Linear Regression

Yutong Wang
London School of Economics

Oct 2023

Contents

1 Data Generating Process - Philosophy

One might have seen this formula thousands of times, since it is the simplest machine-learning model, as well as a widely-used methodology in many subjects that claim they are 'quantitative'. However, it is not as easy to understand it as applying it with a few clicks in E-views/Stata/R/Python/MATLAB, etc.

Here I take the 1-dimensional linear regression model to introduce the technique to deal with it, as well as the underlying philosophy.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i. \tag{1.1}$$

One could treat this population regression function as a data generating process (DGP). Data generating process could be interpreted as a process that super-nature powers (for brevity, I use 'God' to represent it.[1]) use 'magic' ways (or whatever similar in one's mind) to determine the state of the world.

For instance, in a specific data generating process 1.1, to determine y_i, the God has four things to do:

1. Find what is the value of parameters (here the parameter is (β_0, β_1)) for this specific situation in his/her/their[2] dictionary;
2. Observe object i, and obtain the value of x_i;
3. Draw a ε_i from a distribution (which is similar to generating random numbers in our computer, however, the difference is also apparent: we do not know his/her random seed, and even his/her random number generating algorithm);
4. According to the formula 1.1, add β_0, $\beta_1 x_i$ and ε_i together to determine the value of y_i.

Remark 1. For different problems, God may choose different parameters (different in their values as well as the dimensions) in (1) and the distribution of error terms ε in (3). For example, in the case where y_i represents the final score for students at LSE, and x_i represents how many hours they study per week in average, God chooses parameter (θ_0, θ_1) and distribution F to generate the data. In another case, where y_i represents whether one prefers noodles over rices, x_i represents whether one comes from the north of China, parameter (α_0, α_1) and distribution G are applied. We generally assume $(\theta_0, \theta_1) \neq (\alpha_0, \alpha_1)$, and $F \neq G$, which coincides with the intuition.

Remark 2. Here we assume that there is no measurement error, *i.e.* , x_i is the true value of the feature of the object i. This could also be understood as we put the measurement error in the ε, however, such an interpretation allows the ε to have more components, which is unnecessarily more complex.

Remark 3. Given the current technology level and the computing power, it is impossible to know the value of β and each ε_i for any i. Intuitively the β represents the truth of the world, which we would never know. It is somehow disappointing, however, thanks to theorems such as the Law of Large Numbers, we can approximate the truth if we have a sufficiently large amount of data. After all, in most scenarios, we do not need to be extremely accurate (except in astronautical engineering, etc.)

Remark 4. It is merely one way to interpret the data-generating process. I introduce it since I found it compatible with and helpful in dealing with mathematical details. I am NOT suggesting you change your beliefs.

[1] This does not imply that I believe in God or not, or suggest anyone believe in it, or suggest anyone not believing in it.
[2] I do not specify the gender, and you could change the order for those pronounces as you like.

2 About Reference

Most of the materials come from 'my mind' - which is not true. In fact, those basic elements, which are rather common and can be seen anywhere, might come from the textbook I used when I did my undergrad. I believe that I have fully internalized them so that I can arrange the structure to present, and that's the reason I do not cite these books.

There are also some interesting ideas, which may come from some online discussion, such as in Quora or Mathematics Stack Exchange long time ago, so that I am tired to check it out.

In addition, the notes are open access, which does not involve any commercial activity.

For the rest which I found useful recently, I list them here for those interested. I will also cite them in particular sections.

For econometrics, I mainly refer Hansen (2022), and for machine learning, some subsections are from Hastie et al. (2009), which can be freely downloaded here (shared by the author).

3 Simple Linear Regression

3.1 Assumptions

It should be noticed that, regarding whether regressors (a.k.a. features[3]) x_i as random, regression models can be divided two types: fixed design and random design. These two types of models naturally have different assumptions.

3.1.1 Fixed Design

Under fixed design, we treat regressors x_i as fixed, which may be because we have control of x_i. For instance, in the case to find the treatment effect $\mathbb{E}[y_i|x_i = 1]$, where we could fully decide which individual would receive treatment and which not, implying that x_i is not random.

Assumption 1. (Zero-mean Error) $\mathbb{E}[\varepsilon_i] = 0$.

Assumption 2. (Homoscedasticity) $\mathrm{Var}(\varepsilon_i) = \sigma^2$.

Assumption 3. (Uncorrelated Sample) $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j$.

Assumption 4. (Fixed Design) *x_i is not r.v., and takes at least two values.*

Assumption 5. (Normal Error) $\varepsilon_i \sim N(0, \sigma^2)$.

Note that, combining Assumption 3 and Assumption 5 gives $\varepsilon_i \sim N(0, \sigma^2)$ *i.i.d.*, since any uncorrelated normal r.v. are independent. In Assumption 2, 'homoscedasticity' means the same variance, while the counter concept is 'heteroskedasticity', which means the error terms have different variances.

3.1.2 Random Design*

Under random design, we treat regressors x_i as random, intuitively meaning that we can not choose on which value of x_i we observe the corresponding y_i. For example, if our goal is to analyse the effect of an individual's income (x_i) on the party (Demo or Repub) that he is in support for (y_i), and we sample by randomly selecting a phone number, call and get information through the conversation. In this case, we have no control of the income (x_i) of the person before we call him, hence should treat x_i as random.

Note that in this case, x_i is a r.v. and also has its own distribution.

Assumption 6. (Zero-mean Error*) $\mathbb{E}[\varepsilon_i|x_i] = 0$.

Assumption 7. (Homoscedasticity*) $\mathrm{Var}(\varepsilon_i|x_i) = \sigma^2$.

Assumption 8. (Uncorrelated Sample*) $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j$.

Assumption 9. (Random Design*) *x_i is not correlated to ε_i (This is actually implied by $\mathbb{E}[\varepsilon_i|x_i] = 0$), and takes at least two values.*

Assumption 10. (Normal Error*) $\varepsilon_i \sim N(0, \sigma^2)$.

There is no essential difference in theory between fixed design and random design in the sense that all proofs for the random design case could be obtained from the fixed design case by replacing $\mathbb{E}[\cdot]$ by $\mathbb{E}[\cdot|x_i]$. Also,

In the main part of this tutorial, we assume that the model is a fixed design for simplicity, except in chapter 3.3.

[3] 'Regressor' is preferred in economics and econometrics, while 'feature' is mainly used by the machine learning community.

3.2 Sample Regression Function - Estimation of β

3.2.1 Normal Regression

If we know the distribution of ε (*i.e.* , we have the Assumption 5), then the conditional probability $\mathrm{P}(y|x)$ is perfectly known. In the econometrics community, it is called the parametric model. For a parametric model, the maximum likelihood estimation (MLE) method is available. Maximum likelihood estimation is praised for some desired properties (Check it out for those interested.)

The basic idea for MLE coincides with our intuition very well: the parameter in the data-generating process should take the value which makes the probability of seeing the data in hand with the largest probability.

$$\hat{\beta}^{\mathrm{MLE}} \in \arg\max_{\beta}\{\log f(\varepsilon_1, \cdots, \varepsilon_n; \beta)\},$$

where $f(\varepsilon_1, \cdots, \varepsilon_n; \beta)$ is the joint density of r.v.s' $\varepsilon_1, \cdots, \varepsilon_n$, which depends on the parameter β.

Here we apply maximum likelihood estimation, and compare the result with the OLS estimator later. (As far as I know, 'Showing MLE and OLS gives the same estimator' has been used as an interview question for Optiver's quantitative researcher position.)

Since $\varepsilon_1, \cdots, \varepsilon_n$ are independent, we have:

$$\begin{aligned}
\log f(\varepsilon_1, \cdots, \varepsilon_n; \beta) &= \log \prod_{i=1}^{n} f(\varepsilon_i; \beta) \\
&= \sum_{i=1}^{n} \log f(\varepsilon_i; \beta) \\
&= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}{\varepsilon_i}^2} \\
&= \sum_{i=1}^{n} \left(\frac{1}{2}\log(2\pi) - \frac{1}{2}{\varepsilon_i}^2\right) \\
&= \frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n} {\varepsilon_i}^2 \\
&= \frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.
\end{aligned}$$

As a result, maximising $\log f(\varepsilon_1, \cdots, \varepsilon_n; \beta)$

3.2.2 Ordinary Least Square

Once we know there is a data generating process, our interests lie in the value of parameters, and some properties of the distribution of ε (such as the variance $\mathrm{Var}(\varepsilon)$[4]).

One of the information sources used for this purpose (estimation of parameters) that we have access to, is the data (a.k.a. 'sample', or 'realisations'[5]). Suppose now through observation, we have data $\{(x_1, y_1), \cdots, (x_n, y_n)\}$.

[4]Note that the mean and the variance are actually defined for a distribution, rather than a random variable, since any r.v.s with the same distribution have the same mean and variance.

[5]My feeling: 'sample' is mainly used in classic statistics, while 'data' is used in the machine learning community, and 'realisations' is used in both econometrics and statistics, to emphasis both two properties of the object: as a value, and as a random variable.

Generally, once given the data per se, and the data generating process, we have a bunch of methods to estimate the parameter. Naturally, different methods have different properties. The estimation method used in 'Linear Regression' is called Ordinary Least Square (OLS)[6].

Definition 1. Ordinary Least Square (OLS)
Given sample $(x_1, y_1), \cdots, (x_n, y_n)$*, and the data generating process 1.1, the OLS estimator of* (β_0, β_1) *is defined by:*

$$\hat{\beta}^{OLS} = (\hat{\beta}_0^{OLS}, \hat{\beta}_1^{OLS}) \in \arg\min_{(\beta_0,\beta_1)\in\mathbb{R}^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2. \tag{3.2}$$

As this is an introductory level course, we will not go into details about how to evaluate an estimator, however, one should know that this is a core content in classic statistics. Jargon is provided for students who are interested in unbiased, consistent, and efficient. Also, some assumptions in the data-generating process are needed for the estimator to have these nice properties, such as exogeneity ($\mathbb{E}[\varepsilon_i|x_i] = 0$).

We would give explicit expression for the solution to 1. This optimisation problem could be solved by taking derivative of $f(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$, since it is differentiable $\forall(\beta_0, \beta_1) \in \mathbb{R}^2$.

$$\begin{aligned}
\frac{\partial f}{\partial \beta_0} &= \sum_{i=1}^{n} -2(y_i - \beta_0 - \beta_1 x_i) = 0, \\
\frac{\partial f}{\partial \beta_1} &= \sum_{i=1}^{n} -x_i(y_i - \beta_0 - \beta_1 x_i) = 0.
\end{aligned}$$

This is a 2-dimensional linear system of (β_0, β_1), with the solution being:

$$\beta_1^* = \frac{\sum_{i=1}^{n}(y_i - \overline{y_n})(x_i - \overline{x_n})}{\sum_{i=1}^{n}(x_i - \overline{x_n})^2}, \tag{3.3}$$

$$\beta_0^* = \overline{y_n} - \beta_1^* \overline{x_n}. \tag{3.4}$$

Remark 1. Here we write the solution as β_0^* and β_1^*, since we are simply solving an optimization problem, which has nothing to do with randomness. Now let us take the randomness nature of $(x_i, y_i)_{i=1}^n$ into consideration, set $\hat{\beta}_0^{\text{OLS}} = \beta_0^*$, $\hat{\beta}_1^{\text{OLS}} = \beta_1^*$, we obtain the expression of OLS estimator[7].

Remark 2. To be mathematically rigorous, one should also check the second-order derivative to show that (β_0^*, β_1^*) is the minimum point rather than the maximum point.

3.3 Memorise the OLS Solution

At the first glance, 3.3 and 3.4 seem a little bit difficult to memorise. In this section, we would provide a natural way (as a 'plug-in' estimator) to understand and find them.

On one hand, it could be obtained from the 'plug-in' principle, which is widely used in econometrics to construct new estimators. With the Continuous Mapping Theorem, we know that nice properties from the original estimators will be kept.

On the other hand, as a way to interpret the regression formula, it helps solve quick questions in linear regression (such as the question proposed by quant interviewers).

[6] If we have assumption 5 (or 10 for random design case), the model is called parametric model, and the maximum likelihood estimation (MLE) could be applied. However, the results from MLE are the same as the one from OLS estimate.

[7] Recall that estimators are functions of the sample, hence random variables.

3.3.1 Estimation Strategies Based on 'Plug-in' Principle

In (1.1), on one hand, taking expectation on both sides gives: $\mathbb{E}[y_i] = \beta_0 + \beta_1 \mathbb{E}[x_i] + \mathbb{E}[\varepsilon_i]$. According to Assumption 1 that $\mathbb{E}[\varepsilon_i] = 0$, we have:

$$\mathbb{E}[y_i] = \beta_0 + \beta_1 \mathbb{E}[x_i].$$

On the other hand, notice that $\mathrm{Cov}(\cdot, \cdot)$ is a bi-linear function, and if we have the exogeneity assumption ($\mathbb{E}[\varepsilon_i | x_i] = 0$), which implies $\mathrm{Cov}(x_i, \varepsilon_i) = 0$ (apply Law of Iterated Expectation to see this), then:

$$\begin{aligned} \mathrm{Cov}(x_i, y_i) &= \mathrm{Cov}(x_i, \beta_0 + \beta_1 x_i + \varepsilon_i) \\ &= \mathrm{Cov}(x_i, \beta_0) + \mathrm{Cov}(x_i, \beta_1 x_i) + \mathrm{Cov}(x_i, \varepsilon_i) \\ &= 0 + \beta_1 \mathrm{Var}(x_i) + 0 \\ &= \beta_1 \mathrm{Var}(x_i). \end{aligned}$$

Thus, in population level, we have,

$$\beta_1 = \frac{\mathrm{Cov}(x_i, y_i)}{\mathrm{Var}(x_i)}, \tag{3.5}$$

$$\beta_0 = \mathbb{E}[y_i] - \beta_1 \mathbb{E}[x_i]. \tag{3.6}$$

Recall the 'plug-in' principle: when we do not know the value of a (population level) parameter, but need it for further use, we could first estimate it, and then pretend that we 'know' it.

Apply 'plug-in' principle here, we could first estimate $\mathrm{Cov}(x_i, y_i)$ and $\mathrm{Var}(x_i)$ by $\widehat{\mathrm{Cov}}(x_i, y_i)$ and $\widehat{\mathrm{Var}}(x_i)$, then by 3.5 we naturally construct an estimator by plugging-in $\widehat{\mathrm{Cov}}(x_i, y_i)$ to replace $\mathrm{Cov}(x_i, y_i)$:

$$\widehat{\beta}_1 = \frac{\widehat{\mathrm{Cov}}(x_i, y_i)}{\widehat{\mathrm{Var}}(x_i)}.$$

If further we have estimator $\widehat{\mathbb{E}}[y_i]$ and $\widehat{\mathbb{E}}[x_i]$ for parameter $\mathbb{E}[y_i]$ and $\mathbb{E}[x_i]$, we could construct an estimate of β_0 by:

$$\widehat{\beta}_0 = \widehat{\mathbb{E}}[y_i] - \widehat{\beta}_1 \widehat{\mathbb{E}}[x_i].$$

3.3.2 OLS as 'Plug-in' Estimators

Given the philosophy above, we indeed have some natural estimate of population mean, variance, and covariance:

- Sample mean (of x): $\widehat{\mathbb{E}}[x_i] = \frac{1}{n} \sum_{i=1}^n x_i =: \overline{x_n}$;
- Sample mean (of y): $\widehat{\mathbb{E}}[y_i] = \frac{1}{n} \sum_{i=1}^n y_i =: \overline{y_n}$;
- Sample variance: $\widehat{\mathrm{Var}}(x_i) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \overline{x_n})^2$;
- Sample covariance: $\widehat{\mathrm{Cov}}(x_i, y_i) = \frac{1}{n-1} \sum_{i=1}^n (y_i - \overline{y_n})(x_i - \overline{x_n})$.

In fact, if we take these natural estimators[7], and construct estimators as above, we could get the OLS estimators (Ckeck!). This tells us, the OLS estimator could be seen as a 'plug-in' type estimator, if we choose a proper estimate of the population mean, variance, and covariance to plug in.

I believe these natural estimators do not take time to memorise, since one should be familiar with sample mean, sample variance, sample covariance, etc. Basically, to remember the structure of OLS estimator, one needs to memorise the population version 3.6 and 3.5. The former is obtained by computing the covariance between x_i and y_i, while the latter is by taking expectation on both sides of 1.1.

[7] Divided by $(n-1)$ rather than n.

3.4 Fitted Value and Residuals

By utilising the data, now we have the estimate of parameters $\hat{\beta}_0, \hat{\beta}_1$. With those, we could define fitted value[8] $\hat{y}_i$ by:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad \forall i = 1, \cdots n. \tag{3.7}$$

Remark. Fitted value could be defined for any estimated parameter (not only OLS estimate), which is why we do not use the notation $\hat{\beta}_0^{\text{OLS}}$.

For $i = 1, \cdots n$, once we have the fitted value, the residual $\hat{\varepsilon}_i$ is defined as[8]:

$$\begin{aligned} \hat{\varepsilon}_i &:= y_i - \hat{y}_i \\ &= (\beta_0 + \beta_1 x_i + \varepsilon_i) - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \\ &= (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1) x_i + \varepsilon_i. \end{aligned} \tag{3.8}$$

For interpretation in practice, *i.e.* , the viewpoint that we want to use x to explain y, residuals represent the part of the information that cannot be explained by the model (x). From a statistics point of view, $\hat{\varepsilon}_i$ is an estimate of ε_i, since the latter is not observable. Moreover, we calculate the arithmetic average of residuals which will be used later:

$$\begin{aligned} \overline{\hat{\varepsilon}_n} &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\Big((\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)x_i + \varepsilon_i\Big) \\ &= (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)\overline{x_n} + \overline{\varepsilon_n} \\ &= \Big(\beta_0 - (\overline{y_n} - \hat{\beta}_1\overline{x_n})\Big) + (\beta_1 - \hat{\beta}_1)\overline{x_n} + \overline{\varepsilon_n} \\ &= \Big(\beta_0 - (\beta_0 + \beta_1\overline{x_n} + \overline{\varepsilon_n} - \hat{\beta}_1\overline{x_n})\Big) + (\beta_1 - \hat{\beta}_1)\overline{x_n} + \overline{\varepsilon_n} \\ &= 0. \end{aligned} \tag{3.9}$$

One should be able to distinguish $\hat{\varepsilon}_i$ and ε_i. The first one is an estimate of the second. As for their arithmetic average[9], note that $\overline{\hat{\varepsilon}_n} = 0$ for sure, but $\overline{\varepsilon_n} = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i$ is a random variable, and takes 0 with 0 probability generally.[9]

Lemma 1. *For two sequences of real number $\{a_i\}_{i=1}^m$, $\{b_i\}_{i=1}^m$ with the same length m, defining arithmetic average $\overline{a_m} = \frac{1}{m}\sum_{i=1}^{m} a_i$, and $\overline{b_m} = \frac{1}{m}\sum_{i=1}^{m} b_i$ then we have:*

$$\sum_{i=1}^{m}(a_i - \overline{a_m})(b_i - \overline{b_m}) = \sum_{i=1}^{m} a_i(b_i - \overline{b_m}) = \sum_{i=1}^{m}(a_i - \overline{a_m})b_i.$$

Note that we use the notation a, b, m to emphasize that this is a basic algebra result, which does not require any assumptions related to randomness.

[8] In fact, for any estimate of parameter, in any data generating process, we could define fitted value. Here we only consider the OLS estimator. In general setting, we use $f(x)$ to denote the data generating process, and $\hat{f}(x)$ to represent the estimator.

[8] I would like to emphasize that these concepts are defined sequentially.

[9] One should fully understand 'mean', 'expectation' and '(arithmetic) average'

[9] In fact, $\mathrm{P}(\overline{\varepsilon_n} \neq 0) = 1$, since $\overline{\varepsilon_n} \sim N(0, \frac{\sigma^2}{n})$ follows an absolutely continuous probability distribution, which takes any value x with probability 0.

Let's look at the sample covariance between $\{x_1, \ldots, x_n\}$ and the residual $\{\varepsilon_1, \ldots, \varepsilon_n\}$:

$$\begin{aligned}
\sum_{i=1}^{n} x_i \hat{\varepsilon_i} &= \sum_{i=1}^{n} x_i(\hat{\varepsilon_i} - \overline{\varepsilon_n}) \\
&= \sum_{i=1}^{n} (x_i - \overline{x_n})\hat{\varepsilon_i} \\
&= \sum_{i=1}^{n} (x_i - \overline{x_n})\Big((\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)x_i + \varepsilon_i\Big) \\
&= \sum_{i=1}^{n} (x_i - \overline{x_n})\Big(\beta_0 - \overline{y_n} + \hat{\beta}_1\overline{x_n} + (\beta_1 - \hat{\beta}_1)x_i + \varepsilon_i\Big) \\
&= \sum_{i=1}^{n} (x_i - \overline{x_n})\Big(\beta_0 - (\beta_0 + \beta_1\overline{x_n} + \overline{\varepsilon_n}) + \hat{\beta}_1\overline{x_n} + (\beta_1 - \hat{\beta}_1)x_i + \varepsilon_i\Big) \\
&= \sum_{i=1}^{n} (x_i - \overline{x_n})\Big(\varepsilon_i - \overline{\varepsilon_n} + (\beta_1 - \hat{\beta}_1)(x_i - \overline{x_n})\Big) \\
&= \sum_{i=1}^{n} (x_i - \overline{x_n})(\varepsilon_i - \overline{\varepsilon_n}) - (\hat{\beta}_1 - \beta_1)\sum_{i=1}^{n}(x_i - \overline{x_n})^2 \\
&= 0,
\end{aligned}$$

which says that the residual $\hat{\varepsilon}_i$ is uncorrelated with regressor x_i[10]. Similarly, we could show that the residual is uncorrelated with the fitted value:

$$\begin{aligned}
\sum_{i=1}^{n} \hat{y}_i \hat{\varepsilon}_i &= \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i)\hat{\varepsilon}_i \\
&= \hat{\beta}_0 \sum_{i=1}^{n} \hat{\varepsilon}_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i \hat{\varepsilon}_i \\
&= 0
\end{aligned}$$

3.5 Estimation of the Variance σ^2

To estimate $\sigma^2 = \mathrm{Var}(\varepsilon_i)$, a natural way is to use the sample variance[11] of $\{\varepsilon_1, \cdots, \varepsilon_n\}$, *i.e.* ,

$$\hat{\sigma}^2 := \frac{1}{n-1}\sum_{i=1}^{n}(\varepsilon_i - \overline{\varepsilon_n})^2.$$

However, the error terms ε_i are not observable. To get around, by the 'plug-in' principle, we replace them with residuals $\hat{\varepsilon}_i$. Then we obtain:

$$\begin{aligned}
\tilde{\sigma}^2 &:= \frac{1}{n-2}\sum_{i=1}^{n}(\hat{\varepsilon}_i - \overline{\hat{\varepsilon}_n})^2 \\
&= \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i{}^2.
\end{aligned}$$

[10] It would be easier to show if one uses matrix form, however, for exam purposes, scalar form derivation is important per se, with its specific computing trick.

[11] Sample variance is always an unbiased estimator of variance. Check!

Remark 1. One might have noticed that now we divide the summation of squares by $(n-2)$ rather than $(n-1)$. The reason is to make sure $\tilde{\sigma}^2$ is an unbiased estimator of $\sigma^2 = \text{Var}(\varepsilon)$, similar to the reason why the sample variance is divided by $(n-1)$ rather than n (Check! It is important.). We will show this later. For intuition, think of it in this way: to estimate the residuals, we have utilised some amount of information, as if we 'lost' a data point, so $\{\hat{\varepsilon}_1 - \overline{\hat{\varepsilon}_n}, \cdots, \hat{\varepsilon}_n - \overline{\hat{\varepsilon}_n}\}$ could only be treated as $(n-2)$ independent data points. Here we have:

$$\mathbb{E}[\tilde{\sigma}^2] = \sigma^2.$$

Remark 2. We have used the fact 3.9: $\overline{\hat{\varepsilon}_n} = 0$, which holds without any assumptions.

3.6 Sampling Distribution of OLS Estimators - Testing

We have seen that $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables (since they depend on the sample, which is a bunch of independent random variables). Thus, we could analyse the distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$, so that hypothesis testing tasks can be done. For example, we are always interested in $\mathbb{H}_0 : \beta_1 = 0$, in plain English, we hope to test whether parameters are significantly not equal to zero, which means whether there is a relationship between the two variables.

However, one should be aware that, in most cases, it is hard to derive the distribution of an estimator theoretically. Here we could do it because the error term ε_i follows a normal distribution. The famous normal distribution has been thoroughly studied and has lots of nice properties.

Without the normality of ε_i, since the estimator is a linear combination of independent r.v.s which is similar to 'sample mean', we could apply the Central Limit Theorem to draw a conclusion on the asymptotic distribution where we assume sample size n is large enough.

3.6.1 The Distribution of $\tilde{\sigma}^2$

Recall 3.14 and 3.9, do some algebra:

$$\begin{aligned}
\sum_{i=1}^n \hat{\varepsilon}_i^2 &= \sum_{i=1}^n (\hat{\varepsilon}_i - \overline{\hat{\varepsilon}_n})^2 \qquad (3.10)\\
&= \sum_{i=1}^n \Big(\Big[(\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1) x_i + \varepsilon_i \Big] - \Big[(\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)\overline{x_n} + \overline{\varepsilon_n} \Big] \Big)^2 \\
&= \sum_{i=1}^n \Big((\hat{\beta}_1 - \beta_1)(x_i - \overline{x_n}) - (\varepsilon_i - \overline{\varepsilon_n}) \Big)^2 \\
&= \sum_{i=1}^n \Big((\hat{\beta}_1 - \beta_1)^2 (x_i - \overline{x_n})^2 - 2(\hat{\beta}_1 - \beta_1)(x_i - \overline{x_n})(\varepsilon_i - \overline{\varepsilon_n}) + (\varepsilon_i - \overline{\varepsilon_n})^2 \Big) \\
&= (\hat{\beta}_1 - \beta_1)^2 \sum_{i=1}^n (x_i - \overline{x_n})^2 - 2(\hat{\beta}_1 - \beta_1) \sum_{i=1}^n (x_i - \overline{x_n})(\varepsilon_i - \overline{\varepsilon_n}) + \sum_{i=1}^n (\varepsilon_i - \overline{\varepsilon_n})^2 \\
&= \Big(\sum_{i=1}^n w_i \varepsilon_i \Big)^2 \sum_{i=1}^n (x_i - \overline{x_n})^2 - 2 \Big(\sum_{i=1}^n w_i \varepsilon_i \Big) \sum_{i=1}^n \varepsilon_i (x_i - \overline{x_n}) + \sum_{i=1}^n (\varepsilon_i - \overline{\varepsilon_n})^2 \\
&= -\Big(\sum_{i=1}^n w_i \varepsilon_i \Big)^2 \sum_{i=1}^n (x_i - \overline{x_n})^2 + \sum_{i=1}^n (\varepsilon_i - \overline{\varepsilon_n})^2. \qquad (3.11)
\end{aligned}$$

It is easy to show that $\sigma^{-2} \sum_{i=1}^n (\varepsilon_i - \overline{\varepsilon_i})^2 \sim \chi^2(n-1)$, and $\sigma^{-2} \Big(\sum_{i=1}^n w_i \varepsilon_i \Big)^2 \sum_{i=1}^n (x_i - \overline{x_i})^2 \sim \chi^2(1)$ because of 3.15. Multiply the normal random vector $(\varepsilon_1, \cdots, \varepsilon_n)^\top$ with a proper orthonormal matrix

$O \in \mathbb{R}^{n\times n}$, we have $\sigma^{-2}\sum_{i=1}^n \hat{\varepsilon}_i^2 \sim \chi^2(n-2)$, thus, $\mathbb{E}\Big[\sigma^{-2}\sum_{i=1}^n \hat{\varepsilon}_i^2\Big] = n-2$. By definition, $\frac{1}{n-2}\sum_{i=1}^n \hat{\varepsilon}_i^2$ is an unbiased estimator of σ^2.

Thus, replace σ^2 by $\frac{1}{n-2}\sum_{i=1}^n \hat{\varepsilon}_i^2$, we have:

$$\frac{\hat{\beta}_1^{\text{OLS}} - \beta_1}{\sqrt{\frac{1}{n-2}\sum_{i=1}^n \hat{\varepsilon}_i^2}}\sqrt{\sum_{i=1}^n (x_i - \overline{x_n})^2} \sim t_{n-2}. \tag{3.12}$$

3.6.2 The Distribution of $\hat{\beta}_1^{\text{OLS}}$

Now let us derive the sampling distribution of OLS estimators. From 3.3, we do a bit algebra:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\frac{1}{n}\sum_{i=1}^n (y_i - \overline{y_n})(x_i - \overline{x_n})}{\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x_n})^2} \\
&= \frac{\sum_{i=1}^n \big(\beta_1(x_i - \overline{x_n}) + (\varepsilon_i - \overline{\varepsilon_n})\big)(x_i - \overline{x_n})}{\sum_{i=1}^n (x_i - \overline{x_n})^2} \\
&= \beta_1 + \frac{\sum_{i=1}^n (\varepsilon_i - \overline{\varepsilon_n})(x_i - \overline{x_n})}{\sum_{i=1}^n (x_i - \overline{x_n})^2}. \\
&= \beta_1 + \frac{\sum_{i=1}^n \varepsilon_i (x_i - \overline{x_n})}{\sum_{i=1}^n (x_i - \overline{x_n})^2}.
\end{aligned} \tag{3.13}$$

In the first step, by the definition of the model 1.1, we calculate $\overline{y_n} = \frac{1}{n}\sum_{i=1}^n(\beta_0 + \beta_1 x_i + \varepsilon_i) = \beta_0 + \beta_1\overline{x_n} + \overline{\varepsilon_n}$, thus, $y_i - \overline{y_n} = \beta_1(x_i - \overline{x_n}) + (\varepsilon_i - \overline{\varepsilon_n})$. The second step is merely simple algebra. However, the third step is a trick in algebra worth mentioning, since some tutorials in linear algebra use $\frac{1}{n}\sum_{i=1}^n \varepsilon_i = 0$, which is both wrong[12] and not necessary.

Now come back to 3.13. Under fixed design, x_i is not random. We could see the structure more clearly by defining the weight $w_i = \frac{\frac{1}{n}(x_i - \overline{x_n})}{\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x_n})^2}$ for $i = 1, \cdots n$, then those weights satisfies $\sum_{i=1}^n w_i = 0$.

Now let us look at the second term in 3.13,

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^n \varepsilon_i (x_i - \overline{x_n})}{\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x_n})^2} = \beta_1 + \sum_{i=1}^n w_i \varepsilon_i, \tag{3.14}$$

that is, a weighted sum of i.i.d. normal (since $\varepsilon_i \sim_{i.i.d.} N(0, \sigma^2)$). Thus, applying the basic property of normal distribution[13],

Proposition 1. *If $X \sim N(0, \sigma_1^2)$, $Y \sim N(0, \sigma_2^2)$, X and Y are independent, then for any constant $a, b \in \mathbb{R}$, we have $aX_1 + bX_2 \sim N(0, a^2\sigma_1^2 + b^2\sigma_2^2)$.*

We could get:

$$\hat{\beta}_1 - \beta_1 = \sum_{i=1}^n w_i \varepsilon_i \sim N\Big(0, \sum_{i=1}^n w_i^2 \sigma^2\Big). \tag{3.15}$$

[12] Keep it in mind that ε_i's are random variables

[13] This property is implied by a more basic result: the summation of independent normal r.v.s' are still normal. This is one way to characterise normal distribution. Think of it: for example, if we add two uniform distribution, the result is no longer uniform.

Substitute the definition of w_i in, and simple algebra gives:

$$\begin{aligned}
\sum_{i=1}^{n} w_i^2 &= \sum_{i=1}^{n} \Big(\frac{\frac{1}{n}(x_i - \overline{x_n})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x_n})^2} \Big)^2 \\
&= \sum_{i=1}^{n} \frac{\frac{1}{n^2}(x_i - \overline{x_n})^2}{\Big(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x_n})^2\Big)^2} \\
&= \frac{\frac{1}{n^2}\sum_{i=1}^{n}(x_i - \overline{x_n})^2}{\Big(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x_n})^2\Big)^2} \\
&= \frac{\frac{1}{n}}{\Big(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x_n})^2\Big)} \\
&= \frac{1}{\sum_{i=1}^{n}(x_i - \overline{x_n})^2}. \qquad (3.16)
\end{aligned}$$

Thus,

$$\hat{\beta}_1^{\text{OLS}} \sim N(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x_n})^2}). \qquad (3.17)$$

Remark 1. It is also implied that $\hat{\beta}_1^{\text{OLS}}$ is an unbiased estimator of β_1:

$$\mathbb{E}\hat{\beta}_1^{\text{OLS}} = \beta_1.$$

Remark 2. For interpretation, recall that a small variance of an estimator implied that it is a more accurate estimation [14]. In 3.17, we could see that, if the $\sigma^2 = \text{Var}(\varepsilon_i)$ is smaller, or the sample size n is large, or there is more dispersion of observations around the sample mean $\overline{x_n}$, the estimator can be more accurate.

Remark 3. The algebra holds for both cases, where we know the true value σ^2, or we do not. However, if we want to standardise the distribution of $\hat{\beta}^1$, we need to consider these two cases separately.

3.6.2.1 σ^2 is Known

When $\sigma^2 = \text{Var}(\varepsilon_i)$ is Known, 3.17 tell us:

$$\frac{\hat{\beta}_1^{\text{OLS}} - \beta_1}{\sigma} \sqrt{\sum_{i=1}^{n}(x_i - \overline{x_n})^2} \sim N(0,1). \qquad (3.18)$$

If we want to conduct a hypothesis testing: $\mathbb{H}_0 : \beta_1 = 0$ versus $\mathbb{H}_1 : \beta_1 \neq 0$. Under the null, $\beta_1 = 0$, thus, $\frac{\hat{\beta}_1^{\text{OLS}}}{\sigma}\sqrt{\sum_{i=1}^{n}(x_i - \overline{x_n})^2} \sim N(0,1)$. Compute the value of this test statistic, and compare it with the critical value $z_{\alpha/2}$ (from standard normal). The task is done.

3.6.2.2 σ^2 is Unknown - 'Plug-in'

By Basu's Theorem, $\hat{\beta}_1^{\text{OLS}}$ and $\frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$ are independent (This is important to confirm the t-distribution). Thus, we have:

$$\frac{\hat{\beta}_1^{\text{OLS}} - \beta_1}{\sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2}} \sqrt{\sum_{i=1}^{n}(x_i - \overline{x_n})^2} \sim t(n-2).$$

[14] For those interested, check out Cremer-Rao Lower Bound.

3.6.3 Testing

A testing problem is to draw conclusions on whether we should reject the null $\mathbb{H}_0$ (it is a binary decision: reject or not to reject) on behalf of the alternative $\mathbb{H}_1$, based on the data in our hand $D = \{(x_1, y_1), \cdots, (x_n, y_n)\}$.

To make a decision, obviously, we need some criteria to tell us when we should and when we shouldn't. Further, as a theoretical statistician, one might also be interested in why these criteria work well, and whether some criteria are better than others, in which sense are they better, etc.

3.6.3.1 Decision Rules

A decision rule is a function Dec : $D \to$ {reject, not reject}, where D is the space of samples, and {reject, not reject} is the set of two available decisions.

By a function, we mean, as long as we have a bunch of data in hand, based on the decision rule, we should be able to know whether we reject or not.

One thing worth mentioning is that, the 'critical value' we learn in the elementary course, is actually a sufficient feature for decision rules. As a convention, 'critical value' is powerful enough to deal with most of the testing problems. However, in my opinion, the decision rule is more fundamental, which helps foster one's intuition.

3.6.3.2 Two Types of Error

Conventionally, there are two types of criteria in testing: the probability of committing the Type-I error (false positive, false rejection)[15]:

$$\mathrm{P}(\text{reject } \mathbb{H}_0 \mid \mathbb{H}_0 \text{ is true}) \tag{3.19}$$

and the probability of committing the Type-II error (false negative, fail to reject):

$$\mathrm{P}(\text{do not reject } \mathbb{H}_0 \mid \mathbb{H}_1 \text{ is true}) \tag{3.20}$$

Remark. Here in the type-II error, we write 'do not reject' rather than 'accept'. This is mainly because we do not treat the two types of error equally, we will see this later.

For intuition, 'reject' actually means 'we have enough evidence to reject', and 'do not reject' means 'we do not have enough evidence to reject the null', rather than 'we have enough evidence to support the alternative'.

Once we have a decision rule, we can calculate its type-I error and type-II error. And clearly, we want to have a decision rule that makes only a few mistakes, *i.e.* , we want both the type-I error and the type-II error to be as small as possible. However, like everything in the world, there is a trade-off. From a given sample, if we want to reduce the change to make a type-I error, then the change to make a type-II error will increase. If we want to reduce both types of error, there might be only one way - get more samples.

Then, instead of using the long sentences 'the probability to commit type-I error' and 'the probability to commit type-II error', researchers defined 'size' and 'power' for a decision rule as follows to represent the same thing (there is a little bit of difference).

$$\begin{aligned} \text{size of DeRu} &= \mathrm{P}(\text{ reject } \mathbb{H}_0 \mid \mathbb{H}_0 \text{ is true}) =: \alpha \\ \text{power of DeRu} &= \mathrm{P}(\text{ reject } \mathbb{H}_0 \mid \mathbb{H}_1 \text{ is true}) =: 1 - \beta, \end{aligned}$$

where α is the probability of committing type-I error, and β is the probability of committing type-II error.

[15] The word 'positive' comes from the scenario where one is tested for whether infected Covid (for example)

Using the new concept, our purpose is to find a decision rule with a lower size but a higher power. Then, a natural question is proposed: how should we trade-off between α and β? Should we find a test to minimise $\alpha+\beta$? If that is the case, why we do not choose the decision rule that minimises $\alpha+1.2\beta$, or $\alpha+0.9\beta$?

One might have found that there are no objectively correct answers for sure. Let us look at the convention first.

The convention is as follows. Fix a value of α (say, α^*); among all decision rules that have a lower α than α^*, we search for the one with the smallest β (hence with the largest power, $1-\beta$).

The advantage of this method is that, all decision rule found in this way, is the most powerful decision rule at some level of α, so it is some kind of optimal choice. What is still not solved is that, it is hard to compare two most powerful decision rules at different levels of α. For example, if T_1 is the most powerful test[16] at $\alpha=0.05$, and T_2 is the most powerful test at $\alpha=0.01$, it is still unclear which is 'better'.

3.6.3.3 General Procedure

I will first list the general procedure to do testing and then provide examples to elaborate on some detailed questions.

1. Apply your intuition to determine the format of the decision rule;
2. Select a number to be the 'significant level', e.g., 0.05 (typical choices are $0.05, 0.01, 0.005, 0.001$)
3. Specify the decision rule.
4. Compute the value of test statistic, and use it in the decision rule to finalise the decision.

For example, if we want to conduct the hypothesis testing: $\mathbb{H}_0: \beta_1 = 0$ versus $\mathbb{H}_1: \beta_1 \neq 0$ (in plain English: whether feature x significantly affects y) while σ^2 is known. In such situation, recall that $t = \frac{\hat{\beta}_1^{\text{OLS}}-\beta_1}{\sigma}\sqrt{\sum_{i=1}^n (x_i-\overline{x_n})^2} \sim N(0,1)$. Here is the procedure for this concrete example:

1. To determine the decision rule, notice that what we want to test (and conclude) is a proposition about the population parameter β_1, however, we only have its sample version $\hat{\beta}_1$. The intuition (or Law of Large Numbers) tells us that $\hat{\beta}_1$ should be close to β_1. When $\mathbb{H}_0$ is true, $\beta_1=0$, then $|\hat{\beta}_1-\beta_1|$ should not be too large. [17] If it is too large (say, $|\hat{\beta}_1-\beta_1|>c$), then we reject the null. This is what a reasonable decision rule should look like.
2. Select $\alpha=0.05$ for example.
3. We want to control α, that is the probability to commit type-I error, so the decision rule (or c) should satisfy:

$$\begin{aligned}
\text{P}(\text{reject } \mathbb{H}_0 \mid \mathbb{H}_0 \text{ is true}) &= \text{P}\Big(|\hat{\beta}_1| > c \,\Big|\, \frac{\hat{\beta}_1}{\sigma}\sqrt{\sum_{i=1}^n (x_i-\overline{x_n})^2} \sim N(0,1)\Big) \\
&= \text{P}\Big(\frac{|\hat{\beta}_1|}{\sigma}\sqrt{\sum_{i=1}^n (x_i-\overline{x_n})^2} > \frac{c}{\sigma}\sqrt{\sum_{i=1}^n (x_i-\overline{x_n})^2} \,\Big|\, \frac{\hat{\beta}_1}{\sigma}\sqrt{\sum_{i=1}^n (x_i-\overline{x_n})^2} \sim N(0,1)\Big) \\
&= \text{P}_{Z\sim N(0,1)}\Big(|Z| > \frac{c}{\sigma}\sqrt{\sum_{i=1}^n (x_i-\overline{x_n})^2}\Big)
\end{aligned}$$

[16] We are going to use 'decision rule' and 'test' interchangeably, you might have seen that they are simply the same thing.

[17] We cannot directly use the distribution of $\hat{\beta}_1$ to do testing since it contains unknown parameters, that's why we standardise it and get $t^* = \frac{\hat{\beta}_1^{\text{OLS}}}{\sigma}\sqrt{\sum_{i=1}^n (x_i-\overline{x_n})^2} \sim N(0,1)$ under the null.

so we should set $\frac{c}{\sigma}\sqrt{\sum_{i=1}^n (x_i - \overline{x_n})^2} = z_{0.975}$ and solve for c. When c is known, the decision is all clear.

4. Compute $\frac{\hat{\beta}_1}{\sigma}\sqrt{\sum_{i=1}^n (x_i - \overline{x_n})^2}$, and compare it with the critical value (in standard normal distribution). Done.[18]

3.6.4 The Distribution of $\hat{\beta}_0^{\mathbf{OLS}}$

Recall that $\hat{\beta}_0^{\text{OLS}} = \overline{y_n} - \hat{\beta}_1^{\text{OLS}}\overline{x_n}$, combined with 3.14, we have:

$$\begin{aligned}
\hat{\beta}_0^{\text{OLS}} - \beta_0 &= \frac{1}{n}\sum_{i=1}^n (\beta_0 + \beta_1 x_i + \varepsilon_i) - \hat{\beta}_1^{\text{OLS}}\overline{x_n} - \beta_0 \\
&= (\beta_1 - \hat{\beta}_1^{\text{OLS}})\overline{x_n} + \overline{\varepsilon_n} \\
&= -\Big(\sum_{i=1}^n w_i \varepsilon_i\Big)\overline{x_n} + \overline{\varepsilon_n} \\
&= \sum_{i=1}^n \varepsilon_i \Big(\frac{1}{n} - \overline{x_n} w_i\Big) \\
&=: \sum_{i=1}^n \varepsilon_i v_i. \qquad (3.21)
\end{aligned}$$

For these weights v_i, simple algebra gives $\sum_{i=1}^n v_i = \sum_{i=1}^n \left(\frac{1}{n} - \overline{x_n} w_i\right) = 1 - \overline{x_n}\sum_{i=1}^n w_i = 1 - \overline{x_n}\frac{\sum_{i=1}^n (x_i - \overline{x_n})}{\sum_{i=1}^n (x_i - \overline{x_n})^2} = 1$. Since $\sum_{i=1}^n \varepsilon_i v_i$ is a linear combination of independent normal, we have:

$$\sum_{i=1}^n \varepsilon_i v_i \sim N\Big(0, \sigma^2 \sum_{i=1}^n v_i^2\Big).$$

Now solve for $\sum_{i=1}^n v_i^2$:

$$\begin{aligned}
\sum_{i=1}^n v_i^2 &= \sum_{i=1}^n (\frac{1}{n} - \overline{x_n} w_i)^2 \qquad (3.22) \\
&= \sum_{i=1}^n \Big(\frac{1}{n^2} - \frac{2}{n}\overline{x_n} w_i + (\overline{x_n})^2 w_i^2\Big) \qquad (3.23) \\
&= \frac{1}{n} - 0 + (\overline{x_n})^2 \sum_{i=1}^n w_i^2 \qquad (3.24) \\
&= \frac{1}{n} + \frac{(\overline{x_n})^2}{\sum_{i=1}^n (x_i - \overline{x_n})^2}. \qquad (3.25)
\end{aligned}$$

Thus,

$$\hat{\beta}_0^{\text{OLS}} \sim N\Big(\beta_0, \sigma^2 \big(\frac{1}{n} + \frac{(\overline{x_n})^2}{\sum_{i=1}^n (x_i - \overline{x_n})^2}\big)\Big) \qquad (3.26)$$

[18] Since t-distribution has a fatter tail than standard normal, which indicates that we need a larger sample size to achieve the same power. This coincides with the intuition: More information (whether we know the variance) makes the estimation more accurate.

3.6.4.1 σ^2 is Known

$$\frac{\hat{\beta}_0^{\text{OLS}} - \beta_0}{\sqrt{\sigma^2(\frac{1}{n} + \frac{(\overline{x_n})^2}{\sum_{i=1}^n (x_i - \overline{x_n})^2})}} \sim N(0,1). \tag{3.27}$$

Under the null, Compute the value of this test statistic.

3.6.4.2 σ^2 is Unknown

With 'plug-in' principle, replace σ^2 by its estimate $\frac{1}{n-2}\sum_{i=1}^n \hat{\varepsilon}_i^2$, similarly:

$$\frac{\hat{\beta}_0^{\text{OLS}} - \beta_0}{\sqrt{\frac{1}{n-2}\Big(\sum_{i=1}^n \hat{\varepsilon}_i^2\Big)\Big(\frac{1}{n} + \frac{(\overline{x_n})^2}{\sum_{i=1}^n (x_i - \overline{x_n})^2}\Big)}} \sim t_{n-2}. \tag{3.28}$$

3.7 Standard Normal, χ-squared and t-distribution*

We have seen that in different cases, we could do testing for β using either t-distribution or the standard normal, and do testing for $\sigma^2 = \text{Var}(\varepsilon)$ by using χ-squared distribution. Here are some details about these distributions for those interested.

I leave another simple but important question here: why knowing the distribution is essential for testing? Can we do testing without knowing the distribution?

3.7.1 Standard Normal

It is definitely the most well-known distribution. We say X follows the standard normal distribution (denote as $X \sim N(0,1)$), if:

$$\text{P}(X \le x) = \int_{-\infty}^{x} \phi(t)\text{d}t =: \Phi(x),$$

where $\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$. If you just finished your first-year undergrad study, and know a probability distribution should satisfy $\text{P}(X \in \mathbb{R}) = \int_{-\infty}^{\infty} \phi(t)\text{d}t = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\text{d}t = 1$, you might be interested why this happens.

In fact, we do not have a neat expression for $\int_{-\infty}^{x} e^{-t^2}\text{d}t$. However, we have $\int_{-\infty}^{\infty} e^{-t^2}\text{d}t = \sqrt{2\pi}$, from a quite interesting proof:

Proof. Let us look at the square of it:

$$\begin{aligned}\Big(\int_{-\infty}^{\infty} e^{-t^2}\text{d}t\Big)^2 &= \Big(\int_{-\infty}^{\infty} e^{-x^2}\text{d}x\Big)\Big(\int_{-\infty}^{\infty} e^{-y^2}\text{d}y\Big) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)}\text{d}x\text{d}y.\end{aligned}$$

Since the integral converges (as $t \to \infty$, e^{-t^2} goes to 0 more quickly than $\frac{1}{t^2}$, which converges), any parametrisation will yield the same result of the integral. Set $x = r\cos(\theta), y = r\sin(\theta)$, for $r \in [0,\infty)$ and $\theta \in [0, 2\pi)$, then:

$$\frac{\partial(x,y)}{\partial(r,\theta)} = \det\begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{pmatrix} = \det\begin{pmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{pmatrix} = r.$$

Then, together with $x^2+y^2=r^2$, we could continue:

$$\begin{aligned}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)}\mathrm{d}x\mathrm{d}y &= \int_0^{2\pi}\int_0^{\infty} e^{-r^2} r\mathrm{d}r\mathrm{d}\theta = 2\pi\int_0^{\infty} e^{-r^2} r\mathrm{d}r \\ &= 2\pi\int_0^{\infty} e^{-r^2}\frac{1}{2}\mathrm{d}(r^2) \\ &= \pi\int_0^{\infty} e^{-s}\mathrm{d}s = \pi\end{aligned}$$

Thus, $\int_{-\infty}^{\infty} e^{-t^2}\mathrm{d}t = \sqrt{\pi}$, and $\int_{-\infty}^{\infty} e^{-\frac{t^2}{2}}\mathrm{d}t = \sqrt{2}\int_{-\infty}^{\infty} e^{-\frac{t^2}{2}}\mathrm{d}\frac{t}{\sqrt{2}} = \sqrt{2\pi}$. □

The reason why normal distribution is so popular is due to the Central Limit Theorem. Basically, the theorem tells us that the standardised sample mean will be more and more like a standard normal.

Theorem 1. Central Limit Theorem. *Suppose X follows some distribution with finite variance. Denote $\mu = \mathbb{E}X$ and $\sigma^2 = \mathrm{Var}(X)$. If $X_1, X_2, ...$ are independent copies of X, then we have:*

$$\frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1).$$

The convergence is in the sense of distribution, i.e. , For any $t \in \mathbb{R}$, as $n \to \infty$,

$$\mathrm{P}\Big(\frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} \le t\Big) \longrightarrow \Phi(t).$$

It is an amazing theorem in that, we do not restrict the distribution of X too much: as long as the variance exists (*i.e.* , the distribution is not too fat-tailed), the sample mean will eventually (as we have enough many observations) behave like normal.

3.7.2 χ-squared Distribution

Suppose $X_1, X_2, \cdots$ are a sequence of independent standard normal r.v.s', then we define χ-squared distribution with degree-of-freedom n to be the distribution of

$$Y_1 := \sum_{i=1}^{n} X_i^2, \tag{3.29}$$

denoted as $Y_1 \sim \chi^2(n)$ (or $Y_1 \sim \chi_n^2$ in some books).

Due to the explicit relationship with standard normal, we get the mean and the variance: $\mathbb{E}Y = \mathbb{E}\sum_{i=1}^n X_i^2 = \sum_{i=1}^n \mathbb{E}X_i^2 = n$. $\mathrm{Var}(Y) = \mathrm{Var}(\sum_{i=1}^n X_i^2) = \sum_{i=1}^n \mathrm{Var}(X_i^2) = \sum_{i=1}^n \mathbb{E}(X_i^4) - [\mathbb{E}(X_i^2)]^2 = n\mathbb{E}(X_1^4) - n = 2n$.

In the last step, we directly compute the integral or apply moment generating function method, we get $\mathbb{E}(X_1^4) = 3n$ for $X_1 \sim N(0,1)$.

To derive the probability density function (pdf) of χ-squared distribution is a nice practice to test one's grasp of probability theory and basic calculus. Take $n = 1$ as an example, for any $y \ge 0$, the cumulative probability function (cdf) is:

$$\mathrm{P}(Y_1 \le y) = \mathrm{P}(-\sqrt{y} \le X_1 \le \sqrt{y}) = \int_{\sqrt{y}}^{\sqrt{y}} \phi(t)\mathrm{d}t = \int_{\sqrt{y}}^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}\mathrm{d}t.$$

Hence, the pdf is provided by taking the derivative:

$$p_{Y_1}(y) = \frac{\partial}{\partial y}\int_{\sqrt{y}}^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}\mathrm{d}t = \frac{1}{\sqrt{2\pi}} y^{-\frac{1}{2}} e^{-\frac{y}{2}}.$$

For $n = 2$,

$$\begin{aligned}
\mathrm{P}(Y_2 \le y) = \mathrm{P}(X_1^2 + X_2^2 \le y) &= \iint_{x_1^2+x_2^2 \le y} p_{X_1,X_2}(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2 \\
&= \iint_{x_1^2+x_2^2 \le y} \phi(x_1)\phi(x_2)\mathrm{d}x_1\mathrm{d}x_2 \\
&= \iint_{x_1^2+x_2^2 \le y} \frac{1}{2\pi} e^{-\frac{1}{2}(x_1^2+x_2^2)}\mathrm{d}x_1\mathrm{d}x_2 \\
&= \int_0^{2\pi}\int_0^{\sqrt{y}} \frac{1}{2\pi} e^{-\frac{1}{2}r^2} r\mathrm{d}r\mathrm{d}\theta \\
&= 2\pi \int_0^{\sqrt{y}} \frac{1}{2\pi} e^{-\frac{1}{2}r^2}\mathrm{d}(\frac{r^2}{2}) \\
&= \int_0^{\frac{y}{2}} e^{-s}\mathrm{d}s \\
&= 1 - e^{-\frac{y}{2}},
\end{aligned}$$

Thus,

$$p_{Y_2}(y) = \frac{\partial}{\partial y}(1 - e^{-\frac{y}{2}}) = \frac{1}{2} y e^{-\frac{y}{2}}.$$

3.7.3 t-distribution

Still, suppose $X_1, X_2, \cdots$ are a sequence of independent standard normal r.v.s', then we define t-distribution with degree-of-freedom n to be the distribution of

$$Z_n := \frac{X_{n+1}}{\sqrt{\sum_{i=1}^n X_i^2/n}}, \tag{3.30}$$

denoted as $Z_n \sim t(n)$ (or $Z_n \sim t_n$ in some books).

For t-distribution, there is an interesting result which helps foster our intuition:

Proposition 2. *With n goes to infinity, t-distribution with degree of freedom n converges in distribution to standard normal. In notation, suppose Z_n is defined by 3.30 for each n, then for any $t \in \mathbb{R}$, as $n \to \infty$,*

$$\mathrm{P}(Z_n \le t) \to \Phi(t).$$

The Proposition can be proved by Slutsky's Theorem.

Theorem 2. Slutsky's Theorem. *X_n and Y_n are two sequences of random variables. $X_n \xrightarrow{d} X$ (converge in distribution to X), while $Y_n \xrightarrow{p} c$ (converge in probability to a constant c), then:*

- $X_n + Y_n \xrightarrow{d} X + c$;
- $X_n Y_n \xrightarrow{d} Xc$;
- $X_n / Y_n \xrightarrow{d} X/c$, *provided* $c \neq 0$.

Proof. of Proposition 2.

With Slutsky's Theorem, notice that in the numerator of 3.30, it is exactly a normal random variable. We now show that the denominator converges in probability to 1. This follows from the weak Law of Large numbers and the continuous mapping theorem (3).

X_i^2 are a sequence of i.i.d. random variable with $\mathbb{E}X_i^2 = 1$. By weak law of large numbers:

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{p} 1,$$

then by comtinuous mapping theorem, take $g(\cdot)$ as $g(x) = x^{\frac{1}{2}}$:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2} \xrightarrow{p} 1.$$

□

3.8 'Plug-in' Estimators*

Here is a distraction. I would like to provide a bit explanation of how and why 'plug-in' estimators work. Remember that when a parameter is unknown, and we still need to use its value, we apply the 'plug-in' principle - estimate it, and use it as the true value as we know it. This principle works fine in practice, however, in theory, some follow-up questions require further treatment.

To generally explain what are these questions, consider the following scenario: we have sample $\{(x_i, y_i)\}_{i=1}^n$, and want to use the sample to estimate a parameter β in a data generating process $y_i = f_{\beta,\theta}(x_i) + \varepsilon_i$ (There is also another parameter θ in this data generating process).

We construct an estimator $\hat{\beta}$, which requires the value of θ. Denote the estimator as $\hat{\beta} = g(X, Y, \theta)$ for some function g, where $X = (x_1, \cdots, x_n)^\mathsf{T}$, $Y = (y_1, \cdots, y_n)^\mathsf{T}$. In the case where we know θ, the estimator works fine.

In the case where we don't know θ, we could construct an estimator $\hat{\theta} = h(X, Y)$, then plug it in g to form an estimator of β: $\hat{\hat{\beta}} = g(X, Y, \hat{\theta}) = g\big(X, Y, h(X, Y)\big) =: \ell(X, Y)$ which does not require the value of θ any more.

Obviously, $\hat{\beta}$ and $\hat{\hat{\beta}}$ are two different estimators, and we are interested in to what extent they are different. To assess estimators, we need some criteria. There are two widely used criteria of estimators.

3.8.1 Consistency*

Definition 2. Consistency. *For parameter β, an estimator $\hat{\beta}$ is called consistent in estimating β if $\hat{\beta} \xrightarrow{p} \beta$.*

The convergence is in the sense of probability, defined as follows.

Definition 3. Converge in Probability (to a real value) *We say a sequence of random variables $\{X_n\}_{n=1}^{\infty}$ converge in probability to a real value c (denote as $X_n \xrightarrow{p} c$), if $\forall \delta > 0$,*

$$\mathrm{P}(|X_n - c| > \delta) \longrightarrow 0, \quad \text{as} \quad n \to \infty.$$

Theorem 3. Continuous Mapping Theorem. *If $X_n \xrightarrow{p} c$, and $g(\cdot)$ is a continuous function, then $g(X_n) \xrightarrow{p} g(c)$.*

With the Continuous Mapping Theorem, we could prove that if both $\hat{\theta}$ and $\hat{\beta}$ are consistent for parameters θ and β, then the 'plug-in' estimator $\hat{\hat{\beta}}$ is consistent for β. This gives the desired property of a 'plug-in' estimator.

The convergence rate (how fast the convergence is) of estimators is a key sub-field in modern statistics, but we will not go into depth here. To have some kind of intuition, it would be interesting to look at the convergence rate from the information theory point of view. In the first case where we know the value of θ, we have more information, so it is natural to expect a faster convergence.

3.8.2 Distribution*

Generally, replacing a real number with a random variable will change the distribution of that estimator (r.v.). In theory, we could go through and look at how to solve for the distribution of $g(X_1, X_2)$ where X_1 and X_2 are both random variables. However, the integral is typically not basic. For intuition, we give several examples:

1. Suppose $X_1 \sim N(\mu, 1)$. Then $\mu X_1 \sim N(\mu^2, \mu^2)$. Since $\mathbb{E}X_1 = \mu$, X_1 could be treated as an estimator of μ. Replace the μ in μX_1 we get X_1^2. The point here is that X_1^2 is no longer following $N(\mu, 1)$.

2. Given $X \sim N(\mu, \sigma^2)$, and i.i.d. sample $X_1, \cdots, X_n$.
   On one hand, it is clear that $\overline{X_n} = \frac{1}{n}\sum_{i=1}^n X_i \sim N(\mu, \frac{\sigma^2}{n})$, thus,

$$\frac{\overline{X_n} - \mu}{\sqrt{\sigma^2/n}} \sim N(0, 1).$$

   On the other hand, one can use sample variance $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X_n})^2$ as a natural unbiased estimator of σ^2. If we use $\hat{\sigma}^2$ to replace σ^2 above, we would get:

$$\frac{\overline{X_n} - \mu}{\sqrt{\hat{\sigma}^2/n}} \sim t_{n-1}, \tag{3.31}$$

   i.e. , the distribution changes if we replace a real value σ^2 by a random variable $\hat{\sigma}^2$.[10] The proof of result 3.31 takes several pages, for example, see this link provided by Duke University.

[10] The result 3.31 could follow from Basu's theorem, which involves more concepts, such as 'sufficient statistic', 'ancillary statistic', and 'conditional independent'. Terms are left here for those interested.

4 Multivariate Linear Regression

Compared to the simple case where we only have one independent variable (a.k.a. regressors, predictors, covariates, etc.) to explain the dependent variable, here we allow the number of independent variables to be larger than 1. The population regression function becomes:

$$y_i = x_i'\theta + \varepsilon_i,$$

where $x_i = (1, x_i^1, \cdots, x_i^p)' \in \mathbb{R}^{p+1}$, $\theta = (\theta_0, \theta_1, \cdots, \theta_p)' \in \mathbb{R}^{p+1}$,[19] *i.e.* , the number of independent variables is p.

4.1 Prerequisite

4.1.1 DGP in Matrix Form

We need to specify the sample on which we base our estimation to talk about estimation parameters $\theta \in \mathbb{R}^{p+1}$. As in the simple case, we have sample $\{(x_i', y_i)\}_{i=1}^n$. Here is the place to introduce matrix notation which helps to simplify the formula and calculation. We could incorporate all information from the sample in a single line:

$$Y = X\theta + \varepsilon,$$

where $Y = (y_1, \cdots, y_n)' \in \mathbb{R}^n$, $X = (x_1, \cdots, x_n)' \in \mathbb{R}^{n\times(p+1)}$, and $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)' \in \mathbb{R}^n$. Apply the definition of the product of matrix, it is explicitly equivalent to these n number of equations:

$$\begin{aligned} y_1 &= x_1'\theta + \varepsilon_1 \\ y_2 &= x_2'\theta + \varepsilon_2 \\ &\cdots \\ y_n &= x_n'\theta + \varepsilon_n. \end{aligned}$$

We also use $X = (X_0, X_1, X_2, \ldots, X_p)$ to represents the column of matrix X. For $j = 0$, $X_0 = (1, \ldots, 1)^\top \in \mathbb{R}^n$; for $j \geq 1$, each column vector $X_j = (X_{j1}, \ldots, X_{jn})^\top \in \mathbb{R}^n$ represents the value of feature j for all experiment units.

Combining these notations, one can see that $x_i' = (X_{0i}, X_{1i}, \ldots, X_{pi})$.

In a word, x_i' is the i-th row of matrix X, representing the feature of experiment unit i; while X_j is the j-th column of matrix X, representing the value of feature j for all experiment units. One would be more familiar with the notation while using them.

4.1.2 Random Vectors

Suppose $X = (X_1, X_2, \ldots, X_d)^\top \in \mathbb{R}^d$, each X_i is a random variable.

- We extend the definition of '$\mathbb{E}$' to allow it to operate on a vector, by

$$\mathbb{E}X := (\mathbb{E}X_1, \mathbb{E}X_2, \ldots, \mathbb{E}X_d)^\top \in \mathbb{R}^d;$$

- Denote $\mu = \mathbb{E}X$. Recall that, for $d = 1$, variance in 1-D is defined as: $\mathrm{Var}(X) = \mathbb{E}(X-\mu)^2 = \mathbb{E}[X^2] - \mu^2$;
  For $d > 1$, we define 'covariance matrix' in a high-dimensional case similarly:

$$\mathrm{Cov}(X) = \mathbb{E}(X-\mu)(X-\mu)^\top = \mathbb{E}XX^\top - \mu\mu^\top;$$

[19] Normally, by $\mathbb{R}^p$, we mean $\mathbb{R}^{p\times 1}$, *i.e.* , we treat vectors as column vectors by default. This is the convention in both economics and machine learning communities.

The entries in $\mathrm{Cov}(X)$:

$$\big(\mathrm{Cov}(X)\big)_{ij} = \mathbb{E}(X-\mu)_i(X-\mu)_j = \mathrm{Cov}(X_i, X_j)$$

- Examples: $X \in \mathbb{R}^2$,

$$\mathrm{Cov}(X) = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) \end{bmatrix}.$$

- $\mathrm{Cov}(X)$ is a $d \times d$ symmetric PSD (positive semi-definite) matrix.

Proof. $\forall v \in \mathbb{R}^d : v^\top \Sigma v = v^\top(\mathbb{E}XX^\top)v = \mathbb{E}(v^\top X)(X^\top v) = \mathbb{E}\langle X, v\rangle^2 \geq 0.$ □

4.1.3 Multivariate Normal (Prerequisite)

Similar to uni-variate normal, we could define multi-variate normal as[20]:

$$X \sim N(\mu, \Sigma), \tag{4.32}$$

where $X = (X_1, \cdots, X_p)' \in \mathbb{R}^p$ is the normal random vector, $\mu = (\mu_1, \cdots, \mu_p)' \in \mathbb{R}^p$ and $\mu_i = \mathbb{E}X_i$. $\Sigma \in \mathbb{R}^{p\times p}$ is the covariance matrix, $\Sigma_{ij} = \mathrm{Cov}(X_i, X_j)$, thus, Σ is symmetric since $\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i)$.

For the random vector defined in 4.32, the joint probability density function is: [21]

$$\phi(x) = (2\pi)^{-p/2}(\det\Sigma)^{-1/2}\exp\Big\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\Big\} \tag{4.33}$$

For multivariate normal, a nice property is important:

Proposition 3. *If $X \sim N(\mu, \Sigma)$ as in 4.32, and $S \in \mathbb{R}^{m\times p}$ is any $m \times p$ matrix ($m \in \mathbb{Z}_+$). Then,*

$$SX \sim N(S\mu, S\Sigma S').$$

Proof. The proposition follows from three results:

1. Any linear combination of normal (not necessary to be independent) is still normal.
2. For any $a \in \mathbb{R}^p$, $\mathbb{E}a'X = a'\mathbb{E}X$.
3. $\mathrm{Cov}(a'Y, b'Z) = \mathrm{Cov}\Big(\sum_{i=1}^p a_iY_i, \sum_{i=1}^p b_iZ_i\Big) = \sum_{i=1}^p\sum_{i=1}^p a_ib_i\mathrm{Cov}(Y_i, Z_i)$.

By applying the first result, we know each coordinate of SX follows a univariate normal distribution, thus, we only need to make sure what's their mean/variance/covariance. For mean, use the second, for variance and covariance, use the third. □

With this result, we could draw conclusions about the distribution of $\hat\theta$.

$$\hat\theta \sim N\Big(\theta, \sigma^2(X'X)^{-1}\Big)$$

[20] In some books, it is denoted as $X \sim MVN(\mu, \Sigma)$.

[21] The fact that Σ is non-singular coincides with our intuition: If Σ is singular, then there exist one of the X_i being the linear combination of the other, which means we should not embed the vector X into p-dimension.

4.2 Multivariate OLS Estimation

4.2.1 The OLS Estimator

OLS estimator is still defined by the value of θ^* that minimises the Residual sum of squares:

$$\theta^{\text{OLS}} \in \arg\min_{\theta\in\mathbb{R}^{p+1}} (Y - X\theta)^\top (Y - X\theta),$$

Denote $g(\theta) = (Y - X\theta)^\top (Y - X\theta)$, differentiate g w.r.t. θ gives:

$$\begin{aligned} \frac{\partial g(\theta)}{\partial \theta} &= -2X^\top (y - X\theta), \\ \frac{\partial^2 g(\theta)}{\partial\theta\partial\theta^\top} &= 2X^\top X. \end{aligned}$$

Set $\frac{\partial g(\theta)}{\partial\theta} = 0$ yields:[22]

$$\hat{\theta}^{\text{OLS}} = (X'X)^{-1}X'Y, \tag{4.34}$$

and $\frac{\partial^2 g(\theta)}{\partial\theta\partial\theta^\top} = 2X^\top X$ being positive definite ensures that $\hat{\theta}^{\text{OLS}}$ is the minimum point. When dealing with the matrix product, always remember to check the dimension to see whether it is compatible.

Clearly,

$$\begin{aligned} \hat{\theta}^{\text{OLS}} &= (X'X)^{-1}X'(X\theta + \varepsilon) \\ &= \theta + (X'X)^{-1}X'\varepsilon. \end{aligned}$$

So it is the same to prove the consistency property of the estimator under assumptions $\mathbb{E}[\varepsilon] = 0$ for fixed design (or $\mathbb{E}[\varepsilon|X] = 0$ for random design).

One of the important practices to foster one's intuition is to check that the expression 4.34 coincides with the uni-variate case, where $p = 1$, and $X = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix}^\top \in \mathbb{R}^{n\times 2}$.

4.2.2 Fitted Value and Residuals

$$\begin{aligned} \hat{Y} &:= X\hat{\theta}^{\text{OLS}} \\ &= X(X'X)^{-1}X'Y \\ &= \big(X(X'X)^{-1}X'\big)Y \\ &=: P_X Y. \end{aligned}$$

where $P_X = X(X'X)^{-1}X' \in \mathbb{R}^{n\times n}$ is a symmetric idempotent matrix satisfying ${P_X}' = P_X$ and ${P_X}^2 = P_X$ (check!). From matrix theory, we know that,

Proposition 4. *If $f(\cdot)$ is a polynomial, such that $f(A) = 0$ holds for $n \times n$ matrix A. Then, any eigenvalue λ of matrix A satisfies $f(\lambda) = 0$.*

Remark. One should pay attention that the right hand side of $f(A) = 0$, is a $n \times n$ matrix with all elements being 0.

[22] If one is doing machine learning or statistics, it is also important to know how we get it (matrix calculus), towards Lasso regression, Ridge regression, etc.

Proof. To make life easier, we only consider symmetric A so that all eigenvalues of A are real, and A is similar to a diagonal matrix $\Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_n\}$, *i.e.* there exist invertible Q (Q can also be taken as orthogonal, but is not required here) such that $AQ = Q\Lambda$, which implies $A = Q\Lambda Q^{-1}$.
For any integer k,

$$A^k = (Q\Lambda Q^{-1})^k = Q\Lambda^k Q^{-1} = Q \cdot \text{diag}\{\lambda_1^k, \ldots, \lambda_n^k\} \cdot Q^{-1},$$

thus, for any polynomial $f(\cdot)$,

$$f(A) = f(Q\Lambda Q^{-1}) = Qf(\Lambda)Q^{-1} = Q \cdot \text{diag}\{f(\lambda_1), \ldots, f(\lambda_n)\} \cdot Q^{-1}.$$

$f(A) = 0$ implies that $\text{diag}\{f(\lambda_1), \ldots, f(\lambda_n)\} = 0$, *i.e.* $f(\lambda_i) = 0$ for all i. □

Thus, from $P_X^2 = P_X$, we have $\lambda^2 = \lambda$, implying $\lambda = 0, 1$, *i.e.* any eigenvalue of matrix P_X must be 0 or 1.

The residual is defined as:

$$\begin{aligned}
\hat{\varepsilon} &:= Y - \hat{Y} \\
&= Y - X\hat{\theta}^{\text{OLS}} \\
&= Y - X\Big((X'X)^{-1}X'Y\Big) \\
&= \Big(I_n - X(X'X)^{-1}X'\Big)Y \\
&= (I_n - P_X)Y \\
&=: M_X Y.
\end{aligned}$$

One could see that $M_Y^2 = (I_n - P_X)^2 = I_n - P_X + P_X - (P_X)^2 = I_n - (P_X)^2 = I_n - P_X = M_Y$. M_Y is also a projection, and all eigenvalue of M_Y is 0 or 1, similarly.

4.3 Geometric View on the OLS Estimator

4.3.1 Orthogonal Projection and Orthogonal Subspace

The vector space R^n is a natural inner product space with inner product $\langle x, y\rangle = x'y = \sum_{i=1}^n x_i y_i$. If two vectors x, y satisfy $\langle x, y\rangle = 0$, we say they are orthogonal.

Coming back to the linear regression, an important result is that the fitted value and the residual are orthogonal:

$$\begin{aligned}
\langle \hat{Y}, \hat{\varepsilon}\rangle &= \hat{Y}'\hat{\varepsilon} \\
&= (P_X Y)' M_X Y \\
&= Y' P_X' M_X Y \\
&= Y' P_X (I_n - P_X) Y \\
&= Y' O Y \\
&= 0.
\end{aligned}$$

This is to say, P_X and M_X represents an orthogonal projection from $\mathbb{R}^n$ onto its p-dimensional and $(n-p)$-dimensional subspaces.

Furthermore, check whether $\hat{\varepsilon}$ is orthogonal with columns of X:

$$\begin{aligned}
X'\hat{\varepsilon} &= X'\big(I_n - X(X'X)^{-1}X'\big)Y \\
&= X'Y - X'X(X'X)^{-1}X'Y \\
&= X'Y - X'Y \\
&= 0.
\end{aligned}$$

This reveals that any column of X is orthogonal with the residual vector $\hat{\varepsilon}$.

Combining the two arguments, it is now clear that P_X is the operator which projects vectors in $\mathbb{R}^n$ onto the subspace spanned by all column vectors of X; and M_X projects $Y \in \mathbb{R}^n$ into its orthogonal complement. It motivates us to understand the multi-variate case solution as well as the 2SLS (2-stage least square) in a geometric view.

4.3.2 FWL theorem

Consider the case where $p = 2$, and $\beta_0 = 0$ (for simplicity), *i.e.* ,

$$\begin{aligned} Y &= X\beta + \varepsilon \\ &= X_1\beta_1 + X_2\beta_2 + \varepsilon. \end{aligned}$$

To get the OLS estimator β_1^{OLS}, one could either directly apply formula 4.34, or:

1. Regress X_1 on X_2, obtain residual $M_{X_2}X_1$;
2. Regress Y on X_2, obtain residual $M_{X_2}Y$;
3. Regress $M_{X_2}Y$ on $M_{X_2}X_1$.

Theorem 4. *(Frisch-Waugh-Lovell Theorem) The estimator obtained in the three steps above is exactly the OLS estimator.*

Proof. According to the step 3, the estimate obtained is:

$$\begin{aligned} \left((M_{X_2}X_1)'(M_{X_2}X_1)\right)^{-1}(M_{X_2}X_1)'(M_{X_2}Y) &= \left(X_1'(M_{X_2})'M_{X_2}X_1\right)^{-1}X_1'(M_{X_2})'M_{X_2}Y \\ &= (X_1'M_{X_2}X_1)^{-1}X_1'M_{X_2}Y. \end{aligned}$$

To prove the theorem, we need to show $\hat{\beta}_1^{\text{OLS}} = (X_1'M_{X_2}X_1)^{-1}X_1'M_{X_2}Y$, or equivalently:

$$(X_1'M_{X_2}X_1)\hat{\beta}_1^{\text{OLS}} = X_1'M_{X_2}Y.$$

Suppose $Y = X_1\hat{\beta_1}^{\text{OLS}} + X_2\hat{\beta_2}^{\text{OLS}} + \hat{\varepsilon}$ is the fitted-residual decomposition. Left multiply $X_1'M_{X_2}$ on both sides, we get:

$$X_1'M_{X_2}Y = X_1'M_{X_2}X_1\hat{\beta_1}^{\text{OLS}} + X_1'M_{X_2}X_2\hat{\beta_2}^{\text{OLS}} + X_1'M_{X_2}\hat{\varepsilon},$$

where the second term on the RHS is 0 since $M_{X_2}X_2 = 0$, and the third term is also 0 since:

$$\begin{aligned} X_1'M_{X_2}\hat{\varepsilon} &= X_1'(I_n - P_{X_2})\hat{\varepsilon} \\ &= X_1'(\hat{\varepsilon} - P_{X_2}\hat{\varepsilon}) \\ &= X_1'\hat{\varepsilon} \\ &= 0. \end{aligned}$$

□

Remark. Although the theorem gives another way to calculate the OLS estimator, there is almost no reason to use it. It does not make the computation faster in the algorithmic sense. In the situation where the residual for regressing one regressor on another matters, FWL theorem might be useful.

Remark 2. The FWL theorem is useful in 'eliminating the constant β_0' without a cost. Here is how we do it:

1. Regress $Y, X_1, \ldots, X_p$ on X_0 (note that $X_0 = 1$), get residuals $M_0Y, M_0X_1, \ldots, M_0X_p$.

2. Regress M_0Y on those residuals $M_0X_1, \ldots, M_0X_p$.

By the FWL theorem, the estimator we get in the second step is exactly OLS estimators $\hat{\beta}_1^{\text{OLS}}, \ldots, \hat{\beta}_p^{\text{OLS}}$ for regressors except X_0. However, what are those $M_0X_1, \ldots, M_0X_p$?

In fact, regressing X_1 on 1 simply means to find a value $c \in \mathbb{R}$ such that $(\sum_{i=1}^n X_{1i} - c \cdot 1)^2$ is minimised, it is easy to see that $\hat{c} = \frac{1}{n}\sum_{i=1}^n X_{1i} =: \overline{X_1}$, thus, the residual is: $M_0X_1 = X_1 - \hat{c}\mathbf{1} = (X_{11} - \overline{X_1}, \ldots, X_{1n} - \overline{X_1})^\top$.

In plain English, given a dataset, one can always first subtract all features X_j with the sample mean of that feature. By FWL theorem, regress the de-meaned version M_0Y on these de-meaned features will result in the same estimator.

4.4 Variance and its Estimation

You will find it easier to calculate in matrix form than in scalar form since lots of staff just cancel out. In econometrics, there are two 'very-long' words: 'Homoskedasticity' and 'Heteroskedasticity'. The former one says $\text{Var}(\varepsilon_i|X) = \sigma^2$ which means the conditional variance are the same for different X. Under this assumption:

$$\begin{aligned}
\text{Var}(\hat{\theta}|X) &= \text{Var}\Big((X'X)^{-1}X'Y|X\Big) \\
&= \Big((X'X)^{-1}X'\Big) \cdot \text{Var}(Y|X) \cdot \Big((X'X)^{-1}X'\Big)' \\
&= \Big((X'X)^{-1}X'\Big) \cdot (\sigma^2 I_n) \cdot \Big((X'X)^{-1}X'\Big)' \\
&= \sigma^2\Big((X'X)^{-1}X'\Big)\Big((X'X)^{-1}X'\Big)' \\
&= \sigma^2\Big((X'X)^{-1}X'\Big)X(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1},
\end{aligned}$$

where $\text{Var}(Y|X) = \sigma^2 I_n$ is because $\text{Var}(y_i|x_i) = \sigma^2$. Here we have used the associativity of matrix multiplication and commutativity between I and any matrices.

The variance of $\hat{\beta}_j$ is the j-th diagonal element of $\sigma^2(X'X)^{-1}$:

$$\text{Var}(\hat{\beta}_j|X) = \sigma^2(X'X)^{-1}_{jj}$$

4.5 The Efficiency of OLS Estimator

4.5.1 Gauss-Markov theorem

Theorem 5. *For a linear data-generating process $Y = X\beta + \varepsilon$, under the assumptions:*
(A1) No perfect multicollinearity: rank$(X) = k + 1$;
(A2) Strict exogeneity: $\mathbb{E}(\varepsilon|X) = 0$;
(A3) Homoskedasticity: $\text{Var}(\varepsilon_i|X) = \sigma^2$, for $i = 1, \ldots, n$;
(A4) No serial correlation: $\text{Cov}(\varepsilon_i, \varepsilon_j|X) = 0$, for $i \neq j$;
*then the OLS estimator $\hat{\beta}^{OLS}$ is the **BLUE** (**B**est among all **L**inear conditionally **U**nbiased **E**stimators) of β. Furthermore, $\hat{\beta}^{OLS}$ is the **BUE** (Best Unbiased Estimator) of β.*

- The 'BLUE' part is quite obvious:

  Proof. of 'BLUE'.
  For any linear estimator $\tilde{\beta} = A'Y$, where $A \in \mathbb{R}^{n\times(p+1)}$, $\mathbb{E}[\tilde{\beta}|X] = \mathbb{E}[A'Y|X] = A'\mathbb{E}[Y|X] = A'X\beta$, unbiasedness require $A'X = I_{p+1}$.

$\text{Var}(\tilde{\beta}|X) = \text{Var}(A'Y|X) = A'\text{Var}(Y|X)A = \sigma^2 A'A$. To finish the proof, we need to show $A'A \geq (X'X)^{-1}$. Denote $C = A - X(X'X)^{-1}$, we have $X'C = I_{p+1} - I_{p+1} = 0$, then:

$$\begin{aligned} A'A - (X'X)^{-1} &= (C + X(X'X)^{-1})'(C + X(X'X)^{-1}) - (X'X)^{-1} \\ &= C'C + C'X(X'X)^{-1} + (X'X)^{-1}X'C + (X'X)^{-1}X'X(X'X)^{-1} - (X'X)^{-1} \\ &= C'C \succeq 0. \end{aligned}$$

□

- The proof of why $\hat{\beta}^{OLS}$ is the best among all unbiased estimators can be seen from the Cramér–Rao Lower Bound. In CRLB, a parametric model is assumed, and there is a universal/theoretical lower bound for the variance of any unbiased estimators. However, the CRLB argument per se merely concludes that any unbiased estimator cannot do better, but does not provide whether there exists such an estimator, or how to find/construct such an estimator.

Remark. Gauss-Markov theorem and Cramér–Rao Lower Bound are two famous theorems in Econometrics and classic statistics. However, their importance and their relationship with modern statistics/-machine learning can be viewed from the bias-variance trade-off.

4.5.2 Bias-variance Trade-off

Proposition 5. *For a general data-generating process (not necessarily to be linear):*

$$y = f(x) + \varepsilon,$$

we could get an estimator $\hat{f}(\cdot)$ of $f(\cdot)$ from data $\{(x_i, y_i)\}_{i=1}^n$. For any x, one could see that:

$$\begin{aligned} \mathbb{E}(\hat{f}(x) - y)^2 &= MSE(\hat{f}(x)) + \sigma^2 \\ &= \text{Var}(\hat{f}(x)) + [Bias\hat{f}(x)]^2 + \sigma^2. \end{aligned} \tag{4.35}$$

where the mean squared error is defined as: $MSE(\hat{f}(x)) := \mathbb{E}(\hat{f}(x) - f(x))^2$.

Proof.

$$\begin{aligned} (\hat{f}(x) - y)^2 &= (\hat{f}(x) - f(x) - \varepsilon)^2 \\ &= (\hat{f}(x) - f(x))^2 + \varepsilon^2 - 2\varepsilon(\hat{f}(x) - f(x)), \end{aligned}$$

Taking expectation, we have $\mathbb{E}(\hat{f}(x) - y)^2 = \mathbb{E}(\hat{f}(x) - f(x))^2 + \sigma^2$, since $\mathbb{E}\varepsilon(\hat{f}(x) - f(x)) = 0$. Furthermore,

$$\begin{aligned} (\hat{f}(x) - f(x))^2 &= (\hat{f}(x) - \mathbb{E}\hat{f}(x) + \mathbb{E}\hat{f}(x) - f(x))^2 \\ &= (\hat{f}(x) - \mathbb{E}\hat{f}(x))^2 + (\mathbb{E}\hat{f}(x) - f(x))^2 + 2(\hat{f}(x) - \mathbb{E}\hat{f}(x))(\mathbb{E}\hat{f}(x) - f(x)) \end{aligned}$$

Taking expectation for both sides, note that $\mathbb{E}(\hat{f}(x) - \mathbb{E}\hat{f}(x))^2 = \text{Var}(\hat{f}(x))$, and $(\mathbb{E}\hat{f}(x) - f(x))^2 = (\text{Bias}(\hat{f}(x)))^2$, $\mathbb{E}2(\hat{f}(x) - \mathbb{E}\hat{f}(x))(\mathbb{E}\hat{f}(x) - f(x)) = 2(\mathbb{E}\hat{f}(x) - f(x))\mathbb{E}(\hat{f}(x) - \mathbb{E}\hat{f}(x)) = 0$. □

- In 4.35, the σ^2 is the irreducible error, since the distribution of ε is not modelled. Thus, it cannot be reduced by the choice of $\hat{f}$.
- The $\mathbb{E}(\hat{f}(x) - y)^2$, called MSE (mean squared error), is a widely used measure in regression problem of how good an estimate $\hat{f}$ is.[23]

[23] In classification problem, misclassification rate/Gini index/cross entropy serve as the measure.

- It is basically said that, apart from the irreducible error, the MSE can be separated into two parts - (1) the variance, and (2) the square of the bias. In classic statistics, researcher focus more on the unbiased estimators. For unbiased estimators, one can see that the MSE is equal to the variance of the estimator, and thus, an estimator with a smaller variance is preferred. In this sense, Gauss-Markov theorem said that the OLS estimator provide the estimator with the minimum variance, while the CRLB gives a lower bound of the variance for any unbiased estimator.

- However, {unbiased estimator} $\subset$ {all estimator}. Thus, the best unbiased estimator may not be the best (in the sense of MSE) among all estimators. The intuition is that, if we allow the estimator to have a little bias, which makes it more flexible, there might exist a structure to make the variance decrease more.

- In my opinion, the trend that people get more interested in biased estimators is also due to the ability of computers. In the new era when one has a large sample and has the ability to deal with it (it cannot be imagined without computers), unbiasedness is not that important anymore. Instead, consistency takes the place. The explanation of consistency is that one could be sure that the estimator is close to the true parameter with high probability if the sample size is large enough.

4.5.3 Shrinkage Methods - Ridge and Lasso

Originally, Ridge and Lasso are introduced to deal with the colinearity, *i.e.* , $X'X$ is not full-rank ($\text{rank}(X'X) < p$), so its inverse $(X'X)^{-1}$ is not well-defined, and as a result, the OLS estimator $\hat{\beta} = (X'X)^{-1}X'Y$ does not exist.

In Ridge regression, a penalty term is added to the objective function. Now the estimator $\hat{\beta}^{\text{Ridge}}$ is defined to be the solution to the following optimisation problem:

$$\hat{\beta}^{\text{Ridge}} = \arg\min_{\beta\in\mathbb{R}^{p+1}} \Big\{ \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{i=1}^{p} x_{ip}\beta_i)^2 + \lambda \sum_{i=1}^{p} \beta_i^2 \Big\}.$$

The λ is a hyper-parameter, which needs to be set in advance, typically via the cross-validation approach.

Note that the penalty term does not contain β_0^2, It follows from the similar argument in FWL theorem, if we use centred $Y = (y_1 - \overline{Y}, \ldots, y_n - \overline{Y})^\top$ and $X_i = (X_{i1} - \overline{X_i}, \ldots, X_{in} - \overline{X_i})$, the resulting estimator $\hat{\beta_1}, \ldots, \hat{\beta_p}$ will be the same.

For the rest of this section, we will ignore the intercept term β_0. Then the optimisation problem in matrix form is:

$$\hat{\beta}^{\text{Ridge}} = \arg\min_{\beta\in\mathbb{R}^p} (Y - X\beta)'(Y - X\beta) + \beta'\beta,$$

Thus,

$$\hat{\beta}^{\text{Ridge}} = (X'X + \lambda I_p)^{-1}X'Y.$$

- Even if $X'X$ is not full-rank, some value of λ will make $(X'X + \lambda I)$ full-rank, thus the Ridge estimator is well-defined.

- If all features are orthogonal (*i.e.* $X'X = I_p$), then $\hat{\beta}^{\text{Ridge}} = \frac{1}{1+\lambda}\hat{\beta}^{\text{OLS}}$, which implies $\mathbb{E}\hat{\beta}^{\text{Ridge}} \neq \beta$, *i.e.* , $\hat{\beta}^{\text{Ridge}}$ is a biased estimator.

- The variance of $\hat{f}(x) =$

4.6 Orthogonalisation

Orthogonalisation is the noun for the word 'orthogonalise'. The thoughts are that

4.7 Principle Component Regression

4.8 Weighted Least Square

4.9 X is not Full-rank - towards High-dimensional Model

It could be the case that X is not full-rank when the columns of X are linearly dependent. A simple example is that, $x^2 = 4x^1$, where $x^i = (x_1^i, x_2^i, \cdots, x_{n-1}^i, x_n^i)^\top$ is the i-th column of matrix X.

Then, $X'X$ is singular ($\text{rank}(X'X) < p+1$), resulting in ill-defined $(X'X)^{-1}$ as well as $\hat{\theta}^{\text{OLS}}$. The $\hat{\theta}^{\text{OLS}}$ does not uniquely exist and can be easily seen. In fact, suppose $\hat{\theta}^{\text{OLS}} = (\hat{\theta}_0^{\text{OLS}}, \hat{\theta}_1^{\text{OLS}}, \cdots, \hat{\theta}_n^{\text{OLS}})$ minimise the $g(\theta)$, then we could conclude that, for any $m \in \mathbb{R}$, another estimator $\check{\theta} := (\hat{\theta}_0^{\text{OLS}}, \hat{\theta}_1^{\text{OLS}} - 4m, \hat{\theta}_2^{\text{OLS}} + m, \cdots, \hat{\theta}_n^{\text{OLS}})$ also minimises the $g(\theta)$.

However, the fitted value $\hat{Y} = X\hat{\theta}$ is still the projection of Y on the column space of X. The difference is that the column space of X is not $(p+1)$-dimensional, hence does not have a unique way to be represented. In practice, it is implied that there are redundant features (regressors), and we should remove them first before doing the regression.

It is common to see non-full-rank X in lots of applications, such as image analysis. In this case, p is even larger than n. The typical solution is to add more assumptions to the data-generating process, such as the sparsity: among all p regressors, only s (where $s \ll p$) number of them are functioning (only s number of $\theta_i, i = 0, \cdots, p$ are not 0), but we don't know which. To simultaneously estimate which regressor is functioning and its coefficient, researchers have found models with regularization (especially Lasso regression) work perfectly. We will talk about it later (if I have time lol).

References

Hansen, B. (2022). *Econometrics*. Princeton University Press.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.