Advanced Methods in Biostatistics I

Lecture 4

Martin Lindquist

September 7, 2017

Simple linear regression - revisited

- We seek to minimize:

$$||\mathbf{y} - (\beta_0 \mathbf{J}_n + \beta_1 \mathbf{x})||^2$$

  over β_0 and β_1.

- We can do this by finding the projection of $\mathbf{y}$ onto Γ, which is the two dimensional linear subspace of $\mathbb{R}^n$ spanned by the two vectors, $\mathbf{J}_n$ and $\mathbf{x}$.

Projections

Theorem

Let $\mathbf{W}$ be a subspace and $\mathbf{y}$ a vector in $\mathbf{V}$. Assume $\{\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_k\}$ is an orthogonal basis for $\mathbf{W}$. Then the vector

$$\hat{\mathbf{y}} = \sum_{i=1}^{k} \frac{\langle \mathbf{y}, \mathbf{x}_i \rangle}{\langle \mathbf{x}_i, \mathbf{x}_i \rangle} \mathbf{x}_i$$

in $\mathbf{W}$ is the projection of $\mathbf{y}$ onto $\mathbf{W}$.

Simple linear regression - revisited

- Compute an orthogonal basis for Γ, for example $\mathbf{u}_1 = \mathbf{J}_n$ and $\mathbf{u}_2 = \mathbf{x} - \bar{x}\mathbf{J}_n$.
- The projection of $\mathbf{y}$ onto Γ can be expressed as the sum of the individual projections of $\mathbf{y}$ onto $\mathbf{u}_1$ and $\mathbf{y}$ onto $\mathbf{u}_2$, i.e. $\hat{\mathbf{y}} = \hat{\mathbf{y}}_1 + \hat{\mathbf{y}}_2$.

Simple linear regression - revisited

- The projection of $\mathbf{y}$ onto $\mathbf{u}_1$ can be expressed as $\hat{\mathbf{y}}_1 = \hat{\alpha}_0 \mathbf{J}_n$ where $\hat{\alpha}_0 = \bar{y}$.
- The projection of $\mathbf{y}$ onto $\mathbf{u}_2$ can be expressed as $\hat{\mathbf{y}}_1 = \hat{\alpha}_1(\mathbf{x} - \bar{x}\mathbf{J}_n)$ where

$$\hat{\alpha}_1 = \frac{(\mathbf{x} - \bar{x}\mathbf{J}_n)'\mathbf{y}}{(\mathbf{x} - \bar{x}\mathbf{J}_n)'(\mathbf{x} - \bar{x}\mathbf{J}_n)} = \frac{(\mathbf{x} - \bar{x}\mathbf{J}_n)'(\mathbf{y} - \bar{y}\mathbf{J}_n)}{(\mathbf{x} - \bar{x}\mathbf{J}_n)'(\mathbf{x} - \bar{x}\mathbf{J}_n)}.$$

Simple linear regression - revisited

- Note that $\hat{\alpha}_1 = \hat{\beta}_1$ from before.
- Thus, we can write

$$\begin{aligned}\hat{\mathbf{y}} &= \hat{\mathbf{y}}_1 + \hat{\mathbf{y}}_2 \\ &= \bar{y}\mathbf{J}_n + \hat{\beta}_1(\mathbf{x} - \bar{x}\mathbf{J}_n) \\ &= (\bar{y} - \hat{\beta}_1\bar{x})\mathbf{J}_n + \hat{\beta}_1\mathbf{x}\end{aligned}$$

- Setting $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$ provides the familiar solution.

Least squares

- In today's class we will consider the general least squares problem.
- But we begin with some review of matrix algebra.

Matrices

- We are often interested in working with subspaces spanned by a set of vectors.
- Operations on p vectors of length n may be performed by combining them into an $n \times p$ matrix and manipulating this matrix.
- For every matrix $\mathbf{A} \in \mathbb{R}^{n \times p}$, we define a number of fundamental vector spaces.

Column space

Definition

The column space (or range) of a matrix $\mathbf{A}$, denoted $\mathcal{R}(\mathbf{A})$, is defined as the linear space spanned by the columns of $\mathbf{A}$.

Rank

Definition

The rank of the matrix $\mathbf{A}$ is the number of linearly independent columns of $\mathbf{A}$ (i.e., the dimension of $\mathcal{R}(\mathbf{A})$), or equivalently, the number of linearly independent rows of $\mathbf{A}$.

Rank

Theorem

Decreasing property of rank:

$$\mathrm{rank}(\mathbf{AB}) \leq \min\{\mathrm{rank}(\mathbf{A}), \mathrm{rank}(\mathbf{B})\}.$$

Range, Rank, and Null Space

Theorem

$$\operatorname{rank}(\mathbf{A}) = \operatorname{rank}(\mathbf{A}') = \operatorname{rank}(\mathbf{A}'\mathbf{A}) = \operatorname{rank}(\mathbf{A}\mathbf{A}').$$

Null Space

Definition

The null space of a matrix $\mathbf{A}$ is $\mathcal{N}(\mathbf{A}) \equiv \{\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{0}\}$. The nullity of $\mathbf{A}$ is the dimension of $\mathcal{N}(\mathbf{A})$.

Range, Rank, and Null Space

Theorem

$\operatorname{rank}(\mathbf{A}) + \operatorname{nullity}(\mathbf{A}) = p$, the number of columns of **A**.

Linear transformations

Definition

Let $\mathbf{V}$ be a n dimensional vector space and let $\mathbf{W}$ be an p dimensional vector space. A linear transformation $\mathbf{L}$ from $\mathbf{V}$ to $\mathbf{W}$ is a mapping (function) from $\mathbf{V}$ to $\mathbf{W}$ such that

$$\mathbf{L}(\alpha\mathbf{x}+\beta\mathbf{y}) = \alpha\mathbf{L}(\mathbf{x})+\beta\mathbf{L}(\mathbf{y}) \text{ for every } \mathbf{x},\mathbf{y} \in \mathbf{V} \text{ and all } \alpha,\beta \in \mathbb{R}.$$

Linear transformations

- We can regard the $n \times p$ matrix $\mathbf{A}$ as transforming elements of $\mathbb{R}^n$ to $\mathbb{R}^p$ by computing

$$\mathbf{L}(\mathbf{x}) = \mathbf{A}\mathbf{x}$$

- Such a transformation is called a matrix transformation.
- Every linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^p$ is a matrix transformation.

Projections

Theorem

If $\mathbf{V}$ is a vector space and $\mathbf{W}$ is a subspace of $\mathbf{V}$, then $\exists$ two vectors, $\mathbf{w}_1, \mathbf{w}_2 \in \mathbf{V}$ such that

- $\mathbf{y} = \mathbf{w}_1 + \mathbf{w}_2 \quad \forall\, \mathbf{y} \in \mathbf{V}$,
- $\mathbf{w}_1 \in \mathbf{W}$ and $\mathbf{w}_2 \in \mathbf{W}^{\perp}$.
- The vector $\mathbf{w}_1$ is the projection of $\mathbf{y}$ onto $\mathbf{W}$.
- The vector $\mathbf{w}_2$ is the projection of $\mathbf{y}$ onto $\mathbf{W}^{\perp}$.

Projections

- For each subspace $\mathbf{W}$ of $\mathbf{V}$ there exists a projection matrix denoted $\mathbf{P}$.
- This is the matrix that defines the linear transformation of orthogonal projection onto $\mathbf{W}$.
- We can express this as $\mathbf{w}_1 = \mathbf{P}\mathbf{y}$.
- Similarly, $\mathbf{w}_2 = (\mathbf{I} - \mathbf{P})\mathbf{y}$.

Projection Matrices

Theorem

A matrix $\mathbf{P}$ is a projection matrix if and only if it is symmetric ($\mathbf{P}' = \mathbf{P}$) and idempotent ($\mathbf{P}^2 = \mathbf{P}$).

Projections

- Recall for $\hat{\mathbf{y}} = b_1\mathbf{x}_1 + \cdots + b_k\mathbf{x}_p$ to be the projection of $\mathbf{y}$ onto $\mathbf{W} = sp\{\mathbf{x}_1, \ldots \mathbf{x}_p\}$ we need $\langle \mathbf{y} - \hat{\mathbf{y}}, \mathbf{x}_i \rangle = 0$ for all i.
- Let $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 \cdots \mathbf{x}_p \end{bmatrix}$.
- In matrix format we can express the orthogonality constraint as:

$$\mathbf{X}'(\mathbf{y} - \hat{\mathbf{y}}) = 0$$

or

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\hat{\mathbf{y}}$$

Projections

- Similarly, the projection can be written:

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$$

- Thus, we have

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\mathbf{b}$$

- Solving for **b** we obtain:

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

Projections

- Hence, we can express $\hat{\mathbf{y}}$ as follows:

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{Xb} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \end{aligned}$$

- The matrix $P = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is called the projection matrix for the subspace **W.**
- Note: $\mathbf{I} - \mathbf{P}$ is a projection matrix onto $\mathbf{W}^{\perp}$.

Definition

Let $u = f(\mathbf{x})$ be a function of the variables $\mathbf{x} = (x_1, x_2, \ldots x_p)'$, and let $\partial u/\partial x_1$, $\partial u/\partial x_2$, ..., $\partial u/\partial x_p$ be the partial derivatives. Then,

$$\frac{\partial u}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial u}{\partial x_1} \\ \frac{\partial u}{\partial x_2} \\ \vdots \\ \frac{\partial u}{\partial x_p} \end{pmatrix}.$$

Vector and Matrix Calculus

Theorem

Let $u = \mathbf{a}'\mathbf{x} = \mathbf{x}'\mathbf{a}$, where $\mathbf{a} = (a_1, a_2, \ldots a_p)'$ is a vector of constants. Then,

$$\frac{\partial u}{\partial \mathbf{x}} = \frac{\partial(\mathbf{a}'\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial(\mathbf{x}'\mathbf{a})}{\partial \mathbf{x}} = \mathbf{a}.$$

Theorem

Let $u = \mathbf{x}'\mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is a symmetric matrix of constants. Then,

$$\frac{\partial u}{\partial \mathbf{x}} = \frac{\partial(\mathbf{x}'\mathbf{A}\mathbf{x})}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}.$$

General linear model

- Now we seek to develop least squares for the general linear model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, where

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{10} & x_{11} & \cdots & x_{1,p-1} \\ x_{20} & x_{21} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{n,p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Design matrix

- Let $\mathbf{X}$ be a design matrix, notationally its elements and column vectors are given by:

$$\mathbf{X} = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \dots & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix} = [\mathbf{x}_1 \dots \mathbf{x}_p].$$

- We are assuming that $n \geq p$ and $\mathbf{X}$ is of full (column) rank.

Least squares

- Now consider the ordinary least squares criteria:

$$\begin{aligned} f(\beta) &= ||\mathbf{y} - \mathbf{X}\beta||^2 \\ &= (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) \\ &= \mathbf{y}'\mathbf{y} - 2\beta'\mathbf{X}'\mathbf{y} + \beta'\mathbf{X}'\mathbf{X}\beta. \end{aligned}$$

Normal equations

- To minimize $f(\beta)$, we begin by taking the derivative with respect to β:
$$\frac{df}{d\beta} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\beta.$$
- Solving for 0 leads to the so called normal equations:
$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y}.$$

Solution

- Note that the matrix $\mathbf{X}'\mathbf{X}$ retains the same rank as $\mathbf{X}$.
- Thus, it is a full rank $p \times p$ matrix and invertible.
- We can then solve the normal equations as:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Hessian

- The Hessian is $2\mathbf{X}'\mathbf{X}$, which is positive definite.
- This is true because for any non-zero vector, $\mathbf{a}$, we have that $\mathbf{X}'\mathbf{a}$ is non-zero since $\mathbf{X}$ is full rank and then $\mathbf{a}'\mathbf{X}'\mathbf{X}\mathbf{a} = ||\mathbf{X}\mathbf{a}||^2 > 0$.
- Thus, the root of our derivative is indeed a minimum.

R code

- The data set `swiss` in `R` contains data on standardized fertility measure and socio-economic indicators for each of 47 French-speaking provinces of Switzerland in 1888.

```
A data frame with 47 observations on 6 variables,
each of which is in percent, i.e., in [0,100].
[,1]  Fertility  Ig, common standardized fertility measure.
[,2]  Agriculture % males involved in agriculture as occupation.
[,3]  Examination % draftees receiving highest mark on army exam
[,4]  Education  % education beyond primary school for draftees.
[,5]  Catholic  % catholic (as opposed to protestant).
[,6]  Infant.Mortality live births who live less than 1 year.

All variables but Fertility give proportions of the population.
```

R code

```
> y = swiss$Fertility
> x = as.matrix(swiss[,-1])
> solve(t(x) %*% x, t(x) %*% y)
                       [,1]
1                66.9151817
Agriculture      -0.1721140
Examination      -0.2580082
Education        -0.8709401
Catholic          0.1041153
Infant.Mortality  1.0770481

> summary(lm(y ~ x - 1))$coef
                   Estimate  Std. Error   t value     Pr(>|t|)
1                66.9151817 10.70603759  6.250229 1.906051e-07
Agriculture      -0.1721140  0.07030392 -2.448142 1.872715e-02
Examination      -0.2580082  0.25387820 -1.016268 3.154617e-01
Education        -0.8709401  0.18302860 -4.758492 2.430605e-05
Catholic          0.1041153  0.03525785  2.952969 5.190079e-03
Infant.Mortality  1.0770481  0.38171965  2.821568 7.335715e-03
```

Fitted values

- The vector of fitted values is given by

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

- Multiplication by the matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ takes any vector in $\mathbb{R}^n$ and produces the fitted values.

- Typically $\mathbf{H}$ is referred to as the 'hat matrix' since it transforms $\mathbf{y}$ into $\hat{\mathbf{y}}$.

Residuals

- The vector of residuals is given by

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y}.$$

- Multiplication by $(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')$ produces the residuals.

Fitted values & residuals

- Note that because $\hat{\mathbf{y}}$ vector is a linear combination of $\mathbf{X}$, it is orthogonal to the residuals, i.e.

$$\hat{\mathbf{y}}'\mathbf{e} = \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y} = 0.$$

Geometrical perspective

- Consider the column space of the design matrix,

$$\Gamma = \{\mathbf{X}\beta \mid \beta \in \mathbb{R}^p\}.$$

- This p-dimensional space belongs to $\mathbb{R}^n$.

Geometrical perspective

- Consider the vector $\mathbf{y} \in \mathbb{R}^n$.
- Multiplication by the matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ projects $\mathbf{y}$ into Γ.
- That is,
$$\mathbf{y} \to \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$
is the linear projection map between $\mathbb{R}^n$ and Γ.

Geometrical perspective

- The vector $\hat{\mathbf{y}}$ is the point in Γ that is closest to $\mathbf{y}$ and $\hat{\beta}$ is the specific linear combination of the columns of $\mathbf{X}$ that yields $\hat{\mathbf{y}}$.
- The vector $\mathbf{e}$ is the vector connecting $\mathbf{y}$ and $\hat{\mathbf{y}}$, and is orthogonal to all elements in Γ, i.e. it lies in $\Gamma^{\perp}$

Geometrical perspective

- Note that if $\mathbf{W}$ is any $p \times p$ invertible matrix, then the fitted values, $\hat{\mathbf{y}}$ will be the same for the design matrix $\mathbf{XW}$.
- This holds because the spaces $\{\mathbf{X}\beta \mid \beta \in \mathbb{R}^p\}$ and $\{\mathbf{XW}\gamma \mid \gamma \in \mathbb{R}^p\}$ are the same, since if $\mathbf{a} = \mathbf{X}\beta$ then $\mathbf{a} = \mathbf{X}\gamma$ via the relationship $\gamma = \mathbf{W}\beta$.

Geometrical perspective

- Thus, any element of the first space is in the second.
- The same argument implies in the other direction, thus the two spaces are the same.
- Thus, any linear reorganization of the columns of $\mathbf{X}$ results in the same column space and the same fitted values.

Full row rank case

- In the case where **X** is $n \times n$ of full rank, then the columns of **X** form a basis for $\mathbb{R}^n$.
- In this case, $\hat{\mathbf{y}} = \mathbf{y}$, since **y** lives in the space spanned by the columns of **X**.
- All this linear model accomplishes is a lossless linear reorganization of **y**.

Full row rank case

- This is surprisingly useful, especially when the columns of $\mathbf{X}$ are orthonormal ($\mathbf{X}'\mathbf{X} = \mathbf{I}$).
- In this case, the function that takes the outcome vector and converts it to the coefficients is called a ”transform”.
- The most well known versions of transforms are Fourier and wavelet.