Linear Regression

31 Jan 2025

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

— Albert Einstein

Contents

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, $L = \sum_i \left[y_i - \beta_0 - \beta_1 x_i\right]^2$.

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

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

In eq. (1.2.2) the y appearing in the LHS of the equation 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 becomes an average weighted by 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

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

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 rearranging. The results are simply,

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

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

In eq. (1.4.1a) I have introduced the estimators for standard error S and correlation ρ,

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

$$S_x^2 \equiv \frac{1}{N-p}\sum_i (x_i - \overline{x})^2 \tag{1.4.2b}$$

$$\rho_{xy} \equiv \frac{S_{xy}^2}{S_x S_y}\,, \tag{1.4.2c}$$

where k is the number of degrees of freedom. If we regress on both β_0 and β_1 then $k = 2$. If we omit β_0 (i.e. assume it is zero), then $k = 1$. Overlines denote sample means, e.g. $\overline{x} = \frac{1}{N}\sum_{i=1}^N x_i$.

1.5 Properties of estimators

Bias

The first property of the OLS estimators is that they are *unbiased*, when we condition on x. This can be shown with a straightforward calculation that I will carry out below. Note that in the following all expectation values are *conditional on* x. Hence, when I write $\mathbb{E}(y)$ I really

mean $\mathbb{E}[y|x]$ (the expectation of y conditioned on x). This implies that, the expectation of any arbitrary function of $x = (x_1, x_2, ..., x_N)$ is itself when conditioned on x, e.g. $\mathbb{E}[\|x\|^2] = \|x\|^2$.

First, let's evaluate the expectation of $\hat{\beta}_1$. We have,

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

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 present the formulas for quick reference then derive the formula for $\mathrm{Var}\left(\hat{\beta}_1\right)$.

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

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

The variance of $\hat{\beta}_1$ may be written as

$$\mathrm{Var}\left(\hat{\beta}_1\right) = \left[\sum_i (x_i - \overline{x})^2\right]^{-2} \mathrm{Var}\left[\sum_i (x_i - \overline{x})(y_i - \overline{y})\right] , \tag{1.5.6}$$

where I have used the fact that we are conditioning on x to factor the denonimator out a-la the identity $\mathrm{Var}(kY) = k^2\,\mathrm{Var}(Y)$ for constant k and random variable Y. Now let's focus on the variance factor on the right. For convenience, introduce the notation $x_i^* \equiv x_i - \overline{x}$ and notice that $y_i - \overline{y} = \beta_1 x_i^* + \varepsilon_i - \overline{\varepsilon}$, so that when we take the variance only the ε terms will be relevant:

$$\mathrm{Var}\left[\sum_i (x_i - \overline{x})(y_i - \overline{y})\right] = \mathrm{Var}\left(\sum_i x_i^* [\varepsilon_i - \overline{\varepsilon}]\right) \tag{1.5.7a}$$

$$= \mathbb{E}\left[\sum_{i,j} x_i^* x_j^* (\varepsilon_i - \overline{\varepsilon})(\varepsilon_j - \overline{\varepsilon})\right] \tag{1.5.7b}$$

$$= \sum_{i,j} x_i^* x_j^* \left\{\mathbb{E}\left[\varepsilon_i \varepsilon_j - \varepsilon_i \overline{\varepsilon} - \varepsilon_j \overline{\varepsilon} + \overline{\varepsilon}^2\right]\right\} , \tag{1.5.7c}$$

where in the second equality we used the fact that $\mathbb{E}[\varepsilon_i] = \mathbb{E}[\overline{\varepsilon}] = 0$. Using the linearity of expectation to expand eq. (1.5.7c) yields

$$\sum_{i,j} x_i^* x_j^* \left\{ \mathbb{E}\left[\varepsilon_i \varepsilon_j\right] - \mathbb{E}[\varepsilon_i \bar{\varepsilon}] - \mathbb{E}\left[\varepsilon_j \bar{\varepsilon}\right] + \mathbb{E}[\bar{\varepsilon}^2] \right\} . \tag{1.5.8}$$

However, since $\mathbb{E}\left[\varepsilon_i \varepsilon_j\right] = \delta_{i,j}\sigma^2$ and $\bar{\varepsilon} = \frac{1}{n}\sum_k \varepsilon_k$ all of these expectation values can be simplified.

$$\sum_{i,j} x_i^* x_j^* \left\{ \sigma^2 \delta_{ij} - \frac{\sigma^2}{n} - \frac{\sigma^2}{n} + \frac{\sigma^2}{n} \right\} \tag{1.5.9a}$$

$$= \sigma^2 \sum_i \left(x_i^*\right)^2 - \frac{\sigma^2}{n} \sum_{i,j} x_i^* x_j^* . \tag{1.5.9b}$$

In the rightmost term we recognise that $\sum_{i,j} x_i^* x_j^* = \left(\sum_i x_i^*\right)^2$, and furthermore, $\sum_i x_i^* = 0$ by definition, so the term vanishes and we're left with

$$\mathrm{Var}\left(\hat{\beta}_1\right) = \frac{\sigma^2 \sum_i \left(x_i^*\right)^2}{\left(\sum_i \left(x_i^*\right)^2\right)^2} = \frac{\sigma^2}{\sum_i \left(x_i - \overline{x}\right)^2} , \tag{1.5.10}$$

which exactly matches eq. (1.5.5a).

1.6 Significance testing

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

$$\hat{t} = \frac{\hat{\beta}_1}{\mathrm{StdErr}\left(\hat{\beta}_1\right)} = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 / S_x^2}} , \tag{1.6.1}$$

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

$$\hat{\sigma}^2 = \frac{1}{N-k} \sum_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 . \tag{1.6.2}$$

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 with $N-k$ degrees of freedom. This 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.

1.6.1 t-test

Let's explore the t-statistics properties in more detail. First let's discuss the distribution from which $\hat{\beta}_1$ is drawn under the null hypothesis $\beta_1 = 0$ and argue that $\hat{t}$ indeed follows a t-distribution.

The t-distribution with k degrees of freedom arises when you divide a standard normal random variable by a χ_k^2 random variable, normalised so its mean is 1. I.e.,

$$Z \sim N(0,1), \quad X^2 \sim \chi_k^2 \Rightarrow \frac{Z}{\sqrt{X^2/k}} \sim t_k . \tag{1.6.1.3}$$

Under the model given in eq. (1.1.1) we are assuming that the observed values of y fluctuate around the 'true trend' $\beta_1 x$ due to Gaussian noise[1] . If there is no relationship between x and y, then under eq. (1.1.1) this means $\beta_1 = 0$. However, even if $\beta_1 = 0$ our OLS estimate $\hat{\beta}_1$ will generally be nonzero in a given sample. The question is how do we determine if an obtained nonzero $\hat{\beta}_1$ is statistically significant? Assume the null hypothesis $\beta_1 = 0$ and $\varepsilon \sim N(0, \sigma^2)$. Since $\hat{\beta}_1$ is a linear combination of the elements of $y = \varepsilon$, which are normally distributed with mean 0 and variance σ^2, $\hat{\beta}_1$ must also be normally distributed with mean 0 – and we already know its variance from eq. (1.6.1.5a). So $\frac{\hat{\beta}_1}{\sqrt{\mathrm{Var}(\hat{\beta}_1)}} \sim N(0,1)$. Meanwhile, roughly speaking we can write

$$\frac{1}{\sigma^2}\sum_i (y_i - \hat{y}_i)^2 \equiv \frac{1}{\sigma^2}\sum_i \hat{\varepsilon}_i^2 \sim \chi^2_{N-k} \tag{1.6.1.4}$$

so that $\hat{\sigma}^2 = \frac{1}{N-k}\sum_i \hat{\varepsilon}_i^2$ is a scaled χ^2_{N-k} random variable with mean σ^2. Then,

$$\frac{\hat{\beta}_1 / \sqrt{\mathrm{Var}\left(\hat{\beta}_1\right)}}{\sqrt{\hat{\sigma}^2/\sigma^2}} \tag{1.6.1.5}$$

is of the same form as eq. (1.6.1.3). We can simplify by explicitly writing $\mathrm{Var}\left(\hat{\beta}_1\right) = \sigma^2 / \sum (x - \overline{x})^2$ so that we obtain

$$\frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 / \sum_i (x_i - \overline{x})^2}}\,, \tag{1.6.1.6}$$

which is the $\hat{t}$-statistic.

Significance testing is then done by finding the p-value of $\hat{t}$. Let F denote the cdf of the t distribution with $N-k$ degrees of freedom. Then $p = 1 - \left(F(\hat{t}) - \left(F(-\hat{t})\right)\right) = 2\left(1 - F(\hat{t})\right)$ for the two-tailed test, and $p = 1 - F(\hat{t})$ for the one-tailed test (under the null hypothesis that $\beta_1 \leq 0$).

1.7 Multiple regressors

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

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

It's conventional to define the so-called *design matrix* $X \in \mathbb{R}^{N\times(p+1)}$ as[2]

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

In this case the estimator for the regression parameters is

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

[1] eq. (1.1.1) doesn't require the noise to be Gaussian, but this is the most common assumption.

[2] In the literature I mostly see people saying the design matrix is $N \times p$

Its variance-covariance matrix is[3]

$$\mathrm{Var}\left(\hat{\beta}\right) = \sigma^2 (X^T X)^{-1} \,. \tag{1.7.4}$$

Eq. (1.7.4) can be derived easily by using the identity $\mathrm{Var}(A\vec{x}) = A\ \mathrm{Var}(\vec{x})\, A^T$ where $\mathrm{Var}(\vec{x})$ denotes the variance-covariance matrix of the elements of $\vec{x}$: $[\mathrm{Var}(\vec{x})]_{ij} = \mathrm{Cov}\left(x_i, x_j\right)$. Here's a quick derivation of the identity, and then how it can be applied to eq. (1.7.3). I'm going to use the Einstein summation convention and denote the i-th element of x by x^i just for this derivation.

$$\begin{aligned}
[\mathrm{Var}(A\vec{x})]_{ij} &= \mathbb{E}\left[\left(A_{ik}x^k\right)\left(A_{jl}x^l\right)\right] - \mathbb{E}\left[A_{ik}x^k\right]\mathbb{E}\left[A_{jl}x^l\right] \\
&= A_{ik}\mathbb{E}[x^k x^l]A_{jl} - A_{ik}\mathbb{E}[x^k]\mathbb{E}[x^l]A_{jl} \\
&= A_{ik}\mathbb{E}[x^k x^l](A^T)_{lj} - A_{ik}\mathbb{E}[x^k]\mathbb{E}[x^l](A^T)_{lj} \\
&= A_{ik}\left(\mathbb{E}[x^k x^l] - \mathbb{E}[x^k]\mathbb{E}[x^l]\right)(A^T)_{lj} \\
&= A_{ik}\ \mathrm{Cov}\left(x^k, x^l\right)(A^T)_{lj} \\
&= \left[A\ \mathrm{Var}(x)A^T\right]_{ij} \,.
\end{aligned}$$

Applying this identity to eq. (1.7.3) we get

$$\mathrm{Var}\left(\hat{\beta}\right) = \mathrm{Var}\left(\left(X^T X\right)^{-1} X^T y\right) = \left(X^T X\right)^{-1} X^T\ \mathrm{Var}(y) X \left(X^T X\right)^{-1} \,. \tag{1.7.5}$$

The variance-covariance matrix $\mathrm{Var}(y)$ can be written as $\mathrm{Var}(X\beta + \varepsilon) = \mathrm{Var}(\varepsilon) = \sigma^2 \mathbb{1}_{p+1}$. Substituting this into eq. (1.7.5) immediately yields eq. (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 understand their implications. This will help us recognize the fingerprints of each assumption violation.

[3] 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. (1.9.4). 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[4], $\sigma^2 CC^T$.

1.10 Nested model comparison using F-test

An alternative way of determining whether a particular covariate is significant is using an F -test.

2 Consequences of violating Gauss Markov assumptions

2.1 Homoscedasticity

- Significance tests become unreliable
- OLS estimator is no longer the BLUE. The intuitive explanation for this is that it weights more noisy terms the same as less noisy terms. Therefore the strategy should be to downweight the importance of the more noisy samples compared to other samples. This line of reasoning leads us to weighted least squares regression.

Suppose the variance of the residuals is not constant. Assuming there is still no autocorrelation of errors we can write the general case as $\mathrm{Var}(\varepsilon_i) = \sigma_i^2$. Then if we take the linear model

$$y_i = \beta X_i + \varepsilon_i \tag{2.1.1}$$

(where $X_i \in \mathbb{R}^{p+1}$) and normalise both sides by $\frac{1}{\sigma_i}$ we can define

$$\frac{y_i}{\sigma_i} = \beta\left(\frac{X_i}{\sigma_i}\right) + \frac{\varepsilon_i}{\sigma_i}\,. \tag{2.1.2}$$

Since $\mathrm{Var}(\varepsilon_i/\sigma_i) = 1$ the residuals have constant variance. Moreover, the model is still linear, and β is unchanged. So if we do OLS estimation on the augmented data $(X_i/\sigma_i, y_i/\sigma_i)$ we no longer have problems with heteroscedasticity and can use the usual OLS estimator methods. In matrix form, let

$$W \equiv \mathrm{diag}(1/\sigma_i) \in \mathbb{R}^{n\times n}\,. \tag{2.1.3}$$

Then the weighted least squares estimator can be written

Weighted least squares (WLS) estimator:

$$\hat{\beta} = (X^T W^T W X)^{-1} X^T W^T (Wy)\,, \tag{2.1.4}$$

where

$$W = \mathrm{diag}(\sigma_i^{-1})\,. \tag{2.1.5}$$

Of course, we rarely know σ_i^2 precisely, so this also needs to be estimated. If we have reason to believe that the errors are dependent on X one can fit another model to estimate $\sigma^2(X)$, e.g. using another linear model, a decision tree, or a neural network. This might seem to violate the assumption of weak exogeneity, but this is not necessarily the case. You could have $\sigma^2 = \sigma^2(X)$ without violating $\mathbb{E}[\varepsilon|X] = 0$.

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

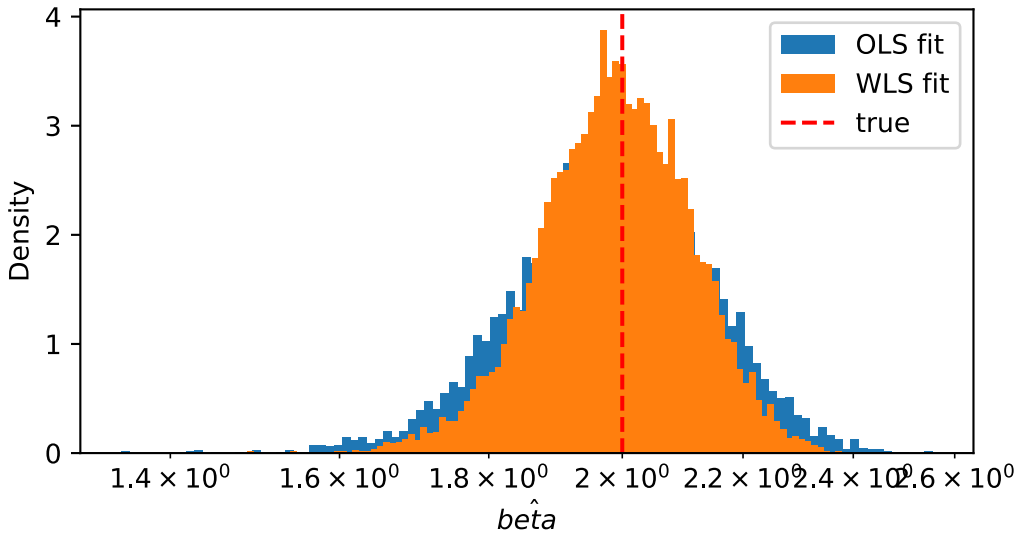
2.1.1 WLS Example

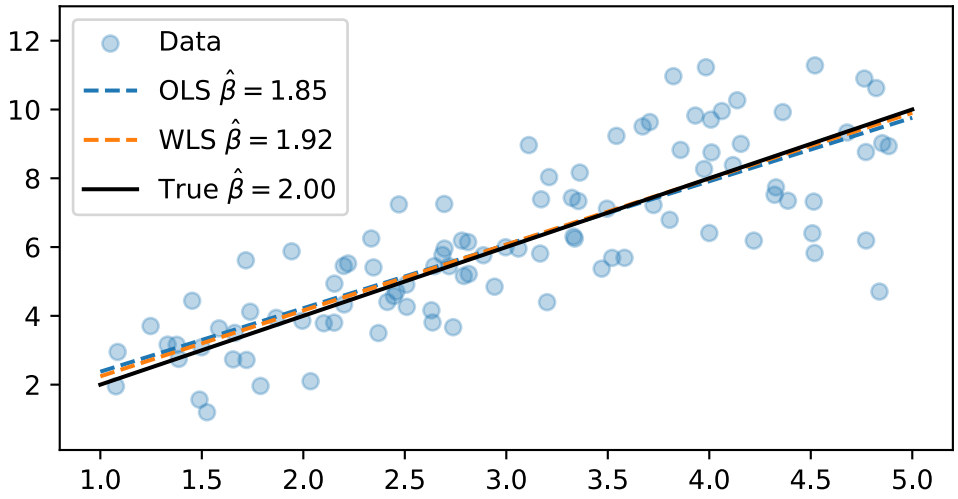
Consider the following setup:

$$x \sim U(1,5) \quad \varepsilon|x \sim \text{LogNorm}\left(0, \frac{1}{4}x^2\right) \tag{2.1.1.6a}$$

$$y|x, \varepsilon = \beta x + \varepsilon\,. \tag{2.1.1.6b}$$

That is, the error term is explicitly dependent on x. In this model $\mathbb{E}[\varepsilon|x] = \exp\left(\frac{1}{4}x^2\right) \neq 0$ so weak exogeneity is also violated.





Need to add:

1. Diagnosing heteroscedasticity (look at residuals vs predictions/vs each feature)
2. Analytic example where $\sigma^2(x) = \beta x + \alpha$
3. Example where you fit σ^2 with a model

2.2 Weak exogeneity

Some terms:

Definition 2.2.1: *Exogeneity* is the assumption that measurement errors are uncorrelated with the covariate x. In other words, $\text{Cov}(x, \varepsilon) = 0$. We often write $\mathbb{E}[\varepsilon|x] = 0$

Definition 2.2.2: *Endogeneity* refers to the errors in measurement of Y being correlated with measurements of x.

The primary issue associated with violation of weak exogeneity in linear regression models is *bias*. The OLS estimator $\hat{\beta}$ no longer satisfies $\mathbb{E}\left(\hat{\beta}\right) = \beta$. Endogeneity arises due to three main reasons:

1. **Omitted variable** (this is a type of model misspecification, so it also violates OL8.)
2. **Errors in measurement of the covariate**
3. **Reverse causality**

Omitted variable bias

Imagine the true data generating model is given by

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \,, \tag{2.2.1}$$

and further suppose x_1 and x_2 are correlated so that (but not perfectly colinear) $\text{Cov}(x_1, x_2) \equiv \rho\sigma_1\sigma_2$, where $\sigma_i \equiv \text{Var}(x_i)$. If we mistakenly assume a model of the form

$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \varepsilon \,, \tag{2.2.2a}$$

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{\varepsilon} \,, \tag{2.2.2b}$$

$$\Rightarrow \tilde{\varepsilon} = \varepsilon + \beta_2 x_2 \tag{2.2.3a}$$

$$\Rightarrow \text{Cov}(x_1, \tilde{\varepsilon}) = \text{Cov}(x_1, \varepsilon) + \beta_2 \ \text{Cov}(x_1, x_2) \neq 0 \,. \tag{2.2.3b}$$

$$\hat{\beta}_1 = \frac{\text{Cov}(x_1, y)}{\text{Var}(x_1)} \to \frac{\tilde{\beta}_1 \ \text{Var}(x_1) + \text{Cov}(x_1, \tilde{\varepsilon})}{\text{Var}(x_1)} \tag{2.2.4a}$$

$$= \tilde{\beta}_1 + \frac{\text{Cov}(x_1, \tilde{\varepsilon})}{\text{Var}(x_1)} \tag{2.2.4b}$$

$$= \tilde{\beta}_1 + \boxed{\rho\, \beta_2 \, \tfrac{\sigma_2}{\sigma_1}} \,. \tag{2.2.4c}$$

The expression in the box is the bias.

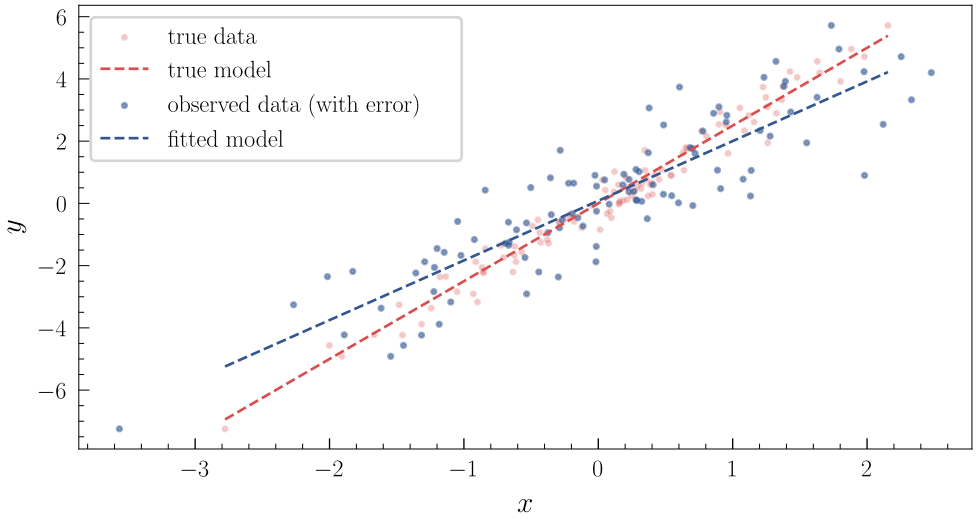


Figure 3: Demonstration of attenuation bias as a result of not accounting for errors in the covariate x.

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 $x \to x_*$,

$$y = \beta_0 + x_*\beta_1 + \varepsilon\,, \tag{2.2.5}$$

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

$$\begin{aligned} y &= \beta_0 + (x - \eta)\beta_1 + \varepsilon \\ &= \beta_0 + x\beta_1 + (\varepsilon - \beta_1\eta) && \text{(2.2.6a)} \\ &\equiv \beta_0 + x\beta_1 + \tilde{\varepsilon}\,, && \text{(2.2.6b)} \end{aligned}$$

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

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2} \to \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)} = \frac{\sigma_{x_*}^2}{\sigma_{x_*}^2 + \sigma_\eta^2}\,\beta_1, \tag{2.2.7}$$

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. From the perspective of these notes we have assumed that (x,y) are drawn from a distribution $f(x,y)$ since the beginning. This framework is natural in econometrics where you may have two time series X_t and Y_t which may both not be "control" variables. On the other hand, in experimental physics we may have more control over X (for example, it could be the length of a wire, which we can choose with good precision). Even this deterministic sampling of X can be modeled probabilistically, e.g. with Dirac deltas.

2.2.1 Diagnosing weak exogeneity

1. Look at the residuals as a function of the features, or the prediction. Is there a trend? Residuals should be 0-centered.

Need to add:

1. Reverse causality explanation

2.3 Multicollinearity

Note: My understanding of the maths of this section are fuzzy. I was following these notes for much of this section, but it seems like they do not include the constant β_0 term in their regression model. Of course, this can be achieved by standardising the target and the covariates e.g. $y \to y - \overline{y}$, $x_i \to x_i - \overline{x}_i$. Wikipedia suggests that the formulas here still stand up when you include β_0. In particular, in the expression below make the replacement $\left(X^T X\right)^{-1} \to \left[\sum (x - \overline{x})^2\right]^{-1}$ to get Wikipedia's expression.

This topic is a favourite in quant finance interviews. Multicollinearity means that two or more variables are linearly dependent (in practice, approximately linearly dependent) so that the covariance matrix $X^T X$ becomes (approximately) singular and some of the regression parameter estimates are undefined (blow up).

The most significant consequence of multicollinearity is *variance inflation*. The basic idea is that since the regression model is effectively trying to find how the target changes with each covariate while holding all but one constant, it isn't able to pick up on degeneracies. We can illustrate this with a simple example. Suppose we have just two covariates x_1 and x_2, and $x_1 = x_2$. Then our regression model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \,. \tag{2.3.1}$$

which can be rewritten as

$$y = \beta_0 + (\beta_1 + \beta_2) x_1 + \varepsilon \,. \tag{2.3.2}$$

Now notice that an increase in β_1 can be compensated by a decrease in β_2 and the equation remains unchanged. Our regression model estimates β_1 and β_2 separately, but they are not uniquely determined, even in the limit of infinite data. So the coefficients β_1 and β_2 will have *high variance*. Yet another way to think about eq. (2.3.2) is that the loss function $L = \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} x_{ij} \beta_j\right)$ will have a flat ridge minima.

Variance inflation factor

Start with eq. (2.3.4) (repeated below for convenience)

$$\mathrm{Var}\big(\hat{\beta}\big) = \sigma^2 (X^T X)^{-1}$$
$$\widehat{\mathrm{Var}}\big(\hat{\beta}\big) = \hat{\sigma}^2 (X^T X)^{-1} .$$

Then by eq. (2.3.4) we have

$$\mathrm{Var}\big(\hat{\beta}_k\big) = \sigma^2 \big(X_{\cdot,k}^T X_{\cdot,k}\big)^{-1} \frac{1}{1-R_k^2} , \tag{2.3.3a}$$

$$\widehat{\mathrm{Var}}\big(\hat{\beta}\big) = \frac{\hat{\sigma}^2}{(n-1)\widehat{\mathrm{Var}}(X_k)} \frac{1}{1-R_k^2} , \tag{2.3.3b}$$

where $X_{\cdot,k}$ is the k-th column of X and R_k^2 is the R-squared obtained by regressing the k -th regressor on all the other regressors. The rightmost factor is known as the *variance inflation factor (VIF)* and it's important enough that I'll enshrine it in a blue box.

$$\mathrm{VIF}_k = \frac{1}{1-R_k^2} \tag{2.3.4}$$

Clearly, if $R_k^2 \approx 1$, then the variance will be large. The important thing is that the VIF characterises how much of x_k can be explained by the other variables. If most of it can, then x_k may be worth removing.

3 Time series analysis

This section is basically just scratch notes at this point. I haven't really figured out structure or presentation. I'm just throwing down useful definitions for quick reference, and derivations that I found useful but weren't in my reference textbook.

The foundation of time series analysis is stationarity.

Definition 3.1: A time series $\{r_t\}$ is said to be weakly stationary if $\mathbb{E}[r_t] = \mu$ is independent of t and the autocovariance $\mathbb{E}[(r_t - \mu)(r_{t-\ell} - \mu)] = \gamma_\ell$ is only a function of the lag.

Stationarity can be checked using the Dickey-Fuller (or augmented Dickey-Fuller) test. To explain this test let's look at a simple example of a linear time series.

3.1 AR(2) model

The AR(2) model for r_t is defined as

$$r_t = \phi_0 + \phi_1 r_{t-1} + \phi_2 r_{t-2} + \varepsilon_t . \tag{3.1.1}$$

Assuming weak stationarity we can derive the mean $\mathbb{E}[r_t]$ as

$$\mathbb{E}[r_t] = \phi_0 + \phi_1 \mathbb{E}[r_{t-1}] + \phi_2 \mathbb{E}[r_{t-2}] + \mathbb{E}[\varepsilon_t] \tag{3.1.2a}$$
$$\Rightarrow \mu = \phi_0 + \phi_1 \mu + \phi_2 \mu \tag{3.1.2b}$$
$$\Rightarrow \mu = \frac{\phi_0}{1-\phi_1-\phi_2} . \tag{3.1.2c}$$

Next we can compute its autocovariance and autocorrelation functions. Rewriting the time series (in terms of deviations from the mean) we get

$$r_t - \mu = \phi_1(r_{t-1} - \mu) + \phi_2(r_{t-2} - \mu) + \varepsilon_t \,. \tag{3.1.3}$$

Now we multiply this by the lagged values on both sides:

$$(r_{t-\ell} - \mu)(r_t - \mu) = \phi_1(r_{t-\ell} - \mu)(r_{t-1} - \mu) + \phi_2(r_{t-\ell} - \mu)(r_{t-2} - \mu) + \varepsilon_t \tag{3.1.4}$$

Taking the expectation,

$$\begin{aligned}\mathbb{E}[(r_{t-\ell} - \mu)(r_t - \mu)] = \phi_1 \mathbb{E}[(r_{t-\ell} - \mu)(r_{t-1} - \mu)] \\ + \phi_2 \mathbb{E}[(r_{t-\ell} - \mu)(r_{t-2} - \mu)] + \mathbb{E}[\varepsilon_t]\end{aligned}$$

Apply stationarity to the expectation values,

$$\Rightarrow \gamma_\ell = \phi_1 \gamma_{\ell-1} + \phi_2 \gamma_{\ell-2} \,. \tag{3.1.5}$$

Divide eq. (3.1.5) by $\sqrt{\mathrm{Var}(r_{t-\ell})\mathrm{Var}(r_t)} = \mathrm{Var}(r_t) = \gamma_0$ to convert to autocorrelation function.

$$\rho_\ell = \phi_1 \rho_{\ell-1} + \phi_2 \rho_{\ell-2} \,. \tag{3.1.6}$$

This gives us a second order recursive relation for the autocorrelation function. It's a second-order difference equation. Introducing the lag operator $L\rho_\ell \equiv \rho_{\ell-1}$ we can write this as,

$$(1 - \phi_1 L - \phi_2 L^2)\rho_\ell = 0 \,. \tag{3.1.7}$$

Introduce an ansatz of the form $\rho_\ell = z^\ell$, then

$$\begin{aligned}(1 - \phi_1 L - \phi_2 L^2) z^\ell &= 0 \\ \Rightarrow \left(z^\ell - \phi_1 z^{\ell-1} - \phi_2 z^{\ell-2}\right) &= 0 \\ \Rightarrow z^{\ell-2}(z^2 - \phi_1 z - \phi_2) &= 0 \,.\end{aligned} \tag{3.1.8}$$

Assume $z \neq 0$ to find non-trivial solutions. This yields the characteristic equation

$$z^2 - \phi_1 z - \phi_2 = 0 \,. \tag{3.1.9}$$

Roots of this polynomial determine the asymptotic properties of the autocovariance. The presence of a unit root implies ρ_ℓ grows exponentially with ℓ.

3.2 AR(p) model

An $AR(p)$ time series is stationary *if and only if* its characteristic equation

$$z^p - \phi_1 z^{p-1} - \cdots - \phi_{p-1} z - \phi_p = 0 \tag{3.2.1}$$

has no unit roots $|z_*|^2 < 1$.

Derivation of AIC for AR(p) models

The likelihood of an AR(p) model, to generate T samples $\{r_t : t = 1, 2, ..., T\}$, given p previous values $\{r_t : t = 0, -1, ..., p-1\}$, and assuming the noise term is Gaussian with zero mean and variance σ^2, is

$$\mathcal{L}(\phi) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{1}{2\sigma^2}((1-\phi[L])r_t)^2\right] . \tag{3.2.2}$$

So the log-likelihood is

$$\ln \mathcal{L}(\phi) = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\left[(1-\phi[L])r_t\right]^2 . \tag{3.2.3}$$

Substituting σ^2 and ϕ with their MLEs $\hat{\sigma}^2 = \text{RSS}/(T-p)$, and $\hat{\phi}$ yields

$$\ln \mathcal{L} = -\frac{T}{2}\ln\hat{\sigma}^2 - \frac{1}{2\hat{\sigma}^2}\sum_{t=1}^{T}\left[\left(1-\hat{\phi}[L]\right)r_t\right]^2 \tag{3.2.4a}$$

$$= -\frac{T}{2}\ln\hat{\sigma}^2 - \frac{1}{2\hat{\sigma}^2}\ \text{RSS} \tag{3.2.4b}$$

$$= -\frac{T}{2}\ln\hat{\sigma}^2 - \frac{1}{2\hat{\sigma}^2}[(T-\ell)\hat{\sigma}^2] \tag{3.2.4c}$$

$$= -\frac{T}{2}\ln\hat{\sigma}^2 - \frac{1}{2}(T-\ell) . \tag{3.2.4d}$$

The last term is a constant and can be dropped, since when we use AIC to compare models the constant terms will be the same across models. Hence,

Definition 3.2.1: *Akaike Information Criterion (AIC)* (smaller is better)

$$\text{AIC} \equiv 2 \times (\text{number of parameters}) - 2\ln(\text{likelihood}) . \tag{3.2.5}$$

For AR(p) models the AIC is given by

$$\text{AIC}(p) = \frac{2}{T}p - \frac{2}{T}\mathcal{L} = \frac{2p}{T} + \ln\hat{\sigma}^2 . \tag{3.2.6}$$

3.2 References

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