Lecture 02
Simple Linear Models: OLS

04 September 2015

Taylor B. Arnold
Yale Statistics
STAT 312/612

Office Hours

- Taylor Arnold:
  - 24 Hillhouse, Office # 206
  - Wednesdays 13:30-15:00, or by appointment
  - Short one-on-one meetings (or small groups)
- Jason Klusowski:
  - 24 Hillhouse, Main Classroom
  - Tuesdays 19:00-20:30
  - Group Q&A style

Website

http://euler.stat.yale.edu/~tba3/stat612

Goals for today

1. calculate the MLE for simple linear regression
2. derive basic properties of the simple linear model MLE
3. introduction to R for simulations and data analysis

Simple Linear Models: MLEs

Considering observing n samples from a simple linear model with only a single unknown slope parameter $\beta \in \mathbb{R}$,

$$y_i = x_i\beta + \epsilon_i, \quad i = 1, \ldots n.$$

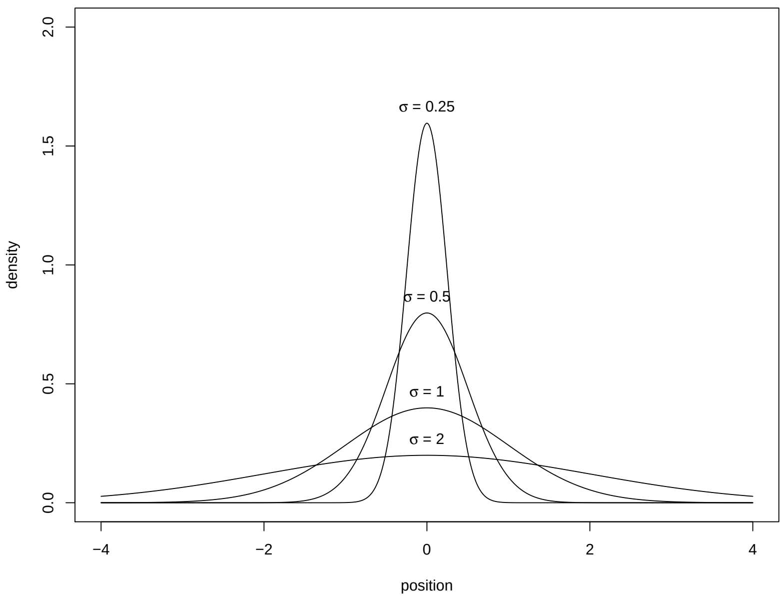
This is, perhaps, the simpliest linear model.

For today, we will assume that the x_i's are fixed and known quantities. This is called a **fixed design**, compared to a **random design**.

The error terms are assumed to be independent and identically distributed random variables with a normal density function:

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

For some unknown variance $\sigma^2 > 0$.


σ = 0.25
σ = 0.5
σ = 1
σ = 2
density
position
0.0
0.5
1.0
1.5
2.0
−4
−2
0
2
4

The density function of a normally distributed random variable with mean μ and variance σ^2 is given by:

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

Conceptually, the front term is just a normalization to make the density sum to 1. The important part is:

$$f(z) \propto exp\left\{-\frac{1}{2\sigma^2}(z-\mu)^2\right\}$$

Which you have probably seen rewritten as:

$$f(z) \propto exp\left\{-0.5 \cdot \left(\frac{z-\mu}{\sigma}\right)^2\right\}$$

Let's look at the maximum likelihood function of our simple regression model:

$$\begin{aligned}\mathcal{L}(\beta, \sigma | x, y) &= \prod_i \mathcal{L}(\beta, \sigma | x_i, y_i) \\ &= \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \times exp\left\{-\frac{1}{2\sigma^2}(y_i - \beta x_i)^2\right\}\end{aligned}$$

Notice that the mean μ from the general case has been replaced by βx_i, which should be the mean of $y_i | x_i$.

We can bring the product up into the the exponent as a sum:

$$\begin{aligned}\mathcal{L}(\beta, \sigma | x, y) &= \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \times exp\left\{-\frac{1}{2\sigma^2}(y_i - \beta x_i)^2\right\} \\ &= (2\pi\sigma^2)^{-n/2} \times exp\left\{-\frac{1}{2\sigma^2} \cdot \sum_i (y_i - \beta x_i)^2\right\}\end{aligned}$$

Let's highlight the slope parameter β:

$$\mathcal{L}(\beta, \sigma | x, y) \quad = \quad (2\pi\sigma^2)^{-n/2} \times exp\left\{ -\frac{1}{2\sigma^2} \cdot \sum_i (y_i - \beta x_i)^2 \right\}$$

What is the MLE for β?

Without resorting to any fancy math, we can see that:

$$\widehat{\beta}_{MLE} = \underset{b \in \mathbb{R}}{\arg\min} \left\{ \sum_i (y_i - b \cdot x_i)^2 \right\} \tag{1}$$

The least squares estimator.

A slightly more 'mathy' approach would be to calculate the the negative log-likelihood:

$$-\log\left\{\mathcal{L}(\beta, \sigma | x, y)\right\} = \frac{n}{2} \cdot \log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_i (y_i - \beta x_i)^2$$

Now the minimum of this corrisponds with the maximum likelihood estimators.

Again, we notice that only the second term depends on β:

$$-\log\left\{\mathcal{L}(\beta,\sigma|x,y)\right\} = \frac{n}{2}\cdot\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_i (y_i - \beta x_i)^2$$

And we can again see without resorting to derivatives that the maximum likelihood estimator is that one that minimizes the sum of squares:

$$\widehat{\beta}_{mle} = \operatorname*{arg\,min}_{b\in\mathbb{R}} \left\{\sum_i (y_i - bx_i)^2\right\}$$

It is possible to directly solve the least squares and obtain an analytic solution to the simple linear regression model.

Taking the derivative of the sum of squares with respect to β we get:

$$
\begin{aligned}
\frac{\partial}{\partial\beta}\sum_i (y_i - \beta x_i)^2 &= -2 \cdot \textstyle\sum_i (y_i - \beta x_i) \cdot x_i \\
&= -2 \cdot \textstyle\sum_i (y_i x_i - \beta x_i^2)
\end{aligned}
$$

Setting the derivative equal to zero:

$$\begin{aligned}
-2 \cdot \sum_i (y_i x_i - \widehat{\beta} x_i^2) &= 0 \\
\sum_i y_i x_i &= \widehat{\beta} \sum_i x_i^2 \\
\widehat{\beta}_{MLE} &= \frac{\sum_i y_i x_i}{\sum_i x_i^2}
\end{aligned}$$

If you have seen the standard simple least squares solution (that is, with an intercept) this should look familiar.

There are many ways of thinking about the maximum likelihood estimator, one of which is as a weighted sum of the data points y_i:

$$
\begin{aligned}
\widehat{\beta} &= \frac{\sum_i y_i x_i}{\sum_i x_i^2} \\
&= \sum_i \left(y_i \cdot \frac{x_i}{\sum_j x_j^2} \right) \\
&= \sum_i y_i w_i
\end{aligned}
$$

One thing that the weighted form of the estimator makes obvious is that the estimator is distributed normally:

$$\widehat{\beta} \sim \mathcal{N}(\cdot, \cdot)$$

As it is the sum of normally distributed variables (y_i).

The mean of the estimator becomes

$$
\begin{aligned}
\mathbb{E}\widehat{\beta} &= \sum_i \mathbb{E}(y_i w_i) \\
&= \sum_i w_i \cdot \mathbb{E}(y_i) \\
&= \sum_i w_i x_i \beta \\
&= \beta \cdot \sum_i x_i \frac{x_i}{\sum_j x_j^2} \\
&= \beta \cdot \frac{\sum_i x_i^2}{\sum_j x_j^2} \\
&= \beta
\end{aligned}
$$

And so we see the estimator is unbiased.

A normally distributed random variable is entirely characterised by its mean and variance. So let us compute the variance of our MLE estimator:

$$
\begin{aligned}
\mathbb{V}\widehat{\beta} &= \sum_i \mathbb{V}(y_i w_i) \\
&= \sum_i w_i^2 \mathbb{V}(y_i) \\
&= \sum_i w_i^2 \sigma^2 \\
&= \sigma^2 \cdot \frac{\sum_i x_i^2}{\left(\sum_i x_i^2\right)^2} \\
&= \frac{\sigma^2}{\sum_i x_i^2}
\end{aligned}
$$

If $\sum_i x_i^2$ diverges, we will get a consistent estimator.

So we now know that the MLE is distributed as:

$$\widehat{\beta} \sim \mathcal{N}(\beta, \frac{\sigma^2}{\sum_i x_i^2})$$

Does this make sense? What if $x_i = 1$ for all i?

The MLE is weighting the data y_i according to:

$$w_i \propto x_i$$

Does this make sense? **Why**?

Simulations

We will be using the R programming language for data analysis and simulations

- Open source software, available at: https://www.r-project.org/
- An implementation of the S programming language
- Designed for interactive data analysis
- For pros/cons, check out the many lengthy internet articles & arguments

Things to know

- everything in R is an object
- indexing starts at 1
- no scalar type, numeric objects are all vectors
- mostly functional language, with some OO features

Gauß-Markov theorem

Many of the nice properties of the MLE estimator result from being unbiased and normally distributed. A natural question is whether another weighted sum of the data points y_i would yield a better estimator.

Formally, if we define:

$$\widehat{\beta}_{BLUE} = \sum_{i} y_i \cdot a_i$$

What values of a_i will minimise the variance of the estimator assuming that we force it to be unbiased? BLUE stands for the Best Linear Unbiased Estimator.

To force unbiaseness, we must have:

$$\begin{aligned} \mathbb{E} \sum_i y_i \cdot a_i &= \beta \\ \sum_i x_i \cdot \beta \cdot a_i &= \beta \\ \sum_i x_i \cdot a_i &= 1 \end{aligned}$$

The variance is given by:

$$\begin{aligned}\mathbb{V}\sum_i y_i \cdot a_i &= \sum_i a_i^2 \cdot \mathbb{V} y_i \\ &= \sum_i a_i^2 \cdot \sigma^2\end{aligned}$$

As we cannot change σ^2, minimising the variance amounts to minimising $\sum_i a_i^2$.

So we have reduced the problem to solving the following:

$$\operatorname*{arg\,min}_{a \in \mathbb{R}^n} \left\{ \sum_i a_i^2 \quad \text{s.t.} \quad \sum_i a_i x_i = 1 \right\}$$

Lagrange multiplier

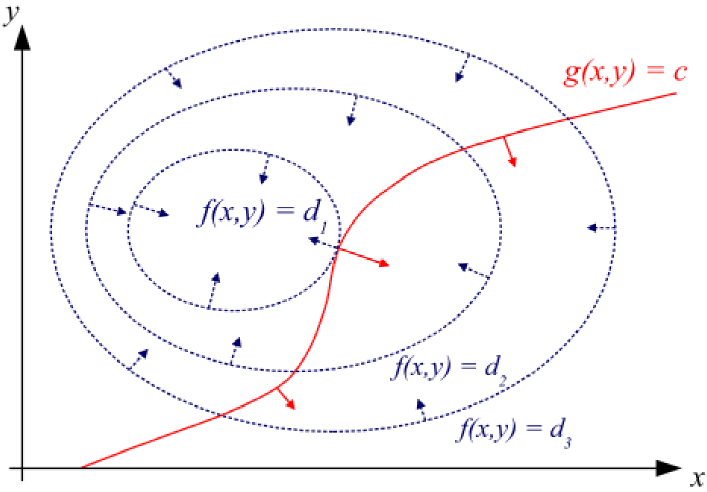
To solve the constrained problem:

$$\underset{x \in \mathbb{R}^p}{\arg\min} \{f(x) \quad \text{s.t.} \quad g(x) = k\}$$

Find stationary points (zero partial derivates) of:

$$L(x, \lambda) = f(x) + \lambda \cdot (g(x) - k)$$

This will give the set of possible minimisers to the original constrained problem.



y
g(x,y) = c
f(x,y) = d1
f(x,y) = d2
f(x,y) = d3
x

For our problem we have:

$$L(a, \lambda) = \sum_i a_i^2 + \lambda \cdot \left(1 - \sum_i a_i x_i\right)$$

Which gives:

$$\begin{aligned}
\frac{\partial}{\partial a_k} L(a, \lambda) &= 2a_k - \lambda x_k \\
2a_k - \lambda x_k &= 0 \\
a_k &= \frac{1}{2} \cdot \lambda \cdot x_k
\end{aligned}$$

The lambda derivative, which is just the constraint, shows the specific value of λ that we need:

$$\frac{\partial}{\partial \lambda} L(a, \lambda) = 1 - \sum_i a_i x_i$$

$$\sum_i a_i x_i = 1$$

Plugging our previous version of a_i:

$$\sum_i \frac{1}{2} \cdot \lambda \cdot x_i \cdot x_i = 1$$

$$\lambda \cdot \sum_i \frac{1}{2} \cdot x_i^2 = 1$$

$$\lambda = \frac{2}{\sum_i x_i^2}$$

Finally, plugging this back in:

$$a_k = \frac{1}{2} \cdot \lambda \cdot x_k$$
$$a_k = \frac{x_k}{\sum_i x_i^2}$$

And this gives:

$$\widehat{\beta}_{BLUE} = \sum_i y_i \cdot \frac{x_i}{\sum_j x_j^2}$$
$$= \widehat{\beta}_{MLE}$$

The MLE estimator has the following properties under our assumptions:

- unbiased
- consistent as long as $\sum_i x_i^2$ diverges
- normally distributed
- is the BLUE estimator
- achieves the Cramér–Rao bound (problem set)
- has an analytic solution

The more common formulation of simple linear models includes an unknown intercept term α. The basic model is then:

$$y_i = \alpha + x_i\beta + \epsilon_i, \quad i = 1, \ldots n.$$

The likelihood function for this revised model is almost the same as before

$$\mathcal{L}(\beta, \sigma | x, y) \quad = \quad (2\pi\sigma^2)^{-n/2} \times exp\left\{-\frac{1}{2\sigma^2} \cdot \sum_i (y_i - \alpha - x_i\beta)^2\right\}$$

Clearly, by the same logic the MLE is given by minimizing the sum of squared residuals.

Solving the least squares problem is only slightly more difficult because now we have two parameters and need to use partial derivatives to solve them. Otherwise the process is the same with a few more terms floating around.

The estimators in this case become:

$$\widehat{\beta} = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$
$$\widehat{\alpha} = \bar{y} - \widehat{\beta}\bar{x}$$

Where $\bar{x} = n^{-1} \sum_i x_i$ and $\bar{y} = n^{-1} \sum_i y_i$.

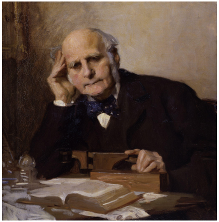
Notice what happens when both means are zero.

All of these properties are maintained jointly for $(\widehat{\alpha}, \widehat{\beta})$

- unbiased
- consistent as long as $\sum_i (x_i - \bar{x})^2$ diverges
- normally distributed
- is the BLUE estimator
- achieves the Cramér–Rao bound
- has an analytic solution

Applications

Sir Francis Galton & Regression



- ‘Co-relations and their measurement, chiefly from anthropometric data’ (1888).
- further ideas in *Natural Inheritance*
    - sweet peas and regression to the mean
    - extinction of surnames (Galton–Watson stochastic processes)
    - ‘Good and Bad Temper in English Families’