Linear models

Rasmus Waagepetersen
Department of Mathematics
Aalborg University
Denmark

October 5, 2016

Outline for today

- linear models
- least squares estimation
- orthogonal projections
- estimation of error variance

Linear regression

Model for random responses Y_i given fixed x_i:

$$Y_i = a + bx_i + \epsilon_i \tag{1}$$

a, b: intercept and slope.

ϵ_i: zero-mean random noise/measurement error/model error.

Matrix form:

$$Y = X\beta + \epsilon$$

where $\beta = (a, b)^\mathsf{T}$ and

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{bmatrix} \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

General linear model

General linear model:

$$Y = X\beta + \epsilon$$

where $Y \in \mathbb{R}^n$, X: $n \times p$ *design* matrix, $\beta \in \mathbb{R}^p$: regression parameter, $\epsilon \in \mathbb{R}^n$: zero-mean noise.

Objective: estimate unknown parameter β and quantify noise variation based on observation y of Y.

NB: $\mu = \mathbb{E}Y$ can be any vector in $L = \text{col}X$ (column space of X).

Instead of specifying the design matrix we may just specify that $\mu \in L$ for some linear subspace L. Then we can use any design matrix X for which $\text{col}X = L$ (that is the columns of X spans L).

Note columns need not be linear independent but if they are then one-to-one correspondence between μ and β !

Example (linear regression): let $\bar{x} = \sum_i x_i / n$. Then

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{bmatrix} \text{ and } \tilde{X} = \begin{bmatrix} 1 & x_1 - \bar{x} \\ 1 & x_2 - \bar{x} \\ \vdots & \\ 1 & x_n - \bar{x} \end{bmatrix}$$

generate the same linear subspace.

However, the columns of $\tilde{X}$ form an orthogonal basis.

Least squares

Suppose linear model is specified by $\mu \in L$. Least squares estimate of μ is

$$\hat{\mu} = \underset{\mu \in L}{\operatorname{argmin}} \|Y - \mu\|^2 = \underset{\mu \in L}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \mu_i)^2$$

If $L = \text{col}X$ where X has full rank this is equivalent to finding

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|Y - X\beta\|^2$$

(unique minimum not available if X not full rank)

Estimation handled easily via orthogonal projections.

Orthogonal decomposition

Suppose L subspace of $\mathbb{R}^n$. Let
$L^\perp = \{v \in \mathbb{R}^n | v \bullet w = 0 \text{ for all } w \in L\}$.

Orthogonal decomposition: each $x \in \mathbb{R}^n$ has a unique decomposition

$$x = u + v$$

where $u \in L$ and $v \in L^\perp$.

Orthogonal projection: u and v above are the orthogonal projections $p_L(x)$ and $p_{L^\perp}(x)$ of x on respectively L and $L^\perp$.

Pythagoras:

$$\|x\|^2 = \|u\|^2 + \|v\|^2$$

Least squares and orthogonal projections

Orthogonal decomposition of data vector Y:

$$Y = p_L(Y) + R$$

where $p_L(Y) \in L$ and $R = Y - p_L(Y) \in L^{\perp}$.

By Pythagoras:

$$\|Y - \mu\|^2 = \|(Y - p_L(Y)) + (p_L(Y) - \mu)\|^2 = \|R\|^2 + \|p_L(Y) - \mu\|^2$$

It thus follows that $\hat{\mu} = p_L(Y)$

We call R the residual

Orthogonal projections

- the orthogonal projection $p_L : \mathbb{R}^n \to L$ on L is a linear mapping. It is thus given by a unique matrix-transformation $p_L(x) = Px$ where P is an $n \times n$ matrix.
- the projection matrix P is symmetric ($P^\mathsf{T} = P$) and idempotent ($P^2 = P$)
- conversely, if a matrix Q is symmetric, idempotent and $L = \mathrm{col}Q$ then Q is the matrix of the orthogonal projection on L.
- if $L = \mathrm{col}X$ and X full rank then $P = X(X^\mathsf{T}X)^{-1}X^\mathsf{T}$ (whereby $\hat{\mu} = X(X^\mathsf{T}X)^{-1}X^\mathsf{T}Y$ and $\hat{\beta} = (X^\mathsf{T}X)^{-1}X^\mathsf{T}Y$)

Note: $\mathbb{E}\hat{\mu} = \mu$ and $\mathbb{E}\hat{\beta} = \beta$ i.e. $\hat{\mu}$ and $\hat{\beta}$ are unbiased.

Implementation in R

Suppose L is spanned by vectors x_i and y is the observed response.

The following `R` call ($p = 3$) fits the linear model:

```
lm(y~x1+x2+x3)
```

Useful methods: `coef`, `resid`, `summary`.

Note: if e.g. x_1 is a categorical variable/factor we typically want one intercept for each category/level of the factor. Then we use:

```
lm(y~factor(x1)+x2+x3)
```

Let $Z = (Z_1, \dots, Z_n)^\mathsf{T}$ where the Z_i are random variables with finite variances.

Then the *variance-covariance* matrix of Z is the $n \times n$ matrix $\mathbb{V}\mathrm{ar} Z$ whose *ij*th entry is $\mathbb{C}\mathrm{ov}(Z_i, Z_j)$.

$\mathbb{V}\mathrm{ar} Z$ is symmetric, positive semi-definite and the diagonal entries are given by the variances of the Z_i.

Variance of $\hat{\beta}$

Suppose the errors $\epsilon_1, \ldots, \epsilon_n$ are uncorrelated with common variance σ^2. I.e.

$$\mathbb{V}\mathrm{ar}\epsilon = \sigma^2 I$$

Then

$$\begin{aligned}\mathbb{V}\mathrm{ar}\hat{\beta} &= \mathbb{V}\mathrm{ar}((X^\mathsf{T} X)^{-1} X^\mathsf{T} Y) \\ &= (X^\mathsf{T} X)^{-1} X^\mathsf{T} \mathbb{V}\mathrm{ar}\epsilon [(X^\mathsf{T} X)^{-1} X^\mathsf{T}]^\mathsf{T} \\ &= \sigma^2 (X^\mathsf{T} X)^{-1}\end{aligned}$$

One can show that $\hat{\beta}$ is BLUE (best linear unbiased estimate) - i.e. it attains minimum variance among all linear unbiased estimates (we will return to this in fourth lecture).

Estimation of σ^2

Suppose that $\mathbb{V}\mathrm{ar}\epsilon = \sigma^2 I$.

Note that

$$R = (Y - \hat{\mu}) = (I - P)Y = (I - P)\epsilon$$

Hence $\mathbb{V}\mathrm{ar}R = \sigma^2(I - P)$. Moreover,

$$\mathbb{E}\|R\|^2 = \sigma^2 \mathrm{trace}(I - P) = \sigma^2(n - p)$$

Thus

$$\hat{\sigma}^2 = \|R\|^2/(n - p)$$

is an unbiased estimate of σ^2.

Exercises

1. (slide 5) Show that the columns in $\tilde{X}$ are orthogonal. Show that $\text{col}X = \text{col}\tilde{X}$.
2. consider the model $y_{ij} = \alpha + \alpha_i + \epsilon_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, k$.
   2.1 is this a linear model ?
   2.2 if so, then write down the design matrix.
   2.3 does the design matrix have full rank ?
   2.4 in case of no to the previous question, can you find an alternative design matrix of full rank ?
3. show that μ and β are in one-to-one correspondence when X has full rank (hint: $(X^\mathsf{T}X)^{-1}X^\mathsf{T}$ is a left inverse to X).
4. (Linear regression): Consider the linear regression model (1). Find the parameter estimates $\hat{a}$ and $\hat{b}$ using
   4.1 differentiation of the least squares criterion.
   4.2 Orthogonal projection in terms of the 'orthogonalized' design matrix where the second column has entries $x_i - \bar{x}$.
   4.3 You may also try orthogonal projection in terms of the original design matrix.

More exercises

5. Consider the insulgas data set on the website. Use R to
   5.1 fit a linear regression model with gas consumption as dependent variable and temperature as independent variable.
   5.2 extend the previous model so that there are different intercepts depending on whether the house is insulated or not.
6. Find the unbiased estimate of the variance σ^2 for a linear regression with $\mathbb{V}\text{ar}\epsilon = \sigma^2 I$.
7. If $Z \in \mathbb{R}^n$ has covariance matrix Σ then show that AZ has covariance matrix $A\Sigma A^\mathsf{T}$. Use this result to show that Σ is positive semi-definite (hint: compute the variance of $a^\mathsf{T} Z$ for an arbitrary vector $a \in \mathbb{R}^n$.)
8. Show that if Z is zero-mean with covariance matrix Σ then $\mathbb{E}\|Z\|^2 = \text{trace}\Sigma$ (sum of diagonal elements of Σ).

More exercises

9. (trace of an orthogonal matrix)
   9.1 Show that the eigen-values of an orthogonal projection matrix P on a p dimensional subspace L are either 1 or 0. What is the multiplicity of the 1's ?
   9.2 Use the previous exercise to show that trace$P = p$ (hint use spectral decomposition of P and the fact that trace$AB =$ traceBA)