Linear Regression

02 Jan 2025

"Everything should be made as simple as possible, but not simpler."

— Albert Einstein

1 Introduction

The most general relationship between variables x and y is a statistical one. Every data point (x, y) is generated by sampling from the joint distribution between x and y, denoted by $p(x, y)$. It is useful to write this relationship in terms of the distribution of y conditioned on x, since often we care about predicting y given observations of x. We therefore write

$$(x, y) \sim p(y|x)\, p(x)\,, \tag{1.1}$$

where $p(x)$ is the marginal distribution of x. In general we want to learn $p(y|x)$ from observed data $\mathcal{D} \equiv \{(x_i, y_i) : i = 1, ..., N\}$. However, we are often limited to learning the conditional mean $\mathbb{E}[y|x]$ (as in the case of minimising an L_2 loss), or median (as in the case of minimising an L_1 loss).

1.1 Linear model

The simplest model is a linear one that assumes y depends linearly on the model parameters β. One example, for the univariate case is

$$y = \beta_0 + \beta_1 x + \varepsilon\,, \tag{1.1.1}$$

where ε is the residual, a random variable which captures the uncertainty in measurements of y. Another example is

$$y = \beta_0 + \beta_1 x^2 + \varepsilon \tag{1.1.2}$$

The only difference is that x has been replaced with x^2, which makes the model non-linear in x. However, since the model is still linear in y and the model parameters β_0, β_1, this is still considered a linear model. **Linearity, in this context, means linear w.r.t y and β.**

Without loss of generality we take eq. (1.1.1) to be our model. Having chosen a model the next obvious question is how we fit the model parameters (in this case β_0 and β_1) given some data? A common approach is to do *ordinary least squares (OLS)* regression, where one quantifies the performance of a set of parameters by the sum of squared differences between predictions and observed values, $L = \sum_i [y_i - \beta_0 - \beta_1 x_i]^2$.

It is certainly reasonable to consider this loss function, but why not the sum of absolute values or sum of 4-th power residuals, or something else entirely? Does it even matter? It turns out it does matter. Since it matters, it's important to motivate this loss function to see what implicit assumptions are being made. We will do this in the next section.

1.2 Deriving the least-squares loss

We start by specifying the conditional distribution $f(y|x)$. Given x, the randomness in y is sourced by the residual ε. If we assume $\varepsilon \sim N(0, \sigma^2)$ then for a single observation we get the log-likelihood

$$-2\log\mathcal{L}(y|x,\beta) = \frac{1}{\sigma^2}(y-\beta_0-\beta_1 x)^2 + \text{constants}\ . \tag{1.2.1}$$

When we have N data this becomes

$$-2\log\mathcal{L}(y|x,\beta) = \frac{1}{\sigma^2}\sum_{i=1}^{N}(y_i-\beta_0-\beta_1 x_i)^2\ . \tag{1.2.2}$$

On the lhs of eq. (1.2.2) y represents the collection $(y_1, y_2, ..., y_N)$, and similarly for x.

We can derive statistical estimators for β_0 and β_1 by finding the values where they maximise the likelihood, or equivalently, minimise the negative log-likelihood. From eq. (1.2.2) we identify that the loss given by the negative loss-likelihood is precisely the least-squares loss function.

In other words, **assuming Gaussian residuals $\varepsilon_i \sim N(0, \sigma^2)$ leads to the least squares loss**.

1.3 Aside: what do we learn by minimising the least-squares loss function?

Suppose we have a flexible model $f(x;\theta)$ with parameters θ that we wish to train to predict y given measurements of x. If we identify the best-fit parameters θ^* as those that minimise the squared difference between our model and true values on our dataset $\mathcal{D}$. That is, by minimising

$$L[f] = \sum_i \left[y_i - f(x_i;\theta)\right]^2\ , \tag{1.3.1}$$

where I've written $L[f]$ to emphasize that the loss function can be interpreted as a functional in terms of the model f. In the limit of infinite data the sum over i is just an average over the joint distribution $f(x,y)$.

$$\begin{aligned} L[f] &\to \iint [y-f(x;\theta)]^2 p(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \iint [y-f(x;\theta)]^2 p(y|x)p(x)\,\mathrm{d}x\,\mathrm{d}y \end{aligned} \tag{1.3.2}$$

Now we minimise L by varing f (we could equivalently varying θ, but doing it this way is more clean, and more fun). Setting $\delta L = 0$ yields

$$\begin{aligned} &\frac{\delta L}{\delta f} = \int (y-f(x;\theta))\,p(y|x)\,p(x)\,\mathrm{d}y = (\mathbb{E}[y|x] - f(x;\theta))\,p(x) = 0 \\ &\quad\Rightarrow f(x;\theta^*) = \mathbb{E}[y|x]\ . \end{aligned} \tag{1.3.3}$$

This is an important result. It tells us that even if we have *infinite data* and an *arbitrarily flexible model*, **the best we can do by minimising a least-squares loss is to learn the conditional expectation of y given x**.

Note: A really nice reference for the content in this section is the introduction of ref. [1].

1.4 OLS estimators

The maximum likelihood estimator is obtained by taking the derivative of eq. (1.2.2) w.r.t. to β_0 and β_1, setting their results equal to zero, and solving for β_0 and β_1. The results are simply,

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}} = \rho_{xy}\frac{S_{yy}}{S_{xx}}, \tag{1.4.1a}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}\,, \tag{1.4.1b}$$

where I have introduced the estimators for standard error S and correlation ρ,

$$S_{xy}^2 \equiv \frac{1}{n-1}\sum_i (x_i - \overline{x})(y_i - \overline{y}) \tag{1.4.2a}$$

$$\rho_{xy} \equiv \frac{S_{xy}}{S_{xx}S_{yy}}\,, \tag{1.4.2b}$$

and overlines denote sample means, e.g. $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$.

1.5 Properties of estimators

Unbiasedness

When conditioned on x we can show the estimators are unbiased. In the following when I write $\mathbb{E}(y)$ I mean the expectation of y conditioned on x. Moreover, the expectation of any arbitrary function of $x = (x_1, x_2, ..., x_N)$ is itself when conditioned on x.

$$\mathbb{E}\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})\mathbb{E}(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}\,, \tag{1.5.1}$$

but,

$$\mathbb{E}(y_i - \overline{y}) = (\beta_0 + \beta_1 x_i - \beta_0 - \beta_1 \overline{x}) = \beta_1 (x_i - \overline{x})\,. \tag{1.5.2}$$

And so,

$$\mathbb{E}\hat{\beta}_1 = \beta_1\,. \tag{1.5.3}$$

Thus, for $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$ we have

$$\mathbb{E}\hat{\beta}_0 = \mathbb{E}\overline{y} - \overline{x}\,\mathbb{E}\hat{\beta}_1 = \beta_0 + \beta_1 \overline{x} - \overline{x}\hat{\beta}_1 = \beta_0\,. \tag{1.5.4}$$

Variance

I'll just state the results, because it is tedious.

$$\mathrm{Var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum_i (x_i - \overline{x})^2} \tag{1.5.5a}$$

$$\mathrm{Var}\left(\hat{\beta}_0\right) = \frac{\sigma^2 \sum_i x_i^2}{n\sum_i (x_i - \overline{x})^2} \tag{1.5.5b}$$

1.6 Signifcance testing

The linear correlation between x and y is typically assessed via the t-statistic,

$$\hat{t} = \frac{\hat{\beta}_1}{\hat{\sigma}/S_{xx}}\,, \tag{1.6.1}$$

where $\hat{\sigma}$ is the estimator for the standard deviation of the residuals and is given by

$$\hat{\sigma} = \sum_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2\,. \tag{1.6.2}$$

If the ε_i are assumed to (1) be Gaussian with mean zero (2) have no autocorrelation (3) exhibit weak exogeneity, then the t-statistic follows a t distribution which can be used to calculate p-values for significance testing. However, if any of these assumptions are violated you can't use the standard p-values. This happens basically all the time in financial time series analysis where, for example, you may model the next time step y_t as a linear combination of lagged values. This introduces autocorrelation in the residuals. The Dickey-Fuller test takes this into account when calculating p-values for the presence of a unit root.

1.7 Multiple regressors

Suppose we want to use p covariates to predict the variate y. We can write down this model as

$$y_i = \beta_0 + \sum_{k=1}^{p} x_{k,i}\beta_k + \varepsilon_k \text{ for } i = 1, 2, ..., N. \tag{1.7.1}$$

Or, if we define $X \in \mathbb{R}^{N\times(p+1)}$ as the matrix

$$X \equiv \begin{pmatrix} 1 & x_{1,1} & x_{2,1} & ... & x_{N,1} \\ 1 & x_{1,2} & x_{2,2} & ... & x_{N,2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1,N} & x_{2,N} & ... & x_{N,2} \end{pmatrix} = \begin{pmatrix} | & | & | & | & | \\ 1 & x_1 & x_2 & ... & x_N \\ | & | & | & | & | \end{pmatrix} \tag{1.7.2}$$

In this case the estimator for the regression parameters is

$$\hat{\beta} = (X^T X)^{-1} X^T y\,. \tag{1.7.3}$$

Its variance-covariance matrix is[1]

$$\mathrm{Var}(\beta) = \sigma^2 X^T X\,. \tag{1.7.4}$$

1.8 Assumptions

Up until this point I haven't gone into much detail about the assumptions we have made. I've just blitzed through the derivation of the estimators. Here we enumerate the assumptions and give them fancy names which I think were popularised by econometrics. Memorising the assumptions is important because they are almost always violated. If they're violated a little then you're probably fine proceeding as usual, but when they're violated a lot we need to introduce ways to fix things.

[1] this can be easily derived by using the identity $\mathrm{Var}(Ax) = A\,\mathrm{Var}(x)A^T$, and using the assumption of homoscedasticity to write $\mathrm{Var}(\varepsilon) = \sigma^2 \mathbb{1}_{p+1}$.

Linear Regression Assumptions

1. **Linearity**: the model is linear in the variate and parameters.
2. **Random sampling**: the data (x_i, y_i) are i.i.d., ensuring that the sample is representative of the population.
3. **No perfect multicollinearity**: the p covariates are linearly independent. This imposes $\text{Rank}(X) = p$.
4. **Weak exogeneity**: no information loss in Y when conditioned on X, $\mathbb{E}[\varepsilon|X] = 0$.
5. **Homoscedasticity** the variance of errors is constant across all values of X.
6. **No autocorrelation**: errors are uncorrelated, $\mathbb{E}\left[\varepsilon_i, \varepsilon_j\right] = 0 \ \forall \ i \neq j$..
7. **Errors follow a distribution (optional)**: here we assumed Gaussian, but they could've been t-distributed
8. **Model specification**: basically "the model is correct". This assumption is often violated if for example there are additional features which have not been included in the model.

1.9 Gauss Markov theorem

One of the most famous results is that the estimator eq. (1.7.3) is the *best linear unbiased estimator (BLUE)*, where best means lowest variance. The derivation is pretty straightforward so I will present it here. First we define an arbitrary linear estimator of β as an estimator of the form

$$\tilde{\beta} = Ay \,, \tag{1.9.1}$$

where $A \in \mathbb{R}^{(p+1)\times N}$. If it's unbiased then,

$$\mathbb{E}\Big(\tilde{\beta}\Big) = \beta \,. \tag{1.9.2}$$

On the other hand substituting eq. (1.9.1) for y and using the fact that $\mathbb{E}(\varepsilon) = 0$ yields

$$\mathbb{E}\Big(\tilde{\beta}\Big) = A\mathbb{E}(X\beta + \varepsilon) = AX\beta \,. \tag{1.9.3}$$

Combining eqs. (1.9.2) and (1.9.3) gives

$$AX\beta = \beta \Rightarrow AX = \mathbb{1}_{p+1} \,. \tag{1.9.4}$$

Eq. (1.9.4) motivates us to decompose A as

$$A = \left(X^T X\right)^{-1} X^T + C \,, \tag{1.9.5}$$

where $C \in \mathbb{R}^{(p+1)\times N}$ is in the null space of X, i.e., $CX = 0$. The first term can't simply be X^{-1} since we need a matrix with the shape $(p+1) \times N$, and X^{-1} would be $N \times (p+1)$.

The variance of $\tilde{\beta}$ can be written as

$$\begin{aligned}
\text{Var}\Big(\tilde{\beta}\Big) &= \text{Var}(A(X\beta + \varepsilon)) = \text{Var}(A\varepsilon) = A \ \text{Var}(\varepsilon) A^T && \text{(1.9.6a)} \\
&= \Big[\left(X^T X\right)^{-1} X^T + C\Big] \sigma^2 \Big[\left(X^T X\right)^{-1} X^T + C\Big]^T \ \text{(using Var}(\varepsilon) = \sigma^2) && \text{(1.9.6b)} \\
&= \sigma^2 \Big\{\left(X^T X\right)^{-1} + X^T X^{-1} X^T C^T + CX\left(X^T X\right)^{-1} + CC^T\Big\} && \text{(1.9.6c)} \\
&= \text{Var}\Big(\hat{\beta}\Big) + \sigma^2 CC^T \,. && \text{(1.9.6d)}
\end{aligned}$$

To go from eq. (1.9.6c) to eq. (1.9.6d) I eliminated the cross terms in the middle via the fact that $CX = 0$ and rewrote the first term using eq. (1.7.4). Since C is a positive semi-definite matrix we have shown that $\mathrm{Var}\left(\tilde{\beta}\right)$ exceeds $\mathrm{Var}\left(\hat{\beta}\right)$ by a positive semi-definite matrix[2], $\sigma^2 CC^T$.

2 Consequences of violating Gauss Markov assumptions

2.1 Weak exogeneity

Some terms:

Definition 2.1.1: *Exogeneity* is the assumption that measurement errors are uncorrelated with the covariate x. In other words, $\mathrm{Cov}(x, \varepsilon) = 0$. We often write $\mathbb{E}[\varepsilon|x] = 0$

Definition 2.1.2: *Endogeneity* refers to the errors in measurement of Y being correlated with measurements of x.

In this section we consider the single-variable model in eq. (1.1.1). We have assumed that there are no errors in our observations of the covariate x, but it's possible there actually are errors. If we naiively use the OLS estimator for $\hat{\beta}$ how does the estimate relate to the true value? Violation of weak exogeneity is sometimes referred to as errors-in-variables. In OLS regression it leads to *attenuation bias*, where $\hat{\beta}$ becomes biased towards 0.

First, let's arrive at the effect using intuition. The OLS estimator for β is S_{xy}/S_x. If there are no errors in the measurement of x then the only thing obscuring our ability to see the true covariance between x and y are the errors in y that we assume in OLS regression. Adding errors to x has the effect of reducing the observed covariance between x and y, so we should expect that if we use the OLS estimator in this case, our estimate would be biased towards zero than the same estimator when used in the case when there are no errors in x.

Now some maths. Denote the true value of x by x^* and let the error in measurements of x_* be η. The model is given by taking eq. (1.1.1) and replacing $x \to x_*$,

$$y = \beta_0 + x_*\beta_1 + \varepsilon\,, \tag{2.1.1}$$

but since we can only measure $x = x_* + \eta$ we have, in practice,

$$\begin{aligned} y &= \beta_0 + (x - \eta)\beta_1 + \varepsilon \\ &= \beta_0 + x\beta_1 + (\varepsilon - \beta_1\eta) && \text{(2.1.2a)} \\ &\equiv \beta_0 + x\beta_1 + \tilde{\varepsilon}\,, && \text{(2.1.2b)} \end{aligned}$$

where $\tilde{\varepsilon} = \varepsilon - \beta_1\eta$ is identified as the "new" residual, which is now correlated with x. The OLS estimator for β_1 then converges to

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2} \to \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)} = \frac{\sigma^2_{x_*}}{\sigma^2_{x_*} + \sigma^2_\eta}\,\beta_1, \tag{2.1.3}$$

[2]To verify that CC^T is positive semi-definite simply write $v = C^T x$, then for any x $|v|^2 = x^T CC^T x \geq 0$.

which is less than or equal to β_1. This effect is called *attenuation damping*. In deriving this expression I used the fact that we condition on the observed x but are uncertain about the true value x_* and the noise η. We have,

$$\begin{aligned}\mathrm{Cov}(x, y) &= \mathrm{Cov}(x_* + \eta, \beta_0 + \beta_1 x_* + \varepsilon) \\ &= \mathrm{Cov}(x_*, \beta_1, x_*) + \mathrm{Cov}(\eta, \beta_1 x_*) + \mathrm{Cov}(\eta, \varepsilon) \\ &= \beta_1 \ \mathrm{Var}(x_*) + 0 + 0 \equiv \beta_1 \sigma_{x*}^2\end{aligned}$$

Note: The first time I encountered this I was very confused about the meaning of $\mathrm{Cov}(x, y)$ because I had the perspective that x is not a random variable and y is.

2.1 References

[1] C. M. Bishop, "Mixture density networks," 1994.