underset{(n\times(k+1))((k+1)\times 1)}{X\beta} + \underset{n\times 1}{\varepsilon}$$

What do we want to know about $\hat{\beta}$? Its distribution!

$E(\hat{\beta}) = \beta$ (Treat Xs as fixed, they come outside of the expectation operator, and it's easy to show.)

$Cov(\hat{\beta}) = \sigma^2(X^TX)^{-1}$ (Again, not too hard to show if you treat the Xs as fixed---details in notes on website.)

And $\hat{\sigma}^2 = \hat{\varepsilon}^T\hat{\varepsilon}/(n-k)$

Statistics---the linear model, multivariate style

Linear model:

$$\underset{n\times 1}{Y} = \underset{(n\times(k+1))((k+1)\times 1}{X\beta} + \underset{n\times 1}{\varepsilon}$$

What do we want to know about $\hat{\beta}$? Its distribution!

$E(\hat{\beta}) = \beta$ (Treat Xs as fixed, they come outside of the expectation operator, and it's easy to show.)

$Cov(\hat{\beta}) = \sigma^2(X^TX)^{-1}$ (Again, not too hard to show if you treat the Xs as fixed---details in notes on website.)

And $\hat{\sigma}^2 = \hat{\varepsilon}^T\hat{\varepsilon}/(n-k)$

And if the errors are normally-distributed, they're also normal.

Statistics---inference in the linear model

Now, finally, we get to inference. Typically, we will want to test hypotheses involving the βs. (The βs are the parameters in our conditional mean function of our outcome variable Y, and the questions we want to answer are usually about the nature of this conditional mean function.)

Sometimes we are only interested in one of the βs. Other times we might want to simultaneously test hypotheses about several of them.

As we saw in the output I showed you earlier, statistical packages typically perform some standard tests for us, but there may be other ones we need to do ourselves.

Statistics---inference in the linear model

Let's start with a pretty general framework for testing hypotheses about β. It's not only quite general and flexible, it's also super intuitive.

Statistics---inference in the linear model

Let's consider hypotheses of the following form:

H_0: $R\beta = c$

H_A: $R\beta \neq c$

R is a $r \times (k+1)$ matrix of restrictions. (If $r = 1$, then we are just testing one restriction, such as $\beta_1 = 0$.)

Statistics---inference in the linear model

Let's consider hypotheses of the following form:

H_0: $R\beta = c$

H_A: $R\beta \neq c$

Almost any hypothesis involving β you can dream up in the context of the linear model can be captured in this framework. You can test whether individual parameters are equal to zero. You can test whether individual parameters are equal to something other than zero. You can test multiple hypotheses simultaneously. You can test hypotheses about linear combinations of parameters. The world is your oyster.

Statistics---inference in the linear model

Let's consider hypotheses of the following form:

$$H_0: R\beta = c$$

$$H_A: R\beta \neq c$$

R is a $r \times (k+1)$ matrix of restrictions.

If, for instance, $R = \{0\ 1\ 0\ .\ .\ 0\}$ and $c = \{0\}$

that corresponds to $H_0: \beta_1 = 0$.

Statistics---inference in the linear model

Let's consider hypotheses of the following form:

H_0: $R\beta = c$

H_A: $R\beta \neq c$

R is a rx(k+1) matrix of restrictions.

If, for instance, $R = \begin{bmatrix} 0 & 1 & 0 & . & . & 0 \\ 0 & 0 & 1 & . & . & 0 \\ . & . & & & & . \\ 0 & 0 & 0 & . & . & 1 \end{bmatrix}$ and $c = \begin{bmatrix} 0 \\ 0 \\ . \\ 0 \end{bmatrix}$

that corresponds to H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$.

Statistics---inference in the linear model

Let's consider hypotheses of the following form:

H_0: $R\beta = c$

H_A: $R\beta \neq c$

R is a $r\times(k+1)$ matrix of restrictions.

If, for instance, $R = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$ and $c = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$

that corresponds to H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$.

Here we're testing k hypotheses simultaneously.

Statistics---inference in the linear model

Let's consider hypotheses of the following form:

$$H_0: R\beta = c$$

$$H_A: R\beta \neq c$$

R is a $r \times (k+1)$ matrix of restrictions.

If, for instance, $R = \begin{bmatrix} 0 & 1 & -1 & . & . & 0 \\ 0 & 0 & 0 & 1 & . & 0 \\ 0 & 0 & 0 & . & . & 1 \end{bmatrix}$ and $c = \begin{bmatrix} 0 \\ 5 \\ -2 \end{bmatrix}$

that corresponds to H_0: $\beta_1 = \beta_2$, $\beta_3 = 5$, and $\beta_k = -2$.

Statistics---inference in the linear model

One thing you cannot do in this framework is test one-sided hypotheses. (We'll get back to those.)

Statistics---inference in the linear model

We have a super intuitive and cool way to test these hypotheses. (First, think of the null as describing a set of restrictions on the model.)

1. We estimate the unrestricted model.
2. We impose the restrictions of the null and estimate that model.
3. We compare the goodness-of-fit of the models. If the restrictions don't really affect the fit of the model much, then the null is probably true or close to true, so we do not want to reject it. If the restrictions really bind, then we do want to reject the null.

Statistics---inference in the linear model

Estimating the unrestricted model should be simple---just run the regression. But how do we estimate the restricted model?

If the restriction is that certain βs = 0, then leave the regressors corresponding to those βs out of the restricted model.

If the restriction is that, say, two βs are equal, create a new regressor, which is the sum of the regressors corresponding to those βs and include that sum in the restricted model in place of the original regressors.

What if the restriction is that some $\beta = c$?

Statistics---inference in the linear model

This is an F-test. (We've mentioned a special case of the F test before. This is a more general formulation.)

$T = ((SSR_R - SSR_U)/r)/(SSR_U/(n-(k+1))$

$T \sim F_{r,n-(k+1)}$ under the null and we reject the null for large values of the test statistic.

Statistics---inference in the linear model

This is an F-test. (We've mentioned a special case of the F test before. This is a more general formulation.)

$T = ((SSR_R - SSR_U)/r)/(SSR_U/(n-(k+1))$

$T \sim F_{r,n-(k+1)}$ under the null and we reject the null for large values of the test statistic.

(Why an F distribution? Well, the reason goes back to one of the facts I told you about special distributions a couple of weeks ago. The ratio of two independent χ^2 random variables divided by their respective degrees of freedom are distributed F.)