Why do "least squares" regression?

Kevin Penner

1 Introduction

- Instead of minimizing the sum of the squared errors, why not the sum of the absolute value of the errors or the sum of the reciprocal of the errors?
- One justification: least squares estimates of coefficients are also *maximum likelihood estimates.*
- **Maximum likelihood method:**
  1. What want to do: Look at sample data. Hypothesize the type of distribution that underlies the data. Choose parameter estimates to be the values for which the probability of getting the sample values is a maximum.
  2. How do this: A likelihood function $L(p)$ is a function that gives the likelihood a set of data is observed given the parameter vector p. Thus,

     $$L(p) = \text{Prob(observe data given value of parameters)}.$$

     We want to get the likelihood as high as possible, so we maximize $L(p)$ subject to p.

Example Given x heads in n coin flips, what is the maximum likelihood estimate of θ, the probability of heads?

1. Distribution underlying the data is the *binomial* distribution.
2. Given this distribution, need to maximize:

   $$L(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

   subject to θ.

3. For convenience, we can maximize $\ln[L(\theta)] = LL(\theta)$ instead of $L(\theta)$:[1]

$$\begin{aligned} LL(\theta) &= \ln\left[\binom{n}{x}\right] + x\ln[\theta] + (n-x)\ln[(1-\theta)] \\ \frac{\mathrm{d}LL(\theta)}{\mathrm{d}\theta} &= \frac{x}{\theta} - \frac{n-x}{1-\theta} = 0. \end{aligned} \tag{1}$$

4. So equation (1) implies $\theta = \frac{x}{n}$.

5. So if we get 3 heads in 10 tosses, the maximum likelihood estimate of the probability of heads is $\frac{3}{10}$.

2 Least Squares and Maximum Likelihood

First we need to make some assumptions about our data. Assume

- There are n data points (x_i, y_i).
- The data points are independent.
- The y_is can be modeled as a linear function of x_is, i.e. $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.
- The noise terms ϵ_i are normally distributed with mean 0 and variance σ^2.
- Randomness in y_i comes from noise, ϵ_i. Thus, at each fixed x_i, the corresponding y_i is normally distributed with mean $\beta_0 + \beta_1 x_i$ and variance σ^2. Using the formula for the normal distribution, we have:

$$\text{Prob}(y_i|x_i) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left[\frac{y_i - (\beta_0 + \beta_1 x_i)}{\sigma}\right]^2}.$$

- Interpret the term $y_i - (\beta_0 + \beta_1 x_i)$ as:

$$\underbrace{y_i}_{\text{actual}} - \underbrace{(\beta_0 + \beta_1 x_i)}_{\text{estimate}} = \underbrace{\epsilon_i}_{\text{noise}}$$

Now, we need to estimate the parameters $p = (\beta_0\ , \beta_1, \sigma)$ with our n data points. We use the log-likelihood function to do this. The probability of getting our set of data with β_0

[1]We can do this because $\ln(\bullet)$ is a monotonically increasing function, so $\ln(x)$ preserves the order of x. For example, assume we have x_1 and x_2 such that $x_1 \geq x_2$. Then, setting $x_1' = \ln(x_1)$ and $x_2' = \ln(x_2)$, we still have $x_1' \geq x_2'$.

and β_1 is

$$\begin{aligned} L(p) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left[\frac{y_i-(\beta_0+\beta_1 x_i)}{\sigma}\right]^2}, \text{ so} \\ LL(p) &= \sum_{i=1}^{n} \frac{-[y_i-(\beta_0+\beta_1 x_i)]^2}{2\sigma^2} + \sum_{i=1}^{n} \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) \\ &= \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} -[y_i-(\beta_0+\beta_1 x_i)]^2\right) - n\ln\left(\sigma\sqrt{2\pi}\right) \qquad (2) \end{aligned}$$

Maximizing equation (2) with respect to β_0 and β_1, we get

$$\begin{aligned} \frac{\mathrm{d}LL(p)}{\mathrm{d}\beta_0} &= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}[y_i-(\beta_0+\beta_1 x_i)]\right) = 0 \\ \frac{\mathrm{d}LL(p)}{\mathrm{d}\beta_1} &= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i[y_i-(\beta_0+\beta_1 x_i)]\right) = 0 \end{aligned}$$

These are the same expressions we get when minimizing the sum of the squared residuals

$$\sum_{i=1}^{n}[y_i-(\beta_0+\beta_1 x_i)]^2 \qquad (3)$$

with respect to β_0 and β_1. Thus, the maximum likelihood estimates of β_0 and β_1 are least-squares estimates!

As a final note, notice the negative sign in the sum in equation (2) - this is why *maximizing* the $LL(p)$ is the same as *minimizing* the RSS.