Least Squares Estimation-Finite-Sample Properties

Ping Yu

School of Economics and Finance
The University of Hong Kong

September 2014

1. Terminology and Assumptions
2. Goodness of Fit
3. Bias and Variance
4. The Gauss-Markov Theorem
5. Multicollinearity
6. Hypothesis Testing: An Introduction
7. LSE as a MLE

Terminology and Assumptions

Terminology

$$\begin{aligned} y &= \mathbf{x}'\beta + u, \\ E[u|\mathbf{x}] &= 0. \end{aligned}$$

- u is called the **error term**, **disturbance** or **unobservable**.

y	$\mathbf{x}$
Dependent variable	Independent Variable
Explained variable	Explanatory variable
Response variable	Control (Stimulus) variable
Predicted variable	Predictor variable
Regressand	Regressor
LHS variable	RHS variable
Endogenous variable	Exogenous variable
.	Covariate
.	Conditioning variable

Table 1: Terminology for Linear Regression

Assumptions

- We maintain the following assumptions in this chapter.
- **Assumption OLS.0** (random sampling): $(y_i, \mathbf{x}_i)$, $i = 1, \cdots, n$, are independent and identically distributed (i.i.d.).
- **Assumption OLS.1** (full rank): $\text{rank}(\mathbf{X}) = k$.
- **Assumption OLS.2** (first moment): $E[y|\mathbf{x}] = \mathbf{x}'\beta$.
- **Assumption OLS.3** (second moment): $E[u^2] < \infty$.
- **Assumption OLS.3**$'$ (homoskedasticity): $E[u^2|\mathbf{x}] = \sigma^2$.

Discussion

- Assumption OLS.2 is equivalent to $y = \mathbf{x}'\beta + u$ (linear in parameters) plus $E[u|\mathbf{x}] = 0$ (zero conditional mean).
- To study the finite-sample properties of the LSE, such as the unbiasedness, we always assume Assumption OLS.2, i.e., the model is linear regression.[1]
- Assumption OLS.3$'$ is stronger than Assumption OLS.3.
- The linear regression model under Assumption OLS.3$'$ is called the **homoskedastic linear regression model**,

$$\begin{aligned} y &= \mathbf{x}'\beta + u, \\ E[u|\mathbf{x}] &= 0, \\ E[u^2|\mathbf{x}] &= \sigma^2. \end{aligned}$$

- If $E[u^2|\mathbf{x}] = \sigma^2(\mathbf{x})$ depends on $\mathbf{x}$ we say u is **heteroskedastic**.

[1] For large-sample properties such as consistency, we require only weaker assumptions.

Goodness of Fit

Residual and SER

- Express
$$y_i = \widehat{y}_i + \widehat{u}_i, \tag{1}$$
where $\widehat{y}_i = \mathbf{x}_i'\widehat{\beta}$ is the **predicted value**, and $\widehat{u}_i = y_i - \widehat{y}_i$ is the **residual**.[2]
- Often, the error variance $\sigma^2 = E[u^2]$ is also a parameter of interest. It measures the variation in the "unexplained" part of the regression.
- Its method of moments (MoM) estimator is the sample average of the squared residuals,
$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i^2 = \frac{1}{n}\widehat{\mathbf{u}}'\widehat{\mathbf{u}}.$$
- An alternative estimator uses the formula
$$s^2 = \frac{1}{n-k}\sum_{i=1}^{n}\widehat{u}_i^2 = \frac{1}{n-k}\widehat{\mathbf{u}}'\widehat{\mathbf{u}}.$$
This estimator adjusts the **degree of freedom** (df) of $\widehat{\mathbf{u}}$.

[2] $\widehat{u}_i$ is different from u_i. The later is unobservable while the former is a by-product of OLS estimation.

Coefficient of Determination

- If $\mathbf{X}$ includes a column of ones, $\mathbf{1}'\widehat{\mathbf{u}} = \sum_{i=1}^{n}\widehat{u}_i = 0$, so $\overline{y} = \overline{\widehat{y}}$.
- Subtracting $\overline{y}$ from both sides of (1), we have

$$\widetilde{y}_i \equiv y_i - \overline{y} = \widehat{y}_i - \overline{y} + \widehat{u}_i \equiv \widetilde{\widehat{y}}_i + \widehat{u}_i$$

- Since $\widetilde{\widehat{\mathbf{y}}}'\widehat{\mathbf{u}} = \widehat{\mathbf{y}}'\widehat{\mathbf{u}} - \overline{y}\cdot\mathbf{1}'\widehat{\mathbf{u}} = \widehat{\beta}\left(\mathbf{X}'\widehat{\mathbf{u}}\right) - \overline{y}\cdot\mathbf{1}'\widehat{\mathbf{u}} = 0$,

$$SST \equiv \left\|\widetilde{\mathbf{y}}\right\|^2 = \widetilde{\mathbf{y}}'\widetilde{\mathbf{y}} = \left\|\widetilde{\widehat{\mathbf{y}}}\right\|^2 + 2\widetilde{\widehat{\mathbf{y}}}'\widehat{\mathbf{u}} + \left\|\widehat{\mathbf{u}}\right\|^2 = \left\|\widetilde{\widehat{\mathbf{y}}}\right\|^2 + \left\|\widehat{\mathbf{u}}\right\|^2 \equiv SSE + SSR, \qquad (2)$$

  where SST, SSE and SSR mean the *total sum of squares*, the *explained sum of squares*, and the *residual sum of squares* (or the sum of squared residuals), respectively.
- Dividing SST on both sides of (2),[3] we have

$$1 = \frac{SSE}{SST} + \frac{SSR}{SST}.$$

- The R**-squared** of the regression, sometimes called the **coefficient of determination**, is defined as

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} = 1 - \frac{\widehat{\sigma}^2}{\widehat{\sigma}_y^2}.$$

[3] When can we conduct this operation, i.e., $SST \neq 0$?

More on R^2

- R^2 is defined only if $\mathbf{x}$ includes a constant.
- It is usually interpreted as the fraction of the sample variation in y that is explained by (nonconstant) $\mathbf{x}$.
- When there is no constant term in $\mathbf{x}_i$, we need to define so-called **uncentered** R^2, denoted as R_u^2,

$$R_u^2 = \frac{\widehat{\mathbf{y}}'\widehat{\mathbf{y}}}{\mathbf{y}'\mathbf{y}}.$$

- R^2 can also be treated as an estimator of

$$\rho^2 = 1 - \sigma^2/\sigma_y^2.$$

- It is often useful in algebraic manipulation of some statistics.
- An alternative estimator of ρ^2 proposed by Henri Theil (1924-2000) called **adjusted** R**-squared** or "R-bar-squared" is

$$\overline{R}^2 = 1 - \frac{s^2}{\widetilde{\sigma}_y^2} = 1 - (1 - R^2)\frac{n-1}{n-k} \le R^2,$$

where $\widetilde{\sigma}_y^2 = \widetilde{\mathbf{y}}'\widetilde{\mathbf{y}}/(n-1)$.

- $\overline{R}^2$ adjusts the degrees of freedom in the numerator and denominator of R^2.

Degree of Freedom

- Why called "degree of freedom"?
- Roughly speaking, the degree of freedom is the dimension of the space where a vector can stay, or how "freely" a vector can move.
- For example, $\widehat{\mathbf{u}}$, as a n-dimensional vector, can only stay in a subspace with dimension $n-k$.
- Why? This is because $\mathbf{X}'\widehat{\mathbf{u}}=0$, so k constraints are imposed on $\widehat{\mathbf{u}}$, and $\widehat{\mathbf{u}}$ cannot move completely freely and loses k degree of freedom.
- Similarly, the degree of freedom of $\widetilde{\mathbf{y}}$ is $n-1$. Figure 1 illustrates why the degree of freedom of $\widetilde{\mathbf{y}}$ is $n-1$ when $n=2$.
- Table 2 summarizes the degrees of freedom for the three terms in (2).

Variation	Notation	df
SSE	$\widehat{\widetilde{\mathbf{y}}}'\widehat{\widetilde{\mathbf{y}}}$	$k-1$
SSR	$\widehat{\mathbf{u}}'\widehat{\mathbf{u}}$	$n-k$
SST	$\widetilde{\mathbf{y}}'\widetilde{\mathbf{y}}$	$n-1$

Table 2: Degrees of Freedom for Three Variations

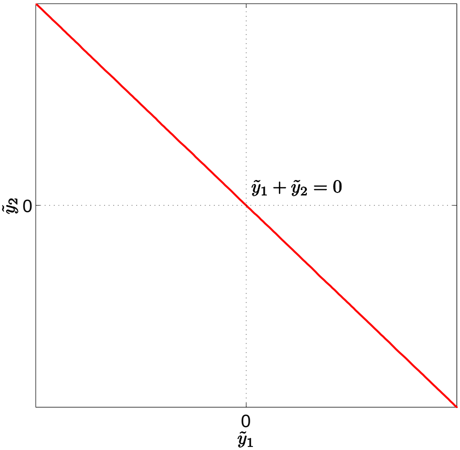

Figure: Although $\dim(\widetilde{\mathbf{y}}) = 2$, $\text{df}(\widetilde{\mathbf{y}}) = 1$, where $\widetilde{\mathbf{y}} = (\widetilde{y}_1, \widetilde{y}_2)$

Bias and Variance

Unbiasedness of the LSE

- Assumption OLS.2 implies that

$$y = \mathbf{x}'\beta + u, E[u|\mathbf{x}] = 0.$$

- Then

$$E[\mathbf{u}|\mathbf{X}] = \begin{pmatrix} \vdots \\ E[u_i|\mathbf{X}] \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ E[u_i|\mathbf{x}_i] \\ \vdots \end{pmatrix} = \mathbf{0},$$

where the second equality is from the assumption of independent sampling (Assumption OLS.0).

- Now,

$$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{u}) = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u},$$

so

$$E\left[\widehat{\beta} - \beta|\mathbf{X}\right] = E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}\right] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\mathbf{u}|\mathbf{X}] = \mathbf{0},$$

i.e., $\widehat{\beta}$ is unbiased.

Variance of the LSE

- $$\begin{aligned} Var\left(\widehat{\beta}|\mathbf{X}\right) &= Var\left(\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}\right) \\ &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'Var\left(\mathbf{u}|\mathbf{X}\right)\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\ &\equiv \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}. \end{aligned}$$

- Note that

$$Var\left(u_i|\mathbf{X}\right) = Var\left(u_i|\mathbf{x}_i\right) = E\left[u_i^2|\mathbf{x}_i\right] - E\left[u_i|\mathbf{x}_i\right]^2 = E\left[u_i^2|\mathbf{x}_i\right] \equiv \sigma_i^2,$$

and

$$\begin{aligned} Cov\left(u_i,u_j|\mathbf{X}\right) &= E\left[u_iu_j|\mathbf{X}\right] - E\left[u_i|\mathbf{X}\right]E\left[u_j|\mathbf{X}\right] \\ &= E\left[u_iu_j|\mathbf{x}_i,\mathbf{x}_j\right] - E\left[u_i|\mathbf{x}_i\right]E\left[u_j|\mathbf{x}_j\right] \\ &= E\left[u_i|\mathbf{x}_i\right]E\left[u_j|\mathbf{x}_j\right] - E\left[u_i|\mathbf{x}_i\right]E\left[u_j|\mathbf{x}_j\right] = 0, \end{aligned}$$

so **D** is a diagonal matrix:

$$\mathbf{D} = \text{diag}\left(\sigma_1^2,\cdots,\sigma_n^2\right).$$

continue...

- It is useful to note that
$$\mathbf{X}'\mathbf{D}\mathbf{X}=\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\sigma_i^2.$$
- In the homoskedastic case, $\sigma_i^2=\sigma^2$ and $\mathbf{D}=\sigma^2\mathbf{I}_n$, so $\mathbf{X}'\mathbf{D}\mathbf{X}=\sigma^2\mathbf{X}'\mathbf{X}$, and
$$Var\left(\widehat{\beta}|\mathbf{X}\right)=\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}.$$
- You are asked to show that
$$Var\left(\widehat{\beta}_j|\mathbf{X}\right)=\sum_{i=1}^{n}w_{ij}\sigma_i^2/SSR_j, j=1,\cdots,k,$$
where $w_{ij}>0$, $\sum_{i=1}^{n}w_{ij}=1$, and SSR_j is the SSR in the regression of x_j on all other regressors.
- So under homoskedasticity,
$$Var\left(\widehat{\beta}_j|\mathbf{X}\right)=\sigma^2/SSR_j=\sigma^2/\left[SST_j(1-R_j^2)\right], j=1,\cdots,k,$$
(why?), where SST_j is the SST of x_j, and R_j^2 is the R-squared from the simple regression of x_j on the remaining regressors (which includes an intercept).

Bias of $\widehat{\sigma}^2$

- Recall that $\widehat{\mathbf{u}} = \mathbf{Mu}$, where we abbreviate $\mathbf{M_X}$ as $\mathbf{M}$, so by the properties of projection matrices and the trace operator, we have

$$\widehat{\sigma}^2 = \frac{1}{n}\widehat{\mathbf{u}}'\widehat{\mathbf{u}} = \frac{1}{n}\mathbf{u}'\mathbf{MMu} = \frac{1}{n}\mathbf{u}'\mathbf{Mu} = \frac{1}{n}\text{tr}\left(\mathbf{u}'\mathbf{Mu}\right) = \frac{1}{n}\text{tr}\left(\mathbf{Muu}'\right).$$

- Then

$$E\left[\widehat{\sigma}^2\middle|\mathbf{X}\right] = \frac{1}{n}\text{tr}\left(E\left[\mathbf{Muu}'|\mathbf{X}\right]\right) = \frac{1}{n}\text{tr}\left(\mathbf{M}E\left[\mathbf{uu}'|\mathbf{X}\right]\right) = \frac{1}{n}\text{tr}\left(\mathbf{MD}\right).$$

- In the homoskedastic case, $\mathbf{D} = \sigma^2\mathbf{I}_n$, so

$$E\left[\widehat{\sigma}^2\middle|\mathbf{X}\right] = \frac{1}{n}\text{tr}\left(\mathbf{M}\sigma^2\right) = \sigma^2\left(\frac{n-k}{n}\right).$$

  Thus $\widehat{\sigma}^2$ underestimates σ^2.

- Alternatively, $s^2 = \frac{1}{n-k}\widehat{\mathbf{u}}'\widehat{\mathbf{u}}$ is unbiased for σ^2. This is the justification for the common preference of s^2 over $\widehat{\sigma}^2$ in empirical practice.
- However, this estimator is only unbiased in the special case of the homoskedastic linear regression model. It is not unbiased in the absence of homoskedasticity or in the projection model.

The Gauss-Markov Theorem

The Gauss-Markov Theorem

- The LSE has some optimality properties among a restricted class of estimators in a restricted class of models.
- The model is restricted to be the homoskedastic linear regression model, and the class of estimators are restricted to be linear unbiased. Here, "linear" means the estimator is a linear function of $\mathbf{y}$.
- In other words, the estimator, say, $\widetilde{\beta}$, can be written as

$$\widetilde{\beta} = \mathbf{A}'\mathbf{y} = \mathbf{A}'(\mathbf{X}\beta + \mathbf{u}) = \mathbf{A}'\mathbf{X}\beta + \mathbf{A}'\mathbf{u},$$

  where $\mathbf{A}$ is any $n \times k$ matrix of $\mathbf{X}$.
- Unbiasedness implies that $E[\widetilde{\beta}|\mathbf{X}] = E[\mathbf{A}'\mathbf{y}|\mathbf{X}] = \mathbf{A}'\mathbf{X}\beta = \beta$ or $\mathbf{A}'\mathbf{X} = \mathbf{I}_k$.
- In this case, $\widetilde{\beta} = \beta + \mathbf{A}'\mathbf{u}$, so under homoskedasticity,

$$Var\left(\widetilde{\beta}|\mathbf{X}\right) = \mathbf{A}'Var(\mathbf{u}|\mathbf{X})\mathbf{A} = \mathbf{A}'\mathbf{A}\sigma^2.$$

- The Gauss-Markov Theorem states that the best choice of $\mathbf{A}'$ is $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ in the sense that this choice of $\mathbf{A}$ achieves the smallest variance.

continue...

Theorem

In the homoskedastic linear regression model, the best (minimum-variance) linear unbiased estimator (BLUE) is the LSE.

Proof.

Given that the variance of the LSE is $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$ and that of $\widetilde{\beta}$ is $\mathbf{A}'\mathbf{A}\sigma^2$. It is sufficient to show that $\mathbf{A}'\mathbf{A}-(\mathbf{X}'\mathbf{X})^{-1}\geq 0$. Set $\mathbf{C}=\mathbf{A}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$. Note that $\mathbf{X}'\mathbf{C}=0$. Then we calculate that

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

□

Limitation and Extension of the Gauss-Markov Theorem

- The scope of the Gauss-Markov Theorem is quite limited given that it requires the class of estimators to be linear unbiased and the model to be homoskedastic.
- This leaves open the possibility that a nonlinear or biased estimator could have lower mean squared error (MSE) than the LSE in a heteroskedastic model.
- MSE: for simplicity, suppose $\dim(\beta) = 1$; then

$$\text{MSE}\left(\widetilde{\beta}\right) = E\left[\left(\widetilde{\beta} - \beta\right)^2\right] \overset{?}{=} Var\left(\widetilde{\beta}\right) + \text{Bias}\left(\widetilde{\beta}\right)^2.$$

- To exclude such possibilities, we need asymptotic (or large-sample) arguments.
- Chamberlain (1987) shows that in the model $y = \mathbf{x}'\beta + u$, if the *only* available information is $E[\mathbf{x}u] = 0$ or ($E[u|\mathbf{x}] = 0$ and $E[u^2|\mathbf{x}] = \sigma^2$), then among *all* estimators, the LSE achieves the lowest asymptotic MSE.

Multicollinearity

Multicollinearity

- If $\text{rank}(\mathbf{X}'\mathbf{X}) < k$, then $\widehat{\beta}$ is not uniquely defined. This is called **strict** (or **exact**) **multicollinearity**.
- This happens when the columns of **X** are linearly dependent, i.e., there is some $\alpha \neq \mathbf{0}$ such that $\mathbf{X}\alpha = \mathbf{0}$.
- Most commonly, this arises when sets of regressors are included which are identically related. For example, if **X** includes a column of ones and both dummies for male and female.
- When this happens, the applied researcher quickly discovers the error as the statistical software will be unable to construct $(\mathbf{X}'\mathbf{X})^{-1}$. Since the error is discovered quickly, this is rarely a problem for applied econometric practice.
- The more relevant issue is **near multicollinearity**, which is often called "multicollinearity" for brevity. This is the situation when the $\mathbf{X}'\mathbf{X}$ matrix is near singular, or when the columns of **X** are *close* to be linearly dependent.
- This definition is not precise, because we have not said what it means for a matrix to be "near singular". This is one difficulty with the definition and interpretation of multicollinearity.

continue...

- One implication of near singularity of matrices is that the numerical reliability of the calculations is reduced.
- A more relevant implication of near multicollinearity is that individual coefficient estimates will be imprecise.
- We can see this most simply in a homoskedastic linear regression model with two regressors
$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + u_i,$$
and
$$\frac{1}{n}\mathbf{X}'\mathbf{X} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$
- In this case,
$$Var\left(\widehat{\beta}|\mathbf{X}\right) = \frac{\sigma^2}{n}\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}^{-1} = \frac{\sigma^2}{n(1-\rho^2)}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}.$$
- The correlation indexes collinearity, since as ρ approaches 1 the matrix becomes singular.
- $\sigma^2/n(1-\rho^2) \to \infty$ as $\rho \to 1$. Thus the more "collinear" are the regressors, the worse the precision of the individual coefficient estimates.

continue...

- In the general model
$$y_i = x_{1i}\beta_1 + \mathbf{x}_{2i}'\beta_2 + u_i,$$
recall that
$$Var\left(\widehat{\beta}_1|\mathbf{X}\right) = \frac{\sigma^2}{SST_1(1-R_1^2)}. \quad (3)$$
- Because the R-squared measures goodness of fit, a value of R_1^2 close to one indicated that $\mathbf{x}_2$ explains much of the variation in x_1 in the sample. This means that x_1 and $\mathbf{x}_2$ are highly correlated. When R_1^2 approaches 1, the variance of $\widehat{\beta}_1$ explodes.
- $1/(1-R_1^2)$ is often termed as the **variance inflation factor** (VIF). Usually, a VIF larger than 10 should arise our attention.
- **Intuition**: β_1 means the effect on y as x_1 changes one unit, holding $\mathbf{x}_2$ fixed. When x_1 and $\mathbf{x}_2$ are highly correlated, you cannot change x_1 while holding $\mathbf{x}_2$ fixed, so β_1 cannot be estimated precisely.
- Multicollinearity is a small-sample problem. As larger and larger data sets are available nowadays, i.e., $n >> k$, it is seldom a problem in current econometric practice.

Hypothesis Testing: An Introduction

Basic Concepts

- null hypothesis, alternative hypothesis
- point hypothesis, one-sided hypothesis, two-sided hypothesis
  - We consider only the point *null* hypothesis in this course.
- simple hypothesis, composite hypothesis
- acceptance region and rejection or critical region
- test statistic, critical value
- type I error and type II error
- size and power
- significance level, statistically (in)significant

Summary

- One hypothesis testing includes the following steps.

1. specify the null and alternative.
2. construct the test statistic.
3. derive the distribution of the test statistic under the null.
4. determine the decision rule (acceptance and rejection regions) by specifying a level of significance.
5. study the power of the test.

- Step 2, 3 and 5 are key since step 1 and 4 are usually trivial.
- Of course, in some cases, how to specify the null and the alternative is also subtle, and in some cases, the critical value is not easy to determine if the asymptotic distribution is complicated.

LSE as a MLE

LSE as a MLE

- Another motivation for the LSE can be obtained from the **normal regression model**:
  Assumption OLS.4 (normality): $u|\mathbf{x} \sim N(0,\sigma^2)$ or $\mathbf{u}|\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_n\sigma^2)$.
- That is, the error u_i is independent of $\mathbf{x}_i$ and has the distribution $N(0,\sigma^2)$, which obviously implies $E[u|\mathbf{x}] = 0$ and $E[u^2|\mathbf{x}] = \sigma^2$.
- The average log likelihood is

$$\begin{aligned} l_n\left(\beta,\sigma^2\right) &= \frac{1}{n}\sum_{i=1}^{n}\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_i-\mathbf{x}_i'\beta)^2}{2\sigma^2}\right)\right) \\ &= -\frac{1}{2}\log(2\pi)-\frac{1}{2}\log\left(\sigma^2\right)-\frac{1}{n}\sum_{i=1}^{n}\frac{(y_i-\mathbf{x}_i'\beta)^2}{2\sigma^2}, \end{aligned}$$

  so $\widehat{\beta}_{MLE} = \widehat{\beta}_{LSE}$.
- It is not hard to show that $\widehat{\beta}-\beta|\mathbf{X} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X} \sim N\left(\mathbf{0},\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right)$.
- But recall the trade-off between efficiency and robustness, which can be applied here.
- Anyway, this is part of the classical theory in least squares estimation.
- We will neglect this section and proceed to the asymptotic theory of the LSE which is more robust and does not require the normality assumption.