Lecture 5 - Multivariate linear models

Data-based Statistical Decision Model

Outline

- Today we will consider:
  1. Foundations of matrix algebra.
  2. Special matrices.
  3. Dependence and inversion.
  4. Connection to regression and sums-of-squares.

Matrices

- A matrix $\mathbf{A}$ is a rectangular collection of scalars (numbers).
- $\mathbf{A}$ is a matrix of size $n \times p$ if it has n rows and p columns.
- $\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1(p-1)} & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2(p-1)} & a_{2p} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ a_{(n-1)1} & a_{(n-1)2} & \dots & a_{(n-1)(p-1)} & a_{(n-1)p} \\ a_{n1} & a_{n2} & \dots & a_{n(p-1)} & a_{np} \end{bmatrix}$
- Often write as $\mathbf{A} = [a_{ij}]$ for $i = 1, ..., n$ and $j = 1, ..., n$.

Examples of Matrices

- A sample 2×3 matrix: $\begin{bmatrix} 1 & 2 & 4 \\ 3 & 10 & 743 \end{bmatrix}$.
- A row vector is a $1 \times p$ matrix: $[1, 3, 5, 10]$.
- A column vector is a $n \times 1$ matrix: $\begin{bmatrix} 1 \\ 10 \end{bmatrix}$.
  - Usually “n-vector" refers to a column vector.
- A scalar can be thought of as a 1×1 matrix: $[190]$.
- A square matrix has $n = p$: $\begin{bmatrix} 2 & 93 \\ 234 & 15 \end{bmatrix}$.

Some Matrix Operations

- Equality: Given two matrices $\mathbf{A}$ and $\mathbf{B}$, we say $\mathbf{A} = \mathbf{B}$ if
  1. Both $\mathbf{A}$ and $\mathbf{B}$ are $n \times p$
  2. and $a_{ij} = b_{ij}$ for $i = 1, ..., n$ and $j = 1, ...p$.
- Transpose: If $\mathbf{A} = [a_{ij}]$ for $i = 1, ..., n$ and $j = 1, ..., p$ then
  - $\mathbf{A}' = [a_{ji}]$ for $j = 1, ..., p$ and $i = 1, ..., n$.
  - $\mathbf{A}'$ is a $p \times n$ matrix.
  - $\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$ then $\mathbf{A}' = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$
  - Also written as $\mathbf{A}^T$.
- There are three main arithmetic operations:
  - Matrix addition.
  - Scalar multiplication.
  - Matrix multiplication.

Matrix Addition and Scalar Multiplication

- If $\mathbf{A}$ and $\mathbf{B}$ are both $n \times p$, then
  - $\mathbf{A} + \mathbf{B} = [a_{ij} + b_{ij}]$ for $i = 1, ..., n$ & $j = 1, ..., p$
  - $\begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 1 & 2 \end{bmatrix}$
- For a scalar c: $c\mathbf{A} = [ca_{ij}]$.
  - $2 \begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 2 & 4 \\ 2 & 4 \end{bmatrix}$.
- Combine matrix addition and scalar multiplication to get matrix subtraction:
  - $\begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 2 \\ 1 & 2 \end{bmatrix}$

Matrix Multiplication

- Assume that **A** is $n \times q$ and **B** is $q \times p$.
- Matrix multiplication is defined as: $\mathbf{AB} = \left[\sum_{k=1}^{q} a_{ik} b_{kj}\right]$.
  - Let $a_{i\bullet} = [a_{i1}, ..., a_{ip}]$ be the i^{th} row of **A**.
  - Let $b_{\bullet j} = [b_{1j}, ..., b_{qj}]'$ be the j^{th} column of **B**.
  - $(ab)_{ij} = a_{i\bullet} b_{\bullet j}$
- **AB** is a $n \times p$ matrix.
- Note that, in general, $\mathbf{AB} \neq \mathbf{BA}$.
  - Can formulate both **AB** and **BA** only if they are square.
- $$\begin{bmatrix} 1 & 2 & 0 \\ 1 & 2 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 1 & 2 \\ 1 & 1 \end{bmatrix} =$$ $$\begin{bmatrix} 1*1 + 2*1 + 0*1 & 1*2 + 2*2 + 0*2 \\ 1*1 + 2*1 + 0*1 & 1*2 + 2*2 + 0*2 \end{bmatrix}.$$

Diagonal Matrices and $\mathbf{I}$

- Symmetric Matrix: $\mathbf{A} = \mathbf{A}'$
  - A symmetric matrix must be square.
- Diagonal Matrix: a square matrix such that $a_{ij} = 0$ when $i \neq j$.
  - $\text{diag}(1, 20) = \begin{bmatrix} 1 & 0 \\ 0 & 20 \end{bmatrix}$
  - Diagonal matrices are symmetric.
- Identity Matrix $\mathbf{I}$ or $\mathbf{I_n}$: diagonal $n \times n$ matrix with $a_{ii} = 1$.
  - $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} 1 * a_{11} + 0 * a_{21} & 1 * a_{12} + 0 * a_{22} \\ 0 * a_{11} + 1 * a_{21} & 0 * a_{12} + 1 * a_{22} \end{bmatrix}$
  - $\mathbf{AI_q} = \mathbf{A}$ and $\mathbf{I_q B} = \mathbf{B}$ when $\mathbf{A}$ is $n \times q$ and $\mathbf{B}$ is $q \times p$.

$\mathbf{1}$, $\mathbf{J}$, and $\mathbf{0}$

- $\mathbf{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$
  - Sometimes denoted as $\mathbf{1_n}$ when it is a n-vector.
  - $\mathbf{1}'\mathbf{1} = \sum_{i=1}^{n} 1 = n$.
  - $\mathbf{1}'\mathbf{Y} = \sum_{i=1}^{n} Y = n\overline{Y}$.
- $\mathbf{J}$ is the matrix of ones.
  - Sometimes denoted as $\mathbf{J}_{np}$ when it is a $n \times p$ matrix.
  - Sometimes denoted as $\mathbf{J}_{n}$ when it is a $n \times n$ matrix.
  - $\mathbf{J}_n = \mathbf{1}\mathbf{1}'$
- $\mathbf{0}$ is the matrix of zeroes.
  - Sometimes denoted as $\mathbf{0}_{np}$ when it is a $n \times p$ matrix.
  - Sometimes denoted as $\mathbf{0}_{n}$ when it is a $n \times n$ matrix.

Why Are We Going Through This?

- $$\begin{bmatrix} \beta_0 + X_1\beta_1 \\ \vdots \\ \beta_0 + X_n\beta_1 \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$
- $$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$
- Simple linear regression model becomes:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

- $\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$, $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$, $\mathbf{X} = \begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}$, $\epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$
- $\mathbf{X}$ is called the design matrix.

Linear Dependance

- Consider the set $\{\mathbf{C_1}, \ldots, \mathbf{C_q}\}$ of q column vectors of length n.
- We say that $\{\mathbf{C_1}, \ldots, \mathbf{C_q}\}$ is linearly independent set of vectors when:
  - $\sum_j k_j \mathbf{C_j} = 0$ only when $k_j = 0$ for $j = 1, ..., q$.
  - $\mathbf{C_i} \neq \sum_{j \neq i} k_j \mathbf{C_j}$ for all sets of scalars k_j.
- A set of vectors that is not independent is said to be linearly dependant.
- $k_1 \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix} + k_2 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + k_3 \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \mathbf{0}$ when
  $k_1 = 1, k_2 = 1, k_3 = -2$
- Can have at most n linearly independent vectors of length n.

Rank of a Matrix

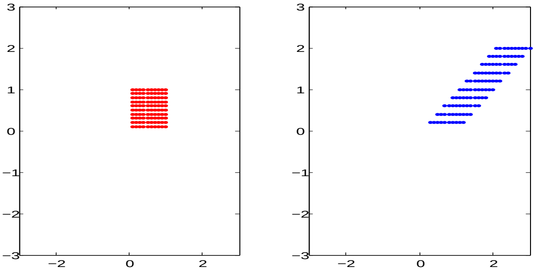
- Consider a $n \times p$ matrix $\mathbf{A}$ as a collection of p column vectors $A_{\bullet j}$.
- The rank of $\mathbf{A}$ is the number of linearly independent columns.
- $\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 3 & 1 & 2 \\ 5 & 1 & 3 \end{bmatrix}$ has rank $(\mathbf{A}) = 2$
- $\text{rank}(\mathbf{A}) \leq \min(n, p)$
- The rank is equivalently the number of linearly independent rows.

Matrices as Operators

- Let R^p be the space of $p-$vectors.
- Can think of the $n \times p$ matrix $\mathbf{A}$ as a map between R^p and R^n.
  - $\mathbf{Ac} = \mathbf{b}$
- The image of $\mathbf{A}$ is the collection of all $n-$vectors $\mathbf{b}$ such that there is a $p-$vector $\mathbf{c}$ where $\mathbf{Ac} = \mathbf{b}$.
  - Note that there might be some $\mathbf{b} \in R^n$ such that there is no $\mathbf{c}$ where $\mathbf{Ac} = \mathbf{b}$.
  - $\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} \neq \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ for all c_1, c_2
- rank($\mathbf{A}$) is the largest set of independent vectors that can be found in the image of $\mathbf{A}$.
  - Intuition: the “dimension" of the image of A.

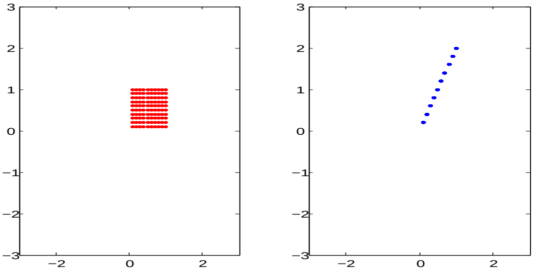
Examples

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 0 & 2 \end{bmatrix}, \text{rank(A)} = 2$$



Examples

$$\mathbf{B} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}, \text{rank(B)} = 1$$

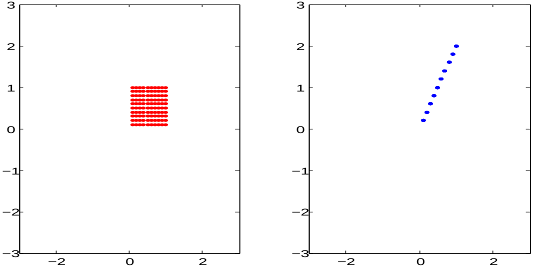


Inverse

- Let $\mathbf{A}$ be a $n \times n$ matrix.
- The inverse of $\mathbf{A}$ is the $n \times n$ matrix $\mathbf{A}^{-1}$ where:
  - $\mathbf{AA^{-1}} = \mathbf{I}$
  - $\mathbf{A^{-1}A} = \mathbf{I}$
- If $\mathbf{Ac} = \mathbf{b}$ then $\mathbf{A^{-1}b} = \mathbf{c}$.
- A matrix only has an inverse if it has “full rank".
  - rank($\mathbf{A}$) = n
- If $\mathbf{A}^{-1}$ does not exist, $\mathbf{A}$ is call singular.
- Inverse of a diagonal matrix diag($a_{11}, ..., a_{nn}$) is diag($a_{11}^{-1}, ..., a_{nn}^{-1}$)

Examples

- $\mathbf{B} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$, rank(B) = 1
- Several red points are mapped onto a single blue point.
- Image will not cover all of R^2



Inverse for 2×2 Matrices

- $\mathbf{A}^{-1} = D^{-1} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}$ where $D = a_{11}a_{22} - a_{12}a_{21}$
- D is called the determinant.
- $D = 0$ for singular matrices.
- $\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} \frac{a_{22}}{D} & \frac{-a_{12}}{D} \\ \frac{-a_{21}}{D} & \frac{a_{11}}{D} \end{bmatrix} =$
$\begin{bmatrix} \frac{a_{11}a_{22}-a_{12}a_{21}}{D} & \frac{-a_{11}a_{12}+a_{12}a_{11}}{D} \\ \frac{a_{21}a_{22}-a_{22}a_{21}}{D} & \frac{-a_{12}a_{21}+a_{11}a_{22}}{D} \end{bmatrix}$

Properties Used in Sums of Squares

- If $\mathbf{c} = [c_1 \ \ldots c_n]'$ then
  - $\mathbf{c}'\mathbf{c} = \sum_{i=1}^{n} c_i^2$.
  - Sums of squares can be written as the product of the transpose of a column vector with itself.
- $\mathbf{X}'\mathbf{X} = \begin{bmatrix} \text{n} & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}$
  - Will see this in the normal equations.
  - $D = n \sum \left(X_i - \overline{X}\right)^2$.
  - When is this singular?
  - $(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} \frac{1}{n} + \frac{\overline{X}^2}{\sum(X_i - \overline{X})^2} & \frac{-\overline{X}}{\sum(X_i - \overline{X})^2} \\ \frac{-\overline{X}}{\sum(X_i - \overline{X})^2} & \frac{1}{\sum(X_i - \overline{X})^2} \end{bmatrix}$

Sum of Squares in Matrix Notation

Recall the simple linear regression model

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon.$$

$$\begin{aligned}
\sum_{i=1}^{n} (Y_i - \beta_0 - X_i\beta_1)^2 &= \sum_{i=1}^{n} \epsilon_i^2 \\
&= \epsilon'\epsilon \\
&= (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta) \\
&= \mathbf{Y}'\mathbf{Y} - \beta'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\beta + \beta'\mathbf{X}'\mathbf{X}\beta \\
&= \mathbf{Y}'\mathbf{Y} - \mathbf{2}\beta'\mathbf{X}'\mathbf{Y} + \beta'\mathbf{X}'\mathbf{X}\beta
\end{aligned}$$

Differentiation of linear and quadratic forms

- Consider $a'x = \sum a_i x_i$ and $x'Ax = \sum_{ij} a_{ij x_i x_j}$.
- $\frac{\delta(a'x)}{\delta x_i} = a_i$ and $\frac{\delta(x'Ax)}{\delta x_i} = \left(\sum_j a_{ij} x_j\right) + (\sum_k a_{ki} x_k)$
- Let $f(x)$ be some function of the $P-$vector x. Define the derivative of x as the P-vector with ith entry $\frac{\delta f(x)}{\delta x_i}$
- $\frac{\delta(a'x)}{\delta x} = a$ and $\frac{\delta(x'Ax)}{\delta x} = (A + A')\,x$

Normal Equations in Matrix Notation

$$\begin{aligned} \sum Y_i &= n\, b_0 + b_1 \sum X_i \\ \sum X_i Y_i &= b_0 \sum X_i + b_1 \sum X_i^2 \end{aligned}$$

- Equivalent to: $\mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{X}B$
  - $\mathbf{B} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$
- Have the solution:
  - $\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{Y}$
  - $\mathbf{H} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'$ is called the hat matrix.

Rewriting the Simple Linear Regression Model

1. The distribution of X is unspecified, possibly even deterministic;
2. $Y \mid X = \beta_0 + \beta_1 x + \epsilon$,where ϵ is a noise variable;
3. ϵ has mean 0, a constant variance σ^2,
4. ϵ is uncorrelated with X and uncorrelated across observations.

With Hints for Multiple Regression

- $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ for $i = 1, ..., n$
- $Y_i = [1\ X_i] \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \epsilon_i$ for $i = 1, ..., n$
- $Y_i = \begin{bmatrix} 1\ X_{i1}\ \dots\ X_{i(p-1)} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \epsilon_i$ for $i = 1, ..., n$

(Take $p = 2$. It's just the simple linear regression.)

- $$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \dots & X_{1(p-1)} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \dots & X_{n(p-1)} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$
- $\mathbf{Y} = \mathbf{X}\beta + \epsilon$
    - $\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & X_{11} & \dots & X_{1(p-1)} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \dots & X_{n(p-1)} \end{bmatrix},$

      $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$

$\mathbf{X}'\mathbf{X}$

- $\mathbf{X}$ is $n \times 2$
- $\mathbf{X}'$ is $2 \times n$
- $\mathbf{X}'\mathbf{X}$ is 2×2

$$
\begin{aligned}
\mathbf{X}'\mathbf{X} &= \begin{bmatrix} 1 & \cdots & 1 \\ X_{11} & \cdots & X_{n1} \\ \vdots & \vdots & \vdots \\ X_{1(p-1)} & \cdots & X_{n(p-1)} \end{bmatrix} \begin{bmatrix} 1 & X_{11} & \ldots & X_{1(p-1)} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \ldots & X_{n(p-1)} \end{bmatrix} \\
&= \begin{bmatrix} 1{*}1 + \ldots + 1{*}1 & 1{*}X_1 + \ldots 1{*}X_n \\ X_1{*}1 + \ldots + X_n{*}1 & X_1 * X_1 + \ldots + X_n X_n \end{bmatrix} \\
&= \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}
\end{aligned}
$$

Least Squares

- We take the least squares estimates b_0, b_1 that minimize $Q(b_0, b_1) = n \times \text{In-Sample MSE}(b_0, b_1)$

$$
\begin{aligned}
Q &= \sum (Y_1 - b_0 - X_i b_1)^2 \\
&= [Y_1 - b_0 - X_1 b_1, \ldots, Y_n - b_0 - X_n b_1] \begin{bmatrix} Y_1 - b_0 - X_1 b_1 \\ \vdots \\ Y_n - b_0 - X_n b_1 \end{bmatrix} \\
&= (\mathbf{Y} - \mathbf{XB})'(\mathbf{Y} - \mathbf{XB})
\end{aligned}
$$

where $\mathbf{B} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix}$

The Normal Equations

- Differentiating Q with respect to b_0 and b_1 finds the minimizers of Q through the normal equations:

$$
\begin{aligned}
nb_0 + b_1 \sum X_i &= \sum Y_i \\
b_0 \sum X_i + b_1 \sum X_i^2 &= \sum X_i Y_i
\end{aligned}
$$

$$
\begin{bmatrix} \text{n} & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}
\begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix}
=
\begin{bmatrix} 1 & \dots & 1 \\ X_{11} & \dots & X_{n1} \\ \vdots & \vdots & \vdots \\ X_{1(p-1)} & \dots & X_{n(p-1)} \end{bmatrix}
\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}
$$

$$\mathbf{X'XB} = \mathbf{X'Y}$$

$$\mathbf{B} = \left(\mathbf{X'X}\right)^{-1} \mathbf{X'Y}$$

Fitted Values

- $\hat{Y}_i = \hat{m}(X_i)$ is the fitted value of the regression curve at X_i
  - Expected value of an observation at X_i
- $\hat{Y}_i = \hat{\beta}_0 + X_i\hat{\beta}_1$ for $i = 1, ...n$
- $\hat{Y}_i = [1\ ,\ X_i] \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_{p-1} \end{bmatrix}$ for $i = 1, ...n$

$$\begin{bmatrix} \hat{Y}_1 \\ \vdots \\ \hat{Y}_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \dots & X_{1(p-1)} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \dots & X_{n(p-1)} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_{p-1} \end{bmatrix}$$

$$\begin{aligned} \hat{\mathbf{Y}} &= \mathbf{XB} \\ \hat{\mathbf{Y}} &= \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} \\ \hat{\mathbf{Y}} &= \mathbf{HY} \end{aligned}$$

- where $\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$ is called the hat matrix.

$\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$

- $\mathbf{H} = [h_{ij}]$ for $i, j = 1, \ldots, n$ where $h_{ij} = \frac{1}{n} + \frac{(X_i - \overline{X})(X_j - \overline{X})}{\sum (X_k - \overline{X})^2}$
- $\mathbf{H}$ is symmetric: $\mathbf{H}' = \mathbf{H}$
- $\mathbf{H}$ is idempotent: $\mathbf{HH} = \mathbf{H}$
- Rank of $\mathbf{H}$ is 2.
- $\mathbf{H}$ is also called influence matrix.

Vector Valued Random Variables

- Consider the vector of random variables $\mathbf{Z} = \begin{bmatrix} Z_1 \\ \vdots \\ Z_n \end{bmatrix}$.
- Assume that
  - $E(Z_i) = \mu_i \qquad \mathrm{Var}(Z_i) = \sigma_i^2 \qquad \mathrm{Cov}(Z_i, Z_j) = \sigma_{ij}$
- We write
  - $E(\mathbf{Z}) = \mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix}$

$$\mathrm{Cov}(\mathbf{Z}) = \Sigma = \begin{bmatrix} \sigma_1^2 & \dots & \sigma_{1j} & \dots & \sigma_{1n} \\ \vdots & & \vdots & & \vdots \\ \sigma_{j1} & \dots & \sigma_j^2 & \dots & \sigma_{jn} \\ \vdots & & \vdots & & \vdots \\ \sigma_{n1} & \dots & \sigma_{nj} & \dots & \sigma_n^2 \end{bmatrix}$$

Properties of Random Vectors

- Let $\mathbf{A}, \mathbf{B}$ be $n \times p$ matrices and $\mathbf{Z}$ a p-vector of random variables.
  - $E(\mathbf{Z}) = \mu$
  - $\text{Cov}(\mathbf{Z}) = \Sigma$
- $E(\mathbf{AZ} + \mathbf{B}) = \mathbf{A}E(\mathbf{Z}) + \mathbf{B} = \mathbf{A}\mu + \mathbf{B}$
- $\text{Cov}(\mathbf{AZ} + \mathbf{B}) = \mathbf{A}\text{Cov}(\mathbf{Z})\mathbf{A}' = \mathbf{A}\Sigma\mathbf{A}'$.
- If $Z_1, \ldots, Z_p$ are normal, then we write $\mathbf{Z} \sim N(\mu, \Sigma)$.
- The $p-$ variate normal distribution has pdf

$$f(\mathbf{Z}) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{Y} - \mu)' \Sigma^{-1} (\mathbf{Y} - \mu)\right]$$

Distribution of Y and $\mathbf{B}$

- Recall: $\mathbf{Y} = \mathbf{X}\beta + \epsilon$.
  - $\epsilon \sim N\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$
  - $\mathbf{Y} \sim N\left(\mathbf{X}\beta, \sigma^2\mathbf{I}\right)$
- Recall: $\mathbf{B} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$

$$\begin{aligned}
\mathbf{B} &\sim N\left(\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\beta, \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left[\sigma^2\mathbf{I}\right]\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right) \\
&\sim N\left(\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{X}\right)\beta, \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{X}\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right) \\
&\sim N\left(\beta, \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)
\end{aligned}$$

- $\mathrm{Var}\left(\mathbf{B}\right) = \sigma^2 \begin{bmatrix} \frac{1}{n} + \frac{\overline{X}^2}{\sum(X_k - \overline{X})^2} & \frac{-\overline{X}}{\sum(X_k - \overline{X})^2} \\ \frac{-\overline{X}}{\sum(X_k - \overline{X})^2} & \frac{1}{\sum(X_k - \overline{X})^2} \end{bmatrix}$

Distribution of $\hat{Y}$

- Recall: $\hat{\mathbf{Y}} = \mathbf{HY}$

$$\begin{aligned}\hat{\mathbf{Y}} &\sim N\left(\mathbf{HX}\beta, \mathbf{H}\left[\sigma^2\mathbf{I}\right]\mathbf{H}'\right)\\ &\sim N\left(\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\beta, \sigma^2\mathbf{HH}\right)\\ &\sim N\left(\mathbf{X}\beta, \sigma^2\mathbf{H}\right)\end{aligned}$$

- $\mathrm{Var}\left(\hat{Y}_i\right) = \sigma^2 h_{ii} = \sigma^2\left(\frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum(X_k - \overline{X})^2}\right)$

Residuals

$$\mathbf{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix} = \mathbf{Y} - \hat{\mathbf{Y}}$$
$$= \mathbf{IY} - \mathbf{HY}$$
$$= (\mathbf{I} - \mathbf{H})\,\mathbf{Y}$$

$$\begin{aligned} E(\mathbf{e}) &= (\mathbf{I} - H)\,E(\mathbf{Y}) \\ &= (\mathbf{I} - H)\,\mathbf{XY} \\ &= \left(\mathbf{X} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\right)\mathbf{Y} \\ &= (\mathbf{X} - \mathbf{X})\,\mathbf{Y} = \mathbf{0} \end{aligned}$$

Covariance Structure of the Residuals

$$\begin{aligned} \text{Cov}(\mathbf{e}) &= (\mathbf{I}-\mathbf{H})'\left(\sigma^2\mathbf{I}\right)(\mathbf{I}-\mathbf{H}) \\ &= \sigma^2\left[\mathbf{I}-\mathbf{H}-\mathbf{H}'+\mathbf{H}'\mathbf{H}\right] \\ &= \sigma^2\left[\mathbf{I}-\mathbf{H}\right] \end{aligned}$$

- $\text{Var}\left(e_i\right) = \sigma^2\left[1 - \frac{1}{n} - \frac{\left(X_i-\overline{X}\right)^2}{\sum\left(X_k-\overline{X}\right)^2}\right]$
- $\text{Cov}\left(e_i, e_j\right) = -\sigma^2\left[\frac{1}{n} + \frac{\left(X_i-\overline{X}\right)\left(X_j-\overline{X}\right)}{\sum\left(X_k-\overline{X}\right)^2}\right]$ when $i \neq j$
- Notice that as $n \to \infty$:
  - $\frac{1}{n} \to 0$
  - $\frac{\left(X_i-\overline{X}\right)\left(X_j-\overline{X}\right)}{\sum\left(X_k-\overline{X}\right)^2} \to 0$
- This means that:
  - $\text{Var}\left(e_i\right) \to \sigma^2$
  - $\text{Cov}\left(e_i, e_j\right) \to 0$

Regression With Two Variable

- Assume that we have:
  - $i = 1, \ldots, n$ observations.
  - Responses Y_i .
  - First Covariates X_{i1} .
  - Second Covariates X_{i2} .
- $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$
  - $\epsilon_i \sim N(0, \sigma^2)$ and ϵ_i and ϵ_j are independent when $i \neq j$.
- First order linear regression model with two predictors.
- $E(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$
  - Regression function is now a function of two variables: $f(X_1, X_2)$.

Assumptions

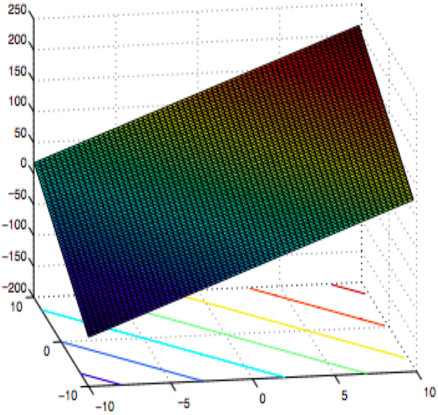
The first-order linear regression model assumes that:

1. $E(Y_i)$ is linear in both X_{i1} and X_{i2}.
    - The regression function is a plane.
    - The regression function is linear in both X_{i1} and X_{i2}.
    - The association between Y and X_1 does not depend on X_2.
2. The error terms have the same variance.
3. The error terms are independent.
4. The error terms are normal.

A regression surface

$$Y = m(X) + \epsilon,$$

where $m(X) = m(X_1, X_2) = 50 + 10X_1 + 7X_2$.



Interpretation

- $E(Y_i|X_{i1} = C_1, X_{i2} = C_2) = \beta_0 + C_1\beta_1 + C_2\beta_2$
- $E(Y_i|X_{i1} = C_1 + 1, X_{i2} = C_2) = \beta_0 + (C_1 + 1)\beta_1 + C_2\beta_2$
- $E(Y_i|X_{i1} = C_1 + 1, X_{i2} = C_2) - E(Y_i|X_{i1} = C_1, X_{i2} = C_2) = \beta_1$
    - This holds regardless of what C_2 is.
    - β_1 is the expected increase in Y for any fixed level of X_2 from an increase in one unit of X_1.
    - Assuming that X_1 and X_2 have an additive effect on Y or that they do not interact.
    - Often called the association between Y and X_1 controlling for X_2.

Regression With Many Variables

- Assume that we are interested in the relationship between the $(p-1)$ variables $X_1, \ldots, X_{p-1}$ and Y.
- Can form the first order linear regression model with $(p-1)$ predictors.

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i(p-1)} + \epsilon_i \\
&= \beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ij} + \epsilon_i
\end{aligned}
$$

- $E(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i(p-1)}$
  - Assumes that the expected value of Y is linear in all of the predictors.
  - Forms a hyperplane.
- β_j is the expected increase in Y for an increase in X_j by one unit while holding all other predictors fixed.
  - Assumes that the relationship between X_j and Y_i does not change as the other predictors change.

Multiple Regression In Matrix Notation

- For $i = 1, \ldots, n$

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i(p-1)} + \epsilon_i \\
&= \left[1,\ X_{i1},\ \ldots,\ X_{i(p-1)}\right] \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}
\end{aligned}
$$

- $\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \ldots & X_{1(p-1)} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \ldots & X_{n(p-1)} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$
- $\mathbf{Y} = \mathbf{X}\beta + \epsilon$
- Notice that $\mathbf{X}$ is $n \times p$.
  - This is why we use the notation that there are (p-1) predictors.

Special Types of Predictors

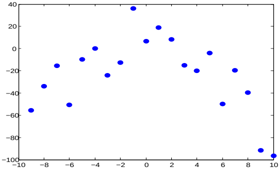
- The predictor variables do not have to be $(p-1)$ unrelated quantitative variables.
- Will now just mention some special types. Will go into greater detail for each type individually later in the semester.
  1. Qualitative Variables: Sex, Race, Car Maker
  2. Polynomial Regression: $X_{i2} = X_{i1}^2$
  3. Transformations: X_{i1} is the log(dosage) of a drug
     - Analogous to transformation in simple linear regression.
  4. Interaction Terms: $X_{i3} = X_{i1}X_{i2}$

Qualitative Variables

- A qualitative variable does not correspond to a particular numeric value.
- Assume we want to know the relationships between age and sex on yearly salary among Philadelphians between the ages of 18-62.
- Let X_{i1}=(age in years) and Y_i=(yearly salary)
- Let $X_{i2} = 1$ for male and 0 for female.
- Assume the model: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$.
- $E\left(Y_i \mid \text{c years old, Male}\right) = \beta_0 + c\beta_1 + \beta_2$
- $E\left(Y_i \mid \text{c years old, Female}\right) = \beta_0 + c\beta_1$
- $\beta_2 = E\left(Y_i \mid \text{c years old, Male}\right) - E\left(Y_i \mid \text{c years old, Female}\right)$
- β_2 is the difference in expect salary between males and females at any given age.

Polynomial Regression

- What if Y_i and X_i have a parabolic relationship rather than a linear.

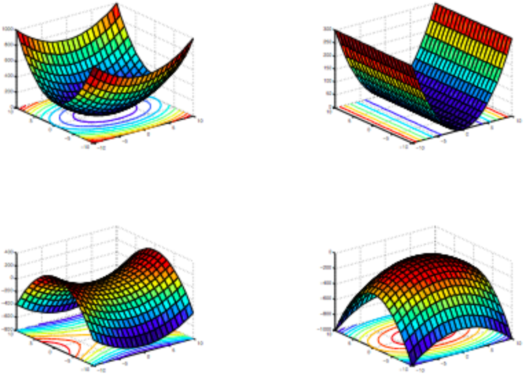


- By visual in inspection, we think $E\left(Y_i\right) = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$
- Let $X_{i1} = X_i$ and $X_{i2} = X_i^2$.
- Fit $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$.

Interaction Effects

- Recall the example where we want to look at the effects of age and sex on yearly salary among Philadelphians between the ages of 18-62.
- Fit $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$.
- β_1 is the expected increase in salary associated with an increase in one year of age for both males and females.
- What if we expect that that amount of extra money earned by a man next year is larger than the amount of money earned by a female next year?
    - The effect of X_{i1} differs for different values of X_{i2}.
    - The variables interact.
- Let $X_{i3} = X_{i1} X_{i2}$ and fit $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3}$
- β_1 is the expected increase in salary for a female associated with an increase in one year of age.
- $\beta_1 + \beta_3$ is the expected increase in salary for a male associated with an increase in one year of age.

A bit more general regression surface



Least Squares Estimation

- We need to estimate $\beta_0, \beta_1, \ldots, \beta_{p-1}$.
- Will use the least squares estimates $b_0, b_1, \ldots, b_{p-1}$ that minimize:

$$
\begin{aligned}
Q &= \sum_{i=1}^{n} \left(Y_i - b_0 - \sum_{k=1}^{p-1} X_{ik} b_k \right)^2 \\
&= (\mathbf{Y} - \mathbf{X}\beta)' (\mathbf{Y} - \mathbf{X}\beta)
\end{aligned}
$$

Finding $b_0, \ldots, b_{p-1}$

- Differentiating, we find the p equations:
  - $\frac{\partial Q}{\partial b_0} = -2\sum_{i=1}^{n}\left(Y_i - b_0 - \sum_{k=1}^{p-1} X_{ik} b_k\right)$
  - For $j = 1, \ldots, (p-1)$:
    $\frac{\partial Q}{\partial b_j} = -2\sum_{i=1}^{n} X_{ij}\left(Y_i - b_0 - \sum_{k=1}^{p-1} X_{ik} b_k\right)$
- Setting these equal to zero, we get the normal equations:
  - $nb_0 + \sum_{k=1}^{p-1} b_k \left(\sum_{i=1}^{n} X_{ik}\right) = \sum_{i=1}^{n} Y_i$
  - For $j = 1, \ldots, (p-1)$:
    $b_0 \left(\sum_{i=1}^{n} X_{ij}\right) + \sum_{k=1}^{p-1} b_k \left(\sum_{i=1}^{n} X_{ij} X_{ik}\right) = \sum_{i=1} n X_{ij} Y_i$

Solving the normal equations

$$\begin{bmatrix} n & \sum X_{i1} & \cdots & \sum X_{i(p-1)} \\ \sum X_{i1} & \sum X_{i1}^2 & \cdots & \sum X_{i1}X_{i(p-1)} \\ \vdots & \vdots & & \vdots \\ \sum X_{in} & \sum X_{i1}X_{i(p-1)} & \cdots & \sum X_{i(p-1)}^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & \cdots & 1 \\ X_{11} & \cdots & X_{n1} \\ \vdots & \vdots & \vdots \\ X_{1(p-1)} & \cdots & X_{n(p-1)} \end{bmatrix} \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$$

- $(\mathbf{X}'\mathbf{X})\,\mathbf{B} = \mathbf{X}'\mathbf{Y}$
- $\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{Y}$
- $\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$ where $\mathbf{H} = \mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'$

When Does **B** Exist?

- If **A** has rank r, then **A′A** also has rank r.
- **X** is $n \times p$ $\begin{bmatrix} 1 & X_{11} & \dots & X_{1(p-1)} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \dots & X_{n(p-1)} \end{bmatrix}$
- For $(\mathbf{X}'\mathbf{X})^{-1}$ to exist, **X** must have rank p.
- $(\mathbf{X}'\mathbf{X})^{-1}$ won't exist if a predictor is a linear combination of other predictors.
  - Called colinearity.
  - There are more than one solution of **B**.
- All of the information contained in one predictor can be obtained from the other $(p-2)$ predictors.
- In (old) practice there is never perfect co-linearity.
  - A predictor can be close enough to a linear combination of the others so that the computer cannot invert $(\mathbf{X}'\mathbf{X})$.

When colinearity can happen?

- If $n < p$, the data are collinear.
- If one of the predictor variables is constant, the data are collinear.
- If two of the predictor variables are proportional to each other, the data are collinear.
- If two of the predictor variables are otherwise linearly related, the data are collinear.

Estimation of σ^2

- Similar to simple linear regression, we use the MSE to estimate σ^2.

$$
\begin{aligned}
SSE &= \sum \left(Y_i - \hat{Y}_i\right)^2 \\
&= (\mathbf{IY} - \mathbf{HY})' (\mathbf{IY} - \mathbf{HY}) \\
&= \mathbf{Y}' [\mathbf{I} - \mathbf{H}]' [\mathbf{I} - \mathbf{H}] \mathbf{Y} \\
&= \mathbf{Y}' [\mathbf{I} - \mathbf{H}] \mathbf{Y}
\end{aligned}
$$

- $\mathbf{H}$ has rank p so $\mathbf{I} - \mathbf{H}$ has rank $n - p$.
- df of SSE is $n - p$
  - Intuition: we have n independent observations but must estimate the p parameters $b_0, \ldots, b_{p-1}$.
- $MSE = SSE/(n - p)$

Inference for $\mathbf{B}$

- Assuming *Gaussianity*, we can use the distributions of the studentized statistics to get $t-$ based inference for b_j.
- Must first find the standard error of b_j.

$$
\begin{aligned}
\text{Var}(\mathbf{B}) &= \text{Var}\left(\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{Y}\right) \\
&= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\text{Var}(\mathbf{Y})\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right) \\
&= \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right) \\
&= \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}
\end{aligned}
$$

- We estimate the standard error of b_j as $s(b_j) = \sqrt{MSE\left(\mathbf{X}'\mathbf{X}\right)^{-1}_{jj}}$

t-test for b_j

- Can form a studentized t-statistic:

$$\frac{b_j - \beta_j}{s(b_j)} \sim t_{n-p}$$

- We now have a $t-$ distribution with $n-p$ degrees of freedom.
- Can invert to get two sided $1-\alpha\%$ confidence intervals:

$$b_j \pm t_{n-p}(1-\alpha/2)\, s(b_j),\ \text{j=0,...,(p-1)}$$

- Can preform a level α $t-$test of $H_0 : \beta_j = C$ vs $H_a : \beta_j \neq C$ by rejecting H_0 when

$$\left|\left(b_j - C\right)/s(b_j)\right| > t_{n-p}(1-\alpha/2)$$

Bonferroni for **B**

- The confidence intervals $b_j \pm t_{n-p}(1-\alpha/2)$ only hold if we want to look at one j.
- If we want to draw inference for more than one coefficient (or all p-1 of them), we will have an inflated type I error rate.
- Example: we are doing a study to explore factors associated with high school truancy. We regress number of school days missed per year on the 10 variables: age, weight, family income, distance to school, teacher's age, size of school, grade, participation in sports, ability to read music, average hrs of tv watched per day
  - Would test if there is an association between days missed and the 10 covariate by testing if $\beta_j = 0$ for $j = 1, \ldots, 10$.
- Can use Bonferroni.
- The simultaneous level α Bonferroni confidence intervals for g coefficients are

$$b_j \pm t_{n-p}(1-\alpha/2g)$$