Paths to least squares.

Goals

- Describe different motivations for the least squares objective
  - Risk minimization
  - Normal maximum likelihood
  - Bayesian inference
  - Method-of-moment estimators

Reading

These notes are a supplement for the following readings:

- Hastie, Tibshirani, and Friedman (2009) Section 2.1 (risk minimization)
- Gelman, Hill, and Vehtari (2021) Section 8.1 (maximum likelihood and Bayesian methods)
- Davidson, MacKinnon, et al. (2004) Section 1.5 (method of moments)

Why least squares?

We have motivated the ordinary least squares (OLS) estimator

$$\widehat{\beta} := \underset{\beta}{\operatorname{argmin}} \frac{1}{N} \sum_{n=1}^{N} \sum_{n=1}^{N} (y_n - \beta^\top x_n)^2$$

by arguing that the squared error is a reasonable measure of "misfit." In this view, the objective is simply to draw a line through the data that is somehow "close" to the datapoints. We discussed how choices other than the vertical squared distance to y_n might give different

fits. For example, minimizing absolute distance, or horizontal distance along some x_n direction, will give different lines through the data.

In this story, there is no probabilistic model. The regression line is simply a summary of a complex dataset. We can also motivate the same objective function by a variety of different probabilistic assumptions, which I'll briefly (and incompletely) survey here.

Risk minimization

Let's assume that the data we observe, (x_n, y_n) are drawn IID (as pairs) from some distribution. Suppose our goal is to predict a new datapoint, y_{new}, using only a new regressor, x_{new}, drawn from the same distribution. Our guess can depend on x_{new}, so let's write $f(x_{\text{new}})$ for the guess. We want to pick an f that is as "good as possible." This is an instance of the machine learning subfield known as "supervised learning."

What does "good as possible" mean? Well, we want $f(x_{\text{new}})$ to be "close" to y_{new}, and one (but not the only) reasonable definition of closeness is the squared distance. Define the "loss function"

$$\mathcal{L}(f(x), y) = (y - f(x))^2.$$

For any particular f, the loss will be different for different x and y. Furthermore, the distribution of y_{new} is random given x_{new}. We want an f that does well over all possible x and y in some sense. A reasonable way to express this is that we want to find f to minimize the *expected loss*, or *risk*:

$$\mathcal{R}(f) = \mathbb{E}\left[\mathcal{L}(f(x), y)\right] = \mathbb{E}\left[(y - f(x))^2\right].$$

Now, we cannot actually calculate $\mathcal{R}(f)$, because we don't know the actual distribution of a future datapoint. But since we are assuming that the distribution is the same as the data we have, we can tentatively invoke the law of large numbers to define the *empirical risk*:

$$\hat{\mathcal{R}}(f) := \frac{1}{N}\sum_{n=1}^{N}(y_n - f(x_n))^2.$$

This is now an objective that we can actually compute. However, it doesn't make sense to minimize this over all possible f. For example, the function

$$f_{\text{bad}}(x) = \begin{cases} y_n \text{ if } x = x_n \text{ for some } n \\ 0 \text{ otherwise} \end{cases}$$

fits the data perfectly (so $\hat{\mathcal{R}}(f_{\text{bad}}) = 0$), but we expect it to provide a poor prediction if we get an x_{new} that we haven't seen before. So we might restrict the form of f to be something that's not too expressive — for example, we might only accept functions of the form $f(x;\beta) = \beta^\top x$. Plugging in this restriction, we can see that empirical risk minimization is the same as the OLS estimator.

Maximum likelihood

In the risk minimization version of OLS, we only assumed that (x, y) were IID (and, implicitly, had certain finite moments). We did not assume any functional form of the distribution. And we didn't assume that the optimal prediction function was linear — the linearity assumption was introduced only in an ad-hoc way to avoid overfitting.

A very different motivation for OLS proceeds by making much stronger statistical assumptions and writing out a maximum likelihood estimator (MLE). In this view, OLS is performing *inference*: that is, identifying the unknown parameters of a probability distribution from which we have samples.

Specifically, let us assume that x_n have some distribution $p(x)$ that we don't care about, and that the y_n are distributed independently as

$$p(y_n|x, \beta^*, \sigma)\mathcal{N}\left(x_n^\top \beta^*, \sigma^2\right) \text{ for some } \beta^* \text{ and } \sigma.$$

There are a few key conceptual differences with risk minimization:

- If we knew β^* and σ, we would know the distribution of $y|x$ completely.
- We're not imagining getting new data, nor defining any kind of prediction error.
- We want to know the parameters, not just make good predictions.

How should we estimate σ and β^*? A commonly used tool is *maximum likelihood estimation*, which finds the values of these parameters that maximize the likelihood of the observed data. Note that by the standard univariate normal distribution formula, up to constants not depending on β and σ,

$$\log p(y|\beta, x, \sigma) = -\frac{1}{2\sigma^2}(y - x^\top \beta)^2 - \frac{1}{2}\log \sigma^2.$$

Note that $p(y, x|\beta, \sigma) = p(y|\beta, x, \sigma)p(x)$. So for any value of σ, the MLE for β is given by

$$
\begin{aligned}
\widehat{\beta} &:= \underset{\beta}{\operatorname{argmax}} \prod n = 1^N p(y_n, x_n|\beta, \sigma) \\
&= \underset{\beta}{\operatorname{argmax}} \sum_{n=1}^{N} \log p(y_n|\beta, x_n, \sigma) \\
&= \underset{\beta}{\operatorname{argmax}} -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y - x^\top \beta)^2 - \frac{N}{2} \log \sigma^2 + \sum_{n=1}^{N} p(x_n) \\
&= \underset{\beta}{\operatorname{argmax}} -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y - x^\top \beta)^2 \\
&= \underset{\beta}{\operatorname{argmin}} \frac{1}{2\sigma^2} \sum_{n=1}^{N} (y - x^\top \beta)^2 \\
&= \underset{\beta}{\operatorname{argmin}} \frac{1}{N} \sum_{n=1}^{N} (y - x^\top \beta)^2.
\end{aligned}
$$

It follows that, no matter what σ is, the MLE for β^* is the OLS estimator.

Bayesian estimates

The maximum likelihood estimator is usually justified in asymptotic terms. That is, as $N \to \infty$, it is consistent and efficient when the model is correctly specified, i.e., when the data actually comes from the specified model for some β^* and σ. Usually that additionally assumes that the dimension of the parameter to be estimated — β^* in this case — is small relative to N.

For inference problems in finite samples, or where P is of the same order as N, then Bayesian statistics provides a way forward, again under correct model specification. We assume everything we assumed for the MLE, but additionally assume

- That σ is known
- In order to generate our data, the universe drew the components of β^* IID from normal distributions with very high variance: ${\beta^*}_p \overset{\text{IID}}{\sim} \mathcal{N}\left(0, \sigma_\beta^2\right)$. For short, we can write $\beta^* \sim p(\beta^*)$ to represent this distribution.

(These assumptions can be relaxed and we can still do Bayesian inference, but the math is more complicated, and the relationship to OLS may be less clear.)

The whole process of choosing a β^* and dataset Y is now random, and there is a joint distribution (keeping X fixed), $p(Y, \beta^*|X) = p(Y|\beta^*, X)p(\beta^*)$. Each new dataset can be imagined as a draw from this joint distribution. On the dataset we have, there was some dependence between what β^* is and what Y is. So knowing what Y is should allow us to guess what β^*

was in our case, even though we don't observe it. This guess takes the form of a *posterior distribution* given by Bayes' rule:

$$p(\beta^*|Y,X) = \frac{p(\beta^*)p(Y|\beta^*,X)}{p(Y|X)}.$$

It turns out that, since both $p(Y|X)$ and $p(\beta^*)$ are normal, this distribution has a closed form with expected value given by

$$\mathbb{E}_{p(\beta^*|Y,X)}\left[\beta^*\right] = (X^\top X + \sigma_\beta^{-2} I_P)^{-1} X^\top Y,$$

and for very large σ_β, we have

$$\mathbb{E}_{p(\beta^*|Y,X)}\left[\beta^*\right] \approx \hat{\beta}.$$

Method of moments estimation

Finally, we describe a way to do inference that allows for only partial specification of the probability model. Note that our MLE assumptions are equivalent to

$$y_n = {\beta^*}^\top x_n + \varepsilon_n,$$

where $\varepsilon_n \overset{\text{IID}}{\sim} \mathcal{N}(0, \sigma^2)$ and independent of x_n. From this it follows that

$$\mathbb{E}\left[x\varepsilon\right] = \mathbb{E}\left[x(y - x^\top \beta^*)\right] = 0.$$

That is, we assume that the residuals are uncorrelated with the regressors.

Suppose now we drop the normal assumption, and simply assume that $\mathbb{E}\left[x\varepsilon\right] = 0$. We can define our estimator $\hat{\beta}$ to be the value of β that makes the population version of this expectation to hold:

$$\hat{\beta} := \beta \text{ such that } \frac{1}{N}\sum_{n=1}^{N} x_n(y_n - x_n^\top \beta) = 0.$$

Rearranging this expression gives that $\hat{\beta}$ is indeed the OLS estimator.

Davidson, Russell, James G MacKinnon, et al. 2004. *Econometric Theory and Methods*. Vol. 5. Oxford University Press New York.

Gelman, Andrew, Jennifer Hill, and Aki Vehtari. 2021. *Regression and Other Stories*. Cambridge University Press.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. "An Introduction to Statistical Learning."