Advanced Methods in Biostatistics I

Lecture 2

Martin Lindquist

August 31, 2017

Single parameter regression

- In today's class we will consider two simple linear models consisting of a single variable: *mean only regression* and *regression through the origin*.
- In both cases we will seek the least squares estimate of the unknown model parameter.
- But we begin with some review of linear algebra.

Vector space

Definition

A (real) vector space consists of a non empty set $\mathbf{V}$ and two operations:

(1) Addition is defined for pairs of elements $\mathbf{x}$ and $\mathbf{y} \in \mathbf{V}$, and yields an element in $\mathbf{V}$, denoted $\mathbf{x} + \mathbf{y}$.

(2) Scalar multiplication is defined for a real number α and an element $\mathbf{x} \in \mathbf{V}$, and yields an element in $\mathbf{V}$ denoted $\alpha\mathbf{x}$.

Vector space

Properties

The following properties must hold for $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbf{V}, \alpha, \beta \in \mathbb{R}$:

(1) $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$

(2) $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$

(3) There is an element in $\mathbf{V}$ denoted $\mathbf{0}$ such that $\mathbf{0} + \mathbf{x} = \mathbf{x} + \mathbf{0} = \mathbf{x}$

(4) For each $\mathbf{x} \in \mathbf{V}$ there is an element in $\mathbf{V}$ denoted $-\mathbf{x}$ such that $\mathbf{x} + (-\mathbf{x}) = (-\mathbf{x}) + \mathbf{x} = \mathbf{0}$

(5) $\alpha(\mathbf{x} + \mathbf{y}) = \alpha\mathbf{x} + \alpha\mathbf{y}$ for all α

(6) $(\alpha + \beta)\mathbf{x} = \alpha\mathbf{x} + \beta\mathbf{x}$ for all α, β

(7) $1\mathbf{x} = \mathbf{x}$

(8) $\alpha(\beta\mathbf{x}) = (\alpha\beta)\mathbf{x}$ for all α, β

Notation

- The elements of $\mathbf{V}$ are called vectors.
- The elements in $\mathbb{R}$ are called scalars.

Example

- Let $\mathbf{V} = \mathbb{R}^{\mathbf{n}}$ be the set of all n-tuples (i.e., ordered set) of real numbers.
  - Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ belong to $\mathbf{V}$.
  - Addition defined as $\mathbf{x} + \mathbf{y} = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n)$.
  - Scalar multiplication defined by $\alpha\mathbf{x} = (\alpha x_1, \alpha x_2, \ldots, \alpha x_n)$.

- $\mathbf{V}$ is a vector space.

Subspace

Definition

A subset $\mathbf{U}$ of a vector space $\mathbf{V}$ is a subspace of $\mathbf{V}$ if and only if the following properties hold:

- $\mathbf{0} \in \mathbf{U}$
- $\mathbf{U}$ is closed under vector addition, i.e., if $\mathbf{u}_1$ and $\mathbf{u}_2$ are in $\mathbf{U}$, then $\mathbf{u}_1 + \mathbf{u}_2 \in \mathbf{U}$.
- $\mathbf{U}$ is closed under scalar products, i.e., if c is a scalar and $\mathbf{u} \in \mathbf{U}$, then $c\,\mathbf{u} \in \mathbf{U}$.

If these properties hold, $\mathbf{U}$ is also a vector space.

Example

- Let **U** be the set of all elements $(x_1, x_2, \ldots, x_n)$ in $\mathbb{R}^n$ such that $x_n = 0$.
- **U** is a vector subspace of $\mathbf{V} = \mathbb{R}^n$.

Span

Definition

If $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ is any set of vectors in a vector space $\mathbf{V}$, the set of all linear combinations of these vectors is called their span.

Linearly independent

Definition

A collection of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ are linearly independent if $\sum_i c_i \mathbf{v}_i \neq 0$ unless $c_i = 0$ for all i.

Linear basis

Definition

A linear basis for a vector space $\mathbf{V}$ is a set of linearly independent vectors which span $\mathbf{V}$.

Dimension

Definition

The dimension of a vector space is the number of vectors in any basis of the vector space.

Inner product

Definition

Inner product of two vectors:

$$< \mathbf{x}, \mathbf{y} >= \mathbf{x}'\mathbf{y} = \sum_{i=1}^{n} x_i y_i$$

A vector space which has an inner product defined for every pair of vectors is called an inner product space.

Inner product

Properties

For points in **V**, the inner product satisfies:

$$\begin{aligned} < \mathbf{x}, \mathbf{y} > &= < \mathbf{y}, \mathbf{x} > \\ < a\mathbf{x}, \mathbf{y} > &= a < \mathbf{x}, \mathbf{y} > \\ < \mathbf{x}_1 + \mathbf{x}_2, \mathbf{y} > &= < \mathbf{x}_1, \mathbf{y} > + < \mathbf{x}_2, \mathbf{y} > \end{aligned}$$

Inner product

Definitions

- The norm $||\mathbf{x}|| = \sqrt{\mathbf{x}'\mathbf{x}}$ gives the length of a vector.
- The distance between $\mathbf{x}$ and $\mathbf{y}$ is given by $||\mathbf{x} - \mathbf{y}||$.
- Two vectors $\mathbf{x}$ and $\mathbf{y}$ are orthogonal if $\mathbf{x}'\mathbf{y} = 0$.
- The angle between two vectors $\mathbf{x}$ and $\mathbf{y}$ is given by:

$$cos(\theta) = \frac{< \mathbf{x}, \mathbf{y} >}{||\mathbf{x}|| \cdot ||\mathbf{y}||}$$

Pythagorean theorem

Theorem

Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be pairwise orthogonal vectors in a Euclidean vector space. Then,

$$||\mathbf{x}_1 + \mathbf{x}_2 + \ldots + \mathbf{x}_n||^2 = ||\mathbf{x}_1||^2 + ||\mathbf{x}_2||^2 + \ldots + ||\mathbf{x}_n||^2$$

Projection

Definition

The projection of a vector $\mathbf{y}$ on a vector $\mathbf{x}$ is the vector $\hat{\mathbf{y}}$ such that

- $\hat{\mathbf{y}} = b\mathbf{x}$ for some constant b.
- $(\mathbf{y} - \hat{\mathbf{y}})$ is orthogonal to $\mathbf{x}$ (or $< \hat{\mathbf{y}}, \mathbf{x} > = < \mathbf{y}, \mathbf{x} >$).

Projection

For two vectors $\mathbf{x}$ and $\mathbf{y}$, the projection of $\mathbf{y}$ onto $\mathbf{x}$ is given by:

$$\hat{\mathbf{y}} = \frac{\mathbf{x}'\mathbf{y}}{\mathbf{x}'\mathbf{x}}\mathbf{x}.$$

Example

- Let $\mathbf{J}_n$ be a vector of ones and $\mathbf{y} = (y_1, y_2, \ldots y_n)'$
- Then, $\hat{\mathbf{y}} = \bar{y}\mathbf{J}_n$.

Projection

Theorem

Among all multiples $a\mathbf{x}$ of $\mathbf{x}$, the projection $\hat{\mathbf{y}}$ of $\mathbf{y}$ on $\mathbf{x}$ is the closest vector to $\mathbf{y}$.

Mean only regression

- Consider the following model:

$$y_i = \mu + \epsilon_i \qquad \text{for } \; i = 1, \dots n.$$

- We seek to minimize the function $f(\mu) = \sum_{i=1}^{n}(y_i - \mu)^2$ with respect to μ.

Mean only regression

- Alternatively we can write the model as $\mathbf{y} = \mathbf{J}_n\mu + \epsilon$.
- Under this formulation we seek to minimize $f(\mu) = ||\mathbf{y} - \mathbf{J}_n\mu||^2$ with respect to μ.
- To do so, we begin by rewriting f as follows:

$$\begin{aligned} f(\mu) &= (\mathbf{y} - \mathbf{J}_n\mu)'(\mathbf{y} - \mathbf{J}_n\mu) \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{J}_n'\mathbf{y}\mu + \mathbf{J}_n'\mathbf{J}_n\mu^2 \\ &= \mathbf{y}'\mathbf{y} - 2n\bar{y}\mu + n\mu^2 \end{aligned}$$

Mean only regression

- Taking derivatives of f with respect to μ we obtain:

$$\frac{df}{d\mu} = -2n\bar{y} + 2n\mu.$$

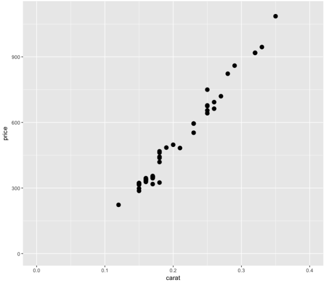
- This has a root at $\hat{\mu} = \bar{y}$.
- Note that the second derivative is $2n > 0$.
- The average is the least squares estimate in the sense of minimizing the Euclidean distance between the observed data and a constant vector.

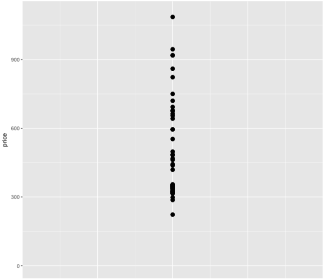
Mean only regression - Geometric interpretation

- Alternatively, we can think of this as projecting our n dimensional data onto the one dimensional subspace spanned by $\mathbf{J}_n$.
- Recall the projection is $\hat{\mathbf{y}} = \bar{y}\mathbf{J}_n$

Diamond data

The `diamond` dataset consists of data on 48 diamond rings containing price in Singapore dollars and size of diamond in carats.





R code

```
> library(UsingR); data(diamond)
> y = diamond$price; x = diamond$carat
> mean(y)
[1] 500.0833
> #using least squares
> coef(lm(y ~ 1))
[1] 500.0833
```

As expected the mean only least squares estimate obtained via `lm` is the empirical mean.

Regression through the origin

- Now consider the regression through the origin problem.
- Consider the following model:

$$y_i = \beta x_i + \epsilon_i \qquad \text{for } i = 1, \ldots n.$$

- We seek to minimize the function $f(\mu) = \sum_{i=1}^{n}(y_i - \beta x_i)^2$ with respect to β.

Regression through the origin

- Note that the pairs (x_i, y_i) form a scatterplot.
- Least squares involves finding the best fitting line of the form $y = \beta_1 x$ by minimizing the sum of the squared vertical distances between the points and the fitted line.
- Note we only consider lines going through the origin.

Regression through the origin

- Let $\mathbf{x} = (x_1, \ldots x_n)'$ be a vector.
- We can write the model: $\mathbf{y} = \mathbf{x}\beta + \epsilon$.
- We seek to minimize $f(\beta) = ||\mathbf{y} - \mathbf{x}\beta||^2$ with respect to β.
- We can re-write the least-squares criteria as follows:

$$f(\beta) = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{x}\beta + \mathbf{x}'\mathbf{x}\beta^2.$$

Regression through the origin

- Taking derivatives with respect to β we obtain:

$$\frac{df}{d\beta} = -2\mathbf{y}'\mathbf{x} + 2\mathbf{x}'\mathbf{x}\beta.$$

- Setting this equal to zero we obtain the solution:

$$\hat{\beta} = \frac{\mathbf{y}'\mathbf{x}}{\mathbf{x}'\mathbf{x}} = \frac{<\mathbf{y},\mathbf{x}>}{<\mathbf{x},\mathbf{x}>} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

Regression through the origin

- Note that the second derivative is $2\mathbf{x}'\mathbf{x} > 0$.
- Thus, $\hat{\beta}$ minimizes the least-squares criteria.

Geometric interpretation

- Alternatively, we can think of this as projecting our *n* dimensional data onto the one dimensional subspace spanned by the single vector $\mathbf{x}$, i.e. $\{\beta\mathbf{x} | \beta \in \mathbb{R}\}$.
- Here the projection given by

$$\hat{\mathbf{y}} = \frac{\mathbf{x}'\mathbf{y}}{\mathbf{x}'\mathbf{x}}\mathbf{x}.$$

R code

Continuing with the diamond example.

```
> yc = y - mean(y)
> xc = x - mean(x)
> sum(yc * xc) / sum(xc * xc)
[1] 3721.025
> coef(lm(yc ~ xc - 1))
      xc
3721.025
```