STAT 153 & 248 - Time Series
Lecture Three

Fall 2025, UC Berkeley

Aditya Guntuboyina

September 4, 2025

1 Simple Linear Regression

We observe data $(x_1, y_1), \ldots, (x_n, y_n)$. In the time series context, $y_1, \ldots, y_n$ is the observed time series. For the covariate, we have two options:

1. $x_i = i$ (covariate is time)
2. $x_i = y_{i-1}$ (covariate is lagged version of the observed time series)

For now, we focus on the first case $x_i = i$ (AutoRegression will be studied in detail later).

The linear regression model is:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{1}$$

where β_0 and β_1 are unknown parameters (their values are to be estimated from the data $(x_1, y_1), \ldots, (x_n, y_n)$) and ϵ_i denotes the error term accounting for the deviation between y_i and the model equation $\beta_0 + \beta_1 x_i$.

We use the function `OLS` in the Python library `statsmodels` to do inference on β_0 and β_1 (inference refers to estimating their values, and also to obtain uncertainty intervals).

There are two main approaches for statistical inference: Frequentist and Bayesian. We shall explore how each of these work for linear regression.

2 Frequentist Inference

The basic ideas behing frequentist inference are as follows:

1. Construct a method for estimating the unknown parameters. In the simple linear regression context, one can use least squares to estimate β_0 and β_1. Probably the most popular estimation strategy is Maximum Likelihood Estimation (this requires writing down a likelihood function).
2. Calculate (exactly or using some approximations) the distribution of the estimators. Use quantiles of the distribution for obtaining interval estimates for the unknown parameters. The quantiles might themselves depend on other unknown parameters (which would then have to be replaced by estimates).

For linear regression, the point estimates are obtained by the method of least squares, where the sum of squares

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{2}$$

is minimized over all values of β_0 and β_1. To minimize $S(\beta_0, \beta_1)$, we take derivatives with respect to β_0, β_1 and equate to zero:

$$\begin{aligned} \frac{\partial}{\partial \beta_0} S(\beta_0, \beta_1) = 0 &\implies \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial}{\partial \beta_1} S(\beta_0, \beta_1) = 0 &\implies \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \end{aligned} \tag{3}$$

It is an exercise to verify that the solution to the above equations is given by:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{and} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \tag{4}$$

where

$$\bar{y} = \frac{y_1 + \cdots + y_n}{n} \quad \text{and} \quad \bar{x} = \frac{x_1 + \cdots + x_n}{n}.$$

We shall refer to (4) as the least squares estimators.

Maximum Likelihood Estimation (MLE) can also be used. To write down the likelihood, one most commonly uses the normality assumption:

$$\epsilon_1, \ldots, \epsilon_n \overset{\text{i.i.d}}{\sim} N(0, \sigma^2). \tag{5}$$

With this normality assumption, the linear regression model can be rewritten as:

$$y_i \overset{\text{independent}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2). \tag{6}$$

The likelihood then becomes:

$$\begin{aligned} f_{y_1,\ldots,y_n|\beta_0,\beta_1,\sigma}(y_1, \ldots, y_n) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right) \\ &= (2\pi)^{-n/2} \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \right) \\ &= (2\pi)^{-n/2} \sigma^{-n} \exp\left(-\frac{S(\beta_0, \beta_1)}{2\sigma^2} \right) \end{aligned} \tag{7}$$

Recall $S(\beta_0, \beta_1)$ above is the sum of squares defined in (2). To write this likelihood, we are assuming that $x_1, \ldots, x_n$ are fixed. This assumption is fine if $x_i = i$ (regression with time as covariate) but not strictly true when $x_i = y_{i-1}$ (auto-regression). We shall see how it is still approximately true in the case of AutoRegression later.

The Maximum Likelihood Estimates of the parameters β_0, β_1, σ are obtained by maximizing the likelihood. As maximizing a function is equivalent to maximizing its logarithm, we attempt to maximize the log-likelihood which leads to an easier maximization. The log-likelihood is:

$$\begin{aligned} \text{log-likelihood} &= \log\left[(2\pi)^{-n/2} \sigma^{-n} \exp\left(-\frac{S(\beta_0, \beta_1)}{2\sigma^2} \right) \right] \\ &= -\frac{n}{2} \log(2\pi) - n \log \sigma - \frac{S(\beta_0, \beta_1)}{2\sigma^2}. \end{aligned}$$

To maximize the log-likelihood, we simply take derivatives with respect to the unknown parameters β_0, β_1, σ and equate those to zero:

$$\frac{\partial}{\partial \beta_0}\text{log-likelihood} = -\frac{1}{2\sigma^2}\frac{\partial}{\partial \beta_0}S(\beta_0,\beta_1) = 0 \implies \frac{\partial}{\partial \beta_0}S(\beta_0,\beta_1) = 0$$
$$\frac{\partial}{\partial \beta_1}\text{log-likelihood} = -\frac{1}{2\sigma^2}\frac{\partial}{\partial \beta_1}S(\beta_0,\beta_1) = 0 \implies \frac{\partial}{\partial \beta_1}S(\beta_0,\beta_1) = 0$$
$$\frac{\partial}{\partial \sigma}\text{log-likelihood} = -\frac{n}{\sigma} + \frac{S(\beta_0,\beta_1)}{\sigma^3} = 0 \implies \sigma = \sqrt{\frac{S(\beta_0,\beta_1)}{n}}.$$

The first two equations coincide with the corresponding equations (3) for minimizing least squares. This shows that the MLEs for β_0 and β_1 coincide with the least squares estimators (4). The MLE for σ is given by the third equation above (with β_0 and β_1 replaced by $\hat{\beta}_0$ and $\hat{\beta}_1$ respectively):

$$\hat{\sigma}_{\text{MLE}} = \sqrt{\frac{S(\hat{\beta}_0,\hat{\beta}_1)}{n}}. \tag{8}$$

The next step is to determine the distribution of the estimators. Let us illustrate this with $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^n (x_i-\bar{x})^2} = \frac{\sum_{i=1}^n y_i(x_i-\bar{x})}{\sum_{i=1}^n (x_i-\bar{x})^2}$$

where the second equality uses $\sum_{i=1}^n \bar{y}(x_i-\bar{x}) = \bar{y}\sum_{i=1}^n (x_i-\bar{x}) = 0$. Under the normality assumption (6), it is easy to check that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i(x_i-\bar{x})}{\sum_{i=1}^n (x_i-\bar{x})^2} \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right).$$

For $\hat{\beta}_0$, one can derive a similar normal distribution (proof omitted):

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i-\bar{x})^2}\right). \tag{9}$$

For $\hat{\sigma}_{\text{MLE}}$, the distribution becomes (proof omitted):

$$\frac{n\hat{\sigma}^2_{\text{MLE}}}{\sigma^2} \sim \chi^2_{n-2} \tag{10}$$

where χ^2_{n-2} is the chi-squared distribution with $n-2$ degrees of freedom. The mean of the chi-squared distribution equals its degrees of freedom which implies that

$$\mathbb{E}\hat{\sigma}^2_{\text{MLE}} = \sigma^2\frac{n-2}{n}.$$

Therefore the MLE for σ^2 is not unbiased (in contrast, the MLEs $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased). It is easy to correct the bias leading to the following unbiased estimator of σ^2:

$$\hat{\sigma}^2_{\text{unbiased}} = \frac{n}{n-2}\hat{\sigma}^2_{\text{MLE}} = \frac{S(\hat{\beta}_0,\hat{\beta}_1)}{n-2}.$$

Usage of $\hat{\sigma}_{\text{unbiased}}$ is much more common than that of $\hat{\sigma}_{\text{MLE}}$ (note that $\hat{\sigma}_{\text{unbiased}}$ is not unbiased for σ; rather the square of $\hat{\sigma}_{\text{unbiased}}$ is unbiased for σ^2).

Another important fact is that $(\hat{\beta}_0,\hat{\beta}_1)$ and $\hat{\sigma}^2_{\text{unbiased}}$ are independent.

These facts are used to derive the following confidence interval for β_1:

$$\left[\hat{\beta}_1 - \frac{\hat{\sigma}_{\text{unbiased}}}{\sqrt{\sum_i (x_i - \bar{x})^2}} t_{n-2,\alpha/2}, \hat{\beta}_1 + \frac{\hat{\sigma}_{\text{unbiased}}}{\sqrt{\sum_i (x_i - \bar{x})^2}} t_{n-2,\alpha/2}\right] \tag{11}$$

where $t_{n-2,\alpha/2}$ is the positive point such that $\mathbb{P}\{t_{n-2} \geq t_{n-2,\alpha/2}\} = \alpha/2$ (i.e., the t-distribution with $n-2$ degrees of freedom assigns probability mass $\alpha/2$ to the right of $t_{n-2,\alpha/2}$). (11) is a valid confidence interval because:

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma_{\text{unbiased}}} \sqrt{\sum_i (x_i - \bar{x})^2} \sim N(0,1) \text{ and } \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\text{unbiased}}} \sqrt{\sum_i (x_i - \bar{x})^2} \sim t_{n-2}$$

where t_{n-2} is the t-distribution with $n-2$ degrees of freedom.

3 Bayesian Inference

Here one treats the unknown parameters also as random variables. The steps are:

1. Assign a probability distribution for the unknown parameters (this represents the **prior**).
2. Treat the likelihood as the conditional probability density for the observed data given the parameters.
3. Use the rules of probability (Bayes rule) to calculate the conditional probability density for the parameters given the observed data (this represents the **posterior**).
4. Use the posterior to answer all inferential questions about the parameters.

Let us do this in the context of our simple linear regression model. the first step is to select a prior for the unknown parameters β_0, β_1, σ. A reasonable prior reflecting ignorance is

$$\beta_0, \beta_1, \log \sigma \overset{\text{i.i.d}}{\sim} \text{Unif}(-C, C)$$

for a large number C (the exact value of C will not matter in the following calculations). Note that as σ is always positive, we have made the uniform assumption on $\log \sigma$ (by the change of variable formula, the density of σ would be given by $f_\sigma(x) = f_{\log\sigma}(\log x)\frac{1}{x} = \frac{I\{-C<\log x<C\}}{2Cx} = \frac{I\{e^{-C}<x<e^C\}}{2Cx}$.

The joint posterior for all the unknown parameters β_0, β_1, σ is then given by (below we write the term "data" for $y_1, \ldots, y_n$):

$$f_{\beta_0,\beta_1,\sigma|\text{data}}(\beta_0, \beta_1, \sigma) \propto f_{y_1,\ldots,y_n|\beta_0,\beta_1,\sigma}(y_1, \ldots, y_n) f_{\beta_0,\beta_1,\sigma}(\beta_0, \beta_1, \sigma).$$

The two terms on the right hand side above are the likelihood:

$$f_{y_1,\ldots,y_n|\beta_0,\beta_1,\sigma}(y_1, \ldots, y_n) \propto \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2\right),$$

and the prior:

$$\begin{aligned} f_{\beta_0,\beta_1,\sigma}(\beta_0, \beta_1, \sigma) &= f_{\beta_0}(\beta_0) f_{\beta_1}(\beta_1) f_\sigma(\sigma) \\ &\propto \frac{I\{-C < \beta_0 < C\}}{2C} \frac{I\{-C < \beta_1 < C\}}{2C} \frac{I\{e^{-C} < \sigma < e^C\}}{2C\sigma} \\ &\propto \frac{1}{\sigma} I\left\{-C < \beta_0, \beta_1, \log\sigma < C\right\}. \end{aligned}$$

We thus obtain

$$f_{\beta_0,\beta_1,\sigma|\text{data}}(\beta_0,\beta_1,\sigma)$$
$$\propto \sigma^{-n-1}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2\right) I\left\{-C<\beta_0,\beta_1,\log\sigma<C\right\}.$$

The above is the joint posterior over β_0,β_1,σ. The posterior over only the main parameters β_0,β_1 can be obtained by integrating (or marginalizing) the parameter σ.

$$f_{\beta_0,\beta_1|\text{data}}(\beta_0,\beta_1)=\int f_{\beta_0,\beta_1,\sigma|\text{data}}(\beta_0,\beta_1,\sigma)d\sigma$$
$$\propto I\{-C<\beta_0,\beta_1<C\}\int_{e^{-C}}^{e^{C}}\sigma^{-n-1}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2\right)d\sigma.$$

When C is large, the above integral can be evaluated from 0 to ∞ which gives

$$f_{\beta_0,\beta_1|\text{data}}(\beta_0,\beta_1)\propto I\{-C<\beta_0,\beta_1<C\}\int_{0}^{\infty}\sigma^{-n-1}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2\right)d\sigma.$$

We will complete this calculation in the next lecture.