476 Statistics, Spring 2022.
Lecturer: Sergey Nadtochiy.

Lecture 10. Linear regression models: general setting, linear regression in dimension one, LS estimator and its properties. (Sections 11.1–11.4)

1 General setting

Consider a sample of n values of a m-dim. vector, called "independent" (or "explanatory", or "predictor") variable:

$$\{X_i := (X_{1i}, \ldots, X_{mi})\}_{i=1}^n,$$

and the associated values of 1-dim. variable, called "dependent" (or "response") variable:

$$\{Y_i\}_{i=1}^n.$$

(Note that the random vectors $\{X_i\}$ may not be independent in the probabilistic sense.)

The response variable Y represents a value that we would ultimately like to predict: e.g., Y may take two values, 0 or 1, indicating whether a given picture contains a cat. The prediction of the response variable is based on the explanatory variable X, which is a vector of relevant values: e.g., this vector could encode a picture, with each of its entries being a numeric representation of a color of the corresponding pixel. Then, Y_i and X_i are the values observed on the i-th trial (i.e., X_i is a digital representation of the i-th picture, and Y_i equals 0 or 1 depending on whether the i-th picture contains a cat or not). We plan to make n trials, and we treat the outcomes as random because these n trials may be repeated in the future (hence, their outcomes are unknown a priori). Once we build a regression model, we are able to predict the value of Y given a value of X: e.g., we can build a "machine" that tells us whether a given picture contains a cat or not, based on a digital representation of the picture.

Our main goal is to find a functional relationship between X and Y, so that we could compute (simulate) Y once we know X.

More specifically, we assume that there exists a vector of parameters $\beta := (\beta_0, \ldots, \beta_k)$ and a function

$$f(x, \varepsilon; \beta), \quad x \in \mathbb{R}^m, \quad \varepsilon \in \mathbb{R},$$

s.t.

$$Y = f(X, \varepsilon; \beta),$$

where ε is a r.v.'s with mean zero, independent of r.v. X. We refer to ε as "noise" or "residual".

Examples of such dependence include: image recognition; decoding of human genome; weather forecasting; prediction of asset prices and economic downturns, etc.

Q 1. *Given the observed values of $\{(X_i, Y_i)\}$, how to choose the parametric family of functions $\{f(\cdot, \cdot; \beta)\}_\beta$?*

One way is to assume that $f(\cdot, \cdot; \beta)$ has a nonlinear dependence on β and is given by an iterated composition of a chosen "activation" function and linear transformations (whose coefficients are the entries of β). This leads to a "neural network". Neural networks work well in some applications. However, in cases where a perfect fit (i.e., $y_i = f(x_i, 0; \beta)$ for all $i = 1, \ldots, n$) is impossible, the neural network representation makes it very challenging to determine which methods of choosing (estimating) β have the desired "good" properties (e.g., consistency).

Another choice of a family of functions $\{f(\cdot,\cdot;\beta)\}_\beta$, which does allow one to construct estimators with desirable properties, is to consider functions that are linear in β. This leads to “linear regression models”. **In this course, we only deal with linear regression models.**

Def 1. *A linear regression model is given by:*

$$f(x,\varepsilon;\beta) = \beta_0 f_0(x) + \cdots + \beta_k f_k(x) + \varepsilon, \quad x \in \mathbb{R}^m,$$

where the basis functions $\{f_j\}_{j=0}^k$ are fixed.

The simplest example is $k = m = 1$, $f_0(x) = 1$, $f_1(x) = x$: i.e.,

$$f(x,\varepsilon;\beta_0,\beta_1) = \beta_0 + \beta_1 x + \varepsilon,$$

which leads to the **1-dimensional linear regression model**:

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

Rem 1. *Linear regression ($k = m = 1$) is nothing else but conditional expectation:*

$$\mathbb{E}(Y|X) = \beta_0 f_0(X) + \cdots + \beta_k f_k(X).$$

The main question we will analyze in the remainder of this course is the following.

Q 2. *Given the observed values of $\{(X_i,Y_i)\}_{i=1}^n$, how to construct a good estimator of the unknown β? what is the precision of such estimator (e.g., via a confidence interval for β)? and does it even make sense to use X as an explanatory variable for Y (e.g. via a hypothesis test on whether $\beta = 0$)?*

2 Least squares (LS) estimator of β, in a 1-dimensional linear regression model

In this section, we assume $k = m = 1$: i.e., we consider the regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

where ε is a zero-mean r.v.'s, independent of X_i.

Q 3. *Given the observed values of $\{(X_i,Y_i)\}_{i=1}^n$, how to estimate β?*

`Draw the sample points and line` $y = 1 + 0.7 \cdot x$.

One popular estimator is the least squares (LS) estimator $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$, which attain the minimum in

$$\min_{\beta_0,\beta_1 \in \mathbb{R}} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2.$$

To find the LS estimator, we differentiate

$$\sum_{j=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2 = \sum_{i=1}^n (Y_i^2 + \beta_0^2 + \beta_1^2 X_i^2 - 2Y_i\beta_0 - 2Y_i\beta_1 X_i + \beta_0\beta_1 X_i)$$

$$= \sum_{i=1}^{n} Y_i^2 + n\beta_0^2 + \beta_1^2 \sum_{i=1}^{n} X_i^2 - 2\beta_0 \sum_{i=1}^{n} Y_i - 2\beta_1 \sum_{i=1}^{n} X_i Y_i + \beta_0\beta_1 \sum_{i=1}^{n} X_i$$

w.r.t. β_0 and β_1 and set these two derivatives to zero.

The resulting estimators are

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X},$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} =: \frac{S_{xy}}{S_{xx}}.$$

A justification of this method is that $\hat{\beta}$ is the MLE for β, if the residuals $\{\varepsilon_i\}$ are normal.

Thm 1. *Assume that $\{\varepsilon_i\}$ are zero-mean i.i.d. normal r.v.'s independent of $\{X_i\}$. Then, the LS estimator $\hat{\beta}$ is the MLE of β.*

Exercise 1. *Assuming, in addition, that $\{X_i\}$ are i.i.d., prove the above theorem.*

Ex 1. *Assume a 1-dimensional linear regression model and consider the following observations:*

$$\{x_i\} : -2, -1, 0, 1, 2$$

$$\{y_i\} : 0, 0, 1, 1, 3$$

Q 4. *Compute the LS estimator $\hat{\beta}$.*

$$\bar{x} = 0, \quad \bar{y} = 1,$$

$$\hat{\beta}_1 = \frac{\sum_{j=1}^{5} x_j(y_j - 1)}{\sum_{j=1}^{5} x_j^2} = \frac{7}{10} = 0.7,$$

$$\hat{\beta}_0 = 1 - 0 = 1$$

`Draw the sample points and line` $y = 1 + 0.7 \cdot x$`.`

2.1 Properties of LS estimator

Denote $\sigma^2 := V(\varepsilon_j)$ and recall that we always assume that $\{\varepsilon_i\}$ are i.i.d. and independent of $\{X_i\}$. Another useful observation is

$$\bar{Y} = \beta_0 + \beta_1 \bar{X} + \bar{\varepsilon}.$$

First, we investigate the **bias** of $\hat{\beta}$. By a simple transformation, we have

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{S_{xx}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(\beta_1 X_i + \varepsilon_i)}{S_{xx}}$$

$$= \beta_1 \frac{\sum_{i=1}^{n}(X_i - \bar{X})X_i}{S_{xx}} + \sum_{i=1}^{n} \frac{X_i - \bar{X}}{S_{xx}} \varepsilon_i = \beta_1 \frac{\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})}{S_{xx}} + \sum_{i=1}^{n} \frac{X_i - \bar{X}}{S_{xx}} \varepsilon_i$$

$$= \beta_1 + \sum_{i=1}^{n} \frac{X_i - \bar{X}}{S_{xx}} \varepsilon_i.$$

Using the above, we obtain

$$\mathbb{E}\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} \mathbb{E}\frac{X_i - \bar{X}}{S_{xx}} \mathbb{E}\varepsilon_i = \beta_1.$$

Hence, $\hat{\beta}_1$ is **unbiased**.

From the above, it is also easy to see that

$$\mathbb{E}\hat{\beta}_0 = \mathbb{E}(\bar{Y} - \hat{\beta}_1 \bar{X}) = \beta_0 + \beta_1 \mathbb{E}\bar{X} - \mathbb{E}\left[\beta_1 \bar{X} + \sum_{i=1}^{n} \frac{(X_i - \bar{X})\bar{X}}{S_{xx}} \varepsilon_i\right] = \beta_0.$$

Hence, $\hat{\beta}_0$ is **unbiased**.

Thus, we have proved the following theorem.

Thm 2. *If $\{\varepsilon_i\}$ are zero-mean i.i.d. and independent of $\{X_i\}$, the LS estimator $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ is unbiased.*

Next, we analyze the question of **consistency** of the LS estimator $\hat{\beta}$. Recall that $\hat{\beta}$ is a MLE of β if $\{\varepsilon_i\}$ are normal, hence, we expect $\hat{\beta}$ to be consistent, at least if $\{\varepsilon_i\}$ are normal. However, it is important to notice that the consistency of MLE relies on the assumption that the sample $\{X_i, Y_i\}$ is i.i.d.. Thus, the consistency of $\hat{\beta}$ holds under the additional assumption that $\{X_i\}$ are i.i.d..

Thm 3. *Assume that $\{\varepsilon_i\}$ are zero-mean i.i.d. and independent of $\{X_i\}$. If, in addition, $\{X_i\}$ are i.i.d. with a finite mean, then $\hat{\beta}$ is consistent.*

Exercise 2. *Prove the above theorem.*

The assumption of i.i.d. $\{X_i\}$ can be relaxed to some extent, but it is important to remember that there are many cases of $\{X_i\}$ that appear in real-world applications, for which the consistency of $\hat{\beta}$ does not hold. Clearly, LS estimator is not a good option for a linear regression is such cases (perhaps, the linear regression itself is not a good model in such cases).

The next question is: how to construct a **confidence interval** for β? Recall that a confidence interval is better than a point estimator as the former provides information about the precision of our estimation. Following the general methodology, we can construct a confidence interval using the pivot

$$\hat{\beta}_j - \beta_j, \quad j = 0, 1,$$

provided we know that distribution of $\hat{\beta}_j$ (and that it does not depend on the unknown parameters).

It turns out that we can find the distribution of $\hat{\beta}_i$ under the assumption of deterministic $\{X_i\}$ and normal $\{\varepsilon_i\}$.

Thm 4. *Assume that $\{\varepsilon_i\}$ are i.i.d. and that $\{X_i = x_i\}$ are deterministic. Then,*

$$V(\hat{\beta}_0) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n S_{xx}}, \quad V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}, \quad cov(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \frac{\bar{x}}{S_{xx}},$$

$$V(a\hat{\beta}_0 + b\hat{\beta}_1) = \sigma^2 \sigma_{ab}^2, \quad \sigma_{ab}^2 := \frac{\frac{a^2}{n}\sum_{i=1}^{n} x_i^2 + b^2 - 2ab\bar{x}}{S_{xx}}, \quad \textit{for any } a, b \in \mathbb{R}.$$

Exercise 3. *Prove the above theorem.*

Thm 5. *Assume that $\{\varepsilon_i\}$ are i.i.d.* **normal** *and that $\{X_i = x_i\}$ are deterministic. Then, for any $a, b \in \mathbb{R}$, the linear combination $a\hat{\beta}_0 + b\hat{\beta}_1$ is* **normal**.

The two theorems above can be used to construct confidence intervals for the unknown $\beta = (\beta_0, \beta_1)$.

Ex 2. *Assume a 1-dimensional linear regression model and consider the following observations:*

$$\{x_j\} : -2, -1, 0, 1, 2$$

$$\{y_j\} : 0, 0, 1, 1, 3$$

Q 5. *Assuming that $\sigma^2 = V(\varepsilon_j)$ is known, that $\{\varepsilon_i\}$ are normal and that $\{X_i = x_i\}$ are deterministic, construct a confidence interval for β_1 of confidence level 0.95.*

Recall

$$S_{xy} = 7, \quad S_{xx} = 10, \quad \hat{\beta}_1 = 0.7, \quad \hat{\beta}_0 = 1 - 0 = 1.$$

Using the above theorems, we obtain:

$$V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}} = \frac{\sigma^2}{10}, \quad \hat{\beta}_1 \sim N(\beta_1, \sigma^2/10).$$

Then, we can use the pivot

$$G = \sqrt{10}\frac{\hat{\beta}_1 - \beta_1}{\sigma} \sim N(0,1)$$

and follow the known strategy:

$$a \leq \beta_1 \leq b \quad \text{if and only if} \quad \sqrt{10}\frac{\hat{\beta}_1 - b}{\sigma} \leq G = \sqrt{10}\frac{\hat{\beta}_1 - \beta_1}{\sigma} \leq \sqrt{10}\frac{\hat{\beta}_1 - a}{\sigma}.$$

Then, to ensure that the confidence interval is of level 0.95, we need

$$0.95 = \mathbb{P}\left(\sqrt{10}\frac{\hat{\beta}_1 - b}{\sigma} \leq G = \sqrt{10}\frac{\hat{\beta}_1 - \beta_1}{\sigma} \leq \sqrt{10}\frac{\hat{\beta}_1 - a}{\sigma}\right),$$

which leads to

$$\sqrt{10}\frac{\hat{\beta}_1 - a}{\sigma} = z_{0.025}, \quad \sqrt{10}\frac{\hat{\beta}_1 - b}{\sigma} = -z_{0.025},$$

and to the confidence interval

$$[a, b] = [\hat{\beta}_1 \pm z_{0.025}\sigma/\sqrt{10}] = [0.7 \pm z_{0.025}\sigma/\sqrt{10}].$$

3 Estimating the variance of residuals, σ^2

Q 6. *How to estimate σ^2?*

A natural estimator is

$$\frac{1}{n-1}\sum_{i=1}^{n}\varepsilon_i^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\beta_0-\beta_1 X_i)^2$$

However, we do not observe β. It turns out that, by replacing the exact coefficients with their estimates, with the adjustment of normalization factor, we still obtain a good estimator of σ^2.

Thm 6. *If $\{\varepsilon_i\}$ are i.i.d. and independent of $\{X_i\}$, then*

$$\tilde{S}^2 := \frac{1}{n-2}\sum_{i=1}^{n}(Y_i-\hat{\beta}_0-\hat{\beta}_1 X_i)^2$$

is an unbiased estimator of σ^2. If, in addition, $\hat{\beta}$ is consistent, then $\tilde{S}^2$ is also consistent.

Proof:

We need the following auxiliary lemma.

Lemma 1.

$$\tilde{S}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left((\varepsilon_i-\bar{\varepsilon})^2-(X_i-\bar{X})^2\varepsilon_i^2/S_{xx}\right).$$

Exercise 4. *Prove the above lemma.*

Using he above lemma, it is easy to show that $\tilde{S}^2$ is unbiased:

$$\mathbb{E}\tilde{S}^2 = \frac{n-1}{n-2}\mathbb{E}\frac{1}{n-1}\sum_{i=1}^{n}(\varepsilon_i-\bar{\varepsilon})^2-\frac{1}{n-2}\mathbb{E}\sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{S_{xx}}\varepsilon_i^2$$

$$=\frac{n-1}{n-2}\sigma^2-\frac{1}{n-2}\sum_{i=1}^{n}\mathbb{E}\frac{(X_i-\bar{X})^2}{S_{xx}}\sigma^2=\frac{n-1}{n-2}\sigma^2-\frac{\sigma^2}{n-2}\mathbb{E}\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{S_{xx}}$$

$$=\frac{\sigma^2(n-1)}{n-2}-\frac{\sigma^2}{n-2}=\sigma^2.$$

Consistency is easy to show via LLN and Chebyshev's inequality, using the above representation of $\tilde{S}^2$. ∎

Exercise 5. *Complete the above proof by showing the consistency of $\tilde{S}^2$ as an estimator for σ^2.*

Recall that, for constructing confidence intervals, it is important to know the distribution of the estimator $\tilde{S}^2$.

Thm 7. *If $\{\varepsilon_j\}$ are i.i.d. $N(0,\sigma^2)$, independent of $\{X_i\}$, then:*

$$\frac{n-2}{\sigma^2}\tilde{S}^2 \sim \chi^2(n-2),$$

and $\tilde{S}^2$ is independent of $\hat{\beta}$. As a consequence, for any $a, b \in \mathbb{R}$,

$$\frac{a\hat{\beta}_0 + b\hat{\beta}_1 - a\beta_0 - b\beta_1}{\tilde{S}\sigma_{ab}} \sim T(n-2),$$

where we recall $\sigma_{ab}^2 = \frac{\frac{a^2}{n}\sum_{i=1}^n X_i^2 + b^2 - 2ab\bar{X}}{S_{xx}}$.

Ex 3. *Assume a 1-dimensional linear regression model and consider the following observations:*

$$\{x_i\} : -2, -1, 0, 1, 2$$

$$\{y_i\} : 0, 0, 1, 1, 3$$

Assume that $\sigma^2 = V(\varepsilon_i)$ is **not known**.

Q 7. *Estimate σ^2 and construct a confidence interval of level 0.95 for σ^2.*

Recall

$$\hat{\beta}_1 = 0.7, \quad \hat{\beta}_0 = 1,$$

and compute the value of the estimator proposed above:

$$\tilde{s}^2 = \frac{1}{5-2}\sum_{i=1}^{5}(y_i - 1 - 0.7 \cdot x_i)^2 \approx 0.367.$$

To construct a confidence interval, we use the pivot

$$G = \frac{3}{\sigma^2}\tilde{S}^2 \sim \chi^2(3)$$

and follow the known strategy:

$$a \leq \sigma^2 \leq b \quad \textit{if and only if} \quad \frac{3}{b}\tilde{S}^2 \leq G = \frac{3}{\sigma^2}\tilde{S}^2 \leq \frac{3}{a}\tilde{S}^2.$$

Thus, we need

$$0.95 = \mathbb{P}(\frac{3}{b}\tilde{S}^2 \leq G \leq \frac{3}{a}\tilde{S}^2),$$

which leads to

$$\frac{3}{a}\tilde{S}^2 = \chi^2_{0.025}, \quad \frac{3}{b}\tilde{S}^2 = \chi^2_{0.975}$$

and to the confidence interval

$$[a, b] = \left[\frac{3}{\chi^2_{0.025}}\tilde{s}^2,\ \frac{3}{\chi^2_{0.975}}\tilde{s}^2\right] = \left[\frac{3}{\chi^2_{0.025}}0.367,\ \frac{3}{\chi^2_{0.975}}0.367\right].$$

Q 8. *Assuming that $\{\varepsilon_i\}$ are normal, construct a confidence interval for β_1 of confidence level 0.95.*

Recall

$$S_{xy} = 7, \quad S_{xx} = 10, \quad \hat{\beta}_1 = 0.7, \quad \hat{\beta}_0 = 1 - 0 = 1.$$

Next, for $a = 0$*,* $b = 1$*, we compute*

$$\sigma_{01}^2 = \frac{1}{S_{xx}} = 1/10$$

and apply the above theorem, to conclude:

$$G = \frac{\hat{\beta}_1 - \beta_1}{\tilde{s}\sigma_{01}} \sim T(n-2)$$

and obtain the confidence interval

$$[a, b] = [\hat{\beta}_1 \pm t_{0.025}\tilde{s}\sigma_{01}] = [0.7 \pm t_{0.025}\sqrt{0.367/10}].$$