Sociology 761

John Fox

Lecture Notes

Linear Models Using Matrices

1. Introduction

- The principal purpose of this lecture is to demonstrate how matrices can be used to simplify the development of statistical models.
- A secondary purpose is to review, and extend, some material in linear models.
- I will take up the following topics:
  - Expressing linear models for regression, dummy regression, and analysis of variance in matrix form.
  - Deriving the least-squares coefficients using matrices.
  - Distribution of the least-squares coefficients.
  - The least-squares coefficients as maximum-likelihood estimators.
  - Statistical inference for linear models.

2. Linear Models in Matrix Form

▶ The general linear model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

where

- y_i is the value of the *response variable* for the ith of n observations.
- $x_{i1}, x_{i2}, ..., x_{ik}$ are the values of k *regressors* for observation i. In linear regression analysis, $x_{i1}, x_{i2}, \ldots, x_{ik}$ are the values of k quantitative *explanatory variables*.
- $\beta_0, \beta_1, \ldots, \beta_k$ are $k+1$ *parameters* to be estimated from the data, including the *constant* or *intercept* term, β_0.
- ε_i is the random *error* variable for the ith observation.

▶ The statistical assumptions of the linear model concern the behaviour of the errors; the standard assumptions include:

- *Linearity*: The average error is zero, $E(\varepsilon_i) = 0$; equivalently, $E(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$.
- *Constant error variance*: The variance of the errors is the same for all observations, $V(\varepsilon_i) = \sigma_\varepsilon^2$; equivalently, $V(y_i) = \sigma_\varepsilon^2$.
- *Normality*: The errors are normally distributed, and so $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$; equivalently, $y_i \sim N(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}, \sigma_\varepsilon^2)$.
- *Independence*: The errors are independently sampled — that is ε_i and ε_j are independent for $i \neq j$; equivalently, y_i and y_j are independent.
- Either the x-values are *fixed* (with respect to repeated sampling) or, if random, the xs are *independent of the errors*.

▶ The linear model may be rewritten as

$$
\begin{aligned}
y_i &= [1, x_{i1}, x_{i2}, ..., x_{ik}] \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \cdot \\ \beta_k \end{bmatrix} + \varepsilon_i \\
&= \underset{(1\times k+1)}{\mathbf{x}_i'} \underset{(k+1\times 1)}{\boldsymbol{\beta}} + \varepsilon_i
\end{aligned}
$$

- There is one such equation for each observation, $i = 1, \ldots, n$.

- Collecting these n equations into a single matrix equation:

$$
\begin{aligned}
\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} &= \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix} \\
\underset{(n\times 1)}{\mathbf{y}} &= \underset{(n\times k+1)}{\mathbf{X}} \underset{(k+1\times 1)}{\boldsymbol{\beta}} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}}
\end{aligned}
$$

  – The $\mathbf{X}$ matrix in the linear model is called the *model matrix* (or the *design matrix*).
  – Note the column of 1s for the constant.

- Similarly, the assumptions of linearity, constant variance, normality, and independence can be recast as

$$\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_n)$$

where $N_n(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_n)$ denotes the multivariate-normal distribution with
  - mean vector $\mathbf{0}$,
  - and covariance matrix

$$\sigma_\varepsilon^2 \mathbf{I}_n = \begin{bmatrix} \sigma_\varepsilon^2 & 0 & \cdots & 0 \\ 0 & \sigma_\varepsilon^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_\varepsilon^2 \end{bmatrix}$$

  - equivalently,

$$\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma_\varepsilon^2 \mathbf{I}_n)$$

2.1 Dummy Regression Models

- The matrix equation $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ suffices not just for linear regression models, but — with suitable specification of the regressors — for linear models generally.
- For example, consider the *dummy-regression model*

$$y_i = \alpha + \beta x_i + \gamma d_i + \delta(x_i d_i) + \varepsilon_i$$

where
  - y is income in dollars,
  - x is years of education,
  - and the dummy regressor d is coded 1 for men and 0 for women.

▶ Recall that this model implies potentially different intercepts and slopes — that is, potentially different regression lines — for the two groups:

- for men,

$$\begin{aligned} y_i &= \alpha + \beta x_i + \gamma 1 + \delta(x_i 1) + \varepsilon_i \\ &= (\alpha + \gamma) + (\beta + \delta)x_i + \varepsilon_i \end{aligned}$$

- for women

$$\begin{aligned} y_i &= \alpha + \beta x_i + \gamma 0 + \delta(x_i 0) + \varepsilon_i \\ &= \alpha + \beta x_i + \varepsilon_i \end{aligned}$$

- and so γ is the difference in intercepts between men and women, and δ is the difference in slopes.
- Because men and women can have different slopes, this model permits gender to *interact* with education in determining income.

▶ Written as a matrix equation, the dummy-regression model becomes.

$$\begin{bmatrix} y_1 \\ \vdots \\ y_{n_1} \\ \hline y_{n_1+1} \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n_1} & 0 & 0 \\ \hline 1 & x_{n_1+1} & 1 & x_{n_1+1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & 1 & x_n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{n_1} \\ \hline \varepsilon_{n_1+1} \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where, for clarity, the n_1 observations for women precede the $n - n_1$ observations for men.

- *Reminder*: When a categorical explanatory variable has more than two (say, m) categories, it generates a set of $m-1$ dummy regressors — that is, one fewer dummy variable than the number of categories.
  - For example, a five-category regional classification might produce the following four dummy regressors:

Region	d_1	d_2	d_3	d_4
East	1	0	0	0
Quebec	0	1	0	0
Ontario	0	0	1	0
Prairies	0	0	0	1
BC	0	0	0	0

  - Here, BC is arbitrarily selected as the *baseline category*, to which other categories will implicitly be compared.

2.2 Analysis of Variance Models

- *Analysis of variance* or *ANOVA* models are linear models in which all of the explanatory variables are *factors* — that is, categorical variables.
- The simplest case is *one-way ANOVA*, where there is a single factor.
  - The one-way ANOVA model is usually written with double-subscript notation as
$$y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$
for *levels* $j = 1, ..., m$ of the factor, and observations $i = 1, \ldots, n_j$ of level j.

▶ The matrix form of the one-way ANOVA model is

$$
\begin{array}{l}
\text{group } 1 \\ \\ \\ \text{group } 2 \\ \\ \\ \\ \text{group } m-1 \\ \\ \text{group } m \\ \\ \\
\end{array}
\begin{bmatrix} y_{11} \\ \vdots \\ y_{n_1,1} \\ \hline y_{12} \\ \vdots \\ y_{n_2,2} \\ \hline \vdots \\ \hline y_{1,m-1} \\ \vdots \\ y_{n_{m-1},m-1} \\ \hline y_{1m} \\ \vdots \\ y_{n_m,m} \end{bmatrix}
=
\begin{bmatrix}
1 & 1 & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots \\
1 & 1 & 0 & \cdots & 0 & 0 \\ \hline
1 & 0 & 1 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots \\
1 & 0 & 1 & \cdots & 0 & 0 \\ \hline
\vdots & \vdots & \vdots & & \vdots & \vdots \\ \hline
1 & 0 & 0 & \cdots & 1 & 0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots \\
1 & 0 & 0 & \cdots & 1 & 0 \\ \hline
1 & 0 & 0 & \cdots & 0 & 1 \\
\vdots & \vdots & \vdots & & \vdots & \vdots \\
1 & 0 & 0 & \cdots & 0 & 1
\end{bmatrix}
\begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{m-1} \\ \alpha_m \end{bmatrix}
+
\begin{bmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{n_1,1} \\ \hline \varepsilon_{12} \\ \vdots \\ \varepsilon_{n_2,2} \\ \hline \vdots \\ \hline \varepsilon_{1,m-1} \\ \vdots \\ \varepsilon_{n_{m-1},m-1} \\ \hline \varepsilon_{1m} \\ \vdots \\ \varepsilon_{n_m,m} \end{bmatrix}
$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

▶ This formulation of the model is problematic because there is a redundant column in the model matrix (which is therefore of *deficient rank* m):

- For example, the first column is the sum of the remaining columns.
- This will create a problem when we try to fit the model by least squares, but more fundamentally, it reflects a redundancy among the parameters of the model.

▶ A common solution to the problem is to reduce the parameters by one. There are many ways to do this, all providing equivalent fits to the data. For example:

- Eliminating the constant, μ, produces a so-called means model,
$$y_{ij} = \alpha_j + \varepsilon_{ij}$$
where α_j now represents the population mean of level j.
- Eliminating one of the α_j produces a dummy-variable solution, with the omitted coefficient corresponding to the baseline category (here category m):

$$\begin{bmatrix} y_{11} \\ \vdots \\ y_{n_1,1} \\ \hline y_{12} \\ \vdots \\ y_{n_2,2} \\ \hline \vdots \\ \hline y_{1,m-1} \\ \vdots \\ y_{n_{m-1},m-1} \\ \hline y_{1m} \\ \vdots \\ y_{n_m,m} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & \cdots & 0 \\ \hline 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 1 & \cdots & 0 \\ \hline \vdots & \vdots & \vdots & & \vdots \\ \hline 1 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 1 \\ \hline 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{m-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{n_1,1} \\ \hline \varepsilon_{12} \\ \vdots \\ \varepsilon_{n_2,2} \\ \hline \vdots \\ \hline \varepsilon_{1,m-1} \\ \vdots \\ \varepsilon_{n_{m-1},m-1} \\ \hline \varepsilon_{1m} \\ \vdots \\ \varepsilon_{n_m,m} \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

▶ Alternatively, we can place a linear constraint on the parameters, most commonly, the *sigma constraint*

$$\sum_{j=1}^{m} \alpha_j = 0$$

- Under this constraint

$$\alpha_m = -\sum_{j=1}^{m-1} \alpha_j$$

need not appear explicitly, producing the model matrix

$$\underset{(n\times m)}{\mathbf{X}} = \begin{array}{ll}
 & \begin{array}{ccccc} (\mu) & (\alpha_1) & (\alpha_2) & \cdots & (\alpha_{m-1}) \end{array} \\
\begin{array}{l} \\ \text{group } 1 \\ \\ \\ \text{group } 2 \\ \\ \\ \\ \text{group } m-1 \\ \\ \\ \text{group } m \\ \\ \end{array} &
\left[\begin{array}{ccccc}
1 & 1 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & & \vdots \\
1 & 1 & 0 & \cdots & 0 \\ \hline
1 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & & \vdots \\
1 & 0 & 1 & \cdots & 0 \\ \hline
\vdots & \vdots & \vdots & & \vdots \\ \hline
1 & 0 & 0 & \cdots & 1 \\
\vdots & \vdots & \vdots & & \vdots \\
1 & 0 & 0 & \cdots & 1 \\ \hline
1 & -1 & -1 & \cdots & -1 \\
\vdots & \vdots & \vdots & & \vdots \\
1 & -1 & -1 & \cdots & -1
\end{array}\right]
\end{array}$$

3. Least-Squares Fit

▶ The fitted linear model is

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

where

- $\mathbf{b} = [b_0, b_1, ..., b_k]'$ is the vector of fitted coefficients.
- $\mathbf{e} = [e_1, e_2, ..., e_n]' = \mathbf{y} - \mathbf{X}\mathbf{b}$ is the vector of residuals.

▶ We want the coefficient vector $\mathbf{b}$ that minimizes the residual sum of squares, expressed as a function of $\mathbf{b}$:

$$\begin{aligned} S(\mathbf{b}) &= \sum e_i^2 = \mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) \\ &= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\mathbf{b} - \mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} \\ &= \mathbf{y}'\mathbf{y} - (2\mathbf{y}'\mathbf{X})\mathbf{b} + \mathbf{b}'(\mathbf{X}'\mathbf{X})\mathbf{b} \end{aligned}$$

- The last line of the equation is justified because $\underset{(1\times n)}{\mathbf{y}'}\underset{(n\times k+1)}{\mathbf{X}}\underset{(k+1\times 1)}{\mathbf{b}}$ and $\underset{(1\times k+1)}{\mathbf{b}'}\underset{(k+1\times n)}{\mathbf{X}'}\underset{(n\times 1)}{\mathbf{y}}$ are both scalars, and consequently equal.

► Noting that $\mathbf{y}'\mathbf{y}$ is a constant (with respect to $\mathbf{b}$), $(2\mathbf{y}'\mathbf{X})\mathbf{b}$ is a linear function of $\mathbf{b}$, and $\mathbf{b}'(\mathbf{X}'\mathbf{X})\mathbf{b}$ is a quadratic form in $\mathbf{b}$,

$$\frac{\partial S(\mathbf{b})}{\partial \mathbf{b}} = \mathbf{0} - 2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b}$$

- Setting the derivative to $\mathbf{0}$ produces the *normal equations* for the linear model

$$\begin{aligned} -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b} &= \mathbf{0} \\ \mathbf{X}'\mathbf{X}\mathbf{b} &= \mathbf{X}'\mathbf{y} \end{aligned}$$

  a system of $k+1$ linear equations in $k+1$ unknowns (i.e., $b_0, b_1, ..., b_k$).

- We can solve the normal equations uniquely for $\mathbf{b}$ if as the $(k+1)\times(k+1)$ matrix $\mathbf{X}'\mathbf{X}$ is nonsingular, which will be the case as long as
  – there are at least as many observations as coefficients — that is, $n \geq k+1$.
  – no column of the model matrix $\mathbf{X}$ is a perfect linear function of the other columns.

- When $\mathbf{X}'\mathbf{X}$ is nonsingular, the least-squares solution is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

- Looking inside of the matrices in the normal equations,
  – the matrix $\mathbf{X}'\mathbf{X}$ contains sums of squares and cross-products for the regressors (including the column of 1s).
  – $\mathbf{X}'\mathbf{y}$ contains sums of products between the regressors and the response.
- The normal equations, therefore, are

$$\begin{array}{llll} b_0 n & +b_1 \sum x_{i1} & + \cdots + b_k \sum x_{ik} & = \sum y_i \\ b_0 \sum x_{i1} & +b_1 \sum x_{i1}^2 & + \cdots + b_k \sum x_{i1}x_{ik} & = \sum x_{i1}y_i \\ \vdots & & & \vdots \\ b_0 \sum x_{ik} & +b_1 \sum x_{ik}x_{i1} & + \cdots + b_k \sum x_{ik}^2 & = \sum x_{ik}y_i \end{array}$$

► An example, using Duncan's regression of occupational prestige on the income and education levels of 45 U.S. occupations:

- Matrices of sums of squares and products:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 45 & 1884 & 2365 \\ 1884 & 105,148 & 122,197 \\ 2365 & 122,197 & 163,265 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 2146 \\ 118,229 \\ 147,936 \end{bmatrix}$$

- The inverse of $\mathbf{X}'\mathbf{X}$:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.1021058996 & -0.0008495732 & -0.0008432006 \\ -0.0008495732 & 0.0000801220 & -0.0000476613 \\ -0.0008432006 & -0.0000476613 & 0.0000540118 \end{bmatrix}$$

- The regression coefficients:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} -6.06466 \\ 0.59873 \\ 0.54583 \end{bmatrix}$$

4. Distribution of the Least-Squares Coefficients

- It is simple to show that least-squares coefficients are *unbiased estimators* of the population regression coefficients:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

and so (assuming a fixed model matrix $\mathbf{X}$),

$$E(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\beta}$$

- The covariance matrix of $\mathbf{b}$ follows from the covariance matrix of $\mathbf{y}$, which is $\sigma_\varepsilon^2\mathbf{I}_n$:

$$\begin{aligned} V(\mathbf{b}) &= \left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right] V(\mathbf{y}) \left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]' \\ &= \left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right] \sigma_\varepsilon^2\mathbf{I}_n \left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]' \\ &= \sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

- Because the error variance σ_ε^2 is an unknown parameter, the covariance matrix of $\mathbf{b}$ must be estimated:

$$\widehat{V}(\mathbf{b}) = s_e^2(\mathbf{X}'\mathbf{X})^{-1}$$

where

$$s_e^2 = \frac{\sum e_i^2}{n-k-1}$$

is the estimated error variance, and e_i is the residual for observation i.

▶ Because the response vector $\mathbf{y}$ is multinormally distributed, so is $\mathbf{b}$; that is

$$\mathbf{b} \sim N_{k+1}\left[\boldsymbol{\beta},\, \sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1}\right]$$

▶ Notice the strong analogy between the formulas for the slope coefficient in least-squares simple regression (i.e., with a single x) and for the coefficients of the linear model in matrix form:

	Simple Regression	*Linear Model*
Model	$y_i = \alpha + \beta x_i + \varepsilon_i$ $y^* = x^*\beta + \varepsilon$	$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
Least-Squares Estimator	$b = \frac{\sum x^* y^*}{\sum x^{*2}}$ $= \left(\sum x^{*2}\right)^{-1} \sum x^* y^*$	$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$
Sampling Variance	$V(b) = \frac{\sigma_\varepsilon^2}{\sum x^{*2}}$ $= \sigma_\varepsilon^2 \left(\sum x^{*2}\right)^{-1}$	$V(\mathbf{b}) = \sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1}$
Distribution	$b \sim$ $N\left[\beta, \sigma_\varepsilon^2 \left(\sum x^{*2}\right)^{-1}\right]$	$\mathbf{b} \sim$ $N_{k+1}\left[\boldsymbol{\beta}, \sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1}\right]$

- In the scalar formulas the following short-hand notation is used:

$$\begin{aligned} x^* &= x_i - \overline{x} \\ y^* &= y_i - \overline{y} \end{aligned}$$

5. Maximum-Likelihood Estimation of the Normal Linear Model

▶ The standard assumptions of the linear model provide a probability model for the data $\mathbf{y}$ (thinking of the model matrix $\mathbf{X}$ as fixed or conditioning on it):

$$\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta},\, \sigma_\varepsilon^2 \mathbf{I}_n)$$

- Then, from the formula for the normal distribution,

$$p(\mathbf{y}) = \frac{1}{(2\pi\sigma_\varepsilon^2)^{n/2}} \exp\left[-\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma_\varepsilon^2}\right]$$

  – *Note*: $\exp(a)$ in a formula means e^a, for the constant $e \simeq 2.718$.

▶ In *maximum-likelihood estimation*, recall, we find the values of the parameters that make the probability of observing the data as high as possible.

- The likelihood function is the same as the probability (or probability-density) of the data, except thought of as a function of the parameters.
- Here,
$$L(\boldsymbol{\beta}, \sigma_\varepsilon^2) = \left(2\pi\sigma_\varepsilon^2\right)^{-n/2} \exp\left[-\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma_\varepsilon^2}\right]$$

▶ As is usually the case, it is simpler to work with the log of the likelihood.
- Whatever values of the parameters maximize the log-likelihood also maximize the likelihood, since the log function is monotone (strictly increasing).
- For the linear model:
$$\log_e L(\boldsymbol{\beta}, \sigma_\varepsilon^2) = -\frac{n}{2}\log_e 2\pi - \frac{n}{2}\log_e \sigma_\varepsilon^2 - \frac{1}{2\sigma_\varepsilon^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$$
- To justify this result, recall that taking logs turns multiplication into addition, division into subtraction, and exponentiation into multiplication; moreover, $\log_e e^a = a$.

▶ To maximize the log-likelihood, we need its derivatives with respect to the parameters.
- Finding the derivatives is simplified by noticing that $(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$ is just the sum of squared errors.
- Differentiating,
$$\begin{aligned}
\frac{\partial \log_e L(\boldsymbol{\beta}, \sigma_\varepsilon^2)}{\partial \boldsymbol{\beta}} &= -\frac{1}{2\sigma_\varepsilon^2}(2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - 2\mathbf{X}'\mathbf{y}) \\
\frac{\partial \log_e L(\boldsymbol{\beta}, \sigma_\varepsilon^2)}{\partial \sigma_\varepsilon^2} &= -\frac{n}{2}\left(\frac{1}{\sigma_\varepsilon^2}\right) + \frac{1}{\sigma_\varepsilon^4}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})
\end{aligned}$$
- Setting the partial derivatives to $\mathbf{0}$ and solving for maximum-likelihood estimates of the parameters produces
$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
\widehat{\sigma}_\varepsilon^2 &= \frac{(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})}{n} = \frac{\mathbf{e}'\mathbf{e}}{n}
\end{aligned}$$
where $\mathbf{e} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$ is the vector of residuals.

- Notice that
  - The MLE $\widehat{\boldsymbol{\beta}}$ is just the least-squares coefficients $\mathbf{b}$.
  - The MLE of the error variance, $\widehat{\sigma}_\varepsilon^2 = \sum e_i^2/n$ is *biased*.
    - The usual unbiased estimator, s_e^2, divides by residual *degrees of freedom* $n-k-1$ rather than by n.
    - The MLE is *consistent*, however, since the bias (along with the variance of the estimator) goes to zero as n get larger.

6. Statistical Inference for Least-Squares Estimation

- Statistical inference for $\boldsymbol{\beta}$ based on the least-squares coefficients $\mathbf{b}$ uses the estimated covariance matrix $\widehat{V}(\mathbf{b}) = s_e^2(\mathbf{X}'\mathbf{X})^{-1}$.
- The simplest case is inference for an individual coefficient, b_j:
  - The standard error of the coefficient is the square root of the jth diagonal entry of the estimated covariance matrix (indexing the matrix from 0):

$$\text{SE}(b_j) = \sqrt{s_e^2[(\mathbf{X}'\mathbf{X})^{-1}]_{jj}}$$

  - Because the error variance has been estimated, hypothesis tests and confidence intervals use the t-distribution with $n-k-1$ degress of freedom.

- For example:
  - To test
$$H_0\colon \beta_j = 0$$
    we compute
$$t_0 = \frac{b_j}{\text{SE}(b_j)}$$
  - To form a 95-percent confidence interval for β_j we take
$$\beta_j = b_j \pm t_{.975,n-k-1}\text{SE}(b_j)$$
    where $t_{.975,n-k-1}$ is the .975 quantile of the t-distribution with $n-k-1$ degrees of freedom.

▶ More generally, suppose that we want to test the linear hypothesis
$$H_0\colon \underset{(q\times k+1)}{\mathbf{L}}\ \underset{(k+1\times 1)}{\boldsymbol{\beta}} = \underset{(q\times 1)}{\mathbf{c}}$$
where the hypothesis matrix $\mathbf{L}$ and the right-hand-side vector $\mathbf{c}$ (usually $\mathbf{0}$) encode the hypothesis.

- For example, in Duncan's regression of prestige on income and education, the hypothesis matrix
$$\mathbf{L} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$
  and right-hand-side vector
$$\mathbf{c} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$
  specify the hypothesis
$$H_0\colon \beta_1 = 0,\ \beta_2 = 0$$

- Likewise, again for Duncan's regression, the one-row hypothesis matrix
$$\mathbf{L} = \begin{bmatrix} 0 & 1 & -1 \end{bmatrix}$$
and right-hand-side $\mathbf{c} = [0]$ correspond to the hypothesis
$$H_0\colon \beta_1 - \beta_2 = 0$$
that is
$$H_0\colon \beta_1 = \beta_2$$
- Under the hypothesis H_0, the statistic
$$F_0 = \frac{(\mathbf{Lb} - \mathbf{c})' \left[\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}'\right]^{-1} (\mathbf{Lb} - \mathbf{c})}{q s_e^2}$$
follows an F-distribution with q and $n - k - 1$ degrees of freedom.

▶ *Example*: For Duncan's regression, the sum of squared residuals is $\mathbf{e}'\mathbf{e} = 7506.699$, and so
$$s_e^2 = \frac{7506.699}{45 - 2 - 1} = 178.7309$$

- The estimated covariance matrix of the least-squares coefficients is
$$\begin{aligned} \widehat{V}(\mathbf{b}) &= s_e^2(\mathbf{X}'\mathbf{X})^{-1} \\ &= 178.7309 \begin{bmatrix} 0.1021058996 & -0.0008495732 & -0.0008432006 \\ -0.0008495732 & 0.0000801220 & -0.0000476613 \\ -0.0008432006 & -0.0000476613 & 0.0000540118 \end{bmatrix} \\ &= \begin{bmatrix} 18.249387 & -0.151844 & -0.150705 \\ -0.151844 & 0.014320 & -0.008519 \\ -0.150705 & -0.008519 & 0.009653 \end{bmatrix} \end{aligned}$$

- The estimated standard errors of the regression coefficients are, therefore,

$$\begin{aligned}\text{SE}(b_0) &= \sqrt{18.249387} = 4.272\\ \text{SE}(b_1) &= \sqrt{0.014320} = 0.1197\\ \text{SE}(b_2) &= \sqrt{0.009653} = 0.09825\end{aligned}$$

- and, a 95-percent confidence interval for β_1 (the income coefficient) is

$$\begin{aligned}\beta_1 &= 0.5987 \pm 2.0181 \times 0.1197\\ &= 0.5987 \pm 0.2416\end{aligned}$$

- To test the hypothesis that both slope coefficients are 0,

$$H_0\text{: } \beta_1 = \beta_2 = 0$$

we have

$$\mathbf{L} = \begin{bmatrix} 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}$$

$$\mathbf{Lb} = \begin{bmatrix} 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix} -6.06466\\ 0.59873\\ 0.54583\end{bmatrix} = \begin{bmatrix} 0.59873\\ 0.54583\end{bmatrix} \text{ (i.e., the two slopes)}$$

$$
\begin{aligned}
F_0 &= \frac{(\mathbf{Lb})' \left[\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}'\right]^{-1} \mathbf{Lb}}{q s_e^2} \\
&= \frac{[0.599, 0.546] \left(\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0.1021 & -0.0008 & -0.0008 \\ -0.0008 & 0.0001 & -0.0000 \\ -0.0008 & -0.0000 & 0.0001 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \right)^{-1} \times \begin{bmatrix} 0.599 \\ 0.546 \end{bmatrix}}{2 \times 178.7309} \\
&= 101.22 \text{ with } 2 \text{ and } 42 \text{ degrees of freedom, } p \simeq 0
\end{aligned}
$$

- To test the hypothesis that the slopes are equal:

$$
\begin{aligned}
\mathbf{L} &= \begin{bmatrix} 0 & 1 & -1 \end{bmatrix} \\
\mathbf{Lb} &= \begin{bmatrix} 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} -6.06466 \\ 0.59873 \\ 0.54583 \end{bmatrix} = 0.05290 \text{ (i.e., the difference in slopes)} \\
F_0 &= \frac{(\mathbf{Lb})' \left[\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}'\right]^{-1} \mathbf{Lb}}{q s_e^2} \\
&= \frac{0.053 \left(\begin{bmatrix} 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 0.1021 & -0.0008 & -0.0008 \\ -0.0008 & 0.0001 & -0.0000 \\ -0.0008 & -0.0000 & 0.0001 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} \right)^{-1} 0.053}{1 \times 178.7309} \\
&= 0.068 \text{ with } 1 \text{ and } 42 \text{ degrees of freedom, } p = .80
\end{aligned}
$$