Chapter 1

DJM

27 January, 2017

The normal linear model

Assume that

$$y_i = x_i^\top \beta + \epsilon_i.$$

1. What are all these things?
2. What is the mean of y_i?
3. What is the distribution of ϵ_i?
4. What is the notation X or Y?

Drawing a sample

$$y_i = x_i^\top \beta + \epsilon_i.$$

Write code which draws a sample form the population given by this model.

```
p = 3
n = 100
sigma = 2
epsilon = rnorm(n,sd=sigma) # this is random
X = matrix(runif(n*p), n, p) # treat this as fixed, but I need numbers
beta = rpois(p+1,5) # also fixed, but I again need numbers
Y = cbind(1,X) %*% beta + epsilon # epsilon is random, so this is
## Equiv: Y = beta[1] + X %*% beta[-1] + epsilon
```

How do we estimate beta?

1. Guess.
2. Ordinary least squares (OLS).
3. Maximum likelihood.
4. Do something more creative.

Method 1: Guess

This method isn't very good, as I'm sure you can imagine.

Method 2. OLS

Suppose I want to find an estimator $\widehat{\beta}$ which makes small errors on my data.

I measure errors with the difference between predictions $X\widehat{\beta}$ and the responses Y.

I don't care if the differences are positive or negative, so I try to measure the total error with

$$\sum_{i=1}^{n} \left| y_i - x_i^\top \widehat{\beta} \right|.$$

This is fine, but hard to minimize (what is the derivative of $|\cdot|$?)

So I use

$$\sum_{i=1}^{n}(y_i - x_i^\top \widehat{\beta})^2.$$

Method 2. OLS solution

We write this as

$$\widehat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n}(y_i - x_i^\top \beta)^2.$$

"Find the β which minimizes the sum of squared errors."

Note that this is the same as

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^\top \beta)^2.$$

"Find the beta which minimizes the mean squared error."

Method 2. Ok, do it

We differentiate and set to zero

$$\begin{aligned}
&\frac{\partial}{\partial\beta}\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^\top\beta)^2 \\
&= \frac{2}{n}\sum_{i=1}^{n} x_i(y_i - x_i^\top\beta) \\
&= \frac{2}{n}\sum_{i=1}^{n} -x_i x_i^\top\beta + x_i y_i \\
0 &\equiv \sum_{i=1}^{n} -x_i x_i^\top\beta + x_i y_i \\
&\Rightarrow \sum_{i=1}^{n} x_i x_i^\top\beta = \sum_{i=1}^{n} x_i y_i \\
&\Rightarrow \beta = \left(\sum_{i=1}^{n} x_i x_i^\top\right)^{-1}\sum_{i=1}^{n} x_i y_i
\end{aligned}$$

In matrix notation...

...this is

$$\hat{\beta} = (X^\top X)^{-1}X^\top Y.$$

The β which "minimizes the sum of squared errors"

AKA, the SSE.

Method 3: maximum likelihood

Method 2 didn't use anything about the distribution of ϵ.

But if we know that ϵ has a normal distribution, we can write down the joint distribution of $Y = (y_1, \ldots, y_n)$:

$$\begin{aligned} f_Y(y;\beta) &= \prod_{i=1}^{n} f_{y_i;\beta}(y_i) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - x_i^\top \beta)^2\right) \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^\top \beta)^2\right) \end{aligned}$$

In M463, we think of f_Y as a function of y with β fixed:

1. If we integrate over y from $-\infty$ to ∞, it's 1.
2. If we want the probability of (a, b), we integrate from a to b.
3. etc.

Turn it around...

...instead, think of it as a function of β.

We call this "the likelihood" of beta: $\mathcal{L}(\beta)$.

Given some data, we can evaluate the likelihood for any value of β (assuming σ is known).

It won't integrate to 1 over β.

But it is "convex", meaning we can maximize it (the second derivative wrt β is everywhere negative).

So let's maximize

The derivative of this thing is kind of ugly.

But if we're trying to maximize over β, we can take an increasing transformation without changing anything.

I choose $\log_e$.

$$\begin{aligned} \mathcal{L}(\beta) &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^\top \beta)^2\right) \\ \ell(\beta) &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^\top \beta)^2 \end{aligned}$$

But we can ignore constants, so this gives

$$\widehat{\beta} = \arg\max_{\beta} -\sum_{i=1}^{n}(y_i - x_i^\top \beta)^2$$

The same as before!

The here and now

In S432, we focus on OLS.

In S420, you look at maximum likelihood (for this and many other distributions).

Here, the method gives the same estimator.

We need to be able to evaluate how good this estimator is however.

Mean squared error (MSE)

Let's look at the population version, and let's forget about the linear model.

Suppose we think that there is **some** function which relates y and x.

Let's call this function f for the moment.

How do we estimate f?

What is f?

Minimizing MSE

Let's try to minimize the **expected** sum of squared errors (MSE)

$$\begin{aligned}
\mathbb{E}\left[(Y - f(X))^2\right] &= \mathbb{E}\left[\mathbb{E}\left[(Y - f(X))^2 \mid X\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[(Y - f(X))^2 \mid X\right]\right] \\
&= \mathbb{E}\left[\mathrm{Var}\left[Y \mid X\right] + \mathbb{E}\left[(Y - f(X))^2 \mid X\right]\right] \\
&= \mathbb{E}\left[\mathrm{Var}\left[Y \mid X\right]\right] + \mathbb{E}\left[\mathbb{E}\left[(Y - f(X))^2 \mid X\right]\right]
\end{aligned}$$

The first part doesn't depend on f, it's constant, and we toss it.

To minimize the rest, take derivatives and set to 0.

$$\begin{aligned}
0 &= \frac{\partial}{\partial f}\mathbb{E}\left[\mathbb{E}\left[(Y - f(X))^2 \mid X\right]\right] \\
&= -\mathbb{E}\left[\mathbb{E}\left[2(Y - f(X) \mid X\right]\right] \\
&\Rightarrow 2\mathbb{E}\left[f(X) \mid X\right] = 2\mathbb{E}\left[Y \mid X\right] \\
&\Rightarrow f(X) = \mathbb{E}\left[Y \mid X\right]
\end{aligned}$$

The regression function

We call this solution:

$$\mu(X) = \mathbb{E}\left[Y \mid X\right]$$

the regression function.

If we **assume** that $\mu(x) = \mathbb{E}\left[Y \mid X = x\right] = x^\top \beta$, then we get back exactly OLS.

But why should we assume $\mu(x) = x^\top \beta$?

The regression function

In mathematics: $\mu(x) = \mathbb{E}[Y \mid X = x]$.

In words: Regression is really about estimating the mean. 1. If $Y \sim \mathrm{N}(\mu, 1)$, our best guess for a **new** Y is μ 2. For regression, we let the mean (μ) **depend** on X. 3. Think of $Y \sim \mathrm{N}(\mu(X), 1)$, then conditional on $X = x$, our best guess for a **new** Y is $\mu(x)$ [whatever this function μ is]

Causality

For any two variables Y and X, we can **always** write

$$Y \mid X = \mu(X) + \eta(X)$$

such that $\mathbb{E}[\eta(X)] = 0$.

- Suppose, $\mu(X) = \mu_0$ (constant in X), are Y and X independent?
- Suppose Y and X are independent, is $\mu(X) = \mu_0$?

Previews of future chapters

Linear smoothers

What is a linear smoother?

1. Suppose I observe $Y_1, \ldots, Y_n$.
2. A linear smoother is any **prediction function** that's linear in $\mathbf{Y}$.

- Linear functions of $\mathbf{Y}$ are simply premultiplications by a matrix, i.e. $\hat{\mathbf{Y}} = \mathbf{W}\mathbf{Y}$ for any matrix $\mathbf{W}$.

3. Examples:

- $\overline{Y} = \frac{1}{n}\sum Y_i = \frac{1}{n}\begin{bmatrix}1 & 1 & \cdots & 1\end{bmatrix}\mathbf{Y}$
- Given X, $\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$
- You will see many other smoothers in this class

kNN as a linear smoother

(We will see **smoothers** in more detail in Ch. 4)

1. For kNN, consider a particular pair (Y_i, X_i)
2. Find the k covariates X_j which are closest to X_i
3. Predict Y_i with the average of those X_j's
4. This turns out to be a linear smoother

- How would you specify $\mathbf{W}$?

Kernels

(Again, more info in Ch. 4)

- There are two definitions of "kernels". We'll use only 1.
- Recall the pdf for the Normal density:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

- The part that depends on the data (x), is a kernel
- The kernel has a *center* (μ) and a *range* (σ)

Kernels (part 2)

- In general, any function which integrates, is non-negative, and symmetric is a kernel in the sense used in the book
- You can think of any (unnormalized) symmetric density function (uniform, normal, Cauchy, etc.)
- The way you use a kernel is take a weighted average of nearby data to make predictions
- The weight of X_j is given by the height of the density centered at X_i
- Examples:
- The **Gaussian** kernel is $K(x - x_0) = e^{-(x-x_0)^2/2}$
- The **Boxcar** kernel is $K(x - x_0) = I(x - x_0 < 1)$

Kernels (part 3)

- You don't need the normalizing constant
- To alter the **support**: take $(x - x_0)/h$ and $K(z) = K(z)/h$
- Now, the range of the density is determined by h
- You can interpret kNN as a particular kind of kernel
- The range is determined by k
- The center is determined by X_i