Maximum Entropy Estimation

To derive a maximum entropy estimator for b_0 and b_1 under noisy constraints, we'll follow these steps:

1. Define the problem with constraints:

Given the linear model:

$$y_i = b_0 + b_1 x_i + \epsilon_i$$

The assumptions are:

- $E[\epsilon_i] = 0$ (errors have zero mean)
- $E[\epsilon_i^2] = \sigma^2$ (errors have constant variance)
- $E[x_i \cdot \epsilon_i] = 0$ (errors are orthogonal to the covariates)

These assumptions hold perfectly in the ideal case. However, since the constraints are noisy, they may not hold exactly. Instead, we'll consider the constraints to be approximately true.

2. Set up the maximum entropy framework:

In the maximum entropy framework, we seek to find the probability distribution for the errors ϵ_i that maximizes the entropy subject to the given noisy constraints.

The entropy $H(p)$ of the distribution $p(\epsilon)$ is given by:

$$H(p) = -\int p(\epsilon) \log p(\epsilon)\, d\epsilon$$

3. Introduce Lagrange multipliers:

We introduce Lagrange multipliers to incorporate the noisy constraints into the maximization of entropy. The Lagrange function is:

$$\mathcal{L} = -\int p(\epsilon) \log p(\epsilon)\, d\epsilon + \lambda_1 (E[\epsilon_i]) + \lambda_2 \left(E[\epsilon_i^2] - \sigma^2\right) + \lambda_3 (E[x_i \cdot \epsilon_i])$$

Here, λ_1, λ_2, and λ_3 are Lagrange multipliers corresponding to the constraints.

4. Solve for the distribution $p(\epsilon)$:

To find the distribution $p(\epsilon)$, we take the functional derivative of $\mathcal{L}$ with respect to $p(\epsilon)$ and set it to zero:

$$\frac{\delta \mathcal{L}}{\delta p(\epsilon)} = -\log p(\epsilon) - 1 + \lambda_1 + \lambda_2 \epsilon + \lambda_3 x_i \cdot \epsilon = 0$$

Solving this, we get:

$$\log p(\epsilon) = -1 + \lambda_1 + \lambda_2\epsilon + \lambda_3 x_i \cdot \epsilon$$

Exponentiating both sides:

$$p(\epsilon) = C\exp(-\lambda_2\epsilon - \lambda_3 x_i \cdot \epsilon)$$

where C is a normalization constant.

5. Determine the form of $p(\epsilon)$:

The form of the distribution $p(\epsilon)$ depends on the values of the Lagrange multipliers:

$$p(\epsilon) = C\exp(-\lambda_2\epsilon - \lambda_3 x_i \cdot \epsilon)$$

For Gaussian errors (which is a common assumption in linear regression), λ_2 and λ_3 would be related to the mean and variance of the errors. Typically, this would correspond to a Gaussian distribution:

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

Here, the mean μ could be determined by the noisy constraints, and the variance σ^2 would correspond to the constraint $E[\epsilon_i^2] = \sigma^2$.

6. Estimate b_0 and b_1:

Given the distribution $p(\epsilon)$, the maximum entropy estimator for b_0 and b_1 can be found by maximizing the likelihood function $L(b_0, b_1)$ under the distribution $p(\epsilon)$.

The likelihood function is:

$$L(b_0, b_1) = \prod_{i=1}^{n} p(y_i - b_0 - b_1 x_i)$$

Taking the log-likelihood and maximizing with respect to b_0 and b_1:

$$\log L(b_0, b_1) = \sum_{i=1}^{n} \log p(y_i - b_0 - b_1 x_i)$$

Substituting $p(\epsilon)$:

$$\log L(b_0, b_1) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2 + \text{constant}$$

Maximizing this log-likelihood function yields the ordinary least squares (OLS) estimators for b_0 and b_1, which are:

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$
$$b_0 = \overline{y} - b_1\overline{x}$$

7. Conclusion:

Under the maximum entropy principle with noisy constraints, the estimators b_0 and b_1 derived are the same as those obtained by the ordinary least squares (OLS) method. The maximum entropy approach confirms that the OLS estimators are optimal under the assumptions of Gaussian errors, even when the constraints are noisy.