STAT 153 & 248 - Time Series
Lecture Two

Spring 2025, UC Berkeley

Aditya Guntuboyina

January 23, 2025

We shall start the first topic of the course: Linear Regression. We shall discuss both frequentist and Bayesian approaches for linear regression. Both approaches end up with identical solutions even though they use very different ideas. We start our discussion with simple linear regression (where there is a single covariate), and then extend to multiple linear regression (where there are multiple covariates).

1 Simple Linear Regression

We observe data $(x_1, y_1), \ldots, (x_n, y_n)$. x_i denotes the covariate value and y_i denotes the response value for the i^{th} observation. In the time series context, in our initial applications, we shall apply linear regression with time as the covariate. For example, in the time series dataset on the population of the United States for each month from January 1959 to December 2024: n denotes the total number of data points, $x_i = i$ and y_i denotes the observed population data for the i^{th} month (first month is January 1959, second month is February 1959 and so on).

In the linear regression model, it is assumed that $x_1, \ldots, x_n$ are fixed deterministic values, and that the response values $y_1, \ldots, y_n$ satisfy the model equation:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \text{with } \epsilon_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2).$$

Another way of writing the model is:

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

There are three parameters in this model: β_0, β_1 and σ^2.

We discuss frequentist and Bayesian approaches for estimating the parameters (as well as uncertainty quantification) from the observed data. A key role in both approaches will be played by the likelihood function which is the joint density of the observations given the

parameter values. The likelihood function is given by:

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

where

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

Note again that we are assuming that $x_1,\ldots,x_n$ are fixed.

2 Frequentist Inference

Frequentist inference is most commonly done via Maximum Likelihood Estimators. The MLEs for β_0,β_1,σ are obtained by maximizing the likelihood. From the expression (1) for the likelihood, the following is a natural strategy for maximizing it: (a) first maximize over β_0,β_1 for fixed σ. This is equivalent to minimizing $S(\beta_0,\beta_1)$ and will lead to the MLEs $\hat{\beta}_0$ and $\hat{\beta}_1$. (b) Plug in the values $\beta_0=\hat{\beta}_0$ and $\beta_1=\hat{\beta}_1$ in (1) and then maximize over σ.

$\hat{\beta}_0$ and $\hat{\beta}_1$ are therefore given by the minimizers of $S(\beta_0,\beta_1)$. It is left as an exercise to verify that

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

where

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

To get the MLE for σ, we need to maximize

$$(2\pi)^{-n/2}\sigma^{-n} \exp\left(-\frac{S(\hat{\beta}_0,\hat{\beta}_1)}{2\sigma^2}\right).$$

It is left as an exercise to show that

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

The quantities $\hat{\beta}_0,\hat{\beta}_1,\hat{\sigma}$ provide point estimates of the unknown parameters β_0,β_1 and σ. More work is needed for uncertainty quantification. For this, one attempts to deduce the distribution of $\hat{\beta}_0,\hat{\beta}_1,\hat{\sigma}$. This can be done in closed form. As an example, for $\hat{\beta}_1$, we have

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

One can also check that, jointly, $\hat{\beta}_0$ and $\hat{\beta}_1$ have the following bivariate normal distribution:

$$\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix} \sim N\left(\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix}, \frac{\sigma^2}{n\sum_{i=1}^{n}(x_i-\bar{x})^2}\begin{pmatrix}\sum_i x_i^2 & -\sum_i x_i\\ -\sum_i x_i & n\end{pmatrix}\right)$$

These formulae are easier to deduce if we use matrix notation (which we shall do when we look at multiple linear regression next week). The distribution of $\hat{\sigma}_{\text{MLE}}$ is given by:

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

where χ^2_{n-2} denotes 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:

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

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 $t_{n-2,\alpha/2}$). (2) is a valid confidence interval because:

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

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

3 Bayesian Inference

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

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

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. We shall do this in the next lecture.